Homework Data Viz Batch 10

Maruko

2024-08-19

Loading Library

```
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
```

```
print("Load library for data visualization")
```

```
## [1] "Load library for data visualization"
```

Veiw top 10 data

```
head(diamonds,10)
```

```
## # A tibble: 10 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
##  2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
##  3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
##  4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
##  5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
##  6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
##  7  0.24 Very Good I     VVS1     62.3    57   336  3.95  3.98  2.47
##  8  0.26 Very Good H     SI1      61.9    55   337  4.07  4.11  2.53
##  9  0.22 Fair      E     VS2      65.1    61   337  3.87  3.78  2.49
## 10  0.23 Very Good H     VS1      59.4    61   338  4     4.05  2.39
```

Preparation of data

filter out outliers

```
set.seed(42)
base <- diamonds %>%
  filter(carat < 2) %>%
  sample_n(1000)
base1 <- base %>% filter(carat < 1)
base2 <- base %>% filter(carat >= 1)

print(base)
```

```
## # A tibble: 1,000 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.4  Good      F     VS2      62    61.3   929  4.69  4.72  2.91
##  2  1.12 Very Good G     SI2      63.3  58    4478  6.7   6.63  4.22
##  3  0.56 Ideal     D     VS2      61.1  56    1963  5.3   5.33  3.25
##  4  0.57 Ideal     D     VS1      61.7  56    2091  5.31  5.33  3.28
##  5  1.23 Ideal     H     SI1      61.5  57    6681  6.92  6.89  4.25
##  6  1.01 Fair      F     SI1      67.2  60    4276  6.06  6     4.05
##  7  0.4  Ideal     D     VS2      61.3  57    1050  4.77  4.75  2.92
##  8  0.9  Ideal     D     SI1      62.1  57    4523  6.18  6.25  3.86
##  9  0.32 Ideal     E     VVS1     61.7  55     917  4.39  4.43  2.72
## 10  0.61 Good      G     VS2      61.2  62.8  1821  5.38  5.42  3.3
## # i 990 more rows
```

```
print(base1)
```

```
## # A tibble: 679 x 10
##    carat cut   color clarity depth table price     x     y     z
##    <dbl> <ord> <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  0.4  Good  F     VS2      62    61.3   929  4.69  4.72  2.91
##  2  0.56 Ideal D     VS2      61.1  56    1963  5.3   5.33  3.25
##  3  0.57 Ideal D     VS1      61.7  56    2091  5.31  5.33  3.28
##  4  0.4  Ideal D     VS2      61.3  57    1050  4.77  4.75  2.92
##  5  0.9  Ideal D     SI1      62.1  57    4523  6.18  6.25  3.86
##  6  0.32 Ideal E     VVS1     61.7  55     917  4.39  4.43  2.72
##  7  0.61 Good  G     VS2      61.2  62.8  1821  5.38  5.42  3.3
##  8  0.57 Ideal H     SI1      61.8  54    1292  5.35  5.37  3.31
##  9  0.33 Ideal E     VS2      60.5  55     738  4.52  4.56  2.74
## 10  0.5  Ideal E     SI1      61.2  56    1555  5.12  5.15  3.14
## # i 669 more rows
```

