Establishing Food Business in Hyderabad

- Applied Data Science Capstone Project by IBM/Coursera

1. Introduction

Hyderabad is one of the fastest growing cities in India, with this rapid growth in and around the city, the doors are open for the huge workforce increasing the city population and also the potential for the new and innovative businesses due to the bigger market availability. One of the fastest growing line of business which is evergreen is the food business. However, considering the diverse population of the city it is essential to understand the tastes and the locations in the city before entering into the business. Before venturing this highly diverse city the ideal location of the establishment and the cuisine liked by most of population must be decided.

2. Business problem

The objective of this project is to analyze and select the best locations in the city of Hyderabad, India, to open a new restaurant. This project is mainly focused on geospatial analysis of the Hyderabad City and to determine the best possible location and cuisine for the new restaurant. Using data science methodology and machine learning techniques like clustering, let's try to answer the business question - What could be the ideal location to establish the new restaurant and what cuisine is mostly welcomed by the residents of Hyderabad?

3. Data acquisition & Cleaning

3.1. Data Requirements

- To answer the above business problem, we need the following data:
  - List of all the areas in Hyderabad City.
  - Geospatial coordinates of the areas in the city
  - Data of different venues in City

3.2. Data Sources

The list of areas/neighborhoods in city can be extracted from this Wikipedia page. The scrapping of the html page can be done using the python requests and beatifulsoup package. The extracted data can then be converted into a data frame using pandas.

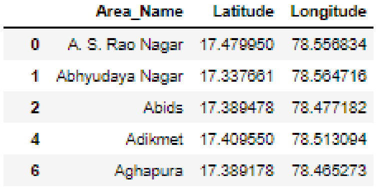
The geospatial coordinates of all the areas/ neighborhoods can be extracted using the geopy and geo-coder packages. Passing the values extracted in scrapping through the geocoder user agent and fetching the geospatial coordinates. Then the entire data set extracted can be converted into a data frame using pandas.

Foursquare API can be used to fetch the venues data in the city and the using clustering on this data the business question can be answered.

4. Methodology

4.1. Using Foursquare API to retrieve the venues data

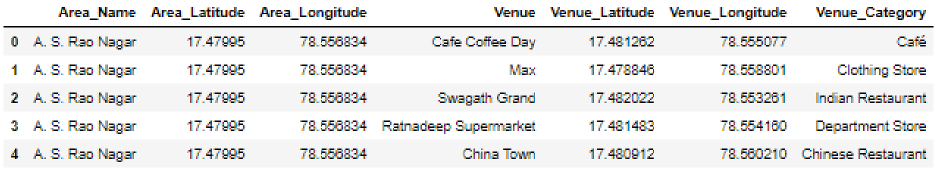
The data prepared by merging the area data from scrapping the Wikipedia page and the geospatial data from the geocoder package is converted into a data frame as below.



	Area_Name	Latitude	Longitude
0	A. S. Rao Nagar	17.479950	78.556834
1	Abhyudaya Nagar	17.337661	78.564716
2	Abids	17.389478	78.477182
4	Adikmet	17.409550	78.513094
6	Aghapura	17.389178	78.465273

There are total of 165 areas identified from through scrapping.

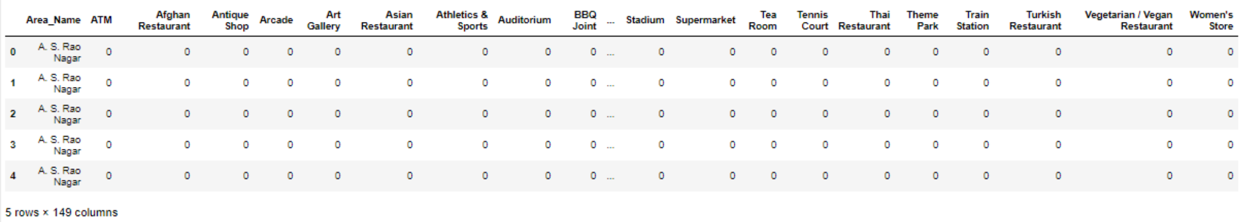
Using the Foursquare API, the venue details of all the areas in the city is extracted and converted into a data frame.



