Capstone Report

Sudeshna Sen

March 7, 2020

Index

1. Introduction …………………………………………………………………. 02
2. Data ……………………………………………………………………………... 02
3. Methodology ………………………………………………………………...11
4. Results ……………………………………………………………………………11
5. Discussion ………………………………………………………………………12
6. Conclusion ……………………………………………………………………. 12
7. References ……………………………………………………………………. 13

**Exploring neighbourhoods in Bangalore for renting apartments**

1. **Introduction**

Bangalore is the hub of all IT jobs in India with population of around **12.5 million**. The city has a population density of **4,381** persons per square kilometre and more than 50% of the population consist of migrants from other states of India.

The rising influx of job seekers in Bangalore has led to a sharp increase in the costs and rental rates of apartments all over the city. Rents vary from place to place depending on the area, accessibility, proximity to marketplaces and restaurants, etc. For those arriving in Bangalore from other states of India, it becomes difficult to make a choice without proper knowledge of the neighbourhoods and the rental rates in different areas.

There are several mobile apps which facilitate the process of apartment hunting but most of them do not provide data about the venues like restaurants, gyms and pubs in that area. With the help of Foursquare data and apartment listings from websites, this project aims to help make the process of apartment hunting easier for the potential client.

**1.1 Target audience**

* People moving to Bangalore from other states, looking to rent apartment.
* Visitors in Bangalore who wish to explore the neighbourhoods for restaurants, gyms and other facilities.
* Property owners looking to rent out apartments who want to get an idea of the price range in their area.

**1.2 Stakeholders**

* Landlords
* Renters
* Real estate agents
* Bangalore Government

1. **Data:**

The following sources of data has been used to create the project:

* 1. **makaan.com**  
     This is a website which provides real estate data of a city. I have used the price trends section of this website to get data about the rental rates across different regions in Bangalore. The information also contains data about the type of apartment (1bhk, 2bhk, 3bhk) and the locality.

Link: <https://www.makaan.com/price-trends/property-rates-for-rent-in-bangalore>

A screenshot of a cell phone

Description automatically generated

Figure 1: Snapshot of the data from the website

* 1. **blr\_neighborhoods.csv**

This file contains the following data: Neighbourhood names of Bangalore, latitude, longitude. This data has been obtained from Kaggle.com.

Link: <https://www.kaggle.com/rmenon1998/bangalore-neighborhoods>

A screenshot of a cell phone

Description automatically generated

Figure 2: Snapshot of the Kaggle Dataset

* 1. **Foursquare location data**

This is a JSON file obtained by calling the Foursquare API. This data consists of the different venues, the type of venues, and its coordinates in each locality.  
  
Link: <https://foursquare.com/>

1. **Methodology**
   1. **Data acquisition:**

**1. Rental rates:**   
  
The data for rental rates was extracted by Web scraping for which I used the BeautifulSoup package. From all the columns, the data that was extracted included the neighborhood name and the average rent for 1 BHK, 2BHK and 3BHK apartments in that area.

A screenshot of a cell phone

Description automatically generated

Figure 3: Rental rates data extracted into dataframe after web scraping

**2. Coordinates data:**   
  
The coordinates data was available on Kaggle.com as a csv file. I downloaded the csv file and read it into a dataframe using the read\_csv() method of Pandas. A screenshot of a cell phone

Description automatically generated

Figure 4: Coordinates data read from csv into a pandas dataframe

**3. Venues data:**

The venue details were obtained by calls to the Foursquare API. The list of latitudes and longitudes were provided and a maximum of 100 venues within a radius of 500m were obtained for each of the geospatial points passed.

A screenshot of a cell phone

Description automatically generated

Figure 5: The venue details obtained from API calls stored in a dataframe

Since I did not make the API call for any particular type of venue, the resultant dataframe has everything from gyms and yoga studios to restaurants and pubs.

* 1. **Data Cleansing**

**1. Rent data:**

* The data that was not available was indicated by a “-”. This data had to be cleaned and all occurrences of “-” were replaced by NaN.
* Next, all occurrences of duplicates were removed. Since the Web scraping was done by keeping the first 20 pages, there were quite few duplicates. The first occurrences of these were kept and the rest were dropped.
* Lastly, all the rental rates were stored as strings. To use them in any kind of analysis, they had to converted to floats.

