IBM Capstone Project:

Melbourne Area Relocation

Christopher van de Vyver

03 March 2020

Table of Figures

Figure 1: Location specified in central Melbourne ..... **Error! Bookmark not defined.**
Figure 2: Train Station Query Database 7
Figure 3: value_counts() Function results 7
Figure 4: Filtered dataframes 7
Figure 5: Unclustered Map 8
Figure 6: Marker popup representation 8
Figure 7: Final folium visualisation 9
Figure 8: Top 10 LGAs for Crime 10
Figure 9: Top 10 LGAs for Median Rent 10
Figure 10: Final LGA Comparison Dataframe 10

Table of Contents

Table of Figures 2
1. Background 4
   1.1. Introduction 4
   1.2. Problem Statement 4
   1.3. Target Market 4
2. Data Management 5
   2.1. Data Required 5
   2.2. Data Sources 5
   2.3. Data Handling 5
3. Methodology 6
4. Results 9
5. Discusssion 11
6. Conclusion 12
7. Recommendations 12

1. Background

1.1. Introduction

I recently graduated from University and moved from South Africa to Melbourne, Australia. By doing so I had to choose a place to live, the greatest contributing factors which influenced my decision were rental prices, crime rates and proximity to facilities such as public transport and grocery stores for the various Local Government Areas (LGA).

When inquiring about the local rental prices and crime rates, my relocation agency directed me to do my own research into the areas by reviewing spreadsheets on government websites, reading published articles and searching rental websites to find the data I required to make an informed decision. I would like to leverage the Foursquare location data to create a visualization (map) which displays the median rent and crime rate on the specific LGA and a dataframe which summarizes the above-mentioned data and ranks the suburbs according to these rates.

I believe the visualization and dataframe would have been extremely useful and allowed me to make an informed decision when I had to relocate, without having to spend countless hours searching through sources of no use or relevance. This leads me to believe other individuals relocating to Melbourne would benefit from this and hence, this data would be able to be sold to travel/relocation agencies to enable them to gain a competitive advantage over other agencies in helping individuals relocate with more ease or success.

The information might even be potentially sold to the individuals by setting up a website or app which displays this information and as time progresses, more data may be added such as number of public transport pickups/stops, grocery stores and restaurants are in each LGA which might influence an individual's decision upon choosing a location to stay.

1.2. Problem Statement

The greatest decision individuals that have chosen to relocate to Melbourne, Australia are faced with when relocating is choosing a place to stay. This decision impacts your family and social life, transport time to work as well as rental costs. The main contributing factors when deciding on a location to stay is the rental cost, safety of the area, proximity to public transport and grocery stores/supermarkets. Proximity to workplace will not be explored at this stage of the project as this is individual specific, however this may be investigated and expanded upon to be included in future along with more facilities in the specific areas in future.

1.3. Target Market

Individuals looking/or in the process of relocating to the Melbourne area in Victoria, Australia would make use of this as it would enable them to make a more informed

decision regarding the place they would like to stay and allow them to find a location that is able to meet both their financial needs and tolerance for crime.

2. Data Management

2.1. Data Required

In order to solve the above-mentioned problem, the data required to solve the initial business problem involving rental prices and crime rates is as follows:

- Foursquare LGA location data
- Crime statistics for the various LGAs
- Median rental rates for the specific LGAs

Methods such as web scraping will be used in order to obtain the information from the various related websites.

2.2. Data Sources

The sources from which the data required will be obtained are as follows:

- www.foursquare.com whereby a request will be sent to the API to search for transport venues and obtain the location data regarding the Local Government Areas (LGAs) required.
- https://www.dhhs.vic.gov.au/sites/default/files/documents/201908/Quarterly%20median%20rents%20by%20local%20government%20area%20-%20June%20quarter%202019.xlsx is the quarterly rental report containing the median rents by LGA.
- https://www.crimestatistics.vic.gov.au/sites/default/files/embridge_cache/emshare/original/public/users/201912/65/2e8549e44/Data_Tables_LGA_Criminal_Incidents_Year_Ending_September_2019.xlsx contains the crime statistics for the LGAs in Victoria, Australia.

