An Exploration of the Relationship between the Partisan-Business Cycle and Economic Inequality within Developed Economies

Richard O'Doherty
Technological University Dublin

Follow this and additional works at: https://arrow.tudublin.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
An Exploration of the Relationship between the Partisan-Business Cycle and Economic Inequality within Developed Economies

Richard Thomas O'Doherty PgD BA

A dissertation submitted in partial fulfilment of the requirements of

Dublin Institute of Technology for the degree of

Master of Science Computing (Data Analytics)

March 2016
Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute’s guidelines for ethics in research.

Signature: __________________________

Richard Thomas O'Doherty

Date: 6th March 2016
Abstract

Recent contributions to the study of inequality have provided strong evidence towards the presence of an established trend, over several decades, of growing economic inequality (with a particular focus on distribution within their tails; i.e. top 10%, 1%) across countries with developed economies and indications of similar trends across developing economies. While the causality and influencing factors to these trends has widely been discussed, and has range from declining domestic growth rates as economies move towards high mass consumption states to globalisation, political decision making and policy application been referred to as both contributory or an instrument for dampening such trends, through the application of Partisan Business Cycles. However contemporary studies have indicated a declining correlation between political orientation of national legislatures or executive branches and domestic economic and financial performance (which by extension, through it strong correlation with growth, inequality). However little exploratory research has been done on the deviation between the observed indicators of inequality and its trend line (residual). As such this paper has explored the relationships between the components of the Partisan Business Cycles and this deviation across a set of advanced economies of a time period between 1975 to 2010. The paper additionally considered political structural and situational constrains, such as governmental systems, minority governments, etc... , on their decision making as variables for consideration.

The analysis was performed using Auto-regression Integrated Moving Average and Multi-Linear Regression modelling. Which strongly indicates that neither the Partisan Business Cycle nor the political situational and structures are explanatory indicators of the variations in inequality from the trend. Nor can previous observed deviations provide strong indication of a future patterns in the deviations from the trend. As such this paper finds no distinguished pattern within these residuals suggesting that such observations can be categorised as white noise around these trends lines or inequality.

**Keywords:** Partisan Business Cycle, Economic Inequality, ARIMA, Regression, Data Quality, Null Handling, Predictive Value Imputation, Interpolation, SAS, VBA, SQL
# TABLE OF CONTENTS

1. **INTRODUCTION** .................................................................................................................. 1  
   1.1 BACKGROUND...................................................................................................................... 1  
   1.2 RESEARCH PROBLEM .......................................................................................................... 5  
   1.3 RESEARCH AIM AND OBJECTIVES ..................................................................................... 7  
   1.4 RESEARCH METHODOLOGY ............................................................................................... 9  
   1.5 SCOPE AND LIMITATIONS ............................................................................................... 10  
   1.6 ORGANISATION OF DISSERTATION .................................................................................. 11  

2. **LITERATURE REVIEW** ..................................................................................................... 13  
   2.1 BUSINESS UNDERSTANDING ........................................................................................... 13  
      2.1.1 PARTISAN-BUSINESS CYCLE .................................................................................. 15  
      2.1.2 PBC’S ECONOMIC AND FINANCIAL IMPACT ......................................................... 17  
      2.1.3 RETURNS AND INEQUALITY ..................................................................................... 20  
      2.1.4 INCOME DISTRIBUTION METRICS ......................................................................... 24  
      2.1.5 LEFT/RIGHT POLITICAL SPECTRUM ....................................................................... 30  
   2.2 CRITICAL EVALUATION OF LITERATURE ...................................................................... 31  
   2.3 RESEARCH QUESTION ....................................................................................................... 36  

3. **DESIGN AND METHODOLOGY** ......................................................................................... 37  
   3.1 CRISP-DM PROCESS MODEL ............................................................................................... 37  
   3.2 DATA UNDERSTANDING ..................................................................................................... 40  
      3.2.1 DEPENDANT VARIABLES ......................................................................................... 42  
      3.2.2 EXPLANATORY VARIABLES .................................................................................... 44  
   3.3 DATA QUALITY - DIMENSIONS AND DEVELOPMENT CYCLE ....................................... 51  
   3.4 DATA CLEANSING/TRANSFORMATION .......................................................................... 55  
      3.4.1 OUTLIERS AND STANDARDIZATION ....................................................................... 55
<table>
<thead>
<tr>
<th>FIGURE 1.1</th>
<th>GOOGLE INCOME/WEALTH INEQUALITY &amp; POLITICAL SEARCHES 2008-2015</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGURE 1.2</td>
<td>REP. OF IRELAND INVERSE PARETO INCOME DISTRIBUTION 1975-2009</td>
<td>6</td>
</tr>
<tr>
<td>FIGURE 1.3</td>
<td>REP. OF IRELAND TOP 10%'S SHARE OF NATIONAL INCOME 1975-2009</td>
<td>6</td>
</tr>
<tr>
<td>FIGURE 1.4</td>
<td>PARETO DISTRIBUTION CURVE</td>
<td>8</td>
</tr>
<tr>
<td>FIGURE 2.1</td>
<td>AVERAGE PERCENTAGE RATE OF PRICE INFLATION VS, YEARS OF SOCIALIST-LABOR PARTIES IN EXCECUTIVES 1960-69</td>
<td>13</td>
</tr>
<tr>
<td>FIGURE 2.2</td>
<td>AVERAGE PERCENTAGE RATE OF UNEMPLOYMENT VS, YEARS OF SOCIALIST-LABOR PARTIES IN EXCECUTIVES 1960-69</td>
<td>14</td>
</tr>
<tr>
<td>FIGURE 2.3</td>
<td>INFLATION VS. UNEMPLOYMENT (PHILIPS CURVE) &amp; INFLATION VS. INCOME</td>
<td>16</td>
</tr>
<tr>
<td>FIGURE 2.4</td>
<td>AVERAGE ANNUAL GROWTH IN REAL PER CAPITA VS, CHANGES IN ECONOMIC FREEDOM (BY QUANTILES)</td>
<td>20</td>
</tr>
</tbody>
</table>
FIGURE 2.5  PROPORTION OF POPULATION VS. INCOME (PARETO DISTRIBUTION) CURVE .............................................................. 26

FIGURE 2.6  INVERSE PARETO DISTRIBUTION COEFFICIENT ............... 27

FIGURE 2.7  SYSTEMS OF GOVERNMENT ............................................. 33

FIGURE 3.1  CRISP - DM PROCESS CYCLES ...................................... 37

FIGURE 3.2  DATA QUALITY DIMENSIONS ....................................... 50

FIGURE 3.3  DATA QUALITY PROJECT LIFE CYCLE .............................. 50

FIGURE 3.4  REP. OF IRELAND DEVIATION OF THE OBSERVED INVERSE PARETO COEFFICIENT FROM THE TREND LINE ......................... 67

FIGURE 4.1  MODULARISED PROGRAMME DESIGN STRUCTURE .......... 75

FIGURE 4.2  LOGICAL DATABASE MODLE ........................................ 84

FIGURE 4.3  INVERSE PARETO COEFFICIENTS' SCATTER PLOTS WITH 95% PREDICTION ELIPSE ................................................................. 102
FIGURE 4.4  TOP 10%’ SCATTER PLOTS WITH 95% PREDICTION ELIPSE
........................................................................................................................................... 104

FIGURE 4.5  INVERSE PARETO COEFFICIENT’S ARIMA TREND AND CORRELATION ANALYSIS .............................................................................................................. 107

FIGURE 4.6  INVERSE PARETO COEFFICIENT’S ARIMA(1,0,0) RESIDUAL NORMALITY DIAGNOSTIC .............................................................. 108

FIGURE 4.7  INVERSE PARETO COEFFICIENT’S ARIMA(0,0,1) RESIDUAL NORMALITY DIAGNOSTIC .............................................................. 109

FIGURE 4.8  INVERSE PARETO COEFFICIENT’S ARIMA(1,0,1) RESIDUAL NORMALITY DIAGNOSTIC .............................................................. 110

FIGURE 4.9  TOP 10% ARIMA TREND AND CORRELATION ANALYSIS ...111

FIGURE 4.10  INVERSE PARETO COEFFICIENT’S FIT CRITERIA ............ 113

FIGURE 4.11  INVERSE PARETO COEFFICIENT’S MALLOW’S CP AND HOCKER’S CRITICA .................................................................................. 114

FIGURE 4.12  TOP 10%’S FIT CRITERIA ...................................................... 118

FIGURE 4.13  TOP 10%’S MALLOW’S CP AND HOCKER’S CRITICA .......... 119
TABLE OF TABLES

TABLE 1.1 OECD INCOME INEQUALITY IN SELECTED COUNTRIES 1820-2000...................................................................................................................... 2

TABLE 1.2 UNITED NATIONS' WORLD ECONOMIC SITUATION AND PROSPECT REPORT OF 2010: DEVELOPED ECONOMIES (DEVELOPED MARKET ECONOMIES) ............................................................................................................................................... 11

TABLE 2.1 PARETO-LORENZ (A) VS. INVERSE PARETO-LORENZ (B) COEFFICIENTS .................................................................................................................. 29

TABLE 2.2 SCORES OF CENTRAL BANK AUTONOMY IN THE LATE 1980'S AND 2003 ................................................................................................................................................. 35

TABLE 3.1 WTID - MEASUREMENTS OF SCALE & TYPE ....................... 43

TABLE 3.2 DPI - MEASUREMENTS OF SCALE & TYPE ......................... 44

TABLE 3.3 EFW - MEASUREMENTS OF SCALE & TYPE ....................... 50

TABLE 3.4 WB - MEASUREMENTS OF SCALE & TYPE ......................... 51

TABLE 3.5 DATA QUALITY DIMENSIONS .............................................. 52
TABLE 4.1 VISUAL BASIC FOR APPLICATIONS (VBA) COMPONENTS ... 72

TABLE 4.2 DPI- VARIABLE DESCRIPTION ......................................................... 98

TABLE 4.3 WB- VARIABLE DESCRIPTION .......................................................... 99

TABLE 4.4 INVERSE PARETO COEFFICIENTS' FINAL SET OF EXPLANATORY VARIABLES FOR MODELING ................................................................. 100

TABLE 4.5 TOP 10% FINAL SET OF EXPLANATORY VARIABLES FOR MODELING ............................................................................................................. 103

TABLE 4.6 INVERSE PARETO COEFFICIENTS ARIMA DICKEY-FULLER UNIT ROOT TEST ........................................................................................................ 106

TABLE 4.7 INVERSE PARETO COEFFICIENTS ARIMA(1,0,0) MAXIMUM LIKELIHOOD ESTIMATION AND AIC/SBC TESTS ......................................................... 107

TABLE 4.8 INVERSE PARETO COEFFICIENTS ARIMA(0,0,1) MAXIMUM LIKELIHOOD ESTIMATION AND AIC/SBC TESTS ......................................................... 109

TABLE 4.9 INVERSE PARETO COEFFICIENTS ARIMA(1,0,1) MAXIMUM LIKELIHOOD ESTIMATION AND AIC/SBC TESTS ......................................................... 110
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.10</td>
<td>Top 10% ARIMA Dickey-Fuller Unit Root Test</td>
</tr>
<tr>
<td>Table 4.11</td>
<td>Inverse Pareto Coefficient's Mallows' Cp &amp; Hocking Critica</td>
</tr>
<tr>
<td>Table 4.12</td>
<td>Inverse Pareto Coefficient's Analysis of Variance</td>
</tr>
<tr>
<td>Table 4.13</td>
<td>Inverse Pareto Coefficient's Parameter Estimates</td>
</tr>
</tbody>
</table>
1. Introduction

1.1 Background

The question of disparities within the distribution of income within societies has re-emerged as a topical subject matter in the political and social discourse of developed states and the academic debates within the fields of economics and social sciences over the last few years. Such discussions have been highlighted by contributions from 'rock-star' economists such as Thomas Piketty (Chassany, 2015) of the Paris School of Economics and international organisations such as the Organisation of Economic Cooperation and Development (OECD), by reporting their findings of strong relationships (Levine & Stephen, 2005)) between economic growth and equality, with a specific distinct trend originating within the later decades of the twentieth century and continuing into the twenty-first (1970's - 2010's) of a growing wealth and income inequality across the modern societies with advanced economies.

Summing up his work in an article for the American Economic Review Thomas Piketty notes that from his analysis "the size of the gap between r and g, where r is the rate of return on capital and g the economy's growth rate, is one of the important forces that can account for the historical magnitude and variations in wealth inequality" (Piketty, 2015). He goes on to emphasize that in advanced economies when g is constrained to low or virtually stagnate level; i.e. steady-state (Center for the Advancement of Steady State Economy, 2015), a prolonged and widening 'r − g' spread over these decades has pronouncedly contributed to the skewing of the income distribution curve, partially during circumstance of shocks; i.e. recessions, deflation, wage decline, etc (Piketty, 2015). Given the general envelope like shape of the distribution curve, the spread will assume a multiplicative form where "unequal
returns on capital are a force for divergence that significantly amplifies and aggravates the effects of the inequality \( r > g \). Indeed, the difference \( r - g \) can be high for large fortunes without necessarily being high for the economy as a whole" (Piketty, 2014a).

Table 1.1 OECD Income Inequality in Selected Countries 1820-2000

<table>
<thead>
<tr>
<th>Year</th>
<th>Western Europe (WE)</th>
<th>Eastern Europe (EE)</th>
<th>Western Offshoots (WO)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GBR</td>
<td>NLD</td>
<td>FRA</td>
</tr>
<tr>
<td>1820</td>
<td>59</td>
<td>56</td>
<td>59</td>
</tr>
<tr>
<td>1830</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1840</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1850</td>
<td>[43]</td>
<td>[48]</td>
<td>[54]</td>
</tr>
<tr>
<td>1860</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1870</td>
<td>[49]</td>
<td>[57]</td>
<td>[56]</td>
</tr>
<tr>
<td>1880</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1890</td>
<td>[37]</td>
<td>[42]</td>
<td>[48]</td>
</tr>
<tr>
<td>1900</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1910</td>
<td>42</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>1920</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1930</td>
<td>43</td>
<td>42</td>
<td>62</td>
</tr>
<tr>
<td>1940</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1950</td>
<td>30</td>
<td>35</td>
<td>58</td>
</tr>
<tr>
<td>1960</td>
<td>29</td>
<td>43</td>
<td>52</td>
</tr>
<tr>
<td>1970</td>
<td>29</td>
<td>36</td>
<td>45</td>
</tr>
<tr>
<td>1980</td>
<td>34</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>1990</td>
<td>39</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>2000</td>
<td>40</td>
<td>32</td>
<td>37</td>
</tr>
</tbody>
</table>

(How was Life?, 2014)

Opponents to this theory has noted concerns ranging from inconsistent use of terminology to criticism surrounding the quality of the data used in the analysis, to a point where the conclusions of such papers have been "undercut by a series of problems and errors. Some issues concern sourcing and definitional problems. Some numbers appear simply to be
constructed out of thin air" (Giles, 2014) with little technical details of the methodologies and techniques where they were applied or the justification for their application. However, such critical analysis have not curtailed the boarder findings of such papers, that there has been a general positive trend in indicators inequality (i.e. GINI, Pareto, Shares of Income, etc... ), nor restrained such subject matter from gaining popularity within the wider public conscience, which is evident by their increasing popularity as Google search terms over the last few years.

This increased interest in inequality has coincided with periods of greater economic uncertainty following the 2008 Liquidity Crisis and the subsequent period of fiscal reform and budgetary austerity within Europe and the wider global economy. These economic crises have been distinct in the level at which they have been politicised, with the swing of electorare preference to parties which were previously been marginal opposition parties or had not been established, as the percentage of such electorates which are un-employed or on lower incomes have increases (Visser, Lubbers, Kraaykamp, & Jaspers, 2014). However throughout this period of accelerating focus in inequality has seen a declining interest in Political Risk; i.e. the risk that returns on investment or economic growth could suffer as a result of domestic political changes or instability (Investopedia, 2015) has seen a decline which is indicated by a declining rate in its use as a Google search term (See Appendix 7.1).
The asymmetry in the trend vectors of these related fields highlight a disparity between the growing interest in the subject of inequality, in its economic sense, and the causal relationship between changes within the political environments of states and the overall economic health or growth within the domestic economy of the state, which is considered to be strongly correlated with income equality (van Zanden, 2014). Part of the logic behind this disconnect is the impact of regionalization of political and monetary authority with the growing proportionate of international market forces, which partially in Europe (Kayser, 2009), has played a significant role in curtailing the perceived perception of the impact nation state governments can influence the growth of their economies and the inflation which they experience.
1.2 Research Problem

The preceding decades to these crises coincided with a decline of the application of Partisan-Business Cycle (PCB) theories’, which theorises the impact of party political orientation on cyclical economic measures, in the analysis of developed markets in some academic circles of political science. Due to the popularisation of monetarist economic principles, which signalled a shift in governments' role in economic policy away from the optimisation of the demand-side of the economy (i.e. Growth through consummation that's facilitated by higher employment or wages) to a refocus towards its supply-side (i.e. Growth achieved by greater and more efficient production), and regionalization following the exposure of advanced economies to simultaneous rise in both unemployment and inflation (stagflation) during the 1970’s under Keynesian orientated policies (Heywood, 2002). As such a reassessment of the contribution that the PBC can offer our understanding of economic policy decision making and inequality is overdue.

As the overall trend in the increase of inequality, spanning the decades succeeding the late 1970’s, is already established few pager have questioned if the orientation of incumbent governments have had little or no affect on the long term trend can the PBC, through its observable elements (e.g. Political orientation, Governmental expenditure, credit/debt consumption, etc...), be used an explanatory component to the residual differences that are seen between this trend line and the reported observations.
To do so this paper looks to review the academic theories and findings that underpin the PCB theory and its affect on the performance of national economies, and by extension gauge
its capability to influence domestic income inequality through the application of data analytical techniques and methods.

1.3 Research Aim and Objectives

While the majority of other studies which investigated this subject matter have looks at the policy orientation of the incumbent government alone and as a single entity this paper will look to access the political environments in which such governments sits and look at the makeup of the governments at a more granular level. This will involve assesses the orientation of Executive and the Legislator which makes up the government, as well as it length in power, age, etc... The political environmental variables can be considered to fall into one of two groups; institutional or situational. The institutional factors will include variable like the electoral system in which the state adheres to (whether it be Proposal Representation, List, etc...) , the institutional system in which it functions (Parliamentary, Presidential or Semi-Presidential), etc... While the situational group will consist of statistics that depict makeup and strength of the government within these national institutions.

While Economic inequality is a heterogeneous concept it is generally surrogate to the distribution of income within the specific subject group, "not least due to the greater availability of data" (Staunton & O’Connor, 2015). To quantify a measurement of income inequality, in addition to the share of income allocated to the top 10% of the population within a country, the paper will also utilise the Pareto indicators and specifically its inverse (B). This indicator is derived from the conventional distribution F function which gives the proportion of a sample which is below or equal to a specified level of a variable (or in the case of B is above that level). Contrasting Income against population on a logarithmic scale a parameter α can be interpolated to give the slope of a fitted line at a point on the curve relative to the proportion of the population (χ). This slope can then be interpreted as a proxy measurement of the income distribution within a segment of a population (i.e. 0.1 - 1%) when
these variables are contrasted on the XY axis, where $\chi_0$ denotes an income level or percentile.

Figure 1.4 Pareto Distribution Curve

This enables the complexity of the distribution of income in its tails to be reduced to a single parameter. The higher the value of this parameter is, the thinner the tail of the frequency distribution is and by association the greater the inequality present within distribution (Cowell, 2011). Using this method this study will look at the distribution of income within the top 1% of the population to complement and contrast the findings regarding the share of national income accumulated by the top 10%.

The main objectives of this dissertation are as follows:

1. Investigate the current views and research conducted to date on PBC and Inequality
2. Identify a set of suitable explanatory variables
3. Determine correlations between the explanatory variables and against the dependent variables.
4. Confirm whether there is a relationship between the PCB and Inequality
5. Identify future areas of research to further knowledge within this field

1.4 Research Methodology

To undertake these objectives the paper will apply the Cross-Industry Standard Process for Data Mining (CRISP-DM) guidelines to this study where the contests allows. Following the 'Business Understanding' and 'Data Understanding' stages, where a Literature Review clarifies and addresses the economic and PCB theories with the variables which contribute to the data mining effort, datasets containing information on economic, political and inequality indicators from several international institutions; i.e. World Banks, Bank of International Settlement, World Top Income Database, etc... will be consolidated into a dimensional modelled database (Objective 1). As part of this data preparation stage the variables within the data sets will be profiled and alternative methods for the handling of outliers and NULL observations will be explored to identify a technique which best servers to cleanse the data and maintain the knowledge held within the data (Objective 2).

