COURSERA CAPSTONE PROJECT

COURSERA IBM DATA SCIENCE CERTIFICATION

BY,
YOGANANDA SRIHARI PRASAD GONE

Coursera Capstone - REPORT :-

1. Introduction :

1.1 Discussion of the "background situation" leading to the problem at hand:

1.2 Problem to be resolved

1.3 Audience for this project.

2. Data :

2.1 Data of Current Situation (current residence place)

2.2 Data required to resolve the problem

2.3 Data sources and data manipulation

3. Methodology :

3.1 Process steps and strategy to resolve the problem

3.2 Data Science Methods, machine learing, mapping tools and exploratory data analysis.

4. Results :

Discussion of the results and how they help to take a decision.

5. Discussion :

Elaboration and discussion on any observations and/or recommendations for improvement.

6. Conclusion :

Desicison taken and Report Conclusion.

1. Introduction

Discussion of the business problem and the people who would be interested in this project.

1.1 Description, Scenario, Background and Explaination of the Problem

Presently, I am pursuing my graduation in Computer Science and Engineering Stream and working on the areas of Data Science. I live in Hyderabad, India. I currently live within walking distance to RTC X roads and I enjoy many ammenities and venues in the area, such as various popular restaurants like Bawarchi, cafes, food shops, Pizza Shops and entertainment. I have been offered a great opportunity to work on the certain areas of Data Science at my college with my fello students. I am very excited and I want to use this opportunity to practice my learnings in Coursera in order to answer relevant questions arisen. The key question is : How can I find a convenient and enjoyable place similar to mine now in Hyderabad? Certainly, I search about that in Google but the idea is to use and apply myself the learned tools during the course. In order to make a comparison and evaluation of the rental options in RTC X roads, I must set some basis, therefore the apartment in RTC X roads must meet the following things:

1. Apartment must be 2 or 3 bedrooms.
2. Desired location is near a metro station in the RTC X roads and within 2km radius price of rent not exceed RS 15,000/- per month.
3. Top ammenities in the selected neighborhood shall be similar to current residence
4. Desirable to have venues such as coffee shops, Family restaurants, Study Halls, gym and food shops as a reference, I have included a map of venues near current residence in Hyderabad.
5. And also to ensure that Super markets, Schools, Study centers, Library are at wakable distance. And to ensure that there is no water problem in that particular locality.

1.2 Challenging problem and Problem to be resolved:

1. Apartment with min 2 bedrooms with monthly rent not to exceed RS 15,000/- permonth
2. Unit located within walking distance (2 km) from a metro station in RTC X Roads
3. Area with ammenities and venues similar to the ones described for current location (See item 2.1)

1.3 Interested and Enthusiast Audience

I believe this is a relevant challenge with valid questions for anyone moving to other large cities in India like Hyderabad, Delhi, Mumbai etc. The same methodology can be applied in accordance to demands as applicable. This case is also applicable for anyone interested in exploring starting or locating a new business in any city. Lastly, it can also serve as a good practical exercise to develop Data Science skills. So Interested Audience can think about it and may show their intrests accordingly. Those people are highly appreciated :).

2. Data

The Description of the data and its sources that will be used to solve the problem are as follows:

1. List of Boroughs and neighborhoods of RTC X Roads with their geodata (lattitude and longitude)
2. List of Hyderabad metro stations in RTC X Roads with their address location
3. List of apartments for rent in RTC X Roads area with their addresses and price
4. Preferably, a list of apartment for rent with additional information, such as price, address, area etc
5. Venues for each RTC X Roads neighborhood (than can be clustered)
6. Venues for subway metro stations, as needed

2.1 Data of Current Situation

Currently reside in the neighborhood of 'RTC X Roads' in Hyderabad. I use Foursquare to identify the venues around the area of residence which are then shown in the Hyderabad map shown in methodology and execution in section 3.0 . It serves as a reference for comparison with the desired future location in RTC X Roads

2.2 Data Required to resolve the problem

In order to make a good choice of a similar apartment in RTC X Roads, the following data is required: List/Information on neighborhoods form RTC X Roads with their Geodata (lattitude and longitude. List/Information about the nearby metro stations in RTC X Roads with geodata. Listed apartments for rent in RTC X Roads area with descriptions (how many beds, price, location, address) Venues and ammenities in the RTC X Roads neighborhoods (e.g. top 10) 2.3 sources and manipulation The list of RTC X Roads neighborhoods is worked out during LAb exercise during the course.

