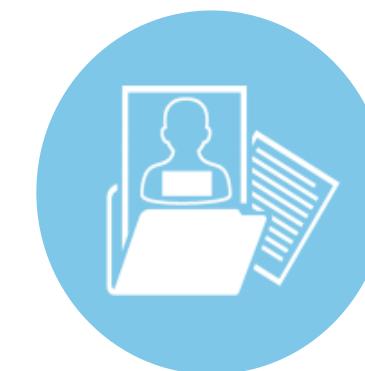


Data Science with R

Lesson 01—Introduction to Business Analytics



After completing
this lesson, you will
be able to:



- Explain the need for Business Analytics
- Discuss different types of Business Analytics
- Discuss Business Analytics case studies
- Explain the importance of Data Science

Some businesses continue to operate in traditional ways.

However, nowadays,

- All aspects of a business collect data and are equipped for data collection.
- The wide availability of data has led to increasing interest in methods to extract useful information from Data Science.
- Organizations in almost every industry focus on exploiting data for competitive advantage.



A business needs to take several decisions, such as:

- What is the possibility that a customer will buy a product P?
- Which should be the next recommended product?
- What is the "realistic" view of opportunity as a customer?
- Do any accounts exist in which there is substantial untapped revenue?



- Which sellers may miss their target orders?
- How to align sellers with customer opportunities to target maximum revenue impact?
- Which factors can influence its product in the marketplace?

A business also needs to take the following decisions:

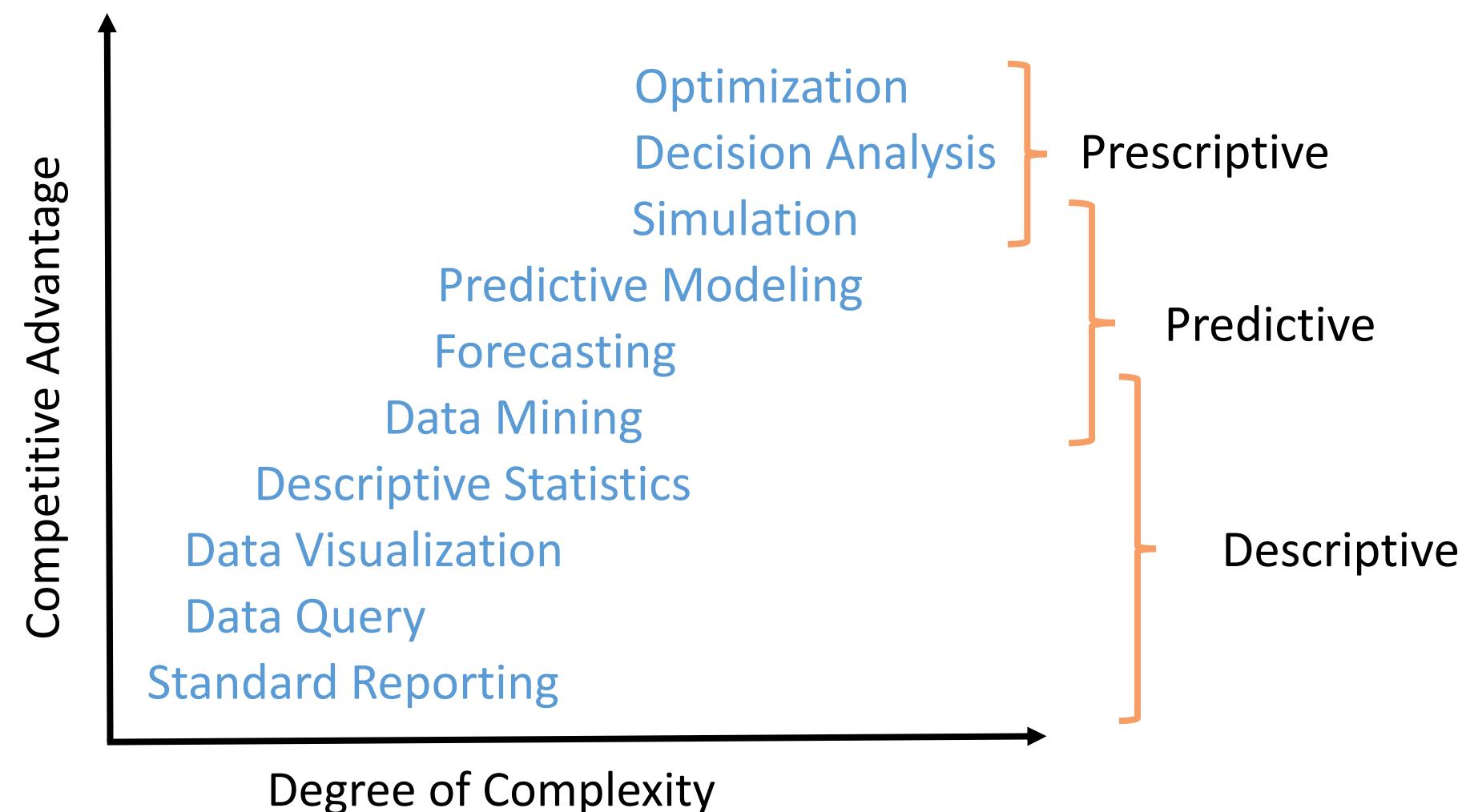
- Which customers are likely to go to its competitors?
- Which type of talent is required to achieve its targets?
- Which business segments are not performing as expected?



- What would be the optimal marketing strategy?
- Which employees can possibly leave the company voluntarily?
- How many employees are required to be hired to achieve its production goals in the next six months?

Business Analytics:

- Is a scientific process that transforms data into insight
- Is used for fact-based or data-driven decision making
- Uses tools such as reports and graphs (simple), optimization, data mining, and simulation (complex)



A few features of Business Analytics are explained below:

Various Methodologies

- Methodologies include concepts from applied probability, applied mathematics, applied statistics, and computer science.
- Data produced helps gain meaningful insights into better business planning and business performance.

Decision Support Systems

- Solutions are primarily used as decision support systems or as their components.
- The solutions help executives, salespeople, and other organizational leaders to make business decisions.

Business Continuity Support

- Business Analytics supports essential business functions.
- It also aids in functions such as hiring, reducing attrition, improving retention, performing staff deployment, and deciding strategy.

In this lesson, you will learn about different types of Business Analytics, which include:

Descriptive Analytics

Health care Analytics

Predictive Analytics

Marketing Analytics

Prescriptive Analytics

Human Resource (HR) Analytics

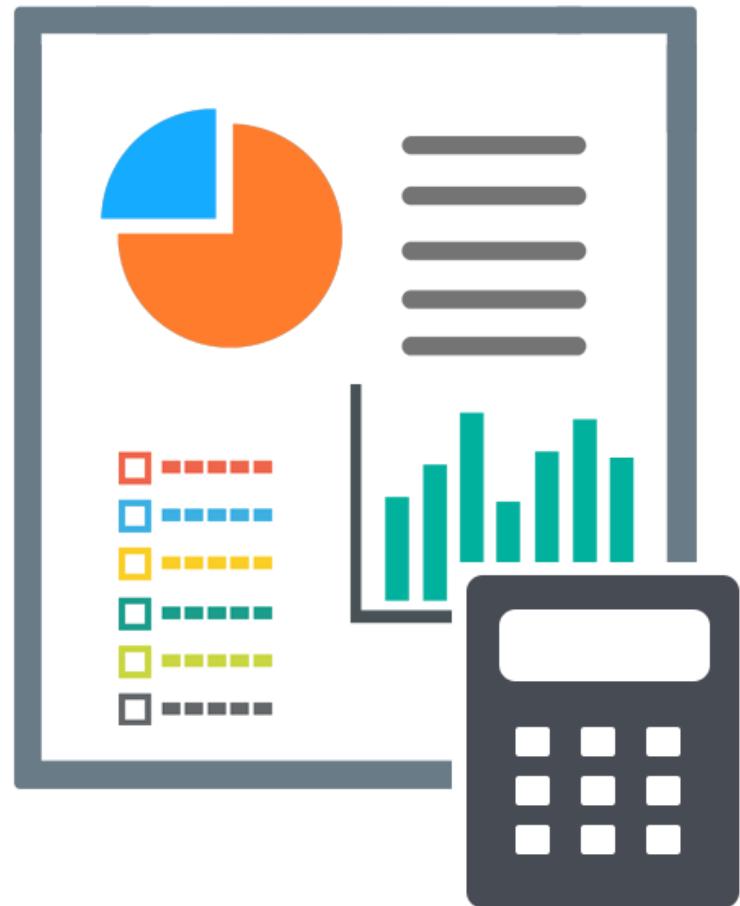
Supply Chain Analytics

Web Analytics



Descriptive Analytics includes techniques that explain what has happened in the past. Examples include:

- Reports
- Data-mining techniques
- Descriptive statistics
- Data queries
- Data dashboards
- Basic what-if spreadsheet models



Predictive Analytics includes techniques that use models created from past data to predict the future or determine the impact of one variable on another. Here are a few examples:

Example 1

A company can create a mathematical model for predicting the future sales by using past data on product sales.

Example 2

A food manufacturing company can estimate the measurement and quantity of unit sales by using the point-of-sale scanner data from retail outlets.

Here are a few more examples of Predictive Analytics:

Example 3

A company can predict the market share of a new product by using the survey data and past purchase behavior.

Other Examples

Companies also use a few other Predictive Analytics techniques, such as time series analysis, linear regression, data-mining techniques, and simulation (risk analysis).

Prescriptive Analytics specifies the best course of action for a business activity in the form of the output of a prescriptive model. The models used in this analytics are called optimization models. They are used in:



Airline Industry

Uses revenue management models and past purchasing data as inputs to get the best pricing strategy across all flights



Finance

Uses portfolio models that utilize historical investment return data to decide the mix of investments

Optimization models are also used in:



Operations

Uses supply network design models that provide the cost-minimizing plant and distribution center locations



Retailing

Uses price markdown models that provide the timing of discount offers and revenue-maximizing discount levels

Analytics tools used in logistics and supply chain management of companies, such as FedEx and UPS, provide benefits like:

- Efficient delivery and optimal sorting of goods
- Vehicle and staff scheduling
- Vehicle routing
- Better inventory and processing control
- Efficient supply chains



Examples:

- **Bernard Chaus, Inc.:** This women's apparel manufacturing company used Descriptive Analytics for presenting its supply chain status to its managers.
- **ConAgra Foods:** This packaged foods company used Prescriptive and Predictive analytics for planning its capacity utilization better by including the inherent uncertainty in commodities pricing.

Analytics in health care is used to:

- Simultaneously control cost and provide more effective treatment to patients
- Improve staff, patient, and facility scheduling
- Control inventory, purchasing, and patient flow



McKinsey Global Institute (MGI) and McKinsey & Company¹⁴ Study

By better utilizing analytics, the United States health care system could save more than \$300 billion each year, which is equivalent to the complete gross domestic product of countries, such as Singapore, Finland, and Ireland.

A better understanding of customer behavior by using data generated from social media and scanner data leads to:

- Better use of advertising budgets
- Effective pricing strategies
- Improved demand forecasting
- Better product line management
- Increased customer satisfaction and loyalty



Automobile Manufacturer Chrysler Team with J. D. Power and Associates

They developed predictive models for supporting their pricing decisions for automobiles as they helped Chrysler to better understand the ramifications of proposed pricing structures. The models generated an estimated annual savings of \$500 million.

The analytics tools used by the Human Resource department ensure that the organization:

- Employs the required skill sets to meet its needs
- Hires the highest quality talent
- Provides an employee-retention environment
- Achieves its organizational diversity objectives

Sears Holding Corporation (SHC)



It created an HR analytics team inside its corporate HR function that used descriptive and predictive analytics to support employee hiring and tracking and influencing retention.

Web Analytics:

- Is the analysis of online activity, such as visiting websites and social media sites (LinkedIn and Facebook)
- Has huge implications for promoting and selling products and services through the Internet

Points to Remember:



- Large companies apply descriptive and advanced analytics to data collected in online experiments to determine the best way to position ads, configure websites, and use social networking sites to promote products and services.
- Online experimentation includes exposing different subgroups to various website versions and then tracking the results.
- These experiments prove to be ineffective as they allow the company to use the trial-and-error method.

Consider an example from a 2004 *New York Times* story:



Introduction



Study by Wal-Mart



Conclusion

The Atlantic coast of Florida was in danger when Hurricane Frances was barreling across the Caribbean and threatening a direct hit.

Consider an example from a 2004 *New York Times* story:



Introduction



Study by Wal-Mart



Conclusion

In Bentonville, Arkansas, the executives at Wal-Mart Stores:

- Considered the situation as an opportunity to apply predictive technology
- Provided forecasts based on:
 - Shopping patterns of customers when Hurricane Charley had struck several weeks earlier
 - Huge amounts of shopper history stored in the Wal-Mart's data warehouse

Consider an example from a 2004 *New York Times* story:



Introduction



Study by Wal-Mart



Conclusion

Wal-Mart Stores could:

- Predict what was going to happen
- Identify the local demand for products
- Anticipate rush stock to the stores much before the hurricane's landfall
- Determine that the stores would also require certain products, not only the usual flashlights
- Observe an unusual increase in the sales of strawberry Pop-Tarts
- Observe that beer was the top-selling, pre-hurricane item

Consider an example from the 1990s:



Introduction



Predictive Modeling



Solution

Earlier, Credit Cards essentially had uniform pricing because:

- The companies did not have appropriate information systems to deal with differential pricing on a large scale.
- The bank management believed that customers would not accept price discrimination.

Consider an example from the 1990s:



Introduction



Predictive Modeling



Solution

Richard Fairbanks and Nigel Morris:

- Approached big banks to offer predictive modeling consulting, finally getting Signet Bank, a regional Virginia bank, to agree
- Convinced the bank that modeling profitability was the right strategy to get ahead

However, the bank had data only to model profitability for the:

- Terms they had offered in the past
- Types of customers who were actually offered credit

Consider an example from the 1990s:



Introduction



Predictive Modeling



Solution

The Signet Bank:

- Experimented by offering different terms to different customers at random, which increased the number of bad accounts
- Changed to about 6% charge-offs from the prevailing charge-off
- Worked toward building predictive models from the data, evaluating them, and then deploying them to increase profits despite continuous losses
- Eventually turned around the credit card portfolio into its most profitable operation

Why is decision making in any business very important? It is because it helps:

- Achieve high revenue
- Lower costs
- Reduce expenses



For increasing revenue while keeping selling expenses the same, Business Analytics suggests changes in the salesforce deployment.



If a company's employees are leaving voluntarily for higher paying jobs and there are high onboarding costs, Business Analytics recommends raises that can help retain existing employees.

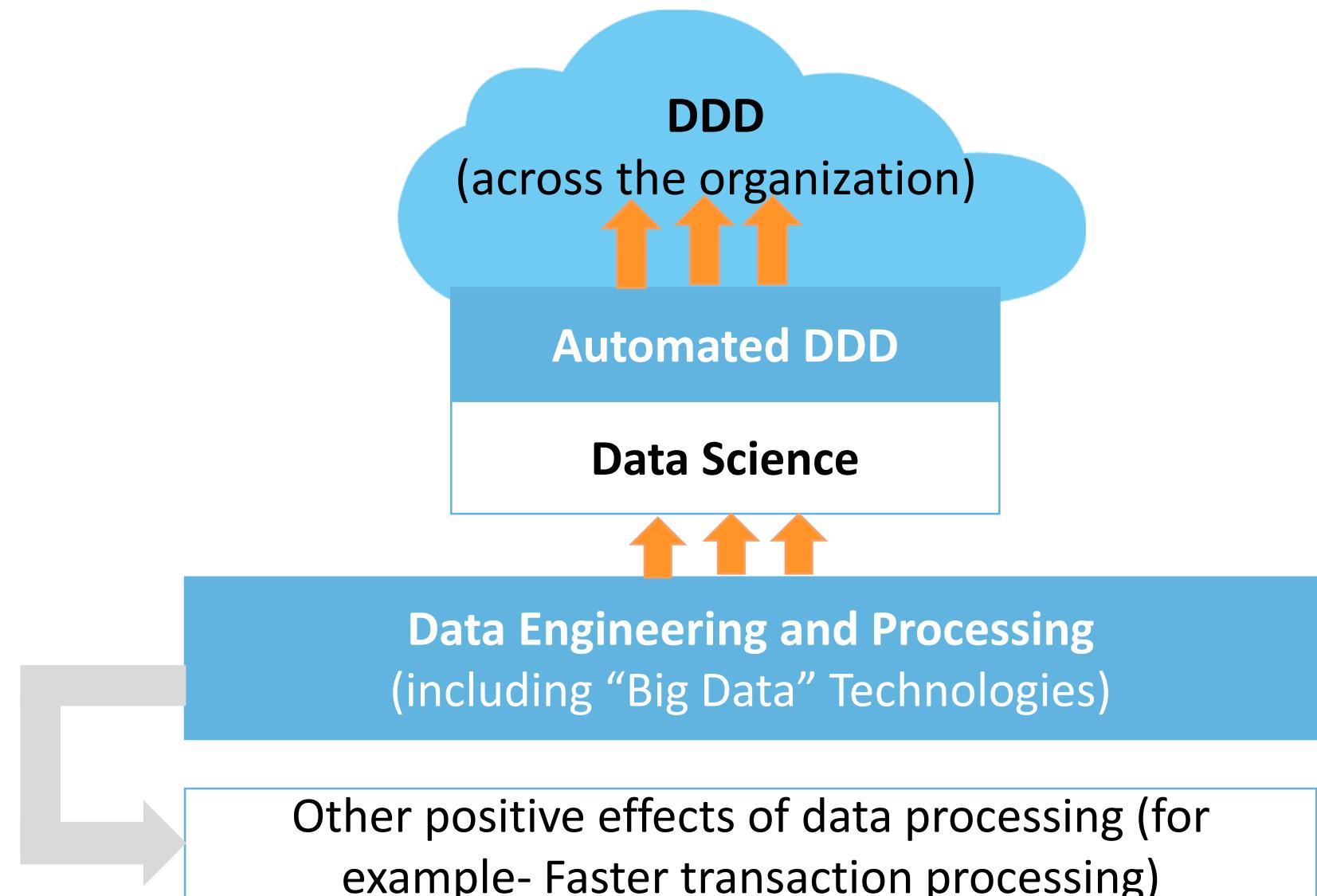
BI:

- Refers to reporting and analytics tools; traditionally used for determining historical data trends
- Includes a set of processes, methodologies, technologies, and architectures that uses the output of the information management processes
- Enables queries by which users can get results by asking data-related questions
- Includes tools that are designed to present the results of analytics in an understandable format



Data Science:

- Includes processes, principles, and methods to understand phenomena through automated data analysis
- Allows Data-Driven Decision Making (DDD), which determines the productivity of an organization



You can consider Data Science a business asset that is about data investment. An understanding of its fundamental concepts is required for data scientists and anyone:

- Working with data scientists
- Employing data scientists
- Investing in data-heavy ventures
- Directing an analytics application in an organization

Points to Remember:



- DDD and big data technologies improve business performance substantially.
- Data Science supports DDD and, at times, performs such decision making automatically.

With respect to Data Science, the key strategic assets are:

- Data
- Ability to extract useful information from data



Points to Remember:



- Many businesses use data analytics to get value from the existing data, and to generally consider if the business has the right analytical talent.
- Incorrect data and unsuitable Data Science talent results in issues such as poor decisions.

Big data typically refers to datasets that are too large for traditional data processing systems. Its technologies are:

- Considered as a popular platform for data analytics and data exploration
- Used to implement data mining techniques and data processing

Nowadays, companies that employ big data technologies have started asking:

Q

What can I now do that I couldn't do before, or do better than I could do before?

Different types of analytical tools available:



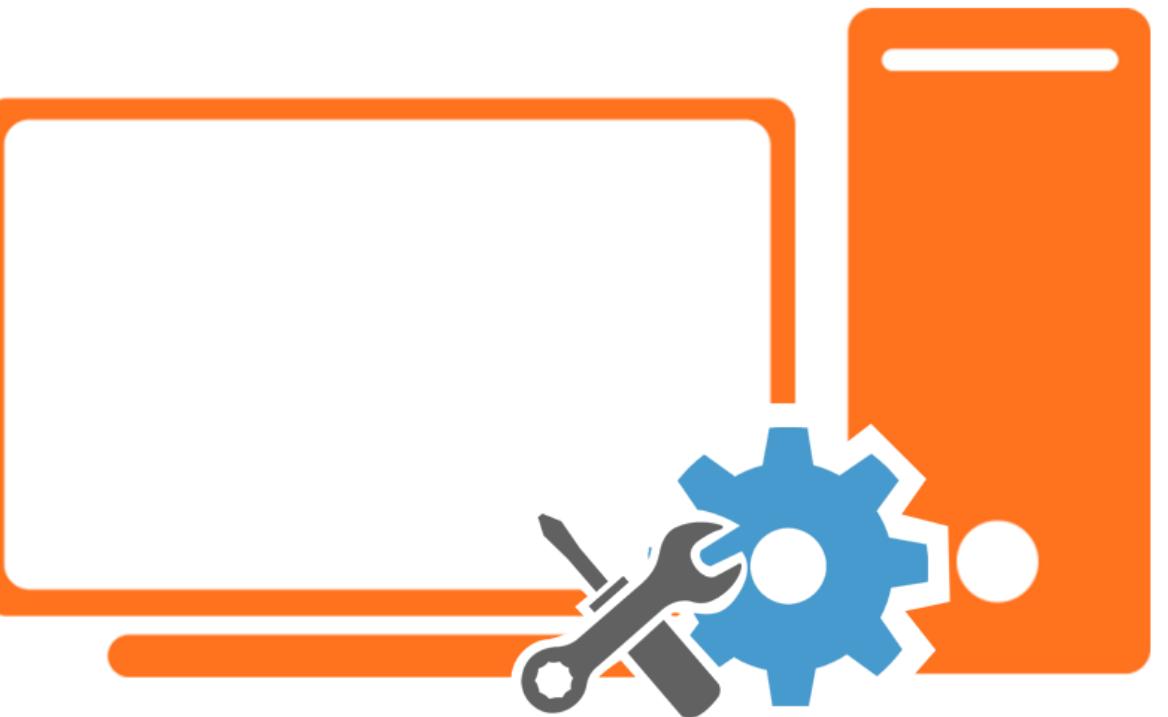
GUI Based: Excel, SPSS, SAS ,Rstudio



Visualization: Tableau, MicroStrategy



Coding Based: SAS, R





QUIZ
1

Which of the following is an example of Prescriptive Analytics? *Select all that apply.*

- a. The airline industry's use of revenue management
- b. The airline industry's use of past purchasing data
- c. The use of portfolio models in finance
- d. The use of price markdown models in retailing



QUIZ
1

Which of the following is an example of Prescriptive Analytics? *Select all that apply.*

- a. The airline industry's use of revenue management
- b. The airline industry's use of past purchasing data
- c. The use of portfolio models in finance
- d. The use of price markdown models in retailing



The correct answer is **a, b, c, and d.**

Explanation: All these examples are of Prescriptive Analytics.

**QUIZ
2**

Which of the following statements is true about Predictive Analytics?

- a. Includes techniques that use models constructed from past data to predict the future
- b. Is defined as the analysis of online activity
- c. Includes techniques that describe what has happened in the past
- d. Helps in supply chain management



QUIZ
2

Which of the following statements is true about Predictive Analytics?

- a. Includes techniques that use models constructed from past data to predict the future
- b. Is defined as the analysis of online activity
- c. Includes techniques that describe what has happened in the past
- d. Helps in supply chain management



The correct answer is **a**.

Explanation: Predictive Analytics includes techniques that use models constructed from past data to predict the future, or ascertain the impact of one variable on another.

QUIZ
3

Which of the following tools or applications is used for data analytics? *Select all that apply.*

- a. C#
- b. R-Programming
- c. SPSS
- d. SAS



QUIZ
3

Which of the following tools or applications is used for data analytics? *Select all that apply.*

- a. C#
- b. R-Programming
- c. SPSS
- d. SAS



The correct answer is **b, c, and d**.

Explanation: R-Programming, SPSS, and SAS are used for data analytics.

QUIZ
4

Which of the following industries uses Business Analytics? *Select all that apply.*

- a. Logistics
- b. IT
- c. Marketing
- d. Retail



QUIZ
4

Which of the following industries uses Business Analytics? *Select all that apply.*

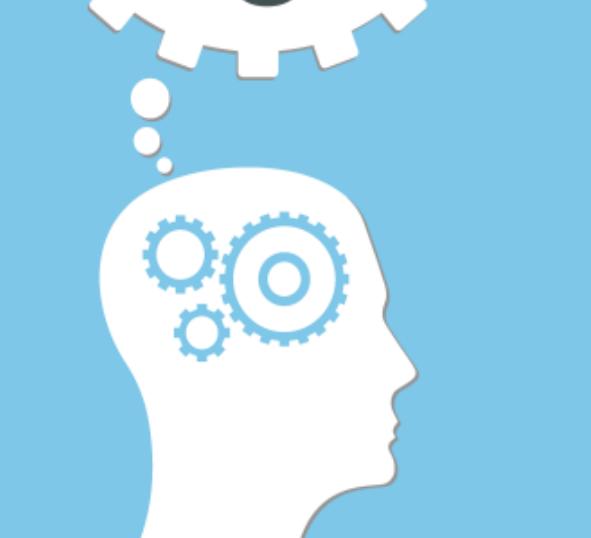
- a. Logistics
- b. IT
- c. Marketing
- d. Retail



The correct answer is **a, c, and d**.

Explanation: Supply chain, marketing, and retail industries use Business Analytics.

Let us summarize the topics covered in this lesson:



- Business Analytics is a scientific process that transforms data into insight.
- Descriptive Analytics includes techniques that explain what has happened in the past.
- Predictive Analytics includes techniques that predict the future by using models created from past data or determine the impact of one variable on another.
- Prescriptive Analytics, the final phase of Business Analytics, specifies the best course of action for a business activity in the form of the output of a prescriptive model. It utilizes the techniques of descriptive and predictive analytics to anticipate the future of a business and to suggest optimal decision options on the basis of such predictions.
- Analytics tools are used in logistics and supply chain management in many companies.

Let us summarize the topics covered in this lesson:



- Analytics in health care is used for purposes such as enabling more effective treatment while controlling costs.
- The analytics tools used by the Human Resource department ensure that the organization has the required skill sets to meet its needs.
- Web Analytics is the analysis of online activity, such as visiting websites and social media sites like LinkedIn and Facebook.
- BI refers to reporting and analytics tools; traditionally utilized for determining the historical data trends.
- Big data implies datasets that are too large for traditional data processing systems.

This concludes “Introduction to Business Analytics.”

The next lesson is “Introduction to R.”

Data Science with R

Lesson 02—Introduction to R



After completing
this lesson, you will
be able to:

- Introduce R
- List the features and drawbacks of R
- Describe the methods to install R
- List the Integrated Development Environments (IDEs) for R
- Explain the various commands in R
- Discuss the R workplace and packages



R is:

- A programming language developed at AT&T Bell Laboratories by Robert Gentleman and Ross Ihaka
- An alternative to S language
- A free, open source language, with highly active community members
- Available across all platforms (Linux, Mac, Windows)

Due to its underlying philosophy and design; R is useful for statistical computation and graphic visualization.



R is a lot more than just a programming language.



Worldwide repository system

R has a worldwide repository system—Comprehensive R Archive Network (CRAN). It can be accessed at <http://cran.r-project.org>.



3,000 packages hosted

As of 2011, there were more than 3,000 such packages hosted on CRAN and many more on other websites.

Do you know there are some limitations of R too?

Steep Learning Curve

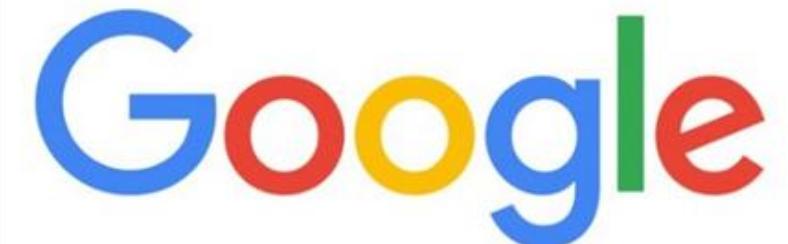
R has a steep learning curve.

Limited RAM Size

R has memory (RAM) limitations when working with large datasets.



R is used by various reputed companies, such as:



Do you want to know more about R?

Here are some pointers about R:

- R statements or commands can be separated by a semicolon (;) or a new line.
- The assignment operator in R is "<-". (Although "=" also works).
- All characters after # are treated as comments.
- There are no multiline or block-level comments.
- The \$ (dollar) operator in R is analogous to a “.” (dot) operator in other languages. The use of the operator is explained below:
 - *Student\$name*
 - *Student\$age*



In this lesson, you will learn about installing R on the following Operating Systems:

Windows

Mac OS X

Linux



For Windows and Mac OS X:

- Simply download a self-installing binary.

For Linux, the installation procedure varies:

- For Debian distribution (including Ubuntu), you can install the R system using its regular package-management tools.
- Since R is open source, you can also compile and install it using source code.



You can install R from the following two sources:

CRAN Website

<http://cran.r-project.org/>

RStudio

<https://www.rstudio.com/products/rstudio/download/>



R is getting updated all the time. Therefore, ensure that you select the most recent version for its installation.

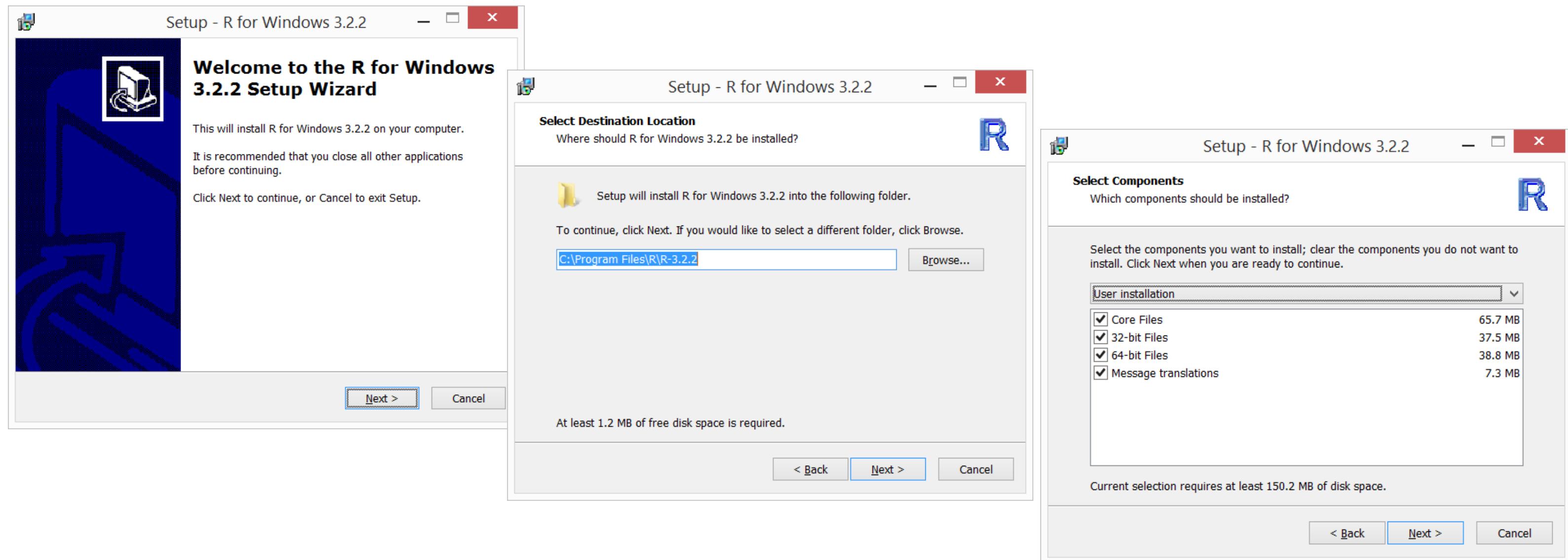
First, you will learn how to install R on Windows from the CRAN website. You can download the package from: <https://cran.r-project.org/bin/windows/base/>.

1. Download R 3.X.X for Windows executable file (.exe).

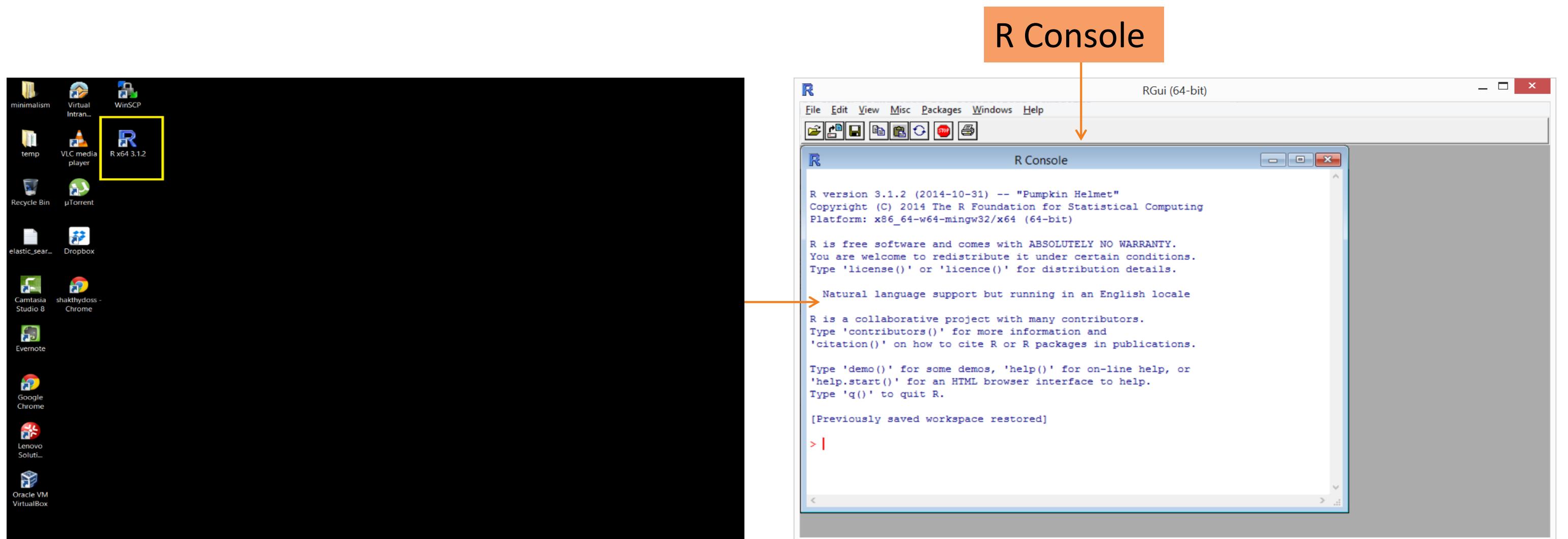


2. Click “Next” in the Setup Wizard.

(You may leave the setting as default.)



3. Finally, open R console (RGUI from your desktop).



The RStudio project currently provides most of the desired features for an IDE in an innovative way, making it easier and more productive to use R. Here are some notable IDEs for R:

Name	Platforms
ESS	Windows, Mac, Linux
Eclipse	Windows, Mac, Linux
SciViews	Windows, Mac, Linux
JGR	Windows, Mac, Linux
Tinn-R	Windows
Notepad++	Windows
Rgui	Windows



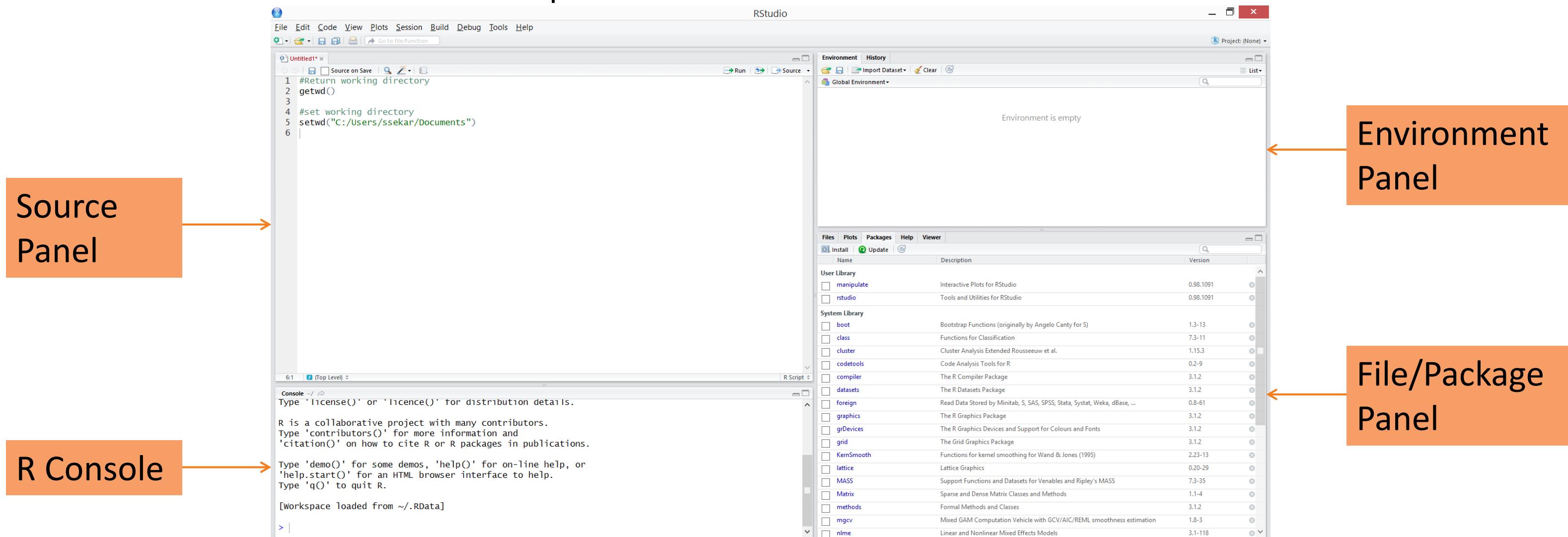
The RStudio program can be run on a desktop or through a web browser. The desktop version is available for Windows, Mac OS X, and Linux platforms.

Installing R on Windows from RStudio Website

Now, you will learn how to install R on Windows from the RStudio website.

1. Download RStudio from <https://www.rstudio.com/products/rstudio/download/>.
2. Run the installation file.
3. Open RStudio.

The different areas of RStudio are depicted below:



When R starts, it undergoes the following process steps:

1

R starts in the working directory, also called the workspace.



2

If present, the .Rprofile file's commands are executed.



3

If present, the .Rdata file is loaded.

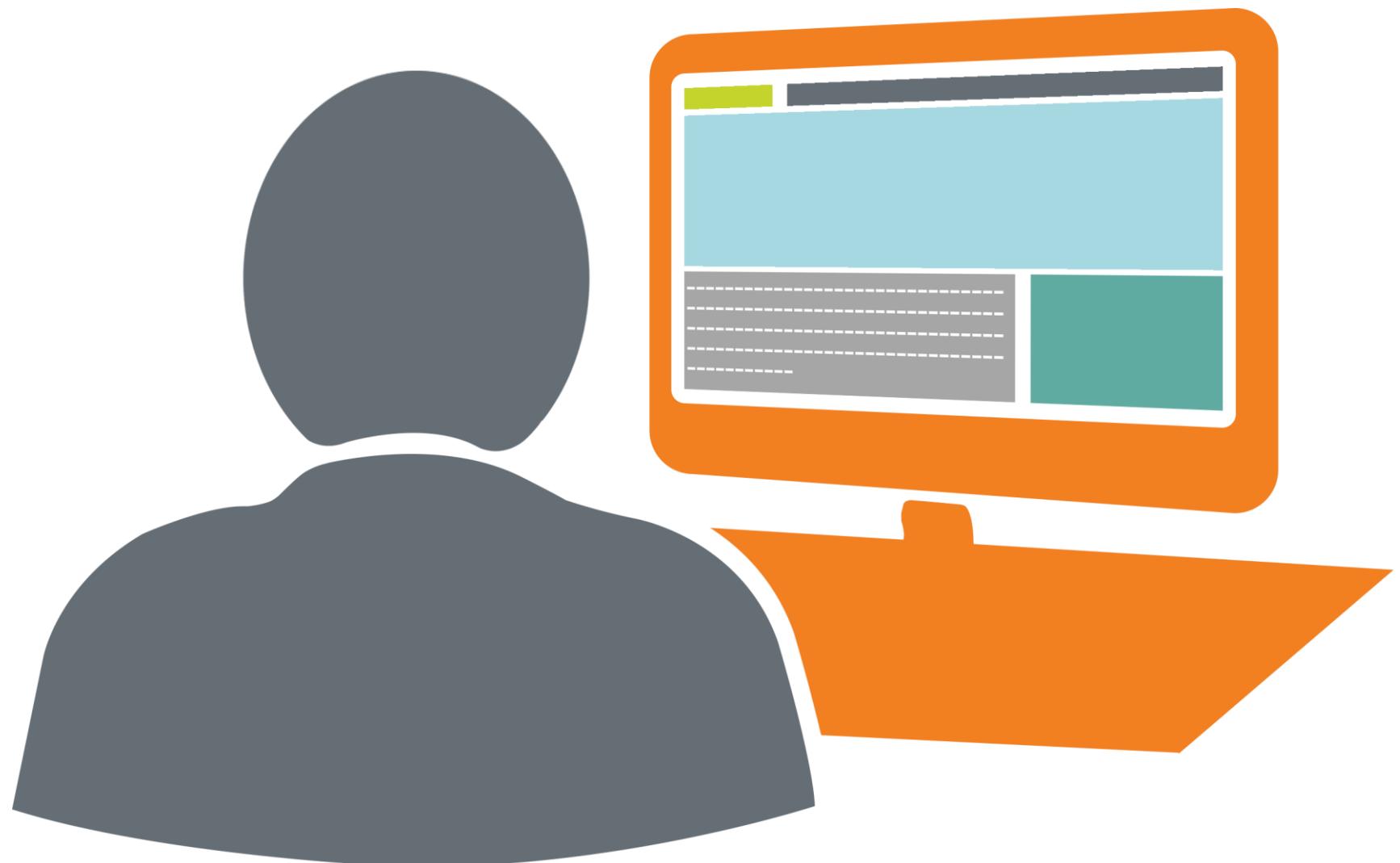
R stores user-defined objects in the workspace.

At the end of an R session, you can save a snapshot of the current workspace. The workspace reloads automatically the next time R starts.

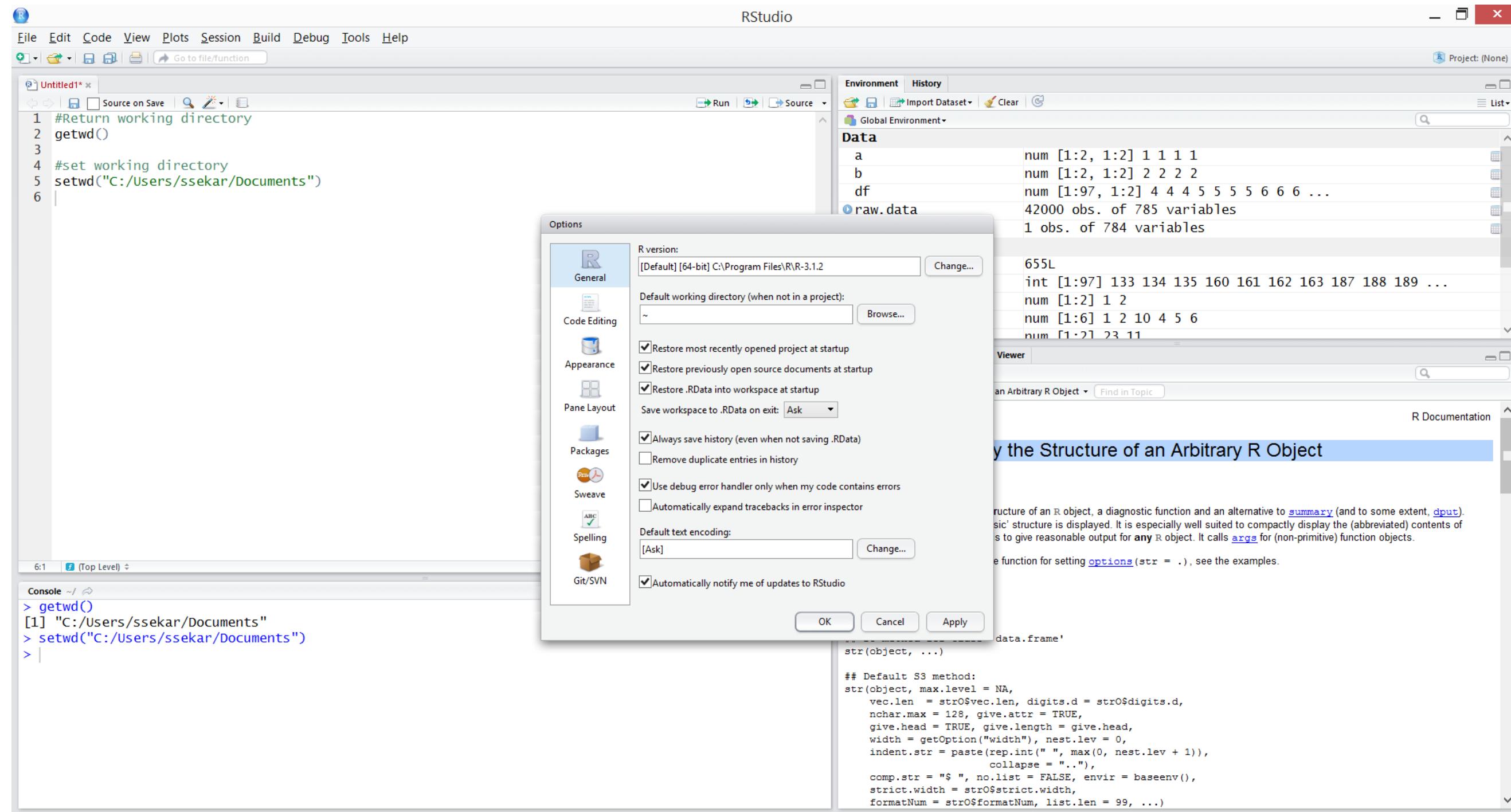


Examples:

- *getwd()* - return working directory
- *setwd()* - set working directory



In RStudio, you can set the workspace by clicking **Tools -> Global Options**.



Understanding Functions of R:

- There are over 1,000 functions at the core of R, and new R functions are created all the time.
- Each R function comes with its own help page. To access a function's help page, type a question mark followed by the function's name in the console.

Getting Help in R:

Here are the commands to get help for some common functions in R:

Help Command	Function
<code>help.start()</code>	# Load HTML help pages into browser
<code>help(package)</code>	# List help page for "package"
<code>?package</code>	# Display short form for "help(package)"
<code>help.search("keyword")</code>	# Search help pages for "keyword"
<code>?help</code>	# Search for more options
<code>help(package=base)</code>	# List tasks in package "base"

Many data scientist programmers and statisticians use R to design tools for analyzing data and to contribute their codes as pre-assembled collections of functions and objects called packages. Each R package is hosted at <http://cran.r-project.org>.

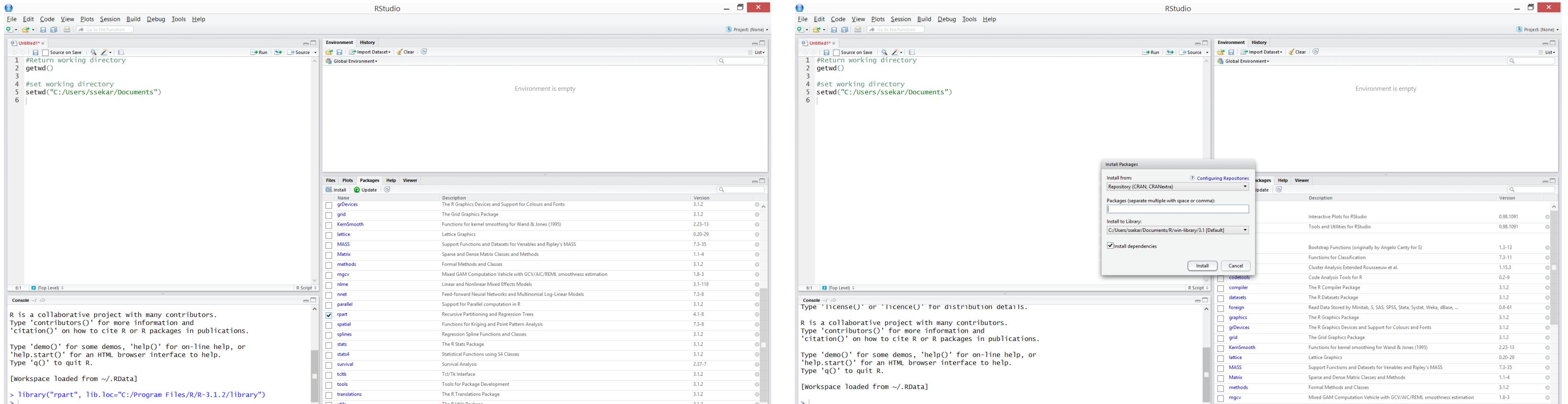
Available R packages are listed here:

R Package	Function
<i>library()</i>	# List available packages to load
<i>library("package")</i>	# Load the package
<i>library(help="package")</i>	# List package contents
<i>detach("package:pkg")</i>	# Unload the loaded package "pkg"
<i>install.packages("package")</i>	# Install the package



Not all packages are loaded by default, but they can be loaded/installed on demand.

You can install an R package by clicking **GUI RStudio -> Tools -> Install Packages.**



Check the package to load



QUIZ
1

Which of the following statements is true about R?

- a. R is a proprietary general purpose language.
- b. R excels in statistical computation and graphic capabilities.
- c. R syntaxes are very similar to the assembly language.



R has no Graphic User Interface (GUI).

QUIZ
1

Which of the following statements is true about R?

- a. R is a proprietary general purpose language.
- b. R excels in statistical computation and graphic capabilities.
- c. R syntaxes are very similar to the assembly language.



R has no Graphic User Interface (GUI).

The correct answer is **b**.

Explanation: R excels in statistical computation and graphic capabilities.

**QUIZ
2**

What does CRAN stand for?

- a. Comprehensive R Network
- b. Computational R Network
- c. Comprehensive R Archive Network



Computational R Archive Network

QUIZ
2

What does CRAN stand for?

- a. Comprehensive R Network
- b. Computational R Network
- c. Comprehensive R Archive Network



Computational R Archive Network

The correct answer is **c.**

Explanation: CRAN stands for Comprehensive R Archive Network.

**QUIZ
3**

Which of the following is not an IDE for R?

- a. RStudio
- b. RGUI
- c. Tinn-R

Dreamweaver



QUIZ
3

Which of the following is not an IDE for R?

- a. RStudio
- b. RGUI
- c. Tinn-R

Dreamweaver



The correct answer is **d**.

Explanation: Dreamweaver is not an IDE for R.

QUIZ
4

Which of the following commands is used to browse and access all help documents in R?

- a. *help.start()*
- b. *help(help)*
- c. *library()*

install.packages(help)



QUIZ
4

Which of the following commands is used to browse and access all help documents in R?

- a. *help.start()*
- b. *help(help)*
- c. *library()*

install.packages(help)



The correct answer is **a.**

Explanation: The command *help.start()* is used to browse and access all help documents in R.

QUIZ
5

Which of the following commands is used for loading an installed package in R?

- a. *load("package")*
- b. *library("package")*
- c. *install.packages("package")*

library()



QUIZ
5

Which of the following commands is used for loading an installed package in R?

- a. *load("package")*
- b. *library("package")*
- c. *install.packages("package")*



library()

The correct answer is **b**.

Explanation: The command *library("package")* is used for loading an installed package in R.

Let us summarize the topics covered in this lesson:



- R is a programming language developed as an alternative to the S language.
- R is available across all platforms—Windows, Mac, and Linux.
- R is most useful for statistical computation and visualization.
- R has a steep learning curve, and its working with large datasets is limited by the RAM size.
- R can be downloaded from either of the following websites:
 - CRAN
 - RStudio
- The RStudio program can run on a desktop or through a web browser.

Let us summarize the topics covered in this lesson:



- R stores user-defined objects in the workspace by allowing the user to take a snapshot of the current workspace and by automatically reloading it the next time R starts.
- Each R function comes with its own help page.
- Not all packages are loaded by default but can be loaded or installed on demand.

This concludes “Introduction to R.”

The next lesson is “R Programming.”



Data Science with R

Lesson 03—R Programming



After completing
this lesson, you will
be able to:



- Explain the four types of R operators
- Describe the different types of conditional statements in R
- Discuss the different types of loops in R
- Explain the commands to run an R script and a batch script
- List the commonly used R functions

R has many operators to perform different mathematical and logical operations. These can be categorized as follows:

Arithmetic operators

Relational operators

Logical operators

Assignment operators



These operators are used for mathematical operations like addition and multiplication. To understand their functions better, look at the table below:

Arithmetic Operators in R	
Operator	Description
+	Addition
-	Subtraction
*	Multiplication
/	Division
^	Exponent
%%	Modulus
%/%	Integer Division

These operators are used to compare two values. To understand their functions better, look at the table below:

Relational Operators in R	
Operator	Description
<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
==	Equal to
!=	Not equal to

These operators are used to perform Boolean operations such as “AND” and “OR.” To understand their functions better, look at the table below:

Logical Operators in R	
Operator	Description
!	Logical NOT
&	Element-wise logical AND
&&	Logical AND
	Element-wise logical OR
	Logical OR

These operators are used to assign values to variables. Variables are assigned using “`<-`”, although “`=`” also works.



Examples:

- `age <- 18` (left assignment)
- `18 -> age` (right assignment)

R supports two types of conditional statements:

If...else

Nested if...else

In if...else statements, when the test expression is True, the code in the “if” block executes; otherwise, the code in the “else” block executes.

If...else

Nested if...else

Example:

```
age <- 20
if(age > 18){
  print("Major")
} else {
  print("Minor")
}
```

In nested if...else statements, only one statement executes, depending on the test expressions in the “if” blocks.

If...else

Nested if...else



Example:

```
x <- 0
if (x < 0) {
  print("Negative number")
} else if (x > 0) {
  print("Positive number")
} else
  print("Zero")
```

This is a vector equivalent form of if...else.



Example:

```
a = c(1,2,3,4)  
ifelse(a %% 2 == 0, "even", "odd")
```

This is similar to a controlled branch of if...else statements.

