Applied Data Science Capstone Project – "The Battle of Neighborhoods"

Part 2a – Data Requirements

With some clarity regarding the questions to be answered via this project, it is evident that the data required would comprise a **listing of the major neighborhoods within the cities of New Delhi, Gurugram, and Noida**, along with their **geospatial location data (i.e., latitude and longitude coordinates)**. A reasonably comprehensive dataset for obtaining the aforementioned geospatial data is the **Zomato API dataset available on Kaggle (Link).**Thereafter, the geospatial location data of the neighborhoods can be utilized to explore different neighborhoods, i.e., retrieve data regarding the neighborhoods in geographical proximity to a particular neighborhood, analyze the most popular cuisines of eateries in different neighborhoods, etc., using **Foursquare API (Link).**

Part 2b – Data Understanding

In this project, the Zomato API dataset available on Kaggle in the comma separated value file **zomato.csv** will be used for obtaining the geospatial data (i.e., latitude and longitude coordinates) of the prominent neighborhoods of New Delhi, Gurugram, and Noida. In this dataset, each restaurant is uniquely identified by its *Restaurant Id*. Every restaurant is assigned the following parameters:

- *Restaurant Id*: Unique id of every restaurant across various cities of the world
- *Restaurant Name*: Name of the restaurant
- *Country Code*: Country in which restaurant is located
- *City*: City in which restaurant is located
- *Address*: Address of the restaurant
- *Locality*: Location in the city
- *Locality Verbose*: Detailed description of the locality
- *Longitude*: Longitude coordinate of the restaurant's location
- *Latitude*: Latitude coordinate of the restaurant's location
- *Cuisines*: Cuisines offered by the restaurant
- *Average Cost for two*: Cost for two people in different currencies
- *Currency*: Currency of the country
- *Has Table booking*: yes/no
- *Has Online delivery*: yes/ no
- *Is delivering*: yes/ no
- *Switch to order menu*: yes/no
- *Price range*: range of price of food
- *Aggregate Rating*: Average rating out of 5
- *Rating color*: depending upon the average rating color
- *Rating text*: text on the basis of rating of rating
- *Votes*: Number of ratings cast by people

After reading in the zomato.csv file into a pandas dataframe, the *City* column will be filtered to include only the rows of data wherein the *City* is either New Delhi, Gurgaon, or Noida. Thereafter, the filtered dataframe will be copied to a new pandas dataframe, consisting of only the data in the columns *Restaurant Name, City, Locality, Longitude,* and *Latitude*. The data in the columns *Locality, Longitude,* and *Latitude* will then be fed into the Foursquare API to explore different neighborhoods, i.e., extract data regarding the neighborhoods in the vicinity of a particular neighborhood (top ‘n’ venues within a radius of ‘m’ meters of a particular neighorhood) and analyze the most popular cuisines of eateries in different neighborhoods (using pandas ‘groupby’ function, and filtering the names of the common venues returned by eatery name, i.e., *restaurant, café, coffee shop, coffee shop*, etc.).

Part 2c – Analytic Approach

With some clarity regarding the data to be used for the project, the analytic approach that will be attempted to solve the business problem can be considered. Since the proximity of different neighborhoods to each other, or even the popular cuisines of restaurants or cafés within neighborhoods is preliminarily difficult to determine via basic exploratory data analysis using pandas, an unsupervised machine learning technique – specifically **K-means clustering** will be used to cluster the neighborhoods based on similar features such as the geographical distance between different neighborhoods, the predominant cuisine of the eateries in the neighborhoods, etc. Depending upon the insights derived via clustering, recommendations regarding Problem A can be made based on minimizing the competition and maximizing the footfall for the neighborhoods within the clusters, whereas Problem B can be addressed via the most common cuisines of the eateries within the clusters.