Capstone Project- The Battle of Neighborhoods

Greater Sydney Area, NSW

By: Riyaj Basukala

1 Introduction/Background

Located on Australia's south-east coast in the state of NSW, Sydney is known as one of the most liveable cities in the world with breathtaking landscapes, beautiful beaches, iconic landmarks, and vibrant culture with multi-diversity. Sydney is the destination of choice for international corporations, business leaders, migrants, tourists and students.

Sydney is the largest and most populous city in Australia and the state capital of New South Wales. Inhabitants of Sydney are also called 'Sydneysiders', comprising of cosmopolitan and international population of people from numerous places around the world. State of NSW has about 128 local government areas and Sydney is made up of 35 local government areas which is further broken down into 658 suburbs [1].

Sydney is home to more than 5.2 million people (as of 2018). The population is growing significantly and it is forecasted to reach about 8 million by 2050.

2 Business Problem and Target Audience

Due to the massive popularity of Sydney as a top destination to live in and its growing population, it is obvious that more residential properties, shopping centers, hospitals, parks, hotels, bars etc. would be needed to be built to meet the demand and to aid this development one would need a wholistic view of different areas to make critical decisions.

This project will look into couple of most populous local government areas in Sydney, NSW and its suburbs to analyze what is the current situation at various neighborhoods in terms of population, crime rates, residents' income and types of venues/facilities available. Project will focus on characterizing similarities and differences between the areas of study.

The outcome of the project is anticipated to be quite useful to local/state government, town planner, property developer, potential startup business owner or even prospective home buyers.

3 Data Description

- List of Local Government Area (LGA) of NSW: This is the list of LGAs in NSW available online. Since geojson file is available only for the whole state; we use this list for plotting the map area of the whole state. The subset of this list will form the scope of the project (focusing on Greater Sydney Area).

- List of State suburbs from Australian Bureau of Statistic: This is the neighborhood data mapped to LGA. It will be used to query Foursquare API to get venues and categories.

- Population data from Australian Bureau of Statistics: This has the absolute 2018 census data with population density per square km. It will be used to identify most populous area of Sydney.

- Income Distribution from Australian Bureau of Statistics: This data based on LGA has mean, median yearly income and also degree of inequality in terms of income share of top 10% earners.

- Crime Data from Australian Bureau of Statistics: This is LGA wise crime data with different types of crime and crime rates per 100,000 population over a period of one year from Oct 2018 to Sep 2019. This, along with income data will be used to profile different LGAs of our concern.

- Locality venues and categories from Foursquare API: This will be used to assist in clustering different neighborhoods based on frequency of occurrence of certain venue categories.

4 Methodology

As mentioned in the above section, this project aimed to find out similarities and differences between different areas of interest in Greater Sydney area in terms of venues, population, income and criminal data. To carry out this project after defining the objective, typical methods of data science were used, such as data collection and preparation, data understanding. Some exploratory data analysis was performed to determine the scope of project and also to understand data. Data visualization along with machine learning algorithm was used to achieve the results.

4.1.1 Data Collection/Preparation/Wrangling

Open data from Australian Bureau of Statistics was used to get the data of Local Government Areas (LGAs) in NSW [2], state suburbs data [3], crime [5], income [6] and population [7] data. Web data scraping using BeautifulSoup package was also performed on wikipedia website data of LGAs in greater Sydney area [4].

To prepare the data for analysis, fair amount of data cleaning/wrangling was done. Two separate data tables of suburbs and LGA were combined based on the key of MB_code_2016 (Mesh block). Every single suburb has its own MB_code. Every LGA has a list of MB_codes belonging to it. With this information, the suburbs were mapped to their corresponding LGAs. For the preparation of data both python pandas and Microsoft excel were used.

Another important step in preparation of data was retrieving coordinates of LGAs and suburbs (neighborhoods). Geocoding package was used for this purpose.

Crime data had many types of *Offence type* and to simplify, they were grouped into 5 main categories.

Apart from the “Others” category, rest of the crime types were considered for computing Total Crime Rates. The mapping is shown in below Table 1.

