Final Report

IMDB Movies Data Analysis

Introduction

```
- State your research questions and explain them. (What is our goal/purpose in this projec
- Describe the provenance of your data. That is, where did you get the data, who collected
- Explain how your data meet the FAIR and/or CARE Principles.
```

Methodology

```
- Describe what attributes you'll focus your analysis on (mention if they are part of your
- How will the analysis be conducted and displayed?
- What did we do to assess our data, pare down the variables, compare them, etc?
  - Talk about old vs new data sets
```

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

```
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")
  )
```

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

```
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")
  )
```

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

```
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")
  )
```

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 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

```
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
```

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

Revenue vs. Runtime with Genre and Average Rating

```
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)
```

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 shows the five number summaries and 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: Summary Statistics

genre	count	revenue_Min	revenue_Q1	revenue_Median	revenue_Avg	revenue_Q3	revenue_M
Action	1525	7	370843.00	16029670	53193780	62678608	936662
Fantasy	472	1228	179182.00	11493182	50022561	54401813	804747
Horror	651	252	41623.00	2245000	20107771	30590174	327481
Mystery	610	87	85258.25	3143231	22451794	30551634	227966

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. 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.

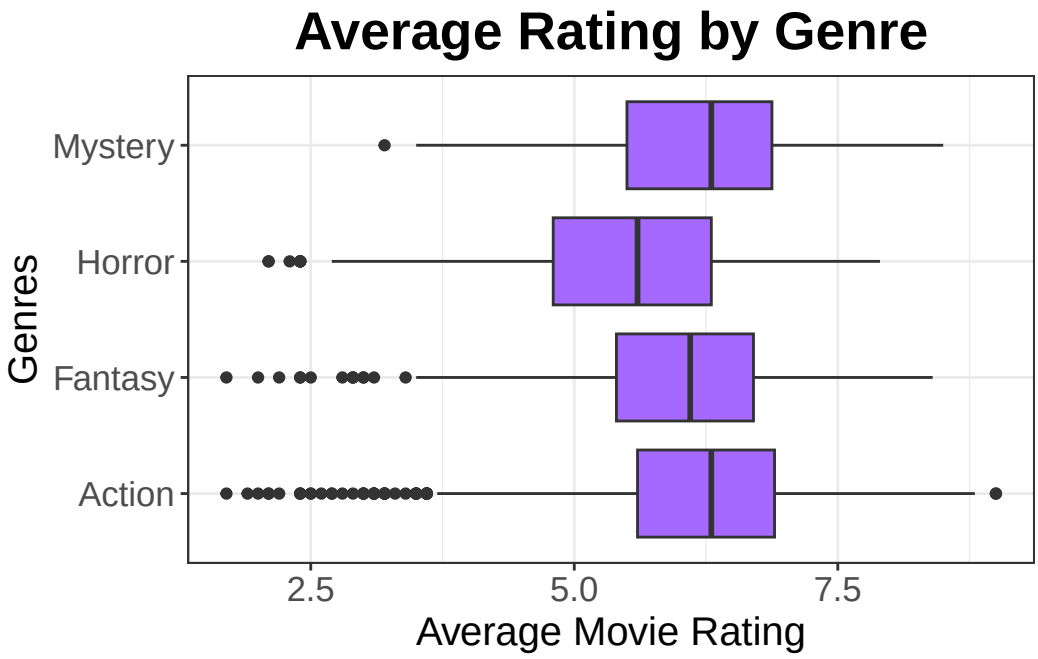

Figure 1

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. Many action movies are highly anticipated and have large fanbases (Marvel, Star Wars, etc.) so this trend makes sense to see.