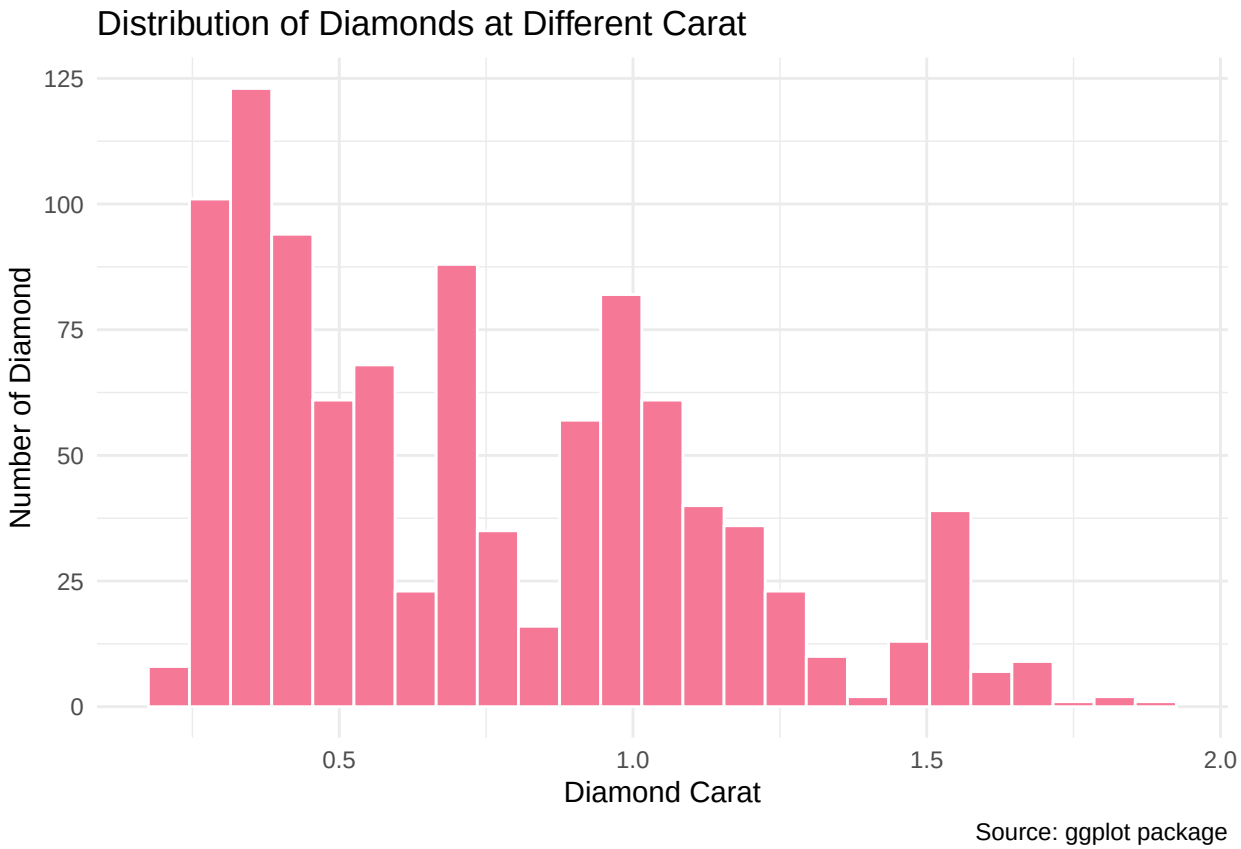
```
print(base2)
```

```
## # A tibble: 321 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  1.12 Very Good G     SI2      63.3    58  4478  6.7   6.63  4.22
##  2  1.23 Ideal     H     SI1      61.5    57  6681  6.92  6.89  4.25
##  3  1.01 Fair      F     SI1      67.2    60  4276  6.06  6     4.05
##  4  1.04 Ideal     G     SI1      61.9    53  5570  6.53  6.54  4.05
##  5  1.55 Ideal     F     SI2      61.9    55 10937  7.44  7.4   4.6
##  6  1.2  Very Good J     VS2      62.6    57  4963  6.72  6.8   4.23
```

```
##  7  1.01 Ideal     H     VS2      61      59  5238  6.51  6.47  3.96
##  8  1.15 Ideal     G     SI1      62      55  6313  6.72  6.76  4.18
##  9  1.26 Fair      I     SI2      64.8    57  4551  6.73  6.69  4.35
## 10  1.52 Premium   G     VVS2     62.1    58 14105  7.4   7.31  4.57
## # i 311 more rows
```

1. Histogram

```
ggplot(base, aes(carat)) +
  geom_histogram(bins = 25,
                 col = "white",
                 fill = "#f57897") +
  theme_minimal() +
  labs(title = "Distribution of Diamonds at Different Carat",
       caption = "Source: ggplot package",
       x = "Diamond Carat",
       y = "Number of Diamond")
```

