Data Visualization Bootcamp Homework

Yossapong Chotchuang

2024-01-01

Direction: Use diamonds dataset from ggplot2 package to create 5 charts.

Understand the dataset and variables

```
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   3.4.4     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(dplyr)
glimpse(diamonds)
```

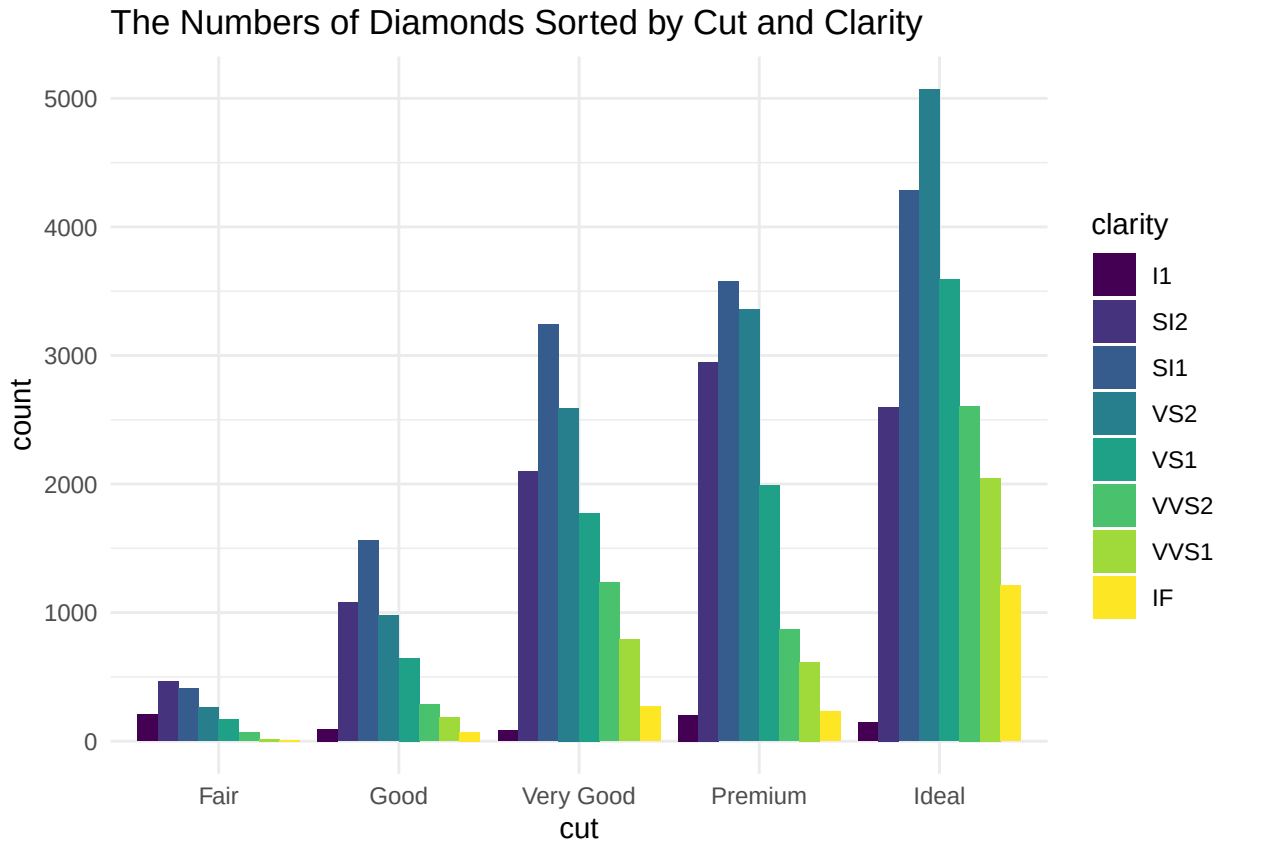
```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord> 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.~
```

This dataset contains 53,940 rows of different diamonds and 10 columns (variables). For clearer understanding, some variables are elaborated as follows: - carat: a unit of weight for diamonds - cut: the quality of how well a diamond is cut represented in 5 levels: “Fair,” “Good,” “Very Good,” “Premium,” “Ideal.” - color: the color of diamond ranging from D to J (best to worst) - clarity: an assessment of how clear a diamond is - x, y, z, depth, and table variables are the values that depict the dimension of the diamond.

