# HW_Data Visualization

Buncha Art

2023-05-24

# Home Work : Live 08 Data Visualization

## - Explore dataframe “Diamonds” and create 5 charts

Image : Diamond Earrings, source : unsplash

In this project, we will explore the dataframe “Diamonds”.

```
library(tidyverse)
library(patchwork)
library(ggplot2)
```

First let’s glimpse() the dataframe.

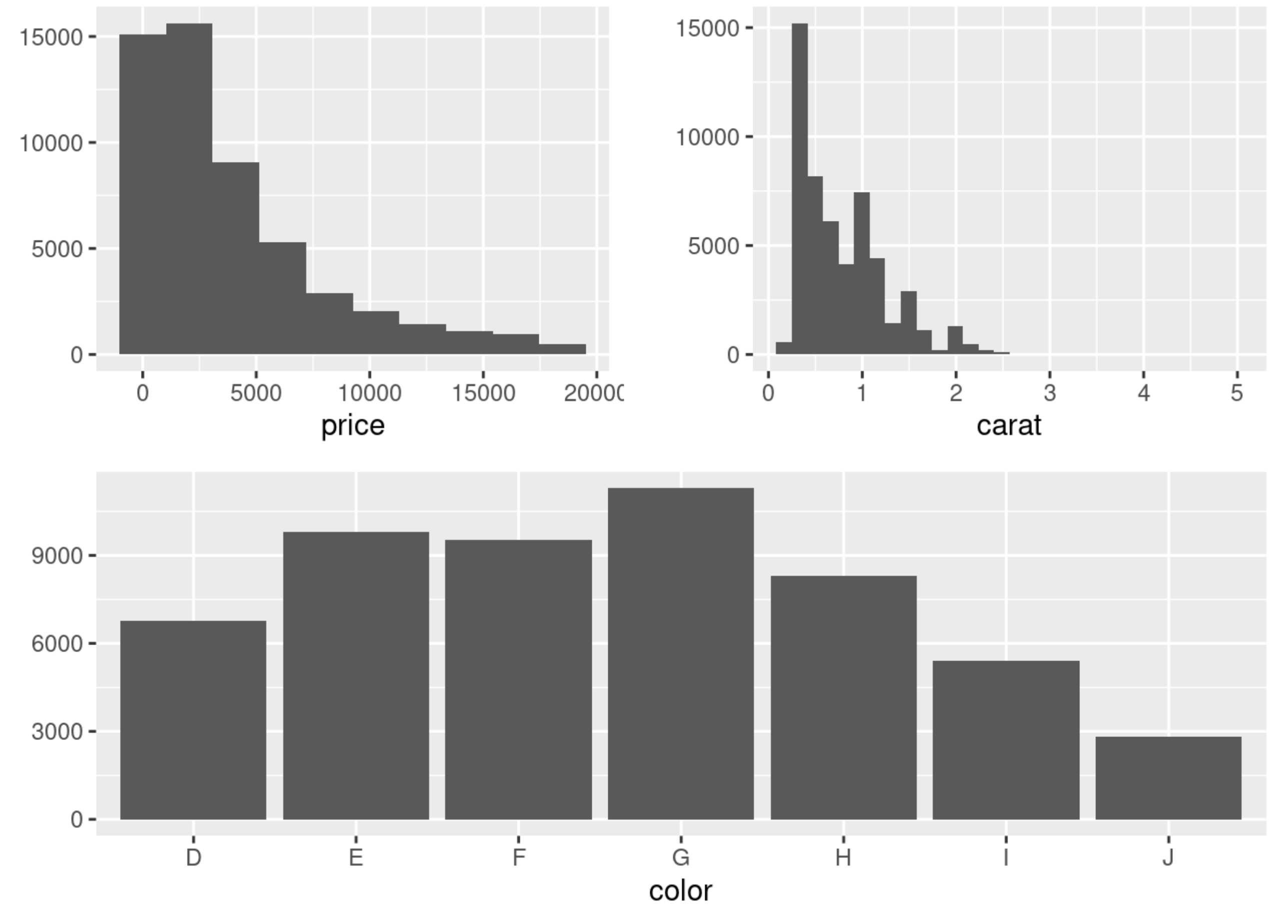
```
glimpse(diamonds)
```

```
## 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.…
```

Diamonds dataframe consists of 53,940 rows and 10 Columns.
Let’s see what we will find from it.

