Lab 4

Aisha Lakshman

2/15/2022

Packages for Lab 4

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
library(broom)
library(ggplot2)
```

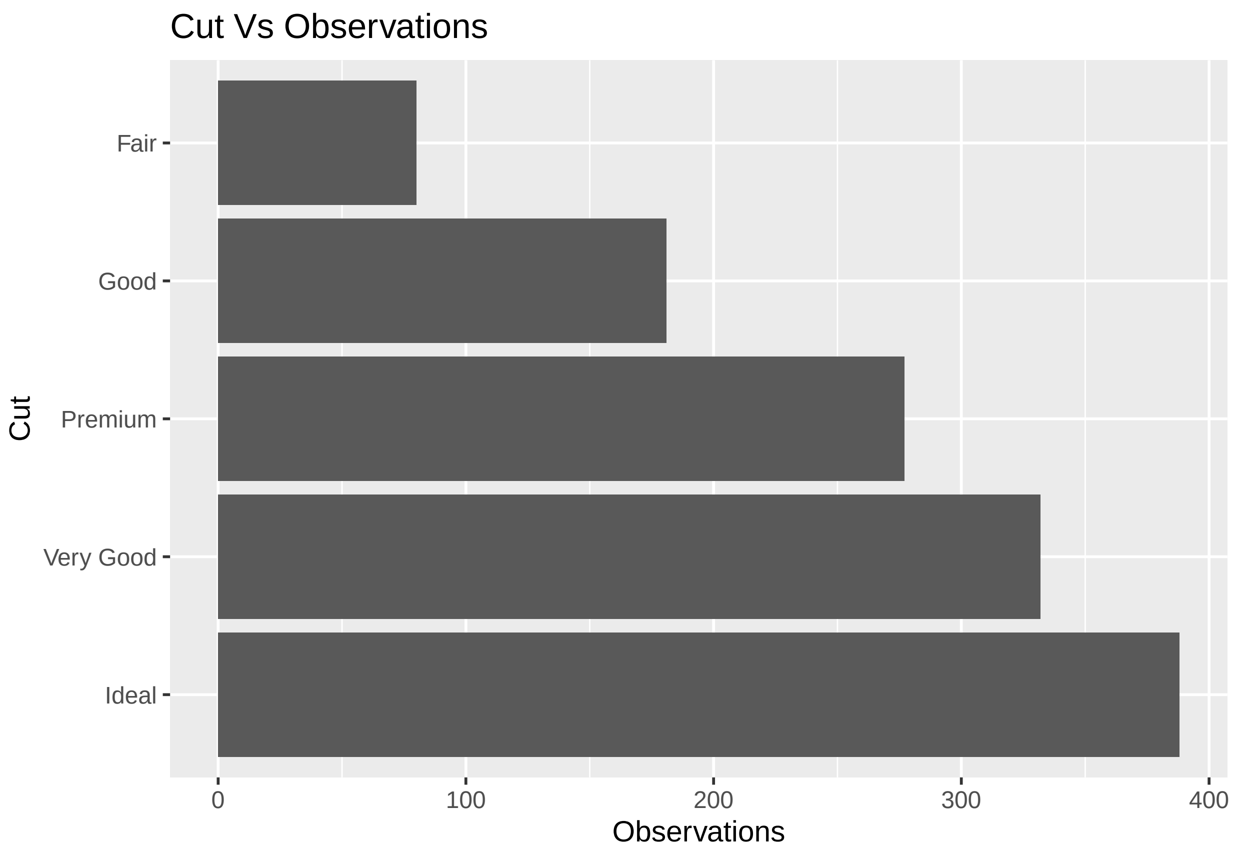
#Exploratory Data Analysis Exercise 1

