```
library(ggplot2)

#load diamonds dataset
data(diamonds)


#data
data("diamonds")


#view first six rows of diamonds dataset
head(diamonds)
```

A tibble: 6 × 10

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

```
#summarize diamonds dataset
summary(diamonds)
```

```
     carat               cut        color        clarity          depth      
 Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00  
 1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
 Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
 Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75  
 3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
 Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
                                    J: 2808   (Other): 2531                  
     table           price            x                y         
 Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720  
 Median :57.00   Median : 2401   Median : 5.700   Median : 5.710  
 Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735  
 3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540  
 Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  

       z         
 Min.   : 0.000  
 1st Qu.: 2.910  
 Median : 3.530  
 Mean   : 3.539  
 3rd Qu.: 4.040  
 Max.   :31.800  
```

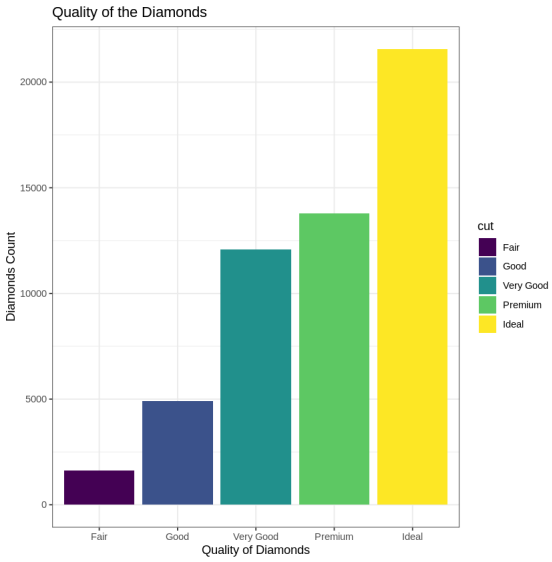
###For each of the numeric variables we can see the following information:

Min: The minimum value.
1st Qu: The value of the first quartile (25th percentile).
Median: The median value.
Mean: The mean value.
3rd Qu: The value of the third quartile (75th percentile).
Max: The maximum value.
For the categorical variables in the dataset (cut, color, and clarity) we see a frequency count of each value.#

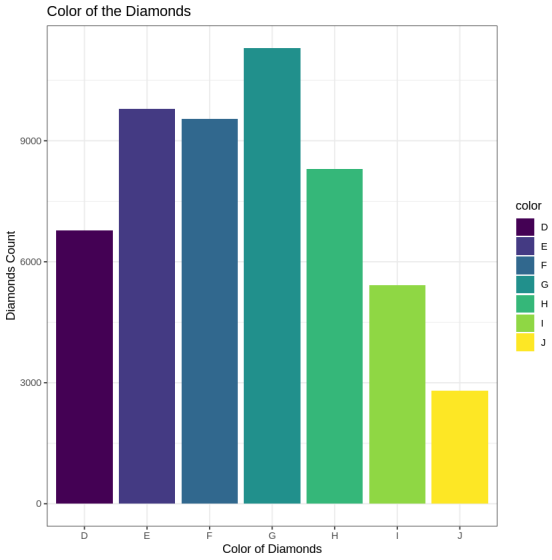
#For example, for the cut variable:

#Fair: This value occurs 1,610 times.
Good: This value occurs 4,906 times.
Very Good: This value occurs 12,082 times.
Premium: This value occurs 13,791 times.
Ideal: This value occurs 21,551 times.###

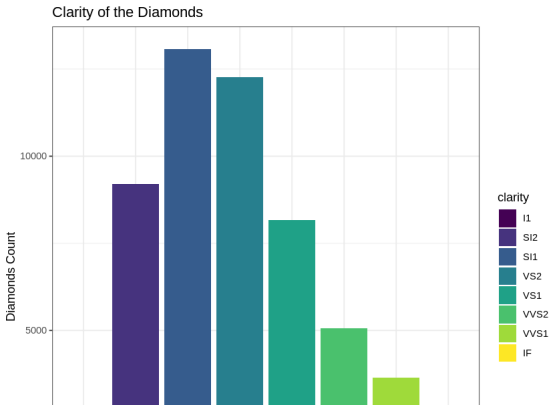
```
# plot the cut (quality) of diamonds (quality <-  cut <- fair,good,very good, premium, ideal)
ggplot(diamonds, aes(x = cut , fill = cut)) +
  theme_bw() +
  geom_bar()+
  labs(x = "Quality of Diamonds",
       y = "Diamonds Count",
       title = "Quality of the Diamonds")
```




```
# plot the color of diamonds (color<- D(best),E,F,G,H,I,J(WORST))
ggplot(diamonds, aes(x = color,fill = color)) +
  theme_bw()+
  geom_bar()+
  labs(x="Color of Diamonds",
       y="Diamonds Count",
       title="Color of the Diamonds")
#There's less of the lower quality diamonds as we would expect! Overall there are less I & J diamonds (lesser quality) than the higher quality diamonds. The
#There seems to be a close to even distribution between diamonds that are colored between E, F, and G. This may mean that customer demand is most for diamon
```

