Diamonds data visualization Homework

Thunpischa Yodkaew

Homework : Explore diamonds data by dplyr and create at least 5 charts. Use RMarkdown and knit to pdf.

Load library

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

Explore diamond dataset

```
glimpse(diamonds)
```

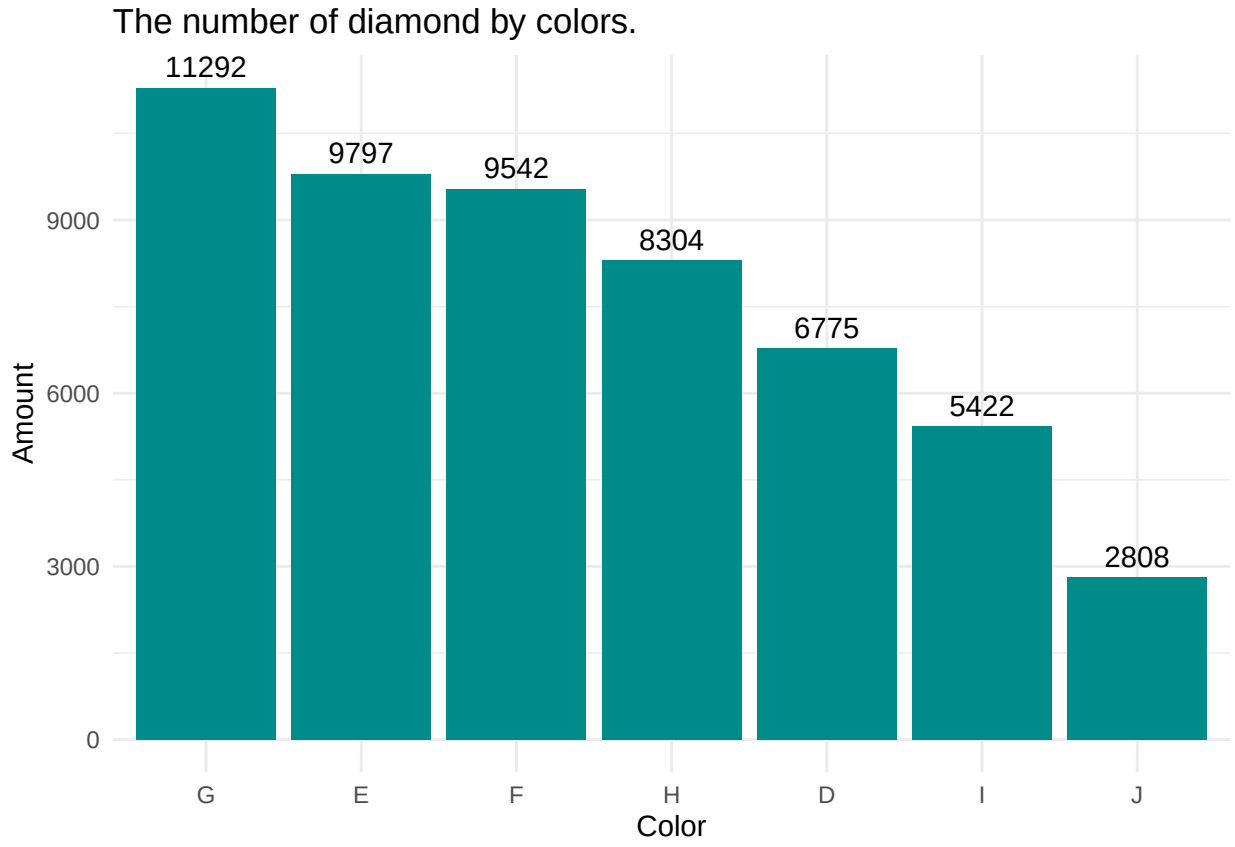
```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord>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

1. How many data in each colour ?

