Homework Data Viz

Monsicha Tame

2023-08-17

Hello World

This project is part of my homework for Data Rockie's Data Science Boot Camp 7 to explore the diamond data set by creating a visualization.

Load required package

Install any required packages.

```
library(tidyverse)
library(corrplot)
```

Explore data frame

Check the following data basics: column names, number of observations, data type, formatting, and missing values.

```
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.…
```

```
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>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

```
which(is.na(diamonds))
```

```
## integer(0)
```

Data Summary

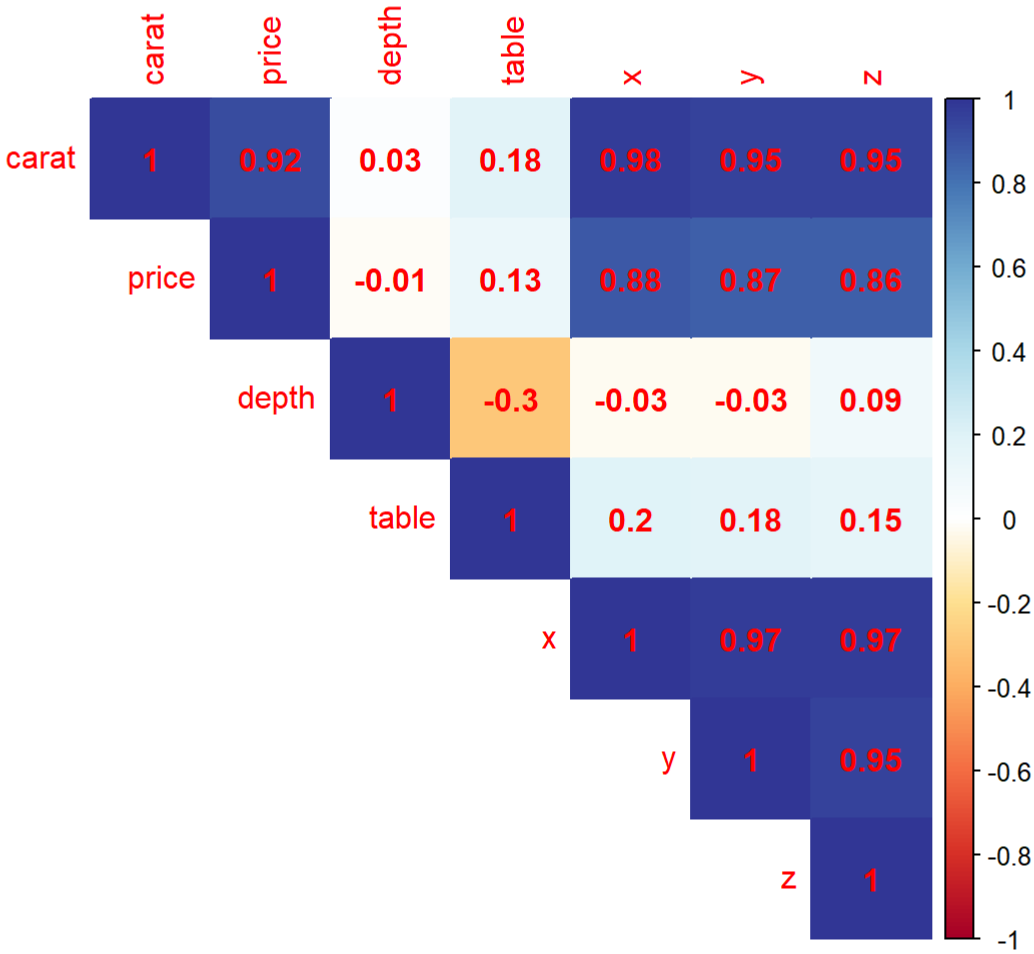
1. There are a total of 53,940 observations.
2. There are no missing value.
3. There are three categorical variables: cut, carat, and clarity.
4. There are seven numerical variables to consider: carat, depth, table, price, x, y, and z.

Explore visualization

Numerical vs Numerical Variables

Check correlation of the numerical variables

