East China City Analysis Report

Allen Wang

2020/01/08

Contents

Introduction **2**

Data **2**

The list of provinces and cities in east China 2

Chinese cities and their coordinates 2

Top venues 2

Methodology **2**

Clustering 2

KMeans 3

Elbow Method in KMeans 3

Descriptive Analysis 4

Results **6**

Uniformity among clusters and within the region 6

Differences among clusters 6

Discussion **6**

The uniformity 6

The differences 6

Conclusion **7**

Introduction

The east region is the richest and most populous area (29.32% of total population according to https://en.wikipedia.org/wiki/Demographics_of_China) in China. This research is aiming for understanding the lifestyle of the cities locate in this area by classifying them into several groups according to the differences on top venues that people visit. The differences include categories of venues as well as the frequencies of those categories. Also, the number and size of city groups provides an insight on the degree of uniformity across the whole region. For example, the more clusters we finally get, the more dispersed this region is.

For companies that are contemplating about setting up new business or expanding current one into the east China region, understanding what venues are available and popular already, and what people in this region like, is critical.

Data

The list of provinces and cities in east China

The list of provinces and cities that locate in the east China region is available at this Wikipedia page. https://en.wikipedia.org/wiki/East_China

Chinese cities and their coordinates

Data of cities and their coordinates are available on the internet and I use the .csv file which is available at this site. https://simplemaps.com/data/cn-cities

Top venues

With city names and coordinates available, I can call Foursquare API to retrieve top 500 venues within the range of 5km in each city, and perform analysis accordingly. https://foursquare.com/city-guide

Methodology

Clustering

This research tries to segregate all the east cities into different groups and find the patterns within them. It is a typical clustering problem. Clustering is the approach where a set of data points are divided into smaller homogenous groups with points in the same subgroup are similar to each other, but differences among subgroups are significant. In the business field, clustering has been used in a large scale to perform market segmentation and customer grouping, which helps companies tailor their promotions and products for different groups of markets/consumers.

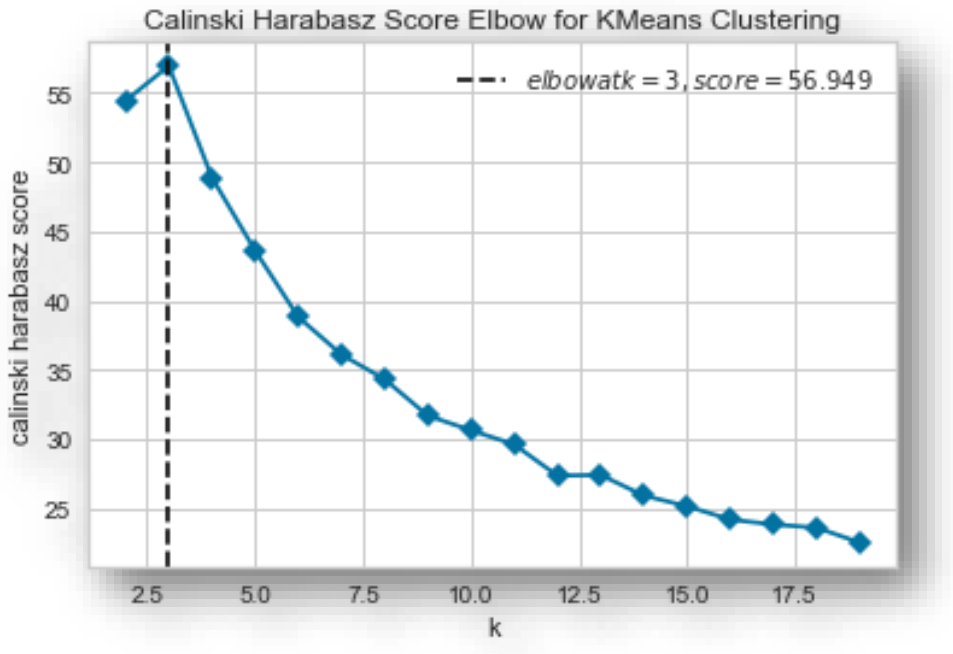
KMeans

Clustering problem is what the *KMeans* algorithm was designed for. Arguably, *KMeans* is one of the most widely used clustering algorithms due to its elegance and simplicity. In this research, I use *scikit-learn* , a popular machine learning Python package, which provides highly integrated APIs for running *KMeans* algorithm.

