HW_Diamonds

Pichaya

2024-07-24

Data Visualization Homework

5 Diamond Charts with Findings

by Pichaya Vetvitavatana

Install Packages and Library

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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 errors
```

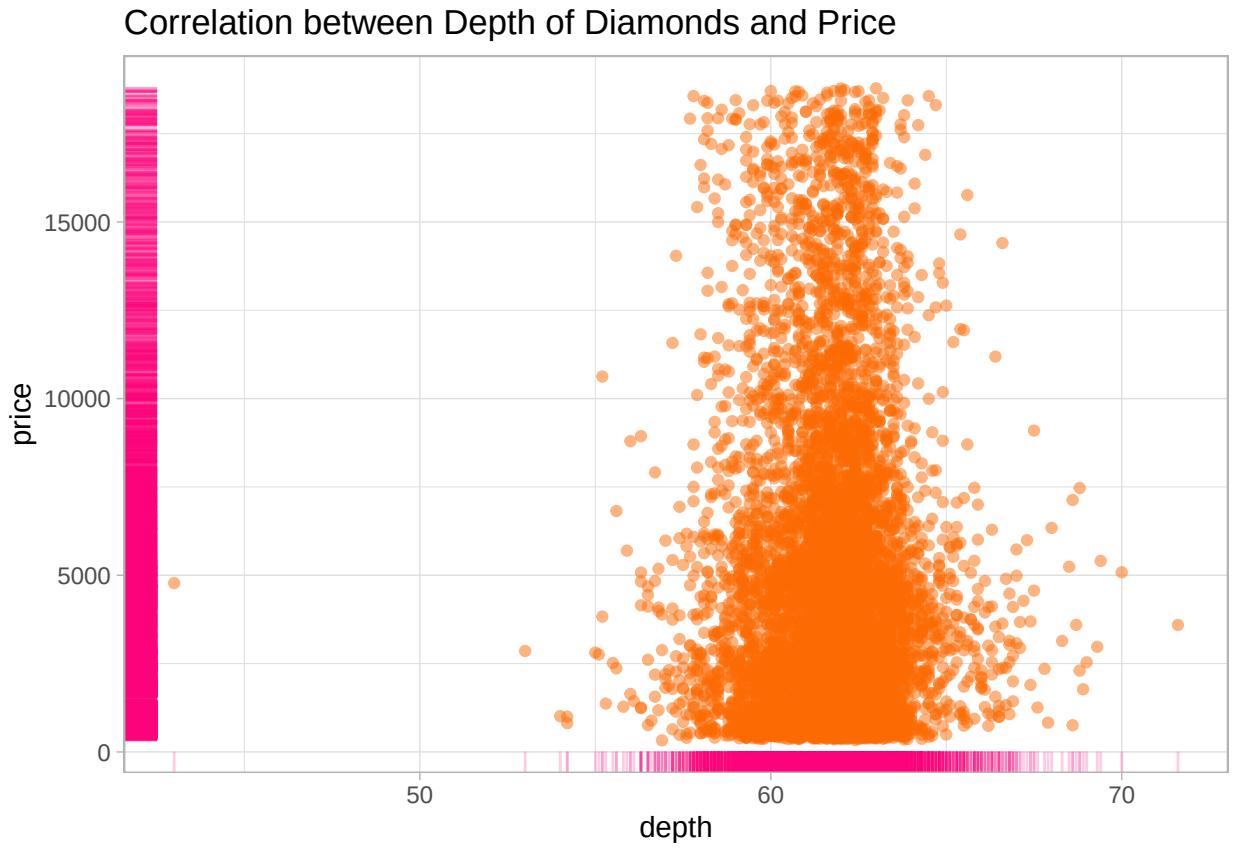
```
library(ggplot2)
library(dplyr)
```

Questions 1

What is the correlation between the dept of the diamonds and the price? Create the chart with random sample of 10,000 data points and find the actual correlation

```
diamond_sample <- sample_n(diamonds,10000)

ggplot(diamond_sample, aes(depth,price)) +
                       geom_point(color = "#fc6b03",
                                  alpha = 0.5) +
                       geom_rug(alpha = 0.2,
                                color = "#fc037b") +
                       theme_light() +
                       labs(title = "Correlation between Depth of Diamonds and Price")
```



Finding

There is no significant correlation between the depth and the price of the diamond.

Questions 2

Generate chart showing the correlation between carat and price and then separate by the top best 3 color

Step 1 Find what colors are there

