The Battle of Neighborhoods in Shanghai

Author: Jane W.J. Goh
Date: 01 November 2020

Introduction

- Background
  - Shanghai is home to 200,000 expats, the largest expat population in China mainland
- Problem
  - Housing information available largely in Chinese, work can be done to close the information asymmetry
- Solution
  - This project explores and clusters the neighborhoods in Shanghai in order to find a suitable community for a new expat in town

Jupyter Project Components

- There are 3 parts to the project:
- 1. Importing and cleaning the Shanghai neighborhoods data
- 2. Calling Foursquare API to find each neighborhood's characteristics- by most common venue
- 3. Using k-means to cluster the neighborhoods & analysis

DATA

Data

- Neighborhood here is defined as within 400 meter/ 5-minute walking distance of a metro station- the main form of transport between home and office.
- Neighborhood characteristics is defined by the type of venues available from the 400 meter/ 5-minute walking distance of the metro station.
- Two main sources of data
  - 1. CSV file containing Shanghai's 423 metro stations in English and Chinese names, locations, latitude, longitude data and
  - 2. Foursquare API which provides venues data for each latitude, longitude data points.

METHODOLOGY

Data Exploration

- In preparing the Shanghai neighborhoods data, duplicated station names were removed.
- After merging the data from Shanghai neighborhoods and Foursquare data, neighborhoods with < 30 venues results were excluded

```
In [566]: # 2.2.4. Tabulate the venues results
          print(shanghai_venues.groupby('StationName_English').count().shape)
          shanghai_venues.groupby('StationName_English').count()
```

```
(278, 6)
```

Out[566]:

	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
StationName_English						
Anshan Xincun	4	4	4	4	4	4
Baiyin Road	2	2	2	2	2	2
Bao'an Highway	6	6	6	6	6	6
Baoshan Road	10	10	10	10	10	10
Baoyang Road	4	4	4	4	4	4

- Analyzed neighborhoods' characteristics:

	StationName_English	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 Most Common Venue
5	Shangcheng Road	Coffee Shop	Hotel	Szechuan Restaurant	Japanese Restaurant	Fast Food Restaurant	Hotpot Restaurant	Pizza Place	Athletics & Sports	Kushikatsu Restaurant	Clothing Store
6	Shanghai Library	Bar	Restaurant	Art Gallery	Bistro	Cocktail Bar	Turkish Restaurant	Nightclub	Hotel	Hualyang Restaurant	Pizza Place
7	South Huangpi Road	Hotel	Café	Chinese Restaurant	New American Restaurant	Cocktail Bar	Coffee Shop	Park	Ice Cream Shop	Taiwanese Restaurant	Shopping Mall
8	Xinzha Road	Fast Food Restaurant	Hotel	Chinese Restaurant	Coffee Shop	Dumpling Restaurant	Lounge	Bakery	Bed & Breakfast	Café	Candy Store
9	Xujiahui	Coffee Shop	Clothing Store	Chinese Restaurant	Sandwich Place	Burger Joint	Pizza Place	Shopping Mall	Fast Food Restaurant	Supermarket	Shanghai Restaurant

Clustering

- It may be good enough to know what is popular in each neighborhood, however, we'll take it a step further and find which neighborhoods are more similar in spirit through clustering analysis.
- K-Means is chosen to learn from the data and cluster the neighborhoods unsupervised.

3.1 Clustering

```
]: # Now we are ready to do the clustering and analysis of neighborhoods
   # 3.1.1. run the clustering algorithm

   # set number of clusters
   kclusters = 5

   # dropping the 'StationName_English' column cause we don't need it for the clustering algorithm
   shanghai_grouped_clustering = shanghai_grouped.drop('StationName_English', 1)

   # run k-means clustering
   kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(shanghai_grouped_clustering)

   # check cluster labels generated for each row in the dataframe
   kmeans.labels_[0:10]
```

```
]: array([4, 2, 3, 2, 0, 1, 4, 3, 1, 2])
```

RESULTS

Cluster 1

- People's square station is filled with variety of Asian cuisine restaurants, hotels, and karaoke bar- indicating that this might be a very bustling place, suitable for someone who prefers a busy environment with easy access to food and entertainment

```
# cluster 1: Tourist area/ Downtown
shanghai_merged.loc[shanghai_merged['Cluster Labels'] == 0, shanghai_merged.columns[[0] + list(range(5, shanghai_merged.shape[1]))]]
```

	StationName_English	Cluster Labels	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 Most Common Venue
12	People's Square	0	Chinese Restaurant	Hotel	Noodle House	Coffee Shop	Sandwich Place	Bookstore	Shanghai Restaurant	Karaoke Bar	Korean Restaurant	Fast Food Restaurant

