The Battle of Neighbourhood

Submission: Sagar Gawali

Contents

- Business Problem
- Data Used
- Methodology
- Results and Observations
- Conclusion

Business Problem

- High demand for residential apartments /bungalows /villas, etc. (both rented/owned)
- Difficult for newcomer in the city or someone who is trying to relocate within the city to identify place which is good for staying and which is based on personal criteria.
- As part of this project, we would try to answer this question - Which area is suitable for different categories of people.
- We will also try to answer whether the area is good for staying for a bachelor's or families.

Data

For this project we will need the following set of data:

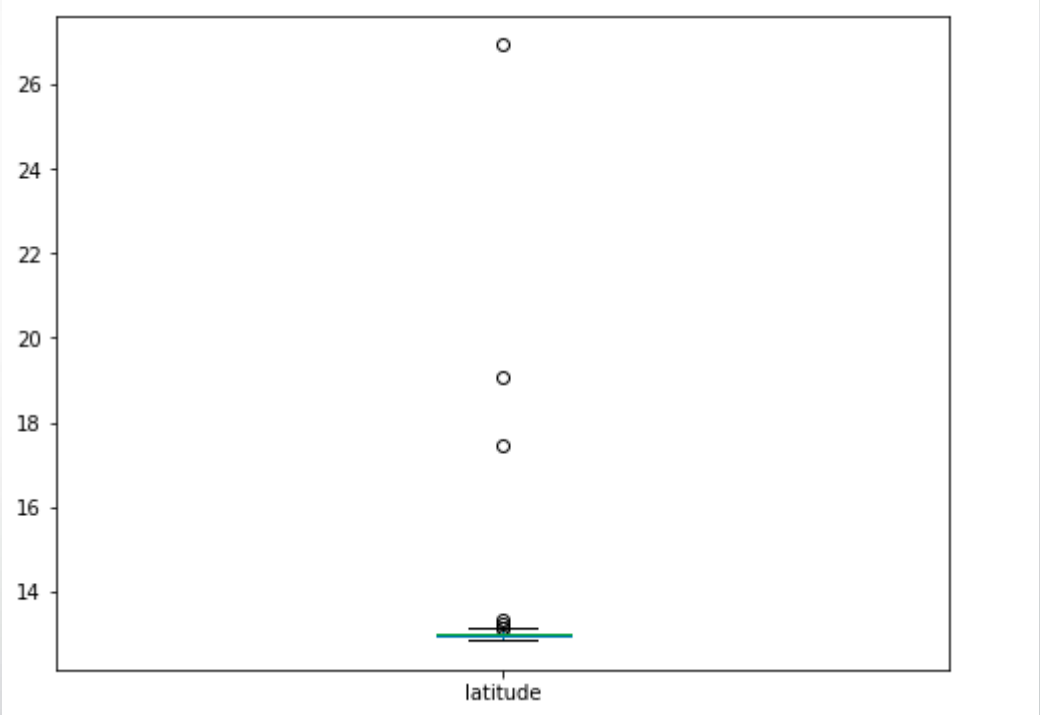
- Classification of areas of Bengaluru city into smaller area i.e. ward (a ward is an administrative unit of the city region;)
- Venue data (Malls, market, Supermarket, Parks, Schools, recreational activities, restaurants)

