



GROUP ASSIGNMENT

House Rent Prediction

Module Code	: CT127-3-2-PDFA – Programming For Data Analysis
Intake Code	: APD2F2209CS(CYB)
Lecturer Name	: Ts.NOR ANIS ASMA BINTI SULAIMAN
Hand in Date	: Week 13 – the 15 th of December 2022

Student ID	Student Name
TP065783	KHALED ELMOATASEM BELLAH ABDALLA MOBARAK MOHAMED AWAD
TP064361	ABDELRAHMAN MOURAD ABDELSATTAR RAMADAN
TP066168	MOHAMED KHAIRY MOHAMED ABDELRAOUF

Contents

1	Introduction:.....	4
2	Install Packages & Libraries	4
3	Source code of .cvs file	5
4	Data Pre-processinwg & Cleaning:.....	5
5	Questions & Analysis	8
5.1	Question 1: What is the relationship between Rent, Area type and point of contact? 8	
5.1.1	Analysis 1.1: The houses that only have (Contact owner) point of contact.	8
5.1.2	Analysis 1.2: The average and max of the rent.....	9
5.1.3	Analysis 1.3: The relationship between Rent and Area type.	9
5.1.4	Analysis 1.4: Average house rents and house sizes for point of contact?	11
5.2	Question 2: What is the relationship between Rent, city, and size?.....	13
5.2.1	Analysis 2.1: The Most preferred city & the least Preferred City.	13
5.2.2	Analysis 2.2: The Rent Per Size.	14
5.2.3	Analysis 2.3: The relationship between house size and house rent.	15
5.2.4	Analysis 2.4: The average house sizes for each city.....	17
5.3	Question 3: What is The Relationship between Rent, City and Furnished Status? ..	18
5.3.1	Analysis 3.1: The most preferred Furnished Status & the least Furnished Status? 18	
5.3.2	Analysis 3.2: The houses that only have (Unfurnished) Furnishing Status in Mumbai City.....	19
5.3.3	Analysis 3.3: The city effect on the rent price.	20
5.3.4	Analysis 3.4: The average house rent and sizes calculated by furnished status.	21
5.4	Q4 What are the most popular houses per category?	22
5.4.1	Analysis 4.1: What are the most popular house sizes?	22
5.4.2	Analysis 4.2: What are the most popular area types?	23
5.4.3	Analysis 4.3: What are the most frequently utilized regions of facilities?	24
5.4.4	Analysis 4.4: What are the most popular house sizes?	25
5.5	Q5 What are the cities which has the highest amounts of each category?.....	26
5.5.1	Analysis 5.1: Which city has the highest total amount of BHK per city?	26
5.5.2	Analysis 5.2: Which city has the highest total amount of each area type per city?	27
5.5.3	Analysis 5.3: Which city has the highest total amount of rent per city?	28
5.5.4	Analysis 5.4: Which city has the highest total amount of furnishing status per city?.....	29
6	Additional Features	30

6.1	Additional Features-1	30
6.2	Additional Features-2	31
6.3	Additional Feature-3	32
7	Conclusion:	33
8	References:	35

1 Introduction:

For the time being, we are concentrating on the dataset on rental housing costs that has been provided to us. The dataset contains 4746 observations across 12 columns. Overall, there are 4,746 different viewpoints. In this case, we use both the numerical and category columns of our spreadsheet. Our primary objective is to identify certain trends and patterns that may highlight useful information for decision making, and our primary target variables in this case are the monthly rent, and the size of the house.

We Have These Main Variables in Mind: We begin by preparing the dataset for analysis by eliminating duplicates and other errors. The next part will go into greater depth on the various data pre-processing stages that follow. Our top priority is to identify certain patterns.

2 Install Packages & Libraries

As Shows in the figure bellow, there is two required packages to be installed and added to the library.

The Packages:

- dplyr: offering a unified collection of verbs that may be used to address many of the issues that arise while manipulating data.
- ggplot2: offers useful commands that may be used to generate complicated graphs using the data contained in a data frame.
- corrplot: corrplot is a visual exploration tool for correlation matrices that provides automatic variable reordering to help uncover hidden patterns across variables.
- plotly: graphing package generates interactive graphs of publishing quality.
- tidyr: helping to create tidy data.
- tidyverse: The core tidyverse comprises the packages you are most likely to use for routine data analysis.
- caTools: Contains a number of fundamental utility functions, including movement (rolling, running) window statistic functions and read/write operations.etc.

```
# Installing Packages.
install.packages("dplyr")
install.packages("ggplot2")
install.packages("corrplot")
install.packages("plotly")
install.packages("tidyr")
install.packages("tidyverse")
install.packages("caTools")

# Loading libraries
library(dplyr)
library(ggplot2)
library(corrplot)
library(plotly)
library(tidyr)
library(tidyverse)
library(caTools)
```

Figure 1: Package installment & importation

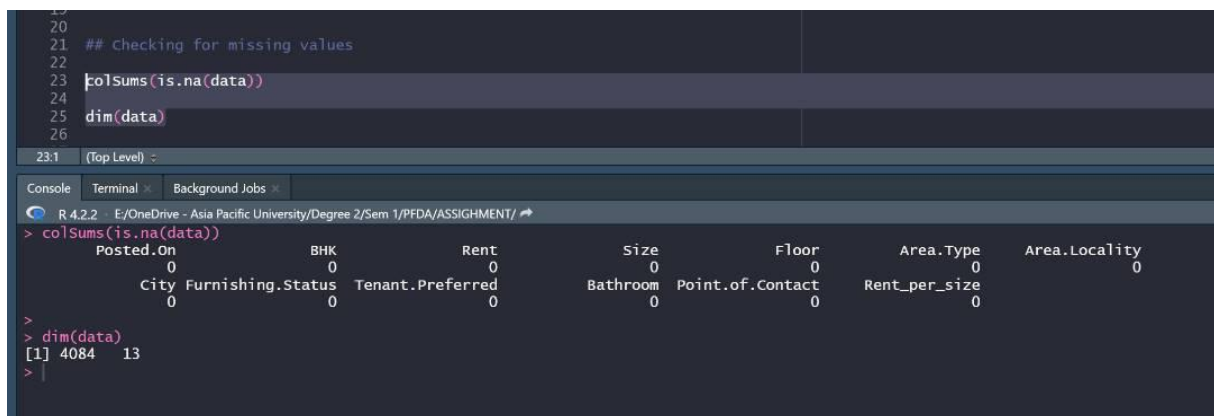
3 Source code of .csv file

```
## Loading the data set.  
  
data <- read.csv("D:\\UNIVERSITY FILES\\3'rd sem\\PFDA\\Assignment(House_Rent)\\House_Rent_Dataset")  
head(data)
```

Figure 2: Dataset source code file

4 Data Pre-processing & Cleaning:

Data pre-processing is an important step in this case where our main goal is to clean the data and make it ready for analysis to get accurate and useful results. The very first step that is being applied after loading the dataset is to check for missing values. The code that is being used for this purpose is attached below, it can be seen that there is no column where we have any missing observations. The columns function along with is.na function is used in this case.



```
21 ## Checking for missing values  
22  
23 colSums(is.na(data))  
24  
25 dim(data)  
26
```

23.1 (Top Level) :-

Console Terminal Background Jobs

R 4.2.2 E:/OneDrive - Asia Pacific University/Degree 2/Sem 1/PFDA/ASSIGHMENT/

```
> colSums(is.na(data))  
Posted.On      BHK      Rent      Size      Floor      Area.Type      Area.Locality  
0              0          0          0          0          0              0  
City Furnishing.Status Tenant.Preferred Bathroom Point.of.Contact Rent_per_size  
0              0          0          0          0          0  
> dim(data)  
[1] 4084 13  
>
```

Figure 3: Checking missing values in the dataset

The second step that is being applied is to check for garbage values in categorical columns. There are some cases when we have two categories with difference in spellings and have same meanings. The unique function is used for this purpose and we can see below that we do not have any garbage values in this case.

```
27
28 |# checking for garbage values
29
30 unique(data$Area.Type)
31 unique(data$City)
32 unique(data$Furnishing.Status)
33 unique(data$Tenant.Preferred)
34 unique(data$Point.of.Contact)
35
36
37 summary(data)
38
39
```

28:1 (Top Level) ▾

Console Terminal × Background Jobs ×

R 4.2.2 · E:/OneDrive - Asia Pacific University/Degree 2/Sem 1/PFDA/ASSIGHMENT/ ➔

```
> ## checking for garbage values
>
> unique(data$Area.Type)
[1] "Super Area" "Carpet Area" "Built Area"
> unique(data$City)
[1] "kolkata" "Mumbai" "Bangalore" "Delhi" "Chennai" "Hyderabad"
> unique(data$Furnishing.Status)
[1] "Unfurnished" "Semi-Furnished" "Furnished"
> unique(data$Tenant.Preferred)
[1] "Bachelors/Family" "Bachelors" "Family"
> unique(data$Point.of.Contact)
[1] "Contact Owner" "Contact Agent" "Contact Builder"
> |
```

Figure 4: Checking garbage values

The summary of the dataset is created and is attached below. It can be seen that the rent column has average rent of 34993 while the maximum value is 3500000. Similarly, the size on average is 967 square foot while the maximum recorded value is 8000 in this case. These values represents that we have outliers in our dataset as there is no possible way to have such a huge difference between maximum and third quartile and mean. Same is the case with bathroom variable as the average bathrooms are 1.9 while the maximum is 10.

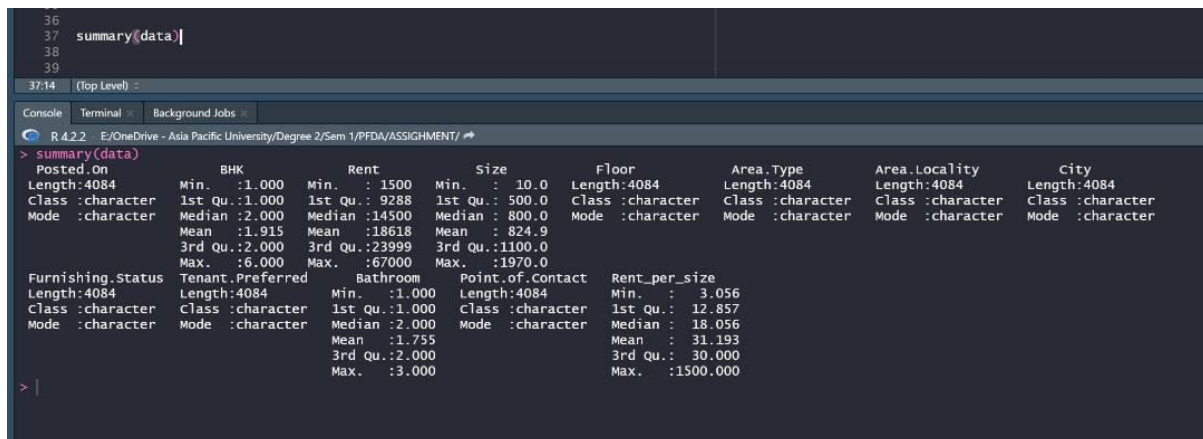


Figure 5: Summary statistics of the dataset

The outliers are removed using the interquartile range method [1] the code used to remove outliers is attached below. A function named outliers is created where we calculated 1st, 3rd quartile and the interquartile range. The next step is calculation of upper and lower limit. Finally, another function to filter outliers is created and is used.

