COURSERA CAPSTONE PROJECT - IBM DATA SCIENCE SPECIALIZATION

HOUSING DATA ANALYSIS

For the Neighborhoods of Mumbai

By: Devesh Poojari

Email: deveshpoojari7@gmail.com

2020

Table Of Contents

About the Report

This report is for the final course of the Data Science Specialization. A 9-course series created by IBM and hosted on Coursera Learning platform, it covers the entire process in Data Science from collecting data to making predictive conclusions based on the data.

The problem and analysis approach are left for the learner to decide and explore on his own, with a requirement to leverage the Foursquare location data to explore or compare neighbourhoods or cities of our choice or to come up with a problem that we can use the Foursquare location data to solve.

Introduction

Mumbai, formerly Bombay, city, capital of Maharashtra state, southwestern India. It is the country's financial and commercial centre and its principal port on the Arabian Sea. Located on Maharashtra's coast, Mumbai is India's most-populous city, and it is one of the largest and most densely populated urban areas in the world. It was built on a site of an ancient settlement, and it took its name from the local goddess Mumba—a form of Parvati, the consort of Shiva, one of the principal deities of Hinduism—whose temple once stood in what is now the southeastern section of the city.

In addition, the city's commercial and financial institutions are strong and vigorous, and Mumbai serves as the country's financial hub. It suffers, however, from some of the perennial problems of many large expanding industrial cities: air and water pollution,

widespread areas of substandard housing, and overcrowding. The last problem is exacerbated by the physical limits of the city's island location. Area about 239 square miles (619 square km). Pop. (2001) 11,978,450; urban agglom., 16,434,386; (2011) 12,478,447; urban agglom., 18,414,288.

Housing is largely privately owned, though there is some public housing built by the government through publicly funded corporations or by private cooperatives with public funds. Mumbai is extremely crowded, and housing is scarce for anyone who is not wealthy.

(For that reason, commercial and industrial enterprises have found it increasingly difficult to attract mid-level professional, technical, or managerial staff.) In an attempt to stem the ongoing immigration of unskilled labour that has increased the city's indigent and homeless population, city planners have encouraged enterprises to locate across Mumbai Harbour—notably in Navi ("New") Mumbai—and have banned the development and expansion of industrial units inside the city; their efforts, however, have been largely unsuccessful.

Because of the limited physical expanse of the city, the growth in Mumbai's population has been accompanied by an astounding increase in population density. By the early 21st century the city had reached an average of some 77,000 persons per square mile (29,500 per square km). The settlement is especially dense in much of the city's older section; the wealthy areas near Back Bay are less heavily populated.

With the above-mentioned problem in the city of Mumbai, it becomes increasingly important to analyse the housing data.

Objective:

The objective of this Capstone Project is to build a system with the help of data to:

1) Analyze property prices and show the property prices in forms of a heatmap to understand the distribution of property prices across the city

2) Cluster similar Neighbourhoods and analyze the types of Neighbourhoods in the city.

3) To check whether the surrounding venues affect the price of a house. If so, what types of venues have the most affect.

Target Audience:

The target audience for this report are:

1) Potential Buyers who can roughly estimate the value of a house based on the surrounding venues and average price.

2) Real Estate makers and planners who can decide what kind of venues to put around their products to maximize selling price or the other way round.

3) House sellers who can optimize their advertisement.

Data Description

The data for this project has been retrieved and processed through multiple sources, giving careful consideration to the accuracy of the methods used to solve the problem, we will need the following data:

- Data containing Neighborhoods of Mumbai and average housing prices. This defines the scope of this project which is confined to Mumbai.
- Latitude and Longitude data of all the above Neighborhoods. This is required in order to plot the map and also get the venue data.
- Venue data, particularly data related to venues in the vicinity of neighbourhoods. We will use this data to perform clustering on the neighbourhoods.

Sources of Data and Method to extract them:

1. The required data of the neighbourhood and their housing prices was collected from 99 acres.com. 99Acres is an Indian real estate database website founded in 2005. At the time of collecting the data, there were 906 neighbourhoods spread out in 9 distinct sub-cities. We will be using Web Scraping techniques to extract the data from the website, with the help of Python requests and BeautifulSoup packages. There were originally 7 columns but by using data transformation steps, only the

required columns were extracted. Columns : **[** sub_city, neighborhood, Avg_price, Avg_1BHK_rental, Avg_2BHK_rental, Avg_3BHK_rental **].** The collected data is stored as CSV for further use in the project.

