Capstone Project - The Battle of the Neighborhoods (Week2)

Applied Data Science Capstone by IBM/Coursera

Table of Contents

1.Introduction: Business Problem

Paris, capital of France is one of the most important and influential cities in the world. In terms of tourism, Paris is the second most visited city in Europe.

The purpose of this project is to provide potential investors the best decisions possible to open a hotel in Paris, France.

In Paris, like in most of the western big cities of Europe, the coronavirus(Covid-19) has a significant impact on tourism market or real estate & property market, but despite this situation, a large number of firms & their analysts believe that the tourism market could make a recovery by 2022.

Regarding our project, to make an optimal recommendation for the site selection for a hotel, we will explore and answer some questions:

1. How does location influence to open an hotel property?
2. How close it is the potential location to the top attractions?

The secondary criteria for the site selection will be also the availability of restaurants, bars and accessibility to public transport and parking.

Finally, we use our data science knowledge using Machine Learning to generate the best option of district based on these criteria.

In order to solve this business problem, we will cluster the Paris areas based on the distance to the attractions and the amenities, i.e. transport, restaurants, bars. We will then compare each cluster with the cluster's characteristics. This will provide valuable information on whether a location is a viable choice for investors.

Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

2.Datasets

Paris is divided in 20 districts (Arrondissements). Each district has its own postal code, starting at 75001 to 75020 and corresponding to district 1 through 20.

The Postal code dataset was downloaded from the datanova.laposte.fr website (French postal service). The json file named laposte_hexasmal was read and used for this analysis.

A geojson file for the limits of each district was downloaded from https://www.data.gouv.fr/fr/datasets/arrondissements-1/

The Foursquare API will be used to access and explore venues and amenities based on the Latitude and Longitude collected using the json file from datanova.laposte.fr website (laposte_hexasmal.json), which will then be read into a data frame for data wrangling and cleaning.

Based on definition of our problem, factors that will influence our decision are:

- number of existing hotels in the neighborhood (any type of hotel category)
- distance from potential location to the top attractions in Paris

Also, the top attractions are taken from www.tripadvisor.com are:

- Musee d'Orsay
- Sainte-Chapelle
- Palais Garnier — Opera
- Notre Dame Cathedral
- Musee de l'Orangerie
- Luxembourg Gardens
- Louvre
- Eiffel
- Pont Alexandre III
- Le Marais

In order to define our potential locations, we decided to use the center of each districts.

District Candidates

We create the latitude & longitude coordinates for centroids of our candidate. As we already mention before, we decided to use the center of each districts of Paris for our potential locations. We are interested just for postal codes 75001–75020 correspond to district 1 through 20 in Paris.

Let's first find the latitude & longitude of our candidates, using specific json file form the French postal service.

After we wrangle and clean our data, keeping just **t**he features relevant for this project 'code_postal' and 'coordonnees_gps', our data frame is looking like this:

	code_postal	lat	lng
0	75004	48.854228	2.357362
1	75007	48.856083	2.312439
2	75011	48.859415	2.378741
3	75015	48.840155	2.293559
4	75020	48.863187	2.400820
5	75006	48.848968	2.332671
6	75012	48.835156	2.419807
7	75013	48.828718	2.362468
8	75018	48.892735	2.348712
9	75002	48.867903	2.344107
10	75003	48.863054	2.359361
11	75008	48.872527	2.312583
12	75014	48.828993	2.327101
13	75016	48.860399	2.262100
14	75017	48.887337	2.307486
15	75019	48.886869	2.384694
16	75001	48.862630	2.336293
17	75005	48.844509	2.349859
18	75009	48.876896	2.337460
19	75010	48.876029	2.361113

As we see our data is looking good.

Foursquare

We will use Foursquare API to get information on hotels in each district. We're interested in venues of category, 'hotel', ‘Museum’, ‘Metro’, ‘Bus Stop’, ‘Bars’,’ Restaurant’ and ‘Parking’.

We then use Foursquare to explore districts in Paris and our table is now like this:

code_postal	lat	lng	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking
75004	48.854228	2.357362	39.0	3.0	2.0	5.0	157.0	48.0	4.0
75007	48.856083	2.312439	30.0	3.0	2.0	6.0	83.0	10.0	2.0
75011	48.859415	2.378741	19.0	2.0	3.0	3.0	95.0	29.0	1.0
75015	48.840155	2.293559	15.0	4.0	5.0	0.0	105.0	7.0	2.0
75020	48.863187	2.400820	4.0	2.0	4.0	1.0	51.0	12.0	4.0
75006	48.848968	2.332671	27.0	5.0	7.0	4.0	109.0	40.0	3.0
75012	48.835156	2.419807	0.0	0.0	2.0	0.0	5.0	0.0	0.0
75013	48.828718	2.362468	8.0	3.0	8.0	1.0	97.0	5.0	4.0
75018	48.892735	2.348712	14.0	3.0	4.0	1.0	75.0	19.0	1.0
75002	48.867903	2.344107	43.0	4.0	2.0	3.0	199.0	99.0	8.0
75003	48.863054	2.359361	28.0	4.0	3.0	5.0	182.0	54.0	4.0
75008	48.872527	2.312583	41.0	4.0	3.0	4.0	143.0	41.0	7.0
75014	48.828993	2.327101	18.0	2.0	5.0	1.0	67.0	5.0	3.0
75016	48.860399	2.262100	0.0	0.0	1.0	1.0	2.0	2.0	0.0
75017	48.887337	2.307486	15.0	2.0	5.0	2.0	84.0	6.0	2.0
75019	48.886869	2.384694	8.0	2.0	5.0	0.0	48.0	12.0	5.0
75001	48.862630	2.336293	55.0	4.0	11.0	11.0	175.0	38.0	9.0
75005	48.844509	2.349859	32.0	5.0	6.0	3.0	136.0	62.0	1.0
75009	48.876896	2.337460	87.0	6.0	6.0	3.0	167.0	57.0	3.0
75010	48.876029	2.361113	63.0	3.0	6.0	1.0	113.0	25.0	9.0

