ANA 515 Assignment 2

Sparsh Kathuria

2022-06-19

# Install libraries  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.5 ✔ purrr 0.3.4  
## ✔ tibble 3.1.6 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(knitr)  
library(rmarkdown)

This data I chose includes movie ratings on Fandango and the data was pulled on August 24, 2015. The data is in the form of csv and contains every film that has a Rotten Tomatoes rating, a RT User rating, a Metacritic score, a Metacritic User score, and IMDb score, and at least 30 fan reviews on Fandango. It has information about both Rotten Tomatoes score and Fandango rating, to provide a good comparison of ratings between the two and determine if the ratings at Fandango are suspicious or not. The dataset has 147 rows and 21 columns. If we think about it, the initial few weeks of the movies being released, the change in IMDB score or rotten tomatoes would be significant, maybe once every hour however as the movie becomes older, there is not a lot of changes in it. The data being pulled is in a .csv format.It is delimited by comma.

Reading Data into R

#Import data from url  
url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/fandango/fandango\_scrape.csv"  
df <- read\_csv(url)

## Rows: 510 Columns: 4  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): FILM  
## dbl (3): STARS, RATING, VOTES  
##   
## 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.

print(colnames(df))

## [1] "FILM" "STARS" "RATING" "VOTES"

Cleaning the Data

# Filter data for votes greater than 10000  
highvotes<-filter(df, VOTES >= 10000)

Characterstics of the data

#number of rows in df  
df\_nrow <- nrow(highvotes)  
df\_ncol <- ncol(highvotes)

This dataframe has 13 rows and 4 columns. The names of the columns and a brief description of each are in the table below:

# declare empty dataframe  
desc <- data.frame(matrix(ncol = 2, nrow = 4))  
  
# get column names  
cols<- c("FILM" , "STARS" , "RATING", "VOTES" )  
  
# create a variable for column descriptions  
desc\_cols <- c('Name of the movie', 'Number of stars presented on Fandango.com' , 'Actual average score the movie obtained' ,'number of people who had reviewed the film')  
  
# add columns to dataframe desc  
desc$colname <- cols  
desc$desc\_cols <- desc\_cols  
  
# create a table using kable from knitr  
knitr::kable(desc, "simple")

| X1 | X2 | colname | desc\_cols |
| --- | --- | --- | --- |
| NA | NA | FILM | Name of the movie |
| NA | NA | STARS | Number of stars presented on Fandango.com |
| NA | NA | RATING | Actual average score the movie obtained |
| NA | NA | VOTES | number of people who had reviewed the film |

Summary Statistics

# get three columns from dataframe  
df\_new <- highvotes[ , c("FILM", "RATING", "VOTES")]   
  
#summary statistics  
summary(df\_new)

## FILM RATING VOTES   
## Length:13 Min. :3.400 Min. :10509   
## Class :character 1st Qu.:4.200 1st Qu.:13055   
## Mode :character Median :4.500 Median :15205   
## Mean :4.323 Mean :20096   
## 3rd Qu.:4.500 3rd Qu.:33538   
## Max. :4.800 Max. :34846

summary\_df <- summary(df\_new)  
  
print(summary\_df)

## FILM RATING VOTES   
## Length:13 Min. :3.400 Min. :10509   
## Class :character 1st Qu.:4.200 1st Qu.:13055   
## Mode :character Median :4.500 Median :15205   
## Mean :4.323 Mean :20096   
## 3rd Qu.:4.500 3rd Qu.:33538   
## Max. :4.800 Max. :34846