Lab 4

Tim Lanthier

2/9/2022

STAT 108: Analysis of Variance

Github Repository: https://github.com/talanthier/lab-04

Exploratory Data Analysis

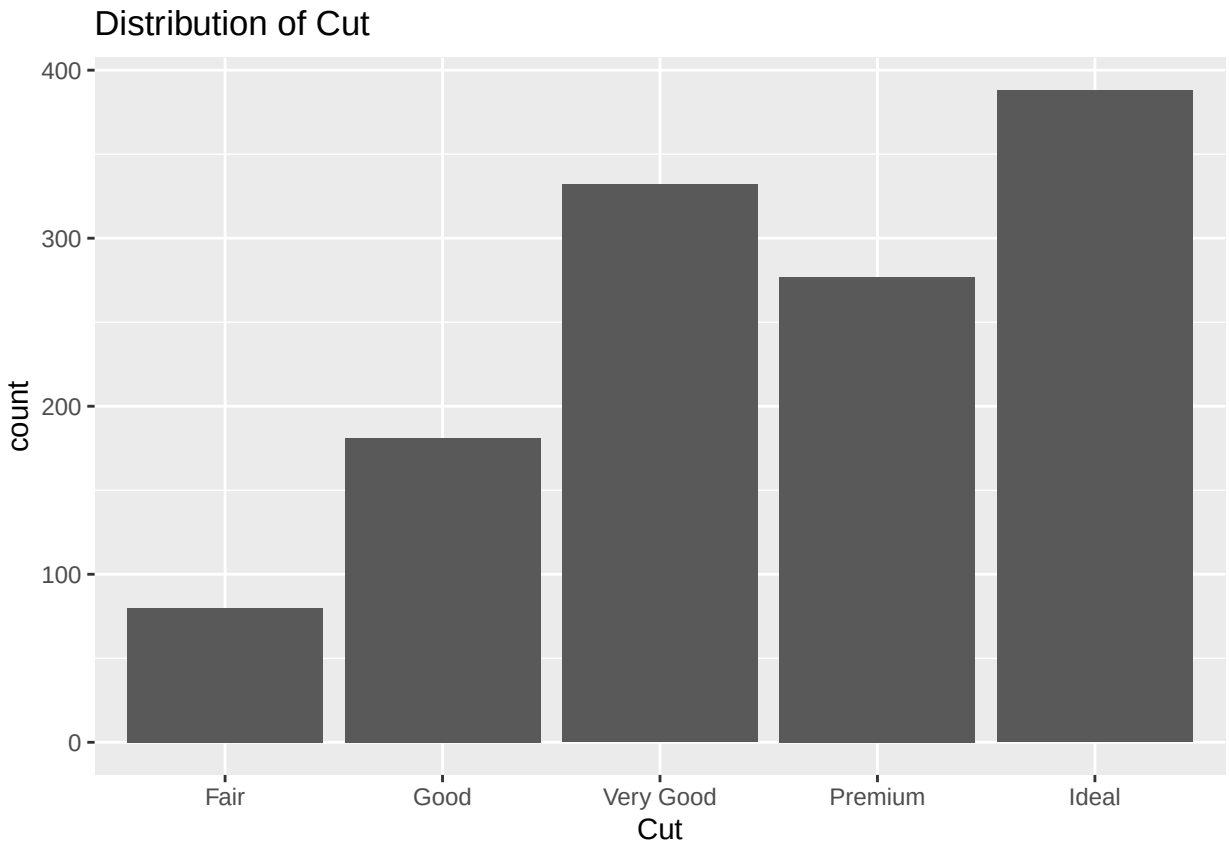
Dataset is the `diamonds` dataset from the ggplot2 package. For this analysis, we will only consider diamonds with a carat weight of 0.5.

```
filtered_data <- diamonds %>% subset(carat == 0.5)
glimpse(filtered_data)
```

```
## Rows: 1,258
## Columns: 10
## $ carat   <dbl> 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.~
## $ cut     <ord> Ideal, Ideal, Good, Good, Very Good, Fair, Fair, Fair, Fair, F~
## $ color   <ord> E, E, D, D, D, F, F, F, F, F, G, F, E, G, G, F, F, E, E, F, E,~
## $ clarity <ord> VVS2, VVS2, VVS2, IF, IF, I1, I1, I1, I1, I1, I1, I1, I1, I1, ~
## $ depth   <dbl> 62.2, 62.2, 62.4, 63.2, 62.9, 69.8, 71.0, 68.4, 67.1, 68.3, 64~
## $ table   <dbl> 54, 54, 64, 59, 59, 55, 57, 54, 57, 58, 60, 58, 61, 57, 56, 60~
## $ price   <int> 2889, 2889, 3017, 3378, 3378, 584, 613, 613, 627, 627, 701, 71~
## $ x       <dbl> 5.08, 5.09, 5.03, 4.99, 4.99, 4.89, 4.87, 4.94, 4.92, 4.91, 5.~
## $ y       <dbl> 5.12, 5.11, 5.06, 5.04, 5.09, 4.80, 4.79, 4.82, 4.87, 4.78, 4.~
## $ z       <dbl> 3.17, 3.17, 3.14, 3.17, 3.17, 3.38, 3.43, 3.35, 3.28, 3.32, 3.~
```