Offence type	Mapped to Category
Murder *	Murder_AttemptedMurder
Attempted murder	Murder_AttemptedMurder
Murder accessory, conspiracy	Murder_AttemptedMurder
Manslaughter *	Murder_AttemptedMurder
Domestic violence related assault	Violence_Assault
Non-domestic violence related assault	Violence_Assault
Assault Police	Violence_Assault
Sexual assault	Violence_Assault
Indecent assault, act of indecency and other sexual offences	Violence_Assault
Abduction and kidnapping	Others
Robbery without a weapon	Robbery_Stealing
Robbery with a firearm	Robbery_Stealing
Robbery with a weapon not a firearm	Robbery_Stealing
Blackmail and extortion	Others
Intimidation, stalking and harassment	Others
Other offences against the person	Others
Break and enter dwelling	BreakingandEntering
Break and enter non-dwelling	BreakingandEntering
Receiving or handling stolen goods	Others
Motor vehicle theft	Robbery_Stealing
Steal from motor vehicle	Robbery_Stealing
Steal from retail store	Robbery_Stealing
Steal from dwelling	Robbery_Stealing
Steal from person	Robbery_Stealing
Stock theft	Robbery_Stealing
Fraud	Others
Other theft	Robbery_Stealing
Arson	Others
Malicious damage to property	Others
Possession and/or use of cocaine	Others
Possession and/or use of narcotics	Others
Possession and/or use of cannabis	Others
Possession and/or use of amphetamines	Others

Possession and/or use of ecstasy	Others
Possession and/or use of other drugs	Others
Dealing, trafficking in cocaine	Others
Dealing, trafficking in narcotics	Others
Dealing, trafficking in cannabis	Others
Dealing, trafficking in amphetamines	Others
Dealing, trafficking in ecstasy	Others
Dealing, trafficking in other drugs	Others
Cultivating cannabis	Others
Manufacture drug	Others
Importing drugs	Others
Other drug offences	Others
Prohibited and regulated weapons offences	Others
Trespass	Others
Offensive conduct	Others
Offensive language	Others
Criminal intent	Others
Betting and gaming offences	Others
Liquor offences	Others
Pornography offences	Others
Prostitution offences	Others
Escape custody	Others
Breach Apprehended Violence Order	Violence_Assault
Breach bail conditions	Others
Fail to appear	Others
Resist or hinder officer	Others
Other offences against justice procedures	Others
Transport regulatory offences	Others
Other offences	Others

Table 1: Summary mapping of Crime type to a Category

4.1.2 Exploratory Data Analysis

To determine the scope of this project, some exploratory data analysis was done. Bar graphs of top 10 LGAs with highest population (figure1) was plotted along with their respective population density (figure2)

