

R for Data Science

Pacifique Nizeyimana, Ph.D

Adventist University of Central Africa

pacifique.nizeyimana@auca.ac.rw

September, 14, 2025

Reference book: R Programming for Data Science
Author: Roger D. Peng

- 1 Review on Descriptive analysis
 - Describing Data with Graphs
 - Description of Data using Numerical measures
- 2 Introduction of R
- 3 Data management
- 4 Data Visualization
- 5 Control Structures
- 6 Functions
- 7 Simulation in R
- 8 Simple linear regression model

Some terms

- **Statistics** is a branch of mathematics that has applications in almost every facet of our daily life.
- Two branches of statistics
 - Descriptive Statistics
 - Inferential Statistics
- **Modelling**: Modelling is the process of creating mathematical representations of real-world systems, phenomena, or processes.
- **Computation**: Computation is the process of using mathematical and logical operations to solve problems, typically performed by a computer or algorithm,
- **Calculus** : Calculus is a branch of mathematics that focus on continuous change (Differentiation and integration)

- **Descriptive statistics** consists of procedures used to summarize and describe the important characteristics of a set of measurements.
- **Inferential statistics** consists of procedures used to make inferences about population characteristics from information contained in a sample drawn from this population.

Objectives

- What is a variable
- Types of variables
- Describing data with graphs

Variable and Data

- A variable is a characteristic that changes or varies over time and/or for different individuals or objects under consideration
- **Examples:** Hair color, white blood cell count, time to failure of a computer component

- **An experimental unit** is the individual or object on which a variable is measured
- A measurement results when a variable is actually measured on an experimental unit.
- A set of measurements, called **data**, can be either a sample or a population.

Example

- **Variable:** Major
- **Experimental unit:** Student
- **Typical Measurements:** Statistics, Economic, etc.

Example

- **Variable:** Time until a light bulb burns out
- **Experimental unit:** Light bulb
- **Typical Measurements:** Light bulb

How many Variable have you measured

- **Univariate data:** One variable is measured on a single experimental unit.
- **Bivariate data:** Two variables are measured on a single experimental unit.
- **Multivariate data:** Multivariate data

TYPES OF VARIABLES

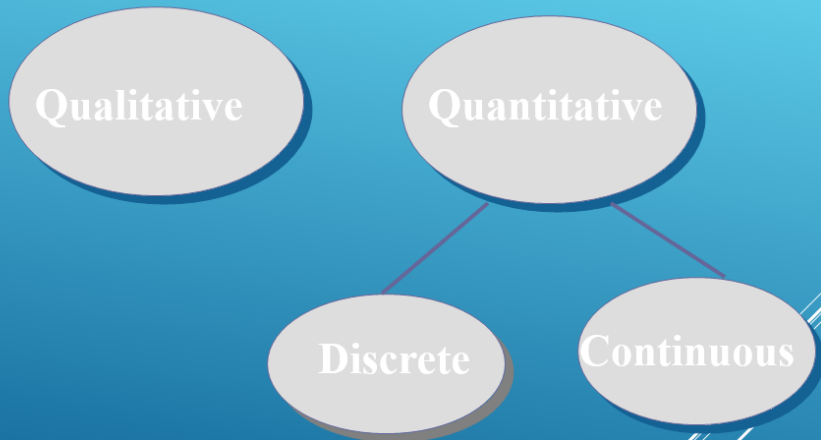


Figure: types of variables

Types of variables

- **Qualitative variables** measure a quality or characteristic on each experimental unit.
- **Examples:** Hair color (black, brown, blonde...) Make of car (Dodge, Honda, Ford...) Gender (male, female) State of birth (California, Arizona,...)

Types of variables

- **Quantitative variables** measure a numerical quantity on each experimental unit.
- **Discrete** if it can assume only a finite or countable number of values. Continuous if it can assume the infinitely many values corresponding to the points on a line interval.
 - Examples: For a particular day, the number of cars entering a college campus is measured.
- **Continuous** if it can assume the infinitely many values corresponding to the points on a line interval.
 - Examples: Height of students in Group 3

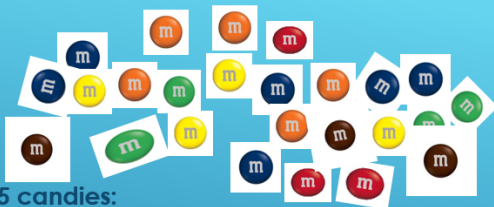
Graphs for qualitative variables

Use a data distribution to describe what values of the variable have been measured.

