# Exploratory Data Analysis

## Initial Phase

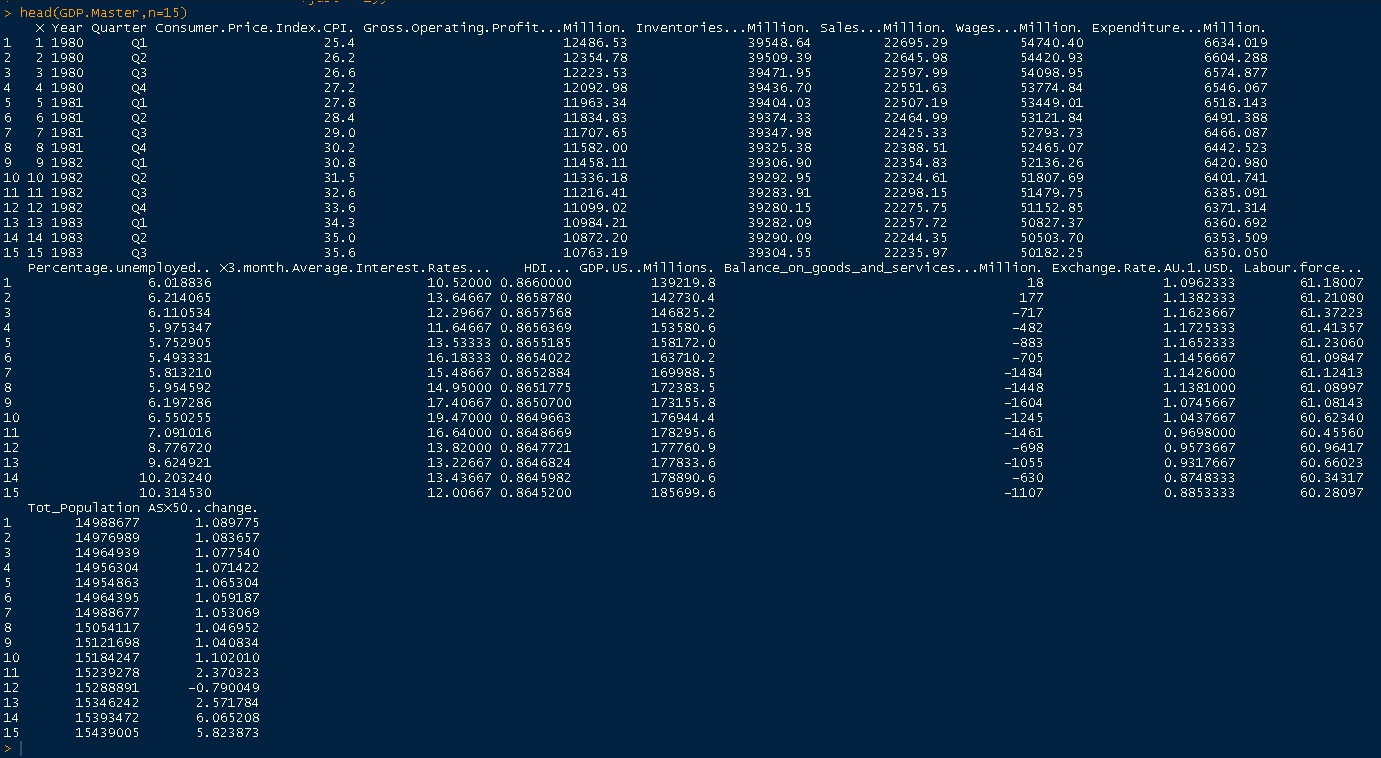
Australian Gross Domestic Product (GDP) is a complex econometric calculation and the starting limitation of this project was not to replicate the real model for calculating GDP but instead, selecting one approach for its measure called the ‘Expenditure approach’ and from this approach, select very few (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:

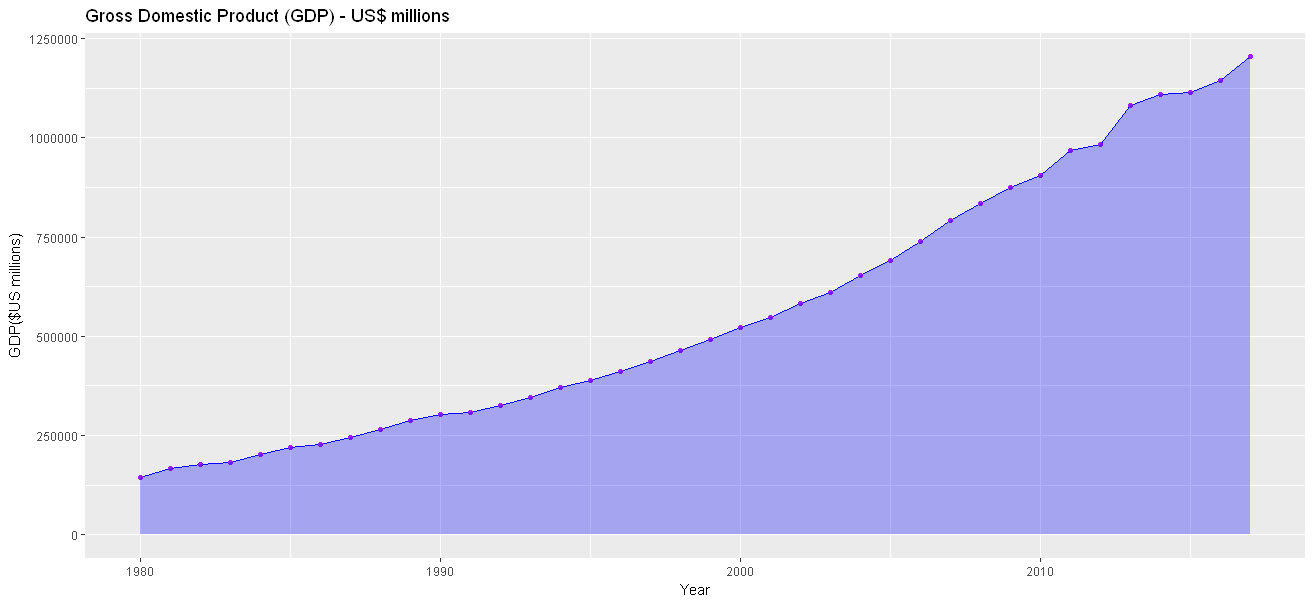
|  |
| --- |
| * Consumer Price Index (CPI) |
| * Gross Operating Profit ($ Million) |
| * Inventories ($ Million) |
| * Sales ($ Million) |
| * Wages ($ Million) |
| * Expenditure ($ Million) |
| * Unemployment (Percentage unemployed %) |
| * Interest Rates (3-month Average %) |
| * Human Development 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.

