Predicting the locality for living based on neighborhoods

Capstone Project - The Battle of the Neighborhoods (Week 2)
Applied Data Science Capstone course by Coursera/IBM

Prepared by Dharmendrasinh Raj

Problem statement

Moving to a new city is challenging when one is not familiar with the locality/areas of the city. In such situation, it is difficult to make a decision on where to live in which fulfills the need of the person.

The person might have needs of specific facilities/neighborhoods in the locality.

This project will be of interest to such individuals who wants to identify possible location/area which fulfills their need of closest proximity to facilities/neighborhoods of their choice.

Scope

The project will focus on the Ahmedabad city of Gujarat, India.

In this project, we will try to find optimal zones/areas in the city where 3 basic facilities i.e. School, Hospital and Indian Restaurant are in the closest proximity of the center of an area.

There are around 48 different zones/areas in Ahmedabad city and hundreds of facilities in the proximity. As part of this project, we will try to solve this problem and recommend 4 optimal areas that satisfy the above criteria of the end user.

Data sources

- FourSquare (https://foursquare.com/)
  Venue search API – this API returns a list of venues near the given location (coordinates) that matches a search criteria.

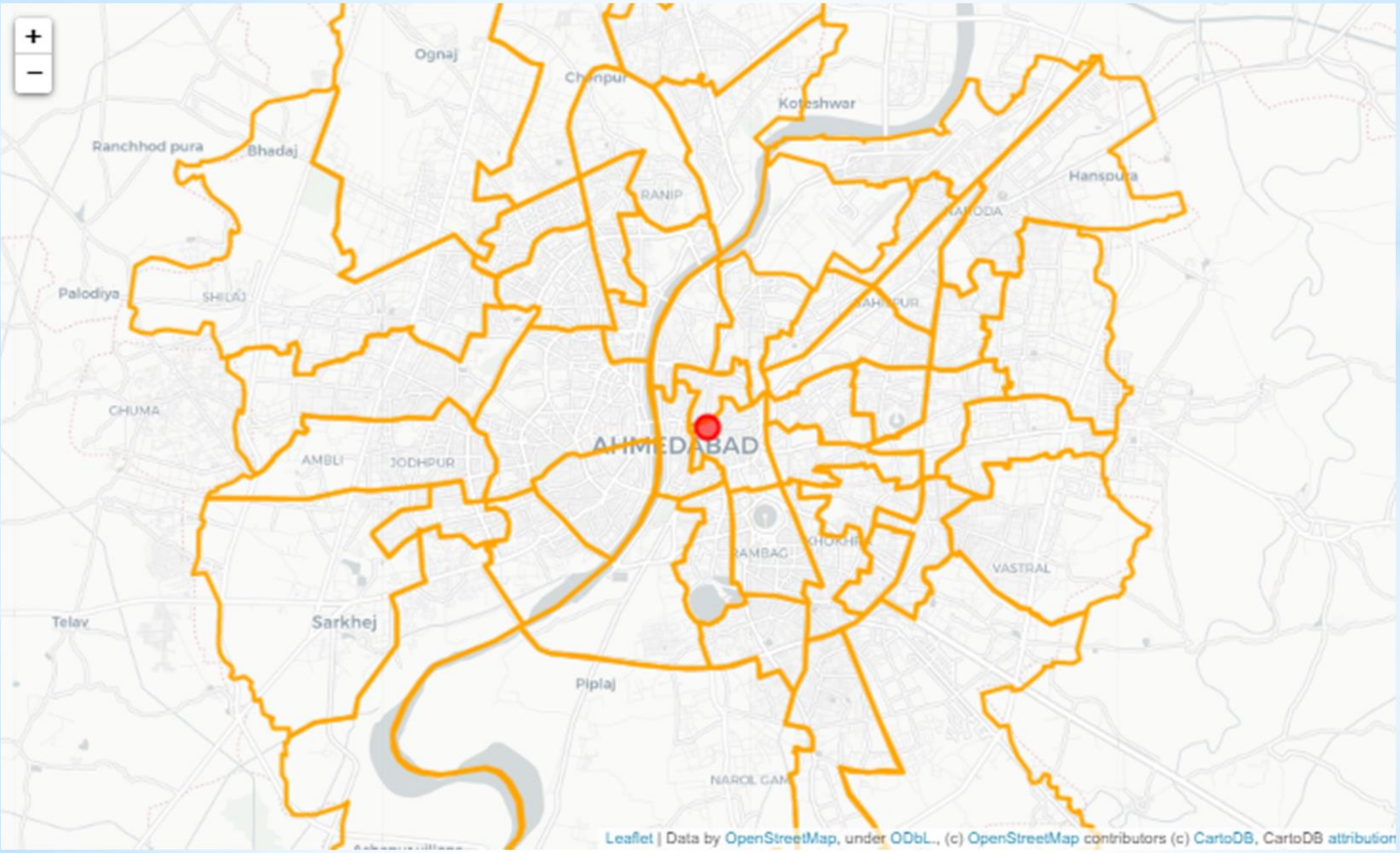
- Geojson file OpenCage Geocoder (http://projects.datameet.org/)
  Wards.geojson – this file provides the list of areas in Ahmedabad city and polygon map coordinates (lat/long) for each of the area.

