L-Università ta' Malta

SOR1232

HYPOTHESIS TESTING AND STATISTICAL MODELLING

STUDY ON DIAMONDS

Compiled by: **Alan Zammit (0297306L)**

Luca Callus (0157906L)

Lecturer: **Derya KARAGOZ**

Department of Statistics and Operations Research

University of Malta

May 2025

CONTENTS

Alan Zammit, Luca Callus

1.0 - Introduction

The data, named, *'diamonds*, was obtained for the package *'ggplot2'* in rStudio, and was first accessed on 12/3/2025. Inside this dataset, data about 50,000 different diamonds was gathered. In this report, we will analyse the data to draw conclusions about diamonds in general as a population.

dependent variable - price

Objectives:

- Gain a general understanding of our data by observing the descriptive statistics
- Analyse more in depth how our data is distributed amongst its variables through the use of graphical representations
- One sample - luca
- Conduct a two-sample Mann-Whitney U Test to test whether carat influences the price
- Conduct a Kruskal-Wallis test to see if there is a significant difference between different cuts
- MLR
- ANOVA
- ANCOVA

Variable	**Description**	**Measurement units**	**Variable Type**
Price	Price of the diamond	US Dollars	Covariate Continuous
Carat	Weight of the diamond (1 carat = 0.2g)	carats	Covariate Continuous
Cut	Quality of the cut	Fair, Good, Very Good, Premium, Ideal	Categorical Nominal
Color	Diamond colour (J is worst, D is best)	From best to worst: D,E,F,G,H,I,J	Categorical Ordinal
Clarity	A measurement of the diamond's clarity (0 = worst, 8 = Best)	(I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))	Categorical Ordinal
Depth	total depth % = 2 * z / (x + y) = z / mean(x, y)	Percentage of depth	Covariate Continuous
Table	% of width of the top of the diamond, relative to its widest point	Table percentage	Covariate Continuous

X	length	mm	Covariate Continuous
Y	width	mm	Covariate Continuous
Z	depth	mm	Covariate Continuous

Table 1.1: Brief descriptions of the dataset variables used from the “diamonds” dataset

2.0 - Exploratory Data Analysis

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Carat	0.2000	0.4000	0.7000	0.7979	1.0400	5.0100
Depth	43.00	61.00	61.80	61.75	62.50	79.00
Table	43.00	56.00	57.00	57.46	59.00	95.00
Price	326	950	2401	3933	5324	18823

Table 2.01: Summary of the covariate variables

Variable	N	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
Carat	53940	0.80	0.47	0.70	0.20	5.01	4.81	1.12	1.26
Depth	53940	61.75	1.43	61.80	43.00	79.00	36.00	-0.08	5.74
Table	53940	57.46	2.23	57.00	43.00	95.00	52.00	0.80	2.80
Price	53940	3932.80	3989.44	2401	326	18823	18497	1.62	2.27
X	53932	5.73	1.12	5.70	3.73	10.74	7.01	0.00	-0.70
Y	53933	5.74	1.14	5.71	3.68	58.90	55.22	2.46	91.72
Z	53920	3.54	0.71	3.53	1.07	31.80	30.73	1.59	47.76

Table 2.02: The descriptive statistics of all covariate variables

From the descriptive statistics that were generated for the covariate variables from table 2.01, it may be noted that not each variable has the same amount of diamonds as missing values were removed before calculating statistics. From these statistics interesting observations can already be made. From the standard deviation of carat (0.47), it can be seen that there are many diamonds on the tails, if you compare it with the Interquartile range, and the skewness shows that most diamonds are very light. Its skewness also follows this assumption as carat is positively skewed (1.12). Price is also interesting as the standard deviation is greater than the mean, which would indicate that there are a fair number of very expensive diamonds, that being said, the high skewness (1.62) and kurtosis (2.18) shows that there are much more diamonds on the cheaper side. The statistics of depth seem to show that there is not a large variety in depth as it is Leptokurtic (kurtosis > 5), which means most values are concentrated around the mean, as is also supported by the low standard deviation (1.43).

2.1 - Graphical Representations:

2.11 - Pie Charts

For all 3 categorical variables, cut, color and clarity, pie charts were generated to observe differences in frequencies between each one.

Cut	Fair	Good	Very Good	Premium	Ideal
Frequency	1610	4906	12082	13791	21551

Table 2.111: Frequency table for cut

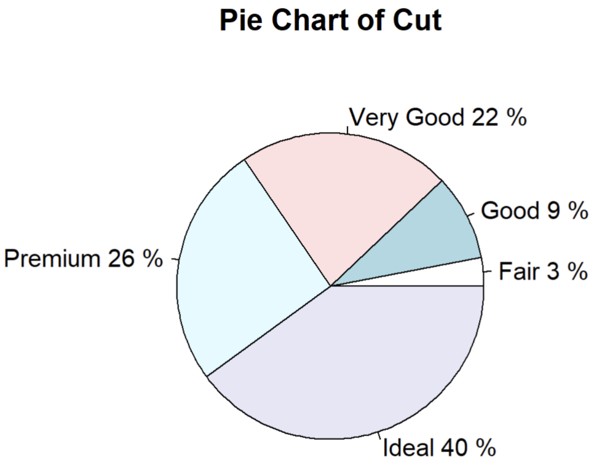


Figure 2.112: Pie chart of cut

The pie chart of cut shows that most of the diamonds are of 'Ideal' cut. Although cut is not a categorical ordinal variable, from the names themselves a sense of order may be instilled, and from this sense it can be observed that the less sought for cuts, 'Good' and 'Fair' or not as common as the others.

Color	D	E	F	G	H	I	J
Frequency	6775	9797	9542	11292	8304	5422	2808

Table 2.113: Frequency table of color

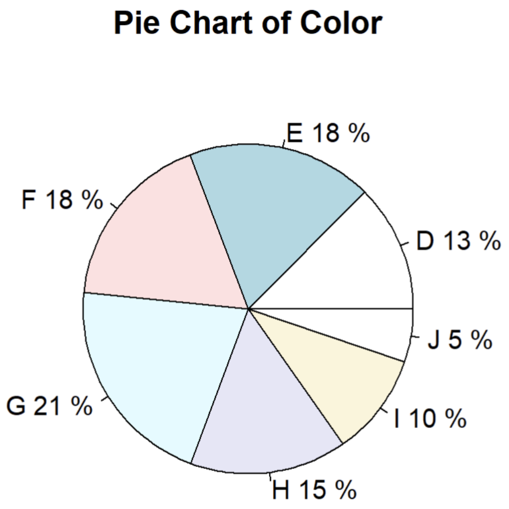


Figure 2.114: Pie chart of color

From Color's pie chart it can be seen that all colors, except for the 2 worst ones, J and I, are evenly distributed amongst diamonds, with the central values G, F and E having a slightly greater occurrence.

Clarity	I1	SI2	SI1	VS2	VS1	VVS2	VVS1	IF
Frequency	741	9194	13065	12258	8171	5066	3655	1790

Table 2.115: Frequency table of clarity

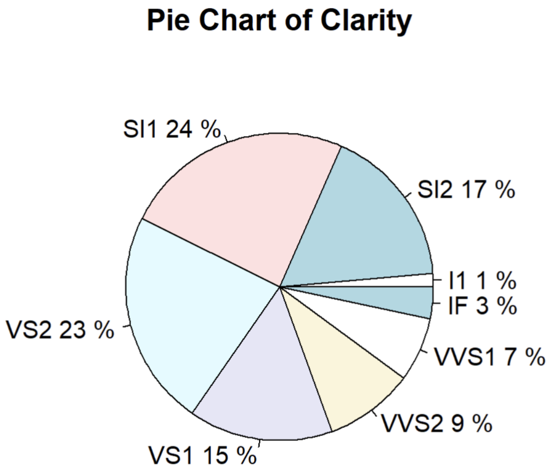


Figure 2.116: Pie chart of clarity

The pie chart of clarity indicates that the better clarities are not the most sought after, as the middle to worse ones, VS2, SI1 and S12 have the largest frequency. The best to worse clarities show a gradual increase in frequency, whereas in the worst, I1, to the one above it, S12, the frequency is increased by 16%. This may show that I1 is either significantly worse than SI2, or not really considered good enough to sell, as they only hold 1% of the sample.

2.2 - Box Plots

First the Box Plots for the categorical variables were generated:

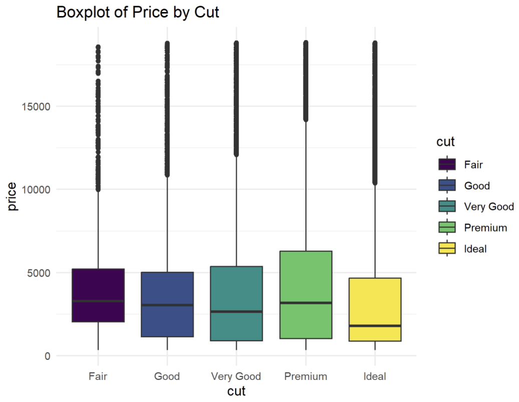


Figure 2.21: Box plot of cut

From the boxplots of figure 2.21, it seems that the difference between cuts isn't significant, and there isn't any specific order., as all the medians, except for Ideal's, are similar.

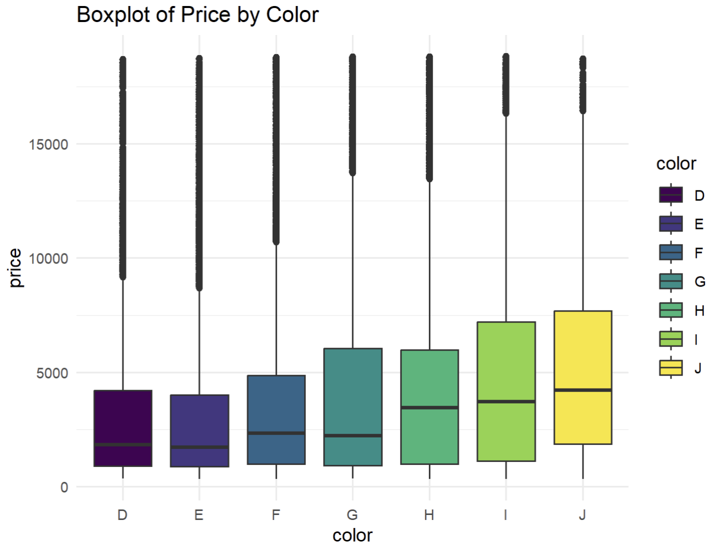


Figure 2.22: Box plot of color

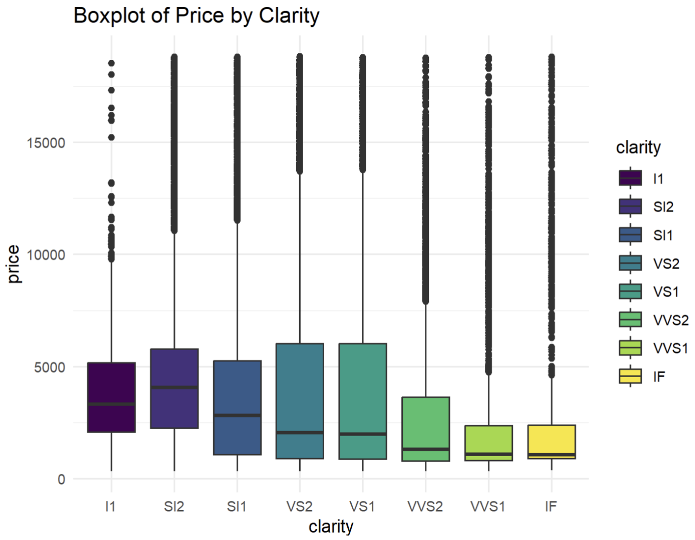


Figure 2.23: Box plot of clarity

The two above box-plots, figures 2.22 and 2.23, are interesting, as they are not as anticipated. The better colors, D,E and F, and the better clarities, IF, VVS1 and VVS2, have smaller medians than those of the worse side of the scale. In fact, the medians slowly increase as the quality gets worse. It stands to reason that the better something is in its aspect, the more expensive it should be. This means that there may be another stronger factor which affects the price of diamonds, as the better clarities and colors are not having the impact it may be expected of them on the price.

