Exploring Data Analysis in R

Sukrita Rojvattanakarn

Install library

```
install.packages(c("tidyverse", "patchwork", "ggthemes"))
```

```
## Installing packages into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
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   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(patchwork)
library(ggthemes)
```

Data Overview (Diamonds)

Diamonds dataset contains 53,940 rows with 10 variables which are carat, cut, color, clarity, depth, price, x(length in mm), y (width in mm) and z (depth in mm).

```
d <- diamonds
glimpse(d)
```

```
## 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.~
```

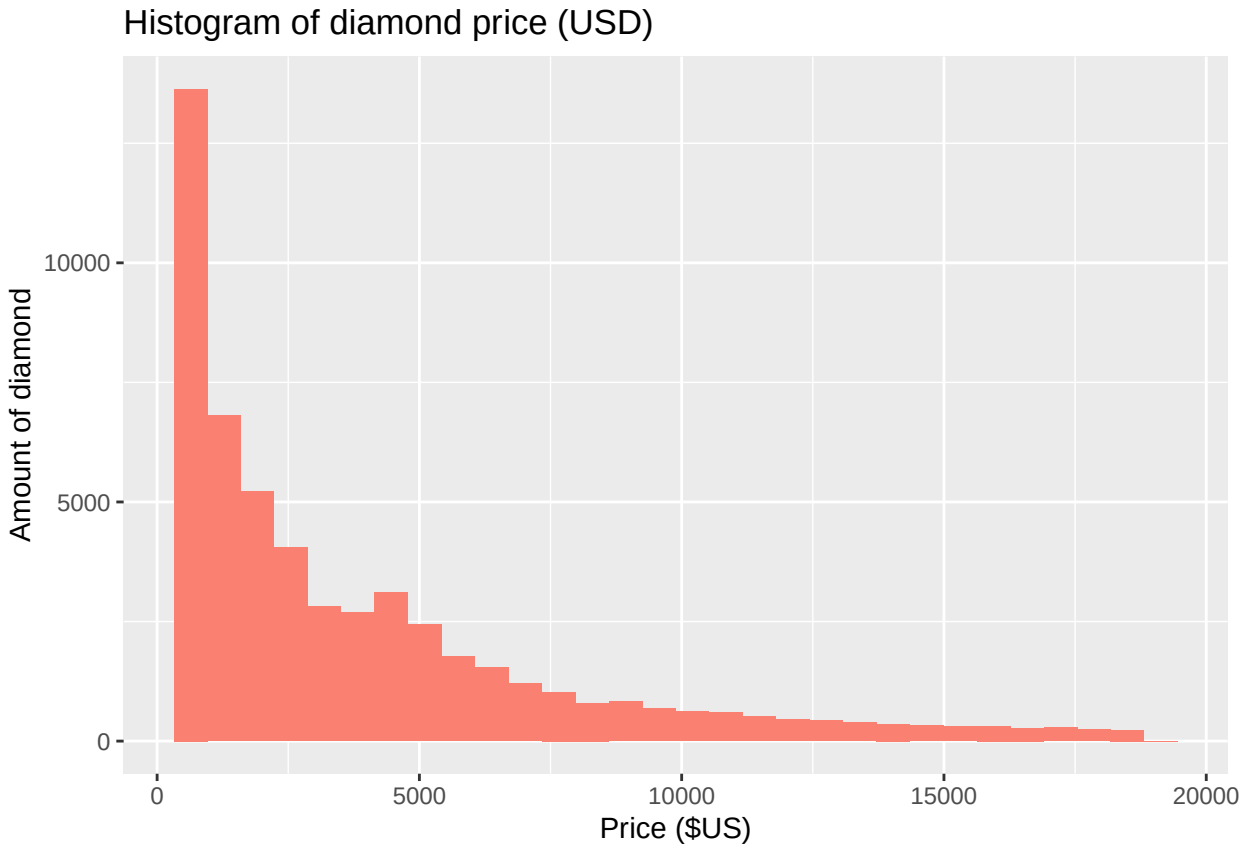
Data Description

- carat = weight of diamond (0.2-5.01)
