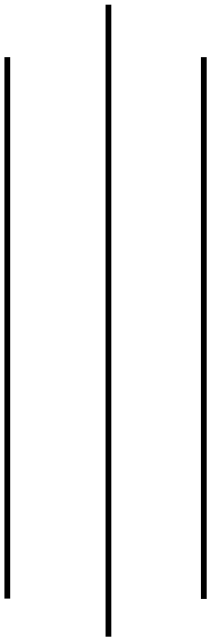
Code Rush

Chakupat Road - 10, Lalitpur



Data Analysis on Houses Rent

Submitted By
Bibek Shyama

Submitted To
Code Rush

Thursday, September 29, 2022

Table of Content

1. Introduction

With the growth of population, people are always looking for a better shelter with good facilities at affordable price. For this purpose, we took house rent dataset of a year (2022) to analyze and predict the rent in the near future.

Objectives:

- Understanding the dataset.
- Analyze/visualize houses with different facilities and their relationship.
- Develop a predictive model for Rent.

2. Dataset

- The size of the dataset is 445.1KB.
- The dataset consist data of 4736 instances with 12 features.
- The features are as follows:

Features	Type	Description
1. Posted On	Object	Full date of the house posted on.
2. BHK	Numeric	Number of Bedrooms, Hall, Kitchen.
3. Rent	Numeric	Rent of the house.
4. Size	Numeric	Size of the house in square feet.
5. Floor	Object	House/Apartments/Flats situated in which floor.
6. Area Type	Object	Size of the house calculated on: i. Super Area ii. Carpet Area iii. Build Area
7. Area Locality	Object	Locality of the house.
8. City	Object	City where the house is located. i. Mumbai ii. Kolkata iii. Delhi iv. Chennai v. Bangalore vi. Hyderabad
9. Furnishing Status	Object	Furnishing status of the house. i. Furnished ii. Semi-Furnished iii. Unfurnished
10. Tenant Preferred	Object	Type of tenant preferred by the owner or agent. i. Bachelor/Family ii. Family iii. Bachelor
11. Bathroom	Numeric	Number of bathrooms in the house.
12. Point of Contact	Object	Whom should the tenant contact for more information regarding the house. i. Contact Owner ii. Contact Agent iii. Contact Builder

3. Methodology

For data analysis, python is one of the famous programming languages. There are plenty python libraries to extract facts and meaningful information from the authentic dataset. Jupyter notebook is a web-based interactive computing platform famous for python programming.
Libraries such as Pandas, NumPy, SciPy, Scikit-learn, etc. are used for analysis or computational purpose while for visualization seaborn, matplotlib, plotly, etc. are used. Inference of the dataset can be drawn using these various tools and technology.

4. Data Loading and Planning

	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
0	2022-05-18	2	10000	1100	Ground out of 2	Super Area	Bandel	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner
1	2022-05-13	2	20000	800	1 out of 3	Super Area	Phool Bagan, Kankurgachi	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
2	2022-05-16	2	17000	1000	1 out of 3	Super Area	Salt Lake City Sector 2	Kolkata	Semi-Furnished	Bachelors/Family	1	Contact Owner
3	2022-07-04	2	10000	800	1 out of 2	Super Area	Dumdum Park	Kolkata	Unfurnished	Bachelors/Family	1	Contact Owner
4	2022-05-09	2	7500	850	1 out of 2	Carpet Area	South Dum Dum	Kolkata	Unfurnished	Bachelors	1	Contact Owner
...	...	...	...	...	...	...	...	...	...	...	...	...
4741	2022-05-18	2	15000	1000	3 out of 5	Carpet Area	Bandam Kommu	Hyderabad	Semi-Furnished	Bachelors/Family	2	Contact Owner
4742	2022-05-15	3	29000	2000	1 out of 4	Super Area	Manikonda, Hyderabad	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Owner
4743	2022-07-10	3	35000	1750	3 out of 5	Carpet Area	Himayath Nagar, NH 7	Hyderabad	Semi-Furnished	Bachelors/Family	3	Contact Agent
4744	2022-07-06	3	45000	1500	23 out of 34	Carpet Area	Gachibowli	Hyderabad	Semi-Furnished	Family	2	Contact Agent
4745	2022-05-04	2	15000	1000	4 out of 5	Carpet Area	Suchitra Circle	Hyderabad	Unfurnished	Bachelors	2	Contact Owner

4746 rows × 12 columns

Figure 1 : Dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4746 entries, 0 to 4745
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype         
---  ------            --------------  -----         
 0   Posted On         4746 non-null   datetime64[ns]
 1   BHK               4746 non-null   int64         
 2   Rent              4746 non-null   int64         
 3   Size              4746 non-null   int64         
 4   Floor             4746 non-null   object        
 5   Area Type         4746 non-null   object        
 6   Area Locality     4746 non-null   object        
 7   City              4746 non-null   object        
 8   Furnishing Status 4746 non-null   object        
 9   Tenant Preferred  4746 non-null   object        
 10  Bathroom          4746 non-null   int64         
 11  Point of Contact  4746 non-null   object        
 12  Month             4746 non-null   int64         
dtypes: datetime64[ns](1), int64(5), object(7)
memory usage: 482.1+ KB
```

