IMDB Movies Data Analysis: Group Final Project

Introduction

Our goal with this project was to understand how the public's perception of a film was related to the film's length, the film's genre, and the film's box office revenue. In order to achieve this, we decided to use the data from a popular film rating website known as the Internet Movie Database, otherwise known as IMDB. IMDB is a vast and widely-used resource, and users can easily rate the film on the site as well as learn relevant information about the film's production and release. The IMDB is easily accessible to any netizen, as is the dataset we are using which can be found on Kaggle. Our dataset was last updated 2 years ago, and the user who scraped the data from the IMDB website received a license to do so. IMDB has a weighted average rating system, and we will be utilizing these weighted averages. In addition to the weighted average, we can access data on how many people rated the film and what the true average rating is for any given film.

Methodology

While it is great that the films and ratings added to IMDB are all user submitted, it also means that films can sometimes be duplicated within the database by accident. To circumvent this, we had to use the duplicated function within a filter function. We also decided as a group that we wanted films of the 21st century, since we felt that users would have left more ratings on films that had been released recently.

One important aspect of our data analysis was determining whether we wanted to have a film count only for their primary genre or for all genres it falls under. We opted for the former, as we felt that if the film qualified to fall under the genre, then the tropes or expectations of the genre would impact the people's perception of the film regardless of the other genres the film fell under.

When we originally read the data, the genres were written in the form of a well-formatted list. Unfortunately for us, a well-formatted list means poorly-formatted data. So, when we had to separate out the genres into individual columns using the comma as a delimiter, some of our genres had a space in front of them. We had to use the case_match function within a mutate to change the inner values so they would be properly compiled later.

Exploratory Data Analysis

First, we wrangled and cleaned our dataset. This process was to make the data more readable and easier to analyze. Additionally, we further wrangled the movies' genres and a few movie franchises into their own data frames so we could consider them as factors in our analysis of the relationship between revenue, average rating, genre, and runtime.

Quantitative

In our quantitative analysis process, we took several combinations of the quantitative factors in our research questions (runtime, average rating, and revenue) and ran visualizations and regression to understand the relationship between every possible combination.

Revenue vs. Average Rating

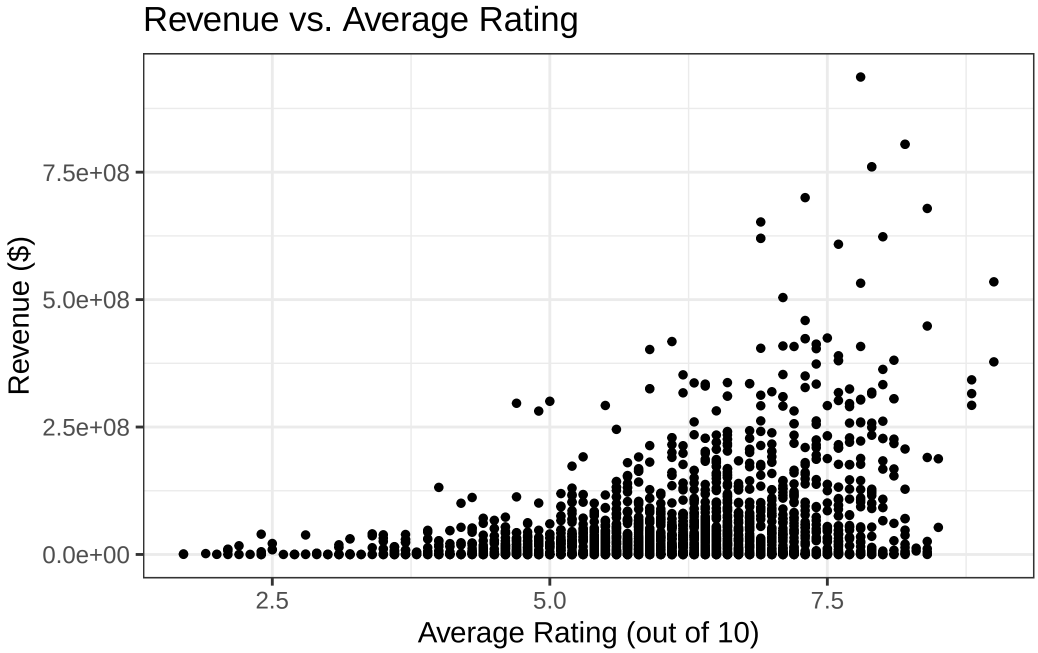


Figure 1

This visualization depicts the relationship between runtime and average rating, which seems to have a positive and linear correlation. Most of the data is clustered towards the center of the graph because it is typical that a movie will run between a little under 100 mins to 150 mins. As a result, since a good portion of movies have that runtime, most of the movies will have a typical rating of anywhere from 5.0 to 7.5. Both of the observable medians are being described through the middle of the graph.

Runtime vs. Average Rating

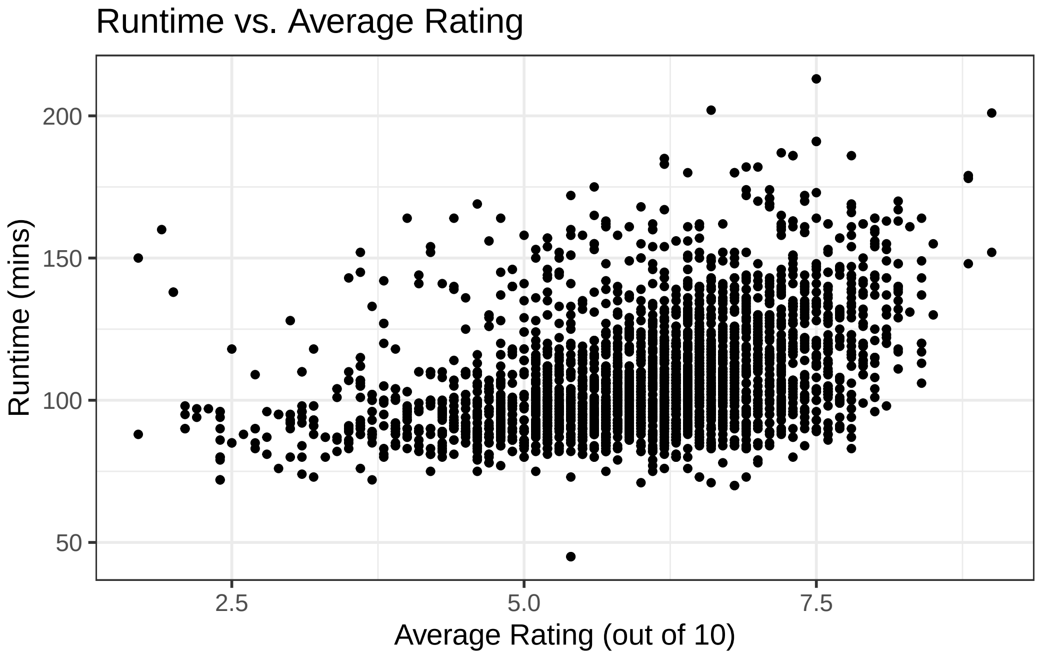


Figure 2

This visualization depicts the relationship between runtime and average rating, which seems to have a positive and linear correlation. Most of the data is clustered towards the center of the graph because it is typical that a move will run between a little under 100 mins to 150 mins. As a result, since a good portion of movies have that runtime, most of the movies will have a typical rating of anywhere from 5.0 to 7.5. Both of the observable medians are being described through the middle of the graph.

