Coursera Applied Capstone Project

Battle of neighborhoods ||

Module – 5 (week 5)

Bengaluru

Author: Anil Kumar Naidu

Date: 16/05/2020

Place: Bengaluru

Introduction

Problem description:

Background:

Bengaluru (or Bangalore), silicon city of India and City of many dreams is a popular booming city, and why?

Refer here: https://www.thenewsminute.com/article/projected-gdp-85-bengaluru-be-world-sfastest-growing-city-report-116556

The city is going to be one of the fastest growing Cities in world with GDP of 8.5%. This means a positive sign for starting many business and real estate.

So, if it is definitely a good destination for investing? Then where to invest? That is very confusing.

Problem:

As almost all areas seem to be better than each other for a human mind. How can a normal human mind even with lot of Experience predict which area would be popular for investing such that the choice results in High Gains?

A confused mind would be wandering around, looking for all possible growth areas in Bangalore. And in such tiresome work, first of all if some areas are missed, it would be a great future loss of profit.

So clearly, the stakes are high due to **two** reasons:

1.) It is difficult to choose given so many parameters to consider 2.)
If some area is missed, there will be huge future loss of profit.

The problem is further diversified to **Major possible investment strategies**:

1.) Residential investments
2.) Fast growth and return investments
3.) Small businesses

How can this problem of Investing in Bengaluru be solved so that the returns are most and which areas are most suited to the possible investment strategies?

Problem Background: Investment in land and business is the best form of earning quick money. Investment is money intensive, a lot of money is needed.

If this money is invested at wrong place, it results in LOSS. This may make the investor Bankrupt also. In some cases due to extreme bankruptcy, the investors have even committed suicide.

Hence a right approach **is must.**

Data Description:

Data required for this method would be Residential and commercial real-estate rates along with all the localities of Bangalore.

One particular website was used to scrap data. The link is listed below.

https://bengaluru.citizenmatters.in/615-real-estate-rates-615

This is a recent data. Hence most applicable. The data consists of Localities in Bangalore with both the Commercial and Residential rates if applicable. Some areas seem to be completely commercial while some are completely residential. This means the currently available Residential/ Commercial spaces are none.

How Data can be used to solve this problem?

The Data covers almost entire localities of Bangalore. First, **data of localities** is created from the scrapping of website. And two dataframes of **Residential and commercial spaces** exclusively are also generated.

Missing values of any data/ inappropriate Rates **are dropped** to create a unison data.

The Data is then used to fetch Coordinates of Each location using **Geocoder**.

After which, **Foursquare** is used to get all the venue data **with in 2km range**.

By using **KMeans** clustering, the Localities are clustered into 5 clusters.

The clusters are observed for **most frequently** occurring venues.

The Residential and Commercial rate data **is then used** to find the **Average** rates of each cluster.

As **land rate and the locality matter** a lot for both Residential and commercial spaces, using the Analysis, meaningful observations about which clusters are best suited for which type of investment is decided upon.

Using the data**, hence, the Best land investment strategy and area(from cluster) can be determined**.

Methodology:

The following steps are performed in the course of the Project:

1.) Requirement analysis
2.) Identifying possible data sources
3.) Collecting data- by using web scrapping and foursquare API
4.) Data preprocessing
5.) Data processing and standardization
6.) Data visualization for better understanding
7.) In case of data redundancy, repeat from Step 3
8.) Applying K-means algorithm with various Cluster values
9.) Machine learning algorithm evaluation
10.) Data re-visualization of cluster data
11.) Inferring meaning out of Data if needed using Graphs
12.) Conclude with results

Data required for solving the aforementioned problem was identified by using **Web scrapping** technique. Data collected was initially covering the City of Bangalore partially. This would mean lower coverage and hence less accurate results.

Then, few **more sources** were identified and a particular source of Data, link below:

https://bengaluru.citizenmatters.in/615-real-estate-rates-615

was decided to be **sufficient** enough.

The Data collected from above link had details about the Localities and prices of Real-Estate there, but **coordinate data was inavailable** then.

