Advanced Course in Marketing 06 Descriptive statistics and data visualization

Takumi Tagashira

Email: takumi.tagashira@r.hit-u.ac.jp

Data summary

- This session focuses on data summary and visualization.
- We will use a package called "ggplot2" for visualization.
  - Included in tidyverse
- Download idpos_customer.csv data via manaba.

```
library(tidyverse)

idpos_cust <- readr::read_csv("data/idpos_customer.csv")
```

Data type and summary

- Quantitative variable
  - Sensible to calculation
  - E.g. Ratio scale (Interval scale)

- Categorical variable
  - Identification and classification of objects
  - Not sensible calculation
  - e.g. Nominal scale, ordinal scale
- We should carefully select methods to summarize data based the variable types

Quantitative variable and summary

- We often use descriptive statistics for quantitative variables.
  - Cross tabulation tables are often used for categorical variables
- Various measurements of descriptive statistics can be shown with summary() function.

```
summary(idpos_cust)
```

Interpretation of Summary() results

- Frequency
  - minimum: 1, maximum: 31, mean: 2.017
- Monetary
  - minimum: 1,503, maximum: 305,665, mean: 16622
- Recency
  - minimum: 1, maximum: 31, mean: 13.95

- R reports DS of other variables automatically, but there are not that informative.
  - It is important to ensure the variable types before the analysis.
- For qualitative variables, DS is informative.
  - We can check whether illogical or unnatural observations are included.

Descriptive statistics: Numerical approach

- We can also numerically summary the data.
- Calculating a value that represents the distribution.
  - Median
  - Mean
  - Variance

Median

- Median is a middle value separating the greater and lesser halves of a dataset.
- e.g. In a dataset [1, 3, 2, 5, 4], the median is 3.
- Put them in order: 1, 2, 3, 4, 5 and the middle number (median) is 3.

Mean

- The most useful representative value.
- Mean of variable x ($x_1, \dots, x_n$)can be defined as:
- $\bar{x} = \frac{1}{n} \sum_i x_i$
- Mean represents center of distribution.
- Sum of deviation from mean is 0.
- $\sum_i (x_i - \bar{x}) = 0$

- Is mean everything?

Is mean everything?

- Suppose the followings are test scores in a class:
  - Mathematics: (3, 3, 5, 5, 5, 5, 5, 7, 7)
  - Japanese: (2, 3, 3, 5, 5, 5, 7, 7, 8)
  - Let the above vectors are the scores of Maths (x) and Japanese (y) exams among 9 students (Full marks: 10).
- The mean values of both exams are 5.
  - $\bar{x}$= 19(3+3+5+5+5+5+5+7+7)=5
  - $\bar{y}$= 19(2+3+3+5+5+5+7+7+8)=5
- x distributes from 3 to 7.
- y distributes from 2 to 8.

Measures of dispersion

- Measurements shows how ‘spread out’ a set of data is.
  - Variance
  - Standard deviation
- Degree of ‘spread out’: Larger deviation from mean represents more ‘spread out’ situation.
- Lower deviation means that observations are distributed near mean.

Variance

- Deviation from mean: $x_i - \bar{x}$
- Sum of deviation from mean is 0.
  - → Positive and negative values off-set each other.
- Why don't we take quadratic value of deviation? That value increases as the sample size increase.
  - Dividing the value by sample size n (or n-1).

- $S^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$
  - S^2: Variance of x
- Variance is squared value and the unit of variance is different from original value. Square root of variance ($\sqrt{\cdot}$) is called standard deviation.

Descriptive statistics in R

- Specifying the first column of dataset X.
  - Median: median(X[,1])
  - Mean: mean(X[,1])
  - Variance: var(X[,1])
    - R returns the calculation of variance with $\frac{1}{(n-1)}\sum_i (x_i - \bar{x})^2$.
    - This definition is called unbiased sample variance and has statistically desirable features.
  - Standard deviation: sd(X[,1])
- When specifying with a variable name (variablename) in the dataset X:
  - X$variablename
  - E.g., Mean: mean(X$variablename)

```
mean(idpos_cust$monetary)
[1] 16622.26
```

Summary of categorical variables

- Descriptive statistics are not informative for categorical variables.
  - We often use frequency of occurrence or cross tabulation tables.
- table() is the fundamental function
- with() function provides better presentation.
  - E.g. with(data, table(varname))

```
table(idpos_cust$gender)

with(idpos_cust, table(gender))
```

Cross turbulence table

- Cross tabulation table can be created by specifying two variables in table() function.

```
with(idpos_cust, table(gender,decile_rank))
```

- We can summarize information of the observations that match a certain condition.
  - Checking the gender ratio within the decile rank ten customers.

```
idpos_cust %>%
  filter(decile_rank == 10) %>%
  with(table(gender))
```

Category and quantitative variables

- Summarizing quantitative variables based on a focal categorical variable.
- You have already conducted this procedure.
  - Combination of group_by() and summarize()

```
idpos_cust %>%
  group_by(gender) %>%
  summarize(mon_m = mean(monetary),
            mon_sd = sd(monetary),
            freq_m = mean(frequency),
            freq_sd = sd(frequency))
```

Data visualization

Introducing several graphs

- This session introduces visualization methods by using a package called “ggplot2” that is included in tidyverse.
- Graphs:
    - Histogram
    - Boxplot
    - Violin plot
- We will use diamonds dataset (included in ggplot2).

```
head(diamonds)
```

