Data analysis: Insights from a Film Dataset

While searching for interesting datasets on Kaggle, I encountered the ‘Letterbox Movie Classification Dataset’, a classification dataset with a detailed collection of metadata for over 10,000 movies.   
<https://www.kaggle.com/datasets/sahilislam007/letterbox-movie-classification-dataset/data>

*Note: This is an American collection, which uses the American word ‘Movie’. When analysing and referencing this data, I shall use the British word ‘film’.*

I wanted to try and answer a few questions based on the data I had available:

* **Are some genres associated with higher average ratings?**
* **Do longer runtimes lead to significantly more likes?**
* **Are films made in English watched more than films in other languages?**

# Step 1: Understanding and Exploring the Dataset

The first step in my analysis was to become familiar with the dataset. The dataset contained the following key columns:

* Film title
* Director
* Average rating: On a scale of 1 to 5.
* Genre
* Runtime: in minutes.
* Original language
* A description of the movie
* Studio
* Watches: Total number of times the film has been watched by users
* List appearances: The number of times the movie appears in user-curated lists.
* Likes
* Fans
* Lowest Rating: The number of 1-star ratings the film has received.
* Medium Rating: The number of 3-star ratings the film has received.
* Highest Rating: The number of 5-star ratings the film has received.
* Total Ratings

Issues revealed in the data

A quick overview of the data revealed some errors in the data.

1. 579 films had an average rating of ['3.2685351873474158], which indicates an error. These films have 201 ratings or less. They also have vastly less engagement than other films in the dataset.
2. There were 9 films with an exact runtime of ['103.16833466773419], all other runtimes were given as whole numbers. This indicates an error.
3. One film had a runtime of 907 minutes. Upon checking, this was revealed as an error.
4. Twenty-four films had a genre listed as ‘Unknown’. More than half of those had a runtime under 28 minutes.
5. Three films had an original language of ‘unknown’.
6. Upon closer inspection it was revealed that within this database of films, there are in fact TV shows.

Key assumptions made in this analysis:

There are some important assumptions that must be made when analysing this dataset.

I cannot verify the quality of this dataset. It was provided by a user on Kaggle, and while I have checked that key elements of the data are correct, I do not know it’s source. I therefore must assume that the data is high-quality and reliable.

# Step 2: Data Cleaning and Preprocessing

Before diving into analysis, I cleaned the dataset to ensure accuracy. Key tasks included:

* Handing missing values: Unknowns were replaced with the correct values.
* Standardising genre classifications: Some movies were listed with multiple genres, so I streamlined the data using the primary genre classification and splitting genres into their own fields (genre\_1, genre\_2, etc).
* Removing outliers: To ensure reliable insights I removed films with low engagement that had incorrect ratings. I also removed TV shows by removing any records over 242 minutes (4 hours) long.

This step was crucial as it laid the foundation for reliable and insightful findings.

# Step 3: Exploratory Data Analysis (EDA)

With the dataset cleaned, I proceeded with exploratory analysis to uncover trends and relationships:

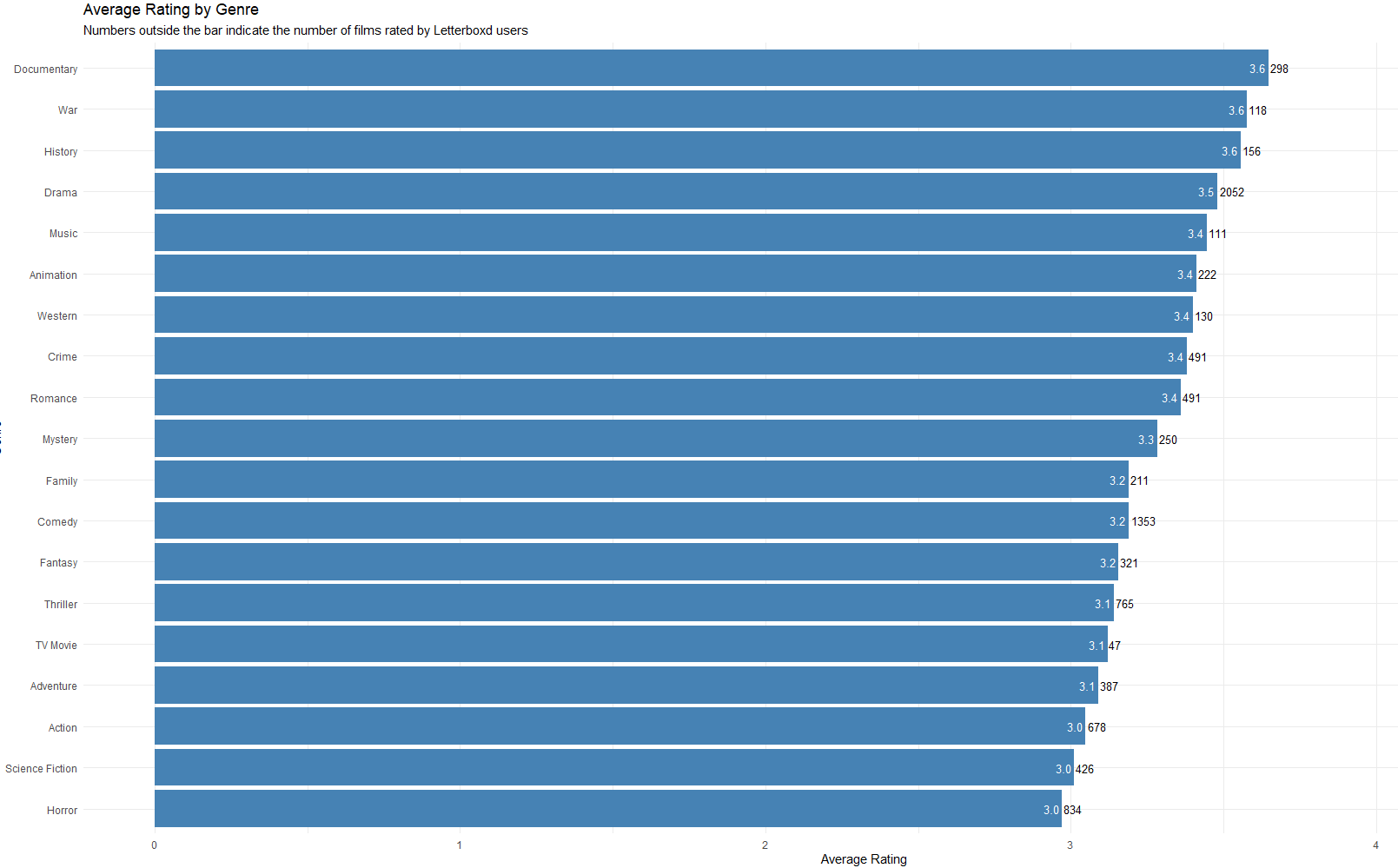
* Comparing genres and average likes: I grouped films by genre and found their average rating and plotted them as a horizonal bar chart.
* Comparing film runtimes with their likes: I created a scatter plot with runtimes and likes to see if there were relationships between the two.
* Comparing watches of a film with their language: I found the top 15 languages of films in the dataset, averaged their likes and created a horizonal bar chart to display the likes of films by language.

