Coursera Capstone Project – Applied Data Science

Vivek Nichenametla

Introduction – Business Problem

- Hyderabad, located in the Southern India is fourth largest city in India and is rapidly growing. The city is famous for its varied cuisine and with existing variety of restaurant chains, starting a restaurant is difficult task.
- This project is to find a suitable place in the city to start a restaurant, which of course is dependant on several factors. Some of them are
  - Similar restaurants around the place
  - The connectivity to the place
  - The affordability of the potential customers around the place
  - Availability and real estate price of the location, etc.,
- This project deals with the fundamental analysis based on location and clustering the restaurants obtained from Foursquare API. The other factors mentioned above are not in the scope of this project though their significance cannot be neglected.

Introduction – Business Problem

- **Business Problem:**

What is the ideal location to start a restaurant in the city?

- **Target Audience:**

This project helps present and future restaurateurs, as location is one of the primary decisions to make while looking to start a business.

Data

Neighborhoods:

The data of neighborhoods in Hyderabad are extracted by web scraping using BeautifulSoup library in Python. In this case the source of the data is

https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India

Geo-coding:

After retrieving the data of neighborhoods into a pandas DataFrame, the latitudes and longitudes are obtained using geocoder package in Python

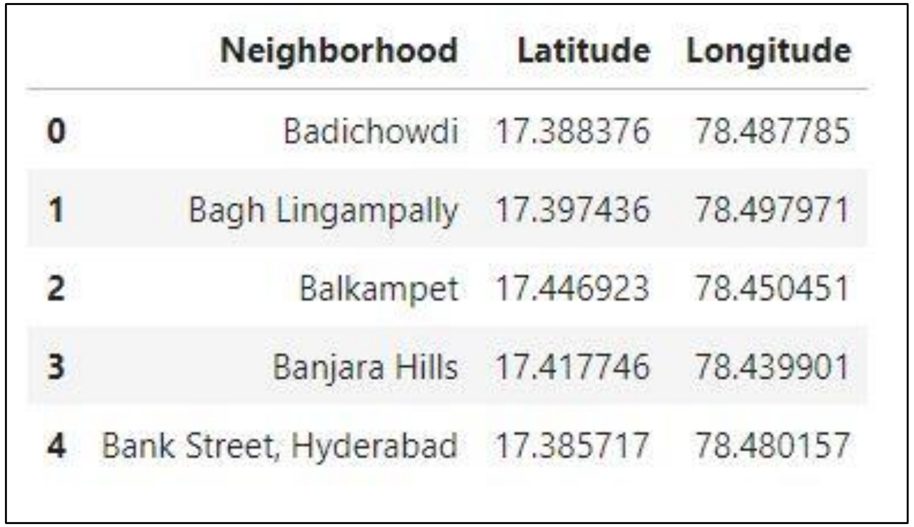
Venue details:

The venue details are obtained using Foursquare API.

https://developer.foursquare.com/

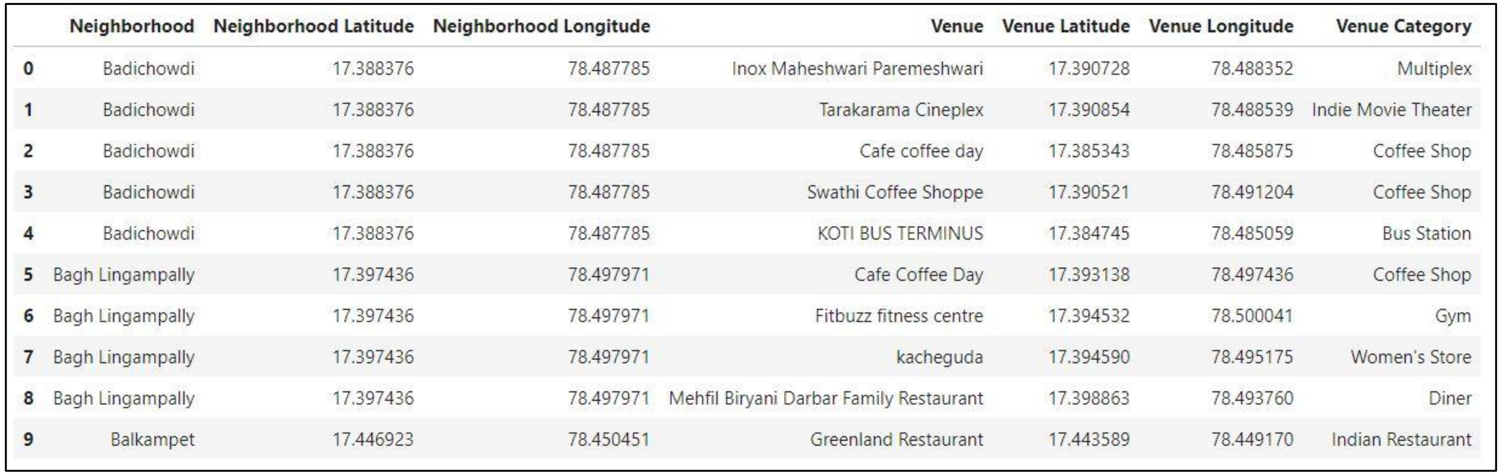
Data - Example

Neighbourhood Data



	Neighborhood	Latitude	Longitude
0	Badichowdi	17.388376	78.487785
1	Bagh Lingampally	17.397436	78.497971
2	Balkampet	17.446923	78.450451
3	Banjara Hills	17.417746	78.439901
4	Bank Street, Hyderabad	17.385717	78.480157

