PS1: K-Drama Dataset Analysis

Barna Bose

September 7, 2025

Dataset

Source: Kaggle / MyDramaList top-ranked Korean dramas. **File:** `kdrama.csv`

Summary of Key Results

Total dramas (#rows)	**250**
Variables (names)	17 (see Q2)
Mean # episodes	**19.064**
# ratings > 9	**8**
Years 2020–2022 (inclusive)	**106**
Type of `Duration`	`character`
Netflix shows (exact match)	**12**
Avg. rating (Netflix exact)	**8.65**

Q1. How many Korean dramas are included?

```
library(readr)
kdrama <- read_csv("C:/Users/barna/Downloads/kdrama.csv")
nrow(kdrama) # -> 250
```

Answer: 250

Q2. List the variables in the dataset

```
names(kdrama)
```

Variables:

1. Name
2. Aired Date
3. Year of release
4. Original Network
5. Aired On

6. Number of Episodes
7. Duration
8. Content Rating
9. Rating
10. Synopsis
11. Genre
12. Tags
13. Director
14. Screenwriter
15. Cast
16. Production companies
17. Rank

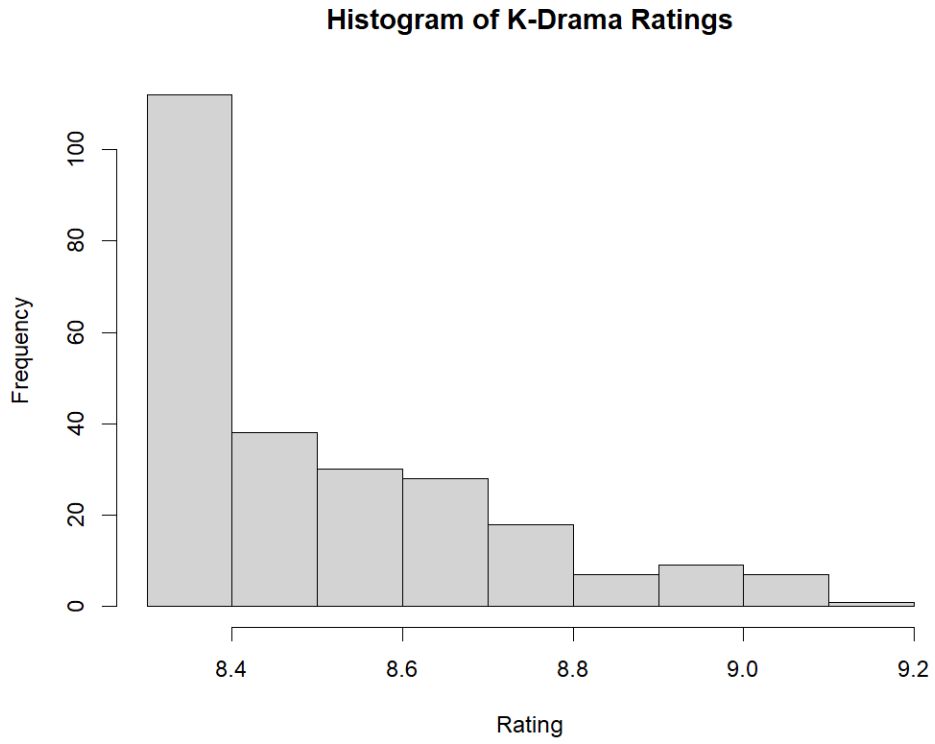
Q3. Mean total number of episodes

```
mean_episodes <- mean(na.omit(kdrama$`Number of Episodes`))
mean_episodes # -> 19.064
```

Answer: 19.064

Q4. Histogram of ratings

```
hist(kdrama$Rating,
     main = "Histogram of K-Drama Ratings",
     xlab = "Rating")
```



Q5. How many shows have rating > 9?

```
sum(kdrama$Rating > 9, na.rm = TRUE) # -> 8
```

Answer: 8

Q6. Rename `Year of release` to `Year` (in place)

```
library(dplyr)
kdrama <- kdrama %>% rename(Year = `Year of release`)

# Base-R alternative (3rd column):
# colnames(kdrama)[3] <- "Year"
```

