Data Visualization Bootcamp Homework

Dhanachote Not

2023-07-05

Introduction

Hi my name is Dhanachote, you can call me by my nickname is Not. I am learning how to use rmarkdown and build my first Data Visualization. By the way, this report is for homework using ggplot2 to create data visualization to build in dataset "diamonds" in R studio with five questions.

rmarkdown-cheat sheet

I would like to share rmarkdown-cheat sheet for everyone use their own project or homework!

rmarkdown-cheatsheet

Minimize dataset before making the Data Visualization

The dataset of diamonds it has 53,940 rows that it can be working with this dataset are not comfortable because it can be running the result slower. So, i will random 10 percents of all samples before marking data visualization.

```
## load library
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- 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
```

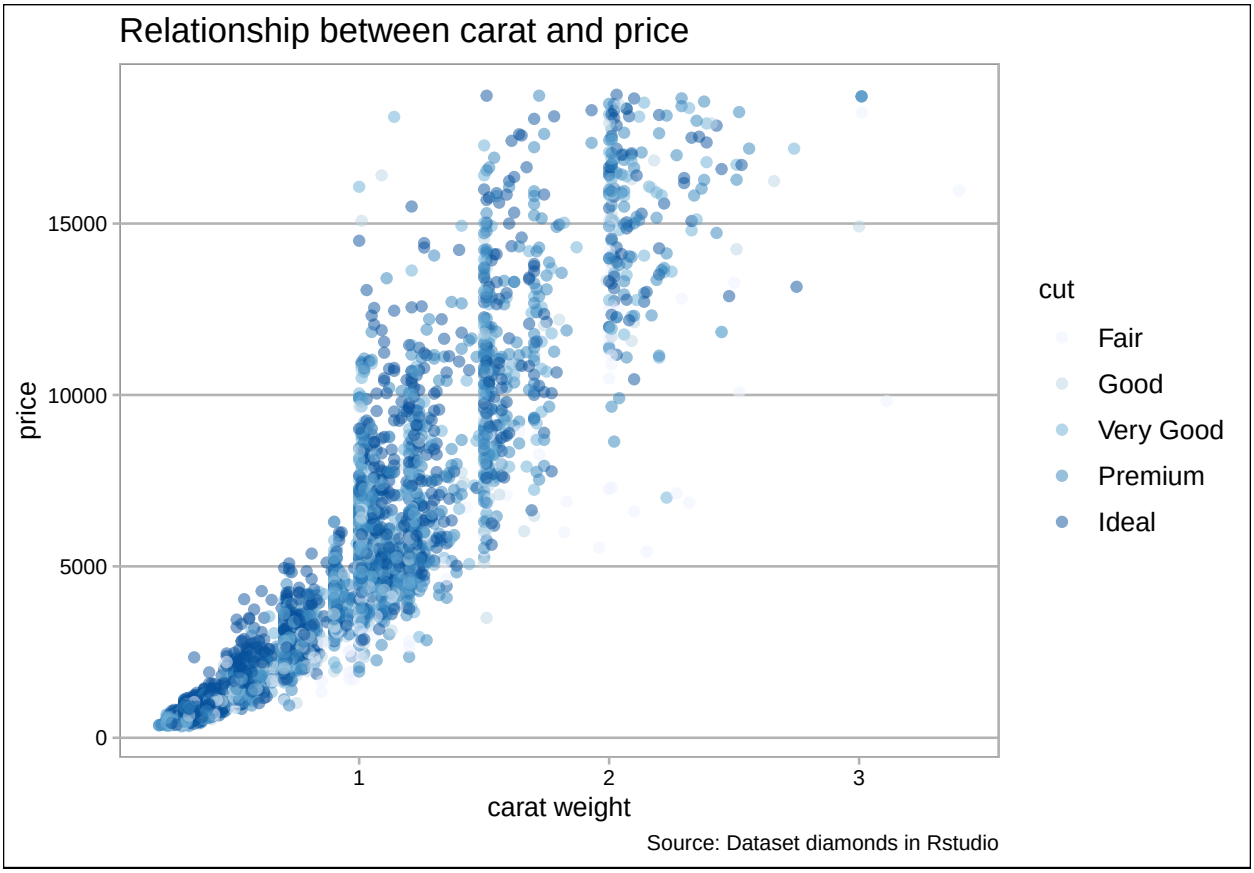
```
library(ggthemes)
## lock random sample for use dataset

set.seed(42)
sample_diamonds <- sample_frac(diamonds, 0.1)
```