```
n_color <- diamonds %>%
  count(color)
n_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
```

```
ggplot(n_color, aes(x = reorder(color, -n), y = n)) +
  geom_col(fill = 'darkcyan') +
  geom_text(aes(label = n), vjust = -0.5) +
  labs(title = 'The number of diamond by colors.',
       x = 'Color',
       y = 'Amount') +
  theme_minimal()
```

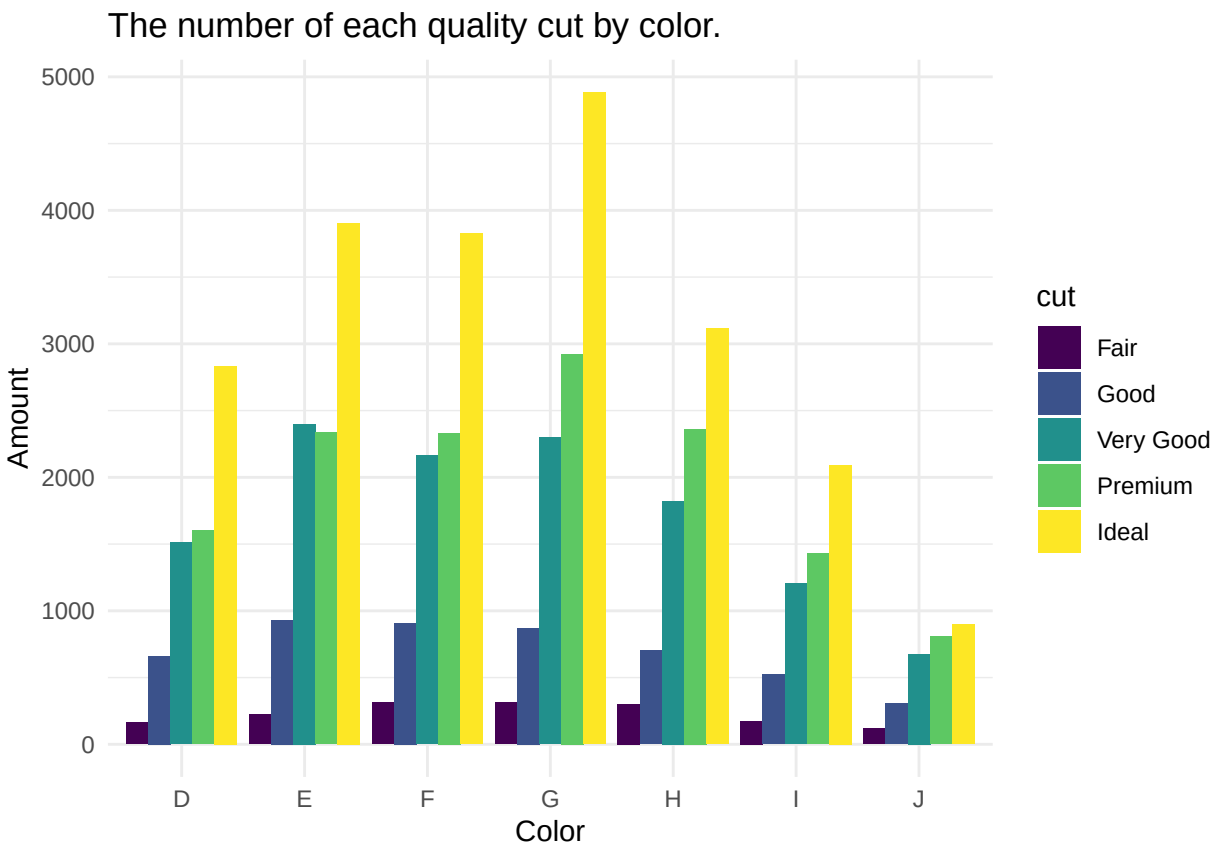


The most three diamond color are 'G', 'E', and 'F', respectively. 'G' is in near colorless group while E and F are in colorless group.

info: https://www.gemsociety.org/article/diamond-color/

2. Amout of Quality cut of each color

