**Capstone Project - Battle of the Neighborhoods in the city of**

**Hyderabad, Telangana, India**

Sandeep Kumar Vengala

**Table of contents:**

* Introduction: Business Problem
* Data
* Methodology
* Analysis
* Results and Discussion
* Conclusion

**Introduction: Business Problem Statement**

In this project we will try to find an optimal location for opening a restaurant. Specifically, this report will be targeted to stakeholders interested in opening any food-joint/restaurant in Hyderabad, Telangana India.

Since there are lots of restaurants in Hyderabad, we will try to detect locations that are not already crowded with restaurants. We are also particularly interested in neighborhoods which are similar to the best neighborhoods in terms of amenities. We would also prefer locations as close to city centre as possible, assuming that first two conditions are met.

We will use our data science technical expertise to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

**Data**

Based on definition of our problem, factors that will influence our decision are:

* number of existing restaurants in the neighborhood (any type of restaurant)
* variety of amenities in the neighborhood, if any
* distance of neighborhood from city centre

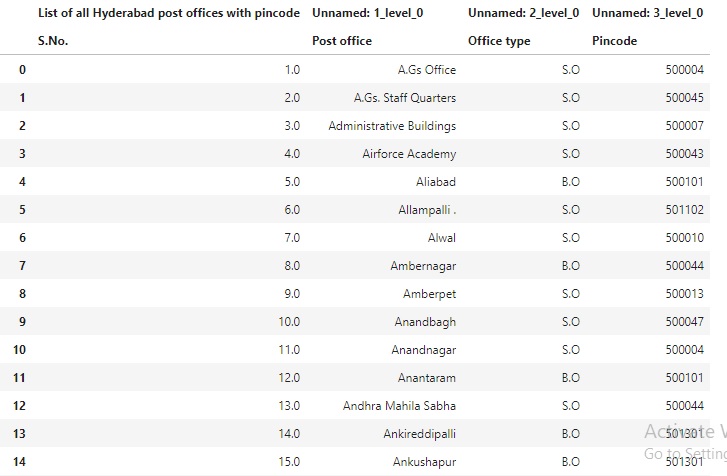
Neighborhoods have been defined based on names of post offices as given in the form of a table in the website: <http://pincode.india-server.com/cities/hyderabad/>

Following data sources will be needed to extract/generate the required information:

* names of all neighbourhoods (post office names) will be extracted from the above-mentioned website using Pandas package in python
* approximate addresses of centres of these neighbourhoods will be obtained using GeoPy Geocoder package in python
* number of restaurants and their type and location in every neighbourhood will be obtained using Foursquare API

**Data Cleaning:**

The table present in the website was procured using the Pandas ‘read\_html()’ method:



**DataFrame gathered from website**

The DataFrame was then cleaned to only include names of the post offices which are the names of the neighborhoods:



**DataFrame after dropping unwanted columns**

The names of neighbors were then extracted from the above DataFrame and a new dataframe was created consisting of the neighborhood names and their latitude and longitude.



**DataFrame after adding latitude and longitude**

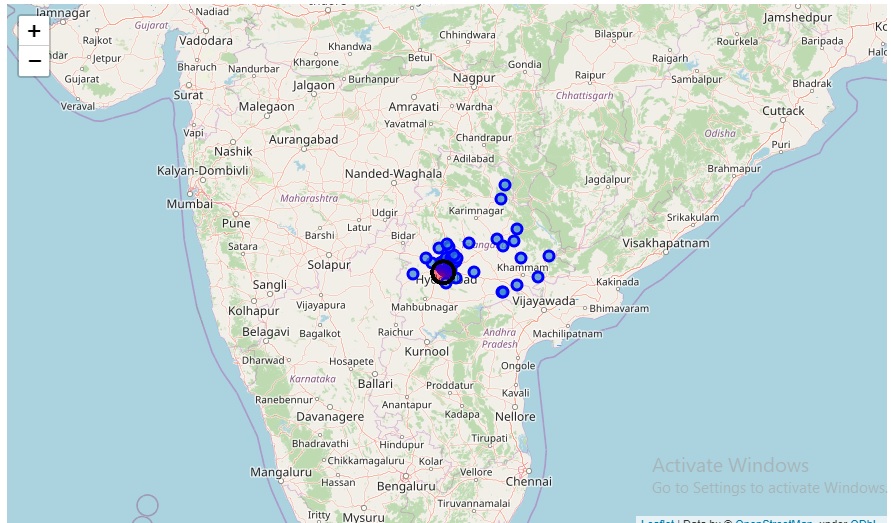
**Methodology**

In the cells below the following format of the project will be followed in order to explore the neighborhoods in Hyderabad and try to analyze and suggest an optimal location for opening a food-joint/restaurant.

