



MH3511 Data Analysis with Computer Group Project

Group 3

Name	Matriculation Number
Khant Nyi Nyi	U2022568J
Choy Xin Yun	U2021811D
Goh Peng Aik	U2022363E
Tan Yi Jie Joel	U2021190G
Chen Wei May (Zeng Weimei)	U2020687E

Table of Contents

1. Introduction	3
2. Data Description	3
3. Exploratory Analysis of Dataset	5
3.1 Summary statistics for <i>price</i>	5
3.2 Summary statistics for <i>carat</i>	6
3.3 Summary statistics for <i>cut</i> , <i>color</i> and <i>clarity</i>	7
3.4 Summary statistics for dimensions (<i>x</i> , <i>y</i> and <i>z</i>)	7
3.5 Summary statistics for proportions (<i>table_percentage</i> and <i>depth_percentage</i>).....	10
4. Statistical Analysis	10
4.1 Relation between <i>log(price)</i> and <i>log(carat)</i>	10
4.2 Relation between <i>log(price)</i> and the categorical variables (<i>cut</i> , <i>color</i> , <i>clarity</i>)	11
4.3 Relation between <i>log(price)</i> and <i>x</i> , <i>y</i> , <i>z</i> respectively	15
4.4 Relation between <i>log(price)</i> and <i>table_percentage</i>	17
4.5 Relation between <i>log(price)</i> and <i>depth_percentage</i>	17
4.6 Relation between <i>log(price)</i> and standardised <i>x</i> , <i>y</i> , <i>z</i> and <i>log(carat)</i> respectively.....	18
5. Conclusion and Discussion	19
6. Appendix	20
7. References	68

1. Introduction

Diamond is the one of the hardest materials on Earth and has long since been recognized for its beauty as a gemstone. Some 142 million carats of diamonds were estimated to have been produced from mines worldwide in 2019 [1]. Although the majority of mined diamonds are not suitable for use as jewellery, the diamond gemstone market is still big, with the value of the United States' diamond jewellery market being 35 billion US dollars in 2020 alone [2]. However, shopping for diamonds can be a tricky business. With prices ranging from a few hundred dollars to tens of thousands of dollars, the cost for a diamond seems volatile. There are also numerous attributes related to the grade of a diamond gemstone. There may also be differences within each grade of a diamond gemstone that may be difficult for the average consumer to distinguish.

Hence, in our project, we aim to find out how much the price of a diamond gemstone depends on its various attributes so that buyers would be better informed of features of a diamond that they should look out for or avoid. The questions we aim to answer in our study are:

- Is price dependent on the 4 'C's (carat, cut, color, clarity) of the diamonds?
- If so, how much does price depend on the 4 'C's (carat, cut, color, clarity) of the diamonds?
- Is price dependent on the dimensions of the diamonds?
- If so, how much does the price depend on the dimensions of the diamonds?
- Is price dependent on the proportion of the dimensions of the gemstones?
- If so, how much does the price depend on the proportion of the diamonds?
- Of all the factors that the price of the diamonds is dependent on, which factor is the most significant?

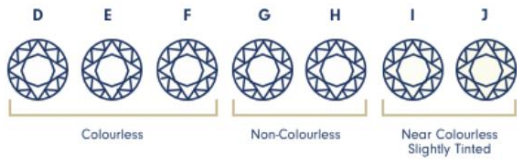

2. Data Description

The 'diamonds' dataset is obtained from the ggplot2 package in R studio. Alternatively, it is also available on kaggle for uses outside of R studio.

The data set contains 53,940 records of individual non-coloured diamond gemstones with 10 unique attributes. Before proceeding with the exploratory data analysis, these steps are carried out to clean the data:

- Variable name 'depth' renamed to 'depth_percentage' to avoid confusion with the variable 'z'.
- Variable name 'table' renamed to 'table_percentage' to avoid confusion with the absolute measurement of table size.

The variables are as such:

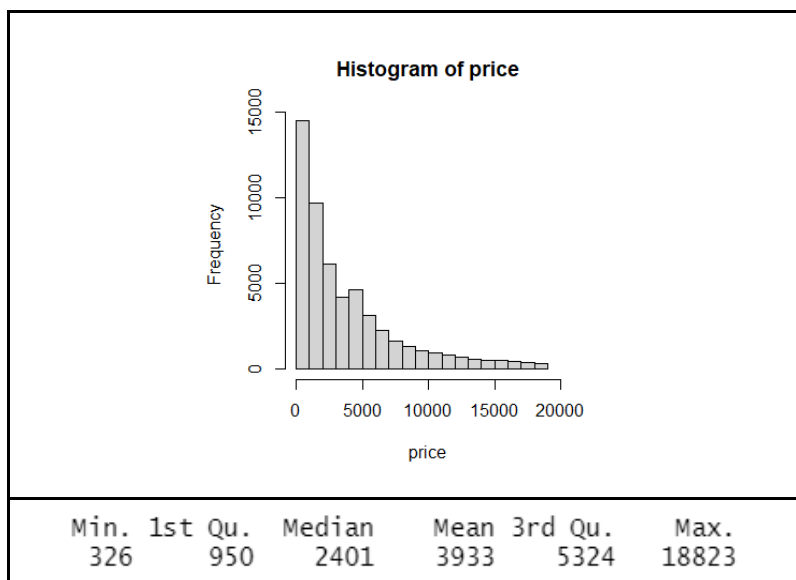
price	The price of the gemstones in US dollars.
carat	The weight of the gemstones, measured in carat. A metric 'carat' is equal to 200 mg.
cut	Descriptive grade given to the quality of cut made on the gemstones. Grade in increasing order: Fair, Good, Very Good, Premium, Ideal.
color	<p>Alphabetical grade given to the quality of color of the gemstones. For non-coloured gemstones, grade is determined by how 'colorless' the gemstones are, with D being the clearest and J being the most tainted[3].</p> 
clarity	<p>Descriptive grade given to the inclusions of internal flaws and surface imperfections in the gemstones[3]. The dataset contains diamonds ranging from IF to I1 clarity grade, where IF has the least internal flaws, and I1 has the most.</p> 
x	Length of the gemstones in mm
y	Width of the gemstones in mm
z	Depth of the gemstones in mm
depth_percentage	<p>The depth of the gemstones' table, expressed as a percentage of its average diameter.</p> $\text{depth percentage} = \frac{z}{\text{mean}(x + y)}$ <p>Depth percentage is a measure of how proportionate the gemstones are.</p>
table_percentage	<p>The width of the gemstones' table (flat surface at the top of the gemstone) expressed as a percentage of its widest length. Table percentage is another measure of how proportionate the gemstones are.</p>

3. Exploratory Analysis of Dataset

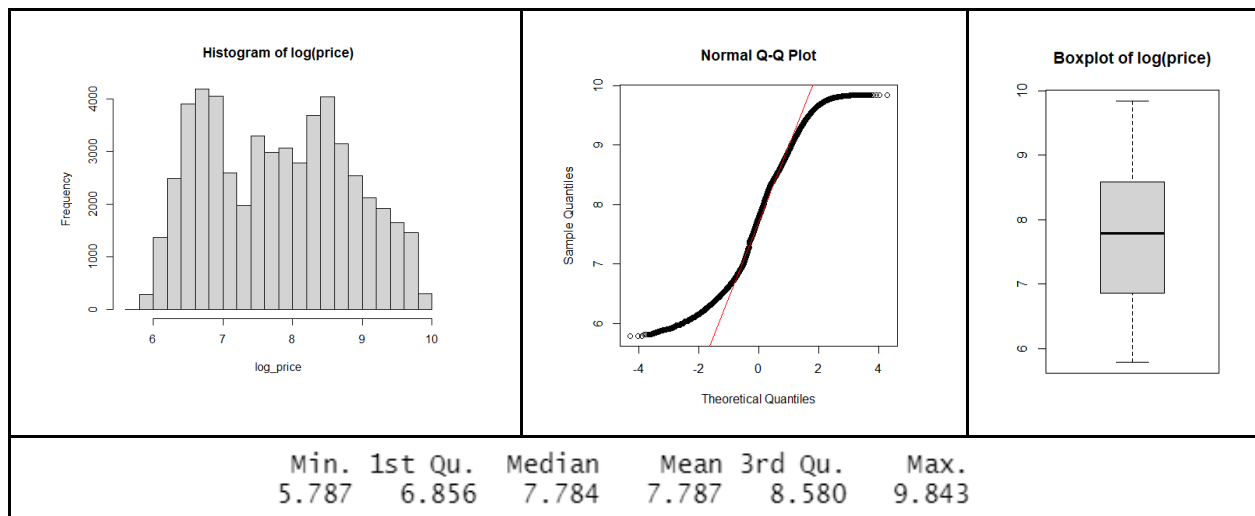
In this section, we look at the individual variables in more detail. The numerical variables were investigated to detect possible outliers, and/or to perform a transformation to avoid highly skewed data. Similarly, the categorical variables were investigated to detect any anomalies.

3.1 Summary statistics for *price*

The data for *price* seems to follow an exponential distribution with a median of 2401 but a max of 18,823. As the data is highly skewed, a log-transformation (of base e) is performed on *price* to make the distribution less skewed.

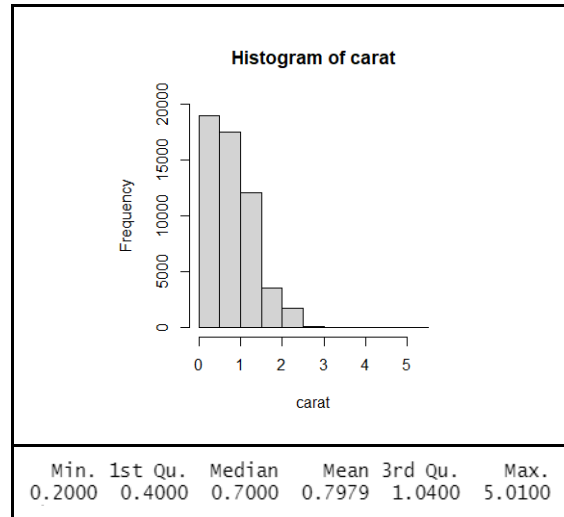


The log transformed price data is more symmetric, has no outliers and is more like a normal distribution. Hence, it would be better suited for use in some models for analysis.

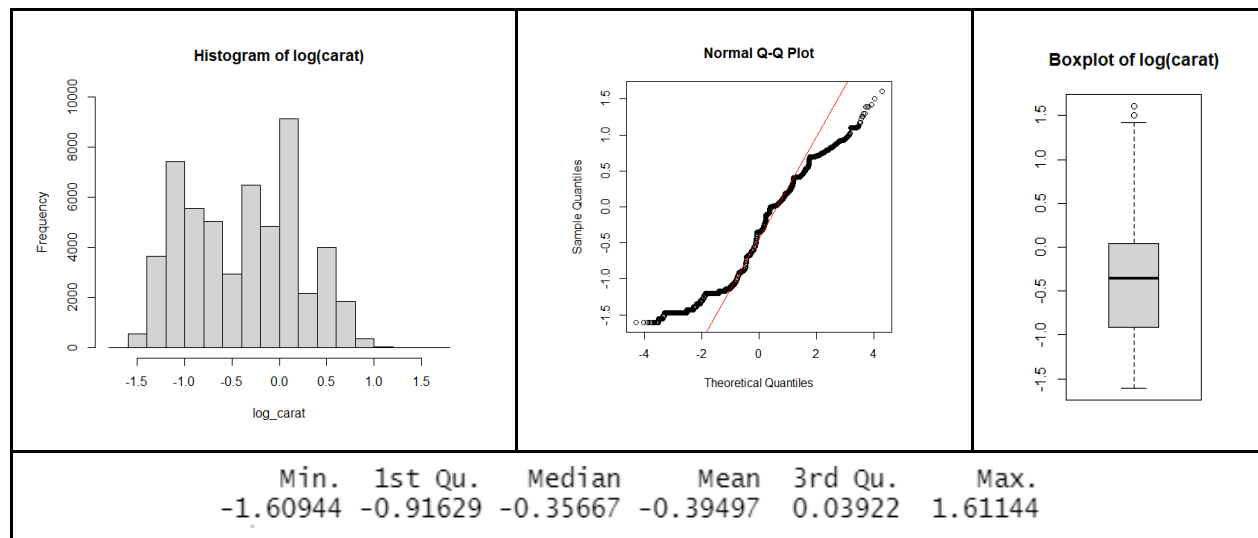


3.2 Summary statistics for *carat*

The values for *carat* also appear highly skewed with close to 20,000 diamonds weighing less than 0.5 carats. As such, a log-transformation (of base e) was also performed.

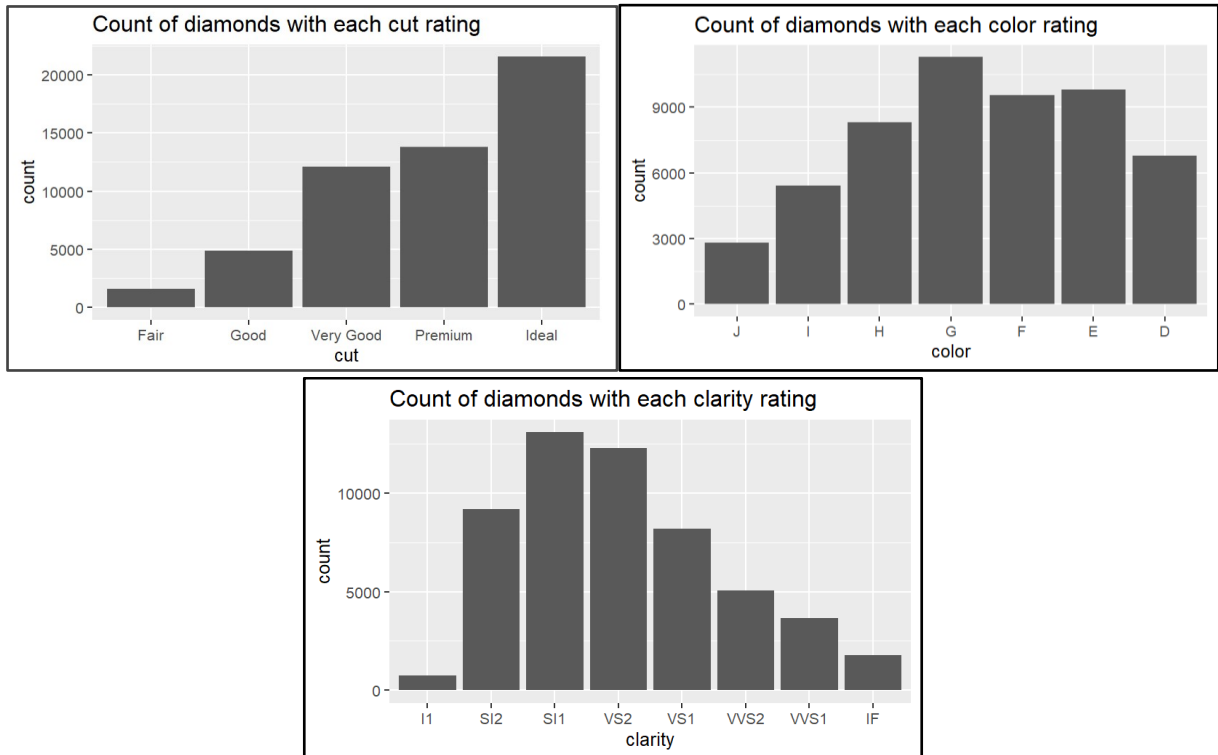


The log transformed *carat* data is less skewed and follows a normal distribution more closely. Hence, it would be better suited for use in some models for analysis. No records were removed.



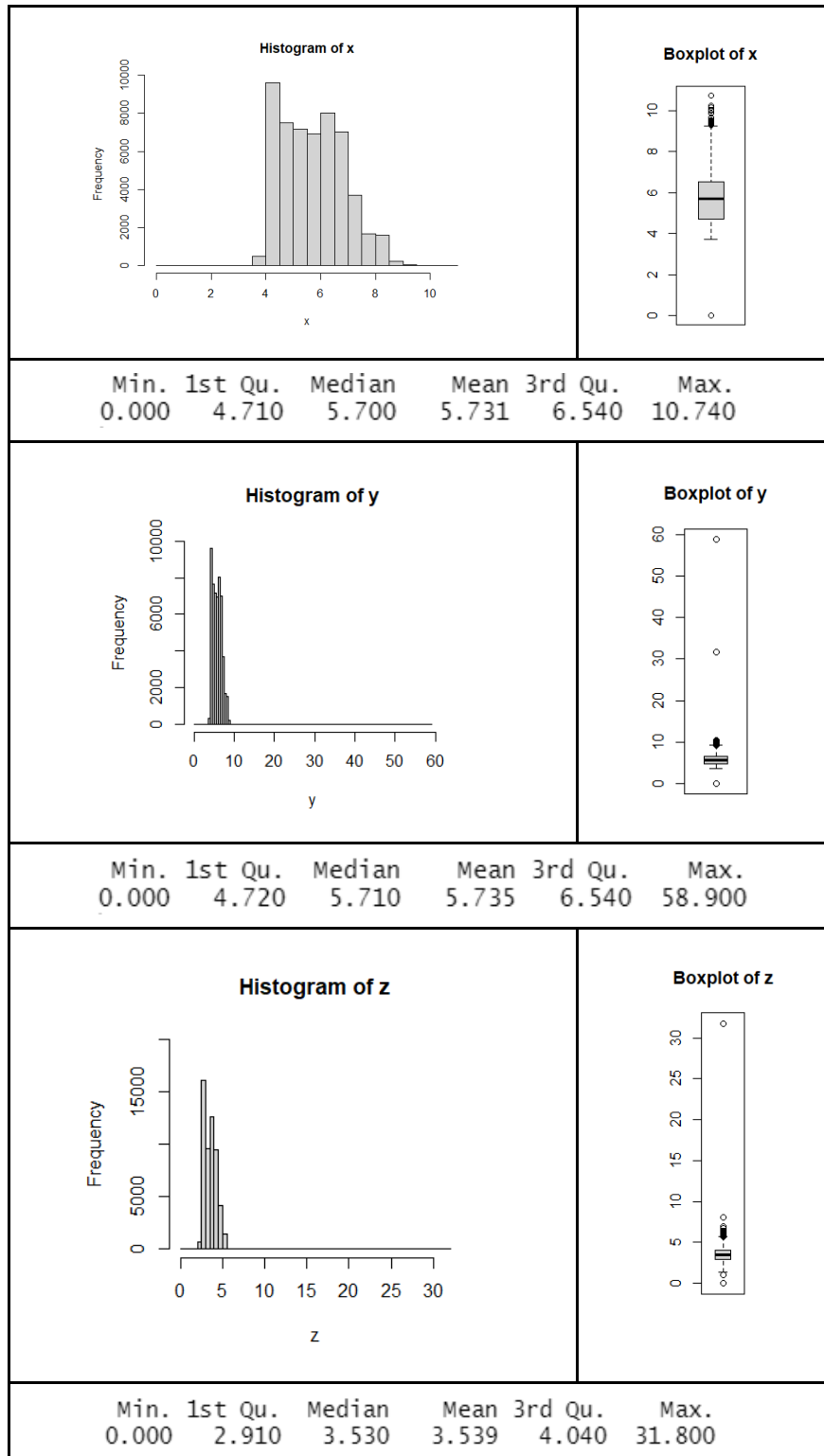
3.3 Summary statistics for *cut*, *color* and *clarity*

The bar charts show the count of the number of diamonds for each categorical variable. For cut, color and clarity, there are no outliers. No records were removed.

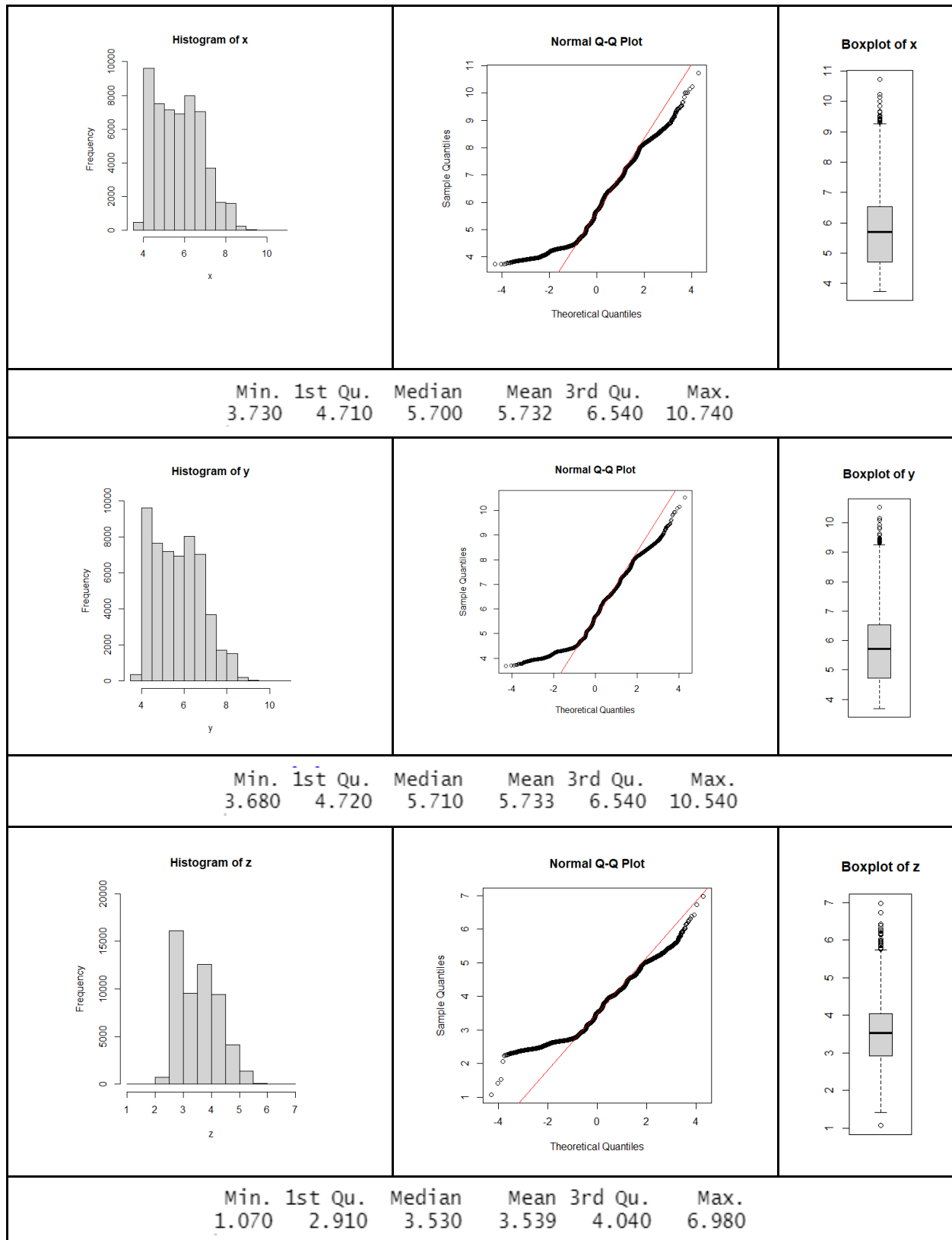


3.4 Summary statistics for dimensions (x, y and z)

Most of the values for each dimension of the diamonds are clustered around their median. However, there are records with one or more of its dimensions having 0 as the value. There are also records with anomalously large y or z values. Upon closer inspection, these values seem to be entered erroneously as the depth percentage value calculated from the given dimension values do not equate to the corresponding *depth_percentage* value. As such, records with dimension values equal to 0 or anomalously large dimension values were removed. A total of 24 records were removed.

