Capstone Project - The Battle of Neighborhoods

Searching for the most environmental district in Paris

Introduction

This final project is designated to help people moving to Paris, France discover which Borough is more Eco-friendly, so they can weigh that consideration in their decision where to rent an apartment. The main goal of this project is to Determine which neighborhood has "Greener" criteria, based on land uses data, venues data and air quality data within each borough.

Data

In order to answer the question above, the twenty arrondissements (boroughs) of Paris data including latitude, longitude, Area, Gardens, Parks, Vegetarian/Vegan restaurants, Organic groceries, Trees, Open spaces and Air Quality index in each borough are necessary. The data sources are as follows:

1. Paris data containing the boroughs, **latitudes**, and **longitudes** and **total area** of each borough will be obtained from the data source: https://opendata.paris.fr/
2. Paris data containing historical **air quality index** during the past few days will be taken from the Breezometer API : https://breezometer.com/historical-air-quality-data using the Request library in Python.
3. Data related to locations of **Public open spaces and Eco-friendly businesses** will be obtained via the Foursquare API by the Request library in Python.
4. Data related to **green spaces** in every borough and the **amount of trees** in every borough will also be taken from the data source : https://opendata.paris.fr/
5. Additional data of Paris **Vegan restaurants** will be taken from the google API : https://developers.google.com/places/

Methodology

In this section I will explore the datasets to characterize the different 20 main districts in Paris. The question whether a borough is eco-friendly or not would be answered by the following steps:

1. Collecting data found in different API's and sites
2. Cleansing the data when necessary
3. Characterizing and visualizing the data

4. K-means clustering method implementation, in aim to group the similar data points together and discover underlying patterns.

Exploratory Data Analysis

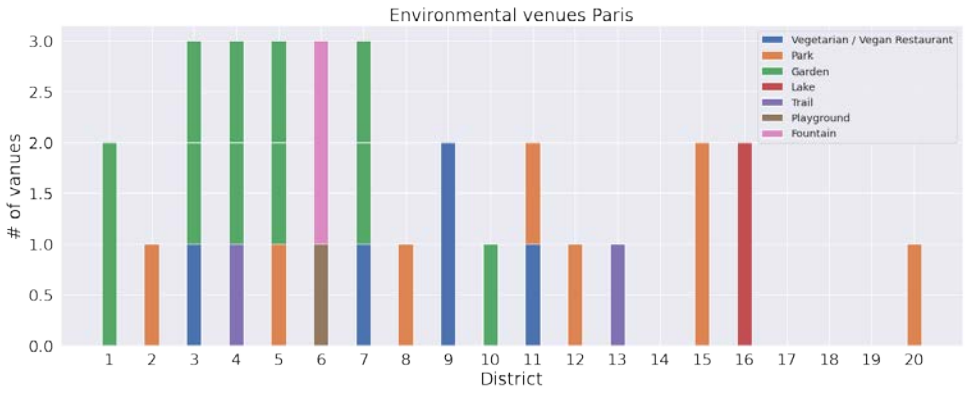
1. latitude-longitude of neighborhoods

Coordinates of each neighborhood were obtained by web scraping from Wikipedia:

	Neighborhood	longitude	latitude
0	2ème Ardt	2.342803	48.868279
1	17ème Ardt	2.306777	48.887327
2	4ème Ardt	2.357630	48.854341
3	8ème Ardt	2.312554	48.872721
4	18ème Ardt	2.348161	48.892569
5	1er Ardt	2.336443	48.862563
6	10ème Ardt	2.360728	48.876130
7	16ème Ardt	2.261971	48.860392
8	3ème Ardt	2.360001	48.862872
9	9ème Ardt	2.337458	48.877164
10	19ème Ardt	2.384821	48.887076
11	11ème Ardt	2.380058	48.859059
12	13ème Ardt	2.362272	48.828388
13	7ème Ardt	2.312188	48.856174
14	14ème Ardt	2.326542	48.829245
15	20ème Ardt	2.401188	48.863461
16	15ème Ardt	2.292826	48.840085
17	5ème Ardt	2.350715	48.844443
18	6ème Ardt	2.332898	48.849130
19	12ème Ardt	2.421325	48.834974

