Homework 2

Caitlin Jagla

03/08/2022

Set-up

Load the tidyverse library and the 3 datasets regarding dog breeds.

```
library(tidyverse)

breed_rank <- read_csv("breed_rank.csv")
breed_traits <- read_csv("breed_traits.csv")
trait_desc <- read_csv("trait_description.csv")
```

Use 2 functions of your choice to investigate the datasets so you know what we are working with.

```
breed_rank |> dim_desc()
```

```
## [1] "[195 x 11]"
```

```
breed_traits |> dim_desc()
```

```
## [1] "[195 x 17]"
```

```
trait_desc |> dim_desc()
```

```
## [1] "[16 x 4]"
```

```
breed_rank |> head()
```

```
## # A tibble: 6 x 11
##   Breed  ‘2013 Rank‘ ‘2014 Rank‘ ‘2015 Rank‘ ‘2016 Rank‘ ‘2017 Rank‘ ‘2018 Rank‘
##   <chr>        <dbl>       <dbl>       <dbl>       <dbl>       <dbl>       <dbl>
## 1 Retri~           1           1           1           1           1           1
## 2 Frenc~          11           9           6           6           4           4
## 3 Germa~           2           2           2           2           2           2
## 4 Retri~           3           3           3           3           3           3
## 5 Bulld~           5           4           4           4           5           5
## 6 Poodl~           8           7           8           7           7           7
## # i 4 more variables: ‘2019 Rank‘ <dbl>, ‘2020 Rank‘ <dbl>, links <chr>,
## #   Image <chr>
```

```
breed_traits |> head()
```

```
## # A tibble: 6 x 17
##   Breed     Affectionate With Fa~1 Good With Young Chil~2 ‘Good With Other Dogs‘
##   <chr>                      <dbl>                  <dbl>                  <dbl>
## 1 Retrieve~                      5                      5                      5
## 2 French B~                      5                      5                      4
## 3 German S~                      5                      5                      3
## 4 Retrieve~                      5                      5                      5
## 5 Bulldogs                       4                      3                      3
## 6 Poodles                        5                      5                      3
## # i abbreviated names: 1: ‘Affectionate With Family‘,
## #   2: ‘Good With Young Children‘
## # i 13 more variables: ‘Shedding Level‘ <dbl>, ‘Coat Grooming Frequency‘ <dbl>,
## #   ‘Drooling Level‘ <dbl>, ‘Coat Type‘ <chr>, ‘Coat Length‘ <chr>,
## #   ‘Openness To Strangers‘ <dbl>, ‘Playfulness Level‘ <dbl>,
## #   ‘Watchdog/Protective Nature‘ <dbl>, ‘Adaptability Level‘ <dbl>,
## #   ‘Trainability Level‘ <dbl>, ‘Energy Level‘ <dbl>, ...
```

```
trait_desc |> head()
```

```
## # A tibble: 6 x 4
##   Trait                     Trait_1               Trait_5              Description
##   <chr>                     <chr>                 <chr>                <chr>
## 1 Affectionate With Family  Independent           Lovey-Dovey          How affect~
## 2 Good With Young Children  Not Recommended       Good With Children   A breed’s ~
## 3 Good With Other Dogs      Not Recommended       Good With Other Dogs How genera~
## 4 Shedding Level            No Shedding           Hair Everywhere      How much f~
## 5 Coat Grooming Frequency   Monthly               Daily                How freque~
## 6 Drooling Level            Less Likely to Drool  Always Have a Towel  How drool-~
```

1. New variable

Let’s say we would like to know which dog breeds increased most in rank from 2013 to 2020.

