Image Source: www.rbnz.govt.nz



**Group Project Report**

# **Kiwi Budget: A DataSET of New Zealand Economic Activities (1999 ~ 2022)**

Jinze (Jerry) Zhou, Ansar Ali, Mingqiu (Jeffery) Chi

DATA422-S2

# Abstract

# This data set contains data covering major economic activities and a key monetary policy indicator (OCR) of New Zealand, from 1999 to 2022. Our analysis of this dataset verifies some correlations between OCR, CPI, HPI and other economic data. All datasets are scraped from online sources which are publicly available. The reuse potential includes training machine learning algorithms that do the forecasting. Individual researchers or organizations involved in policy making can use it for analysing and decision making. This data set is available through API published online.

# Purpose

Our aim is to collect datasets that contain a small but limited number of the most critical economic data indicators, such as monetary policy indicators (OCR only), house price indices (HPI), CPI, unemployment rate, tourism data and import/export data, with the expectation of seeing correlations between the data, which may provide a reference for financial planning for companies or individuals. ~~The reuse potential includes training machine learning algorithms that do the forecasting.~~

# Methodology

Our data collection, processing and analysis process basically follows the following steps as shown in following diagram:

Image Source: monkeylearn.com

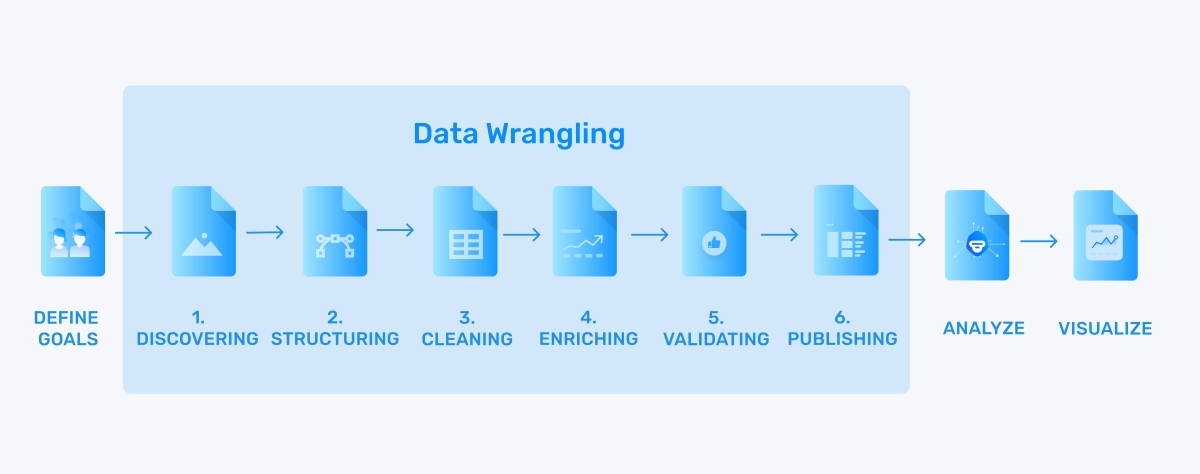


Figure Data Process Pipeline

We have summarized them into the following modules:

* Goals & Model
* Data Collection
* Data Wrangling
* Data Visualize & Analyze

## Model

The goals have been described in the previous section and we briefly describe the economic and financial rationale underlying the data.