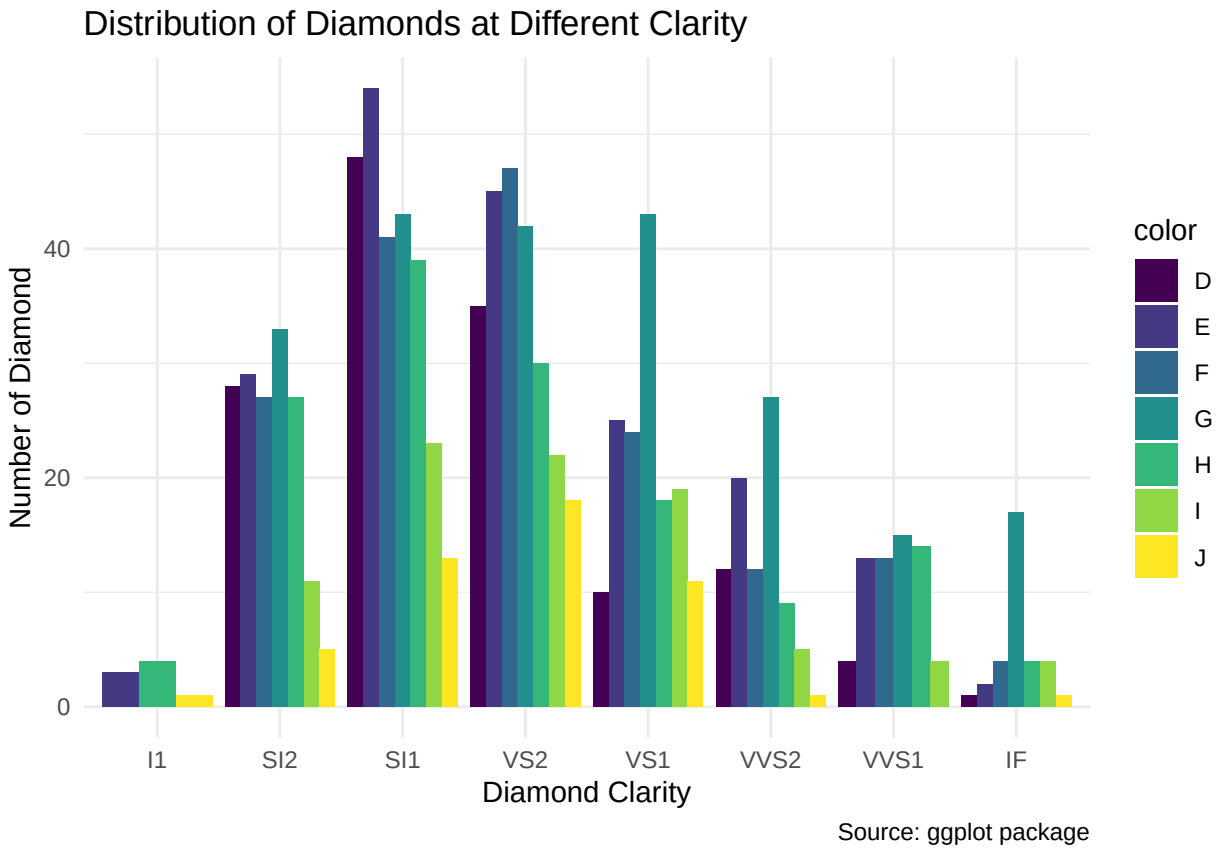


This histogram is shown that 0.2-0.4 is range of high of diamond cut

2. Bar plot by group

```
ggplot(base, aes(clarity, fill = color)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
```

```
  labs(title = "Distribution of Diamonds at Different Clarity",
       caption = "Source: ggplot package",
       x = "Diamond Clarity",
       y = "Number of Diamond")
```

