taha_assignment2

Taha Abbasi-Hashemi

2024-10-08

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## -- 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 become errors
```

```
library(babynames)
```

```
## Warning: package 'babynames' was built under R version 4.3.3
```

```
library(datadictionary)
```

```
## Warning: package 'datadictionary' was built under R version 4.3.3
```

```
library(tinytex)
```

```
## Warning: package 'tinytex' was built under R version 4.3.3
```

```
library(stringr)
```

Part A

1,2. When I only look at the data it isnt clear what is actually being shown here. My first instinct was to assume that this was the number of males or females.

I believe this is an issue of being too wide and not long enough. I would revise this to have multiple rows, where each row indicates the race, gender and a column representing "number of words spoken".

This might also be an instance where the data is too separated. For instance it might make sense to unite the data and have Man be shown as male/female.

I know lord of the rings is famous for having very few women talking, and there is no instance of a woman talking to another woman, but was there actually no human women talking in the first movie? It doesnt seem to be explicitly or implicitly missing, but sometimes I have seen NA values appear as a 0 instead.

Another thing I noticed is that male is sometimes lowercase and sometimes uppercase...

Also the movie The column name movie isnt capitalized.

```
df <- read.csv('lord-of-the-rings-trilogy.csv')

print(df)
```

```
##                        movie elf_female elf_male Hobbit_female hobbit_Male
## 1 The Fellowship of the Ring       1229      971            14        3644
## 2             The Two Towers        183      510             2        2673
## 3     The Return of the King        331      513             0        2463
##   man_Female Man_male
## 1          0     1995
## 2        268     2459
## 3        401     3589
```

3. If the dataset was tidy there would be 4 columns and 18 rows.
4. The columns I would have are Movie, Race, Gender and Words_Spoken. # Part B
5.

```
tidy_df <- df %>%
  pivot_longer(cols = -movie, names_to = c("Race", "Gender"), names_sep = "_", values_to = "Words_Spoke
  mutate(Race = str_to_title(Race), Gender = str_to_title(Gender)) # Make genders and race all title ca
colnames(tidy_df) <- str_to_title(colnames(tidy_df)) # make all columsn title case
print(tidy_df)
```

```
## # A tibble: 18 x 4
##    Movie                      Race   Gender Words_spoken
##    <chr>                      <chr>  <chr>         <int>
##  1 The Fellowship of the Ring Elf    Female         1229
##  2 The Fellowship of the Ring Elf    Male            971
##  3 The Fellowship of the Ring Hobbit Female           14
##  4 The Fellowship of the Ring Hobbit Male           3644
##  5 The Fellowship of the Ring Man    Female            0
##  6 The Fellowship of the Ring Man    Male           1995
##  7 The Two Towers             Elf    Female          183
##  8 The Two Towers             Elf    Male            510
##  9 The Two Towers             Hobbit Female            2
## 10 The Two Towers             Hobbit Male           2673
## 11 The Two Towers             Man    Female          268
## 12 The Two Towers             Man    Male           2459
## 13 The Return of the King     Elf    Female          331
## 14 The Return of the King     Elf    Male            513
```

```
## 15 The Return of the King     Hobbit Female           0
## 16 The Return of the King     Hobbit Male          2463
## 17 The Return of the King     Man    Female         401
## 18 The Return of the King     Man    Male          3589
```

```
#tw <- tidy_df %>%
#  group_by(Movie) %>%
#  filter(Race == "Elf") %>%
#  filter(Gender=="Female")
tw <- tidy_df %>%
  group_by(Race, Gender)
tw <- tw %>% select(-Movie)
tw %>%summarise(sum = sum(Words_spoken))
```

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

```
## # A tibble: 6 x 3
## # Groups:   Race [3]
##   Race   Gender   sum
##   <chr>  <chr>  <int>
## 1 Elf    Female  1743
## 2 Elf    Male    1994
## 3 Hobbit Female    16
## 4 Hobbit Male    8780
## 5 Man    Female   669
## 6 Man    Male    8043
```

```
tr <-tidy_df %>% group_by(Movie, Race)
tr %>%summarise(sum = 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     sum
##   <chr>                       <chr>  <int>
## 1 The Fellowship of the Ring Elf     2200
## 2 The Fellowship of the Ring Hobbit  3658
## 3 The Fellowship of the Ring Man     1995
## 4 The Return of the King     Elf      844
## 5 The Return of the King     Hobbit  2463
## 6 The Return of the King     Man     3990
## 7 The Two Towers             Elf      693
## 8 The Two Towers             Hobbit  2675
## 9 The Two Towers             Man     2727
```

2. Male Hobits :8780 Female Elves: 1743 Male Elves: 1994
3. Yes hobbits in the 1st, but then Man in the 2nd and 3rd movie.
4. Yes the dominant race depends on the movie.