The Battle of Neighborhoods – Buildings Location Predictor

Sunamya Gupta

December 23, 2019

1. INTRODUCTION

1.1 BACKGROUND :

With the increased growth in industrialization along with the fierce competition, it has become important for owners to attain maximum profit along with the customer satisfaction. Many people start their own businesses without analyzing the parameters that help a business to grow, eventually resulting into business failures. Some businesses become successful without even providing good services and some fail regardless of their quality of services. All this depends upon the various parameters that helps an owner to discover and analyze all the aspects before establishing an industry or growing his/her business.

1.2 PROBLEM :

With the rapid increase in customer demands and the fierce competitions among various brands, it has become very important to stay on the top. Not only services, but there are various other aspects that helps a business to become successful. One of the such aspect is“Location”. It is very important to build an infrastructure where people can easily access the resource.

1.3 INTEREST:

The aim of this project is to predict which location is suitable for owners to build the infrastructure. The owner will provide the city and the type of infrastructure he wants to build like restaurant, hospital, etc. Based on these parameters, the model will predict which location is suitable for building the infrastructure so that the owner could attain the maximum profit and customer satisfaction.

2. DATA ACQUISITION AND CLEANING

2.1 DATA SOURCES :

The data was collected from http://download.geonames.org/export/zip/IN.zip. From this source, I obtained the data regarding states, cities, provinces and postal codes of India. Not much data was available for the same, so I had to work accordingly. Maybe my results will not be that accurate but I tried my best to obtain the best data.

2.2 DATA CLEANING :

After the data was obtained, the main task was to clean the data so that it could be used at later stage for processing and predicting. The data contains lots of redundant data, so all the redundant data had to be removed. Attributes like country code, Admin codes, and accuracy values were to be dropped from the dataset as they were not relevant for our application.

The next task was to manage the missing values in the datasets. For many locations, the dataset didn’t contain the latitude and longitude values, so I had to google it and then replace their values with the correct data. For many provinces, the postal code data was also not available so I had to search for the accurate postal code for those provinces.

The third task was to sort the data according to our requirements. Merging the “Place Name” field depending upon the “Postal Codes” and “Provinces” of the city. After this, my next aim was to check for outliers, if any. Correct the data and missing values, if still any exists.

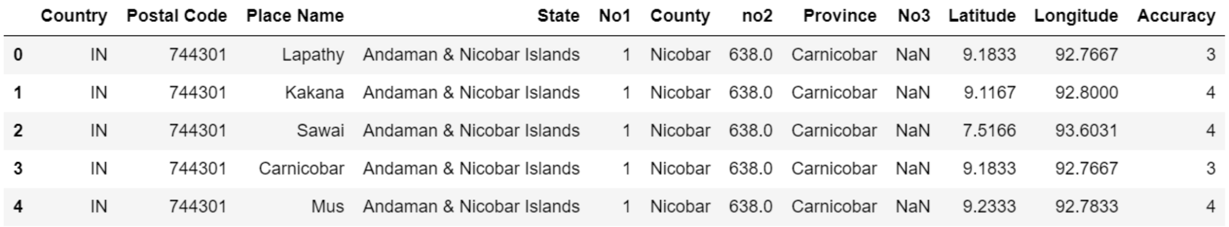
2.3 FEATURE SELECTION :

After the successful preprocessing of data, our aim is to now find those attributes which are going to contribute in the prediction and clustering. Now, our dataset has 8255 samples and 8 features. Our aim is to cluster the data depending on “City” and “Place Name”. These two features are going to help us to cluster the data and determine which location is better for owners to set up their infrastructure accordingly. Right now for us, “State” feature is not an important criteria during clustering process but it is going to help us at initial stage to extract the data related to a particular region. We also have to check which attribute depends on which other attributes. Like “Postal Code” depends on “City” or “Provinces”. So we`ve to choose the features accordingly.

