BATTLE OF NEIGHBOURHOOD – HOLIDAYS IN SINGAPORE

Lee chow jin

1. Introduction

1.1 Background Information

Singapore is centrally located in South East Asia, serving as a hub for major airlines connecting tourists to dream destinations such as Thailand, Malaysia, Indonesia, etc. As a result, many will stop over for a few days to discover the country who is known as a culinary and shopping haven.

After having our daughter, my husband and I have not stopped travelling but our needs for accommodations have changed. For instance, we would need a kitchen to prepare home-cooked food for our baby, so Airbnb has become one of our top options for finding an accommodation. Hence, this project is to help travelers with kids to identify appropriate accommodations to fully enjoy their stay in Singapore.

1.2 Problem Statement

There is 4000+ over accommodations in the Airbnb listing as of December 2020. How can we classify these accommodations into clusters and find more easily the right accommodation that corresponds to the budget and preferences of the travelers? This project will try to answer this question.

1.3 Target Audience

This report is dedicated to travelers who are looking for a short period of stay in Singapore. The results should give them a quick overview of which areas to stay that would satisfy their budget and preferences. The type of accommodations has been limited to "Entire place" as these is the accommodation most suitable for family, but it can be also be relevant for travelers who seek for such type of accommodation.

2. Data

2.1 Data Require

- Listing of Singapore Airbnb accommodation
- Number of venues including restaurants, shopping malls and tourists' attractions in the vicinity of the accommodation's neighborhood

2.2 Data Sources

Below is the link to the dataset.

- Airbnb accommodation listing from Inside Airbnb: http://insideairbnb.com/get-the-data.html (Data compiled in December 2020)
- Foursquare API to extract data on venues in a neighbourhood: https://developer.foursquare.com/

2.3 Data Understanding and preparation

After importing the Singapore's Airbnb accommodation listing, we will perform exploratory data analysis just to have an overview of the data content. We also remove non-necessary column so as to simplify the data for further treatment.

```
#Fetch Singapore listing from AirBnB
df = pd.read_csv('http://data.insideairbnb.com/singapore/sg/singapore/2020-12-29/visualisations/listings.csv', low_memory = False)
df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	numbe
0	49091	COZICOMFORT LONG TERM STAY ROOM 2	266763	Francesca	North Region	Woodlands	1.44255	103.79580	Private room	79	180	1
1	50646	Pleasant Room along Bukit Timah	227796	Sujatha	Central Region	Bukit Timah	1.33235	103.78521	Private room	80	90	18
2	56334	COZICOMFORT	266763	Francesca	North Region	Woodlands	1.44246	103.79667	Private room	66	6	20
3	71609	Ensuite Room (Room 1 & 2) near EXPO	367042	Belinda	East Region	Tampines	1.34541	103.95712	Private room	174	90	20
4	71896	B&B Room 1 near Airport & EXPO	367042	Belinda	East Region	Tampines	1.34567	103.95963	Private room	93	90	24

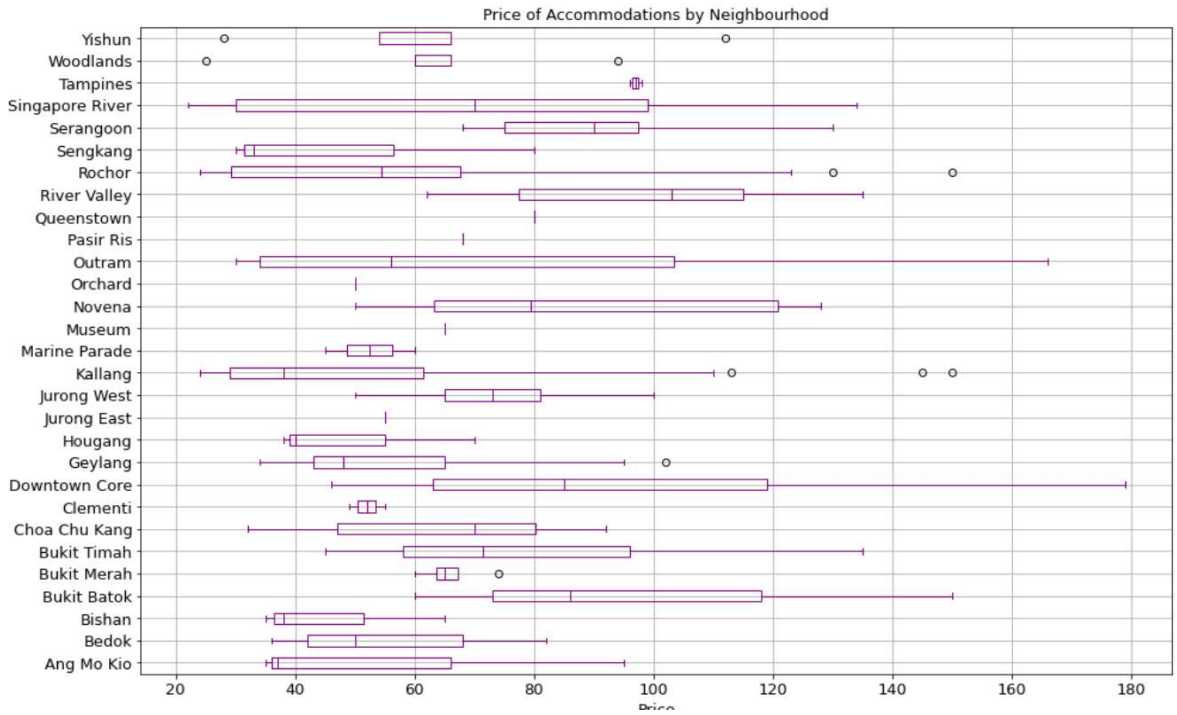
We check if all the rows are correctly filled and the types of data corresponds to the title of the column