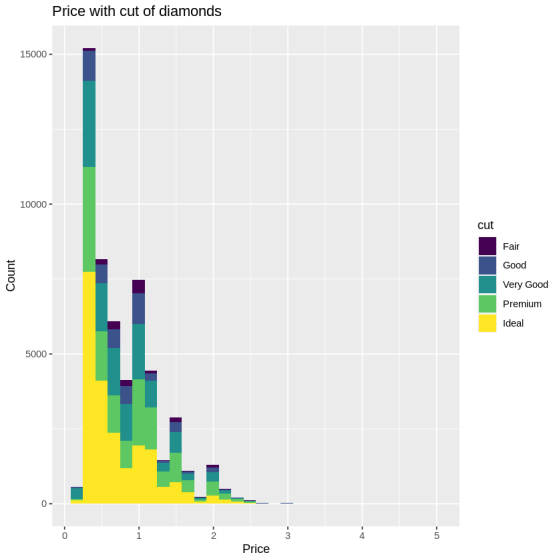



```
# plot the clarity of diamonds (clarity <- l1(WORST), SL2,SL1,VS2,VS1,VVS2,VVS1,LF(BEST))
ggplot(diamonds, aes(x = clarity,fill = clarity)) +
  theme_bw()+
  geom_bar()+
  labs(x = "Clarity of Diamonds",
       y = "Diamonds Count",
       title = "Clarity of the Diamonds")
#There's more of the lesser quality diamonds as we would expect! As the quality increases, the overall demand appears to decrease.
#I think this is due to the fact that only people who have more money can afford to buy the higher priced diamonds.
#That would make sense, since there are less people with more money overall that can afford a more expensive diamond.
```

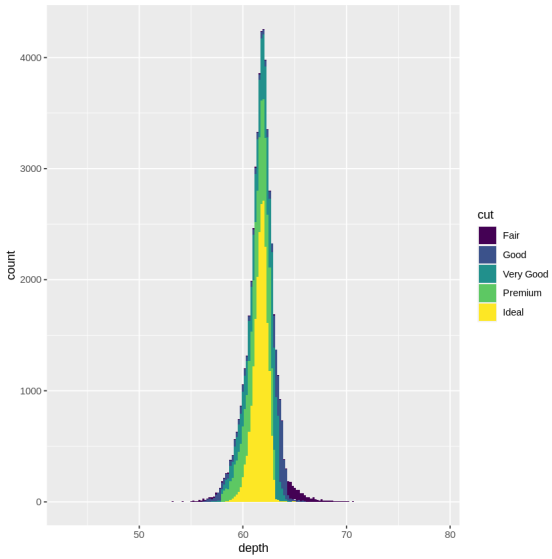



```
##histogram (price with cut of Diamonds )
ggplot(diamonds, aes(x=carat, fill=cut)) +
  geom_histogram()+
  labs(y="Count",
       x="Price",
       title="Price with cut of diamonds")
```

