STAT212
EXAM 1
Jessica Lin

Data Structures

- Vectors hold one type, lists can hold many types
- Data frames and tibbles are lists of equal-length vectors
- Factors are categorical variables with levels and ordering matters

Core Tidyverse Verbs

- select chooses columns
- filter subsets rows
- arrange reorders rows
- mutate creates or transforms variables
- summarize reduces data
- group_by defines groups for summaries

Grouped Data

- Grouping affects summaries and mutations
- summarize drops grouping by default
- Always check grouping before modeling

Missing Data

- MCAR: missing completely at random
- MAR: missing related to observed variables
- MNAR: missing related to unobserved values
- Investigate missingness before removing data
- Avoid extrapolating when missingness is structured

Data Cleaning

- Rename variables for clarity
- Convert variable types when needed
- Trim whitespace and standardize strings
- Identify and handle NA values carefully

Strings and Dates

- Strings are often used to build dates or labels
- Pattern matching helps clean messy text
- Be careful with separators and formatting

Joins

- Joins combine tables using keys
- Left join keeps all rows from the first table
- Inner join keeps only matching rows
- Duplicated keys create repeated rows

Reshaping Data

- Long data is better for analysis and plotting
- Wide data is better for presentation
- Pivoting changes the structure, not the data itself

Conditional Logic

- Use conditions to create categories or flags
- Order of conditions matters
- Always include a default case

Functions

- Functions should do one clear task
- Use arguments instead of hard-coded values
- Output should be predictable and consistent

Modeling Preparation

- Check variable ranges before modeling
- Be cautious with outliers and missing values
- Do not extrapolate beyond observed data