How often each value has occurred

- “How often” can be measured 3 ways:
- Frequency Relative frequency = $\text{Frequency}/n$ Percent = $100 \times \text{Relative frequency}$

EXAMPLE



A bag of M&Ms contains 25 candies:

Raw Data:

Statistical Table:


























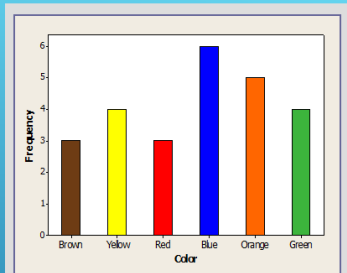
Color	Tally	Frequency	Relative Frequency	Percent
Red	  	3	$3/25 = .12$	12%
Blue	     	6	$6/25 = .24$	24%
Green	   	4	$4/25 = .16$	16%
Orange	    	5	$5/25 = .20$	20%
Brown	  	3	$3/25 = .12$	12%
Yellow	   	4	$4/25 = .16$	16%

Figure: frequency table

GRAPHS



Bar Chart

Pie Chart

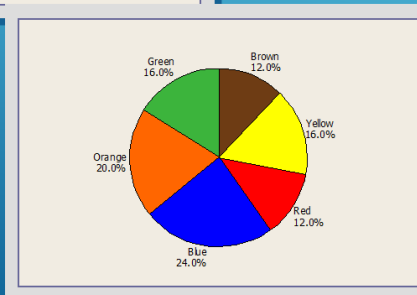


Figure: Graphs for categorical Var.

Graphing quantitative variable

A single quantitative variable measured for different population segments or for different categories of classification can be graphed using a **pie** or **bar chart**.

Example: A Big Mac hamburger costs \$4.90 in Switzerland, \$2.90 in the U.S. and \$1.86 in South Africa.

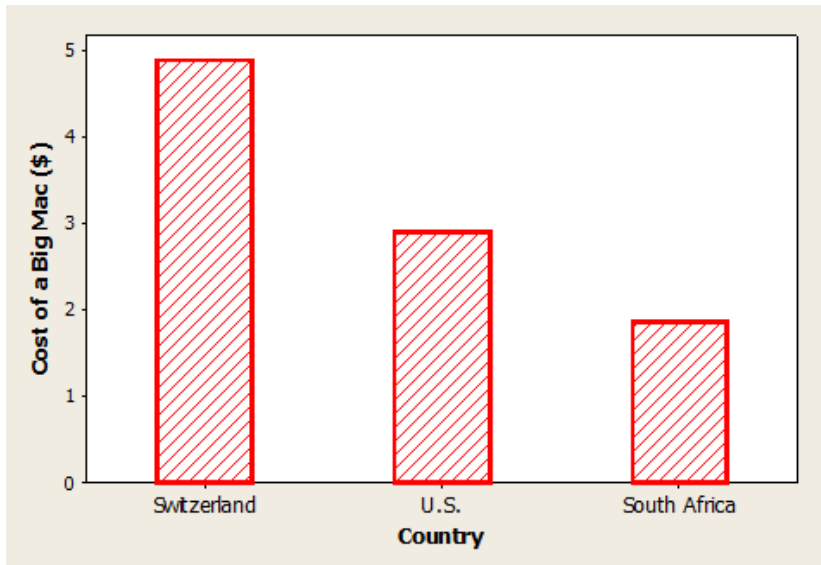


Figure: Graphs for categorical Var.

Graphing quantitative variable

A single quantitative variable measured over time is called a time series. It can be graphed using a line or bar chart.

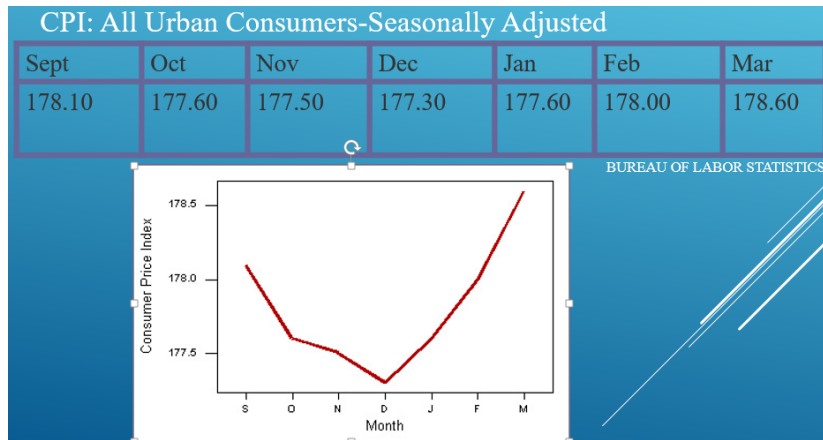


Figure: Line

The simplest graph for quantitative data Plots the measurements as points on a horizontal axis, stacking the points that duplicate existing points.
Example: The set 4, 5, 5, 7, 6