```
diamonds <- diamonds %>% filter(carat == 0.5, !is.na(carat))
```

The new dataset now has 1258 observations.

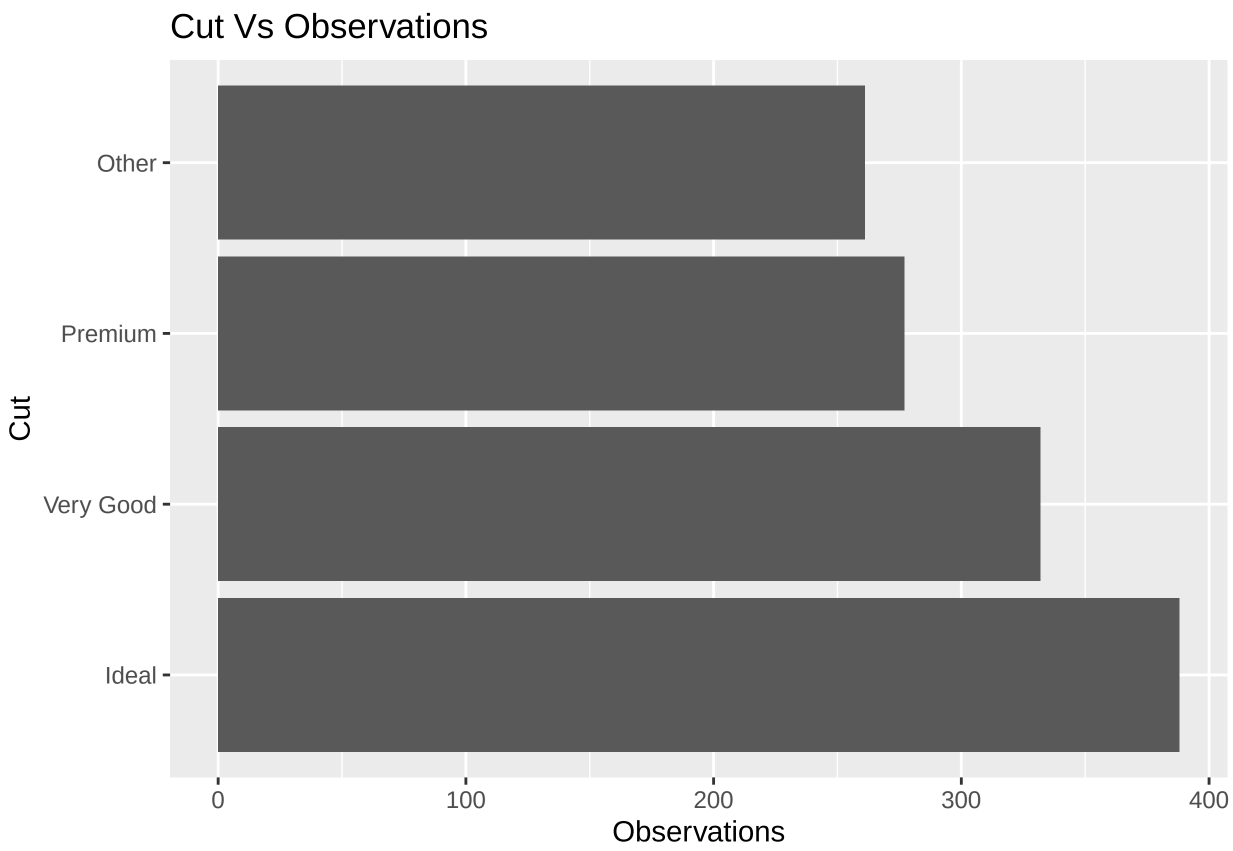
Exercise 2

```
ggplot(diamonds, aes(x = fct_infreq(cut))) +
  geom_bar() +
  labs (x = "Cut", y = "Observations", title = "Cut Vs Observations") +
  coord_flip()
```



The Cut vs Observations plot indicates that cut levels “fair” and “good” have the least amount of observations. Exercise 3

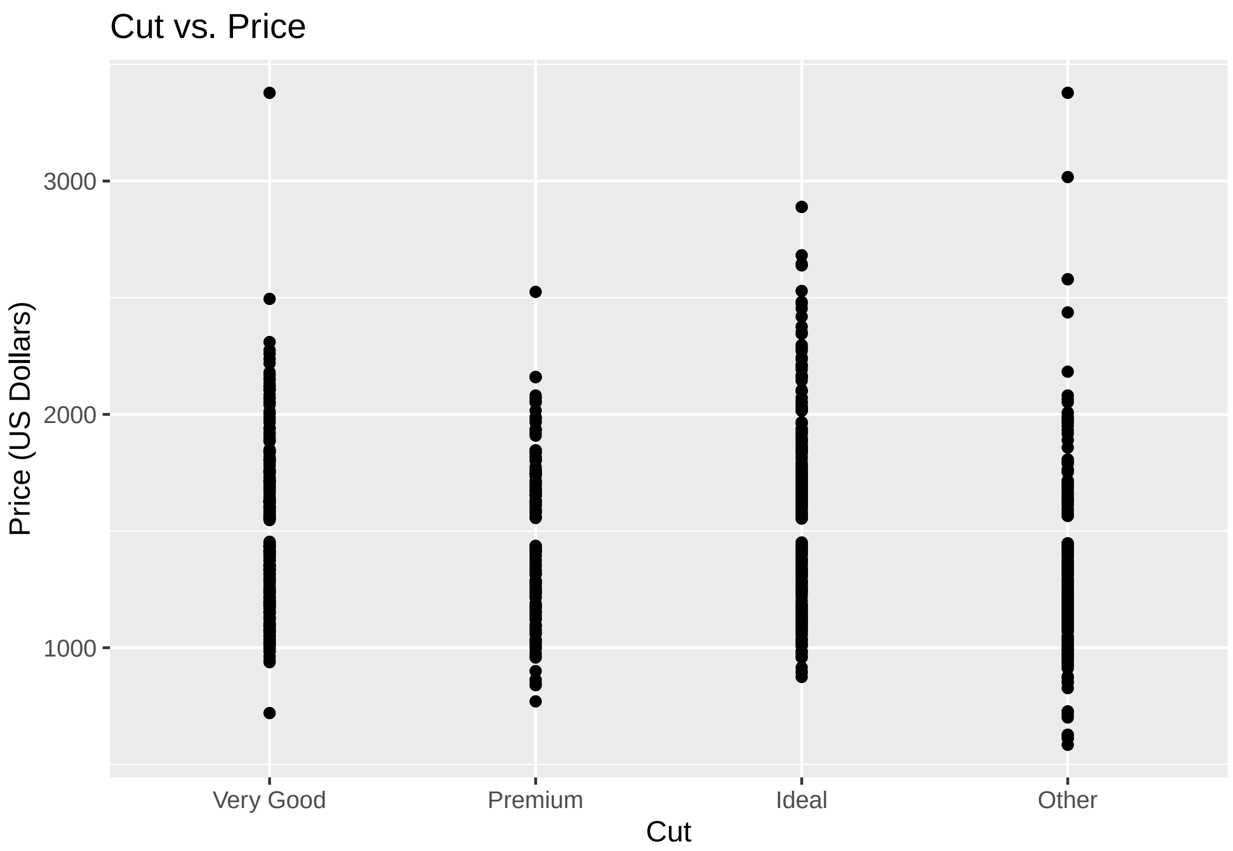
```r
diamonds <- diamonds %>% mutate(cut = fct_lump_n(cut, n = 3))
ggplot(diamonds, aes(x = fct_infreq(cut))) +
  geom_bar() +
  labs(x = "Cut", y = "Observations", title = "Cut Vs Observations") +
  coord_flip()
```



Exercise 4

