Stat 108: Lab 4

Dylan Scoble

2/9/2022

Exercise 1

There are 1,258 observations in the new dataset, after we filter for observations that have a carat weight of 0.5.

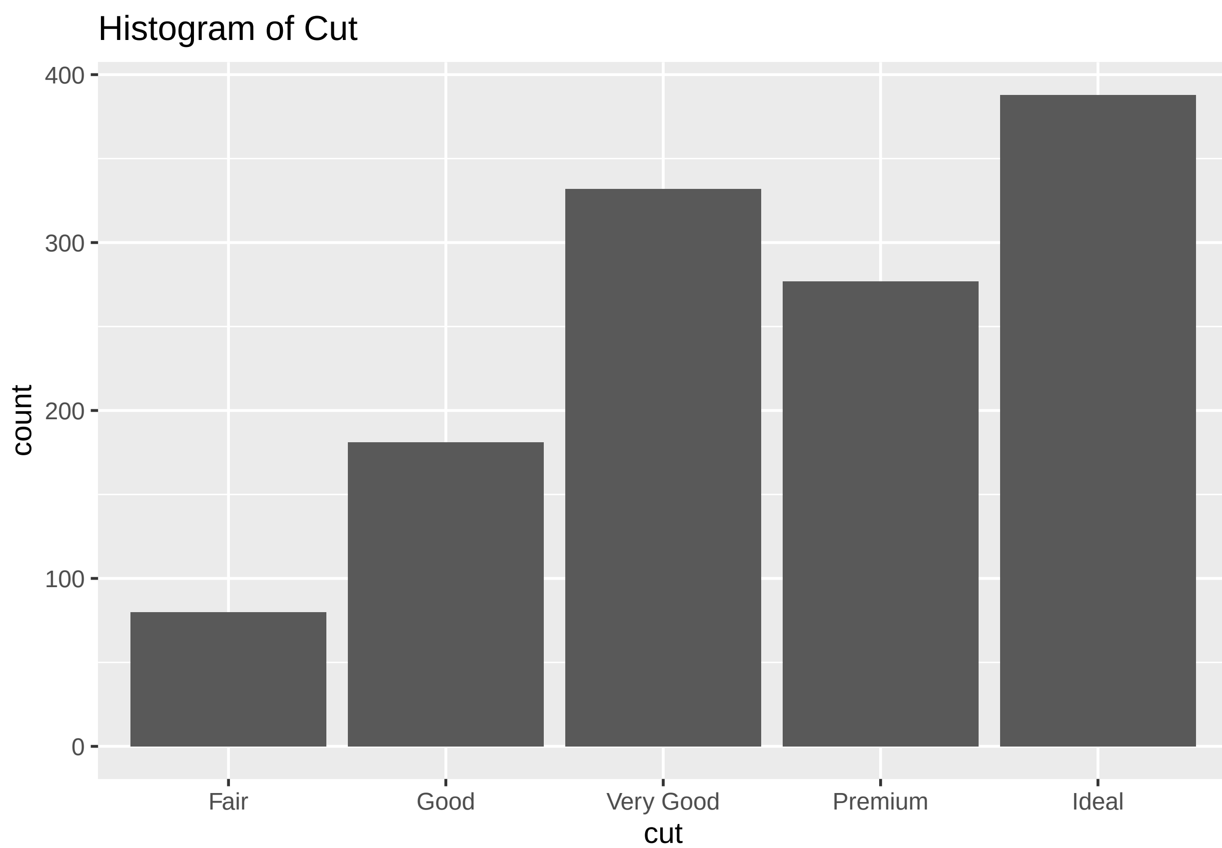
```
data <- diamonds %>%
  filter(carat == 0.5)
glimpse(data)
```

```
## 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.~
```

