Yvonne Martinez, Lyndah Mupfunyah, Brandilyn Hall, and Kwadwo Asante
SMU Data Science Bootcamp
Project 1 Writeup

Netflix Exploratory Analysis

Millions of people watch Netflix from around the world. It's not only a pastime but can also serve as a comfort or escape for many. While Netflix has been popular for quite some time, with the Pandemic, more people are watching than ever. We chose to analyze a Netflix dataset due to the impact it has on the world and our lives. Our aim throughout our analysis was to determine if countries differed in duration length and which country contributed the most movies.

We used exploratory data analysis and data visualizations to turn our data into a story to answer our posed questions and accept or reject our hypotheses. We also performed regression analysis to determine if there were any trends between our identified variables. In the following report, we will introduce you to our hypotheses and analysis questions and then walk through the analysis process by providing the steps we took to answer our analysis and regression questions. We will also provide our visualizations and commentary to help bring our data journey to life. We'll end with limitations and future work aspirations.

Data

Our dataset's name is Netflix Movies and TV Shows and we got it from Kaggle. This dataset contained over 8,800 titles of Movies and TV shows that are available in Netflix from different countries and provides information for each including, duration, rating, listed in, release year, description, cast, director, etc.

We also reviewed a few previous analyses to gather some inspiration for our project. These resources included:

- Kaggle: Audioviz on Netflix Dataset
- Medium: Data Analysis of Netflix Movies and & IMDB rating
- Datacamp: Netflix Movie Data

After reviewing these pages, we were able to make a plan as to how we wanted to conduct our analysis, pose our questions, and structure our visualizations.

Analysis Questions:

- Which countries contributed the most movies to Netflix?
- What genres are most common?
- Which countries have the longest movie durations?
- Is there a correlation between movie duration and release year (are movies getting shorter or longer)?

Hypotheses:

- Null Hypothesis: Genres in each country do not differ in duration.
- Alternative Hypothesis: Genres in each country do differ in duration.

AND

- Null Hypothesis: Countries do not differ in movie lengths.
- Alternative Hypothesis: Countries differ in movie lengths.

Methods

As mentioned above, we conducted exploratory data analysis to answer our posed questions. We conducted our analysis by within a Jupyter notebook and applying python coding to our data and used many methods to manipulate the data to tell our story.

Data Cleaning

Before we could conduct our analysis, we needed to clean up our dataset to make it more readable, easier to use, and ensure accurate results. We began to deduce the information from the dataset by identifying the vital columns such as title, duration, country, listed in, and release year, while dropping the non-applicable ones. We also made sure to turn the columns into the appropriate data type for a balanced transition into a feasible table. We removed TV shows from the dataset to gain a better vantage point and have a more directed focus. Movies also had a more definite duration length than TV shows, so the data was easier to work with.

We renamed our selected columns for easier readability. For example, the column 'listed in', we renamed to 'genre' so it makes more sense to the end user and us as the analysts. We then, reviewed our data for Null values and discovered that our Country column had several. Rather than assign the movie to a country in which it was not released, we collectively agreed to drop all null values.

There were quite a few genres of movies in the original dataset as well. We decided to take the top 7 genres based on the number of movies contained in each genre and dropped the rest. Our countries column needed the most help. We had some columns that had multiple countries in one row. For example, it may have said, "United States, United Kingdom, Egypt, Brazil" all in one row. So, we decided to use code to extract only the 1st country from the list on every row that had this format.

Once that was complete, we took the top 3 countries based on number of movies released, which we identified as the United States, the United Kingdom and India (see fig. 1). The duration column was also in need of some clean up. Initially, the columns contained the number of minutes followed by the word "min", which made the duration value a string. We wanted to make it an integer, so we ran code to remove the word "min" and then cast the numbers to be an integer. We also dropped any duration under 41 mins. The Oscars say anything over 40 minutes is not a short film, therefore would be classified as a movie.

Fig. 1 – Location of top 3 Countries

Finally, there were about 12 different maturity ratings in the initial dataset and many of them were synonymous with others, so we decided to consolidate the ratings that were synonymous together. For example, TV-14 is the same as PG 13, so we made all TV-14 movies PG-13 instead.

Analysis

Number of Movies & Country Contribution

After data cleaning and streaming down our data, we created a data frame that narrowed down our data to the release year and country. We defined the number of rows and columns that our data had and counted the number of movies released in a year. We performed a df.loc to specify only movies released from 2015 and up. United States turned out to be the country with the most movies followed by India and then the United Kingdom. Based on the bar graph (fig. 2.1), we see 2017 had the highest number movies released and 2021 the least. We also created a pie chart to depict the percentage of movies contributed by country. This shows that the United States had the highest contribution of 65.6% followed by India with 22.7% and lastly United Kingdom with 11.7%. (fig. 2.2)

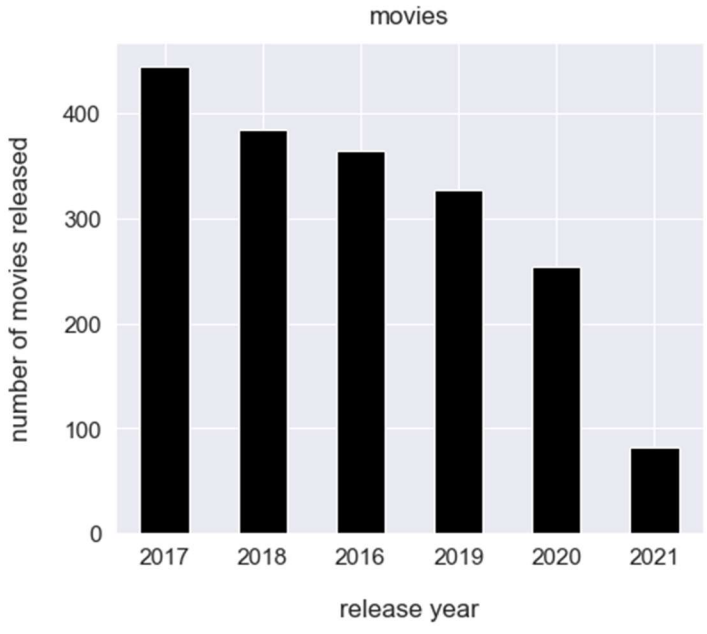

Fig. 2.1

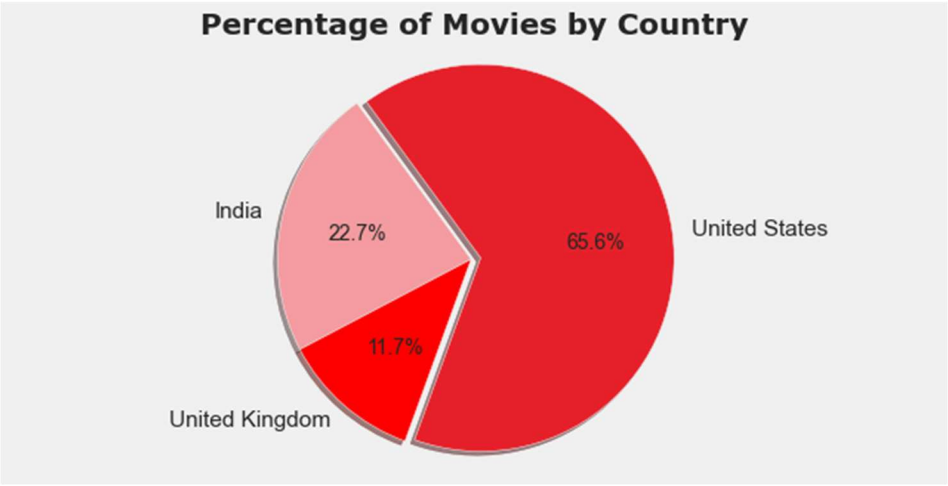

Fig. 2.2

Breaking it down by genre, we determined that within the top 3 countries, Dramas had the most movies in Netflix, followed by Comedies, then Documentaries. We visualized this analysis by a donut chart (fig. 2.3). This could indicate a higher preference of Drama movies but would have to conduct further analysis. We then focused on each country on its own to see if there was a different story. The United States was consistent with our findings of the top 3, with there being more Drama movies, followed by Comedies, and then Action & Adventure (fig. 2.4). The United Kingdom also had Dramas leading the way, with Action & Adventure next and then Comedies (fig. 2.5). India was similar with Dramas, then Comedies, followed by Action & Adventure (fig. 2.6).

