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 (Jeffrey) Chi

DATA422-S2

# Abstract

# This data set contains information on various major economic activities of New Zealand as well as its key monetary policy indicators from the year 1996 to 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 policymaking can use it for analysis and decision-making. This data set is available through API published online.

# Purpose

New Zealand is a relatively small developed free market economy economic entity. It is ranked 50th economy in the world measured by gross domestic product (GDP) [4]. Trade is critical to our economy. Our prosperity relies on overseas business. Kiwi needs to pay for the goods and services imported overseas by selling our products to other countries. Currently, international trade (exports and imports) makes up around 60% of the 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 is not limited to retrospective analysis using data visualization and training machine learning algorithms that do the forecasting.

# Methodology

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

Image Source: monkeylearn.com

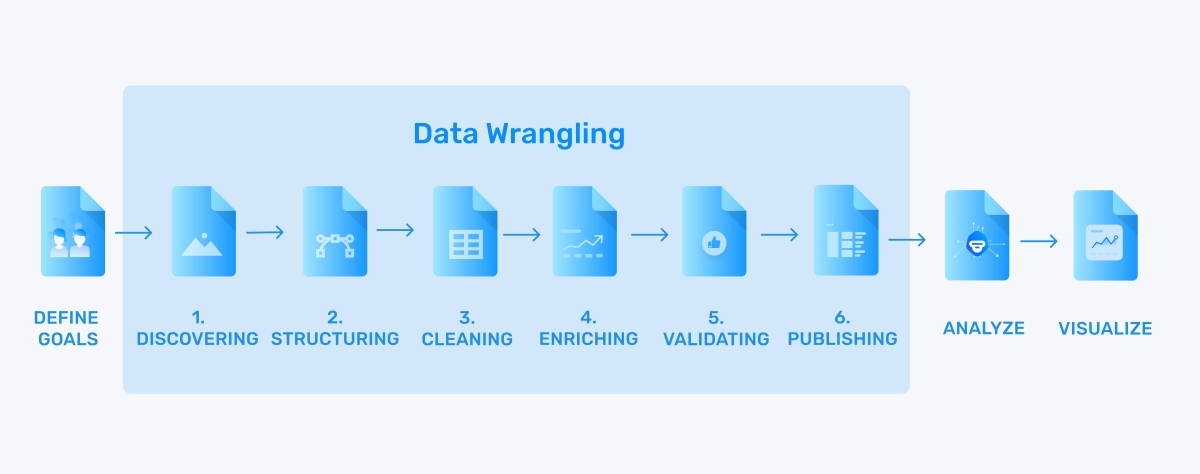


Figure 1 Data Process Pipeline

The main steps can be summarized into the following modules:

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

## Data Model

Monetary policy that are measures employed by governments to influence economic activity, to be more specifically, by manipulating the supplies of money & credit as well as the official cash rate (OCR). The common objectives of such manipulation are to control inflation, maintain a high employment rate and boost 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 the 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 our data. High CPI trends in recent years caused New Zealand Reserved Bank to restrict the money supply and increase the interest rate in an attempt to combat the high inflation rate.

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

When economic activities are low and the unemployment rate is high, the government is under pressure to unleash new monetary policies to stimulate the economy. Hence government usually resolve to lower reserve bank interest rate. With a 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 overseas orders. Subsequently, we should be able to observe an increase in freight transportation activities and oversea passage flights. To stimulate general economic growth, the 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 collect online public data to verify economic principles and examine the 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, visualization and analysis will be described 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 crawl content automatically, 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 provides a free view of the survey data provided by Statistics New Zealand. The website “interest.co.nz” is partnered with “calculate.co.nz”, which has created New Zealand's largest collection of calculators. “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, and the other one is based on Julia collecting data via API. Specifically, we use the R technique to get information from the Infoshare website and use Julia to get information from the other two sites.

Infoshare 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.