2. The Latitude and Longitude Data of the 906 neighbourhoods were obtained by normal Google Search with automated web scraping. Here we will use BeautifulSoup and Python requests to parse data from HTML to the required format and automate the process of searching on Google and scraping the data with the help of Selenium in python and store the data in CSV format for further use in the project.
3. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data and help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (BeautifulSoup), automation(Selenium), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and maps visualization (Folium).

Methodology

1. Firstly, we need to get a list of neighbourhoods in the city of Mumbai. Complete data were available on a real estate site called 99 acres.com. It has almost all neighbourhood of Mumbai along with their housing price. The table consisted many columns but only the ones needed were scraped [SubCity, Neighborhood, Price_range, q/q, Trend, 1BHK rent (range), 2BHK rent (range), 3BHK-rent (range)]. The table was web scraped using Python requests and beautiful soup packages to extract the list of neighbourhoods data. The scarped data was cleaned and brought to the required form.

	sub_city	place	q/q	Avg_price	Avg_1BHK_rental	Avg_2BHK_rental	Avg_3BHK_rental
0	Mumbai Thane	Anand Nagar	1.03	8181.5	11526.0	16830.0	22100.0
1	Mumbai Thane	Ashok Nagar	0.00	9541.0	0.0	19763.0	0.0
2	Mumbai Thane	Balkum	2.24	9371.5	14171.5	19337.5	27431.0
3	Mumbai Thane	Balkum Pada	0.48	8818.5	12571.5	16618.0	0.0
4	Mumbai Thane	Bhaskar Colony	0.00	14747.5	0.0	0.0	0.0
5	Mumbai Thane	Bhayandarpada	0.00	7458.5	0.0	14450.0	0.0
6	Mumbai Thane	Brahmand	0.00	8861.5	12478.5	18360.0	0.0
7	Mumbai Thane	Budhaji Nagar	3.63	7735.0	0.0	0.0	0.0
8	Mumbai Thane	Charai	4.26	11156.5	13923.0	0.0	0.0
9	Mumbai Thane	Chirak Nagar	0.38	11071.5	18925.5	21542.5	0.0

Fig 1. Housing Price Data

2. Second Data required is the coordinates data, which would consist coordinates of all neighbourhood contained in the Housing Price Data. To get this, Automated Web Scraping was implemented. Coordinates were web scraped from Google search result and searching was Automated with Selenium. As Geocoder was not working, this method is not that efficient. Some of the searches didn't give a proper result and required more cleaning. Finally, the data looked like this which was stored in a csv file.

	name	lat	long
0	Ambawadi	19.2411	72.8600
1	Ambedkar Nagar	19.0730	72.8789
2	Anand Nagar	19.2557	72.8656
3	Belpada	19.0395	73.0576
4	Chikholi	19.0760	72.8777
5	Gandhi Nagar	19.0585	72.8479
6	Ganesh Nagar	19.1327	72.8796
7	Geeta Nagar	18.9121	72.8064
8	Hanuman Nagar	19.1954	72.8671
9	Kailash Nagar	19.1973	72.8439
10	Laxmi Nagar	19.1608	72.8323
11	Manda	19.0760	72.8777
12	Moti Nagar	19.1674	72.9350
13	Nehru Nagar	19.0640	72.8826
14	Palm Beach	Not Available	Not Available
15	Pant Nagar	19.0845	72.9105