So in the end, our main features are “City”, “Latitude”, “Longitude”, “Postal Code” and “Place Name”.

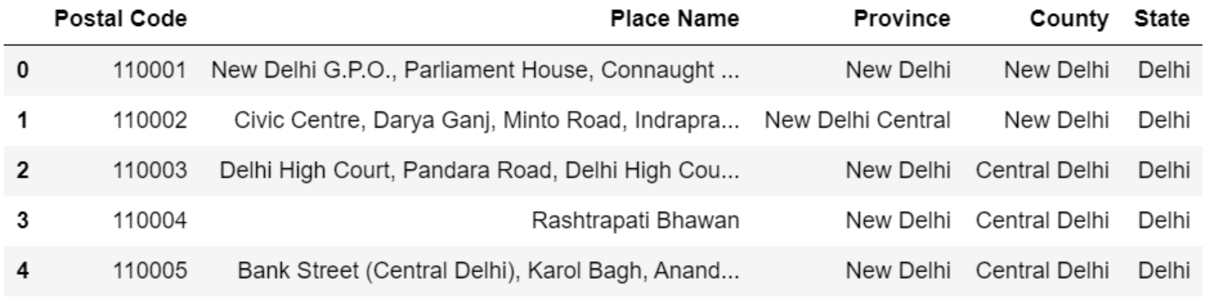
3. EXPLORATORY DATA ANALYSIS

After the collection of data from the source mentioned above, we will load the dataset into our model. The dataset looks like this:



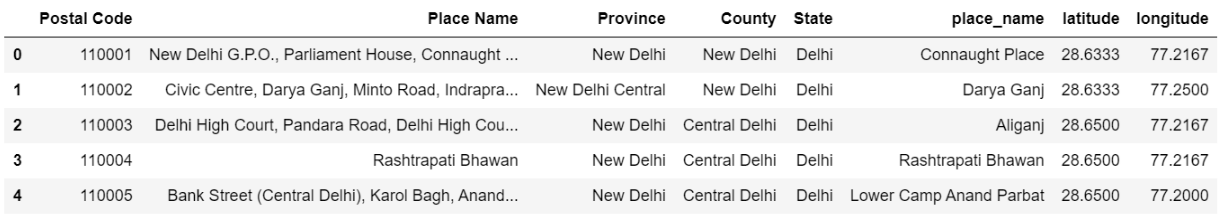
	Country	Postal Code	Place Name	State	No1	County	no2	Province	No3	Latitude	Longitude	Accuracy
0	IN	744301	Lapathy	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.1833	92.7667	3
1	IN	744301	Kakana	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.1167	92.8000	4
2	IN	744301	Sawai	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	7.5166	93.6031	4
3	IN	744301	Carnicobar	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.1833	92.7667	3
4	IN	744301	Mus	Andaman & Nicobar Islands	1	Nicobar	638.0	Carnicobar	NaN	9.2333	92.7833	4

Now we will remove all the unnecessary attributes like “Country”, “No1”, “no2”, “No3”, “Accuracy” as these aren`t helpful for building our model.
Our next aim is to merge all the tuples based on their “Postal Code” values as for many “Place Name”, we have similar “Postal Code” attribute value. After processing the dataset looks like:



	Postal Code	Place Name	Province	County	State
0	110001	New Delhi G.P.O., Parliament House, Connaught ...	New Delhi	New Delhi	Delhi
1	110002	Civic Centre, Darya Ganj, Minto Road, Indrapra...	New Delhi Central	New Delhi	Delhi
2	110003	Delhi High Court, Pandara Road, Delhi High Cou...	New Delhi	Central Delhi	Delhi
3	110004	Rashtrapati Bhawan	New Delhi	Central Delhi	Delhi
4	110005	Bank Street (Central Delhi), Karol Bagh, Anand...	New Delhi	Central Delhi	Delhi