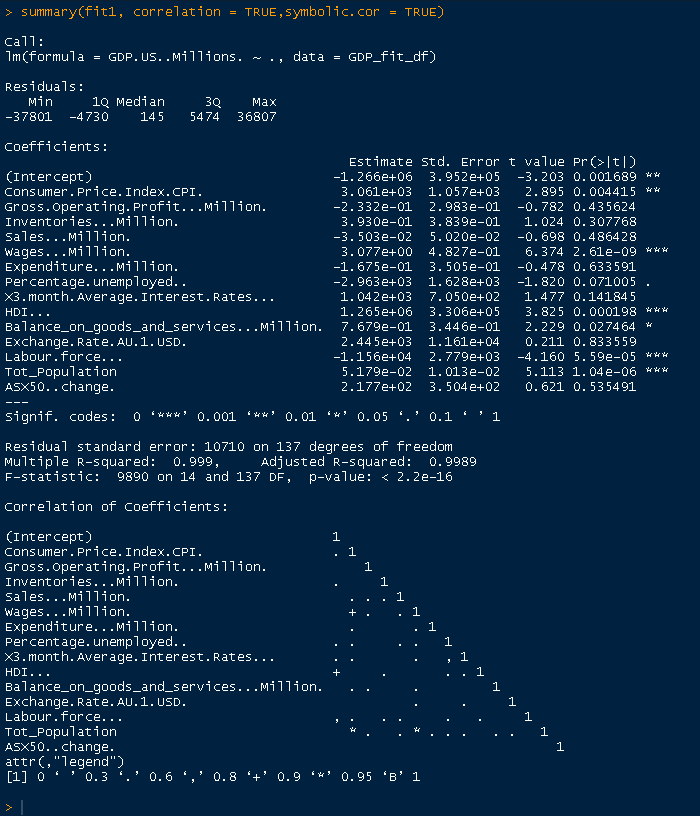


Initially some plots were produced to understand the data a little more. In the graph below the growth (amount) of GDP over the years has been plotted and as observed it has been steadily growing with no noticeable decline.



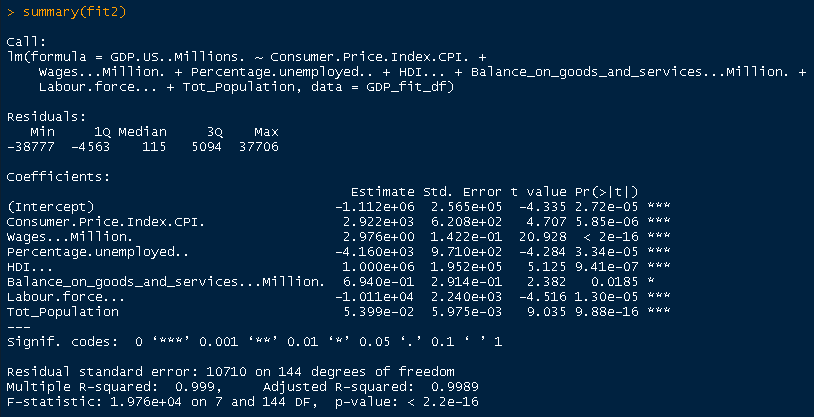
From the plot above, we can observe the growth (amount) of our dependent variable: GDP looks continuous in nature which lead us to the assumption of the relationship between the growth of our dependent variable and our independent indicators may be linear in nature. Since we have a few independent variables, 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:



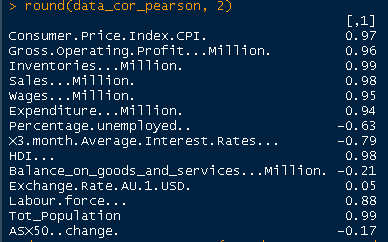
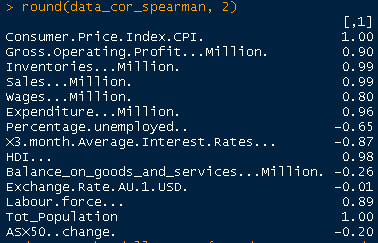
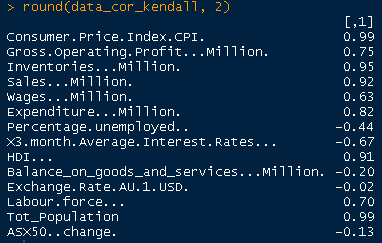
From the summary, the coefficient significance and adjusted R-squared are may be indicative there are too many variables for a proper fit of the linear regression model.

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:

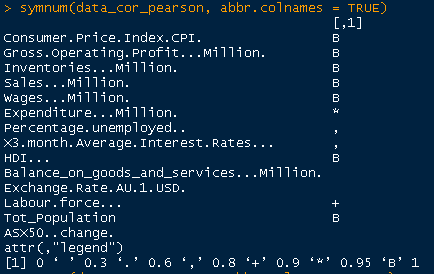
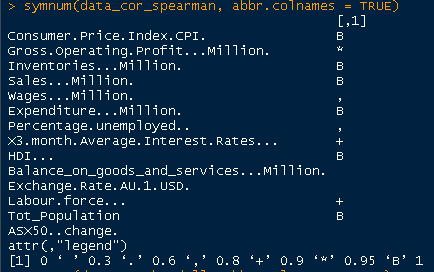
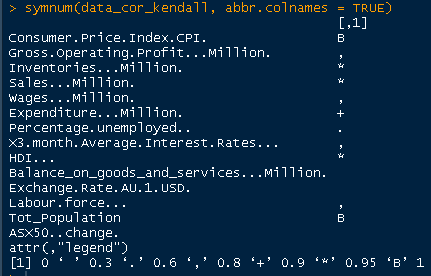