Revenue vs. Runtime

This visualization shows the relationship between revenue and runtime, and though there seems like no there is no linear correlation, but there is a curve similar to a bell. This is understandable as many people do not want to pay for an extremely short or long movie, so

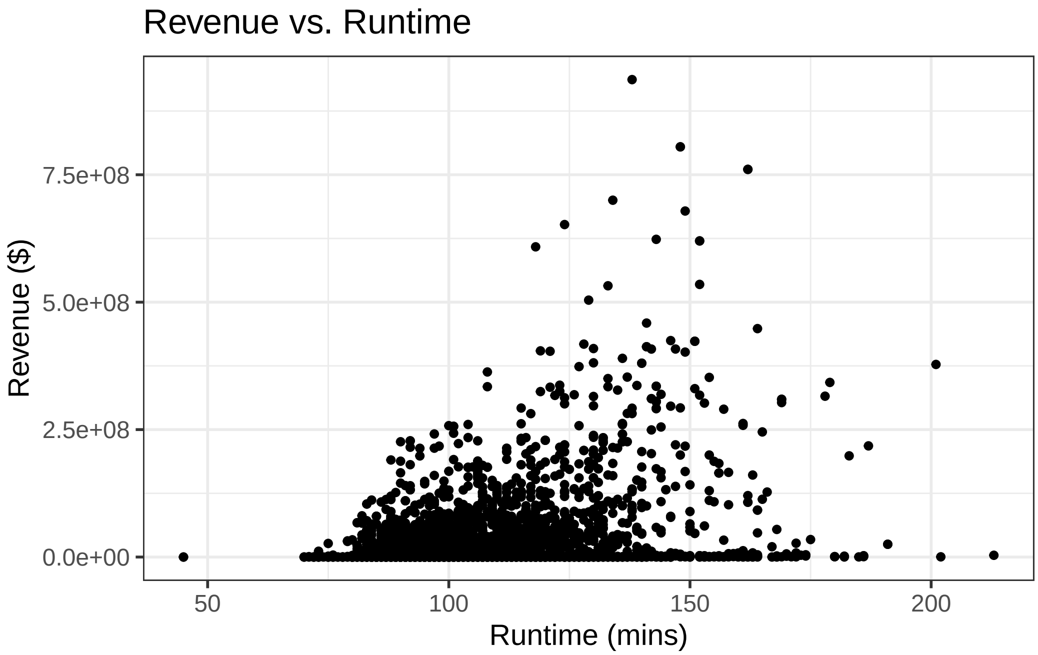


Figure 3

the movies that seemed to make the most money were in that typical less than 100 min to 150 min runtime. This could be also due to an external factor like budget affecting the quality of the movie, because larger movies could afford high-quality editing.

Regression Tables for Quantitative Graphs

```
Warning in data(relevantMovies): data set 'relevantMovies' not found
```

```
===============================================================================
                                             Dependent variable:
                                -----------------------------------------------
                                     revenue       runtime        revenue
                                       (1)           (2)            (3)
-------------------------------------------------------------------------------
rating                          21,513,322.000***  7.120***
                                 (1,195,693.000)   (0.286)

runtime                                                       1,263,630.000***
                                                                (66,892.640)
```

```
Constant                              -89,212,419.000*** 65.319***  -96,362,559.000***
                                       (7,322,218.000)    (1.750)    (7,356,657.000)

-------------------------------------------------------------------------------------
Observations                                3,258          3,258          3,258
R2                                          0.090          0.160          0.099
Adjusted R2                                 0.090          0.160          0.098
Residual Std. Error (df = 3256)       75,441,810.000      18.030     75,095,180.000
F Statistic (df = 1; 3256)                323.724***     620.794***     356.848***
=====================================================================================
Note:                                                    *p<0.1; **p<0.05; ***p<0.01
```

This regression table depicts the relationship between each of the variables in the visualizations depicted above.