```
select(diamonds,color) %>%
  distinct()
```

```
## # A tibble: 7 x 1
##   color
##   <ord>
## 1 E
## 2 I
## 3 J
## 4 H
## 5 F
## 6 G
## 7 D
```

Step 2 Find how many diamonds per color are there

```
select(diamonds,color) %>%
  count(color)
```

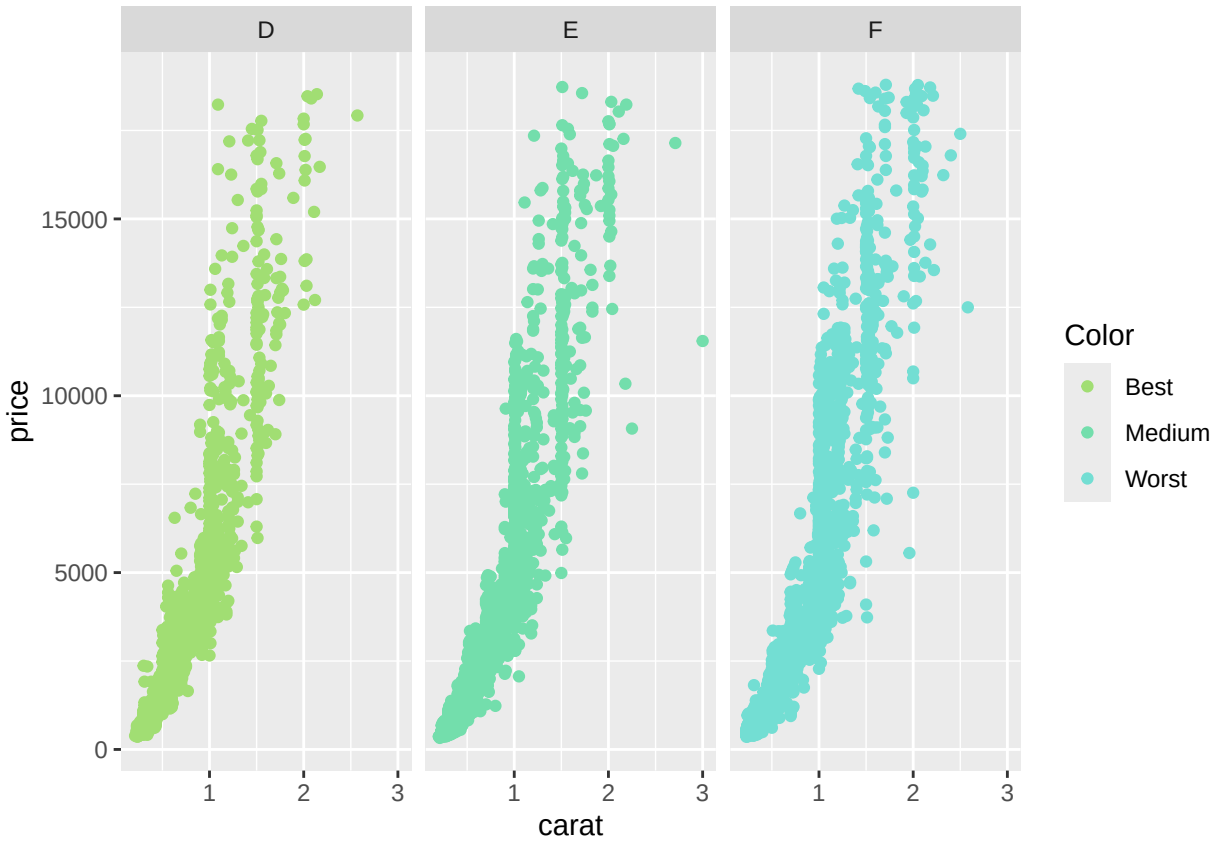
```
## # A tibble: 7 x 2
##   color     n
##   <ord> <int>
## 1 D      6775
## 2 E      9797
## 3 F      9542
## 4 G     11292
## 5 H      8304
## 6 I      5422
## 7 J      2808
```

Step 3 Prepare the data by filtering the top 3 colors out

```
top_3_col <- diamonds %>%
              filter(color == c("D","E","F"))
```

Step 4 Create the chart

