FINAL PROJECT REPORT

Venkata Naga Sai Neeharika Surampudi, Shubham Nilesh Palande, Sindhu Sundararajan, Swetha Lakshmana Perumal

College of Professional Studies, Northeastern University

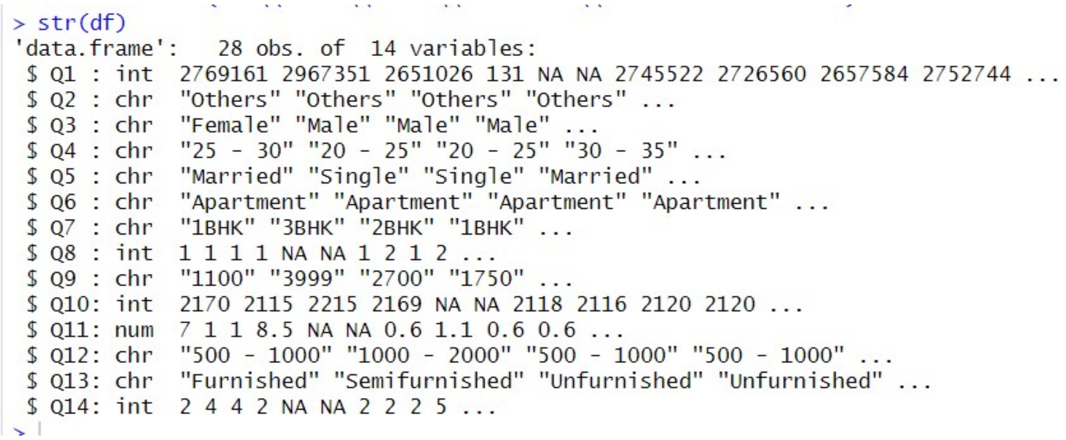
ALY6015: Intermediate Analytics

PROFESSOR: Yun Jiyoung

February 17, 2023

INTRODUCTION

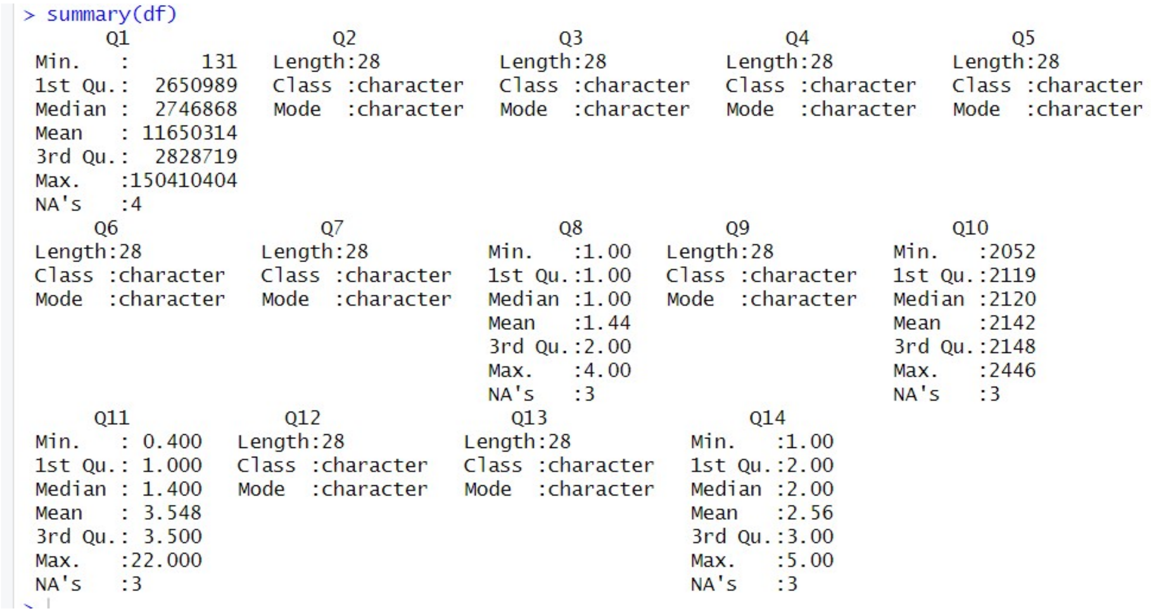
The dataset used is a survey conducted based on the rentals in Boston for Northeastern University students. The dataset contains 28 observations with 14 attributes. The structure of the Boston rentals dataset is shown in Figure 1. Since the data was taken based on asking questions, it can be seen from Figure 1 that the name of the attributes are given as Q1, Q2, etc indicating the questions.



```
> str(df)
'data.frame':   28 obs. of  14 variables:
 $ Q1 : int  2769161 2967351 2651026 131 NA NA 2745522 2726560 2657584 2752744 ...
 $ Q2 : chr  "Others" "Others" "Others" "Others" ...
 $ Q3 : chr  "Female" "Male" "Male" "Male" ...
 $ Q4 : chr  "25 - 30" "20 - 25" "20 - 25" "30 - 35" ...
 $ Q5 : chr  "Married" "Single" "Single" "Married" ...
 $ Q6 : chr  "Apartment" "Apartment" "Apartment" "Apartment" ...
 $ Q7 : chr  "1BHK" "3BHK" "2BHK" "1BHK" ...
 $ Q8 : int  1 1 1 1 NA NA 1 2 1 2 ...
 $ Q9 : chr  "1100" "3999" "2700" "1750" ...
 $ Q10: int  2170 2115 2215 2169 NA NA 2118 2116 2120 2120 ...
 $ Q11: num  7 1 1 8.5 NA NA 0.6 1.1 0.6 0.6 ...
 $ Q12: chr  "500 - 1000" "1000 - 2000" "500 - 1000" "500 - 1000" ...
 $ Q13: chr  "Furnished" "Semifurnished" "Unfurnished" "Unfurnished" ...
 $ Q14: int  2 4 4 2 NA NA 2 2 2 5 ...
>
```