It looks like we now have 1,258 observations in our filtered dataset.

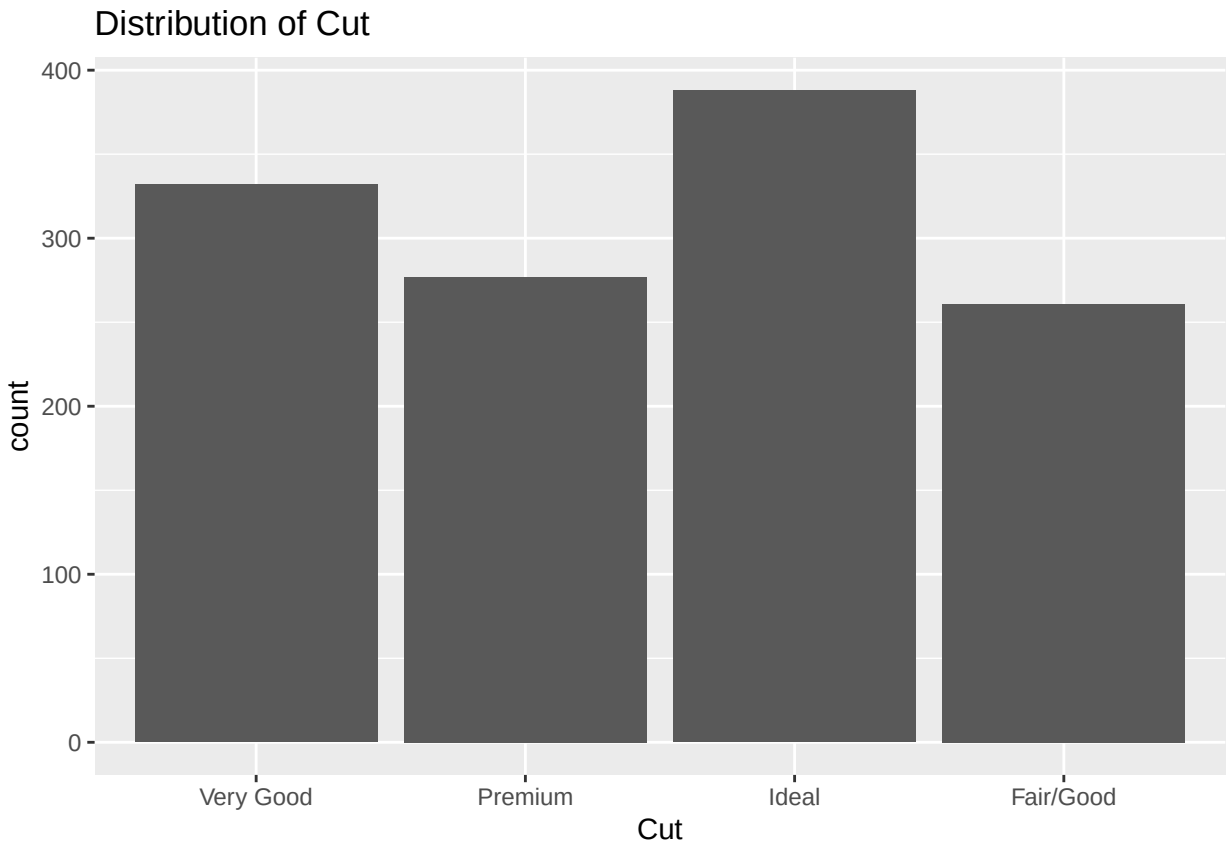
```
ggplot(filtered_data, aes(cut)) +
  geom_bar() +
  labs(x = 'Cut', title = 'Distribution of Cut')
```



