Coursera Capstone Project

Model to predict standard of living in a neighbourhood

Predicting the standard of living will help new dwellers in a city

- There is a lot of movement of people across cities owing to variety of reasons.
  - Job
  - New livelihood
  - Education etc
- People belong to different strata in the society and would like to know which part of the society would suit them well.
- Hence, there is a need for a tool that can guide the newcomers to the city about the varied kinds of neighbourhoods and suggest a suitable one under the following.
  - Expensive b. Normal c.Cheap

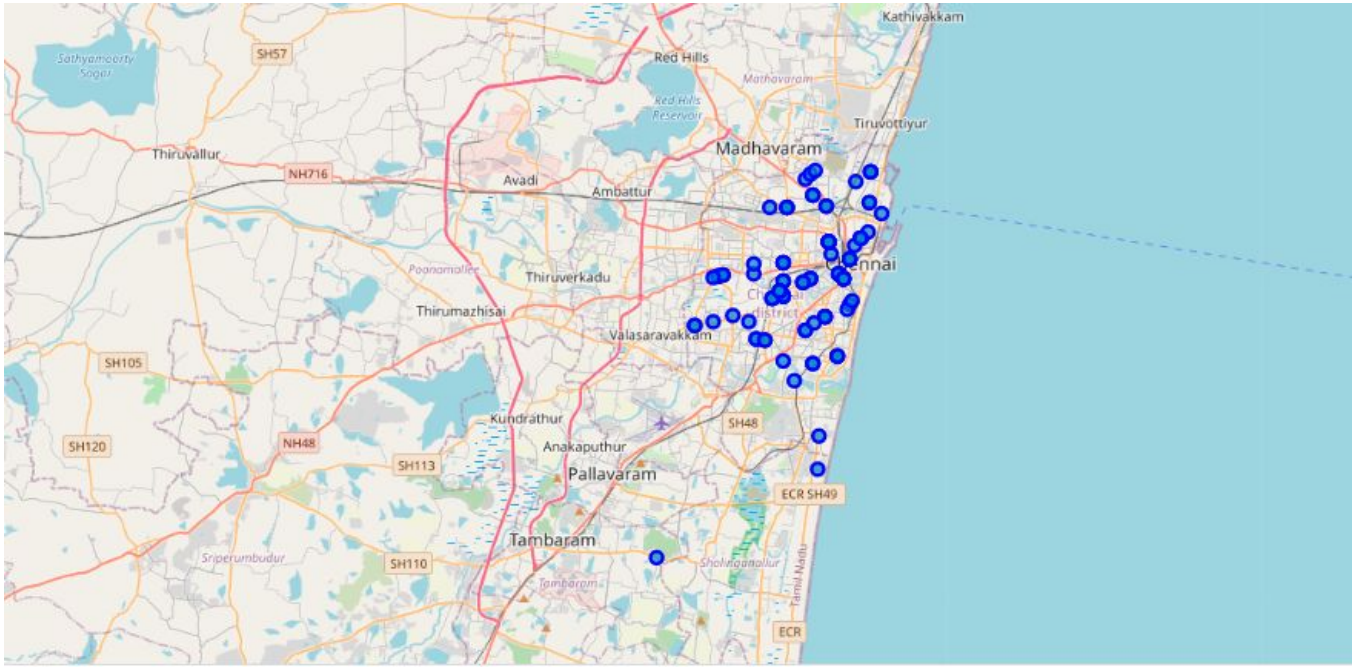
Data Requirements

Following data sources were used.

1. Pincode wise mapping of city, province, state, Latitude, Longitude from the GeoNames Postal Code files.
2. Venues, Venues Category from Foursquare API

Mapping the Localities in Chennai on a Map

This exercise was done to understand if there is sparseness in the distribution of localities.



Setting a hypothesis

Basis empirical knowledge below are some of the venue categories that could define the 3 kinds of localities.

