Predicting the Mumbai suburb venue

Zayneb AMINE

July,1,2020

I : Introduction

I.1 : Background

Mumbai, also known as Bombay, is the capital city of the Indian state of Maharashtra. According to United Nations, as of 2018, Mumbai was the second-most populous city in India after Delhi and the seventh-most populous city in the world with a population of roughly 20 million. As per Indian government population census of 2011, Mumbai was the most populous city in India with an estimated city proper population of 12.5 million living under Municipal Corporation of Greater Mumbai. Mumbai is the centre of the Mumbai Metropolitan Region, the sixth most populous metropolitan area in the world with a population of over 23 million. A such populated city has so many shops and restaurants.

I.2 : Problem

How to search for a specific type of venues, to explore a particular venue, to explore a foursquare user, to explore a geographical location, and to get a trending venues around location. Also how to use the visualization library, folium, and visualize the result. How a person lost in the city can found shops near his position.

What we search ? Top 10 venue of each neighborhoods of Mumbai.

I.3 : Data interest

There is so many people in Mumbai, it could be a real interest for the user to find a shop. Also the city administration can learn how the population live and consumes.

II : Data aquisition and cleaning

II.1 : Data sources

Most mumbai suburb stats are available in the Mumbai Suburb dataset here https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai it is an old dataset dated from 2009.

II.2 : Data cleaning

Data downloaded from source will be combined into one table. Then I will parse data from the html into an object and create a list to store neighborhood data and finally create a dataframe from the list. The I will create another dataframe from the coordinates data. And finally merge it in the first dataframe.

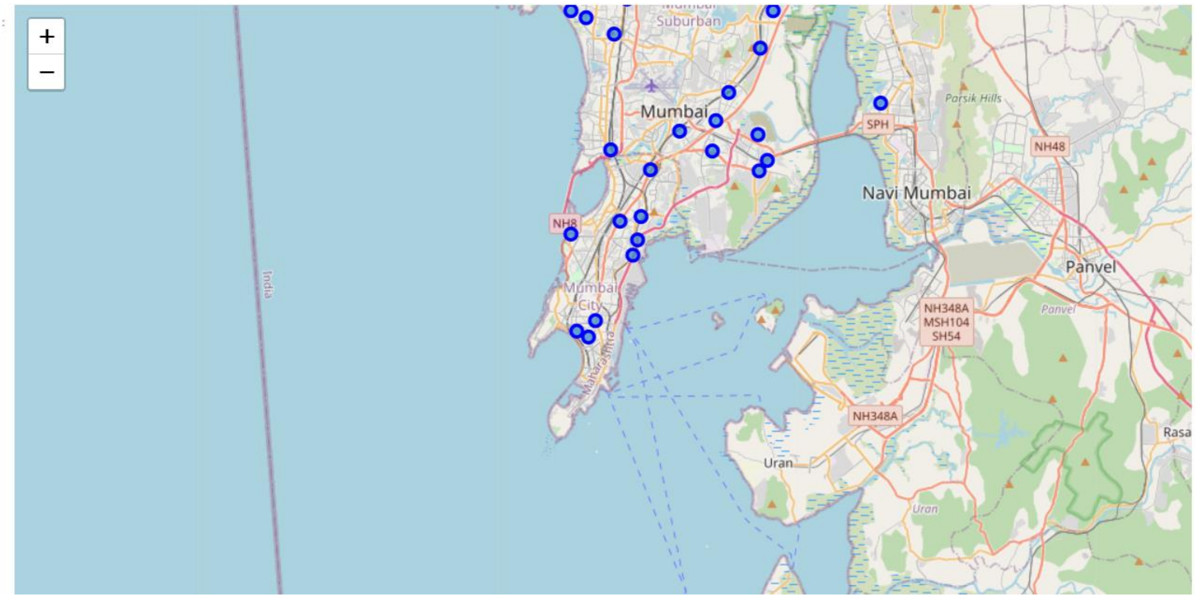
II.3 : Features selection

After cleaning the data there are 199 uniques categories of shop (restaurant). So there is 2738 samples and 199 features.

