

Week-01 (Programming and data analytics)

Key question	Spreadsheets	SQL	R
What is it?	A program that uses rows and columns to organize data and allows for analysis and manipulation through formulas, functions, and built-in features	A database programming language used to communicate with databases to conduct an analysis of data	A general purpose programming language used for statistical analysis, visualization, and other data analysis
What is a primary advantage?	Includes a variety of visualization tools and features	Allows users to manipulate and reorganize data as needed to aid analysis	Provides an accessible language to organize, modify, and clean data frames, and create insightful data visualizations
Which datasets does it work best with?	Smaller datasets	Larger datasets	Larger datasets
What is the source of the data?	Entered manually or imported from an external source	Accessed from an external database	Loaded with R when installed, imported from your computer, or loaded from external sources
Where is the data from my analysis usually stored?	In a spreadsheet file on your computer	Inside tables in the accessed database	In an R file on your computer
Do I use formulas and functions?	Yes	Yes	Yes
Can I create visualizations?	Yes	Yes, by using an additional tool like a database management system (DBMS) or a business intelligence (BI) tool	Yes

* **R**: A programming language frequently used for statistical analysis, visualization, and other data analysis.

→ why R: Accessible, data-centric, open source, and community.

→ Benefits of R:

- ① Reproducing the analysis.
- ② Processing tons of data.
- ③ Creating visualization.

* R-commands :

- ① view(table) — used to view the data of a table.

* Fundamentals concepts of R:

- ① Function.
- ② Comments.
- ③ Variable.
- ④ Data types.
- ⑤ Vectors.
- ⑥ Pipes.

Week-02 (Programming using Rstudio)

* **Function (R)**: A body of reusable code used to perform specific task in R.

* **Argument (R)**: Information that a function in R needs in order to run.

* **Variable (R)**: A representation of a value in R that can be stored for use later during programming.

* **Vectors (R)**: A group of data elements of the same data type stored in a sequence in R.

* **Pipes (R)**: A tools in R for expressing a sequence of multiple operators, represented with "%>%".

* **R-data structure**:

① Vector

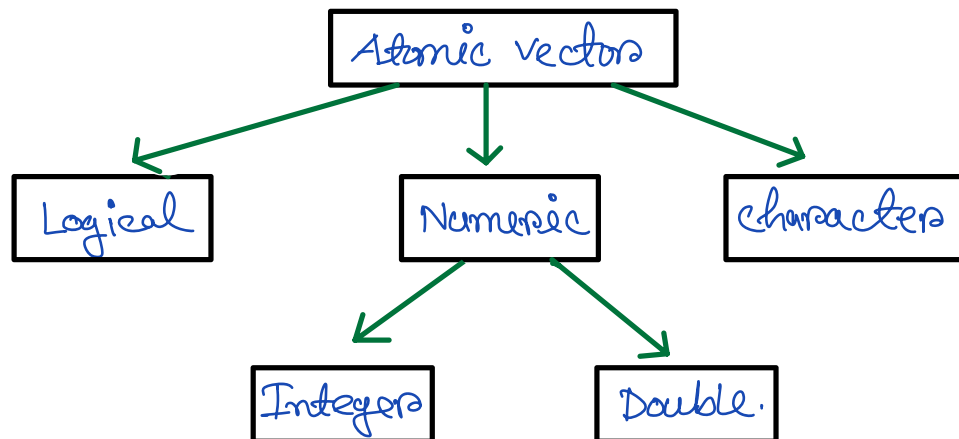
③ Matrix

② Dataframe

④ Array

* Two types of vectors: Atomic vector, and list.

→ Atomic vector: Contains homogenous data.



→ List: Contains series of elements such as single variable, vector, dataframe or matrix.

* R-data and time:

→ R-package: lubridate is R-package for date-and-time.

⦿ today(): function returns today's date.
↳ return "date" object

⦿ now(): function returns the current date-time.
↳ Return "POSIXct" and "POSIXt" object.

R-default date object is "yyyy-mm-dd"

→ converting from string:

- `ymd("2021-01-20")`: return a date object.
- `mdy("January 20th, 2021")`: return a date object.
- `dmy("20-Jan-2021")`: return a date object.
- `ymd("20100708")`: return a date object.

* Operators: A symbol that names the type of operation or calculation to be performed in a formula.

- ① Assignment operator: Used to assign values to a variable or vector.
- ② Arithmetic Operator: Used to complete maths calculation.

* R-package: Unit of reproducible R code.

* Tidyverse (R): A system of packages in R with a common design philosophy for data manipulation, exploration, and visualization.

→ Most essential parts of the workflow for a data analyst:

① `ggplot2()`: Used for data visualization.

② `tidyr()`: Used for data cleaning.

③ `readr()`: Used for importing data.

④ `dplyr()`: Used to perform some common data manipulation tasks.

Week-03 (Working with R)

* **Dataframe**: A collection of columns.

→ **characteristics of dataframe**:

- ① Column should be named.
- ② Each column may store different types of data, but a column only store one type of data.
- ③ Each column should contains same number of data.

→ **data()**: function used to return all the list of data tables that installed R-sessions.

→ **data(dataset-name)**: function load the given data set to the current R-session.

* **R-readr package**:

- ① **read_csv()**: read the csv file.
- ② **read_tsv()**: read the tab separated file.
- ③ **read_delim()**: read general delimited file.

④ `read-buf()`: read fixed-width files.

⑤ `read-table()`: read tabular files where columns are separated by white-space.

⑥ `read-log()`: read web log files.

* Transform data in R:

→ `separate()`: function used to split a column into multiple columns.

→ `unite()`: function merge multiple columns into one single column.

→ `mutate()`: function convert any columns and stored to a new column.

Week-04 (More about visualizations, aesthetics, and annotations)

* some common used R-package for data viz.:

- ① ggplot2.
- ② Plotly
- ③ Lattice
- ④ RGL
- ⑤ Dygraphs.
- ⑥ Leaflet.
- ⑦ Highcharter.
- ⑧ Patchwork
- ⑨ gganimate
- ⑩ ggridge.

→ Core concepts of ggplot2:

- ① Aesthetics: A visual property of an object in the plot i.e. size, shape, and colors of data points.
- ② Geoms: The geometric objects used to represent data i.e. bar, point, line.
- ③ Facets: display smaller group, or subset of the data.
- ④ Labels and annotations: used to customize the plot i.e. title, sub-title, labels, and options.

→ Three basic steps for plotting a ggplot:

- ① starts with `ggplot()` and choose a data set.
- ② Add geom-function to display data.
- ③ Map the variables you want to plot in the arguments of `aes()` function.

Week-05 (Documentation and Reports)

* R Markdown: A file format for making dynamic documents with R.

→ Markdown commands:

- bullet points: Use "*" before each points.
- Add link with path: <url>.
- Add only link: [text](url)
- Add image: ![caption](image-url)