2.3. Data Handling

Upon inspection of the data required (refer to § 2.1), the data will first be required to be “cleaned” in order to be used. The data which is required is in a format deemed by the sources from which it is obtained and requires changes before it may be used or placed into a pandas dataframe.

The foursquare request is a .json file from which the relevant information, stored in the ‘Venues’ category, which then needs to be converted into a dataframe using the json_normalize function. Following which, the data needs to be filtered in order to discard irrelevant information.

The Quarterly Report containing the median rents comprises of multiple years which allows trends to be observed, however for the purposes of this project, only the latest 2019 prices will be used/required. Thus, the remaining information will be discarded

and the latest median prices by LGAs for 1,2,3 bedroom flats and 2,3 and 4 bedroom houses will be averaged used.

The crime statistics includes police regions, offence divisions and subdivisions which are not needed. The data which is of main focus is the LGA incident rate per 100 000 population. Incidents range from homicide, assault to sexual offences and graffiti, thus it is fair to state that not all incidents are of equal severity, but for the purposes of this project, the number of incidents will be used. Perhaps in future, a single or select few categories of incidents may be filtered out and focused on to provide a better picture of the type of crime in the LGA such as comparing homicide rates between LGAs.

3. Methodology

When running the query to the Foursquare API, the location was set to central Melbourne CBD so as for the search radius to encompass as much of Melbourne city as possible.

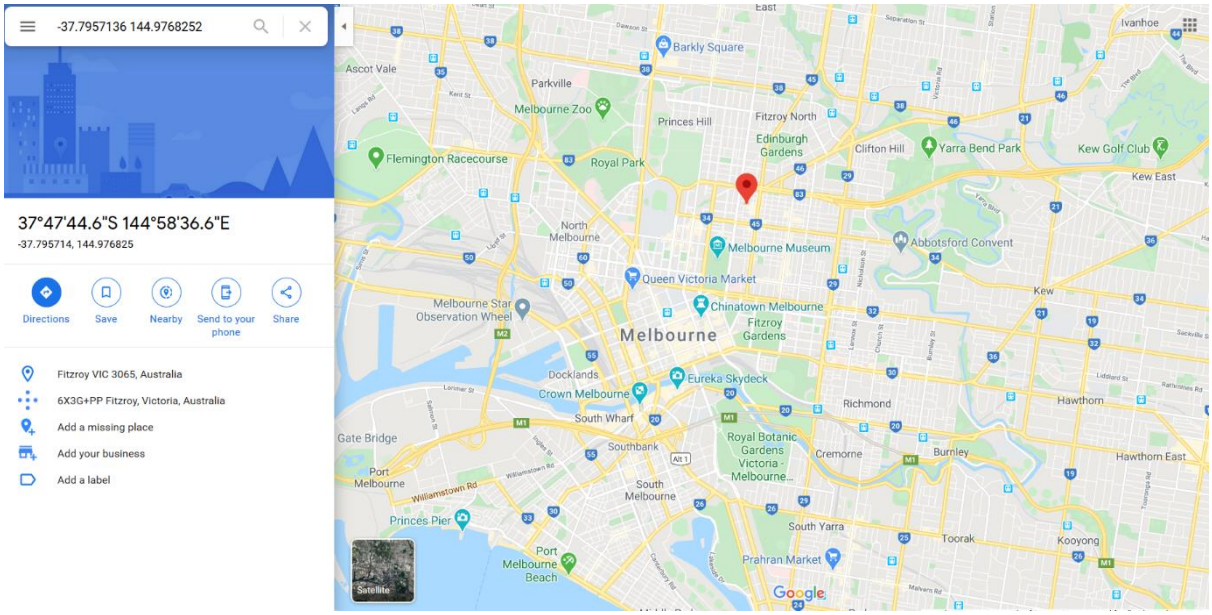

Figure 1: Location specified in central Melbourne

Queries were run for the terms “Bus Stops”, “Train Stations” and “Grocery Stores” in order to obtain the locations of these two public transport options and stores within the city. The main reason these were selected is after performing research, these were the two most used methods of transport by immigrants. The relevant JSON files were pulled from Foursquare and various data handling methods were used in order to filter and extract relevant info from the files. The data was converted from the following:

```
{'id': '4b058754f964a520a38c22e3',
  'name': 'Flinders Street Station',
  'location': {'address': 'Flinders St.',
   'crossStreet': 'at St. Kilda Rd.',
   'lat': -37.81826121222583,
   'lng': 144.96660713480878,
   'distance': 2666,
```