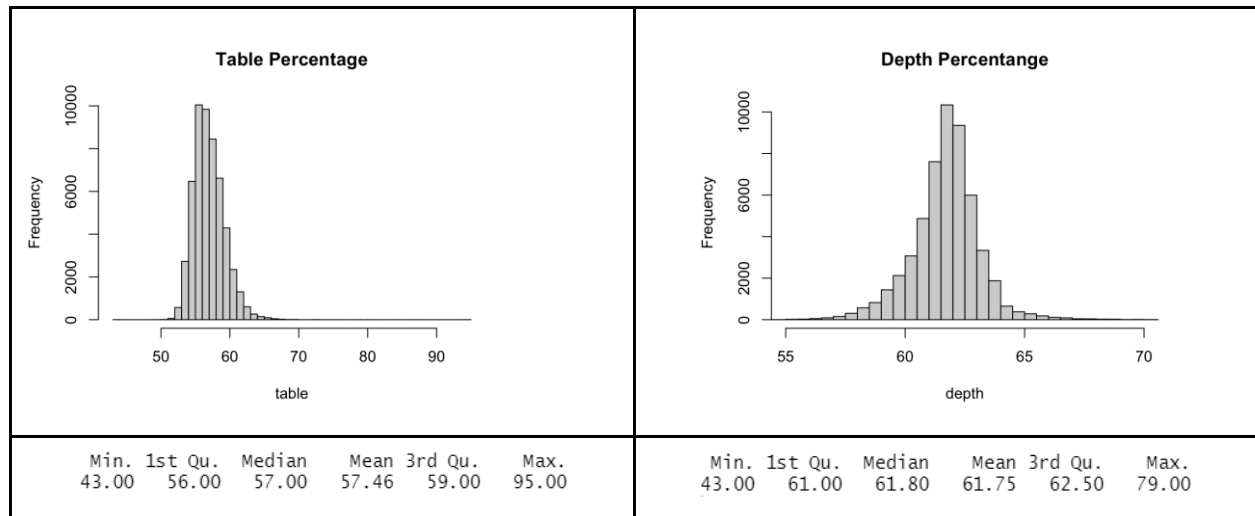


Below are the summary statistics after trimming:



3.5 Summary statistics for proportions (*table_percentage* and *depth_percentage*)

Both *table_percentage* and *depth_percentage* values appear to be distributed symmetrically about the median but both distributions have long tails. However, no records appear to be anomalous. As such, no records were removed.



At the end of the exploratory data analysis, a total of 24 records were removed, leaving us with 53,916 records for the statistical analysis.

4. Statistical Analysis

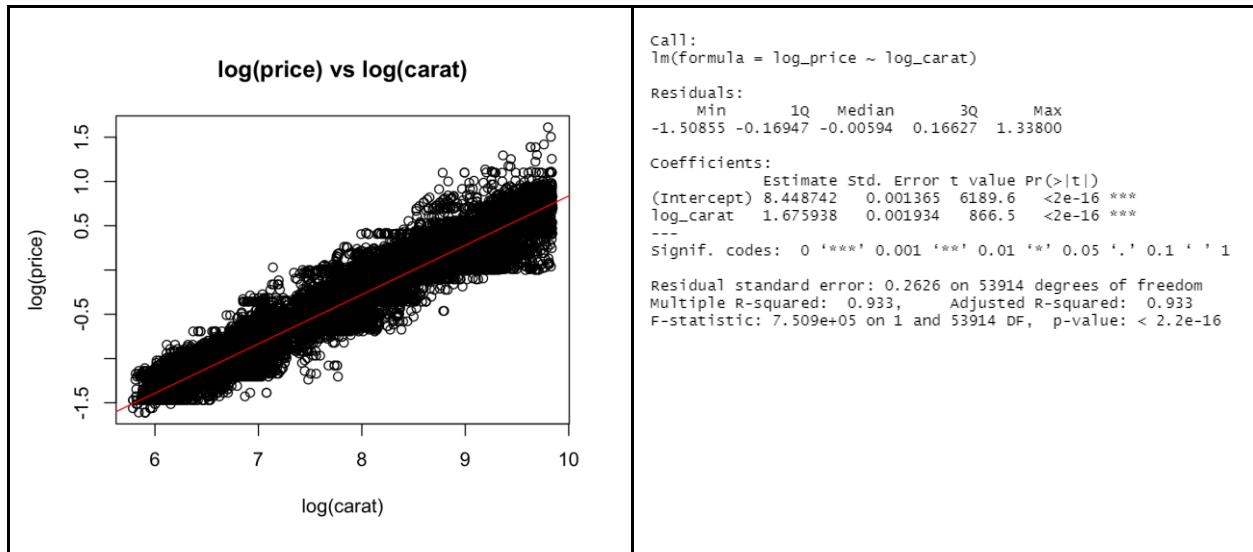
4.1 Relation between *log(price)* and *log(carat)*

In this section, we investigate whether the price of diamonds is dependent on its carat. A simple linear regression of *log(price)* against *log(carat)* was performed and the Pearson correlation coefficient was calculated.

The Pearson correlation coefficient between *log(price)* and *log(carat)* is 0.9659, and the 95% confidence interval calculated is [0.9653, 0.9664]. This indicates that there is likely a strong positive linear relationship between *log(price)* and *log(carat)*.

The linear regression model below returns a p-value ($<2 \times 10^{-16}$) that is much smaller than 0.05. This indicates a statistically significant relationship between *log(price)* and *log(carat)* at 0.05 level of significance. The R-squared value is above 90% (R-squared=0.9330), further supporting the strong positive linear correlation between *log(price)* and *log(carat)*.

As a result, we conclude that the price of a diamond is dependent on its carat.

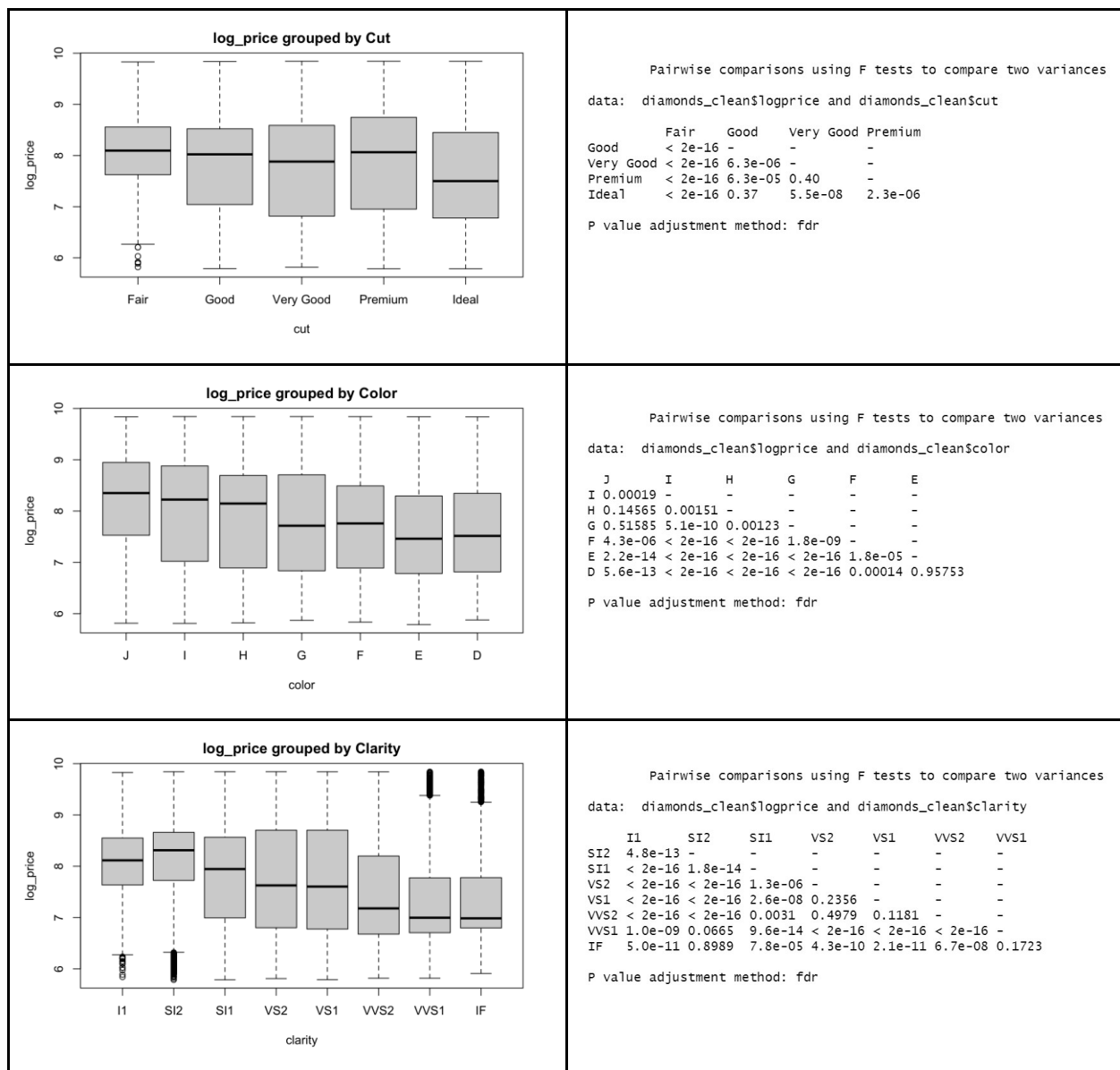


4.2 Relation between $\log(\text{price})$ and the categorical variables (*cut*, *color*, *clarity*)

In this dataset, for each categorical variable, the categories have been ordered as per the description. Given the categories, one would expect that a better *cut*, *color* or *clarity* grade would result in a higher *price*. Below are the boxplots of each variable against $\log(\text{price})$, whereby the grades are arranged from worst to best from left to right.

From the boxplots, there are some slight differences in the median $\log(\text{price})$ across grades for each category. For *cut*, the medians of $\log(\text{prices})$ are roughly similar across grades. For *color*, the median $\log(\text{price})$ generally decreased across the color grades from J to D. For *clarity*, the median $\log(\text{price})$ generally decreased across the grades.

To determine if ANOVA should be performed, a pairwise F-test was carried out between each group to determine if the variances can be assumed to be similar. The p-values are mostly less than 0.05, and hence we conclude that the variances are not the same. As such, we will use the Kruskal-Wallis Test instead of ANOVA to determine if the populations come from identical distributions.



Kruskal-Wallis Test was then performed, as shown below, for each variable *cut*, *color* and *clarity* respectively. Since the p-values for *cut*, *color* and *clarity* each are all much smaller than 0.05, we conclude that the populations do not come from identical distributions.

Kruskal-wallis rank sum test

data: diamonds_clean\$logprice and diamonds_clean\$cut

Kruskal-wallis chi-squared = 975.03, df = 4, p-value < 2.2e-16

Kruskal-Wallis rank sum test

```
data: diamonds_clean$logprice and diamonds_clean$color
Kruskal-Wallis chi-squared = 1333, df = 6, p-value < 2.2e-16
```

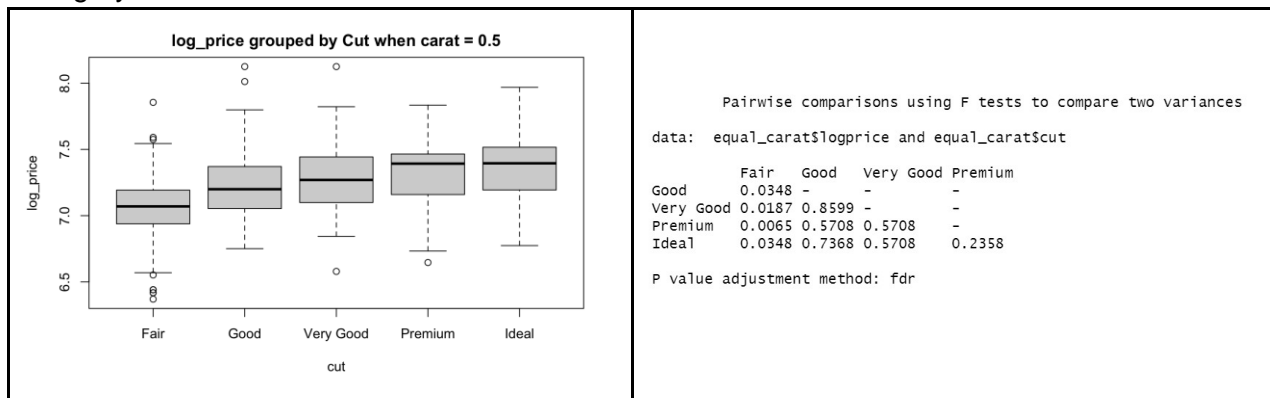
Kruskal-Wallis rank sum test

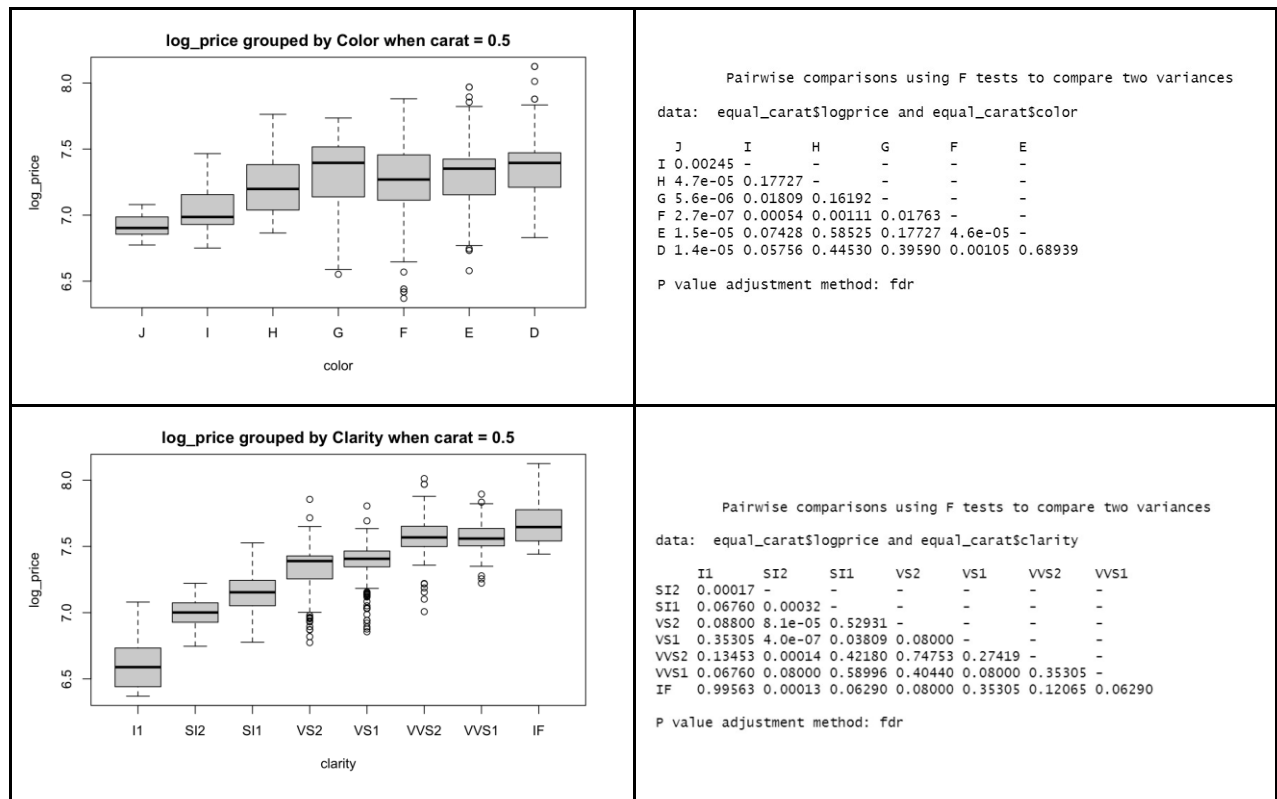
```
data: diamonds_clean$logprice and diamonds_clean$clarity
Kruskal-Wallis chi-squared = 2714.8, df = 7, p-value < 2.2e-16
```

We then calculated the Spearman rank correlation between $\log(\text{price})$ and *cut*, *color* and *clarity* each to determine if there exists a positive correlation as one would expect the price to increase when the *cut*, *color* and *clarity* improves. Since each variable has at least 5 grades that are ordered, correlation coefficients between $\log(\text{price})$ and *cut*, *color* and *clarity* each are calculated by assigning each grade a number: 1 to 5 for *cut*, 1 to 7 for *color*, and 1 to 8 for *clarity*. The Spearman rank correlation values are -0.09294, -0.1500, and -0.2114 for *cut*, *color* and *clarity* respectively, which indicate that there is no strong positive correlation between $\log(\text{price})$ and *cut*, *color*, *clarity*.

However, it can be observed at a fixed carat value, for example $\text{carat}=0.5$, the observed relationships between $\log(\text{price})$ and the categorical variables are more in line with expectations. From the boxplots below, there seems to be an increasing trend with $\log(\text{price})$ as the *cut*, *color*, *clarity* improves.

Like before, pairwise F-tests are carried out to determine if the variances can be assumed to be equal. As shown below, there is at least one p-value that is less than 0.05 for each variable. Thus, we conclude that the variances cannot be assumed to be equal across the grades for each category.





Kruskal-Wallis Test was then carried out for each categorical variable. From the results shown below, the p-values are all much smaller than 0.05 with a reduced sample size of 1257 samples (where the carat = 0.5). As such, we conclude that the populations do not come from identical distributions.

<p>Kruskal-wallis rank sum test</p> <p>data: equal_carat\$logprice and equal_carat\$cut</p> <p>Kruskal-wallis chi-squared = 114.59, df = 4, p-value < 2.2e-16</p>
<p>Kruskal-wallis rank sum test</p> <p>data: equal_carat\$logprice and equal_carat\$color</p> <p>Kruskal-wallis chi-squared = 153.58, df = 6, p-value < 2.2e-16</p>
<p>Kruskal-wallis rank sum test</p> <p>data: equal_carat\$logprice and equal_carat\$clarity</p> <p>Kruskal-wallis chi-squared = 810.64, df = 7, p-value < 2.2e-16</p>

Similar to before, Spearman rank correlation was then calculated between $\log(\text{price})$ and cut , color , clarity each, which are 0.2789, 0.1917 and 0.7883 respectively. From these results, only clarity has a strong positive relationship with $\log(\text{price})$, but it does indicate that there exists a positive relationship between $\log(\text{price})$ and cut , color , clarity when the $\text{carat}=0.5$, which aligns with expectations.

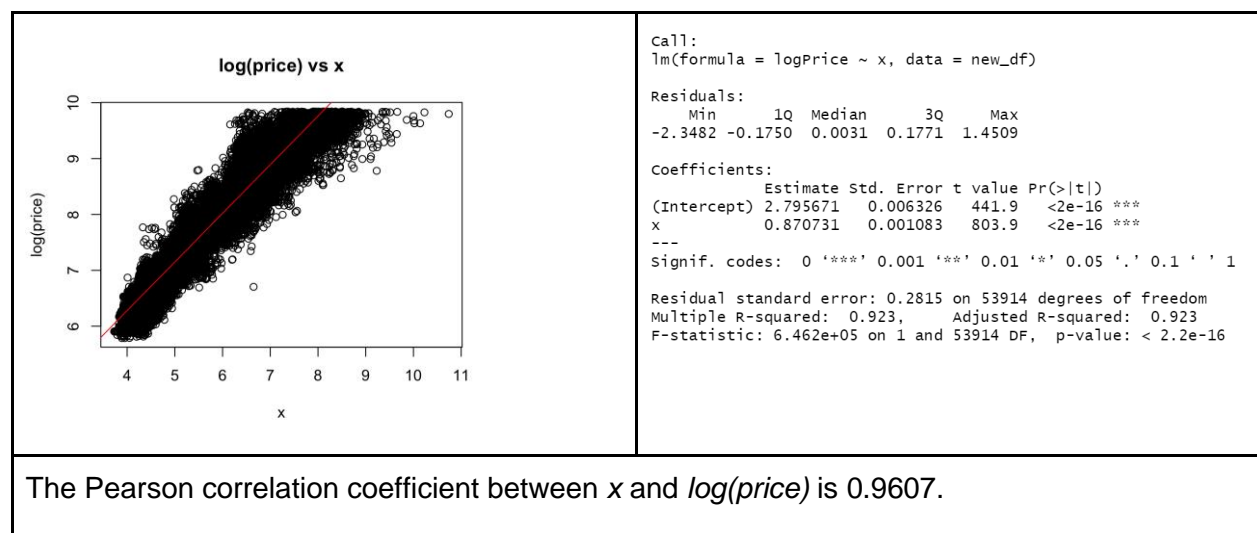
As such, the overall conclusion is that individually, cut , color , and clarity do not affect the price of diamonds. But if the carat of the diamonds is fixed at a value, cut , color and clarity may have a significant effect on price . However, since only one carat value (0.5) was investigated, it is difficult to conclude if a similar effect can also be observed for other carat values.

4.3 Relation between $\log(\text{price})$ and x , y , z respectively

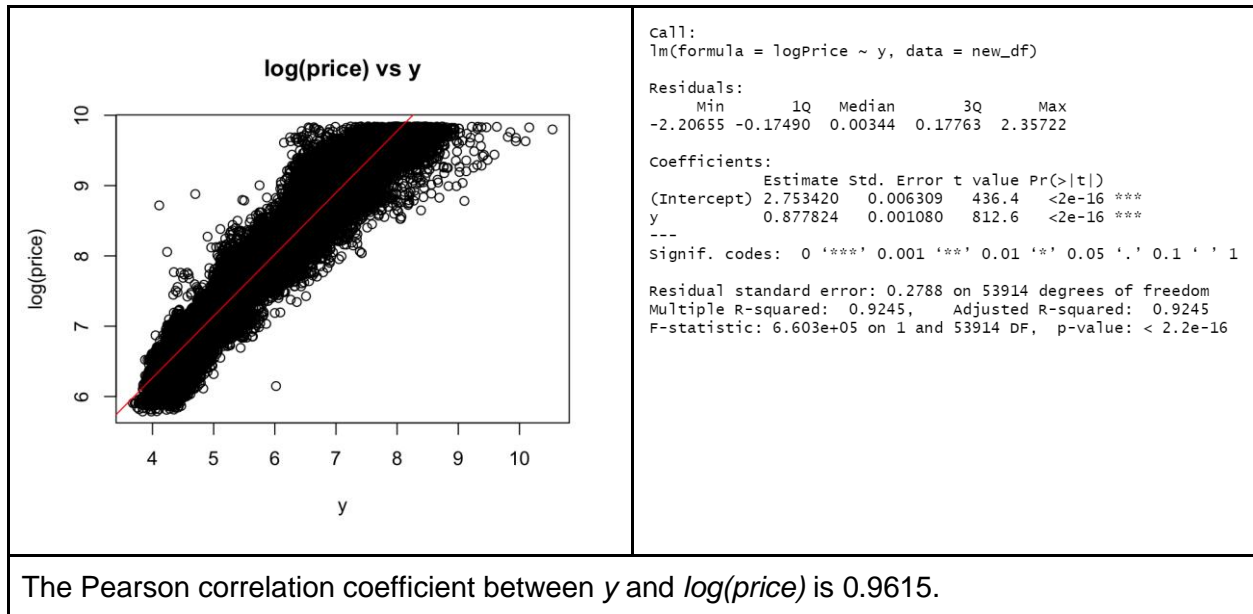
First, the Pearson correlation coefficients were calculated between $\log(\text{price})$ and x , y , z each. The values are 0.9607, 0.9615 and 0.9566 respectively. It appears that the y measurement is the most significant feature in predicting $\log(\text{price})$ among x , y , z measurements because it has the highest correlation coefficient. This is followed by x measurement as the next most significant feature in predicting $\log(\text{price})$, and finally followed by the z measurement.

Next, linear regression of $\log(\text{price})$ against x , y and z each was performed.

