Image Source: www.rbnz.govt.nz



**Group Project Report**

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

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

DATA-422-S2

# Abstract

# This data set contains information of various major social economical activities of New Zealand as well as its key monetary policy indicators from year 1996 to year 2022. All datasets are time series based and scraped from publicly available online sources. The reuse potential includes training machine learning algorithms that do the forecasting. Individual researchers or organizations involved in policy making can use it for analyzing and decision making. This data set is available through API published online.

# Purpose

New Zealand is a relatively small economical entity, it is a highly developed free-market economy. It is the 50th-largest national economy in the world when measured by nominal gross domestic product (GDP). Trade is critical to New Zealand’s economy; it is heavily relying on overseas business. We can only pay for the goods and services we import from overseas by selling exports to other countries. At the moment, international trade (exports and imports) makes up around 60% of New Zealand’s total economic activity.[3]

Our aim is to collect a data set that contains a small but limited number of the most critical economic data indicators, such as monetary policy indicators (OCR only), house price indices (HPI), CPI consumer price index (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. Reuse potential includes but not limited to retrospective analysis using data visualization and training machine learning algorithms that do the forecasting.

# Methodology

We have conducted our project primary based on the steps shown in following diagram:

Image Source: monkeylearn.com

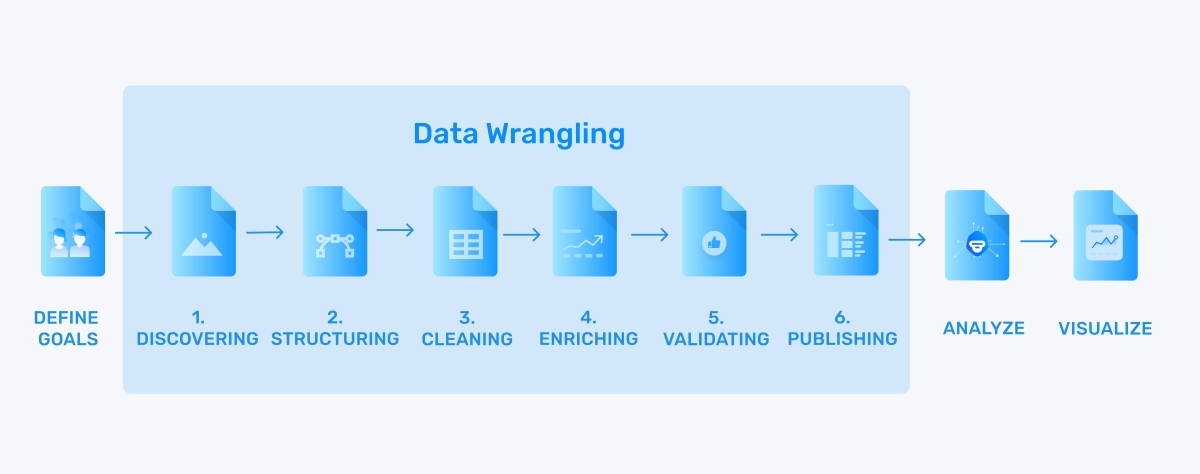


Figure 1 Data Process Pipeline

The main steps can be summarized into the following modules:

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

## Goals & Data Model

Our goal is to collect sufficient data regarding New Zealand economy as well as its Monterey policies, so we are able to produce a good view of the country and analyze the effectiness of it policies via data visualization.

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 common objectives such policy are to control inflation, maintain full employment and achieve high economic growth rates, but these objectives are often in conflict with each other. Inflation control is the RBNZ's only objective under the Banking (Prudential Regulation) Act 1989, but the conflicting objective of guaranteeing employment was also planned in year 2018 and added to the Reserve Bank of New Zealand Act 2021 (RBNZ, 2022). We believe we would be able to observe such conflict in the our 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.

The relationship key between these entities are timestamps (Year, Month, Quarter); Based on economical principles, as New Zealand Reserve Bank sets its interest rate up, business will have difficult of getting cheap loans for further development, it is anticipated that business development will be reduced, which leads to increment of unemployment. On the other hand, increase of Reserve Bank interest rate will attract oversea hot money for short term investment, therefore it is anticipated that exchange rate of New Zealand dollars VS. Other currency mainly US dollars will go up. The high Reserve bank interest rate will lead to lower level of economic activities, subsequently in term of cargo freight statistics, general vehicle activities will drop. At the same tim, raising interest rate usually leads to dropping of the CPI.

