@Author: Galvingwc

Introduction

A Singaporean friend of mine, who has just graduated from university, and is looking to venture into entrepreneurship. Having graduated with a degree in sports science, he plans to open a supplement drinks store as his first local start-up business. However, he could not decide on a location to open his store. Hence, he approached me for help, knowing that I have experience in data science and is hopeful of me using data to locate the best possible location for his store.

Data

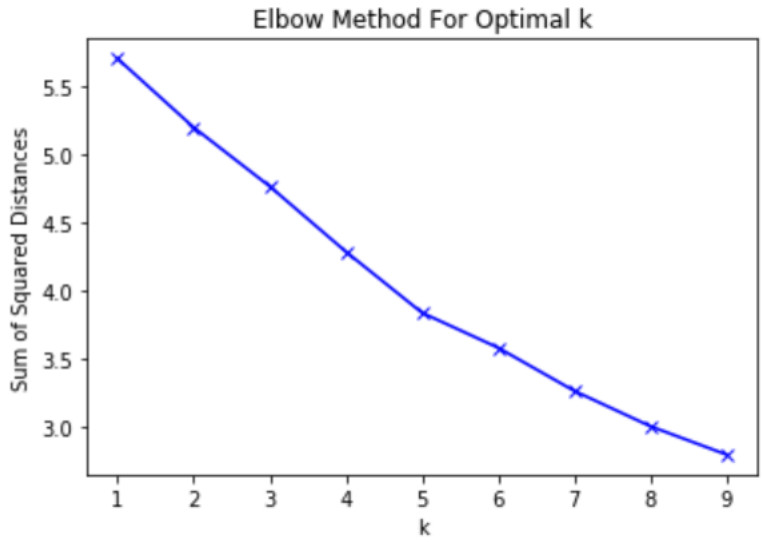
The data set that will be used contains the name of Singapore's neighbourhoods, and their geospatial coordinates. Subsequently, the Foursquare API will be utilised to obtain location and categorical data of nearby venues around the neighbourhoods. The data will be analysed, and the neighbourhoods will be grouped into clusters based on the mean frequency of occurrence of nearby venues by category. Hence, this helps to visualise the clusters characteristics. Noting that the target customers for a supplement drinks store would be fitness enthusiasts, the ideal location would be in clusters which has the highest mean frequency of fitness centres (complementary business) and lowest mean frequency of coffee shops/restaurants (competitive business).

Methodology

1. Import relevant libraries
2. Import data frame from Singapore_Neighbourhoods.csv using pandas.
3. Create a map of Singapore with pop up markers at all neighbourhoods using folium.
4. Create and run function to obtain nearby venues' information using the Foursquare API.
5. Create new data frame of nearby venues using one-hot encoding.
6. Group data frame by neighbourhood, with features containing mean of venue occurrence.
7. Create function to sort and return the mean of venue occurrence in descending order.
8. Create new data frame of top 10 nearby venues by neighbourhood.
9. Find optimal value of k for k-means using the elbow method.
10. Initiate k-means clustering using KMeans.
11. Merge and clean data frame with neighbourhood, cluster label and top 10 nearby venues.
12. Create a map of Singapore with pop up markers of different colours representing different clusters using folium.
13. Examine data frame of the different clusters to determine unique characteristic.

Results

Graph of k against sum of squared distance:



Folium map of 5 different clusters in Singapore:

Cluster 1:

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Ang Mo Kio	0.0	Food Court	Dessert Shop	Coffee Shop
1	Bedok	0.0	Chinese Restaurant	Coffee Shop	Café
2	Bishan	0.0	Food Court	Coffee Shop	Bubble Tea Shop
3	Boon Lay	0.0	Exhibit	Zoo Exhibit	Fishing Spot
4	Bukit Batok	0.0	Coffee Shop	Chinese Restaurant	Fast Food Restaurant

Cluster 2:

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
39	Simpang	1.0	Café	Gay Bar	Zoo Exhibit

Cluster 3:

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
8	Central Water Catchment	2.0	Trail	Ethiopian Restaurant	Food Truck

Cluster 4:

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9	Changi	3.0	Gym / Fitness Center	Ski Chalet	Ethiopian Restaurant

Cluster 5:

	Neighbourhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
17	Kallang	4.0	Coffee Shop	Chinese Restaurant	Food Court
31	Punggol	4.0	Bus Station	Food Court	Coffee Shop
49	Western Islands	4.0	Coffee Shop	Food Court	Zoo Exhibit
50	Western Water Catchment	4.0	Coffee Shop	Bus Station	Café

Discussion

To determine the optimal value of k for the k-means clustering algorithm, a range of values of k is used to iteratively run the algorithm. The sum of squared error for each value of k can then be calculated and a graph of the values of k can be plotted against sum of squared errors. Using the elbow method, the optimal value of k can be observed to be the sharp point joining 2 lines, which is k=5.