```
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_point(aplha = 0.5) +
  labs(x = "Cut", y = "Price (US Dollars)", title = "Cut vs. Price")
```

```
## Warning: Ignoring unknown parameters: aplha
```



Exercise 5

```
diamonds %>%
  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
```

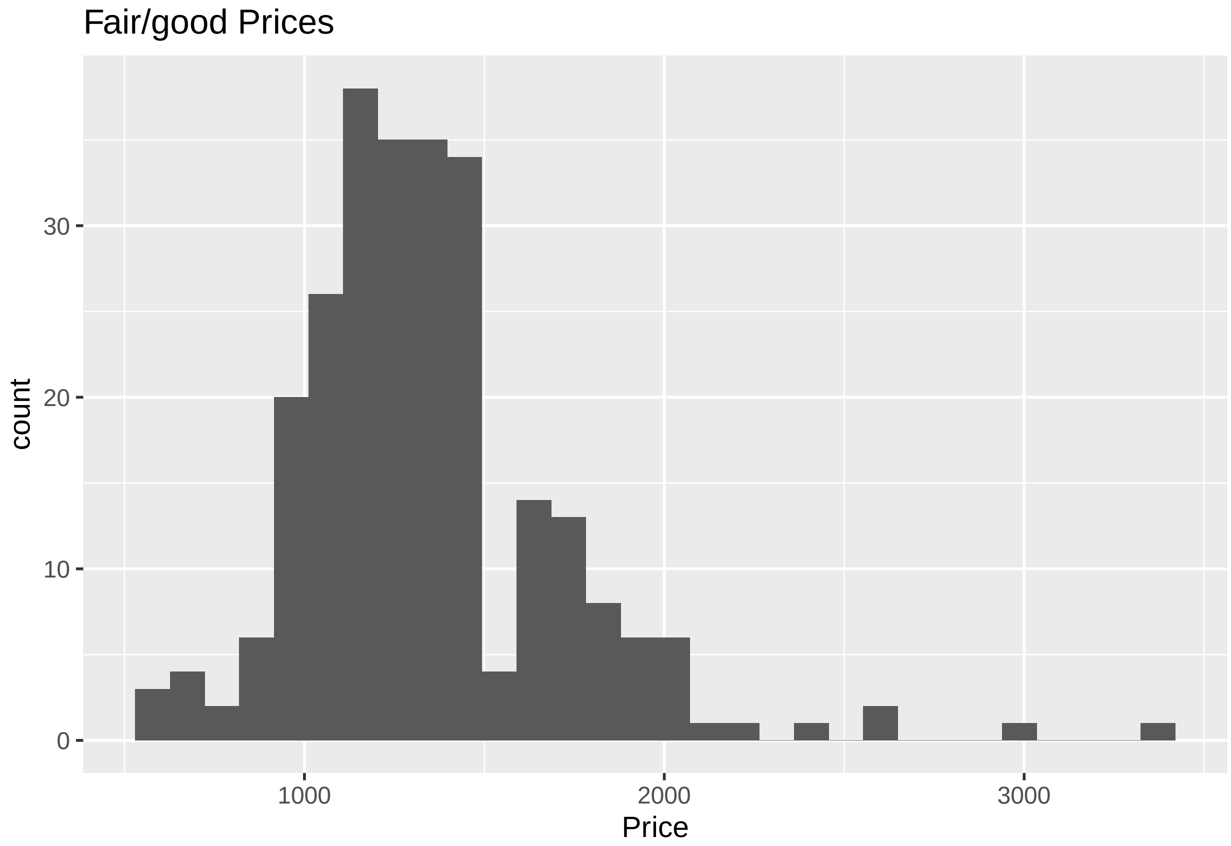
The table above calculates the number of observations along with the mean and standard deviation of price for each level of cut. Exercise 6 For Diamonds that are 0.5 carats, I would say there is a loose linear relationship between price and cut. Mean price increases as cut goes from other(fair/good), very good, ideal, to premium. However, given the context of this problem, I would be hesitant to say there is a clear linear relationship between price and cut.

#Analysis of Variance