**2. Coordinates data**

* The coordinates obtained from Kaggle had quite a few outliers. The first one was that some of the latitudes were positive while some were negative values. This was corrected by changing all latitudes and longitudes to their absolute values.
* Another problem was in that some of the neighborhoods had the suffix of a P.O. (Head post office) or S.O. (Sub post office). These suffixes had to be removed because they were not present in the neighborhood names of the rent data.
* The last problem with the data was that some of the coordinates pointed to regions outside Bangalore. This is visible when we visualize the data points using Folium maps. There seem to be a few points outside Bangalore city and quite a few even outside India. A plausible explanation would be that there exist regions with the same name as that the neighborhoods in Bangalore. So, when the data was extracted using Nominatim, there was an ambiguity in names and the regions outside Bangalore were returned.

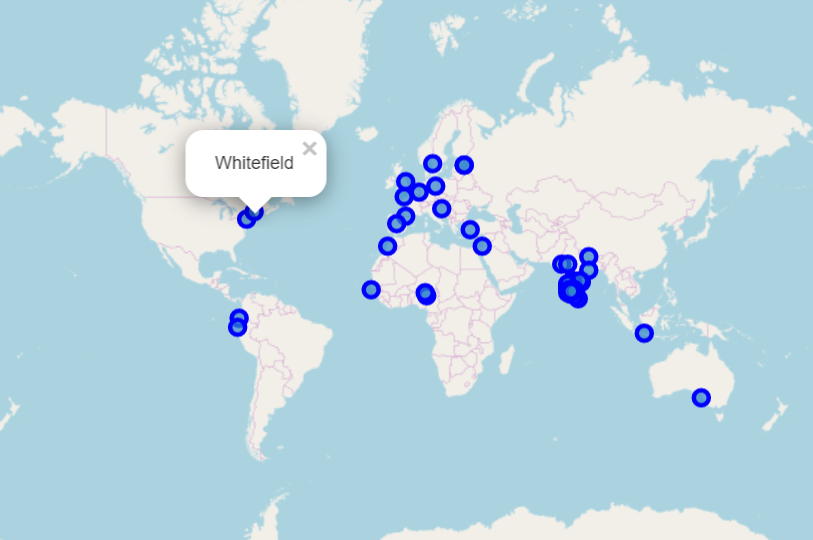


Figure 6: Visualizing outliers in coordinates data

For example, one of these regions happens to be “Whitefield” which lies in North eastern part of Bangalore. Upon a quick Google search, we see that exists a region in the US which shares the same name.

A screenshot of a cell phone

Description automatically generated

Figure 7: Outliers that needed to be treated

Since the coordinates data is needed for the neighborhoods data, I treated only those outliers (Fig. 7) which would impact the rental rates data. This was done by looking up the correct coordinates on Google and plugging them into the dataframe.

* 1. **Exploratory data analysis**

The rental rates data was then combined with the coordinates data to give a consolidated dataframe. Next, I calculated the average rent per head assuming that one person would always go for a single occupancy room. This means that the rent for 2BHKs had to be divided by 2 and for 3BHK, by 3 to get the rental rates per head.

Based on the average rent calculated above, I assigned a ‘rent\_bucket’ or label to each neighborhood. The buckets were assigned as follows:

1. Average rent less than Rs 8000 – **Low**
2. Average rent between Rs 8000 to Rs 10,000 – **Medium**
3. Average rent above Rs 10,000 – **High**

