Capstone Project: The Battle of Neighbourhoods

An analysis of neighbourhoods near Melbourne CBD using data science.

Phang Mun Yun

May 31, 2021

---

1. Introduction

1.1 Background

Buying your first home is exciting as it marks the start of a brand-new chapter in life. However, it is equally stressful as there are numerous factors to consider. These factors include family needs, location and accessibility, facilities and amenities, type of the property, condition of the property, and price of the property. There are many property listing websites available today that consolidates and promotes listing information that many first-time homebuyers can utilise in their research.

In this report, we will use data science methods to help narrow down neighbourhoods based on the needs of the homebuyer. The shortlisted neighbourhoods can then be used as input by the homebuyer when going through listing websites.

1.2 Problem Statement

In this report, we will help Sam to find suitable neighbourhoods in Melbourne (Victoria, Australia) based on his needs. Sam is 30 years old and has been working in Melbourne Central Business District (CBD) for the past 7 years. He is currently renting an apartment near his office with his fiancé. He and his fiancé are fitness freaks, they typically start their day with a morning workout together before work in a gym located near their apartment. After work, they would meet up in the local park for a short run before dinner. They typically cook at home on weekdays and eat out on the weekends with friends and family. Sam and his fiancé share an interest in Japanese cuisine and its always the choice of cuisine when they eat out.

As Sam's marriage ceremony is next year, he wishes to buy his first home and move in before the ceremony. He does not mind moving out of the city as he has a car to travel to and from work but does not wish to drive too long either. It would be a bonus if there's a train station nearby his new home, so he and his fiancé have the option of using public transportation. Hence, he wants his next home to be within 25 kilometres (kms) from the CBD. His budget for his new home is 1,000,000 Australian Dollars (AUD).

There are more than 300 neighbourhoods in Melbourne, and it is daunting to for Sam to go through it and match it to his needs. Summarised, Sam would most likely enjoy neighbourhoods that are within 25 kms from the CBD and has the following venues nearby:

1. Gym
2. Park
3. Grocery store
4. Japanese restaurant
5. Train station

With this information in mind, we will be using data science methods to help Sam identify 5 neighbourhoods that are most suitable to his needs.

1.3 Target Audience

This report is customised to Sam's needs as stated in the previous section. Nevertheless, the approach that we will be using can also be used and tweaked by other homebuyers in their property hunting research. Real estate agents may also be interested in this report as the approach we that will be using can assist them in providing customised data-driven suggestions to their clients.

2. Data

2.1 Description of data

There are 4 datasets required for this analysis:

1. Neighbourhood coordinates
2. Neighbourhood distance from CBD
3. Nearby venues
4. Median house price.

Neighbourhood coordinates dataset is needed as the main reference of location points. This dataset is sourced from Corra [1]. Corra is a private company based in Adelaide, South Australia, and they provide high quality web design and development solutions to their clients. The dataset published by Corra is used as it contains the entire country's latitude and longitude coordinates by postcode, suburb, and state. This dataset can be downloaded in CSV format.