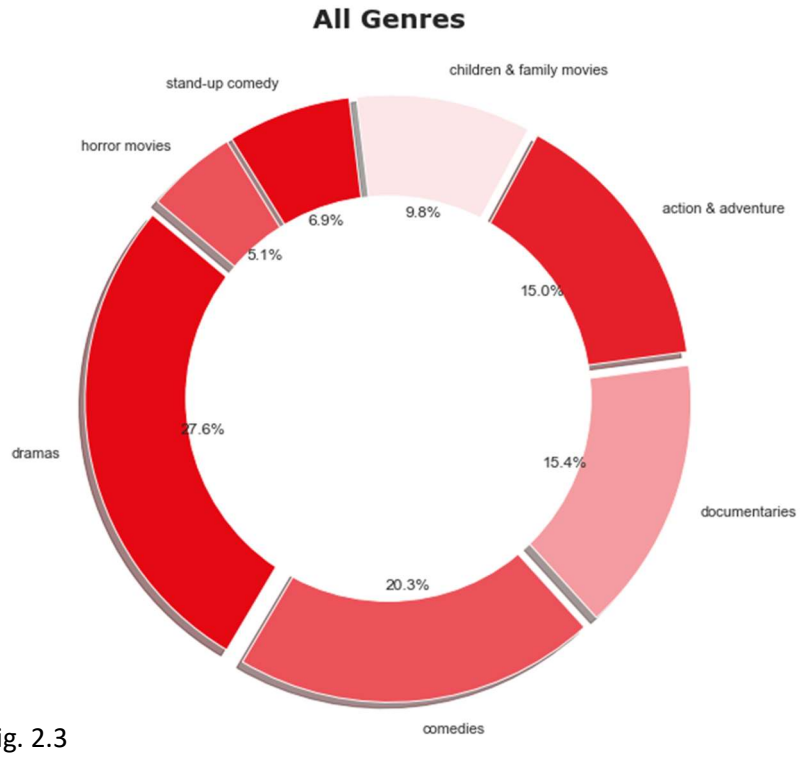

Fig. 2.3

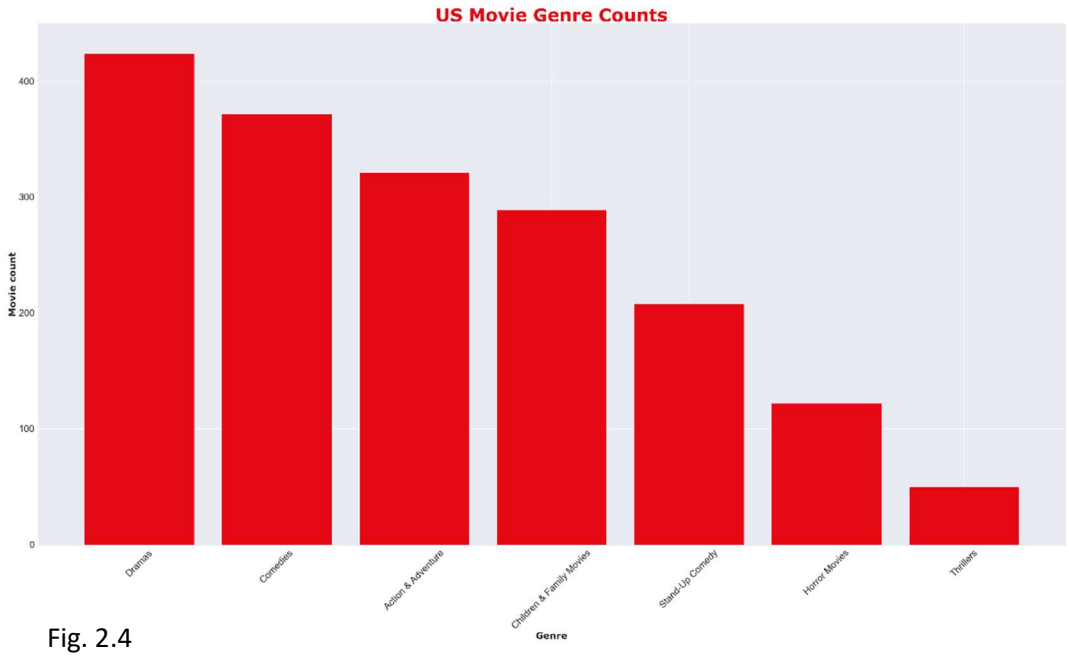

Fig. 2.4

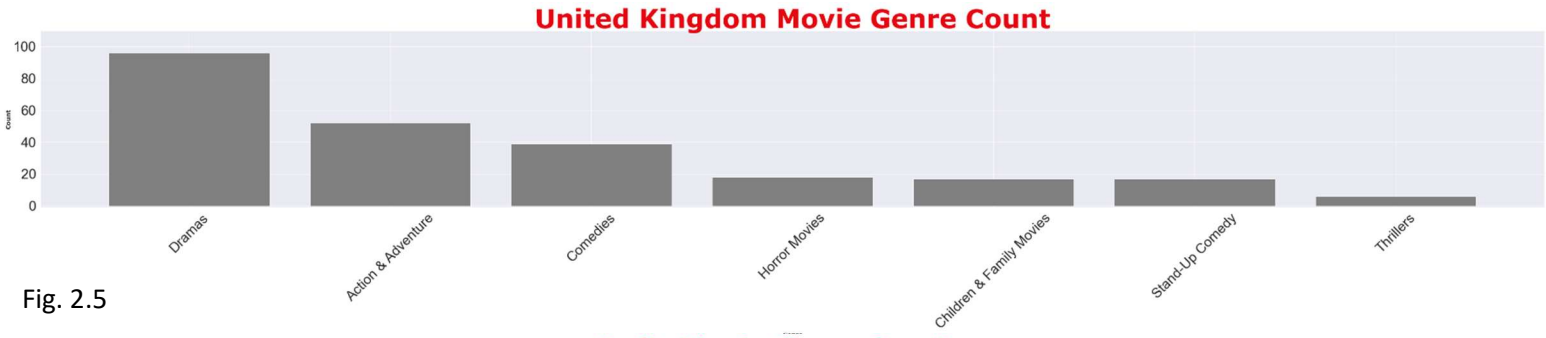

Fig. 2.5

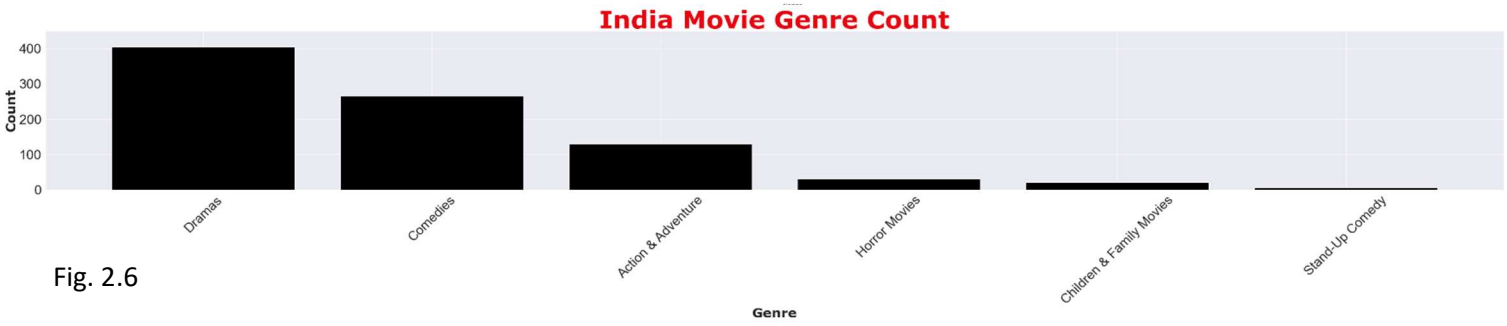

Fig. 2.6

Duration by Country

We then broke it down by country and duration. We grouped by country then found the max duration for each country. The United States lead the way, with a movie at 312 minutes (Black Mirror Bandersnatch) which is a definite outlier but chose to keep it in our dataset because we found it interesting. India was next with a 224-minute movie (Lagaan), the United Kingdom came in last of the 3

