Coursera Applied Data Science Capstone Project

The Battle of Neighborhood: Where to start the business in Guangzhou?

Author: Ziqing Xu

Date: Oct 29, 2020

1. Introduction

In our lives, people are diverse. Some people prefer a cozy lifestyle and choose to work for someone else. However, some people are passionate about challenges, and they start their own businesses to follow their dreams and fulfill their passion. Types of entrepreneurship are diverse, such as opening restaurants, coffee shops, IT companies and so on. Then, a significant problem arises, namely, in which specific location to recommend entrepreneurs to start their businesses. For example, if you want to open a restaurant in Toronto, it is recommended that you open a restaurant on Yonge St, as restaurants are clustered in that area.

The goal of the project is to segment and cluster the neighbourhoods by exploring and comparing the neighbourhoods. By analyzing the clusters, we can figure out the best-recommended location to start a specific type of business. As my hometown is Guangzhou, in this project, we will focus on the neighbourhoods in Guangzhou. Definitely, we can also apply this application to other cities.

2. Data

2.1. Neighbourhood data of Guangzhou extracted from: https://en.wikipedia.org/wiki/List_of_township-level_divisions_of_Guangdong

2.1.1. districts information

2.1.2. subdistricts information

2.2. Geospatial Coordinates of Guangzhou Neighbourhoods will be obtained by using the geocode package. Geospatial Coordinates data is important for getting venues data.

2.3. Venues data of each Neighbourhood will be retrieved by using Foursquare API

3. Methodology

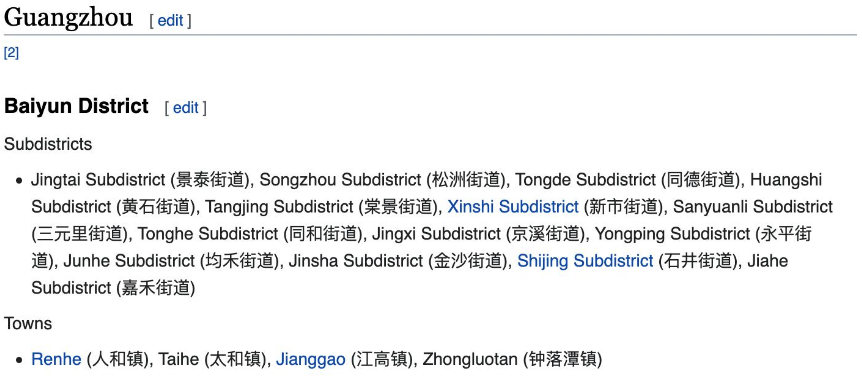
3.1. Data Requirement

The basic requirement of a neighbourhood should contain its borough, city, latitude, longitude.

3.2. Data Collection

Guangzhou’s neighbourhood data can be collected from Wikipedia, as shown in figure 1. Using `requests` library, a HTML file is returned by sending a HTML request to that specific Wikipedia link, listed in the Data section. Then, `BeautifulSoup` library is needed to extract useful information. The neighbourhood data is written into a csv file. Then, the next step is to transform the extracted data into pandas dataframe, as shown in figure 2.

Figure 1. Neighbourhood Data shown in Wikipedia (District is equivalent to Borough; Sub-districts and Towns are equivalent to Neighbourhoods)



Guangzhou [edit]

[2]

Baiyun District [edit]

