20240907_Robbins_ProblemSet1

Caroline Robbins

2024-09-07

Set working directory and load required packages.

```
##setwd("C:/Users/crobb/Downloads")
##install.packages("tinytex")
##install.packages("knitr")
##install.packages("rmarkdown")
library(knitr)
library(rmarkdown)
library(tinytex)
library(tidyverse)
```

Load the data kdrama.csv into the R environment.

```
raw_data <- read.csv("C:/Users/crobb/Downloads/kdrama.csv")
```

1. Download dataset kdrama.csv either directly here or from our shared dropbox folder. This is a dataset with top-ranked Korean Dramas as per the MyDramaList website downloaded from Kaggle. How many Korean Dramas are included in this list? Write an R code to respond to this question.

```
dim(raw_data)
```

```
## [1] 250  17
```

```
nrow(raw_data)
```

```
## [1] 250
```

2. Use an R code to return a list of variables that are included in this dataset.

```
names(raw_data)
```

```
##  [1] "Name"                 "Aired.Date"           "Year.of.release"      "Original.Network"
##  [5] "Aired.On"             "Number.of.Episodes"   "Duration"             "Content.Rating"
##  [9] "Rating"               "Synopsis"             "Genre"                "Tags"
## [13] "Director"             "Screenwriter"         "Cast"                 "Production.companies"
## [17] "Rank"
```

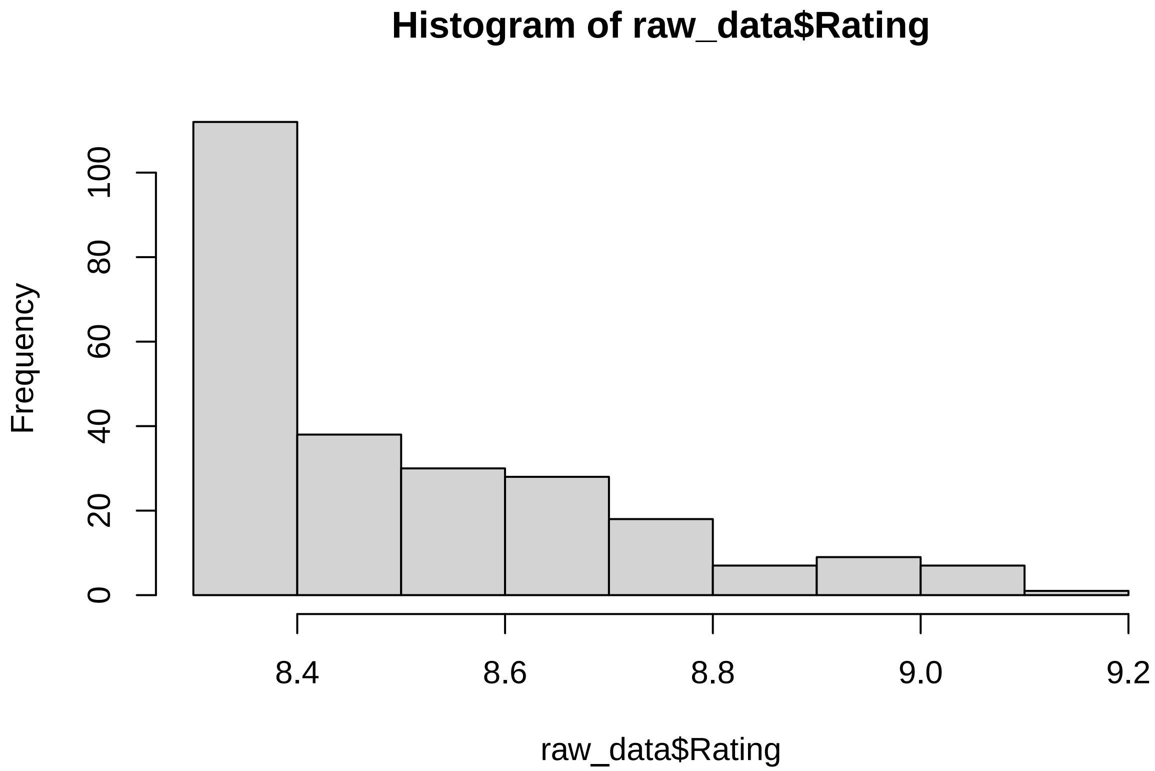
3. Use an R code to identify what the mean value of total number of episodes for all the kdramas in the dataset is.

```
mean(raw_data$Number.of.Episodes)
```

```
## [1] 19.064
```

4. Plot a histogram of the shows rating.

```
hist(raw_data$Rating)
```



5. How many shows have a rating higher than 9 points? Show how you calculated that in R.

```
table(raw_data$Rating > 9)
```

```
##
## FALSE  TRUE
##   242     8
```

6. Rename variable Year.of.release to simply Year without creating a new variable.

```
colnames(raw_data)[3] = "Year"
```

7. How many shows in this dataset were released in 2020-2022? Demonstrate with an R code.

```
table(raw_data$Year <= 2022 &
      raw_data$Year >= 2020)
```

```
##
## FALSE  TRUE
##   144   106
```

8. What type is the variable Duration? Show with an R code.

```
class(raw_data$Duration)
```

```
## [1] "character"
```