```
1  data.duplicated().sum()
```

```
0
```

Figure 2 : Dataset Information.

	BHK	Rent	Size	Bathroom	Month
count	4746.000	4746.000	4746.000	4746.000	4746.000
mean	2.084	34993.451	967.491	1.966	5.756
std	0.832	78106.413	634.202	0.885	0.832
min	1.000	1200.000	10.000	1.000	4.000
25%	2.000	10000.000	550.000	1.000	5.000
50%	2.000	16000.000	850.000	2.000	6.000
75%	3.000	33000.000	1200.000	2.000	6.000
max	6.000	3500000.000	8000.000	10.000	7.000

Figure 3 : Dataset Summary Descriptive Analysis.

Inference:

- The given dataset has 4746 instances with 12 attributes.
- Here ‘Posted on’ is converted from object to date data type. Also, it has value of 2022 year only, so splitting the value to get/add new column called ‘Month’ as integer data type.
- No null and duplicate value are present in the dataset.
- From figure 3, we can observe that the mean of ‘Rent’ is more than twice of median. Hence there are outliers present in ‘Rent’.

4.1 Univariate Analysis

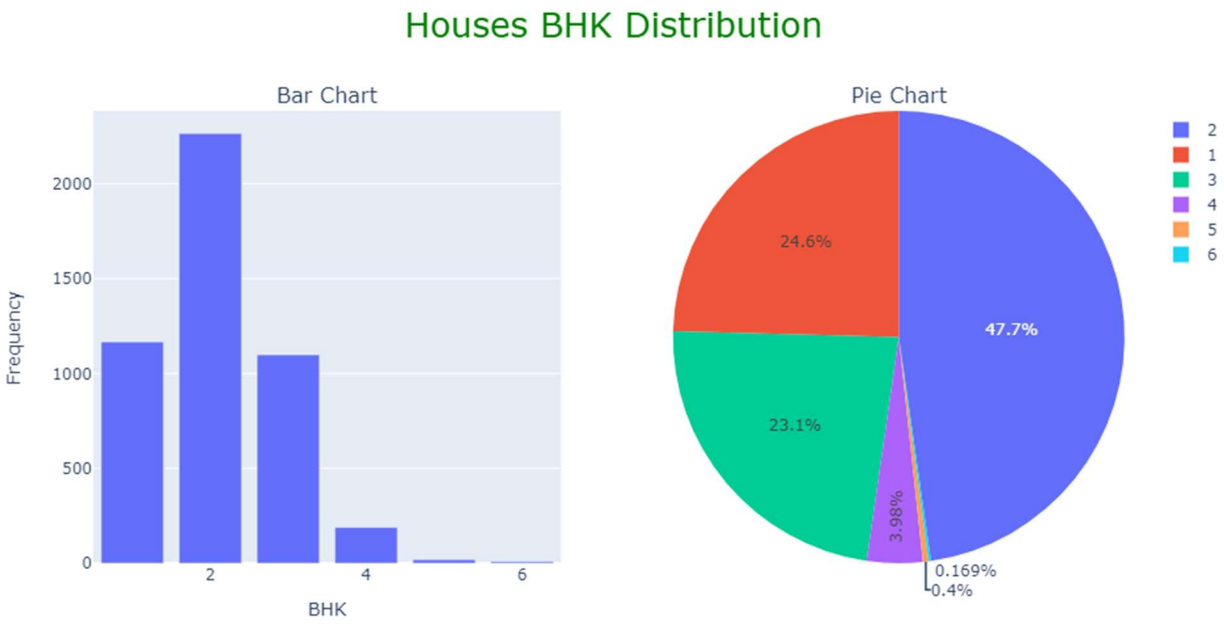


Figure 4 : Visualization of BHK.

Inference:

- Most tenant preferred 2BHK houses, as it is affordable and sufficient for the most family/individuals.
- Above 3 BHK, houses are hardly rented by the people as they are expensive and have very big space.

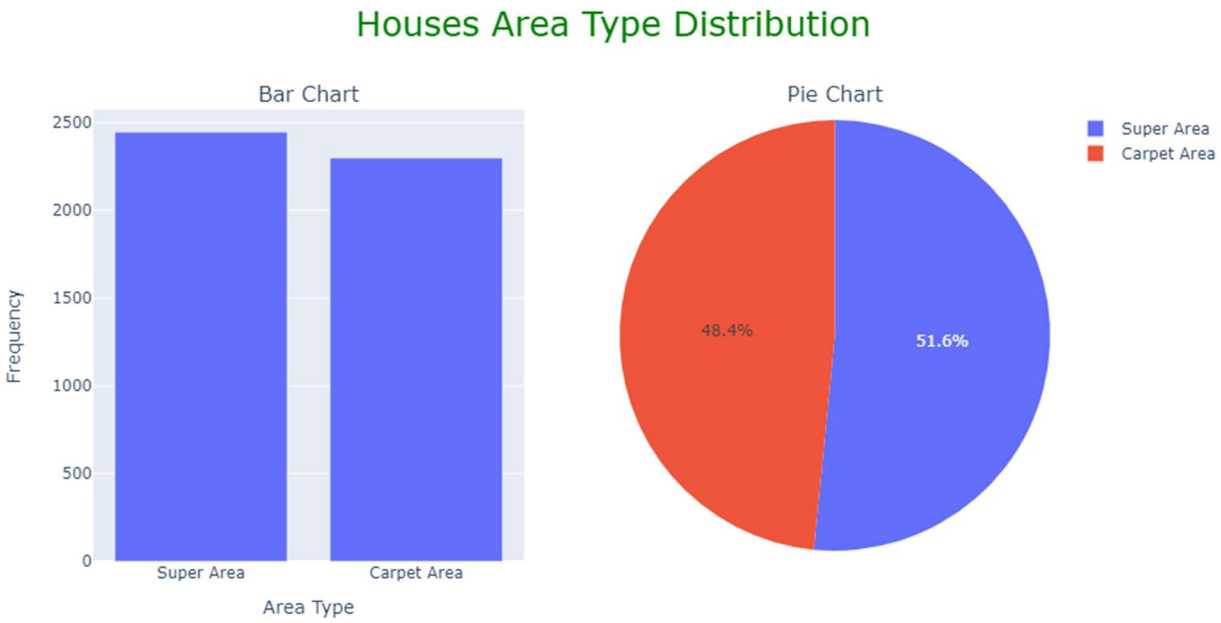


Figure 5 : Visualization of Area Type.

Inference:

- We eliminate the 'Built Area' option as it has only 2 values.
