|  |
| --- |
| Description: 18897 CIC A4 Portrait WordTemp_cropped.jpg |
|  | |
|  | |
| |  |  | | --- | --- | | **SUBJECT NUMBER & NAME** | 36103 Statistical Thinking for Data Science | | **STUDENT NAMES & IDs**  **(SURNAME, FIRST NAME, STUDENT ID)** | Arunachalam, Abishek (SID: 13001262)  Jiang, Benjamin (SID: 12875314)  Pelayre, Anne Gorge (SID: 13102191)  Raghavan, Saminathan (SID:13075597)  Ramal, Miguel (SID: 00060259) | | **TEAM NAME** | SKEPTICS | | **STUDENT EMAIL** | [abishek.arunachalam@student.uts.edu.au](mailto:abishek.arunachalam@student.uts.edu.au), [benjamin.jiang@student.uts.edu.au](mailto:benjamin.jiang@student.uts.edu.au), [annegorge.pelayre@student.uts.edu.au](mailto:annegorge.pelayre@student.uts.edu.au), [saminathan.raghavan@student.uts.edu.au](mailto:saminathan.raghavan@student.uts.edu.au), [miguel.ramal@student.uts.edu.au](mailto:miguel.ramal@student.uts.edu.au) | | **DUE DATE** | 28 May 2018 | | **ASSESSMENT ITEM NUMBER/TITLE** | AT2 Data analysis project, Part B: Report | | * I/We confirm that the work submitted conforms with the university’s guidelines on academic integrity.   *Refer to the UTS policy on ‘Advice to Students on Good Academic Practice’*: <http://www.gsu.uts.edu.au/policies/academicpractice.html>   * I/We am aware of the penalties for plagiarism. This assignment is my/our own work and I/we have not handed in this assignment (either part or completely) for assessment in another subject. * If this assignment is submitted after the due date I/we understand that it will incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.   Please provide details of extensions granted here if applicable \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Signature of Team:** \_\_\_\_\_\_Skeptics\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **­­Date:**  28 / 05 / 2018  If submitted electronically tick here to indicate you agree with the above X | | | |

Contents

[Business Understanding 3](#_Toc515313673)

[Research questions 3](#_Toc515313674)

[Research Approach 3](#_Toc515313675)

[Data collection and wrangling 4](#_Toc515313676)

[CRISP-DM 4](#_Toc515313677)

[Data Sources 5](#_Toc515313678)

[Model 1 5](#_Toc515313679)

[Model 2 5](#_Toc515313680)

[Data Wrangling 7](#_Toc515313681)

[Model 1 9](#_Toc515313682)

[Exploratory Data Analysis - indicators for Model 2 11](#_Toc515313683)

[Initial Phase 11](#_Toc515313684)

[A deeper look into independent variable correlations 14](#_Toc515313685)

[Further Exploration 15](#_Toc515313686)

[Model 2 19](#_Toc515313687)

[Multivariate linear regression: 19](#_Toc515313688)

[Assumptions of linear model: 20](#_Toc515313689)

[Regularization techniques: 22](#_Toc515313690)

[Lasso regression: 22](#_Toc515313691)

[Ridge regression: 23](#_Toc515313692)

[Assumptions of ridge regression: 23](#_Toc515313693)

[Feature extraction models 24](#_Toc515313694)

[Principal Component regression (PCR): 24](#_Toc515313695)

[Assumptions of PCR: 24](#_Toc515313696)

[Partial least squares (PLS): 26](#_Toc515313697)

[Australian GDP Performance 28](#_Toc515313698)

[Calculating GDP 28](#_Toc515313699)

[Australian GDP Data Exploration 28](#_Toc515313700)

[Australian GDP Performance 30](#_Toc515313701)

[Reference List 31](#_Toc515313702)

[Appendix 32](#_Toc515313703)

[Further methods used for displaying variable correlation matrix and use of heatmaps 32](#_Toc515313704)

# Business Understanding

In June 2017, Australia took the world record for the longest run of uninterrupted growth in the developed world (Bagshaw & Massola 2017). It has now been 107 quarters since Australia had a technical recession, defined as two consecutive quarters of negative economic growth.

As one of the most widely used economic indicators, Gross Domestic Product (GDP) is used to gauge the health of a country’s economy (Investopedia 2018). GDP represents the total dollar value of all goods and services a country produced over a specific time period (measured quarterly by the ABS), often referred to as the size of the country’s economy.

Given the importance of having a healthy economy to the wellbeing of a country’s citizens, our team considered it would be important to try and understand what are the factors driving GDP and whether it was possible to predict future GDP of Australia using other information.

Our team viewed choosing this topic as a learning opportunity, to better understand how we as individuals and as a community can contribute to the economy. As a team of data analysts without any formal qualifications in economics, will try to decode the economic jargon and provide insights on the important factors that influence a country’s economy.

# Research questions

As there are two types of GDP that economists use to measure a country’s economy, our regression model will disregard real GDP (economic output adjusted for the effects of inflation) and solely focus on predicting nominal GDP (a country’s economic output without an inflation adjustment). The research questions that we examined with the data:

* **What are the main factors that contribute to GDP in Australia? (Main)**
  + **Does unemployment rate have an effect on GDP?**
  + **Does increase in GDP lead to increase in ASX50 stock prices?**
  + **How does interest rate changes affect GDP?**
  + **If given the right variables can GDP be accurately predicted?**

# Research Approach

There are three ways that a nation’s GDP is traditionally measured, and conceptually they all should deliver the same estimate (Australian Bureau of Statistics 2012). They are:

* The income approach measures income generated by the economy.
* The production approach calculates the sum of the value of goods and services produced by each industry minus the those used in production.
* The expenditure approach measures final expenditures on goods and services.

As it did not matter which method we choose as long as we were consistent in our logic, we choose the expenditure approach to form our starting basis. This decision was not to discount the importance of variables with the other approaches, but due to the expenditure approach for calculating GDP variables having the most readily available information.

In doing so we took a progressive approach to examining factors that contributed to GDP and for model building.

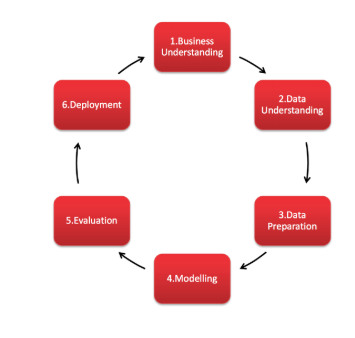
* Model 1 examined the variables of the Expenditure Approach (GDP = C + I + G + NX).
* Model 2 examined and used indicators for the components of the Expenditure Approach e.g. using CPI and Sales as indicators instead of Consumption.
* Model 3 took into consideration other available data sources, not limited to the model e.g. population of Australia.

# Data collection and wrangling

## CRISP-DM

It is necessary to collect data appropriate to the project’s goal and the collected data should be from a credible source and necessary steps are required to convert the raw data to useful information. This is the foremost step for any project as the quality of data is directly proportional to the quality of output. Cross-industry standard process for data mining (CRISP – DM), process model, is used in this project as a sequence of events as mentioned in Figure .

Proper research and brainstorming is done for understanding the different methods for calculating GDP and setting objectives for the project. The three main methods for calculating GDP are the income approach – adding together the factor payments such as total national income, sales taxes; production approach – adding together the factors that contribute the total value of goods and services; and the expenditure approach – adding together the factors that consists of the various types of spending which occur within an economy. A project plan was made to concentrate on calculating GDP by expenditure approach and the indicators that can affect GDP. The expenditure approach of calculating GDP is based on the formula GDP= Consumption or Consumer spending (C) + Government spending (G) + Investment of country (I) + Business capital expenditures (NX).



As per the project plan, the factors affecting GDP are prioritized and grouped to fit the models. The next step followed is the collection of data. Data are collected as components of each of the components (C, I, G and NX) for model one and indicators for the components that might affect GDP for the second model. External variables such as population growth, unemployment rate, Human Development Index (HDI) were also collected. After researching information for economic indicators in Australia, most sources including an e-brief article on the Parliament of Australia website (Woods n.d.) indicate the Australian Bureau of Statistics (ABS) as the main source of economic statistics in Australia. The ABS site provides a free tool: ABS.Stat that offers web browsing and web services interfaces to display and extract data on multiple themes such as Economy, Health, Industry, Labour, People, Census and other snapshots of Australia.

## Data Sources

### Model 1

Data variables for model one is obtained from ABS. The data variables include GDP, Consumption, Investment, Government spending, Exports and Imports to calculate the net exports. The data measures are collected yearly from 1960 to 2017. All the variables are extracted from the same source – ABS and all the variables are measured in same unit – millions. Table 1 provides the source of data for model 1.

|  |  |
| --- | --- |
| **Expenditure Approach** | **Link** |
| GDP | <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/5206.0Dec%202017?OpenDocument> |
| Consumption |
| Investment |
| Government spending |
| Net Exports |

**Table 1**

### Model 2

* **Gross Domestic Product**– Quarterly data in millions from 1960 sourced from The Organisation for Economic Co-operation and Development (OECD). Percentage change in GDP is also extracted.

* **Human Development Index (HDI)** – Yearly data from 1990 sourced from the United Nations Development Programme (UNDP).
* **Interest Rates & Exchange Rates** – Monthly data from 1969 sourced from the Reserve Bank of Australia (RBA). Exchange rates are in USD. Interest rates are in percentage format.
* **Consumer Price Index (CPI), Sales, Expenditure, Labour Force & Balance on Goods and Services**- data sourced from Australian Bureau of Statistics (ABS).
* **Unemployment –** Quarterly data from 1966 sourced from OECD.
* **Total Population** – Quarterly data from 1981 sourced from ABS
* **Stock price** – Monthly data from 1982 sourced from ASX50.

Table 2 provides the data source links:

|  |  |  |
| --- | --- | --- |
| **Expenditure Approach** | **Indicator** | **Link** |
| **GDP** | GDP | <http://stats.oecd.org/restsdmx/sdmx.ashx/GetData/QNA/AUS.B1_GE.CPCARSA.Q/all?startTime=1960-Q1&endTime=2018-Q1> |
| **Consumption** | Consumer Price Index (CPI) | <http://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0> |
|  | Sales | [http://stat.data.abs.gov.au/#](http://stat.data.abs.gov.au/) |
| **Investment** | 3-month Monthly Average Interest Rates(%) | <https://www.rba.gov.au/statistics/historical-data.html#interest-rates> |
|  | Expenditure | [http://stat.data.abs.gov.au/#](http://stat.data.abs.gov.au/) |
|  | Labour Force | <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/6202.0Mar%202018?OpenDocument> |
| **Government spending** | Human Development Index(HDI) | [http://hdr.undp.org/en/data#](http://hdr.undp.org/en/data) |
|  | Unemployment | <https://data.oecd.org/unemp/unemployment-rate.htm> |
| **Net Exports** | Balance on Goods and Services | <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/5368.0Feb%202018?OpenDocument> |
|  | Exchange rates | <https://www.rba.gov.au/statistics/historical-data.html#exchange-rates> |
|  | Total population | <http://stat.data.abs.gov.au/restsdmx/sdmx.ashx/GetData/ERP_QUARTERLY/1.0.3.TT.Q/all?startTime=1981-Q3&endTime=2017-Q3> |
|  | Stock Price | <https://www.asx50list.com/> |

**Table 2 – Data Sources**

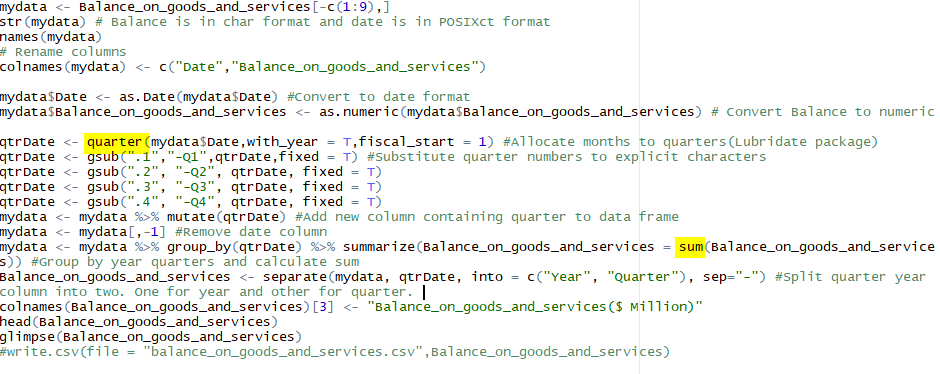
## Data Wrangling

The collected data needs to be transformed or modified for improving the quality of data, commonly known as wrangling of data. The data for model one is collected from the same source and all the data are measured from same period with same measures. Therefore, data for model one is already in good quality. The only checks need to be done are checking for missing values or outliers. The below figure shows that there are no missing values in data for model 1.

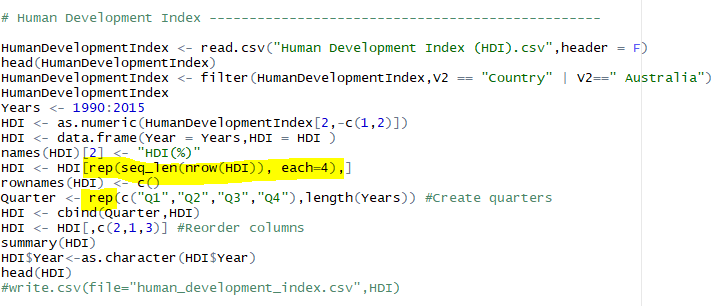


Data for model one is clean and tidy but data for model two is collected from different sources and the variables are of different time periods and intervals. For example, GDP data is available from 1960 and the interval is quarterly data whereas Stock return aggregated value of top 50 companies is available from 1982 and the data is available as monthly data. Therefore it becomes a major challenge in preparing the master dataset.