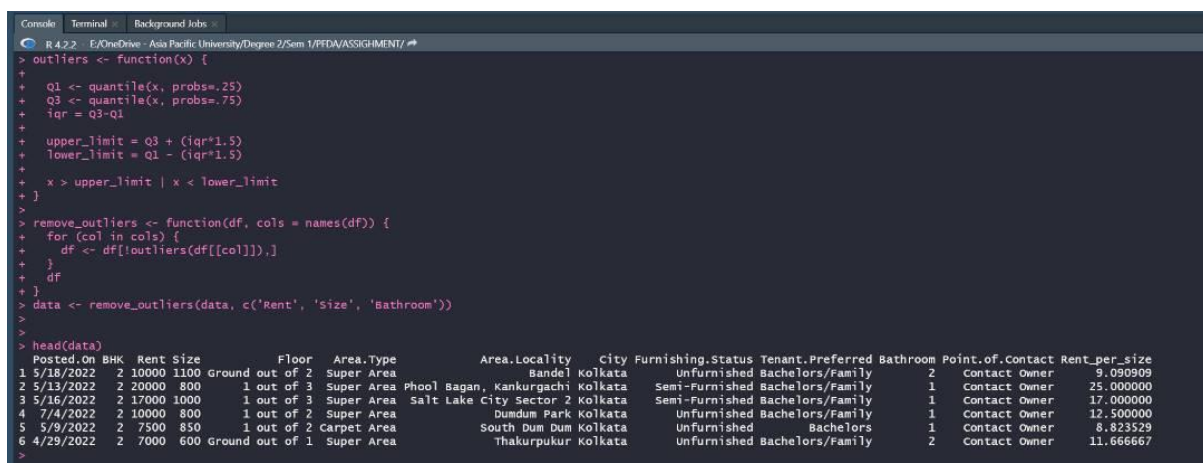


Figure 6: Removing outliers from the dataset

5 Questions & Analysis

5.1 Question 1: What is the relationship between Rent, Area type and point of contact?

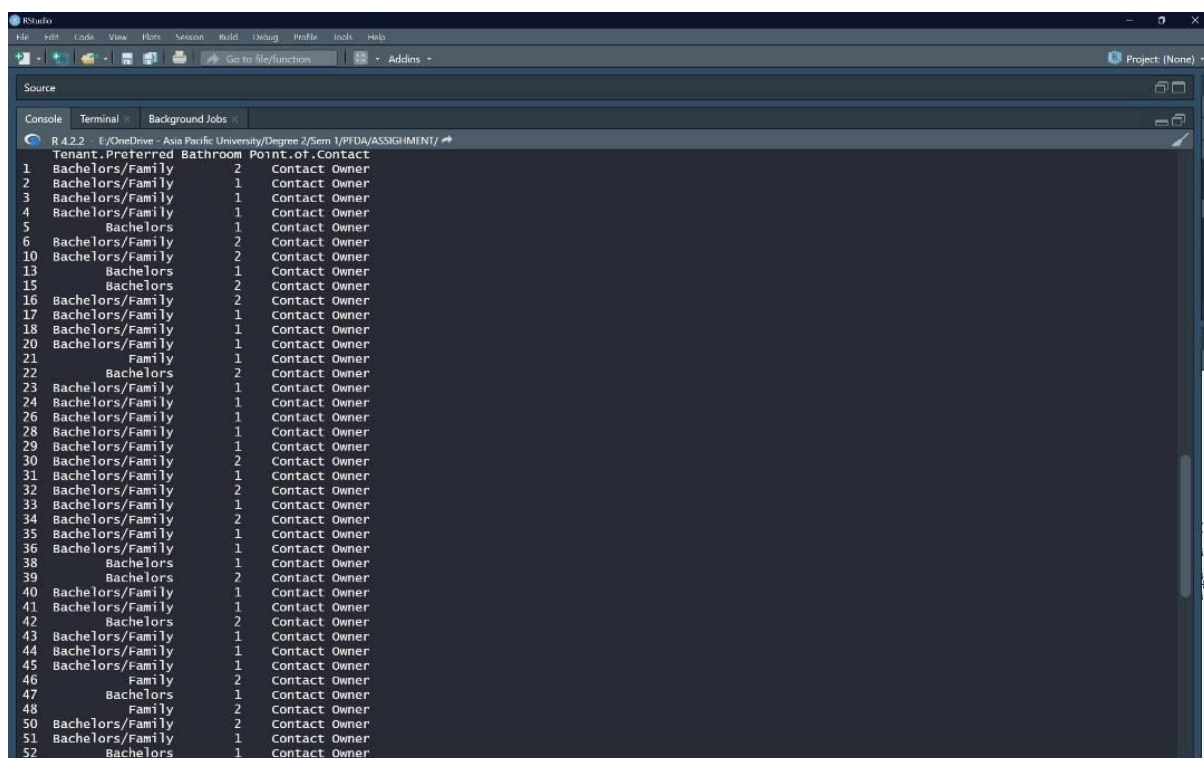
5.1.1 Analysis 1.1: The houses that only have (Contact owner) point of contact.

The very first study is conducted to verify the tenants who rented through direct contact with the property owner. In accordance with the code shown in figure 6. The results are presented in figure 7, which demonstrates that the majority of the properties that are being rented through direct contact with the owner are suitable for both singles and families.

```
# Analysis 1.1
# Q2 what are the houses that only have (Contact owner) point of contact?

data[,12]
data[which(data$Point.of.Contact=="Contact Owner"),]
```

Figure 7: A code for renters who contact the owner

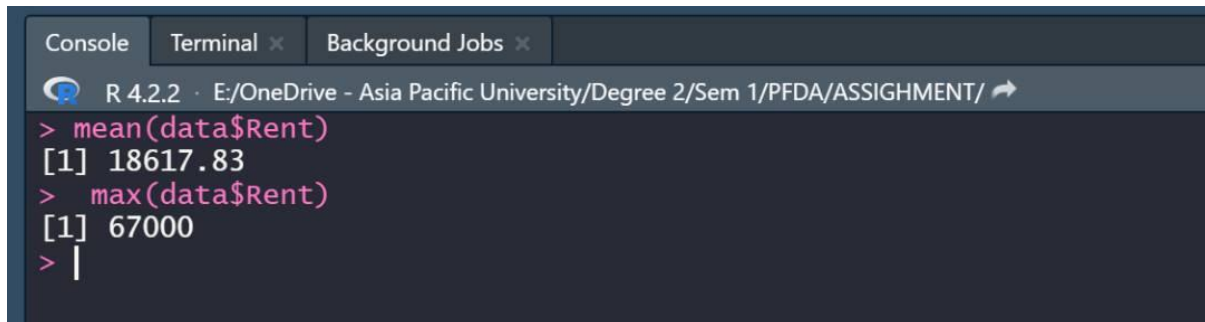


	Tenant	Preferred Bathroom	Point of Contact
1	Bachelors/Family	2	Contact Owner
2	Bachelors/Family	1	Contact Owner
3	Bachelors/Family	1	Contact Owner
4	Bachelors/Family	1	Contact Owner
5	Bachelors	1	Contact Owner
6	Bachelors/Family	2	Contact Owner
10	Bachelors/Family	2	Contact Owner
13	Bachelors	1	Contact Owner
15	Bachelors	2	Contact Owner
16	Bachelors/Family	2	Contact Owner
17	Bachelors/Family	1	Contact Owner
18	Bachelors/Family	1	Contact Owner
20	Bachelors/Family	1	Contact Owner
21	Family	1	Contact Owner
22	Bachelors	2	Contact Owner
23	Bachelors/Family	1	Contact Owner
24	Bachelors/Family	1	Contact Owner
26	Bachelors/Family	1	Contact Owner
28	Bachelors/Family	1	Contact Owner
29	Bachelors/Family	1	Contact Owner
30	Bachelors/Family	2	Contact Owner
31	Bachelors/Family	1	Contact Owner
32	Bachelors/Family	2	Contact Owner
33	Bachelors/Family	1	Contact Owner
34	Bachelors/Family	2	Contact Owner
35	Bachelors/Family	1	Contact Owner
36	Bachelors/Family	1	Contact Owner
38	Bachelors	1	Contact Owner
39	Bachelors	2	Contact Owner
40	Bachelors/Family	1	Contact Owner
41	Bachelors/Family	1	Contact Owner
42	Bachelors	2	Contact Owner
43	Bachelors/Family	1	Contact Owner
44	Bachelors/Family	1	Contact Owner
45	Bachelors/Family	1	Contact Owner
46	Family	2	Contact Owner
47	Bachelors	1	Contact Owner
48	Family	2	Contact Owner
50	Bachelors/Family	2	Contact Owner
51	Bachelors/Family	1	Contact Owner
52	Bachelors	1	Contact Owner

Figure 8: The output of Rental requests to the owner.

5.1.2 Analysis 1.2: The average and max of the rent.

The analysis that is being carried out consists of looking at the available data to determine the average rental price [18617.83] by utilizing the (mean) function, as well as checking the maximum rental price that is currently accessible [67000] by utilizing the (max) function.

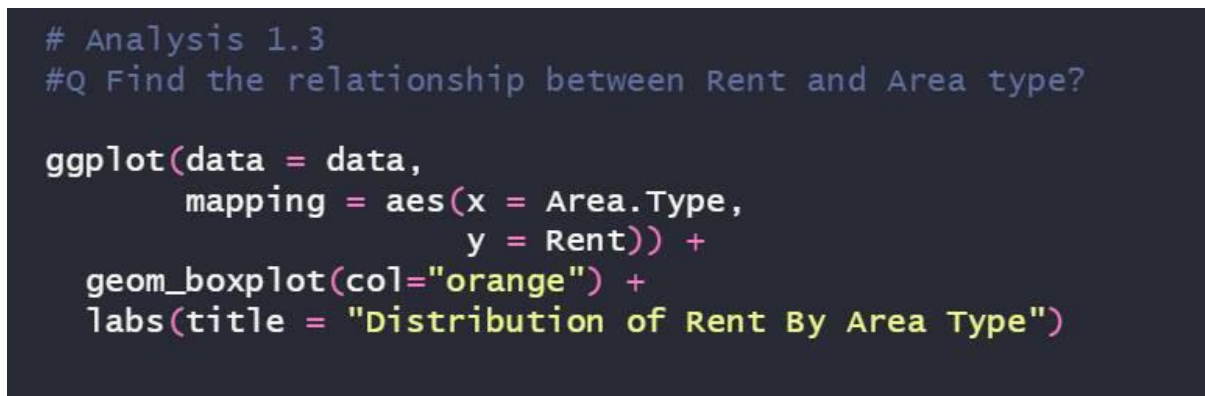


```
Console Terminal Background Jobs
R 4.2.2 · E:/OneDrive - Asia Pacific University/Degree 2/Sem 1/PFDA/ASSIGHMENT/
> mean(data$Rent)
[1] 18617.83
> max(data$Rent)
[1] 67000
> |
```

Figure 9: A code to represent the highest and lowest rental price

5.1.3 Analysis 1.3: The relationship between Rent and Area type.

The analysis that is being done is to see the distribution of house rents by area type. For this purpose, a boxplot [2] is created using the ggplot. The code that is being used is attached below. The geom_boxplot is used to create a boxplot for house rent by area type.



```
# Analysis 1.3
#Q Find the relationship between Rent and Area type?

ggplot(data = data,
       mapping = aes(x = Area.Type,
                     y = Rent)) +
  geom_boxplot(col="orange") +
  labs(title = "Distribution of Rent By Area Type")
```

Figure 10: Code to plot distribution of rent by area type

The graph that is being created in this case is attached below. It can be seen that the highest average rent of houses is observed in “Carpet Area”, the second place is of “Super Area” and the lowest average rent of houses is observed in “Built Area”. This tells us that the cheapest area in term of house rents is the Built Area while the expensive area in terms of house rent is the Carpet Area.

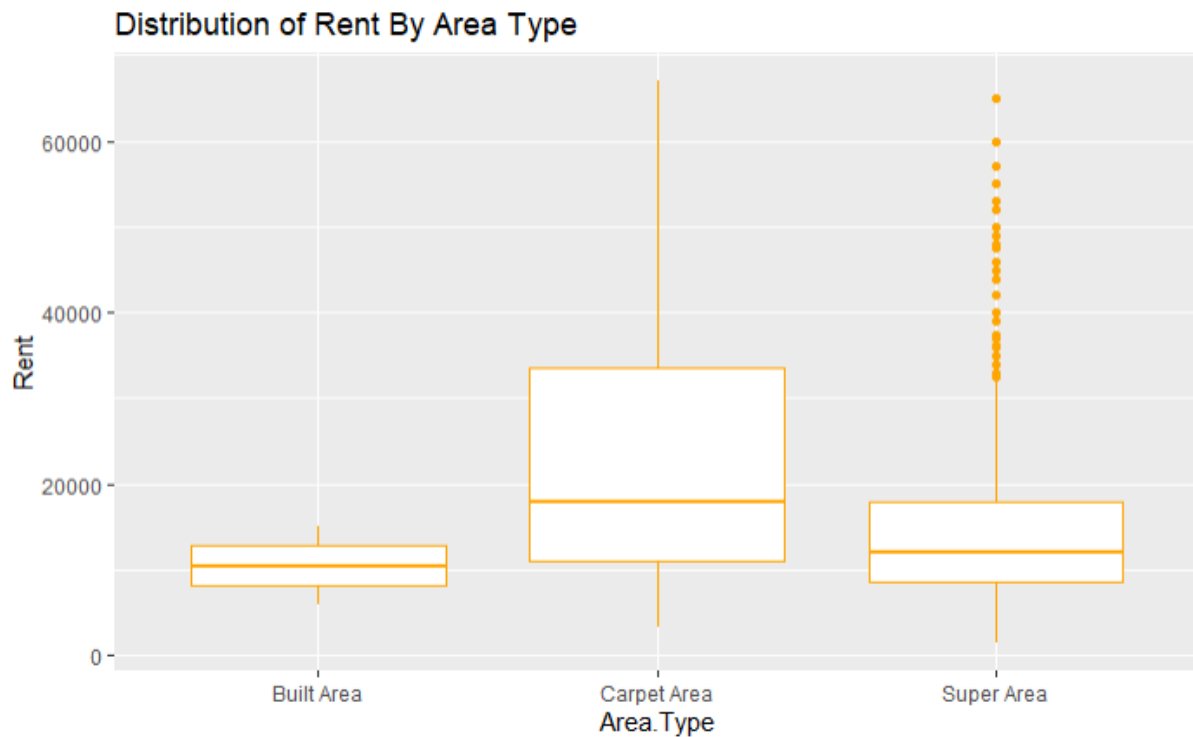


Figure 11: Boxplot for distribution of rent by area type

5.1.4 Analysis 1.4: Average house rents and house sizes for point of contact?

Following this step, the group by and summaries functions are utilised in order to determine the average house rentals as well as the average house sizes for the point of contact variable. The code block that was attached below and used to compute and generate a pie chart may be found below. The plotting of the typical house sizes is accomplished with the help of the same code.

```
# Analysis 1.4
#Q Find average house rents and average house sizes for point of contact?

temp <- data %>%
  group_by(Point.of.Contact) %>%
  summarise(Avg_Size = mean(Size),
            Avg_Rent = mean(Rent))

ggplot(temp, aes(x = "",
                 y = Avg_Rent,
                 fill = Point.of.Contact)) +
  geom_col() +
  geom_text(aes(label = round(Avg_Rent, 2)),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title = "Average Rent By Point of Contact")

ggplot(temp, aes(x = "",
                 y = Avg_Size,
                 fill = Point.of.Contact)) +
  geom_col() +
  geom_text(aes(label = round(Avg_Size, 2)),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title = "Average House Sizes By Point of Contact")
```

Figure 12: Code to calculate average size and rent by point of contact

The resulting plots can be found attached further down. It is clear from looking at the two pie charts that were developed to illustrate the average rent and the average size of houses that the average rent is greatest for those residences in which an agent serves as the primary point of contact. This is also true regarding the average size of houses.