To explore the hypothesis, that PCB does influence domestic income inequality, this study will endeavour to identify the most suitable modelling methodology to quantitatively test this relationship. The first step of this modelling stage will test for and limit the quantity of collinearity between the explanatory variables (Objective 3). If these assumptions are proven to be true between the variables the study will look at modelling the relationship between the explanatory and dependent variables using suitable Regression models; i.e. Auto-Regressive Integrated Moving Average models, Multi-Linear Regression, etc... (Objective 4).
\[ Y_i = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_n X_{i,n} + \epsilon_i \]

Regression based models will be used to compute the coefficient parameters \((\beta_1, \beta_2, \ldots, \beta_n)\) of the explanatory Variables \((X_{i,1}, X_{i,2}, \ldots, X_{i,n})\) with reference to Dependent Variable \((Y_i)\). Following which the suitability of the model and the strength of the correlations/estimate parameters will be critically evaluated as indicators of the relationships between the PBC associated elements and Income Inequality (Objective 5).

### 1.5 Scope and Limitations

The scope of this study will be limited to the time period of 1975 - 2010. The selection of this time frame is partly due to the availability of data, has the reporting of several economic statistics and indices have not been digitally published or were not measured in the periods prior to this time frame. Given that the majority of the variables are macroeconomic indices which are either reported by governmental authorities or are compiled from annual reports by non-governmental organisations and bodies it is quite apparent that majority of these indicators are annual and as such this study will be based on annualised observations, resulting in 35 observations per variable per country.

Such a small number of observations will give rise to issues of collinearity (i.e. explanatory variables within a multiple regression model are highly correlated among themselves) or endogeneity (i.e. variables are correlated with the regression model's error term). And in general will give rise to the risk of over fitting when new observations are applied against a model based on this dataset. To limit such risks the project will look at the economies that share a similar classification of development as one observable grouped entity, as such increasing the sample population of 35 observations per variable by a factor. Where that factor is the number of countries that meet this criteria.
As the stage of economic development and political democraticization varies widely between states at over this period the study will be limited to the observations of developed Economies, to maximise the homogeneity between the observations. The classification of an economy as developed will be taken from the United Nations’ World Economic Situation and Prospect report of 2010 (United Nations, 2010). Any additional reduction of observations from this group will be due to a lack of available variable data.

1.6 Organisation of Dissertation

As the study will follow the CRISP-DM structure of an analytical study the chapter lay out will be constructed to mirror the stages of the process.

Chapter 2:

Business Understanding: This initial phase will consist of a literary review to identify existing studies within these fields and the analytical methods which are associated with them. Justification of the study’s objectives from an academic and theoretical perspective are made in this section and are then translated into a data mining definition.

Data Understanding: Sourced variable dataset undergo an initial examination to allow for familiarity with the frequencies and patterns within the data including the detection or elimination of variables based on these findings,
Chapter 3:

Data Preparation:  This stage incorporated the cleansing and enrichment of the variables' data which progressed from former stages the in preparation for the construction of the final Analytical Base Table.

Modelling: Tests of correlation and significance of the variables are generated. Following which model construction is undertaken upon the explanatory and dependent variables contained within the Analytical Base Table

Chapter 4:

Evaluation: The final constructed model is evaluated to ascertain if significant correlations are present between the explanatory and dependent variables. The evaluated findings from the model are then contracted with are initial research problem's objective to access if new insight are now provided by the model.

Chapter 5:

The final chapter will provide an overview of the study's experiment and its findings. Further areas of investigation and research will be highlighted in order to potentially improve on the results found.
2 Literature Review

2.1 Business Understanding

PBC is an offshoot of the Political-Business Cycle model, in which William Nordhaus observed, in his 1975 empirical case study of Democratic economies, a distinct ‘opportunistic’ tendencies of sitting governments, regardless of their position on the Left-Right Spectrum, to stimulate national economies in the months leading up to elections in order to alter electorate sentiment. Within the PBC theory this artificial stimulation differs distinctly depending on the incumbent’s leanings within the political Left-Right Spectrum and their representative constituencies. And when in government, political parties continue to legislate in favour of the underlining macroeconomic bias towards Growth or Inflation control which best addresses the priorities of their constituencies.

Figure 2.1 Average Percentage Rate of price Inflation vs, Years of Socialist-Labor Parties in Executives 1960-69
Such papers who's focus concerns the interplay between the policy making of public institutions and the performance of associated economies can be defined as been studies of the Political Economy. This Political Economy or the International Political Economy (IPE), as it is commonly referenced, was defined by John Ravenhill as "a subject matter whose central focus is the interrelationship between public and private power in the allocation of scarce resources" (Ravenhill, 2005) through their economic and economic linked activities in his 2005 book on the subject, entitled 'Global Political Economy'. According to IPE studies the interaction between national economies, which aggregated makes up a single global economic entity, are determined by not solely by markets forces but by the policies of nation-states. Such political pursuits, cooperation or competition between states, interact to create a framework of commercial relations within which economic forces operate. While
government policies define the rules that businesses within the Minor and Macro Economies must follow, these policies are in turn influenced by economic and technological forces that are the product of such commerce (Gilpin, 2001). Concluding in an interdependent relationship between political and economic forces, where a study of a change in one force subsequently requires an understanding of the impact on or from the other.

2.1.1 Partisan-Business Cycle

PBC model emerged primarily from the works of Douglas Hibbs (Hibbs, 1977) in which he suggested that economic policies of similar states differ significantly, with relation to their government’s positioning on the Left-Right Spectrum. With central-right or right-of-centre (Right-Wing) governments more lightly to emphasize lower inflation, transaction costs, and tax rates with a modestly balanced budgets in comparison to their counterparts on the left of the spectrum (Left-wing). Who's focus towards lower unemployment and greater income/wealth equalization (with larger state budgets) results in a more expansionary policy than right-wing governments (Schrage, 2002). The PBC categorizes that governmental economic policy agenda is due to the Ideological consistency of their constituency i.e. parties on the left of the spectrum tend to cater for the well-being of a electorate who's primacy source of income is from labour by targeting unemployment (Hibbs, 1977) where as right-wing governments prioritize price stability, viewed as an absolute good, to maintain capital's values (Halls, 1992).
In terms of the A.W. Phillips’ graphical interpretation of macroeconomics’ inverse relationship between the rate of unemployment and the rate of inflation (the Phillips Curve) Hibbs’ model ascertained that left-wing parties are more bias towards an employment (and growth) trade-off with inflation (Phillips, 1958), while right-wing parties are more accepting of the reversal of this trade-off.

Hibbs represents his model as a Loss function where a government j is denoted as Left (L) or Right (R) leaning where $\bar{\pi}^j$ is party j’s target rate of inflation, $\bar{U}^j$ is party j’s target unemployment rate. j’s preference between inflation and growth is represented by $\theta^j$, which is a weigh derived from the difference between inflation and j’s target inflation relative to the deviation of actual unemployment from target.

\[
L^j(U_t, \pi_t) = \frac{(U_t - \bar{U}^j)^2}{2} + \frac{\theta^j (\pi_t - \bar{\pi}^j)^2}{2}
\]

(Heywood, 2002)
Hibbs' study shows that this loss function is characterized by $U^L = < U^R$, $\theta^L = < \theta^R$ and $\tilde{\pi}^L \Rightarrow \tilde{\pi}^R$.

PBC has not been without its criticism. For in periods that economic shocks or stagflation it expressed a significant reduction in its ability to act as an exploitative model. Following the decline of Keynesian economic theory many Neo-liberal economists like Milton Friedman reinterpreted such cycles, and by associated theories, as more closely resembling fluctuations due to the questioning of its cyclical nature (F. Milton, A. Schwartz, 1993).

Yet PBC has continued to be a very applicable tool in determining rational-expectations of economic activity (Alesina, Roubini, & Cohen, 1997) of state actors. Political-business cycle theories have also continued to be widely applied to developing markets, in understanding the oscillation of foreign direct investment and the impact of governmental policy decision making (Block, Ferree, & Singh, 2003). Hibbs’ concluded in his study of Western liberal democracies that “macroeconomic outcomes... are not altogether endogenous to the economy, but obviously are influenced to a significant extent by Long- and Short-Term political choices. The real winners of elections are perhaps best determined by examining the consequences of partisan change rather than simply tallying the votes” (Hibbs, 1977).

2.1.2 PBC’s Economic and Financial Impact

Echoing the relations between $g$ and $\alpha$ that Thomas Piketty was later to conclude, in his 'Capital is Back' and 'Capital in the Twenty-First Century' titled published papers, Hidds observed that during periods of cyclical upswing no evidence that "business expansions bring a general inflation of profits which yield increases in the share of the national income going to capital" as present in real world observation but the contradictory, that growth in the latter half of the business cycle is strongly associated with a decline in share of the national income going contributed to capital returns and an increase in labour costs (Hibbs, 1977).
This relationship between macroeconomic cycles of a state and the return on investments of capital within its markets has been long established. Investment decision making is by definition forward looking and is concerned with developments in the underlying economy’s performance. Most notable of these key indicators are Gross Domestic Product (Growth) and Inflation (O’Loughlin & O’Brien, 2006). Increased GDP will drive demand and profit making within the domestic industries, which will feed back into the financial markets through dividend offering and equity prices. Likewise favorable (low) inflation rates reduce the erosion of capital returns and generally help to maintain the real returns from bond yields at lower coupon rates. However growth and inflation are often observed has having an inverse relationship which is illustrated in the previously discuss Philip’s Curve (Phillips, 1958). As such investors have noted the resulting affect of this trade off in the performance of their portfolios and a correlation to which political party holds office during these specific periods, most notably in the United States.

Jedrzej Bialkowski expresses that such an increased rise is 'conceivable' that in the face of governmental relocation across the Left-Right Spectrum investors adjust upward their perceived market risk on assets. As a rational response to the change of risk ratings and the potential profits from economic growth the stock market will be expected to offer higher incentives or returns (i.e. dividends) under left-wing governments (Bialkowski, Gottschalk, & Wisniewski, 2006).

Roland Fuss and Jana Lenz’s 2011 analysis of the PBC effect within the US during a presidential electoral year noted that a 1% point increase in the winning probability of the Democratic candidate can raise daily market risk on the U.S. stock market by 0.061%. A regression methodology was used to analytically identify this inverse reaction between Democratic surges in electorate opinion polls and financial market risk, as a reflection of economic uncertainty. Their finding suggest that previous studies on partisan politics of financial markets are conservative and that the market's negative reaction haves an under estimated feed back into further opinion polls (Füss & Lenz, 2011).
Mark Andreas Kayser's 2009 study of the effect of international business cycles on voter decisions, observed with the use of spatial time series analysis that with greater correlation of key economic behaviour "cycles in western Europe have converged to the point that they can be considered a single regional cycle" (Kayser, 2009). Such integration of international trade and business cycle has resulted in a 'comovement' of political preferences within regions and with it regional harmonization of partisan economic policies.

Piketty indicates that the variables that constrain the forces behind 'r > g' can logically be classified as either 'natural' and 'non-natural', where 'natural' can bet taken as the economic elements that contribute to growth and 'non-natural' as 'government intervention', driven by the "goals each society sets for itself and the particular challenges each country faces" (Piketty, 2014b). As indicated by the above it is rational to concluded that these 'goals' are subject to and influenced by the partisan orientation of the government of the day and specifically their views on intervention or the openness of their economies. Given that correlation between economic growth and equality which was depicted by the OECD's 2014 'How was Life' study (van Zanden, 2014) and the research by the Heritage Foundation and Frazer Institute that advances in reduced level of State intervention and the promotion of financial/commercial freedoms correlate strongly with "greater economic dynamism and prosperity" (Moore & Griffith, 2015), i.e. growth, within Supply-Sided modelled economies (which most advanced economies have been modelled upon since the late 1970's/early1980's) it is a understandable that a quantifiable index of such freedom should also be modelled as a gauge of government's policy activities within operation of the economy.
2.1.3 Returns and Inequality

As summarised by the 'The Economist', a weekly published business newspaper, the "study of political economy emerged... attempting to understand the dramatic societal and economic changes of the day and to describe their mechanics in a way that would allow them to anticipate future developments. To a great extent they focused on distributional issues—and worried that [unbalanced] distribution spelled serious trouble" for the stability of the economic and political system which over saw these (Economist Book Club, 2014). And as such it is within this line of thought (and to a lesser extent following on from the works of Simon Kuznets in the 1950's) that Thomas Piketty's 2014, study into the history of the
distribution of income and wealth, has continued with the publication of his 'Capital in the Twenty-First Century'.

Summing up his work in an article for the American Economic Review Thomas Piketty notes that from his analysis "the size of the gap between $r$ and $g$, where $r$ is the rate of return on capital and $g$ the economy’s growth rate, is one of the important forces that can account for the historical magnitude and variations in wealth inequality" (Piketty, 2015) which was such a key focus of early IPE studies. He goes on to emphasize that in advanced economies when $g$ is constrained to low or virtually stagnate level; i.e. steady-state (Center for the Advancement of Steady State Economy, 2015), a heightened or prolonged 'r − g' spread will pronouncedly exaggerate the inequality of a wealth distribution, partially during circumstance of shocks; i.e. recessions, deflation, wage decline, etc (Piketty, 2015). While such political and economic shocks will ensure that there are circumstances for downward and upward mobility in the scale of relative wealth acclamation in the long-run magnitude of wealth inequality, on a Macro level, will be a steeply rising function of the 'r − g' spread (Piketty, 2015). Given the general Pareto shape of wealth distribution, the spread will assume a multiplicative form where "unequal returns on capital are a force for divergence that significantly amplifies and aggravates the effects of the inequality $r > g$. Indeed, the difference $r − g$ can be high for large fortunes without necessarily being high for the economy as a whole" (Piketty, 2014a).

Unlike Hidds who based his findings on Regression modelling Pickett's does not reference either inferential statistics nor data analytical methodology in his 'Capital in the Twenty-First Century' to construct his argument but derives a set a mathematical equations from the observations which are in line with econometric principles and that explains the patterns which are observed. These equations with underpin his conclusion are penned as his three 'laws of Capitalism'. The first 'law' ties together the preponderance of wealth in society's total income ($βI$) to its own returns per unit of income ($αI$) and to its capacity to reproduce itself ($r$).

\[ \alpha I = r \times \beta I \text{ (or alternatively } \beta I = \alpha I / r) \text{, where} \]
\[ \beta I = \frac{W}{Y} \text{ is Wealth, a.k.a. Capital (W), as defined as the } \textit{total market value of everything owned} (\text{Piketty, 2014a}), \text{ as a share of the aggregate income } Y \text{ (e.g. GNI)} \]

\[ \alpha I = \frac{R}{Y} \text{ is the ratio of income accruing to Wealth (R) over aggregate income } Y; \text{ and } \]

\[ r = \frac{R}{W} \text{ is the income accruing to Wealth (R) per monetary unit (or $1) of Wealth (W)} \]

The second 'law' attempts to explain the same preponderance of wealth (\( \beta I \)) by linking it to net savings and growth.

**Law 2:** \( \beta I \) rises if \( \sigma > g \)

\[ \sigma = \frac{s}{\beta I} \]

\[ s = \frac{S}{Y} \text{ with } S \text{ representing net aggregate savings} \]

\[ g = \text{ is the proportional rate of change, over time, of aggregate income } Y \]

The last 'law' depicts the manner in which unequal wealth distributions beget even more unequal wealth distributions via the inheritance mechanism. (Varoufakis, 2014)

**Law 3:** \( \beta i \) rises in proportion to \( (\psi - e) d \)

\[ \psi = \frac{i}{Y} \text{ is the ratio of aggregate inheritance transfers (i) over aggregate income } Y \]

\[ e = \frac{W_d}{W_a} \text{ is the ratio of wealth owned by people at the time of their death (Wd) over the mean wealth of those alive (Wa)} \]

\[ d = \text{ the death rate} \]
As Yanis Varoufakis surmises, in his summing up of the implication of these rules, "given the trends established during the past three decades" in the low growth economies of advanced states these laws are self reinforcing of each other, allow one to contribute to the others while it itself is the benefactor others output (Varoufakis, 2014).

However the studies reliance on econometric principles and mathematical proofs, as opposed to data analytical techniques, has not gone unchallenged. Ranged from scepticism of the author's understanding of the dynamics of the supply/demand curve (McCloskey, 2014) to Varoufakis' question of Piketty's choice to treat capital and wealth as one distinct entity as opposed to Varoufakis' argument for a their distinctiveness as "once a problematic definition of aggregate capital is embedded in an analysis, the problems spread out to the definition of the return to capital" which in turn brings Piketty's three laws into scrutiny (Varoufakis, 2014). The most public of these challenges is perhaps that made by the Financial Times who noted the its findings were "undercut by a series of problems and errors. Some issues concern sourcing and definitional problems. Some numbers appear simply to be constructed out of thin air" (Giles, 2014), which lead to legal action been undertaken by Thomas Piketty.

While 'Capital in the Twenty-First Century' may have refrained from the application of analytical or Machine Learning techniques in its mode de operandi several sequent papers which later referenced or built on its finds have applied such methods to test its assumptions. One such study was undertaken by MIT's Daron Acemoglu and James Robinson which constructed a set of regression models to test Piketty's findings. Focusing on the top 1 percent's share of national income as the indicator of inequality and using interest rate and the rate of economic growth as explanatory they lagged inequality variables, ranging from zero to ten, and created a model for each lagged dependent variable. Piketty's hypothesis predicts a positive and significant coefficient for the explanatory variables; i.e. countries with higher g then r the incomes of the bottom 99 percent will grow, however a negative estimate that is statistically insignificant is observed. However in Accemoglu and Roblinson's case study of South Africa and Sweden, they do note that there are other variables that "are quantitatively more important than r - g" and that the "economic and political factors
stressed here [i.e. institutions, de jure political power, etc...] determine the distribution of resources more generally" (Acemoglu & Robinson, 2015).

In furthering the study of the relationship between economic and political factors and Income Inequality Kenneth Scheve and David Stasavage's tested a hypothesis that Partisanship within national governments affect the domestic distribution of income to the extent that governments of the Left are more likely to engage in redistributive policies in the form of progressive tax reform and wealth transfers. Their study specifically focused on the adoption of 'centralized bargaining' of public/private salaries within member states of the OECD, in conjunction with the political ordination of the sitting governments. While their regression models found no significant correlation between incubate Governments who were on the Left of the political spectrum they did observe that the application of policies that established such wage bargaining were correlated with a reduction in Income Inequality (Scheve & Stasavage, 2009). However they do not that in the four case studies which they undertook (i.e. Denmark, Ireland, Netherlands and Sweden) that the negative trend in each state had begun prior to the introduction of such bargaining mechanism.

2.1.4 Income Distribution Metrics

Before proceeding the term in which the above research and this paper discuss 'inequality' should be defined and clarified. As Frank Cowell indicated in his 1977 book titled 'Measuring Inequality' the phase itself is an "awkward word, as well as one used in connection with a number of awkward social and economic problems. The difficulty is that the word can trigger quite a number of different ideas" (Cowell, 2011). If we take 'Capital in the Twenty-First Century' as a starting point Pikett's work primarily focuses on inequality within the distribution of wealth, where wealth made up of income and return on assets. Specifically this research primary (but not exclusively) focused on how much of the wealth on concentrated in the top 1% (bottom 99%) and 10% (bottom 90%) percentiles. However Scheve and Stasavage's research focused on Income inequality and the proportion of income which was received by the top 1%. While the investigation of wealth is a more comprehensive measure as it include both income from labour and capital accumulations from previous time periods it
raises a number of complications. Specifically when viewed under the remit of Data Quality national datasets on capital, and the return from it, issues of availability, accuracy and completeness are apparent. Data on Income, i.e. "the increase in a person's command over resources during a given time period", is more widely available and regularly reported (Cowell, 2011) and as such it is concluded that for the purpose of this paper that it should focus on income, rather than wealth, as a subject matter for inequality.

