3. Basic Data Analytic Methods Using R

Agus Tri Widodo

Review

The previous chapter presented the six phases of the Data Analytics Lifecycle.

- Phase 1: Discovery
- Phase 2: Data Preparation
- Phase 3: Model Planning
- Phase 4: Model Building
- Phase 5: Communicate Results
- Phase 6: Operationalize

Presentasi Tugas Ke-1 (W1-W3)

- Setiap Peserta Presentasi +/- 10 Menit (termasuk tanya jawab)
- Scope 6 Phase Data Analytic Lifecycle

W4- Materi

Key Concepts : Basic features of R Data exploration and analysis with R Statistical methods for eValuation

- Introduction to R
- Exploratory Data Analysis
- Statistical Methods for Evaluation

3.1 Introduction to R

- R is a programming language and software framework for statistical analysis and graphics.
  a. Available for use under the GNU General Public License, R software and installation instructions can be obtained via the Comprehensive R Archive and Network.
  b. This section provides an overview of the basic functionality of R. In later chapters, this foundation in R is utilized to demonstrate many of the presented analytical techniques.
- Example

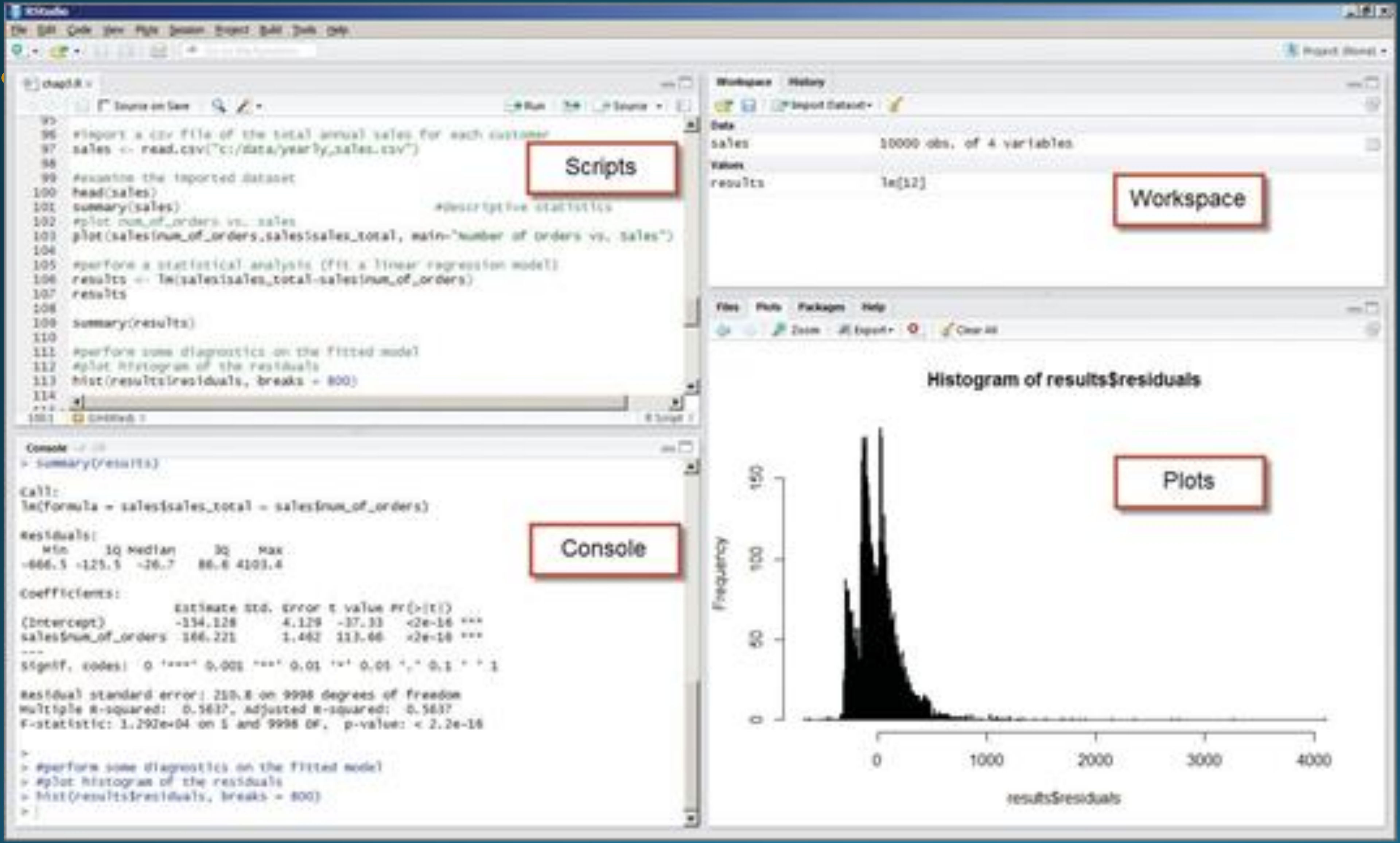
  The read.csv() function is used to import the CSV file. This dataset is stored to the R variable sales using the assignment operator <-.

