Relationship between Manhattan and the city of Paris

Jesús Ponce

November 6, 2020

1 Introduction

1.1 Background

Although Paris was founded years before New York, both cities are very popular for their multicultural spaces including restaurants, cafés, galleries, bars, among others. The real differences between the two cities are about quality of life and the way people view their relationship to both others and the world. You may find this information way more useful if you reside in either Paris or Manhattan.

1.2 Problem

If you are willing relocate, to open a new venue, to visit, to study you may want to know beforehand what type of venues the cities hold and what you are going to be able to visit/find. Therefore, it is interesting to find their similarities and differences in terms of venues. It is always better to know what you will be seeing once you move to the city.

1.3 Interest

Stakeholders analyzing the market location would be interested in knowing what kind of market they will be moving onto, as well as exchange students or vacationers who are interested in either cities.

2 Data acquisition and cleaning

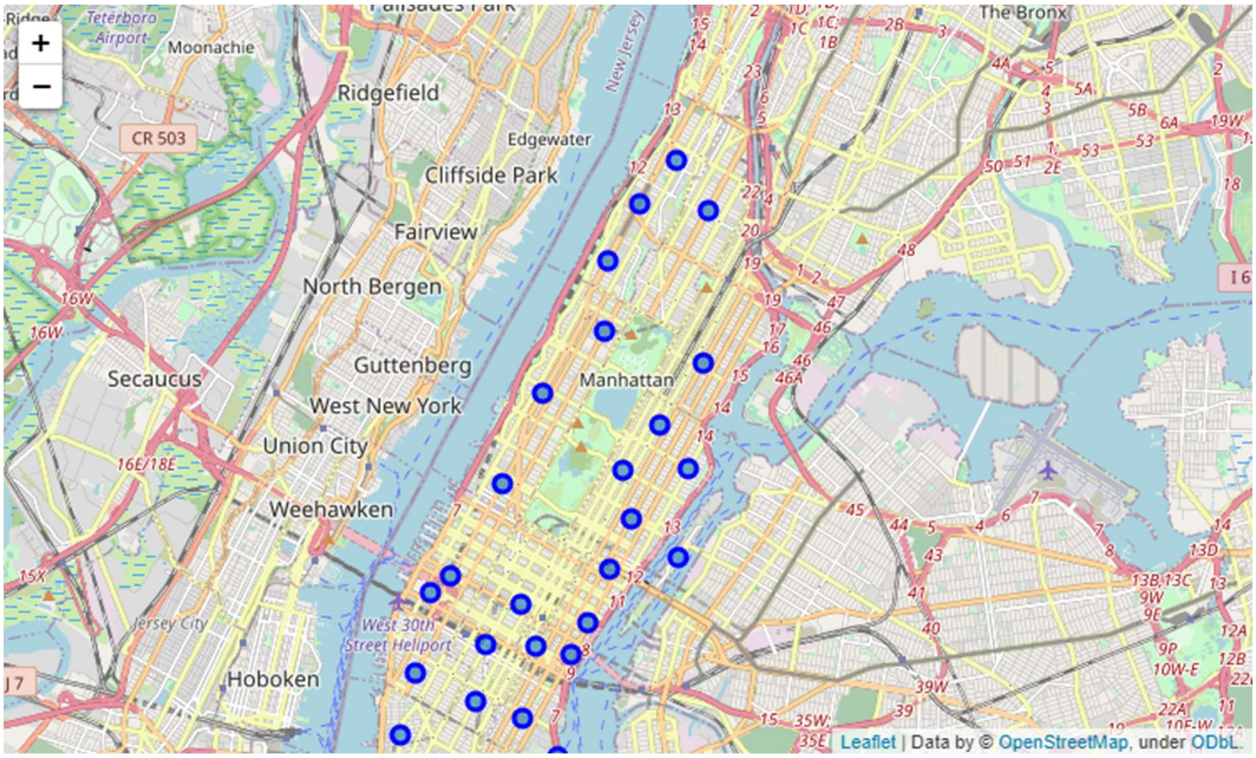
2.1 Data sources

The data used for this analysis was obtained from Coursera's newyork_data.json and from Wikipedia for Paris arrondissements. I had to scrap the data set from Wikipedia and in order to obtain the venues we used the Foursquare API for both cities. Along with this, to finally obtain the coordinates for venues and cities I used the geopy library.