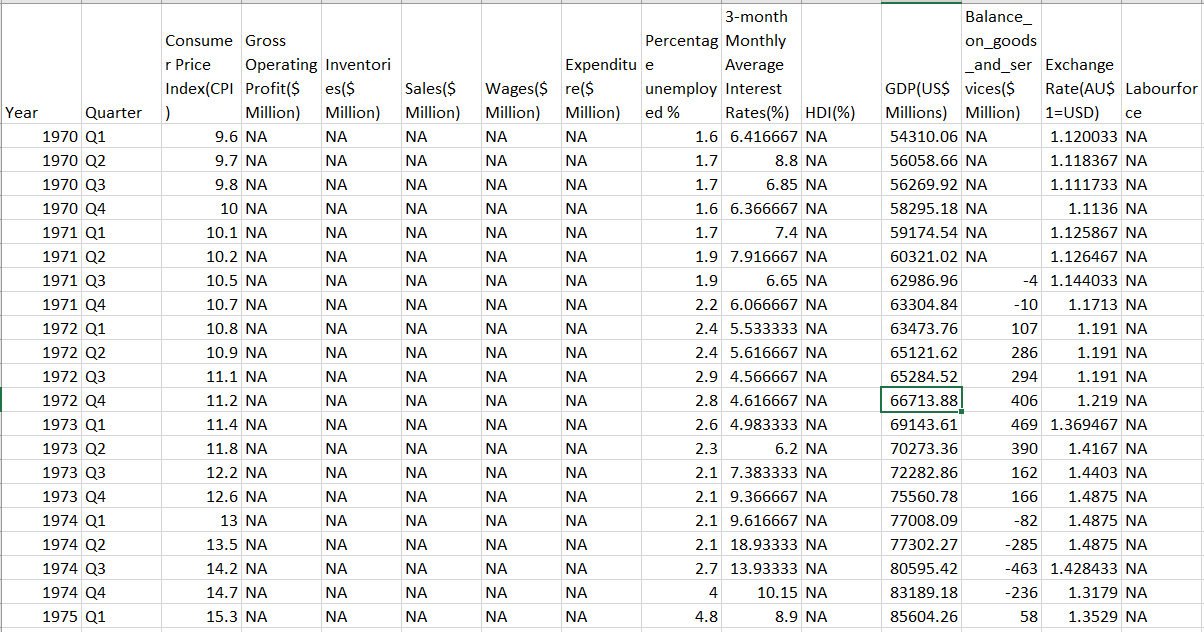
The monthly data of variables are aggregated into quarters and the values for the observations are calculated depending on the type of variable. For example, Stock return aggregated value of top 50 companies are grouped into quarters and value for the variable is calculated as mean stock return value of individual values of respective quarter, similarly Balance of goods and trade variable observations are grouped into quarters but the value for the variable is calculated as sum of individual values of the quarter. A sample of the code snippet is shown in the below figure.



Human Development Index data is available as yearly data. Therefore to maintain data consistency the data is grouped into quarters and the yearly value is repeated for all quarters. The code snippet is shown in the below figure.

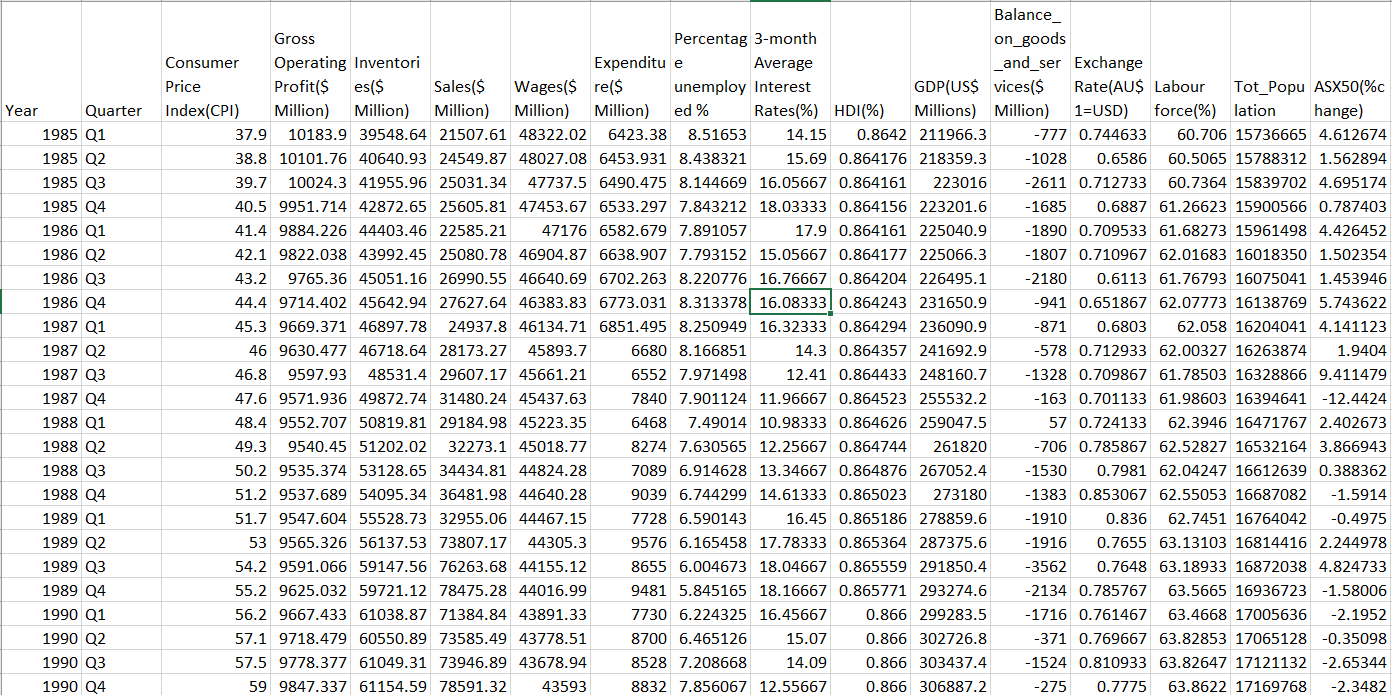


The data is now wrangled such that all the variables for the second model are in quarterly interval. However, the data collected is not available from the same year for all variables. The CPI data which is available from 1960 Q1 is taken as the base dataset and all the other variables are left joined to prepare the master dataset. Since there were missing values for most of the variables decision was made to have master dataset from 1970 Q1. The sample of the master dataset is shown in the below figure.



The master dataset still consists of many missing values and, from above figure, it is clear that the missing values are not at random, but they are at continuous observations. Normally missing values are handled by removing rows or imputing mean/median value for replacing missing values. This is not feasible here because the values are not missed at random and even predicting the missing values from other variables is also nearly not possible because many variables are missing values for the same observations. However a time series imputation can be done which will be appropriate for this time series dataset. The consideration of the 5-10% thumb rule of missing values which is imputation needs to be done only when the number of missing values are below 10% of the original available dataset leads to further stripping of ten years of data that is from 1980 Q1. Time Series imputation is done using a R package imputeTS and Kalman filter algorithm is used for imputation and a clean tidy master dataset is created.

A sample of final wrangled dataset is shown in the below figure.



# Model 1

To start, Model 1 investigated the importance of the variables used in the expenditure formulae (GDP = Consumption(C) + Investment(I) + Government spending(G) + Net Exports(NX)). A simple linear regression model was used to examine the variables.

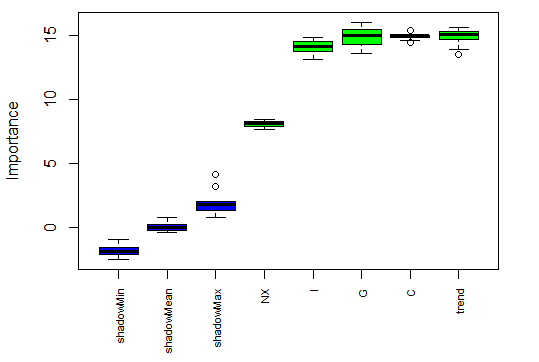
The hold out method (Wikipedia 2018) was used to split the available data into a training set and testing set. As this was a time series, the training and test split was not done randomly, but the training set corresponded to first 70% of the available data and the testing set corresponded to remainder 30% of the available data for each combination. This was to ensure the model would not be built using training set data it would not have had access to at the time of prediction (using test set data). Given this was a time series of annual data, a trend variable was also added.

*Table 1: Model 1 variables*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Average | Pr(>|t|) | varImp score |
| Consumption (C) | 455752 | 8.30e-07 | 5.97 |
| Investment (I) | 143344 | 8.98e-08 | 6.71 |
| Government Spending (G) | 190278 | 0.045887 | 2.07 |
| Net Exports (NX) | -1589 | 0.000576 | 3.79 |

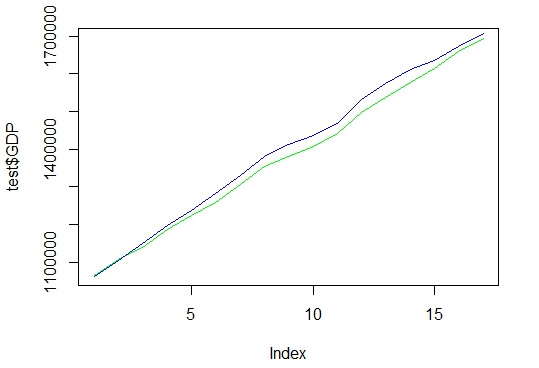
What the simple linear regression model found was that whilst the Consumption variable had the highest average (three times the size of Investment), Investment was found to be the variable that impacted the model the most (based on both P-value and varImp score).

Similarly with Net Exports even though it had a negative average, and was much smaller than government spending average, it was found to have a greater impact on the linear regression model than government spending (based on both P-value and varImp score).



Alternatively, the Boruta package was also applied to the data and it showed a different result. The weighting Boruta package gave to net exports in part reflected the size of this amount when compared to the other variables. It was interesting however, that despite Consumption doubling or tripling the size of Government Spending and Investment respectively, the Boruta package gave similar weightings to all three variables for their importance to GDP.

The linear regression model generated by the training data set was then applied to the test data.



The model fit (as measured by R-squared) was 99.9, whilst accuracy of predictions (as measured by RMSE) was 23384.34. Whilst the simple linear regression model assisted us in better understanding the variables we were dealing with and how they related to GDP, the overly high and almost perfect R-squared highlighted the questionability of using a simple linear regression model to predict GDP with these variables.

# Exploratory Data Analysis - indicators for Model 2

## Initial Phase

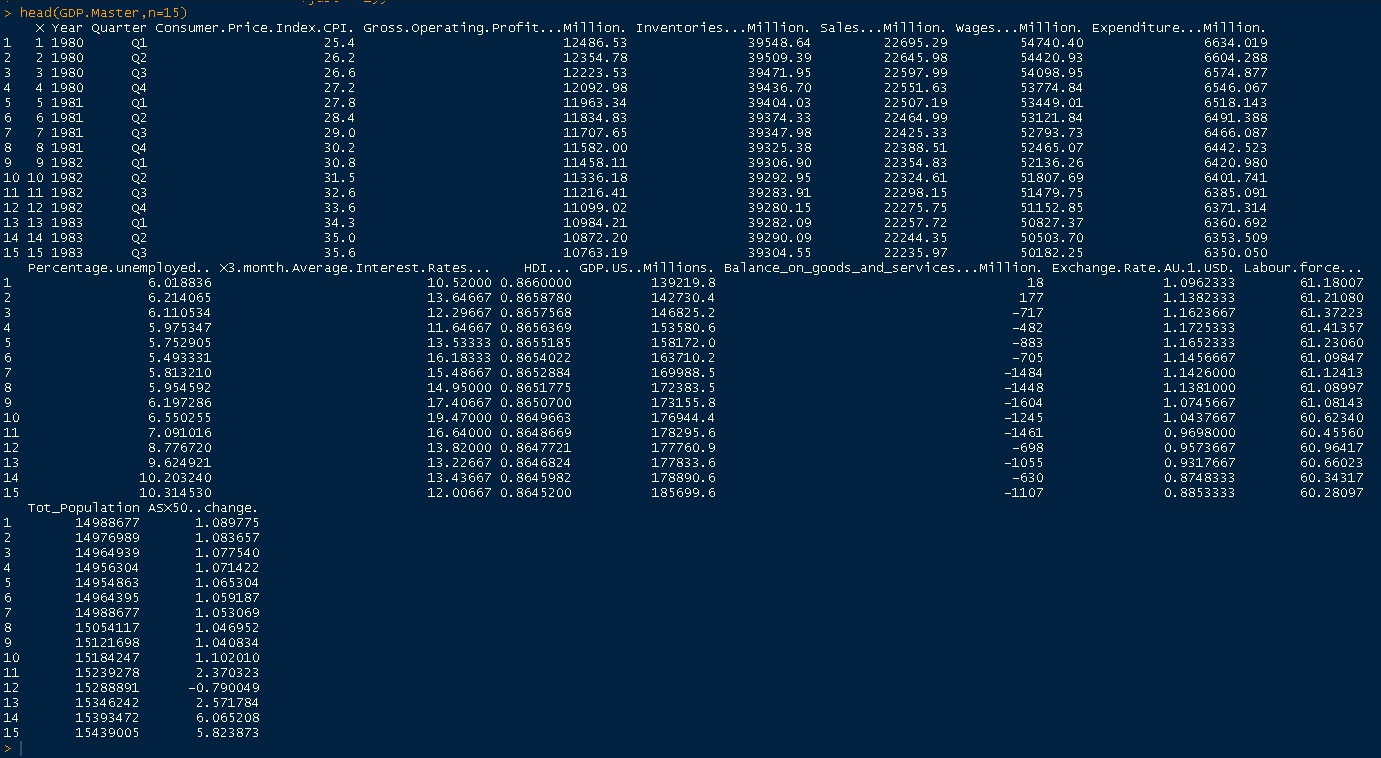
Australian Gross Domestic Product (GDP) is a complex econometric calculation and the starting limitation of this project was not to replicate the real model for calculating GDP but instead, selecting one approach for its measure called the ‘Expenditure approach’ and from this approach, select very few (two or three) independent variables (indicators) from each component (Consumption, Investment, Government Spending, Net Exports) to explore and analyse their influence as factors contributing to the GDP calculation. The variable selection was mainly based on the ease of identifying them as components of GDP and on the availability to obtain the data.