2. Explore and cluster the neighborhoods in Paris - FourSquare data:

Afterwards, venues within 500 meter radius of each neighborhood centroid were collected from the Foursquare API Using the getNearbyVenues function. Venues without environmental significance were dropped, and the remaining venues were summed for each neighborhood. The results are as follows:

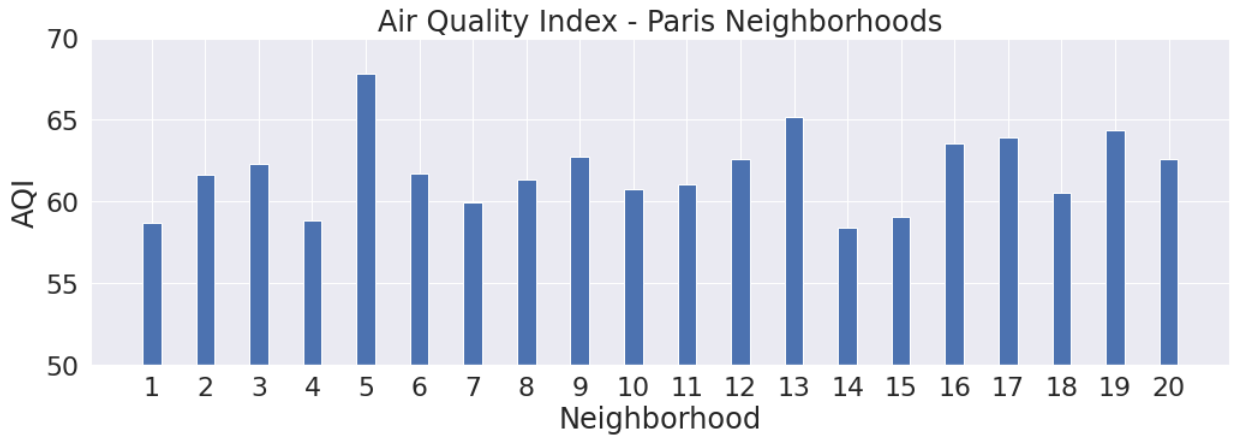


The Foursquare data seems sparse, therefore I searched for more comprehensive data. I found in the Paris City database similar data, with more interesting features such as area of green places which might be more helpful in determining which neighborhood is greener. A database with the number of trees in each neighborhood was also retrieved.

3. Air Quality data

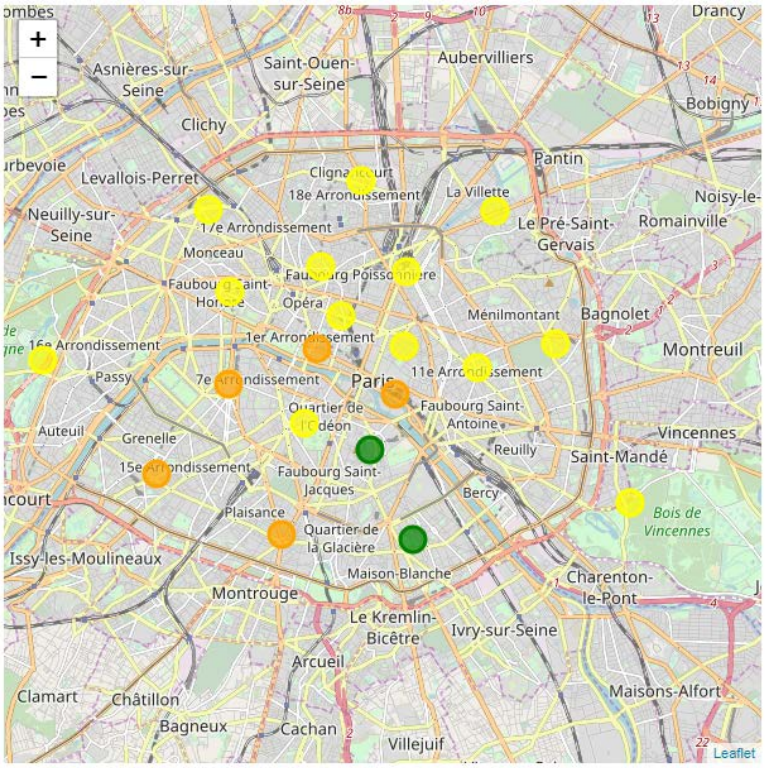
Using the getAQI function, air quality index was retrieved from the Breezometer API. because the number of requests from the API is limited, only 72 hours historical data was retrieved and then averaged for each borough. The AQI index values represent the Air Quality Index, when higher value means better air quality (max value is 100). Ideally, the API would take AQI from numerous points within the 500 meter radius of each neighborhood centroid, but it is not possible considering the free user account of the API.