	Area_Name	Area_Latitude	Area_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	A. S. Rao Nagar	17.47995	78.556834	Cafe Coffee Day	17.481262	78.555077	Café
1	A. S. Rao Nagar	17.47995	78.556834	Max	17.478846	78.558801	Clothing Store
2	A. S. Rao Nagar	17.47995	78.556834	Swagath Grand	17.482022	78.553261	Indian Restaurant
3	A. S. Rao Nagar	17.47995	78.556834	Ratnadeep Supermarket	17.481483	78.554160	Department Store
4	A. S. Rao Nagar	17.47995	78.556834	China Town	17.480912	78.560210	Chinese Restaurant

Details of 978 venues in total are extracted via API call for the 165 areas.

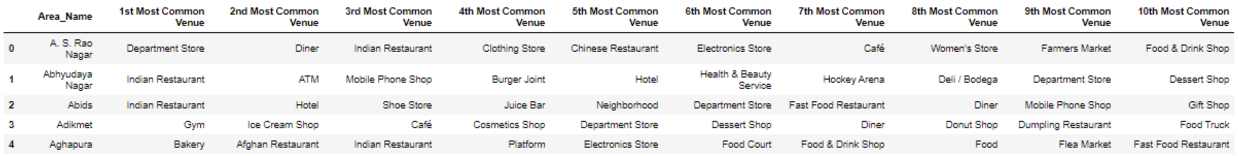
Applying one-hot encoding on this data so that the venue categories can be converted into categorical values. The converted data frame is as below:



	Area_Name	ATM	Afghan Restaurant	Antique Shop	Arcade	Art Gallery	Asian Restaurant	Athletics & Sports	Auditorium	BBQ Joint	...	Stadium	Supermarket	Tea Room	Tennis Court	Thai Restaurant	Theme Park	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant	Women's Store
0	A. S. Rao Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	A. S. Rao Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	A. S. Rao Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	A. S. Rao Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	A. S. Rao Nagar	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 149 columns

As there are multiple entries for the same area, we can group the above data set on area name and then find the top 10 venues of each area so that we can cluster the areas based on the pattern of common venues.



	Area_Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	A. S. Rao Nagar	Department Store	Diner	Indian Restaurant	Clothing Store	Chinese Restaurant	Electronics Store	Café	Women's Store	Farmers Market	Food & Drink Shop
1	Abhyudaya Nagar	Indian Restaurant	ATM	Mobile Phone Shop	Burger Joint	Hotel	Health & Beauty Service	Hockey Arena	Deli / Bodega	Department Store	Dessert Shop
2	Abids	Indian Restaurant	Hotel	Shoe Store	Juice Bar	Neighborhood	Department Store	Fast Food Restaurant	Diner	Mobile Phone Shop	Gift Shop
3	Adikmet	Gym	Ice Cream Shop	Café	Cosmetics Shop	Department Store	Dessert Shop	Diner	Donut Shop	Dumpling Restaurant	Food Truck
4	Aghapura	Bakery	Afghan Restaurant	Indian Restaurant	Platform	Electronics Store	Food Court	Food & Drink Shop	Food	Flea Market	Fast Food Restaurant

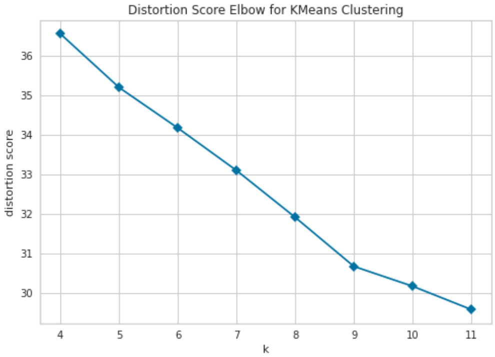
4.2. Clustering of Areas

A fundamental step for any unsupervised clustering algorithm is to determine the optimal number of clusters into which the data may be clustered.