Exercise 2

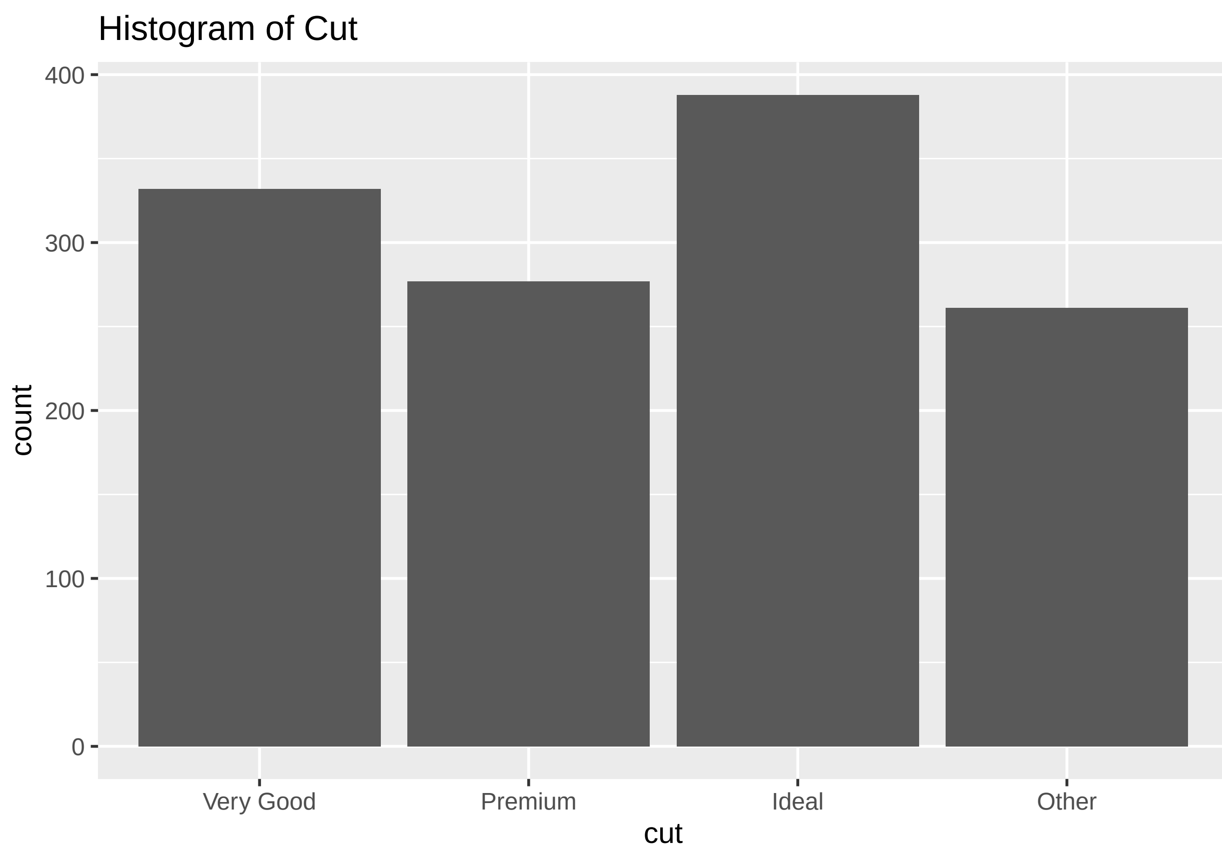
As shown in the histogram, the two levels of cut with the smallest number of observations are "few" and "good".

```
ggplot(data = data, aes(x = cut)) +
  geom_bar() +
  labs(title = "Histogram of Cut")
```



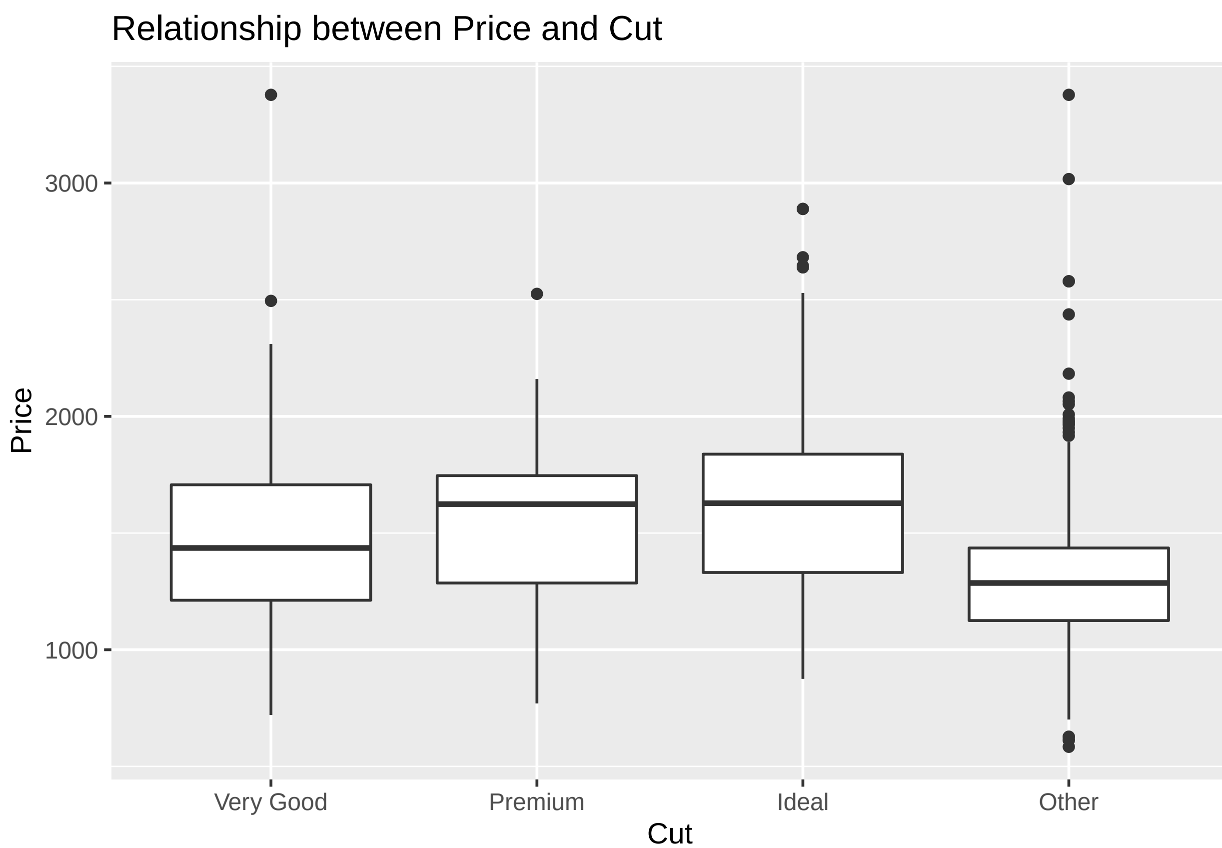
Exercise 3

```
lumpeddata <- data %>%
  mutate(cut = fct_lump_n(cut, n=3))

ggplot(data = lumpeddata, aes(x = cut)) +
  geom_bar() +
  labs(title = "Histogram of Cut")
```