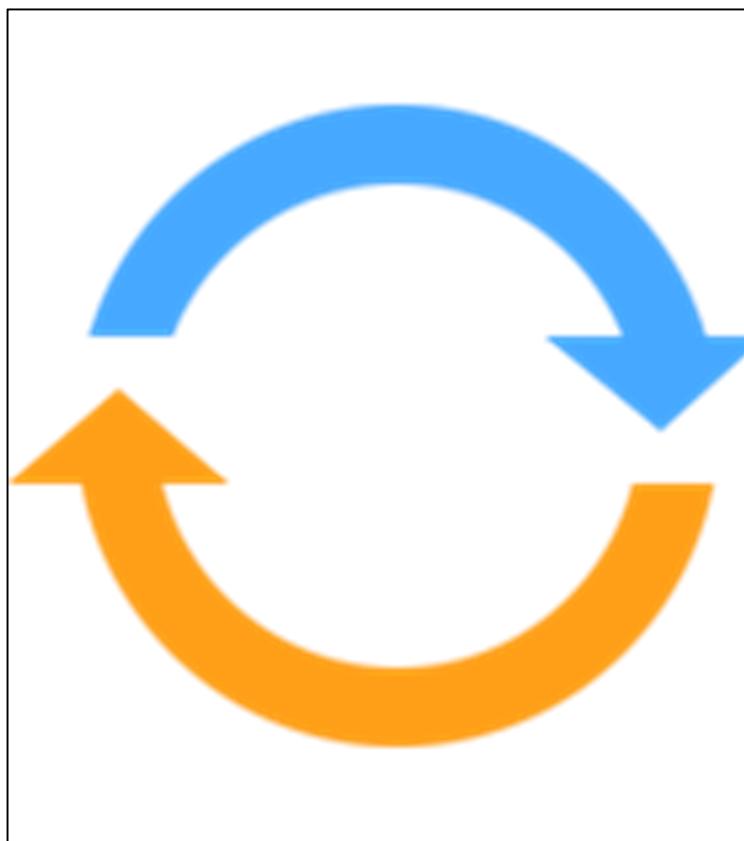
Example 1

```
switch(2, "apple", "ball", "cat")
```

Example 2

```
color = "green"  
switch(color, "red"={print("apple")}, "yellow"={print("banana")}, "green"={print("avocado")})
```

R supports the following types of loops:

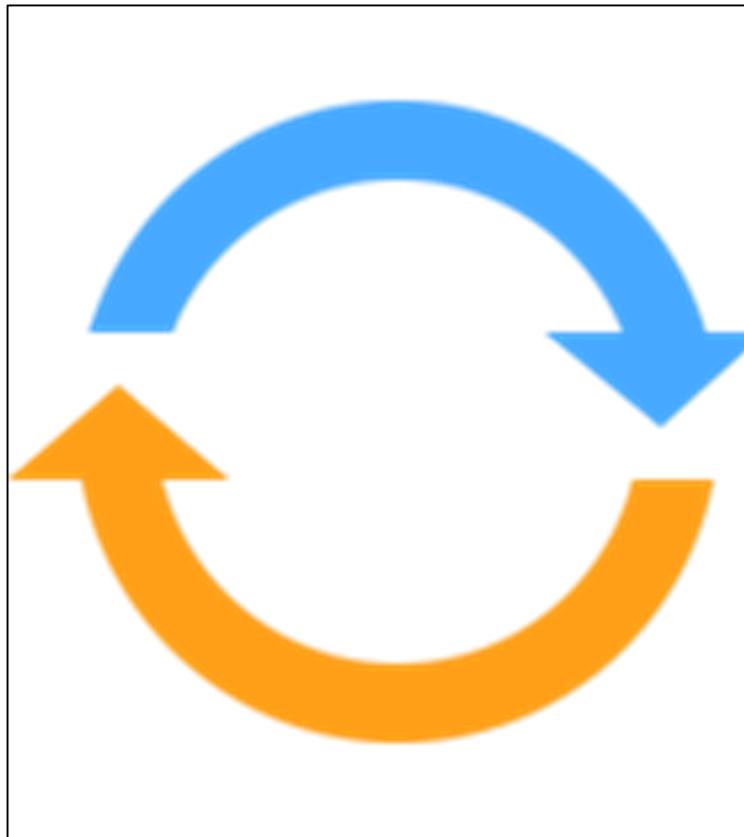


For Loop

While Loop

Repeat Loop

The for loop executes a code for a specific number of times.



For Loop

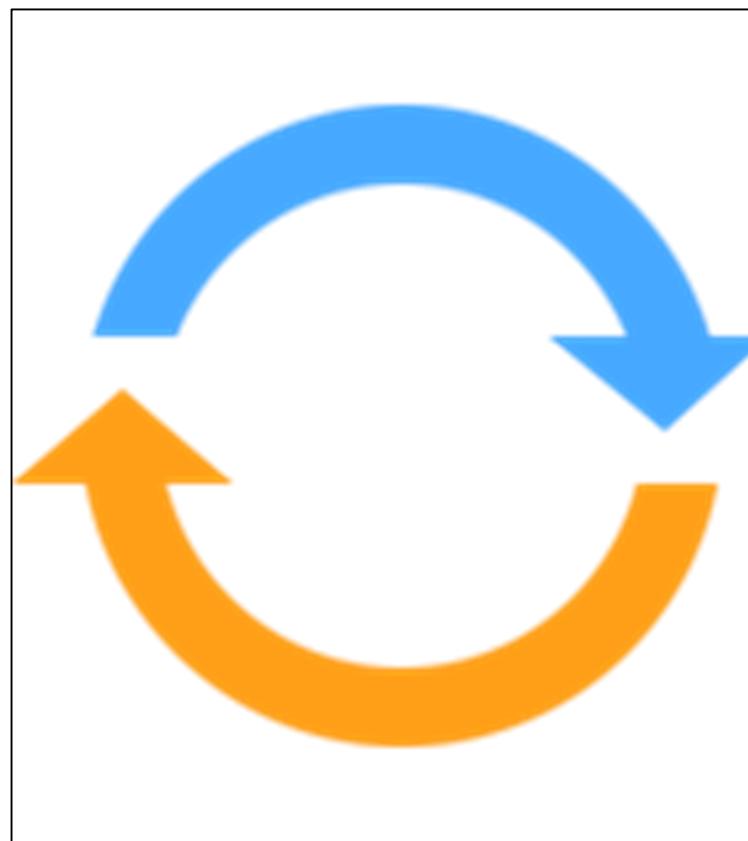
While Loop

Repeat Loop

Example:

```
vec <- c(1,2,3,4,5)
for (val in vec) {
    print(val)
}
```

In the while loop, while the test expression remains True, the code inside the loop keeps on executing.



For Loop

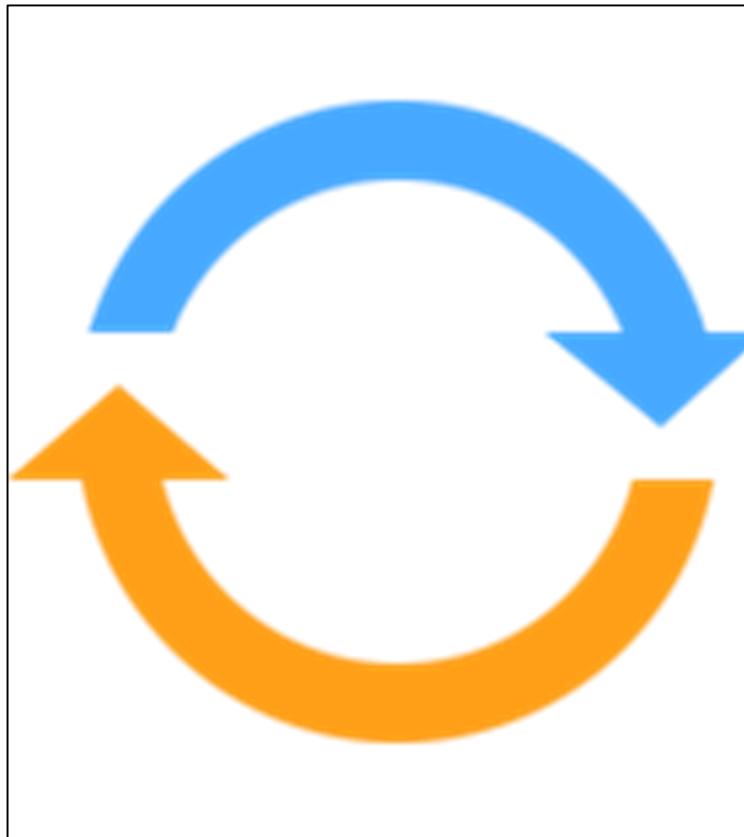
While Loop

Repeat Loop

Example:

```
i <- 1  
while (i < 6) {  
  print(i)  
  i = i+1  
}
```

A repeat loop iterates a code multiple times. Since there is no conditional check to exit the loop, you must specify it inside the body of the loop.



For Loop

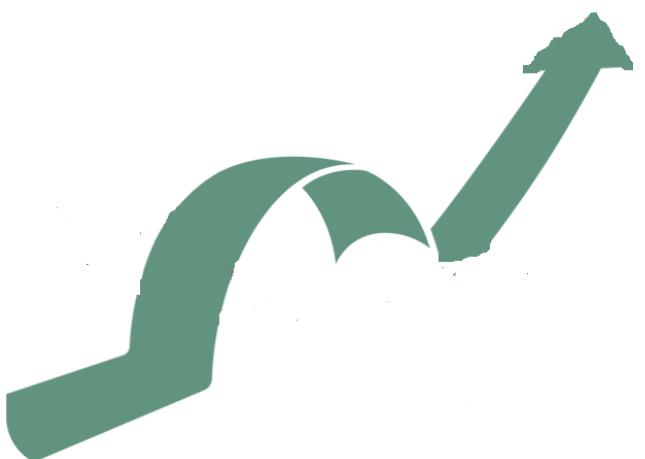
While Loop

Repeat Loop

Example:

```
x <- 1  
repeat {  
  print(x)  
  x = x+1  
  if (x == 6){  
    break  
  }  
}
```

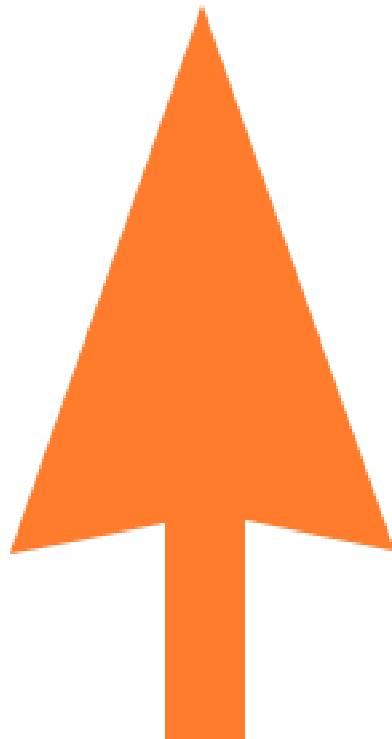
When present inside a loop, it stops the iterations from executing and forces the flow to jump off the loop.



Example:

```
num <- 1:5
for (val in num) {
  if (val == 3){
    break
  }
  print(val)
}
```

It helps in skipping the current iteration of a loop.



Example:

```
num <- 1:5
for (val in num) {
    if (val == 3){
        next
    }
    print(val)
}
```

It helps in reading data from the console or a file.

From the Console

Example:

```
x <- scan()
```

From a File

Example:

```
x <- scan("http://www.ats.ucla.edu/stat/data/scan.txt", what = list(age = 0, name =  
""))
```



Using this function may not always be an efficient way of reading files.

To run an R script, use:

- The “source()” function, which instructs R to read the text file and execute its contents
- The “echo=TRUE” parameter (optional), which echoes the script lines before they are executed



Example:

```
source("myScript.R", echo=TRUE)
```

To run a batch script, use:

- The “R CMD BATCH” command to run the code in batch mode
- The “Rscript” command to send the output to stdout or pass command-line arguments to the script



Examples:

- *R CMD BATCH my_script.R*
- *\$ Rscript myScript.R arg1 arg2*

The table below lists and explains the commonly used R functions:

Function	Description
append()	Add elements to a vector
c()	Combine values into a vector or list
identical()	Test if two objects are exactly equal
length()	Return the length of the R object
ls()	List objects in the current environment
range(x)	Return the minimum and maximum of the vector
rep(x,n)	Repeat the number x, n times
rev(x)	Provide the reversed version of its argument
seq(x,y,n)	Generate regular sequences from x to y, spaced by n
unique(x)	Remove duplicate entries from the vector

Other functions include:

Function	Description
tolower()	Convert a string to lower case letters
toupper()	Convert a string to upper case letters
grep()	Use for regular expressions
summary(x)	Return object summaries
str(x)	Compactly display the structure of an arbitrary R object
glimpse(x)	Compactly display the structure of an arbitrary R object (dplyr package)
class(x)	Return the class of an object
mode(x)	Get or set the type or storage mode of an object
summary(x)	Return object summaries



QUIZ
1

Which of the following is a valid equation?

- a. TRUE %/% FALSE = TRUE
- b. (TRUE | FALSE) & (FALSE != TRUE) == TRUE
- c. (TRUE | FALSE) & (FALSE != TRUE) == FALSE

"A" && "a"



QUIZ
1

Which of the following is a valid equation?

- a. TRUE %/% FALSE = TRUE
- b. (TRUE | FALSE) & (FALSE != TRUE) == TRUE
- c. (TRUE | FALSE) & (FALSE != TRUE) == FALSE



"A" && "a"

The correct answer is **b**.

Explanation: (TRUE | FALSE) & (FALSE != TRUE) == TRUE is a valid equation.

QUIZ
2

Fill in the blank with the correct operator:

$$4 \underline{\quad} 3 = 1$$

a. /

b. *

c. %/%

None of the above



QUIZ
2

Fill in the blank with the correct operator:

$$4 \underline{\quad} 3 = 1$$

a. /

b. *

c. %/%

None of the above



The correct answer is **c.**

Explanation: The correct operator is %/"/>.

QUIZ
3

What will be the output of “age <- 18, 18 -> age, print(age)”?

- a. 18
- b. Error
- c. NA



Binary value of 18

QUIZ
3

What will be the output of “age <- 18, 18 -> age, print(age)”?

- a. 18
- b. Error
- c. NA



Binary value of 18

The correct answer is **a**.

Explanation: The output of “age <- 18, 18 -> age, print(age)” will be 18.

QUIZ
4

Can an “if” statement be used without an “else” block?

- a. Yes
- b. No



QUIZ
4

Can an “if” statement be used without an “else” block?

- a. Yes
- b. No



The correct answer is **a.**

Explanation: Yes. An “if” statement can be used without an “else” block.

QUIZ
5

What will be the output of “`num <- 1:5, for (val in num) {, next, break, print(val), }`”?

- a. The output will be an error.
- b. The output will be numbers: 3, 4, 5.
- c. There will be no output.

There will be an infinite loop.



QUIZ
5

What will be the output of “num <- 1:5, for (val in num) {, next, break, print(val), } ?

- a. The output will be an error.
- b. The output will be numbers: 3, 4, 5.
- c. There will be no output.



There will be an infinite loop.

The correct answer is **c**.

Explanation: There will be no output of the given command; however, the program will run.

Let us summarize the topics covered in this lesson:



- Four types of operators in R are arithmetic, relational, logical, and assignment.
- Two types of conditional statements in R are if...else and nested if...else.
- Three types of loops in R are for loop, while loop, and repeat loop.
- To run an R script, use the “source” function.
- To run a batch script in R, use the “R CMD BATCH” command.
- The commonly used functions in R are given below:

append()	unique(x)
c()	tolower()
identical()	toupper()
length()	grep()
ls()	summary(x)
range(x)	str(x)
rep(x,n)	glimpse(x)
rev(x)	class(x)
seq(x,y,n)	mode(x)

This concludes “R Programming.”

The next lesson is “R Data Structure.”

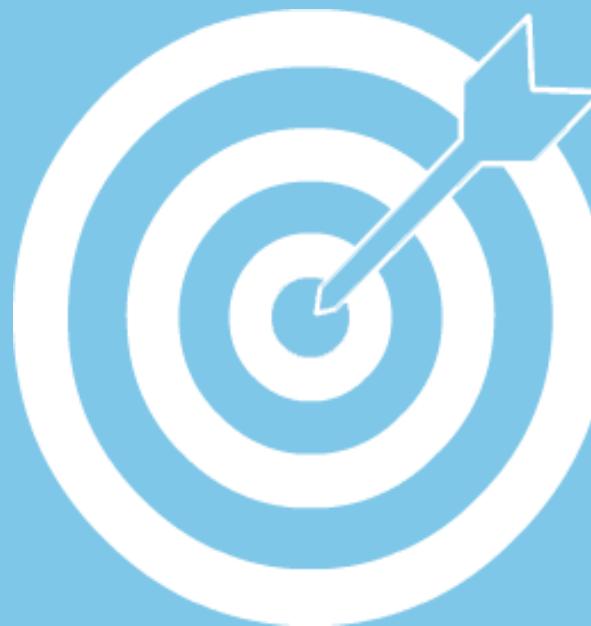


Data Science with R

Lesson 04—R Data Structures

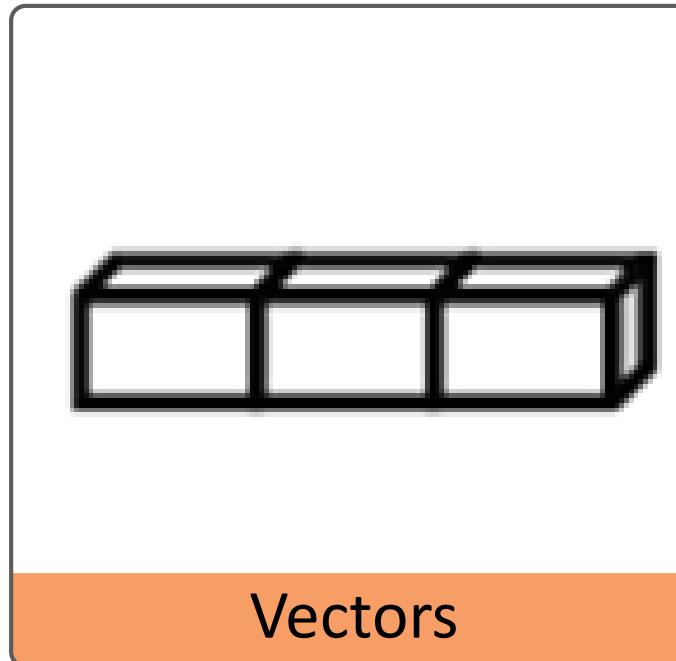


After completing
this lesson, you will
be able to:

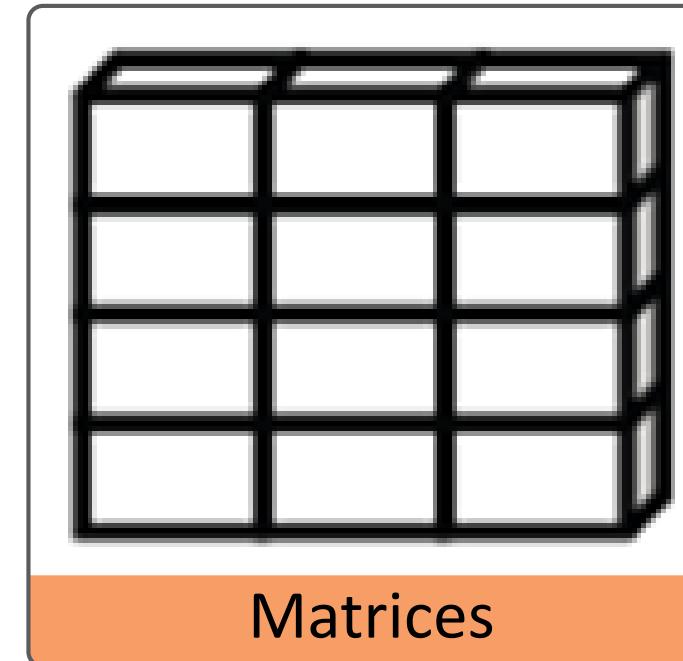


- Explain the different data structures used in R
- Discuss the elements of the different data structures in R
- Explain the acceptable formats to import and export data in R

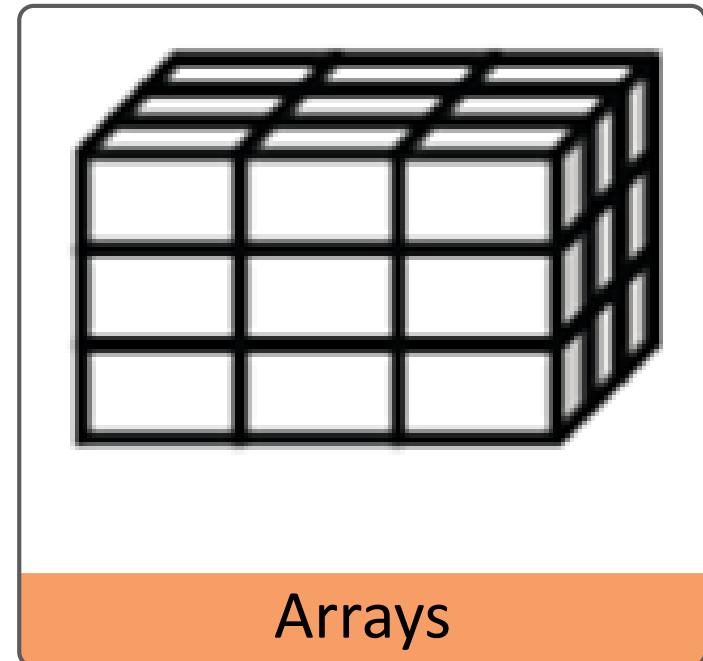
The different data structures used in R are as follows:



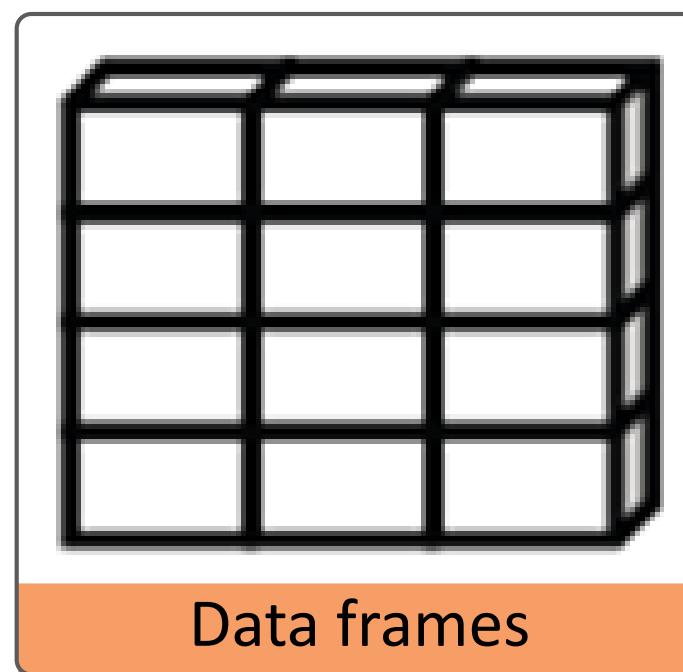
Vectors



Matrices



Arrays



Data frames

- Vectors
- Arrays
- Data frames
- Matrices
- Lists

Lists

These are one-dimensional arrays.



Examples:

- `a <- c(1, 2, 5, 3, 6, -2, 4)`
- `b <- c("one", "two", "three")`
- `c <- c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE)`

Where,

- a is a numeric vector.
- b is a character vector.
- c is a logical vector.

These are one-element vectors used to hold constants.



Example:

`f <- 3`

`g <- "US"`

`h <- TRUE`

It helps create a vector with a shortcut notation.



Example:

`a <- c(1:5)`

This is equivalent to:

`a <- c(1, 2, 3, 4, 5)`

A numeric vector of positions within brackets is used to refer to the various elements of a vector.

Example 1

```
vec <- c("a", "b", "c", "d", "e", "f")
```

Example 2

```
vec[1] # will return the first element in the vector
```

Example 3

```
vec[c(2,4)] # will return the second and fourth elements in the vector
```

These are two-dimensional data structures.



Example:

```
vector <- c(1,2,3,4)  
foo <- matrix(vector, nrow=2, ncol=2)
```



Elements in a matrix must be of the same type, whether a number, character, or Boolean.

Matrices `byrow` is an optional parameter used by matrices.



Example:

- `foo <- matrix(vector, nrow=2, ncol=2, byrow = TRUE)`
- `foo <- matrix(vector, nrow=2, ncol=2, byrow = FALSE)`

You can access a matrix elements by using subscripts or brackets.

Examples:

- `mat <- matrix(c(1:4), nrow=2, ncol = 2)`
- `mat[1,]` *# returns the first row in the matrix.*
- `mat[2,]` *# returns the second row in the matrix.*
- `mat[,1]` *# returns the first column in the matrix.*
- `mat[,2]` *# returns the second column in the matrix.*
- `mat[1,2]` *# returns the element in the first row of the second column.*

Similar to matrices; these can have more than two dimensions.



Examples:

- `a <- matrix(c(1,1,1,1), 2, 2)`
- `b <- matrix(c(2,2,2,2), 2, 2)`
- `foo <- array(c(a,b), c(2,2,2))`

These can be accessed in the same way as matrix elements.



Examples:

- `foo[1,,]` *# returns all elements in the first dimension*
- `foo[2,,]` *# returns all elements in the second dimension*
- `foo[2,1,]` *# returns only the first row element in the second dimension*

These are the most commonly used data structures in R.

A data frame is similar to a general matrix, but its columns can contain different modes of data, such as a number and character.



Examples:

- `name <- c("joe", "jhon", "Nancy")`
- `sex <- c("M", "M", "F")`
- `age <- c(27,26,26)`
- `foo <- data.frame(name,sex,age)`

These can be accessed by column names.



Examples:

- `foo$name` *# returns the name vector in the data frame*
- `foo$age` *# returns the age vector in the data frame*
- `foo$age[2]` *# returns the second element of the age vector in the data frame*

These are the categorical variables in R.



Examples:

- `gender <- c("Male", "Female", "Female", "Male")`
- `status <- c("Poor", "Improved", "Excellent", "Poor", "Excellent")`
- `factor_gender <- factor(gender) # factor_gender has two levels, Male and Female.`
- `factor_status <- factor(status) # factor_status has three levels, Poor, Improved, and Excellent.`



These are the most complex data structures. A list may contain a combination of vectors, matrices, data frames, and even other lists.

**Example:**

```
vec <- c(1,2,3,4)
mat <- matrix(vec,2,2)
foo <- list(vec, mat)
```

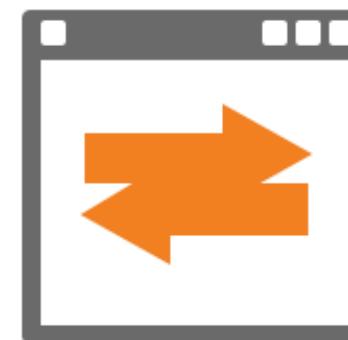
You can import data from four types of files in R:

Excel

Minitab

Table

CSV



Before using the sample data available in an Excel format, you need to import the data into R. Let's understand this process by using the following two examples:

Example 1

```
library(gdata)          # load gdata package  
help(read.xls)         # documentation  
mydata = read.xls("mydata.xls")    # read from first sheet
```

Example 2

```
library(XLConnect)  
wk = loadWorkbook("mydata.xls")  
df = readWorksheet(wk, sheet="Sheet1")
```

Use the function `read.mtp` to import the sample data from a Minitab Portable Worksheet format. This function returns a list of components in the Minitab worksheet.



Example:

```
library(foreign)
help(read.mtp)
mydata = read.mtp("mydata.mtp")
```



A text file can have a data table in it. The cells inside the table are separated by blank characters. Here's an example of a table with four rows and three columns. Let's see how to import this data.

100	a1	b1
200	a2	b2
300	a3	b3
400	a4	b4



Example:

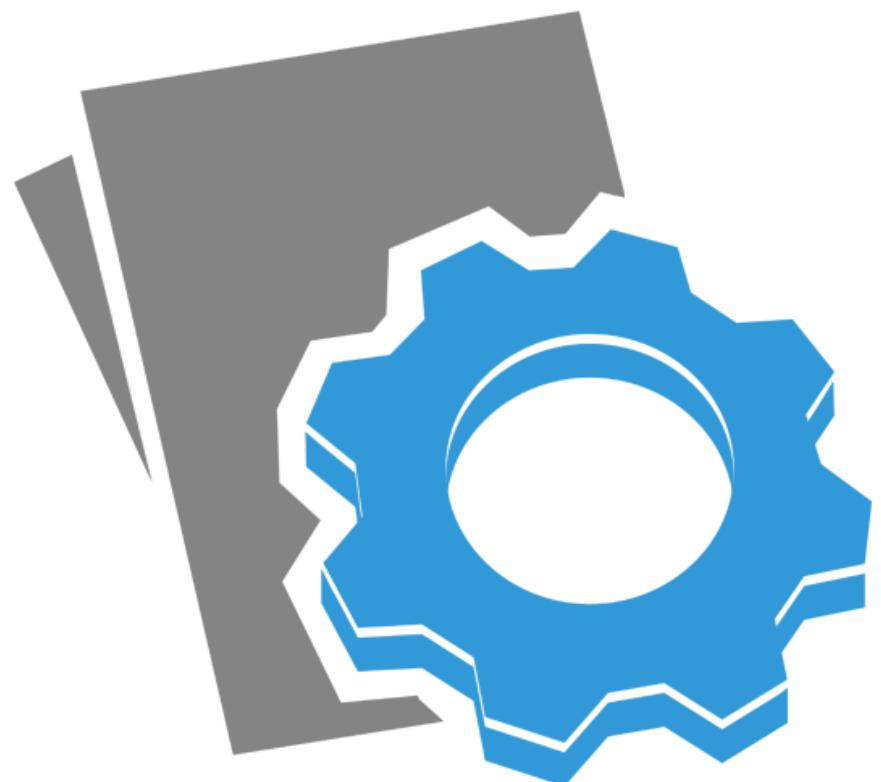
```
help(read.table)  
mydata = read.table("mydata.txt")
```

R allows data import from a Comma Separated Values (CSV) format as well. Each cell inside such a data file is separated by a special character, usually a comma.



Example:

```
help(read.csv)  
mydata = read.csv("mydata.csv", sep=",")
```



R supports data export from three types of files:



Table

Excel

CSV

To export data in a table format from R, refer to the following example:



Table

Excel

CSV

Example:

```
help(write.table)  
write.table(mydata,  
"c:/mydata.txt", sep="\t")
```

To export data in an Excel format from R, refer to the following example:



Table

Excel

CSV

Example:

```
library(xlsx)  
help(write.xlsx)  
write.xlsx(mydata,  
"c:/mydata.xlsx")
```

To export data in a CSV format from R, refer to the following example:



Table

Excel

CSV

Example:

help(write.csv)

write.csv(mydata, file = "mydata.csv")



QUIZ
1

What will be the output of “`vec <- c(1,"hello",TRUE)`”?

- a. vec will be assigned to multiple values.
- b. Nothing will happen.
- c. An error will occur.



vec will have one value, True.

QUIZ
1

What will be the output of “vec <- c(1,"hello",TRUE)”?

- a. vec will be assigned to multiple values.
- b. Nothing will happen.
- c. An error will occur.



vec will have one value, True.

The correct answer is **c**.

Explanation: The output of the given code will be an error.

QUIZ
2

Which statement about R data structures is true?

- a. A matrix is a three-dimensional collection of values of the same type.
- b. A factor can be used to represent a categorical variable.
- c. A vector is a two-dimensional collection of values that can have multiple modes (numeric, character, Boolean).

A data frame can hold only up to 20GB of data.



QUIZ
2

Which statement about R data structures is true?

- a. A matrix is a three-dimensional collection of values of the same type.
- b. A factor can be used to represent a categorical variable.
- c. A vector is a two-dimensional collection of values that can have multiple modes (numeric, character, Boolean).



A data frame can hold only up to 20GB of data.

The correct answer is **b**.

Explanation: A factor is a data structure that can be used to represent a categorical variable.

**QUIZ
3**

For the dataset mentioned below, which data structure will be the most appropriate?

Name	Age	Gender
Jhon	24	M
Joe	24	M
Nancy	25	F

- a. Matrix
- b. Data Frame
- c. Array

List



QUIZ
3

For the dataset mentioned below, which data structure will be the most appropriate?

Name	Age	Gender
Jhon	24	M
Joe	24	M
Nancy	25	F

- a. Matrix
- b. Data Frame
- c. Array

List



The correct answer is **b**.

Explanation: A data frame will be the most appropriate data structure for the given dataset.

QUIZ
4

Identify the correct way to create an array.

- a. `a(vector, dimensions, dimnames)`
 - b. `create(vector, dimensions, dimnames)`
 - c. `array(vector, dimensions, dimnames)`
- `a(vector, dimensions)`



QUIZ
4

Identify the correct way to create an array.

- a. `a(vector, dimensions, dimnames)`
 - b. `create(vector, dimensions, dimnames)`
 - c. `array(vector, dimensions, dimnames)`
- `a(vector, dimensions)`



The correct answer is **c**.

Explanation: `array(vector, dimensions, dimnames)` is the correct way to create an array.

Let us summarize the topics covered in this lesson:



- Five types of data structures in R are vectors, matrices, arrays, data frames, and lists.
- Four types of formats supported by R for importing data are Excel, Minitab, table, and CSV.
- Three types of formats supported by R for exporting data are table, Excel, and CSV.

This concludes “R Data Structure.”

The next lesson is “Apply Functions.”



Data Science with R

Lesson 05—Apply Functions



After completing
this lesson, you will
be able to:



- Explain the various types of apply functions
- Define the dplyr package
- Discuss how to install dplyr
- Describe the various dplyr functions

The apply functions are used to perform a specific change to each column or row of R objects. There are six types of apply functions in R:

apply

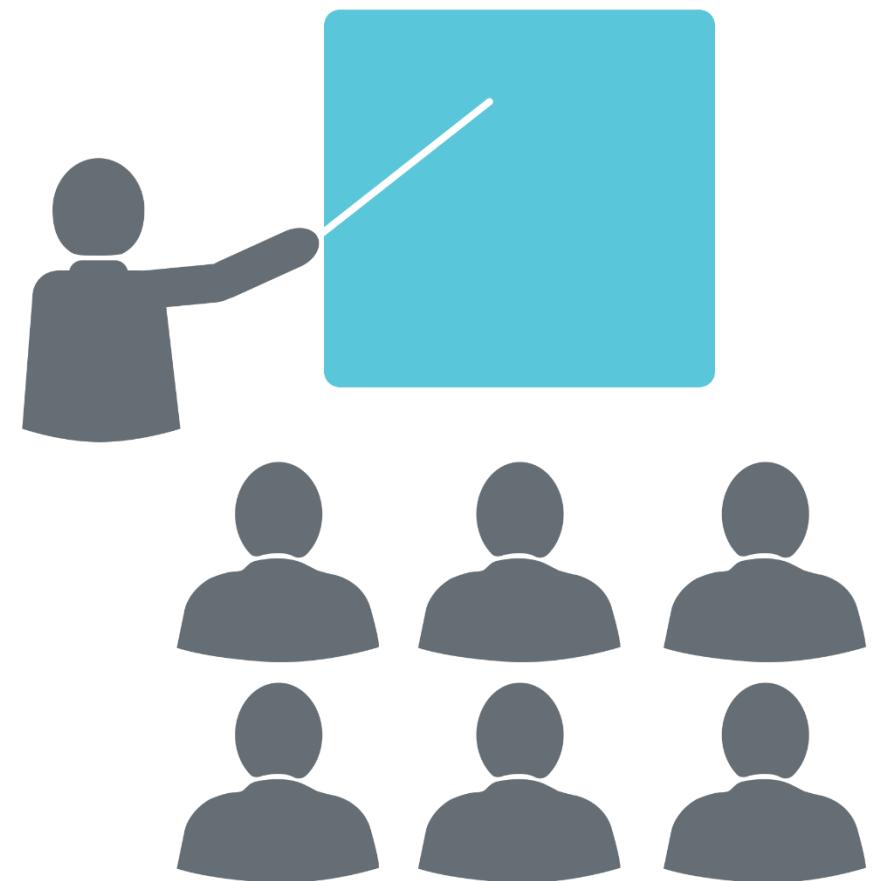
lapply

sapply

tapply

vapply

mapply



It helps apply a function to a matrix row or column and returns a vector, array, or list.

Syntax

Examples

It takes three arguments: matrix/array, margin, and function. The syntax to use this function is:

Syntax

Examples

apply(x, margin, function)

Where,

- *margin* indicates whether the function is to be applied to a row or column.
 - *margin = 1* indicates that the function needs to be applied to a row.
 - *margin = 2* indicates that the function needs to be applied to a column.
- *function* can be any function such as mean, sum, or average.

A few examples to use this function are:

Syntax

Examples

- 
- *m <- matrix(c(1,2,3,4),2,2)*
 - *apply(m,1,sum)*
 - *apply(m,2,sum)*

It takes a list as an argument and works by looping through each element in the list. The output of this function is a list. The syntax to use this function, along with some examples, is:

Syntax

lapply(list, function)

Examples

- *list <- list(a = c(1,1), b=c(2,2), c=c(3,3))*
- *lapply(list,sum)*
- *lapply(list,mean)*

It is similar to lapply(), except that it simplifies the result so that:

- If the result is a list and every element in the list is of size 1, then a vector is returned.
- If the result is a list and every element in the list is of the same size (>1), then a matrix is returned.

Otherwise, the result is returned as a list itself.

The syntax to use this function, along with some examples, are:

Syntax

```
sapply(list, func)
```

Examples

- *list <- list(a = c(1,1), b=c(2,2), c=c(3,3))
sapply(list,sum)*
- *list <- list(a = c(1,2), b=c(1,2,3), c=c(1,2,3,4))
sapply(list, range)*

It works on vectors and groups factors inside those vectors.

Syntax

Example

To define it, you need three arguments: vector, factor of vector, and function. The syntax to use this function is:

Syntax

Example

tapply(x, factor, fun)

An example to use this function is:

Syntax

Example

```
age <- c(23,33,28,21,20,19,34)
gender <- c("m","m","m","f","f","f","m")
f <- factor(gender)
tapply(age,f,mean)
```

Vapply() Function

It works like sapply(), except that you need to specify the type of return value, which could be an integer, a double, or a character.

It helps save time in coercing returned values to fit in a single atomic vector.

The syntax to use this function is:

Syntax

```
vapply(x, function, FUN.VALUE)
```

An example to use this function is:

Example

```
list <- list(a = c(1,1), b=c(2,2), c=c(3,3))
```

```
vapply(list, sum, FUN.VALUE=double(1))
```

Being a multivariate version of `apply()`, `mapply()` applies `FUN` (function to apply) to the first, second, third elements (and so on) of each argument. Arguments are recycled if necessary.

Syntax

Example

The syntax to use this function is given below:

Syntax

Example



mapply(FUN, ...)

Some examples to use this functions are:

Syntax

Example

```
list(rep(1, 4), rep(2, 3), rep(3, 2), rep(4, 1))
```

In this example, the same function (rep) is being repeatedly called out, where:

- The first argument varies from 1 to 4.
- The second argument varies from 4 to 1.

Instead, you can use mapply() as follows:

```
mapply(rep, 1:4, 4:1)
```

This will produce the same result.

It is a powerful R package:

- It transforms and summarizes tabular data with rows and columns.
- It provides simple **verbs**— functions that correspond to the most common data manipulation tasks to help you translate your thoughts into code.

The use of efficient data storage backends by dplyr results in quicker processing speed.



It is referred to as the grammar of data manipulation. It provides the following five verbs (or functions) that are applied on the data set:

- *select*: helps select rows in a table or dataframe
- *filter*: helps filter records in a table or dataframe
- *arrange*: helps rearrange a table or dataframe
- *mutate*: helps add new data
- *summarize*: helps state the data summary



Remember: dplyr is not a part of the default package of R.

To install it separately; use the following command:

```
install.packages("dplyr")
```

To load it into the memory; use the following command:

```
library(dplyr)
```



The dplyr package has the following functions:

- Select()
- Filter()
- Arrange()
- Mutate()
- Summarize()

To understand the use of these functions, let's consider the dataset "mtcars".



This function allows you to select specific columns from large data sets.

To select columns by name

```
select(mtcars, mpg, disp)
```

To select a range of columns by name

```
select(mtcars, mpg:hp)
```

To select columns and rows with string match

```
select(iris, starts_with("Petal"))
select(iris, ends_with("Width"))
select(iris, contains("etal"))
select(iris, matches(".t."))
```

This function enables easy filtering, zoom in, and zoom out of relevant data. The two types of filters are explained below:

Simple filter

```
filter(mtcars, cyl == 8)  
filter(mtcars, cyl < 6)
```

Multiple criteria filter

```
filter(mtcars, cyl < 6 & vs == 1)  
filter(mtcars, cyl < 6 | vs == 1)
```



Comma separated arguments are equivalent to the "And" condition; **Example:** `filter(mtcars, cyl < 6, vs == 1)`

This function helps arrange the data in a specific order. The syntax to use this function, along with some examples, is:

Syntax

arrange(data, ordering_column)

Examples

- *arrange(mtcars, desc(disp))*
- *arrange(mtcars, cyl, disp)*

This function helps add new variables to an existing data set. The syntax to use this function, along with an example, is:

Syntax

```
mutate(data, new_column)
```

Examples

```
mutate(mtcars, my_custom_disp = disp /  
1.0237)
```

This function summarizes multiple values to a single value in a dataset. Here are examples to use this function without and with the group function:

Simple without Group Function

```
summarise(mtcars, mean(disp))
```

Summarize with Group Function

```
summarise(group_by(mtcars, cyl), mean(disp))  
summarise(group_by(mtcars, cyl), m = mean(disp), sd = sd(disp))
```

Here's a list of summary functions that can be used within this function:

- **first**: Returns the first element of a vector
- **last**: Returns the last element of a vector
- **nth(x,n)**: Returns the 'n'th element of a vector
- **n()**: Returns the number of rows in a dataframe
- **n_distinct(x)**: Returns the number of unique values in vector x

In addition, the following functions are also used:

mean	median	mode
max	min	sum
var	length	IQR



QUIZ
1

Which of the following statements is true about the `apply(x, margin, function)`?

- a. When $margin = 2$, the function needs to be applied to a row.
- b. When $margin = 1$, the function needs to be applied to a row.
- c. x must be of type list.



Only arithmetic functions can be passed to the `apply()` function.

QUIZ
1

Which of the following statements is true about the `apply(x, margin, function)`?

- a. When $margin = 2$, the function needs to be applied to a row.
- b. When $margin = 1$, the function needs to be applied to a row.
- c. x must be of type list.



Only arithmetic functions can be passed to the `apply()` function.

The correct answer is **b**.

Explanation: The function given above means that when $margin = 1$, the function needs to be applied to a row.

QUIZ 2

Identify an accurate statement about the lapply() function.

- a. It takes a list as an argument and works by looping through each element in the list.
 - b. It takes a list, an array, or a matrix and loops through each element in the list.
 - c. It is not a standalone function and needs to be applied with the apply() function.
- It is used when the latitude and longitude of an object come into the picture.



QUIZ
2

Identify an accurate statement about the lapply() function.

- a. It takes a list as an argument and works by looping through each element in the list.
 - b. It takes a list, an array, or a matrix and loops through each element in the list.
 - c. It is not a standalone function and needs to be applied with the apply() function.
- It is used when the latitude and longitude of an object come into the picture.



The correct answer is **a**.

Explanation: The lapply() function takes the list as an argument and works by looping through each element in the list.

QUIZ
3

State whether the following statement is true or false.

dplyr is a powerful R package for transforming and summarizing tabular data with rows and columns. It is also referred to as the grammar of data manipulation.

- a. True
- b. False



QUIZ
3

State whether the following statement is true or false.

dplyr is a powerful R package for transforming and summarizing tabular data with rows and columns. It is also referred to as the grammar of data manipulation.

- a. True
- b. False



The correct answer is **a.**

Explanation: dplyr is a powerful R package for transforming and summarizing tabular data with rows and columns. It is also referred to as the grammar of data manipulation.

QUIZ
4

Which dplyr command is used to rearrange the order of columns in a data set?

- a. `order_data(data, ordering_column)`
- b. `sort_data(data,ordering_column)`
- c. `dplyr(data,ordering_column)`

`arrange(data, ordering_column)`



QUIZ
4

Which dplyr command is used to rearrange the order of columns in a data set?

- a. `order_data(data, ordering_column)`
- b. `sort_data(data,ordering_column)`
- c. `dplyr(data,ordering_column)`

`arrange(data, ordering_column)`



The correct answer is **d**.

Explanation: The dplyr command, `arrange(data, ordering_column)`, is used to rearrange the order of columns in a data set.

Let us summarize the topics covered in this lesson:



- The six types of apply functions are apply, lapply, sapply, tapply, vapply, and mapply.
- dplyr is a powerful R package that transforms and summarizes tabular data with rows and columns.
- dplyr is not a part of the default R package and needs to be installed separately.
- The five types of dplyr functions are select, filter, arrange, mutate, and summarize.

This concludes “Apply Functions.”

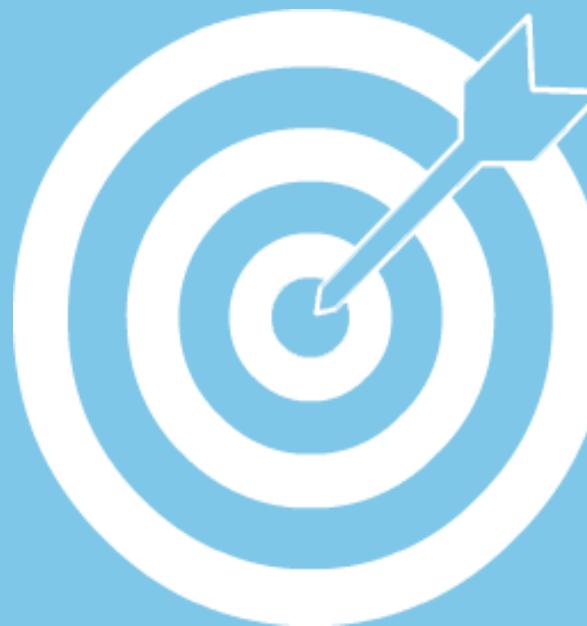
The next lesson is “Data Visualization.”

Data Science with R

Lesson 06—Data Visualization



After completing
this lesson, you will
be able to:



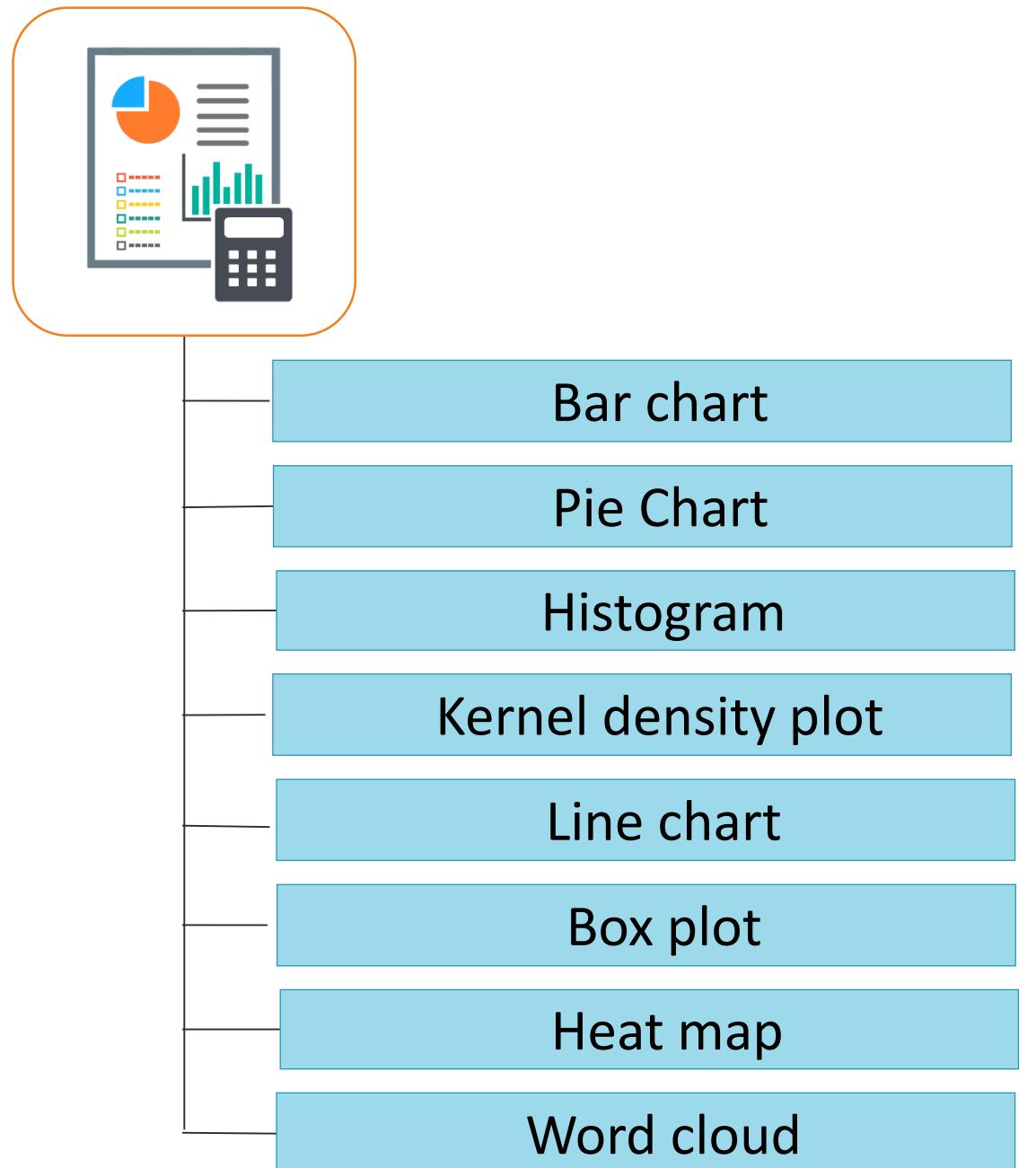
- Explain the various types of graphics available in R
- List the possible file formats of graphic outputs
- Describe the methods to save graphics as files
- Describe the procedure to export graphs in RStudio

R includes powerful packages of graphics that help in data visualization. These graphics can be:

- Viewed on a screen
- Saved in various formats, such as .pdf, .png, .jpg, .wmf, and .ps
- Customized according to the varied graphic needs



R supports eight types of graphics:



Bar charts:

- Are horizontal or vertical bars to show comparisons between categorical values
- Represent lengths, frequency, or proportion of categorical values

The syntax to create a simple bar chart:

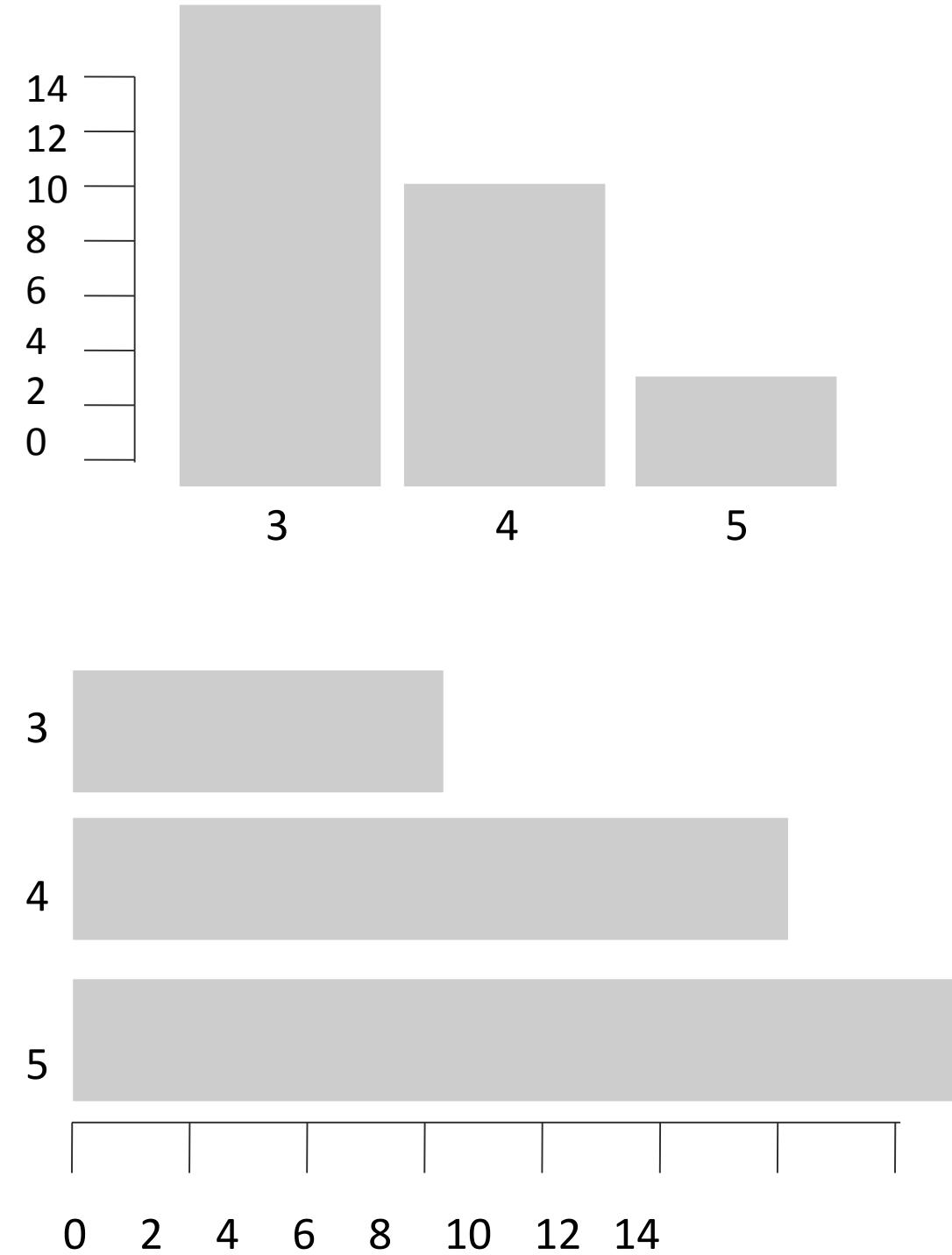
Syntax

```
barplot(x)
```

A few examples to create vertical and horizontal bar charts are as follows:

Examples

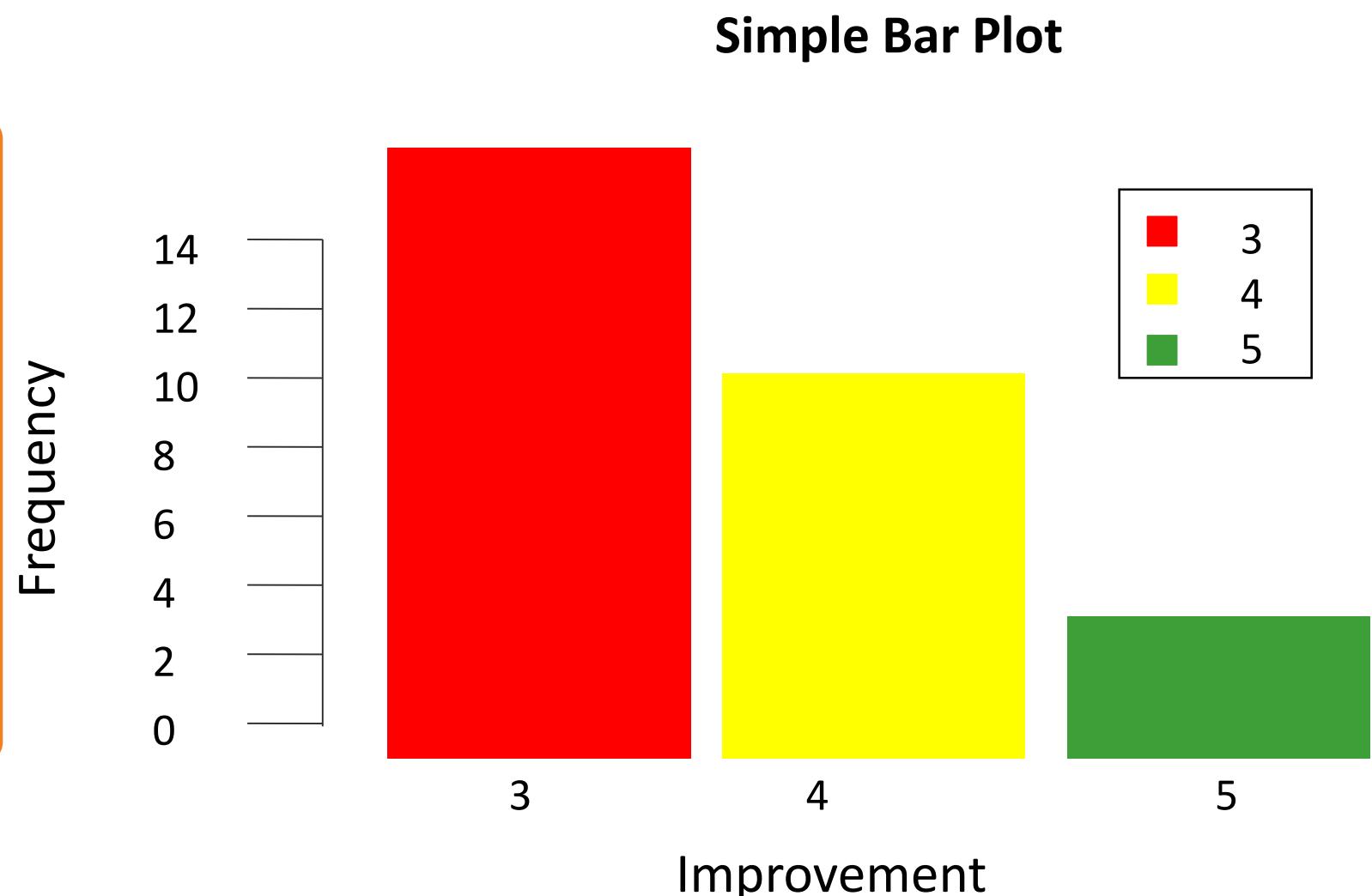
- *counts <- table(mtcars\$gear)
barplot(counts)*
- *#horizontal bar chart
barplot(counts, horiz=TRUE)*



