Kolkata - A Scope for Food Lovers and Businesses

Sayak Chakrabarti

April 16, 2020

2. Data Acquisition and Cleaning

2.1 Data Sources

Data has been obtained from two APIs, Foursquare and Zomato Developers. The centre point of Kolkata was selected vide its latitude-longitude and then a radius of 4 Kilometres was searched for finding out venues using Foursquare API. Thereafter, the latitude-longitude values were used to fetch more details on the venues using Zomato API

2.2 Data Cleaning

Data was downloaded using two API sources as discussed above and finally combined into a single table. The initial data called via Foursquare API comprised of several attributes and was filtered to consider just the venue name-category-latitude-longitude and thereafter, the second set of data was called using Zomato API which is an aggregator for restaurants and food joints, to match the earlier latitude-longitude values and present somewhat-similar data.

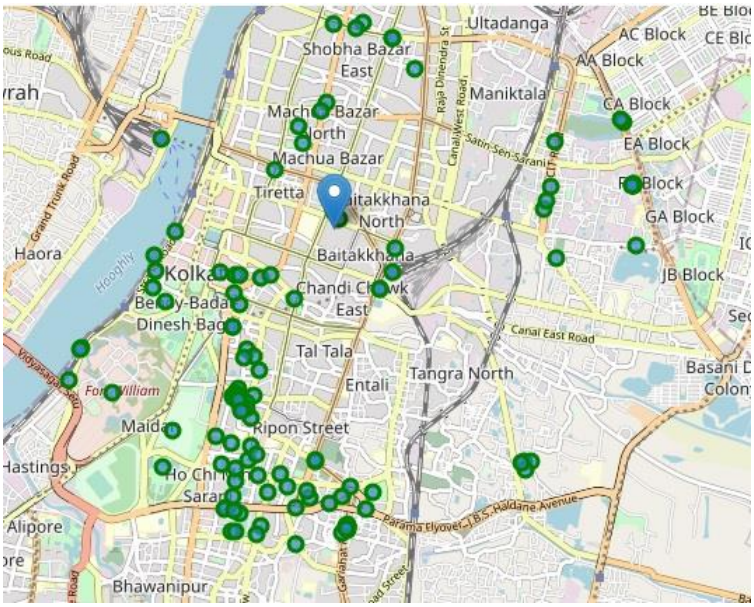

Fig1. List of venues as per Foursquare API

The second dataframe comprised of the venue name, latitude, longitude, price for two, price range, rating and address. Thereafter, both the dataframes were combined to one by matching latitude-longitude values.

2.3 Feature Selection

After data cleaning, a couple of erroneous events were observed and the same rectified sequentially

a. While most of the venues overlapped, some of them did not. Therefore, it was decided to drop all corresponding venues from the two datasets that had their latitude and longitude values different by more than 0.0004. This resulted in 59 rows of data.
b. Post this, it was found in the dataframe that
    i. There are venues like multiplexes which have restaurants inside them
    ii. Two locations are so close that they share the same latitude and longitude values
    iii. There were one or two duplicate data entries

    While it was ok to keep i, it did not make any sense to keep ii and iii and therefore, the corresponding rows were dropped from the dataframe.

c. A new column had to be introduced, called average price, which was an important attribute
d. Redundant columns like latitude, longitude, latitude and longitude differences, price for two were dropped. Name was also dropped as there was already a column called venue which sufficed. This resulted in 41 rows of data.
e. Ratings are an important feature of the dataset and therefore, rows with nil ratings were removed.

The final dataset therefore had 39 rows of data and 8 columns.