Statistics or Machine Learning: What's the Difference?

Statistics and Machine Learning Series Workshops - First workshop
Presented by: Dr. Pascal Tyrrell, Ernest Namdar, Tristal Li

Agenda

- Environment setup
- How to play with my data? (Basic data manipulation)
- How to see my data? (Exploratory data analysis)
    - Summary tables
    - Plots
- What is a model?
    - Fitting a simple linear model

Software Installation & Environment Setup

- Why R?
  - Most commonly used statistical language in academia
  - Pros:
    - Existing packages and functions designed for statistics
    - Freedom in tuning hyperparameters
    - Logistics similar to programming languages (functions, if statements…)
    - Open source software (unlike SAS)
  - Cons:
    - Relatively slow (compared to programming languages like Python)
    - No user-friendly interfaces
- RStudio: IDE for R
- Download RStudio here: RStudio Desktop - Posit
  - RStudio Desktop, Open Source Edition (Free)
  - Follow instructions on this page
- Free statistical software: EZR (Easy R)

R Basics

- Variables
- Simple math
- Lists and indexing
- Loops
- Conditions
- Functions

Basic Data Manipulation

- Play with data using R

- Use Packages:
  - Installation: install.packages(“<package name>”)
  - Usage: library(<package name>)
- Tidyverse package (library(tidyverse))
  - A collection of R packages for data science
    - ggplot2, dplyr, readr,
  - Documentations: Tidyverse
  - Tutorials: R for Data Science
  - Cheat sheets: Posit Cheatsheets

Basic Data Manipulation

- Data cleaning

Tidy Data

1. Each **variable** must have its own **column**
2. Each **observation** must have its own **row**
3. Each **value** must have its own **cell**

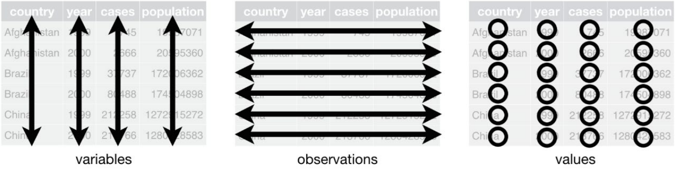

Image from: https://r4ds.had.co.nz/tidy-data.html

Clean column names (follows the same naming convention)

- **snake_case**: consists of lowercase letters, words separated by underscores
- **camelCase**: first letter is lowercase, each new word begins with an uppercase letter

Missing Data (NAs): Some suggestions

- All the missing data need to be NA in R (Rows with NAs will be ignored in many functions)
- Delete the column with NAs if this is not related to your major research objective
- Replace it with a certain value (mean / median / zero / something it means…)
- If >50% of the data are NAs for a certain column, and the dataset is not large enough, delete it

Exploratory Data Analysis

Purpose:

- Visualize the data to help understand patterns
- Check for assumptions, detect outliers and anomalous events
- Find interesting relationships for future analysis

Methods: Summary tables, Graphs, Simple models

Mostly used packages: psych, ggplot2

- **Note**: Here is a tutorial for ggplot2: The Complete ggplot2 Tutorial - Part1 | Introduction To ggplot2 (Full R code)

More information for descriptive statistics: https://mi-data.ca/2023/sas_codes/Descriptive%20Statistics.htm

Statistics or ML?

Based on your purpose…

- For explanability: Stats!
- For prediction and future use: ML!

For more help in Statistics or ML, visit MiData