Predicting the best venues to visit in Bengaluru to experience the Nightlife

Coursera Capstone Project on Applied Data Science

- **Submitted by Parag Ghungrudkar**

1. Introduction:

Bengaluru city is well known for its nightlife. Thousands of Bengalurians look for escape venues on Friday nights and weekends. However, finding the best venues for hangouts is a conundrum, given the plethora of choices available. It always boils down to a tradeoff between budget, rating, likes, and distance. Since the number of choices available are huge, one might make the wrong choice and end up disappointed.

This Data Analytics project intends to help make better choices by clustering similar venues based on three parameters: Rating, Price, and the number of likes. The final cluster will aid in sorting out better venues based on priorities. This project can also be of interest for businesses looking to tap on to the business opportunity from the bustling nightlife of Bengaluru. They could use this information to find out the best venues in each category, and further explore the venues to find out what works for them.

2. Data Acquisition and cleaning:

2.1. Data Acquisition:

I preferred to use the Foursquare database to collect the necessary data for the analysis. The two end points used and their function in data collection is as described below:

- **Venue Categories:**

  Description: Returns a hierarchical list of categories applied to venues.
  Used this end point to collect venues under different categories of nightlife. I used the following categories ID:

Category	**Category ID**
Beer Bar	56aa371ce4b08b9a8d57356c
Beer Garden	4bf58dd8d48988d117941735
Pub	4bf58dd8d48988d11b941735
Hookah Bar	4bf58dd8d48988d119941735'
Sports Bar	4bf58dd8d48988d11d941735
Brewery	50327c8591d4c4b30a586d5d
Lounge	4bf58dd8d48988d121941735

- **Venue Details:**

  Description: Gives the full details about a venue including location, tips, and categories.
  I used this end point to get data about all the venues. This included rating, like counts, and price range.

2.2. Data Cleansing:

The initial data from the FourSquare database was in JSON format. I converted this data into panda's dataframe. From the collated data frame, I observed that metric parameters were empty for many venues. I decided to remove these venues from the analysis since they could not be clustered due to lack of information. As a result of this, the final list of venues filtered down significantly.

2.3. Feature Selection:

Next I needed data to cluster these venues based on some metrics.
With the initial data in hand, I decided to cluster the venues based on following three parameters:

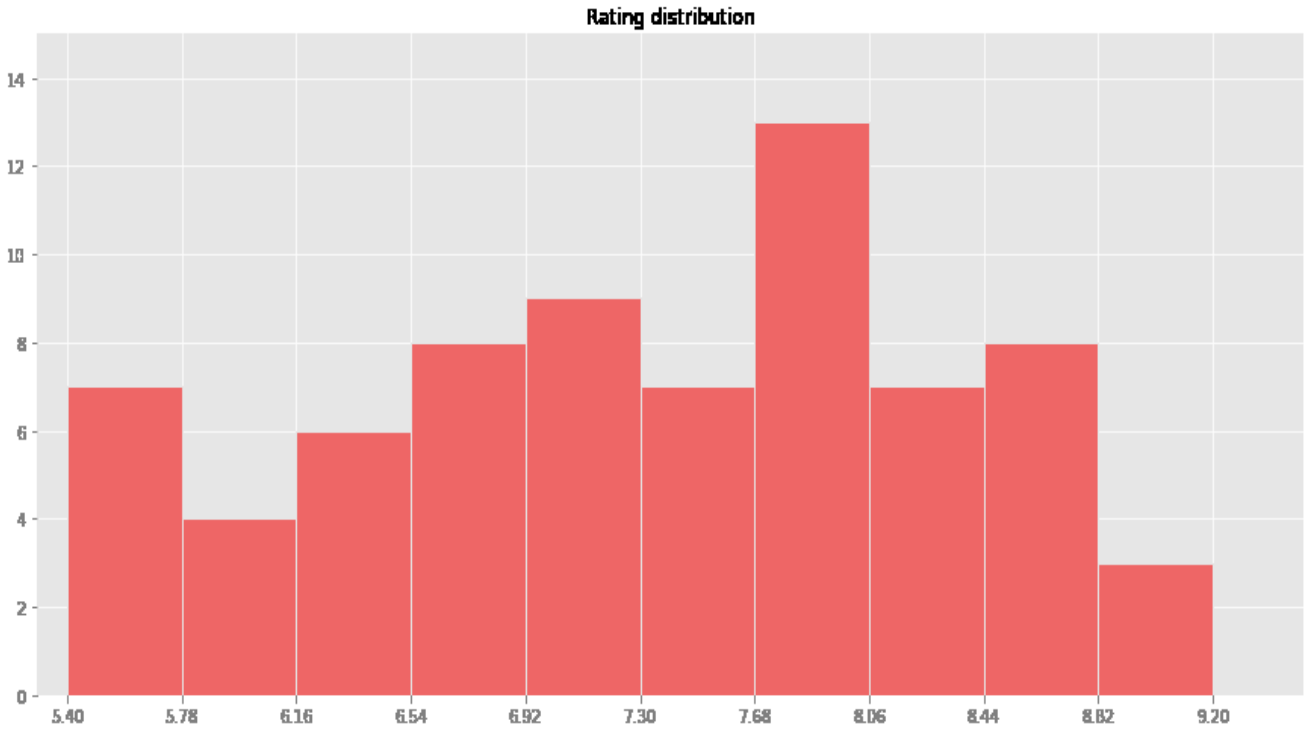
- Price : Cheap, Moderate and Expensive
- Rating : On a scale of 1 to 10
- Likes : Based on user likes from FourSquare