We can take a look at the statistical indicators

	lat	lng	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking
count	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
mean	48.860290	2.344437	27.300000	3.050000	4.500000	2.750000	104.650000	28.550000	3.600000
std	0.018763	0.037009	22.501696	1.538112	2.438723	2.653201	55.678849	26.047477	2.779625
min	48.828718	2.262100	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000	0.000000
25%	48.847853	2.323471	12.500000	2.000000	2.750000	1.000000	73.000000	6.750000	1.750000
50%	48.861515	2.346410	23.000000	3.000000	4.500000	2.500000	101.000000	22.000000	3.000000
75%	48.873403	2.361452	39.500000	4.000000	6.000000	4.000000	146.500000	42.750000	4.250000
max	48.892735	2.419807	87.000000	6.000000	11.000000	11.000000	199.000000	99.000000	9.000000

We will calculation the distance between 2 locations, more exactly between our districts and the top attractions(www.tripadvisor.com website):

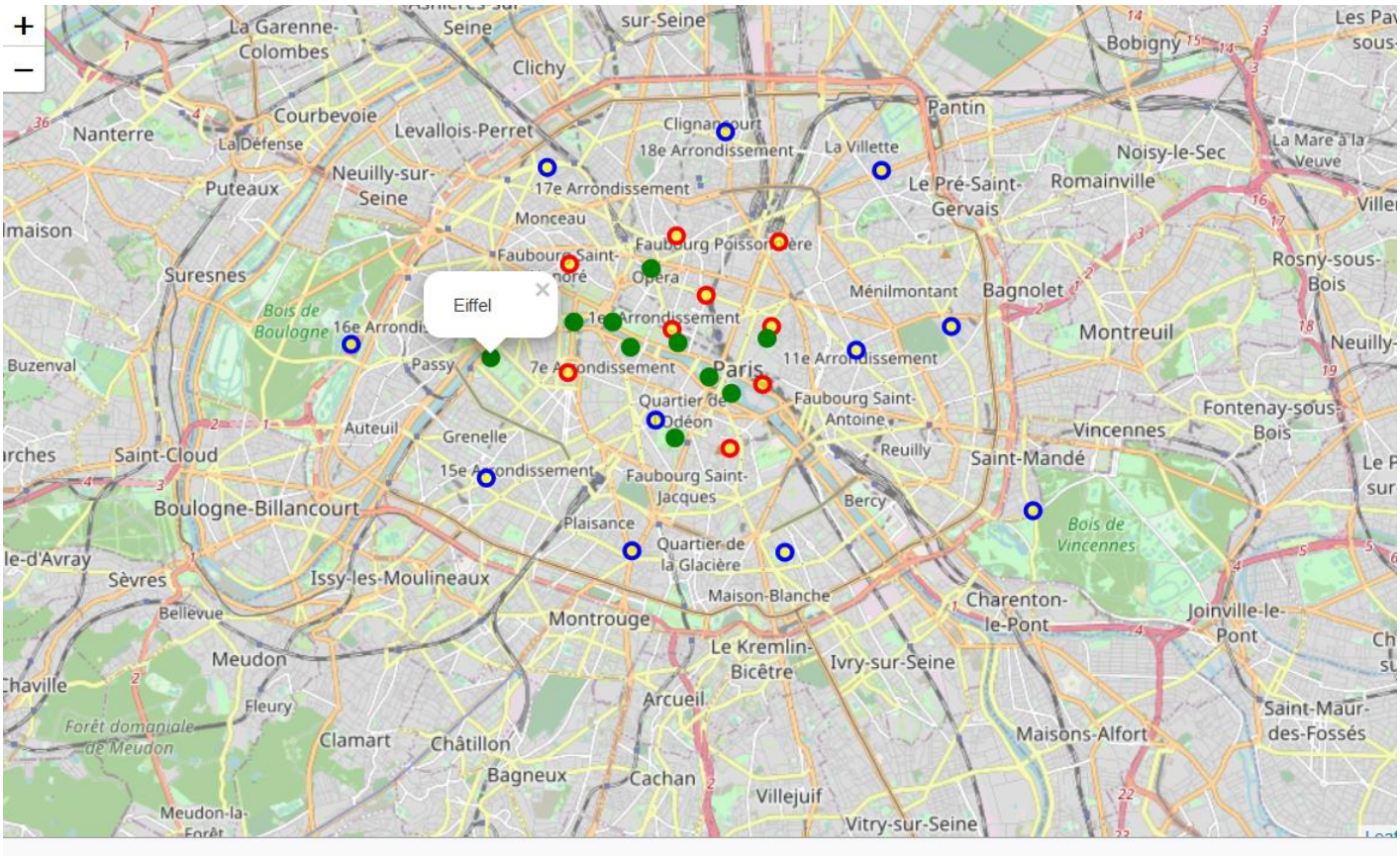
- Musee d'Orsay, 2.326561 , 48.859961
- Sainte-Chapelle, 2.344961 ,48.855375
- Palais Garnier — Opera, 2.331601, 48.871970
- Notre Dame Cathedral, 2.349902, 48.852968
- Musee de l'Orangerie, 2.322672, 48.863788
- Luxembourg Gardens, 2.337160, 48.846222
- Louvre, 2.337644, 48.860611
- Eiffel, 2.294481, 48.858370
- Pont Alexandre III, 2.313559, 48.863900
- Le Marais, 2.358189, 48.861233

Another important feature for the selection of the optimal location of the hotel we also consider the distance to the top 10 attraction in Paris.

We will define function get_distance to calculate the distance between two coordinates, more exactly between our location candidate and the top 10 attractions and we will get a new table with all these distances, including a column for “Total_distance” of that point to all the top ten attractions.

Let's visualize the data we have so far: top ten attraction and our candidate districts.

We created a map centered around Paris and we can play with the zoom level to see how it distributed our possible location link with top attractions and numbers hotels nearby (visible zoom in notebook)



3.Methodology

In this project we will focus to make an optimal recommendation for the site selection of a hotel.

My master data will have the main components: list of districts with chosen features and location of attractions.

In our analysis we will do the calculation and exploration of venues across each districts of Paris.

We will used python folium library to visualize geographic details of Paris.

We will utilize the Foursquare API to explore the districts and segment them.