## A deeper look into independent variable correlations

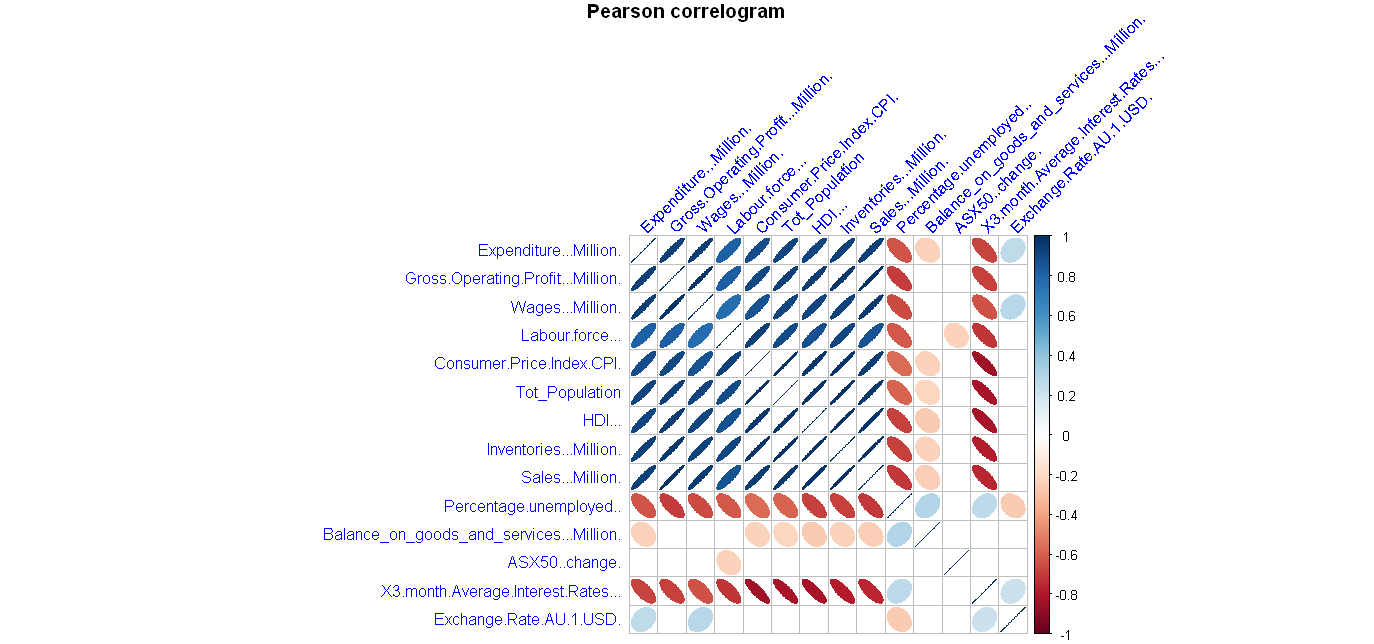
To further explore correlations of the independent variables , we separated the dependent variable: GDP (GDP.US..Millions.) and calculated correlations using y=GDP, x = the rest of variables. Three methods of correlation were run against the data:

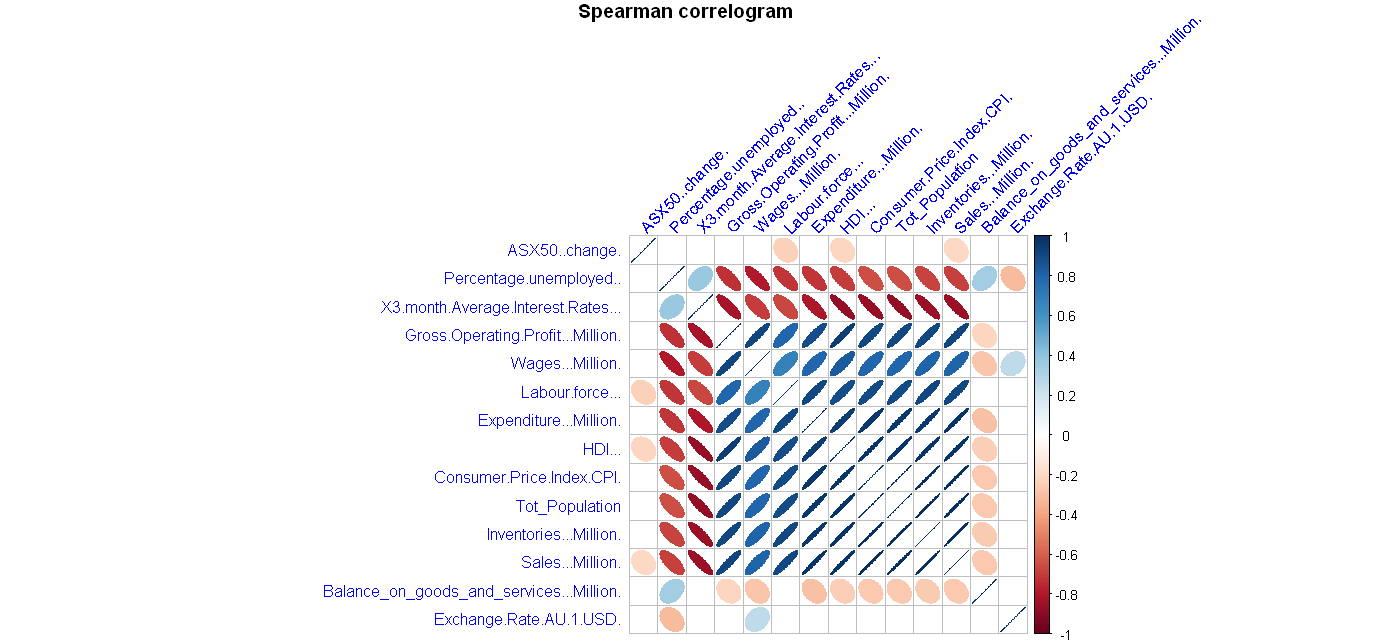
  

replacing correlation coefficients by symbols depending on level of correlation:

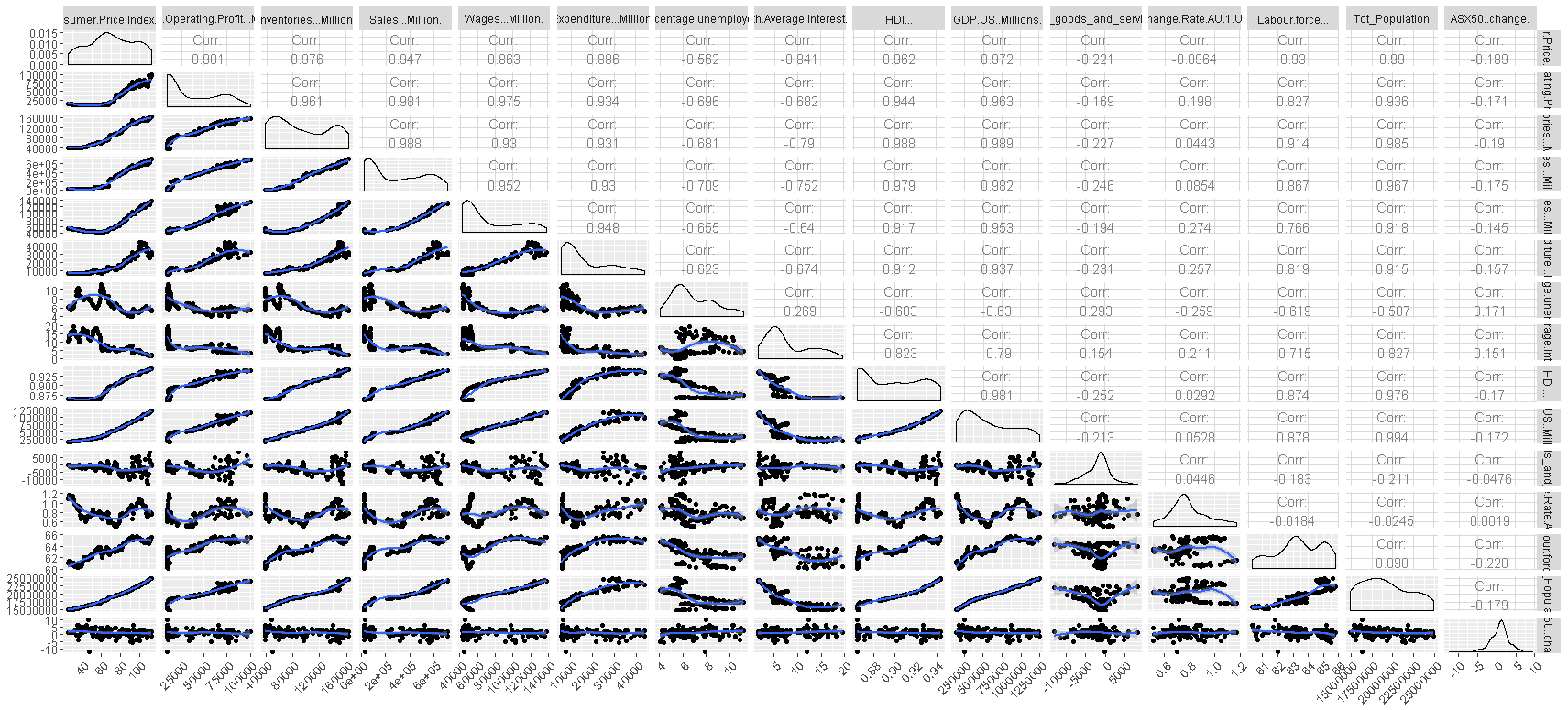
Visualising the correlation matrix, where positive correlations are displayed in blue and negative correlations in red colour:





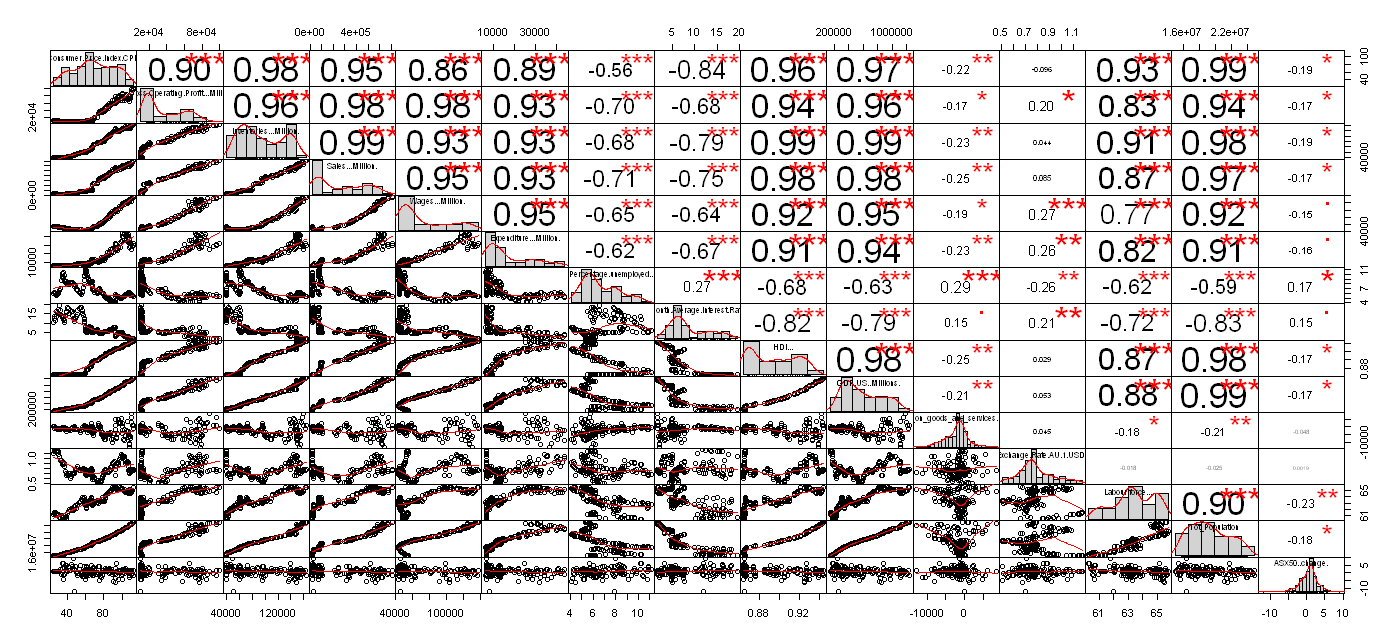
## Further Exploration

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

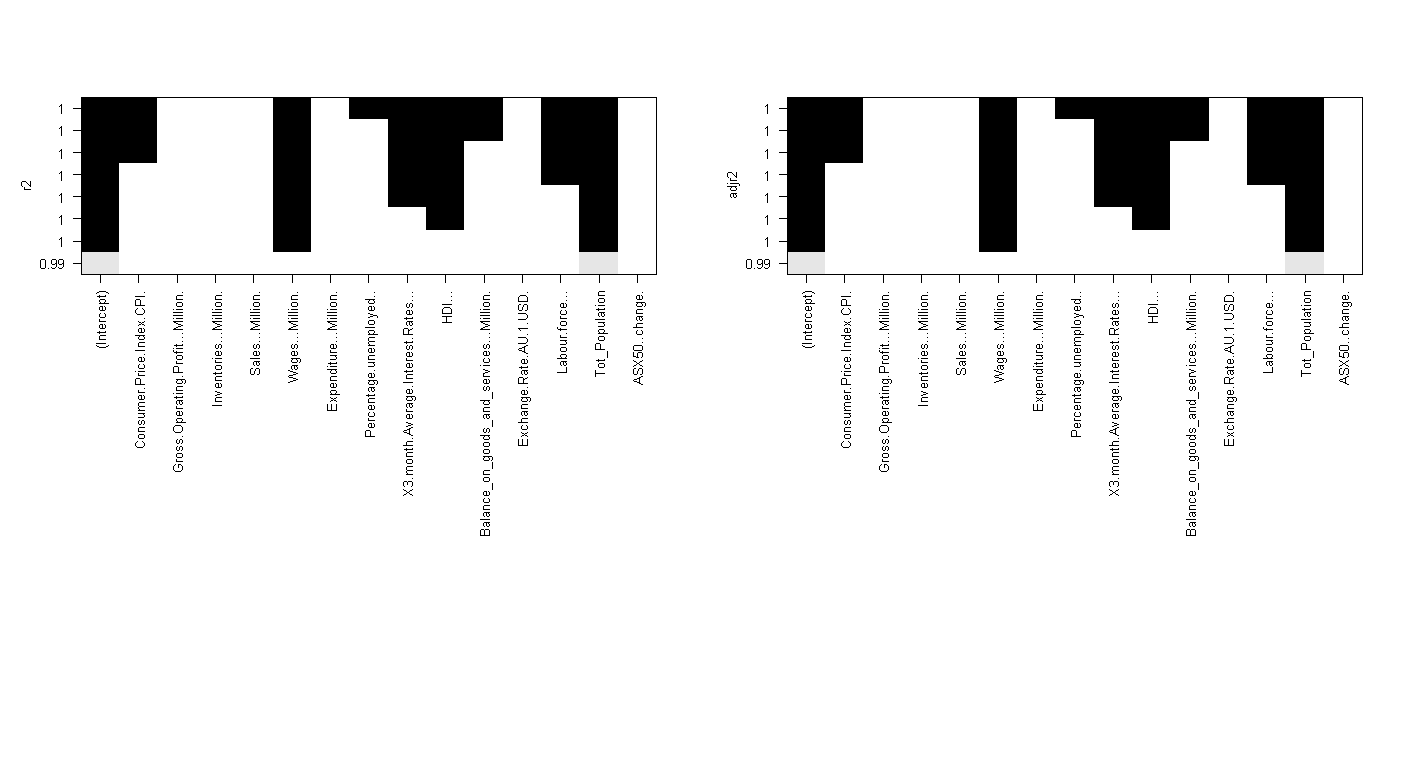


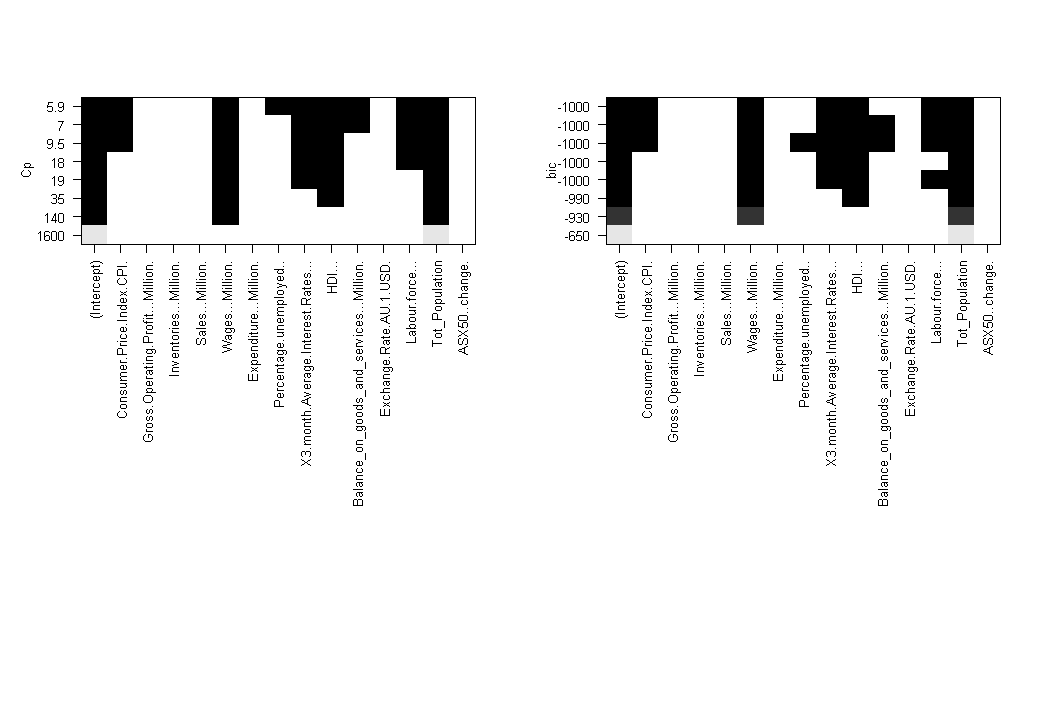
A correlation matrix chart produced by using the function: **chart.Correlation** from the library ("PerformanceAnalytics") gave us a similar display where the distribution of each variable is shown on the diagonal.