Revenue vs. Runtime with Genre and Average Rating

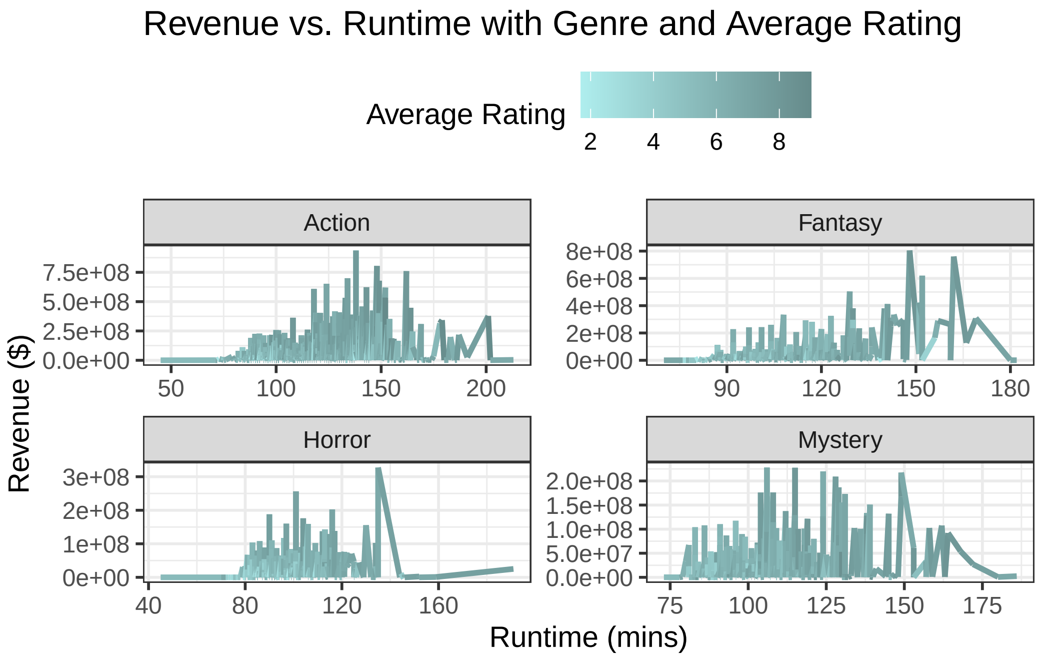


Figure 4

This final visualization is 4D, depicting the relationship between the genre, runtime, revenue, and average rating. Action seems to have the most bell-curve shape, with fantasy, horror, and mystery having large plateaus and peaks throughout their graphs. Horror seemed to have the most movies with the lowest ratings, and mystery and action had several high-rated movies. Additionally, though it seems that mystery had several movies that made a lot of money, the values on the y-axis are significantly smaller than those on the action and fantasy graphs. This means that fantasy must have had the highest-grossing movies, though action has more movies that made consistently more money. Horror and mystery are both movies that have had the shortest runtime, which is understandable as action and fantasy typically requires more budget, and as a result, can afford longer runtime. All in all, there is definitely a relationship between all four variables, because it is clear that there are discrepancies between each variable in every genre. This is why it was crucial to separate the movie by genre, because the values were so different for every movie in the category for the analysis to be as insightful as it currently is.

Qualitative

We also performed qualitative analyses. Our qualitative variable was genre, so we investigated how it impacted revenue, runtime, and ratings by looking at 4 popular movie genres: Action, Horror, Mystery, and Fantasy. We went on to assess how another qualitative variable, movie franchise, impacted the quantitative variables. We looked at the Harry Potter, Pirates of the Caribbean, and Spider-Man franchises to answer this question.

Summary Statistics of Movies by Genre

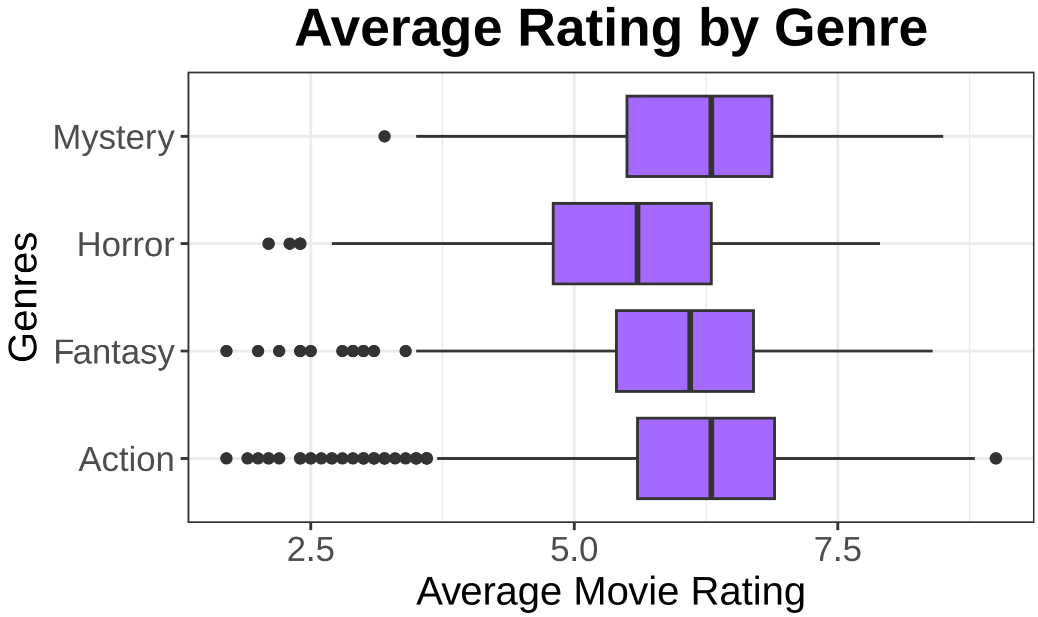
This table (Table 1) shows the means of the movies' runtimes, ratings, and revenues, grouped by genre. This supplements the following visuals to display exact values for the discussed comparisons.

Table 1: Averages of Movies' Revenue, Runtime, and Rating by Genre

Genre	Count	Average Revenue	Average Runtime	Average Rating
Action	1525	53193780	113.72262	6.173246
Fantasy	472	50022561	107.90466	5.994280
Horror	651	20107771	97.58525	5.529032
Mystery	610	22451794	105.97213	6.198033

Genre and Rating

This boxplot shows how viewers' average rating of a movie and that movie's genre are related. We can see that the medians for each genre aren't particularly distinct. There's a lot of overlap between the genres, their IQRs, variations, and more hover around the same area on the graph. Horror is skewed further left than the other genres, indicating viewer's hold generally lower opinions of horror movies, a 5.6/10.0 genre average as compared to the other's approximately 6.0/10.0. It is also important to notice that the 'Action' genre has more outliers than the other genres. (Figure 1) This could be because there are more Action movies in this data set (1525) than other genres, or it could be indicative of a trend in the genre - having some truly standout bad movies, according to viewers.



A box plot comparing the movies' average voter ratings and the movies' genres.

