Tourist Guide in Paris Using K-Means Clustering

1. Introduction

1.1. Background

Paris and its three surrounding departments received over 24 million visitors every year. For centuries, Paris has attracted artists from around the world, who arrive in the city to educate themselves and to seek inspiration from its vast pool of artistic resources and galleries. As a result, Paris has acquired a reputation as the "City of Art".

1.2. Business problem

It usually takes long to plan trips: we have to consider different categories of tourist attractions, art collections, museums, entertainment, outdoor activities, restaurants, etc. There is no easy way to find a visiting plan covering all tourists with different preferences. In this article, I will introduce an approach to make tourist guide for each borough in Paris using K-Means clustering, and I will also make restaurant recommendations according to cuisine types. This project might be of interest to tourists who would like to visit Paris.

2. Data Collection

I collected borough number, borough name, population and area size data from wikipedia website (https://fr.wikipedia.org/wiki/Arrondissements_de_Paris). I used BeautifulSoup library to scrape data from the table. Geographical coordinates of each borough were collected using the arcgis api in the Geocoder Python package. Data collected or scraped from multiple sources were combined into one table. Figure 1 shows the borough information data containing borough number, borough name, area, population, population density, postal code, latitude and longitude.

BoroughNumber	BoroughName	Area	Population	PopulationDensity	PostalCode	Latitude	Longitude
1	Louvre	1.83	16395	8959.016393	75001	48.8634	2.33677
2	Bourse	0.99	21042	21254.545455	75002	48.8677	2.34309
3	Temple	1.17	34389	29392.307692	75003	48.8626	2.35905
4	Hôtel-de-Ville	1.60	28370	17731.250000	75004	48.8543	2.36147
5	Panthéon	2.54	59631	23476.771654	75005	48.8454	2.35189

Figure 1. borough information data

Tourist attraction sites, restaurants and venues data were collected using Foursquare API. In this project, I collected several categories of venues: food, arts & entertainment, outdoors & recreation, spiritual center and clothing store.

3. Methodology

3.1. Restaurant recommendations

3.1.1. Data cleaning

I called Foursquare API to collect restaurant venues data, which contains restaurant name, cuisine category, venue location and borough number. There are several problems with the dataset. I had to clean the dataset before analyzing.

First, there were over a hundred categories; some of them have only a few restaurants, making it impossible to analyze and categorize. Therefore, cuisine categories with less than 15 restaurants were dropped from the dataset.

Second, there were categories which either belong to several cuisines or cannot be categorized as a cuisine such as ‘Diner’, ‘Breakfast’, 'Asian Restaurant', 'Seafood Restaurant', 'Sandwich Place', etc. Therefore, cuisine categories which are ambiguous were dropped from the dataset.

Third, there were some cuisine categories which belong to others: ‘Creperie’ is a type of French cuisine; ‘Sushi ’ is a type of Japanese cuisine; ‘Pizza’ is a type of Italian cuisine, etc. Therefore, I grouped the categories which are the same type of cuisine.

After the procedure of data cleaning, the dataset contains eight cuisine categories: ‘French Restauran t', 'Italian Restaurant', 'Japanese Restaurant', 'Vietnamese Restaurant', 'Thai Restaurant', 'Chinese Re staurant', 'Korean Restaurant', 'Indian Restaurant’.

3.1.2. One hot encoding

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. I converted the dataset using one hot encoding so that I can calculate the percentage of cuisine categories in each borough. Then I wrote scripts to sort the top 3 boroughs with the most number of restaurants for each cuisine category, as shown in figure2.

	1st Borough	2nd Borough	3rd Borough	1st Borough Percentage	2nd Borough Percentage	3rd Borough Percentage
Chinese Restaurant	3	13	20	0.214286	0.214286	0.0714286
French Restaurant	7	5	14	0.0817942	0.0791557	0.0765172
Indian Restaurant	14	15	2	0.2	0.15	0.1
Italian Restaurant	17	2	15	0.100592	0.0769231	0.0769231
Japanese Restaurant	20	9	17	0.0869565	0.0869565	0.0869565
Korean Restaurant	15	10	14	0.238095	0.0952381	0.0952381
Thai Restaurant	13	2	4	0.34375	0.09375	0.09375
Vietnamese Restaurant	13	5	3	0.487179	0.102564	0.0769231

Figure2. Top 3 boroughs with the most number of restaurants for each cuisine category

3.1.3. Data visualization

I plotted the graph of percentage of each cuisine in most frequent boroughs using matplotlib library. As illustrated in figure 3, for each cuisine category, top 3 boroughs are listed with the percentage of the number of restaurants. For example, almost half of Vietnamese restaurants, a third of Thai restaurants and a quarter of Chinese restaurants are located in the borough 13, because there are a lot of Asian neighborhoods located in the borough 13.