```
num <- diamonds%>%
  select(carat, price, depth, table, x, y, z)

corrplot(cor(num), method = "color", type = "upper", addCoef.col =  "red", col = COL2("RdYlBu"))
```



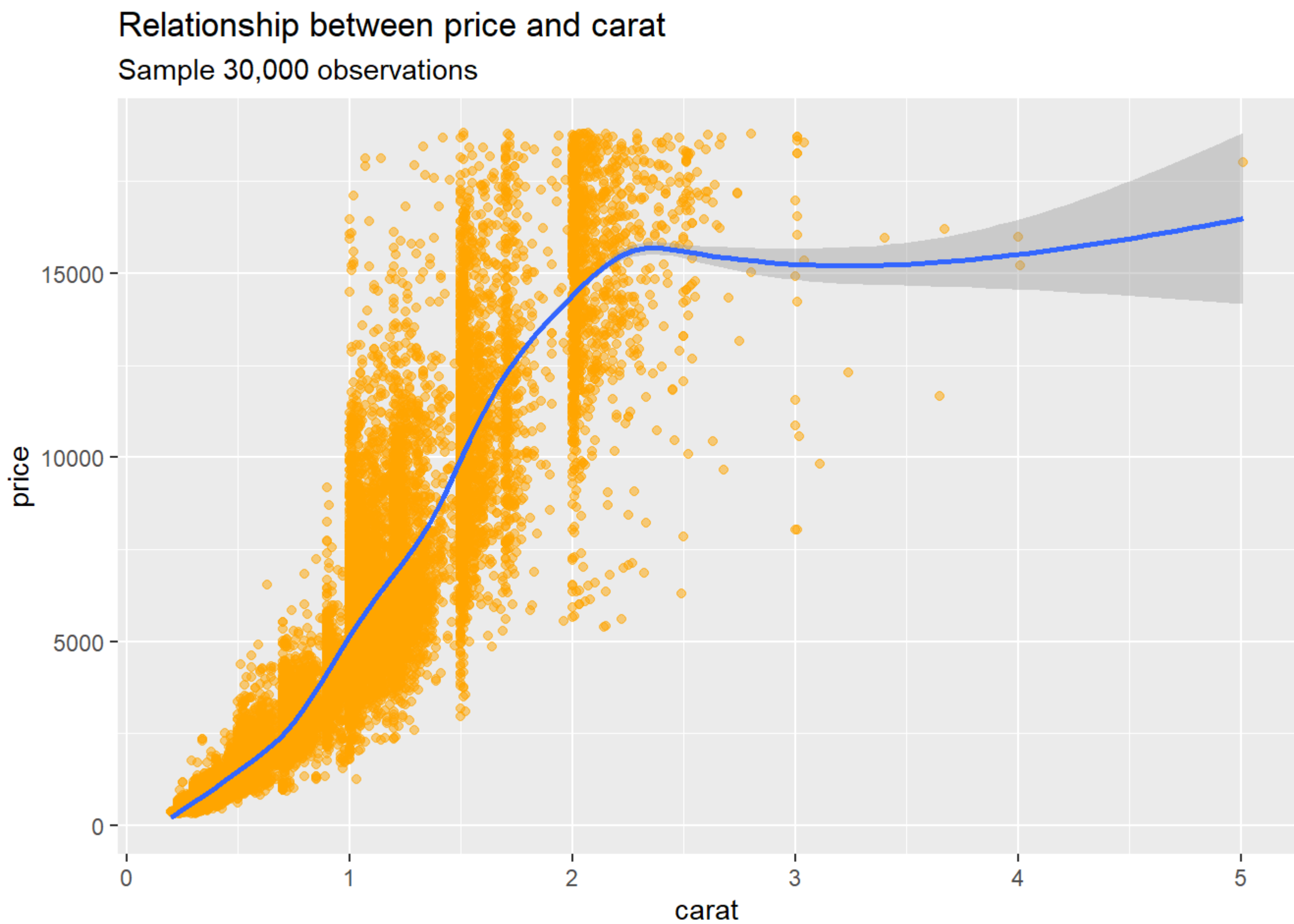
1. There is a positive correlation between carat and x, y, and z, indicating that an increase in carat may predict an increase in diamond size or weight.
2. Price has a positive correlation with carat, x, y, and z, indicating that an increase in carat, x, y, and z results in an increase in price.
3. The depth and the table have a negative correlation, which means that increasing the depth predicts a decrease in the table.

Check relationship between price and carat

I will sampling() function to sampling 30,000 observations and use set.seed() function to guarantee the same random value every time running the code.

