GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APD2F2305CS(DA), APU2F2305CS(CYB)

HAND OUT DATE: **06 JULY 2023**

HAND IN DATE: **01 SEPTEMBER 2023**

LECTURER NAME: **MS. MINNU HELEN JOSEPH**

GROUP MEMBER:

STUDENT ID	STUDENT NAME
TP068687	SOO JIUN GUAN
TP068977	SOONG YAU JOE
TP069429	TEH YUE FENG
TP062972	LIAN JUN ER

Table of Contents

1 Introduction

This assignment's primary goal was to analyse the dataset of predicted house rental rates. The dataset contains details on houses, apartments, and flats that are available for rent, including variables like BHK, rent, size, number of floors, area type, area locality, city, furnishing status, preferred tenant type, number of bathrooms, and point of contact. There are 12 columns and more than 4700 rows in it. It is necessary to clean the data using pre-processing, manipulation, transformation, and visualisation techniques.

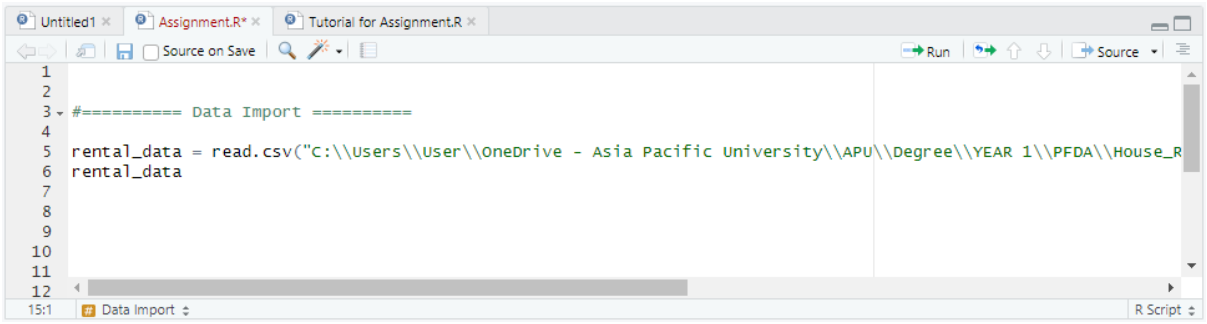
2 Assumption

There are many of assumption have been made for the data set that given such as rental properties posted by agents have a higher average rent compared to properties posted directly by owners. In addition, semi-furnished rental properties have a higher demand among tenants compared to furnished or unfurnished properties etc. Lastly, an assumption of rental properties that only preferred bachelor tenants posted by agents have a higher average rent compared to properties posted directly by owners in Mumbai is agreed as the group's final assumption.

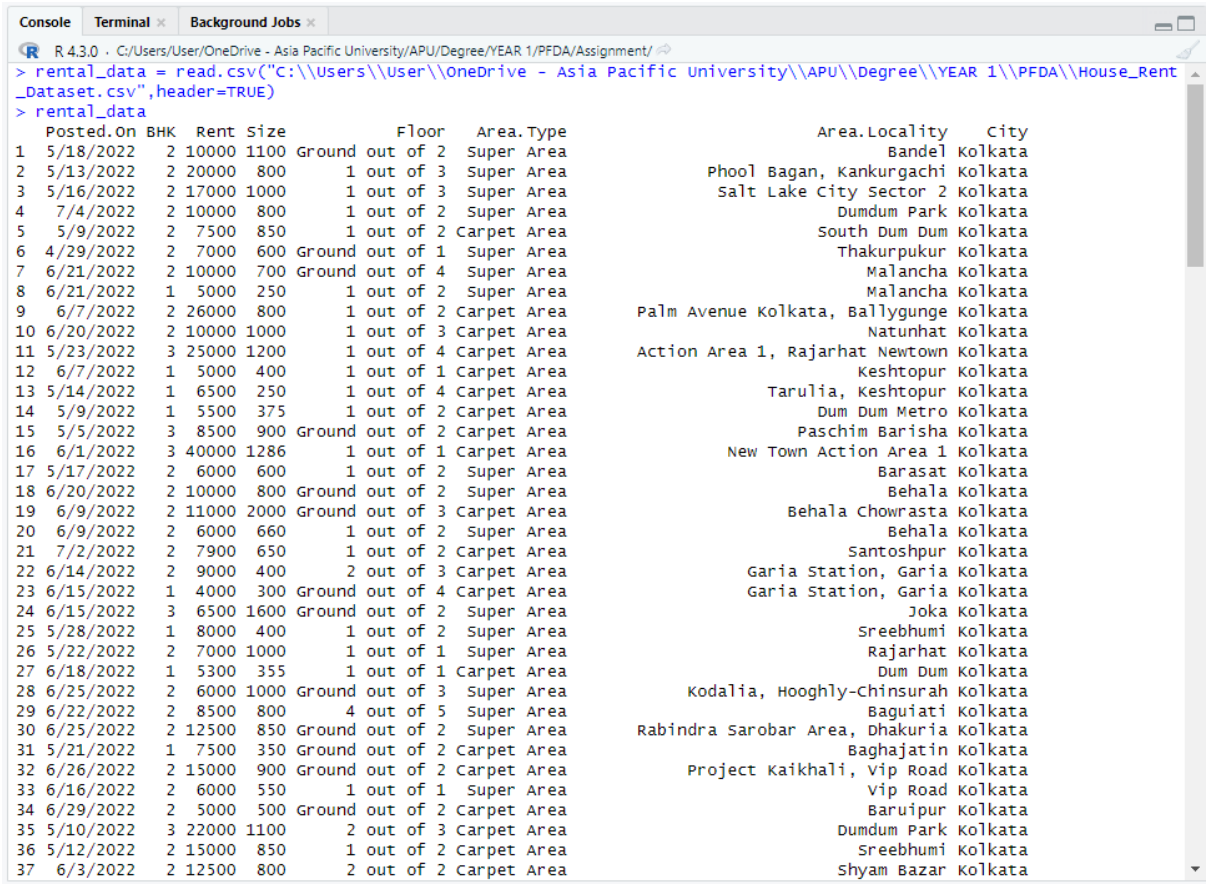
3 Data Process

3.1 Data Import

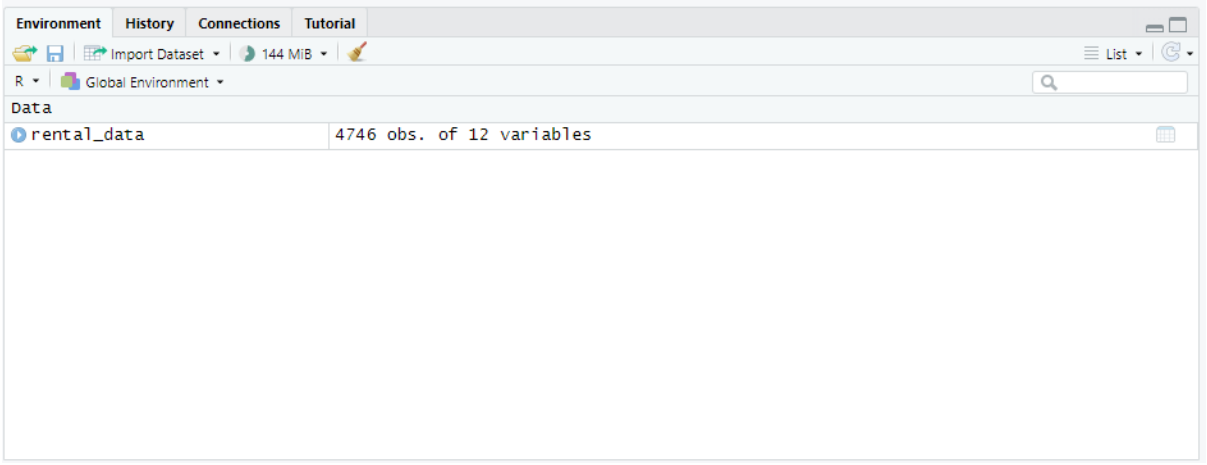
Before do any data analysis in R studio, the first step is to import all the data set that given into the R studio. The data set that given is in CSV file format and named House_Rent_Dataset.csv. In the R script, “read.csv” is used to import data from a csv file format. In the bracket after the “read.csv” is the directory for the data set file and the “header=TRUE” is means that the first row is the header row.



rental_data is being called in the R script and it will list out the data in the console. The output might not list out all the data due to the maximum reached of data.



The data set are imported into the R studio and defined as rental_data. There is 4746 rows and 12 columns in the data set.



3.2 Data Pre-Processing

```
#==========DATA CLEANING==========

#Remove Missing Values
omit_data <- na.omit(rental_data)

#Remove Duplicated Rows
rental_data <- unique(rental_data)

#Check Data
unique(rental_data$Area.Type)
unique(rental_data$Tenant.Preferred)
unique(rental_data$Furnishing.Status)
unique(rental_data$Point.of.Contact)


#REMOVE SPECIAL ROWS
View(filter(rental_data, rental_data$Area.Type == "Built Area"))
View(filter(rental_data, rental_data$Point.of.Contact == "Contact Builder"))
rental_data2 <- filter(rental_data, Area.Type != "Built Area"
                       & Point.of.Contact != "Contact Builder")

# REMOVE OUTLIERS
sort(unique(rental_data$Rent), decreasing=TRUE) # check is there any outliers
rental_data2 <- filter(rental_data, Rent < 1000000)
```

6:42 IMPORT DATASET R Script

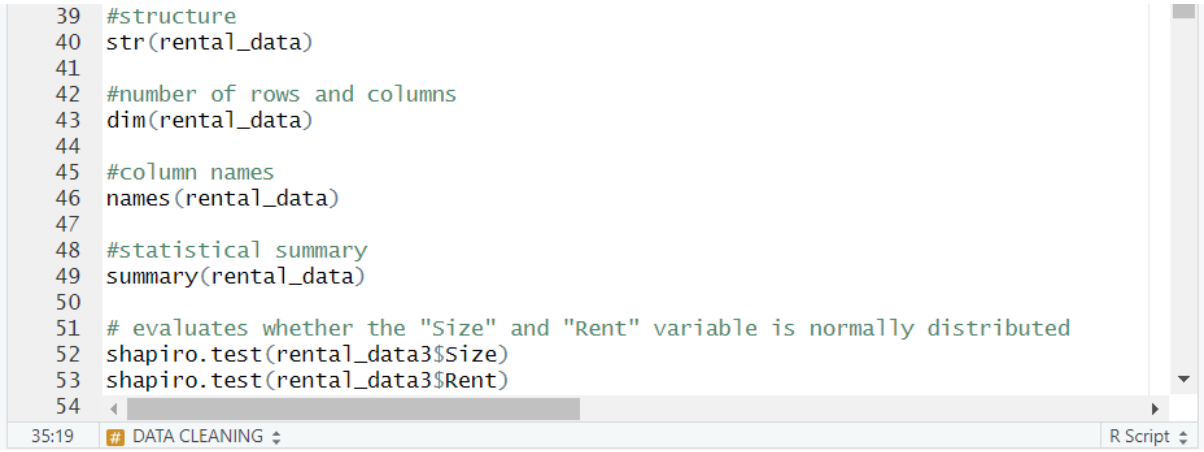
After importing the dataset, there is another important process needs to be conducted, which is data cleaning. Firstly, We use the na.omit() function to remove any rows with missing values from the rental_data. To further refine the dataset, the unique() function is used to help us in removing any duplicated rows in the data. We also use unique() function to examine certain columns within the rental_data. Additionally, we use the filter() function to create a subset that eliminates rows where the Area.Type is "Built Area" and the Point.of.Contact is "Contact Builder". Moving forward, we check for the outliers in the Rent column by using the sort() function that is able to offer a descending arrangement of the rent values. Finally, we use the filter() function to generate a new dataset named rental_data2 that excludes rows where Rent exceeds or is equal to 1,000,000.

rental_data	4746 obs. of 12 variables
rental_data2	4740 obs. of 12 variables

As indicated in the figure above, the rows of the dataset changed from the original 4,746 to 4,740 after data cleaning.

3.3 Data Exploration

Data exploration, one of the data preparations before conducting an in-depth analysis of the dataset. It is a way to let us have a deeper and more comprehensive understanding for the information of the dataset before working with it. An effective exploration assists in recognizing and honing upcoming analytical inquiries and challenges.



At first, we use the str() function to display the structure of rental_data. This function is able to show information about the columns and their data types. We also use dim() function to know the dimensions of rental_data, which include the number of rows and columns. To know the names of the each column, we use the names() function. Furthermore, we use the summary() function to product a statistical summary for each column in rental_data. And finally, we use the shapiro.test() to check whether the variables are normally distributed.

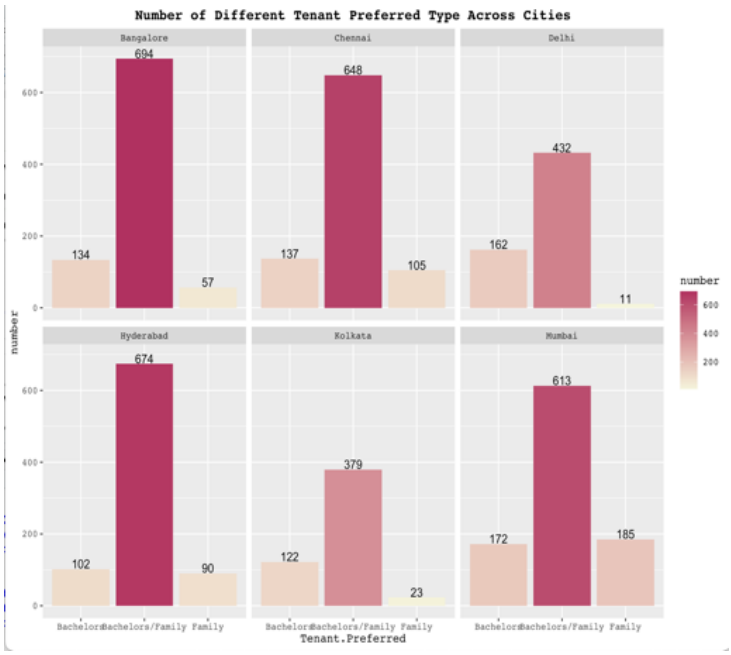
4 Analysis

4.1 Objective 1 (Soong Yau Joe)

To discover the relationship between the tenant preferred by the landlord and the rent of the properties across different cities.

[Please refer to appendix for the code snippets for the analysis]

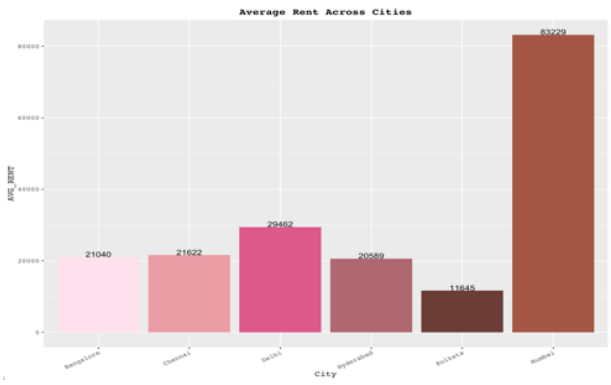
4.1.1 Analysis 1: Which types of tenants are most preferred by the lessor in every city?