```
ggplot(diamonds, aes(x = color, fill = cut)) +
  geom_bar(position = 'dodge') +
  labs(title = 'The number of each quality cut by color.',
       x = 'Color',
       y = 'Amount') +
  theme_minimal()
```



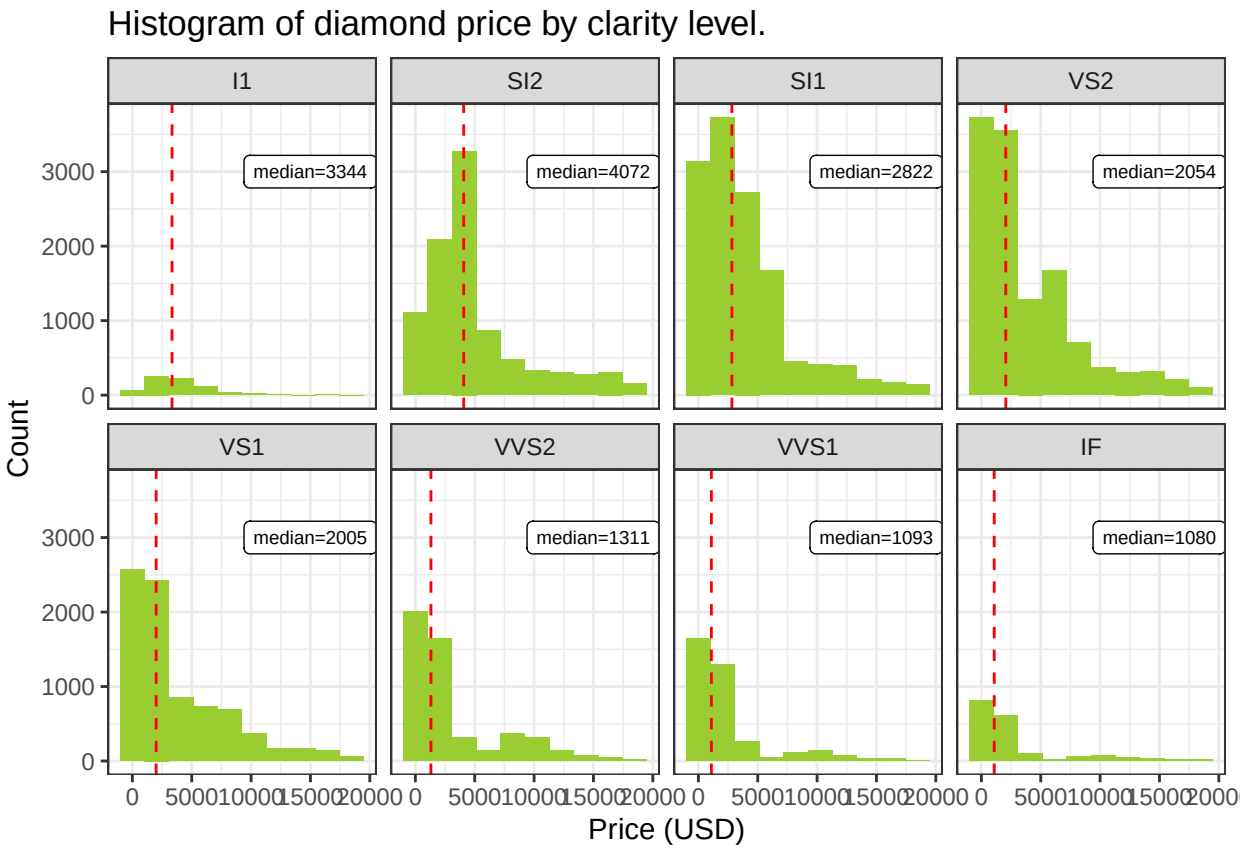
From the chart, all colors distribute similarly. The highest amount of cut quality of all colors is `Ideal` and the lowest is `Fair` cut.

3. Histogram of price for each cut type.

Find median of price for each cut type.

