Coursera Capstone Project

Café Recommendation

Jaynik Gaglani

12 - 02 - 2020

1. Introduction:

1.1 Background:

Opening a new business has to take many factors into consideration. Some of the factors include market, number of competitors, property rent, etc. A place with less competition may result into high profit since the business would be a monopoly. It is very important to get the data about such places for the business to flourish. Hence, it is important to build a system through which we can find the perfect place to open our business.

1.2 Problem Definition:

Creating a system for finding how many businesses of same category are present in an area and doing analysis upon that to acquire the perfect location for starting a business. The implementation is done keeping only cafes in mind but it can be extended to other types of business too.

1.3 Scope:

Current scope of the project finds the density of cafes in the city of Mumbai. It shows the neighborhood with the density of cafes present. The project can be implemented for any type of business by just changing one parameter. Café parameter was selected since the data availability was high.

2. Data Acquisition and Cleaning:

2.1 Data Source:

The data was collected from Wikipedia- https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai.

It consists of list of neighborhoods in Mumbai. The data was scraped using BeautifulSoup library of python. Geocoder library of python was used to obtain the coordinates of each neighborhood. The dataframe consisted of Neighborhood, Latitude and Longitude.

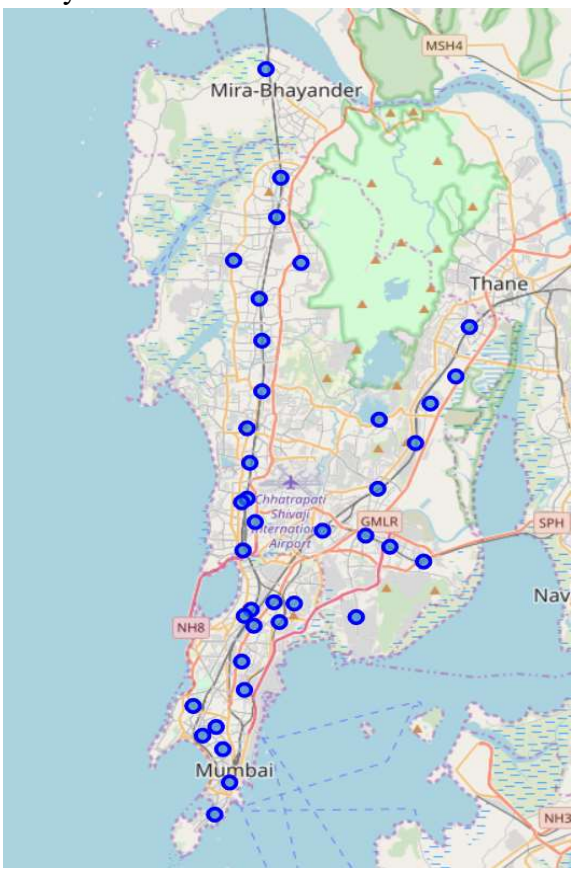
2.2 Data Cleaning:

Before creating the dataframe, the data was cleaned. Since the scraped data contained numbers along with the names, all the numbers were removed from each row. Extra rows like references, others, etc. were also removed. For removing the numbers, a regular expression was used. Hence, re python library was also used. The index was reset after deleting the rows. After cleaning, the coordinates were added to the dataframe with individual columns of Latitude and Longitude.

3. Methodology:

3.1 Generating the Map:

The map was generated using the folium library. Each neighborhood of Mumbai was marked as available in the data. Using the foursquare API, all the venues as per categories were found out. The number of distinct venue categories found in Mumbai were 209. Analysis of each neighborhood was found with respect to the venue category. The mean was taken and each neighborhood were grouped together on the basis of venue. Café parameter had the most non-zero rows. Hence, café was selected for further analysis.