We are able to state that the average house size of those houses for whom an agent serves as the point of contact has the highest rent, and we can also claim that the rent for these houses has the highest value. The rents are the most affordable for properties in which a builder serves as the primary point of contact, and the average square footage of builder-made homes is also the smallest.

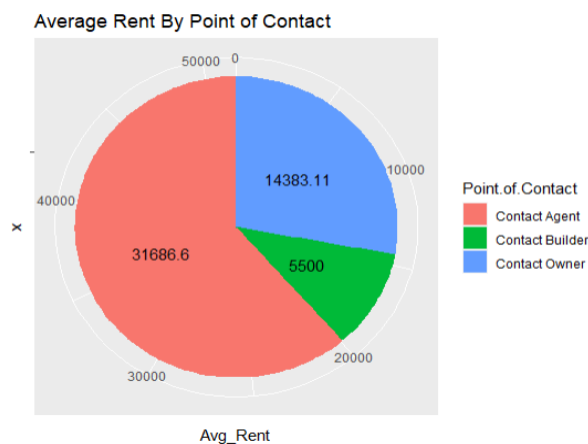


Figure 13: Average rent by point of contact

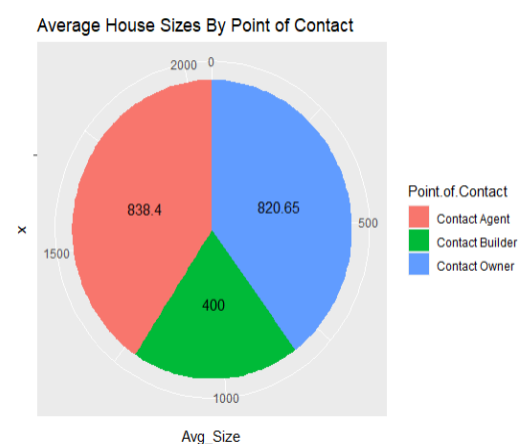


Figure 14: Average house size by point of contact

5.2 Question 2: What is the relationship between Rent, city, and size?

5.2.1 Analysis 2.1: The Most preferred city & the least Preferred City.

This study intends to figure out which area is least well known with settlers and which city is generally famous for foreigners leasing. We need to create a variable called city count, which equals the data filtered by city then summarize the count then equal length pair the date and then arrange to descend the count. The results indicate that "Mumbai" is the least preferred city to rent a property in, while "Chennai" is the most preferred.

```
#Q2 what is the relationship between Rent, city and size?

# Analysis 2.1
# what is the preferred city & what is the least Preferred City?

City_Count<-data%>%group_by(City)%>%summarise(count = length(BHK))%>%arrange(desc(count))
head(City_Count)
ggplot(City_Count,
       mapping = aes(x= City,
                     y= count,
                     fill = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "City", y = "count", title = "Houses counted by cities") + theme_bw()
```

Figure 15: Code to find the most and least cities to live in

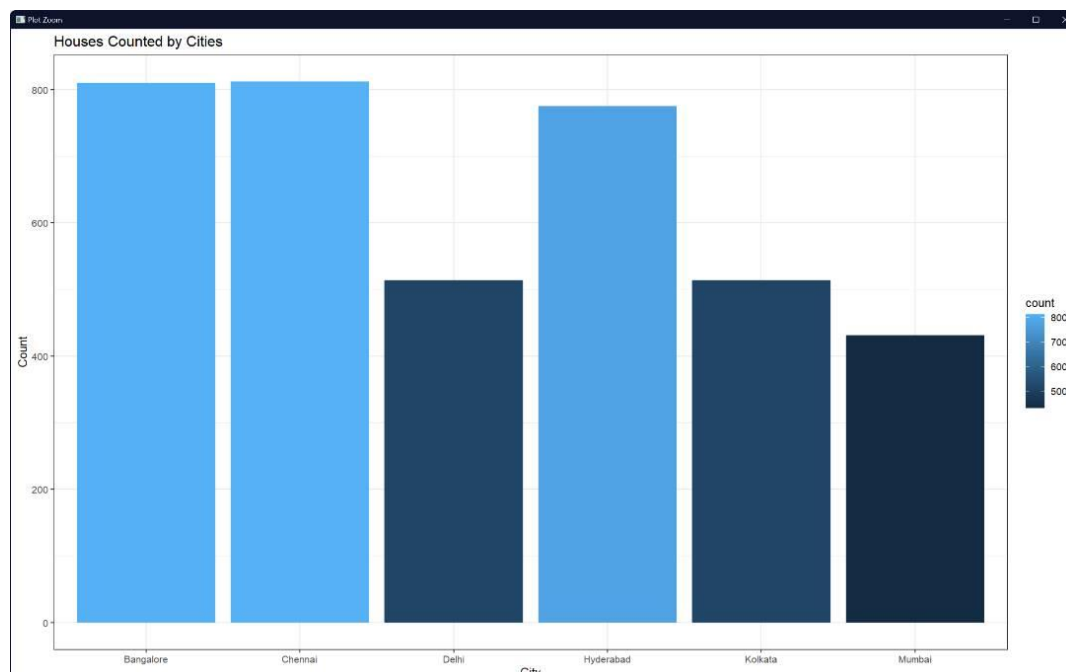


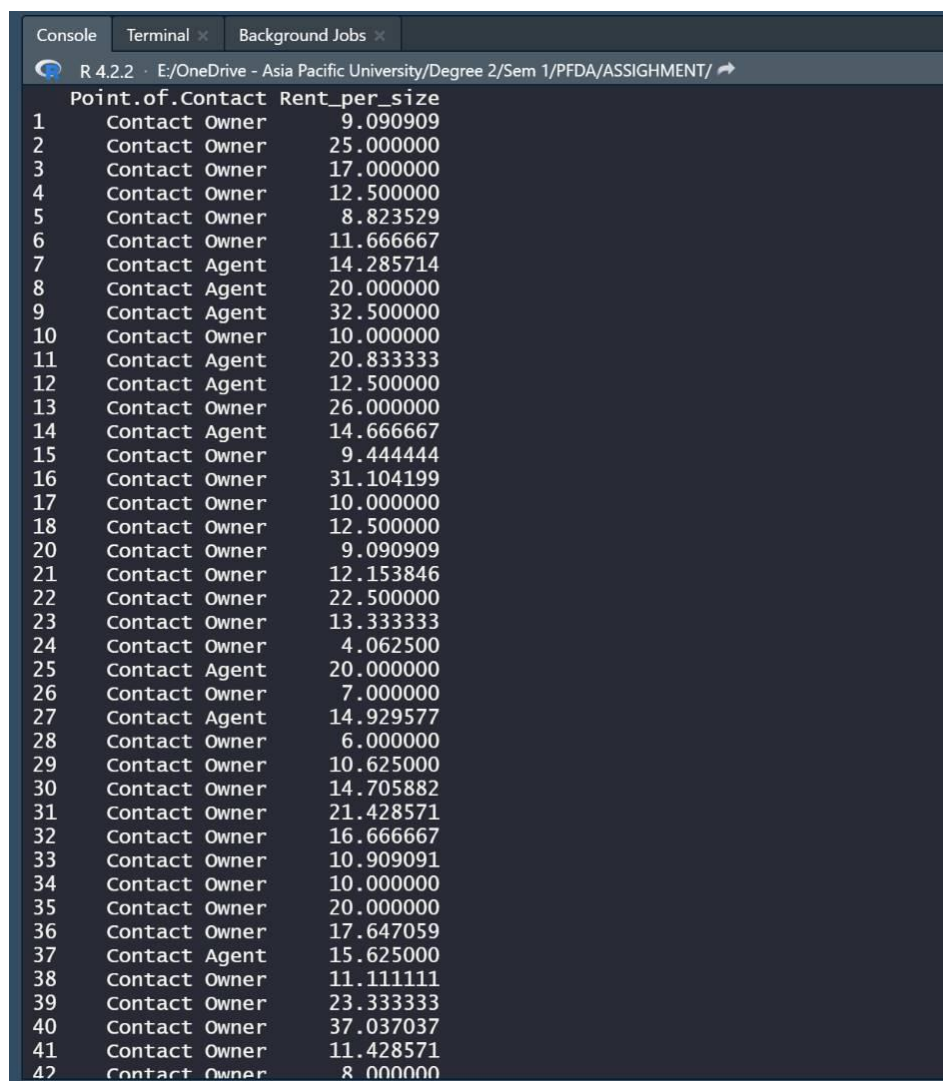
Figure 16: The figure shows that Chennai is preferred to live in and Mumbai the least preferred

5.2.2 Analysis 2.2: The Rent Per Size.

The analysis that is carried out involves displaying all of the potential sizes of the homes that have been rented by making use of the code shown below in Figure 14. This occurs regardless of whether the rent was negotiated with the property owner directly or with the help of a middleman (Agent). Figure 15 displays all of the dimensions that were collected as output.

```
# Analysis 2.2
# Q Find The Rent Per Size?
data$Rent_per_size<-data$Rent/data$Size
data
```

Figure 17: A code shows the size of rented houses



The screenshot shows an R console window with the following output:

	Point.of.Contact	Rent_per_size
1	Contact Owner	9.090909
2	Contact Owner	25.000000
3	Contact Owner	17.000000
4	Contact Owner	12.500000
5	Contact Owner	8.823529
6	Contact Owner	11.666667
7	Contact Agent	14.285714
8	Contact Agent	20.000000
9	Contact Agent	32.500000
10	Contact Owner	10.000000
11	Contact Agent	20.833333
12	Contact Agent	12.500000
13	Contact Owner	26.000000
14	Contact Agent	14.666667
15	Contact Owner	9.444444
16	Contact Owner	31.104199
17	Contact Owner	10.000000
18	Contact Owner	12.500000
20	Contact Owner	9.090909
21	Contact Owner	12.153846
22	Contact Owner	22.500000
23	Contact Owner	13.333333
24	Contact Owner	4.062500
25	Contact Agent	20.000000
26	Contact Owner	7.000000
27	Contact Agent	14.929577
28	Contact Owner	6.000000
29	Contact Owner	10.625000
30	Contact Owner	14.705882
31	Contact Owner	21.428571
32	Contact Owner	16.666667
33	Contact Owner	10.909091
34	Contact Owner	10.000000
35	Contact Owner	20.000000
36	Contact Owner	17.647059
37	Contact Agent	15.625000
38	Contact Owner	11.111111
39	Contact Owner	23.333333
40	Contact Owner	37.037037
41	Contact Owner	11.428571
42	Contact Owner	8.000000

Figure 18: The output of available house sizes

5.2.3 Analysis 2.3: The relationship between house size and house rent.

Using the group by and summary functions of dplyr, we were able to determine the typical dimensions of a home in each municipality. The primary objective of this visualisation is to get insight into the pattern of new residential construction occurring in each city. Please find attached below the code block that was utilised to determine the average house size and then plot it in a bar graph [14]. A bar graph of the results of calculation can be plotted with the help of the Geom_bar programme.

```
# Analysis 2.4
#Q Find The average house sizes for each city?

temp <- data %>%
  group_by(City) %>%
  summarise(Avg_Size = mean(Size))

ggplot(data = temp,
        mapping = aes(x = City,
                      y = Avg_Size,
                      fill = City)) +
  geom_bar(stat="identity", position = "dodge") +
  labs(title = "Average House Sizes By City")
```

Figure 19: Code to calculate average house size by city and plot it

The plot that was obtained in this manner is shown down below. It is clear that Hyderabad possesses the largest house sizes, whereas Delhi is home to the smallest house sizes. In light of the previous analysis graph, we will compare it with this one. We can see that despite the fact that the average house size in Hyderabad is the largest compared to all other cities, the average house rent there is the second lowest. In a similar vein, the average housing rent in Delhi is the second highest in the world, despite having the smallest dwelling sizes. This shows that the location of a house has a significant impact on the amount of money it may be rented for.

