MCU Movies Data Scraping and Exploration

Rahim Jutha

Introduction

The Marvel Cinematic Universe films started being made when I joined high school and since then they have become a part of mainstream media. So far, the movies have been popular all over the world. What makes them popular is the superheroes are the stars, using the characters and storylines they created over the last couple of decades. Due to the sizes of both Marvel Studios and Disney Studios, the resources and payoff for these projects are enormous.

Even though I have experienced these films first-hand, I have many questions about how profitable and how successful they are. I want to know if the production of the MCU films has increased over time plus whether they have generated more revenue since the first film. I also want to identify any trends within the scores and reviews the movies receive.

Scraping and Cleaning the Data from Wikipedia

Data scientists spend 60% of their time cleaning and organizing data. Lucky for us we can scrape the tables from Wikipedia.org with the r package rvest, navigating the HTML code from this URL to extracting the tables is simple. However, some rows are insignificant for our analysis and exist for aesthetic purposes. Also, most of the columns need to be coerced to correct data types so that we can manipulate them later. This includes renaming columns, changing character columns to numeric ones, and removing specific strings from elements.

```
library(knitr)
library(rvest)
library(dplyr)
library(stringr)
library(readr)
library(ggplot2)
library(tidyr)
```

```
marvel_films_url = "https://en.wikipedia.org/wiki/List_of_Marvel_Cinematic_Universe_films"

marvel_html = read_html(marvel_films_url)
marvel_table_nodes =html_nodes(marvel_html,"table")

# CPR = Critical and Public Response
CPR_df = html_table(marvel_table_nodes[[7]], fill =TRUE)
# BOP = Box Office Performance
BOP_df = html_table(marvel_table_nodes[[10]], fill =TRUE)
```

Cleaning the Box Office Performance Table

```
# Rename Columns since there are columns with identical names
names(BOP_df)[c(2,3,4,5,6,7,8,9)] <-
  c('U.S._release_date',
    'Box_office_gross_U.S._and_Canada',
    'Box_office_gross_Other_territories',
    'Box_office_gross_Worldwide',
    'Alltime_Ranking_U.S._and_Canada',
    'Alltime_Ranking_Worldwide',
    'Budget_millions',
    'Ref')

# Remove Ref column since its not significant
BOP_df %>% select(-Ref) -> BOP_df

# Remove insignificant columns and fix numbering
BOP_df[-c(1,2,3,10,17,29),] -> BOP_df
row.names(BOP_df) <- 1:nrow(BOP_df)

# Fix Column Types
BOP_df %>%
  mutate(U.S._release_date = as.Date(U.S._release_date, format = "%B %d, %Y")) %>%
  mutate(Alltime_Ranking_U.S._and_Canada = as.numeric(Alltime_Ranking_U.S._and_Canada)) %>%
  mutate(Alltime_Ranking_Worldwide = as.numeric(Alltime_Ranking_Worldwide)) -> BOP_df

# Fix Money-related Columns
BOP_df$Box_office_gross_U.S._and_Canada =
  str_replace_all(BOP_df$Box_office_gross_U.S._and_Canada, "\\$|,", "")

BOP_df$Box_office_gross_Other_territories =
  str_replace_all(BOP_df$Box_office_gross_Other_territories, "\\$|,", "")

BOP_df$Box_office_gross_Worldwide =
  str_replace_all(BOP_df$Box_office_gross_Worldwide, "\\$|,", "")

BOP_df$Budget_millions =
  str_replace_all(BOP_df$Budget_millions, "[[:blank:]]million|\\$", "")

index = grep(fixed("-"), BOP_df$Budget_millions)
BOP_df$Budget_millions[index] =
  c(round(mean(316,400), digits = 1),
    round(mean(150,175), digits = 1))

BOP_df %>%
  mutate(Box_office_gross_U.S._and_Canada = as.numeric(Box_office_gross_U.S._and_Canada)) %>%
  mutate(Box_office_gross_Other_territories = as.numeric(Box_office_gross_Other_territories)) %>%
  mutate(Box_office_gross_Worldwide = as.numeric(Box_office_gross_Worldwide)) %>%
  mutate(Budget_millions = as.numeric(Budget_millions)) -> BOP_df

# Add new column for release year
BOP_df %>%
  mutate(U.S._release_year = as.integer(substring(as.character(as.POSIXct(U.S._release_date, format = "
```