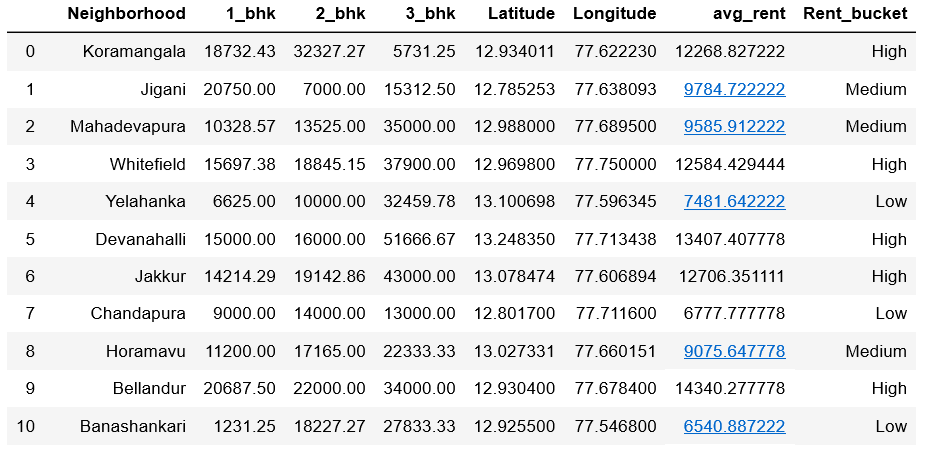


Figure 8: Dataframe after merging

* + 1. **Distribution of rental rates for different types of apartments:**

I plotted the variation in rents in 1BHK, 2BHK and 3BHK apartments respectively. This was done after calculating the rent per head for the three kinds of apartments in all the regions.

A screenshot of a cell phone

Description automatically generated

Figure 9: Rent distribution for different kinds of apartments

It is seen from the plot that the prices for 1BHK flats lie in the range of 8k to 15k while those for the 2BHKs have a lower range of about 6.5k to 10k. The 3BHK flats are the costly ones with prices ranging from 11k to 16k per head.

* + 1. **Count of regions based on rental rates**

The number of regions having low to medium rental rates are almost equal. But the regions having high rental rates surpass them. This observation agrees with the fact that Bangalore is an expensive city.

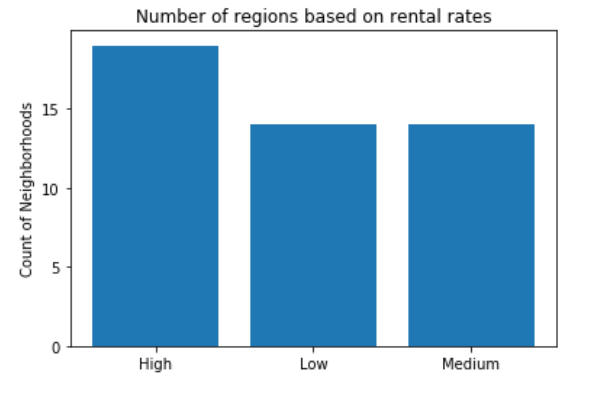


Figure 10: Distribution of neighbourhoods based on rental rates

* + 1. **Top 5 types of venues in Bangalore**

Next, I have explored the top 5 kinds of venues in Bangalore. This gives us an idea of the city. From the plot we can clearly see that the most common venues are restaurants and cafes.

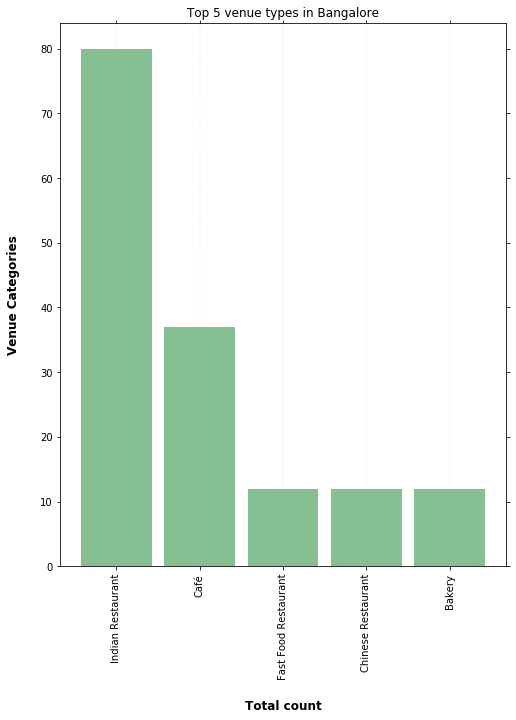


Figure 11: Top 5 venues in Bangalore

* + 1. **Change in rental rates with venue count in the neighborhood**

I plotted the correlation between rental rates and the number of venues in a neighbourhood. For the scatter plot, the prices of each neighbourhood are represented along the Y axis and the venue count along the X axis. To get a decent plot, both the Y and X values have been normalized such that all values lie between 0 and 1.

