DATA SCIENCE CAPSTONE PROJECT

Víctor Pagán Rubio

vpaganrubio@gmail.com

December 27, 2020

Índice de contenido

Executive Summary.

Foursquare data could help differentiate tourism activities to offer to travelers. Paris, Brussels, Rome and Madrid are great cities where a tourist can spend several weeks exploring different activities. A Tourism Agency could be interested in showing different profiles of a city paying attention to its neighborhoods and its possible activities. Each neighborhood and its geolocation data can be used to get Foursquare information to search for interesting activities. The, kMeans is used to generate 5 clusters of similar neighborhoods which can be seen in the city map. All these cicties have a city center and a radius in which all neighbourhoods are located so they could be easy to explore. All of them have a main cluster with most similar neighborhoods.

Introduction.

Foursquare data could help differentiate tourism activities to offer to travelers. Paris, Brussels, Rome and Madrid are great cities where a tourist can spend several weeks exploring different activities.

Business Problem.

Tourist Agency could be interested in showing different profiles of a city paying attention to its neighborhoods and its possible activities.
The target audience would be the Tourism Companies interested to offer interesting trips to travelers.
Also data could be used to recommend cities to visit to tourists.

Data Aquisition and cleaning.

Geo data can be used to map neighborhoods.

Data can be merged to have neighborhoods and its location to explore data in Foursquare.

Four csv files data (Neighborhood, Borough, Latitude, Longitude) are used.

Foursquare Venues Data.

Foursquare Data can be used to find venues, for example :

Paris	**Brussels**	**Rome**	**Madrid**
Bakery	Bakery	Italian Restaurant	Bar
Bar	Bar	Basketball Stadium	Beer Garden
Beer Bar	Bookstore	Boutique	Bistro

Beer Store	Burger Joint	Café	Café
Bistro	Café	Coffee Shop	Coffee Shop
Bourse	Chocolate Shop	College Cafeteria	Comfort Food Restaurant
Brewery	Clothing Store	Concert Hall	Concert Hall
Buttes-Chaumont	Convenience Store	Cosmetics Shop	Convenience Store
Buttes-Montmartre	Cosmetics Shop	Cupcake Shop	Deli / Bodega
Café	Department Store	Dessert Shop	Department Store
Cheese Shop	Diner	Diner	Dessert Shop
Cocktail Bar	Fast Food Restaurant	Dog Run	Diner
Coffee Shop	French Restaurant	Fast Food Restaurant	Dog Run
Convenience Store	Greek Restaurant	Flower Shop	Donut Shop
Creperie	Gym	Fountain	Garden
Entrepôt	Gym / Fitness Center	Fried Chicken Joint	Gastropub
French Restaurant	History Museum	Garden Center	Gym / Fitness Center
Gastropub	Hotel	Gift Shop	Hostel
Hotel	Italian Restaurant	Grocery Store	Hotel
Indian Restaurant	Italian Restaurant	Gym	Ice Cream Shop
Italian Restaurant	Kebab Restaurant	Gym Pool	Japanese Restaurant
Japanese Restaurant	Middle Eastern Restaurant	Hotel	Mediterranean Restaurant
Korean Restaurant	Notary	Ice Cream Shop	Mexican Restaurant
Pizza Place	Pizza Place	Italian Restaurant	Nightclub
Plaza	Plaza	Jewelry Store	Park
Seafood Restaurant	Restaurant	Juice Bar	Pizza Place
Supermarket	Sandwich Place	Nightclub	Playground
Wine Bar	Sandwich Place	Noodle House	Plaza

Methodology.

Research Method.

Kmeans.

Five clusters are generated using kMeans for each city.

Run *k*-means to cluster the neighborhood into 5 clusters.

```
# set number of clusters
kclusters = 5

n_grouped_clustering = n_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(n_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

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

Let's create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.