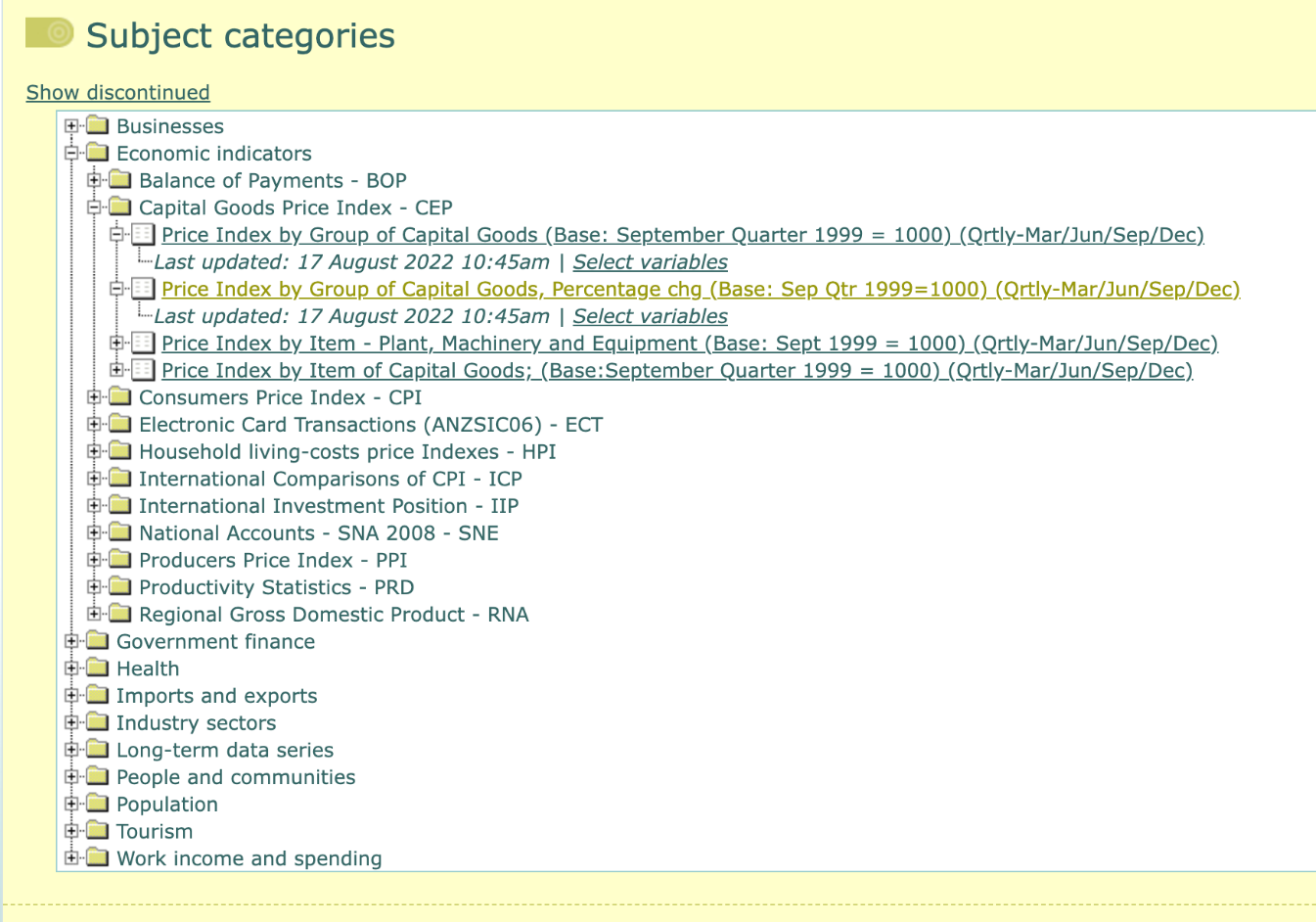


Figure 2 Infoshare

RSelenium offers a set of R libraries for the 'Selenium 2.0 WebDriver' that drives a web browser and automates the 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 the “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 JSON 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" \o "Collector_Juia_API.ipynb)” file.