```
set.seed(42)
ggplot(diamonds%>% sample_n(30000),aes(carat,price))+
  geom_point(alpha = 0.5, col= "orange")+
  geom_smooth()+
  labs(title = "Relationship between price and carat",
       subtitle = "Sample 30,000 observations")
```

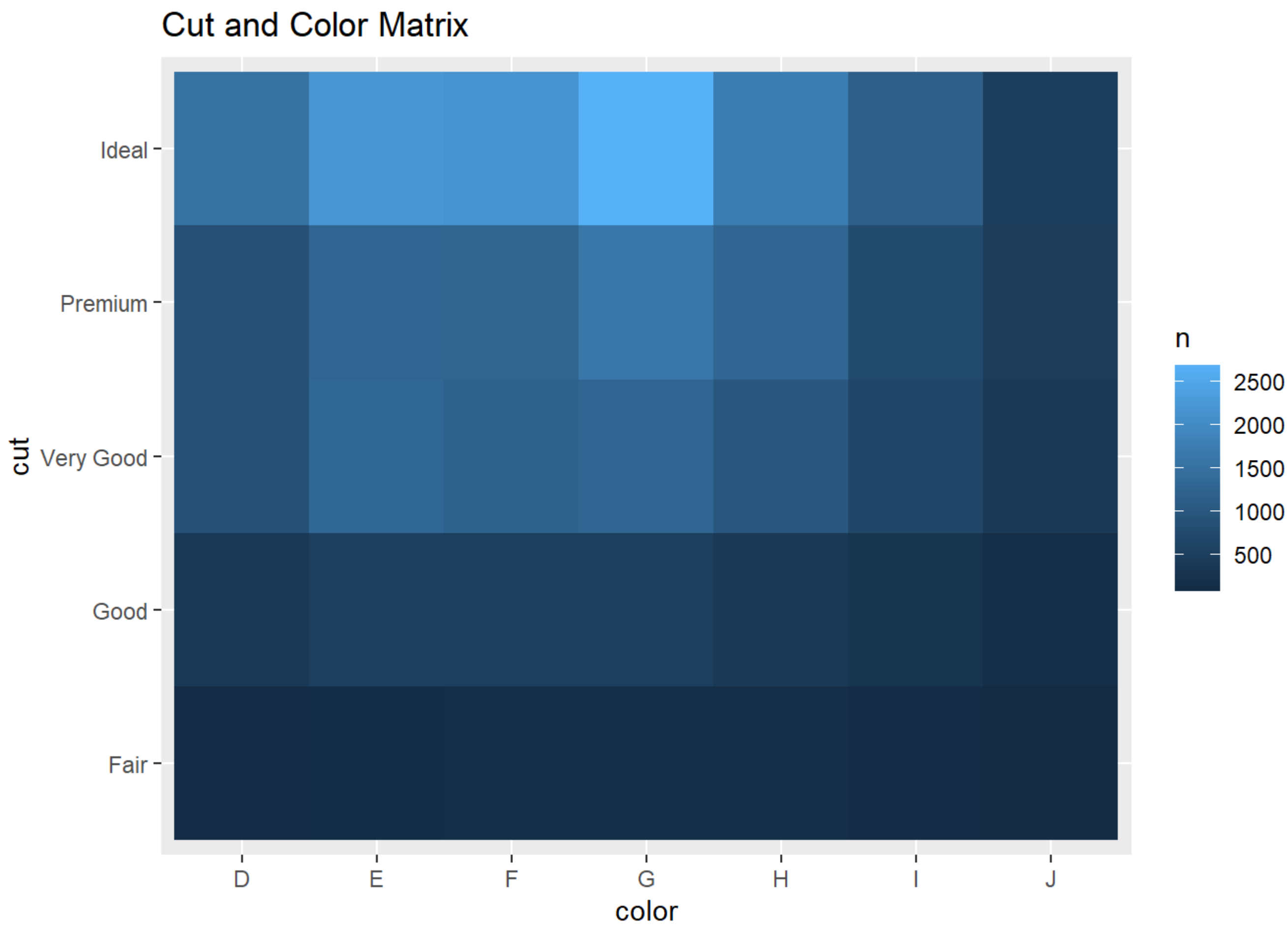