Cleaning the Ratings Table

```
names(CPR_df)[c(2,3,4)] <-
  c('Rotten_Tomatoes','Metacritic','CinemaScore')
CPR_df[-c(1,2,3,10,17),] -> CPR_df
row.names(CPR_df) <- 1:nrow(CPR_df)

CPR_df %>%
  mutate(Rotten_Tomatoes = parse_number(Rotten_Tomatoes)) %>%
  mutate(Metacritic = parse_number(Metacritic)) -> CPR_df

CPR_df$CinemaScore =
  str_replace_all(CPR_df$CinemaScore, "\\[.*", "")
```

Save Data as CSV Files

```
write.csv(BOP_df, "MCU_BOP.csv")
write.csv(CPR_df, "MCU_CPR.csv")
sprintf("Data Collected on %s",Sys.Date())
```

Let's answer my questions!

To do this I will be generating plots to visualize the data I cleaned in a way that answers each question. These will be made using the r package ggplot2 and will include scatterplots, histograms, and trendlines. If you want to use the data download these csv files and run the code below.

```
BOP = read.csv("MCU_BOP.csv")
CPR = read.csv("MCU_CPR.csv")
```

Table of of movies in the Dataset

```
BOP %>%
  select('Film', `U.S._release_year`) %>%
  kable()
```

Film	U.S._release_year
Iron Man	2008
The Incredible Hulk	2008
Iron Man 2	2010
Thor	2011
Captain America: The First Avenger	2011
Marvel's The Avengers	2012
Iron Man 3	2013
Thor: The Dark World	2013
Captain America: The Winter Soldier	2014
Guardians of the Galaxy	2014
Avengers: Age of Ultron	2015

Film	U.S._release_year
Ant-Man	2015
Captain America: Civil War	2016
Doctor Strange	2016
Guardians of the Galaxy Vol. 2	2017
Spider-Man: Homecoming	2017
Thor: Ragnarok	2017
Black Panther	2018
Avengers: Infinity War	2018
Ant-Man and the Wasp	2018
Captain Marvel	2019
Avengers: Endgame	2019
Spider-Man: Far From Home	2019

How profitable are the MCU movies?

To quantify profit for each my movie I calculated the gross profit margin for each movie. To calculate this, you need the revenue and the budget which we have in the data. The gross profit margin is given by the formula

$$\text{Gross Profit Margin} = \frac{Revenue - Budget}{Revenue}. \tag{1}$$

Now that I've calculated I'm going to make a level variable and group the movies by into 3 intervals of gross profit margin to distinguish groups in the graphs below. The levels are: Greater than 80%, Between 60% and 80%, and Less than 60%.

