Data Literacy & Visualization

Student’s Name

Course Number- Name of Course

Instructor’s Name

Date

Data Literacy & Visualization

# 2.1.1 Import required libraries and data

library(tidyverse)

library(ggplot2)

adventure = read.csv("adventure-csv.csv")

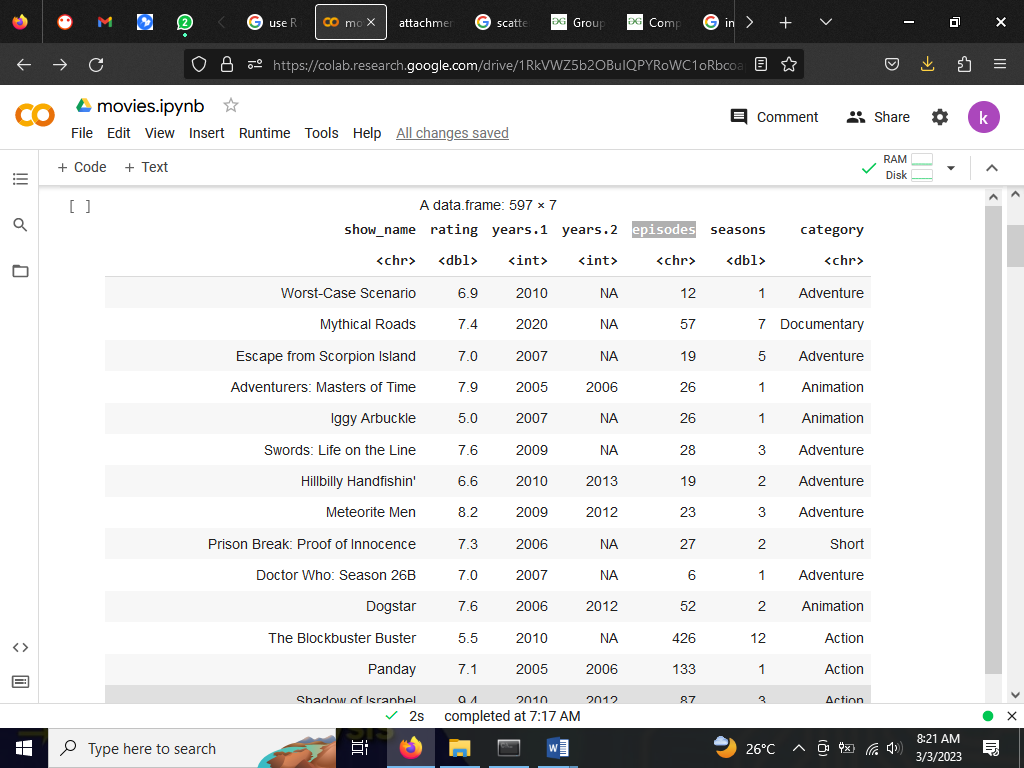
crime = read.csv("crime3-csv-csv.csv")

comedy = read.csv("commedy-csv-(1)-csv.csv")

names(adventure) = names(crime) = names(comedy)

tv\_data = rbind(adventure,crime,comedy)

**Figure 1: Data Snippet**



# 2.1.2 Filter out duplicates and drop missing values

filtered = unique(tv\_data)

filtered=na.omit(filtered)

# 2.1.3 Convert episodes to numeric log transform it and store in a new column

filtered=transform(filtered, episodes=as.numeric(episodes))

filtered$log\_episodes = log(filtered$episodes)

# 2.1.4 Create subset of the data when a show appears for the first time in the data

sub\_data = filtered[!duplicated(filtered$show\_name),]

# 2.2.1 Summary statistics

summary(sub\_data$rating)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.600 6.300 7.300 7.025 7.900 9.400

summary(sub\_data$episodes)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 25.00 43.00 71.54 75.75 2158.00

The average rating of a show is 7.025 and the average number of episodes is 72.

# 2.2.2 Correlation test

rating = sub\_data$rating

episodes = sub\_data$episodes

seasons = sub\_data$seasons

cor(rating,episodes)

0.00933734292361419

cor(rating, seasons)

0.0573723404625935

There is an insignificant positive correlation between show rating and episodes while there a significant positive correlation between show rating and seasons.

