EDA on diamonds dataset

Done by: Sahil Ansari

This dataset contains information about 53,940 round-cut diamonds. There are 10 variables measuring various pieces of information about the diamonds.

Contents of the dataset:

price in US dollars ($326-$18,823)
carat weight of the diamond (0.2-5.01)
cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
color from J (worst) to D (best)
clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
x length in mm (0-10.74)
y width in mm (0-58.9)
z depth in mm (0-31.8)
depth total depth percentage (43-79)
width of top of diamond relative to widest point (43-95)

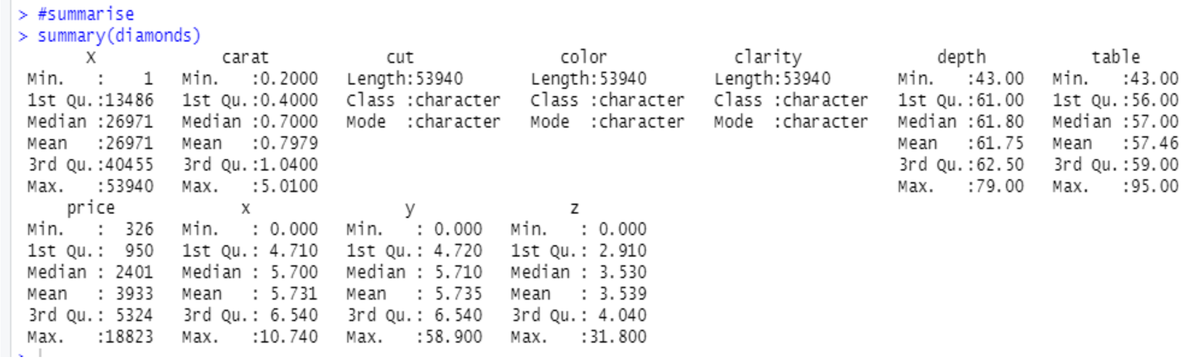
Exploratory Data Analysis :

#load the dataset

diamonds <- read.csv("C://Users//Asus//Desktop//Itvedant lectures//R//diamonds.csv")

# summarize

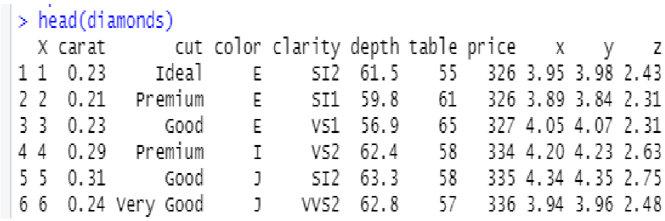
summary(diamonds)



```
> #summarise
> summary(diamonds)
       X             carat            cut               color             clarity             depth           table      
 Min.   :    1   Min.   :0.2000   Length:53940       Length:53940       Length:53940       Min.   :43.00   Min.   :43.00  
 1st Qu.:13486   1st Qu.:0.4000   Class :character   Class :character   Class :character   1st Qu.:61.00   1st Qu.:56.00  
 Median :26971   Median :0.7000   Mode  :character   Mode  :character   Mode  :character   Median :61.80   Median :57.00  
 Mean   :26971   Mean   :0.7979                                                            Mean   :61.75   Mean   :57.46  
 3rd Qu.:40455   3rd Qu.:1.0400                                                            3rd Qu.:62.50   3rd Qu.:59.00  
 Max.   :53940   Max.   :5.0100                                                            Max.   :79.00   Max.   :95.00  
     price             x                y                z         
 Min.   :  326   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median : 2401   Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   : 3933   Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :18823   Max.   :10.740   Max.   :58.900   Max.   :31.800  
```

#preview the dataset

head(diamonds)



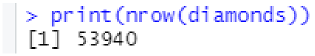
```
> head(diamonds)
  X carat       cut color clarity depth table price    x    y    z
1 1  0.23     Ideal     E     SI2  61.5    55   326 3.95 3.98 2.43
2 2  0.21   Premium     E     SI1  59.8    61   326 3.89 3.84 2.31
3 3  0.23      Good     E     VS1  56.9    65   327 4.05 4.07 2.31
4 4  0.29   Premium     I     VS2  62.4    58   334 4.20 4.23 2.63
5 5  0.31      Good     J     SI2  63.3    58   335 4.34 4.35 2.75
6 6  0.24 Very Good     J    VVS2  62.8    57   336 3.94 3.96 2.48
```

#load the required libraries

library(ggplot2)

library(dplyr)

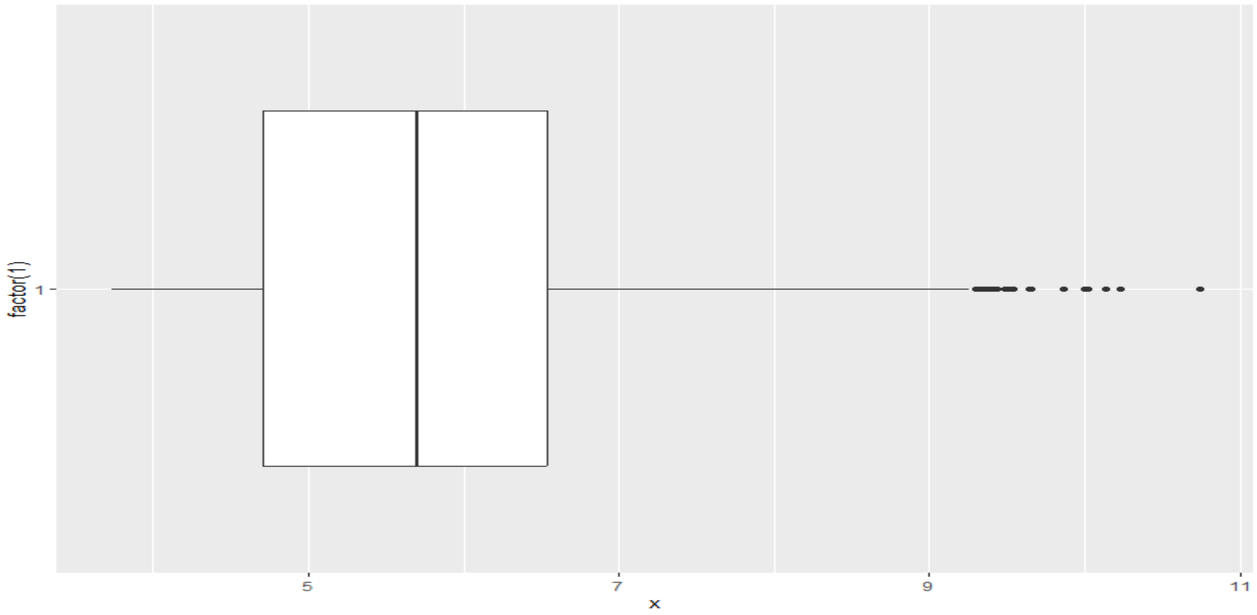
```
#checking the no. of rows in the dataset

print(nrow(diamonds))
```



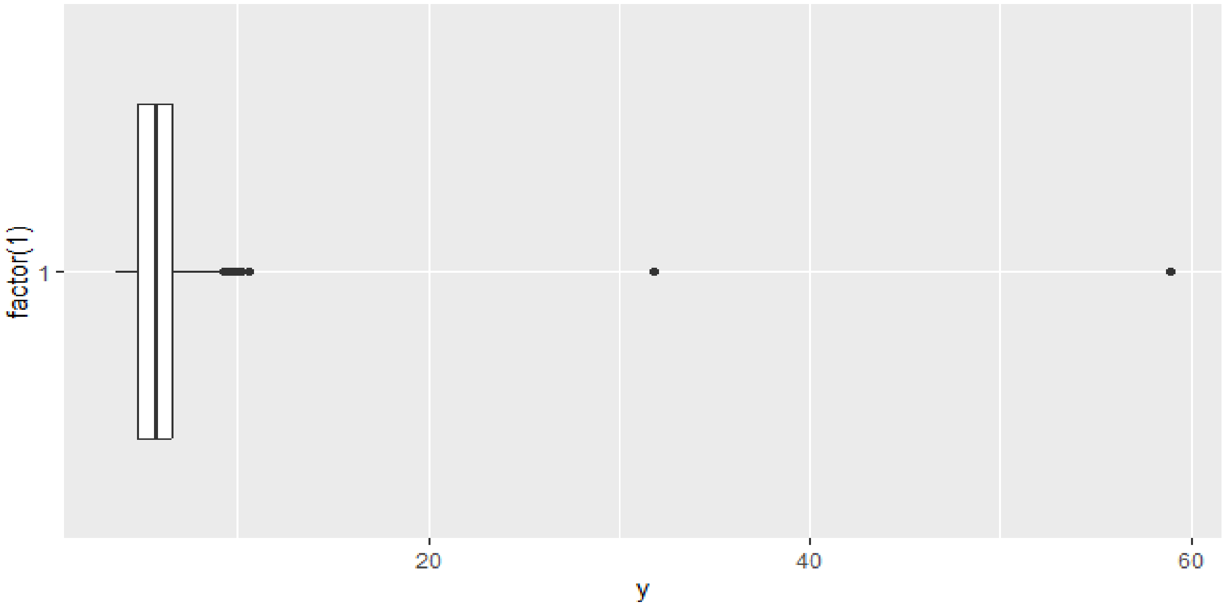
```
> print(nrow(diamonds))
[1] 53940
```

```
#storing dataset into variable df

df <-diamonds
```

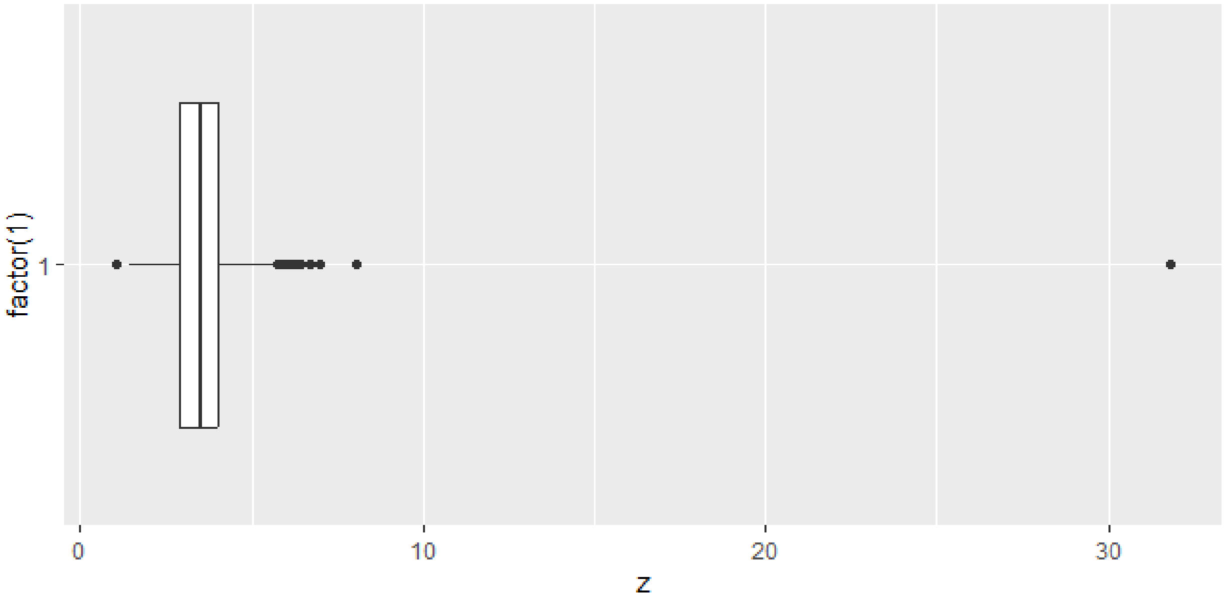
```
#checking for outliers

df %>%

  ggplot(aes(x, factor(1))) +

  geom_boxplot()
```



```
df %>%
  ggplot(aes(y, factor(1))) +
  geom_boxplot()
```



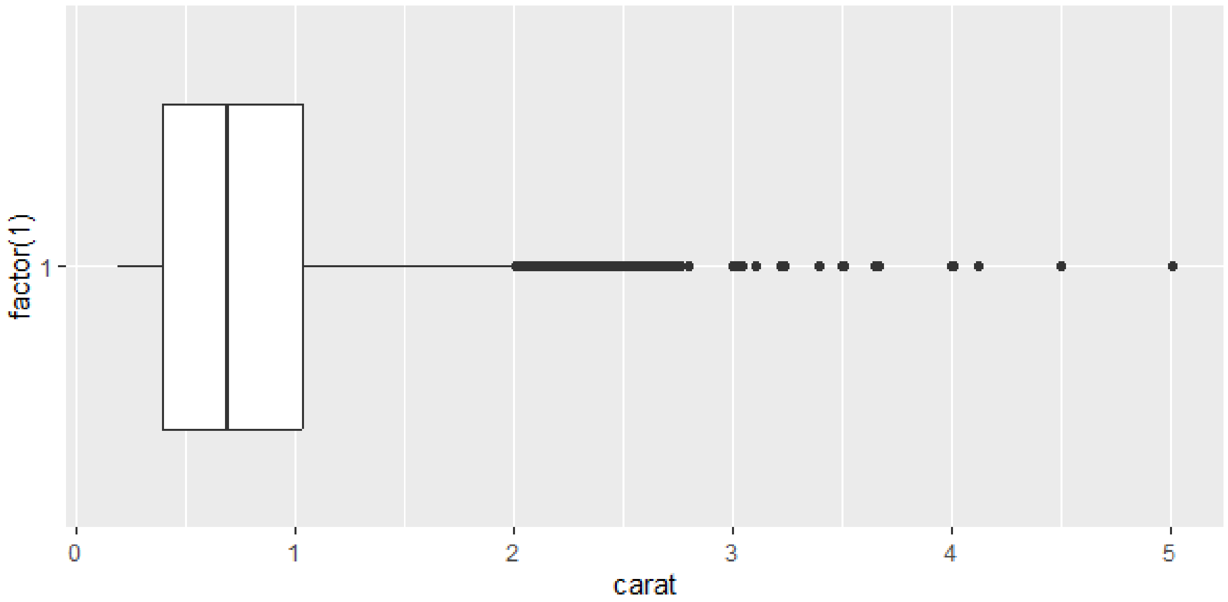
```
df %>%
  ggplot(aes(z, factor(1))) +
  geom_boxplot()
```



df %>%

  ggplot(aes(carat, factor(1))) +

  geom_boxplot()

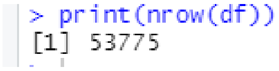


#removing outliers

df <- df %>%

  filter(x<10, y < 20, z < 10, carat < 2.5)

#checking the no. of rows in the dataset

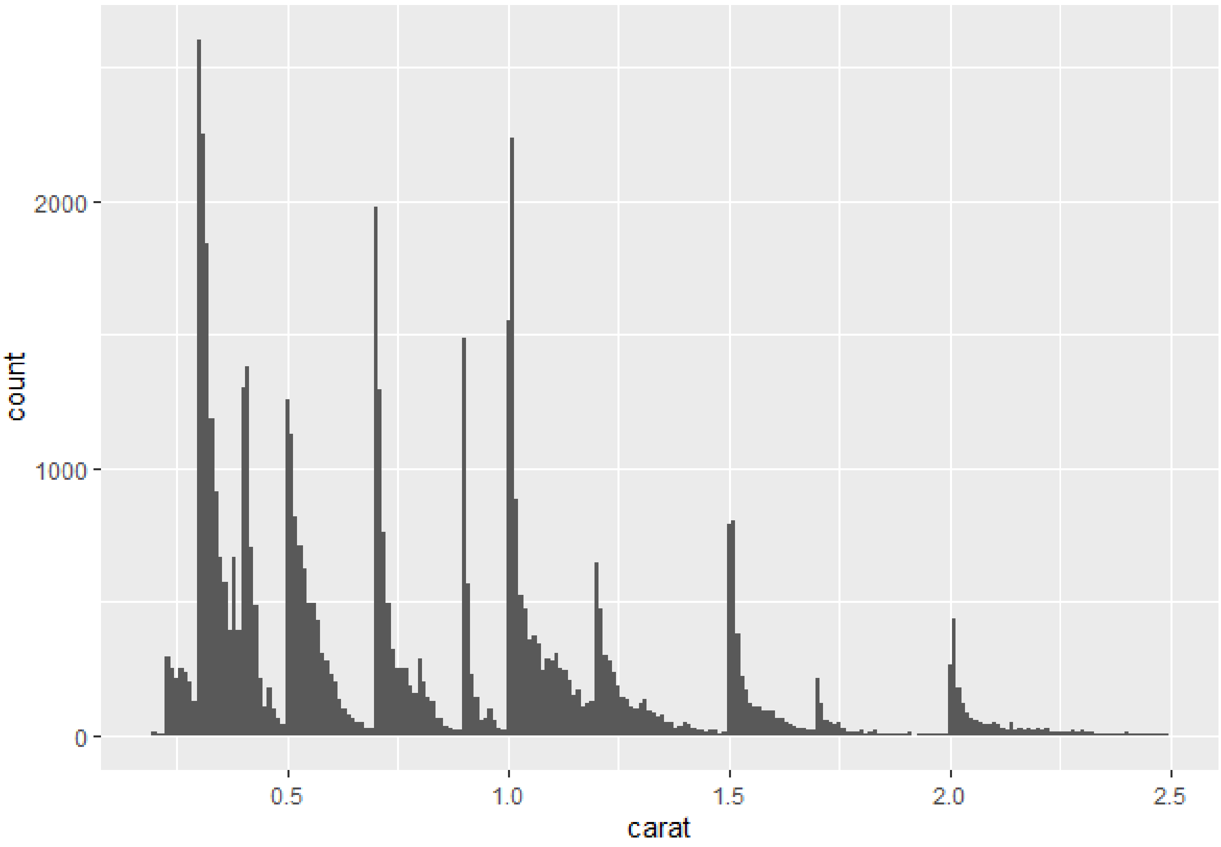


```
> print(nrow(df))
[1] 53775
```

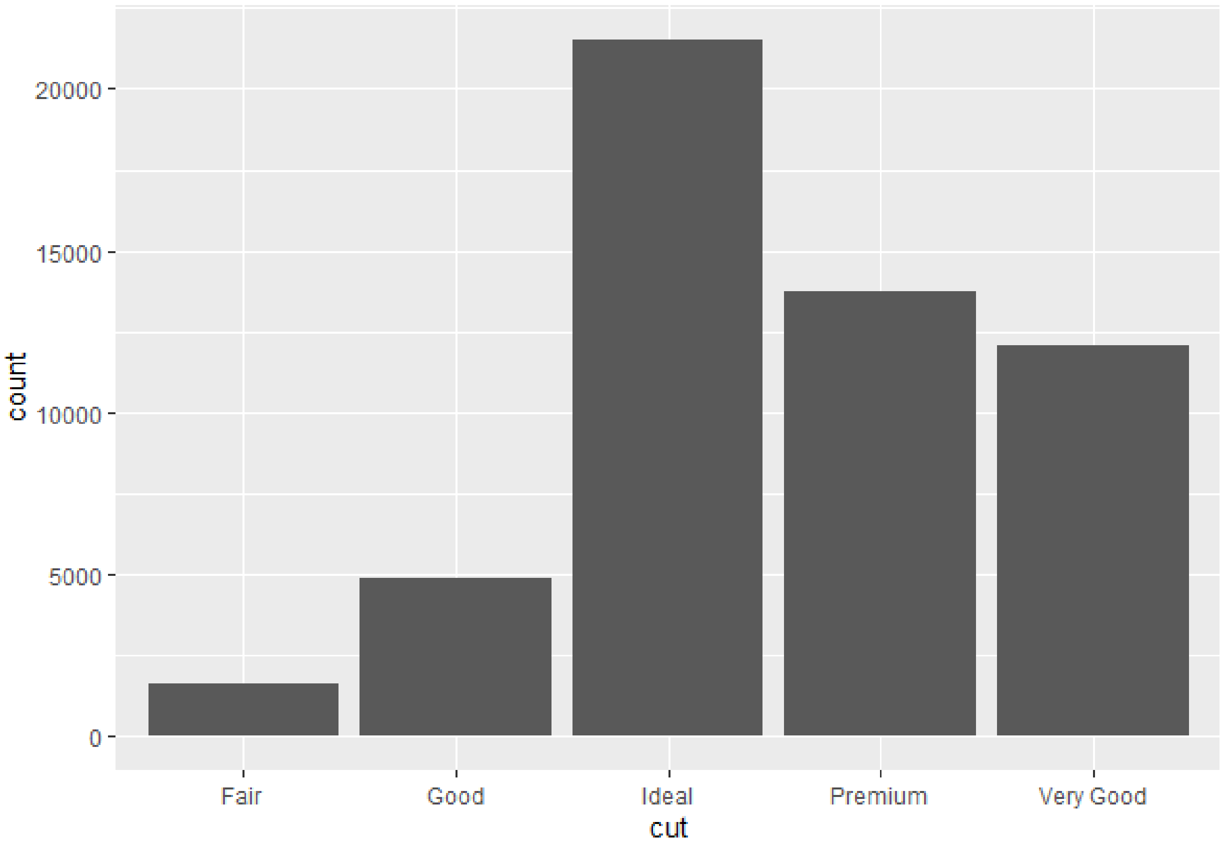
#so 53940-53775=165 rows were removed

Visualizations :

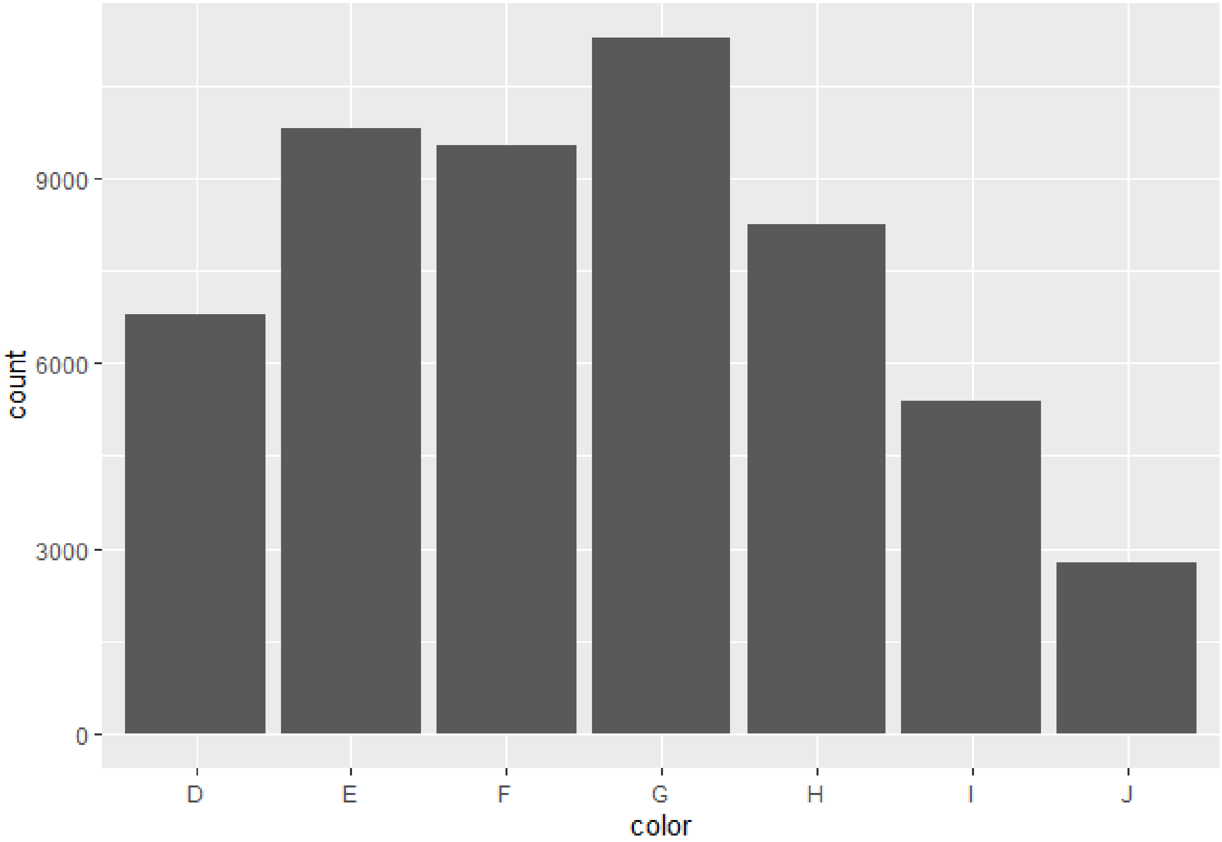
```
df %>%
  ggplot(aes(carat)) +
  geom_histogram(binwidth = 0.01)
```



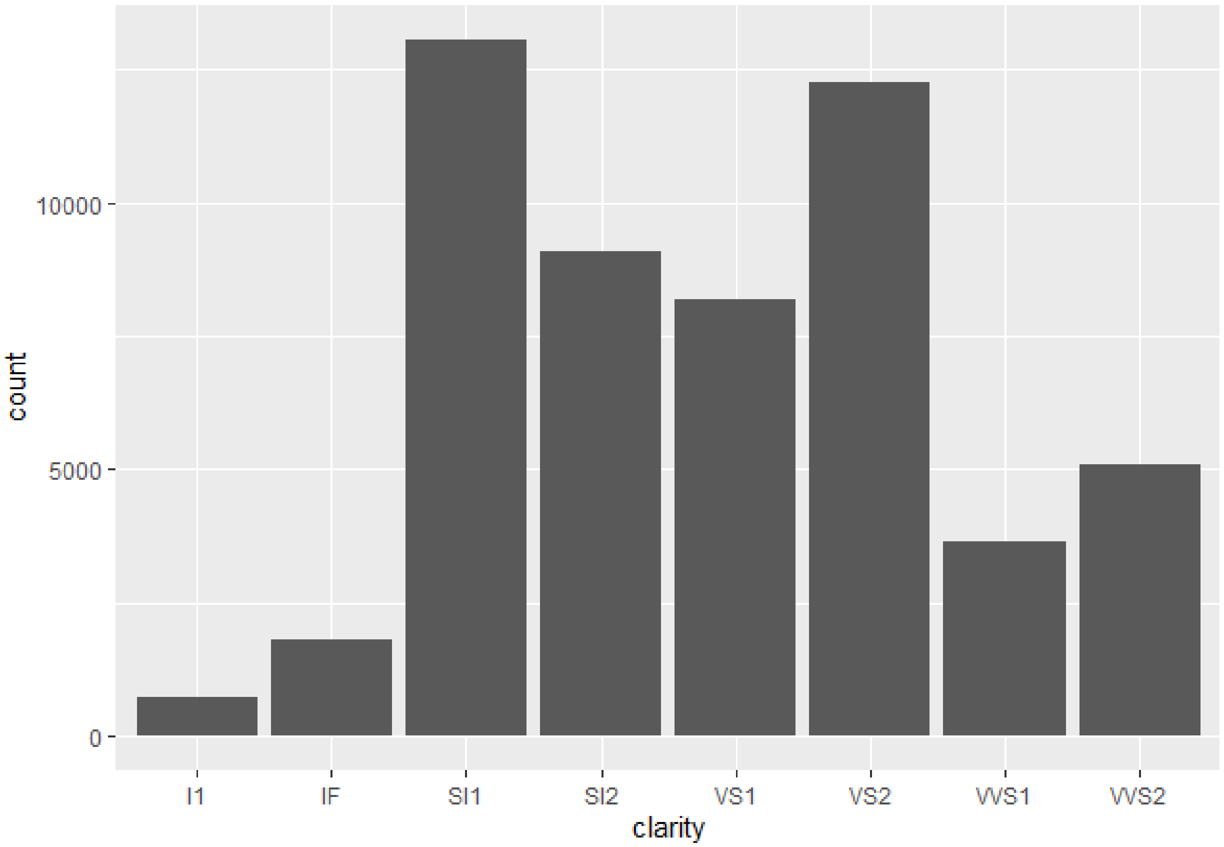
```
df %>%
  ggplot(aes(cut)) +
  geom_bar()
```



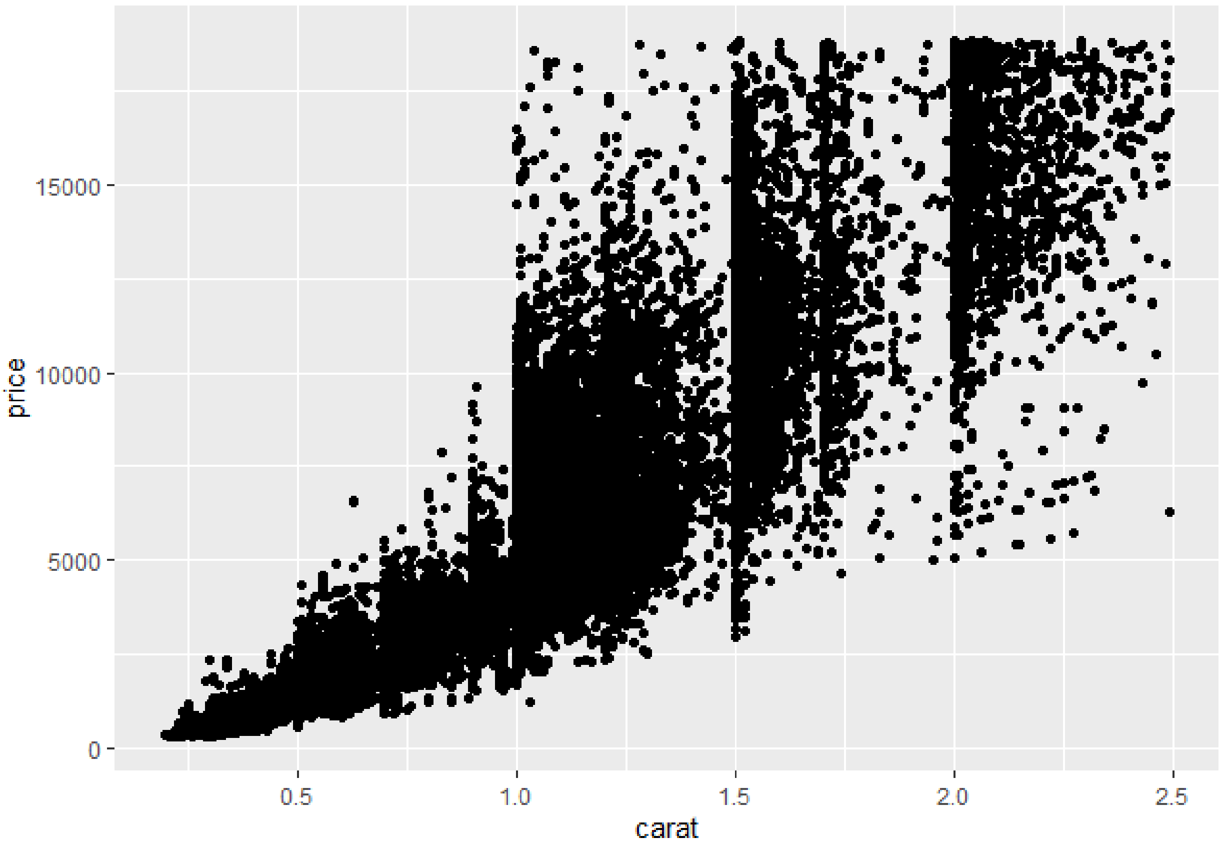
```
df %>%
  ggplot(aes(color)) +
  geom_bar()
```



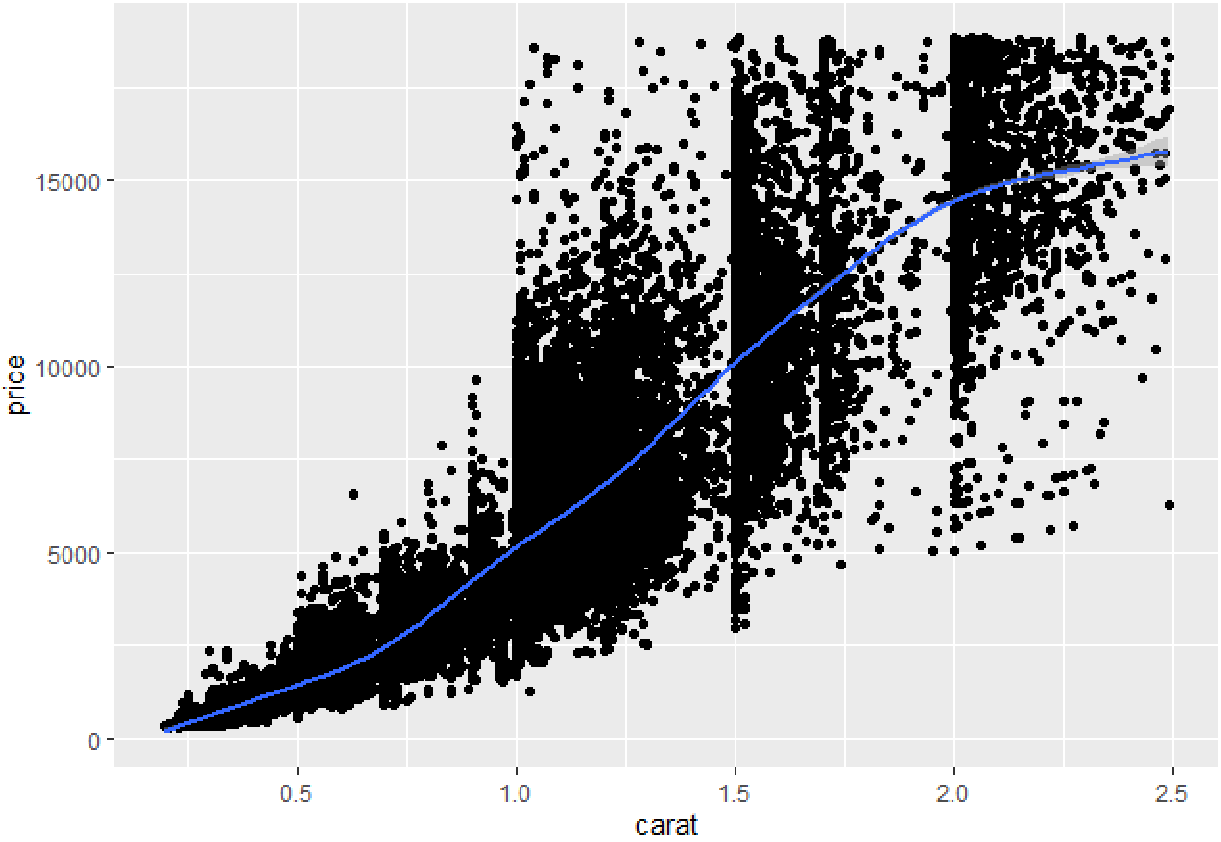
```
df %>%
  ggplot(aes(clarity)) +
  geom_bar()
```



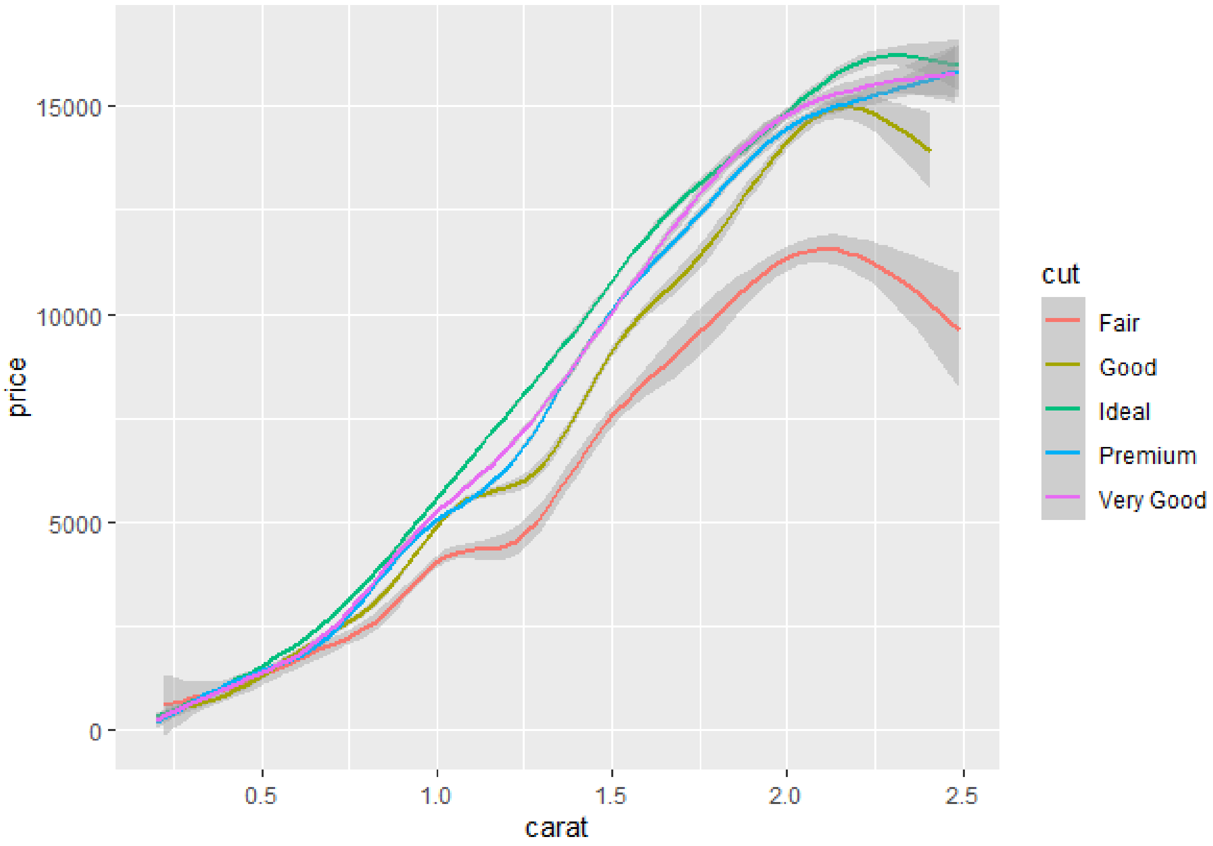
```
ggplot(data = df, mapping = aes(x = carat, y =price)) +
  geom_point()
```



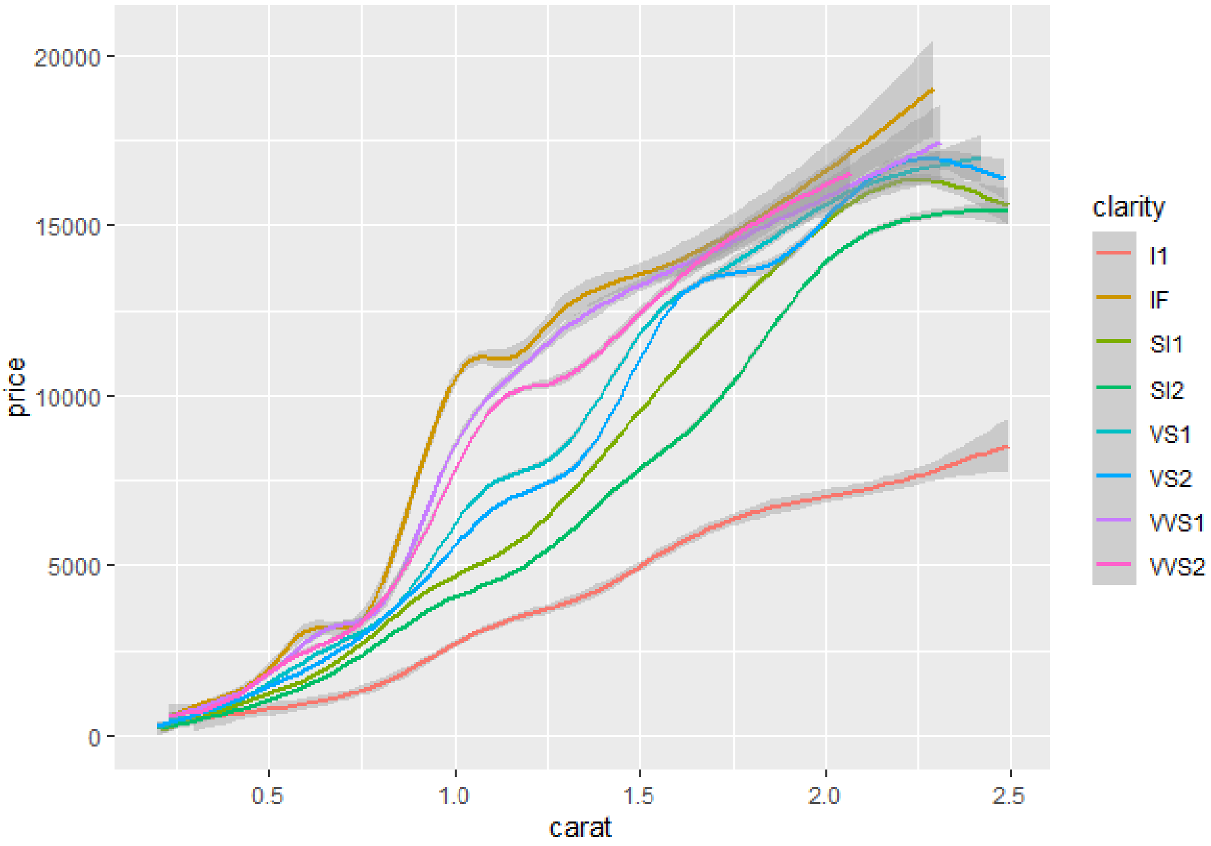
```
ggplot(data = df, mapping = aes(x = carat, y =price)) +
  geom_point() + geom_smooth()
```



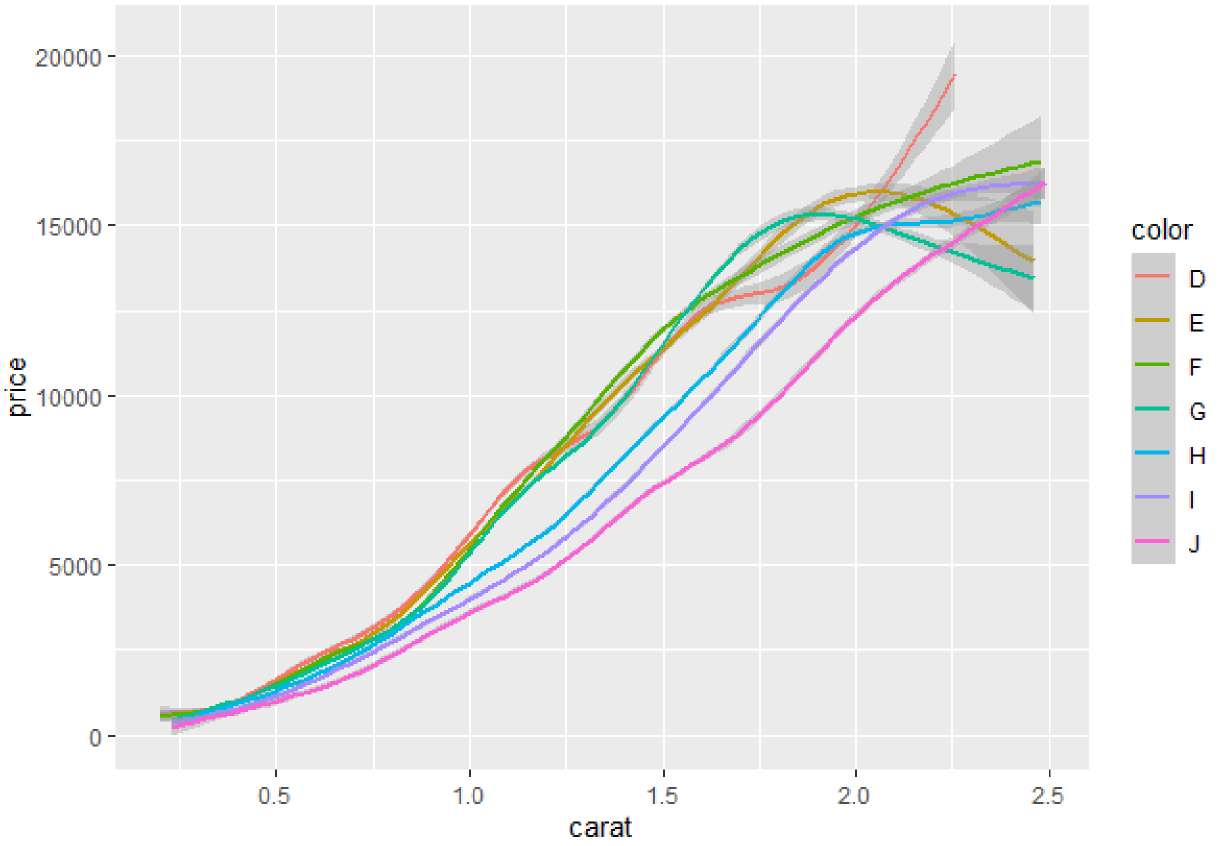
```
df %>%
  ggplot(aes(carat, price, color = cut)) +
  geom_smooth()
```



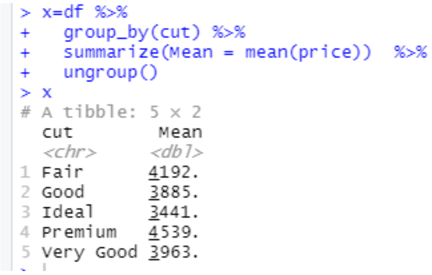
```
df %>%
  ggplot(aes(carat, price, color = clarity)) +
  geom_smooth()
```



```
df %>%
  ggplot(aes(carat, price, color = color)) +
  geom_smooth()
```

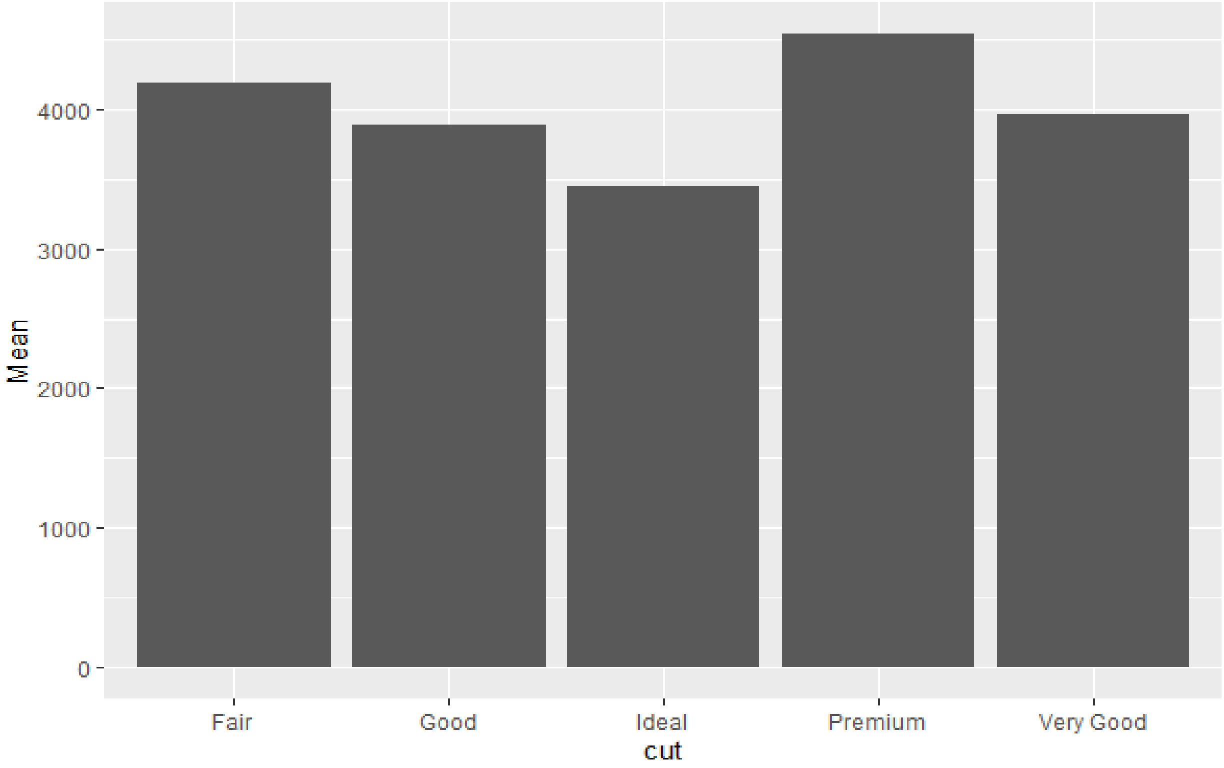


```
#making a sub table to analyze further

x=df %>%

  group_by(cut) %>%

  summarize(Mean = mean(price))  %>%

  ungroup()

x
```

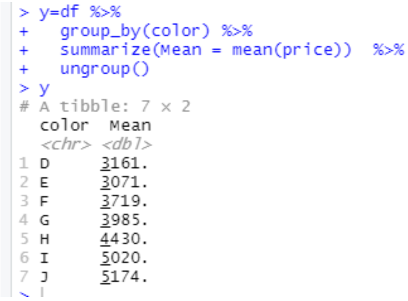


```
> x=df %>%
+   group_by(cut) %>%
+   summarize(Mean = mean(price))  %>%
+   ungroup()
> x
# A tibble: 5 x 2
  cut        Mean
  <chr>     <dbl>
1 Fair      4192.
2 Good      3885.
3 Ideal     3441.
4 Premium   4539.
5 Very Good 3963.
```

```
a<-ggplot(data=x, aes(x=cut, y=Mean)) +

  geom_bar(stat="identity")

a
```

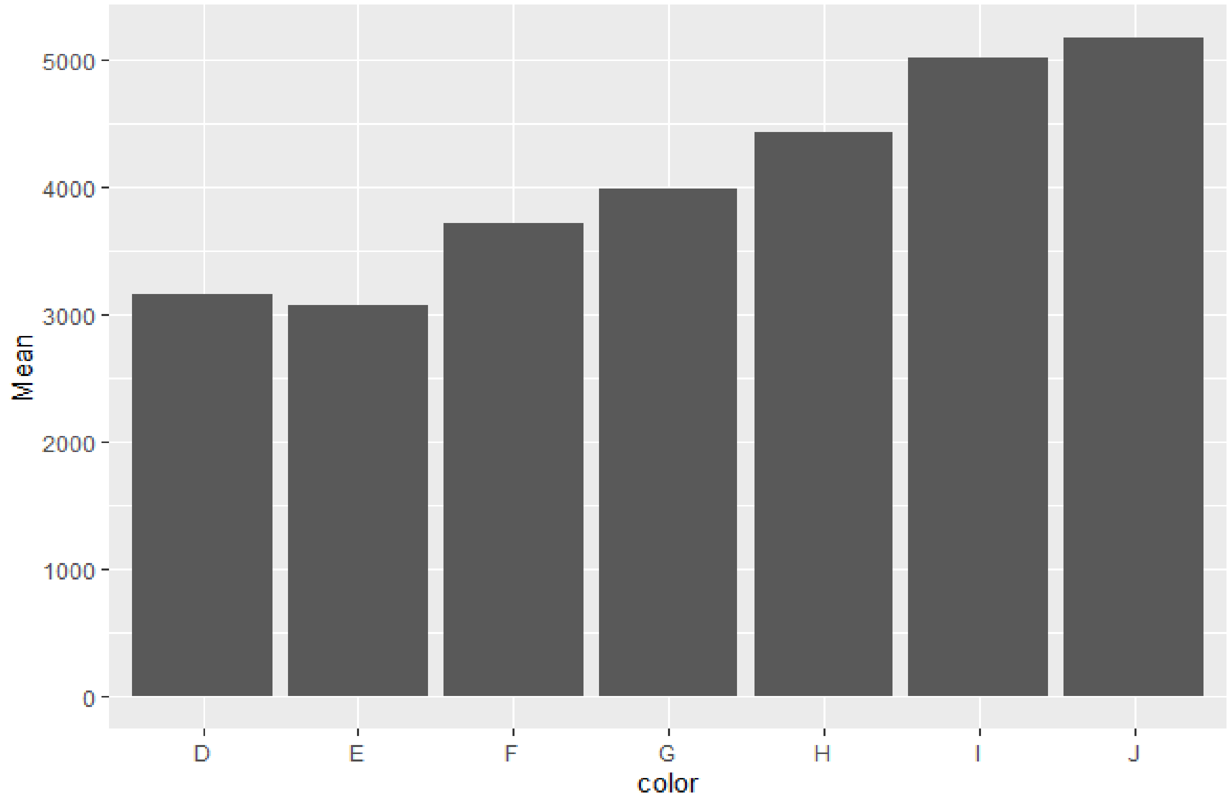


```
y=df %>%

  group_by(color) %>%

  summarize(Mean = mean(price))  %>%

  ungroup()

y
```

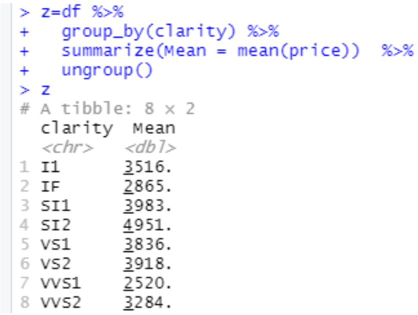


```
> y=df %>%
+   group_by(color) %>%
+   summarize(Mean = mean(price))  %>%
+   ungroup()
> y
# A tibble: 7 x 2
  color  Mean
  <chr> <dbl>
1 D     3161.
2 E     3071.
3 F     3719.
4 G     3985.
5 H     4430.
6 I     5020.
7 J     5174.
>
```

```
b<-ggplot(data=y, aes(x=color, y=Mean)) +

  geom_bar(stat="identity")

b
```

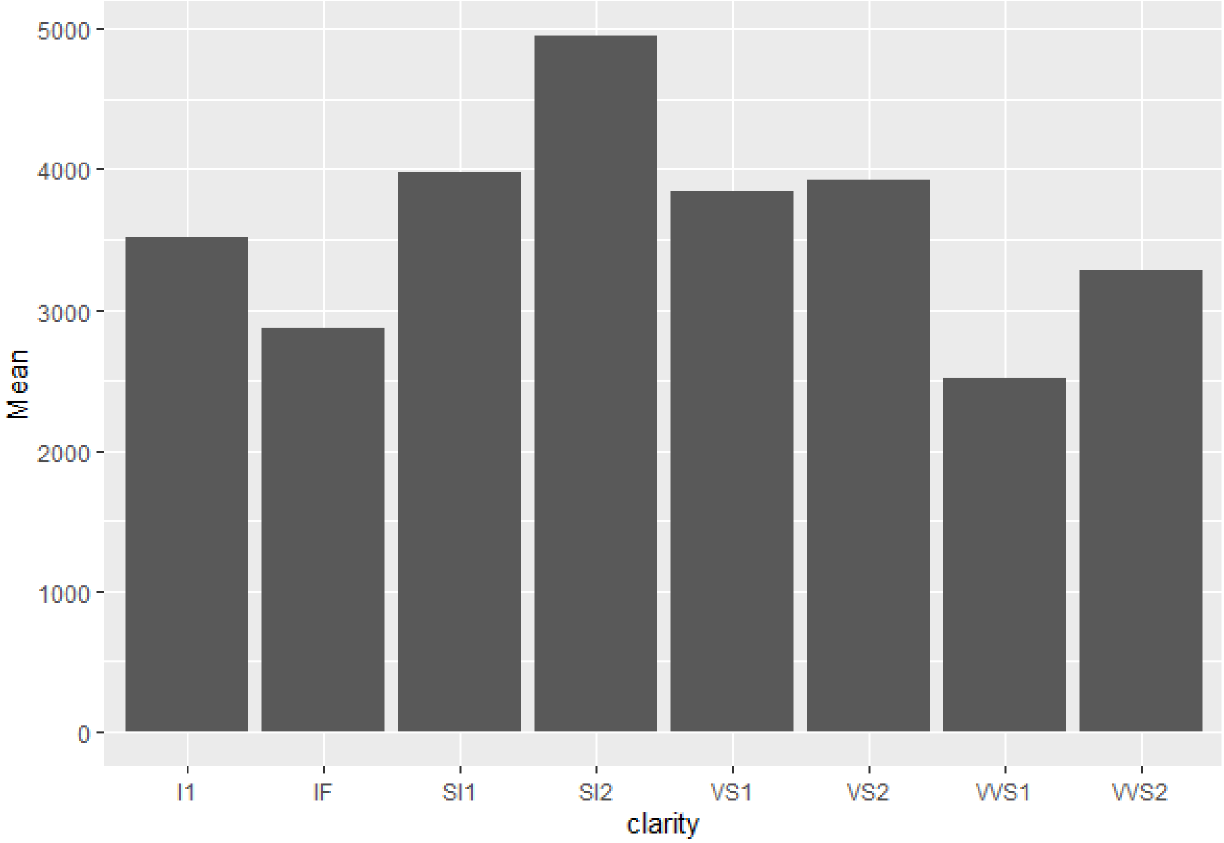


```
z=df %>%
  group_by(clarity) %>%
  summarize(Mean = mean(price))  %>%
  ungroup()
z
```



```
> z=df %>%
+   group_by(clarity) %>%
+   summarize(Mean = mean(price))  %>%
+   ungroup()
> z
# A tibble: 8 x 2
  clarity  Mean
  <chr>    <dbl>
1 I1       3516.
2 IF       2865.
3 SI1      3983.
4 SI2      4951.
5 VS1      3836.
6 VS2      3918.
7 VVS1     2520.
8 VVS2     3284.
```

```
c<-ggplot(data=z, aes(x=clarity, y=Mean)) +
  geom_bar(stat="identity")
c
```



Conclusions :

1. The highest number of diamonds in the dataset is of 0.3 carat followed by 1 carat diamonds.
2. The highest number of diamonds in the dataset is of ideal cut followed by premium cut diamonds.
3. The highest number of diamonds in the dataset is of color G followed by E color diamonds.
4. The highest number of diamonds in the dataset is of clarity SI1 followed by VS2 color diamonds.
5. The price of the diamond is proportional to the carat rating of the diamond.
6. The price of the Ideal cut diamond is more whereas the price of the fair cut diamond is lesser than other cuts for the same carat rating.
7. The price of the IF clarity diamond is more whereas the price of the l1 clarity diamond is lesser than other cuts for the same carat rating.
8. The price of the J color diamond is lesser than other cuts for the carat rating below 2.5.
9. The mean price of premium cut diamonds is highest in the dataset.
10. The mean price of J color diamonds is highest in the dataset.
11. The mean price of SI2 clarity diamonds is highest in the dataset.