  a. *# import a CSV file of the total annual sales for each customer sales <- read.csv("c:/data/yearly_sales.csv")*
  b. *# examine the imported dataset head(sales)*
  c. *Plot num_of_orders vs. sales plot(sales$num_of_orders,sales$sales_total, main="Number of Orders vs. Sales")*
  d. *perform a statistical analysis (fit a linear regression model) results <- lm(sales$sales_total ~ sales$num_of_orders) summary(results)*
  e. *perform some diagnostics on the fitted model # plot histogram of the residuals hist(results$residuals, breaks = 800)*

R Graphical User Interfaces

- R software uses a command-line interface (CLI) that is similar to the BASH shell in Linux or the interactive versions of scripting languages such as Python.
- UNIX and Linux users can enter command R at the terminal prompt to use the CLI.
- For Windows installations, R comes with RGui.exe, which provides a basic graphical user interface (GUI).
- However, to improve the ease of writing, executing, and debugging R code, several additional GUIs have been written for R.
- Popular GUIs include the R commander, Rattle, and RStudio. This section presents a brief overview of RStudio, which was used to build the R examples in this book.

Screenshot of the previous R code example executed in RStudio.



Scripts

```
#import a csv file of the total annual sales for each customer
sales <- read.csv("c:/data/yearly_sales.csv")

#examine the imported dataset
head(sales)
summary(sales)                                  #descriptive statistics
#plot num_of_orders vs. sales
plot(sales$num_of_orders,sales$sales_total, main="Number of Orders vs. Sales")

#perform a statistical analysis (fit a linear regression model)
results <- lm(sales$sales_total~sales$num_of_orders)
results

summary(results)

#perform some diagnostics on the fitted model
#plot histogram of the residuals
hist(results$residuals, breaks = 800)
```

Workspace

```
Data
sales                 10000 obs. of 4 variables
Values
results               lm[12]
```

Console

```
> summary(results)

Call:
lm(formula = sales$sales_total ~ sales$num_of_orders)

Residuals:
   Min     1Q Median     3Q    Max
-666.5 -125.5  -26.7   86.6 4103.4

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -154.128      4.129  -37.33   <2e-16 ***
sales$num_of_orders  166.221      1.462  113.66   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 210.8 on 9998 degrees of freedom
Multiple R-squared:  0.5637,	Adjusted R-squared:  0.5637
F-statistic: 1.292e+04 on 1 and 9998 DF,  p-value: < 2.2e-16

>
> #perform some diagnostics on the fitted model
> #plot histogram of the residuals
> hist(results$residuals, breaks = 800)
> 
```

Plots

Data Import and Export

- In the annual retail sales example, the dataset was imported into R using the read.csv() function as in the following code.

```
        sales <- read.csv("c:/data/yearly_sales.csv")
```

- *To simplify the import of multiple files with long path names, the setwd() function can be used to set the working directory for the subsequent import and export operations, as shown in the following R code.*

```
    setwd("c:/data/")
    sales <- read.csv("yearly_sales.csv")
```