With that premise we collected data on 15 variables:

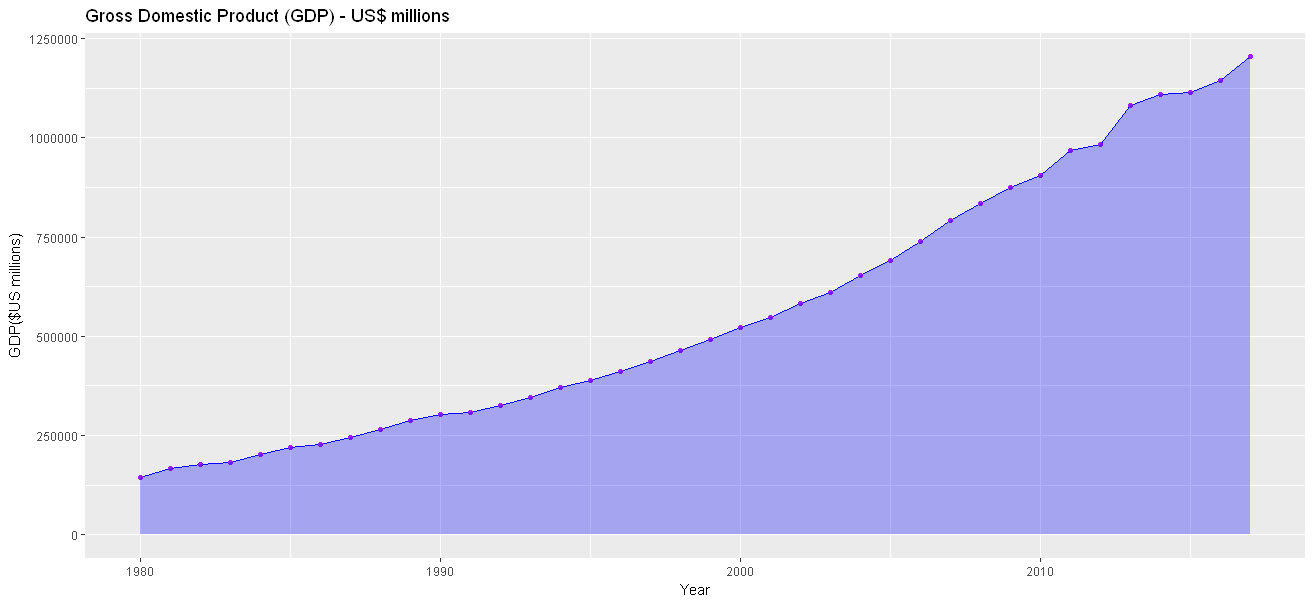
|  |
| --- |
| * Consumer Price Index (CPI) |
| * Gross Operating Profit ($ Million) |
| * Inventories ($ Million) |
| * Sales ($ Million) |
| * Wages ($ Million) |
| * Expenditure ($ Million) |
| * Unemployment (Percentage unemployed %) |
| * Interest Rates (3-month Average %) |
| * Human Development Index (HDI %) |
| * GDP (US$ Millions) |
| * Balance on Goods and Services ($ Million) |
| * Exchange Rate (AU$1=USD) |
| * Labour Force (%) |
| * Total Population |
| * Change in trade value of top 50 performing companies (ASX50 %change) |

During early data exploration, discrepancies were identified with the frequency of data collection over the time period, most of the economic values were found recorded at quartile intervals, however some had monthly or daily frequency and required to be transformed into quartile deductions to maintain homogeneity with other data.

The identification of missing data for various variables between the years 1970 and 1980 were a determining factor to exclude this data from further analysis. Our final dataset contains data from the selected variables in quartile measures since 1980.

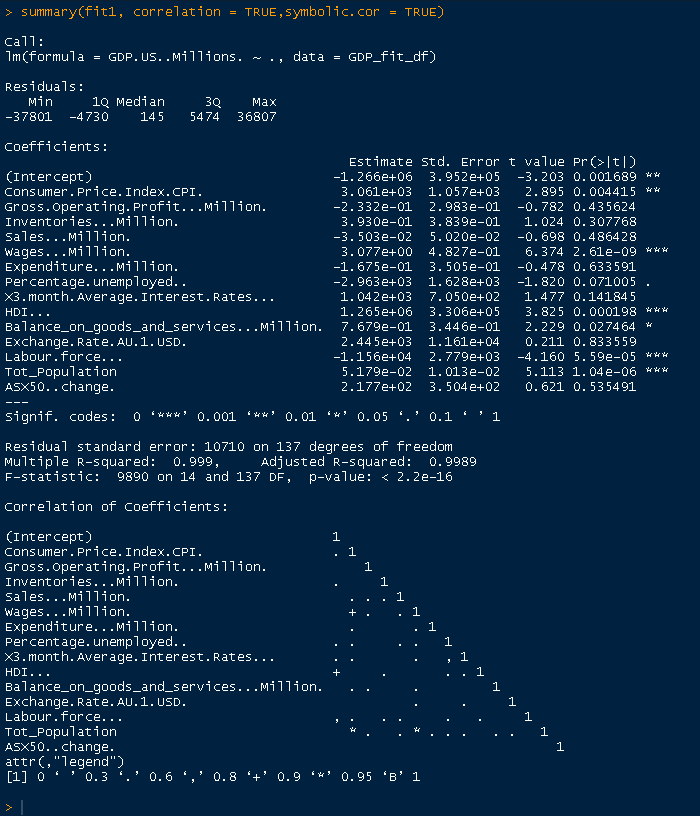


Initially some plots were produced to understand the data a little more. In the graph below the growth (amount) of GDP over the years has been plotted and as observed it has been steadily growing with no noticeable decline.



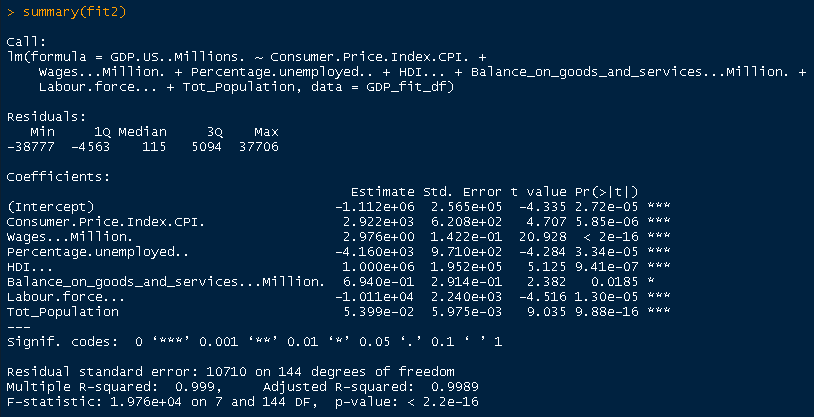
From the plot above, we can observe the growth (amount) of our dependent variable: GDP looks continuous in nature which lead us to the assumption of the relationship between the growth of our dependent variable and our independent indicators may be linear in nature. Since we have a few independent variables, our initial consideration was to look into a Multiple Linear Regression technique for estimating GDP.

A multiple linear regression model was fitted using all independent variables to see how the regression coefficients fit into a linear model. The following summary was obtained:



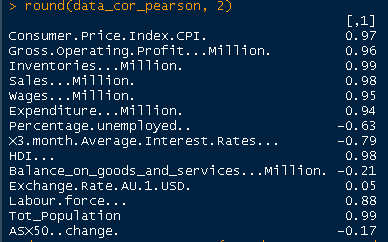
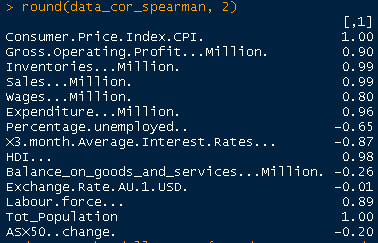
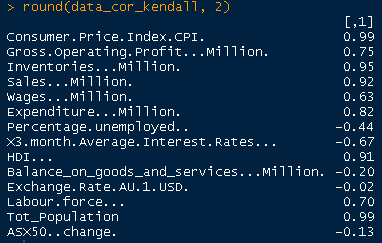
From the summary, the coefficient significance and adjusted R-squared are may be indicative there are too many variables for a proper fit of the linear regression model.

Based on the significance of coefficients from the first model fit, a new regression fit was done by using only those variables indicated with more significance to see the difference in results however it produce similar results:

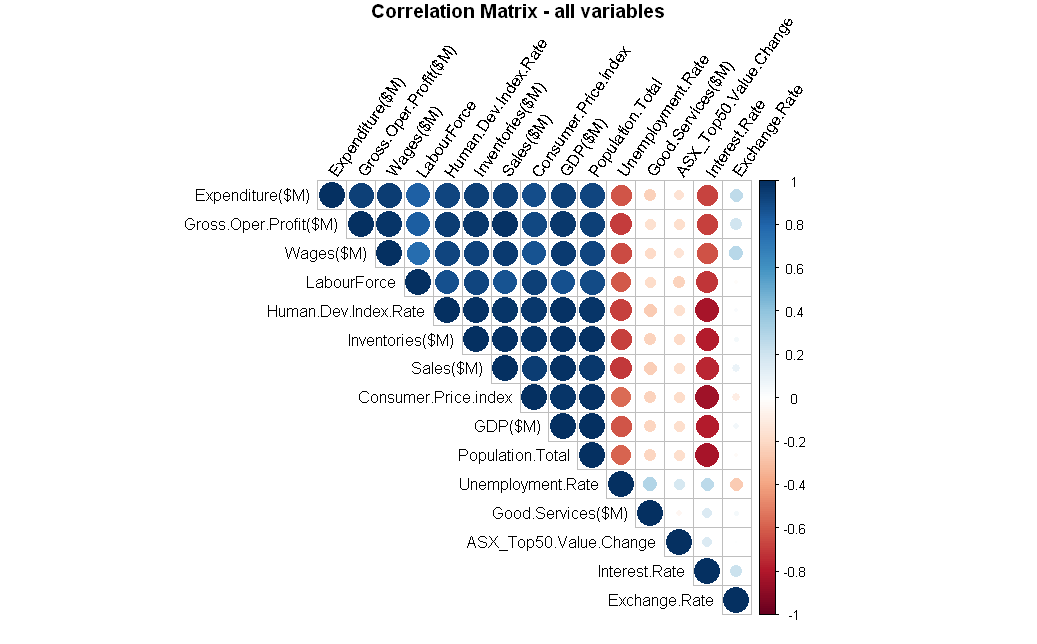


## A deeper look into independent variable correlations

To further explore correlations of the independent variables , we separated the dependent variable: GDP (GDP.US..Millions.) and calculated correlations using y=GDP, x = the rest of variables. Three methods of correlation (Pearson, Spearman and Kendall) were run against the data:

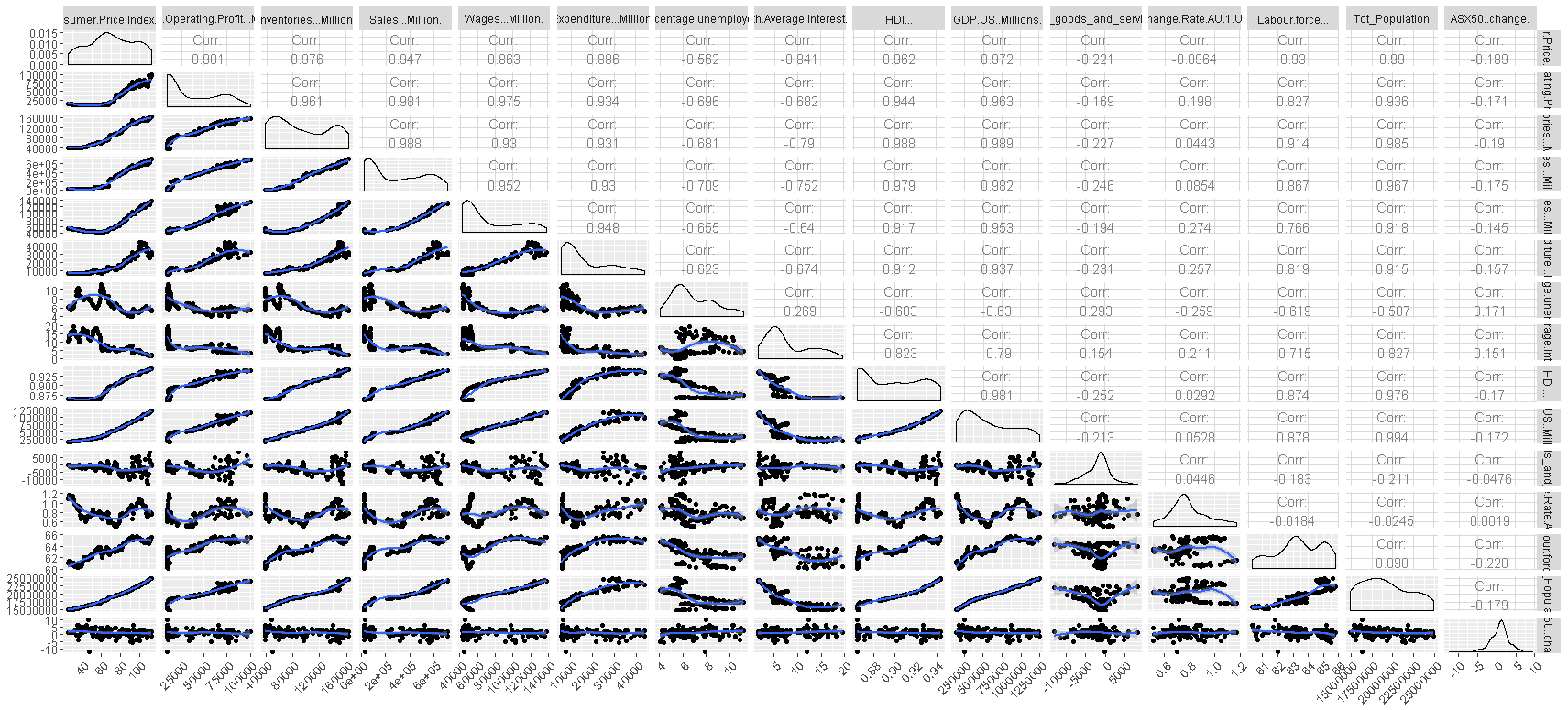
  

In addition we produced a visualisation of the correlation matrix for all our variables, where positive correlations are displayed in blue and negative correlations in red colour (the names of the variables were adjusted for better display in plot below):



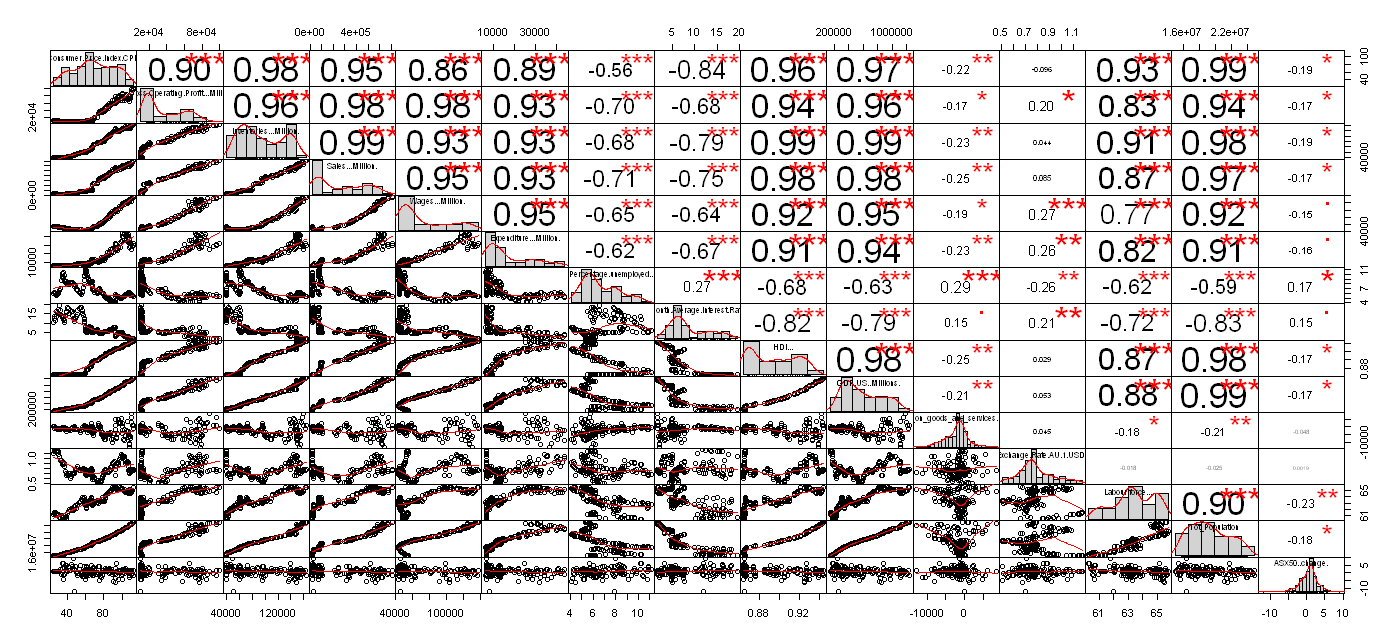
## Further Exploration

To explore the interrelations between the independent variables and any correlations between them, a pairs plot was made to check for visible trends.

