DATA422
Project Report

EDUCATION IN NEW ZEALAND

Kim Van Nguyen – 68569443

Lianyin Liu – 17899438

Swapna Josmi Sam – 74281128

Ancy John – 52770710

Contents

1. Introduction

Right to education is one of the fundamental human rights, in many of the nations. The cultural aspects of a nation mainly depend on its quality of Education. Education is the key to empower a human being and to make him independent. Governments take responsibility of education for its human resources. This is a major step towards progress of the nation. The pathway to reach peace, human well-being and economic growth. are smoother when people are educated. It is the country's will to improve and spread education to everyone. In this 21st century of advanced technology, countries liaise and support each other, enabling easy access of education. Education is one of the issues that countries work together for the betterment. Hence, education is one of government's big concern.

The New Zealand education system entitles free education and free enrolment for every child between 5 and 19 years of age. The whole system is separated as 3 different levels. Primary, secondary and Tertiary levels. This project is an attempt to wrangling, analysing and visualizing of the dataset.

Major focus is on the wrangling part of the data. We collected 4 data-sets from 4 different sources and used R programming to do the wrangling on these data. We used R and Julia to join these data- sets and to visualize it.

Our final data model organizes elements of all these data-set to extract information regarding:

- Education Enrolment
- Employment vs unemployment
- Government expenditure
- School list in New Zealand

2. Data sources

After doing research on sources of education dataset, we decided to use these four sources that we believe they are true, transparent and reliable.

1. World Bank (https://databank.worldbank.org/)

   World Bank is an organization that is combined of five international institutions that aims for ending poverty and boost prosperity. It serves in non--profit sector, so its data is believed to be unbiased and non-driven by money. Further, World Bank is a well-known world institution that has been operating over many decades and has good reputation for its works. It relies on its data to work towards its goals. Not only itself but also many countries and other experts in the world also rely on this data. Therefore, the data should be correct and reliable. In this source, we used "Labour and Unemployment Rate" data.

2. United Nations (http://data.un.org/)

   United Nations is a non-for-profit international organization. It works toward non-financial goals including human rights, so it should have data about education around the world. It is made up of more than 100 countries in the world. Therefore, it has less chance of having manipulated data. It has a history of more than 70 years and is also well-known for its profession and good impact to the change of the world, so we trust its data. It has many offices around the world and large number of staffs, so it has the ability to collect large data in many countries. We obtained "Enrolment" data from this source.

3. UNESCO Institute for Statistics (UIS) (http://data.uis.unesco.org/)

   The UIS is an "official and trusted source of internationally comparable data on education, science, culture and communication" (UNESCO Institute of Statistics, 2019). The organization collects data from trustworthy statistics providers such as national statistical offices, line ministries and other statistical organizations. The data is used to produce reports by large organizations such as UN, World Bank, UNICEF, etc. to work towards development goals especially in the field of education. Additionally, many global indices cannot be calculated without UIS data such as Human Development Index and the World Competitiveness Index (UNESCO Institute of Statistics, 2019). This is where we obtained "Government Expenditure" data.

4. New Zealand government's data website (https://catalogue.data.govt.nz/). The website provides public API to access the data. **The data source which we were after had an API setup and we accessed the data via a web API**. Data.govt.nz is a New Zealand's All-of-Government service with

technical development led by Government Information Services and content production led by Stats NZ (About Data.govt.nz, 2019). It is a warehouse of data released by New Zealand Government organisations. The data is offered to the public for free. Public sector agencies such as Central Government, Local Government, Crown Research Institutes and the Tertiary sector also use this data source. This source was chosen for our project also because it provides more details about education circumstances in New Zealand which is suitable for our New Zealand schools analysis. Based on the reasons above, we think that this is the most reliable, true and informative New Zealand data source compared with other sources. We obtained New Zealand schools "School_list" from this source.

3. Target

Our first target is to see where New Zealand ranks compared with other countries in the world in regard to each category: government expenditure, enrolment, labour and unemployment rate. By knowing these, New Zealand government can know its status compared with the world to improve its policy in improving enrolment, labour and unemployment rate. For this target, we would need to clean up the data, arrange all datasets to be in the same structure (Countries in the first column and Years spread into columns).

Our second target is to create a model which is a combination of New Zealand data from all 4 datasets (Expenditure, Enrolment, Labour and Unemployment). This is a relational data which shows just New Zealand information so we can compare the change over time of the country's expenditure, enrolment, labour and unemployment. There will be plotting for visualisation. Based on the outcomes, we can conclude that whether one variable has influence on the other or not, so policy makers can consider this for future policy adjustments. We let variable "Year" in the first column as observation so observations for later years can be added. To achieve this target, we would need to have all datasets to be wrangled, then select New Zealand data and join.

