Capstone Project - The Battle of Neighborhoods

1. Introduction/Business Problem

The problem is about if it's possible to compare neighborhoods of 5 multicultural cities (London, Sydney, Paris, Amsterdan and San Francisco) in order to find groups of neighborhoods that are similar according to similarities in the categories of businesses that are in them. This information can be used for example for travelers who wish to stay near a commercial area similar to where they come from, it could be also usefull for businesses that want to establish themselves in other cities of the world where there are commercial areas similar to the area where they currently are.

2. Data

The data used for resolve the problem is:

- Foursquare dataset: Foursquare is a local search-and-discovery service mobile app which provides search results for its users. The app provides personalized recommendations of places to go to near a user's current location based on users' previous browsing history, purchases, or check-in history. The Foursquare API allows application developers to interact with the Foursquare platform. The API itself is a RESTful set of addresses to which you can send requests, so there's really nothing to download onto your server.
  The features that will be used from the data are:
  - Bussiness geo location
  - Category
  - Name
- Neigbourhoods geo location data from http://insideairbnb.com/get-the-data.html

  The features that will be used from the data are:
  - Neigbourhood Geo Location
  - Neigbourhood Name

3. Methodology

3.1.Data collection

Dataset were collected using Foursquare API and downloading json files from airbn for neighborhoods geolocation data.

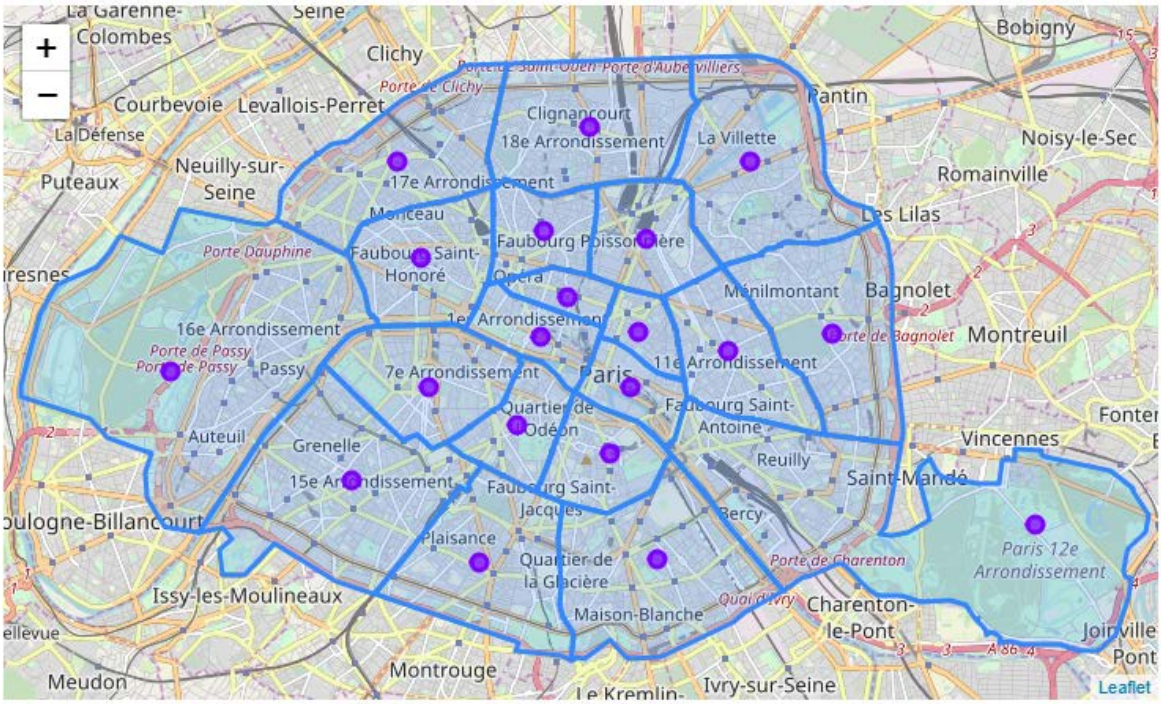
3.2.Data understanding

When carrying out an exploratory analysis of the Airbnb data, it was identified that this data doesn't have a central point for each neighborhood. For this reason it's necessary to perform this calculation in order to have a reference point to explore the venues that exist around.