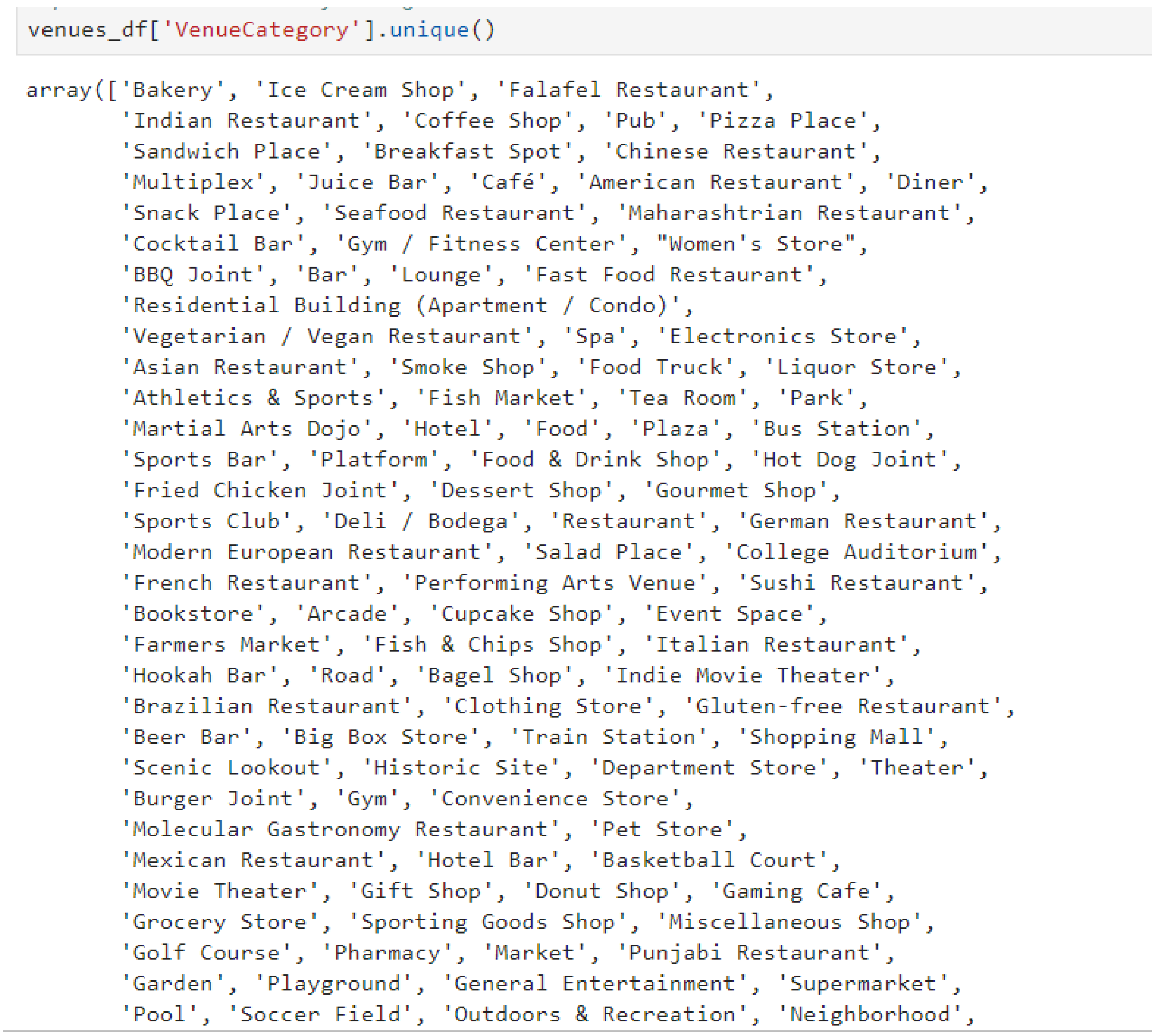
III : Methodology

III.1 : Exploratory Data Analysis

First of all we create a dataframe to populate the coordinates of mumbai, then we merge it in the mumbai suburb dataframe, we put it in a html csv file. And so we can create a map of the neighborhoods of mumbai.