with a movie at 208 minutes (No Direction Home: Bob Dylan). (fig. 3.1) The average duration for each country tells a different story, however. The averages as shown in our Summary Statistics chart, indicate that India actually has the highest average movie length at 126 minutes, then the United Kingdom at 96 minutes, and the United States was last at 94 minutes. (fig. 3.2) Looking at the top 5 movies by duration for all countries, 3 of the 5 movies come from India which is consistent with our average movie lengths for each country (fig. 3.3).

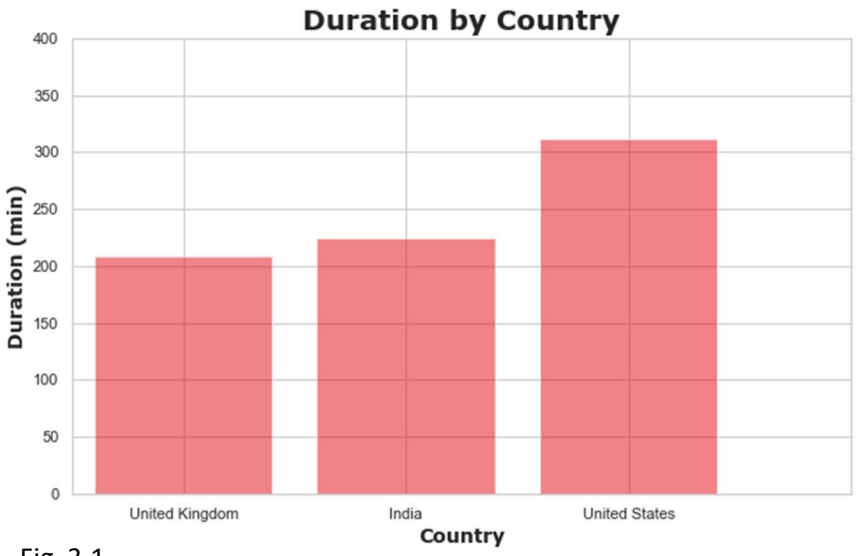

Fig. 3.1

country	mean	max	min	median	std	var
India	126.493151	224.0	41.0	127.0	25.205599	635.322239
United Kingdom	96.611765	208.0	41.0	97.0	23.877190	570.120215
United States	94.167448	312.0	41.0	94.0	21.472312	461.060169

Fig. 3.2

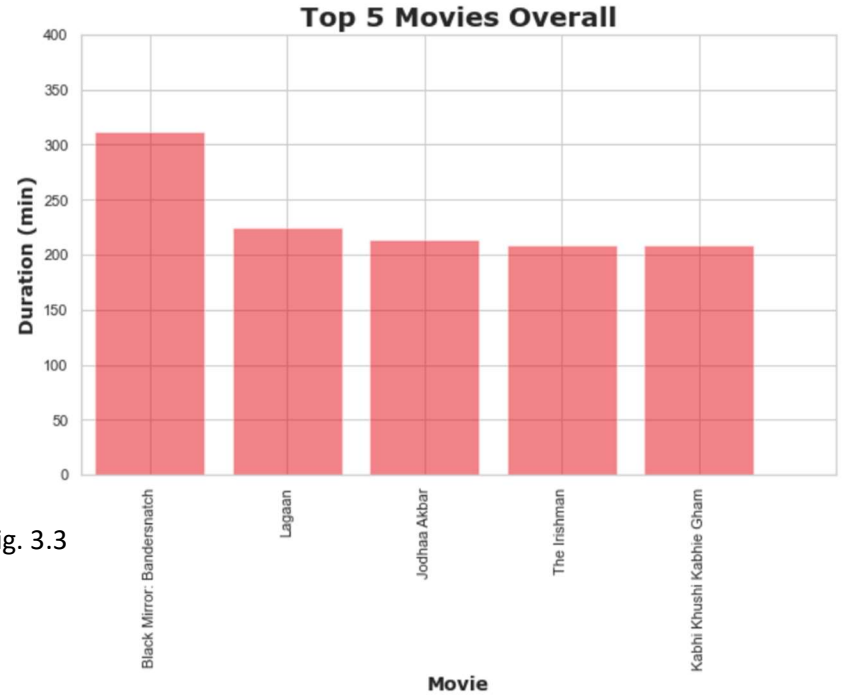

Fig. 3.3

Separating by countries, we visualized our top 5 movies for each country which further confirming our summary statistics table. Black Mirror was clearly an outlier in the US at 312 minutes, with the other 4 movies sitting closer to 200 minutes (fig. 3.4). In the United Kingdom, No Direction Home: Bob Dylan was a slight outlier at 208 minutes with the others running at around 160 minutes (fig. 3.5). In India, no clear outlier was seen, with the longest movie, Lagaan (224 mins), just a bit longer than the 2nd longest movie, Jodhaa Akbar (~215 mins)(fig.3.6).