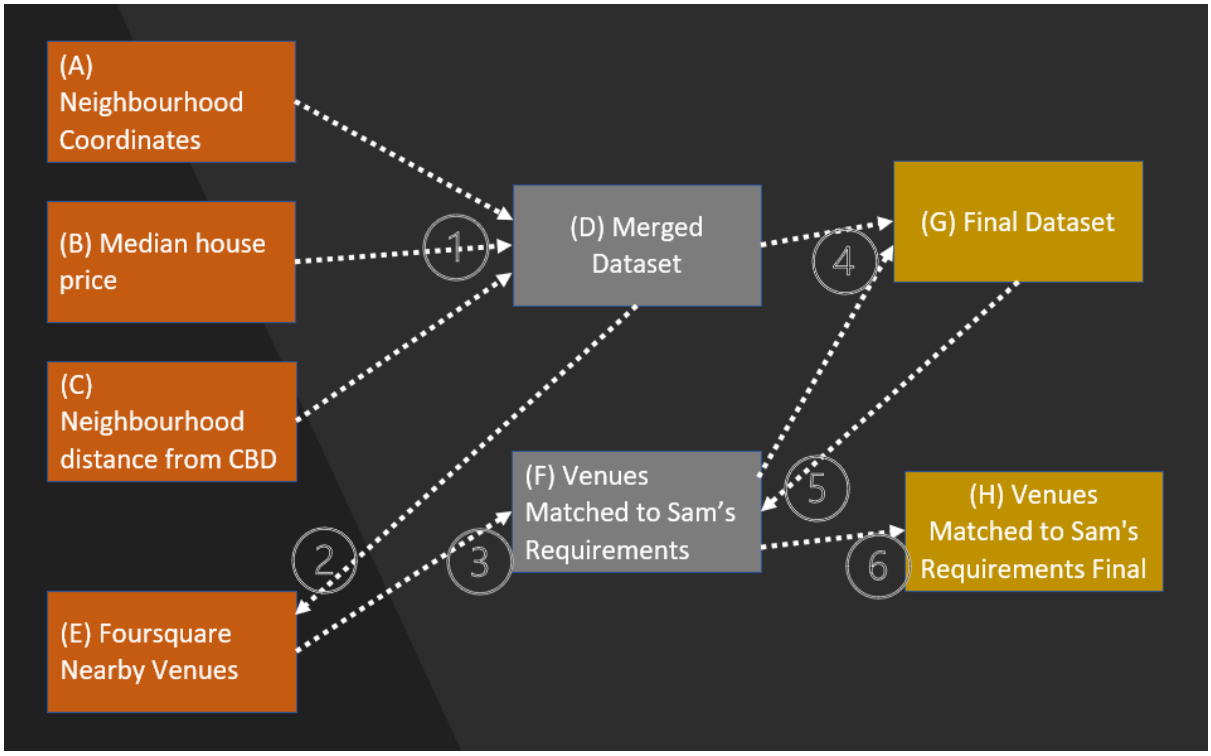
Neighbourhood distance from CBD dataset will be used as reference to Sam's new home requirements of within 25 kms from the CBD. This dataset is sourced from Myboot [2]. Myboot is a website that compiles information on Australian suburbs including demographics and crime statistics. We will be extracting their "Less than 25 km from Melbourne" list using web scraping methods as learned in Python Project for Data Science Course by IBM.

Nearby venues dataset will be used to extract venues information within a specified range (i.e., 500 metres) from a specific location. This dataset will be sourced from Foursquare [3]. Foursquare is an app that provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history. We will extract Foursquare venues data through accessing their Application Programming Interface (API) as learned in Applied Data Science Capstone Course by IBM.

Median house price dataset will be sourced from the Victoria State Government [4]. The dataset will be useful reference to fulfil Sam's budget requirements. This dataset can be downloaded in CSV format and it was recently updated on April 20, 2021.

2.2 Data Preparation

Data downloaded or scraped from various sources mentioned in the previous section were combined into two main datasets, namely: (G) Final Dataset & (H) Venues Matched to Sam's Requirements Final. These datasets will be used for analysis & visualisation.



For Dataset-A, loading it was straightforward with pandas.read_csv function as it is a CSV file downloaded from the source. The dataset contains all coordinates by postcodes in the country, about 16,000 rows. Hence many rows were filtered, and the remaining data only contains the state of Victoria, which is the state of Melbourne.

For Dataset-B, similar to Dataset-A, loading it was seamless with pandas.read_excel function as it is a XLS file downloaded from the source. The headers were a little inconsistent as it is split into many rows. Fixing the header to equals the first row and removing null values allowed us to have a readable header names.

For Dataset-C, the data is scraped using BeautifulSoup library. Minor transformations were made to the data such as removing leading and trailing whitespaces using .str.strip() and change the string to uppercase with .str.upper().

For Dataset-D, pandas.merge() function Datasets A to C using inner join. That also means all non-matched values will not be used.

For Dataset-E, a function was defined to loop through all nearby venues based on the coordinates of neighbourhoods in Dataset-D. In terms of radius for the function, 1,500 meters was used as Sam has a car, and do not mind driving nearby his future home.

