Metro Stations of Shanghai

Hengli Li

June 18, 2020

1. Introduction

1.1 Background

Shanghai, the economic center of China, is one of the well-known cities in the world. As a developed city, Shanghai is blessed with 19 metro lines and 413 stations. As a newly born city, millions of people come in and out from Shanghai in one year. How to look for a suitable home is essential for new-comers, since home is the place where people spend half of their time, also the place where people would like to have a rest. In the meanwhile, new-comers are faced with the problem that Shanghai are new to them, even some people cannot figure out the specific administrative regions of Shanghai. Therefore, it is advantageous for this project to accurately point out every metro station and their surroundings.

1.2 Problem

Data that might contribute to detecting surroundings might include the definite latitude and longitude of metro station, intending to find out the concrete venues, restaurants, shopping malls, and entertainment that enrich the daily life of residents. The aim of this project is to look at the kinds of venues surrounding the metro stations and classify them based on the types of venues near a station the most, as well as taking into consideration the geographic location of the station.

1.3 Interest

As it described before, the target customer of this project is the one who searches for a new home and wants to get well known about one area. Also, this project is able to help visitors to get familiar with Shanghai, and find the comfortable hotels.

2. Data acquisition

2.1 Shanghai Metro Stations of line1

The final data frame used for data analysis contains the line 1 of Shanghai's metro stations' data. The list of all the metro stations we used were retrieved from the Wikipedia page:https://wanweibaike.com/wiki-%E4%B8%8A%E6%B5%B7%E5%9C%B0%E9%93%81%E7%AB%99%E7%82%B9%E5%88%97%E8%A1%A8#1 号线. As shown in the below:

	Number	Station_Name	Opened	Location	Platform_Level	Platform_Type	Transfers
0	1	莘庄	1996年12月28日	Minhang	At-grade	Side platform	5
1	1	外环路	1996年12月28日	Minhang	At-grade	Side platform	NaN
2	1	莲花路	1996年12月28日	Minhang	At-grade	Side platform	NaN
3	1	锦江乐园	1996年12月28日	Xuhui	At-grade	Side platform	NaN
4	1	上海南站	2004年10月30日	Xuhui	Underground	Island platform	3

2.2 Data frame with Latitude and Longitude

The data frame which was retrieved through URL is concise and there is no need to mollify the content. So the next step of this project is to coordinate the latitude and longitude to each row, by looping through the whole list and creating custom Baidu API queries for each row from their cell values. We then save the data frame into a .csv file, so that we can use it for repeated testing and data classification, without having to call the

Baidu API each time. (The coordinate data frame is shown in the below.)

	Number	Station_Name	Opened	Location	Platform_Level	Platform_Type	Transfers	Longitude	Latitude
0	1	莘庄	1996年12月28日	Minhang	At-grade	Side platform	5	121.392186	31.116872
1	1	外环路	1996年12月28日	Minhang	At-grade	Side platform	NaN	121.399614	31.126649
2	1	莲花路	1996年12月28日	Minhang	At-grade	Side platform	NaN	121.409334	31.136734
3	1	锦江乐园	1996年12月28日	Xuhui	At-grade	Side platform	NaN	121.415479	31.145542
4	1	上海南站	2004年10月30日	Xuhui	Underground	Island platform	3	121.435865	31.159439

2.3 Data frame with Foursquare

Using the coordinates, we got from querying with the Baidu API, we can query locations and any nearby venues with the Foursquare Places API. We query each station's location and collect the number of venues, sorted by the top-level categories available from the Foursquare API. We can then classify each station based on the top categories of the total number of nearby venues around a station. Clusters of station locations and their surrounding areas can then be marked on a map of Shanghai and inform users about that part of the city. (Coordinate data frame with Foursquare data)

