

# **Capstone Project – The Battle of Neighborhoods (Week 2)**

Applied Data Science Capstone

## ***Project Report***

***By***

***Prithvijit Gupta***

***29.05.2019.***

## Table of Contents

Applied Data Science Capstone.....	1
A. Introduction: .....	3
B. DATA required:.....	5
C. Methodology:.....	9
D. RESULT / Final Scoring: .....	15
E. ERROR/ EXCEPTION handling:.....	17
F. Discussions: .....	19
G. Conclusion:.....	19

## A. Introduction:

Our client is an EVENT MANAGEMENT entity, and specializes in conducting corporate seminars/conferences.

The project deals with the problem of selecting a suitable area within a host of cities to set up an office. The client wants to buy/ take rent a commercial office at a suitable location in any of the four metro cities in India, i.e. in any of the following cities:

1. Mumbai,
2. Kolkata,
3. New Delhi, or
4. Chennai.



It plans to open up an office which will meet the following parameters/criteria:

Sl. No.	Criteria	Remarks
1	Commercial property rates	The office has to be located in an area where the commercial property rates are preferably low.
2	High number of restaurants	Its customers may have variety of choices for their snacks/ food in case hotel menu is not to their liking.
3	High number of hotels	The client can arrange for both conferences and lodging for the customers.
4	Distance from city airport	The customers may save time in transit

While there are various real estate websites providing the data on commercial properties and various other websites providing location of restaurants and hotels, the client needs to find an optimum location meeting the above 4 criteria.

Such a website or other resource is not readily available. Even if the client scrapes through websites and other sources of information, gathering, aggregating and processing such raw data locating such an area/ location will not be possible.

How will the management decide and finalize such a location ?

Similar situations and problems, as discussed in the introduction, are frequently faced by organizations and corporate houses. How to deal with the such problems ?

### *SOLUTION !!!*

Such problems and situations can be dealt through :

Applied Data Science

The Project will aim to provide a solution based upon  
the USER INPUTS.

### **Target Audience:**

The project will be helpful to

1. Organizations,
2. Corporate houses,
3. Individuals,

who would like to decide and choose a location for setting up a specified facility meeting certain criteria, or is constrained by certain factors.

### **Project Goal:**

To enable the user to take a decision by scoring each location of the chosen city on the basis of the parameters/ criteria and thereby, trying to quantify the result/ output through scoring.

## B. DATA required:

It is to be noted that the project is **USER INTERACTIVE** and the user will have the option of choosing preferences, **including the source of data**.

1. **Choose a city** from the 4 metropolitan cities in India - namely:
  - Mumbai,
  - New Delhi,
  - Kolkata, and
  - Chennai.
2. **Choose** ways of selecting the **database**, namely:
  - Scrape through website and take database from the site :([www.magicbricks.com/](http://www.magicbricks.com/))
  - Use existing database.

In case the website **may not be available** or **block the user** from scrapping, the program will automatically **redirect** the user to the existing database.

3. **Choose the type of property**, i.e:
  - Buy a property under construction,
  - Buy a readily available property, and
  - Take a property on lease rentals.
4. **Assigning Weights** to all the 4 parameters/ criteria on a scale of 100 based upon the preference of the user ('0' being of no importance and '100' being of maximum importance).

### I. The user will need the following data:

Sl. No.	DATA Required
1	Commercial property rates of different areas in a city.
2	Number of restaurants available in a locality/area in the city
3	Number of hotels available in a locality/area in the city.
4	Distance of a location/ area from the city airport.

### II. Source of Data:

The following table illustrates the source of data sourced for this project:

DATA Required	DATA Source
Commercial property rates of different areas in a city.	<b>Option I:</b> Use <b>Existing database</b> already loaded in the notebook. <b>Option II:</b> Top 25 commercial zones for the selected city from

	website ( <a href="http://www.magicbricks.com">www.magicbricks.com</a> ). The Notebook program uses the “Beautiful Soup” package for web-scraping.
Number of restaurants available in a locality/area in the city	For meeting both the criteria, the following are used:  1. The <b>Geolocator</b> library. 2. The <b>Foursquare Location</b> data.
Number of hotels available in a locality/area in the city.	
Distance of a location/ area from the city airport.	The <b>Geolocator</b> library.

### III. Description and using the Data:

This section deals with how the data arrived from the above mentioned sources have been utilized for the project.