The data also had lot of **redundant** cells, which conveyed none information. Those cells were dropped from dataframe.

```
bangalore_boroughs = pd.read_html("https://bengaluru.citizenmatters.in/615-real-estate-rates-615")
bangalore_boroughs[0]
```

	0	1	2
0	Area	Residential rates (All numbers in Rs Sq/ft)	Commercial rates(All numbers in Rs Sq/ft)
1	BLR – Central business Districts	BLR – Central business Districts	BLR – Central business Districts
2	MG Road	NaN	20000
3	Kasturba Rd	NaN	15000
4	Cubbon Rd	12500	NaN
5	Church Street	12500	15000
6	Dickenson Rd	NaN	12500

The Data was initially split into **three data frames**,

1.) **Location** Dataframe (included both residential and commercial localities)
2.) **Residential** location data frame
3.) **Commercial** location dataframe

Our data statistics

```
print(bangalore_residential.head(25))
print(bangalore_commercial.head())
print(bangalore_residential.shape)
print(bangalore_commercial.dtypes)
```

```
          Neighborhood Residential_rate
2            Cubbon Rd            12500
3        Church Street            12500
5          Ashokanagar             2500
6      Victoria Layout             4000
8          Infantry Rd            10000
9         Shivajinagar             2500
10        Cunningham Rd           10000
11           Queens Rd             5000
12          Millers Rd             6000
13         Vasantnagar             7000
14        Langford Town            6000
15        Richmond Town            6000
17          St.Marks Rd           10000
18         Cambridge Rd            5000
19            Ulsoor Rd            8000
20      Vittal Mallya Rd          10000
21           Lavelle Rd           12000
25          Gandhinagar           12000
40    Basaveshwarnagar             4000
41    West of Chord Road           4000
42           Vijaynagar            4000
43          Magadi Road            4000
44       Chandra Layout            3000
45      Dr Rajkumar Road           5000
46   Mahalakshmi Layout            3000
   Neighborhood Commercial_rate
0       MG Road           20000
1    Kasturba Rd          15000
3  Church Street          15000
```

Using **Geocoder** Python API, the data was further collected about the coordinates of each location. This was well tabulated into DataFrame for standard use.

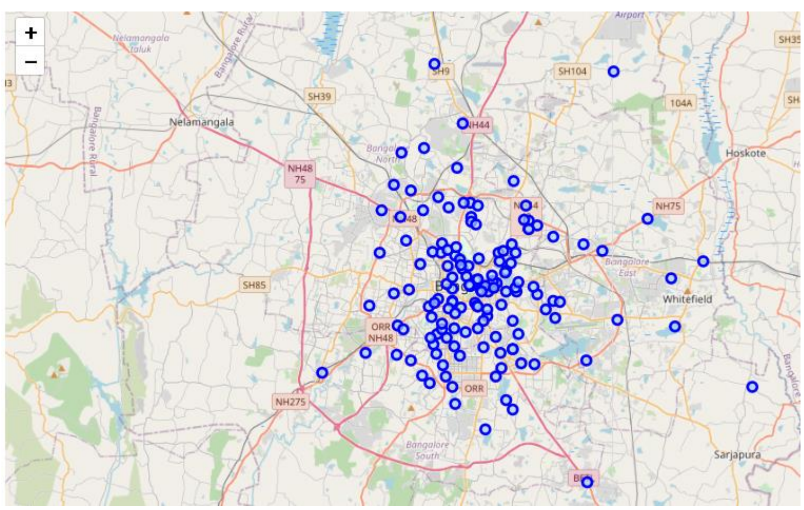
```
bangalore_neighbors.head()
```

	Neighbors	Latitude	Longitude
0	MG Road	12.9777	77.6019
1	Kasturba Rd	12.9767	77.5993
2	Cubbon Rd	12.9778	77.6066
3	Church Street	12.9751	77.6047
4	Dickenson Rd	12.9809	77.6107

For few locations the Geocoder was unable to provide the Coordinates, for those locations**, a manual search** method was used to find the latitude and longitude of the location.

The collected data was visualized using a **Folium City map** along with the markers and popups for verifying the correctness and better coverage.

As shown in the Map, the **blue** marked areas are the localities of interest. It covers complete Bengaluru.



Foursquare API is an extremely useful API for mining data related to **Venues** near by a coordinate.

With a developer account at FourSquare, it is possible to **invoke API calls using Python**.