Hyderabad Venues_Foursquare API



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Badichowdi	17.388376	78.487785	Inox Maheshwari Paremeshwari	17.390728	78.488352	Multiplex
1	Badichowdi	17.388376	78.487785	Tarakarama Cineplex	17.390854	78.488539	Indie Movie Theater
2	Badichowdi	17.388376	78.487785	Cafe coffee day	17.385343	78.485875	Coffee Shop
3	Badichowdi	17.388376	78.487785	Swathi Coffee Shoppe	17.390521	78.491204	Coffee Shop
4	Badichowdi	17.388376	78.487785	KOTI BUS TERMINUS	17.384745	78.485059	Bus Station
5	Bagh Lingampally	17.397436	78.497971	Cafe Coffee Day	17.393138	78.497436	Coffee Shop
6	Bagh Lingampally	17.397436	78.497971	Fitbuzz fitness centre	17.394532	78.500041	Gym
7	Bagh Lingampally	17.397436	78.497971	kacheguda	17.394590	78.495175	Women's Store
8	Bagh Lingampally	17.397436	78.497971	Mehfil Biryani Darbar Family Restaurant	17.398863	78.493760	Diner
9	Balkampet	17.446923	78.450451	Greenland Restaurant	17.443589	78.449170	Indian Restaurant

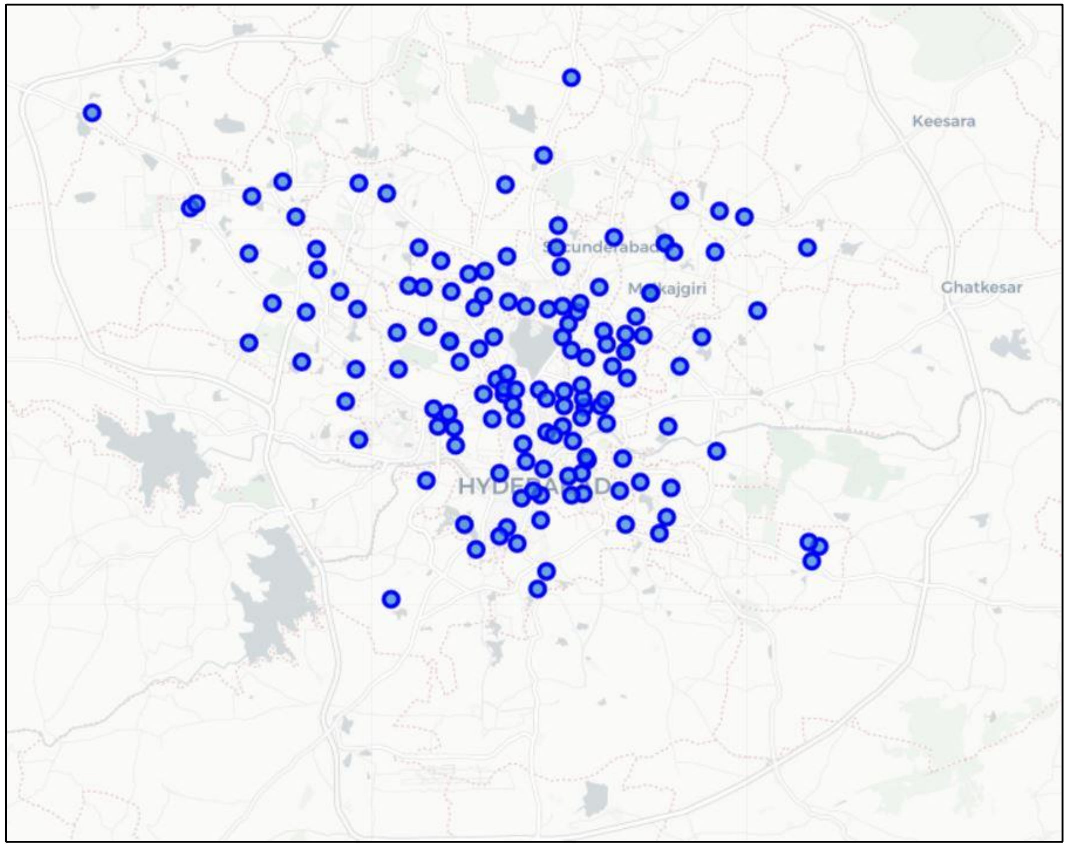
Methodology

BeautifulSoup:

- Data of Neighbourhoods is scraped from the Wikipedia page into Data Frame using BeautifulSoup library.