#### (i) Commercial property rates of different areas in a city.

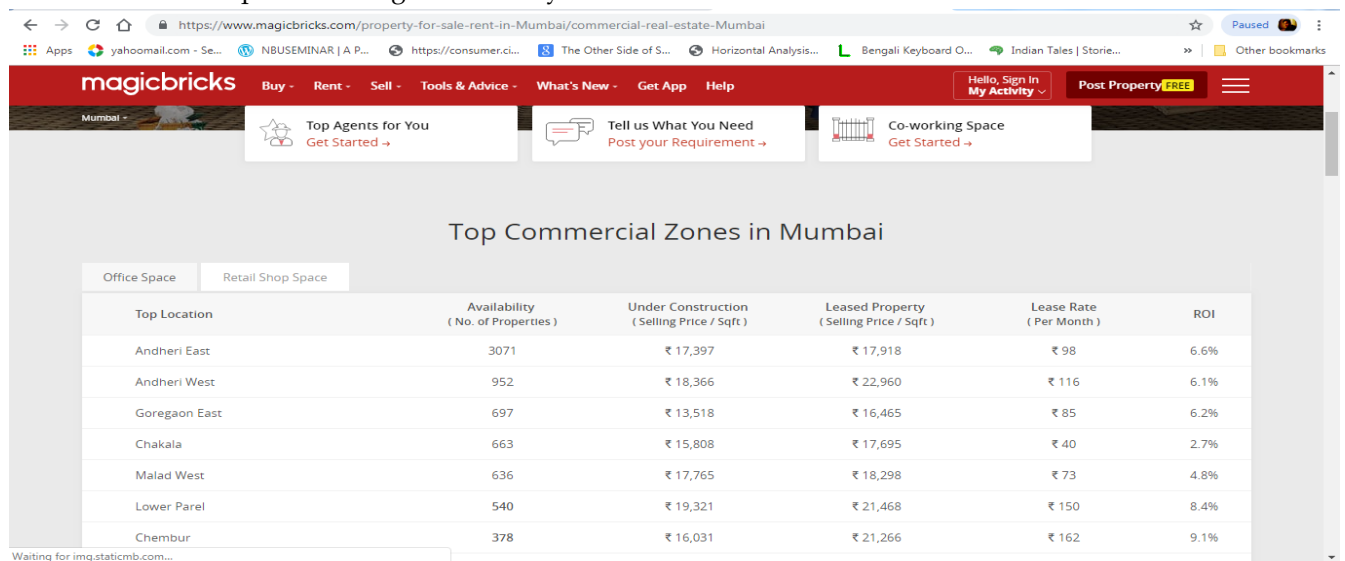
(a) **Option I:** Use **Existing database** already loaded in the notebook.

The client has the option to use the existing databases for the 4 cities, which shows the top 25 commercial zones for each city.

The source of the existing database is the website: [www.magicbricks.com](http://www.magicbricks.com).

(b) **Option II:** Top 25 commercial zones for the selected city from website ([www.magicbricks.com](http://www.magicbricks.com))

The client had the option of using data directly from this website. A screen-shot of this website :



The screenshot shows the Magicbricks website interface. The main heading is 'Top Commercial Zones in Mumbai'. Below this, there are two tabs: 'Office Space' and 'Retail Shop Space'. The 'Office Space' tab is selected. The table below lists the top 25 commercial zones in Mumbai, categorized by availability, under construction, and leased property. The table has six columns: Top Location, Availability (No. of Properties), Under Construction (Selling Price / Sqft), Leased Property (Selling Price / Sqft), Lease Rate (Per Month), and ROI.

Top Location	Availability ( No. of Properties )	Under Construction ( Selling Price / Sqft )	Leased Property ( Selling Price / Sqft )	Lease Rate ( Per Month )	ROI
Andheri East	3071	₹ 17,397	₹ 17,918	₹ 98	6.6%
Andheri West	952	₹ 18,366	₹ 22,960	₹ 116	6.1%
Goregaon East	697	₹ 13,518	₹ 16,465	₹ 85	6.2%
Chakala	663	₹ 15,808	₹ 17,695	₹ 40	2.7%
Malad West	636	₹ 17,765	₹ 18,298	₹ 73	4.8%
Lower Parel	540	₹ 19,321	₹ 21,468	₹ 150	8.4%
Chembur	378	₹ 16,031	₹ 21,266	₹ 162	9.1%

The Notebook program uses the “Beautiful Soup” package for web-scraping.

- Either of the options will result in creating a data-frame providing details of the top 25 commercial zones in the city.
- In case the website refuses to connect for web-scraping, the programme automatically redirects the user to the existing database for the chosen city.
- Difference between the 2 options:

While the option of using web-scraping through Beautiful Soup will result in updated data available on the website, the data available by using the existing database option is as on May 2019 and cannot be changed, hence not dynamic. The program has provided both the options as web-scraping may always not be possible as the website sometimes blocks scraping for security reasons.

- A sample of the dataframe created (for city: Kolkata) is as below:

```
For running on existing database, enter 'A',
For scrapping through website, enter 'B'

Enter your choice B

You have decided to use the scrape through the website
https://www.magicbricks.com/property-for-sale-rent-in-Kolkata/commercial-real-estate-Kolkata
```

[7]:

	Location	No_of_prop	SP_under_cons	SP_leased_prop	Rentals_per_month	ROI
1	Salt Lake City	1489	7460	8109	49	7.3%
2	Rajarhat	329	5353	5956	25	5%
3	New Town	168	4616	5905	44	8.9%
4	Park Street	151	22004	22664	85	4.5%
5	Kalighat	130	7650	8138	52	7.7%
6	A J C Bose Road	106	15500	15965	80	6%
7	Ballygunge	103	10031	11146	39	4.2%
8	Kasba	101	7403	8226	38	5.6%
9	Gariahat	95	20240	22000	41	2.2%
10	Camac St-Park Street area	85	19500	20085	98	5.8%
11	Salt Lake City Sector 2	83	7651	8502	18	2.6%

#### Columns description of the above table:

1. **Location** : Top Commercial Areas/ locations within the city.
2. **No\_of\_prop** : Number of commercial properties available.
3. **SP\_under\_cons** : Selling price of properties under construction.
4. **SP\_leased\_prop** : Selling price of readily available properties.
5. **Rentals\_per\_month** : Existing rates if properties are taken on rent.
6. **ROI** : Return on investment if the property is held for investment purpose.