Then the Box Plots of some covariate variables were generated

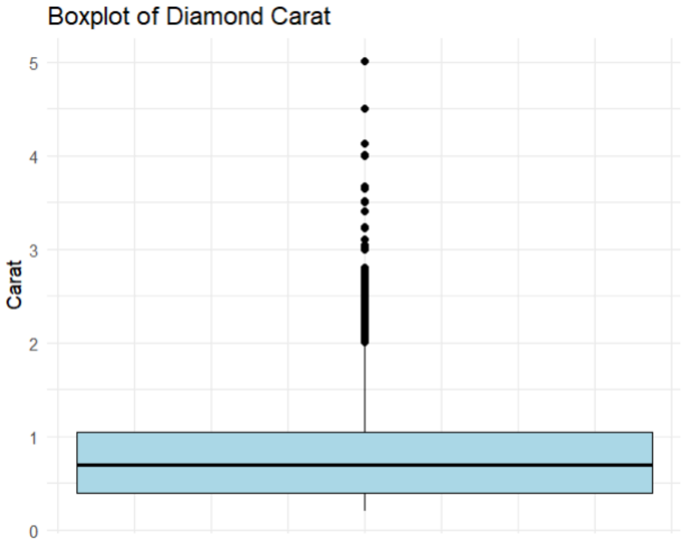


Figure 2.24: Box plot of carat

As can be observed from figure 2.24, the median value of carat is around 0.75, and mainly ranges from around 0.4 to 1.1 carat. However, there are still a significant number of outliers, particularly between the 2 and 3 carat ranges, yet there even exists a very rare 5 carat outlier.

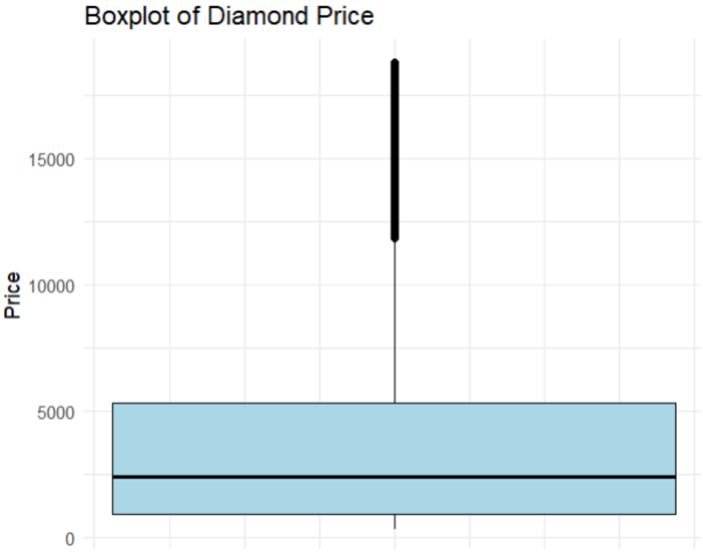


Figure 2.25: Box plot of price

From figure 2.25, it is evident that the median diamond price is about $2500, whilst prices normally range from $1250 to $5100. Yet, there still exist some high-priced outliers, mainly between the $12000 and the $19000 mark.

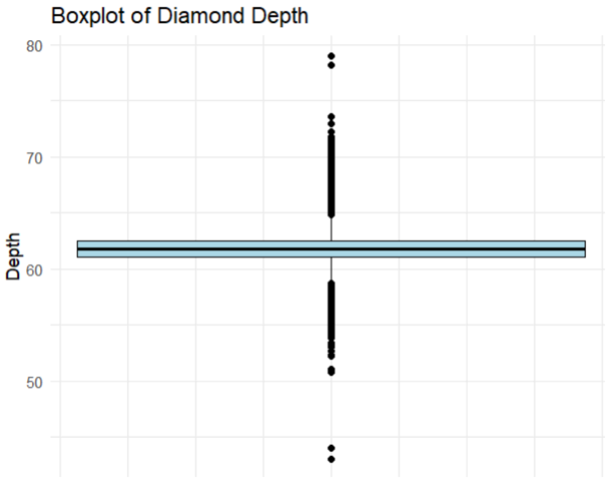


Figure 2.26: Box plot of depth

From figure 2.26, it can be seen that the depth of diamonds differs minimally from the median of around 62%. There still are, however, quite a few outliers, mainly between 50% to 59%, and 64% to 74%.

2.3 - Histograms

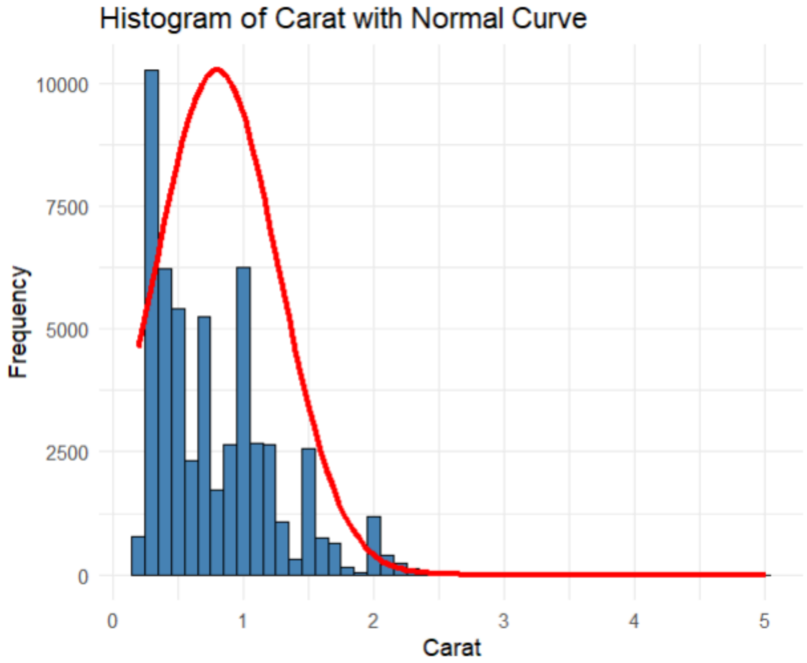


Figure 2.31: Histogram of carat

As was observed in the descriptive statistics, the histogram in figure 2.31 shows that carat is positively skewed and that most diamonds are around 0.2 carats, although there is a good amount of heavier diamonds, mainly at 1 carat. The normal curve indicates that carat isn't normally distributed since it is positively skewed.

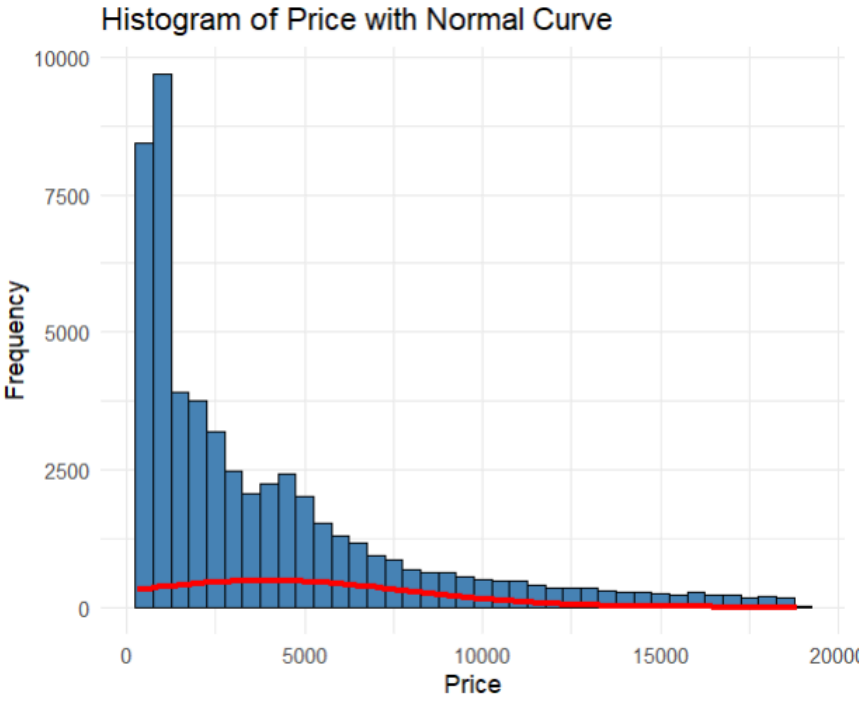


Figure 2.32: Histogram of price

As was also seen from the descriptive statistics, figure 2.32 shows that price is certainly right skewed and diamonds between the $326 and $1000 range are the most common. That being said, there are still a fair amount of expensive diamonds, as after the $1000 mark, the frequency gradually falls. From the attempted fitting of a normal curve to this histogram, it is clear that price does not follow a normal distribution as the curve doesn’t fit the data well.

2.4 - Clustered Bar Charts

These graphs were obtained by splitting the carat into reasonable ranges, and for each range, the average price for each categorical variable was outputted as a clustered bar chart.

Carat Range	0-0.5	0.5-1	1-1.5	1.5-2	2-3	3-5.02
Frequency	17674	17206	12825	4081	2114	40

Table 2.41: Frequency table of defined carat ranges

From the frequency table, there is a good amount of data from the first 3 ranges, whereas in the last range, 3-5.02, there are only 40 instances, so proper observations cannot be made from this range as it is highly dependent on the outliers, which may skew the results.

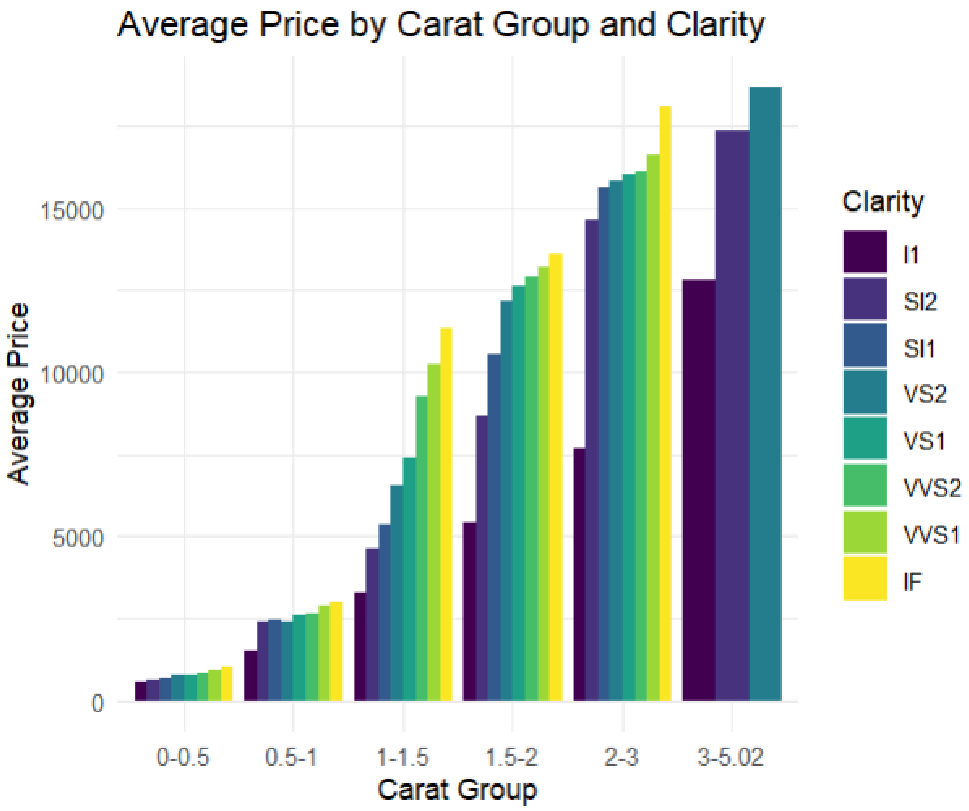


Figure 2.42: Clustered bar chart of carat range split by clarity by average price

As expected, figure 2.42 shows that the higher order clarities have a higher average price than the lower order ones, at the same weight. Therefore, the observations made from the box-plot could mean that there are more heavy-lower clarity diamonds than there are high clarity diamonds.

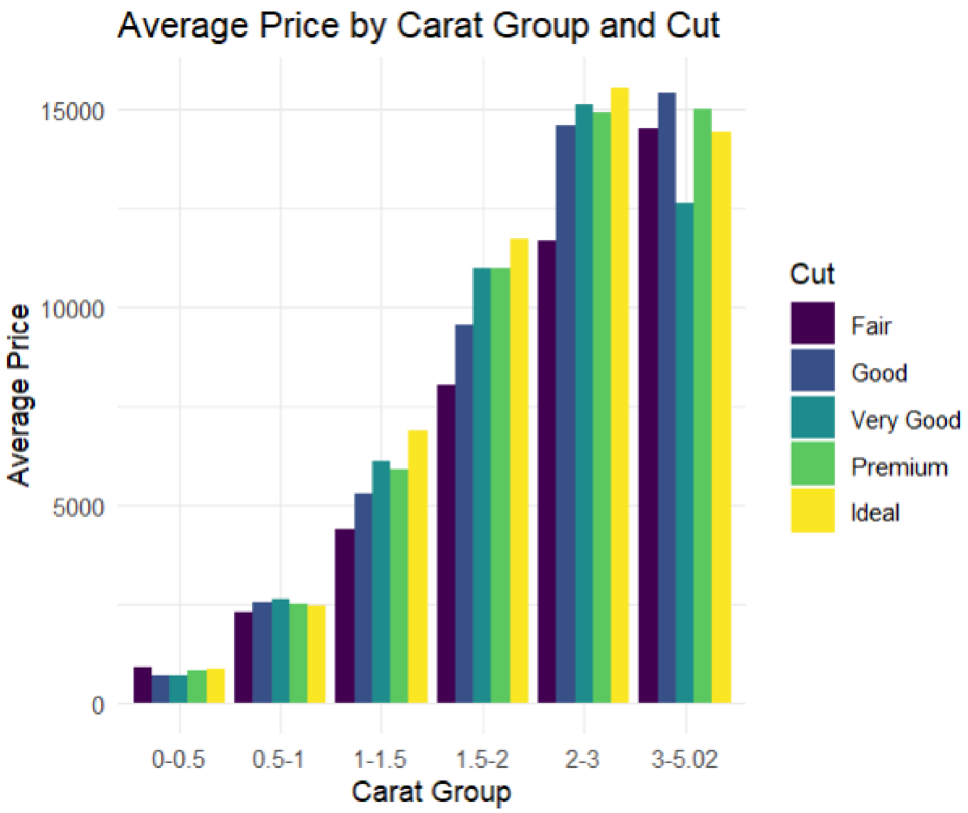


Figure 2.43: Clustered bar chart of carat range split by cut by average price

Figure 2.43 shows that most cuts are evenly priced between each other, with Ideal maybe being the most expensive one.

