Coursera Capstone

IBM Applied Data Science Capstone

Setting Up of Department Store in Hyderabad, India

Ravi Teja

20 July 2020

1. Introduction:

1.1 Background:

Over the years, the goods available for sale to the public have increase in both range and variety and now encompass everything and anything that a customer might possibly need or want. Today it is possible to purchase clothes, toiletries, cosmetics, gardening materials, sporting goods, home appliances and others under one roof that is the Department Store. However, some stores have become renowned from a certain degree of specialization while others for offering goods at discount prices whereas most stores offer a general discount of prices at particular times of the year.

1.2 Business Problem:

The objective of this capstone project is to analyse and select the best locations in Hyderabad, India to open a new Department Store like Ikea, Metro, Walmart. Using data science methodology and machine learning techniques like K-means clustering, this project aims to provide solutions to answer the business question: Which is best place to open a Department Store with less competition and more population?

1.3 Interested Audience:

This project is particularly useful to the Property Developers, Investors and Retailers looking to open or invest in a new Departmental Store in the capital of Telangana that is Hyderabad in India.

2. Data:

2.1 To solve the problem, we will need the following data:

• List of neighbourhoods in Hyderabad. This defines the scope of this project which is confined to the Hyderabad city.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to Department Stores. We will use this data to perform clustering on the neighbourhoods.

2.2 Source to get the required data:

- The Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India consists list of neighbourhood, with a total of 200 neighbourhoods.

- Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods in Hyderabad.

- After that, we will use Foursquare API to get the venue data for those neighbourhoods.

3. Methodology:

Firstly, we need to get the list of neighbourhoods in the city of Hyderabad. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India). We will do web scraping using Python requests and BeautifulSoup package to extract the list of neighbourhood data into a dataframe. However, this is just a list of names which looks like:

```
In [5]: # create a new DataFrame from the list
        df_neighborhood = pd.DataFrame({'Neighborhood' : neighborhoodList})

        df_neighborhood.head()
```

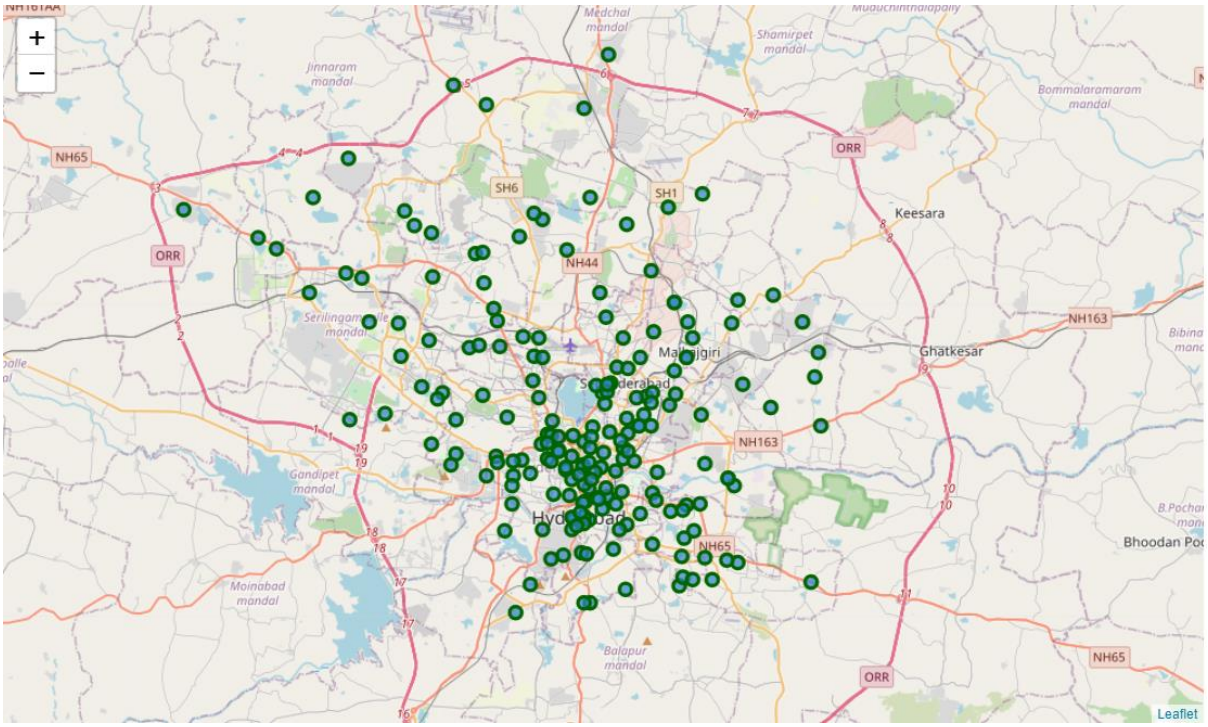
Out[5]:

	Neighborhood
0	A. S. Rao Nagar
1	A.C. Guards
2	Abhyudaya Nagar
3	Abids
4	Adibatla