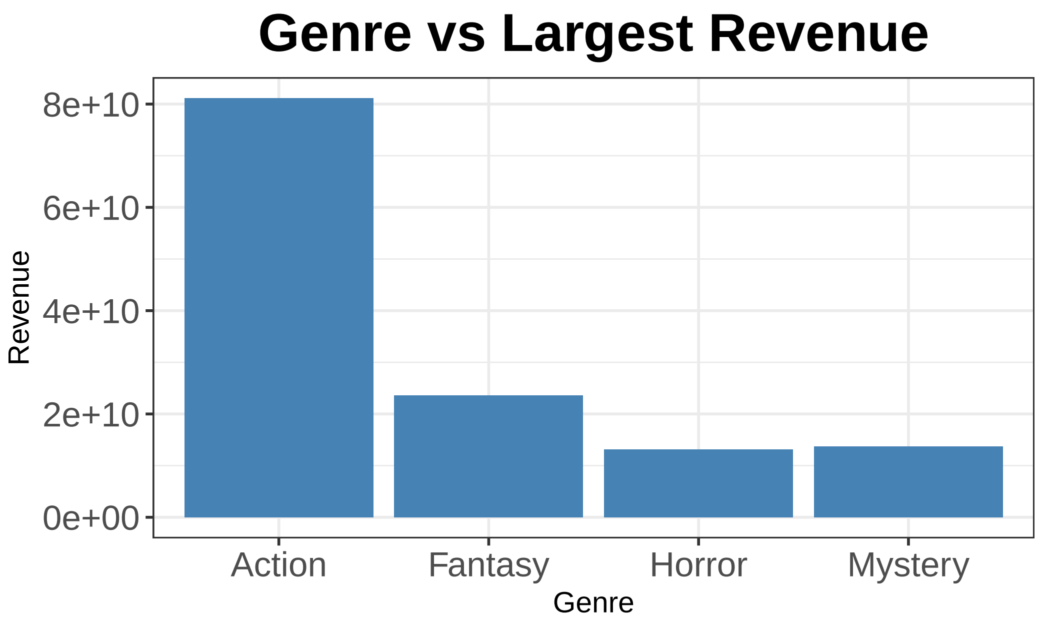
Figure 5

Genre and Revenue

This bar graph shows the largest revenues from each genre, with Action being the largest of these four. Action is nearly four times as large as the next genre down, fantasy. Horror had the least revenue of these four, but it is not far below mystery. Genre clearly has an impact on movie's revenue. (Figure 2) Many action movies are highly anticipated and have large fanbases (Marvel, Star Wars, etc.) so this trend makes sense to see.

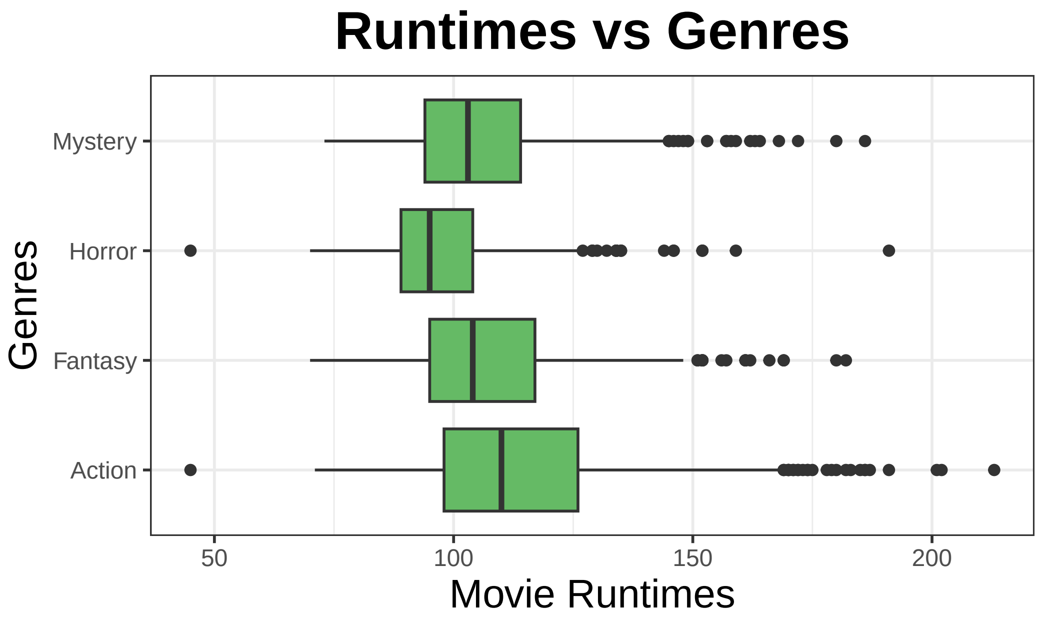
Genre and Runtime

This boxplot shows how movies' runtimes are impacted by their genre. Action movies had the highest median run time, and horror the lowest. There is overlap between all four genres, but there are some clearer trends here. Action movies have more variability in length, and tend to be longer. Fantasy and mystery are similar in their spread, with mystery having more right skewed outliers. Horror movies have the least variability and tend to be shorter. (Figure 3)



A bar plot comparing movies' largest revenues within four genres.

Figure 6

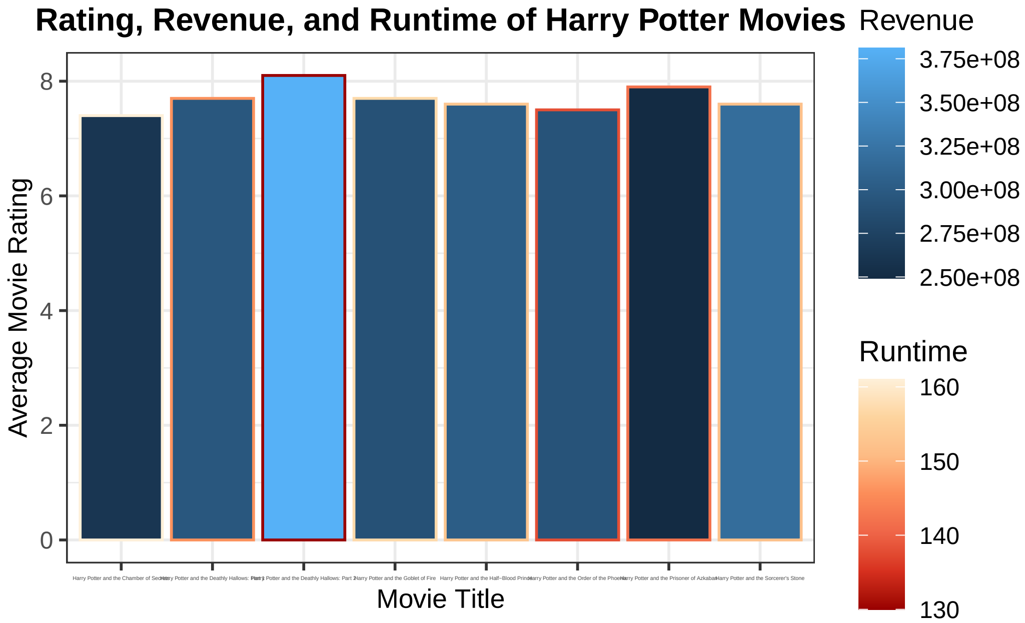


A box plot comparing movies' runtimes and their genres.

Figure 7

Harry Potter Movies

This bar graph displays different Harry Potter movies' average viewer rating, revenue, and runtime. Chamber of Secrets is the lowest rated movie, had the second lowest revenue, and the longest runtime. Interestingly, this correlation holds when looking at the opposite ends of these scales. Deathly Hallows pt. 2 is the highest rated movie, the highest revenue, and the shortest runtime. (Table 2, Figure 4) It is important to note that these movies all have relatively similar runtimes and ratings, so these are likely not significant correlations.



Harry Potter franchise by their average voter ratings, runtimes, and revenues.