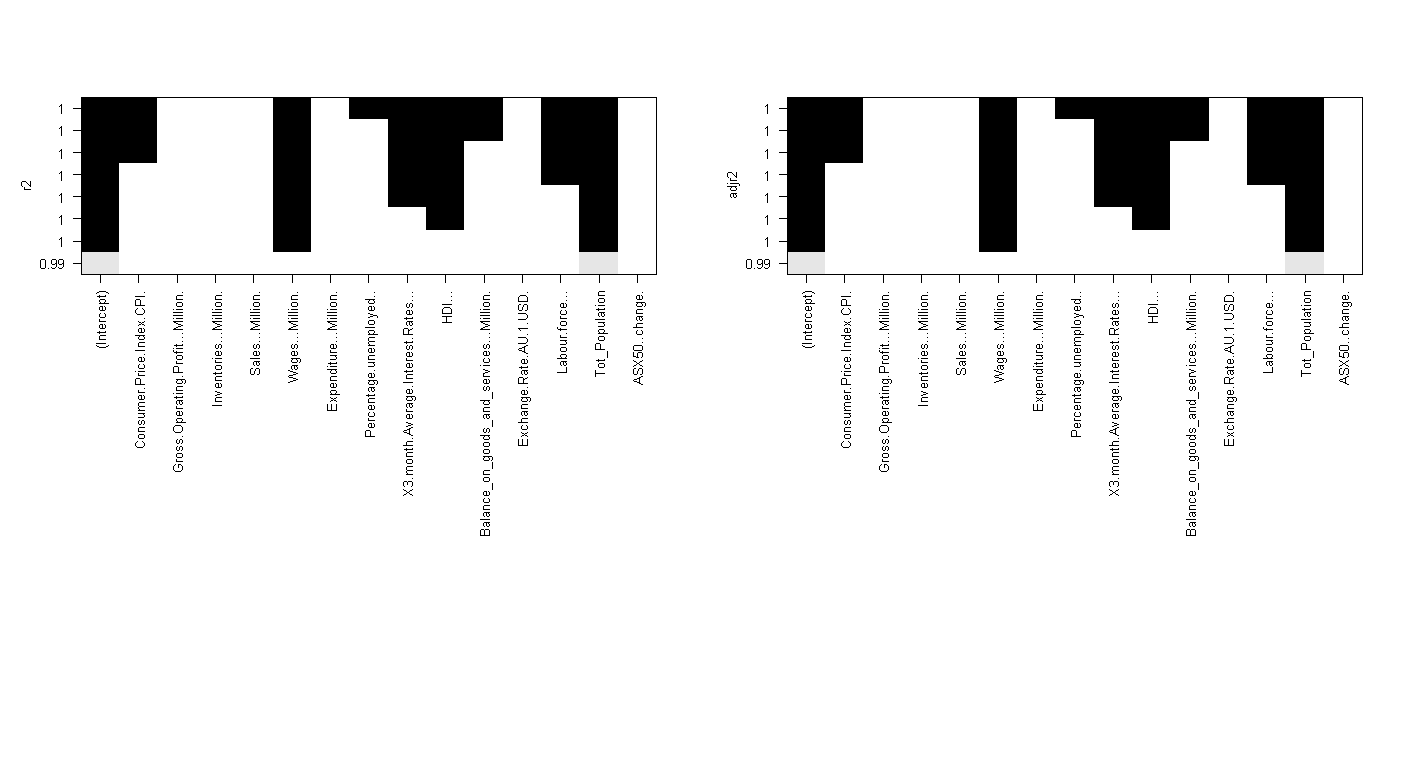


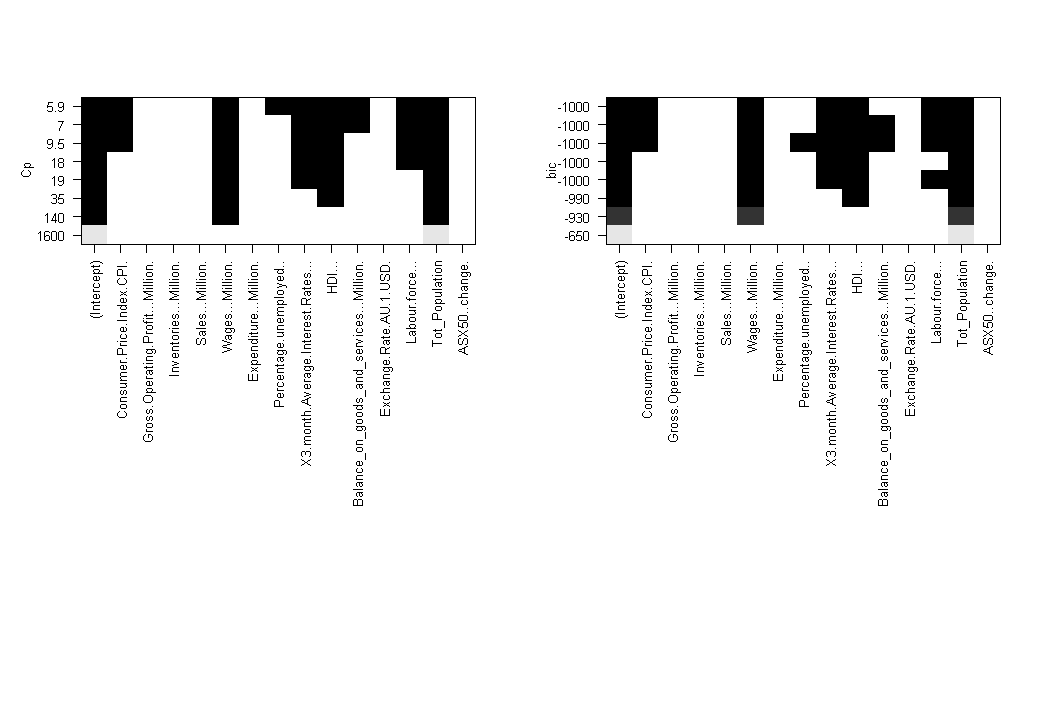
A correlation matrix chart produced by using the function: **chart.Correlation** from the library ("PerformanceAnalytics") gave us a similar display where the distribution of each variable is shown on the diagonal.

* On the bottom of the diagonal, the bivariate scatter plots with a fitted line are displayed
* On the top of the diagonal, the value of the correlation plus the significance level as stars
* Each significance level is associated to a symbol : p-values(0, 0.001, 0.01, 0.05, 0.1, 1) <=> symbols(“\*\*\*”, “\*\*”, “\*”, “.”, " “)



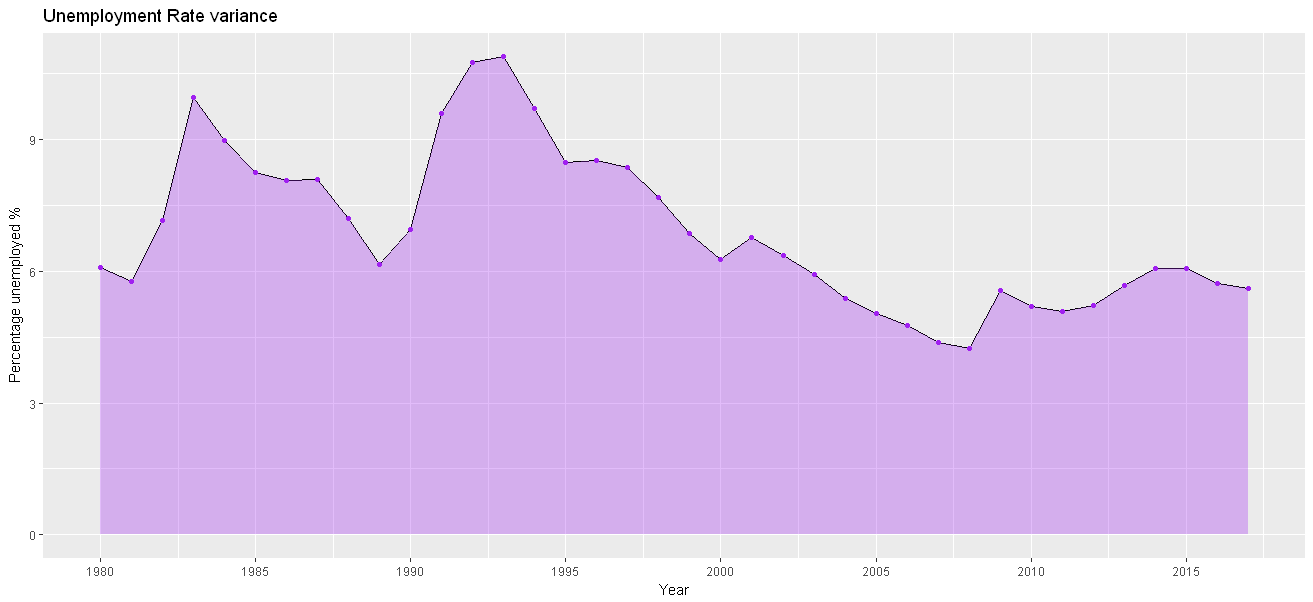
While exploring for subset selection, we tried using function **regsubset** from **library(leaps)** to explore variable selections using “exhaustive” search method. The results were unsurprisingly similar to the significant coefficient from our initial model.





In addition to the exploration for variable selection, we wanted to identify if there was any relation between the yearly change (percentage of growth) of GDP versus un-employment rate.

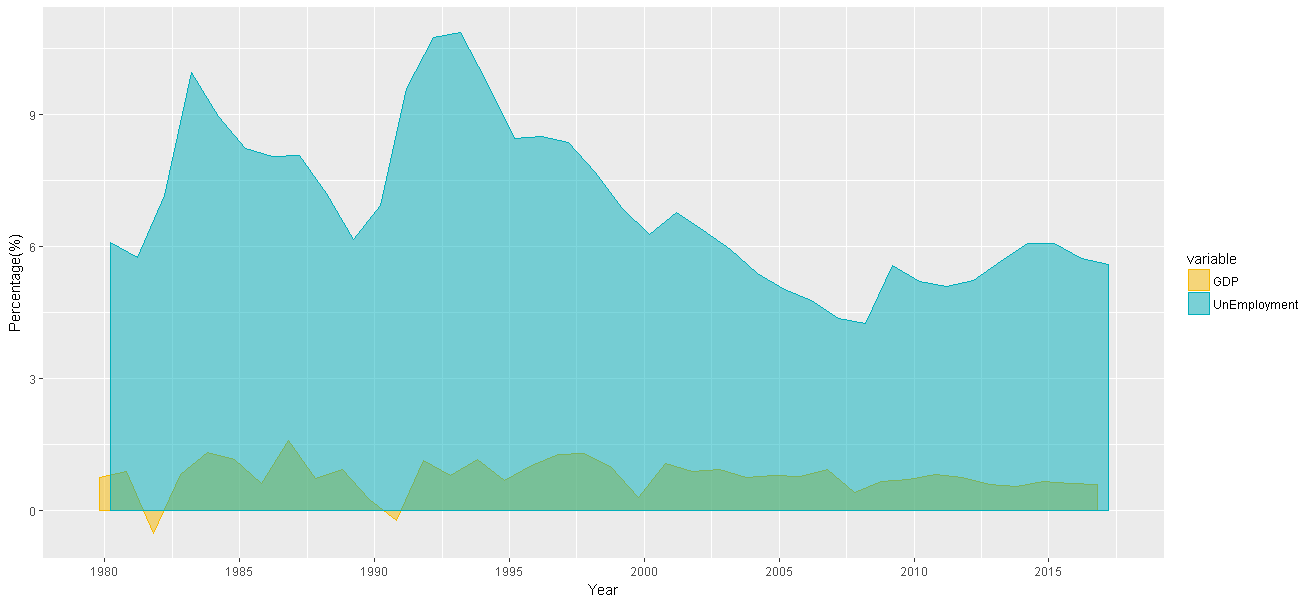
Within our master dataset we brought in the unemployment rate figures and a simple graph shows how this has changed over the years