- Houses built on super area are slightly more than that of carpet area.

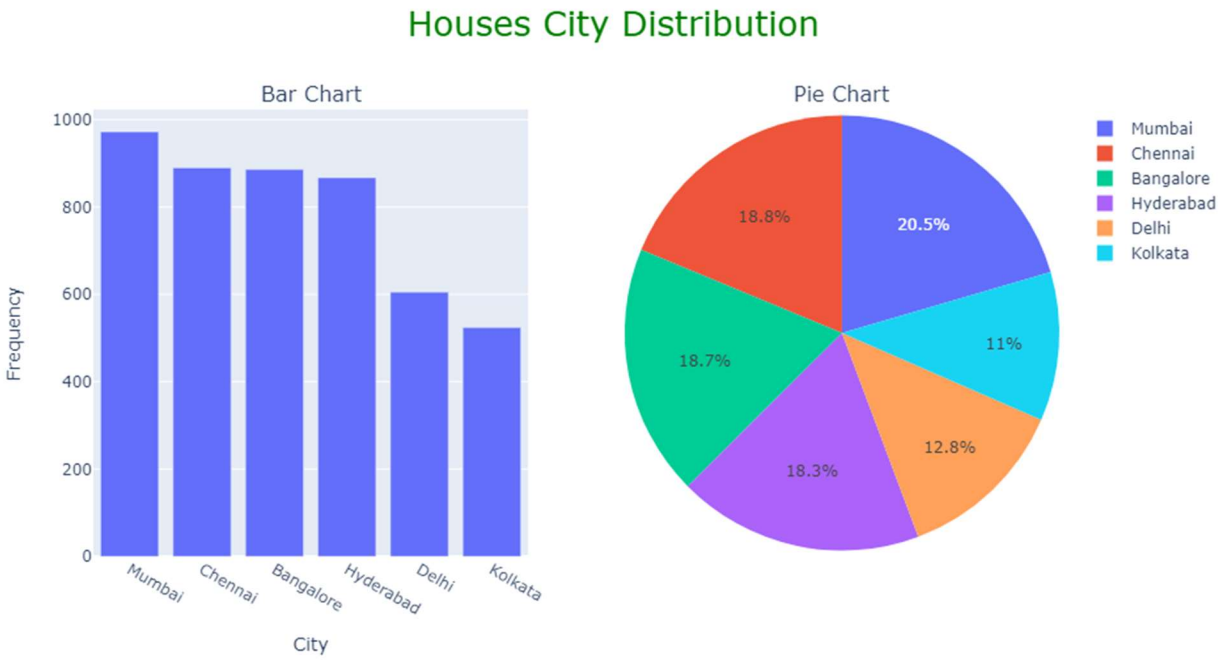


Figure 6: Visualization of City.

Inference:

- Mumbai has the most houses for rent followed by Chennai, Bangalore, Hyderabad and Kolkata having the least number of houses for rent compared to others.

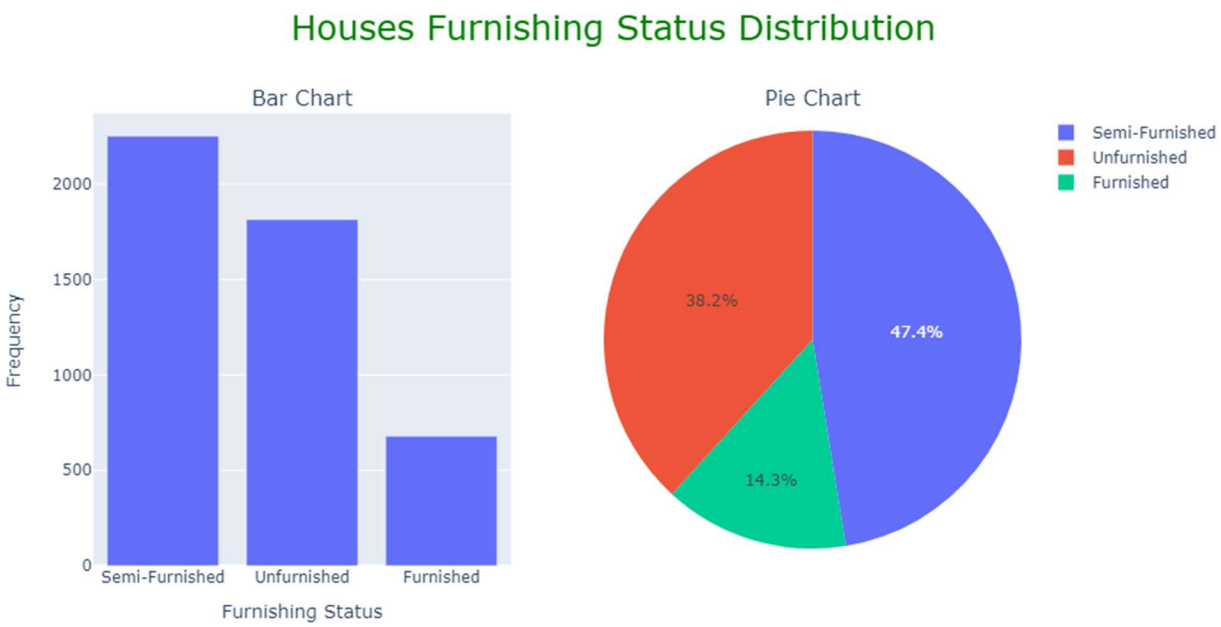


Figure 7: Visualization of Furnishing Status.

Inference:

- It can be interpreted that people usually rent out unfurnished or semi furnished house more. This could be due to negligence of rented personal to not take care of the owners' items. Hence number of fully furnished homes listed are much less as compared to other categories.
  OR
  It can be also interpreted as people usually rent either semi-furnished or unfurnished because they are quite affordable and sufficient for the person.

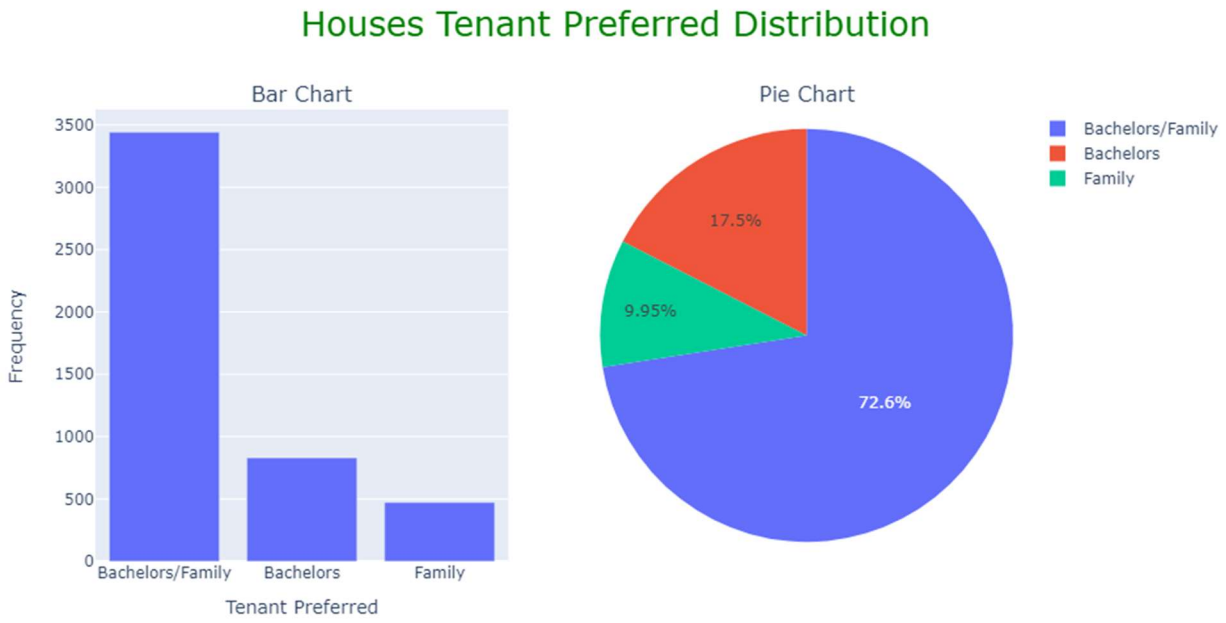


Figure 8: Visualization of Tenant Preferred.

Inference:

- It seems most rental house owner/agent have no special category. They are okay with both Bachelors and Family.

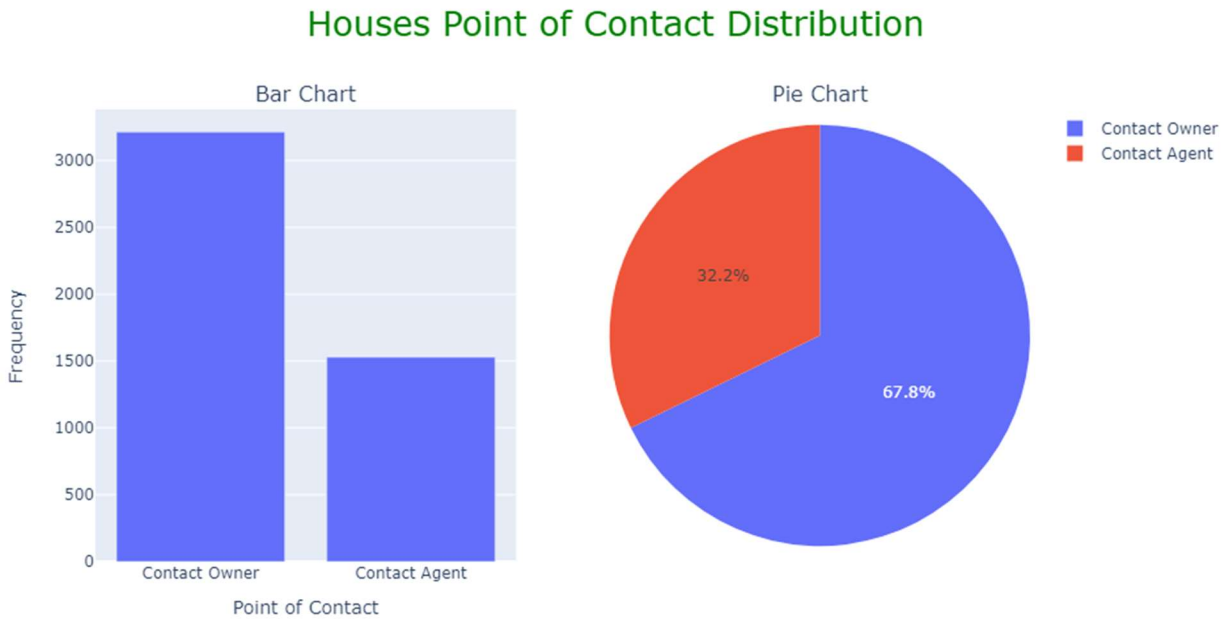


Figure 9: Visualization of Point of Contact.

Inference:

- We have eliminated ‘Contact Builder’ option as it has only 1 value.

- For rental, most people prefer to contact owner rather than agent. This could be because direct contact with the landlord is safe and comparatively cheaper as there is no middle-man commission.