3. Exploratory Data Analysis:

I turned to histograms to understand the frequency distribution of venues for the three metrics. The distribution graphs gave some interesting insights about the data.

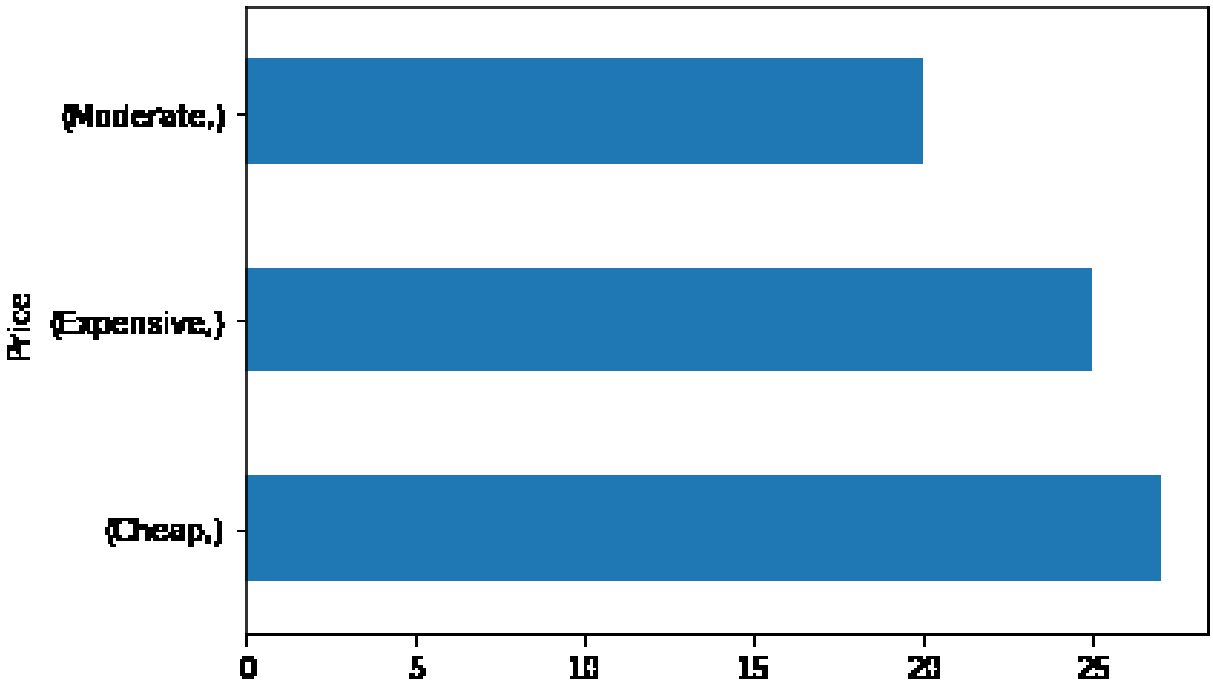
3.1. Rating Distribution:

As in the normal distribution curve, one will expect most of the venues to have average ratings. This was indeed true in this case. Most of the venues had average ratings (here I define average as a decent rating between 7-8 that we usually see as average values). It is interesting to note that more number of venues were on the left side (bad ratings) than on the right (good ratings). This supports the statement - 'The number of good options are always limited'.



