Shanghai Data Analysis

Coursera Capstone Project

Sadasivam Natanakumar

1. Introduction

Shanghai, the most populous city in China, is a perfect amalgamation of Chinese and international cultures. It attracts over 300 million visitors a year, both from within China and outside of it. Businessmen, tourists, students, there are all kinds of people there. The aim of the project is to create a catalogue for people visiting Shanghai to select Hotels to stay at based on their distance from the airport and most frequently visited locations in Shanghai. The project will be using data acquired from Foursquare API services and coded in python to produce the required results.

A view of the streets of Shanghai

Problem Statement/Background Discussion

Travel websites focus on the features within hotels while making suggestions to users. However, people when travelling want to explore the local culture, their cuisine, while also maintaining a certain comfort zone and keeping logistics in mind. This is the area of focus for my project.

The system in this notebook will provide the following use case scenario:

- A person planning to visit Shanghai as a tourist or on a business trip and looking for accommodation.
- For people who want ease of accessibility to the venues flocked by locals.
- To also combine the location of the hotel (distance from airport and metros) and its proximity to venues such as cafes, restaurants and spas.

The view from the Ritz-Carlton in Pudong, Shanghai

To carry out the said objectives, my project makes use of:

- Hotels acquired from Foursquare data.
- Popular venues in the vicinity of the hotels, also acquired from Foursquare data.
- Distance from Airport, using data from open sources.

While using this catalogue, the user has to keep in mind that this is focused on finding the best possible hotels, finding the most frequented venues in the city, and combining the two to give an idea of the best places to stay at that allow you to experience the culture of the city.

2. Data Acquisition

External data was required for the following parts of the project:

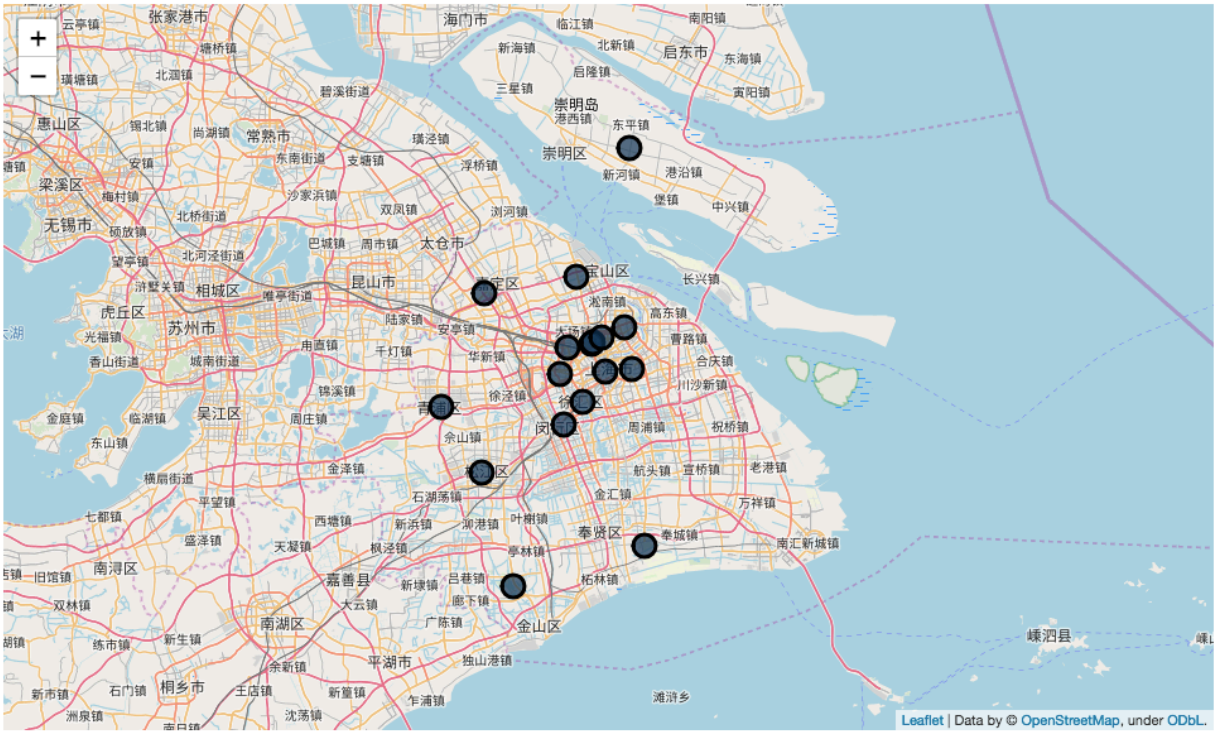
- Postcodes of the districts in Shanghai City.
- Coordinates of the districts in Shanghai City.
- Details about the hotels in Shanghai City.
- Details about the venues in Shanghai City.

