Batch06_Data_Visualization_HomeWork

Thanes Chaiyakul

2022-11-06

Library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)
```

Data Set

- Diamonds
  - dimension: 53,940rows x 10cols
  - price: $326-$18,823
  - carat: 0.2-5.01
  - cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal)
  - color: D (best) to J (worst)
  - clarity: l1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF(best)
  - x: 0-10.74mm
  - y: 0-58.9mm
  - z: 0-31.8mm
  - depth: z/mean(x,y)
  - table: 43-95

```
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.~
```

Check NA

There is no NA value in this data set.

```
apply(diamonds, MARGIN=2, function(col) sum(is.na(col)))
```

```
##   carat     cut   color clarity   depth   table   price       x       y       z
##       0       0       0       0       0       0       0       0       0       0
```

Histogram

Figure1 shows histogram of diamonds' price/carat with different attributes. It can be seen that the high and low numbers of samples for each attributes are listed in the below table:

Attribute	Low	High
Cut	Fair	Ideal
Color	J	G
Clarity	I1	SI1,VS2

Table1: table of sample level for each attribute

```
diamonds %>%
  mutate(pricepercarat = price/carat) %>%
  ggplot(aes(pricepercarat,fill=cut)) +
  geom_histogram(bins=15) +
  xlim(0,10000)+
  theme_minimal()+
  theme(legend.position = "bottom"
        , axis.text.x=element_text(size=rel(.6))
        , axis.text.y=element_text(size=rel(.8)))+
  scale_fill_brewer(type="qual",palette = "Set1")+
  facet_grid(clarity~color)
```

```
## Warning: Removed 617 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 552 rows containing missing values (geom_bar).
```

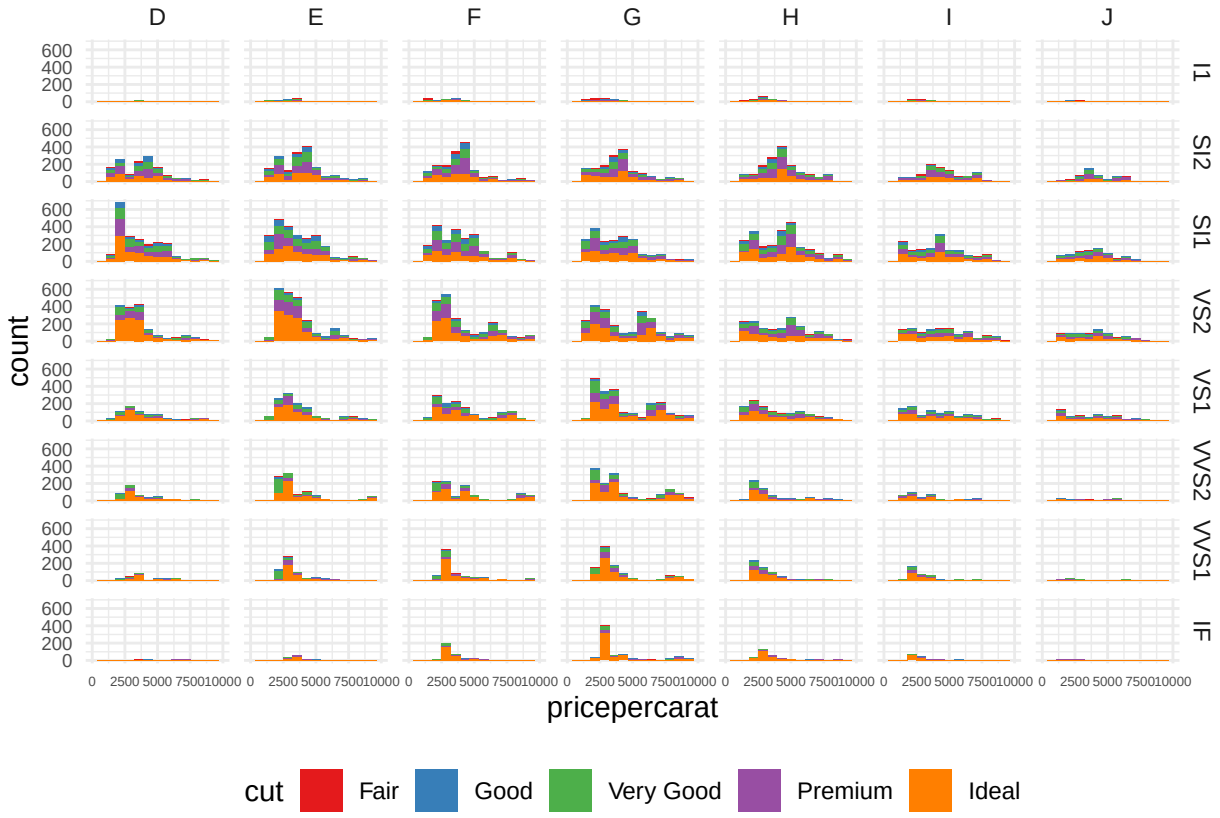

Figure1: Histogram of diamonds' price/carat with different attributes

Mean Price/Carat

In Figure2, means of price/carat for each attribute have been calculated and plotted versus cut for each clarity and color. It shows that diamonds with IF clarity and D color have the highest mean of price/carat no matter the quality of the cut. While diamonds with I1 clarity would have the lowest mean of price/carat among the group.