```
ggplot(top_3_col, aes(carat,price,color = color)) +
  geom_point() +
  facet_grid(~color) +
  scale_color_manual(values = c("#a1de73","#73deac","#73ded3"),
                     name = "Color",
                     labels = c("Best","Medium","Worst"))
```



Finding

The correlation bewteen the price and the carat in different color range is very similar

Question 3

Compare the correlation between price and carat in 2 scenarios: 1) when the carat is less or equal to 3 2) when the carat is above 3.

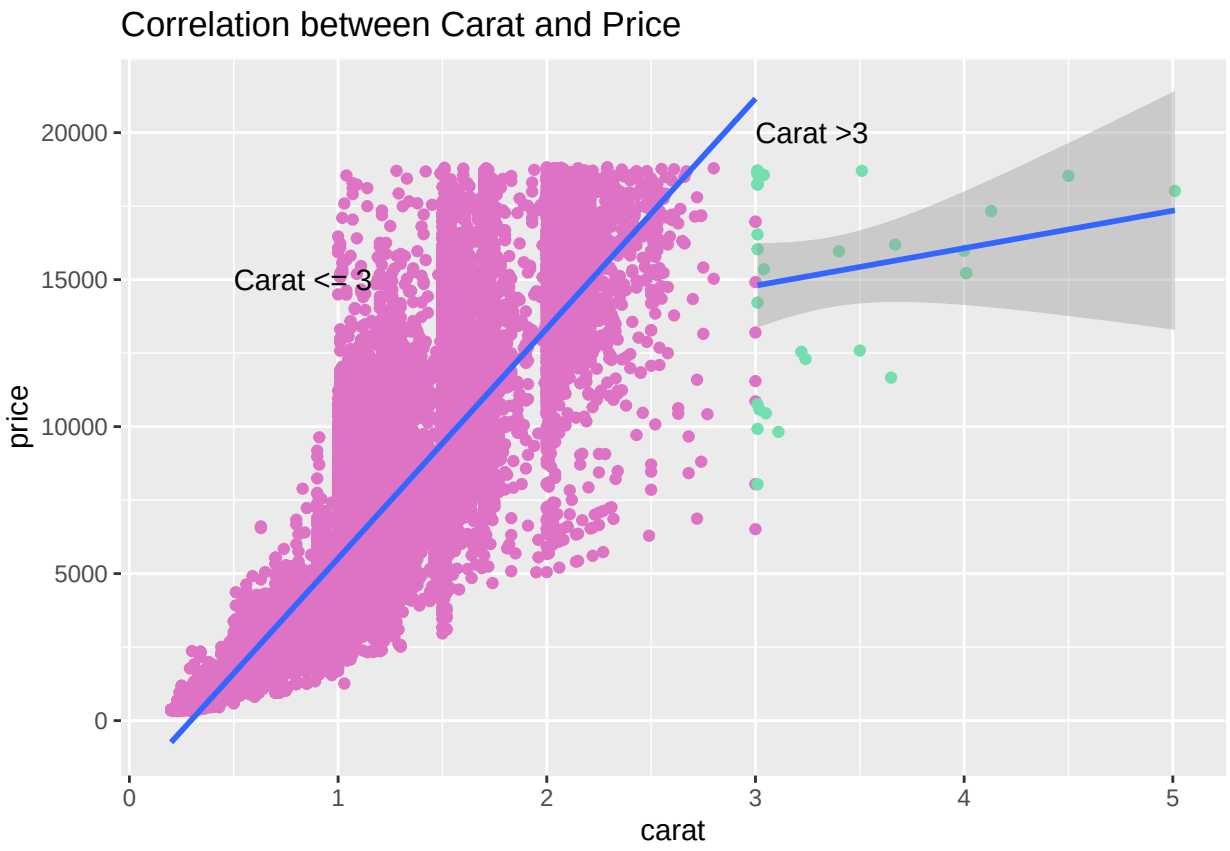
```
car_less3 <- diamonds %>%
                filter(carat <=3)

car_more3 <- diamonds %>%
                filter(carat > 3)


ggplot() +
  geom_point(data = car_less3,
             mapping = aes(carat,price),
             color = "#de73c5") +
  geom_smooth(data = car_less3,
              mapping = aes(carat,price),
              method = lm) +
  geom_point(data = car_more3,
             mapping = aes(carat,price),
```