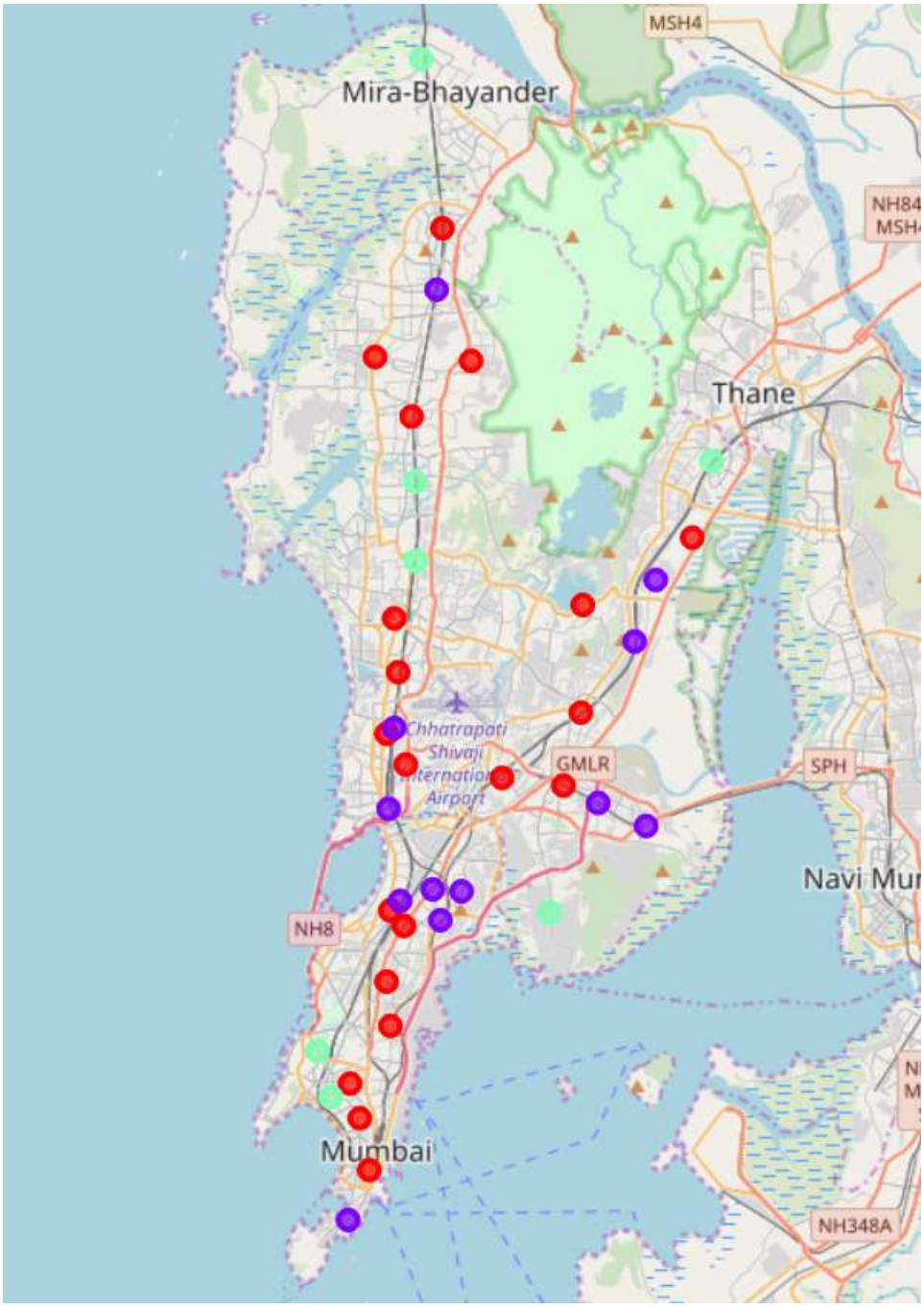
Map of Mumbai Neighborhoods

3.2 Machine Learning Algorithm:

Since the data doesn't produce any output or we are not predicting any values, it was identified that an unsupervised machine learning algorithm would be used. Clustering was done on the data. The K-means Clustering algorithm was selected for clustering the data. We choose the number of clusters. Then we select random points(centroids) that define the clusters. These points are not necessarily from the dataset. Each point is assigned closest centroid and K-clusters are formed. Distance can be Euclidean, etc. as per our requirement. The process is refined by calling it iteratively. New centroid is computed and placed for each cluster datapoints are reassigned. This process helps in refining the model. When there is no reassignment, the model is ready. This algorithm has a random initialization trap (If centroids are wrongly selected, some different cluster is obtained). A WCSS graph (elbow curve) is used to find the optimum number of clusters that should be made. Advantage: No interpretability, overfitting can easily occur, need to choose the number of trees. Disadvantage: Need to choose the number of clusters.

4. Result:

The clustering algorithm used k as three which means number of clusters formed were three. The clusters were formed based on the number of cafés present in the neighborhood. Clusters identified were one with high number of cafes, moderate number of cafes and a smaller number of cafés.



Map after Clustering

- The light blue points are the ones with least number of cafes present.
- The red points are the ones with moderate number of cafes present.
- The dark blue points are the ones with highest number of cafes present

5. Observation and Analysis:

The concentration of owners to build a new café should be in the areas where the number of cafés is least or moderate to get the least competition and earn maximum. Though some places like Trombay may not ne good to open cafés since it is an industrial area and apartments are scarce. Some moderately crowded places like Fort can also be used to open the café since large number of people travel through there everyday of the week. Even if there is some competition, the chance of earning a huge profit is very high.

6. Conclusion:

In the project, it was analyzed which part of Mumbai is best to open a café based in the number of cafes present in the surrounding. The data was obtained by using web scraping. K-Means Clustering algorithm was used effectively to obtain the result by classifying the data into three clusters. The three clusters showed the places as per the number of cafes present in the neighborhood analyzed.