- cut = quality of the cut(fair, good, very good, premium, ideal)
- clarity = measurement of how clear the diamond is (I1 - worst, SI2, SI1, VS2, VS1, VVS1, IF(best))
- depth = total depth percentage (43-79)
- table = width of top of diamond relative to widest point (43-95)
- price = price in $US ($326-$18,823)
- x = length in mm (0-10.74)
- y = width in mm (0.58-9)
- z = depth in mm (0-31.8)

Graph 1: Histogram of diamond price (USD)

```
ggplot(d, aes(price)) +
      geom_histogram(fill="salmon") +
      labs(
        title = "Histogram of diamond price (USD)",
        x = "Price ($US)",
        y = "Amount of diamond"
      )
```

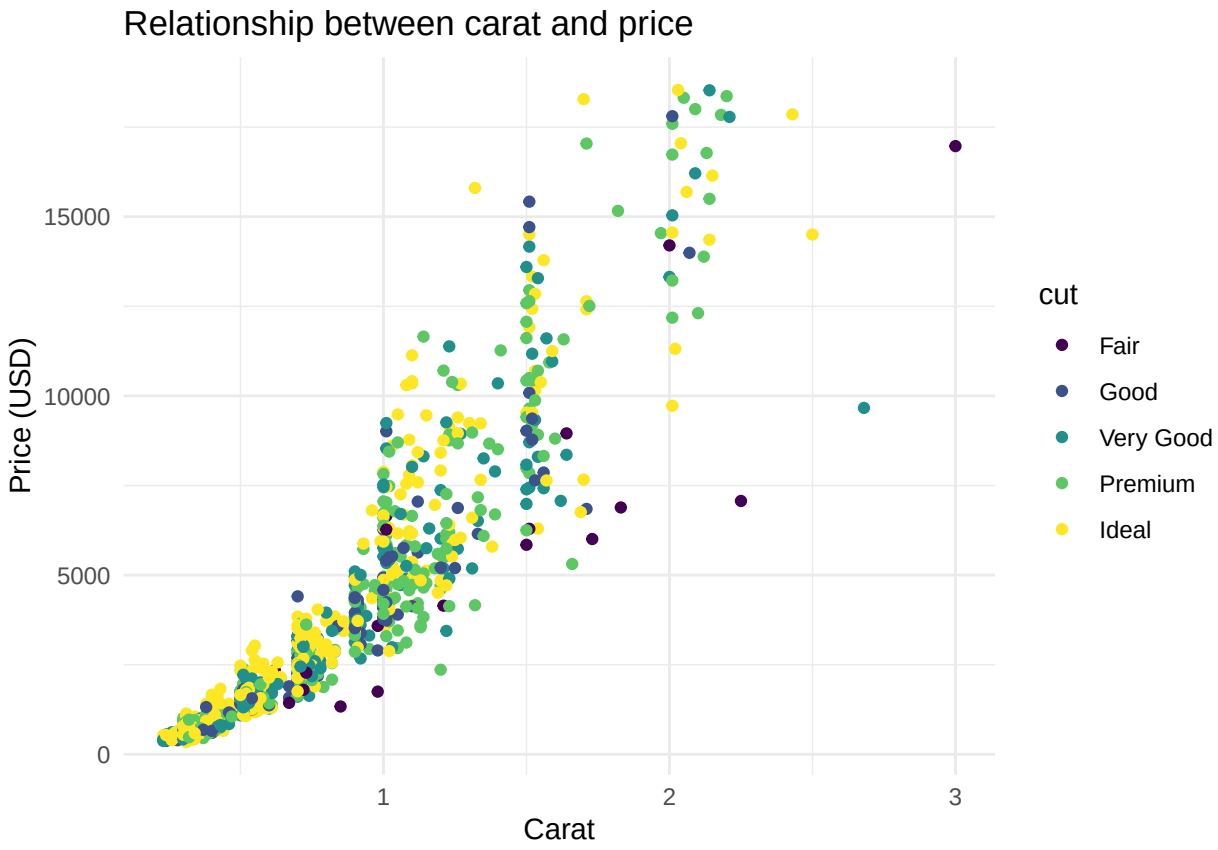
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This graph shows most of diamonds fall in to range between $0-5,000

