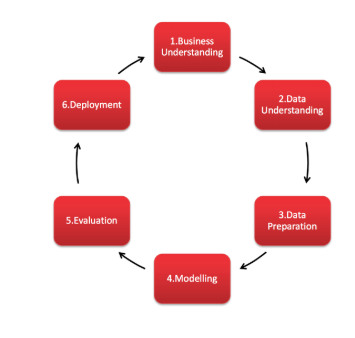
# Data collection and wrangling

## 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**

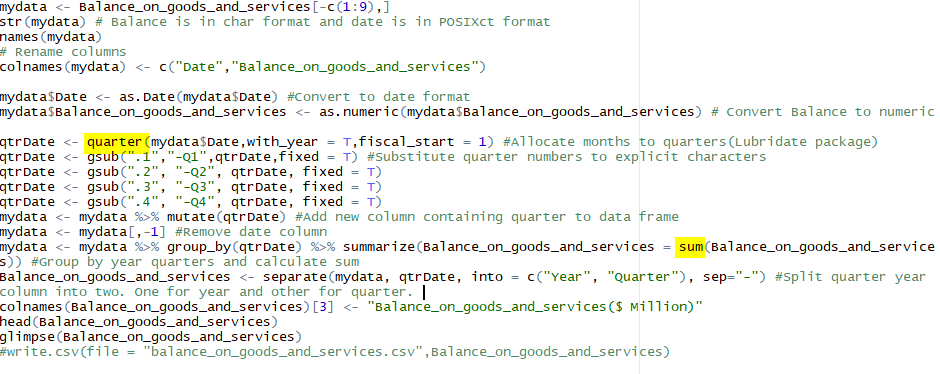
## Data Wrangling

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.

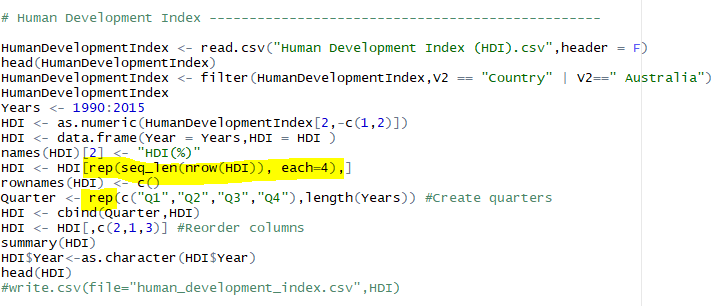


Data for model one is clean and tidy but data for model two is collected from different sources and the variables are of different time periods and intervals. For example, GDP data is available from 1960 and the interval is quarterly data whereas Stock return aggregated value of top 50 companies is available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

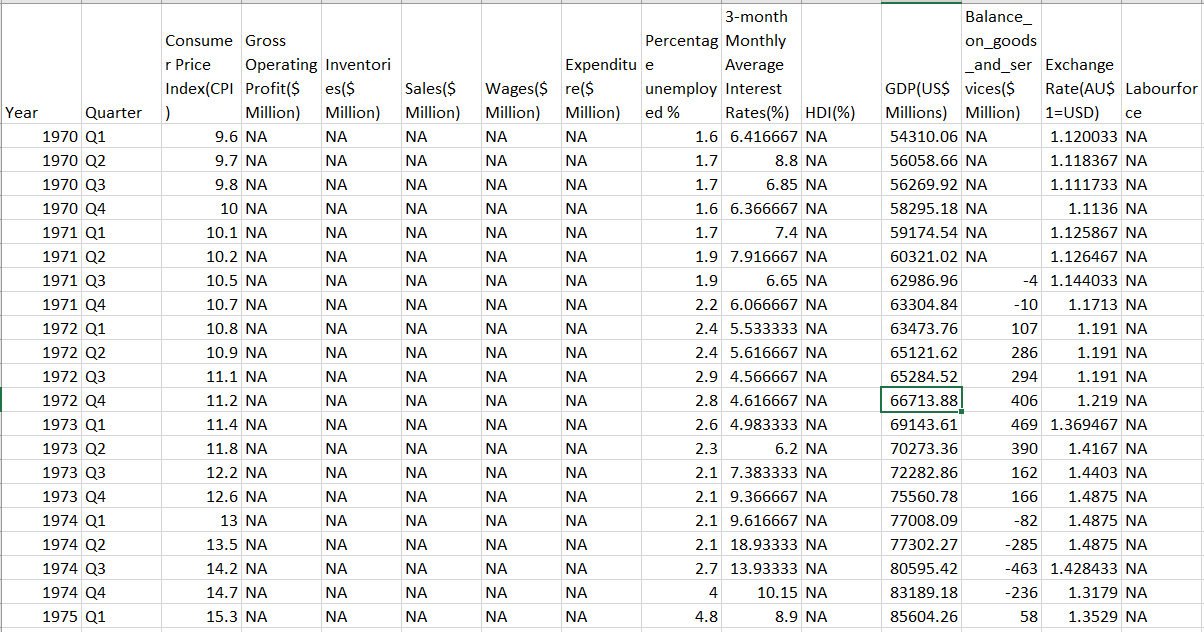
The monthly data of variables are aggregated into quarters and the values for the observations are calculated depending on the type of variable. For example, Stock return aggregated value of top 50 companies are grouped into quarters and value for the variable is calculated as mean stock return value 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.