Question 1 How dose the price of diamonds vary with carat weight?

```
ggplot(sample_diamonds, aes(carat, price, col = cut)) +
  geom_point(alpha  = 0.5) +
  theme_calc() +
  scale_color_brewer(type = "seq",
```

```
                    palette = 1) +
  labs(
    title = "Relationship between carat and price",
    x = "carat weight",
    y = "price",
    caption = "Source: Dataset diamonds in Rstudio"
  )
```




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

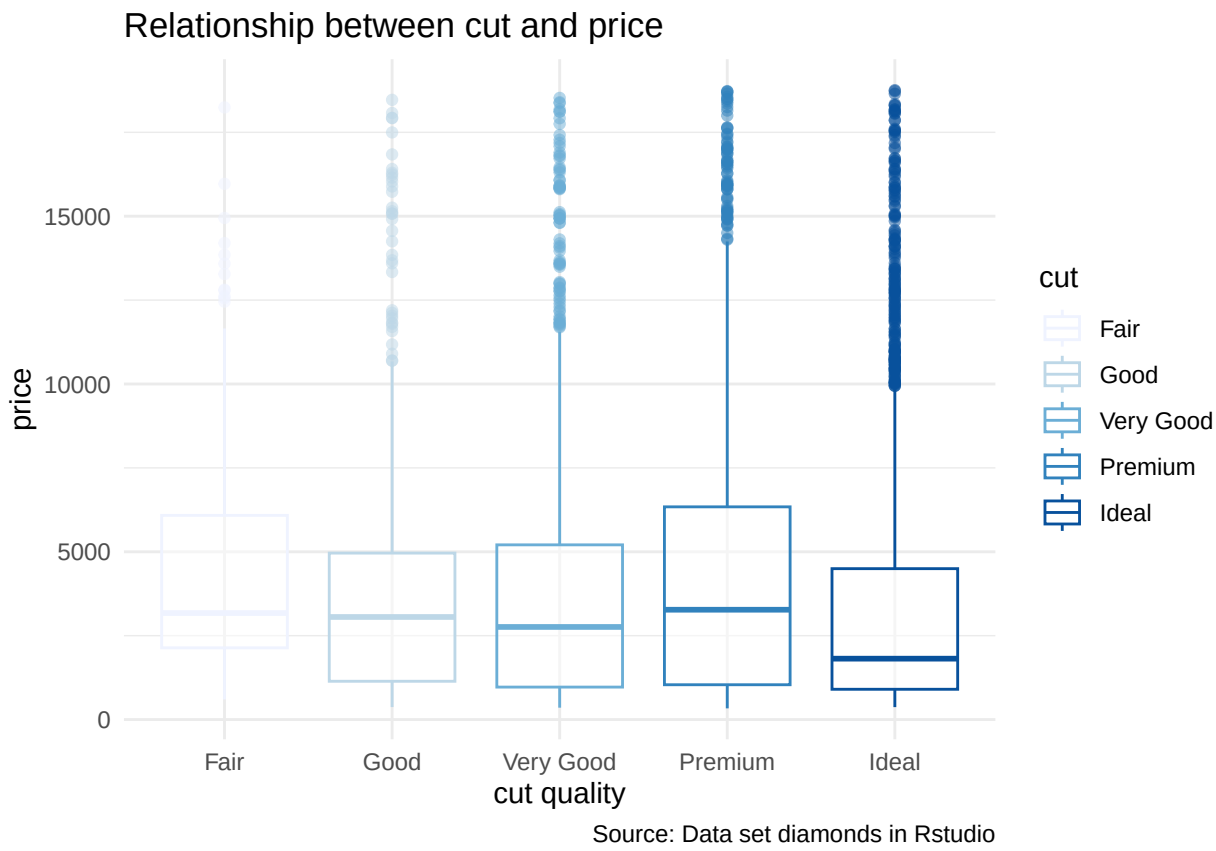
```
## [1] 0.9212468
```

From this plot, we found that the price of diamonds varies with carat weight using correlation = 0.9180853 which means the higher carat might have a higher price.

