Final Capstone Project: Prague vs (Paris+Saint Denis)

Business Problem:

The aim of the Capstone project is to compare two cities, Paris and Prague. I have personal experience from both of these cities. I was born in Prague and I used to live in Paris for 3 years- in a part called Saint Denis. Both cities are the capital cities of their countries the city centers are very similar and very attractive for tourists. For this project I consider Paris as the center of Paris plus a part of Paris called Saint Denis. These both parts should have about 3 million inhabitants. The agglomeration of Paris would be too huge for this project, having more than 9 million of inhabitants. The center of Paris is very specific, it is historical, very expensive for living and there are many cafés, museums, tourists attractions and so on. About 2 million people live in the city Paris center, but it is mostly for tourists and extremely rich inhabitants. Saint Denis is a part of Paris where many “normal” people live. We can find many stadiums, restaurants, cafés, an airport and houses of blocks there. Cheap hotels and hostels are also there. Prague has 1.5 million inhabitants, but the density is a bit lower. In Prague we have also a very expensive city center, and in the parts of the city more far away from the city center there are many houses of blocks, stadiums, old factories and so on. I have experience from both of these cities and I think that they are it will be very interesting to compare them.

Paris and Prague are very beautiful and attractive cities for people all over the world. Many tourists visit these cities every year. This project should help people to decide which city to visit, how many tourist attractions there are and how long they should stay there. It also can be convenient and helpful for people who want to change their neighborhoods within the city. It can be also helpful for people thinking about relocating into one of these cities. Reality Investors could be also interested in this project, because in both of these cities the prices of living are rising. The idea is to look for venues in the different neighborhoods (radius of 500 meters), to cluster them based on the frequency of occurrence and compare them. In the end I will discuss the results.

This project is similar to the previous task we did in the course, clustering Toronto. I use this knowledge and cluster and compare these two European cities. All the necessary steps are described in more detail in the Notebook.

Data analysis:

Data are the essential part of every data science project. We need to have geographical coordinates for the neighborhoods of Paris and Prague. Then we will need to know what venues and how many of them there are in the neighborhoods (radius of 500 meters).

Prague Dataset:

We will use three sources of data. The first source is for the Prague dataset.

For the Prague neighborhood dataset I created an CSV. dataset, which is webscraped from Wikipedia. The dataset is now freely available here: https://www.kaggle.com/konecfil/prague-neighborhoods-dataset.

The first column is name of the neighborhood, the second and third are *Lat* and *Lon*, respectively. These coordinates will be used as centers of the neighborhoods.

To download this dataset we will use od.download("https://www.kaggle.com/konecfil/prague-neighborhoods-dataset").

It will ask us to insert an username and a key. It can be found on your kaggle account (you have to create an account and then it can be found if you click "your account". It will create a new directory with the file.

You also need to add "!pip install opendatasets" and

"import opendatasets as od" to work properly.

There are 57 neighborhoods in Prague.

Paris Dataset:

The Paris dataset is available here: https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e.

The JSON file is for the whole agglomeration of Paris, so we have to limit it to Paris and Saint Denis only.

Columns are : *postal_code:* Postal codes for France, *nom_comm:* Name of Neighborhoods in France, *nom_dept:* Name of the boroughs, *geo_point_2d:* Tuple containing the latitude and longitude of the Neighborhoods.

There are 60 Neighborhoods in Paris (Paris + Saint Denis).