# Chart 1 : Summary of price, carat and cut. (use library(patchwork))

```
d_qp1_hist_price = qplot(price, data=diamonds, geom="histogram", bins=10)
d_qp2_hist_carat = qplot(carat, data=diamonds, geom="histogram")
d_qp3_bar_color = qplot(color, data=diamonds, geom="bar")
(d_qp1_hist_price + d_qp2_hist_carat) / d_qp3_bar_color
```

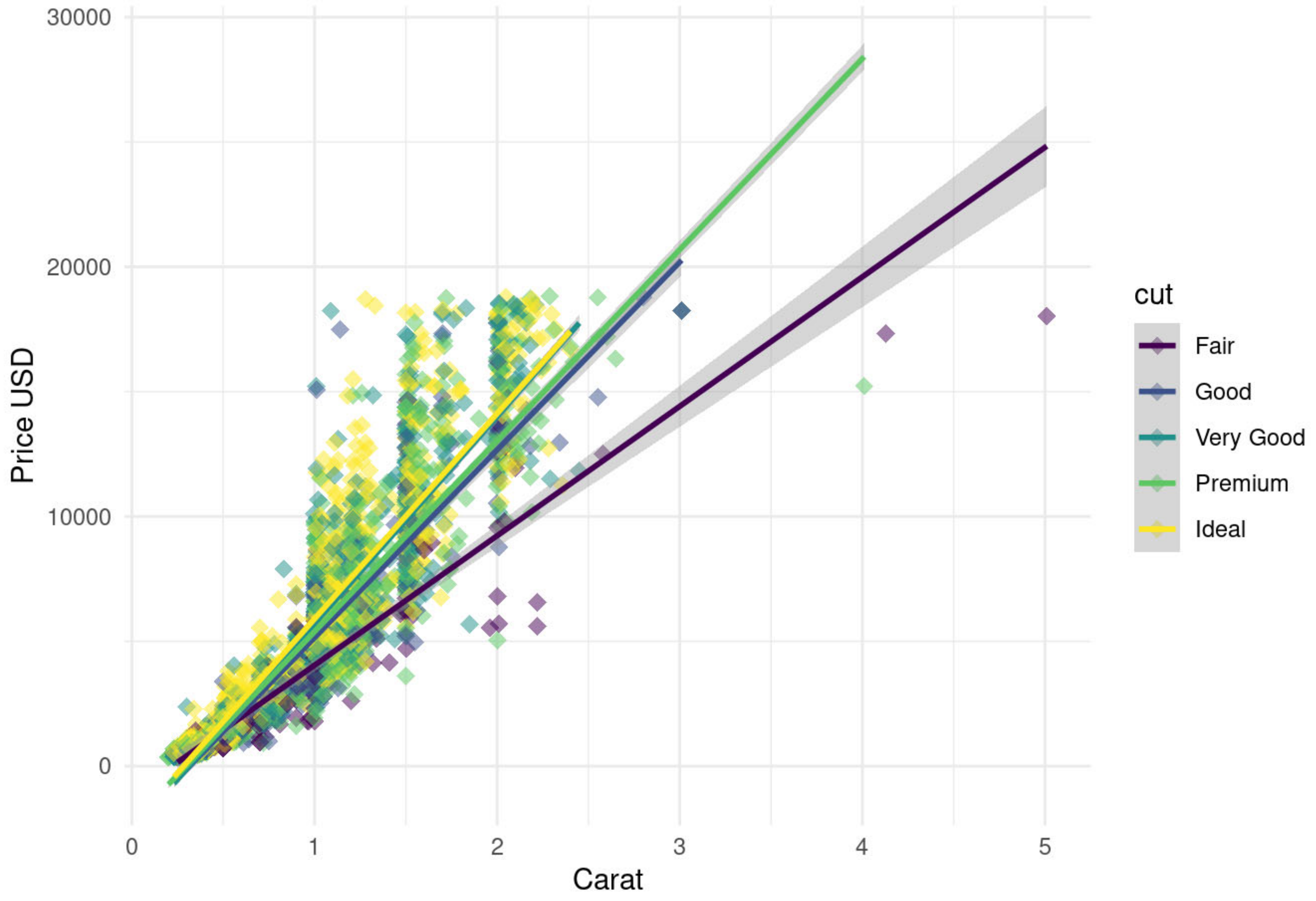


Explanation : We found that :

- Price : main distribution is in segment lower than 5,000 USD
- Carat : main distribution is in segment lower than 0.5 carat
- Color : main distribution is G > E > F > H > D > I > J

## Chart 2 : Relationship between Carat and Price (Sample size = 10%)

```
ggplot(diamonds %>% sample_frac(0.1),
       mapping = aes(carat, price, col=cut)) +
  geom_point(size=3, shape=18, alpha=0.5) +
  geom_smooth(method="lm") +
  theme_minimal() +
  labs(
    x = "Carat",
    y = "Price USD",
    caption = "Datasource : Diamonds ggplot2"
  )
```