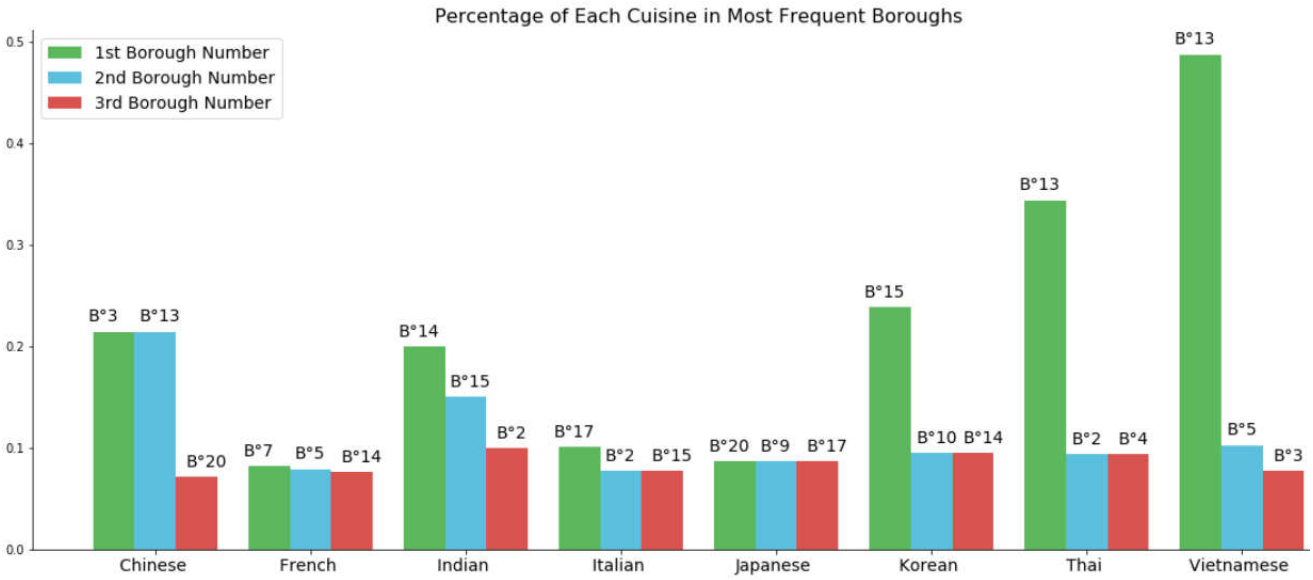

Figure3. Percentage of Each Cuisine in Most Frequent Boroughs

3.2. Tourist attractions

3.2.1. Data cleaning

I used also Foursquare API to collect tourist attractions venues data, which contains venue name, venue category, venue location and borough number. Since I wanted to focus on tourist attractions rather than residential information like supermarkets, laundry service, etc, I collected several categories of venues related to tourist attractions: arts & entertainment, outdoors & recreation, spiritual center and clothing store. There are 224 venues in branch arts and entertainment; 463 venues in branch outdoors and recreation; 86 venues in branch spiritual center; 420 venues in branch clothing store. There are several problems with the dataset. I had to clean the dataset before clustering.

First, there are 102 categories in total; some of them have only a few venues, making it impossible to analyze and categorize. Therefore, categories with less than 5 venues were dropped from the dataset.

Second, there are some categories not related to tourist attractions such as Optical Shop and Multiplex. Therefore, categories which are not related to tourist attractions were dropped from the dataset.