Attribute and Data Types

- In the earlier example, the sales variable contained a record for each customer. Several characteristics, such as total annual sales, number of orders, and gender, were provided for each customer.
- Attributes can be categorized into four types: nominal, ordinal, interval, and ratio.
- Numeric, Character, and Logical Data Types

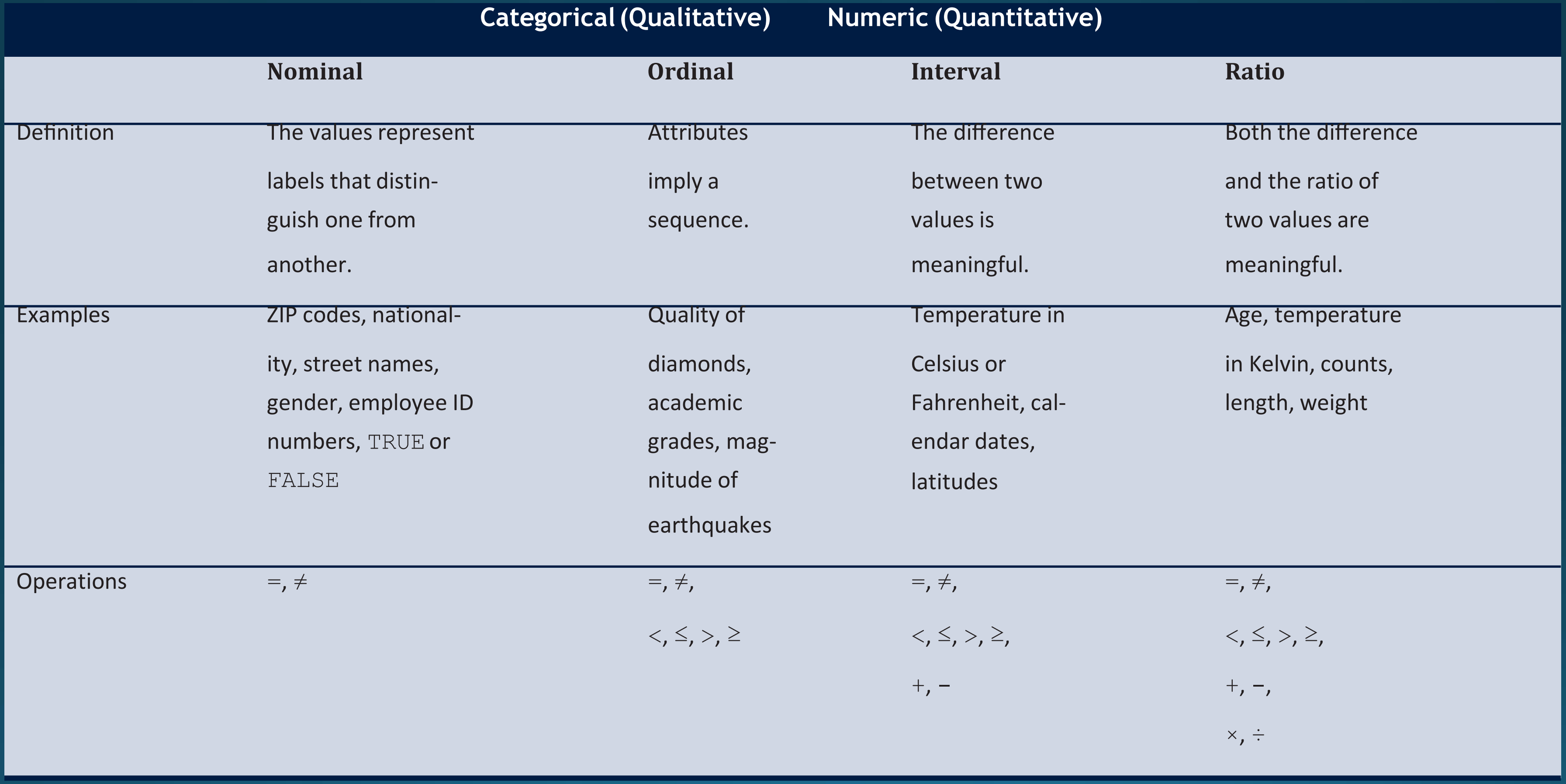
  Like other programming languages, R supports the use of numeric, character, and logical (Boolean) values. Examples of such variables are given in the following R code.

