Data Science Assignment 20174489 Question 1 and Question 2

Purpose

This document serves as my attempt at the Data Science practical on 24 MAy

```
rm(list = ls()) # Clean your environment:
gc() # garbage collection - It can be useful to call gc after a large object has been removed, as this
```

```
##          used (Mb) gc trigger (Mb) max used (Mb)
## Ncells 401430 21.5     827470 44.2   638945 34.2
## Vcells 728377  5.6    8388608 64.0  1633802 12.5
```

```
library(pacman)
p_load(tidyverse, rmsfuns)

# Source in all your functions:
list.files('code/', full.names = T, recursive = T) %>% as.list() %>% walk(~source(.))

# Loading Data

Movies_Data <- read.csv("data/Movies/Movies.csv")
```

Question 1: Profitable Movies

Hi friend

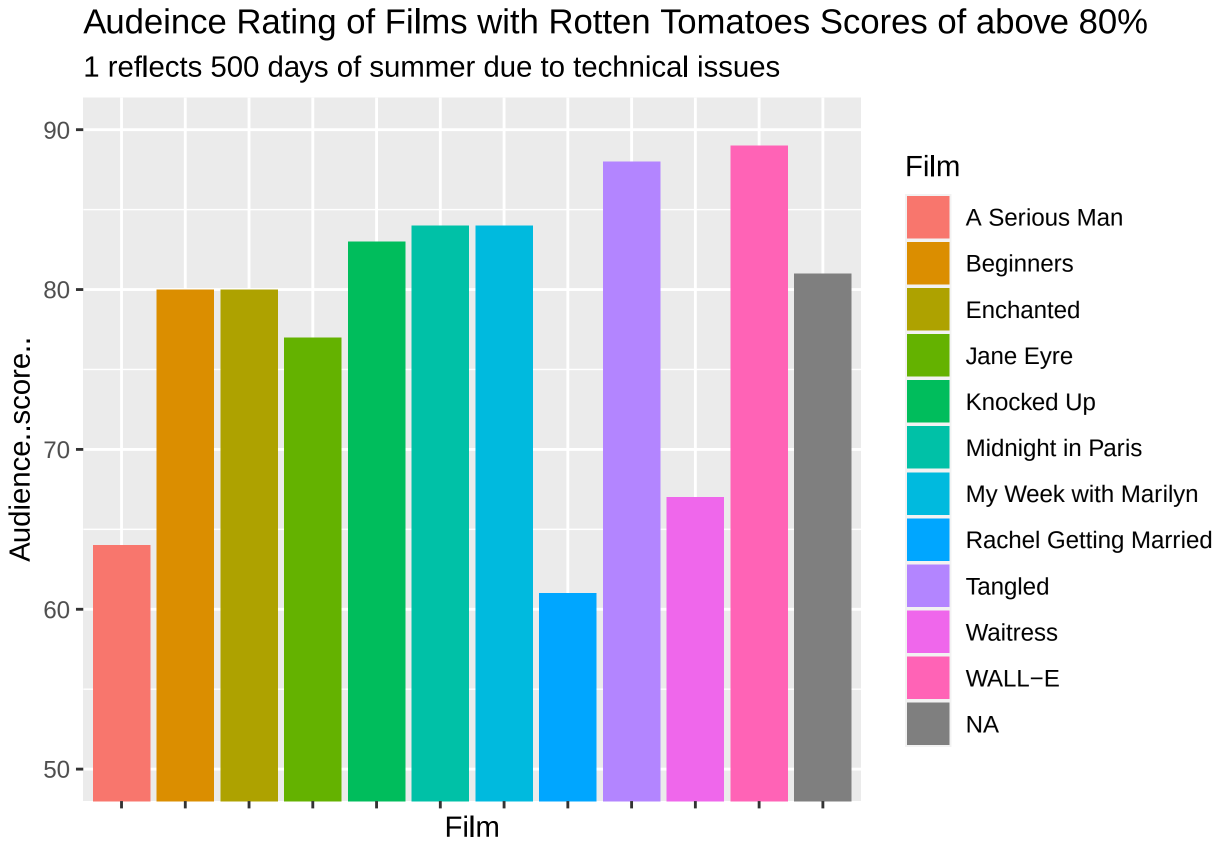
Hope you're doing well. As I am the meticulous person that you know and love, I have decided to test some of the theories that came up in our last conversation.