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



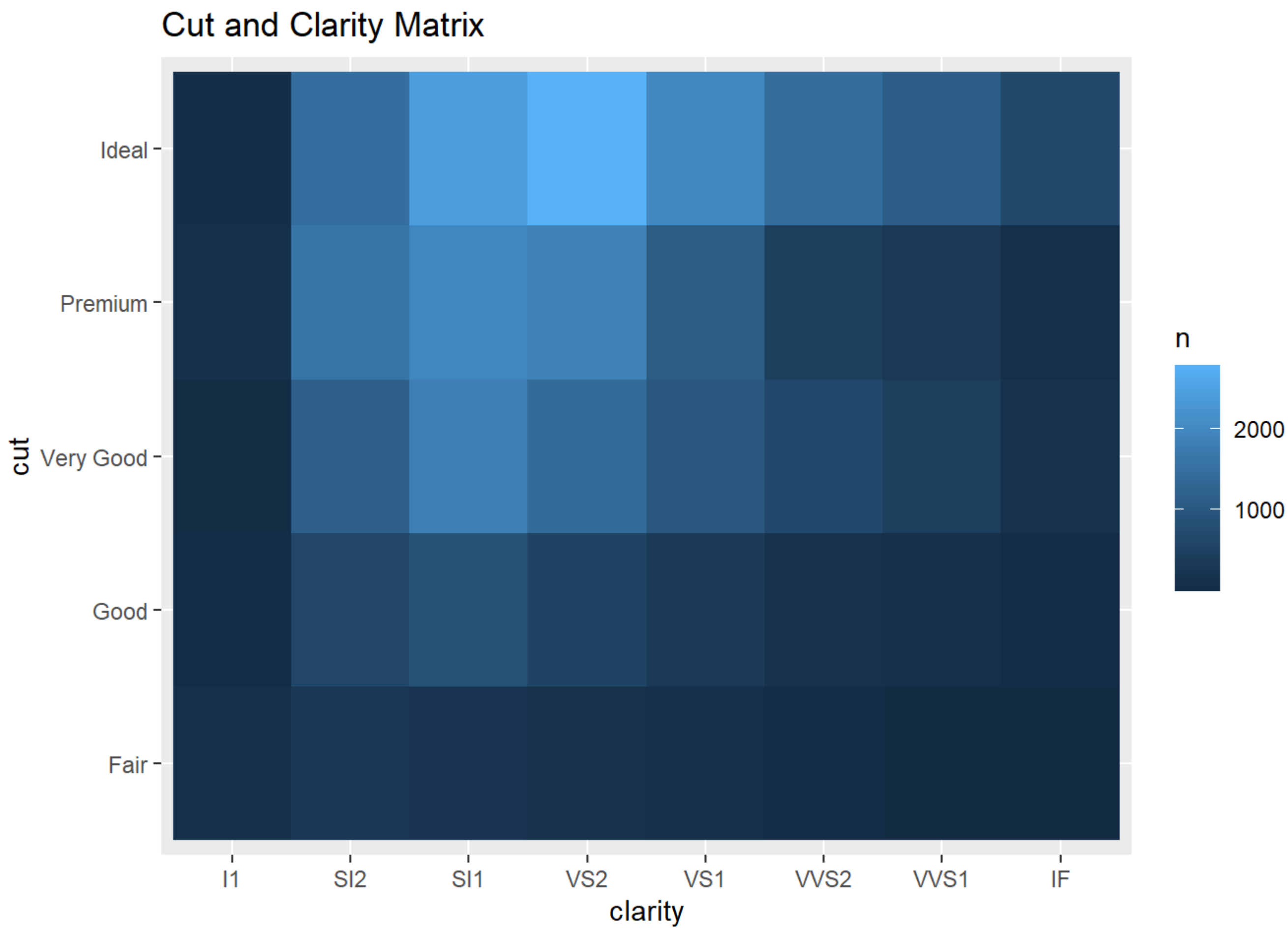
As you can see, there is a positive correlation between price and carat; as the carat increases, so does the price.

Categorical vs Categorical Variables