PS : All the prices/ rates are in *Rs.per sq.ft*

#### USAGE of the above data:

The above data-frame thus arrived will be used to rank/ score the locations based upon the type of property ( *to be chosen by the user*).

(ii) Number of restaurants & hotels available in a locality/area in the city.

For meeting both the criteria, the following are used:

1. The **Geolocator** library.
2. The **Foursquare Location** data.

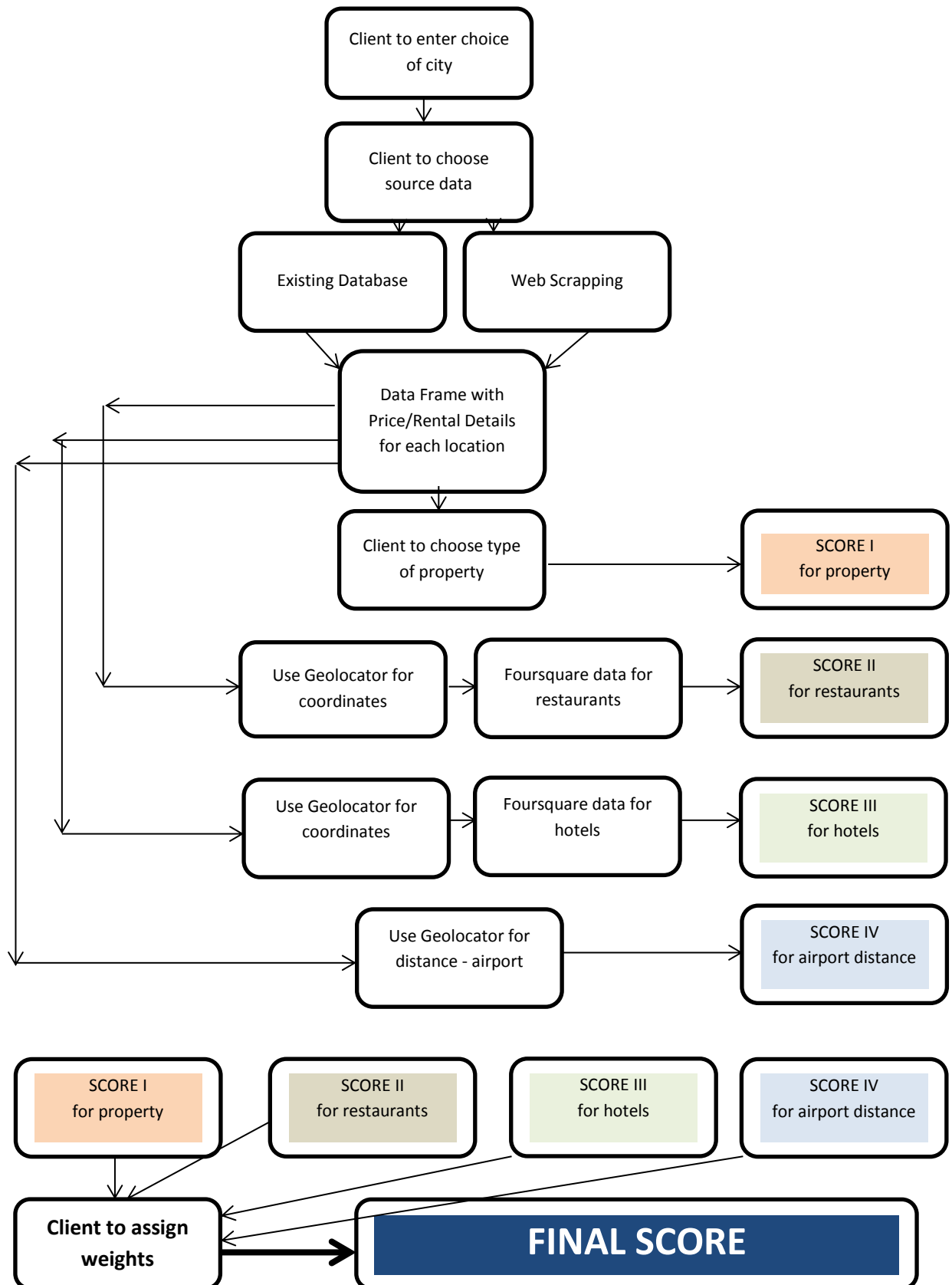
#### USAGE of the above data:

The Geolocator library.	The Foursquare Location data.
Get the coordinates of different commercial locations within the city, helping in: <ol style="list-style-type: none"><li>a. Providing input for Folium Map.</li><li>b. Providing input for Fousquare location data to get venues.</li><li>c. Providing input to calculate distance of each location from the city airport, thus enabling to rank/ score each location.</li></ol>	<p>Get the list of venues and venue categories for each location within a city. Input from the Geolocator library is used for the same.</p> <p>This will be used to determine the number of hotels and restaurants for each location, thus enabling to rank/ score each location.</p>



## C. Methodology:

### Graphical Representation of the model:



## Process Flow/ Methodology:

### I. INITIALIZING:

- Importing and initializing the necessary libraries.
- Creating back-up database of cities which will be used/ routed directly in case website scraping is not available.
- Creating necessary lists and dictionaries, including the list if errors which might occur in case web scraping are not allowed by the website.

### II. TAKING USER INPUTS:

- Prompt the user to enter the choice of city.

**Prompting the user to input his/her choice of city**

**In case the user inputs a wrong entry, the programme will direct the user till a valid choice is chosen.**

```
[*]: print("\n\nAvailable Cities\n\n" , city_list)
city = input("\n\nEnter Your City:").lower()
while city not in city_list:
    print("\n\nINVALID ENTRY ..... TRY AGAIN")
    city = input("\n\nEnter Your City:").lower()
else:
    print("\n\nYou have chosen :\n\n",city)
```

Available Cities

['kolkata', 'chennai', 'new delhi', 'mumbai']

Enter Your City:

- Similarly prompting the user to choose his/her choice of **data source**

In case the website refuses to connect for web-scraping, the programme automatically redirects the user to the existing database for the chosen city. **Errors/ exceptions are handled.**

The data gathered are then subject to Data-cleaning and Data-wrangling and thus creating a dataframe with list of areas/ locations along with property rates

- Similarly prompting the user to choose his/her choice of **property**.

1. For taking property on rent, enter "PR"
2. For buying readily available property, enter "PS"
3. For buying property under construction, enter "PC"

### III. ANALYZING PORPERTY RATES AND SCORING LOCATIONS : CRITERIA - 1

The properties rates are analyzed with the objective of ranking/ scoring each area/ location based on the price/ rentals for the type of property chosen by the user

**Let us suppose that the client has chosen the following:**

1. **City:** Kolkata
2. **Choice of property:** Property on rent.

**The property with the lowest rentals will have the highest score and the resultant dataframe is as below:**

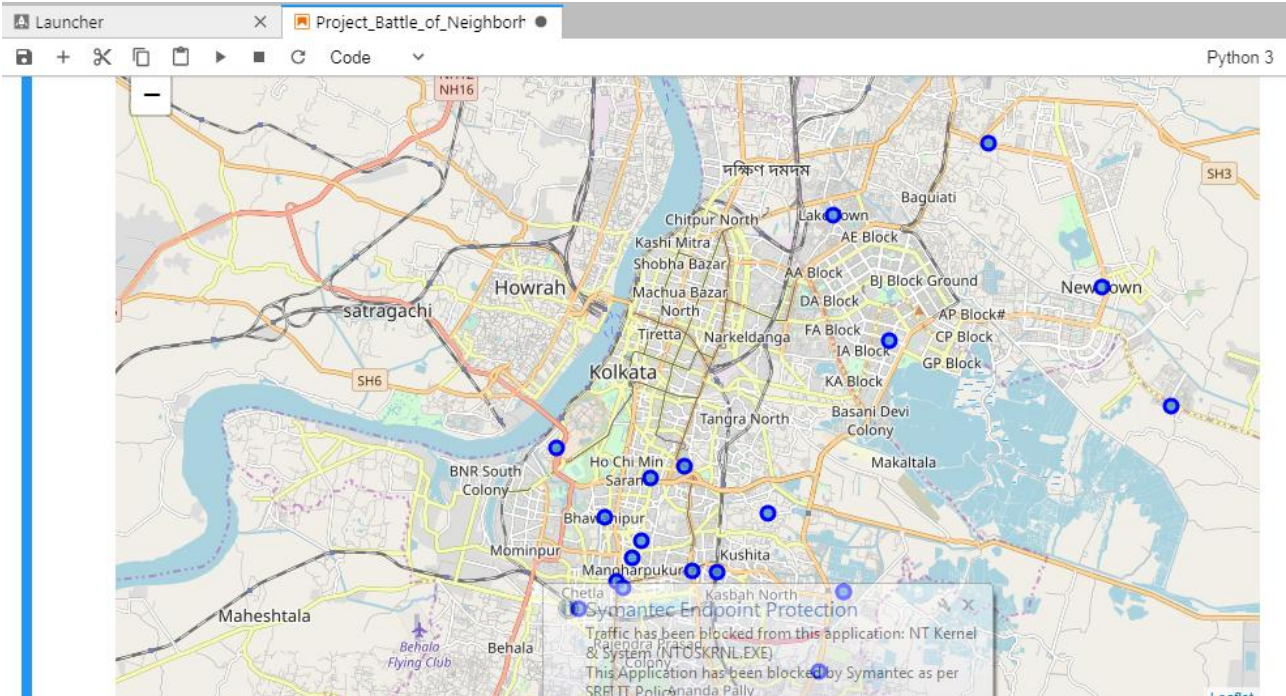
8]:

	Location	No_of_prop	Price/rentals	Score_prop
11	Salt Lake City Sector 2	83	18	25.0
2	Rajarhat	329	25	24.0
19	Garia	58	26	23.0
20	Action Area 2	56	30	21.5
15	New Alipore	76	30	21.5
13	Kasba East	81	33	20.0
8	Kasba	101	38	18.5
17	Bhawanipur	63	38	18.5
7	Ballygunge	103	39	17.0
9	Gariahat	95	41	16.0
3	New Town	168	44	15.0
22	Southern Avenue	53	45	14.0
14	Hazra	78	46	13.0
1	Salt Lake City	1489	49	11.5
12	Rash Behari Ave	81	49	11.5

**Note:**

1. Score is based upon lowest price/ rental criteria - the location with the lowest price/ rentals will get the highest score.
2. For locations where price/ rentals are equal, scoring is divided amongst the number of such locations.

## Mapping the locations with the latitude and longitude using the GEOLOCATOR library:



## IV. ANALYZING/ SCORING LOCATIONS BASED ON RESTURANTS/ HOTELS: CRITERIA-2/3

In this section, each location/ area is analyzed based on the number of restaurants/ hotels available using FOURSQUARE API. Each location is provided a score based upon the number of restaurants/ hotels available in the location/ area.

[24]:

	Nos_of_restaurants	Location	Score_restaurants
0	6	Gariahat	12.0
1	5	Minto Park	10.5
2	5	Park Street	10.5
3	3	Kalighat	8.5
4	3	Rajarhat	8.5
5	2	E M Bypass	5.5
6	2	Hazra	5.5
7	2	New Alipore	5.5
8	2	Rash Behari Ave	5.5
9	1	Ballygunge	2.0
10	1	Southern Avenue	2.0
11	1	Topsia	2.0

[28]:

	Nos_of_hotels	Location	Score_hotels
0	3	Minto Park	7.0
1	2	New Town	6.0
2	1	Gariahat	3.0
3	1	Hazra	3.0
4	1	Kalighat	3.0
5	1	Rajarhat	3.0
6	1	Southern Avenue	3.0

## V. SCORING LOCATIONS BASED ON DISTANCE FROM AIRPORT: CRITERIA -4

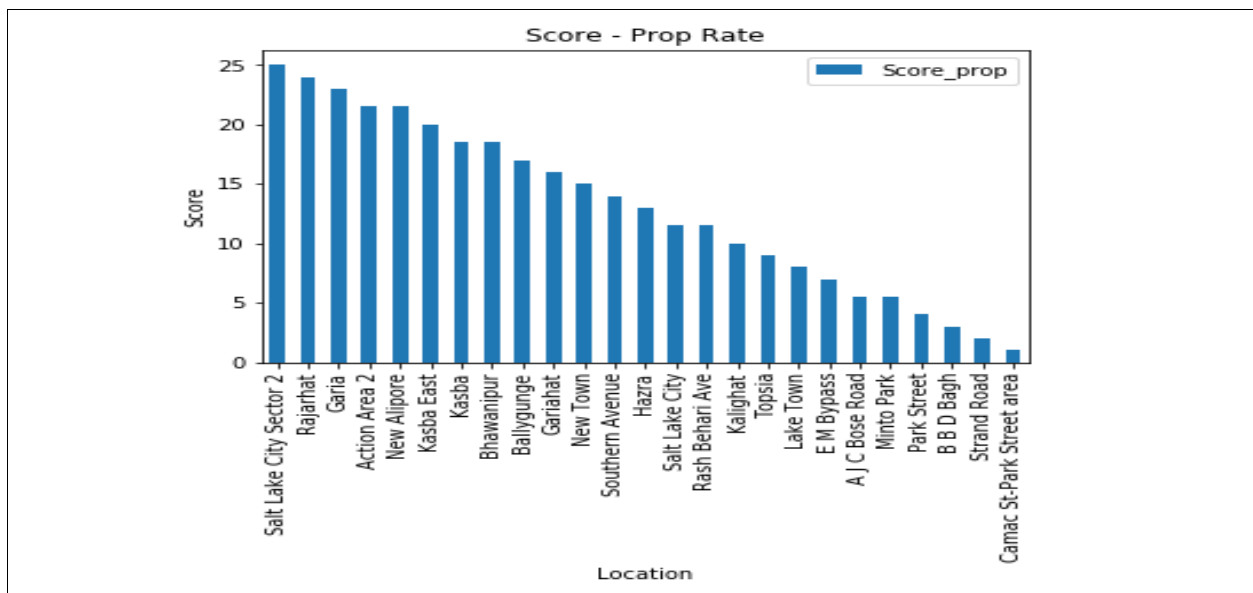
In this section, we would score the locations based on the last criteria, i.e, distance from airport.