When the economic activities are low and unemployment rate is high, the government is under pressure to unleash new monetary policies to stimulate economy. Hence government usually resolve to lower reserve bank interest rate. With high reserve bank interest rate, we should be able to observe weak New Zealand dollars VS over currency. Thus make export of our products cheaper, therefore attracts more oversea orders, subsequently we should be able to observe an increase of freight transportation activities, oversea passage flights. To stimulate general economic growth, government usually starts major infrastructural projects like major road works, etc. Hence it is also anticipated that general road construction activities are increasing.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.

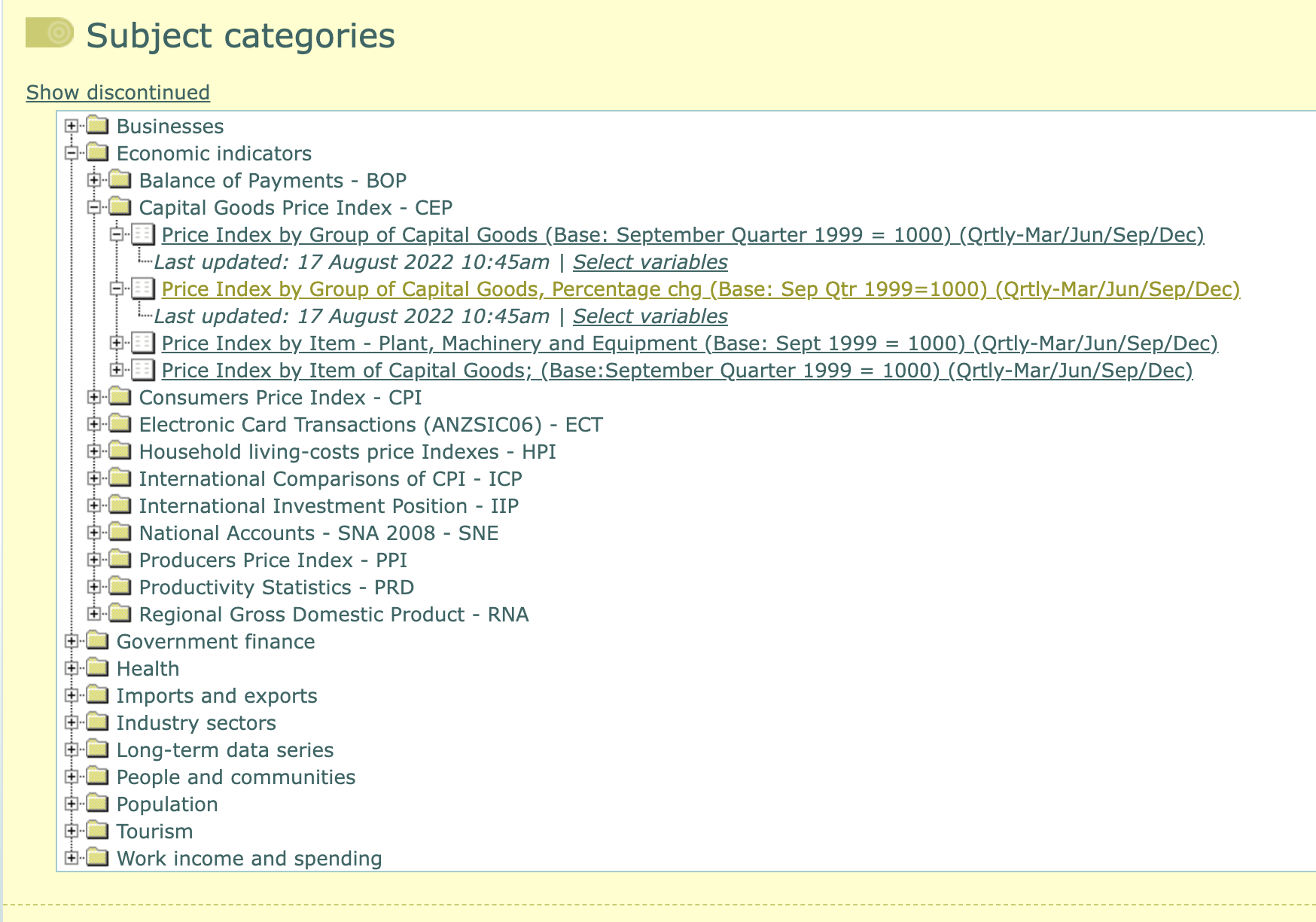


Figure 2 Infoshare

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 3 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 4 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 5 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.