A. Create a new variable called `diff_rank` that is the difference in rank between 2013 and 2020.

```
breed_rank <- breed_rank |> mutate(diff_rank = `2020 Rank` - `2013 Rank`)
```

B. Show the 10 breeds that gained the most interest from 2013 to 2020.

```
breed_rank |>
  arrange(desc(diff_rank)) |>
  head(n=10) |>
  select(Breed, `2013 Rank`, `2020 Rank`, diff_rank)
```

```
## # A tibble: 10 x 4
##    Breed                               ‘2013 Rank‘ ‘2020 Rank‘ diff_rank
##    <chr>                                     <dbl>       <dbl>     <dbl>
##  1 Treeing Walker Coonhounds                   101         153        52
##  2 American English Coonhounds                 146         185        39
##  3 Spaniels (Irish Water)                      141         174        33
##  4 Chinooks                                    156         186        30
##  5 Salukis                                     115         144        29
##  6 Afghan Hounds                                95         122        27
##  7 Kuvaszok                                    150         177        27
##  8 Petits Bassets Griffons Vendeens            138         164        26
##  9 Setters (Irish Red and White)               145         170        25
## 10 Miniature Pinschers                          53          77        24
```

2. Reshape to long form

Begin with the breed_rank dataset and create a long-form dataset where the numeric year is in one column and the numeric rank is in another column. Save only the Breed, year, rank, and diff_rank columns. Save the result into `breed_rank_long` and show it in the report. `breed_rank_long` should have dimensions 1560 x 4.

```
breed_rank_long <-  breed_rank |>
    pivot_longer(cols = `2013 Rank`:`2020 Rank`,
                 names_to = "year", values_to = "rank") |>
    separate(col = "year", into = "year", sep = " ") |>
    select(Breed, year, rank, diff_rank)

dim(breed_rank_long)
```

```
## [1] 1560    4
```