To check the normality assumption, we can create histograms for the four levels of cuts against price (compare mean price across levels of cut). If the plots follow a normal distribution, the normality assumption is satsified.

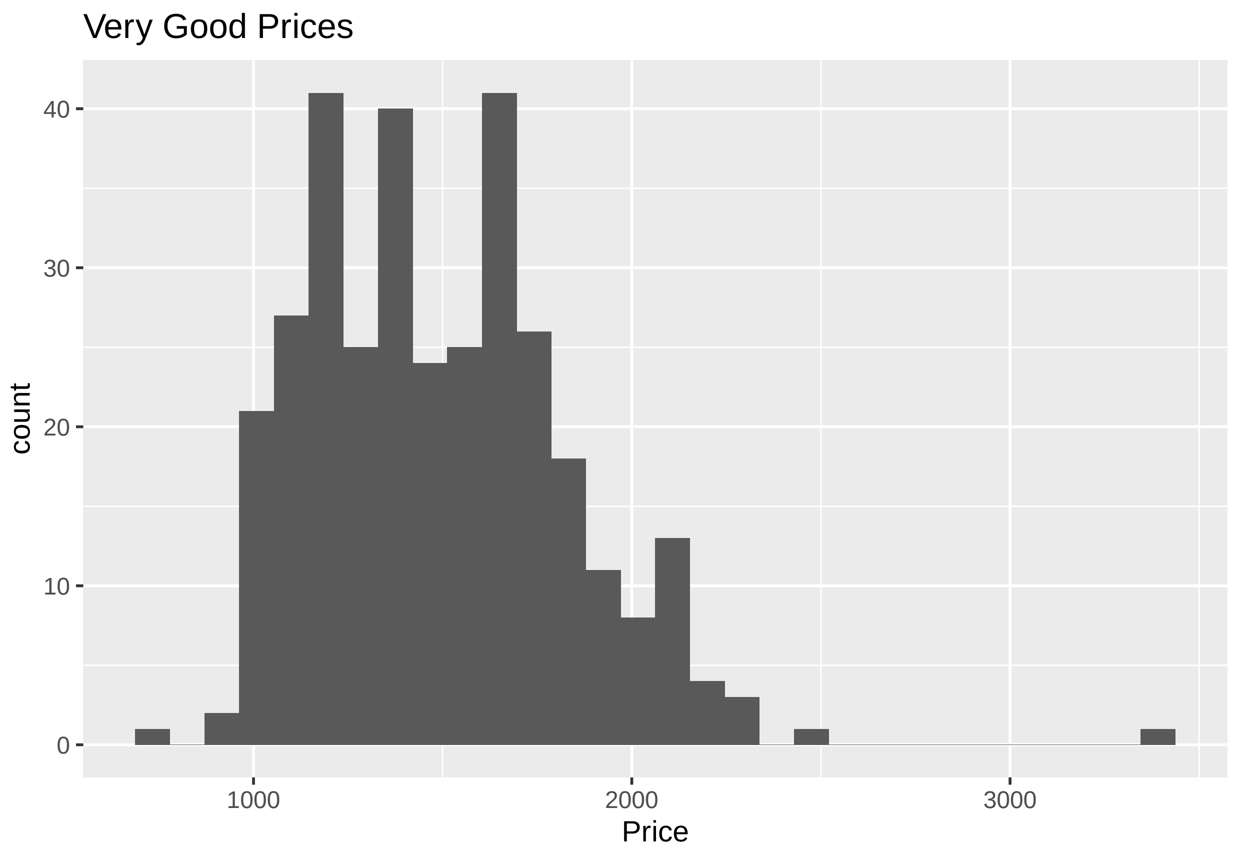
```
fair_good_cut <- diamonds %>%
  filter(cut == "Other")
ggplot(data = fair_good_cut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Fair/good Prices ",
       x = "Price")
```

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



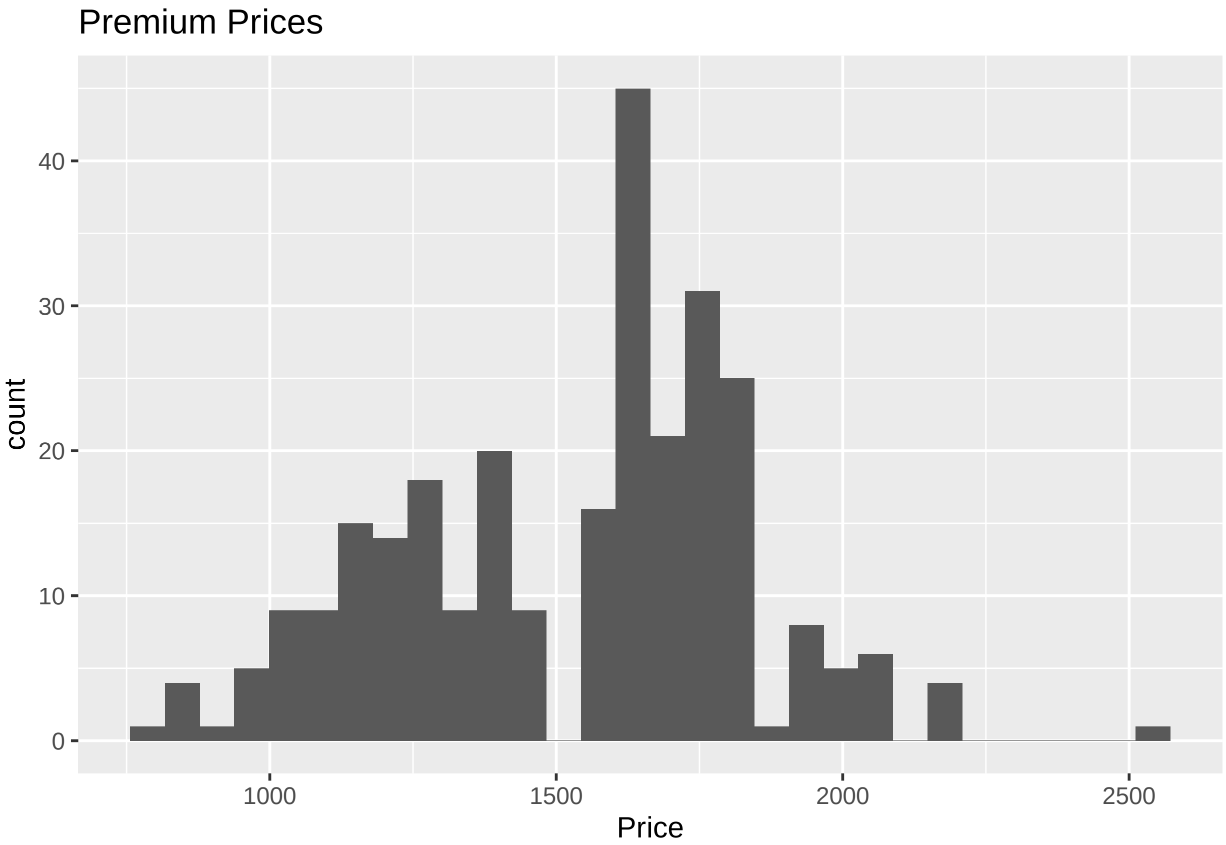
```
very_good_cut <- diamonds %>%
  filter(cut == "Very Good")
ggplot(data = very_good_cut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Very Good Prices",
       x = "Price")
```

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



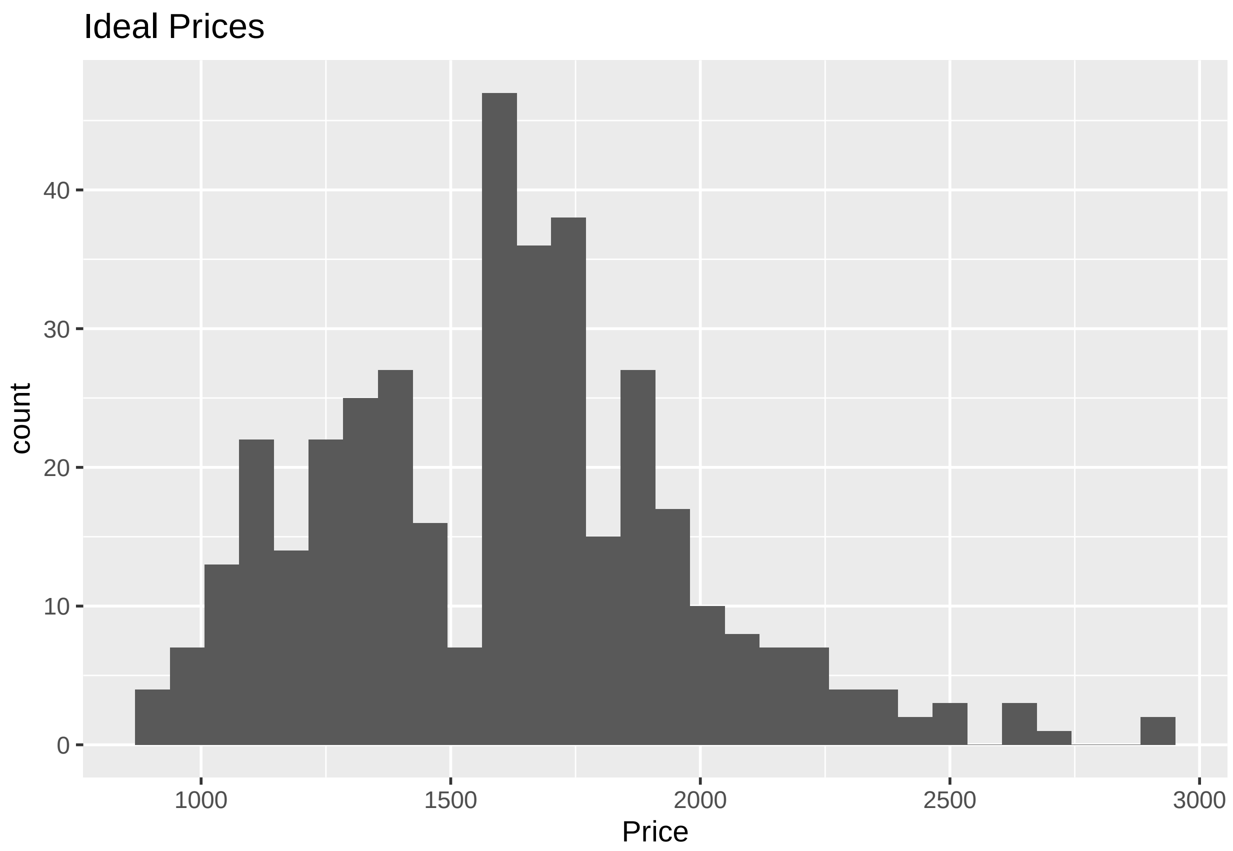
```
premium_cut <- diamonds %>%
  filter(cut == "Premium")