Following on from this the next decision that needs to be made is what indicator best represent the disparity within the distribution of income. Firstly we must make an assumption regarding the commonality in the shape of the distribution of income across states. This assumption is that such distribution generally matches that of a Lorenz curve, which, is a convex curve which represents the share of total income accruing to those below percentile p, as p goes from 0 (bottom of the distribution) to 100 or 1, if a log has been applied, (top of the distribution). Given that income distribution is "fairly homogenous across countries" (Atkinson, Piketty, & Saez, 2011) the Lorenz curve has been found to be quite consistent in accurately representing these distribution across countries and incorporates many of the fundamental principles of inequality measurement (Cowell, 2011).
One of the most common applications of the Lorenz curve of inequality is the derivation of the Gini coefficient. Which represents the ratio of the area of OPD to the area OCD, where the line OD is representative of a perfect income inequality (i.e. a flat distribution) and P is the percentile of the population. While the Gini coefficient can be defined as the difference between two areas under the curve a more accurate and simple definition for it when it is applied for the use of gauging inequality is that it is "the average difference between all possible pairs of incomes in the population, expressed as a proportion of total income" (Cowell, 2011). However, it has been found to be disproportionately sensitive to transfers at the centre rather than in its tails of the distribution (Atkinson et al., 2011). Hence an income transfer between percentiles to the far left of P (i.e. poorer) or between those on the right (i.e. richer) has a much less impact on the coefficient than if the transfer takes place within the middle, as such an alteration near the trough of the convex curve will proportionally have a great impact on the delta between the areas under the curve as this point is pushed out or in towards the OD line.
This may not be a concern if the study was primarily concerned about variance around the mean however the increasing focus of recent research has been that regarding the upper tail of the income distribution (i.e. 1%) as more evidence has shown that the top share has a material impact on overall inequality within the distribution.

"Recent interest in top incomes has focused on the rise in top income shares, but it is also important to examine the distribution within the top income group. Has the changing composition of the top 1 per cent been accompanied by a less concentrated distribution? Or have those at the very top extended their lead?

One answer to this question is provided by the Pareto coefficient, or, more intuitively, the inverse Pareto coefficient, Beta ($\beta$), which measures the relative advantage of those higher up the income scale. Where the upper tail is Pareto in form, the average income of those above income, $y$, is given by $\beta y$"

(Atkinson, 2013)

Figure 2.6 Inverse Pareto Distribution Coefficient

The Pareto coefficient is derived from the conventional distribution $F$ function which gives the proportion of a sample which is below or equal to a specified level of a variable; e.g. the
proportion of the population with incomes less than or equal to defined level. Contrasting Income against population on a logarithmic scale a parameter coefficient can be interpolated to give the slope of a fitted line at a point on the curve, this slope can then be interpreted as a proxy measurement of the income distribution (Cowell, 2011).

The Pareto law for top incomes is given by the following distribution $F$ function $F(y)$ for income $y$:

$$1 - F(y) = \left(\frac{k}{y}\right)^\alpha$$

where $k$ and $\alpha$ are given parameters, $\alpha$ is called the Pareto parameter. The key property of Pareto distributions is that the ratio of average income $y*(y)$ of individuals with income above $y$ to does not depend on the income threshold $y$:

$$y*(y) = \frac{\alpha y}{\alpha - 1}$$

$\therefore \ y* (y) / y = \beta$, with $\beta = \alpha / (\alpha - 1)$

That is, if $\beta = 2$, the average income of individuals with income above $\epsilon n$ is $(\epsilon n \times 2)$ and the average income of individuals with income above $\epsilon m$ is $(\epsilon m \times 2)$. Intuitively, a higher $\beta$ means a fatter upper tail of the distribution. From now on, we refer to $\beta$ as the inverted Pareto coefficient. On the inverted Pareto coefficient $\beta$ (which has more intuitive economic appeal) rather than the standard Pareto coefficient $\alpha$ (Atkinson, 2013). Note that there exists a one-to-one, -monotonically decreasing relationship between the $\alpha$ and $\beta$ coefficients, i.e., $\beta = \alpha / (\alpha - 1)$ and $\alpha = \beta / (\beta - 1)$. 
The advantage of using the $\beta$ coefficient is that a higher $\beta$ coefficient generally means larger top income shares and higher income inequality, while the reverse is true with the more commonly used $\alpha$ coefficient (Atkinson et al., 2011).

Since the coefficients are not affected by differences in the income totals, as they are measures of the shape of the upper part of the distribution and are of greater intuitive economic appeal for this study the Pareto Coefficient and the inverted Coefficient is a more powerful indicator then income proportions and best fits the aim of a quantitative analysis (Atkinson, 2013).
2.1.5 Left/Right Political Spectrum

Although Left-Right spectrum is a recurring feature in the study of political science and governmental behaviour the labels have basically been formulated in an 'ad hoc' fashion. Resulting in a scarcity of empirical foundations and methods in justifying such lateral labelling (Castles, 1997). To greater understand the context and meaning of these labels a review of peer studies into the left-right divide is required that reverted to its measurable elements.

J.A. Laponce's study into the stable landmark elements along the Left-Right spectrum noted that 'cultural bias' in the positioning of such sign posts was an active variable; especially in the cases of 'progress' and 'youth'. The perceptions towards 'Social Hierarchy' and 'Religion' proved to be the most stable indicators but Laponce recognises the outcomes represent a more symbolic structure which tells us more regarding the personal political perception rather than administrable (Laponce, 1972).

Douglas Hibbs, in his establishing work in the PCB, classified the administrable subjects in his principle research by relating the political parties to the demographic and class nature of their voting base. With dichotomous blue and white collar groups voting respectively along the Left-Right Spectrum (Hibbs, 1977). However modern voting patterns, at least in most studied cases, since the 1970's have suggested a break with the credibility of this pattern; with the rise of 'silk-stockling socialists' and 'working-class Tories' (McKenzie & Silver, 1968). With greater electorate focus on leaders' image, economical conditions and issues concerns as the principal deteriorations of voting chose (Lewis-Beck, 1988).

This coordinates with the research of Ian Budge who states that the “meaning of the 'Left-Right' is definite and reasonably constant throughout the [contemporary] analyses. Principally it refers to classic economic policy conflicts- government regulation of the economy through direct controls or takeovers... as opposed to free enterprise” (Budge, 1987). This helps us to recognise that while various former tools to grant labels of left and
right on to governments or parties have come in and out of favour the general recognition of such bodies have remained the same, i.e. the German SPD governments of 1940's or 1990's have remained on the left side of the spectrum regardless of variations to their voter identifies. Anthony Downs focuses this differentiation toward a general variable in the governmental level of economic intervention (Downs, 1957). The theory holds as the most commonly reliant mean used to handle this question within 'Anglo-American' or Western-Liberal social sciences.

As such it will be under taken by this study, to move away from the classification methodology which was initially applied in the study of PCB theory (i.e. the socio-economic demographics of parties' electoral support) and to use data sources for the classification of parties' position on this spectrum which based that decision on their economic policy orientation.

2.2 Critical Evaluation of Literature

As we can observe from the above contributions the majority of such studies which have investigated this subject matter have looked at either the partisan ordination of the incumbent government or the economic indicators of the state as explanatory variables behind inequality. As such the majority of these are based on the assumption that the nation-state Government, with its power to adopt policy to affect the economic environment, is an isolated goal originated decision maker. And are, with the expectation of election cycles, not subject to situational or structural restrains on its ability to act as a rational decision maker. On the subject of structural factors the first fundamental difference that that paper will look to consider is the concept of government as a plural rather the singular term, by that we are referring to the separation of powers within modern states structures between the Executive, the Legislator and the Judicatory (which for our study is of lesser importance due to its role as the interpreter of legislation rather than the creator of such policies) (Heywood, 2002). In the case were the nation-state operates a presidential or semi-presidential structure; e.g.
United States, France, etc... , it is quite plausible that the former two branches of governments can be made up of differing partisan groups. Which inclines to a process of discourse and negotiation to be applied to the ratification of policies which passes through these bodies. While in compared to the alternative two systems the Parliamentary model 'fusions' the executive's and legislative powers allowing for an opportunity for the electorate return a "singular winner who controls almost all of the levers of government" (Fontana, 2009), in that the executive derives its power from the confidence of the assembly ad can be removed by it. However such Parliamentary systems are often associated with muti-partyism and necessity of distinct partisan groups to form coalitions to reach the level of confidence within the assembly for the executive powers to be granted to them. Subsequently, and particularly when Parliamentary assemblies elected through a Proportional Representation method, they are associated with a higher degree of fluidity at which these powers are relocated to personas (Heywood, 2002).
Figure 2.7 Systems of Government

Parliamentary system of Government

Executive

Assembly

Personal Accountability

Electorate

Presidential system of Government

Executive

Checks and Balances

Assembly

Electorate

(Heywood, 2002)
Such multi-partyism plays a distinct role in the situational factors that also may constrain a government's freedom to enact policy, across either system, in as such as within assemblies opposition representatives act as internal 'checks and balances' on the government parties and the executive (Heywood, 2002). Where the opposition is weaker or shares a similar position on the Left/Right spectrum as the incumbent governments, it may have less resistant to the passing of their bills. However if either of these conditions are not meet or if the Government's majority is proportionally negligible (if not an minority) it could be assumed that the freedom enact their optimum policy choice could be restrained and the final ratified policy would be significantly 'watered-down', then that which Douglas Hidds or Kenneth Scheve and David Stasavage's would have suggested.

In relation to the economic indicators, historically the PBC theorists and more recently 'Capital in the Twenty-first Century' have focused these attention upon GDP/GNP, Rate of return (i.e. interest Rates) and Inflation as the main explanatory economic variables of their research. While these economic variables are key indicators of the health and functionality of a national economy, elected policy makers power to direct manage these is in the majority of instances limited or not curtailed and as such their stimulation can be considered as targeted or desired secondary effects of the enactment of such policies. For example the Special Saving Investment Account (SSIA) scheme that was enacted by the Ireland Governments in 2001 with the "specific goal" of encouraging Irish citizens take money out of circulation by saving it in a SSIA scheme with a return of 25% at the end of a five year period. The scheme resulted in circa IE£800 been invested annually through 1,170,208 such accounts, at an expense of over IE£2.5 billion to the Irish Exchequer (Fiscal Policy Division (Tax Policy), 2006). Which was later considered by the then Minister of Finance as one of a set of anti-cyclical mechanisms to constrain "threat of inflation" during the economic boom which the country was experiencing at the time (Joint Committee of Inquiry into the Banking Crisis, 2015). One of the major reasons for this limitation in governments' ability to set policy for the direct manipulation of inflation or interest rates is the establish of the hegemony during the late 1970's and 1980's of the requirement for National Central Banks (NCB) to be politically and economically autonomist from the state. In such that they should have the ability to select the final objectives of monetary policy without mandatory approval or the
participation of government representative and are protect by legislation to pursue monetary stability (i.e. inflation) as a objective in the event of a conflict with government. While economically they are operational independent and does not participate in the primary market for public debt (i.e. Bonds) nor as the primary source of direct credit to the state at preferable rates (Arnone, Laurens, Segalotto, & Sommer, 2009). Since the 1980's these NCB rights have continued to grow in strength (see Table 2.2) and, while the perceived monetary objectives vary from country to country (i.e. United States of America's Federal Reserve set both GDP and Inflation while the European Central Bank (ECB) focus sole on an inflation rate objective of 2%) such independence has become an established norm of advanced economies (Boakes, 2008).

Table 2.2  Scores of Central Bank Autonomy 1980's and 2003

<table>
<thead>
<tr>
<th>Central Banks (number of CB)</th>
<th>Late 1980s (narrow index Cukierman sample)</th>
<th>2003 (narrow index)</th>
<th>2003 (full index)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Political                           Economic</td>
<td>Overall</td>
<td>Political</td>
</tr>
<tr>
<td>All CBs (28)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>GMT Sample (18)</td>
<td>0.36</td>
<td>0.59</td>
<td>0.48</td>
</tr>
<tr>
<td>Cukierman Sample (7)</td>
<td>0.33</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td>ESCB (13) 1f</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

1/ The ESCB includes the ECB and the twelve central banks of countries that were participating in the single currency in 2003.
2/ Includes BCEAO, BEAC and ECCS. Unlike in the ESCB, member countries do not have a national central bank. Source: GMT (1991), Cukierman (1992) and authors’ estimates presented in the Statistical Tables.

As NCBs' grew increasing autonomous governments ability to set target for inflation and growth diminished resulting in the out dating the Hidds's PCB mathematical model (as well as the Philip's curve) as a party’s target rate of inflation ($\pi^T$) became increasing irrelevant as the national inflationary target was now been set by bodies which were outside of the influence of either the legislator or executive.

Given this decrease in direct influence on key parameters of the PCB model and what is commonly regarded in the above mentioned papers at explanatory variables of inequality it is reasonable to consider measurements which governments has direct or increase in as possible variables for this study. Such financial and economic variables would include the level of
government expenditure, rating of savings/investments (coincided Ireland's 2001 SSIA initiative as an example of such a policy which targeted these variables), availability/distribution of Credit, etc... Looking at economic variables at a more granular level (i.e. fiscal, financial, etc... ) would allow for greater insight into the specific indicators which are most correlated with changes in Income Inequality. Thus providing decision makers with valuable information regarding which variables should be targeted to optimise the affect of government policy on inequality.

2.3 Research Question

While the majority of other studies which investigated this subject matter have looked at the policy orientation of the incumbent government alone and as a single entity this paper will look to exploratory access the political situations the Executive and the and the institutional system in which they functions alongside national fiscal and economic indicator as plausible explanatory variables for the variations in the distribution of Income, using the share of income allocated to the top 10% of the population within a country and the inverse Pareto indicators (β), against its trend (as determined through a moving average). As such this paper will look to determine "Can components of the PBC significantly explain the deviations between the observed metrics of economic inequality, through the proxy of national income distributions, and its overarching trend within Developed Economies, with a specific focus on the distribution within the upper tails (i.e. the share of income coordinating to the top 1% and 10% of the population)?".
3 Design and Methodology

3.1 CRISP-DM Process Model

As this is a Data Modelling and Mining project the paper will apply the Cross-Industry Standard Process for Data Mining (CRISP-DM) guidelines to this study where the contests allows. The CRISP-DM is chosen as a process model given its establishment as an industry standard that provides an overview of the data mining life cycle and the dependencies between its stages (IBM, 2011).

Figure 3.1 CRISP - DM Process Cycle
Incorporating and building on the Literature Review (Chapter 2) of the previous section the 'Business Understanding' stage above outlined the key elements of the field of study and the contemporary views and discussions which are present within it. This forms theoretical foundation and the economic reasons for our explanatory data analysis and identify Domain Concepts that will be used to establish the entities for the data model which will be constructed during the modelling stage.

The Business Understanding stage will followed by five distinct stages; Data Understanding, Data Preparation, Modelling, Evaluation and Development stage. Each of which is defined below and then an outlined of how each stage will be undertaken or what is required from it for this study.

**Data Understanding:**

"The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information" (Ncr et al., 1999).

Sourced variable dataset undergo an initial examination to allow for familiarity with the frequencies and patterns within the data including the detection or elimination of variables based on these findings,

**Data Preparation:**

"The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools" (Ncr et al., 1999).
This stage incorporated the cleansing and enrichment of the variables' data which progressed from the data profiling and selection of the former stages the in preparation for the construction of the final Analytical Base Table (ABT).

**Modelling:**

"In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type." (Ncr et al., 1999).

Tests of correlation and significance of the variables are generated, and explanatory variables which are found to represent collinearity are retired from the experiment. Following which model construction is undertaken upon the explanatory and dependent variables contained within the Analytical Base Table.

**Evaluation:**

"At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached" (Ncr et al., 1999).

The evaluated findings from the model are contracted with are initial analytical objectives to access if new insight are now provided by the model.
Deployment:

"Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process" (Ncr et al., 1999).

As there is no 'customer' in a commercial sense the Deployment stage will coincide with final chapter by providing an overview of the study's experiment and its findings. Further areas of investigation and research will be highlighted in order to potentially improve on the results found.

3.2 Data Understanding

Following the 'Business Understanding' stage the CRISP-DM process advocates a 'Data Understanding' to be under taken where the available data sources and attributes which best satisfies the plausible explanatory variables that have been identified during the previous stage. Specifically for this research this relates variables in the domains of Econometrics, Political, Economic Freedom and Income distribution. Following on from the identification of the data sources, and an initial high level review of the datasets, the Type and Scales of Measurement of the variables contained within each of the dataset are defined will be indentified with a description of what the variable is measuring. This initial investigation of the available variables will be become increasing important in the selection of appropriate statistical tests and the interpretation of analytical results, while be undertaken in the 'Modelling' stage of the CRISP-DM and in the design of the database model by reflecting the Data Types (i.e. Char/String, Floater, Integer, etc...) which should be assigned to each attributes that contains the variables' data. The Variable Type will identify if the variable are Quantitative and Categorical in their nature. Quantitative Data (i.e. numerical data) are
variables where arithmetic operations (i.e. addition, subtraction, averaging, etc...) can, meaningfully, be carried out. Alternatively variables can be described as Categorical where the variable denotes groupings or labels (SAS Inc., 2014). For example if we look at political parties' orientation (i.e. Left, Centre, Right) it is not possible to perform to arithmetic calculation on such observation or even if numerical codes were used as keys to indicate such orientations (i.e. 1 = Left, 2= Centre, 3 = Right) the results of such operations as addition, etc... are meaningless (e.g. 1 + 2 = 3 => Left & Centre = Right).

The Scales of Measurement are sub-categories of Variable Type and refer to ways in which variables are defined by their characteristics and properties. Regarding categorical variables the Scales of Measurement can classified as either nominal and ordinal. A nominal categorical variable, are also sometimes referred to as either qualitative or classification variables, exhibits no logical ordering within the levels or groups of its observations; e.g. the categorical variable Gender is classified having properties that are agreeable to a nominal scale of measurement as it cannot meaningfully be arrange in a sequence that reflects intensification or its reverse. While an ordinal scale of measure represents categorical variables in which the arranging of their observes groupings in an orderly fashion implies that the differences between the categories are due to an intensification of magnitude. An example of this is the colour coding of storm warnings issued by Met Eirann (i.e. STATUS YELLOW, STATUS ORANGE and STATUS RED) where the severity of the storm warning can meaningfully be arrange in a specific sequence (Met Éireann, 2015).

Before talking about the Measurements of Scale that relate to quantitative variables, first it is important to address that Quantitative variables can be further defined as belonging to one of two sub-groups; that of Discreet and Continuous. Where Discreet are numerical values that are only a countable whole numbers within a measurement range. For example the number of political parties that present in an assembly cannot include fractions or non zero digits to the right of the decimal place (i.e. 10.5, 11.25, 12.3333, etc...) therefore the variable is restrained to only include whole numbers in its observations. Alternatively quantitative variables can be continuous in that they represent measurements on a scale that has no breaks and contains an infinite range of numerical valuations. The Measurement of Scale of either of these can be classified as either an Interval or a Ratio. An interval scale is similar to the
ordinal Scale of Measurement for a categorical variable in that it can be ordered by magnitude however it also has a "sensible spacing of observations such that differences between measurements are meaningful" (SAS Inc., 2014). Ratio is similar to Interval in many regards (i.e. ordering and spacing) but its range of the scale contains a true zero point which allow for ratios between the variables to be meaningfully calculated. A common example that is used to illustrated this difference is that of the Celsius and Celsius scale of temperature. As the Celsius scale has is no true zero point which would represent a total lack of temperature, measurements made with the Celsius scale cannot be relatively compared to each other to define proportions (i.e. 100°C is not twice the temperature of 50°C) which on the Kelvin scale as there is a zero measure that indicated absolute zero (0°) ratios can be calculated between observations (i.e. 100°K is twice the temperature of 50°K).

Following this initial examination this section will cover the development of an integration and data profiling tool that will allow for the conjoining of the various datasets into a single user friend form while calculating descriptive statistics for each variable and facilitate the generation of graphical representations of the distributions and frequencies of the before mentioned variables.