So, to find the optimum *k* value, we can use the below methods:

- Elbow Curve - Elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

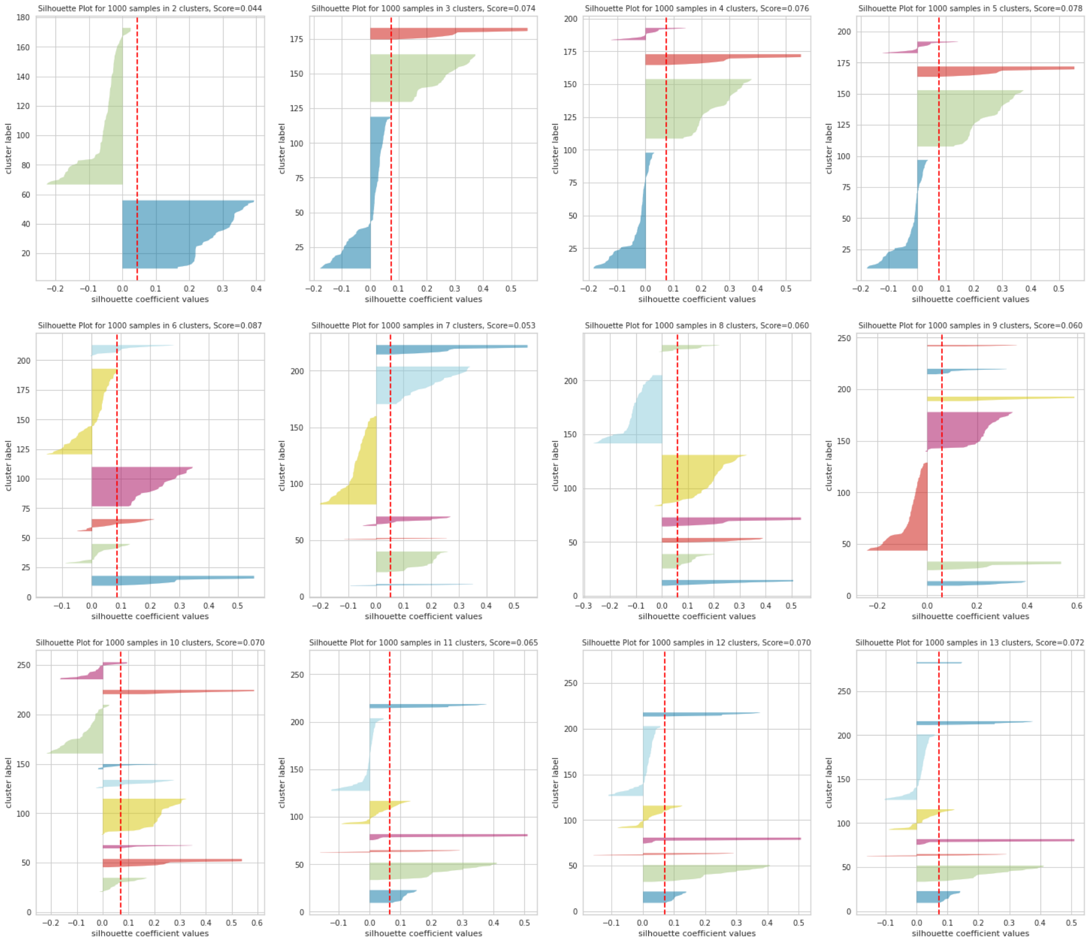
Plotting the Elbow curve using the KElbowVisualizer from yellowbrick for different values of *k* between 4-11 on our data.



As the curve is almost decreasing linearly, we cannot rely on only elbow curve to determine our *k* value.