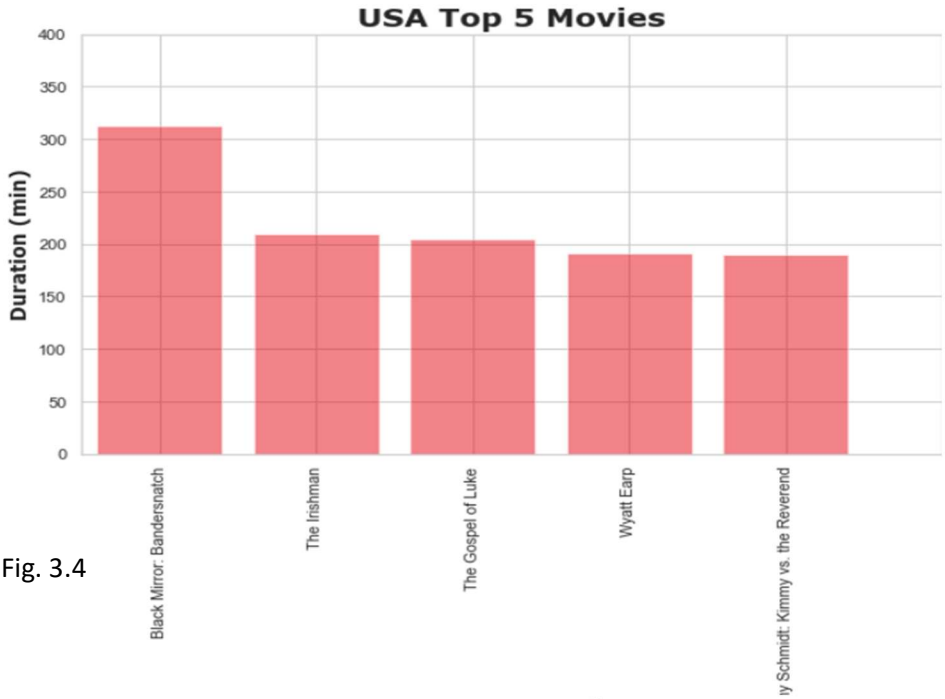

Fig. 3.4

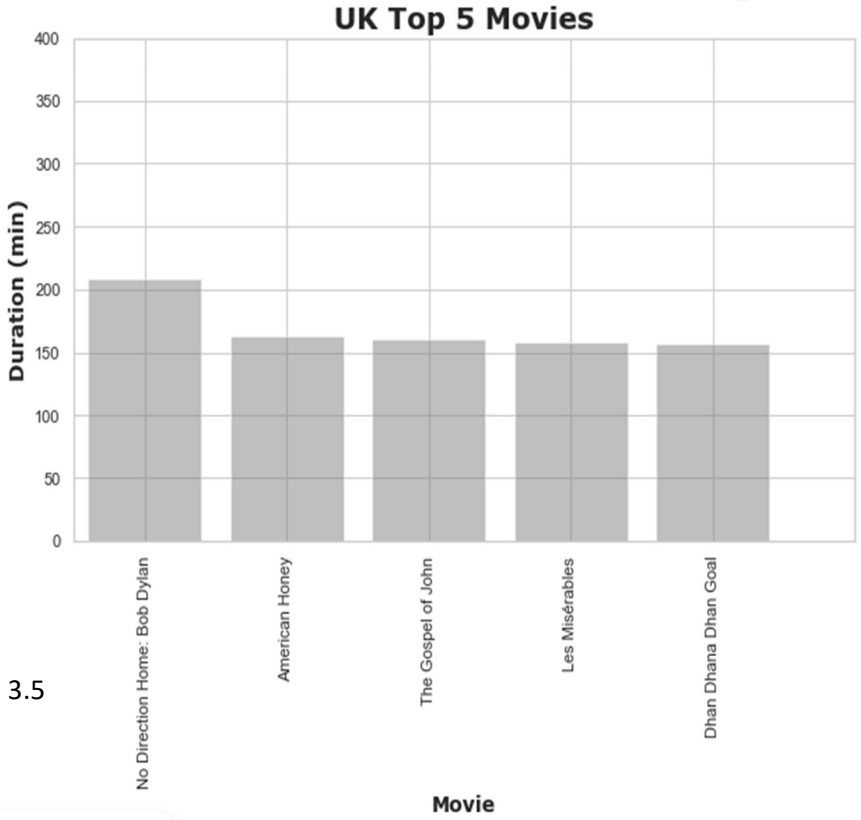

Fig. 3.5

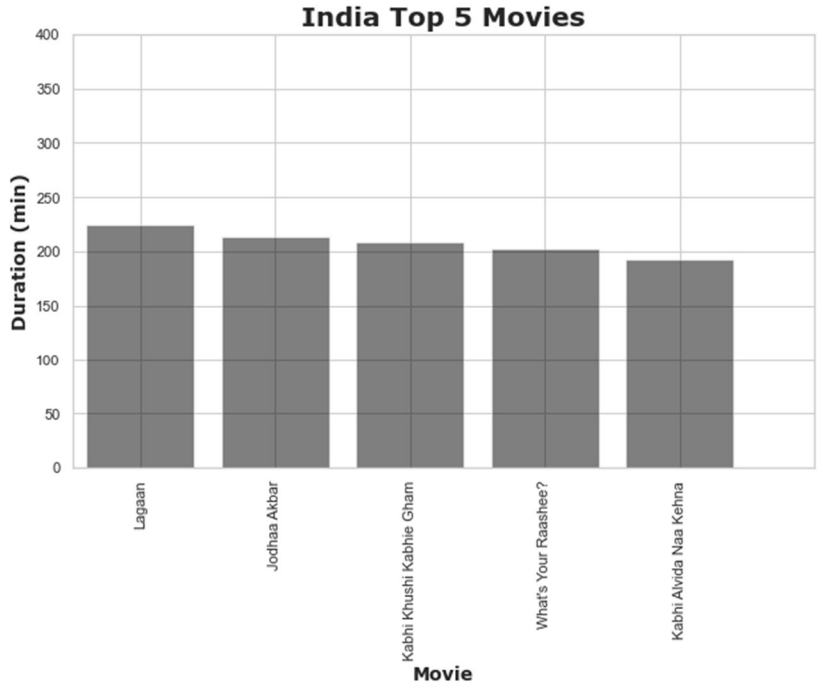

Fig. 3.6

Exploring the data even further, we created a violin plot displaying the duration of movies by each country. This graph shows the United States to have a very long right tail, which indicates a high outlier, which was identified as the 312 Black Mirror Bandersnatch movie. The United Kingdom also had a longer right tail as well, which accounts for the Bob Dylan documentary at 208 minutes. Out of the 3 countries analyzed, India had the most normal distribution. The United Kingdom looked fairly normal but could also be described as slightly bimodal (fig. 4.1).

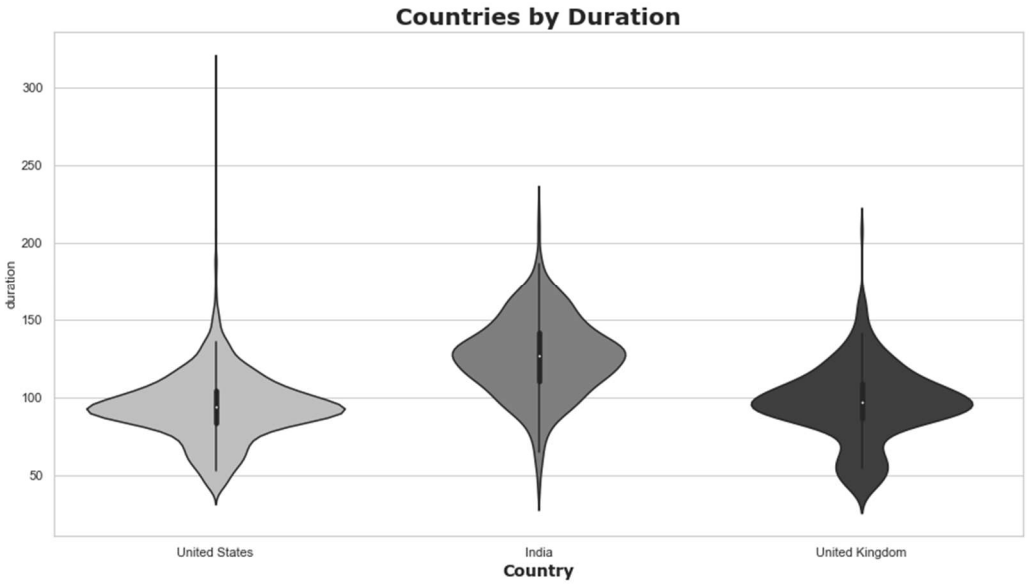

Fig. 4.1

Number of Movies by Duration

The Histogram and KDE graphs we created for each country also shows the shape of the data and show how many movies of each duration there are. Again, we see that both the United States and United Kingdom have a longer right tail, signifying high outliers, and India had a normal curve which is easily seen by this graph. We can also see that in the US and United Kingdom, most movies are around 90-99 minutes, and most movies in India are around 125-130 minutes which coincides with our statistics summary table showing an average move in India being 126 minutes (fig. 4.2, 4.3, 4.4).