To the following database form:

	name	categories	lat	lng
0	Flinders Street Station	Train Station	-37.818261	144.966607
1	Southern Cross Station	Train Station	-37.818396	144.952679
2	Glenferrie Train Station Platform 1 & 2	Train Station	-37.821535	145.036440
3	Melbourne Central Station	Train Station	-37.810294	144.962974
4	Train Station Templestowe	College Gym	-37.757130	145.129390
5	Parliament Station	Metro Station	-37.809841	144.972439
6	Flagstaff Station	Metro Station	-37.811936	144.956364
7	Rail cafe at murrumbeena train station	Café	-37.890570	145.067468
8	Ringwood Train Station (Platform 1 & 2)	Platform	-37.816013	145.228811
9	Richmond Station	Train Station	-37.823859	144.989574
10	Mitcham Train Station Southern Commuter Carpark	Parking	-37.818206	145.193174
11	North Melbourne Station	Train Station	-37.807043	144.942042
12	Mitcham Train Station (Platform 1)	Platform	-37.817932	145.192091

Figure 2: Train Station Query Database

The final data frames which are of concern or focus comprise of the bus stop/train station and grocery store names and their coordinates specifically. As can be observed in Figure 2, not all rows comprise of train or metro station such as index numbers 4 and 7 which refer to a College Gym and Café respectively. Thus, the dataframes need to be filtered and validated further to ensure only relevant items are found in the dataframe. The function value_counts() may be used to assess the values found in the categories column (refer to Figure 3).

```
In [21]: pt_filtered.categories.value_counts()

Out[21]: Train Station    22
         Metro Station     2
         Platform          2
         Parking           1
         Dance Studio      1
         Café              1
         College Gym       1
         Name: categories, dtype: int64
```

Figure 3: value_counts() Function results

The dataframe needs to be filtered to only contain train/metro stations, the remaining items may be discarded.

```
In [14]: train_df.categories.value_counts()

Out[14]: Train Station    22
         Metro Station     2
         Name: categories, dtype: int64

In [15]: bstop_df.categories.value_counts()

Out[15]: Bus Stop       24
         Bus Station     2
         Name: categories, dtype: int64

In [16]: gs_df.categories.value_counts()

Out[16]: Grocery Store    9
         Name: categories, dtype: int64

In [17]: pttransport = pd.concat([train_df,bstop_df], axis = 0)
         pttransport.reset_index(drop = True)
```

Figure 4: Filtered dataframes

Using the Python library Folium, the plugin MarkerCluster and the Nominatim geocoder package from geopy, a map of the Melbourne area was created, and markers were placed at the corresponding locations for train stations, bus stops and grocery stores respectively. The tile type “OpenStreetMap” was selected which allowed for the best visualization of the data.

As can be observed in Figure 5, once the markers have been placed, it is extremely difficult to distinguish between closely placed markers due to the dense nature of the locations at which the markers were placed.