Exercise 4

```
plot <- ggplot(data = lumpeddata, aes(x = cut, y = price)) +
  geom_boxplot() +
  labs(title = "Relationship between Price and Cut",
       x = "Cut",
       y = "Price")

plot
```



Exercise 5

The output of this code is a table that shows the mean, standard deviation, and number of observations of price for each category of cut.

```
lumpeddata %>%
  group_by(cut) %>%
  summarise(mean = mean(price),
            std_dev = sd(price),
            num_observations = n()
  )
```

```
## # A tibble: 4 x 4
##   cut        mean std_dev num_observations
##   <ord>     <dbl>   <dbl>            <int>
## 1 Very Good 1489.    339.              332
## 2 Premium   1532.    304.              277
## 3 Ideal     1609.    368.              388
## 4 Other     1341.    365.              261
```

Exercise 6

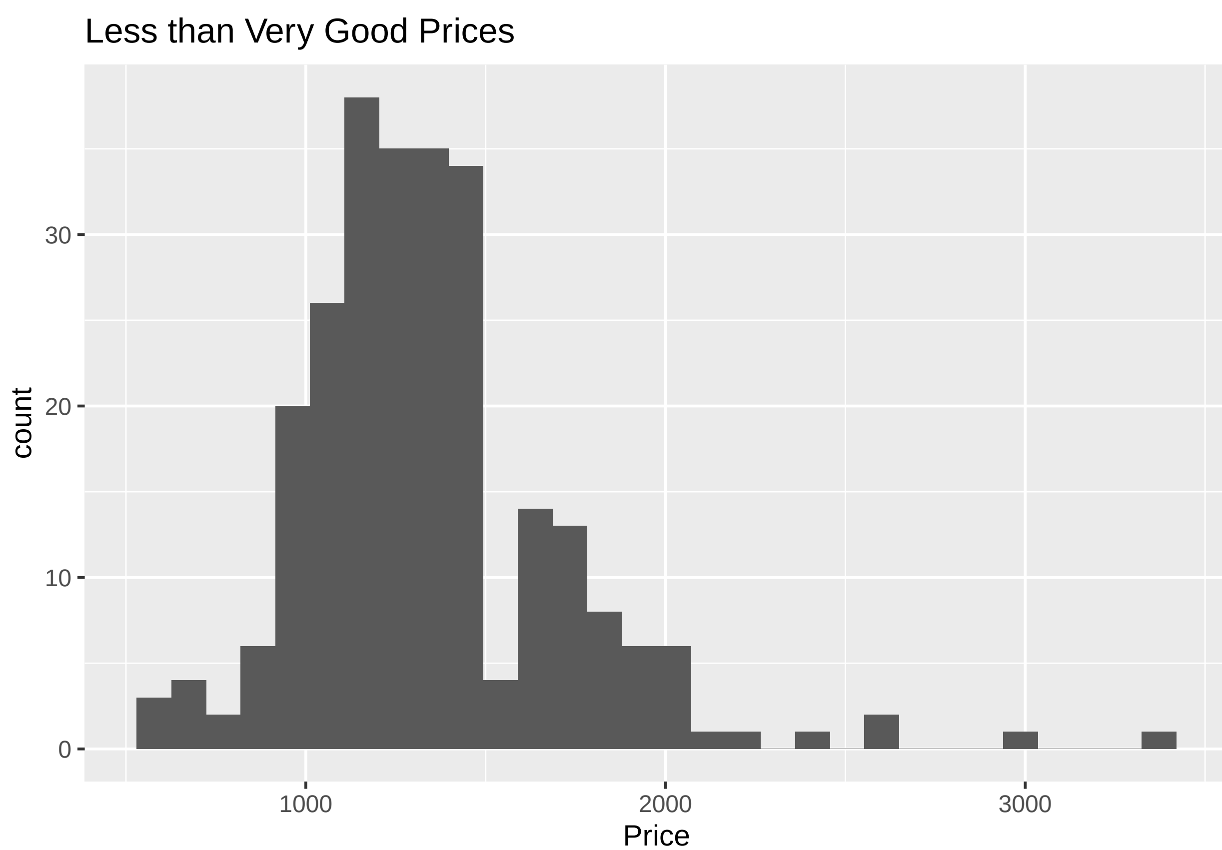
For Diamonds that are 0.5 carats, there is a minor linear relationship between cut and price. As cut increases from less than very good, to very good, to premium, to ideal, the mean price of that category also increases.

Exercise 7

The following code will test the normality assumption. The output of this code allows me to conclude that the normality assumption is satisfied because each plot folllows a relatively normal distribution.