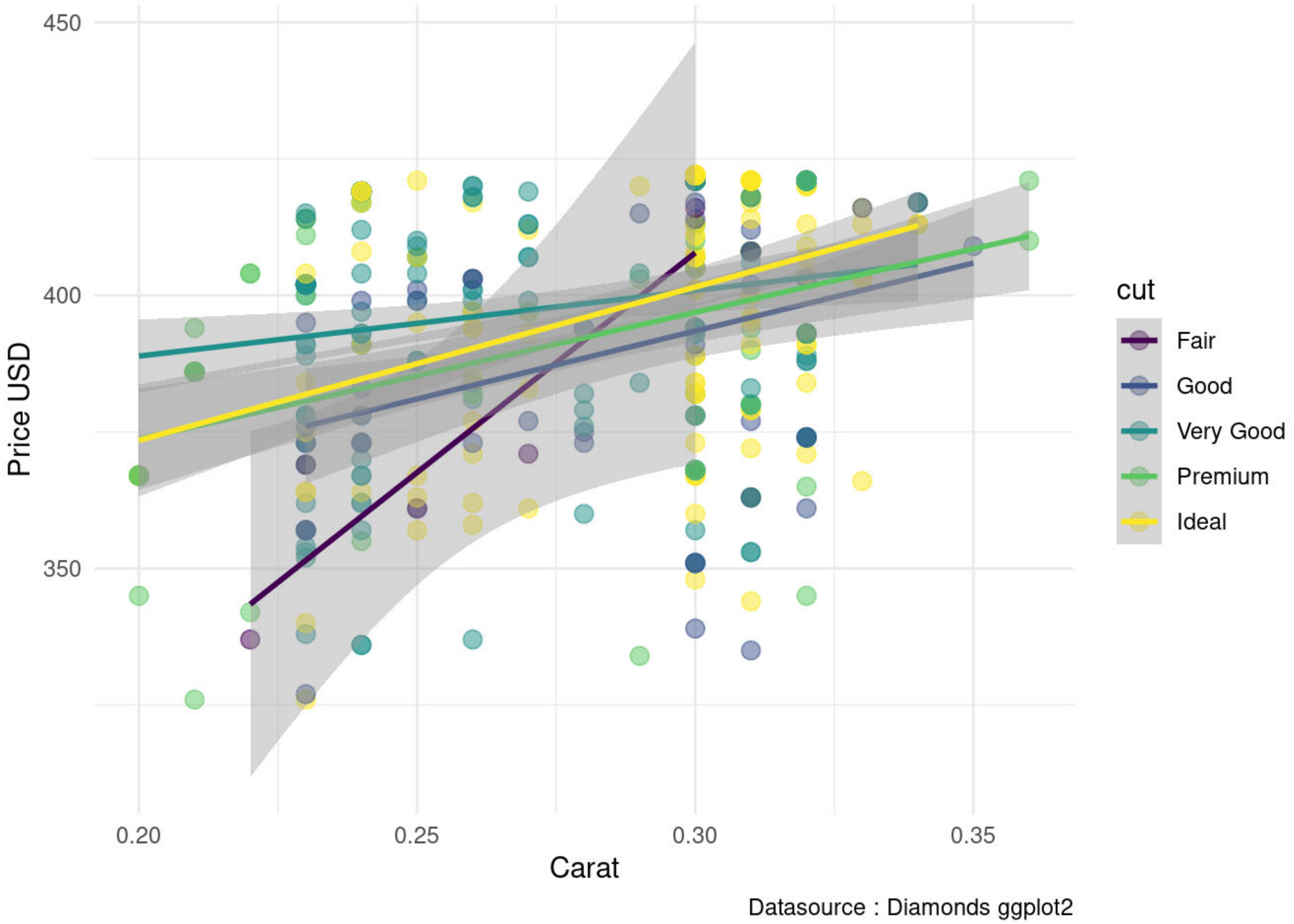
Datasource : Diamonds ggplot2

Explanation : We found the positive correlation between the Carat and Price in every type of cut.

If we look at the top 500 most expensive diamonds and bottom 500 least expensive diamonds, will it show the same result?. Let’s see.