Editing a Simple Bar Chart

You can add titles, legends, and colors to a simple bar chart as shown below:

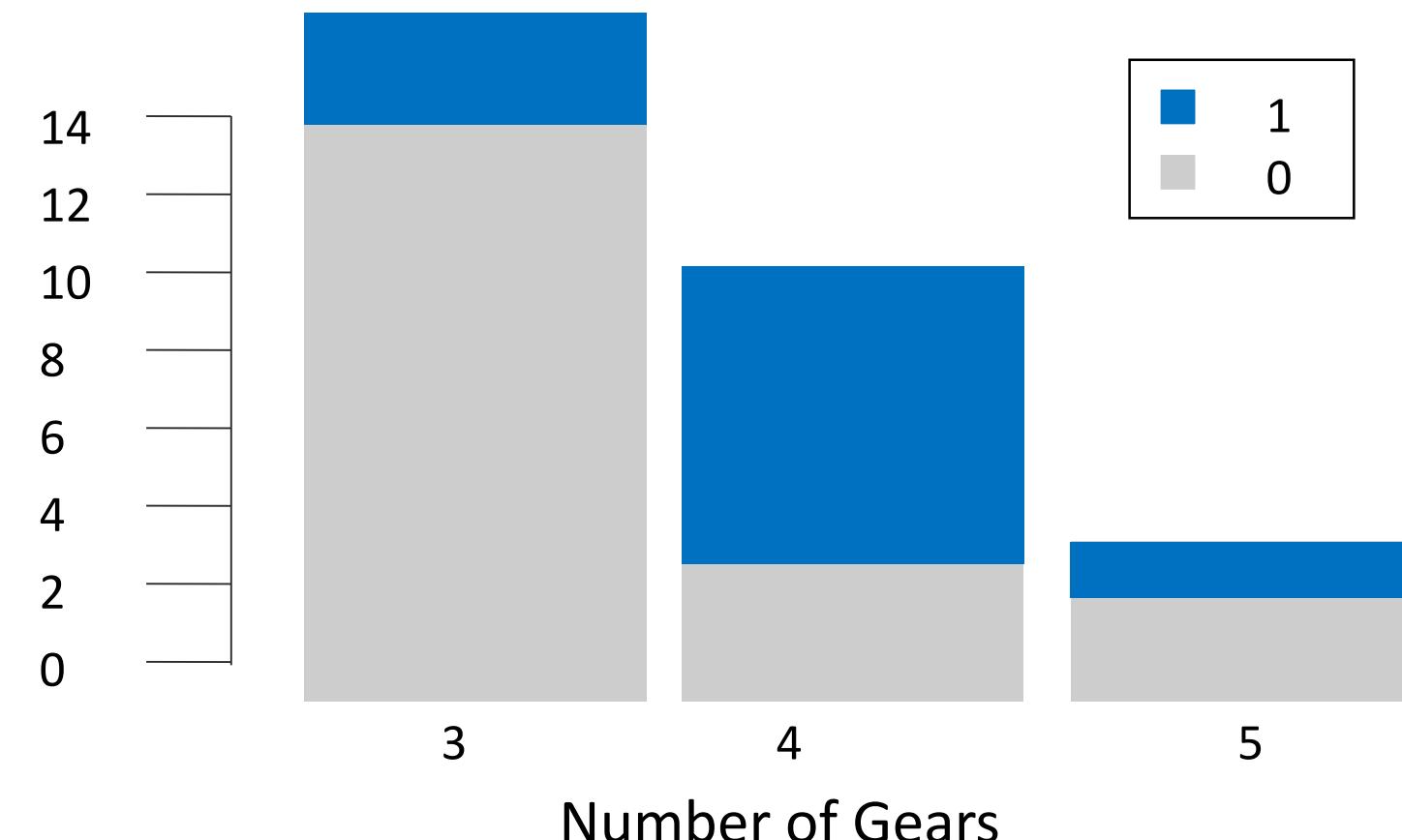
```
counts <- table(mtcars$gear)
barplot(counts,
        main="Simple Bar Plot",
        xlab="Improvement",
        ylab="Frequency",
        legend=rownames(counts),
        col=c("red", "yellow", "green"))
)
```



You can create a stacked bar plot with colors and legends as shown below:

```
counts <- table(mtcars$vs, mtcars$gear)
barplot(counts,
        main="Car Distribution by Gears and VS",
        xlab="Number of Gears",
        col=c("grey","cornflowerblue"),
        legend = rownames(counts))
```

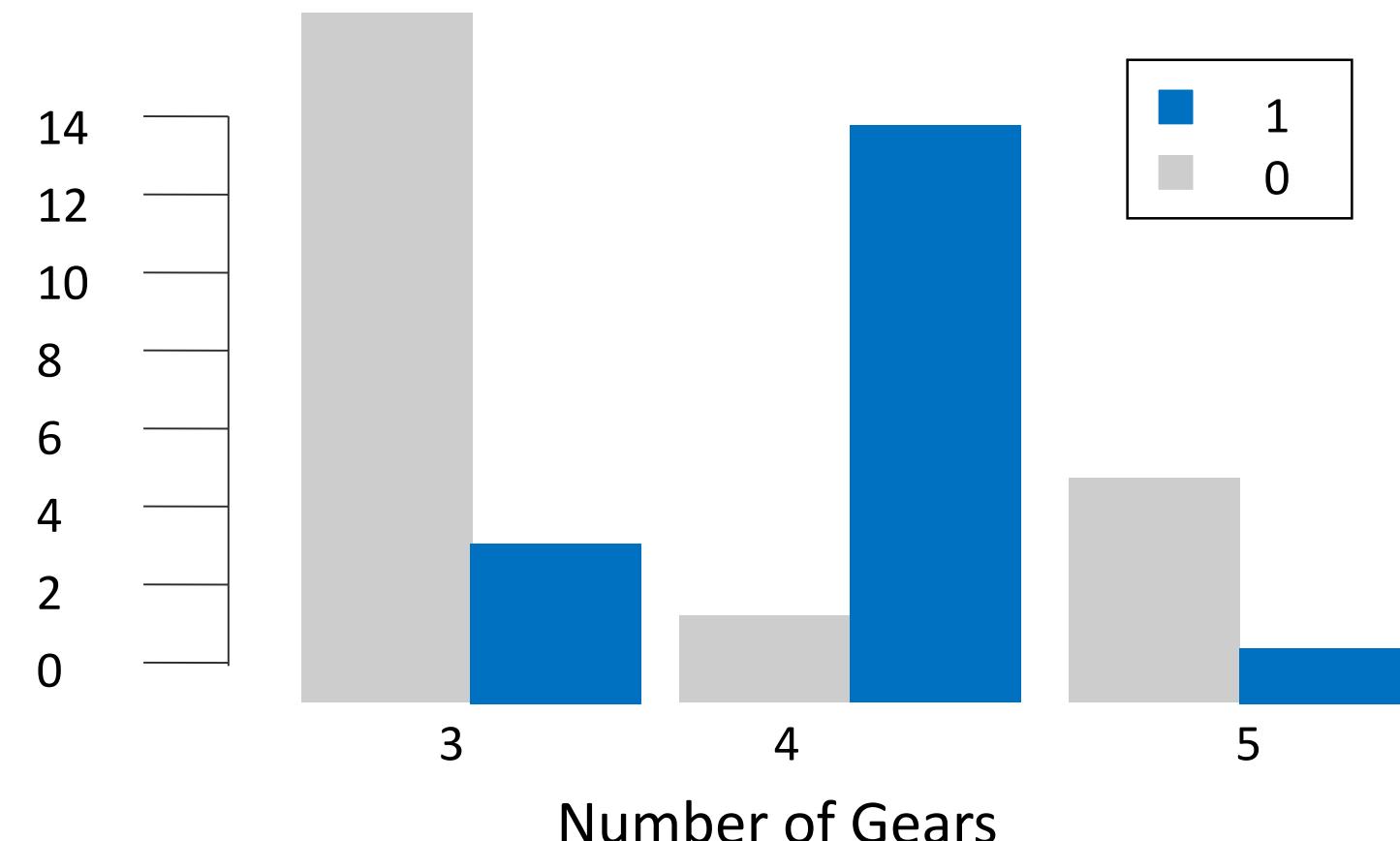
Car Distribution by Gears and VS



You can create a grouped bar plot as shown below:

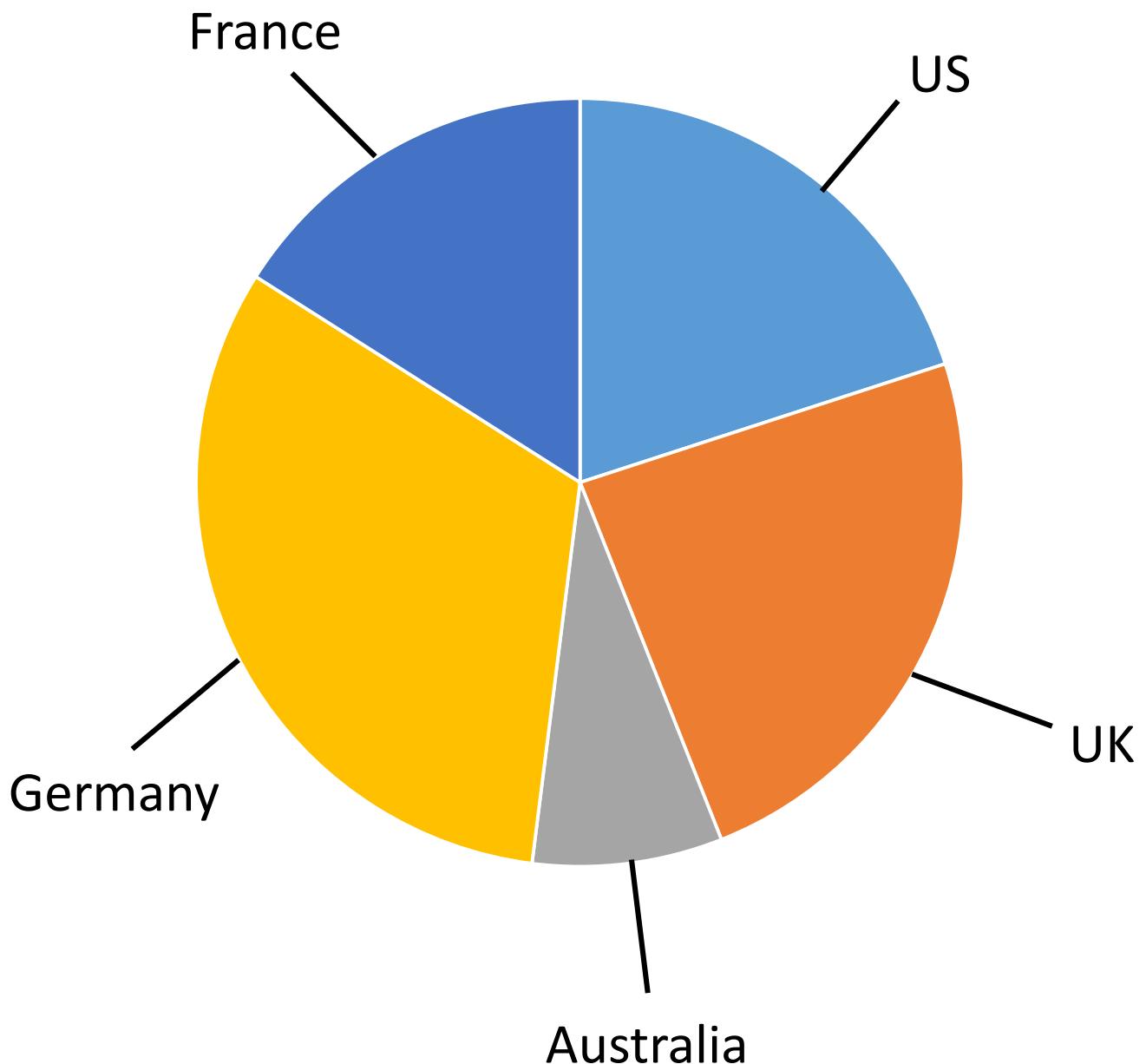
```
counts <- table(mtcars$vs, mtcars$gear)
barplot(counts,
        main="Car Distribution by Gears and VS",
        xlab="Number of Gears",
        col=c("grey","cornflowerblue"),
        legend = rownames(counts),
        beside=TRUE)
```

Car Distribution by Gears and VS



It is a type of graph in which a circle is divided into sectors, each representing a proportion of the whole.

Simple Pie Chart



Example:

```
slices <- c(10, 12, 4, 16, 8)
lbls <- c("US", "UK", "Australia", "Germany",
"France")
pie( slices, labels = lbls, main="Simple Pie Chart")
```

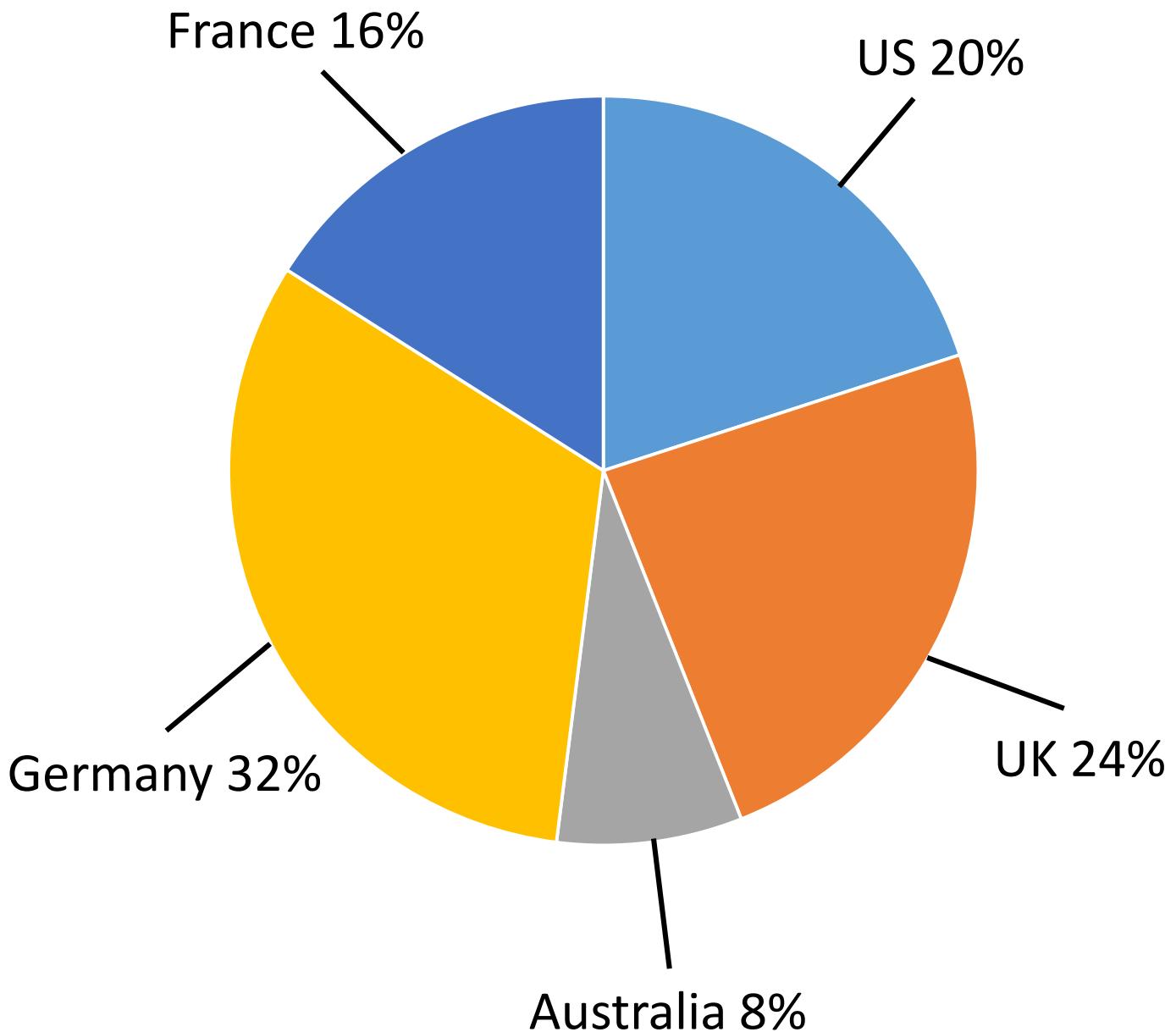
You can add percentages to a pie chart as shown below:



Example:

```
slices <- c(10, 12, 4, 16, 8)
pct <- round(slices/sum(slices)*100)
lbls <- paste(c("US", "UK", "Australia",
"Germany", "France"), " ", pct, "%", sep="")
pie(slices, labels=lbls2,
col=rainbow(5), main="Pie Chart with Percentages")
```

Pie Chart with Percentages



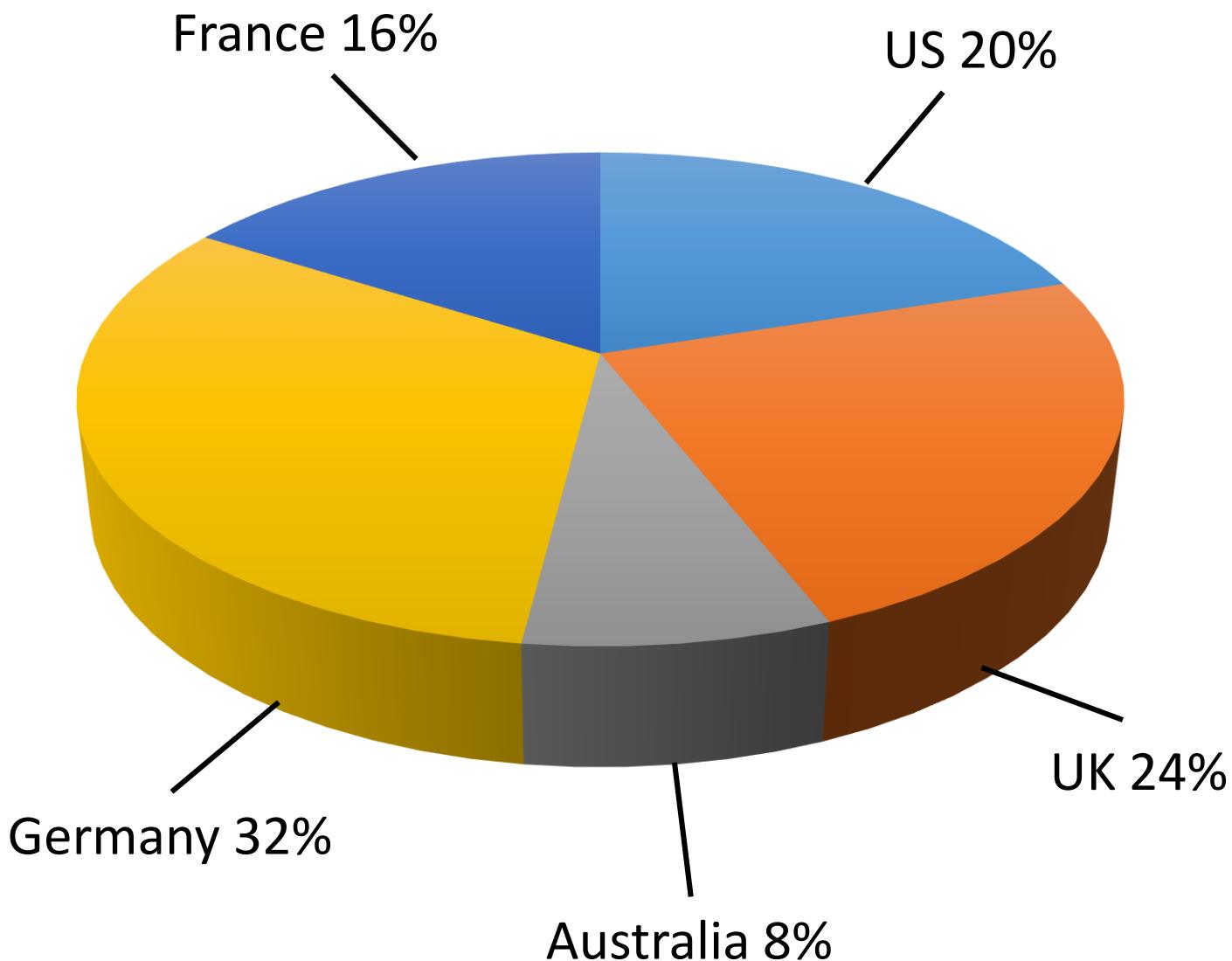
You can also make a 3-dimensional pie chart as shown below:



Example:

```
library(plotrix)
slices <- c(10, 12, 4, 16, 8)
lbls <- paste(
  c("US", "UK", "Australia", "Germany", "France"),
  " ", pct, "%", sep="")
pie3D(slices, labels=lbls, explode=0.0,
      main="3D Pie Chart")
```

3D Pie Chart



These display:

- The distribution of a continuous variable
- The frequency of scores in each bin on the y-axis by dividing the range of scores into bins on the x-axis

The syntax to create a histogram:

Syntax

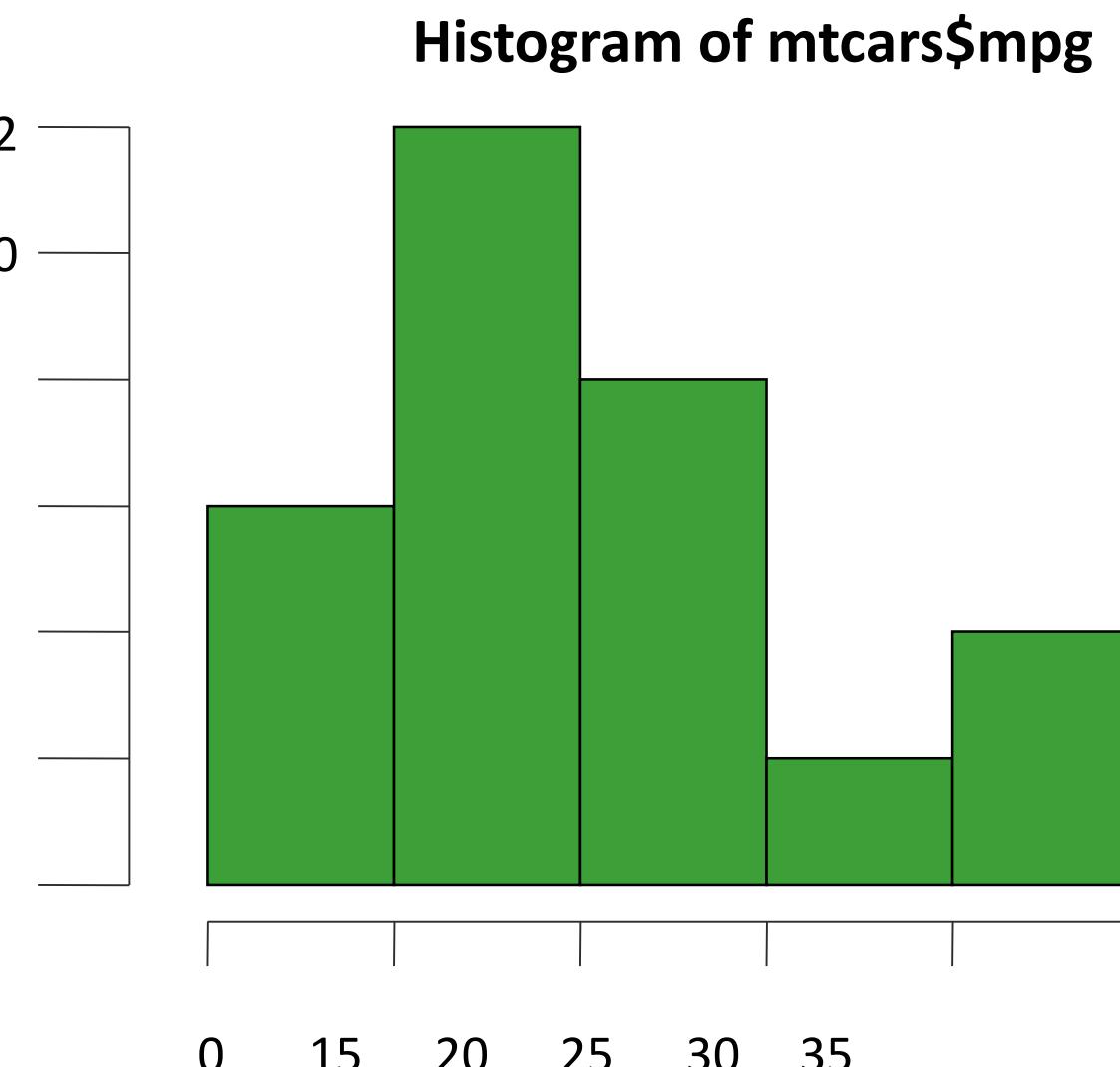
hist(x)

An example to create a histogram:

Example

```
mtcars$mpg #miles per gallon data  
hist(mtcars$mpg)
```

```
# Colored Histogram with Different  
Number of Bins  
hist(mtcars$mpg, breaks=8,  
col="darkgreen")
```



These display the distribution of a continuous variable much more efficiently than histograms.

The syntax to create a kernel density plot:

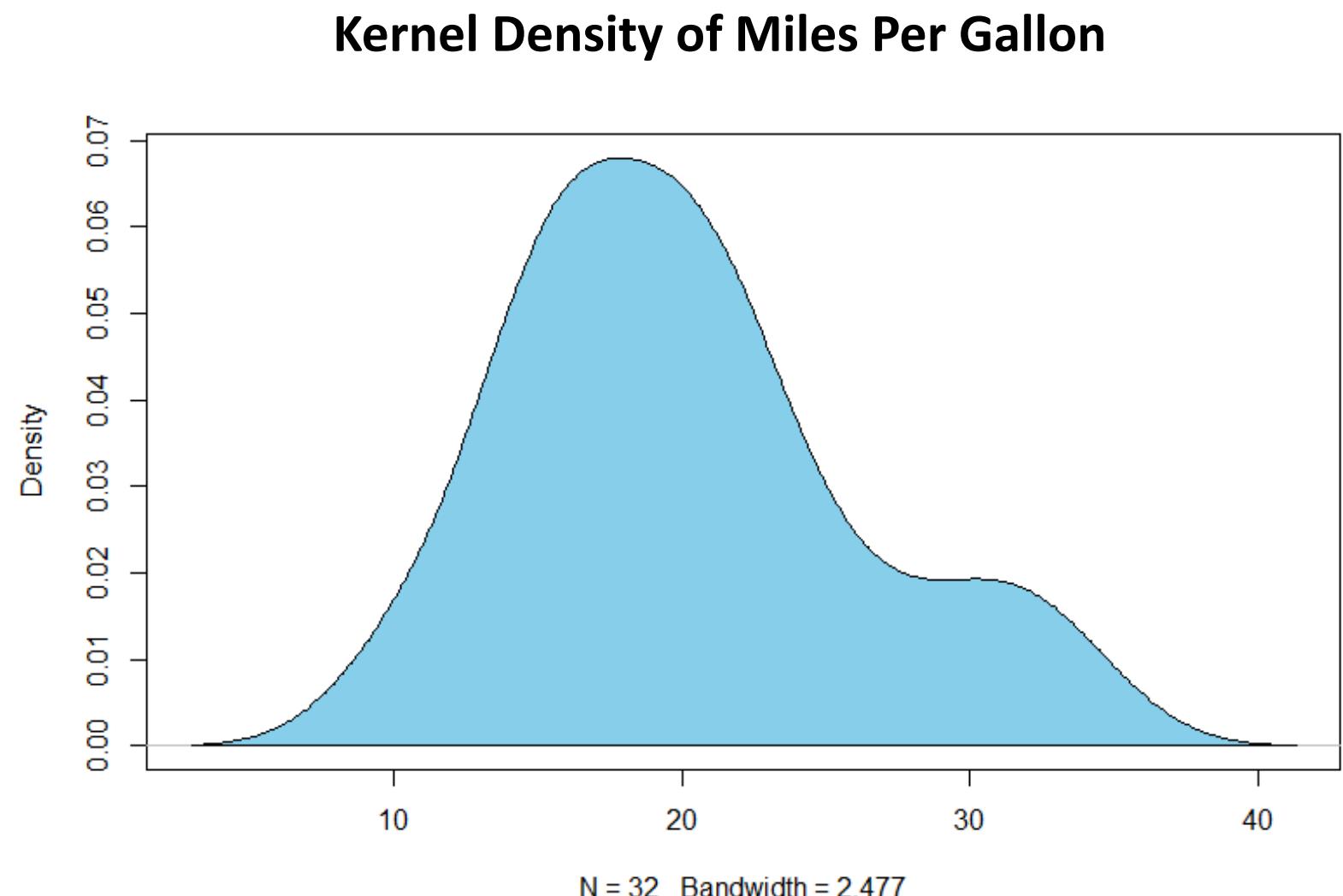
Syntax

plot(density(x))

An example to create a kernel density plot:

Example

```
# kernel Density Plot  
density_data <-  
density(mtcars$mpg)  
plot(density_data)  
  
# Filling density Plot with color  
density_data <-  
density(mtcars$mpg)  
plot(density_data, main="Kernel  
Density of Miles Per Gallon")  
polygon(density_data,  
col="skyblue", border="black")
```



Line Charts:

- Represent a series of data points connected by a straight line
- Are generally used to visualize data that changes over time

The syntax to create a line chart:

Syntax

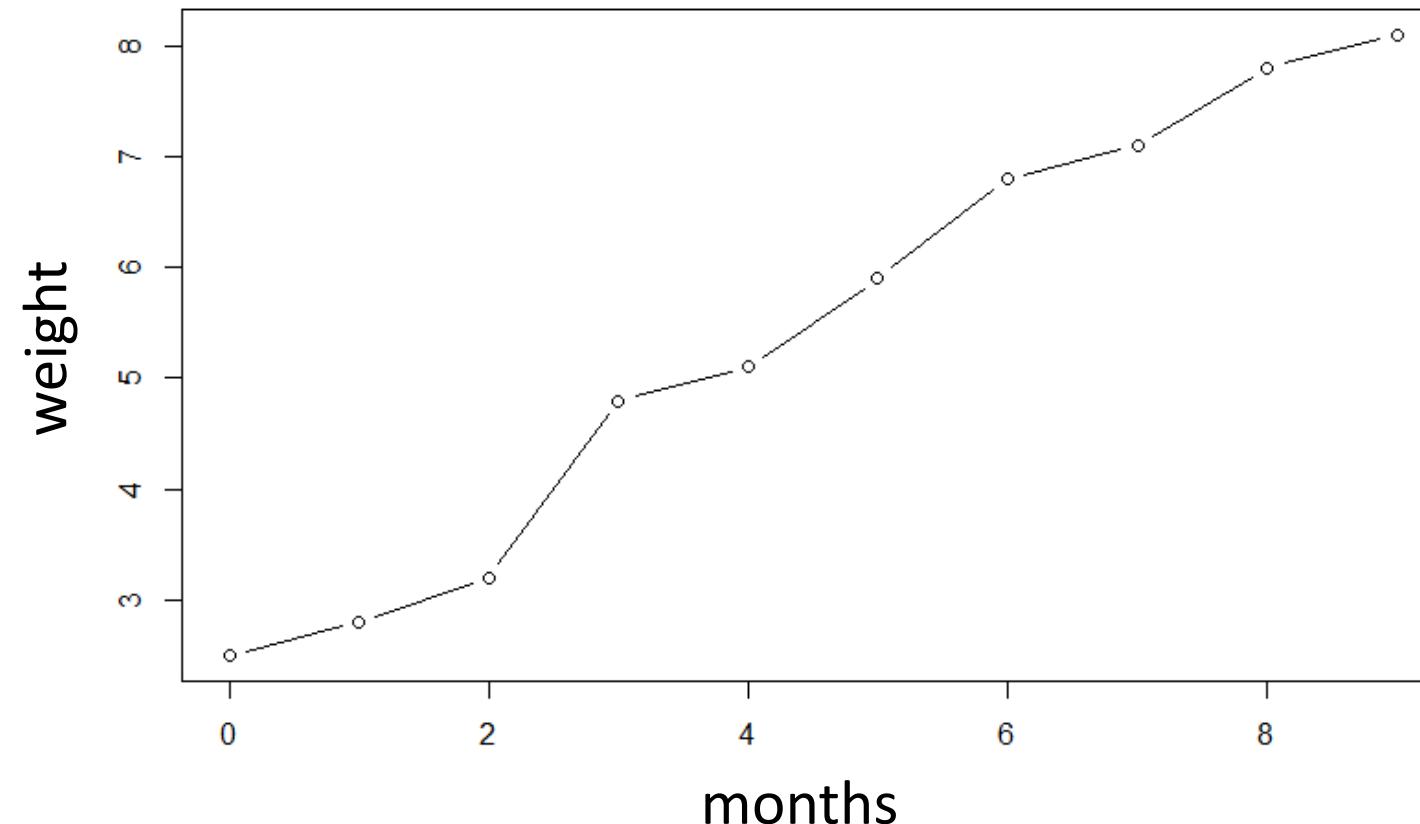
lines(x, y,type=)

An example to create a line chart:

Example

```
weight <- c(2.5, 2.8, 3.2, 4.8, 5.1,  
5.9, 6.8, 7.1, 7.8, 8.1)  
months <- c(0,1,2,3,4,5,6,7,8,9)  
plot(months,  
weight, type = "b",  
main="Baby Weight Chart")
```

Baby Weight Chart



Also known as whisker diagrams, these display the distribution of data that is based on the five-number summary:

- Minimum
- First quartile
- Median
- Third quartile
- Maximum

The syntax to create a box plot:

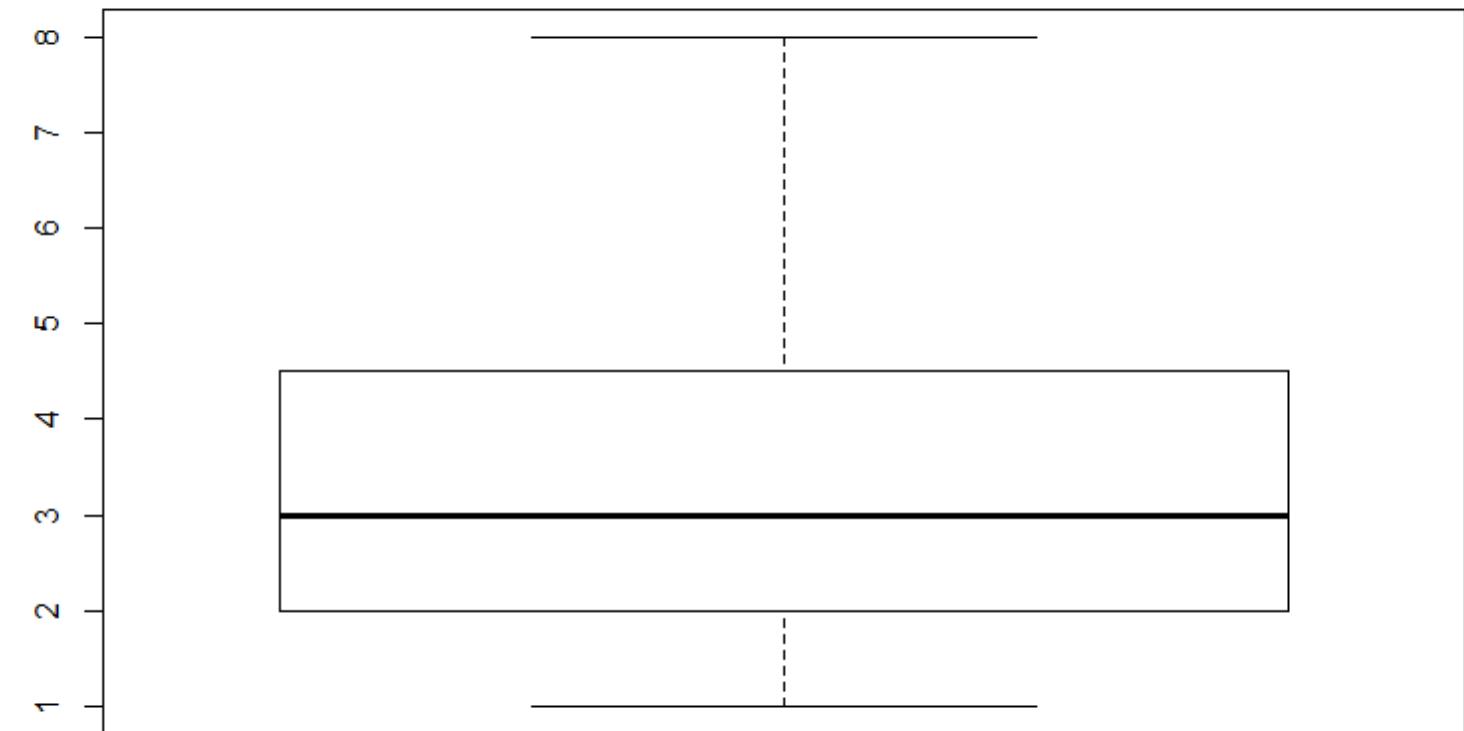
Syntax

boxplot(data)

An example to create a box plot:

Example

```
vec <- c(3, 2, 5, 6, 4, 8, 1, 2, 3, 2, 4)  
summary(vec)  
boxplot(vec, varwidth = TRUE)
```



These are two-dimensional representations of data in which the values are represented by colors. The two types of heat maps are:

Simple Heat Map

Provides an immediate visual summary of information

Elaborate Heat Map

Allows you to understand complex data sets

The syntax to create a heat map:

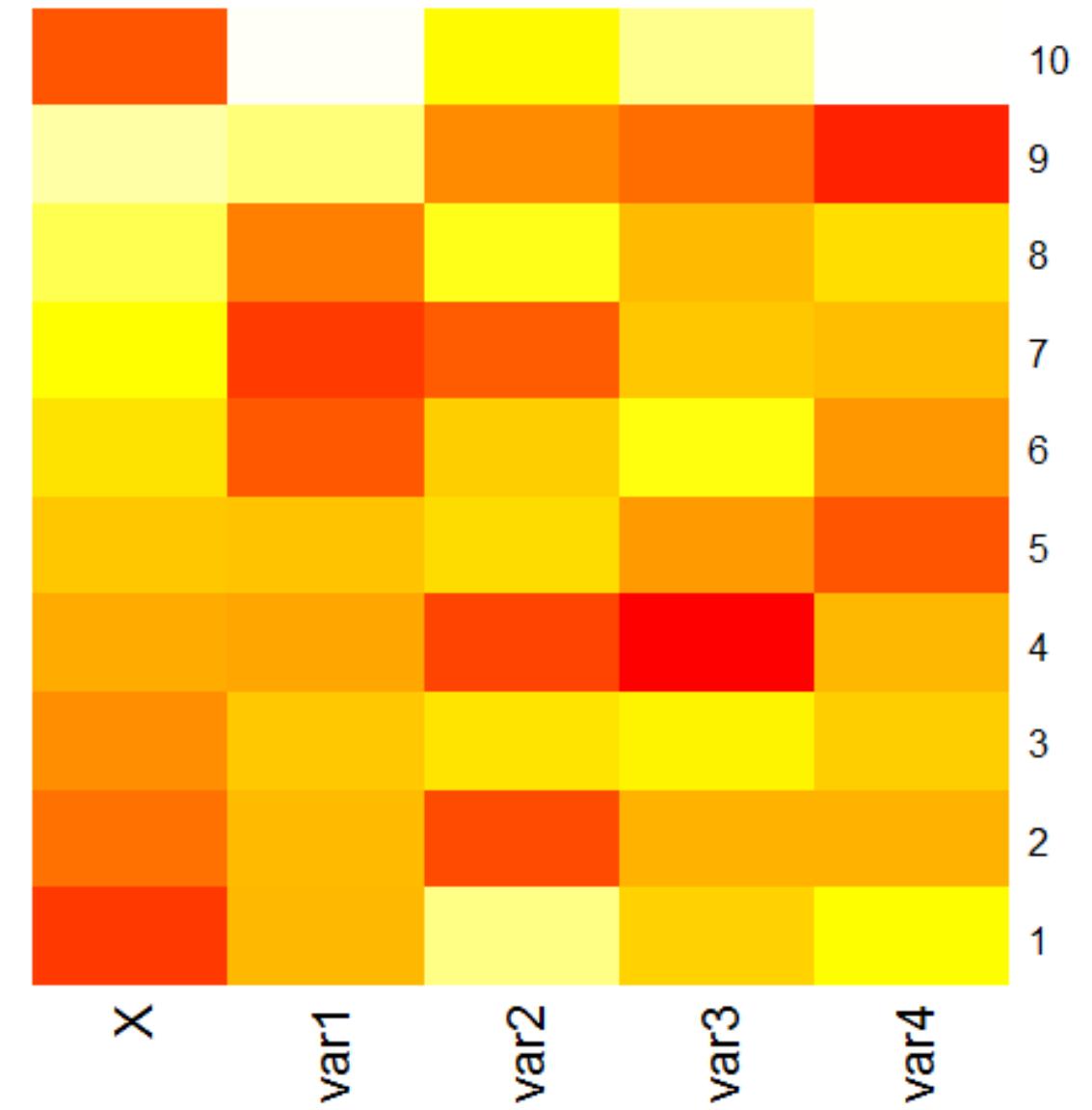
Syntax

heatmap(data, Rowv=NA, Colv=NA)

An example to create a heat map:

Example

```
data <-  
read.csv("HEATMAP.csv",header =  
TRUE)  
#convert Data frame into matrix  
data <- data.matrix(data[,-1])  
heatmap(data,Rowv=NA, Colv=NA,  
col = heat.colors(256),  
scale="column")
```



Also known as tag clouds, these highlight the most commonly cited words in a text using a quick visualization.

The syntax to create a world cloud:

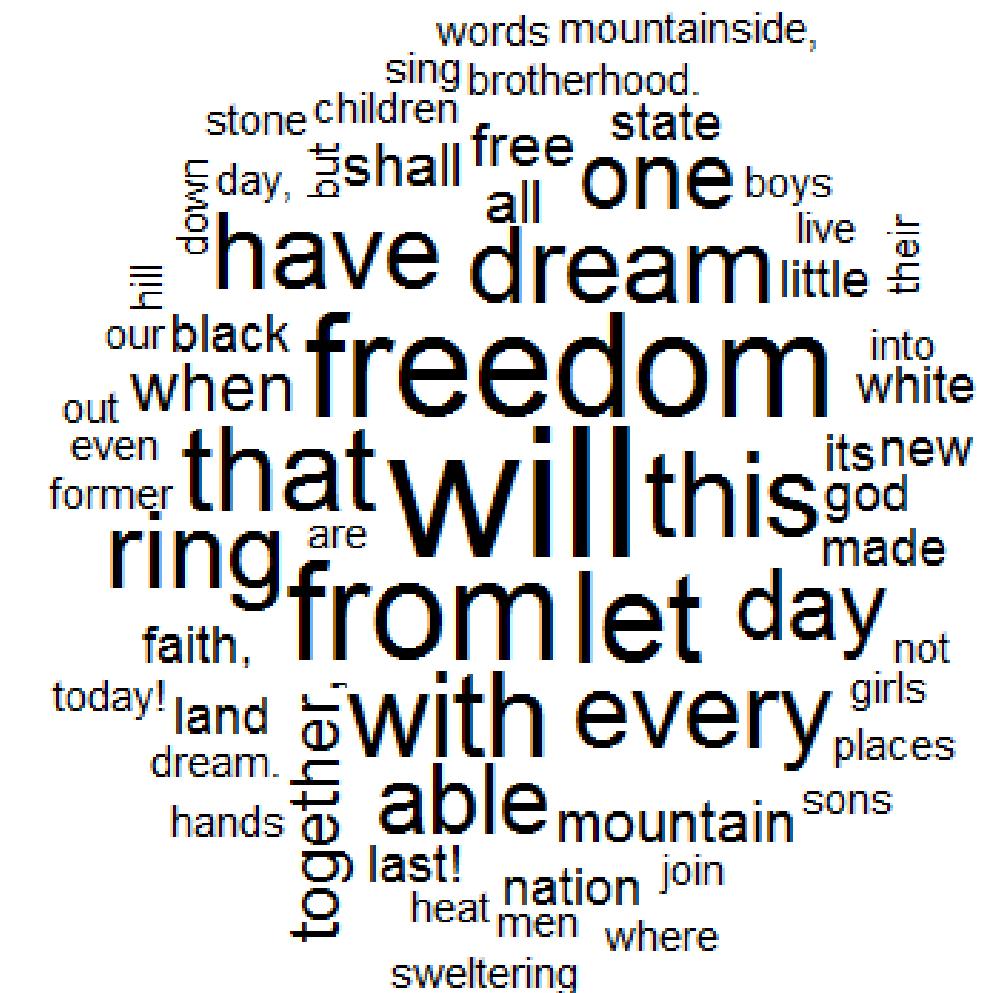
Syntax

```
wordcloud(words = data, freq = freq,  
min.freq = 2,)
```

An example to create a word cloud:

Example

```
install.packages("wordcloud")
library("wordcloud")
data <- read.csv("TEXT.csv",header
= TRUE)
head(data)
wordcloud(words = data$word,
freq = data$freq, min.freq = 2,
max.words=100,
random.order=FALSE)
```

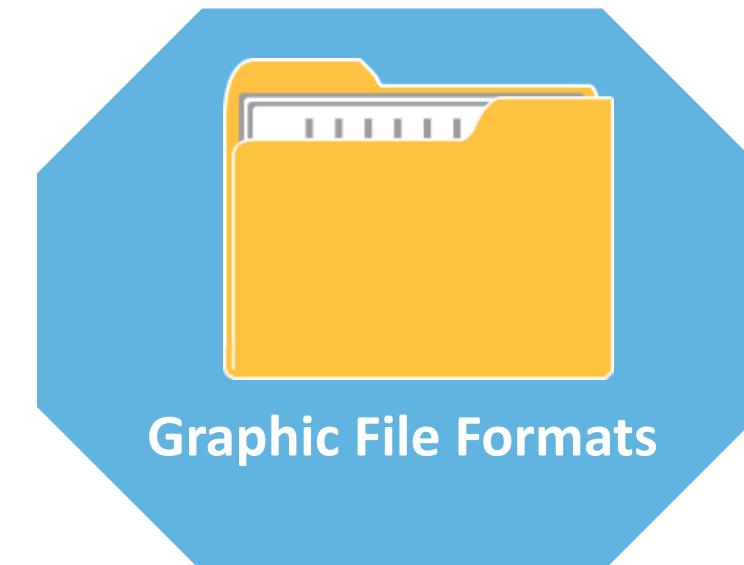


You can save the outputs of graphics in various file formats including:

`pdf("filename.pdf")`
#PDF file

`bmp("filename.bmp")`
#BMP file

`postscript("filename.ps")`
#PostScript file



`win.metafile("filename.wmf")`
#Windows metafile

`png("filename.png")`
#PNG file

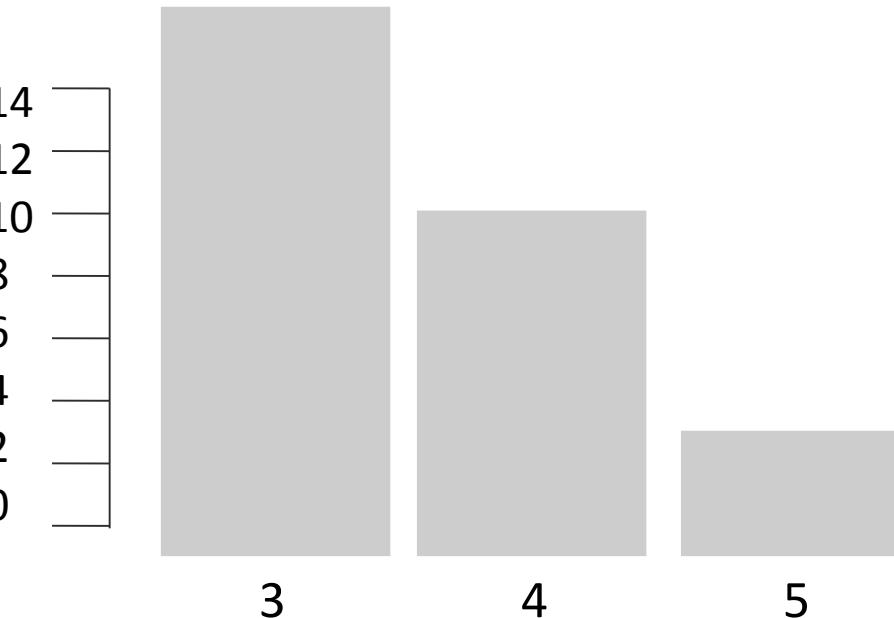
`jpeg("filename.jpg")`
#JPEG file

An example to save a graphic output as a file:



Example:

```
jpeg("myplot.jpg")
counts <- table(mtcars$gear)
barplot(counts)
dev.off()
```



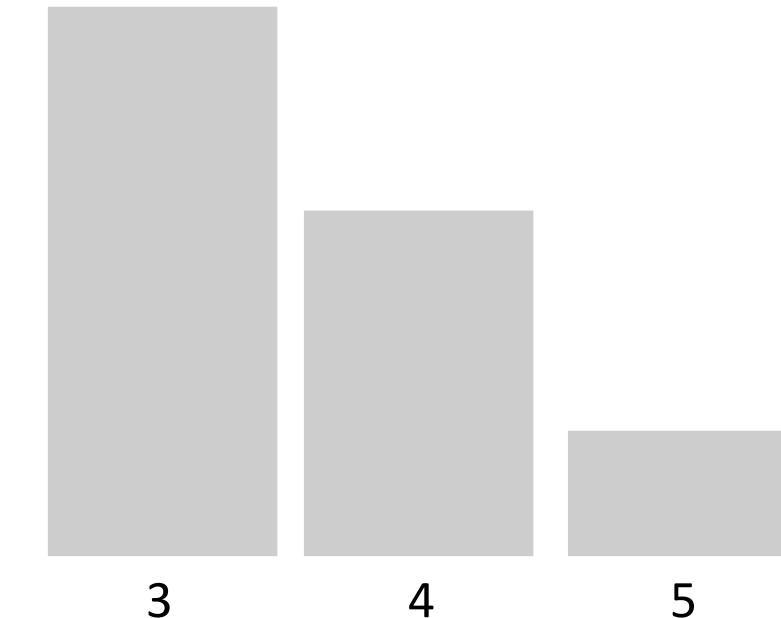
The dev.off() function returns the control back to the terminal.

Alternatively, you can save the same graphic output as a file as shown below:



Example:

```
dev.copy(jpeg, filename="myplot.jpg");
counts <- table(mtcars$gear)
barplot(counts)
dev.off()
```



Exporting Graphs in RStudio

To export graphs in RStudio, follow these steps:

1

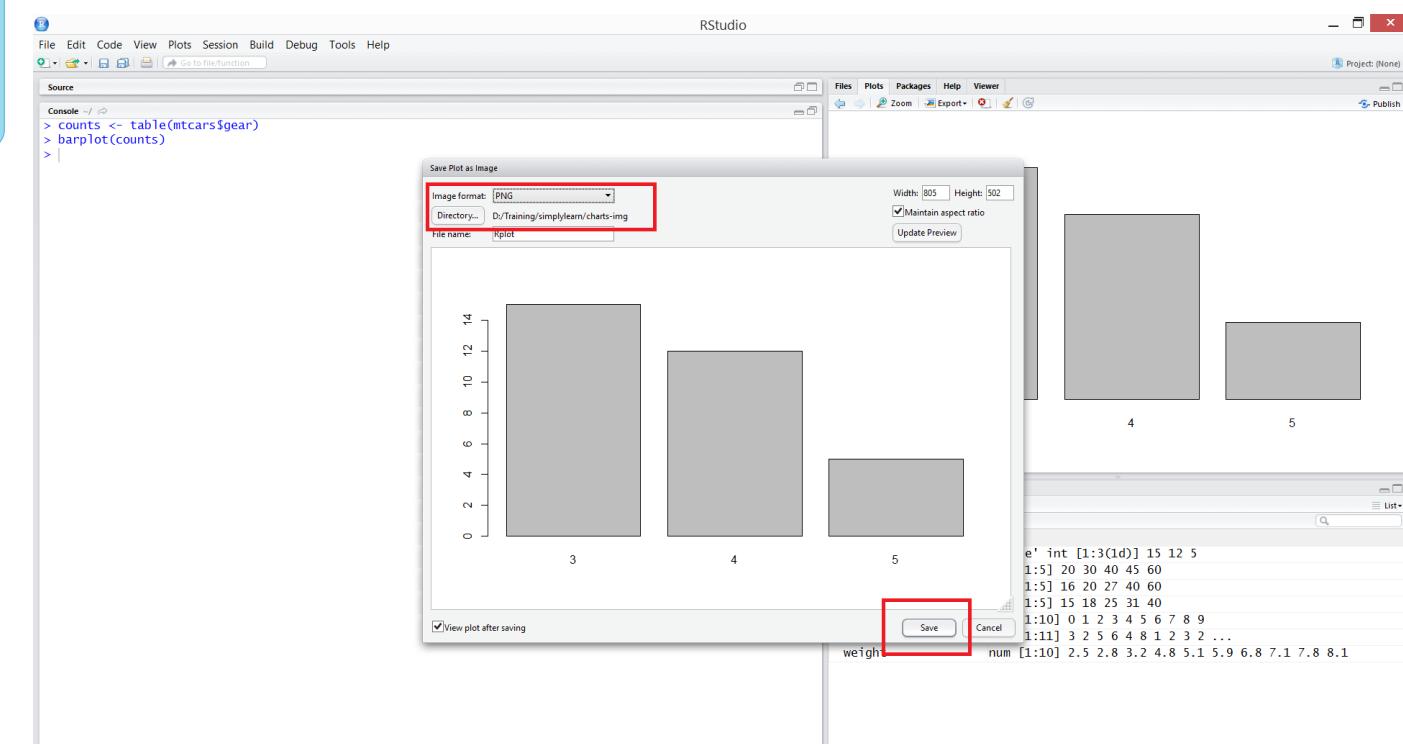
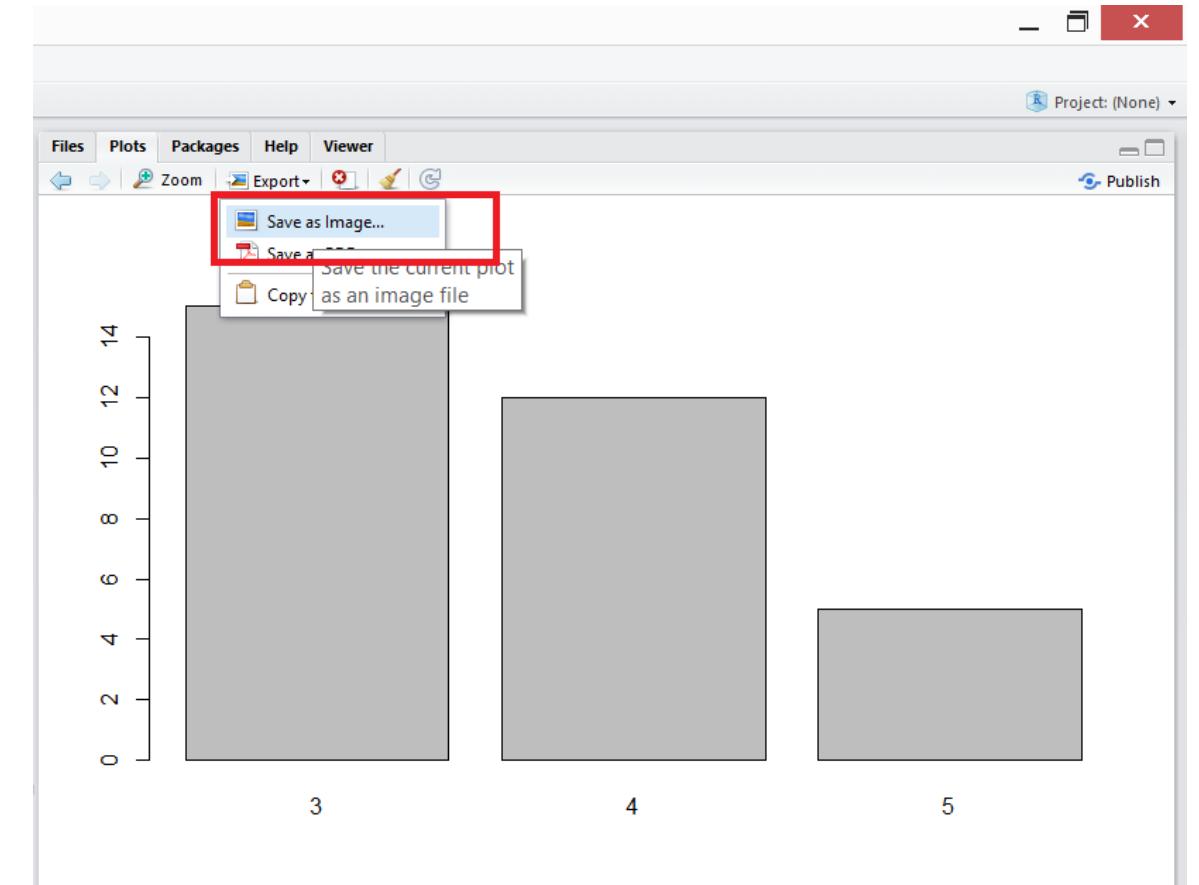
In the **Graphics** panel of RStudio, click the **Plots** tab, click the **Explore** menu, and select the **Save as Image** menu item.

