BATTLE OF THE NEIGHBORHOODS - SINGAPORE

Introduction / Business Problem

A group of young entrepreneurs would like to start a new cafe business in Singapore. However, they are unsure of where to locate their new cafe. They would their cafe to cater to a wide group of people. They have heard that data science might be able to give them better insights in the location and have given me this project to use data science techniques to assist them.

Data

To approach this data science project, firstly, data regarding the neighbors in Singapore are to be retrieved from Wikipedia (https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore)

A screenshot of the relevant table is shown below:

Name (English)	Malay	Chinese	Pinyin	Tamil	Region	Area (km2)	Population[6]	Density (/km2)
Ang Mo Kio		宏茂桥	Hóng mào qiáo	ஆங் மோ கியோ	North-East	13.94	165,710	12,000
Bedok	*	勿洛	Wù luò	பிடோ	East	21.69	281,300	13,000
Bishan		碧山	Bì shān	பீஷான்	Central	7.62	88,490	12,000
Boon Lay		文礼	Wén lǐ	பூன் லே	West	8.23	30	3.6
Bukit Batok	*	武吉巴督	Wǔjí bā dū	புக்கிட் பாத்தோக்	West	11.13	144,410	13,000
Bukit Merah	*	红山	Hóng shān	புக்கிட் மேரா	Central	14.34	151,870	11,000
Bukit Panjang	*	武吉班让	Wǔjí bān ràng	பக்கிட் பஞ்சாங்	West	8.99	140,820	16,000
Bukit Timah	*	武吉知马	Wǔjí zhī mǎ	புக்கிட் டிமா	Central	17.53	77,980	4,400

Unfortunately, this Wikipedia page does not contain latitude and longitude information of the neighborhoods, which will be filled in later.

Firstly, though, the table from the Wikipedia contains information which are not required, hence columns Malay, Chinese, Pinyin and Tamil are dropped. A screenshot of the partial list of data at this stage is shown below:

[7]:

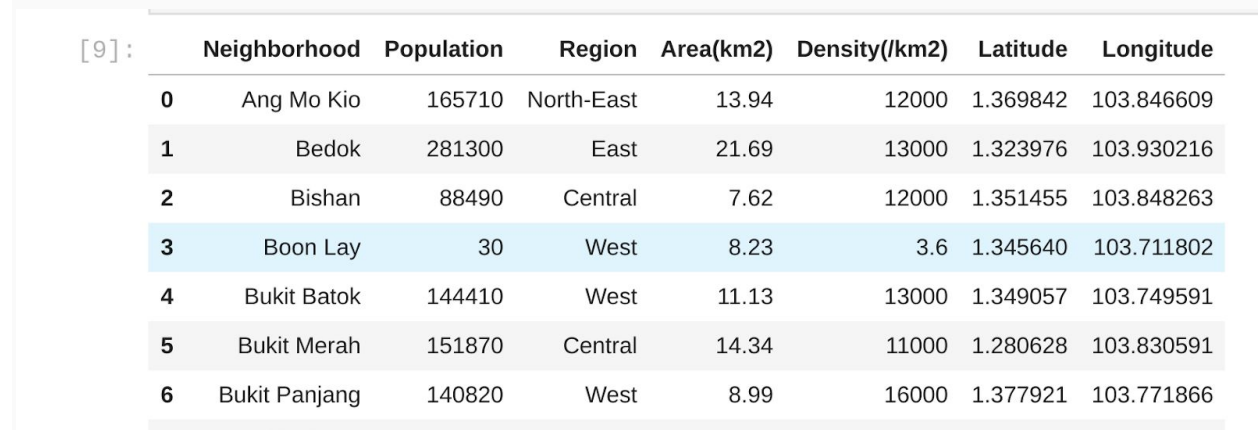|	Neighborhood	Population	Region	Area(km2)	Density(/km2)
0	Ang Mo Kio	165710	North-East	13.94	12000
1	Bedok	281300	East	21.69	13000
2	Bishan	88490	Central	7.62	12000
3	Boon Lay	30	West	8.23	3.6
4	Bukit Batok	144410	West	11.13	13000
5	Bukit Merah	151870	Central	14.34	11000

So back to the issue of missing latitude and longitude data, using nominatim from geopy.gelocator, the required data is retrieved and added to each neighborhood.