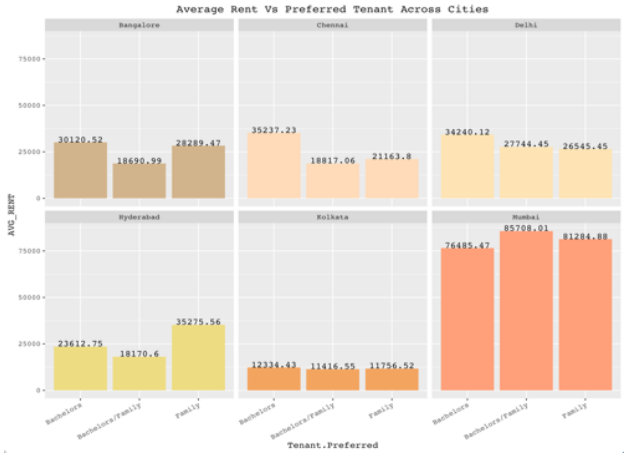


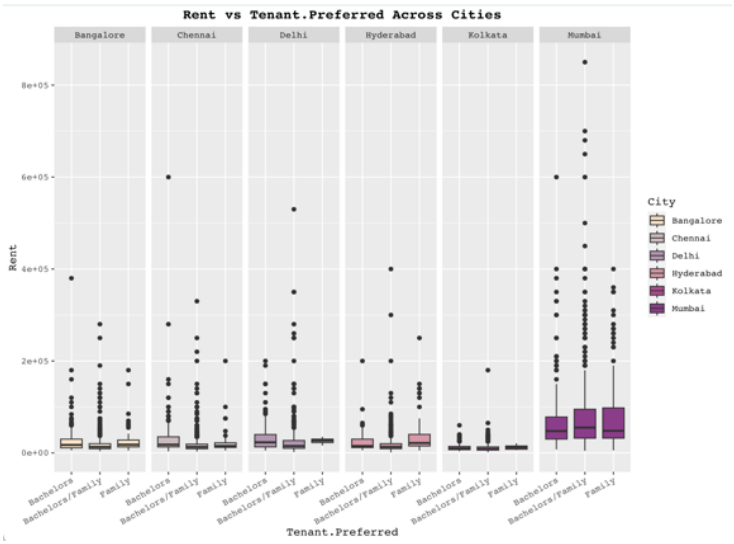

```
[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Kolkata "
[1] "Bachelors/Family --- 72.33"
[1] "Bachelors --- 23.28"
[1] "Family --- 4.39"

[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Mumbai "
[1] "Bachelors/Family --- 63.2"
[1] "Bachelors --- 17.73"
[1] "Family --- 19.07"

[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Bangalore "
[1] "Bachelors/Family --- 78.42"
[1] "Bachelors --- 15.14"
[1] "Family --- 6.44"

[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Delhi "
[1] "Bachelors/Family --- 71.4"
[1] "Bachelors --- 26.78"
[1] "Family --- 1.82"

[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Chennai "
[1] "Bachelors/Family --- 72.81"
[1] "Bachelors --- 15.39"
[1] "Family --- 11.8"

[1] "PERCENTAGE OF EVERY TENANT PREFERRED IN Hyderabad "
[1] "Bachelors/Family --- 77.83"
[1] "Bachelors --- 11.78"
[1] "Family --- 10.39"
```

From the bar chart of the frequency of different type of tenant across the city, the graph is showing us the frequency which the type of preferred tenant is both will always the highest in every city. Next, bachelor type of preferred tenant is taking the second high frequency in most of the city except in Mumbai, but the difference between the second high frequency and the frequency of family type of tenant is relatively small. Furthermore, it is worth mentioning that Mumbai city has the highest frequency of preferring tenant in Family type compared to other cities. Besides, Delhi and Kolkata have a relatively low amount of house rental that is targeting family type of tenant which is only 1.82% and 4.39%

4.1.2 Analysis 2: How does tenant preferred by the lessor affect the rent across different cities?





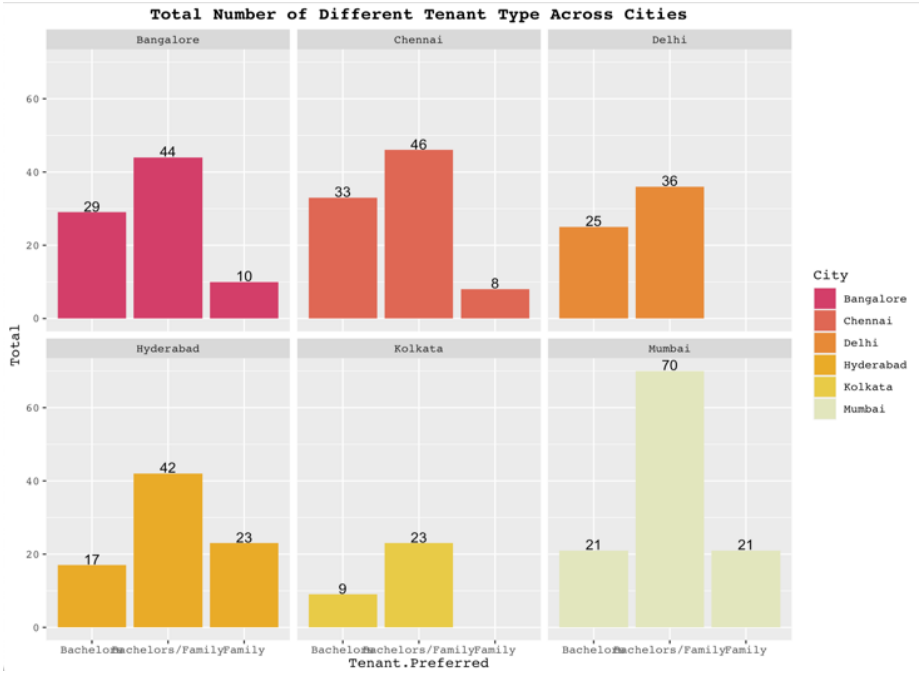



```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  wilcoxon_pairtest_tenant_rent_mumbai$Rent and wilcoxon_pairtest_tenant_rent_mumbai$Tenant.Preferred

                   Bachelors Bachelors/Family
Bachelors/Family 0.41      -
Family           0.94      0.94

P value adjustment method: holm
```

```
[1] "City: Kolkata"
[1] 0.007410566

[1] "City: Mumbai"
[1] 0.3440198

[1] "City: Bangalore"
[1] 3.946855e-08

[1] "City: Delhi"
[1] 6.404913e-08

[1] "City: Chennai"
[1] 7.063419e-11

[1] "City: Hyderabad"
[1] 5.184926e-11
```

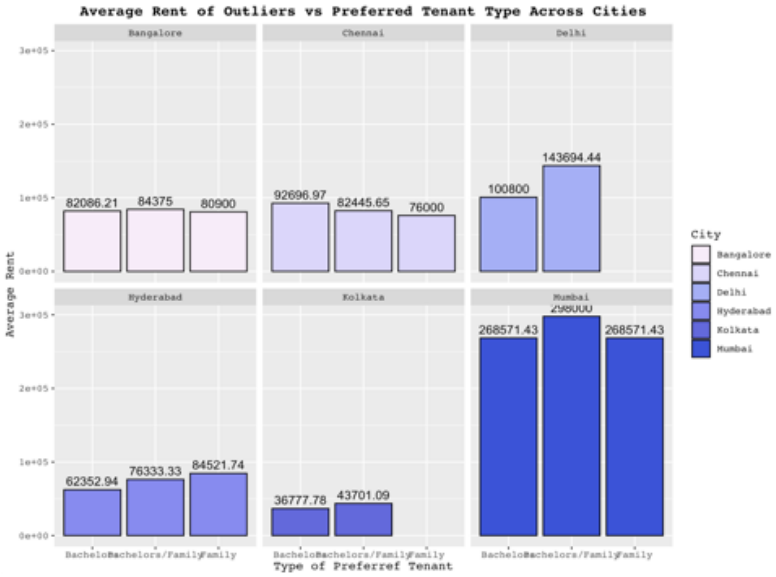
The mean rent in Mumbai is the highest while for Kolkata is the lowest. Speaking about the rental fees, the average rent of preferring Bachelor s Tenant is the highest in majority of the cities except Mumbai and Hyderabad. Next, the average rent for all types of tenants preferred is the highest in Mumbai city compared to other cities. When it comes to determine whether there is a difference between different group of tenants preferred across cities, by assuming the critical p-value for both Kruskal-Wallis Test and Wilcoxon Rank Sum Test is 0.05, it can be known that there is no significant difference between all combination of types of tenants preferred in terms of rent only in Mumbai city. Hence, we can conclude that types of tenants preferred does will not become one of the factors that affect rental price in Mumbai city.

4.1.3 Analysis 3: Which types of preferred tenant does the outliers mostly belongs to?



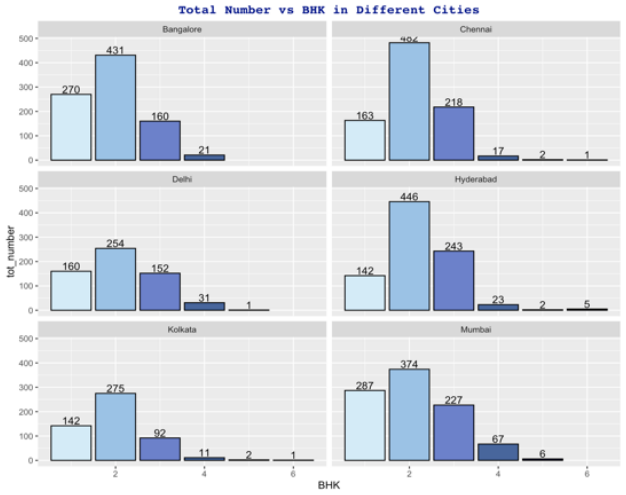
In the context of outliers in terms of rent for every type of tenant preferred across different cities, the number of preferring tenants of Bachelor/Family type is always the highest compared to other types of tenants preferred in every cities with Mumbai city having the highest number of Bachelor/Family type of preferred tenant compared to other cities.

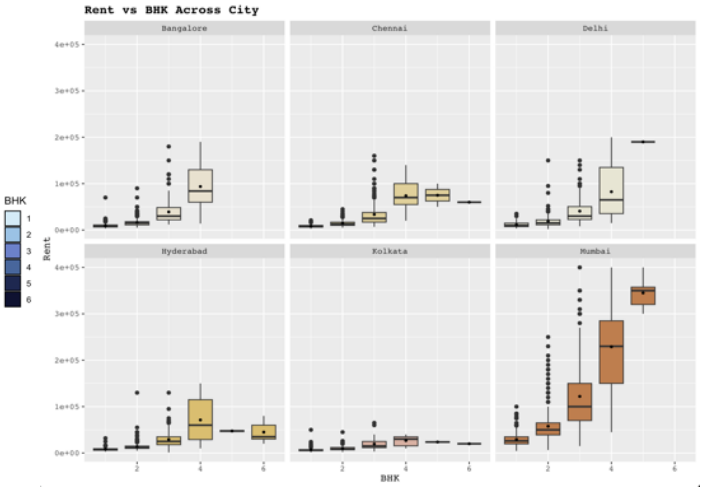
4.1.4 Analysis 4: What is the average rent of those outliers based on the preferred tenant?

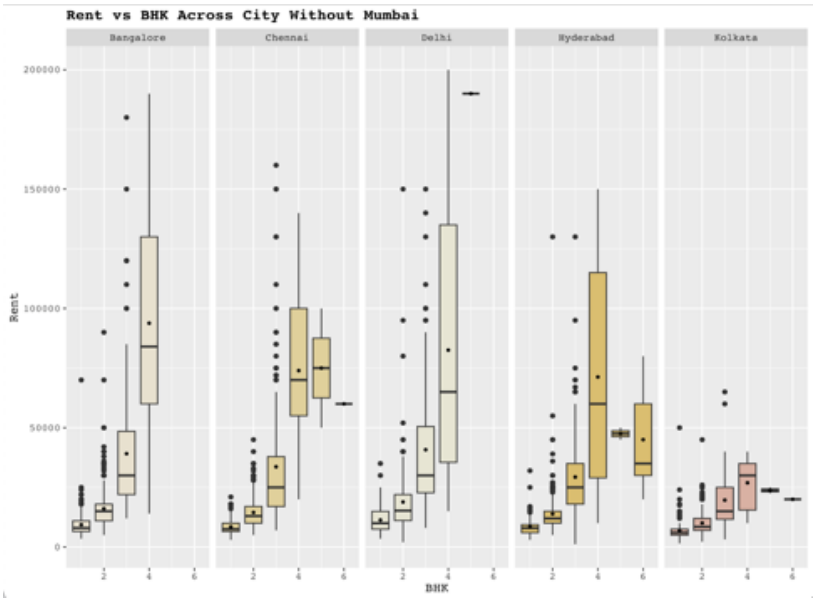


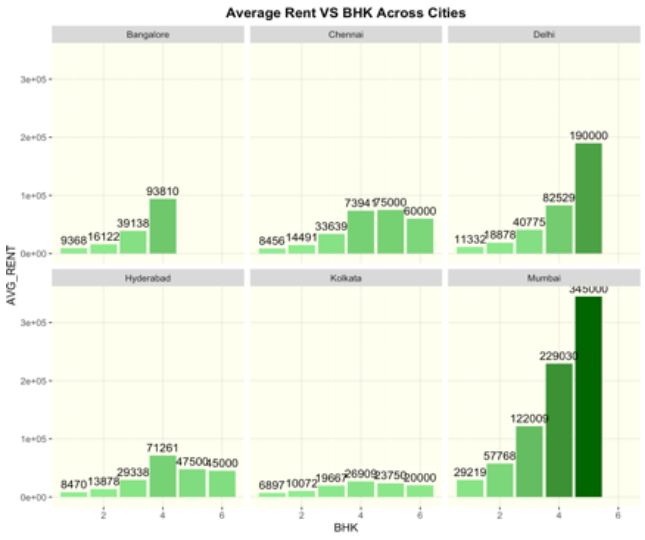
By discovering the average rent of different types of preferred tenants across cities, Mumbai has the highest rent for every type of tenant preferred when compared to other cities with a minimum average rent of approximately 270000 rupees to a maximum average rent to 298000 rupees.

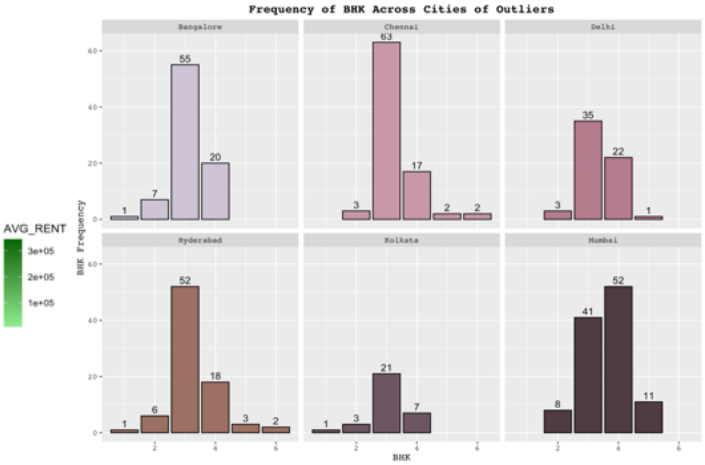
4.1.5 Analysis 5: What is the relation between BHK and rental fees across different cities?

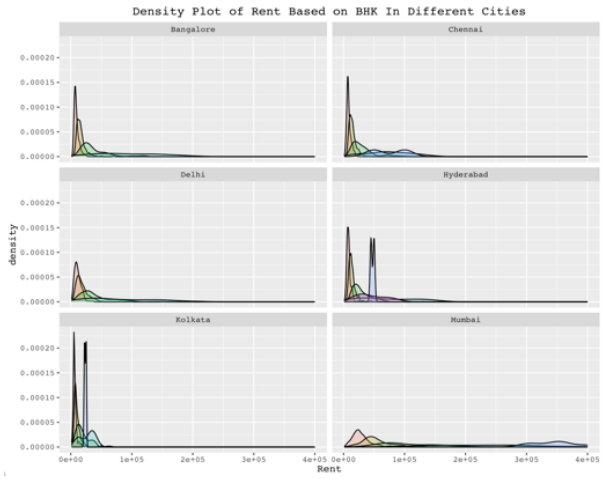


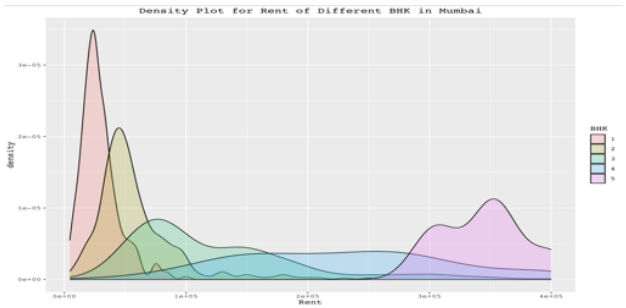


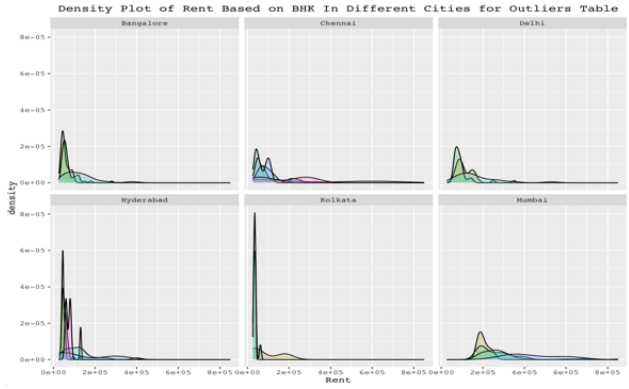


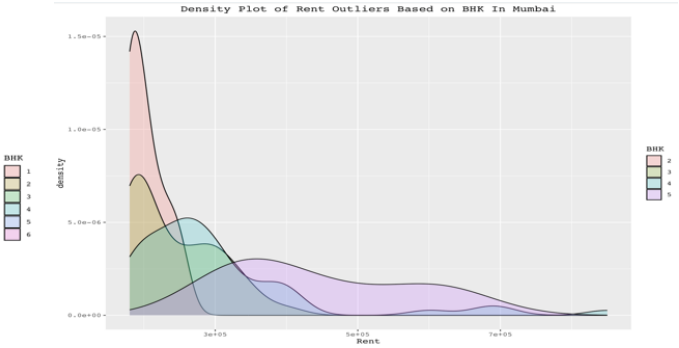










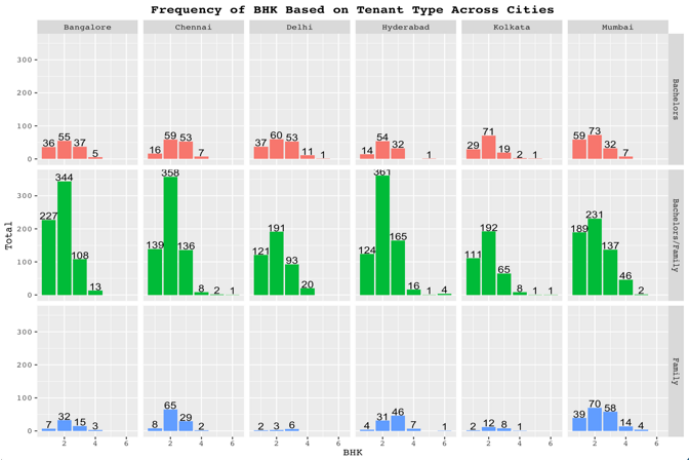



```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  kw_test_bhk_rent_mumbai$Rent and kw_test_bhk_rent_mumbai$BHK

  1       2       3       4
2 <2e-16  -       -       -
3 <2e-16 <2e-16  -       -
4 <2e-16 <2e-16 <2e-16  -
5 0.0001 0.0001 0.0001 0.0019

P value adjustment method: holm
```

```
        Kruskal-Wallis rank sum test

data:  Rent by BHK
Kruskal-Wallis chi-squared = 589.81, df = 4, p-value < 2.2e-16
```