It is to be noted that distance from city airport to each location has been calculated using the **Haversine distance formula**.

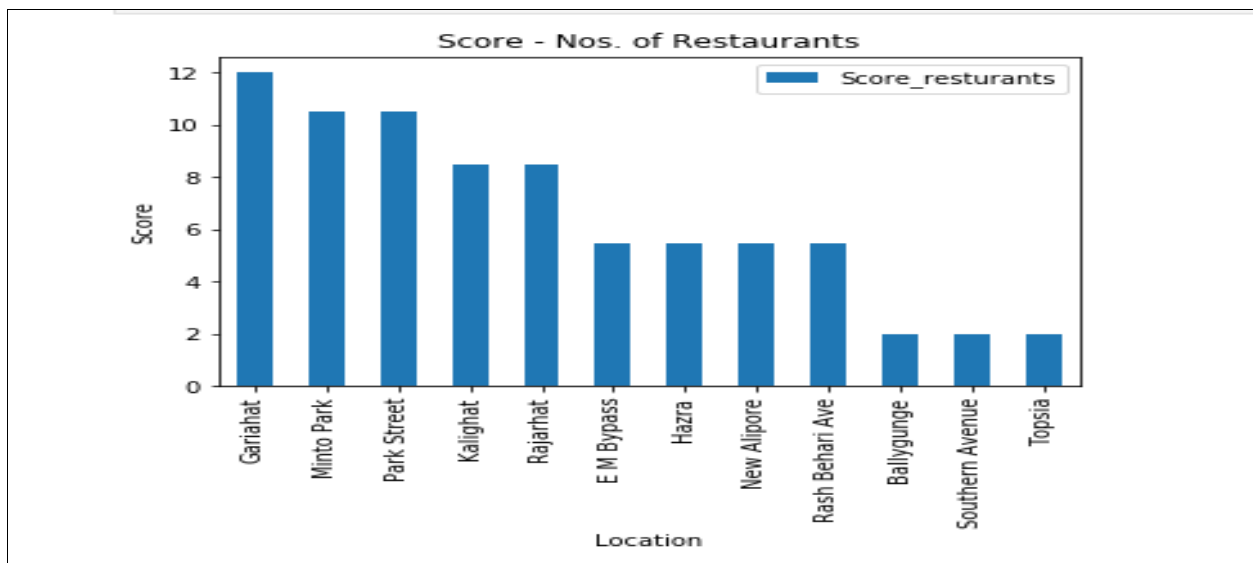
Locations are then ranked upon the distance, with the location closest to the airport getting the maximum score and vice-versa.

### Criteria - Scoring results:

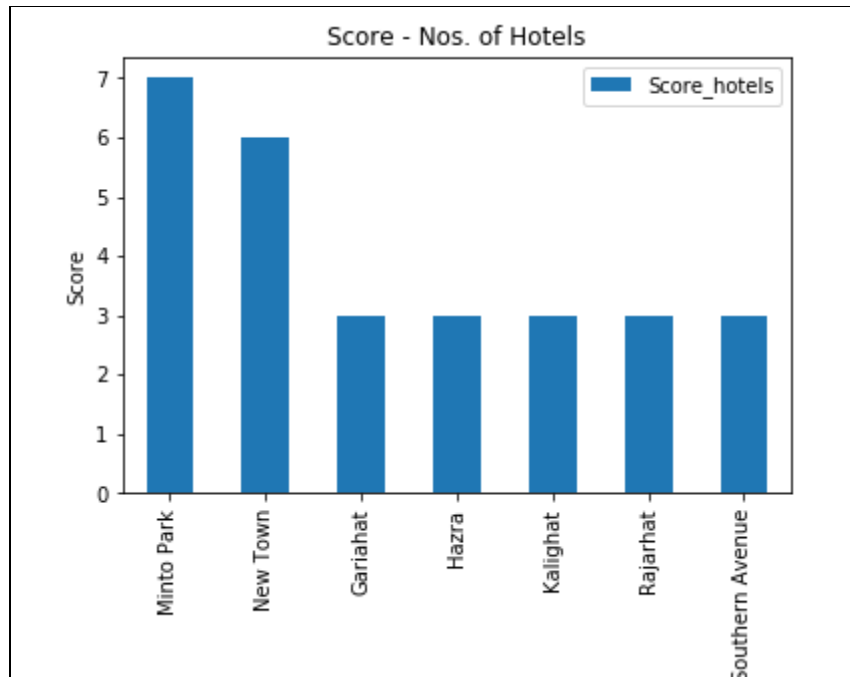
#### (i) Property Criteria:



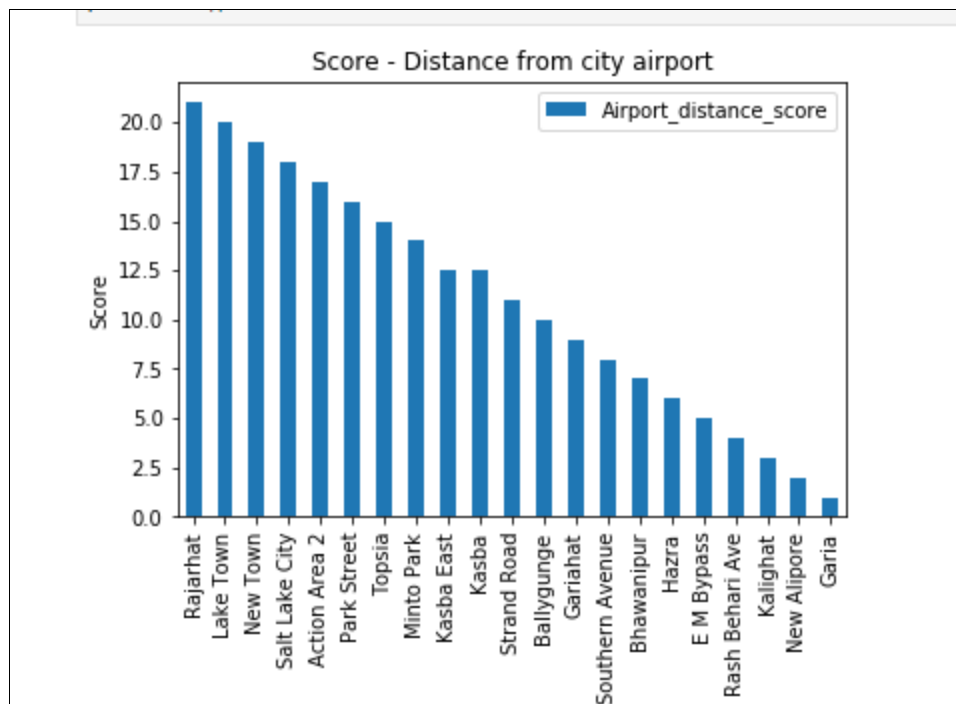
#### (ii) Restaurant Criteria:



(iii) Hotel Criteria:



(iv) Airport Distance Criteria:



## D. RESULT / Final Scoring:

### I. ASSIGNING WEIGHTS FOR EACH CRITERIA

In this section, the user was prompted to assign weights to each of the criteria, i.e for:

1. Property rates.
2. Number of restaurants.
3. Number of hotels.
4. Distance from city airport

Weights were assigned on a scale of 0 -100, of 100 being the highest importance.

Let us assume that the user has assigned the following weights:

#### WEIGHTS ASSIGNED

WEIGHT FOR RESTURANTS : 35.0

WEIGHT FOR HOTELS: 35.0

WEIGHT FOR PROPERTY PRICE/RENTALS : 10.0

WEIGHT FOR DISTANCE FROM AIRPORT : 20.0

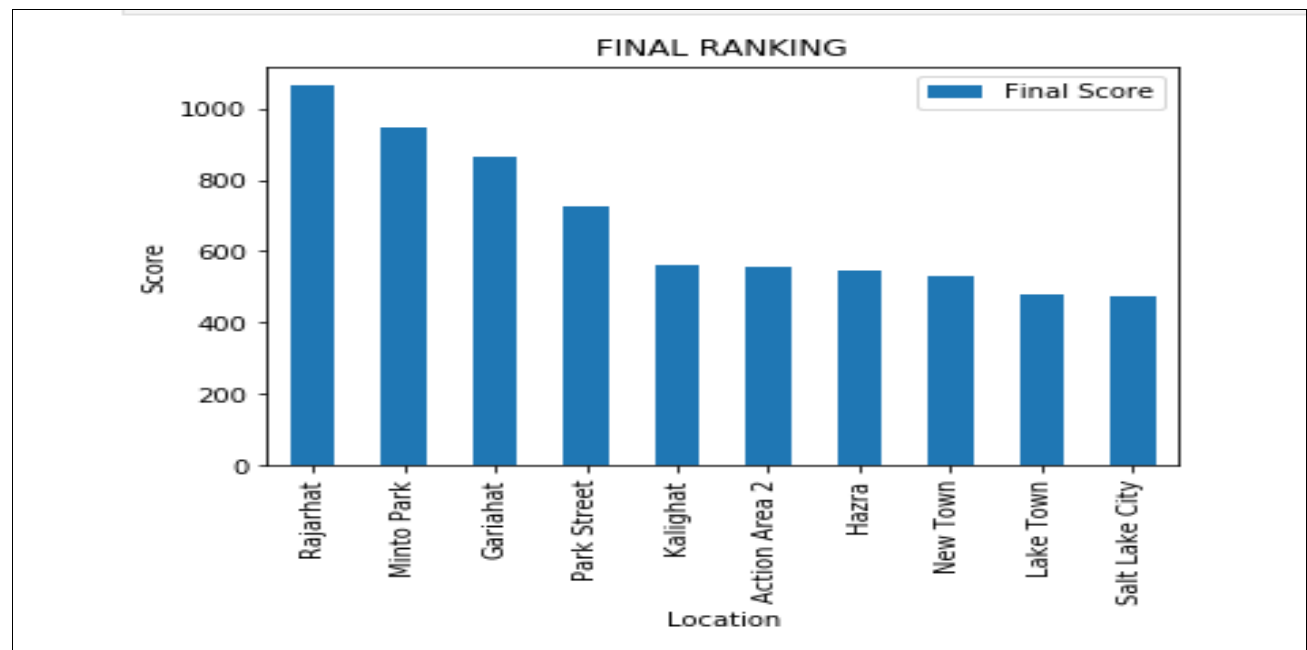
The final score for each location is the summation of the weighted product of each criteria score and weights chosen for those criteria, i.e.

