Data analysis and visualization with R

Outline

String operation
Vector operation
Manipulate data with ‘dplyr’

- Built-in functions
  1. filter/select
  2. arrange/mutate
  3. group by
  4. summarise
- Join data sets

Data Visualization with ‘ggplot2’

- Construct a ggplot
  1. scatterplots
  2. lineplots
  3. bar charts
  4. histograms
  5. boxplots
- Comparing variables

String operation

string:

- is.character(): whether data type is character
- nchar(): count the number of characters
- paste(): combine several character variables into one string
- strsplit(): split 1 string into many shorter strings
- substr(): extract a portion of a string
- grep():/grepl() determine if a pattern of text exists in a character variable
- toupper(): change lowercase to uppercase
- tolower(): change uppercase to lowercase
- gsub(): replace parts of strings

Vector operation

Vector

- append: add a element to specific position of a vector
- match: only return the first encounter of a match in a vector
- str_replace: replace parts of strings in a vector
- table: calculate the frequency of each element of a vector
- sort: sort a vector or factor into ascending or descending order

Manipulate data with ‘dplyr’

dplyr: a set of functions, database query-type operations

dplyr aims to provide a function for each basic verb of data manipulation:

- filter(): select a subset of rows
- arrange(): reorder rows
- select(): select a subset of columns
- mutate(): add new columns
- group_by(): convert to several groups
- summarise(): compute aggregate values

Join data sets

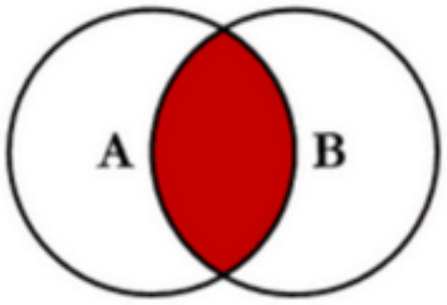
Need information from 2 or more data frames

- Adding Columns: merge two data frames (datasets) horizontally, they have one/multiple same column names

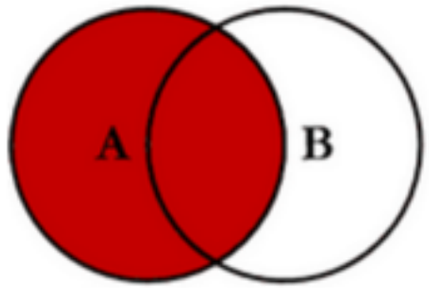
  1. inner_join
  2. left_join
  3. right_join
  4. full_join

- Adding Rows: join two data frames (datasets) vertically, their have all the same column names
  bind

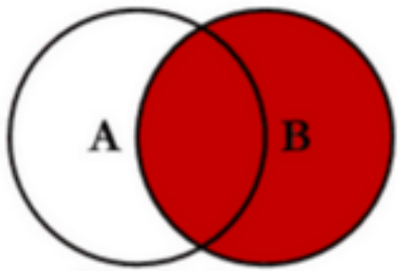
inner_join: keep only match



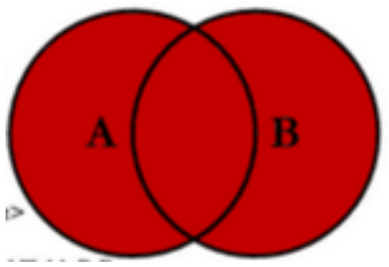
left_join: keep left, find match in right



right_join: keep right, find match in left



full_join: keep all



ggplot2

Advantage of ggplot2:

- plot specification at a high level of abstraction
- very flexible
- theme system for polishing plot appearance
- mature and complete graphics system
- many users, active mailing list

A graph includes:

- data: data frame
- aesthetic mappings:position (i.e., on the x and y axes), color (“outside” color), fill (“inside” color), shape (of points), line type, size
- geometric objects: scatter, line, bar, boxplot, histogram
- label: add labels at x and y axes
- title: add title at the top of the graph
- statistical transformations: boxplot, prediction
- scales: controlling aesthetic mapping
- faceting: create separate graphs for subsets of data