* 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(“\*\*\*”, “\*\*”, “\*”, “.”, " “)



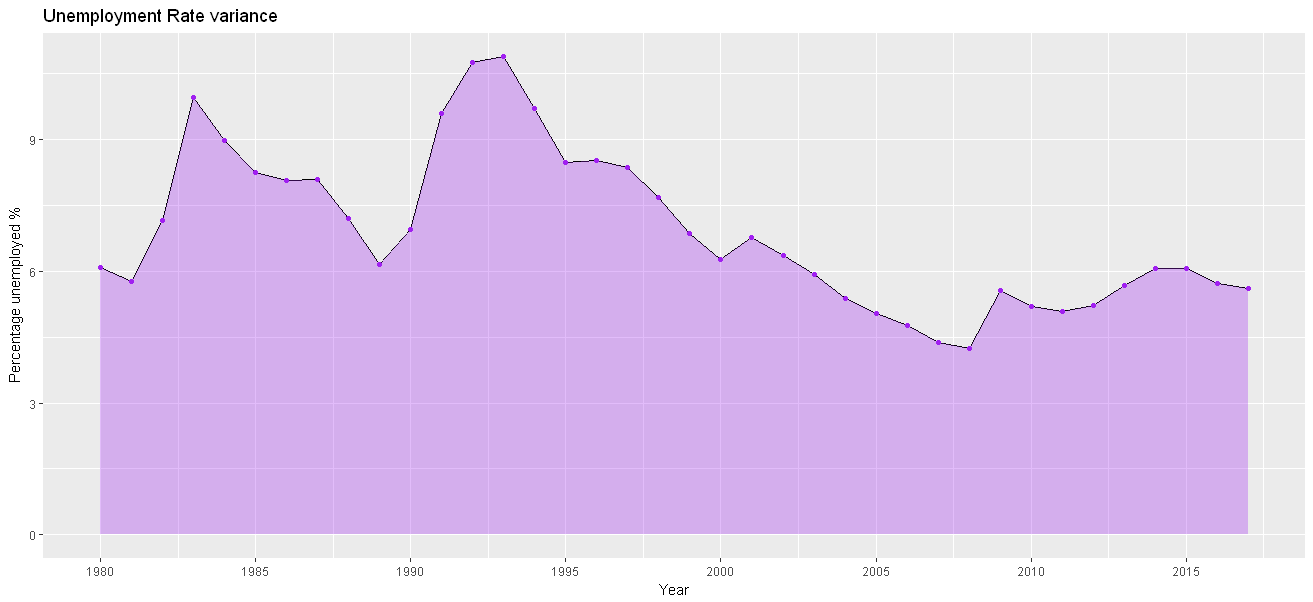
While exploring for subset selection, we tried using function **regsubset** from **library(leaps)** to explore variable selections using “exhaustive” search method. The results were unsurprisingly similar to the significant coefficient from our initial model.





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.

