Data Visualization Bootcamp Homework

Sakon Sukkhapanya Sangiamsri

2023-08-15

Introduction

The purpose of this markdown is to create "Data Visualization" lesson for Data Analysis course only.

Assignment

I. Install and use library "ggplot2" package. ("ggthemes" is optional.)
II. Create any charts by use "diamonds" data set at least 5 ones and describe the insight for each one.
III. Knit into .PDF file and share in Discord.

Assume that you install "ggplot2" already. Then use library(ggplot2) to allow you use these functions for data visualization creating.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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
```

```
library(ggplot2)
library(ggthemes)
```

Scanning a whole data set

In this step, we will use glimpse() and summary() to make some questions and answer them by EDA (Exploratory Data Analysis) with "Data Visualization".

```r
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.~
```

```r
summary(diamonds)
```

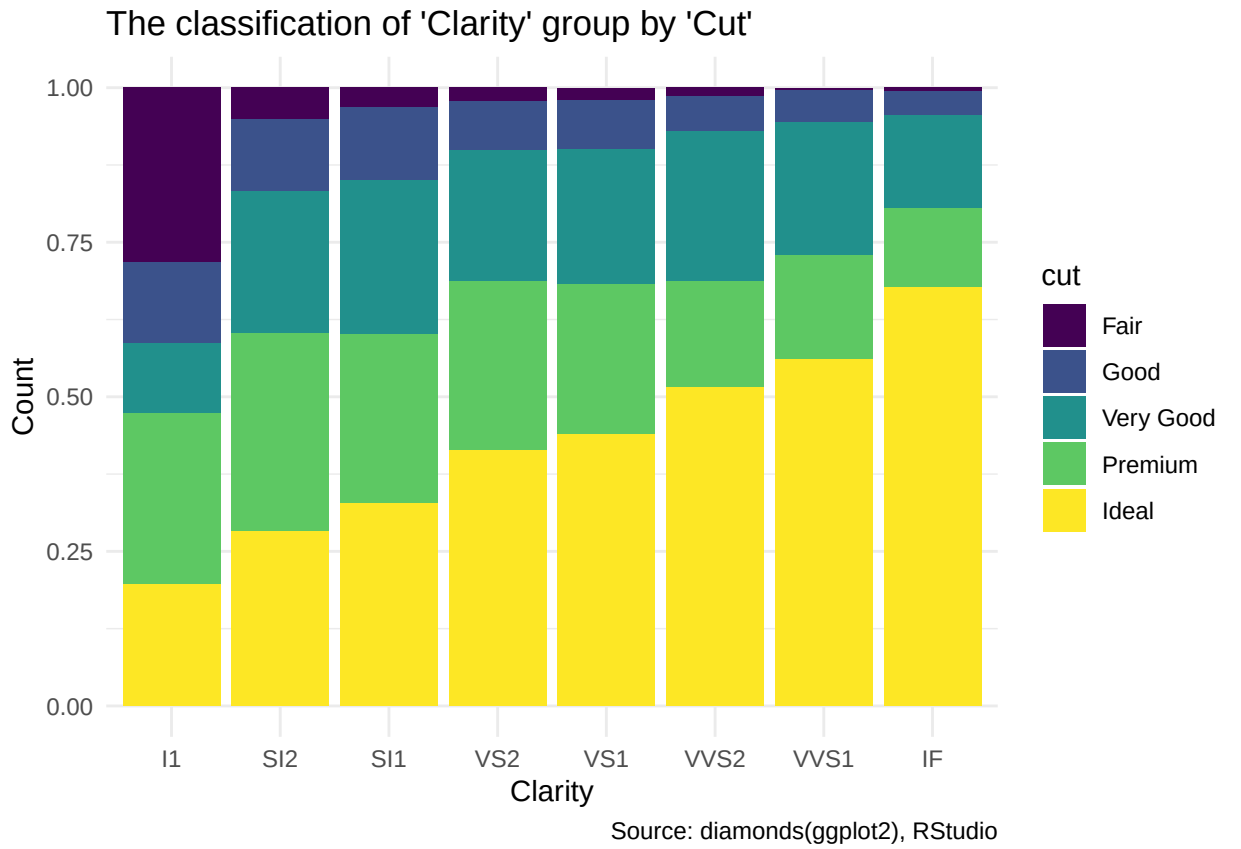
```
##      carat               cut        color        clarity          depth      
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1    :13065   Min.   :43.00  
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2    :12258   1st Qu.:61.00  
##  Median :0.7000   Very Good:12082   F: 9542   SI2    : 9194   Median :61.80  
##  Mean   :0.7979   Premium   :13791   G:11292   VS1    : 8171   Mean   :61.75  
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2   : 5066   3rd Qu.:62.50  
##  Max.   :5.0100                     I: 5422   VVS1   : 3655   Max.   :79.00  
##                                     J: 2808   (Other): 2531                  
##      table           price             x                y         
##  Min.   :43.00   Min.   :  326   Min.   : 0.000   Min.   : 0.000  
##  1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710   1st Qu.: 4.720  
##  Median :57.00   Median : 2401   Median : 5.700   Median : 5.710  
##  Mean   :57.46   Mean   : 3933   Mean   : 5.731   Mean   : 5.735  
##  3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540   3rd Qu.: 6.540  
##  Max.   :95.00   Max.   :18823   Max.   :10.740   Max.   :58.900  
##                                                                   
##        z         
##  Min.   : 0.000  
##  1st Qu.: 2.910  
##  Median : 3.530  
##  Mean   : 3.539  
##  3rd Qu.: 4.040  
##  Max.   :31.800  
##
```

Then, we will create some charts at least 5 charts.

Bar chart

1. Find the distribution of price