The AQI between 100-60 represents good air quality, and when below 60 the air quality index is medium. In the bar plot and also in the map below, it is seen that some of the neighborhood have good air quality, while some neighborhoods air quality is medium (colored in orange in the map).



Air Quality map - Paris (21-24/3/2020)

(Green - AQI > 65 ; Yellow - 06<AQI<65 ; Orange - AQI < 60)



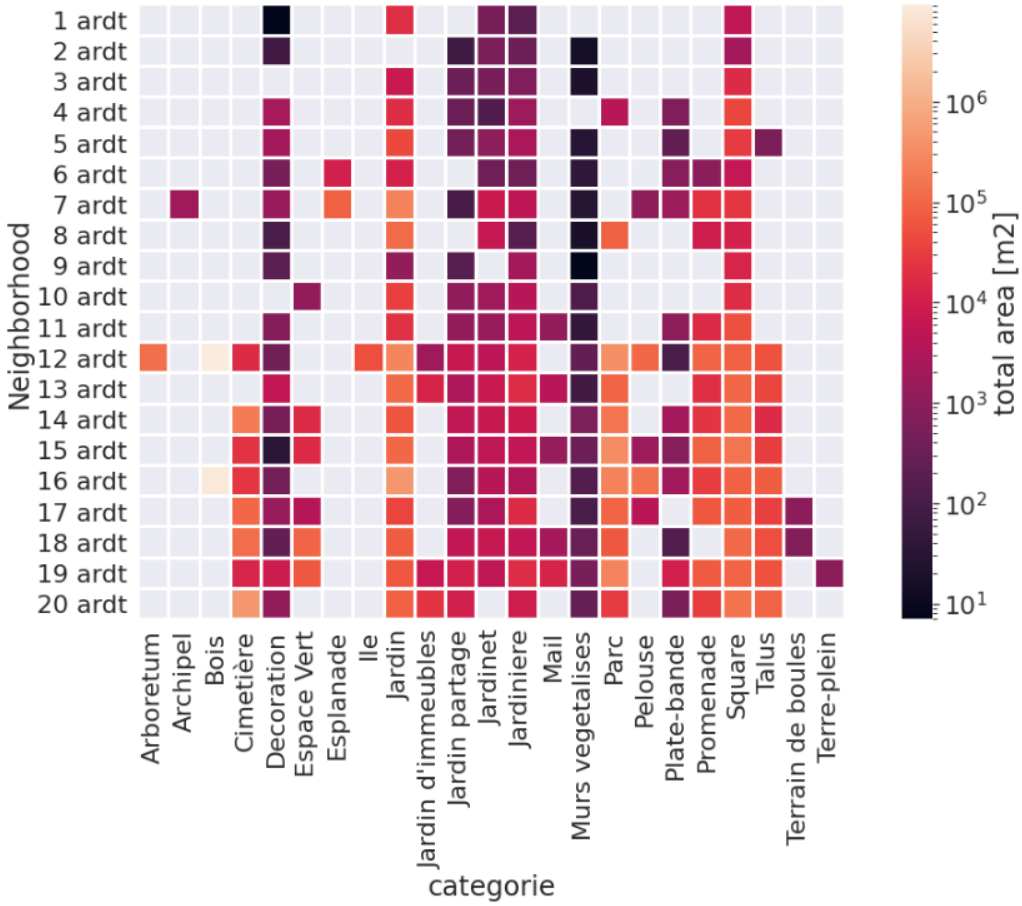
It can be seen that the 5th and 13th districts exhibit the best air quality (above 65), while the 1,4,7,14,15 districts exhibit the worse relative AQI (below 60).

