COMP20008 – Assignment Phase 3
Relationship between Crime and the Number of Liquor Stores in Victoria
William Liandri – 728710

2. Domain

The domains for this project are communities in terms of crime and urban planning in terms of liquor stores. The main focus of this report is on the relation between certain type of crimes and the number of liquor stores in LGAs in Victoria.

3. Questions

This following project will try to answer these following questions:

1. Where are the top 5 LGAs in Victoria with the highest crime rates?
2. What type of crimes that are mostly happened in Victoria?
3. Is there any relationship between the number of liquor stores and the certain type of crimes happening in Victoria? If so, what type of crime is that?

The project will examine the crime rates for 5 types of crimes in Victoria in 2015 and compare them with the number of liquor stores in Victoria in the same year in 79 LGAs in Victoria. If this project find the positive correlation between the number of liquor store and the crime rates for certain type of crimes, the Government can use this information to reduce crime rates in Victoria by reducing the number of liquor stores, especially in the 5 LGAs in Victoria that have the highest crime rates.

4. Datasets

1. "*Crime_location.csv"* (**13804 rows** and **7 columns**)
   This data recorded the number of crimes in each LGA in Victoria between 2012 and 2016 by Crime Statistics Agency in Victoria. This project used only **Table 3** from this data.
   Link:
   https://www.crimestatistics.vic.gov.au/sites/default/files/embridge_cache/emshare/original/public/users/201703/8d/6d93514b8/Crime%20by%20location%20-%20year%20ending%2031%20December%202016.xlsx

2. *"Liquor_Metro_2015.csv"* (1)(**13686 rows** and **12 columns**) and *"Liquor_Region_2015.csv"* (2) (**7275 rows** and **12 columns**)
   This data contains about the list of liquor stores in metropolitan (1) and regional (2) Victoria in 2015. This data was recorded by Victorian Commission for Gambling and Liquor Regulation.
   Link:
   (1) https://www.vcglr.vic.gov.au/sites/default/files/Liquor%20licence%20by%20location%20-%20metro%20-%20excel%20-%20Dec%202015.xls
   (2) https://www.vcglr.vic.gov.au/sites/default/files/Liquor%20licence%20by%20location%20-%20regional%20-%20excel%20-%20Dec%202015.xls

3. *"Population_by_LGA_2015.csv"* (**570 rows** and **23 columns**)
   This data contain about the number of population based on gender and the total population by age and LGA in Australia in 2015. This project only used **Table 3** from this data.
   Link:
   http://www.abs.gov.au/ausstats/subscriber.nsf/log?openagent&32350ds0010_lga_2015.xls&3235.0&Data%20Cubes&21E59889DF4AFE1ECA25801200168299&0&2015&18.08.2016&Latest

5. Pre-Processing

Since all the data in xls format, before beginning data wrangling, all of them needed to be converted to csv file. This process was done using Microsoft Excel by doing copy and paste all the data and then save the file as csv. For the liquor dataset before saving the data as csv file, all commas are replaced with hyphen to avoid error when the data was read using Pandas in Python.

Data reduction was also undertaken during the pre-processing using pandas.DataFrame.ix and index_col function when reading the csv file using Pandas in Python to ignore unnecessary columns.

Dataset	Used Schema
Crime_location.csv	"Jan – Dec Reference Period", "Local Government Area", "CSA Offence Division", "Offence Count"
Liquor_Metro_2015.csv	"Council Name"
Liquor_Region_2015.csv	"Council Name"
Population_by_LGA_2015.csv	"S/T name", "LGA name", "Total person"

Some visualisations were also done during the pre-processing to analyse the data, detect some outliers, noisy data, and missing values on the data, normalised the data and find the correlation between crime in general and the number of liquor stores. The data that has been analysed during this pre-processing were the number of crimes and the number of liquor store for each LGA in Victoria. However, there was a problem raised when doing this part as comparing the number of crimes between LGAs in Victoria was arduous since each LGA has its own number of populations. Thus, normalised the data was needed by calculating the number of crimes every 100,000 population on each LGA using this formula:

$$\textbf{Number of crimes per 100,000 people} = \frac{\textbf{Total number of crimes in that LGA}}{\textbf{Total number of populations in that LGA}} * \textbf{100,000 people}$$

