Project Report

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

Predicting the locality for living based on neighborhoods

Prepared by Dharmendrasinh Raj

1. Introduction / Business Problem

1.1. Background

When one has to move to a new area or city/town that they are not familiar with, it is difficult to make a decision on where to live in which fulfills the need of the person. For example while moving to a totally new city while changing a job.

The person might have different needs such as if someone is having school going kids, they may want to stay to a place where the schools, playgrounds are nearby. Someone who frequently travels by air may want to stay in an area which is in proximity to Airport and has an ease of public transport.

1.2. Problem

One may also rent a place temporary and then look out for preferred areas. But this is a time- consuming process and moving from one place to another can result in excessive cost.

For a small places/towns, it is easy to identify such areas. However, in big cities, this becomes a challenge.

1.3. Audience

The project will be of interest to anyone who wants to identify a location/area which fulfills their need of facilities/neighborhoods of their choice. The project aims to address this problem by recommending the areas that matches the desired facilities criteria of a person.

2. Data Sources

Following datasets/APIs will be used the solve the problem –

1) FourSquare (https://foursquare.com/)
   - Venue search API – this API returns a list of venues near the given location (coordinates) that matches a search criteria.
   - API Request URL – https://api.foursquare.com/v2/venues/search
   - Full API Document Reference – https://developer.foursquare.com/docs/api-reference/venues/search/

2) 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.
   - Reference URL – https://raw.githubusercontent.com/datameet/Municipal_Spatial_Data/master/Ahmedabad/Wards.geojson

3) 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.

Reference to FourSquare link where the venue category data is available:
https://developer.foursquare.com/docs/build-with-foursquare/categories/

The final dataset will look something like below:

Neighborhood facilities	cat_id
Airport	4bf58dd8d48988d1eb931735
Bank	4bf58dd8d48988d10a951735
Bus station	4bf58dd8d48988d1fe931735
Bus Stop	52f2ab2ebcbc57f1066b8b4f
Cinema	4bf58dd8d48988d17f941735
College & University	4d4b7105d754a06372d81259
Community center	52e81612bcbc57f1066b7a34
Fire station	4bf58dd8d48988d12c941735

4) Coordinates dataset – 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 (https://simplemaps.com/data/in-cities)

The dataset will look something like below:

city	lat	lng	country
Mumbai	18.987807	72.836447	India
Delhi	28.651952	77.231495	India
Kolkata	22.562627	88.363044	India
Chennai	13.084622	80.248357	India
Bengaluru	12.977063	77.587106	India
Hyderabad	17.384052	78.456355	India
Ahmedabad	23.025793	72.587265	India
Haora	22.576882	88.318566	India
Pune	18.513271	73.849852	India
Surat	21.195944	72.830232	India

3. Methodology

3.1. Scope of the project

Before we discuss the methodology, let's define the scope of this project.

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.

3.2. Methodology

In the first step, we will extract the required data from various data sources mentioned in the Data section of this document.

Extraction of city coordinates data:

	city	lat	lng	country	iso2	admin	capital	population	population_proper
0	Mumbai	18.987807	72.836447	India	IN	Maharashtra	admin	18978000.0	12691836.0
1	Delhi	28.651952	77.231495	India	IN	Delhi	admin	15926000.0	7633213.0
2	Kolkata	22.562627	88.363044	India	IN	West Bengal	admin	14787000.0	4631392.0

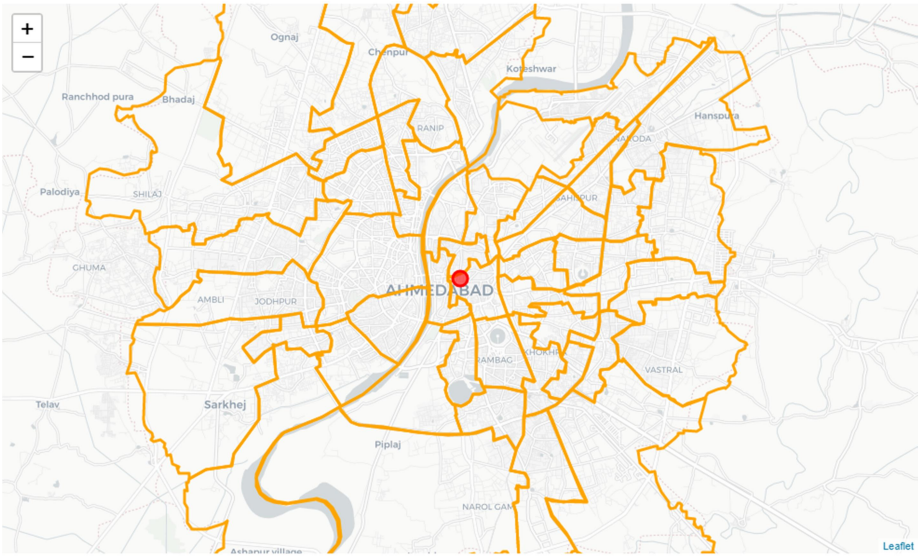
Extraction of facilities/Neighborhood data:

	Neighborhood facilities	cat_id
0	Airport	4bf58dd8d48988d1eb931735
1	Bank	4bf58dd8d48988d10a951735
2	Bus station	4bf58dd8d48988d1fe931735
3	Bus Stop	52f2ab2ebcbc57f1066b8b4f
4	Cinema	4bf58dd8d48988d17f941735

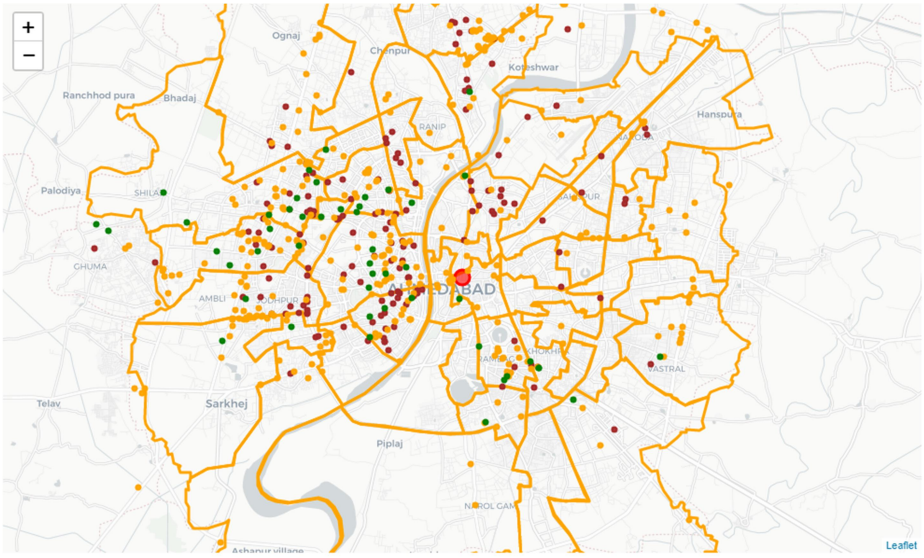
Then we will define the city; Ahmedabad in our case and the facilities criteria which will be School, Hospital and Indian Restaurant.

After gathering all the required data sources, we start the process of processing, analysis and identification of candidate areas.

We will start with displaying the city map and highlighting the city center using plotting libraries. Once done, clearly highlight different areas of the city using the geojson file.



Now that we have clear map view, we will extract the list of nearby facilities within 5km radius of each area using FourSquare API. We will plot these facilities on the map to understand the distribution of venues over the different areas.



The next step will be to go through each of the area and identify if each of the facility type is available in the proximity. Process the available data and find out the distance of nearest facility for each of the facility type. Once the distances are calculated, this will provide us with the areas where each of nearest facility type is in a closest proximity.

