Battle of Neighborhoods

- Prashant Hegde (Data Science Final Project Submission)

Introduction

- We need to be able to get an idea of similarity in neighborhoods in different cities and if we can combine neighborhoods from different cities.
- This scenario helps anyone who is looking to move to a new city that is like the present city they live in. Or anyone who want to get an idea of a new city they want to move to.
- To find out how similar or dissimilar the new city is compared to the present city they are residing currently.

Data

- Data is obtained mostly from websites
- Bangalore data is got from http://www.onlinebangalore.com/guide/pincodes/pincode.html
- Mumbai data is fetched from https://mumbai7.com/postal-codes-in-mumbai/
- Data is cleaned up to some extent, duplicates removed, spaces within zip codes removed, columns names changed to align to string values.
- Cleaned data contains City, area, zip code

Methodology and Models

- First, the Data is understood using Exploratory Data Analysis.
- Done using folium mapping with latitude and longitude coordinates of zip code with venues returned from FourSquare API.
- Data transformation is done using onehot encoding and finding average of frequency of venues among the list of top 100 venues from each zip code.
- This data is passed to K-means clustering model.
- Optimum K is found using elbow method and accurately calculated using math functions.
- Optimum K is used to fit the K-means model and clustered data results are obtained

Conclusions

- Clustering is done once on Bangalore data once on Mumbai data and once on combining Bangalore and Mumbai Data
- Bangalore data shows that the coordinates I chose(live) is among a popular cluster which confirmed my existing knowledge of my surroundings. Mumbai clustering shows that there is a cluster that has more data points than rest clusters combined.
- Combined clustering further confirms the areas that fall in the same cluster as the Bangalore cluster I chose and these are the areas I can consider to be similar.

Future Directions

- It would be good to try this process with Google Places rather than FourSquare API. Google places for one would have much more data and data diversity than FourSquare since google android users easly outnumbers those that use iphone in India especially.
- FourSquare also returns only 100 venues for each call of zip code even after increasing radius and limit.