AMAT 464: Predicting Box Office Revenue

Max Hofstetter & Tobin Cherian

2023-04-28

Introduction

Entertainment is the one thing that brings amusement to people in the world that ranges from films, television shows, art, and music. For our research, we conducted the Top-20 movies of 2010 where we look into their weekly revenues along with their characteristics of MPAA rating and season release. We find this topic sparking interest becasue in the film industry, it can help production studios make logical decisions on what type of movies to create and market based on several characteristics and looking into the weekly revenues to produce a time series model to analyze and predict box office revenues. We input our model into a time series format as for each of the movies, we calculated the probabilities by dividing each of the corresponding weekly revenue by the total amount it has made in box office. We want to diagram the success and categories of which types of movies brought success in the year 2010 on revenue. We can use the graphs coded in R and the data points to find which types of movies found success and why. In the segment of code, we uploaded the necesssary libraries for our project and inputted the dataset that we gathered and collected to help figure out the box office revenue.

```
library(dplyr)
library(tidyverse)
library(forecast)
library(ggplot2)
library(strex)
```

Movie Data and Collecting Revenue

We used 20 popular movies that released in 2010. The data was collected from the Internet Movie Database (IMDb). Data used includes budget, weekly revenue, total revenue, rating, season of release, and # of weeks in theaters. Once we collected all the necessary data we created tables in excel to organize and format the data so it would fit into R nicely. The two vertical tables is what the data looks like transposed in R. Each movie has 28 rows for each week of revenue. As mentioned on how we used IMDB to research onto this project, for each of the movies of 2010, we looked into the amount of money a movie has made in the box office each week. We gather each of the 28 weekly box office revenue and divide it by the exact total revenue the movie has made in order to get the probability that we extract into the "BoxOffice2010" dataset. Based on the table, the revenue probabilities either increase or decrease as the weeks go by for each of the movies.

```
MovieData<-read.csv("/Users/tobincherian/Downloads/BoxOffice2010.csv", header = TRUE, stringsAsFactors =
df <- MovieData[, c(1, 2, 3, 4:31)]
view(df)
```

Purpose of the Time Series Model

We are modeling our dataset as a Time Series model that has different revenue moving from week to week. The revenue of the next week is influenced by the revenue of the previous week and the total revenue is

impacted by many factors. The factors that we are studying in this project is Season in which the movie is released and Rating(G, PG, PG-13, R). The idea behind our Time Series Model is that we will be able to identify patterns and trends in the data and use them to make predictions on the types of movies that can generate continued revenue and success. In the segment of code, we code a dataset known as "df_long" where it compiles a list of the 20 movies and how much they made in revenue for 28 weeks. There are 560 entries in this table and each of the weekly revenue of each of the movies has either increased or decreased as their weeks in theaters went by. The plot demonstrates all the 20 movies and the success on the revenue it has made as time went by.

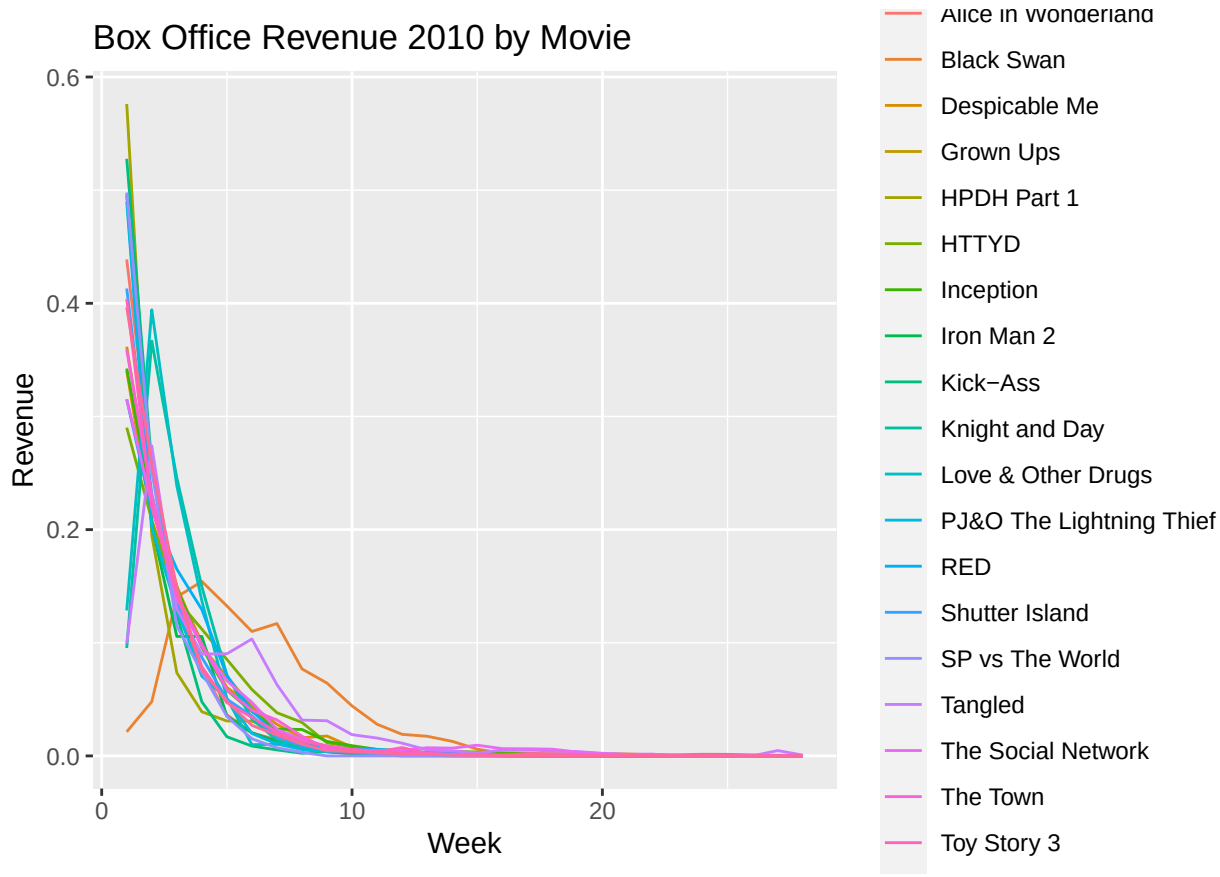
```
df_long <- df %>%
  pivot_longer(cols = 4:31, names_to = "week", values_to = "revenue")
df_long$week <- str_extract_numbers(df_long$week)
str(df_long$week)
```

```
## List of 560
##  $ : num 1
##  $ : num 2
##  $ : num 3
##  $ : num 4
##  $ : num 5
##  $ : num 6
##  $ : num 7
##  $ : num 8
##  $ : num 9
##  $ : num 10
##  $ : num 11
##  $ : num 12
##  $ : num 13
##  $ : num 14
##  $ : num 15
##  $ : num 16
##  $ : num 17
##  $ : num 18
##  $ : num 19
##  $ : num 20
##  $ : num 21
##  $ : num 22
##  $ : num 23
##  $ : num 24
##  $ : num 25
##  $ : num 26
##  $ : num 27
##  $ : num 28
##  $ : num 1
##  $ : num 2
##  $ : num 3
##  $ : num 4
##  $ : num 5
##  $ : num 6
##  $ : num 7
##  $ : num 8
##  $ : num 9
##  $ : num 10
##  $ : num 11
```

```
##  $ : num 12
##  $ : num 13
##  $ : num 14
##  $ : num 15
##  $ : num 16
##  $ : num 17
##  $ : num 18
##  $ : num 19
##  $ : num 20
##  $ : num 21
##  $ : num 22
##  $ : num 23
##  $ : num 24
##  $ : num 25
##  $ : num 26
##  $ : num 27
##  $ : num 28
##  $ : num 1
##  $ : num 2
##  $ : num 3
##  $ : num 4
##  $ : num 5
##  $ : num 6
##  $ : num 7
##  $ : num 8
##  $ : num 9
##  $ : num 10
##  $ : num 11
##  $ : num 12
##  $ : num 13
##  $ : num 14
##  $ : num 15
##  $ : num 16
##  $ : num 17
##  $ : num 18
##  $ : num 19
##  $ : num 20
##  $ : num 21
##  $ : num 22
##  $ : num 23
##  $ : num 24
##  $ : num 25
##  $ : num 26
##  $ : num 27
##  $ : num 28
##  $ : num 1
##  $ : num 2
##  $ : num 3
##  $ : num 4
##  $ : num 5
##  $ : num 6
##  $ : num 7
##  $ : num 8
##  $ : num 9
```

```
##  $ : num 10
##  $ : num 11
##  $ : num 12
##  $ : num 13
##  $ : num 14
##  $ : num 15
##   [list output truncated]
```

```
View(df_long)
df_long$RevenueX <- c(df_long$revenue[-28], 0)
df_long$RevenueX[seq(28, nrow(df_long), by=28)] <- 0
df_long$RevenueY <- c(0, df_long$revenue[-1])
ggplot(df_long, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 by Movie")
```

