|  |  |
| --- | --- |
| Description: 18897 CIC A4 Portrait WordTemp_cropped.jpg | **ASSIGNMENT COVER SHEET** |
|  | |
|  | |
| |  |  | | --- | --- | | **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** | 30 April 2018 | | **ASSESSMENT ITEM NUMBER/TITLE** | AT2 Data analysis project, Part A: Project Plan | | * 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:**  30 / 04 / 2018  If submitted electronically tick here to indicate you agree with the above X | | | |

**Rationale for project**

Gross Domestic Product (GDP) represents the total dollar value of all goods and services a country produced over a specific time period, often referred to as the size of the country’s economy. As one of the most widely used economic indicators, GDP is used to gauge the health of a country’s economy (Investopedia 2018). Given the importance of having a healthy economy to the wellbeing of a country’s citizens, our team considered whether it was possible to predict future GDP of Australia using other historical economic and non-economic information (further discussed in datasets section below).

Our team viewed choosing this topic as a learning opportunity, to better understand how we as individuals and a as a community contribute to the economy. As a team of data analysts without formal qualifications in economics, it will decode economic jargon and provide us with insights on what are the important factors to look at when trying to analyse 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) solely focus on predicting nominal GDP (a country’s economic output without an inflation adjustment).

The main research question our team came up with was “can nominal GDP be accurately predicted from other economic and non-economic information”.

The secondary research question supplementing this was which economic or non-economic factors are most important for influencing nominal GDP.

**Range of datasets examined and chosen for analysis**

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.

Whilst measuring GDP can be complicated, at its most basic mathematically, the calculation can be done in one of three ways: either by adding up what everyone earned in a year (income approach) or by adding up what everyone spent (expenditure method), or by how much everyone produced (production approach). While each approach should, conceptually, deliver the same estimate of GDP, if the three measures are compiled independently using different data sources, then different estimates of GDP will result (ABS 2012). To combat this issue, the estimates in the GDP data sets used had been pre-balanced by the ABS between the three approaches.

As it did not matter then which method we choose as long as we were consistent in our logic, we choose the expenditure method as it had the most readily available information for calculating GDP = Consumption + Investment + Government spending + Net Exports.

In doing so, when choosing our data sets, we were also careful to not choose datasets that were components of each of the methods but rather indicators for the components. For example, we used Consumer Price Index and Retail Trade Turnover as an indicator of the level of Consumption in the economy.

|  |  |  |
| --- | --- | --- |
| Expenditure Approach | Indicator | Link |
| GDP | GDP | http://www.abs.gov.au/ausstats/abs@.nsf/mf/5206.0 |
| Consumption | Consumer Price Index | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0 |
|  | Retail Trade Turnover | http://www.abs.gov.au/ausstats/abs@.nsf/mf/8501.0 |
| Investment | Bank Accepted Bills/Negotiable Certificates of Deposit-3 months | https://www.rba.gov.au/statistics/tables/ |
|  | Employed total persons | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 |
| Government | Unemployed total persons | http://www.abs.gov.au/ausstats/abs@.nsf/mf/6202.0 |
|  | Business Indicators | http://www.abs.gov.au/ausstats/abs@.nsf/mf/5676.0 |
| Net Exports | Trade Weighted Index | https://www.rba.gov.au/statistics/tables/ |
|  | Balance on Goods and Services | http://www.abs.gov.au/ausstats/abs@.nsf/mf/5368.0 |

Additional datasets for analysis:

* Gross Domestic Product – data sourced from The Organisation for Economic Co-operation and Development (OECD) API services: <http://stats.oecd.org/restsdmx/sdmx.ashx/GetData/QNA/AUS.B1_GE.CPCARSA.Q/all?startTime=1960-Q1&endTime=2018-Q1>
* Unemployment – data sourced from The Organisation for Economic Co-operation and Development (OECD): <https://data.oecd.org/unemp/unemployment-rate.htm>
* Human Development Index – data sourced from the United Nations Development Programme (UNDP): <http://hdr.undp.org/en/data>
* Interest rates – data sourced from the Reserve Bank of Australia (RBA): <https://www.rba.gov.au/statistics/historical-data.html#interest-rates>

Using the ABS.Stat tool (<http://stat.data.abs.gov.au/>), most data can be extracted via APIs using Statistical Data and Metadata Standard (SDMX) format, chosen as our preferred mechanism to obtain data from ABS.

**Regression modelling technique**

The data being analysed for this project is made up of historical records of numeric values over a number of years, and multiple predictors to consider for forecasting the GDP. The regression modelling technique to be used will require to handle time-series analysis and time series forecasting.

Although it is early to pin-point a specific modelling technique, there are options for time-series forecasting techniques such as: Auto-Regression (AR) models, Simple Moving Average (SMA), Exponential Smoothing (SES), Autoregressive Integration Moving Average (ARIMA), Neural Network (NN), Holt-winters or Croston.