Third, there are a lot of similar categories that can be grouped in the category list. Therefore, I grouped Shoe Store, Women's Store, Men's Store, Kids Store, Boutique, Lingerie Store, Baby Store,

Accessories Store into Clothing Store; grouped Art Gallery, Art Museum, Museum, History Museum, Spiritual Center, Exhibit into Art and Museum; group Gym / Fitness Center, Gym, Tennis Court, Pool, Yoga Studio, Athletics & Sports, Martial Arts Dojo into Sport; grouped Performing Arts Venue, Theater, Dance Studio, Movie Theater, Music Venue, Indie Movie Theater, Concert Hall, Comedy Club into Entertainment; grouped Garden, Park, Playground, Fountain, Outdoor Sculpture, Lake into Outdoors and Nature; grouped Pedestrian Plaza into Plaza.

After the procedure of data cleaning, the dataset contains seven venue categories: 'Clothing Store', 'Sport', 'Plaza', 'Outdoors and Nature', 'Art and Museum', 'Entertainment', 'Church'.

3.2.2. One hot encoding

I converted the dataset to a form with all categories as columns for clustering. I calculated the percentage of venue categories in each borough. It is a distribution of each venue category in all boroughs, as illustrated in figure 4.

BoroughNumber	Art and Museum	Church	Clothing Store	Entertainment	Outdoors and Nature	Plaza	Sport
1	0.136364	0.032258	0.067708	0.032609	0.038462	0.088235	0.014599
2	0.018182	0.080645	0.085938	0.097826	0.007692	0.051471	0.058394
3	0.163636	0.032258	0.101562	0.010870	0.030769	0.000000	0.014599
4	0.090909	0.016129	0.132812	0.043478	0.107692	0.051471	0.014599

Figure 4. Distribution of each venue category in all boroughs

Then I wrote scripts to sort the most common categories of venues in each borough, as shown in figure 5.

BoroughNumber	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
1	Art and Museum	Plaza	Clothing Store	Outdoors and Nature	Entertainment	Church	Sport
2	Entertainment	Clothing Store	Church	Sport	Plaza	Art and Museum	Outdoors and Nature
3	Art and Museum	Clothing Store	Church	Outdoors and Nature	Sport	Entertainment	Plaza
4	Clothing Store	Outdoors and Nature	Art and Museum	Plaza	Entertainment	Church	Sport
5	Church	Plaza	Entertainment	Art and Museum	Outdoors and Nature	Sport	Clothing Store

Figure 5. The most common categories of venues in each borough

3.2.3. Clustering

I performed clustering with the dataset using K-Means. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. Therefore, average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point. I chose 4 as the optimal value of K in K-Means Clustering.

The map of clustering is illustrated in figure 6. The red color represents boroughs in cluster 1; the blue color represents boroughs in cluster 2; the green color represents boroughs in cluster 3; the yellow color represents boroughs in cluster 4.

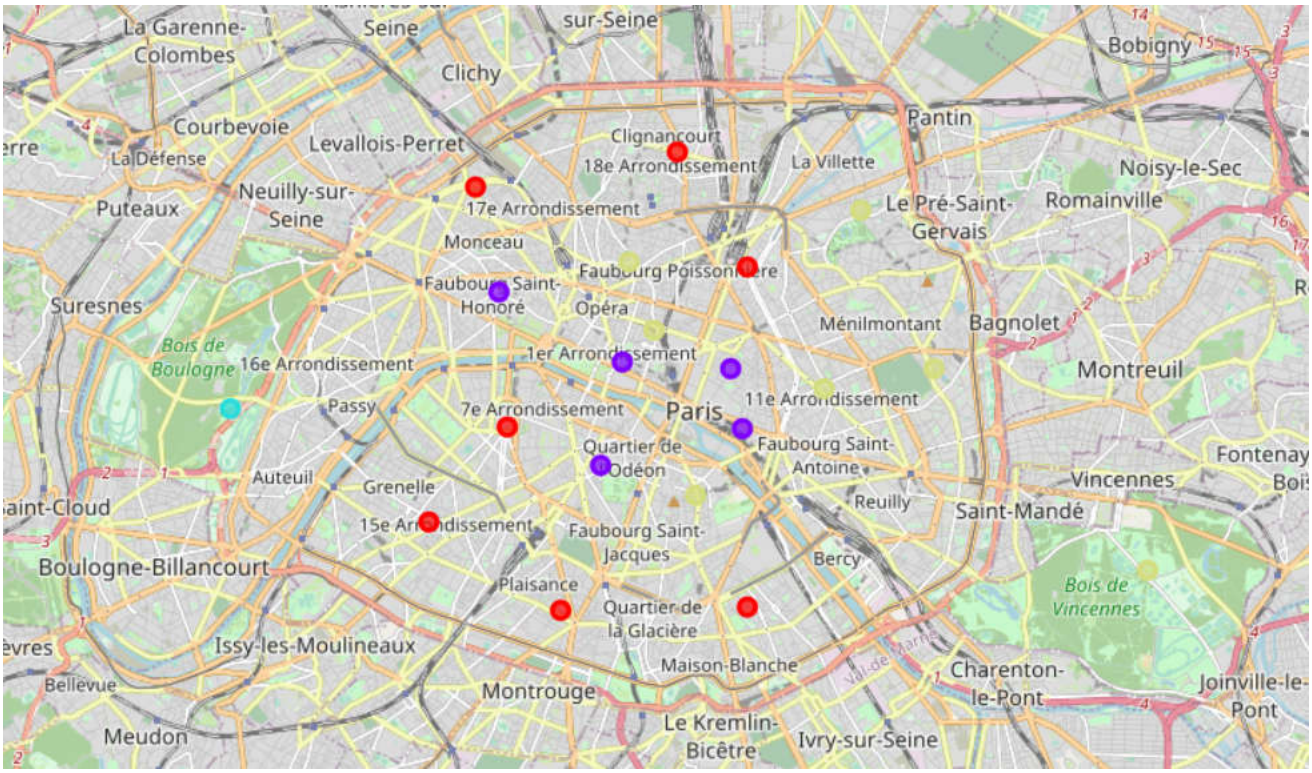

Figure 6. Map of clustering

4. Results

4.1. Restaurant recommendations

As illustrated in figure 3, for each cuisine category, top 3 boroughs are listed with the percentage of the number of restaurants. I displayed the boxplot using seaborn library. Figure 7 shows the percentage distribution for each cuisine category.

According to figure 7, we can tell that on one hand, restaurants for French cuisine, Italian cuisine and Japanese cuisine are evenly distributed in the city because they have small inter quartile range and few outliers; on the other hand, restaurants for Chinese cuisine, Indian cuisine, Korean cuisine, Thai cuisine and Vietnamese cuisine are likely located in specific boroughs in the city because they have large inter quartile range and more outliers.

According to figure 3, we can tell that almost half of Vietnamese restaurants, a third of Thai restaurants and a quarter of Chinese restaurants are located in the borough 13, because there are a lot of Asian neighborhoods located in the borough 13; 20 percent of Indian restaurants are located in the borough 14, and a quarter of Korean restaurants are located in the borough 15.

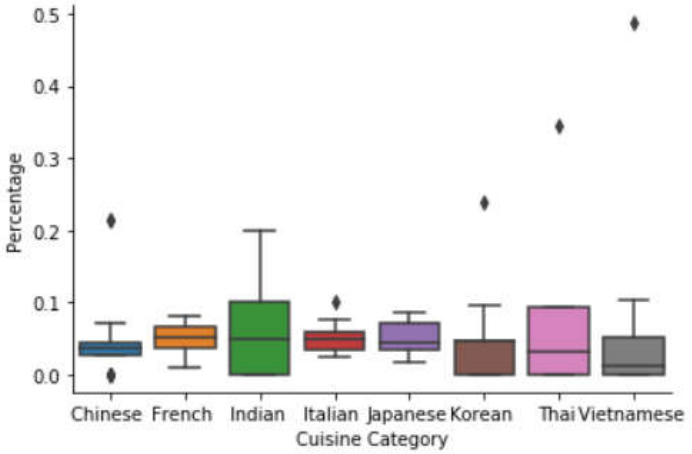

Figure 7. The percentage distribution for each cuisine category

4.2. Tourist attractions

I applied K-Means clustering to the tourist attractions dataset, the results of four clusters are illustrated in figure 8, 9, 10, 11. In cluster 1, boroughs are more likely residential boroughs with a lot of sport venues, plaza venues, and churches. Cluster 2 contains boroughs located in center city with stores, art venues, and museums. Cluster 3 contains the borough 16 which is located in the forest with a lot of sport venues, outdoors and nature venues. Cluster 4 contains boroughs which provide most entertainment venues in Paris such as concert halls, theaters and music venues.

BoroughNumber	ClusterLabels	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
7	0	Art and Museum	Plaza	Outdoors and Nature	Church	Entertainment	Clothing Store	Sport
10	0	Outdoors and Nature	Plaza	Clothing Store	Art and Museum	Church	Sport	Entertainment
13	0	Clothing Store	Sport	Church	Entertainment	Outdoors and Nature	Plaza	Art and Museum
14	0	Sport	Plaza	Outdoors and Nature	Entertainment	Clothing Store	Church	Art and Museum
15	0	Sport	Clothing Store	Plaza	Outdoors and Nature	Church	Entertainment	Art and Museum
17	0	Plaza	Sport	Church	Outdoors and Nature	Clothing Store	Entertainment	Art and Museum
18	0	Church	Sport	Outdoors and Nature	Plaza	Entertainment	Art and Museum	Clothing Store

Figure 8. Cluster 1 tourist attractions

BoroughNumber	ClusterLabels	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
1	1	Art and Museum	Plaza	Clothing Store	Outdoors and Nature	Entertainment	Church	Sport
3	1	Art and Museum	Clothing Store	Church	Outdoors and Nature	Sport	Entertainment	Plaza
4	1	Clothing Store	Outdoors and Nature	Art and Museum	Plaza	Entertainment	Church	Sport
6	1	Clothing Store	Entertainment	Outdoors and Nature	Art and Museum	Plaza	Church	Sport
8	1	Clothing Store	Art and Museum	Entertainment	Plaza	Church	Sport	Outdoors and Nature

Figure 9. Cluster 2 tourist attractions

BoroughNumber	ClusterLabels	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
16	2	Sport	Outdoors and Nature	Plaza	Church	Art and Museum	Clothing Store	Entertainment

Figure 10. Cluster 3 tourist attractions

BoroughNumber	ClusterLabels	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
2	3	Entertainment	Clothing Store	Church	Sport	Plaza	Art and Museum	Outdoors and Nature
5	3	Church	Plaza	Entertainment	Art and Museum	Outdoors and Nature	Sport	Clothing Store
9	3	Entertainment	Art and Museum	Sport	Plaza	Church	Clothing Store	Outdoors and Nature
11	3	Entertainment	Church	Outdoors and Nature	Sport	Art and Museum	Plaza	Clothing Store
12	3	Entertainment	Outdoors and Nature	Sport	Church	Clothing Store	Art and Museum	Plaza
19	3	Church	Sport	Entertainment	Outdoors and Nature	Plaza	Art and Museum	Clothing Store
20	3	Entertainment	Outdoors and Nature	Plaza	Church	Sport	Clothing Store	Art and Museum

Figure 11. Cluster 4 tourist attractions

5. Discussion

The results of restaurant recommendations and tourist attractions clustering are rather reasonable.

According to the study of restaurant recommendation, restaurants for French cuisine, Italian cuisine and Japanese cuisine are evenly distributed in the city. This matches the real situation because those are the most common cuisines in the city. Most Asian restaurants are located in the borough 13, because there are a lot of Asian neighborhoods located in the borough 13.

According to the study of tourist attractions clustering, the clustering results are able to identify the residential boroughs, entertainment boroughs, and boroughs located in center city with stores, art venues, and museums.

However, there are still several aspects for improvement. The venue data collected from Foursquare API are not always precise on venue categories, which can lead to ambiguous categorization. The venue data collected doesn't contain the popularity, the size and the number of visitors per year, which may cause the clustering result not accurate enough (e.g., Musée du Louvre attracts far more tourists than other museums, however, all museums have the same weight in the clustering algorithm). Future research can take into consideration of other variables such as popularity, size and number of visitors of each venue.

6. Conclusion

In this project, I analyzed the distribution of the eight most popular cuisine categories in twenty boroughs of Paris, as well as the distribution of tourist attraction venues in Paris. With the help of graphic illustrations, data tables and visualization maps, I made recommendations on both restoration and tourist attractions for tourists with different preferences. I hope this project can be useful for tourists as well as people who are interested in data science.