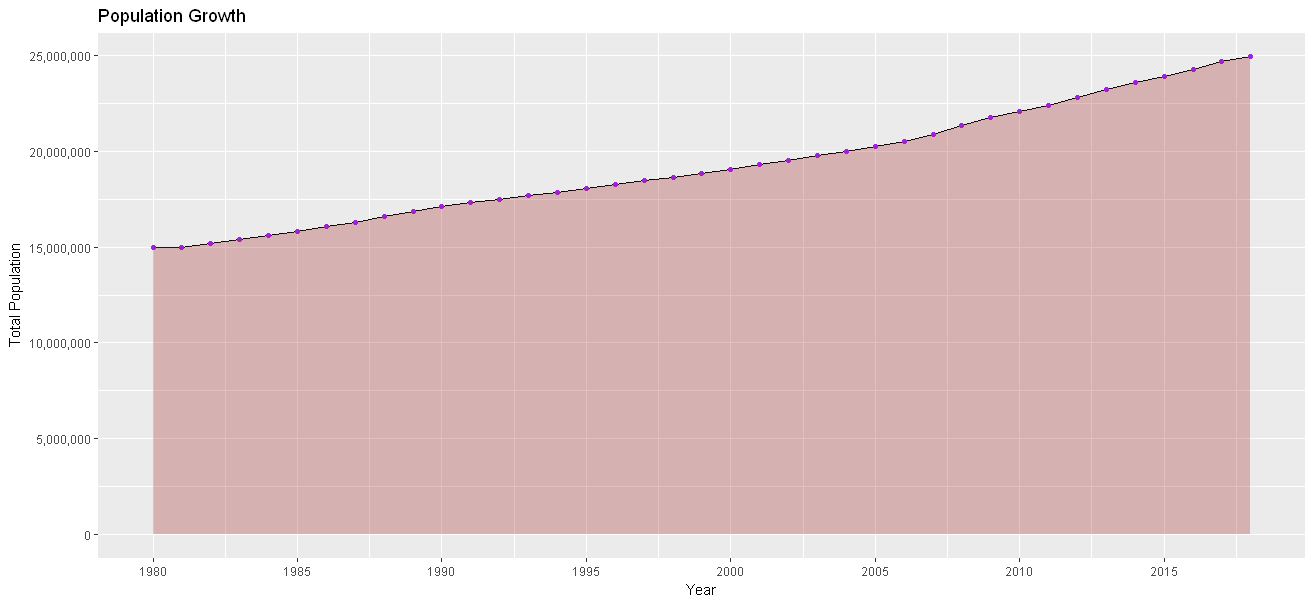
For the purpose of correlating the unemployment rate with GDP variance, a new dataset of quarterly GDP variation percentage was brought into our set. This was used to create a new data frame of aggregated yearly variations as shown below:



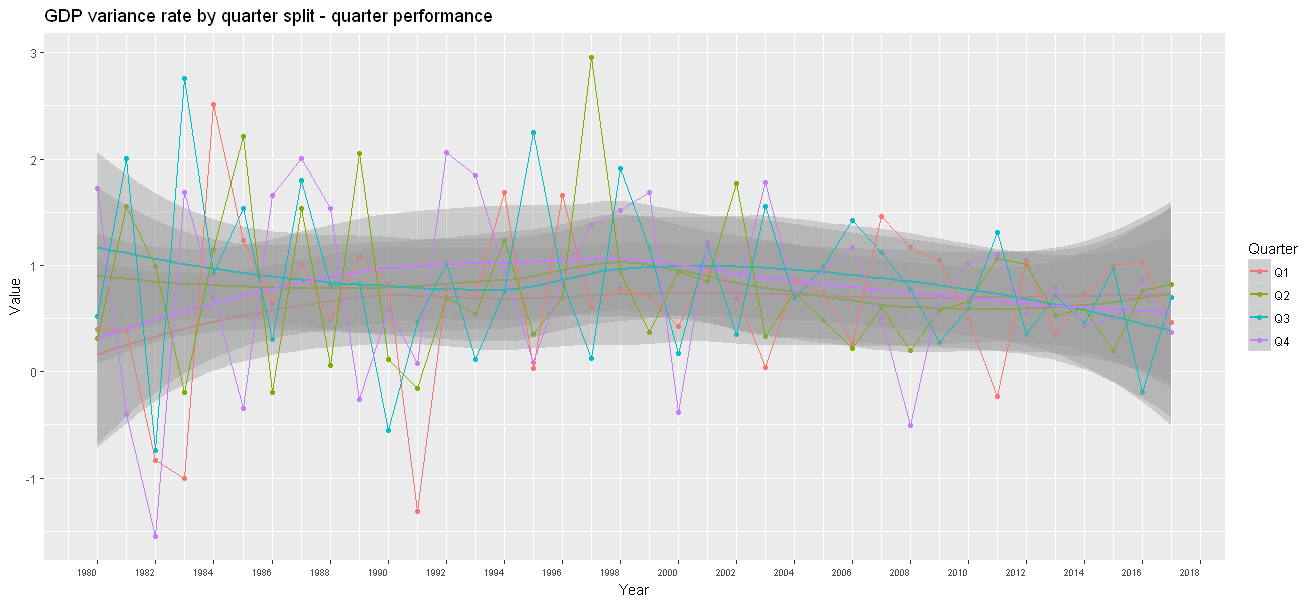
With this data, we were able to graph the relation between GDP change vs un-employment over the years as pictured here were we can observe there is a similar pattern in the yearly changes:



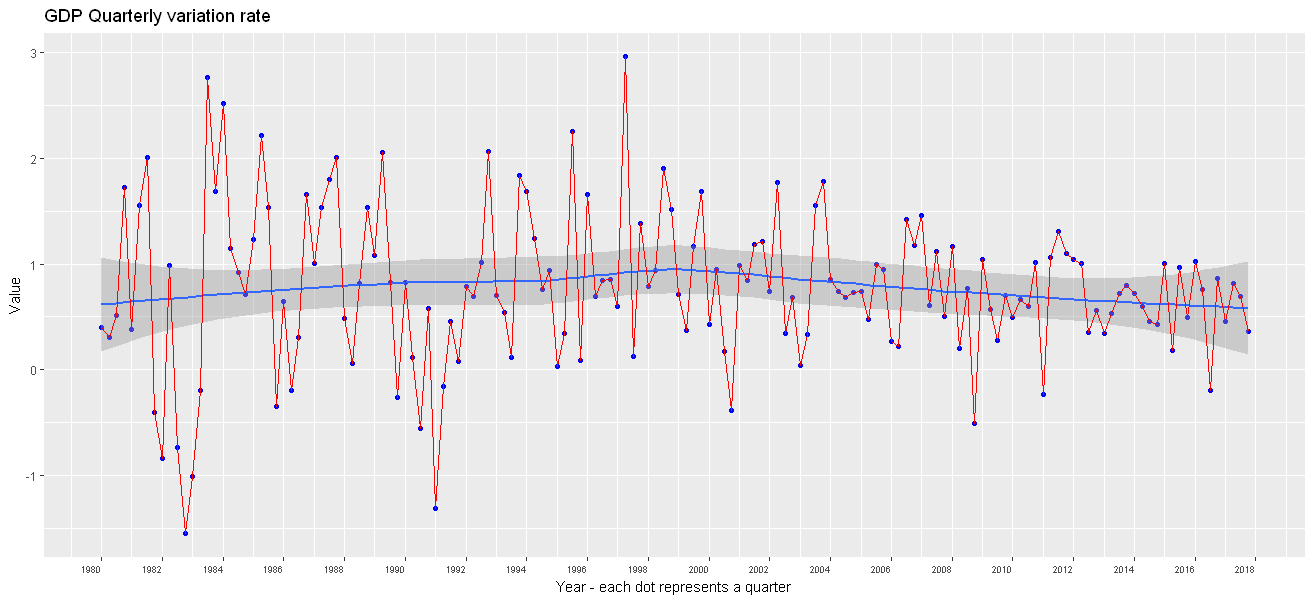
Similarly, a simple plot can provide a nice picture on how steady population in Australia has been growing over the past years. There may be a direct relationship between our growing population and higher growth of GDP. As identified through our initial model, there is a correlation between population growth, consumer price index and wages.

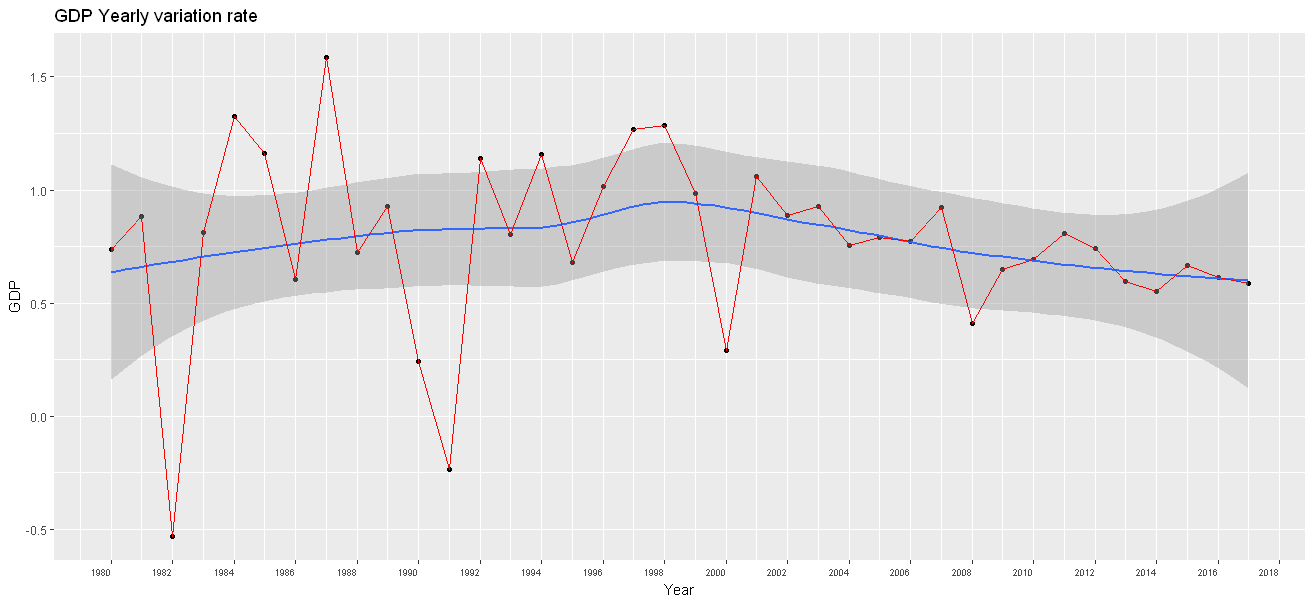


The variations rates of GDP data was also further explored making use of the quarterly and yearly aggregates. By using a split of quartile data we were able to display the quarterly performance of GDP per year since 1980. This was mainly done to explore for any concealed patterns that may be of interest.



And to see the pattern of variations at quarterly and yearly intervals, we produce the following graphs representing the quarterly and yearly variation average for Australian GDP:



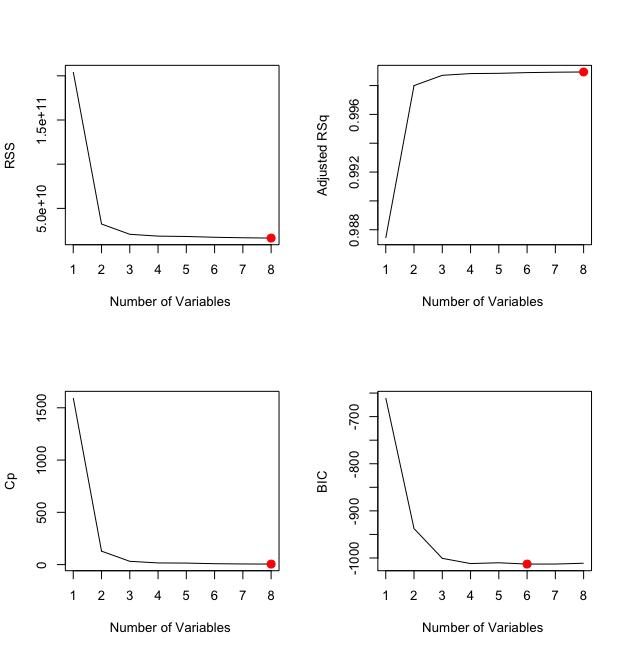


# Model 2

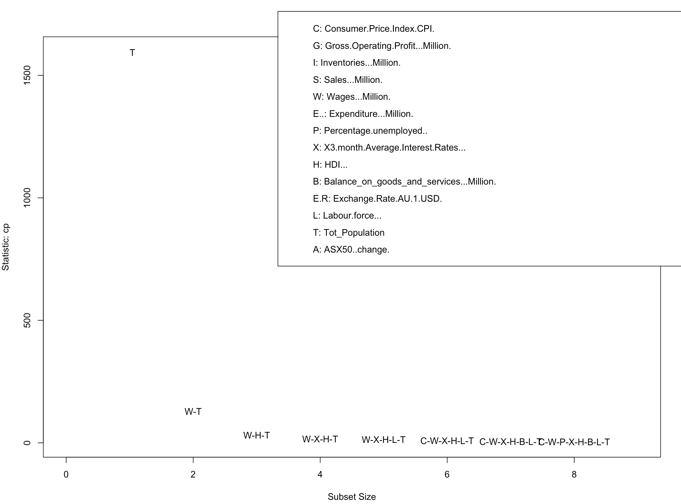
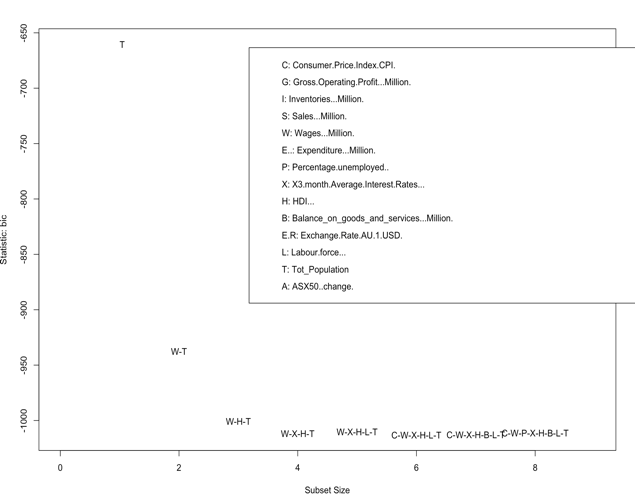
## Multivariate linear regression:

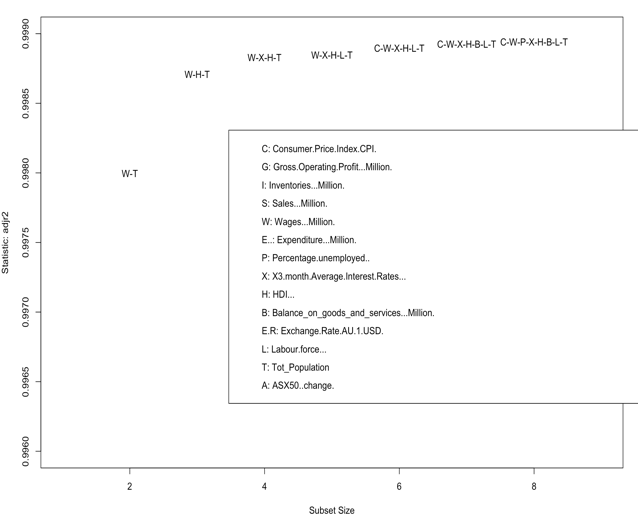
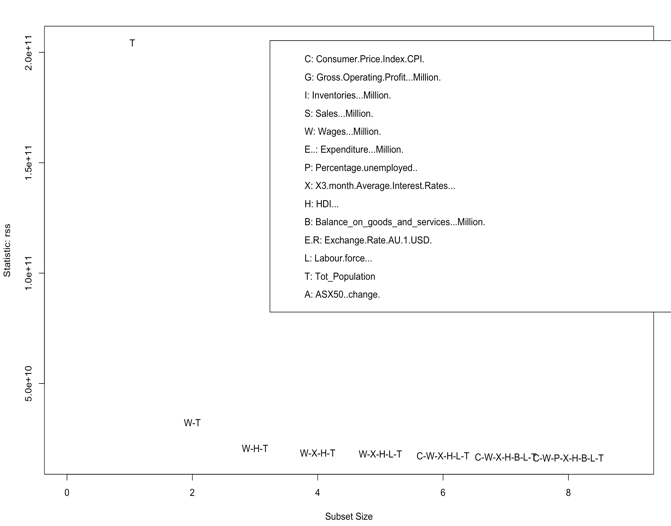
After examining the correlation plot during the exploratory data analysis, it became evident that most of the predictors were correlated with each other leading to multicollinearity. Effects like this could cause coefficient estimates to be erroneous and show significance with target variable even when there is no relationship. Avoiding such a situation, it was decided to drop some of the variables while modelling. To pick the most optimal predictors best subset selection was used. Best subset selection uses various subset of predictors to arrive at the best model based on Mallow’s Cp, Bayesian Information Criterion(BIC), Adjusted-RSquare, Residual sum of squares(RSS).