4. Environmental Data from ParisData.com

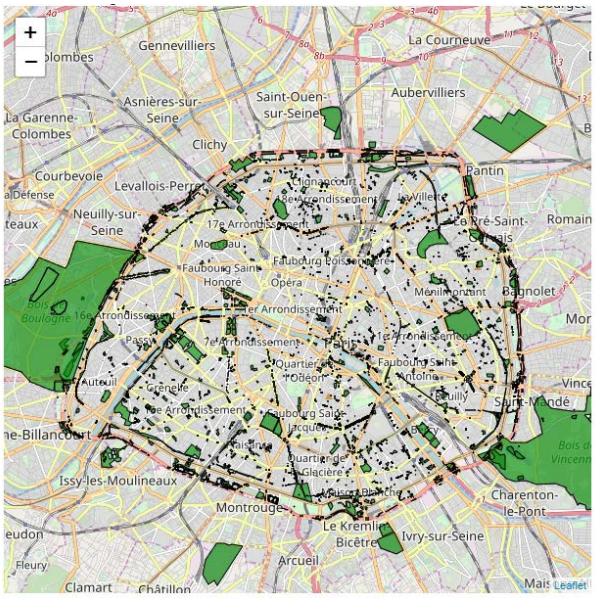
The number of trees for each Neighborhood was summed and then divided by the area of each neighborhood, to produce the density of trees per km^2. Smaller trees density was observed in the neighborhoods located at the center of Paris (Districts 1-10).

	Neighborhood	NBHD_ID	number of trees	area_km2	Trees Density
0	1ème ArdtR	1	1705	1.826	933.734940
1	2ème Ardt	2	538	0.992	542.338710
2	3ème Ardt	3	1244	1.171	1062.339880
3	4ème Ardt	4	2765	1.601	1727.045597
4	5ème Ardt	5	2505	2.541	985.832349
5	6ème Ardt	6	1783	2.154	827.762303
6	7ème Ardt	7	8741	4.088	2138.209393
7	8ème Ardt	8	7247	3.881	1867.302242
8	9ème Ardt	9	1188	2.179	545.204222
9	10ème Ardt	10	3436	2.892	1188.105118
10	11ème Ardt	11	5919	3.666	1614.566285
11	12ème Ardt	12	12601	16.320	772.120098
12	13ème Ardt	13	16904	7.146	2365.519172
13	14ème Ardt	14	12021	5.621	2138.587440
14	15ème Ardt	15	17289	8.502	2033.521524
15	16ème Ardt	16	17119	16.300	1050.245399
16	17ème Ardt	17	11071	5.669	1952.901746
17	18ème Ardt	18	10470	6.005	1743.547044
18	19ème Ardt	19	14428	6.786	2126.142057
19	20ème Ardt	20	15572	5.984	2602.272727

Afterwards, Land uses Data was obtained and then sorted by categories, as seen in the heatmap below:



It is obvious that there is a large area of forests in the 12th and 16th districts, and that most of the neighborhoods have a lot of areas of gardens and squares. cemeteries are abundant in the 12, 14-20 districts, but perhaps cemetery is not considered a green area per se. Below the Green areas in the different districts are displayed in a folium map:



Unfortunately, this dataset is not 100% accurate - for example, "Jardin du Luxemborg" which have a large area, is not included in the dataset. Here the total green area of each neighborhood was summed and then divided by the area of each neighborhood, giving the Percentage of green areas for each district. Not surprisingly, the 12th and 16th districts are on the lead:

	Percentage	NBHD_ID
0	1.394688	1
1	0.331452	2
2	2.202818	3
3	4.233542	4
4	3.161826	5
5	1.476416	6
6	9.366585	7
7	5.810925	8
8	0.815236	9
9	2.095609	10
10	2.798691	11
11	64.073444	12
12	5.770949	13
13	10.520228	14
14	8.942614	15
15	56.124742	16
16	7.954454	17
17	9.025079	18
18	10.139257	19
19	14.776237	20

5. vegan restaurants using Google API:

As a replacement for the sparse Foursquare data, vegan restaurants venues within 500 meter radius of each neighborhood centroid were collected from the Google API Using the getGOOGLEVenues function:

	neighborhood	latitude	longitude	vegan_NBHD
0	1er Ardt	48.862563	2.336443	7
1	2ème Ardt	48.868279	2.342803	14
2	3ème Ardt	48.862872	2.360001	15
3	4ème Ardt	48.854341	2.357630	10
4	5ème Ardt	48.844443	2.350715	6
5	6ème Ardt	48.849130	2.332898	5
6	7ème Ardt	48.856174	2.312188	2
7	8ème Ardt	48.872721	2.312554	5
8	9ème Ardt	48.877164	2.337458	14
9	10ème Ardt	48.876130	2.360728	9
10	11ème Ardt	48.859059	2.380058	11
11	12ème Ardt	48.834974	2.421325	0
12	13ème Ardt	48.828388	2.362272	2
13	14ème Ardt	48.829245	2.326542	2
14	15ème Ardt	48.840085	2.292826	1
15	16ème Ardt	48.860392	2.261971	0
16	17ème Ardt	48.887327	2.306777	2
17	18ème Ardt	48.892569	2.348161	3
18	19ème Ardt	48.887076	2.384821	1
19	20ème Ardt	48.863461	2.401188	4

Contrary to previous results, most of the venues were located at the central districts, maybe due to the large abundance of restaurants there in general.

Results - Clustering

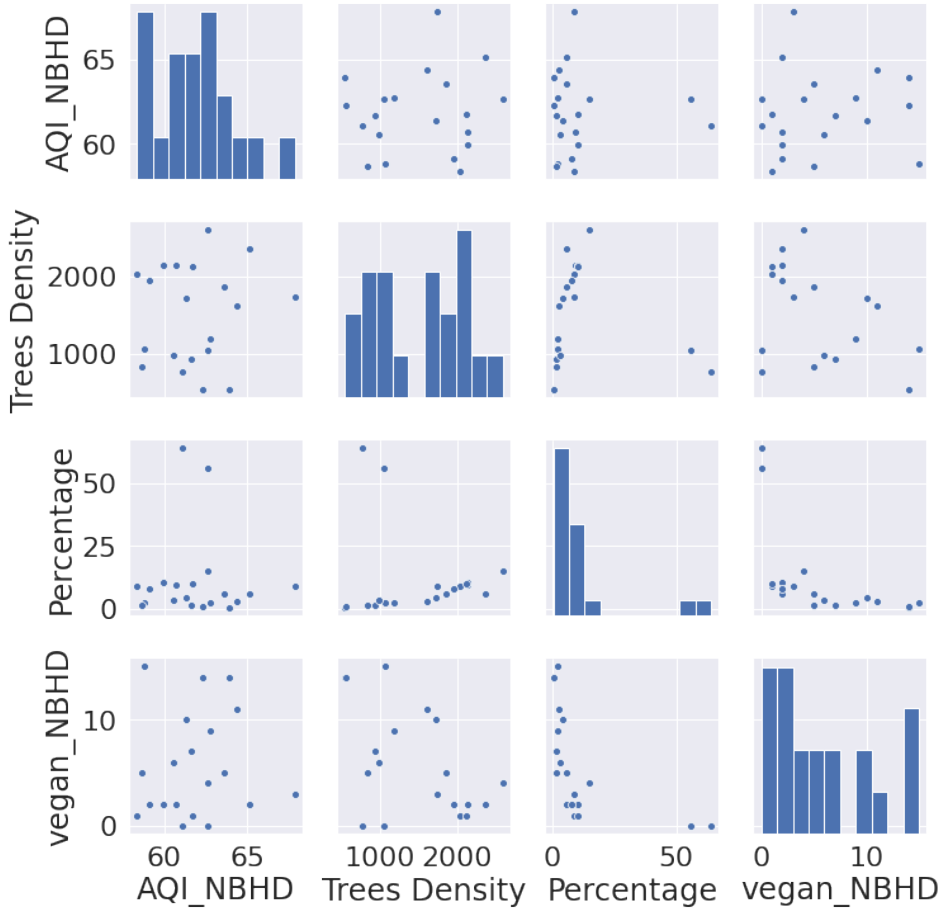
For the clustering, the foresquare API will not be used. Let's see how well correlated are the 4 other datasets - Trees density, green area percentage, vegan restaurants and Air quality:

	AQI_NBHD	Trees Density	Percentage	vegan_NBHD
0	61.657534	933.734940	1.394688	7
1	63.917808	542.338710	0.331452	14
2	58.849315	1062.339880	2.202818	15
3	61.356164	1727.045597	4.233542	10
4	60.561644	985.832349	3.161826	6
5	58.698630	827.762303	1.476416	5
6	60.739726	2138.209393	9.366585	2
7	63.575342	1867.302242	5.810925	5
8	62.301370	545.204222	0.815236	14
9	62.739726	1188.105118	2.095609	9
10	64.383562	1614.566285	2.798691	11
11	61.082192	772.120098	64.073444	0
12	65.150685	2365.519172	5.770949	2
13	59.931507	2138.587440	10.520228	2
14	58.383562	2033.521524	8.942614	1
15	62.630137	1050.245399	56.124742	0
16	59.095890	1952.901746	7.954454	2
17	67.849315	1743.547044	9.025079	3
18	61.739726	2126.142057	10.139257	1
19	62.630137	2602.272727	14.776237	4

Correlation matrix:

	AQI_NBHD	Trees Density	Percentage	vegan_NBHD
AQI_NBHD	1.000000	0.089575	-0.007621	0.071729
Trees Density	0.089575	1.000000	-0.123973	-0.516990
Percentage	-0.007621	-0.123973	1.000000	-0.539352
vegan_NBHD	0.071729	-0.516990	-0.539352	1.000000

Pairplot of the different variables:



A few trends can be seen in this plot and based on the correlation coefficients: There isn't a strong correlation between any of the parameters, except from, somehow, trees density and number of vegan restaurants (Correlation coefficient of 0.52). Secondly, The percentage data histogram is devided to districts with high percentage (12,16) and other districts with much lower percentage of green areas. Furthermore, there are more districts with lower then higher AQI index. Here is a statistical sumary of the variables using the describe method:

	AQI_NBHD	Trees Density	Percentage	vegan_NBHD
count	20.000000	20.000000	20.000000	20.000000
mean	61.863699	1510.864912	11.050739	5.650000
std	2.392522	640.895115	17.281308	4.912658
min	58.383562	542.338710	0.331452	0.000000
25%	60.404110	972.807997	2.176016	2.000000
50%	61.698630	1670.805941	5.790937	4.500000
75%	62.948630	2056.676658	9.559753	9.250000
max	67.849315	2602.272727	64.073444	15.000000