```
              color = "#73deae")+
  geom_smooth(data=car_more3,
              mapping = aes(carat, price),
              method = lm)+
  labs(title = "Correlation between Carat and Price") +
  geom_text(aes(x = c(0.5,3), y = c(15000,20000)), # the axis point to place each label; "Carat <=3" at
            label = c("Carat <= 3", "Carat >3"), # the name of the label
            hjust = 0) # hjust = 0 aligns the text to the left at the specified x-coordinates.
```

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



Finding

The correlation between the carat and the price is higher for diamonds with carate less than 3

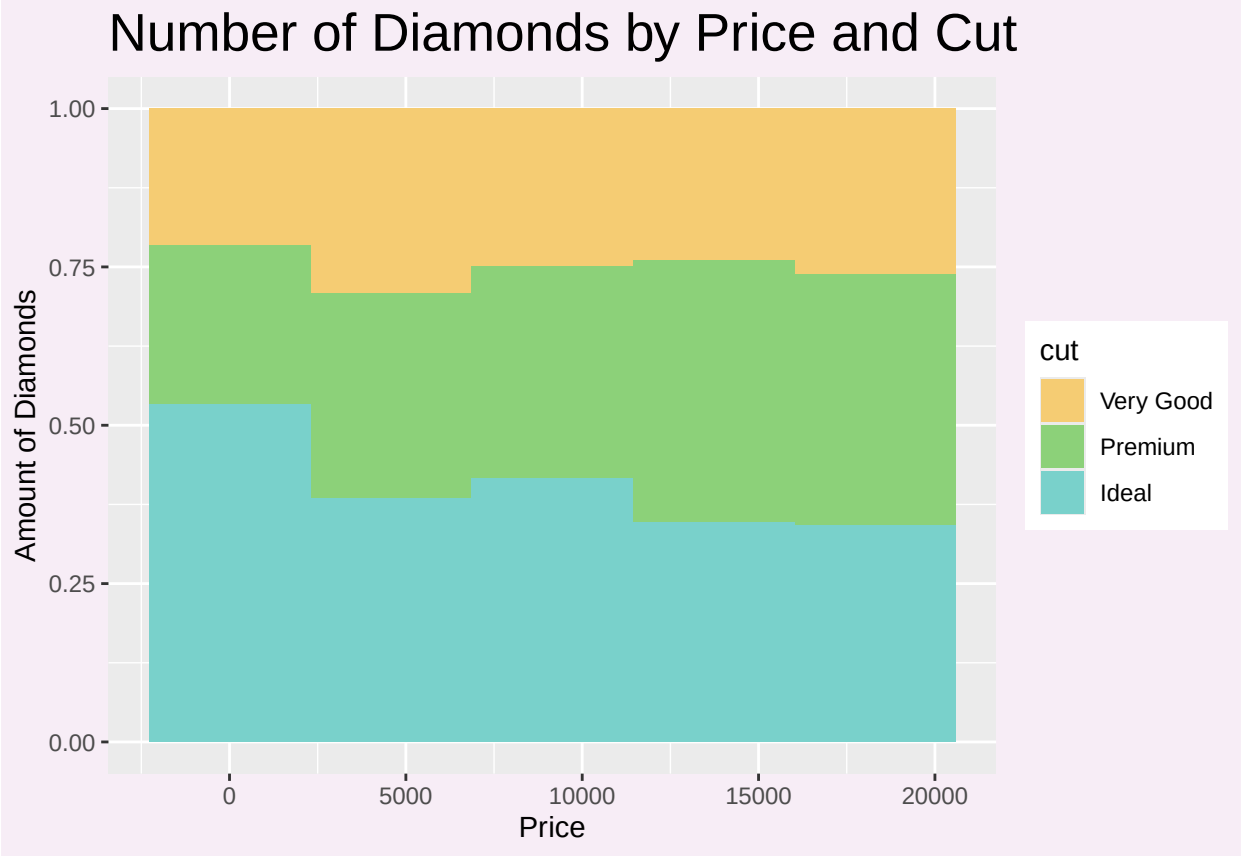
Question 4

How many diamonds in the database which have top 3 cut are there in terms of percentage?