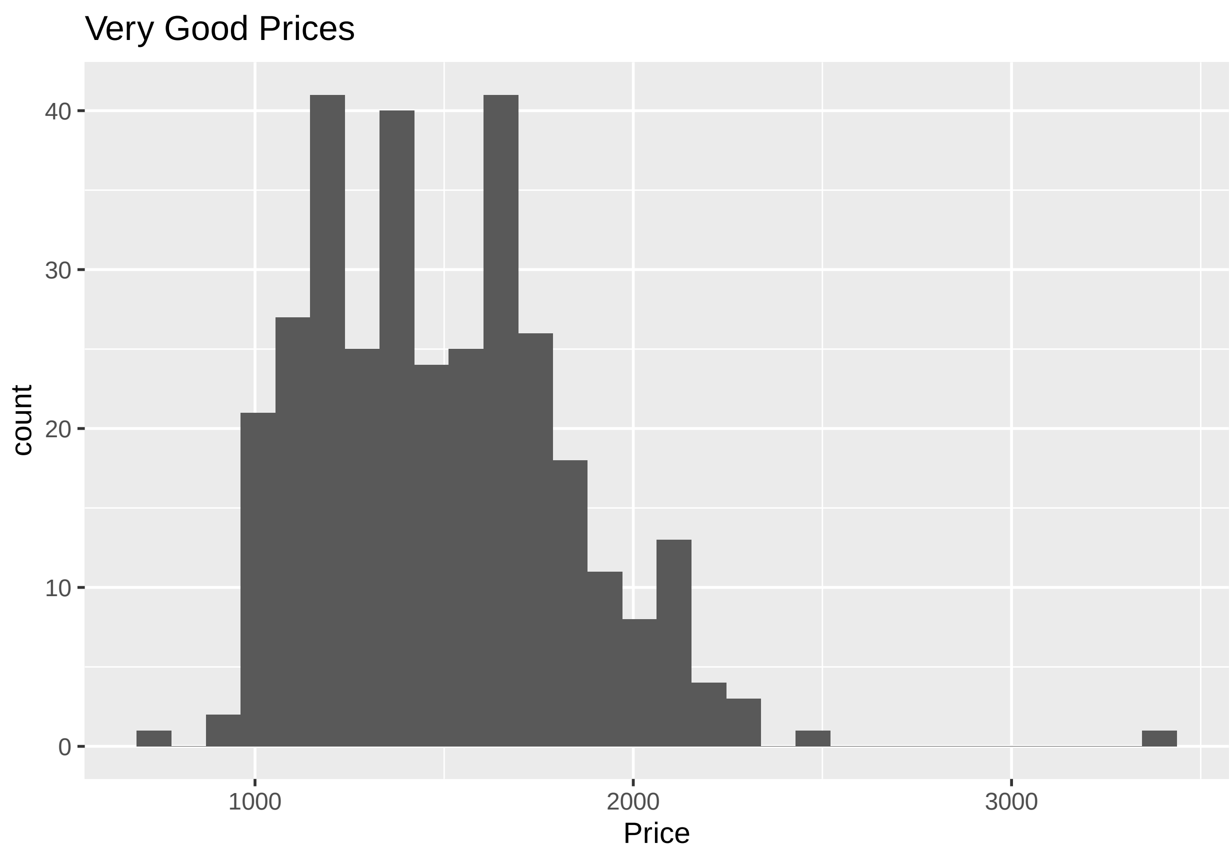
```
badcut <- lumpeddata %>%
  filter(cut == "Other")
ggplot(data = badcut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Less than Very Good Prices",
       x = "Price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



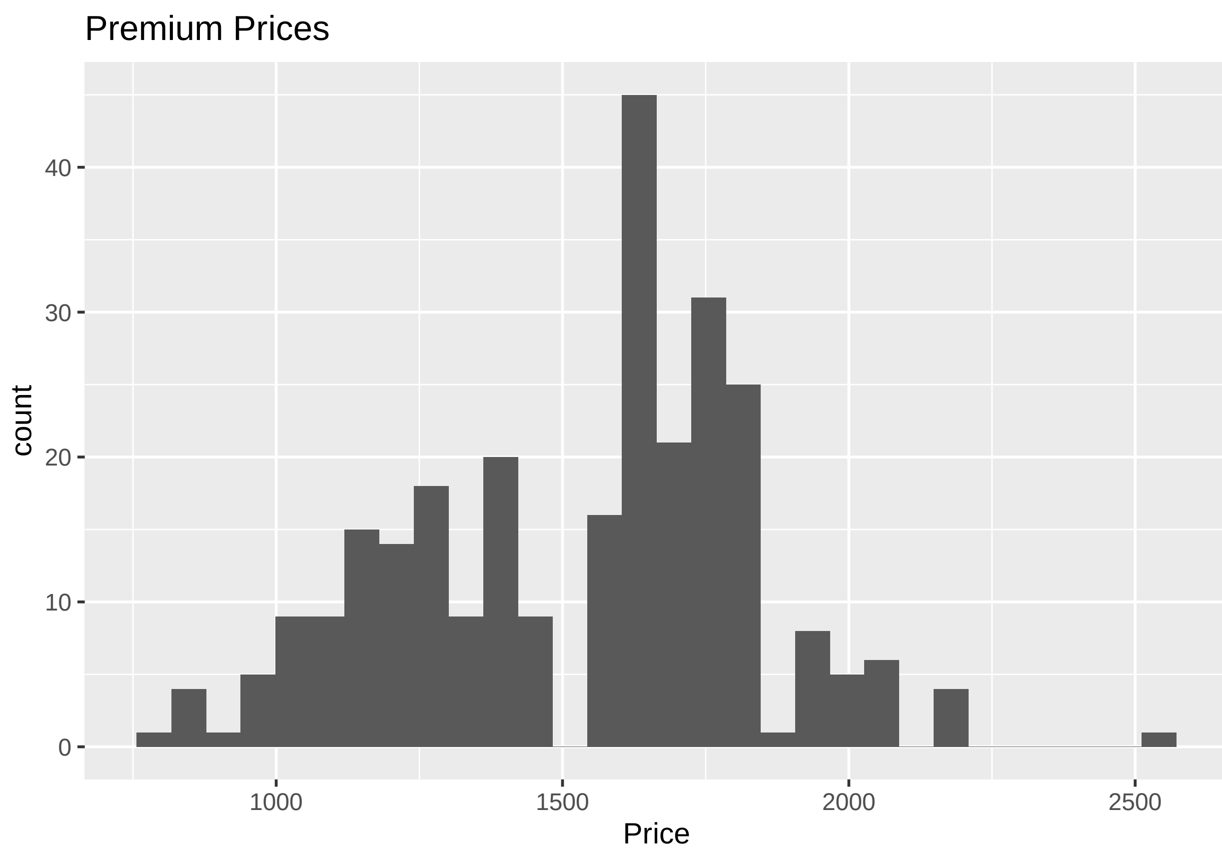