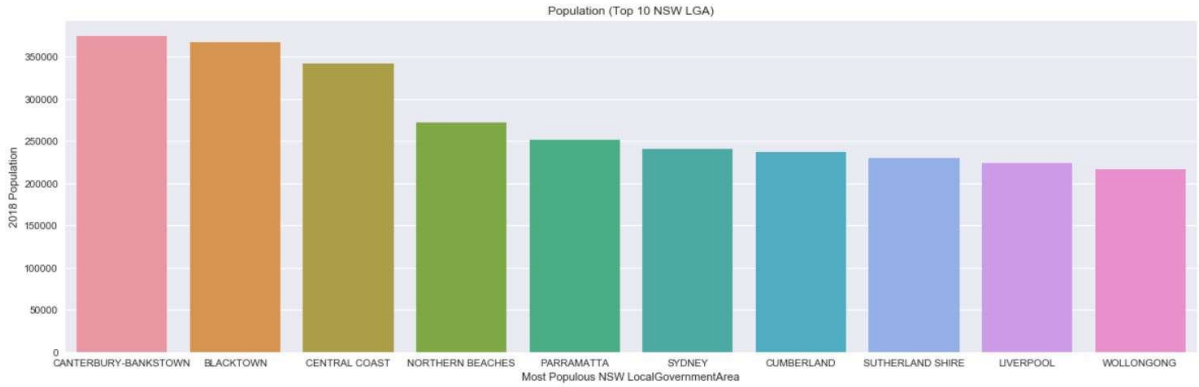

Figure 1: Top 10 LGA population

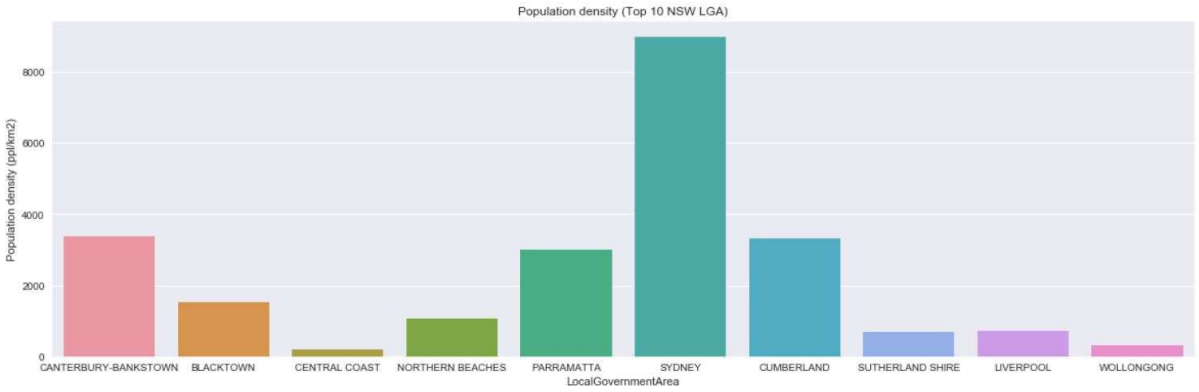

Figure 2: LGA population density

By looking at the figures, there were some LGAs with high population but very low density (per km2). As it was aimed to select the most populous LGAs for the project scope, Sydney, Cumberland, Parramatta and Canterbury-Bankstown were selected to be considered for this project, as they were most densely populated among the top 10 LGAs with highest population.

4.1.3 Inferential Statistical Testing

Regression testing was performed to check if the crime rates in NSW were correlated with any other parameters like population density, mean income, income inequality in the LGAs of NSW. For this, a table in form of pandas dataframe was formed with list of all LGAs in NSW, with corresponding feature parameters of income and population collated from the respective data.

4.1.4 Foursquare location data for Machine Learning

To analyze similarities/differences between neighborhoods/suburbs, Foursquare API is used to query points of interests or venues in the vicinity of those neighborhoods. Based on the frequency of occurrence of various venue types we can characterize a neighborhood. Unsupervised method with no

target variable is used to do partitioning or grouping of various neighborhoods to see if they formed natural groups. k-Means Clustering is a method of unsupervised segmentation that is used in this project.

k-Means Clustering is one of the simplest and popular clustering model used to quickly find insights from unlabeled data. Here we will use this to find out in a certain LGA of our interest what type of clusters of neighborhoods are found. User needs to input the number of clusters (k) for this model. Elbow method of determining best k is used.

After the clustering is applied, data is visualized by using python Folium library where NSW state map along with geojson file of LGAs is used. Folium is also used to generate choropleth maps in order to color code different regions based on the scale/value of certain parameter.

5 Results

5.1 Inferential Statistical Testing

Here, only the LGAs with population density of at least 500 people/km2 were selected mainly due to skewed crime data in LGAs with very low population density. In case of checking any correlation of crime rates existing with LGA population density, mean LGA income, it was found that crime rates showed slight correlation with mean LGA income, with better off LGAs in terms of income having tendency of lower crime rates. However, it has to be mentioned that there were not much data points available for high income LGAs. Most of the LGAs were clustered around $50k to $80k mark with lot of variations in data points.