During this pre-processing, there were some missing values that were found on the data. From the *"Crime_location.csv"*, for type of crime *"F Other offence"*, there were 5 missing values, which came from LGA named *"Hepburn"*, *"West Wimmera"*, *"Hindmarsh"*, and *"Indigo"*. Thus, these missing values were replaced with 0. From the *"Population_by_LGA_2015.csv"*, there was also a missing value on LGA named *"Queenscliffe"*, so it was replaced with the average number of population. Apart from this, the other dataset did not have any noisy data nor missing values.

Detecting some outliers were done by creating 4 boxplots for the crime and the liquor data (2 boxplots before removing the outliers and 2 boxplots after removing the outliers). However, after finding there are 16 LGA names that were the outliers, it was decided not to remove them since it could predict unusual trends in the data, which would give more interesting results. Additionally, the outliers also contained LGAs with the most number of crime rates where the government should be really concern about.

6. Integration

Before doing integration process, finding the similarities in all datasets is necessary. The integration was done based on the LGA name. Since both crime and population dataset already had a LGA name column in their dataset, it was not hard to integrate them. The problem was that the liquor data does not have LGA name column, however, from the *"Council Name"* column in this dataset, the key word was found to integrate this data. One of the examples of the key word was that from the crime and population data, some of the data has LGA named *"Hume"* while some of the liquor data has Council named *"Hume City Council"*. From those data, it could be seen that the key word was *"Hume"* since it could be found in the LGA name and the Council name. Thus, Python code was written to find that key word, which will be used later to store the data.

The integration process was done by using Python defaultdict library to store the number of crimes, the number of crimes for each type of crimes, the number of population, and the number of liquor stores in each LGA. These dictionaries used the LGA name as the key for the dictionary and the value is an integer based on the what they recorded.

7. Results

This project had analysed the data both with and without outliers. Since it had decided to keep the outliers, the data without outliers will not be discuss in this report.

7.1. Top 5 LGAs in Victoria with the Highest Crime Rates

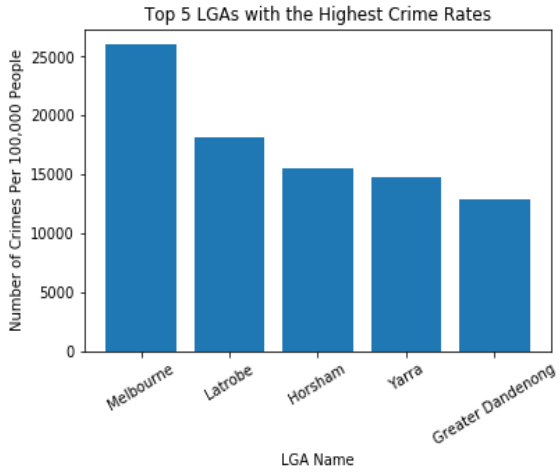

Figure 1. Bar chart of top 5 LGAs with the highest crime rates

Amongst 79 LGAs that were being analysed on this project, there were top 5 LGAs that have the highest crime rates where the Australian Government have to really concern about, those LGAs are Melbourne, Latrobe, Horsham, Yarra and Greater Dandenong. As it can be seen from the **Figure 1**, the highest crime rates was in **Melbourne** with around **26,000** crimes per 100,000 people and this was the only LGA in Victoria that has the crime rates more than 25,000 crimes per 100,000 people since the other LGAs have crime rates below 20,000 crimes per 100,000 people. Given this data, it can draw Government's attention to reduce the number of crimes, especially in Melbourne.