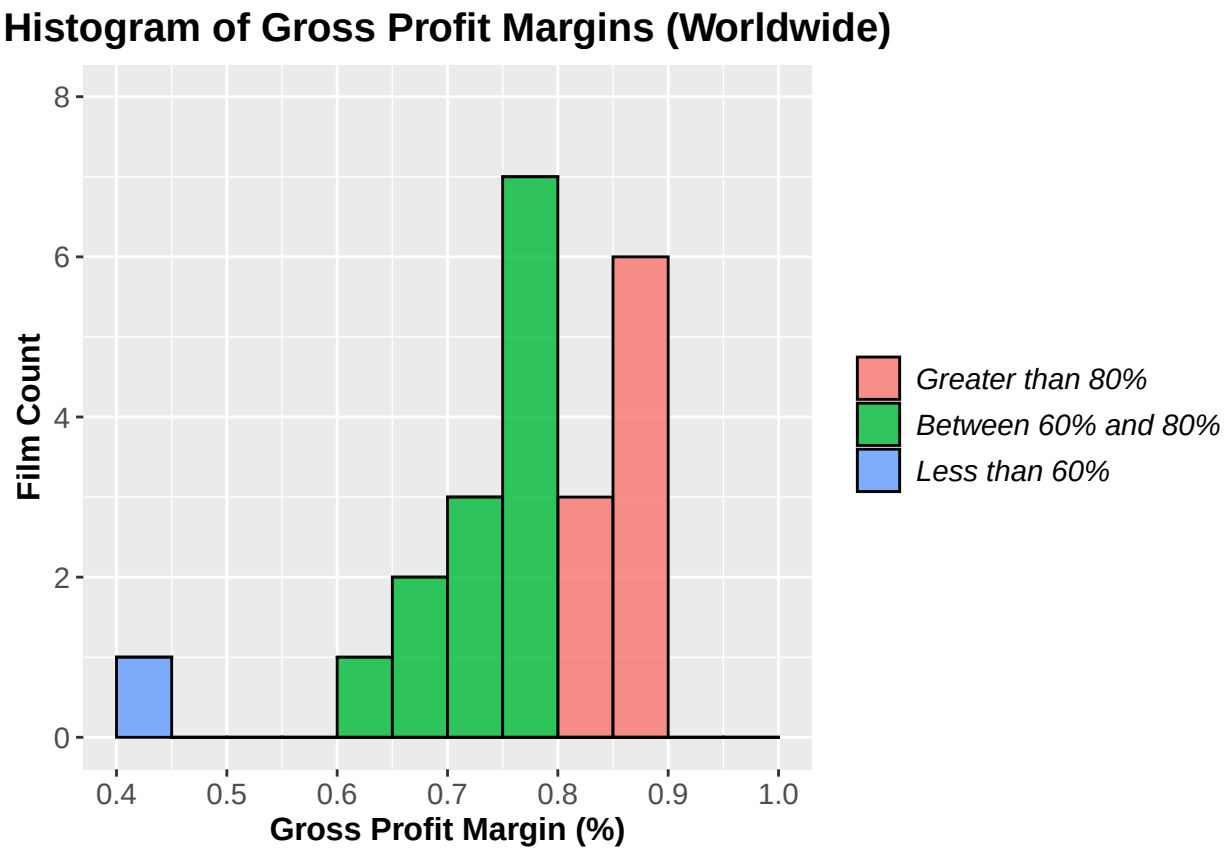
```
BOP %>%
  mutate(GPM_worldwide = (BOP$Box_office_gross_Worldwide - BOP$Budget_millions * 10^6) / BOP$Box_office_

  mutate(
    freq = case_when(
      GPM_worldwide > .8 ~ 'Greater than 80%',
      between(GPM_worldwide, .6, .8) ~ 'Between 60% and 80%',
      GPM_worldwide < .5 ~ 'Less than 60%'
    )
  ) %>%
    mutate(freq = factor(
    freq,
    levels = c(
      'Greater than 80%',
      'Between 60% and 80%',
      'Less than 60%'
    )
  )) -> BOP

  ggplot(BOP, aes(x = BOP$GPM_worldwide, fill=freq)) +
  geom_histogram(breaks=seq(0, 1, by=0.05),
                 col="black",
                 alpha=.8) +
  labs(title="Histogram of Gross Profit Margins (Worldwide)",
       x="Gross Profit Margin (%)", y="Film Count",
       fill = NULL) +
  scale_y_continuous(breaks = c(0,2,4,6,8), limits = c(0, 8)) +
```

```
  scale_x_continuous(breaks = seq(0,1, by=0.1), limits = c(.4,1)) +

  theme_gray() +
  theme(axis.title = element_text(size = 12, face = 'bold'),
        plot.title = element_text(hjust = 0.5, face = 'bold', size = 15),
        legend.text = element_text(size = 11 ,face = 'italic'),
        axis.text = element_text(size = 11)) +
    scale_color_discrete(name = "Gross Profit Margin")
```



From this histogram we observe that most of the films lie in the range of between 60% and 80% gross profit margin. We also notice that there is one outlier indicated by the blue bar. This film is The Incredible Hulk, which was the second MCU movie produced in 2008, from this histogram we can see the movie made a lot less profit than the other MCU movies.

