University of East London

Ain Shams University
Faculty of Computer and
Information Sciences

Big Data Analytics Project

Housing Prices

Team ID: **8**

Team Members:	**ID:**
Ahmed Medhat	20201701703
Hla Hatem	20201701714
Sherry Fady	20201701719
Diaa Eldin Amr	20201701720
Ali Ahmed Shaker	20201701725

A.Project Description

This project aims to analyze a Housing data set to gain valuable insights into the property market. By examining housing data and identifying key factors influencing property prices.

B.Dataset Description

The dataset consists of real estate listings. The list contains attributes that describe the property and its amenities. Variables:

- **Price**: The listing price of the property.
- **Area**: The total size of the property.
- **Bedrooms**: The number of bedrooms in the property.
- **Bathrooms**: The number of bathrooms on the property.
- **Stories**: The number of floors in the property.
- **Main Road**: Binary variable indicates whether the property is located on a main road.
- **Guestroom**: Binary variable indicating the presence or absence of a guest room.
- **Basement**: Binary variable indicating the presence or absence of a basement.
- **HotWaterHeating**: Binary variable indicating the presence or absence of hot water heating.
- **AirConditioning**: Binary variable indicating the presence or absence of air conditioning.
- **Parking**: Binary variable indicating the presence or absence of parking.
- **PreArea** (preferred Area): Binary variable indicating a specific area designation.
- **Furnishing Status**: This variable describes whether the property is furnished, semi-furnished, or unfurnished.

```
> dim(housing)
[1] 545  13
> names(housing)
 [1] "price"            "area"             "bedrooms"         "bathrooms"
 [5] "stories"          "mainroad"         "guestroom"        "basement"
 [9] "hotwaterheating"  "airconditioning"  "parking"          "prefarea"
[13] "furnishingstatus"
> str(housing)
'data.frame':   545 obs. of  13 variables:
 $ price           : int  13300000 12250000 12250000 12215000 11410000 10850000 10150000 10150000 9870000 9800000 ...
 $ area            : int  7420 8960 9960 7500 7420 7500 8580 16200 8100 5750 ...
 $ bedrooms        : int  4 4 3 4 4 3 4 5 4 3 ...
 $ bathrooms       : int  2 4 2 2 1 3 3 3 1 2 ...
 $ stories         : int  3 4 2 2 2 1 4 2 2 4 ...
 $ mainroad        : chr  "yes" "yes" "yes" "yes" ...
 $ guestroom       : chr  "no" "no" "no" "no" ...
 $ basement        : chr  "no" "no" "yes" "yes" ...
 $ hotwaterheating : chr  "no" "no" "no" "no" ...
 $ airconditioning : chr  "yes" "yes" "no" "yes" ...
 $ parking         : int  2 3 2 3 2 2 2 0 2 1 ...
 $ prefarea        : chr  "yes" "no" "yes" "yes" ...
 $ furnishingstatus: chr  "furnished" "furnished" "semi-furnished" "furnished" ...
```

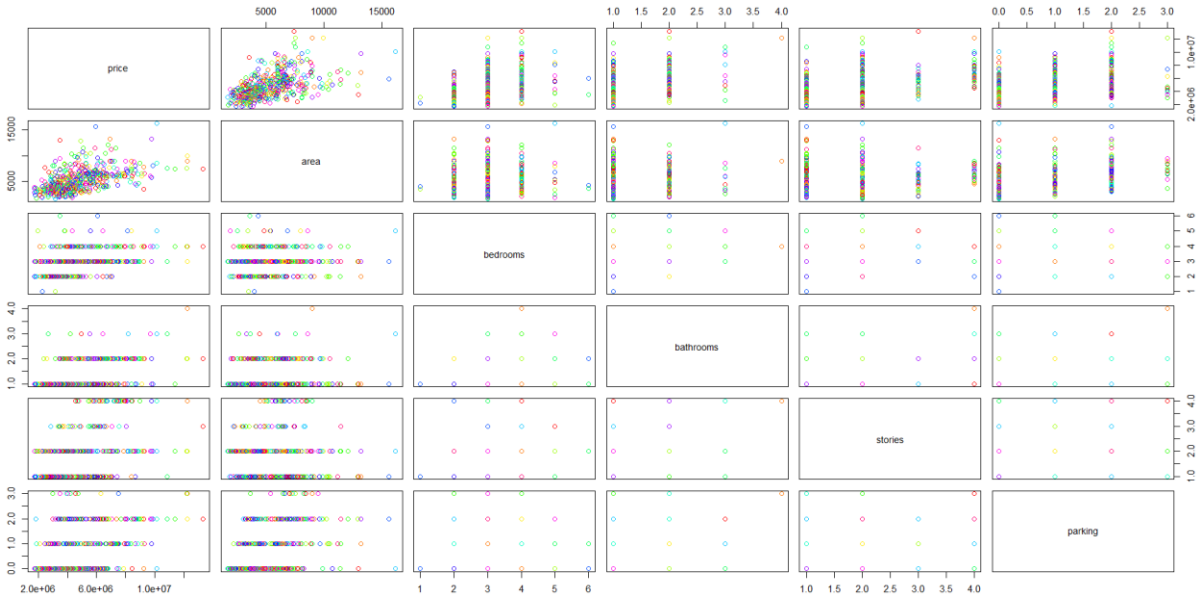
```
> summary(housing)
     price              area          bedrooms        bathrooms       stories
 Min.   : 1750000   Min.   : 1650   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.: 3430000   1st Qu.: 3600   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Median : 4340000   Median : 4600   Median :3.000   Median :1.000   Median :2.000
 Mean   : 4766729   Mean   : 5151   Mean   :2.965   Mean   :1.286   Mean   :1.806
 3rd Qu.: 5740000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :13300000   Max.   :16200   Max.   :6.000   Max.   :4.000   Max.   :4.000
   mainroad          guestroom          basement        hotwaterheating    airconditioning
 Length:545         Length:545         Length:545         Length:545         Length:545
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
    parking          prefarea         furnishingstatus
 Min.   :0.0000   Length:545         Length:545
 1st Qu.:0.0000   Class :character   Class :character
 Median :0.0000   Mode  :character   Mode  :character
 Mean   :0.6936
 3rd Qu.:1.0000
 Max.   :3.0000
```

C.Problem Definition & Project Objectives

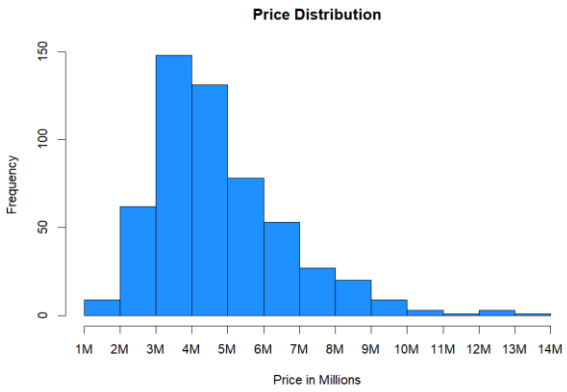
Problem Definition: Accurately predict the asking or selling price of a property based on its characteristics.