Fig 2. Coordinates Data

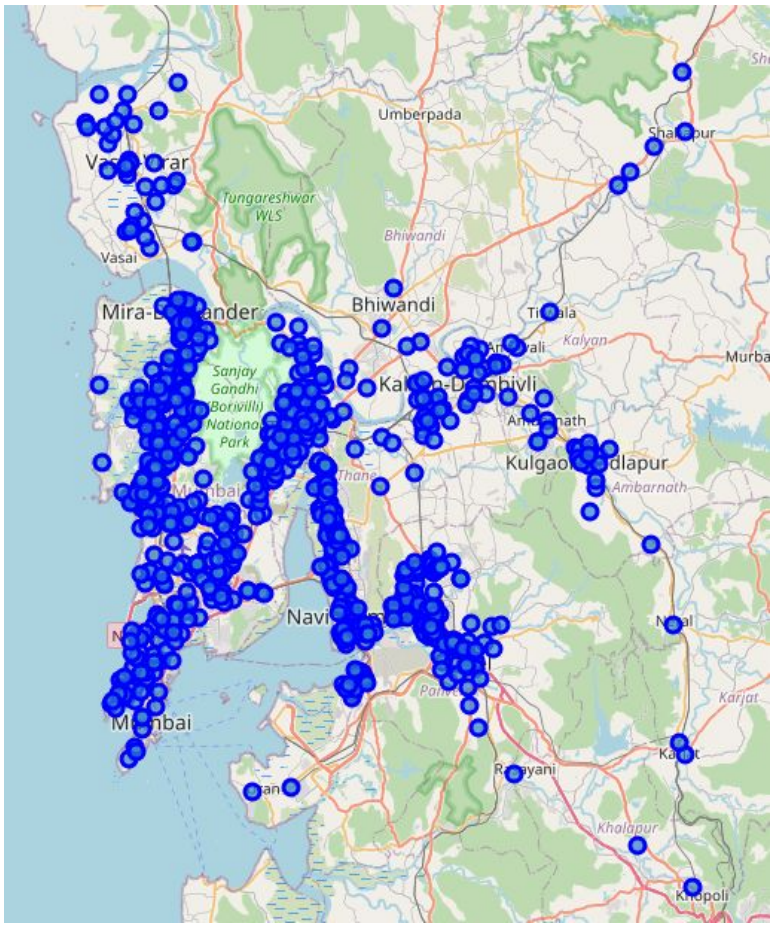

Fig 3. Distribution of Coordinates on Map

The Map was plotted by using Folium, another python library. The data points have a good distribution and cover almost all of Mumbai.

3. There were many outliers in the data as shown.

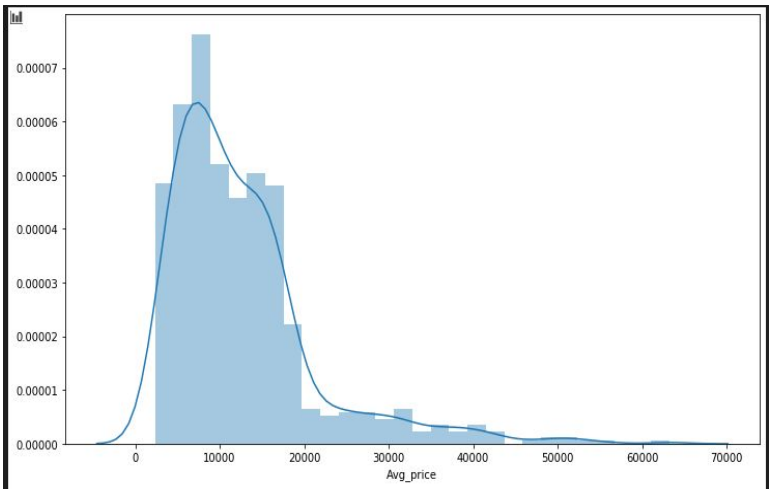

Fig 4. Distribution of Average Price of Houses

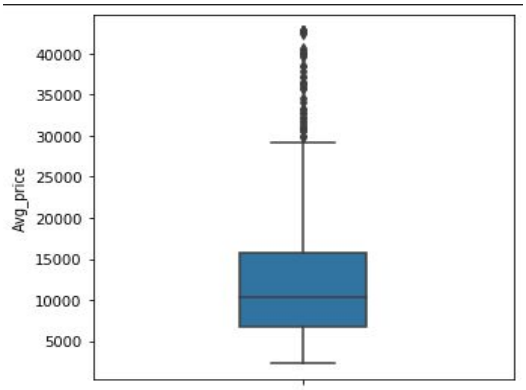

Fig 5. Box Plot of Average Price

On further analysis, these outliers are not errors caused during data collection or anything else, but actual values as we know housing prices downtown are exhorbitantly high. But we can't keep these outliers like this. So we'll treat them by a method called "creating caps". We'll change all outliers average price to the lowest outlier average price as marked in the image below.