```
set.seed(42)
diamonds %>%
  sample_n(30000) %>%
  count(color, cut) %>%
  ggplot(aes(x = color, y = cut)) + geom_tile(aes(fill = n)) +
  labs(title = "Cut and Color Matrix")
```




```
set.seed(42)
diamonds %>%
  sample_n(30000) %>%
  count(cut, clarity) %>%
  ggplot(aes(x = clarity, y = cut)) + geom_tile(aes(fill = n)) +
  labs(title = "Cut and Clarity Matrix")
```

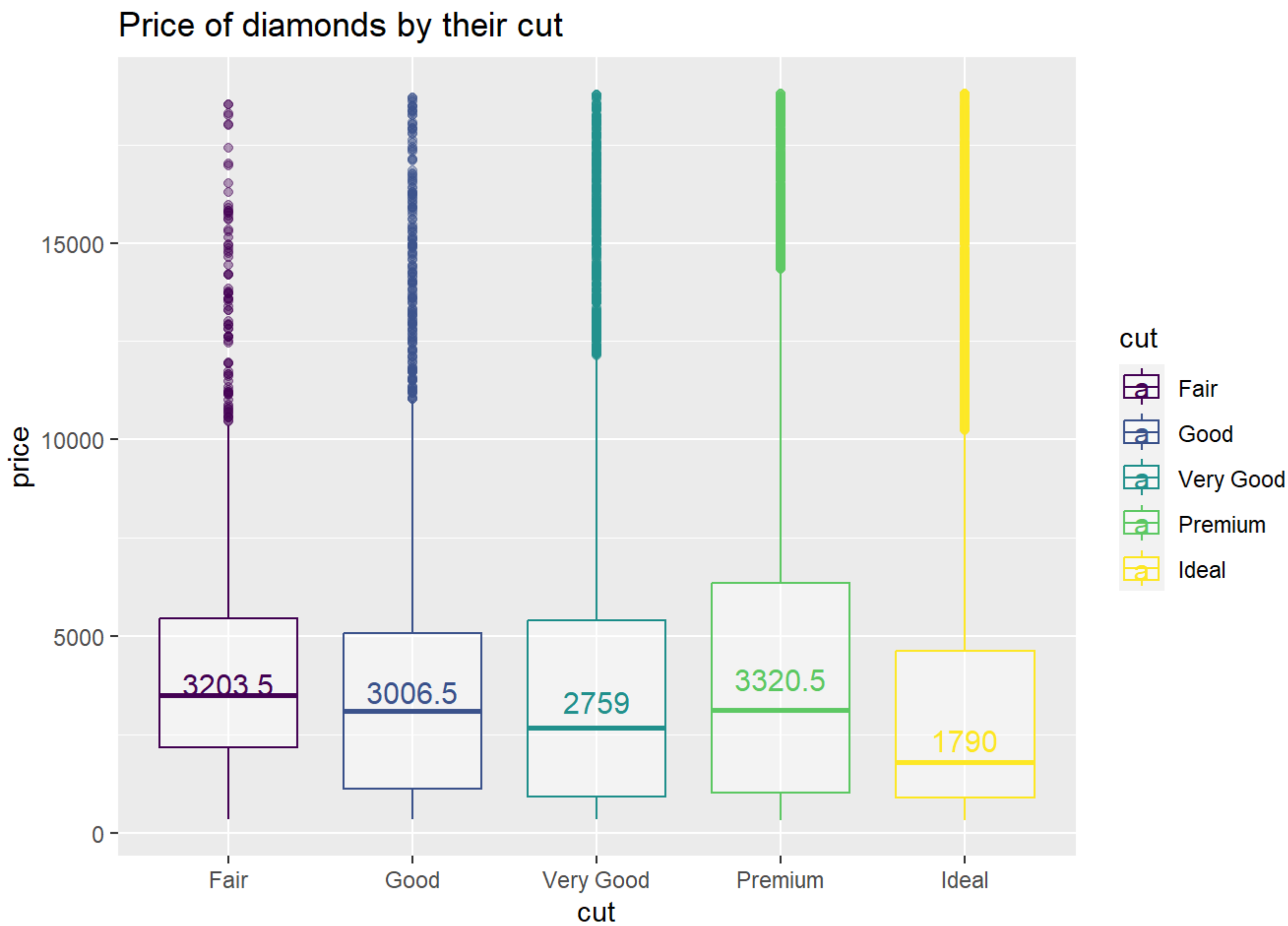


From 30,000 observations, the majority of diamonds are ideal cut diamonds with near-colorless (G scale) color and very small inclusions (VS2 grade).

Categorical vs Numerical Variables

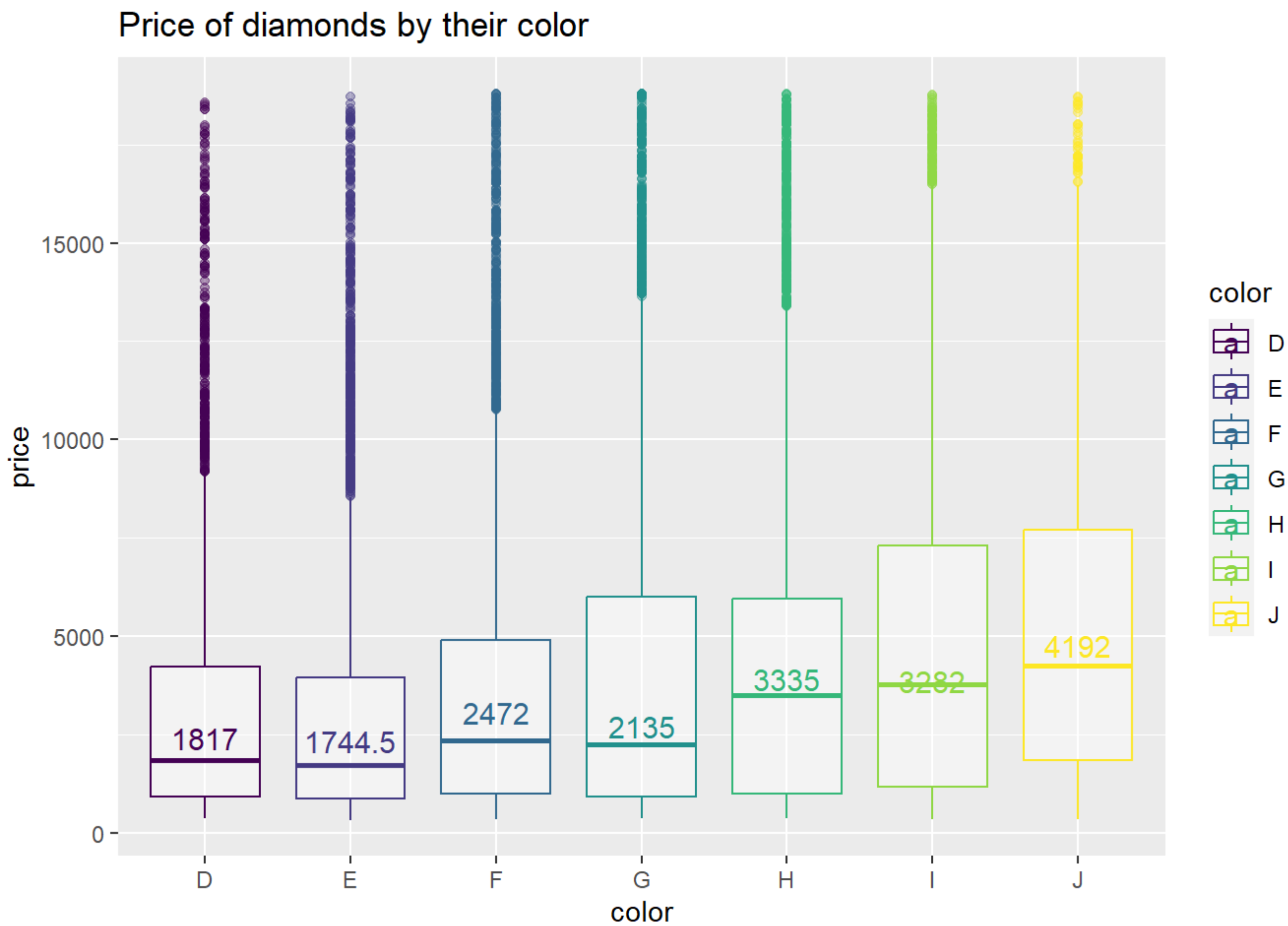
Check the relationship between the price with cut, color, and clarity