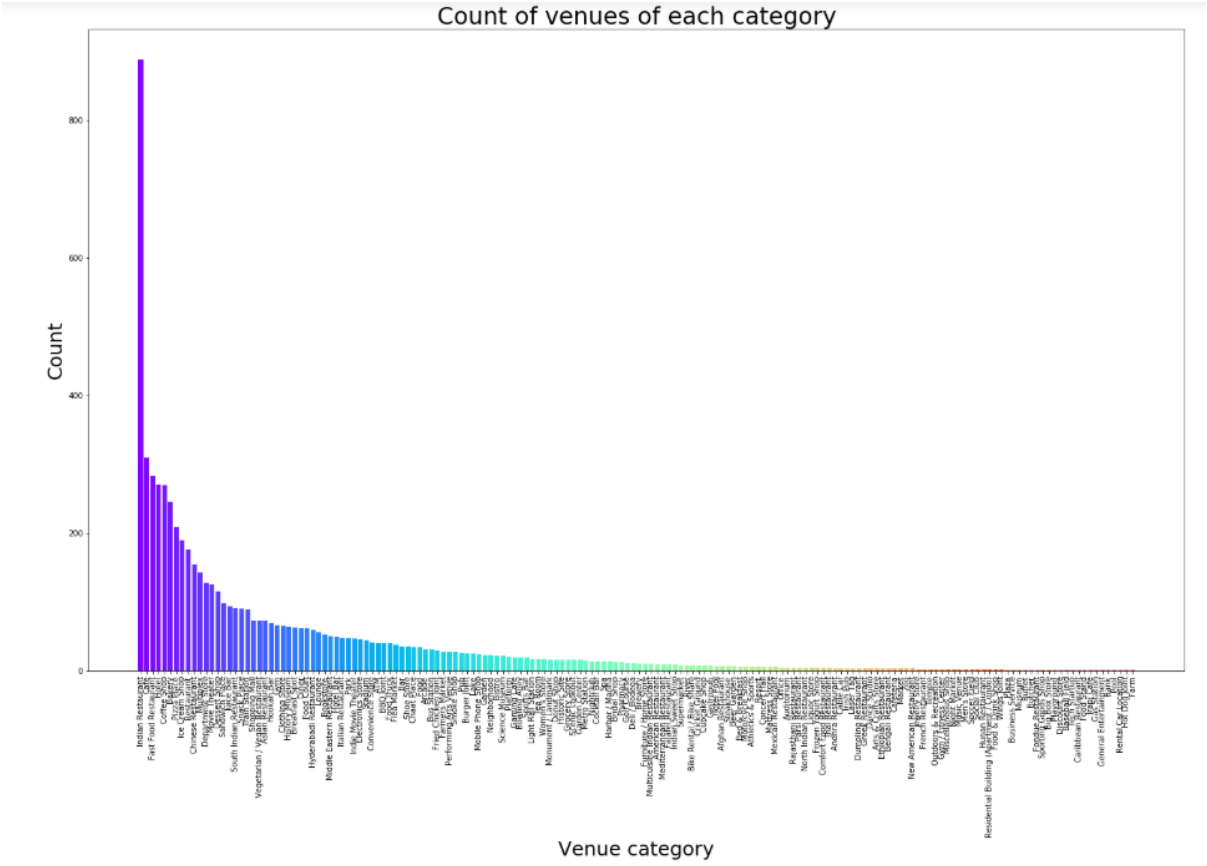
We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. We will also save this to the data frame and now we can apply concatenation for the location data frame and neighbourhood dataframe into one dataframe which looks like:

	Neighborhood	Latitude	Longitude
0	A. S. Rao Nagar	17.411200	78.50824
1	A.C. Guards	17.393001	78.45690
2	Abhyudaya Nagar	17.337650	78.56414
3	Abids	17.389800	78.47658
4	Adibatla	17.235790	78.54132

Now we can visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Hyderabad.



Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. In our case we get 168 unique categories and we can visualize it on a histogram.



Now we can analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the Department Store data, we will filter the “Department Store” as venue category for the neighbourhoods.

	Neighborhoods	Department Store
0	A. S. Rao Nagar	0
1	A.C. Guards	2
2	Abhyudaya Nagar	1
3	Abids	2
4	Adikmet	0