During this process, three of the neighborhoods (Downtown Core, Western Islands and Museum) were not recognized by OpenStreetMap and no latitude and longitude data were returned for these three neighborhoods. Upon further internet search, more detailed data regarding the neighborhood were obtained from the following file from the Department of Statistics Singapore website:
https://www.singstat.gov.sg/-/media/files/find_data/population/statistical_tables/tablea12-2000-2018.xls

From this file, based on the largest population in the subzones of the neighborhood, three other names were chosen (Bugis, Jurong Island and Dhoby Ghaut) to replace the three not recognized by OpenStreetMap.

A screenshot of the partial list of neighborhood with latitude and longitude data added:

	Neighborhood	Population	Region	Area(km2)	Density(/km2)	Latitude	Longitude
0	Ang Mo Kio	165710	North-East	13.94	12000	1.369842	103.846609
1	Bedok	281300	East	21.69	13000	1.323976	103.930216
2	Bishan	88490	Central	7.62	12000	1.351455	103.848263
3	Boon Lay	30	West	8.23	3.6	1.345640	103.711802
4	Bukit Batok	144410	West	11.13	13000	1.349057	103.749591
5	Bukit Merah	151870	Central	14.34	11000	1.280628	103.830591
6	Bukit Panjang	140820	West	8.99	16000	1.377921	103.771866

With the latitude and longitude data obtained for all neighborhoods, the latitude and longitude, as well as the name of the neighborhood, are then used to retrieve the location data of the venues (with a radius setting of 1000m) from Foursquare. A total of 2991 venues in 283 unique categories were retrieved.

Screenshot of the partial list of venues::

	Neighborhood	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Arcade	Art Gallery	...	Waterfall	Waterfront
0	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
1	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
2	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
3	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0
4	Ang Mo Kio	0	0	0	0	0	0	0	0	0	...	0	0

Screenshot of the dataset's shape containing all the venues:

```
[19]:  sg_onehot.shape
```

```
[19]:  (2991, 283)
```

The above dataset will be used for Machine Learning of the Neighborhoods

Methodology

Firstly, the population of each neighborhood is plotted on a Folium to gain a visual representation of the possible locations of the new cafe. The size of the bubbles represents the size of the population of the neighborhood. The Folium Map is shown below:

Figure 1: Map of Neighborhood Population
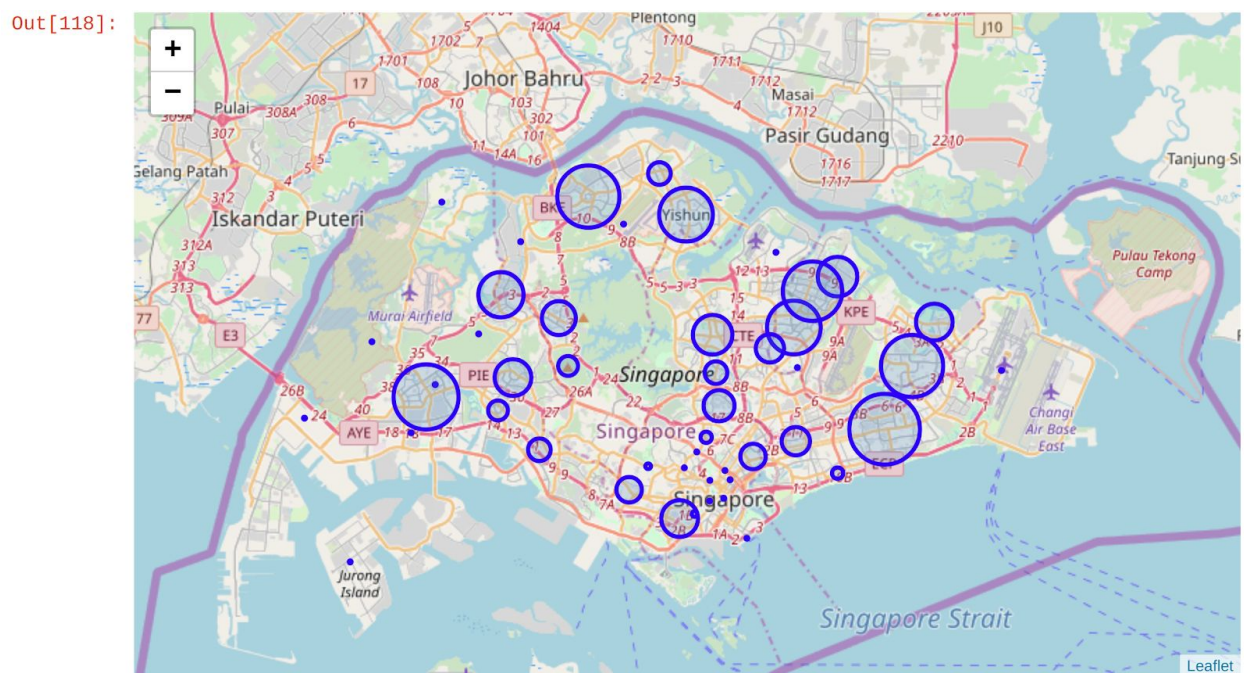
Next, K-Means Clustering is used to cluster neighborhoods into five clusters based on the similarity of venues in the neighborhoods. The resulting neighborhoods' cluster labels were then merged with the neighborhoods' latitude and longitude. A Folium map is then created to show the clusters that are similar to each other. The map is shown below:

Figure 2: Map of Neighborhood Clustering
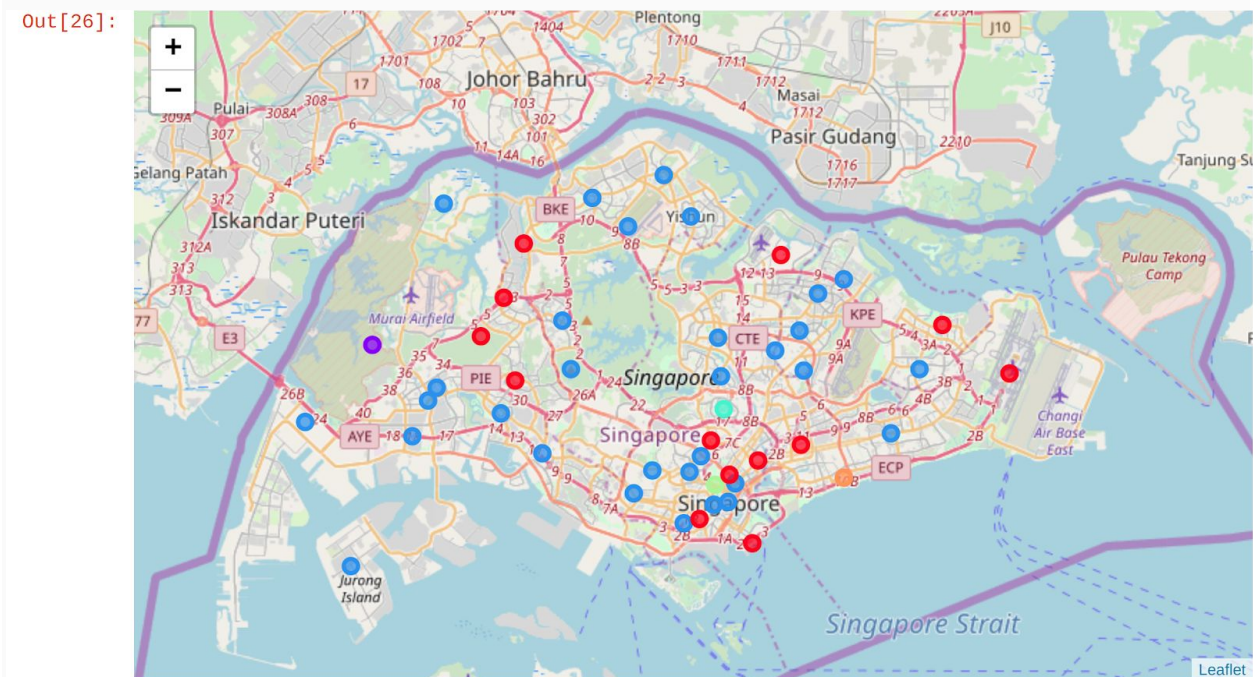
In addition, as the venues returned by Foursquare consists of all categories, while this project is regarding the location of new cafe, further data cleaning was conducted on the dataframe to remove non-food related venues. The venue categories were inspected in batches and non-food categories were removed. Part of the resulting dataframe after this cleanup is shown below:

	Neighborhood	Total Food Venues	Population	Latitude	Longitude
0	Ang Mo Kio	62	165710	1.369842	103.846609
1	Bedok	81	281300	1.323976	103.930216
2	Bishan	43	88490	1.351452	103.848250
3	Boon Lay	55	30	1.345640	103.711802
4	Bugis	74	2510	1.299476	103.855089
5	Bukit Batok	27	144410	1.349057	103.749591
6	Bukit Merah	85	151870	1.280628	103.830591
7	Bukit Panjang	28	140820	1.377921	103.771866
8	Bukit Timah	1	77280	1.354690	103.776372
9	Changi	29	2080	1.352516	103.987007
10	Choa Chu Kang	28	187510	1.389260	103.743728

With the number of food venues found for each neighborhood, another Folium Map is created to visualize this new information:

Figure 3: Map of Total Food Venues in Each Neighborhood

Out[187]:
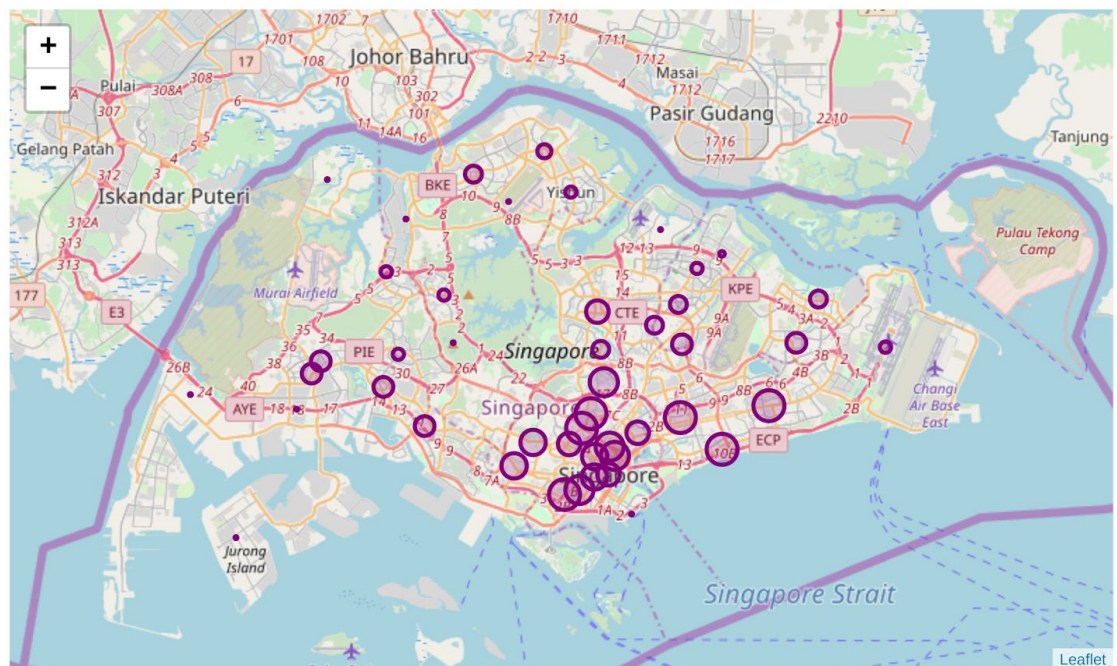
Results

From Figure 1 Population in Singapore neighborhoods, we can see that there are neighborhoods with bigger populations as shown by the bigger bubbles. The top 5 neighborhoods in terms of population are: Bedok, Jurong West,Tampines, Woodlands and Sengkang.