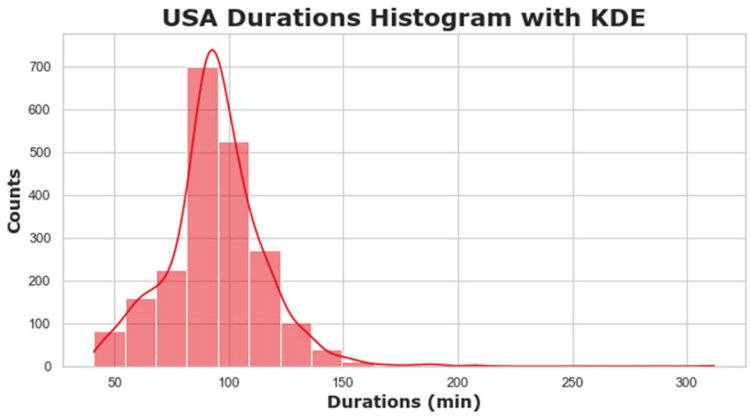

Fig. 4.2

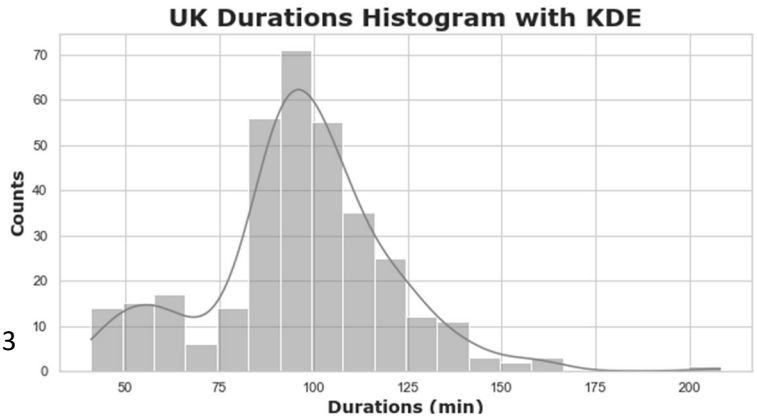

Fig. 4.3

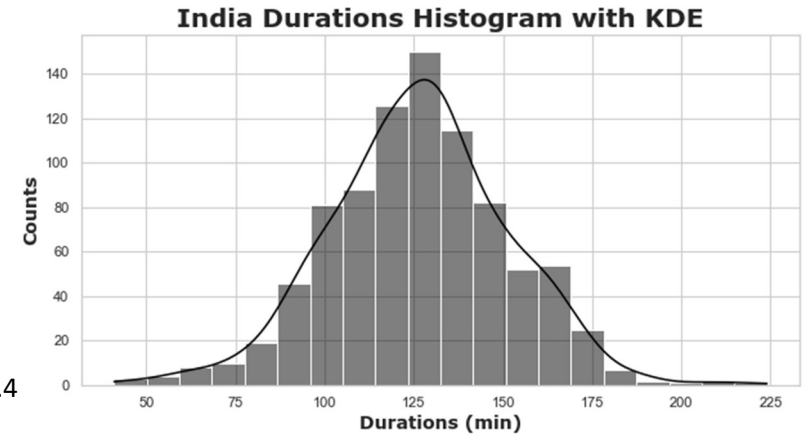

Fig. 4.4

Durations by Genre

Analyzing our data even further, we grouped by genre for each country and then found the average length for each. This was calculated by creating a side-by-side bar chart. This side-by-side Seaborn bar chart shows the average duration by genre for each country and also the SEM for each (4.5). These findings are consistent with the prior analysis we have conducted and also gives us more information. It looks like Action an adventure on average is longer just slightly than dramas in the US. In India, Action and Adventure is quite a bit longer than the 2nd place, Comedies, and in the UNITED KINGDOM, Dramas are on top slightly over Action & Adventure. To wrap up our initial analysis, the US contributed the most movies in Netflix, and has the longest duration between the top 3 countries. On Average, however, India takes the lead with an average movie duration of 126 and the US in 3rd at 94 mins. There are more Drama movies contributed by all 3 countries, with this genre having the highest percent of movies on Netflix. Finally, the lengths of genres in the 3 countries vary, but are overall pretty similar, although India does have a few genres that a quite a bit longer than the other 2 countries.

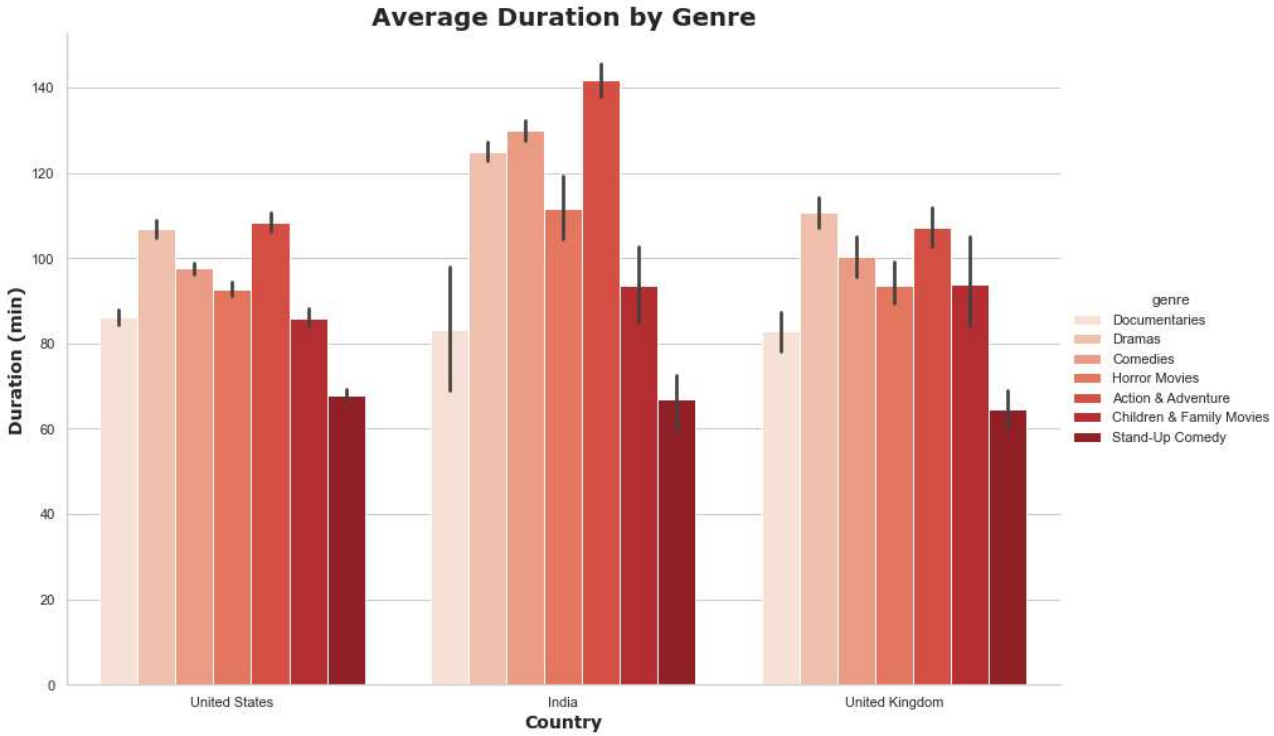

Fig. 4.5

Regression and Correlation

By deciding to narrow our analysis to the top three countries, we also wanted to cancel out some noise and pick three genres from the seven to analyze possible correlation between duration of movies and release year. We compared horror, action & adventure, and drama to United States, India, and United Kingdom.