We will analyze the link between the number of hotels and the distance to attractions (heatmaps, correlation...) to help focus our attention on ideal areas

We will used unsupervised learning K-means algorithm to cluster the districts.

We will present map all potential locations but also create clusters (using k-means clustering) to identify the optimal recommendation for the site.

We will used also, GitHub as our repository.

In the first step in our analysis we have collected the distance from our potential locations to top attractions in Paris, the availability of existing hotels, restaurants, bars and accessibility to public transport and parking. We also cleaned, selected and mapped the features we are interested in.

The second step in our analysis will be exploring our datasets, we will identify the relationships between our features.

In the third step we will focus on creating clusters (using **k-means clustering**) of those locations and their features to identify the optimal districts for our potential location.

Finally, we will identify and optimal propose location to stakeholders

4.Analysis

Now let's calculate two most important things for each location candidate:

- the relationship between number of hotels in vicinity and the distance to the top attractions and total distance to them
- the relationship between the potential location, total distance, numbers restaurants and numbers of bars.

Relation between the principal feature Number_Hotels and Total_distance

In order to measure the relationship between these two variables (Number_Hotels and Total_distance) we will proceed as following:

1. we will calculate the covariance
2. we will plot the graph regplot to shows this relationship
3. we will calculate Pearson's correlation and P-value

1. Covariance

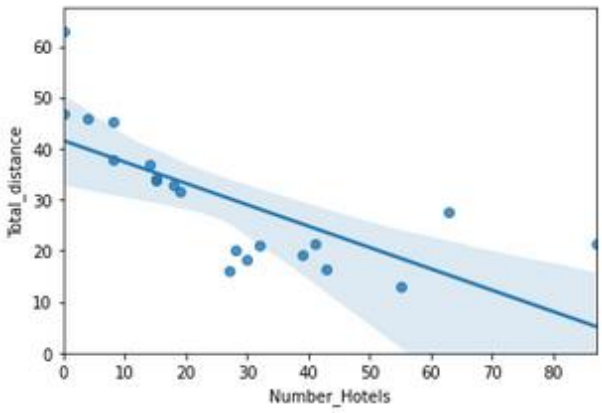
The sign of the covariance can be interpreted as whether the two variables change in the same direction (positive) or change in different directions (negative). A covariance value of zero indicates that both variables are completely independent.

As we can see bellow, the covariance between the two variables is -211.54. The covariance is negative, suggesting the variables change in opposite direction as we expect. If the covariance between the two variables was positive, was suggesting the variables change in the same direction.

```
Covariance between Number_Hotels and Total_distance :

(None,
 array([[ 506.32631579, -211.54089474],
        [-211.54089474,  171.74408921]]))
```

2. Regplot



3. Pearsons correlation and P-value

```
The Pearson Correlation is -0.7173617022958616 with a P_value of P = 0.00037043601871914215
```

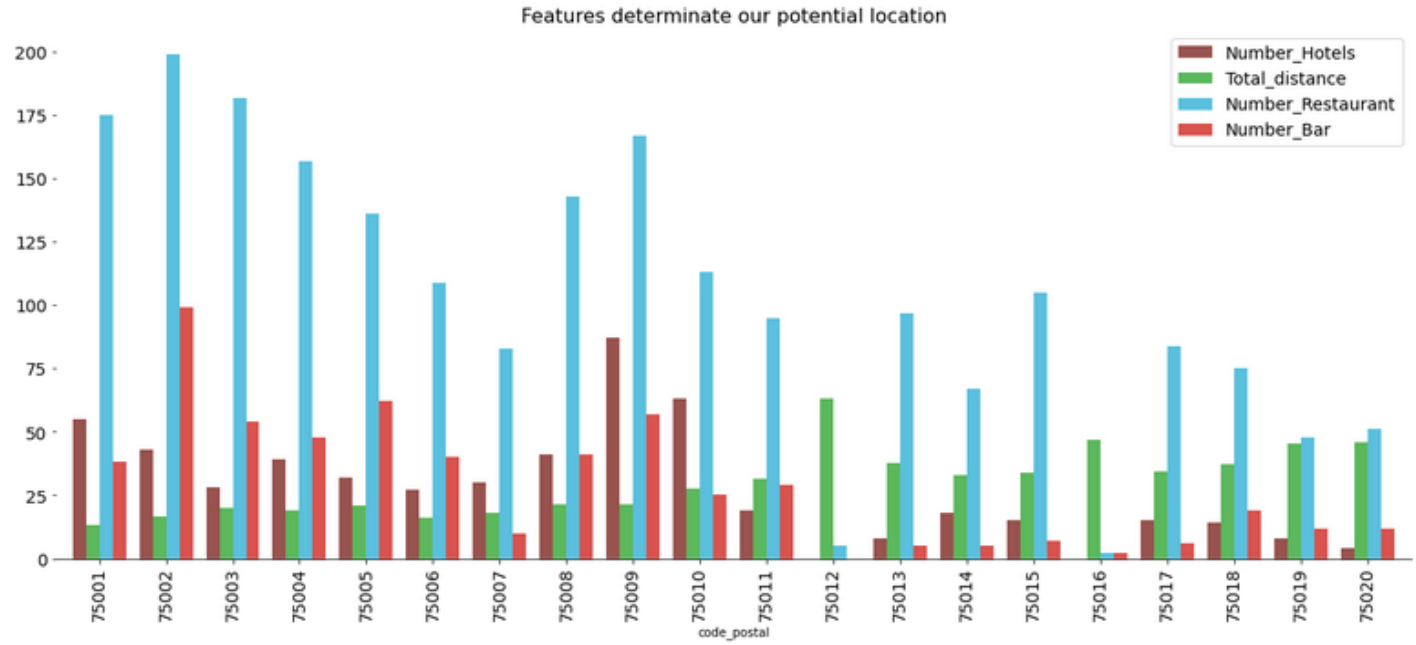
The Pearsons correlation (-0.717) shows a inverse/negative relationship When looking at a regression plot is to pay attention to how scattered the data points are around the regression line , a good indication of the variance of the data.

Correlation coefficient (by default Pearson's correlation coefficient) shows what type of relations is : positive, negative correlation or no correlation between the variables (0), it a measure of the extent of interdependence between variables P-value will tell us how certain we are about the correlation (the relation is supposed be linear relationships) that we calculated between these 2 features.