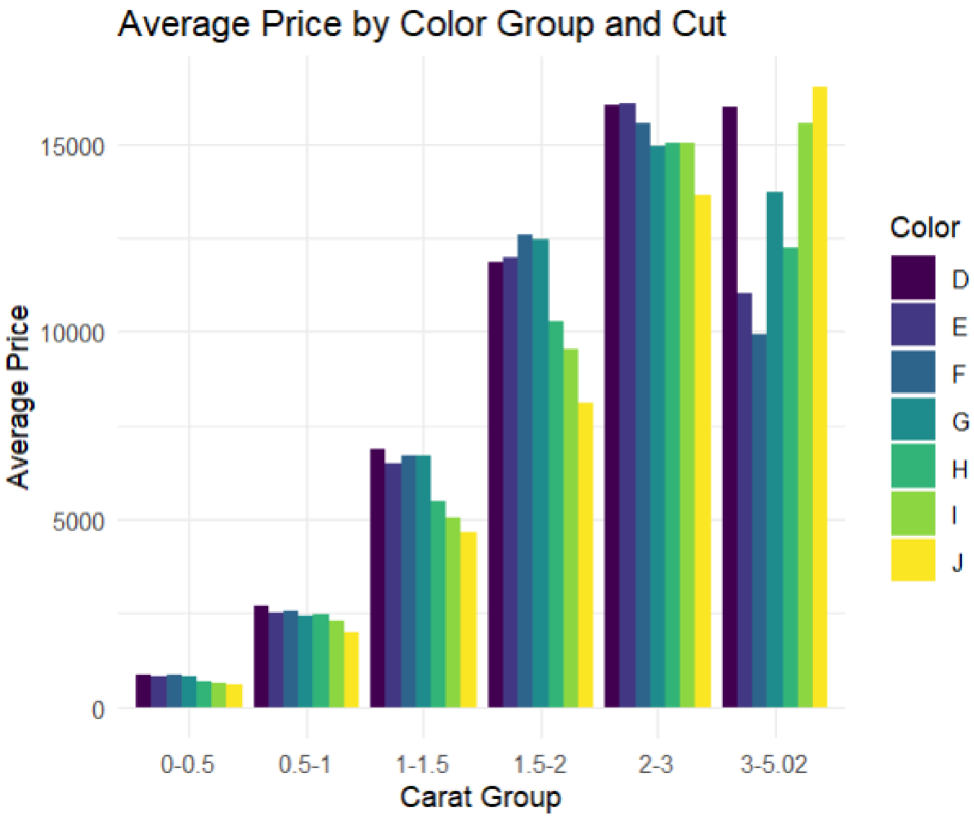


Figure 2.44: Clustered bar chart of carat range split by color by average price

As was with clarity, figure 2.44 shows that the better colored diamonds are on average, more expensive at the same weight. The observations from the final range may be skewed, as was previously said, the sample is only on 40 diamonds, which is not sufficient enough.

2.5 - Scatter Plot

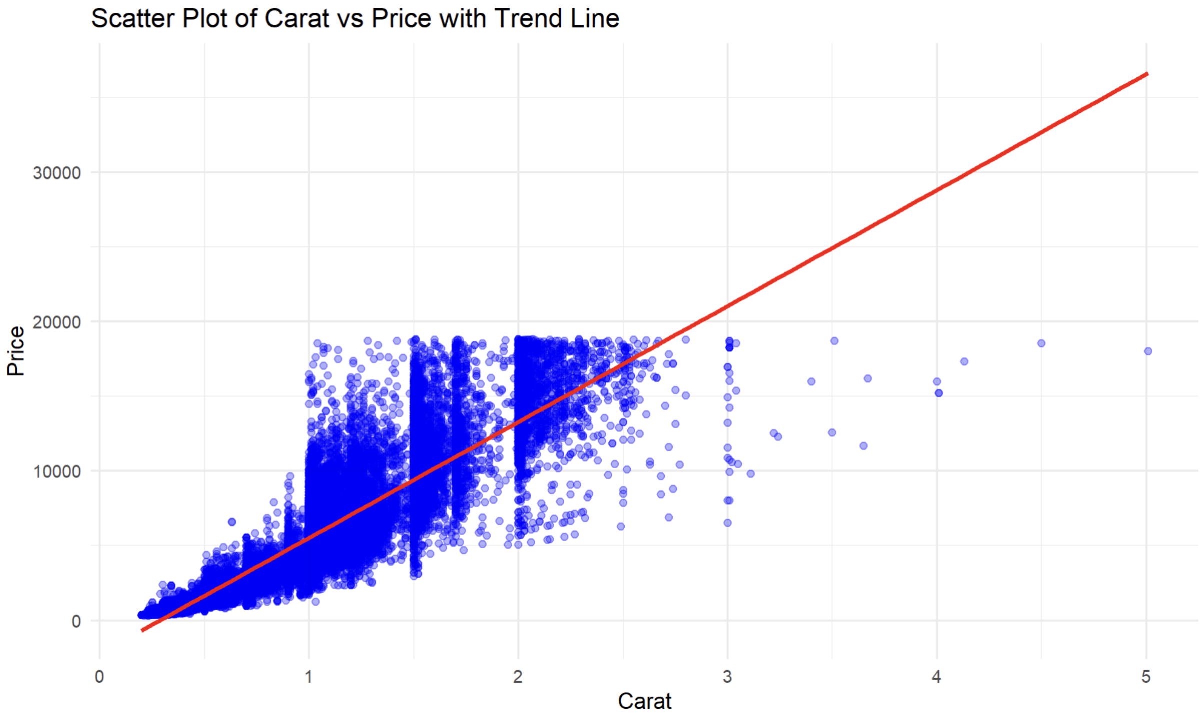


Figure 2.51: Scatter plot of carat by price

The trend line in figure 2.51 shows that there seems to be a linear relationship between Carat and Price. The scatter plot shows that as carat increases, so does the price, but it can be seen that at 1 carat, the price does shoot up for certain diamonds, though they are a small group.

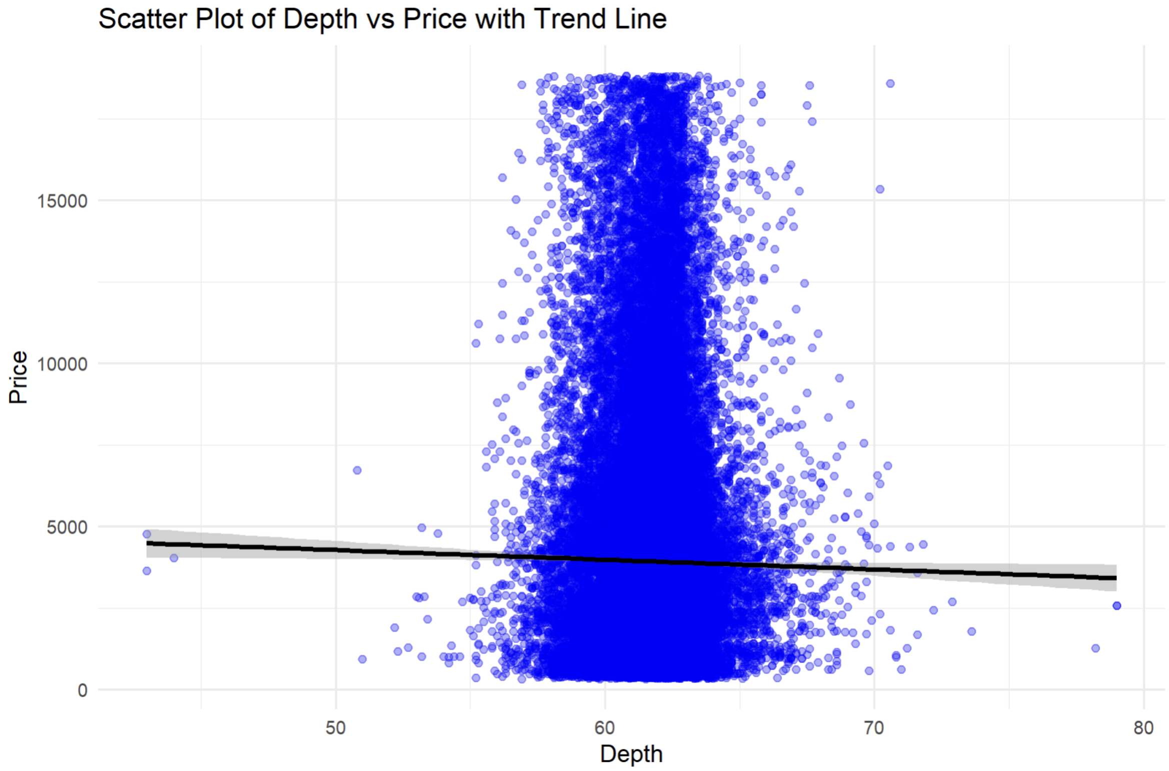


Figure 2.52: Scatter plot of depth by price

From the trend line of figure 2.52, we can see that there is a low correlation between depth and price, since the price almost stays constant, whatever the depth.

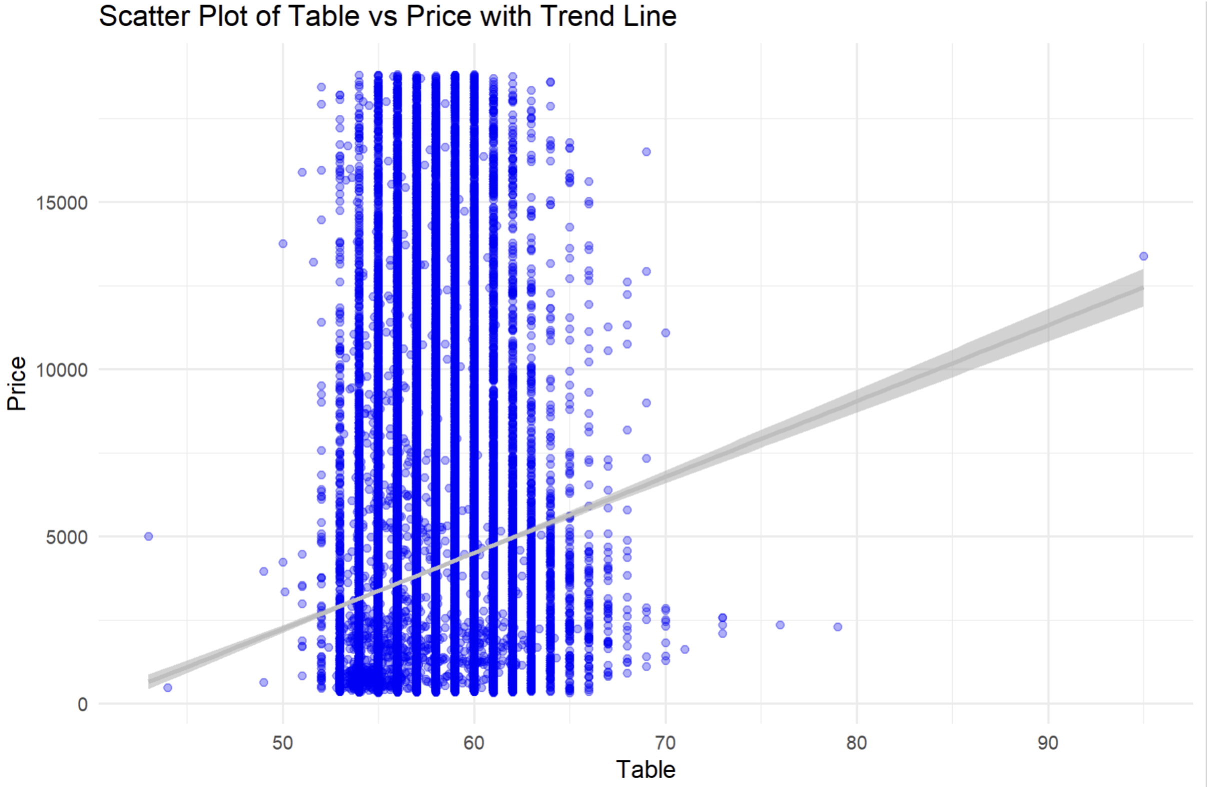


Figure 2.53: Scatter plot of table by price

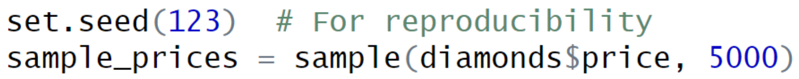
From figure 2.53, we identify that table values tend to cluster between 50% and 70%. Also, the trend line suggests a weak positive correlation between table and price, suggesting that, as table percentage increases, price increases too.