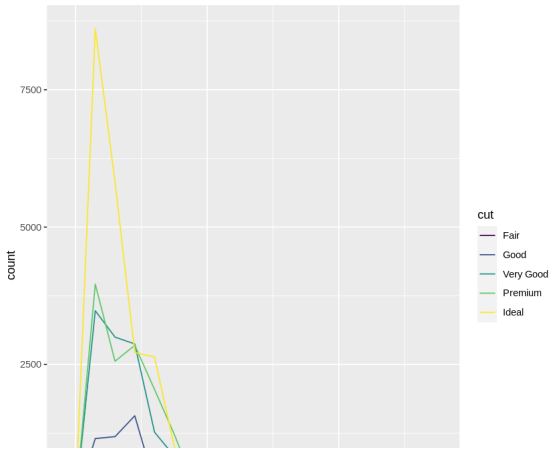
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data = diamonds, aes(x = depth, fill = cut)) +
  geom_histogram(binwidth = 0.2)
```

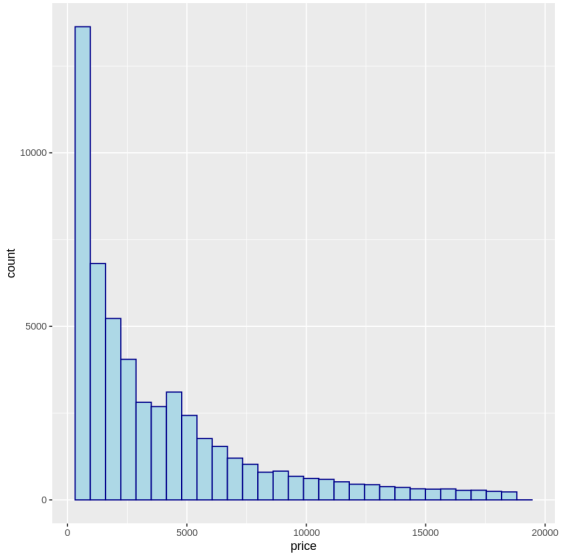



```
ggplot(data = diamonds, mapping = aes(x = carat, colour = cut)) +
  geom_freqpoly(binwidth = 0.3)
```




```
ggplot(data = diamonds, mapping = aes(x=price))+
  geom_histogram(color="darkblue", fill="lightblue")
```

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



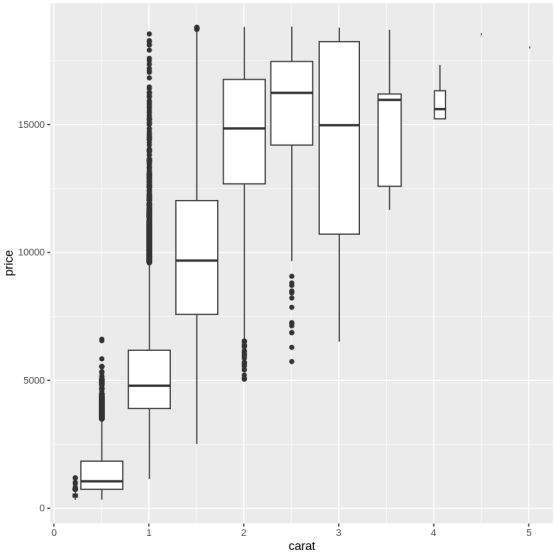

```
mean(diamonds$price) #Mean is $3,932.80
#Most diamonds are priced below $5,000 based on looking at the graphs
```

```
3932.79972191324
```

```
plt <- ggplot(diamonds,aes(x = carat, y = price)) +
  geom_point(
    aes(color = cut, shape = cut),
    size = 1.5,
    alpha = 0.8 # It's nice to add some transparency because there may be overlap.
  ) +
  # Use custom colors
  scale_color_manual(
    values = c("#386cb0", "#fdb462", "#7fc97f","#999999", "#E69F00", "#56B4E9")
  )

plt
#As you can see in the plot, it is obvious that with an increase in carat the price also increases, but due to a large number of data points, it creates an
#Overplot is when there are too many data points in a plot, making it very difficult to summarize the findings from the plot.
#Instead, let's try using a boxplot to divide the continuous data points into quartiles. In this example, you will take carat as a categorical variable and
```

```
Warning message:
“Using shapes for an ordinal variable is not advised”
```

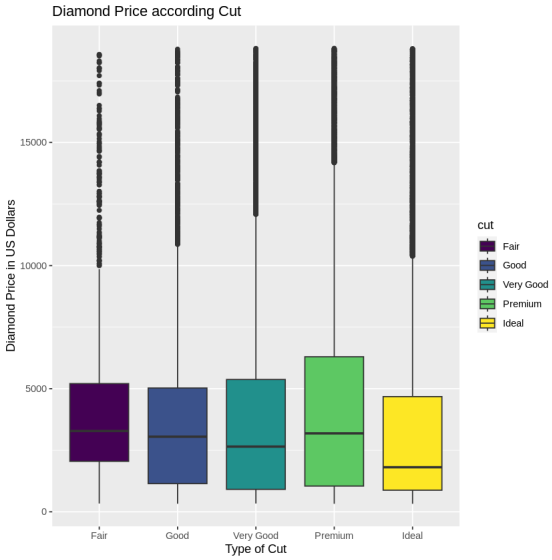