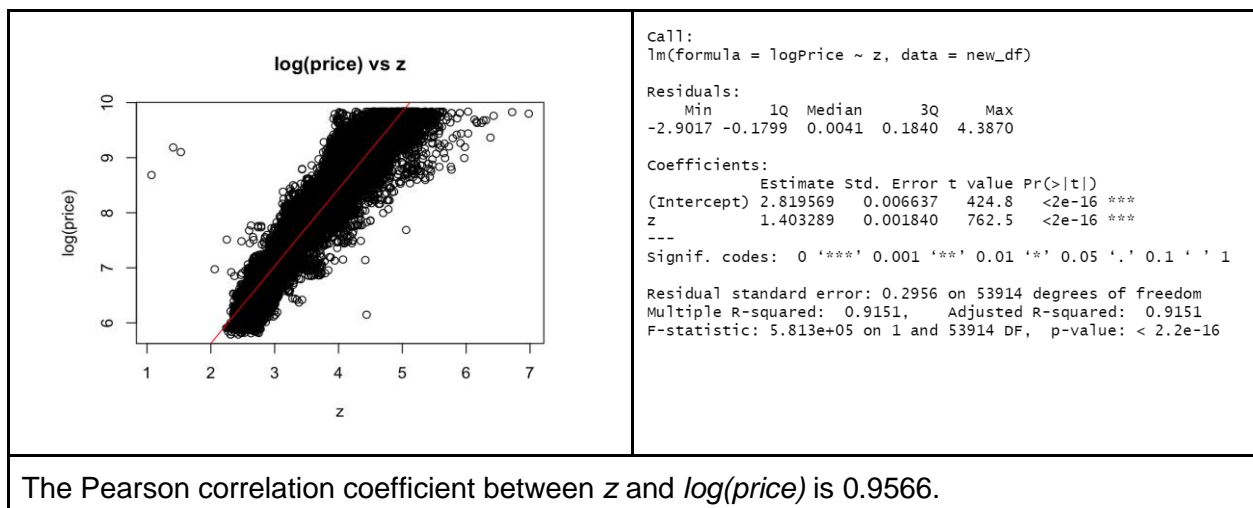
For x , the scatterplot of $\log(\text{price})$ against x shows a positive relationship between the 2 variables. The linear regression model for $\log(\text{price})$ against x provides a p-value of $<2 \times 10^{-16}$, which indicates that there is a statistically significant relationship between $\log(\text{price})$ and x at 0.05 level of significance, with an R-squared value of 0.923. From the linear regression results below, each additional mm for x is associated with an additional 0.8707 increase in $\log(\text{price})$. Hence, we conclude that x statistically affects the $\log(\text{price})$.



For y , the scatterplot of $\log(\text{price})$ against y shows a positive relationship between the 2 variables. The linear regression model for $\log(\text{price})$ against y provides a p-value of $<2 \times 10^{-16}$, which indicates that there is a statistically significant relationship between $\log(\text{price})$ and y at 0.05 level of significance with an R-squared value of 0.9245. From the linear regression results below, each additional mm for y is associated with an additional 0.8778 increase in $\log(\text{price})$. Hence, we conclude that y statistically affects the $\log(\text{price})$.



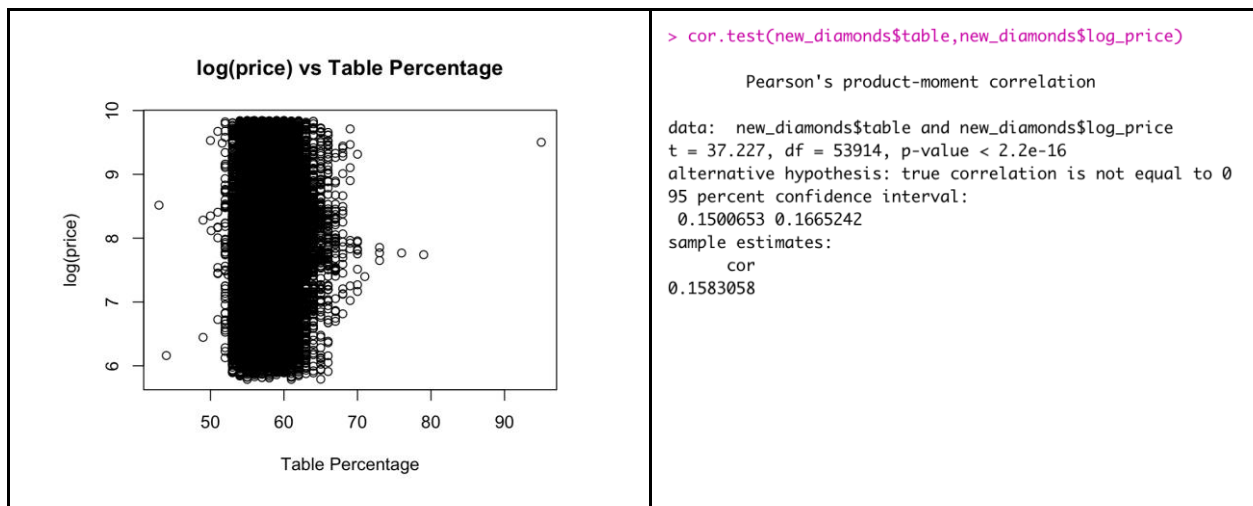
For z , the scatterplot of $\log(\text{price})$ versus z shows a relationship between the 2 variables. The linear regression model for $\log(\text{price})$ against z provides a p-value of $<2 \times 10^{-16}$, which indicates that there is a statistically significant relationship between $\log(\text{price})$ and z at 0.05 level of significance with an R-squared value of 0.9151. From the linear regression results below, each additional mm for z is associated with an additional 1.4033 increase in $\log(\text{price})$. Hence, we conclude that z statistically affects the $\log(\text{price})$.



4.4 Relation between $\log(\text{price})$ and table_percentage

In this section, we investigate the relationship between $\log(\text{price})$ and table_percentage and whether $\log(\text{price})$ is dependent on table_percentage .

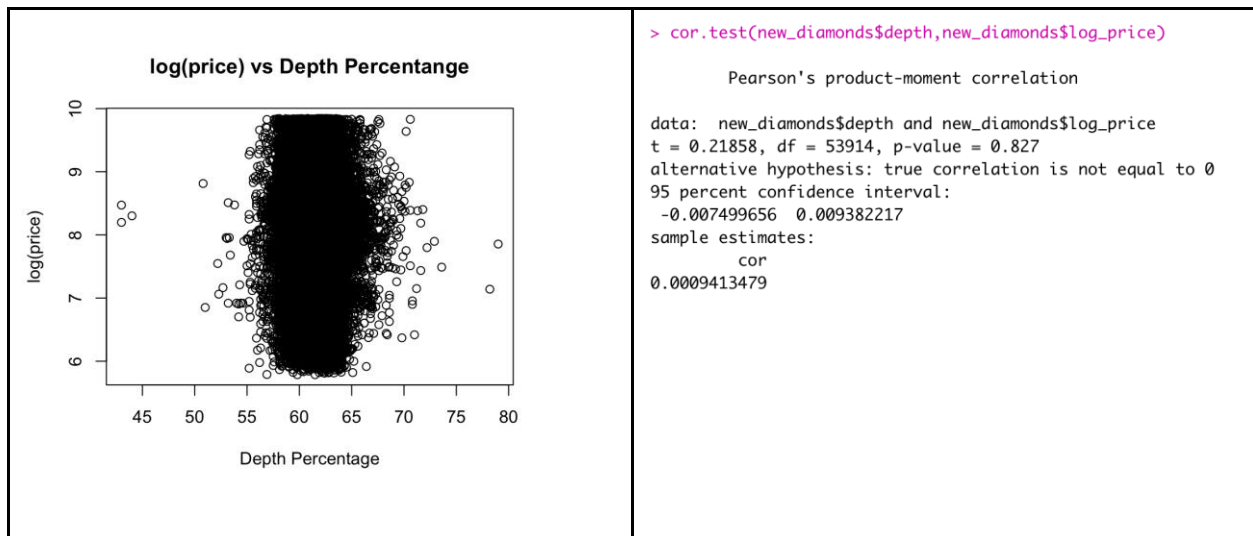
The Pearson correlation coefficient between $\log(\text{price})$ and table_percentage is 0.1583. This suggests that there is no strong linear relationship between $\log(\text{price})$ and table_percentage . Furthermore, the scatterplot of $\log(\text{price})$ versus depth_percentage shows no clear trend as table_percentage increases. Hence, we conclude that the $\log(\text{price})$ of diamonds is not dependent on table_percentage .



4.5 Relation between $\log(\text{price})$ and depth_percentage

In this section, we investigated the relationship between $\log(\text{price})$ and depth_percentage .

The Pearson correlation coefficient between $\log(\text{price})$ and depth_percentage is 0.0009413. This suggests that there is no strong linear correlation between $\log(\text{price})$ and depth_percentage . Furthermore, the scatterplot of $\log(\text{price})$ versus depth_percentage shows no clear trend as depth_percentage increases. Hence, we conclude that the $\log(\text{price})$ of diamonds is not dependent on depth_percentage .



4.6 Relation between $\log(\text{price})$ and standardised x , y , z and $\log(\text{carat})$ respectively

From sections 4.1 to 4.5, the features which are considered significant predictors of $\log(\text{price})$ are x , y , z , and $\log(\text{carat})$. To compare between these features, they were standardised, and multivariate linear regression was performed to identify the most significant predictor of $\log(\text{price})$.

From the multivariate linear regression model below, standardised $\log(\text{carat})$ is the most significant predictor of $\log(\text{price})$ as it has the highest coefficient estimate of 0.9198.

```
call:
lm(formula = logPrice ~ x_norm + y_norm + z_norm + log_carat_norm,
    data = new_df)

Residuals:
    Min       1Q   Median       3Q      Max
-1.47730 -0.16940 -0.00578  0.16583  1.32141

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.786368   0.001121  6947.83  <2e-16 ***
x_norm        -0.277449   0.022615  -12.27  <2e-16 ***
y_norm         0.514573   0.021975   23.42  <2e-16 ***
z_norm        -0.176496   0.009672  -18.25  <2e-16 ***
log_carat_norm 0.919778   0.010813   85.06  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2602 on 53911 degrees of freedom
Multiple R-squared:  0.9342,    Adjusted R-squared:  0.9342
F-statistic: 1.914e+05 on 4 and 53911 DF,  p-value: < 2.2e-16
```

5. Conclusion and Discussion

In conclusion, out of the 4 'C's (*carat*, *cut*, *color*, *clarity*) of the diamonds, *carat* was determined to be the only significant predictor for *price*. The dimensions (*x,y,z*) of diamonds were all found to be significant predictors for *price*. The proportions of the diamonds (*table_percentage* and *depth_percentage*) are not significant predictors of *price*. Out of all the significant predictors for the price of diamonds, *carat* is the most significant.

Although the grades of *color*, *cut* and *clarity* may play a significant role in deciding the price for diamonds of the same *carat*, the discovery that *color*, *cut* and *clarity* individually do not play a significant role in deciding the price of diamonds was surprising. This could be attributed to the fact that differences in grades of *cut*, *clarity* and *color* may not be noticeable to the average consumers and can only be judged by skilled artisans.

It should also be noted that, as discovered in our exploratory data analysis, there were some erroneous records in the dataset. Although some of the erroneous records were removed, there may still be erroneous records undetected by our analysis. Moreover, there could be other attributes of a diamond which may be significant in determining the price of a diamond that are not in the dataset. For example, shape, as defined to be the outline shape of the diamond as looked from the top down [4], could be a significant factor in determining the price of a diamond. However, there was no data on the shape of the diamonds so this could not be verified.

Regardless, we hope that our report proves useful to consumers interested in finding out which aspects of a diamond affect its price.

6.Appendix

```
library(ggplot2)

# outliers from observing the boxplot visually
x_outliers <- which(x<2)
y_outliers <- which(y>20)
z_outliers <- which(z>15)
x_zero_outliers<-which(x<0.1)
y_zero_outliers<- which(y<0.1)
z_zero_outliers <-which(z<0.1)
print("x outliers:")
[1] "x outliers:"
```

```
x_outliers
integer(0)
```

```
print("y outliers:")
[1] "y outliers:"
```

```
y_outliers
integer(0)
```

```
print("z outliers:")
[1] "z outliers:"
```

```
z_outliers
integer(0)
```

```
print("x_zero_outliers:")
```

```
[1] "x_zero_outliers:"
```

```
x_zero_outliers  
integer(0)
```

```
print("y_zero_outliers:")  
[1] "y_zero_outliers:"
```

```
y_zero_outliers  
integer(0)
```

```
print("z_zero_outliers:")  
[1] "z_zero_outliers:"
```

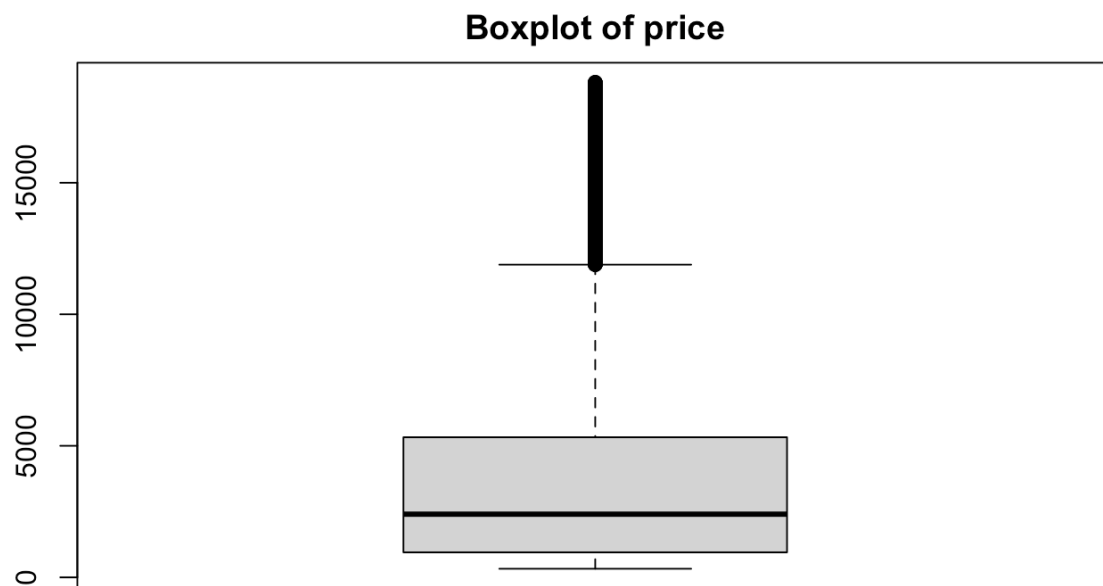
```
z_zero_outliers  
integer(0)
```

```
#removing outliers  
# outliers from observing the boxplot visually  
x_outliers <- which(diamonds$x < 2)  
y_outliers <- which(diamonds$y > 20)  
z_outliers <- which(diamonds$z > 15)  
x_zero_outliers <- which(diamonds$x < 0.1)  
y_zero_outliers <- which(diamonds$y < 0.1)  
z_zero_outliers <- which(diamonds$z < 0.1)  
del_indices <- c(y_outliers, z_outliers, z_zero_outliers) # note: z_zero_outliers includes all x zero outliers and y zero outliers  
  
del_indices <- c(del_indices, 52861)  
#Dataframe cleaning & adding new log(price) and log(carat) as a column  
new_diamonds <- diamonds[-del_indices, ]  
new_diamonds$log_price <- log(new_diamonds$price)
```

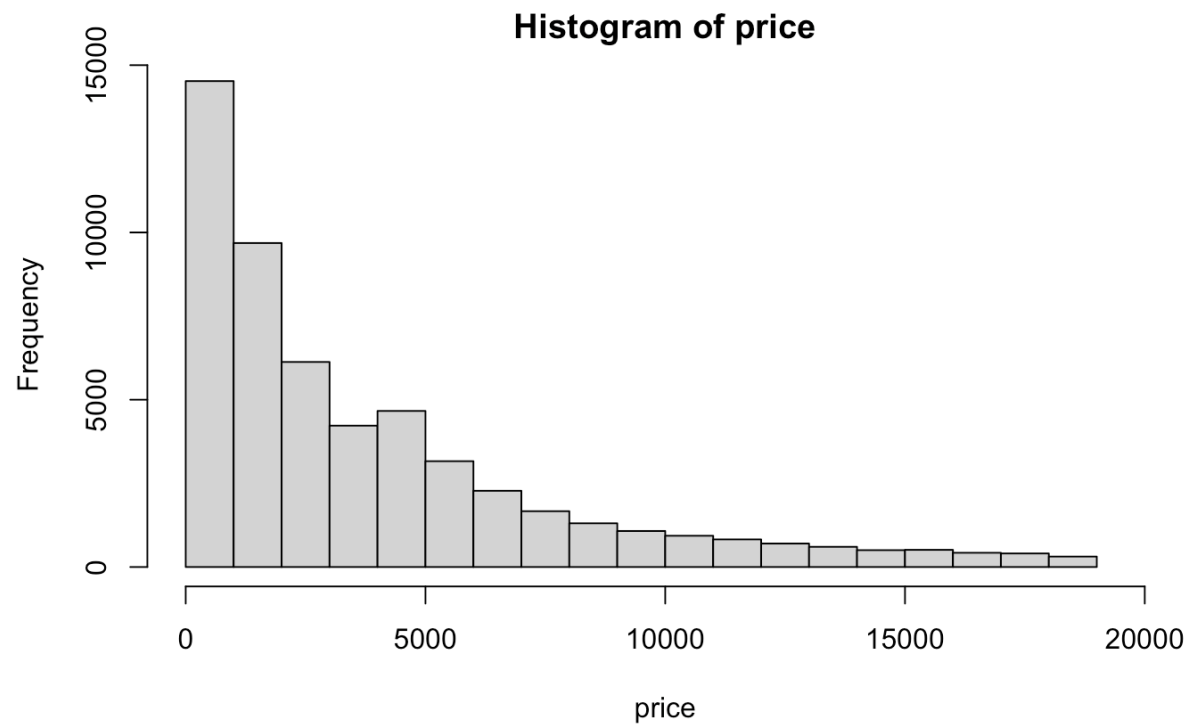
```
new_diamonds$log_carat <- log(new_diamonds$carat)

#Summary Statistics of price,carat, (cut,color,clarity), (x,y,z), (table,depth)

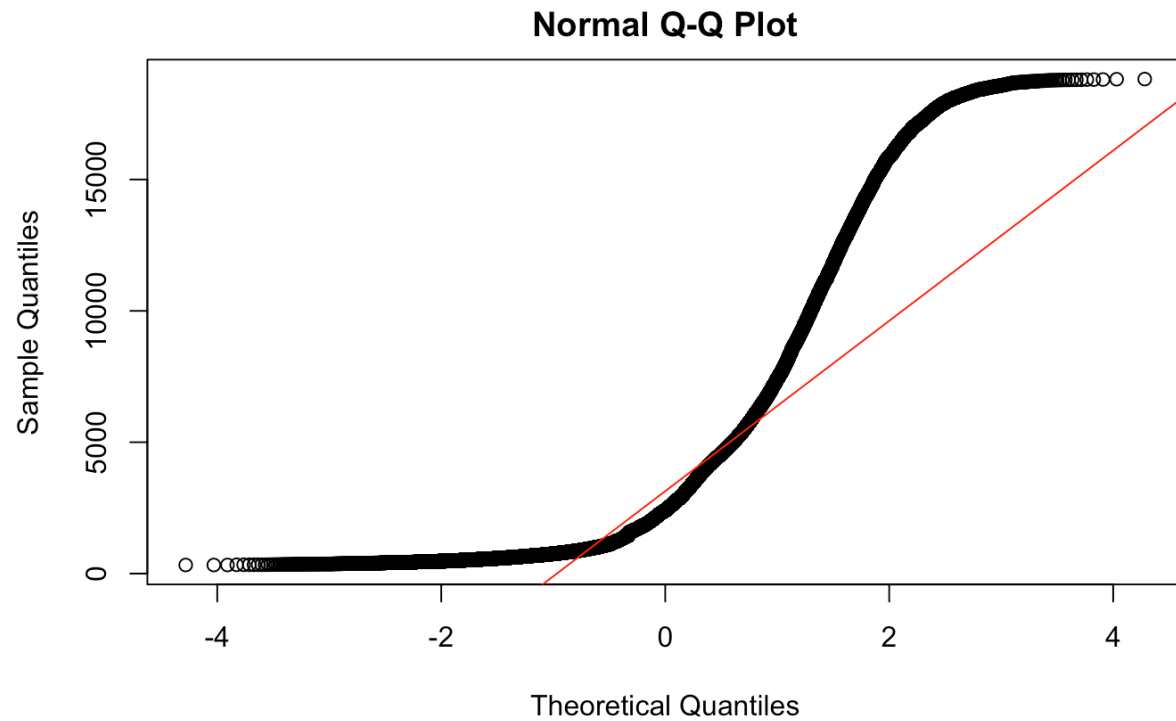
#price
price = diamonds$price
boxplot(price, main = "Boxplot of price")
```



```
hist(price, xlim = c(0, 20000))
```



```
qqnorm(price)
qqline(price, col = 'red')
```

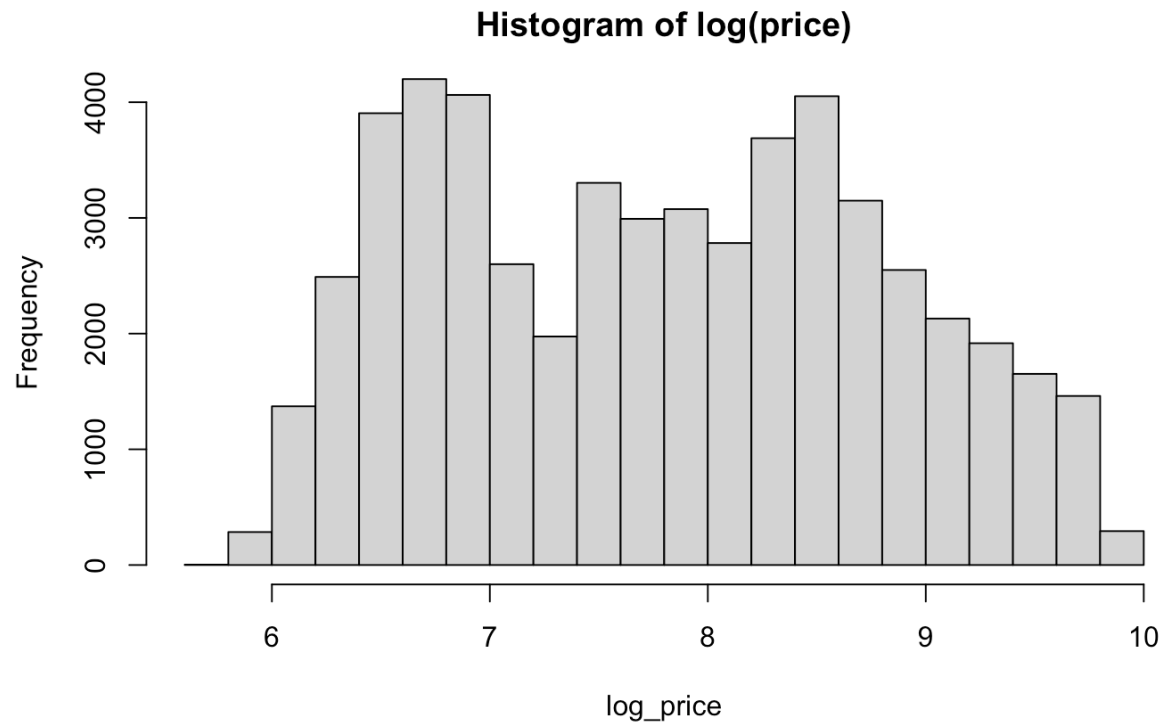


```
summary(price)
```

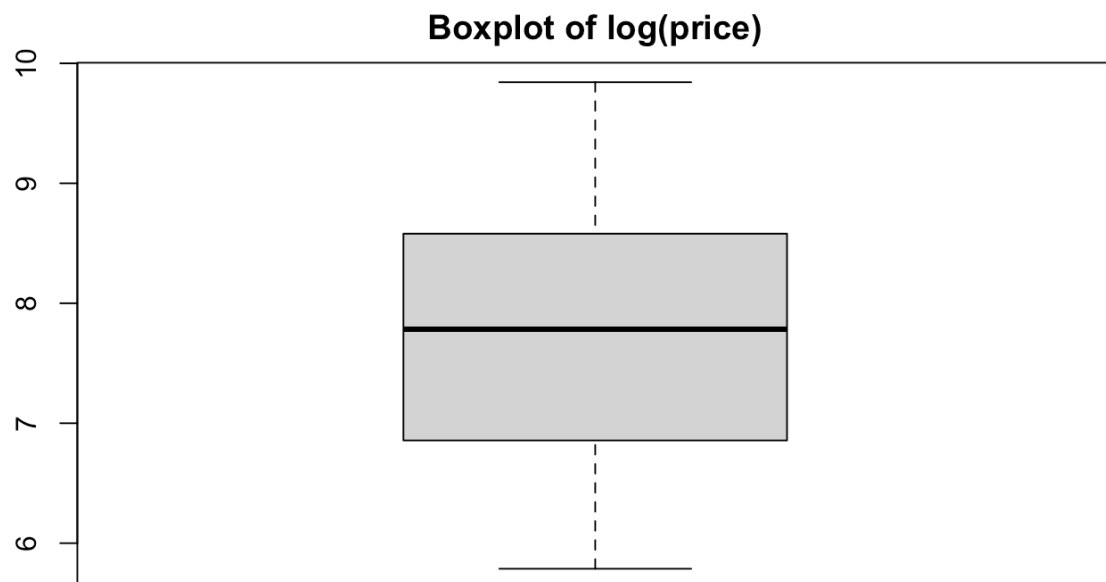
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
326	950	2401	3933	5324	18823

```
log_price = log(price)
```