```
verygoodcut <- lumpeddata %>%
  filter(cut == "Very Good")
ggplot(data = verygoodcut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Very Good Prices",
       x = "Price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



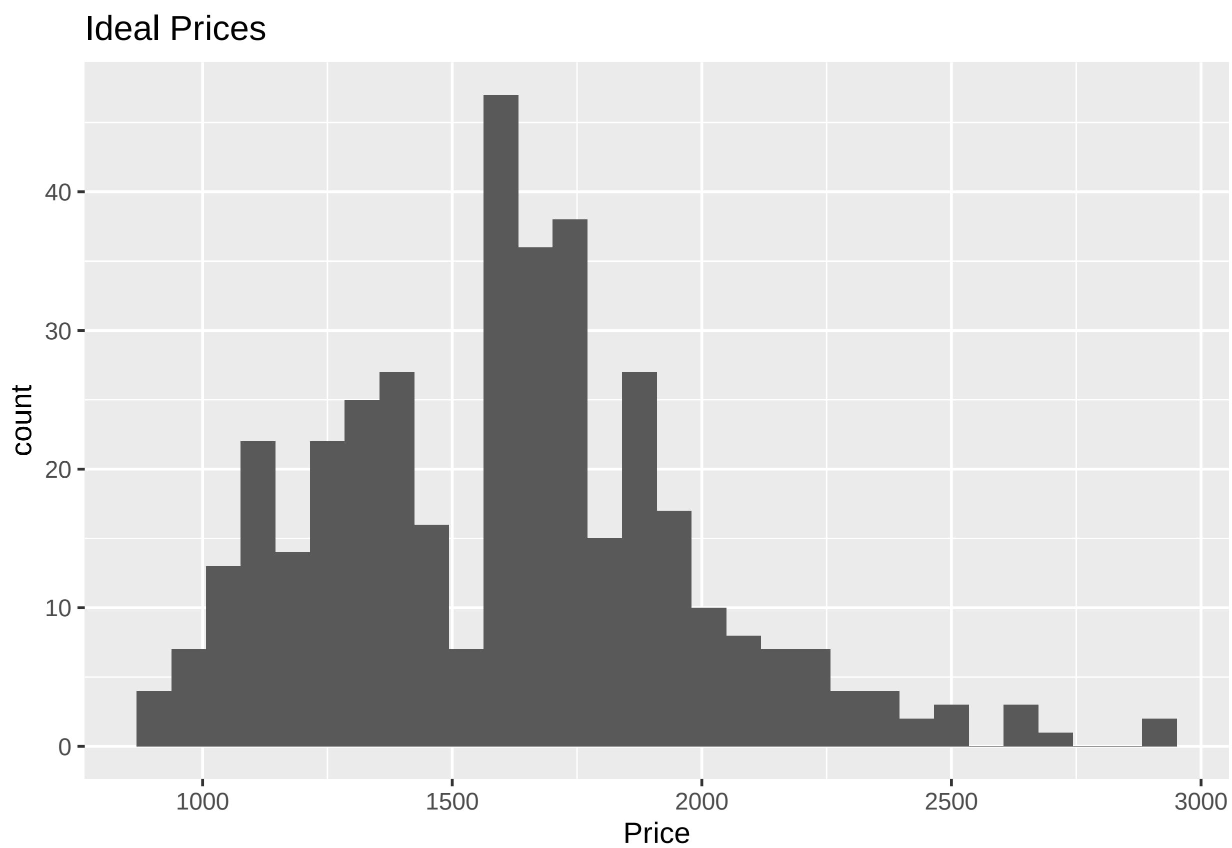

```r
premiumcut <- lumpeddata %>%
  filter(cut == "Premium")
ggplot(data = premiumcut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Premium Prices",
       x = "Price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
idealcut <- lumpeddata %>%
  filter(cut == "Ideal")
