Homework Sprint 5 Data Viz

Phattharachai Maichin

2023-09-15

Built 5 chart from diamonds dataframe in R

[1] Prepared before work with dataset

```
##install package first then call library, ggplot is inside 'tidyverse'
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   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     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 error
```

```
#View(diamonds)
#glimpse(diamonds)

##install package 'patchwork' to to make plot composition in R extremely simple and powerful (In this w
library(patchwork)

## Split data set by 25% to more quick analysis and set.seed() to lock the sampling
set.seed(28); sp_diamonds <- sample_frac(diamonds, 0.25)
```

[2] What each variable mean?

First, we should understand our data variable

Source: Chapter 5 The diamonds dataset

Variable	Description
price	price in US dollars
carat	weight of the diamond
cut	quality of the cut
color	diamond color
clarity	measurement of how clear the diamond is I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
x	length in mm
y	width in mm

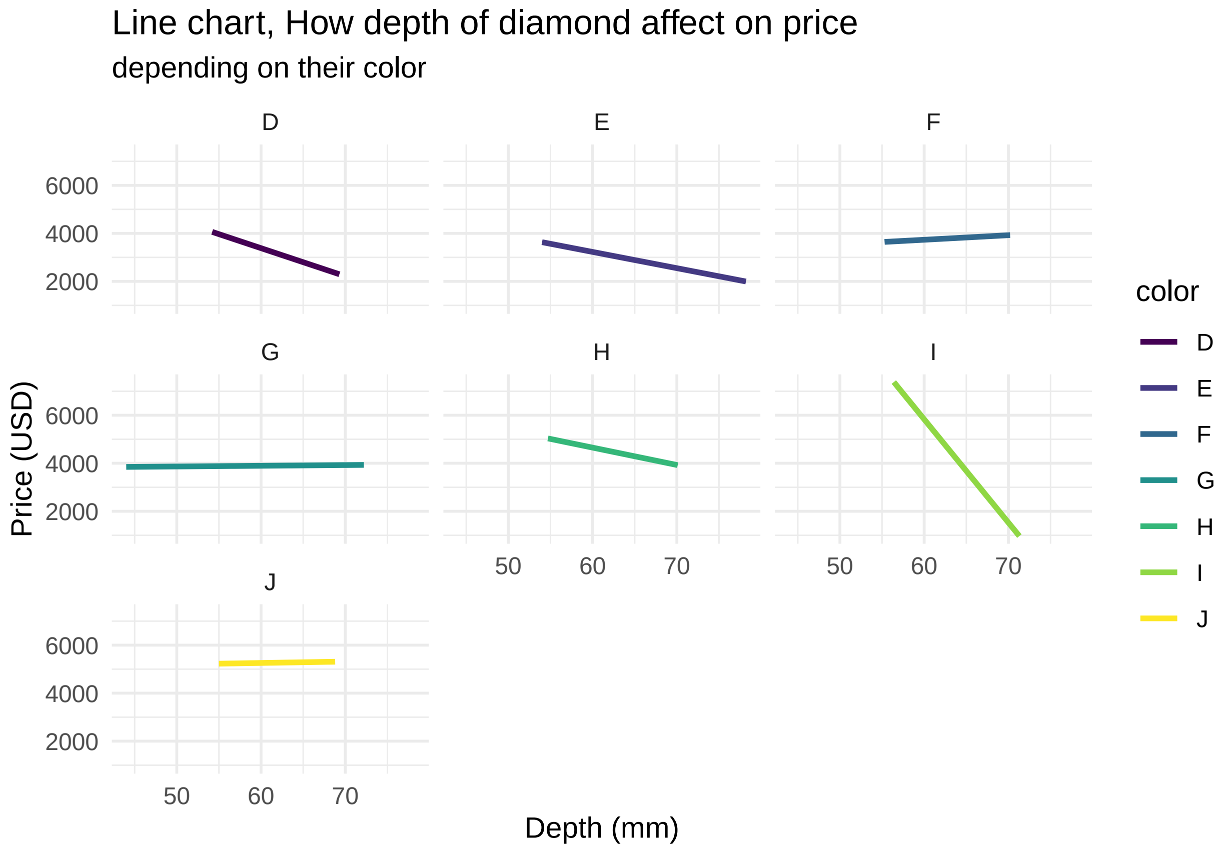
Variable	Description
z	depth in mm
depth	total depth percentage
table	width of top of diamond relative to widest point

