# Coursera - IBM Data Science Capstone Project

The Battle of Neighborhoods

-Karthik Srinivasan

# 1. INTRODUCTION

- **BACKGROUND** : Key objective in opening a new shopping mall is the location and large crowd pullers to provide their products and services to people.

- **PROBLEM** : Project aim to select a high potential location for the opening a new Shopping mall in the neighborhood of Chennai City and also explore the neighborhood in terms of venue category and visualize the view of city and cluster the neighborhood with help of K – means Clustering Algorithm

- **TARGET** : Proprietor and Retail Chain Market holders who are considering to open a new shopping mall in the neighborhoods of Chennai City and explore the common venues of each neighborhood.

# 2. Data Acquisition and Cleaning

**Data Acquisition** : The Data acquired for this project is combination data from three major sources.

- **Wikipedia** Page for Chennai Neighborhoods.
- **Foursquare API**.
- **Geocoder Package** for Latitudes and Longitudes.

**Data Cleaning** :

Web scraping of data from Wikipedia Page of Chennai Neighborhoods and converting into dataframe.

Using Geocoder Package , carving out latitude and longitude of Chennai neighborhoods.

Merging both data into a single dataframe.

# 3. METHODOLOGY

**Foursquare API :** Using Foursquare API , we can carve out venue details of each neighborhood along with different categories in the form JSON file.

**Folium Package :**Folium builds on the data wrangling strengths of the Python ecosystem . With the help of this package we can get the broad perspective view in the form of map with help of coordinates each neighborhood are correctly plotted in the Chennai city region.

**Optimal Number of Clusters :** The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. Based on the Silhouette Score of various clusters, the optimal cluster size is determined.

# 3. METHODOLOGY

**K – Means Clustering :** It is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will cluster the neighborhoods based on the optimal score of clusters and their frequency of occurrence for shopping mall.

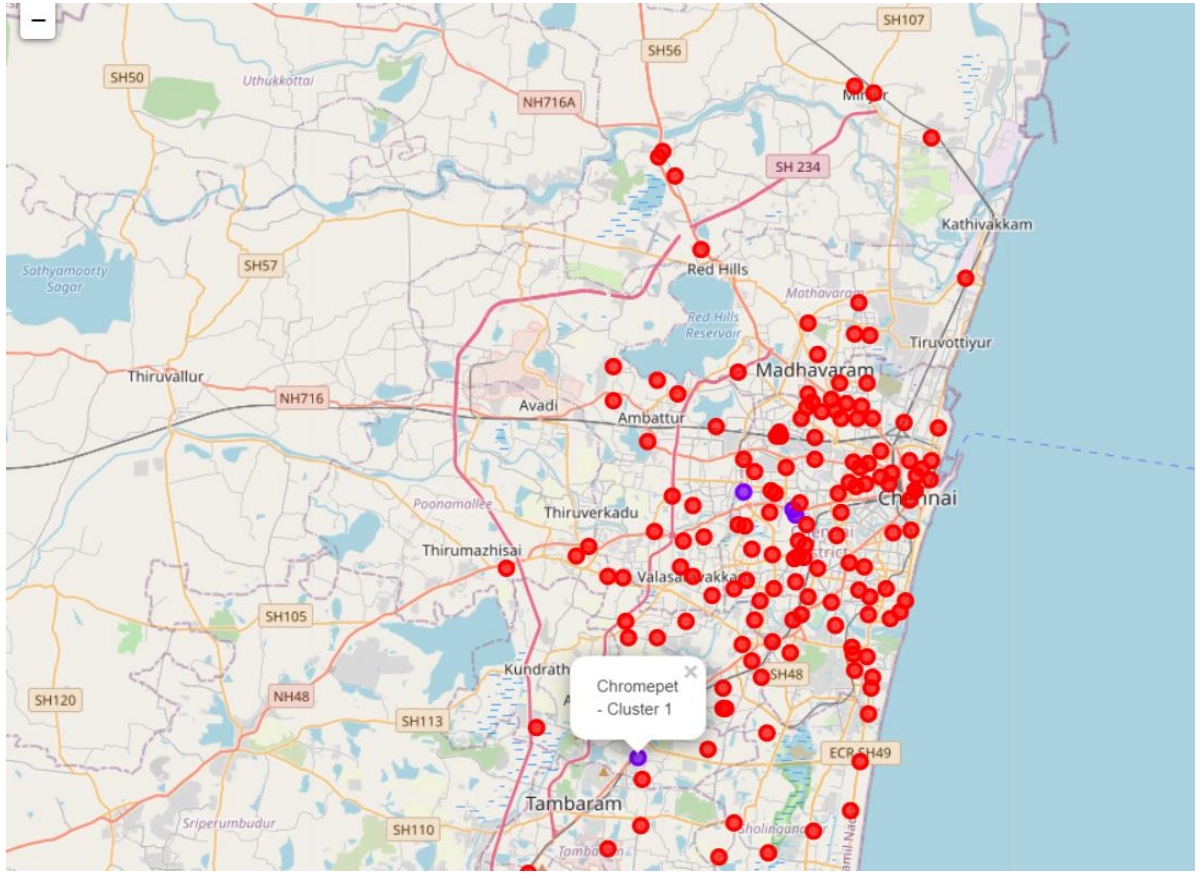
The Clustered neighborhoods are visualized using different colors so as to make them distinguish each other. The results will allow us to identify the neighborhoods have higher concentration of shopping malls to fewer ones.

# 4. RESULTS

The results from the K means clustering show that we can distinguish the neighborhoods into 3 clusters based on the frequency of occurrence for venue category “Shopping Mall”

- **Cluster 0**: Neighborhoods with no existence of Shopping Malls.
- **Cluster 1**: Neighborhoods with low number of Shopping Malls.
- **Cluster 2**: Neighborhoods 2 or more number of Shopping Malls.

# 4. RESULTS




Red points is Cluster0

Blue points is Cluster1

Green points is Cluster2

# 5. DISCUSSIONS

- The aim of the project is to help the proprietor to find a location to set up a Shopping mall in the neighborhoods of Chennai.
- From results, we can carve out that in Cluster 0 represents no existence of Shopping mall.
- Cluster 1which represents a great opportunity and high potential areas to open a shopping mall because it falls under the central location of Chennai city.
- From the findings, Cluster 2 is not advisable for the proprietor to open a new Shopping Mall.

# 6.CONCLUSION

- This project helps a proprietor to get a better understanding of the neighborhoods with respect to frequency of occurrence of shopping malls.
- Findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations to open a new shopping mall.
- Future outcome of this project will be focusing on the most common venues in each neighborhood of Chennai City and filtering out many factors for a perfect living.