```
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

n_merged = p_n

# merge n_grouped with n_data to add latitude/longitude for each neighborhood
n_merged = n_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

n_merged.head() # check the last columns!
```

Data Sources: City, Neighborhood, Latitude, Longitude.

The Data Sources are csv files using Borough for the name of the city, Neighborhood for the neighborhood and it's latitude and longitude.

Paris Data.

Paris

```
p_n=pd.read_csv('Paris.csv')
```

```
p_n.head()
```

3]:

	Borough	Neighborhood	Latitude	Longitude
0	Paris	Louvre	48.862563	2.336443
1	Paris	Bourse	48.868279	2.342803
2	Paris	Buttes-Chaumont	48.887076	2.384821
3	Paris	Luxembourg	48.849130	2.332898
4	Paris	Passy	48.860392	2.261971

Brussels Data.

Brussels

```
b_n=pd.read_csv('Brussels.csv')
```

```
b_n.head()
```

[8]:

	Borough	Neighborhood	Latitude	Longitude
0	Brussels	Bruxelles-Ville	50.850346	4.351721
1	Brussels	Schaerbeek	50.867416	4.377298
2	Brussels	Etterbeek	50.832578	4.388994
3	Brussels	Ixelles	50.833343	4.366629
4	Brussels	Saint Gilles	50.830144	4.340218

Rome Data.

Rome

```
r_n=pd.read_csv('Rome.csv')
```

```
r_n.head()
```

15]:

	Borough	Neighborhood	Latitude	Longitude
0	Rome	Municipio I – Historical Center	41.902860	12.485487
1	Rome	Municipio II – Parioli/Nomentano	41.922397	12.498321
2	Rome	Municipio III – Monte Sacro	41.942542	12.540979
3	Rome	Municipio IV – Tiburtina	41.921630	12.553682
4	Rome	Municipio V – Prenestino/Centocelle	41.891288	12.551022

Madrid Data.

Madrid

```
m_n=pd.read_csv('Madrid.csv')
```

```
m_n.head()
```

18]:

	Borough	Neighborhood	Latitude	Longitude
0	Madrid	Centro	40.411535	-3.707628
1	Madrid	Arganzuela	40.398889	-3.710203
2	Madrid	Retiro	40.411335	-3.674905
3	Madrid	Salamanca	40.428002	-3.686771
4	Madrid	Chamartin	40.461520	-3.686584

Results Sections.

Clusters in Paris.

Examine Clusters in Paris

Cluster 1

In [41]:
```
n_merged.loc[n_merged['Cluster Labels'] == 0, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

Out[41]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	Palais-Bourbon	Hotel	French Restaurant	Italian Restaurant	Plaza	Café	History Museum	Cocktail Bar	Historic Site	Japanese Restaurant	Gourmet Shop
13	Élysée	French Restaurant	Hotel	Bakery	Spa	Department Store	Cocktail Bar	Resort	Corsican Restaurant	Plaza	Italian Restaurant
15	Batignolles-Monceau	Hotel	French Restaurant	Italian Restaurant	Japanese Restaurant	Bakery	Restaurant	Bistro	Plaza	Café	Korean Restaurant
18	Observatoire	French Restaurant	Hotel	Bistro	Italian Restaurant	Bakery	Brasserie	Fast Food Restaurant	Supermarket	Sushi Restaurant	Tea Room

Cluster 2

In [42]:
```
n_merged.loc[n_merged['Cluster Labels'] == 1, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

Out[42]:

		1st Most	2nd Most	3rd Most	4th Most	5th Most	6th Most	7th Most	8th Most	9th Most	10th Most

Clusters in Brussels.

Examine Clusters in Brussels

Cluster 1

```
n_merged.loc[n_merged['Cluster Labels'] == 0, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