Question 2 What is the distribution of diamond prices based on their cut quality?

```
ggplot(sample_diamonds, aes(cut, price, col = cut)) +
  geom_boxplot(alpha = 0.5) +
  theme_minimal() +
  scale_color_brewer( type = "seq",
                      palette = 1) +
  labs(
    title = "Relationship between cut and price",
    x = "cut quality",
    y = "price",
    caption = "Source: Data set diamonds in Rstudio"
```

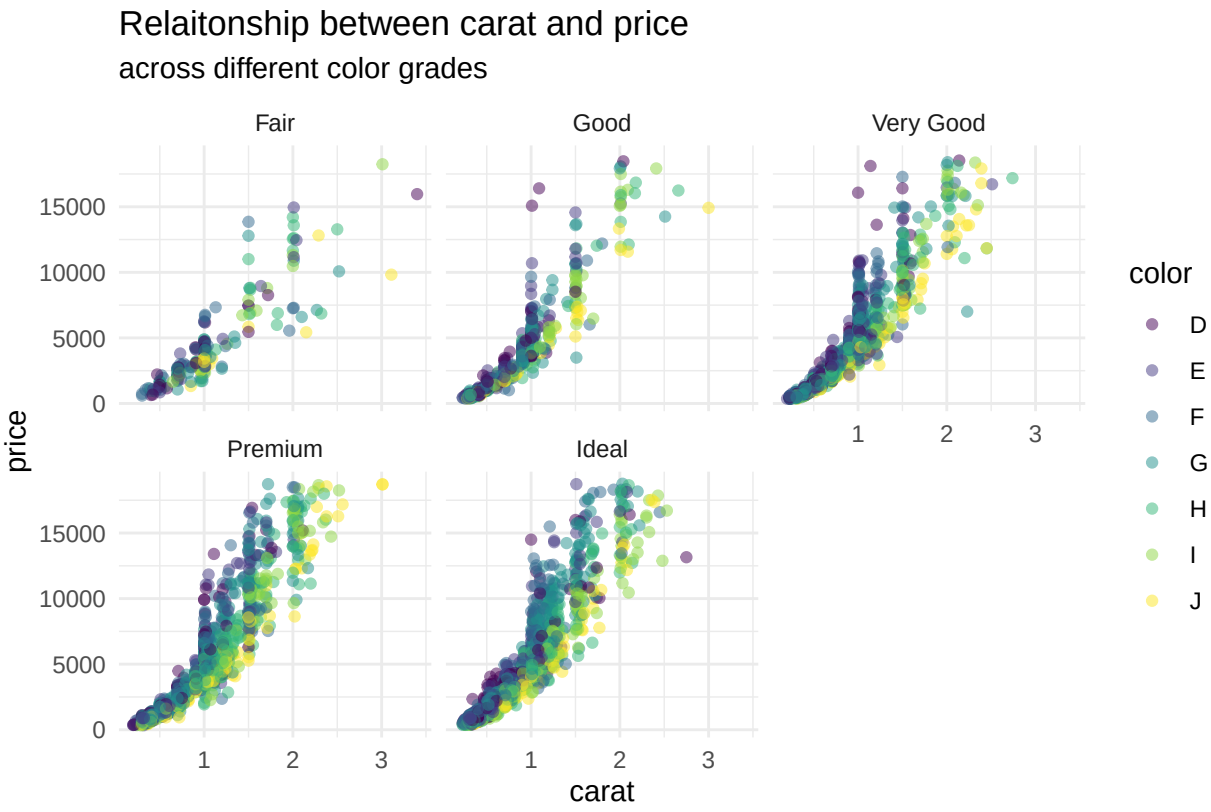
```
)
```



As you can see the relationship between cut and price. A cut quality has a five level is fair, good, very good, premium and ideal as a X-axis, and Y is price. In chart show that higher cut quality it will be higher price

Question 3 How does the relationship between diamond price and carat weight differ across different color grades?

```
ggplot(sample_diamonds,
       aes(carat, price, col = color)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(
    title = "Relaitonship between carat and price",
    subtitle = "across different color grades",
    x = "carat",
    y = "price",
    caption = "Source: Dataset diamonds in Rstudio"
  ) +
  facet_wrap(~ cut)
```




```
## filter level of carat and price by color

sample_diamonds %>%
  select(cut, carat, price, color) %>%
  group_by(color) %>%
  filter(carat > 3, price >= 10000)
```