```
diamonds %>%
  filter(carat<=2 & carat>1) %>%
  mutate(pricepercarat = price/carat) %>%
  group_by(color,cut,clarity) %>%
  summarise(n = n()
            , mean_pricepercarat = mean(pricepercarat)) %>%
  ggplot(aes(cut, mean_pricepercarat, group=color,clarity, color = color)) +
  geom_point()+
  geom_line()+
  theme_minimal() +
  theme(legend.position = "bottom"
        , axis.text.x=element_text(size=rel(.6))
        )+
  scale_color_brewer(type="qual",palette="Set2")+
  facet_wrap(~clarity)
```

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

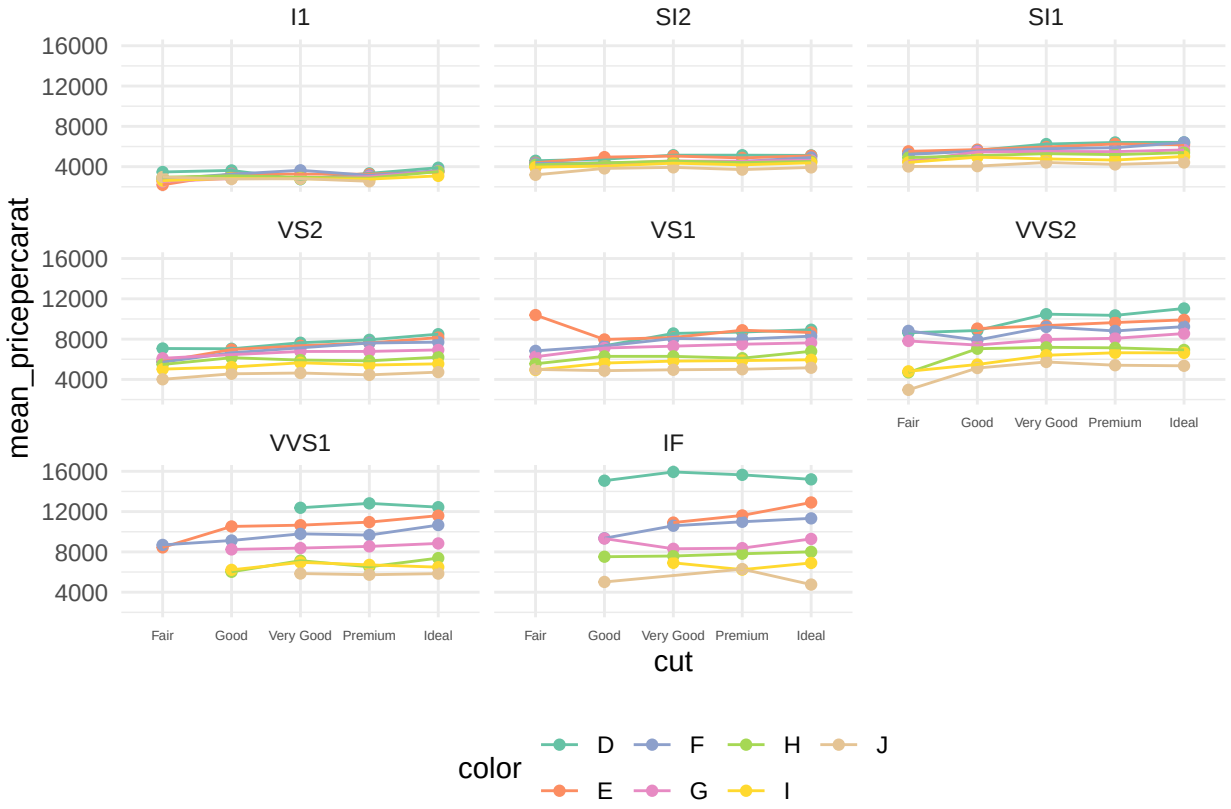

Figure2: Mean Price/Carat for each attribute

Sample Plot Data

Figure3 shows sample plots of carat vs price and trend line group by cut, clarity and color. It can be seen that trend line of cut and clarity are as expected, because the higher quality of cut or clarity the steeper the slope between price and carat. However, in the color-trend-line graph, color F and G have steeper slope than color D and E that have better color quality.