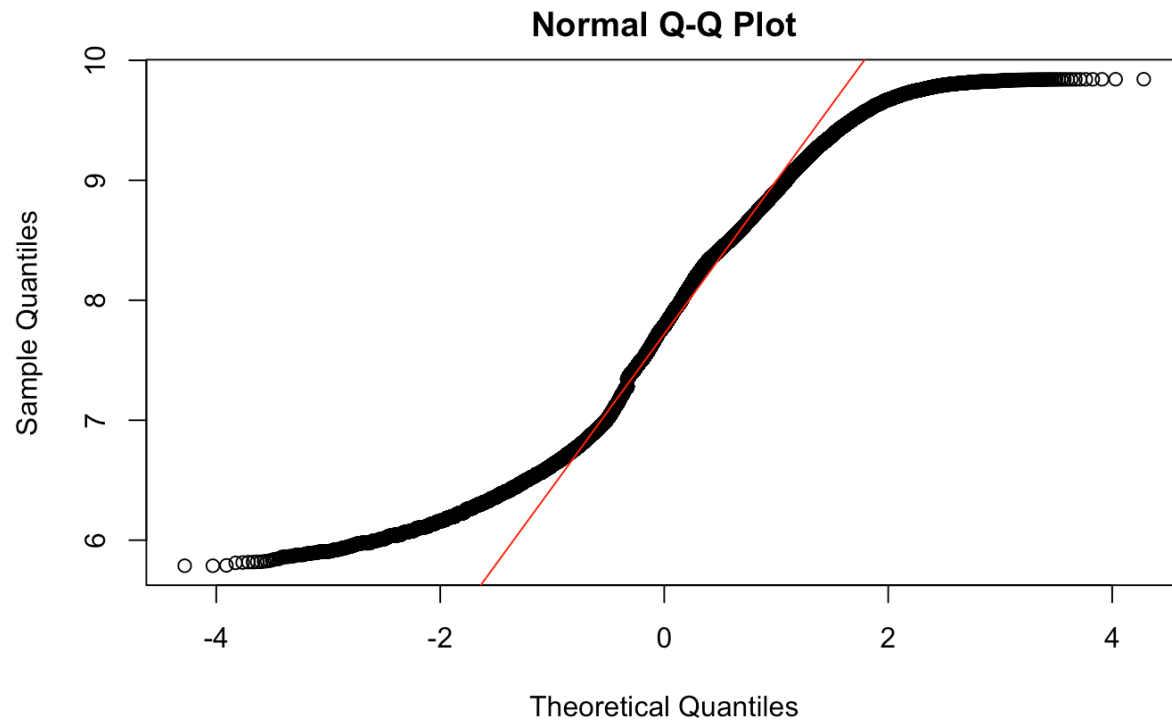
```
hist(log_price, main = "Histogram of log(price)")
```

```
boxplot(log_price, main = "Boxplot of log(price)")
```



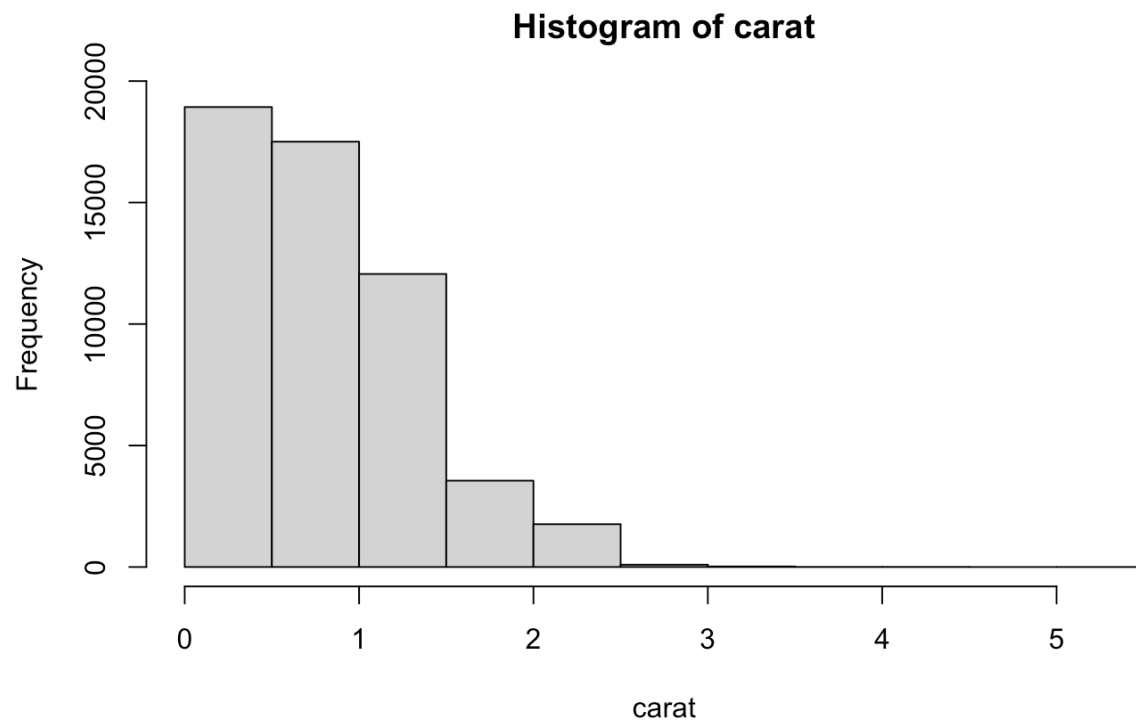
```
qqnorm(log_price)
qqline(log_price, col = 'red')
```



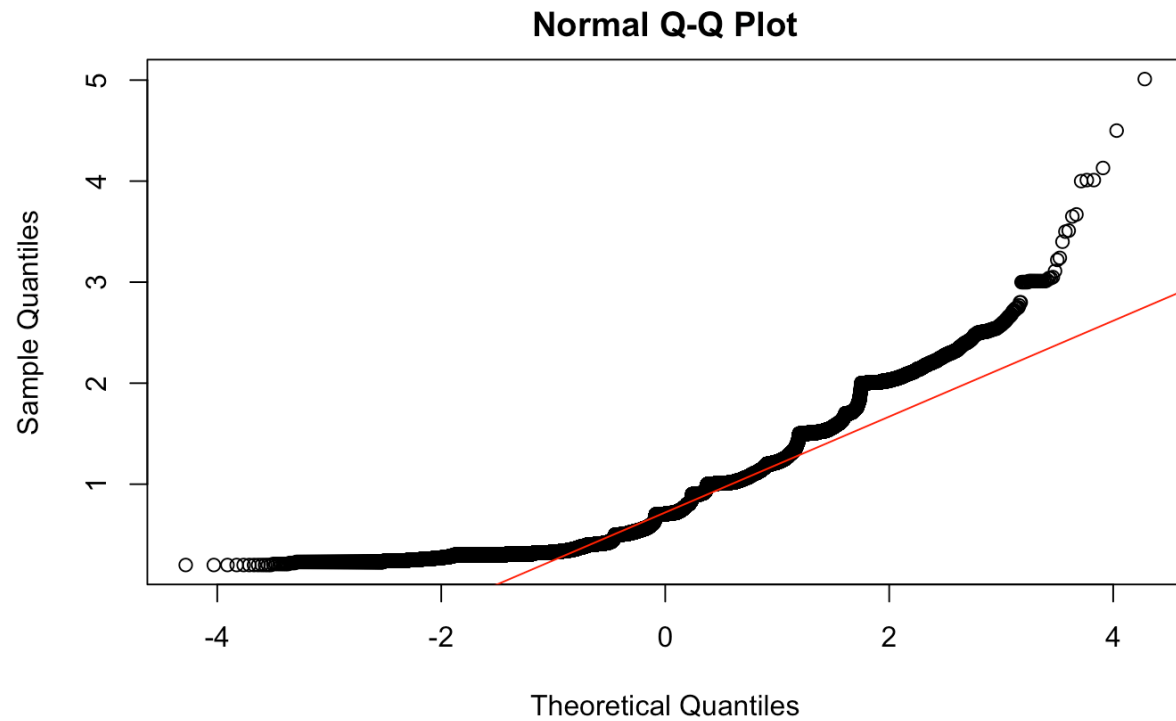
```
summary(log_price)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.787	6.856	7.784	7.787	8.580	9.843

```
#carat
carat = diamonds$carat
hist(carat, ylim = c(0, 20000))
```



```
qqnorm(carat)
qqline(carat, col = 'red')
```

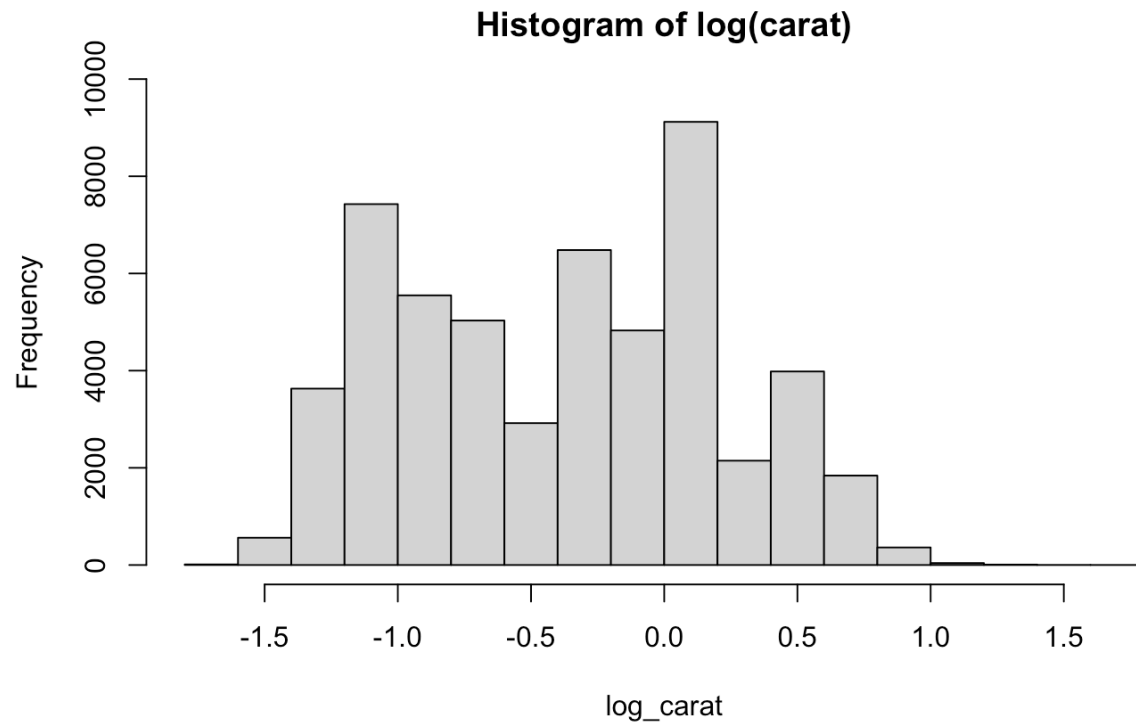


```
summary(carat)
```

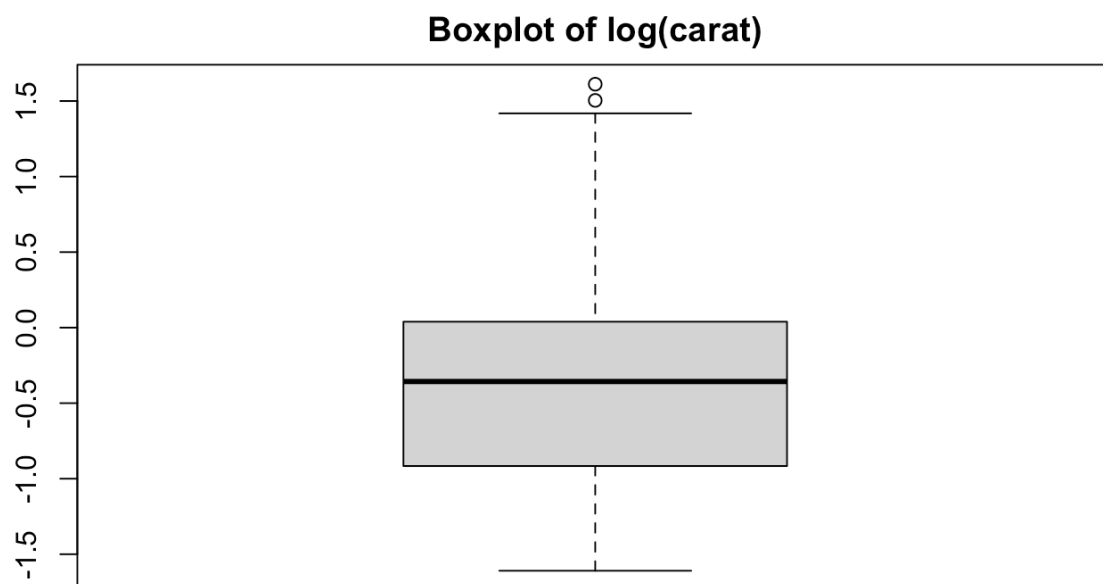
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2000	0.4000	0.7000	0.7979	1.0400	5.0100

```
log_carat = log(carat)
```

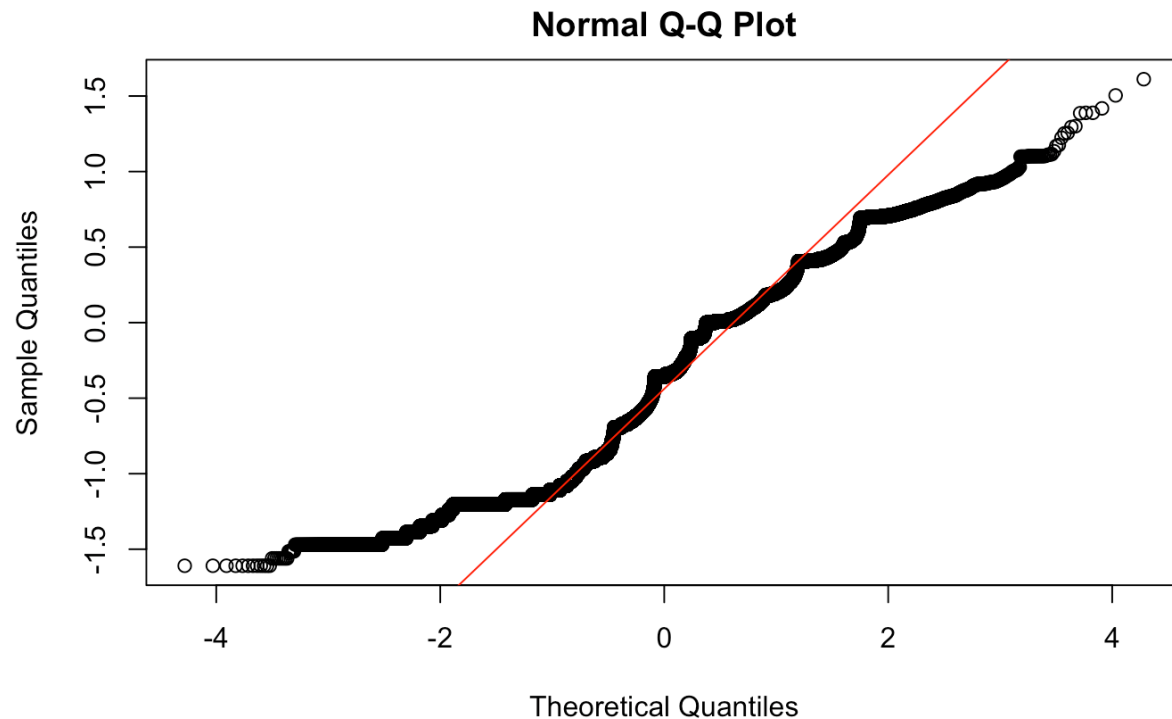
```
hist(log_carat, ylim = c(0, 10000), main = "Histogram of log(carat)")
```



```
boxplot(log_carat, main = "Boxplot of log(carat)")
```



```
qqnorm(log_carat)
qqline(log_carat, col = 'red')
```



```
summary(log_carat)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.60944	-0.91629	-0.35667	-0.39497	0.03922	1.61144

```
#Statistical analysis of log(carat) VS log(price)
print(paste(
  "The correlation coefficient between log(carat) and log(price):",
  cor(new_diamonds$log_carat, new_diamonds$log_price)
))
```

```
[1] "The correlation coefficient between log(carat) and log(price): 0.9659244
92422405"
```

```
plot(
```

```

new_diamonds$log_carat ~ new_diamonds$log_price,
xlab = "log(carat)",
ylab = "log(price)",
main = "log(price) vs log(carat)"
)
#linear regression between log_carat and log_price
fit_log_carat = lm(new_diamonds$log_carat ~ new_diamonds$log_price, data = new_diamonds)
summary(fit_log_carat)

```

Call:

```

lm(formula = new_diamonds$log_carat ~ new_diamonds$log_price,
    data = new_diamonds)

```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.79905	-0.10072	0.00398	0.09991	0.93989

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-4.7299692	0.0050446	-937.6	<2e-16 ***
new_diamonds\$log_price	0.5567093	0.0006424	866.5	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1513 on 53914 degrees of freedom

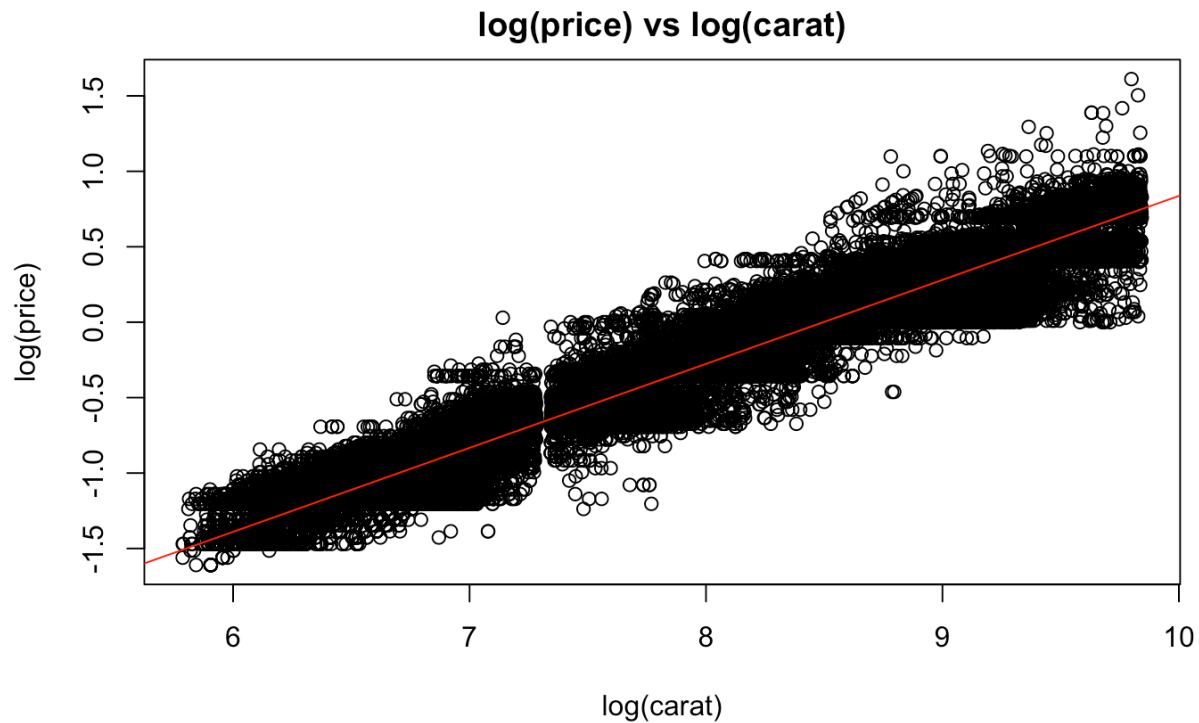
Multiple R-squared: 0.933, Adjusted R-squared: 0.933

F-statistic: 7.509e+05 on 1 and 53914 DF, p-value: < 2.2e-16

```

abline(fit_log_carat, col = "red")