Graph 2 Relationship between of Carat VS. Price

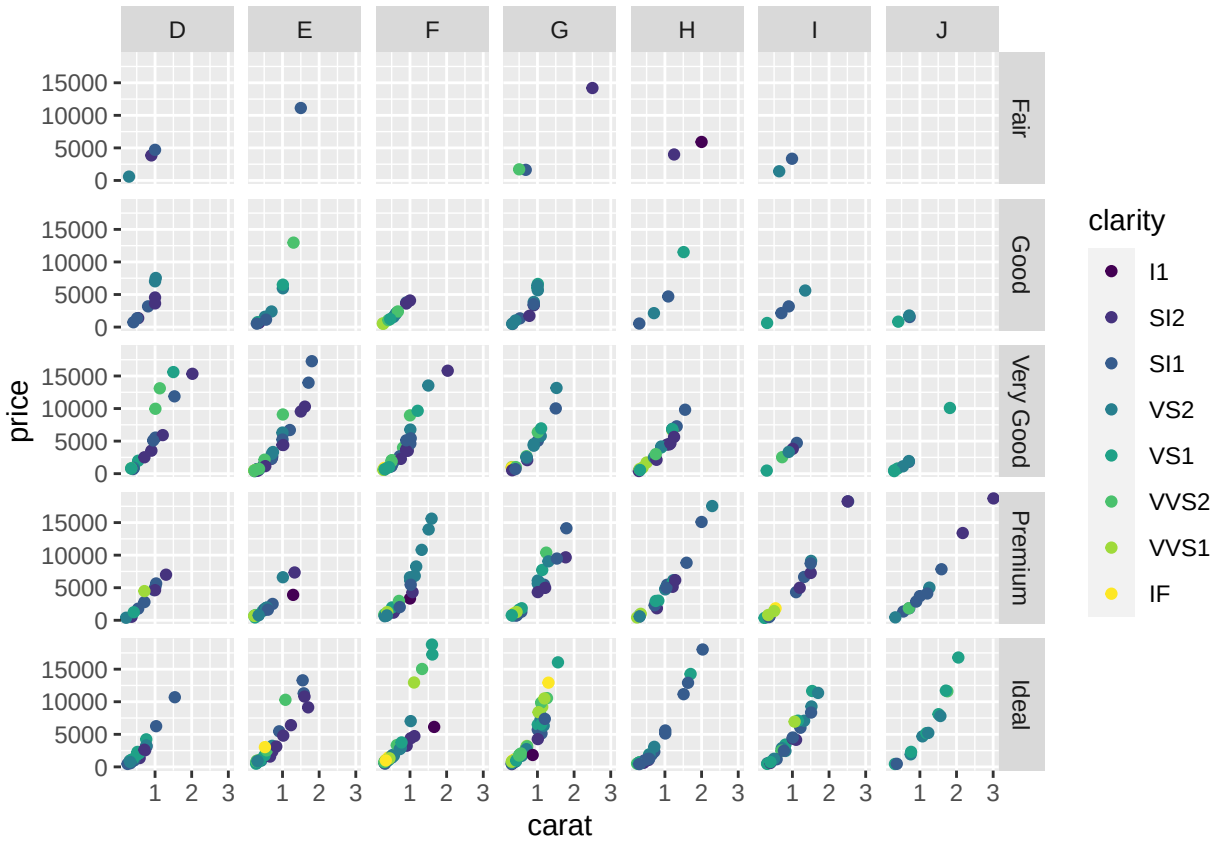
```
ggplot(sample_n(d, 1000),
       aes(carat, price, color=cut)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Relationship between carat and price",
       x = "Carat",
       y = "Price (USD)")
```



This graph shows the more light weight, the more cheaper. Yet, weight and price doesn't relate with the cut.