**Final Score of each location** = (Score for property \* Weight for property) +  
(Score for restaurant \* Weight for restaurant) +  
(Score for hotel \* Weight for hotel) +  
(Score for airport distance \* Weight for city airport distance)

Based on the above calculations, the final score and rank of each location was derived and each location was ranked on the basis of score.

[41]:

	Location	Final Score	Rank
0	Rajarhat	1062.5	1.0
1	Minto Park	947.5	2.0
2	Gariahat	865.0	3.0
3	Park Street	727.5	4.0
4	Kalighat	562.5	5.0
5	Action Area 2	555.0	6.0
6	Hazra	547.5	7.0
7	New Town	530.0	8.0
8	Lake Town	480.0	9.0
9	Salt Lake City	475.0	10.0



#### Final Result:

Rajarhat, with the highest score, seems to be the most appropriate location as per the model, followed by Minto Park and Gariahat.



## E. ERROR/ EXCEPTION handling:

Since the program is **User Interactive** and is dependent upon the inputs from the user, quality/ quantity of data is prone to be affected due to **invalid entries** by the user.

However, the program is designed to handle errors/ exceptions arising out of invalid entries.

The programme will also provide an explanation/ message why the input entered by the user is invalid. The examples below substantiate the above:

### 1. Invalid entry for choice of City:

```
print("\n\nYou have chosen :\n\n",city)

Available Cities

['kolkata', 'chennai', 'new delhi', 'mumbai']

Enter Your City: new york

INVALID ENTRY ..... TRY AGAIN

Enter Your City: kolkata

You have chosen :

kolkata
```

### 2. Invalid entry for choice of dataset:

```
For running on existing database, enter 'A',
For scrapping through website, enter 'B'

Enter your choice c

INVALID ENTRY TRY AGAIN

For running on existing database, enter 'A',
For scrapping through website, enter 'B'

Enter your choice B

You have decided to use the scrape through the website
https://www.magicbricks.com/property-for-sale-rent-in-Kolkata/commercial-real-estate-Kolkata
[7]:
```

	Location	No_of_prop	SP_under_cons	SP_leased_prop	Rentals_per_month	ROI
1	Salt Lake City	1489	7460	8109	49	7.3%

### 3. Invalid entry for choice of property:

```
df_final
Enter Your Choice
r
INVALID ENTRY... PLEASE TRY AGAIN
Enter Your Choice
pr
You have chosen to take property on rent..... getting database.
```

[8]:

	Location	No_of_prop	Price/rentals
1	Salt Lake City	1489	49
2	Rajarhat	329	25
3	New Town	168	44
4	Park Street	151	85

### 4. Invalid entry during assigning weights:

```
Assigning Weights
Enter Weight for resturants g
Input has to a number or float.
Enter Weight for resturants -2

Enter Weight for hotels 34

Enter Weight for property price/rentals 34

Enter Weight for airport distance 34
100.0

ONE OR MORE WEIGHTS IS NEGATIVE....PLEASE ASSIGN WEIGHTS AGAIN
Assigning Weights
Enter Weight for resturants 20

Enter Weight for hotels 20

Enter Weight for property price/rentals 20

Enter Weight for airport distance 20
80.0

SUM OF WEIGHTS IS NOT 100....PLEASE ASSIGN WEIGHTS AGAIN
Assigning Weights
Enter Weight for resturants 35
```

## **F. Discussions:**

The model is dynamic in nature as the results may differ as per the preferences of the user. The user of this project has the options of selection of type of property, assigning weights to each criterion, etc., thus making the notebook user-friendly.

The goal of this project was to choose a locality meeting certain criteria, and if possible, quantify the result giving direction to the stakeholder for further insights. The goal has been achieved where in the final result, we were able to compare locations based on the scoring arrived from the model. In the conclusion section, we explore other possibilities.

## **G. Conclusion:**

The project is an attempt to locate a location meeting certain criteria. The project then tries to develop a scoring model so that available locations are scored on the basis of those criteria. Scoring models form the back-bone of data analytics for many fields like arriving at a Credit score for clients, ratings to corporates, etc. These models are also used for price discovery and alignment.

Based on the available data, a lucid method has been approached in this project. The exercise would have gathered further strength if more information were assimilated and incorporated within the model. This information could have been crime rates in different localities, availability of infrastructure facilities, etc. However, such databases / substantial information are not readily available. If such data are made available in future and incorporated within the model, the analysis is slated to become more robust.

-----END OF REPORT-----