Question 1: How many diamonds in each cut type are there? Which cut type has the most count?

```
ggplot(data = diamonds,
       aes(cut, fill = clarity )) +
  geom_bar(position = "dodge") +
```

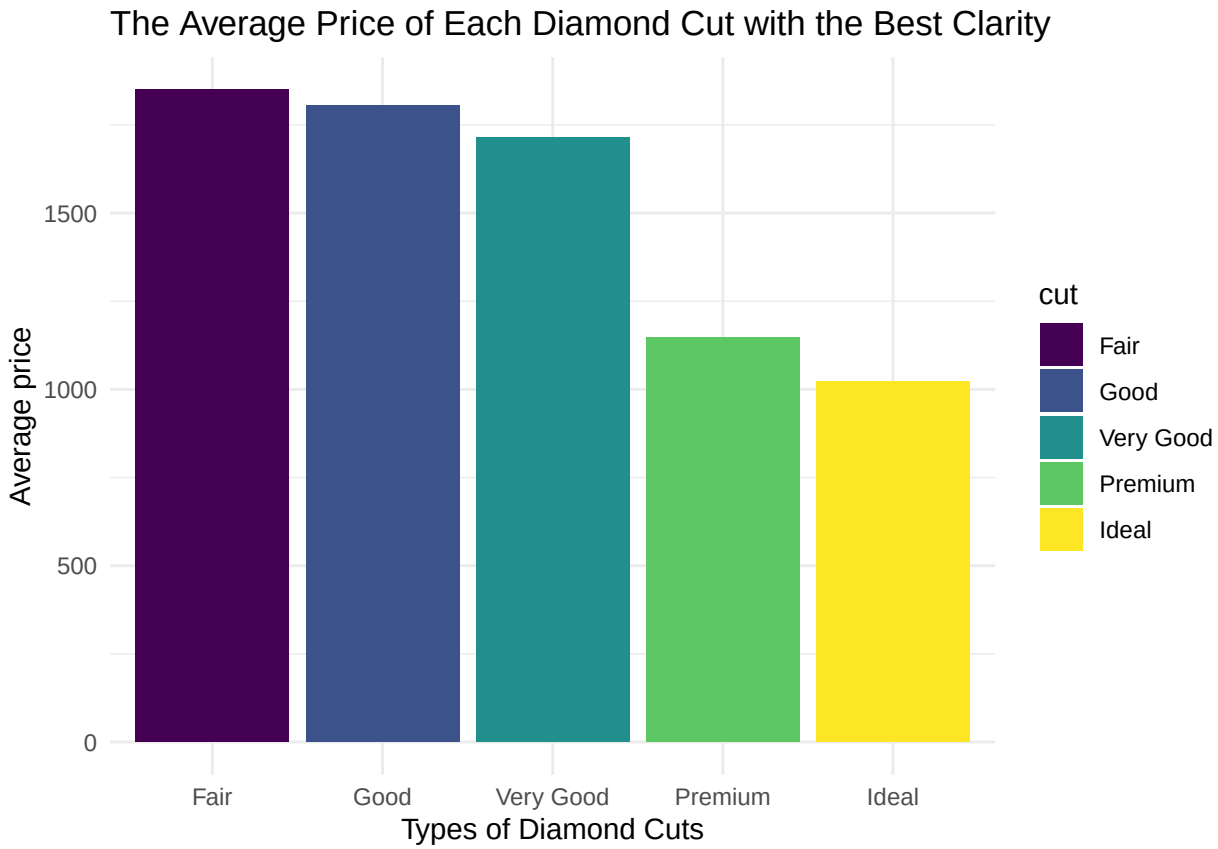
```
  theme_minimal() +
  labs(
    title = "The Numbers of Diamonds Sorted by Cut and Clarity",
    x = "cut",
    y = "count"
  )
```



According to the graph, it can be seen that the number of diamonds with clarity level of SI1 and VS2 is higher compared to the other levels among every cut level. The diamonds with I1 clarity level are lowest in number in almost every cut level while the ones with SI2 and VS1 clarity have close numbers of count. For the diamonds with the clarity level of VVS2, VVS1, and IF, their counts are relatively low across every cut level.

Question 2: What is the average price of the diamonds in each cut type with best clarity level?

