lab04

GITHUB : https://github.com/theeho/lab4

```
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)
```

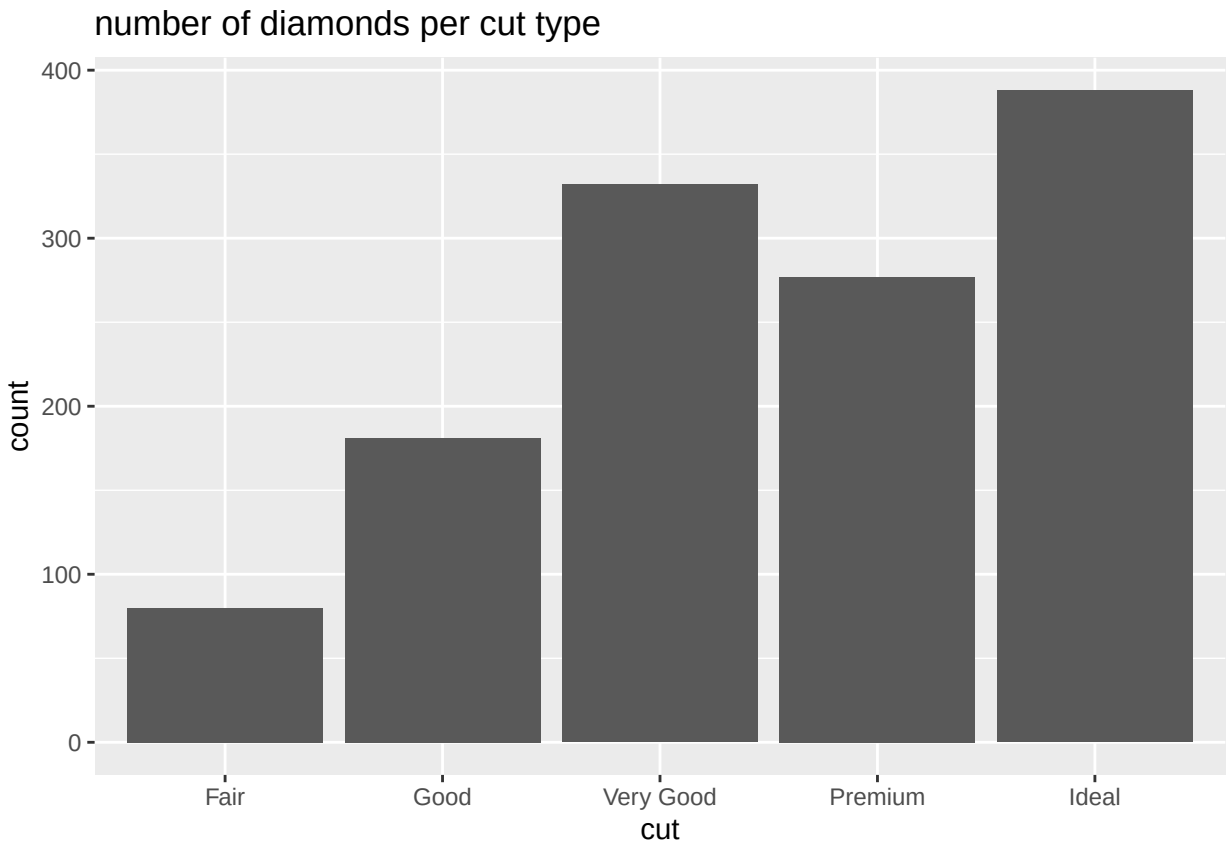
```
data <- diamonds
glimpse(diamonds)
```

```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI1, VS1, ~
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, 59.4, 64~
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54, 62, 58~
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339, 340, 34~
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, 4.00, 4.~
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, 4.05, 4.~
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, 2.39, 2.~
```

```
d_sub <- subset(data,  carat == .5)
```