We will use the following sources for the above-mentioned data categories

- Ward data: https://opencity.in/data/bbmp-wards
- Venue Data: Foursquare APIs

Methodology

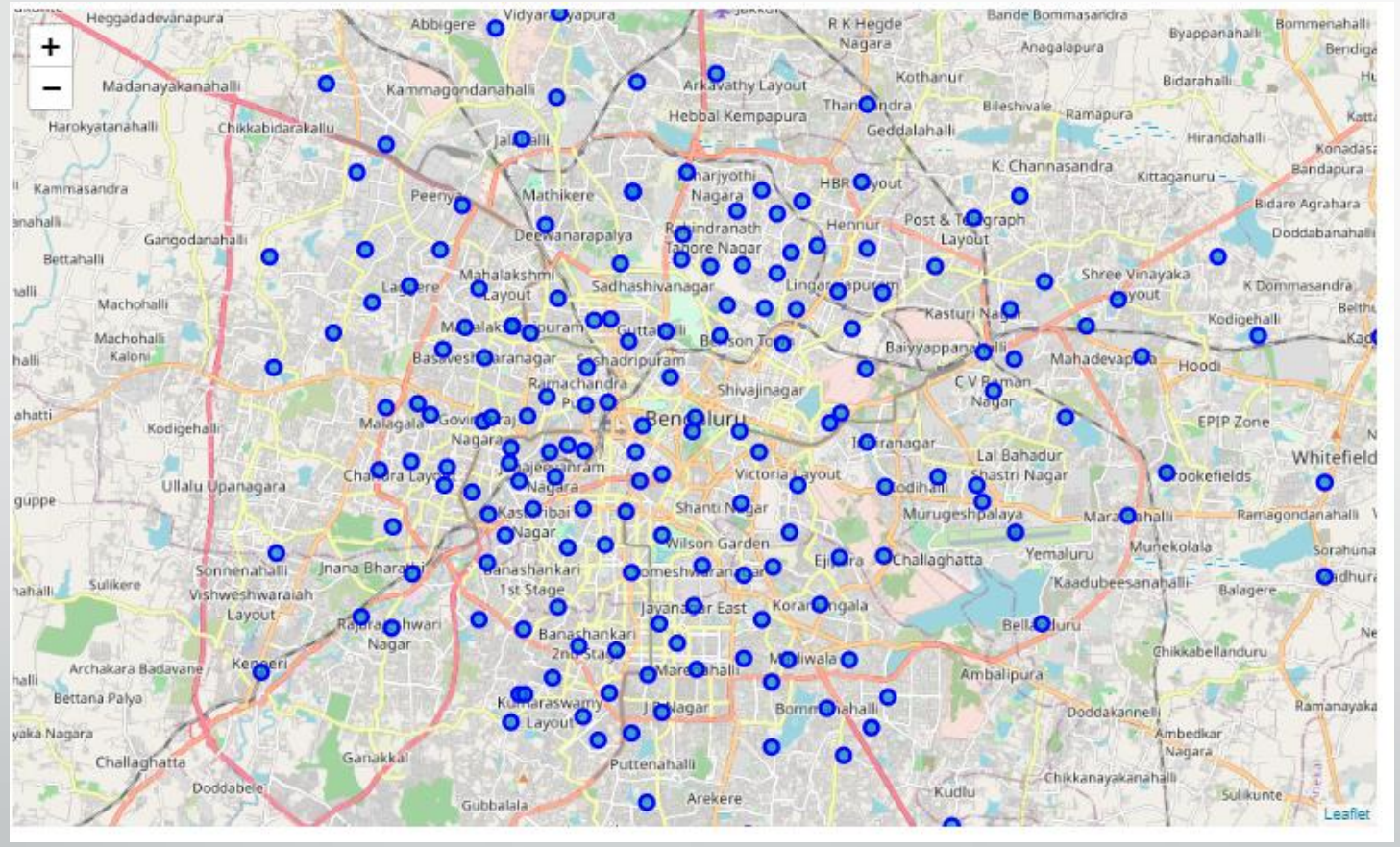
- Following methodology was followed to perform the analysis:
- Data collection from the opencity portal for list of wards.
  - Manual Data cleansing of the list of wards
  - Identifying latitude and longitude of the wards for analysis
    - Remove missing values of latitude and longitude.
    - Identified and Remove outliers in the data
      - Box plot showing outliers in the data (right side plot)