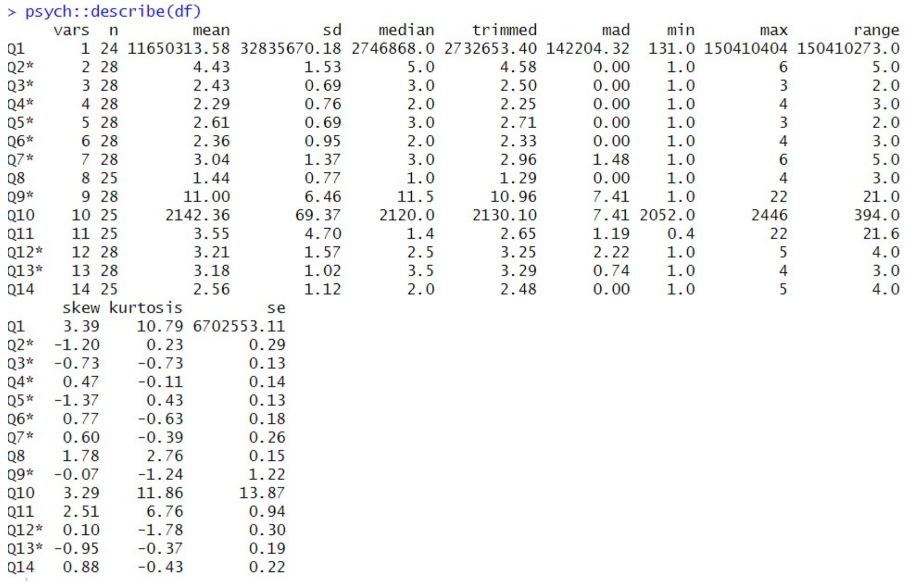
FIGURE 1

The structure and the descriptive statistics of the Boston rentals dataset is illustrated in Figures 2 and 3 respectively. It is clear from the summary statistics that the data set has null values stating that the data set is unclean.



```
> summary(df)
       Q1                  Q2                  Q3                  Q4                  Q5           
 Min.   :      131   Length:28          Length:28          Length:28          Length:28         
 1st Qu.:  2650989   Class :character   Class :character   Class :character   Class :character  
 Median :  2746868   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 11650314                                                                              
 3rd Qu.:  2828719                                                                              
 Max.   :150410404                                                                              
 NA's   :4                                                                                      
      Q6                 Q7                  Q8             Q9                 Q10      
 Length:28          Length:28          Min.   :1.00   Length:28          Min.   :2052  
 Class :character   Class :character   1st Qu.:1.00   Class :character   1st Qu.:2119  
 Mode  :character   Mode  :character   Median :1.00   Mode  :character   Median :2120  
                                       Mean   :1.44                      Mean   :2142  
                                       3rd Qu.:2.00                      3rd Qu.:2148  
                                       Max.   :4.00                      Max.   :2446  
                                       NA's   :3                         NA's   :3     
      Q11              Q12                Q13                 Q14      
 Min.   : 0.400   Length:28          Length:28          Min.   :1.00  
 1st Qu.: 1.000   Class :character   Class :character   1st Qu.:2.00  
 Median : 1.400   Mode  :character   Mode  :character   Median :2.00  
 Mean   : 3.548                                         Mean   :2.56  
 3rd Qu.: 3.500                                         3rd Qu.:3.00  
 Max.   :22.000                                         Max.   :5.00  
 NA's   :3                                              NA's   :3     
>
```

FIGURE 2



```
> psych::describe(df)
     vars  n        mean          sd    median    trimmed       mad    min       max       range
Q1      1 24 11650313.58 32835670.18 2746868.0 2732653.40 142204.32  131.0 150410404 150410273.0
Q2*     2 28        4.43        1.53       5.0       4.58      0.00    1.0         6         5.0
Q3*     3 28        2.43        0.69       3.0       2.50      0.00    1.0         3         2.0
Q4*     4 28        2.29        0.76       2.0       2.25      0.00    1.0         4         3.0
Q5*     5 28        2.61        0.69       3.0       2.71      0.00    1.0         3         2.0
Q6*     6 28        2.36        0.95       2.0       2.33      0.00    1.0         4         3.0
Q7*     7 28        3.04        1.37       3.0       2.96      1.48    1.0         6         5.0
Q8      8 25        1.44        0.77       1.0       1.29      0.00    1.0         4         3.0
Q9*     9 28       11.00        6.46      11.5      10.96      7.41    1.0        22        21.0
Q10    10 25     2142.36       69.37    2120.0    2130.10      7.41 2052.0      2446       394.0
Q11    11 25        3.55        4.70       1.4       2.65      1.19    0.4        22        21.6
Q12*   12 28        3.21        1.57       2.5       3.25      2.22    1.0         5         4.0
Q13*   13 28        3.18        1.02       3.5       3.29      0.74    1.0         4         3.0
Q14    14 25        2.56        1.12       2.0       2.48      0.00    1.0         5         4.0
      skew kurtosis         se
Q1    3.39    10.79 6702553.11
Q2*  -1.20     0.23       0.29
Q3*  -0.73    -0.73       0.13
Q4*   0.47    -0.11       0.14
Q5*  -1.37     0.43       0.13
Q6*   0.77    -0.63       0.18
Q7*   0.60    -0.39       0.26
Q8    1.78     2.76       0.15
Q9*  -0.07    -1.24       1.22
Q10   3.29    11.86      13.87
Q11   2.51     6.76       0.94
Q12*  0.10    -1.78       0.30
Q13* -0.95    -0.37       0.19
Q14   0.88    -0.43       0.22
```

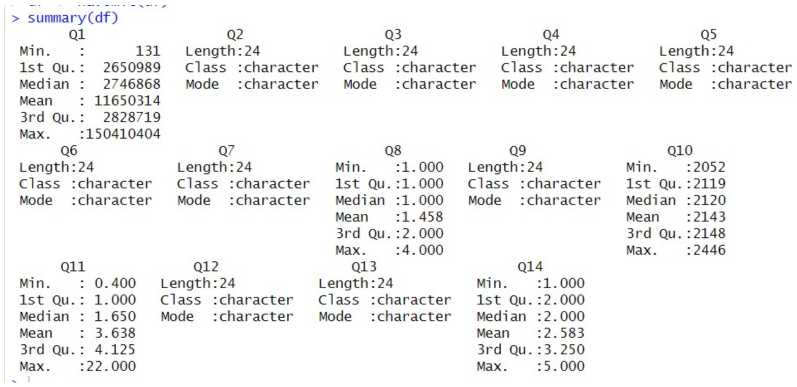
FIGURE 3

The business questions that are to be analyzed using the dataset are:

1. Do the people who live close to the university pay higher or lower rent than those who lives farther away by comparing the two groups of people?
2. Is the apartment or individual housing expensive to rent when there are the same number of bedrooms?
3. Is the number of tenants directly proportional to the average rent?
4. Is the rent for a fully furnished house expensive?