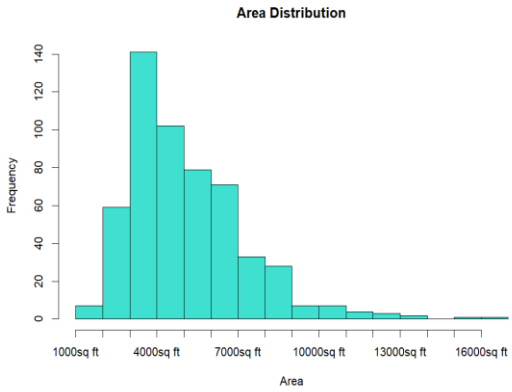
Objective: Develop a pricing model using machine learning and statistical techniques to predict property prices based on the available features in the dataset.

D.Data Visualization

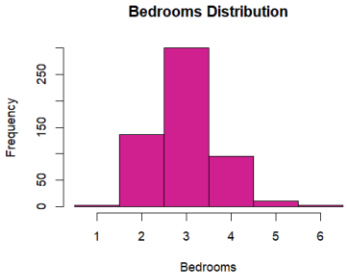


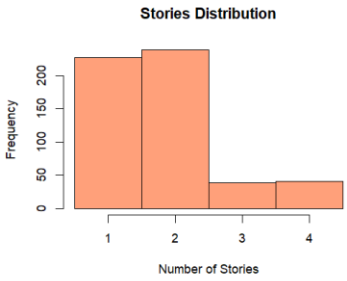
Bar Plots



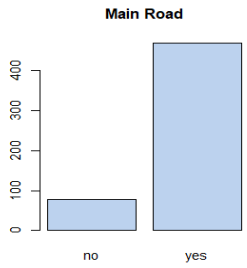


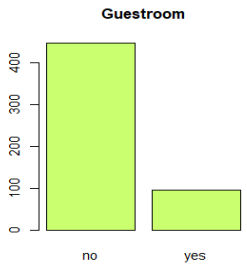
- These bar plots show the distribution of prices and areas
- The most common price range is between 3M to 4M
- The most common area is between 3000 sq ft and 4000 sq ft

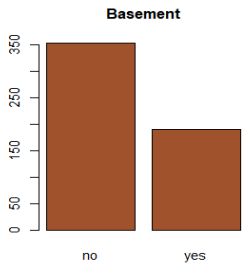


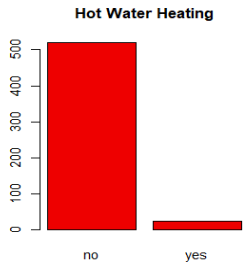


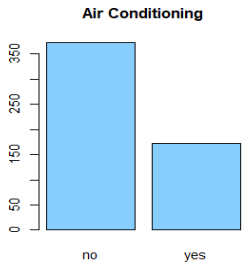
- Most common number of bedrooms is 3
- Most common number of bathrooms is 1
- Most common number of stories is 2
- Most common number of parking is 0

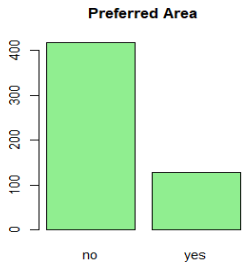






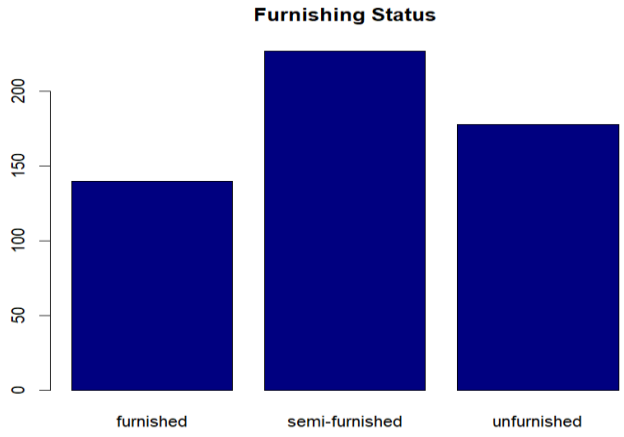




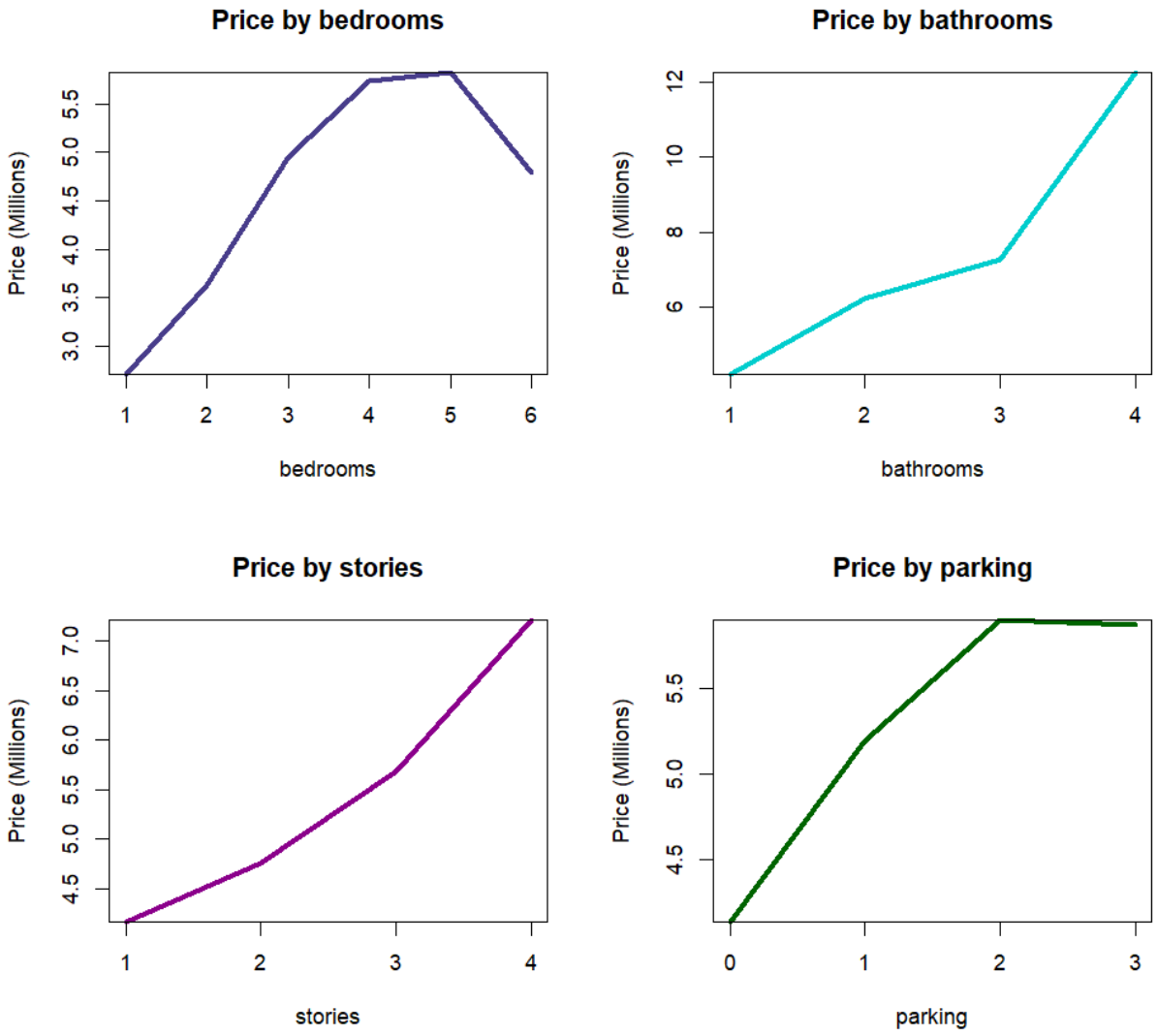


- Most of the houses are on the main road
- Most of the houses do not have guest rooms
- Most of the houses do not have a basement
- Most of the houses do not have hot water heating
- Most of the houses do not have air conditioning
- Most of the houses are not in the preferred area

- The majority of houses appear to be semi-furnished, followed by unfurnished, and then furnished.
- This graph suggests a preference for semi-furnished properties within this project.