Now we load our next dataset, which consist of Latitude and Longitude of the various Postal Codes of India. “Postal Code” is an attribute which is acting as a *Foreign-key* in our new dataset and as a *Primary-key* in our old dataset. Now we have to merge the two datasets, as follows:



	Postal Code	Place Name	Province	County	State	place_name	latitude	longitude
0	110001	New Delhi G.P.O., Parliament House, Connaught ...	New Delhi	New Delhi	Delhi	Connaught Place	28.6333	77.2167
1	110002	Civic Centre, Darya Ganj, Minto Road, Indrapra...	New Delhi Central	New Delhi	Delhi	Darya Ganj	28.6333	77.2500
2	110003	Delhi High Court, Pandara Road, Delhi High Cou...	New Delhi	Central Delhi	Delhi	Aliganj	28.6500	77.2167
3	110004	Rashtrapati Bhawan	New Delhi	Central Delhi	Delhi	Rashtrapati Bhawan	28.6500	77.2167
4	110005	Bank Street (Central Delhi), Karol Bagh, Anand...	New Delhi	Central Delhi	Delhi	Lower Camp Anand Parbat	28.6500	77.2000

4. CLUSTERING :

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance.

4.1 K-Means ALGORITHM:

K-means algorithm is an iterative algorithm that tries to partition the dataset into K - pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

4.1.1 How it works? :

The way k-means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

4.2 SOLUTION TO PROBLEM:

To solve this problem, we are using the K-Means Clustering model. Here we have collected data from the source as mentioned above. Our aim is to help a consumer in predicting and suggesting which location in that city is for which kind of infrastructure and business. Here, we will process the data so that it can be used to build our model. Now, we ask the user to enter the city where he/she wants to build his/her infrastructure. After that we will collect the latitude, longitude, and provinces of the city from the dataset. After collecting these data we will plot the data on the map. Our next task is to use ***FourSquare API*** to collect the business related data of all the provinces in that city. Data obtained from the API is *location, business type, popularity, etc*. Now we have to categorize the data obtained from the API. After that, we will split the data into various clusters depending upon the attributes like popular places and provinces and then we will find out which place is famous in particular province and at how much. Now our aim to cluster the data depending upon the popularity of the businesses and the provinces. In the end, we have to plot the map depending upon the clusters formed in the last step which will help user to locate the area which is suitable for them.