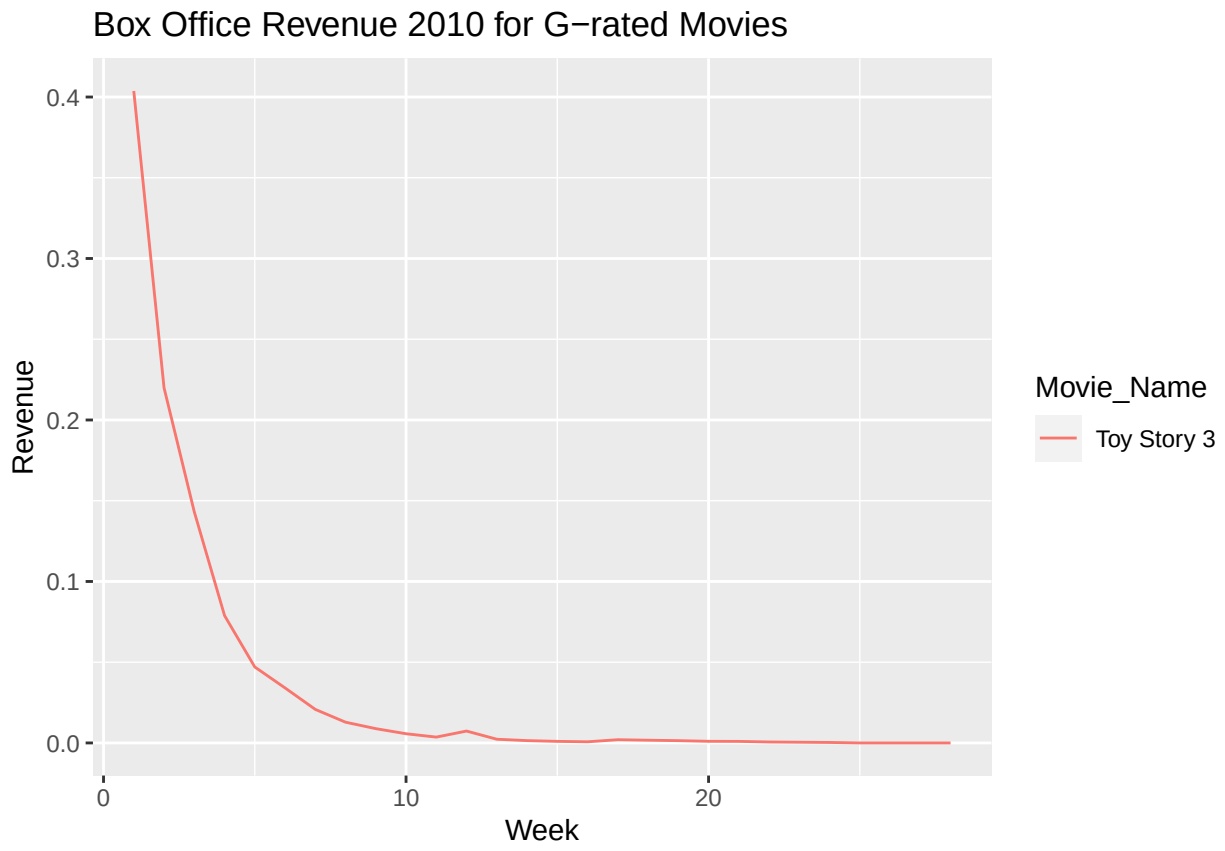


Time Series Plots for MPAAA-Rating

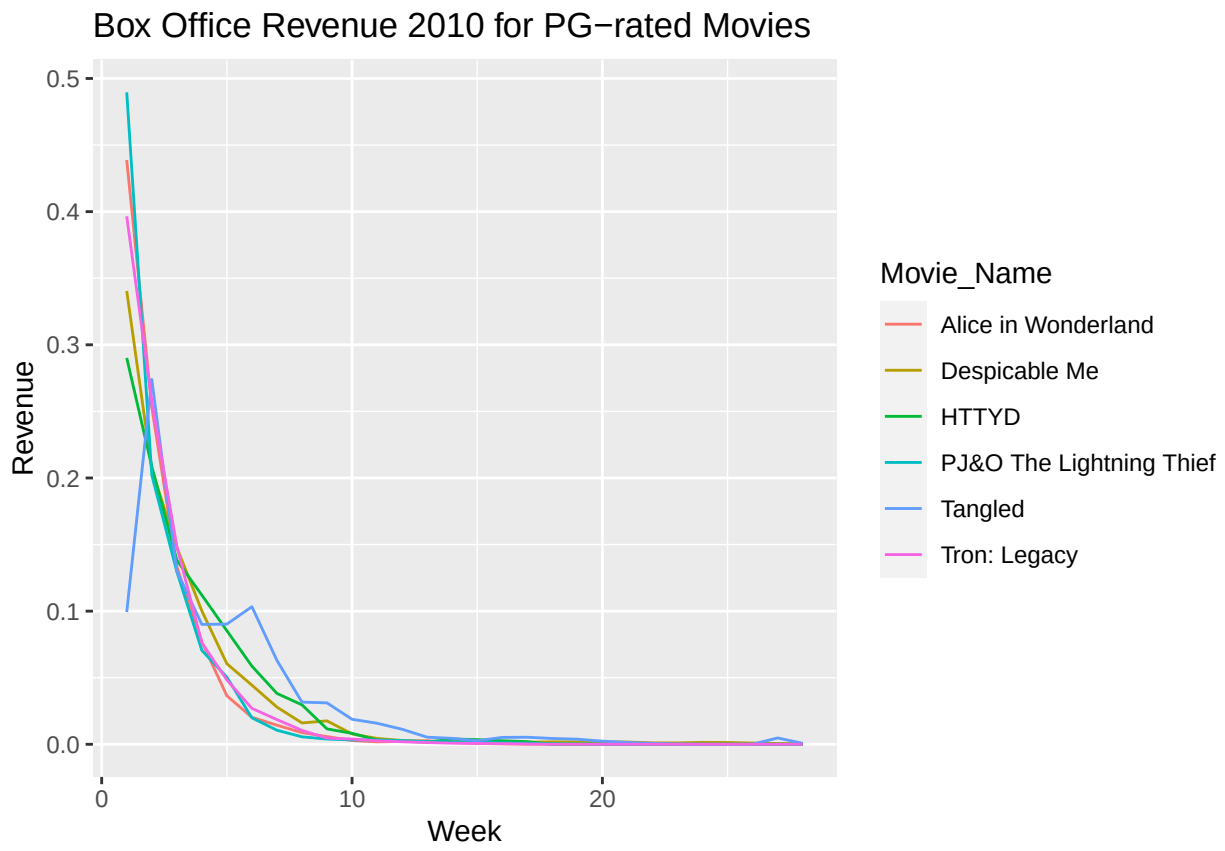
In this segment of code, we want to view if there's a significance of the revenue based on the movie rating for the general audience, such as G, PG, PG-13, and R. As you can see in the Rated-G plot, there's only one movie - Toy Story 3 - that made some significance in revenue around the first 10 out of 28 weeks considering anyone can watch it. The PG category were mostly animated movies so they made the most revenue compared to the G category. The plots of the PG-13 and R category