Brazilian Houses to Rent Data Analysis

Case Study I

K.K.D.S.N.Gunathilaka – s15355

Abstract

This study is based on which contains data on houses to rent in different cities in Brazil. This study centers around an expansive dataset cataloged as "Brazilian_Houses_to_Rent," encompassing a diverse array of information about rental properties across various cities in Brazil. The primary objective is to address the Stakeholders interested in the Brazilian real estate market and those looking to buy Brazilian homes can draw meaningful conclusions here. Exploring this allows for a fine-grained understanding of the factors influencing rental prices and property preferences across different regions of Brazil.

This study focused on 5 main objectives. They are,

- Does the total rent amount significantly affect the floor number?
- Does the house rental amount significantly affect the keeping of a pet or not?
- Does Fire insurance significantly affect whether the house has furniture or not?
- Does the city in which the property is significantly affect the total rent amount?
- Does property tax relate to the area of the property?

The study was carried out to achieve these objectives.
For the Analysis, R Software and Its Output are used here.

The results of this study, we come up with 5 conclusions as follows,

- ✓ Total Rent Amount Significantly affected to the floor number.
- ✓ House Rental Amount Significantly affect keeping of a pet or not.

  (The House Rental Amount for keeping a pet is greater than according to the doesn't keep a pet)
- ✓ Fire insurance significantly affects whether the house has furniture or not.

  (Fire insurance is higher for furnished houses as compared to unfurnished houses.)
- ✓ The city in which the property is significantly affects the total rent amount.
- ✓ Property tax relate to the area of the property.

  (There is Positive correlation ship between Property tax and the Area.)

In conclusion, this analysis offers valuable insights into the Brazilian housing rental market, emphasizing the significance of specific property attributes and regional dynamics in determining rental prices. These findings can assist prospective renters, property owners, and stakeholders in making informed decisions regarding pricing strategies, property investments, and market assessments within the Brazilian real estate landscape.

Contents

List of Figures

Table of Figures

Introduction

This focuses on a big collection of information about houses for rent in different cities in Brazil. It's mainly for people interested in the real estate market in Brazil and those thinking about buying a home there. The goal of this report is to give an overview and analysis of this information. By doing this, it hopes to give useful insights and conclusions that can help people involved in the real estate business and those planning to buy a house in Brazil.

This dataset offers valuable insights into the rental landscape, enabling analysis and exploration of trends, patterns, and correlations within the Brazilian housing market. It comprises structured data fields, including but not limited to property type, area size, number of rooms, parking availability, and associated costs. Amidst this dynamic landscape, this report emerges as a comprehensive repository, delving into an extensive dataset encompassing diverse rental properties across various cities in Brazil. This study is grounded in comprehensive data analysis encompassing various aspects of houses available for rent across diverse cities in Brazil.

According to the following objectives, this analysis was conducted.

Objectives

- Does the total rent amount significantly affect the floor number?
- Does the house rental amount significantly affect the keeping of a pet or not?
- Does Fire insurance significantly affect whether the house has furniture or not?
- Does the city in which the property is significantly affect the total rent amount?
- Does property tax relate to the area of the property?

There are various statistical methods and graphical representations to explore how different aspects affect the rental housing market, as mentioned in our objectives. By using these statistical methods and visual aids, it aims to uncover meaningful insights into how specific features or characteristics influence the rental housing market in different cities across Brazil.

Theory and Methodology

K-S Test for One Sample

Hypothesis: H0: The sample follows a normal distribution.

H1: The sample does not follow a normal distribution.

Test Statistic:

The K-S test statistic (D) measures the maximum absolute difference between the empirical cumulative distribution function (ECDF) of the sample data and the cumulative distribution function (CDF) of the expected normal distribution.

Decision Rule:

If the calculated K-S test statistic (D) is greater than the critical value at a chosen significance level (e.g., $\alpha = 0.05$), reject the null hypothesis.

If the calculated K-S test statistic (D) is less than or equal to the critical value at the chosen significance level, fails to reject the null hypothesis.

p-value is greater than the chosen significance level (e.g., $\alpha = 0.05$), and fails to reject the null hypothesis.

Conclusion:

If the null hypothesis is rejected, it suggests that the sample data does not follow a normal distribution.

If the null hypothesis is not rejected, it indicates that there is insufficient evidence to conclude that the sample data significantly deviates from a normal distribution at the chosen significance level.

Q-Q Plot for Check Normality

If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.

Wilcoxon’s Rank Sum Test

Hypothesis:

Suppose we have independent samples from 2 populations A and B.
H0: There is no difference in location between the two populations A and B ($M_A = M_B$)

The alternative hypothesis may take one of the following forms.
H1: The location of population A is different from the location of population B ($M_A \neq M_B$)
H1: The location of population A is greater than the location of population B ($M_A > M_B$)
H1: The location of population A is less than the location of population B ($M_A < M_B$)

Method:
The test statistic of this test is computed by adding together ranks of the pooled data set. The rank sum statistic is either R1 or R2 as defined in the Mann-Whitney test.

Test Statistic:
The test statistic 'R' is computed depending on the direction of the alternative hypothesis. If the alternative hypothesis is two-sided, 'R' is the observed smallest of R1 and R2. If the alternative hypothesis is one-sided, 'R' is the expected smallest of R1 and R2 when H1 is true.

Decision Rule:
The critical region is of the form, reject H0 if R ≤ critical value.

If the p-value is less than α (e.g., α = 0.05), reject the null hypothesis. If the p-value is greater than or equal to α, fail to reject the null hypothesis.

Conclusion:

If the null hypothesis is rejected, it indicates that there is evidence to suggest that the distributions of the two samples are significantly different.

If the null hypothesis is not rejected, it suggests that there is not enough evidence to conclude that the distributions of the two samples are different.

Kruskal Wallis Test

This test is useful in deciding whether k-independent samples are drawn from different populations.

Hypothesis:
H0: There is no difference in location between the populations from which the k samples have been drawn.
H1: There is a difference in location in one or more populations

Method:
1. Pool all scores for all the k samples.
2. Rank the pooled scores, the smallest score is ranked 1, the next smallest is ranked 2, and so on, and assign the average rank for tied observations.
3. Calculate the sum of the ranks in each sample.

