Problem Set 1

Bryson L

2024-08-29

Rows

For dataset *kdrama*, $n = 250$.

```
pander:: pander(count(kdrama))
```

n
250

Variable Names

```
colnames(kdrama)
```

```
##  [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"
```

Mean Number of Episodes

The mean number of episodes per K-Drama in the dataset is roughly 19.

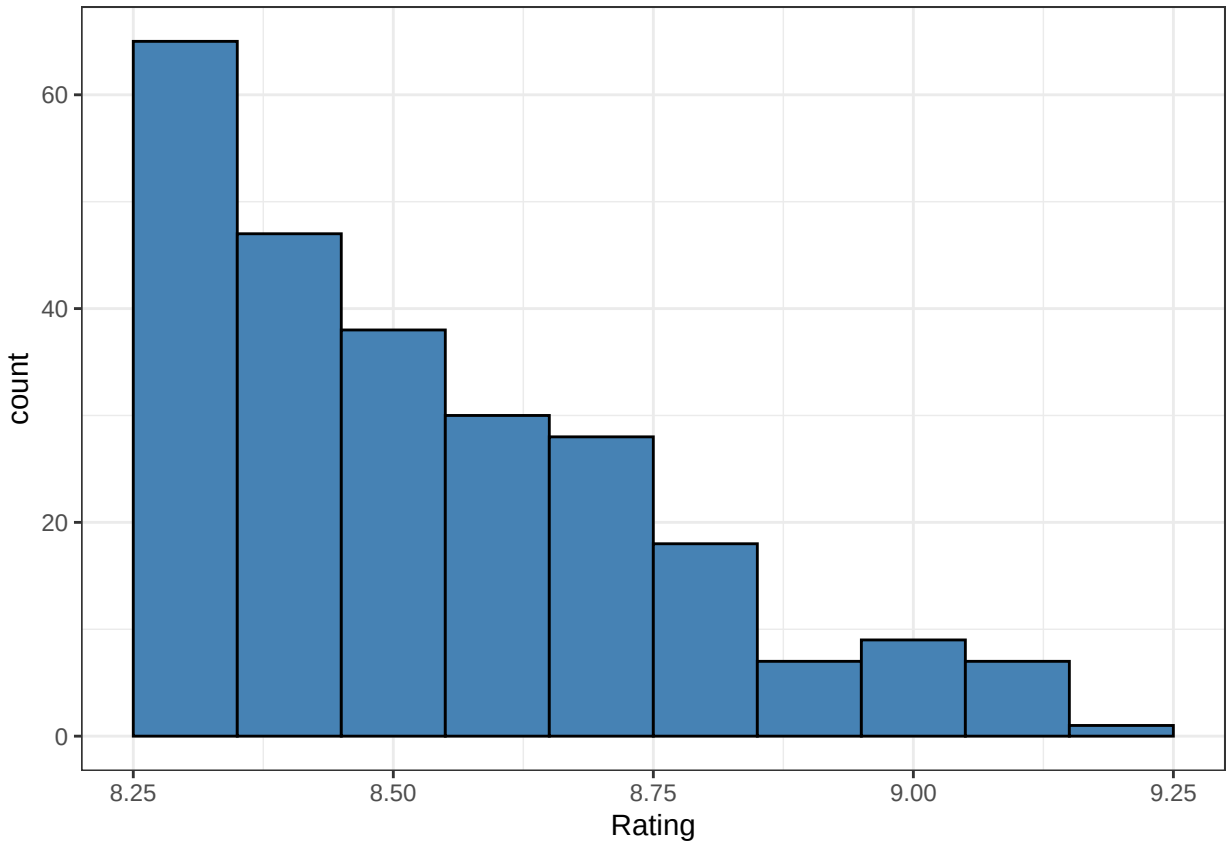
```
mean(kdrama$Number.of.Episodes)
```

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

Show Rating Histogram

The binwidth for the histogram is 0.1.

```
ggplot(kdrama, aes(x=Rating)) +
  geom_histogram(fill='steelblue', color='black', binwidth = 0.10) +
  theme_bw()
```



Rating above 9

17 of the shows in the dataset had a rating of 9 or above. I created a binary variable to evaluate this, with a 1 denoting a rating of 9 or above.

```
kdrama<- kdrama %>%
  mutate(above9 = if_else(Rating >= 9.0, 1, 0))

pander::pander(table(kdrama$above9))
```

0	1
233	17

Year Rename

```
names(kdrama)[names(kdrama) == "Year.of.release"] <- "Year"
```

Released from 2020-2022

106 of the shows in the dataset were released between 2020 and 2022. To do this I created another binary variable, with 1 denoting a “recent release”, or from 2020-2022.

```
kdrama<- kdrama %>%
  mutate(recent.release = if_else(Year >= 2020, 1, 0))

pander::pander(table(kdrama$recent.release))
```

0	1
144	106

Duration

The "*Duration*" variable in the dataset is a character variable.