- Facilities/Neighborhoods dataset:
  this is a local dataset of facilities that contains the list of facilities that the user can decide to choose from e.g. School, Bank, Hospital, Sports center etc. The dataset is created manually with category Ids extracted from FourSquare website.

- Coordinates dataset (https://simplemaps.com/data/in-cities):
  this is a local dataset that contains the list of cities in India and their respective city center coordinates. The data has been extracted from Simple Maps website.

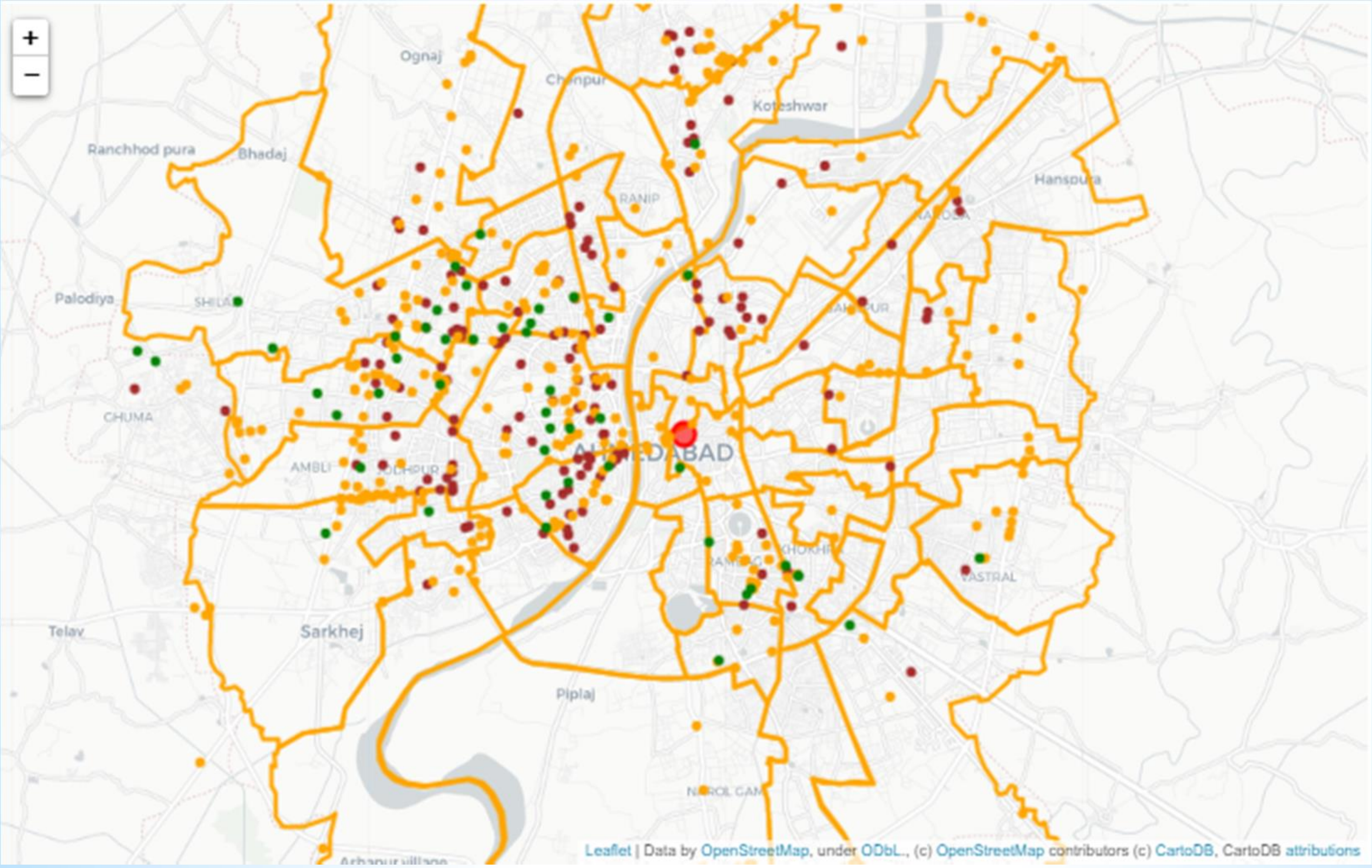
Methodology – Plotting areas

- Extract the city and facilities data from data sources.
- Retrieve the list of areas, their coordinates from Geojson file.
- Clean up the data and plot the zones/areas of the city on the map as below.