Q7. How many shows were released in 2020–2022?

```
sum(na.omit(kdrama$Year) >= 2020 & na.omit(kdrama$Year) <= 2022) # -> 106
```

Answer: 106

Q8. Type of `Duration`

```
class(kdrama$Duration) # -> "character"
```

Answer: `character`

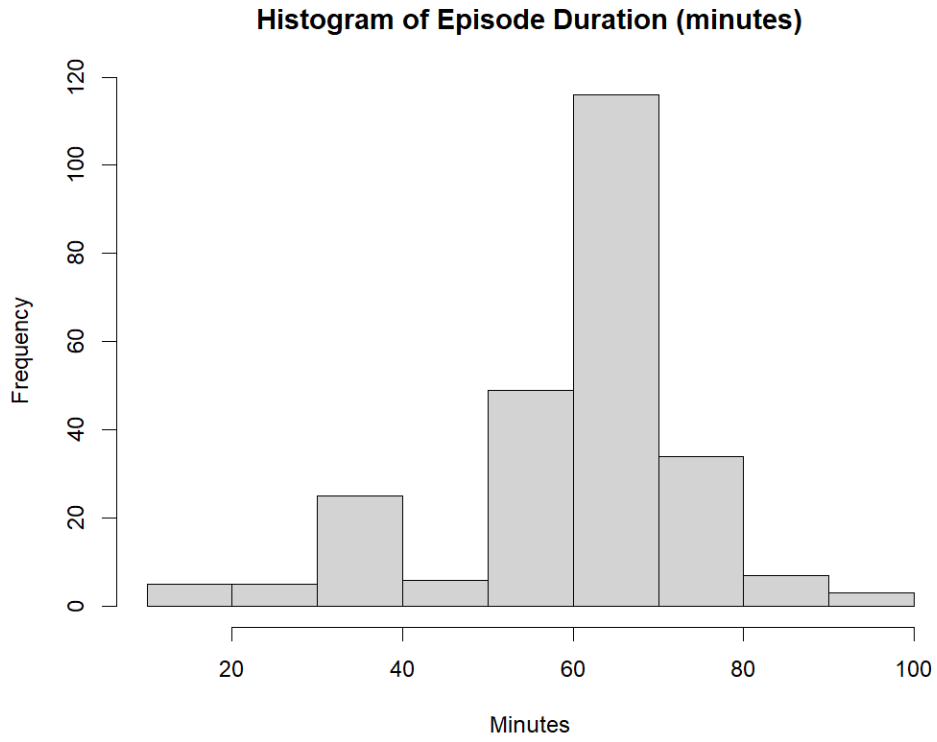
Q9. Recode `Duration` to minutes; plot histogram

Tidyverse (hours + minutes robust extraction):

```
library(dplyr)
library(stringr)

kdrama <- kdrama %>%
  mutate(Duration_min =
           coalesce(as.numeric(str_extract(Duration, "\\d+(?=\\s*hr)")), 0) * 60 +
           coalesce(as.numeric(str_extract(Duration, "\\d+(?=\\s*min)")), 0))

hist(na.omit(kdrama$Duration_min),
     main = "Histogram of Episode Duration (minutes)",
     xlab = "Minutes")
```




```
Duration_to_minutes <- function(x) {
  hrs <- ifelse(grepl("hr", x, ignore.case=TRUE),
                as.numeric(sub(".*?(\\d+)\\s*hr.*", "\\1", x)), 0)
  mins <- ifelse(grepl("min", x, ignore.case=TRUE),
                as.numeric(sub(".*?(\\d+)\\s*min.*", "\\1", x)), 0)
```

```
  hrs[is.na(hrs)] <- 0
  mins[is.na(mins)] <- 0
  60*hrs + mins
}
```

Q10. Subset to Netflix (exact match)

```
library(dplyr)
library(stringr)

netflix_exact <- kdrama %>%
  filter(str_trim(`Original Network`) == "Netflix")

nrow(netflix_exact) # -> 12
```

Answer: 12 shows.

Q11. Average rating for Netflix originals

```
mean(na.omit(netflix_exact$Rating)) # -> 8.65
# or:
mean(netflix_exact$Rating, na.rm = TRUE) # -> 8.65
```

Answer: 8.65