```
median_clarity <- diamonds %>%
  group_by(clarity) %>%
  summarise(median = median(price),
            sd = sd(price))
```

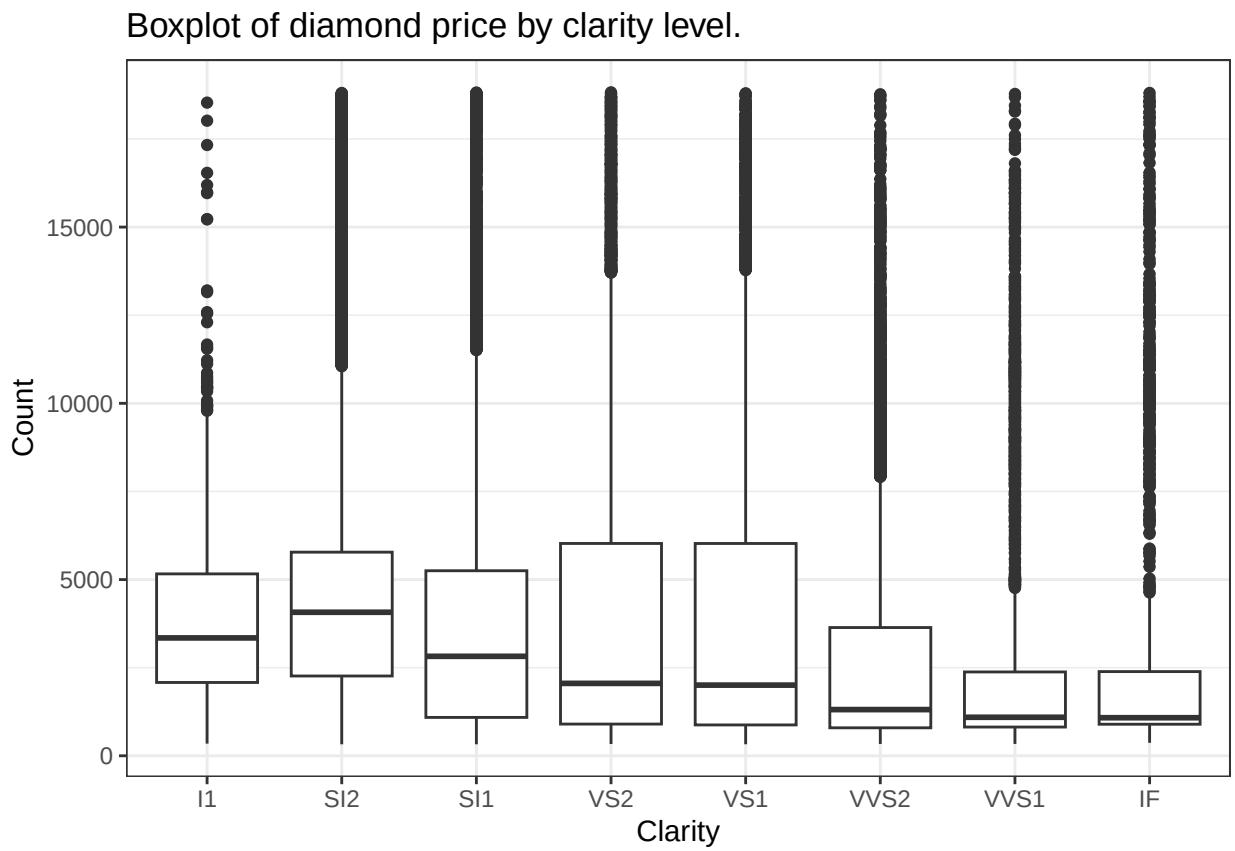
```
ggplot(diamonds, aes(x = price)) +
  geom_histogram(fill = "yellowgreen", bins = 10) +
  geom_vline(data = median_clarity,
             aes(xintercept = median),
             color = "red", linetype = "dashed") +
  geom_label(data = median_clarity,
             aes(x = 15000, y = 3000, label = paste0("median=",median)),
             size = 2.5) +
  facet_wrap(~ clarity, ncol = 4) +
  labs(title = 'Histogram of diamond price by clarity level.',
      x = 'Price (USD)',
      y = 'Count') +
  theme_bw()
```



4. Boxplot of price of each clarity level.

From 3, Distribution of price for each clarity is *right-skewed*, so we will consider **median** price to compare.