Postcodes and coordinates of the districts:

This was done using data from the travelchinaguide.com website which housed a table with the details of the names of districts in Shanghai along with their postcodes. By using simple web scraping techniques with the help of and libraries the data from the table was converted to a dataframe. The coordinates of the districts were found with the help of the library, which were then added to a corresponding dataframe for further reference.

	Zip Code	District/County	Latitude	Longitude
0	200000	Huangpu District	31.218228	121.480302
1	200000	Xuhui District	31.162210	121.432119
2	200000	Jing'an District	31.269891	121.450685
3	200000	Zhabei District	31.272406	121.455084
4	200000	Yangpu District	31.299157	121.521357
5	201900	Baoshan District	31.393338	121.417588
6	200120	Pudong New Area	31.221783	121.538740
7	201600	Songjiang District	31.029593	121.210838
8	201400	Fengxian District	30.893686	121.565314
9	200000	Changning District	31.213015	121.382477
10	200000	Putuo District	31.262501	121.400107
11	200000	Hongkou District	31.280883	121.474403
12	201100	Minhang District	31.119916	121.390940
13	201800	Jiading District	31.364105	121.218458
14	201500	Jinshan District	30.816634	121.279579
15	201700	Qingpu District	31.150700	121.124200
16	202150	Chongming County	31.631339	121.533777

Pandas Dataframe containing details of the Shanghai districts.



Folium map with all the districts indicated using black markers.

Details about the hotels and venues in Shanghai City:

This was acquired using the (foursquare.com). The API allowed me to use its 'explore' feature to get the relevant information required. I first made an API call to get details regarding the hotels, then made another API call to get details about venues near the hotels.

With the first call, details regarding the hotels' name, location, coordinates, etc. was made, using the details of the districts. Each hotel was grouped by the district it was present in.

This was followed by obtaining the details of all the different venues around the hotel, from metro stations to soccer fields to cafes and restaurants, to find which hotels were in close proximity to the most popular places in Shanghai.

	District	Dist_Latitude	Dist_Longitude	Hotel	Hotel_Lat	Hotel_Long
1	Huangpu District	31.218228	121.480302	Fraser Residence Shanghai	31.225639	121.476177
2	Huangpu District	31.218228	121.480302	Andaz Xintiandi, Shanghai (上海新天地安达仕酒店)	31.221941	121.475431
3	Xuhui District	31.162210	121.432119	Pullman Shanghai South (中星铂尔曼酒店)	31.161600	121.425152
4	Jing'an District	31.269891	121.450685	Four Points by Sheraton Shanghai, Daning	31.273794	121.451831
5	Jing'an District	31.269891	121.450685	Shanghai Holand Hotel	31.265858	121.455452
6	Yangpu District	31.299157	121.521357	Hyatt Regency Shanghai, Wujiaochang (上海五角场凯悦酒店)	31.302408	121.514755
7	Yangpu District	31.299157	121.521357	WH Ming Hotel (小南国花园大酒店)	31.291290	121.525227
8	Pudong New Area	31.221783	121.538740	Renaissance Shanghai Pudong Hotel (上海淳大万丽酒店)	31.225783	121.548058
9	Pudong New Area	31.221783	121.538740	Parkview Hotel	31.225966	121.538728
10	Pudong New Area	31.221783	121.538740	Crowne Plaza Century Park Shanghai (上海世纪皇冠假日酒店)	31.225281	121.547203
11	Songjiang District	31.029593	121.210838	Neo-Sunshine Hotel	31.028412	121.206445
12	Songjiang District	31.029593	121.210838	Shanghai Vienna Hotel	31.036080	121.216409
13	Songjiang District	31.029593	121.210838	BaoLong Home Hotel Shanghai	31.030738	121.200523
14	Changning District	31.213015	121.382477	Ruitai Hongqiao Hotel (瑞泰虹桥酒店)	31.212176	121.386333
15	Putuo District	31.262501	121.400107	Radisson Blu Hotel Shanghai Hong Quan	31.260797	121.397849
16	Hongkou District	31.280883	121.474403	SISU Guest House (SISU Guest House Hotel Shang...	31.279578	121.477662
17	Qingpu District	31.150700	121.124200	Crowne Plaza	31.153690	121.132946