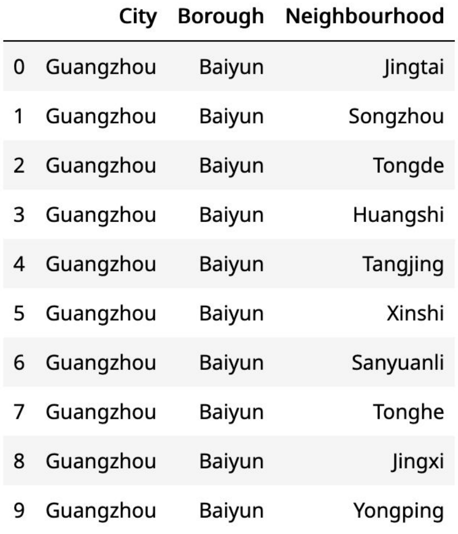
Subdistricts

- Jingtai Subdistrict (景泰街道), Songzhou Subdistrict (松洲街道), Tongde Subdistrict (同德街道), Huangshi Subdistrict (黄石街道), Tangjing Subdistrict (棠景街道), Xinshi Subdistrict (新市街道), Sanyuanli Subdistrict (三元里街道), Tonghe Subdistrict (同和街道), Jingxi Subdistrict (京溪街道), Yongping Subdistrict (永平街道), Junhe Subdistrict (均禾街道), Jinsha Subdistrict (金沙街道), Shijing Subdistrict (石井街道), Jiahe Subdistrict (嘉禾街道)

Towns

- Renhe (人和镇), Taihe (太和镇), Jianggao (江高镇), Zhongluotan (钟落潭镇)

Figure 2. Neighbourhoods in dataframe

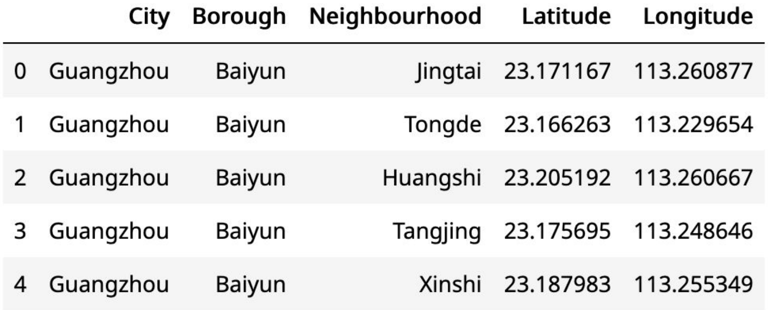


	City	Borough	Neighbourhood
0	Guangzhou	Baiyun	Jingtai
1	Guangzhou	Baiyun	Songzhou
2	Guangzhou	Baiyun	Tongde
3	Guangzhou	Baiyun	Huangshi
4	Guangzhou	Baiyun	Tangjing
5	Guangzhou	Baiyun	Xinshi
6	Guangzhou	Baiyun	Sanyuanli
7	Guangzhou	Baiyun	Tonghe
8	Guangzhou	Baiyun	Jingxi
9	Guangzhou	Baiyun	Yongping

3.3. Data Understanding and Analysis

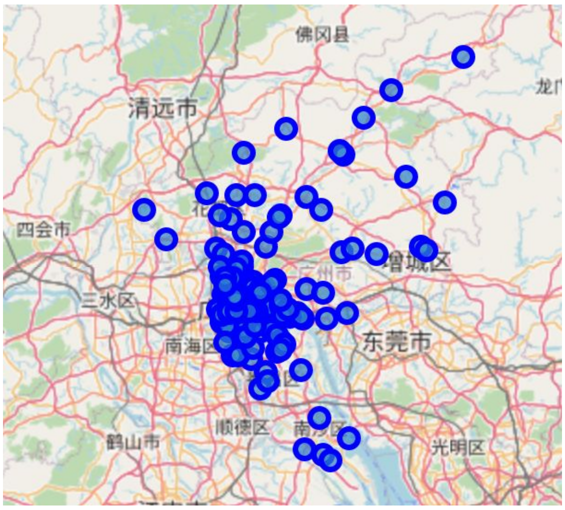
3.3.1. Geospatial data for each neighbourhood is missing for further exploration. The solution is to use geopy library to get geospatial data for each neighbourhood, as shown in figure 3. Then, with the geospatial data, neighbourhoods can be plotted in the map, as shown in figure 4.

