PS1 Barna:K-Drama Dataset Analysis

#q1: download and read the number of rows library(readr) kdrama <- read_csv("C:/Users/barna/Downloads/kdrama.csv") View(kdrama) nrow(kdrama)

ans: Total 250 rows are present

#q2: List of Variables names(kdrama) Ans: [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

#q3. mean value of total number of episodes for all the kdramas mean_episodes <- mean(na.omit(kdrama$`Number of Episodes`)) mean_episodes Ans: 19.064

#q4. histogram of the shows rating hist(kdrama$Rating, main = "Histogram of K-Drama Ratings", xlab = "Rating")

#q5:rating higher than 9 points sum(na.omit(kdrama$`Rating`))$Ans: 2133.5 sum(na.omit(kdrama$`Rating` > 9)) Ans: 8

#q6. Rename variable Year.of.release to simply Year without creating a new variable

library(dplyr)

kdrama <- kdrama %>% rename(Year = `Year of release`) ANother way to create, other than tidyverse, as we know year is in third column, colnames(kdrama)[3] <- "Year"

#q7. dataset were released in 2020-2022 kdrama %>% filter(between(Year, 2020, 2022)) %>% nrow()

couldhave donw like sum(na.omit(kdrama$Year) >= 2020 and na.omit(kdrama$Year) <= 2022) but and symble wasnt working Ans: 106

#q8. type of variable is Duration class(kdrama$`Duration`) Ans: "character"

#q9. Recode variable Duration to a numerical variable measuring duration in minutes. plot histogram of the recoded variable.

#tidyverse way:

library(dplyr) library(stringr) class(kdrama$`Duration`)

kdrama <- kdrama %>% mutate(Duration_min = coalesce(as.numeric(str_extract(Duration, "\d+(?=\shr)")), 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")

#function way: though not sure, took online help..... :(Duration_to_minutes <- function(x) { hrs <- ifelse(grepl("hr", x, ignore.case = TRUE), as.numeric(sub(".?(\d+)\shr.", "\1", x)), 0) mins <- ifelse(grepl("min", x, ignore.case = TRUE), as.numeric(sub(".?(\d+)\smin.", "\1", x)), 0) hrs[is.na(hrs)] <- 0 mins[is.na(mins)] <- 0 60*hrs + mins }

hist(na.omit(dat$Duration_to_min), main = "Histogram of Episode Duration (minutes)", xlab = "Minutes")

10. dataset that will include shows with Original.Network being Netflix. library(dplyr) library(stringr)

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

nrow(netflix_exact) Ans: 12

11. What is the average rating score for the shows that have Netflix as an Original Network. mean(na.omit(netflix_exact‘*Rating*‘))$or, mean(netflix_exactRating$, na.rm = TRUE)

Ans: 8.65