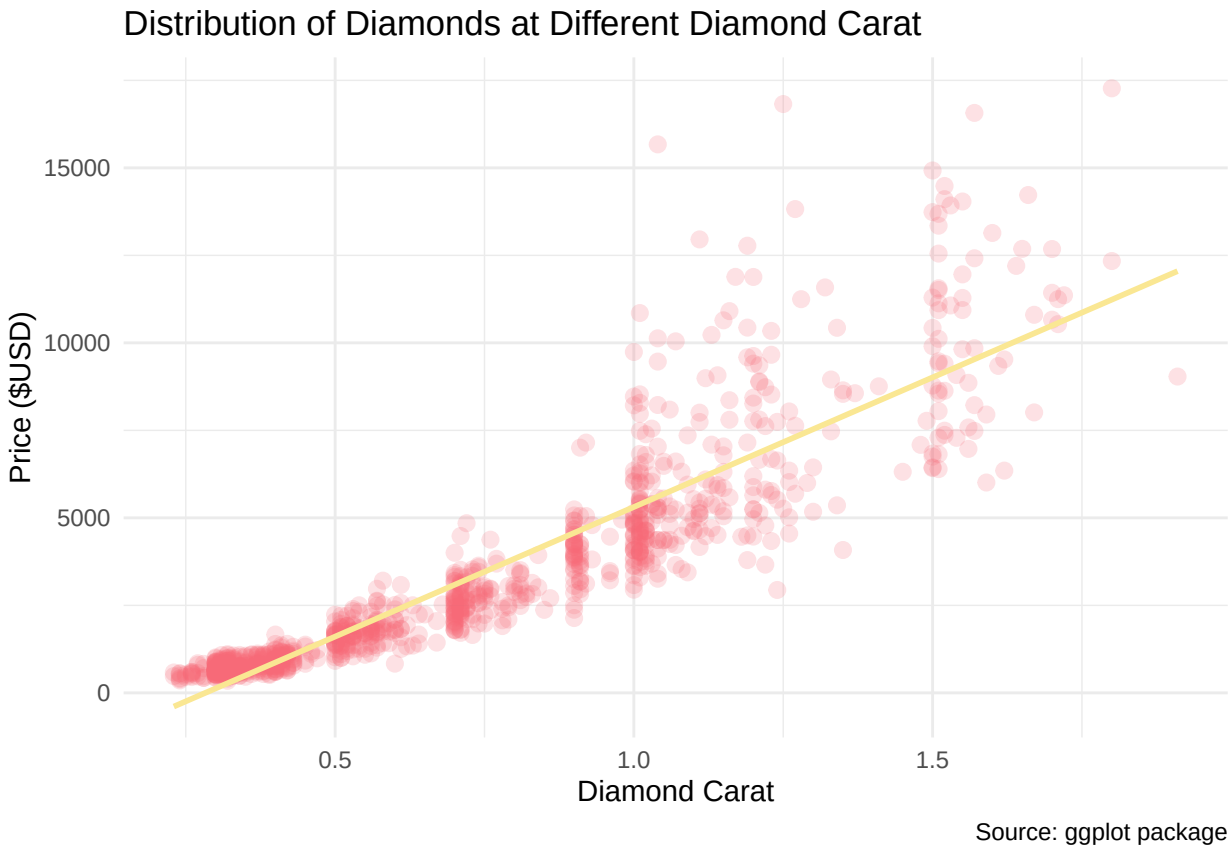


This bar chart is appeared on the highest diamond clarity that is SI1.