The first point I remember you making was that Rotten Tomatoes was always a great review platform, and that, if it had a rating of more than 80% on Rotten Tomatoes, audiences would rate it above 85% every time.

```r
#Plotting the audience rating of films with >80% on Rotten Tomatoes

Above_80_Data <- Above_80(Movies_Data)

Plot_Audience_Rating(Above_80_Data)
```



As you can see, there is no reason to believe that films with a Rotten

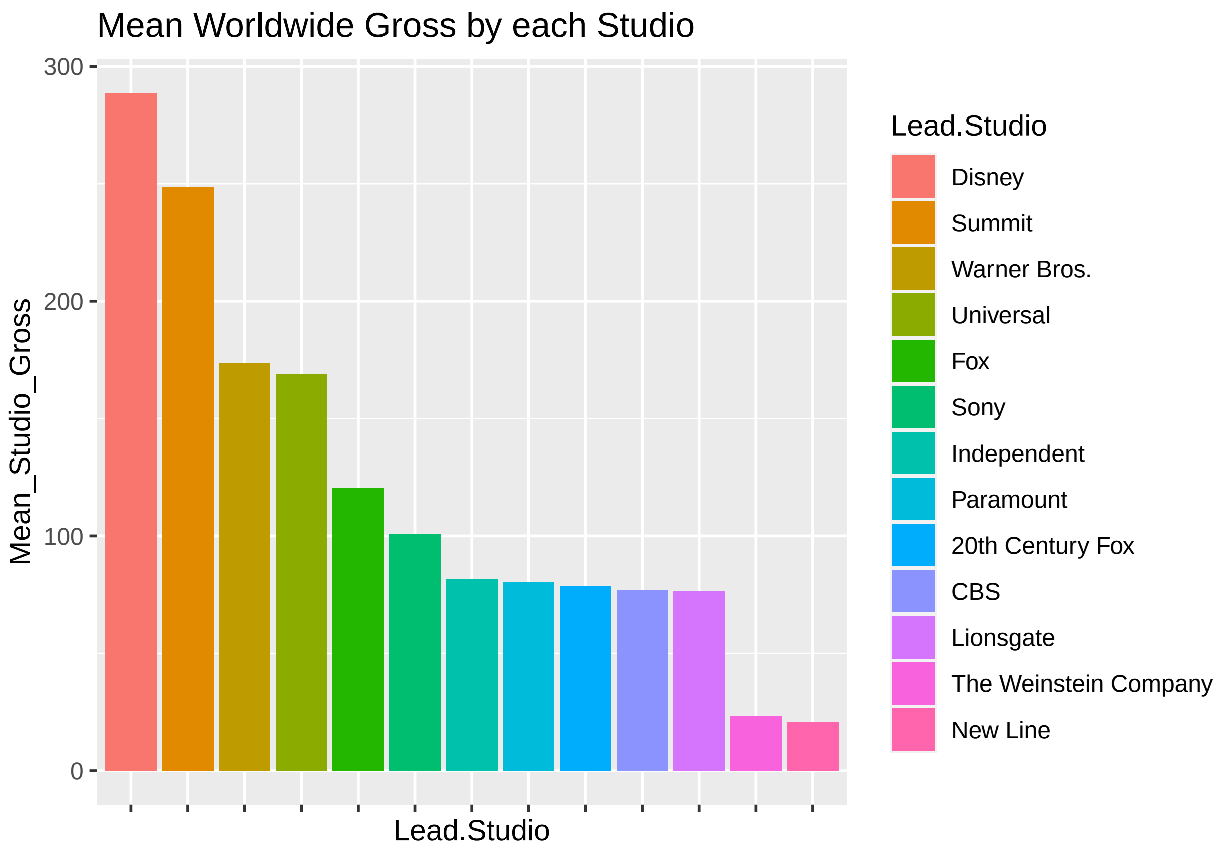
Your next claim was that, whilst Disney does not have the highest grossing numbers, their films are the most profitable.