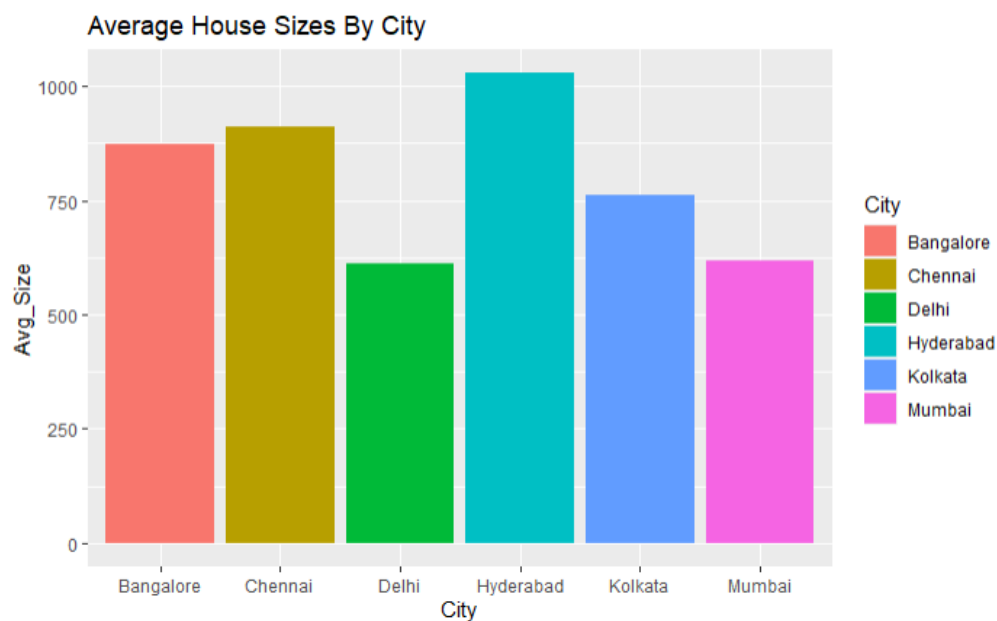


Figure 20: Average house size by city

5.2.4 Analysis 2.4: The average house sizes for each city.

The relationship between a home's square footage and monthly rent will be examined in the future. A regression line is being fitted to a scatter plot so that we can observe the relationship pattern and general trend in order to achieve this objective. The `geom_point` function can be used to create a scatter graph, while the `geom_smooth` function can be used to add a regression line to the graph.

```
# Analysis 2.3
#Q Find the relationship between house size and house rent?

ggplot(data, aes(x=Size, y=Rent)) +
  geom_point()+
  geom_smooth() +
  labs(title = "Relationship Between Size & Rent")
```

Figure 21: Code to plot relationship between size and rent of houses

The scatter plot [3] thus created is attached below. It can be seen that the relationship between both variables is positive. A positive relationship means that with increase in size of a house, the rent of a house increases and with decrease in size of a house the rent of a house decreases as well. The regression line is not linear but as the trend is positive and is upward, hence we can say that the relationship exist between both variables and the relationship is positive.

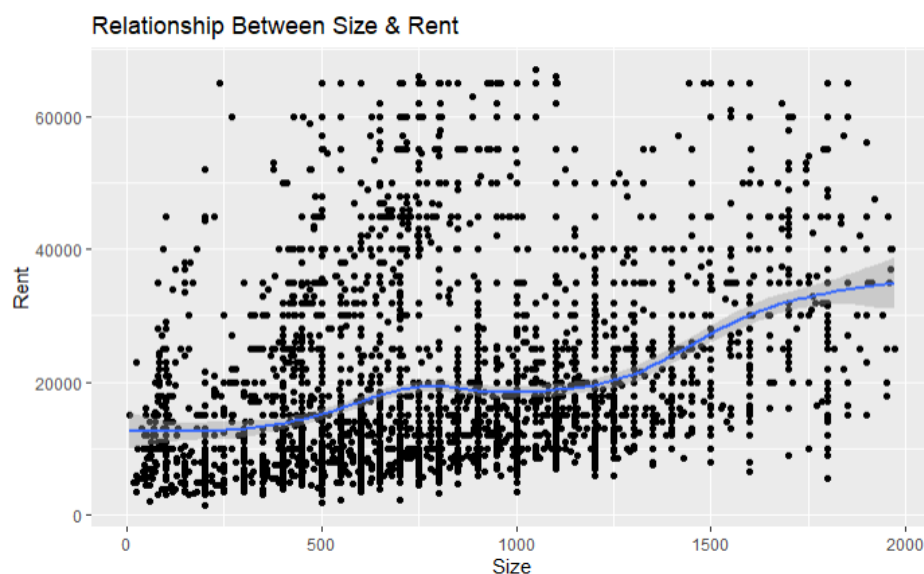


Figure 22: Scatter plot to show the relationship between size and house rent

5.3 Question 3: What is The Relationship between Rent, City and Furnished Status?

5.3.1 Analysis 3.1: The most preferred Furnished Status & the least Furnished Status?

According to the code in Figure 20, the "Semi-furnished" furnishing status is the one that is chosen by persons the most frequently. We need to establish a variable called furnished Status count that will equal the data that has been filtered by furnished Status, followed by the summary count, the date, and the top 5. This variable will be equal to the data that has been filtered by furnished Status. This will show us in the result below that furnished apartments are the least popular, in contrast to unfurnished apartments, which came in second place in terms of demand directly.

```
#Q3 What is The Relationship between Rent,City & Furnished Status?

# Analysis 3.1
# Q What is the preferred Furnished Status & what is the least Furnished Status?

Furnished_Status<-data%>%group_by(Furnishing.Status)%>%summarise(count = length(BHK))%>%top_n(5)
head(Furnished_Status)
ggplot(Furnished_Status,
       mapping = aes(x= Furnishing.Status,
                     y= count,
                     fill = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Furnished Status", y = "count", title = "Count By Facilities") + theme_bw()
```

Figure 23: The code displays the overall preferred furnishing status

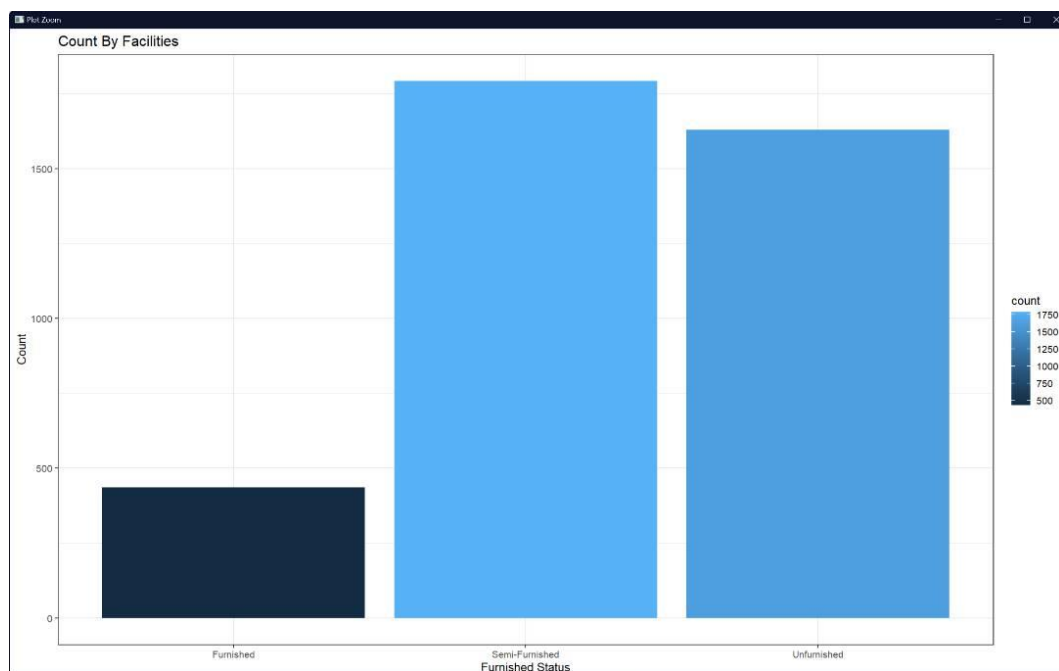


Figure 24: Its displayed that semi-furnished is the most preferred(Furnishing Status)

5.3.2 Analysis 3.2: The houses that only have (Unfurnished) Furnishing Status in Mumbai City.

We determined that Mumbai is the preferred city for residents after compiling statistics on the existing cities and selecting the one that the majority prefers. Another statistic found that the highest-selling apartments are those without furniture. The code displayed in Figure 21 gives data in regards to empty rental homes in a specific city (Mumbai).

This analysis was based on previous research because, as shown in Figure 13, "Mumbai" is the most common location. Figure 20 demonstrates that the majority of homes lack furniture. Both of these results are based on previous research.

```

206
207
208 # Analysis 3.2
209 # Q what are the houses that only have (Unfurnished) Furnishing Status in Mumbai City?
210
211 data[which(data$City=="Chennai" & data$Furnishing.Status=="Semi-Furnished"),]
212
213
214
215
211.1 (Top Level) :

```

Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	
2988	7/6/2022	2	15000	1100	1 out of 2	Super Area	Medavakkam	Chennai	Semi-Furnished	Bachelors	2
2989	5/21/2022	2	6500	1000	Ground out of 1	Super Area	Urapakkam, Vandalur R.F, GST Road	Chennai	Semi-Furnished	Bachelors/Family	2
2992	6/25/2022	1	15000	650	Ground out of 2	Carpet Area	Kambar Colony	Chennai	Semi-Furnished	Bachelors/Family	1
2993	5/18/2022	2	15000	1000	17 out of 31	Carpet Area	Thalambur	Chennai	Semi-Furnished	Bachelors	2
2995	6/14/2022	3	29000	1200	1 out of 3	Carpet Area	Ashok Nagar	Chennai	Semi-Furnished	Bachelors	3
2996	6/19/2022	3	20000	1710	15 out of 29	Super Area	Padur, Old Mahabalipuram Road	Chennai	Semi-Furnished	Bachelors/Family	3
2999	5/26/2022	2	29000	1300	Ground out of 2	Super Area	Annanagar west	Chennai	Semi-Furnished	Bachelors/Family	2
3000	7/10/2022	2	26000	1027	15 out of 19	Super Area	Karapakkam	Chennai	Semi-Furnished	Bachelors	2
3006	6/20/2022	2	23000	920	2 out of 2	Carpet Area	KK Nagar	Chennai	Semi-Furnished	Bachelors/Family	2
3007	6/28/2022	2	10000	600	Ground out of 2	Carpet Area	Sadhanadhapuram	Chennai	Semi-Furnished	Bachelors/Family	2
3009	6/2/2022	2	15000	1200	Ground out of 2	Super Area	Ramapuram	Chennai	Semi-Furnished	Bachelors/Family	2
3010	5/12/2022	1	11000	550	3 out of 4	Super Area	Poongavanapuram	Chennai	Semi-Furnished	Bachelors/Family	1
3015	5/12/2022	2	20000	1000	Ground out of 1	Super Area	Thiruvanniyur	Chennai	Semi-Furnished	Bachelors/Family	2
3017	5/12/2022	2	12000	800	1 out of 3	Super Area	Purasawalkam, PH Road	Chennai	Semi-Furnished	Bachelors/Family	1
3018	5/12/2022	2	27000	1550	2 out of 3	Super Area	Alwarpet	Chennai	Semi-Furnished	Bachelors/Family	3
3021	6/28/2022	2	22000	700	3 out of 4	Super Area	T Nagar	Chennai	Semi-Furnished	Bachelors/Family	2
3023	5/12/2022	2	7000	450	1 out of 2	Super Area	Pammal	Chennai	Semi-Furnished	Bachelors/Family	1
3025	5/12/2022	2	9000	770	1 out of 2	Super Area	Tambaram, GST Road	Chennai	Semi-Furnished	Bachelors/Family	2
3026	5/25/2022	1	4500	150	Ground out of 2	Super Area	Velachery	Chennai	Semi-Furnished	Bachelors/Family	1
3028	6/21/2022	2	18000	1100	Ground out of 3	Carpet Area	Iyyappanthangal	Chennai	Semi-Furnished	Bachelors	2
3032	5/12/2022	1	6500	250	Ground out of 5	Super Area	Adambakkam	Chennai	Semi-Furnished	Bachelors/Family	1
3033	6/23/2022	2	10000	790	2 out of 2	Carpet Area	Kovilancheri	Chennai	Semi-Furnished	Bachelors/Family	2
3034	4/30/2022	2	13000	600	1 out of 3	Super Area	Medavakkam	Chennai	Semi-Furnished	Bachelors/Family	2
3039	5/29/2022	3	35000	1550	3 out of 4	Carpet Area	Thiruvanniyur	Chennai	Semi-Furnished	Bachelors	2
3040	7/10/2022	3	34999	1462	3 out of 4	Carpet Area	Nandambakkam	Chennai	Semi-Furnished	Family	3
3041	6/17/2022	2	11000	970	Ground out of 2	Carpet Area	Pallikaranai	Chennai	Semi-Furnished	Bachelors	2
3046	5/16/2022	2	12000	1000	Ground out of 2	Super Area	Tambaram Sanatorium	Chennai	Semi-Furnished	Bachelors/Family	2
3047	5/26/2022	2	10000	1030	1 out of 2	Carpet Area	Camp Road	Chennai	Semi-Furnished	Bachelors/Family	2
3052	7/6/2022	2	30000	1200	3 out of 3	Carpet Area	Mandaiveli	Chennai	Semi-Furnished	Bachelors/Family	2

Figure 25: The Code displayed the rented houses with a("unfurnished") furnished status.