7.2. Distribution of Liquor Stores in Victoria

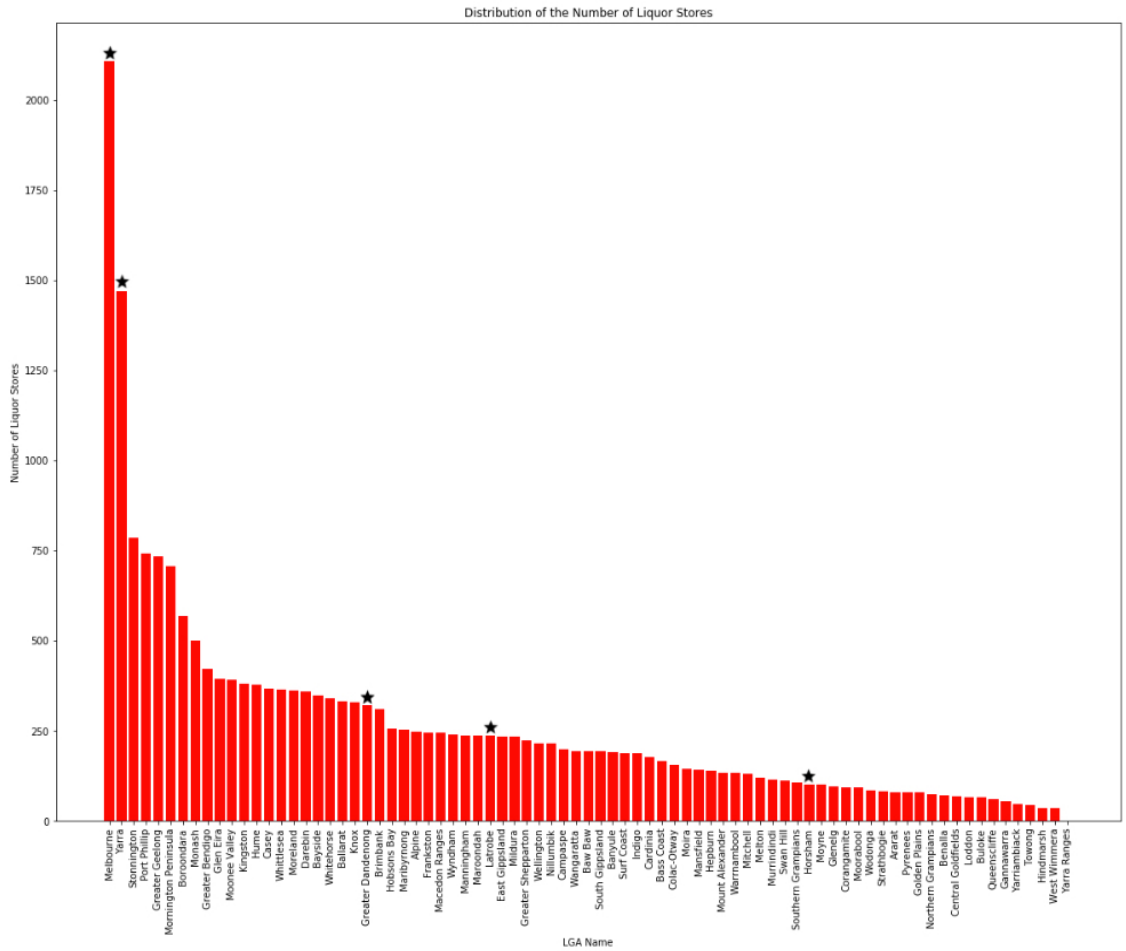

Figure 2. Distribution of the number of liquor stores in 79 LGAs in Victoria

The starred bars are the number of liquor store in 5 LGAs with the highest crime rates. As we can see from the **Figure 2** that **Melbourne** and **Yarra** have the highest number of liquor store with respectively **2,108** and **1,469** number of liquor stores. On the other hand, the other 3 LGAs with the

highest crimes, which are Greater Dandenong, Latrobe, and Horsham lie on the midway in the distribution with respectively **321**, **236**, and **101** number of liquor stores.
Since Melbourne has the highest crime rates and the number of liquor stores, this number of liquor stores can be the main factor that cause crimes in this LGA. However, this result has a limitation since it does not consider another factors that might cause crime as happened in Horsham when it has quite low number of liquor stores, but high crime rates.