1. Get list of all neighborhoods from the website listing names of postal offices all over Hyderabad.
2. Obtain latitudes and longitudes for all these neighborhoods.
3. Check for any outliers, i.e. postal office names which aren't within 25Kms from the centre of Hyderabad and remove these data points.
4. Send GET requests to Foursquare API to get list of venues (maximum 100) within the vicinity of all neighbors (vicinity is defined within a radius of 1000m from the neighborhood).
5. Remove all neighborhoods which do not have more than 10 venues in their vicinities since there isn't enough data for accurate clustering using K-Means Clustering.
6. One-hot encode the data and feed the dataframe to K-Means clustering Algorithm to form 10 clusters from the list of available neighborhoods.
7. Create visualizations using the cluster number of all neighborhoods to better understand the data create valuable insights. Here, we try to find out the best neighborhoods amongst the available list of neighborhoods and try to identify its cluster.
8. Again, send GET requests to Foursquare API to get list of venues (maximum 100) of 'section = food' to obtain only venues related to food joints within the vicinity of all selected neighborhoods (vicinity is defined within a radius of 1000m from the neighborhood).
9. Create visualizations to find out which neighborhoods have maximum food joints and which ones have maximum types of food joints.
10. Suggest optimal locations for opening a restaurant based on all the statistical analysis done.

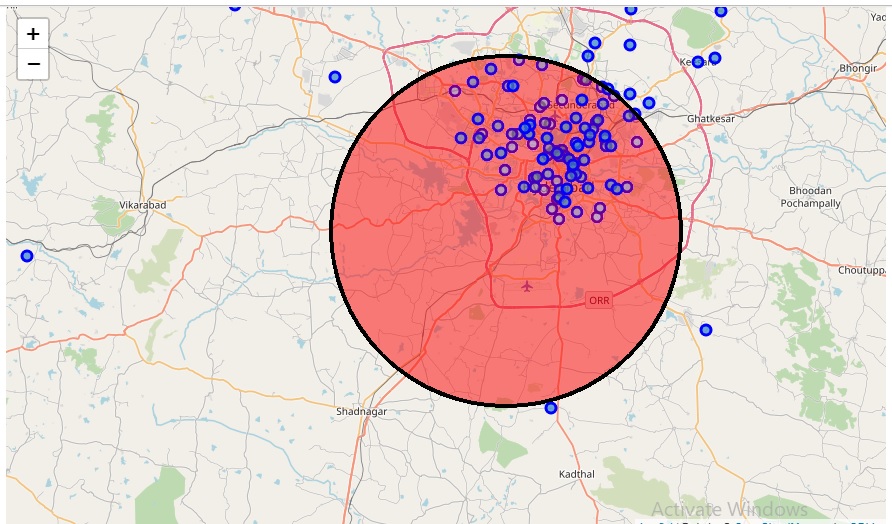
**Data Visualization**

The obtained latitude and longitude values for all the neighborhoods are then used to plot all postal offices/neighborhoods in the map of Maharashtra. It is observed that a lot of these are far away from the centre of the city. The circle with a 25km radius from the centre of the city has been marked for reference.

