Data Exploration and Visualization with R & ggplot

Pongsakorn Tikapichart

Load Library

Install and load following libraries:

- tidyverse
- patchwork
- lubridate

```
# install.packages(c("tidyverse","patchwork","lubridate"))

library(tidyverse)
library(patchwork)
library(lubridate)
```

Examining the Dataset

using glimpse function to examine the data

```
glimpse(diamonds)
```

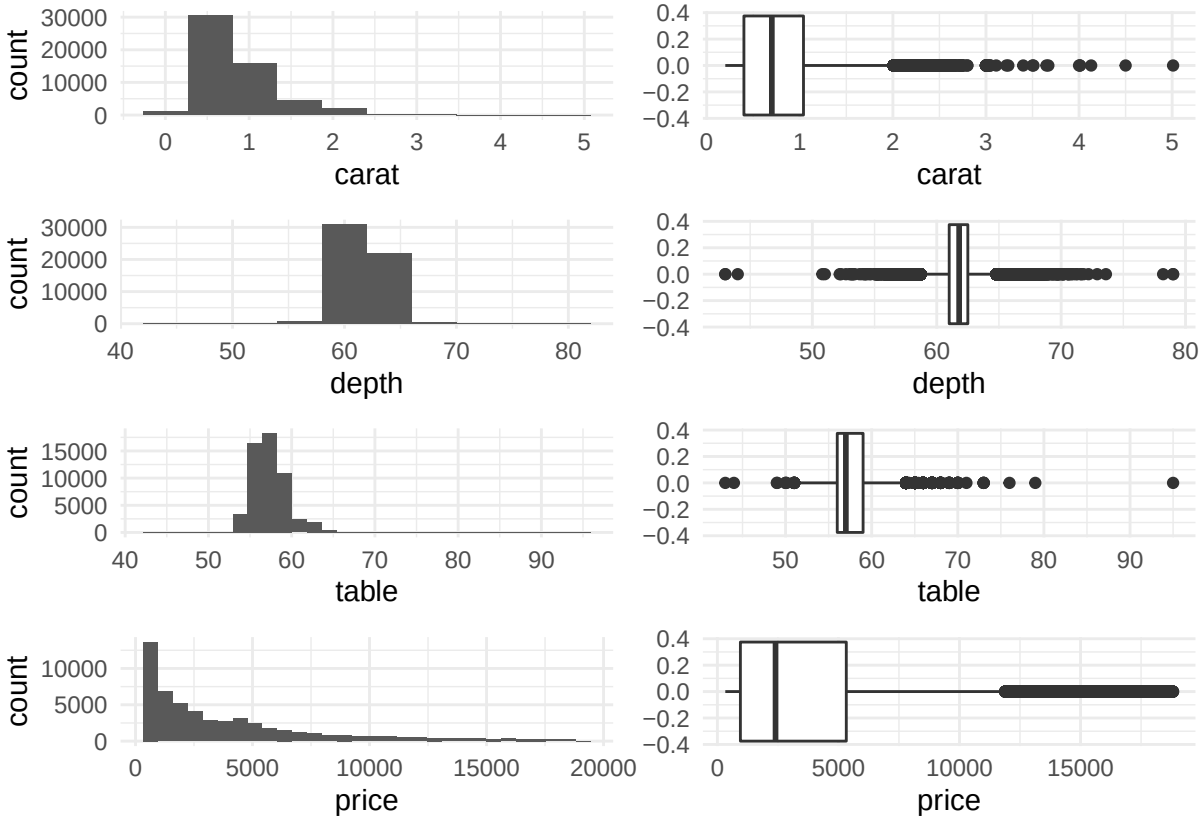
```