2.2 Data cleaning

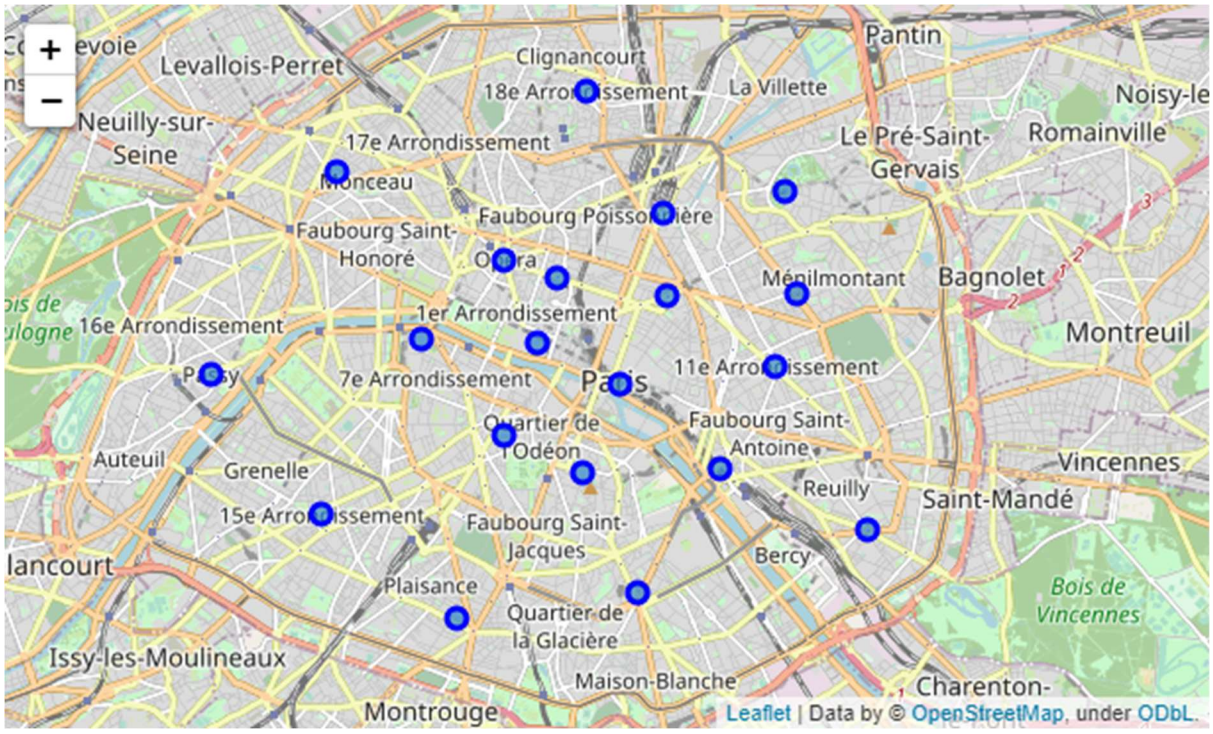
Data cleaning is one of the most important steps before applying a machine learning model. For this study, I had to do different cleaning for each city. In the case of New York, I had to choose a borough since the city has many and listing all the venues of all the boroughs would take all of our API's calls, Manhattan

was my choice. After obtaining all neighborhoods in Manhattan and creating a dataframe, I obtained the latitude and longitude using the geopy library. With this new dataframe I used the Foursquare API to get all the venues for the neighborhoods in the dataframe. This gave me a new dataframe which I used to joined with the first dataframe that contained all of the neighborhoods in Manhattan.



Manhattan Neighborhoods

Data cleaning Paris was a little bit more interesting; I scrapped the data from the Wikipedia page for Paris, the table scrapped included photos from the mayor, name, population, density, and area of each arrondissement. I had to remove unnecessary columns leaving the name, and number of arrondissements. Unfortunately, the dataframe had the first 4 arrondissements in the first row so I had to drop that row and append new rows which included the name and number of the first 4 arrondissements to the dataframe. After that, obtaining the location of each arrondissement was easy with the function used for New York, although after analyzing the resulting dataframe I noticed that the geopy library was not getting the location of the 6th arrondissement accurately. This is a big mistake that has to be fixed or the entire analysis would not be correct. The 6th arrondissement is named *Luxembourg* which is also a country in Europe, geopy took this as me trying to look for Luxembourg's coordinates and returned them. In order to fix this issue, I had to search for the real coordinates for the 6th arrondissement and replace the data for latitude and longitude in the dataframe. The same procedure for getting the venues for Manhattan was used to obtain Paris' venues as well.