Graph 3 Relationship between Carat VS. Clarity

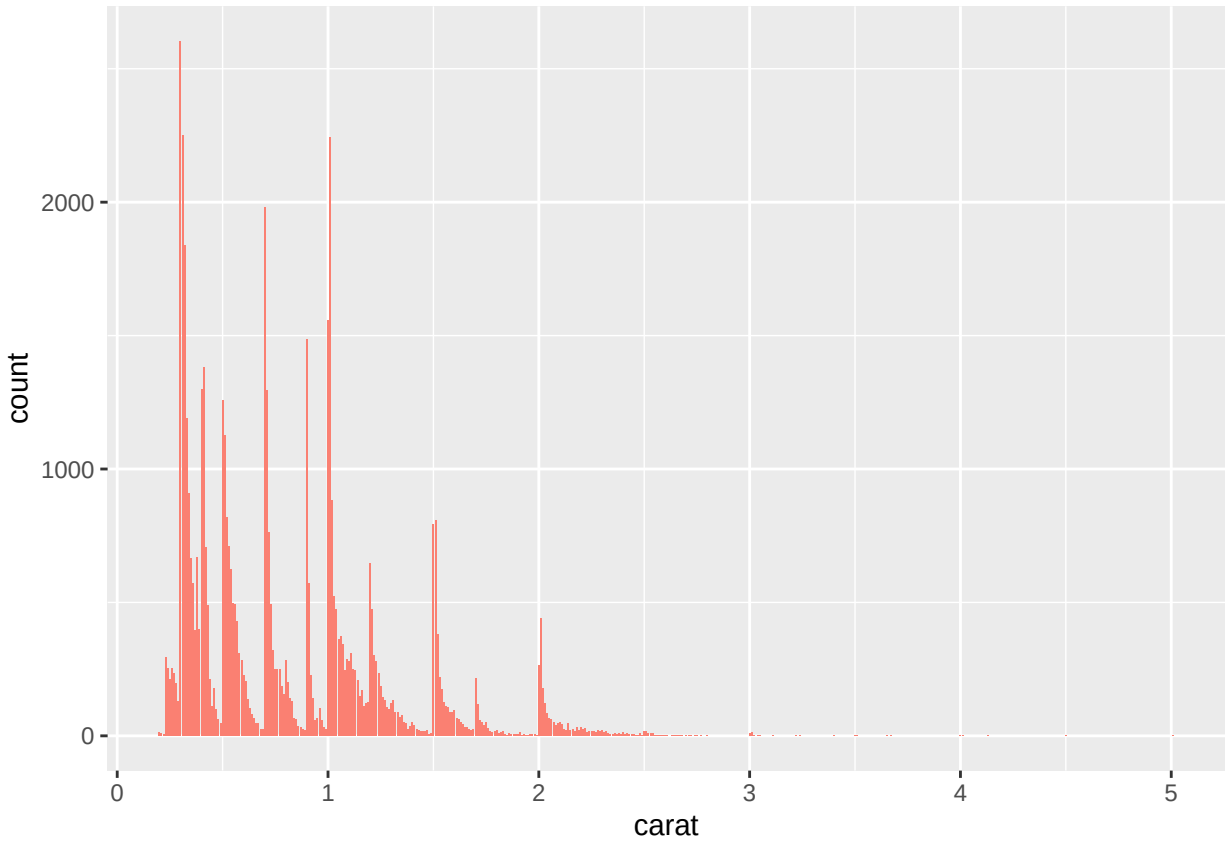
```
ggplot(sample_n(d,500),
       aes(carat, price, color=clarity)) +
       geom_point() +
       facet_grid(cut ~ color)
```



This graph shows that clarity doesn't relate with the price.