# Step 4: Visualising the Results

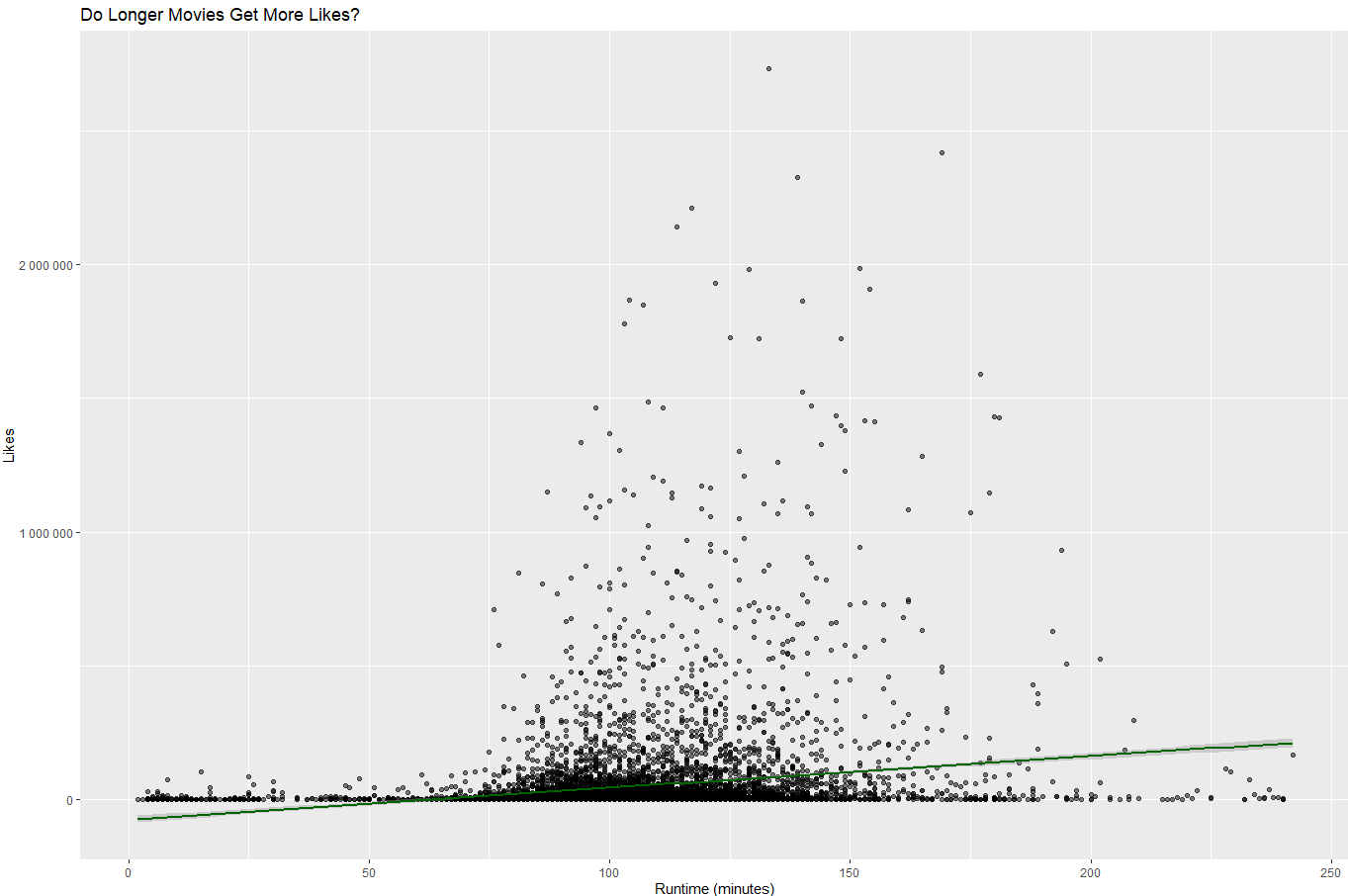
Great analysis requires effective communication. I created visualisations to summarise key findings.