52]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bruxelles-Ville	Chocolate Shop	Plaza	Bar	Bookstore	Hotel	Bakery	Italian Restaurant	Seafood Restaurant	Clothing Store	Sandwich Place
1	Schaerbeek	Tram Station	Supermarket	Plaza	Hookah Bar	Gastropub	Italian Restaurant	Coffee Shop	Middle Eastern Restaurant	Bus Station	Soccer Field
2	Etterbeek	Bar	Sandwich Place	Plaza	Cosmetics Shop	Supermarket	Pizza Place	Snack Place	Diner	Department Store	Kebab Restaurant
3	Ixelles	Bar	Italian Restaurant	Clothing Store	Wine Bar	Art Gallery	Tea Room	Coffee Shop	Bakery	French Restaurant	Plaza
4	Saint Gilles	Bar	Greek Restaurant	Moroccan Restaurant	Bakery	Performing Arts Venue	Pizza Place	Plaza	Brasserie	Friterie	Gym / Fitness Center
5	Anderlecht	Bar	Convenience Store	Plaza	Restaurant	Greek Restaurant	Metro Station	Bakery	Supermarket	History Museum	Italian Restaurant
7	Koekelberg	Gym	History Museum	Bar	Piano Bar	Convenience Store	Sandwich Place	Falafel Restaurant	Soccer Field	French Restaurant	Supermarket

Clusters in Rome.

Examine Clusters in Rome

Cluster 1

:
```
n_merged.loc[n_merged['Cluster Labels'] == 0, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

[84]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Municipio I – Historical Center	Italian Restaurant	Hotel	Plaza	Ice Cream Shop	Boutique	Sandwich Place	Jewelry Store	Fountain	Dessert Shop	Pizza Place
1	Municipio II – Parioli/Nomentano	Italian Restaurant	Seafood Restaurant	Hotel	Plaza	Restaurant	Fountain	Nightclub	Coffee Shop	College Cafeteria	Juice Bar
4	Municipio V – Prenestino/Centocelle	Pizza Place	Café	Supermarket	Italian Restaurant	Noodle House	Sandwich Place	Gym	Basketball Court	Basketball Stadium	Fast Food Restaurant
5	Municipio VI – Roma Delle Torri	Supermarket	Pizza Place	Brewery	Wine Shop	Fried Chicken Joint	Italian Restaurant	Fast Food Restaurant	Middle Eastern Restaurant	Office	Plaza
6	Municipio VII – Appio-Latino/Tuscolano/Cinecittà	Pizza Place	Ice Cream Shop	Pub	Miscellaneous Shop	Italian Restaurant	Gym / Fitness Center	Garden Center	Fried Chicken Joint	Fountain	Flower Shop
7	Municipio VIII – Appia Antica	Pizza Place	Plaza	Diner	Vegetarian / Vegan Restaurant	Bakery	Café	Ice Cream Shop	Supermarket	Italian Restaurant	Soccer Stadium

Clusters in Madrid.

Examine Clusters in Madrid

Cluster 1

]:
```
n_merged.loc[n_merged['Cluster Labels'] == 0, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

[106]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
18	Vicalvaro	Breakfast Spot	Dog Run	Bar	Wine Bar	Dessert Shop	Dumpling Restaurant	Donut Shop	Diner	Department Store	Farmers Market

Cluster 2

]:
```
n_merged.loc[n_merged['Cluster Labels'] == 1, n_merged.columns[[1] + list(range(5, n_merged.shape[1]))]]
```

[107]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Centro	Tapas Restaurant	Plaza	Spanish Restaurant	Bar	Hotel	Hostel	Restaurant	Bistro	Coffee Shop	Dessert Shop
1	Arganzuela	Spanish Restaurant	Bar	Park	Plaza	Playground	Beer Garden	Tapas Restaurant	Ice Cream Shop	Pool	Pizza Place

Empirical Findings.

Most of the neighborhoods have the same different services to offer to the tourist. The city with unique categories greatest number is Paris (204), followed by Brussels (166), next is Madrid (132) and last Rome (80).

Descriptive Stats.

Unique categories in Paris.