are a little identical in their results. The movie "Knight and Day" (PG-13) and "Love and Other Drugs" had a slight increase around the third through fifth week of its release. Another one in Rated-R category is "Black Swan" which peaked later

then most movies. This could mean it had poor marketing but a solid product which had people talking for weeks following the release.

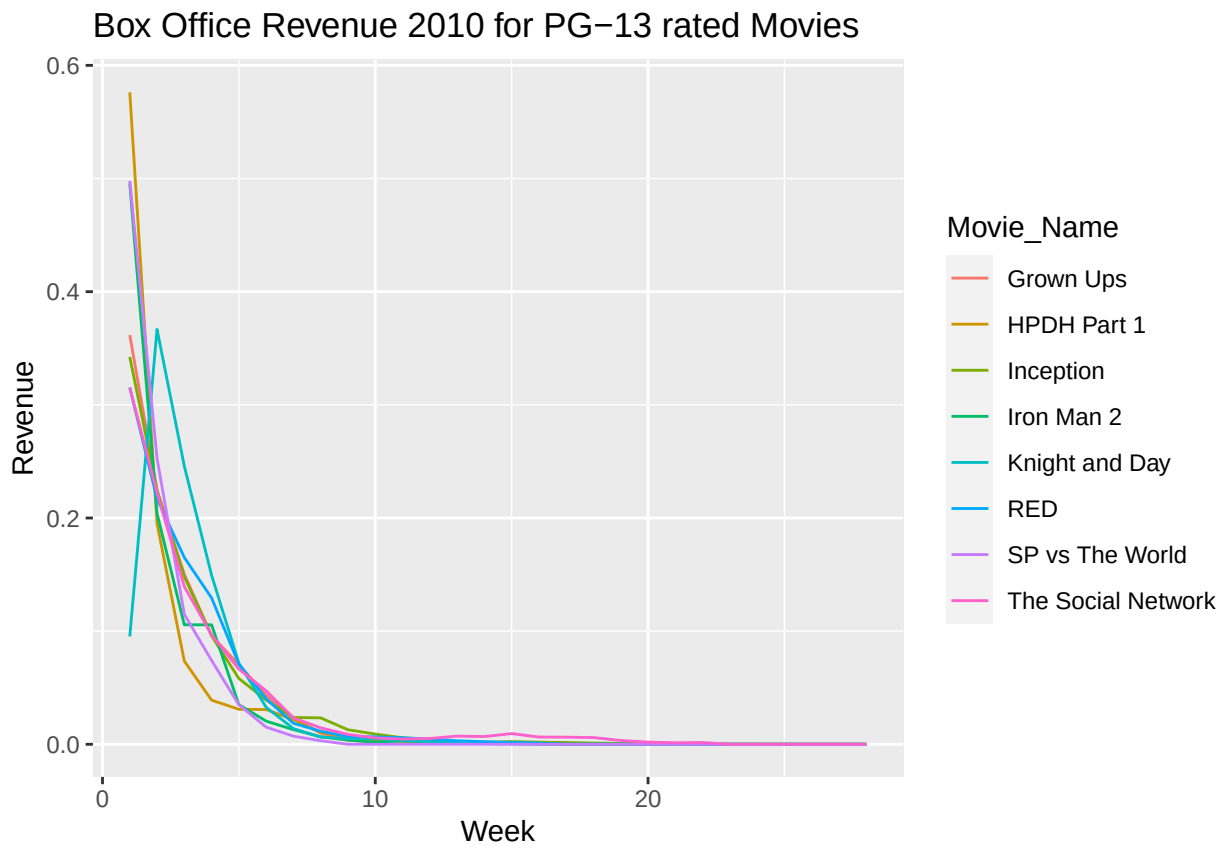
```
# Time series plot for G-rated movies
df_g <- df_long %>%
  filter(Rating == "G")
ggplot(df_g, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for G-rated Movies")
```