5.3.3 Analysis 3.3: The city effect on the rent price.

The rent distributions for each city are plotted once more using boxplot. The utilized code block can be found attached below.

```
# ## Analysis 3.3
#Q Does the city affect on the rent price ?

ggplot(data = data,
       mapping = aes(x = City,
                     y = Rent)) +
  geom_boxplot(col="black") +
  labs(title = "Distribution of Rent By City")
```

Figure 26: Code to plot distribution of rent of houses by city

When the graph that was generated is examined, it is clear that the city of Kolkata has the cheapest average rent for houses, whilst the city of Mumbai has the most expensive average rent for houses. The monthly rent for a property in Delhi is the second highest in the country, while the monthly rent for a house in Hyderabad is the second lowest.

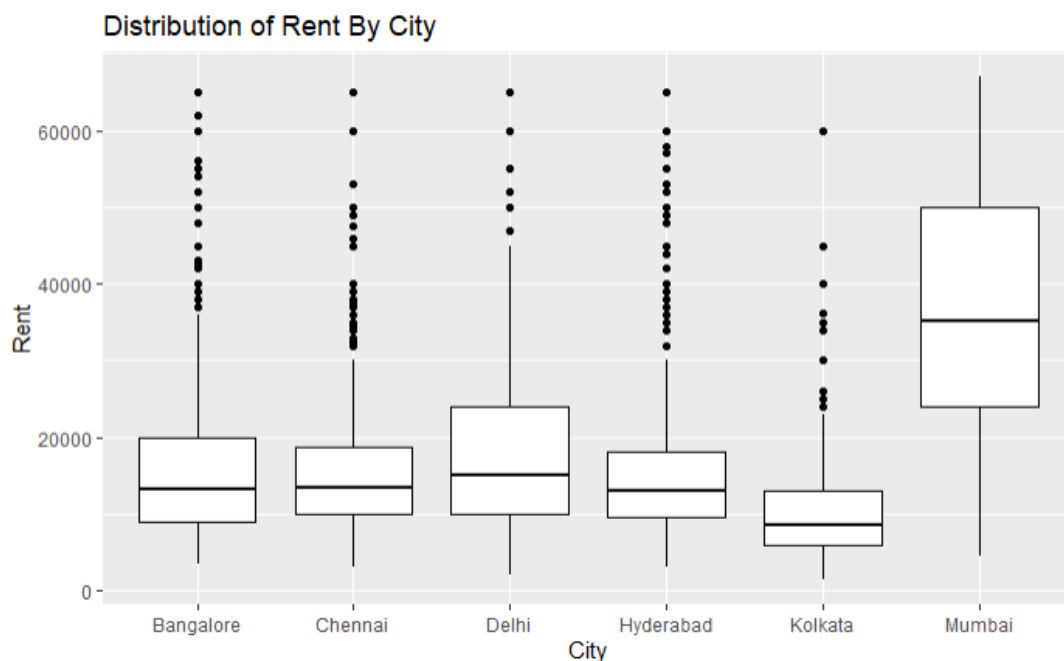


Figure 27: Boxplot for house rents by city

5.3.4 Analysis 3.4: The average house rent and sizes calculated by furnished status.

The average rent for a house and the average size of a house both take into account whether or not a house is furnished. The mean and the group by and summaries functions are utilized in this particular instance. The code block that was used to calculate the data and then plot them is attached.

```
# Analysis 3.4
# Q Find the average house rent and house sizes are calculated by furnished status of a house?

temp <- data %>%
  group_by(Furnishing.Status) %>%
  summarise(Avg_Size = mean(Size),
            Avg_Rent = mean(Rent))

ggplot(temp, aes(x = "",
                 y = Avg_Rent,
                 fill = Furnishing.Status)) +
  geom_col() +
  geom_text(aes(label = round(Avg_Rent, 2)),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title = "Average Rent By Furnishing Status")

ggplot(temp, aes(x = "",
                 y = Avg_Size,
                 fill = Furnishing.Status)) +
  geom_col() +
  geom_text(aes(label = round(Avg_Size, 2)),
            position = position_stack(vjust = 0.5)) +
  coord_polar(theta = "y") +
  labs(title = "Average House Sizes By Furnishing Status")
```

Figure 28: Calculate and plot average house size and average house rent by furnishing status

The plots that were obtained in this manner are shown. The average size of furnished residences is also the largest, making their rent the most expensive overall. Furnished houses also have the highest average rent. On the other hand, the average house size of unfurnished houses is the smallest, and the same can be said about the average rent for these types of homes.

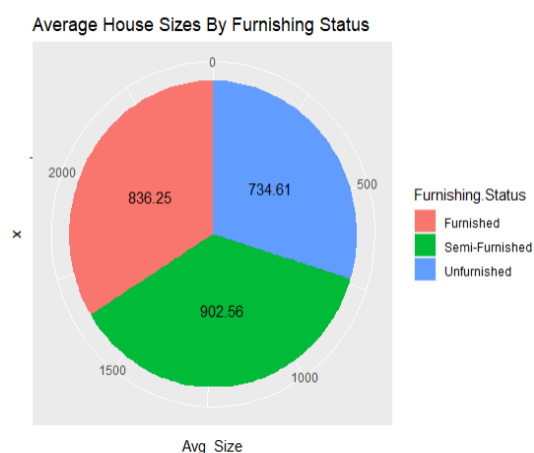


Figure 29: Average house size by furnishing status

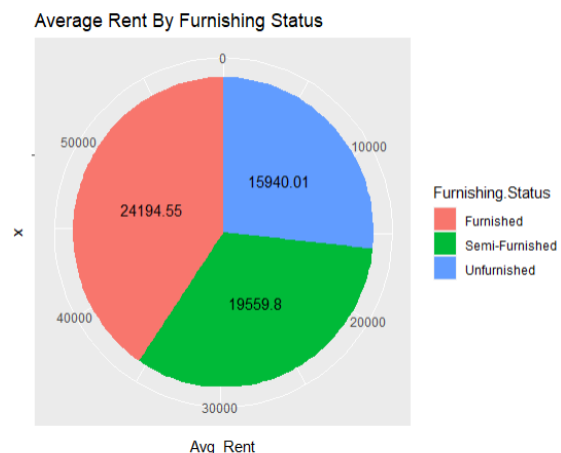


Figure 30: Average house rent by furnishing status

5.4 Q4 What are the most popular houses per category?

5.4.1 Analysis 4.1: What are the most popular house sizes?

```
# Q4 what are the most popular houses per category?

# Analysis 4.1
# Q what are the most popular amount of bathrooms?
Area_Count<-data%>%group_by(Area.Type)%>%summarise(count = length(BHK))
head(Area_Count)
ggplot(Area_Count,
       mapping = aes(x= Area.Type,
                     y= count,
                     fill = count)) +
  geom_bar(stat = "identity") +
  labs(x = "Area Type", y = "Count", title = "Count Of Houses By Area Type") + theme_bw()
```

Figure 31: A code shows the area with the highest amount of rented houses

As we see in figure [32], the (Super area) it's the most preferred area type to rent a house while and the built area is the least preferred area type to rent a house in.

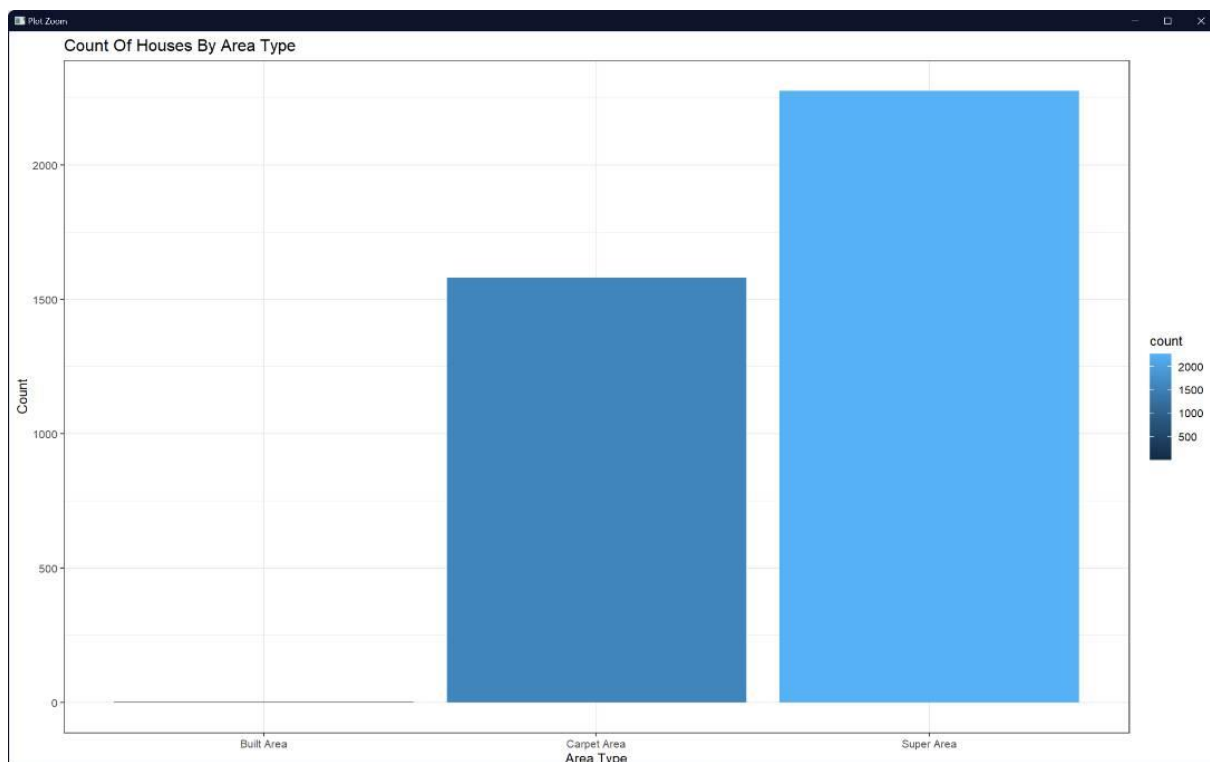


Figure 32: Its displayed that (Super Area) has the highest amount of houses

5.4.2 Analysis 4.2: What are the most popular area types?

```
# Analysis 4.2
# Q what are the most popular area types ?

Bathroom_Count<-data%>%group_by(Bathroom)%>%summarise(count = length(BHK))%>%top_n(5)
head(Bathroom_Count)
ggplot(Bathroom_Count,
       mapping = aes(x= Bathroom,
                      y= count,
                      fill = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Bathrooms" ,y = "Count",title = "Count Of Bathrooms in 1 House") + theme_bw()
```

Figure 33: A code for displaying the average number of bathrooms in one house

According to figure [34] the most popular houses to be rented are those with two bathrooms in there BHK and the least popular is having 3 bathrooms in the same house.