Figure 3.Neighbourhood with Geospatial data



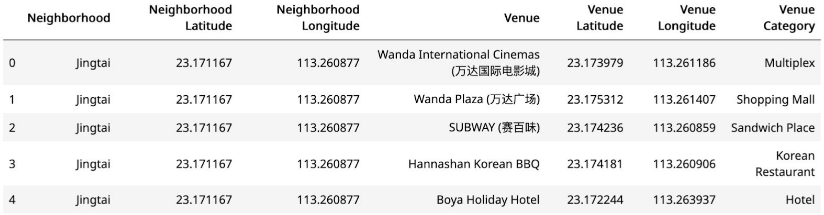
	City	Borough	Neighbourhood	Latitude	Longitude
0	Guangzhou	Baiyun	Jingtai	23.171167	113.260877
1	Guangzhou	Baiyun	Tongde	23.166263	113.229654
2	Guangzhou	Baiyun	Huangshi	23.205192	113.260667
3	Guangzhou	Baiyun	Tangjing	23.175695	113.248646
4	Guangzhou	Baiyun	Xinshi	23.187983	113.255349

Figure 4. Map of Guangzhou



3.3.2. Venues information nearby each neighbourhood is required to make a clustering. In this project, the Foursquare API was used to search for the nearby venues of each neighbourhood in a radius of 500 meters. Only venue name and venue category are extracted, as shown in figure 5.

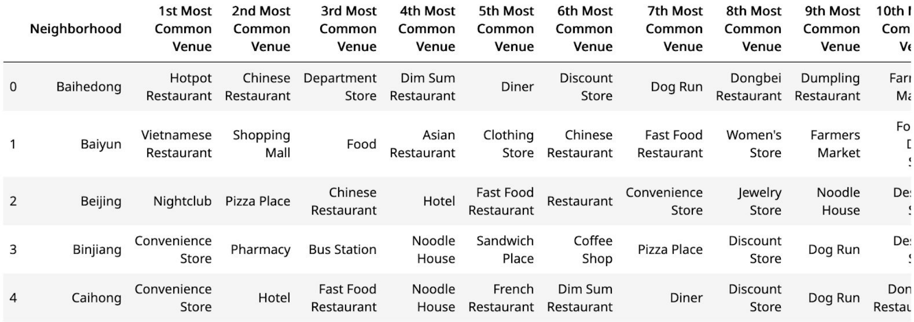
Figure 5. Nearby venues of each neighbourhood



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Jingtai	23.171167	113.260877	Wanda International Cinemas (万达国际电影城)	23.173979	113.261186	Multiplex
1	Jingtai	23.171167	113.260877	Wanda Plaza (万达广场)	23.175312	113.261407	Shopping Mall
2	Jingtai	23.171167	113.260877	SUBWAY (赛百味)	23.174236	113.260859	Sandwich Place
3	Jingtai	23.171167	113.260877	Hannashan Korean BBQ	23.174181	113.260906	Korean Restaurant
4	Jingtai	23.171167	113.260877	Boya Holiday Hotel	23.172244	113.263937	Hotel

3.3.3. Based on the dataset shown in figure 5, creating a new table with top 10 venues for each neighbourhood is required for modeling, as shown in figure 6.