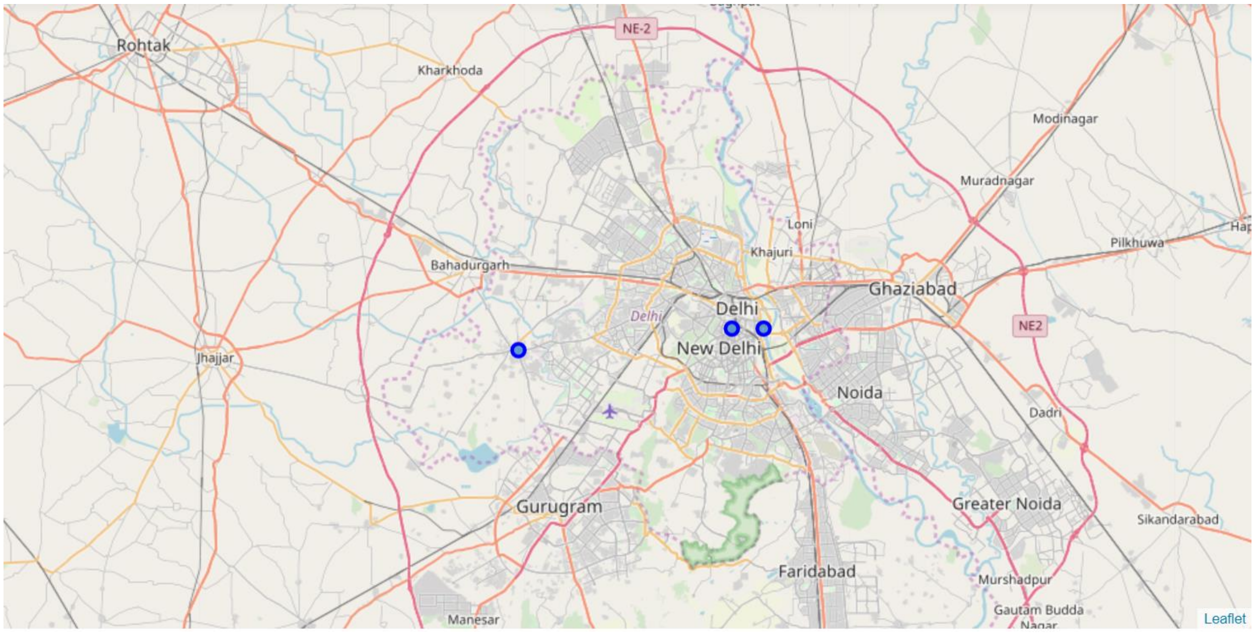
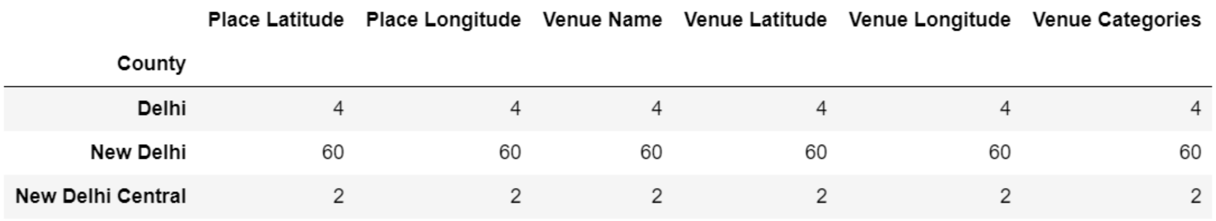


Fig 1: Displaying the sub-locations of New Delhi where the infrastructure can be built



	Place Latitude	Place Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Categories
County						
Delhi	4	4	4	4	4	4
New Delhi	60	60	60	60	60	60
New Delhi Central	2	2	2	2	2	2