```
set.seed(42)
median_cut <- diamonds %>%
  sample_n(3000)%>%
  group_by(cut) %>%
  summarise(median = median(price))

  ggplot(diamonds %>% sample_n(30000),aes(cut, price, color = cut)) +
  geom_boxplot(alpha = 0.4) +
  labs(title = "Price of diamonds by their cut") +
  geom_text(data = median_cut, aes(cut, median, label = median),
            position = position_dodge(width = 0.5), size = 4, vjust = -0.5)
```



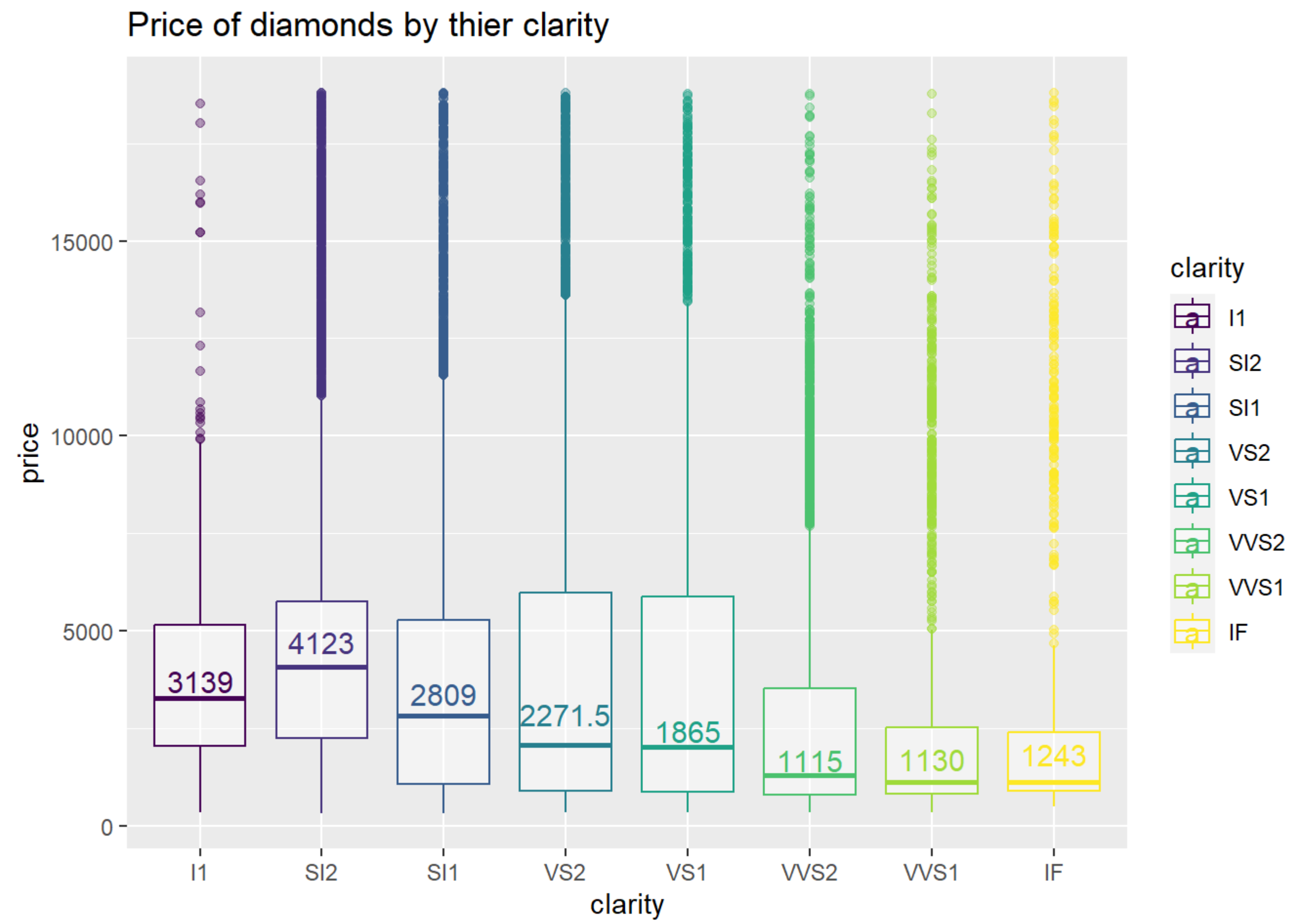
As you can see, the premium cut has the highest median price of 3,320.5 , followed by the fair cut, good cut, very good cut, and ideal cut (Premium> Fair> Good> Very Good> Ideal).