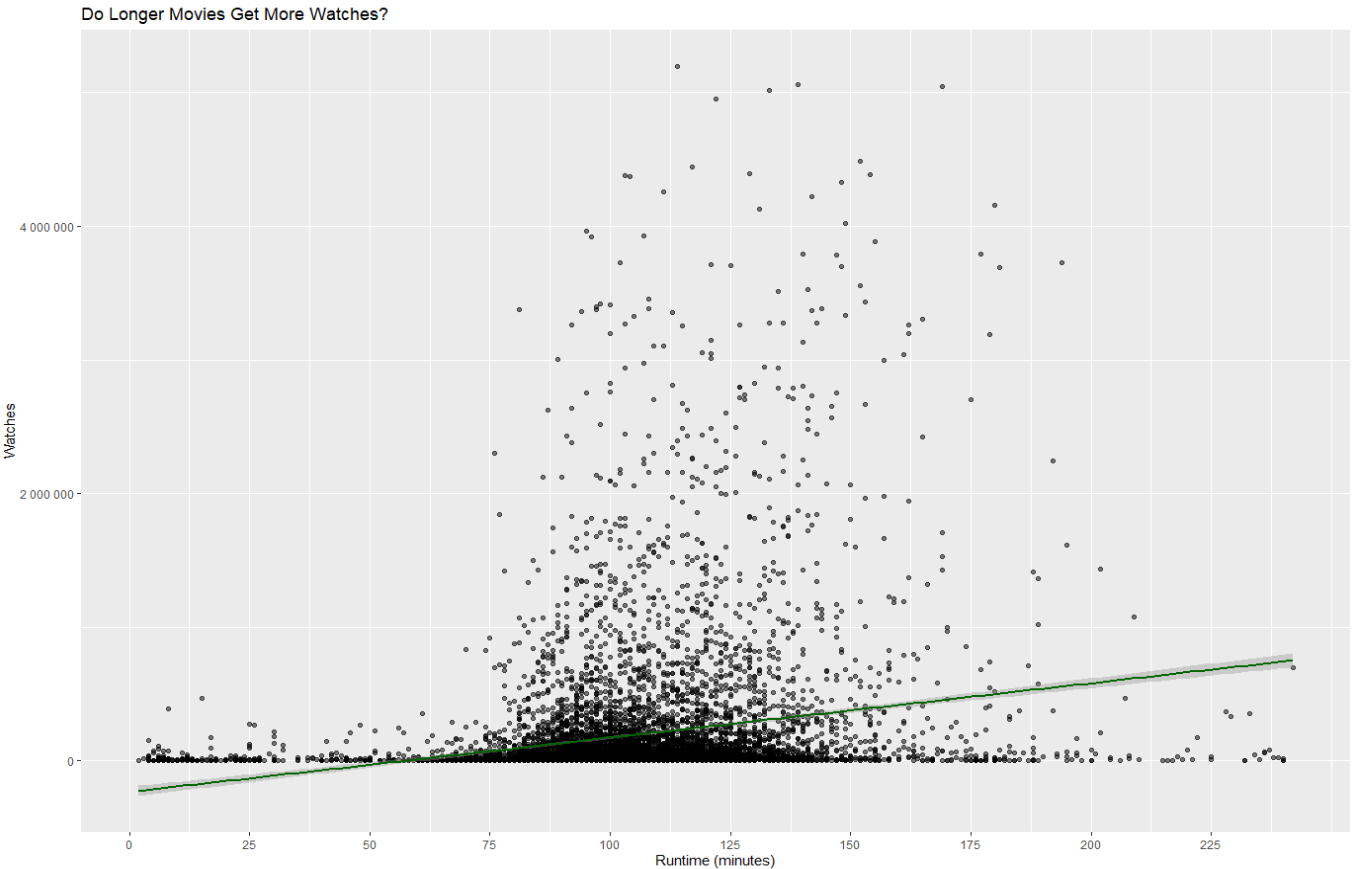
*Note: All visualisations are in full size in the appendix below.*