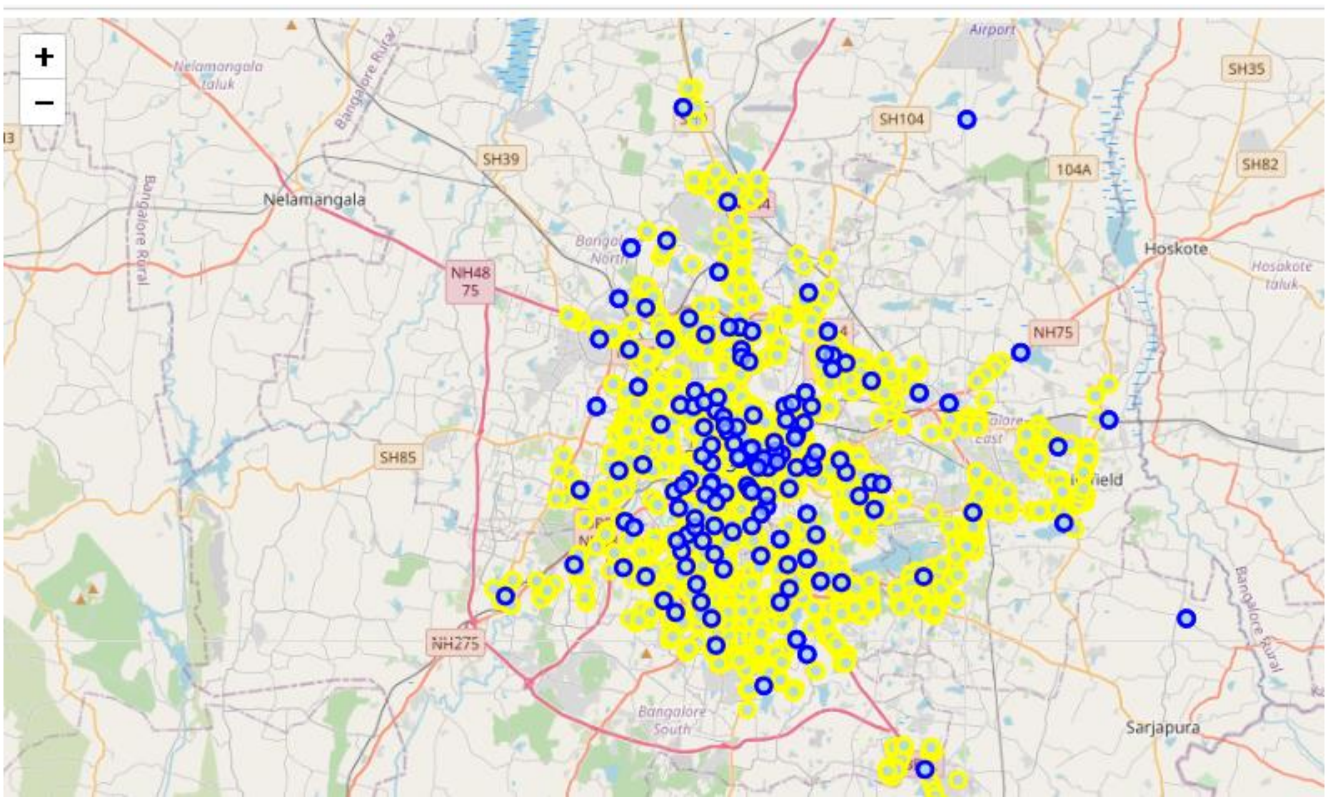
In the project, the calls were made to fetch all the venues near by the localities with a radius of **Two Kilometers.**

The response was a JSON file with venue details and its category. **Category** is our interest, hence using a function, the category pertaining to a locality was retrieved.

The final response and processed data from the FourSquare API is as **shown below**:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	MG Road	12.9777	77.6019	Hard Rock Cafe Bengaluru	12.976389	77.601468	American Restaurant
1	MG Road	12.9777	77.6019	M. Chinnaswamy Stadium	12.978144	77.599223	Cricket Ground
2	MG Road	12.9777	77.6019	M.G Road Boulevard	12.975771	77.603979	Plaza
3	MG Road	12.9777	77.6019	The Entertainment Store	12.975413	77.603045	Toy / Game Store
4	MG Road	12.9777	77.6019	Blossom Book House	12.975042	77.604813	Bookstore

All the localities and their venues were plotted again on a map to visualize better:



The **Blue marked** ones are Localities **and Yellow marked** ones are the venues near by localities, a click on the marker will popout further details about the plot.

Then by use of **One Hot encoding**, **the top ten** most common venues at a localities were found using a function.

Results as follows:

	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	100 Ft / CMH Road	Indian Restaurant	Pub	Café	Ice Cream Shop	Italian Restaurant	Burger Joint	BBQ Joint	Pizza Place	Cupcake Shop	Lounge
1	Avlahalli	Indian Restaurant	Gas Station	Restaurant	Café	Donut Shop	Diner	Discount Store	Dive Bar	Doner Restaurant	Women's Store
2	Bomanahalli	Indian Restaurant	Café	Ice Cream Shop	Sandwich Place	Tea Room	Chinese Restaurant	Pizza Place	Asian Restaurant	Bakery	Hotel Bar
3	Hanumantnagar	Indian Restaurant	Fast Food Restaurant	Café	Breakfast Spot	Ice Cream Shop	Coffee Shop	Park	Snack Place	Sandwich Place	Juice Bar
4	Jakasandra	Indian Restaurant	Café	Pizza Place	Ice Cream Shop	Italian Restaurant	Coffee Shop	Pub	Department Store	Snack Place	Chinese Restaurant