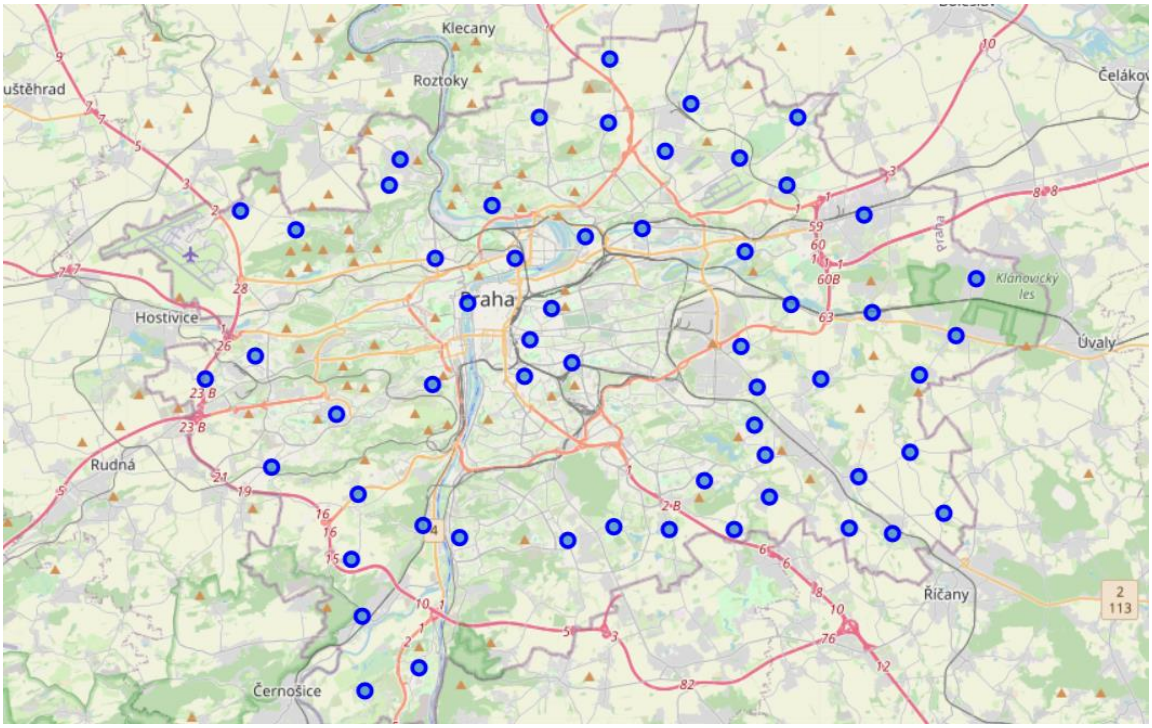
Foursquare API:

For the locations of venues we will use the Foursquare API. Foursquare API provides us with information about venues in the neighborhoods within an area of interest. We will use radius of 500 meters. Foursquare API is the only data source we will be using to obtain these data. To use this API we need to have a Foursquare account.

Data visualization:

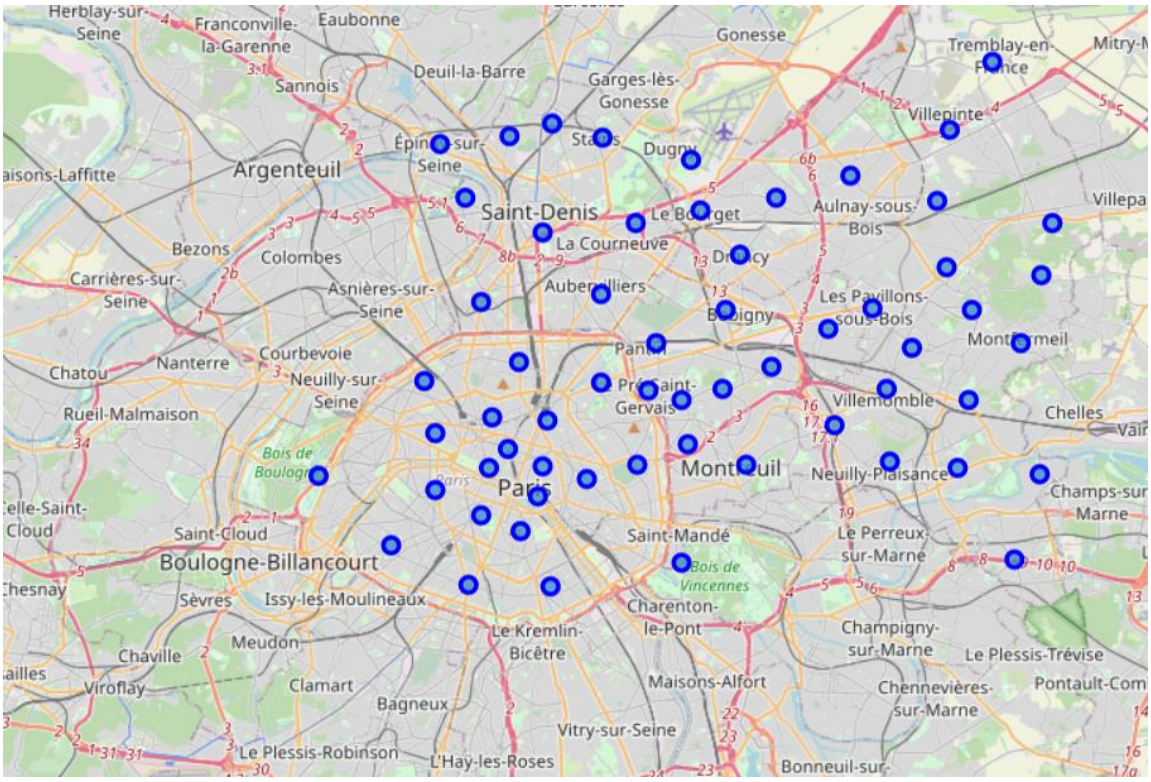
Visualization is also very convenient in our case. we can visualize how distributed the neighborhoods are.