**Are some genres associated with higher average ratings?**

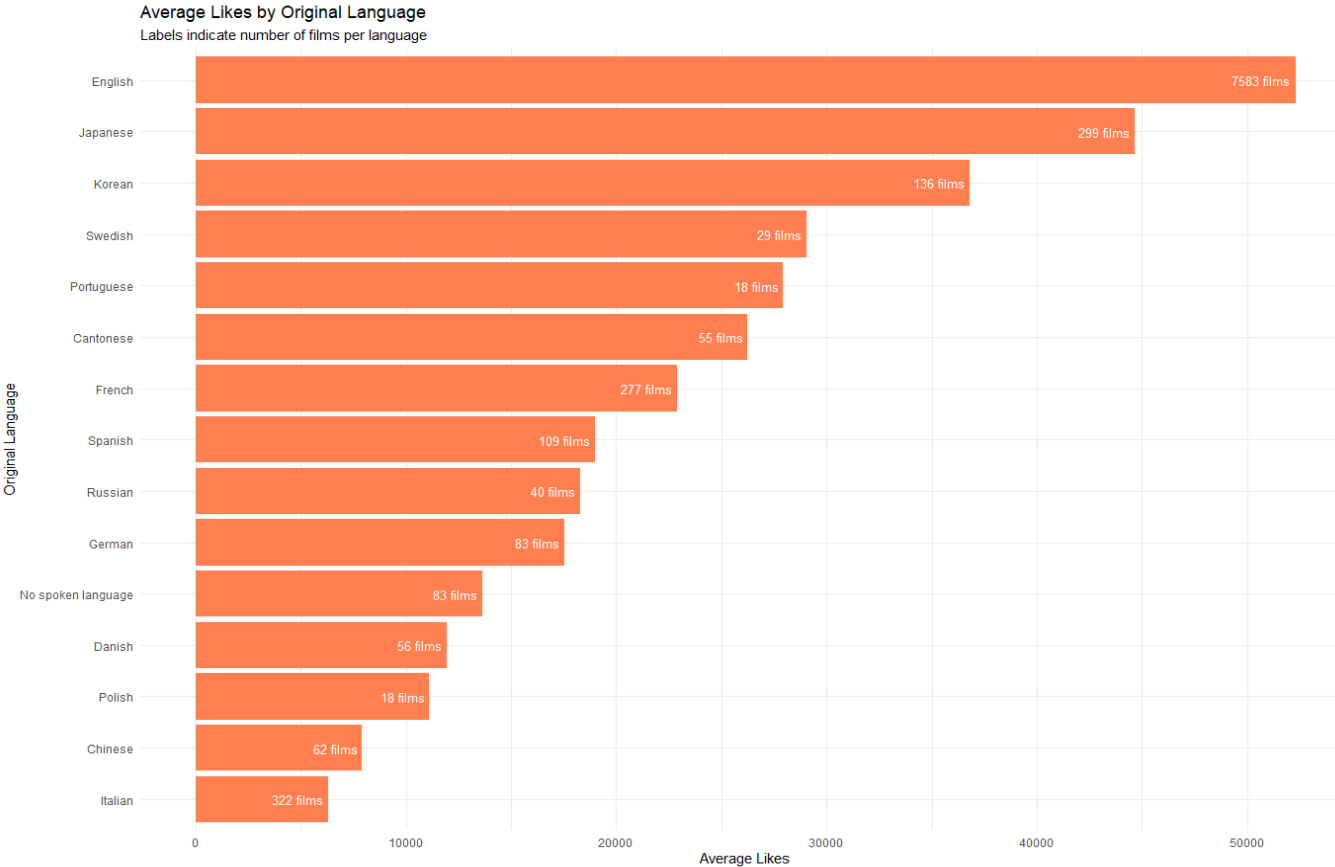


**Do longer runtimes lead to significantly more likes?**

****



**Are films made in English watched more than films in other languages?**



# Step 5: Key Insights and Conclusions

The analysis revealed several interesting insights:

**Are some genres associated with higher average ratings?**

* The largest genre by far is Drama, with 21.9% of films, and despite its large make up as a percentage of the data, it is still rated near the very top of the average ratings at 3.5, with 3.6 being the highest (Documentaries).
* The second largest category, comedy at 14.4% has far more middling average reviews at 3.2
* Horror films make up 8.9% of total films (third largest category) but are the worst average rated films at 3.0 out of 5.
* Documentaries and History films are among the top reviewed by Letterboxd users.

**Do longer runtimes lead to significantly more likes?**

* There is a clear delineation when likes start to significantly pick up for films after a certain runtime. Films past 75 minutes (1h15m) runtime on average begin to experience more likes, with the trend beginning to fall off after 150 minutes (2h30m).
* Films receive less likes when the runtime is short, and less when the runtime is long.