ggplot() function

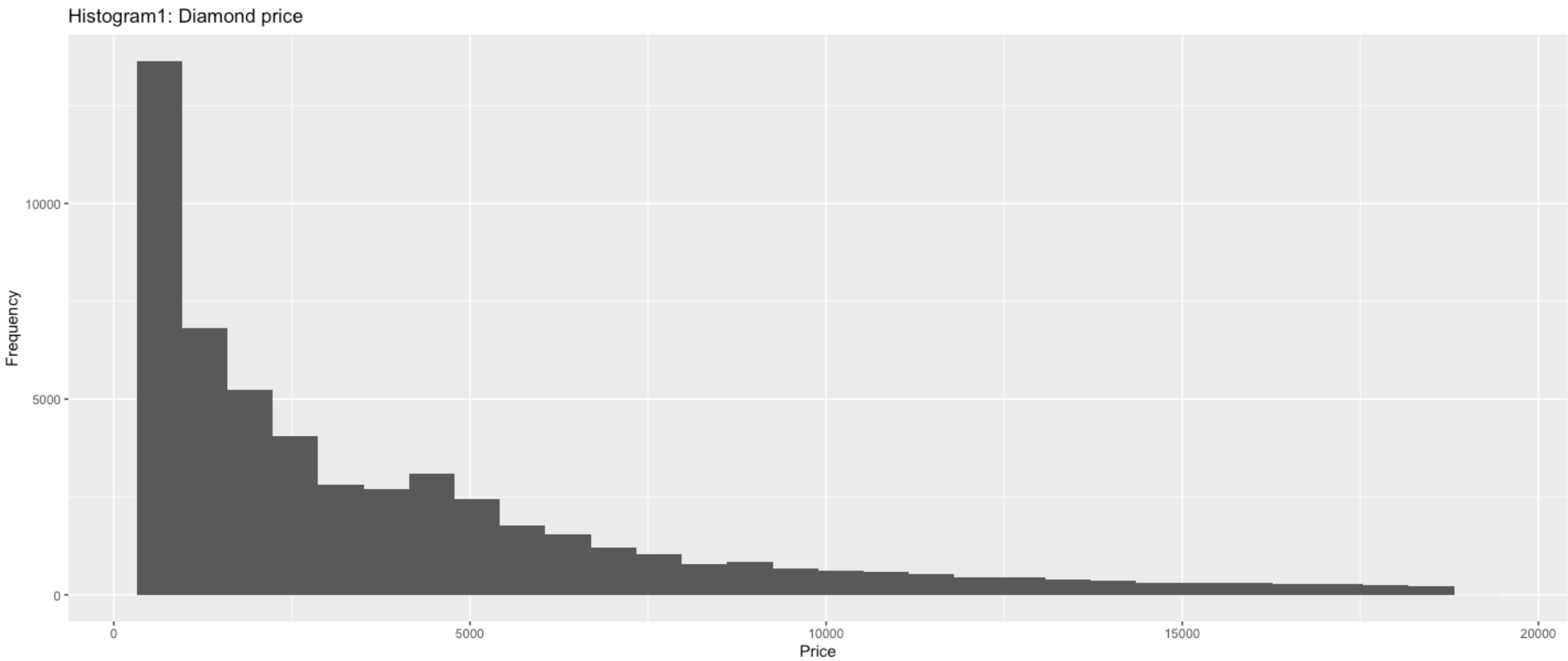
- ggplot() in ggplot2 is the fundamental function.
- ggplot()arguments:
  1. data:referred dataframe
  2. mapping: Arguments specifying the relationship between variables used and the figure to be displayed.
     - aes() function (aesthetics) connects variables and plotting factors within ggplot commands.
- Adding a layer to an object created by ggplot() function.
- Graphical layers are specified as geom_... (geometry).
  - E.g. geom_point(): scatter plot、geom_histogram():histogram

Histogram

- Visualizing the distribution of data in a discretized form.
  - A continuous variable is divided into several classes
- Since only one variable is used, thus aes should specify only x.
- geom_histogram() function is used.
- Visualizing the frequency of the price range of diamonds.

```
p1 <- ggplot(diamonds, mapping = aes(x = price))
p1 + geom_histogram() +
  labs(x = "Price", y = "Frequency",
       title = "Histogram1: Diamond price")
```

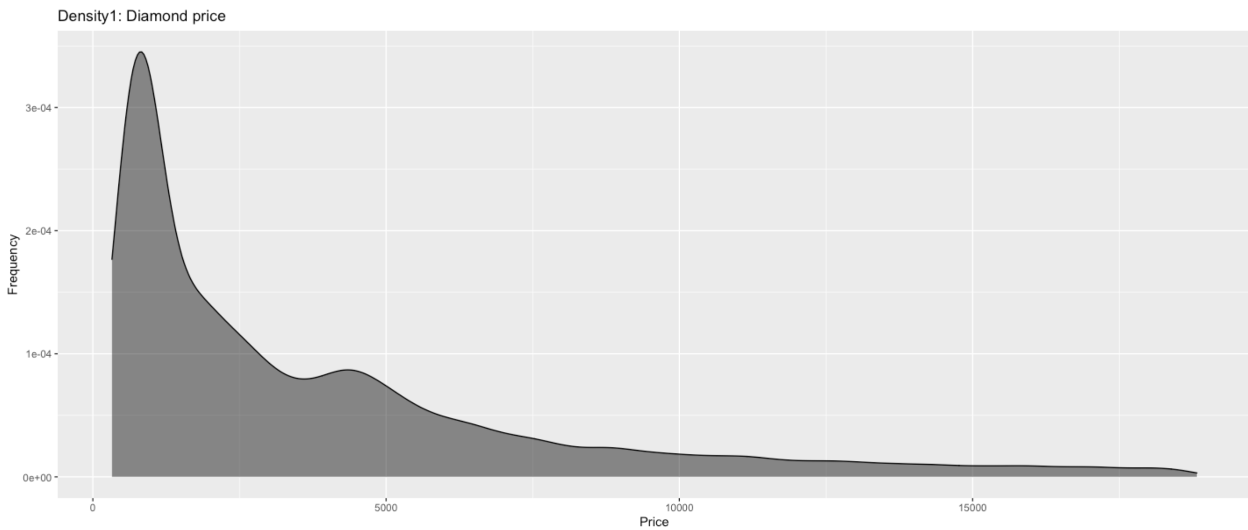
Example: histogram



Density plot

- geom_density() can show a density.
- “fill” argument specifies the color to paint.

```
p1 + geom_density(fill = "black", alpha = 0.5) +
  labs(x = "Price", y = "Frequency",
       title = "Density1: Diamond price")
```