For Dataset-F, a list of venue categories was used as filter to remove other rows in Dataset-E resulting in reducing the dataset size from 11,059 rows to 1,713 rows. Remaining venue categories are also cleaned and grouped into a new venue category. For instance, train station and tram station are

grouped into train/tram station. This allows one hot encoding to be performed next, which is then followed by groupby.() function to count the number of shortlisted venues by neighbourhood. Neighbourhoods that do not have at least 1 shortlisted venues are filtered (from 264 neighbourhoods to just 40).

For Dataset-G, Dataset-F was merged with Dataset-D using inner join to remove non-matched values. This results in just 35 neighbourhoods that met Sam's requirements of nearby venues and have clean coordinates, distance from CBD, and median price data.

For Dataset-H, Dataset-G was merged with Dataset-F to reduce nearby venues data from 1,713 rows to 403 rows. Like Dataset-G, this ensures that this dataset has cleaned references and will not over-populate the map visualisations later.

3. Methodology

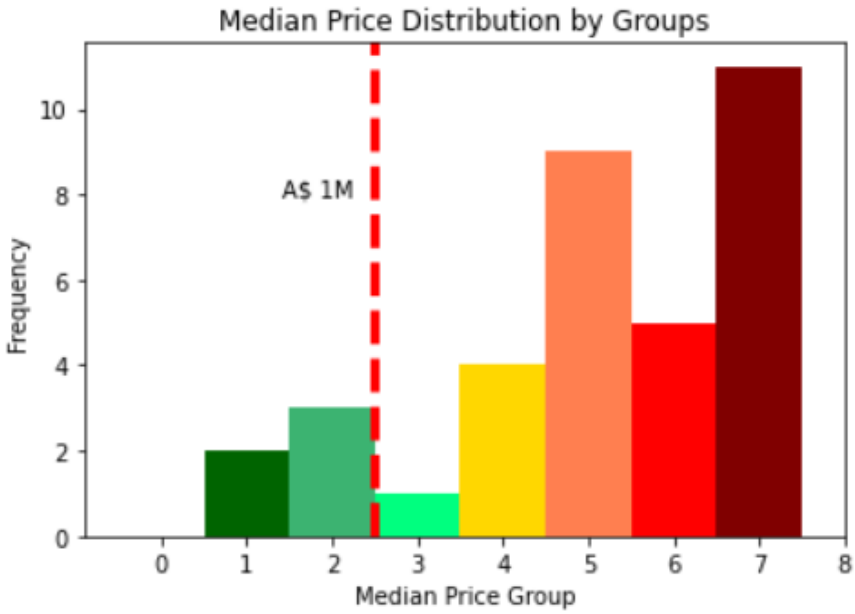
In this section, we will perform data analysis using Dataset-G Final Dataset (df) & Dataset-H Venues Matched to Sam's Requirements (df_venues_sam_final).

The outputs include:

- Median Price Group Distribution
- Map showing shortlisted Neighborhoods
- Map showing shortlisted Neighborhoods + nearby venues
- Cluster Analysis