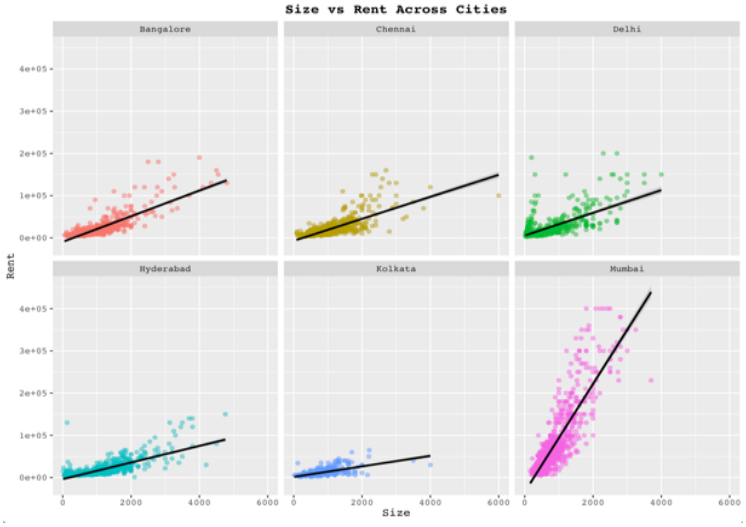
It can be observed that there is a significant total frequency from 1 BHK to 4 BHK where 2 BHK being the most popular in every city whereas the frequency of 5 BHK and 6 BHK is relatively low which means it is rare to find them in any of the cities. From the boxplot, it shows there is an increasing trend of mean rent as the BHK increase from 1 to 4 but it does not hold after 4 BHK for some cities. Next, it is important to notice that majority of the outliers have a 4 BHK in Mumbai city and it is 52 out of 67 in number while the rest of the cities are having majority of their outliers on 3 BHK. Furthermore, the density plot tells us that with a budget of 100000 rupees, tenants will have a high chance to find houses ranging from 1 BHK to 3 BHK in almost every city. The density plot of Mumbai also tells us that it is only possible to rent a 5 BHK house in Mumbai with approximately 337500 rupees budget and most of the outliers start from a minimum 250000 rupees. Based on Pairwise Comparison Wilcoxon Rank Sum Test, it shows that there is significant difference between the number of BHK and rent. This concludes there is a positive relationship between rent and number of BHK in Mumbai.

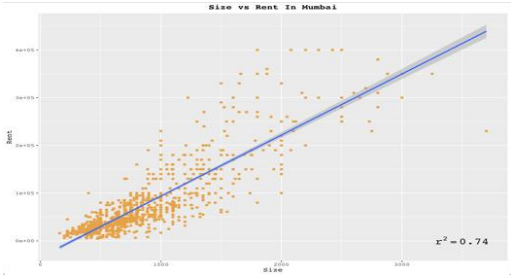
4.1.6 Analysis 6: What number of BHK occurs the most in every type of tenant across different cities?



2 BHK is still being the most popular BHK in every city with the tenant preference of Bachelors/Family, and Bachelors while is an exception happened in Hyderabad, which is 3 BHK being the most popular BHK in the aspect of preferring family tenant.

4.1.7 Analysis 7: What are the relationship between size and Rent across different cities?

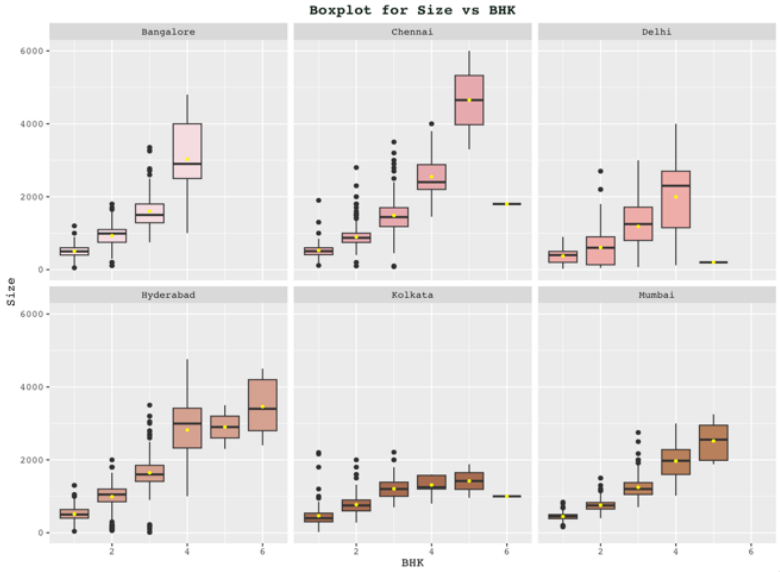




	City	Coeff_Determination
1	Kolkata	0.39
2	Mumbai	0.74
3	Bangalore	0.67
4	Delhi	0.41
5	Chennai	0.55
6	Hyderabad	0.54
7	Kolkata	0.39
8	Mumbai	0.74
9	Bangalore	0.67
10	Delhi	0.41
11	Chennai	0.55
12	Hyderabad	0.54

From the scatter plot with regression line, it shows Mumbai has the greatest rent per unit square foot compared to other city. Speaking about the relationship between rent and size in different cities, the coefficient of determination between rent and size is having a value of at least 0.50 in majority of the cities in which Mumbai is having the highest value 0.74 and this indicates that there are 74% of rent is being affect by the size. Therefore, we can say that size is one of the factors that affect rental fees in Mumbai.

4.1.8 Analysis 8: What is the relationship between Size and BHK?

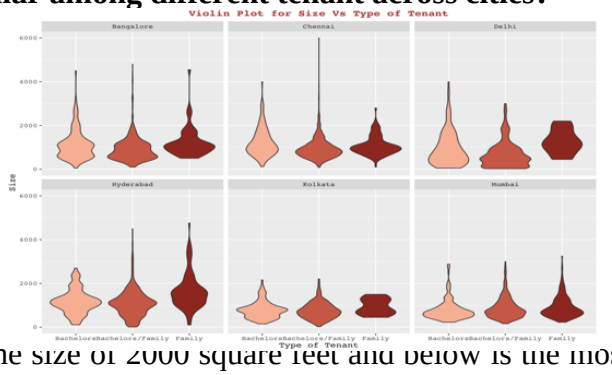



```
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  wilcoxon_test_size_bhk_mumbai$Size and wilcoxon_test_size_bhk_mumbai$BHK

  1      2      3      4
2 <2e-16 -      -      -
3 <2e-16 <2e-16 -      -
4 <2e-16 <2e-16 <2e-16 -
5 0.0001 0.0001 0.0001 0.0322

P value adjustment method: holm
```

The boxplot shows that the average size will increase as the number of BHK increase in every city. Next, it can also show that Chennai is the only one city that have house that could reach up to 6000 square feet with 5 BHK. By using pairwise Wilcoxon rank sum test with assuming a 0.05 as a critical p-value, it shows that all the pair combination of different BHK have a significance different in terms of size in Mumbai. Therefore, this shows that BHK is a factor that affect the size of the rental fees in Mumbai.

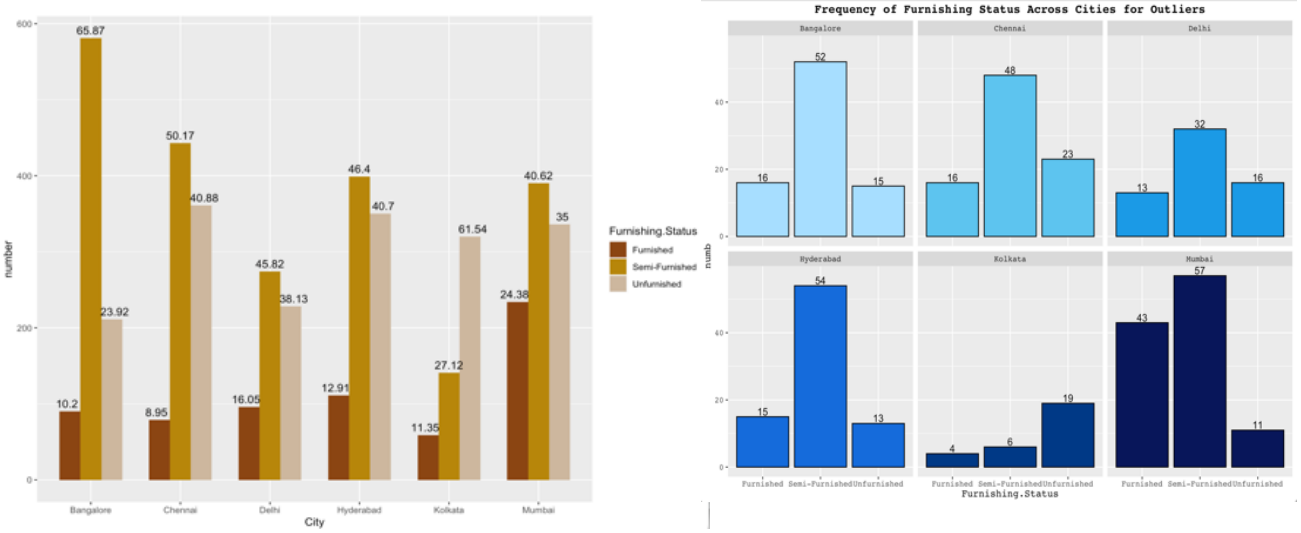
4.1.9 Analysis 9: What size is more popular among different tenant across cities?

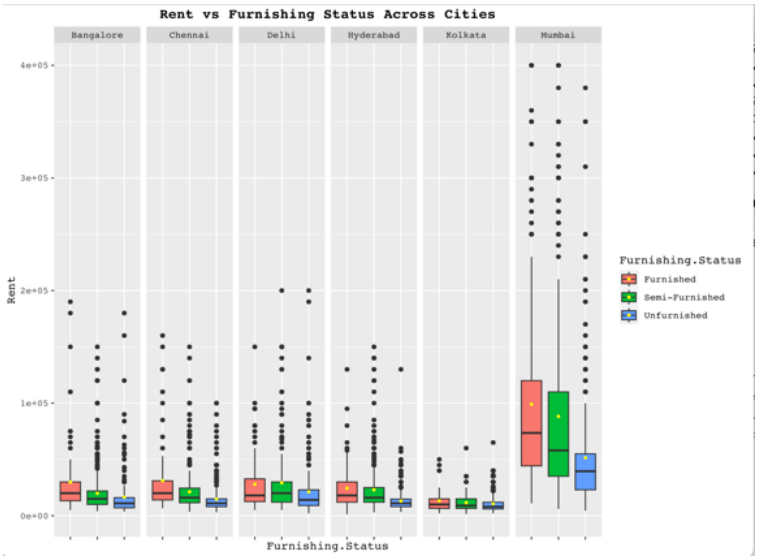
```
        Kruskal-Wallis rank sum test

data:  Size by Tenant.Preferred
Kruskal-Wallis chi-squared = 9.4991, df = 2, p-value = 0.008656

     Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  (rental_data3 %>% filter(City == "Mumbai"))$Size and (rental_data3 %>% filter(City == "Mumbai"))$Tenant.Preferred

                  Bachelors Bachelors/Family
Bachelors/Family 0.0219    -
Family           0.0087    0.2765

P value adjustment method: holm
```



From the violin plot, we are able to notice the size of 2000 square feet and below is the most popular in every type of tenant preference across cities. The Kruskal-Wallis rank sum test also show us there is a significance different between types of tenant preferences in terms of size in Mumbai. However, there is a combination that shows no difference which is Family and Bachelors/Family which mean you will be able to find more or less the same size of rental place when the place is rent to either bachelors/family or family.

4.1.10 Analysis 10: How does a furnishing status affect the rental fees in different city?






```
Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 107.9592, df = 2, p-value = 0


                          Comparison of x by group
                                (No adjustment)
Col Mean-|
Row Mean |   Furnishe   Semi-Fur
---------+----------------------
Semi-Fur |   2.479642
         |     0.0066*
         |
Unfurnis |   9.523901   8.139824
         |     0.0000*   0.0000*

alpha = 0.05
Reject Ho if p <= alpha/2
```

In Mumbai, it has the highest number of furnishing houses compared to other cities while Bangalore will have the highest number of semi-furnished houses instead. Next, the proportion of semi-furnished house is always the highest followed by unfurnished house in almost every city except for Kolkata. When it comes to the outliers of rent, the bar chart shows that there are 43 in total of furnished houses in Mumbai city. Based on the boxplot by excluding Delhi city, it shows the mean rent will be going in a increasing manner when it's from unfurnished house to furnished house. Furthermore, Kruskal-Wallis Test and Dunn Test has shown that there is a significant difference in rent between different types of furnishing status in Mumbai. Therefore, this concludes that furnishing status will be another factors that affect the rent.

4.1.11 Additional Features

1. theme()

   - Use to customize the text style and legends in various graph.

2. function (){}

   - Use to filter out the row of outlier from a city and put it into a new data frame.

3. for (){}

   - Use to automate the function by inputting every city name into the function without doing it manually.

4. xlab()

   - Use in naming the name of x-axis of the graph.

5. summary(lm())$r.squared

   - Use to obtain the coefficient of determination of two continuous variables.

6. right_join()

   - Use to join two tables together with right join.

7. geom_violin()

   - Use to plot a violin plot

8. geom_density()

   - Use to plot a density curve of a continuous variable based on a categorical variable.

9. kruskal.test()

   - Use to perform Kruskal-Wallis Test between a categorical variable and continuous variable.

10. pairwise.wilcox.test()

    - Use to perform Wilcoxon rank sum test for multiple groups in terms of a continuous variable.

4.1.12 Conclusion

As a summary for all the analysis that had been done previously, the tenant preference of the lessor in Mumbai city will not be one of the factors that affect the rental price as tenant preferences have a relatively small variation in average rent among them which makes the hypothesis partially incorrect even though there is a significance difference in terms of the frequency of each type of preferred tenant in Mumbai. From the previous statement, it shows that personnel could just solely choose the place they try to rent based on the lessor preference or their own preferences in other aspects without worrying it in affecting the rental price in Mumbai. Additionally, it also brings us to undergoes further investigation in order to find out which factors could possibly affect the rental price. Hence, BHK, size of rental place, furnishing status has become the first three main concerns among other factors.

Speaking about the BHK, it can be sure there is a strong relation between BHK and rental price in Mumbai as the price will increase when the number of BHK increase and they also can be considered as one of the factors that causes outliers to occur in terms of rent. Additionally, as 2 BHK rental place is the most popular number of BHK in every type of tenant preference in Mumbai especially for the tenant of bachelors/family which they have the highest frequency of 2 BHK houses, this implicitly indicates that it have higher chance of having difference combination with other potential factors to rise up the rent in order to make the average rent of bachelors/family type of tenant preference become almost equal with the other two types of tenant preference's average rent.

Size and furnishing status are the other two factors that is being proved statistically to have an impact to the rental price. It is being known that there will be no size difference of rental place when it's between the lessor that prefer bachelors/family tenant and the lessor who prefer family. For instance, a family who looks for a place to rent in Mumbai will not need to worry about choosing the one that prefer family or the one who prefer both types of tenants will have a huge gap difference in terms of size. Lastly, choosing the rental place based on personnel needs of the furnishing status is rather important as it might cause a difference compared with other choices.

4.2 Objective 2: (Teh Yue Feng)

To investigate the relationship between the total number of different types of stakeholders and the cities they are locate.

Does number of both rental posts by agents and owners varies across cities?

4.2.1 Analysis 1: How many types of Point.of.Contact?

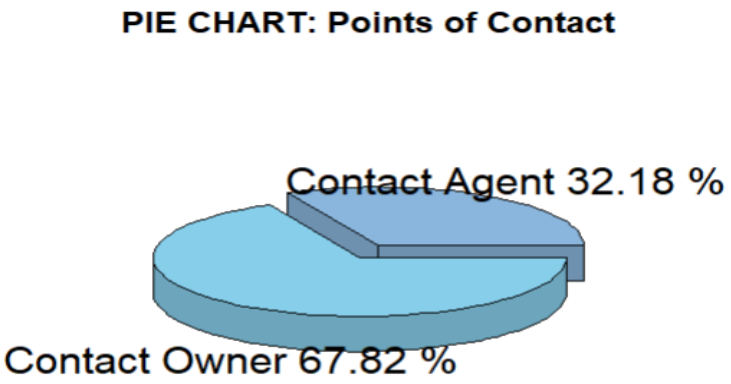
```
    # ---- HOW MANY TYPE OF POINT.OF.CONTACT? ---
  point_contact <- levels(factor(rental_data2$Point.of.Contact))
  point_contact
```

```
>   point_contact <- levels(factor(rental_data2$Point.of.Contact))
>   point_contact
[1] "Contact Agent" "Contact Owner"
```

The output indicates that the "Point.of.Contact" column in the rental_data2 dataset has two unique levels or categories: “Contact Agent” and “Contact Owner”.