When exploring the venues of each neighborhood, it was observed that 386 unique categories were found, many of which are in a very small proportion in the neighborhoods (142 neighborhoods) which results in very dispersed variables and probably will not behave well when we are going to use them to build the clusters

3.3.Data preparation

The calculation of the center of each neighborhood was made using the shapely library and a test of the data was conducted with the city of Paris, resulting in:



A Python function was created to obtain the categories that are at least in a "p" proportion of the neighborhoods, in order to test different proportions when building the clusters and to identify the number of categories that rerult in the best groupings of the data

3.4.Modeling

For modeling, the k-means algorithm were used in order to identify groups of neighborhoods similar to each other in the distribution of categories of places that are in their surroundings.

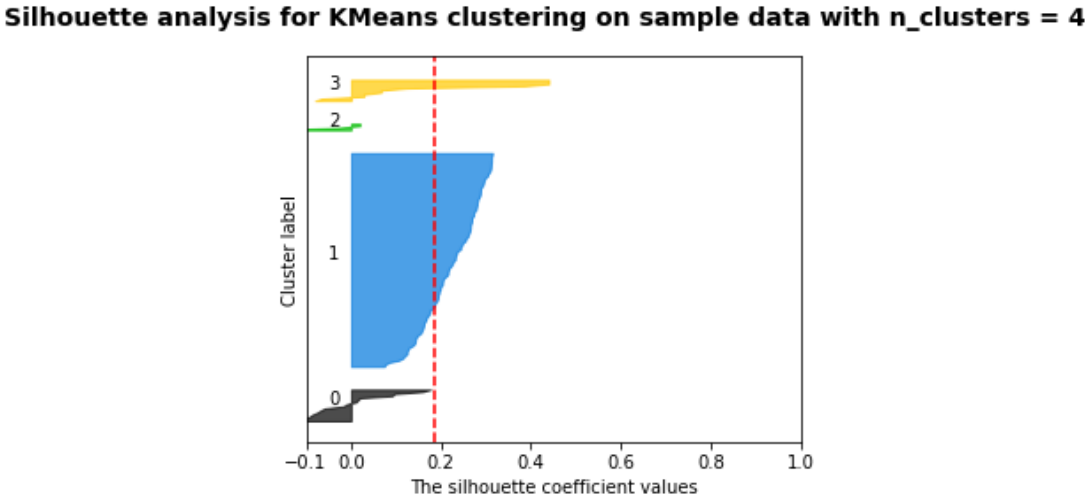
3.5.Evaluation

For the model evaluation the silhouette metric were used wich is a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster.

Likewise, it was taken into account what proportion of neighborhoods for the selection of categories yielded a better distribution of the clusters.

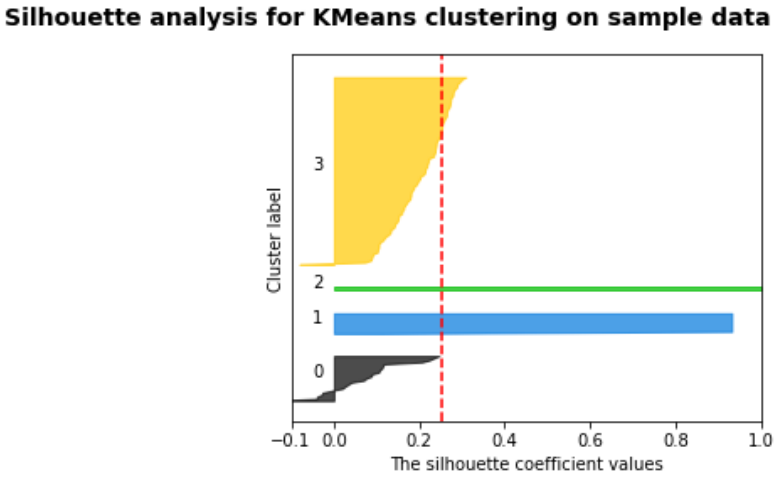
Next, the results for 4 clusters evaluated with different proportions are shown. It can be seen that increasing the minimum proportion improves the distribution of venues in each cluster although some neighborhoods are excluded, we could say that these neighborhoods are atypical and it's preferable to exclude them from the analysis.