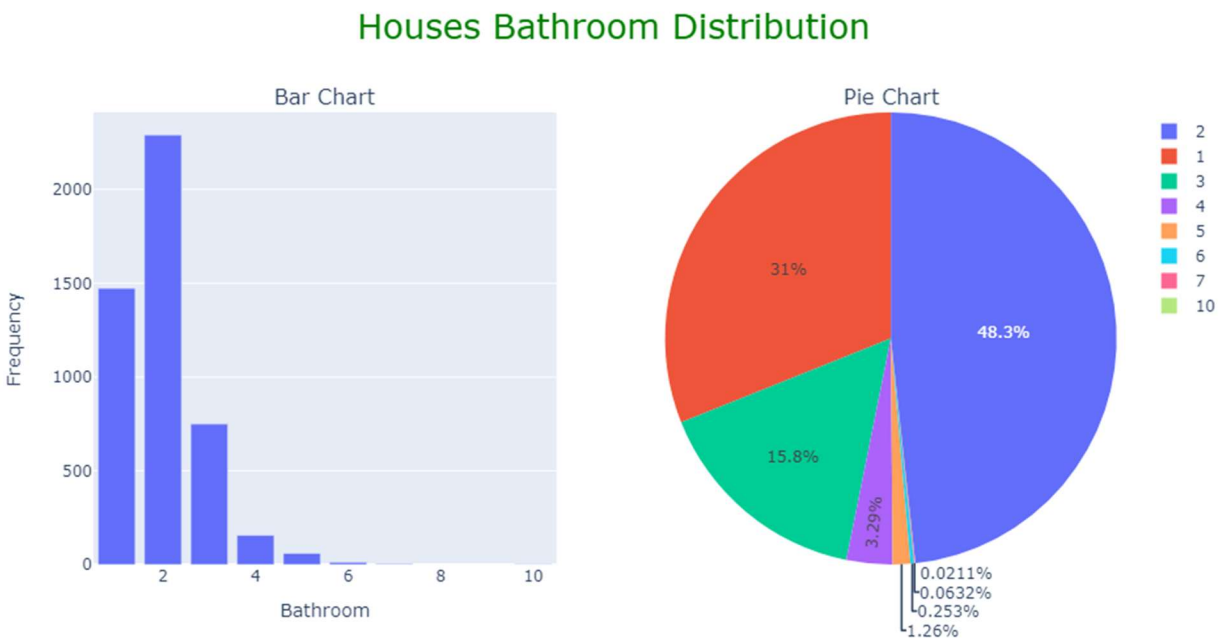


Figure 10: Visualization of Bathroom.

Inference:

- 2-bathroom houses are mostly put up for rental list followed by houses with 1 and 3 bathrooms.
- Houses with more than 3 bathrooms are not put up for rental as it doesn't seem appropriate and much of use.

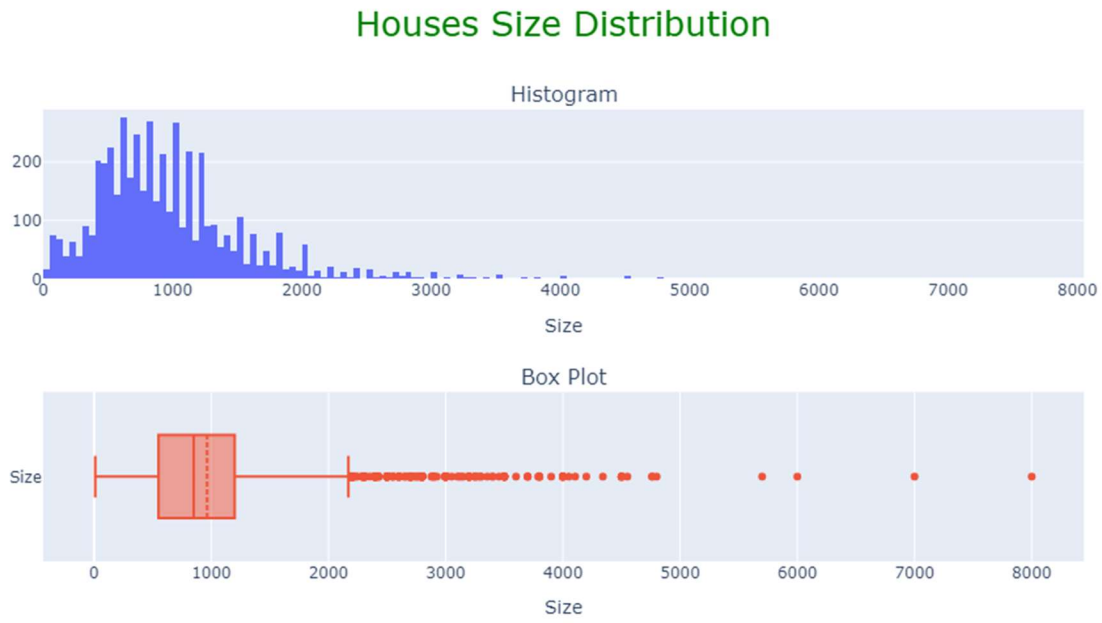


Figure 11: Visualization of Size.

Inference:

- There are large number of outliers present in the size feature. This may be because most houses are normal sized in the dataset. Hence many outliers can be seen in the plot.
- The smallest house rented has size of about 10 sq. feet.
- The biggest house rented has size of about 8000 sq. feet.
- The median of house rented has size of about 850 sq. feet.

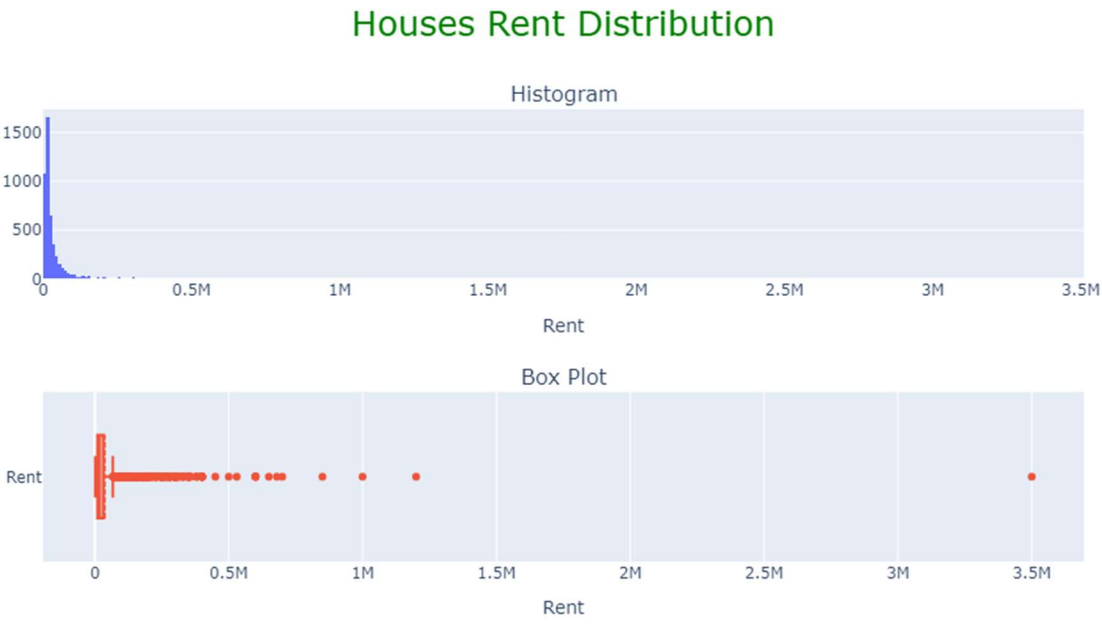


Figure 12: Visualization of Rent.

Inference:

- There are large number of outliers present as can be seen from the box plot. This could be because dataset contain mostly lower price houses.
- Here we can see a house which rent is greater than 0.8M considering most of the houses have rent below 0.5M. So, these extreme outliers are removed.

4.2 Bivariate Analysis

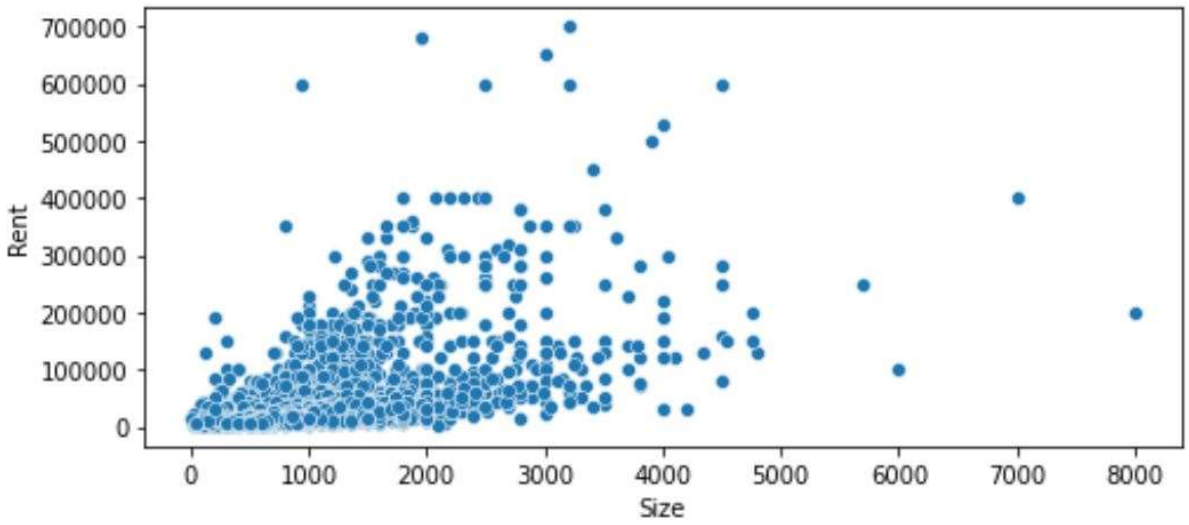


Figure 13: Size vs Rent.

Inference:

- As the size of the house increases, rent also increases somewhat meaning low positive correlation.