DATA CLEANING

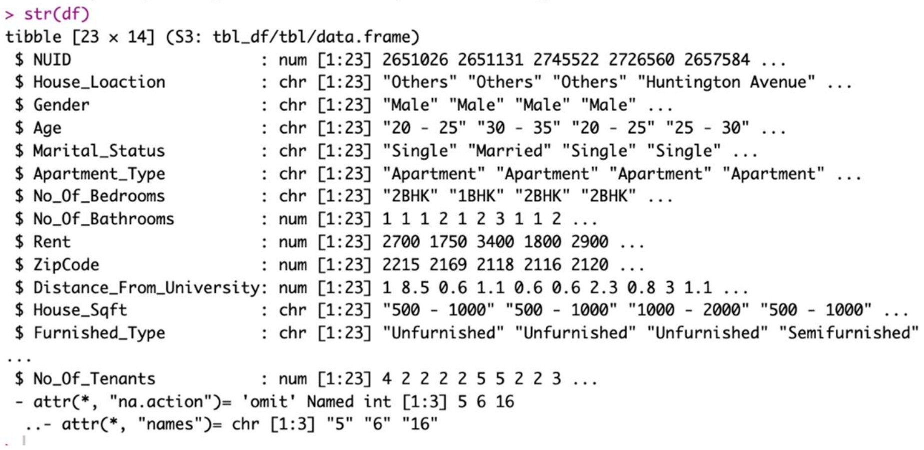
The null values in the dataset are omitted using the drop.na() function. Figure 4 depicts the summary of the cleaned dataset. It is clear from the Figure 4 that there are no more null values.



```
> summary(df)
       Q1                   Q2                  Q3                  Q4                  Q5           
 Min.   :      131   Length:24          Length:24          Length:24          Length:24         
 1st Qu.:  2650989   Class :character   Class :character   Class :character   Class :character  
 Median :  2746868   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   : 11650314                                                                              
 3rd Qu.:  2828719                                                                              
 Max.   :150410404                                                                              
      Q6                  Q7                  Q8              Q9                 Q10      
 Length:24          Length:24          Min.   :1.000   Length:24          Min.   :2052  
 Class :character   Class :character   1st Qu.:1.000   Class :character   1st Qu.:2119  
 Mode  :character   Mode  :character   Median :1.000   Mode  :character   Median :2120  
                                       Mean   :1.458                      Mean   :2143  
                                       3rd Qu.:2.000                      3rd Qu.:2148  
                                       Max.   :4.000                      Max.   :2446  
      Q11              Q12                Q13                 Q14       
 Min.   : 0.400   Length:24          Length:24          Min.   :1.000  
 1st Qu.: 1.000   Class :character   Class :character   1st Qu.:2.000  
 Median : 1.650   Mode  :character   Mode  :character   Median :2.000  
 Mean   : 3.638                                         Mean   :2.583  
 3rd Qu.: 4.125                                         3rd Qu.:3.250  
 Max.   :22.000                                         Max.   :5.000  
```

FIGURE 4

Now, the name of the attributes are renamed from Q1, Q2, Q3, etc to NUID, House_Location, Gender, Age, Marital_Status, Apartment_Type, No_of_Bedrooms, No_of_Bathrooms, Rent, Zipcode, Distance_from_University, House_Sqft, Furnished_Type, and No_of_Tenants. Then, the term 'BHK' is removed from the column Number_of_Bedroom and then it is converted from character to numeric values. The structure of the clean dataset can be seen in Figure 5.



```
> str(df)
tibble [23 × 14] (S3: tbl_df/tbl/data.frame)
 $ NUID                    : num [1:23] 2651026 2651131 2745522 2726560 2657584 ...
 $ House_Loaction          : chr [1:23] "Others" "Others" "Others" "Huntington Avenue" ...
 $ Gender                  : chr [1:23] "Male" "Male" "Male" "Male" ...
 $ Age                     : chr [1:23] "20 - 25" "30 - 35" "20 - 25" "25 - 30" ...
 $ Marital_Status          : chr [1:23] "Single" "Married" "Single" "Single" ...
 $ Apartment_Type          : chr [1:23] "Apartment" "Apartment" "Apartment" "Apartment" ...
 $ No_Of_Bedrooms          : chr [1:23] "2BHK" "1BHK" "2BHK" "2BHK" ...
 $ No_Of_Bathrooms         : num [1:23] 1 1 1 2 1 2 3 1 1 2 ...
 $ Rent                    : num [1:23] 2700 1750 3400 1800 2900 ...
 $ ZipCode                 : num [1:23] 2215 2169 2118 2116 2120 ...
 $ Distance_From_University: num [1:23] 1 8.5 0.6 1.1 0.6 0.6 2.3 0.8 3 1.1 ...
 $ House_Sqft              : chr [1:23] "500 - 1000" "500 - 1000" "1000 - 2000" "500 - 1000" ...
 $ Furnished_Type          : chr [1:23] "Unfurnished" "Unfurnished" "Unfurnished" "Semifurnished"
...
 $ No_Of_Tenants           : num [1:23] 4 2 2 2 2 5 5 2 2 3 ...
 - attr(*, "na.action")= 'omit' Named int [1:3] 5 6 16
  ..- attr(*, "names")= chr [1:3] "5" "6" "16"
```

FIGURE 5

DATA VISUALIZATIONS

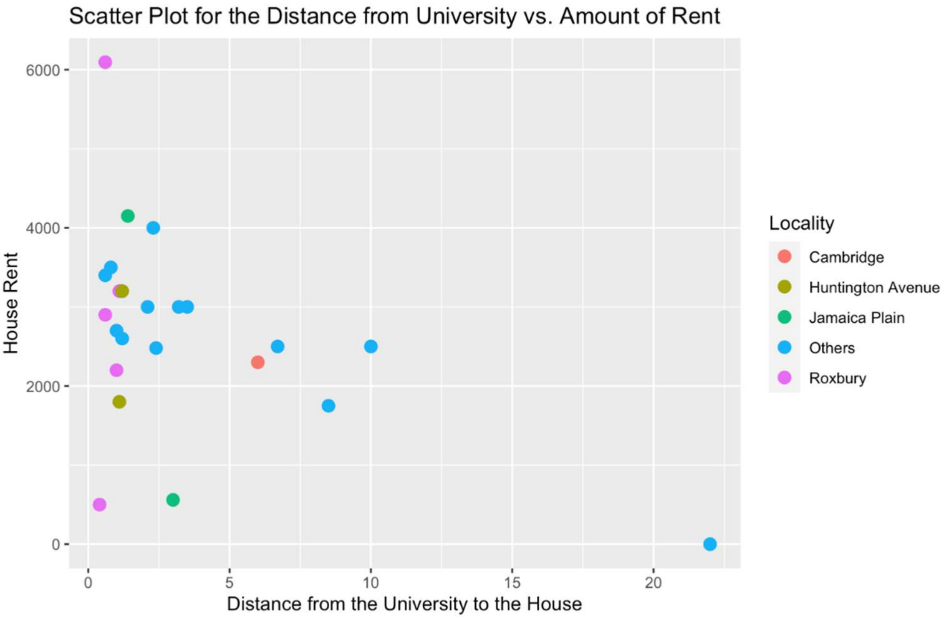


FIGURE 6

Figure 6 is a scatter plot illustrating the distance from the university and the rent amount. It can be inferred from the chart that the majority of the population stays in distance radius of 0-2.5 miles from the University and most of them pay around $1500-$4000 and stay within the radius of 2.5 miles. Hence, it can be concluded that there are people staying closer to the university paying more than the people who are staying far away.