Prague neighborhoods visualization:



Prague Neighborhoods visualization

Paris neighborhood visualization:



Paris Neighborhoods visualization

From the visualization we can see that the distribution of neighborhoods in Paris is much higher.

Then, thanks to the Foursquare API we are able to collect all the venues for the neighborhoods.

Data analysis:

Prague:

There are 1142 venues in Prague neighborhoods. We can use .groupby function to see how many events there are for each neighborhood. The top 3 neighborhoods are Prague 7, Prague 1 and Prague 2, with more than 80 events. It is not surprising at all. Prague 1 and Prague 2 are the neighborhoods in the city center. There are many cafés, hotels, parks, restaurants. Prague 7 is nearby to the city center and this neighborhood is known for many markets, parks, small cafés and restaurants.

Prague Kralovice has only 2 events. It is just a small neighborhood in the suburbs. Several other neighborhoods have only 4 events and it is the same for them, small neighborhoods in the suburbs.

If we use .unique function, it shows us that there are 230 unique events in Prague dataset.

Paris:

There are 1508 events in Paris dataset. There are many events in the center of Paris (7 neighborhoods have more than 80 events). In the Saint Denis part of the city there are not so many events, Montreuil neighborhood has the most events with 19 events.

In Paris there are 237 unique events.

One-hot encoding:

The necessary step to know the frequency of occurrence for each venue in the neighborhoods is one-hot encoding. If the venue is in the neighborhood, the algorithm assigns 1, in other case it assigns 0. Then we apply .mean function to know, how frequent the venue is.

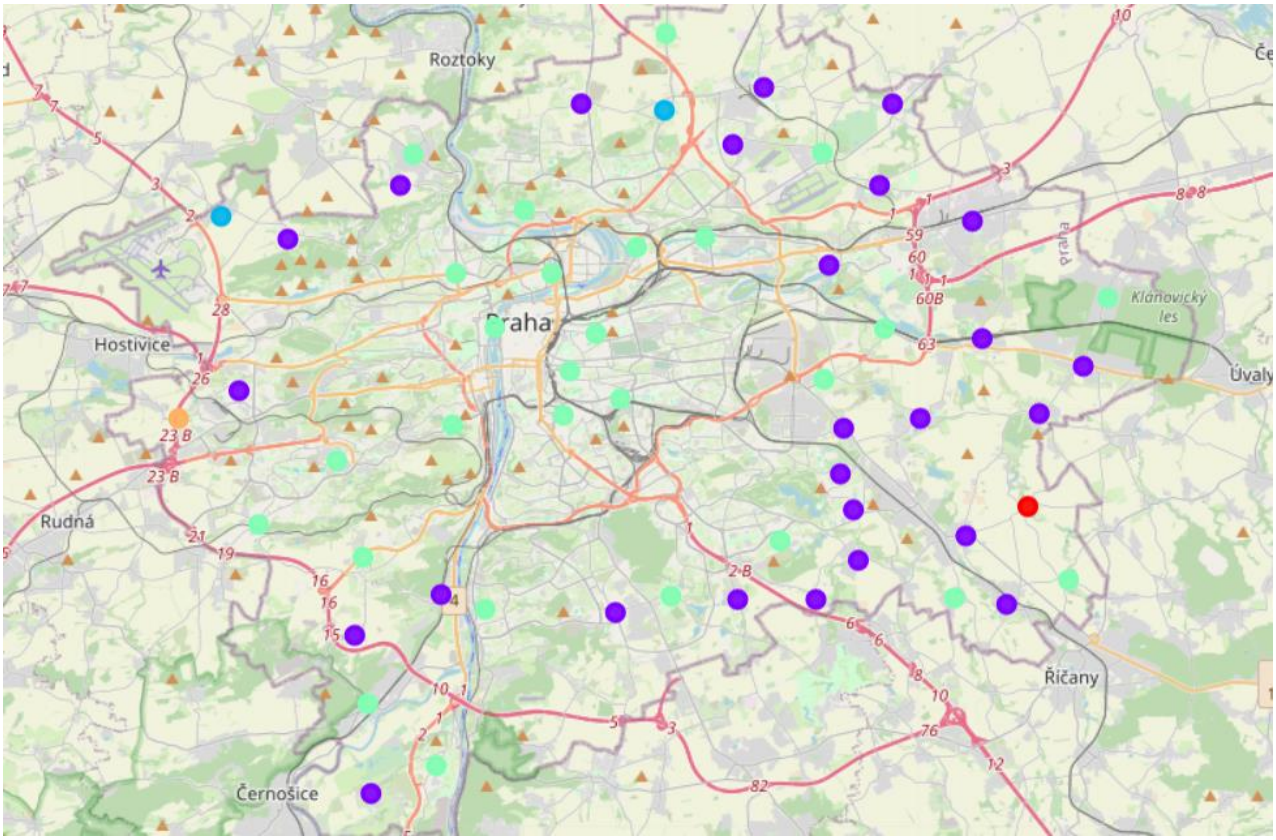
Then we can create the most frequent venues for each neighborhood. It is in a variable called prague_venue_sorted or paris_venues_sorted.