Has the amount of profits and budgets for the MCU movies increased since the first?

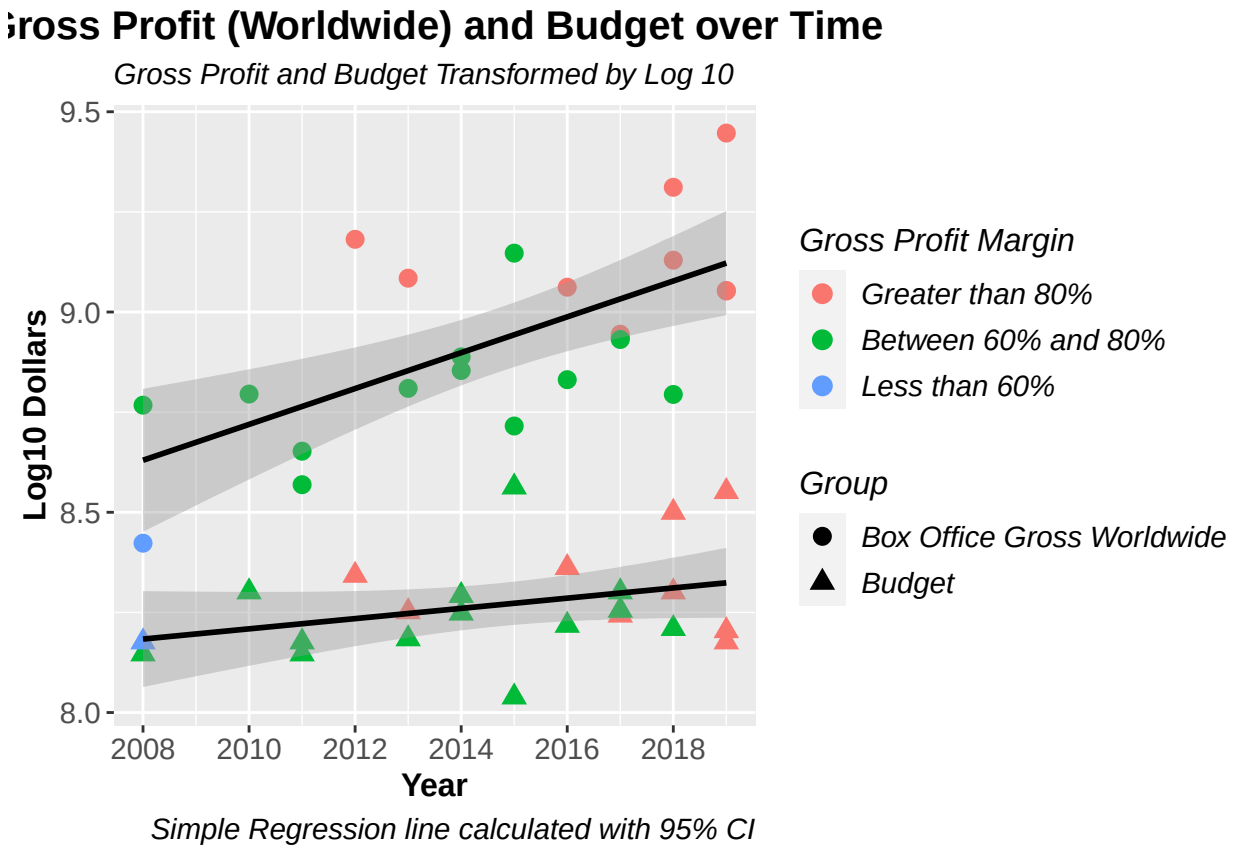
Now I want to see if the movies have been generating more profits since the first movie in 2008. To do this I plotted the revenues and budgets, distinguishing them by circles and triangles. I also used ggplot2 to add a simple regression line to visualize the trend of the revenue and budget for each year from 2008 to 2019.

```
BOP %>% mutate(B_P = log10(Box_office_gross_Worldwide)) %>%
  select(U.S._release_year, freq, B_P) %>%
  mutate(B_P_level = "Box Office Gross Worldwide") -> B_P1
```

```r
BOP %>% mutate(B_P = log10(Budget_millions * 10^6)) %>%
  select(U.S._release_year, freq, B_P) %>%
  mutate(B_P_level = "Budget") -> B_P2

B_P = rbind(B_P1,B_P2)
B_P %>% mutate(B_P_level = factor(B_P_level, levels = c("Box Office Gross Worldwide",
                                          "Budget"))) -> B_P

ggplot(B_P, aes(x=U.S._release_year, y = B_P, group = B_P_level, colour = freq, shape = B_P_level)) +
  geom_point(size = 3, alpha = 1) +
  geom_smooth(method = 'lm',colour = "black") +
  scale_x_continuous(breaks = seq(2008,2019, by=2)) +
  theme_gray() +
  labs(title="Gross Profit (Worldwide) and Budget over Time",
       subtitle = "Gross Profit and Budget Transformed by Log 10",
       x="Year", y="Log10 Dollars", caption = "Simple Regression line calculated with 95% CI") +
  theme(axis.title = element_text(size = 12, face = 'bold'),
        plot.title = element_text(hjust = 0.5, face = 'bold', size = 15),
        legend.text = element_text(size = 11 ,face = 'italic'),
        axis.text = element_text(size = 11),
        legend.title = element_text(size = 12, face = 'italic'),
        plot.subtitle = element_text(face = 'italic'),
        plot.caption = element_text(hjust = 1,size = 11 ,face = 'italic')) +
        scale_color_discrete(name = "Gross Profit Margin") +
        scale_shape_discrete(name  ="Group")
```