ggplot(data = premium_cut, aes(x = price)) +
  geom_histogram() +
  labs(title = "Premium Prices",
       x = "Price")
```

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



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

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



I would say the normality assumption is satisfied because these plots all follow a normal distribution, respectively.

There is no formal test to check for independence, so we have to inspect how our data was collected. I would say the independence assumption is satisfied because each observation are not dependent on another.

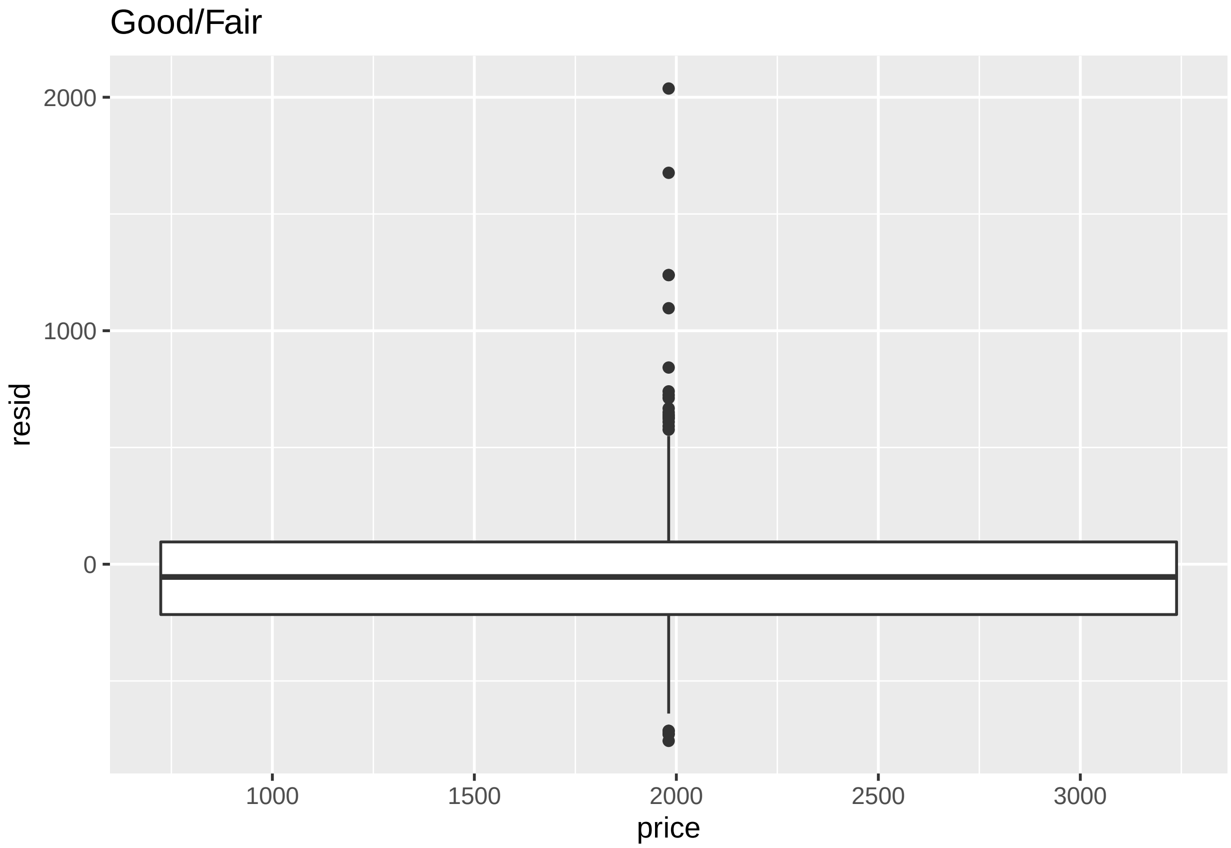
To test the constant variance assumption, we can create a model and adapt our code above to plot price against the residuals. We can visually inspect the variance by inspecting the length of each boxplot.