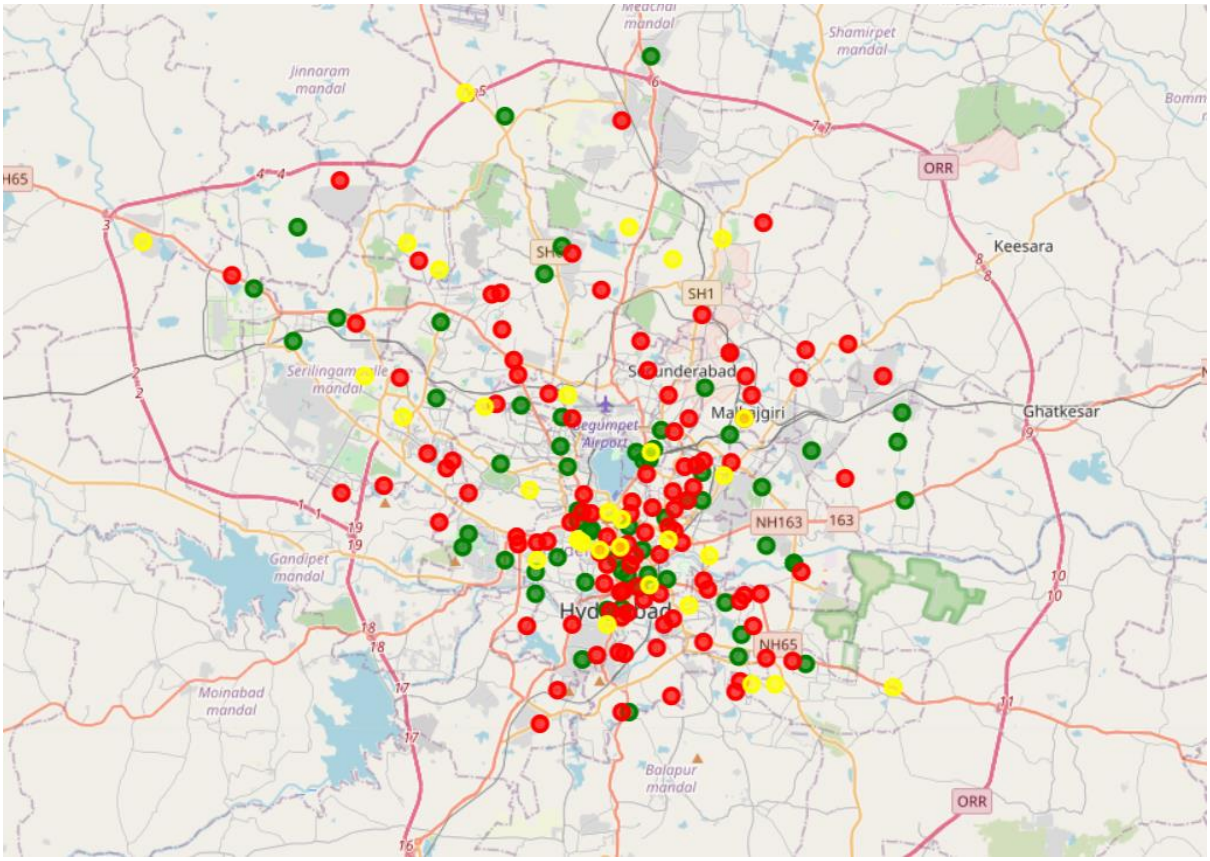
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “Department Store”. The results will allow us to identify which neighbourhoods have higher concentration of department store while which neighbourhoods have fewer number of department store. Based on the occurrence of stores in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Department Store.

	Neighborhood	Department Store	Cluster Labels
0	A. S. Rao Nagar	0	1
1	A.C. Guards	2	2
2	Abhyudaya Nagar	1	0
3	Abids	2	2
4	Adikmet	0	1
5	Afzal Gunj	1	0
6	Aghapura	1	0
7	Aliabad, Hyderabad	0	1
8	Alijah Kotla	0	1
9	Allwyn Colony	0	1

4. Results:

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Department Store”:

- Cluster 0 (green): Neighbourhood with at most 1 Department Store.
- Cluster 1 (red): Neighbourhood with no Department Store.
- Cluster 2 (yellow): Neighbourhood with at most 3 Department Stores.



5. Discussions:

Based on our analysis above, we can draw a number of conclusions that will be useful to aid property developers and investors for setting up of the Department Store. After collecting data from the Wikipedia (data of neighborhoods) & Foursquare (data of venues), we got a list of *168 different venues*. However, as we want to only focus on the data of Department Store, we created a separate DataFrame for the data of department stores. We identified that from the data, there are *91 total department stores* in central Hyderabad. Also, after further analysis we find that there are *at most 3 department stores* in some of the neighborhoods and some doesn't even have one.

Finally, we had separated the neighbourhood into three Cluster with the help of K-means which is an unsupervised Machine Learning Algorithm.

- Cluster - 0: Consists of neighbourhood with at most 1 Department Store.
- Cluster - 1: Consists of neighbourhood with no Department store.
- Cluster - 2: Consists of neighbourhoods with at most 3 Department Store.

6. Conclusion:

A good number of Department Stores are concentrated in the central area of Hyderabad city, with the highest number in Cluster - 2 and moderate number in Cluster-0. This represents a great opportunity and high potential areas to open new department store as there is very little to no competition from existing stores. Meanwhile, department stores in Cluster - 2 are likely suffering from intense competition due to oversupply and high concentration of stores. Therefore, this project recommends property developers to capitalize on these findings to open new department store in neighbourhoods in Cluster - 1 with little to no competition. Property developers with unique selling propositions to stand out from the competition can also open new department store in neighbourhoods in Cluster - 0 with moderate competition. Lastly, Property Developers, Investors and other Retailers are advised to avoid neighbourhoods in Cluster - 2 which already have a high concentration of department stores and suffering from intense competition.