4.2.2 Analysis 2: What are the overall percentage of Point.of.Contact?

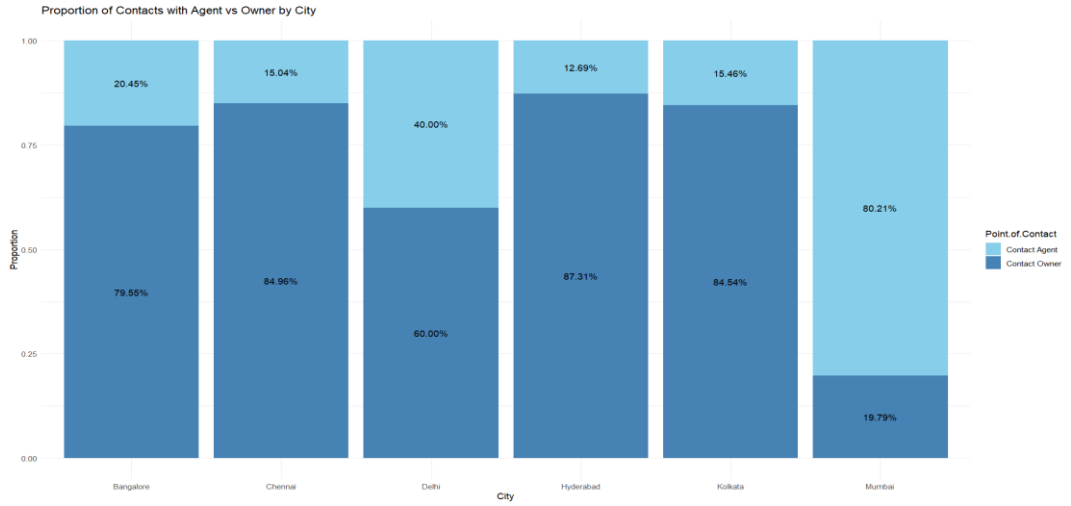
```
  #           --- ANALYSIS 1-2: WHAT ARE THE OVERALL PERCENTAGE OF POINT.OF.CONTACT? ---
total_agents <- sum(rental_data2$Point.of.Contact == "Contact Agent")
total_owners <- sum(rental_data2$Point.of.Contact == "Contact Owner")
total_poc <- total_agents + total_owners

contact_counts <- data.frame(
  Point.of.Contact = c("Contact Agent", "Contact Owner"),
  Count = c(total_agents, total_owners) )

cus_col = c("#8AB5DC","skyblue", "#6082B6", "steelblue", "#426484", "#16212C")

pie3D(contact_counts$Count,
      labels = contact_counts$Label,
      col = cus_col,
      main = "PIE CHART: Points of Contact",
      border = "black",
      explode=0.08)
```



Above you can see the outcome of this pie chart, which reveals that "Contact Owner" owns 67.8% of the contacts while "Contact Agent" only has 32.2%.

4.2.3 Analysis 3: What are the percentage of Point.of.Contact by each city?

```
#           --- ANALYSIS 1-3: WHAT ARE THE PERCENTAGES OF POINT.OF CONTACT BY EACH CITY? ---

proportion_poc <- rental_data2 %>%
  group_by(City, Point.of.Contact) %>%
  summarise(Count = n())

proportion1 <- proportion_poc %>% # CALCULATE PROPORTIONS
  group_by(City) %>%
  mutate(TotalContacts = sum(Count),
         Proportion = Count / TotalContacts)

ggplot(proportion1, aes(x = City, y = Proportion, fill = Point.of.Contact)) +
  geom_bar(stat = "identity") +
  labs(title = "Proportion of Contacts with Agent vs Owner by City",
       x = "City",
       y = "Proportion") +
  geom_text(aes(label = scales::percent(Proportion)),
            position = position_stack(vjust = 0.5)) +
  theme_minimal() +
  scale_fill_manual(values = c("Contact Agent" = "skyblue", "Contact Owner" = "steelblue"))
```



By each city, all of them mostly more prefer contact owner as their channel to find the room. But in Mumbai, tenant more prefer with contact agent, up to 80% of the population here favours Contact Agent.

4.2.4 Analysis 4: The rent ranges of either Point.of.Contact by each city?

```
#       ---   ANALYSIS 1-4: Find the relationship between "Point.of.Contact" & "Rent". ---
#             ---     THE RENT RANGE OF EITHER POINT.OF.CONTACT BY EACH CITY   ---

  rent_ranges <- c(0, 10000, 30000, 60000, 100000, 125000, Inf)
  categorize_rent <- function(rent) {
    cut(rent, breaks = rent_ranges, labels = c("<10k", "10k-30k", "30k-60k",
                                               "60k-100k", "100k-125k", ">125k"), include.lowest = TRUE)
                                    }
  # CATEGORIZE RENTS & CALCULATE PREFERRED POINT OF CONTACT IN EACH CITY
preferred_poc2 <- rental_data2 %>%
  mutate(Rent = categorize_rent(Rent)) %>%
  group_by(City, Rent, Point.of.Contact) %>%
  summarise(TOTAL = n()) %>%
  arrange(City, Rent, desc(TOTAL)) %>%
  select(City, Rent, Point.of.Contact, TOTAL)
  # CREATE A BARPLOT WITH FACETS FOR EACH CITY & EACH CATEGORY
ggplot(preferred_poc2, aes(x = Point.of.Contact, y = TOTAL, fill = TOTAL)) +
  geom_bar(stat = "identity", width = 0.75) +
  labs(title = "Preferred Point of Contact by City and Rent Range",
     x = "Point of Contact",
     y = "RESULT") +
  theme_minimal() +
  geom_text(aes(label = TOTAL)) +
  scale_fill_gradientn("TOTAL", colors = cus_col) +
  facet_grid(City ~ Rent, scales = "free_y", space = "free") +
  theme(axis.text.y.left = element_blank())
```

In the presented bar plot, city rent preferences are shown. The city with the most "Contact Agent" preferences is Mumbai, whereas other cities prefer "Contact Owner". The most popular cities are Bangalore, Chennai, Delhi, and Hyderabad. In the 10k-30k rent range, these cities favor "Contact Owner." In contrast, Kolkata's trend shows a considerable decline in "Contact Agent" preference from 302 to 127 when rent ranges from <10k to 10k-30k. Each city and rental range have its own contact channel preferences, as shown by this analysis.

Hypothesis Testing

```
# CHI-SQUARE TEST
    test1 <- xtabs(~City + Point.of.Contact, rental_data2 )
    chisq.test(test1)
    assocstats(mytable)
```

H0: Both City and Point of contact are not related.

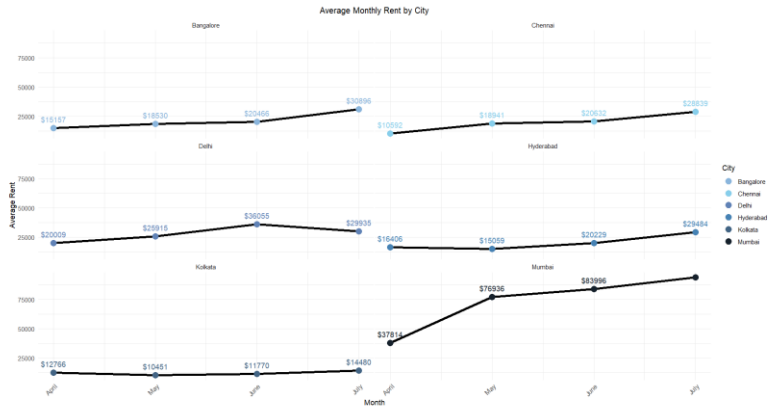
H1: Both City and Point of contact are related.

Since the p-value is lower than the critical value 0.01 that we assumed, we reject H0; Accept H1. Therefore, we will conclude that both City and Point of contact are related.

```
>       assocstats(test1)
                    X^2 df P(> X^2)
Likelihood Ratio 1416.5  5        0
Pearson          1435.9  5        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.482
Cramer's V        : 0.55
```

4.2.5 Analysis 5: How about the average rent in each month by each city?

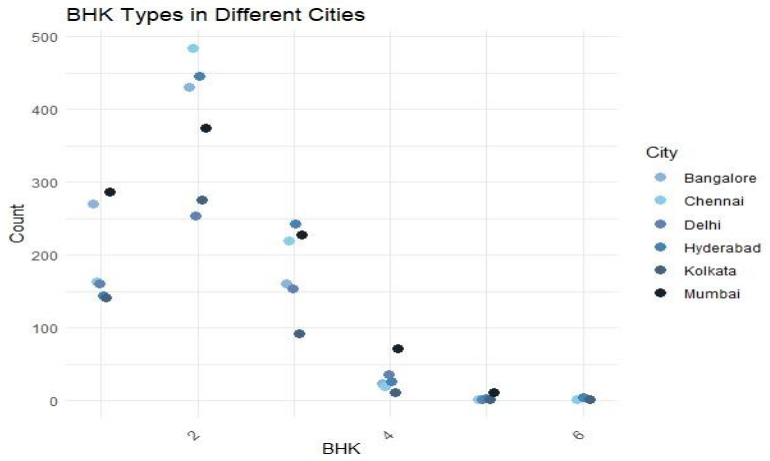
```
#                          INSIDE EACH CITY'S POSTED.ON,
#                  WHAT TYPE OF BHK'S ARE TRENDING IN THE MONTHS OF RENT
#                                        &
#                          WHICH POC DO THESE PEOPLE ASPIRE TO MORE?
#             ANALYSIS 1-5: HOW ABOUT THE AVERAGE RENT IN EACH MONTH IN EACH CITY?

avgrent_month <- rental_data2 %>%
  group_by(Month = month(Posted.On), City) %>%
  summarise(AVG_RENT_MONTHLY = mean(Rent, na.rm = TRUE)) %>%
  ungroup()

ggplot(avgrent_month, aes(x = Month, y = AVG_RENT_MONTHLY, group = City, color = City)) +
  geom_line(size = 1.5, linetype = "solid", color = "black") +
  geom_point(size = 4, shape = 16) +
  geom_text(aes(label = sprintf("$%d", round(AVG_RENT_MONTHLY))), vjust = -1) +
  scale_x_continuous(breaks = 1:12, labels = month.name[1:12]) +
  labs(x = "Month", y = "Average Rent", title = "Average Monthly Rent by City") +
  scale_color_manual(values = cus_col) +
  theme_minimal() +
  theme(  plot.title = element_text(size = 12, hjust = 0.5),
          axis.text.x = element_text(angle = 45, hjust = 1)  ) +
  theme(  plot.margin = margin(1, 1, 1, 1, "cm"),
          panel.spacing = unit(0.3, "lines")  ) +
  facet_wrap(~ City, ncol = 2)
```



The trends observed in the point and line plot, where average rentals are analyzed across different cities, reveal notable variations in rental patterns. While most cities display relatively stable and consistent trends over time, Mumbai stands out with a significant and consistent upward trend in average rents. This substantial increase suggests a distinct and continuous demand-supply dynamic specific to Mumbai's rental market. In contrast, Delhi's trend is characterized by an initial upward movement followed by a subsequent decline. This fluctuation in Delhi's rental rates indicates potential shifts in market conditions or influencing factors. These diverse trends underscore the need for city-specific analyses and the consideration of unique market dynamics that contribute to the observed variations in rental pricing patterns.

4.2.6 Analysis 6: Which type of BHK is the most preferable by each city?

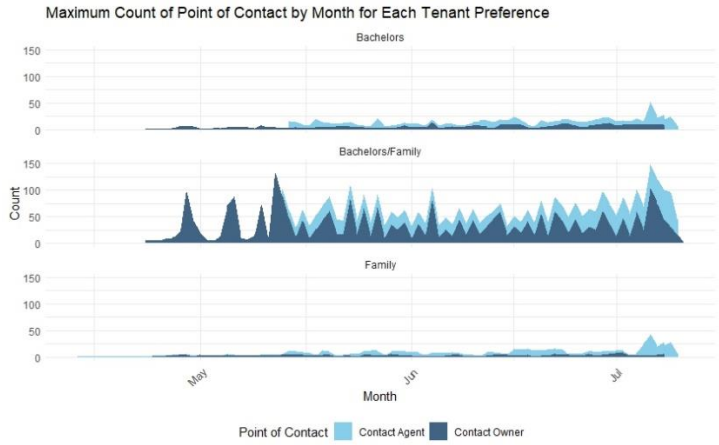
```
#           ---   ANALYSIS 1-6: WHICH TYPE OF BHK IS THE MOST PREFERABLE BY EACH CITY? ---

agg_data <- rental_data2 %>%
  group_by(City, BHK) %>%
  summarize(Count = n())

ggplot(agg_data, aes(x = BHK, y = Count, color = City)) +
  geom_point(position = position_dodge(width = 0.2), size = 3) +
  labs(x = "BHK", y = "Count", title = "BHK Types in Different Cities")+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 2)) +
  scale_color_manual(values = c(cus_col))
```



From the point result, we can know in Chennai BHK quantity as 2 is the most attract. The amount of BHK(4, 5, and 6) , the whole city is not very inclined, probably because it is not very necessary to use so many BHKs.

4.2.7 Analysis 7: In maximum, Tenant more prefers which Point.of.Contact in each month?

```
#          ANALYSIS 1-7: TENANT MORE PREFER WHICH POINT.OF.CONTACT IN EACH MONTH WITH MAXIMUM COUNT?

preferred_poc3 <- rental_data2 %>%
  group_by(Posted.On, Point.of.Contact, Tenant.Preferred) %>%
  summarise(Count = n()) %>%
  arrange(Posted.On, Tenant.Preferred, desc(Count))
maxpoc_monthly1 <- preferred_poc3 %>%
  group_by(Posted.On, Tenant.Preferred) %>%
  filter(Count == max(Count)) %>%
  ungroup()

ggplot(maxpoc_monthly1, aes(x = Posted.On, y = Count, fill = Point.of.Contact)) +
  geom_area(position = "stack") +
  labs(title = "Maximum Count of Point of Contact by Month for Each Tenant Preference",
       x = "Month",
       y = "Count",
       fill = "Point of Contact") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "bottom") +
  scale_fill_manual(values = c("Contact Agent" = "skyblue", "Contact Owner" = "#426484")) +
  facet_wrap(~ Tenant.Preferred, ncol = 1)
```



For Maximum Count, Bachelors/Family is the most fluctuating and the Tenant most inclined to Contact Owner. That's mean most of them more prefer with Contact Owner. But when only Family, they more prefer Contact Agent, but also a small part of them prefer to Contact Agent.

Hypothesis Testing

```
test2 <- xtabs(~Point.of.Contact + Tenant.Preferred, rental_data2)
chisq.test(test2)
assocstats(test2)
View(test2)
```

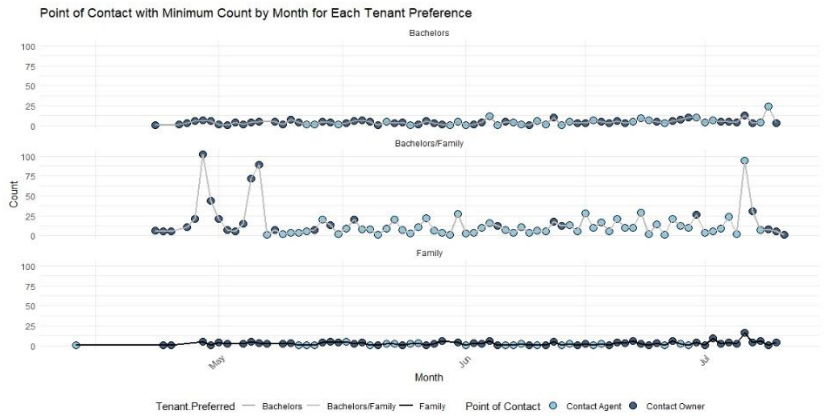
H0 : Both point of contact and tenant preferred are not related.

H1 : Both point of contact and tenant preferred are related.

Since the P-Value is lower than the critical value 0.05 that we assumed; Reject H0; Accept H1. Therefore, we will conclude that both City and Point of Contact are related.

4.2.8 Analysis 8: In Minimum, Tenant more prefers which Point.of.contact in each month?

```
#         ANALYSIS 1-8: TENANT MORE PREFER WHICH POINT.OF.CONTACT IN EACH MONTH WITH MINIMUM COUNT?

preferred_poc4 <- rental_data2 %>%
  group_by(Posted.On, Point.of.Contact, Tenant.Preferred) %>%
  summarise(Count = n()) %>%
  arrange(Posted.On, Tenant.Preferred, Count)
minpoc_monthly <- preferred_poc4 %>%
  group_by(Posted.On, Tenant.Preferred) %>%
  filter(Count == min(Count)) %>%
  ungroup()

ggplot(minpoc_monthly, aes(x = Posted.On, y = Count, fill = Point.of.Contact)) +
  geom_point(stat = "identity", shape = 21, size = 3.5, color = "black") +
  geom_line(aes(x = Posted.On, y = Count, group = Tenant.Preferred,
                color = Tenant.Preferred), size = 0.8) +
  labs(title = "Point of Contact with Minimum Count by Month for Each Tenant Preference",
       x = "Month",
       y = "Count",
       fill = "Point of Contact") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "bottom") +
  scale_fill_manual(values = c("Contact Agent" = "skyblue", "Contact Owner" = "#426484")) +
  scale_color_manual(values = c("Family" = "black", "Bachelors" = "darkgrey",
                                "Bachelors/Family" = "grey")) +
  facet_wrap(~ Tenant.Preferred, ncol = 1)
```



For Minimum Count, I have used the point and line to show that Bachelors/Family also fluctuates the most, but it fluctuates for Contact Agent.