```
ggplot(diamonds, aes(clarity, fill=cut)) +
  geom_bar(position="fill") +
  theme_minimal() +
  labs(title = "The classification of 'Clarity' group by 'Cut'",
       x = "Clarity",
       y = "Count",
       caption = "Source: diamonds(ggplot2), RStudio")
```



From this chart, there are 8 classes of diamonds clarity and group by cut.
For any diamonds clarity, the number of 'ideal cut' is the most one except these classes,
- I1, the most number of diamonds cut is 'Fair cut'.
- SI2, the most number of diamonds cut is 'Premium cut'.

Scatter chart with trend line

2. Find the trend line of diamonds width and price.

```
ggplot(diamonds, aes(y, price, col=price)) +
  geom_point() +
  geom_smooth(col="lightcoral") +
  scale_color_gradient(low='lightblue', high='lightpink') +
```

```
  theme_minimal() +
  labs(title = "The trend line of 'Diamonds Width and Price'",
       x = "Diamonds Width",
       y = "Price (USD)",
       caption = "Source: diamonds(ggplot2), RStudio")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

Next, we try to explore other insights such as the relationship between y (width in mm) and price of diamonds.
Then, we got the new trend line that it is similar like exponential graph and the range of y is around 0-10 mm by approximately. More the width is near to 10 mm, the higher price is.
But there is something that it allow you to see a few outliers which represent the extreme-value of diamonds width in mm unit.

