Moving to Perth

Introduction

In this project we will try to find an optimal suburb to rent a house in. The report is targeted at a couple who have two dogs, love Chinese food, enjoy shopping and looking to move to the city of Perth, Western Australia.

The couple has hired a data scientist to help and the criteria they have given to select a suburb are those suburbs that have the highest combined quantities of parks, Chinese restaurants and shopping malls.

Data

In order to solve this problem a list of Perth suburbs was scraped from the following website: https://en.wikipedia.org/wiki/List_of_Perth_suburbs , using the Pandas' function *read_html().* The table containing the suburbs on the website was then converted to a dataframe with the first five rows shown in Figure 1.

	Suburb	Local government area
0	Alexander Heights	Wanneroo
1	Alfred Cove	Melville
2	Alkimos	Wanneroo
3	Anketell	Kwinana
4	Applecross	Melville

Figure 1: First five rows of Perth Suburbs dataframe

The suburbs in this dataframe were then converted to a list and the *Geocoder* module was used to obtain Latitude and Longitude of each suburb. The Latitude and Longitude values were then added to the dataframe containing the suburbs with the first five rows shown in Figure 2.

	Suburb	Local government area	Latitude	Longitude
0	Alexander Heights	Wanneroo	-31.83101	115.85356
1	Alfred Cove	Melville	-32.03119	115.81566
2	Alkimos	Wanneroo	-31.61629	115.68792
3	Anketell	Kwinana	-32.21018	115.85172
4	Applecross	Melville	-32.01033	115.83483

Figure 2: First five rows of Perth Suburbs dataframe with Latitude and Longitude

Using the Latitude and Longitude of each suburb, the Foursquare API was then leveraged to extract venue category information. A radius of 1km from the Latitude and Longitude of each suburb was utilised. An example of the data extracted from the Foursquare API is given in Figure 3.

	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alexander Heights	-31.83101	115.85356	Domino's Pizza	-31.830509	115.853342	Pizza Place
1	Alexander Heights	-31.83101	115.85356	Coles	-31.830137	115.854053	Supermarket
2	Alexander Heights	-31.83101	115.85356	Alexander Heights Shopping Centre	-31.829769	115.853448	Shopping Mall
3	Alexander Heights	-31.83101	115.85356	Red Rooster	-31.830807	115.853138	Fast Food Restaurant
4	Alexander Heights	-31.83101	115.85356	Woolworths	-31.826764	115.852742	Supermarket

Figure 3: First five rows of Perth venue data

Methodology

In order to perform further analysis, some wrangling was performed on the data. The dataframe containing all the venue data was filtered so it only contained park, shopping mall and chinese restaurant venue categories. The venue categories were then converted into a binary vector using Pandas' *get_dummies()* function. The dataframe containing the binary vectors was then grouped by suburb and the sum of the vectors. This enabled a dataframe containing the total number of each category by suburb to be created as seen in Figure 4.

	Suburb	Chinese Restaurant	Park	Shopping Mall
0	Alexander Heights	0	1	1
1	Alfred Cove	0	4	2
2	Alkimos	0	2	0
3	Applecross	0	3	0
4	Ardross	0	1	0

Figure 4: First five rows of sum of each venue category by suburb

K-means clustering was then thought to be a reasonable machine learning method to aid in finding the optimal suburb. Using this unsupervised learning method is thought to quickly find varying combinations of the venue category data and furthermore find locations with the highest amount of all three venue categories combined. Scikit-learn was used for the clustering of the data using the *KMeans* module of the *sklearn.cluster* library. Five clusters were chosen for this problem as this would be sufficient to find suburbs with the highest combinations of all three categories.

The cluster labels were then added to a dataframe that contained the suburbs and locations. *Folium* was then used to plot the locations of these clusters in the Perth area as seen in Figure 5. From this map it was noted that there was no specific pattern in the geographic location of the clusters but cluster 2 has a small proportion of total suburbs compared to the other clusters.