```
set.seed(42)
df<-sample_n(diamonds, size=5000)
p1<-df %>%
  ggplot(aes(carat, price, color=cut)) +
  geom_point(alpha=0.3) +
  geom_smooth(method="lm", se=F)+
  theme_minimal()+
  scale_color_brewer(type="qual",palette = "Set2")


p2<-df %>%
  ggplot(aes(carat, price, color=clarity)) +
  geom_point(alpha=0.3) +
  geom_smooth(method="lm", se=F)+
  theme_minimal()+
  theme(
    legend.key.size = unit(5, "mm")
    #,legend.spacing = unit(1,'mm')
```

```
    #,legend.spacing.y = unit(1,'mm')
    #,legend.title = element_text(size=6)
    #,legend.text = element_text(size=4)
        )+
  scale_color_brewer(type="qual",palette = "Set2")


p3<-df %>%
  ggplot(aes(carat, price, color=color)) +
  geom_point(alpha=0.3) +
  geom_smooth(method="lm", se=F)+
  theme_minimal()+
  scale_color_brewer(type="qual",palette = "Set2")

(p1+p2)/p3
```

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

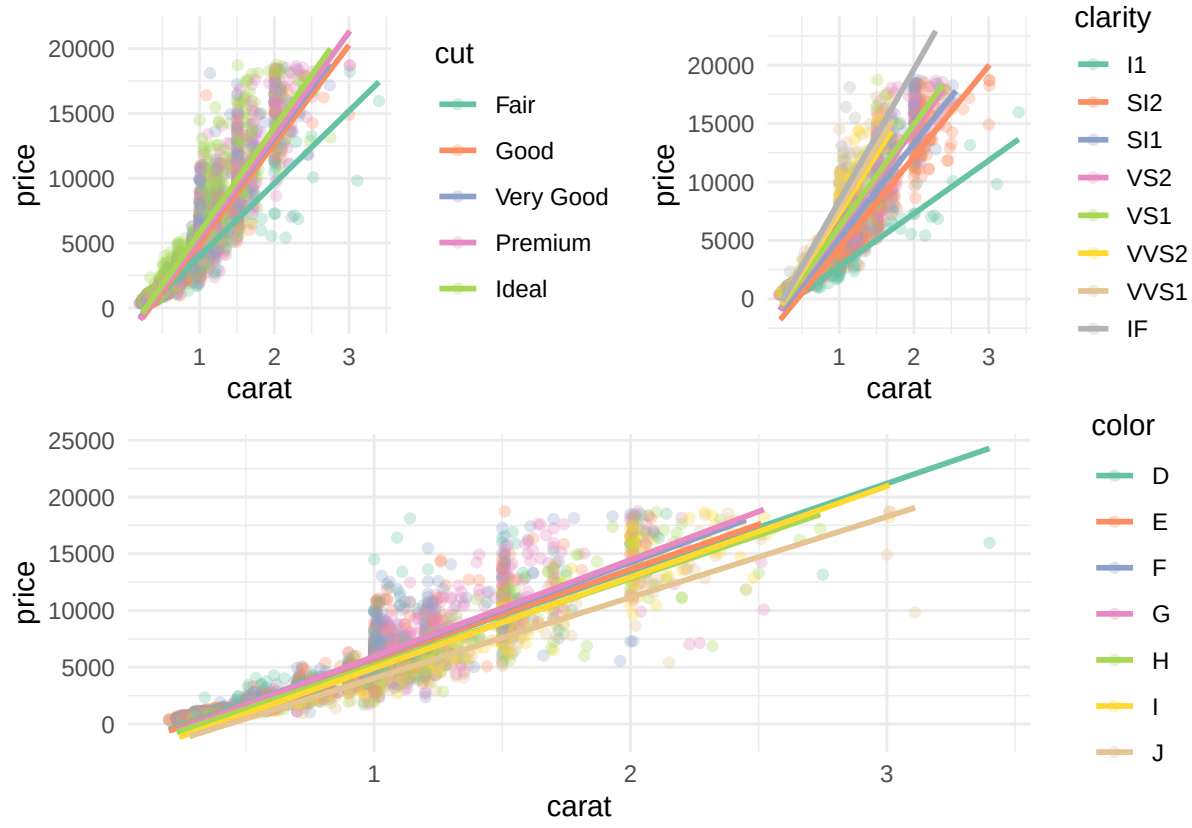

Figure3: Sample plotted of diamonds with trend line for each attribute

Color and Price

In previous section, it can bee seen that diamonds with color F and G have steeper trend line of price per carat than higher quality color counterpart D and E. Below diamonds' weight have been splitted into 3 group to see relation between pricepercarat and color. The top left figure show relation between color and carat to