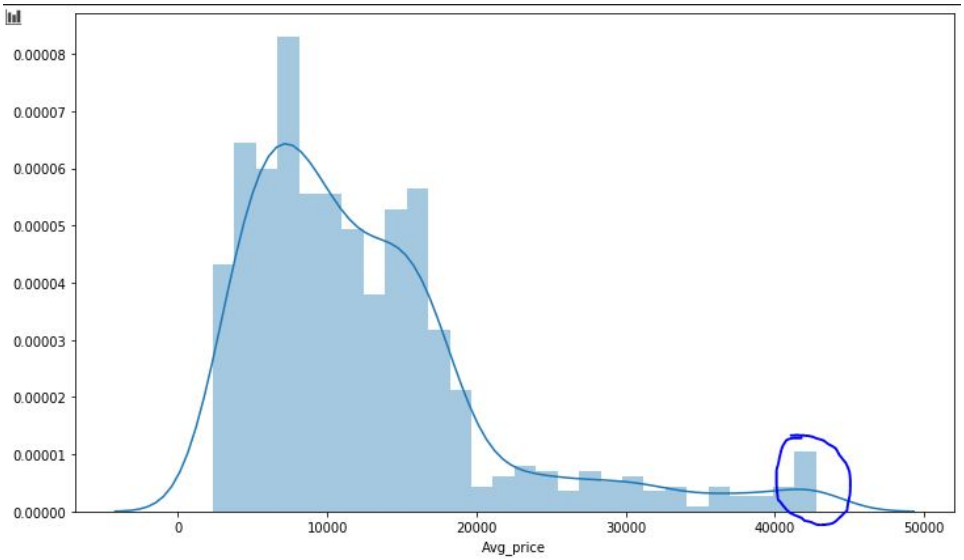

Fig 6. Distribution of Average Price of Houses (After caping)

4. The Heat Map for average housing price is plotted to get a good view of the density of data location and relation between Average Housing Price and density of houses in the same region. We can see that the heat map shows high Average Price of the house mainly in three regions as marked in the image. Out of these, the region in South Mumbai has a very low number of houses as compared to Navi Mumbai and Suburbs in the data hence that concludes that the houses in this region are much costlier than anywhere in Mumbai.

