Coursera Capstone Project

IBM Data Science Professional

Searching for a good area to build a new Hotel in Shanghai, China

Author: Lucas Kanz

10.07.2020

Introduction

Vacation is for many cultures and regions in the World a very important thing. One part of a successful vacation is to be able to book a hotel in your favorite destination. As Shanghai is one of the fastest growing economies and also a venue that's getting more and more important every year for people searching for a interesting and beautiful vacation destination, as well as for business people searching for a place to stay.

On the other hand for business developers it's hard to find a good place to open a new venue as the city is getting more and more crowded, and you don't want to make the mistake of building a new, expensive hotel complex into an area that Is already filled with such. Therefore there is a need for a clear and detailed analysis were the best location for such a new opening should be, as it is crucial for the success of the upcoming business.

Business Problem

The goal of this part of the Capstone Project is to find the perfect spot were such a hotel could be placed to be not in the middle of an area already saturated with such venues. It will enable a minimized business risk, as dependency on competitors should be minimized too for a successful start of the new Hotel business the possible customer wants to open.

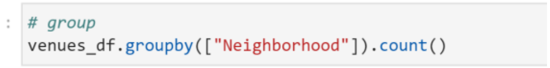
As building a Hotel complex is very cost intensive usually costing between 30 and 200 million depending on the type of Hotel, the success of it needs to be guaranteed to not lead to a financial disaster for the business developer. Gaining insights through valuable data is a important key to lead the path for a successful future.

Audience

This analysis is especially important **for venue developers, hoteliers as well as large business groups looking for additional ways of generating income through a customer serving business like a Hotel or Hotel Chain**. As more regions in the world get access to more and more wealth and start traveling the world the need for more and well designed Hotels is increasing in the last couple of years together with an increase in competition especially in regions of extraordinary growth.

Data

- Scraping Wikipedia sited to get neighborhoods in Shanghai (*)
- add geo data to these scraped neighborhoods via geocoder (add longitudinal and latitudinal coordinates that are needed for Fourspace API)
- clean the dataset and group by neighborhoods to prepare for clustering



```
# group
venues_df.groupby(["Neighborhood"]).count()
```

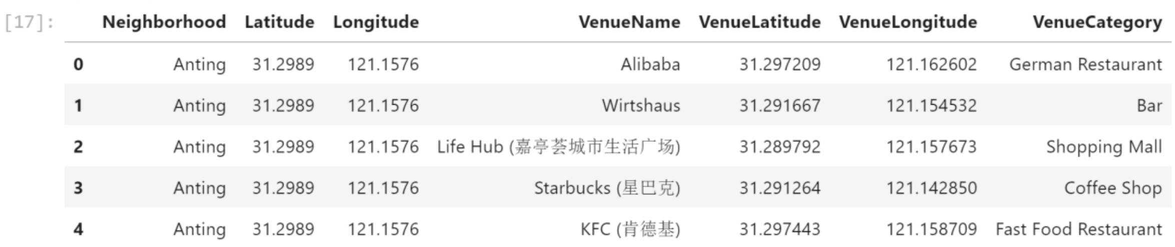
- add venue data by using Fourspace
- take the mean of occurrence of venues
- exclude all other venues except hotels in the dataset
- Cluster neighborhoods using k-means clustering
- Analyse which Cluster Area would be best to open a Hotel to avoid too much competition by creating table and visual data via Folium

We will use web scraping of a Wikipedia Website that's showing neighborhoods in Shanghai:

(*) https://en.wikipedia.org/wiki/Category:Neighbourhoods_of_Shanghai

After that Fourspace will help us to find venues nearby and cluster areas with a high amount of Hotels.

One example what could be gathered by the Foursquared API is different kind of venues around a specific area, like in our case 100 venues in a radius of 2500 meters:



[17]:

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Anting	31.2989	121.1576	Alibaba	31.297209	121.162602	German Restaurant
1	Anting	31.2989	121.1576	Wirtshaus	31.291667	121.154532	Bar
2	Anting	31.2989	121.1576	Life Hub (嘉亭荟城市生活广场)	31.289792	121.157673	Shopping Mall
3	Anting	31.2989	121.1576	Starbucks (星巴克)	31.291264	121.142850	Coffee Shop
4	Anting	31.2989	121.1576	KFC (肯德基)	31.297443	121.158709	Fast Food Restaurant

These can include categories of venues just as German Restaurants, Bars, Shopping Malls, Coffee Shops, Night Clubs, etc. In addition Foursquare is also providing the Venue Latitude and Longitude data for a short comparison where within the Neighborhood the venue is located.