Conclusion

Since the p-value is <0.001, the correlation between (Number_Hotels and Total_distance) is statistically significant, and the negative linear relationship is quite strong (~-0.717).

Plotting relationship between the total distance, numbers restaurants and numbers of bars and potential location



The statistical indicators of our data frame for selected features that we are interested are

	Number_Hotels	Numbers_Metro	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking	Total_distance
count	20.000	20.000	20.000	20.000	20.000	20.000	20.000	20.00	20.000
mean	27.300	3.050	3.050	4.500	2.750	104.650	28.550	3.60	30.134
std	22.502	1.538	1.538	2.439	2.653	55.679	26.047	2.78	13.105
min	0.000	0.000	0.000	1.000	0.000	2.000	0.000	0.00	13.060
25%	12.500	2.000	2.000	2.750	1.000	73.000	6.750	1.75	19.822
50%	23.000	3.000	3.000	4.500	2.500	101.000	22.000	3.00	29.450
75%	39.500	4.000	4.000	6.000	4.000	146.500	42.750	4.25	37.242
max	87.000	6.000	6.000	11.000	11.000	199.000	99.000	9.00	62.960

```
Average distance to closest Hotel from each top: 30.134499999999996
```

We will order the best potential location according with the min distance to the best top attractions in Paris, knowing that the Average distance to closest Hotel from each top: 30.13

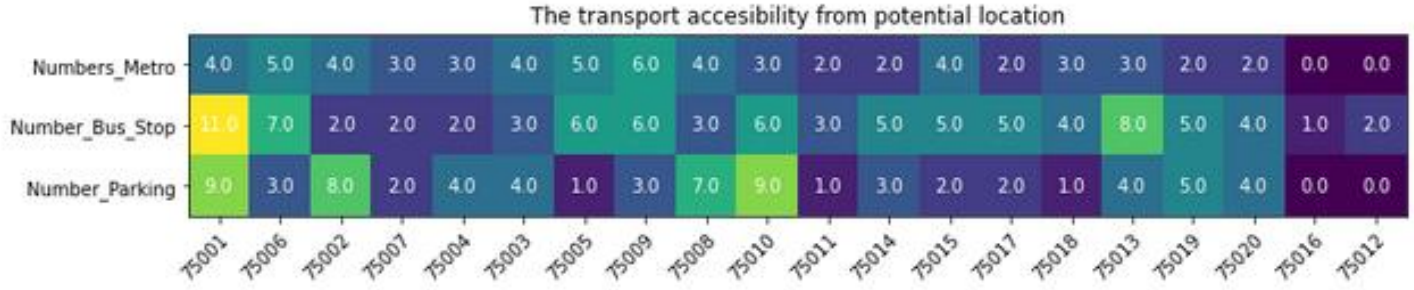
code_postal	Total_distance
75001	13.06
75006	16.11
75002	16.48
75007	18.18
75004	19.08
75003	20.07
75005	21.13
75009	21.35
75008	21.42
75010	27.43
75011	31.47
75014	32.99
75015	33.91
75017	34.19
75018	37.04
75013	37.85
75019	45.31
75020	45.94
75016	46.72
75012	62.96

Relationship between transportation and possible location

The heatmap bellow is give us the accessibility from our potential locations.

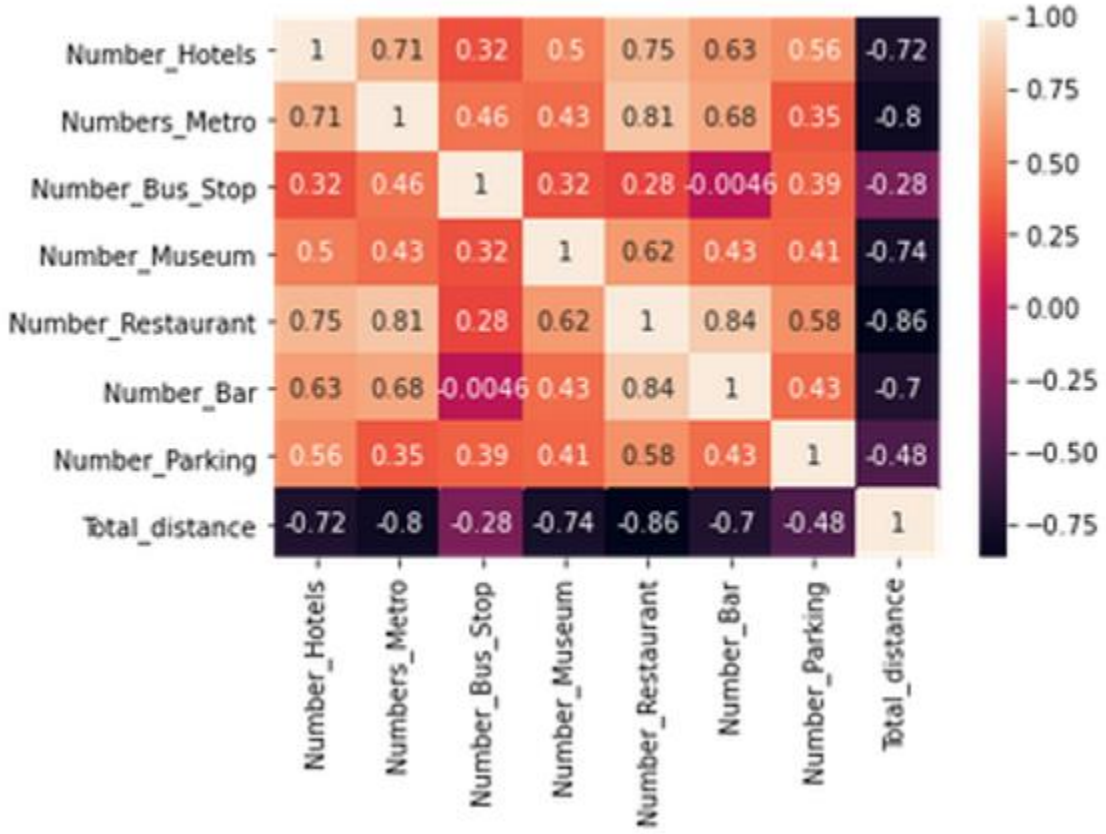
The heatmap plots the target variable postal codes of potential locations' proportional to colour with respect to the variables 'Numbers_Metro','Number_Bus_Stop','Number_Parking" in the horizontal and vertical axis respectively.