For the third target, as we are focusing on New Zealand education, we want to know about the primary and secondary education, number of students enrolled and student enrolment based on origin and school details in New Zealand. The intention was to compare which region of NZ has a greater number of schools and number of students of each origin based on region. This means that we can compare the students count whether they are Maori, European, Asian or International students and reach to a conclusion that which students (Origin) acquiring education in NZ are more in number.

4. Difficulties

[what difficulties you had to overcome to wrangle the data sources into the target data model]

In wrangling process, there are many difficulties we have been through. First difficulty is that there are many useless variables which makes data messy. For example, the "Sources" columns showing the sources of data which is irrelevant to our analysis, and the years that are too old and inapplicable for our analysis. So, we have to delete the irrelevant variables that are not useful such as years before 2000. "Government Expenditure" dataset has many old years, so it is tricky to delete them correctly and not accidently delete needed columns.

The second difficulty is that every dataset is in different structures. The variables in a dataset were recorded as observations instead of variables. Two datasets have the variables in one column called "Series" which makes the country name and year repeated many times. The other dataset also has "Year" in one column which also makes the country name repeated. To tackle this problem, we have to spread the variables of "Series" column, choose one variable useful for our model then spread the "Year" column.

Third difficulty is that values were recorded in different formats. One dataset has quotation marks before and after value but there are no quotation marks in another dataset. So, we have to delete all quotation marks in values that have.

Fourth difficulty is that not all years are available which makes the output incomplete. To interpret the results, we would have to choose an interval of year to see the trend over a long time so we will not be influenced much by the missing years.

Plotting for the relational data (New Zealand joined data) is difficult because each column has different range of value. For example, the unit range of Labour Force is from 2,000,000 to 3,000,000 but Total Unemployment range is from 3 to 7. We also have different type of value such as Government Expenditure's value is different from number of students enrolled in primary education. However, our target is to see if the expenditure affects number of enrolments and other variables, so we still have to plot but in separate graphs.

Using API to extract data and to create a dataframe was little difficult. The resultant dataframe had columns with different number of rows which throwed error and was due to the null values in the data. Therefore, we handled the null values by giving if else statement. Thereby, we created a dataframe without any error. Another, issue faced was the API Key which had not required access to perform which was solved by enabling Geolocate and providing billing.

5. Techniques

Github was used to share our works and track changes to the code.

- In our project, we used 2 languages to do wrangling and plotting: R and Julia. Our CSV formatted files were downloaded from the data sources and imported into Jupyter notebook using *read_csv* function. We used *tidyverse* library for all the wrangling. Other libraries we also used are *readxl*, *visdat*, *ggplot2*, and *dplyr*.
- In wrangling process, first we choose the columns that are relevant and useful for our target by using *select* function. Then we used *rename* function for columns' names. We used *vis_miss()* function to see missing values in the data. We filled missing value with "no data" using *is.na* function. We *spread* the variables and again *select* the variables that are useful for us. We used *arrange* function to sort data as preference. Lastly, we drew boxplots and scatter plots with lines using *boxplot* and *ggplot* function respectively.
- For New Zealand schools' enrolment, we accessed the data via a web API from data.govt.nz. This data source had an API setup. A GET request is submitted within the URL with each parameter separated by an ampersand. We then parse the content returned from the server as text using the content function. Our required data is stored inside the 'records' list in the main list. Required data is fetched and provided as columns of a dataframe. Null values were handled by providing "Data unknown" to Regional_Council. Also removed the 'NA' values from Latitude and Longitude columns by using the *drop.na* function. We used *ggmap* package is used for spatial visualisation which aligns with *ggplot*. *'get_map'* queries Google maps server for the map of New Zealand using personal API key. One of our requirements was to plot the top five regions which has a greater number of schools. We have used the *group_by* function to group the school list based on Regional_Council. We took the mean of Latitude and Longitude and count of schools in each region to form a table using *summarise* function. The five points were plotted on NZ map using *ggplot.* Second requirement was to create a table with the counts of International Students, Maori Students, Asian student, European students and Total Students in each region of New Zealand and use the *mutate function* to find the proportion of these students from the total number of students. The proportion of these different students based on origin in each region was plotted using the *barplot* to find which type of students are more in each region of New Zealand.