Vaugirard	65	65	65	65	65
Élysée	39	39	39	39	39

Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(n_venues['Venue Category'].unique())))
```

```
There are 204 uniques categories.
```

3. Analyze Each Neighborhood in Paris

```
# one hot encoding
```

Unique categories in Brussels.

Watermael-Boitsfort	6	6	6	6	6
Woluwé-St-Lambert	21	21	21	21	21
Woluwé-St-Pierre	14	14	14	14	14

Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(n_venues['Venue Category'].unique())))
```

```
There are 166 uniques categories.
```

Analyze Each Neighborhood in Brussels

```
# one hot encoding
n_onehot = pd.get_dummies(n_venues[['Venue Category']], prefix="", prefix_sep="")
```

Unique categories in Rome.

Municipio XIV – Monte Mario	8	8	8	8
Municipio XV – Cassia/Flaminia	32	32	32	32

Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(n_venues['Venue Category'].unique())))
```

```
There are 80 uniques categories.
```

Analyze Each Neighborhood in Rome

```
# one hot encoding
n_onehot = pd.get_dummies(n_venues[['Venue Category']], prefix="", prefix_sep="")
```

Unique categories in Madrid.

Retiro	28	28	28	28	28	28
Salamanca	77	77	77	77	77	77
San Blas-Canillejas	20	20	20	20	20	20
Tetuan	40	40	40	40	40	40
Usera	12	12	12	12	12	12
Vicalvaro	3	3	3	3	3	3
Villa de Vallecas	9	9	9	9	9	9
Villaverde	5	5	5	5	5	5

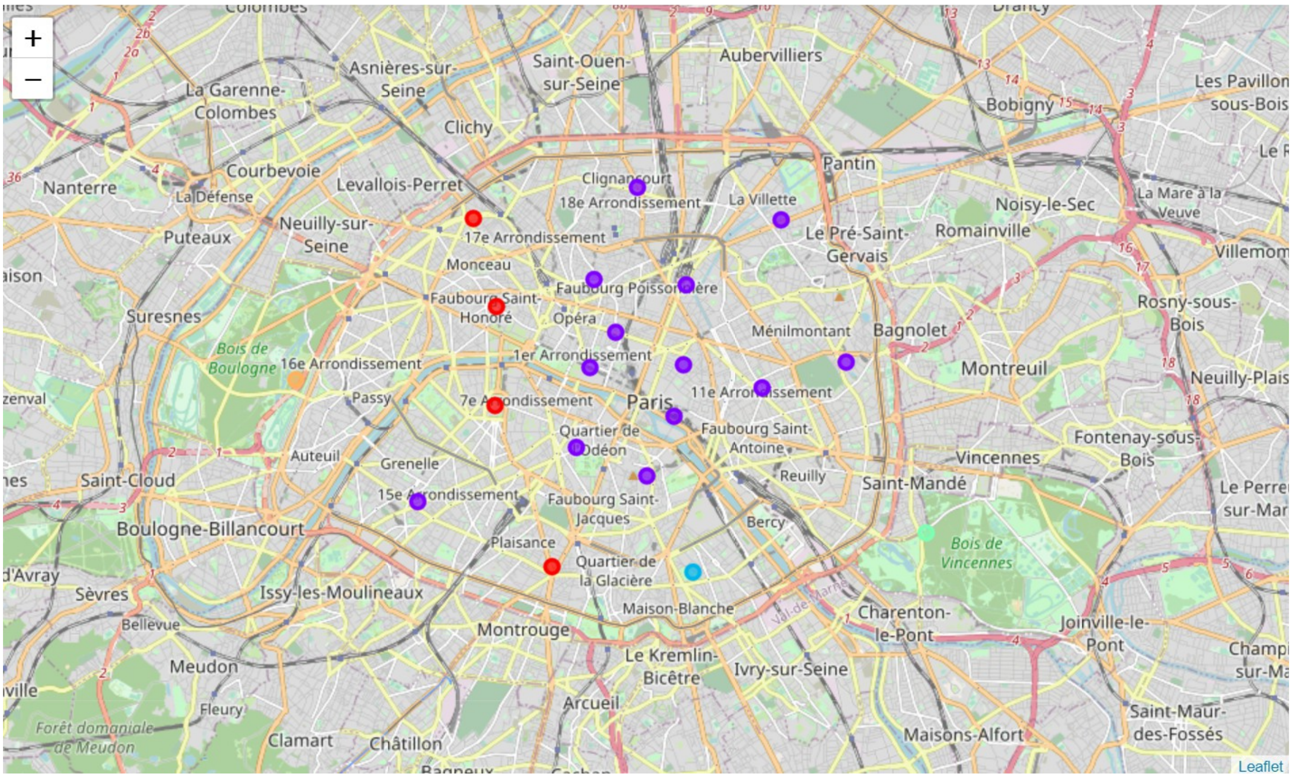
Let's find out how many unique categories can be curated from all the returned venues