Figure 8

Table 2: Summary Statistics of Harry Potter Movies

Movie Title	Year	Runtime	Rating	Votes	Revenue
Harry Potter and the Sorcerer's Stone	2001	152	7.6	792448	317575550
Harry Potter and the Goblet of Fire	2005	157	7.7	636066	290013036
Harry Potter and the Deathly Hallows: Part 2	2011	130	8.1	885463	381011219
Harry Potter and the Prisoner of Azkaban	2004	142	7.9	644151	249358727
Harry Potter and the Chamber of Secrets	2002	161	7.4	645233	261988482
Harry Potter and the Order of the Phoenix	2007	138	7.5	590824	292004738
Harry Potter and the Deathly Hallows: Part 1	2010	146	7.7	557701	295983305

Harry Potter and the Half-Blood Prince	2009	153	7.6	554044	301959197

Pirates of the Caribbean Movies

This bar graph displays different Pirates of the Caribbean movies' viewer rating, revenue, and runtime. The first movie in the series, The Curse of the Black Pearl, is the highest rated movie in the franchise, but this doesn't seem to positively or negatively correlate with revenue or runtime (Table 3, Figure 5). However, the lowest rated movie, Dead Men Tell No Tales, also had the lowest revenue and the shortest runtime. This is a departure from what we saw with Harry Potter, where the best rated movie had the shortest runtime. There is also more variation in voter rating in this franchise than in Harry Potter.

Table 3: Summary Statistics of Pirates of the Caribbean Movies

Movie Title	Year	Runtime	Rating	Votes	Revenue
Pirates of the Caribbean: The Curse of the Black Pearl	2003	143	8.1	1137362	305413918
Pirates of the Caribbean: Dead Men Tell No Tales	2017	129	6.5	317985	172558876
Pirates of the Caribbean: Dead Man's Chest	2006	151	7.3	727169	423315812
Pirates of the Caribbean: On Stranger Tides	2011	136	6.6	534586	241063875
Pirates of the Caribbean: At World's End	2007	169	7.1	658765	309420425

Spider-Man Movies

This bar graph displays different Spider-Man movies' viewer rating, revenue, and runtime. This franchise spans mediums and decades, both of which could play an important role in setting the films apart from one another in these categories. The highest rated movie is Into the Spiderverse, the only animated movie in this analysis. However, it has the lowest revenue and the shortest runtime. The lowest rated movie was Spider-Man 3, but this has no clear correlation with any other variables. (Table 4, Figure 6) This is once again a departure from trends the previous franchise had between these variables, with the exception of the relationship between Harry Potter's highest rated movie and shortest runtime.

Table 4: Summary Statistics of Spider-Man Movies

Movie Title	Year	Runtime	Rating	Votes	Revenue
Spider-Man: No Way Home	2021	148	8.2	770492	804747988
Spider-Man: Into the Spider-Verse	2018	117	8.4	541462	190241310
Spider-Man	2002	121	7.4	826095	403706375
Spider-Man: Homecoming	2017	133	7.4	666223	334201140

Spider-Man 3	2007	139	6.3	598454	336530303
The Amazing Spider-Man	2012	136	6.9	663418	262030663
Spider-Man 2	2004	127	7.4	661539	373585825
The Amazing Spider-Man 2	2014	142	6.6	504903	202853933

Results and Conclusion

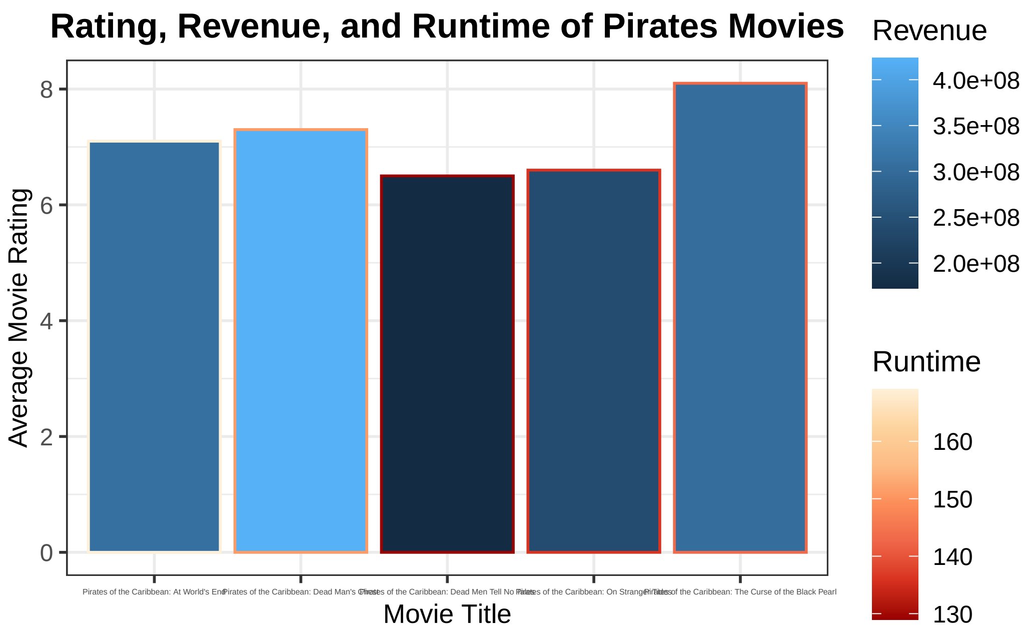
Throughout both quantitative and qualitative EDA, we were able to note the relationship between genre, voting average, revenue, and runtime of movies rated on IMDB. It was crucial to run our analysis in parts like how we did above, as there were several confounding variables like the popularity of the genre or the audience's disinterest in watching long movies. Taking apart movies into famous franchises provided extra analysis within small communities of movies, which was important for minimizing any possible external factors. Within the three chosen movie franchises, correlations between our quantitative variables were unclear but has shown that this kind of testing can be done to further analyze the impact of franchises, directors, etc. Overall, action and fantasy had the highest grossing movies, while the longer the movies were, the lower their rating was. Several notes could be made by every combination of each of the four factors, but this EDA supplemented our questions about the factors that made a movie successful. It also allowed us to evaluate movies within franchises for similarities and differences both with the franchise groups and between them. Within the three chosen movie franchises, correlations between our quantitative variables were unclear but has shown that this kind of testing can be done to further analyze

By using this EDA, movie makers can understand how to optimize their movie in order to get the most revenue and rating by critics on several platforms such as IMDB and Rotten Tomatoes. This would reduce the amount of movies that would not make enough profit, and keep the quality of the film industry high. The trends, as we found in the analysis, though mystery and horror may become more profitable genres with more exposure. All in all, understanding these four factors and their role in creating successful movies will help keep the integrity of films high and make sure that making movies isn't as daunting as it may seem.

References

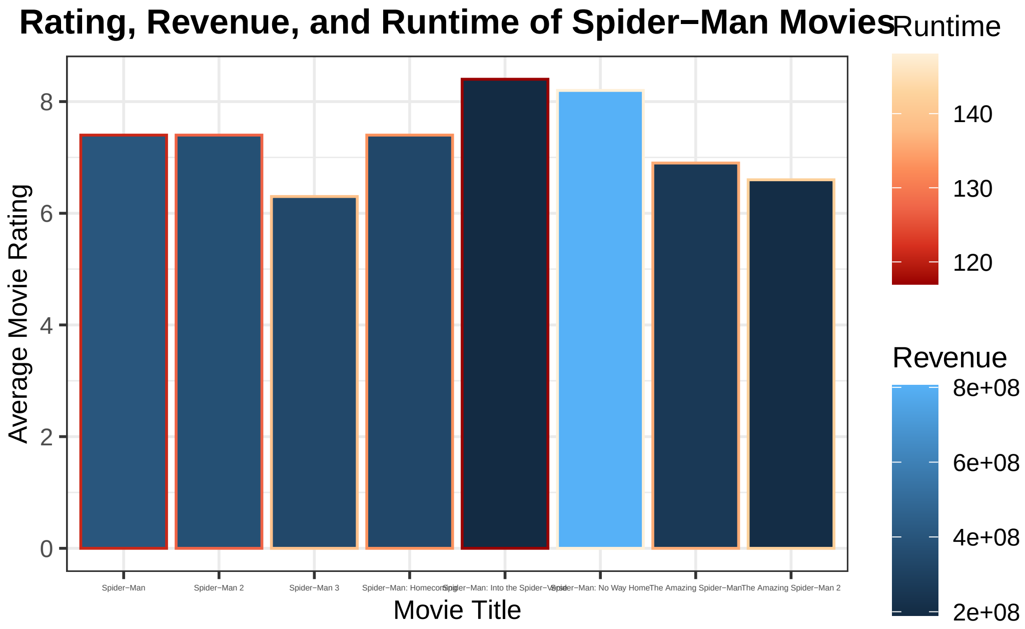
Chidambara, R. (2022). IMDb Movie Dataset: All Movies by Genre [Data set]. https://www.kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre

Code Appendix



A bar plot comparing movies in the Pirates of the Caribbean franchise by their average voter ratings, runtimes, and revenues.

Figure 9



A bar plot comparing movies in the Spider−Man franchise by their average voter ratings, runtimes, and revenues.