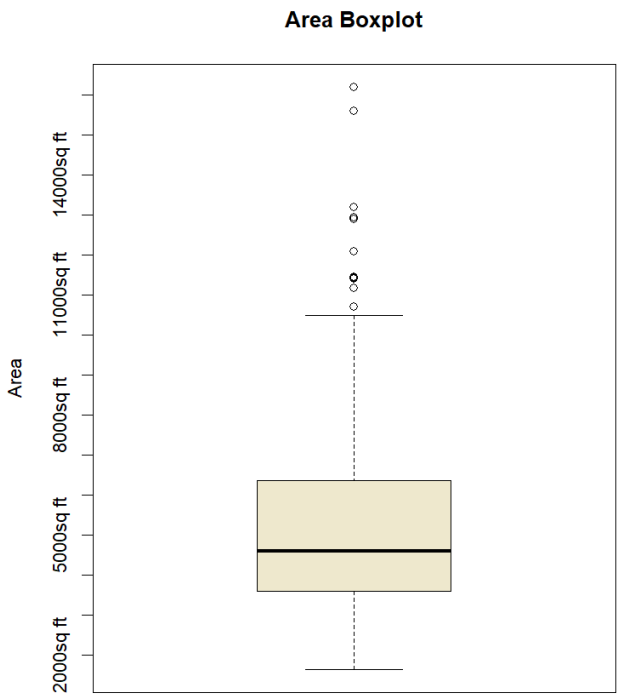
- The number of Bathrooms and Stories have a positive relationship with the price. If the number of bathrooms or stories increases then the house price also increases
- The number of Bedrooms and Parking Spots do not necessarily have a positive relationship with the price as the price increases when they increase, however after reaching a certain threshold it starts decreasing again

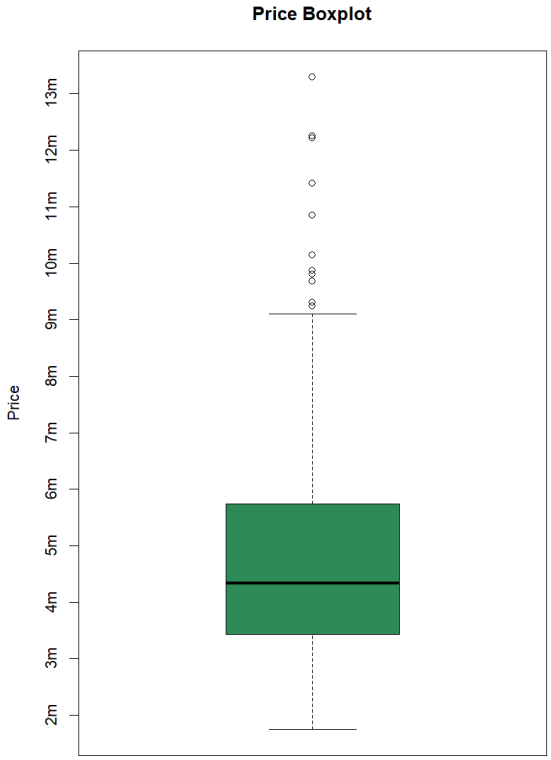


- These Line Plots just confirm the previous observations

Boxplots of columns:

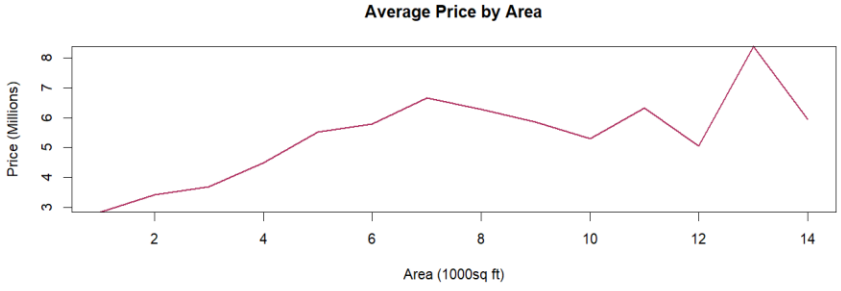
The area is left-skewed with outliers & The Price is slightly left-skewed with outliers

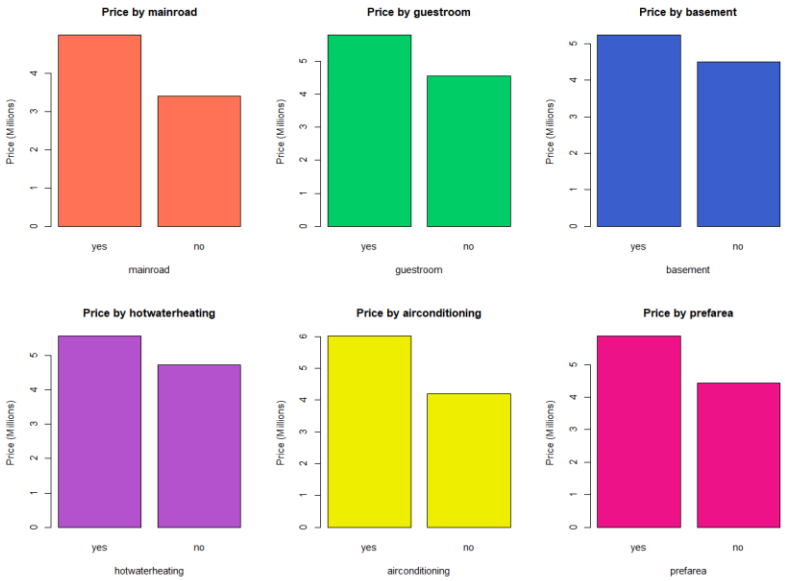




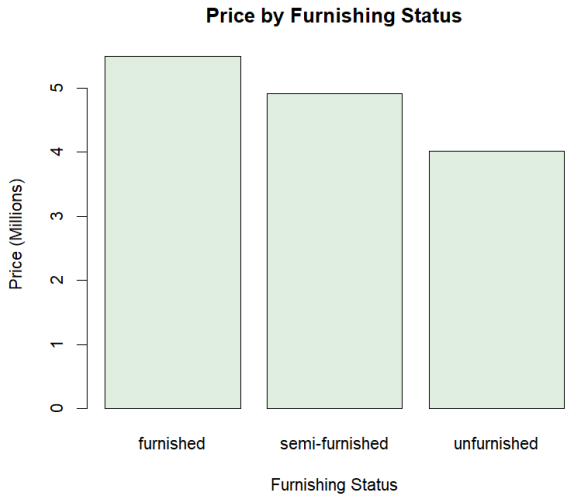
The price increases as the area increases with a drop in prices at 12ft.