```



```
cor.test(new_diamonds$log_price,new_diamonds$log_carat)
```

Pearson's product-moment correlation

data: new_diamonds\$log_price and new_diamonds\$log_carat

t = 866.54, df = 53914, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9653544 0.9664854

sample estimates:

cor

0.9659245

```
#Cut, Color, Clarity Respectively
```

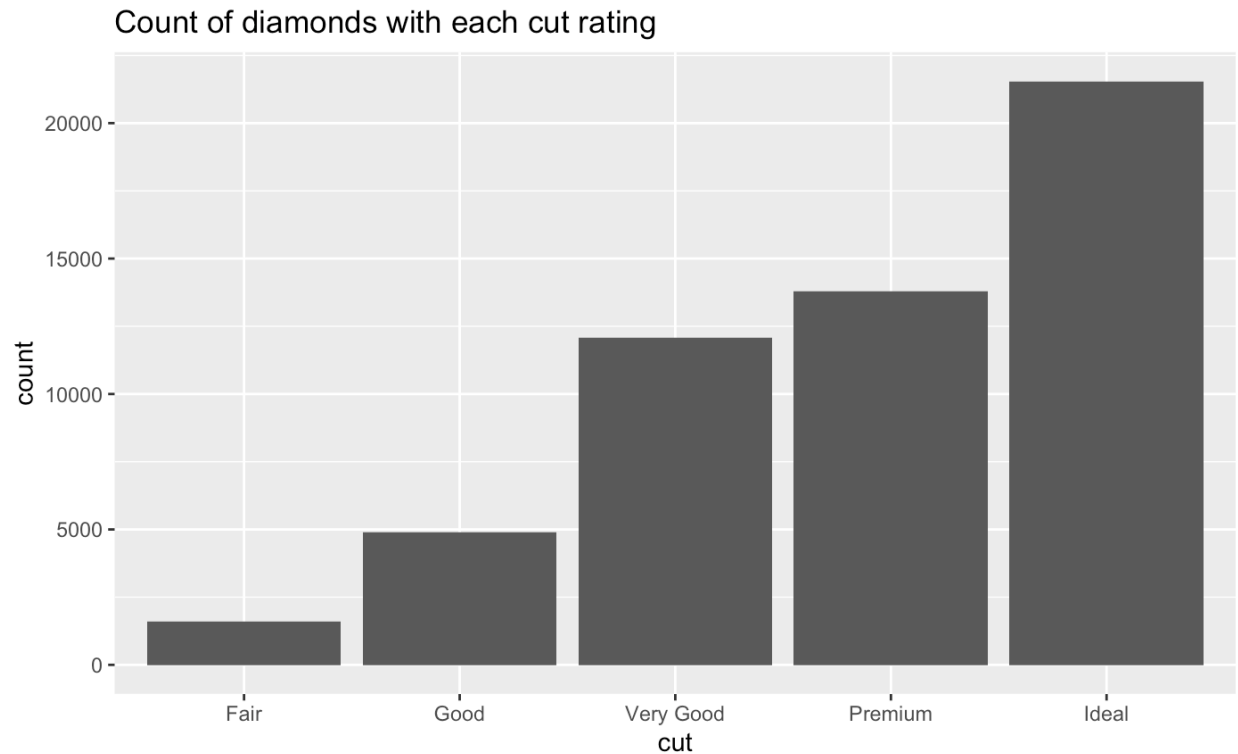
```
diamonds$cut = factor(diamonds$cut,
```



```

        levels = c('Fair', 'Good', 'Very Good', 'Premium', 'Ideal'))
ggplot(data = diamonds, aes(x = cut)) + ggtitle('Count of diamonds with each
cut rating') +
  geom_bar()

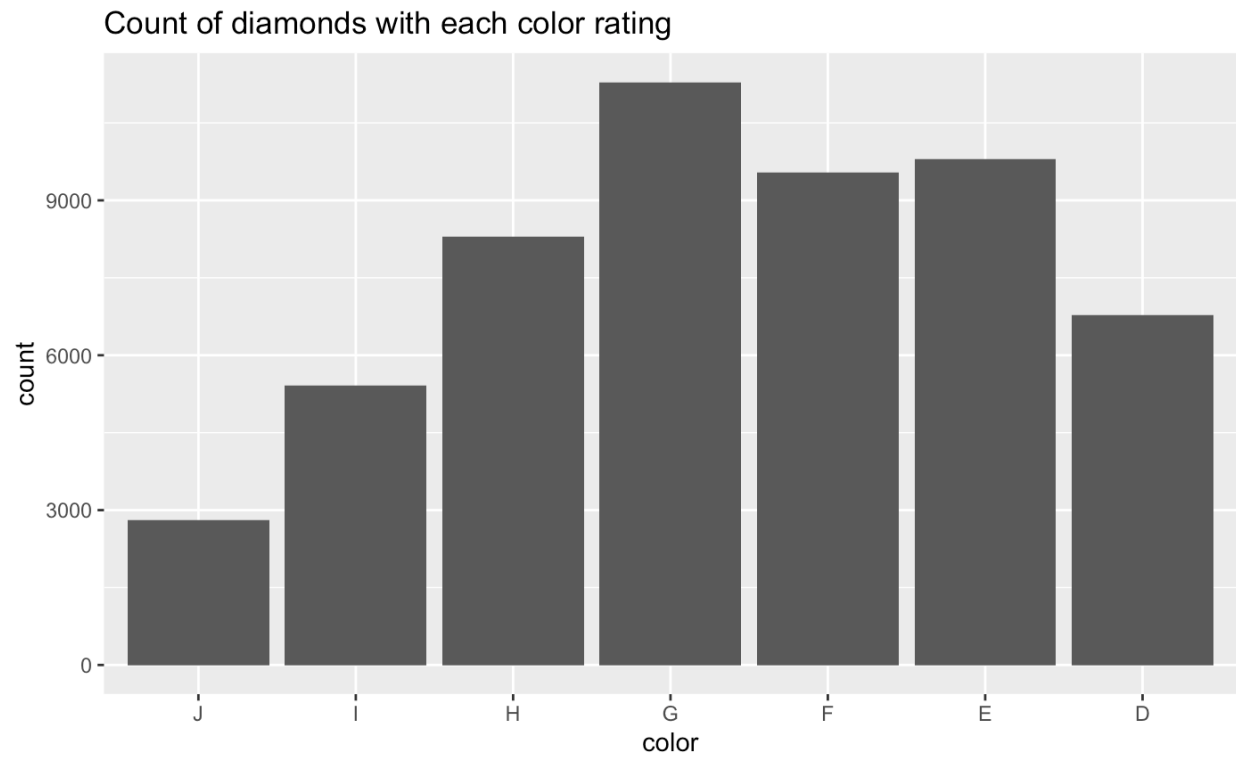
```



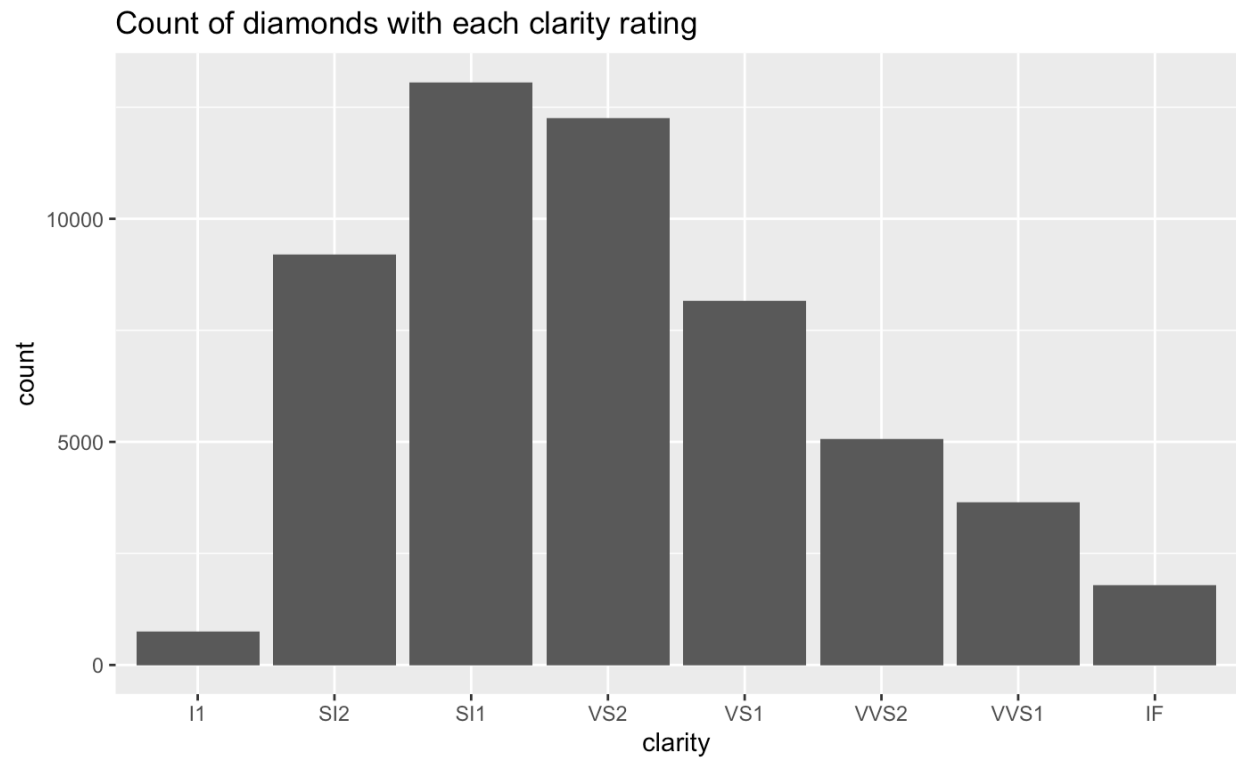
```

diamonds$color = factor(diamonds$color, levels = c('J', 'I', 'H', 'G', 'F', 'E', 'D'))
ggplot(data = diamonds, aes(x = color)) + ggtitle('Count of diamonds with each
color rating') +
  geom_bar()

```

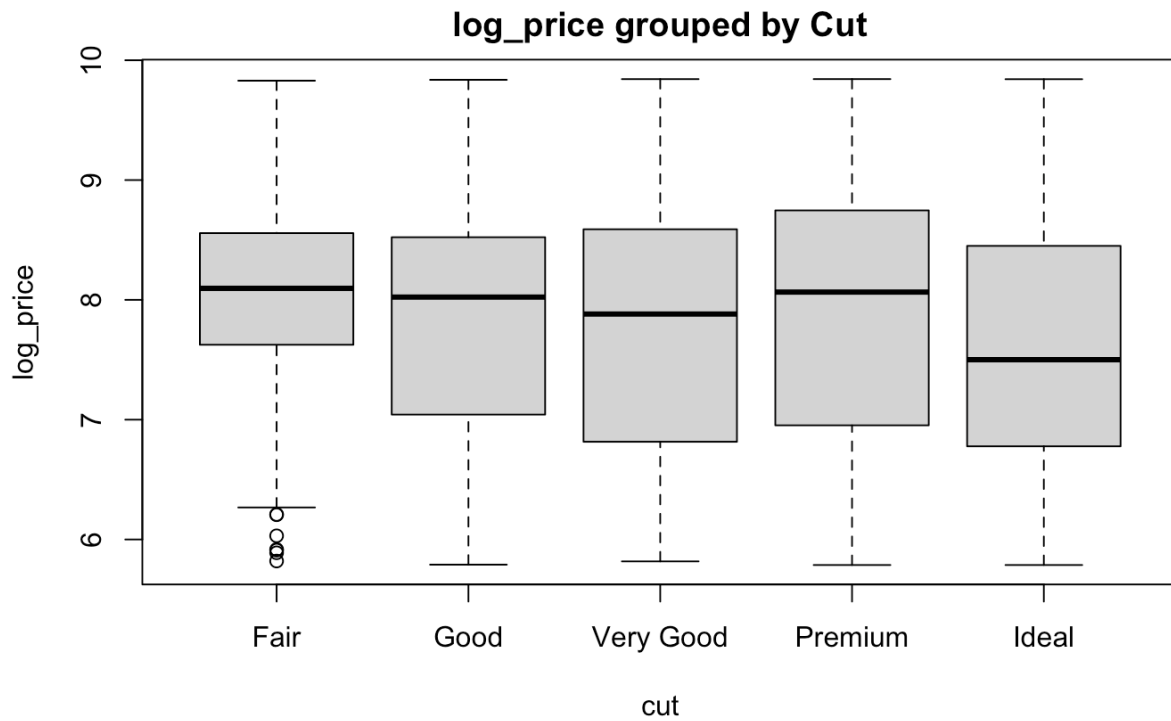


```
diamonds$clarity = factor(diamonds$clarity,  
                           levels = c('I1', 'SI2', 'SI1', 'VS2', 'VS1', 'VVS2',  
                                       'VVS1', 'IF'))  
ggplot(data = diamonds, aes(x = clarity)) + ggtitle('Count of diamonds with e  
ach clarity rating') +  
  geom_bar()
```

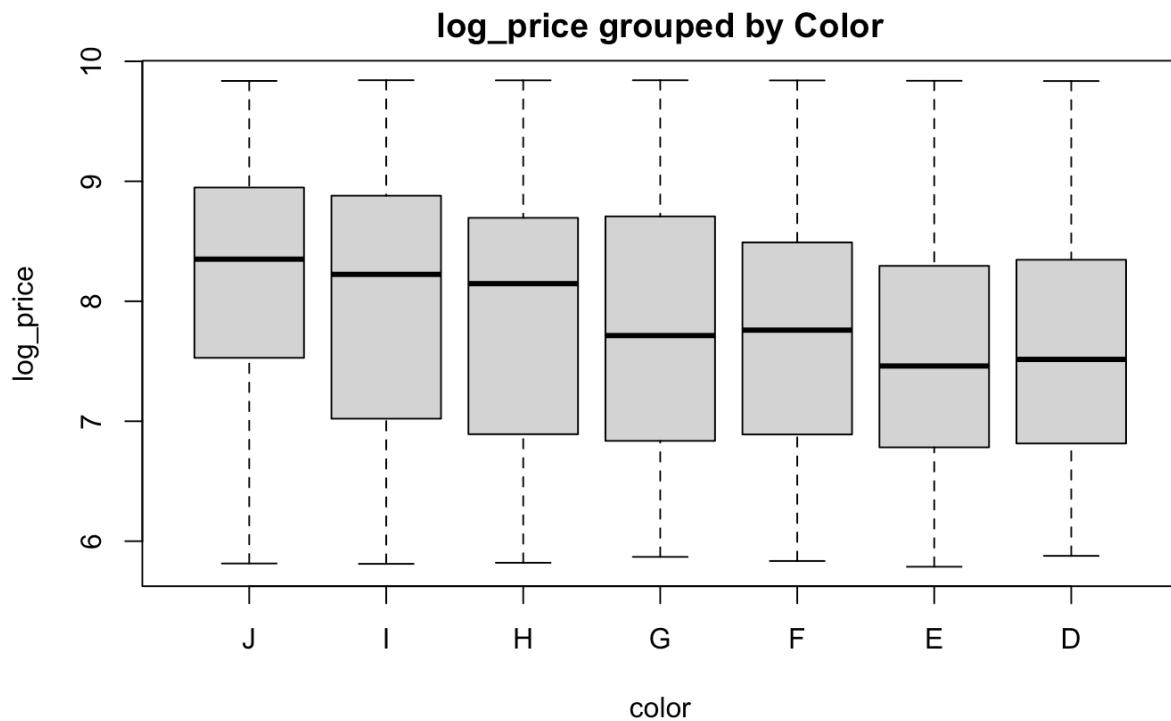


```
##Statistical Analysis of Cut, Color, Clarity

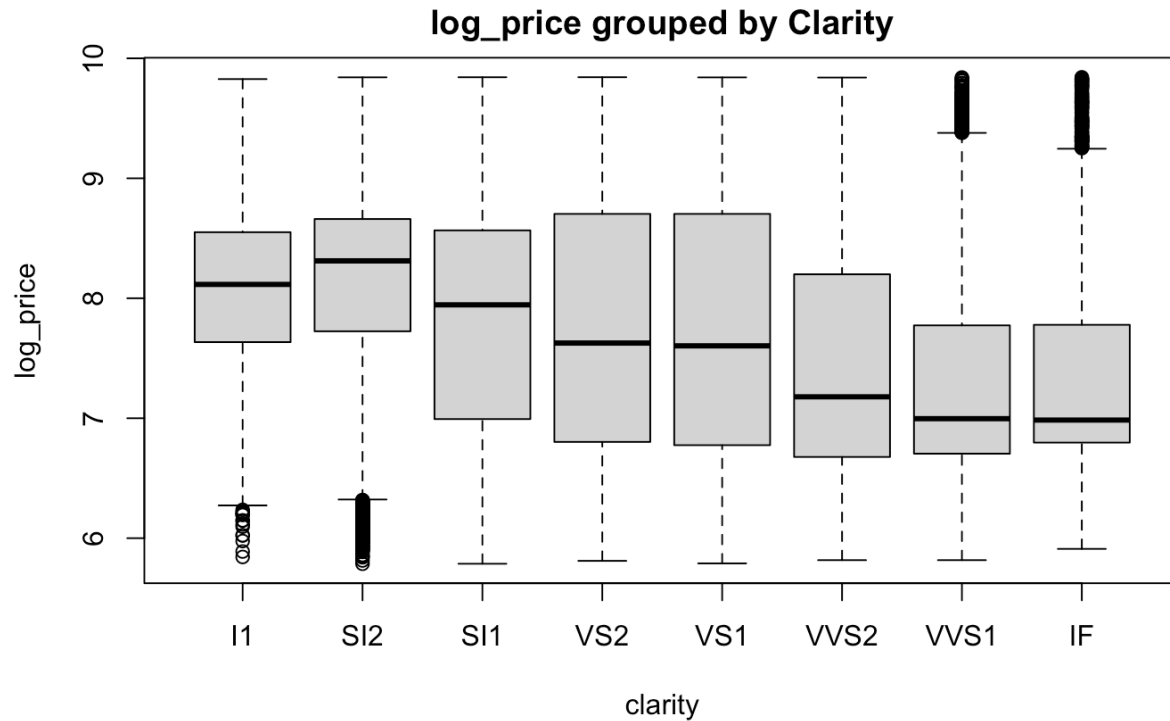
#boxplot
boxplot(log_price ~ cut, data = new_diamonds, main = "log_price grouped by Cut")
```



```
boxplot(log_price ~ color, data = new_diamonds, main = "log_price grouped by Color")
```



```
boxplot(log_price ~ clarity, data = new_diamonds, main = "log_price grouped b
y Clarity")
```



```
#pairwise variance test
library("RVAideMemoire")
pairwise.var.test(new_diamonds$log_price, new_diamonds$cut)
```

Pairwise comparisons using F tests to compare two variances

data: new_diamonds\$log_price and new_diamonds\$cut

	Fair	Good	Very Good	Premium
Good	< 2e-16	-	-	-
Very Good	< 2e-16	6.3e-06	-	-
Premium	< 2e-16	6.3e-05	0.40	-
Ideal	< 2e-16	0.37	5.5e-08	2.3e-06

```
P value adjustment method: fdr
```

```
pairwise.var.test(new_diamonds$log_price, new_diamonds$color)
```

Pairwise comparisons using F tests to compare two variances

data: new_diamonds\$log_price and new_diamonds\$color

	J	I	H	G	F	E
I	0.00019	-	-	-	-	-
H	0.14565	0.00151	-	-	-	-
G	0.51585	5.1e-10	0.00123	-	-	-
F	4.3e-06	< 2e-16	< 2e-16	1.8e-09	-	-
E	2.2e-14	< 2e-16	< 2e-16	< 2e-16	1.8e-05	-
D	5.6e-13	< 2e-16	< 2e-16	< 2e-16	0.00014	0.95753

```
P value adjustment method: fdr
```

```
pairwise.var.test(new_diamonds$log_price, new_diamonds$clarity)
```

Pairwise comparisons using F tests to compare two variances

data: new_diamonds\$log_price and new_diamonds\$clarity

	I1	SI2	SI1	VS2	VS1	VVS2	VVS1
SI2	4.8e-13	-	-	-	-	-	-
SI1	< 2e-16	1.8e-14	-	-	-	-	-
VS2	< 2e-16	< 2e-16	1.3e-06	-	-	-	-
VS1	< 2e-16	< 2e-16	2.6e-08	0.2356	-	-	-
VVS2	< 2e-16	< 2e-16	0.0031	0.4979	0.1181	-	-
VVS1	1.0e-09	0.0665	9.6e-14	< 2e-16	< 2e-16	< 2e-16	-
IF	5.0e-11	0.8989	7.8e-05	4.3e-10	2.1e-11	6.7e-08	0.1723

```
P value adjustment method: fdr
```

```
# kruskal wallis test
kruskal.test(new_diamonds$log_price, new_diamonds$cut)

Kruskal-Wallis rank sum test

data:  new_diamonds$log_price and new_diamonds$cut
Kruskal-Wallis chi-squared = 975.03, df = 4, p-value < 2.2e-16
```

```
kruskal.test(new_diamonds$log_price, new_diamonds$color)

Kruskal-Wallis rank sum test

data:  new_diamonds$log_price and new_diamonds$color
Kruskal-Wallis chi-squared = 1333, df = 6, p-value < 2.2e-16
```

```
kruskal.test(new_diamonds$log_price, new_diamonds$clarity)

Kruskal-Wallis rank sum test

data:  new_diamonds$log_price and new_diamonds$clarity
Kruskal-Wallis chi-squared = 2714.8, df = 7, p-value < 2.2e-16
```

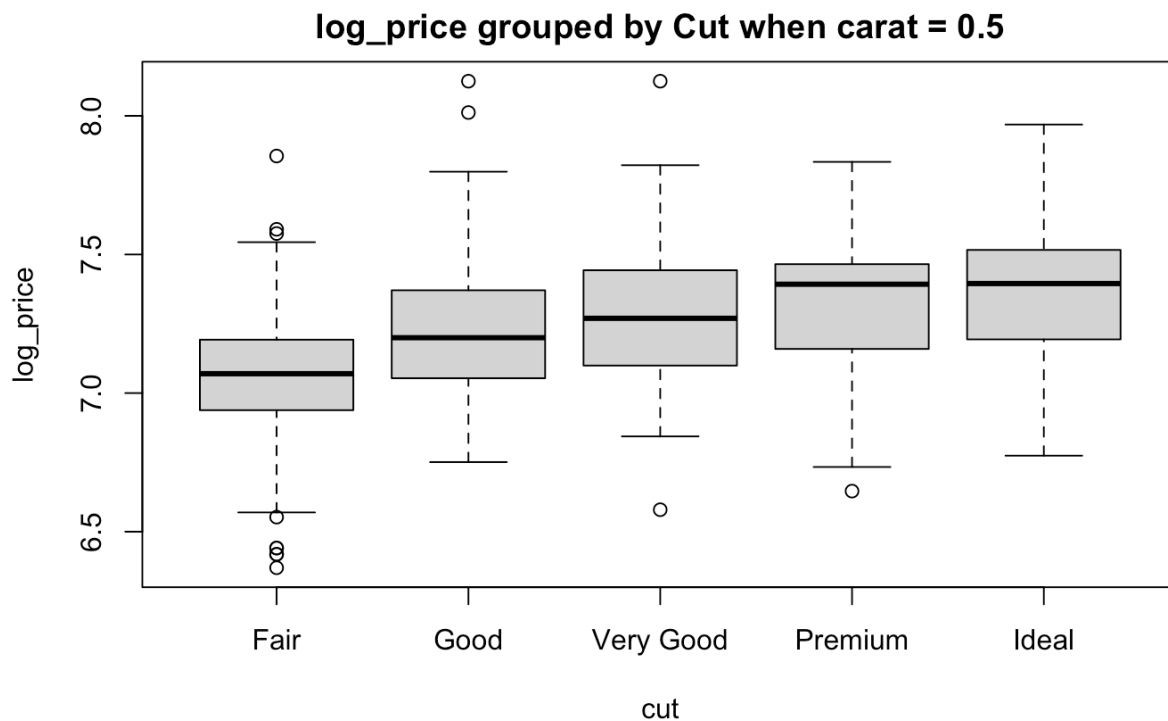
```
# spearman
new_diamonds$cut_num <- unclass(new_diamonds$cut)
new_diamonds$color_num <- unclass(new_diamonds$color)
new_diamonds$clarity_num <- unclass(new_diamonds$clarity)

cor(new_diamonds$log_price, new_diamonds$cut_num, method = 'spearman')
[1] -0.09293967
```