Box plot of latitude

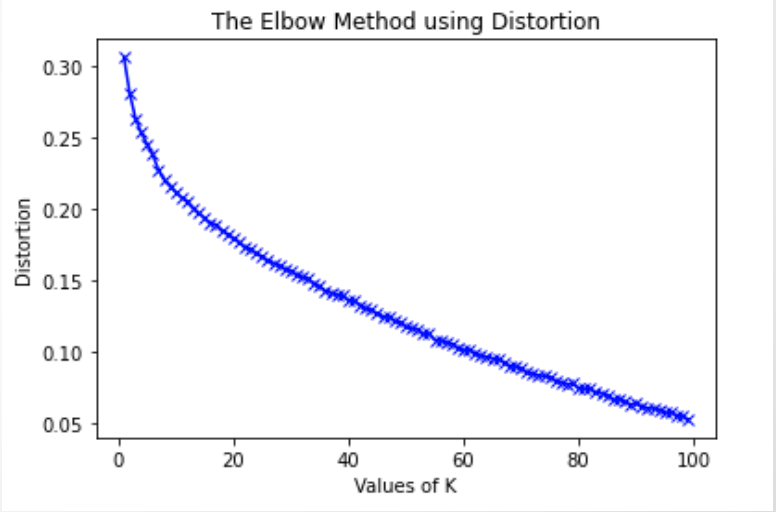
Methodology Cont…

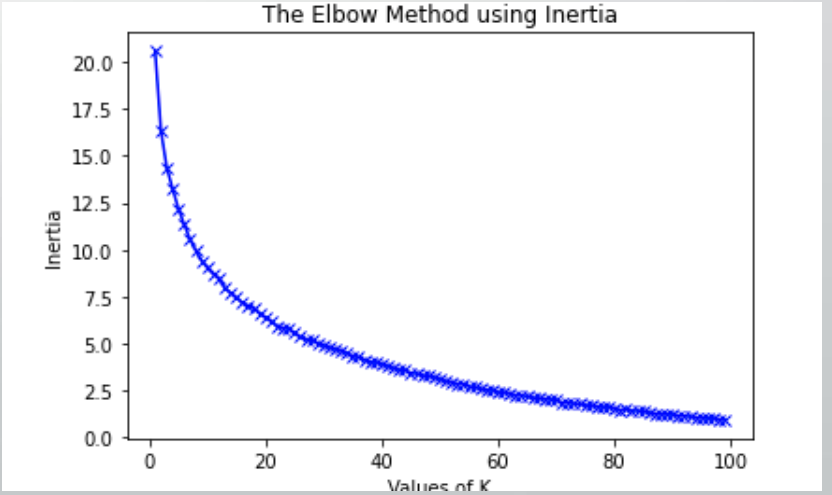
- Plot the ward on folium maps



Methodology Conti…

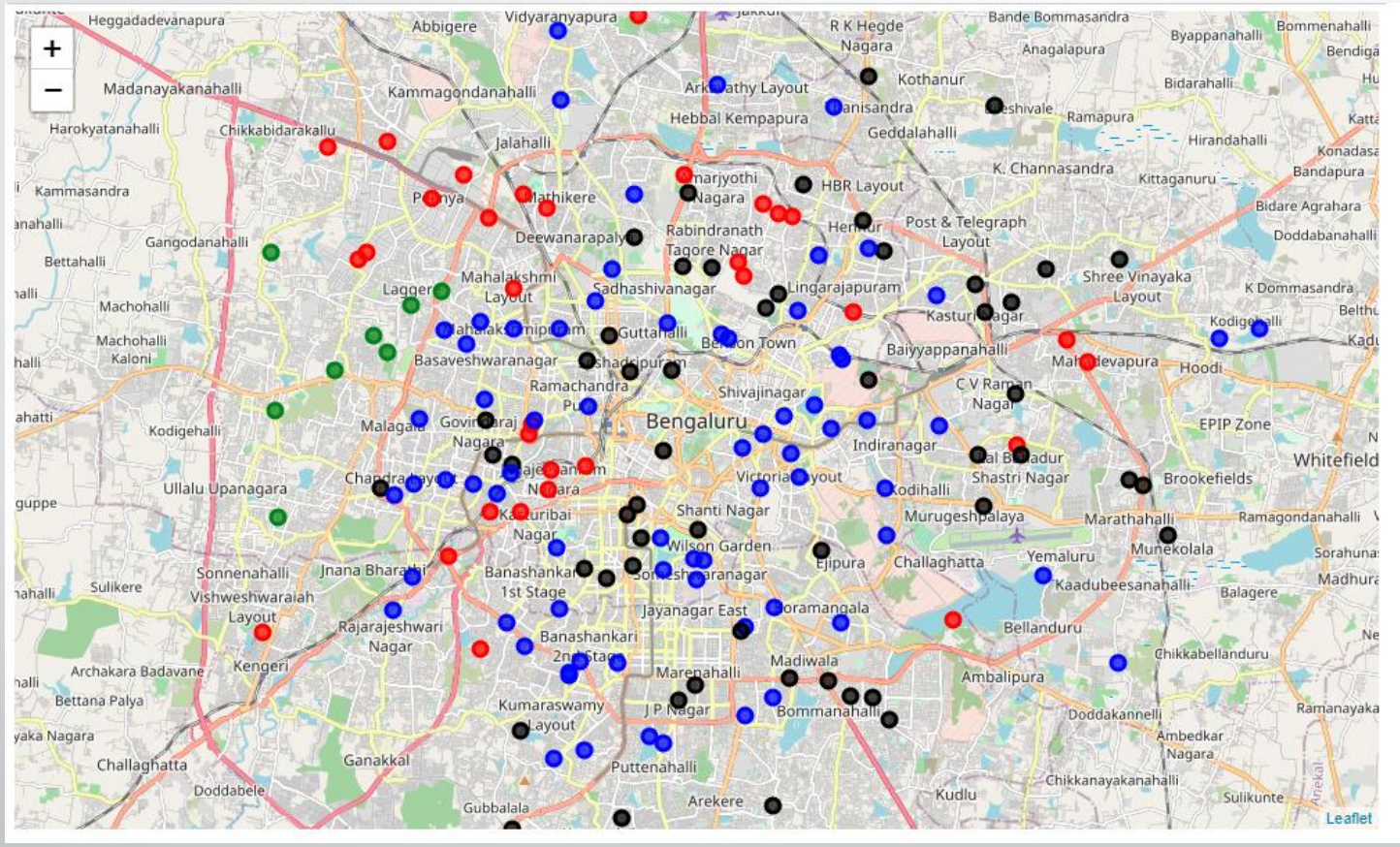
- Based on list of wards, latitude and, longitude, capturing venue data from Foursquare APIs
- Perform ward & Venue level analysis
  - It was identified that there were too many unique venue categories present in the data. Consolidate the venue categories based on business understanding and reduced the categories for further analysis.
- Performed one-hot coding i.e. created dummy variables
- Applied K-means algorithm for the clustering
  - Tried to identify optimal 'K' value for clustering