[3] Question and explore data

Question 1:

```
## How depth of diamond affect on price depending on their color
#Note: use sample 10%
Q1 <- ggplot(data = sp_diamonds,
      mapping = aes(x = depth, y = price, col=color)) +
      theme_minimal() +
      geom_smooth(method = "lm", se = FALSE) +
      facet_wrap(~color) +
      labs(title="Line chart, How depth of diamond affect on price",
           subtitle = "depending on their color",
           x="Depth (mm)",
           y="Price (USD)")
(Q1)
```

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



From this line chart, it can be splited how diamond's depth affects on price based on their color into two groups: first group including "F, G, J" color that means no matter depth of

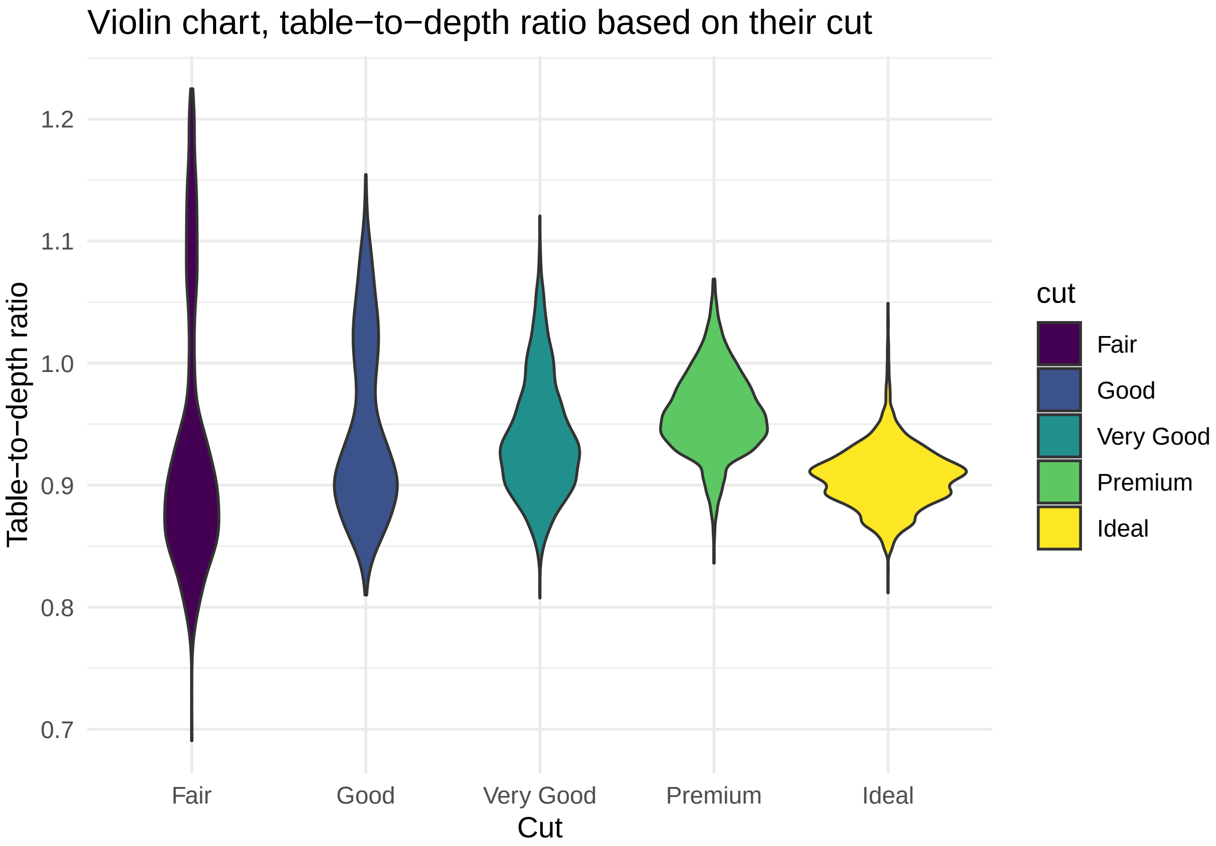
the diamond it is, price are almost the same and other group including "D, E, H, I" color that means size of diamond's depth have a negative correlation effect on price.