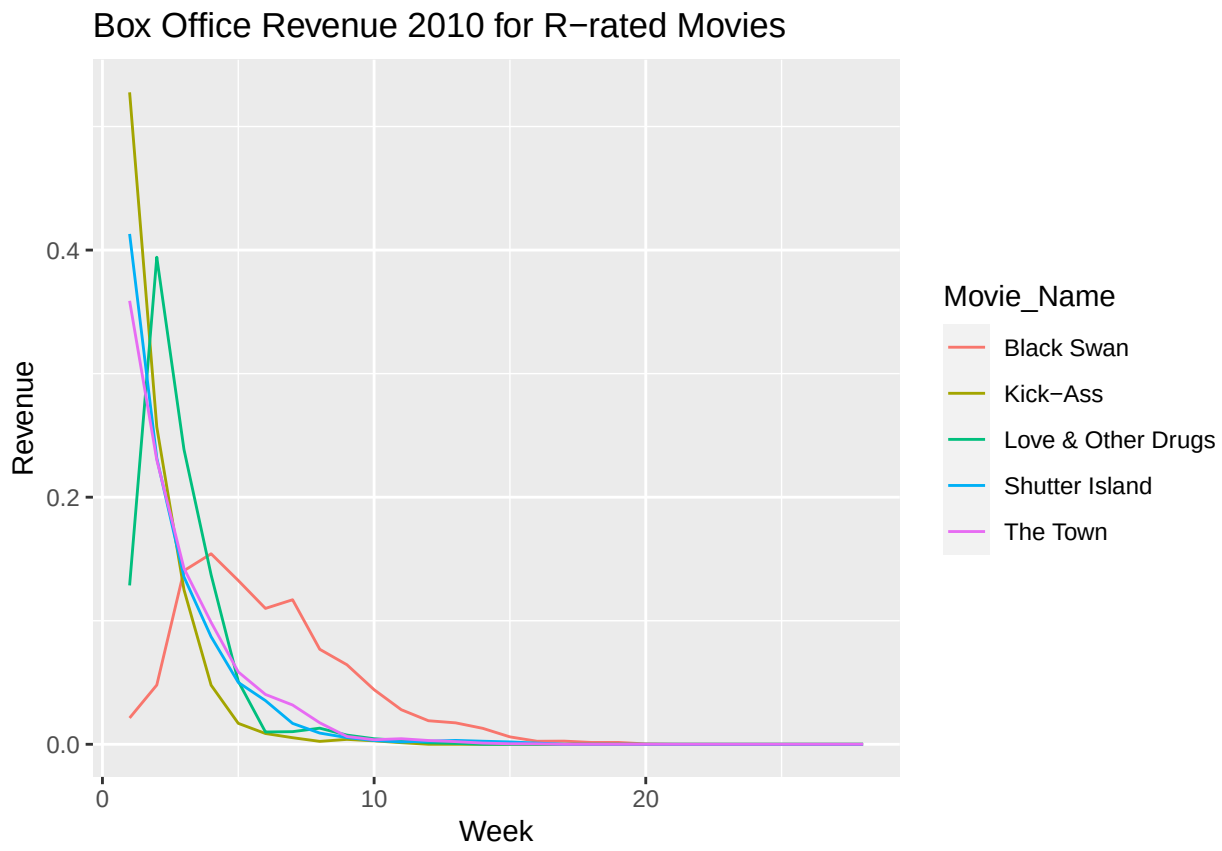

```
# Time series plot for PG-rated movies
df_PG <- df_long %>%
  filter(Rating == "PG")
ggplot(df_PG, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for PG-rated Movies")
```



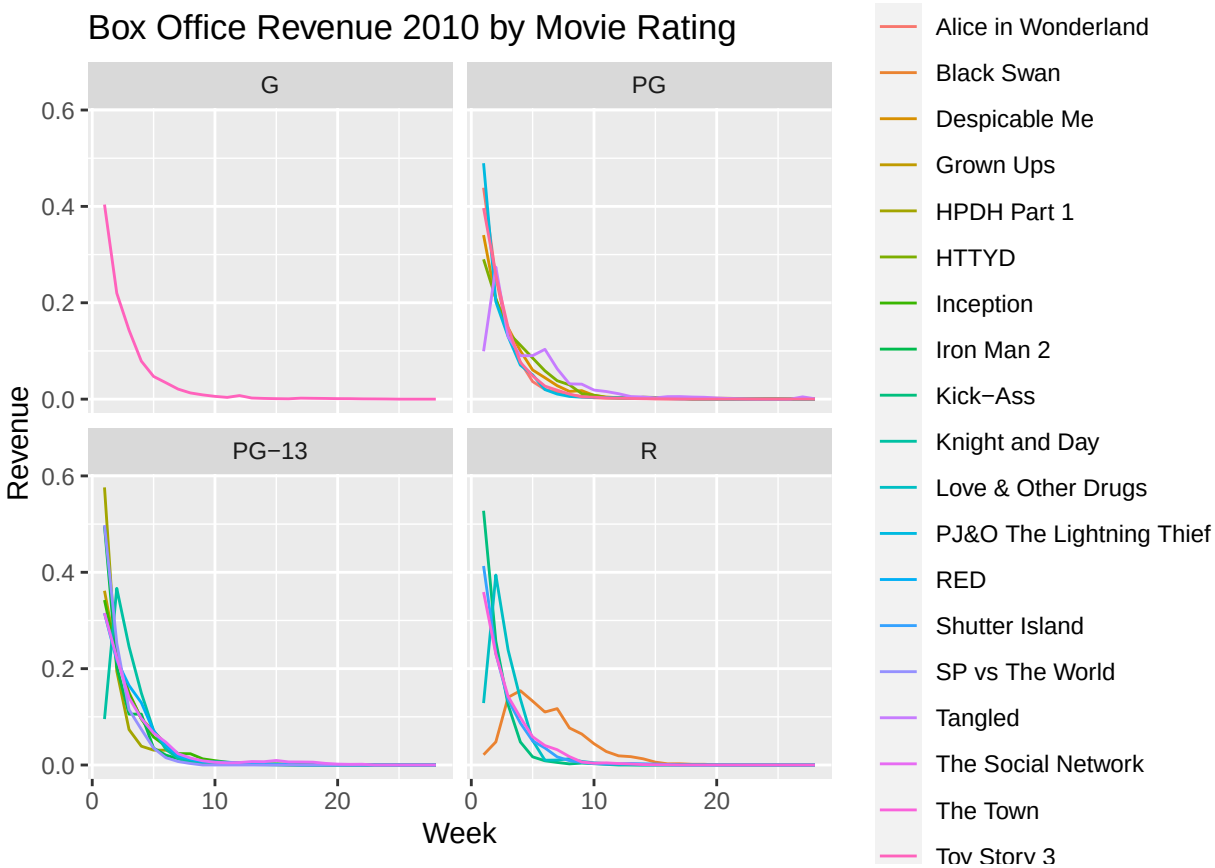

```r
# Time series plot for G-rated movies
df_PG13 <- df_long %>%
  filter(Rating == "PG-13")
ggplot(df_PG13, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for PG-13 rated Movies")
```




```r
# Time series plot for G-rated movies
df_R <- df_long %>%
  filter(Rating == "R")
ggplot(df_R, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for R-rated Movies")
```