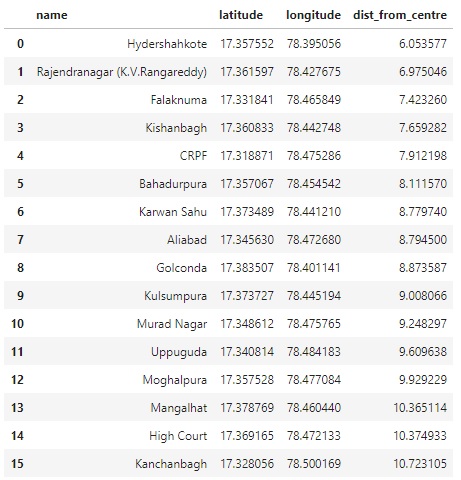


**Map with all postal offices**

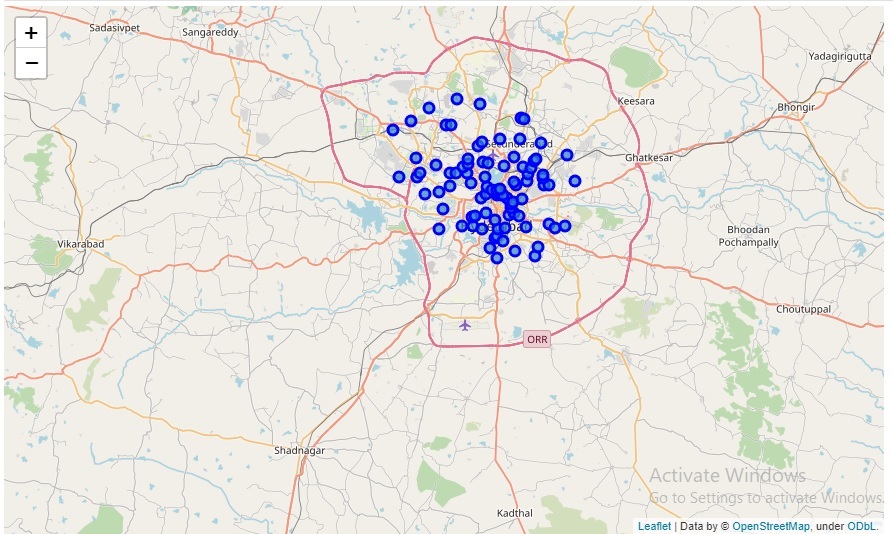
The previous map has been appropriately zoomed to highlight the relevant neighborhoods in Hyderabad.



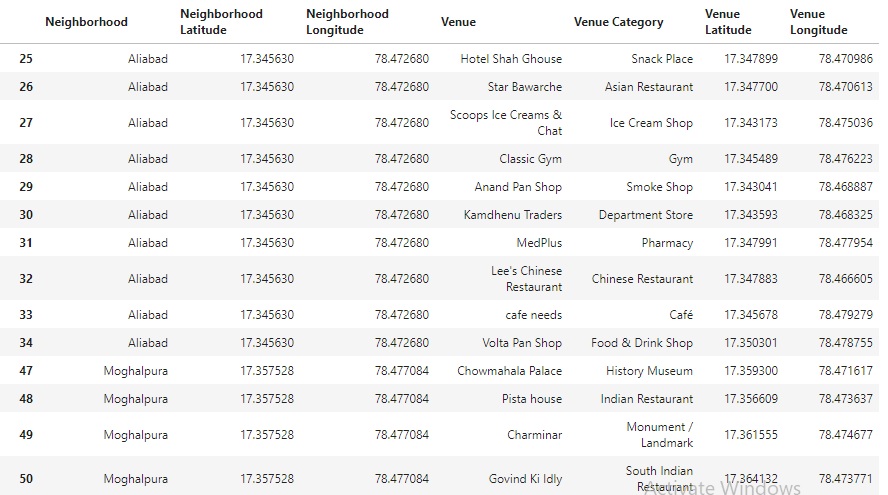
Thus, the dataframe is shortened to only those neighborhoods which are within 25Kms from the city centre and the final dataframe is as follows:



The map of Hyderabad with relevant neighborhoods is as follows:



The dataframe created after getting the response from the Foursquare API as nearest venues within 1000m from a neighbourhood is as follows:



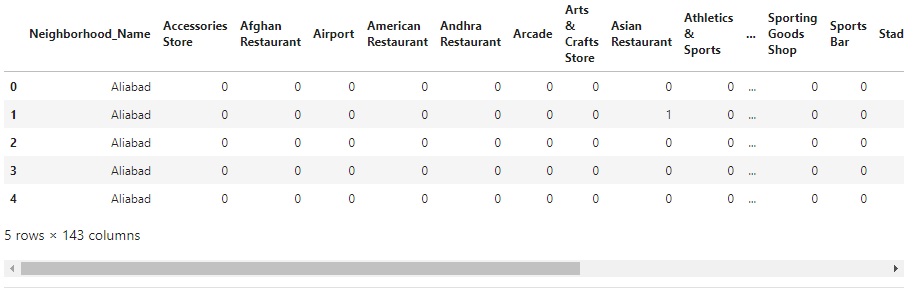
The above table is grouped by neighborhood names and the count of every column is taken to find out the number of venues present in every neighborhood. Neighborhoods which have less than 10 venues are discarded since there isn’t enough data to cluster these neighborhoods properly.