```
df1.isnull().sum()
```

```
id                  0
neighbourhood       0
latitude            0
longitude           0
room_type           0
price               0
minimum_nights      0
availability_365    0
reviews_per_month   0
number_of_reviews   0
dtype: int64
```

```
df1.dtypes
```

```
id                    int64
neighbourhood        object
latitude            float64
longitude           float64
room_type            object
price                 int64
minimum_nights        int64
availability_365      int64
reviews_per_month   float64
number_of_reviews     int64
dtype: object
```



The result dataframe used will be only 309 lines and with 10 columns as shown below:

	id	neighbourhood	latitude	longitude	room_type	price	minimum_nights	availability_365	reviews_per_month	number_of_reviews
0	49091	Woodlands	1.44255	103.79580	Private room	79	180	365	0.01	1
1	50646	Bukit Timah	1.33235	103.78521	Private room	80	90	365	0.22	18
2	56334	Woodlands	1.44246	103.79667	Private room	66	6	365	0.17	20
3	71609	Tampines	1.34541	103.95712	Private room	174	90	365	0.18	20
4	71896	Tampines	1.34567	103.95963	Private room	93	90	365	0.21	24

3. Methodology

The methodology is made up of 4 sections listed below:

Part 1: Exploratory Data Analysis on Accomodation listing

- plot average price per neighbourhood.