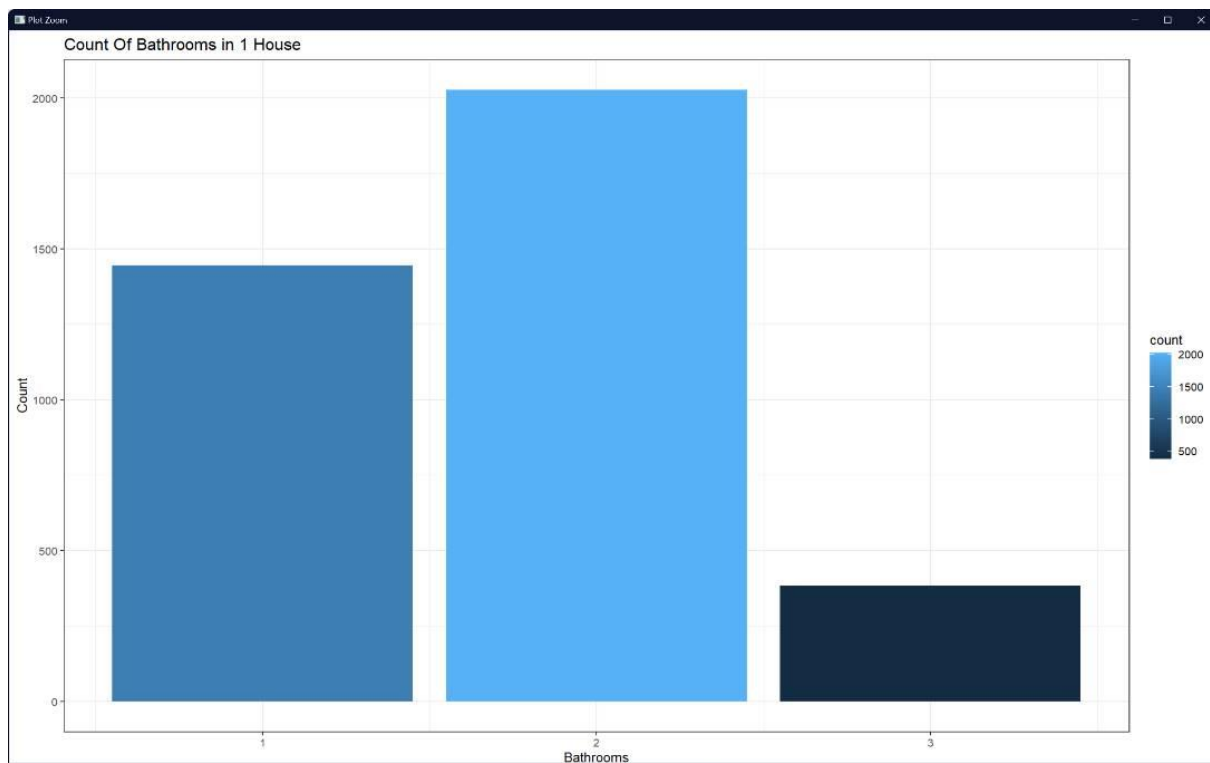


Figure 34: Its shows that a (2 bathrooms) is the most popular choice and (3) is the least

5.4.3 Analysis 4.3: What are the most frequently utilized regions of facilities?

```
# Analysis 4.3
# Q What are the most popular area of Facilities?
Facilities_Count<-data%>%group_by(Area.Locality)%>%summarise(count = length(BHK))%>%top_n(5)
head(Facilities_Count)
ggplot(Facilities_Count,
       mapping = aes(x= Area.Locality,
                     y= count,
                     fill = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Area_Facilities", y = "Count", title = "Count By Facilities") + theme_bw()
```

Figure 35: A code that displays the facilities availability in each city.

As shown in Figure [36], we can see that more tenants from (Electronic Area) has more facilities than the other localities. As there are many other localities in the Mayapur, the top 5 from the list have been presented in the graph to show the most accurate results. However, we cannot fully depend on the graph as there are still hundreds of other localities.

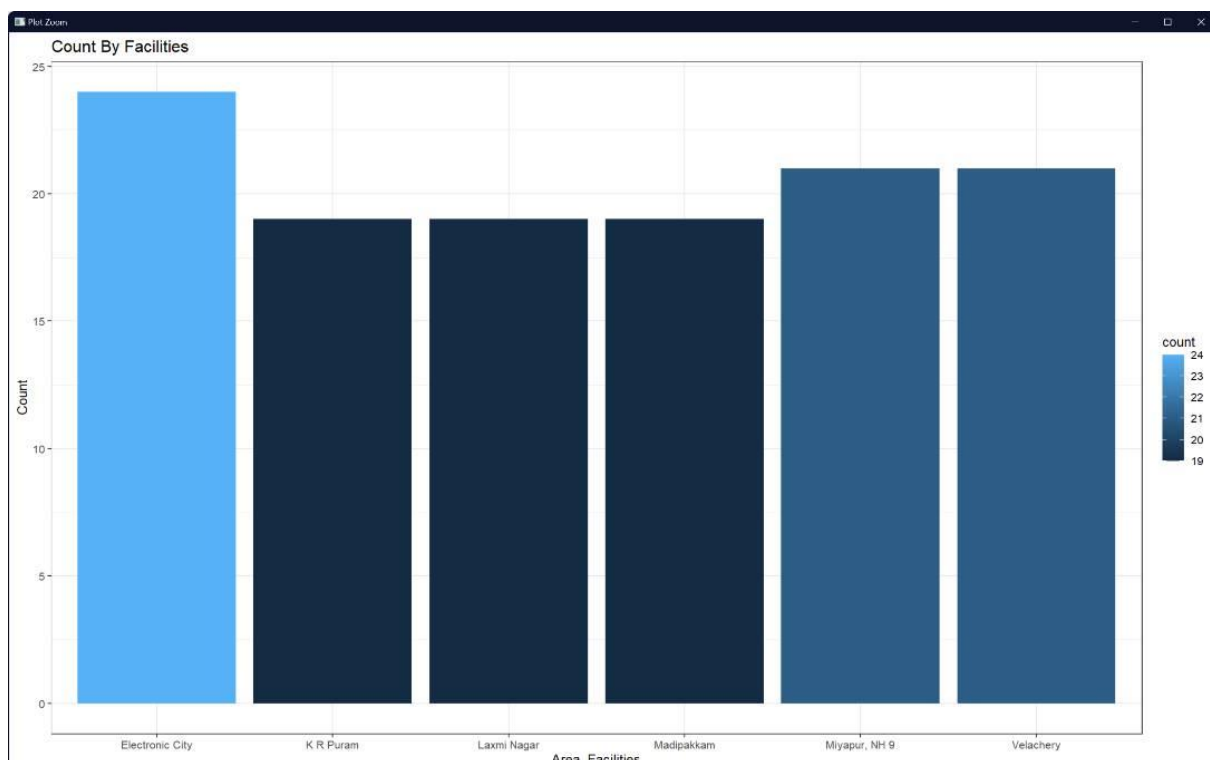


Figure 36: It's shown that (Electronic Area) has the highest amount of facilities

5.4.4 Analysis 4.4: What are the most popular house sizes?

```
# Analysis 4.4
# Q what are the most popular house sizes?
Size_Count<-data%>%group_by(Size)%>%summarise(count = length(Size))%>%top_n(8)
head(Size_Count)
ggplot(Size_Count,
       mapping = aes(x= Size,
                     y= count,
                     fill = count)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Count By Size") + theme_bw()
```

Figure 37: A code to display the count the number houses sizes in each size.

Figure [38] presents us with the number of houses from the same sizes. As there are too many sizes for the graph to display, it displays only the top 8 sizes from the list. We can see that houses with a size of 600 square foot are the most popular compared to the rest of the sizes.

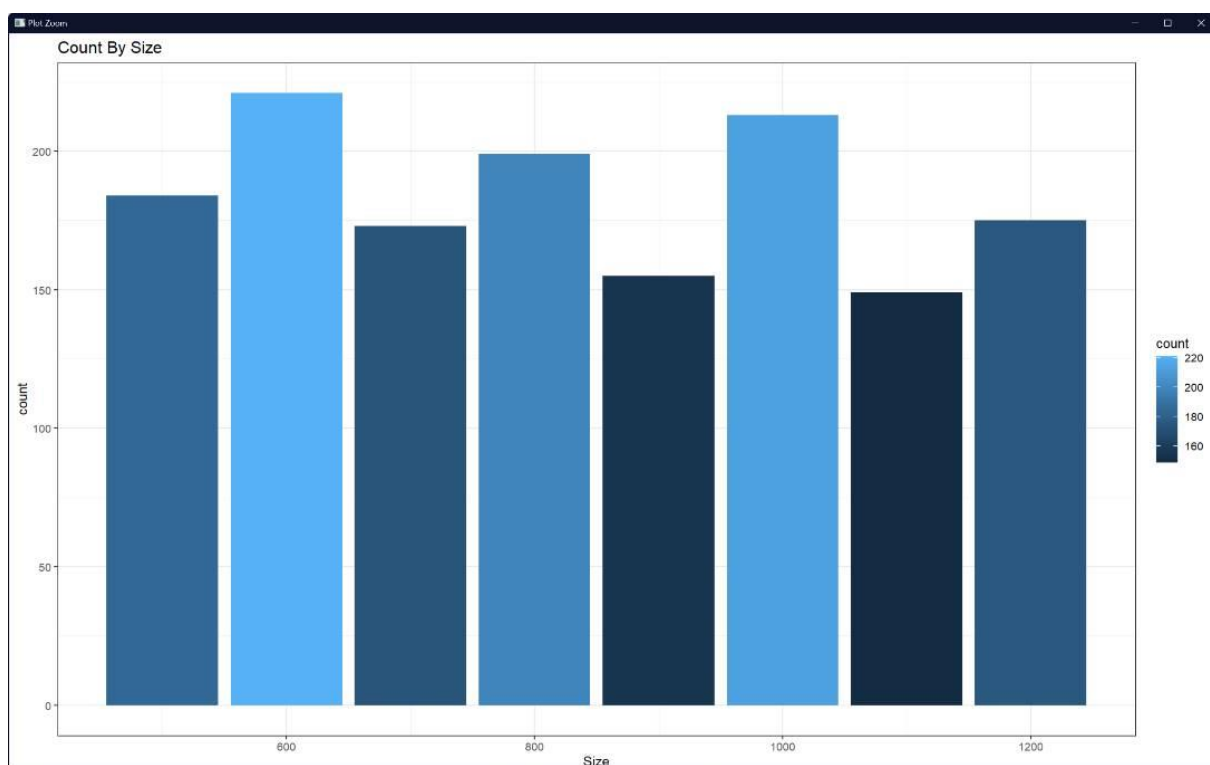


Figure 38: The figure displays that 600 sqft is the most popular house size

5.5 Q5 What are the cities which has the highest amounts of each category?

5.5.1 Analysis 5.1: Which city has the highest total amount of BHK per city?

```
# Q5 what are the cities witch has the highest amounts of each category?  
  
# Analysis 5.1  
# Q which city has the highest total amount of BHK per city?  
  
Total_Amount_BHK_Per_city<-data%>%group_by(City)%>%summarise(Total_BHK = sum(BHK))  
head(Total_Amount_BHK_Per_city)  
ggplot(Total_Amount_BHK_Per_city,  
       mapping = aes(x= City,  
                     y= Total_BHK,  
                     fill = Total_BHK)) +  
  geom_bar(stat = "identity", position = "dodge") +  
  labs(title = "Total Amount Of BHK Per City") + theme_bw()
```

Figure 39: A code to BMK amount per city

According to the graph, Chennai has the most amounts of BHK in their houses ahead of (Hyderabad) by a small margin while (Mumbai) has the least. This means that tenants who are looking for more BHK in their houses should start looking in Chennai for a suitable fit.

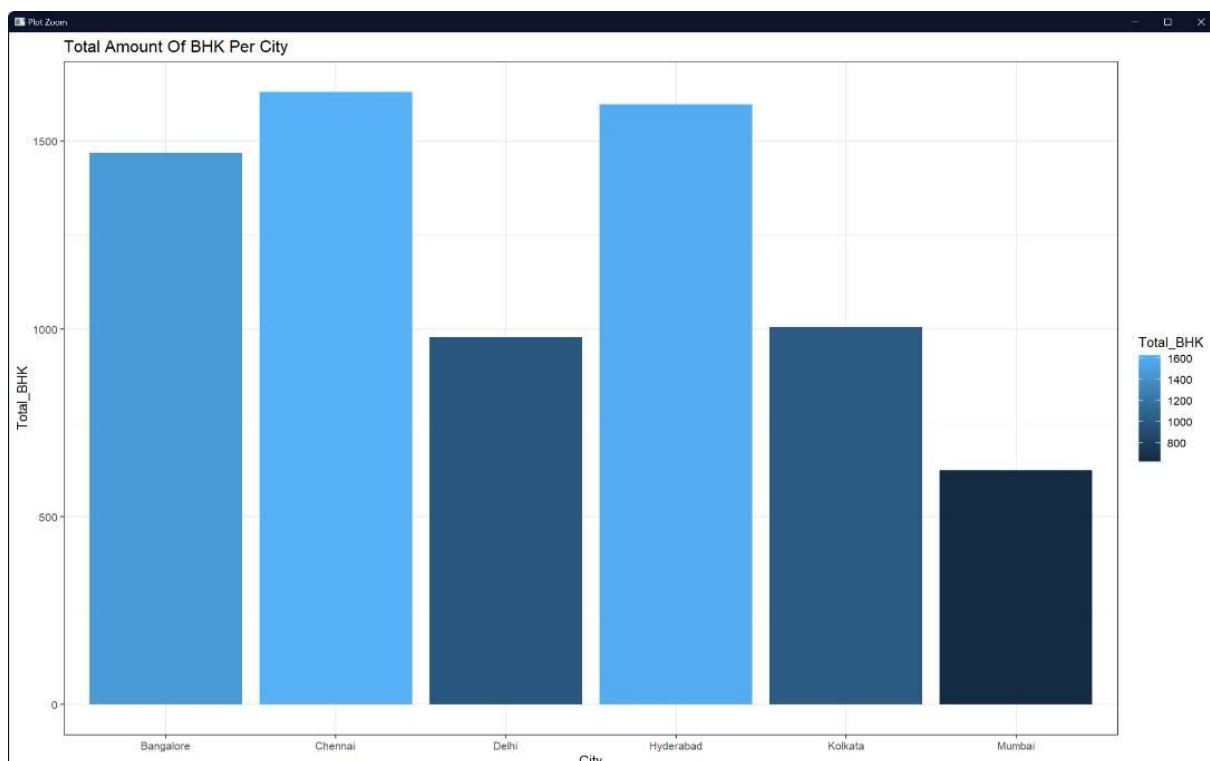


Figure 40: The graph shows that Chennai has the highest amount of BMK

5.5.2 Analysis 5.2: Which city has the highest total amount of each area type per city?

```
# Analysis 5.2
# Q which city has the highest total amount of each area type per city?

Amount_Areatype_Per_city<-data%>%group_by(Area.Type,City)%>%summarise(count = length(City))
head(Amount_Areatype_Per_city)
ggplot(Amount_Areatype_Per_city,
       mapping = aes(x= Area.Type,
                     y= count,
                     fill = City)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Count Of Area Type Per City") + theme_bw()
```

Figure 41: A code the available the count of available areas in each city.