Test Statistic:
Let Ri (i=1,2,. . . , k) denote the sum of the ranks associated with the observations drawn from the i^{th} population. The K-W statistic is given by,

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \left[\frac{R_i^2}{n_i} \right] - 3(N+1)$$

Figure 1: Equation of Kruskal Wallis Test

Decision Rule
Reject H_0 if, H≥ $X^2_{(n-1)}$critical value.

If the p-value is less than the chosen significance level (e.g., α = 0.05): Reject the null hypothesis.

Conclusion:

Reject the null hypothesis Conclude that there is sufficient evidence to suggest that at least one group differs significantly from the others in terms of their central tendencies (medians).

Spearman Correlation Test

Hypothesis:
H0: There is no correlation between two variables.
The alternative hypotheses could be of the following forms;
(i) H1: There is some correlation (either positive or negative) between two variables.
(ii) H1: There is a positive correlation between two variables.
(iii) H1: There is a negative correlation between two variables.

Method:
1. Rank the data of the 2 variables X and Y separately. i.e. compute r(Xi) and r(Yi) for i=1,2,…,n.
2. Then compute the pairwise difference in ranks. i.e., compute di = r(Xi)- r(Yi).

Test Statistic:
If the two variables are positively correlated, then the rankings will also be similar and the rank differences (di) will tend to be small in size. The strongest indication of positive correlation is when the two sets of ranks are identical, in which case all the rank differences will be zero. Now compute;
D2 = $d_1^2 + d_2^2 + d_3^2 + \cdots + d_n^2$
D2 could be directly used as a test statistic. However, as the correlation coefficients are usually defined to be between -1 and +1, we define Spearman's rank correlation coefficient as follows.

$$r_s = 1 - \frac{6D^2}{n^3 - n}$$

Therefore, when D2 = Minimum value = 0, i.e., the strongest evidence of positive correlation then r_s=+1
When D2= Maximum value = (n3 – n)/3, i.e., the strongest evidence of negative correlation then r_s=-1.
Otherwise, r_s lies between –1 and +1.

Decision Rule:
The one-sided alternative hypothesis of positive correlation is supported by values of rs near +1. The required critical region is of the form;
$r_s \geq$ critical value
If we are testing for negative correlation, the critical region is of the form;
$r_s \leq$-critical value
For two sided tests, the critical region is of the form;
$|r_s| \geq$critical value

Data

Consider the dataset "*Brazilian_Houses_to_Rent.xlsx*" which contains data on houses to rent in different cities in Brazil. In this data set, there are 10693 observations. This contains 13 variables. which are 4 quantitative variables and 9 qualitative variables to describe the observations. city, floor, keeping animals, and furniture are the four quantitative variables. All the other variables are Qualitative variables.

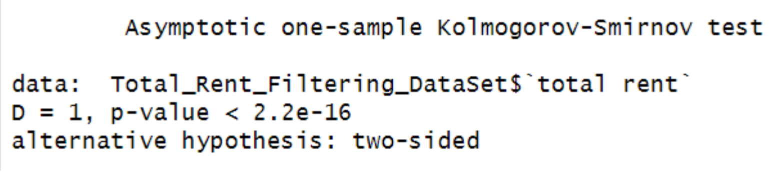
Table 1: Variable Description table

Variable Name	Description
city	The city where the property is located
area	Property area
rooms	Quantity of rooms
bathroom	Quantity of bathrooms
parking spaces	Quantity of rooms
floor	The floor where the property is
keeping animal	Whether pets are allowed or not [If accept animals (1) or not (0)]
furniture	Whether furnished or not [If it is furnished (1) or not (0)]
hoa	Homeowners' association tax (R$)
rent amount	Rent amount (R$)
property tax	Municipal property tax (R$)
fire insurance	Fire insurance value (R$)
total rent	The sum of all values (R$)

Data Analysis

Object 1: Check does the total rent amount significantly affect the floor number.

First, check whether the Total rental amount is normally distributed or not by using the Kolmogorov – Smirnov test.

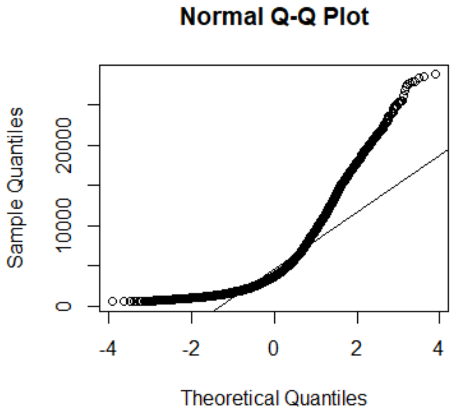


```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Total_Rent_Filtering_DataSet$`total rent`
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 2:K-S Test for Total Rent

According to the above theory and methodology of the K-S test. Here, p-value (2.2×10^{-16}) < Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It Shows, that the sample does not follow a normal distribution.



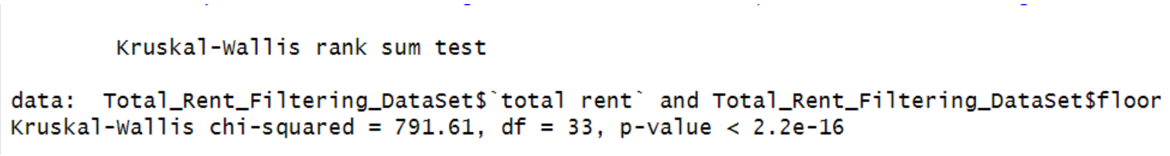
By using this Normal Q-Q plot the points do not fall on the 45-degree reference line. Then the total rent data are not normally distributed, the points will deviate from the reference line.

Therefore, considering the above result we cannot use the Parametric test for in here. Then go through a non-parametric test.

Figure 3:Normal Q-Q plot for Total rental data

By using the Kruskal Wallis Test Check the above objective.

Hypothesis:
H0: There is no difference in location between the populations from which the k samples have been drawn.
H1: There is a difference in location in one or more populations from which the k samples have been drawn.
Where, k=floor numbers (7, 20, 6, 2, 1, 0, 4, 3, 10, 11, 24, 9, 8, 17, 18, 5, 13, 15, 16, 14, 26, 12, 21, 19, 22, 27, 23, 35, 25, 46, 28, 29, 51, 32)