Genre and Runtime

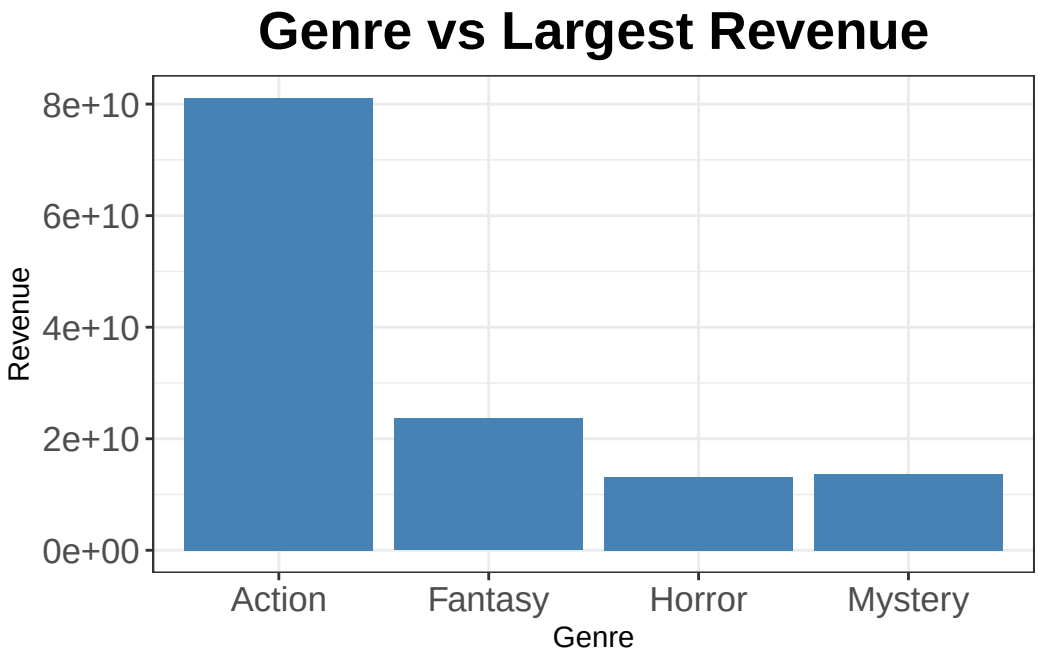

Figure 2

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.

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. It is important to note that these movies all have relatively similar runtimes and ratings, so these are likely not significant correlations.