Test with p=0



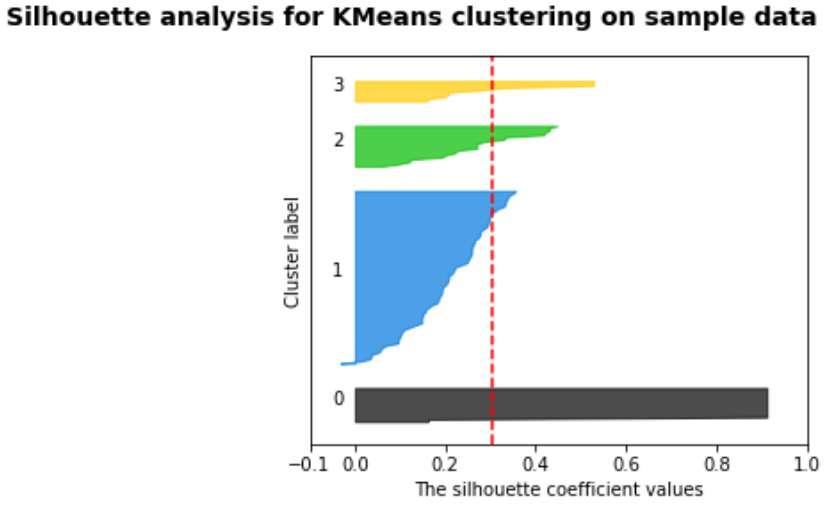

```
Number of items in cluster 0: 9
Number of items in cluster 1: 119
Number of items in cluster 2: 1
Number of items in cluster 3: 13
```

Test with p=0.1



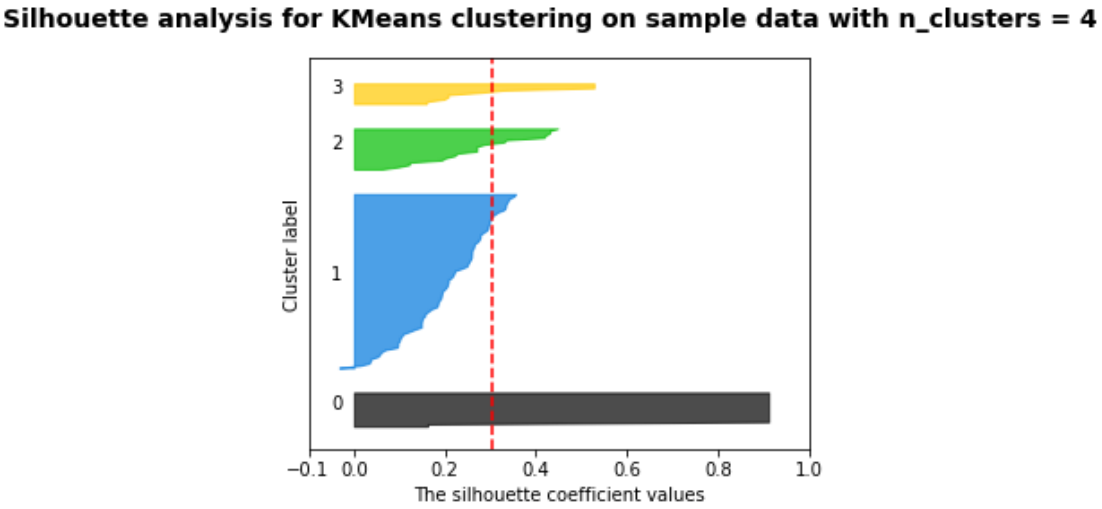

```
Number of items in cluster 0: 16
Number of items in cluster 1: 103
Number of items in cluster 2: 4
Number of items in cluster 3: 11
```