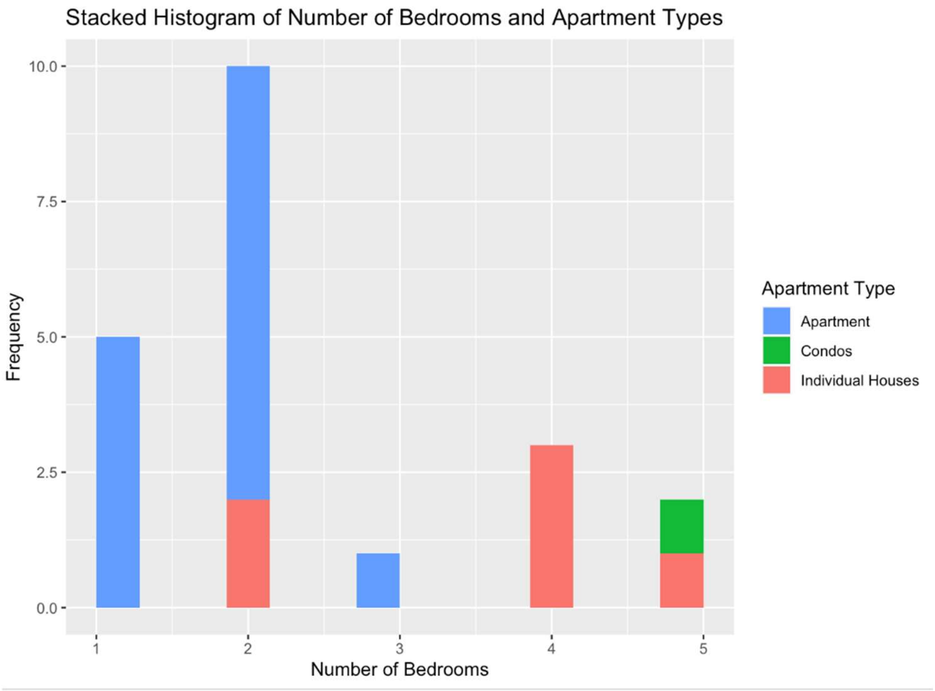


FIGURE 6

Figure 6 is a stacked histogram depicting that until 3BHK people prefer staying in apartments. But for 4BHK or 5BHK people prefer staying in individual houses or condos.

The furnishment level in each locality is illustrates as a tile plot in Figure 7. It could be seen from Figure 7 that the houses in Cambridge are fully furnished. The houses in Jamaica plain, Huntington Avenue, and Roxbury are semi-furnished whereas some houses in Jamaica plain and Roxbury are unfurnished.

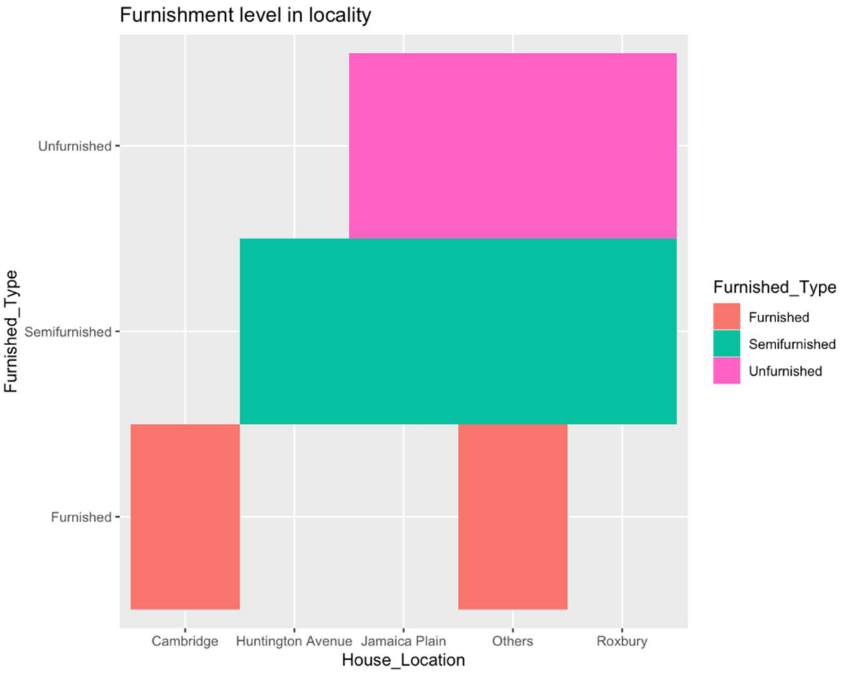


FIGURE 7

The bar plot for Average Rent vs. Apartment Type is shown in Figure 8 where it can be observed that Rent for Condos is expensive as compared to the rent of Apartments and Individual Houses.

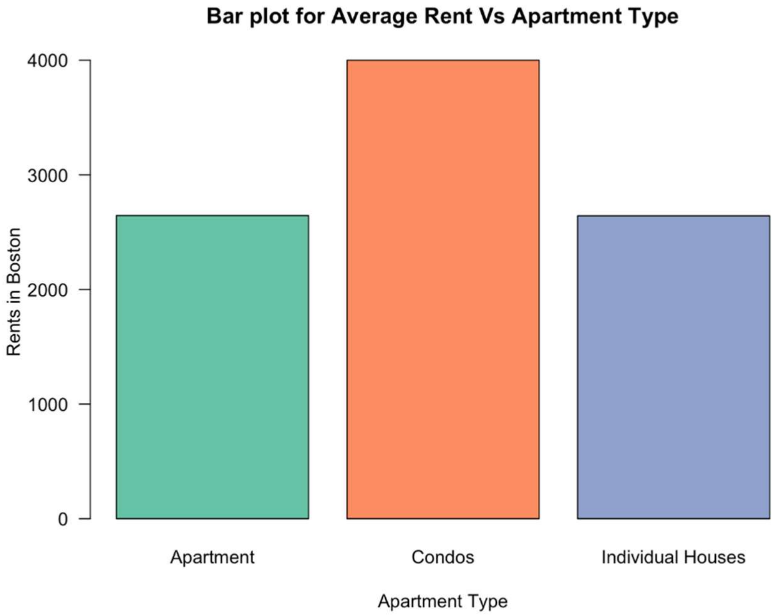
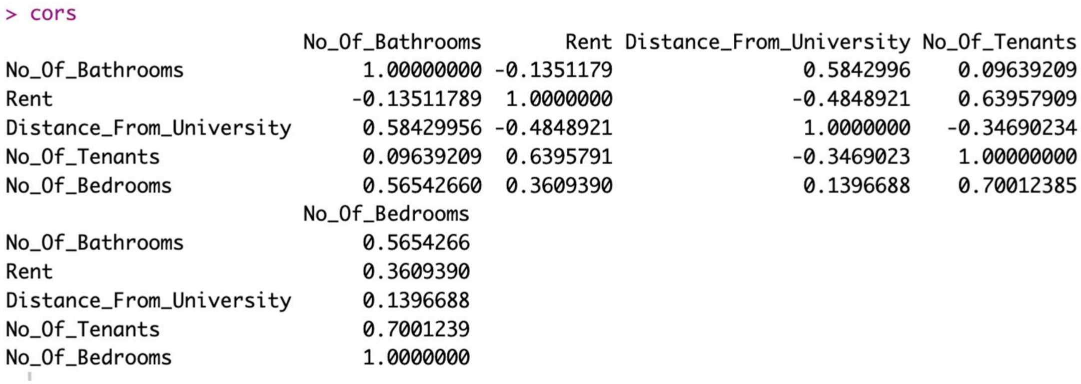


FIGURE 8

INITIAL ANALYSIS

The initial analysis is done to check whether the rents across all the areas is same or not. To determine we conducted chi-square test. Also to find the relation between independent variables we have done correlation analysis.