see how carat is distributed between color. Most of diamond weight is below 2 carat. In the top right of Figure4 with carat<1, diamonds with color F and G have median of pricepercarat closed to color D and E. While diamonds with weight between 1 and 2 carat and color F and G have slightly higher median value than diamonds with color D and E. Finally, the last group of diamonds with weight over 2 carat the median of pricepercarat of color D and E are higher than other color groups.

```
df <- diamonds %>%
      mutate(PriceperCarat=price/carat)

p1<-diamonds %>%
    ggplot(aes(color,carat)) +
    geom_boxplot() +
    theme_minimal()+
    labs(
      subtitle="All Carat"
    )
p2<-df %>%
  filter(carat<=1) %>%
  ggplot(aes(color,PriceperCarat)) +
  geom_boxplot()+
  theme_minimal()+
  labs(
    subtitle="Filtered: Carat <=1"
  )

p3<-df %>%
  filter(carat>1 & carat<=2) %>%
  ggplot(aes(color,PriceperCarat)) +
  geom_boxplot()+
  theme_minimal()+
  labs(
    subtitle="Filtered: 1< Carat <=2"
  )

p4<-df %>%
  filter(carat>2) %>%
  ggplot(aes(color,PriceperCarat)) +
  geom_boxplot()+
  theme_minimal()+
  labs(
    subtitle="Filtered: Carat > 2"
  )