According to the above plot, we have the fewest number of diamonds with a fair or good cut. To address this, we will combine the fair and good cut categories so the distribution between various cuts are more uniform.

```
filtered_data <- filtered_data %>%
  mutate(cut = fct_lump_n(cut, 3, other_level = 'Fair/Good'))
```

```
ggplot(filtered_data, aes(cut)) +
  geom_bar() +
  labs(x = 'Cut', title = 'Distribution of Cut')
```



According to the above bar chart, it looks like we have successfully lumped all fair and good cut diamonds into a single category. Now we will investigate the relationship between diamond cut and its price.

```
ggplot(filtered_data, aes(cut,price)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = 'Cut', y = 'Price', title = 'Diamond Cut vs Price')
```


```
filtered_data %>% group_by(cut) %>%
  summarise(mean_price = mean(price),
            std_dev_price = sd(price),
            num_obs = n())
```

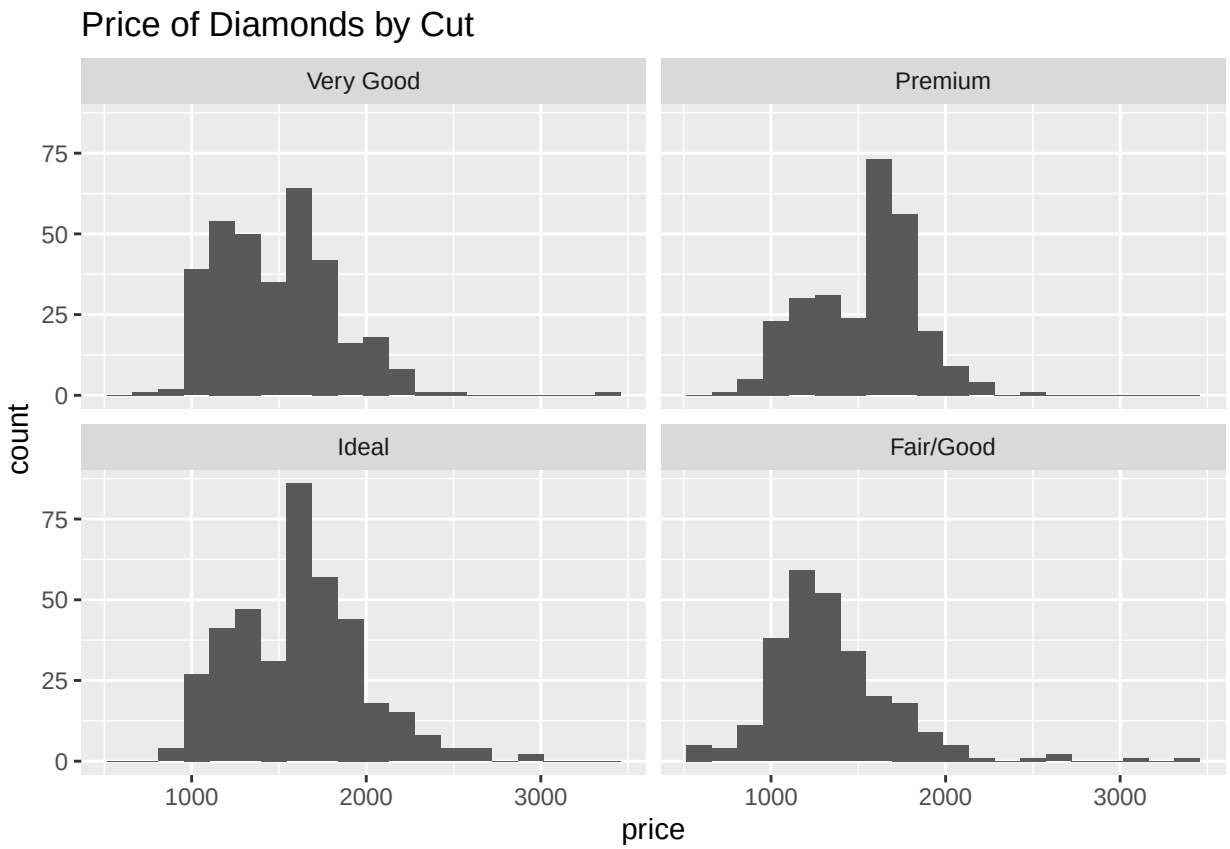
```
## # A tibble: 4 x 4
##   cut       mean_price std_dev_price num_obs
##   <ord>          <dbl>         <dbl>   <int>
## 1 Very Good      1489.          339.     332
## 2 Premium        1532.          304.     277
## 3 Ideal          1609.          368.     388
## 4 Fair/Good      1341.          365.     261
```

Looking at the above plots and summary statistics it is reasonable to assume that there may be some relationship between price of diamonds and its cut. We can see this as the mean prices for diamonds with a premium or ideal cut lie quite far above diamonds with fair or good cuts. This is supported by our boxplot as the upper 75% of diamonds with premium cut have prices which lie above the mean price for diamonds with fair/good cut. Of course this only applies to diamonds which are 0.5 carats since we removed all other diamonds from the dataset.