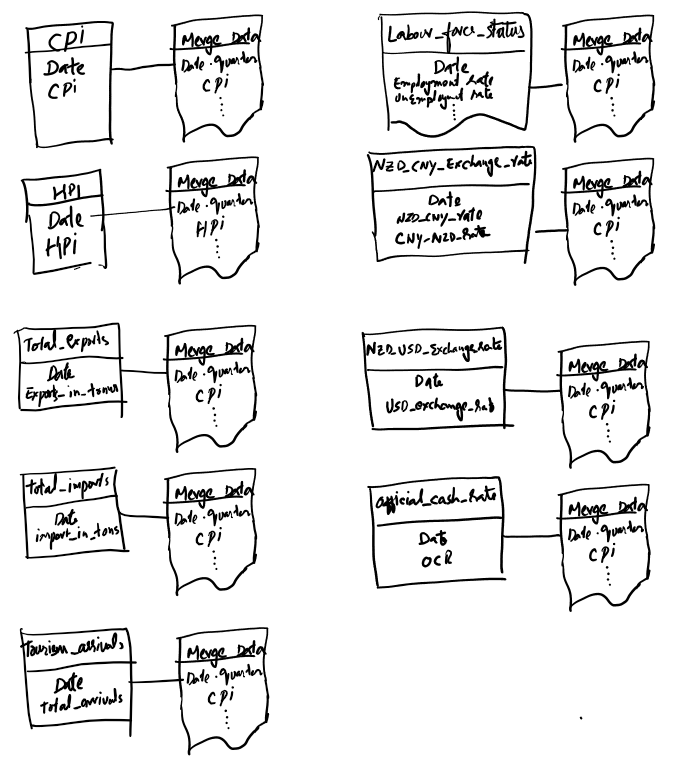
Monetary policy that are measures employed by governments to influence economic activity, specifically by manipulating the supplies of money and credit and by altering official cash rate (OCR). The usual objectives of monetary policy are to control inflation, maintain full employment and achieve high economic growth rates, but these objectives are in conflict. Inflation control is the RBNZ's only objective under the Banking (Prudential Regulation) Act 1989, but the conflicting objective of guaranteeing employment was planned for 2018 and added to the Reserve Bank of New Zealand Act 2021 (RBNZ, 2022). We believe we would see this conflict in the data. High CPI trends in recent years, caused New Zealand Reserved Bank to adopt measures to reduce inflation by restricting growth in the money supply and increasing the interest rate.

Based on economic principles, as New Zealand Reserve Bank sets its interest rate up, businesses will have difficulty getting cheap loans for further development, it is anticipated that business development will be reduced, hence in terms of cargo statistics, housing prices, and tourism counts, we are anticipating dropping in these activities as well as leads to increase of unemployment. On the other hand, an increase in the Reserve Bank interest rate will attract oversea hot money for a short-term investment, therefore it is anticipated that the exchange rate of New Zealand dollars VS. Other currency will go up.

We will collecting online public data to verify economic principles and examine relationship between data. We collected the following data to test the correlation:

* Official cash rate
* CPI
* Unemployment rate
* HPI (Housing price index)
* Exchange rate(NZDUSD/NZDRMB)
* Tourism
* Exports and Imports data.

The data is collected separately and wrangled into a table; the overall structure is shown in the figure below:



Data collection, wrangling, visualize and analyze will be describe in the following section.

# Data Collection

## Data Source

As all data is publicly available on the RBNZ, it is in principle more reliable to collect data from the source where it was originally published, which is the RBNZ website. However, due to the redesign of the RBNZ website, which uses Cloudflare as the CDN and refuses to automatically crawl content, we failed to scrape data from RBNZ although we tried various methods of disguise. As time constraints, we had to choose other crawler-friendly sites to get the data. After much searching and searching we finally settled on three crawler friendly sites to get the data:

* infoshare.stats.govt.nz
  + Labour force status data (Unemployment rate)
  + Tourism visitor data
  + Import & Exports data
* interest.co.nz
  + Official Cash Rate
  + CPI & HPI
  + NZD-USD Exchange Rate
* api.ofx.com
  + NZD-RMB Exchange Rate

Infoshare is a free service provided by Statistics New Zealand to allow the viewing of survey data. The website “interest.co.nz” is partnered with Calculate.co.nz which has created New Zealand's largest collection of calculators. While “ofx.com” is a currency exchange platform regulated by over 50 regulators globally. The backgrounds of these data sources give them high creditability.

## Technologies

The data is collected using two types of technology, one based on R, with RSelenium as a crawler and XML2 as a scraper, the other one is based on Julia collecting data via API. Specifically, we use R technique to get information from Infoshare website and use Julia to get information from the other two sites.