Graph 4 Count Carat

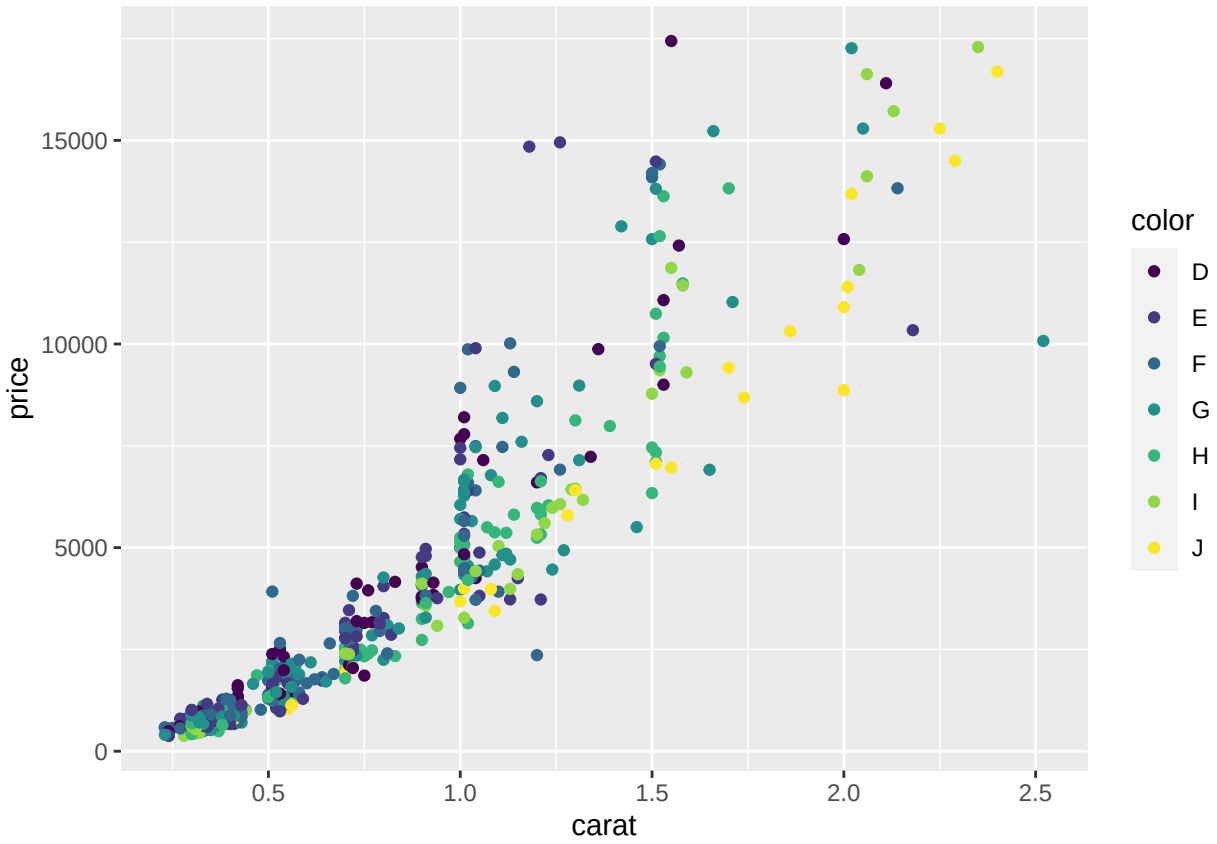
```
ggplot(d, aes(x = carat)) +
  geom_bar(position = "dodge", fill="salmon")
```



Most of carat in this dataset falls into 0-1 weight.

Graph 5 Find relationship between color and price

```
ggplot(sample_n(d,500),
       aes(carat, price, color=color)) +
       geom_point()
```



Color doesn't have much effect on price but weight