3.0 - Parametric / Non-Parametric Tests

3.1 - Shapiro-Wilk test

We first start by performing a normality test on our dependent variable, price. In this case we'll be using the **Shapiro-Wilk Test**, but since our dataset has over 50,000 samples, we selected a random sample of 5,000 diamonds, as this test is sensitive to large sample sizes and may give inaccurate results.

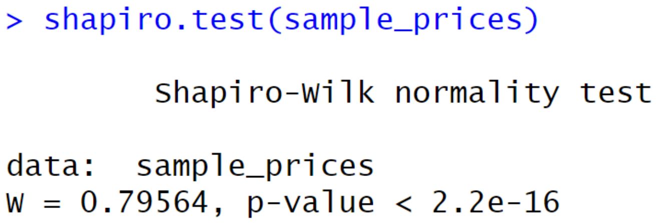
The subset was created with the following code:



```
set.seed(123)  # For reproducibility
sample_prices = sample(diamonds$price, 5000)
```

Hypotheses:
H0: Price follows a normal distribution
H1: Price does not follow a normal distribution

The test was then performed, obtaining the following results:

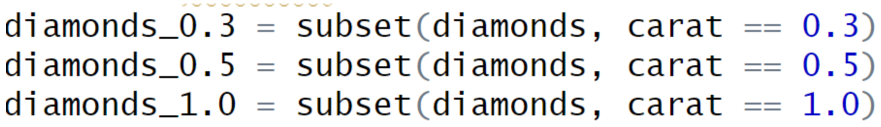


```
> shapiro.test(sample_prices)

        Shapiro-Wilk normality test

data:  sample_prices
W = 0.79564, p-value < 2.2e-16
```

Since the p-value is significantly less than 0.05, we reject H0. Therefore, the price does not follow a normal distribution, which is expected after viewing the graphs obtained during EDA.
This means that we'll be using non-parametric tests below.

3.2 - Wilcoxon signed rank test (Non-parametric alternative of One-Sample T-Test)

Here, we split our dataset into 3 subsets: one for 0.3 carat diamonds, one for 0.5 carat diamonds, and one for 1 carat diamonds. The subsets were created with the following code:



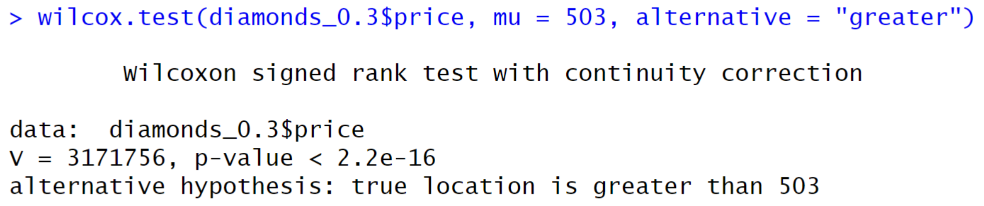
```
diamonds_0.3 = subset(diamonds, carat == 0.3)
diamonds_0.5 = subset(diamonds, carat == 0.5)
diamonds_1.0 = subset(diamonds, carat == 1.0)
```

Our goal was to identify if the median price of each of these subsets were equal to or greater than the current medians of diamonds today, obtained from the following website: Diamond Prices. Particularly, the mean price was $503 for 0.3 carat diamonds, $1042 for 0.5 carat diamonds, and $3678 for 1 carat diamonds (at the time this test was made).
?I think data should be median, whereas data obtained from the site is mean?

For this, we used the **Wilcoxon Signed Rank 1-Tailed Test**. We set hypotheses and obtained the following results:

3.21 - 0.3 carat diamonds

Hypotheses:
H0: The median price of 0.3 carat diamonds is less than or equal to $503
H1: The median price of 0.3 carat diamonds is greater than $503

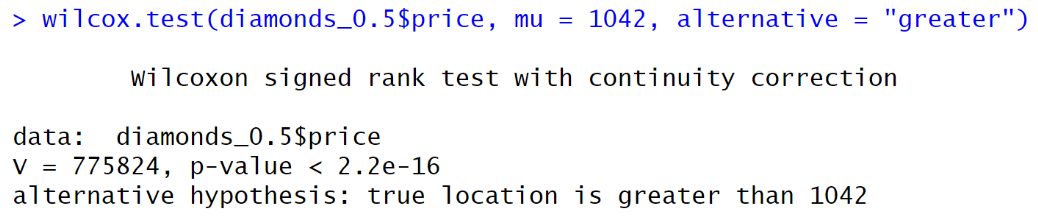


```
> wilcox.test(diamonds_0.3$price, mu = 503, alternative = "greater")

        Wilcoxon signed rank test with continuity correction

data:  diamonds_0.3$price
V = 3171756, p-value < 2.2e-16
alternative hypothesis: true location is greater than 503
```

Since the p-value is significantly less than 0.05, we reject H0. Therefore, the median price of 0.3 carat diamonds is much greater $503.

3.22 - 0.5 carat diamonds

Hypotheses:
H0: The median price of 0.5 carat diamonds is less than or equal to $1042
H1: The median price of 0.5 carat diamonds is greater than $1042

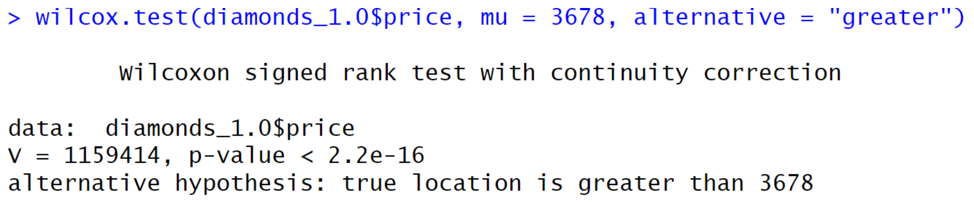


```
> wilcox.test(diamonds_0.5$price, mu = 1042, alternative = "greater")

        Wilcoxon signed rank test with continuity correction

data:  diamonds_0.5$price
V = 775824, p-value < 2.2e-16
alternative hypothesis: true location is greater than 1042
```

Since the p-value is once again significantly less than 0.05, we reject the null hypothesis. Thus, the median price of 0.5 carat diamonds is much greater than $1042.

3.23 - 1 carat diamonds

Hypotheses:
H0: The median price of 1 carat diamonds is less than or equal to $3678
H1: The median price of 1 carat diamonds is greater than $3678



```
> wilcox.test(diamonds_1.0$price, mu = 3678, alternative = "greater")

	Wilcoxon signed rank test with continuity correction

data:  diamonds_1.0$price
V = 1159414, p-value < 2.2e-16
alternative hypothesis: true location is greater than 3678
```

Once again, the p-value is significantly smaller than 0.05, so we reject the null hypothesis and accept the alternative. Thus, the median price of 1 carat diamonds is much higher than $3678.

After obtaining all these results, it seems that the median price of diamonds, irrespective of carat, was much higher in 1998 than they are now, indicating that they must have lost value over the past ~25 years.

I say 1998, as this data was adapted from Fred Cuellar's "Diamonds and Diamond Grading" book which was published in 1998, where the prices recorded were the current diamond prices at the time.

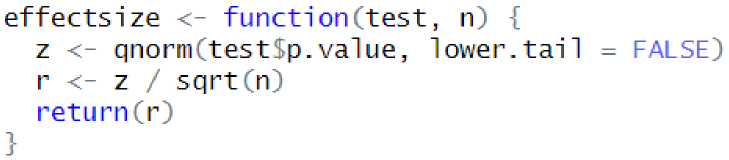
3.24 - Effect sizes

Note on effect sizes:
Since our data does not follow a normal distribution, we had to manually calculate the effect size using this formula:

$$r = \left|\frac{Z}{\sqrt{N}}\right|$$

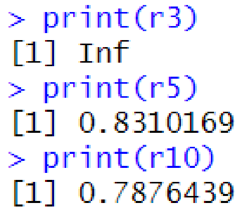
, where Z is the absolute standardized test statistic, and N is the number of observations.

To obtain the effect size of each test, the following function was used:



```
effectsize <- function(test, n) {
  z <- qnorm(test$p.value, lower.tail = FALSE)
  r <- z / sqrt(n)
  return(r)
}
```

The results when using this function on each set were:



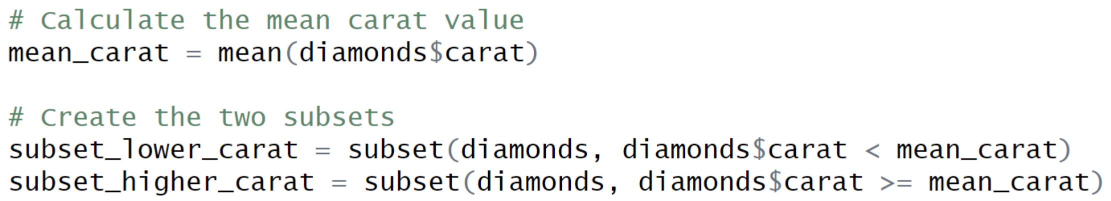
```
> print(r3)
[1] Inf
> print(r5)
[1] 0.8310169
> print(r10)
[1] 0.7876439
```

From these, it can be concluded that there is a very large difference between the claimed mean and the actual mean, since all of the values are greater than 0.5. In the case of r3, the value is infinity because its p-value is very close to 0, so qnorm returns 0, which divided by sqrt(n), returns infinity. So there is a very large difference, but the exact value cannot be calculated using r = Z/root(n).

3.3 - Mann-Whitney U Test (Non-parametric alternative of Independent Samples T-test)

Here, we split our dataset in two, one subset containing diamonds with a carat value less than the mean, and another subset containing diamonds with a carat value equal to or greater than the mean.

The mean of carat was found, and two subsets were created as follows:



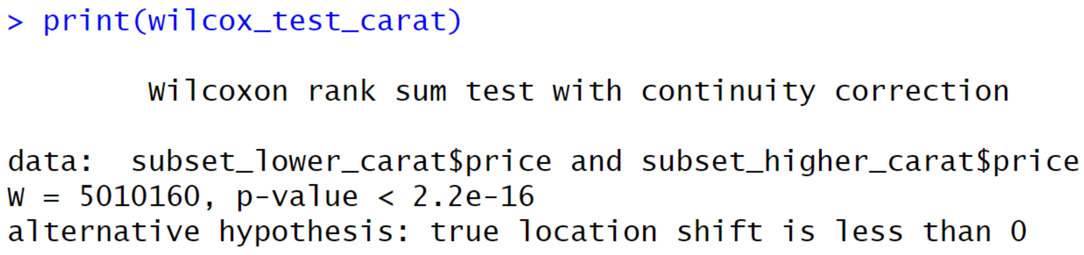
```
# Calculate the mean carat value
mean_carat = mean(diamonds$carat)

# Create the two subsets
subset_lower_carat = subset(diamonds, diamonds$carat < mean_carat)
subset_higher_carat = subset(diamonds, diamonds$carat >= mean_carat)
```

Our aim here was to identify which of the 2 samples has the highest mean price.