```
        Kruskal-Wallis rank sum test

data:  Total_Rent_Filtering_DataSet$`total rent` and Total_Rent_Filtering_DataSet$floor
Kruskal-Wallis chi-squared = 791.61, df = 33, p-value < 2.2e-16
```

Figure 4:Kruskal Wallis test for Total rent in specific floors

Here P-value (2.2×10^{-16}) $<$ Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It can be determined there is a difference in location in one or more populations from which the k samples have been drawn.

Then it shows that the total rent amount impacts the floor number. Figure 5 and Figure 6 show that in the sample of data median values of the Total rent on different floors are different from each other.

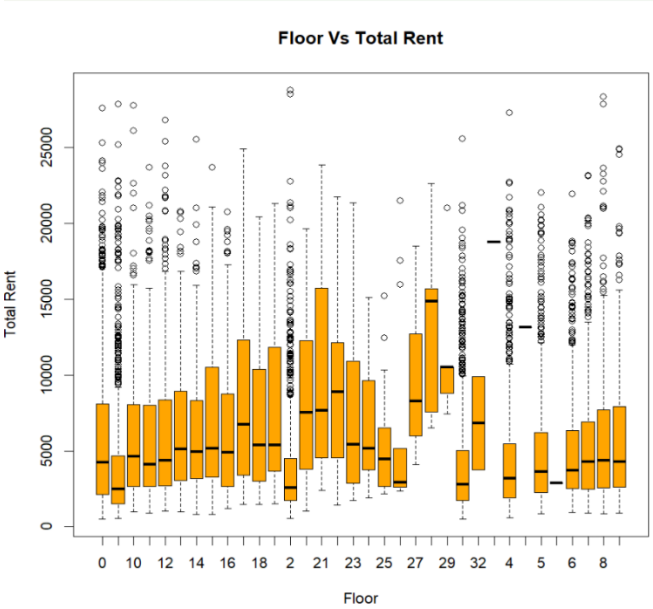
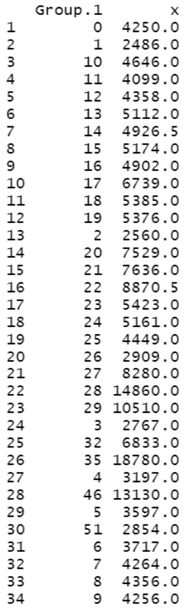


Figure 5:Box Plots for Floor Vs Total Rent

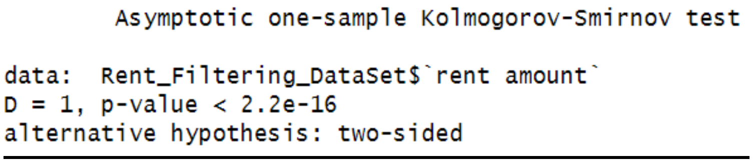


```
   Group.1       x
1        0  4250.0
2        1  2486.0
3       10  4646.0
4       11  4099.0
5       12  4358.0
6       13  5112.0
7       14  4926.5
8       15  5174.0
9       16  4902.0
10      17  6739.0
11      18  5385.0
12      19  5376.0
13       2  2560.0
14      20  7529.0
15      21  7636.0
16      22  8870.5
17      23  5423.0
18      24  5161.0
19      25  4449.0
20      26  2909.0
21      27  8280.0
22      28 14860.0
23      29 10510.0
24       3  2767.0
25      32  6833.0
26      35 18780.0
27       4  3197.0
28      46 13130.0
29       5  3597.0
30      51  2854.0
31       6  3717.0
32       7  4264.0
33       8  4356.0
34       9  4256.0
```

Figure 6: Medium Value according to the sample of floor number

Object 2: Check does the house rental amount significantly affect the keeping of a pet or not.

First, check whether the rental amount is normally distributed or not by using the Kolmogorov – Smirnov test.



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Rent_Filtering_DataSet$`rent amount`
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 7: K-S test for Rental amount

According to the above theory and methodology of the K-S test. Here, p-value (2.2$\times 10^{-16}$) < Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It Shows, the data of the rental amount sample does not follow a normal distribution.

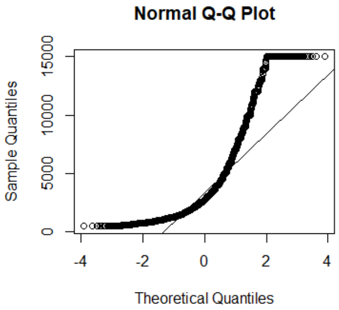


Figure 8: Normal Q-Q plot for Rental amount

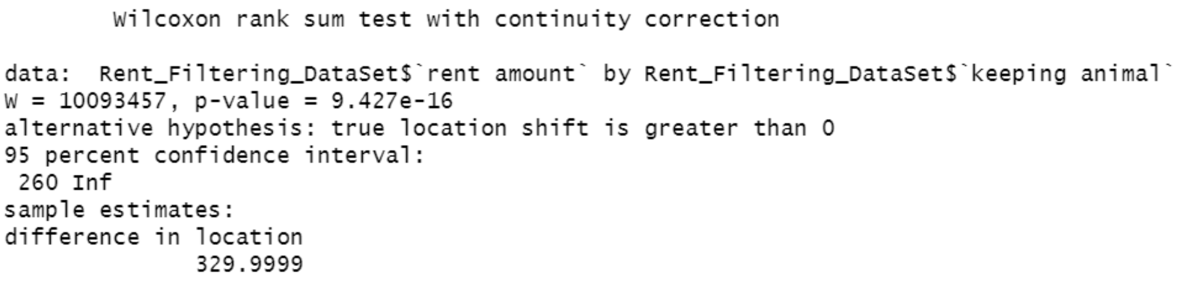
By using this Normal Q-Q plot the points do not fall on the 45-degree reference line. Then the total rent data are not normally distributed, the points will deviate from the reference line.

Therefore, considering the above result we cannot use the Parametric test for in here. Then go through a non-parametric test

By using the Wilcoxon's Rank Sum Test Check the above objective.

H0: There is no difference in location between the two populations of Rent amount whether keeping animals and not keeping animal

H1: The location of the two populations of Rent amount when not keeping an animal is less than the location of the population of Rent amount whether keeping an animal.



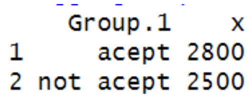
```
        Wilcoxon rank sum test with continuity correction

data:  Rent_Filtering_DataSet$`rent amount` by Rent_Filtering_DataSet$`keeping animal`
W = 10093457, p-value = 9.427e-16
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 260 Inf
sample estimates:
difference in location
              329.9999
```

Figure 9: Wilcoxon rank sum test for a rental amount according to the keeping animal or not

Here, P-value (9.427$\times 10^{-16}$) < Significance value (0.05). Therefore, reject Ho at 5% level of significance. Therefore have enough evidence to conclude that the location of the two populations of Rent amount when not keeping an animal is less than the location of the population of Rent amount whether keeping an animal.

Figure 10 and Figure 11 clearly show that the median value of the Rent value of the data set Keeping animal accept median value is greater than the keeping animal not accept median value. The Figure 11 box plot also shows that the median line of the accepted animal is higher than the median value of the not accepted animal. By the way, the Minimum value and Maximum value of the not-accepted animal are also less than the Minimum value and Maximum value of the accepted animal. The range of the Rent amount when accepting animals is also greater than the Rent amount when not accepting animals.