Folium:

- Folium library which is used to visualize the maps using location data comes very handy in this case. Hence the venues data and various clusters are visualized using Folium library



Neighbourhoods in Hyderabad

Methodology

Foursquare API:

- The venue details are obtained by calling Foursquare developer API for the respective neighborhoods.
- Since, the names of places in India are not segregated well, venue categories which contain the strings such as ‘Restaurant’, ’Food’, ‘Sandwich Place’, ‘Café’ are separated into a Data Frame for the analysis.

One-Hot Coding:

- One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. For the K-means Clustering Algorithm, all unique restaurants are one-hot encoded.

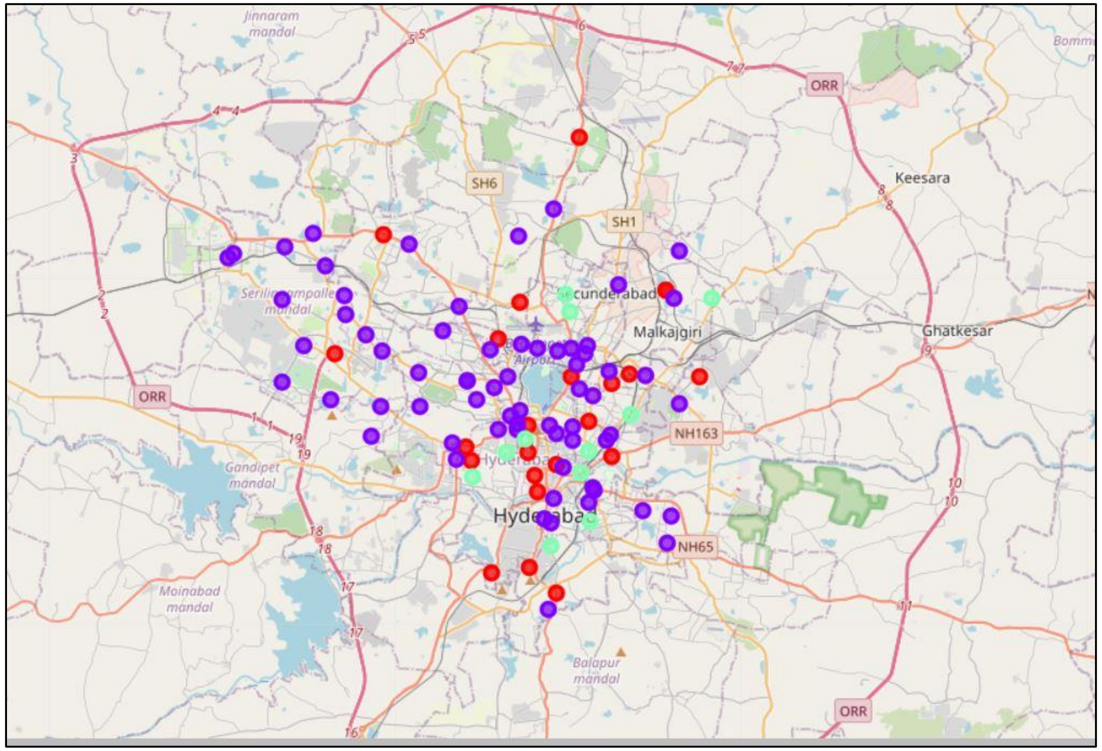
Methodology

K-Means Clustering:

- The venue data is then trained using K-Means Clustering Algorithm to get the desired clusters for the analysis. K-Means is computationally faster in case of large data sets compared to other clustering algorithms.

Results

- The clusters are visualized using the folium library as segregated.
- These clusters are segregated based on the similar restaurants present in the neighborhoods.



Discussions

- The places in the cluster 2 have many restaurants and are similar to one another in terms of availability of restaurants and eat-outs in Hyderabad. People are overloaded with various choice of restaurants in their neighbourhood. Hence, restaurateurs need to be extremely innovative in service and creating environments, which often is capital intensive.
- The majority of the areas in cluster 3 are located in the older part of the city which habitats people from below middle class section of society. This can be a good choice for eat-outs offering services at less cost as preferred by the general public.
- The areas in cluster 1 accommodates middle and above middle class people, in the heart of the city which have shown considerable growth in the past ten to twelve years. This can be a good spot for restaurants to open.

Conclusion

- Cluster 1 is a very good choice to attract customers, as there aren’t many restaurants around. People in these areas need not travel far or depend on online deliveries for good food anymore. In addition, residents who were not keen to travel for food before can now visit the restaurants in this area.
- With minimum capital investment, as these areas are not very expensive, owners can serve the customers. Hence this is a win-win situation for customers as well restaurant owners.