ggplot(data = idealcut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Ideal Prices",
       x = "Price")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



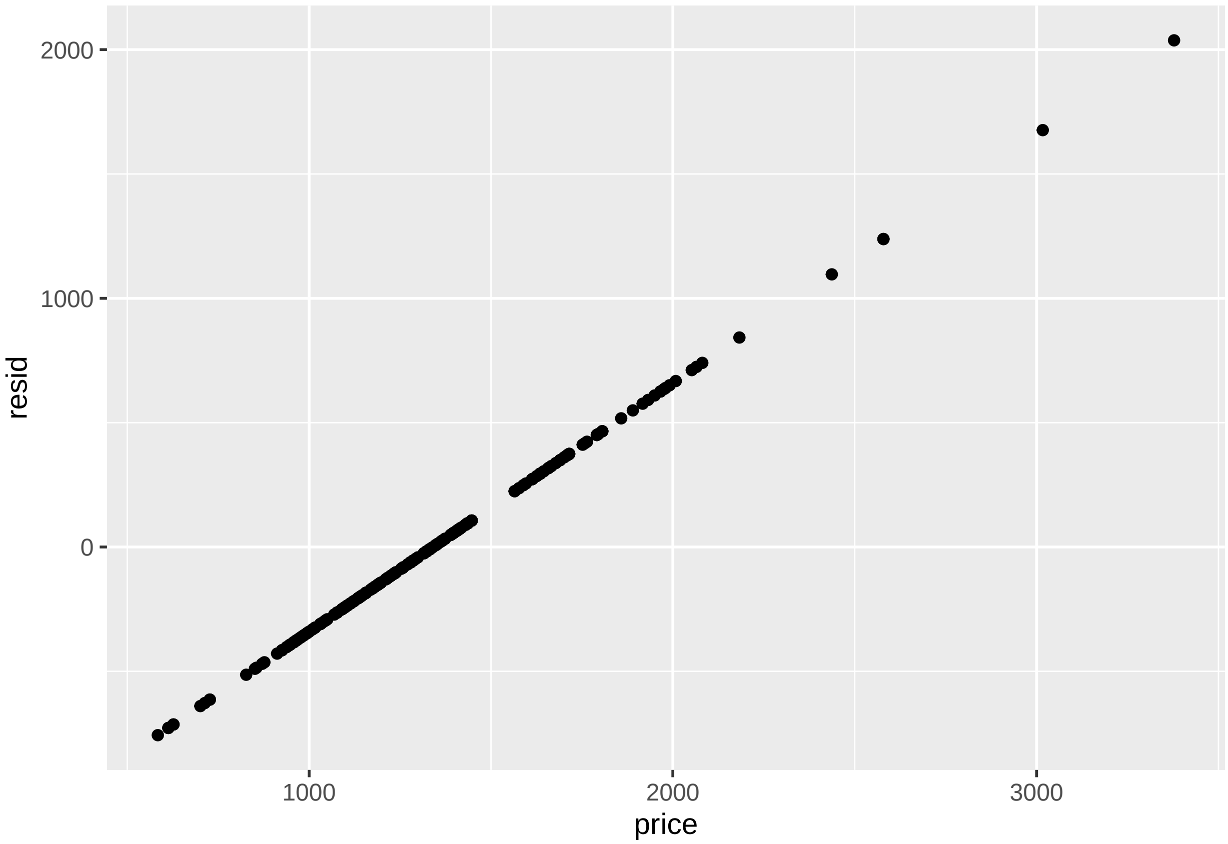
The independence assumption is [not] satisfied because []

The following code will test the constant variance assumption.

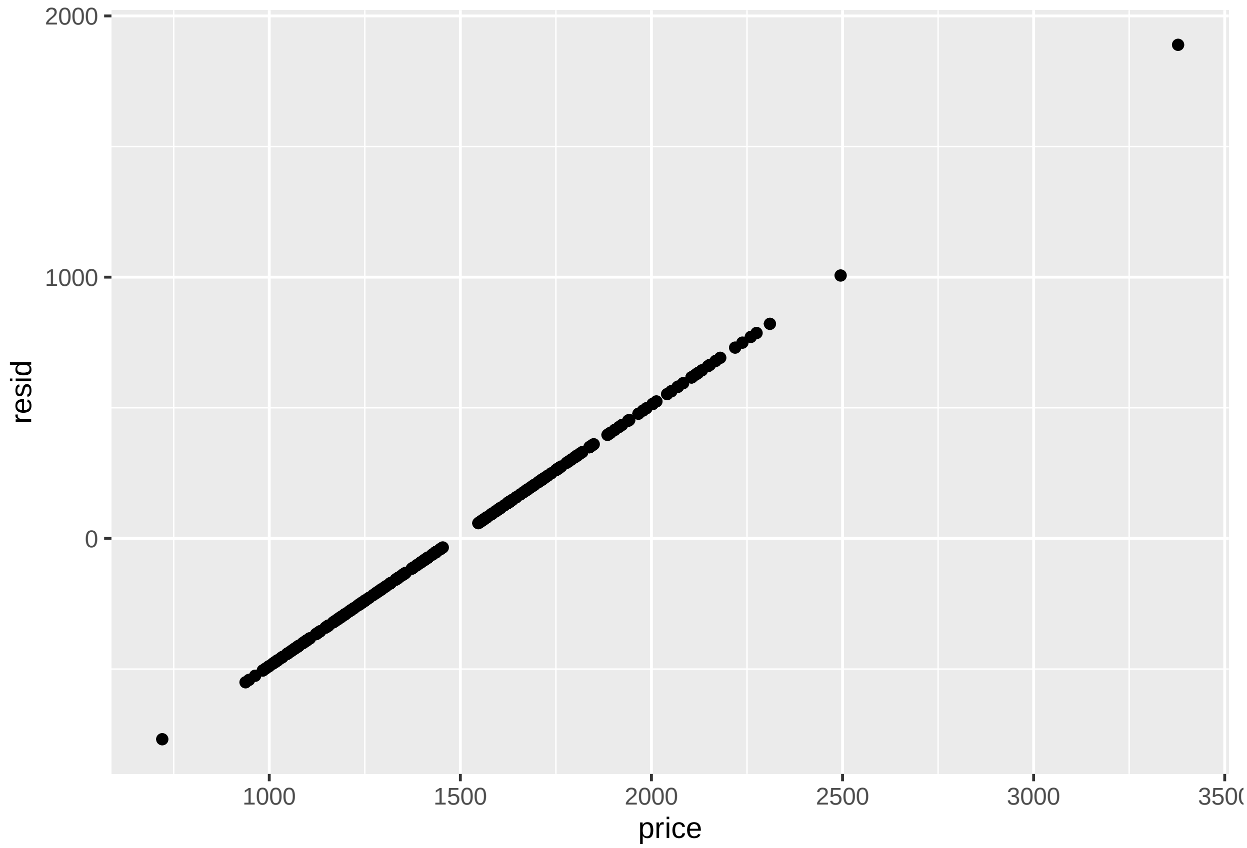
```
model <- lm(price ~ cut, data = lumpeddata)
tidy(model) %>%
  kable(format="markdown", digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1492.438	9.893	150.852	0
cut.L	-82.101	20.181	-4.068	0
cut.Q	-155.569	19.787	-7.862	0
cut.C	-84.678	19.384	-4.368	0