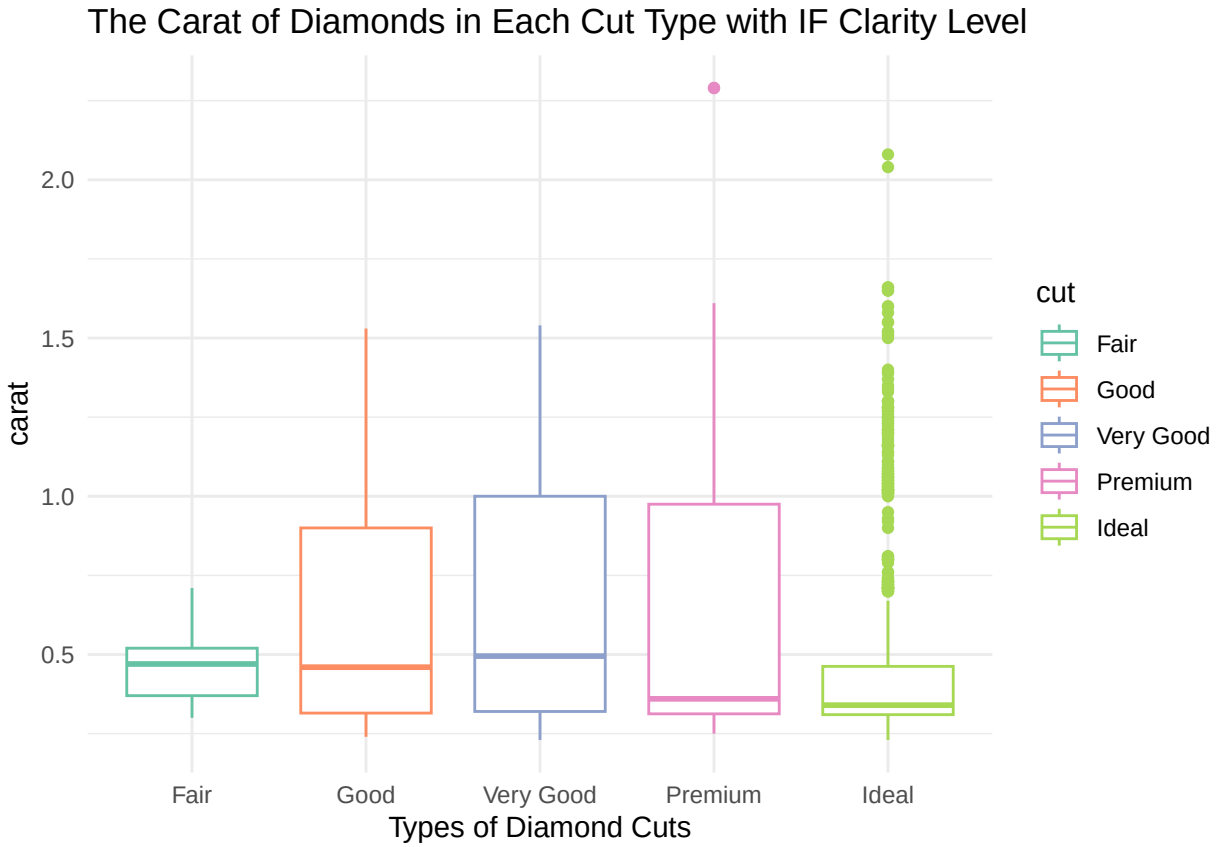
```
diamonds %>%
  filter(clarity == "IF") %>%
  group_by(cut) %>%
  summarise(average_price = median(price)) %>%
  ggplot(aes(cut, average_price, fill = cut)) +
  geom_col() +
  theme_minimal() +
  labs(x = "Types of Diamond Cuts", y = "Average price",
  title = "The Average Price of Each Diamond Cut with the Best Clarity")
```



Based on the graph, the average price of each type of diamond cut from Fair to Ideal ranges from highest to lowest respectively. The result arouses my curiosity of the association between the price and the Ideal level of cut. Is there another factor that may affect the price of the diamonds with such quality. This leads me to the next question.

Question 3: What are the distributions of carat of diamonds in each cut type with IF clarity level?

