Homework Assignment 2

Malaikah Ahmad

PART A

Question 1

```
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
```

```
library(readr)
setwd("~/Downloads")
data <- read.csv("lord-of-the-rings-trilogy.csv")
```

Question 2

This data set is not tidy for the following reasons:

(1) There are multiple variables in the column names (ex: elf_female, elf_male, Hobbit_female, hobbit_Male, man_Female, Man_male). This is combining both race and gender, while in a tidy data set each variable should have its own column.

(2) There is also inconsistent capitalization (Hobbit vs hobbit, Male vs male, and Female vs female) which could cause issues when attempting to manipulate the data set.

(3) This dataset is also in a wide format with different categories of the same variable spread across multiple columns (race and gender). For the set to be tidy, it should be in a longer format with each observation in its own row.

Question 3

If this data set was tidy, it would have 4 columns (movie, race, gender, words_spoken). Since there are 3 movies (The Fellowship of the Ring, The Two Towers, and the Return of the King), 3 races (elf, hobbit, man), and 2 genders (male, female) there will be 18 rows.

Question 4

The column names in tidy format would be:

(1) movie

(2) race

(3) gender

(4) words_spoken

PART B

Question 1: Tidy the Data Set

```
library(tidyr)
library(dplyr)

setwd("~/Downloads")
data <- read.csv("lord-of-the-rings-trilogy.csv")

names(data) <- tolower(names(data))

tidy_data <- data %>%
  pivot_longer(cols = -movie,
               names_to = c("race", "gender"),
               names_sep = "_",
```

```
               values_to = "words_spoken")
print(tidy_data)
```

```
# 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
```

Question 2a: Total Numbers of Words Spoken by Male Hobbits

```
male_hobbits <- tidy_data %>%
  filter(race == "hobbit", gender == "male") %>%
  summarize(total_words = sum(words_spoken, na.rm = TRUE))

male_hobbits
```

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

The total number of words spoken by male hobbits is 8780.

Question 2b: Total Numbers of Words Spoken by Female Elves

```
female_elves <- tidy_data %>%
  filter(race == "elf", gender == "female") %>%
  summarize(total_words = sum(words_spoken, na.rm = TRUE))

female_elves
```

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

The total number of words spoken by female elves is 1743.

Question 2c: Total Numbers of Words Spoken by Male Elves

```
male_elves <- tidy_data %>%
  filter(race == "elf", gender == "male") %>%
  summarize(total_words = sum(words_spoken, na.rm = TRUE))

male_elves
```

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

The total number of words spoken by male elves is 1994.

Question 3

```
race_dominance <- tidy_data %>%
  group_by(movie, race) %>%
  summarize(total_words = sum(words_spoken, na.rm = TRUE))
```

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

```
print(race_dominance)
```

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

Based on this output, the number of spoken words in a movie is dominated by a specific race.

Question 4

```
dominant_race_by_movie <- race_dominance %>%
  group_by(movie) %>%
  top_n(1, total_words)

print(dominant_race_by_movie)
```

```
# A tibble: 3 x 3
# Groups:   movie [3]
  movie                       race   total_words
  <chr>                       <chr>        <int>
1 The Fellowship of the Ring hobbit        3658
2 The Return of the King     man           3990
3 The Two Towers             man           2727
```

Yes, the dominant race does depend on the movie. Specifically, "The Fellowship of the Ring" is dominated by hobbits. While "The Return of the King" and "The Two Towers" are dominated by men.