```
    Group.1    x
1     acept 2800
2 not acept 2500
```

Figure 10: Sample Medium values in rental amount grouping by whether the keeping animal or not

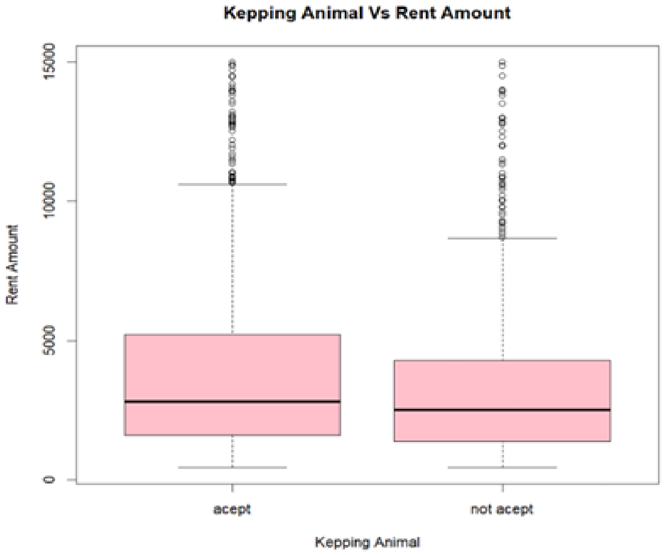
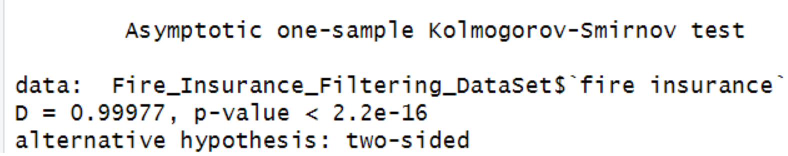


Figure 11:Boxplot for Keeping Animal Vs Rent Amount

Object 3: Check does Fire insurance significantly affect whether the house has furniture or not.

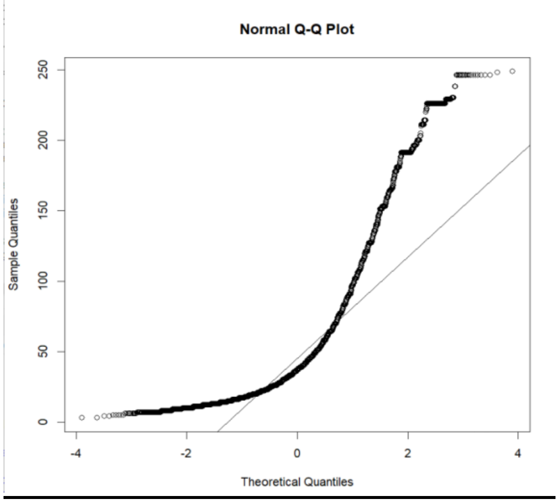
First, check whether the Fire Insurance is normally distributed or not by using the Kolmogorov – Smirnov test.



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Fire_Insurance_Filtering_DataSet$`fire insurance`
D = 0.99977, p-value < 2.2e-16
alternative hypothesis: two-sided
```

According to the above theory and methodology of the K-S test. Here, p-value (2.2×10^{-16}) < Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at 5% level of Significance. It Shows, that the data of the rental amount sample does not follow a normal distribution



By using this Normal Q-Q plot the points do not fall on the 45-degree reference line. Then the total rent data are not normally distributed, the points will deviate from the reference line.

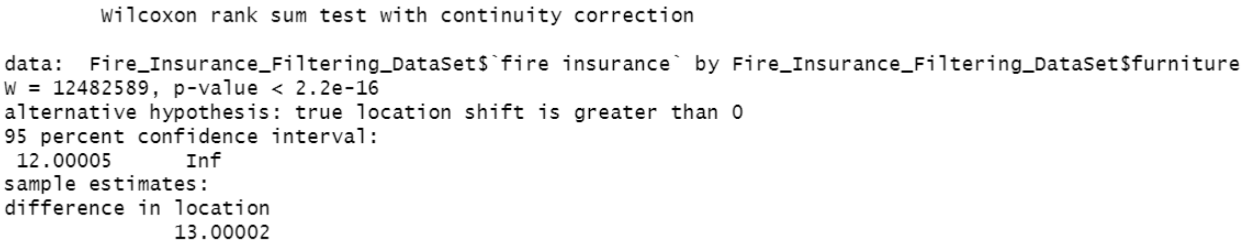
Therefore, considering the above result we cannot use the Parametric test for in here. Then go through a non-parametric test

Figure 12: Normal Q-Q plot for Fire Insurance Data

By using the Wilcoxon's Rank Sum Test Check the above objective.

H0: There is no difference in location between the two populations of Fire Insurance whether the house is furnished or not

H1: The location of the two populations of Fire Insurance when a house is not furnished is less than the location of the population of Fire Insurance when whether house is furnished.



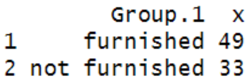
```
        Wilcoxon rank sum test with continuity correction

data:  Fire_Insurance_Filtering_DataSet$`fire insurance` by Fire_Insurance_Filtering_DataSet$furniture
W = 12482589, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 12.00005      Inf
sample estimates:
difference in location
              13.00002
```

Figure 13: Wilcoxon rank sum test for Fire Insurance according to the furnished or not

Here, P-value (2.2×10^{-16}) < Significance value (0.05). Therefore, reject Ho at 5% level of significance. Therefore, have enough evidence to conclude the location of the two populations of Fire Insurance when a house is not furnished is less than the location of the population of Fire Insurance when whether house is furnished.