6. Exploratory Analysis & Outcomes

- <u>Education Enrollment data</u>

```
[13]: Edu_enroll
```

A spec_tbl_df: 8628 × 7

Region/Country/Area	X2	Year	Series	Value	Footnotes	Source
<dbl>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<chr>
1	Total, all countries or areas	2005	Students enrolled in primary education (thousands)	678991.6100	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollement ratio - Primary (male)	104.9360	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollment ratio - Primary (female)	99.9214	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Students enrolled in secondary education (thousands)	509245.7600	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollment ratio - Secondary (male)	65.7318	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollment ratio - Secondary (female)	62.0691	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Students enrolled in tertiary education (thousands)	139950.8800	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollment ratio - Tertiary (male)	23.7177	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2005	Gross enrollment ratio - Tertiary (female)	24.8400	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2010	Students enrolled in primary education (thousands)	697048.8300	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.
1	Total, all countries or areas	2010	Gross enrollement ratio - Primary (male)	105.8406	NA	United Nations Educational, Scientific and Cultural Organization (UNESCO), Montreal, the UNESCO Institute for Statistics (UIS) statistics database, last accessed March 2019.

Figure 6.1: Figure shows the OriginalvEducation Enrollement data

Region co	Region	Year	Gross enr	Gross enr	Gross enr	Gross enr	Gross enr	Gross enr	Students	Students	Students enrolled in tertiary education (thousands)
1	Total, all c	2005	104.936	99.9214	62.0691	65.7318	24.84	23.7177	678991.6	509245.8	139950.9
1	Total, all c	2010	105.8406	102.9678	69.8178	72.3451	30.4946	28.393	697048.8	546102.4	182209.2
1	Total, all c	2014	103.1513	102.8368	75.821	76.8073	37.5289	33.7063	715016.7	580925.1	212931.5
1	Total, all c	2015	102.7414	102.6242	75.9324	76.9851	38.7492	34.7216	720228.5	583315.6	217459
1	Total, all c	2017	103.7417	104.4958	76.1996	76.9813	40.1622	35.7433	746284.9	590725.2	220704.2
4	Afghanist	2004	149.6565	65.002	6.1421	29.417	0.5213	1.8601	4430.14	594.306	27.648
4	Afghanist	2005	126.9444	74.5307	9.3116	28.6331	0	0	4318.82	651.453	0
4	Afghanist	2009	118.4504	78.7441	29.3865	60.9922	1.4252	6.0536	4945.63	1716.19	95.185
4	Afghanist	2010	120.5552	82.7089	34.296	68.6471	0	0	5279.33	2044.16	0
4	Afghanist	2014	125.6708	88.0897	38.5657	68.5591	3.5854	13.0284	6217.76	2602.73	262.874
4	Afghanist	2015	126.2261	87.2409	38.5661	68.412	0	0	6333.7	2698.82	0
4	Afghanist	2017	122.7469	84.1501	39.6317	69.1179	0	0	6358.78	2907.8	0
4	Afghanist	2018	0	0	0	0	0	14.5808	0	0	370.61
8	Albania	2005	101.5816	100.5708	76.3579	80.1788	27.2456	19.1252	237.975	407.403	63.257
8	Albania	2010	94.6319	92.9729	88.1385	89.0627	51.6	38.5663	224.781	355.871	122.326
8	Albania	2014	108.563	106.6985	92.5069	99.9304	86.2626	56.7249	195.72	333.291	173.819
8	Albania	2015	111.1933	108.1993	92.036	99.2919	81.6521	53.3497	188.371	315.079	160.527
8	Albania	2017	111.5243	108.6033	93.0932	98.9879	70.8026	44.4004	174.836	280.378	139.607
9	Oceania	2005	91.6886	89.6841	109.4575	112.7506	72.3974	58.4281	3141.61	3575.04	1642.73
9	Oceania	2010	100.4192	97.4064	108.1662	115.9096	84.5345	62.4603	3556.16	3776.48	2013.95