2

The **Save Image** window is displayed. Select the image format and directory to save the file.

3

Click the **Save** button.



Exporting Graphs as PDFs in RStudio

Follow these steps to export graphs as PDFs in RStudio:

1

In the **Graphics** panel of RStudio, click the **Export** menu and select the **Save as PDF** menu item.



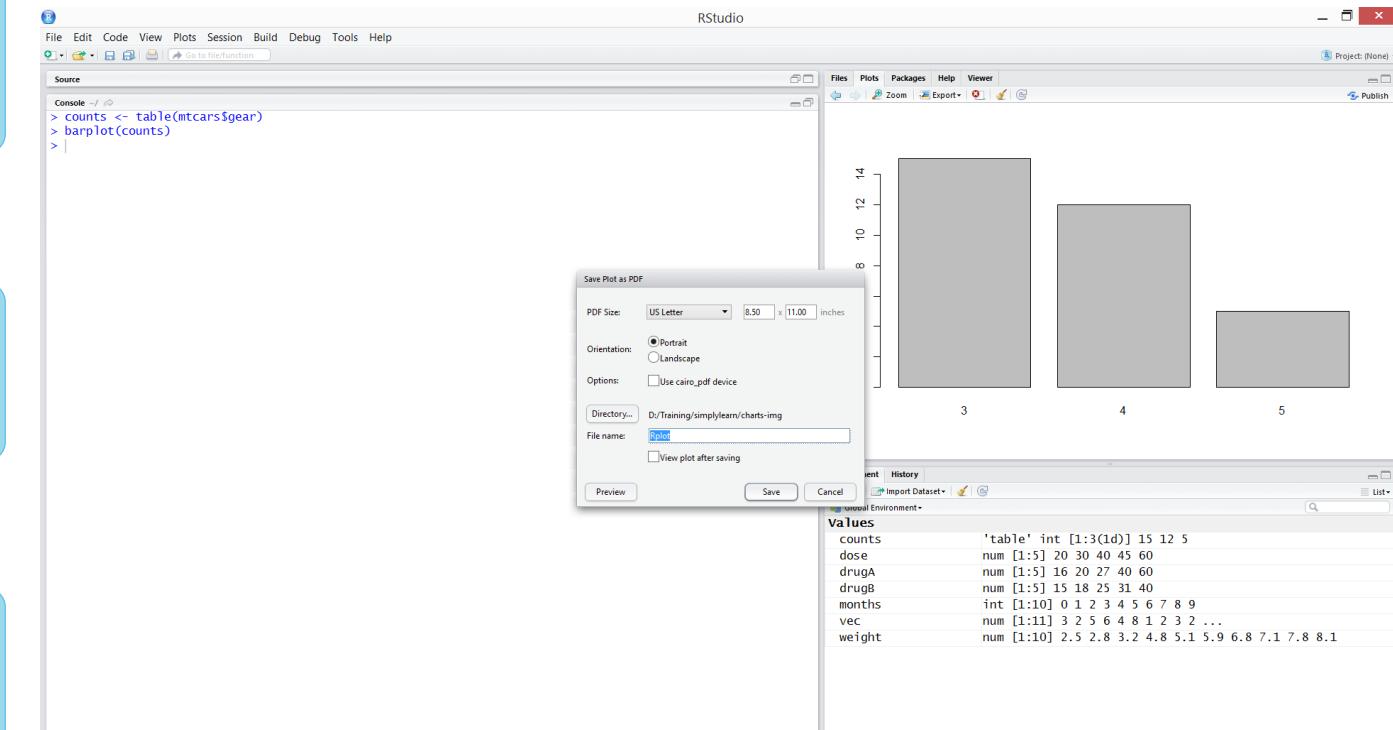
2

Select the directory to save the file.



3

Click the **Save** button.





**QUIZ
1**

Which of the following graphics represent lengths, frequency, or proportion of categorical values?

- a. Line charts
- b. Bar plots
- c. Bar charts

Kernel density plots



QUIZ
1

Which of the following graphics represent lengths, frequency, or proportion of categorical values?

- a. Line charts
- b. Bar plots
- c. Bar charts

Kernel density plots



The correct answer is **c.**

Explanation: Bar charts represent lengths, frequency, or proportion of categorical values.

**QUIZ
2**

Which of the following graphics displays the distribution of data that is based on the five-number summary?

- a. Line charts
- b. Bar plots
- c. Bar charts

Kernel density plots



QUIZ
2

Which of the following graphics displays the distribution of data that is based on the five-number summary?

- a. Line charts
- b. Bar plots
- c. Bar charts

Kernel density plots



The correct answer is **b**.

Explanation: Bar plots display the distribution of data that is based on the five-number summary.

QUIZ
3

State whether the following statement is True or False.

Histograms display the distribution of a continuous variable.

- a. True
- b. False



QUIZ
3

State whether the following statement is True or False.

Histograms display the distribution of a continuous variable.

- a. True
- b. False



The correct answer is **a.**

Explanation: Histograms display the distribution of a continuous variable.

QUIZ

4

Fill in the blank:

Graphic outputs can be saved as files using the _____ function.

- a. `save("filename.png")`
- b. `write.table("filename.png")`
- c. `write.file("filename.png")`

`png("filename.png")`



QUIZ

4

Fill in the blank:

Graphic outputs can be saved as files using the _____ function.

- a. `save("filename.png")`
- b. `write.table("filename.png")`
- c. `write.file("filename.png")`

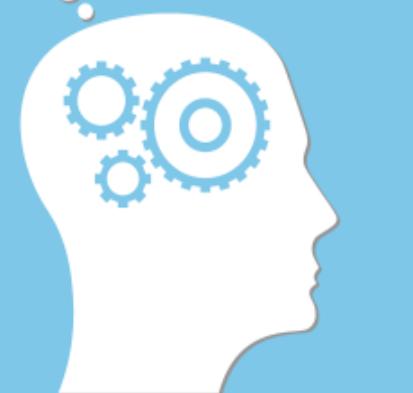
`png("filename.png")`

The correct answer is **d**.

Explanation: Graphic outputs can be saved as files using the `png("filename.png")` function.



Let us summarize the topics covered in this lesson:



- R includes powerful packages of graphics that help in data visualization:
 - Bar chart
 - Pie chart
 - Histogram
 - Kernel density plot
 - Line chart
 - Box plot
 - Heat map
 - World cloud
- You can save the outputs of graphics in various file formats, such as .pdf, .png, and .jpeg.
- You can export graphs in Rstudio by using the Graphics panel.

This concludes “Data Visualization.”

The next lesson is “Introduction to Statistics.”

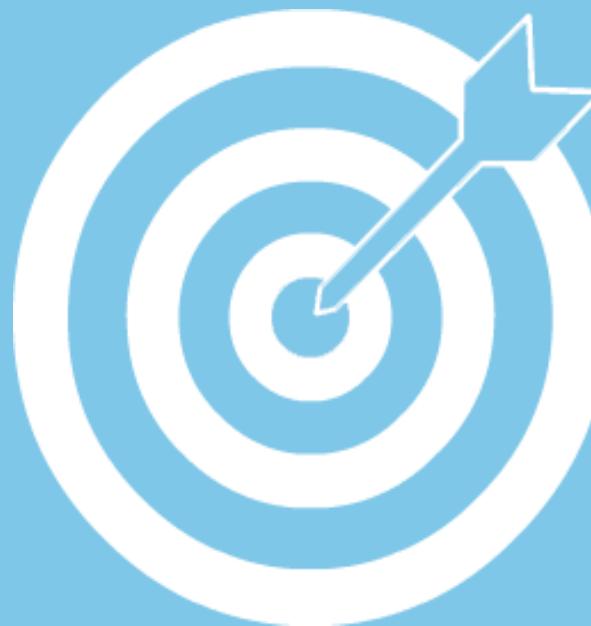


Data Science with R

Lesson 07—Introduction to Statistics



After completing
this lesson, you will
be able to:



- Explain the two types of data
- Describe the types of measurements
- List the steps of statistical investigation
- Discuss the importance of normal distribution in statistics
- Explain the various distance measures
- Describe the various types of correlations
- Explain the use of the dist() function

Let's first understand the basic concepts of statistics.



Statistical Population

A collection of all probable observations of a specific characteristic of interest

Example: All learners taking this course



Sample

A subset of population

Example: A group of 20 learners selected for a quiz



Variable

An item of interest that can acquire various numerical values

Example: The number of defective items manufactured in a factory

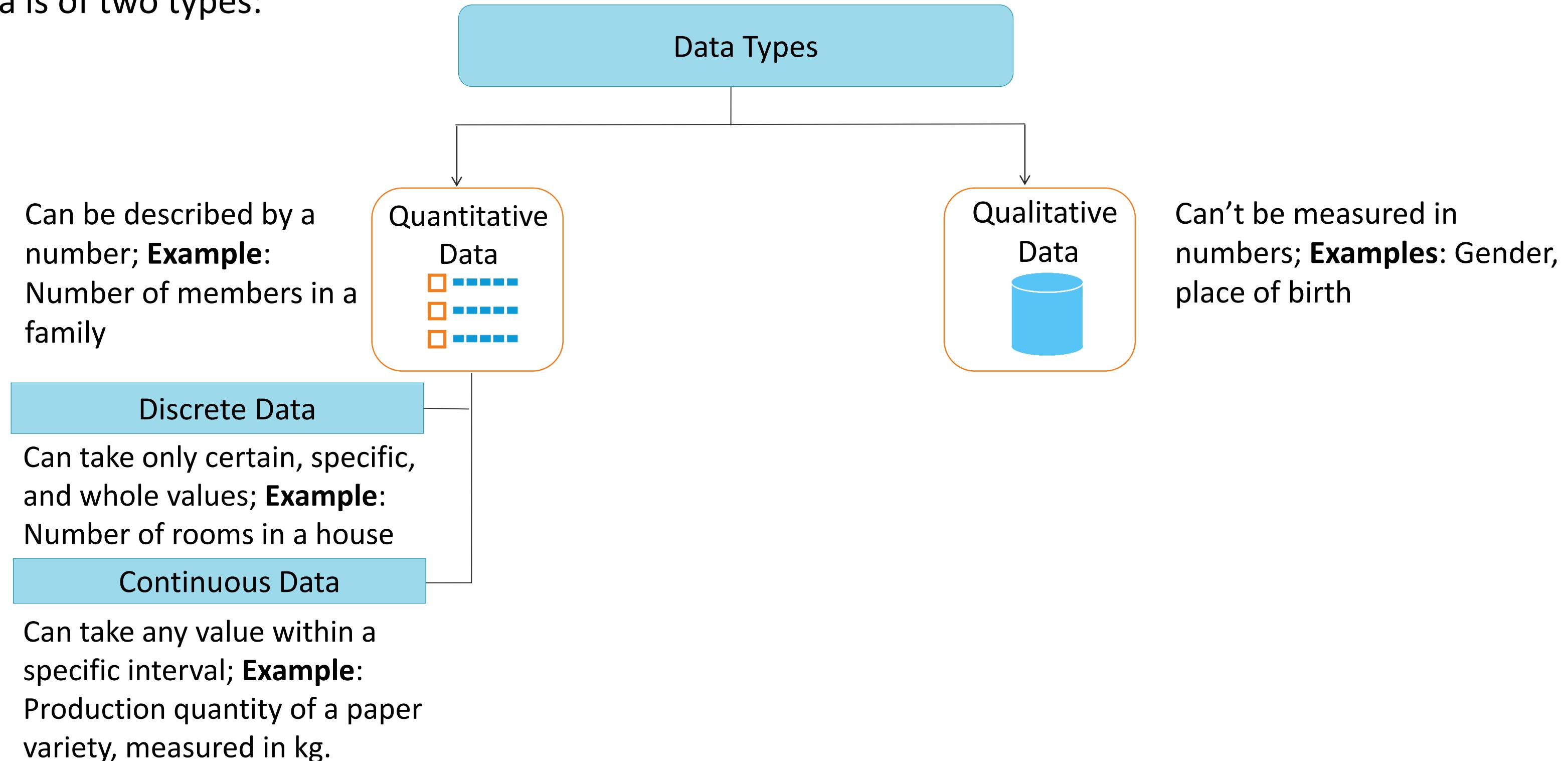


Parameter

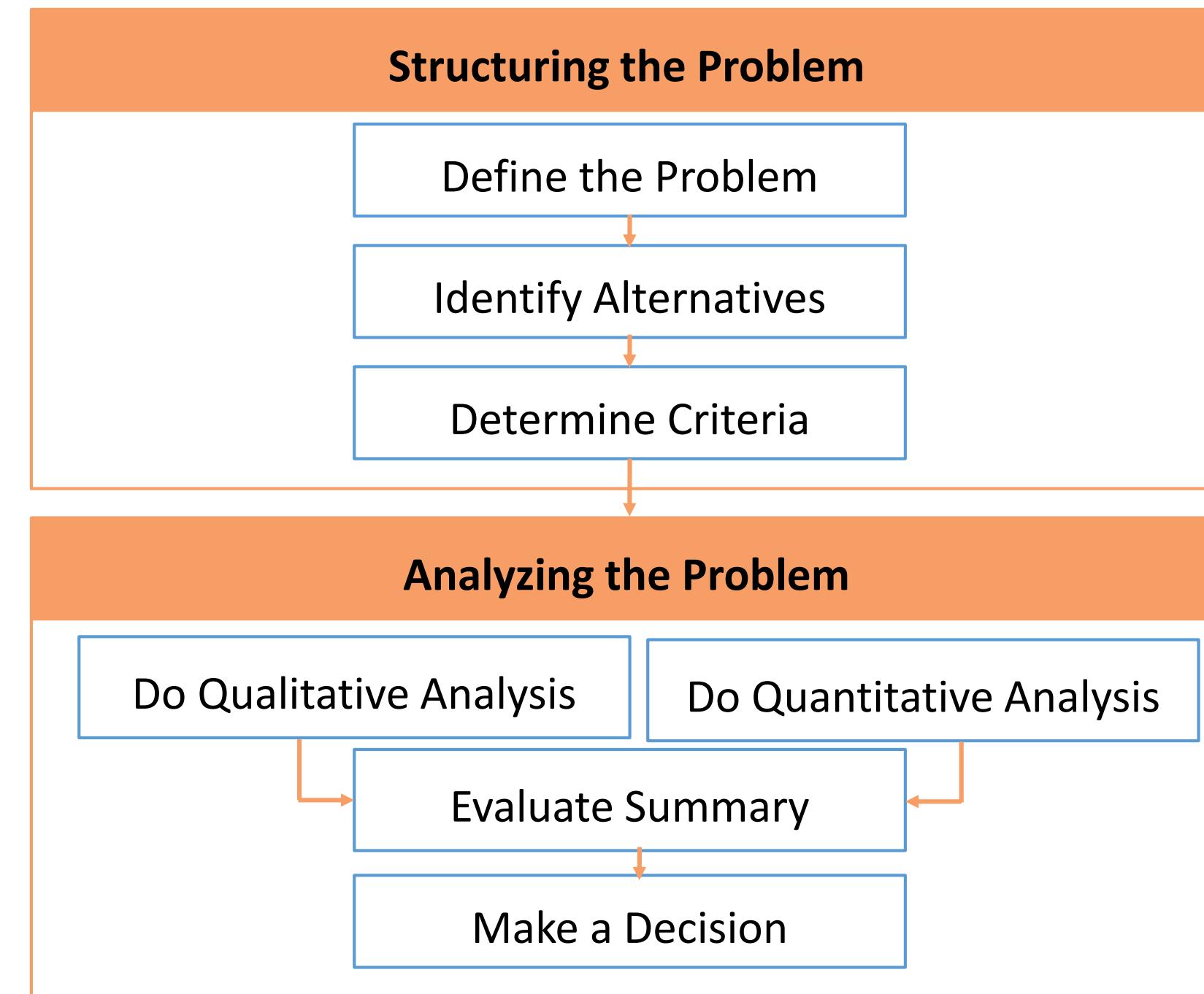
A population characteristic of interest

Example: The average income of a class of people

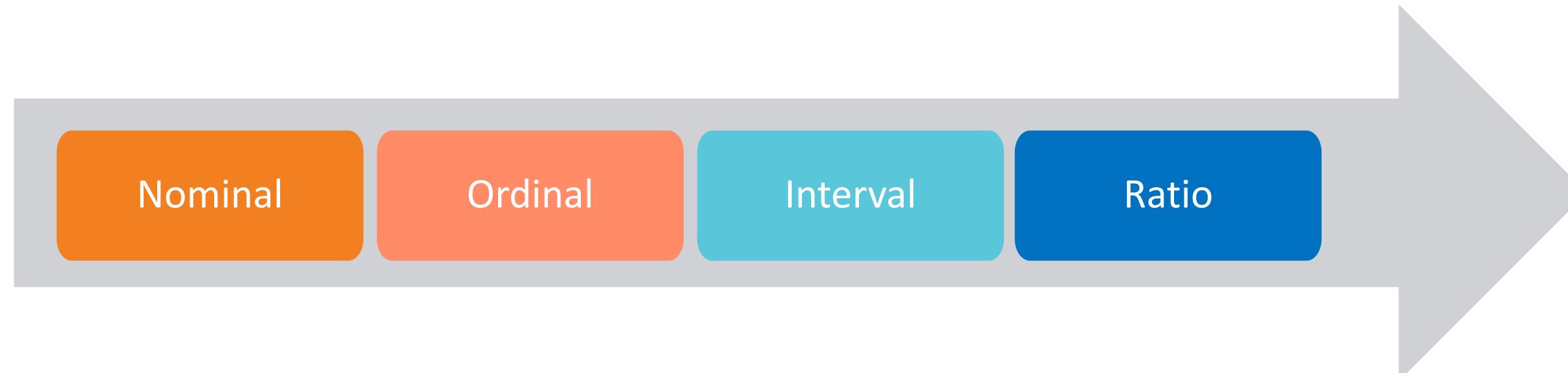
Data is of two types:



You can understand this comparison with the help of the diagram below:



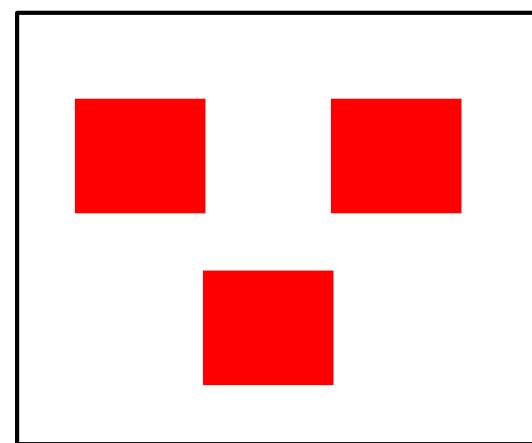
The types of measurements in their increasing order of content are mentioned below:



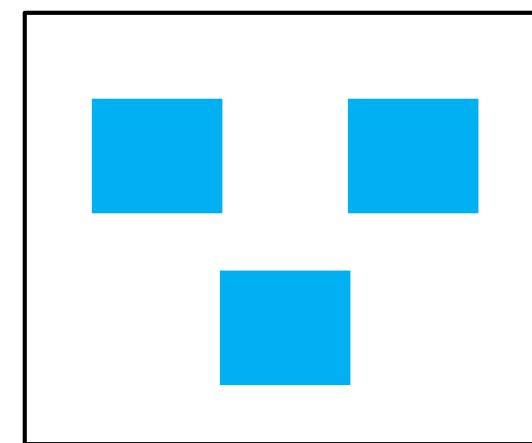
In this measurement:

- Numbers are used to label qualitative items (not quantitative data) in classes or groups.
- The main purpose is to categorize items.
- Values cannot be altered in a numerical manner.
- Arithmetic operations are not possible.

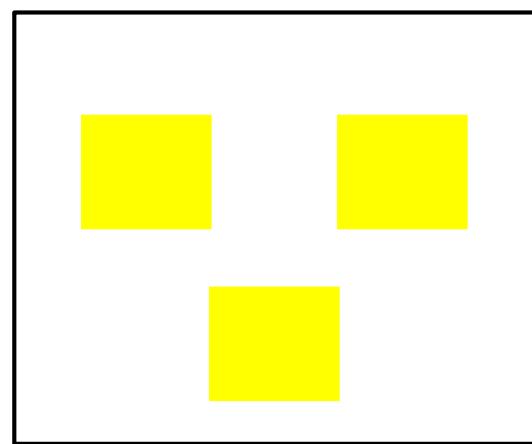
Example: Numbers are used to denote different colored items in a data set.



1



2

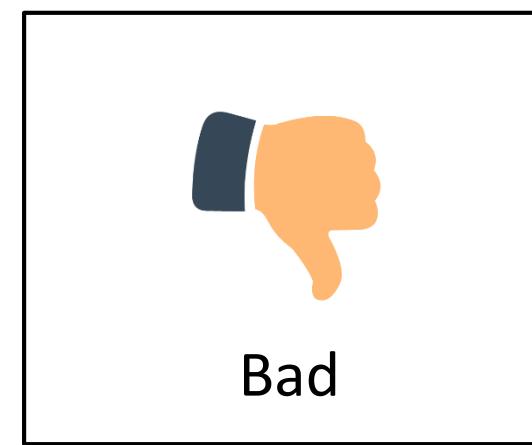


3

In this measurement:

- Numbers are used to rank objects or attributes per their relative size or quality.
- Difference between ranks cannot be measured.
- Arithmetic operations are not possible.

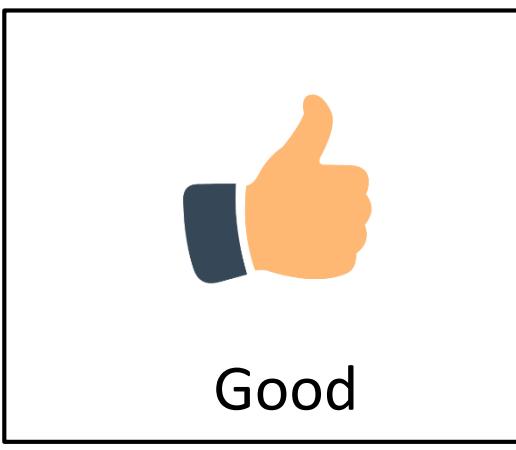
Example: Numbers are used to denote customer preferences for a product.



1



2

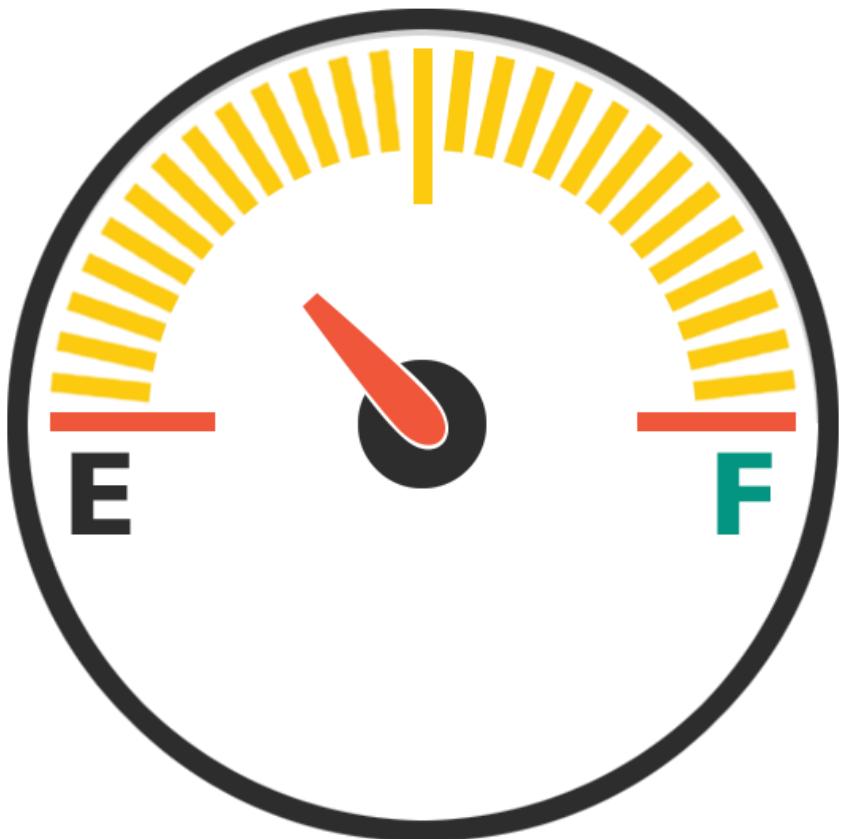


3

In this measurement:

- Values have ordinal properties.
- Difference between ranks can be measured.
- Values have an arbitrary zero point.
- Basic arithmetic operations are possible.

Example: Measurement of the time and temperature



In this measurement:

- Data measured on a ratio scale have a fixed zero point.
- You can take ratios of two measurements that are in a ratio scale.
- Basic arithmetic operations are possible.

Example: Measurement of money, weight, volume, area or length, cost, and profit.

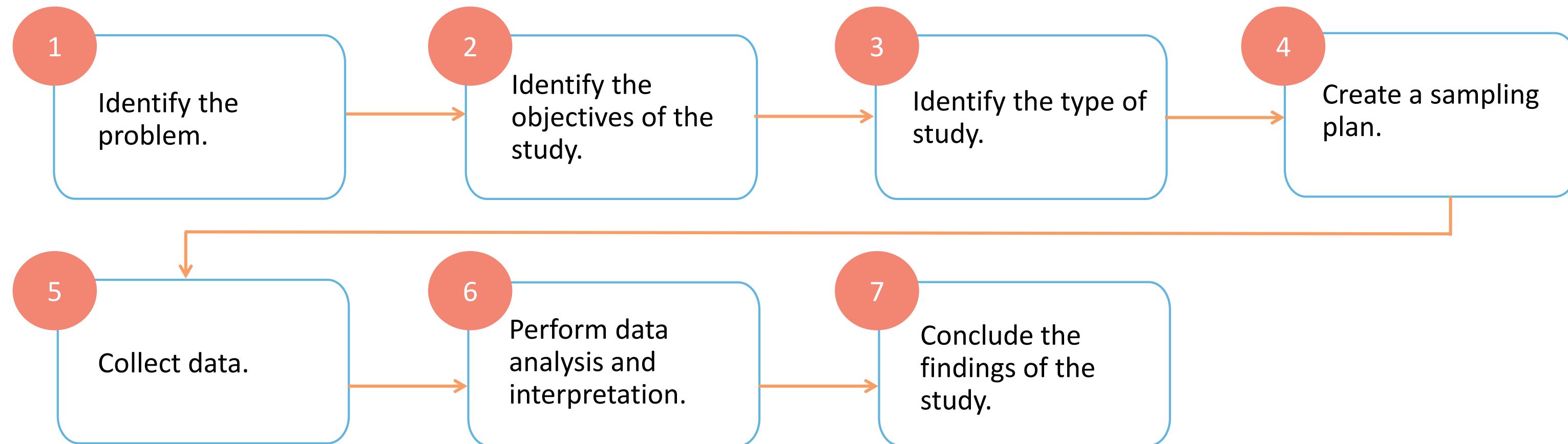


Statistical investigation:

- Lays the structure for decision making
- Provides answers to various management problems
- Provides a set of recommendations to develop and implement the best strategies

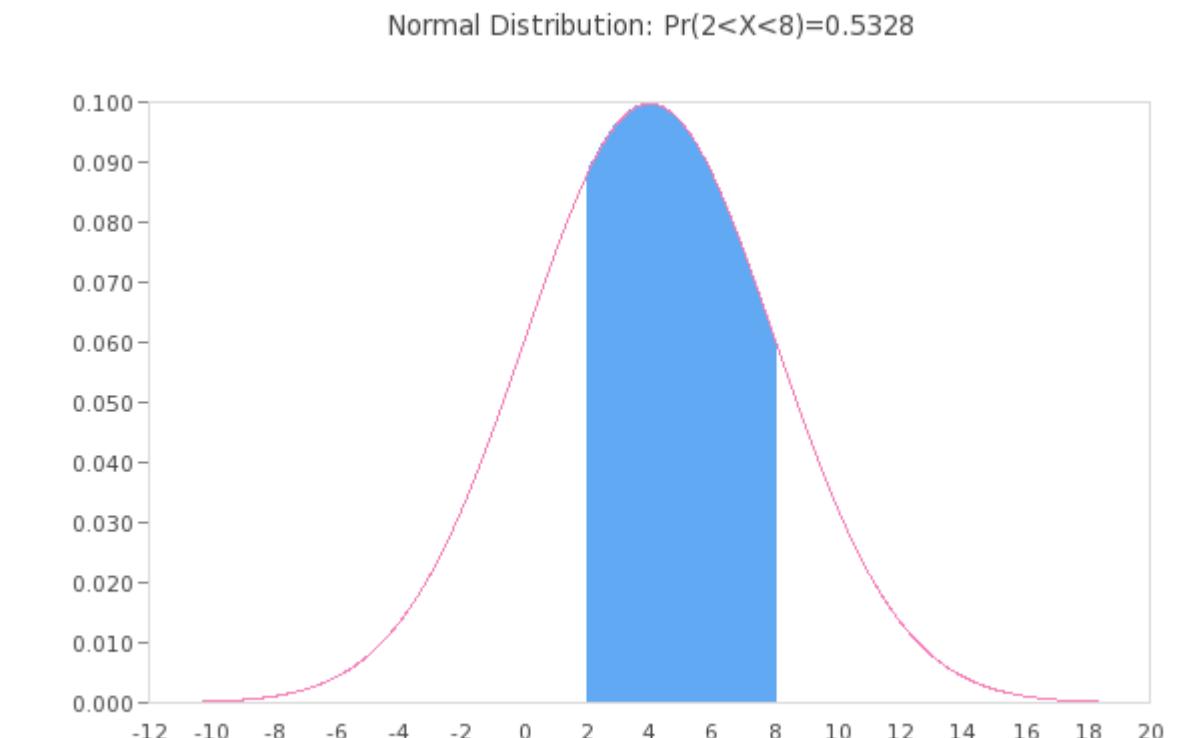


The steps to perform statistical investigation:



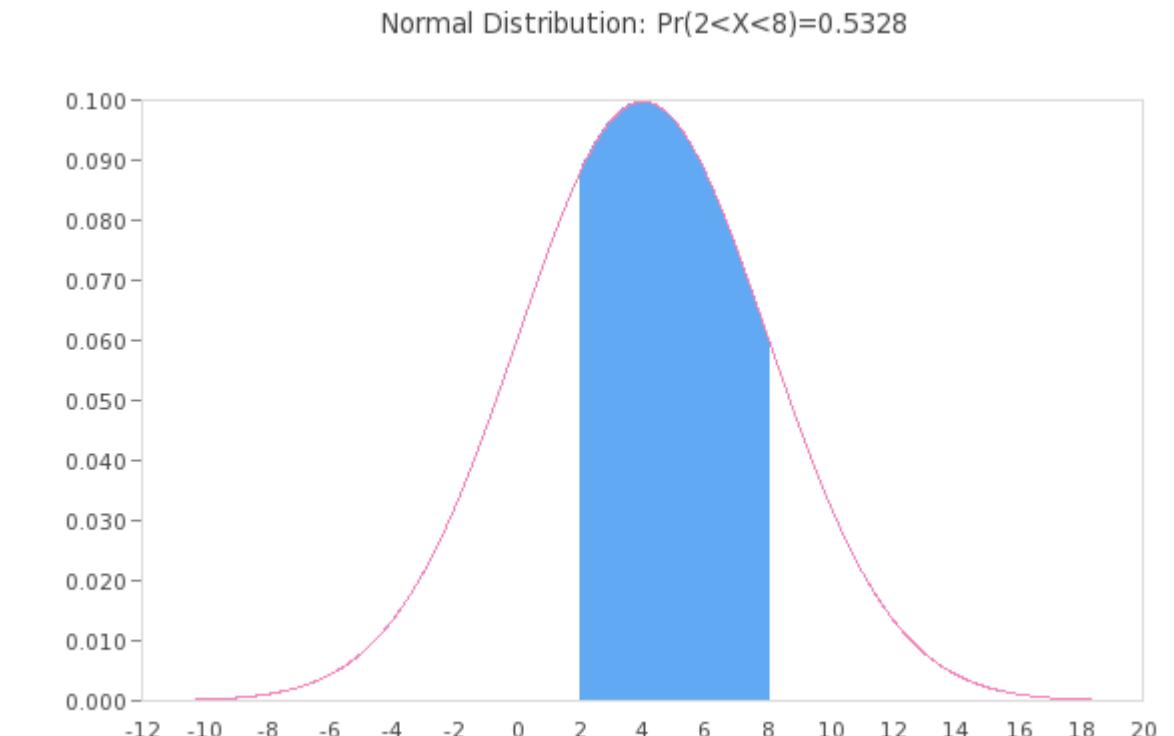
Following are the properties of normal distribution:

- It is a family of bell-shaped and symmetric distributions.
- Each distribution is characterized by a different pair of mean, μ , and variance, σ^2 — $[X \sim N(\mu, \sigma^2)]$.
- Each distribution is asymptotic to the horizontal axis.
- The area under any normally distributed function will be same, regardless of the mean and variance.



A few more properties of normal distribution:

- Measures of central tendency are all identical.
- Its “middle spread” is equal to 1.33 standard deviation.
- Its associated variable has an infinite range ($-\infty < x < \infty$).
- If several independent and random variables are normally distributed, then their sum is also normally distributed.
- The mean of the sum is the sum of all individual means.
- The variance of the sum is the sum of all individual variances.



Assume X_1, X_2, X_3 , and X_4 are independent random variables that are normally distributed with means and variances as shown in the table below:

Variable	Mean	Variance
X_1	12	4
X_2	-5	2
X_3	8	5
X_4	10	1

Let's find the mean and variance of $Q = X_1 - 2X_2 + 3X_3 - 4X_4 + 5$.

$$E(Q) = 12 - 2(-5) + 3(8) - 4(10) + 5 = 11$$

$$V(Q) = 4 + (-2)^2(2) + 3^2(5) + (-4)^2(1) = 73$$

$$SD(Q) = \sqrt{73} = 8.544$$

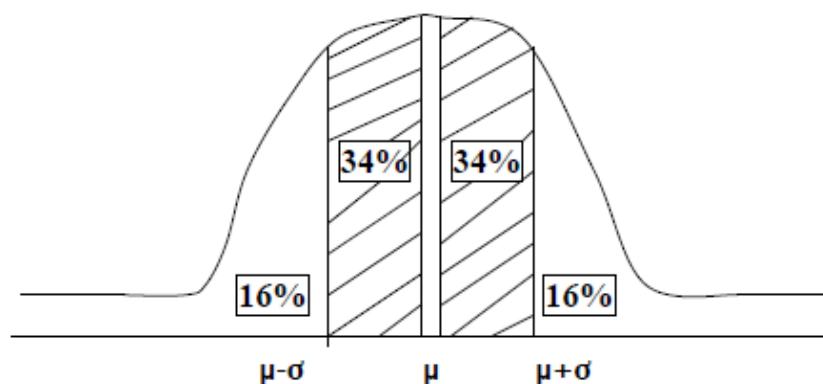
This can be explained as follows:

- Various continuous variables seem to follow normal distribution or can be approximated by normal distribution.
- It can be used to approximate various discrete probability distributions.
- It provides the basis for inferential statistics.



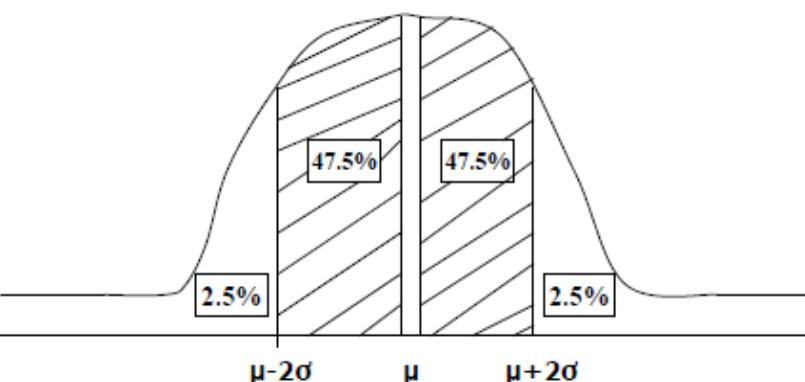
Area under a normal curve has three distinct positions as explained below:

Within one Standard Deviation from mean



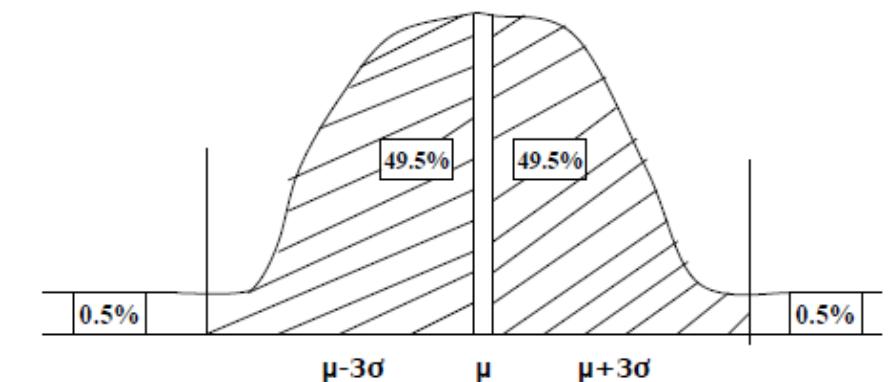
The total area covered within one SD from mean is 68%, which gets divided into 34% on either side of the mean.

Within two Standard Deviations from mean



The total area covered within two SD from mean is 95%, which gets divided into 47.5% on either side of the mean.

Within three Standard Deviations from mean



The total area covered within three SD from mean is 99%, which gets divided into 49.5% on either side of the mean.

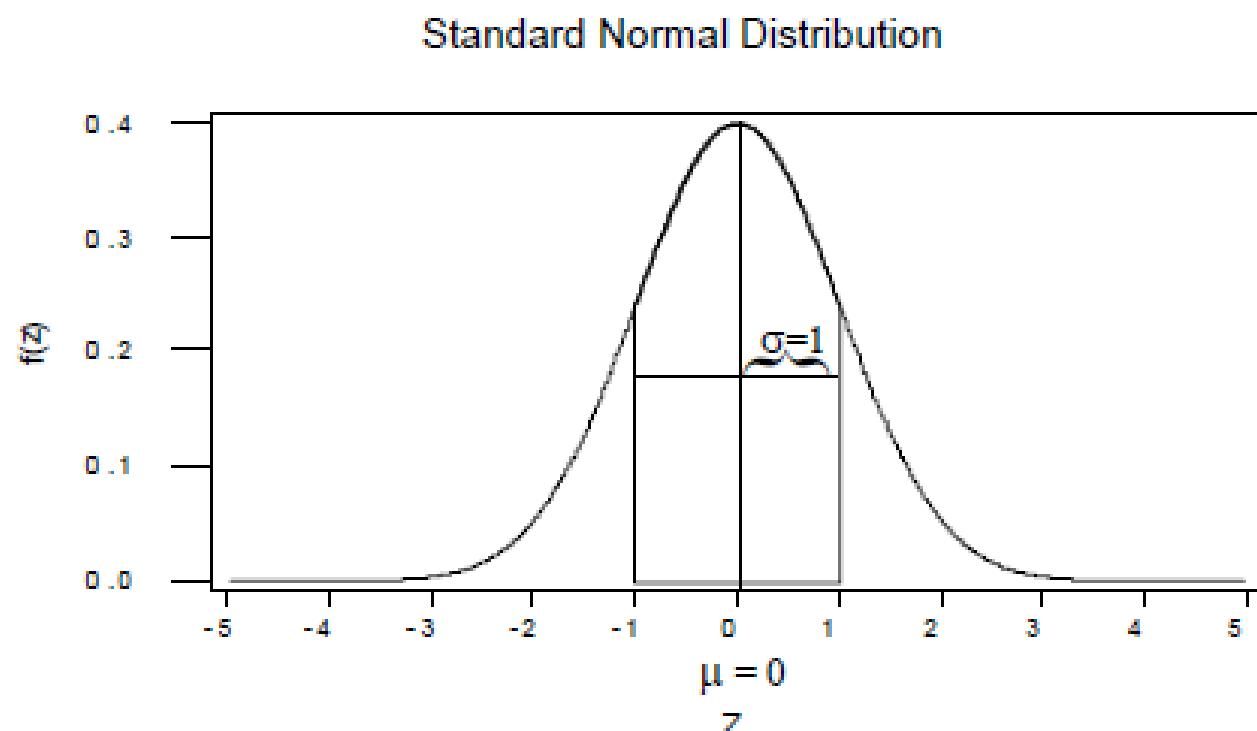
You can convert any normal random variable X to a standard normal variable Z with mean 0 and variance 1 using the given formula:

$$Z = (X - \mu) / \sigma$$

Where,

$$X \sim N(\mu, \sigma^2)$$

Z is the normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$: $Z \sim N(0, 1)$:



These are mathematical methods for measuring the distance between two or more objects. Calculating these measures helps to compare the objects from different perspectives, such as:

Similarity
Measures

Ranges from 0 to 1 [0, 1]

Dissimilarity
Measures

Ranges from 0 to INF [0, Infinity]

Correlation
Measures

Ranges from +1 to -1 [+1, -1]



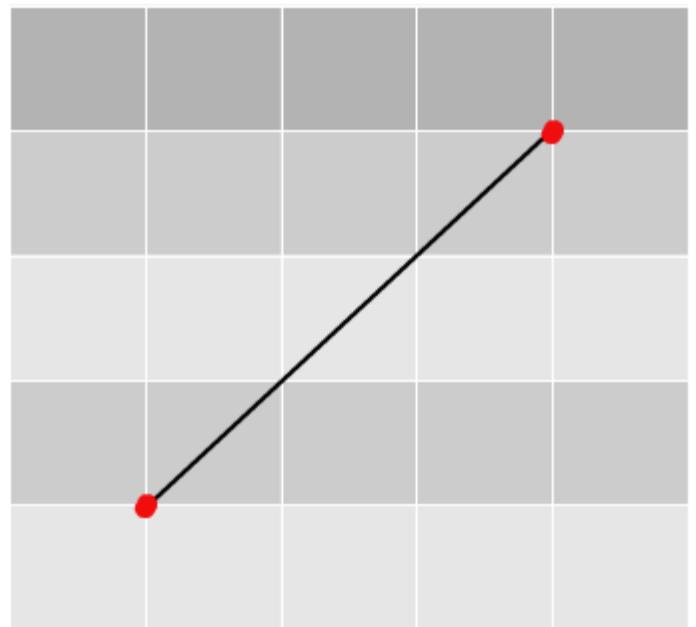
These measures are basic building blocks for activities, such as classification, clustering, and anomaly detection.

Similarity and dissimilarity measures are scores that describe to what extent objects are similar and dissimilar to each other. The table below summarizes the similarity and dissimilarity formulas for simple objects:

Attribute	Similarity (S)	Dissimilarity (D)
Nominal	$S = 1 \text{ if } X = Y$ $S = 0 \text{ if } X \neq Y$	$D = 0 \text{ if } X = Y$ $D = 1 \text{ if } X \neq Y$
Ordinal	$S = 1 - D$	$D = X-Y /(n-1)$ Where, n is the number of values
Interval or Ratio	$S = 1 / (1 + D)$ $S = 1 - (D - \min(D)) / \max(D) - \min(D)$	$D = X - Y $

It is a classical method to calculate the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space). This distance can be calculated by traveling along the line, connecting the points.

Euclidean Distance



You can use the Pythagorean Theorem to compute this distance:

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

You can create a recommendation system by building a similarity metric using:

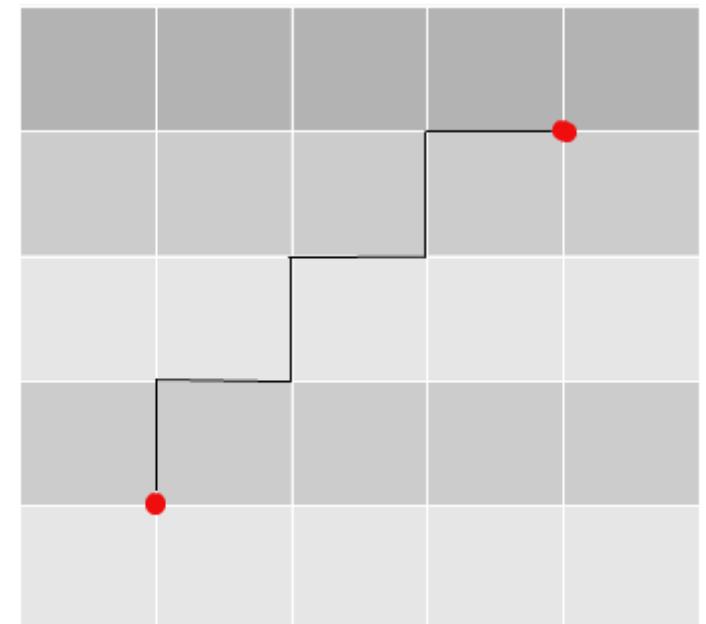
- Users
- Users' item preference data

This metric calculates the Euclidean distance “d” between two user points, similar to what's explained in the table below:

Users	Item 101	Item 102	Item 103	Distance	Similarity to User 1
User 1	5.0	3.0	2.5	0.000	1.000
User 2	2.0	2.5	5.0	3.937	0.203
User 3	2.5	-	-	2.500	0.286
User 4	5.0	-	3.0	0.500	0.667
User 5	4.0	3.0	2.0	1.118	0.472

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

Manhattan Distance



You can use the following formula to compute this distance:

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$

It is a metric on the Euclidean space and can be considered as a generalization of both the Euclidean and Manhattan distances.

You can use the following formula to compute this distance:

$$dist = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

When $r = 1$; it computes the Manhattan distance.

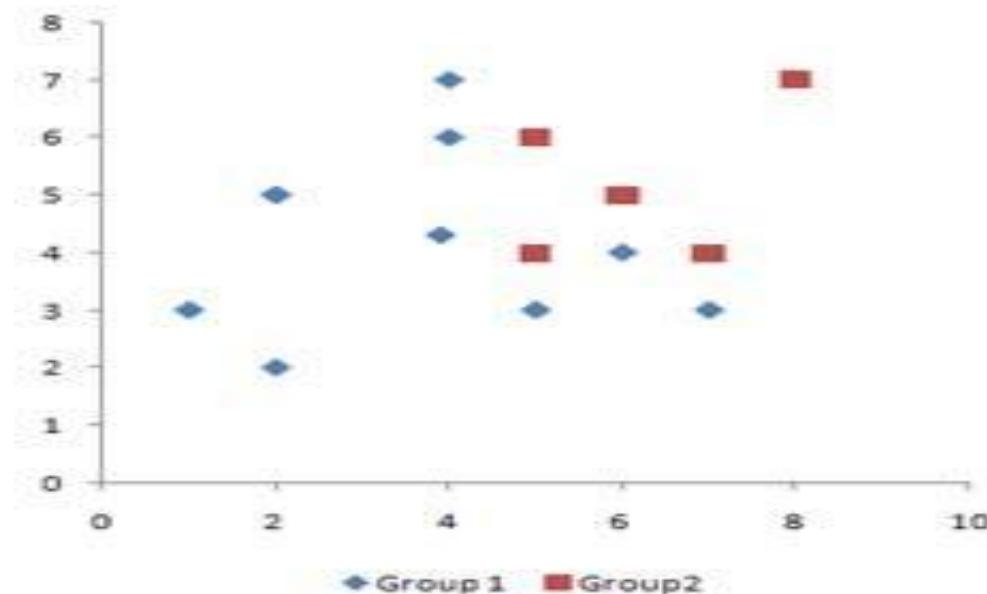
When $r = 2$; it computes the Euclidean distance.

When $r = \infty$; it computes Supremum.

It is used to calculate the distance between two groups of an object for:

- Graphical representation
- Classification and clustering

Mahalanobis Distance



You can use the following formula to compute this distance:

$$D_{ij} = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

It is a measurement:

- That computes the cosine of the angle between two vectors/documents on the Vector Space
- Of orientation and not magnitude

You can use the following formula to compute this similarity:

$$\text{similarity}(\vec{A}, \vec{B}) = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| * \|\vec{B}\|}$$

It is a measure that:

- Provides a number that depicts the relationship between objects
- Describes the bound between the objects
- Is generally represented by “r,” which can range from -1 to +1

If $r \sim 0$

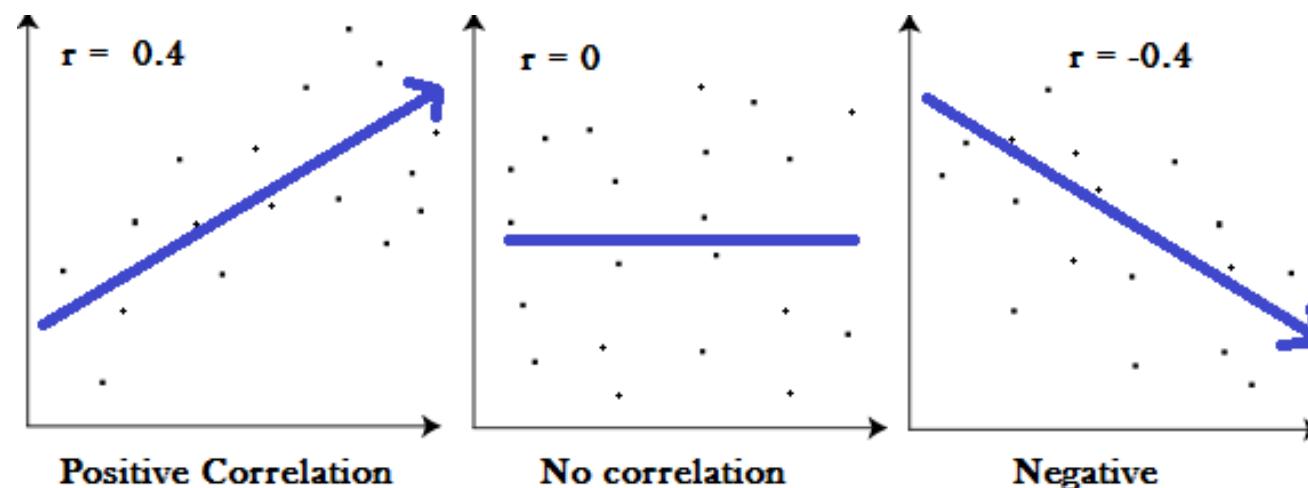
No relationship between the objects

If $r > 0$

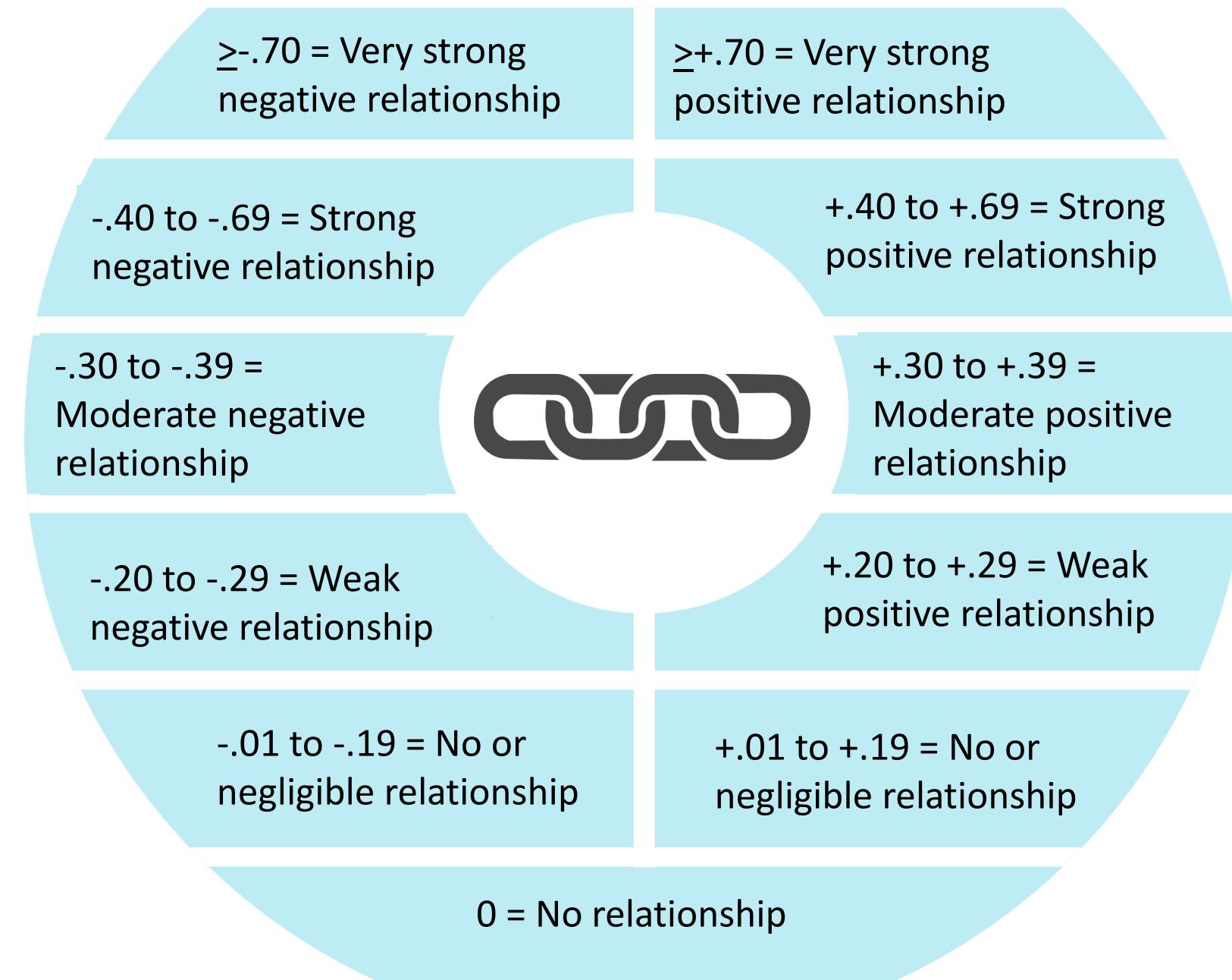
As one object becomes larger, the other also becomes larger

If $r < 0$

As one becomes larger, the other becomes smaller



The way “r” depicts the relationship between objects is explained below:



PPMC:

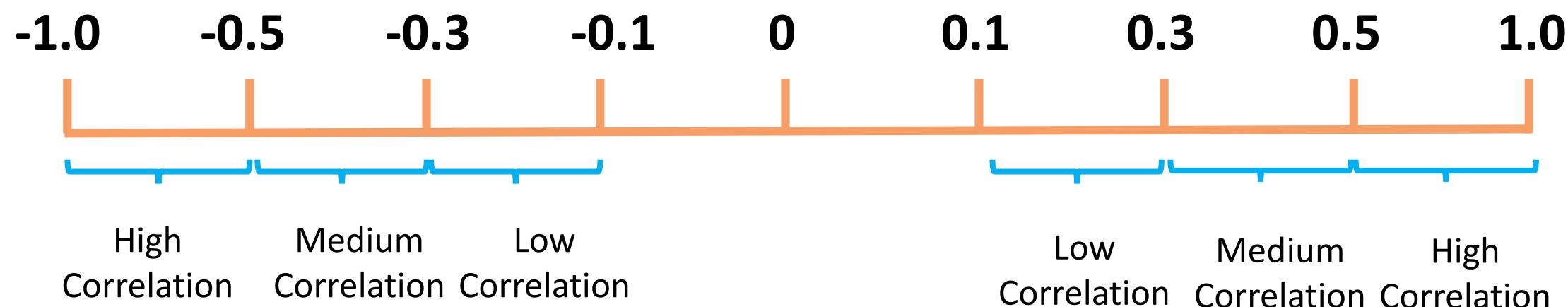
- Shows the linear relationship between two data sets
- Answers the question, “Can I draw a line graph to represent the data?”