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

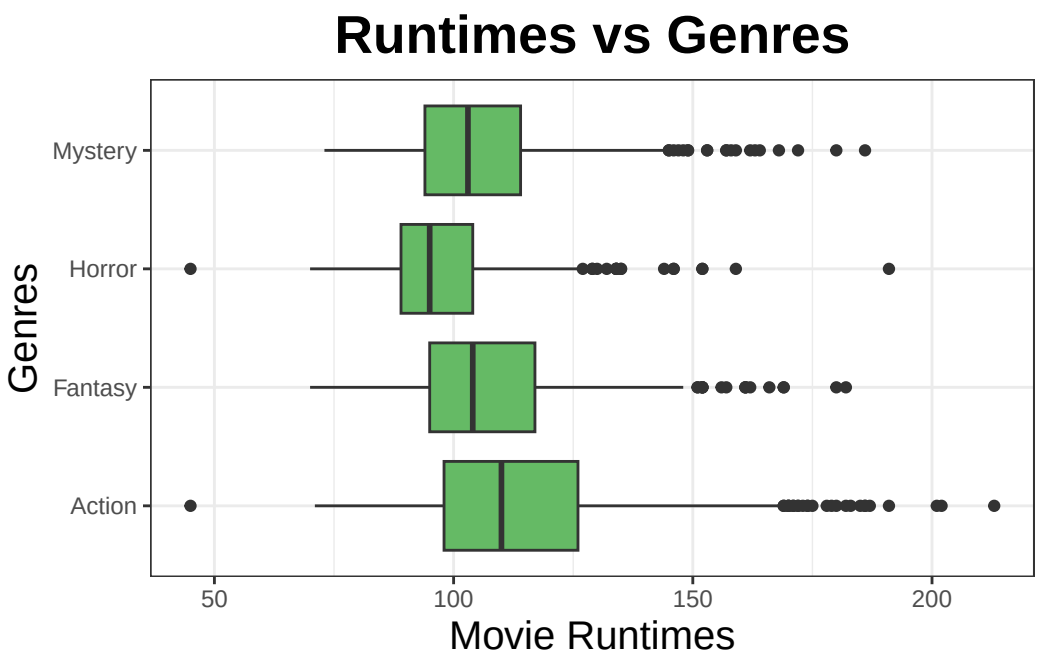

Figure 3

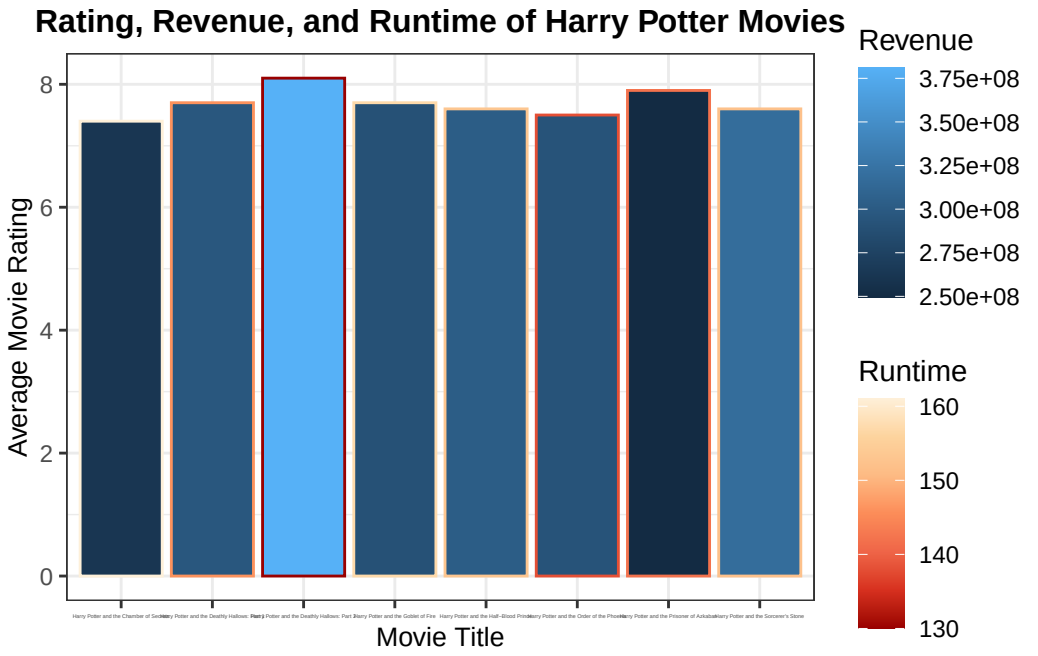

Figure 4

or runtime. 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.