Figure 6.2: Figure shows the Wrangled Education Enrollement data

- Labour Unemployment data

```
[34]: data
```

A tibble: 1061 × 24

Country Name	Country Code	Series	Series Code	1999 [YR1999]	2000 [YR2000]	2001 [YR2001]	2002 [YR2002]	2003 [YR2003]	2004 [YR2004]	...	2009 [YR2009]	201
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	...	<chr>	
Afghanistan	AFG	Labor force, total	SL.TLF.TOTL.IN	6309513	6521151	6836893	7221944	7644720	8065343	...	9244457	
Afghanistan	AFG	Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	3.6059999465942401	3.5169999594568	3.4260001182556201	3.5499999523162802	3.4189999103546098	3.0869998931884801	...	2.4700000286102299	2.275000
Afghanistan	AFG	Unemployment, male (% of male labor force)	SL.UEM.TOTL.MA.ZS	3.8619999855591	3.7699999809265101	3.6760001182556201	3.7839999198913601	3.6489999294281001	3.2639999389648402	...	2.3350000381469700	1.960999
Afghanistan	AFG	Unemployment, female (% of female labor force)	SL.UEM.TOTL.FE.ZS	3.0750000476837198	2.9909999370575	2.9059998989105198	3.0639998912811302	2.9430000782012899	2.72300004959106	...	2.7469999790191699	2.920000
Albania	ALB	Labor force, total	SL.TLF.TOTL.IN	1356347	1351709	1338629	1338094	1319656	1301290	...	1226713	
Albania	ALB	Unemployment, total (% of total labor force)	SL.UEM.TOTL.ZS	18.5690002441406	17.7670001983643	17.4109992980957	17.5100002288818	17.4960002899169 99	17.2719993591308 98	...	13.6739997863769 53	14.08600
Albania	ALB	Unemployment, male (% of male labor force)	SL.UEM.TOTL.MA.ZS	18.2560005187988	17.4409999847412	17.0979995727539	17.2040004730225	17.2019996643066	16.9629993438720 99	...	12.2069997787475 99	12.7625
Albania	ALB	Unemployment, female (% of female labor force)	SL.UEM.TOTL.FE.ZS	19.0090007781982	18.2339992523193	17.8630008697510 01	17.9489994049072 02	17.9139995574950 99	17.7070007324219	...	15.7349996566772	15.881999

Figure 6.3: Figure shows the Original Labour Unemployment data

[47]: data

A tibble: 4220 × 5

Country Name	Country Code	Year	Labor force, total	Unemployment, total (% of total labor force)
<chr>	<chr>	<chr>	<dbl>	<dbl>
Afghanistan	AFG	1999	6309513	3.606
Afghanistan	AFG	2000	6521151	3.517
Afghanistan	AFG	2001	6836893	3.426
Afghanistan	AFG	2002	7221944	3.550
Afghanistan	AFG	2003	7644720	3.419
Afghanistan	AFG	2004	8065343	3.087
Afghanistan	AFG	2005	8459571	2.942
Afghanistan	AFG	2006	8677865	2.825
Afghanistan	AFG	2007	8859372	2.128
Afghanistan	AFG	2008	9034635	2.494
Afghanistan	AFG	2009	9244457	2.470
Afghanistan	AFG	2010	9516457	2.275
Afghanistan	AFG	2011	9940084	1.984
Afghanistan	AFG	2012	10446498	1.692
Afghanistan	AFG	2013	11014922	1.725
Afghanistan	AFG	2014	11611982	1.735
Afghanistan	AFG	2015	12213764	1.679

Figure 6.4: Figure shows the Wrangled Labour Unemployment data

- **School List Data:** One of our requirements was to plot the top five regions which has a greater number of schools. We have used the *group_by* function to group the school list based on Regional_Council. We took the mean of Latitude and Longitude and count of schools in each region to form a table using *summarise* function. The five points were plotted on NZ map using *ggplot.*

	Regional_Council	school_count	Latitude	Longitude
1	Auckland Region	550	-36.88500	174.7835
2	Waikato Region	309	-37.80613	175.4388
3	Canterbury Region	291	-43.63906	172.2818
4	Wellington Region	247	-41.15553	174.9921
5	Manawatu-Wanganui Region	197	-40.06906	175.4507

Table6.1: Table showing the top five regions in New Zealand with a greater number of schools.