4.2.9 Extra Features

a) Theme()

Function is used to customize the appearance and style of your plots. It allows you to modify various aspects of the plot's visual elements, such as text, axis labels, legend, background, and more.

b) Geom_area()

As a stacked area plot, is typically used to display the composition or distribution of a variable across multiple categories or time periods.

c) count():

Use to count the appear time of the values.

d) theme_minimal():

Remove the grey background and show cleaner.

e) theme():

Use to customize elements inside the plot.

f) labs():

Set the title for the plot.

4.2.10 Conclusion

The analysis of minimum count preferences and its relationship with the "Bachelors/Family" category adds another layer of complexity to the understanding of tenant preferences and contact channels, further reinforcing the connections to the previous conclusions. The observation that "Bachelors/Family" exhibits significant fluctuations in minimum count preferences aligns with the trend of "Contact Agent" preference in this category. This suggests that, despite fluctuations, a notable proportion of tenants within this category tend to lean towards "Contact Agent" as their preferred contact channel when the minimum count is considered.

This observation creates an interesting dynamic when compared to the maximum count preferences, where "Bachelors/Family" leaned more towards "Contact Owner." The juxtaposition of these findings highlights the intricate balance between tenant behaviours and their preferred communication methods, which can vary based on different factors such as the urgency of finding a property or the specific needs of the tenant.

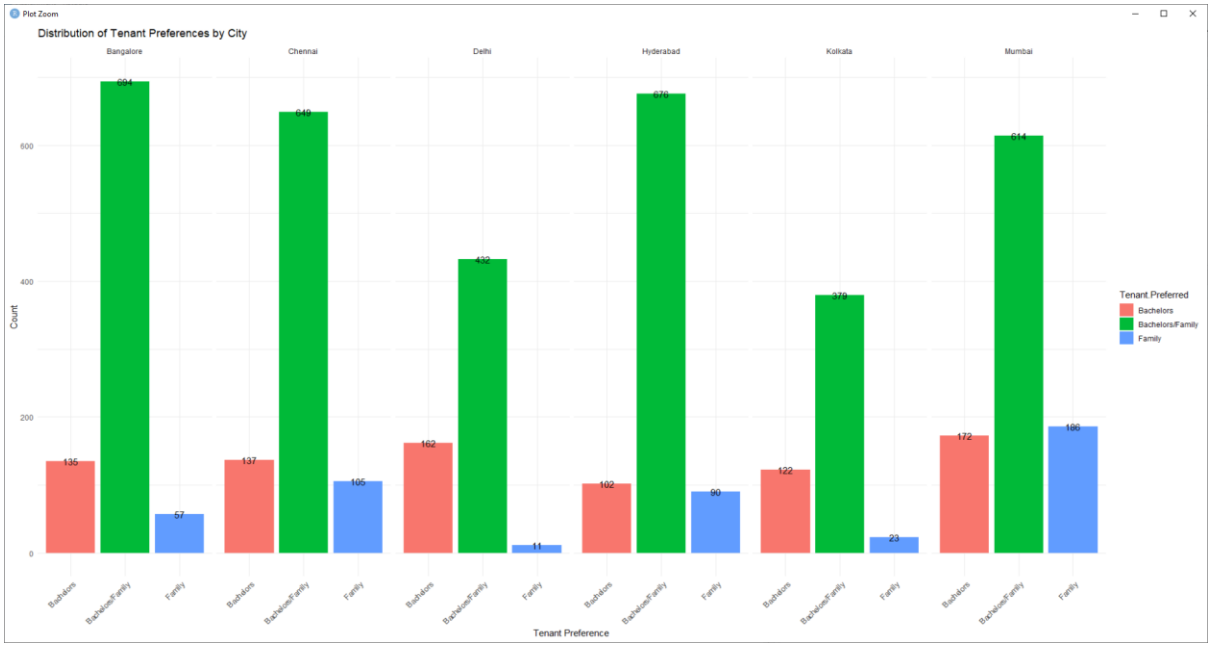
Connecting these findings to the earlier conclusions enriches the overall narrative. The variations in "Bachelors/Family" preferences for both "Contact Agent" and "Contact Owner" underscore the nuanced nature of tenant behaviours, emphasizing that preferences can be influenced by circumstances, time constraints, and individual requirements. These insights guide stakeholders in tailoring their communication strategies, recognizing that tenant preferences can differ not only across cities but also based on the urgency and context of their property search.

4.3 Objective 3 (Lian Jun Er)

To investigate the relationship between the total number of posts for every type of tenant preferred by landlord and the cities they are locate.

4.3.1 Analysis 1: How many times each type of tenant preference appears in posts from that city?

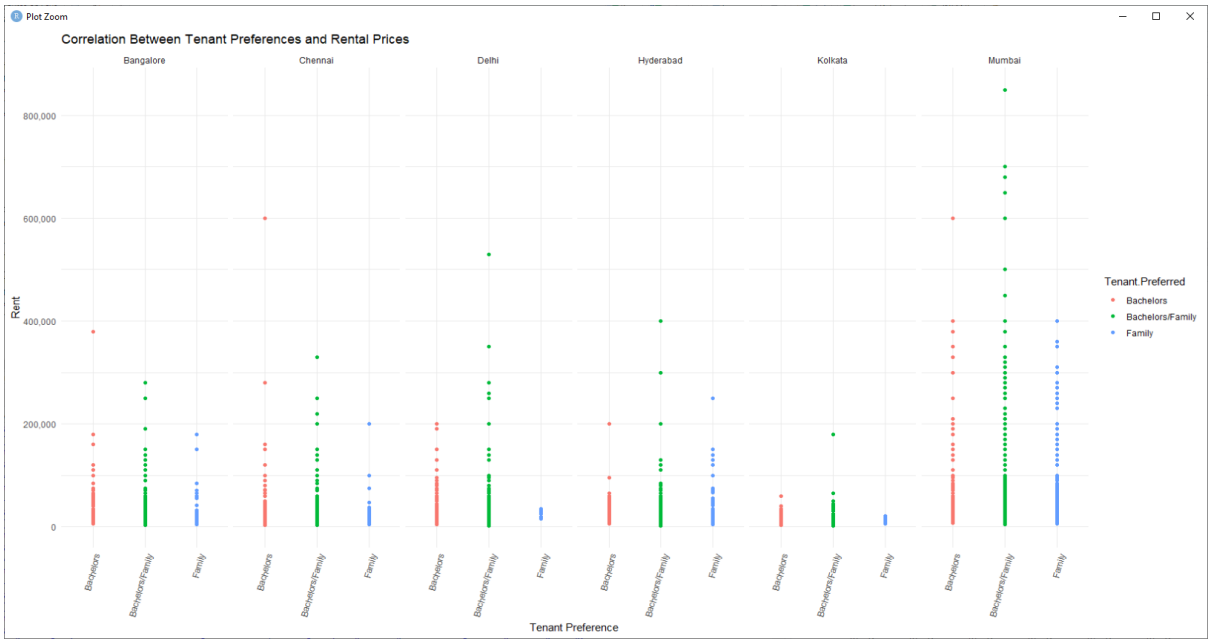
```r
#  How many times each type of tenant preference appears in posts from that city?

tenant_counts <- rental_data %>%
  group_by(City, Tenant.Preferred) %>%
  summarize(Count = n())

ggplot(tenant_counts, aes(x = Tenant.Preferred, y = Count, fill = Tenant.Preferred)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Distribution of Tenant Preferences by City", x = "Tenant Preference", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  facet_wrap(~ City, ncol = 6) +
  geom_text(aes(label=Count))
```



The bar graph above shows that the Distribution of tenant preferences by city. Although the number of Bachelor tenant in Mumbai is not the highest. But it is the highest number compare with other cities.

4.3.2 Analysis 2: What is the correlation between specific tenant preferences and higher rental prices in certain cities?

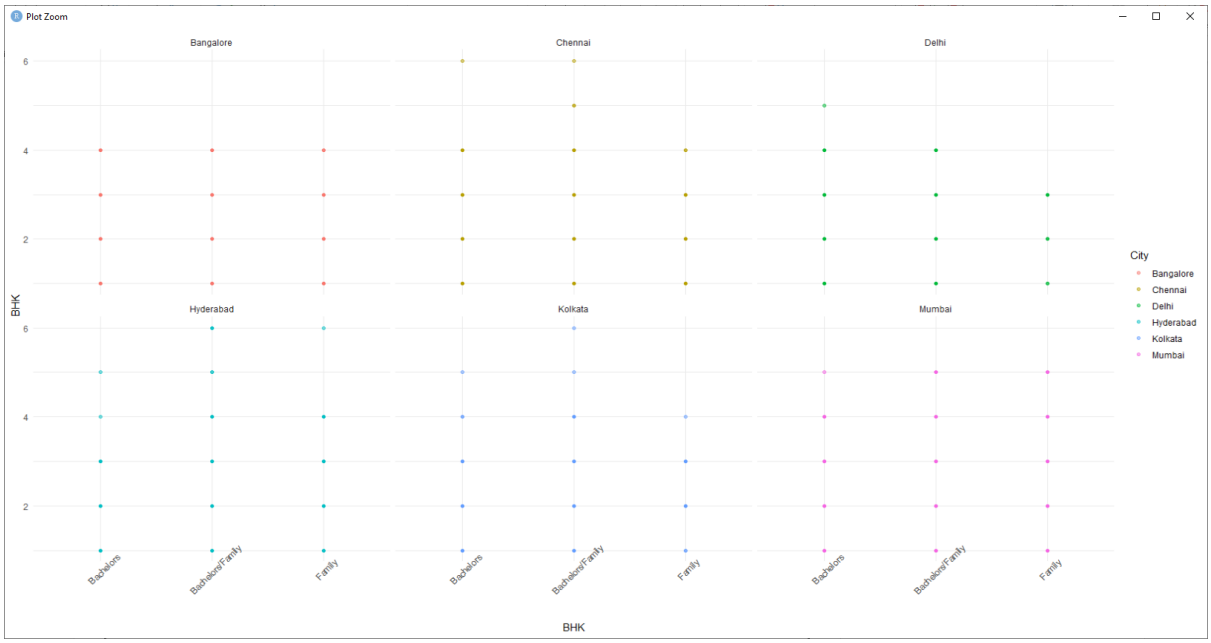
```r
#   What is the correlation between specific tenant preferences and higher rental prices in certain cities?


correlation_plot <- rental_data %>%
  group_by(City, Tenant.Preferred);

ggplot(correlation_plot, aes(x = Tenant.Preferred, y = Rent, col = Tenant.Preferred)) +
  geom_point() +
  scale_y_continuous(labels = comma)+
  theme_minimal() +
  labs(title = "Correlation Between Tenant Preferences and Rental Prices", x = "Tenant Preference", y = "Rent") +
  theme(axis.text.x = element_text(angle = 75, hjust = 1))+
  facet_grid(~ City, scales = "free_x")
```



Most of the tenant's rental are below 200,000. But in Mumbai, there is a few tenants' rental are equal to or more than 200,000. Especially, there is a Bachelor/Family tenant's rental is 800,000.

4.3.3 Analysis 3: How do tenant preferences vary with the number of bedrooms (BHK) in certain cities?

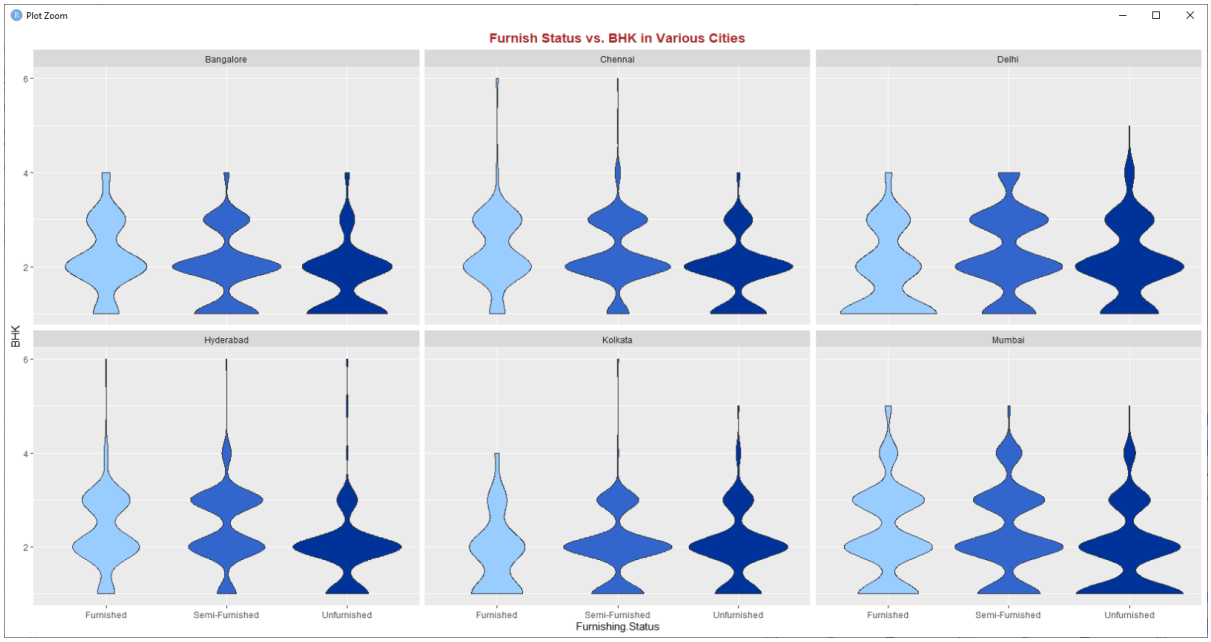
```
61
62
63  #  How do tenant preferences vary with the number of bedrooms (BHK) in certain cities?
64
65
66
67  plot_for_BHK_tenan_prefer <- rental_data %>%
68    group_by(BHK, Tenant.Preferred, City);
69
70  ggplot(plot_for_BHK_tenan_prefer, aes(x=Tenant.Preferred, y=BHK, color=City))+
71    geom_point(alpha=0.5)+
72    labs(x = "BHK", Y= "Count", titlle ="BHK vs Tenant Preferred Across Cities")+
73    facet_wrap(~City)+
74    theme_minimal()+
75    theme(axis.text.x = element_text(hjust = 0.5, angle = 45))
76
77
```



The graph above is showing that how do tenant preferences vary with the number of bedrooms (BHK) in certain cities. In the result, it shows that Bachelor/Family tenant are required a greater number of BHK.

4.3.4 Analysis 4: Is there a correlation between furnish status and different number of BHK in various cities?

```
80
81  #  Is there a correlation between furnish status and different number of BHK in various cities?
82
83  furnish_BHK <- rental_data %>%
84    group_by(BHK, Furnishing.Status, City);
85
86  color_combination <- c("#99CCFF", "#3366CC", "#003399")
87
88  ggplot(furnish_BHK, aes(x=Furnishing.Status, y=BHK, fill=Furnishing.Status))+
89    geom_violin()+
90    scale_fill_manual(values = color_combination)+
91    facet_wrap(~City)+
92    ggtitle("Furnish Status vs. BHK in Various Cities")+
93    theme(plot.title = element_text(face = "bold", color = "brown", hjust = 0.5),
94          legend.position = "none")
95
```



It is clear from the research that there is no statistically significant relationship between the furnished or unfurnished status of rental units, the number of bedrooms (BHK) in the dataset, or the cities under investigation. This implies that the number of bedrooms in a property does not determine its furnishing status.

4.3.5 Extra Features

1) scale_y_continuous( ) : function is used to alter the scales (axis limits, labels, breaks, etc.) on the y-axis of a plot. It is a component of the ggplot2 package.
2) geom_violin( ): is a piece of geometry used to make violin plots.

4.3.6 Conclusion

In conclusion, my investigation has provided some valuable information about tenant preferences and rental properties. Mumbai distinguishes out with a notable presence of Bachelor tenants despite not having the biggest overall number of bachelor tenants. Mumbai also shows a distinct pattern of increased rental rates, including cases where tenants are ready to pay much more, especially in the case of bachelor/family tenants with rentals approaching 800,000.

The findings imply that bachelor/family tenants prefer property that with more BHK units, which is consistent with the association between tenant preferences and the number of bedrooms (BHK) in different cities.

However, the thorough study shows that there is no statistically significant association between the furnished or unfurnished condition of rental units and the number of bedrooms (BHK) throughout the dataset and cities examined. This demonstrates that a rental property's furnishing level is not solely determined by the number of bedrooms.