Infoshare a is a database-driven website, as shown in the figure below, in the process of obtaining our target data we need to click to expand each category, then select specific conditions and submit the request, and finally the site then returns the information. The idea is to automate the process of clicking and selecting and then parsing the html elements after the final information page return using the Xml2 module. So RSelenium was selected.

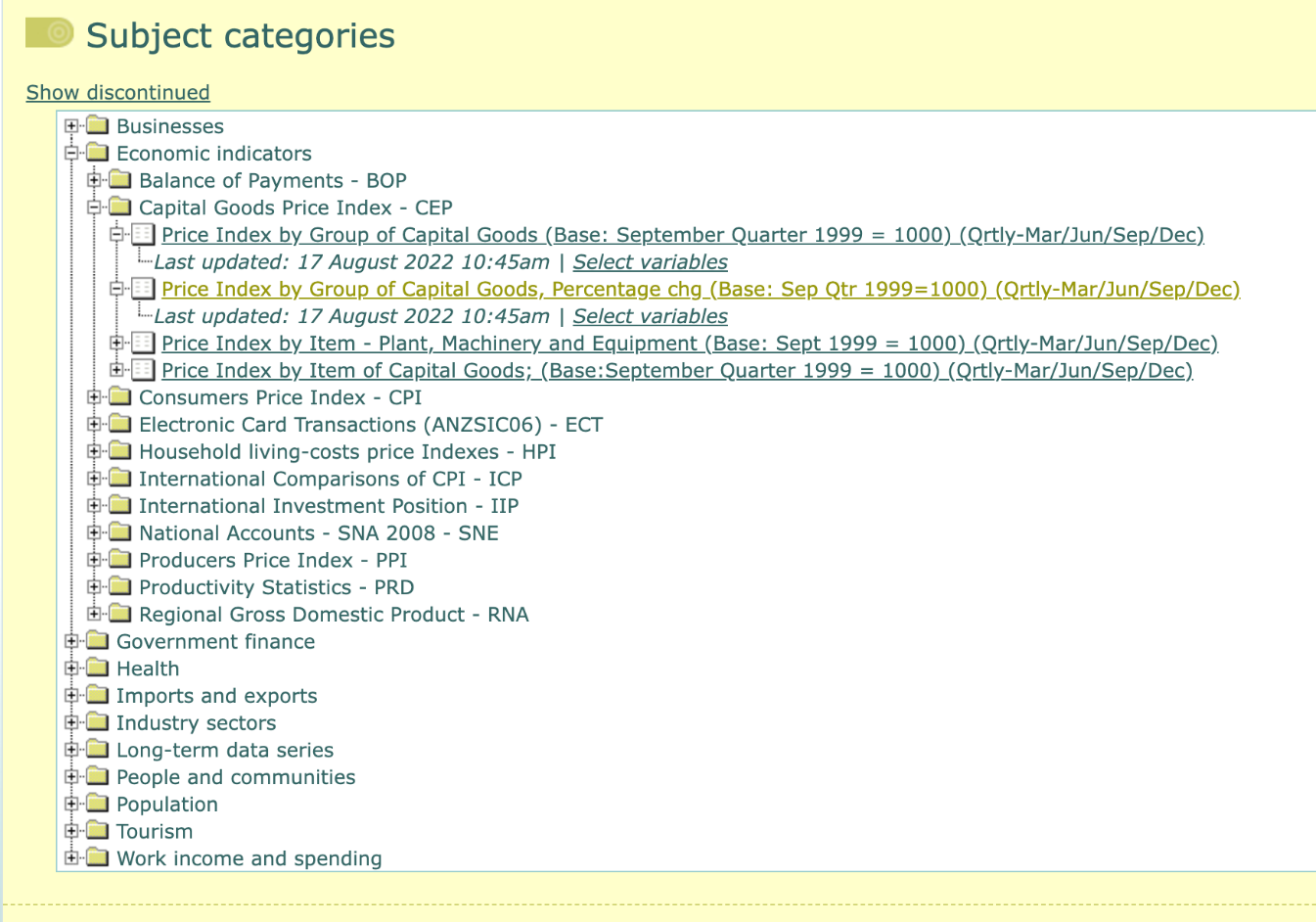


Figure Infoshare

RSelenium Provides a set of R bindings for the 'Selenium 2.0 WebDriver' that allows driving a web browser natively and automating browsers locally or remotely (Kim, 2022). The targeting of a clicked element is determined by a CSS selector, which is submitted to the RSelenium API as a parameter, and the corresponding click and select operations are performed to enable page navigation. The code example is shown below:



Figure RSelenium code example

Xml2 is a wrapper for libxml2 C package that makes working with XML and HTML in R easier. After we get the target information page, we can use Xml2 module to extract data as below:



Figure extract data by Xml2

The code to implement the complete process is in the “[Collector\_R\_WebDriver.ipynb](https://github.com/ansrali/data422-group-project/blob/main/Collector_R_WebDriver.ipynb)” notebook file.

Another technology used to collect data is Julia, with the modules of HTTP, JSON, and DataFrames. All the implementation code for this part is in the “Collector\_Juia\_API.ipynb” notebook file. Firstly, monitor website interaction traffic via Chrome DevTools to find data POST request interface parameters and the format of the returned JOSN data. Secondly, the request process is implemented via the HTTP module to get the returned data in JSON format. Finally, the data is parsed by the JSON module and then converted to DataFrames. The following is a short sample code:



Figure Julia API Code example

All the API data collection code implemented by Julia is in the “[Collector\_Juia\_API.ipynb](https://github.com/ansrali/data422-group-project/blob/main/Collector_Juia_API.ipynb)” file.

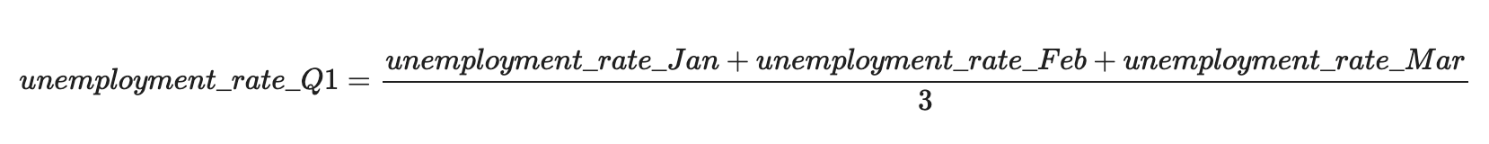
After completing the data acquisition, we are using the **writexl** package in R and the **CSV** module in Julia to save the data to an Excel/CSV file.

# Data Wrangling

## Method

All data are time-series data; the maximum granularity of time is quarterly, and the smaller granularity is monthly, with occasional individual months having a few points. For cases where there are multiple data points in a month, the multi-point average is used as the data for that month so that the smallest granularity can be considered monthly.

And we convert the data to quarterly scale because we consider that economic changes generally occur on a quarterly or even annual basis. The quarterly data of the rate of change can be considered as the product of the months. But inside our data set, there is no such data conversion. Quarterly data for non-rate of change will simply be treated as an arithmetic average of monthly data, such as the unemployment rate data.



Since they are all time-series data, the missing data are filled in using the strategy of filling in from the front to the back and then from the back to the front.

## Technologies

We perform data wrangling and visualization based on R. As described above, there are several tasks required in the wrangling:

* Data Import
* Time Format Transform
* Data Filtering & Combining
* Missing value padding

The packages used to finish these tasks are tidyverse, timeDate, visdat and here packages. Several functions are created based on timeDate package to finish time format converting. With support for various time format conversions, month-based data grouping and averaging, and time-based data filtering are relatively convenient.

The tidyverse is an opinionated set of R packages for data science utilities with s a common design philosophy, language, and data structures (Locke, 2017). We use packages in tidyverse to finish most wrangling task:

* **readr** and **readxl** packages are used to import data from Excel and CSV file, and export of final merged data.
* **purrr** package applied functional programming techniques for row-by-row conversion of time formats in datasets.
* filtering the data outside the selected time range(1999 - 2022) and left join to a complete unified table.
* Fills missing values in selected columns using the next or previous entry.

