R-studio

Making easier to work in R

- Consistent commands

- unified interface

- Navigate & manage

Packages: are bundles of codes that add functions to R

1. Base Packages: Installed with R but not loaded by default
2. Contributed packages: Need to download, installed and loaded separately.

Packages can be downloaded from

1. CRAN- Comprehensive or Archive Network
2. Crantastic.org
3. Github.com/trending/r

dplyr, tidyr, stringr, lubridate, httr, ggvis, ggplot2, shiny, rio- R i/p&o/p, rmarkdown

one package to load them all – pacman(Package Manager)

**R: An intro-**

**Basic graphics –**

1. plot() #command – also called Basic X-Y plotting

?plot is used to get the help for plot()

1. Bar Charts – categorical variables

Simple is good

Most basic graphic for the most basic data

1. Histogram : data is quantitative, scaled, measured, interval or ratio level.

-see what you have (Shape of distribution, Gaps in distribution, Outlier and symmetry in distribution)

1. Scatterplots (Binary distributions)

- For visualizing the associations b/w two quantitative varibles.

-What to look for 1. two variables is linear 2. Consistence spread across the score 3. Outliers checking, 4. Correlation b/w two variables

1. Overlaying Plots:
   1. Why--> Increased information density i.e., more info with same time.
   2. Use that complement & support one another

Basic Statistics:

1. Summary() function in R (get some precision by numerical or statistical information)
2. Counts for Categories
3. Quartile and mean for quantitative variables
4. Describe() function in R (gets more details)
5. Comes from from psych packages ( we’ll get n, mean, SD, median, 10% of trimmed mean, MAD, min/max, range, skewness, kurtosis & SE)
6. **Don’t forget :** This still comes after graphical summaries

**Selecting Cases:**

It allow to focus on our analysis

1. Selecting by category
2. Selecting by value
3. Selecting by both of above

**Accessing Data:**

* + - 1. Data formats:

1. Data types:
   1. Numeric(Integer, single or double precision, char, logical, complex& raw)
2. Data structures:
   1. Vector
      * 1 or more numbers in a 1D array
      * All same data type
      * R’s basic data object
   2. Matrix
      * Two dimensions data
      * Same length
      * Same data class
      * Columns are not named which can be chose by index
   3. Array
      * Identical to matrix but 3 or more dimensional data
   4. Data frame
      * Can have vectors of multiple types (char+int+log)
      * All should be same length
      * Closest R analogue to spreadsheets
      * R has special functions for working with data frame
   5. List
      * Most flexible data format (Put any data)
      * Ordered collection of elements
      * Any class, length or structure
      * List includes list i.e, list inside many list are possible.

**Coercion :** This is changing data object from one type to another. For ex. Char to logical, matrix to data frame, double precision to integer.

1. Factors:

* categories and names of categories
* An attribute of a vector that specifies the possible values and their order

1. Entering Data:

* version of Ad hoc data.
* Many methods are available like colon operator, seq (sequence), c(concatenate), scan and rep, “ <- “ is the assignment operator

1. Importing data:

* CSV file (comma separated values)
* TXT files (text files)
* XLSX files (excel files)
* JSON files (Java script object notation)
* R has built in functions for importing data in many formats. But use one – “rio”(R input output)
* Rio combines all of R’s import functions into one simple utility.

**Modelling Data**:

* + - 1. Hierarchical Clustering (Like with like) – But similarity depends on our criteria
    - Hierarchical vs set k (k – no. of groups)
    - Measures of distance
    - Divisive vs agglomerative.

We’ll keep it simple by using the following methods.

* + - Euclidean distance
    - Hierarchical clustering
    - Divisive method

1. Principal Components:
   * + Less = more (i.e, less noise and fewer unhelpful variables in data = more meaning in data) . Also known as Dimensional reductionality.
     + PCA – Principal Component Analysis

Two variables -> regression Line -> Measure perpendicular distances-->collapse -> Rotate -> PC

* + - 2D to 1D data transformation but maintained the most important information in data set
    - We made analysis and interpretation easier and more reliable.