# Chart 3 : Relationship between Carat and Price of the bottom 500 least expensive diamonds

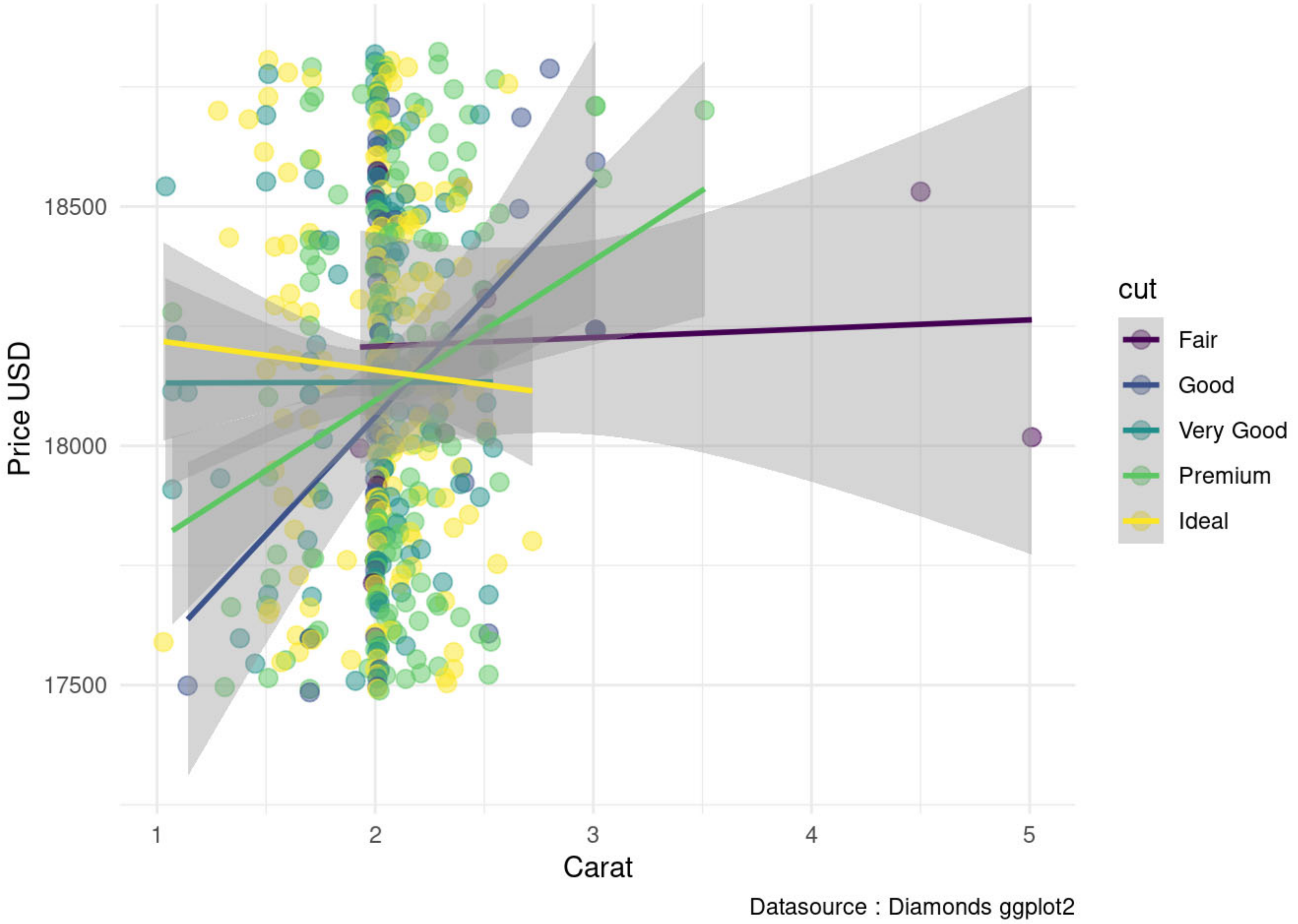
```
bottom_500 <- diamonds %>%
  arrange(price) %>%
  slice(1:500)
ggplot(bottom_500, mapping = aes(carat, price, col=cut)) +
  geom_point(size=3, alpha=0.5) +
  geom_smooth(method = "lm") +
theme_minimal() +
  labs(
    x = "Carat",
    y = "Price USD",
    caption = "Datasource : Diamonds ggplot2"
  )
```



Explanation : for bottom 500 least expensive diamonds, we found positive correlation between carat and price especially for the cut of “Fair”.