There are 1258 observations in the dataset where carats are .5

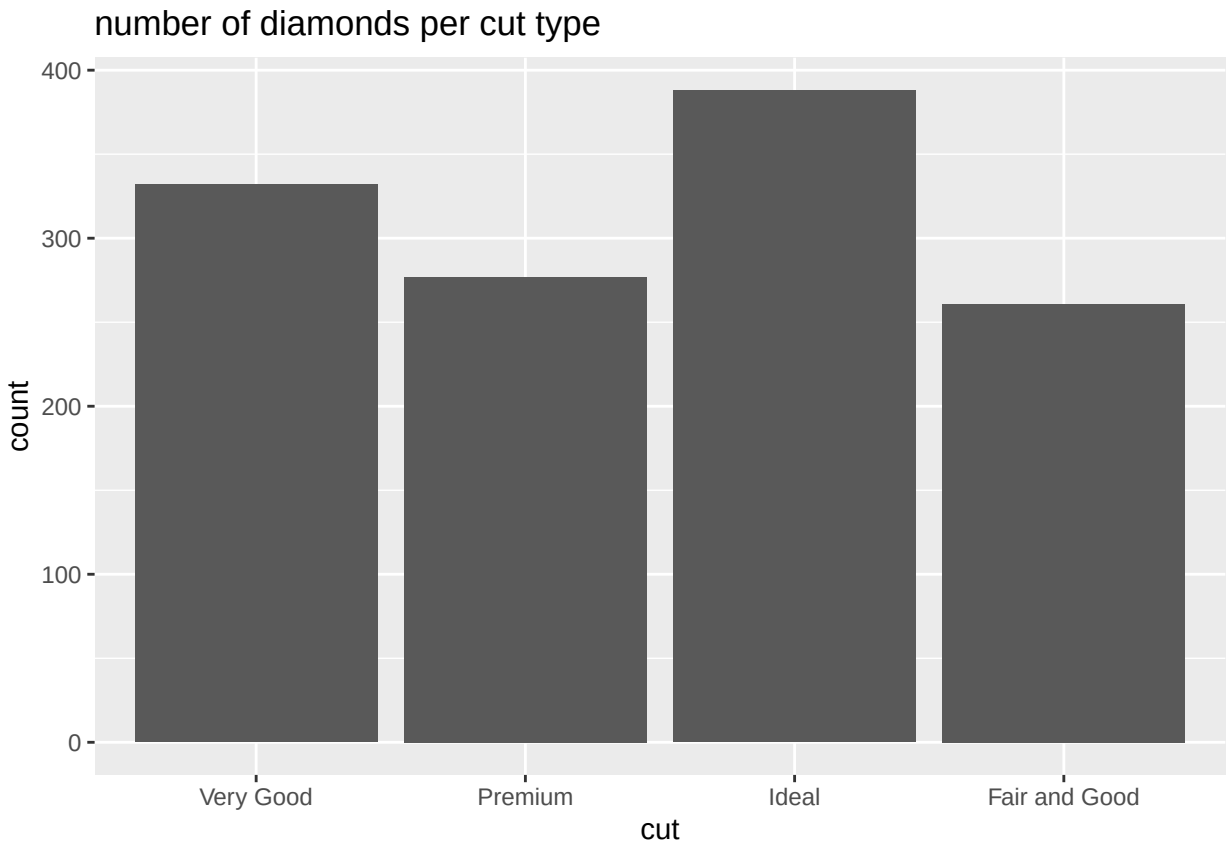
```
ggplot(data = d_sub, aes(x=cut)) +
         geom_bar() + labs(title = "number of diamonds per cut type")
```