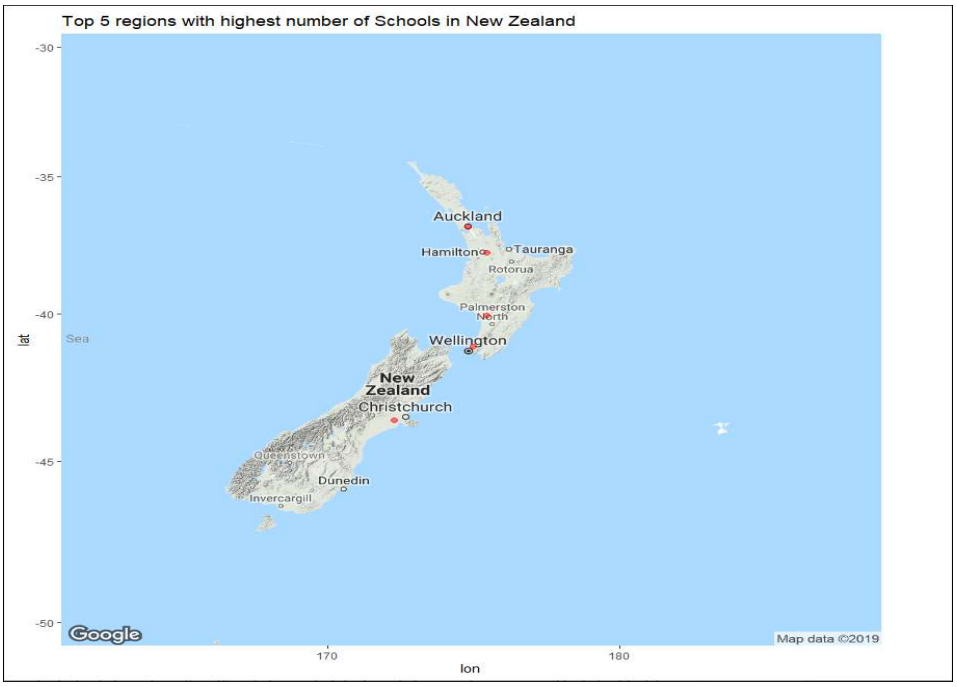


Figure 6.5: Figure showing the top five regions in New Zealand with a greater number of schools.

It is clear from the table and graph that Auckland has a greater number of school when compared to other regions which tells that Auckland provides better education opportunity to students.

- Second requirement was to create a table with the counts of International Students, Maori Students, Asian student, European students and Total Students in each region of New Zealand and use the *mutate function* to find the proportion of these students from the total number of students. The proportion of these different students based on origin in each region was plotted using the *barplot* to find which type of students are more in each region of New Zealand.

	Regional_Council	International_students_count	Maori_students_count	European_students_count	Asian_students_count	Total_Students	Proportion_International
1	Auckland Region	7674	42268	98878	66906	283116	0.0271054974
2	Canterbury Region	1389	13413	65501	10939	98472	0.0141055325
3	Wellington Region	1016	17001	44422	10358	83828	0.0121200554
4	Bay of Plenty Region	837	22864	25622	3635	56082	0.0149245747
5	Waikato Region	778	27868	40905	7215	82644	0.0094138715
6	Otago Region	517	4480	24108	2034	33693	0.0153444336
7	Manawatu-Wanganui Region	502	13735	21565	2592	41398	0.0121261897
8	Hawke's Bay Region	244	11991	15289	1410	31005	0.0078696984
9	Nelson Region	233	1349	6048	794	8946	0.0260451598
10	Northland Region	183	16265	12467	999	31099	0.0058844336
11	Southland Region	127	3701	11383	1002	17084	0.0074338562
12	Tasman Region	124	1085	6565	280	8339	0.0148698885
13	Taranaki Region	93	6124	13251	960	21359	0.0043541364
14	Marlborough Region	46	1450	4621	322	6936	0.0066320646
15	Gisborne Region	10	6302	2571	208	9473	0.0010556318
16	West Coast Region	3	827	3183	159	4363	0.0006876003
17	Area Outside Region	0	50	20	0	71	0.0000000000

Table6.2: Table showing the student distribution based on origins in different regions of New Zealand.