List of Hyderabad Metro Stations :- https://en.wikipedia.org/wiki/List_of_Hyderabad_Metro_stations

Google Map view for the above stations :- https://www.google.com/maps/search/hyderabad+metro+stations/@17.4050226,78.4593429,13z/data=!3m1!4b1

Property Rates & Price Trends in Hyderabad - 2020 :- https://www.makaan.com/price-trends/property-rates-for-buy-in-hyderabad

A csv file was created which will be read in order to create a dataframe and its mapping. The csv file 'roads_data.csv' has the following below data structure. The file will be directly read to the Jupyter Notebook for convenience and space savings. The clustering of neighborhoods and mapping will be shown however. An algorithm was used to determine the geodata from Nominatim.

With the use of geolocator = Nominatim() , it was possible to determine the latitude and longiude for the metro locations as well as for the geodata for each rental place listed. The loop algorithms used are shown in the execution of data in section 3.0 "Great_circle" function from geolocator was used to calculate distances between two points , as in the case to calculate average rent price for units around each subway station and at 2 km radius. Foursquare is used to find the avenues at RTC X Roads neighborhoods in general and a cluster is created to later be able to search for the venues depending of the location shown.

2.3 Data sources and data manipulation

The data will be used as follows: Use Foursquare and geopy data to map top 10 venues for all RTC X Roads neighborhoods and clustered in groups (as per Course LAB) Use foursquare and geopy data to map the location of metro stations , separately and on top of the above clustered map in order to be able to identify the venues and ammenities near each metro station, or explore each location separately Use Foursquare and geopy data to map the location of rental places, in some form, linked to the locations. create a map that depicts, for instance, the average rental price per square ft, around a radious of 1.0 mile (2 km) around each metro station - or a similar metrics. I will be able to quickly point to the popups to know the relative price per area. Addresses from rental locations will be converted to geodata(lat, long) using Geopy-distance and Nominatim. Data will be searched in open data sources if available, from real estate sites if open to reading, libraries or other government agencies such as LIC.
Mapping of the data is done as follows :-

The following maps were created to facilitate the analysis and the choice of the palace to live. RTC X Roads map of Neighborhoods RTC X Roads metro locations RTC X Roads map of places for rent RTC X Roads map of clustered venues and neighborhoods Combined maps of RTC X Roads rent places with subway locations Combined maps of RTC X Roads rent places with desired locations and venues clusters.

3. Methodology

This section represents the main component of the report where the data is gathered, prepared for analysis. The tools described are used here and the Notebook cells indicates the execution of steps.

The analysis and the strategy

The strategy is based on mapping the above described data in section 2.0, in order to facilitate the choice of at least two candidate places for rent. The choice is made based on the demands imposed : location near a metro station, rental price and similar venues to Hyderabad. This visual approach and maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy.

The processing of these DATA and its mapping will allow to answer the key questions to make a decision:

1. what is the cost of available rental places that meet the demands?
2. what is the cost of rent around a mile radius from each metro station?
3. what is the area of RTC X Roads with best rental pricing that meets criteria established?
4. What is the distance from work place and the tentative future rental home?
5. What are the venues of the two best places to live? How the prices compare?
6. How venues distribute among RTC X Roads neighborhoods and around metro stations?
7. Are there tradeoffs between size and price and location?
8. Any other interesting statistical data findings of the real estate and overall data.

Tools :-

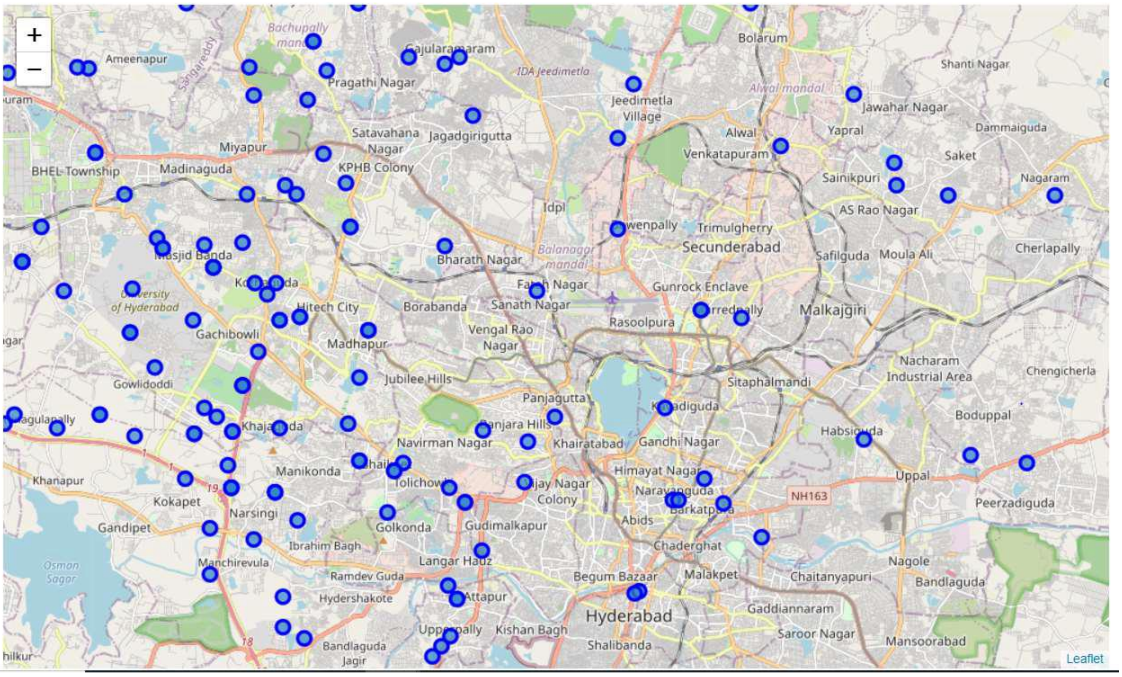
Web-scraping of sites is used to consolidate data-frame information which was saved as csv files for convenience and to simply the report. Geodata was obtained by coding a program to use Nominatimto get latitude and longitude of subway stations and also for each of (144 units) the apartments for rent listed.
Geopy_distanceand Nominatimwere used to establish relative distances.
Seaborn graphic was used for general statistics on rental data.

Maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy.

4. Execution and Results

Current residence Neighborhood in Hyderabad :-

Out[37]:



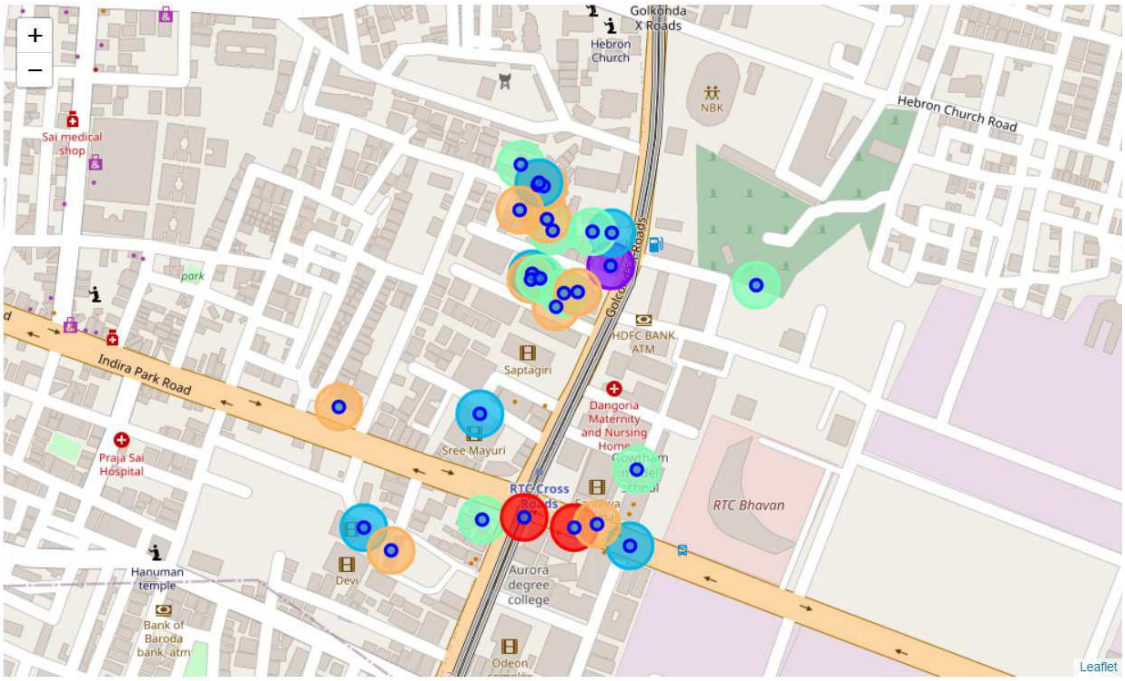
Venues around Neighbourhood :-

Out[9]:

	name	categories	lat	lng
0	Bawarchi	Indian Restaurant	17.406369	78.497662
1	Sudharshan Theatre 35mm	Movie Theater	17.406530	78.495150
2	Devi 70 MM	Movie Theater	17.406329	78.495409
3	Crystal Restaurant	Asian Restaurant	17.406608	78.496268
4	Astoria Restaurant	Indian Restaurant	17.406530	78.497136
5	Sandhya 70 MM	Movie Theater	17.407053	78.497724
6	Sri Mayuri Theatre	Movie Theater	17.407557	78.496241
7	Nandhini Sudha Restaurant	Breakfast Spot	17.403668	78.495264