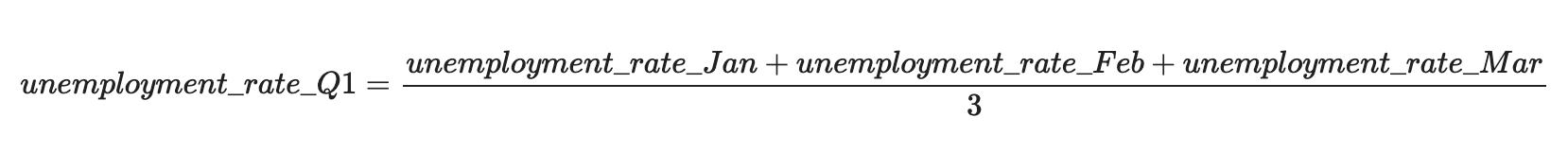
After completing the data acquisition, we are using the **writexl** package in R and the **CSV** module in Julia to save the data to an Excel/CSV file.

# 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 base 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 package, and the majority type of graph is the 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 measurement of data. Corresponding offset, conversion functions have been created to meet this requirement.

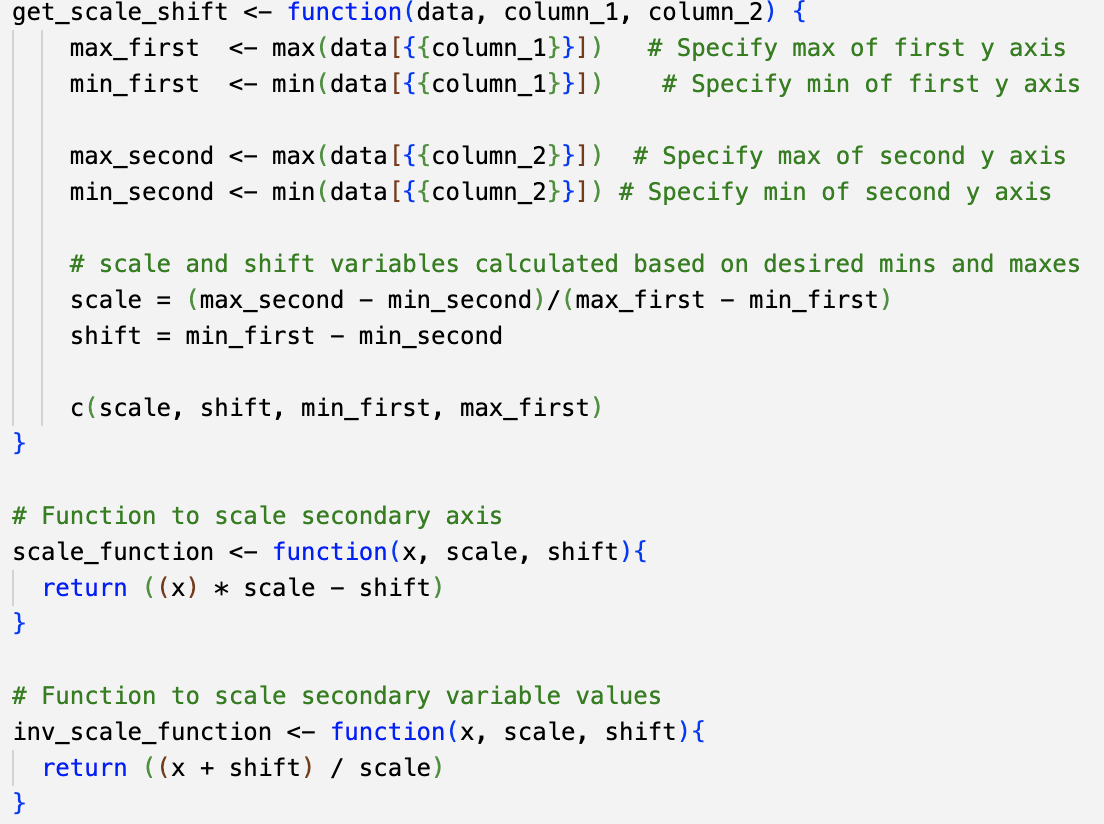


Figure 6 Scale conversion

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

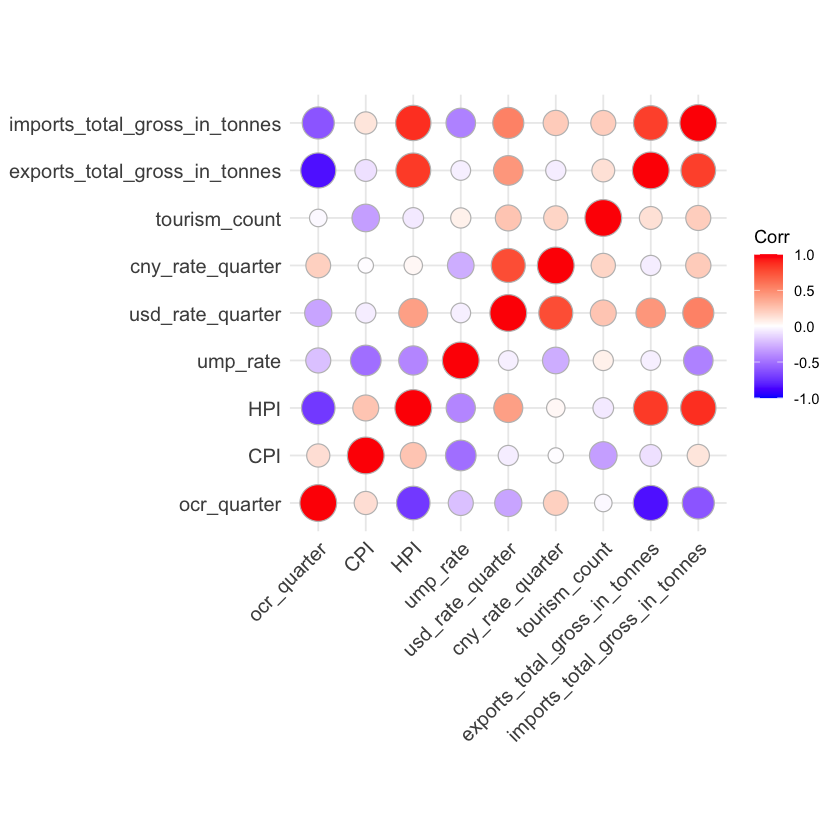


Figure 7 correlation matrix

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

## visualization & Analysis



The above chart reveals the relationship between “Unemployment rate” VS. “Official Cash Rate”; Higher OCR leads to Higher interest rates could lead to rise in unemployment. In the above graph we produced, for the majority of the timeline, the positive correlation between “Unemployment rate”and “OCR” has been observed, which complies to the economic theory. There are exceptions like between year 2003 to year 2008, consider New Zealand economy is heavily rely on export, other factors such as “NZ dollar VS. US dollar exchange rate” which may influence the export business that can be used to explain this phenomenon.



We have plotted “NZ dollar VS. US dollar exchange rate” against “Tourism counts”. It seems in the last three years, the exchange rate has been relatively stable and tourism counts are slowly increasing. Right up until beginning of year 2020 when COVID-19 pandemic breaks up, due to the COVID-19 lockdown, we have observed a sudden drop of the tourism visits.



The official cash rate (OCR) is the interest rate set by the Monetary Policy Committee (MPC). It affects the price of borrowing and saving. Usually, with low inflation, we have low interest rate, Interest rates tend to move in the same direction as inflation but with lags, because interest rates are the primary tool used by central banks to manage inflation. Such positive correlation between “OCR” and “CPI” has been observed from above plot, which proves the effectiveness of the New Zealand monetary policy.