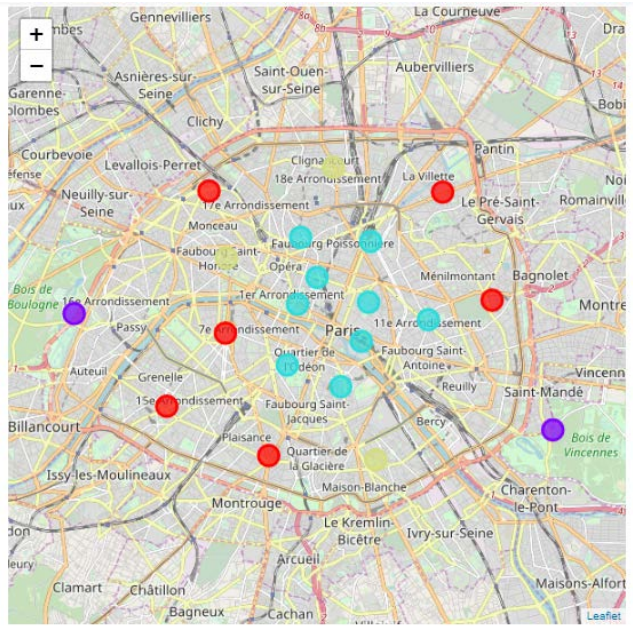
In the final step, the Dataframe was normalized and then clustered. Random state of 0 and 4 cluster groups were defined. The average values for each cluster are as follows:

Labels	AQI	Trees Density	Green area Percentage	Vegan restaurants
0	**61.9**	**911**	**60**	**0**
1	**61.6**	**1047**	**2**	**10**
2	**65.5**	**1992**	**7**	**3**
3	**60.4**	**2165**	**10**	**2**

And finally, below is a map of the different calculated clusters:

Eco-Clusters - Paris

(Cluster 0 - Purple, Cluster 1 - Light Blue, Cluster 2 - Beige, Cluster 3 - Red)



Discussion and Conclusions

In this project, several datasets of the 20 districts of the city of Paris were obtained, analyzed and visualized in order to explore the environmental features of these districts. Afterwards, a k-mean clustering method was applied on the normalized data, so that the districts were categorized to unique categories. The Fouresquare data was not eventually taken into account, due to insufficient data.

It is important to state a few reservations: The air quality data is taken from specific points and for a narrow time window (72 hours), therefore it is not the most accurate assessment of the average air quality

in each district. Besides that, There are some discrepancies between the green areas data and the actual green areas within the districts, as some public gardens and parks are not included in the datasets.

Heaving said that, a few trends were observed from the datasets: It was shown in the AQI dataset that the 5th and 13th districts exhibited the best air quality (above 65), while the 1,4,7,14,15 districts exhibited the worse relative AQI (below 60). It was also shown that the 12th and 16th districts have the largest percentage of green areas, mainly due to the forests within their boundaries. On the other hand, the 2th and 9th districts have a very few green areas. Gardens and Squares were the most common green land uses in all of the districts. The data obtained from the Google API revealed, that most of the vegan restaurants are located in the center of Paris, in the 2nd, 3rd and 9th districts, maybe because these districts have in general more restaurants due to their popularity among tourists. Contrary to that, the number of trees was smaller in the central districts, and larger in the surrounding neighborhoods.

In the k-mean clustering, all of the parameters above were taken into consideration. Random state of 0 and 4 cluster groups were defined. The clusters are:

1. **The 12th and 16th Districts** - These districts are adjacent to forests, but their air quality is relatively poor and have low number of vegan restaurants.
2. **1-6, 9-11 Districts**, located at the center of Paris, have the largest number of vegan restaurants, but suffer from poor air quality and low percentage of green areas and low number of trees.
3. **8th, 13th and 18th Districts** showed the best air quality relative to the other neighborhoods, and had a large number of trees, but lacked green areas and vegan restaurants.
4. **7th, 14th, 15th, 17th, 19th and 20th Districts** all suffer from low air quality, low percentage of green areas and small number of vegan restaurants.

To conclude, the most eco-friendly districts in Paris, according to this analysis, are 12 and 16 in aspects of green areas, 8, 13 and 18 in aspects of air quality and trees, and 1-6, 9-11 in aspects of eco-friendly business vegan restaurants.