Ideal Neighborhood for Renting a House in Mumbai, India

Introduction

Imagine that you are moving to a new city- could be for higher education, for a new job, or to be closer to family and friends. One of the first things you need to figure out is where you will live. You certainly want to rent a house within your budget, but you also want to live in a neighborhood which has venues which you frequently visit- coffee shops, pizza places, gyms, multiplexes, bakeries, etc. Wouldn't it be great to get recommendations on neighborhoods within your budget having all the venues you desire?

Business Problem

The objective of this capstone project is to use the Foursquare API, analyse different neighborhoods in Mumbai, India, and recommend the best neighborhoods to rent a house in based on inputs of budget and venues. Using the data science methodology, this project will answer the following question: If a user is looking to move to Mumbai, what neighborhood would you recommend them to rent a house in?

Data Required

A prospective renter will look for neighborhoods to rent a house based on average property prices and the kind of venues they would want access to. To make a recommendation to a prospective renter, the following data will be useful:

1. Neighborhoods in Mumbai

The data is scraped from this Wikipedia page. The data contains the neighborhoods, the district they fall in, and their geographical coordinates

2. Property Prices

Average neighborhood property prices would help renters shortlist neighborhoods based on their budget. Renters would specifically like to know prices for the house type they are planning on renting- 1 BHK, 2 BHK, and 3 BHK. The data is scraped from 99acres.com

3. Venue Data

Renters would like to know what kind of venues are available in the neighborhoods they have shortlisted and would additionally like to know which of their shortlisted neighborhoods have the venues they frequently visit. The Foursquare API is used to query venue data. Foursquare categorizes venues into restaurants, gyms, parks, multiplexes, dessert shops, etc. Venues in a 1000 metres radius are considered. This radius is chosen because 1000 metres is a reasonable walking distance.

Cleaning and Preparing the Data

1. Mumbai Neighborhoods

Neighborhoods with outlying latitude and longitude values are dropped because Foursquare API relies on geographical location to return accurate venues data.

2. Property Prices

Getting the data

The 99acres.com website is scraped to obtain price data for 1 BHK, 2 BHK, and 3 BHK houses in different neighborhoods. The data obtained is in the form of a price range and neighborhood price is taken as an average of the lower bound and upper bound of this price range.

The following is the first 5 rows of the obtained data set:

	Location	Latitude	Longitude	1 BHK	2 BHK	3 BHK
Neighborhood						
Amboli	Andheri	19.129300	72.843400	NaN	NaN	NaN
Vile Parle	Andheri	19.100000	72.830000	33558.0	51297.5	NaN
Jogeshwari West	Andheri	19.120000	72.850000	23970.0	37952.5	49015.0
Versova	Andheri	19.120000	72.820000	33797.5	51769.5	70987.0
Seven Bungalows	Andheri	19.129052	72.817018	34756.5	49926.5	65747.5

Missing Prices

Missing prices need to be replaced by either some multiple of the other property types in the neighborhood or by the average of the location price. A 2-step approach is employed for this:

First, we consider those neighborhoods which have some property data available (e.g. Vile Parle has 1 BHK and 2 BHK data available while 3 BHK data is missing). We calculate 2 ratios:

$$2\ BHK\ ratio = \frac{\Sigma\left(\frac{1\ BHK\ prices}{2\ BHK\ prices}\right)}{Number\ of\ ratios\ taken}$$

$$3\ BHK\ ratio = \frac{\Sigma\left(\frac{1\ BHK\ prices}{3\ BHK\ prices}\right)}{Number\ of\ ratios\ taken}$$

Missing 1 BHK prices are replaced by the product of the corresponding 2 BHK price and the 2 BHK ratio. If the 2 BHK price is missing, the 1 BHK price is replaced by the product of the corresponding 3 BHK price and the 3 BHK ratio. We have now obtained the 1 BHK prices for all these neighborhoods. Next, we update the 2 ratios using the updated data set. Missing 2 BHK values are obtained by dividing the corresponding 1 BHK price with the 2 BHK ratio. Missing 3 BHK values are obtained by dividing the corresponding 1 BHK price with the 3 BHK ratio. We have now replaced missing prices for all partially filled rows.