3.2. Price Distribution :

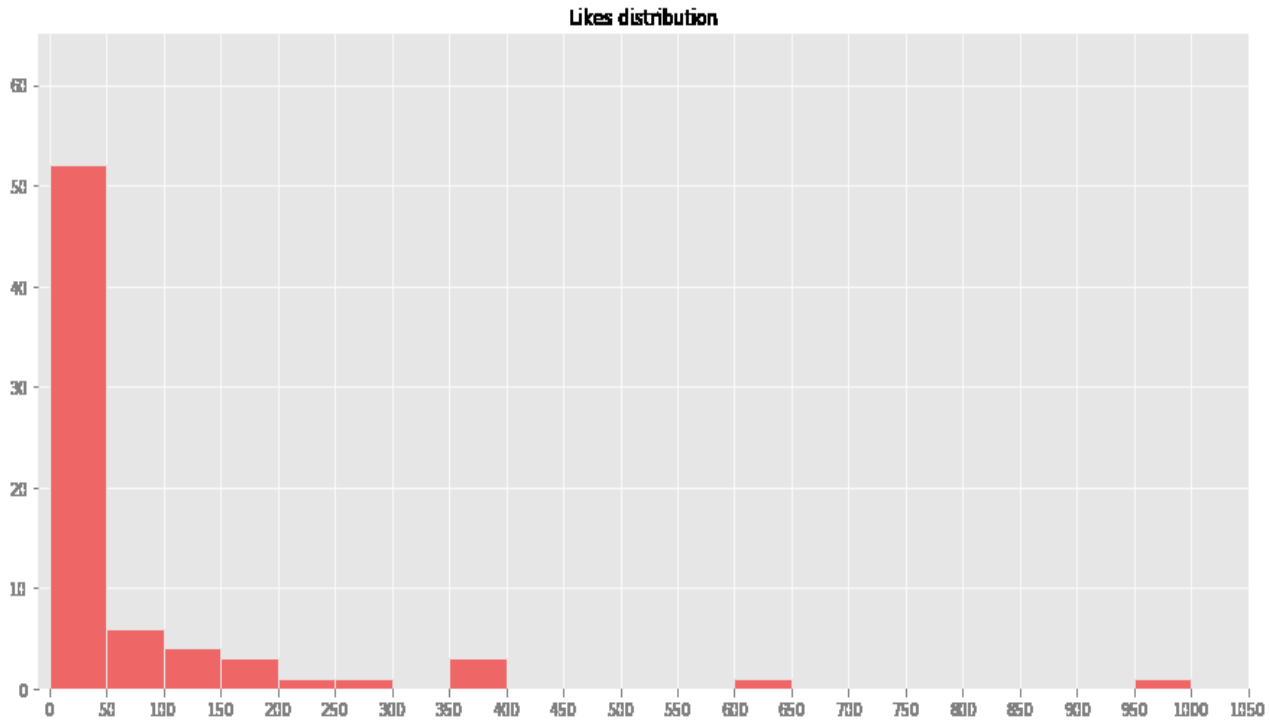
I used horizontal bar graphs to understand the distribution across different price ranges. The distribution across the three categories is almost uniform. This uniformity helped in getting better clusters during the analysis.