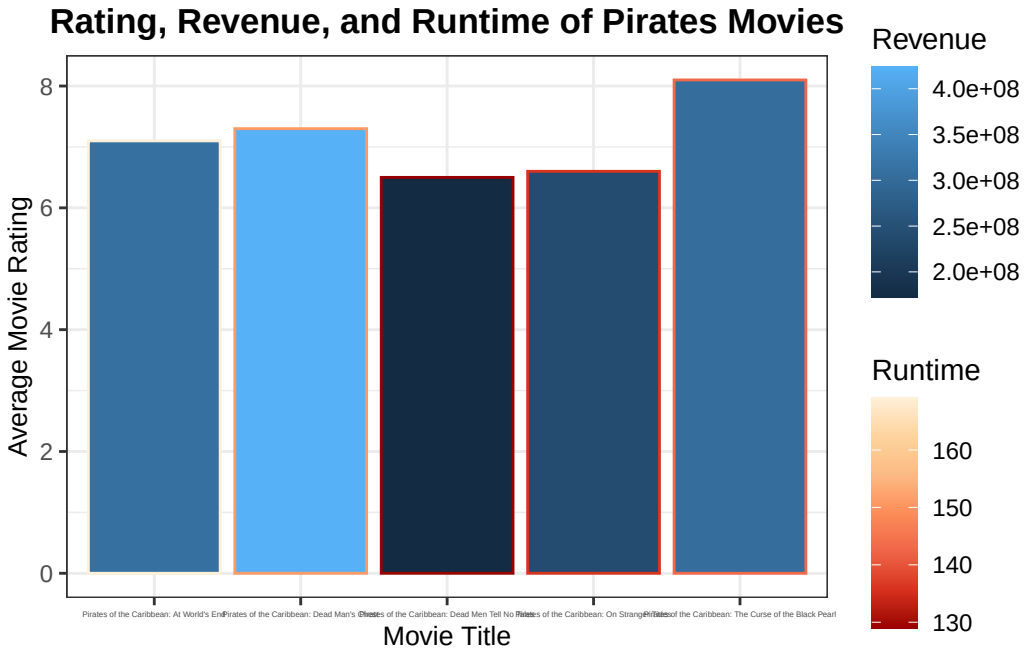

Figure 5

Spiderman 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. 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.

Results and Conclusion

```
- Summary of what we learned from the EDA
- Answer research questions
- Conclusion, future trends, relevance/importance
```

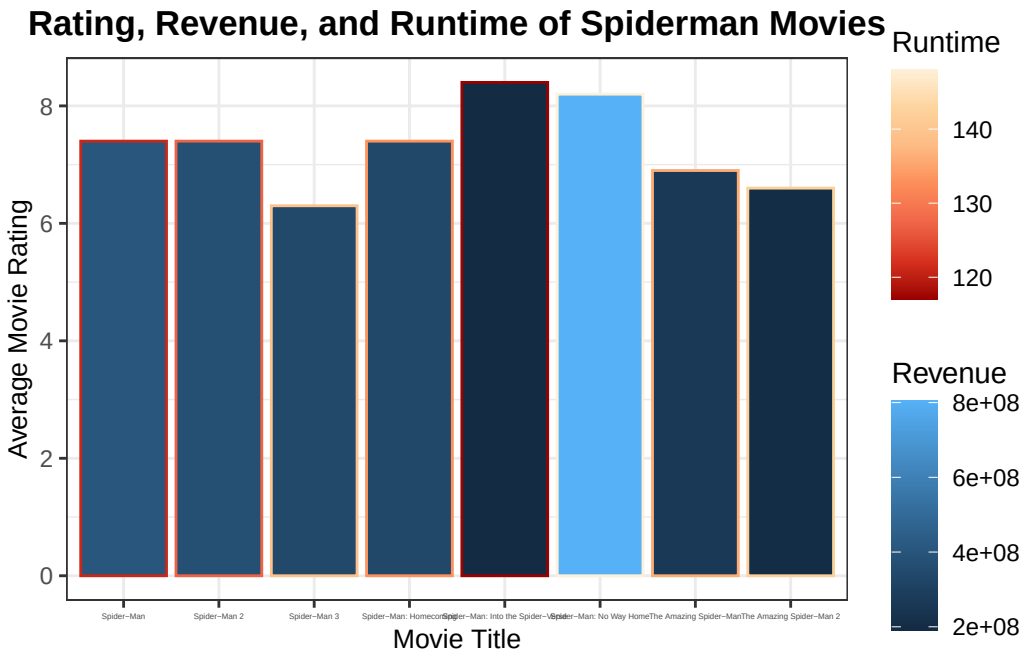

Figure 6

References

- Old data set citation
- New data set citation

Code Appendix

- Code chunk

::: {.cell}

```{.r .cell-code}
library(janitor)
library(knitr)
library(kableExtra)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(tidyr)
library(rvest)
```

```
library(plotly)
library(esquisse)
##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), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)

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

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) %>%
  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) %>%
  filter(!duplicated(movie_name))

spiderSummary <- spiderMovies %>%
  summarize(across(c(revenue,runtime), info)) %>%
  select(-runtime_Count) %>%
  drop_na() %>%
  rename(count = revenue_Count)
moviesSummary %>%
  kable(
    caption = "Summary Statistics of Movies by Genre"
  ) %>%
  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"
  ) +
  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"
  ) +
  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) +
  aes(x = runtime, y = genre) +
  geom_boxplot(fill = "#65BA65") +
  labs(
    x = "Movie Runtimes",
    y = "Genres",
    title = "Runtimes vs 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",
    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)
  )
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",
    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)
  )
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",
    title = "Rating, Revenue, and Runtime of Spiderman Movies",
    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)
  )
```
:::