The resulting table further filters the possible neighborhoods and the final neighborhoods which shall be grouped into clusters.

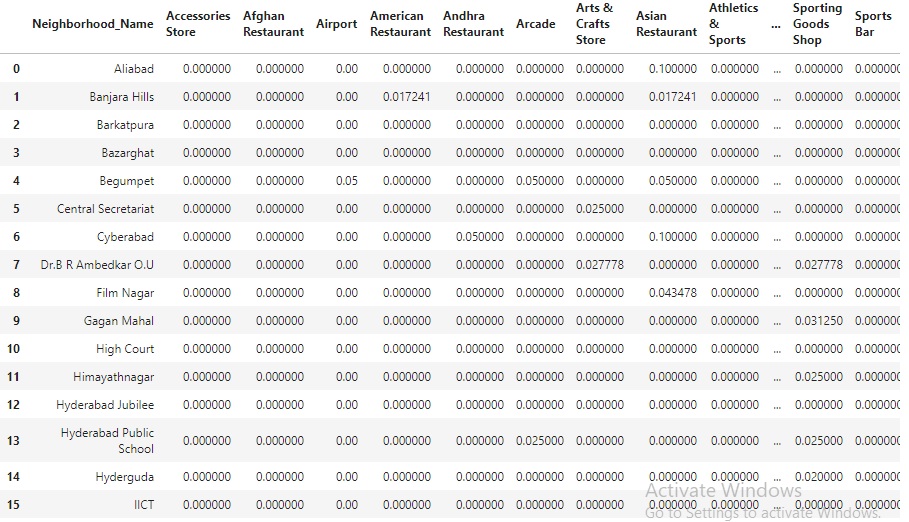
The resulting table is one-hot encoded in order to create all unique ‘Venue Category’ values into columns with a value of 1 or 0.

One(1) meaning that it falls under that category.

The table created after one-hot encoding is as follows:



This dataframe is then grouped by the neighborhood name again and the mean of every column is taken as the values for every row after grouping. The resulting table is as follows:

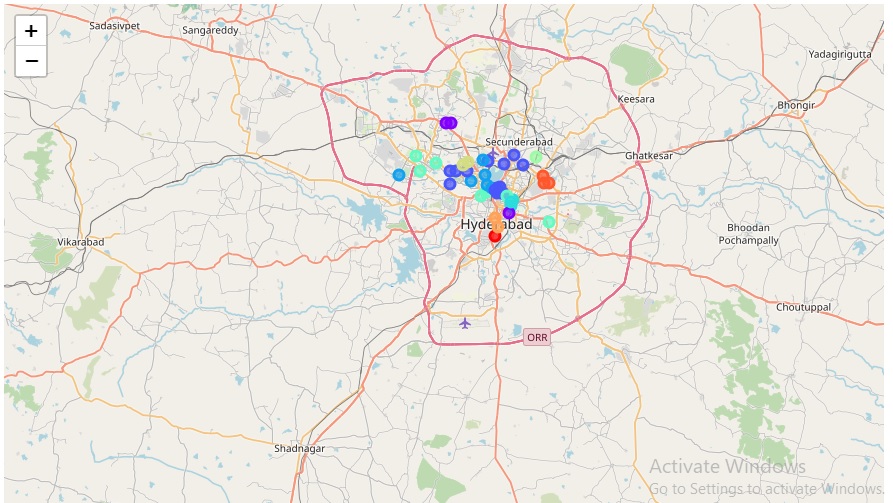


The above dataframe is fed to the K-Means clustering algorithm after dropping the ‘Neighborhood\_Name’ column to create clusters amongst these 41 neighborhoods.