The highest price is at 13ft followed by a decline at 15ft.





Houses by the main road, have a guest room, a basement, hot water heating, air conditioning, and are in a preferred area all have a higher price than the ones that aren't

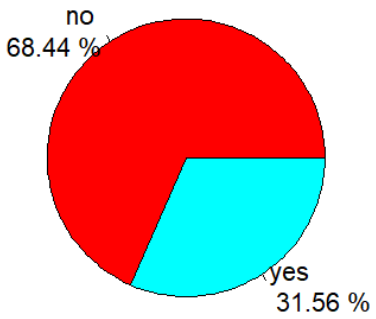


The prices gradually decrease as the furnishing level decreases and the furnished houses are the most expensive

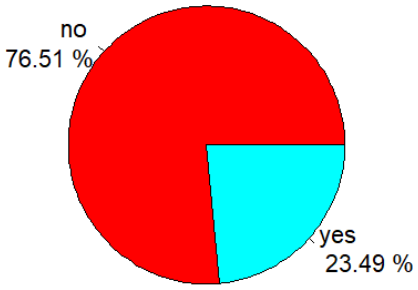
Pie charts:

68% of houses had no air conditioning & 76.5% of houses were not in a preferred area

Pie Chart of airconditioning

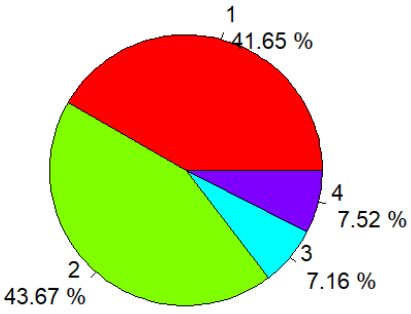


Pie Chart of prefarea

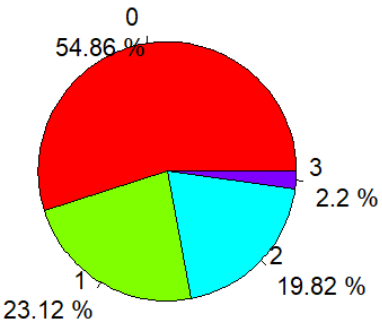


Most houses have 2 stories (43.6%) & Most Houses don't have any parking

Pie Chart of stories

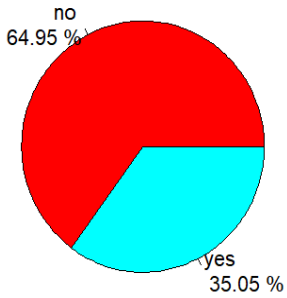


Pie Chart of parking



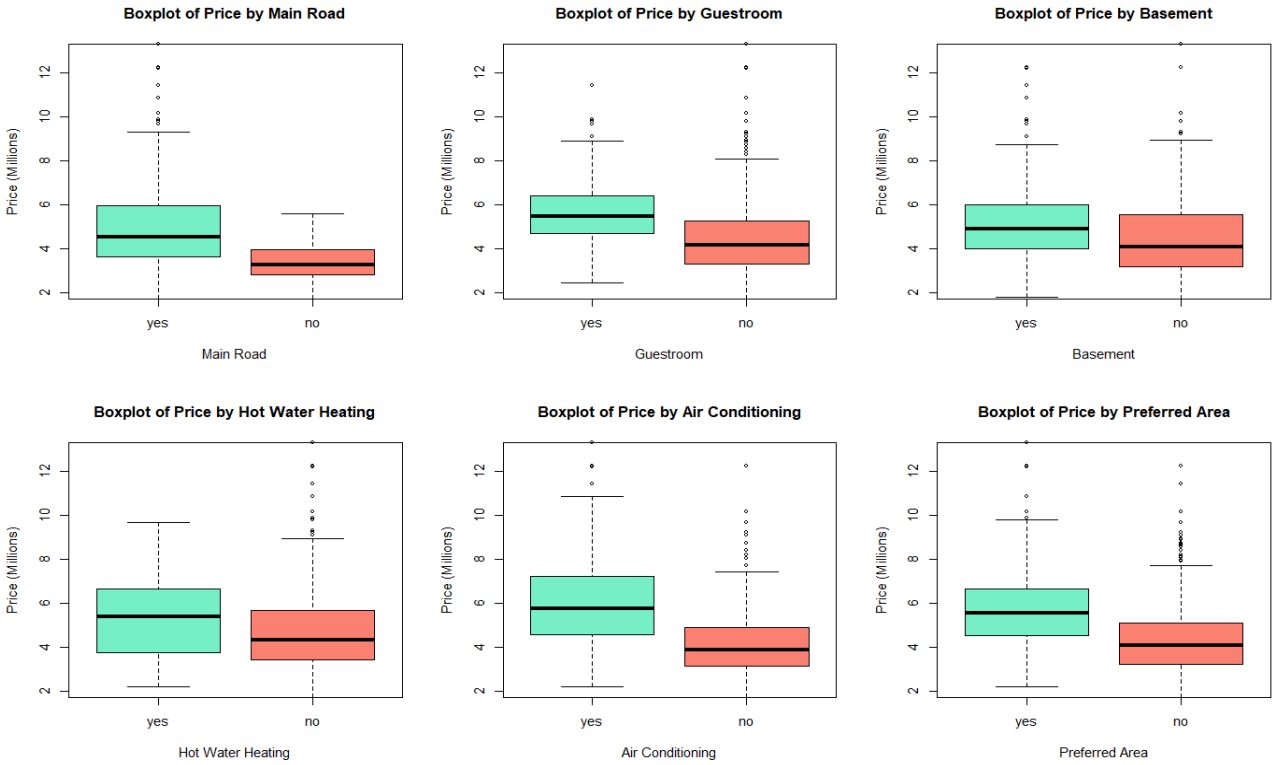
Most Houses didn't have a basement

Pie Chart of basement

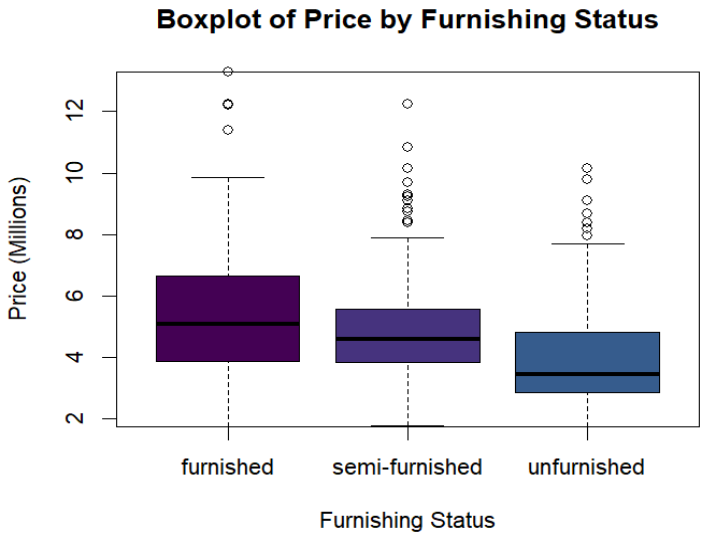


Box Plots:

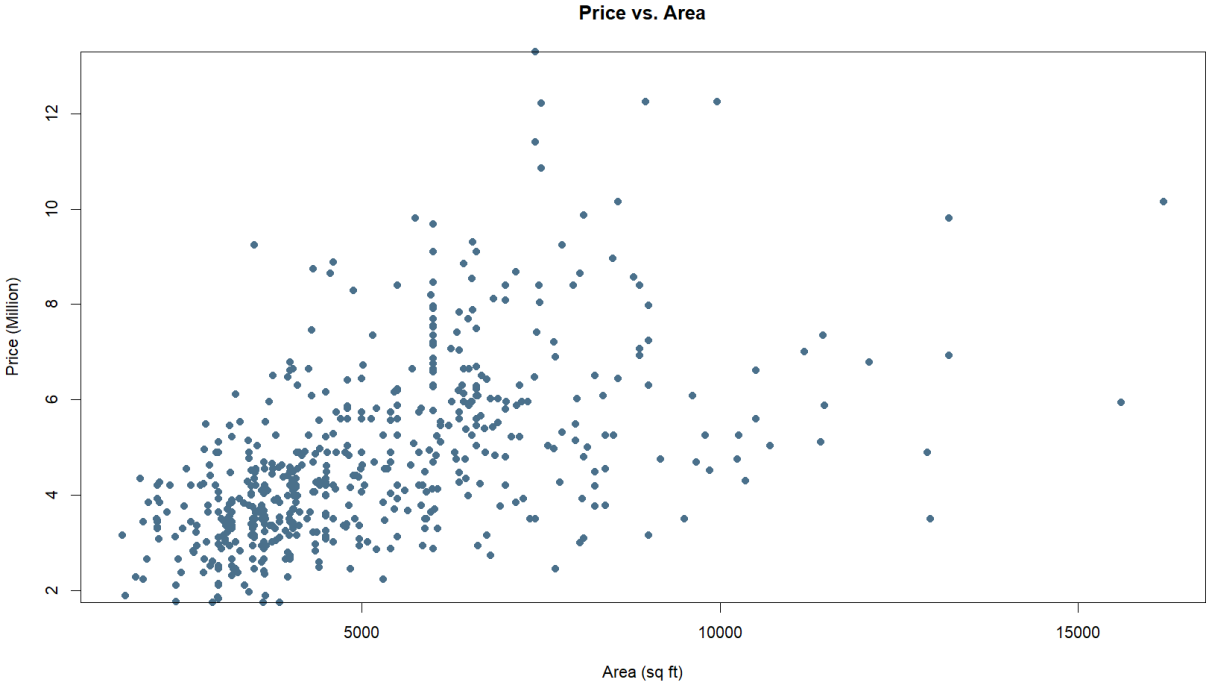
1. **MainRoad: The Yes and No data is left-skewed and yes has outliers.**
2. **Guestroom: Yes and No have a normal distribution and both have outliers.**
3. **Basements: Yes has a normal distribution, No is slightly left skewed, and both have outliers.**
4. **Hot Water: yes is right skewed and no is left-skewed and has outliers.**
5. **Air conditioning: both have normal distributions and outliers.**
6. **Preferred Area: both have normal distributions and outliers.**



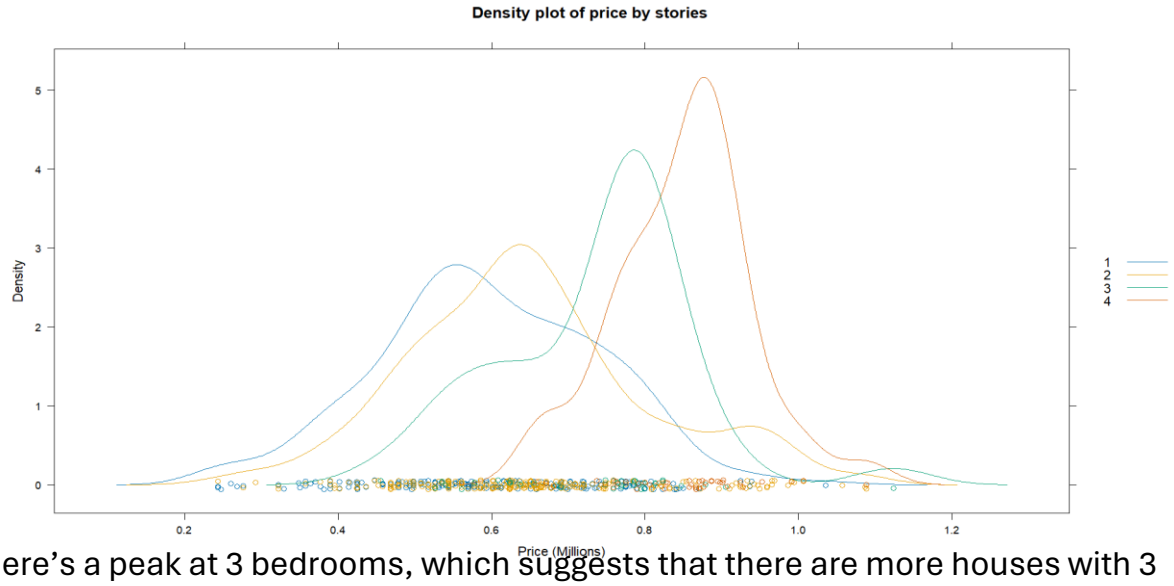
Furnished boxplot: all 3 values have outliers. the first 2 are normally distributed and the third is left skewed.



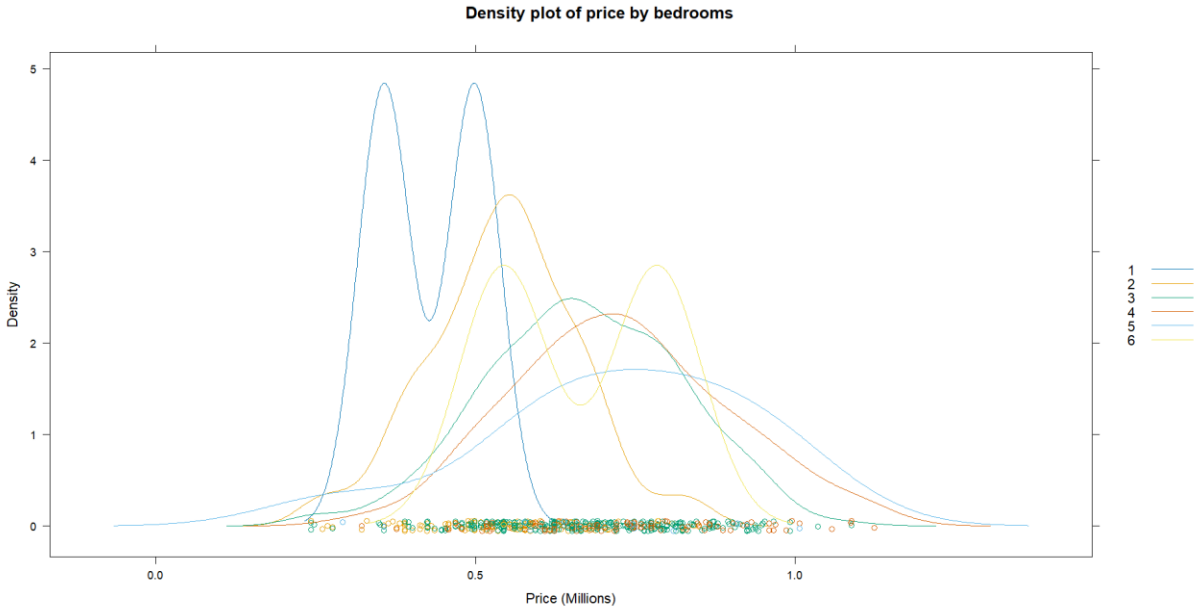
The scatter plot you sent shows a positive correlation between price and area. This means that as the area of a house increases, the price also tends to increase.