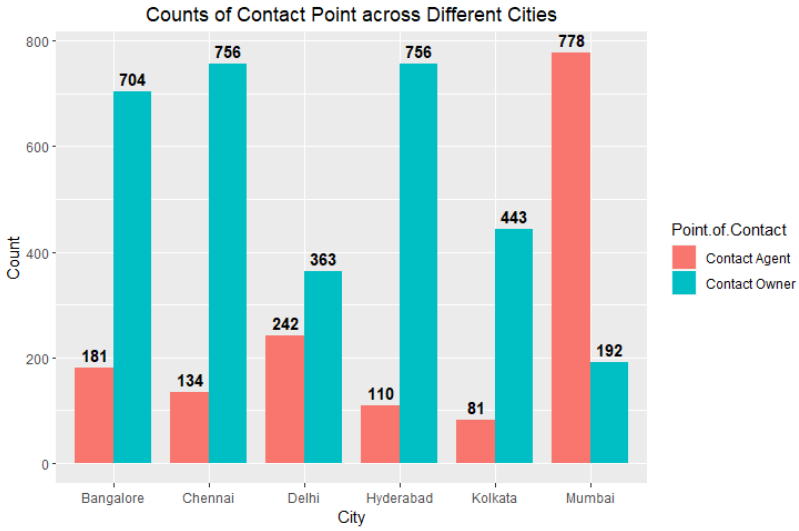
In summary, this study offers important new understandings of tenant preferences and the dynamics of the rental property market. The findings highlight the complex nature of the real estate business, where a variety of factors interact to influence renters' decisions, even though distinct patterns do appear in Mumbai.

4.4 Objective 4: (Soo Jiun Guan)

To discover the relationship between the type of person who post their rental information and rent across different cities.

4.4.1 Analysis 1: How much is the rental information posted by both contact points across different cities?

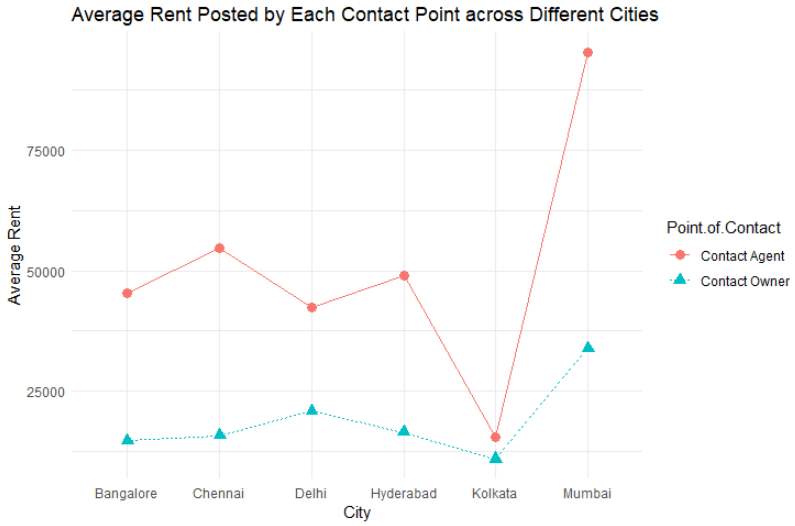
```
contact_count <- rental_data2 %>%
  filter(Point.of.Contact %in% c("Contact Owner", "Contact Agent")) %>%
  count(City, Point.of.Contact)

ggplot(contact_count, aes(x = City, y = n, fill = Point.of.Contact)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.8) +
  geom_text(aes(label = n, fontface = "bold"),
            position = position_dodge(width = 0.8), vjust = -0.5) +
  labs(x = "City", y = "Count",
       title = "Counts of Contact Point across Different Cities") +
  theme(plot.title = element_text(hjust = 0.5))
```



The data presented in the above graph underscores a notable trend. Across most cities, the trend leans towards more rental information posted by owners rather than posted by agents. However, there is a bit different in Mumbai. Over there, the majority of rental information is actually put up by agents, unlike the usual pattern where direct postings by owners are more common.

4.4.2 Analysis 2: How about the average rent posted by both contact points across different cities?

```
average_rent <- rental_data2 %>%
  group_by(City, Point.of.Contact) %>%
  summarise(Average_Rent = mean(Rent))

ggplot(average_rent, aes(x = City, y = Average_Rent, color = Point.of.Contact)) +
  geom_point(size = 3, aes(shape = Point.of.Contact)) +
  geom_line(aes(group = Point.of.Contact, linetype = Point.of.Contact)) +
  labs(x = "City", y = "Average Rent",
       title = "Average Rent Posted by Each Contact Point across Different Cities") +
  theme_minimal()
```



The graph provides visualization of the average rent for properties in different cities based on the point of contact, which can either be “Contact Agent” or “Contact Owner”. In all listed cities, their average rent is higher when properties are posted by agents compared to when they are posted by owners. Especially in Mumbai, the average rent difference between properties posted by agents and owners is substantial, with properties posted by agents having significantly higher average rents, reflecting the agent's influence on raising of the rent from another aspect.

Hypothesis Testing

```
City_Name <- unique(rental_data2$City)
for (i in 1:length(City_Name)){
  filtered_city_row <- rental_data2%>%filter(City==City_Name[i])
  cat("\n")
  print(sprintf("Kruskal Walts Test in %s", City_Name[i]))
  print((wilcox.test(Rent~Point.of.Contact, filtered_city_row))$p.value)
}
```

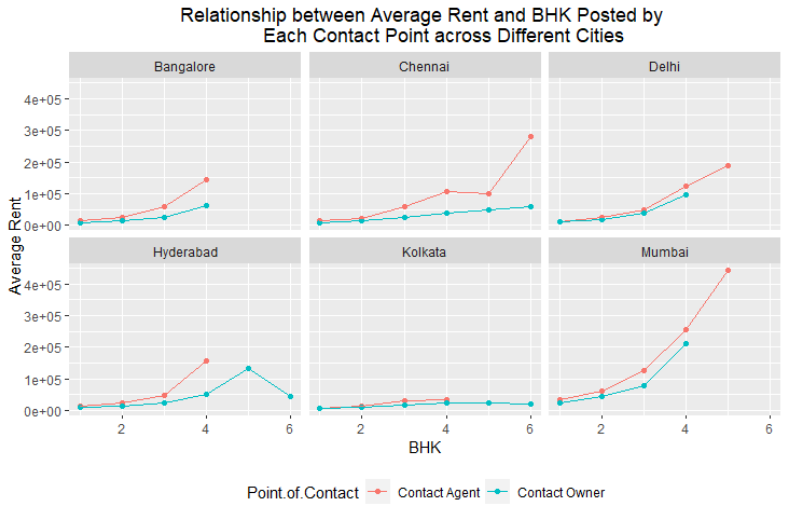
H0: There is no significant difference in rent across different cities based on the point of contact.

H1: There is a significant difference in rent across different cities based on the point of contact.

Since the P-Value is lower than the critical value 0.05 that we assumed; Reject H0; Accept H1. Therefore, there are significant differences in rent between the cities based on the point of contact.

4.4.3 Analysis 3: What is the relationship between average rent and BHK posted by each contact point across different cities?

```
avgrent_bhk <- rental_data2 %>%
  group_by(City, BHK, Point.of.Contact) %>%
  summarise(Average_Rent = mean(Rent))

ggplot(avgrent_bhk, aes(x = BHK, y = Average_Rent, color = Point.of.Contact)) +
  geom_point() +
  geom_line() +
  facet_wrap(~ City, scales = "fixed") +
  labs(x = "BHK", y = "Average Rent",
       title = "Relationship between Average Rent and BHK Posted by
       Each Contact Point across Different Cities") +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5))
```



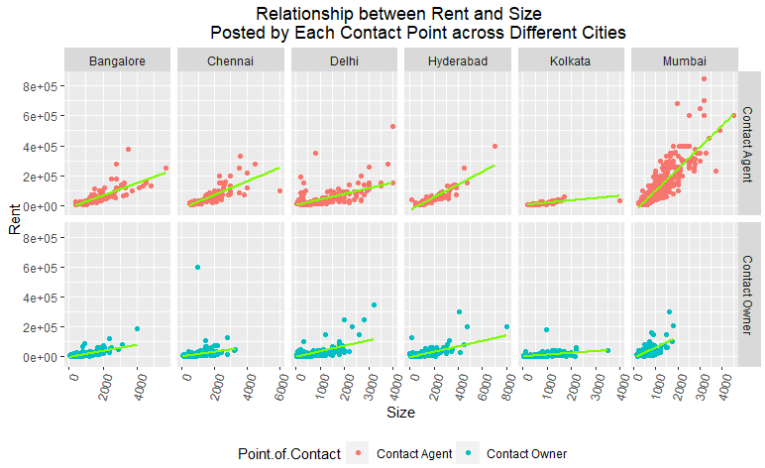
The plot illustrates how the average rent posted by different contact points changes based on the number of bedrooms, bathrooms, and hall-kitchen units (BHK) in various cities. We generally expect rents to increase as the number of bedrooms increases. Interestingly, there are some peculiar observations we can make from the plot.

In few instances, the average rent for properties with a lower number of BHK might surpass the average rent of properties with a higher number of bedrooms. The reason for this phenomenon might be due to some lower BHK properties may be larger in terms of square footage. Additionally, demand for 6 BHK properties might be relatively lower than for 5 BHK properties in these cities, which is able to lead to lower rents due to lower demand.

However, in most cases, there is a noticeable upward pattern in average rent across most cities as the BHK count increases, regardless of who posted it, among which in Mumbai is the most obvious. This outcome is quite predictable and aligns with our expectations as larger properties generally command higher rents. Furthermore, the average rent posted by agents is almost all higher compared to that posted by owners. From this, it can serve as a compelling demonstration of the significant influence agents can wield on rental prices.

4.4.4 Analysis 4: What is the relationship between rent and size posted by each contact point across different cities?

```
ggplot(rental_data2 , aes(x = Size, y = Rent, col = Point.of.Contact)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color ="#7CFC00") +
  facet_grid(Point.of.Contact~ City, scales = "free_x") +
  labs(x = "Size", y = "Rent",
       title = "Relationship between Rent and Size
       Posted by Each Contact Point across Different Cities") +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 75, hjust = 1))
```



The above scatter plot and regression line show the relationship between rent, property size, and points of contact across different cities. In most cities, there is a general trend indicating that larger rental properties tend to have higher rents. This is particularly evident in Mumbai where the rent significantly increases as the property size increases regardless of who posted.

4.4.5 Analysis 5: How about the impact of the interconnection between property size and rent across different cities and lessors?

```
opes_rent_size <- rental_data2 %>%
group_by(City, Point.of.Contact) %>%
summarize(Slope = coef(lm(Rent ~ Size))[2]
ew(slopes_rent_size)
```

	City	Point.of.Contact	Slope
1	Bangalore	Contact Agent	41.32839
2	Bangalore	Contact Owner	20.52595
3	Chennai	Contact Agent	45.42116
4	Chennai	Contact Owner	15.60479
5	Delhi	Contact Agent	37.87030
6	Delhi	Contact Owner	36.66551
7	Hyderabad	Contact Agent	42.70775
8	Hyderabad	Contact Owner	18.07923
9	Kolkata	Contact Agent	15.43620
10	Kolkata	Contact Owner	11.75490
11	Mumbai	Contact Agent	145.57476
12	Mumbai	Contact Owner	69.96666

The above table shows the slope of each regression line and then provides us with insights into how property size and rent are interconnected across various cities, considering different lessors. In all cities, contact with agents results in a pronounced positive relationship, especially in Mumbai.

4.4.6 Analysis 6: How about the strength of linear relationship between property size and rent in different cities based on different point of contact?

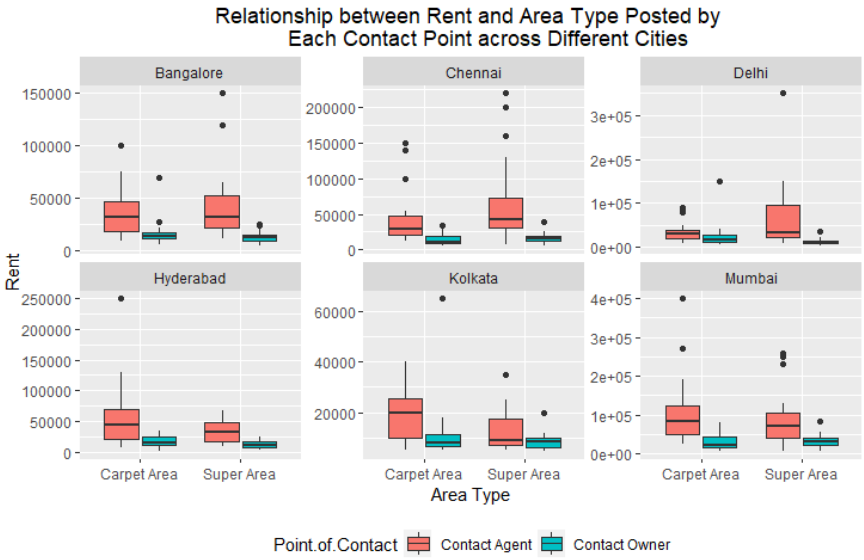
```
correlation_rent_size <- rental_data2 %>%
  group_by(City, Point.of.Contact) %>%
  summarize(Correlation = cor(Size, Rent))
View(correlation_rent_size)
```

	City	Point.of.Contact	Correlation
1	Bangalore	Contact Agent	0.7854012
2	Bangalore	Contact Owner	0.7223133
3	Chennai	Contact Agent	0.7604759
4	Chennai	Contact Owner	0.2722801
5	Delhi	Contact Agent	0.5808388
6	Delhi	Contact Owner	0.6140754
7	Hyderabad	Contact Agent	0.8836031
8	Hyderabad	Contact Owner	0.6538203
9	Kolkata	Contact Agent	0.7408387
10	Kolkata	Contact Owner	0.4223310
11	Mumbai	Contact Agent	0.8608067
12	Mumbai	Contact Owner	0.6208901

The above table shows the correlation coefficient of the size and rent based on different point of contact and cities, which reflects the strength of the linear relationship between these two variables. From the table, there is a constant trend of higher correlation between rent and size in contact agents compare to the contact owner in most of the cities, with the exception of Delhi where is not significant. These findings suggests that, in most cities, properties offered by agents tend to charge higher rents as their size increases.

4.4.7 Analysis 7: What is the relationship between rent and area type posted by each contact point across different cities?

```
sampling <- rental_data2 %>%
  group_by(City,Area.Type,Point.of.Contact) %>%
  sample_n(20, replace = FALSE )

ggplot(sampling, aes(x = Area.Type, y = Rent, fill = Point.of.Contact)) +
  geom_boxplot() +
  facet_wrap(~ City, scales = "free_y") +
  ggtitle("Relationship between Rent and Area Type Posted by Each Contact Point
          across Different Cities") +
  xlab("Area Type") +
  ylab("Rent") +
  theme(legend.position = "bottom", plot.title = element_text(hjust = 0.5))
```




```r
medians_rent_areatype <- sampling %>%
  group_by(City, Area.Type, Point.of.Contact) %>%
  summarise(Median.Rent = median(Rent))

medians_list <- split(medians_rent_areatype, medians$City)

for (city_name in names(medians_list)) {
  cat("Median Rent for", city_name, "\n")
  print(medians_list[[city_name]])
  cat("\n")
}
```

```
Median Rent for Bangalore
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City      Area.Type    Point.of.Contact Median.Rent
  <chr>     <chr>        <chr>                  <dbl>
1 Bangalore Carpet Area  Contact Agent          40000
2 Bangalore Carpet Area  Contact Owner          14000
3 Bangalore Super Area   Contact Agent          35500
4 Bangalore Super Area   Contact Owner           9750

Median Rent for Chennai
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City    Area.Type    Point.of.Contact Median.Rent
  <chr>   <chr>        <chr>                  <dbl>
1 Chennai Carpet Area  Contact Agent          31500
2 Chennai Carpet Area  Contact Owner          14500
3 Chennai Super Area   Contact Agent          55000
4 Chennai Super Area   Contact Owner          14500

Median Rent for Delhi
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City  Area.Type    Point.of.Contact Median.Rent
  <chr> <chr>        <chr>                  <dbl>
1 Delhi Carpet Area  Contact Agent          28500
2 Delhi Carpet Area  Contact Owner          12500
3 Delhi Super Area   Contact Agent          35000
4 Delhi Super Area   Contact Owner          15000

Median Rent for Hyderabad
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City      Area.Type    Point.of.Contact Median.Rent
  <chr>     <chr>        <chr>                  <dbl>
1 Hyderabad Carpet Area  Contact Agent          42500
2 Hyderabad Carpet Area  Contact Owner          14500
3 Hyderabad Super Area   Contact Agent          38000
4 Hyderabad Super Area   Contact Owner          12416.

Median Rent for Kolkata
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City    Area.Type    Point.of.Contact Median.Rent
  <chr>   <chr>        <chr>                  <dbl>
1 Kolkata Carpet Area  Contact Agent          15500
2 Kolkata Carpet Area  Contact Owner           8000
3 Kolkata Super Area   Contact Agent          10000
4 Kolkata Super Area   Contact Owner           7500

Median Rent for Mumbai
# A tibble: 4 × 4
# Groups:   City, Area.Type [2]
  City   Area.Type    Point.of.Contact Median.Rent
  <chr>  <chr>        <chr>                  <dbl>
1 Mumbai Carpet Area  Contact Agent          46000
2 Mumbai Carpet Area  Contact Owner          26000
3 Mumbai Super Area   Contact Agent          55000
4 Mumbai Super Area   Contact Owner          23500
```

The graphs display table about the median rent by different lessors and cities and boxplots that showcase the distribution of rents for different area types within each city, considering different point of contact.