Test with p=0.2



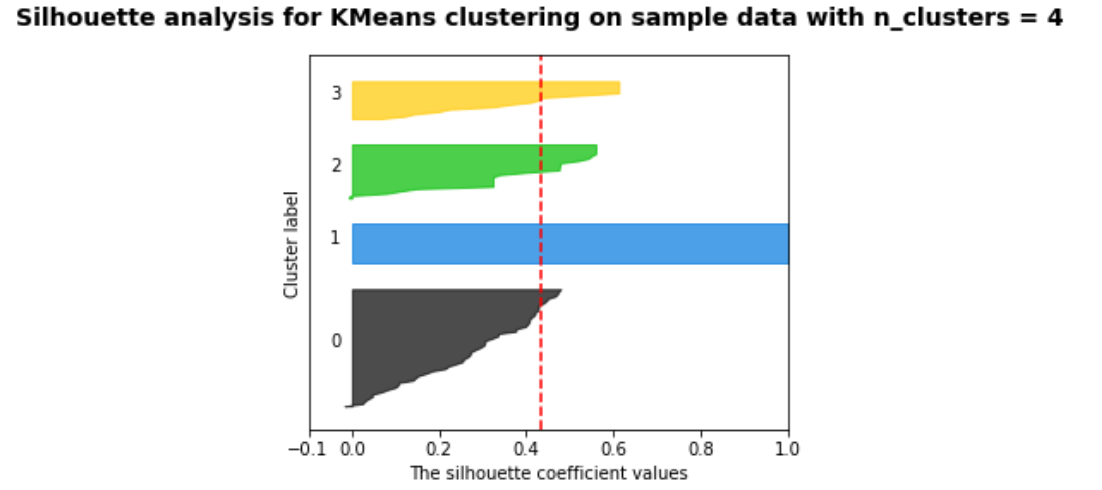

```
Number of items in cluster 0: 23
Number of items in cluster 1: 11
Number of items in cluster 2: 3
Number of items in cluster 3: 93
```

Test with p=0.3




```
Number of items in cluster 0: 16
Number of items in cluster 1: 78
Number of items in cluster 2: 19
Number of items in cluster 3: 10
```

Test with p=0.4




```
Number of items in cluster 0: 51
Number of items in cluster 1: 18
Number of items in cluster 2: 24
Number of items in cluster 3: 17
```

4. Results

The best aproach was 4 clusters with 0.4 proportion.

Proportion of venue categories for each cluster:

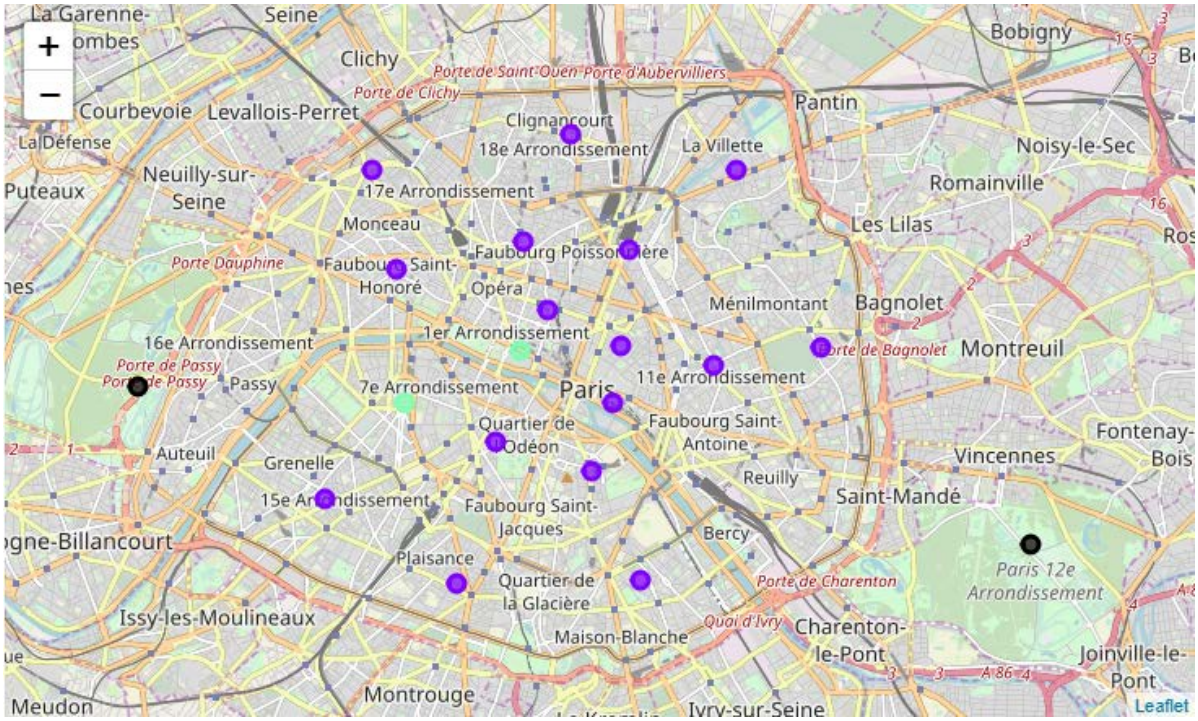
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Bakery	0.224165	0.0	0.056003	0.072941
Café	0.175256	0.0	0.680374	0.100392
Coffee Shop	0.195146	0.0	0.055889	0.720196
Italian Restaurant	0.269075	0.0	0.066020	0.030098
Park	0.136358	1.0	0.141713	0.076373