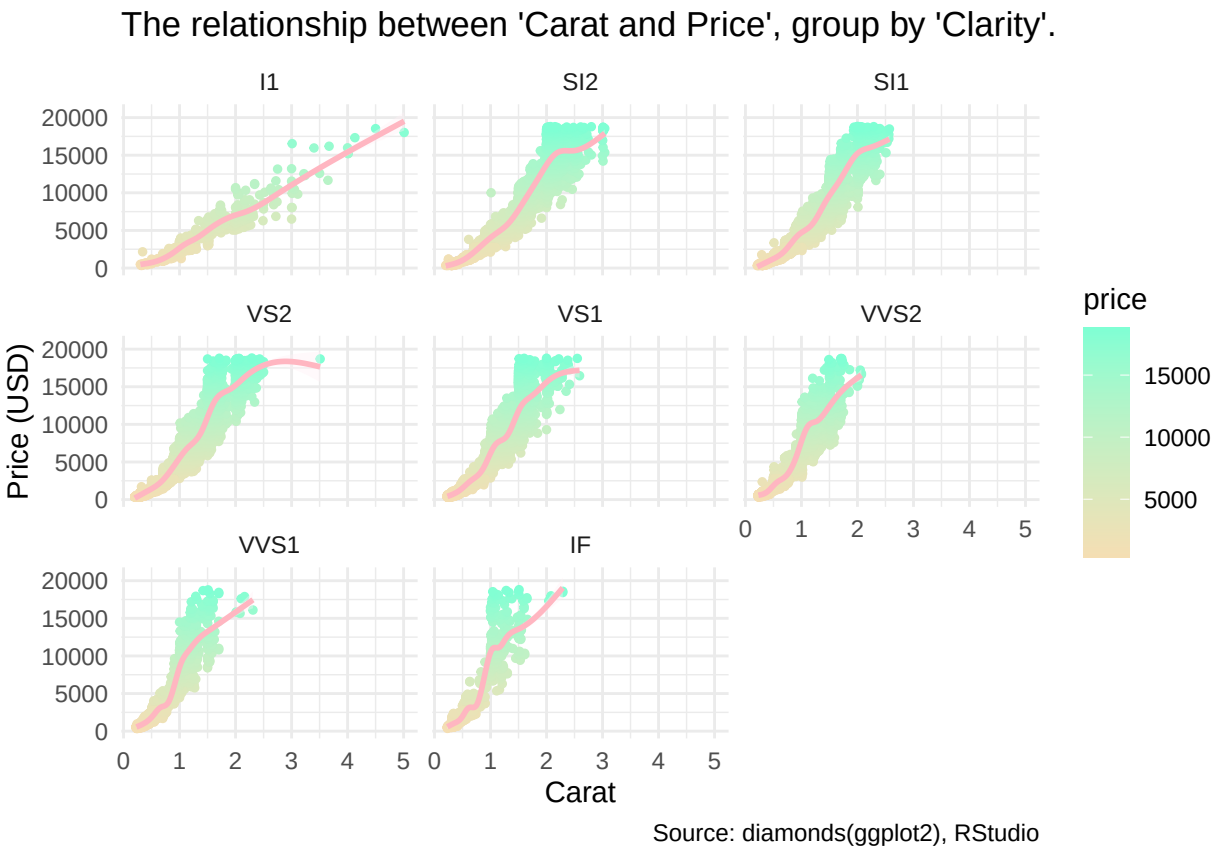
Facet wrap

3. Find the relationship between 'carat and price', group by 'clarity'.

```
ggplot(diamonds, aes(carat, price, col=price)) +
  geom_point(size = 1) +
  geom_smooth(col='lightpink', fill='lavenderblush') +
```

```
  theme_minimal() +
  scale_color_gradient(low='wheat', high='aquamarine') +
  facet_wrap(~clarity) +
  labs(title = "The relationship between 'Carat and Price', group by 'Clarity'.",
       x = "Carat",
       y = "Price (USD)",
       caption = "Source: diamonds(ggplot2), RStudio")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



When we try to explore the relationship between carat and price group by diamonds clarity.
Almost them have similar trend. These patterns tell us that more the carat, more higher the price.
But if we notice these charts thoroughly, the trend line of "VS2" is curving downward when carat is greater than or equal 3.5, and that is an outlier.
Even though "VVS1" and "IF" have some outliers, but the trend and direction between carat and price still related and sensible.

Facet grid

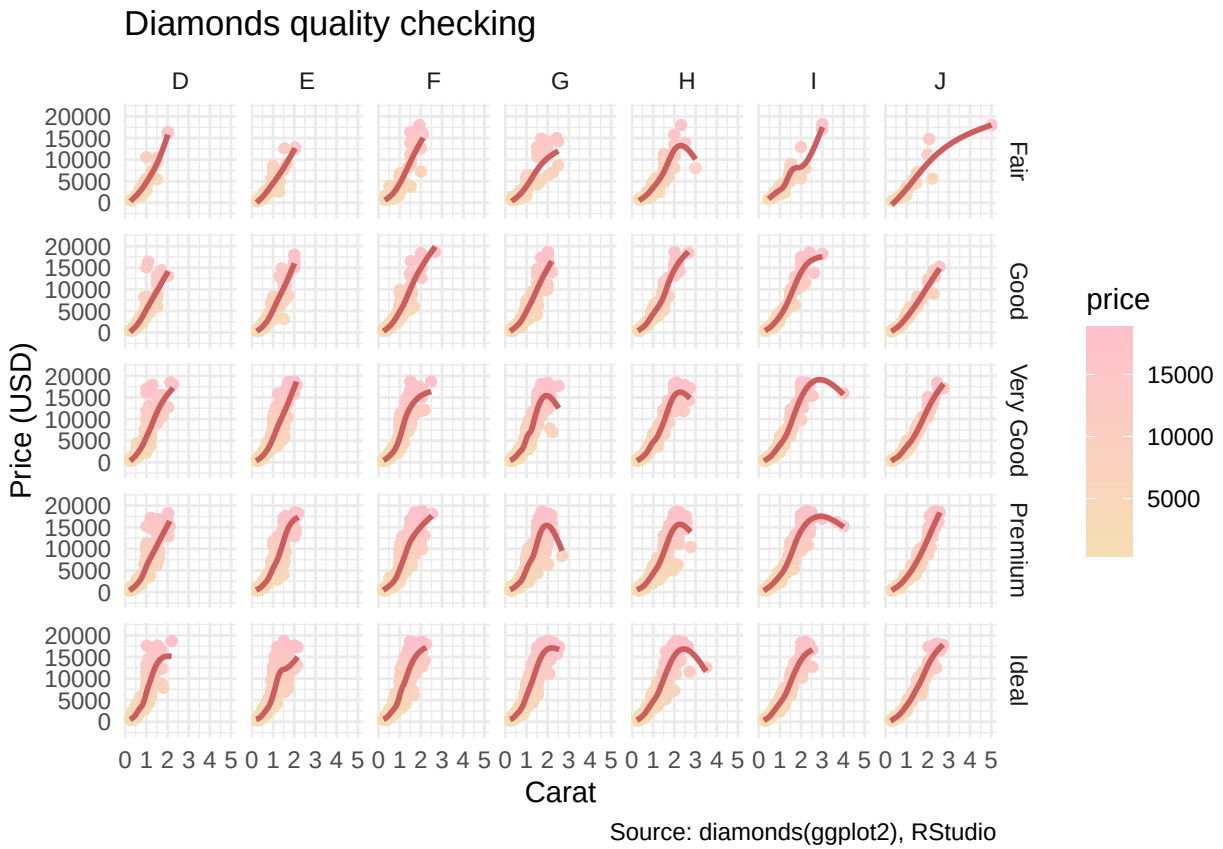
4. 'Diamonds cut' and 'Color' quality checking by random sampling, group by 'Carat and Price'.