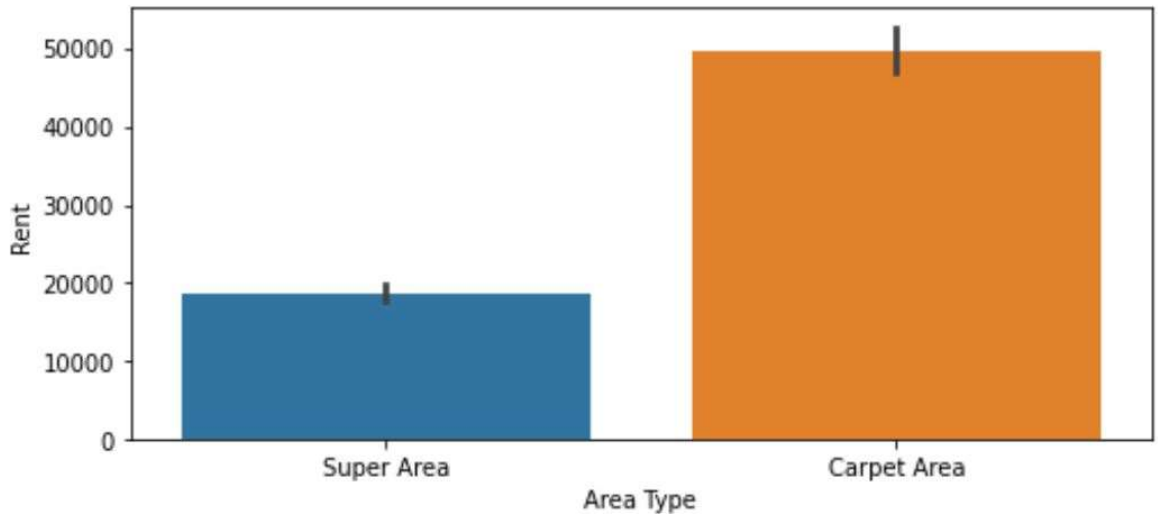


Figure 14 : Area Type vs Rent.

Inference:

- House built on carpet area have highest average rent.

-

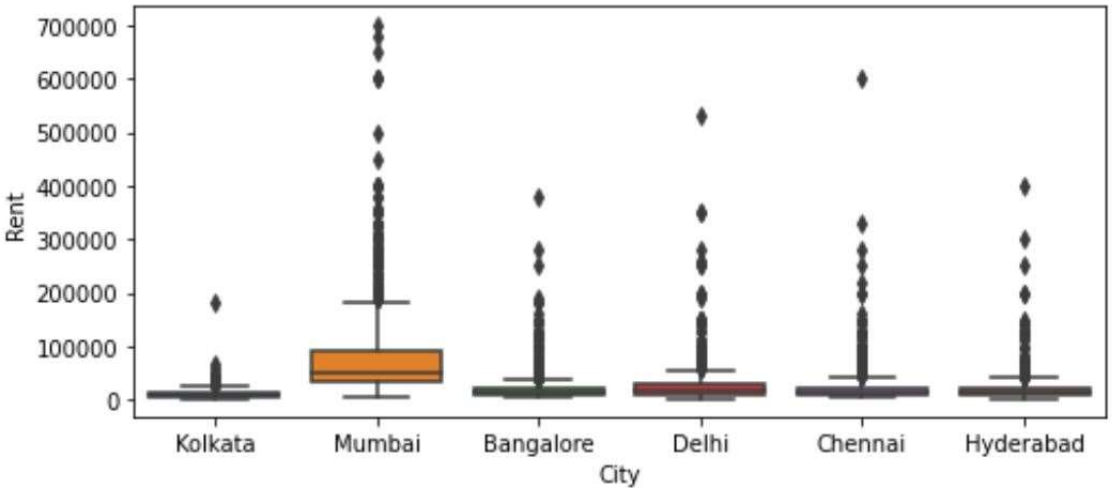


Figure 15: City vs Rent.

Inference:

- The average rent of Mumbai house is the highest than all of the cities and Kolkata has the cheapest rent.

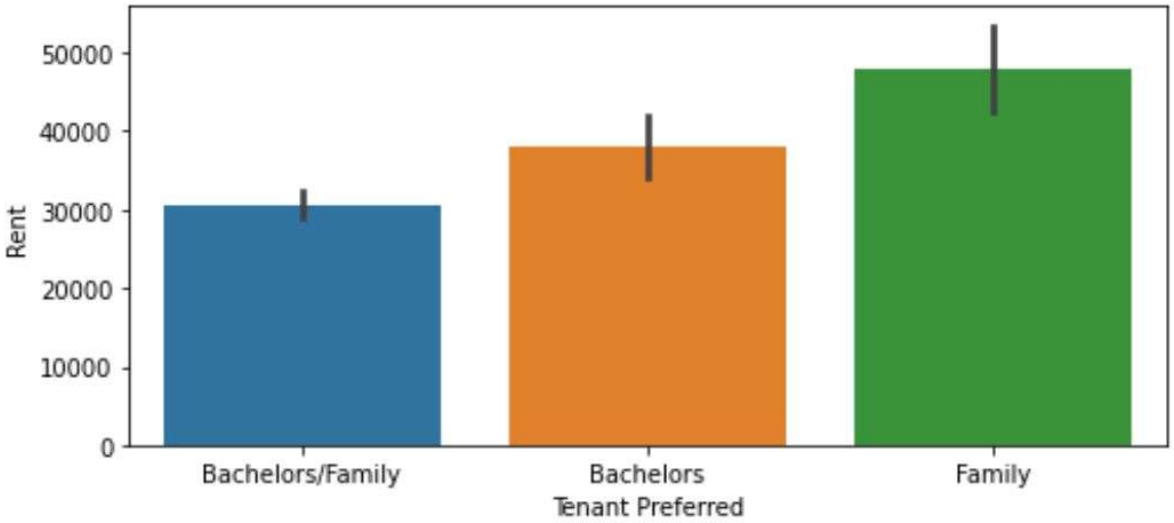


Figure 16: Tenant Preferred vs Rent.

Inference:

- Houses occupied by solely family has the highest average rent.

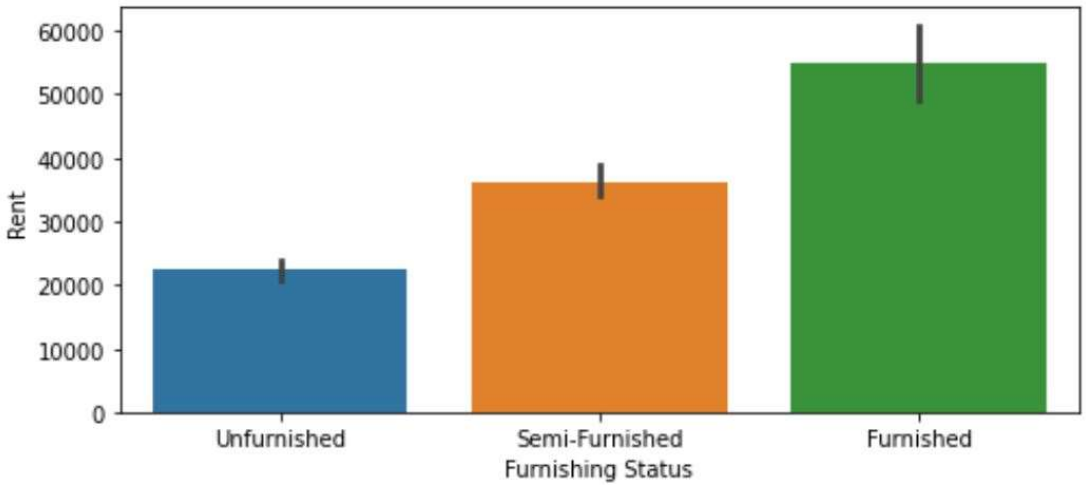


Figure 17: Furnishing Status vs Rent.

Inference:

- Fully furnished houses have highest average rent while unfurnished houses have the lowest. This is because fully furnished house cost more than semi or unfurnished.

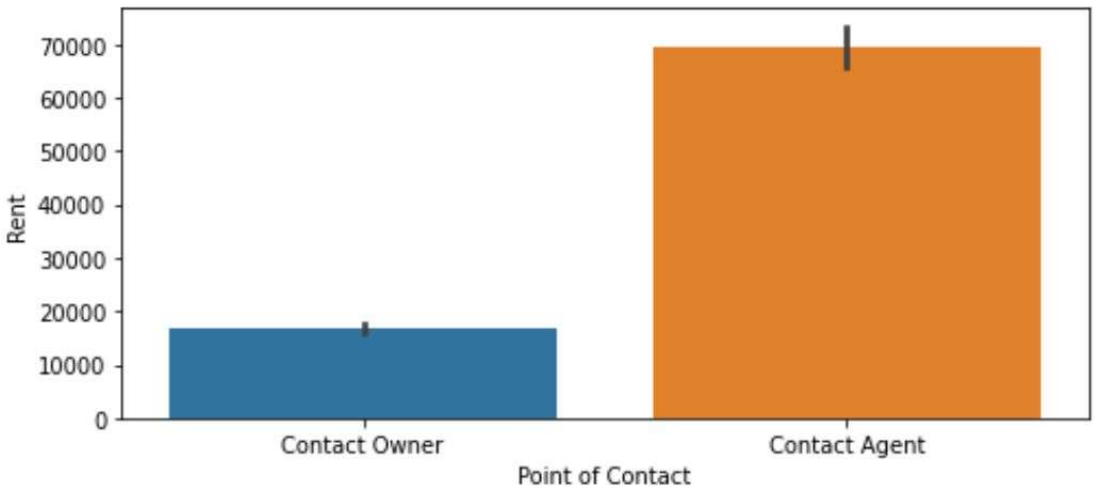


Figure 18: Point of Contact vs Rent.

Inference:

- Houses where point of contact is agent have the highest rent compared owner. This is because of agent-commission.

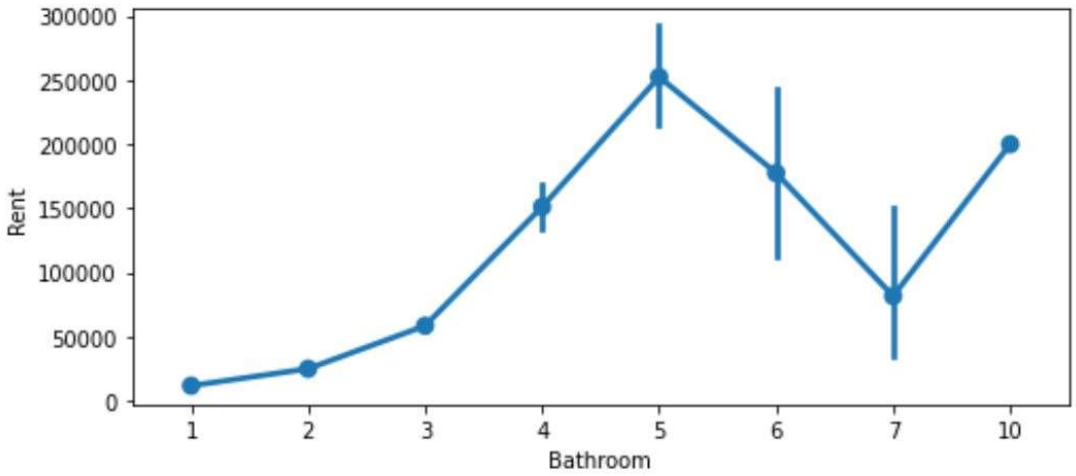


Figure 19: Bathroom vs Rent.

Inference:

- Houses with bathroom from 1 to 5 has increasing average rental price trend whereas houses with more than 5 bathrooms have low rental price. This could be tenant prefers houses with bathroom less than or equal to 5 due to its limited usage.

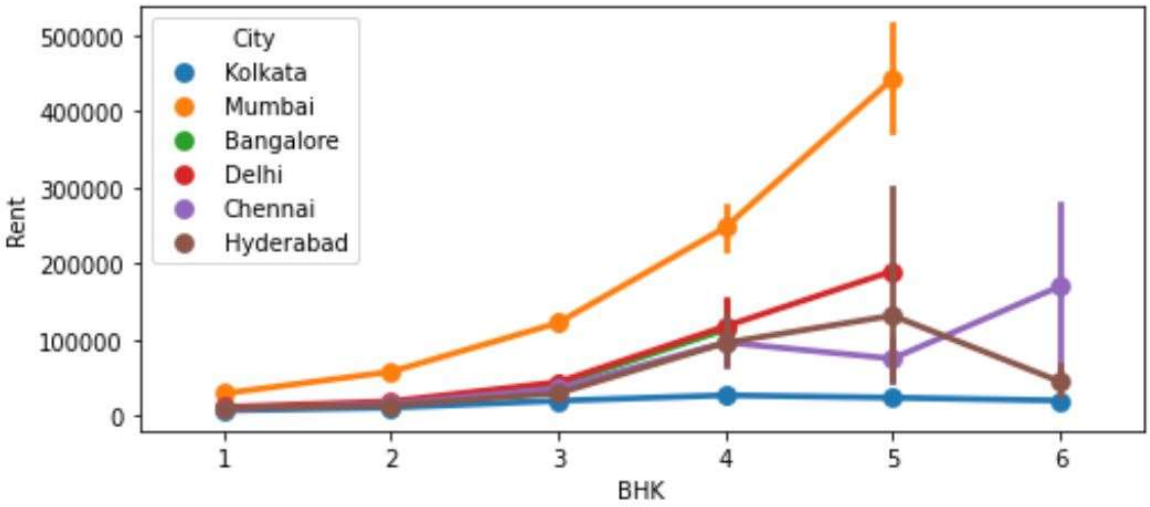


Figure 20: BHK of different city vs Rent.