## Rows: 53,940
## Columns: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, 0.23, 0.~
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very Good, Ver~
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, J, J, I,~
## $ clarity <ord>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

10 features and 53940 diamonds on this data set.

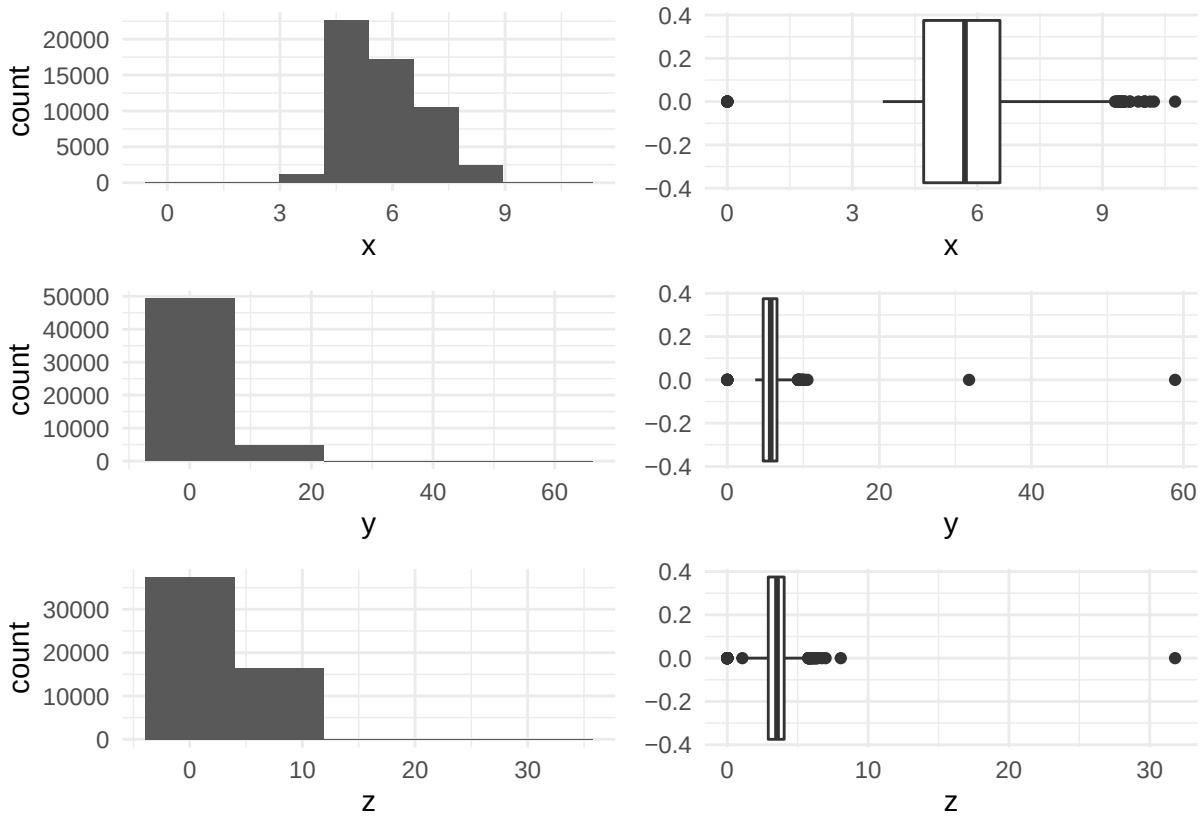
Data distribution

check the data distribution and outlier of continuous data using histogram chart and box chart.

