hw-2_palanisamy-a_400233311

```
# read in file
file_path <- "C:/Users/Palan/OneDrive/Documents/School/McMaster/Phd/Courses/Fall/lord-of-t

lotr_data <- read.csv(file_path)

# Display the first few rows of the data
head(lotr_data)
```

```
                      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
```

```
library(tidyr)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Tidy the data: gather columns to long format
tidy_lotr_data <- lotr_data %>%
  pivot_longer(
    cols = elf_female:Man_male,
    names_to = "race_gender",
    values_to = "lines_spoken"
  ) %>%

  # Separate the race and gender into their own columns
  separate(race_gender, into = c("race", "gender"), sep = "_") %>%

  # Clean up capitalization inconsistencies
  mutate(
    race = tolower(race),
    gender = tolower(gender)
  )
```

```r
# Calculate total number of words spoken by male hobbits
male_hobbits_total <- tidy_lotr_data %>%
  filter(race == "hobbit", gender == "male") %>%
  summarise(total_lines_spoken = sum(lines_spoken))

# Calculate total number of words spoken by female elves
female_elves_total <- tidy_lotr_data %>%
  filter(race == "elf", gender == "female") %>%
  summarise(total_lines_spoken = sum(lines_spoken))

# Calculate total number of words spoken by male elves
male_elves_total <- tidy_lotr_data %>%
  filter(race == "elf", gender == "male") %>%
  summarise(total_lines_spoken = sum(lines_spoken))

# Display results
male_hobbits_total
```

```
# A tibble: 1 x 1
  total_lines_spoken
               <int>
1               8780
```

```r
female_elves_total
```

```
# A tibble: 1 x 1
  total_lines_spoken
               <int>
1               1743
```

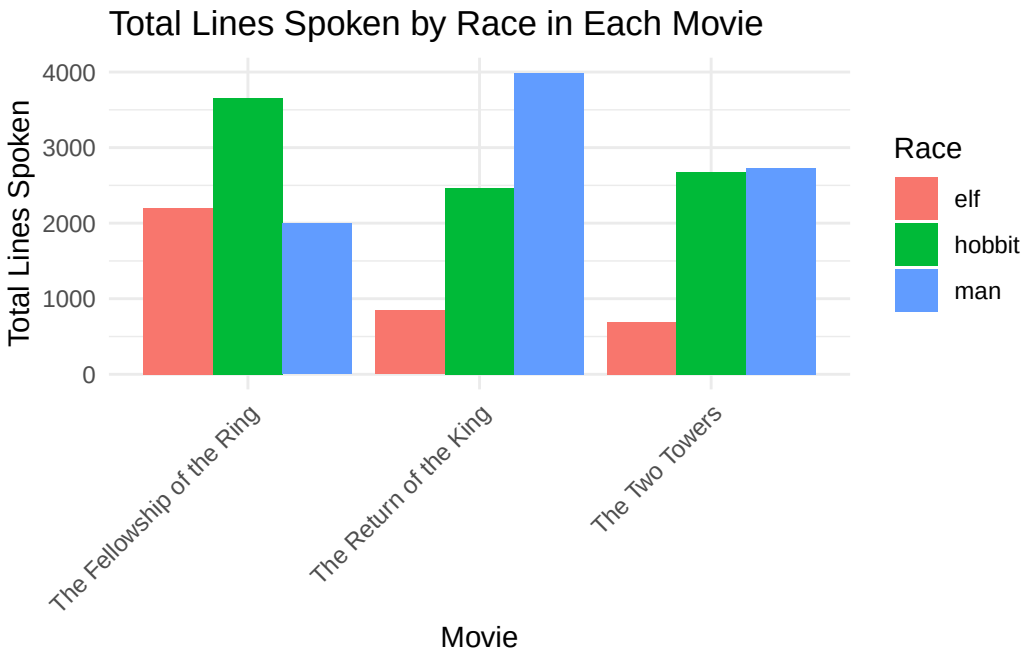
```
male_elves_total
```

```
# A tibble: 1 x 1
  total_lines_spoken
               <int>
1               1994
```

```
library(ggplot2)
# Summarize total lines spoken by race within each movie
race_summary <- tidy_lotr_data %>%
  group_by(movie, race) %>%
  summarise(total_lines_spoken = sum(lines_spoken))
```

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

```
# Create a bar plot
ggplot(race_summary, aes(x = movie, y = total_lines_spoken, fill = race)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Total Lines Spoken by Race in Each Movie",
    x = "Movie",
    y = "Total Lines Spoken",
    fill = "Race"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Part 1

2) a) Currently each column has two variables options race and gender (i.e., elf_male , man_male) but should be broken into two separate columns for increased clarity and analysis (race and gender)

b) Inconsistent capitialization of column names

c) Data would be more readable in long format as this would allow for four more clear column headers (move,race,gender,lines spoken)

3) In a tidy data set there would be 4 columns and 18 rows

4) Column names : movie, gender, race, lines_spoken

Part 2

2: a) male hobbits: 8780 b) female elves: 1743 c) male elves: 1994

3/4: The amount of words are dominated by a specific race but this is movie specific. In 2 out of three movies words are dominated by hobbits (Fellowship of the Ring), by man (The return of the king) and fairly even between hobbit and man in (the two towers)