```
price_cut_dia <- diamonds %>%
              filter(between(price,500,20000) & cut == c("Very Good","Premium","Ideal"))

ggplot(price_cut_dia, aes(price,fill=cut)) +
```

```
  geom_histogram(bins = 5,
                 position = "fill")+
  labs(title = "Number of Diamonds by Price and Cut",
       x = "Price",
       y = "Amount of Diamonds")+
  theme(plot.title = element_text(size = 20),
        plot.background = element_rect(fill = "#f7edf6")) +
  scale_fill_manual(values = c("#f5cc73","#8cd179","#79d1cb"))
```



Finding

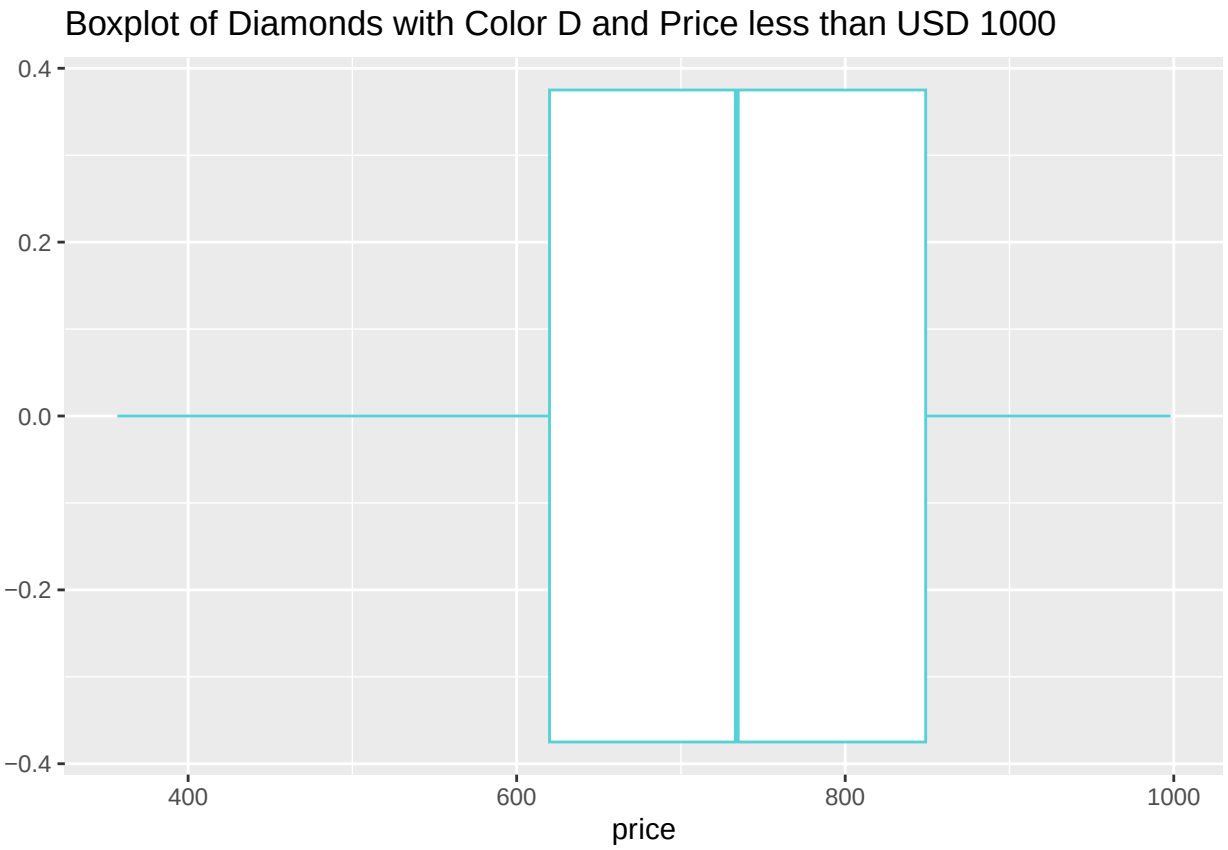
There is a higher share of diamonds with premium cut when the price is higher

Question 5

Find the max,min and average price of the diamonds which is cheaper than USD 1K and with color D are there? Create a box plot to visualize this data.

```
D_less1000 <- diamonds %>%
                filter(price<1000 & color == "D")

ggplot(D_less1000, aes(price)) +
  geom_boxplot(color = "#55d2d9") +
  labs(title = "Boxplot of Diamonds with Color D and Price less than USD 1000")
```



Finding

The lowest price is approx. USD 200, the highest approx. USD 1000, averagely at USD 740.