```
df <- diamonds
h_plot_carat <- ggplot(df, aes(carat)) + geom_histogram(bins = 10) + theme_minimal()
h_plot_depth <- ggplot(df, aes(depth)) + geom_histogram(bins = 10)  + theme_minimal()
h_plot_table <- ggplot(df, aes(table)) + geom_histogram()  + theme_minimal()
h_plot_price <- ggplot(df, aes(price)) + geom_histogram()  + theme_minimal()


box_plot_carat <- ggplot(df, aes(carat)) + geom_boxplot() + theme_minimal()
box_plot_depth <- ggplot(df, aes(depth)) + geom_boxplot() + theme_minimal()
box_plot_table <- ggplot(df, aes(table)) + geom_boxplot() + theme_minimal()
box_plot_price <- ggplot(df, aes(price)) + geom_boxplot() + theme_minimal()

(h_plot_carat + box_plot_carat) /
  (h_plot_depth + box_plot_depth) /
  (h_plot_table + box_plot_table) /
  (h_plot_price + box_plot_price)
```

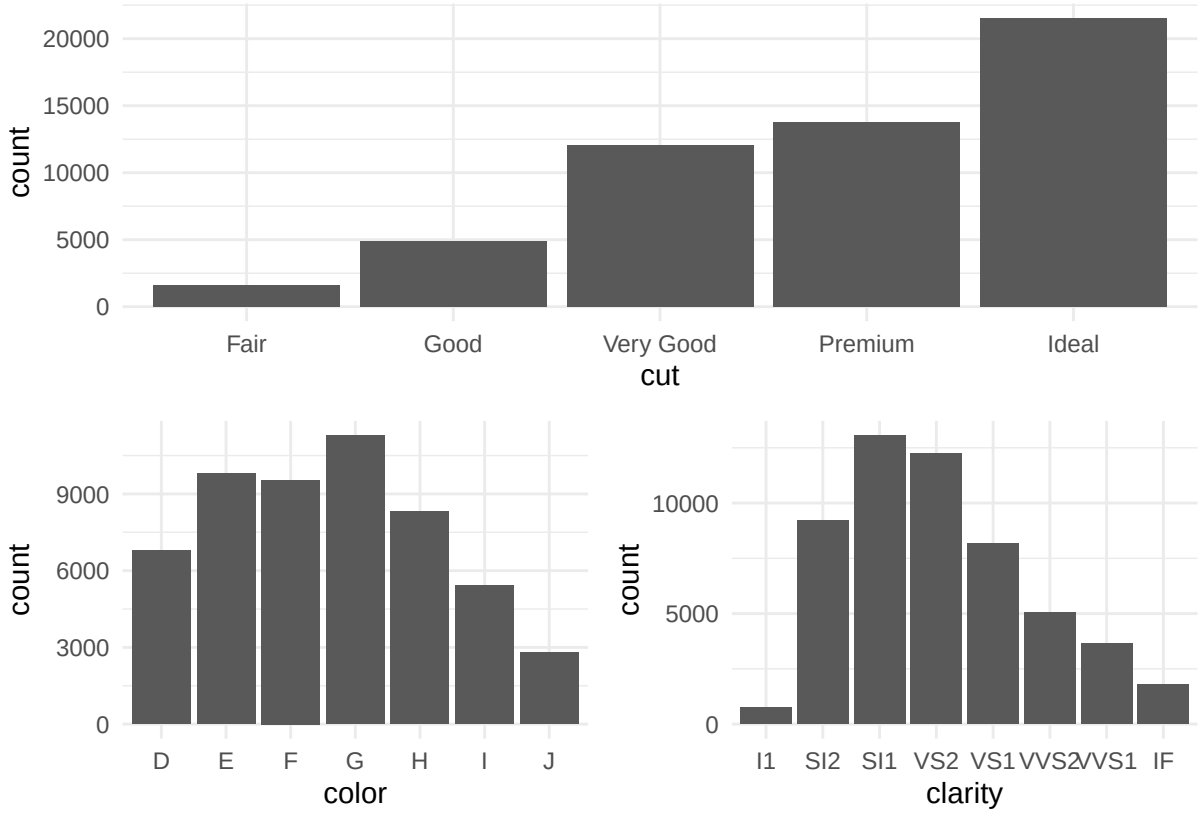


```r
h_plot_x <- ggplot(df, aes(x)) + geom_histogram(bins = 10) + theme_minimal()
h_plot_y <- ggplot(df, aes(y)) + geom_histogram(bins = 5)  + theme_minimal()
h_plot_z <- ggplot(df, aes(z)) + geom_histogram(bins = 5)  + theme_minimal()

box_plot_x <- ggplot(df, aes(x)) + geom_boxplot() + theme_minimal()
box_plot_y <- ggplot(df, aes(y)) + geom_boxplot() + theme_minimal()
box_plot_z <- ggplot(df, aes(z)) + geom_boxplot() + theme_minimal()


(h_plot_x + box_plot_x)/ (h_plot_y +  box_plot_y) /(h_plot_z + box_plot_z)
```



check the data distribution of discrete data using bar chart.