The best model for each of the criterion is picked and plotted in the figure below. Majority of them have selected model with 8 predictors to be the best model.



As we have the problem of collinearity and our aim was to pick lesser predictors without losing prediction accuracy consequent sub-optimal predictor combinations are examined. The below plots show the best predictor combination in each category. The model with four predictors is chosen as the measures are negligibly different from the best model while reducing the predictor count.



Each alphabet of the 4-variable model denotes:

**W – Wages**

**X – Interest rates**

**H- HDI**

**T- Total population**

There is no prediction feature for best subset regression. So, the selected predictors were used once again in the linear model. Variance inflation factor(VIF) was measured and the values for the predictors had reduced compared to the previous models. A check of the assumptions of linear model for the data was done.

### Assumptions of linear model:

**Linearity –** There exists a linear relationship between X and Y.

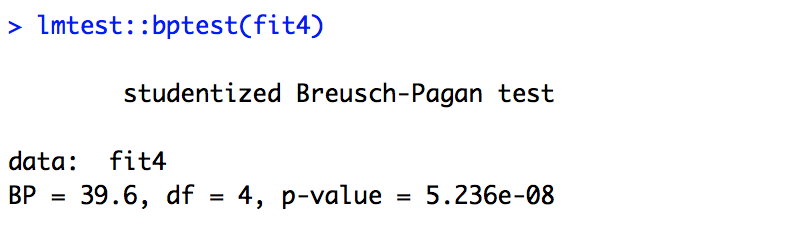
**Homoscedasticity-** The variance of the residuals is same for all values of X.

**Independence –**Residuals are independent to each other.

**Normality –** The values of X and Y are normally distributed.

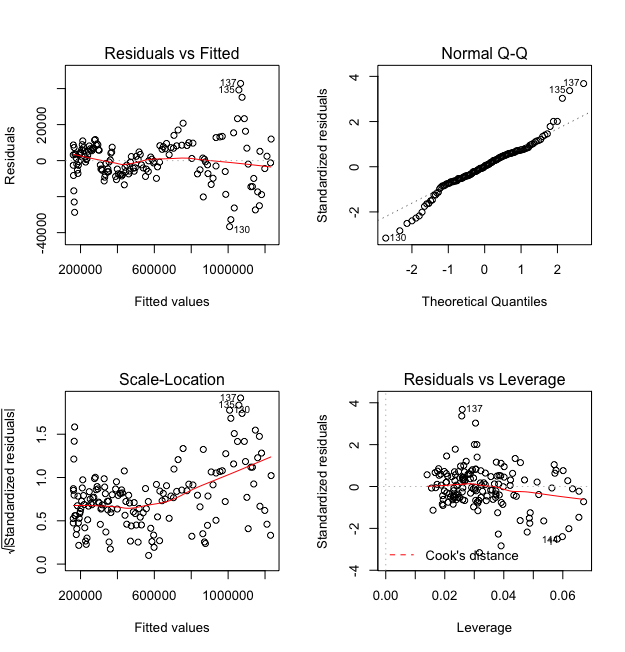
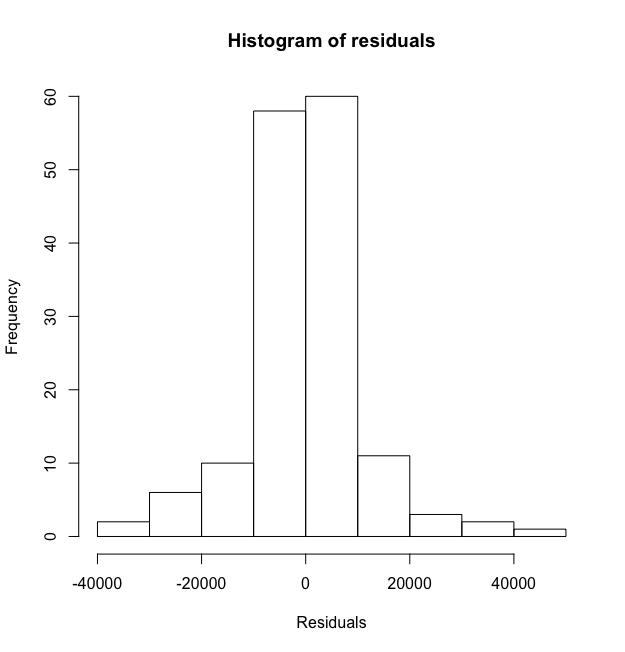
The left-side plot on the first row is a check for linear relationship between X and Y and homoscedasticity. The red straight line show there exists a linear relationship. But the residuals are not evenly spread, and the variance of the residuals get broader towards the end. This give intuition that there is heteroscedasticity in data and one of the model assumption in violated.

To verify the intuition **Breush Pagan Test is run on the fitted model. A value as low as 5.236e-08 was observed which is very less than the significance level of 0.05. So we reject the null hypothesis that the variance of the residuals are constant and infer heteroscedasticity is present in data.**



The right-side plot on the first row is a check for normal distribution of residuals. Residual of each observation is plotted along a 45degree QQ-line. Most of the residuals are along the line except for a few in the tails. To visualize better, a histogram is plotted with the residuals. The bell-shaped curve of the histogram shows the residuals are normally distributed satisfying the assumption. We can see both the plots have identified observations 135 and 137 as outliers.

The first plot in the bottom row shows scaled linear relationship between residuals and fitted values and the second plot has picked observation 137 to be influential leverage point that affect the slope.

Heteroscedasticity in the data can be rectified using Box-Cox transformation. But since each observation in the data is about the economic indicators of Australia they depend on the previous observations and do not satisfy the Independence assumption.

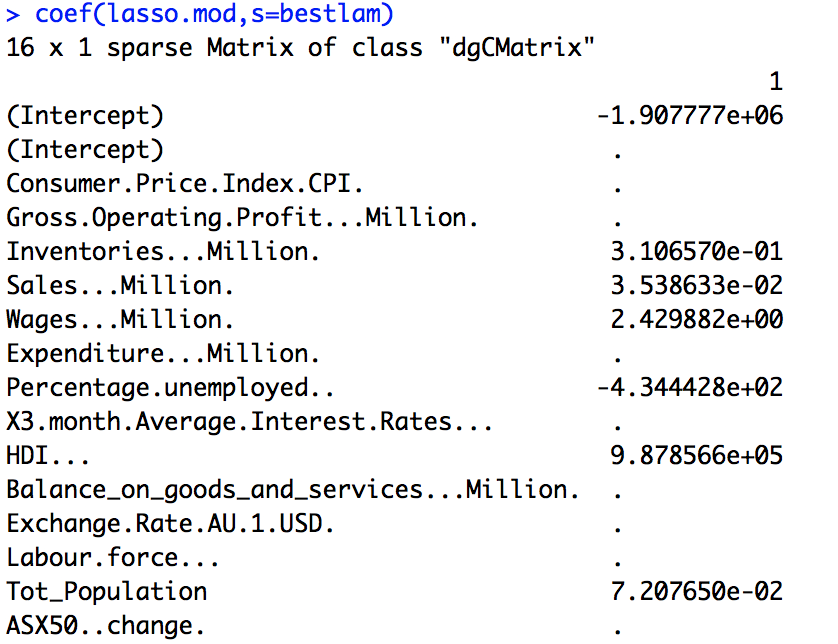
Having said that we try to run regularization models on the dataset.

# Regularization techniques:

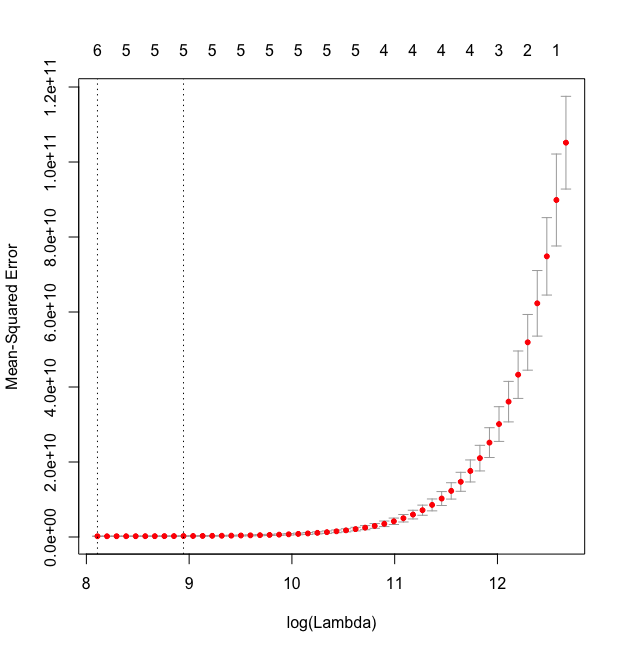
## Lasso regression:

Lasso regression is a regularization technique that shrinks the coefficient estimates of some of the predictors to zero thereby reducing the number of dimensions in the model. They handle multicollinearity well and do not make any prior assumptions about the residuals in the data. Lasso considers the joint distribution of Y on X and does not assume Y to be generated from a linear model. So, it considered to be a robust model for this setting.

Lasso regression was run on the data and the best lambda (regularization parameter) chose the following model.



The model gave more weightage to Human Development Index(HDI), Total population and percentage unemployed while aggressively shrinking and eliminating other predictors. The plot below shows how lasso has reduced the predictor count as lambda increases. The two vertical lines shows the best lambda and lambda value with 1 standard error.



A prediction was done with the best lambda model and a **Mean Absolute Error (MAE)** of **9991.27** was estimated on the test set. Considering the values in millions scale for GDP the error value seemed very low. Lasso did a good job on the dataset.

## Ridge regression:

Ridge regression is another type of regularization technique which uses L2 Norm to shrink the coefficients of the predictors near to zero but does not eliminate any predictor in the process.

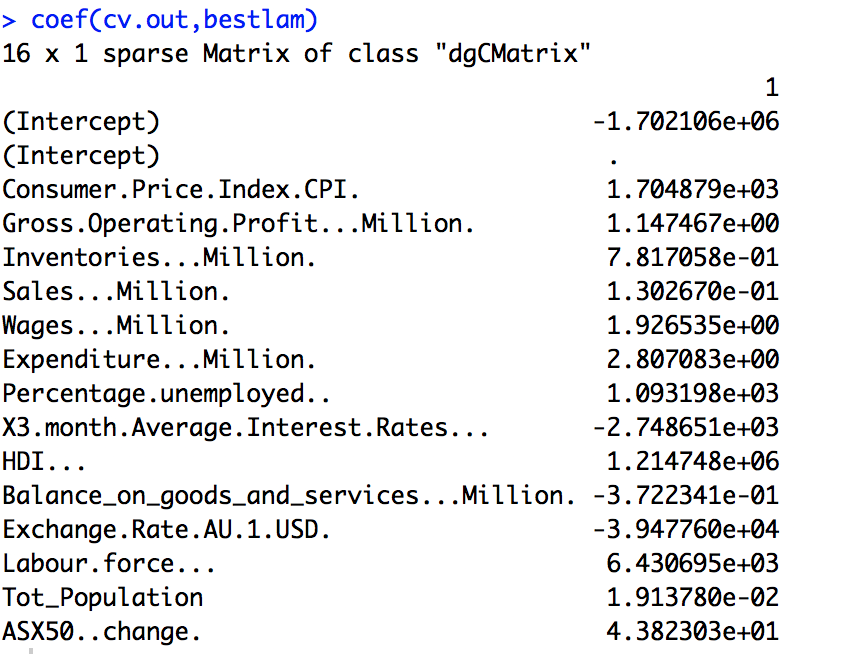
### Assumptions of ridge regression:

**Linearity –** There exists a linear relationship between X and Y.

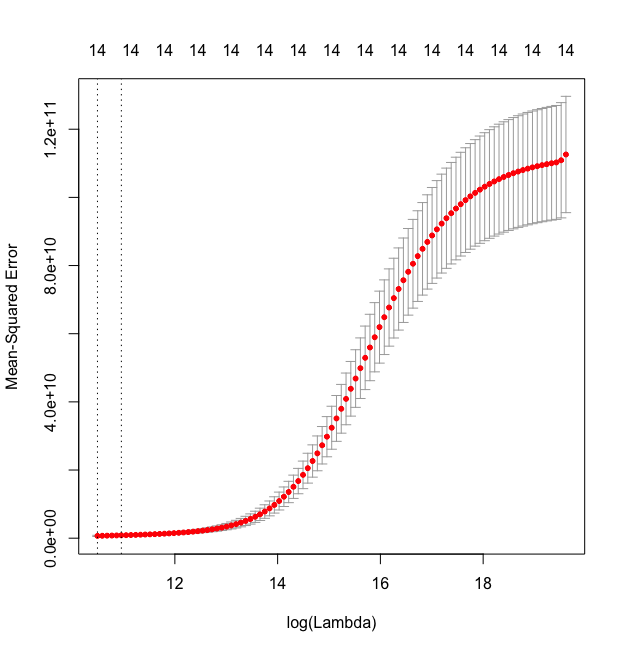
**Homoscedasticity-** The variance of the residuals is same for all values of X.

**Independence –**Residuals are independent to each other.

Ridge regression does not make assumption that the residuals are normally distributed. Various research shows that ridge is good at handling collinearity in data similar to Lasso (Confronting Multicollinearity In Ecological Multiple Regression, 2003). Subsequently data was modelled with ridge regression.