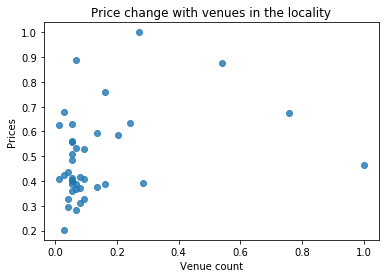


Figure 12: Distribution of rental rates with number of venues in that region

It was observed was that there is not much relation between the number of venues in a locality and the rental rates of the locality. This can be due to several factors:

* Foursquare does not have information about all the venues in an area.
* The rental rates are dependent on more factors like transportation, demographics of the region, how old that part of Bangalore is, etc.
* The radius for API calls is too small to include proper data.  
    
  + 1. **Most common venue by neighbourhood**

The final data exploration done was to determine the most frequently occurring venues in each region which was stored in a dataframe.

A screenshot of a cell phone

Description automatically generated

Figure 13: Dataframe containing the top 3 venues of each neighbourhood

* 1. **Clustering**

The neighborhoods have been clustered into 4 distinct groups depending on the rental rates and the most common types of venue in that region.

For proper clustering the data had to be encoded using the one hot encoding scheme. The encoding has been done for the rent bucket column too so that it is taken into consideration while creating the clusters.

First a map of Bangalore with the neighborhoods has been created using Folium. The geographical coordinates of Bangalore were obtained using Nominatim.

A picture containing text, map

Description automatically generated

Figure 14: Bangalore map created using Folium

Then the data was clustered using **K-means algorithm** which is one of the most common clustering techniques used in Data Science. The total number of clusters was set to 4.

1. **Results**

Once the clusters were created, each cluster was examined to find an appropriate label for it. One approach of labelling a cluster would be based on the most frequent venue. But since that would become very specific, the clusters were labelled based on the rent bucked and by manually examining each cluster. Finally, these clusters were visualized by superimposing them on the map created earlier.

A close up of a map

Description automatically generated

Figure 15: Visualization of clusters

We can see (Fig. 15) that the neighbourhood have been segregated into clusters. On clicking any of the cluster points, the cluster label appears which gives an information about the rental rates and the kind of venues common in that area.

1. **Discussion**

As mentioned before, Bangalore has a huge population density. This project only used data from one single property website. There are several other apps and websites providing real time data on rental rates that can be leveraged to make this project more dynamic. Moreover, the different neighbourhoods differ in size and it does not make justice to return venues within the same radius for all the neighbourhoods irrespective of their sizes.

The only criteria considered were the venues. Hence, we did not find anything conclusive about venue and rent dependency. But the complexity can be increased by taking into consideration factors like bus and metro routes, offices and IT parks in the locality, etc.

The project can be made more dynamic and complex and provides a huge scope for further development by considering the above factors.

1. **Conclusion**

In this project I clustered the neighbourhood in Bangalore base on the rental rates and the venues in that neighbourhood. From the clustering we can conclude that:

* Regions like Whitefield and Indiranagar have high rental rates and are places with a lot of Cafes and high-end restaurants.
* Regions like Banashankri, Kangeri and Malathalli have low rents and local eateries offering South Indian dishes.
* Regions like Hoodi and Mahadevpura have medium rent and have a good mix of different kinds of stores.
* The last cluster has Devanahalii and Jakkur. These regions have high rental rates and mainly have ATMs and Indian restaurants.

We can also say that Bangalore has a lot of eateries and restaurants compared to other types of venues and that it is cheaper to rent a single room in a 2BHK than a 1BHK.

Since people's budgets and tastes differ, this clustering visualization will help them to select a proper area to rent a place. It can also help them choose the most optimum place to rent a flat based on the proximity to their workplace.

1. **References**

[1] Rental rates: <https://www.makaan.com>

[2] Coordinates of Bangalore neighbourhoods: [https://www.kaggle.com](https://www.kaggle.com/)

[3] Venues data: <https://foursquare.com/>

[4] Bangalore – <https://www.wikipedia.org/>

[5] Neighborhoods coordinates for outlier correction: [www.google.com/maps/](https://www.google.com/maps/)