ANA 515 Assignment 02

Rashmitha Pasupureddy

2022-07-04

Description

This dataset consists of script from all Lord of the Rings Trilogy - Fellowship of the Rings, Two Towers and Return of the King. This dataset has 2390 observations for 4 variables. It was collected from https://www.kaggle.com/datasets/paultimothymooney/lord-of-the-rings-data. The dataset decribes the script for each character and which movie it is from. The dataset is a delimited flat file while the delimiter is a space

Loading Packages

```
    library(tidyverse)
    library(dplyr)
    library(knitr)
    library(bslib)
    library(readr)
    library(stringr)
    library(DT)
```

Reading the data with read.csv function from the package readr

```
        lotr <- read.csv("C:/Users/rashm/Desktop/ANA 515 Fundamentals of Data Storage/LOTR/lotr_scripts.csv")
```

Cleaning/Pre-Processing Data

```
    lotr_df <- lotr %>%
        rename(character_name = char) #Renaming the column char to character_name
        str_to_title(lotr_df$character_name) #Converting the character names to title case
        trimws(lotr_df$dialog) #I have noticed some of the dialogues have a leading whitespace. So, we're trimming

        ## I have hidden the output for this code chunk as it produces a large amount of data. Below is another chu
```

```
        datatable(lotr_df, rownames=TRUE, filter="top", options=list(pageLength=6, scrollX=T)) #This is reduce the
```

Show 6 entries   Search:

	X	character_name	dialog	movie
	All	All	All	All
1	0	DEAGOL	Oh Smeagol Ive got one! , Ive got a fish Smeagol, Smeagol!	The Return of the King
2	1	SMEAGOL	Pull it in! Go on, go on, go on, pull it in!	The Return of the King
3	2	DEAGOL	Arrghh!	The Return of the King
4	3	SMEAGOL	Deagol!	The Return of the King
5	4	SMEAGOL	Deagol!	The Return of the King
6	5	SMEAGOL	Deagol!	The Return of the King

Showing 1 to 6 of 2,390 entries   Previous 1 2 3 4 5 … 399 Next

Characteristics of Data

```
        observations <- nrow(lotr_df)
        variables <- ncol(lotr_df)
```

This data frame has 2390 rows and 4 columns. The names of the columns and a brief description of each are in the table below:

Table

```
        kable(str(lotr_df))
```

```
## 'data.frame':    2390 obs. of  4 variables:
##  $ X             : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ character_name: chr  "DEAGOL" "SMEAGOL" "DEAGOL" "SMEAGOL" ...
##  $ dialog        : chr  "Oh Smeagol Ive got one! , Ive got a fish Smeagol, Smeagol!    " "Pull it in! Go on, go
##  $ movie         : chr  "The Return of the King " "The Return of the King " "The Return of the King " "The Retur
```

|| || || ||

Summary Statistics

```
#Since my data set only has one variable with numeric values, I chose to apply the required functions to it

    FrodoFilter <- filter(lotr_df, character_name=="FRODO")
    FrodoMin <- min(FrodoFilter$X)
    FrodoMax <- max(FrodoFilter$X)
    FrodoMean <- mean(FrodoFilter$X)
    FrodoMV <- colSums(is.na(FrodoFilter))
    summary(FrodoFilter) #Using summary function
```

```
##        X        character_name        dialog             movie
##  Min.   :  16   Length:225         Length:225         Length:225
##  1st Qu.: 574   Class :character   Class :character   Class :character
##  Median :1356   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1255
##  3rd Qu.:1922
##  Max.   :2337
```

Saving summary stats

```
    FrodoSummary <- data.frame(FrodoMin, FrodoMax, FrodoMean, FrodoMV)
    print(FrodoSummary)
```

```
##                FrodoMin FrodoMax FrodoMean FrodoMV
## X                    16     2337  1254.516       0
## character_name       16     2337  1254.516       0
## dialog               16     2337  1254.516       0
## movie                16     2337  1254.516       0
```