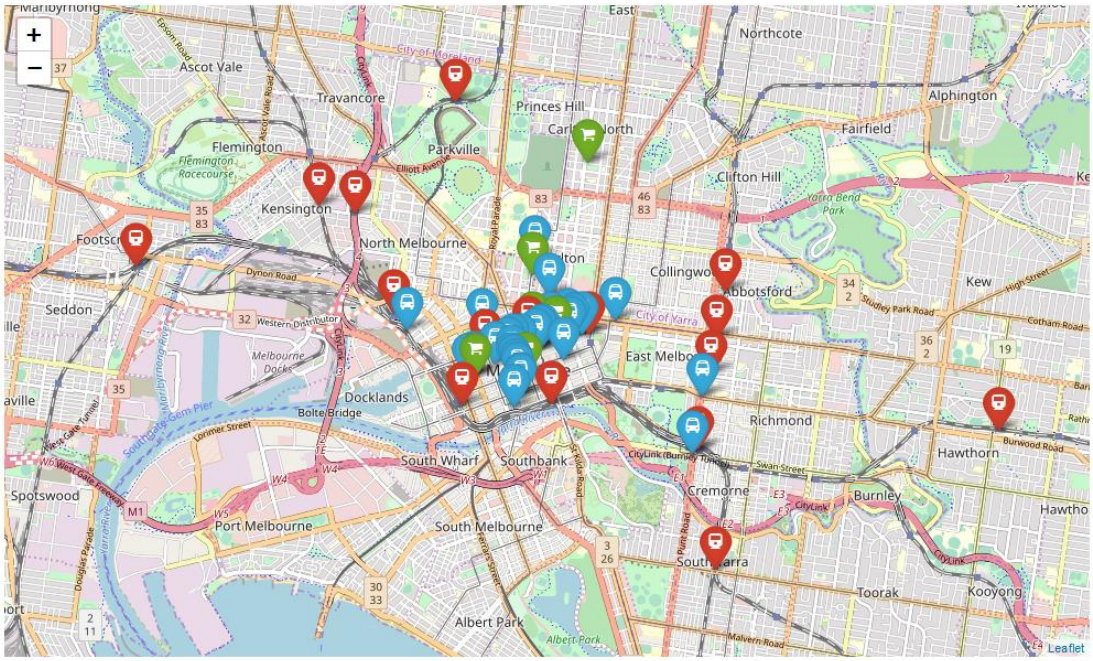

Figure 5: Unclustered Map

To address this issue, clusters were introduced into the map to allow for better visualization of the data points. Popups were introduced to display the corresponding marker name and venue type to enable the marker information to be displayed and observed with a single click. A representation may be observed in Figure 6.

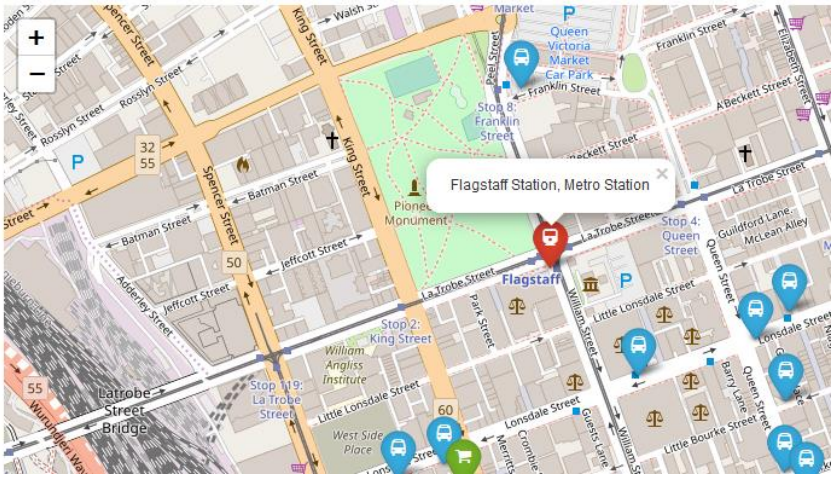

Figure 6: Marker popup representation

The crime and rental data were read in from the corresponding files. The crime data contained a white space infront of the words which did not allow the two dataframes to be merged. Pandas built in function str.strip() was used in order to remove the white space and allow the dataframes to be merged.

Once the dataframes had been merged on the column “LGA” corresponding to the Local Government Areas, ranks were assigned to the LGAs with regards to their crime rate and median rental rate. An overall combined ranking was assigned based on the other two rankings and the order was sorted according to this ranking. This allows individuals to see which LGAs were the best with regards to lowest crime and rental rates.

4. Results

The final output using Folium is a map containing clustered markers, each corresponding to a location of a train station, bus stop or grocery store. Each of these venues need has been assigned a different icon and colour to allow for easier distinction. The final map is shown in Figure 7. Individuals may zoom in and focus on a desired cluster such as shown in Figure 6.