Table 1 below shows the Top 5 neighborhoods with the largest population

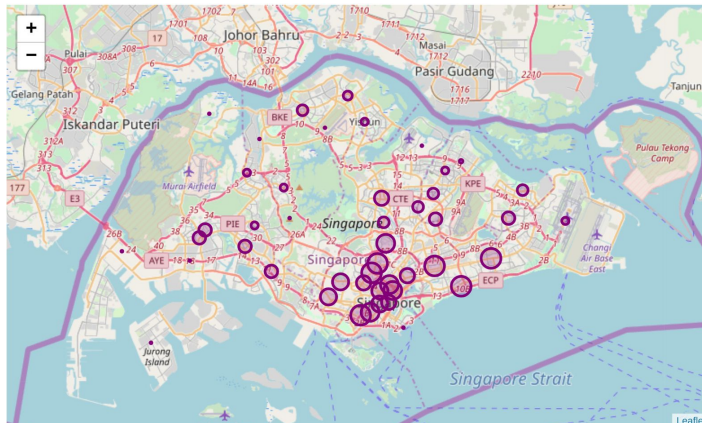| [18]:		Neighborhood	Population
1		Bedok	281300
15		Jurong West	266720
39		Tampines	257110
46		Woodlands	252530
34		Sengkang	240640

Next, from the K Means Clustering, it was found that the Top 5 most populated neighborhoods belongs to Cluster 2, indicating that they are highly similar in terms of types of venue category.

**Table 2** below shows Top 5 most populated neighborhoods in Cluster 2:

Out[28]:

	Cluster Labels	Neighborhood	Population
1	2	Bedok	281300
15	2	Jurong West	266720
39	2	Tampines	257110
46	2	Woodlands	252530
34	2	Sengkang	240640

After obtaining the total food venues in each Neighborhood, a side-by-side comparison with the population in each Neighborhood was made. Looking at the size of bubbles, we can tell that for some neighborhoods, the number of food venues looks small compared to the population.
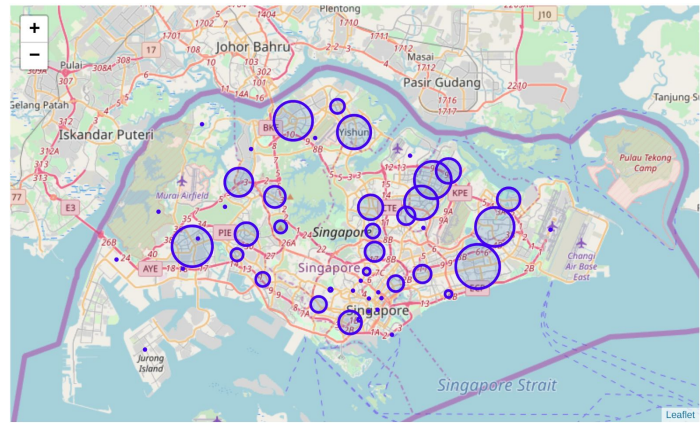
No. of food venues   Population in each Neighborhood

A calculation was then made to derive the ratio of population to food venue in each neighborhood.

Table 3 below shows this ratio at last column. Note that this table is sorted by population size of neighborhood in descending order

`Out[81]:`

	Neighborhood	Total Food Venues	Population	Latitude	Longitude	Pop_to_Food_Ratio
1	Bedok	81	281300	1.323976	103.930216	3472
17	Jurong West	50	266720	1.339636	103.707339	5334
40	Tampines	51	257110	1.354653	103.943571	5041
44	Woodlands	43	252530	1.436897	103.786216	5872
35	Sengkang	31	240640	1.390949	103.895175	7762

Discussion.

Through the analysis and visualization of the dataset, it was observed that there are several neighborhoods with considerably larger populations than the rest. These neighborhoods are namely: Bedok, Jurong West, Tampines, Woodlands and Sengkang. Since the clients would like to target a wide group of customers, these five neighborhoods would have to be in their shortlist of locations.

At the same time, after clustering based on similarity of venue categories, it was found that the same five neighborhoods also belong to the same cluster. Thus, any of the five neighborhoods would be highly similar. As the venues in the neighborhoods are there to provide service to the population, it can be inferred that the population in these neighborhoods are highly similar as well.

Next, through the analysis of the ratio of population to the number of food venues (Table 3), we can see Sengkang has highest ratio amongst the five neighborhoods. With more people than food venues, in comparison with other neighborhoods, it could mean that opening a cafe that would attract a large crowd.

Conclusion

Based on the findings of the exploratory data analysis and clustering, we conclude that the top five neighborhood that we would recommend to the clients are: Bedok, Jurong West, Tampines, Woodlands and Sengkang. In particular, we would also highlight that Sengkang could be the best of the five as it has the largest population to food venues ratio.