Similarity of neighbourhoods in Bangalore

# Introduction

This analysis has been done by exploring various venues like café shop, food court, hospitals, bus depot, and other public outlets, in city of Bangalore, India, to understand their local areas of interest.

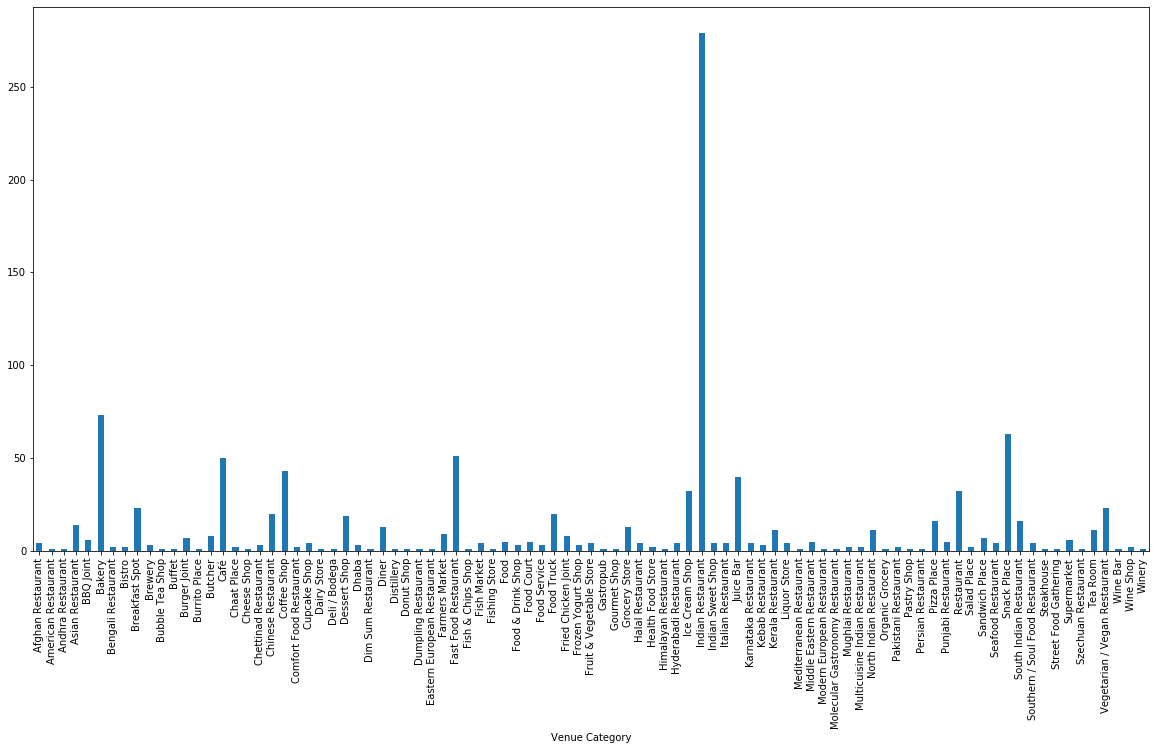
The vast dataset generated using FourSquare API has been clustered into various categories to simplify the study and then plot on a map to have a clear insight on the analysis.

This report summarizes the similarity of neighbourhood calculated on basis of venue within food category but approach can be extended for any available venue category.

# Data acquisition and cleaning

The dataset collected on the various areas of city have been further bucketed according to the postal code along with its latitude and longitude values to enable a connection with data collected from FourSquare API using latitude and longitude.

The study is limited to “food” category from the intersected dataset. The filtered dataset is then plotted on a bar graph. By analysis of the frequency of each food category, top 9 food categories were selected to obtain the basic dataframe for this study.



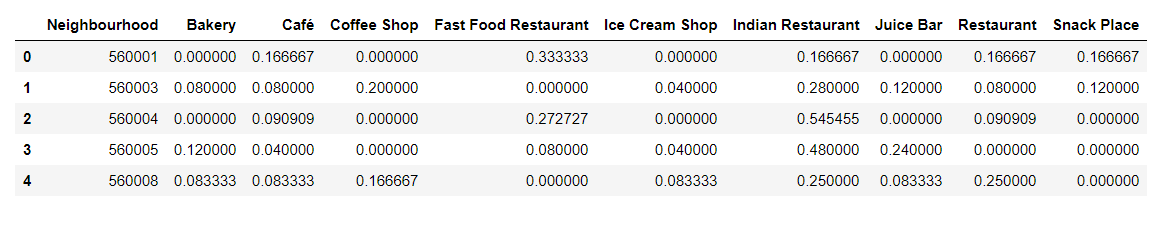
Distribution of food categories across all neighbourhoods

# Data Analysis

To analyze the data, a dataframe with “Neighbourhood” and “Venue Category” is created. The “Venue Category” contains only top 9 food categories identified earlier.

Subsequently, category information is converted using dummies command. All postal codes are displayed row-wise. Each row depicting one neighbourhood with their corresponding dummy (0/1) against the listed top 9 categories column-wise. Dummy “1” represents the presence and dummy “0” represents the absence of a category at a given postal code.

