Quantitative Reasoning II

Data, Variation, Visualization, and Trends
Introduction to Key Concepts in R

What is R?

The R Programming Language

Key Features of R

1. Data Manipulation:

- R provides tools to manage, clean, and manipulate large datasets using libraries like `dplyr` and `tidyr`.

2. Statistical Analysis:

- It supports a wide range of statistical techniques, such as linear and nonlinear modeling, hypothesis testing, time-series analysis, and clustering.

3. Data Visualization:

- R excels at creating high-quality, customizable visualizations using libraries like `ggplot2`, `lattice`, and base plotting.

The R Programming Language

Key Features of R

4. Extensive Libraries:

- Thousands of packages are available on CRAN (Comprehensive R Archive Network) for specialized tasks, including bioinformatics, machine learning, and financial modeling.

5. Interactive Analysis:

- R works well with interactive environments like RStudio and Jupyter notebooks, providing an intuitive workflow for coding and exploring data.

6. Community and Open Source:

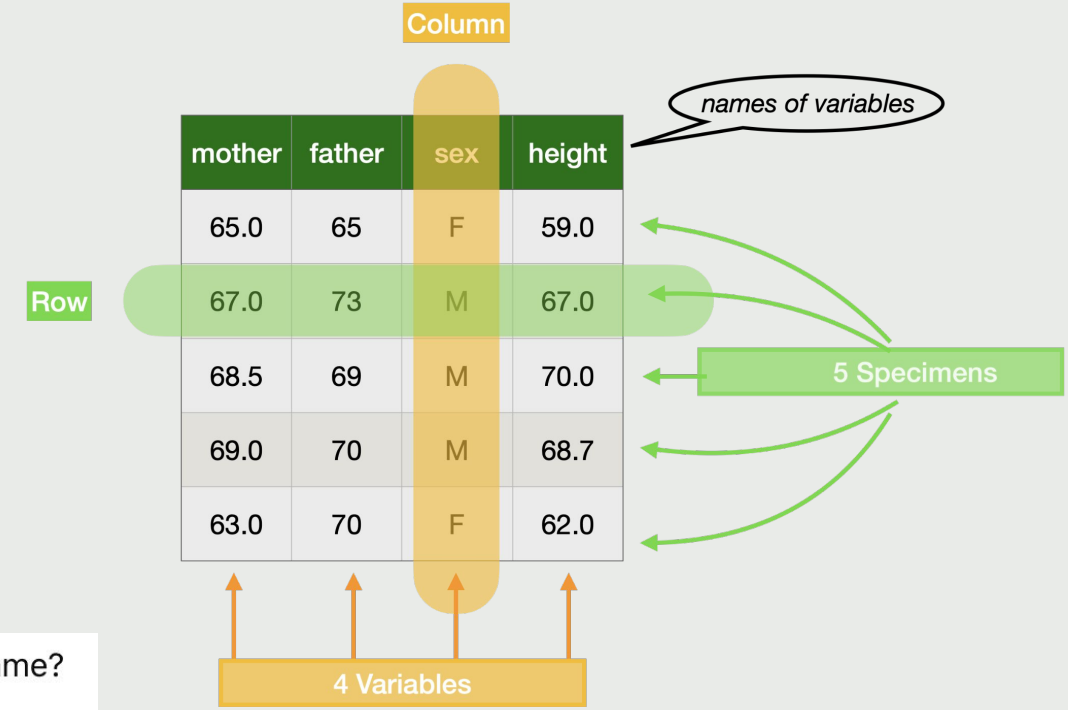
- R is free and open source, with a strong community of contributors who continuously develop and maintain packages and resources.

What is a Data Frame?

Definition: A data frame stores data as rows (observations) and columns (variables).

Ex: Galton - a dataframe from the 1880's on human heights.

Key Terms: Specimen, variable, unit of observation, quantitative, categorical.



Which type of variable is `sex` in the Galton data frame?

A. Quantitative

B. Categorical

Working with Data Frames in R

Which function would you use to preview the first few rows of a data frame?

A. `tail()`
B. `head()`
C. `names()`
D. `nrow()`

Raise your hand if you know:
What do the other functions do?

Arguments in Functions

Arguments: are additional pieces of information provided to a function to control its behavior or output.

- Some require or support arguments (e.g., `head(3)`).
- Some do not (e.g., `nrow()`).

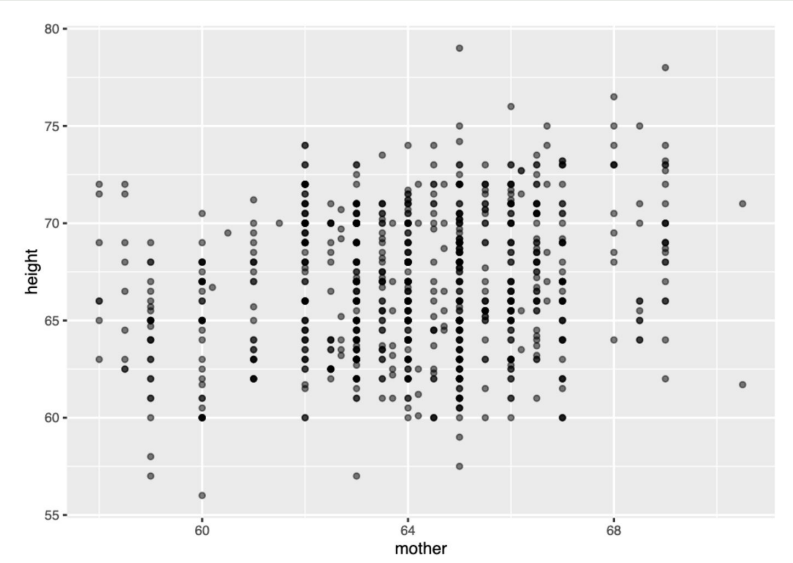
Data Visualization

When plotting points on a graph ("Annotated Point Plots"), there are explanatory and response variables.