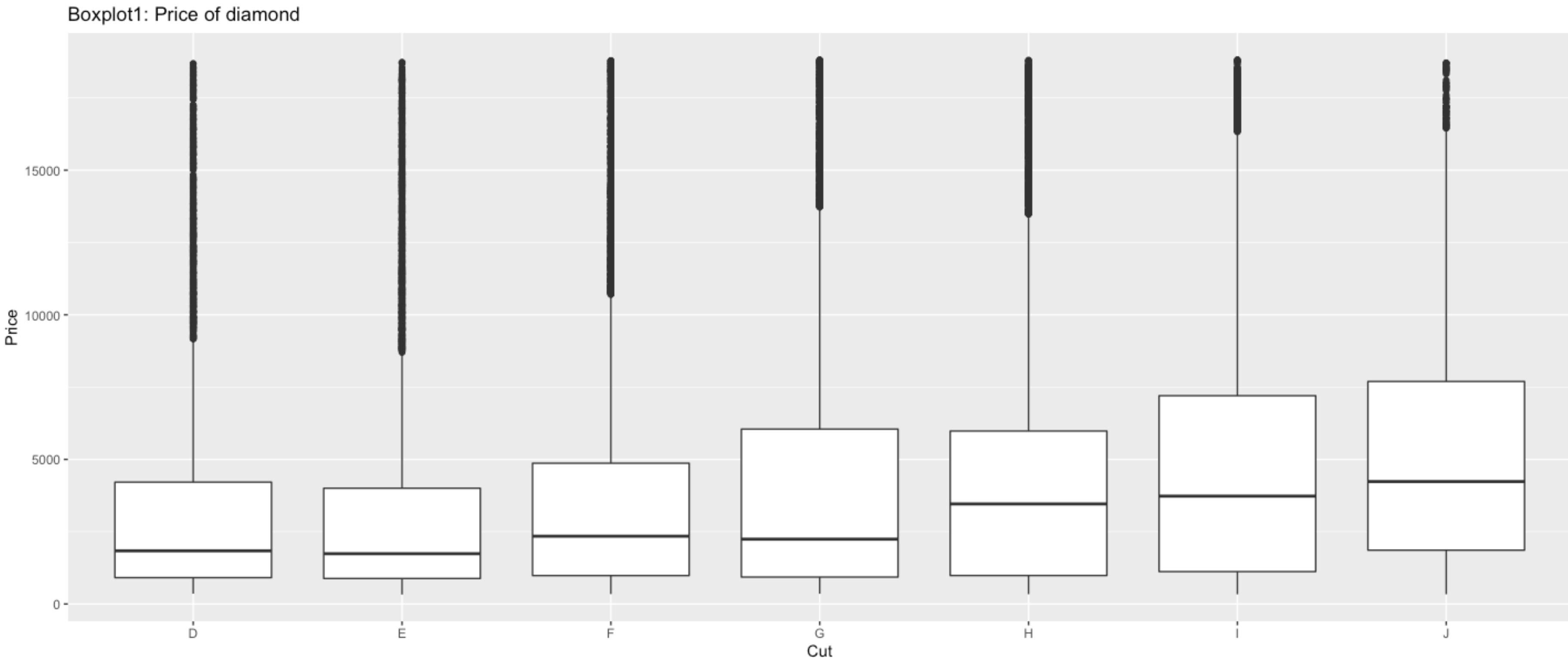


Boxplot

- Boxplot is a graphical representation of quartiles, quartile ranges.
- Quartiles represent the values that divide data into four equal part.
  - First quartile: Q1, Second quartile: Q2, Third quartile : Q3, Maximum value: Q4
  - Quartile range: A range between from Q3 to Q1.
- Checking price distribution for each cut quality (Fair, Good, Very Good, Premium, Ideal).

```
p2 <- ggplot(diamonds, mapping = aes(x = color, y = price))
p2 + geom_boxplot() +
  labs(x = "Cut", y = "Price",
       title = "Boxplot1: Price of diamond")
```

Boxplot



Boxplot1: Price of diamond
Price
15000
10000
5000
0
D
E
F
G
H
I
J
Cut

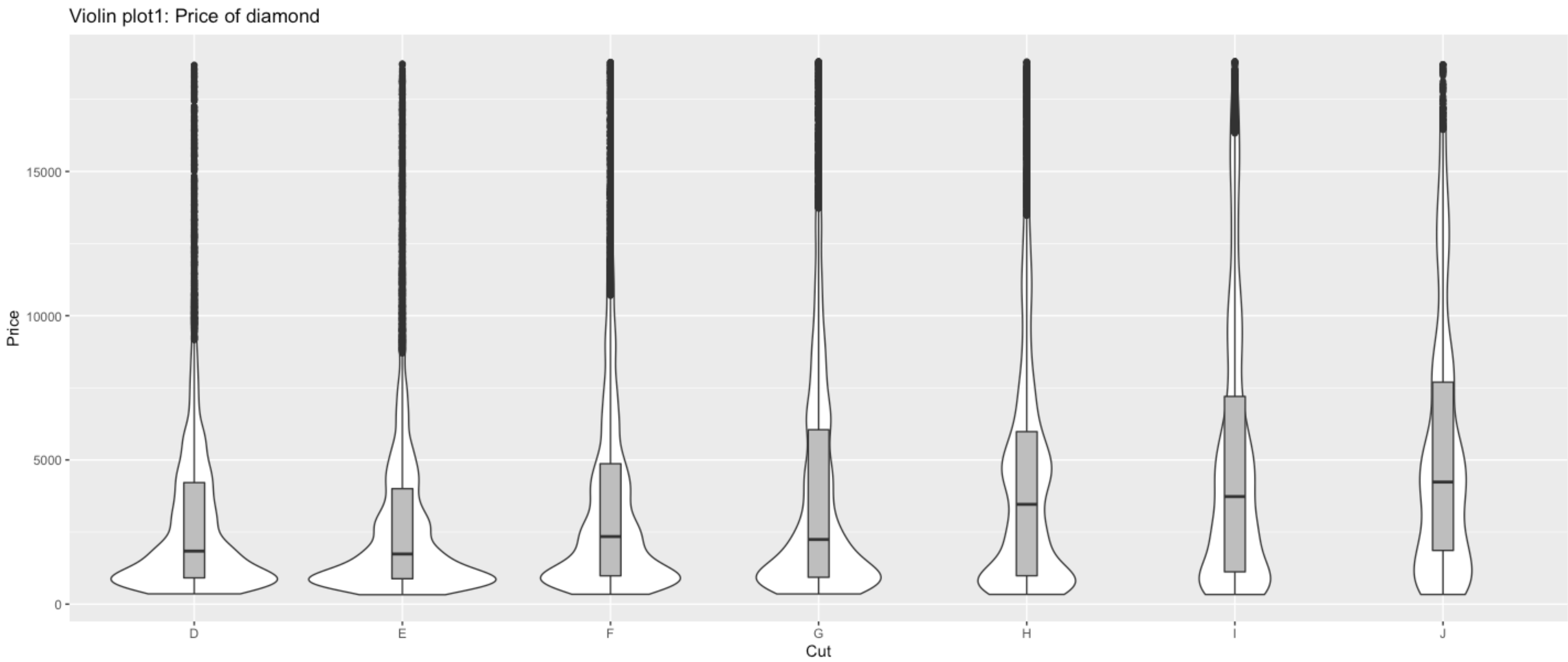
Violin plot

- Violin plots contain more detailed information about the distribution than box-plot.
- Using geom_violin() function
- Width of the plot shows the frequency of occurrence.