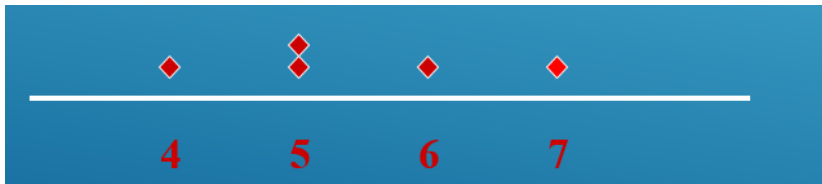


Figure: Line

Steam and Leaf

Uses the actual numerical values of each data point.

- Divide each measurement into two parts: the stem and the leaf.
- List the stems in a column, with a vertical line to their right.
- For each measurement, record the leaf portion in the same row as its matching stem.
- Order the leaves from lowest to highest in each stem.
- Provide a key to your coding.

EXAMPLE

The prices (\$) of 18 brands of walking shoes:

90	70	70	70	75	70	65	68	60
74	70	95	75	70	68	65	40	65

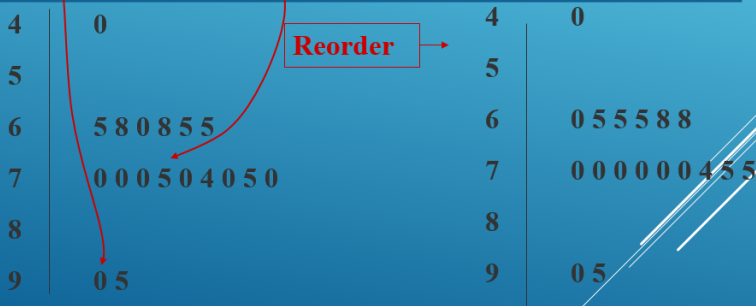


Figure: Line

Relative frequency, Histogram

A relative frequency histogram for a quantitative data set is a bar graph in which the height of the bar shows “how often” (measured as a proportion or relative frequency) measurements fall in a particular class or subinterval.

- Divide the range of the data into 5-12 subintervals of equal length.
- Calculate the approximate width of the subinterval as $\text{Range}/\text{number of subintervals}$.
- Use the method of left inclusion, including the left endpoint, but not the right in your tally.
- Create a statistical table including the subintervals, their frequencies and relative frequencies

Relative frequency, Histogram

Draw the relative frequency histogram, plotting the subintervals on the horizontal axis and the relative frequencies on the vertical axis.

- The height of the bar represents
- The proportion of measurements falling in that class or sub-interval.
Calculate the approximate width of the sub-interval as $\text{Range}/\text{number of sub-intervals}$.
- The probability that a single measurement, drawn at random from the set, will belong to that class or sub-interval

EXAMPLE

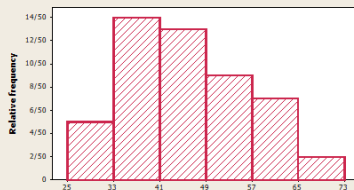
- The ages of 50 tenured faculty at a state university.

▶	34	48	70	63	52	52	35	50	37	43	53	43	52	44
▶	42	31	36	48	43	26	58	62	49	34	48	53	39	45
▶	34	59	34	66	40	59	36	41	35	36	62	34	38	28
▶	43	50	30	43	32	44	58	53						

- We choose to use **6** intervals.
- Minimum class width = $(70 - 26)/6 = 7.33$
- Convenient class width = **8**
- Use **6** classes of length **8**, starting at **25**.

Histogram

Age	Tally	Frequency	Relative Frequency	Percent
25 to < 33	1111	5	$5/50 = .10$	10%
33 to < 41	1111 1111 1111	14	$14/50 = .28$	28%
41 to < 49	1111 1111 111	13	$13/50 = .26$	26%
49 to < 57	1111 1111	9	$9/50 = .18$	18%
57 to < 65	1111 11	7	$7/50 = .14$	14%
65 to < 73	11	2	$2/50 = .04$	4%



Box plot

Two methods to describe dataset

- Measure of center
- Measure of the spread

Measure of Center

- Mean
- Mode
- Median

Measure of spread

- Range
- Variance
- Standard of deviation

Other concept on Spread

- Tchebysheff's Theorem
- Empirical rule

Describing Bivariate data

- Correlation
- Scatter plot

Data science project

- Import
- Tidy (Storing in a consistent form): Column is variable and row is an observation.
- Understanding data (Transform, visualize, Model)
- Communicate

- Installation of R
- Installation of R studio
- Calculator in R
- Creating Dataset
- Data importation
- Data description in R

Install R and R studio

- For R: go to CRAN site "Comprehensive R Archive Network" or [https:// cloud.r-project.org](https://cloud.r-project.org)
- For R studio: <http://www.rstudio.com/download>.