This allows us to visualize how the postal code of potential locations is related to the transport connections as Numbers_Metro , Number_Bus_Stop , Number_Parking.We can see that the code 75001 is seems to be a good option, but also 75010 or 75013.



Relationship between multiple variables at the same time

In order to find this, we can use correlation matrix build with the Pearson's correlation coefficient show us both the strength of the relationship and its direction (positive or negative correlations).We will use seaborn's heatmap() method to plot the matrix.



As we can know already and we can see in this heatmap, the correlation between number hotels and total distance is negative (- 0,72) and an other negative correlation is between number

parking and total distance (-0,48) or number of bus stop and total distance suggesting the variables change in opposite direction as we expect.

For the rest of variables, the covariance between the two variables were positive, suggesting the variables change in the same direction.

Modeling

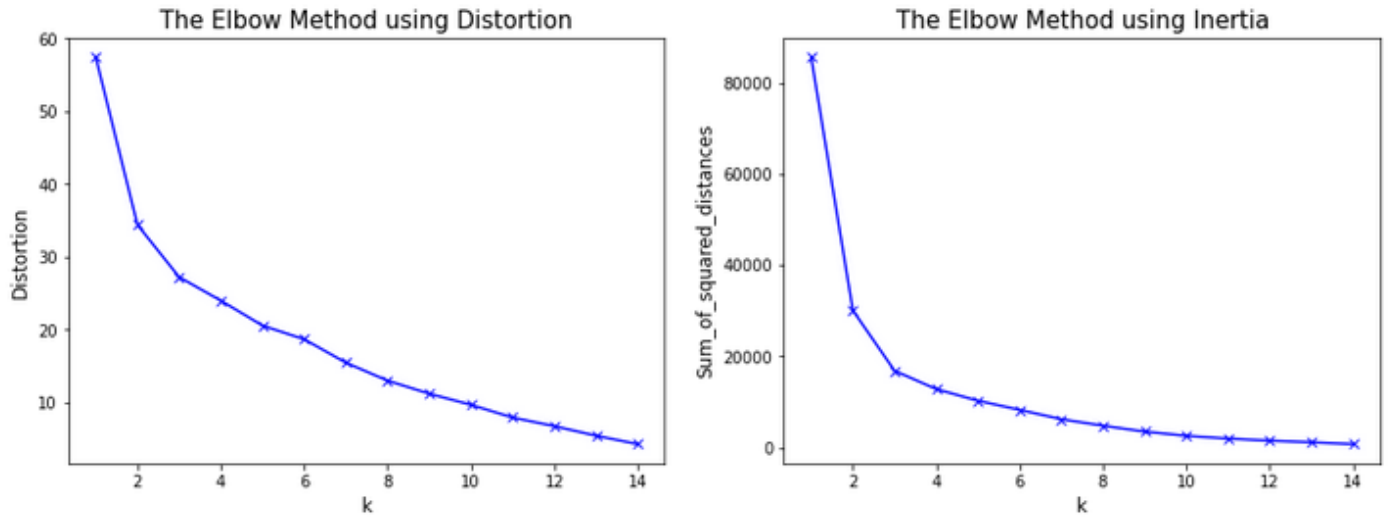
Optimising K - Elbow Method

The Elbow Method is used to determine the optimal value of k as this is one of the most popular methods.

We will be using 2 metric values calculated from a range of k values in order to determine the 'elbow point', i.e. the point after which the metrics starts decreasing linearly.

Those 2 metric values are:

-Distortion: Calculated as the average of the squared distances from the cluster centres of the respective clusters where typically the Euclidean distance is used.
-Inertia: The sum of squared distances of samples to their closest cluster centre.



So, we deduce that **k=3 is the optimal K**

Clustering using KMeans

Clustering or cluster analysis is the process of dividing data into groups (clusters) in such a way that objects in the same cluster are more similar to each other than those in other clusters.

Grab the labels for each point in the model using Kmean

```
array([1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2])
```

Our new data frame sorted by Cluster Labels and with selected columns it is looking like this:

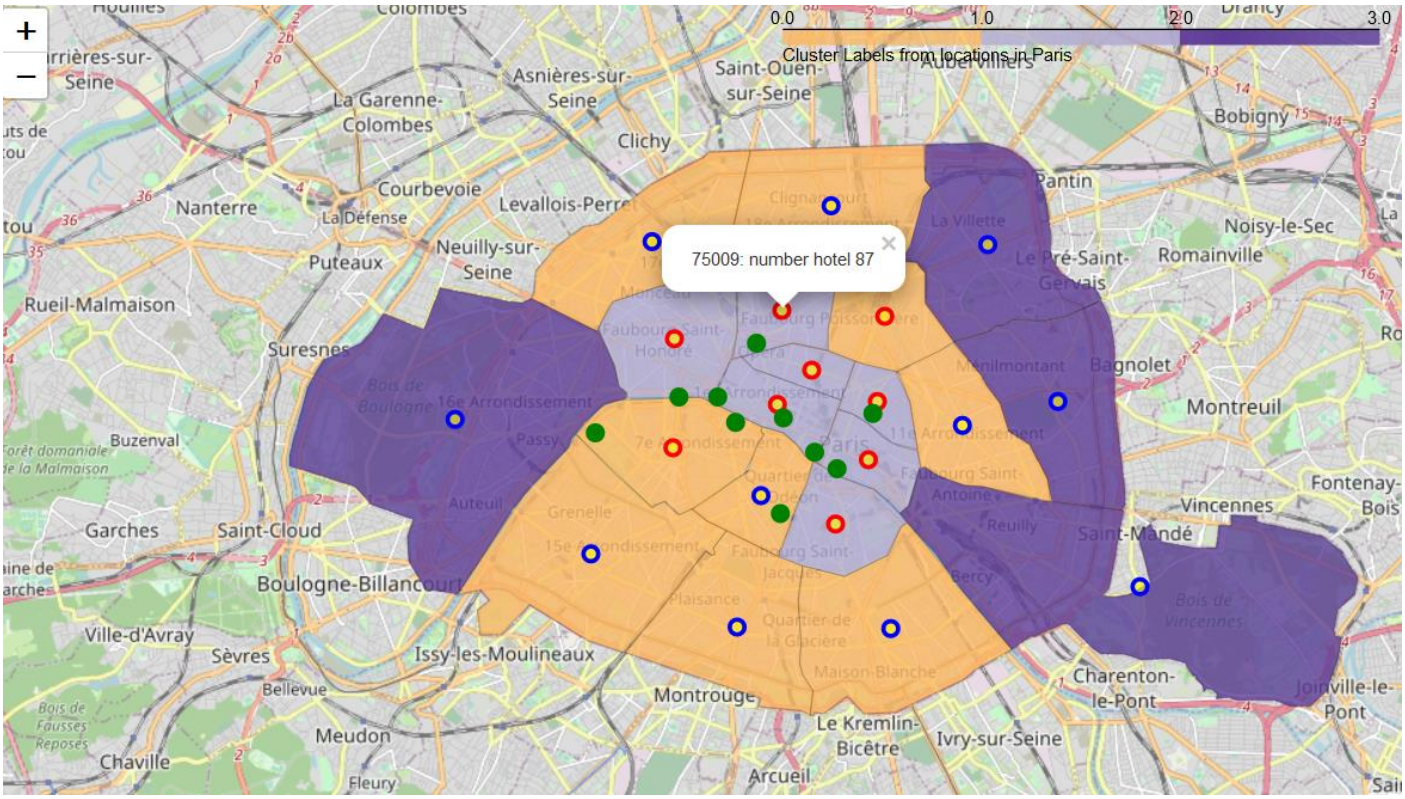
Cluster Labels	code_postal	Number_Hotels	Total_distance
0	75010	63.0	27.43
0	75006	27.0	16.11
0	75013	8.0	37.85
0	75007	30.0	18.18
0	75018	14.0	37.04
0	75017	15.0	34.19
0	75015	15.0	33.91
0	75014	18.0	32.99
0	75011	19.0	31.47
1	75001	55.0	13.06
1	75008	41.0	21.42
1	75009	87.0	21.35
1	75005	32.0	21.13
1	75003	28.0	20.07
1	75004	39.0	19.08
1	75002	43.0	16.48
2	75019	8.0	45.31
2	75020	4.0	45.94
2	75016	0.0	46.72
2	75012	0.0	62.96

Creating a Choropleth map

We will use the choropleth method with the following main parameters:

- geo_data, which is the GeoJSON file from https://www.data.gouv.fr/fr/datasets/arrondissements-1/
- data, which is the data frame containing the data.
- columns, which represents the columns in the data frame that will be used to create the Choropleth map.
- key_on, which is the key or variable in the GeoJSON file that contains the name of the variable of interest.

We load the geoJson file for Paris districts and we visualise the clusters using an interactive Plotly map



This map doesn't indicate the possible location, but it indicates that there is a higher density of top attractions in cluster 1 but they also have more the average of number of hotel (bullet red outer color). Moreover, Cluster 2 is furthest from the attractions, but they also have the least number of hotels.

Based on this we will now focus our analysis on areas of Cluster 1 (close to attractions) and Cluster 2 (least number of hotels).

Examine Clusters

The principal features that will influence our selection of Cluster are as the total distance to the top of attractions and Number Hotels around our potential location to be minimum.

To each of our cluster data frame, we will add a row for the cluster's mean values of each features.

Cluster 0 (average distance and number of hotels) - excluded

	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking	Total_distance
5	27.00	5.00	7.00	4.00	109.00	40.00	3.00	16.11
1	30.00	3.00	2.00	6.00	83.00	10.00	2.00	18.18
19	63.00	3.00	6.00	1.00	113.00	25.00	9.00	27.43
2	19.00	2.00	3.00	3.00	95.00	29.00	1.00	31.47
12	18.00	2.00	5.00	1.00	67.00	5.00	3.00	32.99
3	15.00	4.00	5.00	0.00	105.00	7.00	2.00	33.91
14	15.00	2.00	5.00	2.00	84.00	6.00	2.00	34.19
8	14.00	3.00	4.00	1.00	75.00	19.00	1.00	37.04
7	8.00	3.00	8.00	1.00	97.00	5.00	4.00	37.85
mean	23.22	3.00	5.00	2.11	92.00	16.22	3.00	29.91

Cluster 1 (close to attractions)

	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking	Total_distance
16	55.00	4.00	11.00	11.00	175.00	38.00	9.00	13.06
9	43.00	4.00	2.00	3.00	199.00	99.00	8.00	16.48
0	39.00	3.00	2.00	5.00	157.00	48.00	4.00	19.08
10	28.00	4.00	3.00	5.00	182.00	54.00	4.00	20.07
17	32.00	5.00	6.00	3.00	136.00	62.00	1.00	21.13
18	87.00	6.00	6.00	3.00	167.00	57.00	3.00	21.35
11	41.00	4.00	3.00	4.00	143.00	41.00	7.00	21.42
mean	46.43	4.29	4.71	4.86	165.57	57.00	5.14	18.94

Cluster 2 (least number of hotels)

	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking	Total_distance
15	8.00	2.00	5.00	0.00	48.00	12.00	5.00	45.31
4	4.00	2.00	4.00	1.00	51.00	12.00	4.00	45.94
13	0.00	0.00	1.00	1.00	2.00	2.00	0.00	46.72
6	0.00	0.00	2.00	0.00	5.00	0.00	0.00	62.96
mean	3.00	1.00	3.00	0.50	26.50	6.50	2.25	50.23

We will take a look at the mean of the number of hotels and total distance of our clusters

Cluster Labels	Number_Hotels	Total_distance
0.0	23.222	29.908
1.0	46.429	18.941
2.0	3.000	50.232

Because of the covariance between the primary features are negatively correlated, we consider the minimum of these two features are decisional for our proposal, so we can assume the cluster 1 and 2 are the best location