You can use the following formula to compute this correlation:

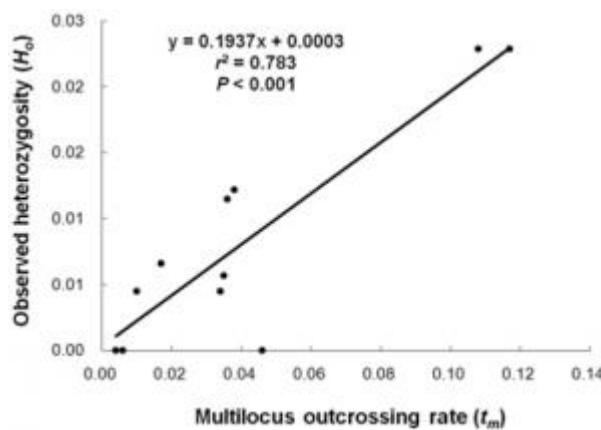
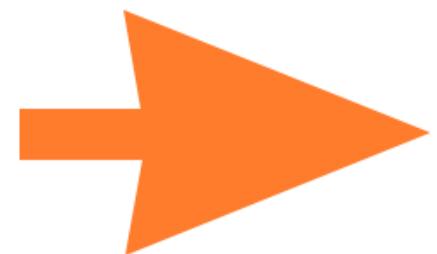
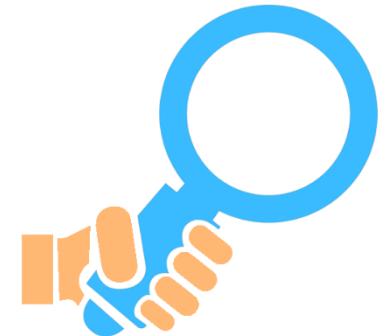
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The value “r”:

- Lies between -1 and 1
 - Gets closer to 0; meaning that the data points around the line of best fit variate more



Consider the case study below:



Research Topic:

Scientists in China wanted to analyze a genetic relationship between the different weedy rice populations.

Purpose:

To find out the evolutionary potential of the rice

Result:

Scientists analyzed the Pearson's correlation between the two groups and found a positive correlation of between 0.783 and 0.895 for weedy rice populations.

It calculates and returns the:

- Auto-distance/similarity matrix between:
 - rows or columns of a matrix/data frame or
 - a list.
- Cross-distance matrix between:
 - two matrices,
 - data frames, or
 - lists.

Let's see how to apply this function to calculate different distance measures:

Euclidean Distance

```
dist(test[1:2,])
```

Manhattan Distance

```
dist(test[1:2],m  
ethod="manhatt  
an")
```

Minkowski Distance, p = 2

```
dist(test[1:2],m  
ethod="minkows  
ki",p=2)
```



QUIZ

1

The following formula is used to calculate the _____.

- a. Manhattan distance
- b. Euclidean distance
- c. Minkowski distance

Mahalanobis distance



QUIZ

1

The following formula is used to calculate the _____.

- a. Manhattan distance
- b. Euclidean distance
- c. Minkowski distance

Mahalanobis distance



The correct answer is **b**.

Explanation: The given formula is used to calculate the Euclidean distance.

**QUIZ
2**

Which relation exists between objects when the correlation value is +.70 ?

- a. Weak and negative
- b. Weak and positive
- c. Strong and positive

Strong and negative



QUIZ
2

Which relation exists between objects when the correlation value is +.70 ?

- a. Weak and negative
- b. Weak and positive
- c. Strong and positive

Strong and negative



The correct answer is **c.**

Explanation: A strong and positive relation exists between objects when the correlation value is +.70.

**QUIZ
3**

The range of the Dissimilarity Measure lies from ____.

- a. 0 to infinity
- b. negative infinity to positive infinity
- c. -1 to 1

0 to 1



QUIZ
3

The range of the Dissimilarity Measure lies from ____.

- a. 0 to infinity
- b. negative Infinity to positive infinity
- c. -1 to 1

0 to 1



The correct answer is **a**.

Explanation: The range of the Dissimilarity Measure lies from 0 to INF [0, Infinity].

**QUIZ
4**

Which of the following types of measurements gives you the highest information content?

- a. Interval
- b. Ordinal
- c. Ratio

Nominal



QUIZ
4

Which of the following types of measurements gives you the highest information content?

- a. Interval
- b. Ordinal
- c. Ratio



Nominal

The correct answer is **c.**

Explanation: The ratio measurement gives the highest information content.

Let us summarize the topics covered in this lesson:



- Data is of two types: quantitative data and qualitative data.
- The types of measurements in their increasing order of content are nominal, ordinal, interval, and ratio.
- Statistical investigation lays the structure for decision making.
- Normal distribution is a family of bell-shaped and symmetric distributions.
- Calculating distance measures helps compare the objects from different perspectives, such as similarity, dissimilarity, and correlation.

Let us summarize the topics covered in this lesson:



- Euclidean Distance is the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space).
- Manhattan Distance is calculated by traversing vertical and horizontal lines in the grid-based system.
- Minkowski Distance can be considered as a generalization of both the Euclidean and Manhattan distance.
- Correlation is a measure that provides a number, which depicts the relationship between objects.
- PPMC shows the linear relationship between two data sets.

This concludes “Introduction to Statistics.”

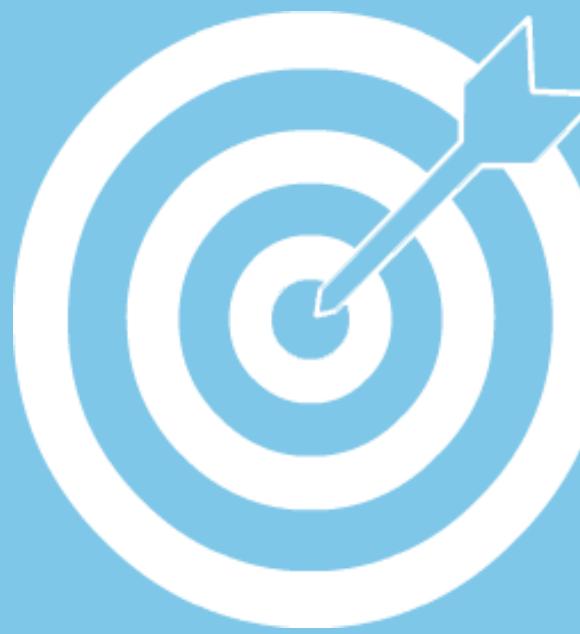
The next lesson is “Hypothesis Testing I.”

Data Science with R

Lesson 08—Hypothesis Testing I



After completing
this lesson, you will
be able to:



- Discuss the need of hypothesis testing in businesses
- Differentiate between null and alternate hypotheses
- Describe the two types of errors in sampling
- Interpret the confidence level, significance level, and power of a test
- Explain the types of hypothesis tests

It is an assertion or a statement about:

- The state of nature
- The true value of an unknown population parameter



Each hypothesis:

- Implies its contradiction or alternative
- Is either true or false
- Can be rejected on the basis of trial testimony and evidence and sample data



Examples:

- The accused is innocent.
- $\mu = 100$

Businesses collect and analyze data to help managers make optimal decisions that maximize profit at minimum risk. This depends on the acceptance or rejection of a hypothesis. For example, suitable hypothesis formulation and testing can help in the situation described below:



The product will be launched if the company gets a market share of 15% or more.

The product will not be launched if the company gets a market share of less than 15%.

Null Hypothesis:

- Is the first step in hypothesis testing
- Is usually a hypothesis of “no difference”
- Is denoted by H_0
- Is performed for a possible rejection under a true assumption
- Always refers to a specified value of the population parameter, such as μ



Example:

The population mean is 100.

Or

$$H_0: \mu = 100$$

This hypothesis:

- Commonly represents the status quo situation
- Is held to be true until a test results in its rejection
- Is accepted as “true” or rejected as “false” based on a consideration of a test statistic



This hypothesis is:

- Complementary to null hypothesis
- Denoted by H_1
- Used to decide whether to employ a single-tailed test or two-tailed test

Example:



For $H_0: \mu = 100$, the alternative hypothesis could be:

- $H_1: \mu \neq 100$
- $H_1: \mu > 100$ (right tailed)
- $H_1: \mu < 100$ (left tailed)

The differences between these hypotheses are given in the table below:

Null	Alternate
An assertion about one or more population parameters	An assertion of all situations that are not covered by the null hypothesis
Denoted by H_0	Denoted by H_1



Common Features of H_0 and H_1 :

- Mutually exclusive
- Exhaustive

The sampling theory draws valid inferences about the population parameters on the basis of sample results. In practice, a lot is accepted or rejected after examining its sample.



Example: A quality inspector accepts or rejects hardware components supplied by a vendor, generally, on the basis of test results of a random sample.

Until such statistical decisions are taken on the basis of evidence that does not provide complete confidence, there will be chances of errors.

The errors in statistical decisions are of two types:

Types of Errors

Type I Errors

- Reject H_0 , when it is true
- Probability is denoted by α

Type II Errors

- Accept H_0 , when it is wrong or H_1 is true
- Probability is denoted by β



In practice, a Type I error means rejecting a lot when it is good (producer's risk) and Type II error means accepting a lot when it is bad (consumer's risk).

It lists the possible outcomes of a statistical hypothesis test, as depicted below:

		State of Nature	
Decision	H_0 True	H_0 False	
Accept H_0	Correct	Type II Error (β)	
Reject H_0	Type I Error (α)	Correct	

A decision can be correct or incorrect.

Correct Decisions are:

- Fail to reject a true H_0
- Reject a false H_0



Incorrect Decisions are:

- Fail to reject a false H_0
- Reject a true H_0



The sampling distribution of a test statistic has two regions—a region of rejection (critical region) and a region of acceptance. The critical region amounts to rejection of H_0 , corresponding to the test statistic t in the sample space S .

Derivation:

If $w = \text{critical region}$ and $t = t(X_1, X_2, \dots, X_n)$ (Based on a random sample of size n)

Then,

- $P(t \in w | H_0) = \alpha$, $P(t \in w | H_1) = \beta$ (where, complementary set of w is the acceptance region)
- $W \cup w = S$, $w \cap w = \emptyset$



It is:

- The probability of a Type I error (α), that is, a random value of statistic t belongs to the critical region
- Usually set at 5% or 1% when employed in hypothesis testing

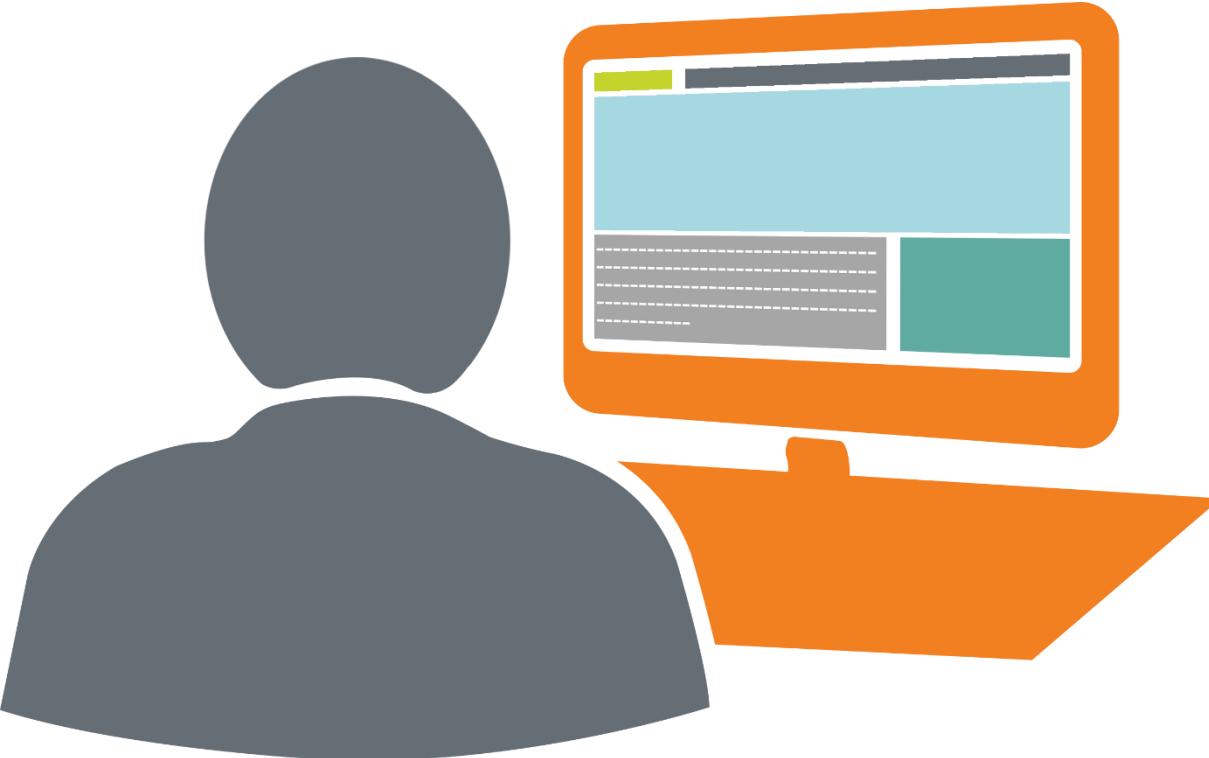
Important Points:

- If $\alpha = 0.05$ and you reject H_0 , then there is a 5% probability that you have rejected H_0 when it is true.
- The desired level of significance depends on the amount of risk you want to take in rejecting H_0 when it is true.



It:

- Is the complement of the probability of a Type I error ($1-\alpha$)
- Yields confidence level, when multiplied by 100%
- Represents the probability of concluding that a specific value of parameter being tested under H_0 is possible when, in fact, it is true



It is:

- The probability of committing a Type II error
- Depends on the difference between the hypothesized and actual values of the population parameter
- Inversely proportional to α



It is:

- The complement of the probability of a Type II error ($1-\beta$)
- The probability of rejecting H_0 , when it is false
- Required to be as powerful as possible for all critical regions of the same size



The power of test depends on the following factors:

Population Standard Deviation

Inversely proportional

Sample Size Used

Directly proportional

Level of Significance

Directly proportional

In a test, critical region is an area under the probability curve of a sampling distribution of a test statistic. There are two types of statistical hypothesis tests:

One tailed

- In this test, H_1 is one tailed (left tailed or right tailed).
- In a right-tailed test, critical region lies in the right tail of a sampling distribution, while for a left-tailed test, it lies in the left tail of distribution.

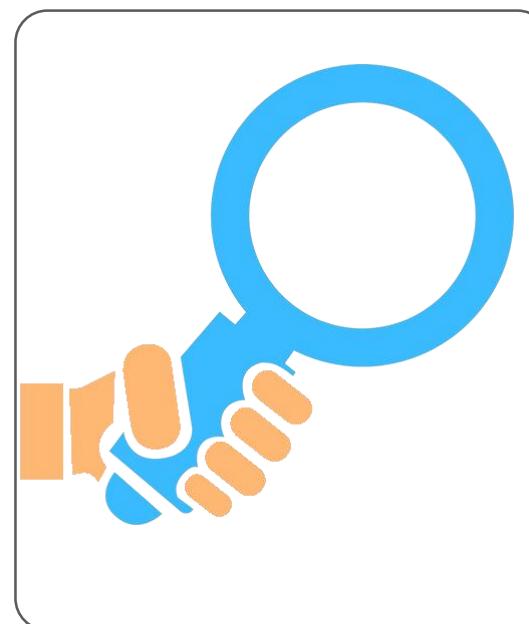
Two tailed

- In this test, H_1 is two tailed.
- Critical region is the area lying in both tails of the probability curve of the test statistic.

Assume that there are two population brands of bulbs:

- One has been manufactured using the standard process (mean life μ_1)
- The other has been manufactured using a new technique (mean life μ_2)

Now, there are three possibilities for testing bulbs:

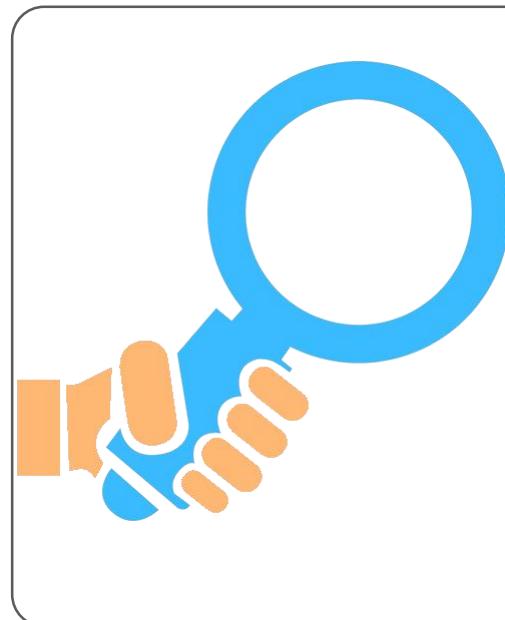


Test if the bulbs differ significantly

Test if the new process bulbs are better than the standard ones

Test if the new process bulbs are inferior to the standard ones

Let's analyze the results of these possibilities:



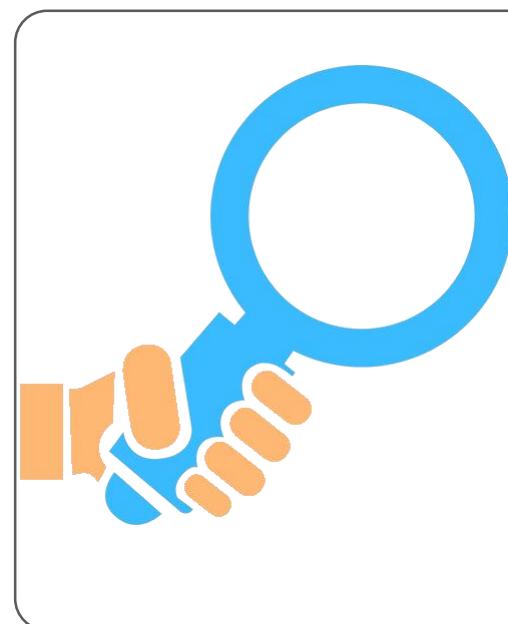
Test if the bulbs differ significantly

Test if the new process bulbs are better than the standard ones

Test if the new process bulbs are inferior to the standard ones

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Therefore, it is a two-tailed test.

Let's analyze the results of these possibilities:



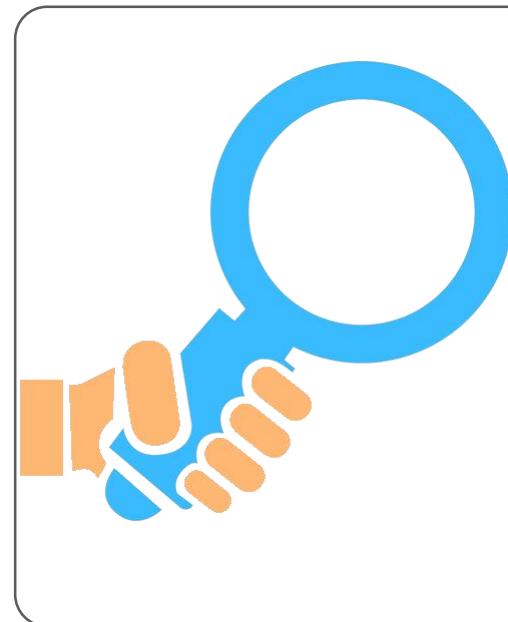
Test if the bulbs differ significantly

Test if the new process bulbs are better than the standard ones

Test if the new process bulbs are inferior to the standard ones

- $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$
- Therefore, it is a left-tailed test.

Let's analyze the results of these possibilities:



Test if the bulbs differ significantly

Test if the new process bulbs are better than the standard ones

Test if the new process bulbs are inferior to the standard ones

- $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$
- Therefore, it is a right-tailed test.

Let's understand this with the help of an example:



Problem



Diagnosis



Answer

For a chips brand:

- The label on a chips packet mentions that the maximum amount of saturated fat in a packet is 2 grams
- A test on a sample of 35 packets reveals that the mean saturated fat per packet is 2.1 grams

Assuming that the population standard deviation is 0.25 grams and significance level is .05, should the claim on the label be rejected?

Let's understand this with the help of an example:



Problem



Diagnosis



Answer

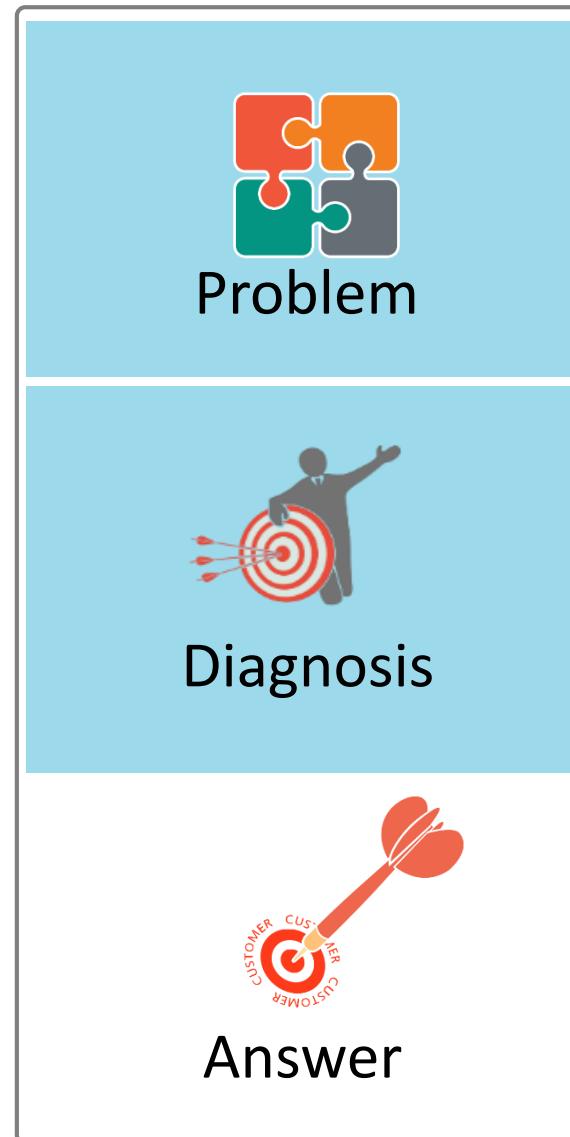
According to the null hypothesis, $\mu \leq 2$. Let's start by computing the test statistic.

```
> xbar = 2.1 # sample mean  
> mu0 = 2 # hypothesized value  
> sigma = 0.25 # population standard deviation  
> n = 35 # sample size  
> z = (xbar-mu0)/(sigma/sqrt(n))  
> z # test statistic  
[1] 2.3664
```

Let's then calculate the critical value at .05 significance level.

```
> alpha = .05  
> z.alpha = qnorm(1-alpha)  
> z.alpha # critical value  
[1] 1.6449
```

Let's understand this with the help of an example:



The diagnosis shows that the test statistic $2.3664 >$ the critical value 1.6449 .

Therefore, at .05 significance level, the claim of at most 2 grams of saturated fat in a chips packet should be rejected.

Test statistic:

- Is a function of sample observations that help in making the final decision about the acceptance or rejection of H_0
- Requires a knowledge of sampling distribution under H_0 to frame decision rules

Example:



The standardized variable for large samples, corresponding to statistic t , is:

$$Z = (t - E(t)) / (S.E(t)) \sim N(0,1) \dots \dots \dots \quad (1)$$

asymptotically as $n \gg \infty$

Here, the value of Z under H_0 is the test statistic.

The value of the test statistic, separating critical and acceptance regions, depends on the following two factors:

- Level of significance
- Alternative hypothesis H_1 (if it is two tailed or single tailed)

The critical value of the test statistic at level of significance α is determined by Z_α (where, $P(|Z| > Z_\alpha) = \alpha$). It depends on the type of tailed test as follows:

Single-Tailed Test

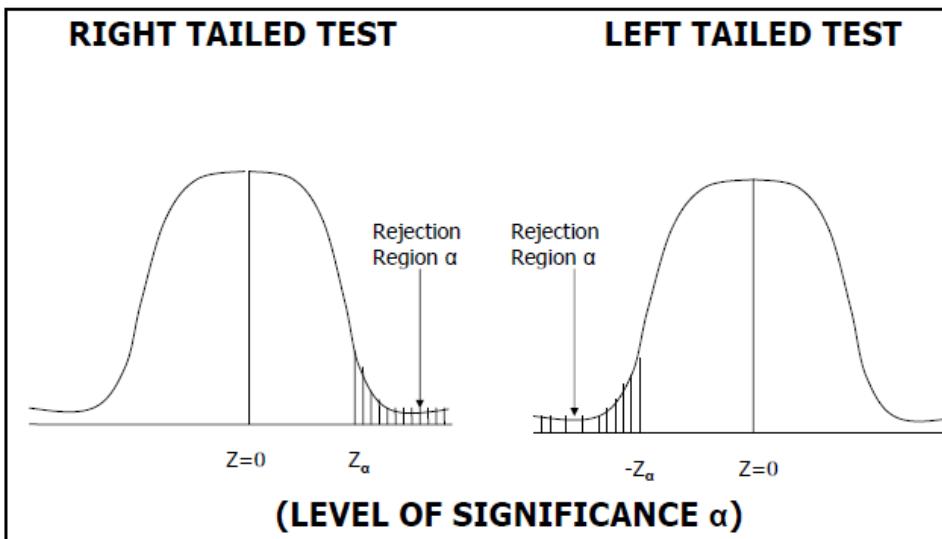
- Z_α is such that the total area for the right-tailed test, to the right of it is α and for the left-tailed test, to the left of $-Z_\alpha$ is α :
 - For right-tailed test: $P(Z > Z_\alpha) = \alpha$
 - For left-tailed test: $P(Z < -Z_\alpha) = \alpha$
- The critical value of Z at level of significance α is the same as the critical value of Z for a two-tailed test at level of significance 2α .

Two-Tailed Test

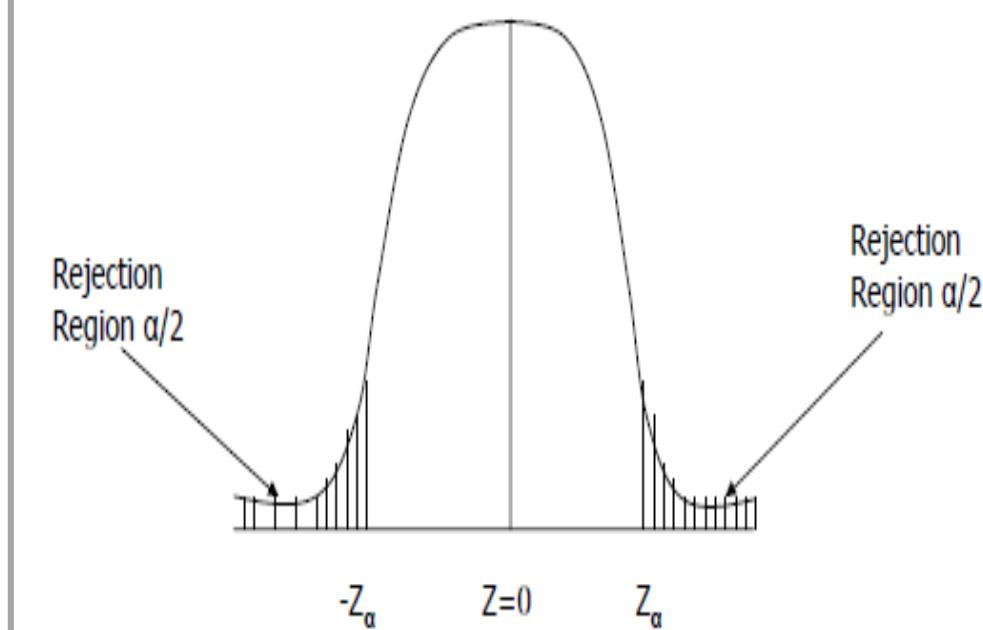
- Z_α is such that the total area of critical region on both tails is α :
 - $P(Z > Z_\alpha) + P(Z < -Z_\alpha) = \alpha$
 - $P(Z > Z_\alpha) + P(Z > Z_\alpha) = \alpha$ (normal probability curve is symmetrical)
- $2 P(Z > Z_\alpha) = \alpha$ or $P(Z > Z_\alpha) = \alpha/2$ (area of each tail is $\alpha/2$)

The graphs below show a comparison of rejection regions in two types of tailed tests:

Single-Tailed Test



Two-Tailed Test



The critical value of Z_α of Z using the normal probability table is shown below:

	State of Nature		
Critical Value (Z_α)	1%	5%	10%
Two-Tailed Test	$Z_\alpha = 2.58$	$Z_\alpha = 1.96$	$Z_\alpha = 1.645$
Right-Tailed Test	$Z_\alpha = 2.33$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left-Tailed Test	$Z_\alpha = -2.33$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$



Test statistic Z will not be normal for small n (<30) and you need to use significant values.



QUIZ
1

State whether the following statement is True or False.

In a statistical test, any observed difference between specified populations is due to sampling or experimental errors.

- a. True
- b. False



QUIZ
1

State whether the following statement is True or False.

In a statistical test, any observed difference between specified populations is due to sampling or experimental errors.

- a. True
- b. False



The correct answer is **a.**

Explanation: In a statistical test, any observed difference between specified populations is due to sampling or experimental errors.

QUIZ
2

Which of the following draws valid inferences about the population parameters on the basis of sample results?

- a. Power Test
- b. Error Sampling
- c. Sampling Theory

Confidence Interval



QUIZ
2

Which of the following draws valid inferences about the population parameters on the basis of sample results?

- a. Power Test
- b. Error Sampling
- c. Sampling Theory



Confidence Interval

The correct answer is **c.**

Explanation: Sampling theory draws valid inferences about the population parameters on the basis of sample results.

QUIZ
3*Fill in the blank:*

_____ refers to the percentage of all possible samples that can be expected to include the true population parameter.

- a. Significance level
- b. Confidence level
- c. Power test

Null hypothesis



QUIZ
3*Fill in the blank:*

_____ refers to the percentage of all possible samples that can be expected to include the true population parameter.

- a. Significance level
- b. Confidence level
- c. Power test

Null hypothesis



The correct answer is **b**.

Explanation: Confidence level refers to the percentage of all possible samples that can be expected to include the true population parameter.

QUIZ
4

Which of the following statements about β risk is correct ?

- a. It is inversely proportional to α .
- b. It is the probability of committing a Type I error.
- c. It is the probability of rejecting H_0 when it is false.

It provides the confidence level when multiplied by 100%.



QUIZ
4

Which of the following statements about β risk is correct ?

- a. It is inversely proportional to α .
- b. It is the probability of committing a Type I error.
- c. It is the probability of rejecting H_0 when it is false.



It provides the confidence level when multiplied by 100%.

The correct answer is **a**.

Explanation: β risk is inversely proportional to α .

Let us summarize the topics covered in this lesson:



- Each hypothesis implies its contradiction or alternative.
- Null hypothesis is the first step in hypothesis testing.
- Alternate hypothesis is complementary to null hypothesis.
- The errors in statistical decisions are of two types: Type I errors and Type II errors.
- The critical region amounts to rejection of H_0 , corresponding to statistic t in the sample space S .
- Level of significance is the probability of a Type I error (α), which means that a random value of statistic t belongs to the critical region.
- β risk is the probability of committing a Type II error.
- Power of test is the complement of the probability of a Type II error ($1-\beta$).
- There are two types of statistical hypothesis tests: One Tailed and Two Tailed.
- Test statistic is a function of sample observations that help in making the final decision to accept or reject H_0 .

This concludes “Hypothesis Testing I.”

The next lesson is “Hypothesis Testing II.”



Data Science with R

Lesson 09—Hypothesis Testing II



After completing
this lesson, you will
be able to:



- Explain the various parametric tests
- Discuss the types of null hypothesis tests
- Describe the Chi-Square Test
- Discuss the ANOVA test

In these tests, inferences are based on the assumptions made about the nature of the population distribution. Usually, the population considered for these tests is normal. These tests are of two types:

Z-Test or T-Test

In these tests, two population means or proportions are compared and tested.

Analysis of Variance (ANOVA) Test

In these tests, equality of several population means is tested.

It is performed in cases in which the test statistic is t, σ is known, and:

- The population is normal.
- The sample size is at least 30. (The population does not need to be normal.)

To calculate Z, use this formula:

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}} \right)}$$

Let's understand this with the help of a case study:



Problem

The test scores of an entrance exam fit a normal distribution with the mean test score of 72, and a standard deviation of 15.2. Let's compute the percentage of students scoring 84 or more.



Calculation in R

Let's use the `pnorm` function to find the required percentage of students and the upper tail of the normal distribution.

```
pnorm(84, mean = 72, sd = 15.2, lower.tail = FALSE)  
[1] 0.21492
```



Answer

The required percentage is 21.5%.

It is performed cases in which the test statistic is t and σ is unknown, but:

- The sample standard deviation is known.
- The population is normal.

To calculate t, use this formula:

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}} \right)}$$

Let's understand this with the help of a case study:



Problem

Let's find out the 2.5th and 97.5th percentiles of the Student's t-distribution, assuming 5 degrees of freedom.



Calculation in R

Let's use the quantile function "qt" against the decimal values 0.025 and 0.975.

```
qt(c(.025, .975), df = 5) # 5 degrees of freedom
```

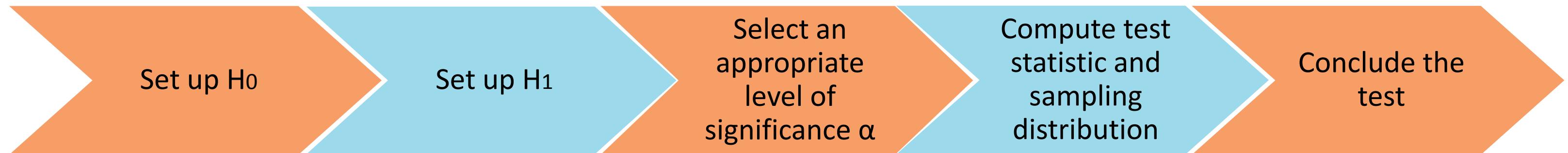
```
[1] -2.5706 2.5706
```



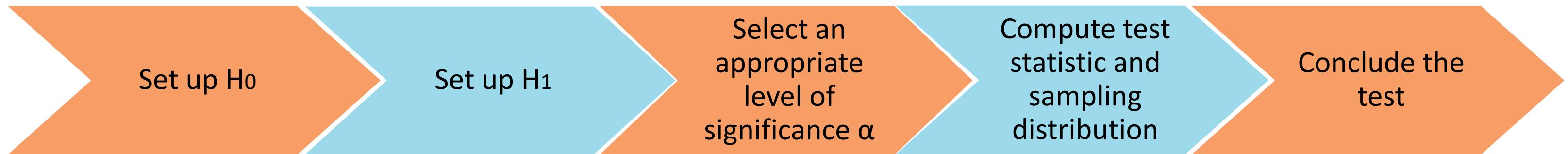
Answer

The required 2.5th and 97.5th percentiles are -2.5706 and 2.5706, respectively.

To test the null hypothesis, follow these steps:

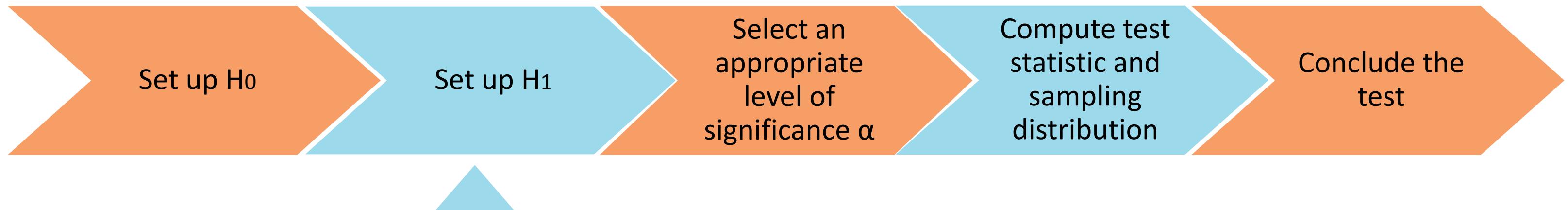


To test the null hypothesis, follow these steps:



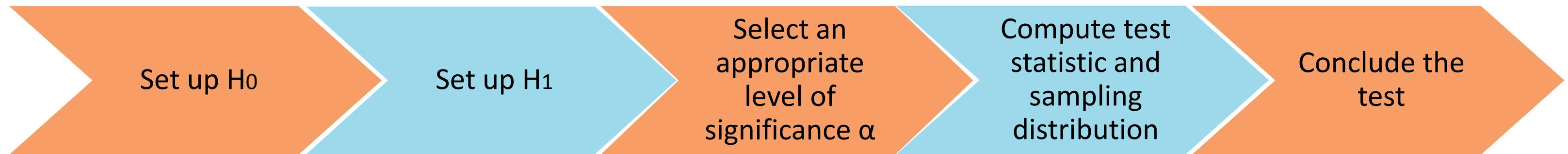
Set up H_0 by using population parameters.

To test the null hypothesis, follow these steps:



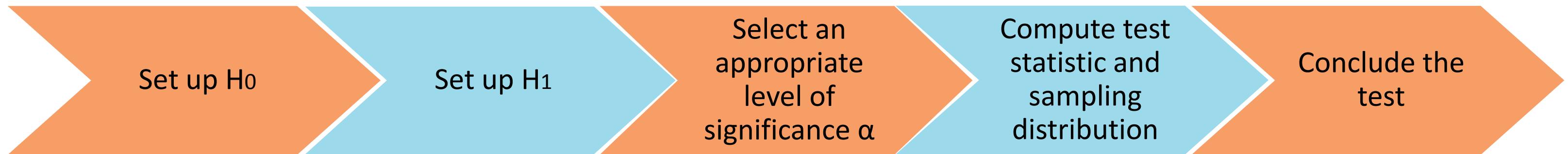
Set up H_1 to decide whether to use a single-tailed or a two-tailed test.

To test the null hypothesis, follow these steps:



Select α depending on the reliability of estimates and permissible risk. Note that α is fixed before a sample is drawn.

To test the null hypothesis, follow these steps:



Compute these when H_0 is true. For large sample tests: $Z = (t - E(t)) / S.E(t)$ and for small ones, assume the population to be normal and use various test statistics that follow t- or F-distribution.

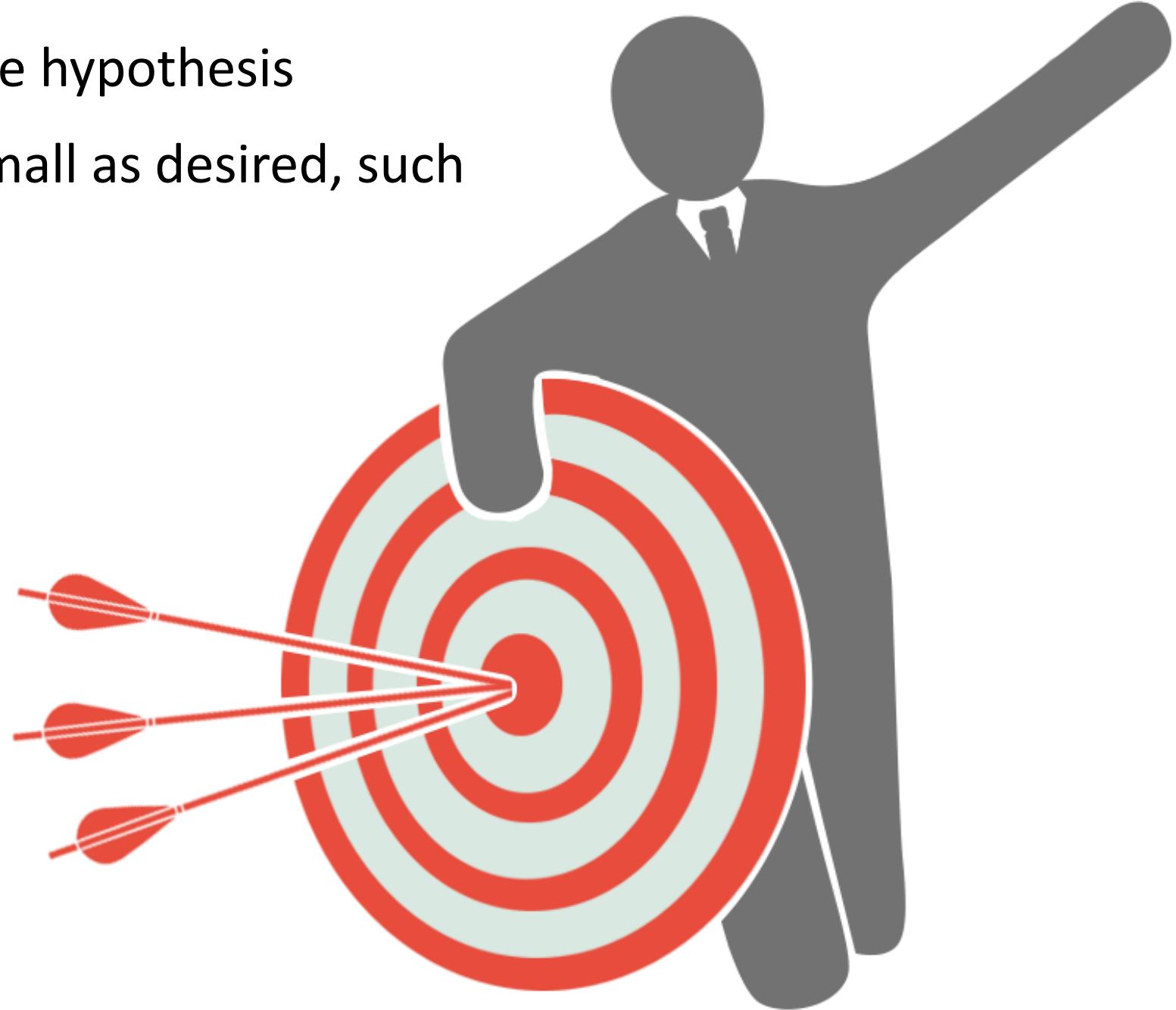
To test null hypothesis, follow these steps:



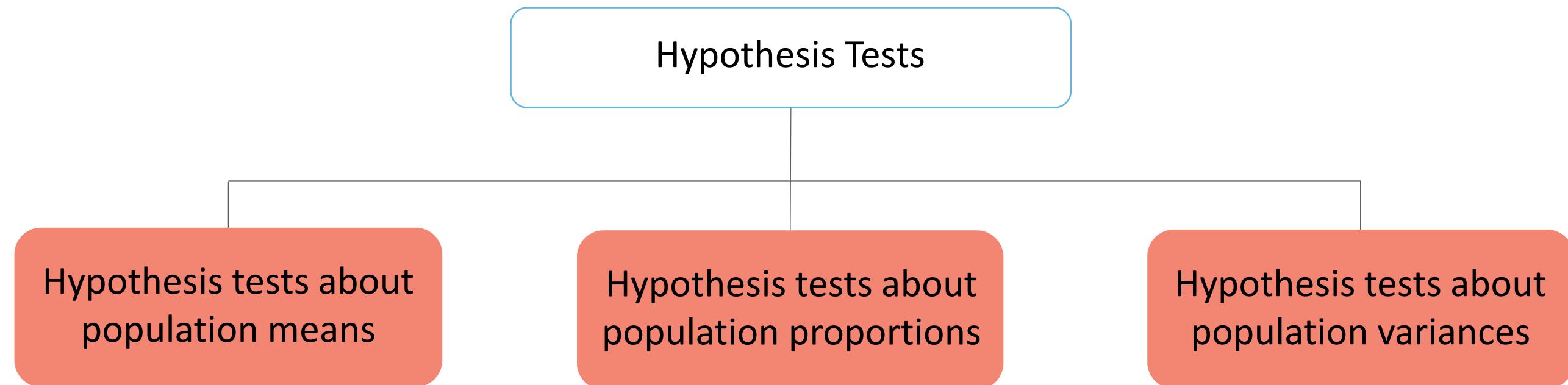
Compare Z with $Z\alpha$. If $|Z| < Z\alpha$, Z is not significant and the null hypothesis may, therefore, be accepted. If $|Z| \geq Z\alpha$, it is significant and the null hypothesis is rejected.

These include:

- Disproving a null hypothesis in favor of an alternative hypothesis
- Determining and regulating α , making its value as small as desired, such as 0.01 or 0.05
- Gaining a high level of confidence in decisions



These are:



The objective and assumptions of this test are described below:



Objective: To test the hypothesis that compares the population mean of interest with a specified value.



Assumptions: Assume X_1, X_2, \dots, X_n is a sample of size n from a normal population with mean μ and variance σ^2 . The mean X is distributed normally with the mean μ and variance σ^2/n ($X \sim N(\mu, \sigma^2/n)$). If n is large, X will be calculated similarly, even if the sample is from a non-normal population.

Therefore, for large samples, the standard normal variable corresponding to X bar is Z (as calculated in the Z-test).

Now, consider a random large sample of size n , giving a sample mean as the \bar{X} bar.



Test: You need to test the hypothesis that the sample mean X has been drawn from a population with the mean μ and a specified value μ_0 , that is:

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$
- $H_1 : \mu > \mu_0$
- $H_1 : \mu < \mu_0$

Under null hypothesis, $Z = (\bar{X} - \mu_0)/S.E.(X)$ follows Standard Normal Distribution approximately.

The critical region of the test has been explained in the table below:

H_1	Critical Region	
	5% Level	1% Level
$\mu \neq \mu_0$	$ Z \geq 1.96$	$ Z \geq 2.58$
$\mu > \mu_0$	$Z \geq 1.645$	$Z \geq 2.33$
$\mu < \mu_0$	$Z \leq -1.645$	$Z \leq -2.33$

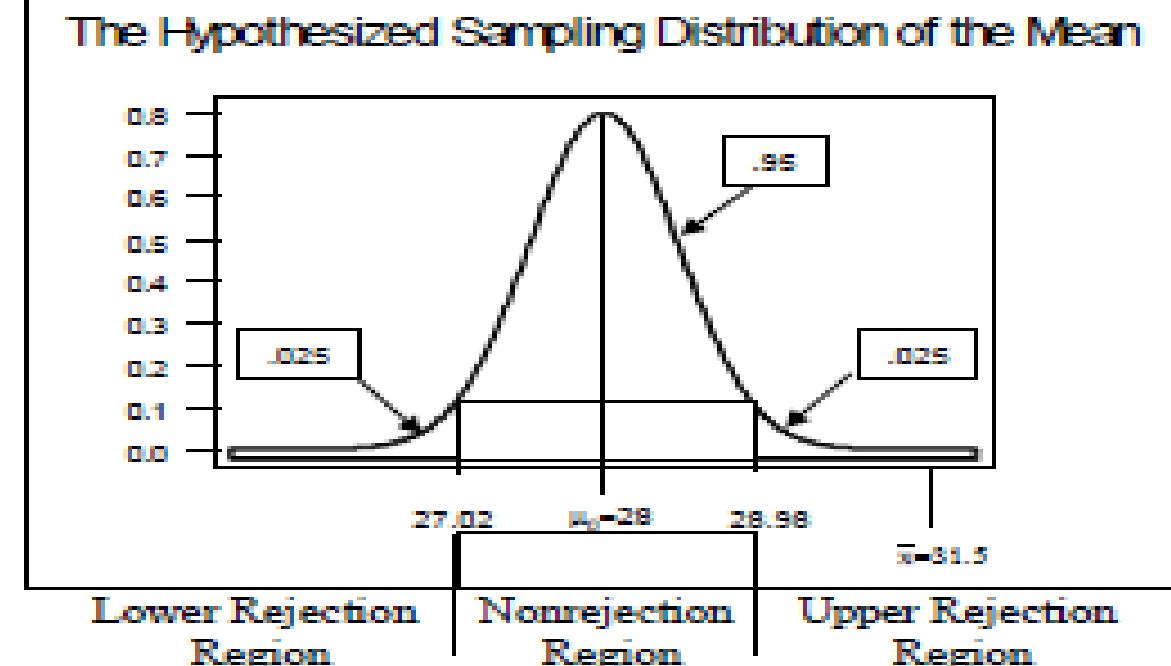


Important Points:

- 95% confidence limit for population mean μ : $\bar{X} \pm 1.96$ (S.E. (\bar{X}))
- 99% confidence limit for population mean μ : $\bar{X} \pm 2.58$ (S.E. (\bar{X})), where S.E. (\bar{X}) = σ/\sqrt{n} .
- If population variance σ^2 is unknown, its estimate is used as provided by sample variance S^2 .
- S.E. (X) = $S\sqrt{n}$, where population S.D. σ is unknown.

The rules are:

- Create a $(1-\alpha)$ non-rejection region around the hypothesized population mean.
 - Do not reject H_0 if the sample mean falls within the non-rejection region.
- Decide between the null and alternative hypotheses.
 - If p value $\leq \alpha$, reject H_0 .



An automatic machine fills an aerated drink in 2000 cc bottles. A tester:

- Needs to test H_0 that the average amount being filled in a bottle is at least 2000 cc
- Selects a random sample of 40 bottles and records the exact content of the bottles
- Finds the sample mean to be 1999.6 cc
- Considers the population standard deviation as 1.30 cc

Let's test the null hypothesis at the significance level of 5%.

$H_0: \mu \geq 2000$
 $H_1: \mu < 2000$
 $n = 40$
For $\alpha = 0.05$, the critical value of z is -1.645
The test statistic is: $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
Do not reject H_0 if: $[z > -1.645]$
Reject H_0 if: $[z \leq -1.645]$

$n = 40$
 $\bar{x} = 1999.6$
 $\sigma = 1.3$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{1999.6 - 2000}{1.3 / \sqrt{40}} = -1.95 \Rightarrow \text{Reject } H_0$$

For an insurance company:

- The average liability insurance for each board seat in small companies has been \$2000.
- α is 0.01.

Let's test this hypothesis using the Growth Resources, Inc. survey data.

$$H_0: \mu = 2000$$

$$H_1: \mu \neq 2000$$

For $\alpha = 0.01$, critical values of z are ± 2.576

$$\text{The test statistic is: } z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Do not reject H_0 if: $[-2.576 < z < 2.576]$

Reject H_0 if: $[z \leq -2.576]$ or $[z \geq 2.576]$

$$n = 100$$

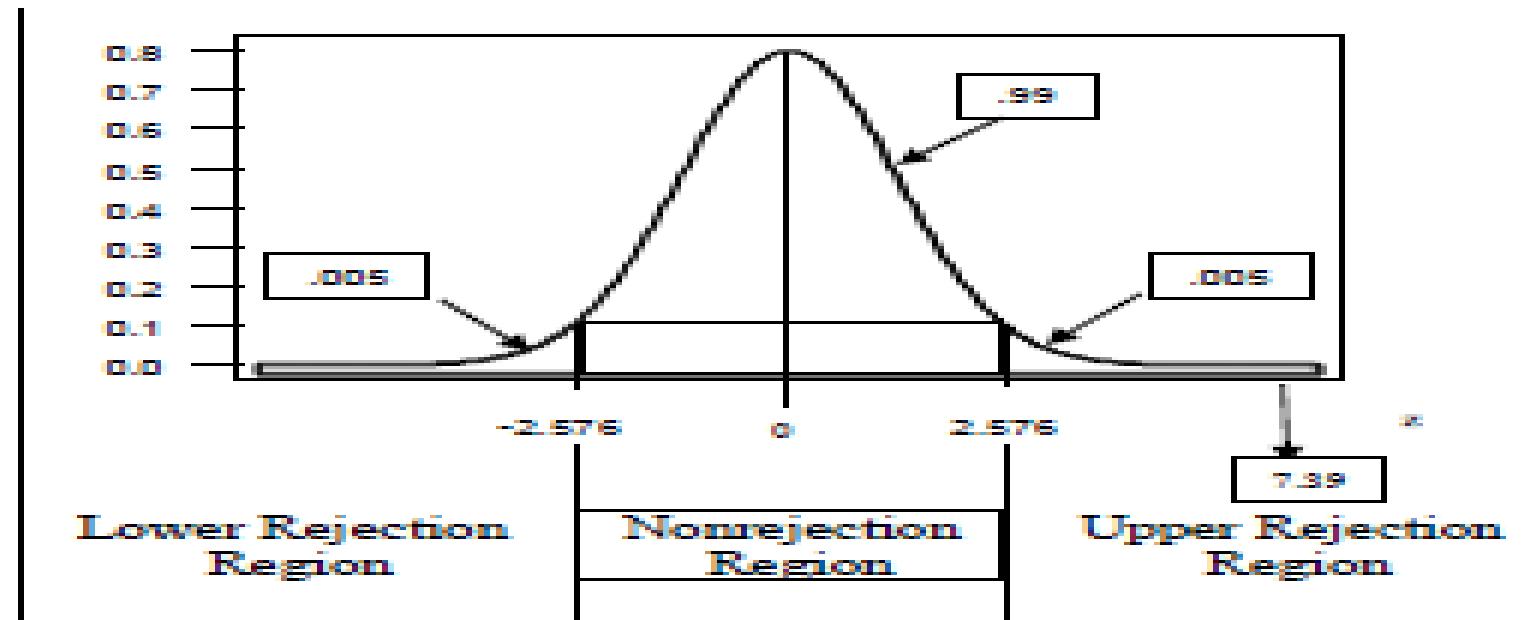
$$\bar{x} = 2700$$

$$s = 947$$

$$z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{2700 - 2000}{\frac{947}{\sqrt{100}}}$$

$$= \frac{700}{94.7} = 7.39 \Rightarrow \text{Reject } H_0$$

The Standard Normal Distribution for the insurance company is shown below:



Conclusion: The test statistic falls in the upper rejection region; therefore, H_0 is rejected and the average insurance liability is more than \$2000.

The objective of this test is mentioned below:



Objective: To test the hypothesis that compares the population proportion of interest with a specified value

Now, consider a random sample of the size n and the proportion of members with a certain attribute p .



Test: You need to test the hypothesis that the proportion P in the population has a specified value P_0 , that is:

- $H_0 : P = P_0$
- $H_1 : P \neq P_0$
- $H_1 : P > P_0$
- $H_1 : P < P_0$

For a large sample, $Z = (p - P_0)/S.E.(p) \sim N(0,1)$ (under H_0)

Where,

$p = X/n$ = Number of successes in sample / Sample size

P_0 = Hypothesized proportion of successes in the population

The critical region of the test has been explained in the table below:

H_1	Critical Region	
	5% Level	1% Level
$P \neq P_0$	$ Z \geq 1.96$	$ Z \geq 2.58$
$P > P_0$	$Z \geq 1.645$	$Z \geq 2.33$
$P < P_0$	$Z \leq -1.645$	$Z \leq -2.33$



Important Points:

- 95% confidence limit for population mean P : $p \pm 1.96$ (S.E. (p))
- 99% confidence limit for population mean P : $p \pm 2.58$ (S.E. (p))

$$\text{Where S.E.} (p) = \sqrt{\left(\frac{P(1-P)}{n}\right)}.$$

Let's consider a case study.