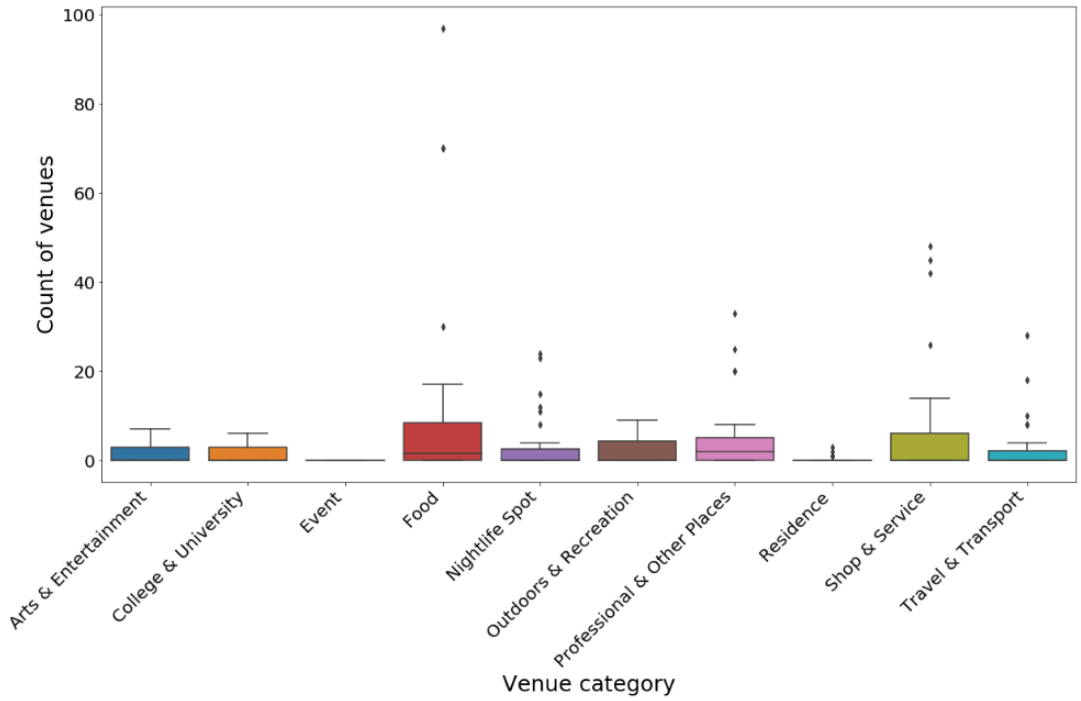
	Number	Station_Name	Opened	Location	Platform_Level	Platform_Type	Transfers	Longitude	Latitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spo
0	1	莘庄	1996年12月28日	Minhang	At-grade	Side platform	5	121.392186	31.116872	0	0	0	1	0
1	1	外环路	1996年12月28日	Minhang	At-grade	Side platform	NaN	121.399614	31.126649	0	0	0	2	1
2	1	莲花路	1996年12月28日	Minhang	At-grade	Side platform	NaN	121.409334	31.136734	0	0	0	6	2
3	1	锦江乐园	1996年12月28日	Xuhui	At-grade	Side platform	NaN	121.415479	31.145542	0	0	0	4	1
4	1	上海南站	2004年10月30日	Xuhui	Underground	Island platform	3	121.435865	31.159439	0	1	0	3	0

3. Exploratory Data Analysis

3.1 Boxplot of Venue Category

Scrapping data from Foursquare, I can find that there are ten categories of venue,

including Arts & Entertainment, College & University, Event, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service, Travel & Transport. To visualize the results of data frame, I choose box plot to vividly displace the distribution of different venue category. (The box plot is shown in the below.)