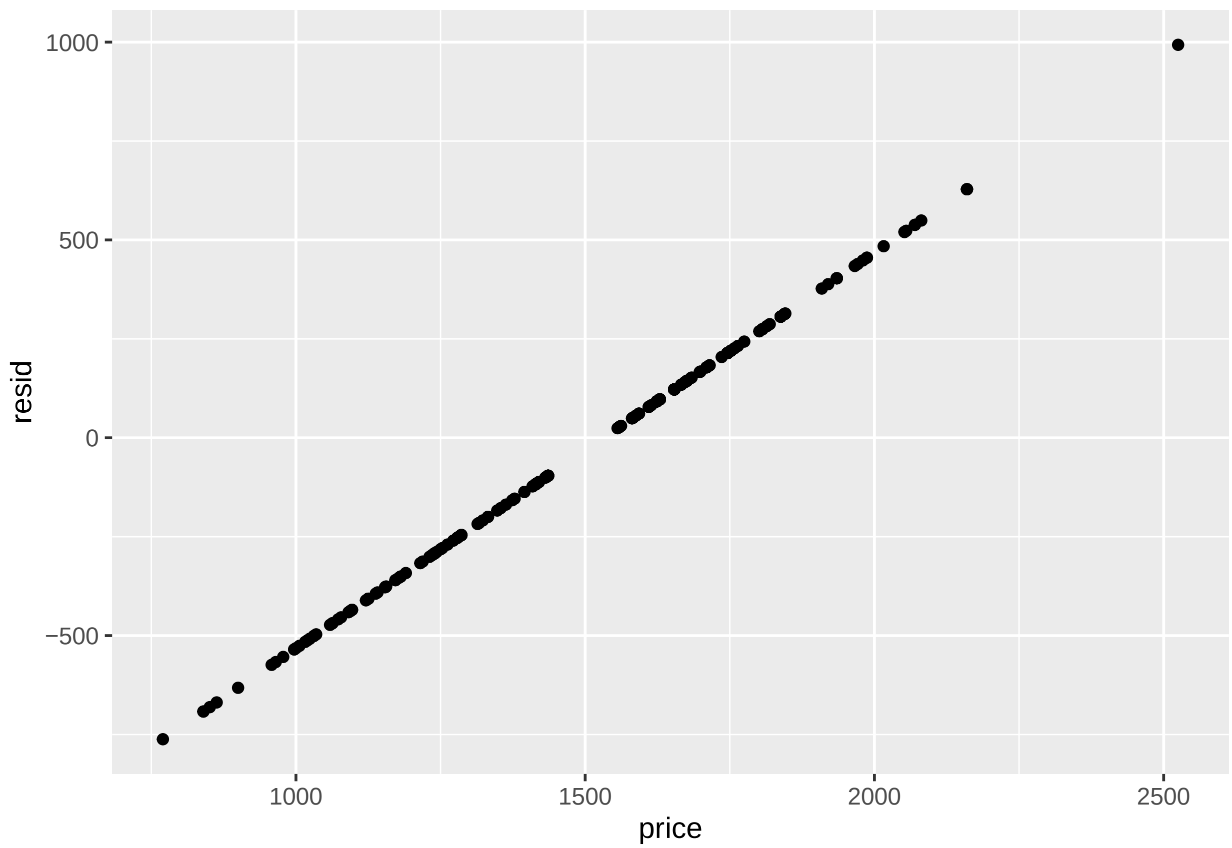
```
lumpeddata <- lumpeddata %>%
  mutate(resid = residuals(model))

badcut <- lumpeddata %>%
  filter(cut == "Other")
ggplot(data = badcut, aes(x = price, y = resid)) +
  geom_point()
```