3.3. Like counts Distribution :

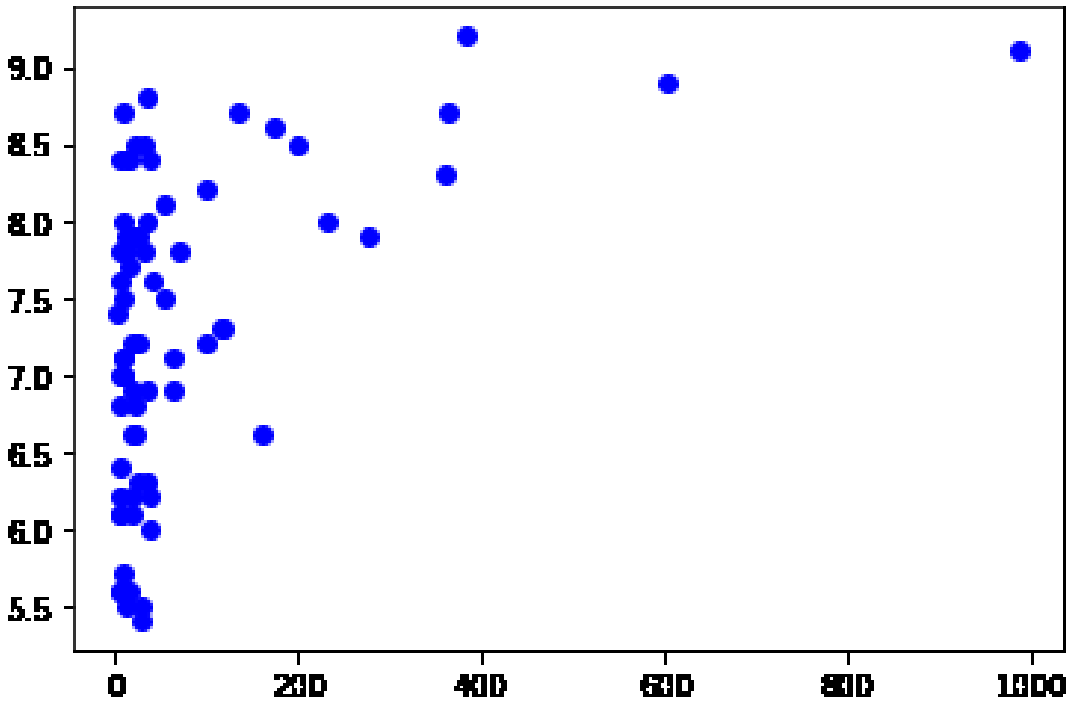
This particular histogram is quite interesting. The frequency distribution is sporadic which was not in the ratings case that we saw earlier. Main highlights from the above histogram are as follows:

Near about 50 venues fell in the range of 0-50. These venues could be either less popular or less like by the visitors. To investigate this further, I plotted a scatter plot for 'number of likes Vs. Rating'. There are few outliers in the histogram that have likes in the range of 950-1000. I used scatter plots to see if they are the same venues with maximum ratings.