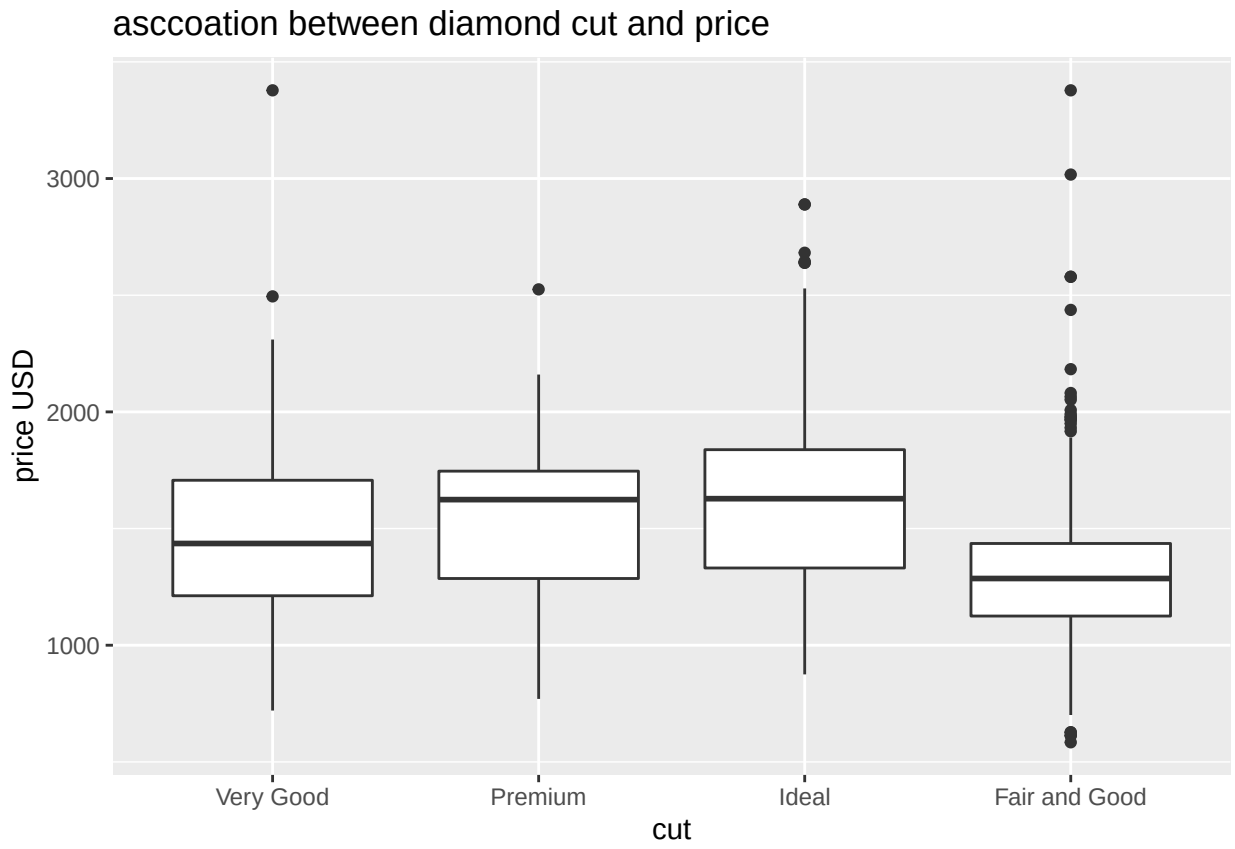
Cuts of fair and good have the fewest observations.

```
combine <- d_sub %>%
  mutate(cut = fct_lump(cut, n = 3, other_level = "Fair and Good"))
```

```
ggplot(data = combine, aes(x=cut)) +
         geom_bar() + labs(title = "number of diamonds per cut type")
```




```
ggplot(data = combine, aes(x=cut, y = price)) + geom_boxplot() + labs(title = "asccoation between diamo
```




```
m1 <- lm(price~cut, data = combine)
summary(m1)
```

```
##
## Call:
## lm(formula = price ~ cut, data = combine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -768.66 -251.41   14.83  214.22 2037.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1492.438      9.893 150.852  < 2e-16 ***
## cut.L        -82.101     20.181  -4.068 5.03e-05 ***
## cut.Q       -155.569     19.787  -7.862 8.08e-15 ***
## cut.C        -84.678     19.384  -4.368 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 346.7 on 1254 degrees of freedom
## Multiple R-squared:  0.07094,    Adjusted R-squared:  0.06872
## F-statistic: 31.92 on 3 and 1254 DF,  p-value: < 2.2e-16
```

```
vg <- combine %>% filter(cut == "Very Good")
p <- combine %>% filter(cut == "Premium")
i <- combine %>% filter(cut == "Ideal")
fg <- combine %>% filter(cut == "Fair and Good")
summary(combine)
```

```
##      carat               cut        color       clarity        depth      
##  Min.   :0.5   Very Good    :332   D:233   VS2    :361   Min.   :55.30  
##  1st Qu.:0.5   Premium      :277   E:362   SI1    :323   1st Qu.:61.20  
##  Median :0.5   Ideal        :388   F:248   VS1    :184   Median :62.10  
##  Mean   :0.5   Fair and Good:261   G:238   VVS2   :149   Mean   :62.04  
##  3rd Qu.:0.5                       H:107   SI2    :147   3rd Qu.:62.90  
##  Max.   :0.5                       I: 44   VVS1   : 56   Max.   :79.00  
##                                    J: 26   (Other): 38                  
##      table           price             x               y               z        
##  Min.   :52.00   Min.   : 584   Min.   :4.850   Min.   :4.750   Min.   :2.940  
##  1st Qu.:56.00   1st Qu.:1235   1st Qu.:5.050   1st Qu.:5.050   1st Qu.:3.130  
##  Median :57.90   Median :1436   Median :5.080   Median :5.090   Median :3.160  
##  Mean   :57.81   Mean   :1504   Mean   :5.085   Mean   :5.085   Mean   :3.155  
##  3rd Qu.:59.00   3rd Qu.:1746   3rd Qu.:5.120   3rd Qu.:5.130   3rd Qu.:3.180  
##  Max.   :73.00   Max.   :3378   Max.   :5.360   Max.   :5.360   Max.   :5.060  
##
```