Figure 9 shows the correlation between number of bathrooms, bedrooms, rent, distance from the university, and the number of tenants.



```
> cors
                         No_Of_Bathrooms       Rent Distance_From_University No_Of_Tenants
No_Of_Bathrooms               1.00000000 -0.1351179                0.5842996    0.09639209
Rent                         -0.13511789  1.0000000               -0.4848921    0.63957909
Distance_From_University      0.58429956 -0.4848921                1.0000000   -0.34690234
No_Of_Tenants                 0.09639209  0.6395791               -0.3469023    1.00000000
No_Of_Bedrooms                0.56542660  0.3609390                0.1396688    0.70012385
                         No_Of_Bedrooms
No_Of_Bathrooms               0.5654266
Rent                          0.3609390
Distance_From_University      0.1396688
No_Of_Tenants                 0.7001239
No_Of_Bedrooms                1.0000000
```

FIGURE 9

The correlation plot is illustrated in Figure 10. Basically correlation is calculated based on the size if the size and color (blue for positive correlation and orange for negative correlation with the size and color depends on the strength of the correlation) It is clear from the plot that the number of bedrooms and the number of tenants and the number of Tenants and rent are highly correlated with a correlation value of 0.7. The least correlation is between the rent and the distance from the university.

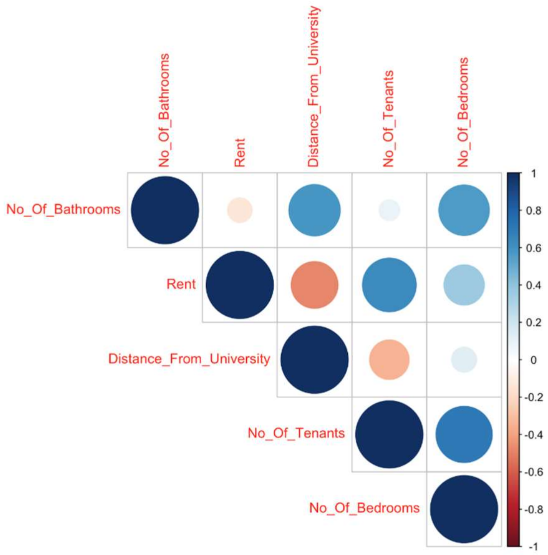
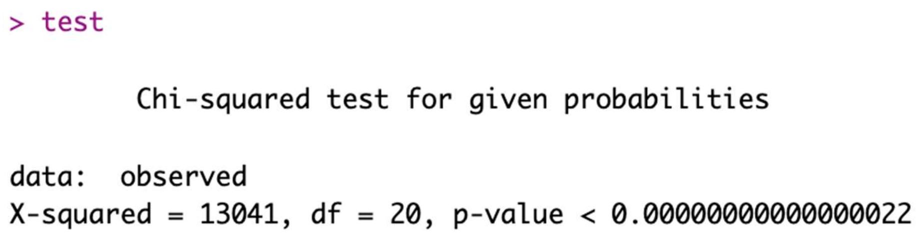


FIGURE 10

Chi square test:

Null Hypotheses (H0): The rent of all areas in Boston is the same.
Alternate Hypotheses (Ha): The rent of all areas in Boston is not the same.
Alpha: 0.05



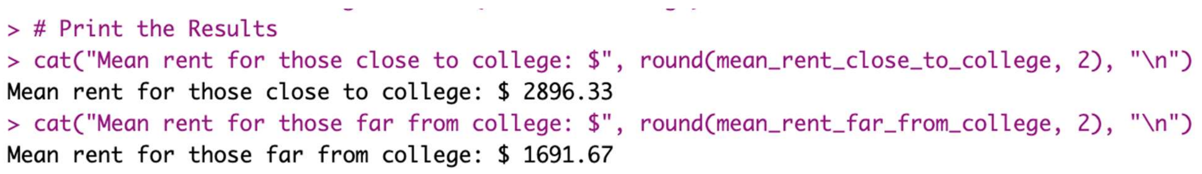
```
> test

	Chi-squared test for given probabilities

data:  observed
X-squared = 13041, df = 20, p-value < 0.00000000000000022
```

FIGURE 11

From the result it is evident that the p value is 2.2*10^16 which less than alpha value of 0.05. Hence, the null hypothesis is rejected. From the initial analysis, it can be concluded that the rent of all the areas in Boston are not same.

Question 1: Do the people who live close to the university pay higher or lower rent than those who lives farther away by comparing the two groups of people?

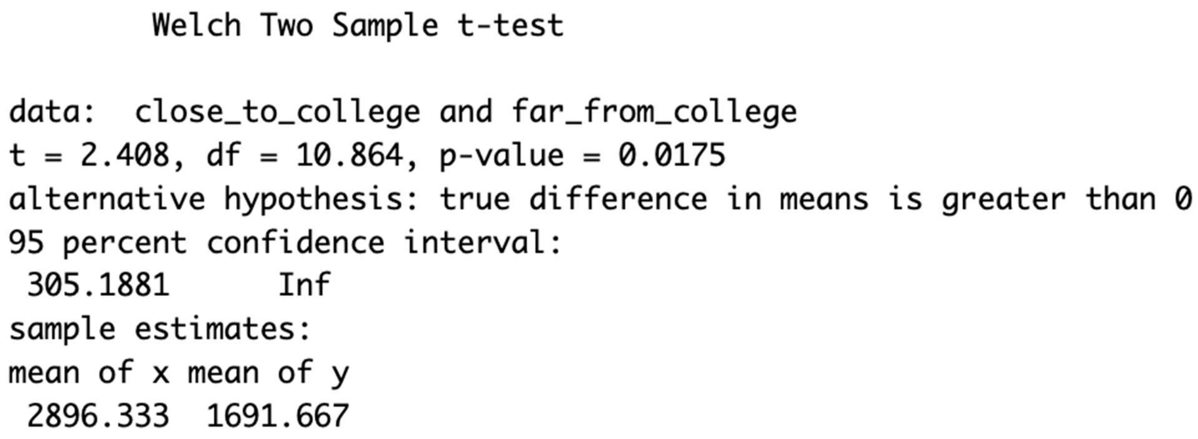
Here, two categories are compared, i.e., the people who are staying near the university (within 4 miles) and the people staying farther from the university (greater than 4 miles). So, two-sample t test is conducted. The mean values for people staying near the university and for the people staying farther from the university are calculated and can be seen in Figure 12.



```
> # Print the Results
> cat("Mean rent for those close to college: $", round(mean_rent_close_to_college, 2), "\n")
Mean rent for those close to college: $ 2896.33
> cat("Mean rent for those far from college: $", round(mean_rent_far_from_college, 2), "\n")
Mean rent for those far from college: $ 1691.67
```