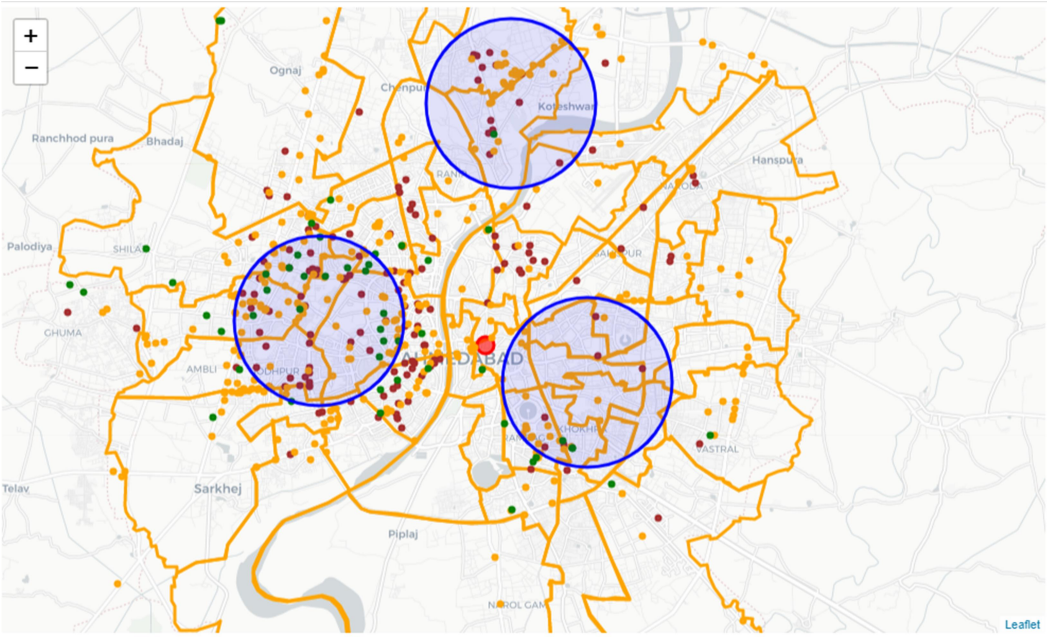
Retrieve 4 optimal areas based on the density and closest proximity of facilities and plot them on the map.

3.3. Exploratory data analysis

Kmeans machine learning algorithm will be used for prediction and exploratory data analysis to verify the results in the next section.

Using this algorithm, we will generate clusters of facilities based on their density over the city and analyze the most densed areas/zones.

The analysis shows that the most densed areas are Bodakdev, Navrangpura and Jodhpur.



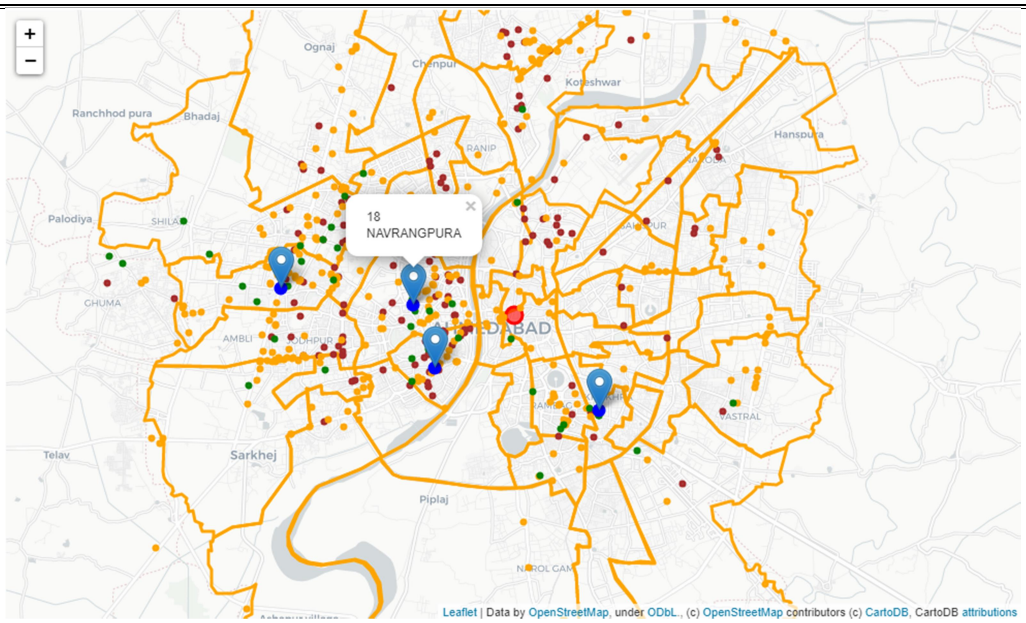
4. Result

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

As per the analysis, following 4 areas are the possible city areas that has the hospital, school and Indian restaurant in the closest proximity:

1. Bodakdev
2. Navrangpura
3. Khokhra
4. Paldi

Note: the result is based on the distance of facilities from the **center of the area** and not from each other.



5. Discussion

The data of all facilities could not be retrieved fully due to FourSuqare API restriction (with free develooper product version, there is a limit on the number of records returned in API and number of API calls that can be made in a day). This could impact the result since we don't get the full list of venues/facilities through the API request.

The result can be further optimized and better recommendation can be given if the full data is made available.

In this project, we have focused on the proximity to the center of the zone/area. The results can be different if the proximity between individual facilities are taken into consideration.

6. 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 final decision on optimal area will be made the end user based on various other factors e.g. the location of the office, type of school such as nursery, primary, high school etc.