We can see that in both cases there are many cafés, restaurants, pubs in the centers of both of these cities. In the suburbs the most common venues are bus stops, supermarkets and gas station.

Clustering Neighborhoods:

We will use k-means algorithm to cluster the neighborhoods. The number of clusters will be 5. The best way to see the clusters is to visualize them. I will again use folium library for visualization.

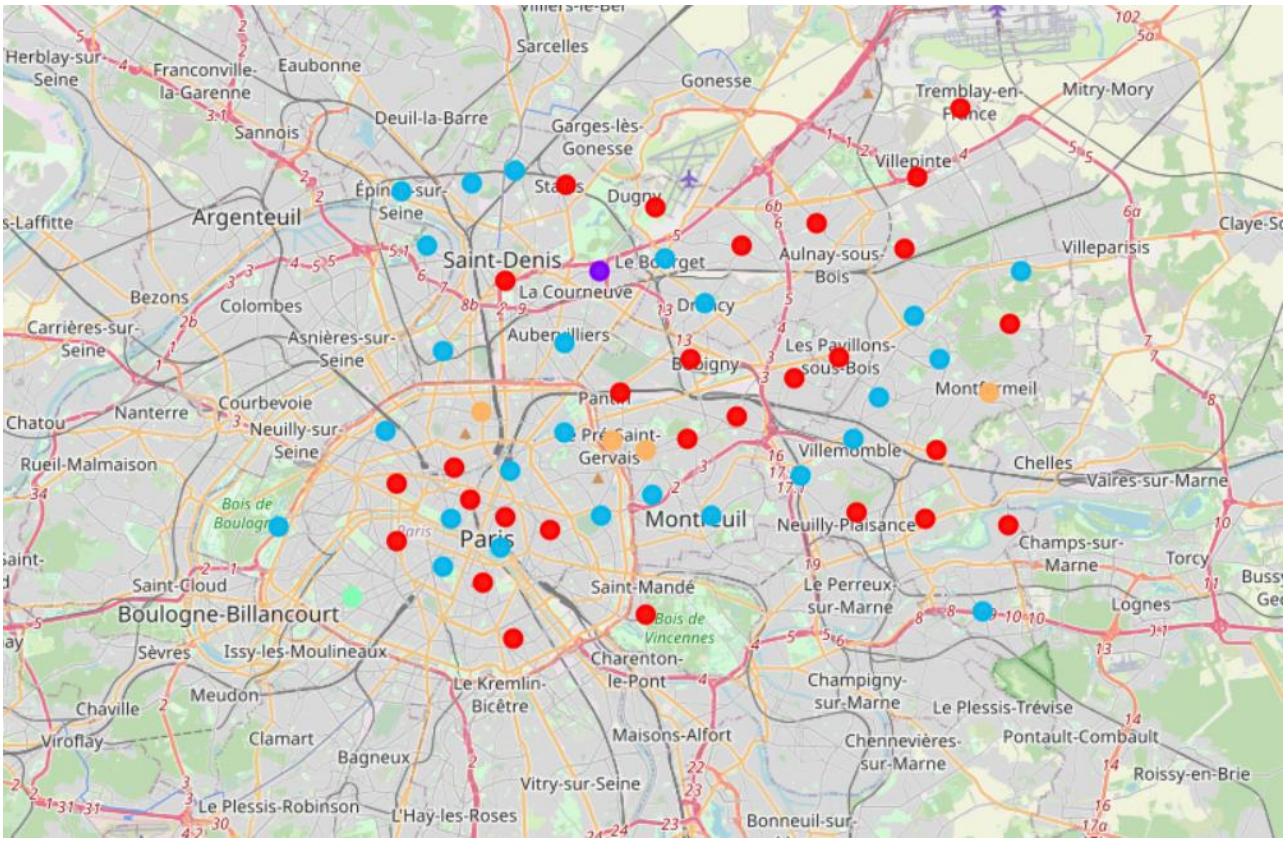
Prague Clusters:



Prague Clusters

We can see that the most frequent clusters are dark blue and green. The green is mainly in the center. Dark blue then surrounds it.

Paris Clusters:



Paris Clusters

The main clusters in Paris are light blue and red. In this case they are mixed.

Exploring the clusters, discussion:

Prague Clusters:

Cluster 0:

The first cluster consists of only one neighborhood- Prague Kralovice. It is located in the suburbs and three the most frequent events there are: Field, Auto Workshop and a Zoo.

Cluster 1:

Cluster 1 is much more interesting. This cluster has 26 neighborhoods. It is the dark blue cluster in the map. The most frequent events there are Bus stops, gardens, restaurants and so on. It is obvious that this is a cluster of the suburbs.

Cluster 2:

Cluster 2 consists of only 2 neighborhoods. Prague Dablice and Prague Predni Kopanina. Both of these clusters have in the top 5 events: Soccer field, Bus stop, Zoo and Restaurant. In the map it is light blue.

Cluster 3:

Cluster 3 is the most interesting one. It is the yellow cluster in the city center. It has 27 neighborhoods. There are restaurants, pubs, cafés, hotels and similar. This cluster is the most interesting one for tourists and investors or business person. The most interesting places are in this cluster.

Cluster 4:

This cluster consist of Prague Zlicin only. This cluster is very specific as well, we can see Playground, Trail and Outdoor and Recreation there. It is the yellow cluster.

Paris Clusters:

Cluster 0:

There are 28 neighborhoods in Cluster 0 (red one). There are many restaurants, banks, hotels, pools and shops. This cluster is mixed in the whole Paris.

Cluster 1:

The dark blue cluster consists of 1 neighborhood only. The most frequent venues are Flea market, Zoo Exhibit and Exhibit. This cluster is very specific.

Cluster 2:

There are 25 neighborhoods in Cluster 2. It is the light blue one. There are many supermarkets, parks, gas stations and shops. I think that this cluster is very convenient for living.

Cluster 3:

There is only one neighborhood in cluster 3. The most common venue is Middle Eastern Restaurant, so I suppose it is a Middle Eastern Neighborhood. Interesting is that this cluster is in the center of Paris (green one).

Cluster 4:

It is the brown cluster. The most common venue there is Supermarket (in all cases). This cluster is probably good for living.

Conclusion:

The aim of the project was to compare two cities, Prague and Paris (center of Paris and Saint Denis). In both cases we made 5 clusters using the same algorithm- kmeans.

In Prague the there were 2 main clusters. One of them was very strictly for the center of the city. There were many pubs, cafés, restaurants, banks and so on. The second main cluster had many bus stops, gardens, restaurants. It is the cluster convenient for living. Then we had 3 clusters, located in the suburbs, which were very specific and had few neighborhoods.

The situation in Paris was much more interesting. We also had 2 main clusters, but these clusters were mixed in the city. The first main cluster had many restaurants, banks, hotels, pools and shops. I would say this is the “business” cluster. The second main cluster had many supermarkets, parks, gas stations and shops. From my point of view this cluster is for living. Then we had 3 very specific clusters. One with many Middle Eastern Restaurants and one with the most supermarkets.