- If a plot distributes wide horizontally, it can be interpreted as most data gathered in a narrow area.

```
p2 + geom_violin() +
  geom_boxplot(fill = "gray", width = 0.1) +
  labs(x = "Cut", y = "Price",
       title = "Boxplot1: Price of diamond")
```

Violin plot



Violin plot1: Price of diamond
15000
10000
5000
0
Price
D
E
F
G
H
I
J
Cut

Relationship between two variables

Relationship between two variables

- Graphical approach
  - Scatter plot

- Numerical approach
  - Covariance: $S_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$
  - Correlation: $\rho_{xy} = \frac{S_{xy}}{\sqrt{S_x^2} \cdot \sqrt{S_y^2}}$
    - S_x^2 and S_y^2 are variance of x and y respectively
    - i.e. Standardized value of covariance

Standardization

- Standard deviation represents the scatter of distribution, and its unit is the same as original value.
- Z (standardized value of x) is defined as follows.

- $Z_x = \frac{(x_i - \bar{x})}{\sqrt{S^2}}$

- Standardized value contributes to compare values of different variables.
- Z indicates the relative position of each observation from mean per standard deviation.

Standardization

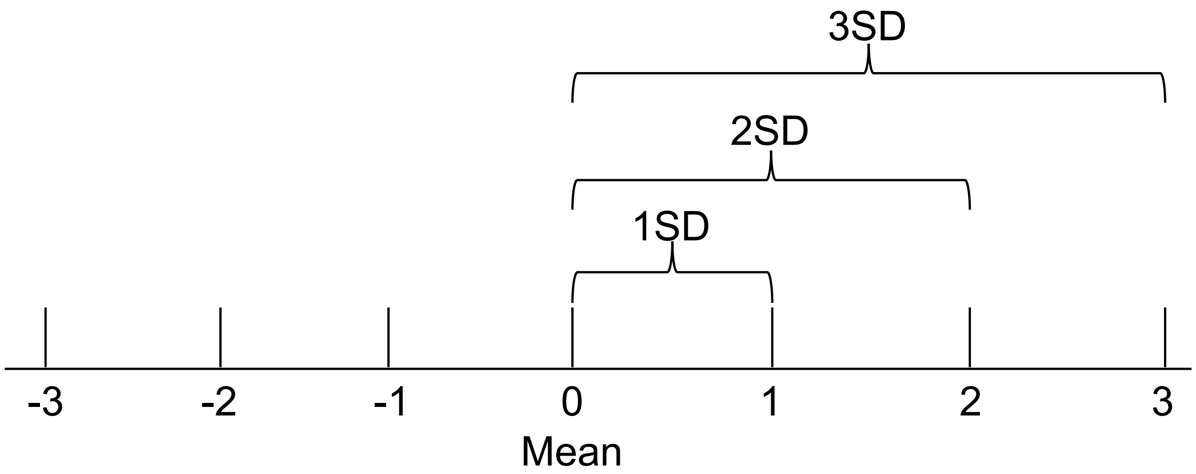
- Standard deviation represents the scatter of data and its unit is the same as original value.
- Z (standardized value of x) is defined as follows.

- $Z_x = \frac{(x_i - \bar{x})}{\sqrt{S^2}}$

- Standardized value contributes to compare values of different variables.
- Mean of Z is 0 and the standard deviation is 1.

Standardization

- Standard scores show relative position of the individuals in the data.



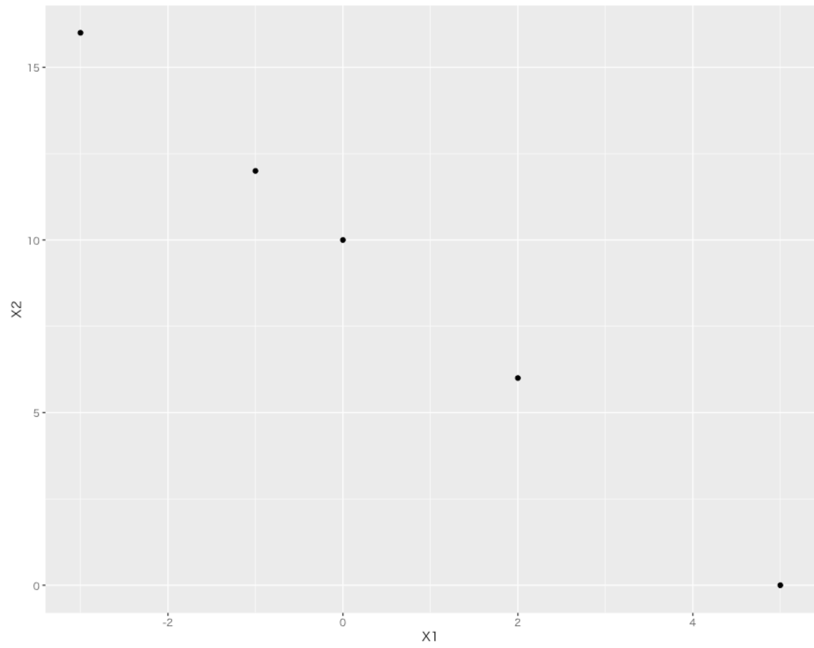
Correlation coefficient

- Interpretation of covariance is not easy:
  →Correlation coefficients (ρ)
- Correlation coefficient ρ takes value between $-1 \leq \rho \leq 1$.
  - Positive correlation ↔ $0 \leq \rho_{xy} \leq 1$
  - Uncorrelated ↔ $\rho_{xy} = 0$
  - Negative correlation ↔ $-1 \leq \rho_{xy} \leq 0$

Correlation and scatter plot

- The following dataset represents a correlation coefficient of -1.
- All the observations follows a linear function $(y = -2x + 10)$.
- Correlation indicates how the observations close to linear function.