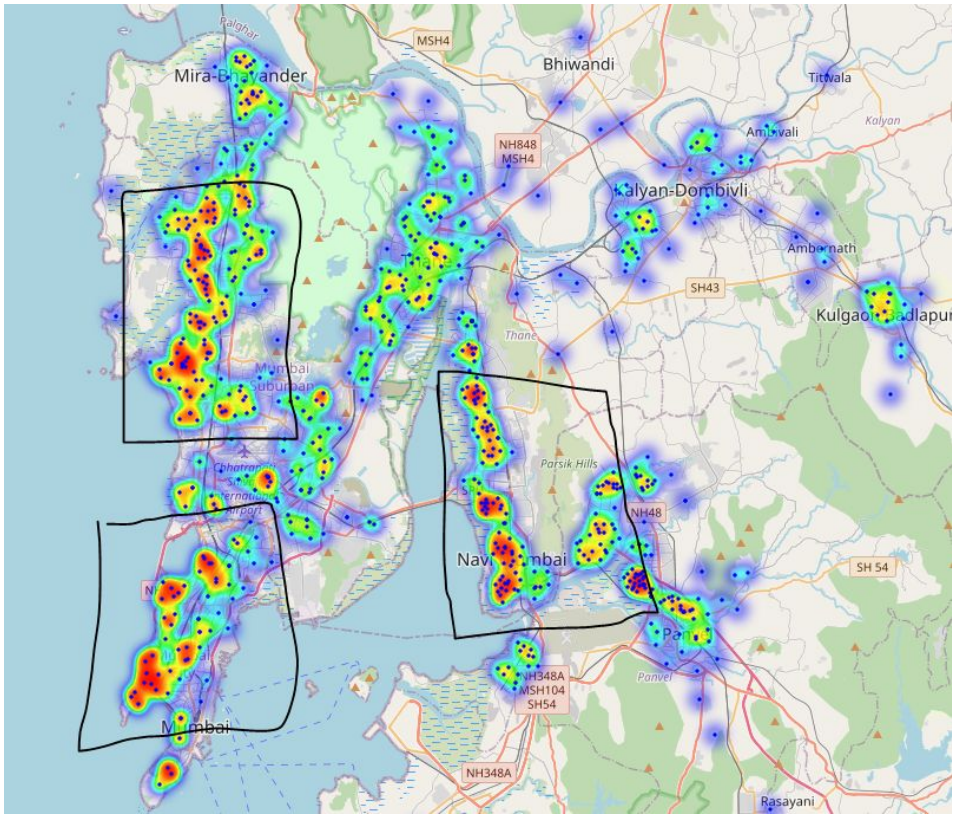

Fig 7. Average Price Heat Map

5. The Housing Price was binned and then plotted on the Map. This map verifies our above conclusion of housing prices in South Mumbai as compared to other parts of Mumbai. Furthermore, as it was in Introduction, the region of Navi Mumbai has been planned for housing only with no factories and industries in the city hence we can see a high volume of housing data in the database which are cheap, which again

supports the major cause of developing the project of building houses for mid-level professional, technical, or managerial staff working in the main city.

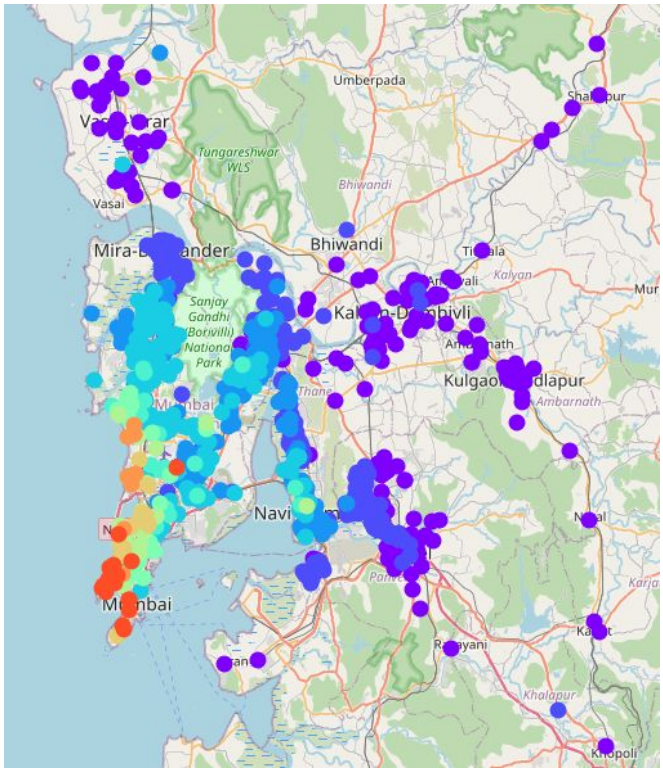

Fig 8. Distribution of Binned Prices

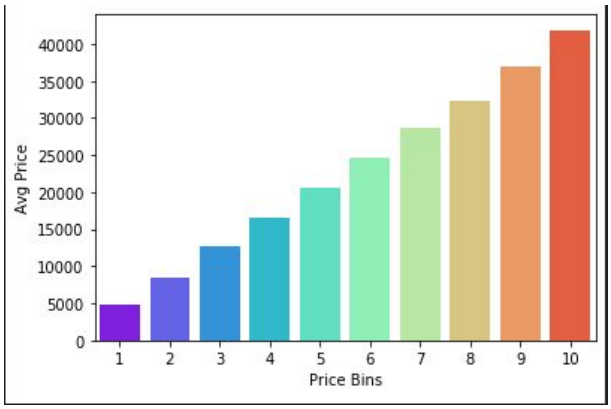

Fig 9. Legend for the Map

From the Above Figures, we can conclude that “Prices increases as we go South”.

6. Next, we will use the Foursquare API to get the top 100 venues that are within a radius of 1000 meters. There are a lot of Venues in Navi Mumbai yet they have cheapest housing prices in the city. This could mean that venue doesn’t have a major impact on housing prices. This can be a strong point and help in further analysis.

