Problem set 2

Put your name here

Table of contents

### Getting started

library(tidyverse)
library(janitor)

### Read the data

Rows: 195 Columns: 9
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Breed
dbl (8): 2013 Rank, 2014 Rank, 2015 Rank, 2016 Rank, 2017 Rank, 2018 Rank, 2...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.
Rows: 195 Columns: 17
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (3): Breed, Coat Type, Coat Length
dbl (14): Affectionate With Family, Good With Young Children, Good With Othe...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

breed\_rank <- read\_csv("data/breed\_rank.csv")
breed\_traits <- read\_csv("data/breed\_traits.csv")

### Clean the data

Display variables.

names(breed\_rank)

[1] "Breed" "2013 Rank" "2014 Rank" "2015 Rank" "2016 Rank" "2017 Rank"
[7] "2018 Rank" "2019 Rank" "2020 Rank"

names(breed\_traits)

 [1] "Breed" "Affectionate With Family"
 [3] "Good With Young Children" "Good With Other Dogs"
 [5] "Shedding Level" "Coat Grooming Frequency"
 [7] "Drooling Level" "Coat Type"
 [9] "Coat Length" "Openness To Strangers"
[11] "Playfulness Level" "Watchdog/Protective Nature"
[13] "Adaptability Level" "Trainability Level"
[15] "Energy Level" "Barking Level"
[17] "Mental Stimulation Needs"

Make better names.

breed\_traits <- breed\_traits |>
 clean\_names()

### Manipulate the data using dplyr

Maka a summary.

breed\_traits |>
 group\_by(shedding\_level) |>
 summarise(n = n())

# A tibble: 6 × 2
 shedding\_level n
 <dbl> <int>
1 0 1
2 1 27
3 2 41
4 3 109
5 4 16
6 5 1

Filter the shedding\_level 0.

breed\_traits <- breed\_traits |>
 filter(shedding\_level != 0)

Check if manipulation was successful.

breed\_traits |> count(shedding\_level)

# A tibble: 5 × 2
 shedding\_level n
 <dbl> <int>
1 1 27
2 2 41
3 3 109
4 4 16
5 5 1

Make an untidy data frame.

untidy\_scores <- breed\_traits |>
 mutate(untidy\_score = shedding\_level +
 coat\_grooming\_frequency + drooling\_level) |>
 select(breed, untidy\_score)

Arrange scores in descending order.

untidy\_scores |>
 arrange(desc(untidy\_score))

# A tibble: 194 × 2
 breed untidy\_score
 <chr> <dbl>
 1 Bernese Mountain Dogs 11
 2 Leonbergers 11
 3 Newfoundlands 10
 4 Bloodhounds 10
 5 St. Bernards 10
 6 Old English Sheepdogs 10
 7 Dogues de Bordeaux 10
 8 Neapolitan Mastiffs 10
 9 Black Russian Terriers 10
10 Tibetan Mastiffs 10
# ℹ 184 more rows

### Tidying the data

How does this this data set fail to meet the criteria for tidy data?

There are three interrelated rules which make a dataset tidy:

1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

We have a year and a rank variable, but neither of these variables have their own column. Shown above is one observation, by dog breed. But that “one” observation is actually eight separate observations: the rank in 2013, the rank in 2014, etc. Each observation needs to have its own row.

Make pivoted data with a year and a rank variable.

ranks\_pivoted <- breed\_rank |>
 pivot\_longer(`2013 Rank`:`2020 Rank`,
 names\_to = "year",
 values\_to = "rank")

Rename breed and make the year variable numeric.

ranks\_pivoted <- ranks\_pivoted |>
 rename(breed = Breed) |>
 mutate(year = parse\_number(year))

Filter data to only Bernese Mountain Dogs.

ranks\_pivoted <- ranks\_pivoted |>
 filter(str\_detect(breed, "Bernese"))

Plot rankings across time.

ranks\_pivoted |>
 ggplot(aes(x = year, y = rank, label = rank)) +
 geom\_point(size = 3) +
 geom\_text(vjust = 2)