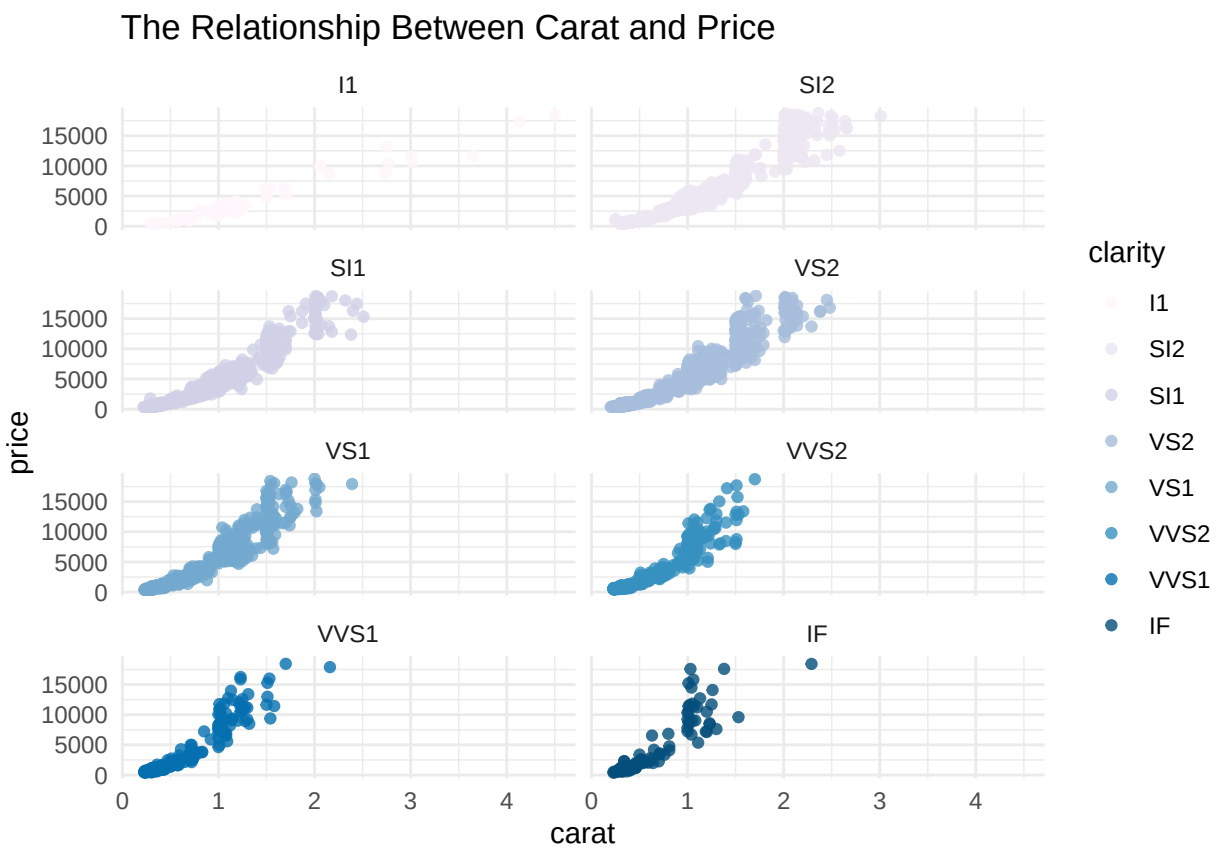
```
diamonds %>%
  filter(clarity == "IF") %>%
  group_by(cut) %>%
  ggplot(aes(x = cut, y = carat, col = cut)) +
  geom_boxplot() +
  theme_minimal() +
  scale_colour_brewer(type = "qual", palette = 7) +
  labs(x = "Types of Diamond Cuts", y = "carat",
       title = "The Carat of Diamonds in Each Cut Type with IF Clarity Level")
```



From the result of the box diagram, the IF-clarity diamonds with ideal cut quality have the smallest size and lowest median in terms of carat. Compared to the diamonds with other types, the data of ideal-cut diamonds also has the lowest dispersion, which is very different to the ones with good, very good, and premium levels but close to the fair diamonds. With the aforementioned explanation, I would like to explore more about the relationship between the carat and price of diamonds.

Question 4: What is the relationship between carat and price of the diamonds? If a diamond is clearer, does it mean it will have more price?