Cluster Labels	Number_Hotels	Total_distance
1.0	46.428571	18.941429

Cluster Labels	Number_Hotels	Total_distance
2.0	3.0	50.2325

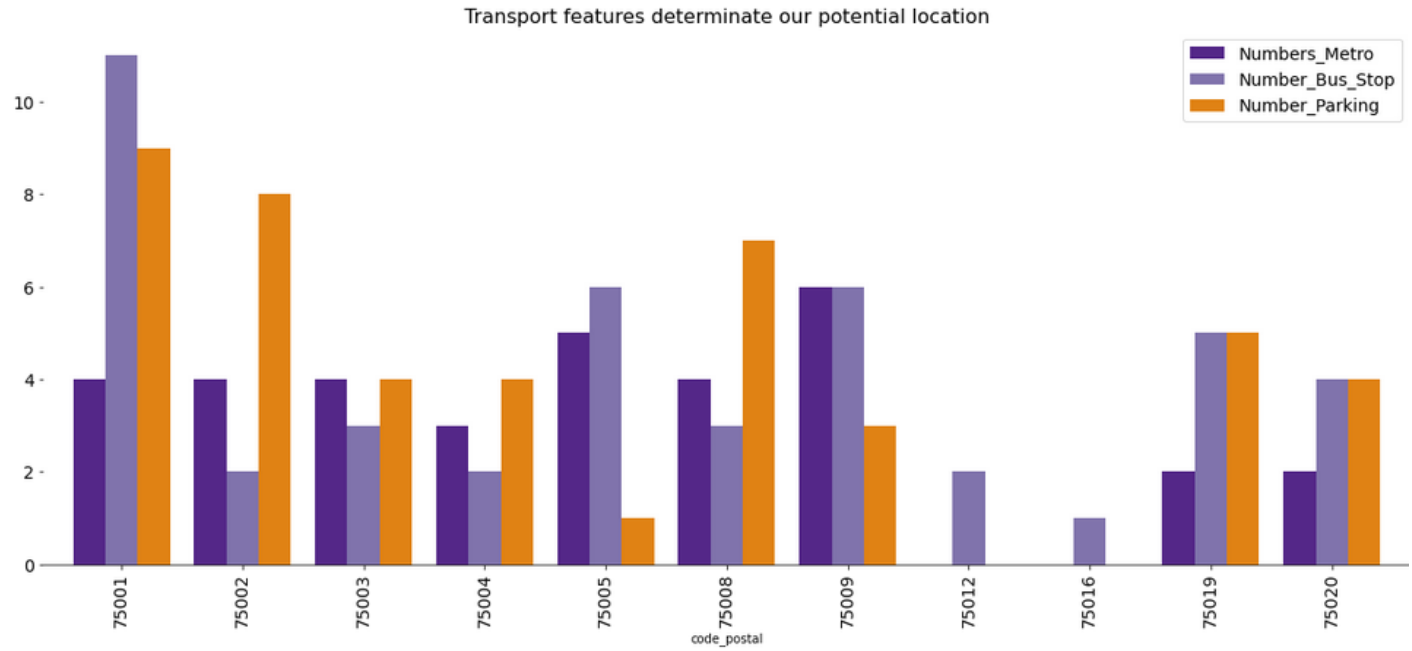
Finnaly, our new dataframe with the selected columns with label 1 and 2 is:

code_postal	Number_Hotels	Numbers_Metro	Number_Bus_Stop	Number_Museum	Number_Restaurant	Number_Bar	Number_Parking	Total_distance
75001	55.0	4.0	11.0	11.0	175.0	38.0	9.0	13.06
75002	43.0	4.0	2.0	3.0	199.0	99.0	8.0	16.48
75004	39.0	3.0	2.0	5.0	157.0	48.0	4.0	19.08
75003	28.0	4.0	3.0	5.0	182.0	54.0	4.0	20.07
75005	32.0	5.0	6.0	3.0	136.0	62.0	1.0	21.13
75009	87.0	6.0	6.0	3.0	167.0	57.0	3.0	21.35
75008	41.0	4.0	3.0	4.0	143.0	41.0	7.0	21.42
75019	8.0	2.0	5.0	0.0	48.0	12.0	5.0	45.31
75020	4.0	2.0	4.0	1.0	51.0	12.0	4.0	45.94
75016	0.0	0.0	1.0	1.0	2.0	2.0	0.0	46.72
75012	0.0	0.0	2.0	0.0	5.0	0.0	0.0	62.96

We split our features that influence our decision in two group: one consider principal group that are related with Transport (metro, bus, parking) and secondary group related with the Foods and Drinks features (Restaurant/bars).

Let’s create a data frame of transport related features in the selected group of potential location and we build a Histogram of the transport features that help determinate our potential location.

From the graph below we can see that majority of the transport features are below 10 per each postal code but only postal code 75001 is with all the features higher than others.



