hw-2_notebook

Owen Bruce

1. Read in the file lord-of-the-rings-trilogy.csv that contains data on the number of words spoken in the Lord of the Rings movies for males and females of three of the main races of Middle Earth.

   ```
   library(tidyverse)
   ```

   ```
   -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
   v dplyr     1.1.4     v readr     2.1.5
   v forcats   1.0.0     v stringr   1.5.1
   v ggplot2   3.5.1     v tibble    3.2.1
   v lubridate 1.9.3     v tidyr     1.3.1
   v purrr     1.0.2
   -- Conflicts ------------------------------------------ tidyverse_conflicts() --
   x dplyr::filter() masks stats::filter()
   x dplyr::lag()    masks stats::lag()
   i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
   ```

   ```
   lotr_dialogue <- readr::read_csv(
     "./lord-of-the-rings-trilogy.csv"
     )
   ```

   ```
   Rows: 3 Columns: 7
   -- Column specification --------------------------------------------------------
   Delimiter: ","
   chr (1): movie
   dbl (6): elf_female, elf_male, Hobbit_female, hobbit_Male, man_Female, Man_male

   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.
   ```

2. List all the ways this dataset is not tidy.

   1. Some columns have multiple variables - male's and female's should be separated into a gender variable and the race should be separated into a race variable so that each column is one variable.

   2. Each row should have one observation - currently each row represents one movie, it should be organized with every row as a unique combination of movie race and gender

   3. The columns have inconsistent capitalization in the naming and races

3. How many columns and rows would you have if this dataset was tidy?

Columns: Movie, Gender, Race, Words Spoken = 4 columns

Rows: 3 movies * 2 genders * 3 races = 18 rows

4. What would the column names be in tidy format?

movie_name, gender, race, words_spoken

5. Tidy the dataset you read in from A1.

```r
#rename the columns to have consistent capitalization
lotr_dialogue_r <- lotr_dialogue |>
  dplyr::rename(
    hobbit_female = Hobbit_female,
    hobbit_male = hobbit_Male,
    human_female = man_Female,
    human_male = Man_male,
  )
#pivot the data
tidy_lotr <- lotr_dialogue_r |>
  tidyr::pivot_longer(
    cols = !"movie",
    names_to = c("race", "gender"),
    names_sep = "_",
    values_to = "words_spoken"
    )
```

6. What's the total number of words spoken by: a) male hobbits, b) female elves, and c) male elves?

```r
#A:
tidy_lotr |>
  dplyr::filter(
    race == "hobbit",
    gender == "male"
    ) |> tally(words_spoken)
```

```
# A tibble: 1 x 1
      n
  <dbl>
1  8780
```

```
#B:
tidy_lotr |>
  dplyr::filter(
    race == "elf",
    gender == "female"
    ) |> tally(words_spoken)
```

```
# A tibble: 1 x 1
      n
  <dbl>
1  1743
```

```
#C:
tidy_lotr |>
  dplyr::filter(
    race == "elf",
    gender == "male"
    ) |> tally(words_spoken)
```

```
# A tibble: 1 x 1
      n
  <dbl>
1  1994
```

7. Is the number of spoken words in a movie dominated by a single race? 8. Does the dominant race depend on the movie?

```
tidy_lotr |> dplyr::group_by(
  movie,
  race
  ) |>
  summarize(
    total = sum(words_spoken)
    )
```

```
`summarise()` has grouped output by 'movie'. You can override using the
`.groups` argument.
```

```
# A tibble: 9 x 3
# Groups:   movie [3]
```

```
  movie                      race   total
  <chr>                      <chr>  <dbl>
1 The Fellowship of the Ring elf     2200
2 The Fellowship of the Ring hobbit  3658
3 The Fellowship of the Ring human   1995
4 The Return of the King     elf      844
5 The Return of the King     hobbit  2463
6 The Return of the King     human   3990
7 The Two Towers             elf      693
8 The Two Towers             hobbit  2675
9 The Two Towers             human   2727
```

Hobbits dominate the first movie in terms of words spoken, while humans dominate the second.

The third is fairly balanced between humans and hobbits.