RTC X Roads, Hyderabad Map -Neighborhoods and Cluster of Venues :-

Out[14]:



GeoDataHyderabad appartmentsfor rent :

In [52]:
```
# csv files with rental places with basic data but still wihtout geodata ( latitude and longitude)
# pd.read_csv(' le.csv', header=None, nrows=5)
RTC_rent=pd.read_csv('roads2_data_new.csv')
RTC_rent.head()
```

Out[52]:

	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
0	BhagyalaxmiNagar,Kavadiguda	Upper West Side	2.94	5	3400	10000	NaN	NaN
1	Gachibowli,OuterRingRoad	Upper East Side	3.57	3	2100	7500	NaN	NaN
2	Gachibowli	Upper West Side	1.89	4	2800	5300	NaN	NaN
3	Moosapet,NH	West Village	3.03	2	1650	5000	NaN	NaN
4	RaghavendraColonykondapur	Chelsea	3.45	2	1450	5000	NaN	NaN

In [53]:
```
RTC_rent.tail()
```

Out[53]:

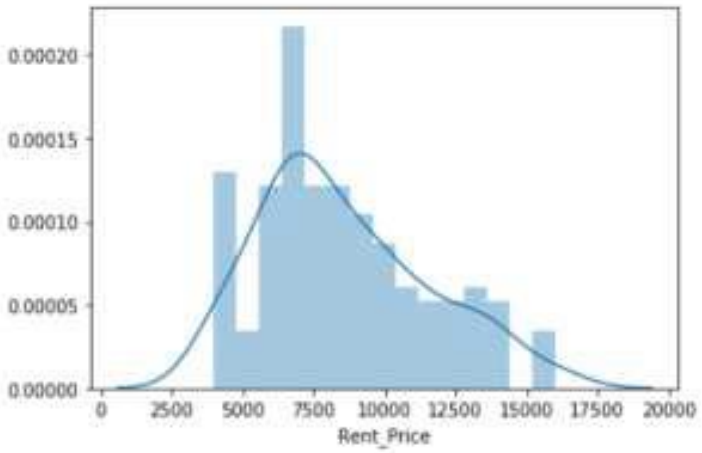
	Address	Area	Price_per_ft2	Rooms	Area-ft2	Rent_Price	Lat	Long
139	KukatpallyHousingBoardColony,NH	Rental in Lenox Hill	5.15	3	1700	8750	NaN	NaN
140	KavuriHills,Madhapur	No fee rental in Tribeca	7.11	2	1223	8700	NaN	NaN
141	SMRVinayIconia,Kondapur	No fee rental in Midtown East	3.87	3	2100	8118	NaN	NaN
142	TrendsetWinz,Gachibowli,OuterRingRoad	No fee rental in Central Park West	5.06	2	1600	8095	NaN	NaN
143	Gachibowli,OuterRingRoad	Rental in Greenwich Village	6.67	2	1500	10000	NaN	NaN

Rental Price Statistics of Appartments:-

Budget 15000/month is around the mean

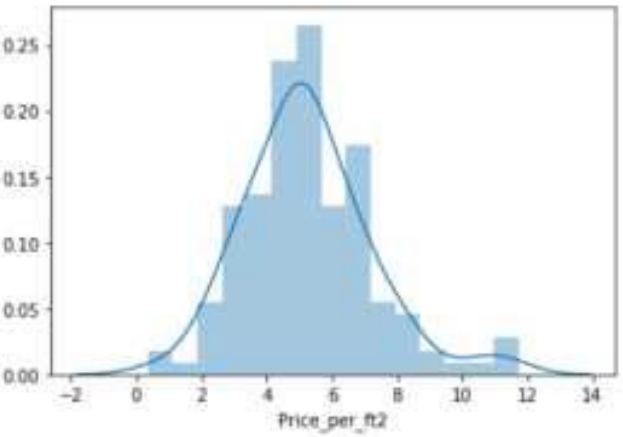
```
In [56]: import seaborn as sns
         sns.distplot(RTC_rent['Rent_Price'],bins=15)

Out[56]: <matplotlib.axes._subplots.AxesSubplot at 0x1e58b53db00>
```