  a. i <- 1 # create a numeric variable

  b. sport <- "football" # create a character variable

  c. flag <- TRUE # create a logical variable

Distinguishes these four attribute types and shows the operations they support



	Categorical (Qualitative)		Numeric (Quantitative)	
	Nominal	**Ordinal**	**Interval**	**Ratio**
Definition	The values represent labels that distinguish one from another.	Attributes imply a sequence.	The difference between two values is meaningful.	Both the difference and the ratio of two values are meaningful.
Examples	ZIP codes, nationality, street names, gender, employee ID numbers, `TRUE` or `FALSE`	Quality of diamonds, academic grades, magnitude of earthquakes	Temperature in Celsius or Fahrenheit, calendar dates, latitudes	Age, temperature in Kelvin, counts, length, weight
Operations	=, ≠	=, ≠, <, ≤, >, ≥	=, ≠, <, ≤, >, ≥, +, −	=, ≠, <, ≤, >, ≥, +, −, ×, ÷

Descriptive Statistics

- It has already been shown that the summary() function provides several descriptive statistics, such as the mean and median, about a variable such as the sales data frame.
- The results now include the counts for the three levels of the spender variable based on the earlier examples involving factors.
- Summary Sales

```
     cust_id          sales_total          num_of_orders          gender          spender
  Min.   :100001    Min.   :  30.02      Min.   : 1.000         F:5035       small :3382
  1st Qu.:102501    1st Qu.:  80.29      1st Qu.: 2.000         M:4965       medium:5469

  Median :105001    Median : 151.65      Median : 2.000                      big   :1149
  Mean   :105001    Mean   : 249.46      Mean   : 2.428
  3rd Qu.:107500    3rd Qu.: 295.50      3rd Qu.: 3.000
  Max.   :110000    Max.   :7606.09      Max.   :22.000
```