Elbow Method in *KMeans*

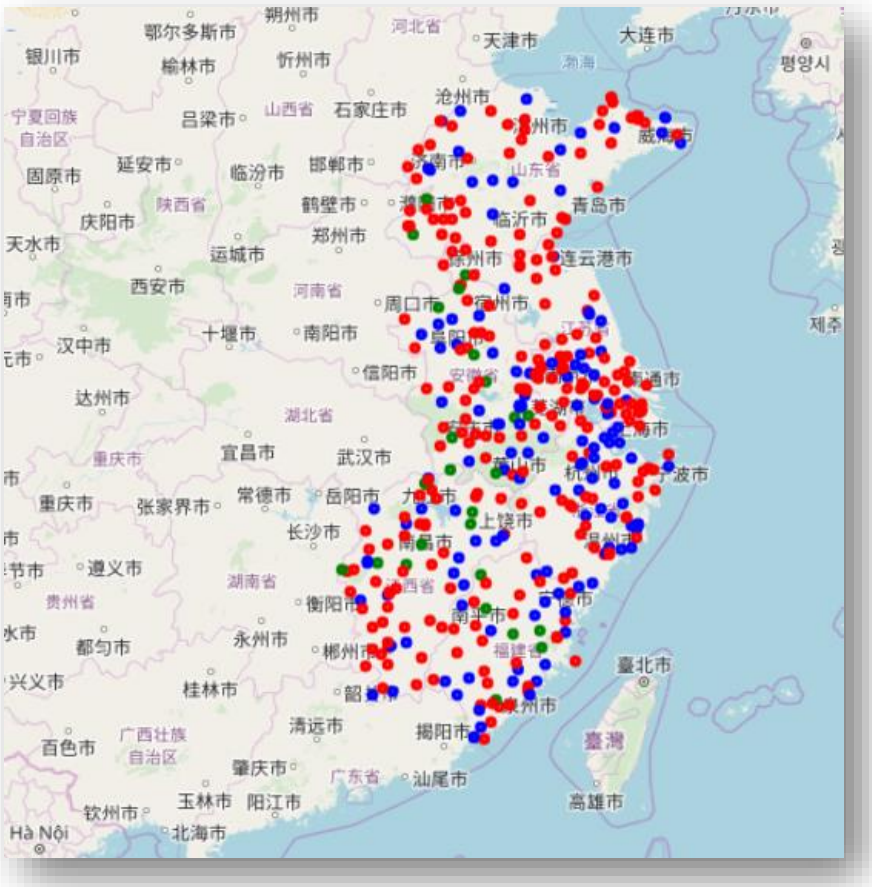
KMeans is used for unsupervised learning. In this particular clustering problem, the optimal number of clusters that fits this dataset is unknown, and that's where elbow method comes to the rescue. One way to implement the elbow method is calculating the distortion value (the average of squared distances from points to their cluster centers). As the number of clusters increase, the value of distortion drops significantly at the beginning phase, then display a smooth trend of slower declining. The number of cluster (dented by k) on the turning point is where we can strike a balance between accuracy of clustering and efficient utilization of computational power, and avoid overfitting.

With the python package *yellowbrick* in place, we don't have to the process stated above by ourselves, just simply call the function *KElbowVisualizer* from the package and it will automatically calculate the optimal number of K.



Geospatial Data Visualization

After finding the best k, I applied *KMeans* algorithm and successfully divided cities into 3 groups. With group labels and coordinates of cities available, I used *folium,* a user-friendly and powerful geospatial data visualization Python package, to mark the cities in red/green/blue color according to the labels. Below is the output.



Descriptive Analysis

After retrieving top 500 venues for every city, I applied one hot encoding technique on the categories of venues and calculated the frequency of venues for every city, then sort in descending order and get the top 10 most common venues in each city, as shown below.

city	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
Andongwei	Hotel	Harbor / Marina	Chinese Restaurant	Boat or Ferry	Shandong Restaurant	Zoo	Football Stadium	Food Service	Food Truck	French Restaurant
Anqing	Fast Food Restaurant	Hotel	Train Station	Shopping Plaza	Food Truck	Food & Drink Shop	Food Court	Food Service	Football Stadium	Flea Market
Aoyang	Bus Station	Zoo	German Restaurant	Gastropub	Garden	Furniture / Home Store	Fujian Restaurant	Fried Chicken Joint	French Restaurant	Fountain
Aoyang	Bus Station	Zoo	German Restaurant	Gastropub	Garden	Furniture / Home Store	Fujian Restaurant	Fried Chicken Joint	French Restaurant	Fountain
Baiguan	Train Station	Hotel	Fast Food Restaurant	Coffee Shop	Fountain	Food Court	Food Service	Food Truck	Football Stadium	Zoo
Baoshan	Coffee Shop	Port	Stadium	Shopping Mall	Park	Fast Food Restaurant	Football Stadium	Fountain	Food	Food Truck
Bashan	Market	Plaza	Zoo	Food	Gastropub	Garden	Furniture / Home Store	Fujian Restaurant	Fried Chicken Joint	French Restaurant
Beidajie	Coffee Shop	Fast Food Restaurant	Hotel	Shopping Mall	Mexican Restaurant	Garden	Burger Joint	General Entertainment	Grocery Store	Chinese Restaurant

After clustering into 3 groups, first I calculated the number of cities that each cluster includes, as shown in table below.

Cluster	Number of cities
0	267
1	25
2	13

Then I aggregated all the common venues within each cluster and listed the top 10 venues together with their percentage, which gives us a deeper insight into the features of the 3 different clusters.