- Silhouette Score - Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.

Plotting the Silhouette curve using the SilhouetteVisualizer from yellowbrick for different values of *k* between 2-13 on our data.

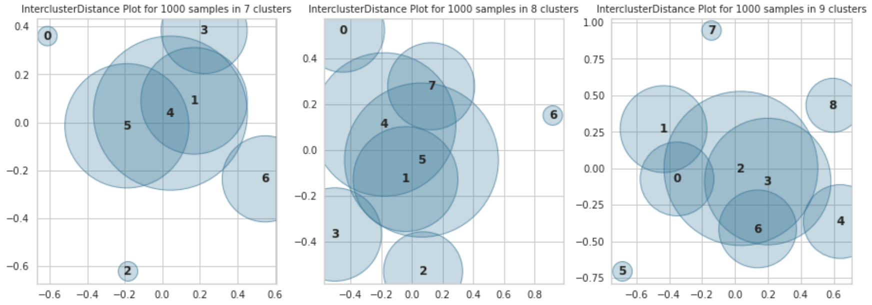


From the subplots above for the *k* values from 2-13 we can clearly see the most optimal values for *k* will be between 7-9 as the spikes are more consistent and the score is nearer to +1.

Now, let us try to find the optimal *k* value between 7-9 using the inter cluster distance method.

- Inter cluster distance – Inter cluster distance is the distance between two objects belonging to two different clusters.

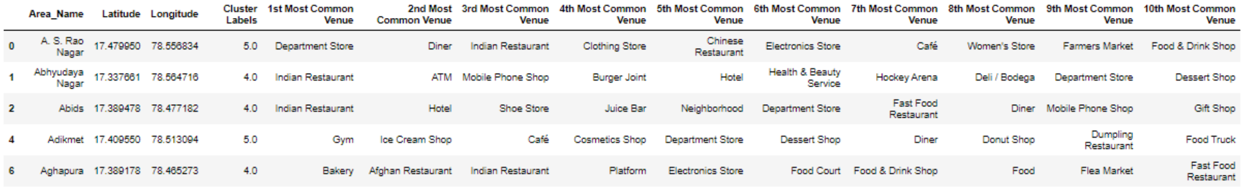
Plotting the Inter cluster Distance plot using the InterclusterDistance from yellowbrick for different values of *k* between 7-9 on our data.



Looking at the above inter cluster distance plot for the values of *k* between 7-9, we can see that the graph for *k*=8 is more aligned to center and also more loosely packed than the rest.

Hence, we can pick the *k* value as 8 to apply *k*-means clustering on our data.

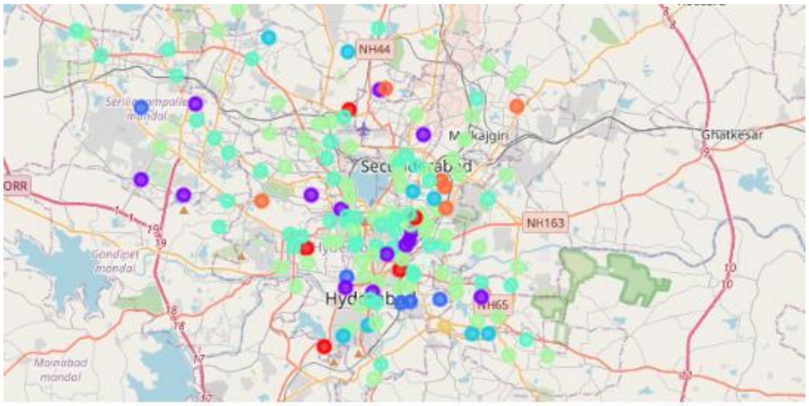
Now, clustering the venues data into 8 clusters we create a data frame as below.