3.31 - 2-Sample non-parametric test

Hypotheses:
H0: The price distribution for diamonds with carat < mean(carat) is greater than or equal to the price distribution for diamonds with carat ≥ mean(carat)
H1: The price distribution for carat < mean(carat) is less than that of diamonds with carat ≥ mean(carat)

We then performed the **Mann-Whitney 1-tailed test** and obtained the following results:

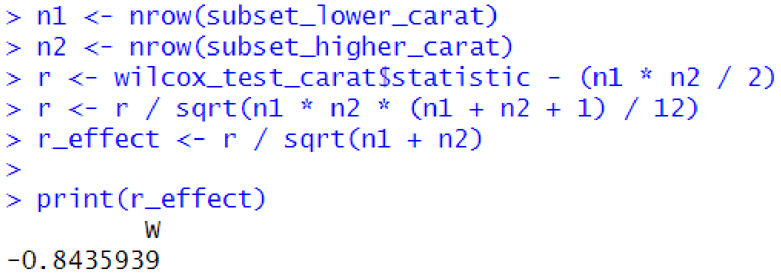


```
> print(wilcox_test_carat)

        Wilcoxon rank sum test with continuity correction

data:  subset_lower_carat$price and subset_higher_carat$price
W = 5010160, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

Since the p-value is 0.05, we reject the null hypothesis and accept the alternative, thus concluding that diamonds with carat ≥ mean(carat) have a higher mean price than carat < mean(carat). This was expected, after analysing the graphs obtained in EDA.

3.32 - Effect size

We then obtain the effect size:



```
> n1 <- nrow(subset_lower_carat)
> n2 <- nrow(subset_higher_carat)
> r <- wilcox_test_carat$statistic - (n1 * n2 / 2)
> r <- r / sqrt(n1 * n2 * (n1 + n2 + 1) / 12)
> r_effect <- r / sqrt(n1 + n2)
> 
> print(r_effect)
         W 
-0.8435939 
```

The function varied from the previous one as the output using the previous formula was infinity, so to get an exact value, a variation was used.

From the previous test, we know that the higher-carat diamonds have the greater prices, so the absolute value of the effect size is greater than 0.8, which means that there is a very large difference between the two groups.

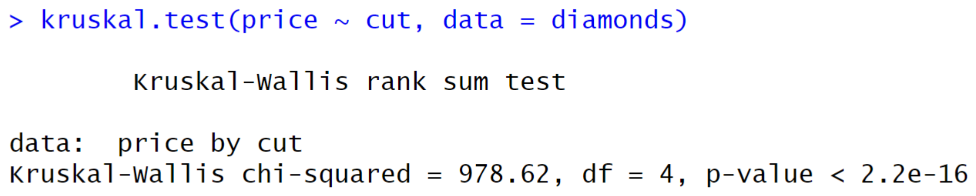
3.4 - Kruskal-Wallis Test on Price (x) and Cut (y) (Non-parametric alternative of One-Way Anova)

Finally, we wanted to find out whether the median price of diamonds with different cuts is the same or not. For this, the **Kruskal-Wallis test** was used, taking the y value as the price, and x value as the cut.

3.41 - Non-parametric Anova

Hypotheses:
H0: The median price is the same across all cut categories
H1: At least one cut category has a significantly different median price

The test was then performed and the following results were obtained:

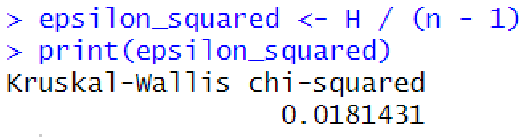


```
> kruskal.test(price ~ cut, data = diamonds)

        Kruskal-Wallis rank sum test

data:  price by cut
Kruskal-Wallis chi-squared = 978.62, df = 4, p-value < 2.2e-16
```

Since the p-value is less than 0.05, we reject H0 and accept the alternative hypothesis. We can thus conclude that, no, median prices are not the same across all cut categories, which is expected after analysing the boxplot of price by cut in EDA.

3.42 - Effect size

We then obtained the effect size using **epsilon squared**, as it less biased than **eta-squared**:



```
> epsilon_squared <- H / (n - 1)
> print(epsilon_squared)
Kruskal-Wallis chi-squared
                 0.0181431
```

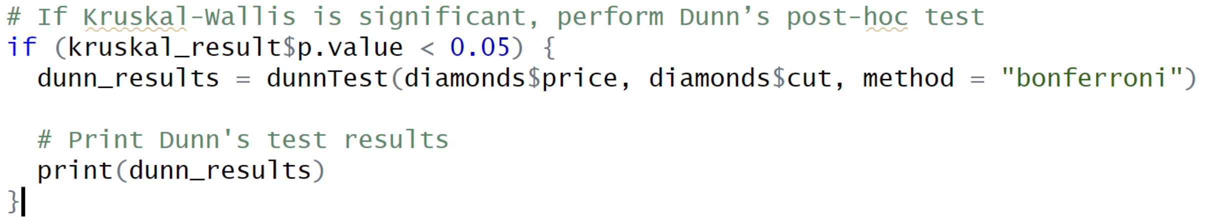
,where H is the Kruskal-Wallis test statistic, and n is the number of observations.

ε^2 is greater than 0.01, so there is a significant, but small effect size. From this we can also conclude that only 1.8% of the variance in price can be explained by the difference in cut.

Due to this, we then conducted the **Dunn Test** to identify which specific cuts have different medians. For this, we used the *Bonferroni* method.

3.43 - Post-hoc test

This was done with the following code:



```
# If Kruskal-Wallis is significant, perform Dunn’s post-hoc test
if (kruskal_result$p.value < 0.05) {
  dunn_results = dunnTest(diamonds$price, diamonds$cut, method = "bonferroni")

  # Print Dunn's test results
  print(dunn_results)
}
```

After running the test, we obtained the following results:

#	Comparison	Z	P.unadj	P.adj
1	Fair - Good	8.904490	5.363185e-19	5.363185e-18
2	Fair - Ideal	17.631040	1.423243e-69	1.423243e-68
3	Good - Ideal	12.628560	1.469412e-36	1.469412e-35
4	Fair - Premium	5.698925	1.205650e-08	1.205650e-07
5	Good - Premium	-6.356191	2.068175e-10	2.068175e-09
6	Ideal - Premium	-28.009093	1.259098e-172	1.259098e-171
7	Fair - Very Good	11.426910	3.068331e-30	3.068331e-29
8	Good - Very Good	2.800555	5.101482e-03	5.101482e-02
9	Ideal - Very Good	-13.405493	5.614897e-41	5.614897e-40
10	Premium - Very Good	12.284149	1.101981e-34	1.101981e-33

Table 3.431: Table showing the results from the Dunn test

From the results of the **Kruskal-Wallis test**, we can conclude that each cut group has a significant difference between each other. Good and Very Good have the least difference between them, as the P-adj value (0.051) is slightly higher than our alpha (0.05). The cut with the most significant differences is ‘Ideal’, as the comparisons with it have the largest absolute values of Z.

4.0 - Statistical Modelling

4.1 - Multiple Linear Regression

Our original MLR Model, which contains all covariate variables:

Yi is our dependent/response variable: **Price**

Price = μ + B1carat + B2depth + B3table + B4x + B5y + B6z + ε

In order to fit a MLR model, the analysed data needs to satisfy the following assumptions:

1. Response variable and explanatory variables are covariates.
2. A linear relationship exists between the dependent variable and each of the independent variables.
3. There must be no multicollinearity - presence of multicollinearity is detected by inspecting a number of multicollinearity diagnostics.
4. Residuals should be independent of each other.
5. There must be no influential outliers
6. Residuals must follow a normal distribution
7. Residuals must be homoscedastic

4.11 - Assumption 1

All the variables in the model are covariates

4.12 - Assumption 2

The scatterplots of the dependent variable with the independent variables were observed to check for a linear relationship. Note: since our dataset contains over 50,000 observations, a sample of 1000 were used for the scatter plots.

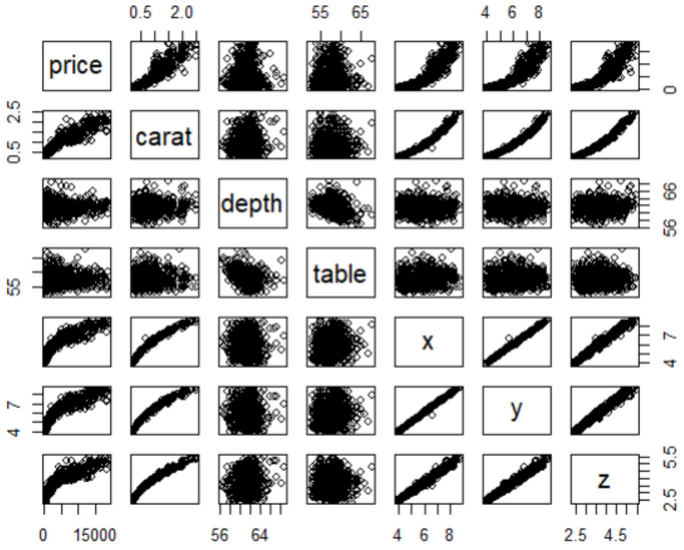


Figure 4.121: Scatter plot of each variable with all other variables

Spearman correlation analysis:

The **spearman correlation test** was done to check if the variables are linearly dependent on price. A correlation statistic of 0 means that there is no correlation between the two variables, but as the statistic tends to +1, or -1, the correlation between the two variables increases.

```
> rcorr(data, type="spearman")
```

	price	carat	depth	table	x	y	z
price	1.00	0.96	0.01	0.17	0.96	0.96	0.96
carat	0.96	1.00	0.03	0.19	0.97	0.97	0.97
depth	0.01	0.03	1.00	-0.25	-0.02	-0.03	0.10
table	0.17	0.19	-0.25	1.00	0.20	0.20	0.16
x	0.96	0.97	-0.02	0.20	1.00	1.00	0.99
y	0.96	0.97	-0.03	0.20	1.00	1.00	0.99
z	0.96	0.97	0.10	0.16	0.99	0.99	1.00

Table 4.122: Spearman test output, showing the correlation statistic between each variable

From this we conclude that carat, x, y, and z all have a positive linear relationship with price, whilst depth and table do not. Therefore, depth and table will be removed from our model.

Price = μ + B1carat + B2x + B3y + B4z + ε

4.13 - Assumption 3

4.131 - Spearman test

The **spearman test** was run again on the remaining variables to check if the independent variables have a linear relationship between themselves or not, i.e. checking for multicollinearity.

	price	**carat**	**x**	**y**	**z**
price	1.00	0.96	0.96	0.96	0.96
carat	0.96	1.00	0.97	0.97	0.97
x	0.96	0.97	1.00	1.00	0.99
y	0.96	0.97	1.00	1.00	0.99
z	0.96	0.97	0.99	0.99	1.00

Table 4.1311: Spearman test output, showing the correlation statistic between each variable

As can be seen from Table 4.1311, there is a lot of very high correlation between each variable.

Hypothesis:
H0: The 2 variables are not correlated, Pxy = 0
H1: The two variables are correlated, Pxy != 0

Note: Pxy is the correlation coefficient/statistic

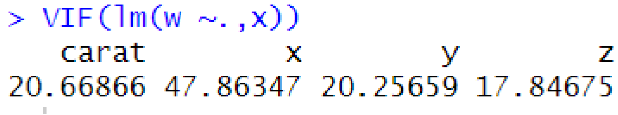
	price	**carat**	**x**	**y**	**z**
price		0	0	0	0
carat	0		0	0	0
x	0	0		0	0
y	0	0	0		0
z	0	0	0	0	

Table 4.1312: P-values for the spearman test

As you can see, all the values are under the significance level, 0.05, so H0 is rejected in all cases, and H1 is accepted, which means that multicollinearity exists between the variables.

4.132 - VIF test

Next, another check was made using the **VIFs**.



```
> VIF(lm(w ~.,x))
   carat        x        y        z 
20.66866 47.86347 20.25659 17.84675 
```

Variable	carat	x	y	z
VIF	20.66866	47.86347	20.25659	17.84675

Table 4.1321: Table of VIF for each variable

Since all values are greater than 5, this indicates that there is a significant amount of multicollinearity that is affecting the variance of the parameter estimates corresponding to these variables.

4.133 - Condition Indices

Finally, the **Condition Indices** were checked to see which variables suffer the most from collinearity.

#	**Eigenvalues**	**CI**	**Intercept**	**carat**	**x**	**y**
1	4.8477	1.0000	0.0002	0.0005	0.0000	0.0001
2	0.1474	5.7355	0.0084	0.0574	0.0000	0.0001
3	0.0025	44.4296	0.7235	0.6756	0.0033	0.1959
4	0.0018	51.3533	0.0029	0.0030	0.0019	0.4208
5	0.0006	89.9449	0.2651	0.2636	0.9947	0.3832

Table 4.1331: Table showing the Condition Indices for each variable

CI>30 indicates a lot of multicollinearity, and CI>80 indicates severe multicollinearity, and from Row 5, you can see that carat, x and y have high variance proportions, and Row 4 shows that y and z have high variance proportions.

After considering all the multicollinearity diagnostics we can conclude that carat, x, y, and z are all correlated with each other, so the model will contain only 1 independent variable.

Price = μ + B1var + ε, var = carat/x/y/z

4.14 - Assumption 4

To obtain the fitted model, the **summary of the model** and the ANOVA (analysis of variance) table were obtained.