3.2 Exploratory Data Analysis

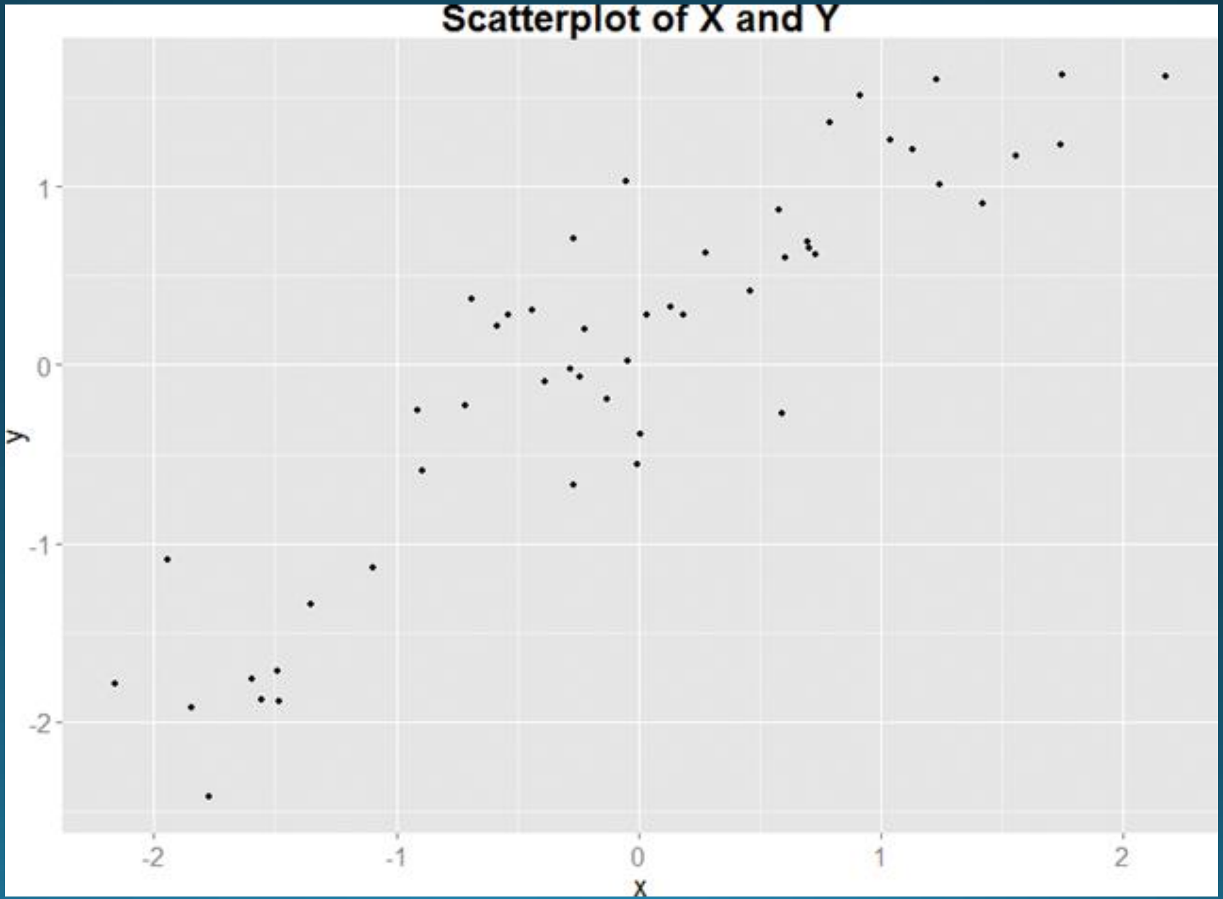
- Visualization Before Analysis
- Dirty Data
- Visualizing a Single Variable
- Examining Multiple Variables
- Data Exploration Versus Presentation

3.2 Exploratory Data Analysis

- Importing and exporting data in R, basic data types and operations, and generating descriptive statistics