Fig 2: Showing the count of already existing businesses

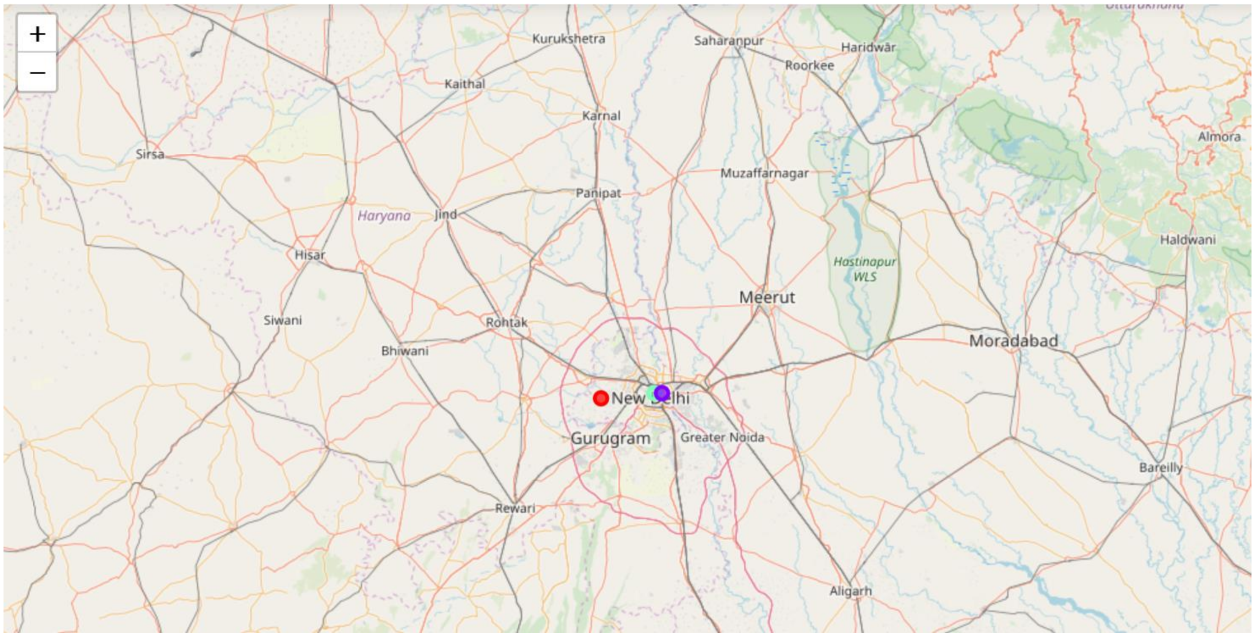
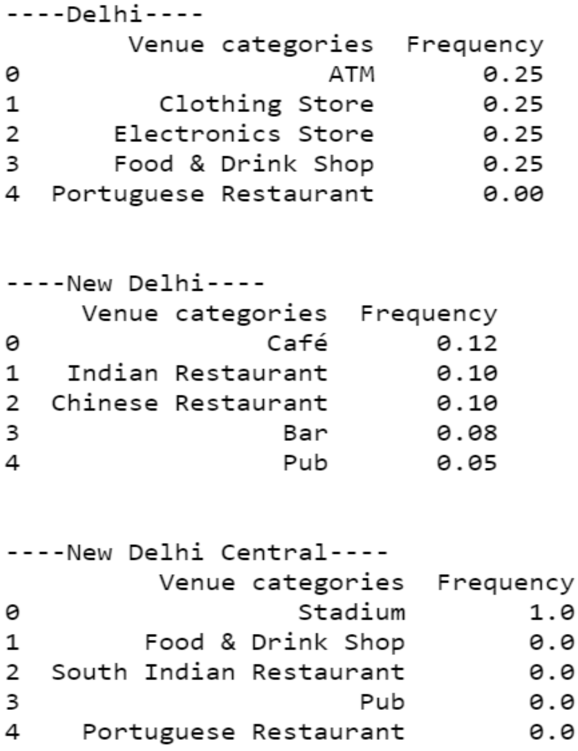


Fig 3: After the successful processing, these areas have been classified under various clusters



```
----Delhi----
          Venue categories  Frequency
0                      ATM       0.25
1           Clothing Store       0.25
2        Electronics Store       0.25
3        Food & Drink Shop       0.25
4   Portuguese Restaurant        0.00


----New Delhi----
       Venue categories  Frequency
0                  Café       0.12
1    Indian Restaurant       0.10
2   Chinese Restaurant       0.10
3                   Bar       0.08
4                   Pub       0.05


----New Delhi Central----
            Venue categories  Frequency
0                    Stadium        1.0
1          Food & Drink Shop        0.0
2    South Indian Restaurant        0.0
3                        Pub        0.0
4      Portuguese Restaurant        0.0
```

Fig 4: Showing the frequency of most popular businesses in an area from top to bottom

5. RESULTS :

We processed the data successfully and provided the user with the data as in where he/she can open or build his/her infrastructure or business. Results must not be accurate due to lack of data collection, but the result obtained from the processing was quite precise.

6. CONCLUSION :

In this study, we analyzed the location data and were able to cluster the sub-location depending upon the popularity of the business and were able to suggest the user as in where to build the infrastructure for maximum profit and customer satisfaction. I built the model using K-Means Clustering Algorithm which is an unsupervised learning method. This model will help businessmen in maximizing the profit and increase the customer satisfaction which in turn will yield the better result for the business. Due to lack of data available online, the result obtained aren`t that accurate but it the model is trained with good data sets, then results would have been more accurate and precise.