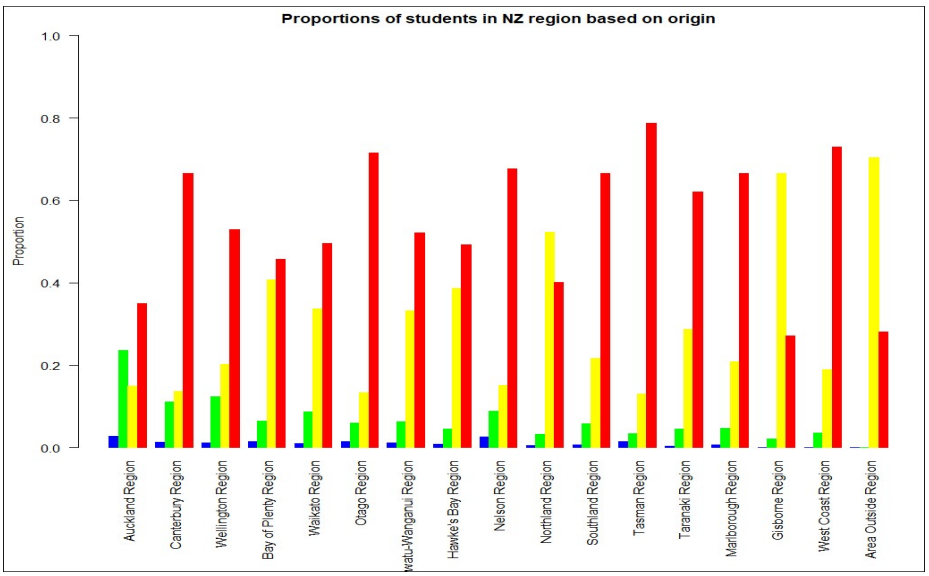


Figure 6.6: Figure showing the student proportion in each region based on their origin.

From the graph it is clear that European student have a greater proportion when compared to the others in most of the regions.

The final data model

A spec_tbl_df: 5 × 5

Year	Primary_enrollment	Government_Expenditure	LabourForce	TotalUnemployment
<int>	<dbl>	<chr>	<dbl>	<dbl>
2010	348.492	15.68851	2324565	6.557
2005	352.845	16.25796	2178594	3.807
2014	360.206	16.27686	2464873	5.752
2017	384.251	no data	2694635	4.702
2015	368.306	16.37508	2513183	5.365

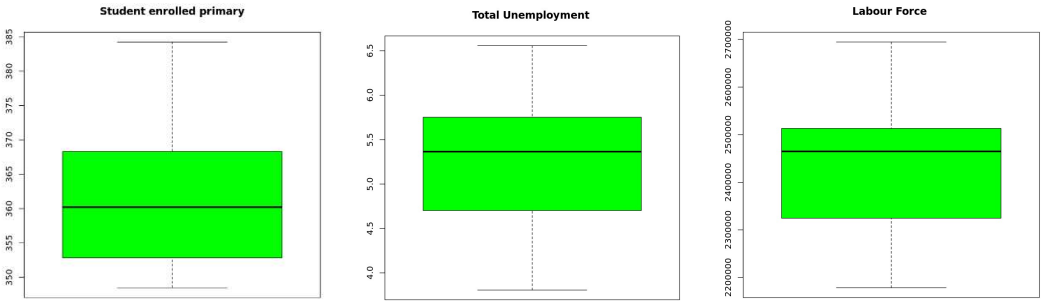


Figure 6.7 box plot of the final joint data

The first plot shows that the median of student enrolled in primary median is 360, the minimal is under 350 and maximal is over 380.

The second plot shows that the median of unemployment in New Zealand is between 5.0 and 5.5, the minimal is under 4.0 and maximal is over 6.5

The third plot shows that the median of labour force between 2400000 and 2500000, the minimal is under 2200000 and maximal is roughly 2700000.

There is no outlier in the boxplots.