Yvonne Martinez, Lyndah Mupfunyah, Brandilyn Hall, and Kwadwo Asante
SMU Data Science Bootcamp
Project 1 Writeup

First beginning with Action & Adventure. Our r-squared values were significantly low. (United Kingdom: 0.188, United States: 0.054, India: 0.015). A low r-squared value gives us reasoning to believe there is not a strong correlation when comparing this genre across the three countries. As a team we decided to keep our outliers to point out during our presentation. We noticed an interesting outlier in India. Jodhaa Akbar, which was released in 2008, is the longest action & adventure movie in our Netflix dataset at 214 minutes. Due to our low r-squared value we could assume that our prediction models would be weak. All countries were far from being along the diagonal line, proving our assumption (fig. 5.1 – 5.9)

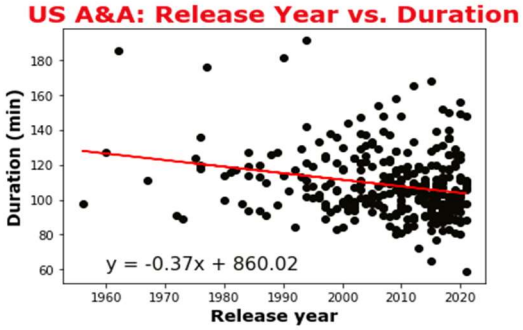

Fig. 5.1

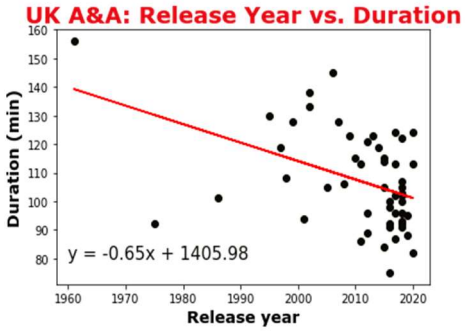

Fig. 5.2

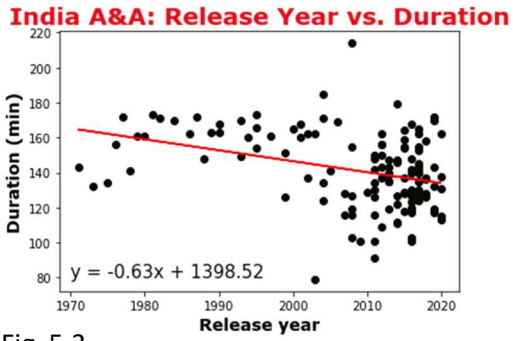

Fig. 5.3

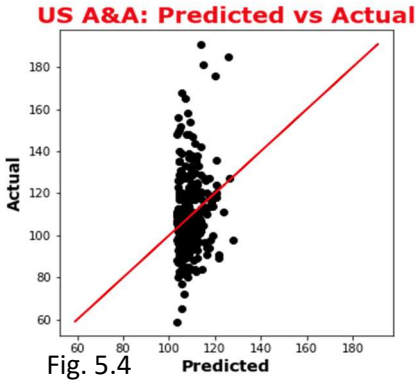

Fig. 5.4

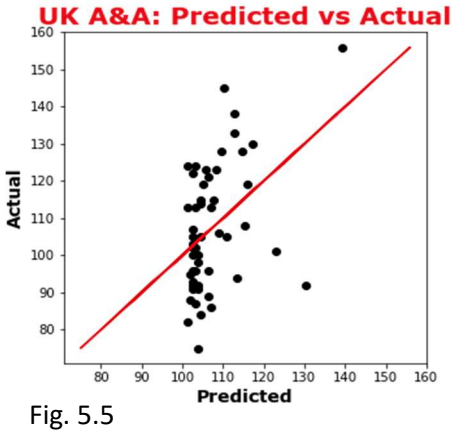

Fig. 5.5

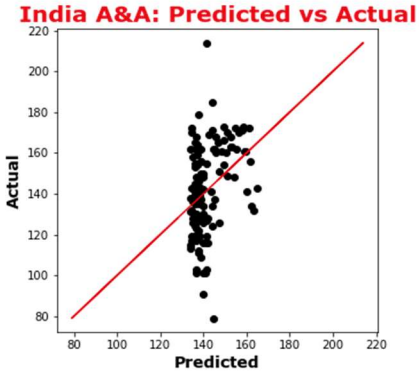

Fig. 5.6

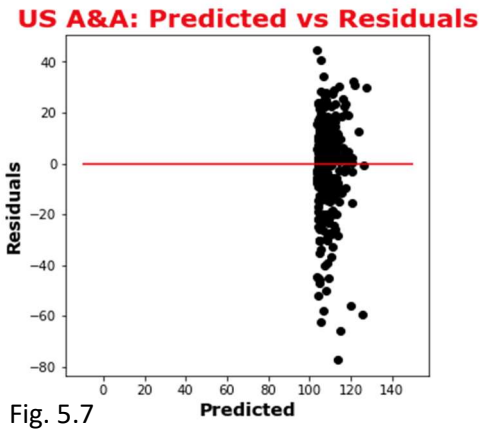

Fig. 5.7

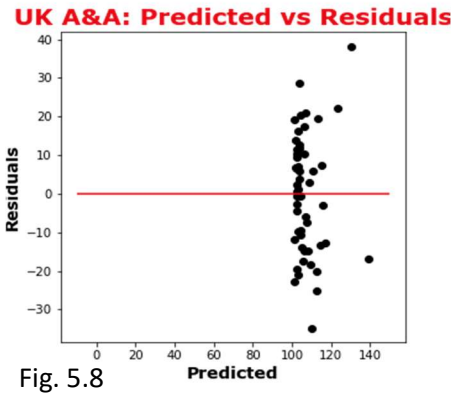

Fig. 5.8

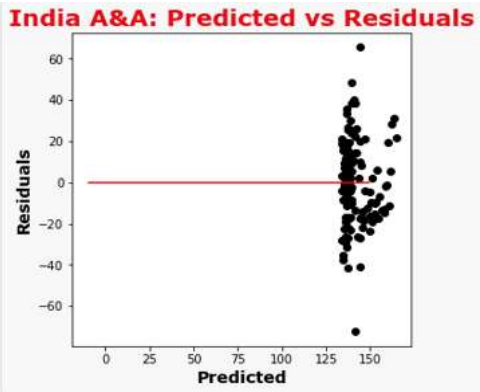

Fig. 5.9

Second comparison we noticed a similar situation. Drama had very low r-squared values for all countries. (United Kingdom: 0.069, United States: 0.035, India: 0.015). United Sates visually showed not much difference in duration of drama movies over the course of years. Black Mirror released in 2018 was a great outlier at 312 minutes. United Kingdom by far showed he weakest correlation due to a very scattered plot. Once again, with the low r-squared value we could assume that our prediction models would be weak, which was proven in predicted vs actual and predicted vs. residual (fig. 6.1-6.9).