	Area_Name	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	A. S. Rao Nagar	17.479950	78.556834	5.0	Department Store	Diner	Indian Restaurant	Clothing Store	Chinese Restaurant	Electronics Store	Café	Women's Store	Farmers Market	Food & Drink Shop
1	Abhyudaya Nagar	17.337661	78.564716	4.0	Indian Restaurant	ATM	Mobile Phone Shop	Burger Joint	Hotel	Health & Beauty Service	Hockey Arena	Deli / Bodega	Department Store	Dessert Shop
2	Abids	17.389478	78.477182	4.0	Indian Restaurant	Hotel	Shoe Store	Juice Bar	Neighborhood	Department Store	Fast Food Restaurant	Diner	Mobile Phone Shop	Gift Shop
4	Adikmet	17.409550	78.513094	5.0	Gym	Ice Cream Shop	Café	Cosmetics Shop	Department Store	Dessert Shop	Diner	Donut Shop	Dumpling Restaurant	Food Truck
6	Aghapura	17.389178	78.465273	4.0	Bakery	Afghan Restaurant	Indian Restaurant	Platform	Electronics Store	Food Court	Food & Drink Shop	Food	Flea Market	Fast Food Restaurant

4.3. Choosing the correct cluster for setting up the Food Business

Using folium map we can plot the cluster on the city map.

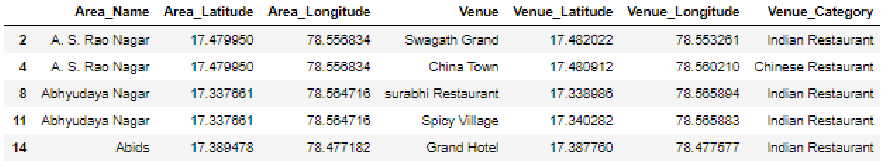


As seen from the map, cluster 4 (cyan) appears to be more centered to the city and is also dense with more locations. Assuming the population to be more at the center of the city, cluster 4 would be the best recommended zone for the new business.

4.4. Choosing best cuisines in the selected cluster for setting up the Food Business

Now, as we have narrowed down the recommended location to a zone, we need to find the best cuisines in the zone for the restaurant.

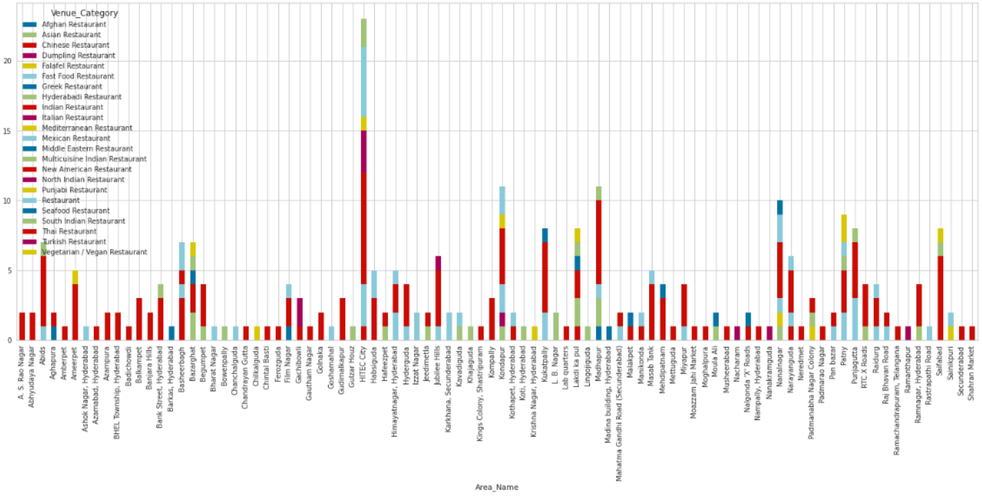
Let's filter out the venues data retrieved through foursquare API to only the venue categories of restaurants.