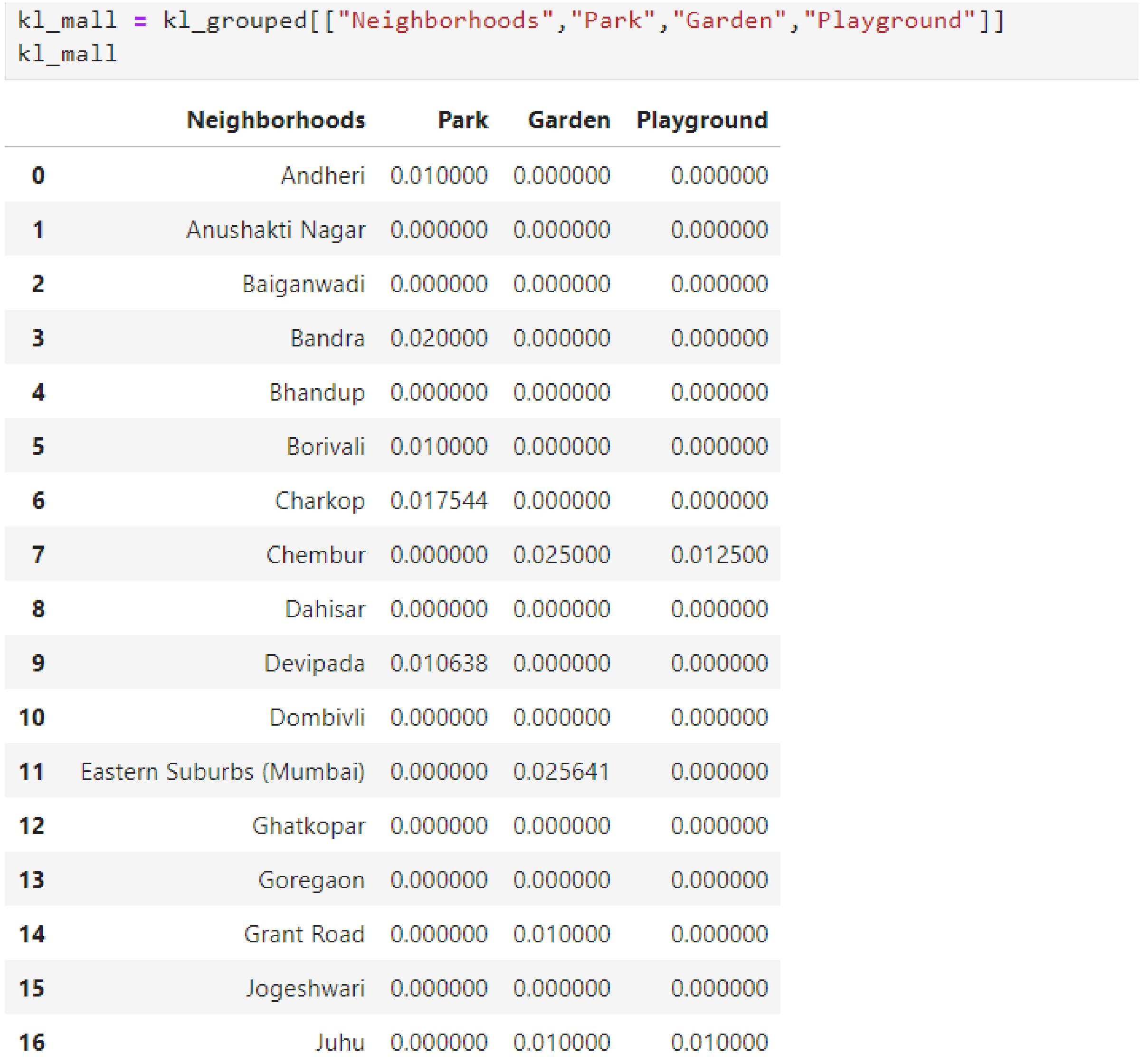
I recovered venues into a dataframe with 7 column : Neighborhood, Latitude, Longitude, VenueName, VenueLatitude, VenueLongitude, VenueCategory. I count for each neighborhood the different columns with the .count function. And I search for the unique cathegory of venue that exists with the .unique function.



```
venues_df['VenueCategory'].unique()
```

```
array(['Bakery', 'Ice Cream Shop', 'Falafel Restaurant',
       'Indian Restaurant', 'Coffee Shop', 'Pub', 'Pizza Place',
       'Sandwich Place', 'Breakfast Spot', 'Chinese Restaurant',
       'Multiplex', 'Juice Bar', 'Café', 'American Restaurant', 'Diner',
       'Snack Place', 'Seafood Restaurant', 'Maharashtrian Restaurant',
       'Cocktail Bar', 'Gym / Fitness Center', "Women's Store",
       'BBQ Joint', 'Bar', 'Lounge', 'Fast Food Restaurant',
       'Residential Building (Apartment / Condo)',
       'Vegetarian / Vegan Restaurant', 'Spa', 'Electronics Store',
       'Asian Restaurant', 'Smoke Shop', 'Food Truck', 'Liquor Store',
       'Athletics & Sports', 'Fish Market', 'Tea Room', 'Park',
       'Martial Arts Dojo', 'Hotel', 'Food', 'Plaza', 'Bus Station',
       'Sports Bar', 'Platform', 'Food & Drink Shop', 'Hot Dog Joint',
       'Fried Chicken Joint', 'Dessert Shop', 'Gourmet Shop',
       'Sports Club', 'Deli / Bodega', 'Restaurant', 'German Restaurant',
       'Modern European Restaurant', 'Salad Place', 'College Auditorium',
       'French Restaurant', 'Performing Arts Venue', 'Sushi Restaurant',
       'Bookstore', 'Arcade', 'Cupcake Shop', 'Event Space',
       'Farmers Market', 'Fish & Chips Shop', 'Italian Restaurant',
       'Hookah Bar', 'Road', 'Bagel Shop', 'Indie Movie Theater',
       'Brazilian Restaurant', 'Clothing Store', 'Gluten-free Restaurant',
       'Beer Bar', 'Big Box Store', 'Train Station', 'Shopping Mall',
       'Scenic Lookout', 'Historic Site', 'Department Store', 'Theater',
       'Burger Joint', 'Gym', 'Convenience Store',
       'Molecular Gastronomy Restaurant', 'Pet Store',
       'Mexican Restaurant', 'Hotel Bar', 'Basketball Court',
       'Movie Theater', 'Gift Shop', 'Donut Shop', 'Gaming Cafe',
       'Grocery Store', 'Sporting Goods Shop', 'Miscellaneous Shop',
       'Golf Course', 'Pharmacy', 'Market', 'Punjabi Restaurant',
       'Garden', 'Playground', 'General Entertainment', 'Supermarket',
       'Pool', 'Soccer Field', 'Outdoors & Recreation', 'Neighborhood',
```