After data conversion and merging, the visdat package can be used to examine the integrity of data, then the complete merged table is saved as a CSV file for subsequent visualization and analysis. All data wrangling relevant codes are in the "[Data\_Wrangling.ipynb](https://github.com/ansrali/data422-group-project/blob/main/Data_Wrangling.ipynb" \o "Data_Wrangling.ipynb)" notebook file.

# Data Visualization and Analyses

## Technologies

The data visualization uses R's ggplot2, and the main type of graph is a time-based multi-curve presentation. The two-axis graph is used to display two measures of data, such as OCR data and import/export data. The second Y-axis corresponding to the second measure of data needs to be offset and transformed so that the second measure can be presented simultaneously with the first measure of data. Corresponding offset, conversion functions have been created to meet this requirement.

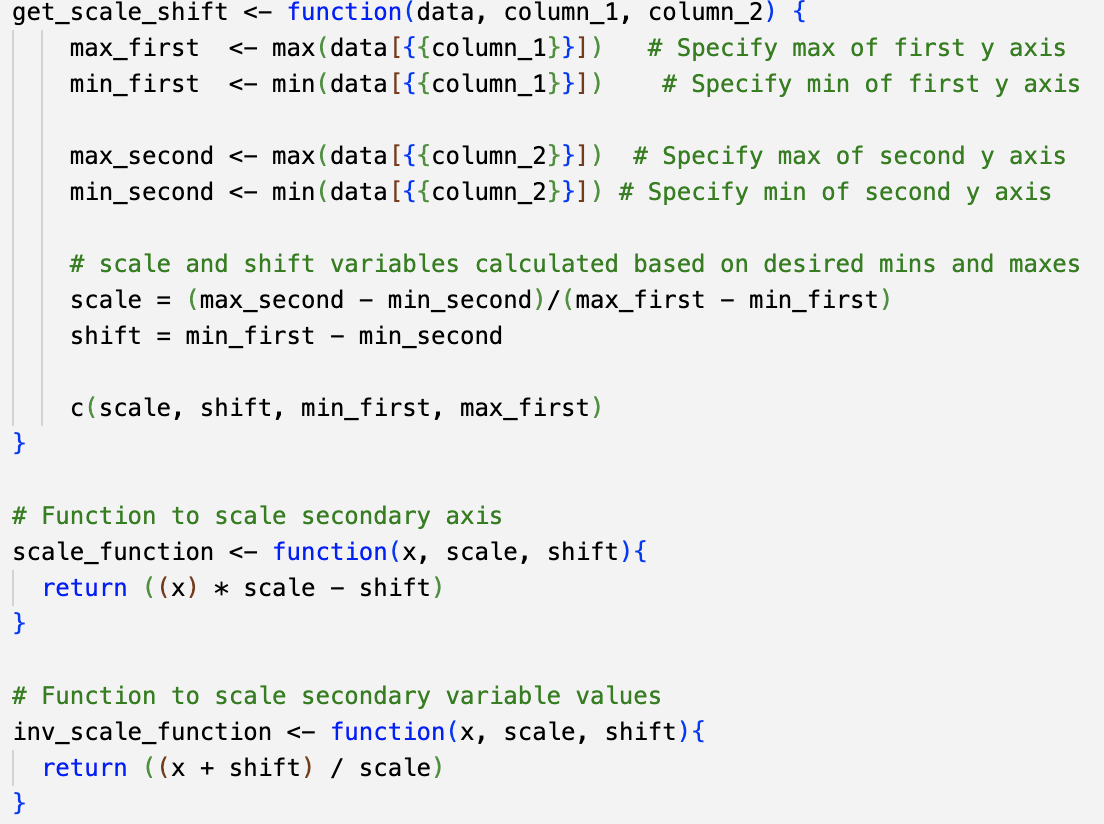


Figure Scale conversion

The ggcorrplot package is used to present the correlation in dataset.