3.2.1 Dependant Variables

With regards to the dependent variable; i.e. The Top 10%’s share of National Income (1110101) and Inverse Pareto Coefficient (7100201), such data on Income Distributions will primarily be sourced from the World Top Income database (WTID). As one of the primary sources used by Piketty, the selection of the WTID for income data is a natural chose as it will reduce the risk of misinterpretation and error in taxonomy of the variable. The data is primary base on Tax statistic, which is in line with earlier studies of Income Inequality, as undertaken by Simon Kuznets, and is in accordance with international guidelines, as set forth in the "System of National Accounts (SNA) which established the rules and concepts for national wealth accounts and balance sheets" where possible (Piketty & Zucman, 2014). Presently it contains data on over 30 countries with timeseries spanning over a century for
some states. Built to accompany the publishing of the two books 'Top Incomes over the XXth Century' (2007) and 'Top Incomes : a Global Perspective' (2010), the WTID claim to offers the 'most comprehensive set of historical series on high-income inequality'. For the needs of this study the WTID provided a comprehensive set of observations for the Pareto Coefficient and Inverse Pareto Coefficient across multiple countries, and specifically those of advanced economies which satisfy the requirements of this study. When we perform an initial investigation of the dataset we observe that of the 404 variable the majority of the variables are Quantitative with a Continuous Ratio Measurements of Scale, such classification indicates that during the designing of the Database that a Floater (or Decimal) data type are be best suited to the attributes which will contain this data (i.e. the Dependent Variables). There are two other variables who's properties don't fall into these classifications, that of 'COUNTRY' and 'YEAR'. County matches the criteria of an Ordinal Categorical as it functions as an identifier for the countries which the income inequality data refers to and cannot be meaningfully ranked in a sequence. A similar statement can be said for YEAR which fits Measurement of Scale of a Discreet Interval. It is evident that from this investigation that some combination of these two identifying variables can be used during the modelling of the database and analytical experiment at unique keys.

Table 3.1 WTID - Measurements of Scale & Type

<table>
<thead>
<tr>
<th>WTID</th>
<th>Measures of Scale</th>
<th>Categorical</th>
<th>Quantitative</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous Interval</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Continuous Ratio</td>
<td></td>
<td>402</td>
<td>402</td>
</tr>
<tr>
<td></td>
<td>Discrete Interval</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Discrete Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Nominal</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ordinal</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grand Total</td>
<td>1</td>
<td>403</td>
<td>404</td>
</tr>
</tbody>
</table>
3.2.2 Explanatory Variables

The Database of Political Institutions (DPI) will be utilised to source a time series dataset of classification data of different political structures within and between countries. The DPI is a comprehensive and detailed source of quantitative data on 113 variables, from 1975-2011, on political institutions in 180 states which was compiled by the Development Research Group of the World Bank for research into political economy activity and comparative political institutions, with the investigation of the "existence and importance" of Political Business Cycles and the "political conditions under which governments promote rather than retard economic development" as cited fields to which its development was targeted by providing election and institutional data in a large sample of countries (Beck, Clarke, Groff, Keefer, & Patrick, 2001). It was initially used in 'New Tools and New Tests in Comparative Political Economy' where it was utilised to demonstrate the impact of divided government on public debt and the impact of presidential vs. parliamentary style institutions on democratic consolidation and has been periodically updated through to 2012. In regards to the classification of the variables it is far more varied then that of the WTID. Of the 108 variables 65 are categorical nominal, these include variables that define what form of a government system and electoral method is operated (discussed above), the classification of the Executive and Legislator's Left-Right orientation, etc.

Table 3.2 DPI- Measurements of Scale & Type

<table>
<thead>
<tr>
<th>DPI</th>
<th>Measures of Scale</th>
<th>Categorical</th>
<th>Quantitative</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous Interval</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous Ratio</td>
<td></td>
<td></td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Discrete Interval</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Discrete Ratio</td>
<td></td>
<td></td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>Nominal</td>
<td></td>
<td>65</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td></td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>67</td>
<td>46</td>
<td>113</td>
</tr>
</tbody>
</table>
The Continuous Ratio variables in addition to measurement of the proposition of seat held by parties or governments also include several derived measure of the dominance and majority of the incumbent government within the legislator. These include the Herfindahl Governments Index (HERFGOV) and Opposition Index (HERFOPP) which are the sum of the squared seat shares (S) of all parties in the government or opposition (n) respectively (Keefer, 2012).

\[
HERFGOV = S1^2 + S2^2 + \cdots + Sn^2
\]

\[\therefore HERFGOV = \sum_{i=1}^{n} S_i^2\]

This index is derived from the Herfindahl-Hirschman Index (HHI) which is used by market regulators (e.g. United States of America's Federal Trade Commission, etc...) to gauge the affect on competition within a market of the economy by the auctioning of a merger of two or more firm that trade within that market. The HHI is calculated by summing the squares of the individual firms’ market shares, and in so doing gives proportionately greater weight to the larger market shares, to return a value between 0 and 10,000 basis points (Bp) and is rerun again using the market share that would exist in a post-merger environment. Using this such regulators access concentration within the market where HHI's less than 1500Bp indicate a low concentrated/highly competitive Markets while an index above 2500Bp is indicative of a high concentration/low competition within a market (U.S. Department of Justice & Federal Trade Commission, 2010). In the DPI instead of looking at a market share which firms hold the database substitutes the share of seats each party holds within the incumbent and its opposition to determine the concentration or dominance of parties on either side of the aisle. Another variable that is a gauge of political dominance is that of FRAC (and GOVFRAC) which is the probability that two assembly members picked at random from the legislature are of different parties (Keefer, 2012).
Of the categorical variables, the methodology which DPI's implements to classify the orientation of the executive and the government and opposition within the legislature along the Left-Right political spectrum is based on the naming conventions of the political parties or grouping against which each below and that convenst association to policy preferences for greater or less state control of the economy. As such, unless an alternative classification is identified through a peered source, partisan parties which have the terms 'conservative', 'christian democratic', etc... in their names are categorised are 'Right' by the DPI. Similarly, parties are classified as 'Left' if their names associate them to 'communist', 'socialist', 'social democratic', etc... Centrist are associated with 'social-liberalism' and hence the advocating of greater involvement of private enterprise within the economy while supporting a redistributive role for government. Where none of the above condition are meet the variable defaults to 'Other' (Beck et al., 2001). Within the database the categories are given numeral identifiers where Right = 1, Left = 3 and Center = 2.

Economic freedom, as a measureable "degree to which the policies and institutions of countries are supportive" of restrained intervention in their domestic labour market, the movement of goods and capital and constraint on freedoms that could otherwise constrain commerce, will be sourced through the Fraser Institute's 'Economic Freedom of the World' annual indexes (EFW). The research data is overseen by an Editorial Advisory Board and incorporates 42 peer reviewed distinct variables of data to measure economic freedom in 141 states. The variables are separated into five major subject areas; Size of Government, Legal System and Security of Property Rights, Sound Money, Freedom to Trade Internationally and Regulation. Within the five major areas, there are 24 components many of whom are themselves made up of several sub-components. Each component and sub-component is placed on a scale from 0 to 10 that reflects the distribution of the underlying data. The subject areas index ratings is derived from the MEAN of their components, where the components are linked to sub-components they too are derived from the MEAN of these sub-components. Finally the total Economic Freedom Index for each country is derived from the average of the major subject areas (James, Robert, & Joshua, 2014).
1. Size of Government

"The four components of Area 1 indicate the extent to which countries rely on the political process to allocate resources and goods and services. When government spending increases relative to spending by individuals, households, and businesses, government decision-making is substituted for personal choice and economic freedom is reduced" (James et al., 2014).

A. Government consumption

B. Transfers and subsidies

C. Government enterprises and investment

D. Top marginal tax rate
   (i) Top marginal income tax rate
   (ii) Top marginal income and payroll tax rate

2. Legal System and Property Rights

"Protection of persons and their rightfully acquired property is a central element of economic freedom and a civil society. Indeed, it is the most important function of government... The key ingredients of a legal system consistent with economic freedom are rule of law, security of property rights, an independent and unbiased judiciary, and impartial and effective enforcement of the law" (James et al., 2014).

A. Judicial independence

B. Impartial courts

C. Protection of property rights

D. Military interference in rule of law and politics
E. Integrity of the legal system

F. Legal enforcement of contracts

G. Regulatory restrictions on the sale of real property

H. Reliability of police

I. Business costs of crime

3. Sound Money

"Sound money is essential to protect property rights and, thus, economic freedom. Inflation erodes the value of property held in monetary instruments. When governments finance their expenditures by creating money, in effect, they are expropriating the property and violating the economic freedom of their citizens" (James et al., 2014).

A. Money growth

B. Standard deviation of inflation

C. Inflation: most recent year

D. Freedom to own foreign currency bank accounts

4. Freedom to Trade Internationally

"Voluntary exchange is a positive-sum activity: both trading partners gain and the pursuit of the gain provides the motivation for the exchange. Thus, freedom to trade internationally also contributes substantially to our modern living standards" (James et al., 2014).

A. Tariffs

(i) Revenue from trade taxes (% of trade sector)
(ii) Mean tariff rate

(iii) Standard deviation of tariff rates

B. Regulatory trade barriers

(i) Non-tariff trade barriers

(ii) Compliance costs of importing and exporting

C. Black-market exchange rates

D. Controls of the movement of capital and people

(i) Foreign ownership/investment restrictions

(ii) Capital controls

(iii) Freedom of foreigners to visit

5. Regulation

"When regulations restrict entry into markets and interfere with the freedom to engage in voluntary exchange, they... limit the freedom of exchange in credit, labor, and product markets" (James et al., 2014).

A. Credit market regulations

(i) Ownership of banks

(ii) Private sector credit

(iii) Interest rate controls/negative real interest rates
B. Labour market regulations

(i) Hiring regulations and minimum wage
(ii) Hiring and firing regulations
(iii) Centralized collective bargaining
(iv) Hours regulations
(v) Mandated cost of worker dismissal
(vi) Conscription

C. Business regulations

(i) Administrative requirements
(ii) Bureaucracy costs
(iii) Starting a business
(iv) Extra payments/bribes/favoritism
(v) Licensing restrictions
(vi) Cost of tax compliance

Table 3.3 EFW - Measurements of scale & type

<table>
<thead>
<tr>
<th>EFW Measures of Scale</th>
<th>Categorical</th>
<th>Quantitative</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous Interval</td>
<td></td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>Continuous Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete Interval</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Discrete Ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td>1</td>
<td>56</td>
<td>57</td>
</tr>
</tbody>
</table>
The economic time-series data on nation macroeconomic indicators (i.e. Gross Domestic Products, National Income, etc...) and on financial and budgetary figures (Governments' Expenditure, External Debt, Interest Rates, Consumer Price Indexes/Inflation, etc...) will be sourced from the World bank's World Databank. The World Databank is an online web resource of indicator and metadata that has been made publically available by the World Bank. Some of the fiscal and financial variables are published in both current US Dollar and the Local Currency Units, as identified by the postscripts CD and CN in the variable titles.

Table 3.4  WB - Measurements of scale & type

<table>
<thead>
<tr>
<th>WB - Economic Data</th>
<th>Measures of Scale</th>
<th>Categorical</th>
<th>Quantitative</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous Interval</td>
<td></td>
<td></td>
<td>607</td>
<td>607</td>
</tr>
<tr>
<td>Continuous Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete Interval</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Discrete Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ordinal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>1</td>
<td>608</td>
<td>609</td>
</tr>
</tbody>
</table>

3.3  Data Quality - Dimensions and Development Cycle

For the purpose of this study definition Data Quality (DQ) can be defined as "a set of measures that determine if data are independently understandable for informed reuse" (Peer, Green, & Stephenson, 2014), for this study such a definition confirmed that the available variable data is fit for purpose. The failure to monitor and curtail attribute that represent questionable standards can lead to technical issues ranging from the loss of stored data within incompatible databases to processing errors during the data cleansing and manipulation if the data is inapplicable with the attributes' declared data types or contains inconsistencies. While such issues within a un-enterprise database or data mining process may be inconvenient on a national scale such issues are quotes as cost the United States' economy six hundred billion
Hence the design of an adequate DQ management system is one that meets and supports a specific set of goals (i.e. authenticity, verity, etc...) set out by the developer. As whenever a set of objectives are declared an arbitrary order of preference may also be outline to determine the priorities of the end user requirements weather they "prize the completeness of the data while others their accessibility" (Peer et al., 2014). As a by product of such judgement calls the best suit DQ system may not necessarily be the one that guarantee's zero defects at the cost of data availability or reliability as in a pragmatic and appreciable DQ mechanism "quality is conformance to valid requirements" that meet the business's or project's needs, such requirements can be classified as belonging to one of a set of dimensions in order to systematically "define, measure, and manage the quality of the data and information" for the purpose of DQ development and governance (Geiger, 2004).

Table 3.5  Data Quality dimensions

<table>
<thead>
<tr>
<th>Data Specifications</th>
<th>Data Integrity Fundamentals</th>
<th>Duplication</th>
<th>Accuracy</th>
<th>Consistency and Synchronization</th>
<th>Timeliness and Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A measure of the existence, completeness, quality, and documentation of data standards, data models, business rules, metadata, and reference data.</td>
<td>A measure of the existence, validity, structure, content and other basic characteristics of the data.</td>
<td>A measure of unwanted duplication existing within or across systems for a particular field, record, or data set.</td>
<td>A measure of the correctness of the content of the data (which requires an authoritative source of reference to be identified and accessible).</td>
<td>A measure of the equivalence of information stored or used in various data stores, applications, and systems, and the processes for making data equivalent.</td>
<td>A measure of the degree to which data are current and available for use as specified and in the timeframe in which they are expected.</td>
</tr>
<tr>
<td>Ease-of-Use and Maintainability</td>
<td>Data Coverage</td>
<td>Presentation Quality</td>
<td>Perception, Relevance, and Trust</td>
<td>Data Decay</td>
<td>Transactability</td>
</tr>
<tr>
<td>A measure of the degree to which data can be accessed and used and the degree to which data can be updated, maintained, and managed.</td>
<td>A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest.</td>
<td>A measure of how information is presented to and collected from those who utilize it. Format and appearance support appropriate use of information.</td>
<td>A measure of the perception of and confidence in the quality of the data; the importance, value, and relevance of the data to business needs.</td>
<td>A measure of the rate of negative change to the data.</td>
<td>A measure of the degree to which data will produce the desired business transaction or outcome.</td>
</tr>
</tbody>
</table>

Hence on an enterprise scale a DQ mechanism operates on a both a reactive and proactive level. The proactive elements consist of Specialized tools to enforce the integrity of the roles and responsibilities of data attribute that are in line with the quality expectations and the
requirements of the supporting practices. While the reactive component rectified inherent DQ inconsistencies within legacy data tables and attributes (Geiger, 2004). As there are not specified universal industry standard DQ management systems are regularly bespoke in design so to prioritise the required dimensions.

Figure 3.3 Data Quality Project Life Cycle

The development cycle of the DQ mechanism begins with a 'Justification' stage where the required dimensions are priority as per the project's need, this is followed by a 'Planning' stage where the data quality requirements using the dimensions at a high level are contracted against a profile of the data to identify structural differences or data issues and provide input to the selection criteria and cleans/transformation requirements. Using the information obtained from the Business Understanding (i.e. Chapter 2) and the Data Understanding (Chapter 3.1) stages the dimensions which are of most significance for this research are that of the 'Data Integrity Fundamental' and 'Data Coverage' class. The justification for this is that as the variables are measured and provided through several different institutions (i.e. data sources) and that the majority of the variables are macroeconomic indices which are either reported by governmental authorities or are compiled from annual reports by non-governmental. As such the capability of specific variables to be reported is dependent on the efforts and capability on behalf of the source to gather sufficient raw data to derive calculated or compounded variables (i.e. Consumer Price Indexes, Government Educational Expenditure, Value added by Industry, etc...) and determine if a variable exists (Short-term Economic Statistics Expert Group, 2002). Or if the content of the variables differ significantly in their range or data type (i.e. GDP priced in local currency units). While as noted in the case of the EFW's data the structure of the available dataset differ to other sources as in only one variable is reported per relational table.
Given that the study is constrained to Advanced Economies and a time frame of the 1975 to 2010 the availability of observations for this range is also of concern to the DQ and determining variable selection. As the variable are time-series in nature the existence of NULLs within the observation will also need to be addressed as it will question the comprehensiveness of the variables.

The access these dimensions, undertaking the profiling the variable/attribute's data is a "process of gaining an understanding of the existing data relative to the quality specifications" (Geiger, 2004) by quantifying their characteristics by calculating numerical summaries of such features. Where such feature summaries allow for cross comparisons and contract to be made between individual or sets of variables. These features break into two descriptive statistical categories; that of Measurements of Location and Measures of Variability (SAS Inc., 2014).

Statistics that locate the centre of the data are called measures of central tendency (i.e. Mean [\(\mu\)], Median, and Mode) and form a subcategory of the Measurements of Location. Other example of Measurements of Location include methods of defining of proportional range in accordance to declared reference points, such as percentiles. In effect these measure returns a value at which a certain fraction (quarters in the case of quartiles, 1/5s for quintiles, 1/6s for sextiles, etc...) or percentage of the observed variables fall below.

Measures of Variability quantifies the distribution or spread within the population of observed data points within the variable and are constructed from the Measurements of Location. Its measurements include the measure of such variability around the \(\mu\) (i.e. Variance [\(\sigma^2\)]) while the Standard Deviation (\(\sigma\)) quantifies how spread out the variable's observations by calculating in the same units of measurement how much variation there is from its \(\mu\). As the above measures are reported in either the variables' unit of measurement or its square neither are optima for cross comparison of variables' variance, to accommodate this a deviation of the \(\sigma\) measure is the Coefficient of Variation (CV) which measures the \(\sigma\) expressed as a percentage of the \(\mu\), hence creating a ratio Measure of Scale that allows for contracting such variation between explanatory or dependent variables. As such many
Measures of Variability are constructed from the Measurements of Location (i.e. μ, etc... ) including such as the variables' range (i.e. the difference between the maximum and minimum observed values) and the its Inter-quartile Range (i.e. the difference between the 1st and 3rd quartiles of the observed values) which determines the range of the middle 50% of observation.

These profiled then contribute to identifying the action that is then required by the DQ management system to safeguard the dimensional conformity of the data in the 'Requirements and Analysis' stage. This conformity is enacted through the either Transforming the data, from the "format of a source system into the required format of a new destination system" (Techopedia.com, 2016b), or Cleansing, "altering data in a given storage resource to make sure that it is accurate and correct" (Techopedia.com, 2016a), it through queries developed and designed during the subsequent Design and Development/Testing phases. As we are primary concerned with the 'Data Integrity Fundamental' and 'Data Coverage' dimension we are two aspects that are of particular interest to this DQ system is that of NULLs and outliers. The development of such processed faces a number of challenges and must account for including Performance Risks; additional drain on processing power (due to high volumes of calculation, temporary data storage, etc... ) slows down operation, Traceability Risks; derived data cannot be linked back to or reflective of the source data, etc... (Trillium Software, 2015). As such during these stages choice of a suitable infrastructure and scripting languages (i.e. Talend DQ, Visual Basic, SQL etc...) that best meets these requirements should be made with such considerations.

3.4 Data Cleansing/Transformation

3.4.1 Outliers and Standardization

An Outlier is an 'extreme value' which is located at a significant distant from μ of the of the dataset's population (Levine, Stephan, Krehbiel, & L., 2008) and can significantly influence
the μ and other statistics or model that are built from such parameters where are outlier may have been included in its calculation (SAS Inc., 2014). To prevent these outliers from influencing the models output their identification and examination should be a consideration of the DQ management system. Given the possible diversity in the ranges of the variables (i.e. nominal LCU, percentages, basis points, etc...) restructuring the range so that it systematically easier to identify the position of observations within the distribution of the variables’ observations is recommended process for cleansing data in this regards (Levine et al., 2008).