```
randomsampling <- sample_frac(diamonds, 0.30)
randomsampling
```

```
## # A tibble: 16,182 x 10
##    carat cut       color clarity depth table price     x     y     z
##    <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
##  1  1    Very Good D     SI1      62.1    61  5390  6.3   6.32  3.92
##  2  0.91 Ideal     G     SI1      62.6    56  3991  6.27  6.3   3.81
##  3  0.41 Premium   G     VS2      61.9    59   827  4.73  4.77  2.94
##  4  0.51 Ideal     E     VS2      61.6    55  1985  5.15  5.2   3.19
##  5  0.35 Premium   J     VS1      62.2    58   591  4.54  4.49  2.81
##  6  1.7  Premium   H     SI2      62.1    56  7673  7.66  7.55  4.72
##  7  0.71 Good      G     VS2      64.2    58  2599  5.59  5.62  3.6
##  8  0.27 Fair      E     VS1      66.4    58   371  3.99  4.02  2.66
##  9  0.41 Very Good F     SI1      62.3    60   679  4.73  4.78  2.96
## 10  0.71 Good      H     SI1      63.4    54  2056  5.6   5.67  3.57
## # i 16,172 more rows
```

```
ggplot(randomsampling, aes(carat, price, col=price)) +
  geom_point() +
  geom_smooth(method='gam', col='indianred', fill='lavenderblush') +
  theme_minimal() +
  scale_color_gradient(low='wheat', high='pink') +
  facet_grid(cut~color) +
  labs(title = "Diamonds quality checking",
      x = "Carat",
      y = "Price (USD)",
      caption = "Source: diamonds(ggplot2), RStudio")
```