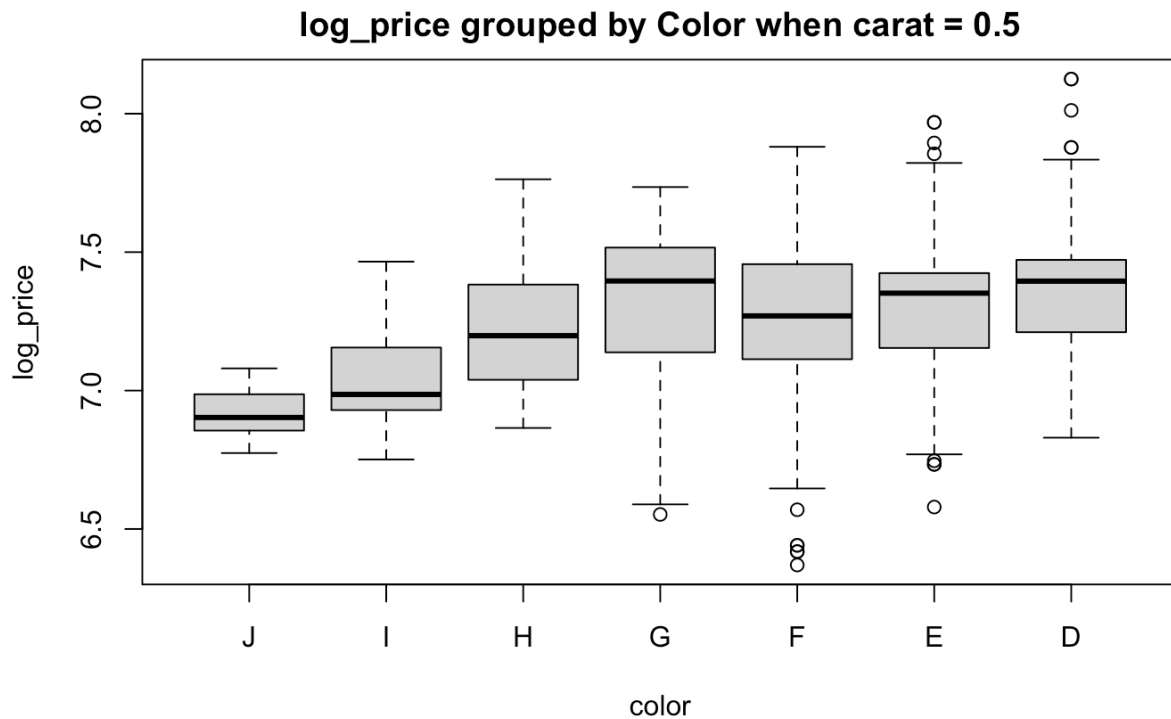
```
cor(new_diamonds$log_price, new_diamonds$color_num, method = 'spearman')  
[1] -0.1499864
```

```
cor(new_diamonds$log_price, new_diamonds$clarity_num, method = 'spearman')  
[1] -0.2114244
```

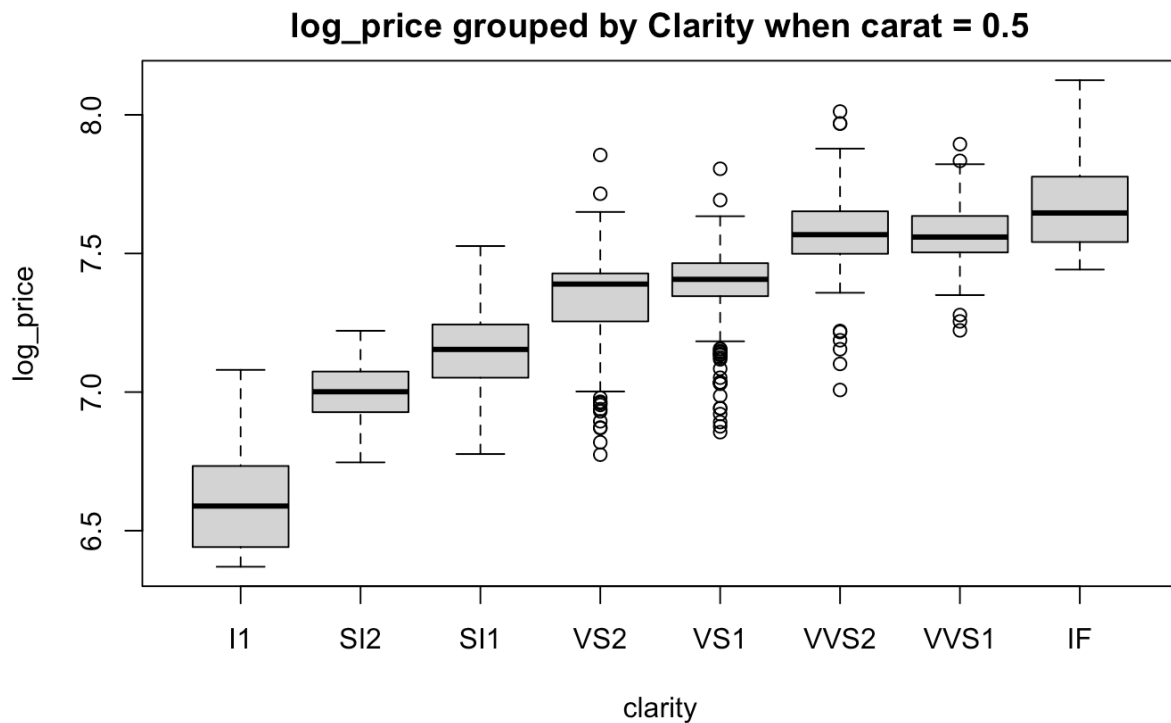
```
# for SUBSET where carat = 0.5  
equal_carat = subset(new_diamonds, carat == 0.5)  
  
#plotting boxplot  
equal_carat$color = factor(equal_carat$color, levels = c('J', 'I', 'H', 'G',  
'F', 'E', 'D'))  
  
boxplot(log_price ~ cut, data = equal_carat, main = "log_price grouped by Cut  
when carat = 0.5")
```



```
boxplot(log_price ~ color, data = equal_carat, main = "log_price grouped by C  
olor when carat = 0.5")
```

```
boxplot(log_price ~ clarity, data = equal_carat, main = "log_price grouped by  
Clarity when carat = 0.5")
```



```
#pairwise variance
library("RVAideMemoire")
pairwise.var.test(equal_carat$log_price, equal_carat$cut)
```

Pairwise comparisons using F tests to compare two variances

data: equal_carat\$log_price and equal_carat\$cut

	Fair	Good	Very Good	Premium
Good	0.0348	-	-	-
Very Good	0.0187	0.8599	-	-
Premium	0.0065	0.5708	0.5708	-
Ideal	0.0348	0.7368	0.5708	0.2358

P value adjustment method: fdr

```
pairwise.var.test(equal_carat$log_price, equal_carat$color)
```

Pairwise comparisons using F tests to compare two variances

data: equal_carat\$log_price and equal_carat\$color

	J	I	H	G	F	E
I	0.00245	-	-	-	-	-
H	4.7e-05	0.17727	-	-	-	-
G	5.6e-06	0.01809	0.16192	-	-	-
F	2.7e-07	0.00054	0.00111	0.01763	-	-
E	1.5e-05	0.07428	0.58525	0.17727	4.6e-05	-
D	1.4e-05	0.05756	0.44530	0.39590	0.00105	0.68939

P value adjustment method: fdr

```
pairwise.var.test(equal_carat$log_price, equal_carat$clarity)
```

Pairwise comparisons using F tests to compare two variances

data: equal_carat\$log_price and equal_carat\$clarity

	I1	SI2	SI1	VS2	VS1	VVS2	VVS1
SI2	0.00017	-	-	-	-	-	-
SI1	0.06760	0.00032	-	-	-	-	-
VS2	0.08800	8.1e-05	0.52931	-	-	-	-
VS1	0.35305	4.0e-07	0.03809	0.08000	-	-	-
VVS2	0.13453	0.00014	0.42180	0.74753	0.27419	-	-
VVS1	0.06760	0.08000	0.58996	0.40440	0.08000	0.35305	-
IF	0.99563	0.00013	0.06290	0.08000	0.35305	0.12065	0.06290

P value adjustment method: fdr

```
# kruskal-wallis test
```

```
kruskal.test(equal_carat$log_price, equal_carat$cut)
```

Kruskal-Wallis rank sum test

data: equal_carat\$log_price and equal_carat\$cut

Kruskal-Wallis chi-squared = 114.59, df = 4, p-value < 2.2e-16

```
kruskal.test(equal_carat$log_price, equal_carat$color)
```

Kruskal-Wallis rank sum test

data: equal_carat\$log_price and equal_carat\$color

Kruskal-Wallis chi-squared = 153.58, df = 6, p-value < 2.2e-16

```
kruskal.test(equal_carat$log_price, equal_carat$clarity)
```

Kruskal-Wallis rank sum test

data: equal_carat\$log_price and equal_carat\$clarity

Kruskal-Wallis chi-squared = 810.64, df = 7, p-value < 2.2e-16

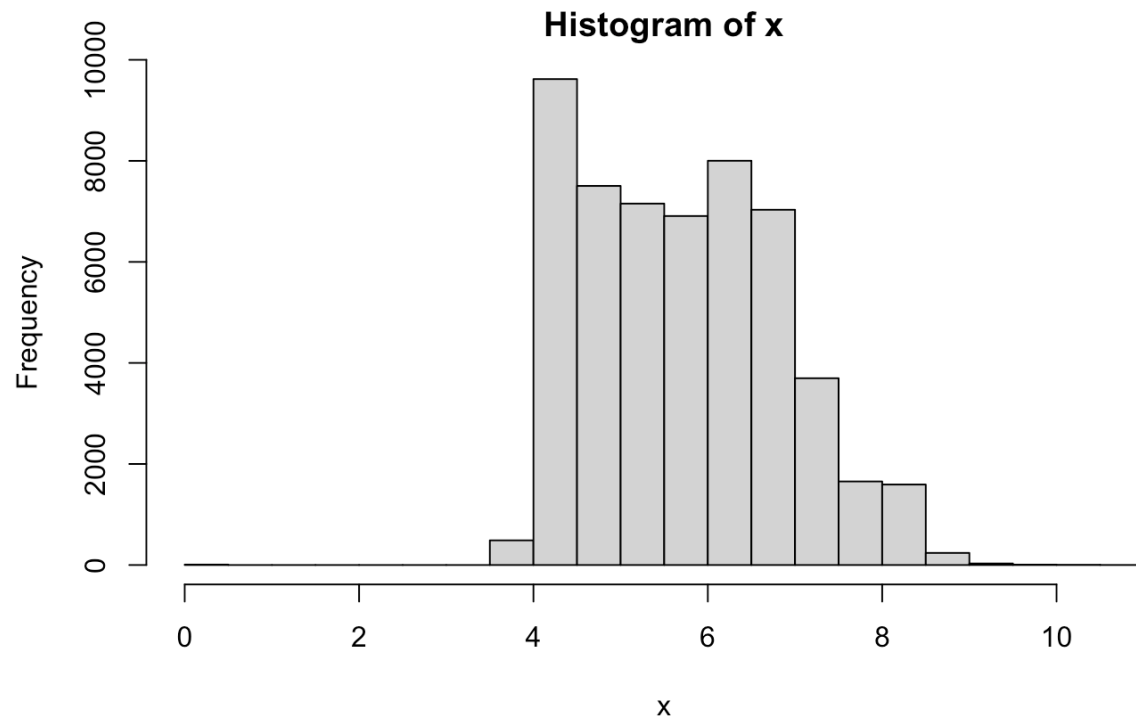
```
#spearman
equal_carat$cut_num <- unclass(equal_carat$cut)
equal_carat$color_num <- unclass(equal_carat$color)
equal_carat$clarity_num <- unclass(equal_carat$clarity)

cor(equal_carat$log_price, equal_carat$cut_num, method = 'spearman')
[1] 0.2788528
```

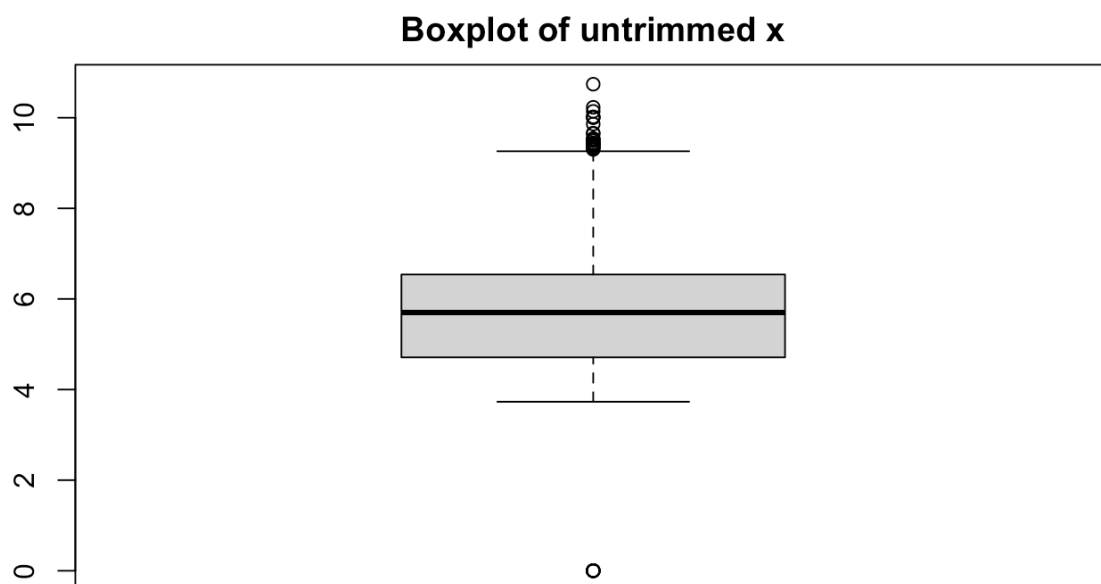
```
cor(equal_carat$log_price, equal_carat$color_num, method = 'spearman')
[1] 0.1916593
```

```
cor(equal_carat$log_price, equal_carat$clarity_num, method = 'spearman')
[1] 0.788343
```

```
#x
x = diamonds$x
hist(x)
```



```
boxplot(x, main = "Boxplot of untrimmed x")
```



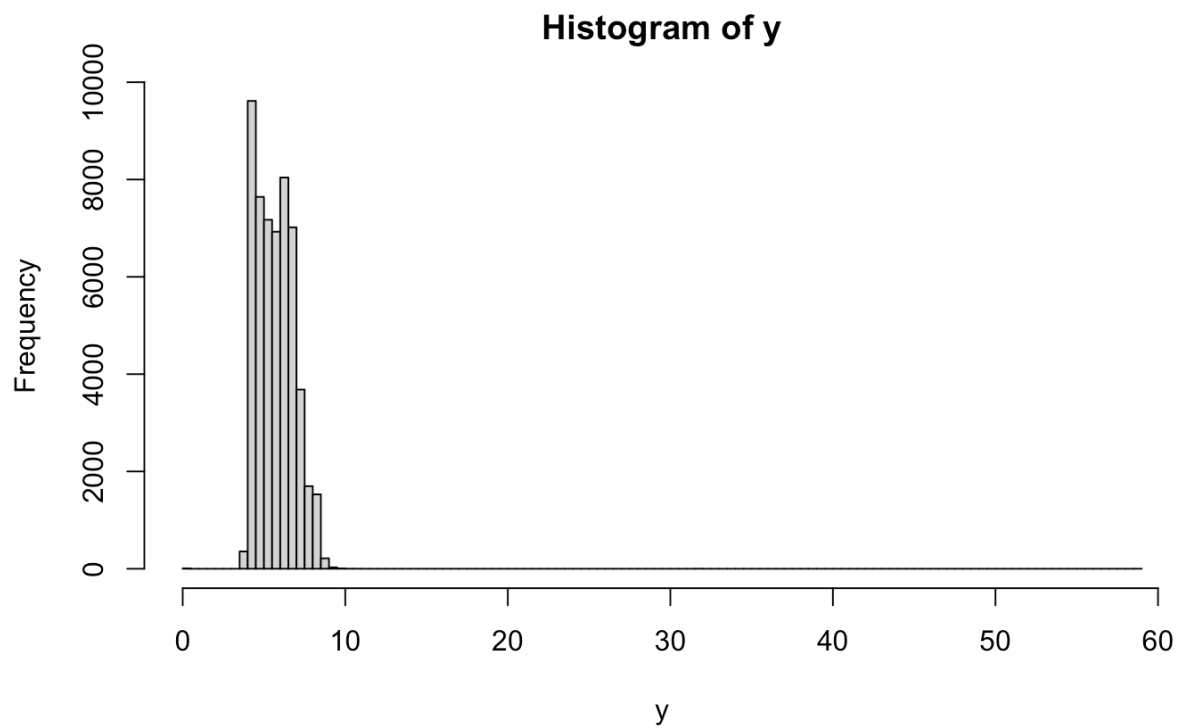
```
summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	4.710	5.700	5.731	6.540	10.740

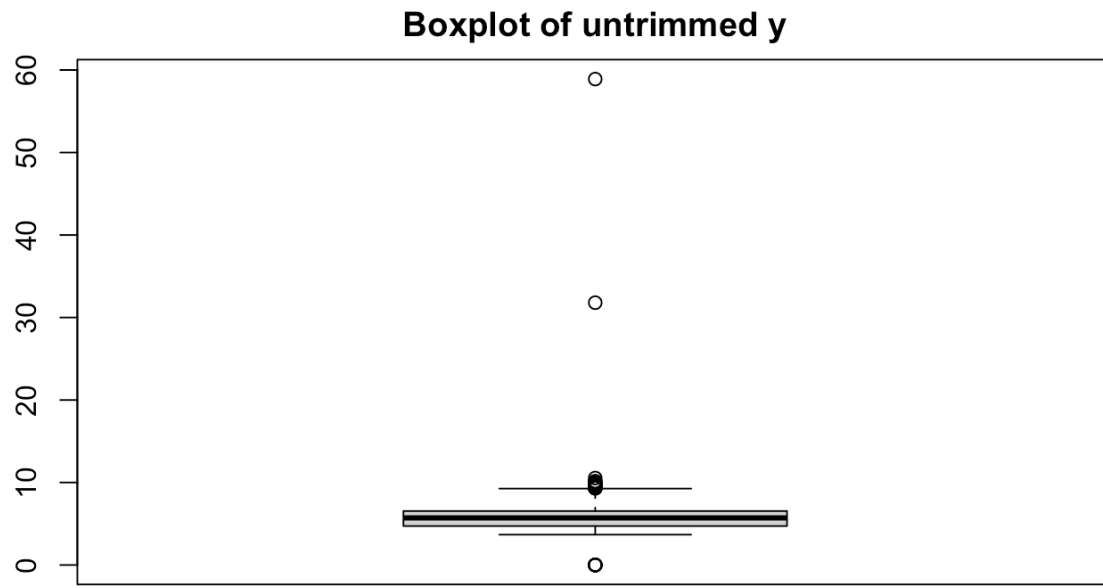
```
#y
```

```
y = diamonds$y
```

```
hist(y, breaks = 100, ylim = c(0, 10000))
```



```
boxplot(y, main = "Boxplot of untrimmed y")
```



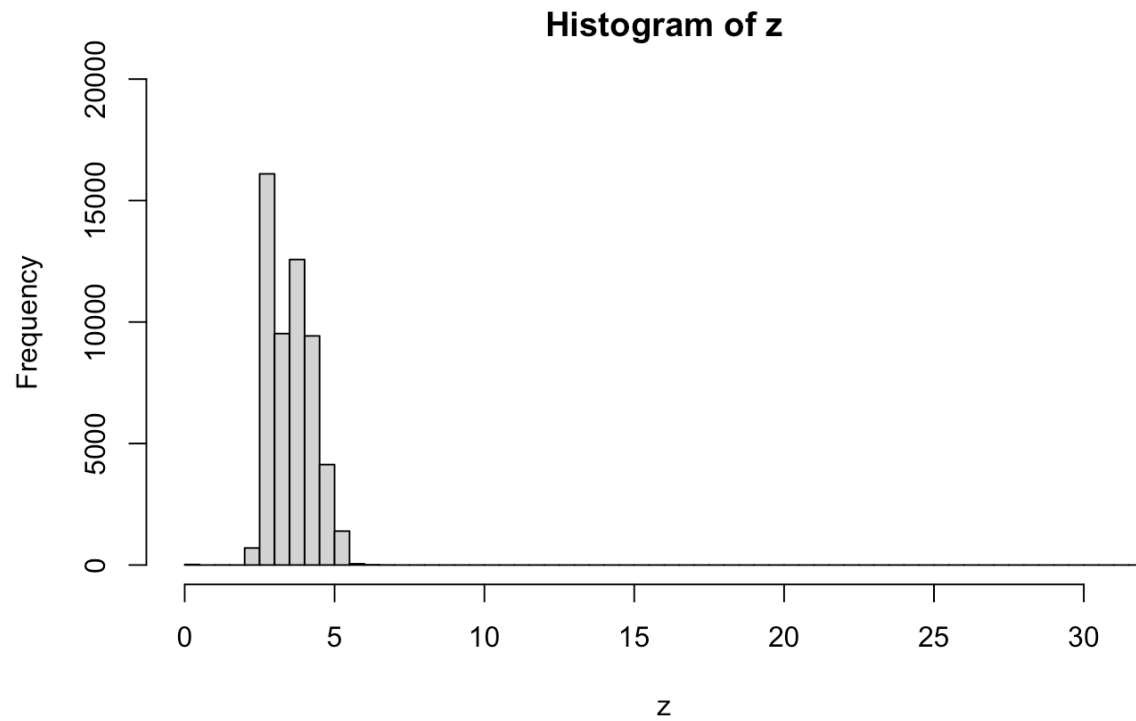
```
summary(y)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	4.720	5.710	5.735	6.540	58.900

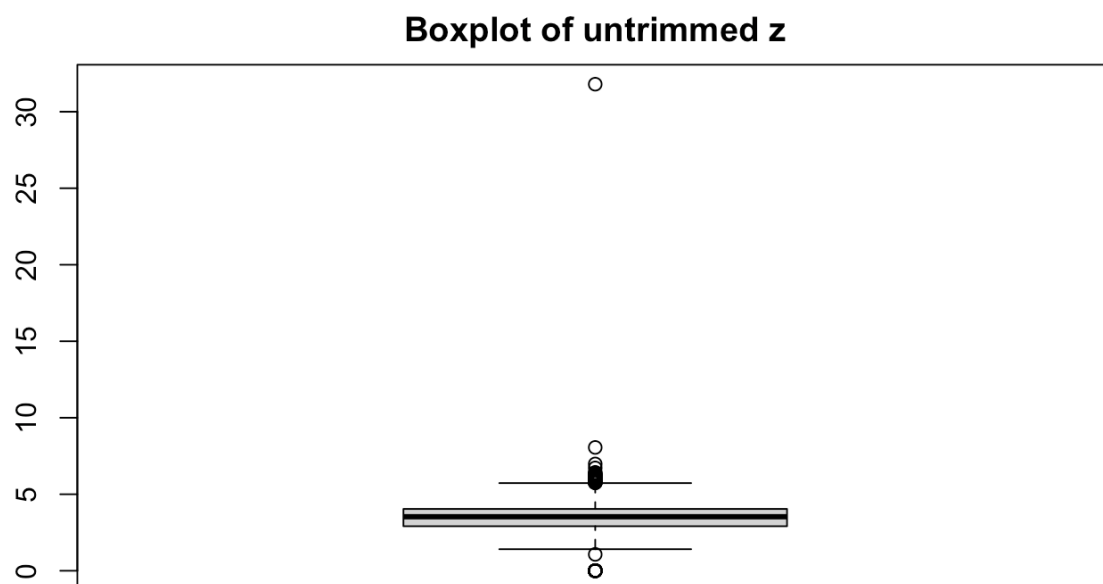
```
#z
```

```
z = diamonds$z
```

```
hist(z, breaks = 100, ylim = c(0, 20000))
```



```
boxplot(z, main = "Boxplot of untrimmed z")
```




```
summary(z)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.910	3.530	3.539	4.040	31.800

```
#z outlier
```

```
diamonds[48411, ]
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.51	Very Good	E	VS1	61.8	54.7	1970	5.12	5.15	31.8

1 row

```
#y outlier
```

```
diamonds[24068, ]
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
2	Premium	H	SI2	58.9	57	12210	8.09	58.9	8.06

1 row