Analysis of Variance

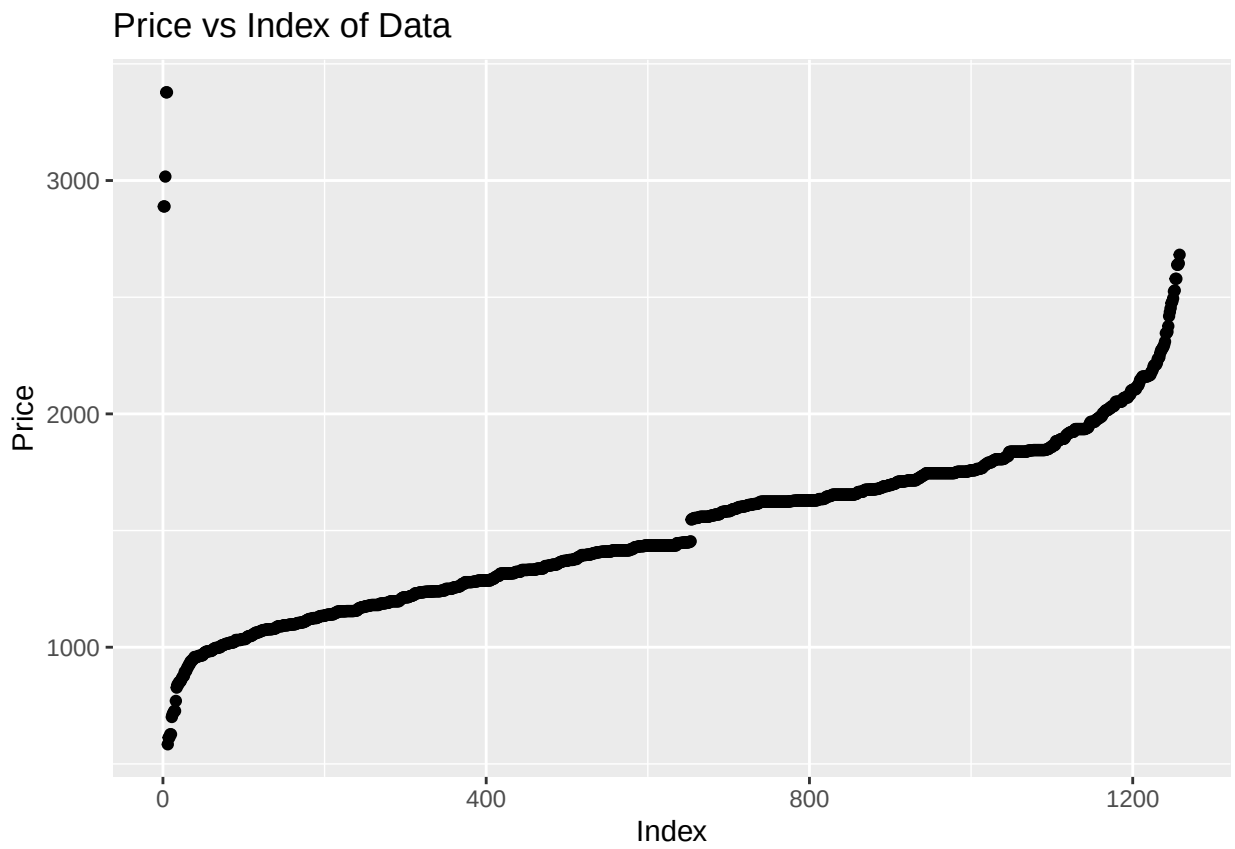
We will start with the normality assumption. Looking at the plot below, grouped by cut, most of the distributions look reasonably bell curve shaped. All four histograms are unimodal but a few lack the symmetry. For example, 'Very Good' cut diamonds have a spike around the $1200 price mark as well as around $1500. That being said, we would say the shape is mostly bell curve like so we would say the normality condition is satisfied.

```
ggplot(filtered_data, aes(x = price)) +
  geom_histogram(bins = 20) +
  facet_wrap(~cut) +
  labs(title = 'Price of Diamonds by Cut')
```



