Clustering Bangalore Neighborhoods

Applied Data Science Capstone

Introduction

Background

- Bangalore (Bengaluru) is a bustling metropolis of India
- large tech industry that creates thousands of jobs a year
- huge influx in population

Problem

- Every person that moves in is faced with the question of where to stay
- One major factor in determining that is what facilities and attractions a neighborhood offers

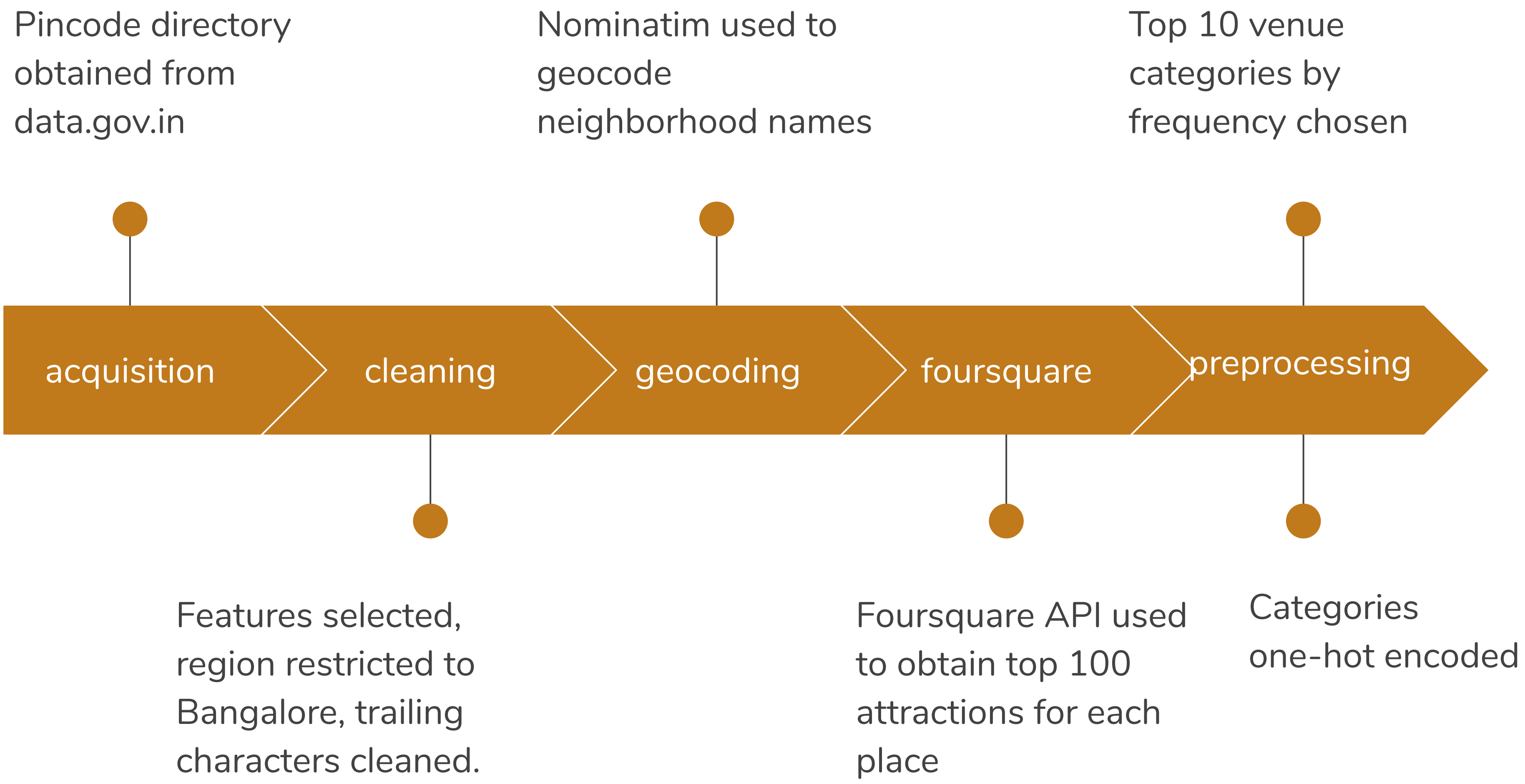
- When business owners determine the location of their store/business they need to look into what other attractions offer.
- Lack of similar businesses may offer a competitive advantage but might indicate a lack of consumer interest.