The plot below shows how ridge works. The X axis of the plot shows that the number of predictors has not reduced after training the model with the data.



The trained model was used to predict on the test dataset and MAE value calculated was **18240.82**. The error estimate from the ridge regression is a bit higher than measure for lasso regression.

# Feature extraction models

## Principal Component regression (PCR):

PCR is a feature extraction model and helps in reducing the feature space to principal components that are orthogonal to each other. PCR usually performs better in collinear setting of predictors and hence selected.

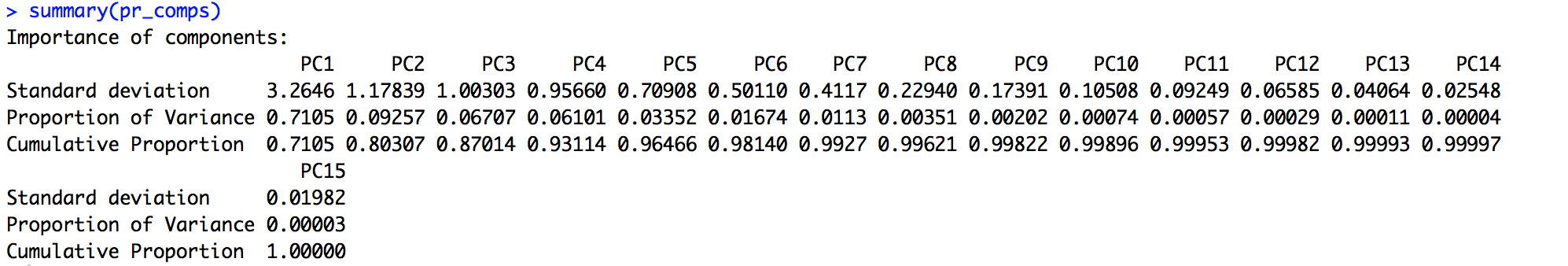
### Assumptions of PCR:

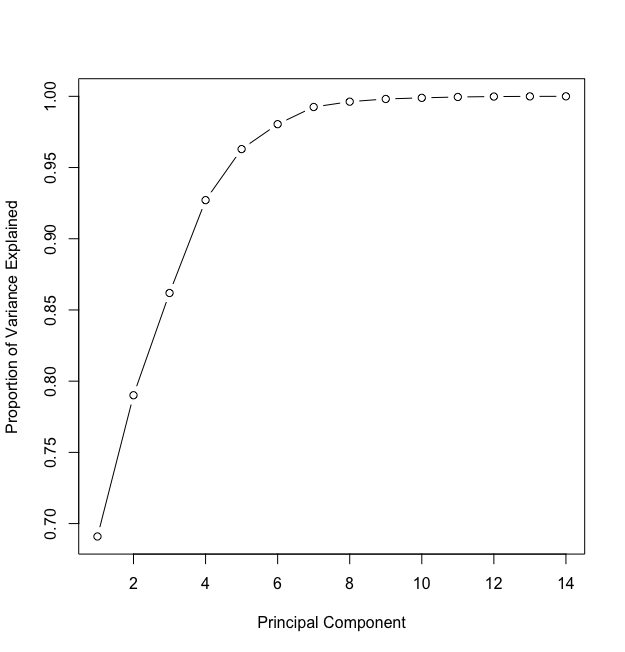
**Linearity –** There exists a linear relationship between X and Y.

**Homoscedasticity-** The variance of the residuals is same for all values of X.

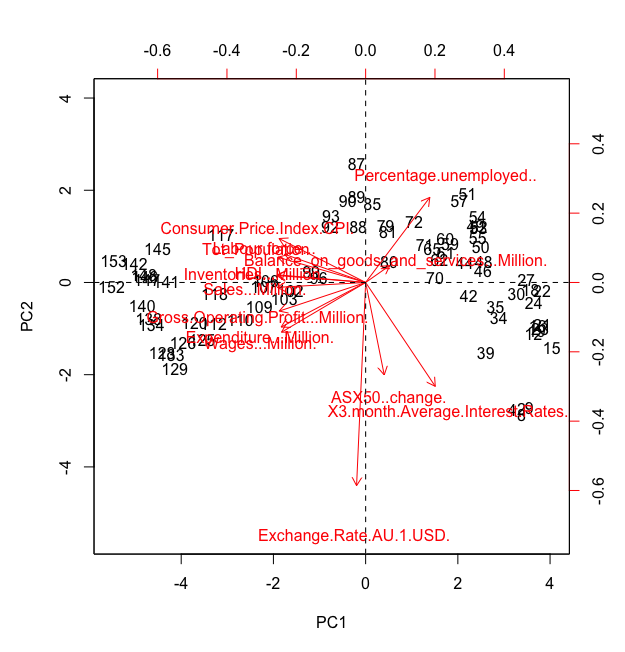
**Independence –**Residuals are independent to each other.

The predictors for principal components are picked based on the proportion of variance explained. Usually the first few principal components explain most of the variance in data.





In our dataset the first 6 principal components explained about 98% of variance in data. Each principal component takes data from different dimension and hence are not affected by multicollinearity.



The predictors taken into consideration for the first two principal components are shown in the bi-plot above. Predictors Inventories and Sales explain most of the variance for PC1. Percentage unemployed and Exchange rates explain the most of variance in PC2.

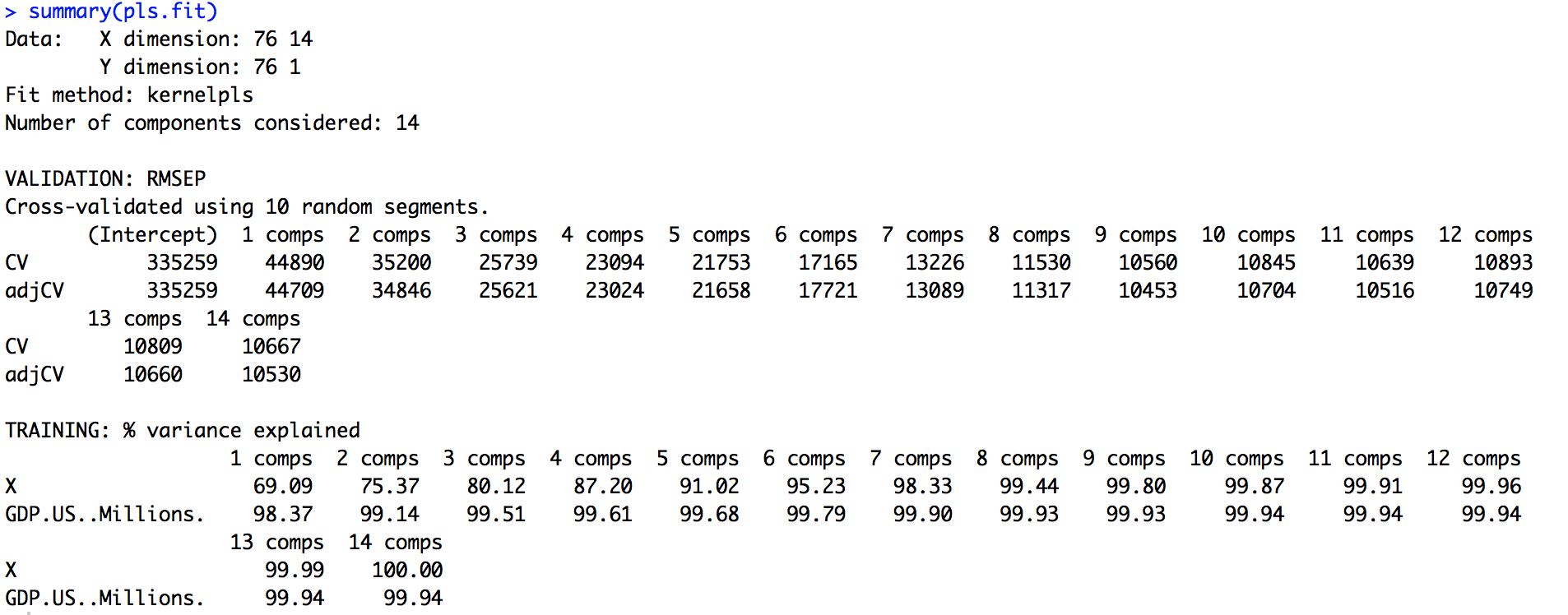
A prediction was done with the first 6 principal components and a MAE value of **62064.09**

was calculated which is very much higher than regularization techniques.

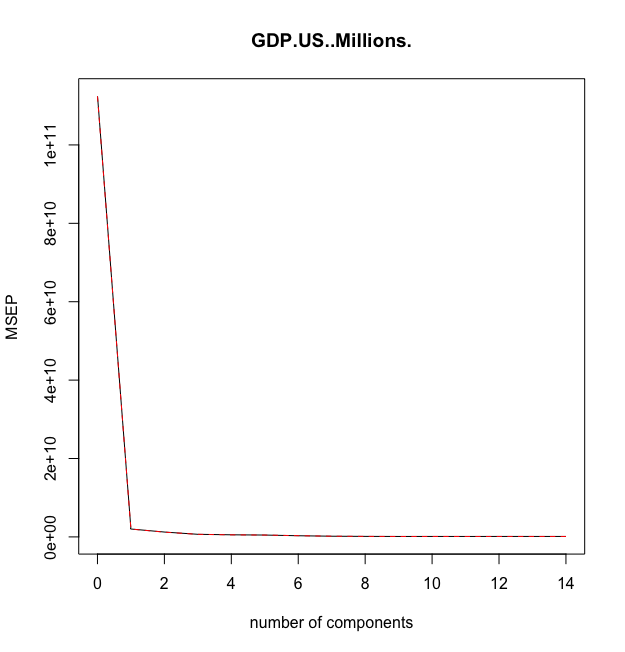
## Partial least squares (PLS):

PLS is another type of feature extraction technique. The main difference between PCR and PLS is PLS considers the dependent variable and hence would be able to arrive at a higher proportion of variance explained within a few components. The assumptions of PCR apply to PLS as well.

PLS was run on the GDP dataset and the following components were returned. As stated 99.5% of variance is explained within the first 3 components. The cross validated error estimates slowly reduce as more components are added.



The plot shows reduction in the Mean Squared Error(MSE) as PLS components increase. A very big drop in the MSE is observed when the first component is added. The first 4 components were chosen and used for modelling.



MAE estimate for PLS is calculated to be **21418.03.** It is lesser than the value calculated by PCR but still high compared to regularization techniques.

# Australian GDP Performance

Available data from Organization for Economic Co-operation and Development (OECD) allows for a comparative analysis on how the Australian quarterly GDP performs over the years. Australia holds the current record worldwide as the economy with the longest economic growth among developed countries.

How GDP growth rate is calculated?

GDP is calculated first on a quarterly basis with the following formula:

Quarter-on-quarter GDP Growth Rate = (Current Qtr GDP – Previous Qtr GDP)/Previous Qtr GDP

The resulting growth rate for the quarter is then annualized using the following formula:

Growth Rate = [ (1 + Quarter-on-quarter GDP Growth Rate)4 – 1 ] \* 100

The resultant Growth Rate is rounded 1 decimal number. GDP Growth rate is a quick indicator on how the country’s economic performance compares with the previous quarter and with the same quarter the previous year.

## Calculating GDP

There are several so-called approaches in calculating GDP. These are the Output Approach, Expenditure Approach, and the Income Approach. Output Approach is characterized by computing the GDP using the total value of goods and services produced by the country. The Expenditure Approach uses government spending, consumption, net exports and investments whilst Income Approach measures the GDP by the total income generated in producing goods and services.

Australian GDP is calculated as:

GDP = C (Private Consumption) + I (Private Investment) + G (Government Expenditure) + (EX – IM)

where EX stands for export of goods and services and IM stands for import of goods and services.

## Australian GDP Data Exploration

For our group, the Skeptics, we decided to use the Expenditure Approach as the GDP method by which we work on this paper. We sourced our data from OECD. Annual and quarterly GDP data from 1960 for all the 3 GDP approaches are available in OECD. For the purpose of comparing Australian GDP vs the top 10 OECD member economies, we used the GDP Expenditure Approach data. GDP data of Canada, Spain, UK, Ireland, Japan, South Korea, The Netherlands, France, and USA were chosen for comparison with Australia. It is worth mentioning that on this research, we are using the Expenditure Approach data only, which may not show exactly the same figures or results as that of the other approaches in terms of GDP behaviors or performance measurement. Secondly, that continuous periods of positive or negative growth vary when using one over the other approaches.

Counting all the quarter periods with uninterrupted positive growth for the 10 countries from 1960, Table 1 shows that Australia emerged as the country with the most positive growth periods.