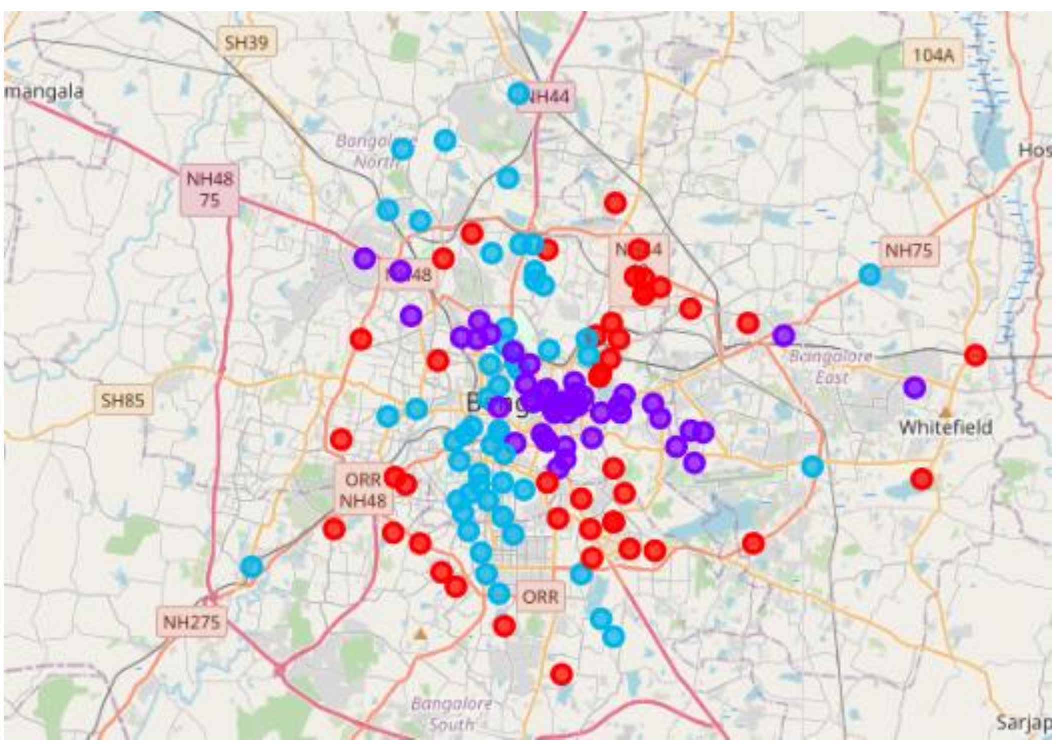
The corresponding one-hot encoded and averaged values were further used for applying **Machine Learning technique to better understand** and group data.

K-Means clustering algorithm was the most preferred classification algorithm, this is because it scales to large datasets with accurate results and it is simple to implement.

With K-Means various cluster values from **3 to 12 were tried**, the clusters with **KClusters=5** was the best cluster observed. In other clusters one cluster had more weightage and covered majority portion, this shows poor classification.

This data was further concatenated **with Coordinates** data to plot it on a Map.

Results were as follows:

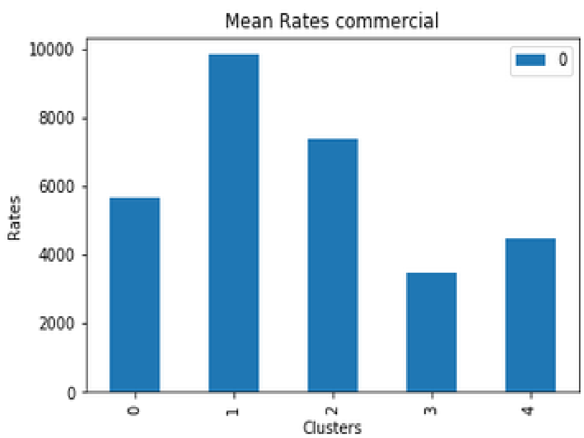


The classification data was now **used along with Residential dataframes** and Commercial data frames that were split up initially.

