# Problem Set 1

Shaka Y.J. Li

2024-08-28

```
###Load the data we need
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.4      v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

```
###Load the data
data <- read_csv("/Users/shakali/FSU work/POS5737/MethodsII/kdrama.csv")
```

```
## Rows: 250 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): Name, Aired Date, Original Network, Aired On, Duration, Content Ra...
## dbl  (3): Year of release, Number of Episodes, Rating
##
## 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.
```

```
###Q1
nrow(data)
```

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

```
###Q2
colnames(data)
```

```
##  [1] "Name"                "Aired Date"          "Year of release"
##  [4] "Original Network"    "Aired On"            "Number of Episodes"
##  [7] "Duration"            "Content Rating"      "Rating"
## [10] "Synopsis"            "Genre"               "Tags"
## [13] "Director"            "Screenwriter"        "Cast"
## [16] "Production companies" "Rank"
```

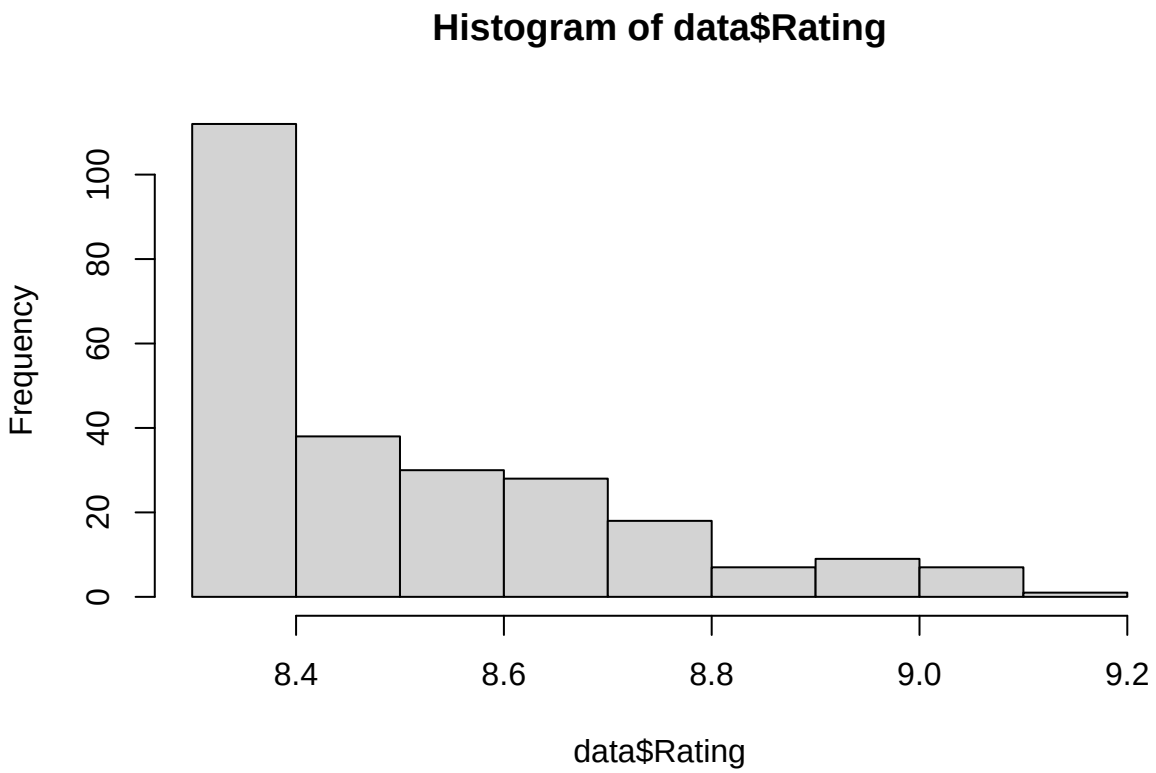
```
###Q3
sum(is.na(data$`Number of Episodes`))###Checking if there are any missing values
```

```
## [1] 0
```

```r
mean(data$`Number of Episodes`)###Claculate the mean
```

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