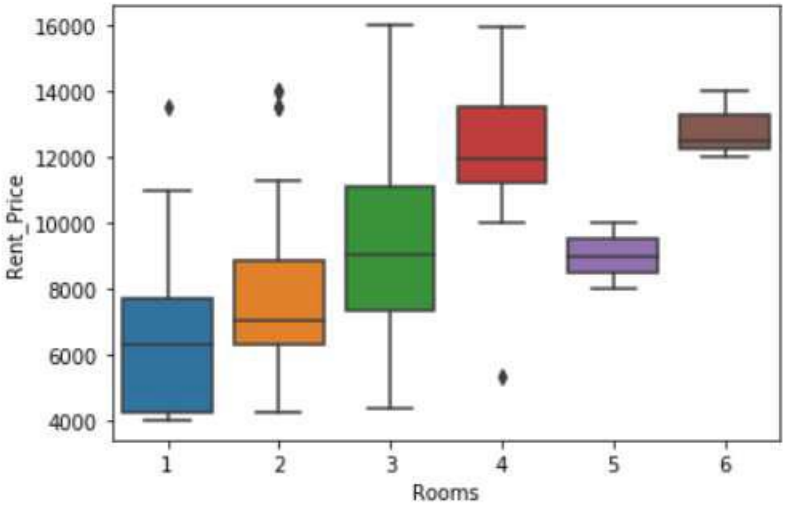



```
In [57]: import seaborn as sns
         sns.distplot(RTC_rent['Price_per_ft2'],bins=15)

Out[57]: <matplotlib.axes._subplots.AxesSubplot at 0x1e589c09f98>
```



```
In [58]: sns.boxplot(x='Rooms', y= 'Rent_Price', data=RTC_rent)
Out[58]: <matplotlib.axes._subplots.AxesSubplot at 0x1e589ac1f28>
```



Venues of cluster 3 :-

```
In [61]: ## kk is the cluster number to explore
kk = 3
RTC_X_Roads_merged.loc[RTC_X_Roads_merged['Cluster Labels'] == kk, RTC_X_Roads_merged.columns[[1] + list(range(5, RTC_X_Roads_mer
```

Out[61]:

	Neighborhood	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
3	Crystal Restaurant	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	Park	Frozen Yogurt Shop	Spanish Restaurant
5	Sandhya 70 MM	Deli / Bodega	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Sushi Restaurant	Beer Garden	Coffee Shop	Falafel Restaurant	Bike Trail	Other Nightlife
10	Chikkadpally Police Station	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Deli / Bodega	Gym	Sporting Goods Shop	Thai Restaurant
12	Satya Residency	Italian Restaurant	Bar	Bakery	Vegetarian / Vegan Restaurant	Indian Restaurant	Coffee Shop	Cosmetics Shop	Wine Bar	Mexican Restaurant	Sushi Restaurant
16	ICICI Bank	Sandwich Place	Hotel	Japanese Restaurant	Gym / Fitness Center	Coffee Shop	Salon / Barbershop	Burger Joint	French Restaurant	Bar	Italian Restaurant
17	Azaan Bait Al Mandi Musherabad	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	Theater	Art Gallery	Seafood Restaurant	American Restaurant	Hotel
18	Telugu CEEFI	Italian Restaurant	Sushi Restaurant	French Restaurant	Clothing Store	Chinese Restaurant	Café	Indian Restaurant	Bakery	Seafood Restaurant	Electronics Store
27	Twin City Speech& Hearing Centre	Italian Restaurant	Restaurant	Thrift / Vintage Store	Cocktail Bar	Bagel Shop	Coffee Shop	Pizza Place	Mexican Restaurant	Grocery Store	Wine Shop
29	Muthoot Fincorp	Coffee Shop	Hotel	Gym	Wine Shop	Steakhouse	Bar	Italian Restaurant	Pizza Place	Park	Gym / Fitness Center

Hyderabad Metro Stations geodata

In [64]:
```
RTC=pd.read_csv('roads5_data.csv')
print(RTC.shape)
RTC.head()
```

```
(48, 4)
```

Out[64]:

	sub_station	sub_address	lat	long
0	Miyapur	Miyapur Metro Station, Nadigada Tanda, Miyapur...	17.496454	78.370747
1	JNTU College	JNTU College Metro Station, near JNTU College,...	17.498566	78.386677
2	KPHB Colony	KPHB Colony Metro Station, Bhagya Nagar Colony...	17.493797	78.399483
3	Kukatpally	Kukatpally Metro Station, APHB Colony, Kukatpa...	17.485115	78.409369
4	Balanagar	Dr B.R. Ambedkar Balanagar Metro Station, IDA ...	17.476826	78.419926

In [65]:
```
# removing duplicate rows and creating new set mhsub1
RTCsub1=RTC.drop_duplicates(subset=['lat','long'], keep="last").reset_index(drop=True)
RTCsub1.shape
```

