Data Science Notes

Data Wrangling

Verb	Action
arrange()	arrange the rows by column
filter()	filter a subset of rows
select()	select a subset of columns
mutate()	mutate or create a column
summarize()	calculate a numerical sumary
group_by()	group rows by column

Goals of Wrangling

- keep certain variables
- define new variables
- reformat/clean variables
- combine datasets
- Process strings / text
- numerically explore datasets

Filtering

Symbol	meaning
==	equal to
!=	not equal to
>	greater than
>=	" " and equal to
<	less than
<=	"|" and equal to
%in%, c(...,...)	a list of mutiple variables

Dates

mutate to dates from strings

year(today)
month(today) ↓
week(today) label = TRUE
mday(today)
yday(today)
wday(today) ↓

Reshaping

Pivoting

Pivot_longer(data, cols, name_to = "name", values_to = "value") Combine values into 1 variable

Pivot_wider(data, names_from = "name", values_from = "value") Spread values across new variables

Joining

Mutating Joins

left_join() keeps all observations from the left but discards right ones that dont match

Inner_join() keep only observations that match in both right and left

full_join() keeps all observations

Filtering Joins

semi_join() discards observations in left without match in right. If multiple matches in left keeps only 1.

anti_join() Discards observations in left that dont match in right.

Factors

order of factor levels + labels

fct_relevel() manually reorder levels

fct_reorder() reorder by another variable

fct_infreq() highest to lowest freqyency

fct_rev() reverse current order

fct_recode() manualy change levels

fct_lump() group least comon values

Strings

String functions

str_replace() finds first part and replaces

str_replace_all() replaces all instances

str_to_lower() Upper → lower case

str_sub() only keeps a subset of the string

str_length() number of characters

str_detect() if "-" is in the string = True

Data Import: read_csv(), read_delim(), read_sheet(), st_read()