# Chart 4 : Relationship between Carat and Price of the top 500 most expensive diamonds

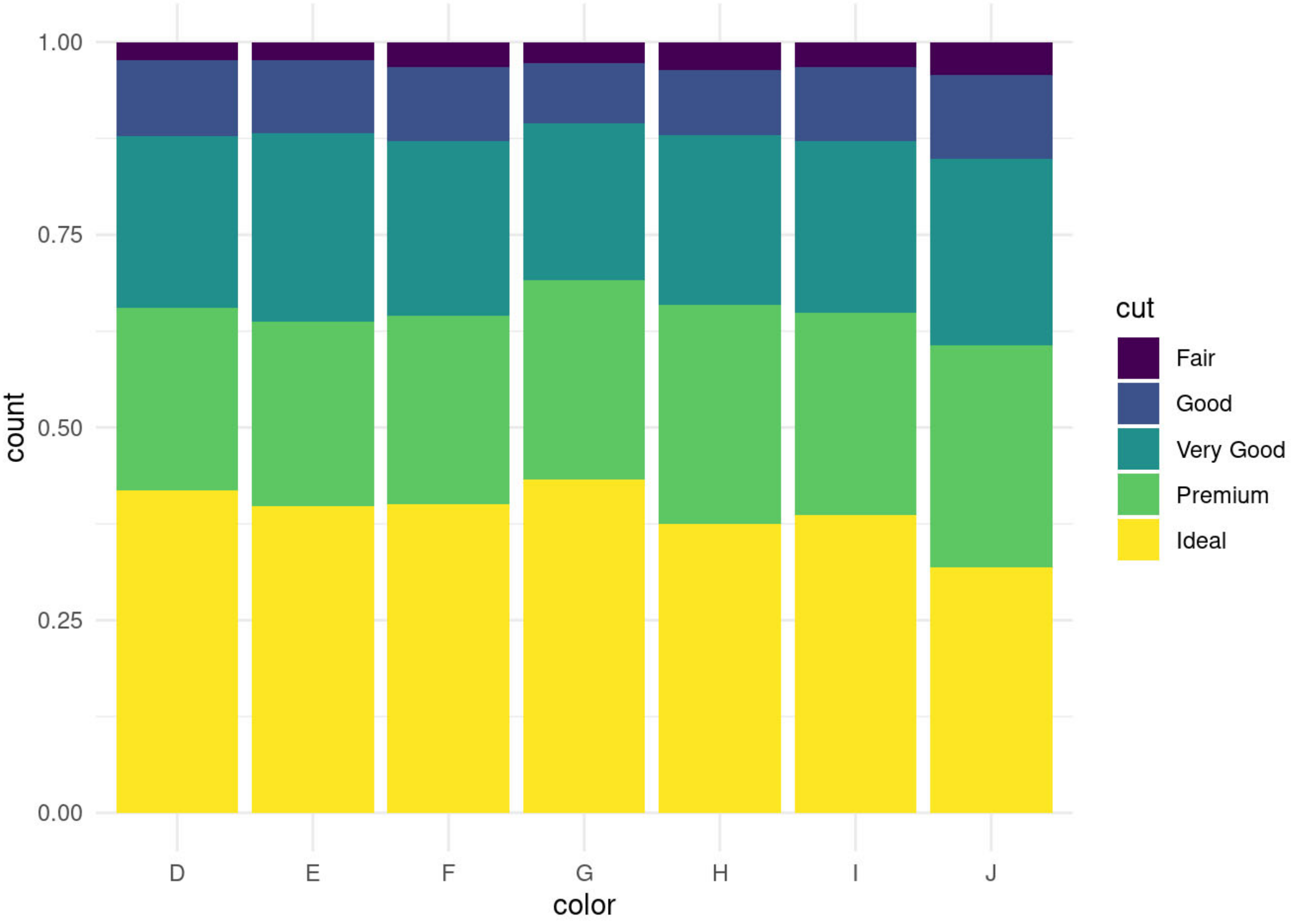
```
top_500 <- diamonds %>%
  arrange(desc(price)) %>%
  slice(1:500)
ggplot(top_500, mapping = aes(carat, price, col=cut)) +
  geom_point(size=3, alpha=0.5) +
  geom_smooth(method = "lm") +
  theme_minimal() +
  labs(
    x = "Carat",
    y = "Price USD",
    caption = "Datasource : Diamonds ggplot2"
  )
```



Explanation : for top 500 most expensive diamonds, we found positive correlation between carat and price for the cut of “Good” and “Premium”, found slightly neutral correlation for cut of “Fair”and “Very Good” and found a slightly negative correlation for cut of “Ideal”.