- plot accommodations on folium to see their distribution on Singapore's map.

Part 2: Get Venues around accommodation

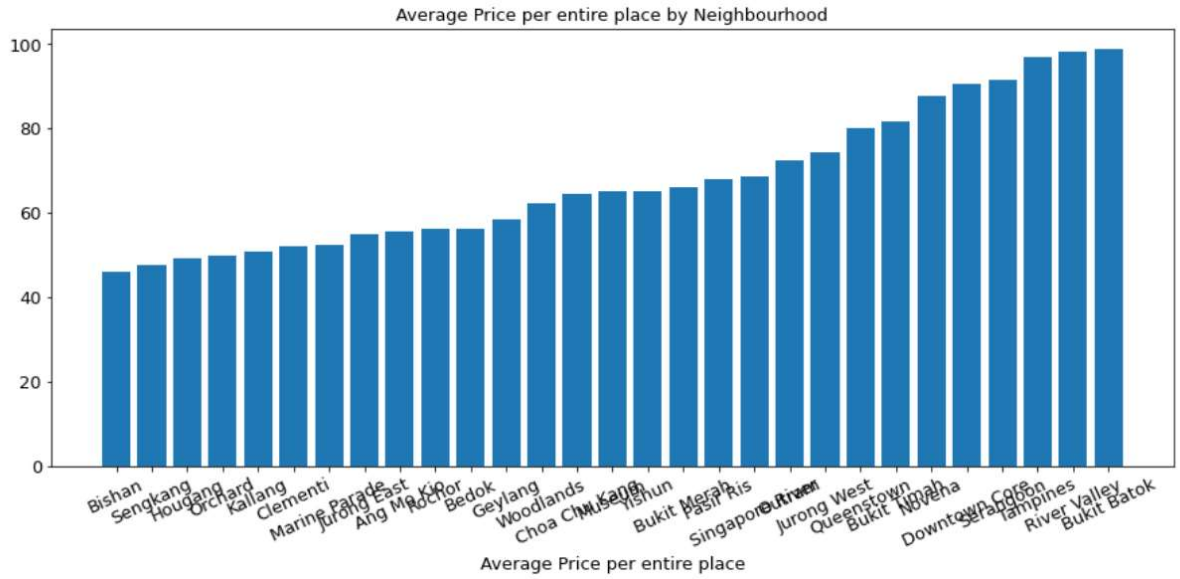
- use the foursquare API to find the venues around the accommodations.
- explore the dataframe of the venues collected.
- classify the venues them into main categories (Food, Shopping, Transport, etc.) that are of interest to travellers.

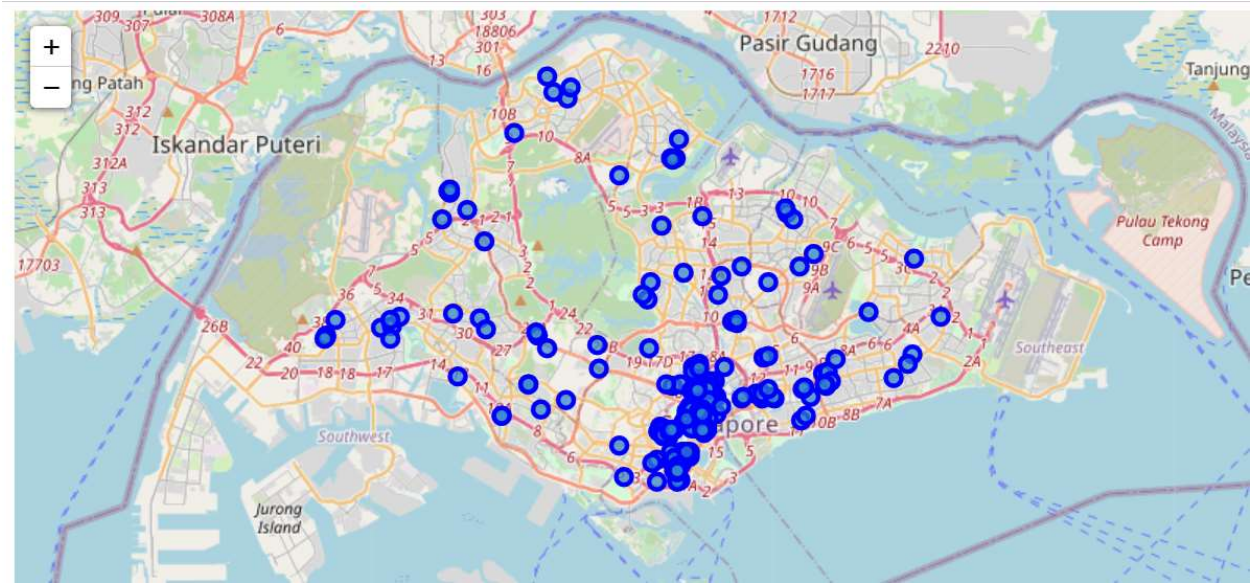
Part 3: One Hot Encoding and K-means Clustering.

Part 4: Further analysis on each cluster¶

3.1 Exploratory Data Analysis on Accommodation listing

We will plot bar graph to have an overview of the distribution of price and a folium map of the geographical locations of the selected accommodations.





3.2: Get Venues around accomodation

Using the Foursquare API, we gather the venues around the neighborhood. First, we define a getNearbyVenues function which is used to call Foursquare API to gather the venues. Next, we perform an exploratory analysis on the venues dataframe using head, shape and groupby to have an overview of the contents in the dataframe. Lastly, we classify all the venue categories into 5 main categories (food, entertainment, shopping, outdoor, transport and culture) for more efficient clustering.