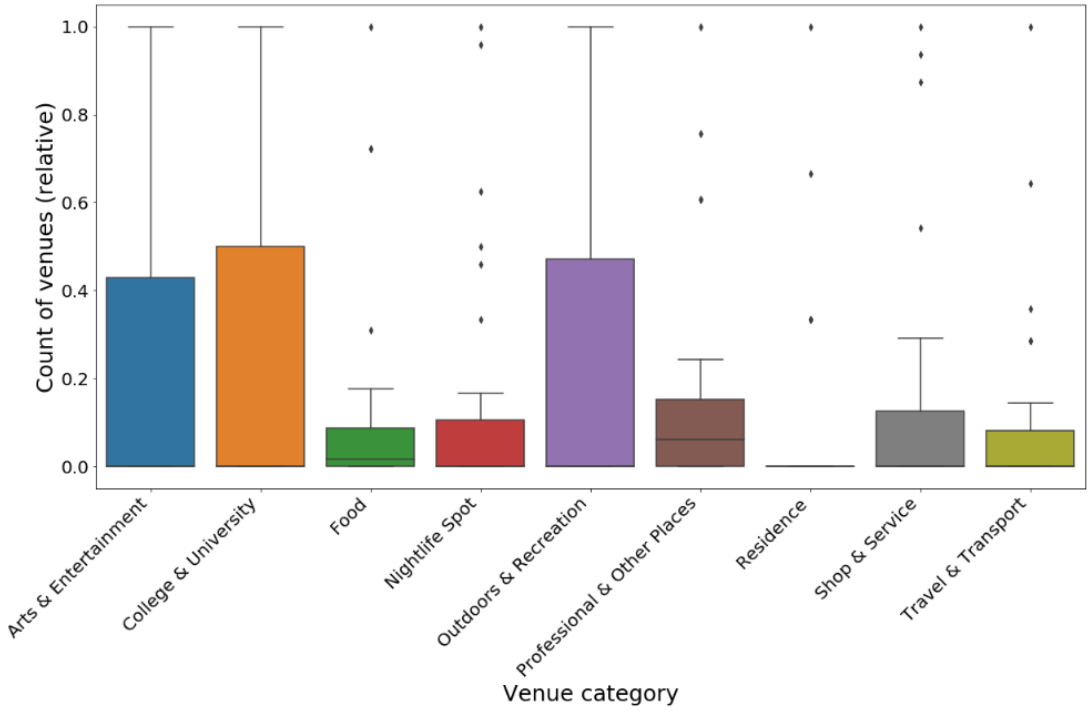


3.2 Normalize Data

It looks like the most frequent venue categories are Shop&Service, Professional&OtherPlaces, Travel&Transport, and Food. Event has very little data, so let's discard it from both the data frame and the list of categories. Normalize the data using MinMaxScaler (scale from 0 to 1), this scales the data and provides an easy to interpret score at the same time and separate the columns to be normalized from the rest of the data :

	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	0.0	0.000000	0.010309	0.000000	0.111111	0.030303	0.0	0.000000	0.035714
1	0.0	0.000000	0.020619	0.041667	0.111111	0.151515	0.0	0.041667	0.035714
2	0.0	0.000000	0.061856	0.083333	0.000000	0.090909	0.0	0.083333	0.000000
3	0.0	0.000000	0.041237	0.041667	0.000000	0.090909	0.0	0.062500	0.035714
4	0.0	0.166667	0.030928	0.000000	0.000000	0.121212	0.0	0.062500	0.071429

Using boxplot to show the normalized data:



4. Methodology--K-Means Clustering

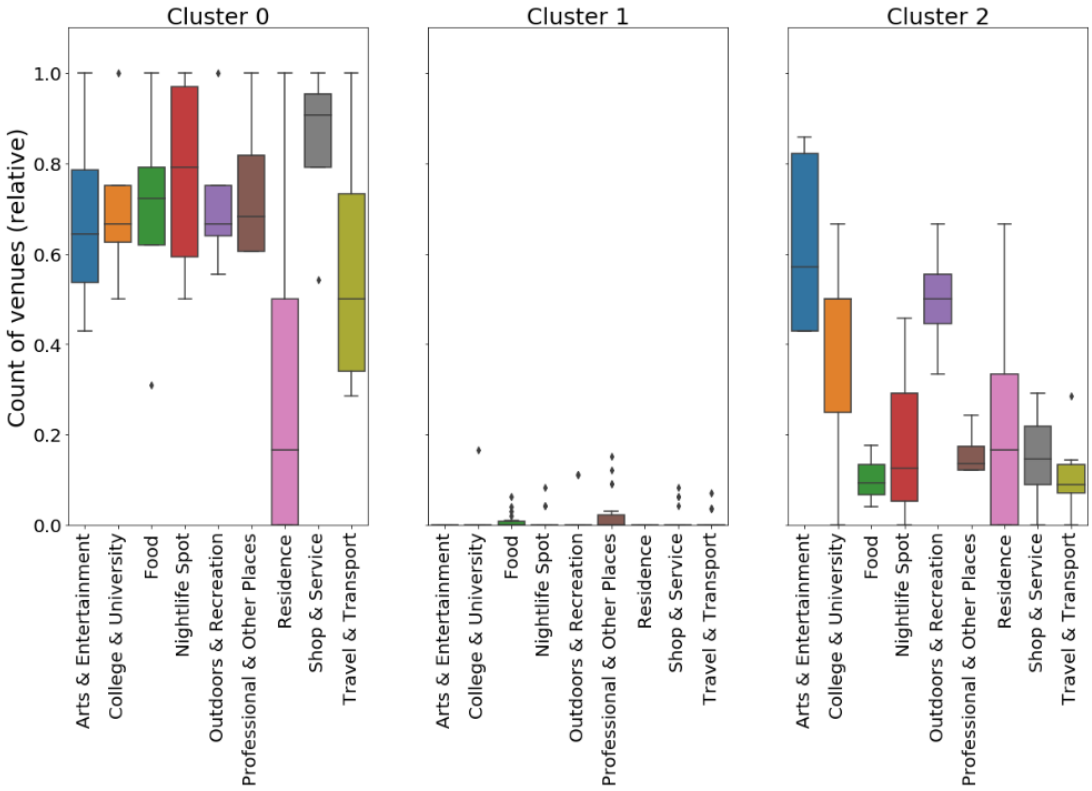
The classifier of project will use K-Means Clustering. Using different numbers of clusters, the initial results can be seen below :

