Description of the Problem:

The city of Hyderabad(https://en.wikipedia.org/wiki/Hyderabad) is a global IT hub in India and is developing quite fast. Due to the young and high income work force(typically IT, SW and Engg companies spread across the city) there is a need for lot of shopping areas where the demand is quite high and the spending capacity is quite good.

How do companies or potential investors identify the areas in which they can build new malls attracting the customers. This has always been a big question. Due to the fast urbanization and people migrating to this city, there is a need for this kind of analysis and this capstone project.

Target Users:

This report will analyze the neighborhoods and recommends the areas where there is a need for the malls. Potentials investors, real estate companies, infrastructure developers are the possible target users of this report.

Data Section:

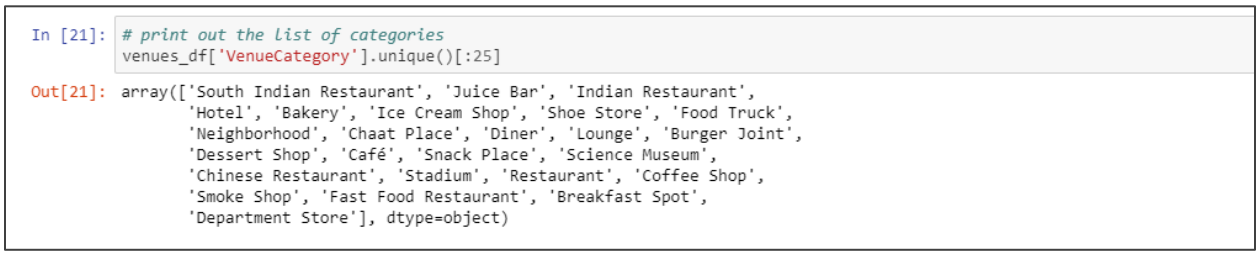
We use the list of suburbs from the city of Hyderabad from Wikipedia(https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India)

It has 54 suburbs.

The Lat-Long will be retrieved by the 'geocoder' library for these 54 neighborhoods.

We also use the Foursquare API to get the venues in these neighborhoods. Foursquare has a rich source of data and some of the categories are top categories are listed here.

```
'South Indian Restaurant', 'Juice Bar', 'Indian Restaurant','Hotel', 'Bakery',
'Ice Cream Shop', 'Shoe Store', 'Food Truck', 'Neighborhood', 'Chaat Place', 'D
iner', 'Lounge', 'Burger Joint', 'Dessert Shop', 'Café', 'Snack Place', 'Scienc
e Museum', 'Chinese Restaurant', 'Stadium', 'Restaurant', 'Coffee Shop', 'Smoke
Shop', 'Fast Food Restaurant', 'Breakfast Spot',       'Department Store']
```



```
In [21]: # print out the list of categories
         venues_df['VenueCategory'].unique()[:25]

Out[21]: array(['South Indian Restaurant', 'Juice Bar', 'Indian Restaurant',
                'Hotel', 'Bakery', 'Ice Cream Shop', 'Shoe Store', 'Food Truck',
                'Neighborhood', 'Chaat Place', 'Diner', 'Lounge', 'Burger Joint',
                'Dessert Shop', 'Café', 'Snack Place', 'Science Museum',
                'Chinese Restaurant', 'Stadium', 'Restaurant', 'Coffee Shop',
                'Smoke Shop', 'Fast Food Restaurant', 'Breakfast Spot',
                'Department Store'], dtype=object)
```

Methodology:

1. For this analysis, we need to find the list of neighborhoods in Hyderabad. This is accessible from here. https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India

2. This data is in the format of tables and we use the Beautifulsoup package to scrape the data from the html page. Using this package, we get the list of neighborhoods in the city of Hyderabad.

3. In order to get the latitude and longitude of these Neighborhood points, we use the geocoder library to convert these addressed into lat-long coordinates.

4. Now we have the Neighborhood data and the lat-long coordinates of this data available.

5. Next, we load this data into pandas dataframes for better handling. Using the Folium API, we visualize these Neighborhoods in the map.

6. Using the Foursquare API, we try to get the top 100 venues within the 200 meters. Foursquare API is called using the 'explore' method and the necessary parameters (CLIENT ID, SECRET ID, API version etc.).