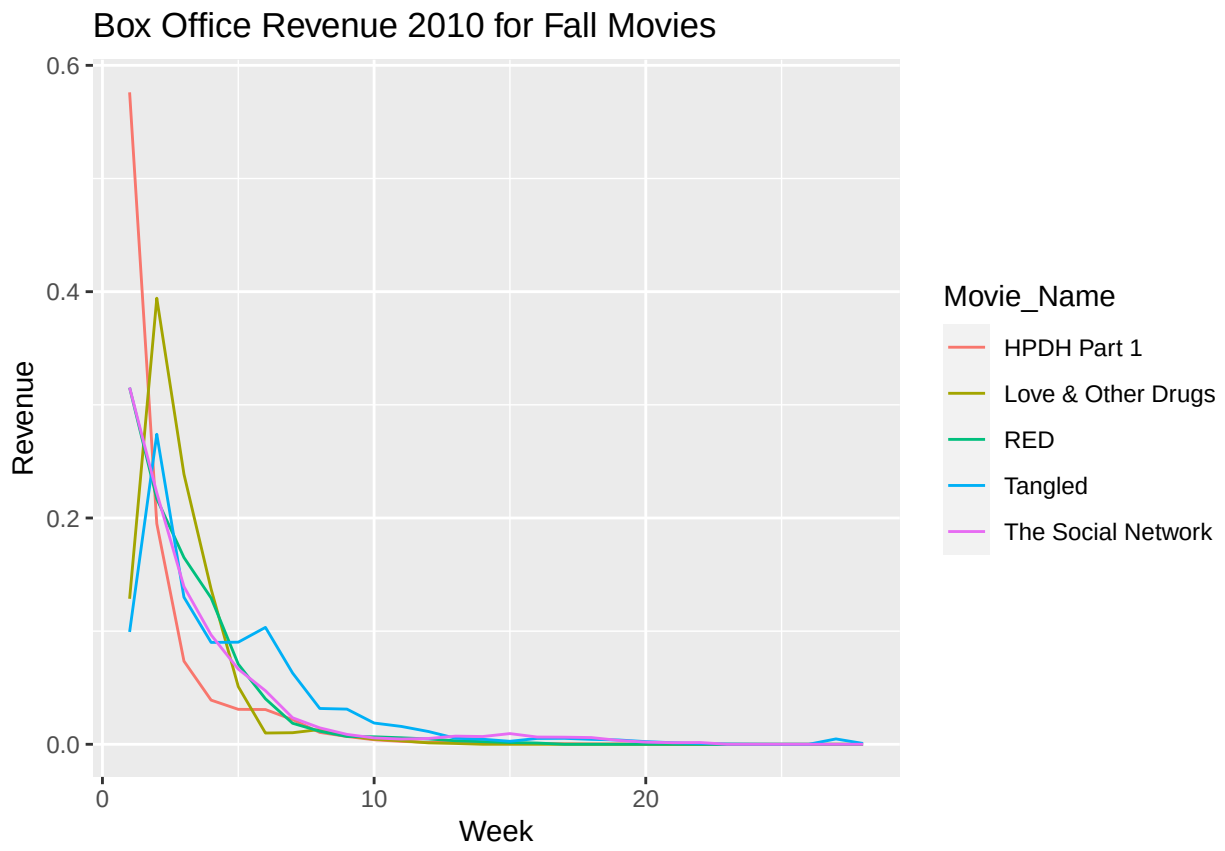

```
# Time series plot by Ratingss
ggplot(df_long, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  facet_wrap(~ Rating) +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 by Movie Rating")
```