1. Expensive - 'Shopping Mall', 'Japanese Restaurant', 'Russian Restaurant', 'Mexican Restaurant', 'Gym / Fitness Center', 'Spa', 'Pet Store','Pub', 'Yoga Studio'
2. Normal - 'Indian Restaurant', 'Furniture / Home Store', 'Department Store', 'Ice Cream Shop','Bookstore','Clothing Store','Flower Shop'
3. Cheap - 'Market', 'Train Station', 'Restaurant', 'Pier', 'Movie Theater', 'Bakery', 'Beach', 'Flea Market'

We will use the above definition to classify the clusters once the output is available

Choosing the right modelling method

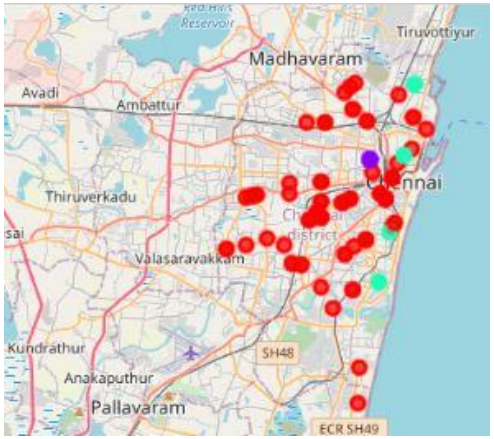
- The exercise at hand was to segregate the localities into 3 kinds.
- With no prior information available to us on the kind of localities, the scope of the problem statement falls under ‘Unsupervised Learning’.
- Clustering seems to be the obvious choice for the situation.
- Two types of clustering will be performed.
  - K-Means
  - Hierarchical

K- Means Clustering- Number of Clusters

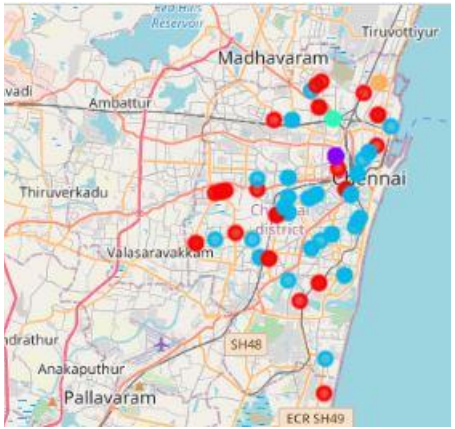
- Started with 3 clusters and iterated till 10 clusters to look at the clustering profile.
- Given that it is unsupervised, the decision to arrive at the right cluster was done using the grouping of localities.
- Given that clustering is done on the kind of venues around localities, the clusters had to be geographically close to one another.