Methodology cont…

- Performed clustering by keeping value of K = 4.
  - Plotted clusters on folium map
- Examined and analysed the cluster to provide observation, results and conclusion.



Cluster Plot on Folium Map

Results and Observations

Based on the data, every ward in the city was classified into one of the four clusters. This was achieved through unsupervised learning algorithm of clustering using K- means. These four cluster were created from the data fetched from the foursquare API's related to venues.

- **Cluster 1:** In this cluster, first 3 'most common venues' are related to the food.
- **Cluster 2:** In this cluster, first 3 'most common venues' are related to Sports, Gym & Fitness Centre, Entertainment, Travel, Supermarket and restaurant, park etc
- **Cluster 3:** In this cluster, first 3 'most common venues' are related to food and more specifically to Indian Restaurants, snack & refreshments and Groceries
- **Cluster 4:** In this cluster, first 4 'most common venues' are related to ATM, Sports, Gym & Fitness Centre, Automobile.
-

Conclusion

1. Based on clusters obtained and observations made, following conclusion can be drawn:

Cluster	Suitable for
1	For the people loving different types food or bachelors who love to eat outside instead of cooking at home
2	For the people who are extroverts, who love to go out and enjoy or families who need all rounded facilities like Gym, sports, entertainment, parks, easy travel arrangement for schools and offices
3	For the people who love traditional food or native food, mostly to the higher age people who are accustomed to the native/traditional food items
4	For the people who enjoys food prepared at home but like to go out for sports, gym and fitness center, have personal vehicles

2. Though these conclusions are drawn based on cluster, these don't feel distinctive clusters. Some of these clusters feel overlapping and clear demarcation is absent.
3. Also, when we tried to identify optimal 'K' value for clustering using K-means, there was no clear output using elbow method.
4. Foursquare venue data had different categories present, however some of the key categories were missing e.g. Hospitals, Schools, Banks, Parks, Day Care Centers and Amusement Parks in the locality. There is high presence of food related categories. Thus, clustering is heavily biased towards food category. More elaborate venue data is required for better analysis of the neighborhood.

Thank you