Paris Arrondissements

Finally, data had to be normalized. Both data frames contained all of the venues for each arrondissement and neighborhood so I used one hot encoding to normalize the data and for each neighborhood/arrondissement the frequency of the venue was converted into a column, creating a new dataframe which contained the name of the arrondissement/neighborhood as the row and for the columns the type of venue nearby (one hot encoded).

After applying the mean and grouping by neighborhood, I can order by top 10 venues categories for each neighborhood, which meant I was able to cluster them to see how close these neighborhoods are to each other in terms of venues.

3 Clustering Modeling

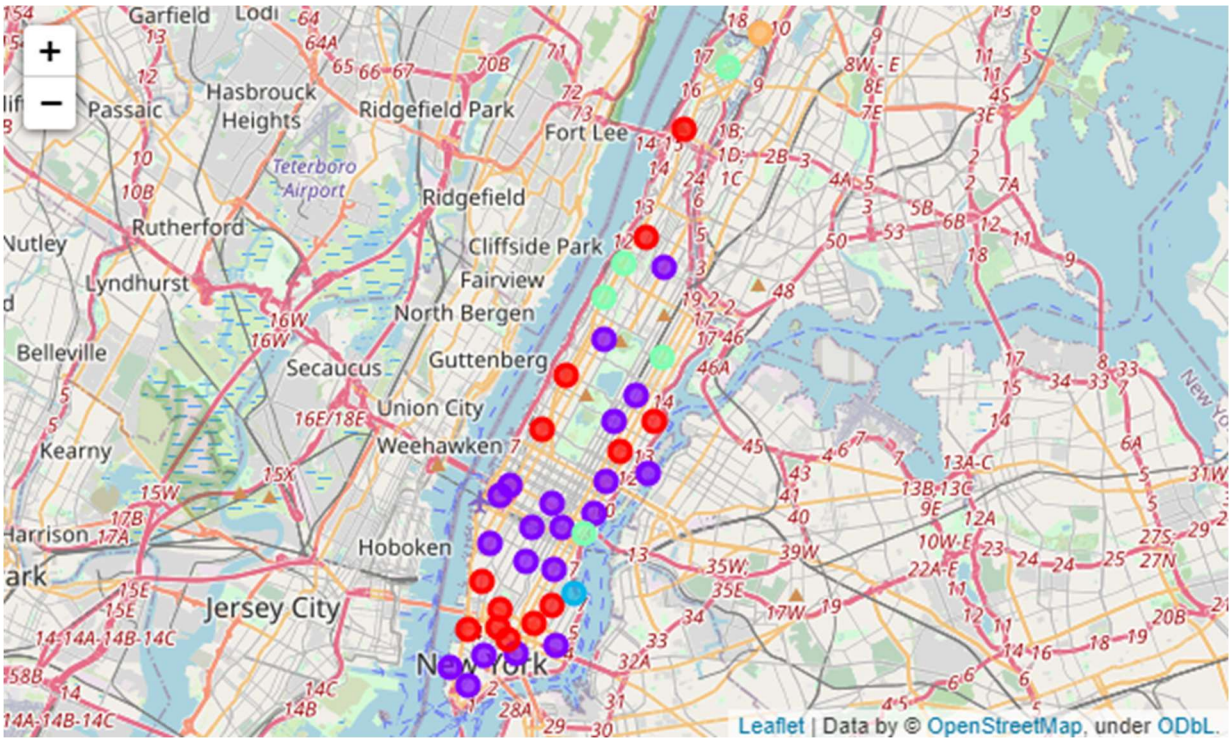
For both Manhattan and Paris, the neighborhoods were clustered into five. Five is a good number since we can differentiate way better for the number of neighborhoods both cities have, the larger the cluster the more differences we would notice and since we are looking for more similarities in both cities five cluster let us visually notice the similarities.

Clustering the neighborhoods meant clustering the frequency of type of venues each neighborhood had in common with others. We can see how the different types of venues interact with each other and also with the location in the map.