3.1 Median Price Group Distribution

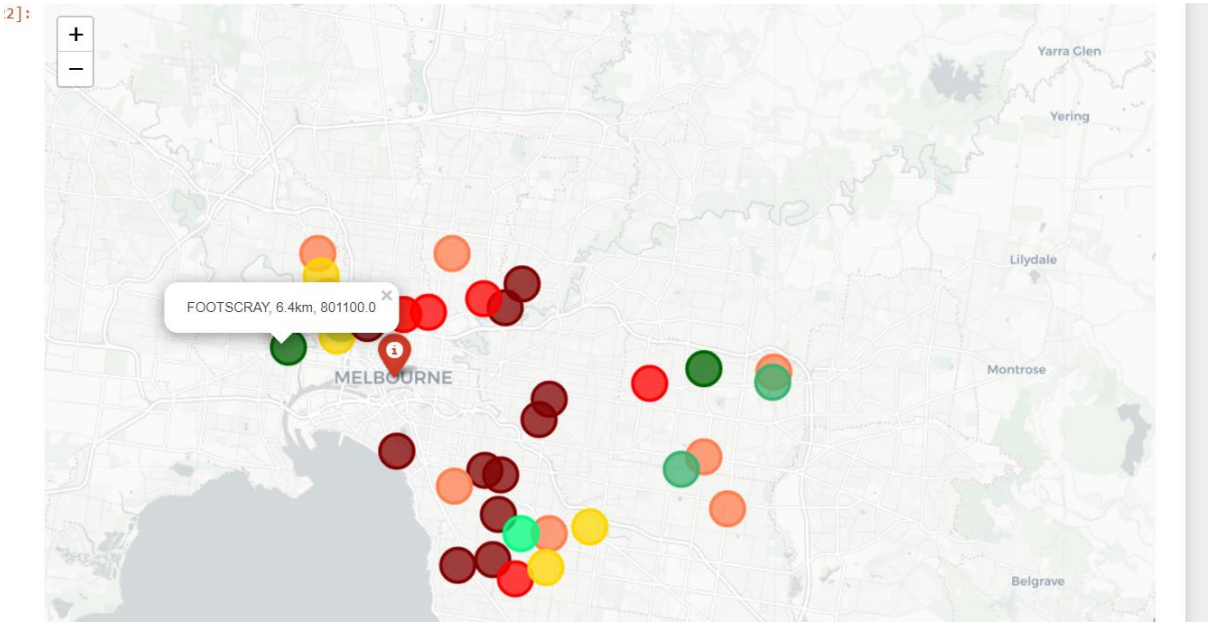
To understand the distribution of median prices across 35 neighbourhoods in Dataset-G, we first need to bin the prices into larger groups. Since Sam's budget is A$ 1M, I have binned the median price around this figure.




```
[1, # <900k
 2, # 900k to 1M
 3, # 1M to 1.1M
 4, # 1.1M to 1.2M
 5, # 1.2M to 1.4M
 6, # 1.4M to 1.6M
 7] # 1.6 to 5M
```

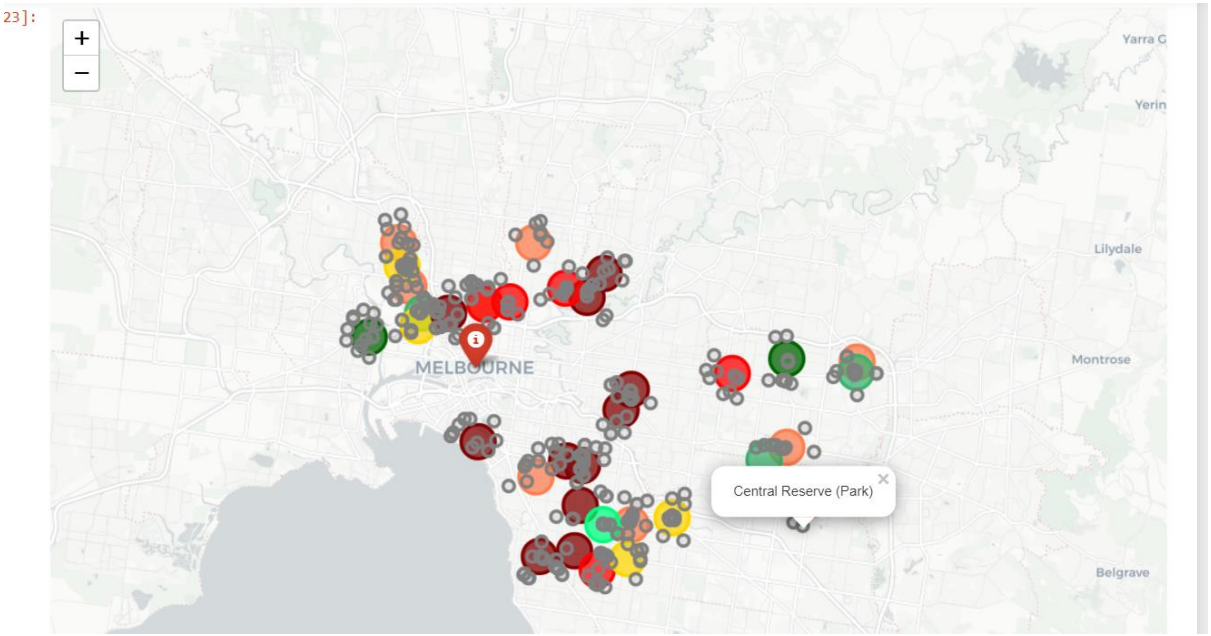
3.2 Map showing shortlisted neighbourhoods

To understand the distribution of neighbourhoods around Melbourne CBD by median price group, I have prepared a map using the Folium library to visualise it. The colours are reflecting the same theme as showed in the histogram previously, dark green being the lowest group and dark red being the highest group.