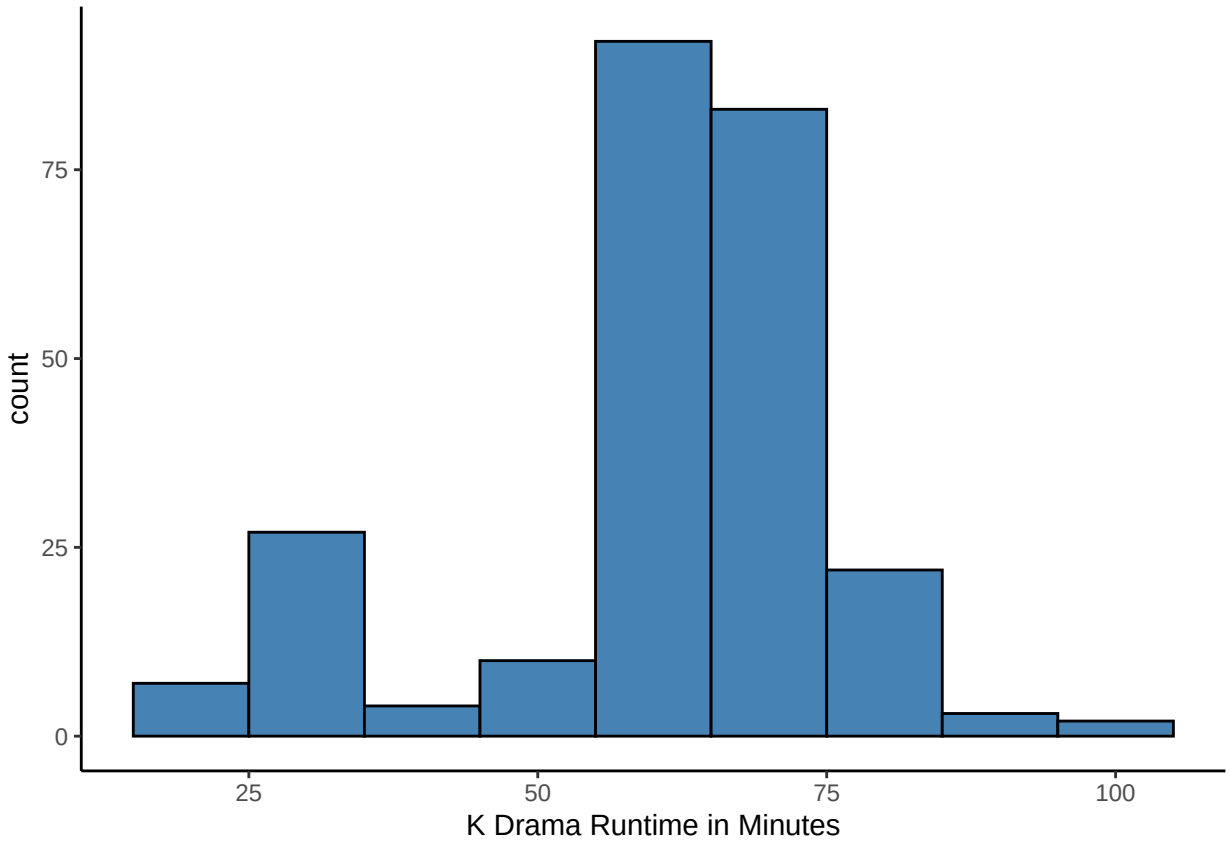
```
class(kdrama$Duration)
```

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

Recoding Duration

To recode the variable, I had to split the strings, which I accomplished with *strsplit*(). The histogram depicting show duration by minutes is below. The binwidth for the histogram is 10 minutes.

```
kdrama_split<- data.frame(do.call("rbind", strsplit(kdrama$Duration, "hr. | min.")))

kdrama_split<- kdrama_split %>%
  mutate(hour = as.numeric(X1))

kdrama_split<- kdrama_split %>%
  mutate(minutes = as.numeric(X2))

kdrama_hour<- kdrama_split %>%
  mutate("hour" = if_else(hour==1, 1,0))

kdrama_minute<- kdrama_hour %>%
  mutate("min.duration" = hour*60 + minutes)

kdrama_minute_duration<- select(kdrama_minute, c(min.duration))

kdrama_min<- data.frame(kdrama, kdrama_minute_duration)

ggplot2::ggplot(kdrama_min, aes(x=min.duration)) +
         geom_histogram(binwidth=10,fill="steelblue", color = "black")+
  labs(x="K Drama Runtime in Minutes")+
         theme_classic()
```



Netflix Original

I created a subset of the data with all shows that included Netflix as an original network, so this subset includes shows that only appeared on Netflix and shows that listed Netflix as one of their multiple original networks.

```
kdrama_netflix<- filter(kdrama_min, grepl("Netflix", Original.Network))
```

Netflix Mean Rating

Shows that included Netflix as an original network had an average rating of about 8.7

```
mean(kdrama_netflix$Rating)
```

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