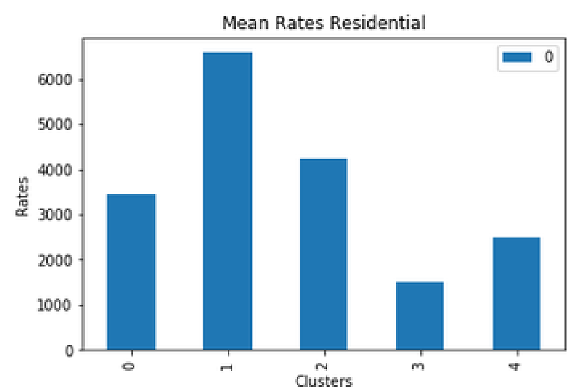
The data in the Residential and commercial dataframes were assigned with clusters to find out correlation

Also **the counts of clusters** in Residential and commercial category was found as follows:

```
         Area  Commercial_rate
Cluster
0          37               37
1          36               36
2          46               46
3           1                1
4           1                1
```




```
         Area  Residential_rate
Cluster
0          32                32
1          29                29
2          34                34
3           1                 1
4           1                 1
             0
0  3437.500000
1  6579.310345
2  4247.058824
3  1500.000000
4  2500.000000
```



The Bar plots above show the Mean land rates for commercial and residential lands respectively.

Further for each cluster, the **top venues were analysed** as shown below:

```
In [492]: #Cluster1
          Neighborhood_venues_sorted.loc[Neighborhood_venues_sorted['Cluster Labels'] == 1, Neighborhood_venues_sorted.co
          lumns[[0] + list(range(4, Neighborhood_venues_sorted.shape[1]))]]
```

Out[492]:

	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	100 Ft / CMH Road	Indian Restaurant	Pub	Café	Ice Cream Shop	Italian Restaurant	Burger Joint	BBQ Joint	Pizza Place	Cupcake Shop	Lounge
5	Krishanrajpuram	Fast Food Restaurant	Coffee Shop	Clothing Store	French Restaurant	Donut Shop	Café	Sporting Goods Shop	Shopping Mall	Lounge	Bowling Alley
7	Murgeshpalya	Indian Restaurant	Restaurant	Café	Ice Cream Shop	Bar	Pub	Fast Food Restaurant	Burger Joint	Hotel	Korean Restaurant
16	Airport Rd	Indian Restaurant	Hotel	Lounge	Café	Brewery	Ice Cream Shop	Park	Asian Restaurant	Breakfast Spot	Pub
20	BLR – SOUTH	Coffee Shop	Café	Airport Service	Airport Terminal	Brewery	Doner Restaurant	French Restaurant	Beer Bar	Sandwich Place	Taxi Stand
21	BLR – SOUTH-WEST	Coffee Shop	Café	Airport Service	Airport Terminal	Brewery	Doner Restaurant	French Restaurant	Beer Bar	Sandwich Place	Taxi Stand
31	Cambridge Rd	Indian Restaurant	Café	Pub	Chinese Restaurant	Hotel	Tea Room	Andhra Restaurant	Ice Cream Shop	Bar	Brewery
35	Church Street	Hotel	Indian	Lounge	Pub	Ice Cream	Brewery	Café	Shopping Mall	Japanese	Tea Room

Results:

As per the analysis of cluster data and the Residential/commercial real estate rates, the following observations were made:

- **Cluster 1** has many eateries, all round facilities.
- **Cluster 3** The residential and commercial rates are cheap and locality also is fine
- **Cluster 0** An average City area, this means, its a perfect combination of Residential and commercial space.
- **Cluster 4** Facilities are less and rates are more
- **Cluster 2** The rates are below average and adequate facilities.

Discussions:

On a whole, if a person wants to start with small business, Cluster 3 locality is preferred. Cluster 2 is slightly costly but has adequate facilities than Cluster 3,this can be second preferred option for starting small businesses.

If a person wants to start with Fast moving business that requires more visitors, Cluster 1 or Cluster 0 is most preferred. Cluster 1 has many eateries and businesses, this means it is a Commercial Business District. Starting Pubs or restaurants there would be advantageous

Cluster 4 is not a preferred locality for Residential purpose as facilities are less and Rates are also more.

Cluster 2 is most preferred for Residential Purpose followed by cluster 3 and cluster 0.

Conclusion:

In this Report, the Booming city of Bengaluru and the problem of finding out best investment locality with various investment strategies was discussed. By using data visualization on Maps, Foursquare APIs, and analysis, the resulting Data was applied to K-Means algorithm. Results showed proper classification of localities into clusters. Further, the real-estate rates was used to find correlation and better solving of the aforementioned problem. The solution showed what each cluster was best suited for. Using this data analysis any naïve investor can easily invest capital at the right locality. With this, the report is concluded.

Thank you.