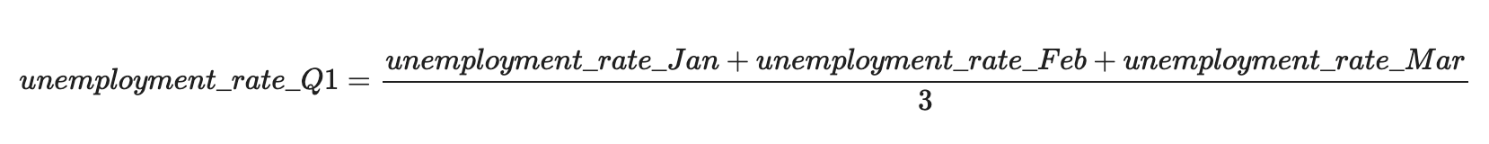
After completing the data acquisition, we are using the “**writexl”** package in R and the “**CSV”** module in Julia to save the data to an Excel/CSV file.

# 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 a 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 the “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 tasks:

* **“readr”** and “**readxl”** packages are used to import data from Excel and CSV files and export final merged data.
* **The “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.

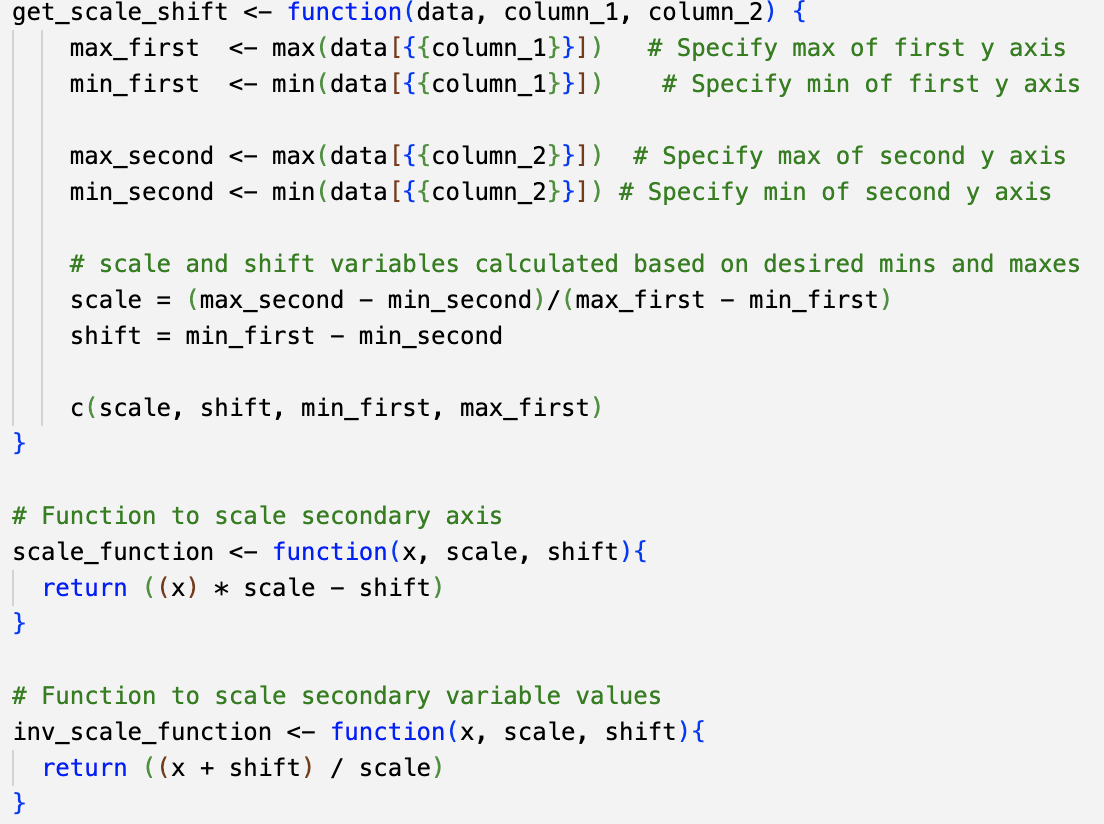


Figure 6 Scale conversion

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

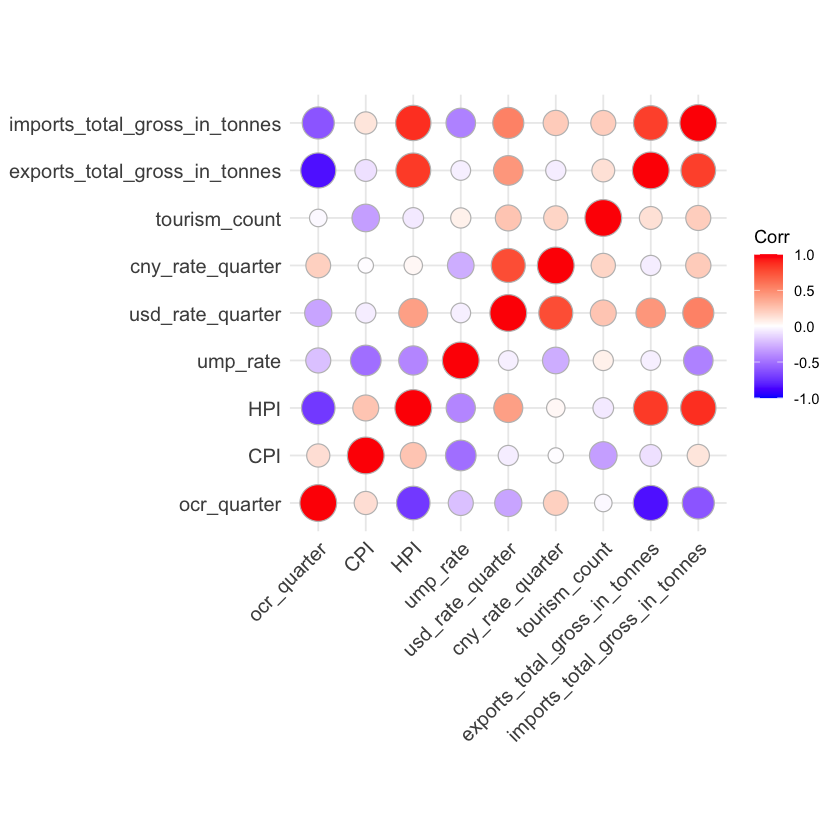


Figure 7 correlation matrix

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 the “Unemployment rate” VS “Official Cash Rate”; Higher OCR leads to Higher interest rates could lead to a rise in unemployment. In the above graph we produced, for most of the timeline, the positive correlation between “Unemployment rate” and “OCR” has been observed, which complies with the economic theory. There are exceptions, like between the years 2003 to the year 2008, considering New Zealand's economy heavily relies on export. Other factors, such as the “N.Z. dollar VS US dollar exchange rate”, may influence the export business, which can be used to explain this phenomenon.



We have plotted the “N.Z. 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 the beginning of the year 2020, when the COVID-19 pandemic breaks up due to the COVID-19 lockdown, we observe a sudden drop in tourism visits.



The monetary Policy Committee (MPC) of New Zealand sets the interest rate set, which is also named the official cash rate (OCR), which affects the costs of borrowing and saving. Interest rates trends to lag behind inflation trends but move in the same direction, as interest rates are the primary tool used by central banks to manage inflation. Such a positive but delayed correlation between “OCR” and “CPI” has been observed from the above plot, which proves the effectiveness of the New Zealand monetary policy.



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



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

1. Tourism counts have a relatively weak negative relationship with CPI, indicating that rising costs and inflation does NOT significantly impact tourism vising New Zealand, at least not for the year from 1999 ~ to 2020.
2. CNY\_Rate\_Quarter appears to have a strong positive relationship with USD\_rate\_quarter. This phenomenon occurs for a good reason. China does not have a marketed-driven floating exchange rate. Instead, it pegs its currency to the U.S. dollar.
3. The unemployment rate appears to have a relatively strong negative relationship with CPI, which complies with the economic principle that when high inflation usually goes with a low-interest rate, which leads to a higher employment rate (low unemployment). The same principle also applies backwards.
4. HPI appears to have a strong negative relationship with OCR; this phenomenon can be explained by the 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