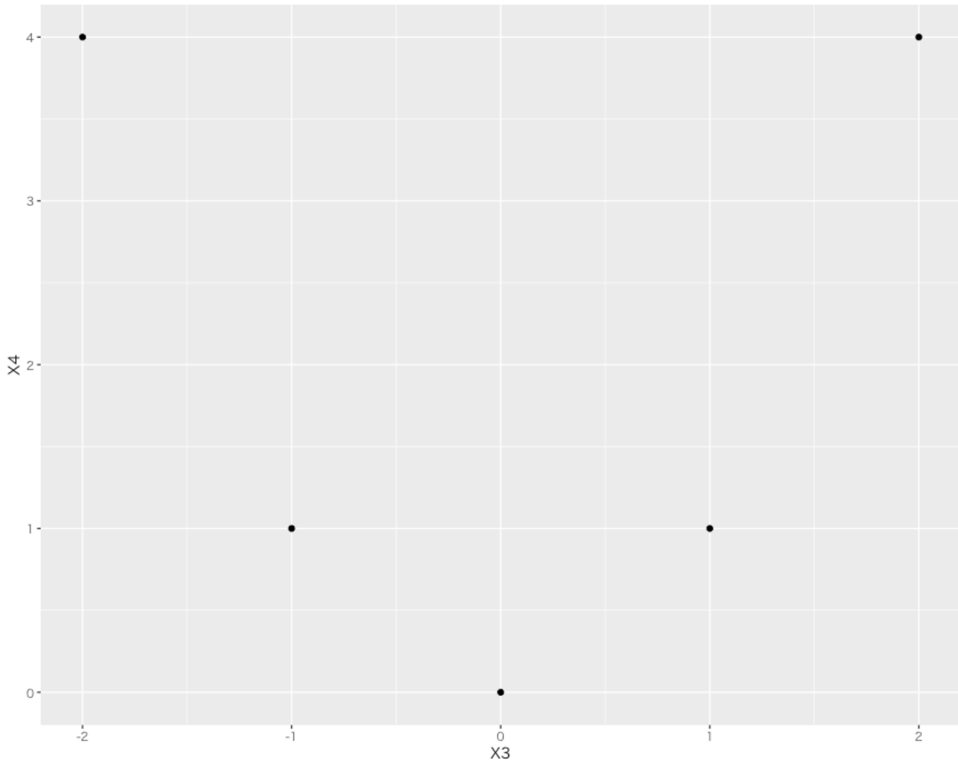
	1	2	3	4	5
x	-3	-1	0	2	5
y	16	12	10	6	0



Correlation and scatter plot

- What's the correlation coefficient of this dataset?

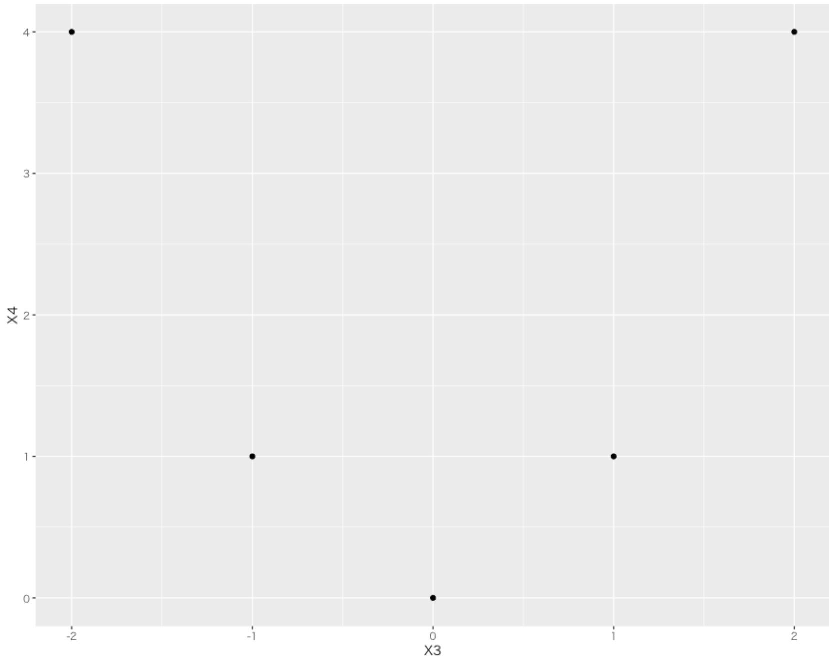
	1	2	3	4	5
x	-2	-1	0	1	2
y	4	1	0	1	4



Correlation and scatter plot

- Following dataset represents: $\rho = 0$
- The correlation equals zero, but can you conclude that there are not inter-related?
  - $y = x^2$

	1	2	3	4	5
x	-2	-1	0	1	2
y	4	1	0	1	4



Summary: correlation

- Correlation coefficient ρ takes value between $-1 \leq \rho \leq 1$.
  - Positive correlation $\leftrightarrow$ $\mathbf{0} \leq \rho_{xy} \leq \mathbf{1}$
  - Uncorrelated $\leftrightarrow$ $\rho_{xy} = \mathbf{0}$
  - Negative correlation $\leftrightarrow$ $-\mathbf{1} \leq \rho_{xy} \leq \mathbf{0}$

- Correlation represents linear relationship between two variables.
- Zero correlation coefficient does not mean there is no interrelation between variables.

R exercise

- Data: diamonds
- Calculating covariance and correlation

```
#Covariance
cov(diamonds$carat, diamonds$price)
[1] 1742.765

#Correlation
cor(diamonds$carat, diamonds$price)
[1] 0.9215913
```

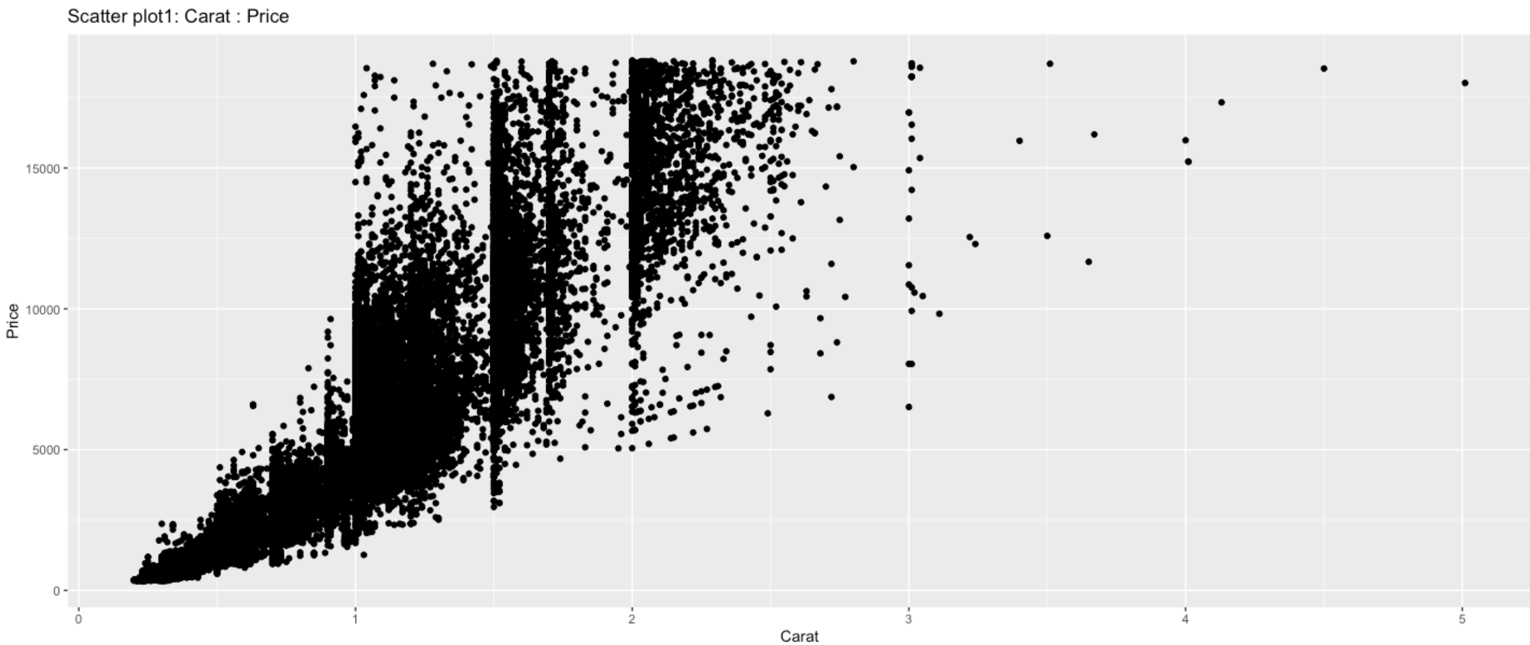
- Correlation coefficient shows these two variables have strong linear relationship.