Inference:

- The average rent of the 5BHK is the highest compared to all in all cities.

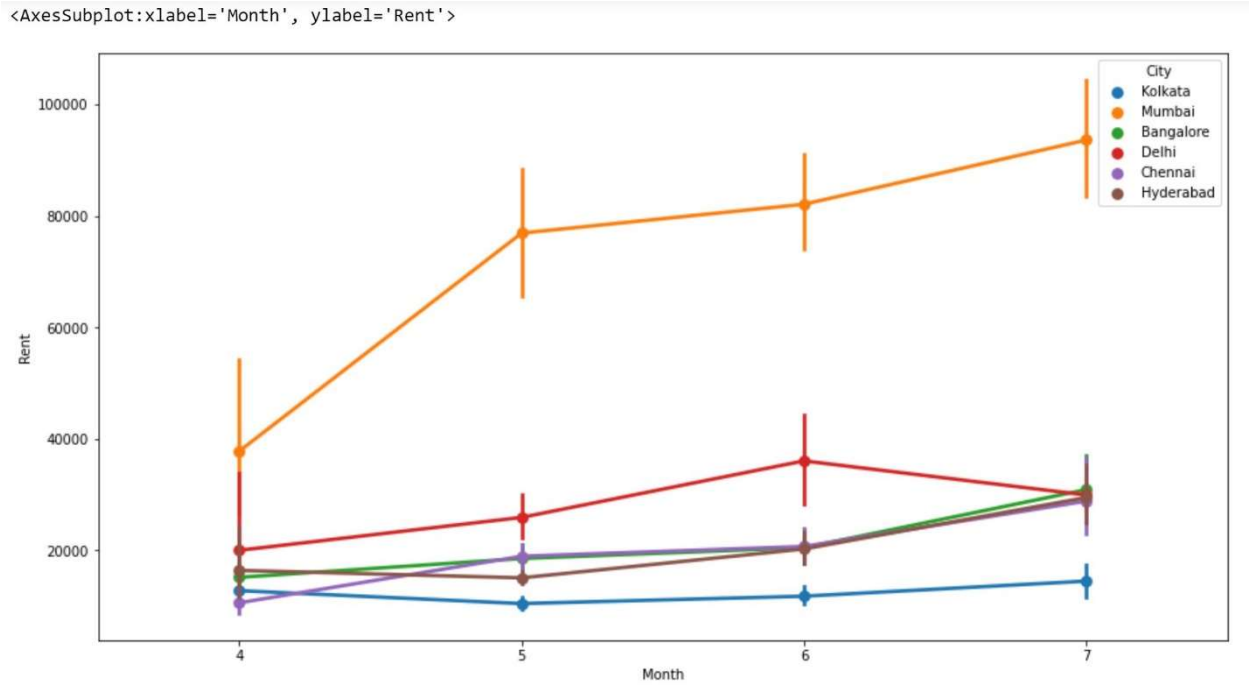


Figure 21: Monthly Rent in different Cities.

Inference:

- There is overall increasing trend observed in rent prices over months in every city.
- Mumbai rental price is increasing rapidly compared to other cities.

4.3 Correlation

Figure 22: Correlation Coefficient of different attributes.

Observation:

- Rent has moderate correlation with bathroom (0.59), size (0.51), BHK (0.49 ~0.5) and point of contact (0.45).
- Rent and Area Type have low correlation with 0.28.
- Rent and City have negative low correlation with 0.32.

5. Data Pre-Processing

For data pre-processing, we will be using mostly sklearn.

- First, we have drop unnecessary columns. They are ‘Posted On’, ‘Floor’, ‘Area Locality’ and ‘Month’. We will not be using these features for our analysis.
- After eliminating, we are left with 4735 instances and 9 features which is not quite sufficient but still we will work on this dataset.
- Since we need to predict rent. So, extracting features into ‘x’ variable without the rent feature and extracting rent feature into ‘y’ variable.
- We will use 70% of the data for training and 30% for testing.

6. Model Evaluation

We use linear regression, decision tree and random forest as our model.

6.1 Linear Regression