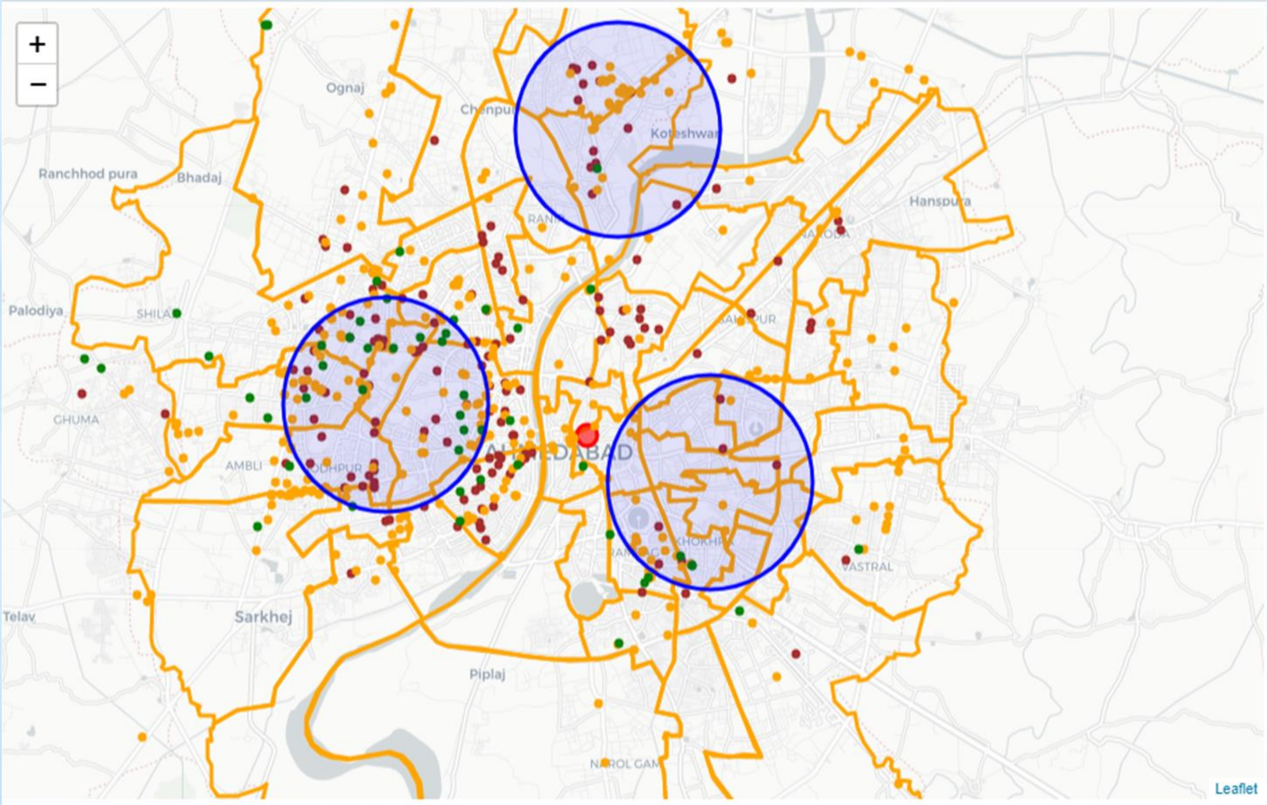
Methodology – Plotting facilities

- Retrieve the facilities data using FourSquare API for each of the area.
- Normalize and clean up the the retrieved data.
- Plot the facilities on the map as below in different colors based on facility type.