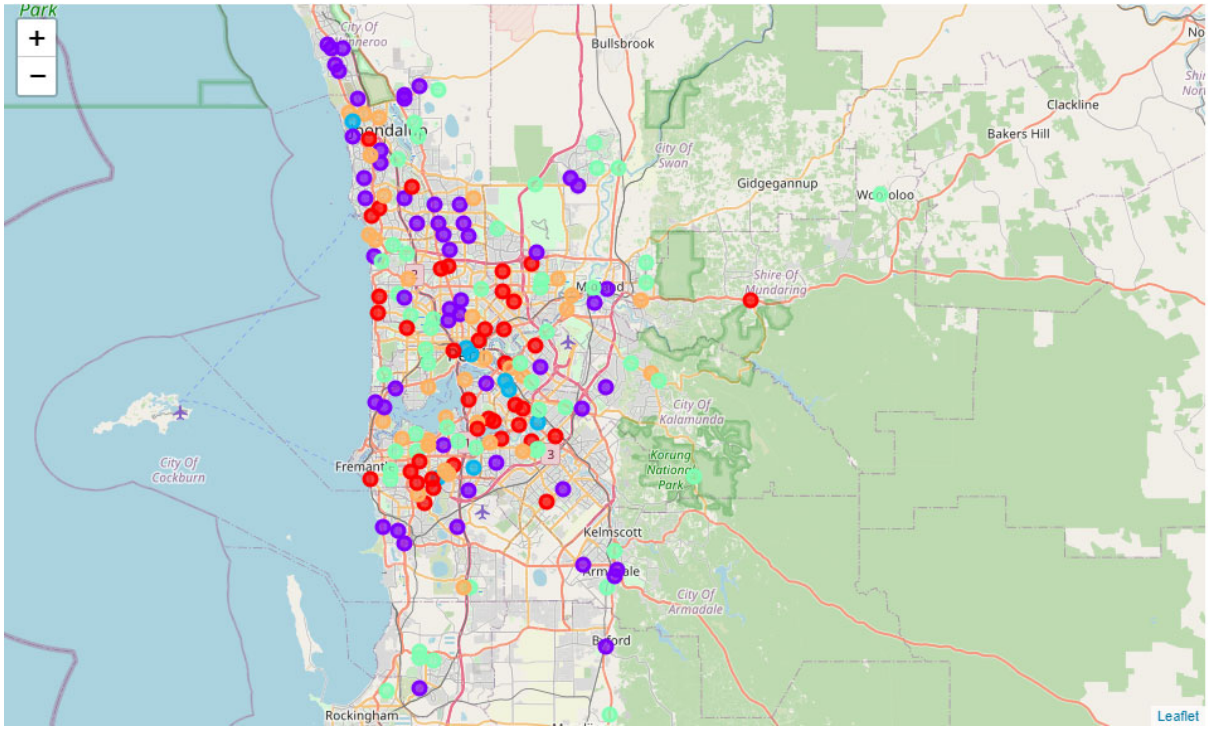

Figure 5: Location of clusters – 0 -red, 1 – purple, 2 – blue, 3 – green and 4 - orange

Results

Matplotlib was then used to further complete data analysis of the clustered data. A bar plot was used to show the mean and maximum number of locations within each cluster as seen in Figure 6 and Figure 7.

Figure 6: Mean number of locations by cluster

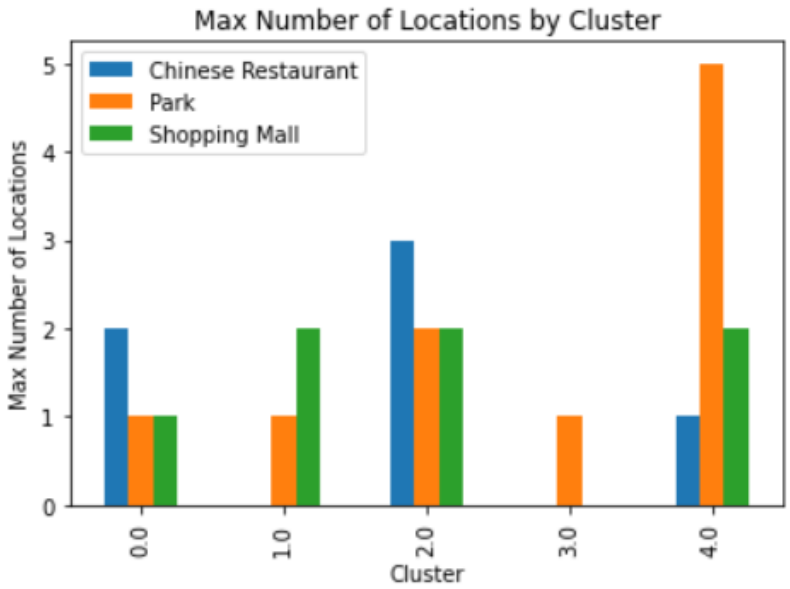

Figure 7: Mean number of locations by cluster

From these plots it can be noted that:

- Cluster 2 has the highest mean number of Chinese restaurants and shopping malls
- Cluster 2 contains suburbs that have the highest number of Chinese restaurants and shopping malls.
- Cluster 4 has the highest mean and maximum number of parks
- Cluster 4 has low mean number of Chinese restaurants and shopping malls

Discussion

Through analysis of the clustered data given in Figure 6 and Figure 7 it can be concluded that the suburbs contained in cluster 2 have the highest quantities of combined parks, Chinese restaurants and shopping malls. Cluster 4 does have a higher mean and maximum for parks however the mean and maximum shopping mall and Chinese restaurant values are lower or equal to that of cluster 2. From this further analysis of cluster 2 was completed in order to find the optimum suburb.

A Matplotlib bar plot was created to show the number of different localities in each suburb of cluster 2 as seen in Figure 8. From the bar plot it is noted that Bull Creek has the highest number of parks and shopping malls and the second highest number of Chinese restaurants. Cannington, East Victoria Park, Northbridge and Victoria Park all have one more Chinese restaurant than Bull Creek but as they have less parks and/or shopping malls, Bull Creek still seems like the optimal suburb to rent a house in. All suburbs excluding Northbridge and Victoria Park in cluster 2 seem like reasonable choices however as they provide at least one option for each venue category. Northbridge and Victoria Park are not recommended as they have no parks.