Most dwellings in the dataset have a super area, as seen by the statistics below. In addition, we can observe that among the other cities, Hyderabad has the greatest number of super-sized homes, while Mumbai has the least. In contrast, Mumbai has more residences than any other city in the dataset in terms of carpet area. This dataset would assist renters in determining which city to search for a home in case they want a property with a carpet or super space. Moreover, the (built area) count is behind the other areas by a big margin.

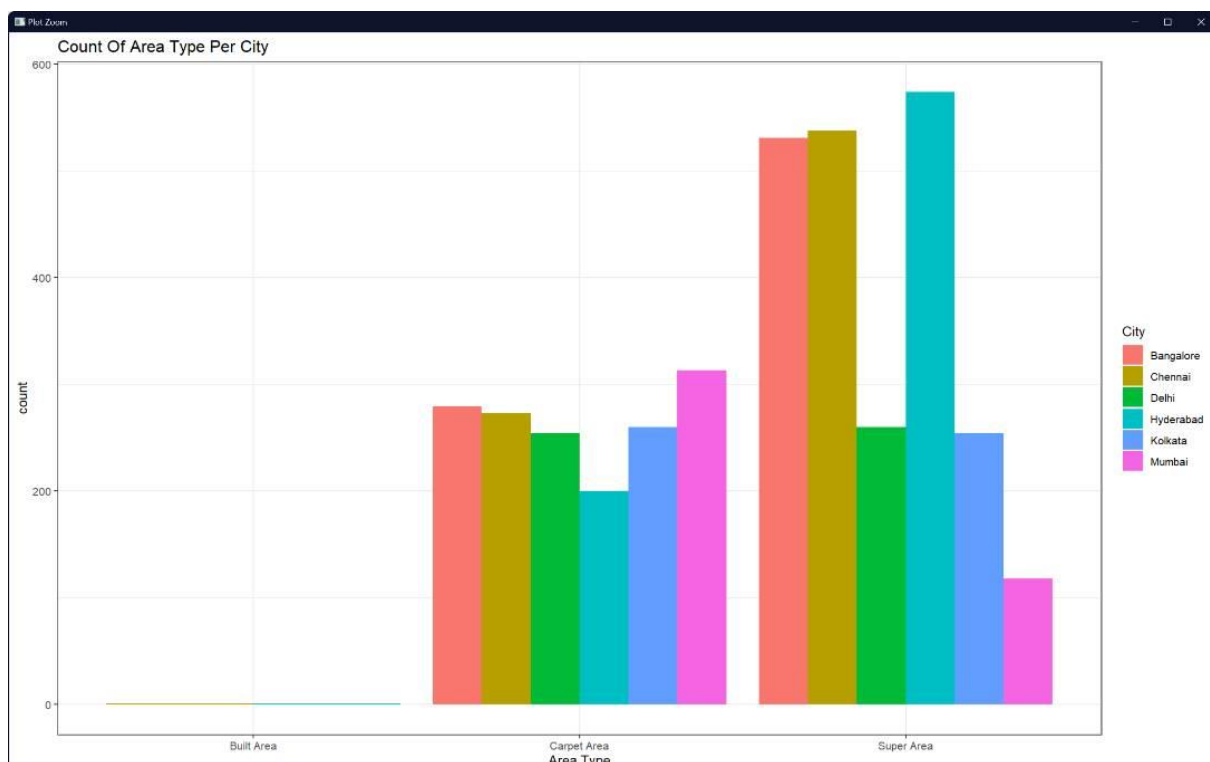


Figure 42: The figure displays the area count in each city

5.5.3 Analysis 5.3: Which city has the highest total amount of rent per city?

```
# Analysis 5.3
# Q which city has the highest total amount of rent per city?

Total_Amount_Rent_Per_city<-data%>%group_by(City)%>%summarise(Total_Rent = sum(Rent))
head(Total_Amount_Rent_Per_city)
ggplot(Total_Amount_Rent_Per_city,
       mapping = aes(x= City,
                     y= Total_Rent,
                     fill = Total_Rent)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total amount Of rent per City") + theme_bw()
```

Figure 43: A code that shows the total rent price paid in each city

In Figure [44], we can see that the total rent in Mumbai is not significantly higher than the rest of the cities especially(Chennai). With this data in mind, we can assume that Mumbai is the most expensive city when it comes to living expenses and Delhi has the lowest.

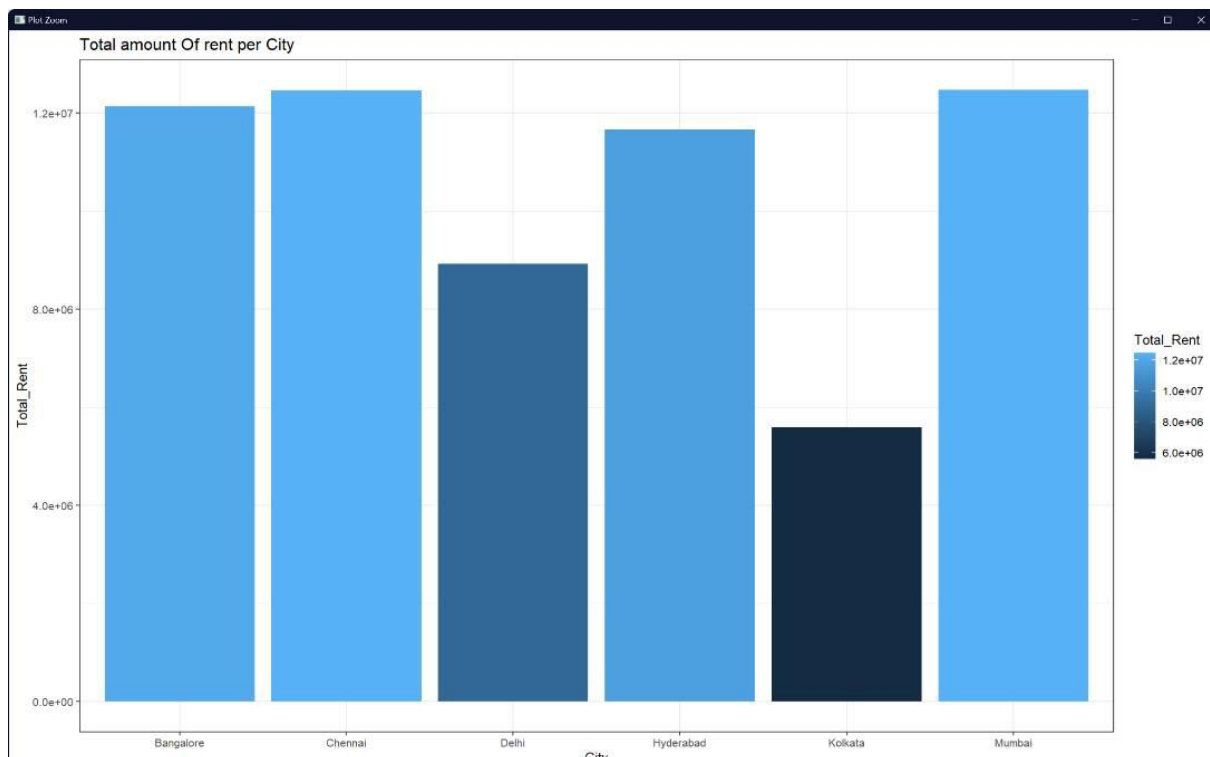


Figure 44: The graph shows that (Chennai & Mumbai) are close to each other for the total rent paid.

5.5.4 Analysis 5.4: Which city has the highest total amount of furnishing status per city?

```
# Analysis 5.4
# Q which city has the highest total amount of each furnishing status per city?

Total_Furnishing_Per_city<-data%>%group_by(Furnishing.Status, City)%>%summarise(count = length(City))
head(Total_Furnishing_Per_city)
ggplot(Total_Furnishing_Per_city,
       mapping = aes(x= Furnishing.Status,
                     y= count,
                     fill = city)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Amount Of Each furnishing Status Per City") + theme_bw()
```

Figure 45: A code that shows the most popular furnishing status per each city.

According to Figure [46], most dwellings in the dataset are just semi furnished. In addition, Bangalore has the greatest number of semi-furnished homes, while Kolkata has the fewest. Because Chennai has the most unfurnished properties, prospective renters seeking unfurnished housing should begin their search in Chennai. On the other hand, renters searching for furnished residences should select Hyderabad, since it has the most furnished houses compared to Kolkata, which has the least. This graph makes it easy for renters seeking a certain area type to choose the city in which they want to reside.

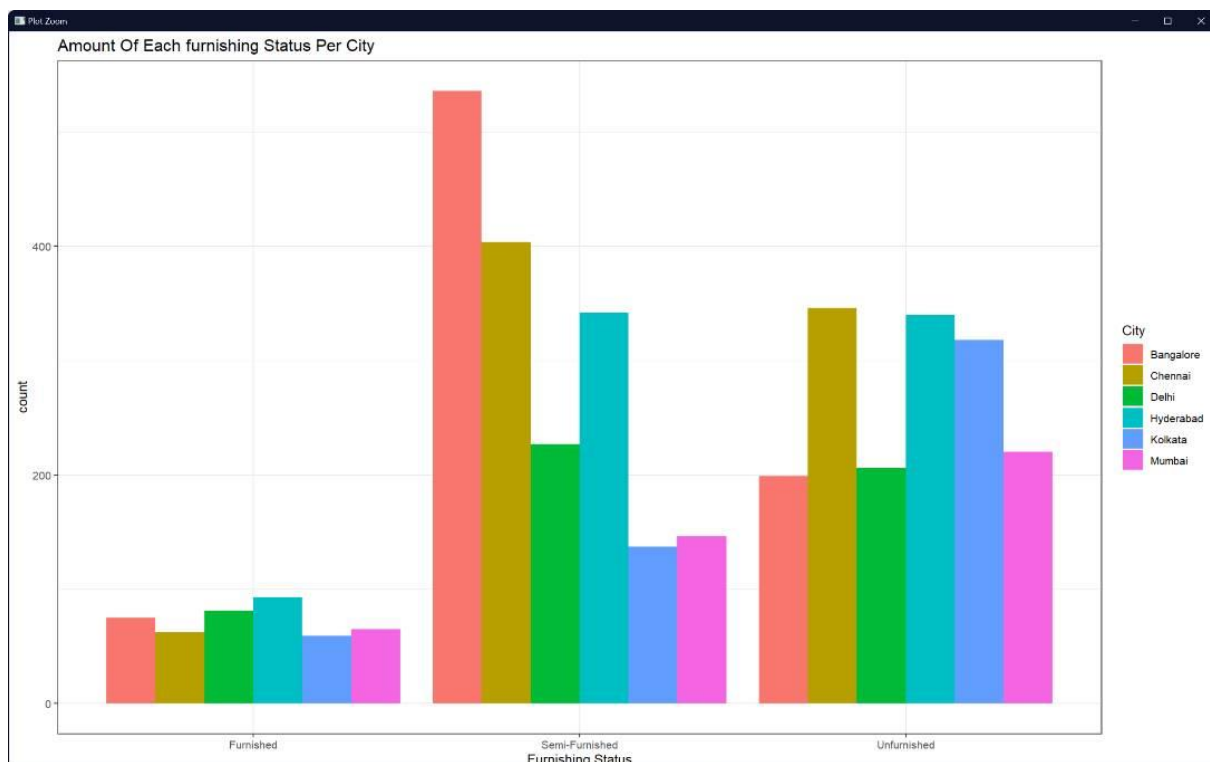


Figure 46: The graph shows that the most popular furnishing status is (Semi-Furnished)

6 Additional Features

Figure 47 contains a correlogram matrix graph that illustrates the relationships between the most important aspects of the dataset, including the number of bedrooms, square footage, and bathrooms. As can be seen in the graph that is located above, the proportion rises with the complexity of the elements that are being considered. This graph allows us to see how each element interacts with one another and gives us a better picture of how our dataset will perform as a result. As a result of this, we can see how our dataset will perform.