From above plot, we can see that the growth of the HPI starts to exceed the growth of the CPI from year 2012 onwards, from this diagram, we can appreciate the occurrence of the housing affordability issues. CPI usually relate to affordability of food, drink and other commodities. From reading this chart, we can see from year 2017, both CPI and HPI are surging, indicates affordability issues of both the commodities and housing. Under such circumstances Government should step to ease the surging of the inflation in short term by raise interest rate, but in long term, the solution lies in building more houses and increase our productivity as well as finding cheaper source of purchasing affordable commodities.



On the above chart, purple means negative strong relationship, Red is strong positive relationship. Based on the correlation diagram that we have produced, the following relationships have been identified.

1. Tourism counts has a relatively weak negative relationship with CPI, indicates that rising of the costs and inflation does NOT significantly impacts on tourism vising New Zealand, at least not for the year from 1999 ~ 2020.
2. CNY\_Rate\_Quarter appears to have a strong positive relationship with USD\_rate\_quarter, this phenomena occurs for a good reason. China does not have a floating exchange rate that is determined by market, instead it pegs its currency, the yuan (or renminbi), to the U.S. dollar at 8.28 for more than a decade starting in 1994.
3. Unemployment rate appears to heave a relatively strong negative relationship with CPI, which complied to the economical principle that when high inflation usually goes with low interest rate which leads to higher employment rate (low unemployment). The same principle also applies backwards.
4. HPI appears to have a strong negative relationship with OCR, this phenomena can be explained by that high mortgage rate increases the cost of purchasing housing as it is more expensive to use mortgage financing. Thus, the demand for housing falls so as the housing price fall.

# 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

To conclude our study, by following a clearly defined goal, we have carried out in-depth discussion and analysis of the topic. An extensive researches on the available data sources has been performed. After some trial and error, reliable data sources have been identified and we have coded for web contents scraping, followed by data wrangling and visualization. The analytic results on the data visualization complied to the economic theory and principle as well as major international events has further proved our data collection to be successful.

# 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.