0.15ct
0.20ct
0.25ct
0.30ct
0.35ct
0.40ct
0.50ct
0.60ct
0.70ct
1.00ct

Scatter plot

- Visualizing carat vs. price

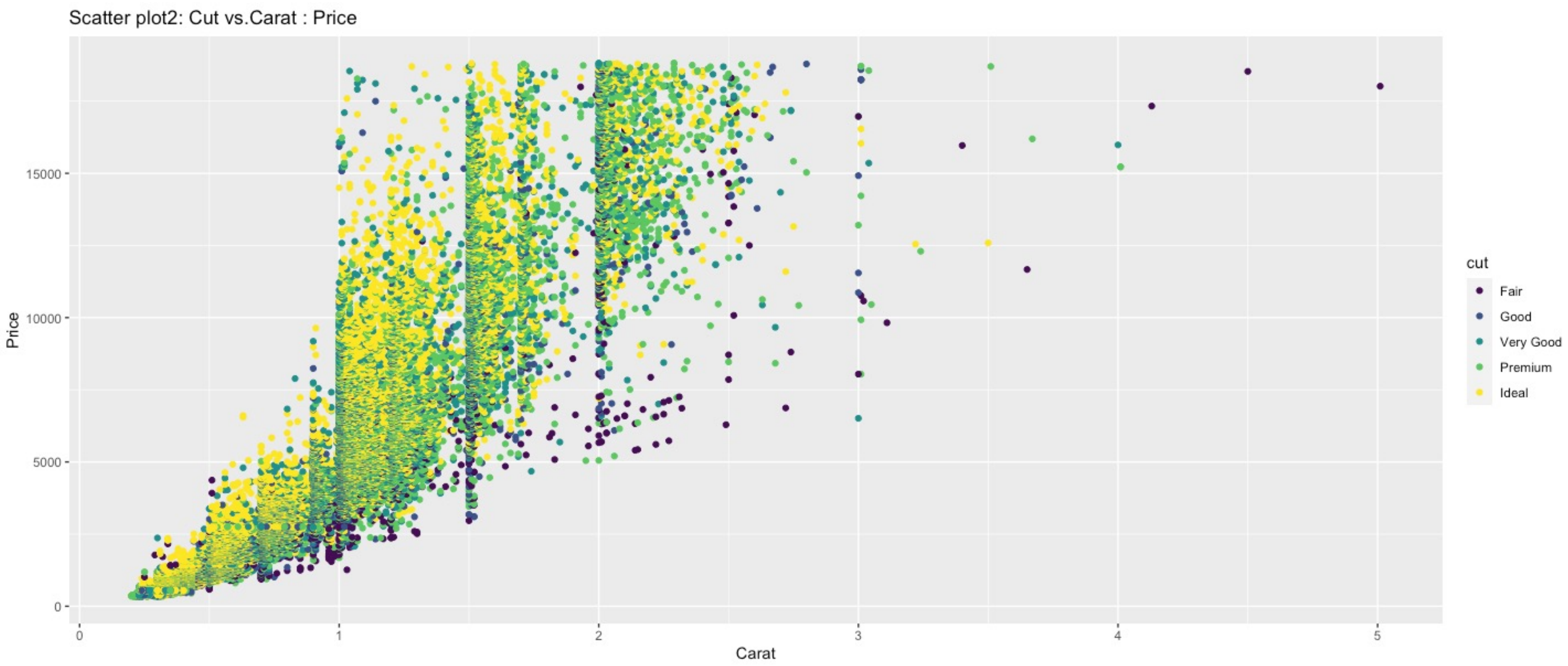
```
p3 <- ggplot(diamonds, mapping = aes(x = carat, y = price))
p3 + geom_point() +
  labs(x = "Carat", y = "Price",
       title = "Scatter plot1: Carat : Price")
```