Interest

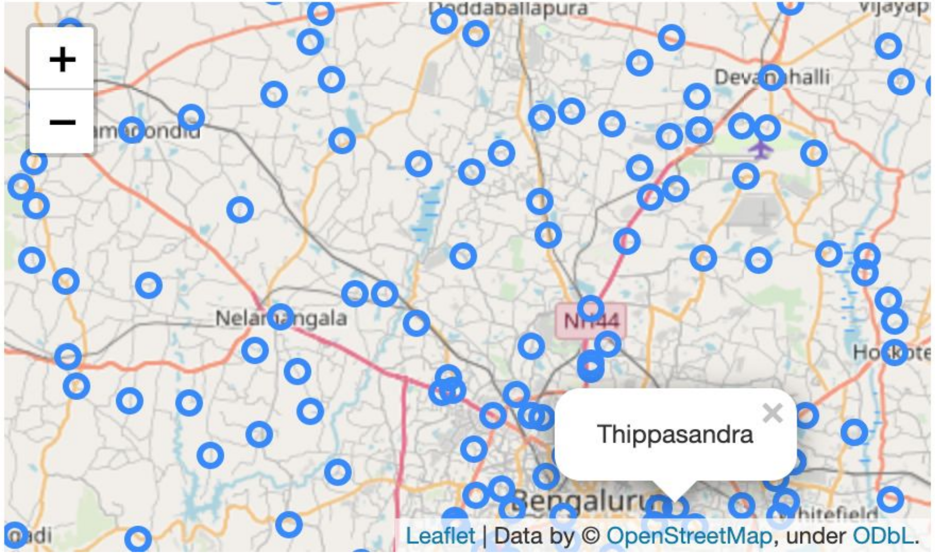
Neighborhoods analysis is can be an important tool for businesses in determining where to open their shop. It can also help people moving into a new area find similar neighborhoods that they can choose from

We will be clustering the different neighborhoods in Bangalore based on their 10 most popular attractions.

Data


Pincode directory obtained from data.gov.in
Nominatim used to geocode neighborhood names
Top 10 venue categories by frequency chosen
acquisition
cleaning
geocoding
foursquare
preprocessing
Features selected, region restricted to Bangalore, trailing characters cleaned.
Foursquare API used to obtain top 100 attractions for each place
Categories one-hot encoded