Such a process can be achieved through the method of normalising range to a reasonable scale, weather that be a defined range (e.g. 0-1), a percent, ratio, etc... The normalisation of an observed variable \(x_i\) to a value \(z_i\) that is limited to a specific range with a maximum and minimum possible value can be achieved thought the following equation:

\[
z_i = \frac{x_i - \text{Min}(x)}{\text{Max}(x) - \text{Min}(x)}
\]

While such methods are of use (Gelman & Hill, 2009) as they do not refer to the inherent distribution within the variable as such the alternative method of scaling the variable with reference to the μ and σ of the dataset can better communicated the relative positioning of an observation to its μ. A common method of doing this is though the application of \(z\)-Score

\[
z_i = \frac{x_i - \mu}{\sigma}
\]

Under this approach the standardised observation is "interpreted in units of standard deviations" (Gelman & Hill, 2009) and as such can be regarding examined to find indications
of it been a outlier. Under the Empirical Rule, for a normal distribution, only about 1 out of 20 values will be beyond two units of standard deviations from the mean in either direction and as such any standardised observation not found in the interval \( \mu \pm 2\sigma \) may be conceded a potential outliers.

While the Empirical Rule referrers specifically to a normal or bell-shaped distribution a more general rule which is applicable to all distributions is the Chebyshev Rule. Which "states that for any data set, regardless of shape, the percentage of values that are found within distances of \( k \) standard deviation from the mean must be a least":

\[
\left(1 - \frac{1}{k^2}\right) \cdot \left(\frac{100}{1}\right)
\]

(Levine et al., 2008)

As such in the event that the distribution is normally distributed the Empirical Rule declares that 95% of observation will contained within the first two deviation from \( \mu \), while in the case that the curve is any other shape other than a normal bell-shape (i.e. skewed ) at least 75% of will be found within those bounds.

Hence the design of the DQ mechanism will be designed to include an application of the \( z \)-Score Standardisation method to identity outlier which may be present within the variables of the dataset and cleans them from the dataset prior to model construction.

### 3.4.2 NULL Handling

Based on the insights from the Data Understanding stage of the CRISP-DM model and the selection of DQ dimensions which lead on from it NULL handling is an element which will need to be address by the DQ management system, based on a technique which is adequate
for the experiment in hand. In traditional exploratory analysis, where modelling techniques such as regressions, etc... observation where the dependent variable was unknown were typically excludes (i.e. complete-case analysis) but even where the dependent variable is observed similar missing data (i.e. NULLs) in the body of explanatory variables can "limit the amount of information available in the analysis" (Gelman & Hill, 2009). The present of NULLs within the datasets can classified as belonging to one of four classes, depending on their cause or pattern, of 'Missingness Mechanisms'.

- **Missingness Completely at Random:**
  
  "A variable is missing completely at random if the probability of missingness is the same for all units".

- **Missingness at Random:**
  
  "Most missingness is not completely at random, as can be seen from the data themselves... A more general assumption, missing at random, is that the probability a variable is missing depends only on available information".

- **Missingness that Depends on Unobserved Predictors:**
  
  "Missingness is no longer 'at random' if it depends on information that has not been recorded and this information also predicts the missing values... If missingness is not at random, it must be explicitly modeled, or else you must accept some bias in your inferences".

- **Missingness that Depends on the Missing Value Itself:**
  
  "The probability of missingness depends on the (potentially missing) variable itself".

  (Gelman & Hill, 2009)
Given the sources (World Bank, Bank of International Settlements, etc.) and the nature of the variables and the data which they are derived from (Fiscal data, National Income Distributions, Macro-Econometrics, etc...) it is common that collection and reporting of such aggregated information is dependent on parameter such as bureaucratic structure and reporting cycles. And as such the NULLs which are present within this study's dataset are unlikely to be random but have 'Root Causes' (McGilvray, 2009) outside of the reported variables and as such a suitable model should be developed to account for such NULLs. Such a model responds to NULLS either by 'Discarding' the observation, 'Acquire the Missing Values', 'Imputation' or 'Reduced-feature' (Saar-Tsechansky & Provost, 2007). As the dataset which is been utilised by this study is already limit in the span of its observations and as such the removal of additional instances, as is what is prescribed by the 'Discarding' the Observation, would further limit the information that is available to the modelling stage this option for NULL Handling is not appreciable in this circumstance. Similarly while the present of NULLs within variables will be a determining factor in Variable selection following the profiling stage the knowledge that for certain states the present of NULLs are due to differing reporting cycling (e.g. Annual vs. Bi-Annual) a removal of a whole variable by the 'Reduced-Feature' action is also excluded as this will have already taken place to a supervised degree. As acquiring the missing values can require the additional costs and open the variables to risk surrounding differences in methodology such an approach is impracticable given the restraints of this study. As such in order to maintain the information which is contained by the known observations within the variables an imputation approach is most consistent with the aims of this study.

Imputation required the replacement of the NULL observation with an estimation of its value base on a generated prediction of it from a specified model. Such models vary in complexity and methodology, for the sake of this research it will be limited to method of Predictive Value Imputation (PVI). The most simplest PVI approach is to replace the NULL with $\mu$ or the Mode value of the variable for Quantitative and Categorical variable types respectively. The benefits of this approach is that the implementation of this method will not have an inverse affect on the Central Tendency as it is expanding the observations with values at the centre of the distribution or the table of frequency in the case of categorical. Additionally from a design and implementation view point this approach requires limited processing power as the values for imputation can be calculated and store once during the DQ processed
and does not require recalculation for each NULL observation. However from a modelling perspective this approach can limit or distort the impact of the significance of explanatory variables on the dependent by not accounting for trends or patterns within the data.

More robust imputations can be predicted with the use of interpolation models. Such model look to return prediction of a suitable estimated value for a NULL value by observing its neighbouring points in its sequence, line or curve (Egelston, 2016). Three common approaches is that of Step, Linear and Cubic Interpolation.

Step Interpolations involves the replacement of the NULL with the previous or next observed value within the sequence. As in the first discussed approach this requires limits processing resources but does need the recalculation of the imputation value for each encountered NULL. Can be of notable use when handling NULLs in categorical variables, specifically if they are not ordinal, as it can guarantee that the estimated value belongs to the categorical set but still takes in to consideration its location where as the MODE would not. However with regards to quantitative, and especially continuous Measurements of Scale, Step Interpolation can distort patterns within the variables' sequence and this affect can be exaggerated if a number of NULLs follow each other.

Linear Interpolation proceeds on from the Step method by accessing a weight to the values \( (V) \) of its two neighbouring points (i.e. \( a, b \)) in the sequence base on the location (\( T \)) of NULL (\( \chi \)) and estimating a imputation value (\( V_\chi \)).

\[
V_\chi = \frac{(T_\chi - Ta)Vb + (Tb - T_\chi)Va}{Tb - Ta}
\]

As such the returned estimate is located along a straight line which has the observed neighbouring values as end points. The Linear PVI as such is a more predictive estimation of the missing observation as it is proportional to its neighbours and hence conforms more
closely to trends within the variable. However this model require additional processing resources as for each encounter with a NULL the numerator and denominator of the equation must be computed and the estimator calculated. As the Linear PVI calculation is independent of values in the sequence beyond the NULLs two known neighbours it is characterised by non-smoothness or sharp changes in the curves direction due to this discontinuity at each point.

Cubic Interpolation resolve some of these questions regarding smoothness by allowing for 'continuity between the segments'. This continuity is achieved by expending the inputs of the interpolation equation from the two neighbouring observations to four, where similarly to the Linear PVI each of these points is weighted according to its relative distance from \( T \chi \) and where the first/second derivative is also continuous, so ensure a smoothness to the curve on which the \( V \chi \) can be found. As this Cubic interpolation also required the derivatives to be calculated the processing requirements are significantly increased when compared to the previous alternative PVIs. In return for this cost however the estimated value for NULLs can be conceded to be more representative of the flow of the trend within the variable but can bring into question the consistency of the interpolation at the start or end of the sequence of observation given that the 4th observed variable is no present within the population (Bourke, 1999).

Of these PVI options no one is perceived as the market standard and as such depending on the nature of the variables and the subject matter either (or another) interpolation rule may be applied. For example the OECD elects to incorporates the Linear method to "fill the gaps in the time series" for its 'Indicators of Regulation Impact' (OECD, 2015) while in the same year it concluded to avoid the use of the Linear PVI in favour of a 'backward- and forward-projection' technique (van Zanden, 2014) in its 'How was Life?' study. When specifically consulting the data sources which this study is focusing, Linear PVI has been opted for by the BIS.
"For most countries, data meeting the target specification are available over the entire series. [but] for countries where data with targeted properties are available only at lower frequency (ie annually), data are linearly interpolated." (BIS, 2015)

Similarly many governmental institutions and agencies (e.g. Australia) from whom "both high-quality aggregate (macro) and distributional data (micro) on household economic well-being" is been coincided, via the WTID, in this study apply linearly interpolate and extrapolate method for the imputation of missing values (La Cava, 2014). As such given the processing demands and the opting for the linear interpolation method by several of the data sources the Linear PVI will be implemented into the DQ design for Null Handling as opposed to alternative methods.

This leads on to the concept of extrapolation; i.e. the process of finding a value outside of the range of the observed data set, that is required if the NULLs occur at the start or the end of the timeseries. As La Cava notes the Australian Bureau of Statistics (ABS) in addition to applying a Linear PVI method to handle NULLs embodied within the dataset but also to use Linear Interpolation to predict to extrapolate such bordering NULLs. This is achieved by switching the neighbouring observed values (i.e. a, b) with the two observation which preceded or follows the NULLs depending on if the NULL is at the end or state of the sequence. However with such a process the further forward or back a value is extrapolated the more inaccurate such a predictions will be (Botts, 2016). Hence the development of an extrapolation mechanism should take this into consideration and attempt to account for such concerns.

3.5 Modelling Techniques

3.5.1 Multi-Linear Regression
As noted during the literary review the primary tool that has been used to measure the causality behind Income Inequality has been the application of regression modelling. Part of the reasoning behind this, and why regression is the "most widely used tool in econometrics", is that for explanatory aims the production of mathematical formula which represents the influence which the explanatory variables exerts on a dependent variable with a measure of error makes the interpretation and acquisition of knowledge from the experiment intuitive and when compared to Machine Learning techniques (Schmidheiny, 2010). Such techniques range from a wide verity of supervised tools; including Support Vector Machines (SVM), Random Forest, Artificial Neural Networks (ANN), etc... but which "the interpretation of the calibration models from ANNs is often very difficult" for exploratory purposes, as such these analytical mechanisms are often given the title 'Black Box' solutions (Goodacre, 2003). As such so to keep with the standards already set by previous IPE studies and to allow for clear examination of the strength of explanatory variable the design of the experiment will opt to utilise a set of Multiple Linear-Regression models to undertake this research.

The depend variable \((Y)\) of a regression model must be a numerical values which represents a natural order and where the differences between two values \((Y_i, Y_j)\) are meaningful, a interval or ratio Measurement of Scale, with a suitable level of accuracy. As such ordinal categorical variables (e.g. Left, Centre, Right) are not optimal variables structure for this model (Schmidheiny, 2010), as such where similar variables are present within the sourced dataset an adequate transformation will need to be applied.

\[
Y_i = \alpha + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_n X_{i,n} + \varepsilon_i
\]

The coefficient parameters (e.g. \(\beta_n\)) defines the assumed relationship between dependent variable and explanatory variables (i.e.\(X_n\)) based on a fitted line that is as close as possible to all the data points. The optimisation of this lines position, and the calibration of the coefficients, is achieved through the calculation of the line equation, from an array of candidate parameters (e.g. \(\hat{\beta}_n\)), that minimizes the sum of the squared vertical distances between the observed gridded/graphed data points and the fitted line (i.e. Least Squares
method). This method should produce the Best Linear Unbiased Estimators (BLUE) of the population with minimum variance (SAS Inc., 2014). While the Error Term (ε) represents the remaining unexplained difference between \( Y \) and the fitted line (as defined by the rest of the equation). The \( \epsilon \) is representative of the effects of the variables that were omitted from the equation (Freedman, 2001), Nonlinearity within the variables, inaccuracy in the recording of the observations and stochasticity within their distribution (Ramanathan, 2002). With the assumption that all variances of \( \epsilon \) are equal (i.e. \( \text{var}(\epsilon) = \sigma \)), approximately normally distributed and are independent (Shedden, 2015).

When undertaking an explanatory analysis through Multiple Linear-Regression the testing of the relationship between the explanatory and dependent variables are concluded through the testing of the statistical significance (\( p \)-Value) of the parameter coefficients (\( \beta \)) along with their mathematical signs (+/-) and magnitude to determine whether a relationship exists (SAS Inc., 2014). Linear regression models are based four assumptions (LINE) along with the fifth implied condition of minimal correlation between explanatory variables (collinearity). The assumptions are:

**Linearity:**

The relationship between explanatory and dependent variables are linear.

**Independence of Errors:**

The errors (i.e. \( \hat{\epsilon}_{x_1}, \hat{\epsilon}_{x_2}, ..., \hat{\epsilon}_{x_n} \)) are independent of one another.

**Normality of Error:**

The errors (i.e. \( \hat{\epsilon}_{x_1}, \hat{\epsilon}_{x_2}, ..., \hat{\epsilon}_{x_n} \)) are normally distributed at each value of \( x \).

**Homoscedasticity:**
The variance of the errors (i.e. \(\hat{\epsilon}_{x_1}, \hat{\epsilon}_{x_2}, \ldots, \hat{\epsilon}_{x_n}\)) are constant for all values of \(x\).

While these assumptions are universal for all linear regression models, such analysis are 'fairly robust against departures from the assumption' but for extreme cases (Levine et al., 2008). Time series modelling adds an additional the assumption that the properties such as 'measures of location' and 'measures of variance', etc... of the variables remain constant over time (stationarity).

### 3.5.2 Auto-regression Integrated Moving Average (ARIMA)

As the method utilised by Hibbs in his key works into the PBC field the Auto-Regression Integrated Moving Average (ARIMA) will be used as a technique for the study of the data in a time series structure for this research. The ARIMA or ARIMA(p,d,q) is made up of three elements, first is the 'Auto-regressive' (p) model element which accounts for the value of a variable in one period (\(y_t\)) been related to its values in previous periods (\(y_{t-1}\)). The second element 'Integration' (d) refers to the degree of differenc in the model while the final is the 'Moving Average' (q) that account for the possibility of a relationship between a variable and the residuals from previous periods (\(y - y_{t-1}\)) (Katchova, 2013).

\[
y_t = \mu + \sum_{i=1}^{p} (1 - \rho) \cdot y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \cdot \varepsilon_{t-i}
\]

Where \(\theta\) is the coefficient for the lagged error term and \((1 - \rho)\) is the coefficient for the lagged variable. The model requires \(\mu\) to be a constant for all periods and hence assumes that statistical parameter of the dependent variable is stable (stationarity). This assumption can be enforced through the application of 'first differences' where the variable is transformed to reflect the changes from one period to the next (Nau, 2016). The application of an analytical
time series method such as ARIMA can account for the relationships between variables through time where Regression models (without Lags) can overestimates the relationship between the such variables.

3.6 Data Preparation & Modelling Design

With the above insight from the Business Understanding and Data Understanding stages it is concluded that the cleansing of the variables' data, which was further expanded during the Data Preparation Stage, will require the development of a bespoke tool to integrate the variables data from the 4 distinct data source. Following which a Data Quality and Profiling task will be required. This task, in addition to the generation of descriptive statistic of each variable, will allow for the identify inconsistencies, completeness issues and outliners.

Such findings have guided the requirements of the 'Data Preparation' and later the 'Modelling' stages of the project. To facilitate this a suitable Dimensional database model will be designed to store the integrated datasets using ERWin Data Modeller. In such a model the dimensional tables will set up to contain the categorical and reference attribute of the observations within the fact table. The developed database will be constructed in a MySQL Database Management System, where within metadata for each attribute will be generates and stored.

To cleanse and transform the data a number specifically developed subroutines and algorithms written in structured query language (SQL) will be run again the integrated dataset to remove outlier, through the standardization of the observations of each variable and the identifying of confidence interval (i.e. CI 95%) boundaries, and derive Step and Linear PVIs to handle the Nulls that are then produced from the removal of the outliers and that were inherently within the original categorical and quantitative variables. So to allow for scalability wherever possible functions and procedure will be incorporated in these scripts so
to allow for the looping of the algorithms against the attributes whom metadata meets the
criteria. Following the cleansing of the data the dependent variables; i.e. WT7100201 and
WT1110101 will go through a process of data enrichment which involves calculating a
measurable indicator of inequality that can be quantitatively measured against a generated a
trend line (i.e. Moving Average) of each county's income distribution. The reasoning for this
requirement is that as noted in the literary review/business understanding segment that over
the time period of this study's focus there has been a general upward movement within the
measurements of Income Inequality across all advance economies. As such in order to gauge
the impact of politically orientated policy decisions the dependent variables need to represent
changes relative to this trend. The moving averages' period is set to 4 (i.e. 4 years), as this is
representative of the general constitutional lifecycle of a government and the 'Economic and
Investment' cycle (O’Loughlin & O’Brien, 2006).

Figure 3.4  Rep. of Ireland deviation of the observed Inverse pareto coefficient from the trend
line

As for the explanatory variables, it’s has been noted during the initial Data Understanding
stage that within the continuous quantitative the size of the numerically observation can vary
greatly depending on the variables' subject, as in representing values into the billions when
considering Government expenditure quotations in nominal currency units to percentages when referring to Economic growth. In order to make these standardise the time-series variables at \( t \) will be converted into percentage change in comparison to the previous observation (i.e. \( t_{-1} \)).

Based on the research of D.Acemoglu and J.Robinson that the prevailing political institutions at a certain time \((t)\) determine the distribution of de jure and de facto political power which influence economic institutions. Through the distribution of resources as designated by institution policy the level of inequality within the economy is impacted at \( t_{+1} \). As such a lag of 1 is applied to the ABT, where the explanatory variables for \( t \) is contrasted against the dependant variable at \( t_{+1} \).

\[
\text{political institutions}_t \quad \rightarrow \quad \begin{cases}
\text{de jure political power}_t \\
\text{de facto political power}_t
\end{cases} \quad \rightarrow \quad \begin{cases}
\text{economic institutions}_t \\
\text{technology}_t, \text{skills}_t, \text{prices}_t
\end{cases}
\rightarrow \begin{cases}
\text{economic performance}_t \\
\text{inequality}_{t+1}
\end{cases}
\rightarrow \text{political institutions}_{t+1}
\]

(Acemoglu & Robinson, 2015)

On completion of these data preparation, cleansing and enrichment tasks an ABT will be exported to SAS Enterprise Guide to perform statistical modelling, via ARIMA and Multi-Linear Regression techniques. The modelling procedure will involve the generation on correlation matrixes between the explanatory variables and the removal such variable where such a correlation if found to exist, in order to limit the present of collinearity (i.e. deterministic variables within a multiple regression model are highly correlated among themselves) and endogeneity (i.e. variables are correlated with the regression model's error term) with the model generation process. As this is an explanatory (rather than a predictive) task the study is specifically interested in the parameter coefficients as to determine whether a relationship exists between the dependent variable and the explanatory variables, the study
will focus on the development of regression based models as these offer greater granularity surrounding the strength of factors on an dependent, through the use of \( p \)-Values, Coefficients' signs and magnitude then classification or machine learning techniques. The final regression models will be evaluated, not only on the \( p \)-Values and Coefficients but, by analysing the resulting strength of their Mallow’s \( Cp \) statistic and their suitability for explanatory purposes against the Hockings model.

\[
SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

\[
Cp = \frac{SSE}{\sigma^2} + 2k - n
\]

Under this approach as increases in the number of explanatory variables (\( k \)) are offset with a decreased \( SSE \). Hence the model with the smallest \( Cp \) is represents the best linear model (Beal, 2013). Select of the optimum model for parameter estimation can be identified through the Hocking criterion. Under which bounds a \( Cp \) score is deemed as been suitable for the pursuit of explanatory analysis (SAS Inc., 2014). Where \( p \) is the number of parameters (including the intercept) used within the model and \( p_{full} \) is the total number of parameters available for model construction.

\[
\text{Hocking Criterion} \implies Cp \leq 2p - p_{full} + 1
\]

### 3.7 Design Limitations

- The research will be constrained by the limitations of the available data, which be require the research to be implemented on a relatively small data set. Such
limitations are a reoccurring obstruction in the field of in-country inequality. As such "data will often make it necessary to narrow the scope of inquiry rather severely" by limiting the research in our case to a set of advance economies over a defined scope (Piketty, 2014a)

- As a continuation of the prior limitation the present of NULLs within data sets also poses its own distinct challenges to this research. By introducing the requirement for NULL Handling method such interpolation mechanisms for missing observations.