7.3 Correlation between Certain Type of Crimes and the Number of Liquor Stores

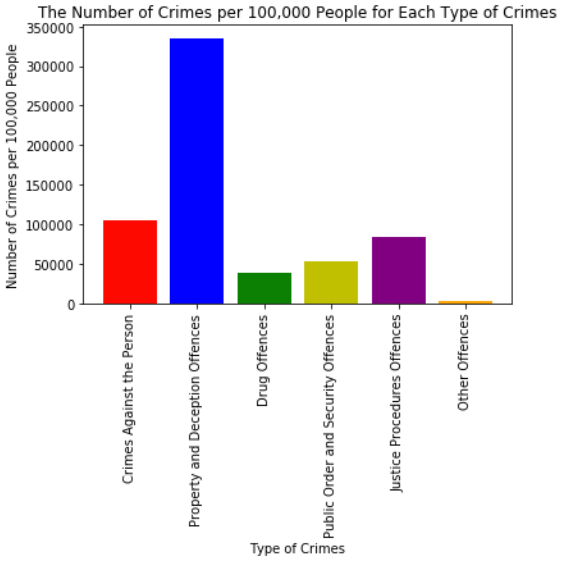

Figure 3 shows that the type of crime that has the highest crime rates that happened in Victoria in 2015 is "**Property and Deception Offences"**, which was **335,509** crimes per 100,00 people in Victoria. Conversely, the lowest crime rates was "**Other Offences"** with **3,017** crimes per 100,000 people in Victoria. The other 3 type of crimes lie between 40,000 and 100,000 crimes per 100,000 people in Victoria. As it can be seen that there is high difference between the crime rate for propery and deception offences with the other type of crimes.