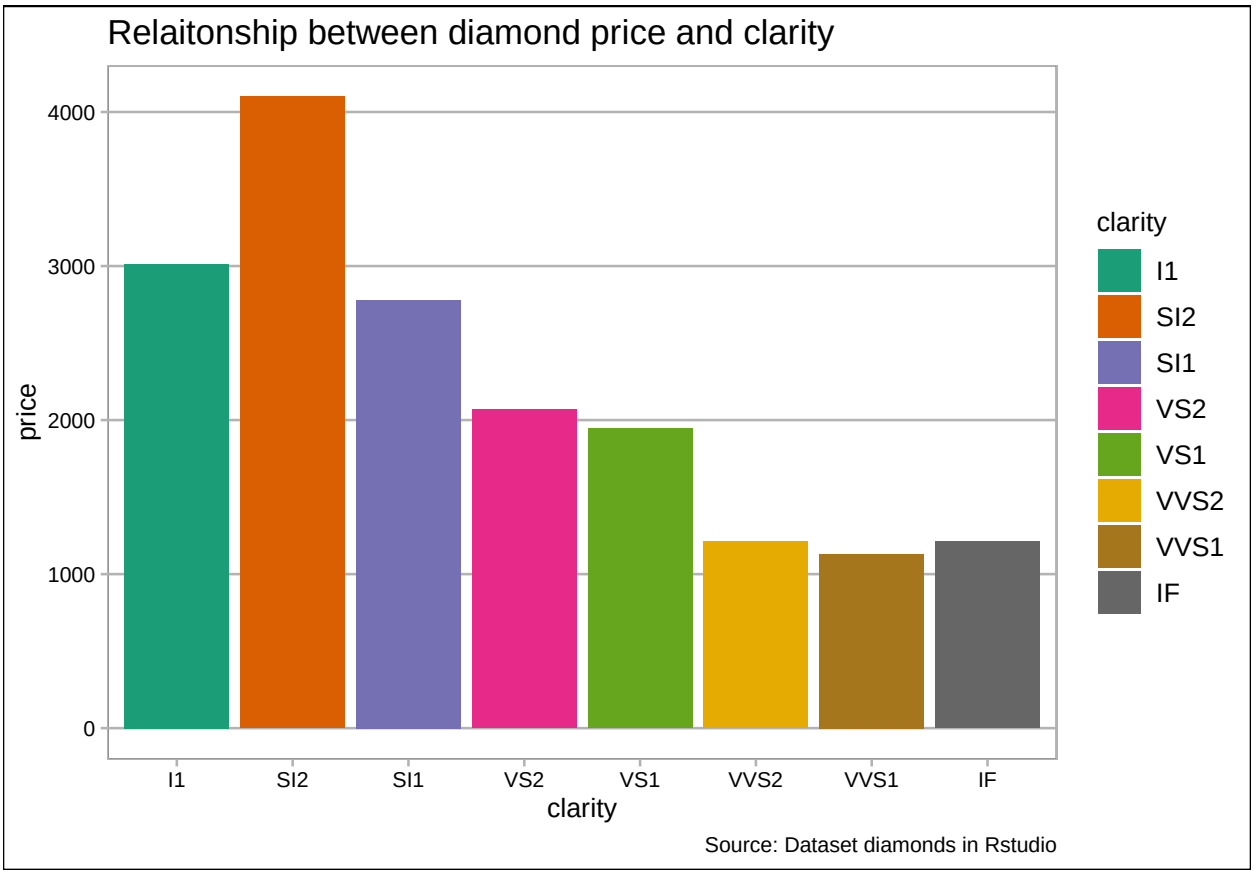
```
## # A tibble: 4 x 4
## # Groups:   color [3]
##   cut     carat price color
##   <ord>   <dbl> <int> <ord>
## 1 Fair     3.01 18242 I
## 2 Premium  3.01 18710 J
## 3 Premium  3.01 18710 J
## 4 Fair     3.4  15964 D
```

In this chart, it show how does the relationship between diamond price and carat weight by color grades. I decide use `facet_wrap(~ cut)` show the chart easier for understanding that the price will be increasing by color in every level of cut. In the summary, Ideal are the highest price in 3.01 carat weight by J color.

Question 4 Can we observe any relationship between diamond price and clarity?

```
agg_price_by_clarity <- sample_diamonds %>%
  group_by(clarity) %>%
  summarise(
    med_price = median(price)
  )
```

```
ggplot(agg_price_by_clarity,
       aes(clarity, med_price, fill = clarity)) +
  geom_col() +
  theme_calc() +
  scale_fill_brewer( type = "qua",
                     palette  = 2) +
  labs(
    title = "Relaitonship between diamond price and clarity",
    x = "clarity",
    y = "price",
    caption = "Source: Dataset diamonds in Rstudio"
  )
```




```
## median diamonds sample price

median(sample_diamonds$price)
```

```
## [1] 2415
```

In this bar chart `geom_col()` will explain the relationship between diamond price and clarity. They have SI2 clarity is the highest price followed by I1 and SI1. Also, the median of sample diamonds is 2,415.