2 clusters divided the area into just the downtown central area and the outer city surrounding area;

3 clusters yield the most intuitive result consisting of high density areas, medium venue density areas, and low density suburbs;

4 or more clusters are difficult to interpret, or need a more in-depth analysis to explain the cluster results;

For the scope of this class project, we will use 3 clusters in our analysis. Then use boxplots to view the classified clusters :



5. Conclusion

5.1 Results

We can briefly summarize each of our classified clusters by looking at the boxplot showing the normalized values for the venues nearby each group of stations :

Cluster 0 has the highest number of venues nearby, especially for Shop&Service, Nightlife Spot, and Travel&Transport;

Cluster 1 on average has the least number of venues near its stations, and appears as the lowest density area;

Cluster 2 has the lowest number of Residence venues, and is between the other 2 clusters in nearby venue density.

After coloring and plotting most of the stations on a map of the Shanghailine1 Metropolitan Region, we can see that :

Cluster 0 most likely has the highest number of people passing by and creating venues and check-ins, as they are in the densely populated areas of the city (offices and department stores i.e. Xujiahui);

Cluster 1 marks stations that are not in areas as developed as in the other two clusters;

Cluster 2 seems to mark stations where there are populated by different universities and colleges.

The final map is still quite informative as stations nearby were rendered and appear to be correctly classified, so for our project's scope (looking at the area as a whole) this is still deemed quite satisfactory.

5.2 Discussion

There are some factors to consider when analyzing the results of this data science project. First, using the Foursquare database to get the number of venues around each station can make our results a bit biased towards the Food and Travel&Transport categories, as these 2 types of locations are the most commented and checked-into places

(see the paper at

https://www.researchgate.net/publication/261060627_Exploring_venue_popularity_in_Foursquare for more details). The significance of a location or building also is not shown, so some key landmarks or important areas might not be highlighted. However, with the main theme being density and having users being able to click and reveal the top 3 categories of each rendered location on the map, we were able to answer the questions and challenges asked in the beginning of the project. Users can use the interactive map of Shanghai's Metro to find out more about the surroundings of each station by clicking on a circled area, and a popup will inform them about the top 3 types of venues around the station. By familiarizing themselves with the color scheme of the map (blue for high-density, white for medium-density, red for sparsely-dense areas), users can view the overall status of the Shanghai Metropolitan Region.

5.3 Conclusion

We have shown how to use the Baidu API, the Foursquare Places API, and the Python Folium library to retrieve the locations and nearby number of venues around each of Shanghai's line1 metro stations, and plot most of them onto an interactive map of the Shanghai Metropolitan Region. The data collected can be useful to others in the future in other areas of research, especially if combined with more data from other sources, such as social media feeds or census data.