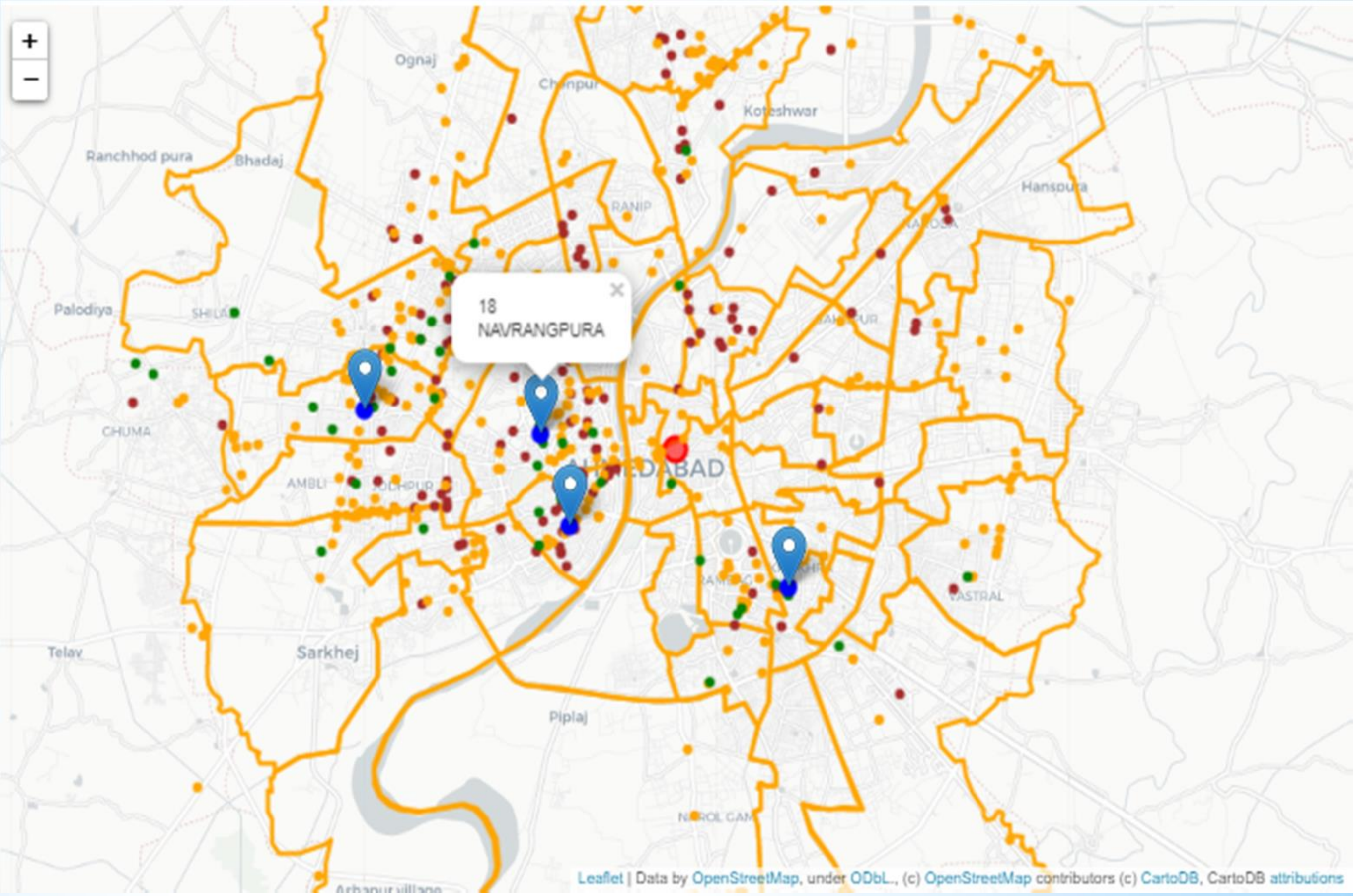
Exploratory data analysis

- Prediction and data analysis carried out using Kmeans machine learning algorithm.
- Generated the clusters of facilities
- Analysis shows that the most densed areas are Bodakdev, Navrangpura and Jodhpur.



Result

- Process the data to find the distribution of facilities around different areas and their relative distance from area center.
- After processing the data, following 4 areas were recommended based on the criteria set:
  - Bodakdev
  - Khokhra
  - Navrangpura
  - Paldi



The results we have achieved with our methodology aligns with this exploratory data analysis.

Conclusion

Purpose of this project was to recommend zones/areas that has facilities of School, Hospital, Indian restaurant close to center of a zone with lowest distance in order to aid the person in narrowing down the search for optimal location for living.

Using FourSquare data, first we identified the list of facilities within the 5km radius of each of the area that satisfy the criteria of end user. Then identified the lowest distance for each of the facility type from center of the area.

The data of all facilities could not be retrieved fully due to FourSuqare API restriction. The result can be further optimized by getting full data. The final decision on optimal area will be made the end user based on various factors e.g. the location of the office, type of school whether nursery, primary, high school etc.