Second, we consider neighborhoods which have no property data available (e.g. Amboli has no data for 1 BHK, 2 BHK, or 3 BHK prices). We find the average property prices for each district. Missing prices in the average property prices for districts data set are obtained in the same way as above. We now replace the remaining missing values with the district average (e.g. Amboli's property prices will be replaced by the average district prices for Andheri).

The following is the properties data after cleaning:

	Neighborhood	Location	Latitude	Longitude	1 BHK	2 BHK	3 BHK
0	Amboli	Andheri	19.129300	72.843400	30786.65	49194.215259	74945.481401
1	Vile Parle	Andheri	19.100000	72.830000	33558.00	51297.500000	89363.675711
2	Jogeshwari West	Andheri	19.120000	72.850000	23970.00	37952.500000	49015.000000
3	Versova	Andheri	19.120000	72.820000	33797.50	51769.500000	70987.000000
4	Seven Bungalows	Andheri	19.129052	72.817018	34756.50	49926.500000	65747.500000

3. Venue Data

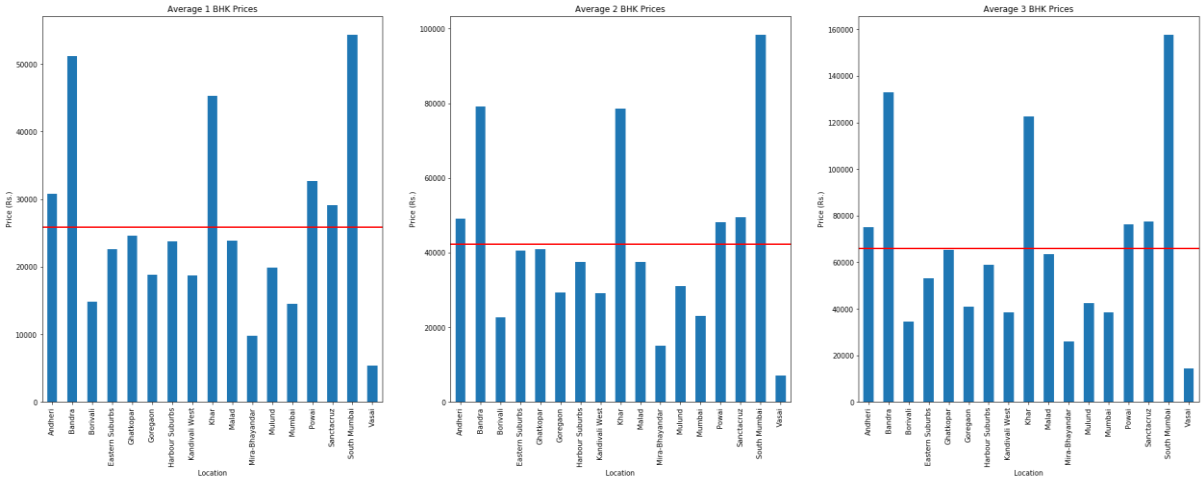
The Foursquare API is queried to get venue data. We first make a call to get the top 100 venues in a radius of 2 kilometres. All the neighborhoods which return less than 20 venues in a 2 kilometres radius are dropped. Another call is made to Foursquare to get the top 100 venues in a 1000 metres radius. We take a 1000 metres radius because this is a reasonable walking distance. Next, the venues data is cleaned up by combining certain venue categories (e.g. Burrito place and Tex-Mex place are merged with Mexican restaurant). Next, venue categories that appear less than 16 times in the entire city are removed. Finally, one-hot encoding is done for the data set: for each neighborhood, the value of a corresponding venue category is taken as 1 if the venue category is present in the neighborhood, and 0 otherwise.

Exploratory Data Analysis

In order to better understand the data that we will be working with, let's do an exploratory data analysis.