```
b_plot_cut <- ggplot(df, aes(cut)) + geom_bar()  + theme_minimal()
b_plot_color <- ggplot(df, aes(color)) + geom_bar()  + theme_minimal()
b_plot_clarity <- ggplot(df, aes(clarity)) + geom_bar()  + theme_minimal()

b_plot_cut / (b_plot_color + b_plot_clarity)
```



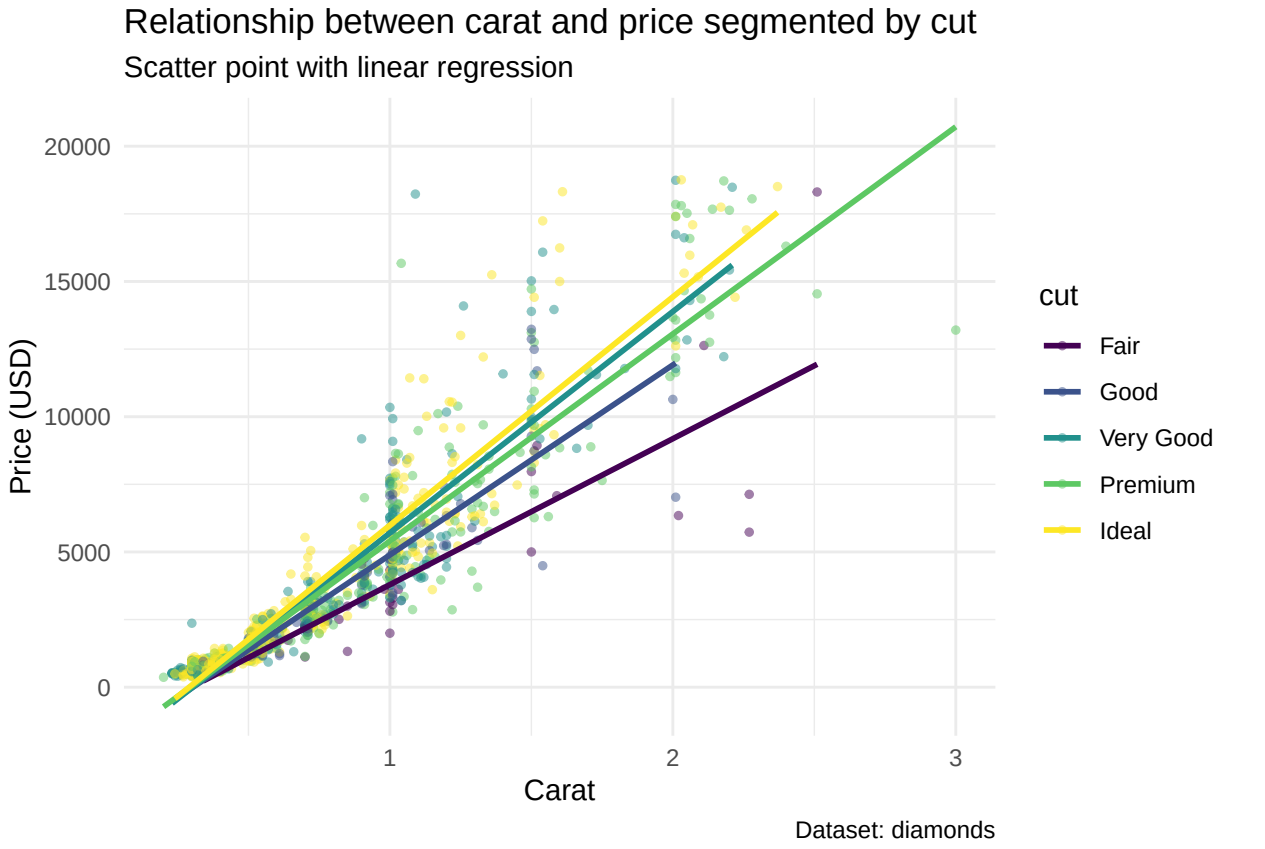
Find the relationship between each variable.

Carat, price, and cut

The following graph shows the relationship between the carat and price segmented by cut.

According to 'Too many data point' to illustrate in the chart, only 1,000 sample is randomly chosen.

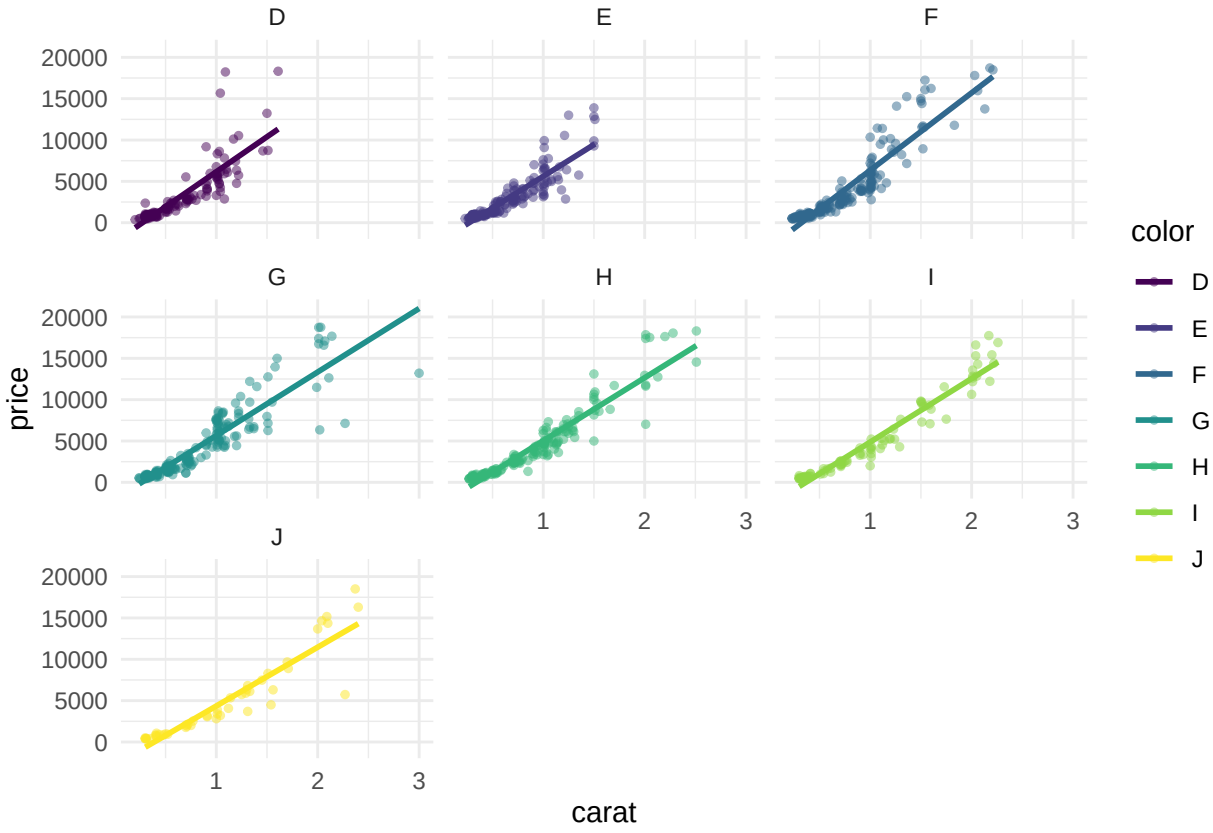
```
set.seed(1234)
ggplot(sample_n(df,1000), aes(carat, price, color= cut)) +
  geom_point(alpha=0.5, size=1) +
  geom_smooth(method="lm", se=F) +
  theme_minimal() +
  labs(
    title = "Relationship between carat and price segmented by cut",
    subtitle = "Scatter point with linear regression",
    x = "Carat",
    y = "Price (USD)",
    caption = "Dataset: diamonds"
  )
```



This graph shows the more carat, the more price.In addition, the more quality of cut would, the more price is set. However, the Premium quality diamonds' price is lower than the 'Very Good' quality diamond's price. That's need to be further investigation.

Carat, price, and color

```
set.seed(1234)
ggplot(sample_n(df,1000), aes(carat, price, color= color)) +
  geom_point(alpha=0.5, size=1) +
  geom_smooth(method="lm", se=F) +
  theme_minimal() +
  facet_wrap(~ color, ncol=3)
```



Carat, price, cut, clarity

```
set.seed(1234)
ggplot(sample_n(df,5000), aes(carat, price, color= cut)) +
  geom_point(alpha=0.5, size=1) +
  geom_smooth(method="lm", se=F) +
  theme_minimal() +
  facet_grid(cut ~ clarity)
```