# 2.2.3 What is the average rating for shows that ran for 2 seasons? 3 seasons? 4 seasons?

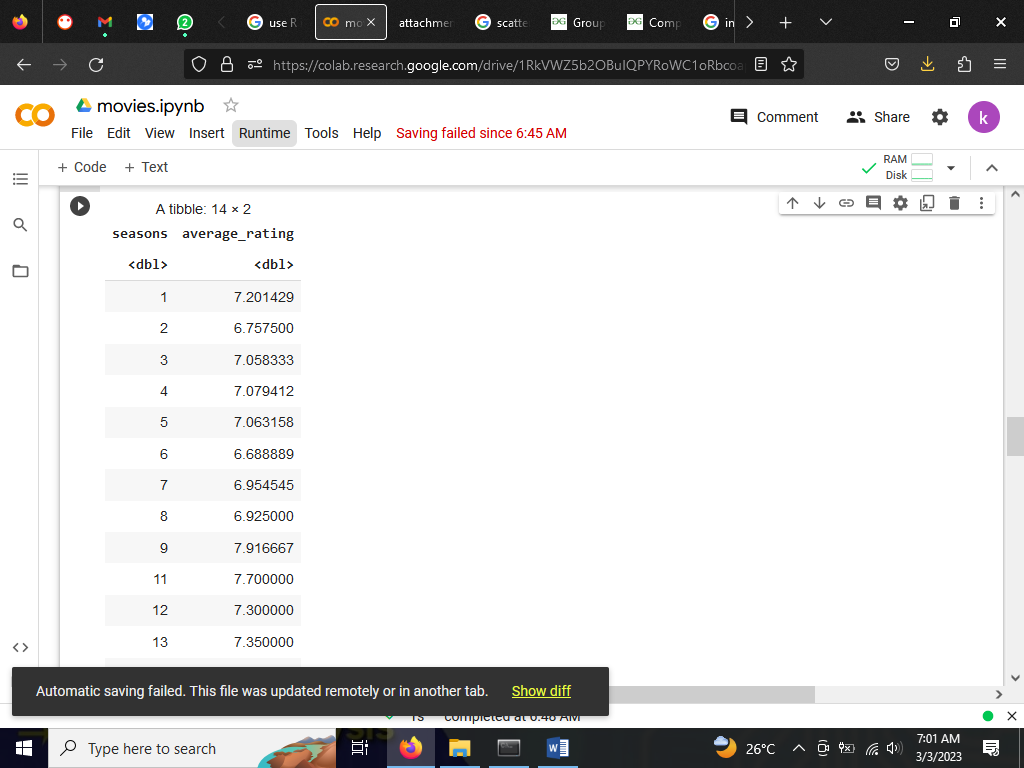
rating\_by\_seasons=sub\_data %>% group\_by(seasons) %>%

  summarise(

  average\_rating = mean(rating))

View(rating\_by\_seasons)

**Figure 2: Average Rating Grouped by Seasons**



The average rating of shows with one season is 7.2, shows with two seasons have an average rating of 6.8. Shows with three seasons are rated 7.1 on average, and shows with four seasons are rated 7.1 on average.

# 3.1 Correlation between Ratings and Episodes

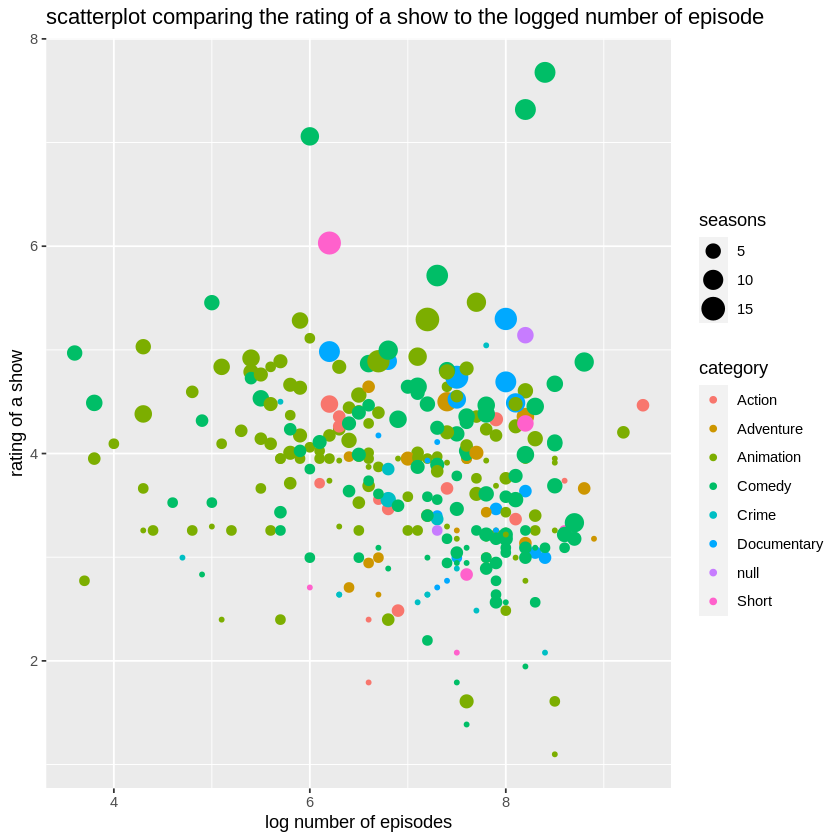
ggplot(sub\_data,aes(rating,log\_episodes)) +