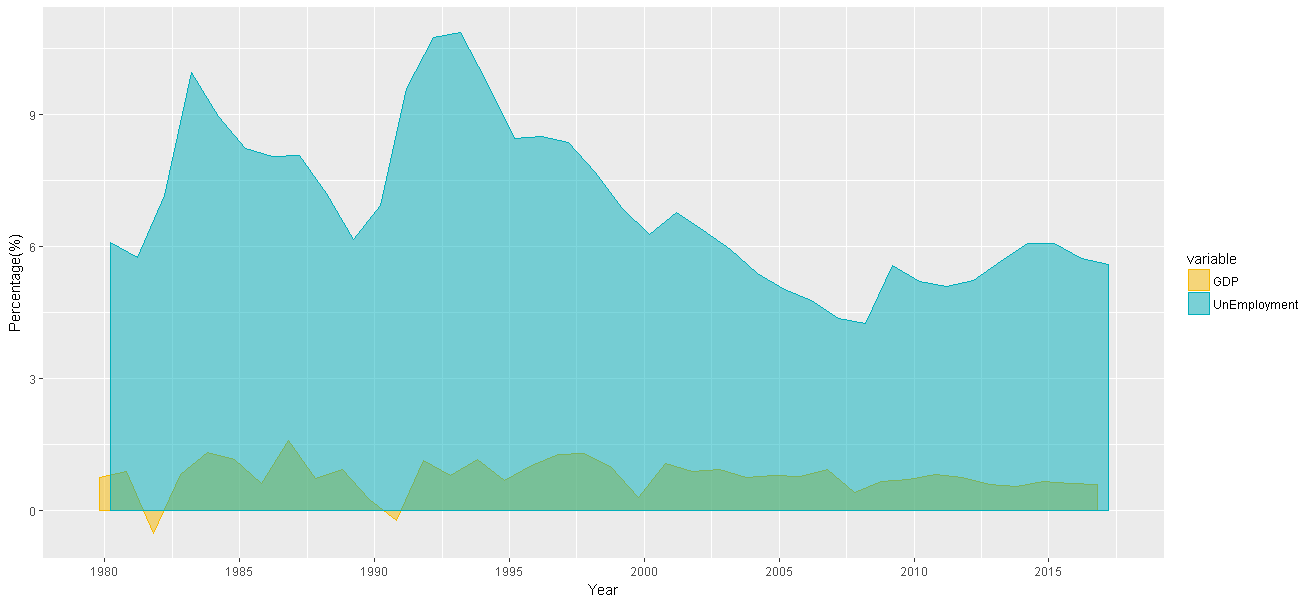
Within our master dataset we brought in the unemployment rate figures and a simple graph shows how this has changed over the years



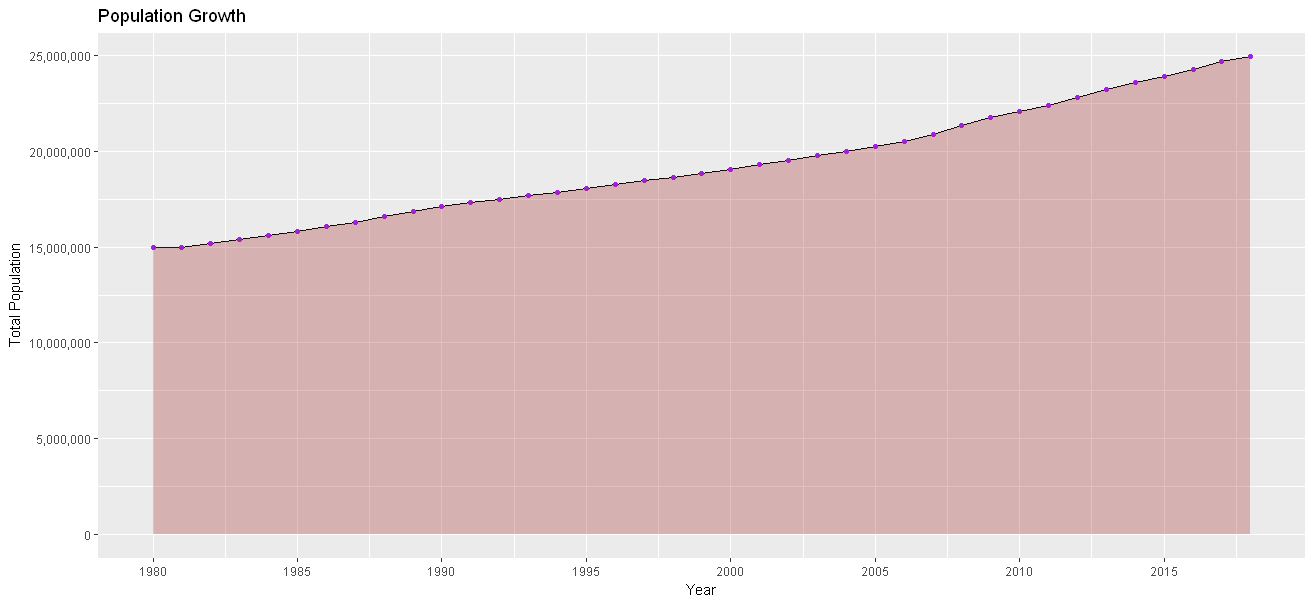
For the purpose of correlating the unemployment rate with GDP variance, a new dataset of quarterly GDP variation percentage was brought into our set. This was used to create a new data frame of aggregated yearly variations as shown below:



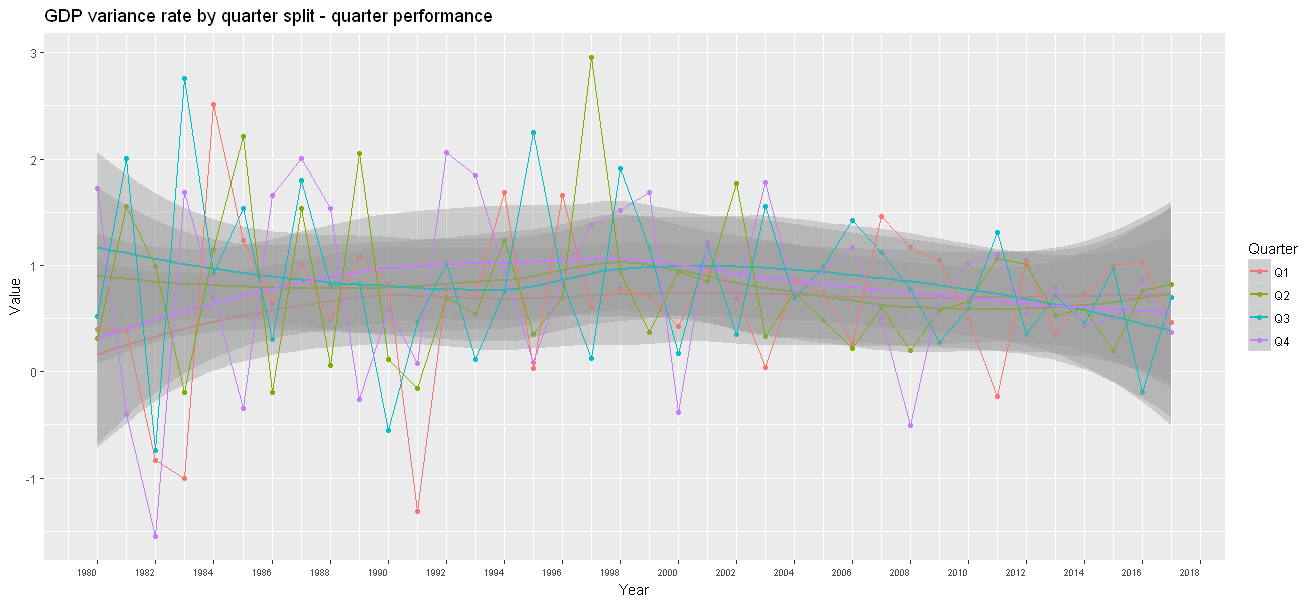
With this data, we were able to graph the relation between GDP change vs un-employment over the years as pictured here were we can observe there is a similar pattern in the yearly changes:



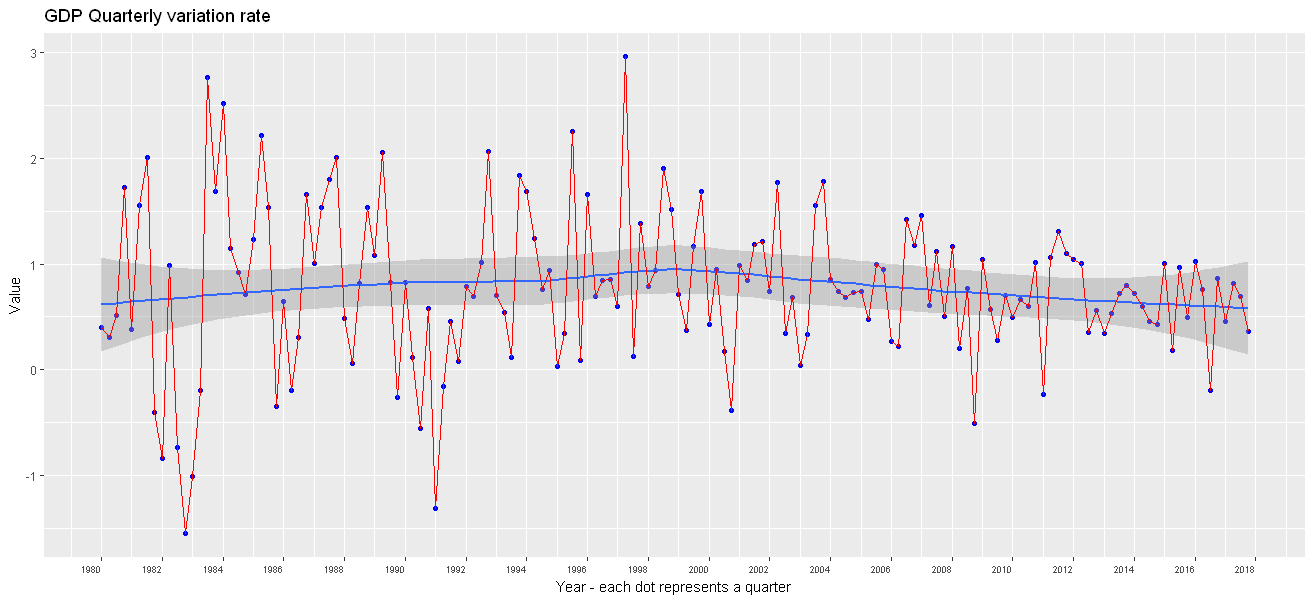
Similarly, a simple plot can provide a nice picture on how steady population in Australia has been growing over the past years. There may be a direct relationship between our growing population and higher growth of GDP. As identified through our initial model, there is a correlation between population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

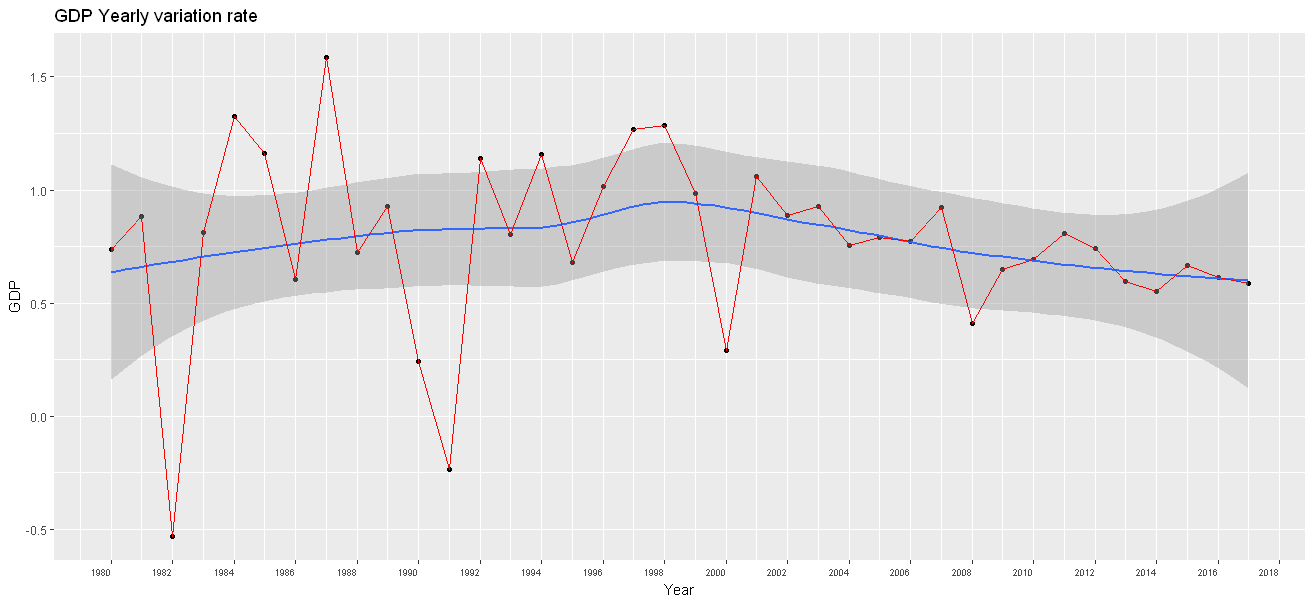


Looking more into the GDP data, using a split by quarterly data we were able to display the quarterly performance of GDP per year since 1980. This was mainly done to explore for any concealed patterns that may be of interest.