Figure 6. top 10 venues

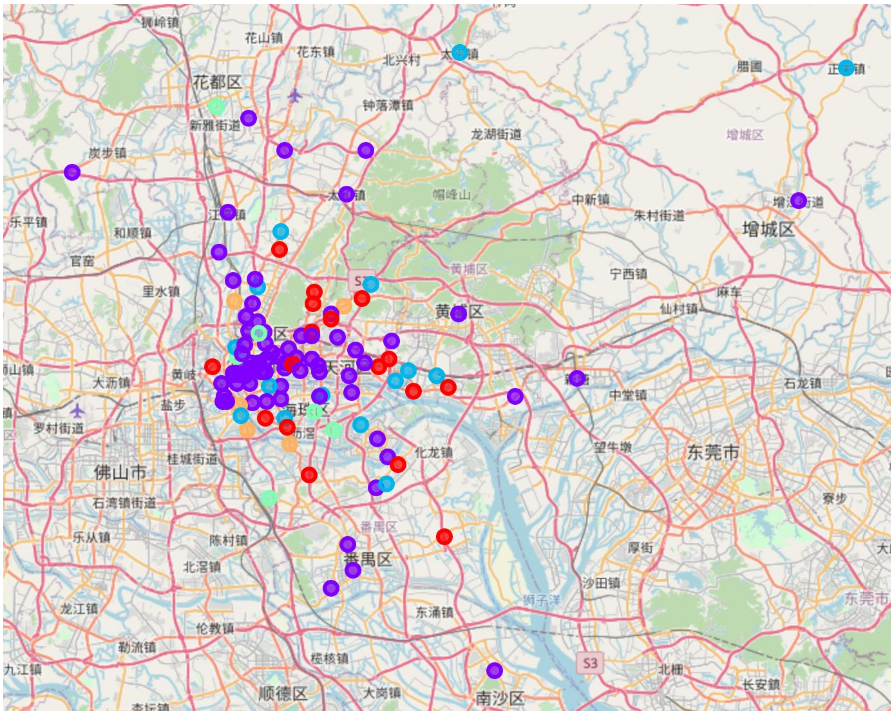


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th I Com Ve
0	Baihedong	Hotpot Restaurant	Chinese Restaurant	Department Store	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Dongbei Restaurant	Dumpling Restaurant	Fari Ma
1	Baiyun	Vietnamese Restaurant	Shopping Mall	Food	Asian Restaurant	Clothing Store	Chinese Restaurant	Fast Food Restaurant	Women's Store	Farmers Market	Fo D S
2	Beijing	Nightclub	Pizza Place	Chinese Restaurant	Hotel	Fast Food Restaurant	Restaurant	Convenience Store	Jewelry Store	Noodle House	De: S
3	Binjiang	Convenience Store	Pharmacy	Bus Station	Noodle House	Sandwich Place	Coffee Shop	Pizza Place	Discount Store	Dog Run	De: S
4	Caihong	Convenience Store	Hotel	Fast Food Restaurant	Noodle House	French Restaurant	Dim Sum Restaurant	Diner	Discount Store	Dog Run	Don Restau

3.4. Modeling and Evaluation

Since the goal is to determine the location to start up a specific type of business, K-MEANS could be a good option. K-MEANS will divide all neighbourhoods into K clusters based on venue categories. In each cluster, neighbourhoods are similar to each other and dissimilar to objects in other clusters. We can analyze the data of each cluster to get insight of the pattern of venues. Then, we could recommend a location for a specific type of business. In the project, I choose K to be 5.

Figure 7. Clusters



4. Results

4.1. Number of Neighbourhoods in Each Clusters

Cluster #	# of neighbourhoods
1	17
2	71
3	15
4	8
5	5

4.2. Description of Each Cluster (Distribution is shown in Figure 7)

4.2.1. Cluster 1 (Red): In this cluster, restaurants (Chinese, Dim Sum, Cantonese, Fast Food), metro station, and stores(shopping mall, convenience store) are most recommended.

4.2.2. Cluster 2 (Purple): In this cluster, the types of business are diverse. Based on the result, it is likely to be an area for entertainment. Cafe, Restaurant, Shopping Mall, Resort, Park, and Sport Court are recommended.

4.2.3. Cluster 3 (Blue): In this cluster, Opening a Chinese Restaurant is most recommended as the first most common venues are Chinese Restaurants for most of the neighbourhood in this cluster. Department Store and Women's Store are also recommended.

4.2.4. Cluster 4 (Mint Green): In this cluster, Opening a hotel is most recommended as the first most common venues are hotels for most of the neighbourhood in this cluster. From the result, the neighbourhoods in this cluster are likely to be areas for tourism. Starting up a tourism-relative business is recommended.

4.2.5. Cluster 5: In this cluster, Opening a Restaurant is most recommended as the most of the common venues are about dining. However, Cantonese Food is most popular.

5. Discussion

As we see in the result section, the number of neighbourhoods in cluster 2 is much more than in the other clusters. The type of venues are diverse in cluster 2, so it is not easy to make a recommendation in the area covered by cluster 2. Using a bigger K may get a better result.

6. Conclusion

Overall, as demonstrated in the result section, we can effectively make a recommendation for entrepreneurs to start up different types of business. I hope this can be applied to different cities worldwidely.