Data Description

Restaurant Location Analysis in Delhi-NCR

Divyanshi Paliwal

September 30, 2019

Data Acquisition

The data to be used in the project has been taken from one of the publically available Zomato Dataset on Kaggle. Zomato is a restaurant finder and online booking application operational in India. However, the datasets are not exhaustive and misses many restaurants. Nevertheless, has enough data to be analysed (close to 5000 rows for Delhi, Noida and Gurgaon combined).

Data Cleaning

Data downloaded has records of restaurants throughout India. Since the study focuses only on Delhi-NCR, the data needs to be filtered. For simplicity, in Delhi-NCR I have considered New Delhi, Noida and Gurgaon only. According to the Indian Government other cities like Sonepat, Greater Noida, Indirapuram etc. also belong to Delhi-NCR, but with ground knowledge it can be easily said that New Delhi, Noida and Gurgaon have more restaurants, thus it was apt to consider only these three cities as a part of Delhi-NCR.

The dataset contains comma separated locality or locations. By looking at the data, it could be seen that the value after the comma was the main locality and before the comma was not needed, hence the column ‘locality’ needs to be modified accordingly. Also, the names are not consistent, for example – ‘Greater Kailash 1 (GK)’ and ‘Greater Kailash (GK) 1’, they both address to the same locality and hence need to be corrected. The column ‘highlights’ contains a set of all the features provided by the restaurant, though these features are not needed as such, but their count is important to know if more number of features attract more customers. Thus, the column will be modified to contain the number of features each restaurant offers. Another column, ‘establishments’ contains values in square brackets, therefore the brackets need to be removed as they are not adding to the data.

It is also observed that some of the restaurants have not specified their ‘establishment’ (the type of restaurant). These rows will not be removed as they are needed for other analysis, however, the empty value will be replaced with a hyphen.

Also, to get latitude and longitude data Foursqaure API will be used and the result will be appended to the dataset.

Parameter Selection

The dataset contains columns, which are not needed as they are not adding to the data and hence, they need to be removed. The table below shows what parameters will be retained and which parameters will be removed.

Table 1: Parameters to Retain

Parameter	*Reason to retain*
Name	It should be kept for identification only, though it does not add to the analysis.
Establishment	This is required to identify type of restaurant that may work in a locality.
City	For identification if a locality belongs to New Delhi, Noida or Gurgaon
Locality	This is required for identification of a suitable area to open a restaurant
Address	This is required for getting Latitude and Longitude data
Cuisines	This is required to identify what cuisines are popular
Average_cost_for_two	This is required to analyse suitable price range
Highlights	This column shows the features provided and was needed for analysis
Aggregate_rating	This is needed to identify highly rated localities and other impacts
Votes	This column is synonymous to number to people visited, which is needed for analysis

Table 2: Parameters to Remove

Parameter	*Reason to remove*
Res_id	This parameter has no value in the analysis
url	This parameter has no value in the analysis
City_id	City is enough to distinguish between places and city_id is redundant
Zipcode	This parameter has no value in the analysis
Country_id	This parameter has no value in the analysis
Locality_verbose	This is a redundant column

Timings	This parameter has no value in the analysis
Currency	This parameter has no value in the analysis
Rating_text	This is a redundant column
Photo_count	This parameter has no value in the analysis
Opentable_support	This is a redundant column
Delivery	This is a redundant column
Takeaway	This is a redundant column