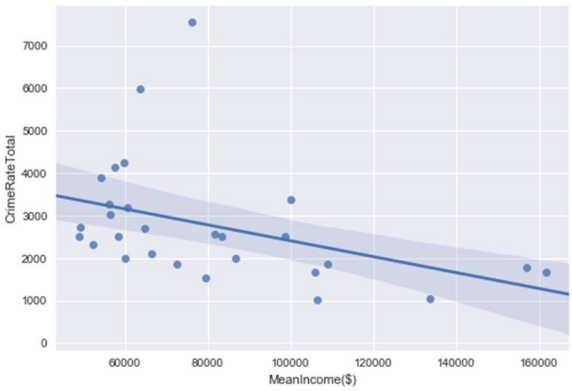

Figure 3: Regplot of CrimeRate vs Mean Income

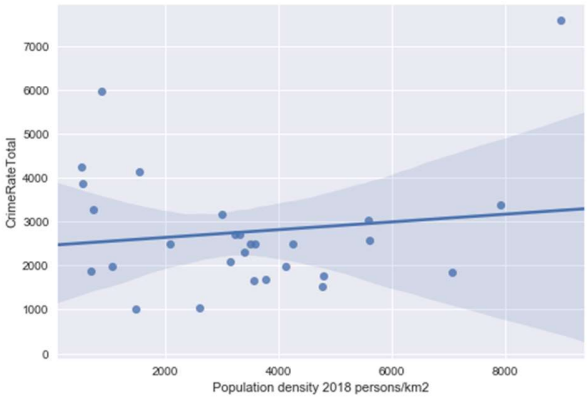

Figure 4: Regplot of CrimeRate vs Population density

In terms of population density, here also very little positive correlation was seen. This implied that apart from population density there will be multitude of other factors that can drive the crime rates in certain localities.

5.2 Comparison of 4 LGAs (most populous)

Figure 5: Population density with 4 LGAs under study

Above figure shows the 4 LGAs of our concern, Sydney, Parramatta, Cumberland and Canterbury-Bankstown, in choropleth map with LGA wise population density. We can see that Sydney has much higher population density than the other 3 LGAs.

Income wise Sydney and Parramatta fall into same range ($57k – $78k) while the other two LGAs have lower mean income.

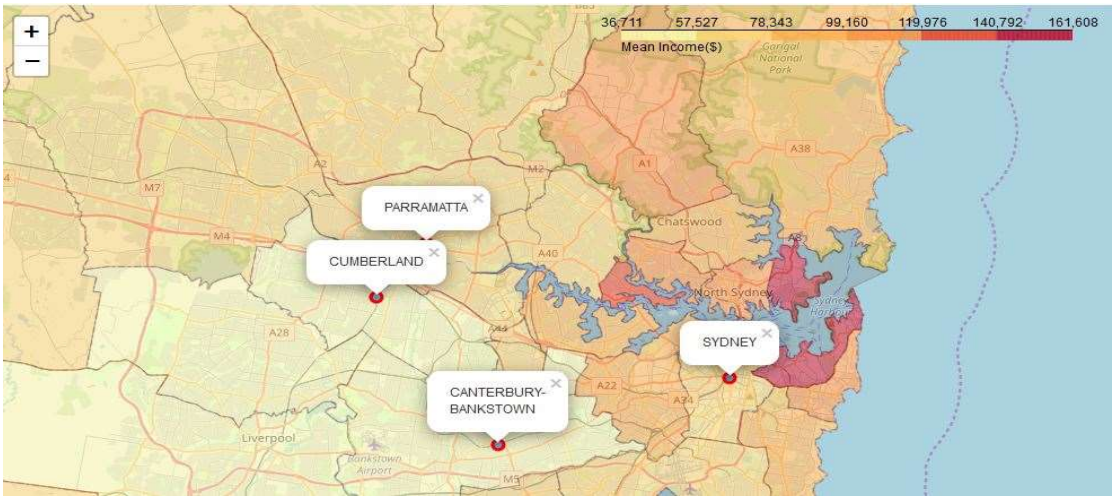

Figure 6: Mean Income with 4 LGAs under study

In terms of crime rates, Sydney is seen to be having higher crime rate than the other three. With Canterbury-Bankstown being in the lowest zone and Parramatta and Cumberland in next higher zone.

Figure 7: Crime Rate with 4 LGAs under study

K value for the k-Mean cluster was selected as 4. The elbow method did not quite give a clear elbow point. So manual check if each cluster was done for different k values to see if generated clusters made a logical sense.

When k-Mean clustering was run for neighborhoods in four LGAs of our concern, following groups were formed.