Out[65]:
```
(48, 4)
```

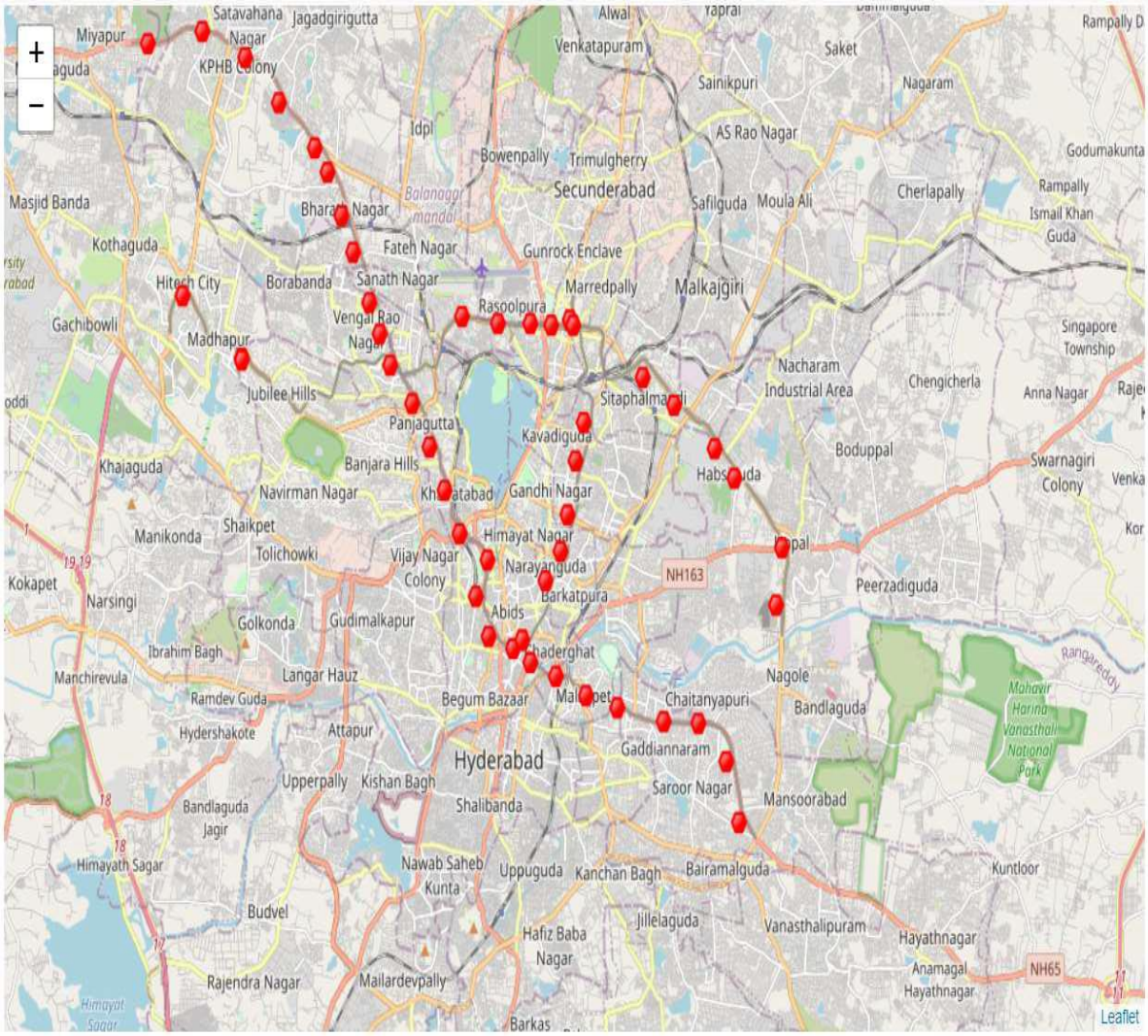
In [66]:
```
RTCsub1.tail()
```

Out[66]:

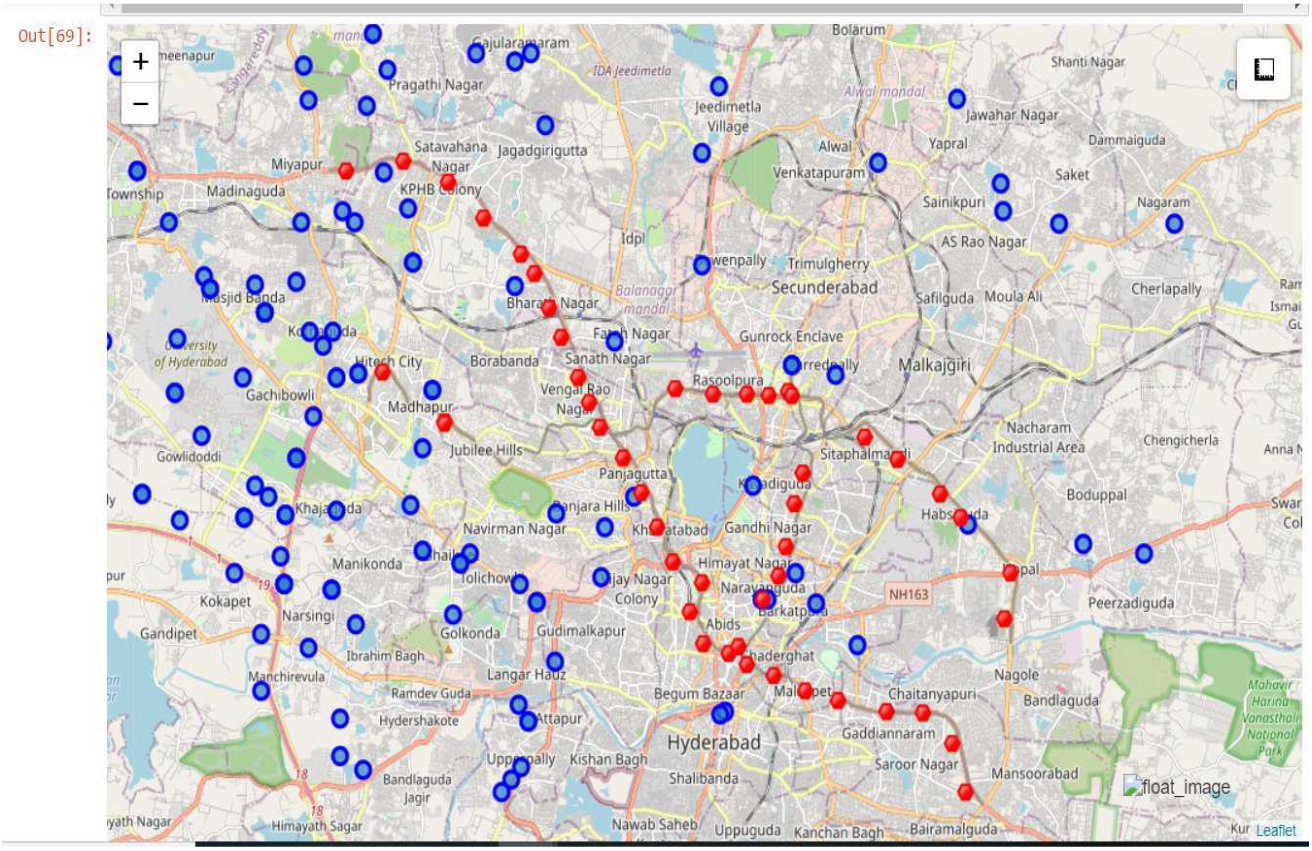
	sub_station	sub_address	lat	long
43	Rasoolpura	Rasoolpura Metro Station, Gun Bazar, Rasoolpur...	17.443597	78.474176
44	Prakash Nagar	Prakash Nagar Metro Station, Chikoti Gardens, ...	17.445083	78.463742
45	Madhapur	Madhapur Metro Station, Road Number 36, Aditya...	17.437244	78.398296
46	Durgam Cheruvu	Durgam Cheruvu Metro Railway Station, CBI Colo...	17.436957	78.398443
47	Hitech City	Hitech City Metro Station, Silicon Valley, Mad...	17.448826	78.381001

MAP of Hyderabad showing the location of subway stations:

Out[67]:



Appartmentsfor rent (blue color) and Metro Stations (redcolor)



Select the apartment for rent

Venues in Cluster 2

Venues for Apartment 1 - Cluster 2

In [71]:
```
## kk is the cluster number to explore
kk = 2
RTC_X_Roads_merged.loc[RTC_X_Roads_merged['Cluster Labels'] == kk, RTC_X_Roads_merged.columns[[1] + list(range(5, RTC_X_Roads_mer
```

Out[71]:

	Neighborhood	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	Bawarchi	Coffee Shop	Discount Store	Yoga Studio	Steakhouse	Supplement Shop	Tennis Stadium	Shoe Store	Gym	Bank	Seafood Restaurant
1	Sudharshan Theatre 35 mm	Chinese Restaurant	Cocktail Bar	Dim Sum Restaurant	American Restaurant	Vietnamese Restaurant	Salon / Barbershop	Noodle House	Bakery	Bubble Tea Shop	Ice Cream Shop
6	Sri Mayuri Theatre	African Restaurant	Seafood Restaurant	French Restaurant	American Restaurant	Cosmetics Shop	Chinese Restaurant	Event Space	Liquor Store	Beer Bar	Gym / Fitness Center
9	Nandhini sudha Reastaurant	Coffee Shop	Gym	Bar	Italian Restaurant	Sushi Restaurant	Pizza Place	Mexican Restaurant	Deli / Bodega	Japanese Restaurant	Pub
14	Kala Jyothi Process Pvt. Ltd	Theater	Italian Restaurant	Coffee Shop	American Restaurant	Gym / Fitness Center	Hotel	Wine Shop	Spa	Gym	Indie Theater
23	District Industries Centre	Clothing Store	Boutique	Women's Store	Shoe Store	Men's Store	Furniture / Home Store	Italian Restaurant	Mediterranean Restaurant	Art Gallery	Design Studio
26	R.K.Enterprises Mobile Shoppee	Coffee Shop	American Restaurant	Park	Bookstore	Pizza Place	Sandwich Place	Burger Joint	Café	Deli / Bodega	Tennis Court