3.3 Map showing shortlisted neighbourhoods + nearby venues

Nearby venues data also provide more information to the map. As some neighbourhoods are close to each other, the share the same nearby venues as shortlisted based on Sam's requirements.



3.4 Cluster Analysis using K-Means Clustering

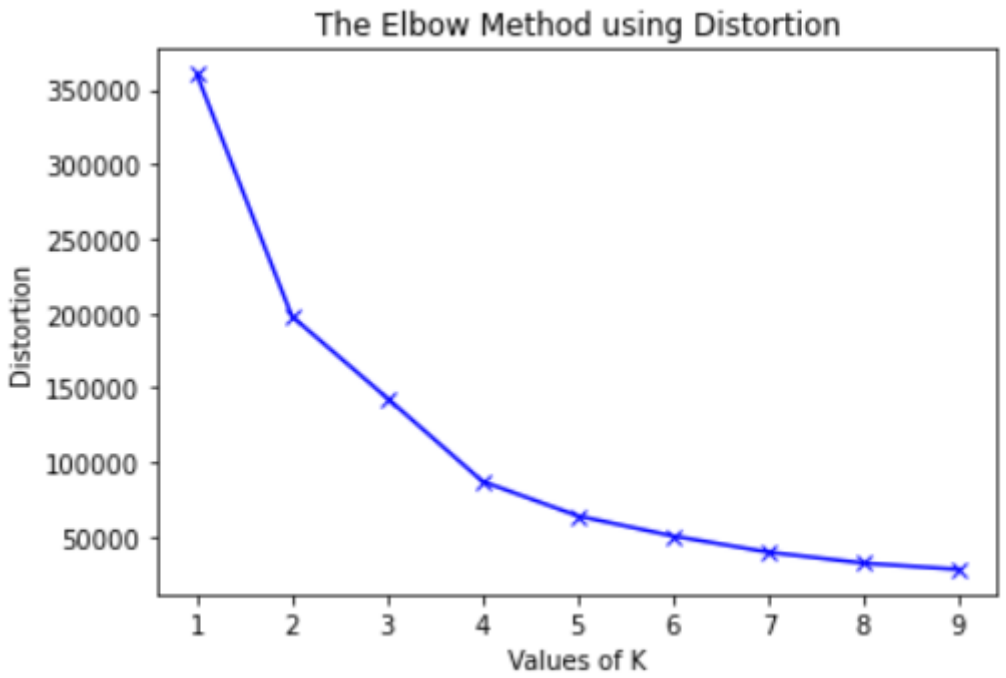
The aim of K-Means clustering technique is to segregate groups with similar traits and assign them into clusters. We will use this method to help us further segregate the 35 neighbourhoods into clusters and this would help Sam in focusing his in-depth research in property listings.

The following columns would be used as features to run the K-Means clustering technique.

]:

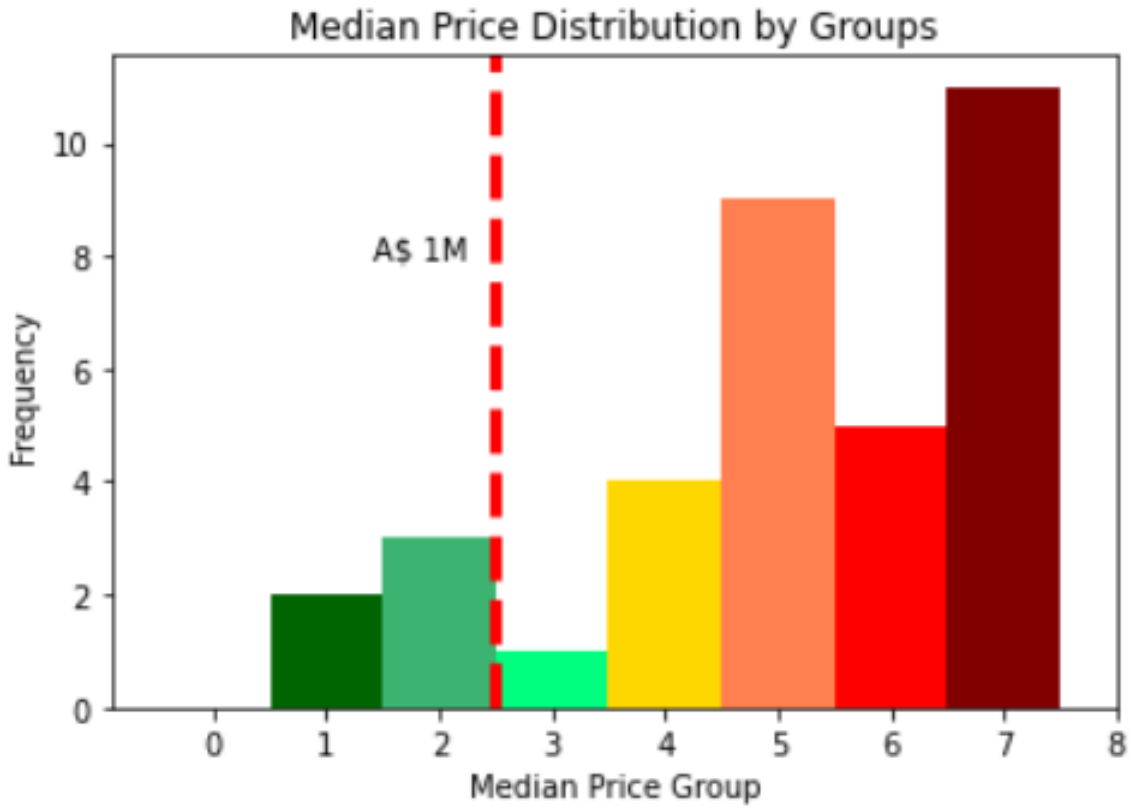
	Grocery Store	Gym	Japanese Restaurant	Park	Train/Tram Station	km	Median_Price
0	2.0	1.0	2.0	3.0	3.0	6.4	1865000.0
1	4.0	1.0	4.0	2.0	6.0	6.3	2054000.0
2	7.0	5.0	2.0	1.0	2.0	5.9	1328800.0
3	3.0	1.0	2.0	2.0	1.0	15.9	872000.0
4	3.0	2.0	1.0	1.0	3.0	13.1	1429000.0

As K-Means clustering technique requires the user to specify the number of clusters before it runs the algorithm, we used the Elbow Method to find the optimal value of k in K-means.



Based on the Elbow Method, we have selected 4 as the optimum value of clusters.

