Capstone Project: Exploring Restaurants in Britain

1. Introduction

Britain includes an interesting diversity of cultures of multiple nationalities and ethnicities. The goal of this project is to explore that diversity with respect to the restaurants and cuisines that exist in Britain. Further, a part of the analysis is focused on London, as the most populous city of Britain.

Here are the list of objectives/questions to be explored:

1) Geo-Exploration of Restaurants Clusters in Britain.
2) Cities of Top-Rated Restaurants in Britain.
3) Where are the Top-Rated Restaurants in London?
4) Where are the Lowest Rated Restaurants in London?
5) Where are the best Indian restaurants in London?

2. Data Description

The dataset was part of a Kaggle competition:

https://www.kaggle.com/shrutimehta/zomato-restaurants-data

The collected data has been stored in the Comma Separated Value file, Zomato.csv. Each restaurant in the dataset is uniquely identified by its Restaurant Id. Every Restaurant contains the following variables:

• 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 casted by people

3. Methodology

3.1 K-Means Clustering

The partitional clustering approach was applied in this project. Partitioning methods relocate data points by moving them from one cluster to another starting from an initial, usually random, composition. The basic idea is to find a clustering structure that minimises a certain error criterion measuring the distance (e.g. Euclidian distance) of each instance to its representative value. The K-Means algorithm is a typical example of the partitional clustering approach, which was used in this study.

Due to its conceptual simplicity, the K-Means algorithm has been widely used for clustering tasks. As its name suggests, the algorithm partitions data into a number of clusters (K) represented by their centroids. The centroid of each cluster is calculated as the mean of all instances belonging to that cluster. The Sum of Squared Error (SSE) is used to measure the total squared distance from points to their centroids, as in the equation below. The SSE may be globally optimised by exhaustively enumerating all partitions, or by giving an approximate solution based on heuristics. The algorithm converges to a solution when meeting at least one of these conditions: i) The cluster assignments no longer change, or ii) The specified number of iterations is complete.

$$J(C_k) = \sum_{X_i \in C_K} \|X_i - \mu_K\|^2$$

Where μ_K is the mean of cluster C_k, and $J(C_k)$ is the squared error between $\mu_K$and the points in C_k [1].

3.2 Exploratory Visualizations

Visualization was ideally described as the transformation of the symbolic into the geometric [2]. Various benefits can be attributed to data Visualization. In contrast to text-based means, the interpretation of visual formats happens immediately in a pre-attentive manner. Further, the pictorial representation of data can help answer or discover questions. The usefulness of exploratory data analysis using visual techniques was early introduced in John Tukey's landmark textbook Exploratory Data Analysis [3].

In this respect, the project applied a set of exploratory visualization to answer the question provided in the introduction. Examples includes geographic visualization, and bar chart plots. The next section presents the visualizations produced. The visualizations were implemented using the Python libraires of Folium [4] and Matplotlib [5].

4. Results

Figure 1: K-Means-based clusters of restaurants based on geographic locations (K=4).

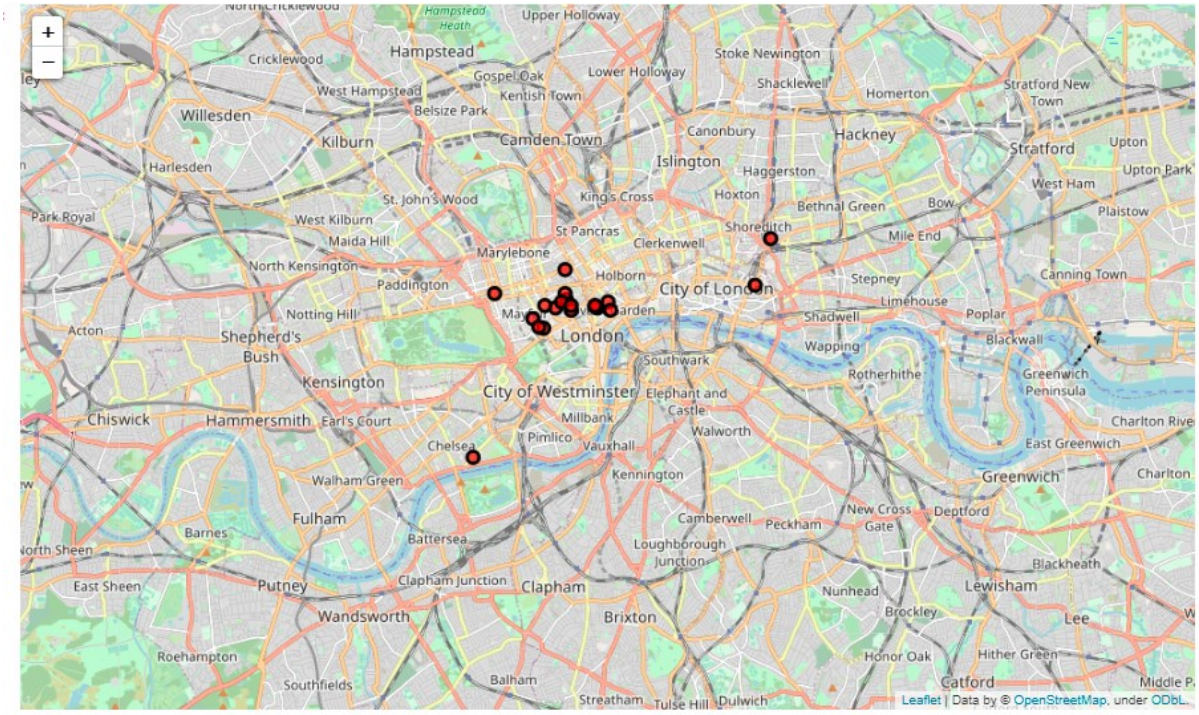

Figure 2: A zoomed-in view of the cluster of restaurants in London.