- Classification of categorical variables particularly in regards to political subjects (i.e. Governmental System, etc.) are open to interpretation of de Facto reality and de Jura structure. For example the Republic of Ireland is classified as operating under a Parliamentary structure but has an elected President with considerable constitutional powers including the requirement for their signature to enact legislation. However in the practice of government this right to obstruct the legislator by refusal to sign acts of legislation is not enacted hence the all executive powers are then to be found within the Government cabinet.
4 Implementation and Results

4.1 Data Preparation

The data preparation stage of the CRISP-DM is recognised as the most time and resource heavy of all the stages of the data mining process as it involves several tasks which ranged from Merging datasets, Attribute selection, Deriving new attributes, Aggregating records through to Null handling (IBM, 2011), which has been discussed during the DQ design segment of the previous section and which will need be enacted during the overarching Extract, Transform and Load (ETL) process. As part of this process the quality of the data of the available variables will be ascertained through a process of data profiling and statistical description.

To enact the DQ design which has been outlined above for this paper it has decided to develop a bespoke application to meet these requirements, based on Visual Basic for Applications (VBA) through Microsoft Excel (Appendix 7.2). VBA is an object orientated programming language based on BASIC, and as such as a syntax that follows a Object.Method/Property or in the sense of English Grammar a Noun.Verb/Adjective structure where a Method refers to an action which is to be undertaken or applied to the declared object. Greater levels of granularity or detail can be applied to Methods through the use of Parameters where the code can specify non-defaulted setting of the used Methods. Similarly a Property describes characteristic of the object which through a similar syntax as above can be altered accordingly. The object orientation of VBA allow for such Methods and Properties to be applied across the entire object or set of objects (Collection) as such for the purposes of integration and profiling of the data sets and the variables within them VBA allow for an efficient use of coding of the manipulation these data sets and the variables when declared as objects (Jelen & Syrtad, 2014).
Table 4.1  Visual basic for applications (vba) components

<table>
<thead>
<tr>
<th>VBA Component</th>
<th>Analogous To</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object</td>
<td>Noun</td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td>Plural noun</td>
<td>Usually specifies which object: Worksheets(1).</td>
</tr>
<tr>
<td>Method</td>
<td>Verb</td>
<td>Object.Method.</td>
</tr>
<tr>
<td>Parameter</td>
<td>Adverb</td>
<td>Lists parameters after the method. Separate the parameter name from its value with :=.</td>
</tr>
<tr>
<td>Property</td>
<td>Adjective</td>
<td>You can set a property activecell.height = 10 or query the value of a property x = activecell.height.</td>
</tr>
</tbody>
</table>

Scripts of VBA statements are grouped into subroutines (macros) which are applied to complete complex or resource heavy tasks such as creating new set of data from the sourced raw data, computing the descriptive statistical characteristics of the integrated data sets to visualising the patterns and distributions of the data in appreciate charts, by cycling through the coded statements in a procedural or 'step-by-step' fashion (Microsoft Corp., 2009). To complete the DQ task of integrating, profiling and visualising the variables' data the design of the VBA program will utilises a set of such macros with the development of another type of subroutine call a function. Custom functions differ from macros in that they perform calculations instead of taking actions (Dodge & Stinson, 2007). Specifically in this software project, a function (i.e. FIND_COL()) was developed to located the position of an declared variable in an array or matrix by utilising the .FIND method.

So to make the processing of the data more intuitive and linear an interface was designed to structure and ease the users' declaration of the data sets of the variables raw data and the selection of parameters (i.e. Time range, Variable selection or Variable to be visualised), and declare the files that contain the variable datasets that the tool is to use as inputs. This is facilitated through the use of ActiveX button objects which call VBA macros (e.g.
FIND_OPEN_WB(), etc...) within the application to complete desired tasks (e.g. Identify and create a list of open data files from which the user can select, etc...).

The selected file and variables names for these inputs are used within the macros to populate declared variables (dimensions), in these cases of string data type, which will then be used to define objects or create arrays respectfully (Microsoft Corp., 2009).

Following the select on the file names the remaining parameters that the program requires from the used are:

**Declare the time range** in years (i.e. YRS_RNG). As a sourced datasets contain observations from differing data ranges the YRS_RNG field allow for a specific range to be selected to determine the range of the integrated dataset.
Declare the field within the ISO CTY data table that contains the numeric code that's representative of a country. As to facilitate the integration of the datasets a unique identifier will required for each observation. that can be matched across the sourced datasets. The generation of this key will be constructed through a combination of the numerical version of the date stamp and a identifier of the country of the observation. In order for this the stated country in the raw data sources need to be converted into a numeric code. Such codes are published by the International Organization for Standardization in their ISO 3166-1 list (ISO, 2015).

Select the names of the variables (as titled in the sourced data file) from each of the datasets that have been selected. These variables have been selected due to the inputs of the Business Understanding (Literary Review) and Data Understanding stages of the CRISP-DM.

Given the number of individual tasks which are required in the processing of the data so to facilitate the integration of the dataset from different sources and profile the selected variables, a modular design has been elected as a preferable programming design pattern as opposed to the creation of a single overarching macro.
While the design of a signal macro would have had the benefits of a simpler design and reduced redundancy in regards the declaring and defining of variables, by divided the programme into small, manageable procedures its development is more manageable (e.g. debugging, etc...) and versatile. An effective modularised programme should satisfy three requirements that a main procedure always needs to be created to coordinate and integrate the work of all individual procedures. As such in this tools GET_DATA() is created to coordinated the calling of the sub-procedures/macros and exit on their successful completion. Error Handling is managed within the individual macros so to allow for more accurate reporting of errors and efficient debugging.
While "secondly, all the sub-procedures only communicate with their direct superior procedures" and finally "any 'cross-talk' among the sub-procedures at the same level is not allowed" (Sun et al., 2012) as such no parallel level communication has been incorporated into the design.

```vba
Sub GET_DATA()
    Call ATTR_INPUT
    Call NM_FIELDS
    Call CREATE_TAB
    Call MAP_CTRYCODE
    Call SEL_VARIABLES
    Call VAR_HEADERS
    Call PROFILE_TAB
    Call UPD_GRAPHIC
End Sub
```

On running the tool the GET_DATA() macro calls its sub-macros in sequence, where each subsequent macro is run when the prior macros has been successfully completed. The first task will remove all worksheets from the Excel tool a par from the INPUT and OUTPUT sheets and clear all the content from OUTPUT.
The OUTPUT sheet will contain the integrated data and as such will start reconstructing the table by cycling through the countries that have been included in the ISO input and assign a date in the YYYYMMDD format for each specified, using 1231 for the MMDD segment. Combining these two attributes, ISO Country code and YYYYMMDD, a third attribute is constructed which is unique.

The sourced datasets are then copied into the tool and assigned an individual worksheet and similar three fields, as the previously discussed, are added. Using the inherent Country field the Numeric Code from the ISO 3166-1 list is lookup'd and inputted into the non-integrated worksheets with the construction of a YYYYMMDD datestamp. As before with the deriving of the unique identifier (UI) in the OUTPUT worksheet the macros now construct comparable keys in each of the non-integrated worksheets by combining these two fields.

New fields are then created for each of the selected variables in the OUTPUT worksheet and the variable observations are then populated by using the UI to join the worksheets. By doing so an integrated dataset of the variables from the different sources is constructed.

The penultimate sub-macro (PROFILE_TAB() ) is tasked with the job of creating a profiling statistics for each of the variables within the integrated data set. These statistics included the Measurements of Location and Measures of Variability as discussed in Chapter 3 as well as the percentage count of the number of NULLs within each of the variables (Appendix 7.3) to meet the profiling requirements as set out by out DQ dimension selection.
The profile also includes the identification of its Modes and the percentage of dominance of its 1st mode, which can be specifically useful in handling non-continuous numerical data types (e.g. categorical, strings, discrete numerical, etc...)
 Measures of Variability:

Using the integrated dataset with a UI a selection of graphic chart are updated with the new data to visualise the patterns and distribution within each of these variables (Appendix 7.4).

The Time series Scatter Plot (Numerical Variables) plots each values of a variable on a two-dimensional graph such that which each point represents the value of a numerical variable (Y Axis) at a specific time (X Axis). The time series plots can help reveal patterns over time, patterns that are not apparent from the descriptive statistics or by reviewing the array of numerical observations (Levine & Stephen, 2005).
The Residuals Plot utilises the $z$-Scores of the variables to generate a plot that visualises the standardised data with their deviations, indicated along the Y-Axis, from their $\mu$ (X Axis). This plot can support the identification of the quantity of outliers or the spread of the distribution of the variable.

Frequency Table/Column Chart (Categorical Variables) reflects the quantity at which different values occur within the variable and as such the distribution or shape of the data (SAS Inc., 2014).
As noted in the 'Data Understanding' stage due to the availability of data and the reporting schedules/methodology of national statistics and econometrics, specifically in the decades prior to the 1990's the presence of NULLs is one of the main obstacle in undertaking this analysis. With this taken into account and the ineffectiveness of imputation over specific thresholds it is decided to perform a simplified feature-reduction where any variable (i.e. feature) with a percentage of NULL within its total number of observations (i.e. population) that is equal to or greater than 40% will be removed from the analysis regardless of any other Data Quality or Correlation features.

While we will implement this as a general rule there are some incidences where an exception to this will be made. The expectations are based on variable by variable case.
In the case of the of several of the variables provided within the World Bank’s Database of Political Institutions EXELEC, OPPMAJH, MILITARY, etc... that behave like Boolean indicators were '1' is comparable to a TRUE flag. As such we are seeing a large proportion of NULLs within these variables where there is no '1' flag present.

An exception is also made for the Economic Freedom of the World index that is provided by the Frazer Institute. Prior to 2001 the institute published data on the a country’s Economic Freedom index on a semi-decadal basis due to the lack of comprehensive or complete at those time periods. As such this creates an "unique pattern of missing features" (Saar-Tsechansky & Provost, 2007) within these variables and as such an imputation method may be used to interpolate observations for the regular periods of NULL. With the above findings all variables with 40% or more NULLs (excluding the above noted exceptions) are removed from the study and the remaining variables within the integrated dataset are deemed to be the selected variables that will contribute to the final ABT.
4.2 Data Cleansing/Transformation

4.2.1 Database model Design

To store the data and facilitate the cleaning phase of the DQ process a purpose built Database model is designed. This raises the option as to whether an Normalised Rational (ER) model or a Dimensional model design best suits the requirements as set out by the research question and DQ mechanism. The ER model divides the data into discreet entities which is represented by a table in the physical schema and as such supports the normalisation of data by removing the need for redundancy in that data attributes are only cited once in the model within its specific table (Connolly & Begg, 2010). However the inherent need for multiple tables in the symmetric layout increases the complex that is required from queries to navigate the tables.

The resolve this constraint the Dimensional Database model represents data in a standardised arrangement of tables to maximise the intuitiveness of the data paths and relationships. The Dimensional model (most commonly) follows a star schema design where a single table that contains observed aspects/activities or ‘facts’ occupies the centre of the model with a number of tables that contain descriptive information or reference data that provide context and constrains. The data held within the Fact Table are numerical data types while string based information is positioned within the Dimensional Tables which are connected back to the fact within the Fact Tables through their Primary Keys (PK). While this model can contain redundancy by not enforcing normality within the Dimensional tables this simpler layout and division of attributes allows for high-performance by queries to access, and in our case cleans and transform data as part of our DQ mechanism in the production of the ABT.

To facilitate this the CA ERwin Data Modeller software package is used to add in the design. A modelling solution product, such as ERwin, is utilised as opposed to manually designing and translating the design into code as it allows for the creations of database design through an visual interface, increasing efficiency and development time (CA Technologies, 2011).
To achieve this design a choice of the Dimensional tables must be selected, the first of these is the table to contain Date stamp information that referrers to the facts. The table contains descriptive temporal information (i.e. Date stamps, Quarters, Dates of Quarter Ends, Weekday indicators, etc...) which are linked to the Fact table using the YRS_ID attribute, the numerical version of the YYYYMMDD date stamp, which was constructed during in the integration process. Secondly as all facts should be restrained to be of a numerical type the categorical variable are each assigned a Dimensional table where the descriptive string is linked using an numerical Foreign key (FK) in the Fact Table that refers to its categorical class. In the case of categorical variables within the PID such numerical keys are already provide and as such Database model simply integrate these indexes and build upon them.

Figure 4.2   Logical Database mode
4.2.2 Data Cleansing

The data model is then constructed in a MYSQL Database Management System (DBMS), through which the cleansing and transformation segments of the DQ mechanism will be undertaken. As such these operation will be constructed through a 'Database Language S-Q-L' (SQL) based programme (ISO/IEC, 2003). Unlike VBA, which is a procedural, SQL is a declarative language which means that it "lacks traditional control-flow constructs" which required the use of loops (For, With, etc..) which were required during the Integration and Profiling subroutines (Fehily, 2008). SQL instead relies on the DBMS’s optimizer to determine the most efficient method for the clauses of a statement to be accomplished.

The interactive (non-embedded) SQL programme that has been developed to cleanse and transform the integrated dataset, and finally construct the ABT, is divided into eleven stages that consist of Data Definition (DDL), for the creation of the designed Database Model and the creation of temporary tables/attributes, and Data Manipulation languages (DML), to retrieve and edit data within the tables, statements (Appendix 7.5).

The first two stages oversee the construction of the Fact and Dimensional tables, with the attributes which each contains, and their constrains and data types. This is the followed by the creations of the relationship between the tables and their PK and FKs to complete the Database model. During the population of the tables the default values (i.e. -999, -888) which are present in the sourced dataset are removed and replaced with NULLs so to prevent these values from interfering with the Step and Linear PVI. To enhance the efficiency of the PVI algorithms, and the following calculation of the first differences, for variables where observations are available prior to 1975 the 1974 observation.(or the MEAN value of the last 5 years for the dependent variable to support calculation of their Moving Average) is included.

On completion of the data input, sets of Metadata is created for table and its attributes, which contains information on the constrains and data types of each of the attributes held within the
tables. Using the calculation of $z$-Scores of the variables on a country by country based any observation noted to greater than 1.96 or less than -1.96, which corresponds to the boundary of a 95% of the distribution of a normal distribution, is flagged as an outlier and replaced with a NULL value.

### 4.2.3 Data Transformation

As variables consist of both categorical and continuous numerical scales of measurement the NULL Handling mechanism utilises the Meta data, that has been generated for each of the attributes within the database which stores the variables data, to identify the best suited PVI algorithm (i.e. Step or Linear interpolation) for the variables accordingly. The algorithms are built using the mathematical formulas noted in Chapter 3 where curser and nested cursers are then applied for the PVIs to be applied on a country by country basis within each variable. Which calls on the required PVI method, that have been packages in functions, with the country code and variable name as requires input parameters of the function. The constructed script is then called through procedures which executes the statements.

- **PVI - Linear Interpolation**

The application of the Linear PVI's algorithm is structurally subdivided into four stages. The first stage of the algorithm creates a table through a set of cross and left joins with the Fact Table (FACT TAB) that contains each observation of the country in question (i.e. CTY) ordered by DATESTAMP with the observed variable and empty user-defined variables (e.g. @DIFF, etc...) which are populated by the stages of the algorithm which follows, for use in the interpolation calculation.
The second stage of the process populate the user defined variables that have been created in the prior stage with related to the observations from the date stamp prior to the stamp which is presently being the subject of the algorithm (i.e. \( t_{-1} \)).
Building on the previous two stages the third stage repeats that of the second but instead of populating user-defined variables for \( t_{-1} \) it focuses on the next date stamp (i.e. \( t_{+1} \)).

The final stage updated observations within the fact table with an interpolated value which has been calculated from the linear PVI equation with values for the numerators and denominators been sourced from the generated table. It is at this stage that the customisation of the algorithm to prevent the 'run away' forward or retrospective casting in the event that there is a sequence of NULL observations at the end or start of the array of observations. In such an event the algorithm will linearly forecast (or retrospectively cast) an observation for the next year but following that will plateau an additional interpolation. This is achieved by manipulating the multiplier (i.e. DATEDIFF(YRS_ID, PREV_YRS_ID)) of the equation to
return a zero value and hence the output of the equation will be that of the previous observation in the time series. A similar manipulation is under taken of retrospective interpolation which are subject to a similar circumstance.

- **PVI - Step Interpolation**

To enable the Step Interpolated PVI first a set of tables are created to stored the earliest observation in the time series of each country's variables. As it may occur that the earliest observation may not coincide with the date stamp of the start of the time series (i.e. 1975-12-31) the script is designed to account for such events or where the earliest observation is a default values which the vender/data compiler has used to identify missing or not applicable (i.e. N/A) observation (e.g. -999, -888, etc...). Where categorical variables are in the form of boolean data types zeros are used to replace NULLs as from the data understanding that these denote false for these variables.
On completion of the DQ process the numerical explanatory variables are then standardised, using the First Difference percentage. The First Difference is the series of changes from one observation to the next in the time series. Converting this into a percentage of the previous observation converts the unit of measurements of each variable (i.e. LCU, CD, percent of GDP, etc...) to a common scale which allows for fairer comparison between them variable for the purpose of modelling. The First Difference is seen to re-enforce the assumption of stationarity (i.e. properties such as 'measures of location' and 'measures of variance', etc...) remain constant over time) within the time series (Nau, 2016).
For the dependent variable the process is a little different as the study is concerned with the change in comparison to the trend the subtractor and denominator is substituted by a lagged indicator to represent the over the 4 year period. For this indicator a Simple Moving Average, so all previous observation within the period carry the same weight, is derived using a bespoke algorithm and is executed using a similar process as the PVI (e.g. cursors and procedures).

\[ x_{Diff} = \frac{(x_t - x_{t-1})}{x_{t-1}} \]

\[ M_{Ay} = \sum_{i=1}^{n} y_i \cdot \left( \frac{1}{n} \right) \]
\[ y_{\text{Diff\%}} = \frac{(y_t - MA_y)}{MA_y} \]

/* Create a procedure to calculate the Moving Average of the
Pareto-Lorenz and inverse Pareto-Lorenz coefficients, the residual
of the observed vs MA at T and the Boolean flag if the observation
is above or below the MA*/
PROCEDURE IF EXISTS MA_INPUT;
DELIMITER $$
CREATE PROCEDURE MA_INPUT() BEGIN

/* Declare Variables */
DECLARE c INT;
DECLARE done BOOLEAN DEFAULT 0;

/* Create the CURSOR */
DECLARE CURSOR_MA CURSOR
FOR
SELECT CTY FROM MA_CTY_TAB;

/* Set a Handler to Loop through the observations within CTY*/
/* Declare a condition when the variable 'done' is set to '1' */
DECLARE CONTINUE Handler FOR SQLSTATE '02000' SET done = 1;
/* SQLSTATE '02000' is a "NOT FOUND" (i.e. next row) Error Code*/

/* Create the table to store the data */
DROP TABLE IF EXISTS MA_VAR_TAB;
CREATE TABLE IF NOT EXISTS
MA_VAR_TAB
(
  CTY_YRS BIGINT,
  CTY INT,
  YRS_ID BIGINT,
  YRS_Plus5 DATE,
  WT1110101_MA FLOAT,
  WT1110101_VAR FLOAT,
  WT1110101_BOOL BOOLEAN,
  WT1110301_MA FLOAT,
  WT1110301_VAR FLOAT,
  WT1110301_BOOL BOOLEAN,
  WT7100101_MA FLOAT,
  WT7100101_VAR FLOAT,
  WT7100101_BOOL BOOLEAN,
  WT7100201_MA FLOAT,
  WT7100201_VAR FLOAT,
  WT7100201_BOOL BOOLEAN
) ;

92 | P a g e
FACT_TAB.CTY_YRS, FACT_TAB.CTY, FACT_TAB.YRS_ID, FACT_TAB.WT1110101, FACT_TAB.WT1110301, FACT_TAB.WT710C101, FACT_TAB.WT710C201 FROM FACT_TAB WHERE FACT_TAB.CTY_YRS, FACT_TAB.CTY, FACT_TAB.YRS_ID, FACT_TAB.WT1110101, FACT_TAB.WT1110301, FACT_TAB.WT710C101, FACT_TAB.WT710C201 FROM FACT_TAB WHERE FACT_TAB.CTY = 0 GROUP BY FACT_TAB.YRS_ID ) AS y ON x.CTY = y.CTY WHERE STR_TO_DATE(CAST(x.YRS_ID AS CHAR),'%Y%m%d') BETWEEN STR_TO_DATE(CAST(y.YRS_ID AS CHAR),'%Y%m%d') AND DATE_ADD(STR_TO_DATE(CAST(y.YRS_ID AS CHAR),'%Y%m%d'), INTERVAL 4 YEAR) GROUP BY x.YRS_ID ORDER BY x.CTY, x.YRS_ID ;

/* End the Loop when 'done' is '1' */
UNTIL done END REPEAT:

/* Close the CURSOR */
CLOSE CURSOR_NA;

END##

DELIMITER ;
The categorical explanatory variables in turn require recoding into a number of separate, dichotomous variables so to allow the Multi-Regression model to calculate R coefficients for the strength of these variables in influencing the dependent variable (Gelman & Hill, 2009). This recoding converts the categorical into boolean 'dummy' variables where a new attributes are created for each class of a categorical variable and '1' or '0' is assigned using IF statements to flag if an observation belongs to that Boolean variable.