4. Results & Discussion

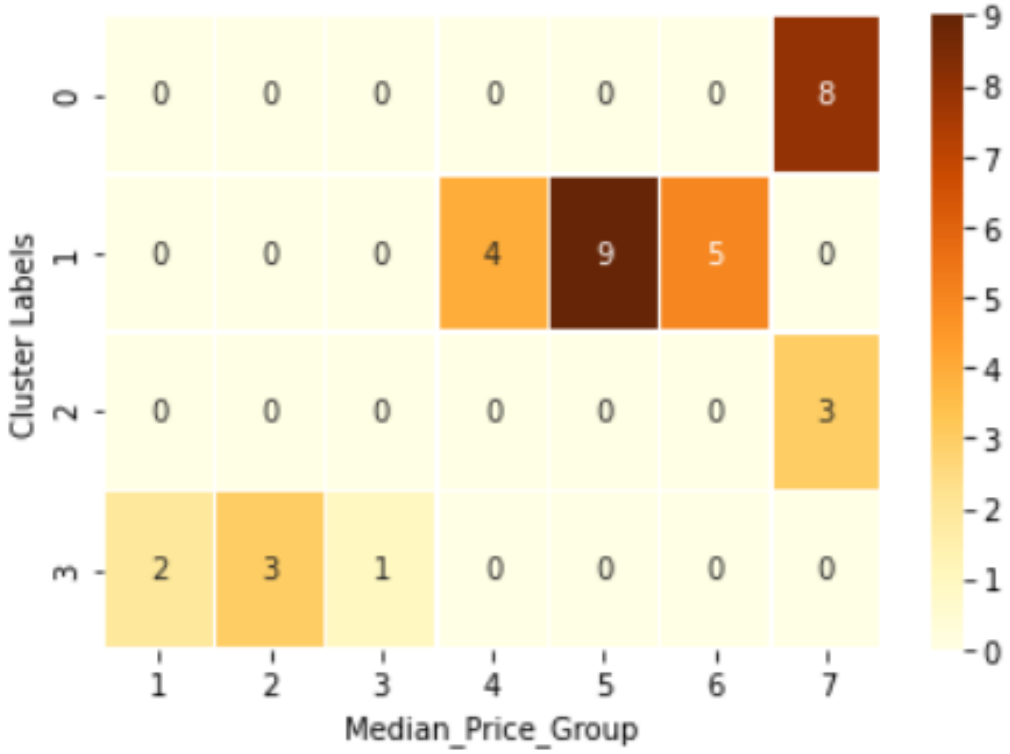


]:

```
[1, # <900k
 2, # 900k to 1M
 3, # 1M to 1.1M
 4, # 1.1M to 1.2M
 5, # 1.2M to 1.4M
 6, # 1.4M to 1.6M
 7] # 1.6 to 5M
```

Median_Price_Group	Neighborhood
1	2
2	3
3	1
4	4
5	9
6	5
7	11

The first observation we can make is from the median price group distribution as shown above. Only a minority (5/35 = 15% = 1 out of 7) of the neighbourhoods shortlisted are within Sam's budget. There are 2 neighbourhood groups that might be too expensive for Sam, that is neighbourhoods in Group 6 & 7. Nevertheless, we will not remove these datapoints as they are median prices only, and there might be properties priced around Sam's budget, when he continues his research on the listing websites.



Cluster Labels	Neighborhood
0	8
1	18
2	3
3	6

Cluster Labels	Grocery Store	Gym	Japanese Restaurant	Park	Train/Tram Station	km	lat	lon	Median_Price	Median_Price_Group
3	4.000000	2.000000	1.666667	1.666667	1.500000	12.200000	-37.827170	145.059342	9.179333e+05	1.833333
1	3.444444	1.888889	2.111111	2.000000	2.277778	8.500000	-37.821884	145.027613	1.305606e+06	5.055556
0	3.000000	1.000000	1.750000	2.250000	3.000000	7.337500	-37.838766	145.018720	1.853500e+06	7.000000
2	3.000000	1.666667	2.333333	2.333333	3.666667	5.366667	-37.848668	145.015247	2.532500e+06	7.000000

Based on the table above, we can observe that as km from CBD is greater, the median price decreases. Cluster 1 looks interesting, as its distribution of shortlisted venues are greater than 1.8 across all categories, suggesting a more balanced cluster than the others. I would recommend him to start with cluster 1 & 3 first, as these neighbourhoods are within his budget (around median price).

5. Conclusion

The purpose of this project is to prepare a shortlist of neighborhoods for Sam to consider based on his needs. Data from multiple sources were transformed & combined to achieve this purpose. Using K-Means clustering technique, I was able to provide a more refined list of neighborhoods that I feel Sam would like. From over 300 neighborhoods near Melbourne to less than 25 neighborhood recommendations, Sam will be able to find his dream home very soon!

Some potential improvement data points that could increase the value of this project:

1. Friends and family location by neighborhood data

2. Available homes for sale & price data

3. Crime statistics by neighborhood data

4. Neighborhood demographics data

6. References

[1] *Corra*. URL: https://www.corra.com.au/australian-postcode-location-data/ (visited on 2021/05/23)

[2] *Myboot*. URL: http://myboot.com.au/VIC/25/suburblist.aspx (visited on 2021/05/23)

[3] *Foursquare*. URL: https://foursquare.com/ (visited on 2021/05/23)

[4] *Victoria State Government*. URL: https://discover.data.vic.gov.au/dataset/victorian-property-sales-report-median-house-by-suburb (visited on 2021/05/23)