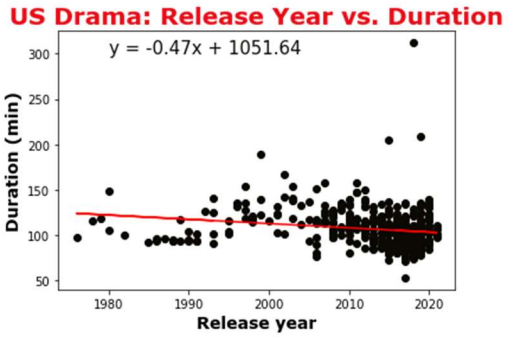

Fig. 6.1

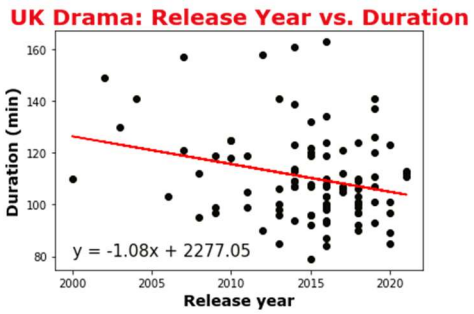

Fig. 6.2

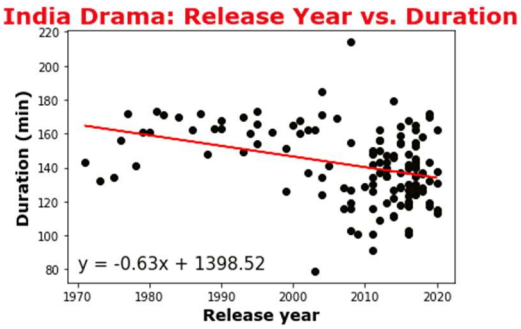

Fig. 6.3

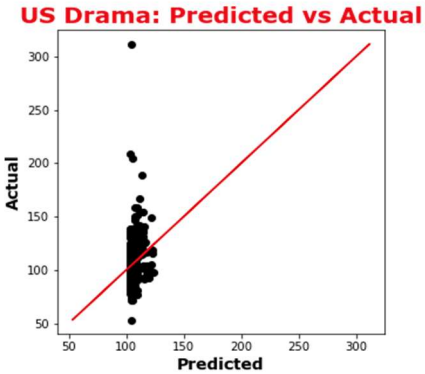

Fig. 6.4

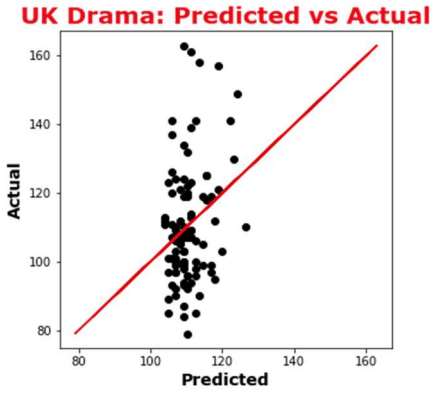

Fig. 6.5

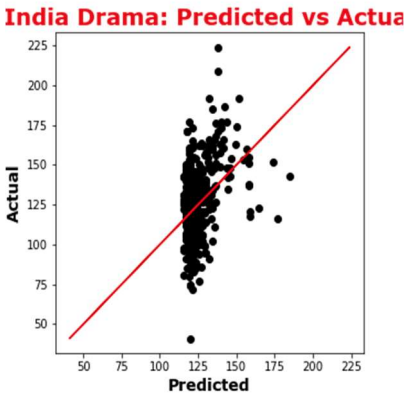

Fig. 6.6

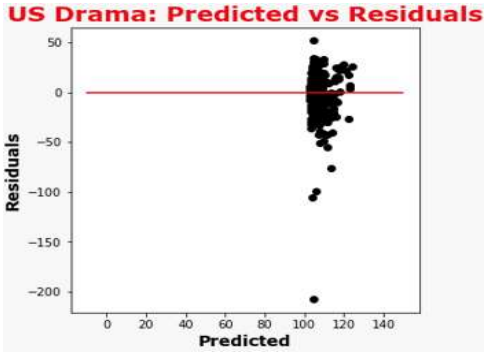

Fig. 6.7

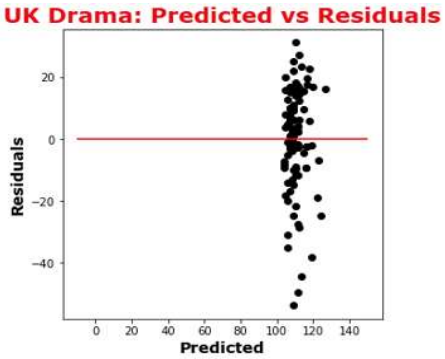

Fig. 6.8

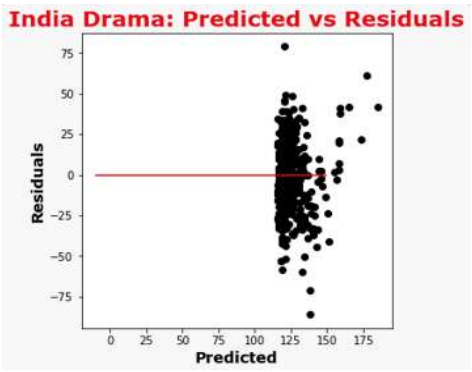

Fig. 6.9

Finishing our last comparison with Horror movies, United States was clearly a country that preferred horror movies. Both India and United Kingdom lacked releases prior to the 2000s. All three countries once again had a low r-squared value and weak predicted vs actual and predicted vs. residual models. (United Kingdom: 0.003, United States: 0.176, India: 0.015). Visually we could see horror movies in the United States were becoming shorter over time. Although the two other countries lacked releases it was interesting to point out outliers like Apostle at 130 minutes in 2018 and Andhakaaram at 171 minutes in 2020. Apostle was just as long as US outlier What Lies Beneath released in 2000. (fig.7.1-7.9)