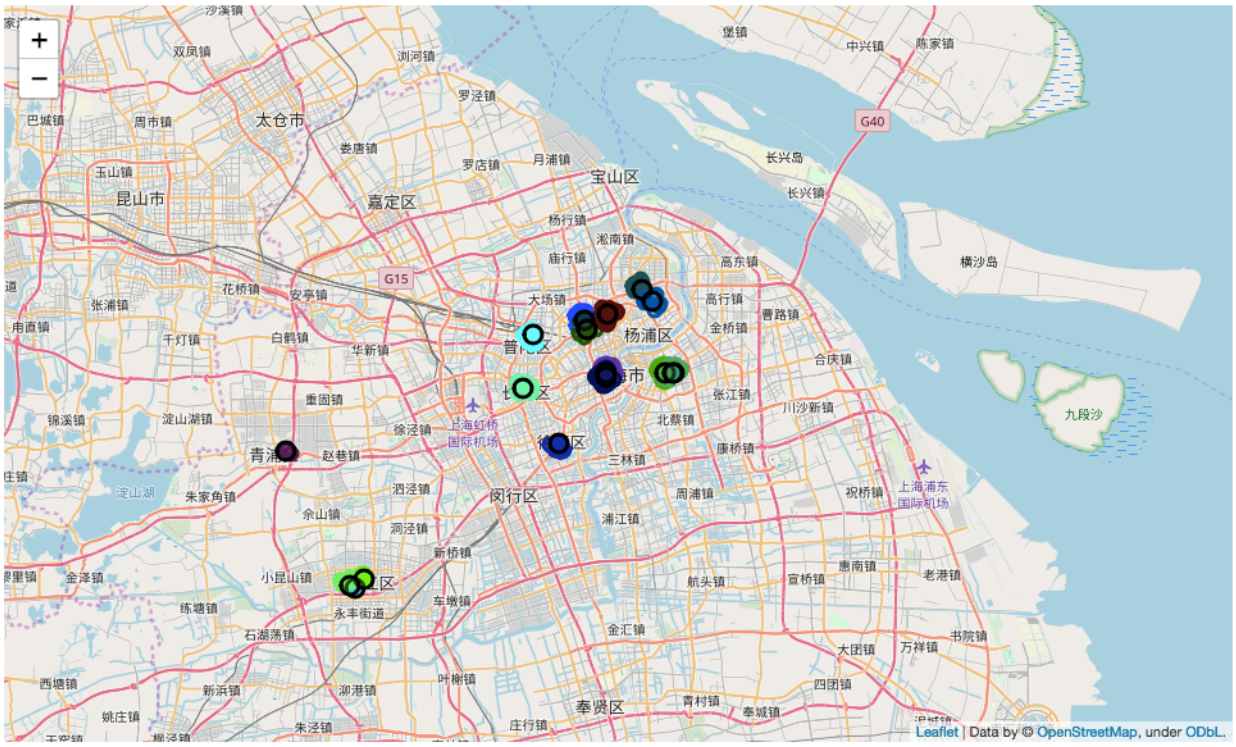
Details of the best hotels in Shanghai, grouped by their districts.

	Hotel	Hotel_Lat	Hotel_Long	Venue	Venue_Lat	Venue_Long	Venue_Category
1	Fraser Residence Shanghai	31.225639	121.476177	Mayita	31.226250	121.476113	Mexican Restaurant
2	Fraser Residence Shanghai	31.225639	121.476177	Green Massage (青籁养身)	31.223859	121.472873	Massage Studio
3	Fraser Residence Shanghai	31.225639	121.476177	Starbucks Reserve (星巴克臻选)	31.222805	121.474812	Coffee Shop
4	Fraser Residence Shanghai	31.225639	121.476177	city'super	31.226773	121.473824	Grocery Store
5	Fraser Residence Shanghai	31.225639	121.476177	Open Kitchen by Hunter Gatherer (Hunter Gather...	31.222577	121.474145	Restaurant