Figure 10

```r
library(janitor)
library(knitr)
library(kableExtra)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidyr)
library(rvest)
library(plotly)
library(esquisse)
library(readr)
library(jtools)
library(stargazer)
##Importing the Data----
fantasyRaw <- read_csv(
  file = "~/Desktop/STAT184/fantasy.csv"
)

actionRaw <- read_csv(
  file = "~/Desktop/STAT184/action.csv"
)

horrorRaw <- read_csv(
  file = "~/Desktop/STAT184/horror.csv"
)

mysteryRaw <- read_csv(
  file = "~/Desktop/STAT184/mystery.csv"
)

##Merging the Data----
moviesRaw <- full_join(
  x = fantasyRaw,
  y = actionRaw
) %>%
  full_join(
    y = horrorRaw
  ) %>%
  full_join(
    y = mysteryRaw
  )
```

```
##Cleaning the Data----
moviesCleaned <- moviesRaw %>%
  rename(revenue = `gross(in $)`
  ) %>%
  dplyr:: select(-movie_id, -description, -director_id, -star_id
  ) %>%
  drop_na() %>%
  filter(!grepl('19', year)) %>%
  filter(!duplicated(movie_name)) %>%
  mutate(runtime = readr::parse_number(runtime))

##Listing Only Relevant Movies----
relevantMovies <- moviesCleaned %>%
  separate_wider_delim(
    cols = genre,
    delim = ",",
    names = c("Genre1", "Genre2", "Genre3"),
    too_few = "align_start"
  ) %>%
  pivot_longer(
    cols = starts_with("Genre"),
    names_to = "genreNumber",
    values_to = "genre"
  ) %>%
 mutate(genre = case_match(
   genre,
   " Action" ~ "Action",
   " Mystery" ~ "Mystery",
   " Fantasy" ~ "Fantasy",
  " Horror" ~ "Horror",
  .default = genre
 )) %>%
  drop_na() %>%
  filter(
    genre == "Action" |
    genre == "Horror" |
    genre == "Mystery" |
    genre == "Fantasy") %>%
  select(-genreNumber)
```

```r
##Getting Summary Statistics----
info <- list(
  Count = ~as.double(n()),
  Min = ~as.double(min(.x)),
  Q1 = ~as.double(quantile(.x,probs = 0.25, na.rm = TRUE)),
  Median = ~as.double(median(.x)),
  Avg = ~as.double(mean(.x)),
  Q3 = ~as.double(quantile(.x,probs = 0.75, na.rm = TRUE)),
  Max = ~as.double(max(.x))
)

moviesSummary <- relevantMovies %>%
  group_by(genre) %>%
  summarize(across(c(revenue,runtime,rating), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

##Film Franchises----
harryPotterMovies <- relevantMovies %>%
  filter(grepl('Harry Potter', movie_name)) %>%
  select(-star, -genre, -director, -certificate)

harryPotterSummary <- harryPotterMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

piratesMovies <- relevantMovies  %>%
  filter(grepl('Pirates of the Caribbean:', movie_name)) %>%
  select(-star, -genre, -director, -certificate) %>%
  filter(!duplicated(movie_name))

piratesSummary <- piratesMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

spiderMovies <- relevantMovies  %>%
  filter(grepl('Spider-Man', movie_name)) %>%
  select(-star, -genre, -director, -certificate) %>%
  filter(!duplicated(movie_name))

spiderSummary <- spiderMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = rating,
    y = revenue
  )
)+
  geom_point(size = 1)+
  labs(
    x = "Average Rating (out of 10)",
    y = "Revenue ($)",
    title = "Revenue vs. Average Rating"
  )+
  theme_bw()
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = rating,
    y = runtime
  )
)+
  geom_point(size = 1)+
  labs(
    x = "Average Rating (out of 10)",
    y = "Runtime (mins)",
    title = "Runtime vs. Average Rating"
  )+
  theme_bw()
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = runtime,
    y = revenue
  )
```

```r
)+
  geom_point(size = 1)+
  labs(
    x = "Runtime (mins)",
    y = "Revenue ($)",
    title = "Revenue vs. Runtime"
  )+
  theme_bw()
data(relevantMovies)
revenue_rating_regression <- lm(revenue ~ rating, data = relevantMovies)
runtime_rating_regression <- lm(runtime ~ rating, data = relevantMovies)
revenue_runtime_regression <- lm(revenue ~ runtime, data = relevantMovies)

stargazer(revenue_rating_regression,runtime_rating_regression,revenue_runtime_regression,  ty
scatterplot <- ggplot(
  data = relevantMovies,
  mapping = aes(
    x = runtime,
    y = revenue,
    color = rating,
    #color = genre,
    #shape = averageRating
  )
)+
  geom_line(linewidth = 1)+
  labs(
    x = "Runtime (mins)",
    y = "Revenue ($)",
    color = "Average Rating",
    #color = "Genre",
    #shape = "Average Rating
    title = "Revenue vs. Runtime with Genre and Average Rating"
  )+
  scale_color_gradient(low = "#AFEEEE", high = "#668B8B")
facet_scatter <- scatterplot + facet_wrap(~genre, scales = "free")+
  theme_bw()+
  theme(
    legend.position = "top"
  )
print(facet_scatter)
moviesSummary %>%
  select(genre, count, revenue_Avg, runtime_Avg, rating_Avg) %>%
  kable(
    caption = "Averages of Movies' Revenue, Runtime, and Rating by Genre",
    col.names = c("Genre", "Count", "Average Revenue", "Average Runtime", "Average Rating")
  ) %>%
  kableExtra::kable_classic()
ggplot(relevantMovies) +
  aes(x = rating, y = genre) +
  geom_boxplot(fill = "#A569FF") +
  labs(
    x = "Average Movie Rating",
    y = "Genres",
    title = "Average Rating by Genre",
    caption = "A box plot comparing the movies' average voter ratings and the movies' genres
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 15L),
    axis.title.x = element_text(size = 15L),
    axis.text.y = element_text(size = 13L),
    axis.text.x = element_text(size = 13L)
  )
ggplot(relevantMovies) +
  aes(x = genre, y = revenue) +
  geom_bar(stat = "summary", fun = "sum", fill = "#4682B4") +
  labs(
    x = "Genre",
    y = "Revenue",
    title = "Genre vs Largest Revenue",
    caption = "A bar plot comparing movies' largest revenues within four genres."
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.text.y = element_text(size = 13L),
    axis.text.x = element_text(size = 13L)
  )
ggplot(relevantMovies) +
```

```r
  aes(x = runtime, y = genre) +
  geom_boxplot(fill = "#65BA65") +
  labs(
    x = "Movie Runtimes",
    y = "Genres",
    title = "Runtimes vs Genres",
    caption = "A box plot comparing movies' runtimes and their genres."
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 15L),
    axis.title.x = element_text(size = 15L)
  )
ggplot(harryPotterMovies) +
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
    title = "Rating, Revenue, and Runtime of Harry Potter Movies",
    caption = "A bar plot comparing movies in the Harry Potter franchise by their average vo
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 12L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 2L)
  )
harryPotterMovies %>%
  kable(
    caption = "Summary Statistics of Harry Potter Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()
ggplot(piratesMovies) +
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
    title = "Rating, Revenue, and Runtime of Pirates Movies",
    caption = "A bar plot comparing movies in the Pirates of the Caribbean franchise by thei
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 13L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 3L),
    plot.caption = element_text(size = 5L)
  )
piratesMovies %>%
  kable(
    caption = "Summary Statistics of Pirates of the Caribbean Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()
ggplot(spiderMovies) +
```

```r
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
    title = "Rating, Revenue, and Runtime of Spider-Man Movies",
    caption = "A bar plot comparing movies in the Spider-Man franchise by their average voter
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 13L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 3L),
    plot.caption = element_text(size = 5L)
  )
spiderMovies %>%
  kable(
    caption = "Summary Statistics of Spider-Man Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()
##Load Packages
library(janitor)
library(knitr)
library(kableExtra)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidyr)
library(rvest)
library(plotly)
library(esquisse)
library(readr)
library(jtools)
library(stargazer)

##Importing the Data----
fantasyRaw <- read_csv(
  file = "~/Desktop/STAT184/fantasy.csv"
)

actionRaw <- read_csv(
  file = "~/Desktop/STAT184/action.csv"
)

horrorRaw <- read_csv(
  file = "~/Desktop/STAT184/horror.csv"
)

mysteryRaw <- read_csv(
  file = "~/Desktop/STAT184/mystery.csv"