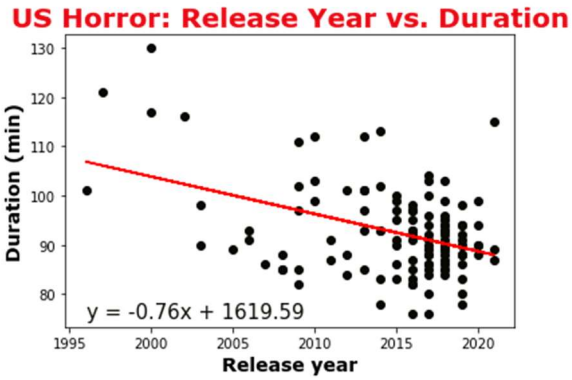

Fig. 7.1

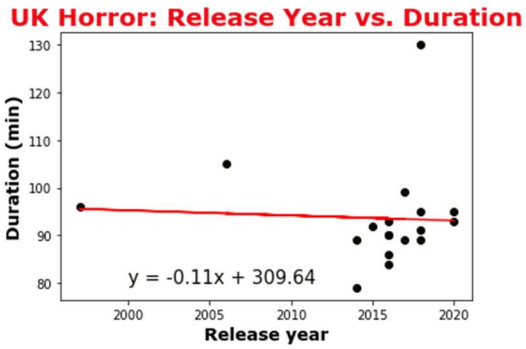

Fig. 7.2

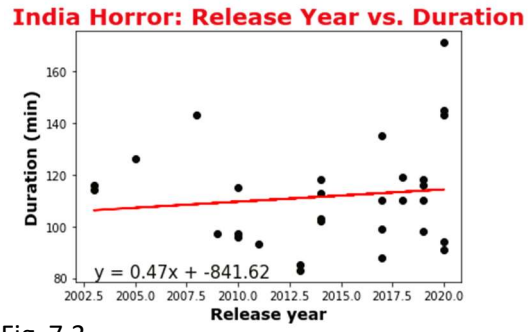

Fig. 7.3

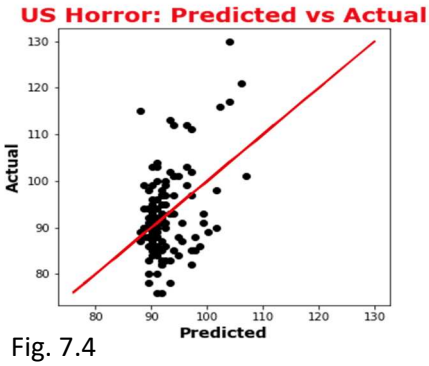

Fig. 7.4

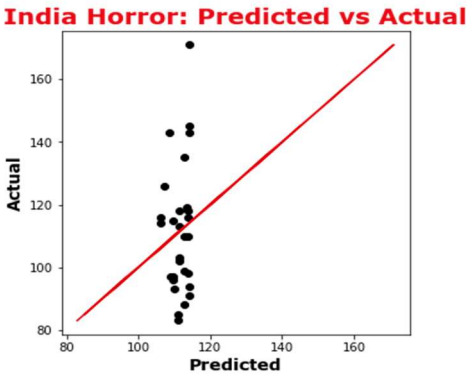

Fig. 7.5

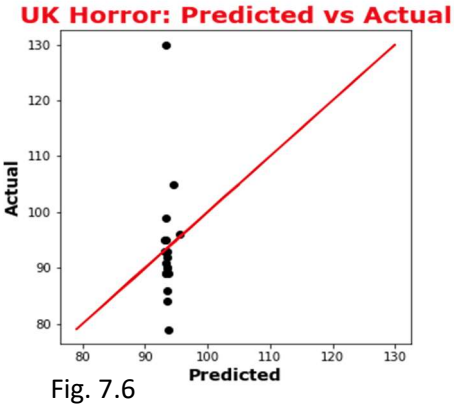

Fig. 7.6

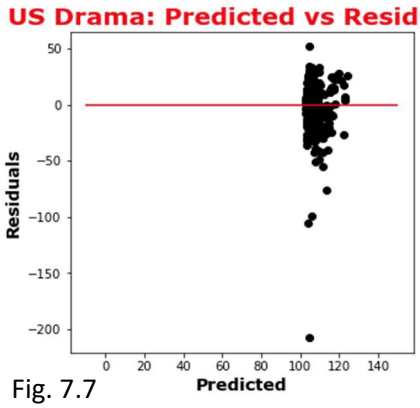

Fig. 7.7

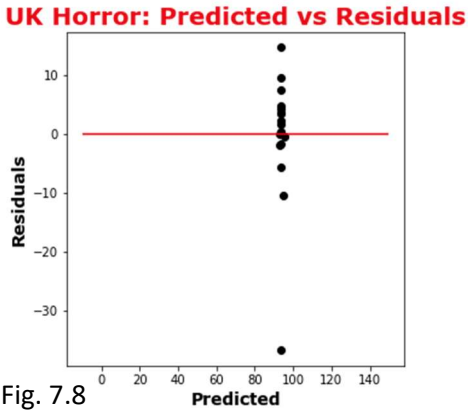

Fig. 7.8

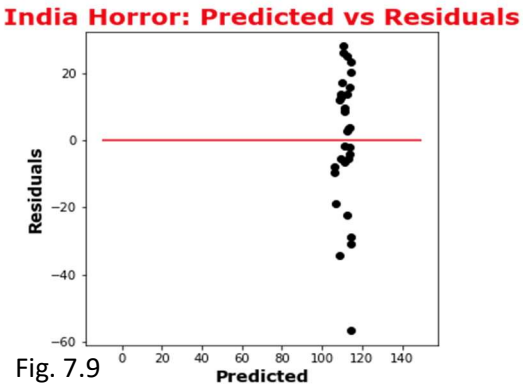

Fig. 7.9

In conclusion none of our genres showed strong correlation when comparing whether movies durations have changed over time.

Anova

Our last analysis consisted of completing an Anova. We decided this was the best fit as we were comparing 3 different groups. Each p-value output was over .05 meaning no significant difference exists (Action & adventure:0.684, Drama: 0.108, Horror 0.695). The box plots also reiterated the outliers identified before. (fig. 8.1, 8.2, 8.3)

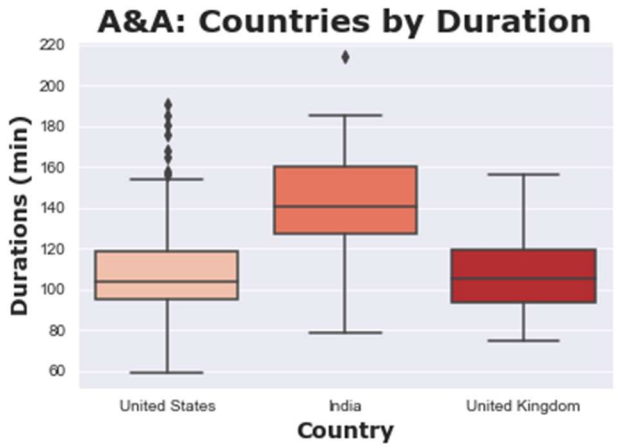

Fig. 8.1

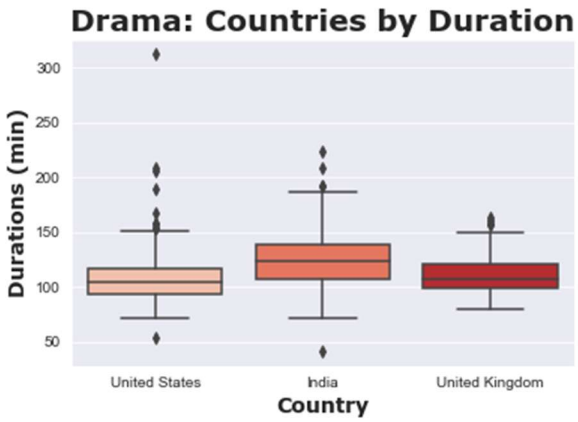

Fig. 8.2

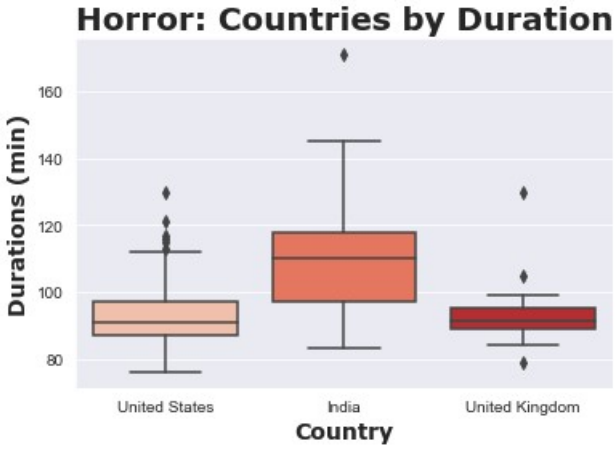

Fig. 8.3

Conclusions

From our analysis we came to the following conclusions:

- The United States contributed more movies to Netflix than the other 2 countries studied.
- Out of all movies contributed by the top 3 countries to Netflix, the Drama genre had the most. This was evident in all countries.
- On average, India produces the longest movies out of the top 3 countries analyzed.
- The United States produced the longest movie overall, however, with a 312- minute feature (which is an outlier).

- In each country, Action & Adventure had the longest durations in 2 of our countries (US and India) and the United Kingdom had Dramas as slightly longer.
- There is not a correlation between movie length and release year for any genre, which indicates that there is no evidence proving movies have gotten longer or shorter over the years with the exception of Horror movies in the United States being visually shorter in recent years.

Limitations

As with any dataset and analysis there were some limitations. This dataset provided does not account for all movies in production, but only focuses on movies that are available on Netflix. This dataset, therefore, may not represent all the movies each country releases. The dataset also doesn't contain all movies added to Netflix to present date, so there are some movies missing that are currently streaming. Additionally, our dataset is not representative of every country and therefore, we aren't able to analyze the movies they release as they are not a part of Netflix. Another deficit of our dataset was that it didn't have a popularity rating on the movies. This would have given us an even deeper look into duration and popularity, and well and genre and popularity. Given more time and practice, more analysis could be done and more questions answered about our data.

Future Work

In the future, we would like to conduct a deeper analysis on genre, country and duration to gain even more insight on the trends that different countries have regarding these variables. Also, we would like to obtain an IMDB dataset (or use the OMDB API) and merge it with this dataset to determine which movies are most popular and which countries have the most popular movies. Finally, we would like to bring back TV shows into our dataset and conduct similar analysis on them.

Sources

Netflix Movies and TV Shows Dataset - https://www.kaggle.com/shivamb/netflix-shows

Kaggle: Autoviz on Netflix Dataset (inspiration)- https://www.kaggle.com/rsesha/autoviz-on-netflix-dataset

Medium: Data Analysis of Netflix Movies & IMDB rating using Python (inspiration) - https://medium.com/geekculture/data-analysis-of-netflix-movies-imdb-rating-using-jupyter-notebook-d923186da6c7

Datacamp: Netflix Movie Data (inspiration) - https://www.datacamp.com/workspace/templates/dataset-python-netflix-movie-data