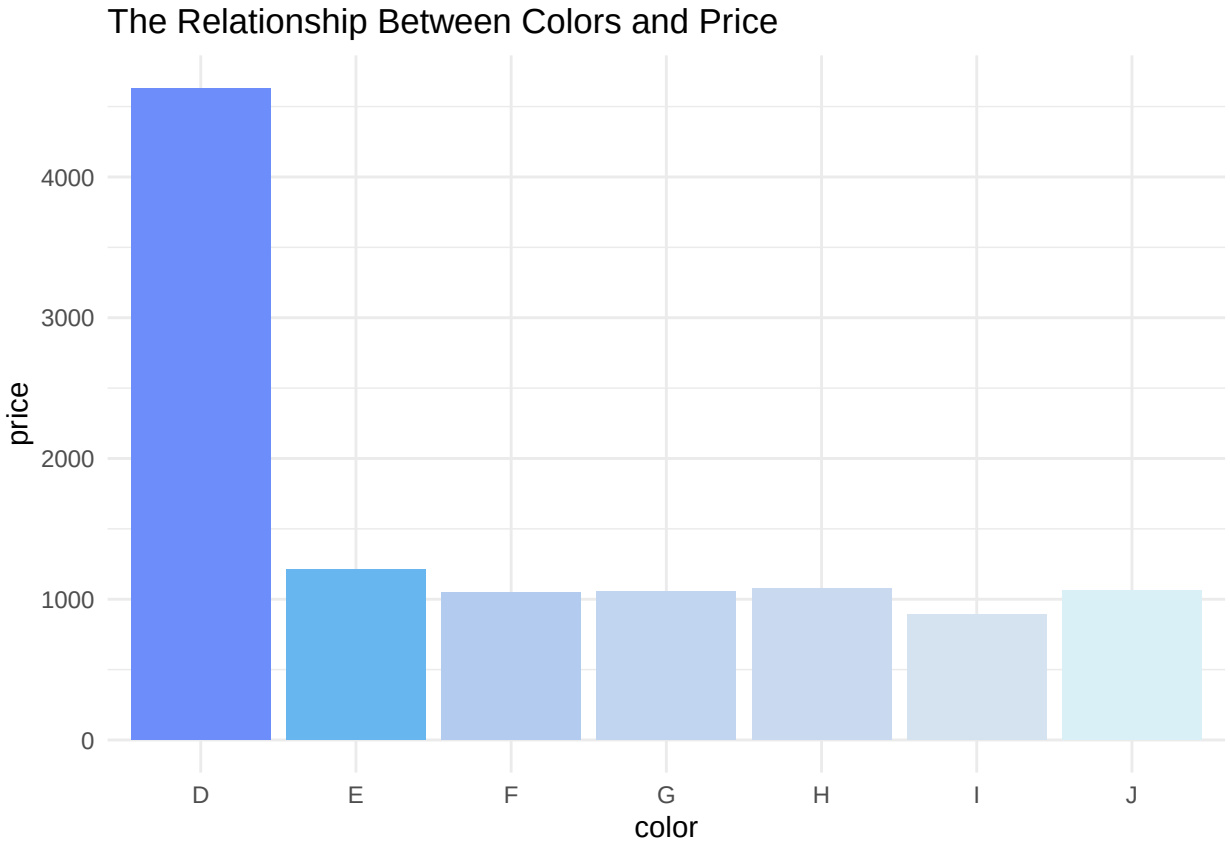
```
set.seed(45)
mini_diamonds <- sample_frac(diamonds, 0.1)
ggplot(mini_diamonds,
       aes(carat, price, col=clarity)) +
  geom_point(alpha=0.8) +
  theme_minimal() +
  scale_color_brewer(type = "seq",
                     palette = 9) +
  facet_wrap(~clarity, ncol=2) +
  labs(
    title = "The Relationship Between Carat and Price",
    x = "carat",
    y = "price"
  )
```



According to the graph, it can be concluded that the higher the carat, the higher the price. The clarity factor is also associated the price of the diamonds. The better of clarity of a diamond is, the higher its price will be. Other than that, it can be clearly noticed that the IF-clarity diamonds with 0.5 carats have the average price similar to the ones in other types but have expensive price when appearing with 1 carat. However, most of diamonds with IF clarity come in a very small size, resulting in inevitably low price.

Question 5: Which colors of Diamonds with IF Clarity Level Has the Highest Average Price?

```
set.seed(45)
mini_diamonds <- sample_frac(diamonds, 0.1)
diamonds %>%
  filter(clarity == "IF") %>%
  group_by(color) %>%
  summarise(average_price = median(price)) %>%
  ggplot(aes(color, average_price, fill = color)) +
  geom_col(fill = c("#6c8dfa", "#68b6ef", "#B4CBF0", "#C1D5F0", "#C8D9F0", "#D5E3F0", "#DAF0F7")) +
  theme_minimal() +
  labs(
    title = "The Relationship Between Colors and Price",
    x = "color",
    y = "price"
  )
```



Clearly demonstrated by the chart, the color of the diamonds with IF clarity level that have the highest price is D color. As for the ones with other color groups, their average prices are close to each other.