A marketing manager needs to decide whether to launch a new product.



The product will be launched if the product acceptance rate (measured in a blind comparison) exceeds 30%.



The product will not be launched if the product acceptance rate does not exceed 30%.

The manager considers a random sample of 200 consumers, which shows the acceptance rate as 32%. Assuming the level of significance α of 0.05, let's perform hypothesis testing to conclude an action.



Test:

$$H_0: P \leq 0.30$$

$$H_1: P > 0.30$$

$$\alpha = 0.05$$

$$Z = (p - P) / \sqrt{(P(1-P)/n)}$$

$$= (0.32 - 0.30) / \sqrt{(0.30(1-0.30)/200)}$$

$$= 0.62$$

Critical value of Z for a right-tailed test at $\alpha = 0.05$ is $Z\alpha = 1.645$

Since $|Z| = 0.62 < Z\alpha = 1.645$

Accept H_0 at 5% level of significance.



Recommended Action:

Manager should not introduce the new product in the market.

This test:

- Considers the square of a standard normal variate
- Is appropriate for nominal level measurements
- Evaluates if frequencies observed in different categories vary significantly from the frequencies expected under a specified set of assumptions
- Determines how well an assumed distribution fits the data
- Uses contingency tables (in market researches, these tables are called cross-tabs)



Example:

If $x \sim N(m, \sigma^2)$, then $Z = (x-m)/\sigma \sim N(0,1)$.

Here, $Z^2 = ((x-m)/\sigma)^2$ is a chi-square variate with 1 degree of freedom.

To perform the Chi-Square Test, follow these steps:

Hypothesize about a population by stating H_0 and H_1 .

→ Compute the occurrence frequencies of certain events expected under H_0 .

→ Note the observed counts of data points falling in different cells.

→ Consider the difference between observed and expected frequencies.

→ Compare the statistic values with critical points of Chi-Square distribution.



Note:

- The difference between observed and expected frequencies provides a computed value of the Chi-Square statistic:

$$\chi^2 \text{ statistic: } \chi^2 = \sum_1^k (O_i - E_i)^2 / E_i,$$

Where,

E_i = Expected frequency in category i

O_i = Observed frequencies in category I, if H_0 is true

- The test statistic depends on the square of differences, so sample values are always positive.



These are:

- $\sum(O_i - E_i) = 0$
- χ^2 test is a non-parametric test.
- χ^2 distribution can be used as long as the expected count in every cell is at least 5.0.

It is:

- The number of independent variates that make up the statistic (for example, χ^2)
- Usually denoted by v or df
- Computed as the total number of observations less the number of independent constraints

Note:

To use data for estimating parameters of probability distribution stated in H_0 , for every parameter, an additional degree of freedom will be lost.

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

This follows χ^2 distribution with $(k-1)$ degrees of freedom.

Where, $\sum O_i = \sum E_i = N$, total frequency



This test is suitable for situations involving:

- One categorical variable
- Two categorical variables, where you will need to determine the relation between them
- Cross-tabulations, where you need to find out the relation between two categorical variables
- Non-quantifiable variables (For example, answers to questions like, do employees in different age groups choose different types of health plans?)



In case of non-quantifiable information, instead of correlation, the word “association” or “dependence,” and instead of variables, the word “factors” is used.

Some situations that require you to identify the relation between two attributes are:



Credit worthiness of borrowers for their age groups and personal loans

Relation between the performance of salesmen and training received

Return on a single stock and on stocks of a sector like pharmaceutical and banking

Level of job satisfaction and salary grades of employees

Category of viewers and impact of a TV campaign



In such cases, H_0 and H_1 are tested, which show the dependence and independence of the two factors, respectively.

The manager of a restaurant needs to know the relation between customer satisfaction and the salaries of the people waiting tables. She:

- Takes a random sample of 100 customers asking if the service was excellent, good, or poor.
- Categorizes the salaries of the people waiting as low, medium, and high.

Her findings are shown in the table below:

	Salary			
Service	Low	Medium	High	Total
Excellent	9	10	7	26
Good	11	9	31	51
Poor	12	8	3	23
Total	32	27	41	100

Assume the level of significance as 0.05. Here, H_0 and H_1 will denote the independence and dependence of the service quality on the salaries of people waiting tables.



Test: Therefore:

- $DF = (3-1)(3-1) = 4$
- Under H_0 , expected frequencies are:
 - $E_{11} = (26 \times 32)/100 = 8.32, E_{12} = 7.02, E_{13} = 10.66$
 - $E_{21} = 16.32, E_{22} = 13.77, E_{23} = 20.91$
 - $E_{31} = 7.36, E_{32} = 6.21, E_{33} = 9.41$

Therefore, $\chi^2(\text{calculated}) = (9-8.32)^2/8.32 + (10-7.02)^2/7.02 + (7-10.66)^2/10.66 + (11-16.32)^2/16.32 + (9-13.77)^2/13.77 + (31-20.91)^2/20.91 + (12-7.36)^2/7.36 + (8-6.21)^2/6.21 + (3-9.43)^2/9.43 = 18.658$

- $\chi^2_{0.05,4} = 9.48773$
 $\chi^2(\text{Calculated}) > \chi^2(\text{Tabulated})$
- Reject H_0 , accept H_1 .



Conclusion:

Service quality is dependent on the salaries of the people waiting.

To perform this test in R, let's consider a table, which is a result of a survey conducted among students about their smoking habits. This table has:

- “Smoke” variables, which record the smoking habits of students (Allowed values: "Heavy," "Regul," "Occas," and "Never")
- “Exer” variables, which record the exercise levels of smoking (Allowed values: "Freq," "Some," and "None")

Let's build the contingency table in R:

```
library(MASS) # load the MASS package  
head(survey)  
tbl = table(survey$Smoke, survey$Exer)  
tbl  
      Freq  None Some  
    Heavy 7 1 3  
    Never 87 18 84  
    Occas 12 3 4  
    Regul 9 1 7
```



Problem



Calculation in R



Answer

Assuming .05 as the significance level, let's test the hypothesis whether the smoking habits of students are independent of their exercise levels or not.

Let's use the `chisq.test` function for the contingency table and find the value of p.
`chisq.test(tbl)`

Output: data: table(survey\$Smoke, survey\$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828

As $p >$ significance level, H_0 is not rejected. This means that the smoking habits of students are independent of their exercise levels.

Various statistical applications in social science, psychology, natural sciences, and business administration include several groups, such as:

- An environmentalist might need to know if the average amount of pollution varies in several bodies of water.
- A sociologist might want to find out if a person's income varies according to his/her upbringing.

The ANOVA test is used for such hypothesis tests that compare the averages of two or more groups.



This test:

- Determines if a statistically significant difference exists among several group means or not by using variances
- Tests $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ (where, μ = group mean and k = number of groups)

Five Assumptions:

- All samples are random and independent.
- Each population is normal.
- The factor is a categorical variable.
- The populations have equal standard deviations.
- The result is a numerical variable.

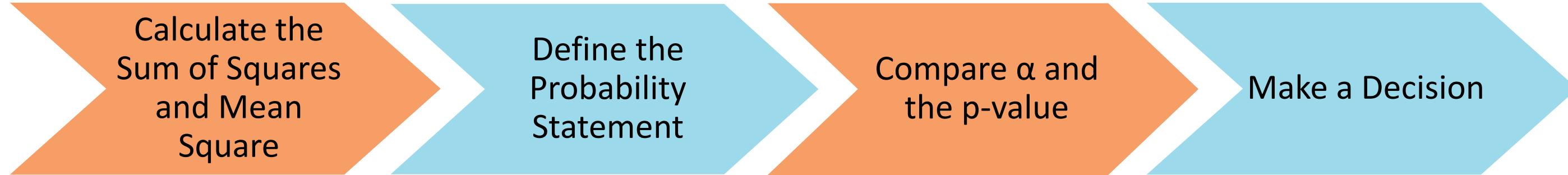


However, the test provides a significant result, and H_1 is accepted.

These are explained below:

- **F-Distribution:** The distribution for the test with two different degrees of freedom
- **F-Ratio:** Value derived from two estimates of the variance, as described below:
 - **Variance between samples (SSbetween):** An estimate of σ^2 : variance of the sample means * n, when the sample sizes are the same. When sizes are different, the variance is weighted to account for different sample sizes.
 - **Variance within samples (SSwithin):** An estimate of σ^2 : average of sample variances. When sizes are different, the variance within samples is weighted.

To perform this test, perform these steps:





Problem

Let's find the 95th percentile of the F-distribution with (5, 2) degrees of freedom.



Calculation in R

Let's use the quantile function "qf" of the F-distribution against the decimal value 0.95.

```
qf(.95, df1 = 5, df2 = 2)
```

```
[1] 19.296
```



Answer

The required percentile 19.296.

Four sororities considered a random sample of grade means of a few sisters based on their past terms. Assuming a significance level of 1%, the results are given in the table below:

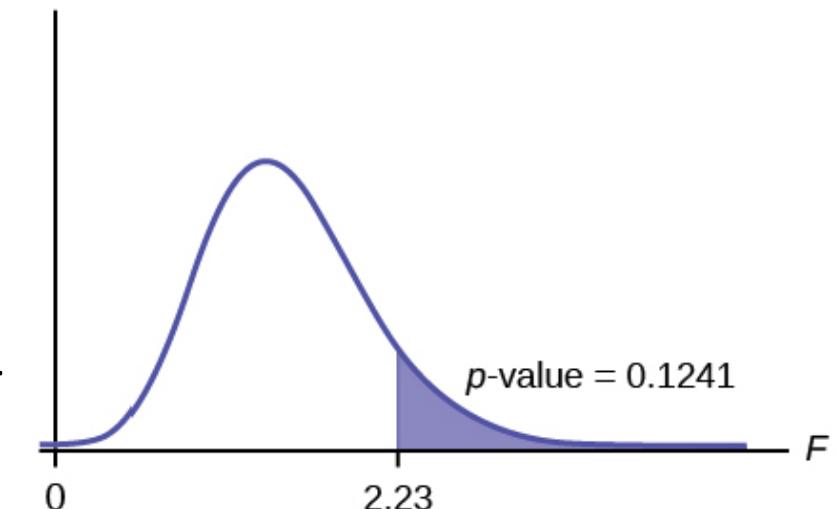
Sorority 1	Sorority 2	Sorority 3	Sorority 4
2.17	2.63	2.63	3.79
1.85	1.77	3.78	3.45
2.83	3.25	4.00	3.08
1.69	1.86	2.55	2.26
3.33	2.21	2.45	3.18

Let's find out if there is a difference in the mean grades among the sororities, assuming μ_1, μ_2, μ_3 , and μ_4 are the population means of the sororities.



Test: Therefore:

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
- $H_1: \text{Not all of the means } \mu_1, \mu_2, \mu_3, \text{ and } \mu_4 \text{ are equal}$
- **Distribution for the test:** $F_{3,16}$
 - $df(\text{num}) = k - 1 = 4 - 1 = 3$
 - $df(\text{denom}) = n - k = 20 - 4 = 16$
- **Calculate the test statistic:** $F = 2.23$
- **Define probability statement:** $p\text{-value} = P(F > 2.23) = 0.1241$
- **Compare α and the p-value:** $\alpha = 0.01$
 - $p\text{-value} = 0.1241$
 - $\alpha < p\text{-value}$
- **Decide:** Since $\alpha < p\text{-value}$, you cannot reject H_0 .



Conclusion:

Without sufficient evidence, you cannot conclude that there is a difference among the mean grades for the sororities.



Example

A fast food chain is testing and marketing three of its new menu items. To analyze if they are equally popular, consider:

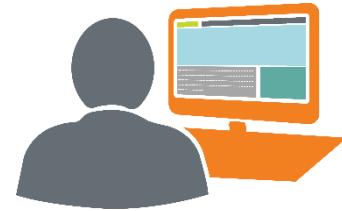
- 18 random restaurants for the study
- 6 of the restaurants to test market the first menu item, another 6 for the second one, and the remaining 6 for the last one



Problem

After a week of testing, assume that the table below shows the sales figures of the menu items in the 18 restaurants. At .05 level of significance, you need to test whether the mean sales volumes for these menu items are equal.

Item 1	Item 2	Item 3
22	52	16
42	33	24
44	8	19
52	47	18
45	43	34
37	32	39



Calculation in R

Follow these steps:

1. Copy and paste the sales figures in a table file "fastfood-1.txt" using a text editor.
2. Load the file into a data frame df1 using the read.table function.

```
df1 = read.table("fastfood-1.txt", header = TRUE); df1
```

Item1 Item2 Item3

	Item1	Item2	Item3
1	22	52	16
2	42	33	24
3	44	8	19
4	52	47	18
5	45	43	34
6	37	32	39

3. Concatenate the data rows of df1 into a single vector r.

```
r = c(t(as.matrix(df1))) # response data
```

```
r
```

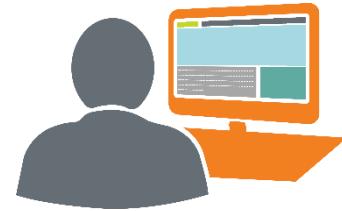
```
[1] 22 52 16 42 33 ...
```

4. Assign new variables for the treatment levels and number of observations.

```
f = c("Item1", "Item2", "Item3") # treatment levels
```

```
k = 3      # number of treatment levels
```

```
n = 6      # observations per treatment
```



Calculation in R

Further steps are as follows:

4. Create a vector of treatment factors, corresponding to each element of r in step 3, using the gl function.

```
tm = gl(k, 1, n*k, factor(f)) # matching treatments
```

```
Tm
```

```
[1] Item1 Item2 Item3 Item1 Item2 ...
```

```
tm = gl(k, 1, n*k, factor(f)) # matching treatments
```

```
tm
```

```
[1] Item1 Item2 Item3 Item1 Item2 ...
```

Apply the function aov to a formula that describes the response r by the treatment factor tm.

```
av = aov(r ~ tm)
```

Print out the ANOVA table with the summary function.

```
summary(av)
```

Df	Sum Sq	Mean Sq	F value	Pr(>F)
----	--------	---------	---------	--------

tm	2	745	373	2.54	0.11
----	---	-----	-----	------	------

Residuals	15	2200	147		
-----------	----	------	-----	--	--



Solution

p-value of 0.11 > .05 significance level. Do not reject H_0 , which means that the mean sales volumes of the new menu items are all equal.



QUIZ
1

The formula to calculate the F-ratio is ____.

- a. $M_s_{\text{between}} - M_s_{\text{within}}$
- b. $M_s_{\text{between}} / M_s_{\text{within}}$
- c. $M_s_{\text{within}} / M_s_{\text{between}}$

$M_s_{\text{within}} * M_s_{\text{between}}$



QUIZ
1

The formula to calculate the F-ratio is ____.

- a. $M_s_{\text{between}} - M_s_{\text{within}}$
- b. $M_s_{\text{between}} / M_s_{\text{within}}$
- c. $M_s_{\text{within}} / M_s_{\text{between}}$

$M_s_{\text{within}} * M_s_{\text{between}}$



The correct answer is **b**.

Explanation: The formula to calculate the F-ratio is $M_s_{\text{between}} / M_s_{\text{within}}$.

QUIZ
2

Which of the following statements about the one-way ANOVA Test is true?

- a. H_0 is accepted if $\alpha < p\text{-value}$.
- b. H_1 is accepted if $\alpha < p\text{-value}$.
- c. H_0 is rejected if $\alpha = p\text{-value}$.

H_1 is rejected if $\alpha = p\text{-value}$.



QUIZ
2

Which of the following statements about the one-way ANOVA Test is true?

- a. H_0 is accepted if $\alpha < p\text{-value}$.
- b. H_1 is accepted if $\alpha < p\text{-value}$.
- c. H_0 is rejected if $\alpha = p\text{-value}$.



H_1 is rejected if $\alpha = p\text{-value}$.

The correct answer is **a**.

Explanation: H_0 is accepted if $\alpha < p\text{-value}$.

QUIZ
3

Which of the following statements about the Chi-Square Test is correct ?

- a. The answers in this test cannot be measured on a numerical scale.
- b. The related variables are not quantifiable.
- c. It evaluates whether observed frequencies differ significantly from frequencies expected under a specified set of assumptions.



All of the above statements are correct.

QUIZ
3

Which of the following statements about the Chi-Square Test is correct ?

- a. The answers in this test cannot be measured on a numerical scale.
- b. The related variables are not quantifiable.
- c. It evaluates whether observed frequencies differ significantly from frequencies expected under a specified set of assumptions.



All of the above statements are correct.

The correct answer is **d**.

Explanation: All the given statements about the Chi-Square Test are correct .

QUIZ
4

Which of the following tests is used when σ is unknown, the sample standard deviation is known, and the population is normal?

- a. Z-Test
- b. T-Test
- c. Chi-Square Test



F-Ratio Test

QUIZ
4

Which of the following tests is used when σ is unknown, the sample standard deviation is known, and the population is normal?

- a. Z-Test
- b. T-Test
- c. Chi-Square Test

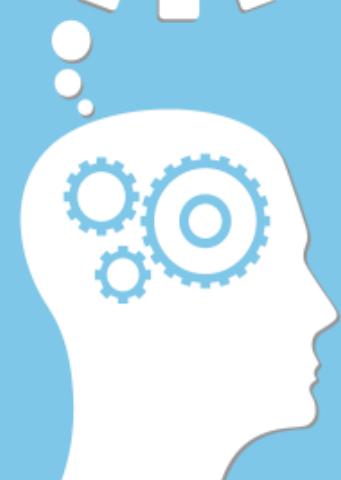


F-Ratio Test

The correct answer is **b**.

Explanation: The T-test is used when σ is unknown, the sample standard deviation is known, and the population is normal.

Let us summarize the topics covered in this lesson:



- There are two types of parametric tests: Z-Test or T-Test and ANOVA Test.
- The Z-test is performed in cases in which the test statistic is t and σ is known.
- The T-test is performed in cases in which the test statistic is t and σ is unknown.
- To test the null hypothesis, the steps involve setting up H_0 , setting up H_1 , selecting an appropriate level of significance α , computing the test statistic and sampling distribution, and concluding the test.
- Three types of hypothesis tests are hypothesis tests about population means, hypothesis tests about population proportions, and hypothesis tests about population variances.
- The degree of freedom is the number of independent variates that make up the statistic.
- The Chi-Square Test considers the square of a standard normal variate.
- The ANOVA test is used for such hypothesis tests that compare the averages of two or more groups.
- The one-way ANOVA test determines, by using variances, if a statistically significant difference exists among several group means.

This concludes “Hypothesis Testing II.”

The next lesson is “Regression.”

Data Science with R

Lesson 10—Regression Analysis



After completing
this lesson, you will
be able to:



- Explain regression analysis
- Describe the different types of regression analysis models
- List the functions to convert non-linear models to linear models

It is a technique used to:

- Estimate a relationship between variables
- Predict the value of one variable (dependent variable) on the basis of other variables (independent variables)



Example:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Here, Y is a dependent variable, whereas β_0 , β_1 , x, and ε are independent variables.

Regression analysis is used in several situations, such as those described below:

Example 1:

Using the data given below, analyze the relation between the size of a house and its selling price for a real state agent.



Size in Sq. Ft.	Price in \$
1400	24500
1600	31200
1700	27800

Example 2:

With the help of the data given below, predict the exam scores of students who study for 7.2 hours.



Hours of Studies	Exam Score
6	53
8	79
7.3	69
7.4	80

Some more examples are:

Example 3:

Based on the expected number of customers and the previous days' data given below, predict the number of burgers that will be sold by a KFC outlet .



No. of Customers	Burgers Sold
798	444
450	324
5067	1054

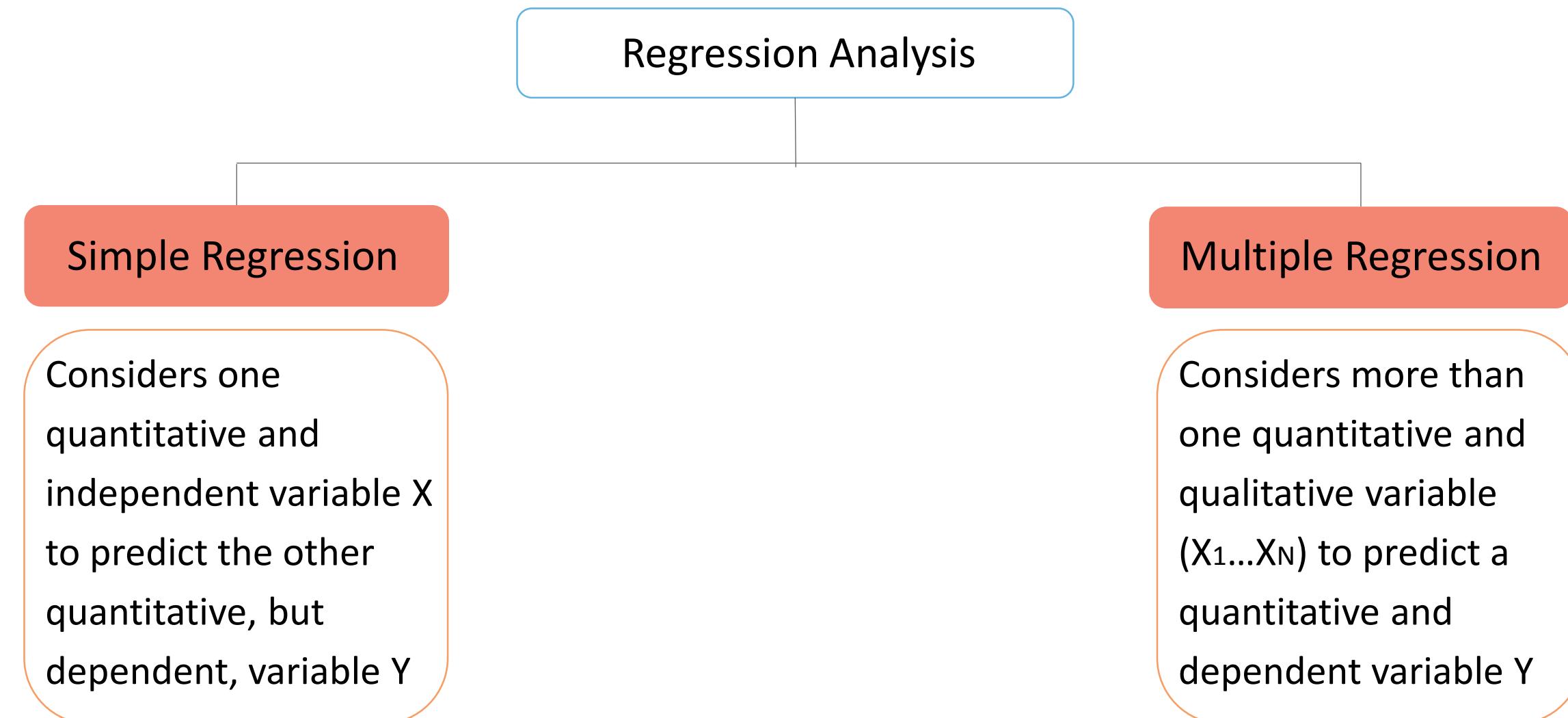
Example 4:

Calculate the life expectancy for a group of people with the average length of schooling based on the data given below:



Average Length of Schooling (in years)	Life Expectancy
14	63
20	80.4
10	57.1

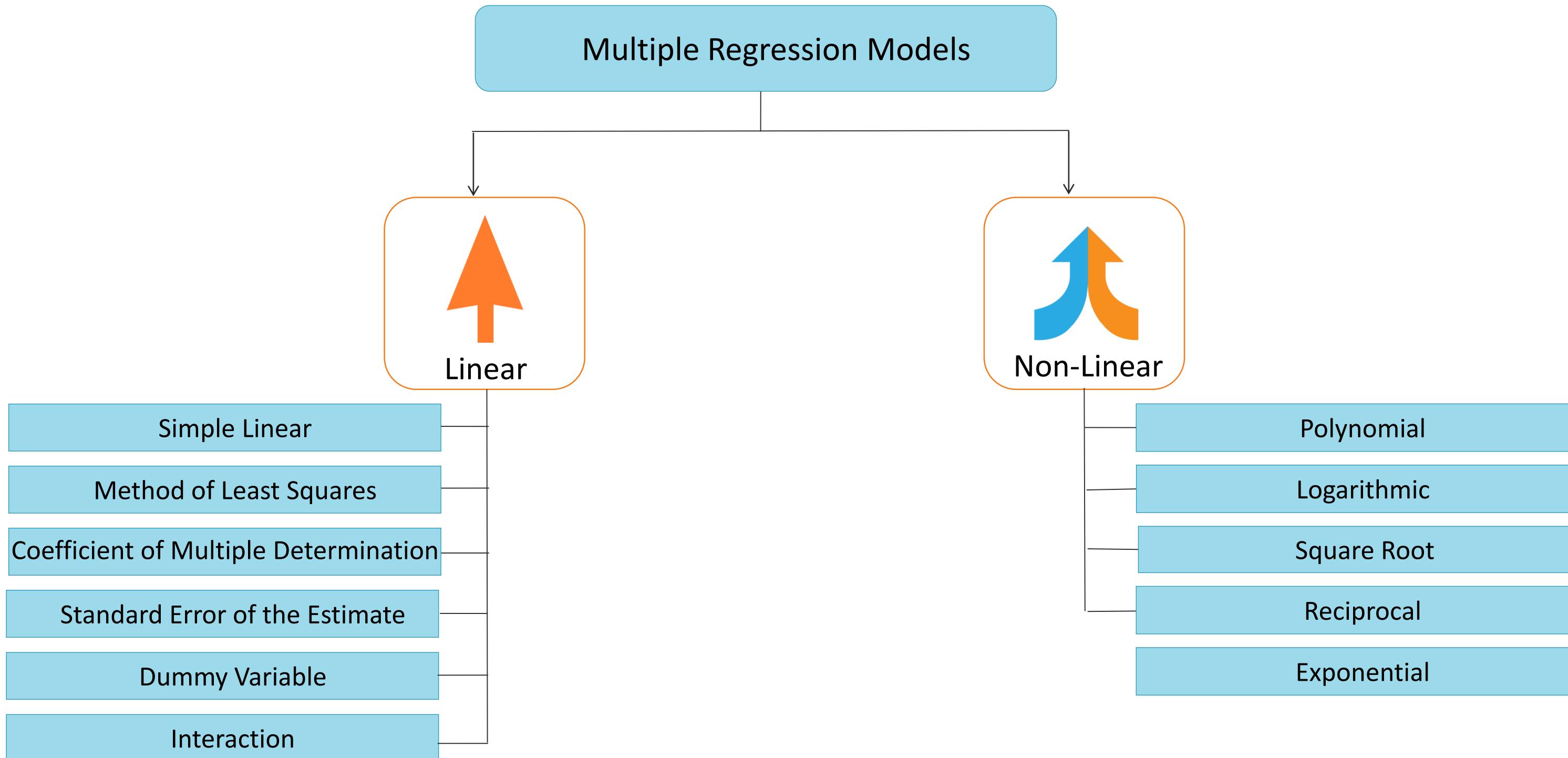
These are:



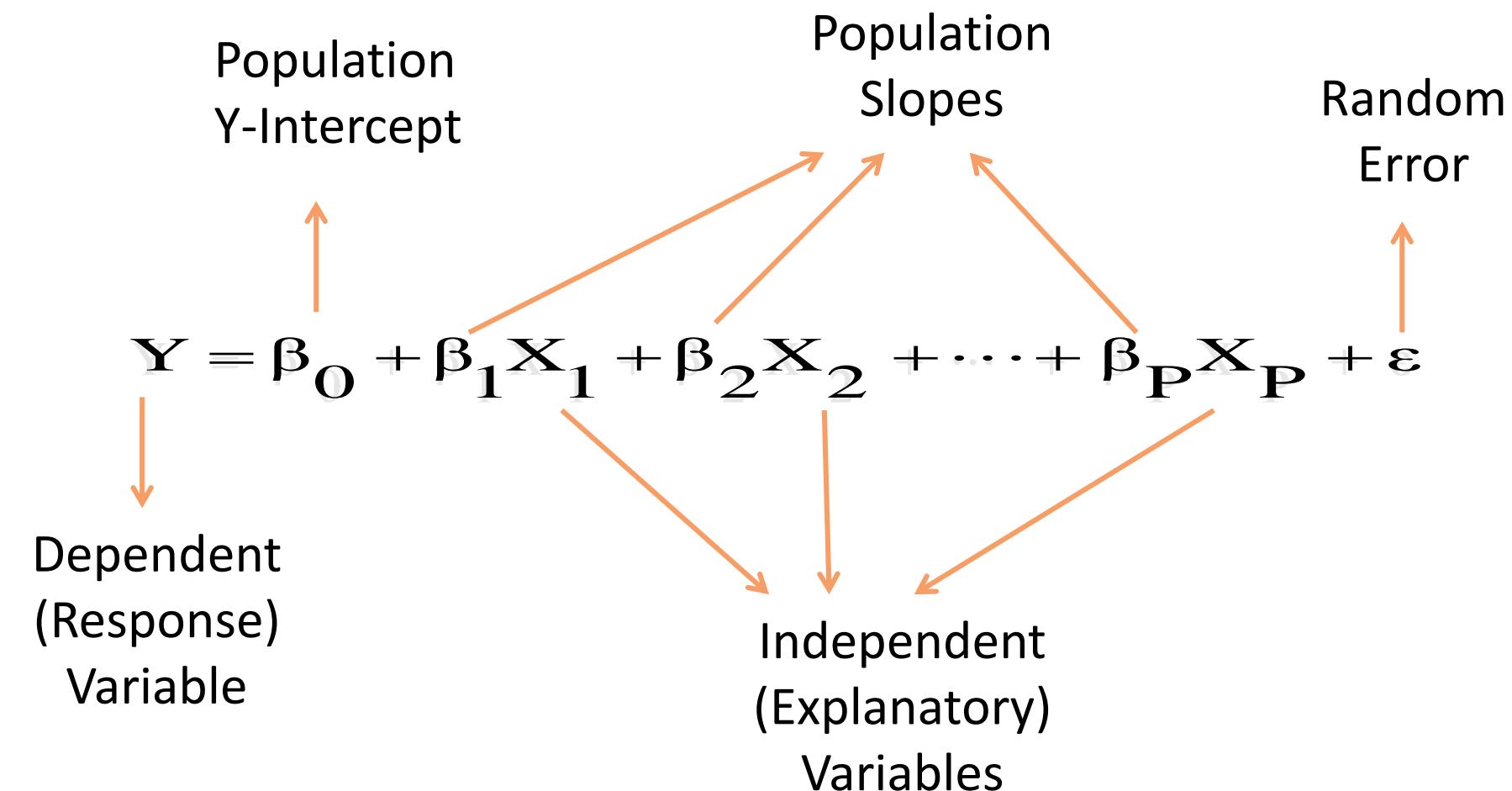
In this analysis:

- A relationship between a dependent variable and an independent variable is modeled
- The output is a function to predict the dependent variable on the basis of the values of independent variables
- A straight line is fit to the data

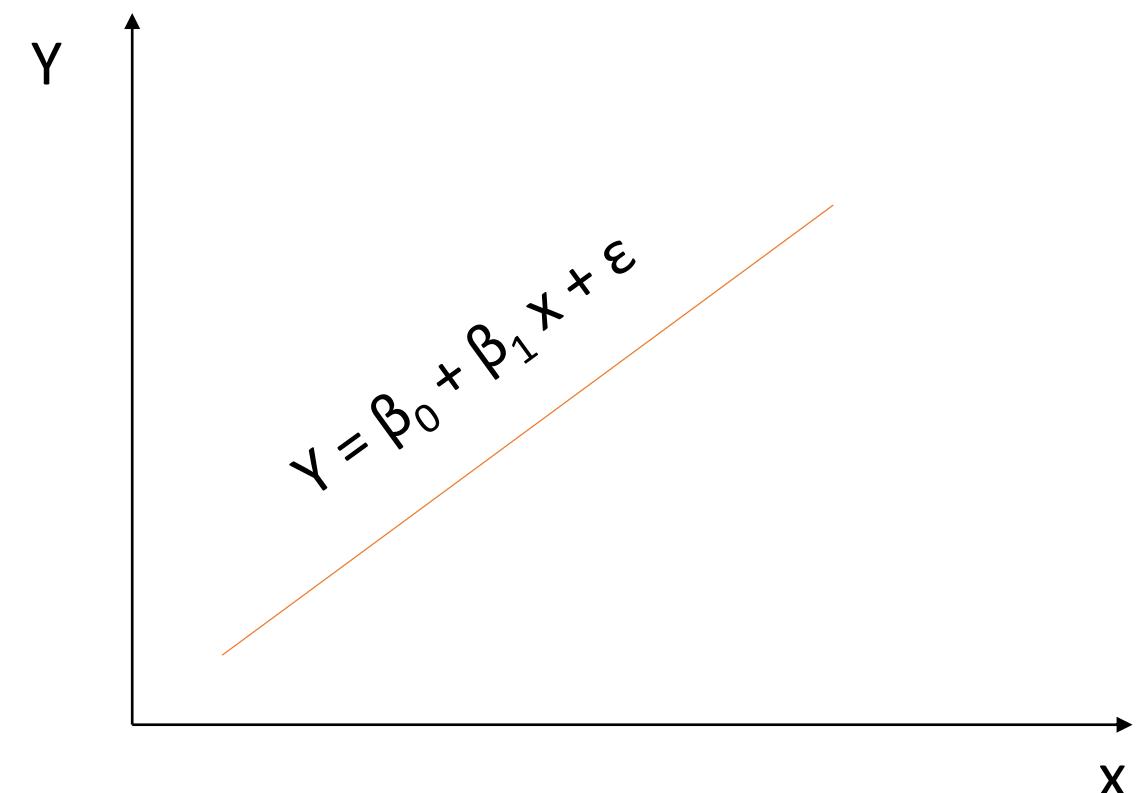
These are:



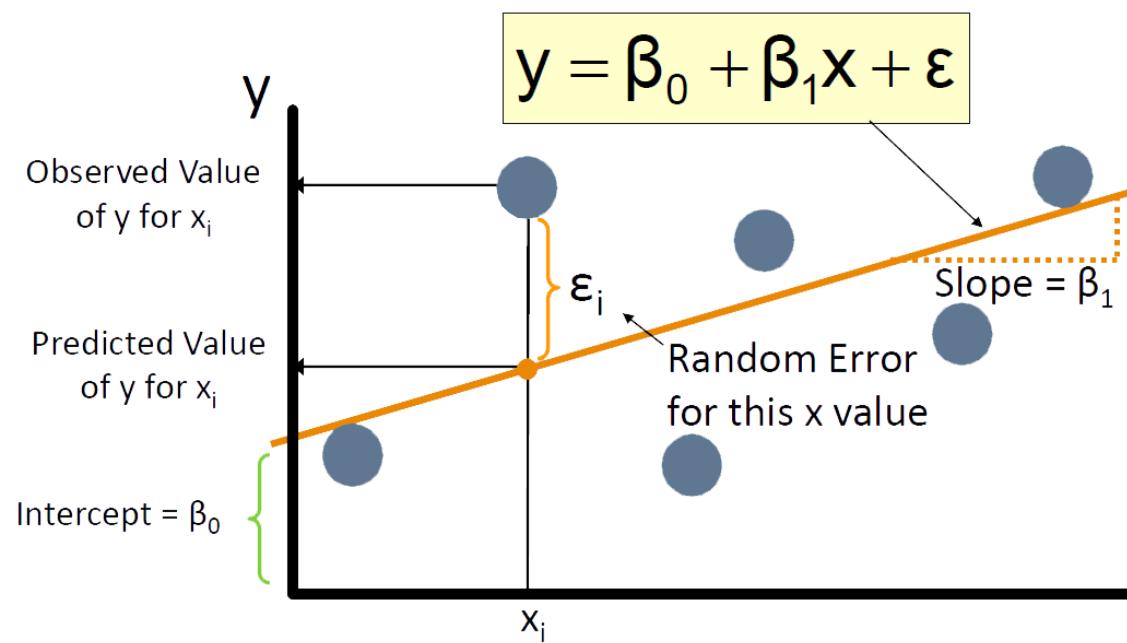
It depicts the relationship between one dependent and two or more independent variables. An example and its components are explained below:



Example of Linear Regression Model Relationships



A more descriptive graphical representation of simple linear regression is given below:



Here:

- β_1 represents the slope. A slope of two variables implies that each one-unit change in x results in a two-unit change in y .
- β_1 represents the estimated change in the average value of y as a result of a one-unit change in x .
- β_0 represents the estimated average value of y when the value of x is zero.

This demo will show the steps to do simple linear regression in R.



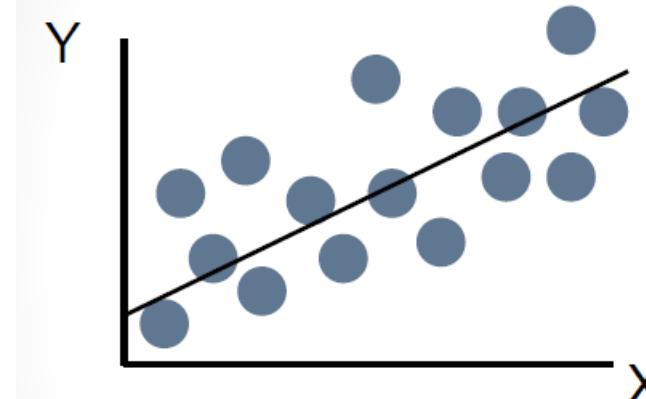
It is:

- Defined as the examination of a linear relationship between the independent (x) and dependent variable (y)
- Denoted by “r”

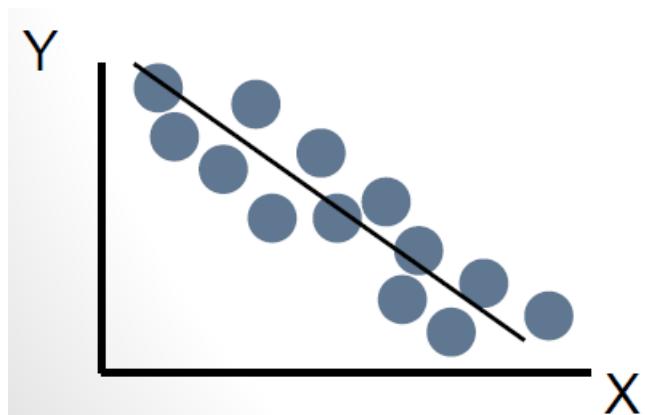
Formula in R: $r = \text{cov}(XY)/sd(x)*sd(y)$

X and Y can exist in three different types of relations:

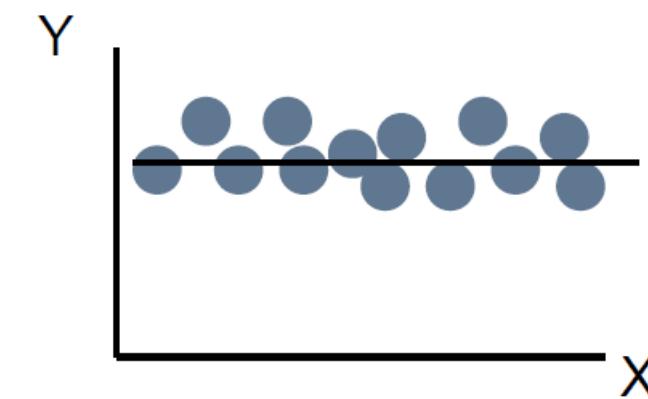
Positive Relation



Negative Relation

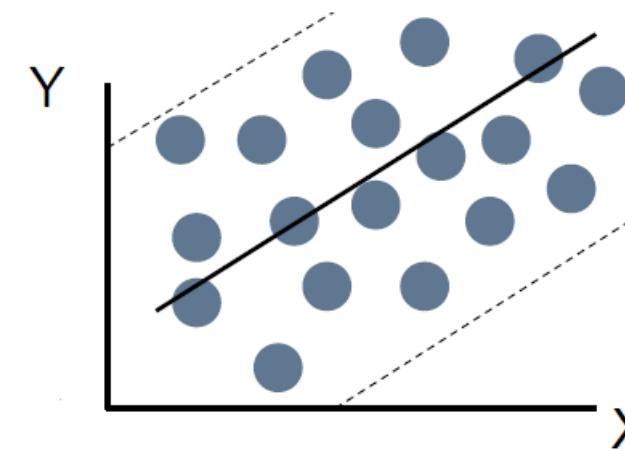


No Relation

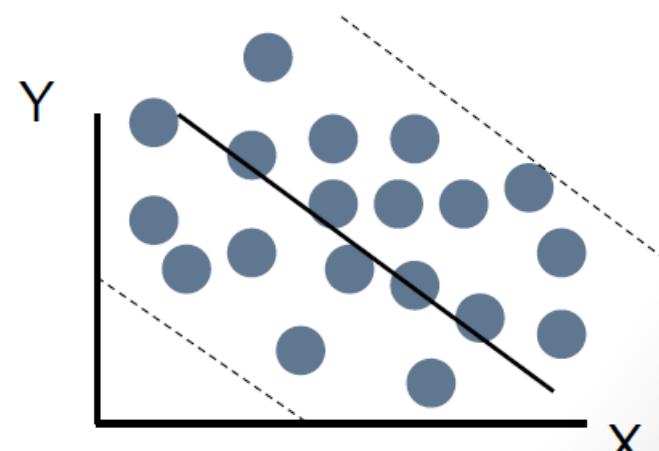


They can also exist in a weak relation:

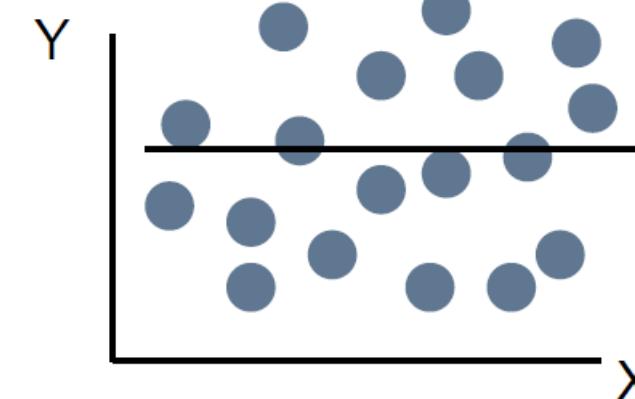
Weak Relation Type I



Weak Relation Type II



Weak Relation Type III



This demo will show the steps to find the correlation between two variables.

Method of Least Squares Regression Model

It selects the line with the lowest total sum of squared prediction errors (Sum of Squares of Error, or SSE).

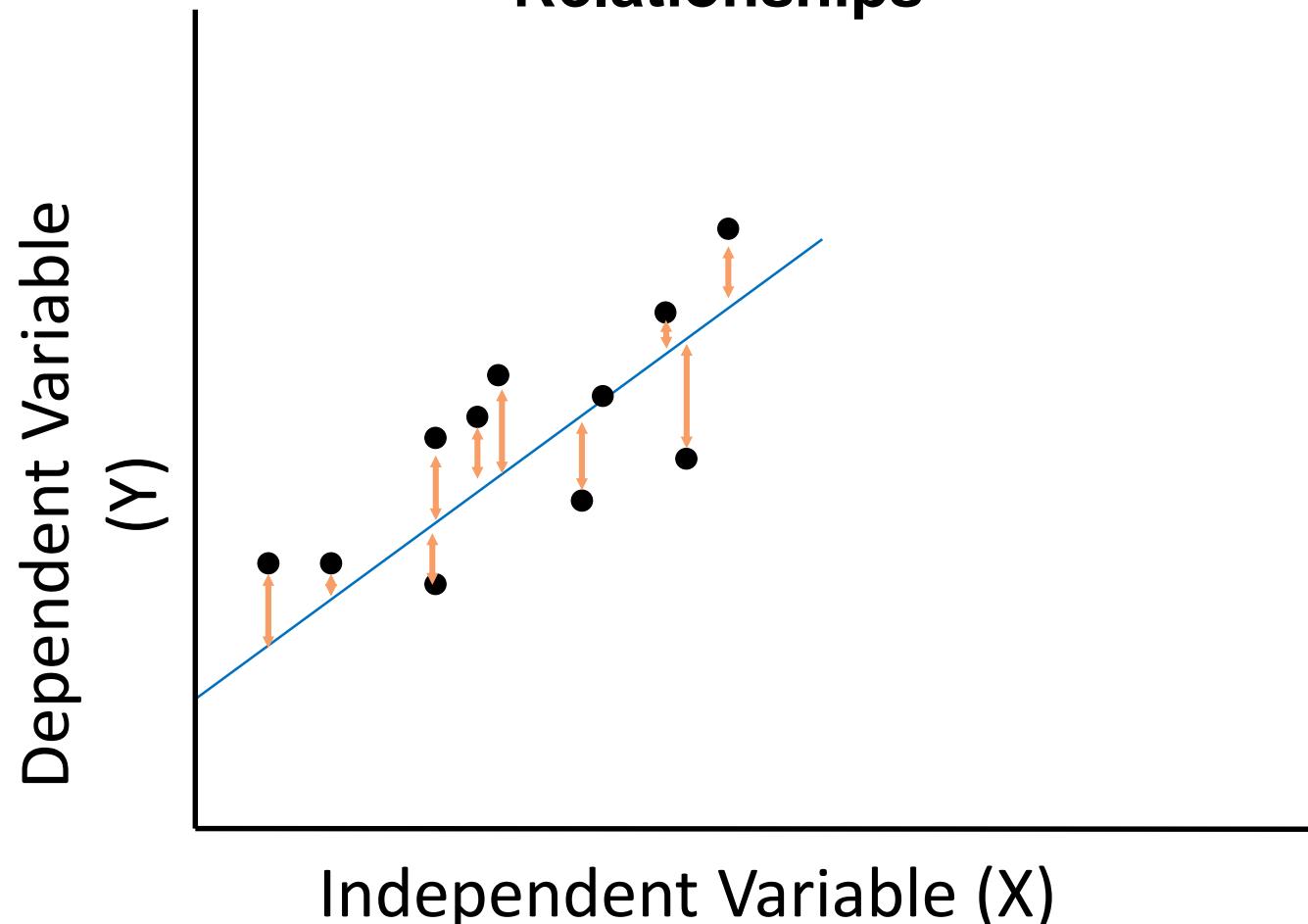
Mathematically,

$$SSR = \sum (\hat{y} - \bar{y})^2 \text{ (measure of an explained variation)}$$

$$SSE = \sum (y - \hat{y})^2 \text{ (measure of an unexplained variation)}$$

$$SST = SSR + SSE = \sum (y - \bar{y})^2 \text{ (measure of the total variation in } y)$$

Example of Method of Least Squares Regression Model Relationships



It:

- Determines the relation between X and Y when you reject H_0
- Is often referred by R

$$R^2 = \frac{SSR}{SST} = \frac{(SSR)}{(SSR+SSE)}$$

$$0 \leq R^2 \leq 1$$

The higher the value, the more accurate the regression model

Standard error of a regression ($S_{y.x}$)

- Is the measure of variability around the line of regression
- Can be used the same way as SD

$$\text{Standard Error} = \sqrt{SSE/(n - k)}$$

Where,

n = Number of observations in the sample

k = Total number of variables in the model

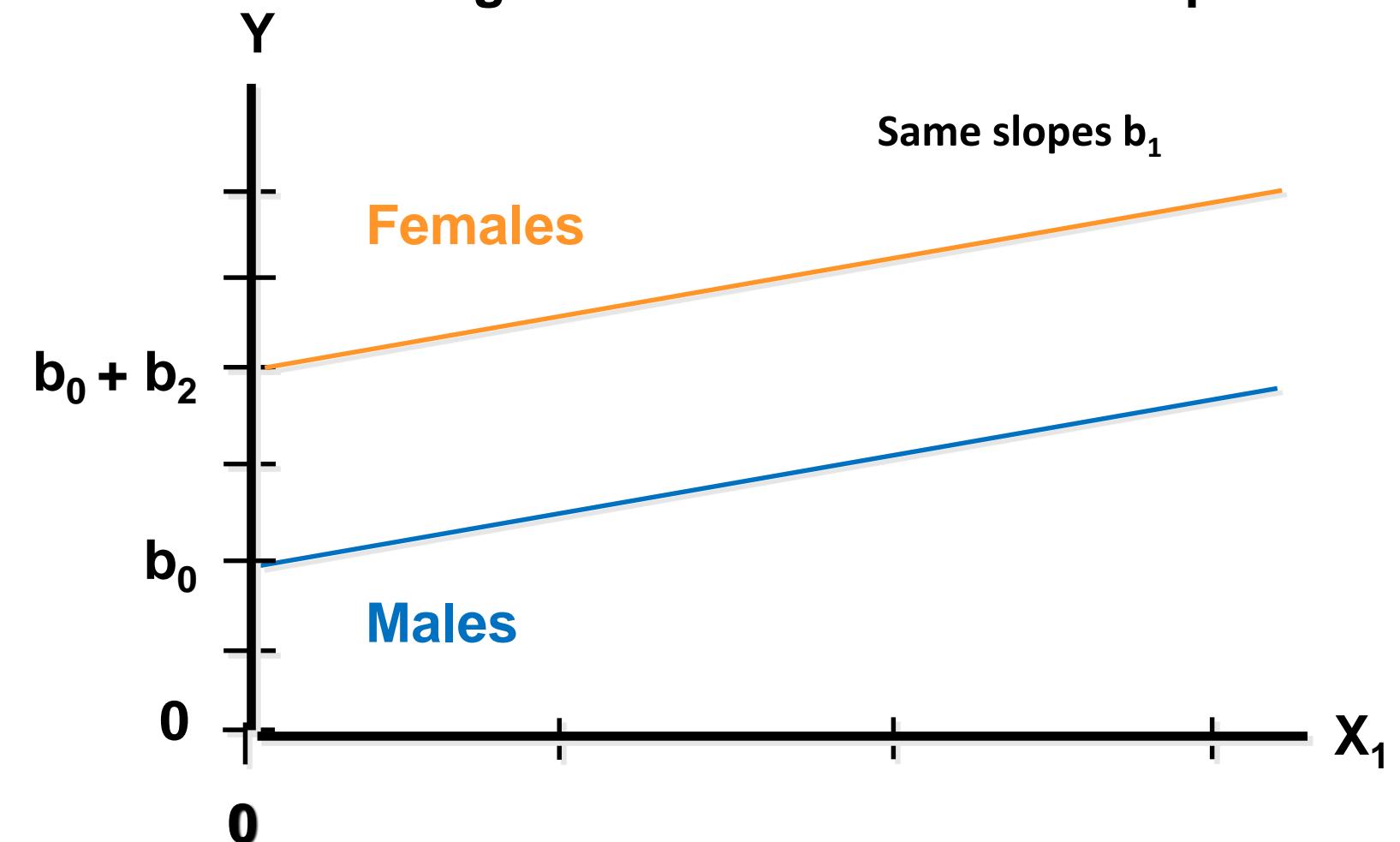


$Y \pm 2$ standard errors provide approximately 95% accuracy. However, 3 standard errors provide a 99% confidence interval.

This model:

- Includes variables, each with two levels coded 0 and 1
- Assumes that only the intercept is different
- Includes slopes that are constant across categories
- Permits the use of qualitative data
- Incorporates large residuals and influence measures

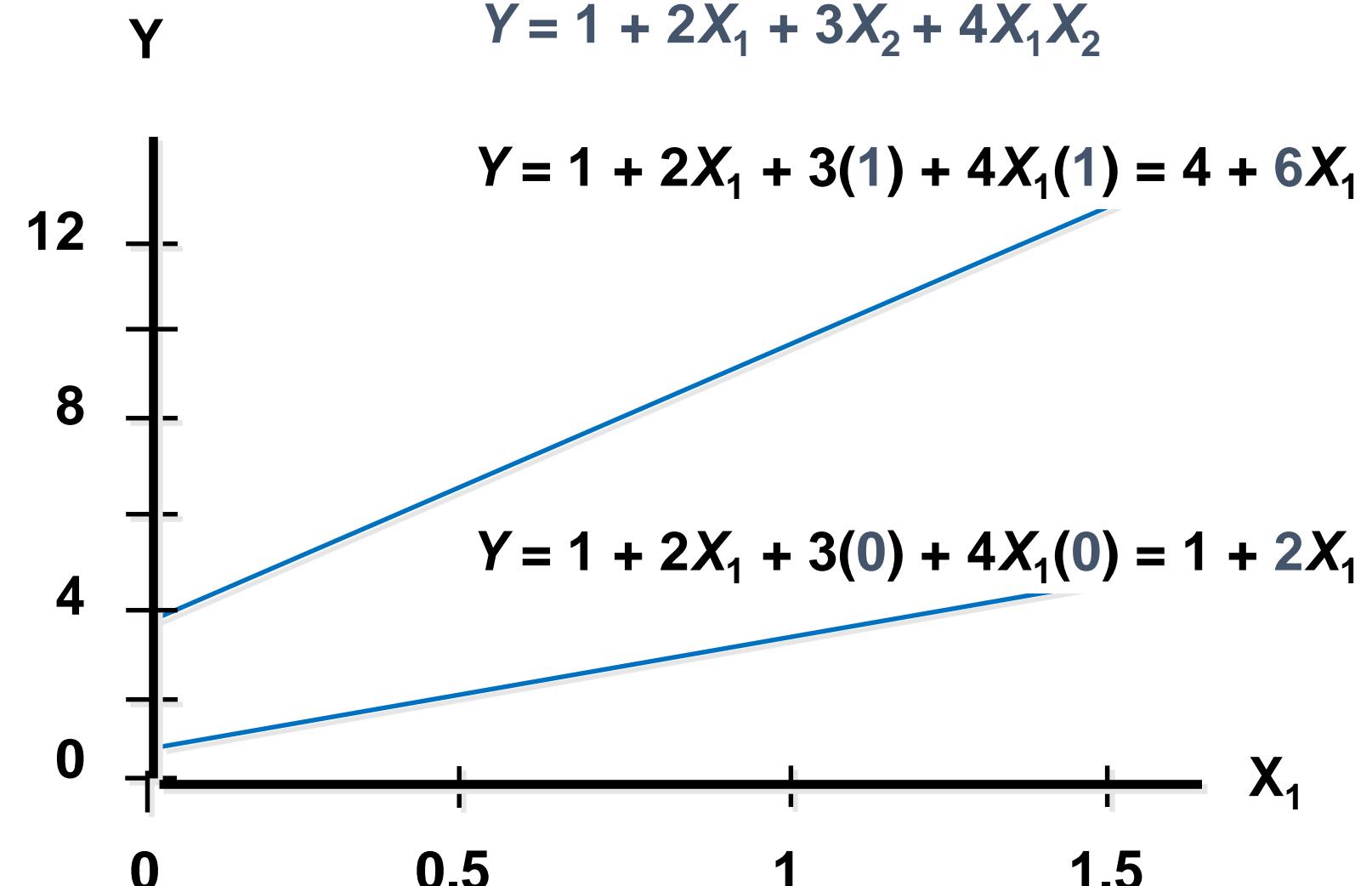
Example of Dummy Variable Regression Model Relationships



This model:

- Assumes the interaction between pairs of X variables
- Includes two-way cross product terms
- Can be combined with other models, such as a dummy variable regression model

Example of Interaction Model Relationships



In $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \varepsilon_i$, without and with the interaction term; the effect of X_1 on Y is measured by β_1 and $\beta_1 + \beta_3 X_2$, respectively.