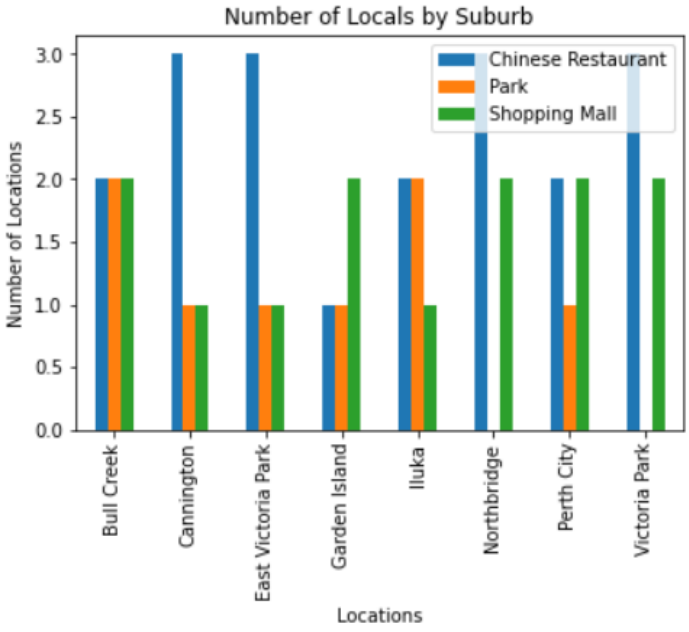

Figure 8: Number of different locations within cluster 2 suburbs

Conclusion

It can be concluded that Bull Creek is the optimum suburb in Perth, Western Australia, for the couple to look to rent a house in as it has high quantities of parks, Chinese restaurants and shopping malls. The location of Bull Creek is given by the map created using Folium in Figure 9. Further reasonable choices that have slightly less combined quantities than Bull Creek are Cannington, East Victoria Park, Garden Island, Iluka and Perth City.

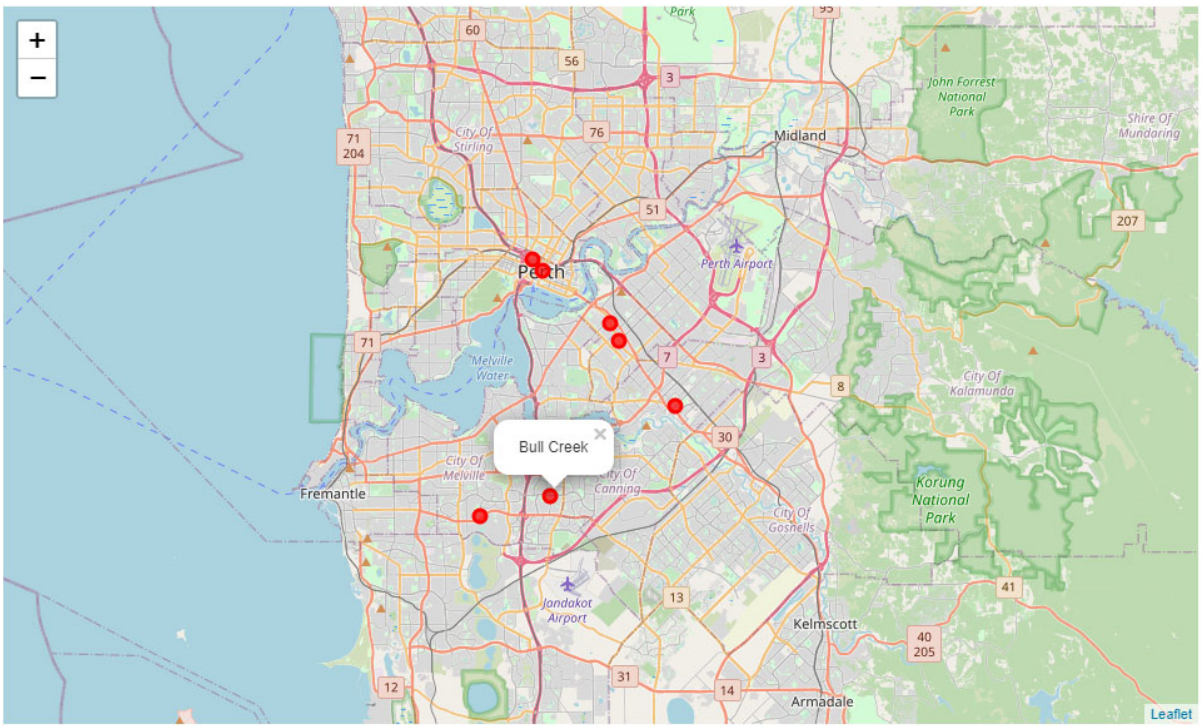

Figure 9: Location of suburbs in cluster