```
diamonds[49190, ]
```

carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
0.51	Ideal	E	VS1	61.8	55	2075	5.15	31.8	5.12

1 row

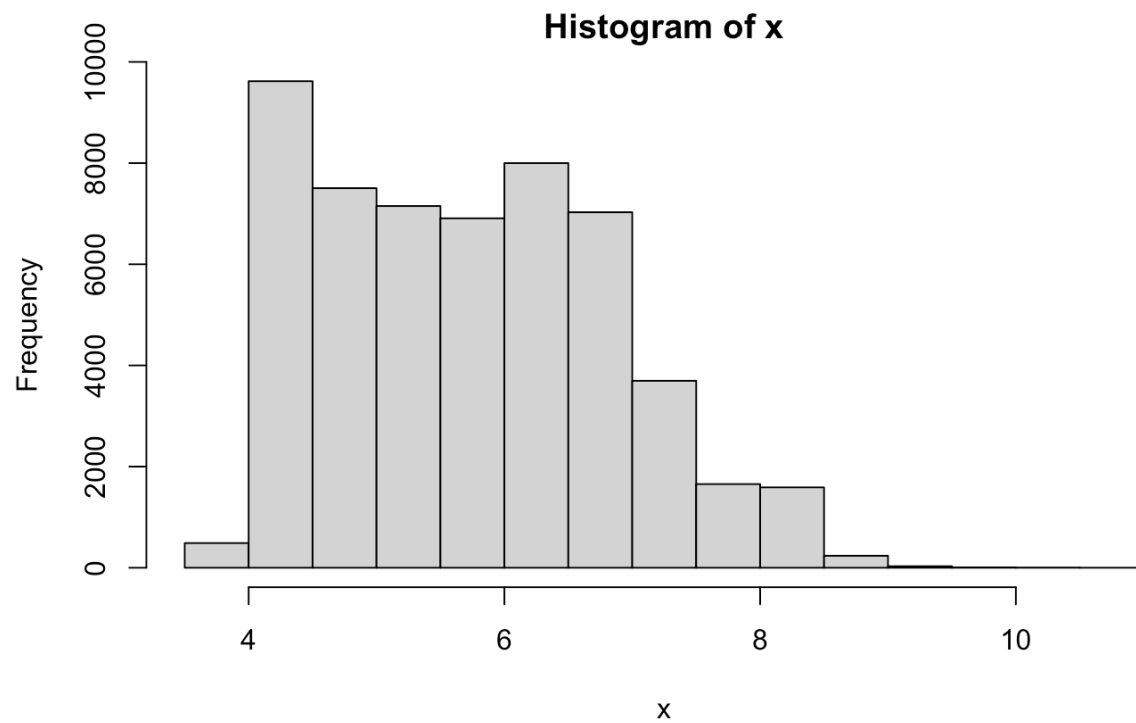
```
#x outlier
```

```
diamonds[11183, ]
```

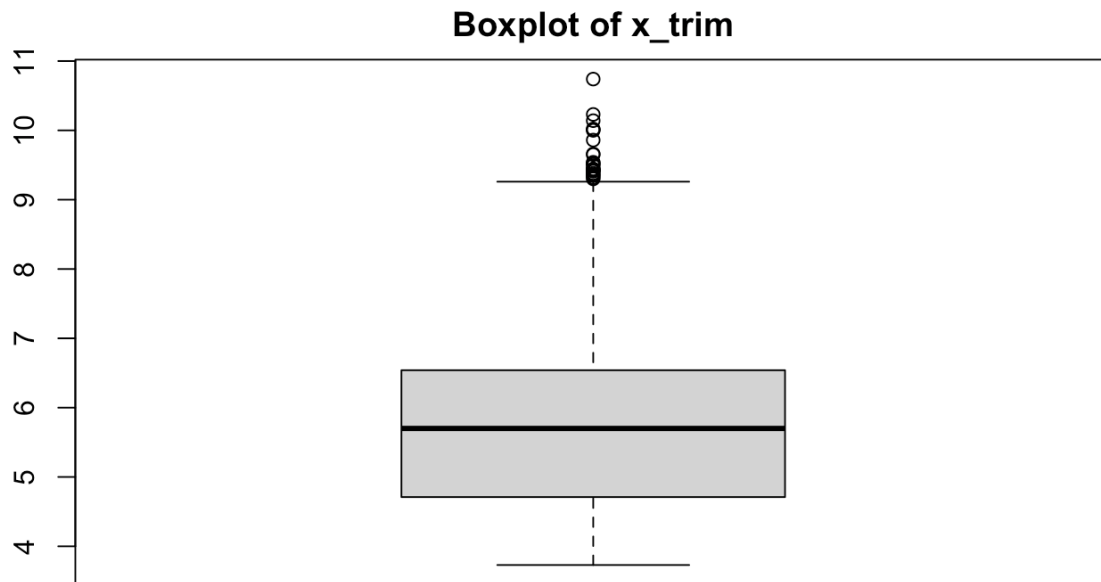
carat <dbl>	cut <ord>	color <ord>	clarity <ord>	depth <dbl>	table <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>
1.07	Ideal	F	SI2	61.6	56	4954	0	6.62	0

1 row

```
#new x  
x = new_diamonds$x  
hist(x)
```



```
boxplot(x, main = "Boxplot of x_trim")
```



```
summary(x)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.730	4.710	5.700	5.732	6.540	10.740

```
#Statistical analysis of new x
```

```
plot(
```

```
  new_diamonds$log_price ~ new_diamonds$x,
```

```
  xlab = "x",
```

```
  ylab = "log(price)",
```

```
  main = "log(price) vs x"
```

```
)
```

```
print(paste(
```

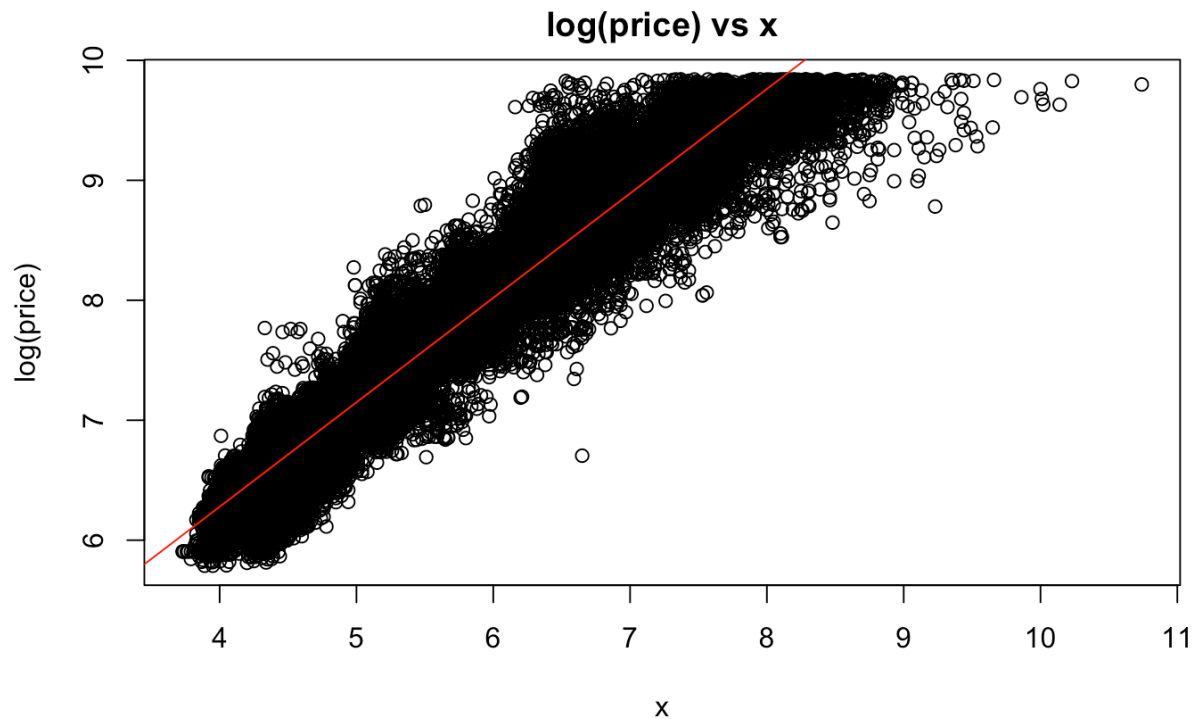
```
  "The correlation coefficient between x_trim and log(price):",
```

```
  cor(new_diamonds$x, new_diamonds$log_price)
```

```
))
```

```
[1] "The correlation coefficient between x_trim and log(price): 0.960723864418426"
```

```
fit_x <- lm(log_price ~ x, data = new_diamonds)
abline(fit_x, col = "red")
```



```
summary(fit_x)
```

Call:

```
lm(formula = log_price ~ x, data = new_diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.3482	-0.1750	0.0031	0.1771	1.4509

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.795671	0.006326	441.9	<2e-16 ***
x	0.870731	0.001083	803.9	<2e-16 ***

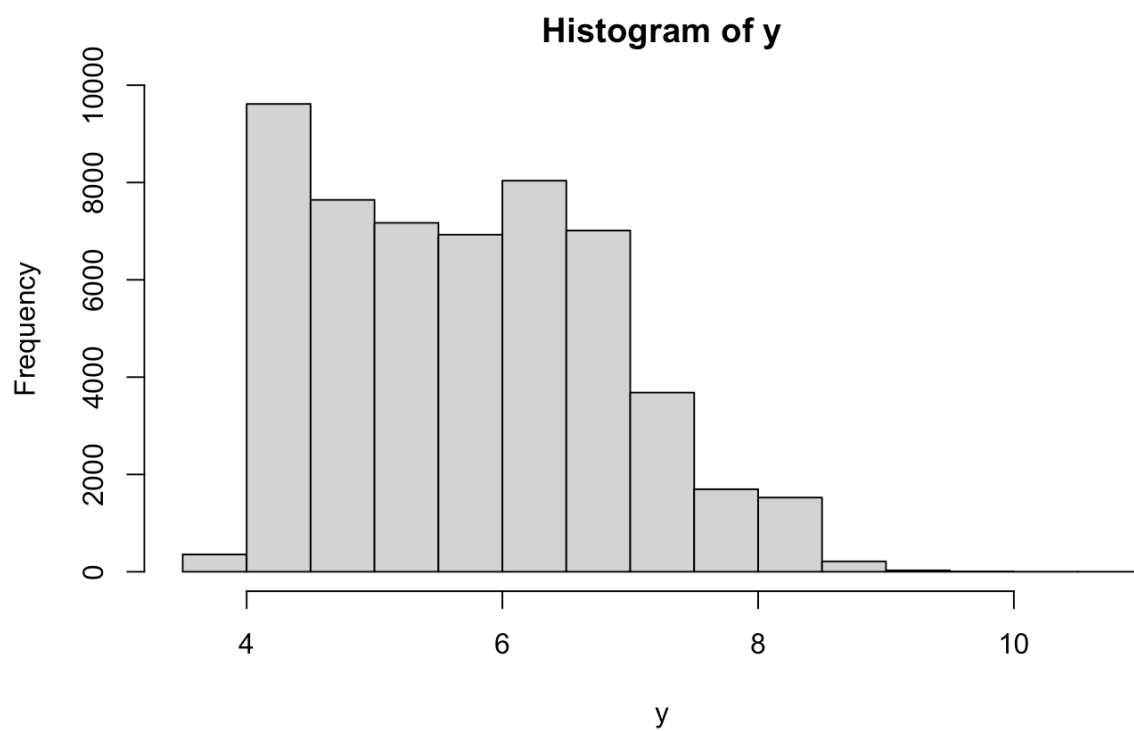
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2815 on 53914 degrees of freedom
```

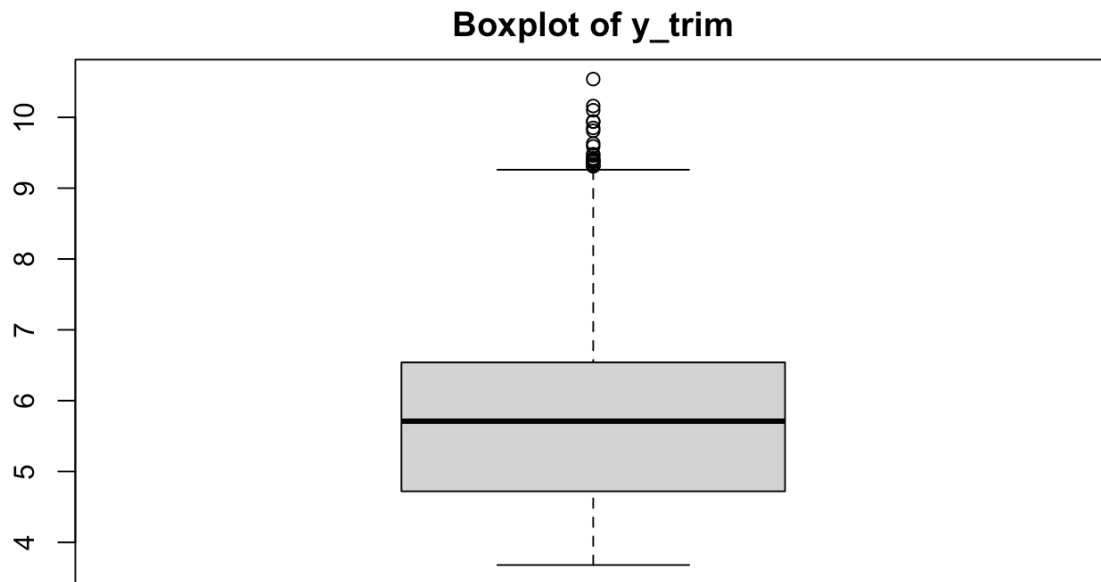
```
Multiple R-squared:  0.923, Adjusted R-squared:  0.923
```

```
F-statistic: 6.462e+05 on 1 and 53914 DF,  p-value: < 2.2e-16
```

```
#new y  
y = new_diamonds$y  
hist(y, ylim = c(0, 10000))
```



```
boxplot(y, main = "Boxplot of y_trim")
```



```
summary(y)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.680	4.720	5.710	5.733	6.540	10.540

```
#Statistical analysis of new y
```

```
plot(
```

```
  new_diamonds$log_price ~ new_diamonds$y,
```

```
  xlab = "y",
```

```
  ylab = "log(price)",
```

```
  main = "log(price) vs y"
```

```
)
```

```
print(paste(
```

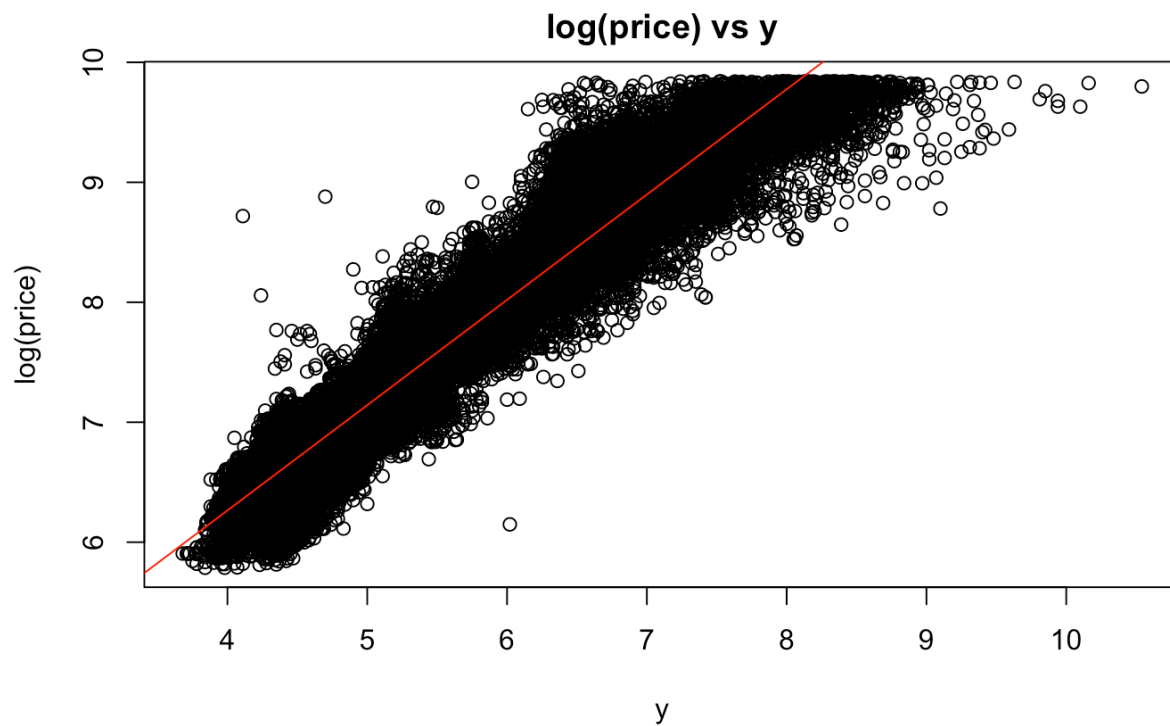
```
  "The correlation coefficient between y_trim and log(price):",
```

```
  cor(new_diamonds$y, new_diamonds$log_price)
```

```
))
```

```
[1] "The correlation coefficient between y_trim and log(price): 0.961514262054473"
```

```
fit_y <- lm(log_price ~ y, data = new_diamonds)
abline(fit_y, col = "red")
```



```
summary(fit_y)
```

Call:

```
lm(formula = log_price ~ y, data = new_diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.20655	-0.17490	0.00344	0.17763	2.35722

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.753420	0.006309	436.4	<2e-16 ***
y	0.877824	0.001080	812.6	<2e-16 ***

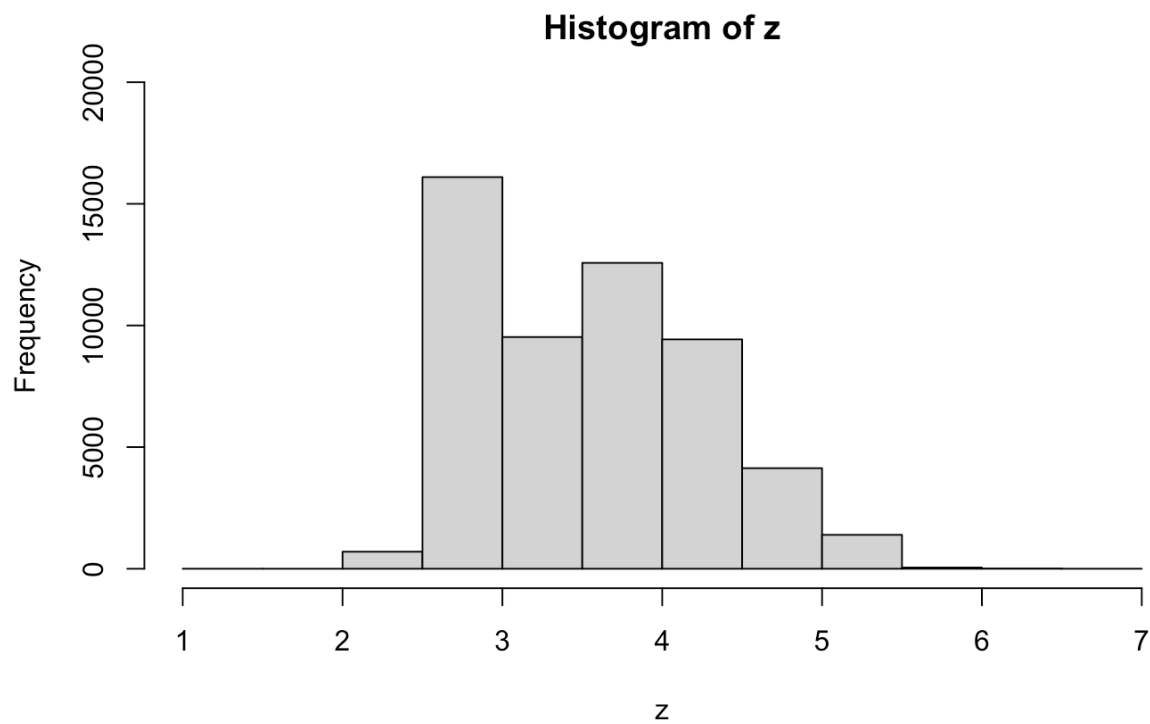
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2788 on 53914 degrees of freedom
```

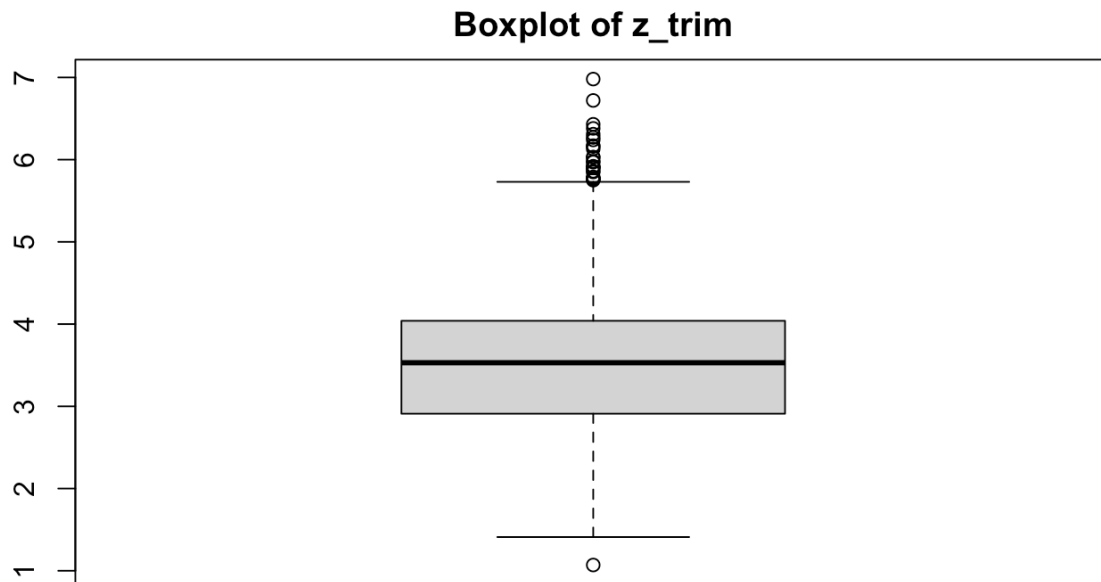
```
Multiple R-squared:  0.9245,    Adjusted R-squared:  0.9245
```

```
F-statistic: 6.603e+05 on 1 and 53914 DF,  p-value: < 2.2e-16
```

```
#new z  
z = new_diamonds$z  
hist(z, ylim = c(0, 20000))
```



```
boxplot(z, main = "Boxplot of z_trim")
```

```
summary(z)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.070	2.910	3.530	3.539	4.040	6.980

```
#Statistical analysis of new z
```

```
plot(
```

```
  new_diamonds$log_price ~ new_diamonds$z,
```

```
  xlab = "z",
```

```
  ylab = "log(price)",
```

```
  main = "log(price) vs z"
```

```
)
```

```
print(paste(
```