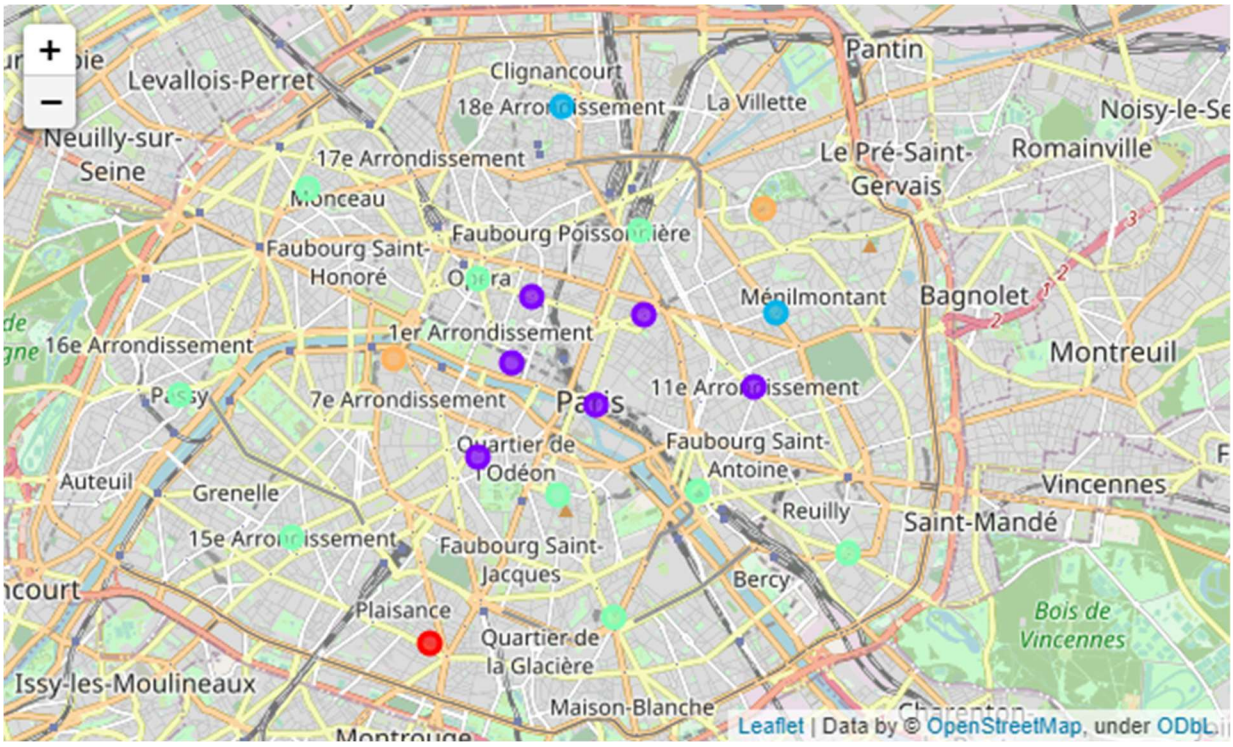
I used K-means clustering, which gave a label to each venue of the neighborhood. After obtaining each label, I inserted a new column with the information of the labels to the dataframe and plotted a new map

showing the neighborhoods in a different color depending of the number of the cluster (this cluster was decided with the frequency and top of venues for each neighborhood)

3.1 Manhattan



3.2 Paris



4 Conclusion

In this study, I analyzed the relationship between the venues of the cities of New York (specifically Manhattan) and Paris. I identified their neighborhoods/arrondissements with their corresponding venue and the top venues for each neighborhood/arrondissement. In order to obtain the relationship, I used k-means clustering which gave a label to each neighborhood. This model is very useful when trying to identify clusters within maps because it gives you a good idea on how close different places of the world can really be even if they are very far apart. This can help any person trying to identify where, what type, of venue they want to open, any student or vacationer visiting the city.

5 Insights

The city of Paris and Manhattan share a lot of similar places. Looking at the map with all the clusters, we are able to see that the purple cluster are dominant. This particular cluster ins interesting because it holds the more interesting spots in the cities, (Eiffel Tower, popular cafés for Paris and Central Park, 5th Avenue for Manhattan) meaning since a lot of visitors may be interested there, venues should be different and interesting for anyone; this same cluster include many restaurant for both places, restaurant that serve any type of food, from French, to Italian, to Korean; Ice-cream shops and plazas are also noticeably predominant in this cluster. These clusters also show work life venues, you will find more population since companies settled in the center of the city.

The green cluster is a cluster which for each city have a different significance. For Paris, the green one has a lot of hotels, this also interesting because as you can see in the map, the green cluster wraps up the purple one which makes a lot of sense. In the case of Manhattan, the green cluster are mainly parks and restaurants that aren't that recurrent: Mexican and seafood. These restaurants aren't that famous because Korean, Italian are dominant in the city of New York.

Finally, it is interesting to notice that Manhattan has a lot of red dot clusters in comparison with Paris. Upon analyzing this, I can see that the red cluster in Manhattan are nightlife venues. Nightlife in Manhattan isn't as vivid as Las Vegas, but these venues really hold up to Manhattan's nightlife making it a good and relaxing borough to visit when going to New York. Please take a look at this image, showing the population number for Manhattan at night (you can see the resemblance with the cluster map).