Figure 14 and Figure 15 clearly show that the median value of the Fire Insurance of the data set Furnished house median value is greater than the data set Not Furnished house median value. The Figure 15 box plot also shows that the median line of the furnished house Fire Insurance is higher than the median value of the not furnished house Fire Insurance. By the way, the Minimum value and Maximum value of the not furnished house Fire Insurance are also less than the Minimum value and Maximum value of the furnished house Fire Insurance. The range of the Fire Insurance when the Furnished house is also greater than the range of the Fire Insurance when not Furnished house.



```
        Group.1  x
1     furnished 49
2 not furnished 33
```

Figure 14:Sample Median values for fire insurance

grouped by Furnished or not

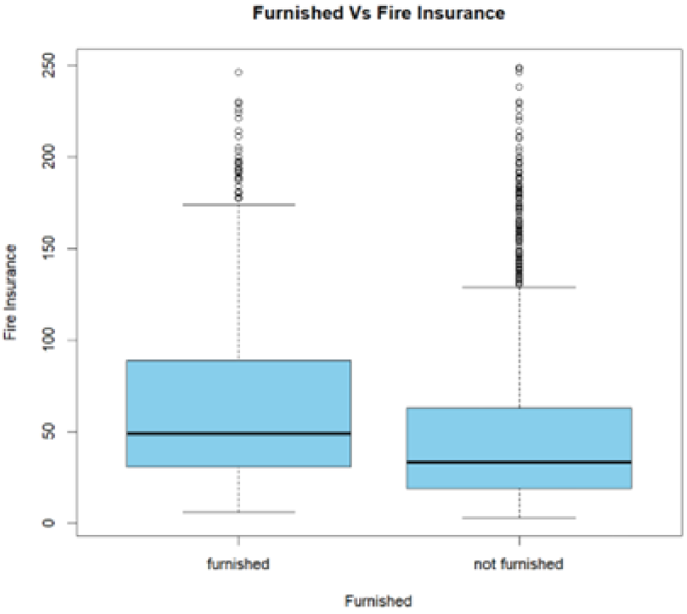


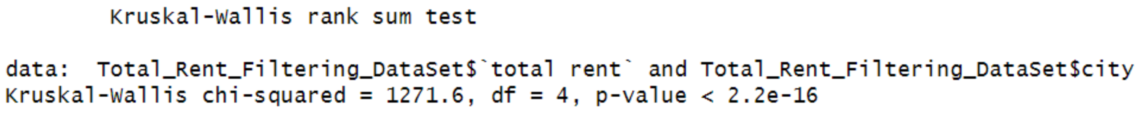
Figure 15: Boxplot for Furnished Vs Fire Insurance

Object 4: Check does the city in which the property is significantly affect the total rent amount.

By using the Kruskal Wallis Test Check the above objective 4.

Hypothesis:

H0: There is no difference in location between the populations from which the k samples have been drawn.
H1: There is a difference in location in one or more populations from which the k samples have been drawn.
Where, k=(Belo Horizonte city, Campinas city, Porto Alegre city, Rio de Janeiro city, Saeo Paulo city)



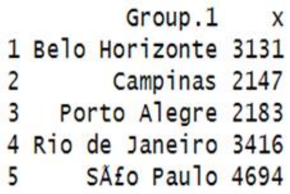
```
        Kruskal-Wallis rank sum test

data:  Total_Rent_Filtering_DataSet$`total rent` and Total_Rent_Filtering_DataSet$city
Kruskal-Wallis chi-squared = 1271.6, df = 4, p-value < 2.2e-16
```

Figure 16: Kruskal Wallis test for total rent grouping by the city

Here P-value (2.2×10^{-16}) $<$ Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It can be determined there is a difference in location in one or more populations from which the k samples have been drawn. Then it shows that the city in which the property impacts the total rent amount.

Figure 17 and Figure 18 show that in the sample of data median values of the rent amount in different cities are different from each other. In the Figure 18 box plots, median lines are different from each other, and the range of the different box plots is different from each other. Also, Figure 17 and Figure 18 indicate the Median value of the total rent amount Saeo Paulo city > Rio de Janeiro city>Belo Horizonte city>Porto Alegre city>Campinas city.



```
          Group.1    x
1 Belo Horizonte 3131
2       Campinas 2147
3   Porto Alegre 2183
4 Rio de Janeiro 3416
5     SÃ£o Paulo 4694
```

Figure 17: Sample Medium values for Total rent grouping according to the city

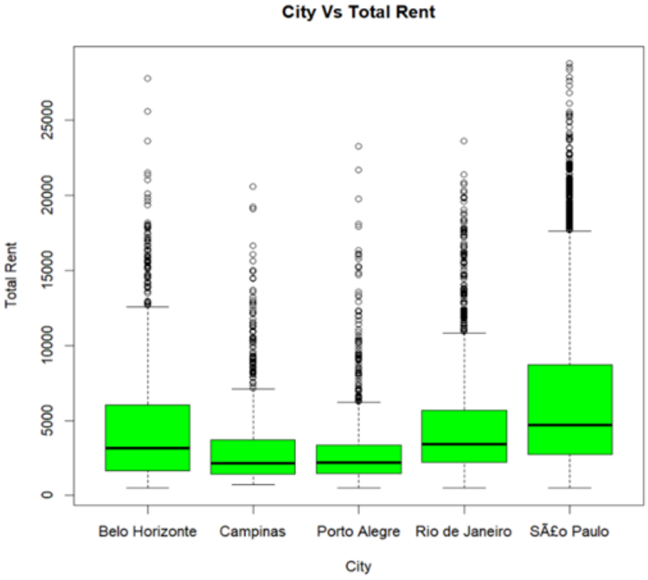
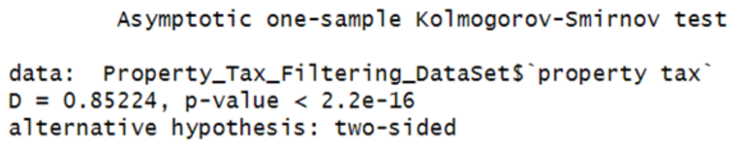


Figure 18: Box plot for City vs. Total Rent

Objective 5: Check does property tax relate to the area of the property.