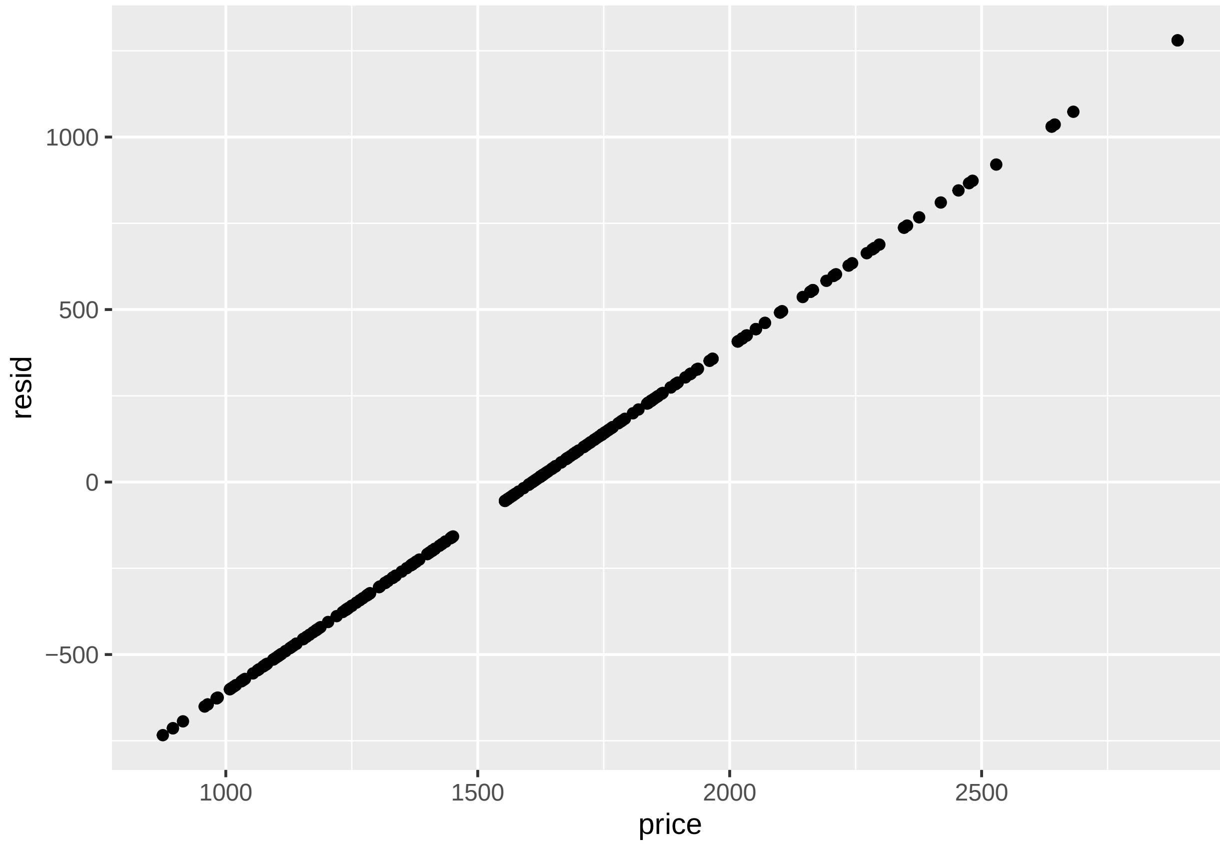

```r
verygoodcut <- lumpeddata %>%
  filter(cut == "Very Good")
ggplot(data = verygoodcut, aes(x = price, y = resid)) +
  geom_point()
```




```
premiumcut <- lumpeddata %>%
  filter(cut == "Premium")
ggplot(data = premiumcut, aes(x = price, y = resid)) +
  geom_point()
```




```
idealcut <- lumpeddata %>%
  filter(cut == "Ideal")
ggplot(data = idealcut, aes(x = price, y = resid)) +
  geom_point()
```



The constant variance assumption is not satisfied because for each level of price, there is not a cloud-shaped pattern to represent the relationship between cut and price.

Exercise 8

```
model <- lm(price ~ cut, data = lumpeddata)
kable(anova(model), format="markdown",digits=6)
```

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

Exercise 9

According to the ANOVA table above, the sample variance for price is 120180.6.

Exercise 10

The output of the code below shows the variance of price for each level of cut. For an ideal cut, the variance of price is 9.893. For a premium cut, the variance of price is 19.384. For a very good cut, the variance of price is 19.787. For any other cut, the variance of price is 20.181.

```r
tidy(model) %>%
  kable(format="markdown", digits=3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1492.438	9.893	150.852	0
cut.L	-82.101	20.181	-4.068	0
cut.Q	-155.569	19.787	-7.862	0
cut.C	-84.678	19.384	-4.368	0

```r
summary(model)$coef[,2]
```

```
## (Intercept)       cut.L       cut.Q       cut.C
##    9.893372   20.181275   19.786744   19.384185
```

Exercise 11

The null hypothesis states that there is not a statistically significant difference in price between each level of cut. The alternative hypothesis states that there is a statistically significant difference in price between each level of cut. In a mathematical sense, $H_0 : mean(price : cut = verygood) = mean(price : cut = premium) = mean(price : cut = ideal) = mean(price : cut = other)$

H_A :!$(mean(price : cut = verygood) = mean(price : cut = premium) = mean(price : cut = ideal) = mean(price : cut = other))$

Exercise 12

Because the F statistic is very high in our ANOVA analysis, the p-value is low. Therefore, we can reject the null hypothesis that cut does not affect price and accept the alternative hypothesis.

Exercise 13

If the cut is less than very good, we see that this is a major indicator of a lower price. There is also more outliers and a higher variance among diamonds that havea cut of less than very good.