)

##Merging the Data----
moviesRaw <- full_join(
  x = fantasyRaw,
  y = actionRaw
) %>%
  full_join(
    y = horrorRaw
  ) %>%
  full_join(
    y = mysteryRaw
  )

##Cleaning the Data----
moviesCleaned <- moviesRaw %>%
  rename(revenue = `gross(in $)`
  ) %>%
  dplyr:: select(-movie_id, -description, -director_id, -star_id
  ) %>%
  drop_na() %>%
  filter(!grepl('19', year)) %>%
  filter(!duplicated(movie_name)) %>%
  mutate(runtime = readr::parse_number(runtime))

##Listing Only Relevant Movies----
relevantMovies <- moviesCleaned %>%
  separate_wider_delim(
    cols = genre,
    delim = ",",
    names = c("Genre1", "Genre2", "Genre3"),
    too_few = "align_start"
  ) %>%
  pivot_longer(
    cols = starts_with("Genre"),
    names_to = "genreNumber",
    values_to = "genre"
  ) %>%
 mutate(genre = case_match(
   genre,
   " Action" ~ "Action",
   " Mystery" ~ "Mystery",
   " Fantasy" ~ "Fantasy",
  " Horror" ~ "Horror",
  .default = genre
 )) %>%
  drop_na() %>%
  filter(
    genre == "Action" |
    genre == "Horror" |
    genre == "Mystery" |
    genre == "Fantasy") %>%
  select(-genreNumber)



##Getting Summary Statistics----
info <- list(
  Count = ~as.double(n()),
  Min = ~as.double(min(.x)),
  Q1 = ~as.double(quantile(.x,probs = 0.25, na.rm = TRUE)),
  Median = ~as.double(median(.x)),
  Avg = ~as.double(mean(.x)),
  Q3 = ~as.double(quantile(.x,probs = 0.75, na.rm = TRUE)),
  Max = ~as.double(max(.x))
)

moviesSummary <- relevantMovies %>%
  group_by(genre) %>%
  summarize(across(c(revenue,runtime,rating), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

##Film Franchises----
harryPotterMovies <- relevantMovies %>%
  filter(grepl('Harry Potter', movie_name)) %>%
  select(-star, -genre, -director, -certificate)

harryPotterSummary <- harryPotterMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

piratesMovies <- relevantMovies  %>%
  filter(grepl('Pirates of the Caribbean:', movie_name)) %>%
  select(-star, -genre, -director, -certificate) %>%
  filter(!duplicated(movie_name))

piratesSummary <- piratesMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

spiderMovies <- relevantMovies  %>%
  filter(grepl('Spider-Man', movie_name)) %>%
  select(-star, -genre, -director, -certificate) %>%
  filter(!duplicated(movie_name))

spiderSummary <- spiderMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
```

```r
  rename(count = revenue_Count)

##Revenue and Rating
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = rating,
    y = revenue
  )
)+
  geom_point(size = 1)+
  labs(
    x = "Average Rating (out of 10)",
    y = "Revenue ($)",
    title = "Revenue vs. Average Rating"
  )+
  theme_bw()+
  theme(
    text = element_text(family="Times New Roman")
  )

##Runtime and Rating
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = rating,
    y = runtime
  )
)+
  geom_point(size = 1)+
  labs(
    x = "Average Rating (out of 10)",
    y = "Runtime (mins)",
    title = "Runtime vs. Average Rating"
  )+
  theme_bw()+
  theme(
    text = element_text(family="Times New Roman")
  )

##Runtime and Revenue
ggplot(
  data = relevantMovies,
  mapping = aes(
    x = runtime,
    y = revenue
  )
)+
  geom_point(size = 1)+
  labs(
    x = "Runtime (mins)",
    y = "Revenue ($)",
    title = "Revenue vs. Runtime"
  )+
  theme_bw()+
  theme(
    text = element_text(family="Times New Roman")
  )

##Regression Table
data(relevantMovies)
revenue_rating_regression <- lm(revenue ~ rating, data = relevantMovies)
runtime_rating_regression <- lm(runtime ~ rating, data = relevantMovies)
revenue_runtime_regression <- lm(revenue ~ runtime, data = relevantMovies)

stargazer(revenue_rating_regression,runtime_rating_regression,revenue_runtime_regression,  ty

##4D
scatterplot <- ggplot(
  data = relevantMovies,
  mapping = aes(
    x = runtime,
    y = revenue,
    color = rating,
    #color = genre,
    #shape = averageRating
  )
)+
  geom_line(linewidth = 1)+
  labs(
    x = "Runtime (mins)",
    y = "Revenue ($)",
    color = "Average Rating",
    #color = "Genre",
    #shape = "Average Rating
    title = "Revenue vs. Runtime with Genre and Average Rating"
  )+
  scale_color_gradient(low = "#AFEEEE", high = "#668B8B")
facet_scatter <- scatterplot + facet_wrap(~genre, scales = "free")+
  theme_bw()+
  theme(
    legend.position = "top",
    text = element_text(family="Times New Roman")
  )
print(facet_scatter)

##Summary Stats Movies
moviesSummary %>%
  select(genre, count, revenue_Avg, runtime_Avg, rating_Avg) %>%
  kable(
    caption = "Averages of Movies' Revenue, Runtime, and Rating by Genre",
    col.names = c("Genre", "Count", "Average Revenue", "Average Runtime", "Average Rating")
  ) %>%
  kableExtra::kable_classic()

##Genre and Rating
ggplot(relevantMovies) +
  aes(x = rating, y = genre) +
  geom_boxplot(fill = "#A569FF") +
  labs(
    x = "Average Movie Rating",
    y = "Genres",
    title = "Average Rating by Genre",
    caption = "A box plot comparing the movies' average voter ratings and the movies' genres
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 15L),
    axis.title.x = element_text(size = 15L),
    axis.text.y = element_text(size = 13L),
    axis.text.x = element_text(size = 13L)
  )
```

```r
##Genre and Revenue
ggplot(relevantMovies) +
  aes(x = genre, y = revenue) +
  geom_bar(stat = "summary", fun = "sum", fill = "#4682B4") +
  labs(
    x = "Genre",
    y = "Revenue",
    title = "Genre vs Largest Revenue",
    caption = "A bar plot comparing movies' largest revenues within four genres."
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.text.y = element_text(size = 13L),
    axis.text.x = element_text(size = 13L)
  )

##Genre and Runtime
ggplot(relevantMovies) +
  aes(x = runtime, y = genre) +
  geom_boxplot(fill = "#65BA65") +
  labs(
    x = "Movie Runtimes",
    y = "Genres",
    title = "Runtimes vs Genres",
    caption = "A box plot comparing movies' runtimes and their genres."
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 20L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 15L),
    axis.title.x = element_text(size = 15L)
  )

##Harry Potter Figure
ggplot(harryPotterMovies) +
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
    title = "Rating, Revenue, and Runtime of Harry Potter Movies",
    caption = "A bar plot comparing movies in the Harry Potter franchise by their average vo
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 12L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 2L)
  )

##Harry Potter Table
harryPotterMovies %>%
  kable(
    caption = "Summary Statistics of Harry Potter Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()

##PoTC Figure
ggplot(piratesMovies) +
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
    title = "Rating, Revenue, and Runtime of Pirates Movies",
    caption = "A bar plot comparing movies in the Pirates of the Caribbean franchise by thei
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 13L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 3L)
  )

##PoTC Table
piratesMovies %>%
  kable(
    caption = "Summary Statistics of Pirates of the Caribbean Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()

##Spider-Man Figure
ggplot(spiderMovies) +
  aes(
    x = movie_name,
    y = rating,
    fill = revenue,
    colour = runtime
  ) +
  geom_bar(stat = "summary", fun = "sum") +
  scale_fill_gradient() +
  scale_color_distiller(palette = "OrRd") +
  labs(
    x = "Movie Title",
    y = "Average Movie Rating",
```

```r
    title = "Rating, Revenue, and Runtime of Spider-Man Movies",
    caption = "A bar plot comparing movies in the Spider-Man franchise by their average vote
    fill = "Revenue",
    color = "Runtime"
  ) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 13L,
    face = "bold",
    hjust = 0.5),
    axis.title.y = element_text(size = 10L),
    axis.title.x = element_text(size = 10L),
    axis.text.x = element_text(size = 3L)
  )

##Spider-Man Table
spiderMovies %>%
  kable(
    caption = "Summary Statistics of Spider-Man Movies",
    col.names = c("Movie Title", "Year", "Runtime", "Rating", "Votes", "Revenue")
  ) %>%
  kableExtra::kable_classic()
```