(p1+p2)/(p3+p4)
```

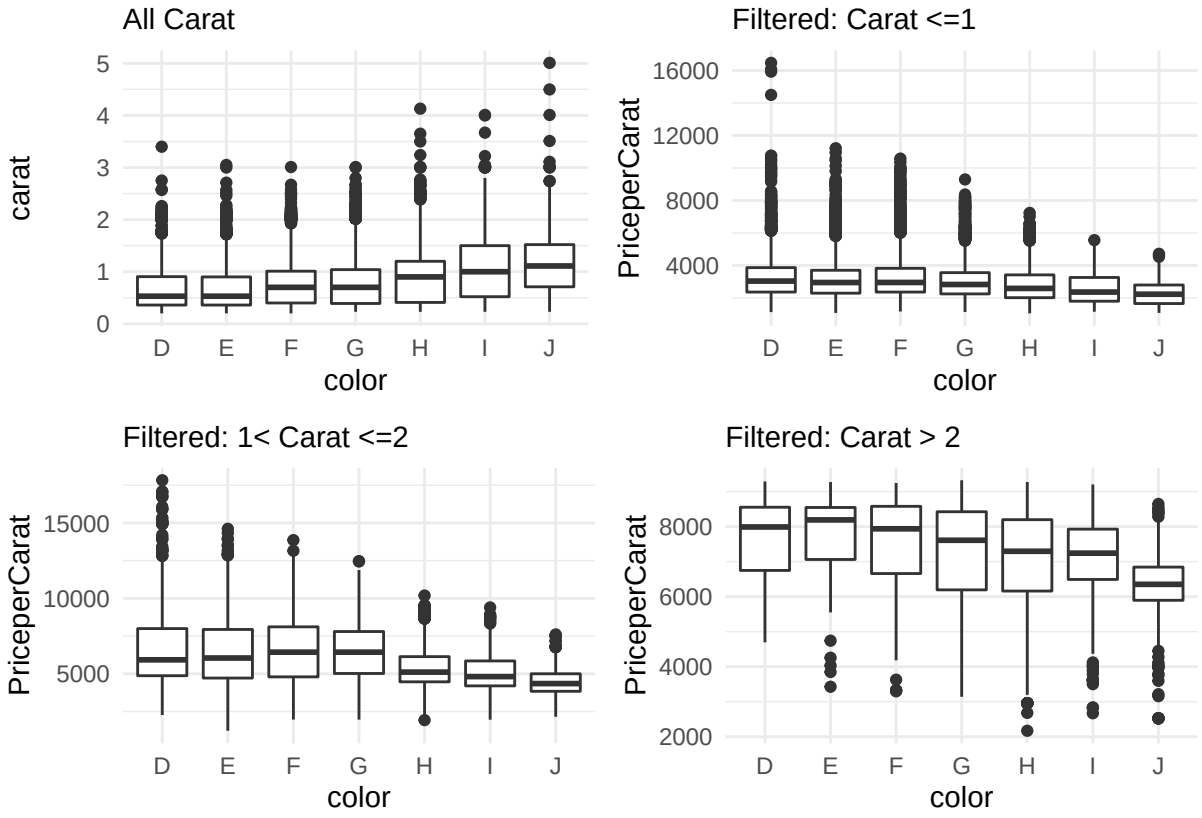

Figure4: boxplot of color and carat, color and price/carat with various range of carat

Color and Price/Carat for 1< carat <=2

In Figure5, a plot of color and price/carat for 1< carat <=2 has been further investigated by adding mean as red dot and error bar with min as mean - sd and max as mean + sd. It can be seen that diamonds with color F and G has mean value closed to median than color D and E, because color F and G has less outliers.

```
##plot relation between color and pricepercarat for carat between 1 and 2
##1<=carat<=2 17,171rows
diamonds %>%
  filter(
    carat<=2 &
      carat>1) %>%
  mutate(pricepercarat=price/carat) %>%
  ggplot(aes(color, pricepercarat)) +
  geom_boxplot()+
  stat_summary(color = "red")+
  stat_summary(
    fun.min = function(x) mean(x) - sd(x)
    ,fun.max = function(x) mean(x) + sd(x)
    ,geom =  "errorbar"
    ,color = "red"
    ,width = .3
  )+
  theme_minimal()+
  labs(
```