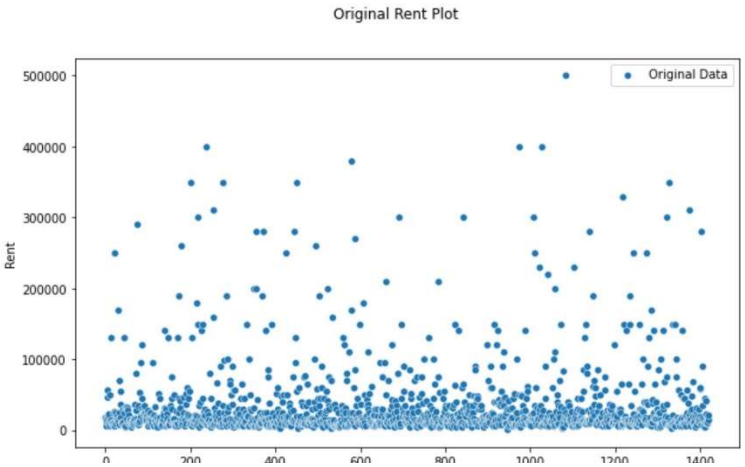


Figure 23: Original Data.

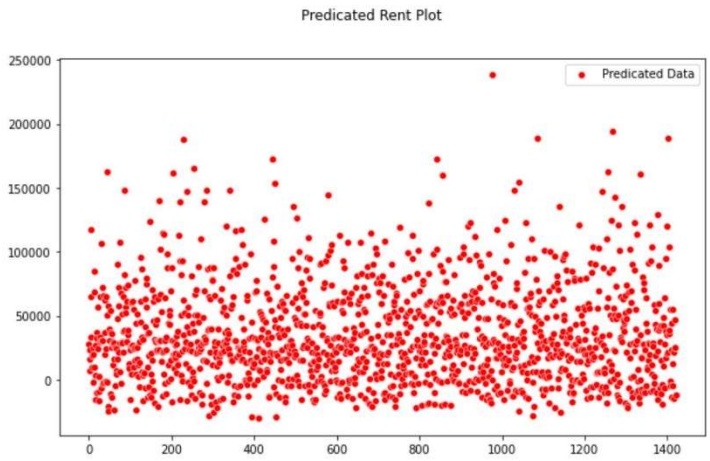


Figure 24: Predicated Data.

Mean Absolute Error = 23311.88
Root Mean Square Error = 38025.75
R Square = 0.5085
Test Accuracy Score: 50.85%

NOTE: Here the predicated plot has a difference of 50000 on y-axis where as original data has difference of 10000. That is why the model plot doesn't seem similar.

6.2 Decision Tree

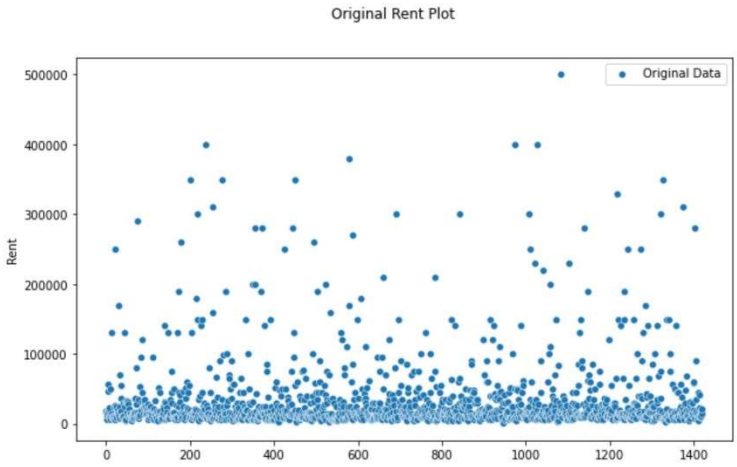


Figure 26: Original Data.

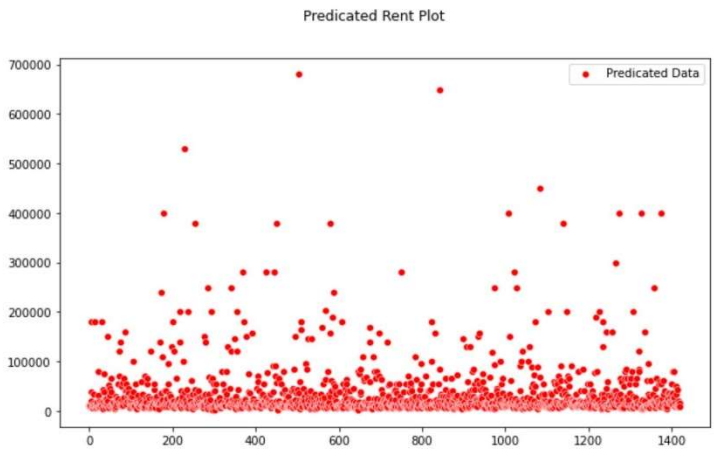


Figure 25: Predicated Data.

Mean Absolute Error = 14135.97
Root Mean Square Error = 35430.81
R Square = 0.573
Test Accuracy Score: 57.33%

6.3 Random Forest

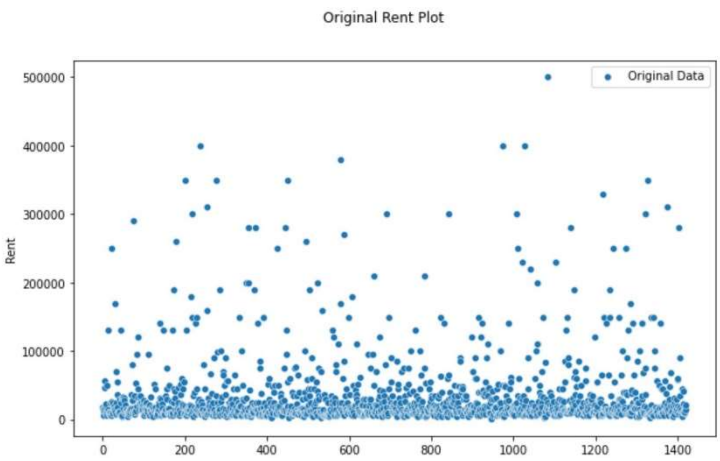


Figure 28: Original Data.

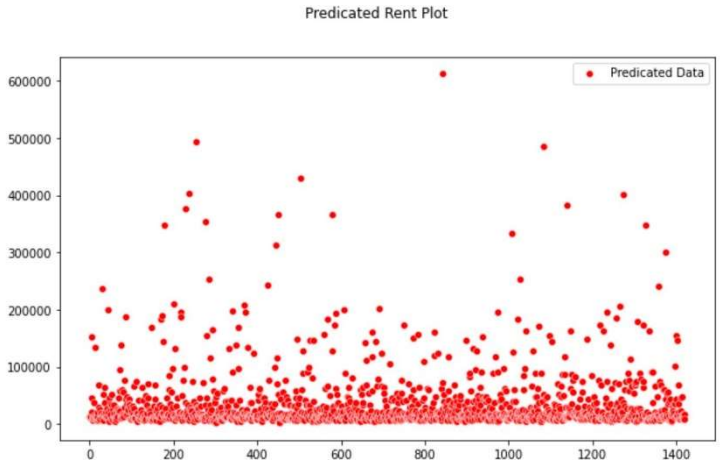


Figure 27: Predicated Data.

Mean Absolute Error = 12318.18
Root Mean Square Error = 28722.25
R Square = 0.7196
Test Accuracy Score: 71.96%

Inference:
From the plots, we can observe that Random Forest has most similar plot with the original rent plot compared to other 2.
-The rent price prediction can have error of +ve or -ve:

23311.88 for Linear Regression Model.
14135.97 for Decision Tree Model.
12318.18 for Random Forest Model.

Client Rent Prediction

Given Case:

You are working as an analyst at Real Estate Consulting firm at New Delhi. A married couples approaches your consulting agency. They need your consultation for renting a flat due to recent job transfer of the husband from Mumbai. You presented them with the options you have; based on the available dataset. Client chose their favorable options and wants to know the approximate expenses for the rent.

The option chosen by the clients are:

- They are looking for fully furnished flat of area around 850 sq. foot.
- They are family of 4 members, husband/wife and 2 kids.
- They want an extra room and bathroom for the guests as well.

Here given client requirements are as follows:

- Looking for a flat at New Delhi
- Full furnished flat.
- Area of the flat around 850 sq.ft.
- Extra room and bathroom for guest.

Input features for our model are as follows:

1. BHK
2. Size
3. Area Type
4. City
5. Furnishing Status
6. Tenant Preferred
7. Bathroom
8. Point of Contact

So, we will make necessary assumption regarding the input features.

Here input for our model is as follow:

1. BHK = 3
2. Size = 850
3. Area type = Super Area
4. City = Delhi
5. Furnishing Status = Furnished
6. Tenant Preferred = Family
7. Bathroom = 3
8. Point of Contact = Contact Owner

Rent Prediction

From Linear Regression Model, Rent is predicated to be around 23885.83.

.

From Decision Tree Model, Rent is predicated to be around 18000.00.

From Random Forest Model, Rent is predicated to be around 27012.51.