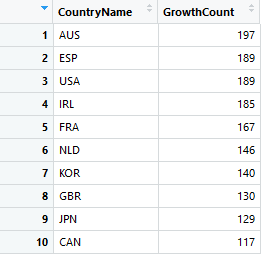
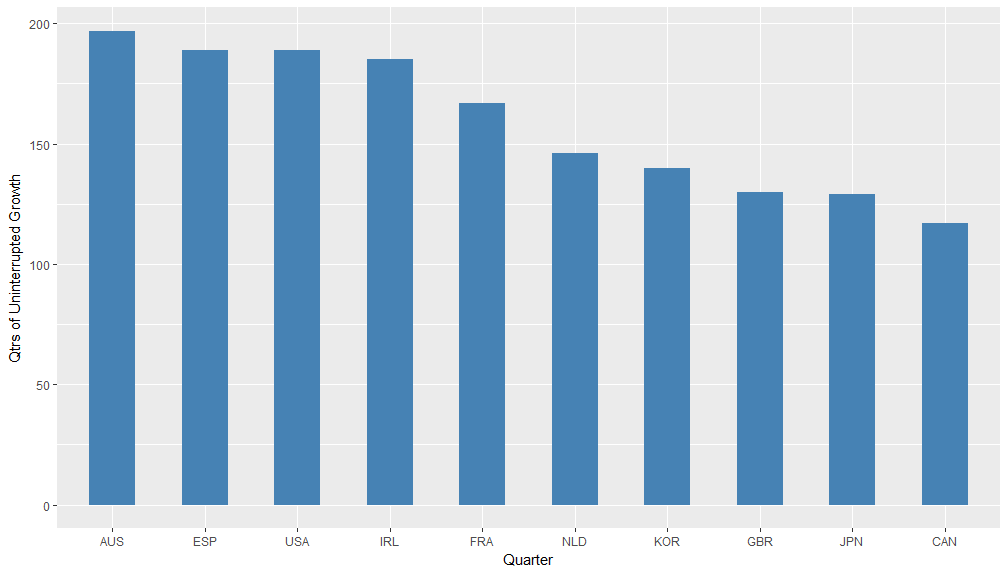


Table 1: Continuous Growth Periods (Data Source: OECD)

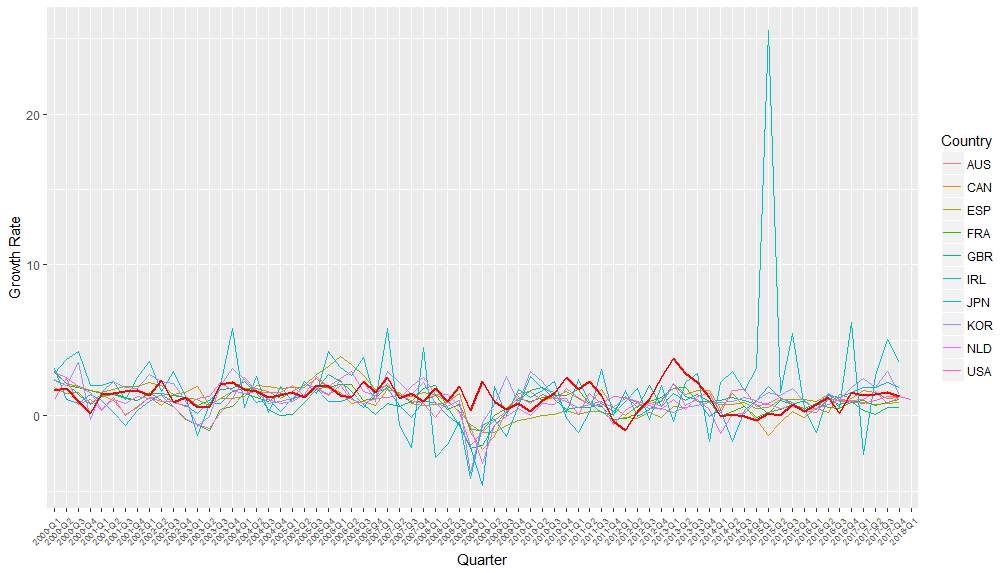
Graph 1 below shows the bar chart on the most number of consecutive periods or quarters the countries enjoyed uninterrupted GDP growth.



Graph 1: Periods of uninterrupted growth of 10 countries (Data Source: OECD)

## Australian GDP Performance

Graph 2 shows the GDP growth performance of 10 countries including that of Australia (highlighted in red). The Australian economy has shown to be resilient during the past two decades. Australia was able to weather the many global economic crisis including the Asian crisis of 1997-1998. Gone through the US stock market crash and recession of 2001 and performed well above others during the Global Financial crisis (GFC) of 2009. Australia’s impressive economic performance in the past two decades has been a result of recent economic reforms, booming mining sector, and a strong population growth rate.



Graph 2: GDP growth of 10 countries from 2000 - 2017 (Data Source: OECD)

# Reference List

Australian Bureau of Statistics 2012, *Defining and measuring GDP*, viewed 28 April 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1301.0~2012~Main%20Features~Defining%20and%20measuring%20GDP~221>.

##### Bagshaw, E. & Massola, J. 2017, *GDP: Australia grabs record for longest time without a recession*, SMH, viewed 15 May 2018, < https://www.smh.com.au/business/the-economy/gdp-australia-grabs-record-for-longest-time-without-a-recession-20170607-gwm0o2.html >.

*Confronting Multicollinearity In Ecological Multiple Regression - Graham - 2003 - Ecology - Wiley Online Library* 2003,viewed 28 May 2018, <<https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/02-3114>>.

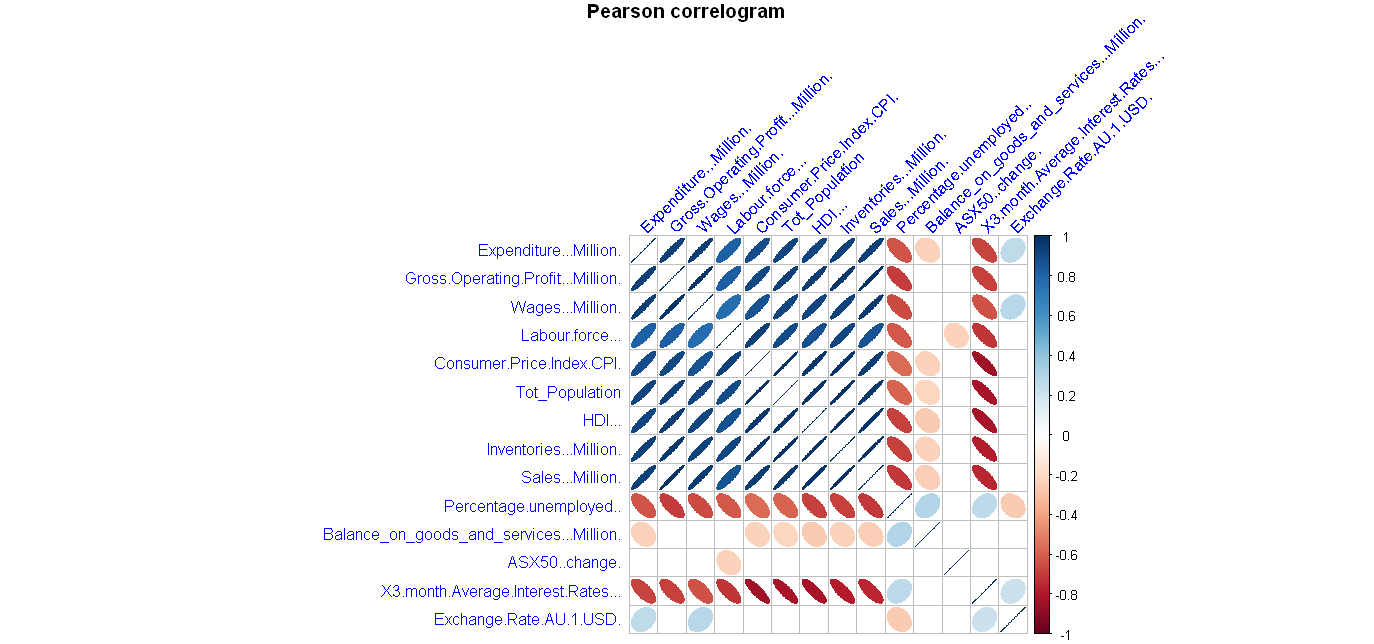
Investopedia 2018, *What is GDP and why is it so important to economists and investors?*, viewed 28 April 2018, < https://www.investopedia.com/ask/answers/199.asp>.

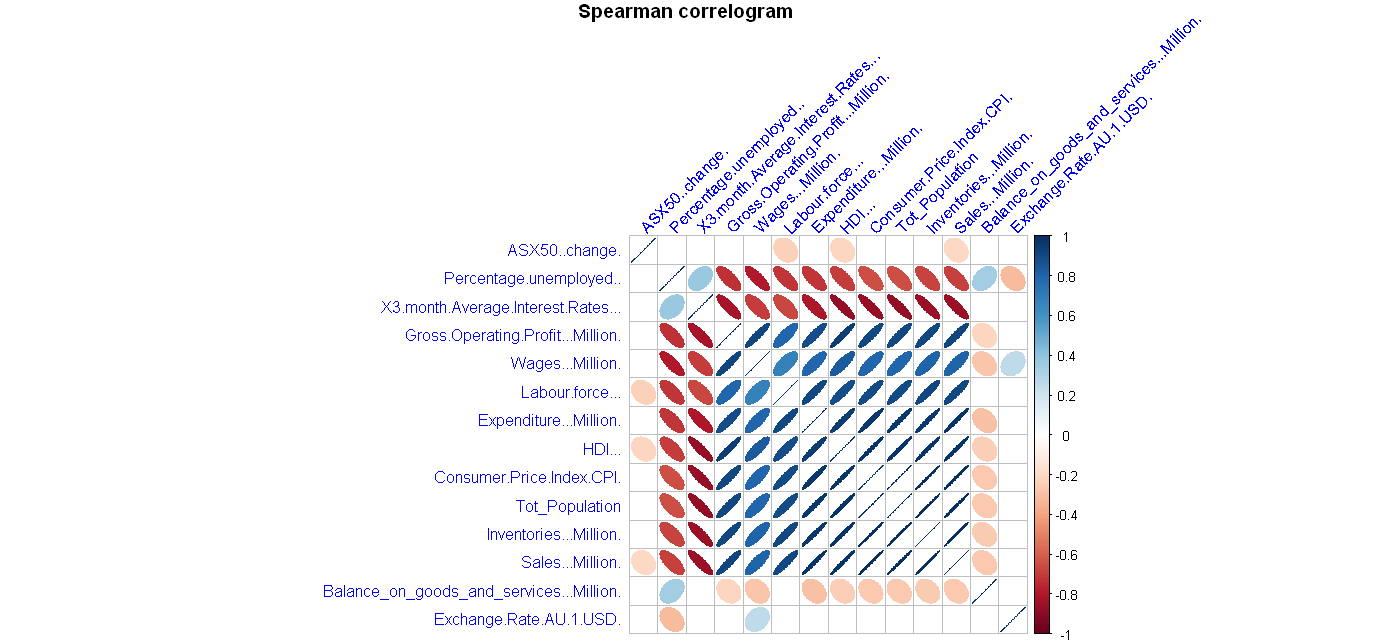
Wikipedia. 2018, *Training, test, and validation sets*, San Francisco, viewed 10 May 2018, < https://en.wikipedia.org/wiki/Training,\_test,\_and\_validation\_sets >.

# Appendix

## Further methods used for displaying variable correlation matrix and use of heatmaps

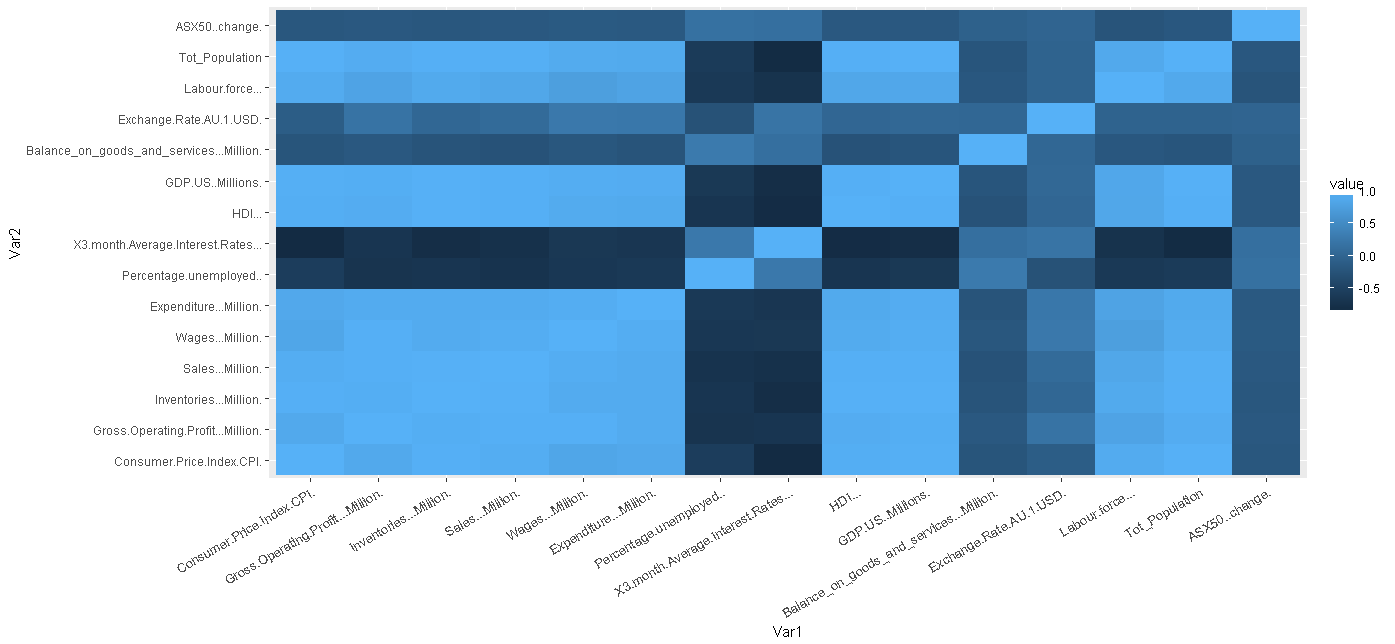
Alternative was to display the correlation matrix was explored, by separating the GDP (dependent variable) and the independent variables and running Pearson and Spearman correlation matrix visualisations:



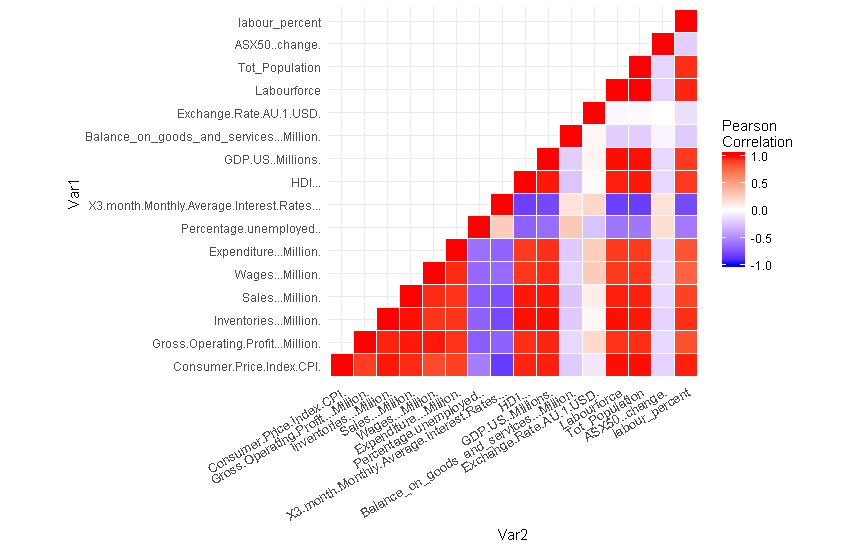


Positive correlations are displayed in blue and negative correlations in red colour. Colour intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend colour shows the correlation coefficients and the corresponding colours.

And by using additional heatmap functions within ggplot2, we are able to display the correlations as a heatmap.



Trying out different techniques for displaying correlation led to using other techniques such as applying ‘melt’ to correlation matrix and eliminating redundant information



* the figure above shows negative correlations in blue colour and positive correlations in red.

Finalizing the experimentation of heatmap techniques, the graph produced above has been re-ordered based on correlation values and the correlation coefficient has been added to heatmap producing a better looking display:

