Data Visualization Bootcamp Homework

Nirucha P

2023-12-23

Assignment

1. Use the diamonds dataset to plot and explain the charts.
2. Use the ggplot2 (or tidyverse) library in R Language.
3. Export this document to pdf with R Markdown.

Install packaged

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Load clean data

```
data("diamonds")
str(diamonds)
```

```
## tibble [53,940 x 10] (S3: tbl_df/tbl/data.frame)
##  $ carat  : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
##  $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
##  $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
##  $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
##  $ depth  : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
##  $ table  : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
##  $ price  : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
##  $ x      : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
##  $ y      : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
##  $ z      : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

Prepare data

Check some missing values in the diamonds dataset.

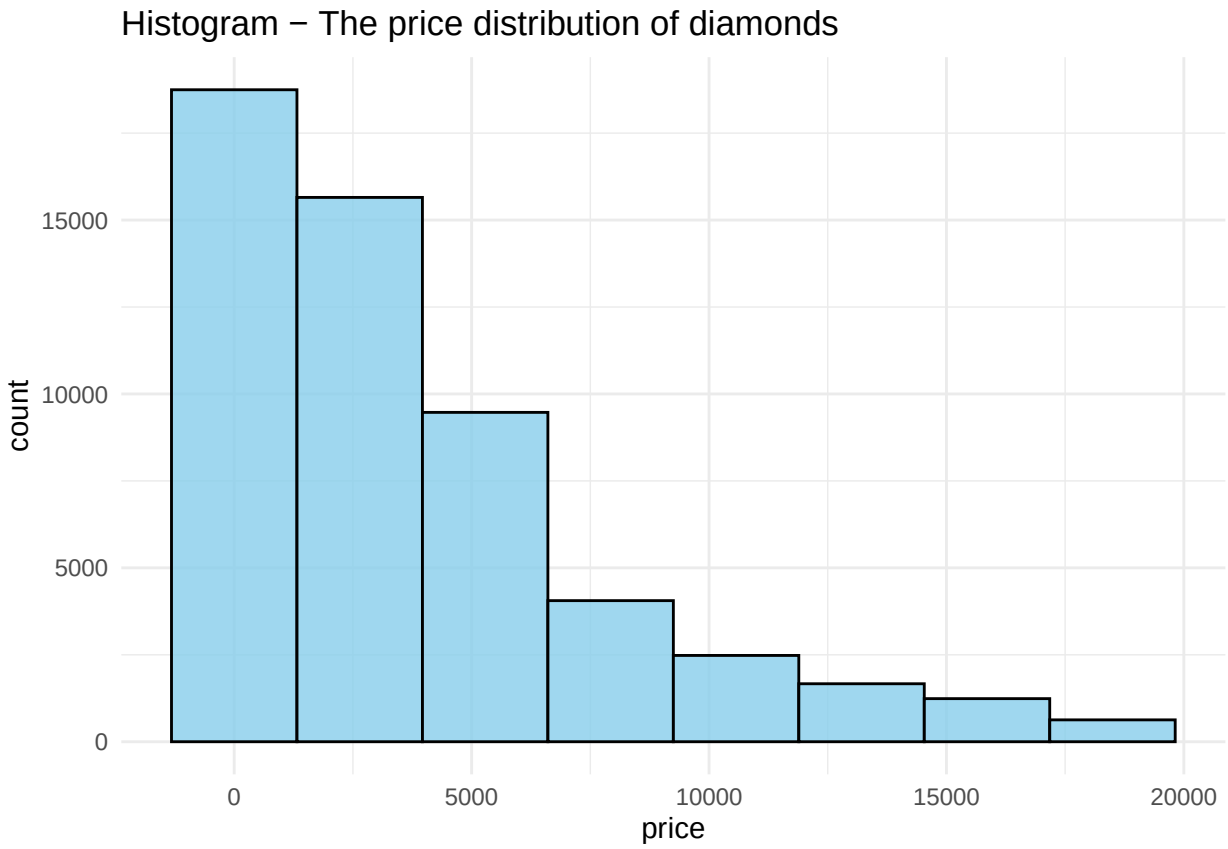
```
if(sum(is.na(diamonds)) > 0){
  print("This dataset has some missing values.")

} else{
  print("This dataset doesn't have any missing values.")
}
```

```
## [1] "This dataset doesn't have any missing values."
```

Plot and explain the charts

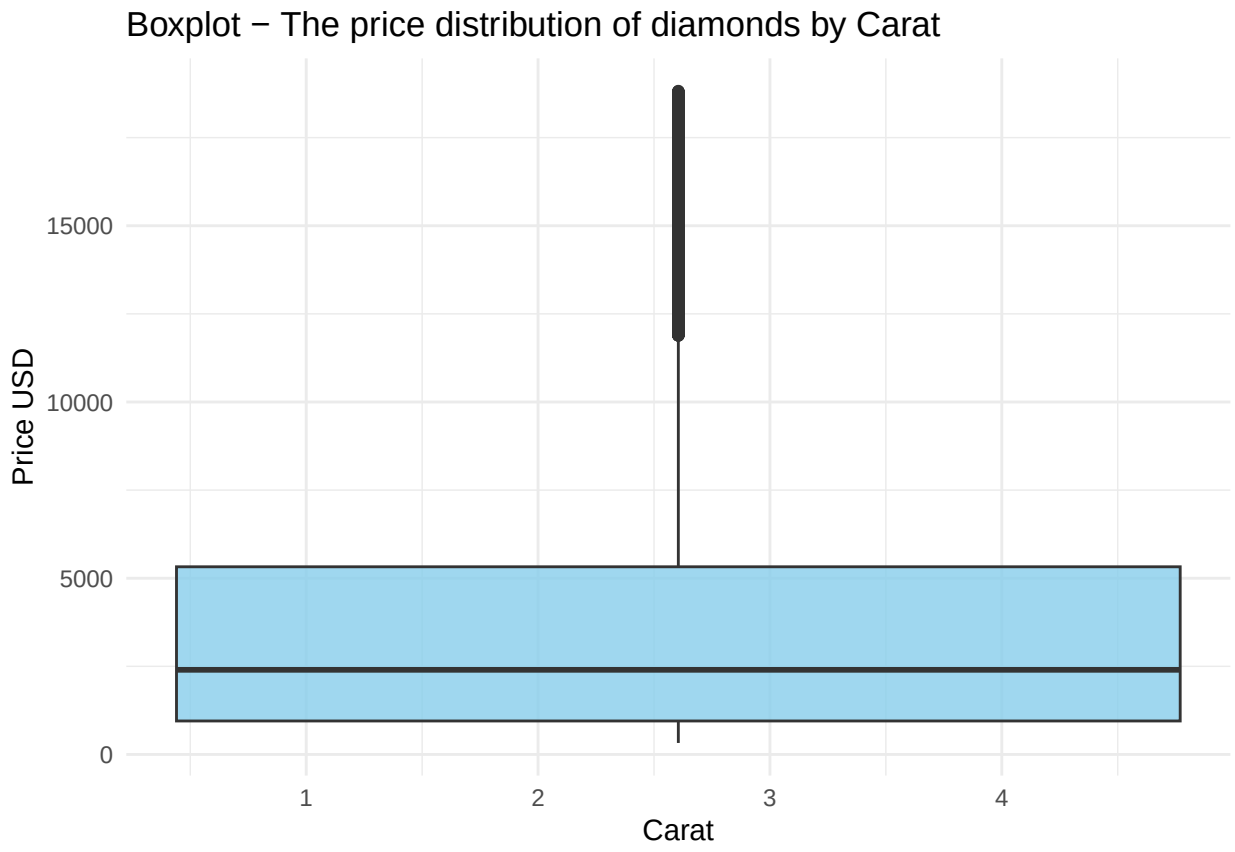
[1] What is the price distribution of diamonds?

```
ggplot(data = diamonds,
       mapping = aes(x = price)) +
  geom_histogram(bins = 8, fill = "skyblue", col = "black", alpha = 0.8) +
  theme_minimal() +
  labs(title = "Histogram - The price distribution of diamonds")
```



```r
ggplot(diamonds, aes(x = carat, y = price)) +
  geom_boxplot(fill = "skyblue", alpha = 0.8) +
  theme_minimal() +
  labs(title = "Boxplot - The price distribution of diamonds by Carat",
       x = "Carat",
       y = "Price USD")
```

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




```r
summary(diamonds$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     326     950    2401    3933    5324   18823
```

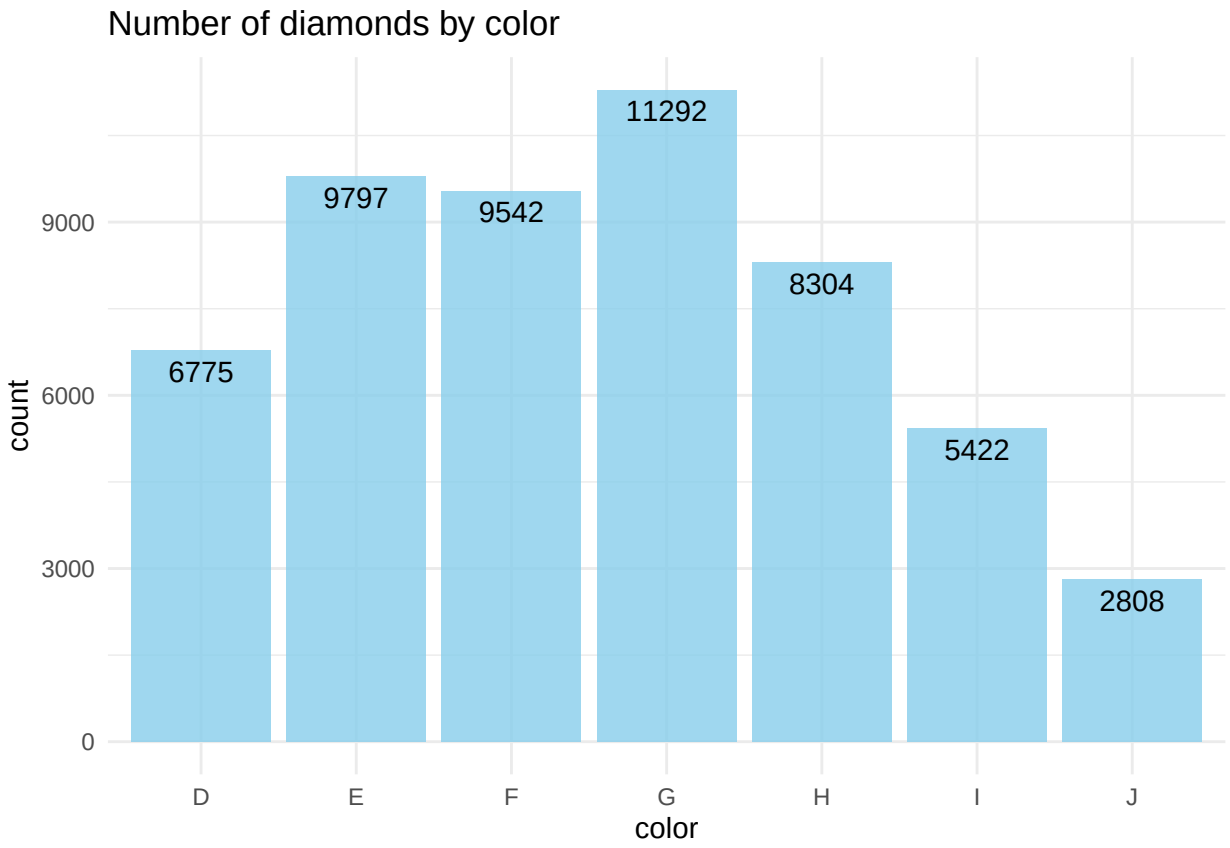
The price distribution of diamonds is right-skewed, with a median price of around 2,400 USD. Boxplot shows the value of the outlier at the top.

It shows that most diamond prices are below 5,000 USD.

[2] What is the most color of diamonds

```
ggplot(data = diamonds,
       mapping = aes(x = color)) +
  geom_bar(fill = "skyblue", alpha = 0.8) +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "black") +
  theme_minimal() +
  labs(title = "Number of diamonds by color")
```

```
## Warning: The dot-dot notation (‘..count..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(count)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



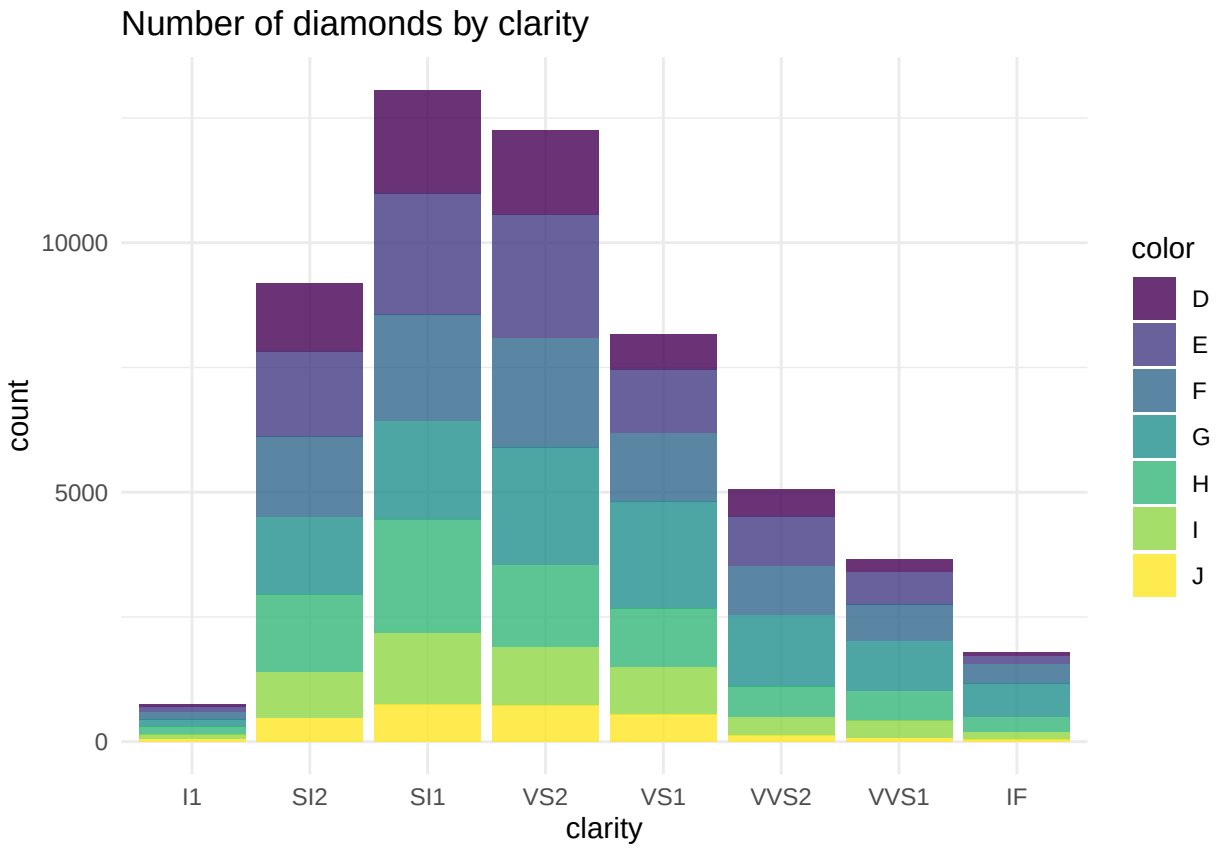
The color of diamonds is separated into three groups (best to worst), as follows:

1. Colorless : D, E, F
2. Near Colorless : G, H (Typically look colorless to the naked eye)
3. Near Colorless Slightly Tinted : I, J

From the chart above, the number of diamonds with a color grade of G is the highest, and J is the lowest.

[3] What is the most clarity of diamonds?

```
ggplot(data = diamonds,
       mapping = aes(x = clarity, fill = color)) +
  geom_bar(alpha = 0.8) +
  theme_minimal() +
  labs(title = "Number of diamonds by clarity")
```



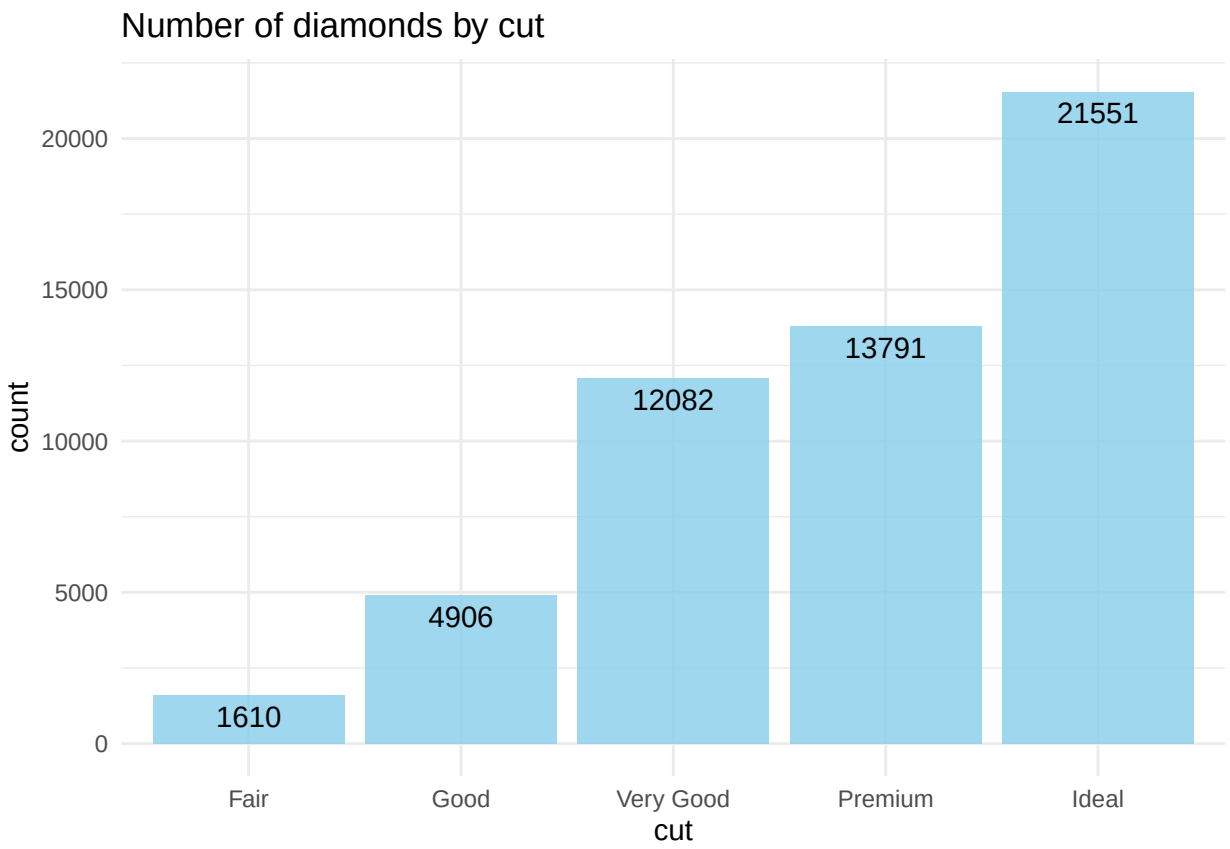
Since diamonds are natural gemstones, it is simple to find impurities in diamonds. However, the clarity of a diamond affects its sparkle.

The order of clarity of diamonds is as follows (worst to best): I1 < SI2 < SI1 < VS2 < VS1 < VVS2 < VVS1 < IF

From the chart above, most diamonds found may be in the SI1 category (Slightly Included).

[4] What is the most cutting of diamonds?

```
ggplot(data = diamonds,
       mapping = aes(x = cut)) +
  geom_bar(fill= "skyblue", alpha = 0.8) +
  geom_text(aes(label = ..count..), stat = "count", vjust = 1.5, colour = "black") +
  theme_minimal()  +
  labs(title = "Number of diamonds by cut")
```

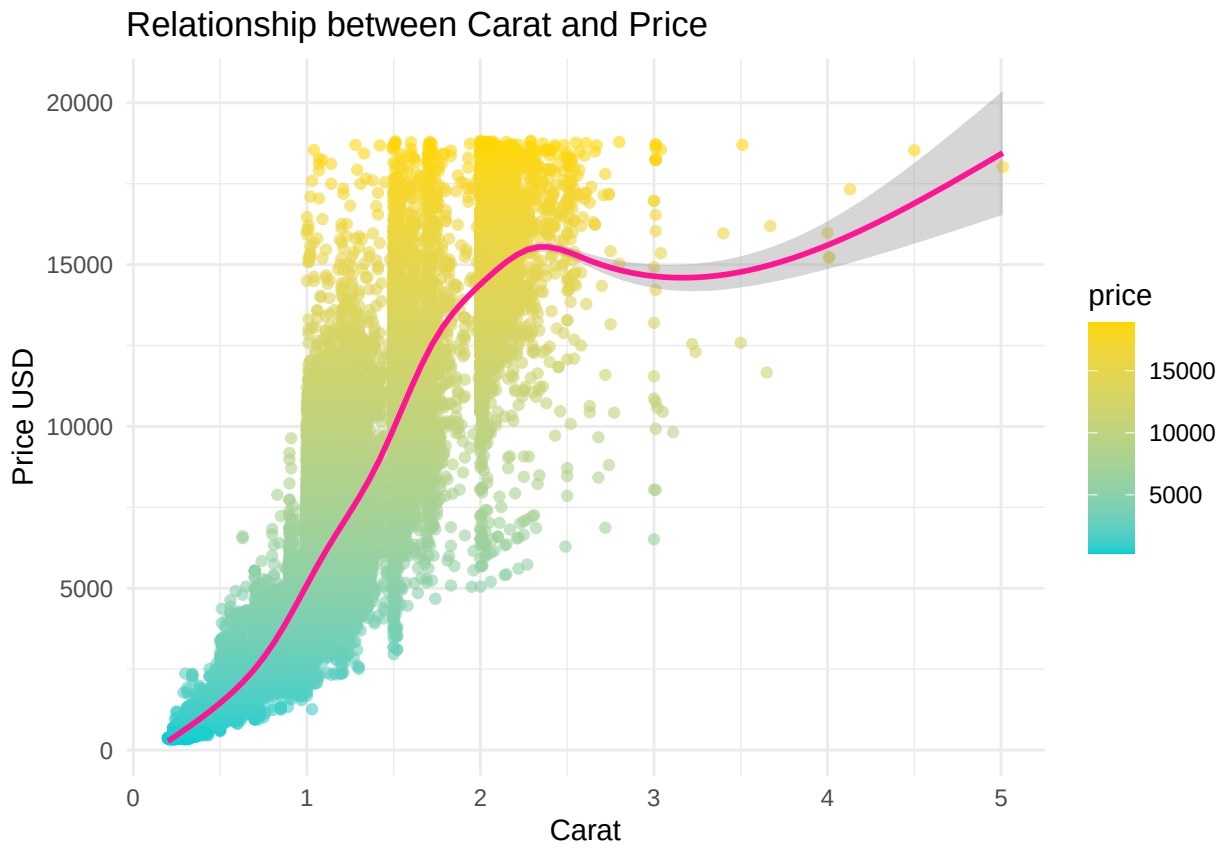


The process of diamond cutting is a factor that jewelers can control themselves because both the color and clarity of the diamonds occur naturally. From the chart above, the ideal level (the best of cutting) has the highest number of diamonds.

[5] What is the relationship between carat and price?

```
ggplot(diamonds, aes(x = carat, y = price, col = price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(col = "deeppink") +
  scale_color_gradient(low = "darkturquoise", high = "gold") +
  theme_minimal() +
  labs(title = "Relationship between Carat and Price",
       x = "Carat",
       y = "Price USD")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
cor(diamonds$carat, diamonds$price)
```

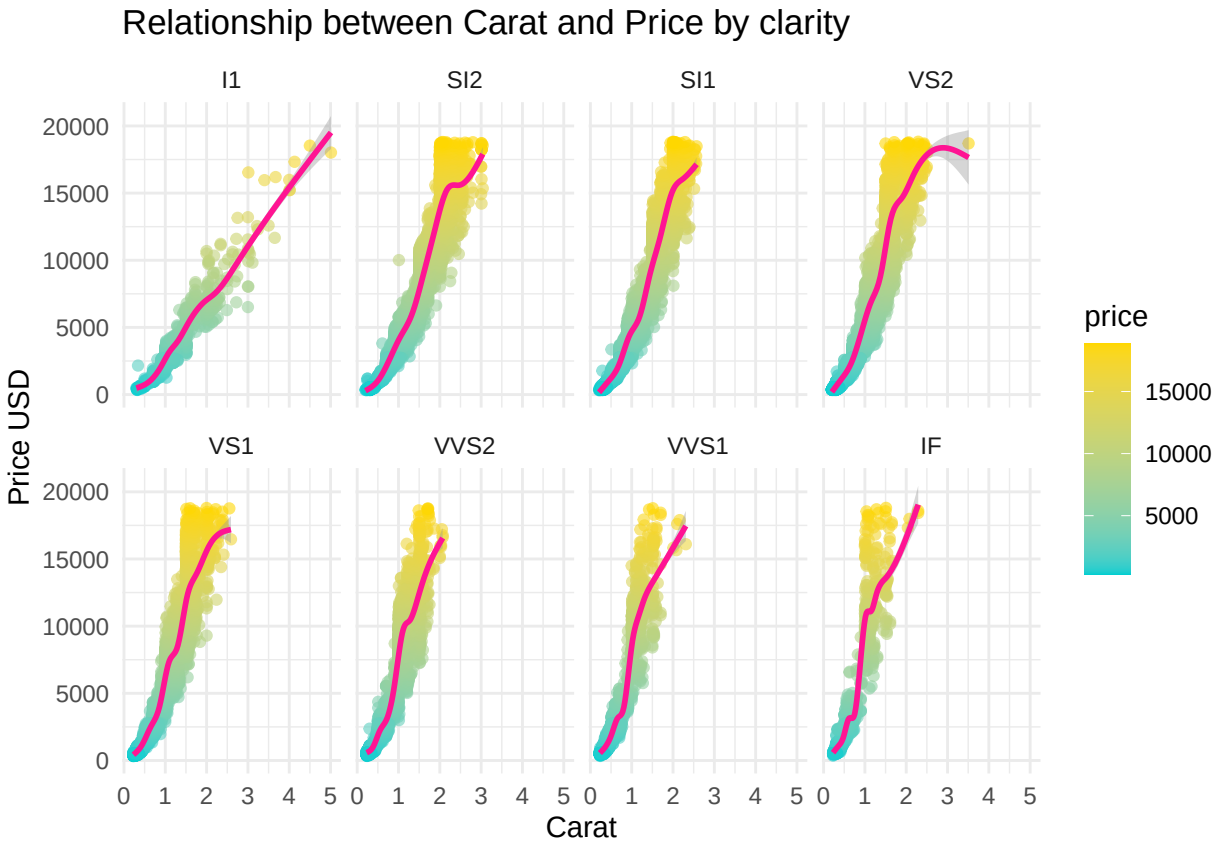
```
## [1] 0.9215913
```

From the scatter plot and a correlation coefficient value of around 0.92, there is a correlation between carat and prices trending in the same direction and positively correlated. Therefore, it is possible that as the carat weight increases, there is a tendency for the price to increase as well.

[6] What is the relationship between carat and price, categorized by the clarity?

```
ggplot(diamonds, aes(x = carat, y = price, col=price)) +
  geom_point(alpha = 0.6) +
  geom_smooth(col = "deeppink") +
  facet_wrap(~ clarity, ncol =4 ) +
  scale_color_gradient(low = "darkturquoise", high = "gold") +
  theme_minimal() +
  labs(title = "Relationship between Carat and Price by clarity",
       x = "Carat",
       y = "Price USD")
```

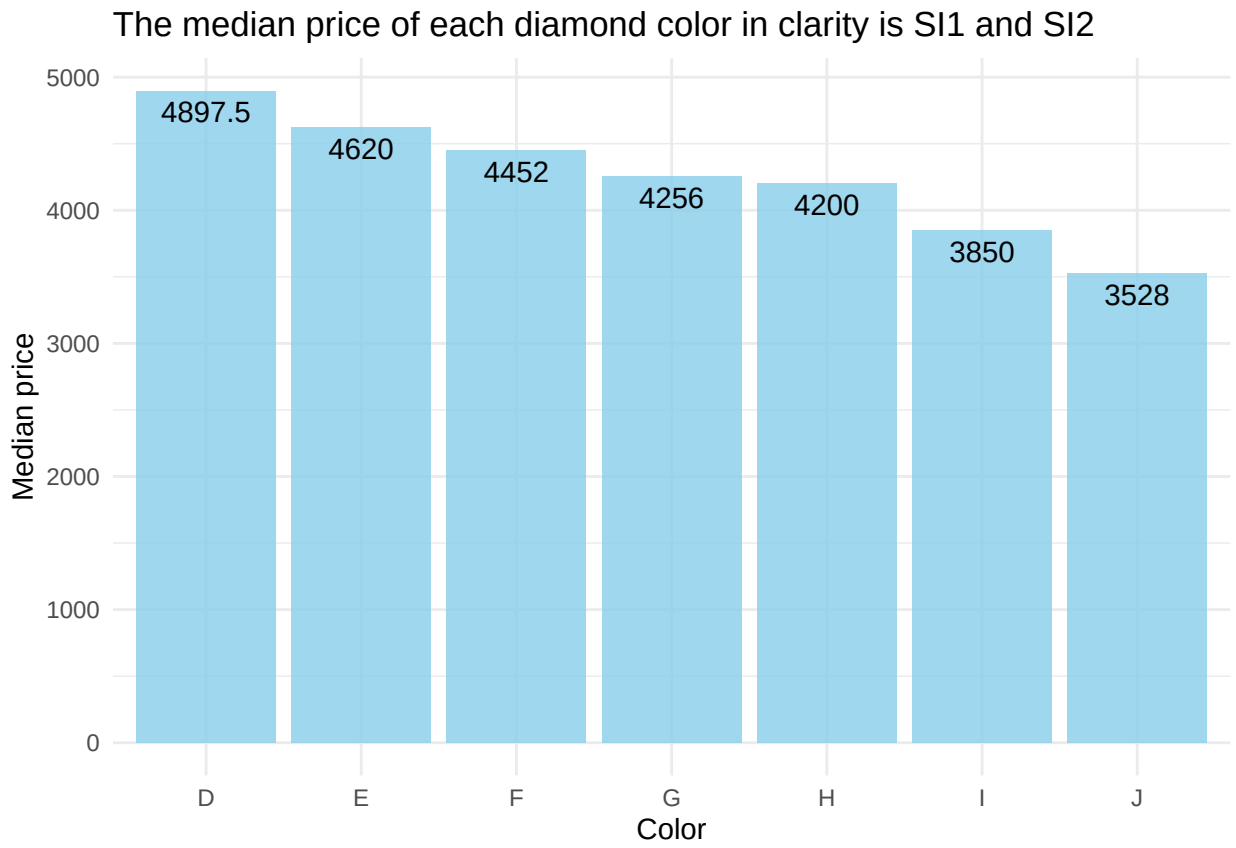
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



From the above graph, in the relationship between carat and price, categorized by the clarity of the diamond, we can observe that for diamonds with IF clarity level, weighing 1 carat, the price is more than 15,000 USD. As the clarity level decreases, diamonds must weigh more than 1 carat to have a price exceeding 15,000 USD. Therefore, the clarity level of the diamond tends to higher prices.

[7] What is the median price of each diamond color in clarity is SI1, SI2?

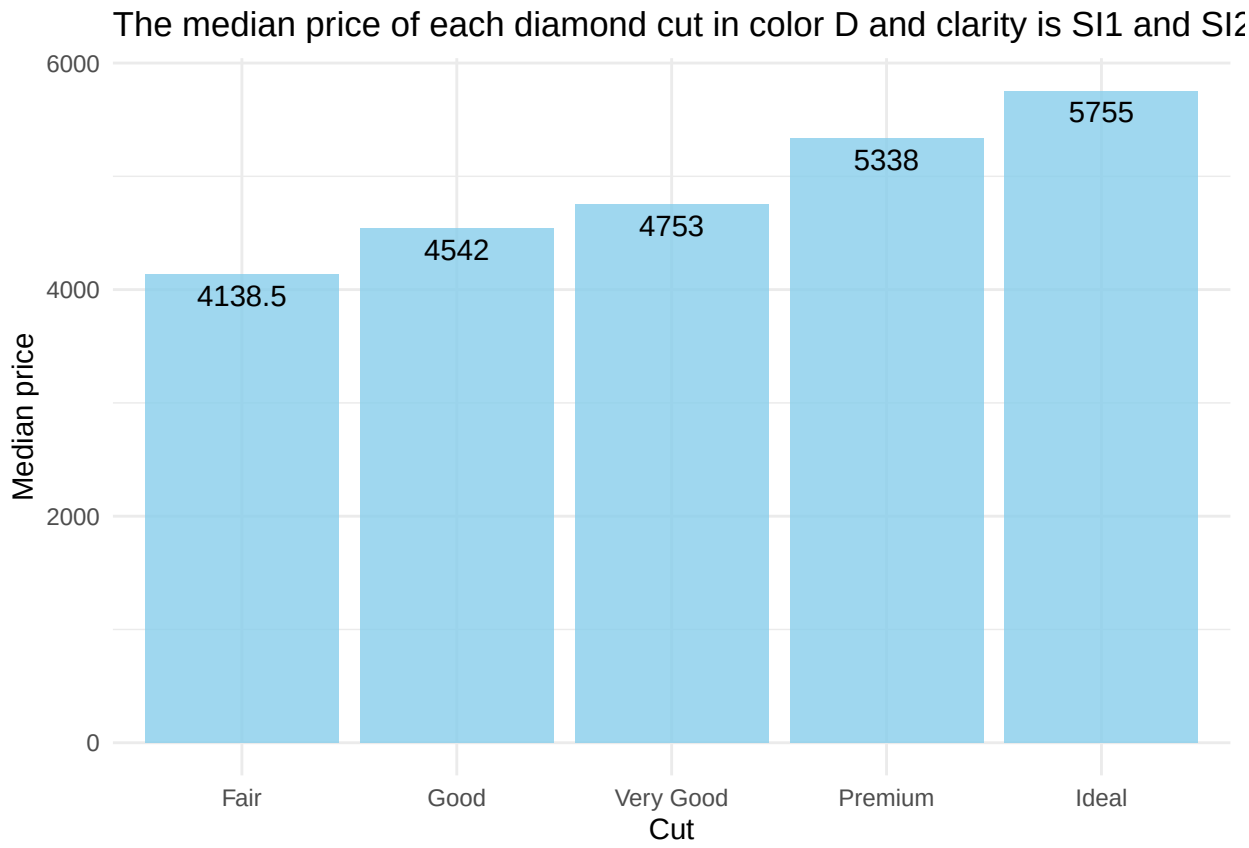
```
diamonds %>%
  filter(carat == 1, (clarity == "SI1" | clarity == "SI2")) %>%
  group_by(color) %>%
  summarise(med_price = median(price)) %>%
  ggplot(aes(color,med_price))+
  geom_col(fill = "skyblue", alpha = 0.8) +
  geom_text(aes(label = med_price), vjust = 1.5, colour = "black") +
  theme_minimal() +
  labs(title = "The median price of each diamond color in clarity is SI1 and SI2",
    x = "Color",
    y = "Median price")
```



From the above graph, in the case where diamonds weigh 1 carat and are in the clarity range of SI1 and SI2 (which is a relatively lower clarity level), it is found that the color of the diamonds affects the diamond's price.

[8] **What is the median price of each diamond cut in color D and clarity is SI1, SI2**

```
diamonds %>%
  filter(color == "D", carat == 1, (clarity == "SI1" | clarity == "SI2")) %>%
  group_by(cut) %>%
  summarise(med_price = median(price)) %>%
  ggplot(aes(cut,med_price))+
  geom_col(fill = "skyblue", alpha = 0.8) +
  geom_text(aes(label = med_price), vjust = 1.5, colour = "black") +
  theme_minimal() +
  labs(title = "The median price of each diamond cut in color D and clarity is SI1 and SI2",
    x = "Cut",
    y = "Median price")
```



From the above graph, in the case where a diamond weighs 1 carat, has a color grade of D, and has clarity in the range of SI1 and SI2. It found that the diamond cut affects the diamond's price.

Conclusion

From the diamonds dataset, we found that carat weight is a significant factor in the price of diamonds. In terms of color and clarity, that cannot be controlled, as diamonds are natural occurrences. However, in cases where diamonds have lower color grades or are in a clarity range with noticeable imperfections, diamond cutting is a factor that can enhance the price of the diamond, and this factor jewelers can control to manage the quality themselves.