The ABT is the constructed from these attributes which contain transformed explanatory and dependent standardised data and the dummy attributes which represent the categorical data in boolean form, and finally exported as a CSV file to be pick up by SAS Enterprise Guide (SAS-EG) for the construction on the regression model.

4.3 Modelling

4.3.1 Collinearity/Correlation Matrixes

As discussed Regression models are based on a number of assumptions, one of which is that there should be no strong correlation between that explanatory variables within the model the first task undertaken within SAS-EG is to check for collinearity within the populations of explanatory variables. This is accomplished through the generation of correlation matrixes, a table with n by n dimension where the value at the cross sections of $x_i$ and $y_j$ indicate the strength of the linear association (dependence) between the variable $i$ and $j$. The SAS-EG offers several differing methodologies for the calculation of correlation coefficient; including 'Pearson (parametric) r' (which measures a linear dependence between two variables), Kendall and Spearman Test (measures correlation coefficient based on rank). Which are produce alongside a p-value to test the Null-Hypothesis that the correlation coefficient is not significantly different from the baseline (chance). Across all methods the coefficient evaluation ranges between $1^\pm$. A positive association where the pair of variables tend to
increase or decrease in tandem (monotomic) results in coefficient which is greater than 0. Where as if the relationship is reversed so that one increases in once variable generally is mirrored with a decrease in the other; a negative monotonic association, is reflected by a correlation coefficient which is less than 0. While a coefficient of 0 indicated an absence of a monotonic association between the pair (Shong, 2010).

**Pearson r Coefficient:**

The ratio ($\rho$) of the covariance ($Cov$) of the two variables to the product of their respective standard deviations.

$$
Cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n - 1}
$$

$$
\rho(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}
$$

**Spearman Rank-order Coefficient:**

Is a rank based version of the Pearson r Coefficient which measures the strength of association between two ranked or ordered variables ($r_s$). Unlike the Pearson r Coefficient the ranking of the Spearman Coefficient allow for non-linear correlation to be accounted for in the resulting score between $1^\pm$.

$$
r_s = \frac{\sum_{i=1}^{n}((rank(x_i) - \overline{rank(x)}). (rank(y) - \overline{rank(y)}))}{\sqrt{\sum_{i=1}^{n}((rank(x_i) - \overline{rank(x)}))^2 \cdot \sum_{i=1}^{n}((rank(y_i) - \overline{rank(y}))^2}}
$$
Kendall’s TAU Coefficient:

The Tau coefficient (τ) quantifies the discrepancy between the number of ranked concordant and discordant pairs, and is specifically suited to gauge association between two ordinal variables. Unlike the Pearson r Coefficient which only returns a score of 1 for a perfect linear relationship Tau can return a maximum value for "wider range of scenarios" (Shong, 2010).

\[
sgn(x_i - x_j) = \begin{cases} 
1 & \text{if } (x_i - x_j) > 0 \\
0 & \text{if } (x_i - x_j) = 0 \\
-1 & \text{if } (x_i - x_j) < 0 
\end{cases}
\]

\[
sgn(y_i - y_j) = \begin{cases} 
1 & \text{if } (y_i - y_j) > 0 \\
0 & \text{if } (y_i - y_j) = 0 \\
-1 & \text{if } (y_i - y_j) < 0 
\end{cases}
\]

\[
\tau = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_i - x_j) \cdot sgn(y_i - y_j)}{n(n - 1)}
\]

As the research which was reviewed during the Business Understanding stage of the CRISP-DM most focused on Linear models and the it was decided during the design chapter to focus this experiments on the development of Linear Regression models the Pearson r Coefficient of correlation (ρ) best suits these requirements. As such Correlation Matrixes of the variables are built using these methodology.

From the Literary Review a significant coefficient of determination which is has been using within the field of IPE to identify significant association between variable has been circa 0.6+, with correlations of 0.74 or more been marked out are been 'very high' (Scheve & Stasavage, 2009). As such all collinearity which the matrixes identify with pairs of variables of correlations above 0.74 or below -0.74 is removed by the dropping of one of the pairs from the study. Any remaining correlation on either side of 0.6+ is investigated on a case by case basis with one of the pair of explanatory variables removed.
From this analysis it is observed that all EFW variables are strongly correlated with each other, indicating a positive association between the category variables and their Sub-category variables of $\rho = +0.86167$. Similarly the Total EFW score and the category variables are shown to have 'perfect linear relationship' with a $\rho = +1.0$. This is not surprising when it is conceded that methodology which the Frazer Institute use to derive the 'EFW_TOTAL’ valuation is based on the calculation of the MEAN of the values of the county score in each of the EFW categories. As such it is deemed that all the EFW variable can be removed from the study with the exception of 'EFW_TOTAL'.

Within the BIS variables the Correlation Matrix reveals that Credit provided to the private non-financial sector from all sources independent of the country of origin (CREDIT_PRIV_NONFIN_SECTOR) is highly correlated with those that source their credit from Domestic banks (CREDIT_PRIV_DOM_BANK) with a $\rho$ of $+0.96957$. Additionally credit provided to Households and Non-profit organisations (NPO) serving households (CREDIT_PRIV_HOUSE_NPISH) was also reflects a strong positive relationship with the lending been made to private non-financial corporations (CREDIT_PRIV_NONFIN_CORP) that operate within the states' economy. As the credit provided by domestic banks to the non-financial sector and that which is lend to household and NPOs is not significantly correlated (and a low $p$-value of $0 < .0001$ rejects the Null-Hypothesis that $\rho$ is not significantly different from the baseline) CREDIT_PRIV_DOM_BANK and CREDIT_PRIV_HOUSE_NPISH are retained for the study.

The matrix of the DPI reveals that a strong negative association is present between the HHI Indexes, HERFGOV and HERFOPP, and the FRAC variables, GOVFRAC and OPPFRAC, which measure the probability that two assembly members picked at random from the legislature are of different parties with $\rho$ of -0.79572 and -0.95269 respectively. This indicated that as the HHI increased, in reflection of high concentration, the corresponding probability of two randomly selected members from that parliament been of the different party becomes increasingly unlikely. This relationship is logical as the decrease in competition within an assembly is synonymous with a reduction in the number of political
parties taking seats within the aforementioned assembly, as such the HHI variables, with their precedence of been utilised across agencies for the quantifying of such states of concentration, are retained and the FRAC variables are dropped from the study.

Additionally the matrix reveals that the dummy variable for the use of a parliamentary structure (GOVT_SYS_PARL) is has a negatively 'perfect linear relationship' \( (\rho = -1.0) \) with the variable to indicate if the lower assembly majority was held by an opposition party (OPPMAJH). As parliamentary structures already dominates the population, and the other governmental structure irrelevant in this sample OPPMAJH is retained while the variables of governmental structures are removed. Other DPI variables that removed at this stage are:

Table 4.2 DPI - Variable description

<table>
<thead>
<tr>
<th>Variable Title</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXECAGE</td>
<td>Executive Time since formation under this name</td>
</tr>
<tr>
<td>GOV1VOTE</td>
<td>Largest Government Party Share of the Vote</td>
</tr>
<tr>
<td>H_ELECT_SYS_IND</td>
<td>Electoral system of the House: IND</td>
</tr>
<tr>
<td>H_ELECT_SYS_MIX</td>
<td>Electoral system of the House: Mixed</td>
</tr>
<tr>
<td>NUMVOTE</td>
<td>Total vote share of all government parties</td>
</tr>
<tr>
<td>OPP1VOTE</td>
<td>Largest Opposition Party Share of the Vote</td>
</tr>
<tr>
<td>TOTALSEATS</td>
<td>Total Seats within the Assembly</td>
</tr>
<tr>
<td>MILITARY</td>
<td>Is Chief Executive a military officer?</td>
</tr>
</tbody>
</table>

As in regards to the econometrics variables the first noted result from it is the significant positive relationship between the variables which are measured using the local currency units (LCU) of the country in question and the same variables which is instead using a constant United States Dollar (CD) as the denominated unit of measurement. In the variables that uses LCU changes to the currency, as in the conversion to the Euro (€) in 2002 by Germany, France, Italy, etc... , are apparent in the data and could cause unintended bias or error in a model so therefore it is decided to opt for the CD versions of these variables. In other circumstances where significant correlations are identified between pairs of variables the variable that best relates to policy decisions are given preference (e.g. Government
Expenditure, Deposit Interest Rate, etc...). Using such method in conjunction with the correlation matrix the 609 original variables are reduced to 50 (below).

Table 4.3  WB - Variable description

<table>
<thead>
<tr>
<th>Variable Title</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NY_ADJ_DKAP_GN_ZS</td>
<td>Adjusted savings: consumption of fixed capital (% of GNI)</td>
</tr>
<tr>
<td>NY_ADJ_AEDU_CD</td>
<td>Adjusted savings: education expenditure (current US$)</td>
</tr>
<tr>
<td>NY_ADJ_DNGY_GN_ZS</td>
<td>Adjusted savings: energy depletion (% of GNI)</td>
</tr>
<tr>
<td>NY_ADJ_DRES_GN_ZS</td>
<td>Adjusted savings: natural resources depletion (% of GNI)</td>
</tr>
<tr>
<td>NY_ADJ_NNAT_GN_ZS</td>
<td>Adjusted savings: net national savings (% of GNI)</td>
</tr>
<tr>
<td>TX_VAL_AGRI_ZS_UN</td>
<td>Agricultural raw materials exports (% of merchandise exports)</td>
</tr>
<tr>
<td>TM_VAL_AGRI_ZS_UN</td>
<td>Agricultural raw materials imports (% of merchandise imports)</td>
</tr>
<tr>
<td>NV_AGR_TOTL_ZS</td>
<td>Agriculture, value added (% of GDP)</td>
</tr>
<tr>
<td>NV_AGR_TOTL_KD_ZG</td>
<td>Agriculture, value added (annual % growth)</td>
</tr>
<tr>
<td>NE_GDI_STKB_CD</td>
<td>Changes in inventories (current US$)</td>
</tr>
<tr>
<td>FS_AST_CGOV_GD_ZS</td>
<td>Claims on central government, etc. (% GDP)</td>
</tr>
<tr>
<td>FR_INR_DPST</td>
<td>Deposit interest rate (%)</td>
</tr>
<tr>
<td>FS_AST_PRVT_GD_ZS</td>
<td>Domestic credit to private sector (% of GDP)</td>
</tr>
<tr>
<td>NE_EXP_GNFS_ZS</td>
<td>Exports of goods and services (% of GDP)</td>
</tr>
<tr>
<td>NE_EXP_GNFS_KD_ZG</td>
<td>Exports of goods and services (annual % growth)</td>
</tr>
<tr>
<td>NE_RSB_GNFS_ZS</td>
<td>External balance on goods and services (% of GDP)</td>
</tr>
<tr>
<td>NE_CON_TETC_KD_ZG</td>
<td>Final consumption expenditure, etc. (annual % growth)</td>
</tr>
<tr>
<td>TX_VAL_FOOD_ZS_UN</td>
<td>Food exports (% of merchandise exports)</td>
</tr>
<tr>
<td>TM_VAL_FOOD_ZS_UN</td>
<td>Food imports (% of merchandise imports)</td>
</tr>
<tr>
<td>BX_KLT_DINV_WD_GD_ZS</td>
<td>Foreign direct investment, net inflows (% of GDP)</td>
</tr>
<tr>
<td>TX_VAL_FUEL_ZS_UN</td>
<td>Fuel exports (% of merchandise exports)</td>
</tr>
<tr>
<td>TM_VAL_FUEL_ZS_UN</td>
<td>Fuel imports (% of merchandise imports)</td>
</tr>
<tr>
<td>NY_GDP_MKTP_KD_ZG</td>
<td>GDP growth (annual %)</td>
</tr>
<tr>
<td>NY_GDP_PCAP_KD_ZG</td>
<td>GDP per capita growth (annual %)</td>
</tr>
<tr>
<td>NE_CON_GOVT_ZS</td>
<td>General government final consumption expenditure (% of GDP)</td>
</tr>
<tr>
<td>NE_CON_GOVT_KD_ZG</td>
<td>General government final consumption expenditure (annual % growth)</td>
</tr>
<tr>
<td>NY_GNP_PCAP_KD_ZG</td>
<td>GNI per capita growth (annual %)</td>
</tr>
<tr>
<td>NY_GDI_TOTL_KD_ZG</td>
<td>Gross capital formation (annual % growth)</td>
</tr>
<tr>
<td>NY_GDS_TOTL_ZS</td>
<td>Gross domestic savings (% of GDP)</td>
</tr>
<tr>
<td>NE_GDI_FTOT_ZS</td>
<td>Gross fixed capital formation (% of GDP)</td>
</tr>
<tr>
<td>NE_DAB_TOTL_ZS</td>
<td>Gross national expenditure (% of GDP)</td>
</tr>
<tr>
<td>NY_GDP_FCST_CD</td>
<td>Gross value added at factor cost (current US$)</td>
</tr>
<tr>
<td>NE_IMP_GNFS_KD_ZG</td>
<td>Imports of goods and services (annual % growth)</td>
</tr>
<tr>
<td>NV_IND_TOTL_ZS</td>
<td>Industry, value added (% of GDP)</td>
</tr>
<tr>
<td>FP_CPI_TOTL_ZG</td>
<td>Inflation, consumer prices (annual %)</td>
</tr>
<tr>
<td>TX_VAL_MANF_ZS_UN</td>
<td>Manufactures exports (% of merchandise exports)</td>
</tr>
</tbody>
</table>
These EFW, BIS, DPI and WB variables are then compared to the two dependent variables to identify suitable variables to be included in the last modelling stage.

### 4.3.2 Variable Selection

**Inverse Pareto Coefficient (β)**

When compared to the Inverse Pareto Coefficient (β) ten variables are identified as have correlation coefficients that are significant from the baseline as defined by a $p$-value which is below 0.05.
Table 4.4  Inverse Pareto Coefficient's' Final set of explanatory variables for modeling

| Variable                | Pearson Correlation Coefficients, N = 576 | Prob > |r| under H0: Rho=0 |
|-------------------------|--------------------------------------------|--------|------------------|
| NY_GDP_PCAP_KD_ZG      | -0.20106                                   | 0.0001 |
| NV_IND_MANF_KD_ZG      | -0.1962                                    | 0.0001 |
| NE_DAB_TOTL_ZS         | -0.14844                                   | 0.0004 |
| HERFGOV                | -0.12523                                   | 0.0029 |
| NE_CON_GOVT_ZS         | 0.12152                                    | 0.0035 |
| PARTYAGE                | -0.11684                                   | 0.005  |
| OPPMAJS                 | -0.11349                                   | 0.0064 |
| NE_EXP_GNFS_KD_ZG      | -0.11113                                   | 0.0076 |
| TM_VAL_FOOD_ZS_UN      | 0.10574                                    | 0.0111 |
| OPPMAJH                | -0.09782                                   | 0.0189 |

Of these variables 'GDP per capita growth' (NY_GDP_PCAP_KD_ZG) and Value add by Manufacturing (NV_IND_MANF_KD_ZG) are seen as having the strongest association with the dependent variable $\beta$, but are considerably tighter then the bounds of the significant coefficient of determination that have be determined are $0.6^\pm$. 
Figure 4.3  Inverse Pareto Coefficients' scatter plots with 95% prediction ellipse

Figure 4.4  Top 10%' scatter plots with 95% prediction ellipse
Income Top 10%

When compared to the distribution of income associated to the top 10% of the population, 30 variables are identified as having correlation coefficients that are significant from the baseline as defined by a p-value which is below 0.05.

Table 4.5  Top 10%  Final set of Explanatory Variables for Modeling

| Variable                                    | Pearson Correlation Coefficients, N = 576 | Prob > |r| under H0: Rho=0 |
|---------------------------------------------|----------------------------------------|--------|------------------|
| EXEC_NO                                     | 0.38718                                | 0.0001 |
| FP_CPI_TOTL_ZG                              | 0.33405                                | 0.0001 |
| LIEC_1                                      | 0.19697                                | 0.0001 |
| CREDIT_PRIV_HOUSE_NPISH                     | 0.17211                                | 0.0001 |
| GOV2_C                                      | -0.17137                               | 0.0001 |
| NE_CON_GOVT_KD_ZG                           | 0.1553                                 | 0.0002 |
| EXEC_R                                      | -0.15668                               | 0.0002 |
| EFW_TOTAL                                   | -0.15685                               | 0.0002 |
| GOV2_L                                      | -0.14127                               | 0.0007 |
| H_ELECT_SYS_PR                              | 0.13765                                | 0.0009 |
| H_ELECT_SYS_PLU                             | -0.13765                               | 0.0009 |
| LIEC_7                                      | -0.13802                               | 0.0009 |
| EIEC_7                                      | -0.13802                               | 0.0009 |
| NUMGOV                                      | -0.13262                               | 0.0014 |
| GOV1_R                                      | -0.13116                               | 0.0016 |
| FI_RES_TOTL_CD                              | 0.12766                                | 0.0021 |
| GOVT_SYS_PARL                               | 0.12265                                | 0.0032 |
| OPPMAJH                                     | -0.12265                               | 0.0032 |
| EIEC_2                                      | 0.11198                                | 0.0071 |
| SSH                                         | -0.11062                               | 0.0079 |
| PARTYAGE                                    | -0.10981                               | 0.0083 |
| GOV1_L                                      | 0.10639                                | 0.0106 |
| HERFOPP                                     | -0.1015                                | 0.0148 |
| MAJ                                         | -0.09974                               | 0.0166 |
| TX_VAL_MMTL_ZS_UN                           | -0.09381                               | 0.0243 |
| TM_VAL_AGRI_ZS_UN                           | 0.08987                                | 0.031  |
| NE_EXP_GNFS_KD_ZG                           | -0.08706                               | 0.0367 |
| NY_ADJ_AEDU_CD                              | 0.083                                  | 0.0465 |
| LIEC_6TO7                                   | 0.08253                                | 0.0477 |
| EIEC_6TO7                                   | 0.08253                                | 0.0477 |
Of these variables the dummy variable for the indication that no executive been present (EXEC_NO) and 'Inflation, consumer prices' index (FP_CPI_TOTL_ZG) are seen as having the strongest association with the dependent variable, but like the analysis of the explanatory variables relationships with β these coefficient (+0.38718 and +0.33405 respectively) are considerably tighter then the bounds of the significance that have be determined are 0.6⁺.

Figure 4.4 Top 10%' scatter plots with 95% prediction ellipse
4.3.3 ARIMA Modelling

Inverse Pareto Coefficient (β)

The non-seasonal time series modelling of Inverse Pareto Coefficient (β), the dependent variable, shows that the dependent variable is largely stationarity with little increase in variance over time as according its observation chart, this is due to the application of the First Difference during the transformation stage. This observation is confirmed through the Dickey-Fuller test which indicates if a modelled variable has a unit root (i.e. non-stationarity) by returning a $|p| \approx 1.00$ (Katchova, 2013).

$$y_t = \rho y_{t-1} + \varepsilon_t$$
\[ y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + \varepsilon_t \]

\[ \therefore \Delta y_t = (\rho - 1)y_{t-1} + \varepsilon_t \]

Tau is the test statistic of the Dickey-Fuller test of \( \mu \) and of the trend, which are stationarity as indicated by the highly significant \( p \)-Value (i.e. <0.0001). Hence rejecting the Null Hypothesis of stationary for both parameters.