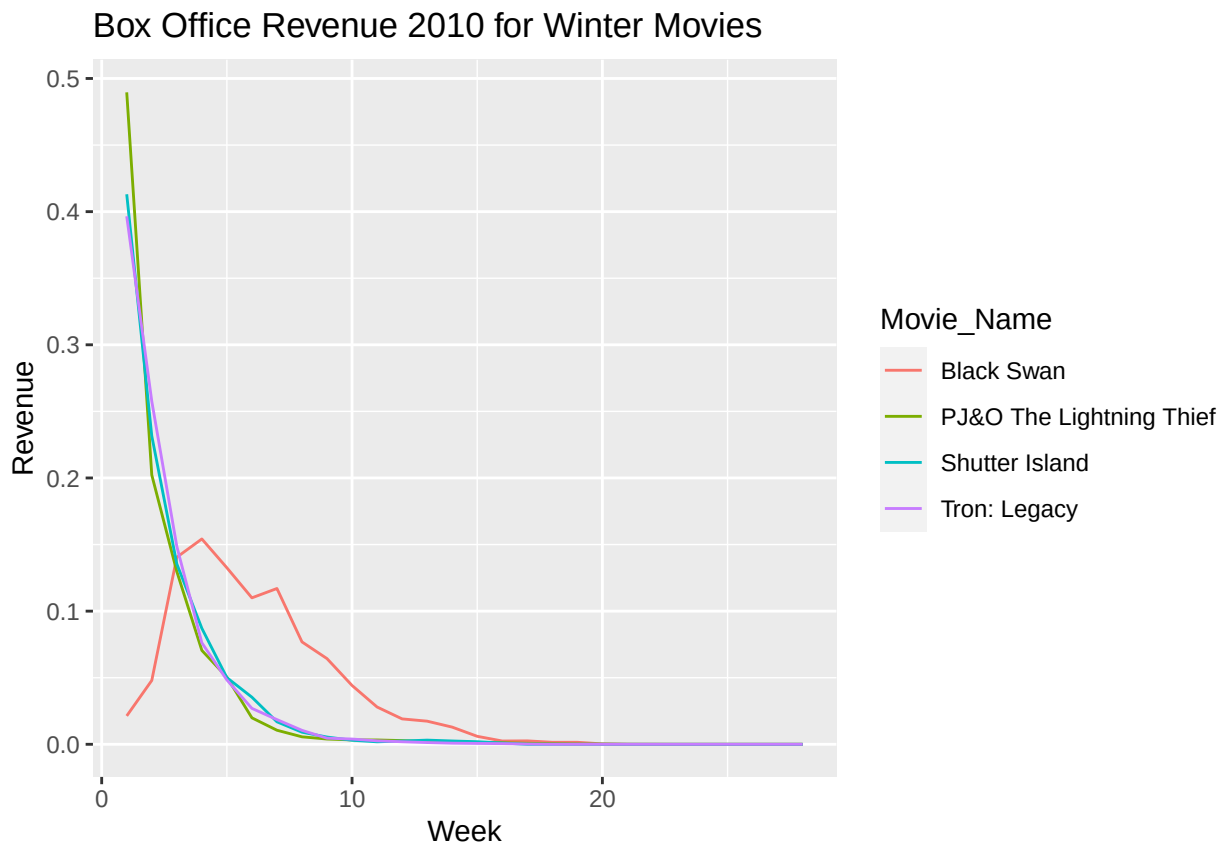
Time Series Plots for Season Release

Seasons can make an influence on movie releases, like scary movies mostly being released during the Fall season because of Halloween. Or even majority of animated movies being released over the spring or summer since children will be nearly out of school and find time to watch it. In this segment of code, we analyze the weekly revenues on the movies that were released based on the season they were released in. The Winter and Spring plot has four movies each that made revenues in an equivalent manner. The Fall plot has five movies though the Harry Potter movie dominated the most in revenue compared to the other seasons. The Summer has more movies in all range of genres and generated much in revenue.

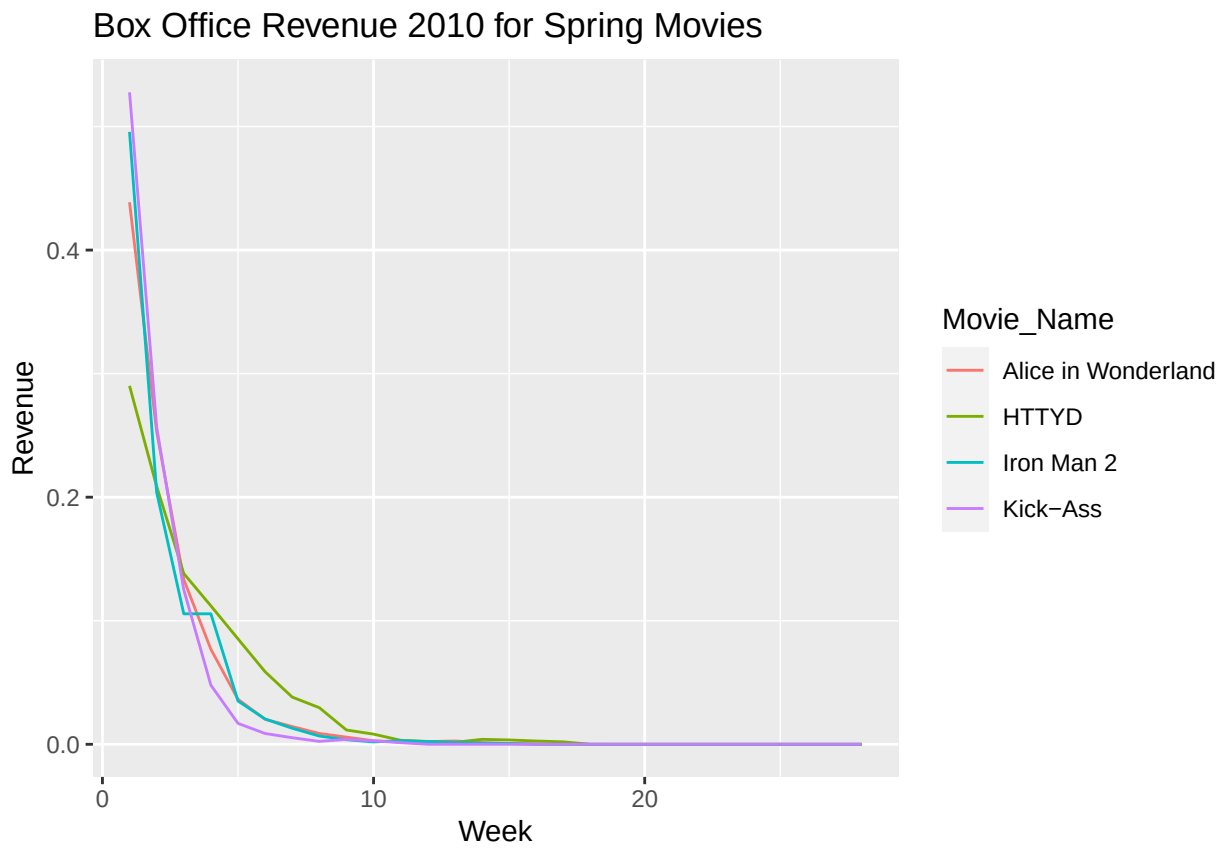
```
# Time series plot for movies released in Fall
df_Fall <- df_long %>%
    filter(Season == "Fall")
ggplot(df_Fall, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for Fall Movies")
```