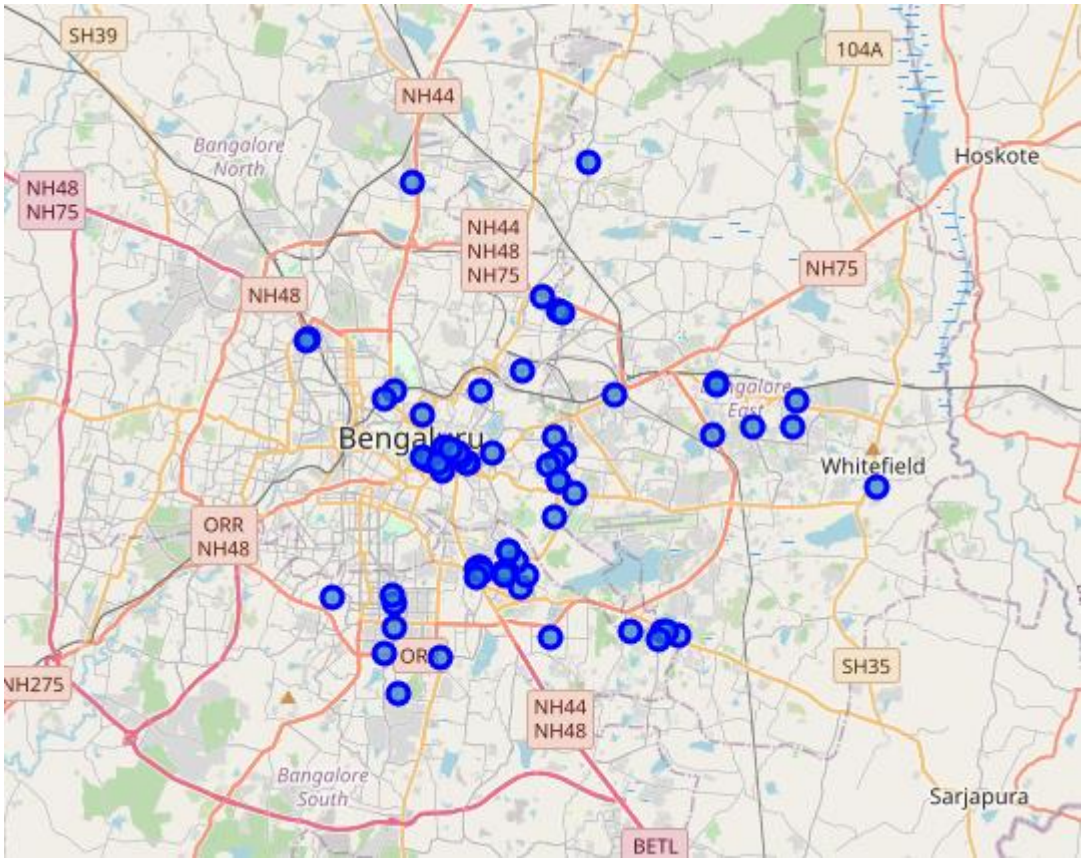
3.4. Likes counts Vs. Rating

I used a scatter plot to understand the correlation between the two parameters. The plot supported by the previous stated hypothesis that Venues with more rating indeed have more number of likes. Though there are many venues that have less number of likes but good rating. My hypothesis is that these might be new venues which might not be popular yet.



3.5. Plotting venues using folium library:

Using a folium library, I plotted the venues on the map. I can be noted from the map that most of the venues are clustered together. These locations must be the city centers. I preferred not to digress from the current investigation, but it would be interesting to see if location has any effect on the number of likes and Rating**.**



4. Clustering Model:

I settled on the Kmeans clustering algorithm from the sci-kit learn library to cluster the venues into different groups.
In the initial revisions, I executed the clustering of the venues without normalizing the parameters. However, it occurred to me that the clusters were biased towards the number of likes on analysis. Though the cluster gave good insights, I decided to normalize the data to remove the bias. One unintended but fortunate effect of this was that pricing had a more significant bias now. As a result, the venues sorted themselves out based on price category.

I decided the number of clusters based on the following assumption:

- Under the pricing column, the venues fall under Cheap, Moderate, and Expensive.
- Under the Rating column, the venues fall under: Bad, Decent, and Good.
- From high-level permutation and combinations, we can have 3*3 = 9 possibilities.

5. Results

The Kmeans clustering algorithm resulted in 9 clusters. On analyzing the type of venues in each cluster based on the three metrics, I labeled each cluster as follows:

5.1. Cluster 0: 'Moderate, good but not well known'

Venues under this cluster are under moderate price range category. They have good ratings (7.5 to 8.5). However, these venues have less number of likes which points out that they might not be well known.

name	Cluster Labels	Rating	Price	Likes
Windmills Craftswork	0	8.7	Moderate	9
Big Brewsky	0	8.6	Moderate	172
The Druid Garden	0	8.5	Moderate	20
XOOX Brewmill	0	8.4	Moderate	11
Uru Brewpark	0	8.4	Moderate	5
153 Biere Street	0	8.2	Moderate	97
Bangalore Brew Works	0	8.1	Moderate	54
Communiti	0	8.0	Moderate	33
Biergarten	0	7.9	Moderate	23
Byg Brewski	0	7.8	Moderate	10
The Pump House	0	7.5	Moderate	7
Brahma Brews	0	7.4	Moderate	3

5.2. Cluster 1: 'Cheap and Decent'

name	Cluster Labels	Rating	Price	Likes
House of Commons	1	8.4	Cheap	14
The Permit Room	1	8.4	Cheap	36
Dublin	1	7.9	Cheap	19
Hammered	1	7.9	Cheap	17
The Local	1	7.8	Cheap	32
Vapour - Pub and Brewery	1	7.8	Cheap	6
Guzzlers Inn	1	7.5	Cheap	53
The Open Box	1	7.2	Cheap	19
Harry's Singapore	1	7.1	Cheap	8
Le Rock	1	7.1	Cheap	62
Prost	1	7.0	Cheap	6
Gilly's Redefined	1	7.0	Cheap	9
Jimi's Beer Cafe	1	6.9	Cheap	62
J Cubez	1	6.9	Cheap	34
Barleyz - The Brew House	1	6.6	Cheap	161