**Issues that may arise during project**

Due to the nature of a time series dataset, some identified potential issues are:

Not all the datasets gathered have the same amount of available data over a similar period span. Data will require cleaning, scaling and transformation as it is often required with time-series data. At this early stage, there are discrepancies found on the starting year range of available data.

During data exploration phase, we have encountered discrepancies with the frequency of data collection over the time period, most of our economic values have been found recorded at quartile intervals, however some have monthly or daily frequency and require to be transformed into quartile deductions to maintain homogeneity with other data. This carries a risk of potential analyst errors during data manipulation.

Forecasting a model for GDP implies making decision on the time horizon of predictions, a shorter time horizon would be easier to predict with higher confidence. This also leads into another aspect of the forecast in how frequent could the forecast be updated over time as new information becomes available (assuming latest information would imply more accurate predictions).

**References**

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>.

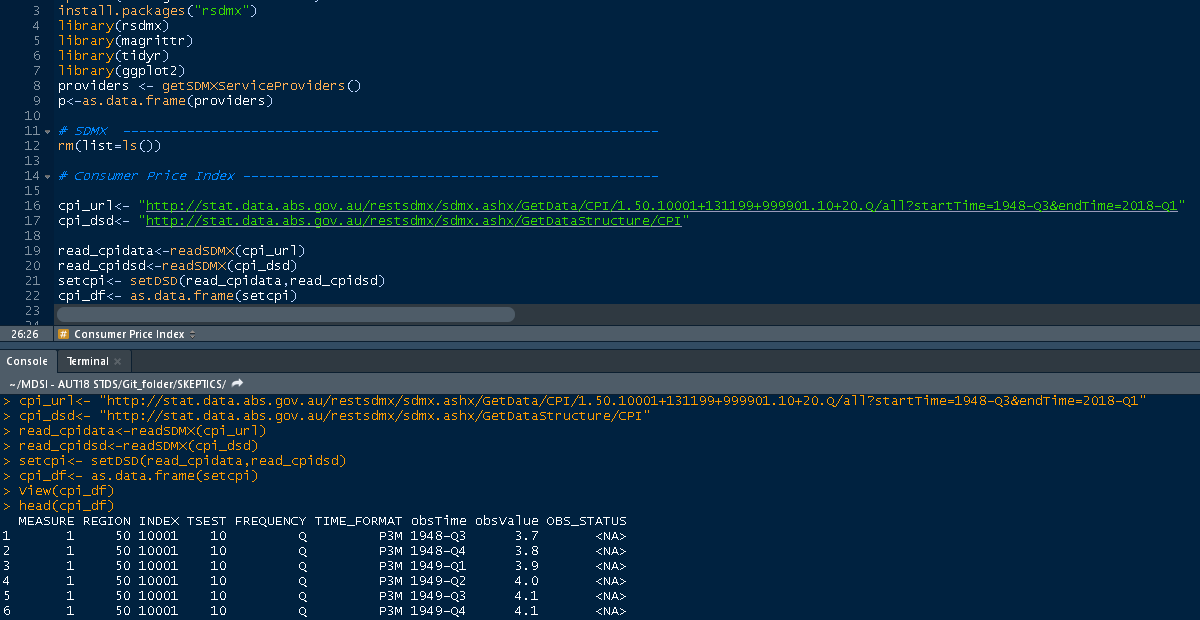
Investopedia 2018, *What is GDP and why is it so important to economists and investors?*, viewed 28 April 2018, < https://www.investopedia.com/ask/answers/199.asp>.

Woods, G. n.d., ‘Economic Indicators on the Internet’, *Economic Indicators on the Internet*, E-Brief, viewed 26 April 2018, <https://www.aph.gov.au/About\_Parliament/Parliamentary\_Departments/Parliamentary\_Library/Publications\_Archive/archive/ecindicators>.