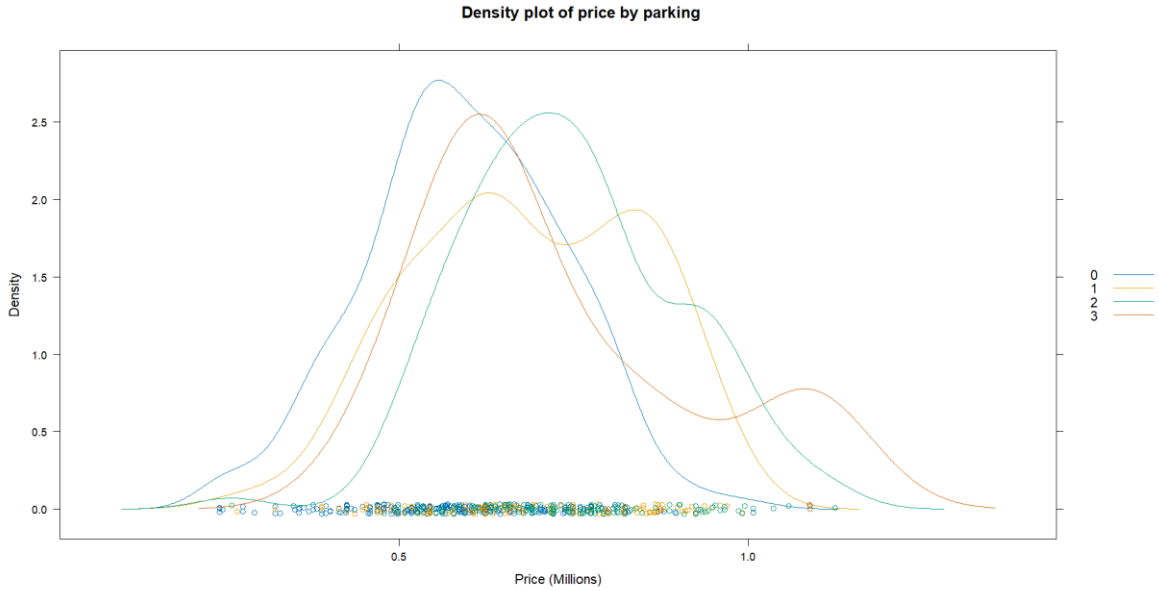
The Density plot suggests that the number of stories is a factor in house price.



There's a peak at 3 bedrooms, which suggests that there are more houses with 3 bedrooms than any other number of bedrooms.



The density graph shows a peak at 2 rooms



E. Hypothesis Testing:

ANOVA (Analysis of Variance):

1st: Null hypothesis: #H0: There is no significant difference in the average price of houses based on furnishing.

Examining Data:

```
         Group.1       x
1      furnished 5495696
2 semi-furnished 4907524
3    unfurnished 4013831
```

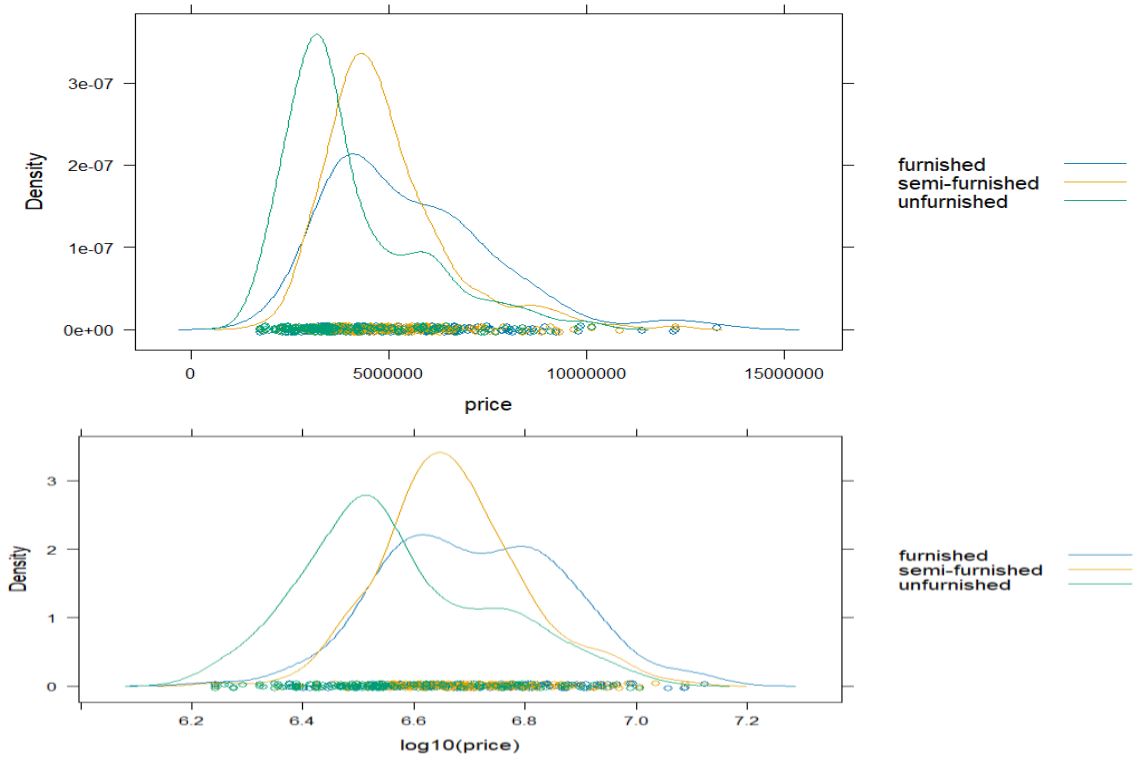
The null hypothesis is **rejected** since the p < significant value:

```
                   Df    Sum Sq  Mean Sq F value   Pr(>F)
furnishingstatus    2 1.798e+14 8.99e+13   28.27 2.09e-12 ***
Residuals         542 1.723e+15 3.18e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tukey’s test(Pair-wise comparison of means): All Null hypotheses are rejected since their P values are all less than the significant level of 0.01

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = (price ~ furnishingstatus), data = df)

$furnishingstatus
                                  diff       lwr        upr     p adj
semi-furnished-furnished     -588171.8 -1038522  -137822.0 0.0063642
unfurnished-furnished       -1481864.5 -1955270 -1008458.9 0.0000000
unfurnished-semi-furnished   -893692.8 -1313257  -474128.5 0.0000023
```

Density plot:



2nd: Null hypothesis: #H0: There is no significant difference in the average price of houses based on the number of parking.
The null hypothesis is **rejected** since the p < significant value:

```
[1] "0" "1" "2" "3"
                Df    Sum Sq   Mean Sq F value Pr(>F)
housing[[var]]   3 2.939e+14 9.797e+13   32.93 <2e-16 ***
Residuals      541 1.609e+15 2.975e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = housing$price ~ housing[[var]])

$`housing[[var]]`
          diff        lwr     upr     p adj
1-0 1054372.17   582292.6 1526452 0.0000001
2-0 1760311.43  1261322.0 2259301 0.0000000
3-0 1731149.94   422584.1 3039716 0.0038950
2-1  705939.26   123095.2 1288783 0.0102120
3-1  676777.78  -666003.4 2019559 0.5640377
3-2  -29161.48 -1381637.9 1323315 0.9999384
```

Tukey's test: there's a difference between having 2 parking spots compared to 1, a difference between having 3 parking spots compared to 1, and a difference between having 3 compared to 2.