And displaying the quarterly and yearly variation average for Australian GDP:

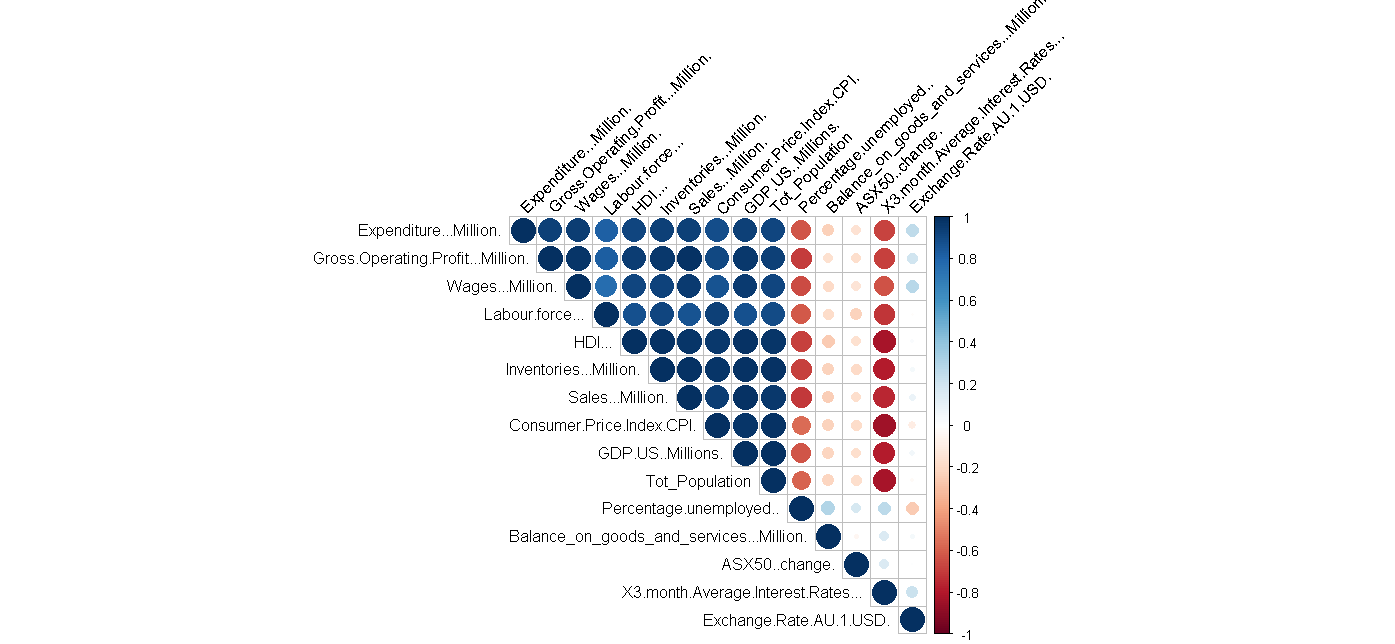




# Appendix:

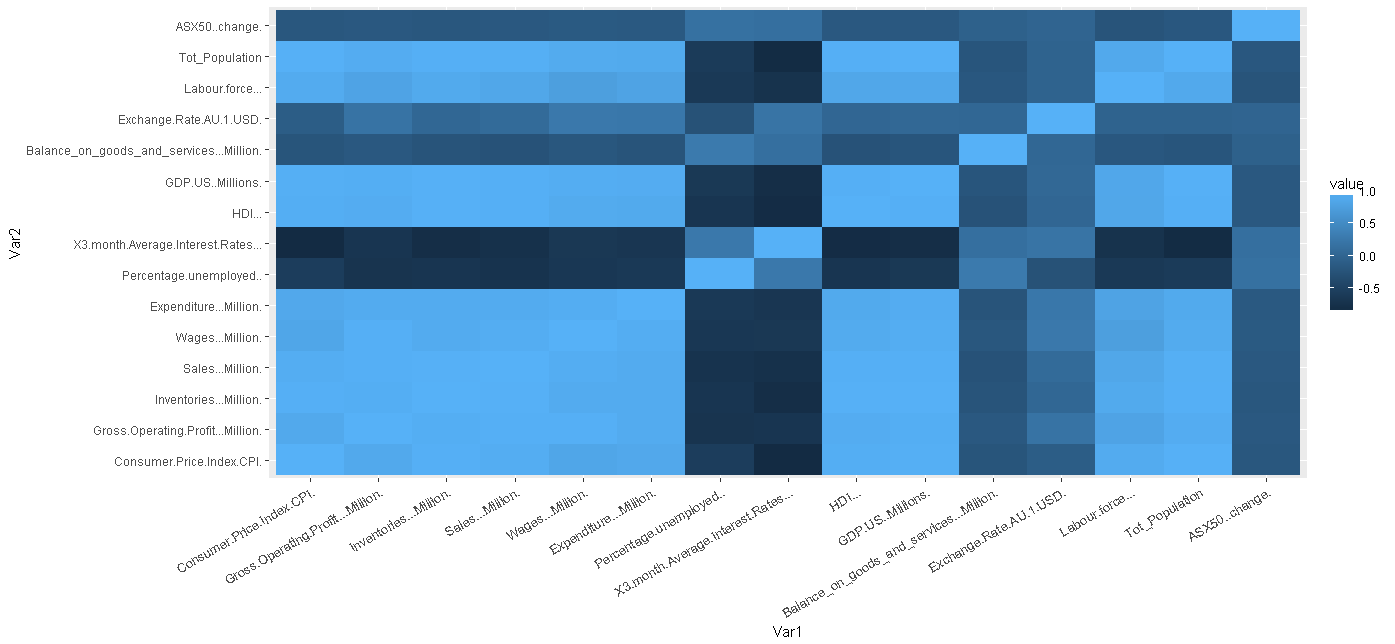
## Further methods used for displaying variable correlation using heatmaps

An alternative display of the correlation matrix, using all dataset variables without separation of GDP and the independent variables allows for the full correlation matrix to be visualised:

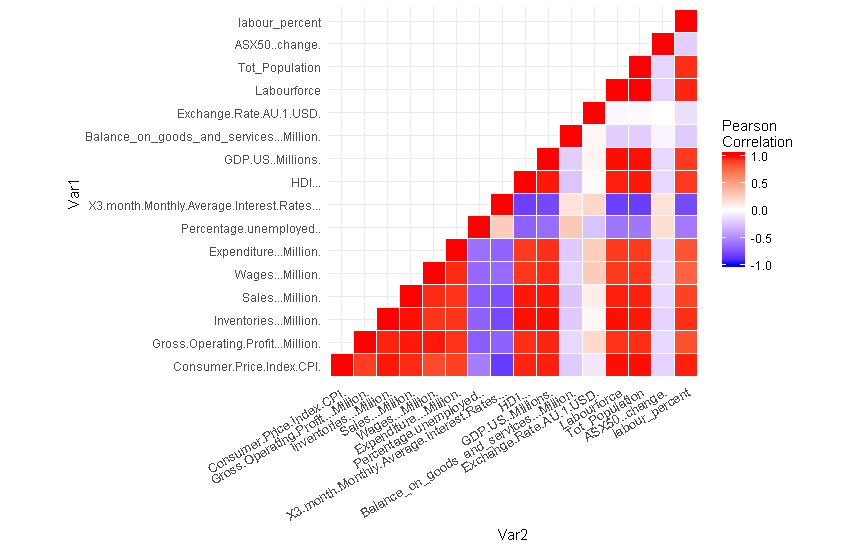


Positive correlations are displayed in blue and negative correlations in red colour. Colour intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend colour shows the correlation coefficients and the corresponding colours.

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:

