Coursera - IBM Data Science Professional Certificate

Assignment - The Battle of Neighborhoods
Paris vs Manhattan

Introduction

Imagine we have a friend who lives in Paris and who plans to travel over to US for a business trip and stop-over in New York for a long weekend and enjoy the city.

When in Paris, he likes to hang out in Saint Germain, a district where you can find plethora of cafes, brasseries, bars, and restaurants. You also find art galleries, antique stores, theaters and cinemas, street markets, parks, clothing retailers and other shops. Saint Germain in quite central of Paris and located in the 6th arrondissement.

Now, he would like to find the same in New York but doesn't know much about that city. Where is the Saint Germain of New York? What neighborhood of New York is similar to Saint Germain?

To answer questions like the one asked by our friend, we analyze here data from Foursquare, a popular location-based social network. Foursquare enables users to share their current location with friends, rate and comment on venues they visit (places such as restaurants, hotels, cafeterias, bookshops, and museums) and read reviews of venues that other users have left.

For the purposes of this project, we use geographical datasets of the two cities New York, more specifically Manhattan, and Paris. We take the neighborhoods to be areas on a city's geographic map. We then inquire the venues recommendations from Foursquare through their API. Each area is associated with the set of venues within predefined boundaries, based on our manually collected ground-truth data. In turn, each venue is associated with a name, a category (e.g. restaurant, coffee shop, bar, museum, art gallery, etc.), and a geographic location (latitude and longitude). We got about 1,800 venues listed for Paris and 3,300 venues for Manhattan.

Data

Paris geographical data

Paris is divided into 20 administrative districts, referred to as arrondissements. The 20 arrondissements are arranged in the form of a clockwise spiral (often likened to a snail shell), starting from the center of the city.
Among landmarks of Paris, we can find the **Louvre Museum** in the 1st arrondissement, the **Museum of Modern Art (Georges Pompidou)** in the 3rd arrondissement, **Notre Dame Cathedral** in the 4th arrondissement, the **Eiffel Tower** in the 7th arrondissement, the **Champs-Elysées Avenue** in the 8th arrondissement, the **Arc of Triumph** at the limits of the 8th, 16th, and 17th arrondissements, and so on.

New York geographical data

Geographically, New York is a city with 5 boroughs, which are the Bronx, Brooklyn, Manhattan, Queens, and Staten Island, and hundreds of neighborhoods.

When we think of New York City, Manhattan is often the first place we picture. The borough is home to well-known attractions, such as **Central Park**, the **Empire State Building**, world-class museums such as **MOMA**, the bright lights of **Times Square** and **Broadway**. Manhattan contains big-name neighborhoods, international restaurants, classy boutiques, trendy bars and more.

Venues recommendations API by Foursquare

Foursquare features a developer API that lets developpers applications make use of Foursquare's location data. Their API powers various geo-enabled searches of venues with sophisticated details (e.g. tips, hours, menus, stats over time), searches of users, checkins, etc.

In the scope of the resources allocated to this project, we just use the API feature of exploring top recommended venues nearby a location that returns basic venue data (name, location, etc.), category, and ID.

Combining the geographical cities data with Foursquare venues

The approach is to gather venues data from Foursquare based on our two cities neighborhoods. We then use the **venues categories** feature for clustering the neighborhoods and look for basic similarities between these neighborhoods of Paris and Manhattan.

Datasets Exploration and Setting

In this section, we perform **data wrangling**, as below:

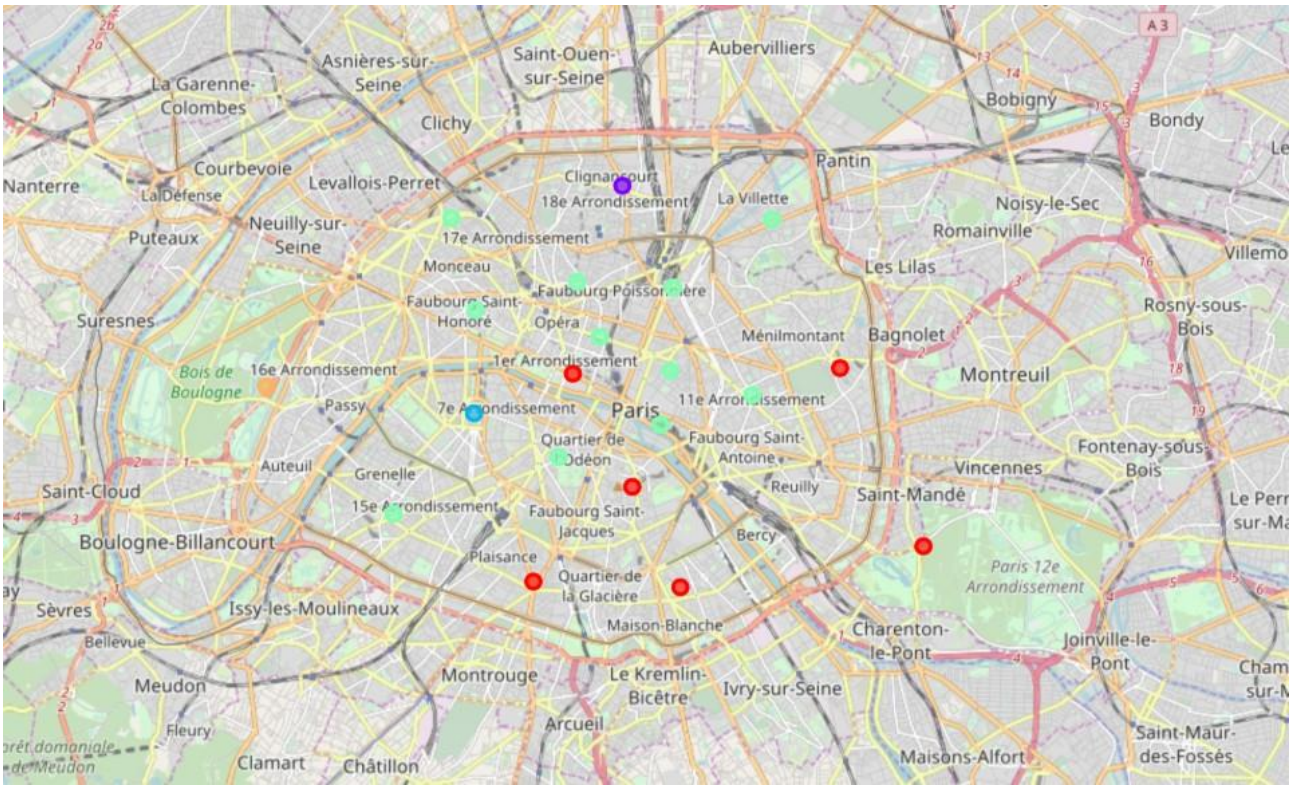
1. Geographical data of Paris arrondissements and New York boroughs/neighborhoods.
    - Extracting neighborhood geographical information for the cities of Paris and New York from json files (refer to the links provided in the Data section above).
    - Creating Pandas dataframes that include:
        - names of the boroughs/arrondissements,
        - names of the neighborhoods,
        - locations of the neighborhoods in terms of latitude and longitude.
    - Visualizing the neighborhoods in on a Leaflet map via the Folium library.

1. Foursquare recommended venues data based on neighborhoods geolocalisation.
    - Gathering venues recommendations through the explore endpoint API.
    - Creating Pandas dataframes that include:
        - names of the boroughs/arrondissements,
        - names of the neighborhoods,
        - locations of the neighborhoods in terms of latitude and longitude,
        - the names of the venues,
        - locations of the venues in terms of latitude and longitude,
        - categories of the venues.
        -

Discussion

Now that we have the contents of our clusters within each city, we want to identify these clusters by discriminating the venues that distinguish each cluster.

About Paris



We have five clusters in Paris with three of them, namely clusters 3, 4 and 5, containing only 1 neighborhood.

Based on the image above, the colors code is as follows:

```
cluster 1: red
cluster 2: purple
cluster 3: blue
cluster 4: green
cluster 5: orange (west of Paris)
```

1. City of culinary experience
   Restaurants and bistros, which are typical Parisian restaurants serving mostly daily French food, are the most recommended common venues **by Foursquare users**.

We can find restaurants everywhere, in every cluster, accounting for about 40-45% of the top 10 recommended venues in the two main clusters, i.e. clusters 1 and 2. French cuisine is obviously predominant (note that every cluster has 10% of French restaurants) followed by Italian cuisine and Japanese cuisine. There is no surprise about French and Italian food, which could be considered as local taste, being popular but Japanese cuisine seems to be also trendy in Paris.

1. Paris, a cool place to drink too

   This city is famed for its dining but it's also a cool place to drink too.

   The bar scene is diverse with wine bars, beer bars, cocktail bars, pubs and so on, as recommended venues. Except in cluster 4 (only 1 neighborhood, which by the way is remote from downtown), we can also find bars in every cluster. The two main clusters have similar rate of recommendations for bars, at about 12%.

1. Café and baguette

   Paris has also a plethora of cafés/coffee shops and bakeries. These venues account for 14-17% of recommendations in clusters 1 and 2. Here also, except in cluster 4, we can find cafés and bakeries in every cluster.

1. Accommodation

   Clusters 1 and 2 cover 85% (17 arrondissements out of 20) of the Paris area, so not surprisingly cover almost all the hotels in our recommendations, amounting about 7-8% of the venues in each cluster.

Discriminating clusters.

- Cluster 1 and cluster 4:

  These are the two main clusters with 6 neighborhoods for cluster 1 and 11 neighborhoods for cluster 4. We can observe that these clusters are alike with regard to the rate of restaurants, cafés and accommodation. So we can barely distinguish them. The main difference is that cluster 4 contains museums, historic landmarks, which are also famous features of Paris, while cluster 1 has none.

  Cluster 1: Intellectual, lively and trendy travelers.

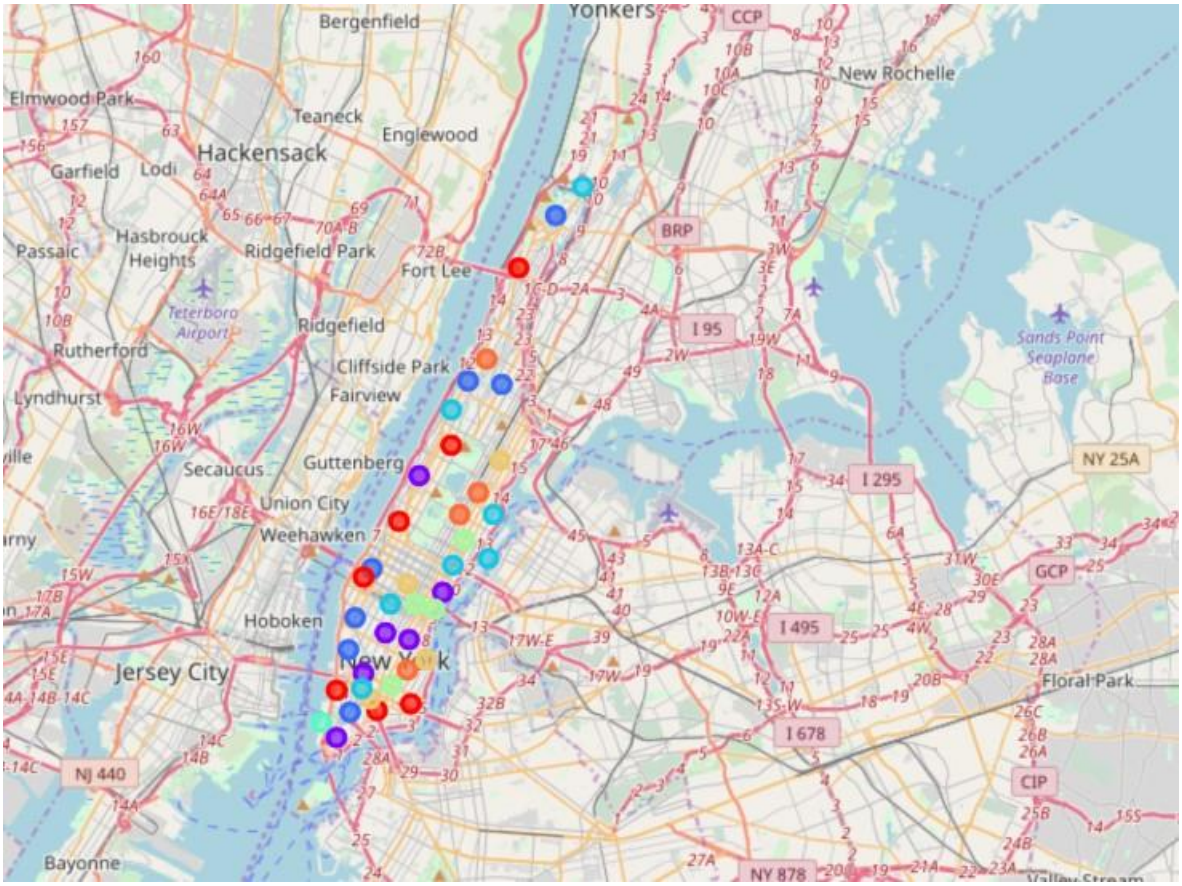
  Cluster 4: History, art and fashion lovers.

- Cluster 2: Buttes-Montmartre could be good for bohemian travelers because of the various styles of food venues.

- Cluster 3: Only 1 neighborhood, Palais-Bourbon in the 7th arrondissement. A cluster with history and culture.

- Cluster 5: Passy neighborhood at the west side of Paris, quite remote, with a few venues for sport enthusiasts.

About Manhattan



We have eight clusters in Manhattan. Based on the image above, the colors code is as follows:

```
cluster 1: red
cluster 2: purple
cluster 3: blue
cluster 4: blue-green
cluster 5: light green
cluster 6: green
cluster 7: light orange
cluster 8: orange
```

1. Not surprisingly, restaurants and food places are largely rated among clusters:

```
Cluster 1: 36%
Cluster 2: 51%
Cluster 3: 38%
Cluster 4: 37%
```

```
Cluster 5: 20%
Cluster 6: 50%
Cluster 7: 40%
Cluster 8: 30%
```

We may note that a part (24% in cluster 4) of the food venues are pizza, sandwich and burger places, which also tells about the Manhattan style of eateries.

1. Bars are also in every cluster, except in cluster 5, which has only 1 neighborhood. The average of bars then is about 7% of the recommendations, reaching 12.5% in cluster 8.

1. Cafés/coffee shops, bakeries and alike shops cover between 10 and 20% of the venues.

1. Accommodation recommendations are a bit lower compared to Paris. Intuitively, we may think that Paris is more rated by travelers.

1. Gym / Fitness centers are largely more rated than in Paris. This confirms that rating in Manhattan is more from locals.

Discriminating clusters.

- Cluster 1: A large part of recommendations for cultural and entertainment venues, at about 9%.
-
- Cluster 2: More than 50% of the recommendations are about eateries with 43% of restaurants, mainly Italian cuisine.

- Cluster 3: A mix of everything with a wide variety of restaurants.

- Cluster 4: Lots of stores in this cluster.

- Cluster 5: Only 1 neighborhood, Battery Park City at the far south-west of Manhattan.

- Cluster 6: Similar to Cluster 2, with about 50% of eateries. With 3/4 neighborhoods on the shores, seafood and sushi restaurants are popular.

- Cluster 7: Similar to cluster 6 but with less restaurants (30% here vs 40% in cluster 6)

- Cluster 8: It could be a more residential cluster with the highest rate of gym venues amounting 12.5% of the cluster.

Conclusion

In this report, we looked at the problem of matching neighborhoods across Paris and Manhattan using Foursquare venues recommendations data. We just used simple metrics such as rates of venues listed in neighborhoods for comparison. When evaluating against ground truth data, we found that there are no clear similarities between clusters found in Paris and those found in Manhattan.

There would be a few reasons for these limitations.

- Featuring.

  Our Foursquare dataset include only the venues categories as one feature for clustering. We could use more features such as checkins and possibly user profiles.

  We could also add more datasets including for example demographics, transportation, and real estate.

- Better use of geographical data.

  For example, the Paris dataset is based on segmentation by arrondissement, which is not the best approach as actual neighborhoods with regard to venues could be located on more than one (overlapping) arrondissement. Also the administrative center of the arrondissement may not be the actual center of the neighborhood.

  So, we should provide a better match for the selected neighborhoods with the ground truth neighborhoods.

- Choosing better clustering algorithm.

  In this report, we used k-Means model but we could also try other models such as DBSCAN and hierarchical clustering. A few other approaches could be used too but they are beyond the level of this report.

  Note that with the k-Means model, we can empirically adjust the radius of the neighborhood but the results won't be significantly improved.

Having this said, the approach shown here in this report would be better used for comparing neighborhoods within a city but not comparing across cities.