# Chart 5 : Proportion of diamonds based on color sort by cut

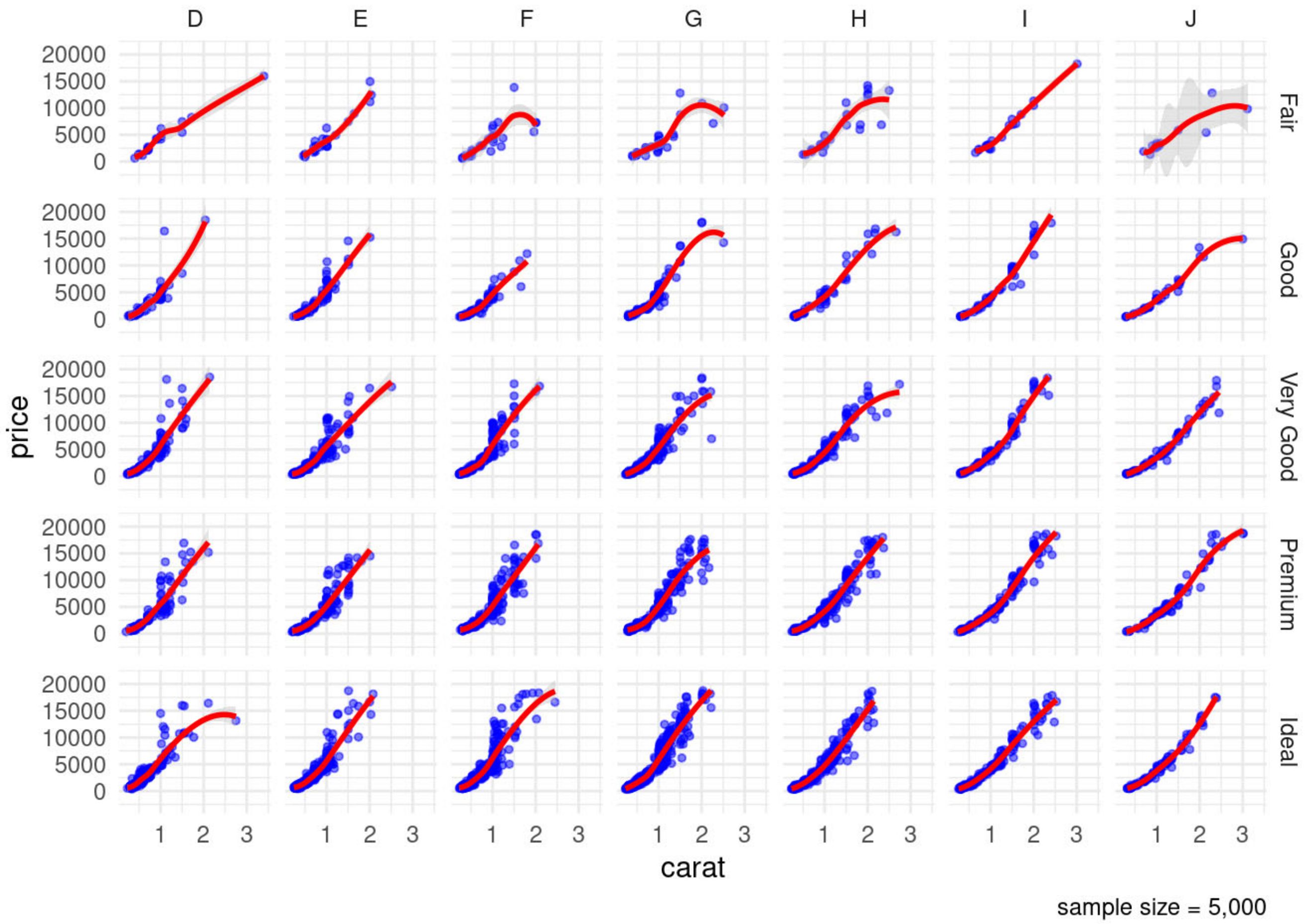
```
ggplot(diamonds, aes(color, fill=cut)) +
  geom_bar(position = "fill") +
  theme_minimal()
```



Explanation : we found that when categorized by color, the proportion of cut type sort from greatest to least is Ideal > Premium > Very Good > Good > Fair.

# Chart 6 : Relationship between carat and price (grid of cut, color)

```
set.seed(42)
ggplot(diamonds %>% sample_n(5000), aes(carat, price)) +
  geom_point(size=1, alpha=0.5, col="blue") +
  geom_smooth(method="loess", col="red", fill="grey") +
  theme_minimal() +
  facet_grid(cut ~color)+
  labs(
    caption = "sample size = 5,000"
  )
```



Explanation : we found positive correlation between carat and price in all type of cut and color. (Sample size = 5,000)

Thank you very much for your interest.

May your life shine bright like a diamond.