A sample from the dataframe, showing the venue details for places next to Fraser Residence hotel

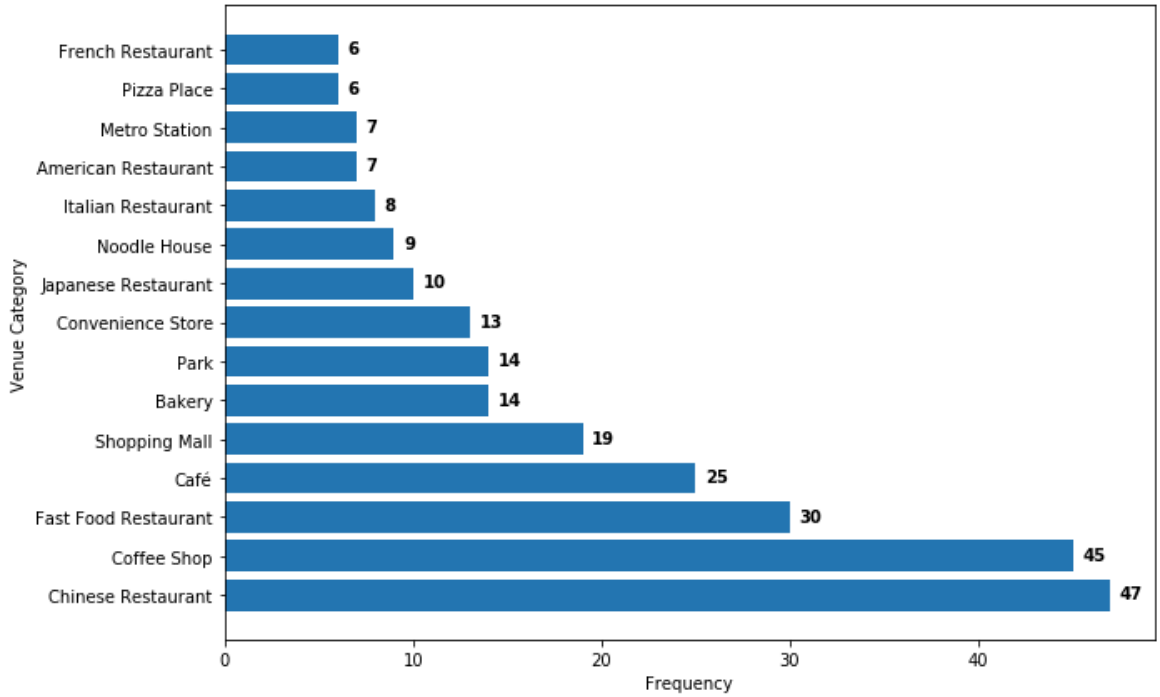
3. Methodology

- The first task was to acquire details regarding the different postcodes of different districts using the aforementioned website. The data was scraped and saved to a dataframe.
- The second task was to get the coordinates of the district using library and then displaying them on a map.
- Following this, the first Foursquare API call was made, to get the hotels in Shanghai. This was done by first finding out all the venues, and then filtering out the ones that weren't a hotel.
- The relevant details of the hotel were stored in a dataframe, from which the coordinate details were used to make a second API call.
- In the second API call, venues nearby the hotels were found and stored in a new dataframe. Using this data, a list of the top 15 most popular places was created.



A folium map showing all the venues (in multicolour) near the hotels (in black rings)

- Using this list, onehot encoded dataframe was created to compare which hotels had more popular places nearby and the number was totalled and added to a new 'Sum' column.