Number of neighborhoods for city for cluster

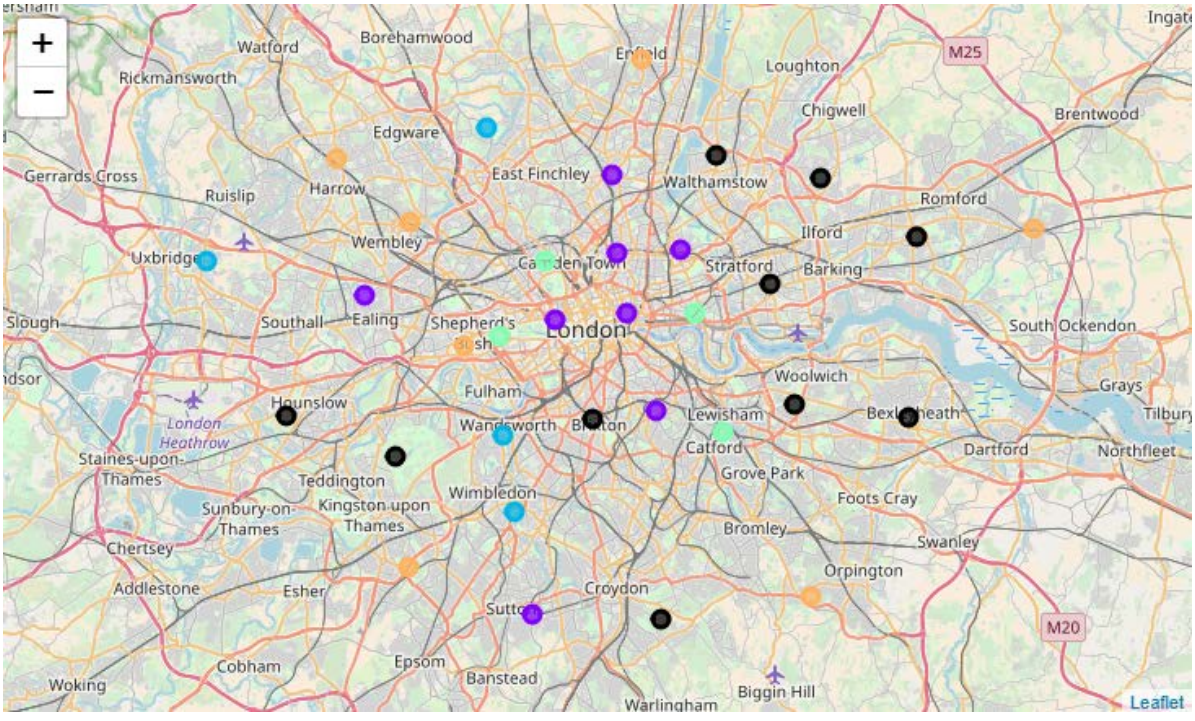
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Amsterdam	9	3	3	1
London	8	4	4	7
Paris	16	0	2	0
San Francisco	15	8	4	8
Sydney	3	3	11	1

Clusters distribution:

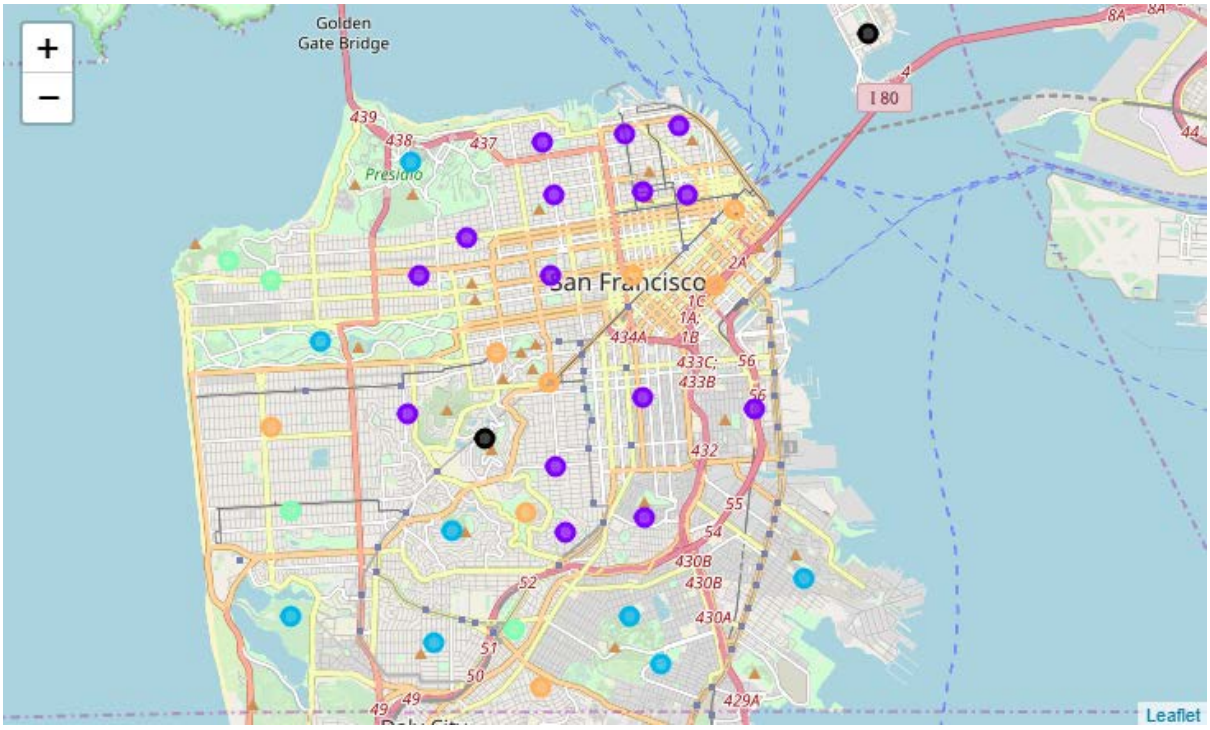
Paris



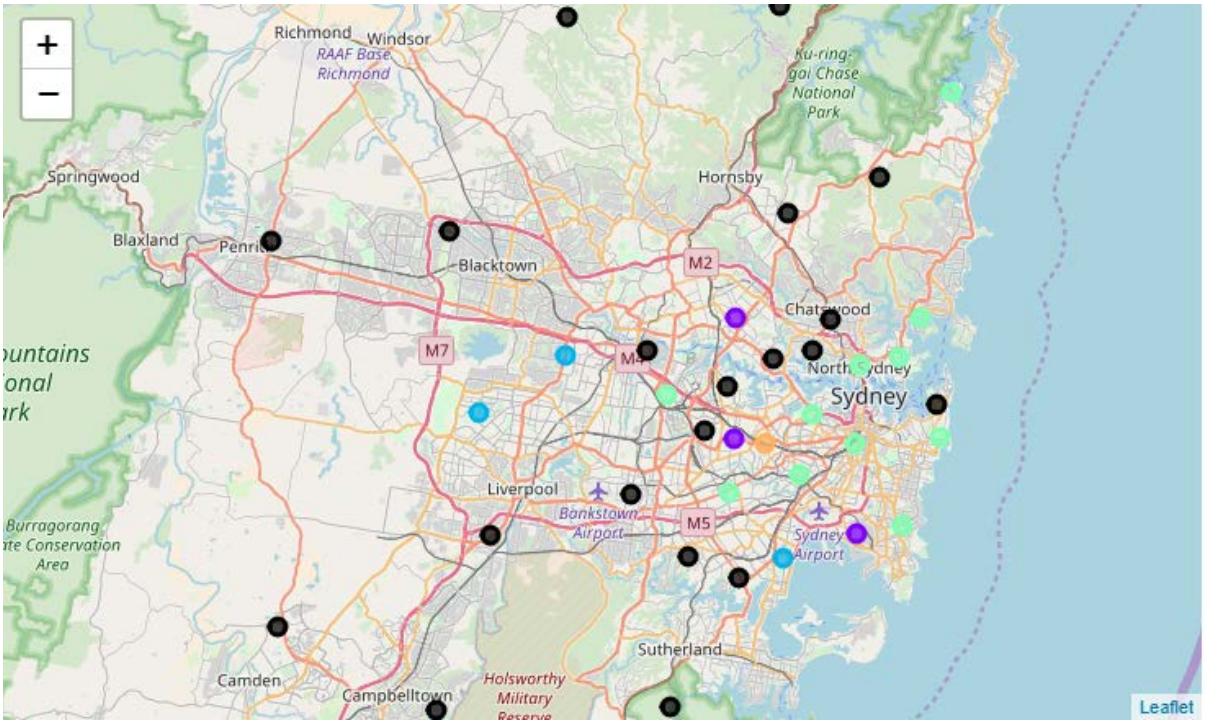
London



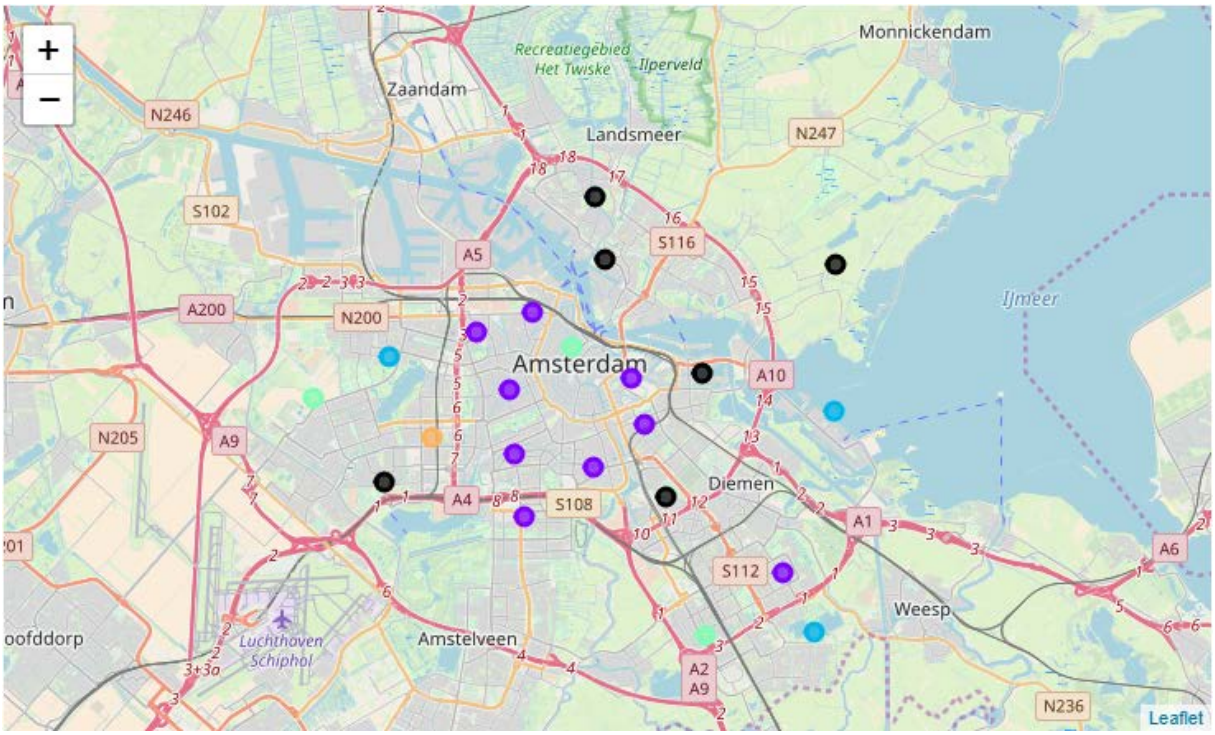
San Francisco



Sidney



Amsterdam



5. Discussion

It’s possible to observe that Cluster 1 has a great variety of categories of venues, almost 100% of the neighborhoods of Paris belong to this cluster, this means that Paris has a very good variety of categories of venues.

Sydney is the city with the largest number of outliers, allowing us to conclude that it is the most different city from the other cities evaluated.

London and San Francisco have a very similar distribution of neighborhoods belonging to the different clusters, although London is the second with the highest number of outliers.

Likewise Amsterdam is very similar to Sydney with respect to the distribution of clusters.

Cluster 2 corresponds to neighborhoods with 100% of parks. Cluster 3 is more represented by neighborhoods that own cafes and parks, while Cluster 4 by coffee shops and cafes.

6. Conclusion

When applying machine learning techniques, it is very important to perform exploratory data analyzes in order to identify possible problems that may result in low performance of the models.

Georeferencing data are very useful to identify interesting patterns that allow comparing different cities and making important decisions when establishing new businesses. Foursquare is a valuable tool that offers free georeferencing data that can be very useful for companies.