- Cluster 0, red dots on the map

Venues	Percentage%
Hotel	15.02
Zoo	11.58
Fast Food Restaurant	10.95
Coffee Shop	9.68
Fried Chicken Joint	9.32
French Restaurant	9.14
Garden	8.87
Fujian Restaurant	8.69
Furniture/ Home Store	8.42
Shopping Mall	8.33

- Cluster 1, green dots on the map

Venues	Percentage%
Train Station	12.14
Zoo	11.65
Fried Chicken Joint	10.19
Furniture/Home Store	10.19
Gastropub	10.19
Fujian Restaurant	10.19
Garden	10.19
French Restaurant	9.22
Food & Drink Shop	8.74
General Entertainment	7.28

- Cluster 2, blue dots on the map

Venues	Percentage%
Hotel	16.96
Zoo	11.73
Fried Chicken Joint	10.20
Fujian Restaurant	10.08
Furniture/Home Store	9.44
French Restaurant	9.31
Garden	9.31
Food	9.06
Gastropub	7.53
Food Truck	6.38

Results

Uniformity among clusters and within the region

As we can see from the table and the cluster map, red dots (cluster 0) account for the majority (63%) of east Chinese cities and scatter across the whole map. Also, when drilling down to top 10 popular venues in each cluster, it is obvious that 6 venues – 'Zoo', Fried Chicken Joint', 'Fujian Restaurant', 'French Restaurant', 'Garden', 'Furniture / Home Store' – out of the total 10 are identified in all three clusters.
These two facts indicate that there is an apparent uniformity of city types and lifestyles in the east China region.

Differences among clusters

Let's shift our focus from similarity to differences. The first thing that stands out from the map is the distribution of green dots (cluster 1). It is obvious that green cities are located mainly in the regions which are relatively far away from coastline. Interestingly, when looking at the top 10 most common venues for three clusters, 'Train Station' only appears in the green group and it ranks on the very top of that list.

Interesting facts about food

'Fried Chicken Joint', 'Fujian Restaurant', and 'French restaurant' are listed in all three city groups, and 'coffee shop' only appears among the top 10 in the red cluster, whereas gastropub are only identified in the rest 2 groups.

Discussion

The uniformity

The uniformity pattern, which is backed by the dominant distribution of red cities, and the same 6 popular venues among the top 10 list for all groups, suggest that a considerable extent of consistency across among cities in east China, in terms of the venues people visit, which can be interpreted as one aspect of lifestyle. The existence of uniformity can potentially be explained by the high level of average GDP per capita, convenience of public transportation, and increasing level of intra-regional cooperation, which significantly contribute to the communication between different cities, thus reducing differences in lifestyles.

The differences

However, significant differences among city groups were still identified by this research, especially the geographical distribution of the green cities, which are relatively far away from coastline, whose group has 'train station' listed on the top of the most common venues. Being closer to the sea usually is a privilege in terms of easier access to resources provided by the ocean, and better opportunities to international trade, which is part of the reason why east China became the richest area in China. Inner land cities lack those advantages mentioned and the economy there are not be as robust as their coastal neighbors, and the fact that 'train station' being on the top may indicate that people frequently travel to or from these cities, for

example, young generation frequently travel outbound for work opportunities or tourist travel inbound for a day trip.

What is more, the difference between food venues among three groups is worth discussing. It is noticed that all three groups have 'Fujian Restaurant' and 'French restaurant' in common listed in the top 10, but only in the red group that 'French restaurant' precedes 'Fujian restaurant'. French restaurants are usually considered high-end and luxurious, whereas Fujian restaurant is relatively normal and cheap. The former one being more popular over the latter demonstrated considerable consuming power for red cities, and the dominant percentage of the red means that this group can represent the whole east China region. And take a closer look at the list, 'Coffee Shop' ranks 4th in the list in red group, while it did not even make to the top 10 in other 2 groups. Chinese coffee market is still premature, which means lack of supply and high prices, thus coffee shops' appearing on the top of venues echoes with the leading position of French restaurants.

Conclusion

We now understand that cities in the east China can be separated into 3 different groups.

By looking at the representative red cluster, we can gain a deeper insight into the lifestyle of the whole region. The fact that 'French Restaurant' is more popular than 'Fujian Restaurant' and 'Coffee Shop' being on the top indicates that people in east China are happy and able to accept new lifestyles, like drinking a cup of currently overpriced coffee. As of the time when I write this report, *Luckin Coffee*, a Chinese domestic franchise, has expanded all over the east China as well as other major Chinese cities, accompanied by their skyrocketing share prices. I think this research could offer an interesting perspective to help explain why.

Apart from the apparent uniformity, variations among cities cannot be ignored. Based on the analysis I have done, the differences among city groups can be attributed to people's buying power, which is related to the robustness of local economy. Although east China as a whole is by the East Sea and is a rich area, inner land cities are placed in relatively underprivileged position.