FIGURE 12

Null Hypothesis (H0): The mean differences of rent for houses closer to college is more than, that of the houses far from the college
Alternate Hypothesis (Ha): The mean differences of rent for houses closer to college is not same as the houses far from the college.
Alpha: 0.05



```
        Welch Two Sample t-test

data:  close_to_college and far_from_college
t = 2.408, df = 10.864, p-value = 0.0175
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 305.1881      Inf
sample estimates:
mean of x mean of y 
 2896.333  1691.667 
```

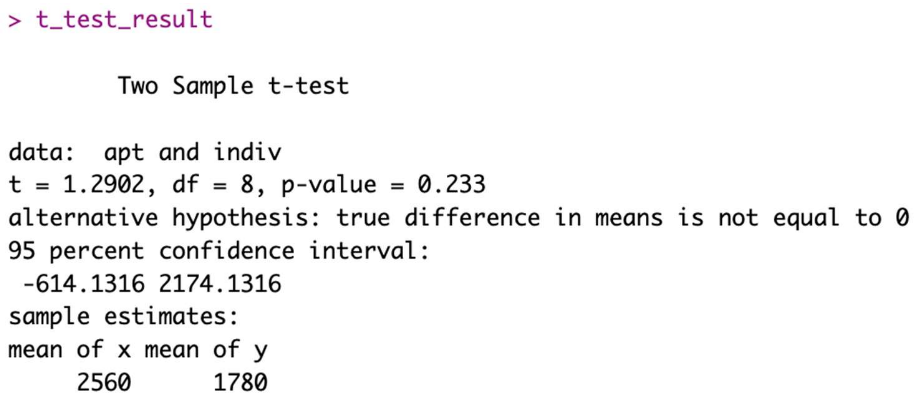
FIGURE 13

It can be seen from Figure 13 that the p-value is 0.0175, which is greater than the alpha value of 0.05. So, the null hypothesis is rejected. There is enough evidence to support the claim that the mean differences of rent for houses closer to college is not same as the houses far from the college.

Question 2: Is the apartment or individual housing expensive to rent when there are the same number of bedrooms?

As two groups are involved i.e., people living in apartments and people living in individual houses, two-sample t-test is conducted for this question. Here, subset() function is used to subset the data of individual houses and apartments from the dataset. For the number of bedrooms, the houses with 2 bedrooms are considered as majority of the people live in 2 bedrooms in the dataset.

Null Hypothesis (H0): The rent for the apartment is same as that of individual houses for 2BHKs.
Alternate Hypothesis (Ha): The rent for the apartment is different as that of individual houses for 2BHKs.
Alpha: 0.05



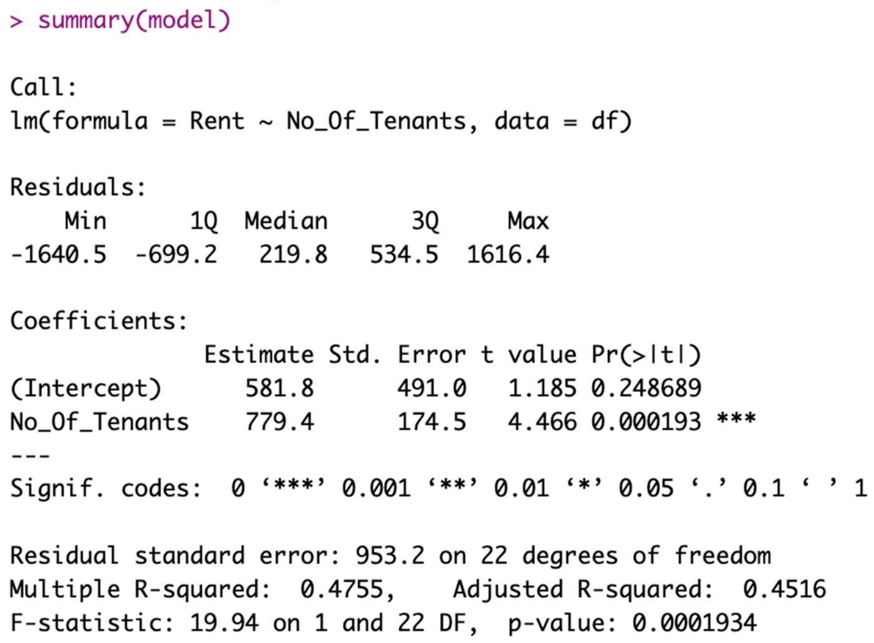
```
> t_test_result

	Two Sample t-test

data:  apt and indiv
t = 1.2902, df = 8, p-value = 0.233
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -614.1316 2174.1316
sample estimates:
mean of x mean of y 
     2560      1780 
```

FIGURE 14

From Figure 14, it is evident that the p-value is 0.233, which is less than the alpha value of 0.05. There is not enough evidence to support the and so, the null hypothesis is failed to be rejected. Hence, it can be concluded that the rent for the apartment is same as that of individual houses for 2BHKs.

Question 3: Is the number of tenants directly proportional to the average rent?

Linear regression analysis is done to predict the answers for question 3. The summary of the linear regression analysis can be seen in Figure 15.



```
> summary(model)

Call:
lm(formula = Rent ~ No_Of_Tenants, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1640.5  -699.2   219.8   534.5  1616.4 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)      581.8      491.0   1.185 0.248689    
No_Of_Tenants    779.4      174.5   4.466 0.000193 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 953.2 on 22 degrees of freedom
Multiple R-squared:  0.4755,	Adjusted R-squared:  0.4516 
F-statistic: 19.94 on 1 and 22 DF,  p-value: 0.0001934
```