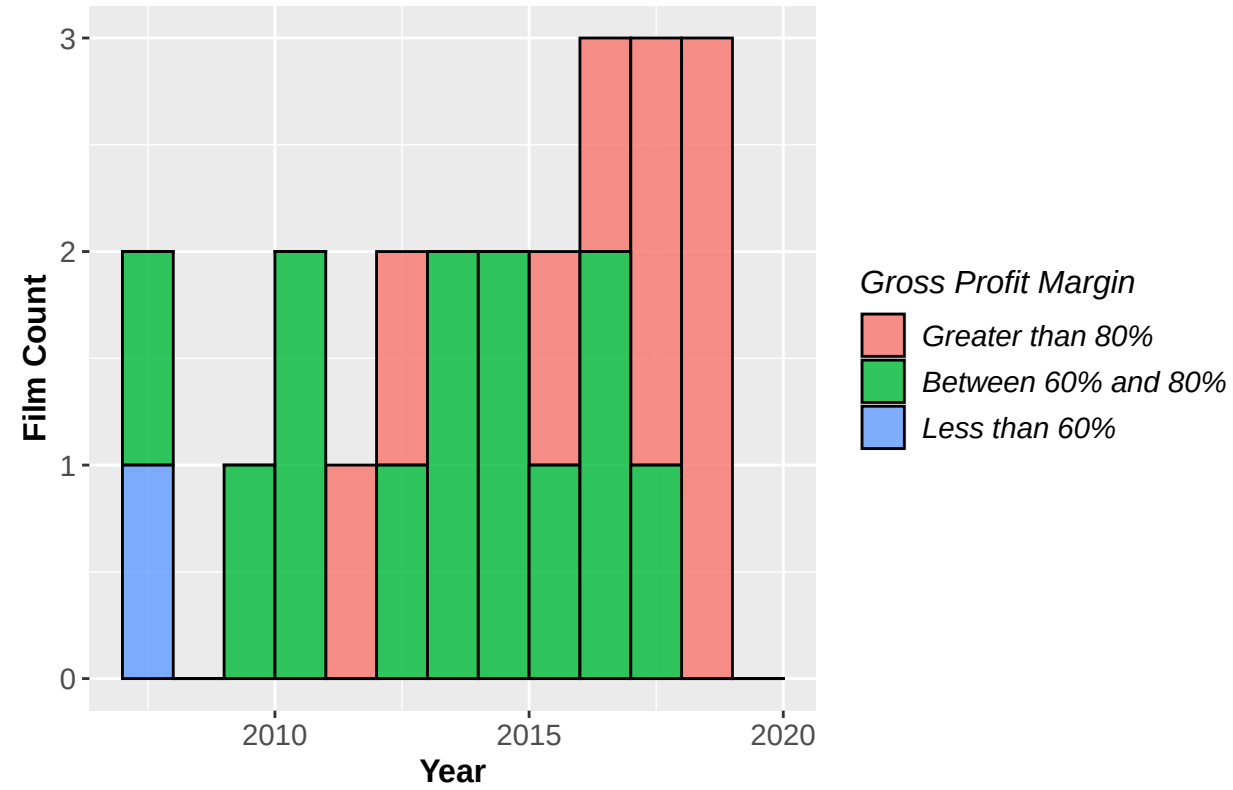


From this plot we can see that the revenues for the MCU films has been increasing over the years and the budgets for the movies have stayed relatively constant. From this we can conclude that the movies have been making more money over time without having to increase the budget of the movies. This makes sense because as more movies are produced, more people are exposed to them and hear about them and eventually go watch them.

Have productions for the movies increased?

Next, I wanted to know if they started making more MCU movies as they got more popular. To answer this, I made a histogram to count the number of movies produced each year.

```
ggplot(BOP, aes(x=U.S._release_year, fill = freq)) +
  geom_histogram(breaks=seq(2007, 2020, by=1),