Analysis



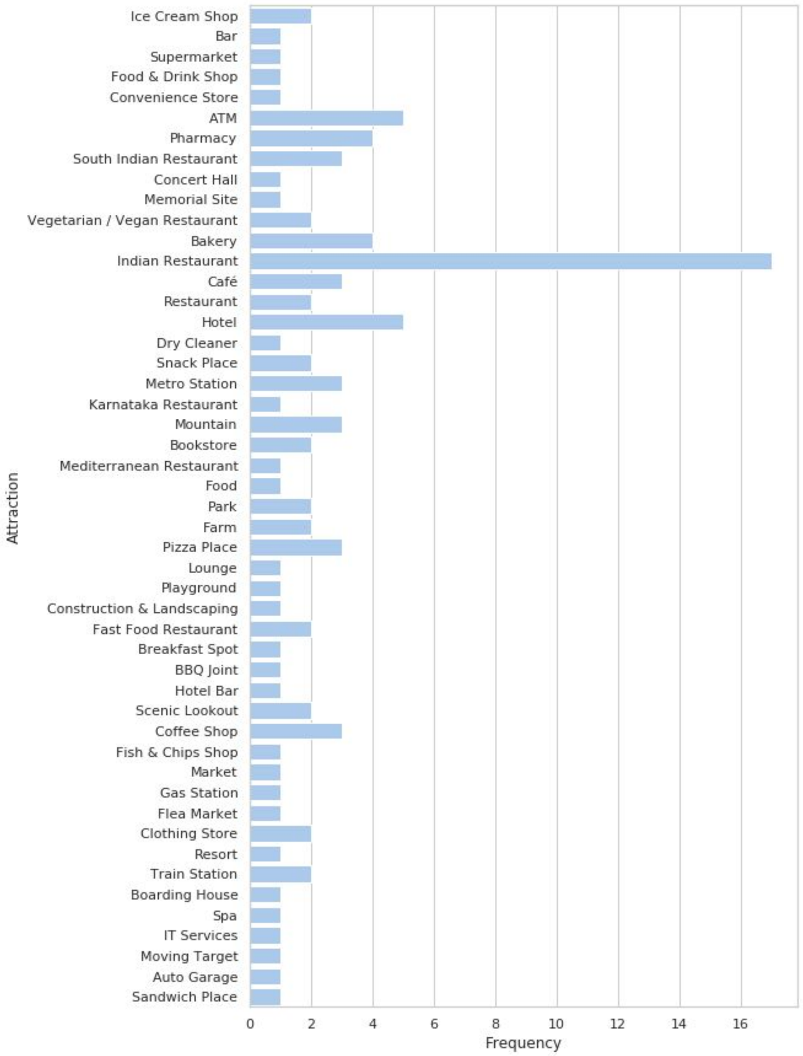
Initial map of neighborhoods

Analysis

```
Neighborhood
Adugodi            5
Agram            100
Akkur             28
Alahalli           1
Amruthahalli       3
```

172 unique categories

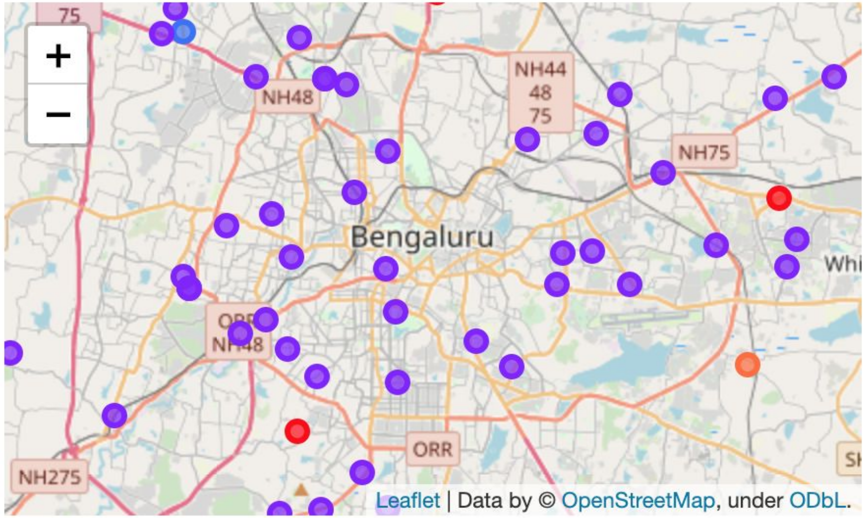
Distribution of most popular attraction by frequency



Methodology

K Means Clustering

Elbow Method and Silhouette Score method used to determine optimal value of k is 8



K Medoids Clustering

- Elbow Plot showed two distinct elbows at k=2 and k=6
- k=6 used to cluster the dataset. Each cluster analyzed for patterns

DBSCAN Clustering

- Density Based Spatial Clustering of Applications with Noise
- Robust to outliers
- Takes epsilon and minpts as inputs
- Experimented with various values of minpts and epsilon but data never clustered into more than 2 clusters.