In a point plot, which variable is usually mapped to the x-axis?

A. Response

B. Explanatory



Run Code

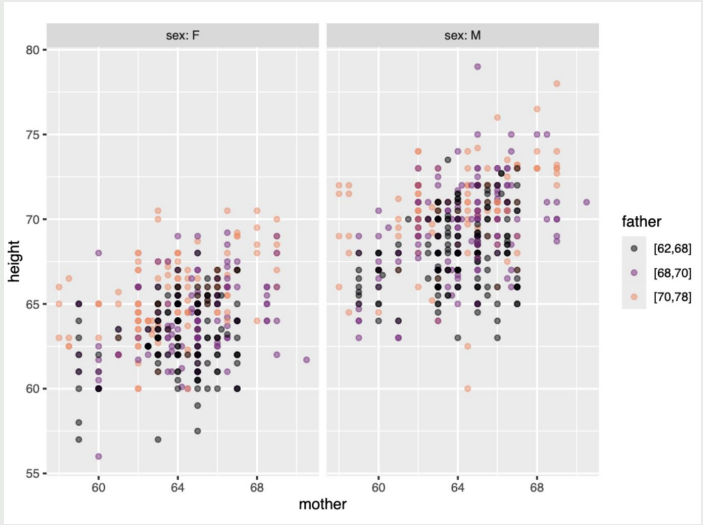
```
1  Galton |> point_plot(height ~ mother)
```

Data Visualization with Multiple Explanatory Variables

What does this line of code plot (multiple explanatory variables)?

Run Code

```
1  Galton |> point_plot(height ~ mother + father + sex)
```



Additional explanatory variables get incorporated into the visual in the following order: x-axis, color, facet

Graphical Annotations

- Model vs Violin

Model:

Adds a statistical model layer to the point plot to show the relationship or trend between variables.

Commonly used for regression-like trends or general patterns in the data.

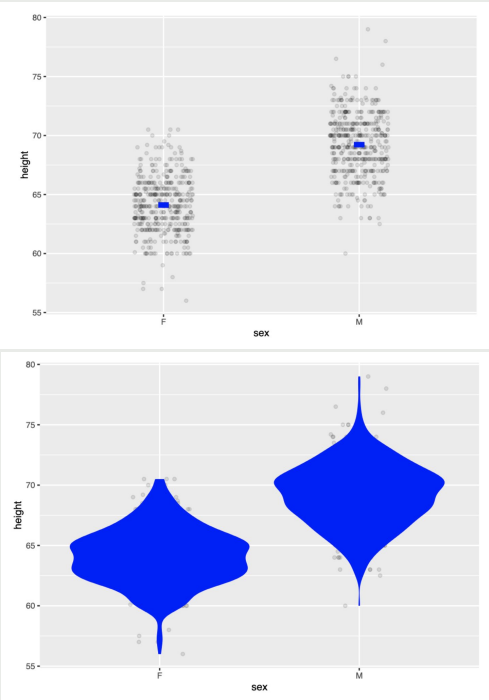
Violin:

Adds a **violin plot** layer to the point plot to visualize the distribution of the data.

Suitable when an explanatory variable mapped to `x` is categorical (e.g., `sex` or `species`).

What is the graphical technique “jittering”?

```
Run Code
1  Galton |> point_plot(height ~ sex,
2                       annot = "model",
3                       point_ink = 0.1,
4                       model_ink = 1.0)
```



Data Wrangling

Data wrangling, also known as data cleaning or data preprocessing, is the process of transforming raw data into a structured, clean, and usable format for analysis.

This involves manipulating and organizing data to make it more accessible and ready for exploration, visualization, and modeling.

`arrange()`: Sorting rows based on one or more variables.
`filter()`: Selecting rows based on specified conditions.
`mutate()`: Creating new variables or modifying existing ones.
`select()`: Choosing specific columns to keep or remove.
`summarize()`: Aggregating data to generate summary statistics (e.g., mean, variance).
`.by =`: Grouping data for operations (e.g., calculating summaries for each group).

Data Wrangling Exercise

Step 1

Install RStudio

Step 2

Work through data wrangling exercises.
(Ungraded, but useful skills for your final project)

Dataset

Data Set: heartrate.csv

- This data is from an exercise study on maximum heart rates.
- Each row is a person.
- The two variables are the person's age in years and their maximum heart rate (hrmax) in beats per minute, as measured by a treadmill test.

Data is located on the class Github.

Step 1: RStudio Installation

https://posit.co/download/rstudio-desktop/

You will first install R, then the RStudio Environment.