                 color = "black", alpha = 0.8) +
  theme_gray() +
  labs(title="Gross Profit (Worldwide) and Budget over Time",
       x="Year", y="Film Count") +
  theme(axis.title = element_text(size = 12, face = 'bold'),
        plot.title = element_text(hjust = 0.5, face = 'bold', size = 15),
        legend.text = element_text(size = 11 ,face = 'italic'),
        axis.text = element_text(size = 11),
        legend.title = element_text(size = 12, face = 'italic')) +
        scale_fill_discrete(name = "Gross Profit Margin")
```



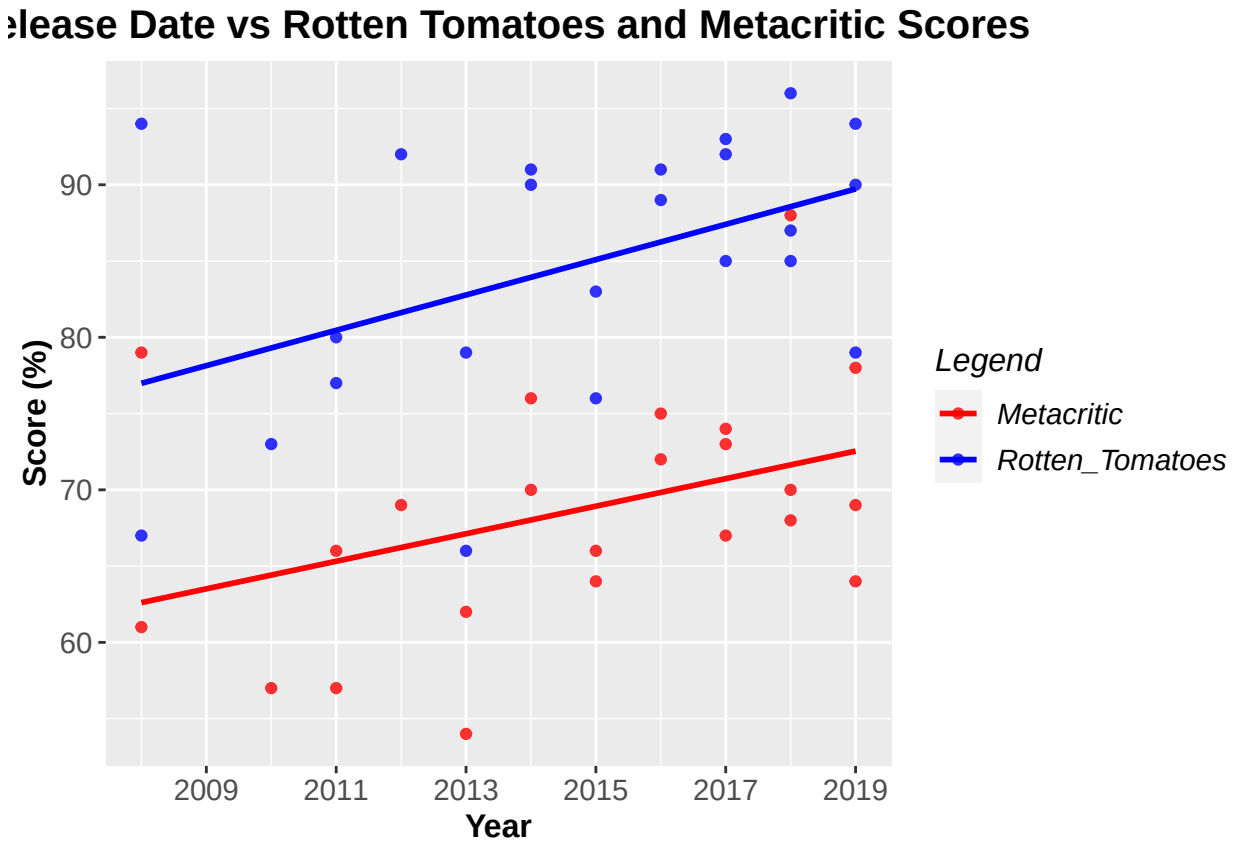
From this histogram we can see that from 2008 to 2016 we have about 2 movies per year being produced. However, from 2017-2019, there have been 3 movies produced each year which shows that the production