Summary Stats:

Very Good

```
mean(vg$price)
```

```
## [1] 1488.663
```

```
sd(vg$price)
```

```
## [1] 339.363
```

```
nrow(vg)
```

```
## [1] 332
```

Premium

```
mean(p$price)
```

```
## [1] 1531.776
```

```
sd(p$price)
```

```
## [1] 304.1443
```

```
nrow(p)
```

```
## [1] 277
```

Ideal

```
mean(i$price)
```

```
## [1] 1608.668
```

```
sd(i$price)
```

```
## [1] 368.3448
```

```
nrow(i)
```

```
## [1] 388
```

Fair

```
mean(fg$price)
```

```
## [1] 1340.644
```

```
sd(fg$price)
```

```
## [1] 364.5216
```

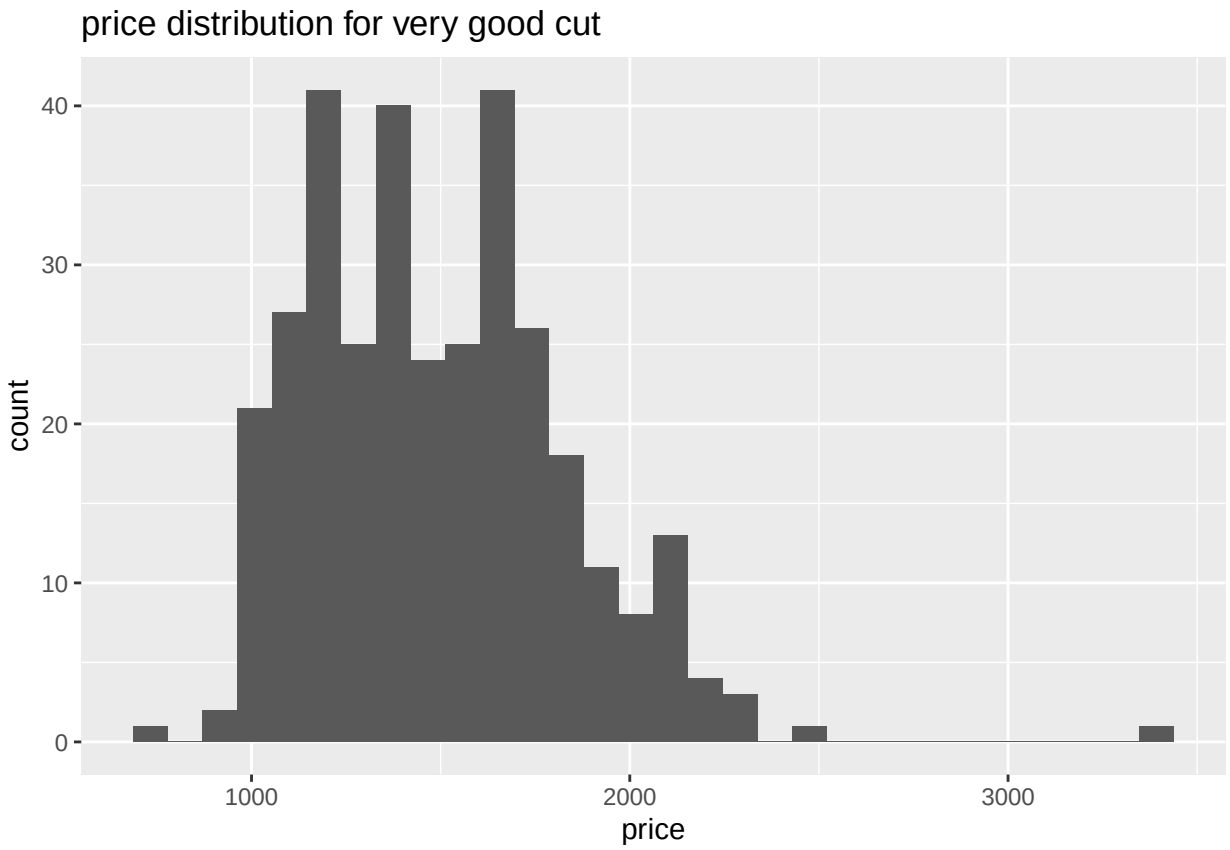
```
nrow(fg)
```

```
## [1] 261
```