Cluster 2

- Xinzha Road and Shangcheng Road stations are concentrated in largely Asian/some Western restaurants and cafes, hotels and b&b's, and shopping (sports & clothes)- an indication that it could be an area that is convenient for tourist access.

```
# cluster 2: Higher-end residential cluster
shanghai_merged.loc[shanghai_merged['Cluster Labels'] == 1, shanghai_merged.columns[[0] + list(range(5, shanghai_merged.shape[1]))]]
```

	StationName_English	Cluster Labels	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 Most Common Venue
13	Xinzha Road	1	Fast Food Restaurant	Hotel	Chinese Restaurant	Coffee Shop	Dumpling Restaurant	Lounge	Bakery	Bed & Breakfast	Café	Candy Store
246	Shangcheng Road	1	Coffee Shop	Hotel	Szechuan Restaurant	Japanese Restaurant	Fast Food Restaurant	Hotpot Restaurant	Pizza Place	Athletics & Sports	Kushikatsu Restaurant	Clothing Store

Cluster 3

- Xujiahui, Lujiazhui and Huamu Road stations are first and foremost about coffee shops, followed by a mix of Asian/Western restaurants, and access to shopping malls/ convenience stores/ supermarket, these are indications of the convenient city lifestyle

```
# cluster 3: Central Business District/ Office/ High-rise residential cluster
shanghai_merged.loc[shanghai_merged['Cluster Labels'] == 2, shanghai_merged.columns[[0] + list(range(5, shanghai_merged.shape[1]))]]
```

	StationName_English	Cluster Labels	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 Most Common Venue
7	Xujiahui	2	Coffee Shop	Clothing Store	Chinese Restaurant	Sandwich Place	Burger Joint	Pizza Place	Shopping Mall	Fast Food Restaurant	Supermarket	Shanghai Restaurant
41	Lujiazui	2	Coffee Shop	Hotel Bar	Scenic Lookout	Hotel	Chinese Restaurant	Japanese Restaurant	Italian Restaurant	Convenience Store	Dumpling Restaurant	Electronics Store
192	Huamu Road	2	Coffee Shop	Cantonese Restaurant	Burger Joint	Pizza Place	Fast Food Restaurant	Clothing Store	Shanghai Restaurant	Sandwich Place	Restaurant	Noodle House

Cluster 4

- South Huangpi Road and Jian'an Temple stations are the first clusters where we see gym and park making it to the top 10. Along with a mix of Asian/Western restaurants and cocktail bars, this cluster looks like a good base for work/life balance.

```
# cluster 4: The Entertainment/ TikTok Influencers' Hangout cluster
shanghai_merged.loc[shanghai_merged['Cluster Labels'] == 3, shanghai_merged.columns[[0] + list(range(5, shanghai_merged.shape[1]))]]
```

	StationName_English	Cluster Labels	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 Most Common Venue
11	South Huangpi Road	3	Hotel	Café	Chinese Restaurant	New American Restaurant	Cocktail Bar	Coffee Shop	Park	Ice Cream Shop	Taiwanese Restaurant	Shopping Mall
37	Jing'an Temple	3	Japanese Restaurant	Cocktail Bar	Coffee Shop	Shanghai Restaurant	Burger Joint	Gym	Cantonese Restaurant	Food Court	Café	Lounge

Cluster 5

- East Nanjing Road and Shanghai Library stations are defined by its lounge/bistro/deli, largely Western restaurants. And the art gallery and jazz club? This cluster looks like one fine lifestyle neighborhood

```
# cluster 5: Sophisticated Hangout Cluster
shanghai_merged.loc[shanghai_merged['Cluster Labels'] == 4, shanghai_merged.columns[[0] + list(range(5, shanghai_merged.shape[1]))]]
```

	StationName_English	Cluster Labels	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 Most Common Venue
40	East Nanjing Road	4	French Restaurant	Chinese Restaurant	Hotel	Italian Restaurant	Lounge	Shopping Mall	Seafood Restaurant	Restaurant	Deli / Bodega	Jazz Club
271	Shanghai Library	4	Bar	Restaurant	Art Gallery	Bistro	Cocktail Bar	Turkish Restaurant	Nightclub	Hotel	Huaiyang Restaurant	Pizza Place

DISCUSSION

Improvements

- This project could be further improved with the following:
- In part 1: Introduce a web crawler and more comprehensive data cleaning codes.
- In part 1, 2: Can include other potential data sources such as average property price, surround building types, and residential demographic to make neighborhood characteristics identification even more meaningful.
- In part 3: Instead of K-Means, maybe can explore DBSCAN density-based clustering.

CONCLUSION

Thank you

- We have made a headway into exploring of different neighborhoods in Shanghai, a different cluster for everyone with different lifestyles.
- For people who are looking to a new place, hope you find a place to call home. For my fellow classmates, Happy Learning!