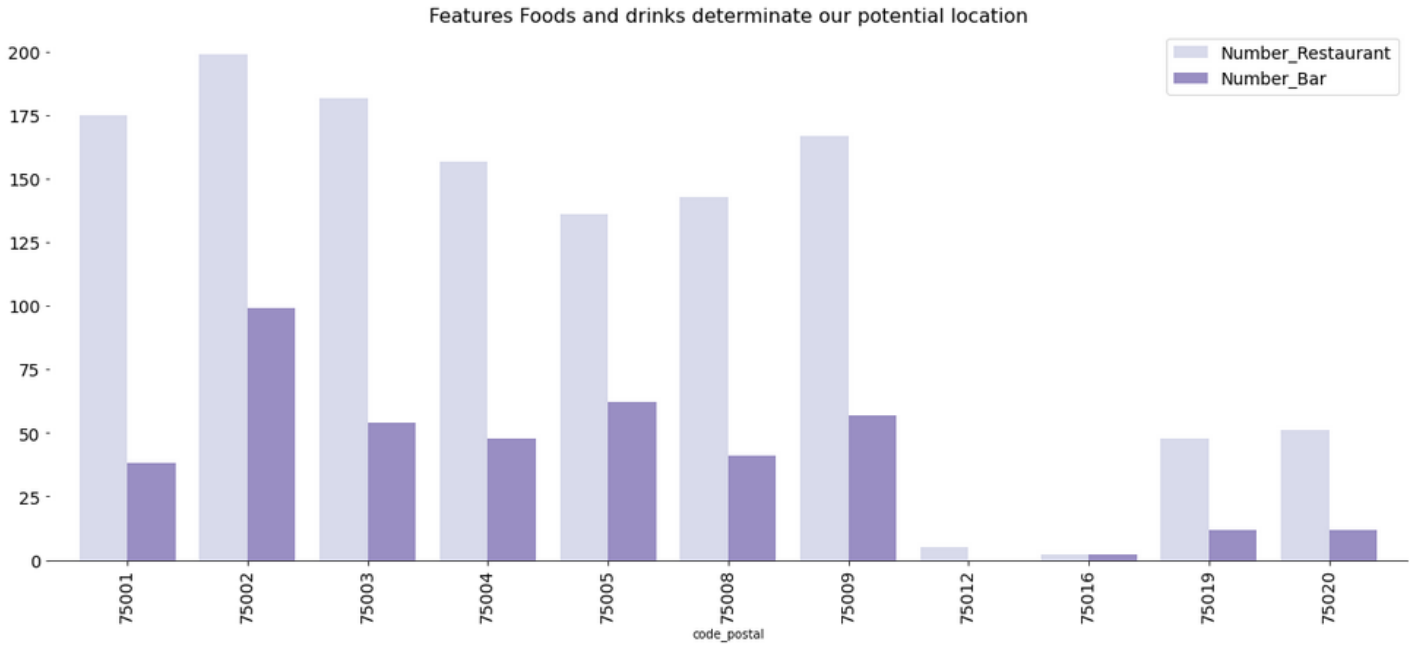

```
Mean of Numbers_Metro is : 3.090909090909091
Mean of Number_Bus_Stop is: 4.090909090909091
Mean of Number_Parking is: 4.090909090909091
```

code_postal	Numbers_Metro	Number_Bus_Stop	Number_Parking	Numbers_Metro_over_mean	Number_Bus_Stop_over_mean	Number_Parking_over_mean
75001	4.0	11.0	9.0	True	True	True
75002	4.0	2.0	8.0	True	False	True
75004	3.0	2.0	4.0	True	False	True
75003	4.0	3.0	4.0	True	False	True
75005	5.0	6.0	1.0	True	True	False
75009	6.0	6.0	3.0	True	True	False
75008	4.0	3.0	7.0	True	False	True
75019	2.0	5.0	5.0	False	True	True
75020	2.0	4.0	4.0	False	True	True
75016	0.0	1.0	0.0	False	False	False
75012	0.0	2.0	0.0	False	False	False

The table above shows that postal code 75001 to have all the features higher than each mean of transport's features.

Let's create a data frame of Foods and Drinks features (Restaurant/bars) in the selected group of potential location.

Bellow we can see that majority of Foods and Drinks features are below 200 per each postal code but postal code 75001 has one of the highest values.




```
Mean of Number_Restaurant is : 115.0
Mean of Number_Bar is: 38.63636363636363
```

code_postal	Number_Restaurant	Number_Bar	number_Restaurant_over_mean	Number_Bar_over_mean
75001	175.0	38.0	True	True
75002	199.0	99.0	True	True
75004	157.0	48.0	True	True
75003	182.0	54.0	True	True
75005	136.0	62.0	True	True
75009	167.0	57.0	True	True
75008	143.0	41.0	True	True
75019	48.0	12.0	False	False
75020	51.0	12.0	False	False
75016	2.0	2.0	False	False
75012	5.0	0.0	False	False

The table above shows that code postal 75001 continue to be a better option location having the restaurants over mean of category.

5.Results and Discussion

Our analysis shows that beside the fact there is a great number of hotels in Paris, there are places for hotels close to city center.

We targeted the tourists that want to explore the attractions and enjoy French cuisine and night life.

The highest concentration of hotels was detected in the first 10 district (at the center of Paris), but also the total distance to the principal top attractions in these districts is lower the mean of 30 km, also in plus is a low of density of restaurants, bars and very good connexions in Paris.

After analysing the relations between the features, we found that our potential districts (postal code) have a positive correlations between the transport connections, number hotels and negative correlations between total distance directing our attention to this more narrow area of interest, taking as primarily feature total distance, so close to attraction and second number of hotels.

These location candidates were then clustered to create zones of interest which contain fewer number of locations. I used unsupervised learning K-means algorithm to cluster the districts. K-Means algorithm is one of the most common clustering methods of unsupervised learning.

First, I will run K-Means to cluster the districts into 3 clusters because analyzing the K-Means with Elbow method had ensured us the 3 degree is the optimum k of the K-Means.

We visualise the clusters using an interactive Plotly map that doesn't indicate the possible location, but it indicates that there is a higher density of top attractions in cluster 1, also they have more the average of number of hotel (bullet red outer color). Moreover, Cluster 2 is furthest from the attractions, but they also have the least number of hotels.

The principal features that will influence our selection of Cluster are the total distance to the top attractions and the Number of Hotels around our potential location to be minimum.

Based on this we will now focus our analysis on areas of Cluster 1 (close to attractions) and Cluster 2 (least number of hotels).

Examining our clusters also, we found the cluster 1 and 2 are the best location

To select our possible location, we will analyse the others features. We split those features in two group (Transport and Foods & Drinks). After plot the histograms we found that 75001 is the better optional for transport and second for restaurants.

Purpose of this analysis was to make a recommendation where will be good for opening an hotel having as target the tourists.

6.Conclusion

Purpose of this project was to identify Paris areas close to top attractions with a low number of hotels which would be an optimal location for a new hotel.

Using Foursquare data, we visualise the clusters using an interactive Plotly map and we identify general districts that justify further analysis of cluster 1.

I use machine learning and use Kmean to help us to choose the better location. We selected cluster 1 and 2 as the best locations because of the negative correlation between number of hotels and distance to attractions we keep the minimum of these two features.

We explore cluster data frame 1 and 2 in order to create a major zone of interest and visually inspect the selected Clusters, plotting data.

Final decision on optimal restaurant location will be made by investors based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to attraction), proximity to metro, parking, bars, restaurants, etc.