Table 4.6   Inverse Pareto Coefficient's ARIMA Dickey-fuller unit root test

\[ \text{Table 4.6} \quad \text{Inverse Pareto Coefficient's ARIMA Dickey-fuller unit root test} \]

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Type} & \text{Lags} & \text{Rho} & \text{Pr} < \text{Rho} & \text{Pr} < \text{Tau} & \text{Tau} & \text{Pr} > \text{F} \\
\hline
\text{Zero Mean} & 0 & -436.608 & 0.0001 & -18.72 & <.0001 & 0 \\
& 1 & -316.352 & 0.0001 & -12.54 & <.0001 & 0 \\
& 2 & -223.104 & 0.0001 & -7.62 & <.0001 & 0 \\
\text{Single Mean} & 0 & -469.560 & 0.0001 & -19.85 & <.0001 & 197.00 \times 0.0010 \\
& 1 & -369.869 & 0.0001 & -13.54 & <.0001 & 91.65 \times 0.0010 \\
& 2 & -279.477 & 0.0001 & -10.53 & <.0001 & 55.47 \times 0.0010 \\
\text{Trend} & 0 & -469.704 & 0.0001 & -19.84 & <.0001 & 196.72 \times 0.0010 \\
& 1 & -370.072 & 0.0001 & -13.53 & <.0001 & 91.52 \times 0.0010 \\
& 2 & -279.659 & 0.0001 & -10.52 & <.0001 & 55.38 \times 0.0010 \\
\hline
\end{array}
\]

However while the charted Autocorrelation Function (ACF), the proportion of the covariance between \( y_t \) and \( y_{t-n} \) to the variance of a dependent variable, does not reflect a gradual linear decay which would have been apparent in a non-stationarity model, both it and the Partial Autocorrelation Functions (PACF) show that largest correlation is present at the Lag of 1, which would suggest the use of a First-order model, but which is ultimately insignificant (i.e. it is within the bounds of 0.6\( ^{\pm} \)).

\[
ACF(y_n) = \frac{\text{Cov}(y_t, y_{t-n})}{\sigma^2}
\]
This insignificance is reinforced with the analysis of the First-order Autoregressive, First-order Moving Average and First-order Autoregressive/ Moving Average model. The measurements of the Akaike Information (AIC) and the Schwarz-Bayesian Criteria (SBC), which measures the trade-off between the model's fit and its simplicity, are high (i.e. far from zero) indicating the level of complexity required to fit to such model. While the residuals are largely distributed in a bell-shaped/normal distribution the quartile - quartile (Q-Q) plot reflects oscillation/curving around the prediction line which is indicative of lightness in the distributions' tailed.
- First-order Autoregressive model [ARIMA(1,0,0) or AR(1)]:

Table 4.7  Inverse Pareto Coefficient's ARIMA(1,0,0) maximun likehood estimation and AIC/SBC tests

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Approx Pr &gt;</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU</td>
<td>-0.02432</td>
<td>0.0044826</td>
<td>-5.43</td>
<td>&lt;.0001</td>
<td>0</td>
</tr>
<tr>
<td>AR1,1</td>
<td>0.18310</td>
<td>0.04111</td>
<td>4.45</td>
<td>&lt;.0001</td>
<td>1</td>
</tr>
</tbody>
</table>

| Constant Estimate | -0.01987 |
| Variance Estimate | 0.00773 |
| Std Error Estimate | 0.087921 |
| AIC              | -1164.23 |
| SBC              | -1155.52 |
| Number of Residuals | 576 |

Figure 4.6  Inverse Pareto Coefficient's ARIMA(1,0,0) residual normality diagnostic
• First-order Moving Average model [ARIMA(0,0,1) or MA(1)]:

Table 4.8  Inverse Pareto Coefficient's ARIMA(0,0,1) maximum likelihood estimation and AIC/SBC tests

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Approx Pr &gt;</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU</td>
<td>-0.02434</td>
<td>0.0042323</td>
<td>-5.75</td>
<td>&lt; 0.001</td>
<td>0</td>
</tr>
<tr>
<td>MA1.1</td>
<td>-0.15186</td>
<td>0.04134</td>
<td>-3.67</td>
<td>0.0003</td>
<td>1</td>
</tr>
</tbody>
</table>

| Constant Estimate | -0.02434 |
| Variance Estimate  | 0.00778  |
| Std Error Estimate | 0.088203 |
| AIC                | -1160.57 |
| SBC                | -1151.86 |
| Number of Residuals | 576      |

Figure 4.7  Inverse Pareto Coefficient's ARIMA(0,0,1) residual normality diagnostic
First-order Autoregressive and Moving Average model [ARIMA(1,0,1) or ARMA(1,1)]:

Table 4.9 Inverse Pareto Coefficient's ARIMA(1,0,1) maximum likelihood estimation and AIC/SBC tests

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Approx Pr &gt;</th>
<th>Lag</th>
</tr>
</thead>
<tbody>
<tr>
<td>MU</td>
<td>-0.01679</td>
<td>0.01361</td>
<td>-1.23</td>
<td>0.2180</td>
<td>0</td>
</tr>
<tr>
<td>MA1,1</td>
<td>0.91150</td>
<td>0.03175</td>
<td>28.71</td>
<td>&lt;.0001</td>
<td>1</td>
</tr>
<tr>
<td>AR1,1</td>
<td>0.97735</td>
<td>0.01805</td>
<td>54.16</td>
<td>&lt;.0001</td>
<td>1</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Estimate</td>
<td>-0.00038</td>
</tr>
<tr>
<td>Variance Estimate</td>
<td>0.007453</td>
</tr>
<tr>
<td>Std Error Estimate</td>
<td>0.086333</td>
</tr>
<tr>
<td>AIC</td>
<td>-1184.27</td>
</tr>
<tr>
<td>SBC</td>
<td>-1171.2</td>
</tr>
<tr>
<td>Number of Residuals</td>
<td>576</td>
</tr>
</tbody>
</table>

Figure 4.8 Inverse Pareto Coefficient's ARIMA(1,0,1) residual normality diagnostic
**Income Top 10%**

Similarly to the analysis of the Inverse Pareto Coefficient while stationarity is confirmed by the Dickey-Fuller test the ACF and PACF indicated that there is no Lag of significant correlation (i.e. it is within the bounds of $0.6^{±}$).

Figure 4.9  Top 10% ARIMA trend and correlation analysis
As such, the ARIMA analysis have indicated no significance as indicated by ACF or PACF for the dependent variables. Hence no significant ARIMA models can be built that would accurately represent a study of the explanatory variable so the study will focus on utilising Multi-Linear Regression to the exploratory parameter analysis.
4.3.4 Multi-linear Regression Modelling

**Inverse Pareto Coefficient ($\beta$)**

Using the final set of ten explanatory variables (NV_IND_MANF_KD_ZG, NY_GDP_PCAP_KD_ZG, NE_DAB_TOTL_ZS, HERFGOV, NE_CON_GOVT_ZS, PARTYAGE, OPPMAJS, NE_EXP_GNFS_KD_ZG, TM_VAL_FOOD_ZS_UN, OPPMAJH) a number of candidate Multi-Linear Regression models are generated with their performance statistics.

Figure 4.10 Inverse Pareto Coefficient’s fit criteria

![Fit Criteria for WT7100201_VAR](image)

The $p_{full}$ parameter of the Hockings model is calculated to be 11 (i.e. 10 possible explanatory variables with the addition of 1 for the intercept). This allow for the computing of a Hocker criterion, which is graphed on to the below dot plot as a red line, where a $Cp$ statistic which scores below this line is most suitable for parameter estimation ($Cp \leq 2p - p_{full} + 1$).
Using the Hockings model to calibrate the Mallow's $C_p$ statistics in this case the best fitting model that satisfies this criterion and that is optimum in regards to its parsimonious is determined to consist of 8 variables. Of these models one that consist of the Economic value added by Manufacturing (NV_IND_MANF_KD_ZG), GDP per capita growth (NY_GDP_PCAP_KD_ZG), Gross national expenditure (NE_DAB_TOTL_ZS), HHI of the Government (HERFGOV), General government final consumption expenditure (NE_CON_GOVT_ZS), Average Age of Parties (PARTYAGE), Exports of goods and services (NE_EXP_GNFS_KD_ZG) and Opposition majority (OPPMAJH) is deemed to be the best fit.
Table 4.11  Inverse Pareto Coefficient’s Mallow’s Cp & Hocking CRITicA

<table>
<thead>
<tr>
<th>$p_{\text{full}}$</th>
<th>#Variables</th>
<th>$p$</th>
<th>Hocker</th>
<th>C(p)</th>
<th>C(p) &lt; Hocker</th>
<th>R-Square</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7.6706</td>
<td>TRUE</td>
<td>0.1195</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE NE_EXP_GNFS_KD_ZG OPPMAJH</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>7.6804</td>
<td>TRUE</td>
<td>0.1195</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE OPPMAJS OPPMAJH</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9.0868</td>
<td>TRUE</td>
<td>0.1205</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE OPPMAJS NE_EXP_GNFS_KD_ZG OPPMAJH</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9.5957</td>
<td>TRUE</td>
<td>0.1196</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE NE_EXP_GNFS_KD_ZG TM_VAL_FOOD_ZS_UN OPPMAJH</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>9.6721</td>
<td>TRUE</td>
<td>0.1195</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE OPPMAJS TM_VAL_FOOD_ZS_UN OPPMAJH</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>11</td>
<td>TRUE</td>
<td>0.1206</td>
<td>NV_IND_MANF_KD_ZG NY_GDP_PCAP_KD_ZG NE_DAB_TOTL_ZS HERFGOV NE_CON_GOVT_ZS PARTYAGE OPPMAJS NE_EXP_GNFS_KD_ZG TM_VAL_FOOD_ZS_UN OPPMAJH</td>
</tr>
</tbody>
</table>
On examination of the model's Analysis of Variance the overall $p$-Value for this explanatory model (>0.0001) is highly significant. However the F-Value, which is the ratio of Mean Square of the Model to that of $\varepsilon$, and the difference between the $r^2$ (The ratio of the regression Sum of Squares to the total Sum of Squares) and Adjusted $r^2$ (which takes into account the number of terms in the model and only increases if the addition of variables significantly improve the model) is marginal which indicated that model is not over burdened with variables. However when the measurement $r^2$ and Adjusted $r^2$ are individually examined neither of these are found to be significant (0.1195 and 0.1059 respectively) which indicates that the overall model's strength is not suitability strong.

Table 4.12 Inverse Pareto Coefficient's Analysis of variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>0.50906</td>
<td>0.06363</td>
<td>8.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>518</td>
<td>3.75081</td>
<td>0.00724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>526</td>
<td>4.25987</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

"In an explanatory or analytic model, our interest is in the parameter estimates" (SAS Inc., 2014) however the Parameter Estimates table for this model shows that five of the variable weights are not significant, with $p$-Values that are greater than 0.05. Of the remaining Parameter Estimates ($\beta$) the weight of the 'Gross national expenditure' variable (NE_DAB_TOTL_ZS) is reflecting a much stronger influence on the dependent variable ($\beta = -0.96881$) then was indicated in by its correlation ($\rho = -0.14844$). This suggests that one percentage unit change in 'Gross national expenditure' is on average the mirrored with a -0.96881 move in the dependent variable (i.e. $\beta$). However this significance is only observes as part of this overall model which itself is not significant with measurements of $r^2$ and Adjusted $r^2$ of 0.1195 and 0.1059 respectively.
Table 4.13  Inverse Pareto Coefficient's Parameter estimates

| Variable                        | DF | Parameter Estimate | Standard Error  | t Value | Pr > |t| |
|---------------------------------|----|--------------------|-----------------|---------|------|---|
| Intercept                       | 1  | 0.04960            | 0.01587         | 3.12    | 0.0019 |
| NV_IND_MANF_KD_ZG               | 1  | -0.00325           | 0.00096713      | -3.36   | 0.0009 |
| NY_GDP_PCAP_KD_ZG               | 1  | -0.00423           | 0.00240         | -1.76   | 0.0794 |
| NE_DAB_TOTL_ZS                  | 1  | -0.96881           | 0.26196         | -3.70   | 0.0002 |
| HERFGOV                         | 1  | 0.037633           | 0.01524         | 2.47    | 0.0137 |
| NE_CON_GOVT_ZS                  | 1  | 0.16503            | 0.12001         | 1.38    | 0.1697 |
| PARTYAGE                        | 1  | -0.0004818         | 0.0001319       | -3.62   | 0.0003 |
| NE_EXP_GNFS_KD_ZG               | 1  | -0.00074648        | 0.00099413      | -0.75   | 0.4531 |
| OPPMAJH                         | 1  | 0.02199            | 0.02017         | 1.09    | 0.2760 |

**Income Top 10%**

Using the final set of 30 explanatory variables (EXEC_NO, FP_CPI_TOTL_ZG, LIEC_1, CREDIT_PRIV HOUSE_NPISH, GOV2_C, NE_CON_GOVT_KD_ZG, EXEC_R, EFW_TOTAL, GOV2_L, H_ELECT_SYS_PR, H_ELECT_SYS_PLU, LIEC_7, EIEC_7, NUMGOV, GOV1_R, FI_RES_TOTL_CD, GOVT_SYS_PARL, OPPMAJH, EIEC_2, SSH, PARTYAGE, GOV1_L, HERFOPP, MAJ, TX_VAL_MMTL_ZS_UN, TM_VAL_AGRI_ZS_UN, NE_EXP_GNFS_KD_ZG, NY_ADJ_AEDU_CD, LIEC_6TO7, EIEC_6TO7) a number of candidate Multi-Linear Regression models are generated with their performance statistics.
The $p_{full}$ parameter of the Hockings model is calculated to be 31 (i.e. 30 possible explanatory variables with the addition of 1 for the intercept). This allow for the computing of a Hocker criterion, which is graphed on to the below dot plot as a red line, where a $Cp$ statistic which scores below this line is most suitable for parameter estimation ($Cp \leq 2p - p_{full} + 1$).

Using the Hockings model to calibrate the Mallow's $Cp$ statistics in this case the no model is deemed as satisfying this criterion Given the weak correlations that were computed using the Pearson coefficient the lack of a suitable candidate model not exceptional.
4.4 Evaluation and Analysis

4.4.1 Discussion of Results

The analysis has contributed to the research of the PCB by indicating, through ARIMA and Multi-Linear Regression techniques, that of the elements researched by this paper no significant relationship is observed with economic inequality, in the form of Income distribution and specifically when discussing Income Inequality as measure by the distribution of income accountable to the top 10% and 1% of the population. The political structure/electoral systems and freedom at which governments are able to design and legislate
policy due to increased concentration of political parties within the Government or assembly are shown to be insignificant due to their marginal correlations and coefficients of determinations with reference to income distribution. Likewise the selected Microeconomic and fiscal indicators' explanatory powers are also limited, with no variables' correlation that is greater than the level of significance (i.e. 0.6\(^+\)). Both these coefficients and the large Error terms (in relation to the mean square of the models) suggests that variables which may shed greater insight into the behaviour of observed metrics of inequality to its general trend were omitted from this experiment. As such the finding of this analysis must conclude that, given the available data, such variation from the trend cannot be accurately explained to a significant degree by the political orientation of governmental institutions and therefore such ideological positioning has had little if any affect on the proportional quality of wealth (in the form of income) which is allocated to the upper tail of its distribution within these nation states.

### 4.4.2 Strength of findings

- This findings of this experiment have been based on an application of the CRISP-DM model that provides a structured approach for the processing of analytical experiments and model building with the aim that such findings are of a high standard and fit for purpose.

- This study has spent specific focus on the issues of Data Quality and the application of cleansing or transformation techniques where required. A focus which is notably lacking in a number of previous works in this field.
4.4.3 Limitation of Findings

- The research is limited by the lack of data in the terms of longevity and density. The quality of the findings would be greatly increased if the time series data for each variable per country could be sourced from earlier periods prior to 1975 (longevity) or if observations were reported at increased regularity of intervals (density).

- Similarly the lack of longevity and density within the data constrains the ability to perform lag of several years without significantly depreciating the number of observations which are available for model construction.

- The use of PVI introduce bias into the model as it is based on the assumption that observation within the variable follow a linear trend.
5 Conclusion

5.1 Research Overview

Unlike many of other studies which have looked upon the policy orientation of the incumbent government under the Partisan-Business Cycle (PCB) as an objective single entity this paper accessed the governmental makeup and the political environments in which such governments sat. While overlapping such orientation and policy decision making with a quantitative measure this socioeconomic outcome in the form of Income Inequality. To quantify this yard stick a two measurements of income inequality (i.e. the share of income allocated to the top 10% of the population within a country and the Inverse Pareto indicator (B)).

5.2 Problem Definition

This paper looked to research the PCB theories’, which theorises the impact of party political orientation on cyclical economic measures, capability to affect trend in domestic income inequality. And to determine if institutionalised political structures or the context on the governmental environment (i.e. polarization of parties, partisan competition within assemblies, etc...) in which such policies were enacted could also impact movement around the trends of income inequality. To gauge these entities datasets of applicable variables were sources from several institution; including the Bank of International Settlements, Frazer Institute and the World Bank. Preliminary review of the availability of the variables that are associated to the subject matter indicate several limitations and data quality issues which limit the research to a sample group of developed economies. Based on existing research in
the field and an initial review of the characteristics of the data a regression based solution was deemed as the most suitable suite of analytical techniques to investigate this problem.

5.3 Design/Experimentation, Evaluation & Results

To undertake this research a CRIPS-DM structure of six stages was followed (consisting of Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment) with included the development of a Data Quality management mechanism to answer the issues of Outliers and NULL handling that were identified in the initial review and profiling of the raw data. A lag of $t_{+1}$ is applied to the Income Inequality (dependent) variable to allow for the modelling of the PCB (explanatory) variables influence on a future observation. Through a process of evaluating Pearson Coefficients and modelling Multi-Linear Regression models it is found that for neither of the indicators of Income Inequality do any one (correlation) of set (Regression model) PCB variables represent a statistically significantly strong associate. Hence this paper's findings indicates that the PCB has reflected little explanatory power over the oscillation in domestic income inequality along its trend line and that, given the available data, the ideological positioning of nation state governments has had little if any affect on the proportional quality of wealth which is allocated to the upper tail of their national distribution of income.

5.4 Contributions to Body of Knowledge

The findings of this paper add to the conclusion of such researcher as Daron Acemoglu, James Robinson, Kenneth Scheve and David Staasavage that the PCB does not significantly explain economic inequality but mores specifically that its elements also cannot be deemed as been explanatory for variations form their trend. Which means that not only is the political
orientation of governments express limited influence on the overall long term trend in regards to inequality but also that it cannot decidedly explain observed deviations from this trend. Additionally this study also shows that such insignificance is not due either structural nor situation contexts, such as the need for Governments to water-down or negotiate the intensity of policy due to presence of competition within the government, as indicated by the weak correlation/parameter scoring of its Herfindahl-Hirschman Index and the polarization indicators.

Additionally this paper has provide insight into the application of the CRISP-DM structured process for explanatory projects and parameter analysis in the field of PCB and Inequality research. And the issue regarding Data Quality, including Outlier and NULL handling which future endeavours into this research may encounter.

### 5.5 Future Work & Recommendations

Due to the availability of data this paper was limited to the analysis of Income inequality and excluded metrics or Wealth distribution. As this is the key elements in emerging research in the field on Inequality future research opportunity into the relationship into the PCB relation with the distribution of wealth may offer greater insight into overall inequality within a society. Similarly if additional data was made available similar research focusing on Developing Economies could provide insight into the relationship between PBC and inequality in such economies.

Disparity in Tobin's Q (a ratio of a company's assets in relation to its capitalised market value) between economies with differing rule and regulations that govern the operation of the firm and its relations with its stockholder has often been associated with the orientation or makeup of policy decision makers. Using the techniques and DPI data utilized in this page an investigation into this could contribute to the knowledge behind such disparities.
6 Bibliography


02400710610309107211001204403604500002708709806510907910508407805408409

302509200908906606606803009510300811


Sun, W., Wang, X., & Sun, X. (2012). USING MODULAR PROGRAMMING STRATEGY TO PRACTICE COMPUTER PROGRAMMING. In *American Society for Engineering Education* (pp. 1 – 11).