Question 2:

```
## What is distribution of (Table to depth ratio) based on their cut?
ratio <- sp_diamonds$table/sp_diamonds$depth

Q2 <- ggplot(data = sp_diamonds,
      mapping = aes(x = cut, y = ratio, fill=cut)) +
      theme_minimal() +
      geom_violin() +
      labs(title="Violin chart, table-to-depth ratio based on their cut",
      x="Cut",
      y="Table-to-depth ratio")

(Q2)
```



Violin chart showed that median of table to depth ratio of all diamonds' cut are around 0.7-0.9.

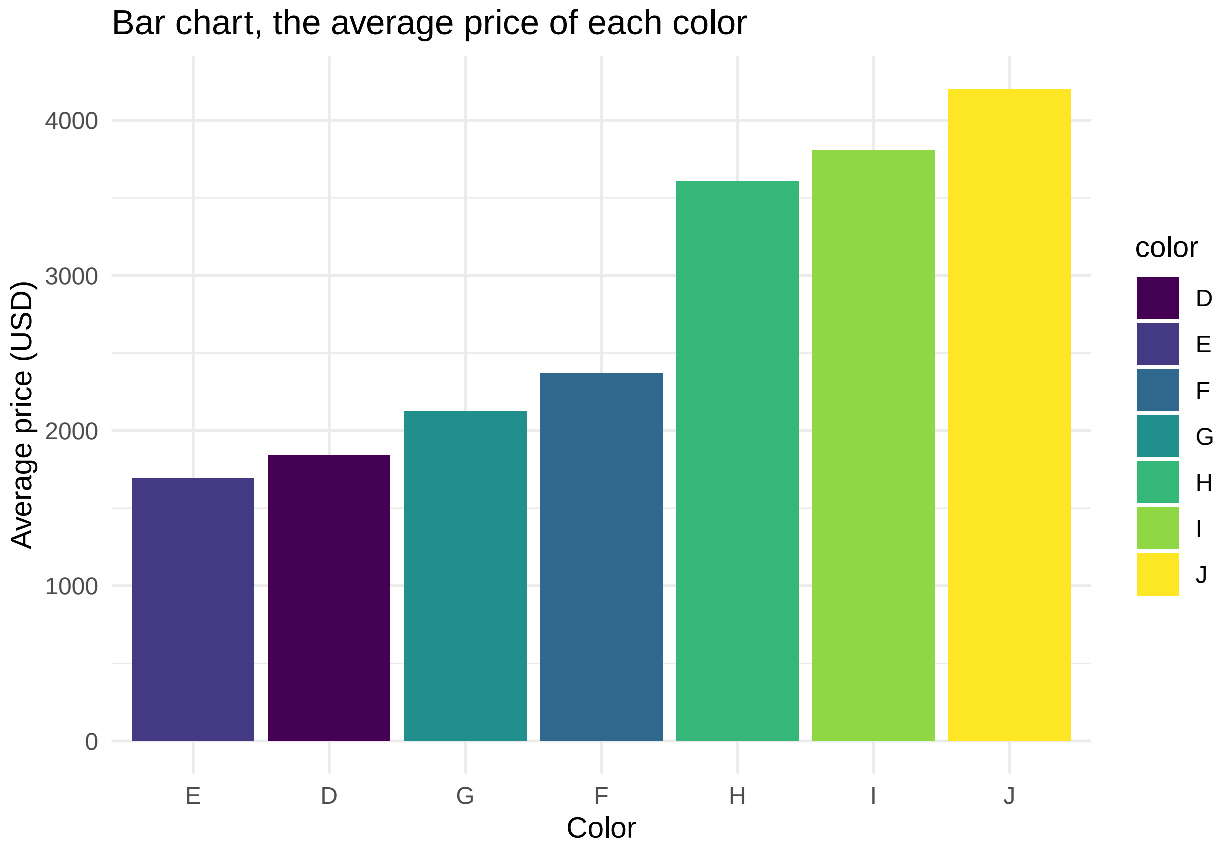
Question 3:

```
## What is the average price of each color?
#Note: Reordering geom_col by value using fct_reorder(group, value)

sp_diamonds2 <- sp_diamonds %>%
                group_by(color) %>%
                summarise(avg_price = median(price))

Q3 <- ggplot(data = sp_diamonds2,
             mapping = aes(x = fct_reorder(color, avg_price), y = avg_price, fill=color)) +
             theme_minimal() +
             geom_col() +
             labs(title="Bar chart, the average price of each color",
             x="Color",
             y="Average price (USD)")

(Q3)
```



Bar chart showed that the sorting color "E, D, G, F, H, I, J" by the cheapest to the expensivest average price of diamonds, respectively.