Next we make the one hot encoding and we group it per neighborhood. We just want to keep Park, Garden an Playground.

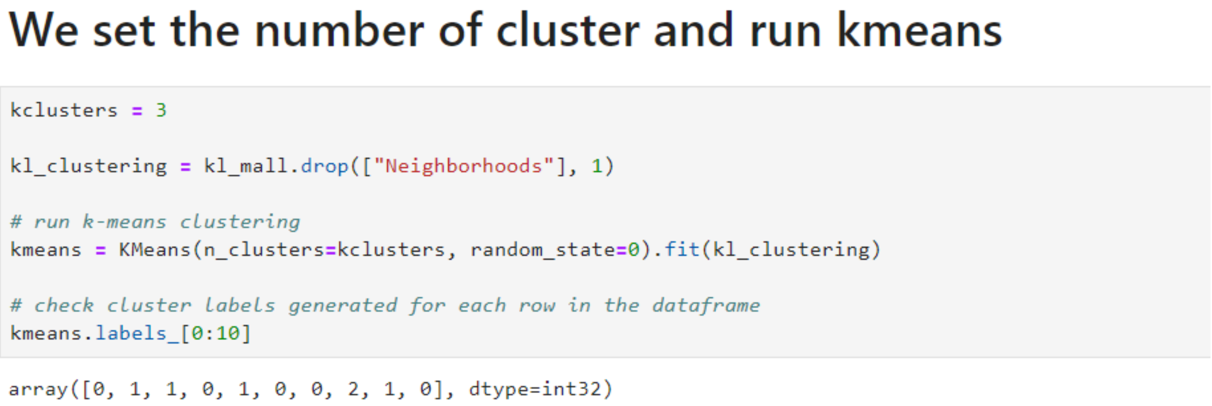


```
kl_mall = kl_grouped[["Neighborhoods","Park","Garden","Playground"]]
kl_mall
```