```
> model1<-lm(y ~.,x)
> summary(model1)

> model2 <- aov(y ~.,x)
> summary(model2)
```

Coefficients:

	Estimate Std.	**Error**	**T value**	**Pr(>|t|)**
Intercept	-2256.36	13.06	-172.8	<2e-16
carat	7756.43	14.07	551.4	<2e-16

Table 4.141: Table showing the coefficient for the fitted model

Adjusted R-squared: 0.8493

Source	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
carat	**1**	**7.291e+11**	**7.291e+11**	**304051**	**< 2e-16**
Residuals	**53938**	**1.293e+11**	**2.398e+06**		

Table 4.142: ANOVA Table

The multiple R-squared statistic (0.8493) shows that 84.93% of the variance in price can be determined from the carat value. This test was also run with models including x, y, and z separately, but the model with carat had the highest Adjusted R-squared value, so it was deemed the best from the available models.

Price = μ + B1carat + ε

For the ANOVA table, we present the following hypothesis:
H0: Model with only a constant term (intercept only model) is a good fit for the data
H1: Model fitted (which includes covariates) fits better than the model with only the intercept term

As can be seen from Table 4.142, the p-value is less than 0.05, so we reject H0, and accept H1, and hence, carat is significant for our fitted model.

To find the estimated value for B0 and B1, we present the following hypothesis:

H0: Bi = 0
H1: Bi != 0

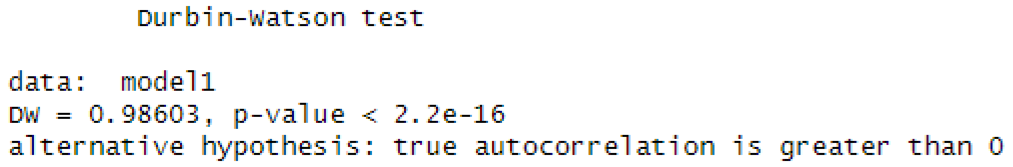
From the “Coefficients” table the p value is less than 0.05 for both the intercept and carat, so we reject the null hypothesis for both cases, and accept H1, which mean the fitted model will look like:

Price = -2256.36 + 7756.43carat

Though this model is not quite accurate, as from it, a 0.3carat diamond is only worth $70.57, which is quite low. This could mean that there are outliers affecting our model, or that carat is not enough to determine the price, there are other variables to consider, those being the categorical variables. Thus the fitted model from the ANCOVA will be more accurate.

Finally, the **Durbin-Watson test** was run to check if the residuals are independent of each other:

H0: The residuals are independent
H1: The residuals are not independent

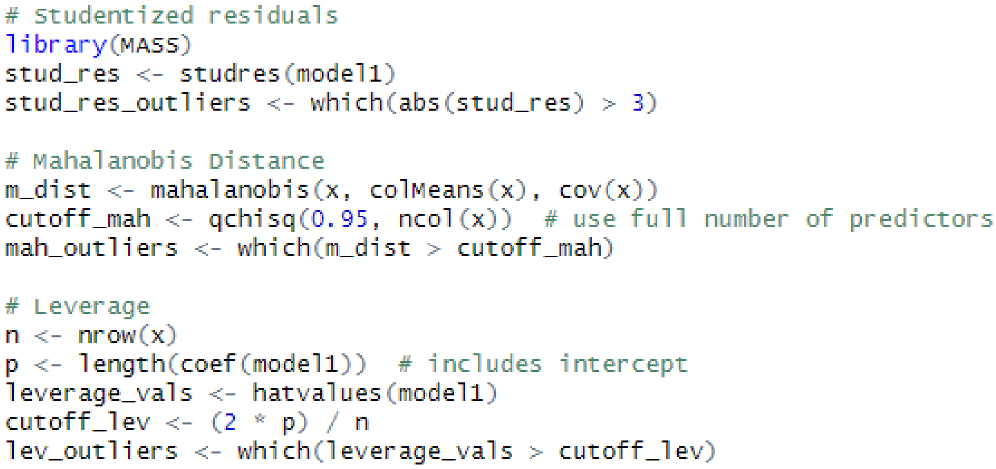


```
        Durbin-Watson test

data:  model1
DW = 0.98603, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Since the p-value is less than 0.05 we reject the null hypothesis and accept H1, and also, the DW statistic is out of the 1.5-2.5 range, which is a cause for concern, which shows that Assumption 4 is not satisfied.

4.15 - Assumption 5

Next, 4 tests were made to check for potential outliers: **Studentized residuals**, **Mahalanobis Distance**, **Leverage test**, and **Cook's distance**. Note: Cook's distance returned no outliers.



```
# Studentized residuals
library(MASS)
stud_res <- studres(model1)
stud_res_outliers <- which(abs(stud_res) > 3)

# Mahalanobis Distance
m_dist <- mahalanobis(x, colMeans(x), cov(x))
cutoff_mah <- qchisq(0.95, ncol(x))  # use full number of predictors
mah_outliers <- which(m_dist > cutoff_mah)

# Leverage
n <- nrow(x)
p <- length(coef(model1))  # includes intercept
leverage_vals <- hatvalues(model1)
cutoff_lev <- (2 * p) / n
lev_outliers <- which(leverage_vals > cutoff_lev)
```

After the tests, 2626 diamonds were flagged as outliers by at least two of the tests, so the regression analysis will be run again, and if the parameter estimates change only slightly, the outliers can be retained.

Coefficients:

	Estimate Std.	Error	T value	Pr(>|t|)
Intercept	-2127.98	13.24	-160.7	<2e-16
carat	7527.50	16.02	469.9	<2e-16

Table 4.151: Coefficients table for the updated model

Adjusted R-squared: 0.8114

The summary now indicates that the fitted model shows that 81.14% of the variance in price can be determined from the carat value, as well as the null hypothesis is rejected again for both cases B0, B1. The parameter estimates are considerably different, so those diamonds are influential points and will not be included in the analysis. Thus, the new fitted mode is:

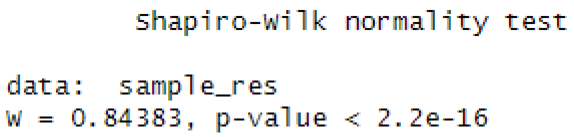
Price = -2127.98 + 7527.5carat

Without the outliers, this model shows that a 0.3carat diamond is worth $130.27, which is also too low, so as said before, the ANCOVA model will be a more accurate model.

4.16 - Assumption 6

The **Shapiro-Wilk normality test** was run on the residuals:

Hypothesis:
H0: Residuals follow a normal distribution
H1: Residuals do not follow a normal distribution



```
	Shapiro-Wilk normality test

data:  sample_res
W = 0.84383, p-value < 2.2e-16
```

Since the p-value is less than 0.05, H0 is rejected and H1 is accepted, which means that Assumption 6 is not met.

To confirm this, a Histogram and Q-Q Plot were generated:

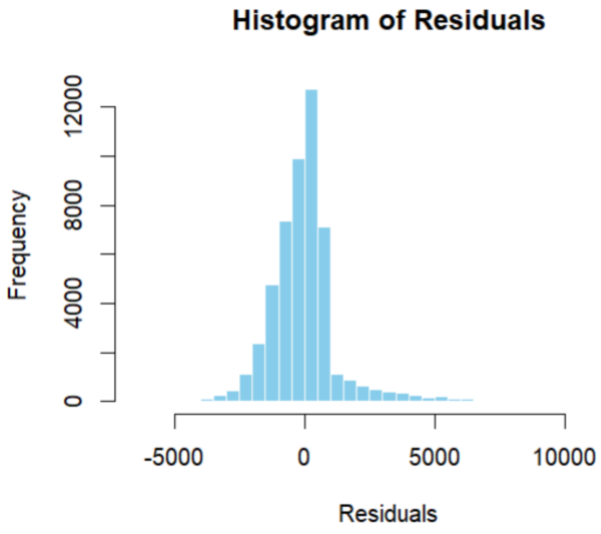


Figure 4.161: Histogram of residuals

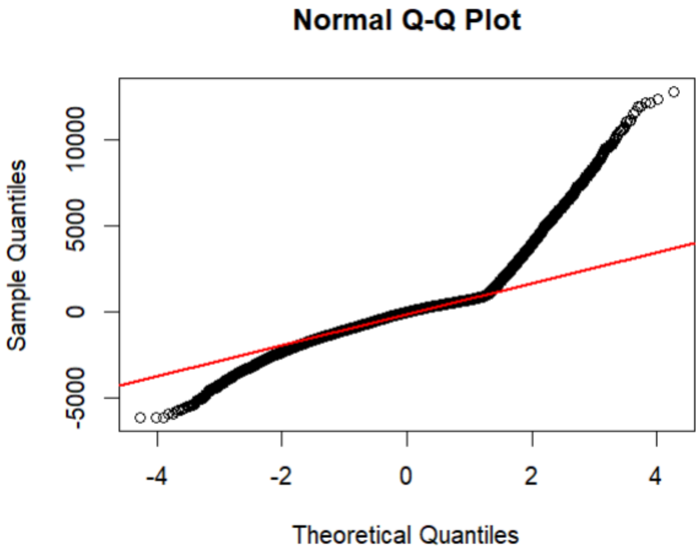


Figure 4.162: Q-Q plot

The histogram shows that the residuals aren't normally distributed as the graph is positively skewed, and this is confirmed by the Q-Q plot, which shows substantial deviation from the line, particularly in the tails.

4.17 - Assumption 7

A scatter plot of the residuals vs the fitted values was generated. If residuals are homoscedastic, one can see a constant variation in the y-axis.

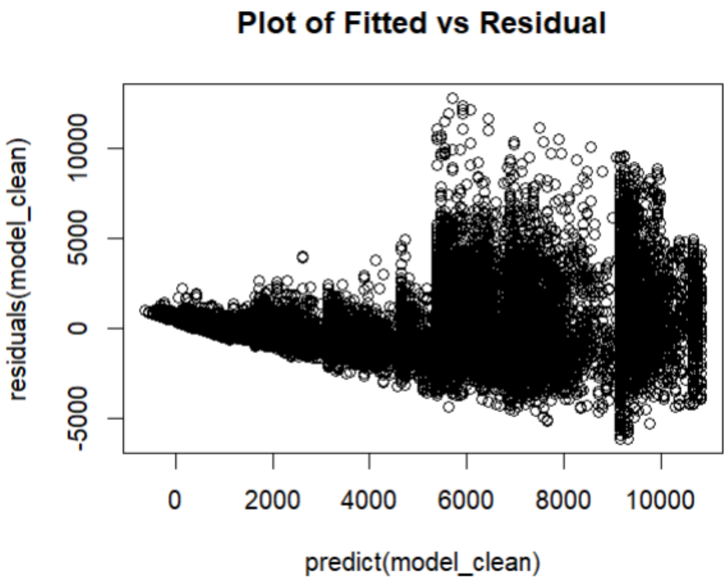


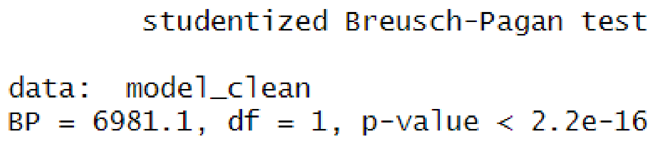
Figure 4.171: plot of fitted vs residual

As can be seen from the plot, there is not a constant variation in the y-axis, so the residuals are not homoscedastic.

The **Breusch Pagan** test was also conducted to be sure:

Hypothesis:

H0 : Residuals are homoscedastic
H1 : Residuals are heteroskedastic



```
        studentized Breusch-Pagan test

data:  model_clean
BP = 6981.1, df = 1, p-value < 2.2e-16
```

Since the p-value is less than 0.05, the null hypothesis is rejected, and the alternative is accepted, so Assumption 7 is also not satisfied.

4.2 - N-Way Anova

The 3-Way Anova model is made up of the following categorical variables: Cut, Color, Clarity.

$$Yijkl = \mu + \alpha i + \beta j + \gamma k + (\alpha\beta)ij + (\alpha\gamma)ik + (\beta\gamma)jk + (\alpha\beta\gamma)ijk + \varepsilon ijkl$$

Where:
Yijkl: the response variable (e.g., price)

μ: overall mean

αi: effect of the ith level of cut (5 levels)

βj: effect of the jth level of clarity (7 levels)

γk: effect of the kth level of color (7 levels)

$(\alpha\beta)ij$: interaction between cut and clarity

$(\alpha\gamma)ik$: interaction between cut and color

$(\beta\gamma)jk$: interaction between clarity and color

$(\alpha\beta\gamma)ijk$: three-way interaction between cut, clarity, and color

$\varepsilon ijkl$: random error term

4.21 - Generating the ANOVA table

The Anova table was generated using:

```
> model_anova <- aov(price ~ cut * color * clarity, data = diamonds)
> summary(model_anova)
```

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cut	4	1.104e+10	2.760e+09	189.755	< 2e-16
color	6	2.551e+10	4.251e+09	292.230	< 2e-16
clarity	7	2.000e+10	2.857e+09	196.381	< 2e-16
cut:color	24	1.835e+09	7.648e+07	5.257	8.01e-16
cut:clarity	28	2.425e+09	8.659e+07	5.952	< 2e-16

color:clarity	42	1.272e+10	3.027e+08	20.811	< 2e-16
cut:color:clarity	164	4.282e+09	2.611e+07	1.795	1.96e-09

Table 4.121: ANOVA Table

We present the following hypothesis
H0: The variable is not significant for our model
H1: The variable is significant for our model

As can be seen from the Pr(>F) column from Figure 4.121, the p-value for cut, color and clarity are all smaller than our significance level (0.05), so we reject H0, and accept H1 for all 3 variables. The same applies for 2-way and 3-way interactions between the variables, as all of their p-values are smaller than 0.05, so H1 is accepted in each case. Hence, our model stays the same.