I did further analysis into the number of watches for films which revealed that:

* Watches & Likes are very strongly correlated. R² = 0.9035
* Films receive many more watches past 75 minutes (1h15m) runtime and fall off after roughly 137 minutes (2h17m).
* There are outliers, and these films tend to be classic box office hits, receiving millions of views. The greatest being ‘Fight Club’ with 5 million watches and a runtime of 139 minutes, all the way to The Godfather Part II with 1.4 million watches and a runtime of 202 minutes and ‘The Titanic’ with 3.7 million watches and a runtime of 194 minutes.

**Are films made in English watched more than films in other languages?**

* As expected, the vast number of films within the dataset are in English (7,583 films, 81.1%). More English films are made than that of any other language in the world.
* Despite only 299 Japanese films in the dataset (3.2%), they have the second highest likes with just under 45,000 average likes.
* Korean also surprises with 136 films (1.4%) at almost 37,000 likes
* The second largest language in the dataset, ‘Italian’ with 322 films (3.4%) has the lowest average likes at 6000 average likes

# Conclusions:

* **Are some genres associated with higher average ratings?**

Yes, although it is incredibly complex and varies wildly by quality of film. There is an enduring love of drama which has the largest review relative to number of films. Documentary films are surprisingly well reviewed.

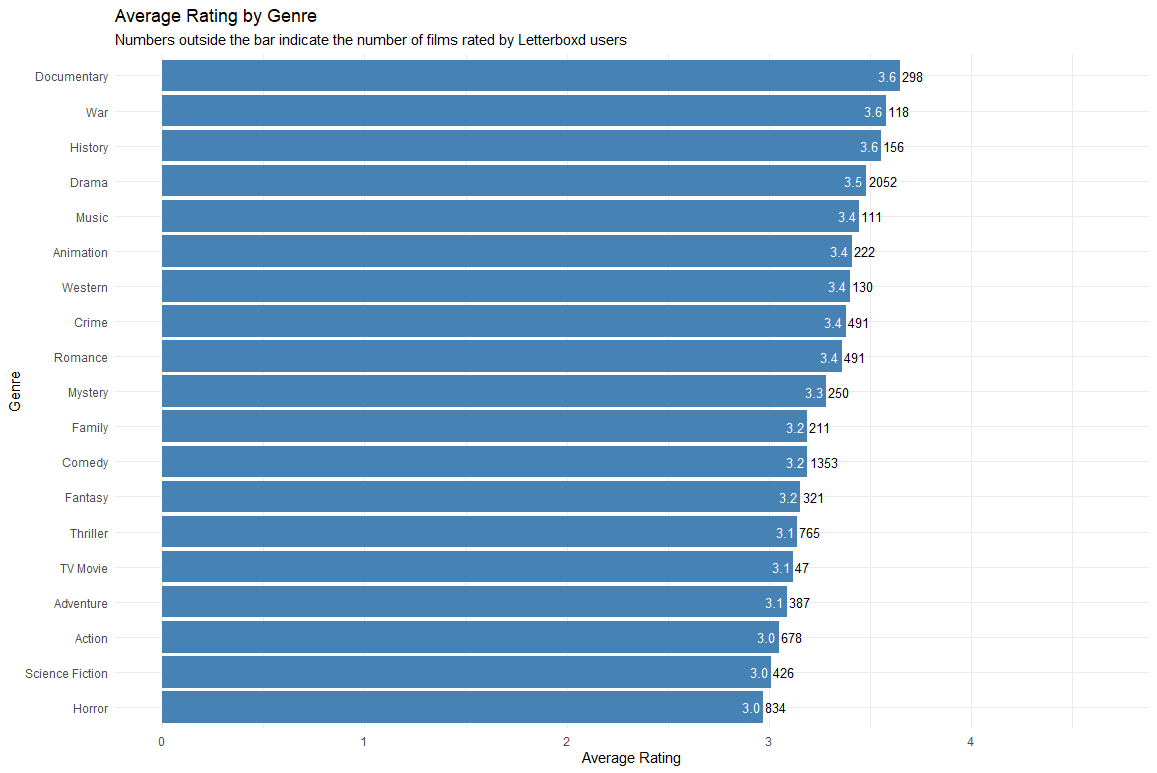
* **Do longer runtimes lead to significantly more likes?**

