COURSE GLOSSARY

# Introduction to the Tidyverse

Aesthetic (aes): A mapping in ggplot2 that links data variables to visual properties of a plot (for example x, y, color, or size) so information is encoded visually

arrange: A dplyr verb that sorts the rows of a dataset in ascending or descending order based on one or more variables

Categorical variable: A variable whose values belong to a limited set of discrete groups or categories, for example continent or country names

Data frame: A rectangular data structure in R that stores observations as rows and variables as columns, similar to a spreadsheet or SQL table

dplyr verb: A core function in the dplyr package that performs a single, atomic data transformation (e.g., filter, arrange, mutate, summarize, group_by)

filter: A dplyr verb that returns a subset of rows (observations) from a dataset that satisfy one or more logical conditions

Gapminder: A public dataset (and R package) that contains country-level social and economic indicators over time, such as life expectancy, population, and GDP per capita

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

Hallucination: When a model produces confident but incorrect or fabricated information, often due to gaps or biases in its training data or reasoning process

ggplot2: An R package for layered, grammar-of-graphics plotting that builds visualizations by specifying data, aesthetic mappings, and geometric layers

group_by: A dplyr verb that defines groups of rows so subsequent summary operations are performed within each group rather than on the entire dataset

mutate: A dplyr verb that creates new columns or transforms existing ones by computing values from other variables in the dataset

Numeric variable: A variable that represents quantitative values which can be measured and used in arithmetic, such as population or life expectancy

Observation: A single row in a dataset representing one unit of analysis, such as a country-year record in Gapminder

Pipe: An operator (commonly %>%) that passes the output of one expression directly as the input to the next, enabling readable chains of data transformations

R package: A bundle of R functions, data, and documentation created by others to extend R's base capabilities and make common tasks easier to reuse

summarize: A dplyr verb that collapses groups or an entire dataset into summary statistics (like mean or sum), producing one row per group or overall

Tibble: A modern variant of a data frame that prints more compactly and behaves predictably in the tidyverse, often used as the default table format

Tidyverse: A coherent collection of R packages designed for data science tasks such as importing, transforming, and visualizing data, built to work together with consistent syntax and philosophy

Variable: A column in a dataset that stores values of a particular type (numeric, categorical, etc.) describing an attribute of observations