```
ggplot(diamonds, aes(x = clarity, y = price)) +
  geom_boxplot() +
  labs(title = 'Boxplot of diamond price by clarity level.',
       x = 'Clarity',
       y = 'Count') +
  theme_bw()
```



`SI2` level has the most expensive price which is 4040 USD, while `IF` , the best clarity level, has 1073 USD as a median price which is the lowest median price among all clarity levels. However, `IF` group has a smallest median price, its s.d. is huge and there are a lot of outliers.

5. Relationship between carat and price diamonds in each clarity level

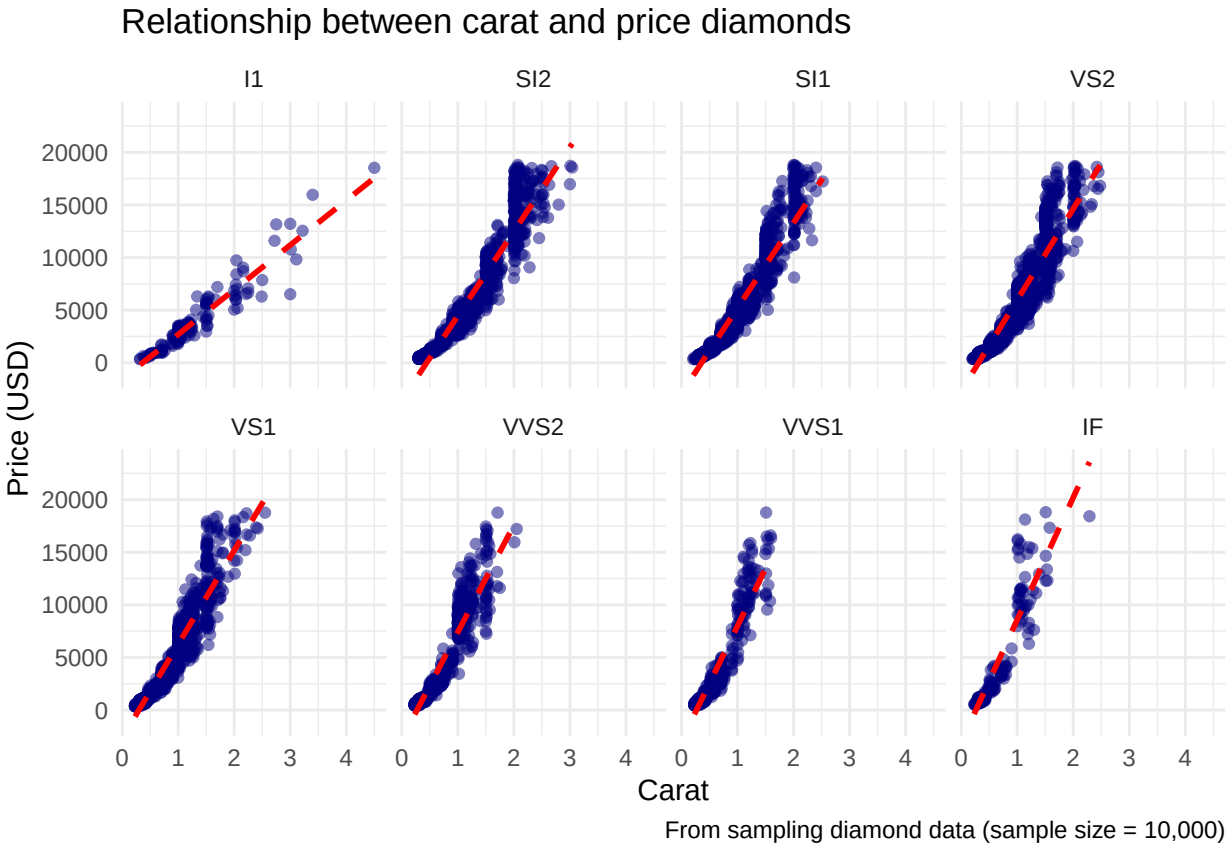
Sample 10000 data points

```
set.seed(10)
diamonds_samp <- diamonds %>%
  sample_n(10000)
```

Create chart from sample data `diamonds_samp`

```
ggplot(diamonds_samp, aes(x = carat, y = price)) +
  geom_point(color = 'navy', alpha = 0.5) +
  geom_smooth(method = 'lm', se = F, color = 'red', linetype = 'dashed') +
  facet_wrap(~ clarity, ncol = 4) +
  labs(
    title = "Relationship between carat and price diamonds",
    x = "Carat",
    y = "Price (USD)",
    caption = "From sampling diamond data (sample size = 10,000)"
  ) +
  theme_minimal()
```

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



By all clarity levels, price and carat has a positive relationship. `IF` group's price rises sharply when carat increases, while `I1`'s price rises steadily when carat increases.