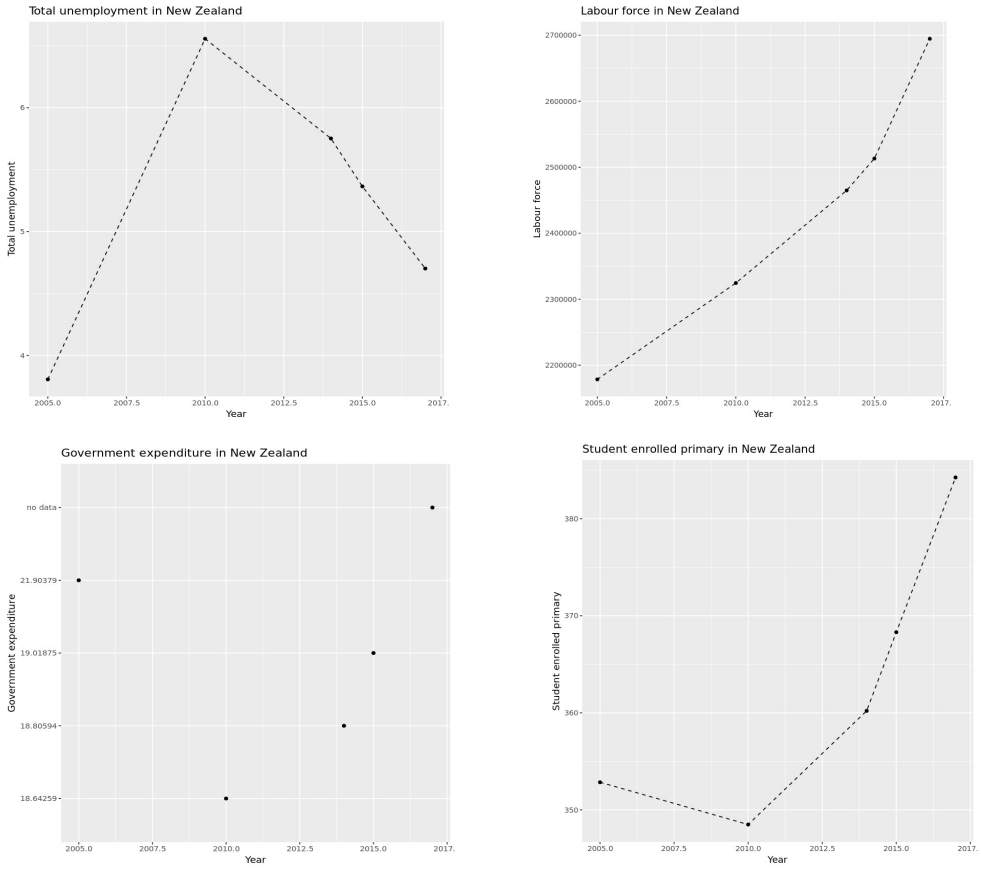


Figure 6.8time series trend of the final joint data

From the time series plot we can see that the unemployment rate in New Zealand rises sharply from 2005 to 2010 and goes down until 2017.

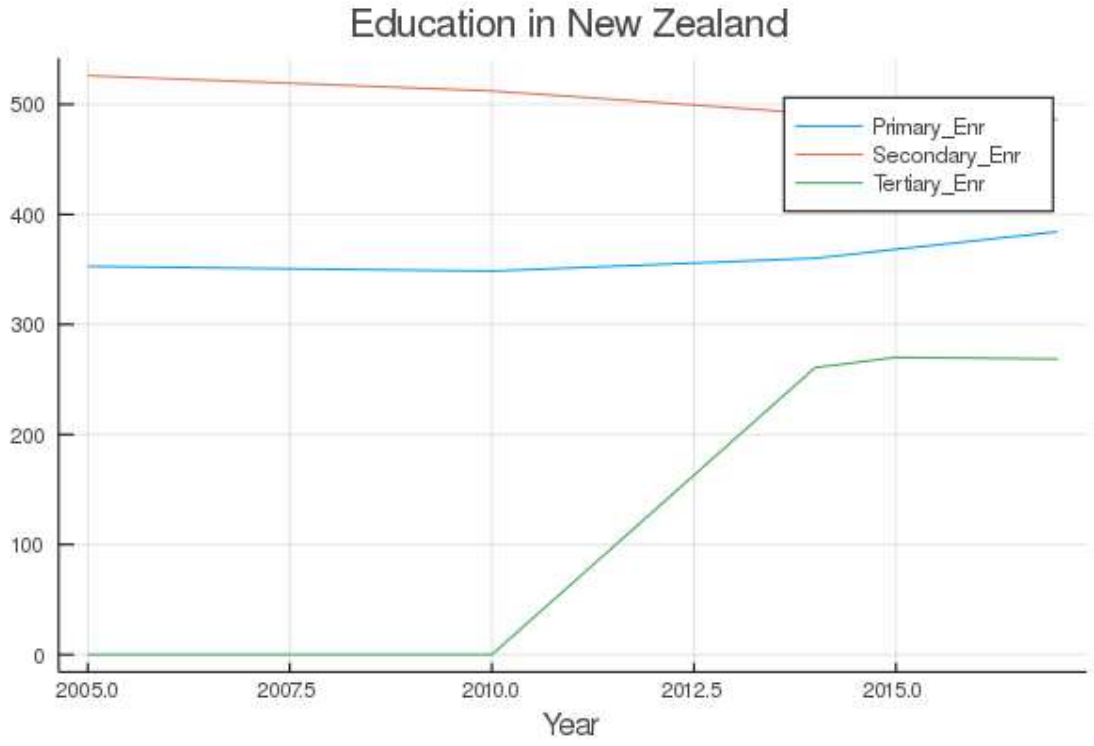
The labour force in New Zealand increases from 2005 to 2017 sharper and sharper.

The student enrollment in primary in New Zealand goes down from 2005 to 2010, then increases rapidly.