A dataframe is created which consists of the names of the neighborhoods, their latitudes and longitudes as well as their assigned cluster number. The dataframe is as follows:

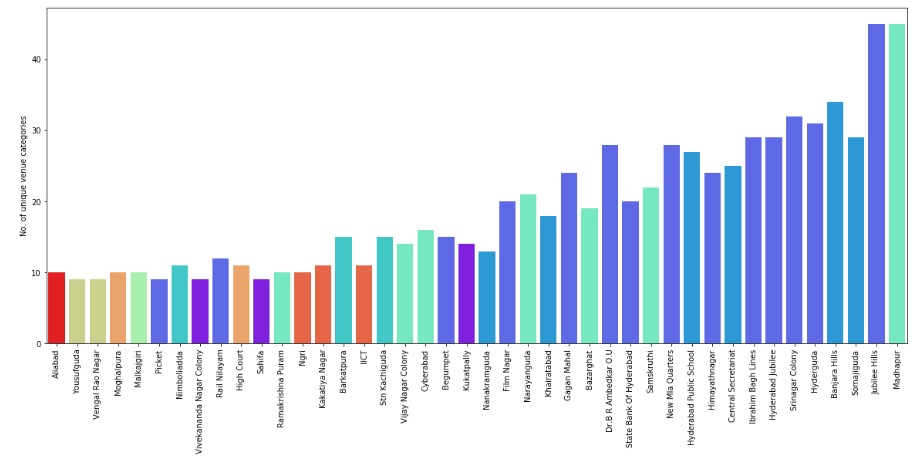


The map of Hyderabad is plotted with circle markers for every neighborhood which are color coded according to the cluster number of every neighborhood. The map is as follows:



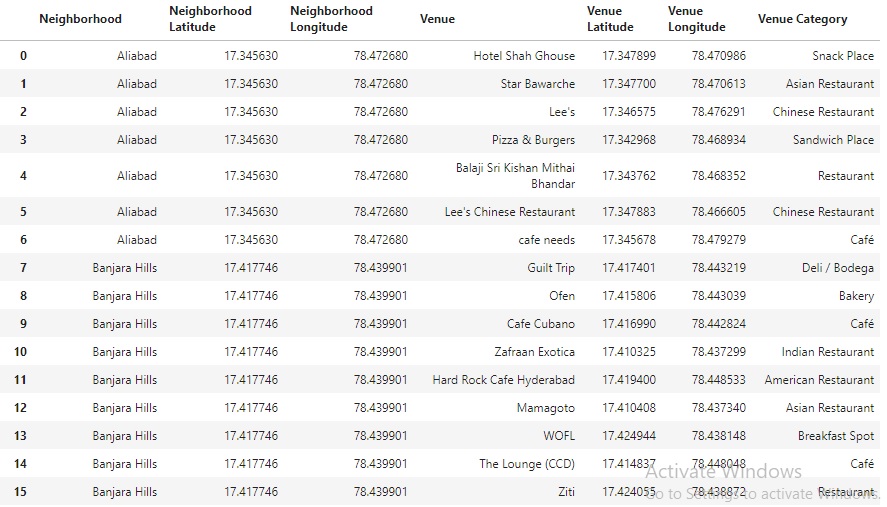
**Analysis**

A bar graph was plotted using the number of unique venue categories along the Y-axis with the corresponding neighborhood name on the X-axis. This is done to estimate the quality of a neighborhood based on the assumption that the best neighborhood will have the highest diversity of amenities which in this case is the number of unique venue categories. The bar plot has been color coded such that all bars of the same color correspond to neighborhoods in the same cluster. The bar graph is as follows:

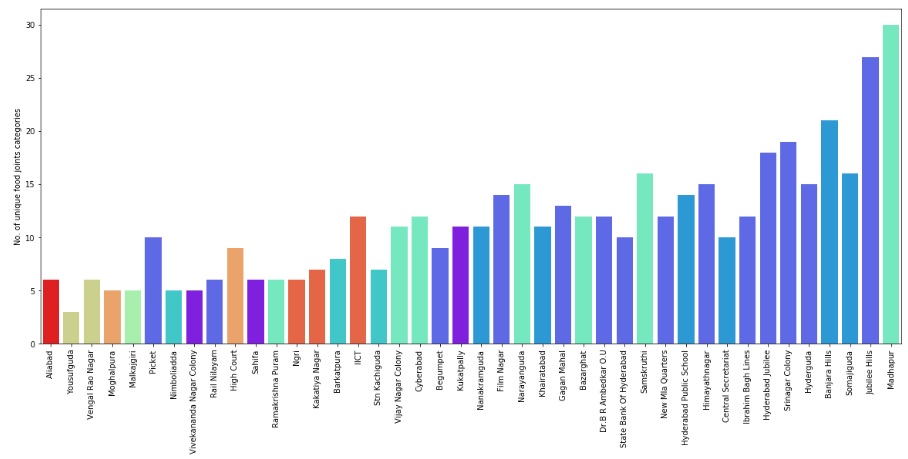


From the above plot, we can conclude that clusters pertaining to Jubilee Hills and Madhapur, i.e. cluster number 2 and 5 respectively can definitely be considered the best clusters since they consist of neighborhoods with the highest diversity of amenities.

Now, in order to judge neighborhoods on the basis of food-specific venues, the Foursquare API is again used to request for venues pertaining to foods, for all the final candidate neighborhoods and resulting dataframe is as follows:



Now, the bar graph is plotted with the total number of food-joints/restaurants in a neighborhood on the Y-axis with the corresponding neighborhood name on the X-axis. This is done in order to estimate the saturation of a neighborhood with respect to the number of food-joints present in the neighborhood since we want to detect locations that are not already crowded with restaurants. The bars are color-coded according to the cluster number here as well. The bar graph is as follows:



**Results and Discussion**

Based on the bar graph showing the number of unique venue categories for each neighborhood, we concluded that clusters pertaining to Jubilee Hills and Madhapur, i.e. cluster numbers 2 and 5 respectively can definitely be considered the best clusters since they consist of neighborhoods with the highest variety of amenities.

Thus, prospective stakeholders would like to open up their restaurant in similar neighborhoods since they have the largest customer base and also have a huge demand. However, it is possible that these neighborhoods might be saturated with venues. Therefore, we also look at the plot showing the number of unique venues for each and every neighborhood. Here we see that the while the best neighborhoods in terms of diversity of amenities such as Jubilee Hills, Banjara Hills, Madhapur etc. are saturated with a large number of venues, there exist neighborhoods which are part of the same cluster yet have lesser number of venues. These neighborhoods might be optimal for opening up restaurants since they fall to the same cluster and are therefore similar to neighborhoods with highly diverse amenities yet have a smaller number of venues.

Instead of looking at the total number of venues of a certain neighborhood to determine the saturation, it would be better to simply look at the total number of food joints / restaurants in that neighborhood. Thus, we look at the bar graph showing the number of venue categories pertaining to food for each neighborhood. We can see that our initial assumption is still correct and neighborhoods in cluster 2 and 5 have the highest variety of food joints are thereby considered the best neighborhoods.

So, the optimal location to open up restaurants in Hyderabad would be neighburhoods in the **best clusters** i.e. clusters 2 and 5 with the **least number of existing food joints**. The ideal candidate neighborhoods can be seen from the plot of number of unique venues related to food for every neighborhood. Here we see that the neighborhoods with the least number of food joints belonging to cluster 2 and 5 are:

1. Rail Nilayam
2. RamaKrishna Puram
3. Picket
4. Begumpet
5. Cyberabad

**Conclusion**

The purpose of this project was to identify areas/neighborhoods in Hyderabad close to the city-centre with low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for opening of a food-joint/restaurant. By calculating venue density distribution from Foursquare data we have first identified general clusters that justify further analysis and created groups of similar neighborhoods, and then generated extensive visualizations to justify how good a certain cluster is based on characteristics (diversity of amenities) of the neighborhoods in the clusters.

Then the Foursquare data was again used in order to find venues pertaining to food for all the neighborhoods to determine the saturation of a neighborhood with respect to number of food joints. **Optimal locations were identified as those neighborhoods which are part of the best clusters (having neighborhoods with most diverse amenities) yet have the least number of food joints.** The optimal locations obtained are:

1. Rail Nilayam
2. RamaKrishna Puram
3. Picket
4. Begumpet
5. Cyberabad

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighborhoods and their locations, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.