I decided to_____

```
# Plotting Mean Worldwide Gross by each Studio

Gross_Data <- Rank_Gross(Movies_Data)

Gross_Plot(Gross_Data)
```

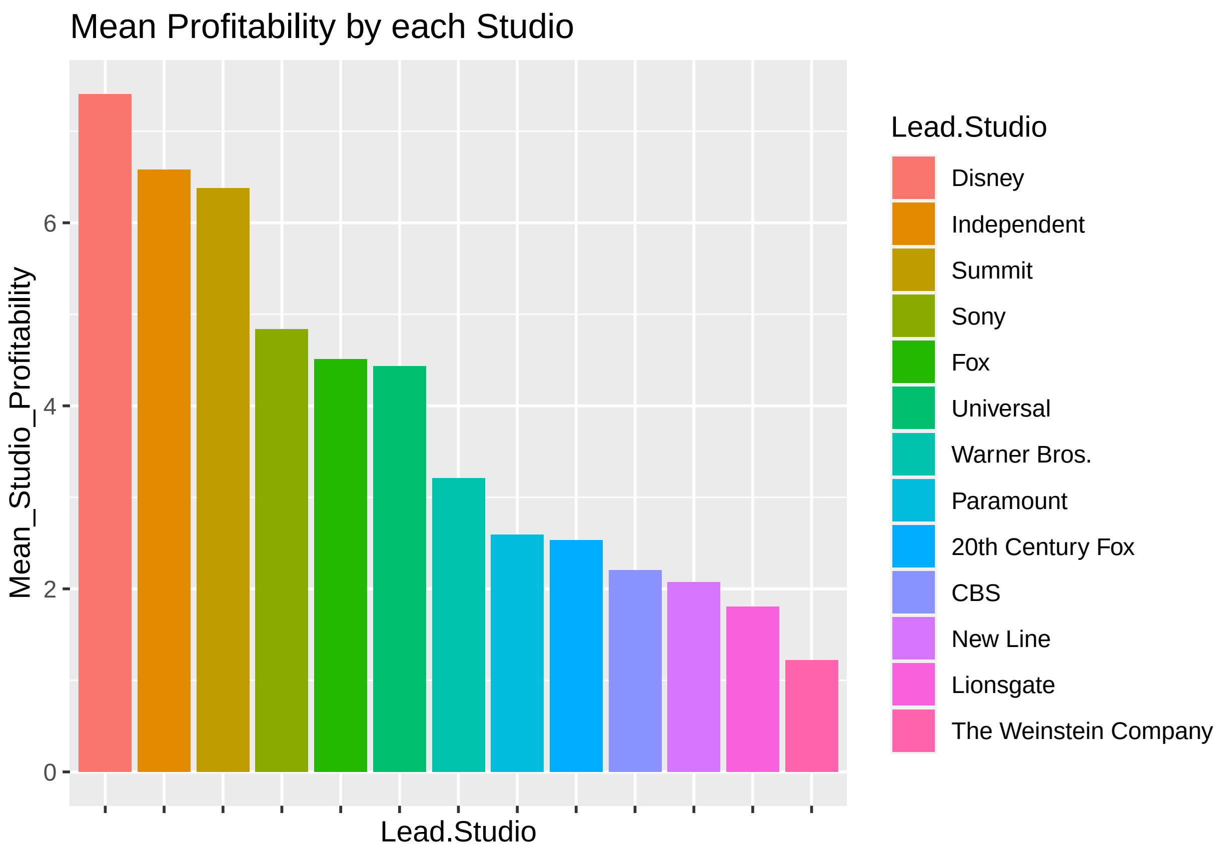


As you can see, Disney has the highest mean gross out of all the studios in my dataset. That already invalidates your first claim, but let's look at profitability.

```
# Plotting movie profitability by studio


Profit_Data <- Rank_Profit(Movies_Data)

Profit_Plot(Profit_Data)
```



So I do concede that Disney is the most profitable. That means you've scored half a point out of two so far. Let's see if you can get a passing mark.

```
#Showing correlation

Cor_AW <- Movies_Data %>% summarise(Corr = cor(Audience..score.., Worldwide.Gross,
                                               method = "spearman", use = "pairwise.complete.obs"))

Cor_RW <- Movies_Data %>% summarise(Corr = cor(Rotten.Tomatoes.., Worldwide.Gross,
                                               method = "spearman", use = "pairwise.complete.obs"))
```

From the data, I have observed that the correlation between audience score and worldwide gross is 0.291. This indicates weak positive correlation, going against what you said.

Looking at whether Rotten Tomatoes' ratings are any better, I observe that the correlation between the RT score and gross is -0.075. This means that there is pretty much no correlation.

So you are correct that audience rating is a better indicator, but it is not a good one.

Overall, you were pretty wrong with your theories. I do, however, appreciate the discussion and hope that we can have more in the future.

Question 2: Billionaires

```
# Reading and combining data
```