Figure 3: Cities of Top-Rated Restaurants in Britain.

Figure 4: Locations of top-rated restaurants in London.

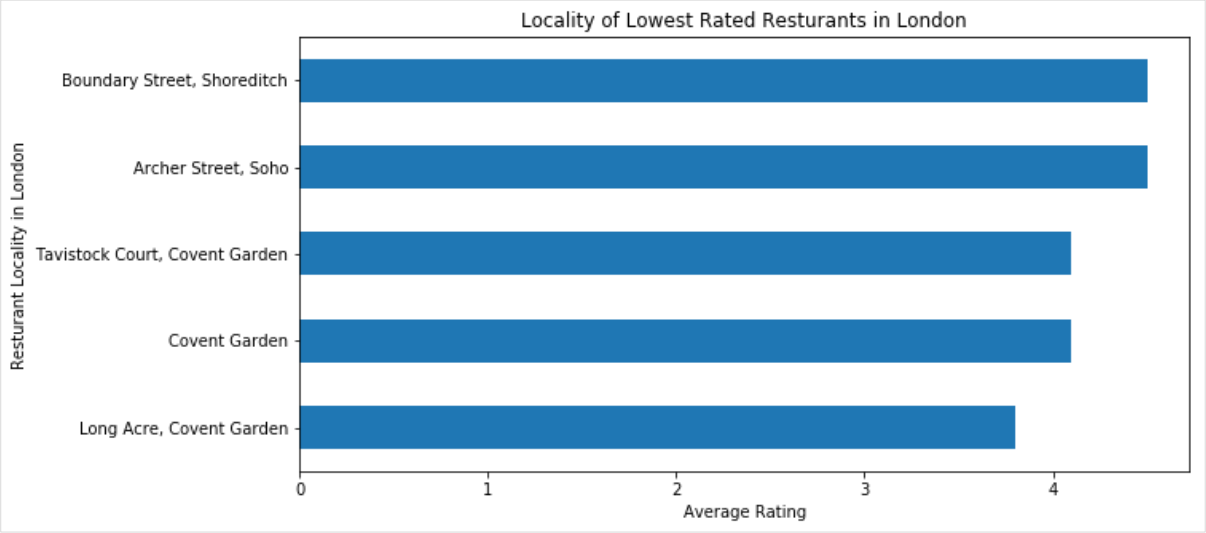

Figure 5: Locations of lowest rated restaurants in London.

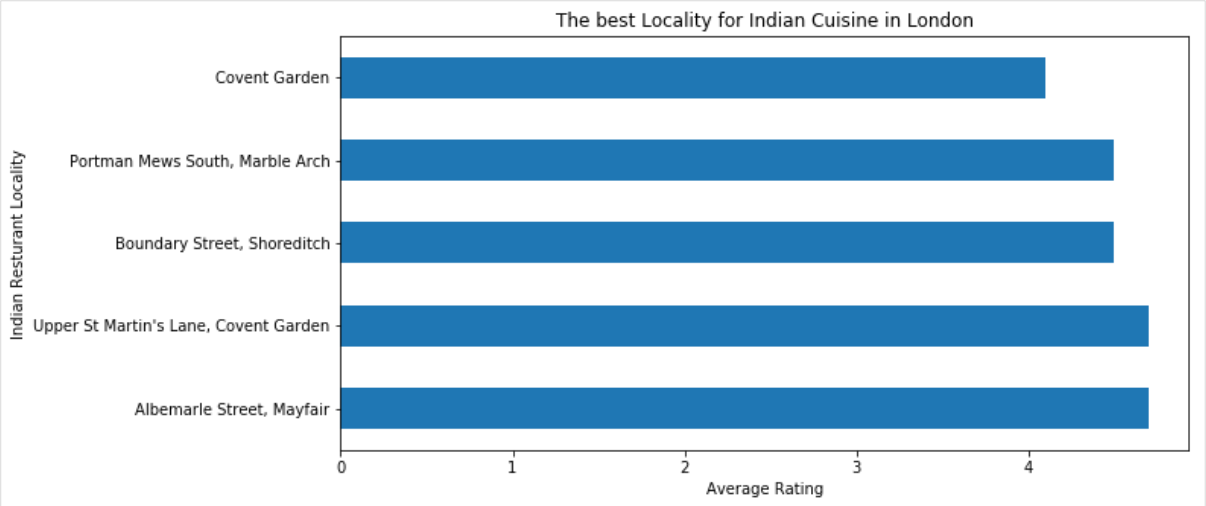

Figure 6: Locations of best Indian restaurants in London.

References

[1] Jain, A.K. Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 2010, 31(8):651-666.

[2] McCormick, Bruce Howard, Thomas A. DeFanti, and Maxine D. Brown. "Visualization in scientific computing." IEEE Computer Graphics and Applications 7, no. 10 (1987): 69-69.

[3] Tukey, John W. "Exploratory data analysis." (1977): 2-3.

[4] https://python-visualization.github.io/folium/

[5] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. IEEE Annals of the History of Computing, 9(03), 90-95.