Methodology

To begin with we extract the neighbourhoods in Shanghai, China. This can be performed vie data scraping from the Wikipedia page mentioned above. Web scraping is done via Python request as well as with BeautifulSoup Library.

Secondly, as the ouput of the data scraping will be a list without any geographical data we will latitude and longitude coordinates as a preparation to use the Foursquare API. This can be achieved via the Geocoder Library installed into Python. This package converts addresses into coordinates that can then be added into a Pandas Dataframe .

Thirdly this data can then be visualized via folium (python library) in a beautiful map. This is basically needed to check in between if the data gathered is what we're expecting.

After that we're using the Foursquare API to find 100 venues in an area of 2500 meters around. The Data is returned in an JSON File Format from which we extract some key data. We then can check how many unique venues are available in the region. Then the Dataset will be grouped by neighborhood as well as showcasing the mean of the occurrence of each venue in that area as a pre-process step to prepare the data for clustering via k-means clustering.

Clustering will split the data into 3 based in the occurrence if "Hotel" in the dataset. This will enable us to see which neighborhoods have a high occurrence of Hotels (or low occurrence).

This is basically the key information we need to decide in which neighborhood a Hotel opening would make the most sense from a business perspective only looking into minimizing competition nearby.

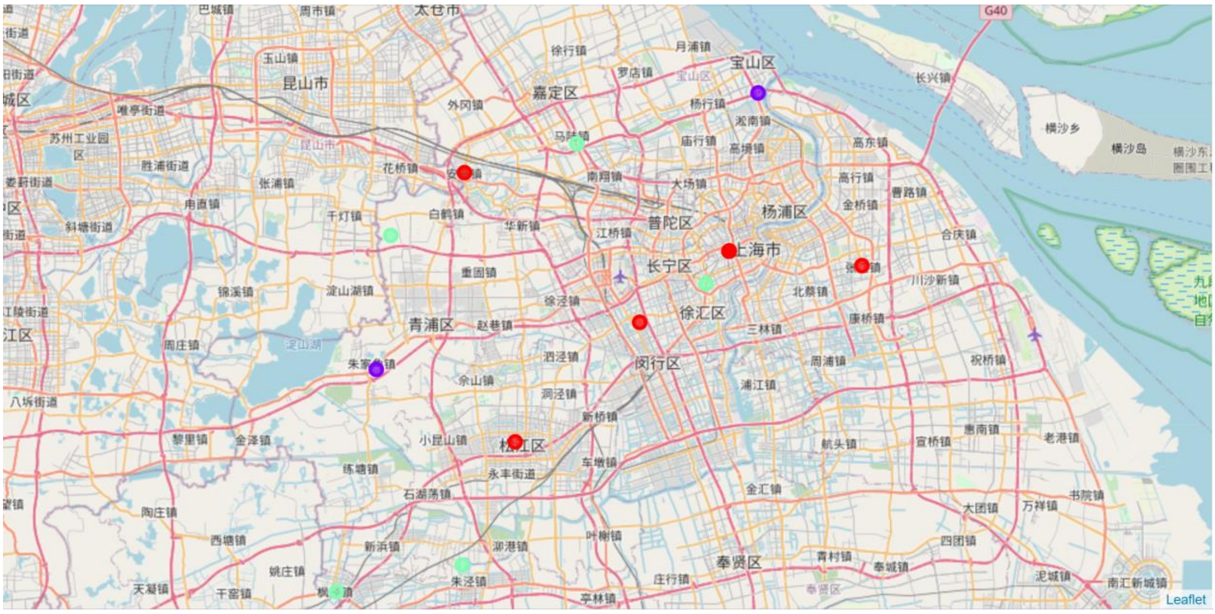
Results

The clustering resulted into three different clusters as specified below:

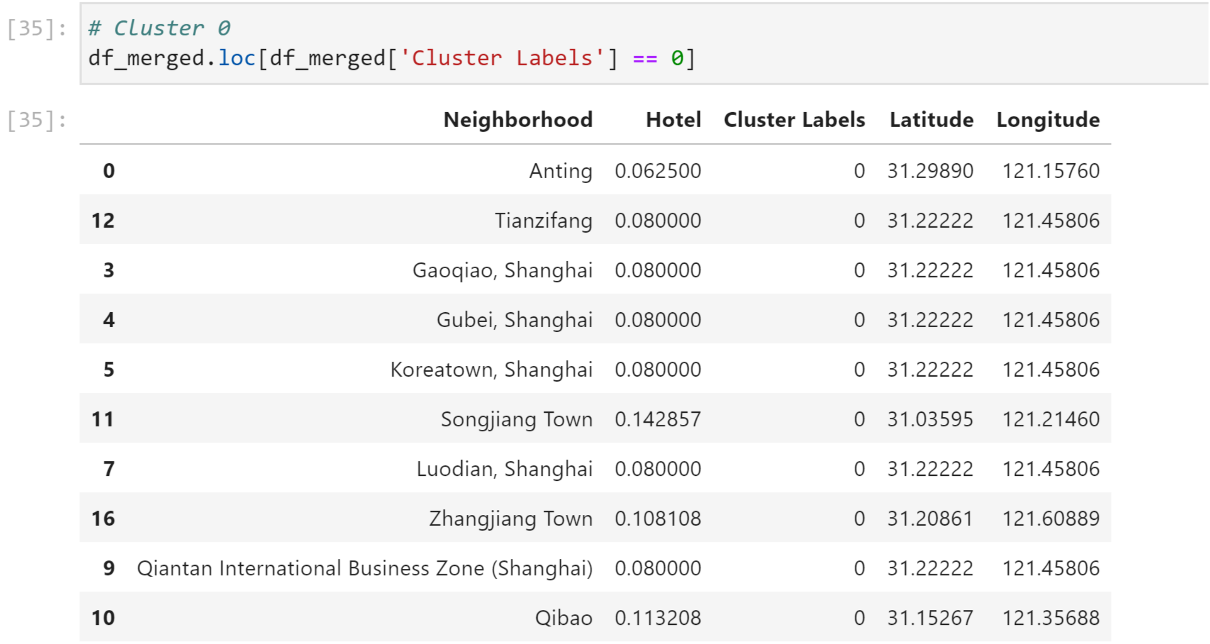
Cluster 0 (red): Neighborhoods with moderate hotel amount

Cluster 1 (purple): Neighborhood with low to almost no existing Hotels

Cluster 2 (green): Neighborhood with high amounts of Hotels in that area



Cluster 0 Details:

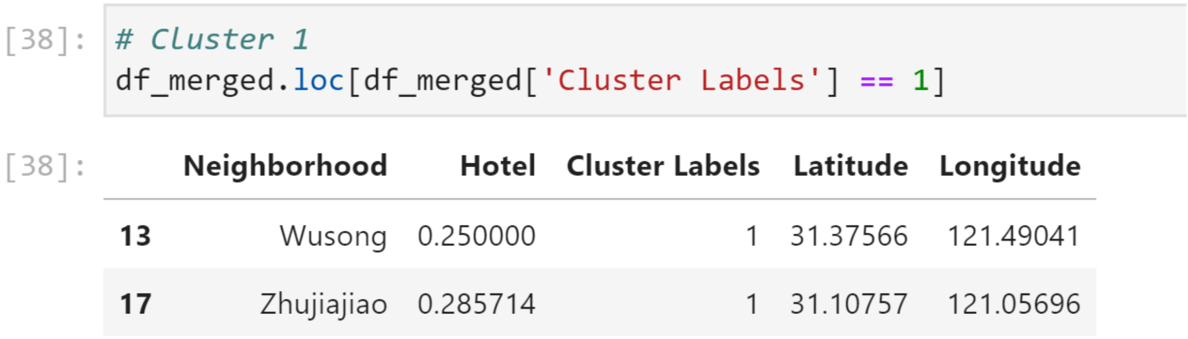


```
[35]: # Cluster 0
      df_merged.loc[df_merged['Cluster Labels'] == 0]
```

[35]:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
0	Anting	0.062500	0	31.29890	121.15760
12	Tianzifang	0.080000	0	31.22222	121.45806
3	Gaoqiao, Shanghai	0.080000	0	31.22222	121.45806
4	Gubei, Shanghai	0.080000	0	31.22222	121.45806
5	Koreatown, Shanghai	0.080000	0	31.22222	121.45806
11	Songjiang Town	0.142857	0	31.03595	121.21460
7	Luodian, Shanghai	0.080000	0	31.22222	121.45806
16	Zhangjiang Town	0.108108	0	31.20861	121.60889
9	Qiantan International Business Zone (Shanghai)	0.080000	0	31.22222	121.45806
10	Qibao	0.113208	0	31.15267	121.35688

Cluster 1 Details:

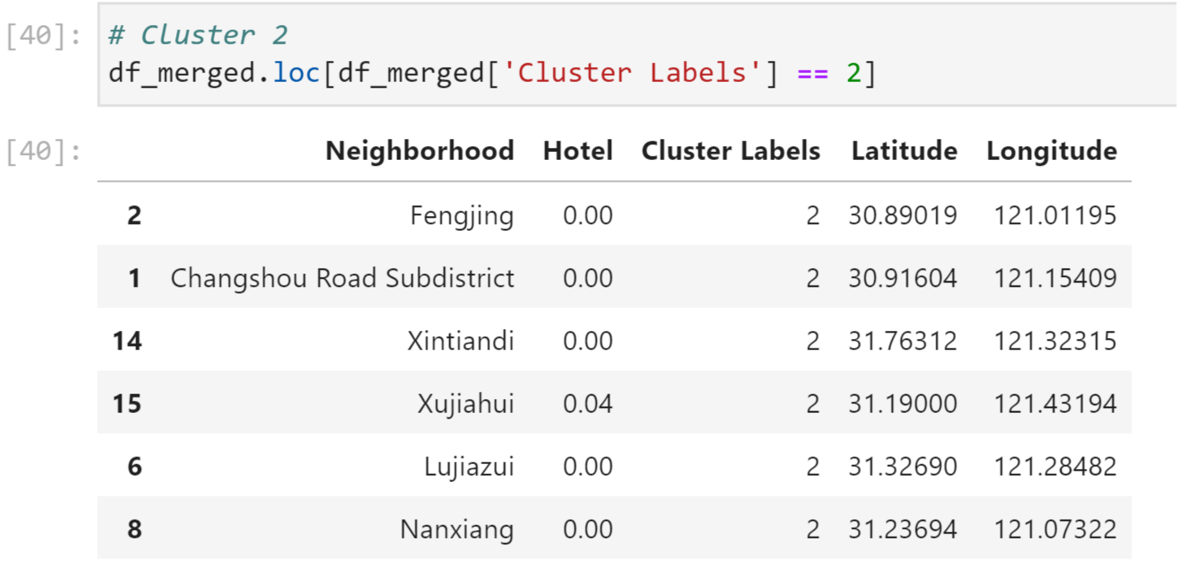


```
[38]: # Cluster 1
      df_merged.loc[df_merged['Cluster Labels'] == 1]
```

[38]:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
13	Wusong	0.250000	1	31.37566	121.49041
17	Zhujiajiao	0.285714	1	31.10757	121.05696

Cluster 2 Details:



```
[40]: # Cluster 2
      df_merged.loc[df_merged['Cluster Labels'] == 2]
```

[40]:

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
2	Fengjing	0.00	2	30.89019	121.01195
1	Changshou Road Subdistrict	0.00	2	30.91604	121.15409
14	Xintiandi	0.00	2	31.76312	121.32315
15	Xujiahui	0.04	2	31.19000	121.43194
6	Lujiazui	0.00	2	31.32690	121.28482
8	Nanxiang	0.00	2	31.23694	121.07322

Discussion

As it’s visible in the Map above, most of the Hotels are located in the center and the far west of Shanghai. Central is a very important location for many hotels as customers can reach easy into all direction within the city, west is important as it’s near the airport. So Cluster 2 (green) seems to be a bad choice setting up a hotel, looking that the competition there is already quite high. On the other hand looking on Cluster 1 (purple) we can clearly see that in general the coast region seems to be not yet as used for hotels even having beautiful look into the Shanghai bay. This might be a very good location with key selling points and a low competition.

This project suggests therefore using the coast area within Cluster 2 for potential business openings in the Hotel Business as it’s most likely the area with the least amount of competition.

Disclaimer

In this project we only considered the low amount of competition as an important point for an opening of a Hotel. In reality, for sure this is not the only key value for a decision on where a good location would be. Facts as urban surroundings, restaurants, coffee shops, view from the hotel, accessibility as well as a good reachability are also important factors to consider (among others). Unfortunately this data was not available for this project, as it's partially also depending on taste of the business owner, and the overall theme of the Hotel. Also using a bigger dataset (using a paid account for your choice of API) could improve this results.

Conclusion

With this project we went through every step that's needed in a Data Science Project. From identifying the business problem/business case to specifiying data that's needed, extracting it, cleaning and preparing it for further analysis to finally perform machine learning to find clusters within the data. All this leading to get us into the position to make data driven recommendations to our customers that enable them for a high chance of business success.