First, check whether the Property tax data are normally distributed or not by using the Kolmogorov – Smirnov test.



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Property_Tax_Filtering_DataSet$`property tax`
D = 0.85224, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 19: K-S Test for Property Tax

According to the above theory and methodology of the K-S test. Here, p-value (2.2$\times 10^{-16}$) < Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It Shows, that the data of the property tax data sample does not follow a normal distribution

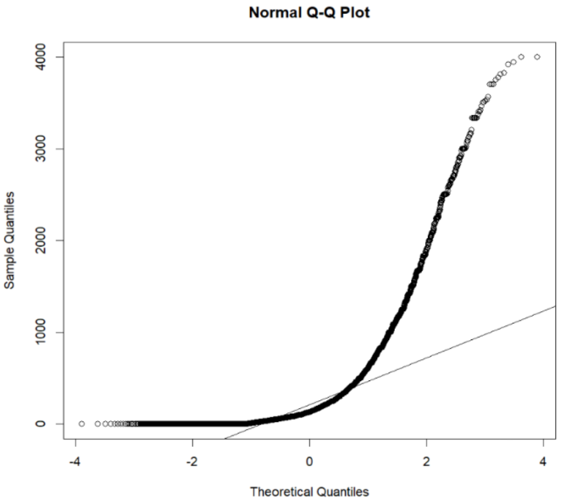
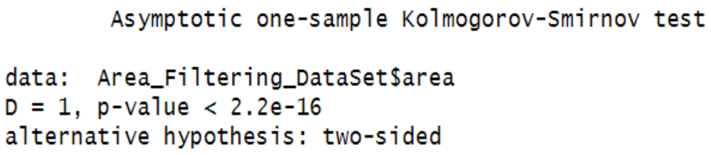


Figure 20: Normal Q-Q plot for property tax

By using this Normal Q-Q plot the points do not fall on the 45-degree reference line. Then the total rent data are not normally distributed, the points will deviate from the reference line.

Therefore, considering the above result we cannot use the Parametric test for in here. Then go through a non-parametric test

Secondly, check whether the Area data are normally distributed or not by using the Kolmogorov – Smirnov test.



```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Area_Filtering_DataSet$area
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 21: K-S Test for Area

According to the above theory and methodology of the K-S test. Here, p-value (2.2× 10^{-16}) < Significance value ($\alpha = 0.05$). Therefore, Reject H_0 at a 5% level of Significance. It Shows, that the data of the Area data sample does not follow a normal distribution.

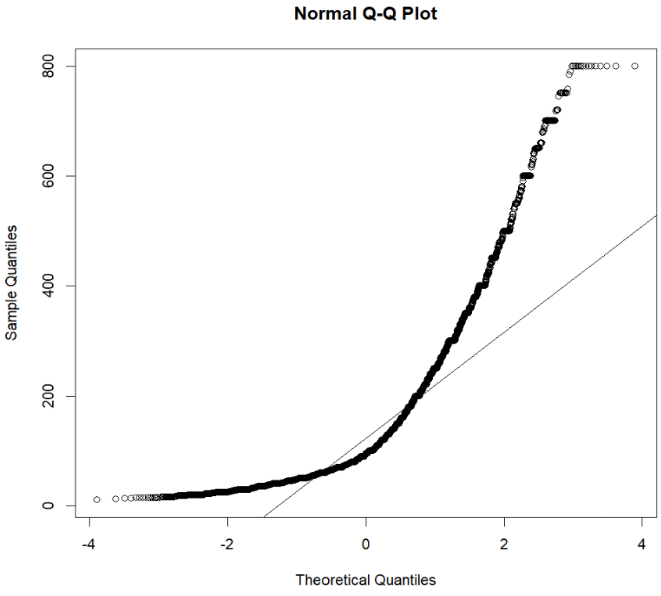


Figure 22: Normal Q-Q plot for area

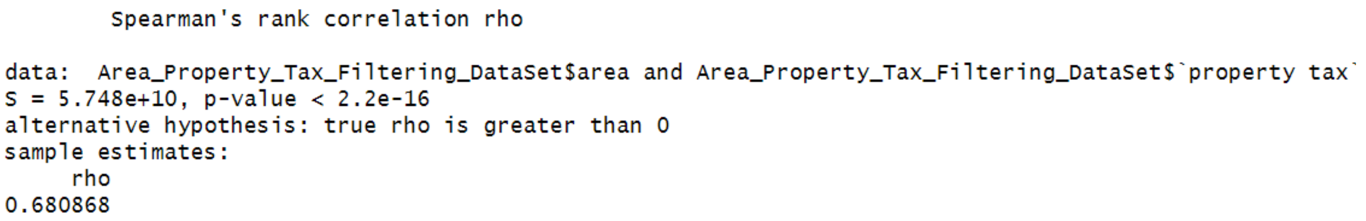
By using this Normal Q-Q plot the points do not fall on the 45-degree reference line. Then the total rent data are not normally distributed, the points will deviate from the reference line.

Therefore, considering the above result we cannot use the Parametric test for in here. Then go through a non-parametric test

In this case have to use a non-parametric correlation test for testing the correlation between Property tax and Area. In this case can use Spearman's rank Correlation test to test the correlation between Property tax and Area.

Hypothesis:

H0: There is no correlation between Property tax and Area.
H1: There is a positive correlation between Property tax and Area.