FIGURE 15

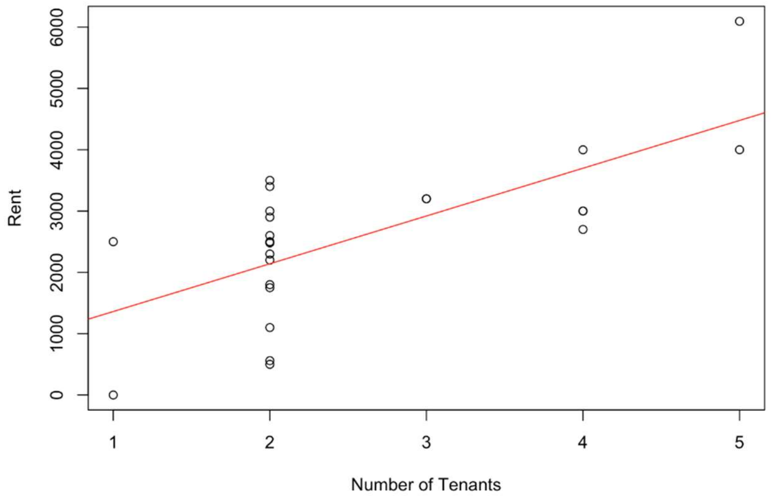


FIGURE 16

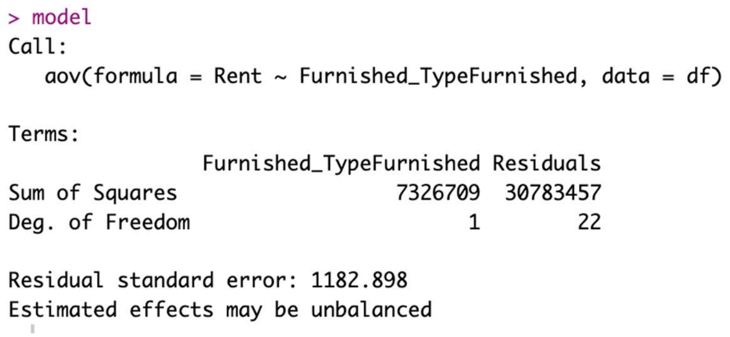
Figure 16 depicts the regression analysis that involves fitting a line to the data points, where the dependent variable is the rent, and the independent variable is the number of tenants. The slope of the line would indicate the change in rent for each additional tenant. If the slope is positive and statistically significant, then we can say that there is a relationship between the number of tenants and rent, and that having a higher number of tenants is associated with higher rent.

Since the slope is positive for the 1st quadrant, it can be concluded that the rent and the number of tenants are directly proportional to each other.

Question 4: Is the rent for a fully furnished house expensive?

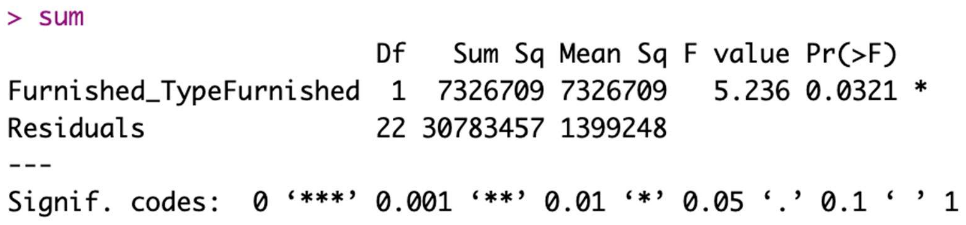
ANOVA test is performed for this question to compare the mean of the three furnishment levels namely furnished, unfurnished, and semi-furnished. Dummy variables are created for the furnished_type attribute using the model.matrix() function.

Null Hypothesis (H0): The variances of rents for furnished, semi furnished, and fully furnished houses are same.
Alternate Hypothesis (Ha): The variances of rents for furnished, semi furnished, and fully furnished houses are different.
Alpha: 0.05



```
> model
Call:
   aov(formula = Rent ~ Furnished_TypeFurnished, data = df)

Terms:
                Furnished_TypeFurnished Residuals
Sum of Squares                  7326709  30783457
Deg. of Freedom                       1        22

Residual standard error: 1182.898
Estimated effects may be unbalanced
```

FIGURE 17



```
> sum
                        Df   Sum Sq Mean Sq F value Pr(>F)
Furnished_TypeFurnished  1  7326709 7326709   5.236 0.0321 *
Residuals               22 30783457 1399248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

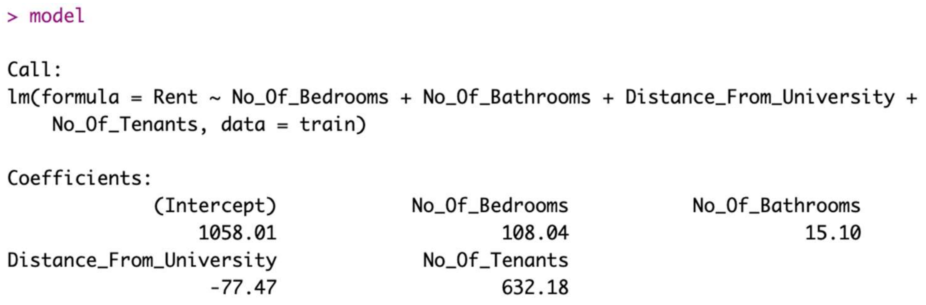
FIGURE 18

It can be inferred from Figure 18 that the p-value is 0.0321, which is greater than the alpha value of 0.05. So, the null hypothesis is rejected. There is enough evidence to support the claim that the variances of rents for furnished, semi furnished, and fully furnished houses are different.

It can be concluded that rent varies when compared to furnished, semi furnished and fully furnished houses.

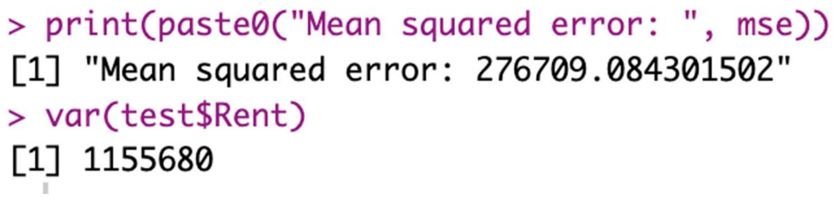
LINEAR REGRESSION

By using data partition, the dataset is split into train and test data where the train dataset contains 70% of the values and the test dataset contains 30% of the values. Linear regression model is done to predict the rent where the No_Of_Bedrooms, No_Of_Bathrooms, Distance_From_University, and No_Of_Tenants are used as predictors.



