Assignment - Week 4

Humpreet Sahans

2024-06-14

## #Section 1: Description of the data

This assignment uses the Biopics dataset, which is sourced from IMDB and publicly available on Github via FiveThirtyEight.

It is a comprehensive collection of information about 761 biographical movies. Each row contains a distinct film title, a URL to its IMDB page, the country of origin, release dates, actors cast, gross earnings at the US box office and information about the subjects on which the films are based. The data is stored in a CSV file i.e. structured as plain text with comma delimiters.

The purpose of collecting this data was to analyze trends and patterns in the production of biographical films, especially in terms of diversity and representation. Researchers can use this data to explore the racial and gender diversity among biopic subjects, examine the correlation between box office earnings and the subject’s characteristics, or address other related research questions.

## #Section 2: Reading the data

#Reading data into a table from the csv file using the function 'read.csv'. This is a base R function that is used to read comma-separated values. There were no empty rows in our dataset and all 761 entries were read.  
  
url <- "https://github.com/fivethirtyeight/data/raw/master/biopics/biopics.csv"  
biopics <- read.csv(url)

## #Section 3: Cleaning the data

#Renaming the data for consistency and better clarity  
re\_biopics <- biopics %>%  
 rename(  
 "year\_of\_release" = year\_release,  
 "subject\_gender" = subject\_sex,  
 "name\_of\_lead\_actor" = lead\_actor\_actress  
 ) %>%  
  
#Filtering the renamed data table by removing columns that don't add much value  
 select(-site, -race\_known) %>%  
  
#Replacing the '-' with '0' in the column 'box\_office' & Converting currency data to numeric values by removing alphabets; the values are in Millions  
 mutate( box\_office = ifelse(box\_office == '-', "0", box\_office),  
 box\_office = parse\_number(box\_office))

## #Section 4: Characteristics of the (filtered) data

This dataframe has 761 rows and 12 columns. The names of the columns and a brief description of each are in the table below:

columns\_summary <- data.frame(  
 Columns = c(colnames(re\_biopics)),  
 Description = c("Title of the film", "Country of origin", "Year of release", "Gross earnings at U.S. box office", "Director of the film", "Number of subjects featured in the film", "Actual name of the featured subject", "The Occupation or reason for the recognition of the subject","Race of the subject", "Dummy variable that indicates person of color", "Gender of the subject", "Name of the lead actor or actress who played the subject")  
)  
  
kable(columns\_summary, caption = "Biopics Variables & Definitions")

Biopics Variables & Definitions

| Columns | Description |
| --- | --- |
| title | Title of the film |
| country | Country of origin |
| year\_of\_release | Year of release |
| box\_office | Gross earnings at U.S. box office |
| director | Director of the film |
| number\_of\_subjects | Number of subjects featured in the film |
| subject | Actual name of the featured subject |
| type\_of\_subject | The Occupation or reason for the recognition of the subject |
| subject\_race | Race of the subject |
| person\_of\_color | Dummy variable that indicates person of color |
| subject\_gender | Gender of the subject |
| name\_of\_lead\_actor | Name of the lead actor or actress who played the subject |

## #Section 5: Summary Statistics

#Selecting three numeric columns to summarize  
data\_pick3 <- select(re\_biopics, box\_office, number\_of\_subjects, person\_of\_color)  
  
#Creating summary statistics and Printing it  
summary\_stats <- summary(data\_pick3)  
print(summary\_stats)

## box\_office number\_of\_subjects person\_of\_color   
## Min. : 0.00 Min. :1.000 Min. :0.0000   
## 1st Qu.: 0.00 1st Qu.:1.000 1st Qu.:0.0000   
## Median : 2.42 Median :1.000 Median :0.0000   
## Mean : 45.60 Mean :1.268 Mean :0.1314   
## 3rd Qu.: 30.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :872.00 Max. :4.000 Max. :1.0000

#Extra  
#Creating a new table with the extracted summary statistics & Printing it  
  
sum\_table <- data.frame(  
   
 Statistic = c("Minimum Value", "Maximum Value", "Mean Value", "No. of Missing Values"),  
   
 Box\_Office = c(  
 min(data\_pick3$box\_office), max(data\_pick3$box\_office), mean(data\_pick3$box\_office), sum(is.na(data\_pick3$box\_office))),  
   
No\_of\_subjects = c(  
 min(data\_pick3$number\_of\_subjects), max(data\_pick3$number\_of\_subjects), mean(data\_pick3$number\_of\_subjects), sum(is.na(data\_pick3$number\_of\_subjects))),  
  
Person\_of\_color = c(  
 min(data\_pick3$person\_of\_color), max(data\_pick3$person\_of\_color), mean(data\_pick3$person\_of\_color), sum(is.na(data\_pick3$person\_of\_color)))  
  
 )  
   
print(sum\_table)

## Statistic Box\_Office No\_of\_subjects Person\_of\_color  
## 1 Minimum Value 0.00000 1.000000 0.000000  
## 2 Maximum Value 872.00000 4.000000 1.000000  
## 3 Mean Value 45.59938 1.268068 0.131406  
## 4 No. of Missing Values 0.00000 0.000000 0.000000