```
print('There are {} uniques categories.'.format(len(n_venues['Venue Category'].unique())))
```

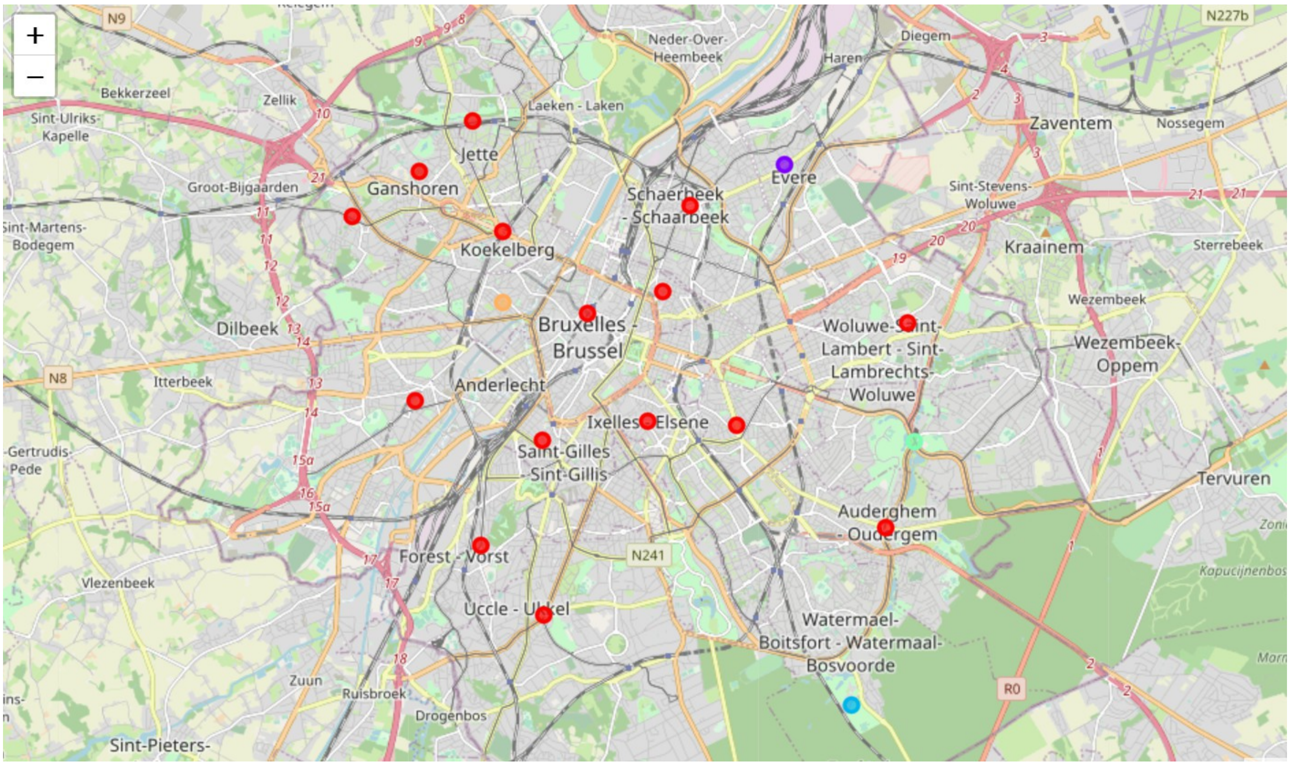
```
There are 132 uniques categories.
```

Illustrative Graphics

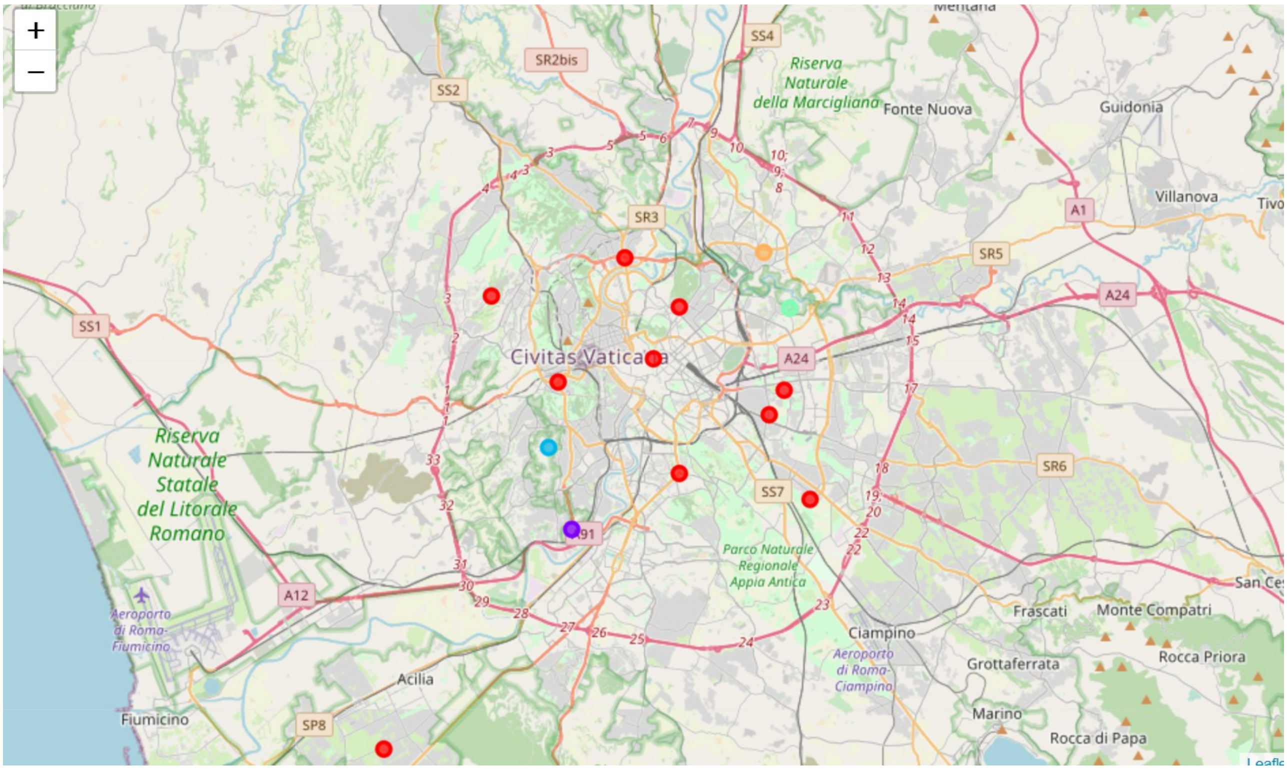