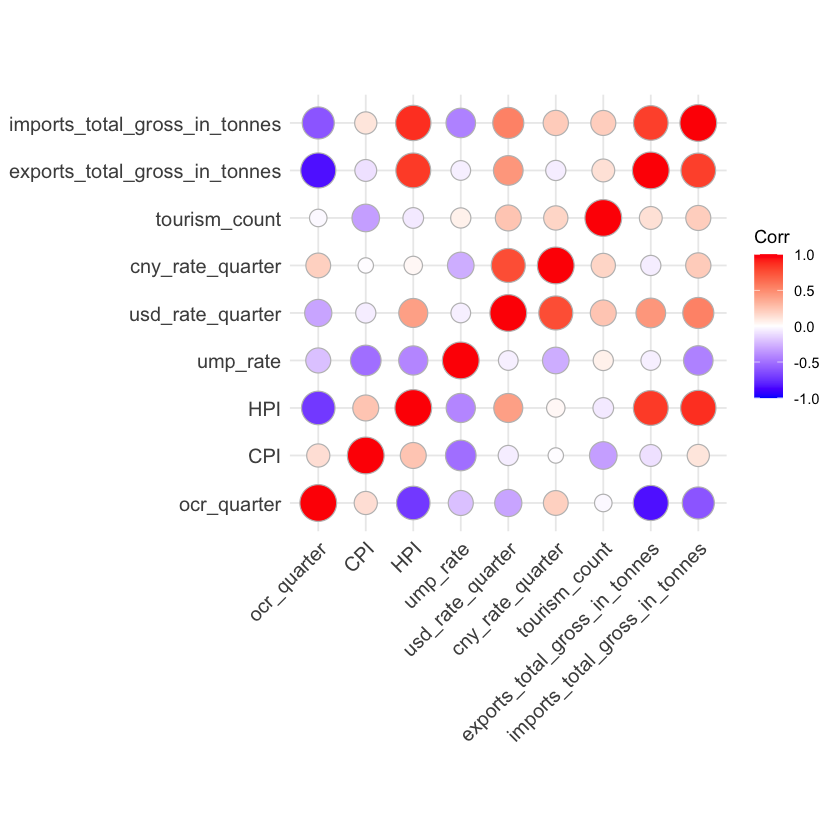


Figure correlation matrix

All data visualization code is in the " [Data\_Visualisation.ipynb](https://github.com/ansrali/data422-group-project/blob/main/Data_Visualisation.ipynb)" file

# Data API publishing

## Technologies

The final cleaned and wrangled dataset has been published for the public in the form of REST-API on AMAZON (AWS) EC2 instance. To achieve this R and Plumber are used on the Ubuntu distribution. To acquire data following endpoints are developed:

Endpoints:

*~/data\_summary:* This endpoint provides a summary of each variable in the dataset. e.g., mean, quartiles, count of each factor etc.

URL:

http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/data\_summary

*~/data\_variables:* This endpoint provides names of all columns in the dataset.

URL:

http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/data\_variables

*~/data\_all:* This endpoint sends back the whole dataset in Json format.

URL:

http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/data\_all

*~/plot:* This endpoint prints the plot of all indicators in one plot similar to the visualization section in this report.

URL:

http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/plot

*~/data\_selected:* This endpoint provides data between selected dates. Date range can be specified by using '*year\_from'* and '*year\_to*' parameters in url.

Example:

http://ec2-3-27-14-25.ap-southeast- 2.compute.amazonaws.com:8000/data\_selected?year\_from=1999&year\_to=2001

# Summary

Limitation

# Reference

RBNZ (2022). *Our legislation*. https://www.rbnz.govt.nz/about-us/responsibility-and-accountability/our-legislation.

Kim, J. Y. (2022). *RSelenium package*. RDocumentation. <https://www.rdocumentation.org/packages/RSelenium/versions/1.7.9>

Locke, S. (2017). *Data Manipulation in R*. Locke Data.

Total word counts excluding the tables, reference and appendix are xxxx.