```r
###Q4
hist(data$Rating)
```




```r
###Q5
data_over9 <- ###Subset the rating value whose value is larger than 9
  data %>%
  group_by(Name) %>%
  subset(Rating > 9 )

nrow(data_over9) ###Calculate how many of them
```

```
## [1] 8
```

```r
###Q6
data <- data %>%
  rename("Year" = "Year of release", ###Rename the original one to simply Year
         "original_network" = "Original Network") ###Rename the variable I need
```

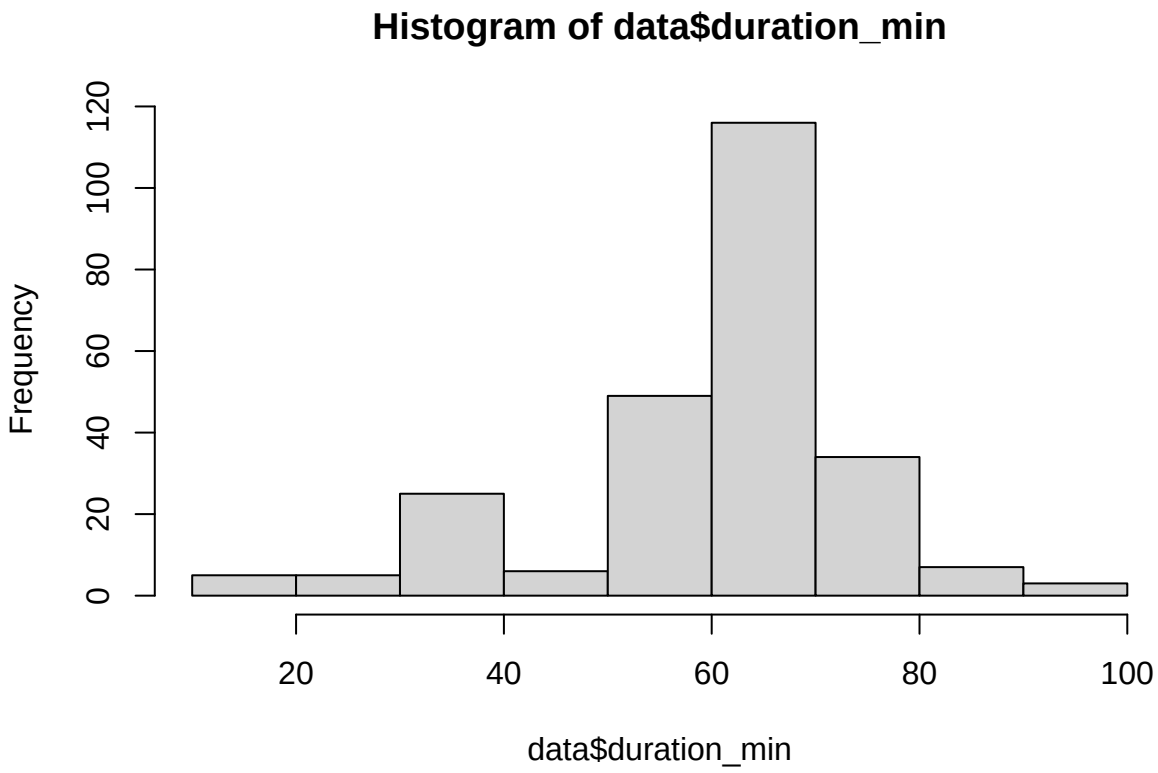
```r
###Q7
data_2020_2022 <- data %>% ###Subset the data from 2020 to 2022 first
  filter(Year %in% c(2020:2022)) %>%
  group_by(Name) %>%
  arrange(Year)

nrow(data_2020_2022) ###Calculate the amount of the dataset
```

```
## [1] 106
```

```
###Q8
class(data$Duration)
```

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

```
###Q9
data <- data %>%
  mutate(duration_min = gsub("hr.", "*60 +", Duration) %>% #Create a new column called minutes and mult
                        gsub("min.", "", .) %>% #delete "minutes" and add with previous results
                        gsub("^\\s+", "", .) %>% # removing any remaining leading space in the string
                        sapply(function(x) eval(parse(text = x))) %>% #converting to numeric
                        unlist())
hist(data$duration_min) ###create histogram
```




```
###Q10
data_netflix <- data %>% filter(str_detect(original_network, "Netflix"))
```

```
###Q11
mean(data_netflix$Rating)
```

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