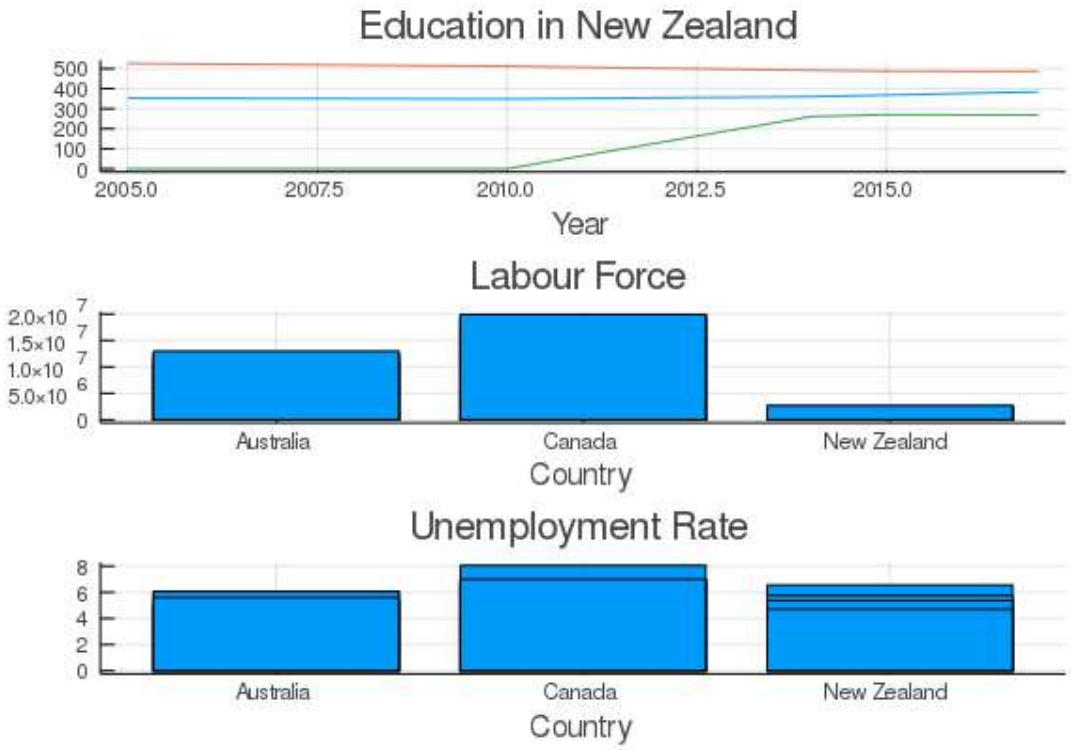
The Government expenditure in primary in New Zealand goes down sharply from 2005 to 2010, then increases rapidly

Julia part

We used Julia programming language in some part, for joining and plotting of the data. We have joined the education enrolment and labour unemployment data frames for this.



The first plot is the education enrolment in three different levels, primary, secondary and tertiary. We have maximum enrolment at the secondary level. Enrolment in the primary and tertiary levels are showing improvement over the years. One difficulty found in this plot is the missing information. Since there is no data until 2010 for the tertiary level the value is assigned to zero.



In the combined plot above, we did a comparison of Labour force and the un-employment rate of New Zealand with that of Australia and Canada. Labour force is the least in New Zealand compared to the others. Yet, the unemployment rate in New Zealand is higher than that of Australia.

7. Conclusions

The group project helped us understand how to deal with the wrangling of a messy data, and its further analysis &visualization. Now we have answers to some interesting questions and are able to do follow-up and deep dive analysis.

8. References

Data.govt.nz. 2019. *About.* Retrieved from https://www.data.govt.nz/about/about-data-govt-nz/

Reddy, C. n.d. *Why Education is Important? Top 13 Reasons.* Retrieved from https://content.wisestep.com/education-important-top-reasons/

UNESCO Institute of Statistics. 2019. *About Us | UNESCO UIS.* Retrieved from http://uis.unesco.org/en/about-us

UNESCO Institute of Statistics. 2019. *Data to Transform Lives.* Retrieved from http://uis.unesco.org/en/data-transform-lives

World Bank. 2019. *About the World Bank.* Retrieved from https://www.worldbank.org/en/about

https://www.rdocumentation.org/packages/ggmap/versions/3.0.0/topics/get_map

https://catalogue.data.govt.nz/api/3/action/datastore_search?resource_id=bdfe0e4c-1554-4701-a8fe-ba1c8e0cc2ce&limit=2556