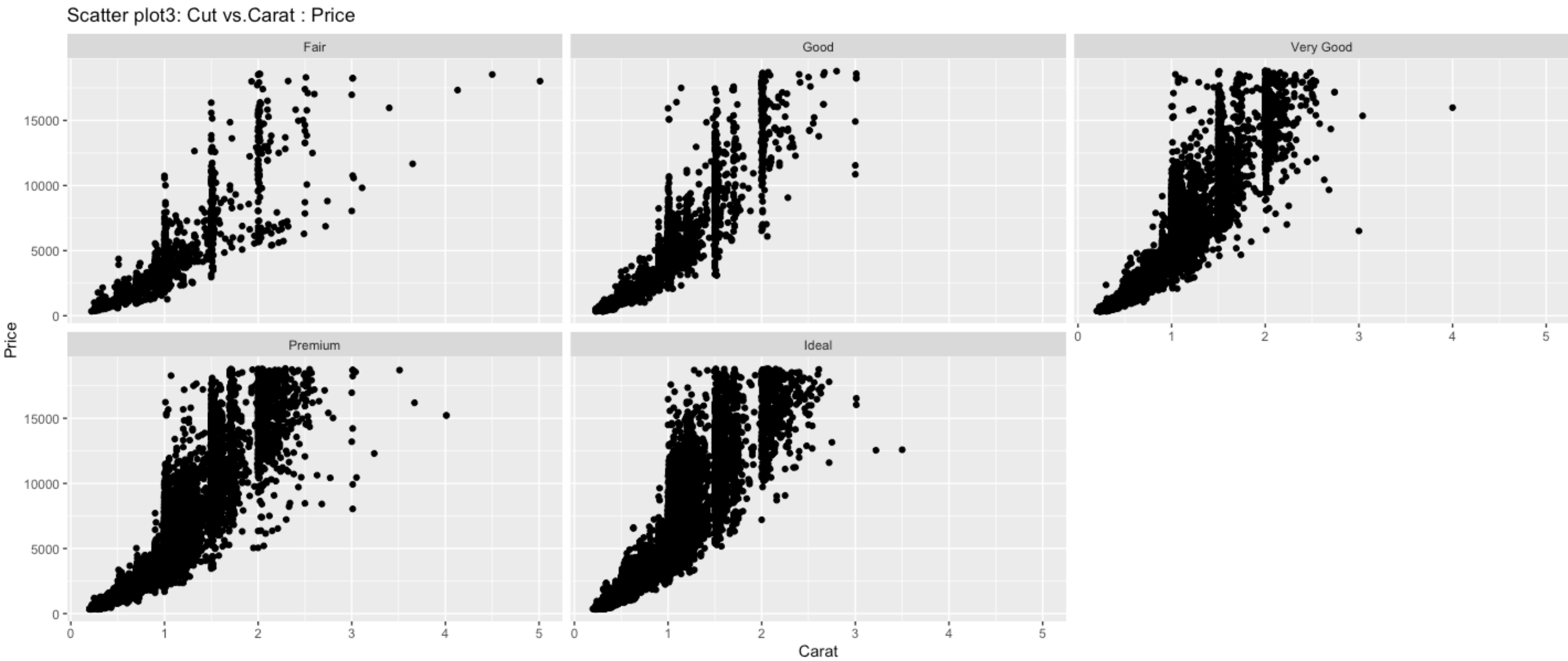
Categorical variable and scatter plot

- Summarizing relationship between two numerical variables and a categorical variable.
  - E.g. Does the relationship between carat and price depend on cut quality?
- Approach:
  1. Painting different colors for each category within a same plot
     - Using the “color = categ_var” argument in “Mapping = aes()”
  2. Drawing different plots for each category
     - Using facet_grid() or facet_wrap()

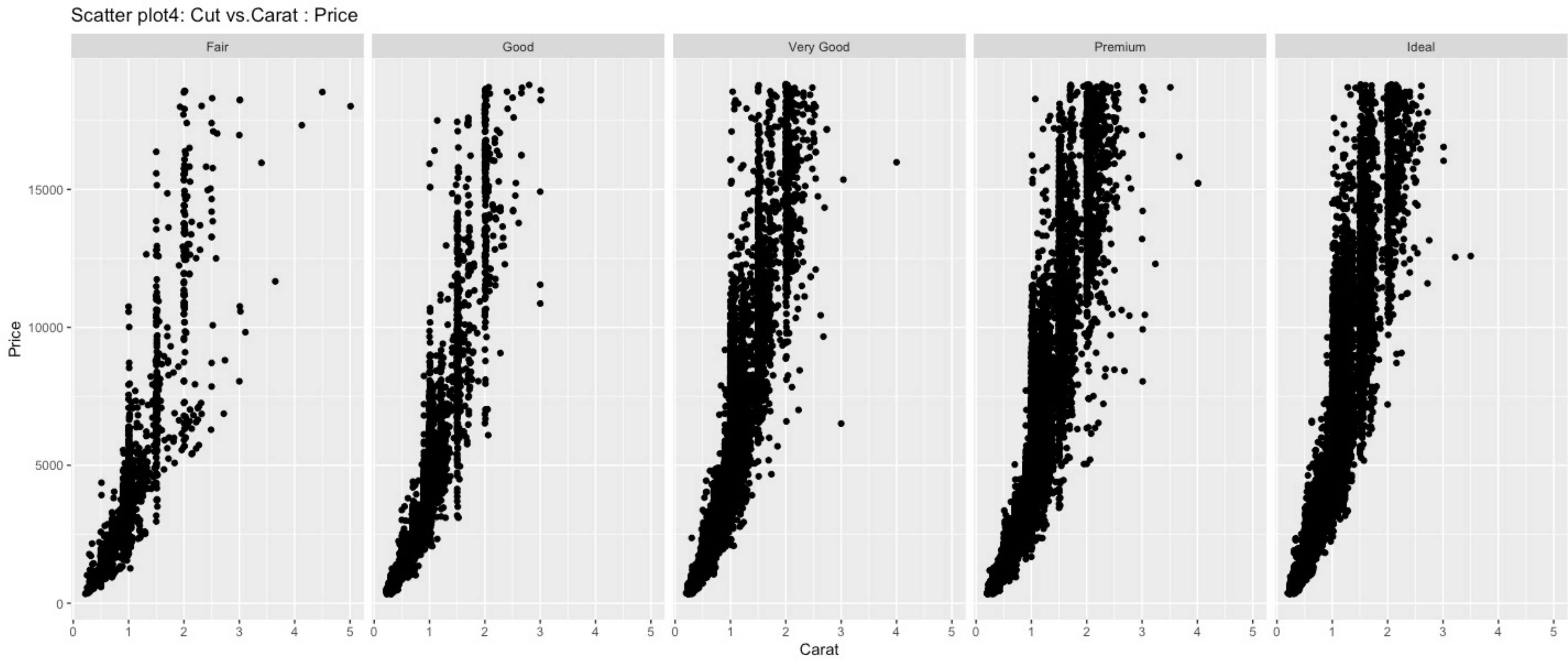
```r
p4 <- ggplot(diamonds, mapping = aes(x = carat, y = price, color = cut))
p4 + geom_point() +
  labs(x = "Carat", y = "Price", color = "cut",
       title = "Scatter plot2: Cut vs.Carat : Price")
```



```
p3 + geom_point() + facet_wrap(~cut) +
  labs(x = "Carat", y = "Price",
       title = "Scatter plot3: Cut vs.Carat : Price")
```



```
p3 + geom_point() + facet_grid(~ cut) +
  labs(x = "Carat", y = "Price",
       title = "Scatter plot4: Cut vs.Carat : Price")
```