```
ggplot(data = diamonds, mapping = aes(x = carat, y = price)) +
  geom_boxplot(mapping = aes(group = cut_width(carat, 0.5)))

#From the above we see a few unusual data points:

#Some one carat diamonds have an exceptionally high price,

#The average price of three carat diamonds is relatively low.

#The data points above three carats can be ignored because they are not contributing much to the analysis.

#With this plot, we find the relationship between two categorical variables or one categorical and one continuous variable.
```



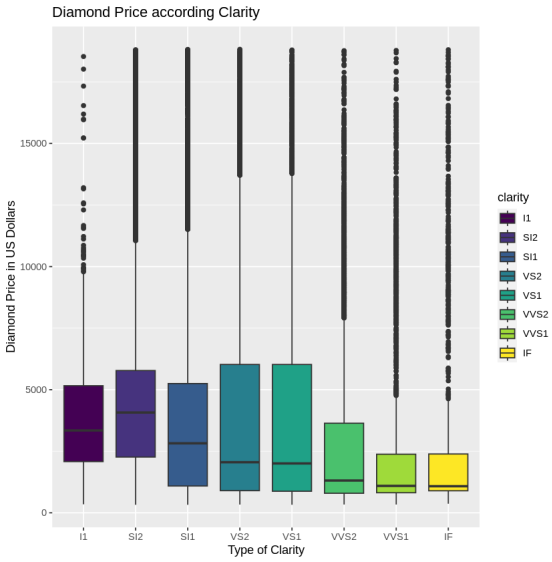

```
ggplot(data=diamonds, aes(x=carat, y=price)) +
  # get rid of top percentile as they could skew the data
  scale_x_continuous(lim=c(0,quantile(diamonds$carat,0.99))) +
  scale_y_continuous(lim=c(0,quantile(diamonds$price,0.99))) +
  geom_point(fill=I('#dd3333'), color= I("black"), aes(alpha=1/10),shape=21) +
  stat_smooth(method='lm')
#It appears that there is a positive, linear relationship between price and carat weight. We need to further investigate this
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
ggplot(diamonds, aes(factor(cut), price, fill=cut)) +
geom_boxplot() +
ggtitle("Diamond Price according Cut") +
xlab("Type of Cut") +
ylab("Diamond Price in US Dollars")
#It doesn't appear that Cut is a good way to determine the quality or whether or not a diamond will be expensive.
#Maybe it's a marketing thing or branding effect? It's yet another way to make folks feel like they are picking a higher quality diamond?
#Maybe it's a "feel better" about your purchase metric? It's interesting why it's included when it does not seem to affect price that much.
```




```
##Price of Diamonds by Clarity##
ggplot(diamonds, aes(factor(clarity), price, fill=clarity)) +
geom_boxplot() +
ggtitle("Diamond Price according Clarity") +
xlab("Type of Clarity") +
ylab("Diamond Price in US Dollars")
#Clarity is a meaningful variable as compared to cut based on the above.
```



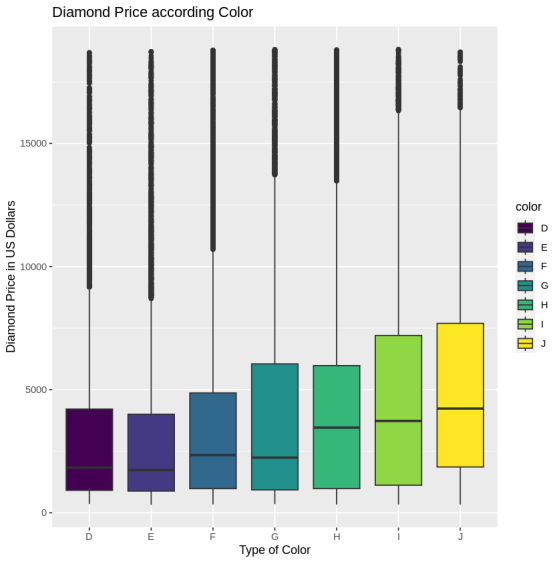

```
by(diamonds$price, diamonds$clarity, summary)
#It seems from the data above that maximum price of diamonds are quite similar among all based on clarity
```

```
diamonds$clarity: I1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    345    2080    3344    3924    5161   18531
------------------------------------------------------------
diamonds$clarity: SI2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326    2264    4072    5063    5777   18804
------------------------------------------------------------
diamonds$clarity: SI1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326    1089    2822    3996    5250   18818
------------------------------------------------------------
diamonds$clarity: VS2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334     900    2054    3925    6024   18823
------------------------------------------------------------
diamonds$clarity: VS1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
by(diamonds$price, diamonds$color, summary)
#Similary, minimum price of the diamonds are quite similary priced among diamonds based on color
```

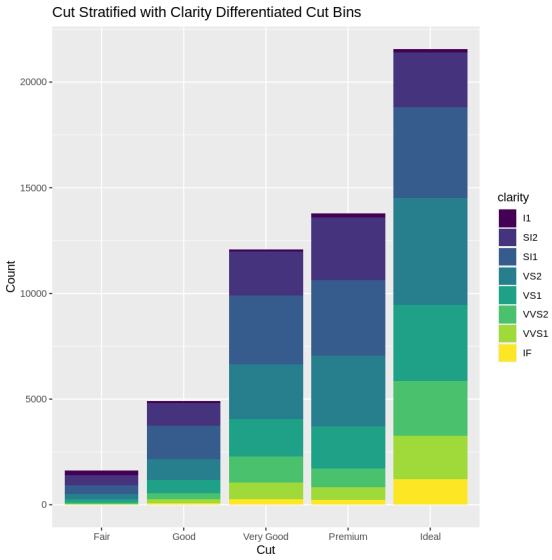
```
diamonds$color: D
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    357     911    1838    3170    4214   18693
------------------------------------------------------------
diamonds$color: E
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     882    1739    3077    4003   18731
------------------------------------------------------------
diamonds$color: F
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    342     982    2344    3725    4868   18791
------------------------------------------------------------
diamonds$color: G
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    354     931    2242    3999    6048   18818
------------------------------------------------------------
diamonds$color: H
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    337     984    3460    4487    5980   18803
------------------------------------------------------------
diamonds$color: I
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1120    3730    5092    7202   18823
------------------------------------------------------------
diamonds$color: J
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    335    1860    4234    5324    7695   18710
```

```
##Price per carat by color##
ggplot(diamonds, aes(factor(color), price, fill=color)) +
geom_boxplot() +
ggtitle("Diamond Price according Color") +
xlab("Type of Color") +
ylab("Diamond Price in US Dollars")
#Color looks like it makes a difference in the quality or whether or not a diamond will be expensive - as we would expect.
#Color is a meaningful variable as compared to cut.
```



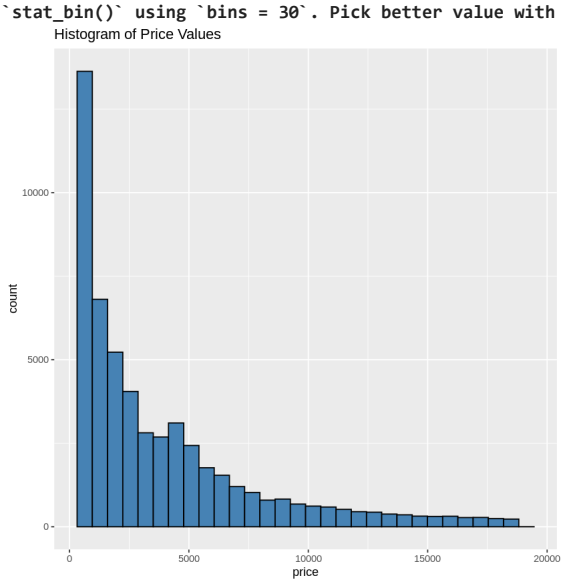

```
ggplot(data = diamonds) +
geom_bar(mapping = aes(x = cut, fill = clarity)) +
ggtitle("Cut Stratified with Clarity Differentiated Cut Bins") +
```

```
xlab("Cut") +
ylab("Count")
##This graph is useful in showing by types of Cut, what the distribution is by clarity.
#As we would expect, as the quality increases in clarity, it gets harder to find or becomes more rare.
#Though it's interesting that this graph could possibly show that maybe jewelers who cut diamonds target an ideal cut for all diamonds,
#but maybe for some reason if it doesn't work out to be an ideal cut, then it becomes a lesser cut?
```

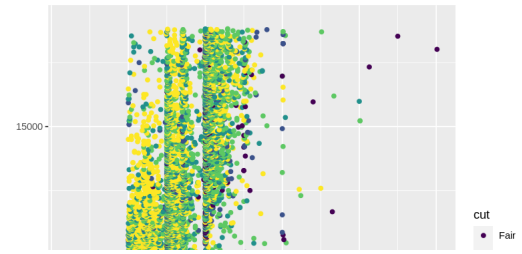



```
#After looking through these plots, it looks like the lesser quality diamonds seem to be more expensive?
#What explains this?
#I think the Carat weight comes into play here. Carat weight seems to be the single most determining factor in deciding the price of a diamond.**


#create histogram of values for price
ggplot(data=diamonds, aes(x=price)) +
  geom_histogram(fill="steelblue", color="black") +
  ggtitle("Histogram of Price Values")
```

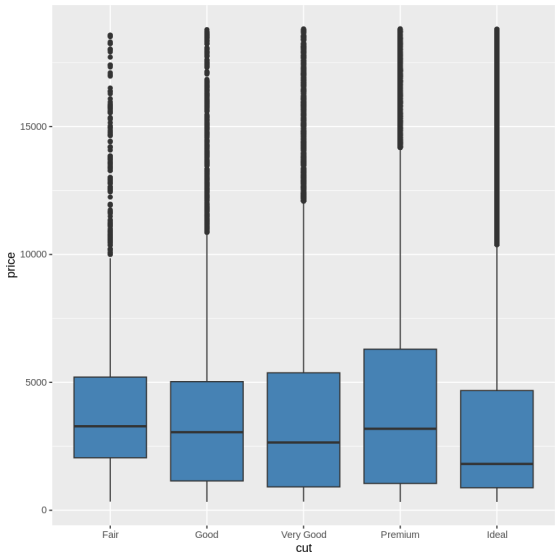
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.




```
#create scatterplot of carat vs. price, using cut as color variable
ggplot(data=diamonds, aes(x=carat, y=price, color=cut)) +
  geom_point()
##We can also use the geom_point() function to create a scatterplot of any pairwise combination of variables:##
```




```
#create scatterplot of price, grouped by cut
ggplot(data=diamonds, aes(x=cut, y=price)) +
  geom_boxplot(fill="steelblue")
##We can also use the geom_boxplot() function to create a boxplot of one variable grouped by another variable:##
```




```
#We can also use the cor() function to create a correlation matrix to view the correlation coefficient between each pairwise combination of numeric variable
#create correlation matrix of (rounded to 2 decimal places)
round(cor(diamonds[c('carat', 'depth', 'table', 'price', 'x', 'y', 'z')]), 2)
```

A matrix: 7 × 7 of type dbl

	carat	depth	table	price	x	y	z
carat	1.00	0.03	0.18	0.92	0.98	0.95	0.95
depth	0.03	1.00	-0.30	-0.01	-0.03	-0.03	0.09
table	0.18	-0.30	1.00	0.13	0.20	0.18	0.15
price	0.92	-0.01	0.13	1.00	0.88	0.87	0.86
x	0.98	-0.03	0.20	0.88	1.00	0.97	0.97
y	0.95	-0.03	0.18	0.87	0.97	1.00	0.95
z	0.95	0.09	0.15	0.86	0.97	0.95	1.00

```
#Conclusion
#From the analysis, there are four factors that affect the price of diamonds.
#These main factors are diamond's carat, its color, cut and clarity. However, the carat seems to be one factor that has the highest influence on the price o
#Colorless diamonds are rare which makes them expensive
#Premium cut diamonds have high prices and Fair cut diamonds generally have lower prices.
#Flawless Diamonds have high prices as expected and in general they are smaller in size (low carat value). They also turn to have ideal to premium cuts
```