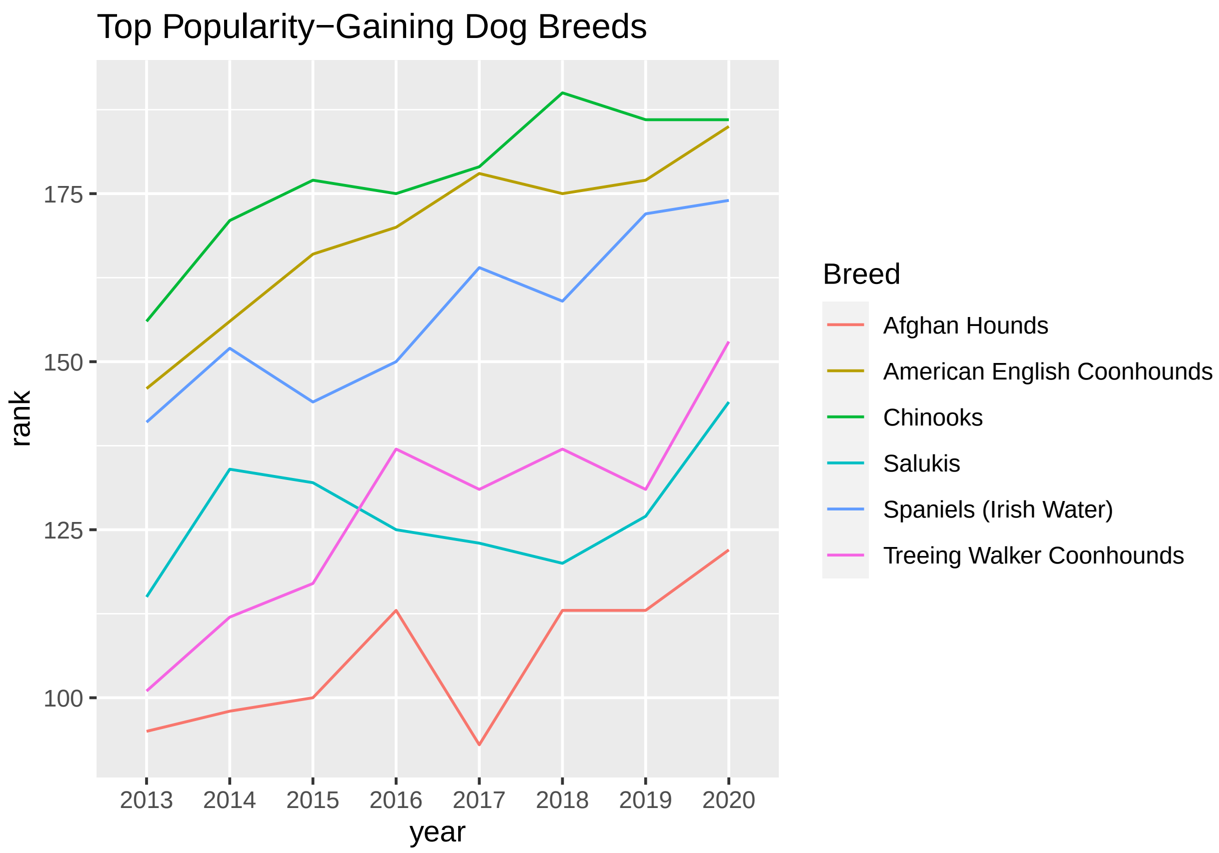
3. Plot

Use a line graph to see the rank of each breed over time for the 5 breeds that gained the most popularity from 2013 to 2020.

```
top_diff <- breed_rank |>
            arrange(desc(diff_rank)) |>
            head() |>
            pull(Breed)



breed_rank_long |>
  filter(Breed %in% top_diff) |>
  ggplot(aes(color = Breed, x = year, y = rank)) +
  geom_line(aes(group = Breed)) +
  labs(title = "Top Popularity-Gaining Dog Breeds")
```



4. Merge

A. Start with the `breed_rank_long` dataset and create a new dataset that preserves only rows for 2013 and 2020. Call this new dataset `dogs`. Show `dogs` in the report.

```
dogs <- breed_rank_long |>
  filter(year == c(2013, 2020))

dogs |> head()
```

```
## # A tibble: 6 x 4
##   Breed                   year   rank diff_rank
##   <chr>                   <chr> <dbl>     <dbl>
## 1 Retrievers (Labrador) 2013      1         0
## 2 Retrievers (Labrador) 2020      1         0
## 3 French Bulldogs       2013     11        -9
## 4 French Bulldogs       2020      2        -9
## 5 German Shepherd Dogs  2013      2         1
## 6 German Shepherd Dogs  2020      3         1
```

B. Use a `left_join()` with `dogs` on the left and `breed_traits` on the right. Save the resulting dataset into `dogs` and show it in the report.

```r
dogs <- left_join(dogs, breed_traits)
```