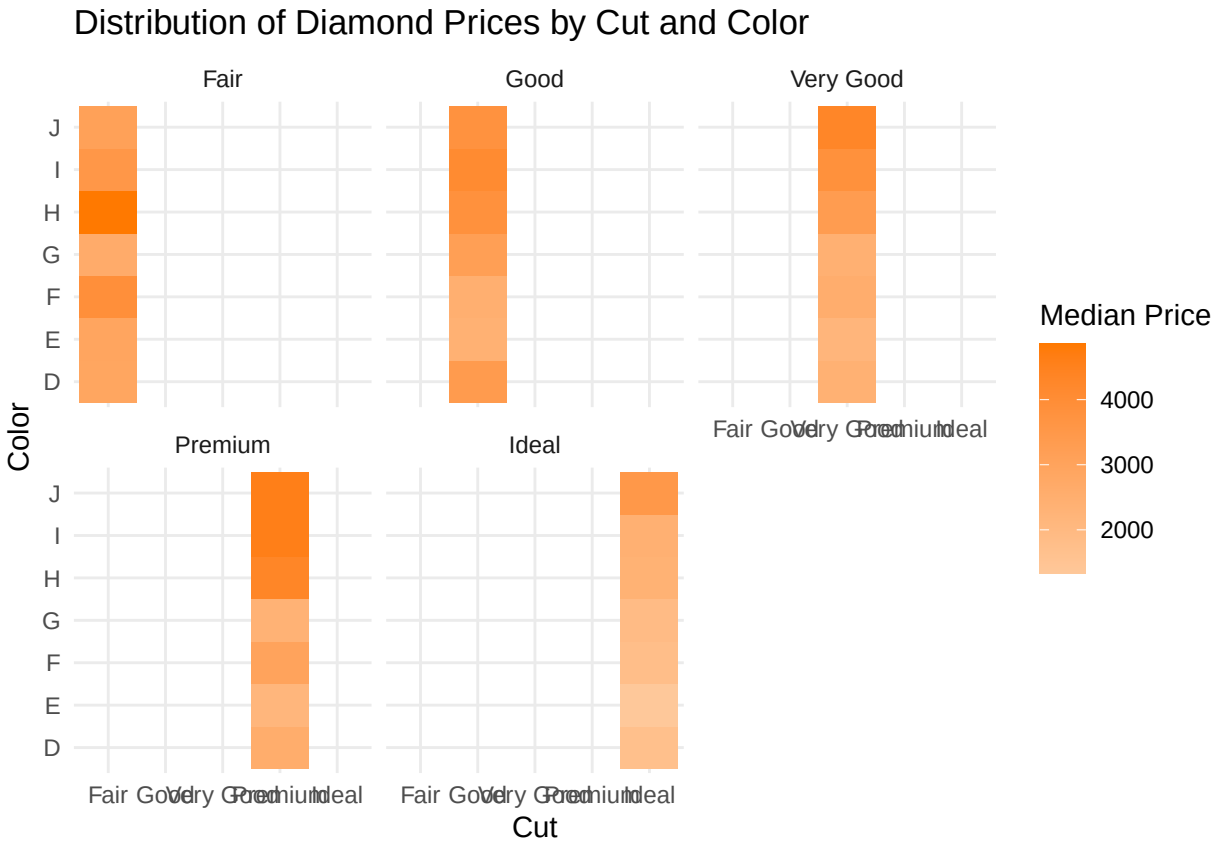
Question 5 What is the distribution of diamond prices based on their cut and color grades?

```
agg_price_by_cut_color <- sample_diamonds %>%
  group_by(cut, color) %>%
  summarise(
```

```
    med_price = median(price)
  )
```

```
## `summarise()` has grouped output by 'cut'. You can override using the `.groups`
## argument.
```

```
ggplot(agg_price_by_cut_color, aes(cut, color, fill = med_price)) +
  geom_tile() +
  scale_fill_gradient(low = "#fec89a", high = "#ff7900") +
  theme_minimal() +
  labs(
    title = "Distribution of Diamond Prices by Cut and Color",
    x = "Cut",
    y = "Color",
    fill = "Median Price"
  ) +
  facet_wrap(~ cut, ncol = 3, nrow = 2)
```



In this chart is distribution of diamonds prices based on their cut and color grades. If you see on the heatmap chart, it show level of cut relationship with color and distribution by average price.

Summary

In this report, I learn a lot of `geom` chart, how to create my data visualization with the dataset by diamonds and use rmarkdown to build web or export file to PDF. Thank you.