```
model <- lm(price ~ cut, data = diamonds)
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

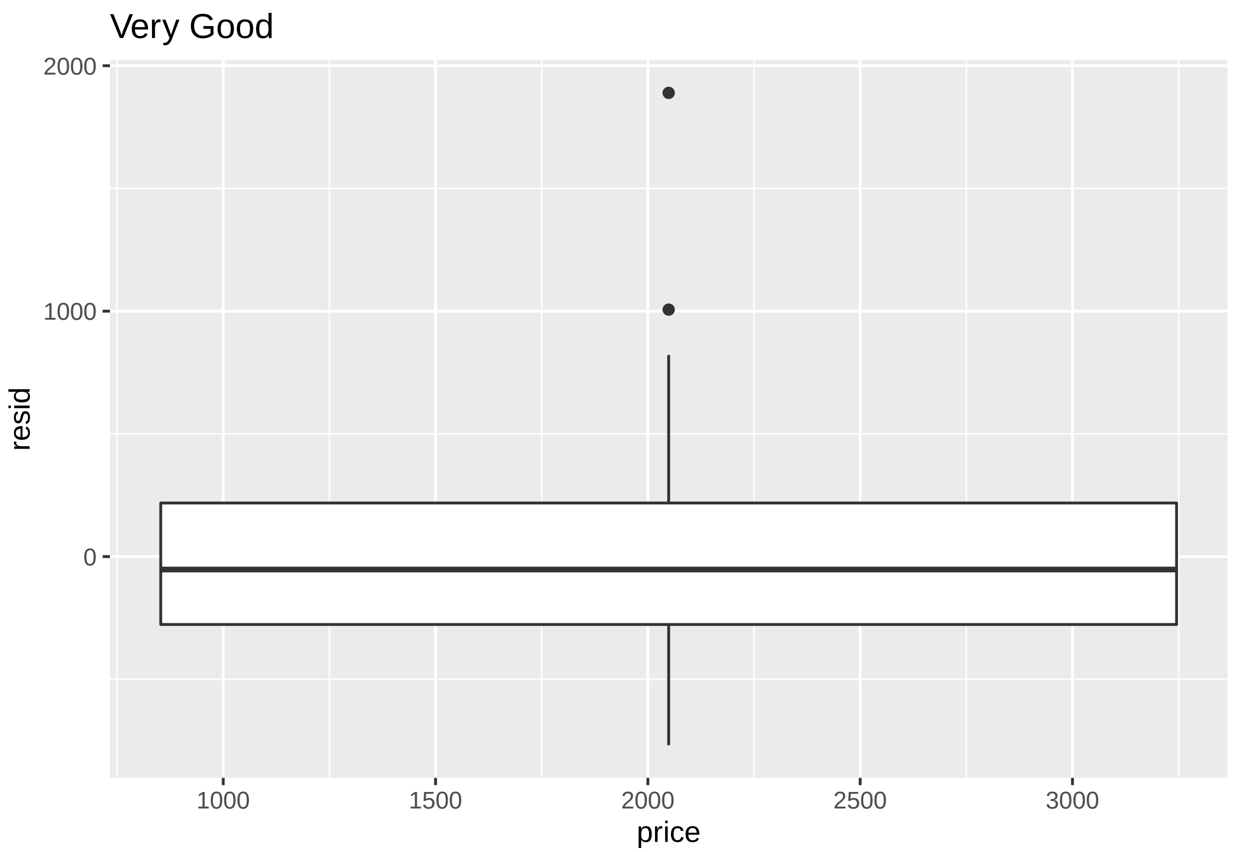
```
diamonds <- diamonds %>%
  mutate(resid = residuals(model))

fair_good_cut <- diamonds %>%
  filter(cut == "Other")
ggplot(data = fair_good_cut, aes(x = price, y = resid)) +
  geom_boxplot() + labs(title = "Good/Fair")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



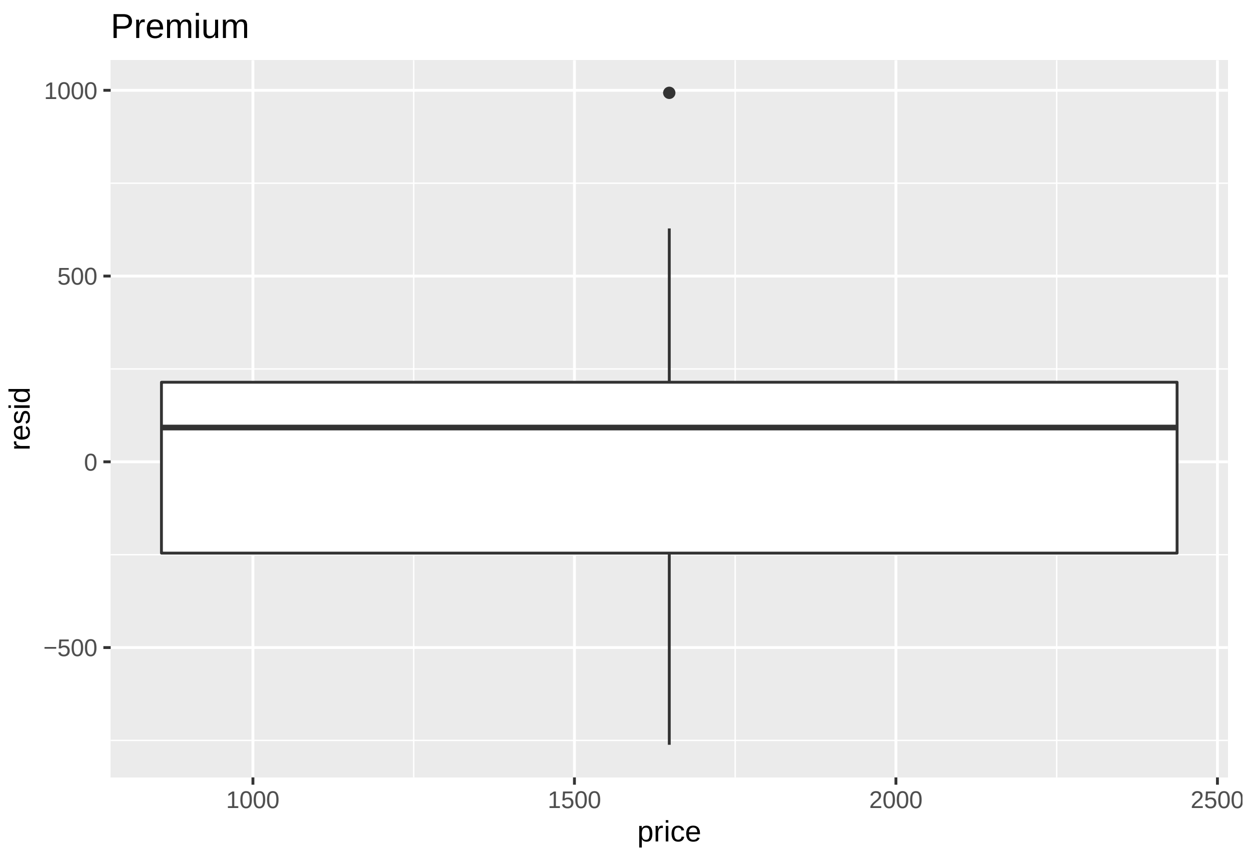
```
very_good_cut <- diamonds %>%
  filter(cut == "Very Good")
ggplot(data = very_good_cut, aes(x = price, y = resid)) +
  geom_boxplot() + labs(title = "Very Good")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



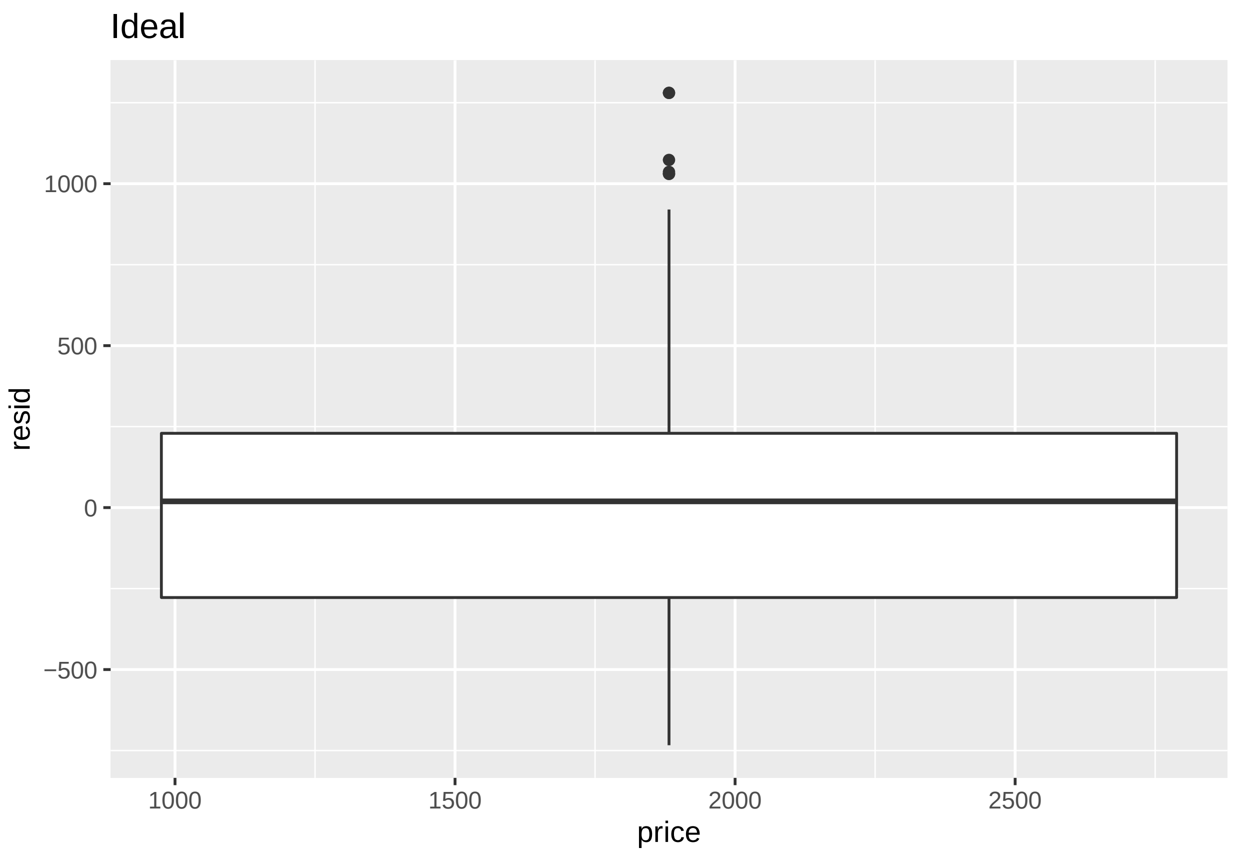
```
premium_cut <- diamonds %>%
  filter(cut == "Premium")
ggplot(data = premium_cut, aes(x = price, y = resid)) +
  geom_boxplot() + labs(title = "Premium")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