  geom\_point(aes(colour = category,size = seasons)) +

   ggtitle("scatterplot comparing the rating of a show to the logged number of episode") +

  xlab("log number of episodes") + ylab("rating of a show")

**Figure 3: Correlation Plot between Episodes and Show Rating**



The show ratings increases as the number of seasons increases. Shows with 15 seasons have the highest ratings. Shows with 5 seasons are concentrated around the lowest ranks. The scatter plot indicated that genre has no impact on show ratings. Some shows with many episodes have low ratings while some shows with few episodes have high ratings. Therefore, we conclude the number of episodes has little to no effect on show rating.

# 3.2 Popularity by Genre

hundred=head(sub\_data, 150)

genres <- hundred$category

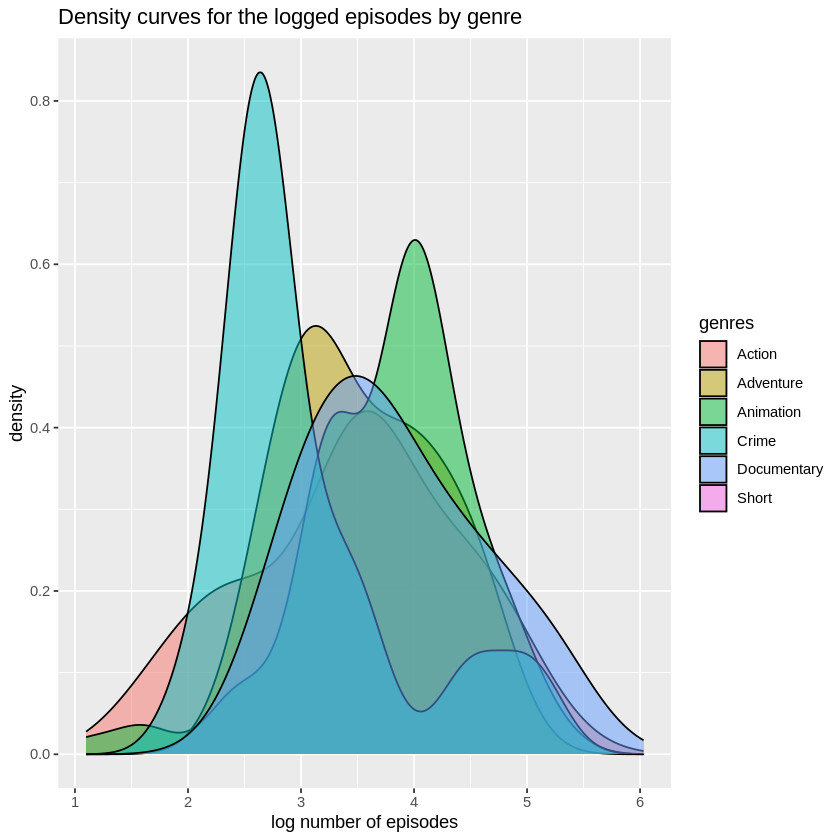
log\_epi <- log(hundred$episodes)

ggplot(hundred) + geom\_density(aes(x=log\_epi,fill = genres),alpha = 0.5) +

   ggtitle("Density curves for the logged episodes by genre") +

  xlab("log number of episodes") + ylab("density")

**Figure 4: Density Curves of Logged Episodes by Genre**



Crime shows with two to three shows are the most popular. Animation shows are the second most popular shows in the data. Adventure shows are the third most popular shows. Short shows are less popular in the sample.