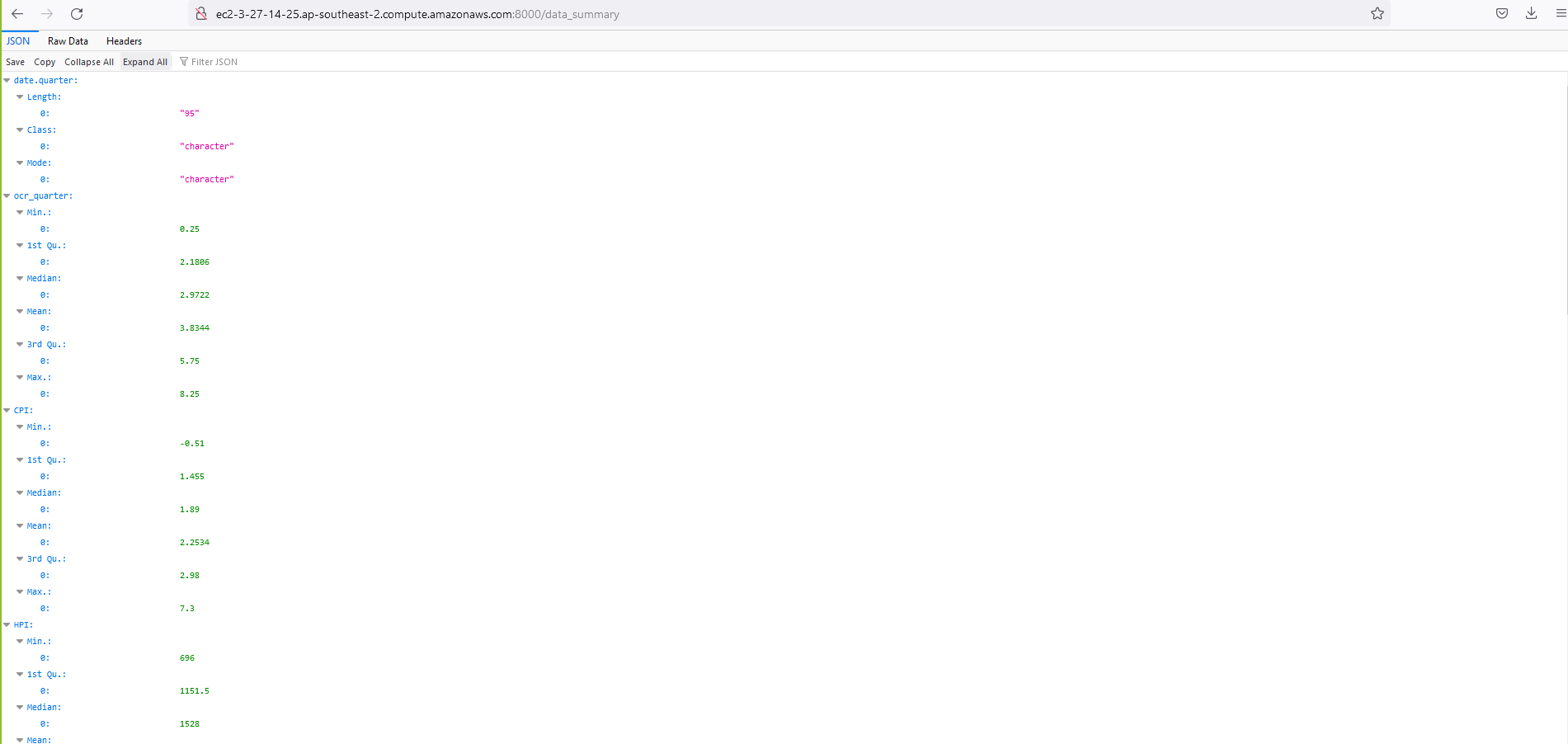
## Technologies

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:

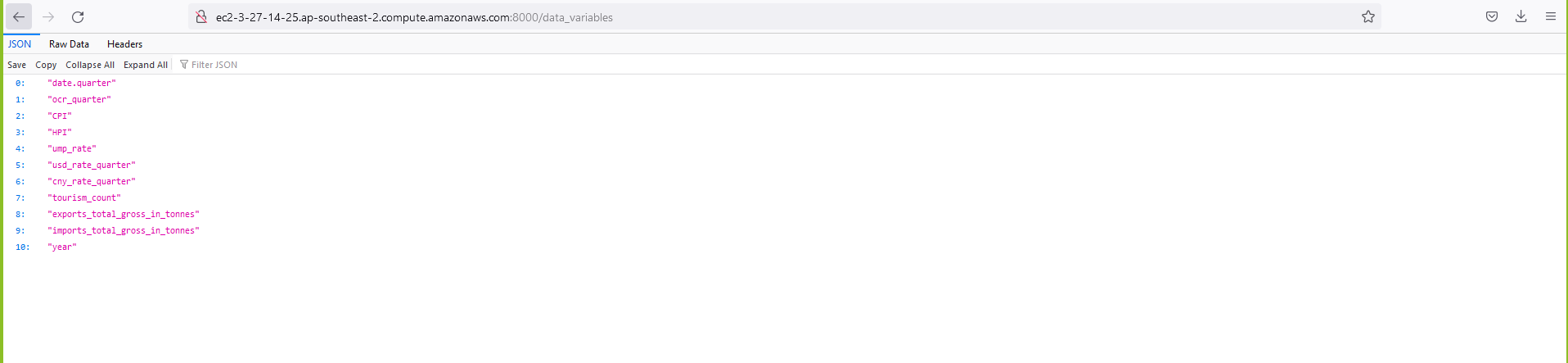
*~/data\_summary:* This endpoint provides a summary of each variable in the dataset. e.g., mean, quartiles, count of each factor etc.

URL: <http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/data_summary>



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

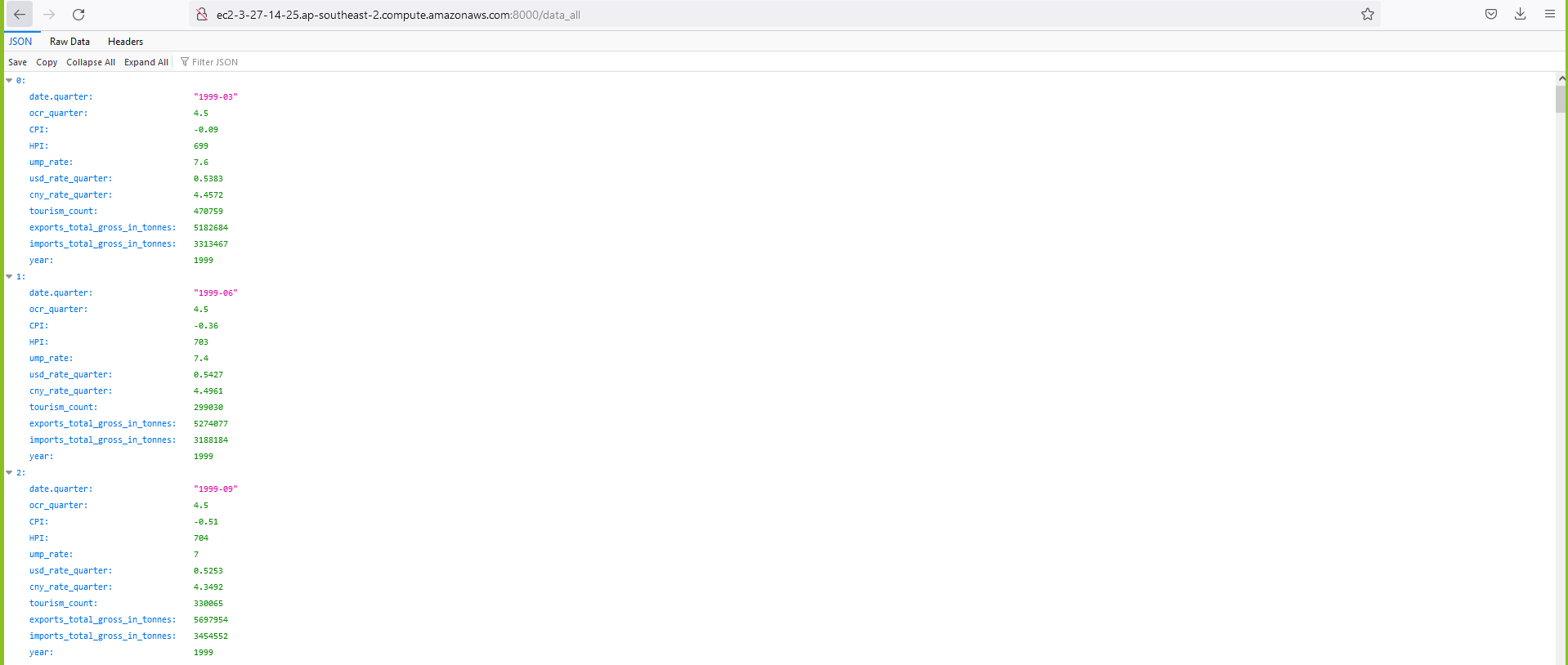
URL: http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/data\_variables