$Yijkl = \mu + \alpha i + \beta j + \gamma k + (\alpha\beta)ij + (\alpha\gamma)ik + (\beta\gamma)jk + (\alpha\beta\gamma)ijk + \varepsilon ijkl$

4.22 - Checking normality for each categorical level

To check for normality, the following hypothesis was used:
H0: The categorical variable level follows a normal distribution
H1: The categorical variable level does not follow a normal distribution

Cut	Sample Size (n)	Shapiro-Wilk p-value
Fair	1,610	6.60×10^{-40}
Good	4,906	1.57×10^{-59}
Very Good	12,082 (sampled)	3.37×10^{-60}
Premium	13,791 (sampled)	1.77×10^{-58}
Ideal	21,551 (sampled)	1.29×10^{-66}

Figure 4.221: Normality table for Cut

Color	Sample Size (n)	Shapiro-Wilk p-value
D	6,775 (sampled)	3.00×10^{-66}
E	9,797 (sampled)	1.36×10^{-67}
F	9,542 (sampled)	6.02×10^{-63}
G	11,292 (sampled)	6.89×10^{-61}
H	8,304 (sampled)	1.37×10^{-57}
I	5,422 (sampled)	6.40×10^{-56}
J	2,808	6.11×10^{-41}

Figure 4.222: Normality table for Color

Clarity	Sample Size (n)	Shapiro-Wilk p-value
I1	741	4.45×10^{-25}
SI2	9,194 (sampled)	1.41×10^{-58}
SI1	13,065 (sampled)	2.90×10^{-60}
VS2	12,258 (sampled)	1.33×10^{-61}
VS1	8,171 (sampled)	3.27×10^{-62}
VVS2	5,066 (sampled)	4.97×10^{-68}
VVS1	3,655	3.26×10^{-68}
IF	1,790	5.13×10^{-54}

Figure 4.223: Normality table for Clarity

As can be seen from the p-values of Figures 4.221, 4.222, and 4.223, none of the categorical variable levels follow a normal distribution, as every p-value is less than the significance level, 0.05, so H0 is rejected in every case, and H1 is accepted instead.

4.23 - Check for homogeneity of variances for each categorical level

To check for homogeneity we present the following hypothesis:
H0: Variances are equal across groups
H1: Homogeneity violated

Source	Df	F value	Pr(>F)
group	4	123.6	$< 2.2 \times 10^{-16}$

Figure 4.231: Levene's Test for Homogeneity of Variance for Cut

Source	Df	F value	Pr(>F)
group	6	219.12	$< 2.2 \times 10^{-16}$

Figure 4.231: Levene's Test for Homogeneity of Variance for Color

Source	Df	F value	Pr(>F)
group	7	77.809	$< 2.2 \times 10^{-16}$

Figure 4.231: Levene's Test for Homogeneity of Variance for Clarity

As can be seen from the p-values of Figures 4.231, 4.232, and 4.233, homogeneity is violated for each variable, as every p-value is less than the significance level, 0.05, so H0 is rejected in every case, and H1 is accepted instead.

(Please note that the fitted model was not created since due to the many categorical levels, the formula would have included over 300 parameters, so to simplify we found which 1-way, 2-way and 3-way parameters are significant for the model)

4.3 - ANCOVA

The ANCOVA model consists of a combination of Carat, Cut, Color, and Clarity.

Yi is our dependent/response variable: **Price**

Price = μ + Bcarat + C1cutGood + C2cutVeryGood + C3cutPremium + C4cutIdeal +D1colorE + D2colorF + D3colorG + D4colorH + D5colorI + D6colorJ + E1claritySI2 + E2claritySI1 + E3clarityVS2 + E4clarityVS1 + E5clarityVVS2 + E6clarityVVS1 + E7clarityIF + ε

(Please note that the interactions between categorical variables were not included in the model due to the sheer number of terms produced, so the model given is a simplified version)

To create this model, the following assumptions must be satisfied:

1. Response variable and explanatory variables are covariates.
2. A linear relationship exists between the dependent variable and each of the independent variables.
3. There must be no multicollinearity - presence of multicollinearity is detected by inspecting a number of multicollinearity diagnostics.
4. Residuals should be independent of each other.
5. There must be no influential outliers
6. Residuals must follow a normal distribution
7. Residuals must be homoscedastic

4.31 - Assumption 1

The response variable is covariate. Carat is covariate, and dummy variables were created for Cut, Color, and Clarity to make them covariates.

4.32 - Assumption 2

The **spearman correlation test** was done to check if the variables are linearly dependent on price. A correlation statistic of 0 means that there is no correlation between the two variables, but as the statistic tends to +1, or -1, the correlation between the two variables increases.

	price	**cut**	**color**	**clarity**	**carat**
price	1.00	-0.09	0.15	-0.21	0.96
cut	-0.09	1.00	-0.02	0.19	-0.14
color	0.15	-0.02	1.00	0.03	0.25

clarity	-0.21	0.19	0.03	1.00	-0.37
carat	0.96	-0.14	0.25	-0.37	1.00

Figure 4.321: Spearman test output, showing the correlation statistic between each variable

From this table, we conclude that price has a positive correlation with color and carat, and a negative correlation with clarity. However, cut doesn't seem to have any correlation with price, so it is removed.

Price = μ + Bcarat +D1colorE + D2colorF + D3colorG + D4colorH + D5colorI + D6colorJ +
E1claritySI2 + E2claritySI1 + E3clarityVS2 + E4clarityVS1 + E5clarityVVS2 +
E6clarityVVS1 + E7clarityIF + ε

4.33 - Assumption 3

The **spearman test** was run again on *cut, color, clarity*, *carat* to check if the independent variables have a linear relationship between themselves or not, i.e. checking for multicollinearity.

Hypothesis:
H0: The four variables are not correlated, Pxy = 0
H1: The four variables are correlated, Pxy != 0

Note: Pxy is the correlation coefficient/statistic

	price	**cut**	**color**	**clarity**	**carat**
price	1.00	-0.09	0.15	-0.21	0.96
cut	-0.09	1.00	-0.02	0.19	-0.14
color	0.15	-0.02	1.00	0.03	0.25
clarity	-0.21	0.19	0.03	1.00	-0.37
carat	0.96	-0.14	0.25	-0.37	1.00

Figure 4.331: Spearman test output, showing the correlation statistic between each variable

The correlation statistics between color and carat, and between clarity and carat are not large enough to remove one of them.

Price = μ + Bcarat +D1colorE + D2colorF + D3colorG + D4colorH + D5colorI + D6colorJ + E1claritySI2 + E2claritySI1 + E3clarityVS2 + E4clarityVS1 + E5clarityVVS2 + E6clarityVVS1 + E7clarityIF + ε

Hypothesis:
H0: The 2 variables are not correlated, Pxy = 0
H1: The two variables are correlated, Pxy != 0

	price	**cut**	**color**	**clarity**	**carat**
price		0	0	0	0
cut	0		0	0	0
color	0	0		0	0
clarity	0	0	0		0
carat	0	0	0	0	

Figure 4.332: P-values for the spearman test

According to the results, all the values are under the significance level, 0.05, so we reject H0 in all cases, thus accepting H1, which means that multicollinearity exists between the variables.

VIF Test

Next, another check was made using the **VIFs**.

Variable	color	clarity	carat
VIF	1.180269	1.352806	25.391630

Figure 4.332: Table of VIF for each variable

Since the value of carat is greater than 5, this indicates that there is a significant amount of multicollinearity affecting the variance of the parameter estimates. However, since the value of color and clarity are less than 5, there isn't a significant amount of multicollinearity.

Condition Indices

Finally, the **Condition Indices** were checked to see which variables suffer the most from collinearity.

Coefficients:

	CI
1	1.000000
2	1.286185
3	1.405428
4	1.464014
5	1.545684
6	1.611681
7	1.677207
8	1.715387
9	1.89408
10	2.021531
11	2.314869

12	2.553492
13	2.886158
14	3.915955
15	5.844831

Figure 4.333: Table showing the Condition Indices for each variable

Since almost all values are less than 5, this shows that there is a high level of multicollinearity. The multicollinearity mainly comes from carat, but we are going to refrain from removing any variables to avoid losing the point of having the Ancova model.

4.34 - Assumption 4

To obtain the fitted model, the **summary of the model** was obtained.

Coefficients:

	Estimate Std.	**Pr(>|t|)**
Intercept	-3506.64	< 2e-16
carat	8856.23	< 2e-16
colorE	-1916.89	< 2e-16
colorF	-629.34	< 2e-16
colorG	-181.84	< 2e-16
colorH	18.58	0.184
colorI	-86.19	7.25e-11
colorJ	-56.21	2.96e-06
claritySI2	4417.66	< 2e-16
claritySI1	-1939.31	< 2e-16
clarityVS2	1009.46	< 2e-16
clarityVS1	-410.68	< 2e-16
clarityVVS2	242.64	< 2e-16
clarityVVS1	-12.13	0.393
clarityIF	121.32	< 2e-16

Figure 4.341: Table showing the coefficient for the fitted model

Adjusted R-squared: 0.9139

Source	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
carat	1	7.291e+11	7.291e+11	532320	< 2e-16
color	6	1.256e+10	2.093e+09	1528	< 2e-16
clarity	7	4.292e+10	6.132e+09	4477	< 2e-16
Residuals	53925	7.386e+10	1.370e+06		

Figure 4.342: ANOVA Table

The multiple R-squared statistic (0.9139) shows that 91.39% of the variance in price can be determined from the carat, color, and clarity value.

For the coefficients table, the following hypothesis is presented:
H0: The categorical level is not significant for the fitted model
H1: The categorical level is significant for the fitted model

So for each categorical level whose p-value is less than 0.05, its coefficient is placed in the fitted model, the others, like clarityVVS1 (p-value: 0.393), are omitted from the fitted model.

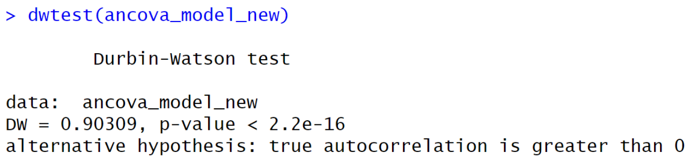
$\text{Price} = -3506.64 + 8856.23 \cdot \text{carat} - 1916.89 \cdot \text{colorE} - 629.34 \cdot \text{colorF} - 181.84 \cdot \text{colorG}$
$-86.19 \cdot \text{colorI} - 56.21 \cdot \text{colorJ} + 4417.66 \cdot \text{claritySI2} - 1939.31 \cdot \text{claritySI1} +$
$1009.46 \cdot \text{clarityVS2} - 410.68 \cdot \text{clarityVS1} + 242.64 \cdot \text{clarityVVS2} + 121.32 \cdot \text{clarityIF}$

For the ANOVA table, we present the following hypothesis:
H0: Model with only a constant term (intercept only model) is a good fit for the data
H1: Model fitted (which includes covariates) fits better than the model with only the intercept term

Since all the p-values are less than 0.05, we reject H0. Thus we accept H1.