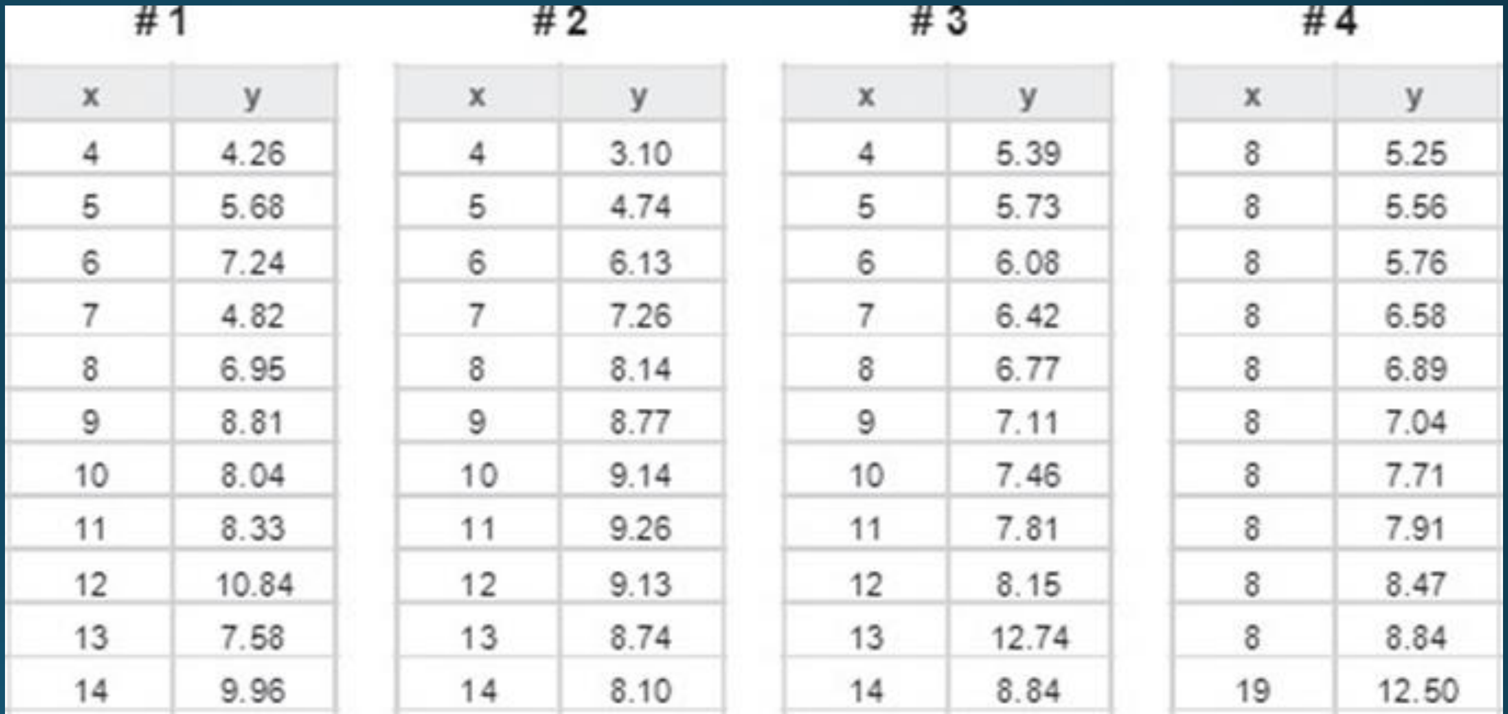
The code to generate data :

```
x <- rnorm(50)
y <- x + rnorm(50, mean=0, sd=0.5) data <-
as.data.frame(cbind(x, y))
```



Visualization Before Analysis

Statistical Property Value
Mean of 9
Variance of y 11
Mean of y 7.50 (to 2 decimal points)



# 1		# 2		# 3		# 4	
x	y	x	y	x	y	x	y
4	4.26	4	3.10	4	5.39	8	5.25
5	5.68	5	4.74	5	5.73	8	5.56
6	7.24	6	6.13	6	6.08	8	5.76
7	4.82	7	7.26	7	6.42	8	6.58
8	6.95	8	8.14	8	6.77	8	6.89
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	7.71
11	8.33	11	9.26	11	7.81	8	7.91
12	10.84	12	9.13	12	8.15	8	8.47
13	7.58	13	8.74	13	12.74	8	8.84
14	9.96	14	8.10	14	8.84	19	12.50

3.3 Statistical Methods for Evaluation

- Hypothesis Testing
- Difference of Means
- Wilcoxon Rank-Sum Test
- Type I and Type II Errors
- Power and Sample Size
- ANOVA

Summary

- R is a popular package and programming language for data exploration, analytics, and visualization.
- As an introduction to R, this chapter covers the R GUI, data I/O, attribute and data types, and descriptive statistics. This chapter also discusses how to use R to perform exploratory data analysis, including the discovery of dirty data, visualization of one or more variables, and customization of visualization for different audiences.
- Finally, the chapter introduces some basic statistical methods.
- The first statistical method presented in the chapter is the hypothesis testing.
- The Student's t-test and Welch's t-test are included as two example hypothesis tests designed for testing the difference of means.
- Other statistical methods and tools presented in this chapter include confidence intervals, Wilcoxon rank-sum test, type I and II errors, effect size, and ANOVA.

Exercises

1. How many levels does fdata contain in the following R code?

```
data = (1,2,2,3,1,2,3,3,1,2,3,3,1)
fdata = factor(data)
```

2. Two vectors, v1 and v2, are created with the following R code:

```
v1 <- 1
v2 <- 2
```

What are the results of :

a. cbind(v1,v2) and rbind(v1,v2)?

b. rbind(c(v1=v1, v2=v2), c(v1, v2), c(4, 5))