```
set.seed(42)
median_color <- diamonds %>%
  sample_n(3000)%>%
  group_by(color) %>%
  summarise(median = median(price))

  ggplot(diamonds %>% sample_n(30000),aes(color, price, color = color)) +
  geom_boxplot(alpha = 0.4) +
  labs(title = "Price of diamonds by their color") +
  geom_text(data = median_color, aes(color, median, label = median),
            position = position_dodge(width = 0.5), size = 4, vjust = -0.5)
```



This section shows the J color, which is the lowest grade of color from ‘D’ to ‘J’ and has the highest median of 4,192 prices, followed by the H color and then the I, F, G, D, and E colors (J> H> I> F> G> D> E).

```
set.seed(42)
median_clarity <- diamonds %>%
  sample_n(3000)%>%
  group_by(clarity) %>%
  summarise(median = median(price))

  ggplot(diamonds%>%sample_n(30000),aes(clarity, price, color = clarity)) +
  geom_boxplot(alpha = 0.4) +
  labs(title = "Price of diamonds by thier clarity") +
  geom_text(data = median_clarity, aes(clarity, median, label = median),
            position = position_dodge(width = 0.5), size = 4, vjust = -0.5)
```



This section shows that SI2 has the highest median of 4,123 prices, followed by I1 with 3139 prices, SI1 with 2,809 prices, then VS2, VS1, IF, VVS1, and VVS2 (SI2> I1> SI1> VS2> VS1> IF> VVS1> VVS2> VVS2).

Conclusion

I divided the exploration into three parts to investigate numerical vs. numerical variables, category vs. category variables, and category vs. numerical variables by creating a visualization with ggplot. The analysis reveals: 1. There is a strong positive relationship between price and carat, x, y, and z, which means that diamonds with a higher carat, larger size, and heavier weight tend to be more expensive. 2. The most popular diamonds are ideal-cut diamonds with near-colorless and very small inclusions clarity.