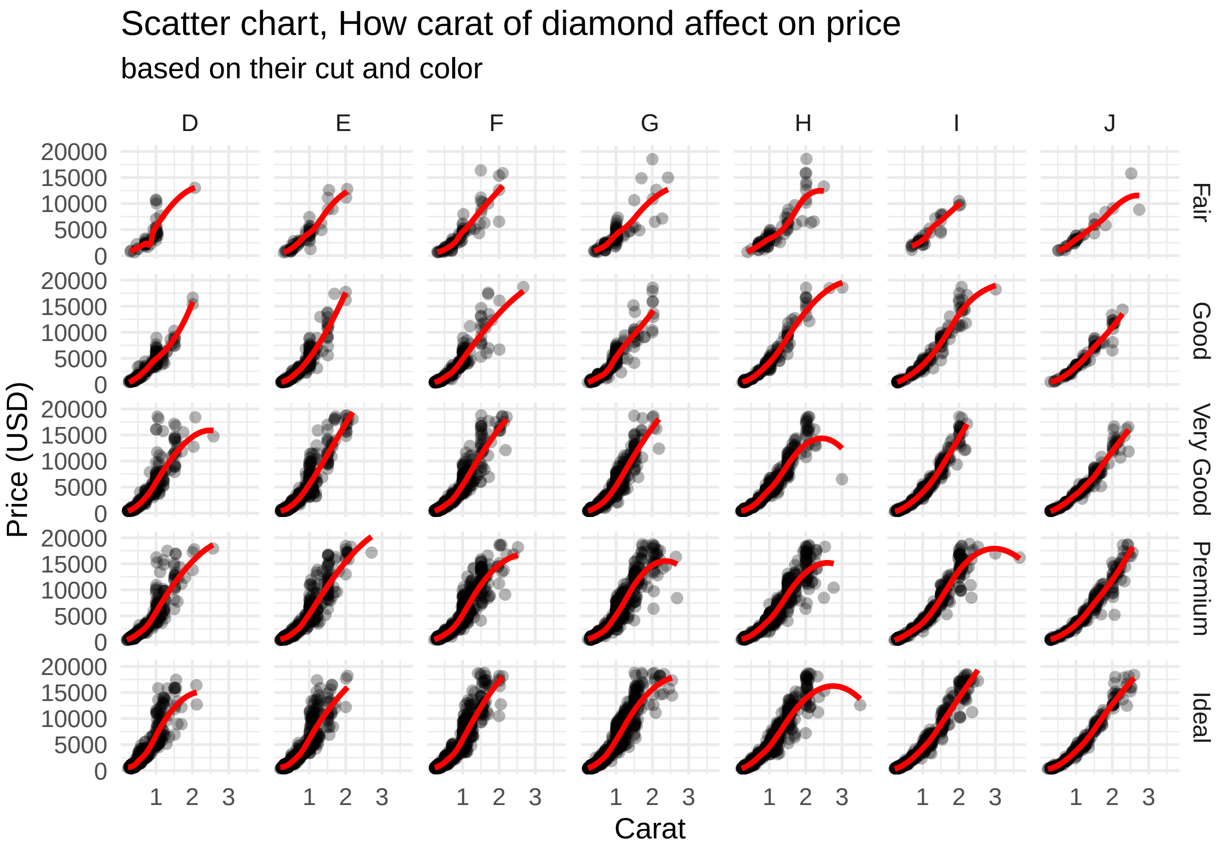
Question 4:

```
## How carat of diamond affect on price based on their cut and color

Q4 <- ggplot(data = sp_diamonds,
             mapping = aes(x = carat, y = price)) +
             theme_minimal() +
             geom_point(alpha = 0.3, col = "black") +
             geom_smooth(method = "loess", se = FALSE, col = "red") +
             facet_grid(cut ~ color) +
             labs(title="Scatter chart, How carat of diamond affect on price",
             subtitle = "based on their cut and color",
             x="Carat",
             y="Price (USD)")

(Q4)
```