```
        Spearman's rank correlation rho

data:  Area_Property_Tax_Filtering_DataSet$area and Area_Property_Tax_Filtering_DataSet$`property tax`
S = 5.748e+10, p-value < 2.2e-16
alternative hypothesis: true rho is greater than 0
sample estimates:
     rho
0.680868
```

Figure 23: Spearman's rank correlation test for checking the association between Property tax and Area

Here P-value (2.2× 10^{-16})<Significance level (0.05). Then, reject Ho and a 5% level of confidence level. Therefore, have enough evidence to conclude that There is a Positive correlation between Property tax and Area at a 5% level of significance. By the way, the Spearman correlation coefficient (0.680868) indicates a positive correlation also.

The following Figure 24 Scatterplot indicates the Relationship between Area and Property tax. It also clearly shows a positive correlation between Area and Property Tax.

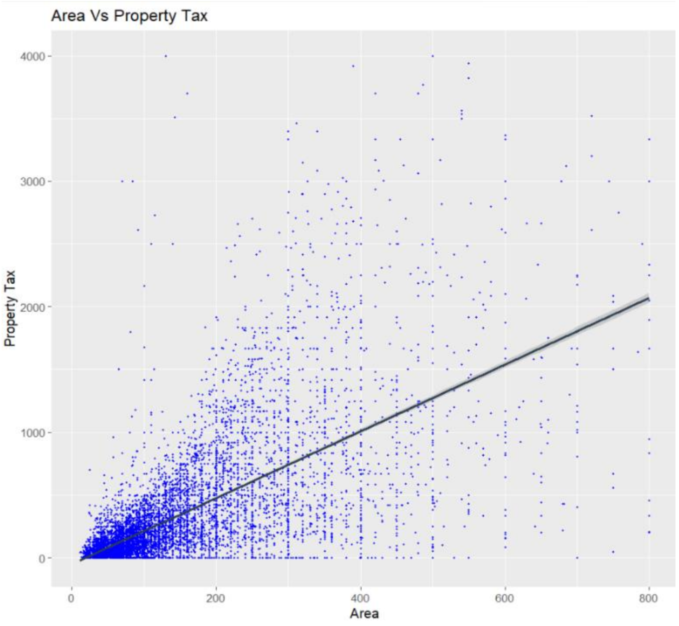


Figure 24: Scatter plot for Area vs. property tax

General Discussion and Conclusion

According to the above study we can conclude that,

- ✓ Total Rent Amount Significantly affected to the floor number.
- ✓ House Rental Amount Significantly affect keeping of a pet or not.

  (The House Rental Amount for keeping a pet is greater than according to the doesn't keep a pet)
- ✓ Fire insurance significantly affects whether the house has furniture or not.

  (Fire insurance is higher for furnished houses as compared to unfurnished houses.)
- ✓ The city in which the property is significantly affects the total rent amount.
- ✓ Property tax relate to the area of the property.

  (There is Positive correlation ship between Property tax and the Area.)

Assessing these factors helps buyers find a residence that meets their requirements without overspending, ensuring a comfortable living space that suits their lifestyle and financial capabilities. This approach empowers individuals to invest wisely, obtaining a property that not only fulfills their immediate needs but also aligns with their future goals and aspirations.

By diligently evaluating these elements, individuals can make a prudent investment that aligns with their current and future housing needs while staying within their financial means.

Appendix

R Code

```
attach(Brazilian_Houses_to_Rent)
#%%%%%%%%%%%%%%%%%%%%  CLEAR DATA SET  %%%%%%%%%%%%%%%%%%%%
#install.packages("tidyverse")
library(tidyverse)
#view(Brazilian_Houses_to_Rent)
glimpse(Brazilian_Houses_to_Rent)
#__________________________ Check Missing Values in Data Set ____________________________
sum(is.na(Brazilian_Houses_to_Rent))

#__________________________ Check Duplicated Rows in Data Set ___________________________
sum(duplicated(Brazilian_Houses_to_Rent))

#__________________________ Remove Duplicated Rows in Data Set __________________________
Brazilian_Houses_to_Rent=unique(Brazilian_Houses_to_Rent)
sum(duplicated(Brazilian_Houses_to_Rent))

#__________________________ Summary of Data Set __________________________________
summary(Brazilian_Houses_to_Rent)

#__________________________ Check Unique in variables _________________________________
unique(Brazilian_Houses_to_Rent$floor)
Brazilian_Houses_to_Rent$floor[Brazilian_Houses_to_Rent$floor == '-']=0
unique(Brazilian_Houses_to_Rent$floor)
New_DataSet = subset(Brazilian_Houses_to_Rent, Brazilian_Houses_to_Rent$floor != "301")
New_DataSet
unique(New_DataSet$floor)
unique(city)
unique(rooms)
unique(bathroom)
unique(`parking spaces`)
unique(`keeping animal`)
unique(furniture)

#_________________ Objective1 [Does the total rent amount significantly affect the floor number?]___________

#Remove Outline Data in Total Rent
plot(New_DataSet$floor,New_DataSet$`total rent`)
boxplot(New_DataSet$`total rent`)
summary(`total rent`)
Outlier_Data_total_rent=filter(New_DataSet,New_DataSet$`total rent`>30000)
Outlier_Data_total_rent
Total_Rent_Filtering_DataSet=filter(New_DataSet,New_DataSet$`total rent`<=30000)
Total_Rent_Filtering_DataSet
boxplot(Total_Rent_Filtering_DataSet$`total rent`)

#Check for normality Total rent
qqnorm(Total_Rent_Filtering_DataSet$`total rent`)
qqline(Total_Rent_Filtering_DataSet$`total rent`)
ks.test(Total_Rent_Filtering_DataSet$`total rent`,"pnorm")
unique(New_DataSet$floor)

#Non Parametric Test (Kruskal Wallis Test)
kruskal.test(Total_Rent_Filtering_DataSet$`total rent`,Total_Rent_Filtering_DataSet$floor)

#Check median in the sample data set
```

```
aggregate(Total_Rent_Filtering_DataSet$`total rent`,
list(Total_Rent_Filtering_DataSet$floor), FUN=median)


#Draw box plot

boxplot(Total_Rent_Filtering_DataSet$`total rent`~Total_Rent_Filtering_DataSet$floor,
main="Floor Vs Total Rent", col="orange", xlab="Floor", ylab = "Total Rent")


#_________________Check Objective2 [Does the house rental amount significantly affect the
keeping of a pet or not?]___________


#Remove Outline Data in the Rent amount

boxplot(New_DataSet$`rent amount`)

summary(New_DataSet$`rent amount`)

Outlier_Data_rent_amount=filter(New_DataSet,New_DataSet$`rent amount`>15000)

Outlier_Data_rent_amount

Rent_Filtering_DataSet=filter(New_DataSet,New_DataSet$`rent amount`<=15000)

Rent_Filtering_DataSet

boxplot(Rent_Filtering_DataSet$`rent amount`)


#Checking for normality

qqnorm(Rent_Filtering_DataSet$`rent amount`)

qqline(Rent_Filtering_DataSet$`rent amount`)

ks.test(Rent_Filtering_DataSet$`rent amount`,"pnorm")