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

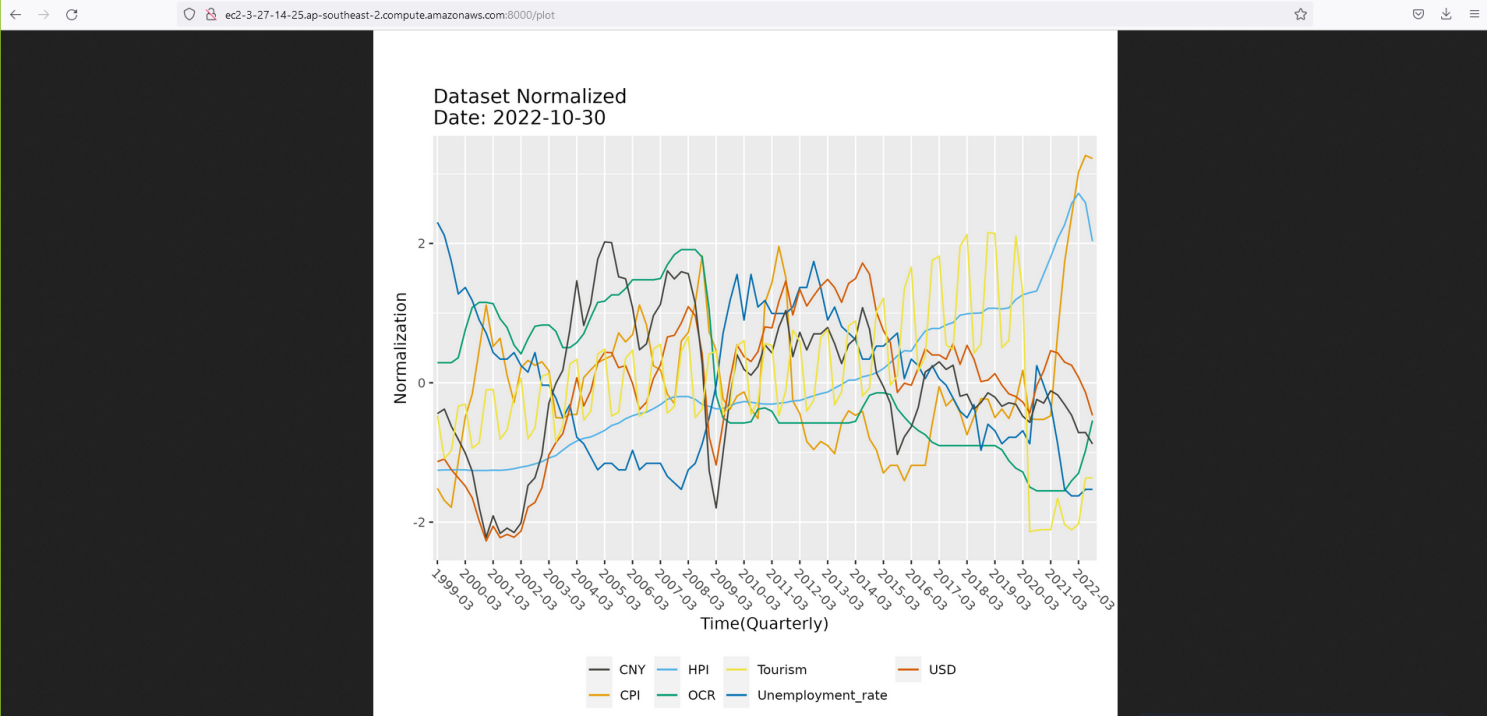
URL:

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, like the visualization section in this report.