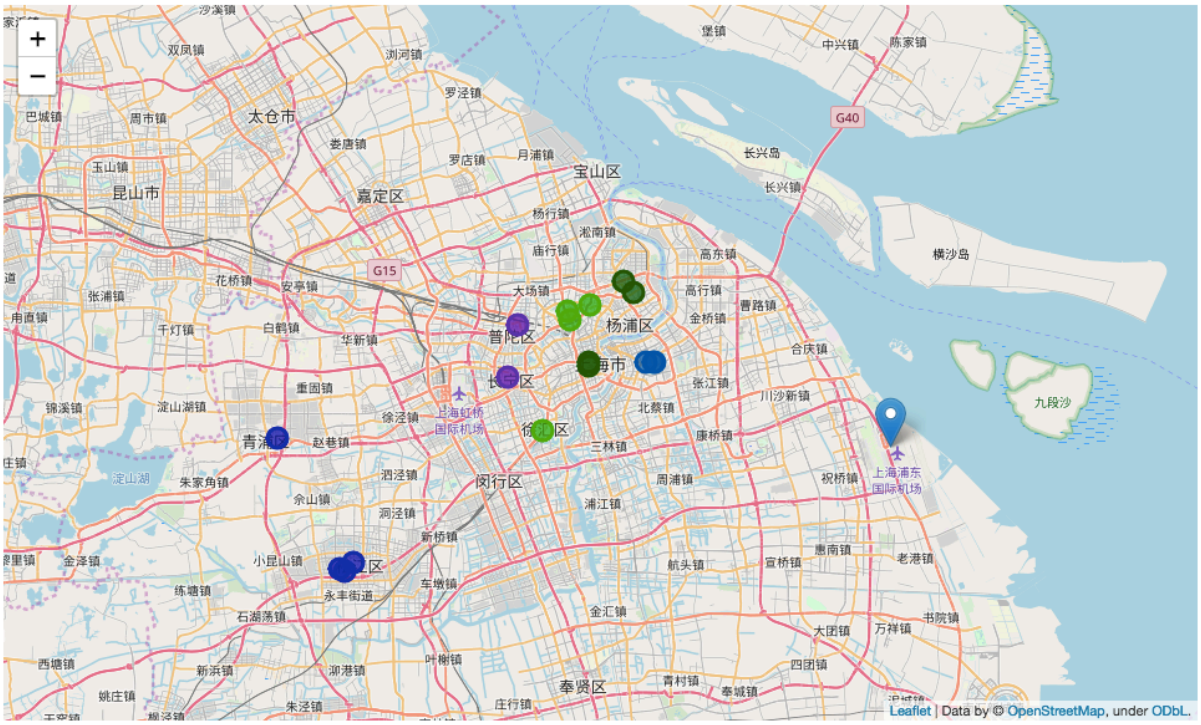
A bar graph of the 'Top 15 Most Popular Venues' in Shanghai

- After this, the distance of each hotel from the Shanghai airport was calculated. This was done because many a times businesspersons prefer to reduce transport time to and from the airport. Using these two parameters, clustering analysis was conducted and displayed on a map.

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
1	Andaz Xintiandi, Shanghai (上海新天地安达仕酒店)	0.419355	32.876467	2	31.221941	121.475431	Chinese Restaurant	Café	Restaurant
2	BaoLong Home Hotel Shanghai	0.500000	59.343134	1	31.030738	121.200523	Neighborhood	Coffee Shop	Lake
3	Crowne Plaza	0.666667	64.407039	1	31.153690	121.132946	Coffee Shop	Clothing Store	Shopping Mall
4	Crowne Plaza Century Park Shanghai (上海世纪皇冠假日酒店)	0.588235	26.457728	3	31.225281	121.547203	Fast Food Restaurant	Coffee Shop	Bakery
5	Four Points by Sheraton Shanghai, Daning	0.703704	36.879304	4	31.273794	121.451831	Chinese Restaurant	Coffee Shop	Bakery

A part of the final dataframe used to conduct the cluster analysis

4. Results



A folium map rendered with the 5 clusters and the Shanghai Intl Airport

- The distance from the airport influences clusters. If we followed an imaginary radius with the airport as a concurrent centre, we can identify that all the hotels fit in ideal clusters that can be imagined to be on a circle.

```
No of Hotels in Cluster Label 0: 2
```

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	...	6th Most Common
1	Radisson Blu Hotel Shanghai Hong Quan	0.222222	41.194022	0	31.260797	121.397849	Seafood Restaurant	Soccer Field	Food Truck	Train Station	...	Metro Station
2	Ruitai Hongqiao Hotel (瑞泰虹桥酒店)	0.615385	40.922399	0	31.212176	121.386333	Japanese Restaurant	Chinese Restaurant	Coffee Shop	Café	...	Fast Fooc Restaurar