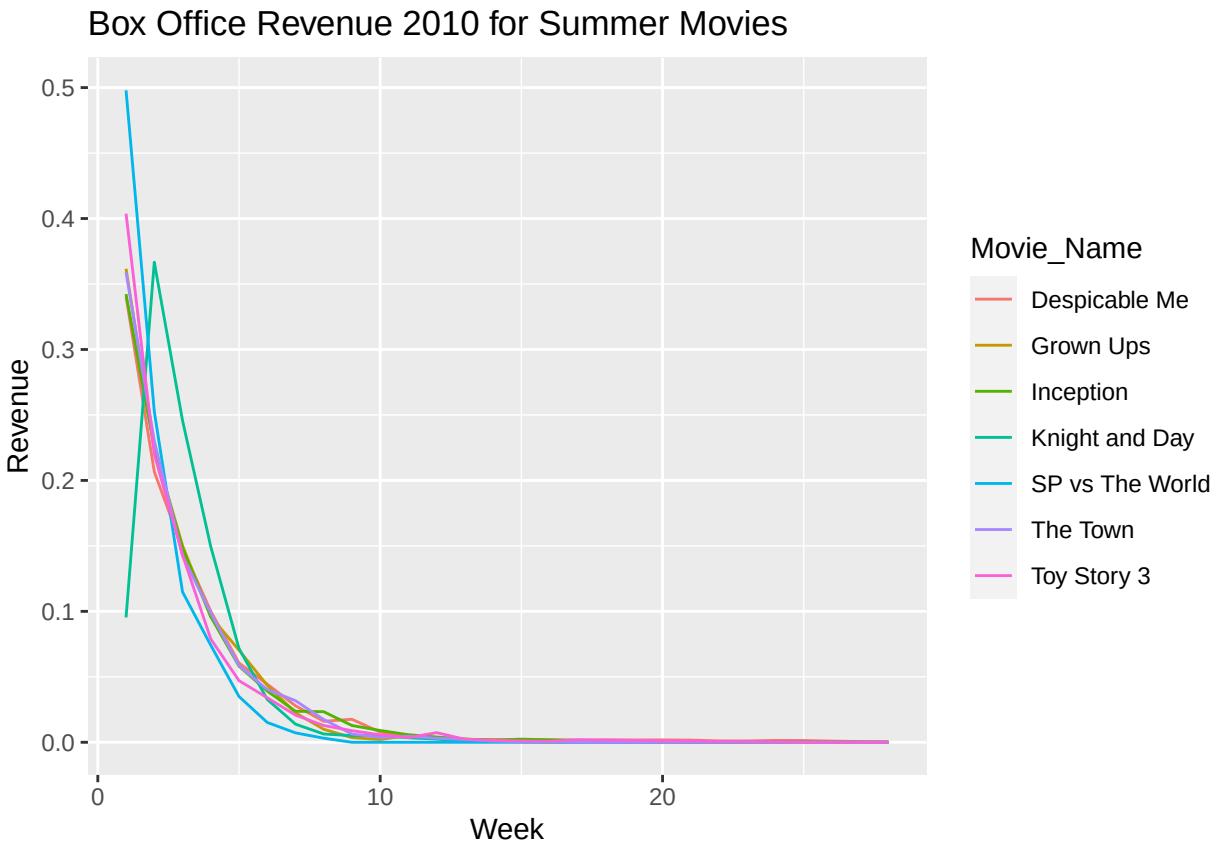

```
# Time series plot for movies released in Winter
df_Winter <- df_long %>%
  filter(Season == "Winter")
ggplot(df_Winter, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for Winter Movies")
```



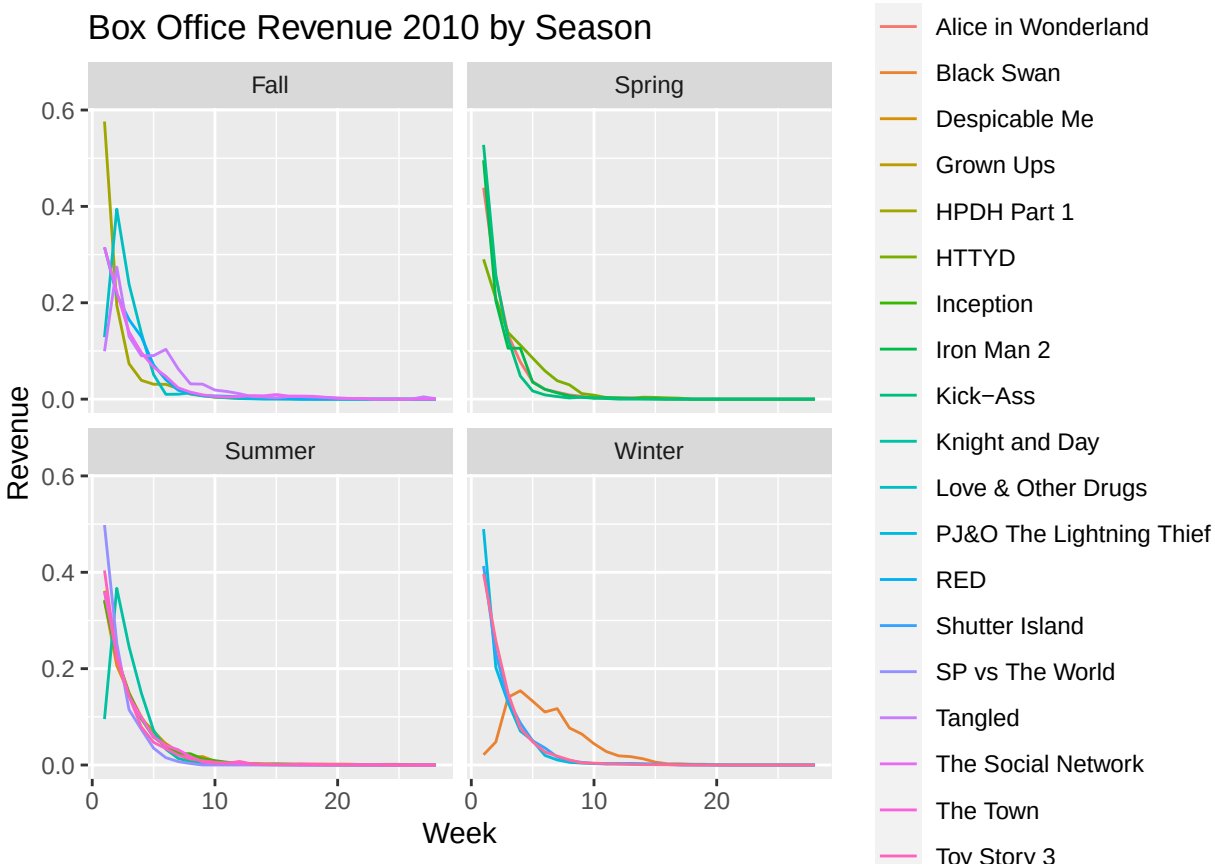

```
# Time series plot for movies released in Spring
df_Spring <- df_long %>%
  filter(Season == "Spring")
ggplot(df_Spring, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for Spring Movies")
```




```
# Time series plot for movies released in Summer
df_Summer <- df_long %>%
  filter(Season == "Summer")
ggplot(df_Summer, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 for Summer Movies")
```