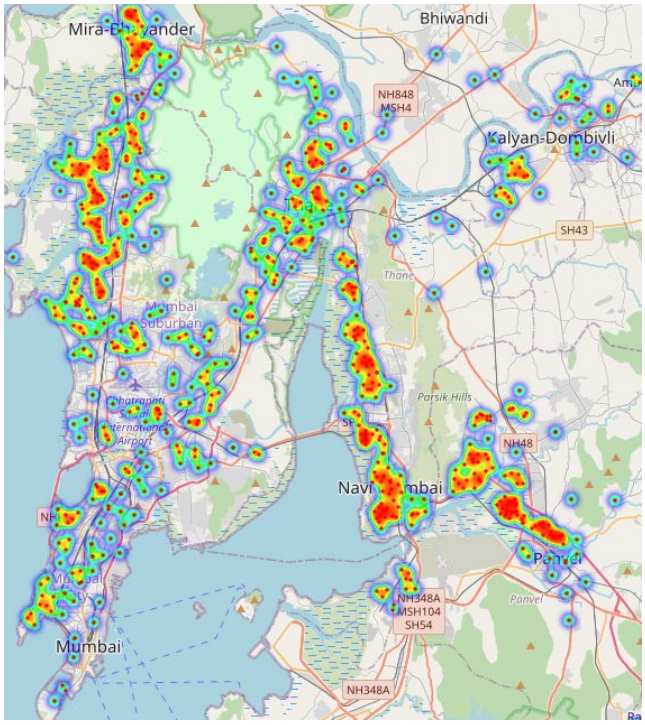

Fig 10. Heat Map of the number of Venues

7. The total types of venues received from the Four-Square API were 268. These types are often overlapping and similar, in terms of their ability to contribute to clustering of Neighbourhoods. Mapping venue category to a more general type of category to group same kind of venue category into one category.

0	stores_daily_conveniences	7	cafe_fastfoods
1	regular_restaurants	8	bars_nightlife
2	transport_vicinity	9	cuisine_restaurants
3	arts_culture_recreation	10	tourist_interest
4	shopping	11	kids_family_residential
5	nature_view	12	business_hub
6	sports_fitness	13	education_colleges

Fig 11. Venue Mapping

8. After Mapping the venues, top 10 venues of each neighbourhood is extracted. This can give good information about the type of neighbourhood. Further can be visualized for better understanding the clusters formed in the next step

9. Next Clustering is performed using K-means Clustering taking the value of K = 10. Clustering of Neighborhood gives the 10 clusters of neighbourhood based on the count of a type of venues in each neighbourhood. Clustering will bring all the same types of the neighbourhood together.