Based on the graph and the summary statistics. There is some evidence to support an association between cut and price for diamonds that are .5 carrats as the means differ from eachother. However, more analysis of the assumptions of normality, independence and constant variance is needed.

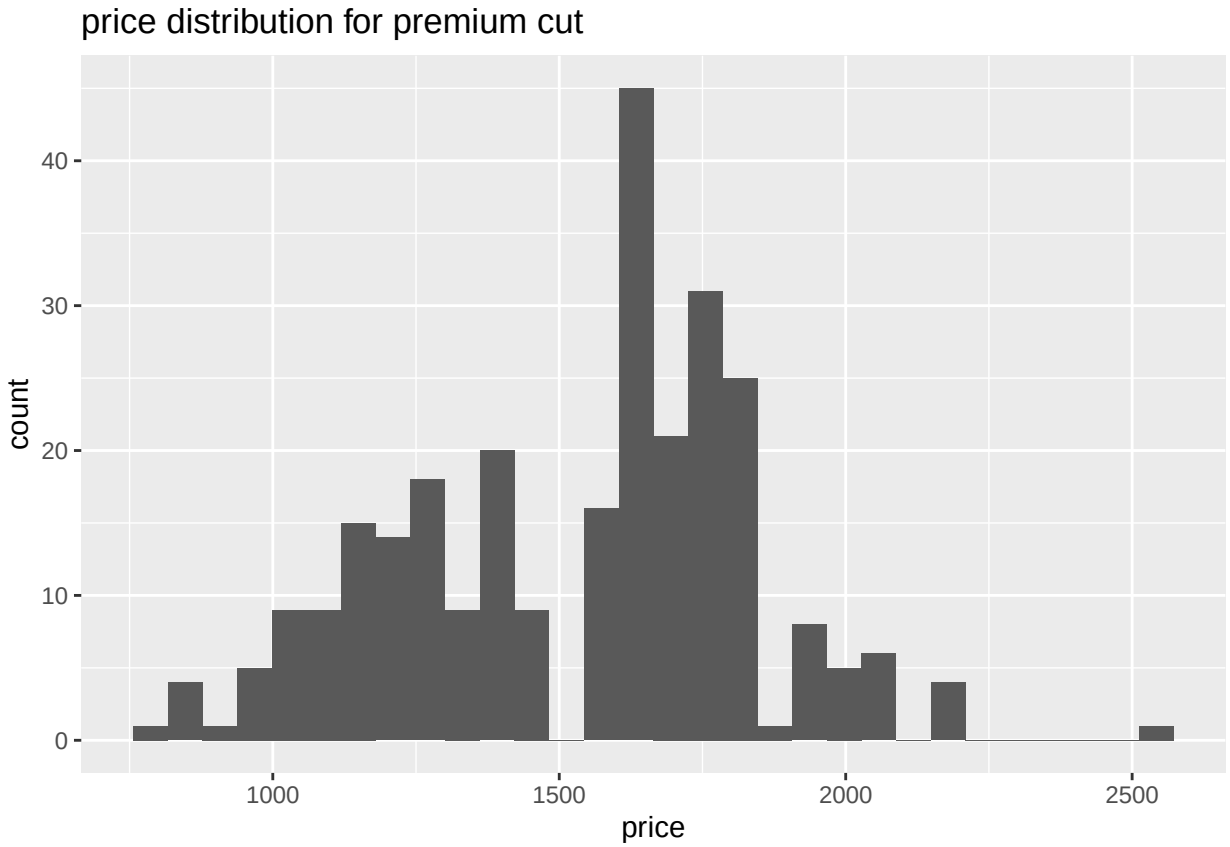
```
mvg <- ggplot(data = vg, aes(x=price)) +
         geom_histogram() + labs(title = "price distribution for very good cut")
mp <- ggplot(data = p, aes(x=price)) +
         geom_histogram() + labs(title = "price distribution for premium cut")
mi <- ggplot(data =i, aes(x=price)) +
         geom_histogram() + labs(title = "price distribution for ideal cut")
mfg <- ggplot(data = fg, aes(x=price)) +
         geom_histogram() + labs(title = "price distribution for fair cut")

mvg
```

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



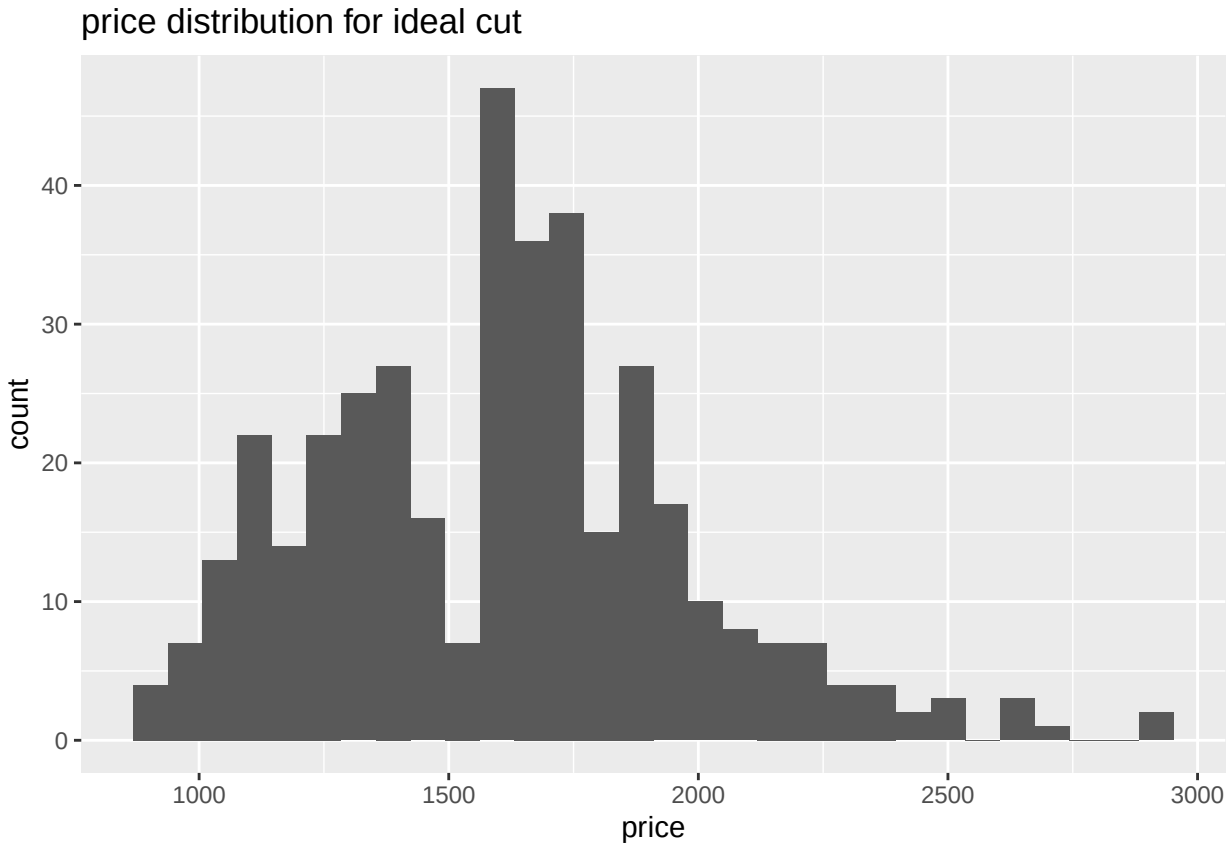

```
mp
```

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



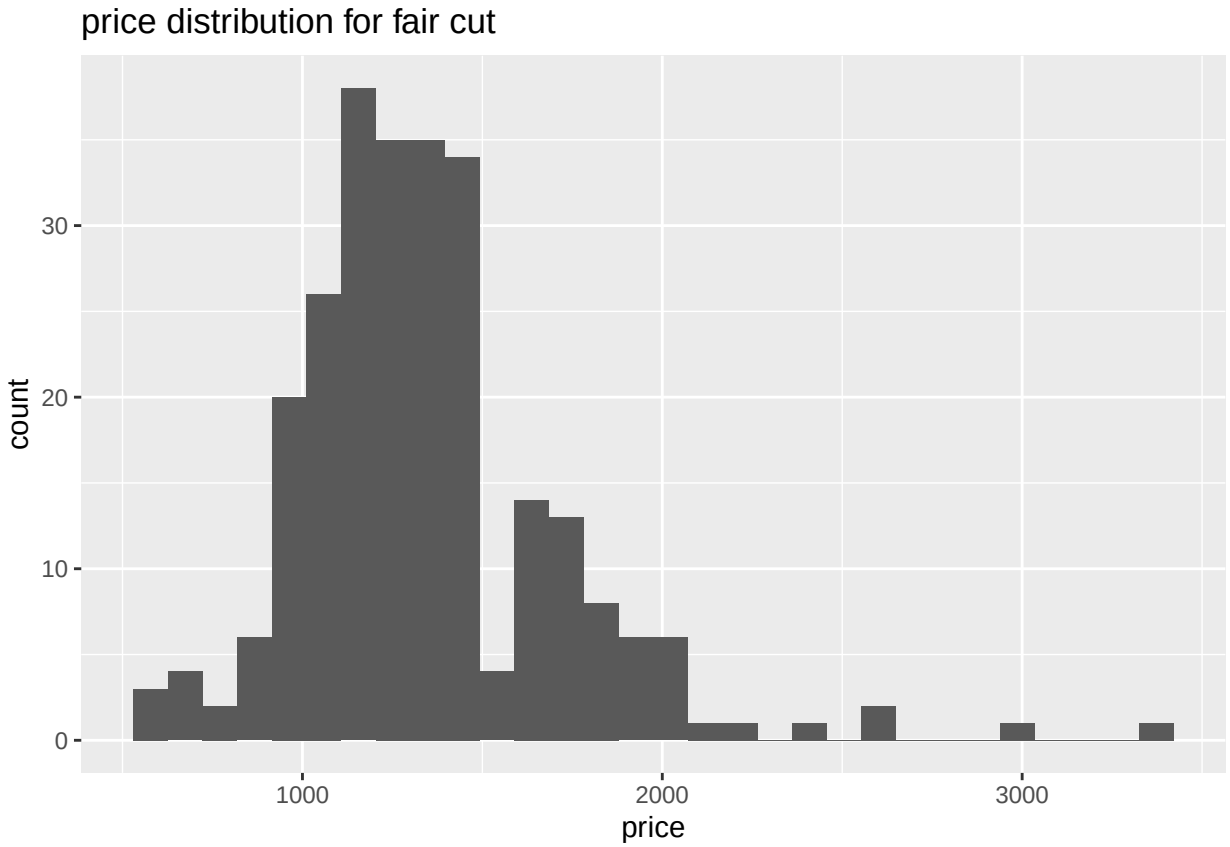

```
mi
```

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




```
mfg
```

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



Normality: All of the distributions seems to be fairly normal. Although, cuts that are fair and very good seem to be more uniform than normal.

Independence: Based on the quality of the dataset and context that diamond prices are usually independent it can be assumed that independence assumption is satasfied.

Constant Variance: The variance differs slightly between the groups. The range of standard deviations is [304, 368]. But considering the range of prices is [584, 3387] I think this assumption of constant variance can be satasified.

```
anova <- aov(price ~ cut, data = combine)
summary(anova)
```

```
##               Df    Sum Sq Mean Sq F value Pr(>F)
## cut            3  11507056 3835685   31.92 <2e-16 ***
## Residuals   1254 150706506  120181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9) Total sum of sqauares = 11507056 + 150706506 = 162213562 Sample variance = 162213562/(n-1) = 129048

10) fair sd = 364, variance = 132496 very good sd = 339, variance = 114921 ideal sd = 368, variance = 135424 premium sd = 304, variance = 92416

11) Null hypothesis: There is no association between cut and price. That is, the mean for price is the no different between different cuts.

Hypothesis: There is an association between cut and price. That is, the mean for price is different for different cuts.

12)

I conclude that there is evidence which supports my hypothesis. The P value is near zero for the anova model. The variance between the model and the data supports the hypothesis that there is some difference in price ascociated with cut.

13)

The anova does not give us any information about the individual levels. It doesn't tell us the difference between the price between the levels, in only tells us how well our model accounts for the variance in the data.