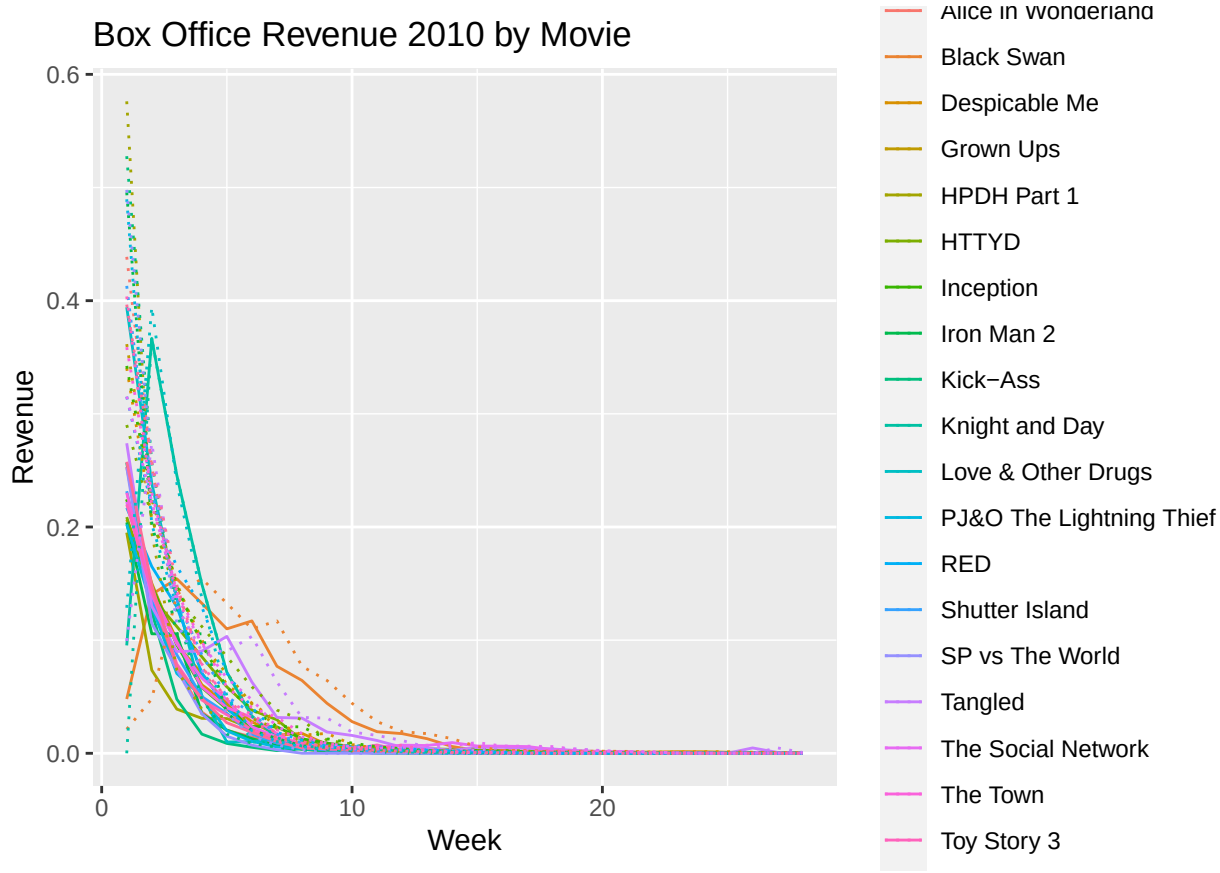

```
# Time series plot by Seasons
ggplot(df_long, aes(x = as.numeric(week), y = revenue, color = Movie_Name)) +
  geom_line() +
  facet_wrap(~ Season) +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 by Season")
```



Time Series Plot on RevenueX and RevenueY

This is the last plot that we coded where it takes the RevenueX (current week of the movie being a solid line) and RevenueY (following week of the movie being a dotted line) from the table "df_long". There were some significance and analyzation that we come across from this plot among with others. Based on the graph, there were some correlation on how the weeks went by just distinguishing the two revenues where it was able to predict on how well a movie could make as the weeks went by.

```
ggplot(df_long, aes(x = as.numeric(week))) +
  geom_line(aes(y = RevenueX, color = Movie_Name), linetype = "solid") +
  geom_line(aes(y = RevenueY, color = Movie_Name), linetype = "dotted") +
  labs(x = "Week", y = "Revenue", title = "Box Office Revenue 2010 by Movie") +
  scale_color_discrete(name = "Movie Name")
```



Literature Review

Predicting Movie Revenue from IMDb Data, by Steven Yoo, Robert Kanter, David Cummings TA: Andrew Maas. These are students at Stanford who created a model that tackles the following question - "Given the information known about a movie in the week of its release, can we predict the total gross revenue for that movie?" They eventually found that the movie's budget is the strongest individual indicator of a movies eventual gross revenue. The students compared a linear regression model and a logistic regression model. They found that the models were very similar but the features they used were insufficient to make strong predictions of gross revenue.

https://cs229.stanford.edu/proj2011/YooKanterCummings-PredictingMovieRevenuesUsingImdbData.pdf

Summary and Analysis

In concluding our study we observed several patterns in our time series model. Movies that are released in the Summer tend to stay in theaters longer and in turn generate more revenue. The most successful movies were movies rated PG and PG-13 and this can be attributed to the wide audience that these movies target. The most obvious conclusion from our study is that the most important factor in a movies success is the overall product and execution of the production. This is an easy conclusion to make because of the fact that many movies from different ratings to release dates found success. This study would be able to find stronger patterns/trends using a larger data set and by collecting data from movie releases across many different years instead of just one - which would be beneficial for the film production industry to make clear decision on what movies to make that will target audience and reach a level of success in the box office.

rmarkdown::render("Stochastic_Box.Rmd", output_format = "pdf_document")