K Means Clustering - Output

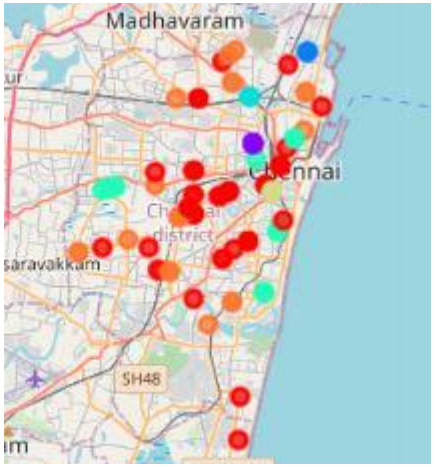
K =3



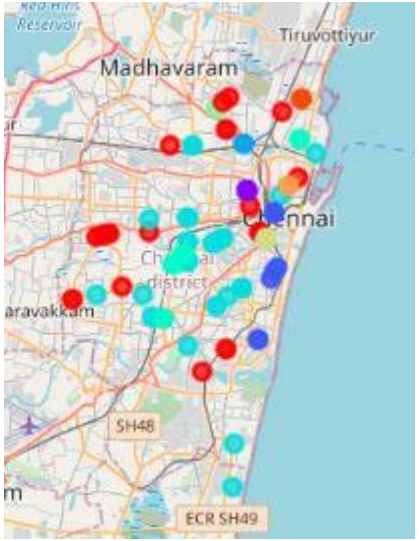
K =5



K =7

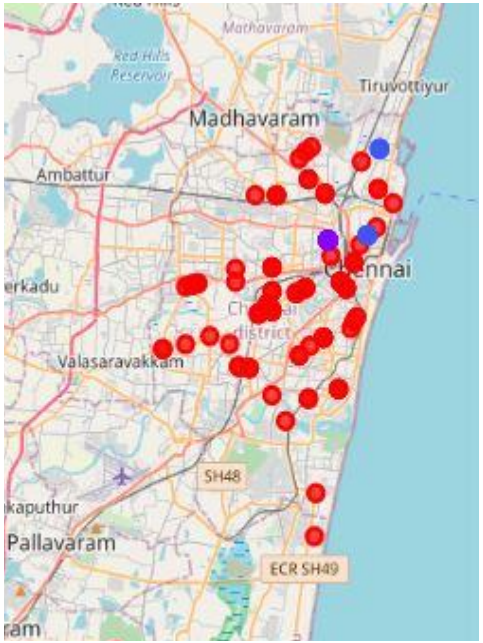


K =10

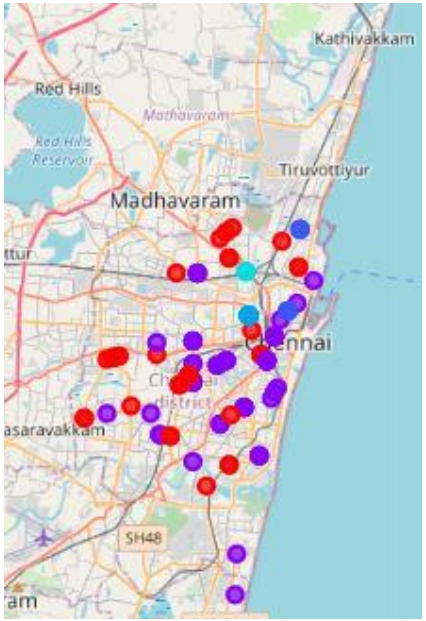


Hierarchical Clustering - Output

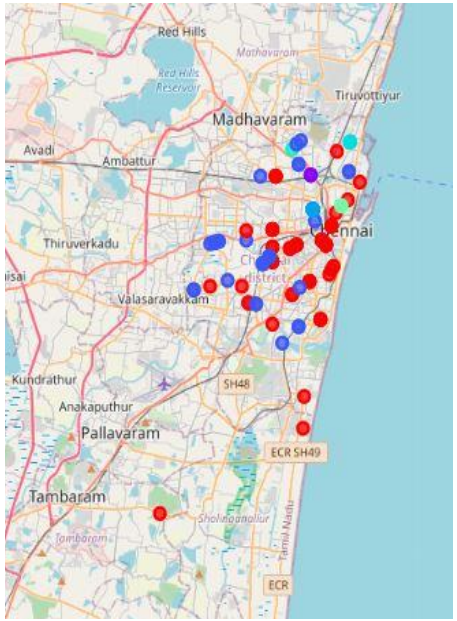
L=3



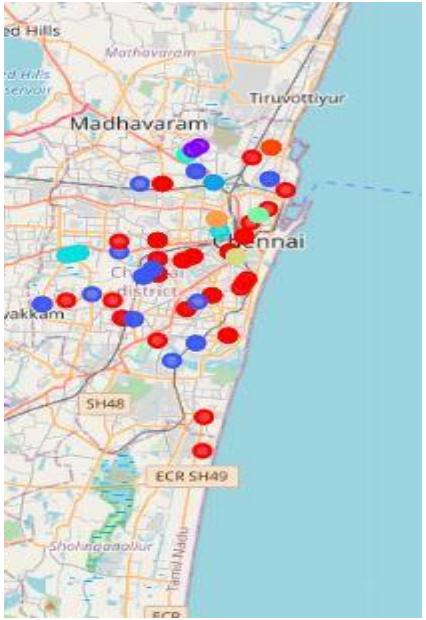
L =5



L=7



L=10



Results

Basis the clustering output, below are how the clusters pan out

K Means

- Expensive - Cluster 1
- Normal - Cluster 6,
- Cheap - Cluster 4
- Expensive / Normal - Cluster 0.

The other clusters are relatively small and not worth exploring.

Hierarchical Clustering

- Expensive - Cluster 3, Cluster 2
- Normal - Cluster 0
- Cheap - Cluster 1

Conclusion & Next Steps

- The areas that are generally known to expensive, normal & cheap basis empirical data have fitted into the right clusters.
- Model has done a fair approximation of different kind of neighbourhood using venues around them
- Next Steps would be to add additional variables like availability of schools, colleges, hospitals etc around the localities.
- This would help crystallize the localities much further and hence take a better call.