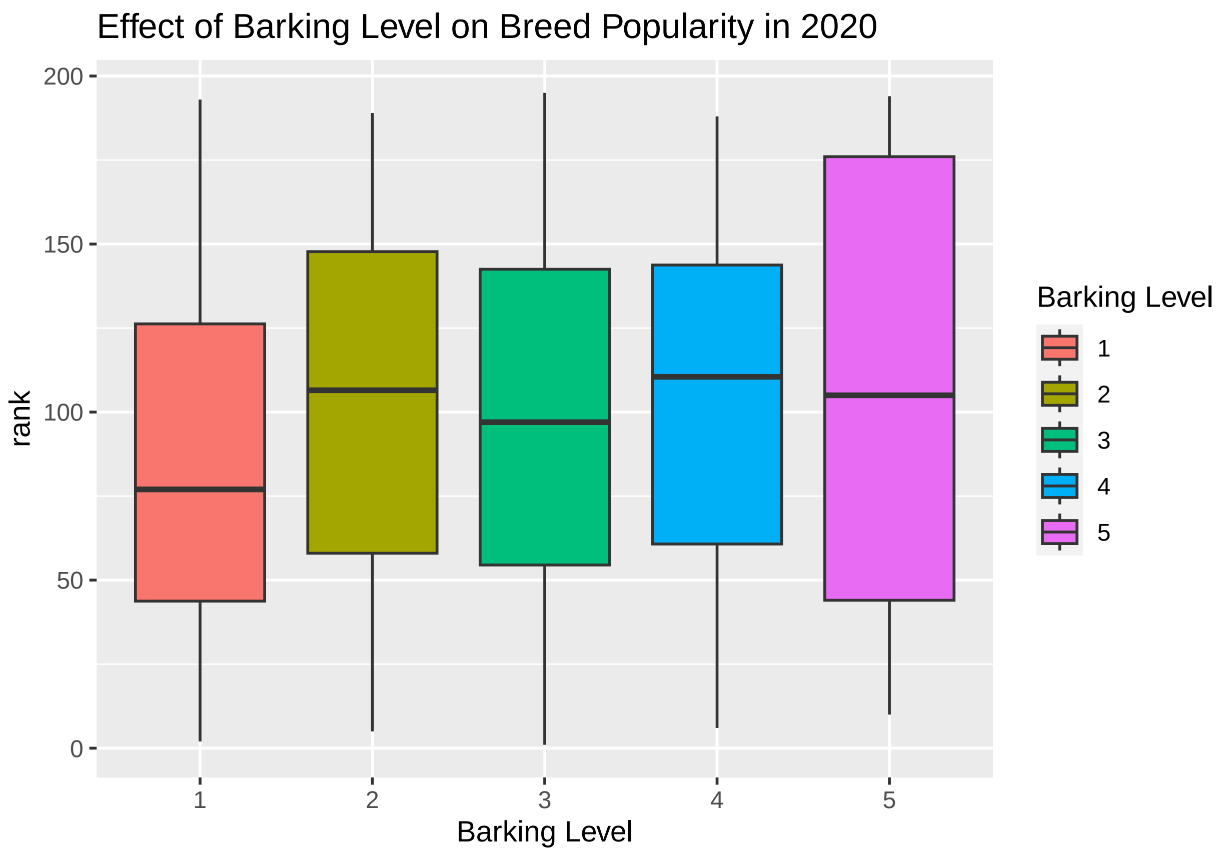
```
## Joining with ‘by = join_by(Breed)‘
```

```r
head(dogs)
```

```
## # A tibble: 6 x 20
##   Breed      year   rank diff_rank Affectionate With Fa~1 Good With Young Chil~2
##   <chr>      <chr> <dbl>     <dbl>                  <dbl>                  <dbl>
## 1 Retriever~ 2013      1         0                      5                      5
## 2 Retriever~ 2020      1         0                      5                      5
## 3 French Bu~ 2013     11        -9                      5                      5
## 4 French Bu~ 2020      2        -9                      5                      5
## 5 German Sh~ 2013      2         1                      5                      5
## 6 German Sh~ 2020      3         1                      5                      5
## # i abbreviated names: 1: ‘Affectionate With Family‘,
## #   2: ‘Good With Young Children‘
## # i 14 more variables: ‘Good With Other Dogs‘ <dbl>, ‘Shedding Level‘ <dbl>,
## #   ‘Coat Grooming Frequency‘ <dbl>, ‘Drooling Level‘ <dbl>, ‘Coat Type‘ <chr>,
## #   ‘Coat Length‘ <chr>, ‘Openness To Strangers‘ <dbl>,
## #   ‘Playfulness Level‘ <dbl>, ‘Watchdog/Protective Nature‘ <dbl>,
## #   ‘Adaptability Level‘ <dbl>, ‘Trainability Level‘ <dbl>, ...
```

C. Now that rank and breed traits are in the same dataset, create a plot of your choice to show the relationship between `Barking Level` and 2020 ranking. Write a sentence to interprete your plot. Remember that high rank = more popular.

```r
dogs |> filter(year == 2020 & `Barking Level`!=0) |> mutate_at("Barking Level", factor) |>
  ggplot() +
  geom_boxplot(aes(fill = `Barking Level`, x = `Barking Level`, y = rank)) +
  labs(title = "Effect of Barking Level on Breed Popularity in 2020")
```



There does not appear to be any relationship between barking level and breed popularity in 2020.