	Neighborhoods	Park	Garden	Playground
0	Andheri	0.010000	0.000000	0.000000
1	Anushakti Nagar	0.000000	0.000000	0.000000
2	Baiganwadi	0.000000	0.000000	0.000000
3	Bandra	0.020000	0.000000	0.000000
4	Bhandup	0.000000	0.000000	0.000000
5	Borivali	0.010000	0.000000	0.000000
6	Charkop	0.017544	0.000000	0.000000
7	Chembur	0.000000	0.025000	0.012500
8	Dahisar	0.000000	0.000000	0.000000
9	Devipada	0.010638	0.000000	0.000000
10	Dombivli	0.000000	0.000000	0.000000
11	Eastern Suburbs (Mumbai)	0.000000	0.025641	0.000000
12	Ghatkopar	0.000000	0.000000	0.000000
13	Goregaon	0.000000	0.000000	0.000000
14	Grant Road	0.000000	0.010000	0.000000
15	Jogeshwari	0.000000	0.000000	0.000000
16	Juhu	0.000000	0.010000	0.010000

III.2 : Machine Learning used

I decided to use the Kmeans method to analysed the data. For that, I create 3 clusters and I applied the kmeans method on the dataframe where I drop the neighborhood column.

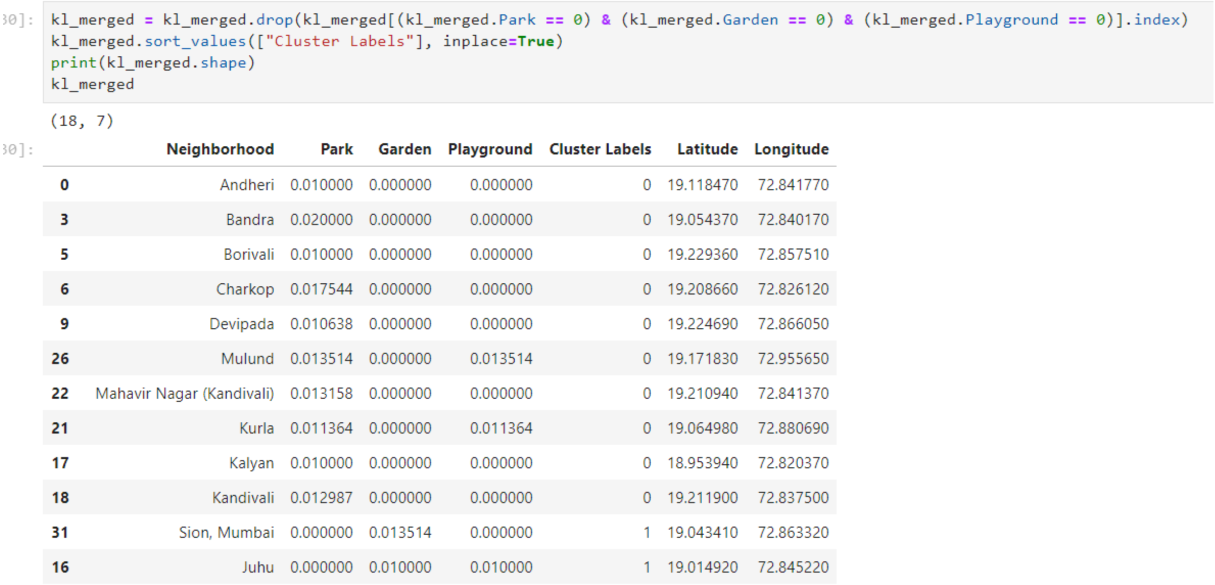


We set the number of cluster and run kmeans

```
kclusters = 3

kl_clustering = kl_mall.drop(["Neighborhoods"], 1)

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

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

```
array([0, 1, 1, 0, 1, 0, 0, 2, 1, 0], dtype=int32)
```

In order to create the top 10 venues for each neighborhood I create a new dataframe that is the same as the precedent and I add the cluster column.