**Appendix**

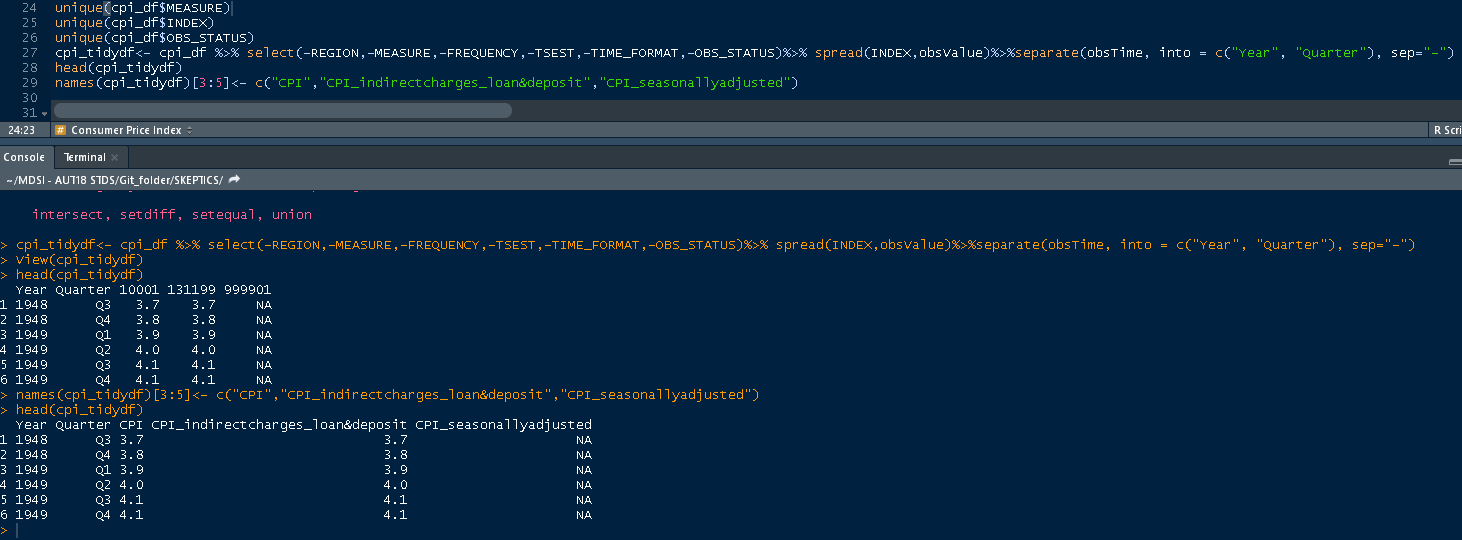
Below are some snippets of scripts using API methods for data extraction from the Australian Bureau of Statistics, we are using SDMX to bring data into R.

Obtaining Consumer Price Index data:

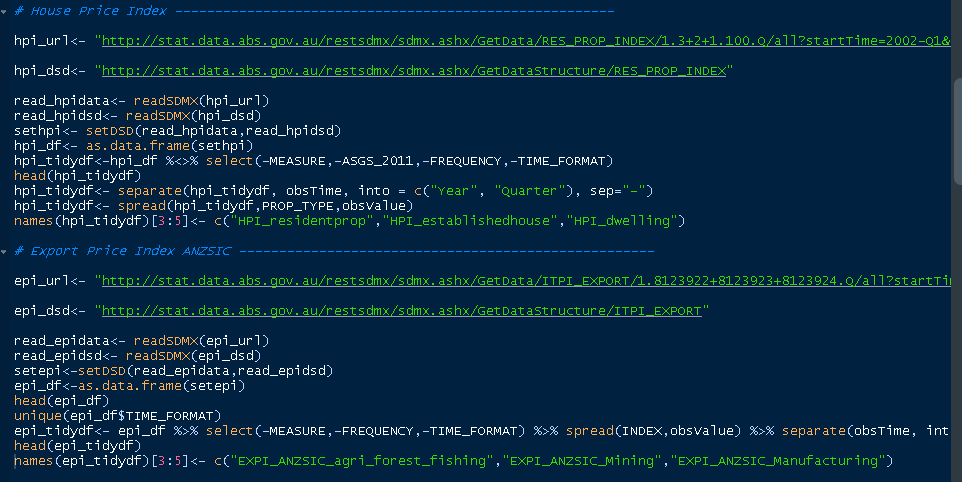


As data usually does not come in the format required to merge or analyse properly (see data format on screenshot above), we are required to transform obtained datasets to prepare data-frames ready for analysis and data merging.

Transformation of data (consumer price index) into usable data frame for project use