3rd: Null hypothesis: #H0: There is no significant difference in the average price of houses based on the number of bedrooms.
The null hypothesis is **rejected** since the p < significant value:

```
[1] "1" "2" "3" "4" "5" "6"
                Df    Sum Sq   Mean Sq F value Pr(>F)
housing[[var]]   5 2.933e+14 5.867e+13   19.64 <2e-16 ***
Residuals      539 1.610e+15 2.987e+12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = housing$price ~ housing[[var]])

$`housing[[var]]`
           diff        lwr     upr     p adj
2-1   919522.06 -2601130.2 4440174 0.9758673
3-1  2242098.13 -1264579.9 5748776 0.4481284
4-1  3017257.89  -514387.6 6548903 0.1433633
5-1  3107300.00  -721332.4 6935932 0.1872987
6-1  2079000.00 -2863743.1 7021743 0.8354139
3-2  1322576.07   811622.8 1833529 0.0000000
4-2  2097735.84  1436825.2 2758647 0.0000000
5-2  2187777.94   568300.1 3807256 0.0017381
6-2  1159477.94 -2361174.3 4680130 0.9354184
4-3   775159.76   193265.4 1357054 0.0021376
5-3   865201.87  -723667.7 2454071 0.6270700
6-3  -163098.13 -3669776.1 3343580 0.9999942
5-4    90042.11 -1553197.5 1733282 0.9999869
6-4  -938257.89 -4469903.4 2593388 0.9739916
6-5 -1028300.00 -4856932.4 2800332 0.9727226
```

Tuckey's test shows that the rows with p > significant value and there's a significant difference between both values

4th: Null hypothesis: #H0: There is no significant difference in the average price of houses based on the number of Bathrooms.
The null hypothesis is **rejected** since the p < significant value:

Tuckey's test shows that each different number of bathrooms doesn't have effects against each other.

```
[1] "1" "2" "3" "4"
                 Df    Sum Sq   Mean Sq F value Pr(>F)
housing[[var]]    3 3.580e+14 1.193e+14   41.78 <2e-16 ***
Residuals       541 1.545e+15 2.856e+12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = housing$price ~ housing[[var]])

$`housing[[var]]`
         diff       lwr       upr     p adj
2-1  593414.9  189362.2  997467.7 0.0009804
3-1 1514777.3  759844.0 2269710.6 0.0000020
4-1 3037791.2 2298738.2 3776844.1 0.0000000
3-2  921362.4  168991.2 1673733.5 0.0091363
4-2 2444376.2 1707940.7 3180811.7 0.0000000
4-3 1523013.9  548846.0 2497181.7 0.0003721
```

T-Test:

- We are evaluating the hypothesis of the effect of some categories on the price using T-Test
- Null hypothesis: #H0: There is no significant difference in the average price of houses based on the following categories (main road, guestroom, basement, hot water heating, air-conditioning, and preferred area) .
- The null hypothesis are all **rejected** since the T > 0 (far enough from zero) and the p is very small

```
[1] "mainroad"

        Two Sample t-test

data:  y and n
t = 7.2451, df = 543, p-value = 1.49e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1161003 2024742
sample estimates:
mean of x mean of y
  4991777   3398905

[1] "guestroom"

        Two Sample t-test

data:  y and n
t = 6.1586, df = 543, p-value = 1.429e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  850177.7 1646524.9
sample estimates:
mean of x mean of y
  5792897   4544546

[1] "basement"

        Two Sample t-test

data:  y and n
t = 4.4372, df = 543, p-value = 1.104e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  408305 1056994
sample estimates:
mean of x mean of y
  5242615   4509966

[1] "hotwaterheating"

        Two Sample t-test

data:  y and n
t = 2.1783, df = 543, p-value = 0.02982
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
   81649.17 1581084.52
sample estimates:
mean of x mean of y
  5559960   4728593
```

```
[1] "airconditioning"

        Two Sample t-test

data:  y and n
t = 11.839, df = 543, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1519092 2123469
sample estimates:
mean of x mean of y
  6013221   4191940

[1] "prefarea"

        Two Sample t-test

data:  y and n
t = 8.1399, df = 543, p-value = 2.718e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1102927 1804567
sample estimates:
mean of x mean of y
  5879046   4425299
```

- The air conditioning has the highest value of T (11.839) which means that the presence of air conditioning has the highest effect on the pricing

Dataset preparation (Test, Train)

```
housing$mainroad <- factor(housing$mainroad, levels = c("yes", "no"))
housing$guestroom <- factor(housing$guestroom, levels = c("yes", "no"))
housing$basement <- factor(housing$basement, levels = c("yes", "no"))
housing$hotwaterheating <- factor(housing$hotwaterheating, levels = c("yes", "no"))
housing$airconditioning <- factor(housing$airconditioning, levels = c("yes", "no"))
housing$prefarea <- factor(housing$prefarea, levels = c("yes", "no"))
housing$furnishingstatus <- factor(housing$furnishingstatus, levels = c("furnished", "semi-furnished", "unfurnished"))
```

Convert categorical variables to factors

Data Analytics Technique:

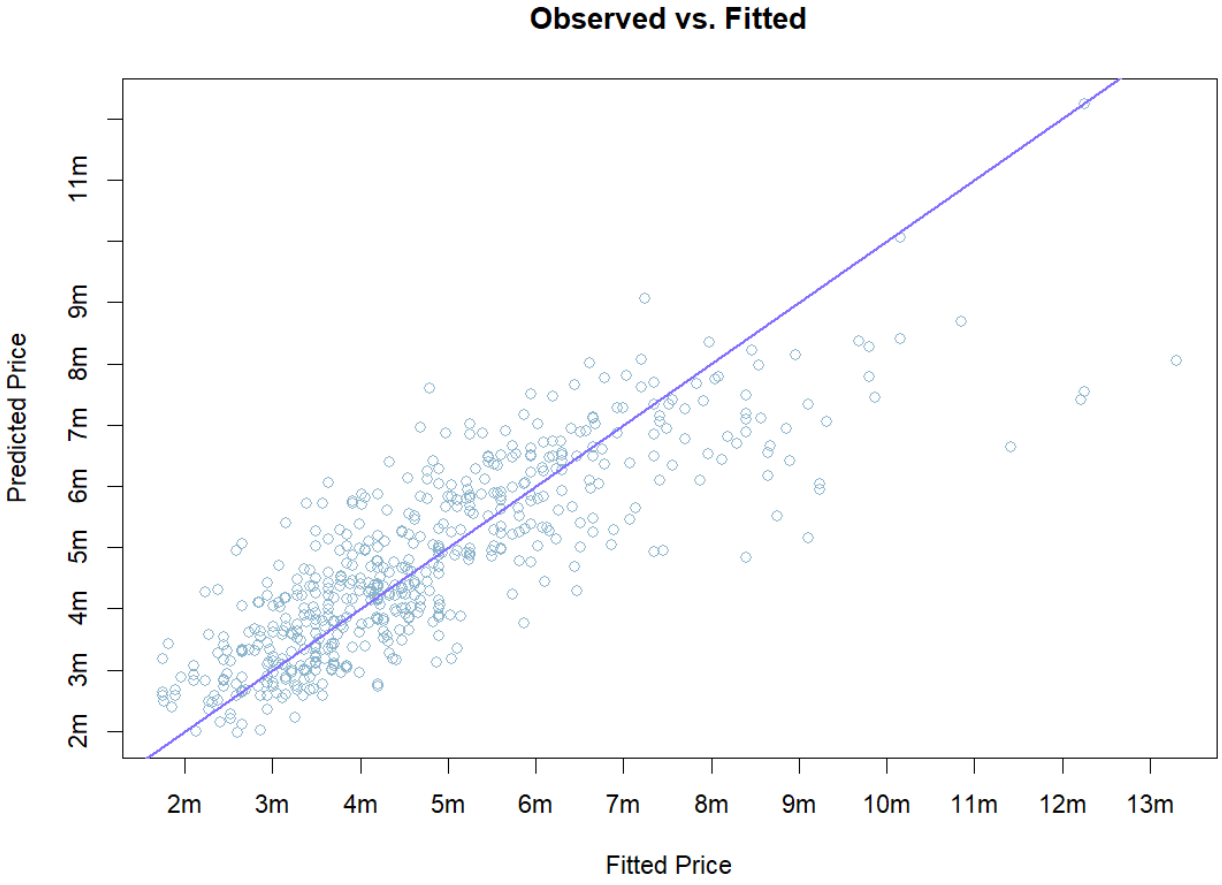
Linear Regression was used to predict house prices.

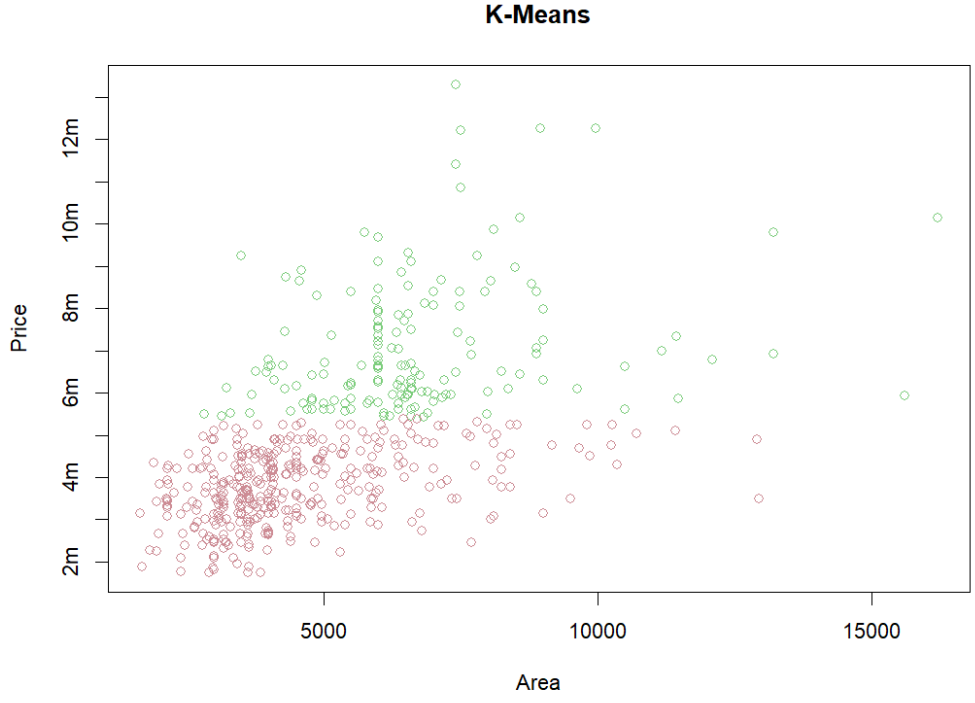
The predicted values are continuous so we can't use a classification model and our data is labeled so K-means shouldn't be used as well.

Performance Measures

```
> attributes(fit)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "contrasts"     "xlevels"       "call"          "terms"
[13] "model"

$class
[1] "lm"


Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5251832.51  828091.85   6.342 4.92e-10 ***
area                              235.62      24.96   9.442  < 2e-16 ***
bedrooms2                     -121224.21  763298.00  -0.159 0.873875
bedrooms3                      119897.61  766050.81   0.157 0.875689
bedrooms4                      204453.78  776535.34   0.263 0.792431
bedrooms5                      352331.32  843096.88   0.418 0.676191
bedrooms6                      861907.29 1076888.59   0.800 0.423862
bathrooms2                     889309.60  121853.03   7.298 1.10e-12 ***
bathrooms3                    2087419.84  357299.19   5.842 9.08e-09 ***
bathrooms4                    5309187.73 1132696.27   4.687 3.54e-06 ***
stories2                       300434.37  121020.21   2.483 0.013360 *
stories3                       832268.81  208078.72   4.000 7.26e-05 ***
stories4                      1406036.05  216192.56   6.504 1.84e-10 ***
mainroadno                    -433968.54  143672.89  -3.021 0.002647 **
guestroomno                   -284657.48  132188.26  -2.153 0.031743 *
basementno                    -373520.17  112121.12  -3.331 0.000925 ***
hotwaterheatingno             -839134.50  226392.66  -3.707 0.000233 ***
airconditioningno             -843080.22  109913.76  -7.670 8.48e-14 ***
parking1                       396938.75  119101.07   3.333 0.000921 ***
parking2                       622112.73  130978.04   4.750 2.64e-06 ***
parking3                       112217.86  336032.76   0.334 0.738552
prefareano                    -647611.82  118363.86  -5.471 6.95e-08 ***
furnishingstatussemi-furnished  -27717.11  118038.66  -0.235 0.814445
furnishingstatusunfurnished    -417022.33  127026.27  -3.283 0.001096 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1066000 on 521 degrees of freedom
Multiple R-squared:  0.6889,    Adjusted R-squared:  0.6752
F-statistic: 50.17 on 23 and 521 DF,  p-value: < 2.2e-16
```

```
> summary(fit)

Call:
lm(formula = price ~ ., data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-2798873  -663461   -59293   509704  5249000
```


Observed vs. Fitted
Predicted Price
2m
3m
4m
5m
6m
7m
8m
9m
11m
2m
3m
4m
5m
6m
7m
8m
9m
10m
11m
12m
13m
Fitted Price


K-Means
Price
2m
4m
6m
8m
10m
12m
5000
10000
15000
Area