We summarize a typical value using the median as opposed to the mean when the value of the mean can be distorted by the outliers (*FAQs on Measures of Central Tendency - Mean, Mode and Median | Laerd Statistics*, n.d.). Therefore, the decision of using median as measure of central tendency has been made here. According to the median rent table, it is evident that

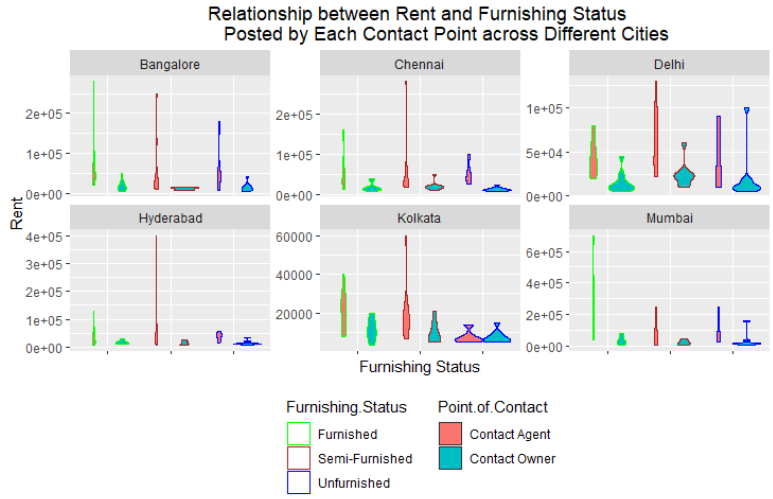
properties listed by agents with different area types tend to have higher median rents compared to those listed by owners across most cities. This consistent trend aligns seamlessly with our earlier analysis, which consistently highlighted the propensity of "Contact Agents" to list properties with elevated rents.

In addition, the boxplots reveal an intriguing aspect, which is the larger interquartile range (IQR) for properties listed by agents. This observation signifies a wider spread of rent values within the "Contact Agent" group. In essence, it reveals that the range of rents for properties listed by agents is more diverse than those listed by owners. On the contrary, the IQR for properties listed by owners presents a less discrete pattern.

The presence of outliers within various cities and area types is indicative of the diverse factors affecting rent prices. This insight underscores that rents are shaped by multiple variables beyond the type of person posting the property. Although the type of person posting the property does influence rent, it's vital to recognize that other factors play equally significant roles in determining rent values.

4.4.8 Analysis 8: What is the relationship between rent and furnishing status posted by each contact point across different cities?

```
sampling2 <- rental_data2 %>%
  group_by(City,Furnishing.Status,Point.of.Contact) %>%
  sample_n(10, replace = FALSE )

ggplot(sampling2, aes(x = Furnishing.Status, y = Rent, fill = Point.of.Contact,
                      color = Furnishing.Status)) +
  geom_violin() +
  facet_wrap(~ City, scales = "free_y") +
  ggtitle("Relationship between Rent and Furnishing Status
          Posted by Each Contact Point across Different Cities") +
  xlab("Furnishing Status") +
  ylab("Rent") +
  scale_color_manual(values = c("green", "brown", "blue")) +
  theme(legend.position = "bottom", legend.direction = "vertical",
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_blank())
```



The violin plots provide a visual representation of the relationship between rent, furnishing status, and the type of person posting the property across different cities. Notably, the widest parts of the violin plots for "Contact Agent" consistently exceed that of "Contact Owner" in most furnishing statuses across different cities, indicating where the data has the greatest frequency.

In addition, the IQR for the "Contact Agent" group is visibly larger, signifying a more diverse distribution of rents within this category. In contrast, the IQR for "Contact Owner" tends to be narrower, indicating a relatively more compact range of rent values within this category.

These findings strongly suggest that the involvement of "Contact Agent" as intermediaries in the property listings appears to wield a considerable influence. Their presence leads to a broader spectrum of rent values, implying that properties listed by agents cater to a wider array of preferences and market dynamics.

4.4.9 Extra Features

a) count(): use to count the how much times each unique value appears
b) labs(): set the title, as well as add labels to the x-axis and the y-axis of the plot
c) theme(): customize elements of the plots like the colors, fonts, legends, labels, and so on
d) theme_minimal(): remove the background elements and decorations of the plot
e) wilcoxon.test(): perform a Wilcoxon rank sum test to compare two paired groups
f) split(): divide a data frame into separate subsets based on the unique values in the specific column

g) geom_violin(): use to create a violin plot

4.4.10 Conclusion

In summary, the comprehensive analysis conducted on various aspects of rental properties posted by different lessors across different cities has yielded valuable insights into how the type of person posting a rental property influences its rent.

Across the majority of cities, there is a consistent trend emerges, which is rental properties listed by agents generally command higher average rents compared to those posted by owners. This pattern holds true across diverse scenarios, including variations in area type, furnishing status, property size, and even the BHK. These findings conform to the common perception that properties managed by agents often carry higher rental values due to the expertise they bring, as well as the effective marketing strategies and additional services.

It's important to note that while this analysis has identified a consistent trend, there are nuances within individual cities. In Mumbai, an exceptionally higher rent distribution unfolds, causing an exceedingly pronounced disparity compared to other cities. This is because most rental properties in Mumbai are posted by agents, contributing to an elevated average rent in comparison to other cities, thereby emphasizing the exceptional role that agents play in shaping the local rental scene.

The evidence gathered from the analyses and observations underscores the tangible influence of the type of person responsible for listing a rental property on its rent. The presence of contact agents typically leads to higher rental rates, with Mumbai offering a distinctive case where agent-listed properties dominate the rental market. This comprehensive analysis effectively supports and partially validates the initially proposed hypothesis.

5 Conclusion

As all the analysis had been done, the team has come to a conclusion where the hypothesis that is being stated at the beginning is only partially correct since the type of tenant preferred by the lessor is not one of the factors that influence the rental fees in Mumbai city. On the other side, the team also proved that properties which require for contacting agents will always have an average rent that is higher than contacting the owner in every city and the gap difference becomes relatively obvious in Mumbai. The teams decided to dive in further with other possible attributes in the dataset to find out the potential aspects that affect the rental fees in Mumbai even though the hypothesis is being proven.

After in-depth exploration, the teams find out that furnishing status of a properties become one of the causes that influence the rental fees in Mumbai and combining this fact with another statistical truth that Mumbai has the highest amount of furnished properties compared to other cities brings the team into concluding the higher amount of furnished properties in Mumbai is part of the reasons that makes the average rent of the properties in Mumbai becomes higher compared to other cities. Furthermore, the analysis shows that time is also one of the factors that change the average rent in Mumbai as it increases from month to month. Therefore, for tenant who is looking for properties to rent, he or she can start from looking the latest post to get a great deal.

It is worth mentioning that there are still other possible factors in the dataset the teams have not discover yet which may be a moderator, mediator, or even a confounder to any of the variables that the team used to interpret. As a conclusion for this project, the hypothesis has been proofed to be partially correct, and the group are also able to realize the complexity of relating variables together as the dimensions continue to increase and being taken into consideration for the analysis due to the increase of combination number between variables. Last but not least, the team have realized the importance of statistical testing instead of solely visualizing a graph as the testing parameters could become one of the beneficial tools in strengthening your proof or assumption.

6 Workload Matrix

	Name	Introduction	Data Import	Data Pre-Processing	Data Exploration	Conclusion
1.	SOONG YAU JOE	**25%**	**25%**	**25%**	**25%**	**25%**
2.	SOO JIUN GUAN	**25%**	**25%**	**25%**	**25%**	**25%**
3.	LIAN JUN ER	**25%**	**25%**	**25%**	**25%**	**25%**
4.	TEH YUE FENG	**25%**	**25%**	**25%**	**25%**	**25%**
TOTAL:		**100%**	**100%**	**100%**	**100%**	**100%**

7 Reference

FAQs on Measures of Central Tendency - Mean, Mode and Median | Laird Statistics. (n.d.). https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median-faqs.php

8 Appendix

Analysis 4.1.1

```
#     Which types of tenant are most preferred by the lessor in every cities?
# The code below shows the information of rent across different cities

# The code below is to assign distinct values of certain variable into a new variable
# for the use in for loop section
City_Name <- unique(rental_data2$City)
Tenant_Type <- unique(rental_data2$Tenant.Preferred)
```

```
# The code shows the total number of every type of tenant preferred across different cities
numb_tenant_acity <- select(rental_data2, c("Rent", "City", "Tenant.Preferred")) %>%
  group_by(City, Tenant.Preferred) %>%
  summarise(number=n(), .groups = 'drop')

ggplot(numb_tenant_acity, aes(x=Tenant.Preferred, y=number, fill=number))+
  geom_bar(stat = 'identity')+
  scale_fill_gradient(low = "beige", high = "maroon")+
  facet_wrap(~City)+
  geom_text(aes(label=number, vjust=-0.1))+
  ggtitle("Number of Different Tenant Preferred Type Across Cities")+
  theme(plot.title = element_text(face="bold", hjust = 0.5))+
  theme(text = element_text(family = "Courier"))
```

```
# The code below is the FUNCTION to calculate the percentage of different tenant preferred
# in a specific city
tenant_percentage <- function(numbOfCity, numbOfTenant)
{
  tenantOfCityTable <- filter(rental_data2, City == numbOfCity)
  tenantOfCityTable2 <- filter(rental_data2, City==numbOfCity & Tenant.Preferred==numbOfTenant)

  total_tenant <- nrow(tenantOfCityTable)
  specific_tenant <- nrow(tenantOfCityTable2)

  tenant_proportion <- round((specific_tenant / total_tenant) *100 , digits = 2)

  print(sprintf("%s --- %s", numbOfTenant, tenant_proportion))
}


# The code below displays the proportion of every tenant preferred across different cities
for (i in 1:length(City_Name))
{
  cat("\n")
  print(sprintf("PERCENTAGE OF EVERY TENANT PREFERRED IN %s ", City_Name[i]))
  for(j in 1:length(Tenant_Type))
  {
    tenant_percentage(City_Name[i], Tenant_Type[j])
  }
}
```

Analysis 4.1.2

```
avg_rent = select(rental_data2, c("Rent", "City")) %>% group_by(City) %>%
  summarise(AVG_RENT = round(mean(Rent), digits = 0) )

color_combination3 = c("#ffe1ed", "#ec9ca6", "#db598c", "#af6671", "#6b3c37", "#a45745")

ggplot(avg_rent, aes(x=City, y=AVG_RENT, ylab="Average Rent")) +
  geom_bar(stat = "identity", fill=color_combination3) +
  geom_text(aes(label=AVG_RENT, vjust=-0.05))+
  theme(axis.text.x = element_text(angle = -30, hjust = 1))+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))+
  ggtitle("Average Rent Across Cities")+
  theme(text = element_text(family = "Courier"))+
  theme(axis.text.x = element_text(angle = 30, vjust = 1, hjust = 1))
```

```
TP_vs_Rent <- rental_data2%>%select(c("City", "Tenant.Preferred","Rent"))%>%
  group_by(City, Tenant.Preferred)%>%
  summarise(AVG_RENT = mean(Rent), .groups = 'drop')

color_combination <- c("tan","peachpuff1", "moccasin","lightgoldenrod","sandybrown", "lightsalmon")

ggplot(TP_vs_Rent, aes(x=Tenant.Preferred, y=AVG_RENT, fill=City))+
  geom_bar(stat = 'identity')+
  facet_wrap(~City)+
  scale_fill_manual(values = color_combination)+
  geom_text(aes(label=round(AVG_RENT, digits = 2), vjust=-0.05, family="Courier"))+
  ggtitle("Average Rent Vs Preferred Tenant Across Cities")+
  theme(axis.text.x = element_text(angle = 30, vjust = 1, hjust = 1))+
  theme(text = element_text(family = "Courier"))+
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

```
TP_vs_Rent2 <- rental_data2%>%select(c("City", "Tenant.Preferred","Rent"))%>%
  group_by(City, Tenant.Preferred)

color_combination4 <- c("#f9e3c9", "#ccbcbf", "#b296b1", "#d799a9", "#ad4d8f", "#8a3c8a")

ggplot(TP_vs_Rent2, aes(x=Tenant.Preferred, y=Rent, fill=City))+
  geom_boxplot()+
  facet_grid(~City)+
  scale_fill_manual(values =  color_combination4)+
  theme(axis.text.x = element_text(angle = 30, hjust = 1))+
  ggtitle("Rent vs Tenant.Preferred Across Cities")+
  theme(text = element_text(family = "Courier"))+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```

```
# Data trimming to remove outliers based on the boxplot
# List out the top 10 highest rent on every city to help in data trimming
rental_data3 %>%
  select(c("Rent", "Tenant.Preferred", "City"))%>%
  filter(City=="Kolkata")%>%
  arrange(desc(Rent))%>%
  head(10)

# Trim by using filter
rental_data3 <- rental_data2 %>%
  filter(City!="Mumbai"| Rent<450000)%>%
  filter(City!="Chennai"| Rent<280000)%>%
  filter(City!="Chennai"| Rent<450000)%>%
  filter(City!="Chennai"| Rent<200000)%>%
  filter(City!="Bangalore"| Rent<250000)%>%
  filter(City!="Hyderabad"| Rent<200000)%>%
  filter(City!="Delhi"| Rent<250000)%>%
  filter(City!="Kolkata"| Rent<180000)
```

Analysis 4.1.3

```
# The function below is use to generate a table to fill in outliers data from a city
outliers_empty_table <- subset(rental_data3, Rent == 0)

outliers_table <- function(numbOfCity, emptyTable)
{
  city_table <- subset(rental_data3, City == numbOfCity, select = c(Rent, Tenant.Preferred))

  LB = quantile((city_table$Rent), probs = 0.25) - (1.5*IQR(city_table$Rent))
  UB = quantile((city_table$Rent), probs = 0.75) + (1.5*IQR(city_table$Rent))

  temp_outliers_table <- filter(rental_data2, (Rent<LB|Rent>UB) & City==numbOfCity)

  outliers_table <- rbind(emptyTable, temp_outliers_table)

}


# The for loop code below is to combine all the outliers from different cities together
for (i in 1:length(City_Name)){
  outliers_empty_table <- outliers_table(City_Name[i], outliers_empty_table)
}
```

```
# The code below will be use to display a bar chart of number vs Tenant.Preferred
outlier_tenant_preferred <- outliers_empty_table%>%
  select(c("City", "Tenant.Preferred", "Rent"))%>%
  group_by(Tenant.Preferred, City)%>%
  summarise(number=n(), .groups = 'drop')

color_combination2 <- c(colorspace::heat_hcl(6))

ggplot(outlier_tenant_preferred, aes(x=Tenant.Preferred, y=number, fill=City))+
  geom_bar(stat = 'identity')+
  facet_wrap(~City)+
  geom_text(aes(label=number, vjust=-0.1))+
  scale_fill_manual(values=color_combination2)+
  ggtitle("Total Number of Different Tenant Type Across Cities")+
  theme(text = element_text(family = "Courier"))+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))+
  ylab("Total")

# Trim dataset based on the previous analysis
outliers_empty_table <- outliers_empty_table %>%
  filter(City!="Mumbai"| Rent<450000)%>%
  filter(City!="Chennai"| Rent<280000)%>%
  filter(City!="Chennai"| Rent<450000)%>%
  filter(City!="Chennai"| Rent<200000)%>%
  filter(City!="Bangalore"| Rent<250000)%>%
  filter(City!="Hyderabad"| Rent<200000)%>%
  filter(City!="Delhi"| Rent<250000)%>%
  filter(City!="Kolkata"| Rent<180000)
```

Analysis 4.1.4

```
# Perform a Kruskal-Wallis Test to find out whether there is significance difference
# between the type of tenant preferred in terms of rent in different city
for (i in 1:length(City_Name)){
  kw_test_tenant_rent <- rental_data3%>%filter(City==City_Name[i])
  cat("\n")
  print(sprintf("City: %s", City_Name[i]))
  print((kruskal.test(Rent~Tenant.Preferred, kw_test_tenant_rent))$p.value)

}

kw_test_tenant_rent <- rental_data3

wilcoxon_pairtest_tenant_rent_mumbai <- kw_test_tenant_rent %>%
  filter(City == "Mumbai")

pairwise.wilcox.test(wilcoxon_pairtest_tenant_rent_mumbai$Rent,
                     wilcoxon_pairtest_tenant_rent_mumbai$Tenant.Preferred)
```

```
outlier_tenant_rent <- outliers_empty_table%>%
  select(c("City", "Tenant.Preferred", "Rent"))%>%
  group_by(City, Tenant.Preferred)%>%
  summarize(AVG_RENT = mean(Rent), .groups = 'drop')

color_combination5 <- c("#f7ecf9", "#dbd5fa", "#a5affb", "#878beb", "#636adc", "#3b52d7")

ggplot(outlier_tenant_rent, aes(x=Tenant.Preferred, y=AVG_RENT, fill=City))+
  geom_bar(stat = 'identity', color="black")+
  facet_wrap(~City)+
  scale_fill_manual(values = color_combination5)+
  geom_text(aes(label=round(AVG_RENT, digits = 2)), vjust=-0.5)+
  ggtitle("Average Rent of Outliers vs Preferred Tenant Type Across Cities")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5))+
  theme(text = element_text(family = "Courier"))+
  ylab("Average Rent") + xlab("Type of Preferred Tenant")
```