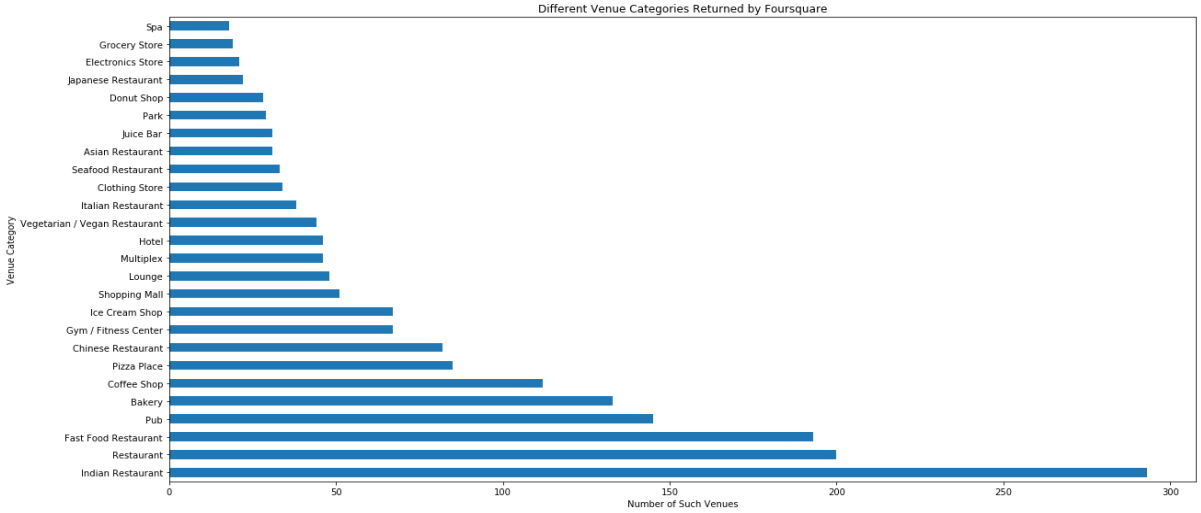
Property Prices in Mumbai

As seen in the graphs, rental prices are highest for South Mumbai and lowest for Vasai. A 3-bedroom hall kitchen apartment in South Mumbai will cost around Rs. 1,60,000/ month on average while the same property can be rented in Vasai for less than Rs. 20,000/ month. The red horizontal line indicates the average rental price for each type of house.



Venues as Returned by the Foursquare API

As seen in the graph, most of the venues returned by Foursquare fall into the category of restaurants- these have been divided into their different cuisines and specialties for a more refined search. A user can also choose from dessert shops, gyms, parks, multiplexes, spas, and shopping malls.



Using the Model

A user can enter the property type they are looking for (1 BHK, 2 BHK, 3 BHK). The model will provide the price range of that property type and ask the user to enter their budget (which should be within the price range). The model will take this budget and return all neighborhoods who's property price falls in the range of 75%-125% of the user's budget. The user is then prompted to enter a list of venues they would prefer to have in their neighborhood based on a list of possible venue choices. The model will then return neighborhoods (if any) that fit the user's specifications.

Limitations of the Model

Firstly, Foursquare data isn't all encompassing. Most of the data falls into the categories of food. Thus, the model fails to provide valuable information such as proximity to schools, public transport, pharmacies, and other amenities which are crucial to consider when looking for a house. Furthermore, a number of venues in Mumbai are unavailable on the Foursquare database and are therefore excluded from the analysis, preventing us from getting a completely accurate picture. Lastly, some neighborhoods had to be dropped because Foursquare just didn't have enough venues data for these neighborhoods. As a result, the sample set is shrunk and does not cover the entire city of Mumbai.

Conclusion

Although the model fails to cover the entire city of Mumbai, it provides valuable information regarding prices for most neighborhoods in Greater Mumbai (Mumbai Suburban District and Mumbai City District) which can help point a user in the right direction so that they can begin their apartment search. Furthermore, it can help identify neighborhoods having a lot of restaurants, pubs, dessert shops, multiplexes, and other such venues which would appeal to someone looking for a lively neighborhood.

Data on schools, public transport, and healthcare in addition to Foursquare data would allow the model to make a stronger recommendation and would also enable it to cater to a wider category of users.

References:

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

https://medium.com/r/?url=https%3A%2F%2Fwww.99acres.com%2Fproperty-rates-and-price-trends-in-mumbai