5.3. Cluster 2: 'Expensive and Decent'

name	Cluster Labels	Rating	Price	Likes
Whitefield Social	2	8.8	Expensive	34
The 13th Floor	2	8.7	Expensive	133
Koramangala Social	2	8.5	Expensive	30
Skyye	2	8.5	Expensive	198
Fenny's Lounge & Kitchen	2	8.0	Expensive	230
eclipse lounge	2	8.0	Expensive	8
JW Marriott Executive Lounge	2	7.9	Expensive	12
High Ultra Lounge	2	7.8	Expensive	71
Bang	2	7.7	Expensive	14
The Whitefield Arms Café	2	7.6	Expensive	6
i-BAR	2	7.6	Expensive	40
LOFT 38	2	7.3	Expensive	116
The Fisherman's Wharf	2	7.3	Expensive	119
Sathya's	2	7.2	Expensive	100

5.4. Cluster 3:'Expensive, Excellent, and Very well known'

This particular cluster has only one venue. This is because the venue is outstanding when compared to others. It has a high rating and along with it maximum number of likes.

	name	Cluster Labels	Rating	Price	Likes
1	Toit Brewpub	3	9.1	Expensive	986

5.5. Cluster 4: 'Moderate, Good, and well known'

name	Cluster Labels	Rating	Price	Likes
Windmills Craftworks	4	9.2	Moderate	382
The Biere Club	4	8.3	Moderate	359
Big Pitcher	4	7.9	Moderate	276

5.6. Cluster 5: 'Moderate and Bad'

name	Cluster Labels	Rating	Price	Likes
The Pallet - Brewhouse & Kitchen	5	6.8	Moderate	6
The Local - Terrace Drinkery	5	6.2	Moderate	37
Brooks And Bonds	5	6.1	Moderate	5
Brewmeister	5	6.1	Moderate	6
3 Monkeys BrewPub	5	5.4	Moderate	29

5.7. Cluster 6: 'Expensive, good, and well known'

	name	Cluster Labels	Rating	Price	Likes	
2	Arbor Brewing Company	6	8.9	Expensive	603	1
6	Church Street Social	6	8.7	Expensive	362	1

5.8. Cluster 7: 'Cheap and bad'

name	Cluster Labels	Rating	Price	Likes
Extreme Sports Bar (Play Arena)	7	6.6	Cheap	18
Sherlock Holmes	7	6.6	Cheap	21
Hungry Hippie	7	6.4	Cheap	4
Jimis Beer Cafe	7	6.3	Cheap	33
Enigma - The Pub	7	6.2	Cheap	19
SH 17 Restaurant	7	6.2	Cheap	5
Mockaholic Restro Beer Cafe	7	6.0	Cheap	36
Cafe Coffee Day	7	5.6	Cheap	6
Dug Out Sports Bar	7	5.6	Cheap	14
SBX Sports & Music Cafe	7	5.6	Cheap	5
Harry's	7	5.5	Cheap	11
Coconut Grove	7	5.5	Cheap	27

5.9. Cluster 8:'Expensive, bad, and not well known'

name	Cluster Labels	Rating	Price	Likes
The Studio Bar	8	7.2	Expensive	24
Upbeat	8	7.1	Expensive	8
Firehouse	8	7.0	Expensive	9
Hoot	8	6.9	Expensive	18
Rasta S02E01	8	6.8	Expensive	20
Sutra	8	6.3	Expensive	24
Eden Park - The Lounge	8	6.1	Expensive	18
Loveshack	8	5.7	Expensive	8

6. Conclusion :

From the above analysis, I was able to sort venues based on these three metrics: Price, Rating, and likes. This analysis could be used to select the most appropriate venue for an outing based on preferences. It also forms a basis for the recommendation system.

From a business perspective, this analysis can be used as a starting point to research what an ideal venue should be to attract customers. Further, more libraries could be used to cluster the preferred venues based on venue, ambiance, etc. One interesting research would be to find whether the location has any effect on the three metrics. However, this analysis's scope is to cluster the venues based on the three metrics that have been achieved successfully.