3. Scatter plot

```
ggplot(base, aes(carat, price)) +
  geom_point(size = 2.5,
             alpha = 0.2,
             col = "#f76a78") +
  geom_smooth(method = "lm",
              col = "#fae793",
              se = FALSE) +
  theme_minimal() +
  labs(title = "Distribution of Diamonds at Different Diamond Carat",
       caption = "Source: ggplot package",
       x = "Diamond Carat",
       y = "Price ($USD)")
```

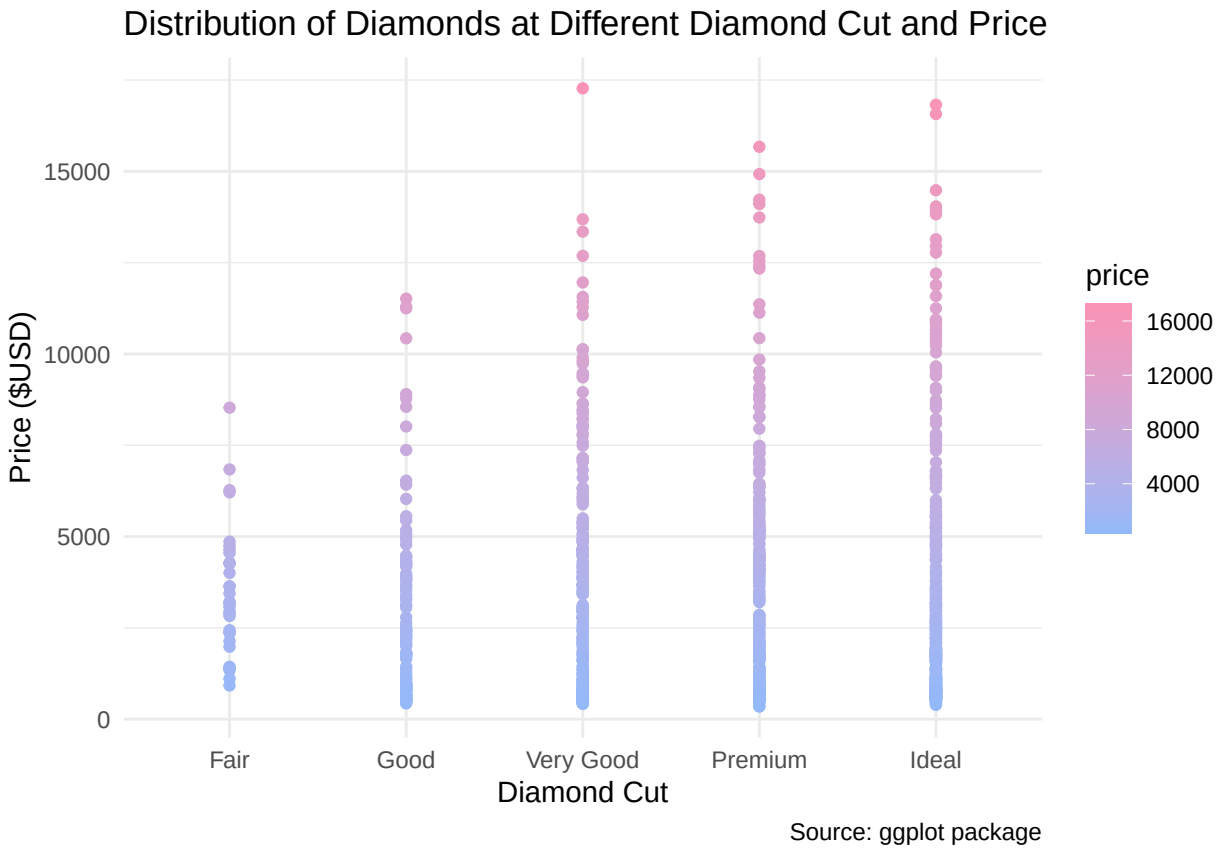
```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatter plot show relation between price and carat, increasing of diamond carat increase price of diamond.

4. Scatter plot with color gradient

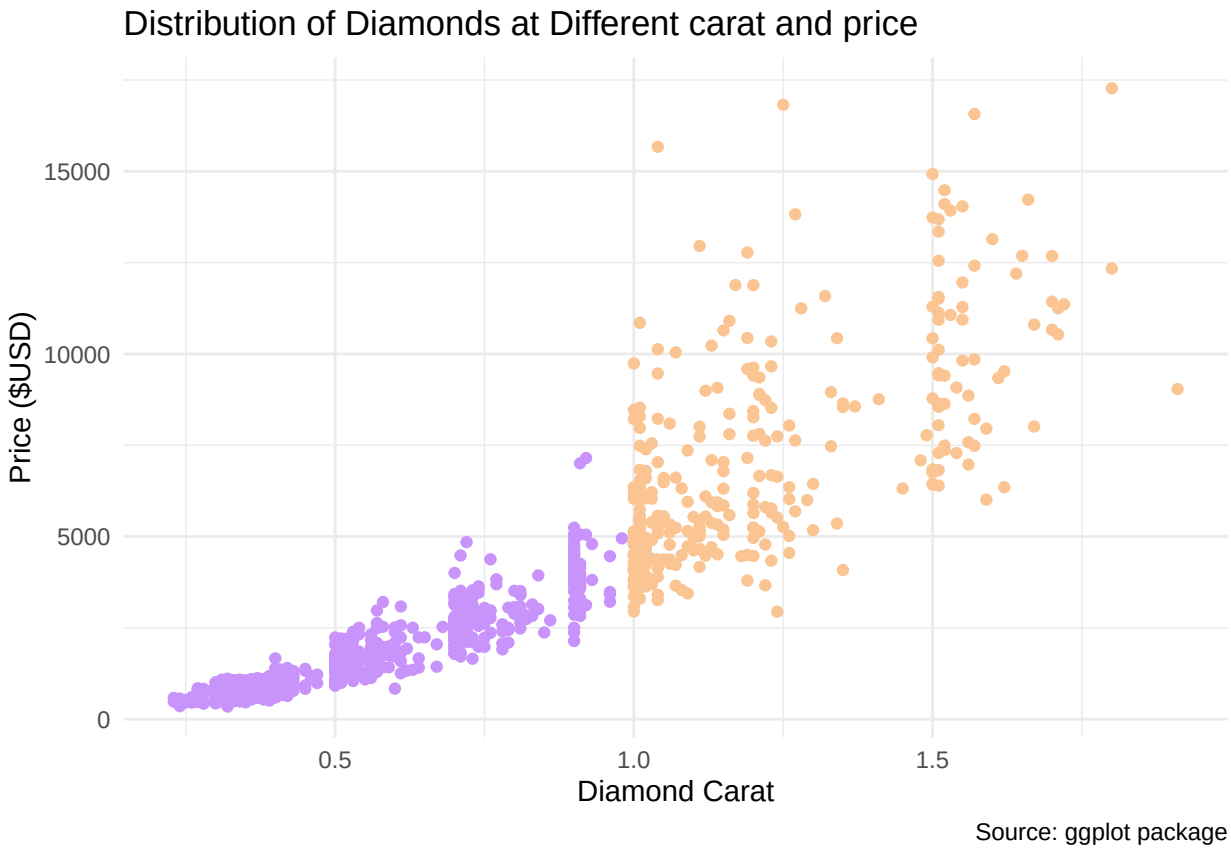
```
ggplot(base, aes(cut, price, colour = price)) +
  geom_point() +
  scale_color_gradient(low = "#93b9fa", high = "#fa93b4") +
  theme_minimal() +
  labs(title = "Distribution of Diamonds at Different Diamond Cut and Price",
       caption = "Source: ggplot package",
       x = "Diamond Cut",
       y = "Price ($USD)")
```



This scatter plot with color gradient is shown that the high number of diamond relate with increasing price.

5. Scatter plot with multiple dataframe

```
ggplot() +
  geom_point(data = base1,
       mapping = aes(carat, price),
       color = "#c893fa") +
  geom_point(data = base2,
             mapping = aes(carat, price),
             color = "#fac593") +
  theme_minimal() +
  labs(title = "Distribution of Diamonds at Different carat and price",
       caption = "Source: ggplot package",
       x = "Diamond Carat",
       y = "Price ($USD)")
```



This scatter plot with multiple dataframe is appeared on less diamond carat that is less price and large diamond carat that is large price.