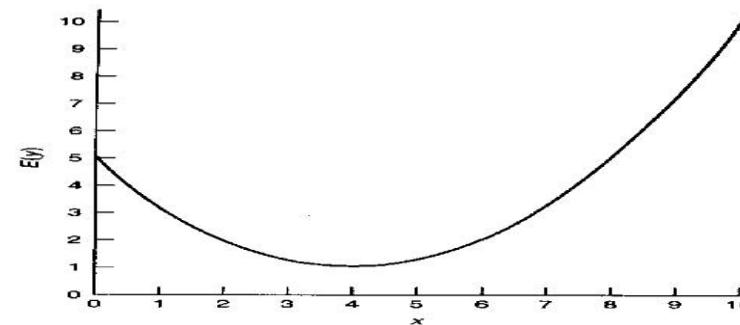
As opposed to linear regression, in this type of regression:

- Number of predictors is large
- Prior to fitting a regression model with all the predictors, stepwise or best subsets model-selection methods are used. This is done to screen out predictors that are not associated with the responses



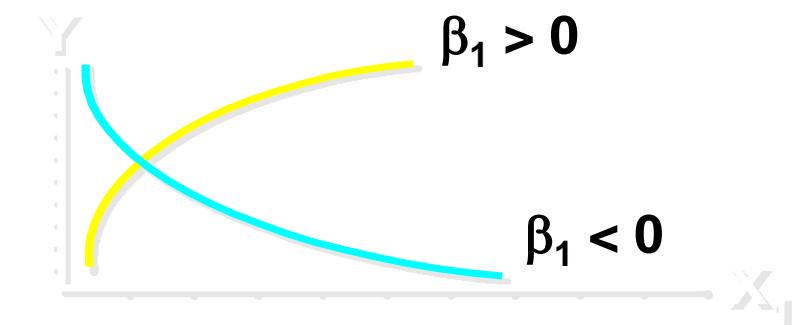
Let's see these models that are explained graphically here, along with their equations:

Polynomial Regression Model



$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

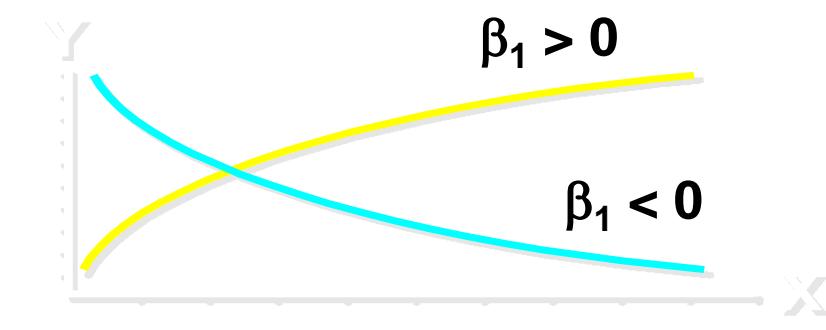
Logarithmic Regression Model



$$Y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \varepsilon$$

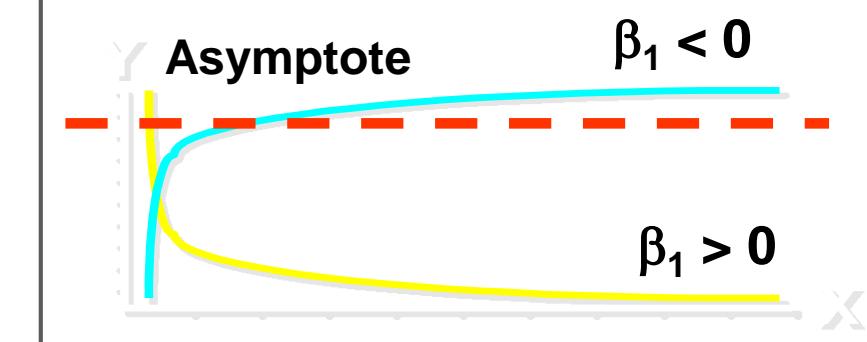
Here, square root and reciprocal models are explained graphically, along with their equations:

Square Root Regression Model



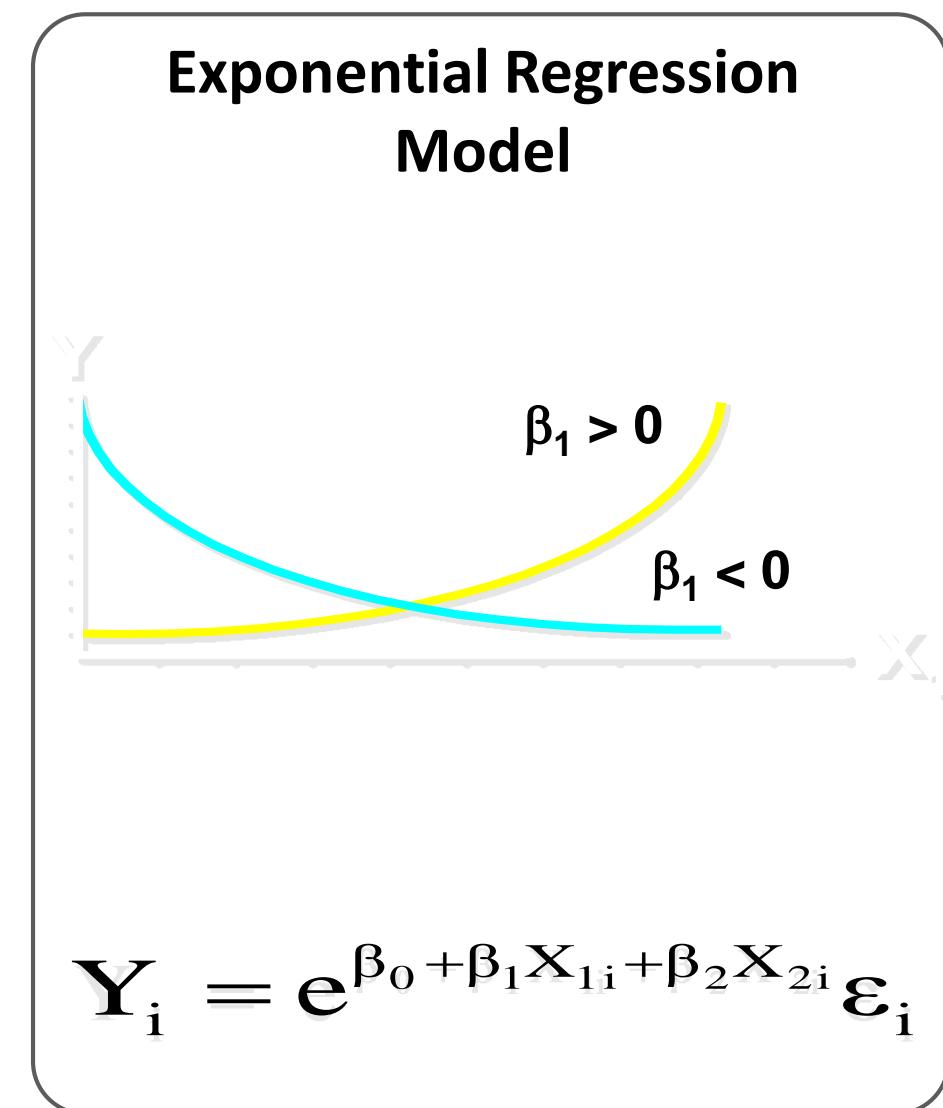
$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \beta_2 \sqrt{X_{2i}} + \varepsilon_i$$

Reciprocal Regression Model



$$Y_i = \beta_0 + \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{1}{X_{2i}} + \varepsilon_i$$

Following is the graphical explanation of an Exponential model, along with its equation:



This demo will show the steps to do regression analysis with multiple variables.

You can convert non-linear models to linear models by applying the following functions:

Exponential

Logarithmic

Trigonometric

Power

You can use the following iterative algorithms for the estimation parameters of complex non-linear models:

Gauss Newton Algorithm

Gradient Descent Algorithm

Levenberg – Marquardt Algorithm



QUIZ
1

The independent variable x is also called the _____.

- a. explanatory variable
- b. coefficient
- c. residual

target variable



QUIZ
1

The independent variable x is also called the _____.

- a. explanatory variable
- b. coefficient
- c. residual

target variable



The correct answer is **a.**

Explanation: The independent variable x is also called the explanatory variable.

QUIZ
2

State whether the following statement is “True” or “False.”

Linear regression is possible when the target variable is a categorical variable.

- a. True
- b. False



QUIZ
2

State whether the following statement is “True” or “False.”

Linear regression is possible when the target variable is a categorical variable.

- a. True
- b. False



The correct answer is **b**.

Explanation: Linear Regression analysis is used when one variable is dependent on another variable.

QUIZ
3*Fill in the blank:*

The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the _____.

- a. explanatory variable
- b. coefficient
- c. residual

target variable



QUIZ
3*Fill in the blank:*

The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the _____.

- a. explanatory variable
- b. coefficient
- c. residual

target variable



The correct answer is **c.**

Explanation: The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the residual.

QUIZ
4

Which of the following statements about the Method of Least Squares Regression Model is correct? *Select all that apply.*

- a. It selects the line with the mean total sum of squared prediction errors.
- b. It is a linear regression model.
- c. It selects the line with the lowest total sum of squared prediction errors.



It determines the relation between X and Y, when you reject H₀.

QUIZ
4

Which of the following statements about the Method of Least Squares Regression Model is correct? *Select all that apply.*

- a. It selects the line with the mean total sum of squared prediction errors.
- b. It is a linear regression model.
- c. It selects the line with the lowest total sum of squared prediction errors.



It determines the relation between X and Y, when you reject H₀.

The correct answers are **b** and **c**.

Explanation: Method of Least Squares Regression Model is linear regression model that selects the line with the lowest total sum of squared prediction errors.

Let us summarize the topics covered in this lesson:

- Regression analysis is used to estimate the relationship between variables.
- Simple regression considers one quantitative and independent variable X to predict the other quantitative, but dependent, variable Y.
- Multiple regression considers more than one quantitative and qualitative variable ($X_1 \dots X_N$) to predict a quantitative and dependent variable Y.



Let us summarize the topics covered in this lesson:



- Multiple regression has the following types of models:
- Linear Models:
 - Simple Linear
 - Method of Least Squares
 - Coefficient of Multiple Determination
 - Standard Error of the Estimate
 - Dummy Variable
 - Interaction
- Non-Linear Models:
 - Polynomial
 - Logarithmic
 - Square Root
 - Reciprocal
 - Exponential
- You can express non-linear models to linear models by applying functions.

This concludes “Regression Analysis.”

The next lesson is “Classification.”

Data Science with R

Lesson 11—Classification



After completing
this lesson, you will
be able to:



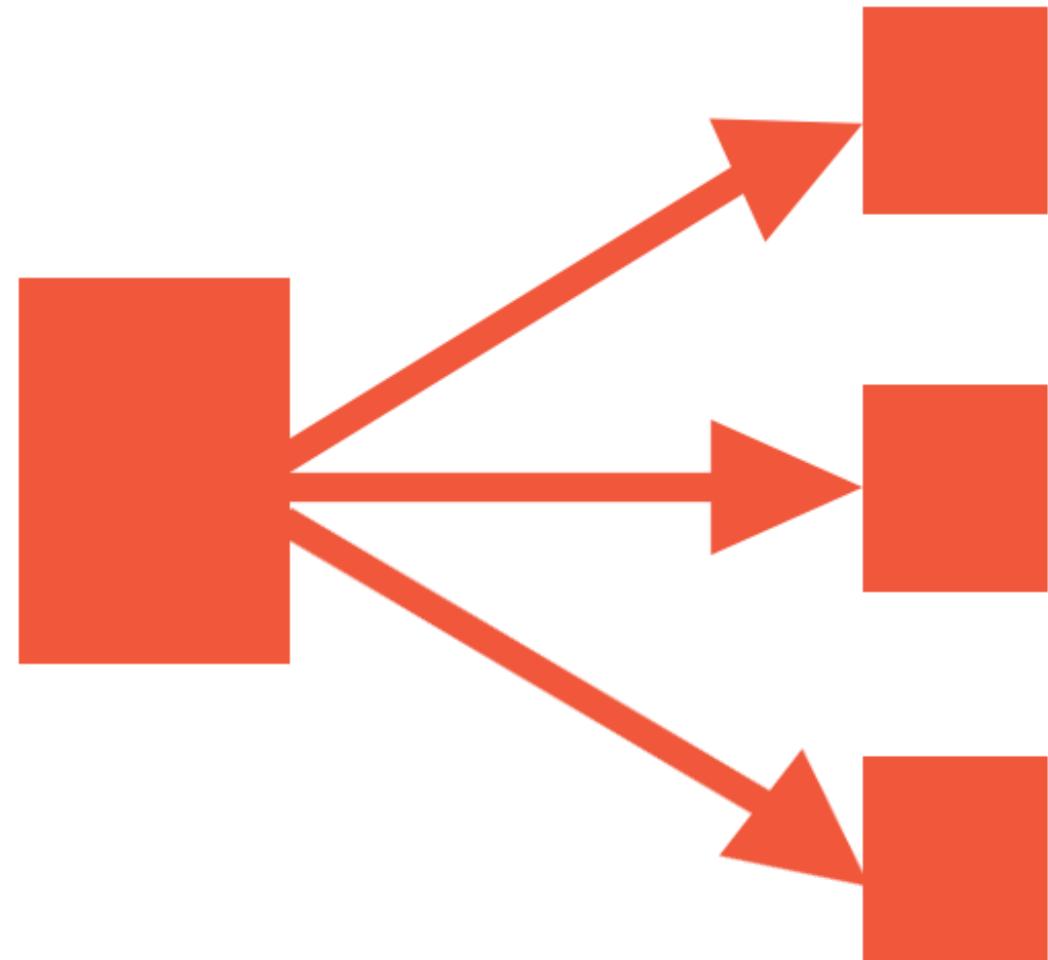
- Explain classification
- Describe the classification system and process
- List the various issues related to classification and prediction
- Explain the various classification techniques

It is a technique to:

- Determine the extent to which a thing will or will not be a part of a category or type
- Use specific information (input) to select a single response (output) from a list of predetermined potential responses

Typical Applications:

- Target marketing
- Credit approval
- Fraud detection
- Medical diagnosis

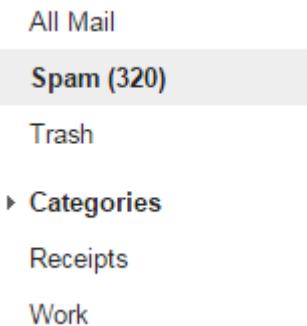


A few examples of classification are:



Apple iTunes

Uses classification to categorize songs into potential playlists



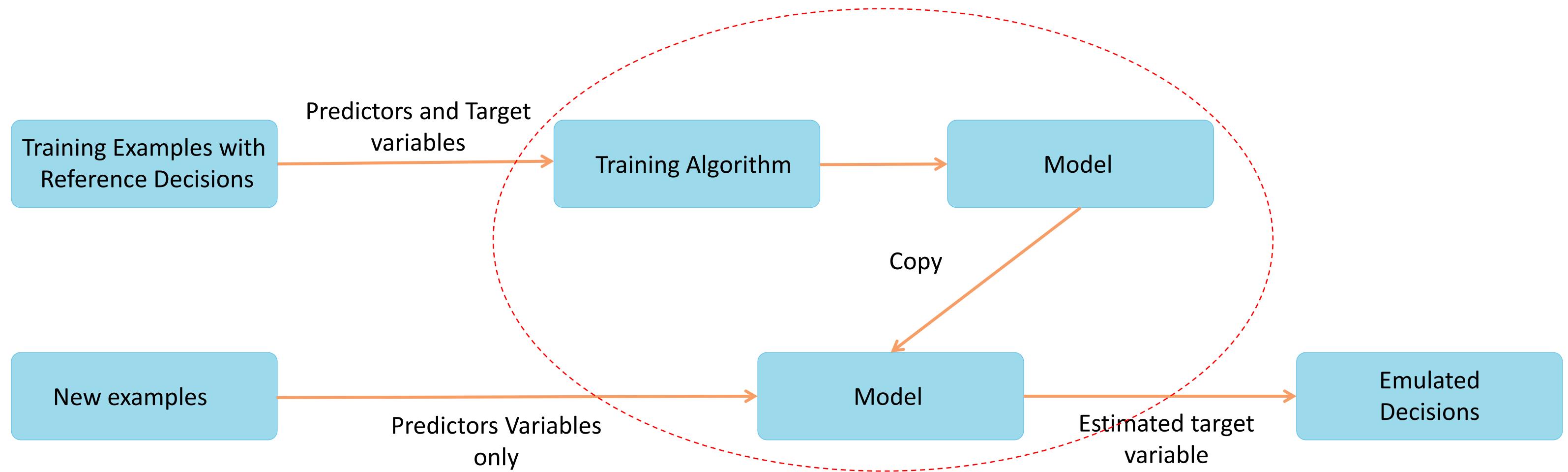
Google and Yahoo

Use classification to decide if a message is spam, based on prior emails, spam reports, and email characteristics

Here's the difference between classification and prediction:

Classification	Prediction
Predicts categorical class labels, categorizes data on the basis of the training sets and class labels, and uses them to classify new data	Foretells unknown or missing values

The classification system is depicted below:



The classification process includes model construction and model usage as its two techniques for prediction:



Model Construction

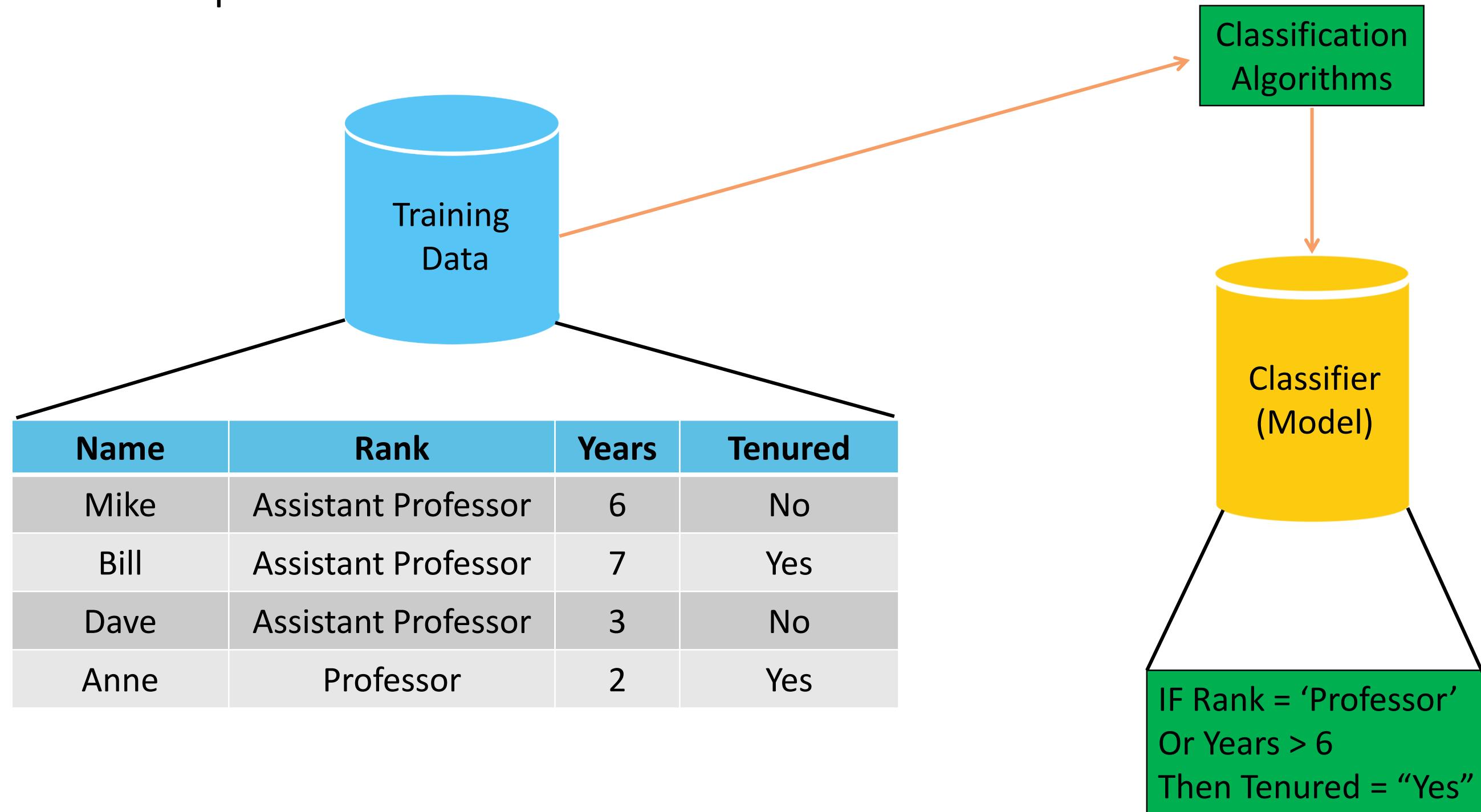
It is done to describe a set of predetermined classes. Every tuple or sample belongs to a predefined class. The model is represented as decision trees, classification rules, or mathematical formulae.



Model Usage for Prediction

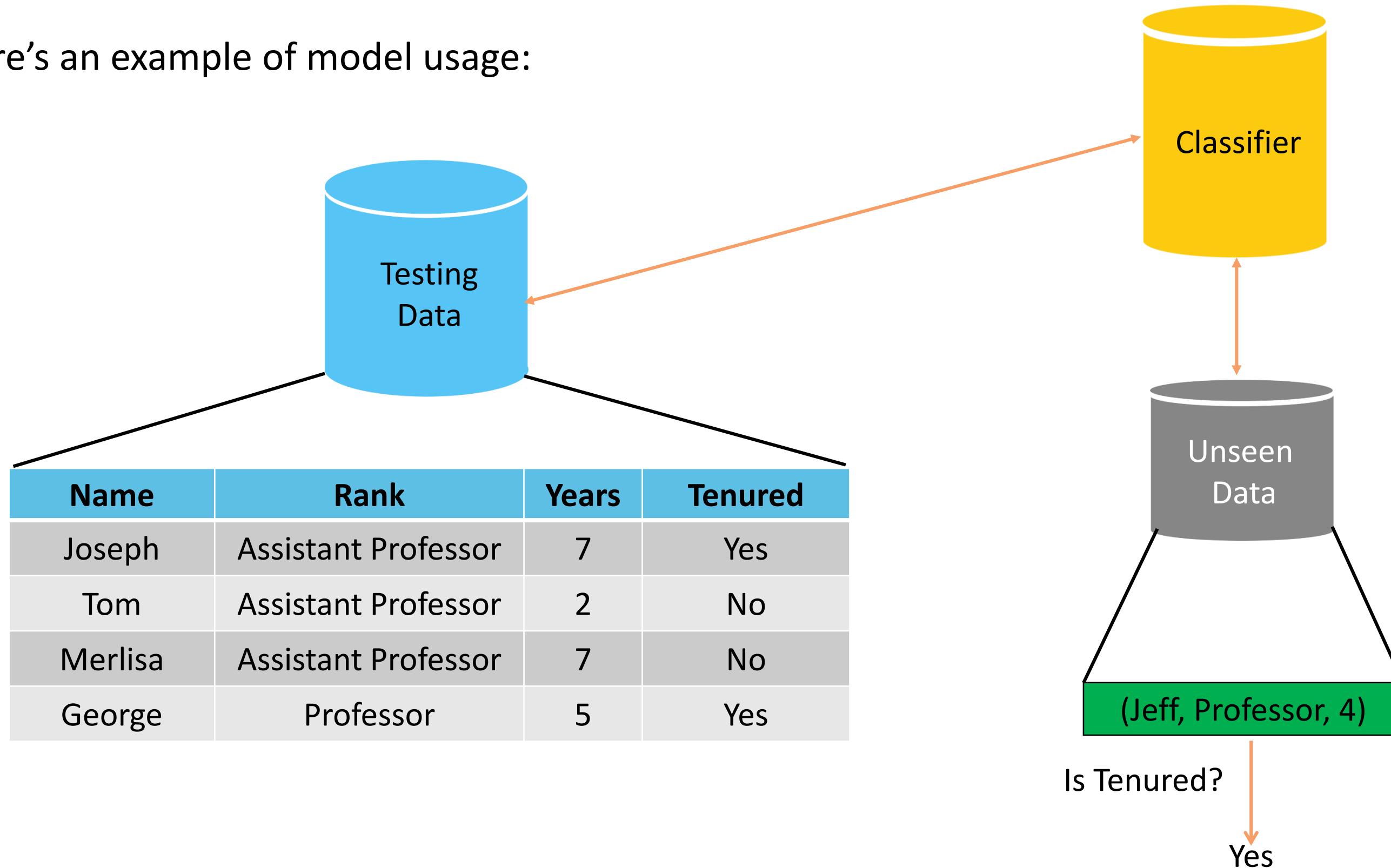
It is done to classify unknown or future objects and to estimate the accuracy of a model. The accuracy rate is the percentage of the test set samples correctly classified by the model. In case of acceptable accuracy, the model is used for classifying data tuples with unknown class labels.

Here's an example of model construction:



Classification Process—Model Usage in Prediction

Here's an example of model usage:



The two issues regarding the classification and prediction of data are:

- Data Preparation Issues
- Evaluating Classification Methods Issues

These issues are related to:

Data Cleaning

Preprocess data to handle missing values and reducing noise

Relevance Analysis

Eliminate the redundant or irrelevant attributes

Data Transformation

Normalize and/or generalize data

These issues are related to:

Accuracy

Determine the predictor and classifier accuracy

Speed and Scalability

Handle time to construct and use the model

Robustness

Handle missing values and noise of values

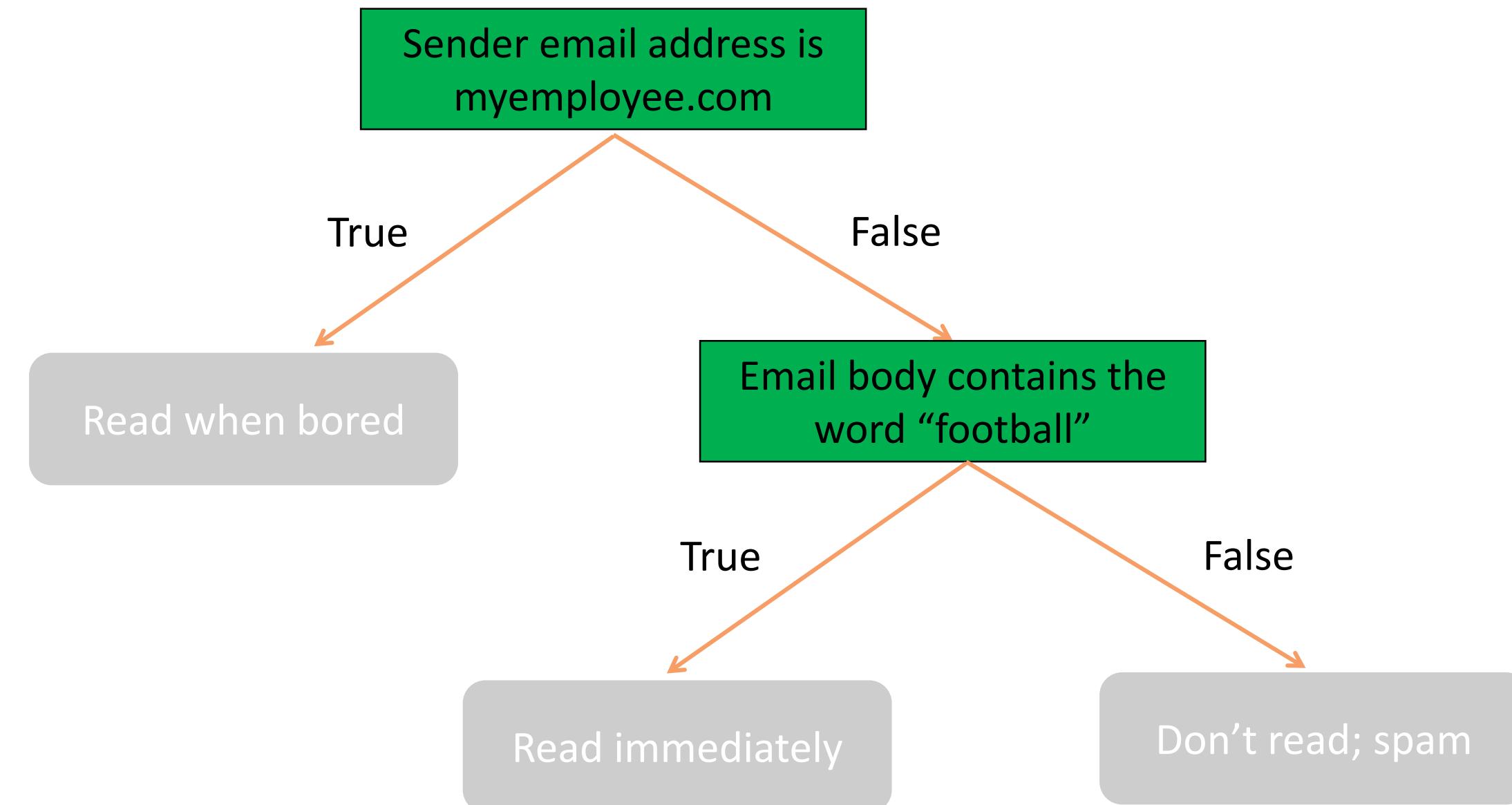
Scalability

Evaluate efficiency in disk-resident databases

Interpretability

Understand the insight provided by the model

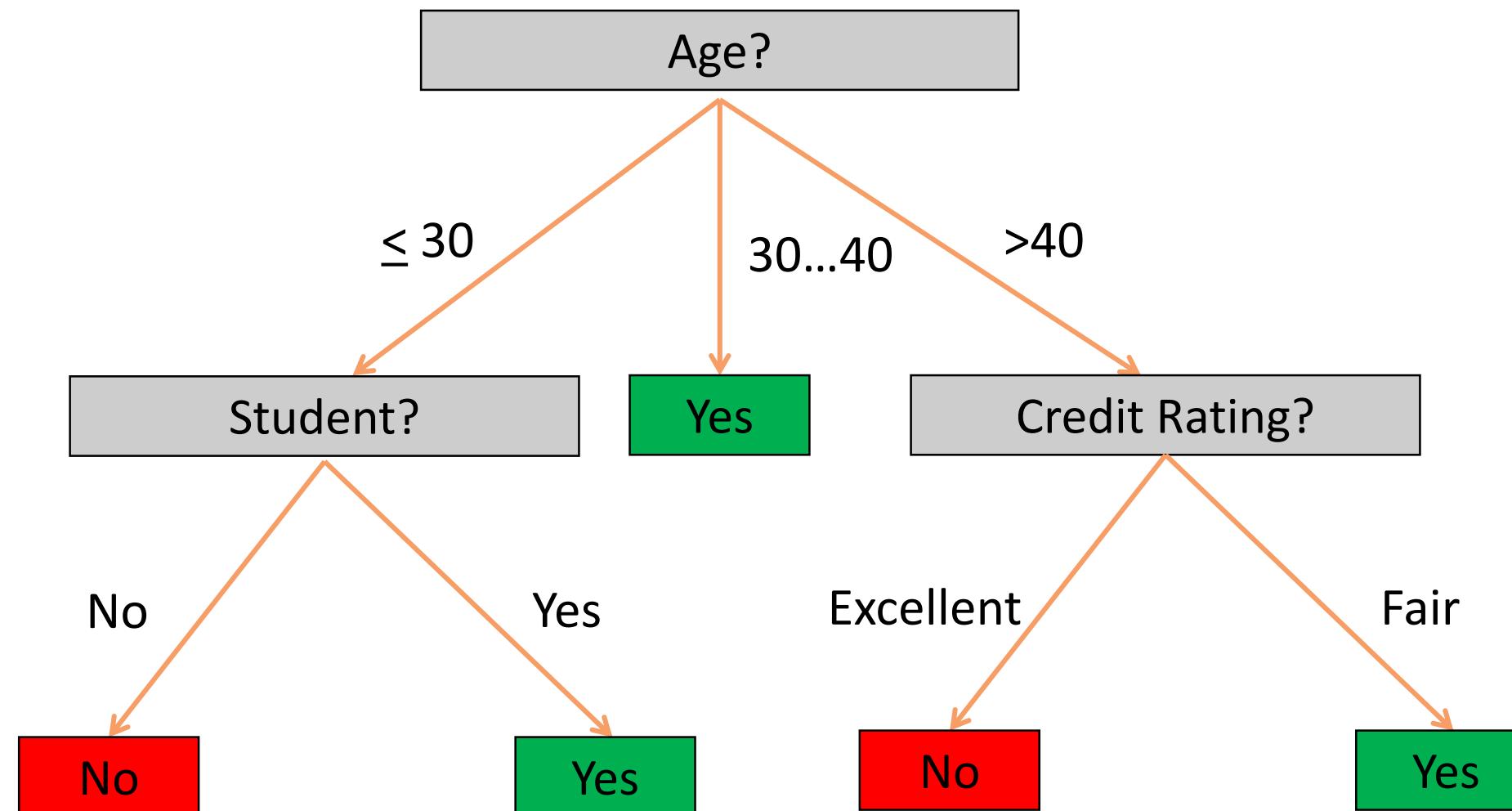
It is one of the most common classification techniques. An example is given below:



To understand a decision tree better, consider the given “Buy Computer” dataset:

Age	Income	Student	Credit Rating	Buys Computer
≤ 30	High	No	Fair	No
< 30	High	No	Excellent	No
31...40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

As an output of the dataset, the following decision tree can be created:



In these rules:

- The statements are represented as IF-THEN rules
- There is, at least, one rule for every path from the root to a leaf in a tree
- A conjunction is formed for every attribute-value pair along a path in a tree
- The class prediction is held by the leaf node in a tree



Let's apply these rules on the "Buy Computer" dataset:

IF Age = " ≤ 30 " AND Student = "No" THEN buys_computer = "No"

IF Age = " ≤ 30 " AND Student = "Yes" THEN buys_computer = "Yes"

IF Age = "31...40" THEN buys_computer = "Yes"

IF Age = ">40" AND Credit Rating = "Excellent" THEN buys_computer = "Yes"

IF Age = " ≤ 30 " AND Credit Rating = "Fair" THEN buys_computer = "No"

Sometimes, a tree may overfit the training data, which can lead to issues, such as:

- Too many branches
- Less accurate and unseen samples



How to avoid overfitting?

There are two approaches:

- **Prepruning:** Stop the construction of a tree early. If the goodness measure is falling below a threshold, do not split the node.
- **Postpruning:** In case, selecting an appropriate threshold is difficult, remove branches from a fully-developed tree by getting a progressively pruned trees' sequence.

The three tips to determine the final tree size are:

Tip 1

Separate training (2/3) and testing (1/3) sets

Tip 2

Apply cross-validation

Tip 3

Use a statistical test (for example, chi-square) to determine whether pruning or expanding a node can improve the distribution

A tree is constructed in a top-down manner and includes the following steps:

Place all training examples at the root

→ Categorize the attributes

→ Partition examples recursively based on the selected attributes

→ Select test attributes on the basis of a heuristic or statistical measure

Conditions to stop partitioning:

- For a node, all samples belong to the same class.
- No attributes are left for further partitioning.
- No samples are left for classification.

You need to select an attribute with the highest information gain, which is defined as:

$$\text{Gain}(A) = I(s_1, s_2, \dots, s_m) - E(A)$$

Where,

S contains s_i tuples of class C_i for $i = \{1, \dots, m\}$

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m \frac{s_i}{S} \log_2 \frac{s_i}{S}$$

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{S} I(s_{1j}, \dots, s_{mj}) \quad (\text{entropy of attribute A with values } \{a1, a2, \dots, av\})$$

Now, let's consider the “Buy Computer” dataset, and calculate $\text{Gain}(A)$.

Assume:

Class P: Buy Computer = “Yes”

Class N: Buy Computer = “No”

The attributes of the “Buy Computer” table can be categorized as:

Age	p_i	n_i	$I(p_i, n_i)$
≤ 30	2	3	0.971
>40	3	2	0.971
31...40	4	0	0

$$\text{So, } I(p, n) = I(9, 5) = 0.940$$

The entropy to identify the age will be calculated as follows:

$$\begin{aligned} E(\text{age}) &= \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) \\ &\quad + \frac{5}{14} I(3,2) = 0.694 \end{aligned}$$

Therefore, Gain(Age) will be calculated as follows:

$$\text{Gain}(\text{age}) = I(p,n) - E(\text{age}) = 0.246$$

Similarly,

$$\text{Gain}(\text{income}) = 0.029$$

$$\text{Gain}(\text{student}) = 0.151$$

$$\text{Gain}(\text{credit_rating}) = 0.048$$

Assume that A is a continuous-valued attribute. To calculate its Information Gain, you must determine the best split point for A (threshold on A) by:

1. Sorting the values of A in an increasing order
2. Selecting the midpoint between each pair of adjacent values
3. Selecting a point with the minimum expected information requirement for A as the split point

Important Points:

- D_1 represents the set of tuples in D that satisfy $A \leq$ split point
- D_2 represents the set of tuples in D that satisfy $A >$ split point



You can enhance a basic tree using the following methods:

Use continuous-valued attributes

Define new discrete-valued attributes dynamically to partition the continuous valued attributes into a discrete set of intervals

Manage missing attribute values

Assign the most common attribute value and probability to each feasible value

Create attributes

Create new attributes on the basis of existing and sparsely represented attributes to reduce repetition, replication, and fragmentation

Data trees are used in data mining because they:

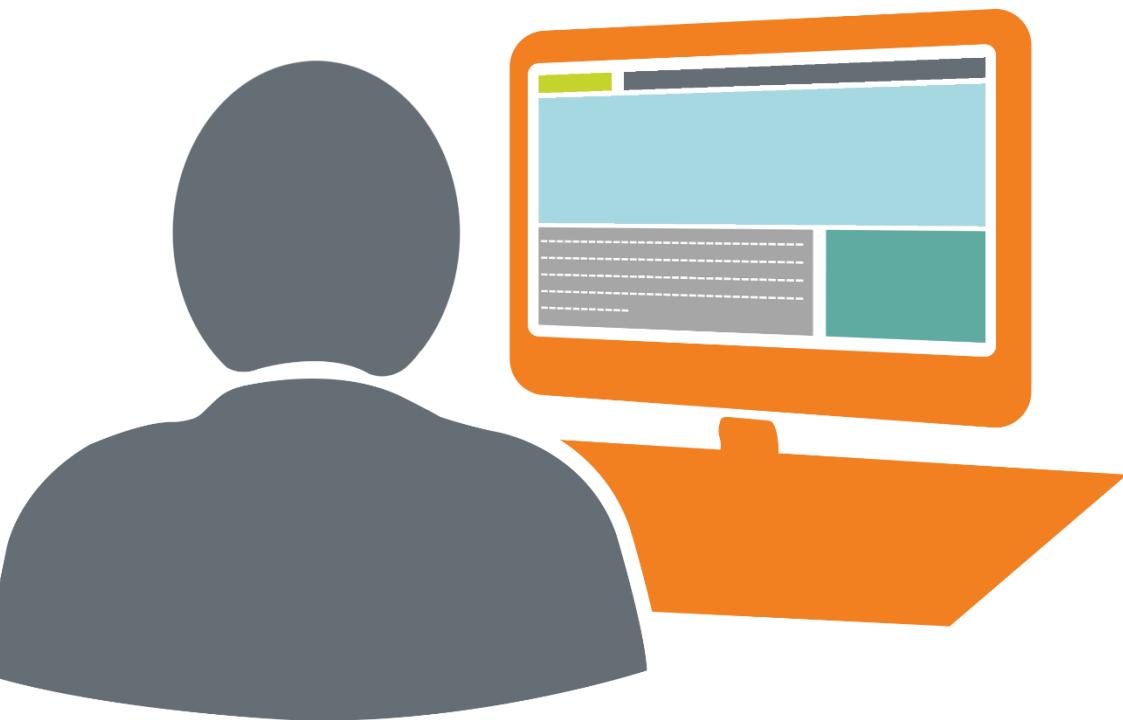
- Have a faster learning speed than other classification methods
- Can be converted to easy and simple classification rules
- Can use SQL queries
- Have a high classification accuracy



This demo will show the steps to model a decision tree.

This:

- Is a probabilistic model
- Assumes conditional independence between features



This model has the following features:

Probabilistic Learning

Determines explicit probabilities for hypothesis

Incremental

Allows each training example to incrementally increase or decrease the probability that a hypothesis is correct

Standard

Provides a standard of optimal decision making to measure other methods

Probabilistic Prediction

Predicts various hypotheses that are weighted by their probabilities

Assume:

X = Data sample with unknown class label

H = A hypothesis that X belongs to class C

For classification, you need to determine:

- $P(H|X)$: Probability that the hypothesis holds, given the observed data sample X
- $P(H)$: Prior probability of hypothesis H
- $P(X)$: Probability that the sample data is observed
- $P(X|H)$: Probability of observing the sample X, given that the hypothesis holds



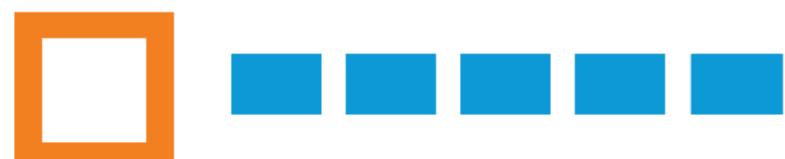
Following the Bayesian Theorem:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$



Informally:

posteriori = likelihood \times prior/evidence



MAP (maximum posteriori) hypothesis:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D) = \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h).$$



This theorem has certain practical difficulties, such as the requirement of initial knowledge of many probabilities and a huge computational cost.

Assuming that attributes are conditionally independent, Naive Bayes Classifier can be calculated as:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$



Points to Remember:

- The product of occurrence of two elements can be defined as:
 $P([y_1, y_2], C) = P(y_1, C) * P(y_2, C)$
- The classifier reduces the computation cost.
- Once $P(X|C_i)$ is known, you can assign X to the class with maximum $P(X|C_i) * P(C_i)$.

Applying Naive Bayes Classifier—Example

To understand the classifier, consider the “Buy Computer” dataset:

Age	Income	Student	Credit Rating	Buys Computer
≤ 30	High	No	Fair	No
< 30	High	No	Excellent	No
31...40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31...40	Low	Yes	Excellent	Yes
≤ 30	Medium	No	Fair	No
≤ 30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
≤ 30	Medium	Yes	Excellent	Yes
31...40	Medium	No	Excellent	Yes
31...40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

Consider the data sample given below:

$X = (\text{Age} \leq 30, \text{Income} = \text{Medium}, \text{Student} = \text{Yes}, \text{Credit Rating} = \text{Fair})$

Let's compute $P(X|C_i)$ for each class:

$$P(\text{Age} = \text{"<30"} | \text{buys_computer} = \text{"Yes"}) = 2/9 = 0.222$$

$$P(\text{Age} = \text{"<30"} | \text{buys_computer} = \text{"No"}) = 3/5 = 0.6$$

$$P(\text{Income} = \text{"Medium"} | \text{buys_computer} = \text{"Yes"}) = 4/9 = 0.444$$

$$P(\text{Income} = \text{"Medium"} | \text{buys_computer} = \text{"No"}) = 2/5 = 0.4$$

$$P(\text{Student} = \text{"Yes"} | \text{buys_computer} = \text{"Yes"}) = 6/9 = 0.667$$

$$P(\text{Student} = \text{"Yes"} | \text{buys_computer} = \text{"No"}) = 1/5 = 0.2$$

$$P(\text{Credit Rating} = \text{"Fair"} | \text{buys_computer} = \text{"Yes"}) = 6/9 = 0.667$$

$$P(\text{Credit Rating} = \text{"Fair"} | \text{buys_computer} = \text{"No"}) = 2/5 = 0.4$$

$$X = (\text{Age} \leq 30, \text{Income} = \text{Medium}, \text{Student} = \text{Yes}, \text{Credit Rating} = \text{Fair})$$

$$P(X|C_i) : P(X|\text{buys_computer} = \text{"Yes"}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$

$$P(X|\text{buys_computer} = \text{"No"}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$$

$$P(X|C_i) * P(C_i) : P(X|\text{buys_computer} = \text{"Yes"}) * P(\text{buys_computer} = \text{"Yes"}) = 0.028$$

$$P(X|\text{buys_computer} = \text{"No"}) * P(\text{buys_computer} = \text{"No"}) = 0.007$$

Hence, X belongs to class "buys_computer="Yes"



Naive Bayes Classifier has the following advantages and disadvantages:



Advantages:

- Easy implementation
- Good results



Disadvantages:

- Less accuracy because of class conditional independence
- Practical dependencies among variables



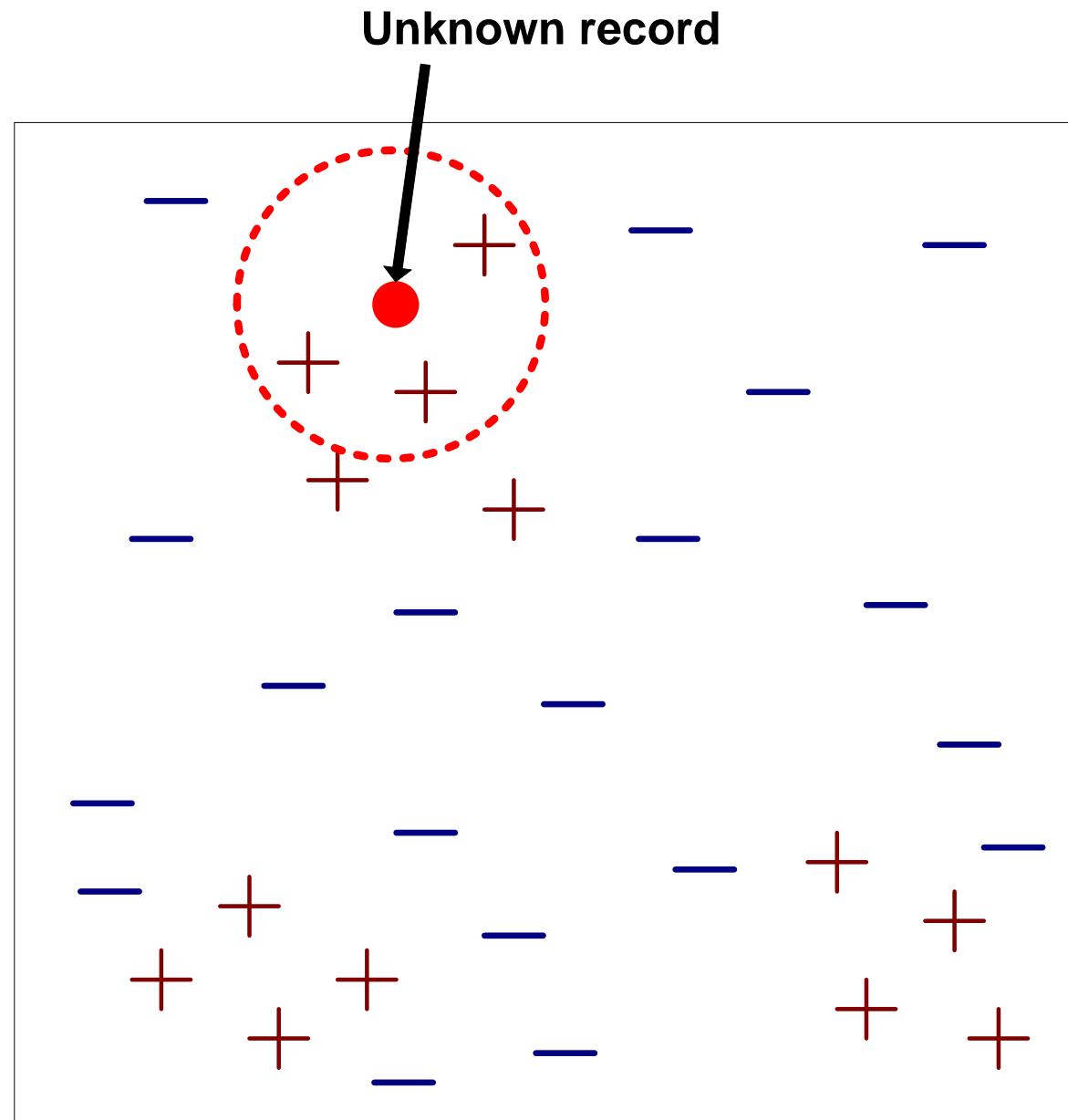
You can deal with dependencies using Bayesian Belief Networks.

Demo—Perform Classification Using the Naive Bayes Method

This demo will show the steps to do classification using the Naive Bayes method.

These:

- Use k “closest” points to perform classification
- Need three parameters:
 - Stored records’ set
 - Distance metric
 - The value of k (the number of nearest neighbors to retrieve)



The steps to classify an unknown record are:

1

Calculate the distance of the unknown record with other training records



2

Identify k-nearest neighbors

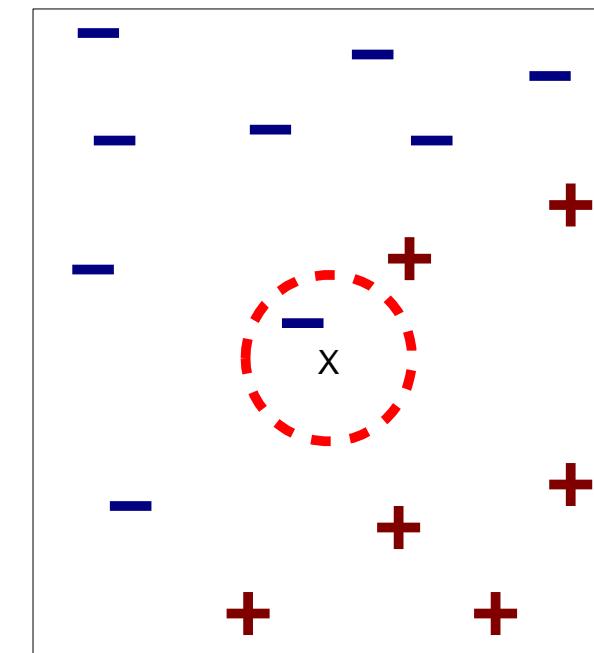


3

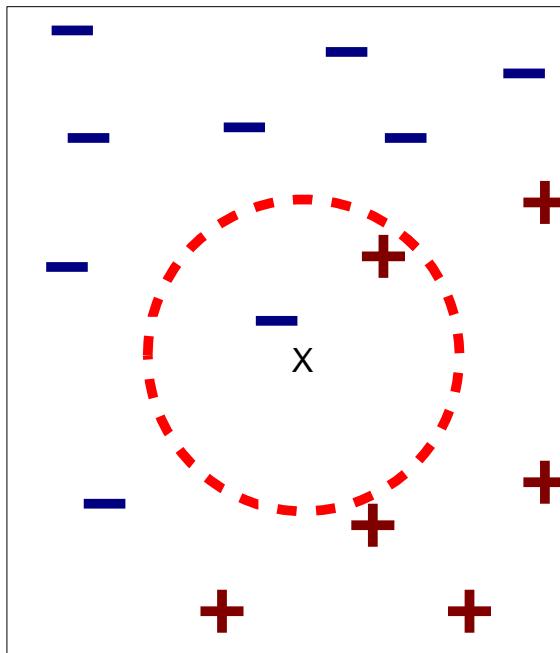
Use class labels of nearest neighbors to find the class label of an unknown record (for example, by taking the majority number of votes)

K-nearest neighbors are defined as the data points of a record x that are at the smallest distance from x . The graphical representation of these neighbors is given below:

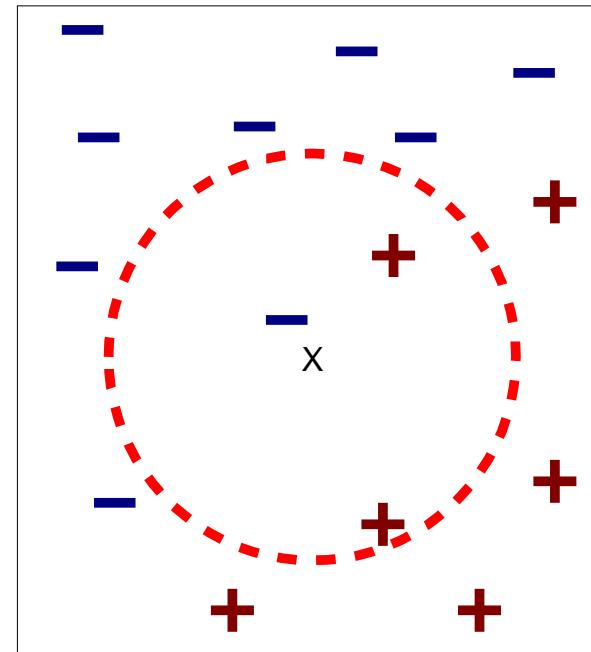
1-Nearest Neighbor



2-Nearest Neighbor



3-Nearest Neighbor



For the Nearest Neighbor Classifiers, the distance between two points is expressed in the form of Euclidean distance, which is calculated by:

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

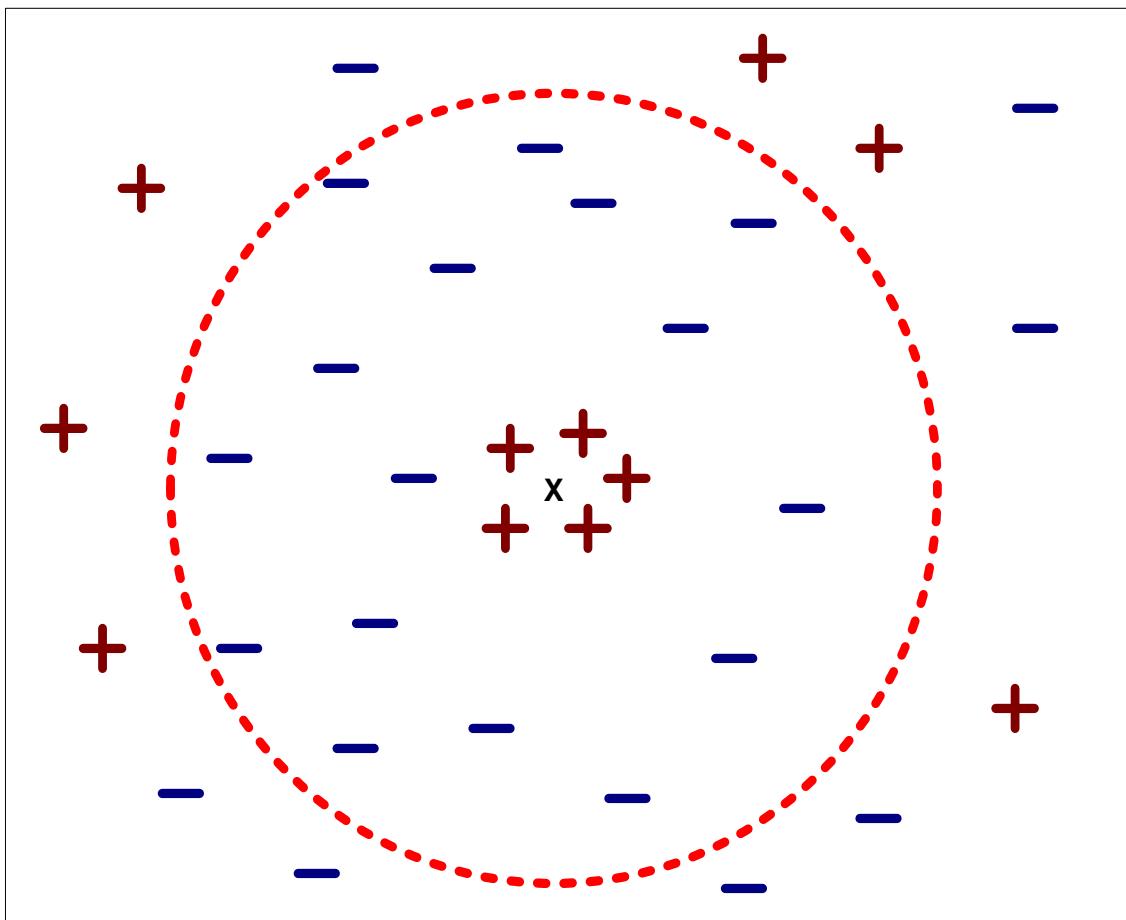


You can determine the class from the nearest neighbor list by:

- Taking the majority number of votes of class labels among the k-nearest neighbors
- Weighing the vote according to the distance
 - Weight factor, $w = 1/d^2$

When choosing the value of k, keep the following points in mind:

- If its value is too small, neighborhood is sensitive to noise points
- If its value is too large, neighborhood may include points from other classes



Attributes may have to be scaled for preventing distance measures to be dominated by any attribute.