Figure 3. Number of crimes for each type of crimes

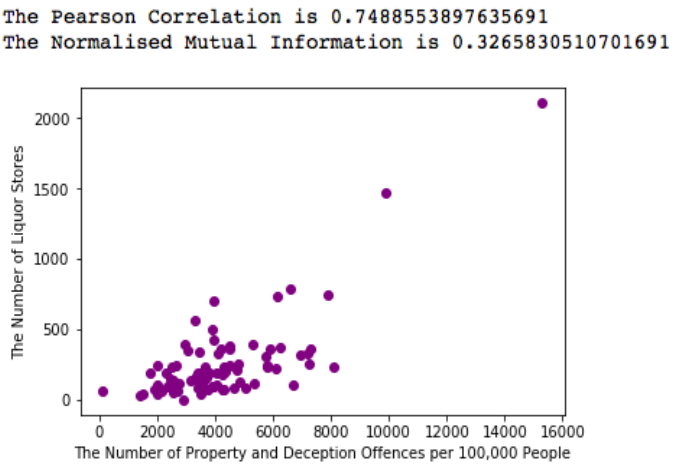

Figure 4. Scatter plot of property and deception offences

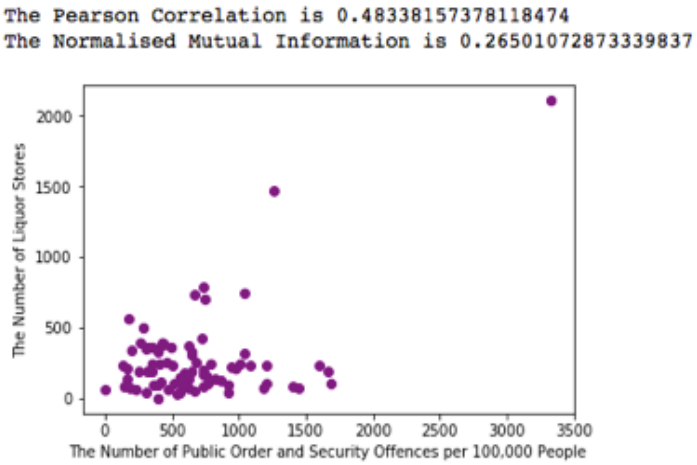

Figure 5. Scatter plot for public order and security offences

Amongst 6 type of crimes that were being analysed on this project by creating scatter plots, calculating Pearson correlation and Normalised Mutual Information (NMI) for each type of crimes, the property and deception offences and public order and security offences were two type of crimes that have the highest correlation with the number of liquor store. As shown on **Figure 4** and **Figure 5** that the pearson correlation for **"Property and Deception Offences"** is **0.75**, which indicates there is strong relationship whereas the correlation for **"Public Order and Security Offences"** is **0.48**, which converts to moderate relationship (Cohen, 1988). Since there is also non-linear relation in the data and the NMI is above 0.25, therefore there is high possibility that there is correlation between the number of crimes and the number of liquor stores.

Even though it can find the correlation between the certain type of crimes and the number of liquor stores, it does not mean that reducing the number of liquor store will always reduce the crime rates in

a LGA since the causality is not covered in this project. For example, this project does not cover the liquor consumption as a result of high number of liquor stores and also liquors that are bought and consumed in different places. Furthermore, calculating Pearson correlation and NMI also have some limitations since Pearson correlation cannot be used for non-linear scatter plot whereas NMI can give different result when different bin is used.

Value

The dataset that were used on this project were useless on its own. By looking at the raw data, it only gives information about the list of crimes that happened between 2012 and 2015 and also the list of complete address of liquor stores in Victoria. Since this project involved some processes amongst those dataset through the pre-processing, integration, and analysis or visualisation processes, those data could give some values. The pre-processing was important since it could help to process the data in Pandas by converting the data to the csv file. For the integration, it helped to link between the dataset to give information about the number of crimes, liquor stores, and population in each LGA. Nevertheless, the analysis or visualisation was also important since it helped to seek answer for the proposed question by creating scatter plots, table, bar charts and calculating Pearson's Correlation and Normalised Mutual Information (NMI).

Challenges and Reflections

Finding suitable topic and datasets were the most challenging steps on this project. I had changed the topic for 3 times before deciding to use this topic. Some researches were done to ensure that the chosen topic would give valuable, reliable, and innovative information.

Normalising the data was somewhat important to give more reliable information. However, it caused to change the datasets and the codes. It used to use the datasets in 2016 since this is the latest data, it can give more relevant information. However, the problem raised when normalising the data was going to be done and the population data in 2016 could be found after doing some researchers for few days. Therefore, the dataset in 2015 was preferred because it had population data, which was used to normalised the data. Thus, some codes were also changed and there were some new codes to be written to suit the new dataset.

Question Resolution

There are 4 factors that have been discussed on this project, which help to answer the proposed question. Firstly, the 5 LGAs with the highest crime rates were Melbourne, Latrobe, Horsham, Yarra, and Greater Dandenong. Secondly, the highest number of liquor stores were in Melbourne. Furthermore, the highest crime rate was in the category "Property and Deception Offences", which also have a strong correlation with the number of liquor stores. From the results of this project, it can be concluded that the government can reduce the number of crimes, specifically crime with category **"Property and Deception Offences",** by reducing the number of liquor stores, especially in those 5 LGAs. Thus, it can benefit the community as a whole as the safety increases.

Code

There were around 800 lines of code in total that were written from scratch on this project. Around 400 lines of code were used during the pre-processing and around 400 lines of code to visualise the results of this project. The code to read the data from the csv and slice the data were cited from week 2 tutorial while plotting charts and plots were cited from week 5 tutorial with some modifications to suit the purpose.

This project used some libraries in Python, such as pandas, matplotlib, numpy, collections and math. The pandas library was used to read the csv file and slice the data. The matplotlib and numpy were used to help creating charts and plots on this project. The collections library was used to use defaultdict in Python to store the crime, liquor store, and population data. The math library was used to use log function to calculate Normalised Mutual Information (NMI).

Bibliography

Cohen, J., 1988. Statistical power analysis for the behavioural sciences. Hillside. NJ: *Lawrence Earlbaum Associates*