Red: Cluster 0 (home services and European restaurants)

Dark Blue: Cluster 1 (mostly with fast food, Middle Eastern/Asian restaurants, playground)

Light Blue: Cluster 2 (mostly near train station, park)

Greenish yellow: Cluster 3 (mostly with Café, pub/bar, pizza place, Italian restaurant)

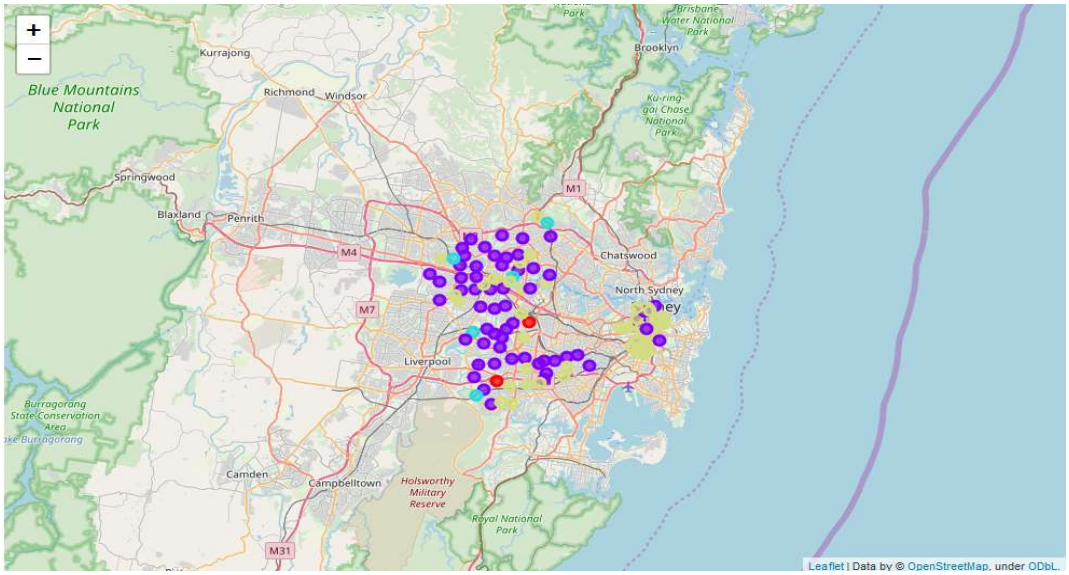

Figure 8: Clustering the Neighborhoods of 4 LGAs under study

6 Discussion

In this project, capital city of NSW Australia, Sydney was in focus. Various Local Government Areas (LGAs) were considered along with the suburbs/neighborhoods in each LGA. The scope of the project was limited to the highly populous LGAs in the Greater Sydney area. Top 10 highly populated LGAs were considered and then out of them, 4 LGAs with highest population density were taken. In this group of 4 LGAs, City of Sydney, City of Parramatta, City of Canterbury-Bankstown and Cumberland council made it.

Among the four LGAs, Sydney has the highest crime rate and population density. Sydney also had highest mean income. It was found that crime rate had very mild correlation with population density (positive) and mean income (negative). For the LGAs with mid to lower level of income range, it was seen that crime rate was quite independent of the mean income. In terms of characteristics based on venue categories, neighborhoods in Sydney LGA had highest concentration of cluster type three consisting of café, bar/pubs, pizza places. Both Parramatta and Cumberland councils, which had similar population density and crime rate, can be seen as having neighborhoods belonging to cluster type 1, having mainly fast food joints, Asian/Middle Eastern restaurants. Canterbury area had good mix of both cluster type 1 and type 3.

7 Conclusion

This project has gone into characterizing four of the most populous areas of Greater Sydney. Income data, crime rate, population density and most common type of venues nearby were used to view the characteristics of those areas. This project has made use of online available data sources including web data and online location-based service API, foursquare, to gain insights into the neighborhoods of different Local Government Areas in Greater Sydney.As an enhancement or further works on the project, house sales data can be brought in and cross checked with neighborhood venues in order to examine sale price depending on location.

References

[1] https://www.cityofsydney.nsw.gov.au/learn/research-and-statistics/the-city-at-a-glance/greater-sydney

[2] Link to list of all NSW Local Government Areas (LGAs)

File link:All NSW Local Government Areas (LGAs)

[3] Link to list of all NSW suburbs (neighborhoods)

File link: List of all NSW suburbs (neighborhoods)

[4] Wiki link to list of Greater Sydney LGAs

[5] Crime data from NSW Bureau of Crime Statistics and Research

[6] Total Income Distribution (table 2f.5)

File link: File for Total Income Distribution (table 2f.5)

[7] Population Data (Table 1. Estimated Resident Population, Local Government Areas, New South Wales)

[8] Geojson file of NSW LGAs