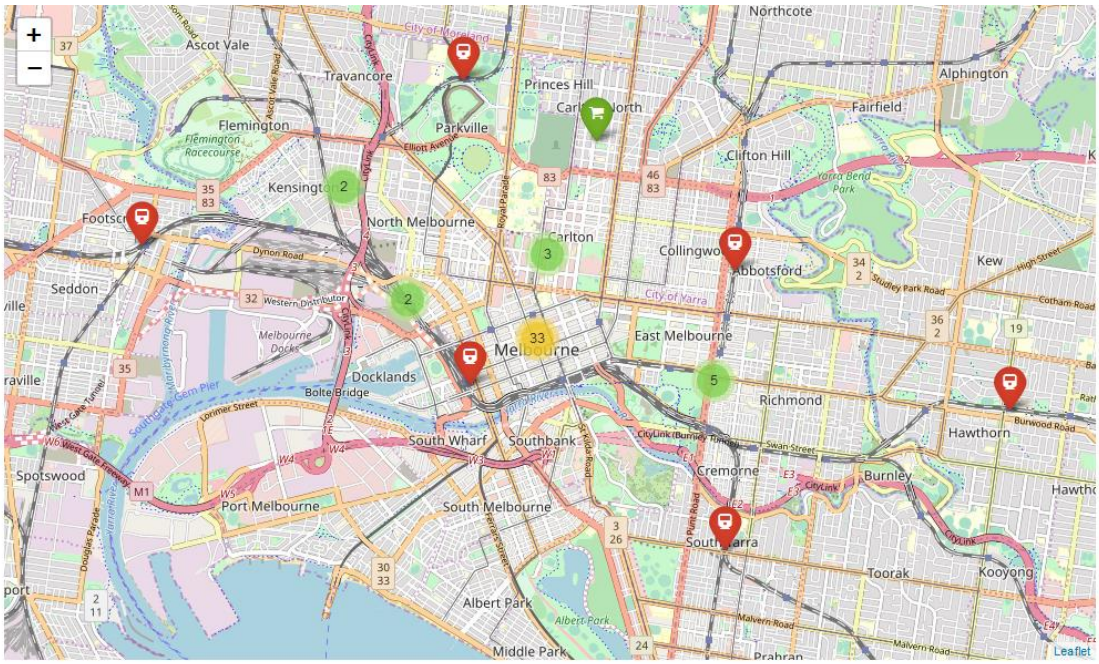

Figure 7: Final folium visualisation

Dataframes (Figures 8 and 9) containing the median rental income and crime incidents, by local government areas were created and ranking columns were created based on the median rental costs as well as the crime rates in the areas. The table was finally combined and ranked according to the combined ranking in order to determine which areas were best overall (Figure 10). The areas which placed in the top 10 for lowest median rental income, lowest crime incidents and lowest overall rating can be observed in the following figures.

	LGA	Rate per 100,000 population	Crime Ranking
0	Golden Plains	2122.823293	1.0
1	Indigo	2304.524120	2.0
2	Nillumbik	2636.618088	3.0
3	Alpine	2654.325843	4.0
4	Manningham	2719.907510	5.0
5	Moyne	2814.332520	6.0
6	West Wimmera	2827.978852	7.0
7	Towong	2926.456118	8.0
8	Buloke	3150.480530	9.0
9	Macedon Ranges	3188.647985	10.0

Figure 8: Top 10 LGAs for Crime

	LGA	Median	Rental Ranking
0	Yarriambiack	180	1.0
1	West Wimmera	195	2.0
2	Hindmarsh	200	3.0
3	Buloke	203	4.0
4	Gannawarra	220	5.0
5	Loddon	225	6.0
6	Northern Grampians	230	7.0
7	Latrobe	250	9.5
8	Central Goldfields	250	9.5
9	Pyrenees	250	9.5