#Non Parametric Test(Wilcoxon's Rank Sum Test)

wilcox.test(Rent_Filtering_DataSet$`rent amount`~Rent_Filtering_DataSet$`keeping
animal`,mu=0,alternative="two.sided",conf.int=T,Conf.level=0.95,paired=F,exact=F,correct=
T)

wilcox.test(Rent_Filtering_DataSet$`rent amount`~Rent_Filtering_DataSet$`keeping
animal`,mu=0,alternative="greater",conf.int=T,Conf.level=0.95,paired=F,exact=F,correct=T)


#Check median in the sample data set
```

```
aggregate(Rent_Filtering_DataSet$`rent amount`, list(Rent_Filtering_DataSet$`keeping animal`), FUN=median)


#Draw box plot

boxplot(Rent_Filtering_DataSet$`rent amount`~Rent_Filtering_DataSet$`keeping animal`,main="Kepping Animal Vs Rent Amount", col="pink", xlab = "Kepping Animal", ylab = "Rent Amount")


#__________________________ Objective 3 [Does Fire insurance significantly affect whether the house has furniture or not?] ____________________


#Remove Outline Data in the fire insurance

summary(New_DataSet$`fire insurance`)

boxplot(New_DataSet$`fire insurance`)


Outlier_Data_fire_insurance=filter(New_DataSet,New_DataSet$`fire insurance`>250)

Outlier_Data_fire_insurance

Fire_Insurance_Filtering_DataSet=filter(New_DataSet,New_DataSet$`fire insurance`<=250)

Fire_Insurance_Filtering_DataSet

boxplot(Fire_Insurance_Filtering_DataSet$`fire insurance`)


#Check for normality Fire insurance

qqnorm(Fire_Insurance_Filtering_DataSet$`fire insurance`)

qqline(Fire_Insurance_Filtering_DataSet$`fire insurance`)

ks.test(Fire_Insurance_Filtering_DataSet$`fire insurance`,"pnorm")


#Non Parametric Test (Wilcoxon's Rank Sum Test)

wilcox.test(Fire_Insurance_Filtering_DataSet$`fire insurance`~Fire_Insurance_Filtering_DataSet$furniture,mu=0,alternative="two.sided",conf.int=T,Conf.level=0.95,paired=F,exact=F,correct=T)

wilcox.test(Fire_Insurance_Filtering_DataSet$`fire insurance`~Fire_Insurance_Filtering_DataSet$furniture,mu=0,alternative="greater",conf.int=T,Conf.level=0.95,paired=F,exact=F,correct=T)
```

```
#Check median in the sample data set

aggregate(Fire_Insurance_Filtering_DataSet$`fire insurance`,
list(Fire_Insurance_Filtering_DataSet$furniture), FUN=median)


#Draw box plot

boxplot(Fire_Insurance_Filtering_DataSet$`fire
insurance`~Fire_Insurance_Filtering_DataSet$furniture, main="Furnished Vs Fire
Insurance", col="sky blue", xlab = "Furnished", ylab="Fire Insurance")


#_____________________ Objective 4 [Does the city in which the property is significantly
affect the total rent amount?] ______________________________


#Checking for normality Total rent

qqnorm(Total_Rent_Filtering_DataSet$`total rent`)

qqline(Total_Rent_Filtering_DataSet$`total rent`)

ks.test(Total_Rent_Filtering_DataSet$`total rent`)



#Non Parametric Test (Kruskal Wallis Test)

kruskal.test(Total_Rent_Filtering_DataSet$`total rent`,Total_Rent_Filtering_DataSet$city)


#Check median in the sample data set

aggregate(Total_Rent_Filtering_DataSet$`total rent`,
list(Total_Rent_Filtering_DataSet$city), FUN=median)


#Draw box plot

boxplot(Total_Rent_Filtering_DataSet$`total rent`~Total_Rent_Filtering_DataSet$city,
col="green", main="City Vs Total Rent", xlab = "City", ylab = "Total Rent")


#_____________________ Objective 5 [Does property tax relate to the area of the
property?] ___________________

#Remove Outline Data in Property tax
```

```
summary(New_DataSet$`property tax`)
boxplot(New_DataSet$`property tax`)

Outlier_Data_Property_Tax=filter(New_DataSet,New_DataSet$`property tax`>4000)
Outlier_Data_Property_Tax
Property_Tax_Filtering_DataSet=filter(New_DataSet,New_DataSet$`property tax`<=4000)
Property_Tax_Filtering_DataSet
boxplot(Property_Tax_Filtering_DataSet$`property tax`)

#Check normality in Property tax
qqnorm(Property_Tax_Filtering_DataSet$`property tax`)
qqline(Property_Tax_Filtering_DataSet$`property tax`)
ks.test(Property_Tax_Filtering_DataSet$`property tax`,"pnorm")

#Remove Outline data in Area
summary(New_DataSet$area)
boxplot(New_DataSet$area)

Outlier_Data_Area=filter(New_DataSet,New_DataSet$area>800)
Outlier_Data_Area
Area_Filtering_DataSet=filter(New_DataSet,New_DataSet$area<=800)
Area_Filtering_DataSet
boxplot(Area_Filtering_DataSet$area)

#Check normality in Area
qqnorm(Area_Filtering_DataSet$area)
qqline(Area_Filtering_DataSet$area)
ks.test(Area_Filtering_DataSet$area,"pnorm")

#Filer Outline data of Property tax in the Area Filtering data set
```

```
Area_Property_Tax_Filtering_DataSet=filter(Area_Filtering_DataSet,Area_Filtering_DataSet$`property tax`<=4000)

Area_Property_Tax_Filtering_DataSet



#Non Parametric Test (Spearman Correlation Test)

cor.test(Area_Property_Tax_Filtering_DataSet$area,Area_Property_Tax_Filtering_DataSet$`property tax`,method="spearman",exact = FALSE,alternative = "two.sided")

cor.test(Area_Property_Tax_Filtering_DataSet$area,Area_Property_Tax_Filtering_DataSet$`property tax`,method="spearman",exact = FALSE,alternative = "greater")




#Draw Scatter Plot

library(ggplot2)

ggplot(Area_Property_Tax_Filtering_DataSet, aes(x =
Area_Property_Tax_Filtering_DataSet$area, y =
Area_Property_Tax_Filtering_DataSet$`property tax`)) +

  geom_point(color = "blue", size = 0.5) +

  ggtitle("Area Vs Property Tax") +

  xlab("Area") + ylab("Property Tax")+

  geom_smooth(method=lm, color='#2C3E50')
```