Agglomerative Clustering

- Hierarchical clustering
- 8 clusters
- Clusters were observed for similarity

Results

Silhouette Score

K Means	0.19610746739247847
K Medoids	0.07733267946208815
DBSCAN	0.1391318150805085
Agglomerative	0.137552821016789

Davies-Bouldin Score

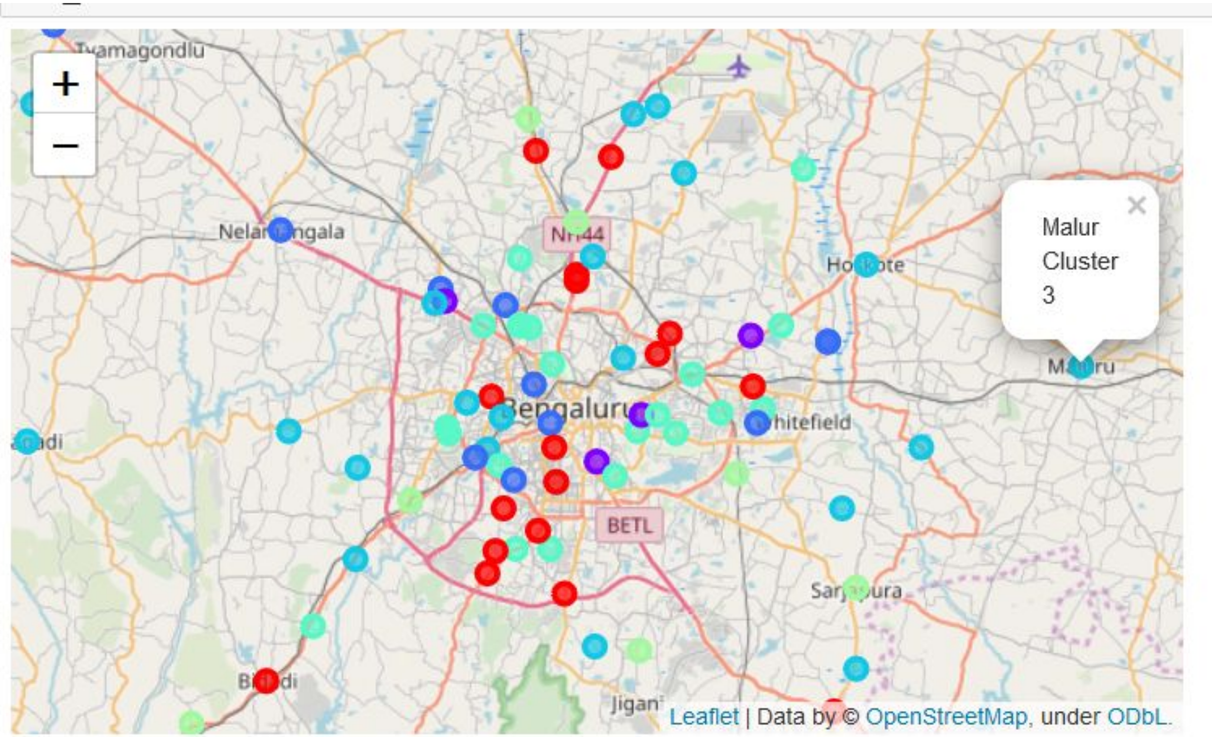
K Means	1.1792534290426284
K Medoids	3.916227564389721
DBSCAN	3.4678885613218626
Agglomerative	1.3956784752393547

Inference

DBSCAN seems to do best according to metrics as it has the best Silhouette Score as well as the second best Davies-Bouldin Score. However, it splits the dataset into 2 clusters which is too generalized.

K Medoids offers the best Davies-Bouldin Score so it its results can be considered for further analysis.

Clustered Map



Discussion

- We can analyze the resulting clusters based on what our needs are
- For aspiring entrepreneurs it can serve as a useful metric to predict the success of their venture based on what are the other popular attractions in the area
- For people considering moving into a neighborhood they can check what other neighborhoods are similar to t

Results

- Neighborhood Analysis can be a valuable step in making decisions for business and personal reasons.
- There are several different clustering algorithms. There is no 'one-size-fits-all' approach to clustering.
- Once we cluster neighborhoods we can try to identify patterns within the cluster or perform data analysis and visualization techniques to make better sense of the clusters.
- We can then use this information to guide our decision making process.

The End