Venues in Cluster 3

Venues for Apartment 2 - Cluster 3 ¶

In [72]:
```
## kk is the cluster number to explore
kk = 3
RTC_X_Roads_merged.loc[RTC_X_Roads_merged['Cluster Labels'] == kk, RTC_X_Roads_merged.columns[[1] + list(range(5, RTC_X_Roads_mer
```

Out[72]:

	Neighborhood	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
3	Crystal Restaurant	Mexican Restaurant	Lounge	Pizza Place	Café	Wine Bar	Bakery	American Restaurant	Park	Frozen Yogurt Shop	Spanish Restaurant
5	Sandhya 70 MM	Deli / Bodega	Italian Restaurant	Seafood Restaurant	Mexican Restaurant	Sushi Restaurant	Beer Garden	Coffee Shop	Falafel Restaurant	Bike Trail	Other Nightlife
10	Chikkadpally Police Station	Sushi Restaurant	Italian Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Deli / Bodega	Gym	Sporting Goods Shop	Thai Restaurant
12	Satya Residency	Italian Restaurant	Bar	Bakery	Vegetarian / Vegan Restaurant	Indian Restaurant	Coffee Shop	Cosmetics Shop	Wine Bar	Mexican Restaurant	Sushi Restaurant
16	ICICI Bank	Sandwich Place	Hotel	Japanese Restaurant	Gym / Fitness Center	Coffee Shop	Salon / Barbershop	Burger Joint	French Restaurant	Bar	Italian Restaurant
17	Azaan Bait Al Mandi Musherabad	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Nightclub	Theater	Art Gallery	Seafood Restaurant	American Restaurant	Hotel
18	Telugu CEEFI	Italian Restaurant	Sushi Restaurant	French Restaurant	Clothing Store	Chinese Restaurant	Café	Indian Restaurant	Bakery	Seafood Restaurant	Electronics Store
27	Twin City Speech& Hearing Centre	Italian Restaurant	Restaurant	Thrift / Vintage Store	Cocktail Bar	Bagel Shop	Coffee Shop	Pizza Place	Mexican Restaurant	Grocery Store	Wine Shop
29	Muthoot Fincorp	Coffee Shop	Hotel	Gym	Wine Shop	Steakhouse	Bar	Italian Restaurant	Pizza Place	Park	Gym / Fitness Center

Apartment Selection :-

Using the "one map" above, I was able to explore all possibilities since the popups provide the information needed for a good decision. Apartment 1 rent cost is 15,000 slightly above the budget. Apt 1 is located 400 meters from subway station. One can walk to work place and use subway for other places aroung. Venues for this apt are as of Cluster 2 and it is located in a fine district in the East side of RTC_X_Roads.

Apartment 2 rent cost is 12,000, just under the budget. Apt 2 is located 60 meters from subway station but one will have to ride the subway daily to work , possibly 40-60 min ride. Venues for this apt are as of Cluster 3. Based on current Hyderabad venues, I feel that Cluster 2 type of venues is a closer resemblance to the current place. That means that APARTMENT 1 is a better choice since the extra monthly rent is worth the conveniences it provides.

5. Discussion

In general, I am positively impressed with the overall organization, content and lab works presented during the Coursera IBM Certification Course

I feel this Capstone project presented me a great opportunity to practice and apply the Data Science tools and methodologies learned.

I have created a good project that I can present as an example to show my potential.

I feel I have acquired a good starting point to become a professional Data Scientist and I will continue exploring to creating examples of practical cases.

6. Conclusion

I feel rewarded with the efforts, time and money spent. I believe this course with all the topics covered is well worthy of appreciation.

This project has shown me a practical application to resolve a real situation that has impacting personal and financial impact using Data Science tools.

The mapping with Folium is a very powerful technique to consolidate information and make the analysis and decision thoroughly and with confidence. I would recommend for use in similar situations.

One must keep abreast of new tools for DS that continue to appear for application in several business fields

Thank You Very Much…