```
ideal_cut <- diamonds %>%
  filter(cut == "Ideal")
ggplot(data = ideal_cut, aes(x = price, y = resid)) +
  geom_boxplot() + labs(title = "Ideal")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



I would say that the constant variance assumption is not satisfied because the lengths of each boxplot are not the same. If the sample sizes for each cut level were the same, the variance assumption could be satisfied, but in this case they are not.

Exercise 8

```
model <- lm(price ~ cut, data = diamonds)
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, the sample mean for price = The sample variance can be calculated by dividing the sum of squares by the number of observations. According to the ANOVA table, the sum of squares for price is = 150706506. In exercise 5, we calculated the number of observations for each level of cut. Total observations = 332 + 277 + 388 + 261 = 1258. The code below calculates the sample variance for price based off values in the ANOVA table

```
sum_of_sqaures <- 150706506
obs_num <- 1258
sample_variance_price <- sum_of_sqaures/obs_num
sample_variance_price
```

```
## [1] 119798.5
```

Therefore, the sample variance for price is 119798.5.

Exercise 10

```
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

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

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

ideal cut variance = 9.893372, premium cut variance = 19.384185, very good cut variance = 19.786744, good/fair cut variance = 20.181275.

Exercise 11 Math notation null_hypothesis = (mean(price: cut = ideal) = (mean(price: cut = premium) = (mean(price: cut = very good) = (mean(price: cut = other). alternative_hypothesis = !(mean(price: cut = ideal) = (mean(price: cut = premium) = (mean(price: cut = very good) = (mean(price: cut = other) The null hypothesis states that means of price are the same among each level of cut. The alternative hypothesis states that at least one price mean is not equal to other price means among each level of cut.

Exercise 12 I would reject the null hypothesis and accept the alternative hypothesis because there seems to be a statistically significant relationship between price and level of cut.

Exercise 13 I would say no further statistical analysis is required because there is a statistically significant relationship between price and level of cut.