```
## `geom_smooth()` using formula = 'y ~ s(x, bs = "cs")'
```



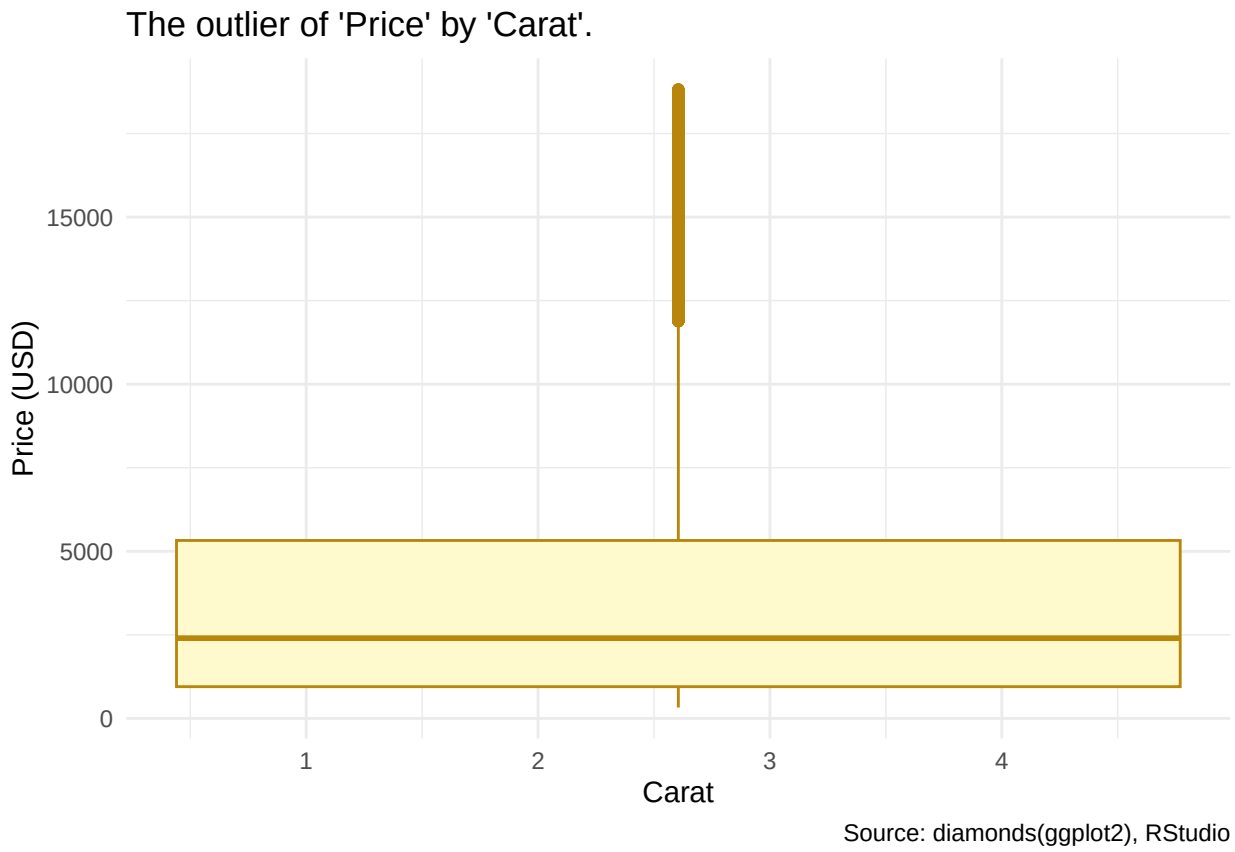
To verify diamond quality, we need to know the factor of 'carat', 'cut', and 'color' are sensible to cost the price or not.
Then, we used the random sampling to got the sample represented a whole diamonds data set.
The factors we need to check some diamonds quality included 'carat', 'color', and 'cut', these factors we expected that they affected to the price.
After we plot the scatter plot by facet grid and add trend line, we noticed that:
- Fair cut: H and J color are not sensible.
- Good cut: I color is not sensible.
- Very good cut: H and J color are not sensible. But G color is ambiguous, we need to discover more data to verify the quality.
- Premium cut: F, H and J color are not sensible.
- Ideal cut: D, H and I color are not sensible. But G color is ambiguous, we need to discover more data to verify the quality.

Boxplot

5. Detect the outlier of diamonds price by carat

```
ggplot(diamonds, aes(carat, price, group='price')) +
  geom_boxplot(fill="lemonchiffon",col="darkgoldenrod") +
  theme_minimal() +
  labs(title = "The outlier of 'Price' by 'Carat'.",
       x = "Carat",
```

```
      y = "Price (USD)",
      caption = "Source: diamonds(ggplot2), RStudio")
```



This boxplot tell us that a bunch of outliers is above the body of box (around 12,000 USD by approximately). And median is around 2,500 USD by approximately that it is near the minimum value.
If we plot this chart into histogram chart to see dispersion, it will be "Right skewed distribution".

Conclusion

There are many factors that affects to diamonds price such as carat, cut, color, clarity, width, length, and depth. But the outliers of this data set allow us to verify other questions like the quality of diamonds. When we try to plot some scatter charts to see the relationship between carat and price group by color and cut, some trend lines curving are unreasonable.

For instance, most of trend lines are similar like exponential graph and the line is close to 2.5 carat. But some trend lines are downward when carat is more than 2.5 and the price is cheaper than other data points. That means some diamonds flawed by cut, color, clarity, and etc. which affects to the price and reflect to the outliers.