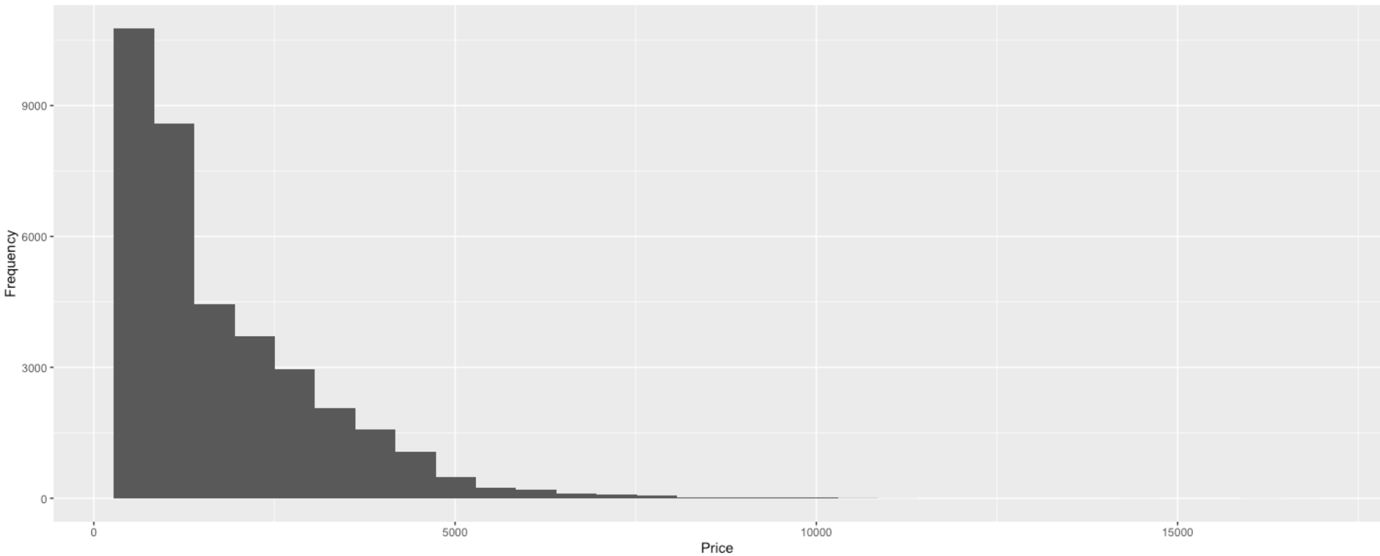
Pipes and ggplot2

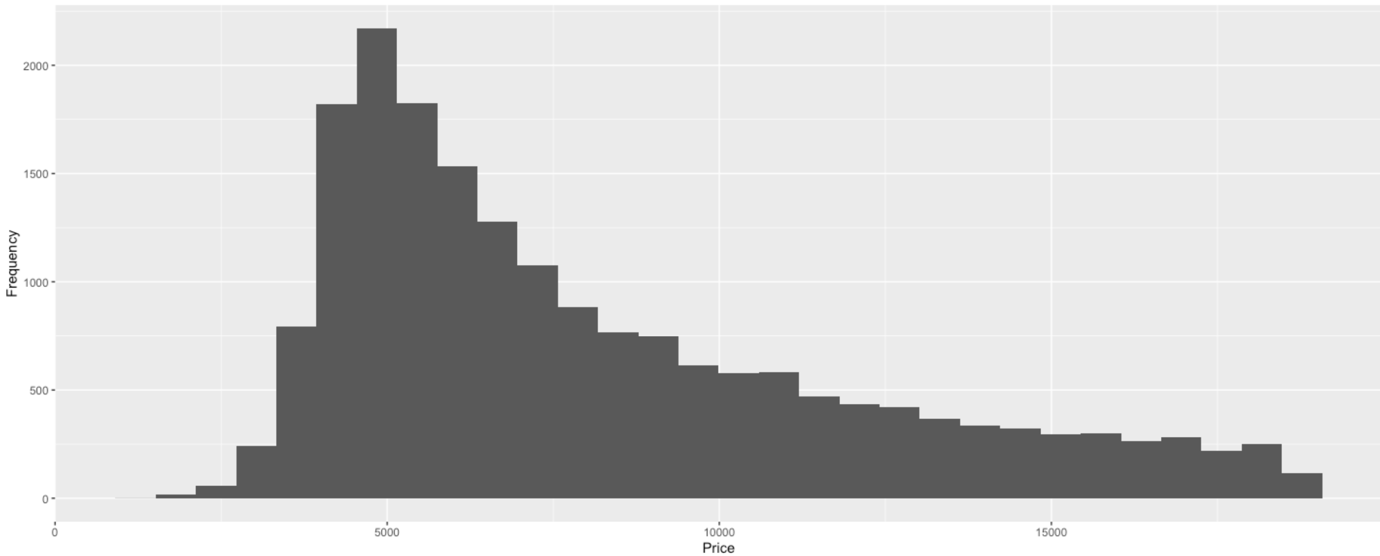
- Combination of pipe operators and ggplot2
  - E.g. Dividing observations into two groups:
  - 1.00 carat or more
  - 1.00 carat or less

```
p5 <- diamonds %>%
  filter(carat >= 1.0) %>%
  ggplot(mapping = aes(x = price))
p5 + geom_histogram() +
  labs(x = "Price", y = "Frequency",
       title = "Histogram:1.00 carat/more")

p6 <- diamonds %>%
  filter(carat <= 1.0) %>%
  ggplot(mapping = aes(x = price))
p6 + geom_histogram() +
  labs(x = "Price", y = "Frequency",
       title = "Histogram:1.00 carat/less")
```

Histogram:1.00 carat/less





Saving figure

- Save the created figures
- ggsave() function is used
  - Created figures should be defined as objects

```
ggsave(filename = "plot1.pdf",
 plot = plot1, width = 10, height = 5, units = "cm")
```

- You can also save your figures cy clicking on the plots tab.
  - Plots -> Export -> Save as Image/ Save as PDF -> Directory -> File name

Exampled data for exercise

- tips_UTF.csv is a dataset of amounts of expense and tip in a restaurant.
- This dataset includes the following variables:
  - Total expense by a customer （total_bill ）
  - Amount of tip by a customer （tips）
  - gender （sex）
  - if the customer smoke or not （smoker）
  - day （day）
  - time-zone of the visit （time ）
  - number of people as a group （size）

Exercise

- Download “tips_UTF.csv” in your data folder and conduct the following analyses.
  1. Draw histograms of total bill and tips.
  2. Calculate mean values of total bill and tips.
  3. Calculate variances of total bill and tips.
  4. Calculate the covariance and correlation of total bill and tips.
  5. Draw a scatterplot of total bill and tips
- No submission required.