Section 1: Description of the data.

Description: The dataset used in this analysis measures gender representation and diversity in the comic book industry. It includes data collected from Marvel Wikia and DC Wikia, capturing information such as character names, publishers, appearances, gender, alignment, and more. This dataset allows for investigating research questions related to gender dynamics in comic books, exploring patterns across different publishers, analyzing the relationships between gender and other character attributes, and assessing overall diversity within comic book narratives. The dataset is saved in a structured format as a CSV (Comma-Separated Values) file, which is commonly used for storing tabular data. CSV files are delimited, with commas serving as the delimiter to separate data values. This format enables easy integration with data analysis tools and programming languages like R.

```
# Reading and combining the data from two CSV files. Mainly using readr and dplyr here
library(readr)
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(knitr)

file1 <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-da
file2 <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/dc-wikia-data.

data1 <- read_csv(file1)
```

```
## Rows: 16376 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): name, urlslug, ID, ALIGN, EYE, HAIR, SEX, GSM, ALIVE, FIRST APPEAR...
## dbl  (3): page_id, APPEARANCES, Year
##
## 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.
```

```
data2 <- read_csv(file2)
```

```
## Rows: 6896 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): name, urlslug, ID, ALIGN, EYE, HAIR, SEX, GSM, ALIVE, FIRST APPEAR...
## dbl  (3): page_id, APPEARANCES, YEAR
##
## 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.
```

```
combined_data <- bind_rows(data1, data2)

# Cleaning function
clean_column_names <- function(data) {
  renamed_data <- data %>%
    rename(
      Gender = SEX,
      Alignment = ALIGN,
      `Mortality Status` = ALIVE,
      Name = name,
      `Total Appearances` = APPEARANCES
    )

  # Return cleaned data
  renamed_data
}

# Cleaning the data
clean_data <- combined_data %>%
  select(SEX, ALIGN, ALIVE, name, APPEARANCES) %>%
  clean_column_names()

# Output the sentence
cat("This dataframe has", nrow(clean_data), "rows and", ncol(clean_data), "columns.\n\n")
```

```
## This dataframe has 23272 rows and 5 columns.
```

```
# Creating the table
column_names <- c("Column Name", "Description")
column_desc <- c(
  "ALIGNMENT" = "Alignment of the character (good, bad, or neutral)",
  "SEX" = "Gender of the character",
  "MORTALITY STATUS" = "Indicates if the character is alive or deceased",
  "NAME" = "Name of the character",
  "TOTAL APPEARANCES" = "Total number of comic book appearances"
)
column_table <- data.frame(Column_Names = names(column_desc), Description = unname(column_desc))

# Output the table using kable
kable(column_table, format = "markdown")
```

Column_Names	Description
ALIGNMENT	Alignment of the character (good, bad, or neutral)
SEX	Gender of the character
MORTALITY STATUS	Indicates if the character is alive or deceased
NAME	Name of the character
TOTAL APPEARANCES	Total number of comic book appearances

```
# Reading and combining the data from two CSV files. Mainly using readr and dplyr here
library(readr)
library(dplyr)

file1 <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/marvel-wikia-da
file2 <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/comic-characters/dc-wikia-data.c

data1 <- read_csv(file1)
```

```
## Rows: 16376 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): name, urlslug, ID, ALIGN, EYE, HAIR, SEX, GSM, ALIVE, FIRST APPEAR...
## dbl  (3): page_id, APPEARANCES, Year
##
## 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.
```

```
data2 <- read_csv(file2)
```

```
## Rows: 6896 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): name, urlslug, ID, ALIGN, EYE, HAIR, SEX, GSM, ALIVE, FIRST APPEAR...
## dbl  (3): page_id, APPEARANCES, YEAR
##
## 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.
```

```
combined_data <- bind_rows(data1, data2)

# Selecting three columns
selected_columns <- combined_data %>%
  select(SEX, APPEARANCES, ALIGN)

# Calculate summary statistics
summary_stats <- combined_data %>%
  select(SEX, APPEARANCES, ALIGN) %>%
  summarize(
    Minimum_SEX = min(SEX, na.rm = TRUE),
    Maximum_SEX = max(SEX, na.rm = TRUE),
    Mean_APPEARANCES = mean(APPEARANCES, na.rm = TRUE),
    Num_Missing_ALIGN = sum(is.na(ALIGN))
  )
```

```r
# Create a data frame to display the summary results
summary_df <- data.frame(
  Column = c("Minimum_SEX", "Maximum_SEX", "Mean_APPEARANCES", "Num_Missing_ALIGN"),
  Summary = as.character(summary_stats),
  stringsAsFactors = FALSE
)

# Display the summary table
knitr::kable(summary_df, format = "markdown")
```

Column	Summary
Minimum_SEX	Agender Characters
Maximum_SEX	Transgender Characters
Mean_APPEARANCES	19.0093029650321
Num_Missing_ALIGN	3413