As for constant variance, this condition is also satisfied. Looking at the summary statistics we calculated earlier, the standard deviations for price within each group lies relatively close to each other. For example, the price of diamonds with very good, ideal, and fair/good cuts have standard deviations 339.363, 368.345, and 364.522 respectively. While premium diamonds have a lower standard deviation of 304.144, this is still quite close to the standard deviations of the other groups. So since the standard deviations (and hence the variances) between groups are similar, we would conclude that the constant variance condition is satisfied.

```
ggplot(data = filtered_data, aes(x = seq.int(nrow(filtered_data)), y = price)) +
  geom_point() +
  labs(x = 'Index', y= 'Price', title = 'Price vs Index of Data')
```



For checking independence, we have no knowledge as to how the data was collected. Just looking to see if there is any relationship as to how the data was ordered, we can plot the price against the rows in which the observations appeared. The plot is shown above and shows there is a clear relationship between how the data was ordered and the diamond price. Of course, it is possible that the dataset was sorted based off of price after the data was collected, but since we do not know how the data was collected, we cannot be sure that the observations are independent. Hence we shall consider the independence assumption violated. If the observations were put into the dataset as it was collected, then the observations are dependent on adjacent observations.

```
cut_model <- lm(price ~ cut, data = filtered_data)

anova(cut_model) %>% kable(format = 'markdown', digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cut	3	11507056	3835685.3	31.916	0
Residuals	1254	150706506	120180.6	NA	NA

The sample variance for price includes the variation within each group as well the variation between groups. From the ANOVA table we find the total sum of squares to be $\text{SST} = 11507056 + 150706506 = 162213562$ and the total degrees of freedom is $\text{DFT} = 1254 + 3 = 1257$. Hence we can find the sample variance of price to be

$$s_y^2 = \frac{\text{SST}}{\text{DFT}} = \frac{162213562}{1257} = 129048.2$$

We can quickly verify this

```
var(x = filtered_data$price)
```

```
## [1] 129048.2
```

```
filtered_data %>% group_by(cut) %>% summarize(Variance = var(price))
```

```
## # A tibble: 4 x 2
##   cut       Variance
##   <ord>        <dbl>
## 1 Very Good  115167.
## 2 Premium     92504.
## 3 Ideal      135678.
## 4 Fair/Good  132876.
```

Looking at the above table, we can find the sample variance for price within each of the 4 levels of cut. As for the ANOVA, we have the two following hypotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \mu_i \neq \mu_j \text{ for some } i \neq j$$

where $\mu_1, \mu_2, \mu_3, \mu_4$ are the true population means for price within each of the 4 levels of cut. Intepreting these, the null hypothesis states that the true population mean price of diamonds is the same between groups of diamond cut. The alternate hypothesis states that there is at least one group whose true mean diamond price differs from the true mean diamond price of another diamond cut.

Looking at the ANOVA table, we have a large F value and a very small corresponding p-value. Our p-value is small enough that we cannot express it with 3 digits of precision. Since the p-value is so small, we have sufficient evidence to reject the null hypothesis that there is no difference in true mean price between groups of diamonds based on cut.

So we have evidence that the mean price for diamonds within a level of cut is different from the mean price within another level of cut. Now we will investigate which means significantly differ from one another.

Further Analysis

In order to verify which means in particular are different from one another, we will conduct a few t tests pairwise. Seeing as we have 4 levels for `cut` (Fair/Good,Very Good, Ideal, Premium) we will need to conduct $\binom{4}{2} = 6$ t tests in order to cover all cases. We can start by filtering the diamonds into 4 separate datasets based on `cut`.

```
fair_good_diamonds <- filtered_data %>% filter(cut == 'Fair/Good')
very_good_diamonds <- filtered_data %>% filter(cut == 'Very Good')
ideal_diamonds <- filtered_data %>% filter(cut == 'Ideal')
premium_diamonds <- filtered_data %>% filter(cut == 'Premium')
```

Fair/Good vs Very Good Cut Diamonds

Now we will conduct the t tests. We will start with comparing the groups Fair/Good and Very Good for cut.