Analysis 4.1.5

```
#Plot the density curve of Rent based on BHK in Mumbai
rent_density_plot_mumbai <-rental_data3 %>%
  filter(City=="Mumbai")%>%
  mutate(BHK = as.factor(BHK))

plot_mumbai_rent_bhk_density<-ggplot(rent_density_plot_mumbai, aes(x=Rent, fill=BHK))+
  geom_density(alpha=0.25)+
  ggtitle("Density Plot for Rent of Different BHK in Mumbai")+
  theme(plot.title = element_text(hjust = 0.5), text = element_text(family = "Courier"))

plot_mumbai_rent_bhk_density
```

```
# The code below shows the frequency of post based on the number of BHK across different cities
fr_bhk_table <- rental_data3 %>%
  select(c("City", "Rent","BHK")) %>%
  group_by(City, BHK) %>%
  summarise(tot_number=n(), .groups = 'drop')

color_combination11 = c("#d3ebf8","#9bc2e5", "#6c80ca", "#48669d", "#1e284f", "#0f1331")

plot_for_BHK_Number<- ggplot(fr_bhk_table, aes(x=BHK, y=tot_number, fill=factor(BHK)))+
  geom_bar(stat='identity', color="black")+
  scale_fill_manual(values = color_combination11)+
  geom_text(aes(label=tot_number, vjust=-0.2))+
  facet_wrap(~City, ncol = 2)+
  ggtitle("Total Number vs BHK in Different Cities")+
  theme(plot.title = element_text(family = "Courier", colour = "darkblue", hjust = 0.5, face = "bold"))+
  labs(fill="BHK")
plot_for_BHK_Number
```

```
# Plot the density curve of Rent based on BHK for outliers table
rent_density_plot_outliers <- outliers_empty_table %>%
  mutate(BHK=as.factor(BHK))

plot_rent_bhk_density_outliers <-ggplot(rent_density_plot_outliers, aes(x=Rent, fill=BHK))+
  geom_density(alpha=0.25)+
  facet_wrap(~City)+
  ggtitle("Density Plot of Rent Based on BHK In Different Cities for Outliers Table")+
  theme(plot.title = element_text(hjust = 0.5), text = element_text(family = "Courier"))

plot_rent_bhk_density_outliers

ggplot(rent_density_plot_outliers%>%filter(City=="Mumbai"), aes(x=Rent, fill=BHK))+
  geom_density(alpha=0.25)+
  ggtitle("Density Plot of Rent Outliers Based on BHK In Mumbai")+
  theme(plot.title = element_text(hjust = 0.5), text = element_text(family = "Courier"))
```

```
# Conduct a Kruskal-Wallis Test to show whether there is a significance difference between BHK
# in Mumbai City in terms of rent
# H0: There is no significant difference between BHK interms of rent in Mumbai
# H1: There is a significant difference between BHK interms of rent in Mumbai
# Assumption of critical p-value will be 0.01
kw_test_bhk_rent_mumbai <- rental_data3%>%filter(City=="Mumbai")
kruskal.test(Rent~BHK, data = kw_test_bhk_rent_mumbai)
# Use pairwise wilcoxon test to find out which group combination have a significant difference
pairwise.wilcox.test(kw_test_bhk_rent_mumbai$Rent, kw_test_bhk_rent_mumbai$BHK)
```

```
# Boxplot to show BHK vs Rent
bhk_and_rent_boxplot <- rental_data3 %>% select(c("City", "BHK", "Rent")) %>%
  group_by(City, BHK)

bhk_and_rent_boxplot_except_mumbai <- rental_data3 %>% select(c("City", "BHK", "Rent")) %>%
  filter(City!="Mumbai")%>%
  group_by(City, BHK)

color_combination6 <- c("#e9e1cf", "#e0d19b", "#e7e5d3", "#d9be71", "#dab2a3", "#be7e4e")

ggplot(bhk_and_rent_boxplot, aes(x=BHK, y=Rent, fill=City))+
  geom_boxplot(aes(group=BHK))+
  scale_fill_manual(values = color_combination6)+
  facet_wrap(~City)+
  stat_summary(fun=mean, geom = "point", shape=20)+
  ggtitle("Rent vs BHK Across City")+
  theme(plot.title = element_text(face = "bold"), legend.position = "none")+
  theme(text = element_text(family = "Courier"))

ggplot(bhk_and_rent_boxplot_except_mumbai, aes(x=BHK, y=Rent, fill=City))+
  geom_boxplot(aes(group=BHK))+
  scale_fill_manual(values = color_combination6)+
  facet_wrap(~City, nrow = 1)+
  stat_summary(fun=mean, geom = "point", shape=20)+
  ggtitle("Rent vs BHK Across City Without Mumbai")+
  theme(plot.title = element_text(face = "bold"), legend.position = "none")+
  theme(text = element_text(family = "Courier"))
```

```
# Second round of data trimming
# List out the top 10 highest rent based on condition
rental_data3 %>% filter(City=="Kolkata" & BHK ==6) %>%
  arrange(desc(Rent)) %>%
  head(10)

# Trim using filter
rental_data3 <- rental_data3 %>%
  filter(City!="Hyderabad"| Rent!=130000 | BHK!=2) %>%
  filter(City!="Bangalore"| Rent!=70000 | BHK!=1) %>%
  filter(City!="Chennai"| BHK!=6) %>%
  filter(City!="Kolkata"| BHK!=6)  %>%
  filter(City!="Kolkata"| BHK!=5)  %>%
  filter(City!="Hyderabad"| BHK!=5)  %>%
  filter(City!="Delhi"| BHK!=5)

outliers_empty_table <- outliers_empty_table %>%
  filter(City!="Hyderabad"| Rent!=130000 | BHK!=2) %>%
  filter(City!="Bangalore"| Rent!=70000 | BHK!=1) %>%
  filter(City!="Chennai"| BHK!=6) %>%
  filter(City!="Kolkata"| BHK!=6)  %>%
  filter(City!="Kolkata"| BHK!=5)  %>%
  filter(City!="Hyderabad"| BHK!=5)  %>%
  filter(City!="Delhi"| BHK!=5)
```

```
# The code below shows the bar chart of average rent for different BHK which are
# being grouped by City
bhk_and_rent = select(rental_data3, c("Rent", "BHK", "City")) %>%
  group_by(City, BHK) %>%
  summarise(AVG_RENT=round(mean(Rent), digits = 0), .groups = 'drop')
bhk_and_rent

plot_for_bhk_avgRent<-ggplot(bhk_and_rent, aes(x=BHK, y=AVG_RENT, fill=AVG_RENT)) +
    geom_bar(stat = 'identity')+
    scale_fill_gradient(low='lightgreen', high='darkgreen')+
    facet_wrap(~City)+
    ggtitle("Average Rent VS BHK Across Cities")+
    theme(plot.title = element_text(hjust = 0.5, face = "bold"))+
    theme(panel.background = element_rect("ivory"), panel.grid.major = element_line("darkgrey", size = 0.1))+
    geom_text(aes(label=round(AVG_RENT, digits = 2), vjust=-0.5))
plot_for_bhk_avgRent
```

```
# The code below shows the frequency of post based on the number of BHK across
# different cities from the outliers table
outliers_table_BHK <- outliers_empty_table %>%
  select(c("City", "Rent", "BHK"))%>%
  group_by(BHK, City)%>%
  summarise(FREQ_BHK=n(), .groups='drop')
outliers_table_BHK

color_combination7 <- c("#cec4dd", "#c798a9", "#b47b8d", "#9c7161", "#6d5662", "#513a44")

plot_for_BHK_Number_outliers<-ggplot(outliers_table_BHK, aes(x=BHK, y=FREQ_BHK, fill=City))+
  geom_bar(stat = 'identity', color="black")+
  scale_fill_manual(values = color_combination7)+
  geom_text(aes(label=FREQ_BHK, vjust=-0.4))+
  facet_wrap(~City)+
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5, face = "bold"))+
  ggtitle("Frequency of BHK Across Cities of Outliers")+
  ylab("BHK Frequency")+
  theme(text = element_text(family = "Courier"))

plot_for_BHK_Number_outliers
```

Analysis 4.1.6

```
bhk_and_tenant <- rental_data3 %>%
  select(c("Tenant.Preferred","BHK", "City", "Rent"))%>%
  group_by(City, Tenant.Preferred, BHK)%>%
  summarise(tot_bhk=n(), .groups = 'drop')


plot_tenantnumb_vs_bhk <- ggplot(bhk_and_tenant, aes(x=BHK, y=tot_bhk, fill=Tenant.Preferred))+
  geom_bar(stat = 'identity')+
  facet_grid(Tenant.Preferred~City) +
  ggtitle("Frequency of BHK Based on Tenant Type Across Cities")+
  ylab("Total")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5),
        text = element_text(family = "Courier"),
        legend.position = "none")+
  geom_text(aes(label=tot_bhk), vjust=-0.1)


plot_tenantnumb_vs_bhk
```

Analysis 4.1.7

```
# The code below is use to plot a scatter plot along with regression line
# between size and rent across different cities
plot_for_size_rent <- ggplot(rental_data3, aes(x=Size, y=Rent, color=City))+
  geom_point(alpha=0.5)+
  stat_smooth(method = lm, color="black")+
  facet_wrap(~City)+
  ggtitle("Size vs Rent Across Cities")+
  theme(text = element_text(family = "Courier"),
        plot.title = element_text(hjust = 0.5, face = "bold"), legend.position = "none")

plot_for_size_rent
```

```
# Plot a scatter plot with regression line in Mumbai
dat_for_scatplot_mumbai <- filter(rental_data3, City=="Mumbai")

sum_test <- summary(lm(Rent~Size, dat_for_scatplot_mumbai))
sum_test

ggplot(dat_for_scatplot_mumbai, aes(x=Size, y=Rent))+
  geom_point(color="#e8a248")+
  stat_smooth(method = lm)+
  annotate("text", x=3500, y = 1000,label = "r^2==0.74", parse=TRUE, size=6, family="Courier")+
  ggtitle("Size vs Rent In Mumbai")+
  theme(plot.title = element_text(face = "bold", hjust = 0.5),
        text = element_text(family = "Courier"))
```

```
# Display all the coefficient of determination of rent vs size in every city
City <- c()
Coeff_Determination <-c(City=character(0), Coeff_Determination=numeric(0))

table_for_coeff_determination <- data.frame()

display_rsquare <- function(City_Name){
  City <- City_Name
  dat_for_scatplot <- filter(rental_data3, City==City_Name)
  rsquare_result <- summary(lm(Rent~Size, dat_for_scatplot))
  Coeff_Determination <-round(rsquare_result$r.squared, digits=2)
  combine_value <- data.frame(City=City, Coeff_Determination=Coeff_Determination)
  table_for_coeff_determination<<-rbind(table_for_coeff_determination, combine_value)

}

for (i in 1:length(City_Name)) {
  display_rsquare(City_Name[i])
}
```

Analysis 4.1.8

```
color_combination8 <- c("#f4dce1", "#e6a9ae", "#efaeaa", "#d5a190", "#a86d50", "#ba815b")

ggplot(rental_data3, aes(x=BHK, y=Size, fill=City))+
  geom_boxplot(aes(group=BHK))+
  facet_wrap(~City)+
  ggtitle("Boxplot for Size vs BHK")+
  scale_fill_manual(values = color_combination8)+
  theme(plot.title = element_text(hjust = 0.5, face = "bold", colour = "#16281d"),
        text = element_text("Courier"),
        legend.position = "none")+
  stat_summary(fun = mean, geom = "point", shape=20, color="yellow")

# Third round of data trimming
# List out the top 10 highest rent based on condition
rental_data3 %>% filter(City=="Kolkata", BHK==3) %>%
  arrange((Size)) %>%
  head(10)

# Trim using filter
rental_data3 <- rental_data3 %>%
  filter(City!="Kolkata"| Size<3500) %>%
  filter(City!="Hyderabad"| Size!=100 | BHK!=4)%>%
  filter(City!="Mumbai"| Size!=3700 | BHK!=4) %>%
  filter(City!="Kolkata"| Size!=120 | BHK!=3)

outliers_empty_table <- outliers_empty_table %>%
  filter(City!="Kolkata"| Size<3500) %>%
  filter(City!="Hyderabad"| Size!=100 | BHK!=4)%>%
  filter(City!="Mumbai"| Size!=3700 | BHK!=4) %>%
  filter(City!="Kolkata"| Size!=120 | BHK!=3)

# Perform pairwise wilcoxon test to test out the relation between size and bhk in Mumbai
wilcoxon_test_size_bhk_mumbai<-rental_data3%>%filter(City=="Mumbai")
pairwise.wilcox.test(wilcoxon_test_size_bhk_mumbai$Size, wilcoxon_test_size_bhk_mumbai$BHK)
```

Analysis 4.1.9

```
color_combination9 <- c("#f6ae94", "#c55745", "#8d251d")

ggplot(rental_data3, aes(x=Tenant.Preferred, y=Size, fill=Tenant.Preferred))+
  geom_violin()+
  scale_fill_manual(values = color_combination9)+
  facet_wrap(~City)+
  ggtitle("Violin Plot for Size Vs Type of Tenant")+
  xlab("Type of Tenant")+
  theme(text = element_text("Courier"),
        plot.title = element_text(face = "bold", color = "brown", hjust = 0.5),
        legend.position = "none")
```

```
# Conduct hypothesis testing to find out whether there are any difference in terms of rent
# between different tenant group in Mumbai
# Assume the critical p-value to be 0.05
kruskal.test(Size~Tenant.Preferred, data=rental_data3%>%filter(City=="Mumbai"))
pairwise.wilcox.test((rental_data3%>%filter(City=="Mumbai"))$Size,
                     (rental_data3%>%filter(City=="Mumbai"))$Tenant.Preferred)
```

Analysis 4.1.10

```
# Find out the total number of different furnishing status in different cities

numbOfFurnishingStatus <- select(rental_data3, c("City", "Furnishing.Status", "Tenant.Preferred")) %>%
  group_by(City, Furnishing.Status) %>%
  summarise(number=n(), .groups='drop')
numbOfFurnishingStatus


totNumbInCity <- rental_data3 %>% select(c("City", "Furnishing.Status")) %>%
  group_by(City) %>%
  summarise(tot_number=n(), .groups='drop')
totNumbInCity

joint_table <- (right_join(numbOfFurnishingStatus, totNumbInCity)) %>%
  mutate(percentage = round((number/tot_number)*100, digits = 2))


ggplot(joint_table, aes(x=City, y=number, fill=Furnishing.Status)) +
  geom_col(width=0.7, position="dodge")+
  scale_fill_manual(values = c("chocolate4", "darkgoldenrod", "bisque3"))+
  geom_text(aes(label=percentage), position=position_dodge(0.8), vjust=-0.5)
```

```
# Find out whether most of the outliers in Mumbai comes from fully furnished house.
outlier_numb_vs_furnish_status <-
  outliers_empty_table %>%
  group_by(City, Furnishing.Status)%>%
  summarise(numb=n(), .groups = 'drop')

color_combination10 <- c("#a6ddf4", "#5ec4ef", "#1a9ae0", "#146bdc", "#003986", "#081559")

ggplot(outlier_numb_vs_furnish_status, aes(x=Furnishing.Status, y=numb, fill=City))+
  geom_bar(stat = 'identity', color="black")+
  scale_fill_manual(values = color_combination10)+
  facet_wrap(~City)+
  geom_text(aes(label=numb, vjust=-0.2))+
  ggtitle("Frequency of Furnishing Status Across Cities for Outliers")+
  theme(legend.position = "none",
        text = element_text(family = "Courier"),
        plot.title = element_text(hjust = 0.5, face = "bold"))
```

```
# Find the relation between rental fees and furnishing status across different city
ggplot(rental_data3, aes(x=Furnishing.Status, y=Rent, fill=Furnishing.Status))+
  geom_boxplot(aes(group=Furnishing.Status))+
  facet_grid(~City)+
  stat_summary(fun = mean, geom = "point", shape=20, color="yellow")+
  theme(axis.text.x = element_blank(),
        plot.title = element_text(face = "bold", hjust = 0.5),
        text = element_text(family = "Courier"))+
  ggtitle("Rent vs Furnishing Status Across Cities")
```

```
# By focusing in Mumbai city, conduct a test to test out whether there is a significance difference
# between furnishing status in terms of rent
# H0: There is no significance different between furnishing status in terms of rent
# H1: There is significance different between furnishing status in terms of rent
data_for_test_furnish_vs_rent_mumbai <- rental_data3%>%filter(City=="Mumbai")

# Conduct Kruskal-Wallis Test and Post-hoc test
kruskal.test(Rent~Furnishing.Status, data=data_for_test_furnish_vs_rent_mumbai)

dunn.test(data_for_test_furnish_vs_rent_mumbai$Rent,
          g=data_for_test_furnish_vs_rent_mumbai$Furnishing.Status)
```