Within each neighbourhood, local mean of each category is calculated. This new dataframe represents weight distribution matrix of each category against respective postal code. This formulates the feature dataset for further segregation to various clusters depending on their similarities.



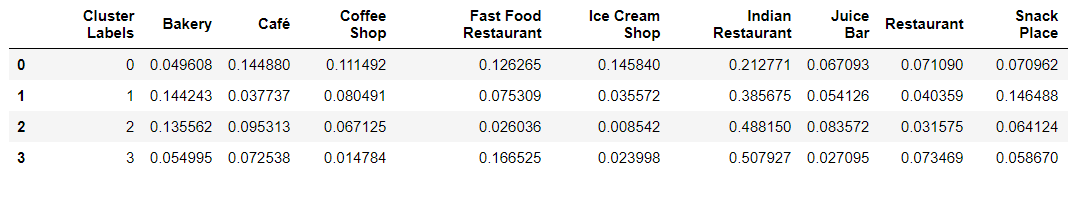
Feature dataset of each neigbourhood with top 9 food categories

# Clustering of data

With the feature set of each postal code, k-means clustering is performed to determine the similarity across neighbourhoods. To identify best homogeneity within clusters, the number of clusters are varied between 4 to 8. The cluster label identified is added to dataframe with postal code and latitude and longitude details. With this dataframe, multiple rows represent same cluster label. To have best representation of each cluster, local mean of each category against cluster is performed.

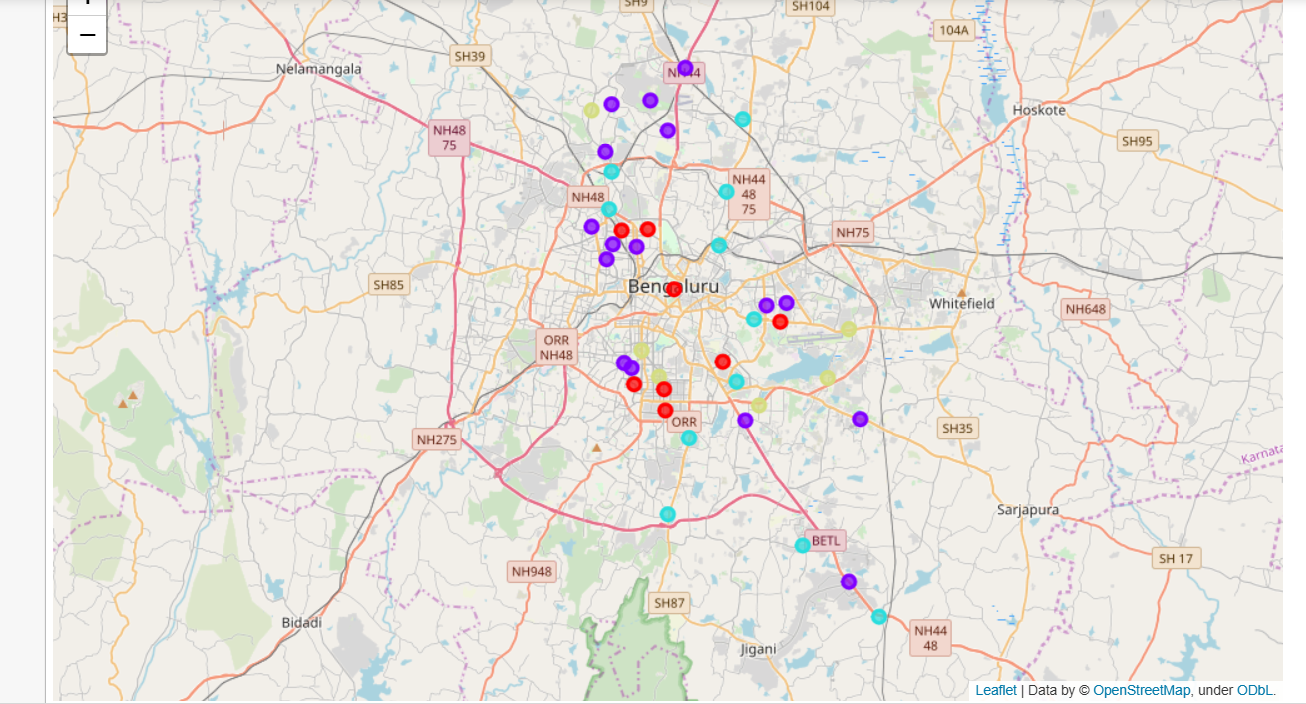
# Discussions

The dataframe with cluster labels and features (food category) depicts the frequency distribution of top 9 food categories within each cluster. With reference to cluster labels, similar neighbourhoods are identified. Also, feature set of each cluster indicates the presence of various food category venues in those neighbourhoods. For example, neighbourhoods in Cluster 1 has “Indian Restaurant” (38%) along with “Bakery” (13%) whereas Cluster 3 has “Indian Restaurant” (50%) with “Fast food Restaurant” (16%).



Cluster labels with average feature set

For quick insight of neighbourhood similarity, each cluster is assigned a unique color. This color is used to create markers for neigbourhood on a map.



Neighbourhood similarity basis K-means clustering