```
In [24]: df_v.head()
```

Out[24]:

	id	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	56334	1.44246	103.79667	Asia Ghani	1.437431	103.795388	Malay Restaurant
1	56334	1.44246	103.79667	Kampung Admiralty Hawker Centre	1.439939	103.800774	Food Court
2	56334	1.44246	103.79667	Starbucks	1.439761	103.800659	Coffee Shop
3	56334	1.44246	103.79667	McDonald's	1.445931	103.798101	Fast Food Restaurant
4	56334	1.44246	103.79667	NTUC Fairprice	1.439955	103.800761	Supermarket

```
In [25]: df_v.shape
```

Out[25]: (12641, 7)

```
In [26]: df_v1=df_v.groupby("Venue Category").count()
         df_v1
```

Out[26]:

	id	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
Venue Category						
Accessories Store	7	7	7	7	7	7
African Restaurant	1	1	1	1	1	1
American Restaurant	14	14	14	14	14	14
Argentinian Restaurant	9	9	9	9	9	9
Art Gallery	46	46	46	46	46	46

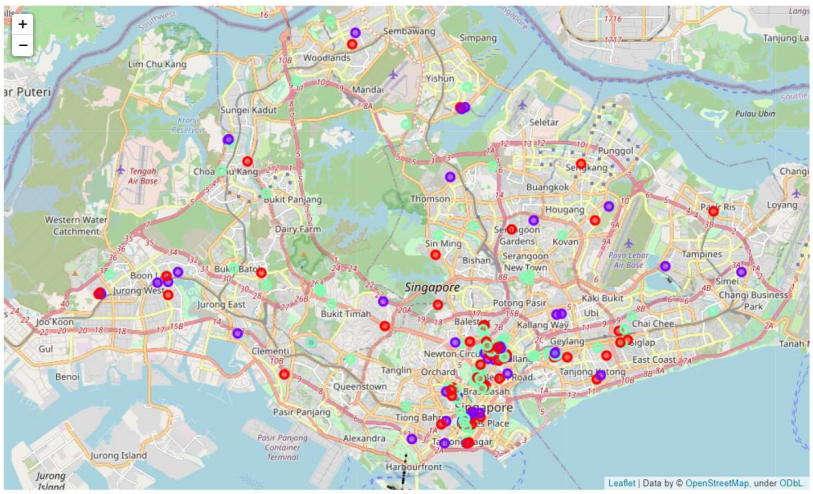
Final table used for k-means clustering in next step.

	id	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	custom_category
0	56334	1.44246	103.79667	Asia Ghani	1.437431	103.795388	Malay Restaurant	Food
1	56334	1.44246	103.79667	Kampung Admiralty Hawker Centre	1.439939	103.800774	Food Court	Food
2	56334	1.44246	103.79667	Starbucks	1.439761	103.800659	Coffee Shop	Food
3	56334	1.44246	103.79667	McDonald's	1.445931	103.798101	Fast Food Restaurant	Food
4	56334	1.44246	103.79667	KFC	1.437409	103.795428	Fast Food Restaurant	Food

3.3 One Hot Encoding and K-means Clustering

First using get_dummies, we replace the custom_categorys columns with dummies so that we transform categorical information into integer (1 or 0). Next, We use the k-means to create clusters and we set k=3. The following map displays the results.

Out[48]:



4.0 Results and Discussion

Following are the characteristic of the three clusters.

Cluster 0 (red):

- Average price $67

- Max price $104

Top Venues nearby: Food, Entertainment, SHopping

Cluster 1 (purple):

- Average price $64

- Max price $120

Top Venues nearby: Food, Entertainment, SHopping

Cluster 2 (green):

- Average price $66

- Max price $165

Top Venues nearby: Food, Shopping, SHopping

The cluster are not so well separated as maybe the radius of venue explore is set too high for the scale of Singapore at 700 and I could not manage to re-run the codes with a smaller radius. However, we can see that the prices are similar whether it is downtown or further away from city centre, all ameninties are very close be it restaurant, transport, outdoor or culture.

5. Conclusion

I set off to identify the best accommodations for families looking for short stay in Singapore.

Through compiling of data from Airbnb and venues from Foursquare, I used the K-means clustering to form 3 clusters. However, I found that the clusters are overlapping and this may be due to small size of the city state. Nevertheless, with this result, we can see that the different neighbour offers similar amenities (restaurants and malls), outdoor activities and convenience of transport. Hence, we conclude base on this finding that any cluster is suitable for families.

To further improve the analysis, we could try with a smaller radius in the parameters used in the Foursquare API to see with it, the clusters formed maybe more distinct.