Examples:

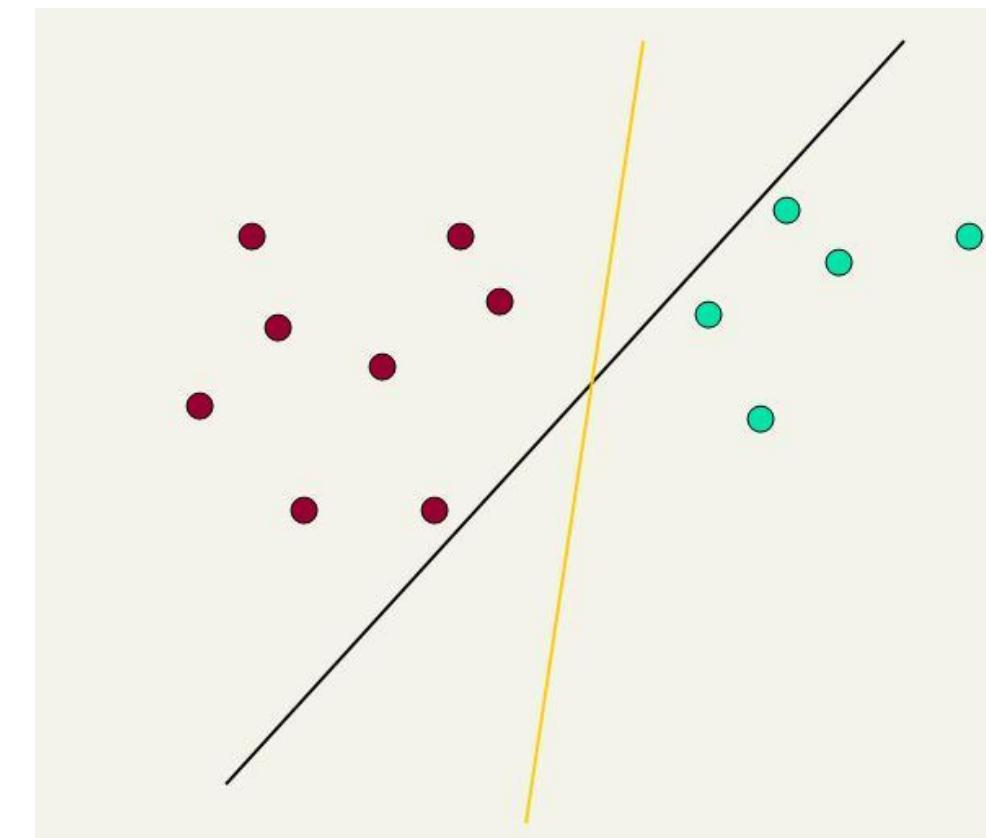
- A person's height may vary from 1.0m to 2.0m
- A person's weight may vary from 80lb to 280lb
- A person's income may vary from \$15K to \$2M

There can be lots of possible solutions for variables a, b, and c.

Methods, such as perception, can help find a separating hyperplane but do not provide an optimal solution.

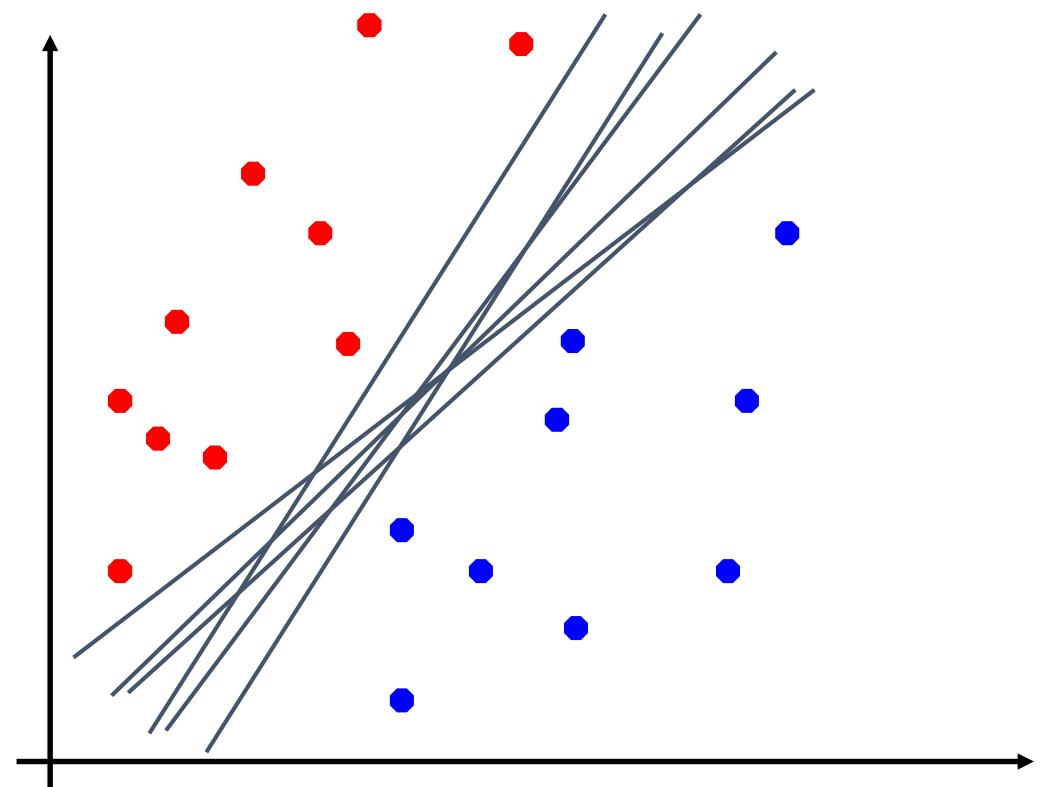
Support Vector Machine (SVM) provides an optimal solution, which:

- Maximizes the distance between the “difficult points” and the hyperplane
- Provides one intuition: There are not many uncertain classification decisions if there are no points near the decision surface



SVMs:

- Maximize the margin around the separating hyperplane
- Allow the decision function to be fully specified by a subset of training samples



Distance from the considered example to the separator is denoted by r , which is derived as follows:

Unit vector is $w/|w|$, so line is $rw/|w|$.

$$x' = x - yrw/|w|$$

x' satisfies $w^T x' + b = 0$.

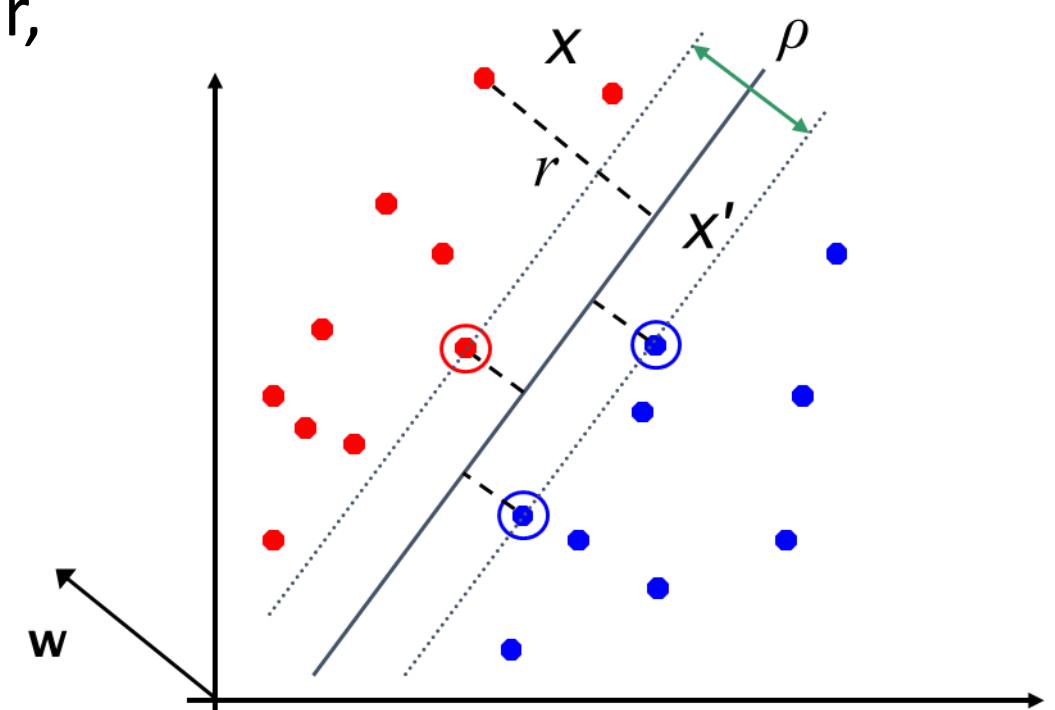
$$\text{So, } w^T(x - yrw/|w|) + b = 0$$

Recall that $|w| = \sqrt{w^Tw}$

$$\text{So, } w^T x - yr|w| + b = 0$$

Therefore:

$$r = y(w^T x + b)/|w|$$



Margin ρ is the width of separation between the support vectors of classes.

Assume that the entire data is, at least, at distance 1 from the hyperplane, then for a training set

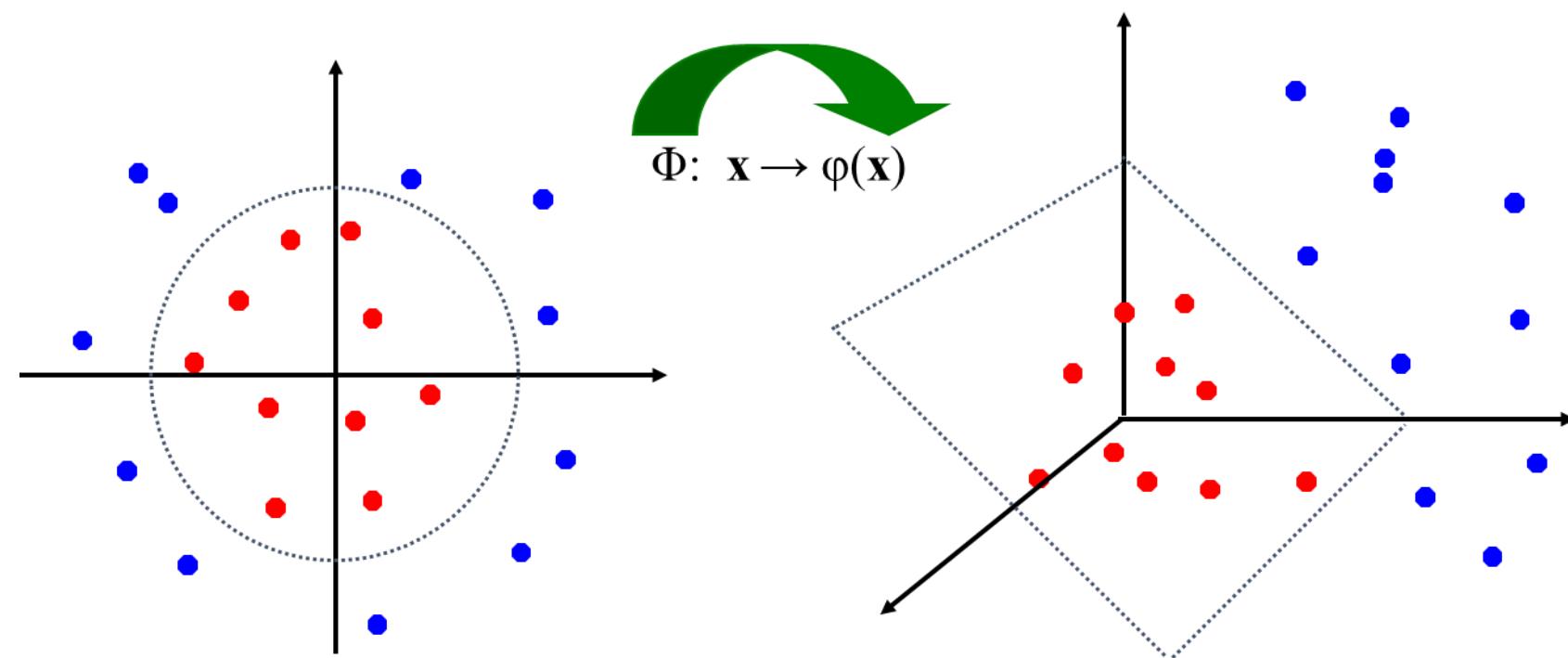
$\{(x_i, y_i)\}$:

- $w^T x_i + b \geq 1$ if $y_i = 1$
- $w^T x_i + b \leq -1$ if $y_i = -1$

For support vectors, the inequality becomes an equality. Therefore:

$$r = \frac{2}{\|w\|}$$

You can always map the original feature space to a higher-dimensional feature space, where the training set is separable.



This demo will show the steps to support a vector machine.



**QUIZ
1**

Which of the following is not a classification technique?

- a. Decision Tree
- b. SVM
- c. Bayesian Classifiers

Linear Regression



QUIZ
1

Which of the following is not a classification technique?

- a. Decision Tree
- b. SVM
- c. Bayesian Classifiers

Linear Regression



The correct answer is **d**.

Explanation: Linear Regression is not a classification technique.

QUIZ
2

State whether the following statement is “True” or “False.”

Discriminating between spam and ham e-mails is a classification task.

- a. True
- b. False



QUIZ
2

State whether the following statement is “True” or “False.”

Discriminating between spam and ham e-mails is a classification task.

- a. True
- b. False



The correct answer is **a.**

Explanation: Discriminating between spam and ham e-mails is a classification task.

QUIZ

3

Fill in the blank:

Predicting the cost of a land in sq.ft. is a(n) _____ problem.

- a. classification
 - b. prediction
 - c. cluster
- association



QUIZ

3

Fill in the blank:

Predicting the cost of a land in sq.ft. is a(n) _____ problem.

- a. classification
- b. prediction
- c. cluster

association



The correct answer is **b**.

Explanation: Predicting the cost of a land in sq.ft. is a prediction problem.

QUIZ

4

State whether the following statement is “True” or “False.”

A decision tree is a tree in which every node is either a leaf node or a decision node.

- a. True
- b. False



QUIZ

4

State whether the following statement is “True” or “False.”

A decision tree is a tree in which every node is either a leaf node or a decision node.

- a. True
- b. False



The correct answer is **a.**

Explanation: A decision tree is a tree in which every node is either a leaf node or a decision node.

Let us summarize the topics covered in this lesson:



- Classification is a technique to determine the extent to which a thing will or will not be a part of a category or type.
- For prediction, the classification process includes model construction and model usage as its two techniques.
- The two kinds of issues regarding classification and prediction of data are:
 - Data Preparation Issues
 - Evaluating Classification Methods Issues
- Different classification techniques include decision tree, Naive Bayes Classifier Model, Nearest Neighbor Classifiers, and SVMs.

This concludes “Classification.”

The next lesson is “Clustering.”

Data Science with R

Lesson 12—Clustering



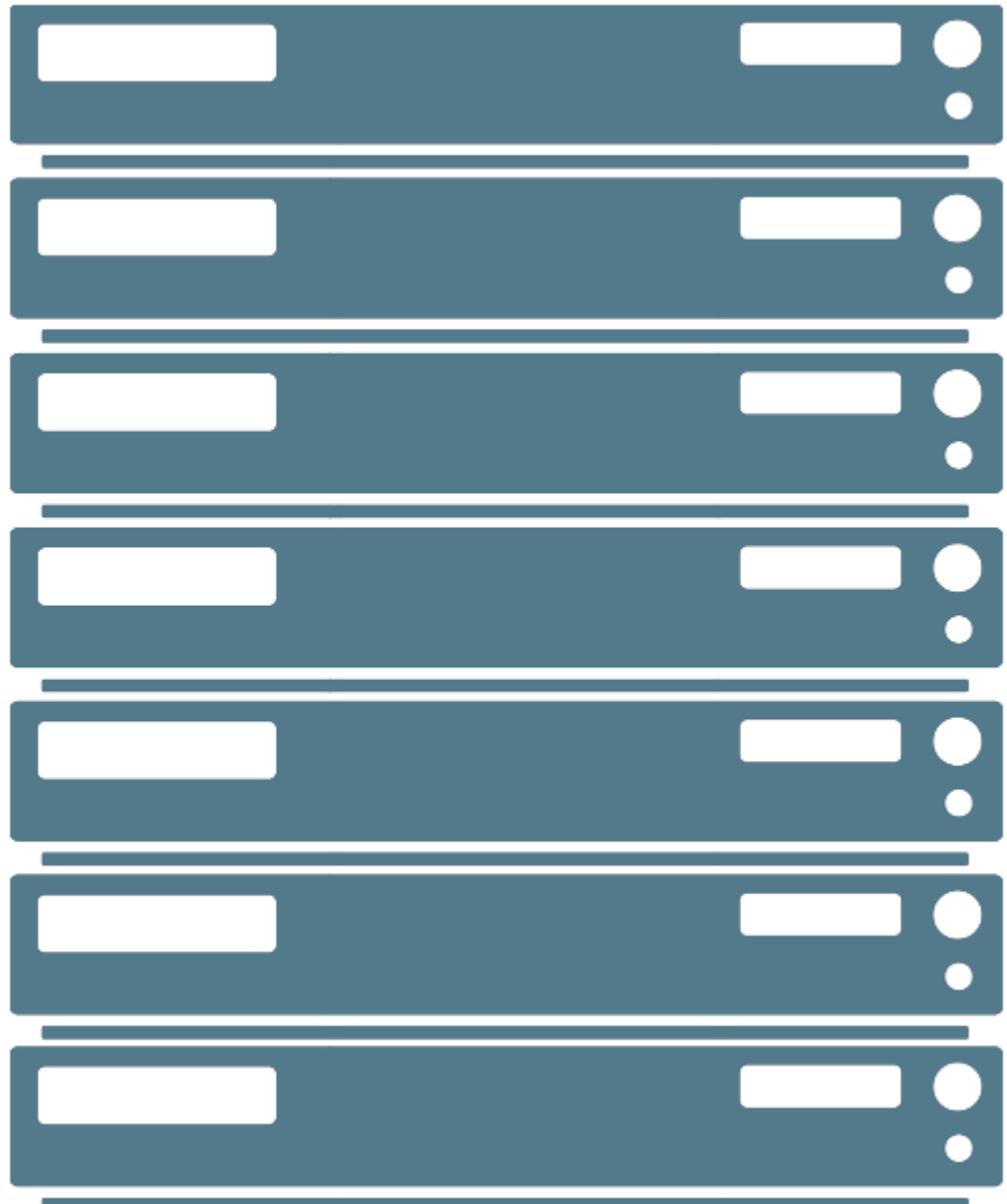
After completing
this lesson, you will
be able to:



- Explain clustering
- Describe clustering use cases
- Discuss clustering models

It is a type of unsupervised learning that:

- Forms clusters of similar objects automatically
- Segments the data so that each training example is assigned to a segment



Here's how clustering and classification are different from each other:

Feature	Clustering	Classification
Type of Learning Method	Unsupervised	Supervised
Purpose	Segment the data to assign each training example to a segment called a cluster	Predict the class to which a given training example belongs

These include:

- Grouping the content of a website or product in a retail business
- Segmenting customers or users in different groups on the basis of their metadata and behavioral characteristics
- Segmenting communities in ecology
- Finding clusters of similar genes
- Creating image segments to be used in image analysis applications



Some examples of clustering models are:

- K-means clustering
- Hierarchical Clustering
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Clustering

K-means tries to:

- Partition a set of data points into K distinct clusters
- Find clusters to minimize the sum of squared errors (WCSS) in every cluster

Calculation:

$$\sum_{i=1}^n \sum_{j=1}^n (x(j) - u(i))^2$$

It includes the following steps:

The k centroids are assigned to a point randomly

Every point in the dataset is assigned to a cluster

All centroids are updated by taking the mean of all the points in that cluster

It is as follows:

- *Create k points for starting centroids*
- *While any point is changing a cluster assignment*
 - *for every point in our dataset:*
 - *for every centroid*
 - *calculate the distance between the centroid and point*
 - *assign the point to the cluster with the lowest distance*
 - *for every cluster calculate the mean of the points in that cluster*
 - *assign the centroid to the mean*



To perform K-means clustering, use the **k-means** function in the R package stats. The two important features of this function are:

nstart

A parameter used for reducing the algorithm sensitivity to the random selection of the initial clusters/cluster means

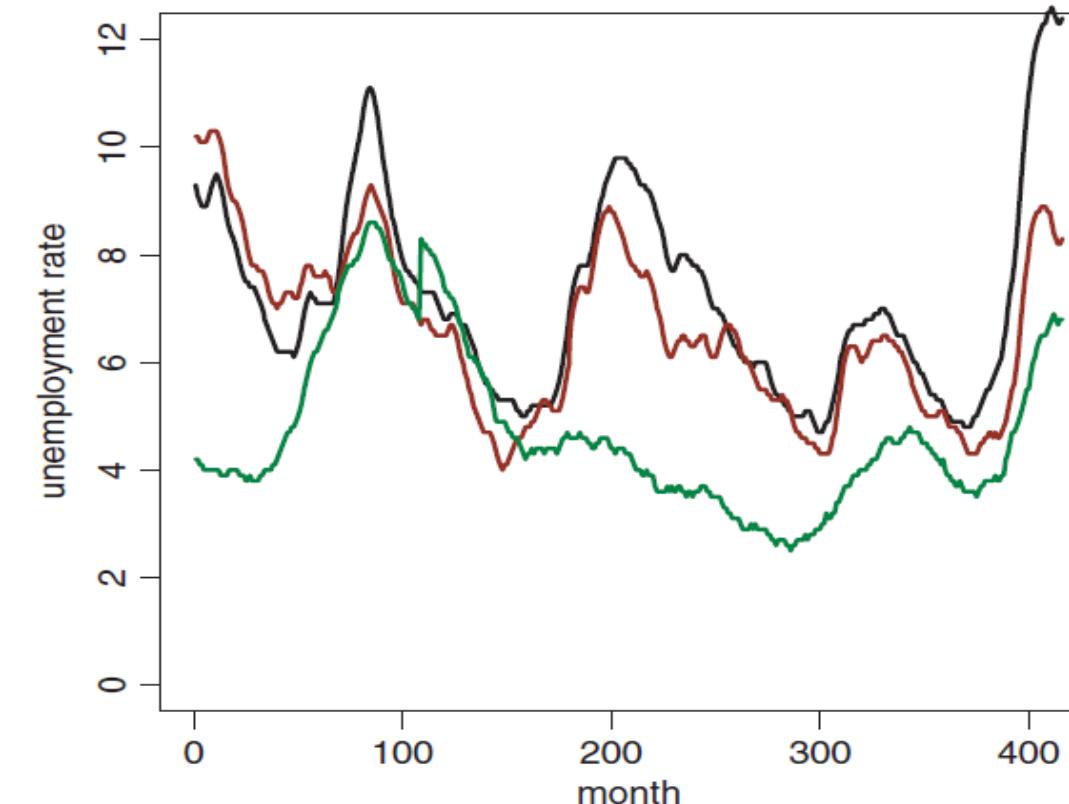
Cluster labels

Labels change from one run to the other

Consider the given case study:



The monthly and seasonal adjusted unemployment rates, from January 1976 to August 2010, for 50 U.S. states were captured. The graph below shows the time series plots of three states: Iowa (green), New York (red), and California (black).



Consider the given case study:



Problem

You need to cluster states group wise.

Assume:

- Each state is characterized by a feature vector with $p = 416$.
- New York and California form a cluster.

You need to calculate the 416 monthly averages with two observations each.

Consider the given case study:



Solution in R

```
## read the data; series are stored column-wise with labels in first ## row
raw <- read.csv("C:/DataMining/Data/unempstates.csv")
## transpose the data then we have 50 rows (states) and 416 columns (time periods)
rawt=matrix(nrow=50,ncol=416)
rawt=t(raw)
## k-means clustering in 416 dimensions
set.seed(1)
grpunemp2 <- kmeans(rawt, centers=2, nstart=10)
sort(grpunemp2$cluster)
grpunemp3 <- kmeans(rawt, centers=3, nstart=10)
sort(grpunemp3$cluster)
grpunemp4 <- kmeans(rawt, centers=4, nstart=10)
sort(grpunemp4$cluster)
grpunemp5 <- kmeans(rawt, centers=5, nstart=10)
sort(grpunemp5$cluster)
```

This demo will show the steps to do clustering using k-means.

It:

- Clusters n units/objects, each with p features, into smaller groups
- Creates a hierarchy of clusters as a dendrogram



Important Points about Dendograms:

- Units in the same cluster are joined by a horizontal line.
- The leaves at the bottom represent individual units.
- They are useful as they provide a visual representation of clusters.

They are of two types:

Agglomerative Algorithms

Method: Start at the individual leaves and successively merge clusters together

Approach: Bottom-up

Divisive Algorithms

Method: Start at the root and recursively split the clusters

Approach: Top-down

Two requirements are:

Distance Measure

For categorical variables, it can be defined from their number of matches and mismatches.

Linkage Creation

It determines the choice of clusters to be merged.



A distance measure between two objects with feature vectors $X_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ is non-negative and symmetric and satisfies $d(x_i, x_j) \leq d(x_i, x_k) + d(x_j, x_k)$.

In this process:

- An $n \times n$ distance matrix is considered, where the number in the i^{th} row and j^{th} column is the distance between the i^{th} and j^{th} units.
- The distance matrix is symmetric with zeros in the diagonal.
- Rows and columns are merged as clusters and the distances between them are updated.

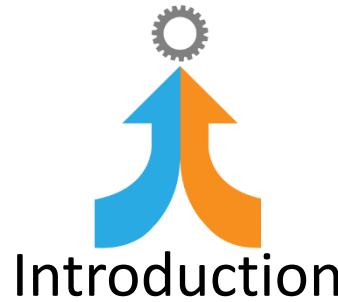
For the R package cluster

Use the “agnes” function

For the stats package

Use the “use the hclust” function

Consider the given case study:



The protein intakes in 25 European countries were captured from 9 food sources, as given in the table below:

Country	Red Meat	White Meat	Eggs	Milk	Fish	Cereals	Starch	Nuts	Fr&Veg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
United Kingdom	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

Consider the given case study:



Problem

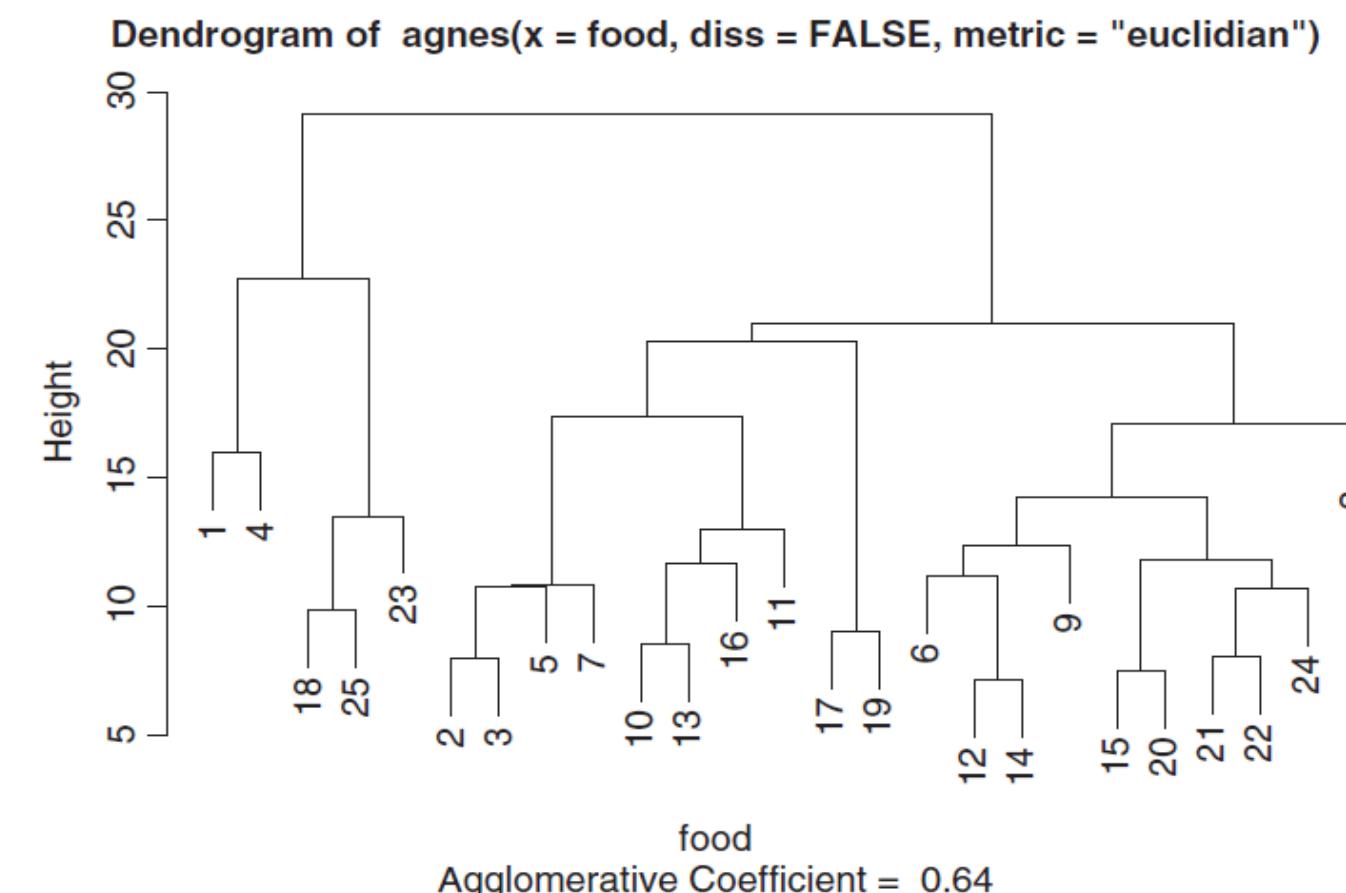
You need to determine whether the listed 25 countries can be separated into a smaller number of clusters.

Consider the given case study:



Solution in R

```
library(cluster)
food <- read.csv("C:/DataMining/Data/protein.csv")
foodagg=agnes(food,diss=FALSE,metric="euclidian")
plot(foodagg) ## dendrogram
```



This demo will show the steps to do hierarchical clustering.

This:

- Is inspired from the human natural clustering approach
- Allows clusters (therefore, classes) to be easily and readily identifiable
- Yields single points scattered around the datasets as outliers
- Requires two parameters at most: a density metric and the minimum size of a cluster

Database elements can be of two types: border points and core points. Other concepts of DBSCAN are:

Eps

A predetermined value that determines the distance measure of all points in the neighborhood of a point p

MinPts

A constant that determines the neighborhood size of a point q, which is directly density-reachable from p

A cluster D can be defined as follows if q belongs to D and:

q is density-reachable from p; p's neighborhood is greater than MinPts threshold

p belongs to D

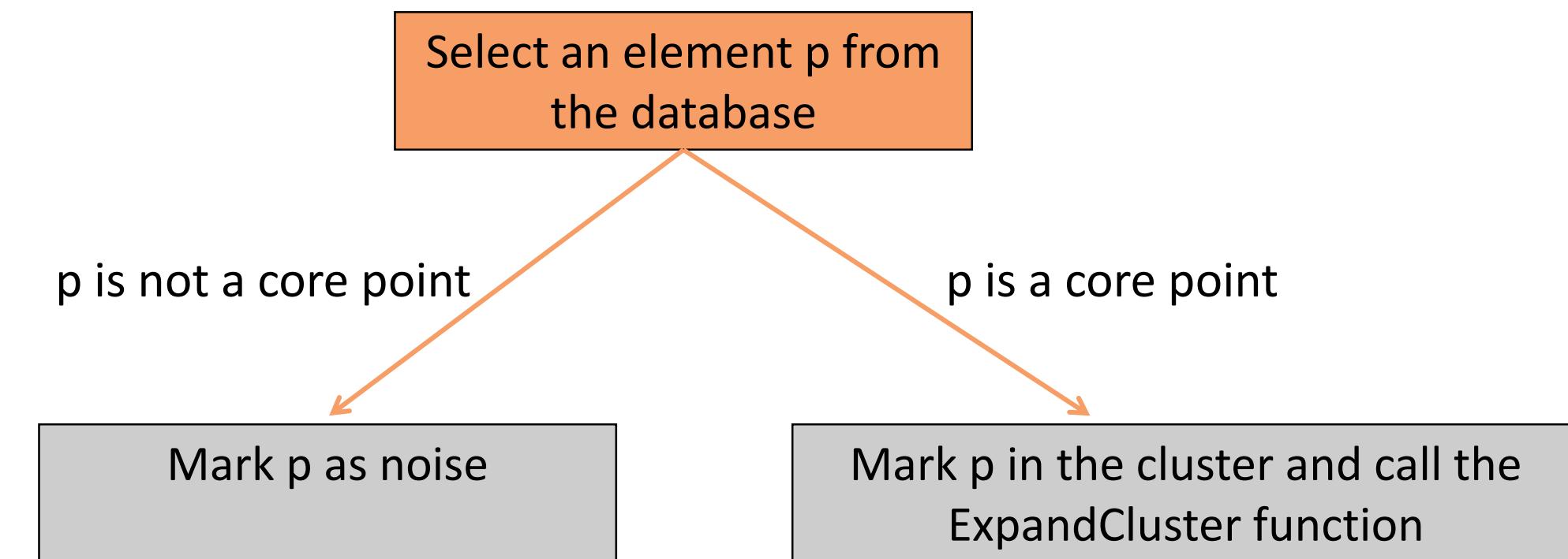
p belongs to D

q and p are directly reachable from a point t



A cluster D is created by a set of points that respects a certain degree of concentration, which is set by the MinPts and Eps constraints.

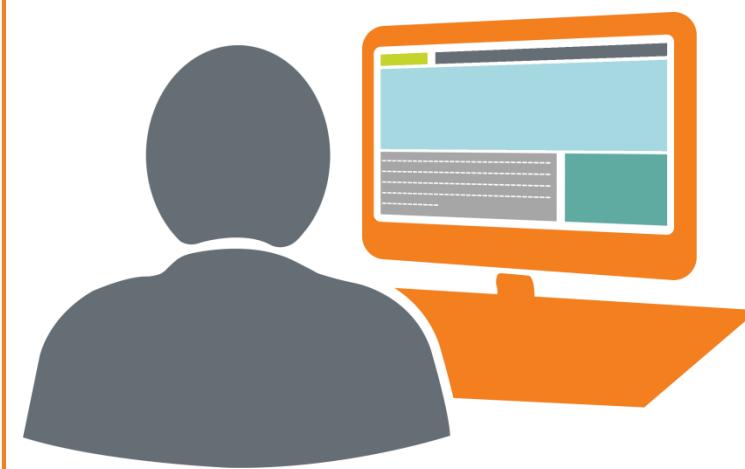
Here's the depiction of how this algorithm works:



The `ExpandCluster` is a recursive function that finds all the points that are density-reachable from p and are currently marked as unclassified or noise.

In R, DBSCAN is available through the fpc package, which provides the following advantages:

- High configurability
- A visualization interface
- Availability of dbscan procedure: `dbscan(data, eps, MinPts, scale, method, seeds, showplot, countmode)`



How to install fpc?

Run the following code:

install.packages("fpc", dependencies = TRUE)

library('fpc')

Consider the given case study:



An astronomical entity can be a planet, star of any kind or age, galaxy or other exotic entity that could be unidentified previously. Nowadays, astronomers can capture the electromagnetic intensity emitted from an entity at a given range of the spectrum in a grid, noise caused by the sensors, and diffuse emission from the atmosphere and space.

Consider the given case study:



Problem

You need to identify the various celestial entities existing in the different clusters.

Consider the given case study:



Solution in R

Step 1: Remove noise and diffuse emission by using a threshold

Step 2: Perform DBSCAN clustering on individual pixels by using the eps parameter to link a complete emission area together at the images, for each channel of the electromagnetic spectrum:

```
x <- matrix(scan("file.dat",1214), nrow=1214, ncol=2, byrow=TRUE);  
dbscan(x,5,showplot = 2);
```



QUIZ
1

Identify correct statements about the k-means clustering. *Select all that apply.*

- a. It attempts to partition a set of data points into k distinct clusters.
 - b. It allows to use k-means in the R package named “stats.”
 - c. It yields single points scattered around the datasets as outliers.
- It presents a hierarchy of clusters as a dendrogram.



QUIZ
1

Identify correct statements about the k-means clustering. *Select all that apply.*

- a. It attempts to partition a set of data points into k distinct clusters.
- b. It allows to use k-means in the R package named “stats”.
- c. It yields single points scattered around the datasets as outliers.



It presents a hierarchy of clusters as a dendrogram.

The correct answers are **a and b**.

Explanation: The k-means clustering technique attempts to partition a set of data points into k distinct clusters and allows to use k-means in the R package names as “stats”.

QUIZ
2

Which of the following statements are true of Hierarchical clustering? *Select all that apply.*

- a. It is inspired by the natural clustering approach.
 - b. Its algorithms are only of the agglomerative type.
 - c. It yields single points scattered around the datasets as outliers.
- It presents a hierarchy of clusters as a dendrogram.



QUIZ
2

Which of the following statements are true of Hierarchical clustering? *Select all that apply.*

- a. It is inspired by the natural clustering approach.
- b. Its algorithms are only of the agglomerative type.
- c. It yields single points scattered around the datasets as outliers.



It presents a hierarchy of clusters as a dendrogram.

The correct answer is **d**.

Explanation: The hierarchical clustering presents a hierarchy of clusters as a dendrogram.

QUIZ
3

Identify accurate statements about divisive hierarchical procedures. *Select all that apply.*

- a. As we move down the hierarchy, all units start in one cluster and splits are performed recursively.
- b. They represent a top-down approach.
- c. They represent a bottom-up approach.



They offer high-configurability.

QUIZ
3

Identify accurate statements about divisive hierarchical procedures. *Select all that apply.*

- a. As we move down the hierarchy, all units start in one cluster and splits are performed recursively.
- b. They represent a top-down approach.
- c. They represent a bottom-up approach.



They offer high-configurability.

The correct answers are **a and b.**

Explanation: These follow a top-down procedure and as we move down the hierarchy, all units start in one cluster and splits are performed recursively.

QUIZ
4

Which statements about the DBSCAN clustering technique are true? *Select all that apply.*

- a. It is available on R's fpc package.
- b. It offers high configurability.
- c. It yields single points scattered around the datasets as outliers.



A neighborhood of a point p is a set of all points that have a distance measure less than a predetermined value, called Eps.

QUIZ
4

Which statements about the DBSCAN clustering technique are true? *Select all that apply.*

- a. It is available on R's fpc package.
- b. It offers high configurability.
- c. It yields single points scattered around the datasets as outliers.



A neighborhood of a point p is a set of all points that have a distance measure less than a predetermined value, called Eps.

The correct answers are **a, b, c, and d**.

Explanation: All the given statements are true for the DBSCAN clustering technique.

Let us summarize the topics covered in this lesson:



- Clustering is a type of unsupervised learning that forms clusters of similar objects automatically.
- K-means clustering tries to partition a set of data points into K distinct clusters.
- Hierarchical clustering clusters n units/objects, each with p features, into smaller groups.
- DBSCAN clustering is inspired by the natural clustering approach.

This concludes “Clustering.”

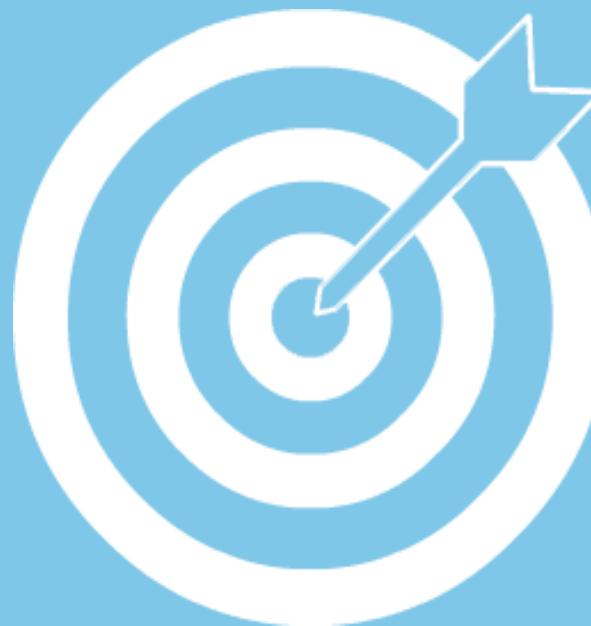
The next lesson is “Association.”

Data Science with R

Lesson 13—Association



After completing
this lesson, you will
be able to:



- Explain association rule mining
- Describe the parameters of interesting relationships
- Discuss the strength measures of association rules
- Explain the Apriori algorithm

This is a classical Data Mining technique that:

- Finds out interesting patterns in a dataset
- Assumes all data elements as categorical
- Is not suitable for numeric data



Brute-force solutions cannot solve the problem of finding different combinations of items in less time and computing power.

Some examples are:



Market Basket Data Analysis



Purchase Data Analysis



Website Traffic Analysis

Interesting relationships have two parameters:

- **Frequent item sets:** Collection of items occurring together frequently
- **Association rules:** Indicators of a strong relationship between two items

Example:



In the “Items” table below, {wine, diapers, soy milk} is the frequent item set and diapers → Wine is an association rule:

Transaction number	Items
0	soy milk, lettuce
1	lettuce, diapers, wine, chard
2	soy milk, diapers, wine, orange juice
3	lettuce, soy milk, diapers, wine
4	lettuce, soy milk, diapers, orange juice

An association rule is a pattern that states when X occurs, Y occurs with a certain probability. A transaction t contains X, a set of items (item set) in I, if X is a subset of t.

An association rule is an implication of the form:

$$X \rightarrow Y$$

Where, $X, Y \subset I$, and $X \cap Y = \emptyset$

The measures of the strength of association rules are explained below:

Support

For an item set, it is the percentage of the dataset that contains this item set.

The rule holds with support sup in T, if sup% of transactions contain $X \cup Y$.

$$sup = Pr(X \cup Y).$$

Example: In the “Items” table, the support of {soy milk} is 4/5 and of {soymilk, diapers} is 3/5.

Confidence

The confidence for the rule $\{diapers\} \rightarrow \{wine\}$ is defined as $support(\{diapers, wine\})/support(\{diapers\})$.

The rule holds in T with confidence conf if conf% of transactions that contain X also contain Y.

$$conf = Pr(Y | X)$$

Example: In the “Items” table, the confidence for $diapers \rightarrow wine$ is $3/5/4/5 = 3/4 = 0.75$.

While support and confidence can help you quantify the success of association analysis, for thousands of sale items, the process of finding them can be really slow.

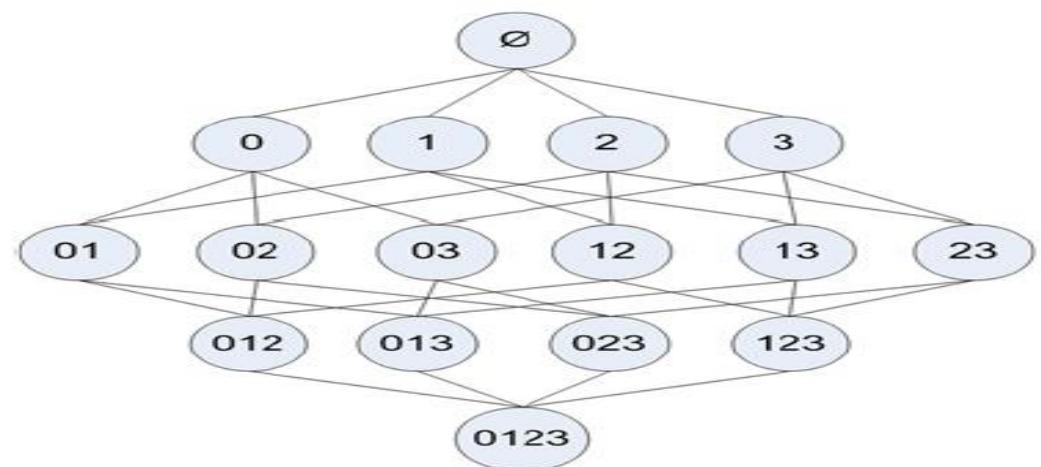
In such cases, you can use algorithms such as Apriori.



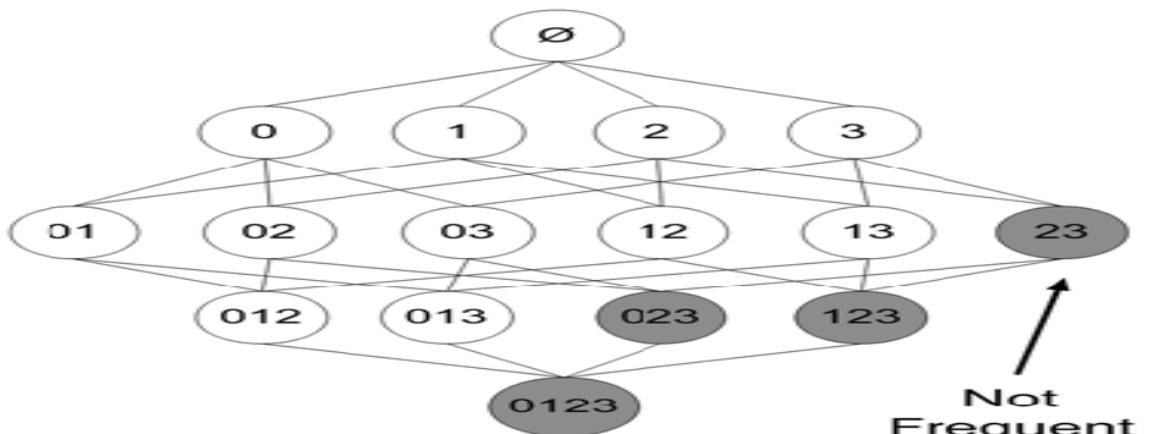
All possible item sets from the set {1, 2, 3}

This algorithm:

- Helps reduce the number of possible interesting item sets
- Assumes that if an item set is frequent, all of its subsets are also frequent



With infrequent item sets highlighted



To understand its application, consider the below “Shopping Baskets” items set, which ignores some important parameters, such as quantities of items and price paid:

- t1: Beef, Chicken, Milk
- t2: Beef, Cheese
- t3: Cheese, Boots
- t4: Beef, Chicken, Cheese
- t5: Beef, Chicken, Clothes, Cheese, Milk
- t6: Chicken, Clothes, Milk
- t7: Chicken, Milk, Clothes

It includes two steps:

Mine all frequent item sets

Generate rules from frequent item sets



Assume:

- $\text{minsup} = 30\%$
- $\text{minconf} = 80\%$

An example frequent item set:

{Chicken, Clothes, Milk} [sup = 3/7]

Association rules from the item set:

Clothes \rightarrow Milk, Chicken [sup = 3/7, conf = 3/3]

...

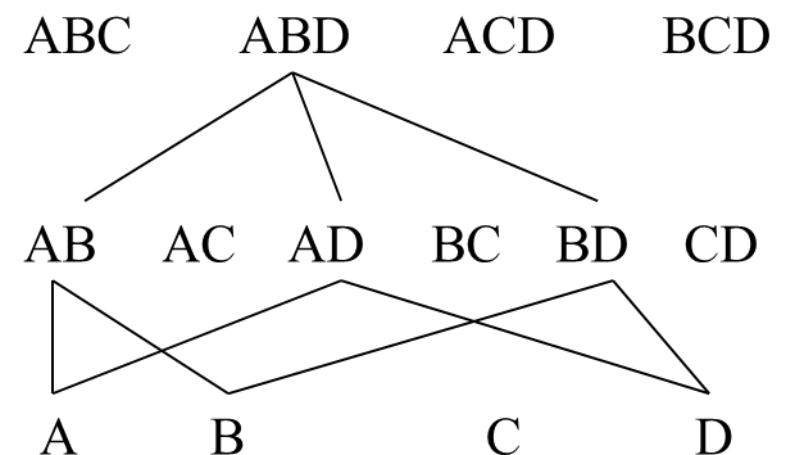
...

Clothes, Chicken \rightarrow Milk, [sup = 3/7, conf = 3/3]

A frequent item set is:

- The one with $\text{sup} \geq \text{minsup}$
- Any subset of a frequent item set

Visual Depiction



Also called level-wise search, it includes the following steps:

Find all 1-item frequent item sets; then all 2-item frequent item sets, and so on

In each iteration k , consider item sets that contain some $k-1$ frequent item sets

Find frequent item sets of size 1: F_1

!

With $k = 2$, C_k = item sets of size k that could be frequent, given F_{k-1} , and F_k = item sets that are actually frequent, $F_k \subseteq C_k$.

Consider the below dataset T with minsup = 0.5:

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

itemset:count

1. scan T → $C_1: \{1\}:2, \{2\}:3, \{3\}:3, \{4\}:1, \{5\}:3$

→ $F_1: \{1\}:2, \{2\}:3, \{3\}:3, \{5\}:3$

→ $C_2: \{1,2\}, \{1,3\}, \{1,5\}, \{2,3\}, \{2,5\}, \{3,5\}$

2. scan T → $C_2: \{1,2\}:1, \{1,3\}:2, \{1,5\}:1, \{2,3\}:2, \{2,5\}:3, \{3,5\}:2$

→ $F_2: \{1,3\}:2, \{2,3\}:2, \{2,5\}:3, \{3,5\}:2$

→ $C_3: \{2, 3, 5\}$

3. scan T → $C_3: \{2, 3, 5\}:2 \rightarrow F_3: \{2, 3, 5\}$

The items in I are sorted in lexicographic order (total order).

- In each item set, it is used throughout the algorithm.
- $\{w[1], w[2], \dots, w[k]\}$ represents a k-item set, where w consists of items $w[1], w[2], \dots, w[k]$, where $w[1] < w[2] < \dots < w[k]$.

The algorithm for ordering items is:

```
C1 ← init-pass(T);  
F1 ← {f | f ∈ C1, f.count/n ≥ minsup}; // n: no. of transactions in T  
for (k = 2; Fk-1 ≠ ∅; k++) do  
    Ck ← candidate-gen(Fk-1);  
    for each transaction t ∈ T do  
        for each candidate c ∈ Ck do  
            if c is contained in t then  
                c.count++;  
            end  
        end  
    end  
    Fk ← {c ∈ Ck | c.count/n ≥ minsup}  
end  
return F ←  $\bigcup_k F_k$ ;
```

The candidate-gen function takes F_{k-1} and returns candidates as the superset of the set of all frequent k item sets. It includes two steps:

1

Join: Generate all possible candidate item sets C_k of length k



2

Prune: Remove the candidates in C_k that cannot be frequent

The algorithm for candidate generation is:

```
Function candidate-gen( $F_{k-1}$ )
     $C_k \leftarrow \emptyset;$ 
    forall  $f_1, f_2 \in F_{k-1}$ 
        with  $f_1 = \{i_1, \dots, i_{k-2}, i_{k-1}\}$ 
        and  $f_2 = \{i_1, \dots, i_{k-2}, i'_{k-1}\}$ 
        and  $i_{k-1} < i'_{k-1}$  do
             $c \leftarrow \{i_1, \dots, i_{k-1}, i'_{k-1}\};$  // join  $f_1$  and  $f_2$ 
             $C_k \leftarrow C_k \cup \{c\};$ 
            for each  $(k-1)$ -subset  $s$  of  $c$  do
                if ( $s \notin F_{k-1}$ ) then
                    delete  $c$  from  $C_k;$  // prune
                end
            end
        return  $C_k;$ 
```

Assume $F_3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{1, 3, 5\}, \{2, 3, 4\}\}$, then:

After join

$$C_4 = \{\{1, 2, 3, 4\}, \{1, 3, 4, 5\}\}$$

After prune

$$C_4 = \{\{1, 2, 3, 4\}\}$$

For each frequent item set X and proper nonempty subset A of X , assume $B = X - A$.

$A \rightarrow B$ is an association rule if:

$$\text{Confidence}(A \rightarrow B) \geq \text{minconf}$$

$$\text{support}(A \rightarrow B) = \text{support}(A \cup B) = \text{support}(X)$$

$$\text{confidence}(A \rightarrow B) = \text{support}(A \cup B) / \text{support}(A)$$

Assume $\{2,3,4\}$ is frequent with $\text{sup} = 50\%$ and proper nonempty subsets: $\{2,3\}$, $\{2,4\}$, $\{3,4\}$, $\{2\}$, $\{3\}$, $\{4\}$, with $\text{sup} = 50\%$, 50% , 75% , 75% , 75% , respectively.

Association rules:

$2,3 \rightarrow 4$, confidence = 100%

$2,4 \rightarrow 3$, confidence = 100%

$3,4 \rightarrow 2$, confidence = 67%

$2 \rightarrow 3,4$, confidence = 67%

$3 \rightarrow 2,4$, confidence = 67%

$4 \rightarrow 2,3$, confidence = 67%

Support of all rules = 50%

This demo will show the steps to do association using the Apriori algorithm.

This demo will show the steps to do visualization on associated rules.

Some problems related with association mining are:

Single minsup

It assumes that all data items have similar frequencies and/or are of the same nature.

False Items

Some items appear very frequently, whereas others appear rarely.

Items Frequencies Variation

If minsup is high, rules with rare items are not found; if minsup is set low, it may cause combinatorial explosion.



**QUIZ
1**

Association rules are interesting:

- a. if they satisfy both minimum and maximum iterations.
- b. if they satisfy both minimum support and minimum confidence thresholds.
- c. if they satisfy both association correlations.

if they satisfy Apriori constants.



QUIZ
1

Association rules are interesting:

- a. if they satisfy both minimum and maximum iterations.
- b. if they satisfy both minimum support and minimum confidence thresholds.
- c. if they satisfy both association correlations.

if they satisfy Apriori constants.



The correct answer is **b**.

Explanation: Association rules are interesting if they satisfy both minimum support and minimum confidence thresholds.

QUIZ
2

What is the formula to calculate support?

- a. $\Pr(X | Y)$
- b. $\Pr(X \cup Y)$
- c. $\Pr(X * Y)$

$\Pr(X / Y)$



QUIZ
2

What is the formula to calculate support?

- a. $\Pr(X | Y)$
- b. $\Pr(X \cup Y)$
- c. $\Pr(X * Y)$

$\Pr(X / Y)$



The correct answer is **b.**

Explanation: The formula to calculate Support is $\Pr(X \cup Y)$.

**QUIZ
3**

Which of the following algorithms can be used to solve the problem of support and confidence?

- a. Candidate generation
- b. Classification
- c. Apriori



Item set

QUIZ
3

Which of the following algorithms can be used to solve the problem of support and confidence?

- a. Candidate generation
- b. Classification
- c. Apriori



Item set

The correct answers are **c.**

Explanation: The Apriori algorithm can be used to solve the problem of support and confidence.

QUIZ
4

Which of the following conditions is true for mining frequent item sets?

- a. $\text{sup} \leq \text{minsup}$
- b. $\text{sup} < \text{minsup}$
- c. $\text{sup} = \text{minsup}$

$\text{sup} \geq \text{minsup}$



QUIZ
4

Which of the following conditions is true for mining frequent item sets?

- a. $\text{sup} \leq \text{minsup}$
- b. $\text{sup} < \text{minsup}$
- c. $\text{sup} = \text{minsup}$

$\text{sup} \geq \text{minsup}$



The correct answer is **d**.

Explanation: $\text{sup} \geq \text{minsup}$ is true for mining frequent item sets.

Let us summarize the topics covered in this lesson:



- Association rule mining finds out interesting patterns in a dataset.
- The interesting relationships can have two parameters: frequent item sets and association rules.
- An association rule is a pattern that states when X occurs, Y occurs with a certain probability.
- The measures of the strength of association rules are support and confidence.
- While support and confidence can help quantify the success of association analysis, for thousands of sale items, the process can be really slow, which is solved by algorithms, such as Apriori.
- The Apriori algorithm includes two steps: mining all frequent item sets and generating rules from frequent item sets.

This concludes “Association.”

This is the last lesson of the course



Thank You