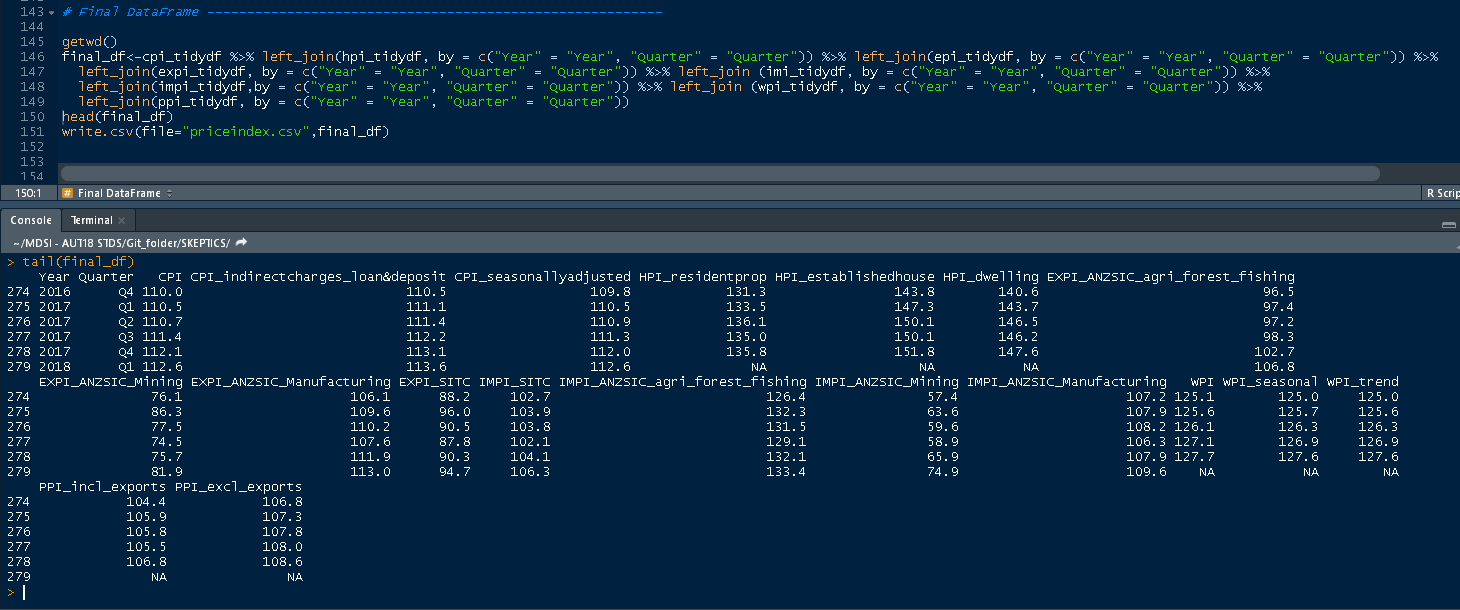


In similar fashion, other datasets have been obtained and transformed using same methods:



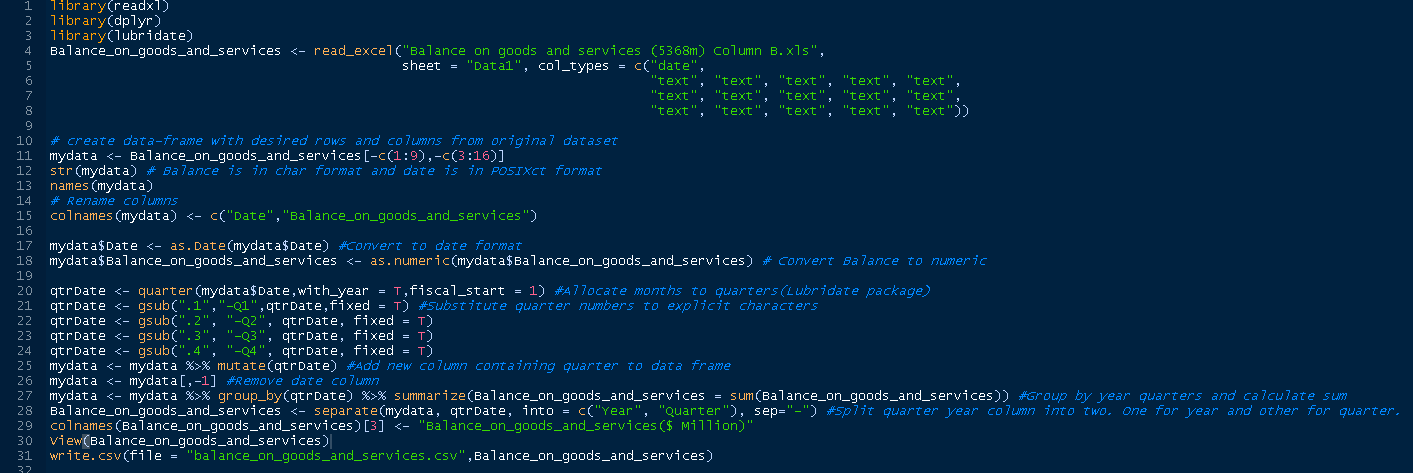
As GDP is derived from multiple economic indicators, and each is composed of multiple indices (derived from multiple datasets), we are employing data-merging techniques using left joins on year and quarter to match corresponding indices over a time-series spread of data.

Merging multiple data-frames on consumer price index into a master data-frame for project



Another sample of data-transformation can be observed on script below where dataset obtained on balance of goods and services was recorded on a monthly frequency and we are required to transform the values into quartile equivalents in order to merge with other datasets

Transforming a monthly frequency index into quartiles



Some of the data required for analysis does not come from data sources offering API interfaces to obtain them therefore we are required to download and import the csv, xls dataset and manipulate as required.

Importing downloaded (.xls) dataset and transformation to quartile frequency