```
kl_merged = kl_merged.drop(kl_merged[(kl_merged.Park == 0) & (kl_merged.Garden == 0) & (kl_merged.Playground == 0)].index)
kl_merged.sort_values(["Cluster Labels"], inplace=True)
print(kl_merged.shape)
kl_merged
```

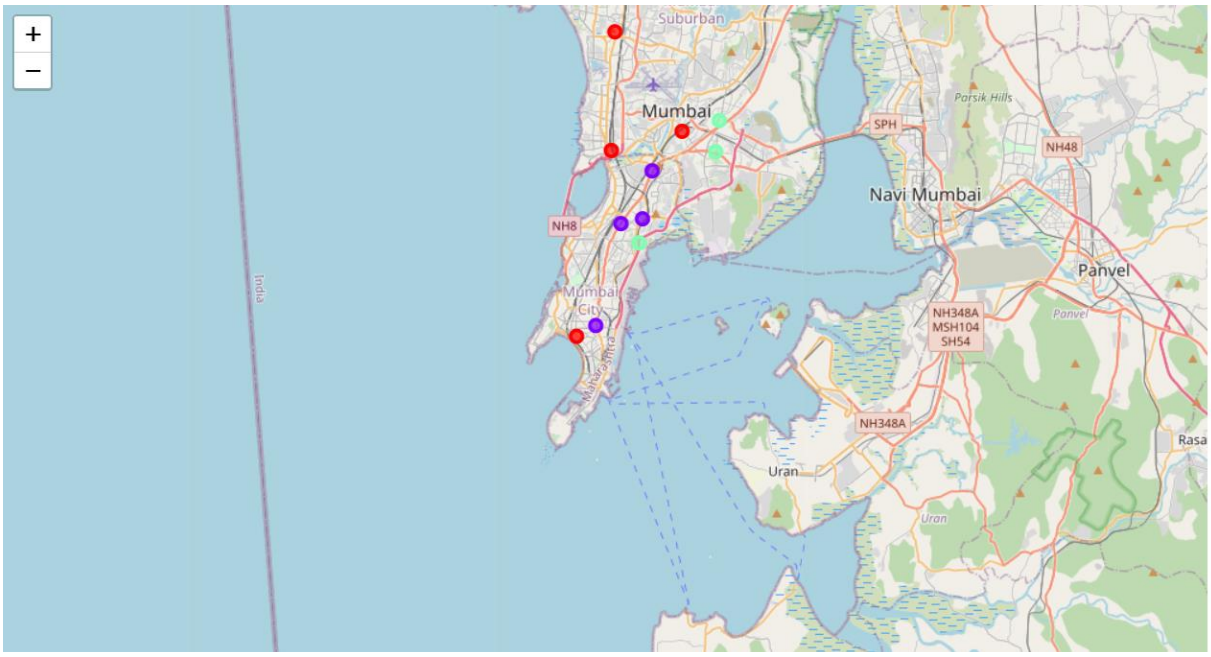
```
(18, 7)
```

	Neighborhood	Park	Garden	Playground	Cluster Labels	Latitude	Longitude
0	Andheri	0.010000	0.000000	0.000000	0	19.118470	72.841770
3	Bandra	0.020000	0.000000	0.000000	0	19.054370	72.840170
5	Borivali	0.010000	0.000000	0.000000	0	19.229360	72.857510
6	Charkop	0.017544	0.000000	0.000000	0	19.208660	72.826120
9	Devipada	0.010638	0.000000	0.000000	0	19.224690	72.866050
26	Mulund	0.013514	0.000000	0.013514	0	19.171830	72.955650
22	Mahavir Nagar (Kandivali)	0.013158	0.000000	0.000000	0	19.210940	72.841370
21	Kurla	0.011364	0.000000	0.011364	0	19.064980	72.880690
17	Kalyan	0.010000	0.000000	0.000000	0	18.953940	72.820370
18	Kandivali	0.012987	0.000000	0.000000	0	19.211900	72.837500
31	Sion, Mumbai	0.000000	0.013514	0.000000	1	19.043410	72.863320
16	Juhu	0.000000	0.010000	0.010000	1	19.014920	72.845220

Then we merge it with the dataframe that contains the longitude and latitude data. Next we sort the value in function of cluster labels. Finally we display the map of this dataframe.

IV : Result

The final map we display represent the cluster per neighborhoods



V : Discussion

I have conscious that my project looks like what we did with toronto but I am novice and I thought it was beter to train myself.

VI : Conclusion

In conclusion, we can see that there is 11 important cluster in mumbai suburb where there is the top venues of the neighborhoods.