```
t.test(fair_good_diamonds$price, very_good_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  fair_good_diamonds$price and very_good_diamonds$price
## t = -5.0592, df = 538.6, p-value = 5.784e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -205.49148  -90.54646
## sample estimates:
## mean of x mean of y
##  1340.644  1488.663
```

According to our 2 sample t test, seeing as we have a very small p-value, we have sufficient evidence to reject the null hypothesis that the difference in means is nonzero. This is supported by our 95% confidence interval which says we are 95% confidence that the difference in the true mean price between the groups lies in the interval (-205.491, -90.546). This confidence interval suggests that we are confident that the true mean price of diamonds with Very Good cut is higher than the true mean price of Fair/Good cut diamonds. We can repeat this analysis for the other pairs of groups.

Fair/Good vs Ideal Cut Diamonds

```
t.test(fair_good_diamonds$price, ideal_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  fair_good_diamonds$price and ideal_diamonds$price
## t = -9.146, df = 561.77, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -325.5848 -210.4629
## sample estimates:
## mean of x mean of y
##  1340.644  1608.668
```

Once again, we have an extremely small p-value. Thus we reject the null hypothesis that the mean price for fair/good cut diamonds is the same as the mean price for ideal cut diamonds. This is supported by the 95% confidence interval (-325.585, -210.4629) as 0 does not lie within the confidence interval. Also, by this confidence interval we are 95% confident that the true mean difference in price of fair/good cut diamonds and ideal cut diamonds lies between (-325.585, -210.4629).

Fair/Good vs Premium Cut Diamonds

```
t.test(fair_good_diamonds$price, premium_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  fair_good_diamonds$price and premium_diamonds$price
## t = -6.5827, df = 507.33, p-value = 1.155e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -248.1768 -134.0881
## sample estimates:
## mean of x mean of y
##  1340.644  1531.776
```

Seeing as we have such a small p-value we have sufficient evidence to reject the hypothesis that the true difference in mean price between fair/good cut diamonds and premium cut diamonds is 0. From the 95% confidence interval, we can conclude that the true difference lies in the interval (-248.177, -134.0881). So we are confident that the true mean price for premium cut diamonds is higher than the true mean price for fair/good cut diamonds.

Very Good vs Ideal Cut Diamonds

```
t.test(very_good_diamonds$price, ideal_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  very_good_diamonds$price and ideal_diamonds$price
## t = -4.5469, df = 714.07, p-value = 6.394e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -171.82144  -68.18831
## sample estimates:
## mean of x mean of y
##  1488.663  1608.668
```

Seeing as we have a very small p-value, we have sufficient evidence to reject the null hypothesis that the mean price of very good cut diamonds does not differ from ideal cut diamonds. Looking at the confidence interval, we are 95% confident that the true difference in mean price between the 2 groups lies between -171.821 and -68.188. Hence we have sufficient evidence that the true mean price in ideal diamonds lies above the true mean price of very good diamonds.

Very Good vs Premium Cut Diamonds

```
t.test(very_good_diamonds$price, premium_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  very_good_diamonds$price and premium_diamonds$price
## t = -1.6523, df = 603.88, p-value = 0.09899
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -94.357312   8.130267
## sample estimates:
## mean of x mean of y
##  1488.663  1531.776
```

Here we have a p-value of 0.099. At the $\alpha = 0.05$ significance level, since $0.099 > 0.05$, we do not have sufficient evidence to reject the null hypothesis that the true mean price of very good cut diamonds is different from the true mean price of premium cut diamonds. We also have a 95% confidence interval of (-94.357, 8.13). This supports our rejection of the null hypothesis as our confidence interval contains 0. So we cannot state that there is a statistically significant difference in price between very good cut diamonds and premium cut diamonds.

Ideal vs Premium Cut Diamonds

```
t.test(ideal_diamonds$price, premium_diamonds$price)
```

```
##
##  Welch Two Sample t-test
##
## data:  ideal_diamonds$price and premium_diamonds$price
## t = 2.9408, df = 649.08, p-value = 0.00339
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##   25.54963 128.23308
## sample estimates:
## mean of x mean of y
##  1608.668  1531.776
```

Seeing as we have a p-value of 0.003, at the $\alpha = 0.05$ significance level, we have sufficient evidence to reject the null hypothesis that the mean price of ideal cut diamonds is the same as the true mean price of premium cut diamonds. Looking at our confidence interval, we have a 95% confidence interval of (25.55, 128.233). Hence, we are 95% confident that the true mean price of ideal cut diamonds is higher than the true mean price of premium cut diamonds with a difference between \$25.549 and \$128.233.

Now we have covered all of the pairs of groups. To summarize, the only groups whose means are not significantly different from one another are Ideal and Premium cut diamonds as well as Very Good and Premium Cut diamonds.