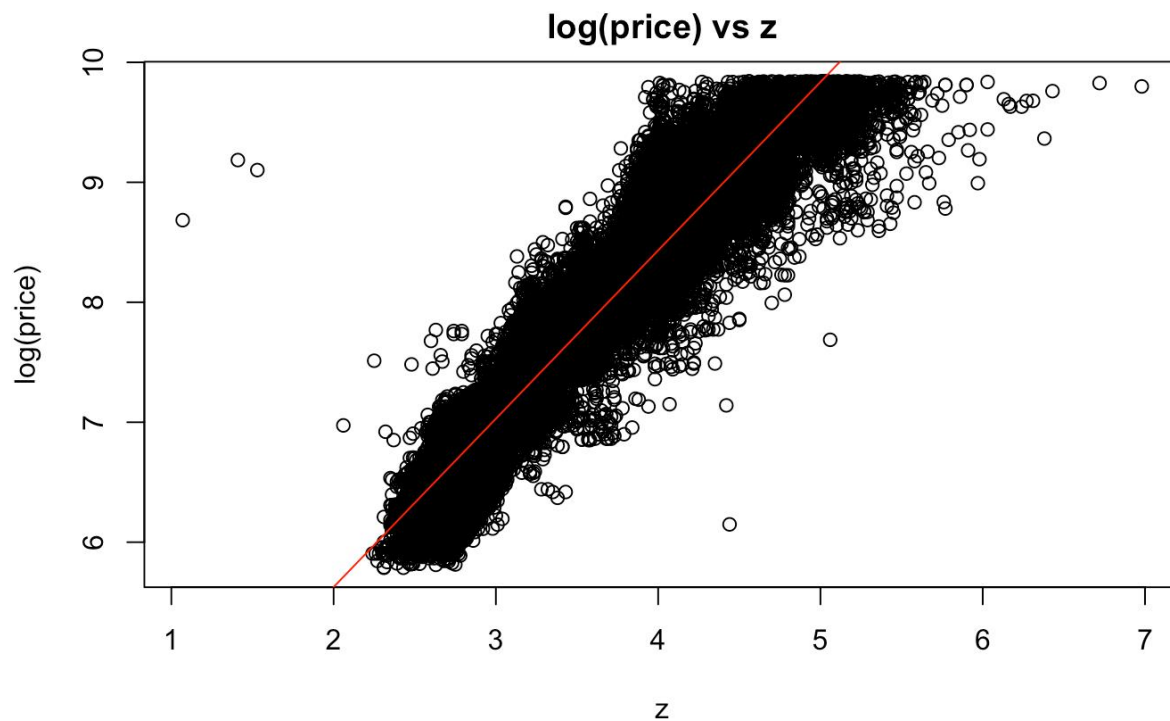
```
  "The correlation coefficient between z_trim and log(price):",
```

```
  cor(new_diamonds$z, new_diamonds$log_price)
```

```
))
```

```
[1] "The correlation coefficient between z_trim and log(price): 0.95662410713  
1078"
```

```
fit_z <- lm(log_price ~ z, data = new_diamonds)
abline(fit_z, col = "red")
```



```
summary(fit_z)
```

Call:

```
lm(formula = log_price ~ z, data = new_diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.9017	-0.1799	0.0041	0.1840	4.3870

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.819569	0.006637	424.8	<2e-16 ***
z	1.403289	0.001840	762.5	<2e-16 ***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.2956 on 53914 degrees of freedom

Multiple R-squared: 0.9151, Adjusted R-squared: 0.9151

F-statistic: 5.813e+05 on 1 and 53914 DF, p-value: < 2.2e-16

```
# Multivariate linear regression between log(price) and (x,y,z)
new_diamonds$x_norm <-
  (new_diamonds$x - mean(new_diamonds$x)) / sd(new_diamonds$x)
new_diamonds$y_norm <-
  (new_diamonds$y - mean(new_diamonds$y)) / sd(new_diamonds$y)
new_diamonds$z_norm <-
  (new_diamonds$z - mean(new_diamonds$z)) / sd(new_diamonds$z)

fit_xyz_plus_norm <-
  lm(log_price ~ x_norm + y_norm + z_norm, data = new_diamonds)
summary(fit_xyz_plus_norm) # show results
```

Call:

```
lm(formula = log_price ~ x_norm + y_norm + z_norm, data = new_diamonds)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.36374	-0.17168	0.00166	0.17585	1.87111

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.786368	0.001194	6523.869	<2e-16 ***
x_norm	0.054494	0.023724	2.297	0.0216 *
y_norm	0.707319	0.023278	30.385	<2e-16 ***
z_norm	0.215764	0.009054	23.830	<2e-16 ***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2771 on 53912 degrees of freedom
Multiple R-squared: 0.9254, Adjusted R-squared: 0.9254
F-statistic: 2.229e+05 on 3 and 53912 DF, p-value: < 2.2e-16
```

```
# Multivariate linear regression between log(price) and (x,y,z,log_carat) (section 4.6)
new_diamonds$log_carat_norm <-
  (new_diamonds$log_carat - mean(new_diamonds$log_carat)) / sd(new_diamonds$log_carat)
fit_multi <-
  lm(log_price ~ x_norm + y_norm + z_norm + log_carat_norm, data = new_diamonds)
summary(fit_multi) # show results
```

Call:

```
lm(formula = log_price ~ x_norm + y_norm + z_norm + log_carat_norm,
    data = new_diamonds)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.47730	-0.16940	-0.00578	0.16583	1.32141

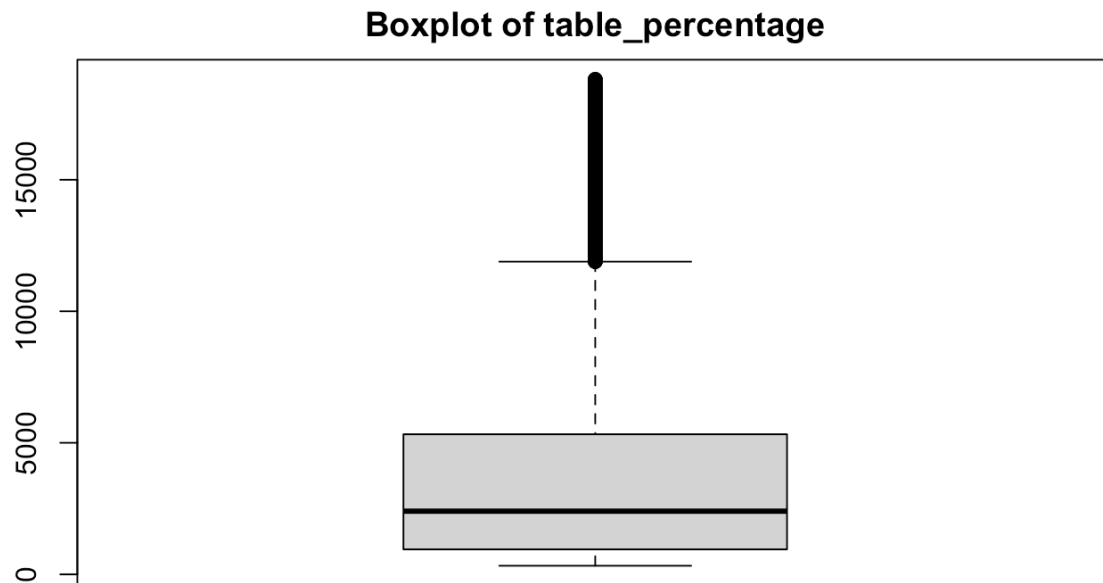
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.786368	0.001121	6947.83	<2e-16 ***
x_norm	-0.277449	0.022615	-12.27	<2e-16 ***
y_norm	0.514573	0.021975	23.42	<2e-16 ***
z_norm	-0.176496	0.009672	-18.25	<2e-16 ***
log_carat_norm	0.919778	0.010813	85.06	<2e-16 ***

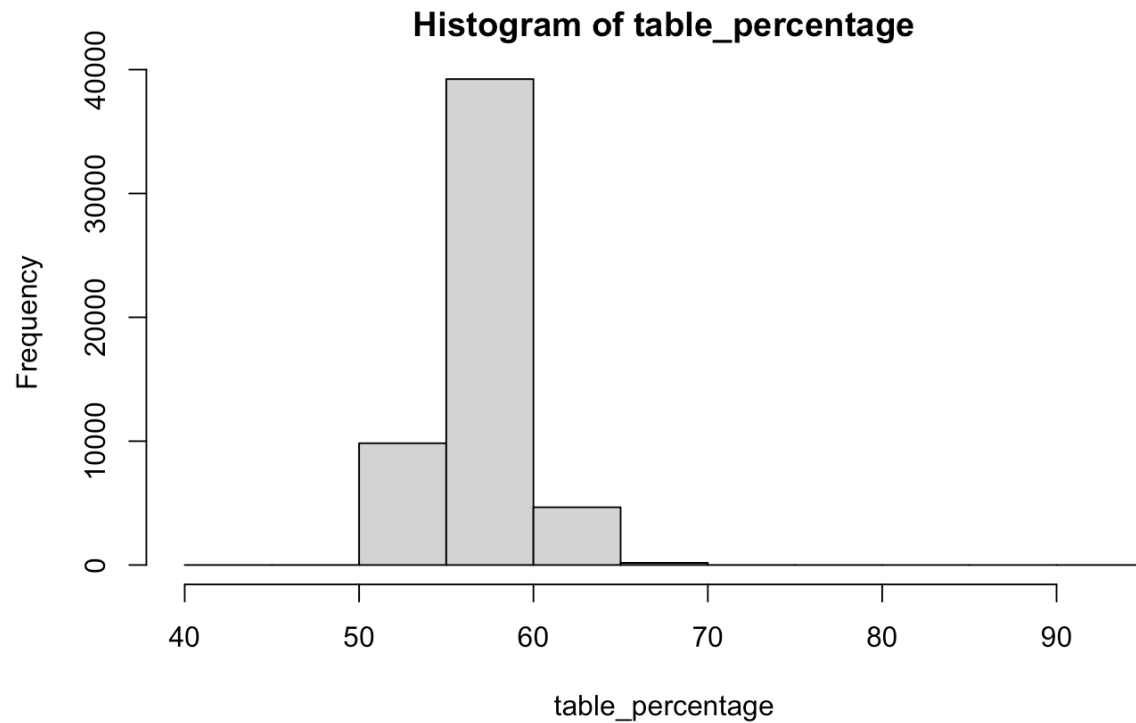
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 0.2602 on 53911 degrees of freedom
Multiple R-squared: 0.9342, Adjusted R-squared: 0.9342
F-statistic: 1.914e+05 on 4 and 53911 DF, p-value: < 2.2e-16
```

```
#table_percentage  
table_percentage = new_diamonds$table  
boxplot(price, main = "Boxplot of table_percentage")
```



```
hist(table_percentage)
```



```
summary(table_percentage)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
43.00	56.00	57.00	57.46	59.00	95.00

```
#table_percentage statistical analysis [Linear Regression]
plot(
  new_diamonds$log_price ~ new_diamonds$table,
  main = "log(price) vs Table Percentage",
  xlab = "Table Percentage",
  ylab = "log(price)"
)
fit_table = lm(formula = log_price ~ table , data = new_diamonds)
summary(fit_table)
```

Call:

```
lm(formula = log_price ~ table, data = new_diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.53895	-0.88522	-0.00377	0.78319	2.42837

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.653930	0.111090	32.89	<2e-16 ***
table	0.071923	0.001932	37.23	<2e-16 ***

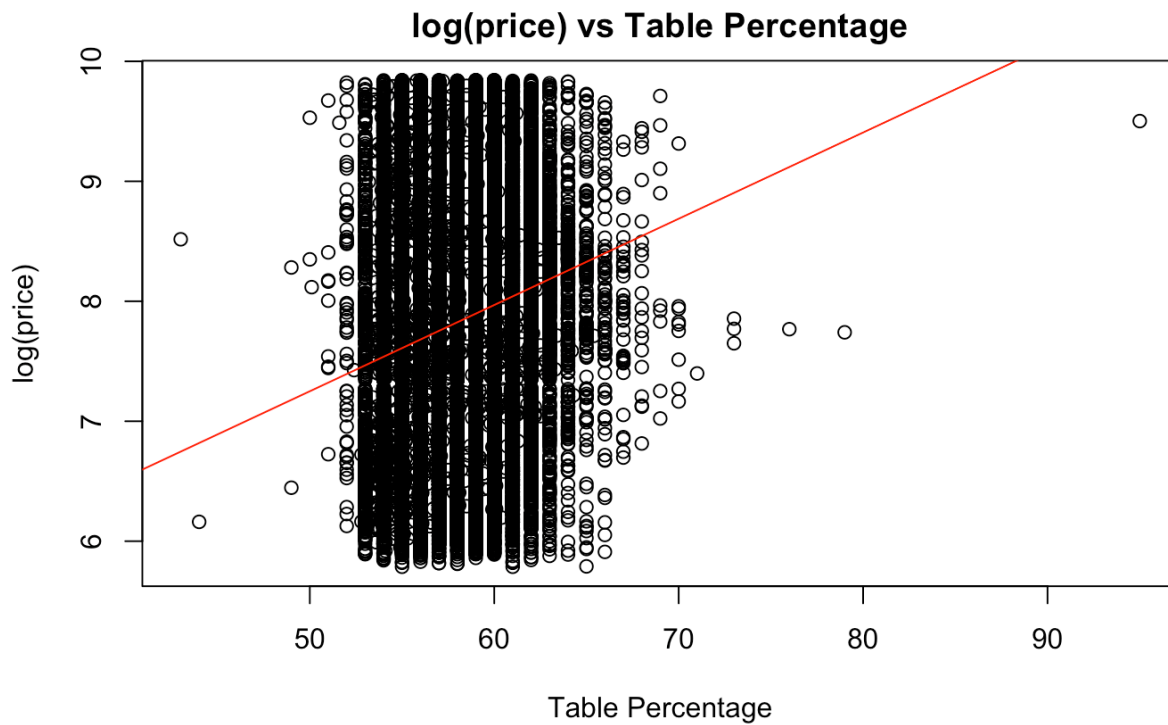
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 53914 degrees of freedom

Multiple R-squared: 0.02506, Adjusted R-squared: 0.02504

F-statistic: 1386 on 1 and 53914 DF, p-value: < 2.2e-16

```
abline(fit_table, col = "red")
```



```
print(paste(
```

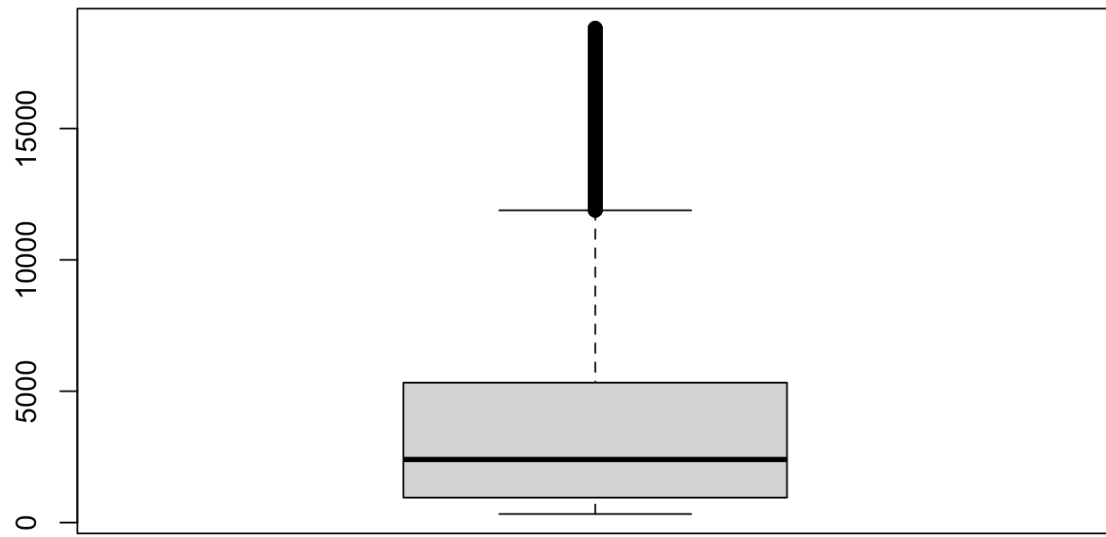
```
"The correlation coefficient between table_percentage and log_price:",  
cor(new_diamonds$table, new_diamonds$log_price)  
)  
[1] "The correlation coefficient between table_percentage and log_price: 0.15  
8305764001349"
```

```
cor.test(new_diamonds$table, new_diamonds$log_price)
```

```
Pearson's product-moment correlation  
  
data: new_diamonds$table and new_diamonds$log_price  
t = 37.227, df = 53914, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.1500653 0.1665242  
sample estimates:  
      cor  
0.1583058
```

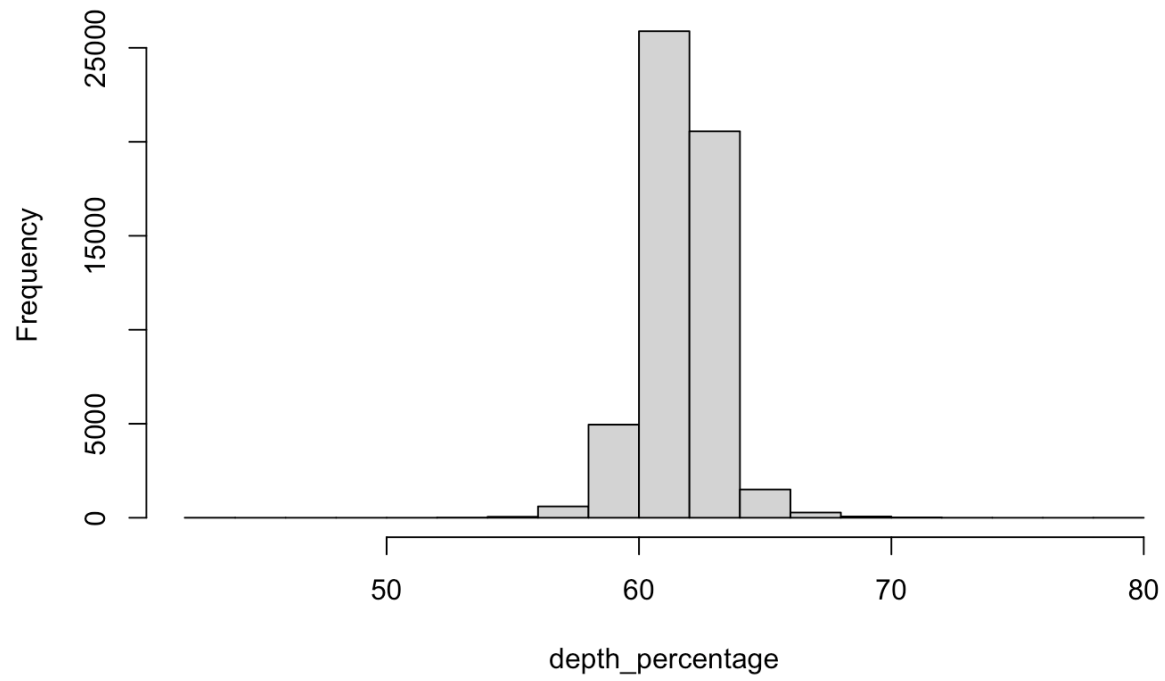
```
#depth_percentage  
depth_percentage = diamonds$depth  
boxplot(price, main = "Boxplot of depth_percentage")
```


Boxplot of depth_percentage



```
hist(depth_percentage)
```

Histogram of depth_percentage



```
summary(depth_percentage)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
43.00	61.00	61.80	61.75	62.50	79.00

```
#depth_percentage statistical analysis [Linear Regression]
```

```
plot(
```

```
  new_diamonds$log_price ~ new_diamonds$depth,
```

```
  main = "log(price) vs Depth Percentange",
```

```
  xlab = "Depth Percentage" ,
```

```
  ylab = "log(price)"
```

```
)
```

```
fit_depth = lm(formula = log_price ~ depth, data = new_diamonds)
```

```
summary(fit_depth)
```

Call:

```
lm(formula = log_price ~ depth, data = new_diamonds)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.99930	-0.93040	-0.00269	0.79336	2.05741

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.7451393	0.1886746	41.050	<2e-16 ***
depth	0.0006677	0.0030547	0.219	0.827

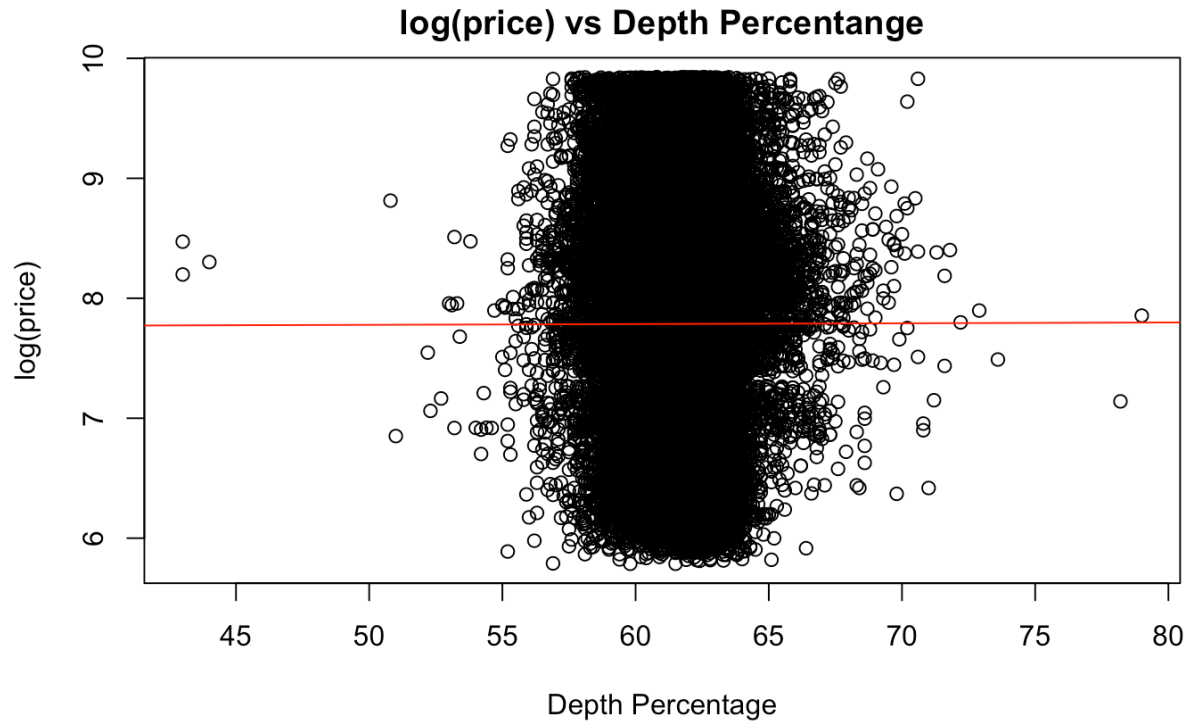
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.015 on 53914 degrees of freedom

Multiple R-squared: 8.861e-07, Adjusted R-squared: -1.766e-05

F-statistic: 0.04778 on 1 and 53914 DF, p-value: 0.827

```
abline(fit_depth, col = "red")
```



```
print(paste(
  "The correlation coefficient between depth_percentage and log_price:",
  cor(new_diamonds$depth, new_diamonds$log_price)
))
```

```
[1] "The correlation coefficient between depth_percentage and log_price: 0.00
094134786202324"
```

```
cor.test(new_diamonds$depth, new_diamonds$log_price)
```

Pearson's product-moment correlation

```
data: new_diamonds$depth and new_diamonds$log_price
t = 0.21858, df = 53914, p-value = 0.827
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.007499656 0.009382217
```

```
sample estimates:
```

```
cor
```

```
0.0009413479
```

7. References

[1] M. Garside. c2022. Statista [Internet]. [cited 2022 March 03]. Available from:

<https://www.statista.com/topics/1704/diamond-industry/#dossierKeyfigures>

[2] M. Garside. c2021. Statista [Internet]. [cited 2022 March 03]. Available from:

<https://www.statista.com/statistics/585103/diamond-jewelry-market-value-worldwide-by-region/>

[3] c2019. Q Report [Internet]. [cited 2022 March 03]. Available from:

<https://www.qreport.com.au/blog/the-beginners-guide-to-buying-diamonds>

[4] c2022. Diamond shape and diamond cut guide[Internet] [cited 2022 March 29]. Available from:

[https://www.forevermark.com/en/now-forever/guides/diamond-engagement-guide/diamond-shape-and-cut-](https://www.forevermark.com/en/now-forever/guides/diamond-engagement-guide/diamond-shape-and-cut-guide/#:~:text=The%20diamond%20shape%20refers%20to,to%20as%20'fancy%20shapes')

[guide/#:~:text=The%20diamond%20shape%20refers%20to,to%20as%20'fancy%20shapes'](https://www.forevermark.com/en/now-forever/guides/diamond-engagement-guide/diamond-shape-and-cut-guide/#:~:text=The%20diamond%20shape%20refers%20to,to%20as%20'fancy%20shapes')