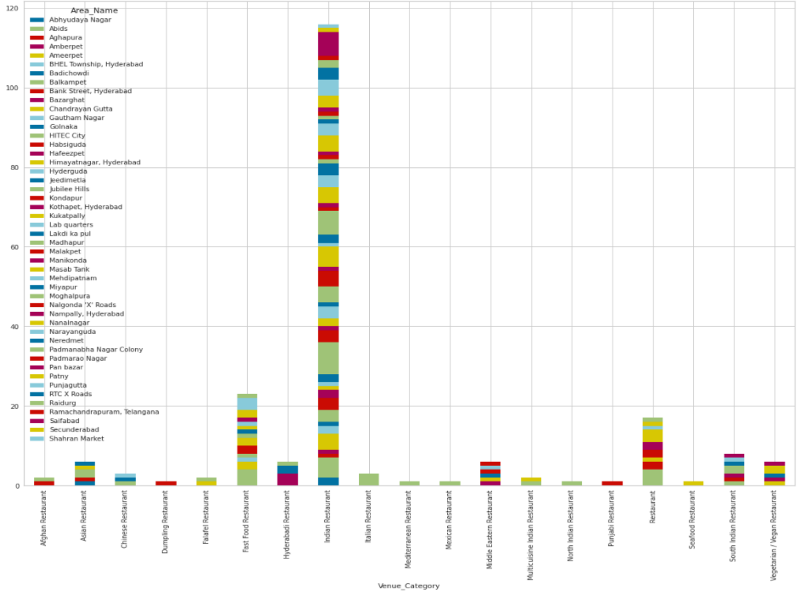
	Area_Name	Area_Latitude	Area_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
2	A. S. Rao Nagar	17.479950	78.556834	Swagath Grand	17.482022	78.553261	Indian Restaurant
4	A. S. Rao Nagar	17.479950	78.556834	China Town	17.480912	78.560210	Chinese Restaurant
8	Abhyudaya Nagar	17.337661	78.564716	surabhi Restaurant	17.338986	78.565894	Indian Restaurant
11	Abhyudaya Nagar	17.337661	78.564716	Spicy Village	17.340282	78.565883	Indian Restaurant
14	Abids	17.389478	78.477182	Grand Hotel	17.387760	78.477577	Indian Restaurant

There are total of 23 unique cuisines served in the 165 areas of the city.

Let’s visualize the cuisines available area-wise as below.



Again, filtering the venues data for only the areas in the cluster 4 (our recommended zone) and plot the graph to visualize the no. of restaurants with different cuisines in the area.



From the above bar graph, we can see top 3 cuisines in the cluster are:

- Indian Cuisine
- Chinese Cuisine
- South Indian Cuisine

5. Results and Discussion

In this study, we have extracted the area/neighborhood data from the Wikipedia page using the python packages beatifulsoup and requests. The resultant data set has a total of 165 areas. These using the foursquare API we have fetched the venues data for the 165 areas in the city. There are total of 978 venues retrieved for all the areas in the city.

Using k-means clustering we tried to group the areas based on the venue categories. For determining the optimum k value, the elbow curve, silhouette curve and inter cluster distance techniques are used. Then, plotting the clusters using folium map the most appropriate cluster is picked. Assuming that the population will be high at the center of the city, the cluster which is dense and more aligned to the city center is considered the optimal choice.

Filtering the venue categories for only restaurants, we can see the total unique cuisines in the city are 23. Further narrowing the search to the cluster picked, the number unique cuisines is 19 for the 48 areas in the cluster.

Using the bar plot, we can visualize the most common cuisines and pick the top cuisines.

6. Conclusion

Purpose of this project is to identify the best possible location and cuisine to set up a food business/restaurant in the Hyderabad city. By calculating the venues density and the alignment of the different similar venues to the center of the city we had narrowed down the search for the appropriate location (cluster). Also, the top cuisines with that cluster are identified which will help the stakeholder to decide on the type of restaurant.

The final decision on the optimal location can be made by the stakeholders considering other factors such as real-estate, population, levels of noise / proximity to major roads of the locations in the recommended cluster.