```
> model

Call:
lm(formula = Rent ~ No_Of_Bedrooms + No_Of_Bathrooms + Distance_From_University +
    No_Of_Tenants, data = train)

Coefficients:
             (Intercept)            No_Of_Bedrooms           No_Of_Bathrooms
                 1058.01                    108.04                     15.10
Distance_From_University             No_Of_Tenants
                  -77.47                    632.18
```

FIGURE 19



```
> print(paste0("Mean squared error: ", mse))
[1] "Mean squared error: 276709.084301502"
> var(test$Rent)
[1] 1155680
```

FIGURE 20

It can be inferred from Figure 20 that as MSE is much smaller than variance. Hence, it can be concluded that, this model is a good fit for predicting rent.

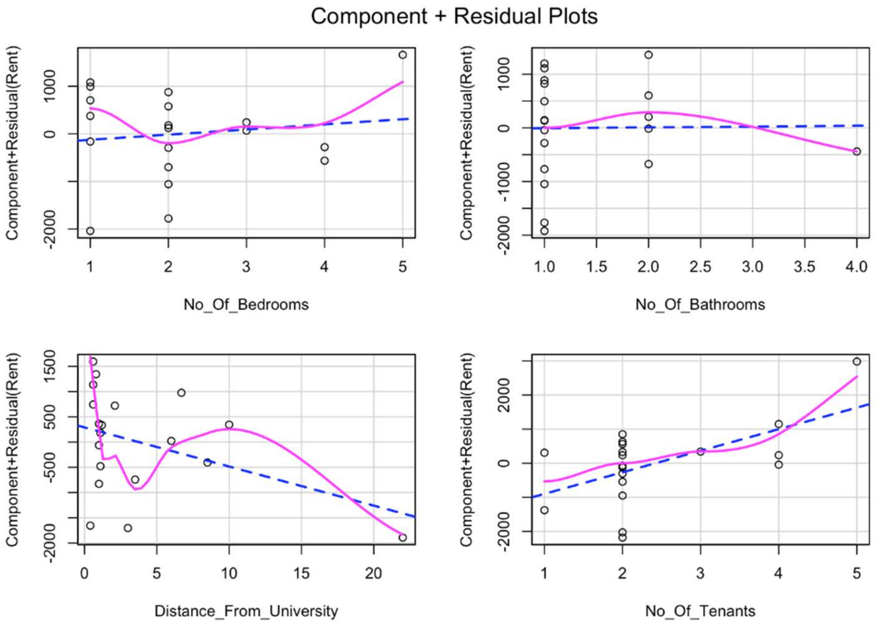


FIGURE 21

Figure 21 shows components and residual plots for rent with respect to the attributes No_of_Bedrooms, No_Of_Bathrooms, Distance_From_University, and No_Of_Tenants mentioned in the regression model. From these graphs, we got to know that there are only few residual values that are deviated from the original plot and we ignored them as we have less data to work with.

Spread level Plot **Hat Plot**

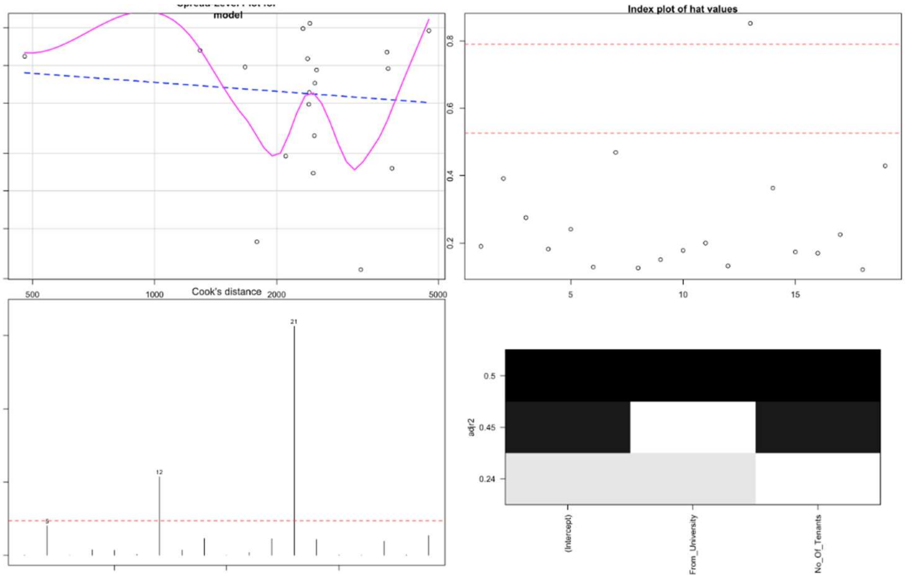
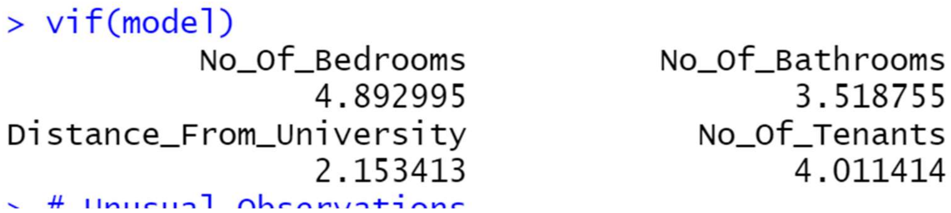


FIGURE 22

From the cook’s distance graph we got to know the influential data points and we found out that there are few outliers. From this spread level plot, the values in a spread level plot are not concentrated on the ideal line and there is a decreasing trend in the plot, this indicates that the spread of the data decreases as the level of the predictor variable increases. This means that the variance of the data is not constant across the different levels of the predictor variable, and may indicate a violation of the assumption of equal variance.

We calculated the VIF for the model to check **Multicollinearity** for these variables



```
> vif(model)
           No_Of_Bedrooms           No_Of_Bathrooms
                 4.892995                  3.518755
Distance_From_University             No_Of_Tenants
                 2.153413                  4.011414
> # Unusual Observations
```

All the values for the variables are less than 5 hence we can say that **Multicollinearity** exists for these variables and correction of multicollinearity is not required.

SUMMARY

From all the analysis we can say that those who are living near to the college are paying more rent when compared to those who are far and also based on the based on the furnishing type of the house's rents are varying, the houses in Cambridge are fully furnished. The houses in Jamaica plain, Huntington Avenue, and Roxbury are semi-furnished whereas some houses in Jamaica plain and Roxbury are unfurnished. We also got to know that condos are little expensive when compared to furnished and unfurnished houses.

Our linear regression model is as follows.

Rent = 1058.01 + (108.04 * No_Of_Bedrooms) + (15.10 * No_Of_Bathrooms) - (77.47 * Distance_From_University) + (632.18 * No_Of_Tenants)

RECOMMENDATIONS

- Consider the location of the house when determining the rent amount: Since those who are living near the college are paying more rent when compared to those who are far, it may be a good idea to adjust the rent amount based on the proximity to the college.
- Consider the furnishing type of the house when determining the rent amount: Since the rent amounts vary based on the furnishing type of the house, it may be a good idea to adjust the rent amount accordingly. Fully furnished houses in Cambridge can command higher rent amounts compared to semi-furnished or unfurnished houses in Jamaica plain, Huntington Avenue, and Roxbury.
- Consider the type of housing when determining the rent amount: Condos are more expensive when compared to furnished and unfurnished houses. Thus, it may be a good idea to adjust the rent amount for condos accordingly.

REFERENCES

Bevans, R. (2020, February 25). Linear Regression in R | A Step-by-Step Guide & Examples. In *Scribbr*. https://www.scribbr.com/statistics/linear-regression-in-r/

R Handbook: Hypothesis Testing and p-values. (n.d.). In *R Handbook: Hypothesis Testing and p-values*. https://rcompanion.org/handbook/D_01.html

Kabacoff, R. (2015, June 4). *R in Action*.