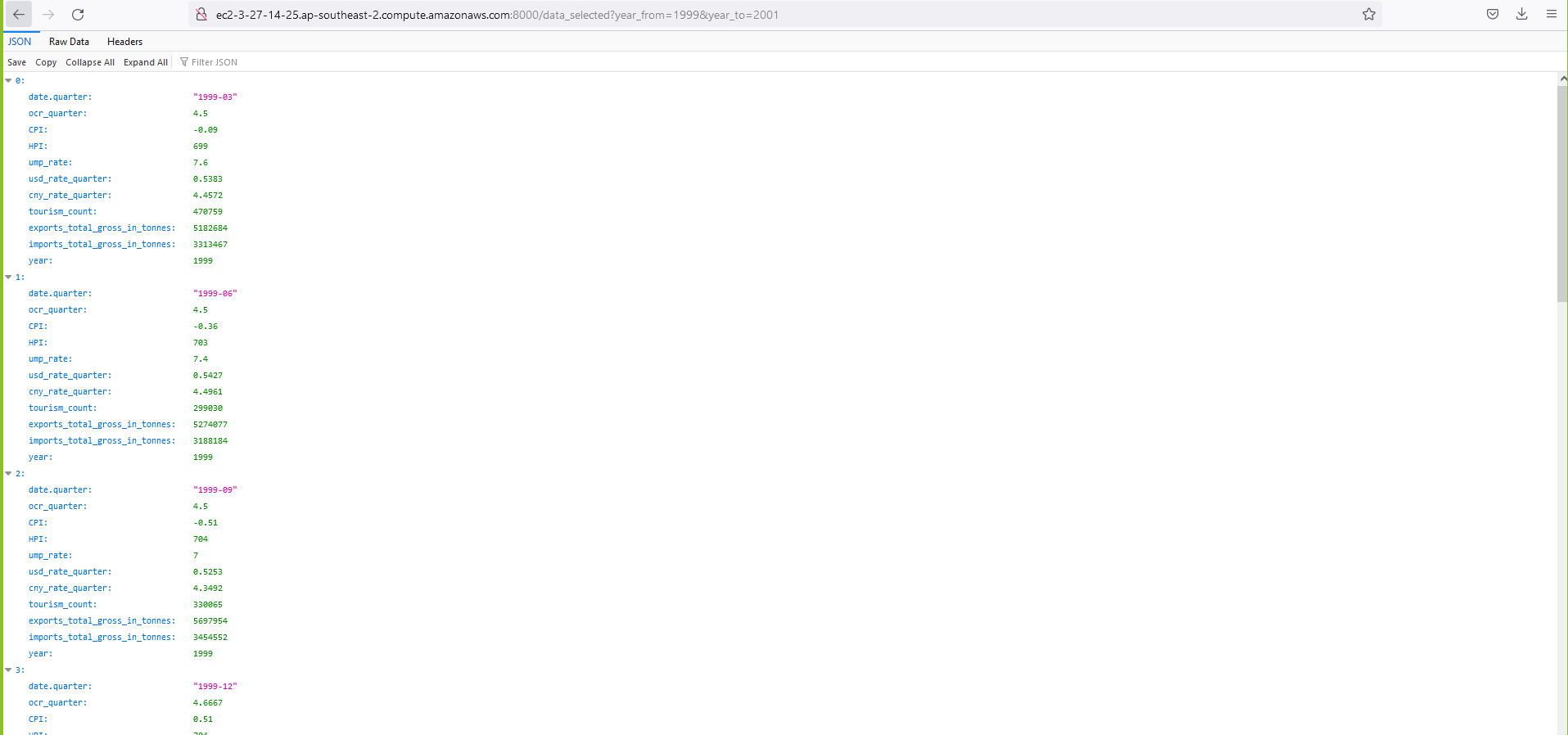
URL: http://ec2-3-27-14-25.ap-southeast-2.compute.amazonaws.com:8000/plot



*~/data\_selected:* This endpoint provides data between selected dates. The date range can be specified by using '*year\_from'* and '*year\_to*' parameters in the 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 an in-depth discussion and analysis of the topic. Extensive research 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 content scraping, followed by data wrangling and visualization. The analytic results of the data visualization complied with the economic theory and principle, and major international events have further proved our data collection to be successful.

# Reference

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

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

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

4. Economy of New Zealand. (2022, October 27). In *Wikipedia*. https://en.wikipedia.org/wiki/Economy\_of\_New\_Zealand