2 rows × 21 columns

```
No of Hotels in Cluster Label 2: 4
```

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	...	6t Co Ve
1	BaoLong Home Hotel Shanghai	0.500000	59.343134	1	31.030738	121.200523	Neighborhood	Coffee Shop	Lake	Chinese Restaurant	...	Di Re
2	Crowne Plaza	0.666667	64.407039	1	31.153690	121.132946	Coffee Shop	Clothing Store	Shopping Mall	Yoga Studio	...	De Sh
3	Neo-Sunshine Hotel	0.600000	58.847578	1	31.028412	121.206445	Convenience Store	Chinese Restaurant	Taiwanese Restaurant	Supermarket	...	Fo
4	Shanghai Vienna Hotel	0.571429	57.735268	1	31.036080	121.216409	Coffee Shop	Movie Theater	Big Box Store	Park	...	Su

4 rows × 21 columns

```
No of Hotels in Cluster Label 3: 4
```

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	...	6th M Com Venu
1	Andaz Xintiandi, Shanghai (上海新天地安达仕酒店)	0.419355	32.876467	2	31.221941	121.475431	Chinese Restaurant	Café	Restaurant	Coffee Shop	...	Amer Resta
2	Fraser Residence Shanghai	0.426966	32.917222	2	31.225639	121.476177	Chinese Restaurant	Coffee Shop	Café	Park	...	Nooc Hous
3	Hyatt Regency Shanghai, Wujiaochang (上海五角场凯悦酒店)	0.489362	33.008692	2	31.302408	121.514755	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Shopping Mall	...	Stadi
4	WH Ming Hotel (小南国花园大酒店)	0.625000	31.514159	2	31.291290	121.525227	Park	Ice Cream Shop	Pizza Place	Asian Restaurant	...	Metro Static

4 rows × 21 columns

```
No of Hotels in Cluster Label 4: 3
```

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	..
1	Crowne Plaza Century Park Shanghai (上海世纪皇冠假日酒店)	0.588235	26.457728	3	31.225281	121.547203	Fast Food Restaurant	Coffee Shop	Bakery	Convenience Store	..
2	Parkview Hotel	0.642857	27.243991	3	31.225966	121.538728	Coffee Shop	Convenience Store	Fast Food Restaurant	Chinese Restaurant	..
3	Renaissance Shanghai Pudong Hotel (上海淳大万丽酒店)	0.633333	26.399932	3	31.225783	121.548058	Coffee Shop	Fast Food Restaurant	Convenience Store	Pub	..

3 rows × 21 columns

```
No of Hotels in Cluster Label 5: 4
```

	Hotel	Sum	Distance	Cluster Label	Hotel_Lat	Hotel_Long	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	...	6th M Comn Venue
1	Four Points by Sheraton Shanghai, Daning	0.703704	36.879304	4	31.273794	121.451831	Chinese Restaurant	Coffee Shop	Bakery	Fast Food Restaurant	...	Groce Store
2	Pullman Shanghai South (中星铂尔曼酒店)	0.500000	36.583998	4	31.161600	121.425152	Fast Food Restaurant	Bus Station	Coffee Shop	Supermarket	...	Cosm Shop
3	SISU Guest House (SISU Guest House Hotel Shang...	0.613636	34.895072	4	31.279578	121.477662	Coffee Shop	Fast Food Restaurant	Café	Chinese Restaurant	...	Pizza Place
4	Shanghai Holand Hotel	0.500000	36.226524	4	31.265858	121.455452	Fast Food Restaurant	Metro Station	Road	Tea Room	...	Noodl House

4 rows × 21 columns

- The venues that are nearby hotels equally influence the cluster analysis. No two hotels in a cluster (bar the first one, which indicates an outlier) have a difference of more than 0.25 in terms of their sums.
- Only 4 hotels have a sum of less than 0.5, indicating that popular venues in Shanghai are also located near the hotels.

Discussion/Conclusion

Using Foursquare API, we have collected a decent amount of hotel recommendations in Shanghai. However, it has its limitations as we do not have an exhaustive list of all the available hotels, restaurants, cafes, etc.

The generated clusters indicate that there are a lot of popular venues in close proximity of the Hotels. However, this doesn't mean that one influences the other (**Correlation doesn't imply causation**). Travellers who are on a decent budget and want to explore the local culture will find this catalogue the most useful. Businesspersons can also make use of this catalogue to understand how far their hotel will be from the airport, and if they have any spare time, which places they can visit.

Since this also includes a graph of most popular venues, people can identify those places and manually choose to visit them. The most popular venues are Chinese restaurants, Cafes in general and fast food outlets.

In conclusion, I would like to thank Coursera and IBM for allowing me to create this informative and helpful project and hone my skills for the future.