Paris clusters Map



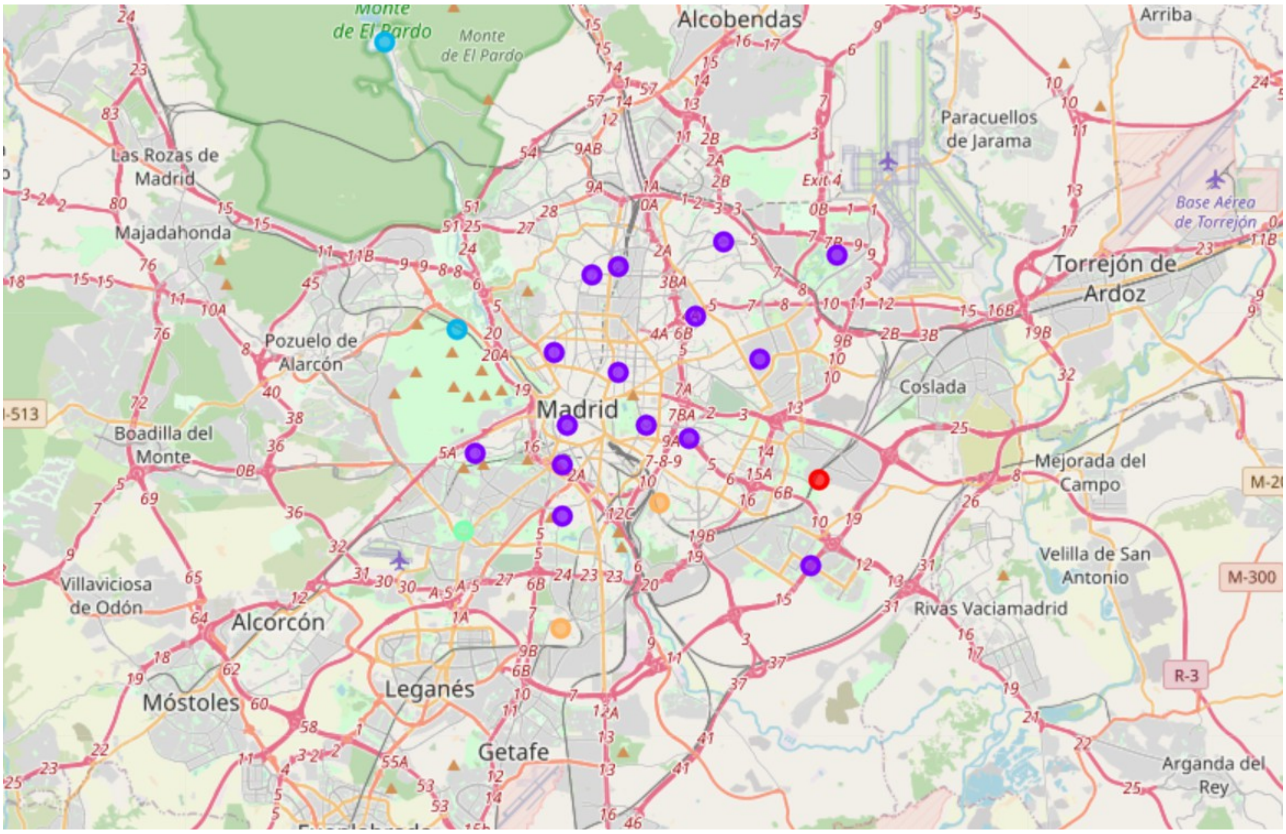
Brussels clusters Map



Rome clusters Map



Madrid clusters Map



Discussion Section

All these cicties have a city center and a radius in which all neighbourhoods are located. All of them have a main cluster with most similar neighborhoods.

Conslusion section.

Similar cities with similar services some of then with more unique categories than others.

References,Acknowledgements and Appendices.

Foursquare.

Google Maps.

Wikipedia.