of films has increased recently. In this histogram we can also see that the movies in this time period have performed well based on their Gross Profit Margin.

Counts of Scores Lets take a look at some histograms showing how Rotten CinemaScore, Rotten Tomatoes, and Metacritic Rank the films.

```
BOP %>% inner_join(CPR, by="Film") %>%
  select(Film,U.S._release_year, Rotten_Tomatoes, Metacritic) %>%
  gather(Type, value = Score, Rotten_Tomatoes, Metacritic) %>%
  ggplot(aes(x = U.S._release_year, y = Score, colour = Type)) +
  geom_point(alpha = 0.8) +
  geom_smooth(se = FALSE, method = 'lm', alpha = 0.8) +
  scale_x_continuous(breaks = seq(2007,2019,2), limits = c(2008,2019)) +
  theme_gray() +
  labs(title="Release Date vs Rotten Tomatoes and Metacritic Scores",
       x="Year", y="Score (%)", color = "Legend") +
  theme(axis.title = element_text(size = 12, face = 'bold'),
        plot.title = element_text(hjust = 0.5, face = 'bold', size = 15),
        legend.text = element_text(size = 11 ,face = 'italic'),
        axis.text = element_text(size = 11),
        legend.title = element_text(size = 12, face = 'italic')) +
        scale_fill_discrete(name = "Gross Profit Margin") +
  scale_color_manual(values = c('red', 'blue'))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We observe that there is some upward trend to the scores over time which makes sense since the movies as whole are getting bigger and better.

Conclusion

By doing this short analysis I was able to learn a lot about the MCU films. It was interesting to see that all the movies had a gross profit margin greater than 60% except The Incredible Hulk. I remember which is also one of the the very first films. We also found that the budgets were staying the same but producing increasing amounts of profit. This makes sense cause they have become so popular. This also supports the fact that we found that they have been producing even more movies in the last few years.

From observing the histograms from the 3 scores for each movie we see that generally the ratings of the movies are high, which is what we expect because they are so popular. Lastly we observed the trend of the ratings of the movies overtime and see that they are increasing, meaning the general publics opinion of them have been increasing since they started.

Overall, this small project allowed me to take a peek at some of the data revolving the MCU movies. Although it was basic, the exploration of the data was fruitful and I learned a ton about R and some meaningful information abou the films.