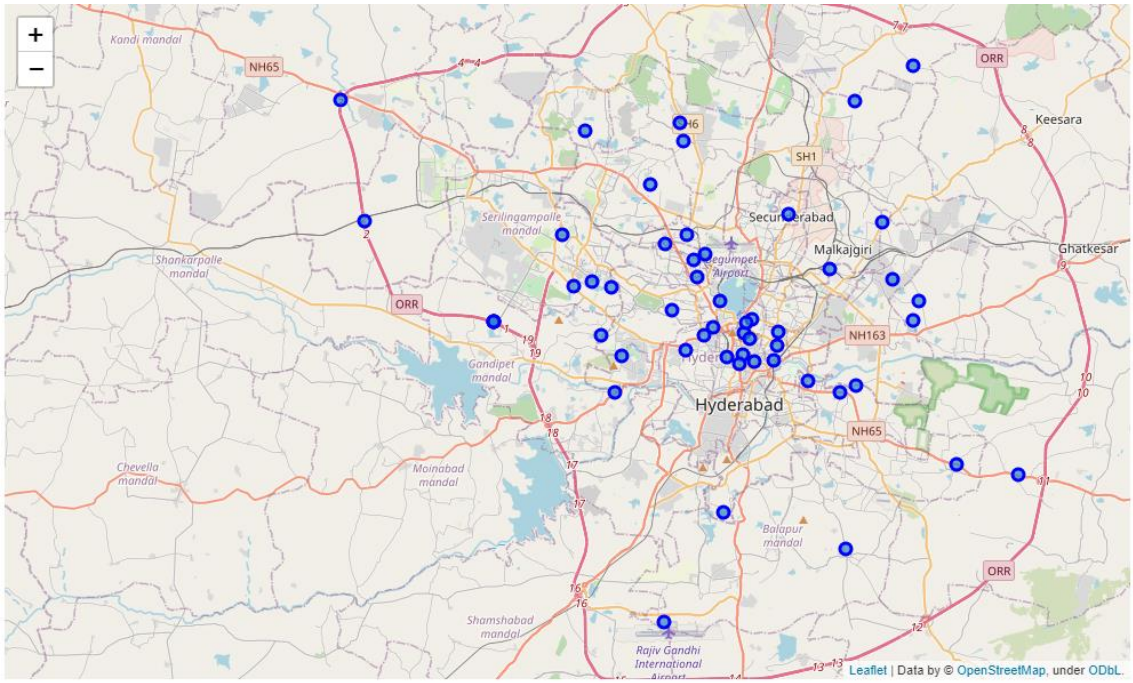
7. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.

8. We will analyze each neighborhood by grouping the rows by neighborhood 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 analyzing the "Shopping Mall" data, we will filter the "Shopping Mall" as venue category for the neighborhoods.

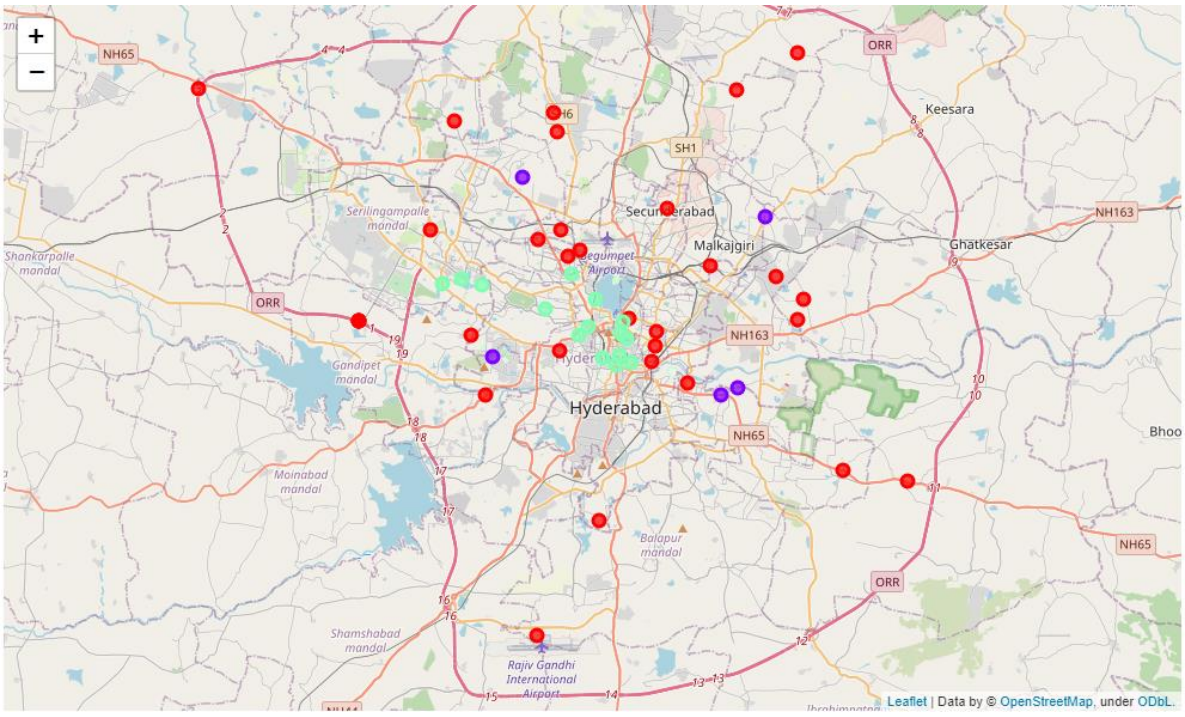
9. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised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

Results:

Th initial plot of the Hyderabad city with the neighborhood superimposed is mentioned below.



Later post the KNN clustering algorithm, the neighborhoods are clustered in the following ways.



Discussion & Conclusion:

As you can see from the above picture most of the malls are in cluster 0(red color) and the cluster 1(green color) has a low concentration of malls and ideally that would be a better place to plan for any mall.

However, we have not considered any other details like the proximity of the livable area and residential area, income levels into consideration.

With some other details baked into this solution, this could become a comprehensive guide to the overall opportunities in the Metropolitan area of the City of Hyderabad.