```
    title="Plot between Color and Price-per-Carat for 1< Carat <=2"
  )
```

```
## No summary function supplied, defaulting to `mean_se()`
```

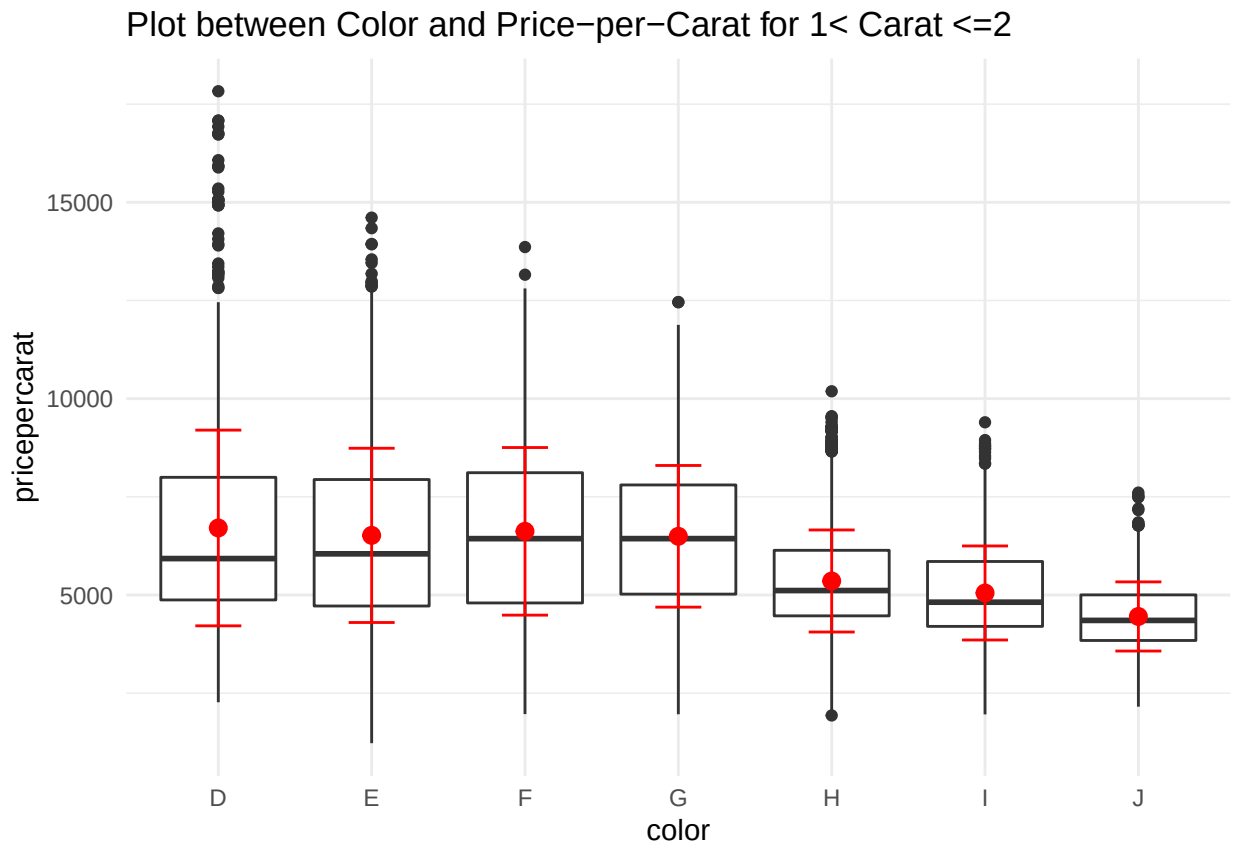

Figure5: Boxplot shows price-per-carat, red dot shows mean, errorbar shows range of mean - sd and mean + sd

Conclusion

There are many factors that affect the price of diamonds such as weight (carat) and quality (color, clarity, cut). As expected, diamonds with higher carat or quality tend to have higher price. If a diamond have exceptional qualities such as clarity and color, it's price/carat will be very high compared to others. There might be some case that diamonds with not the highest quality have higher price/carat, but that occurs at certain range of carat.