9. Recode variable Duration so that it would be a numerical variable measuring duration in minutes. Plot the histogram of the recoded variable.

```r
##Remove the character values of hr. and min. from the variable
split_duration <- data.frame(do.call("rbind", strsplit(as.character
                                                       (raw_data$Duration),
                                                       c("hr. | min."))))

##Convert character numbers into numeric form to combine later
numeric_data <- split_duration %>%
  mutate_if(is.character, as.numeric)

##Convert the hours variable so that it contains the minute value (60)
##and set all others equal to 0 so only the minute value in X2 remains.
only_minutes <- data.frame((X1 = case_when(numeric_data$X1 == 1 ~ 60,
                                           numeric_data$X1 ==
                                             numeric_data$X2 ~ 0,
                                TRUE ~ numeric_data$X2)),
                           (X2 = numeric_data$X2))
##Check the column names for the next steps
colnames(only_minutes)
```

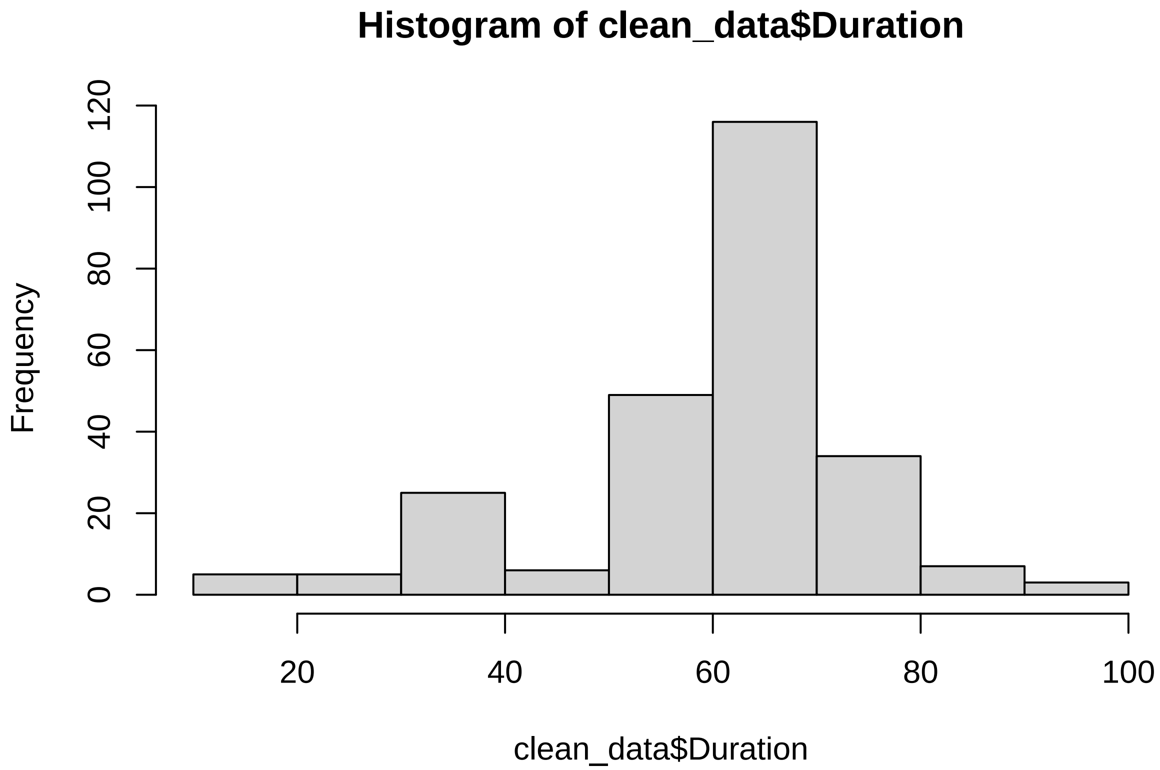
```
## [1] "X.X1...case_when.numeric_data.X1....1...60..numeric_data.X1...."
## [2] "X.X2...numeric_data.X2."
```

```r
##Fix the errors in the column names
almost_minutes <- only_minutes %>%
  rename(hours = X.X1...case_when.numeric_data.X1....1...60..numeric_data.X1....
         ) %>%
  rename(minutes = X.X2...numeric_data.X2.)

##Combine hours and minutes to have a pure minute value
duration_minutes <- almost_minutes %>%
  mutate(MinDuration = hours + minutes)
duration_minutes

##Add the new duration variable to the old Duration variable in the dataset
##and remove the other variables that have no purpose now
new_data <- data.frame(raw_data, duration_minutes)
new_data
new_data$hours <- NULL
new_data$minutes <- NULL
clean_data <- new_data %>%
  mutate(Duration = MinDuration)
clean_data$MinDuration <- NULL

clean_data
hist(clean_data$Duration)
```



10. Create a new dataset that will include shows with Original.Network being Netflix.

```
#Shows only showing on Netflix exclusively
netflix_only1 <- filter(clean_data, Original.Network == "Netflix")
#Shows on Netflix and other networks
netflix_only2 <- filter(clean_data, str_detect(Original.Network, "Netflix"))
```

11. What is the average rating score for the shows that have Netflix as an Original Network.

```
mean(netflix_only1$Rating)
```

```
## [1] 8.65
```

```
mean(netflix_only2$Rating)
```

```
## [1] 8.6625
```