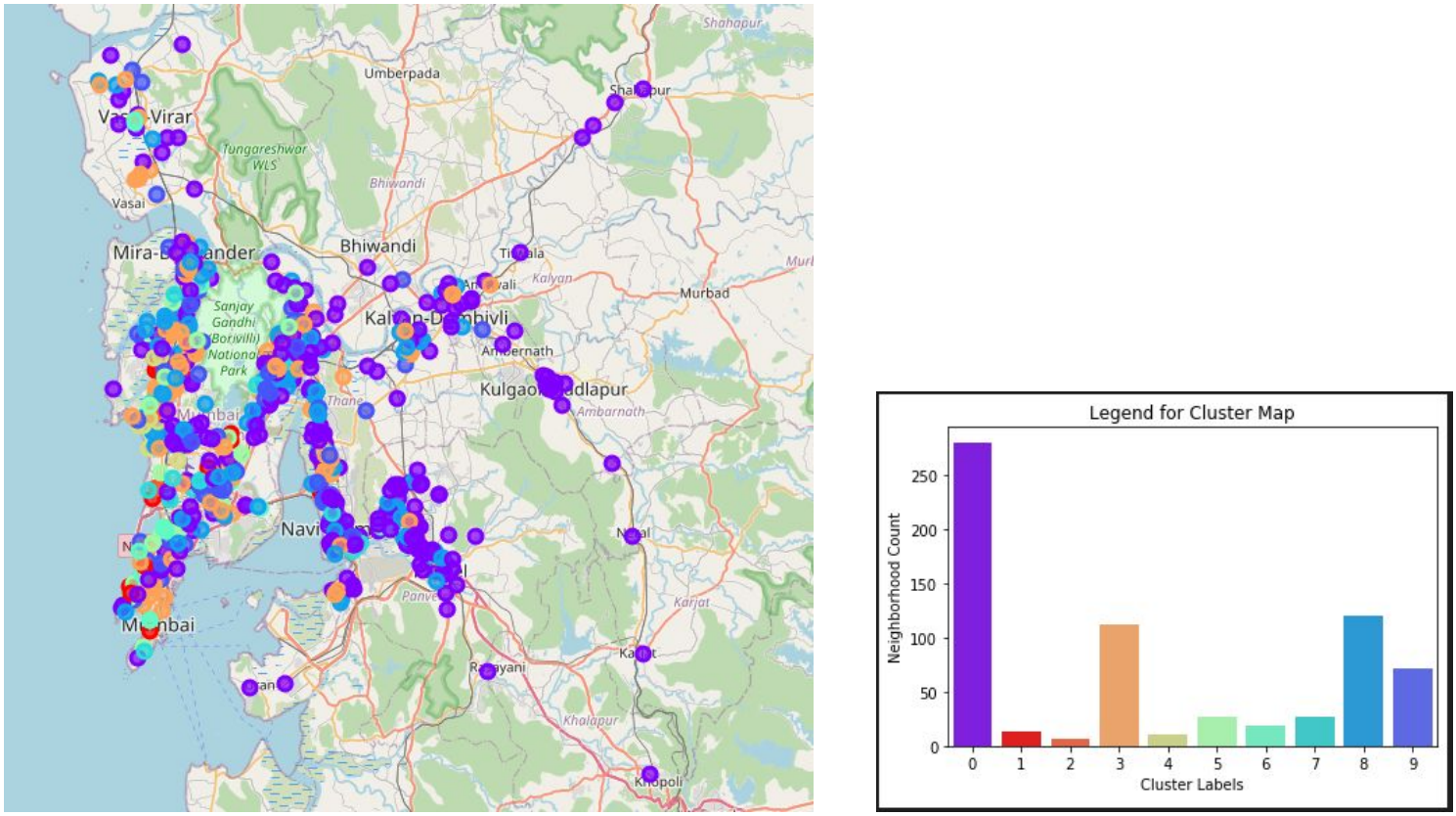

Fig 12.Clustering Map and Legend

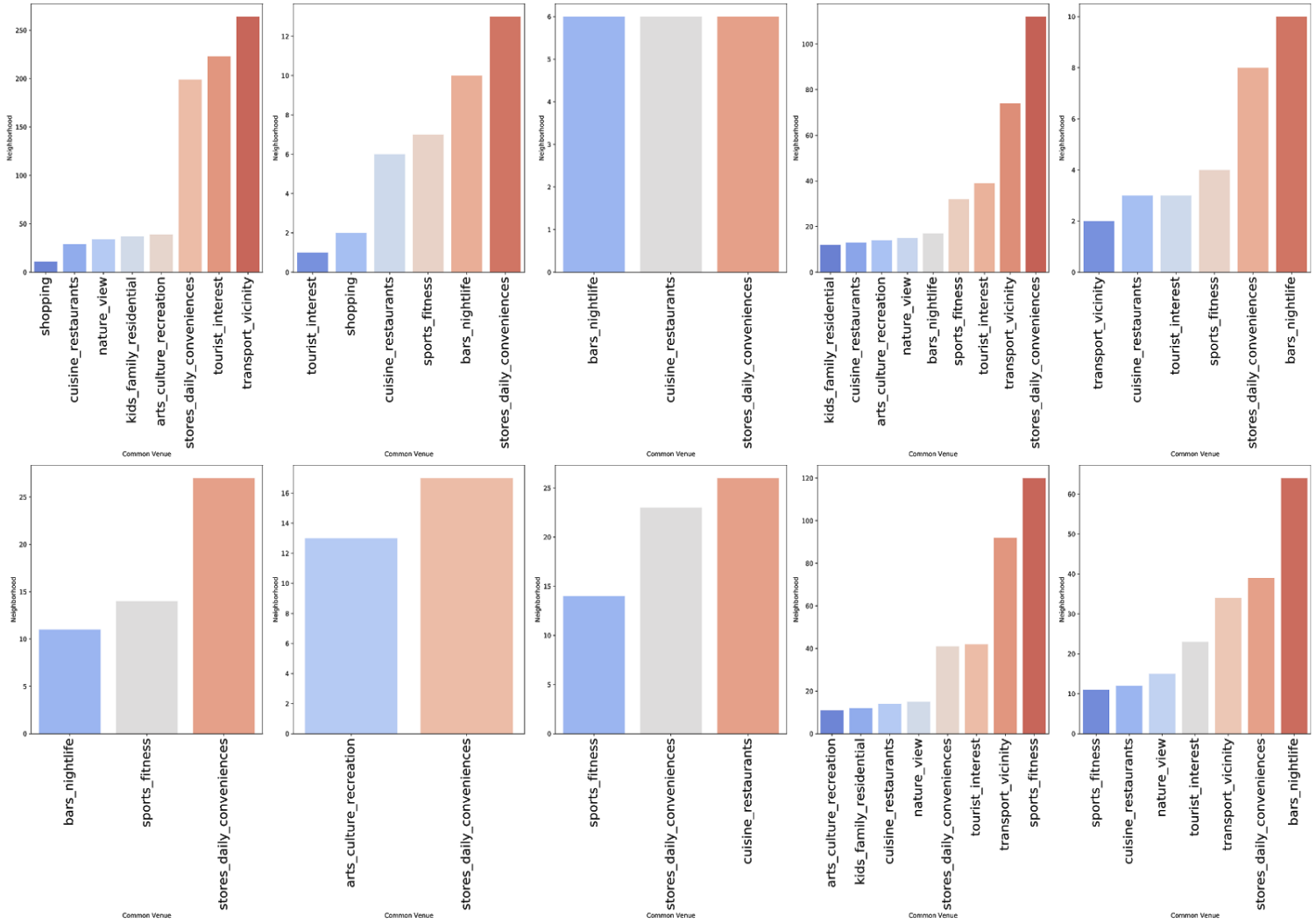

Fig 13. Venue-Clusters

```
Cluster_0 = Residential area Friendly neighborhood
Cluster_1 = Area for explorers
Cluster_2 = Commercial Area
Cluster_3 = Posh Area
Cluster_4 = Area to hangout
Cluster_5 = Area for bachelors(young generation)
Cluster_6 = Area for arts and culture exporers
Cluster_7 = Area for sports and fitness enthusiats
Cluster_8 = Area for High Energetic Residents
Cluster_9 = Area for high profile Residents
```

10. Applying Machine Learning model to predict Housing Price based on the count of venues in the neighbourhood of a house. The R2 value for machine learning models used is very small which implies that the model may not be suitable for the data.

Results

1. We were able to analyze prices of houses as well as understand the distribution of the average price and were able to abstract the following results:-
    a. South Mumbai has the lowest density of houses yet most expensive houses. Whereas as we go away from the mainland of Mumbai Prices go down.
    b. Navi Mumbai has the highest number of housing data followed by Andheri-Dahisar, whereas South Mumbai and Harbour has the lowest housing data. This clearly matches with their existence as the region of Navi Mumbai has been planned for housing only which are cheap, which again supports the major cause of developing the project of building houses for mid-level professional, technical, or managerial staff working in the main city.
2. We could Cluster Similar Neighborhood and could decide what type of neighbourhood it belonged and what type of people would prefer.
3. Even though the scores seem to be not that promising, but may improve on using the more sophisticated method. The reason for its poor prediction could be
    a. The real estate price is hard to predict with the available features and data.
    b. The data is incomplete for example the type of venues that have a major influence on housing prices was not captured.
    c. The machine learning techniques were very basic.

Discussions

1. It can be seen that housing prices in some areas are exorbitantly high.
2. The housing prices as we ascended towards downtown as it mostly a commercial area with fewer houses.
3. One of the most challenging tasks was constructing the data. Because housing data was not available easily. When combining data from multiple sources, inconsistency can happen. As from the notebook it can be seen lots of efforts are required to check, research and change the data before using in the analysis.
4. A better Machine Learning technique is required to successfully predict the prices of houses.

Conclusion

We can get meaningful and logical insights from the result. Doing this project helps to practice every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into. This report also explains that further development can be made to improve the analysis like using different Clustering Method to cluster the neighbourhood or using more advanced Machine Learning technique to predict housing prices. Moreover extra analysis can be done like we can analyse rental prices the same way as rental prices were available in the raw dataset. Towards the person that went through this project, many thanks for the time and patient.

Keep Learning and Keep Growing.