6.1 Additional Features-1

```
#Additional feature 1 Create Correlogram Matrix  
Correlogram_Matrix<-cor(data[,c(2,3,4,11)])  
head(Correlogram_Matrix)  
corrplot(Correlogram_Matrix,addCoef.col = TRUE)
```

Figure 47:A code that generates correlogram Matrix

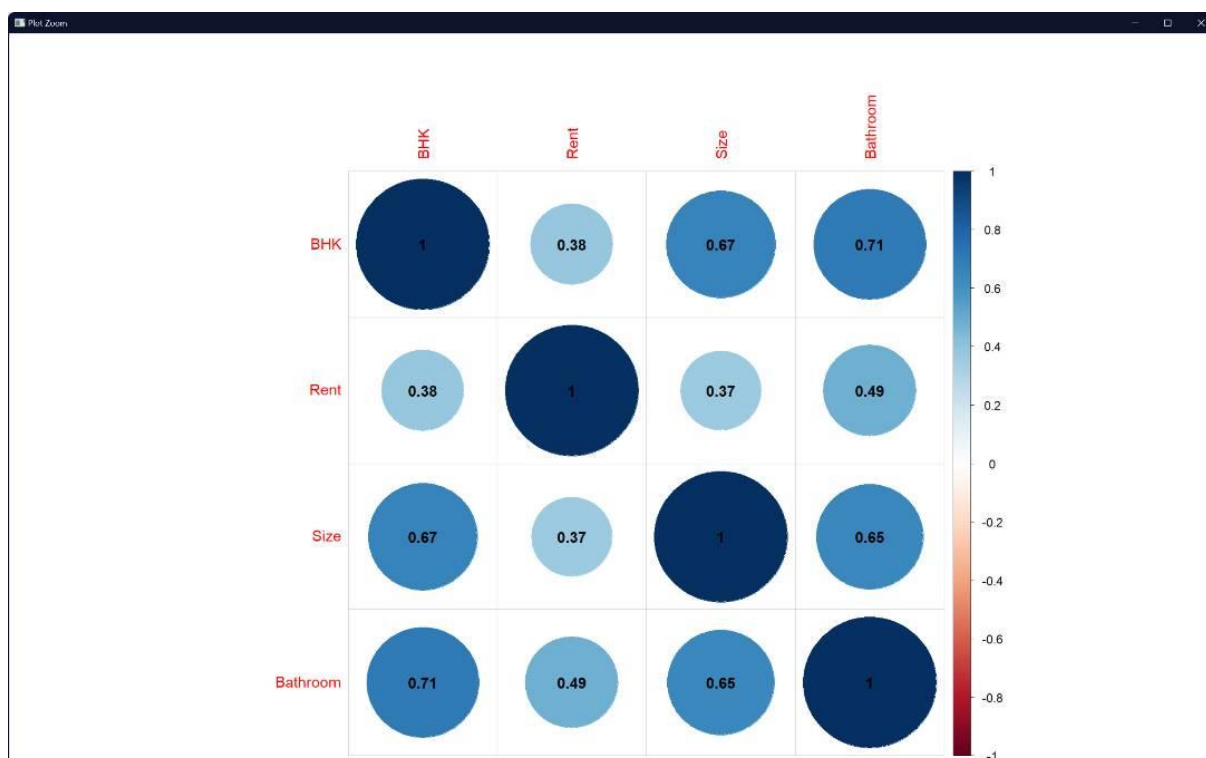


Figure 48:

6.2 Additional Features-2

Figure [49] displays the scatterplot graph that is produced by the functions "abline" and "plot" regarding the regression that examines the relationship between rent and the size of the dwellings in the dataset. We can see that Edline creates a blue line demonstrating the increasing lease amount as the aspect develops. This indicates that house size unquestionably has a role in determining house leases.

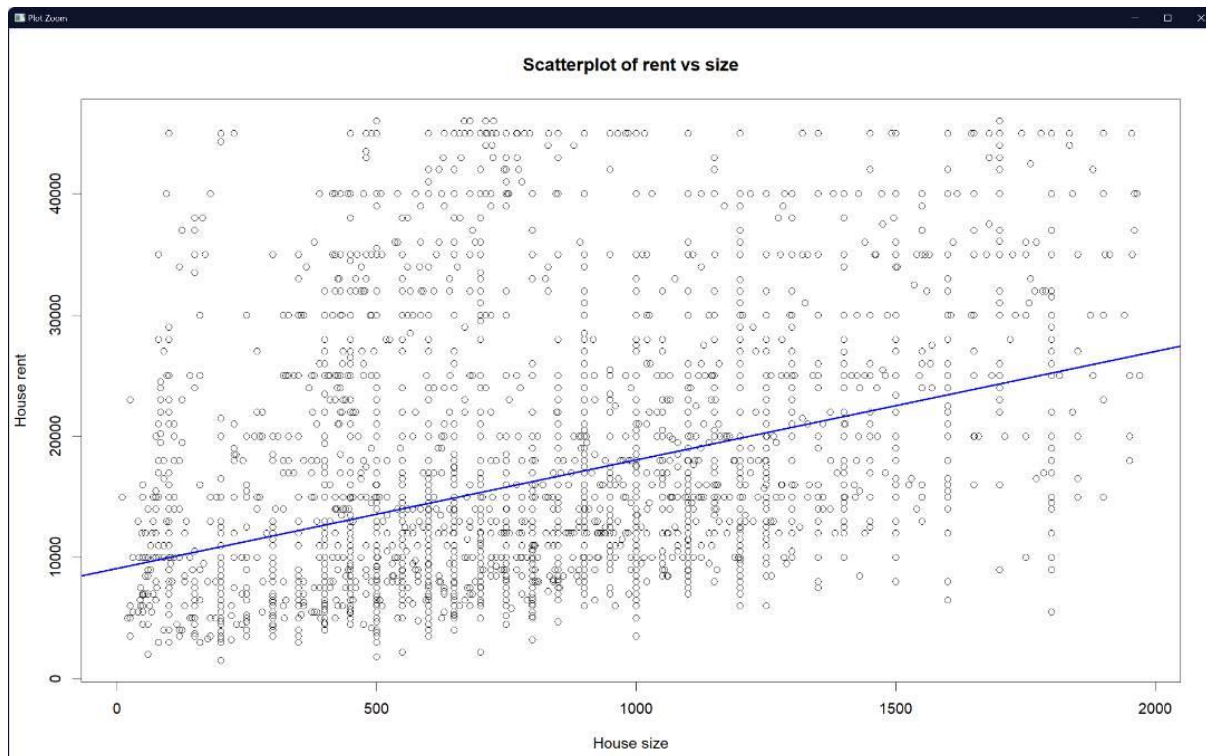


Figure 49: Scatter plot to show the connection between rent & size

6.3 Additional Feature-3

```
#Additional feature 3 create a violin graph for rent by size  
rent_size <- ggplot (data, aes (x = Rent, y = Size)) + geom_violin()  
head(rent_size)  
ggplot (data, aes (x = Rent, y = Size)) + geom_violin(trim = FALSE)
```

Figure 50: A code that generates a violin graph with the connection between Rent & Size

Figure [50] displays a violin graph that was generated by the `geom_violin()` function. On this graph, you can observe the range of rent prices for each different size of house. We can see that the scope of the lease has the greatest reach when the size of the house is approximately 2,000 square feet, whereas the scope of the lease has a more limited reach when the house is larger. Because of this, there is a greater range of possible rent prices for homes with smaller floor plans.

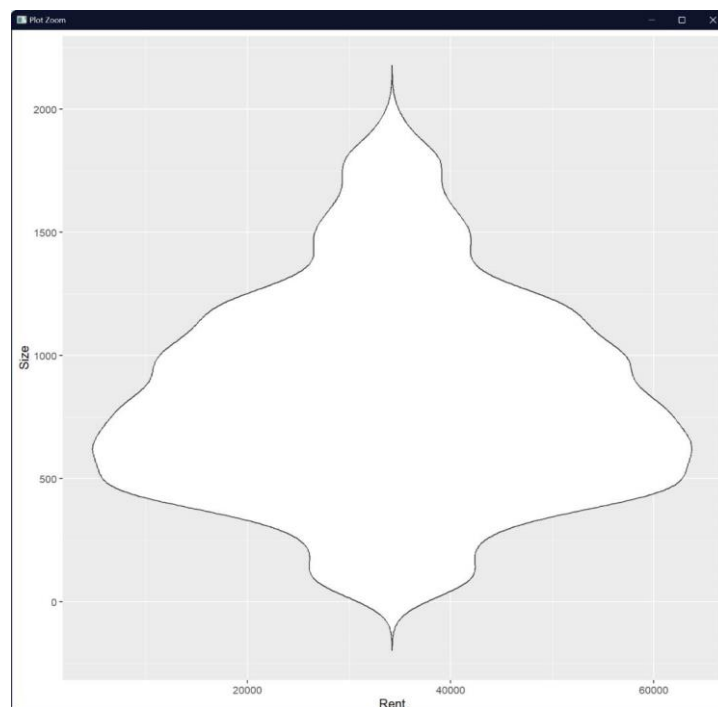


Figure 51: A violin graph that shows the relationship between rent and size

7 Conclusion:

We were able to complete this task and gain some new insights after organizing, cleaning, and preparing the data for analysis. Our investigation led us to the conclusion that, in comparison to other cities, Hyderabad has the largest average home size and the lowest average rent. In this scenario, the Built Area is the least expensive while the Carpet Region is the most expensive.

A home's monthly rent varies depending on its size, and agency homes are significantly more expensive than owner and builder homes. The average rent for a home in Mumbai is significantly higher than in any other metropolis.

We also found that there is a slight proportionality, and it can be explained by the fact that Mumbai is the most costly location, and as a direct result, it has the lowest population

density, both in terms of immigrants and the overall population. After performing a large number of studies, we have arrived at the conclusion that certain cities may have higher costs, but this is likely due to the fact that residents do not particularly enjoy living there.

We also found that there is a slight proportionality, and it can be explained by the fact that Mumbai is the most costly location, and as a direct result, it has the lowest population density, both in terms of immigrants and the overall population. In terms of rent, the most sought-after locations are ones that come partially furnished.

Because communication with the agent is comparable to 1529 and communication with the owner is equivalent to 3216, it would appear that direct communication with the owner is the most valuable form of communication. This indicates that the combined rents paid by persons who are in direct contact with the King account for close to 67% of the total.

In addition to this, it was found that out of the cities that were stated earlier, Chennai has the greatest population in the region.

8 References:

1. *How to Remove Outliers in R*. (2021, September 27). R-bloggers. Retrieved December 3, 2022, from <https://www.r-bloggers.com/2021/09/how-to-remove-outliers-in-r-3/>
2. Coder, R. (2021c, March 24). *Box plot by group in ggplot2*. R CHARTS | a Collection of Charts and Graphs Made with the R Programming Language. Retrieved December 3, 2022, from <https://r-charts.com/distribution/box-plot-group-ggplot2/>
3. *ggplot2 scatter plots: Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA*. (n.d.). www.sthda.com. Retrieved November 29, 2022, from <http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization>
4. Chon, W. (2021, December 15). *8 Tips for Better Data Visualization - Towards Data Science*. Medium. Retrieved November 25, 2022, from <https://towardsdatascience.com/8-tips-for-better-data-visualization-2f7118e8a9f4>
5. GeeksforGeeks. (2021b, February 6). *R Tutorial*. Retrieved October 5, 2022, from <https://www.geeksforgeeks.org/r-tutorial/>
6. *Comprehensive Guide to Data Visualization in R*. (2020, July 12). Analytics Vidhya. Retrieved November 15, 2022, from <https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/>
7. *Tidyverse packages*. (n.d.). TidyVerse. Retrieved November 30, 2022, from <https://www.tidyverse.org/packages/>
8. *Predicting House Prices using R*. (2017, September 3). Kaggle. Retrieved November 20, 2022, from <https://www.kaggle.com/code/pradeeptripathi/predicting-house-prices-using-r>
9. *Example R code / analysis for housing data*. (n.d.). Retrieved November 28, 2022, from https://pages.pomona.edu/%7Ejsh04747/courses/math58/Final_examp.html
10. *How to check data type in R -*. (n.d.). ProjectPro. Retrieved November 30, 2022, from <https://www.projectpro.io/recipes/check-data-type->
11. *How to check data type in R -*. (n.d.-b). ProjectPro. Retrieved November 20, 2022, from <https://www.projectpro.io/recipes/check-data-type-r>

12. *Data Types and Structures – Programming with R*. (n.d.). Software Carpentry - Programming With R. Retrieved December 1, 2022, from <https://swcarpentry.github.io/r-novice-inflammation/13-supp-data-structures/>

13.