Figure 9: Top 10 LGAs for Median Rent

	LGA	Median	Rate per 100,000 population	Rental Ranking	Crime Ranking	Overall Ranking
0	West Wimmera	195	2827.978852	2.0	7.0	1.0
1	Buloke	203	3150.480530	4.0	9.0	2.0
2	Hindmarsh	200	3503.447621	3.0	15.0	3.0
3	Corangamite	265	3402.127650	13.0	12.0	4.0
4	Yarriambiack	180	4176.377419	1.0	27.0	5.0
5	Indigo	300	2304.524120	26.5	2.0	6.5
6	Towong	280	2926.456118	20.5	8.0	6.5
7	South Gippsland	278	3429.262378	18.0	14.0	8.0
8	Loddon	225	4233.231586	6.0	28.0	9.5
9	Alpine	310	2654.325843	30.0	4.0	9.5

Figure 10: Final LGA Comparison Dataframe

5. Discusssion

The final Folium map shown in Figure 7 contains all the locations in Melbourne or grocery stores, metro and train stations as well as bus stops. These specific venues were selected as they are the main concern for an individual selecting a place to stay. One ideally would select a location that is as safe as possible whilst being close enough to transport options that allow commutes to work and the city which are as easy and fast as possible. The location has to satisfy all these criteria whilst still being able to be affordable for the individual/fit into each individual's respective budget.

The map was designed to be used as a tool in conjunction with the final database which contains all the Local Government Areas (LGAs) and their respective rates and rankings. The individual is ideally meant to consult the dataframe to see which LGAs fall into their budget and then refer to the map to explore the corresponding LGA to determine the grocery store, bus stop and train station locations.

After a number of LGAs within the budget and crime tolerance of the individual have been explored and the individual has assessed all relevant venue locations, one may begin to consult/search property websites for listings in the desired area with a lot more peace of mind.

The difficulty with a tool of this nature, is it is difficult to quantify the true amount of benefit or value it brings to an individual as it is saving the individual a large amount of time and frustration searching through online sources to try find locations scattered through many LGAs across Melbourne. Once listings have been found, unless the individual has a friend or family member living in the area or who is quite knowledgeable about the city, they will not know whether or not the area is safe, and the degree of safety in comparison to other LGAs. This tool provides the user with peace of mind by using recent statistical data from verified governmental agencies so the user does not have to question the validity of the information. The tool allows the user to make an informed choice and safeguard themselves as much as possible against incidents of crime in their residential area.

The tool also allows the user to assess listings found online by their location and proximity to transport options and grocery stores. After reviewing a few property websites, this information is not available for each listing and thus will enable the user to make a more informed and an objectively better decision.

After creating the respective dataframes and ranking the LGAs by the respective categories, the following LGAs were deemed the best by each ranking method: Yarriambiack had the lowest median rental rate, followed by West Wimmera and Hindmarsh. Golden Plains had the lowest crime incident rate followed by Indigo and Nillumbik. The top three LGAs by combined ranking are West Wimmera followed by Buloke and Hindmarsh.

These LGAs scored the best due to their location and population density. These areas are located further away from Melbourne CBD and have fewer people residing in them. The closer the LGA is to the Melbourne CBD, the higher the average rental rate and number of crime incidents is in that respective LGA. It would be disingenuous to state that the above mentioned LGAs are the best for an immigrant moving to Melbourne as their work is expected to be located in close proximity to the Melbourne CBD. Even though the above named best-scoring LGAs would not necessarily be the best areas to reside, the tool may still be used effectively. The user would be able to use the tool to select among a range of areas further down the ranking list which are in closer proximity to Melbourne CBD and make an informed decision.

6. Conclusion

- The best scoring LGAs may not necessarily be most optimal LGA for the individual's requirements.
- Even though the best scoring LGAs may not suit the user's criteria, the tool is still extremely useful and will yield benefits to the user by allowing them to select the most optimal LGA.
- The visualisation is an effective tool to be used in conjunction with the ranking system dataframes. One should not be used without the other.

7. Recommendations

- Additional venues may be added to increase the comprehensiveness of the mapping tool and enable users to make a more informed decision.
- More statistics such as number of restaurants or hospitals can be included in venue search. Even schools for couples or families moving.