Finally, the **Durbin-Watson test** was run to check if the residuals are independent of each other:

H0: The residuals are independent
H1: The residuals are not independent



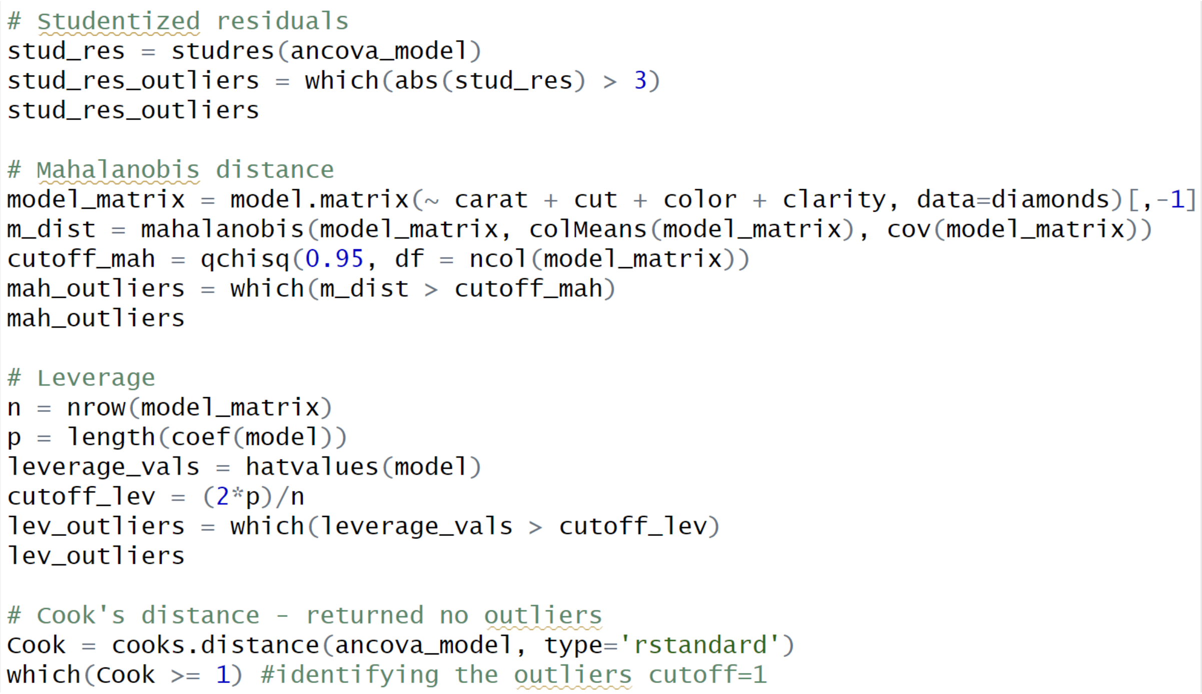
```
> dwtest(ancova_model_new)

	Durbin-Watson test

data:  ancova_model_new
DW = 0.90309, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Since the p-value is less than 0.05 we reject the null hypothesis and accept H1, and also, the DW statistic is out of the 1.5-2.5 range, which is a cause for concern, which shows that Assumption 4 is not satisfied.

4.35 - Assumption 5

Next, these 4 tests were performed to check for potential outliers: **Studentized residuals**, **Mahalanobis Distance**, **Leverage test**, and **Cook's distance**.

They were implemented as follows:



```
# Studentized residuals
stud_res = studres(ancova_model)
stud_res_outliers = which(abs(stud_res) > 3)
stud_res_outliers

# Mahalanobis distance
model_matrix = model.matrix(~ carat + cut + color + clarity, data=diamonds)[,-1]
m_dist = mahalanobis(model_matrix, colMeans(model_matrix), cov(model_matrix))
cutoff_mah = qchisq(0.95, df = ncol(model_matrix))
mah_outliers = which(m_dist > cutoff_mah)
mah_outliers

# Leverage
n = nrow(model_matrix)
p = length(coef(model))
leverage_vals = hatvalues(model)
cutoff_lev = (2*p)/n
lev_outliers = which(leverage_vals > cutoff_lev)
lev_outliers

# Cook's distance - returned no outliers
Cook = cooks.distance(ancova_model, type='rstandard')
which(Cook >= 1) #identifying the outliers cutoff=1
```

After the tests, 2895 diamonds were flagged as outliers by at least two of the tests, so the regression analysis will be run again, and if the parameter estimates change only slightly, the outliers can be retained.

Note: Cook's distance returned no outliers.

Note also that most of the outliers have clarity IF and I1.

Coefficients:

	Estimate Std.	**Pr(>|t|)**
Intercept		
carat	8944.49	< 2e-16
colorE	-175.87	< 2e-16
colorF	-280.69	< 2e-16
colorG	-448.37	< 2e-16
colorH	-931.57	< 2e-16

colorI	-1402.89	< 2e-16
colorJ	-2315.21	< 2e-16
claritySI1	2061.51	< 2e-16
clarityVS2	-584.45	< 2e-16
clarityVS1	78.74	4.11e-09
clarityVVS2	-91.64	4.40e-13
clarityVVS1	-102.43	< 2e-16

Figure 4.351: Coefficients table for the updated model

Adjusted R-Squared: 0.9231

The summary now indicates that the fitted model shows that 92.31% of the variance in price can be determined from the carat, color, and clarity value, as well as the null hypothesis is rejected again for all of the coefficients of the new fitted model. The parameter estimates are considerably different, so those diamonds are influential points and will not be included in the analysis. Thus, the new fitted mode is:

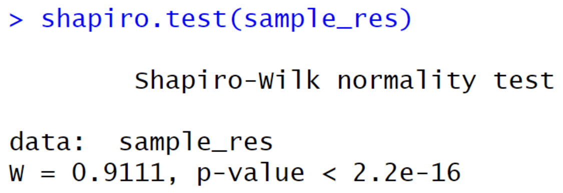
$$\text{Price} = -3506.64 + 8944.49 \cdot \text{carat} - 175.87 \cdot \text{colorE} - 280.69 \cdot \text{colorF} - 448.37 \cdot \text{colorG} - 931.57 \cdot \text{colorH} - 1402.89 \cdot \text{colorI} - 2315.21 \cdot \text{colorJ} + 2061.51 \cdot \text{claritySI1} - 584.45 \cdot \text{clarityVS2} + 78.74 \cdot \text{clarityVS1} - 91.64 \cdot \text{clarityVVS2} - 102.43 + \cdot \text{clarityVVS1}$$

Although this model is more accurate than the MLR mode, the adjusted R-Squared is higher, if a random diamond is chosen from the dataset, it will not return the same price +/- $300, as the model is based off 50,000+ diamonds, so it will not return highly accurate prices for each of them. Though this model can be used by someone to price diamonds if he didn't have an idea on how to do so.

4.36 - Assumption 6

The **Shapiro-Wilk normality test** was run on the residuals:

Hypothesis:
H0: Residuals follow a normal distribution
H1: Residuals do not follow a normal distribution



```
> shapiro.test(sample_res)

        Shapiro-Wilk normality test

data:  sample_res
W = 0.9111, p-value < 2.2e-16
```

Since the p-value is less than 0.05, H0 is rejected and H1 is accepted, which means that Assumption 6 is not met.

To confirm this, a Histogram and Q-Q Plot were generated:

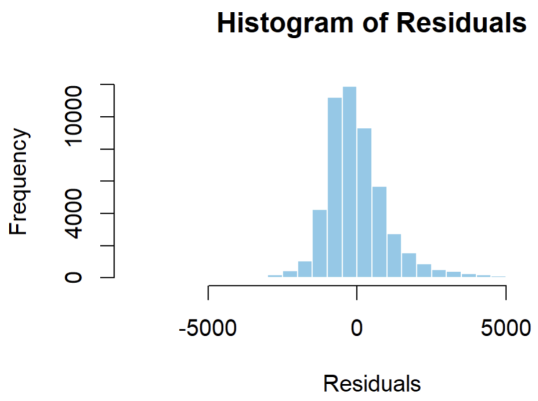


Figure 4.361: Histogram of residuals

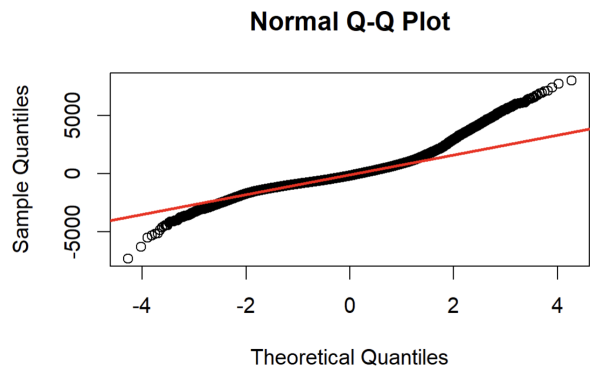


Figure 4.362: Q-Q plot

The histogram shows that the residuals aren’t normally distributed as the graph is positively skewed, and this is confirmed by the Q-Q plot, which shows substantial deviation from the line, particularly in the tails.

4.37 - Assumption 7

A scatter plot of the residuals vs the fitted values was generated. If residuals are homoscedastic, one can see a constant variation in the y-axis.

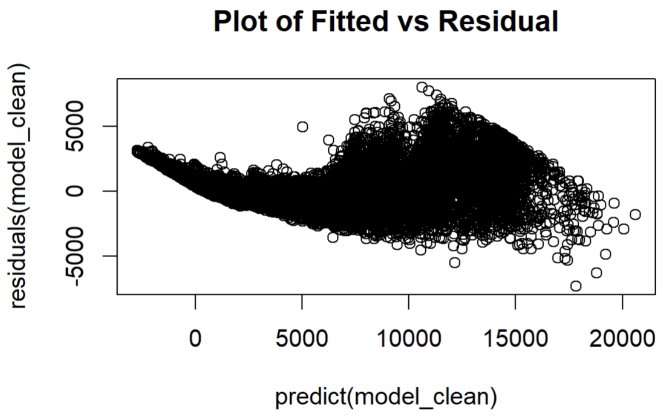
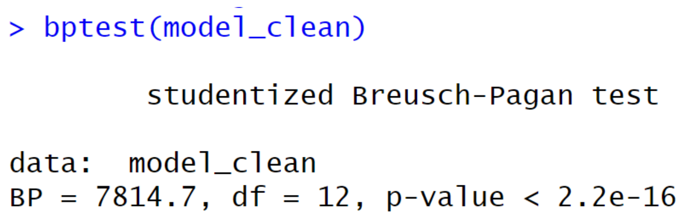


Figure 4.371: Plot of fitted vs residual

As can be seen from the plot, there is not a constant variation in the y-axis, so the residuals are not homoscedastic.

The **Breusch Pagan** test was also conducted to be sure:

Hypothesis:
H0 : Residuals are homoscedastic
H1 : Residuals are heteroskedastic



```
> bptest(model_clean)

        studentized Breusch-Pagan test

data:  model_clean
BP = 7814.7, df = 12, p-value < 2.2e-16
```

Since the p-value is less than 0.05, the null hypothesis is rejected, and the alternative is accepted, so Assumption 7 is also not satisfied.

5.0 - Conclusion

A lot can be concluded about diamonds from all the tests done and results obtained. The main points are:

- Carat it the most significant variable that affects the price of a diamond, but the price between same carat diamonds will vary, mostly by color and clarity, with the more ordinal colors and clarities being generally more expensive, as was shown in the clustered bar charts of Figures 2.44 and 2.42.
- Although higher ordinal color and clarity diamonds are more expensive for the same carat diamonds, the most expensive diamonds are generally from the lower ordinal colors and clarities. This is shown in the box-plots of Figures 2.22 and 2.23, where the median price lowered, as the ordinality became higher. Hence, lower ordinal color and clarity diamonds are found in bigger/ heavier diamonds, and since carat is the most significant variable, this makes these diamonds more expensive.
- The 2-tailed non-parametric t-test shows that heavier diamonds are more expensive, and the non-parametric ANOVA shows that, although cut is not an ordinal categorical variable, the cut level is significant to the price, with certain levels, like “Ideal”, being more significant than others.
- From the MLR, it can be shown that variables depth and table do not affect the price of diamonds significantly, and that carat is more significant than the diamonds dimensions, although the dimensions still are still significant for the price, mostly because the dimensions are correlated with the weight of a diamond.
- The MLR also showed that, although carat explains the majority of the variance in price, it is not sufficient enough to obtain the exact price, hence the ANCOVA model is more accurate. Hence, that is why ANCOVA has a higher adjusted R-Squared than MLR.
- ANOVA shows that the model with interactions is more accurate than the model without interactions, so the price of a diamond is also affected by the interactions between its categorical variables.

Declaration of Authorship

I, Alan Zammit[1] , declare that this assignment entitled:

"Hypothesis Testing and Statistical Modelling on Diamonds"[2]

and the work presented in it are my own.

I confirm that:

1. Where any part of this assignment has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
2. This work is submitted in ~~partial~~[3] fulfillment of the requirements of the credit SOR1232 [4]offered by the Department of Statistics and Operations Research, Faculty of Science, University of Malta.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the works of others, the source is always given. With the exception of such quotations, this assignment is entirely my own work.
5. I have acknowledged all sources used for the purpose of this work.
6. I have read the guidelines are regulations of the University of Malta regarding plagiarism and understand that the penalties for committing a breach of the regulations include the loss of marks, cancellation of examination results; enforced suspension of studies; or expulsion from the degree program.

Signature:

Date: 30/05/2025

[1] Insert name surname and identity card number.

[2] Insert title of Assignment.

[3] remove the word partial where appropriate.

[4] Insert code of credit.

Declaration of Authorship

I, Luca Callus 157906L[1] , declare that this assignment entitled:

“Hypothesis Testing and Statistical Modelling on Diamonds”[2]

and the work presented in it are my own.

I confirm that:

1. Where any part of this assignment has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
2. This work is submitted in ~~partial~~[3] fulfillment of the requirements of the credit SOR1232 [4]offered by the Department of Statistics and Operations Research, Faculty of Science, University of Malta.
3. Where I have consulted the published work of others, this is always clearly attributed.
4. Where I have quoted from the works of others, the source is always given. With the exception of such quotations, this assignment is entirely my own work.
5. I have acknowledged all sources used for the purpose of this work.
6. I have read the guidelines are regulations of the University of Malta regarding plagiarism and understand that the penalties for committing a breach of the regulations include the loss of marks, cancellation of examination results; enforced suspension of studies; or expulsion from the degree program.

Signature: ____________________

Date: 30/05/2025

[1] Insert name surname and identity card number.

[2] Insert title of Assignment.

[3] remove the word partial where appropriate.

[4] Insert code of credit.