Data structures

- Vector: One dimension array (numerical data)
- Matrix: Two dimension array
- Array: Like matrix but can have more than two dimension
- Factor: Represent categorical variable
- Data frame: More general than a matrix bcz it can accommodate characters and are used to store tabular data in R.
- List: an ordered collection of objects (components) and contain elements of different classes

Calculating Memory Requirements for R Objects

- How much memory is being used up by all of the data objects residing in your workspace
- Suppose I have a data frame with 1,500,000 rows and 120 columns, all of which are numeric data
- Roughly, how much memory is required to store this data frame?
- Well, on most modern computers double precision floating point numbers³ are stored using 64 bits of memory, or 8 bytes
- If the memory is not enough, freeze up your computer (or at least your R session) like kill the R process, in the best case scenario, or reboot your computer, in the worst case.

- Give light and the people will find their own way by Ella baker
- Sometimes we make the process more complicated than we need to. We will never make a journey of thousands miles by fretting about how long it will take or how hard it will be. we make the journey by each day step by step and repeating it again and again until we reach our destination. by Joseph B. W.

Data management

- Use of \$, [, or [[]] for subset extraction
- Sorting data
- Subset (selecting variables, dropping variables, selecting observation)
- Use of dplyr package

dp_{lyr} functions

- `select()`
- `filter()`
- `arrange()`
- `rename()`
- `mutate()`
- `group_` `by`
- `%>%`

- Basic plots
- use of ggplot2 package

Key functions in ggplot2

- `geom_bar()`
- `geom_boxplot()`
- `geom_density()`
- `geom_histogram()`
- `geom_hline()`
- `geom_jitter()`
- `geom_line()`

key functions in ggplot2

- `geom_point()`
- `geom_rug()`
- `geom_smooth()`
- `geom_smooth()`
- `geom_text()`
- `geom_violin()`
- `geom_vline()`
-

No matter what you are going through, there is a light at the end of the tunnel and it may seem hard to get to it but you can do it and just keep working towards it and you will find the positive side of things. by **Demi Lovato**

This allows the use to control the flow of execution of a series of R expressions. Put some logic in your R code

- **if and else:** Testing a condition and acting on it
- **for:** execute a loop while a fixed number of times
- **while:** execute a loop while a condition is true
- **repeat:** execute an infinite loop(must break out of it to stop)
- **break:** Break the execution of a loop
- **Next:** skip an iteration of a loop

if(condition) do something Continue with the rest of the code.

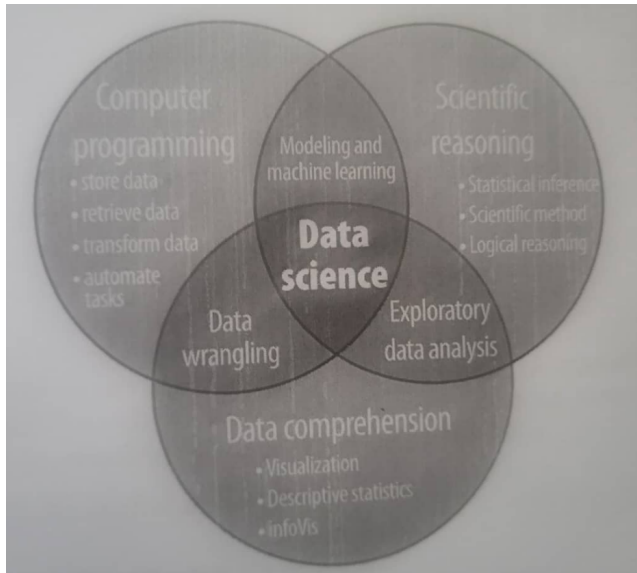
- Grouping
- Loops
- Condition execution

Function are fundamental building block of R that is why we need a solid foundation of functions work. Three main components of a function

- The `body()` : Code inside the function
- The `formals ()`: List of arguments which controls how you can call the function
- The `environment ()`: The map of the location of the function's variables

Education is not the learning of facts, but the training of the mind to think, by **Albert Einstein**

Data science



God bless you