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



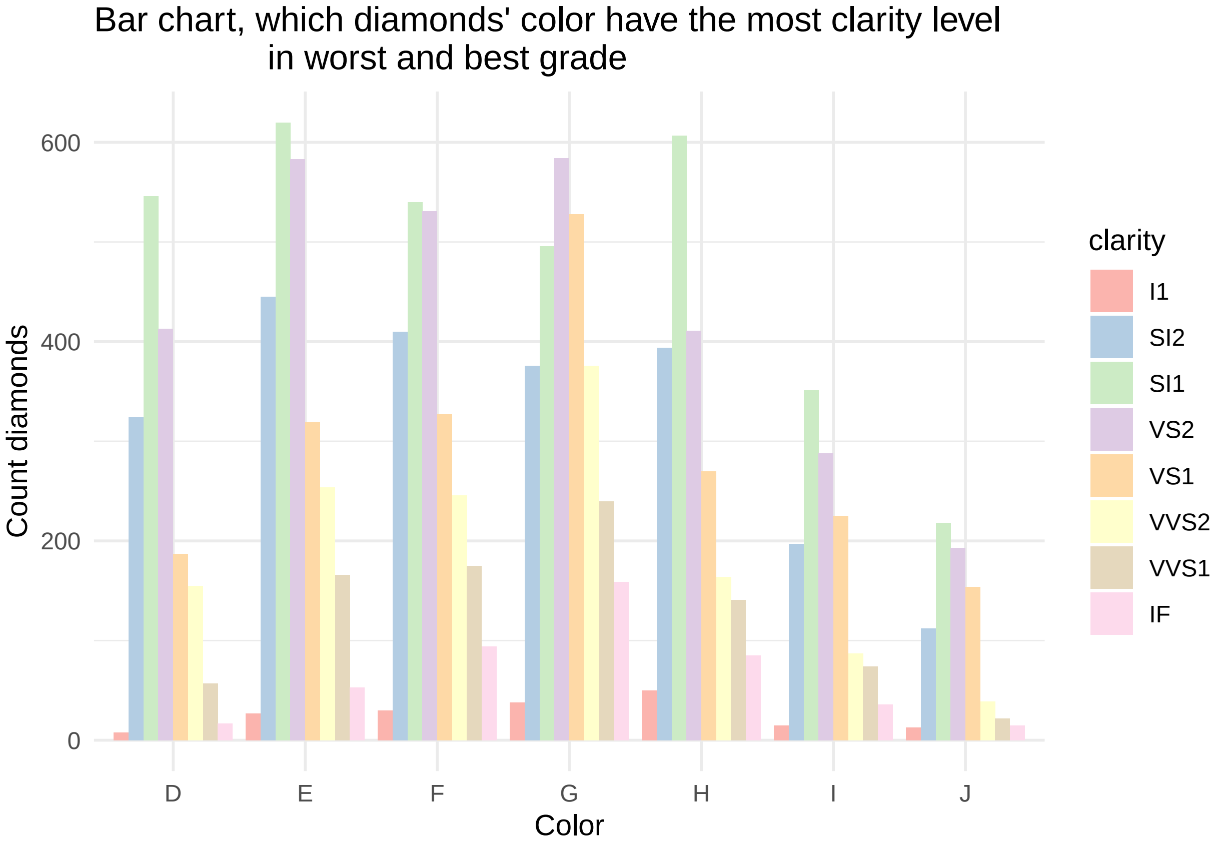
No matter cut or color of diamond it is, there are all the same trend when the carat increased, the price also increased.

Question 5:

```
## Which color has the most clarity level of 'I1' (worst) and 'IF' (best) ?

Q5 <- ggplot(data = sp_diamonds,
             mapping = aes(x = color, fill = clarity)) +
             theme_minimal() +
             geom_bar(position="dodge") +
             scale_fill_brewer(type = 'qual', palette = 4) +
             labs(title = "Bar chart, which diamonds' color have the most clarity level
                  in worst and best grade",
             x="Color",
             y="Count diamonds")
```

```
(Q5)
```



"J" color diamond has the most quantity of "I1" that considered to be the worst clarity quality of diamond. But "G" color diamond has the most quantity of "IF" that considered to be the best clarity quality of diamond.