Using k=5 for the k-means clustering algorithm, a folium map displaying the 5 different clusters of neighbourhoods in Singapore can be created. From the map, it can be observed that majority of the neighbourhoods fall under the red cluster (cluster 1). The remaining clusters are the purple, blue, teal and orange cluster (cluster 2, 3, 4 and 5) which contains 1, 1, 1, 4 neighbourhoods respectively.

By extracting out the different cluster's data frames, the unique characteristic of each cluster can be determined upon closer inspection. After inspecting top 3 nearby venues for each neighbourhood in the 5 clusters, they can be grouped into:

Cluster 1: Food court/Restaurant

Cluster 2: Café/Bar

Cluster 3: Park/Trail

Cluster 4: Fitness/Recreation

Cluster 5: Coffee shop/Bus station

Based on these groupings and keeping the business problem in mind, the best cluster to locate the supplement drinks store can finally be decided. Since cluster 1, 2 and 5 contains many venues with competing businesses, it is only wise to avoid them. The remaining cluster 3 and 4 are both suitable and complementary to the drinks business. However, it can be noted that cluster 3 contains parks and nature trails whereas cluster 4 contains gyms and fitness centres. With target customers who are fitness enthusiasts, the most ideal cluster to locate a supplement drinks store would be cluster 4. Furthermore, nature parks and trails more isolated compared to gyms and fitness centres which are more accessible. Hence, since cluster 4 contains only 1 neighbourhood, it can be decided that the best location to locate the supplement drinks store would be in Changi.

Conclusion

By utilising the available data sets online and location data from Foursquare API, data analysis can be done using machine learning algorithms. This reduces the risk of decision-making and increases the rate of success. Data analysis for decision-making can be applied to many other fields and are not limited to retail businesses. Hence, data analysis is an important and beneficial skill to have.

Appendix

Singapore_Neighbourhoods.csv:

Neighbourhood	Latitude	Longitude
Ang Mo Kio	1.37173	103.8476
Bedok	1.32466	103.9324
Bishan	1.3504	103.8487
Boon Lay	1.3238	103.7061
Bukit Batok	1.35002	103.7493
Bukit Merah	1.28265	103.8187
Bukit Panjang	1.38006	103.7643
Bukit Timah	1.34242	103.7766
Central Water Catchment	1.35586	103.7953
Changi	1.35558	103.975
Choa Chu Kang	1.38511	103.745
Clementi	1.31435	103.7652
Downtown Core	1.28944	103.85
Geylang	1.32061	103.8869
Hougang	1.37258	103.8937
Jurong East	1.32949	103.7383
Jurong West	1.35223	103.7113
Kallang	1.3241	103.8705
Lim Chu Kang	1.44399	103.7247
Mandai	1.42066	103.8167
Marina East	1.29517	103.8712
Marina South	1.23776	103.8521
Marine Parade	1.30128	103.9049
Museum	1.28944	103.85
Newton	1.31368	103.8362
Novena	1.31696	103.8442
Orchard	1.30257	103.8347
Outram	1.284	103.8426
Pasir Ris	1.37244	103.9496
Paya lebar	1.3463	103.8996
Pioneer	1.32514	103.6794
Punggol	1.39849	103.9079
Queenstown	1.29883	103.804
River Valley	1.2939	103.8353

Rochor	1.30544	103.8539
Seletar	1.409	103.8816
Sembawang	1.4482	103.8195
Sengkang	1.38931	103.8995
Serangoon	1.35345	103.8724
Simpang	1.44435	103.8427
Singapore River	1.28854	103.8493
Southern Islands	1.23776	103.8521
Straits View	1.27978	103.8535
Sungei Kadut	1.40912	103.7509
Tampines	1.35248	103.9446
Tanglin	1.29908	103.8164
Tengah	1.3575	103.7315
Toa Payoh	1.33146	103.8495
Tuas	1.33645	103.6471
Western Islands	1.32863	103.6687
Western Water Catchment	1.34218	103.6785
Woodlands	1.43605	103.7861
Yishun	1.42972	103.8359