Yes, to a point. There are significantly more likes for films above 75 minutes and below 150. This indicates that short films are less popular film types and the general film viewing audience prefers films that aren’t excessively long.

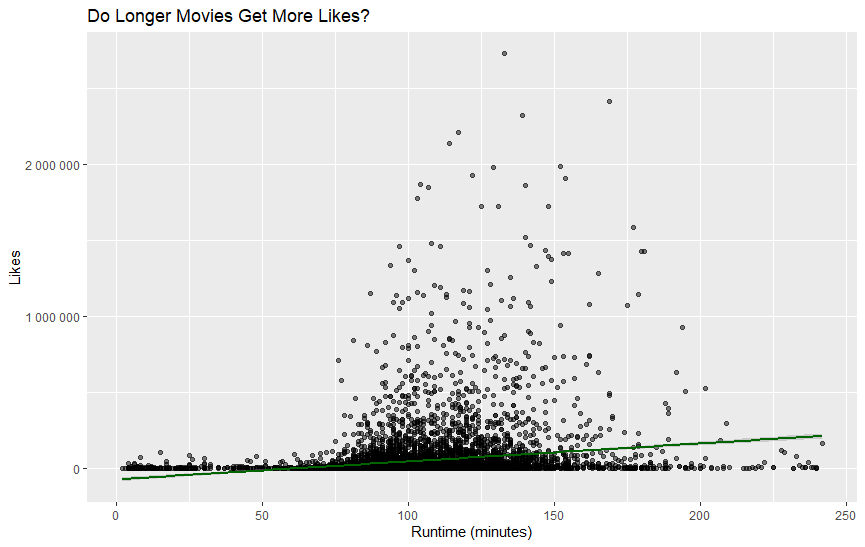
* **Are films made in English watched more than films in other languages?**

Yes, although there are many more films made in English, so this is not surprising. What is surprising is that Japanese and Korean films receive on average many more likes than other language films.

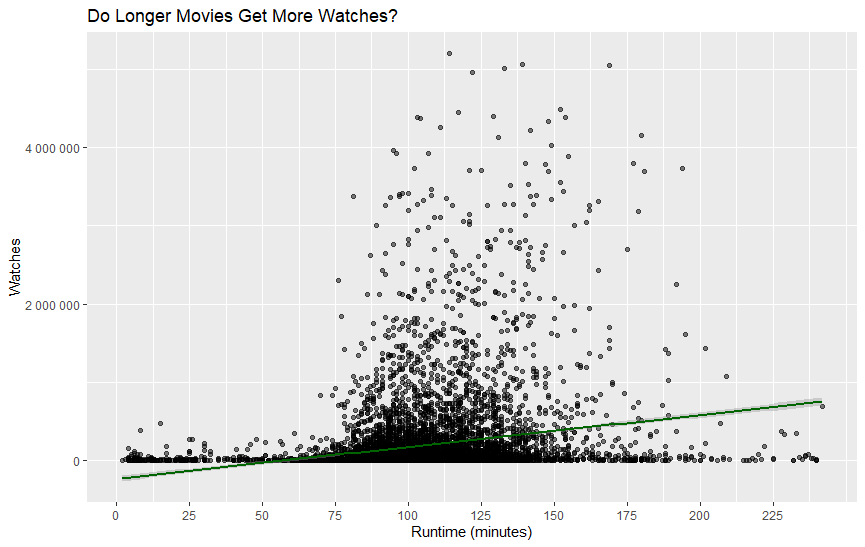
**Appendix A**: Average Rating by Genre



**Appendix B**: Do longer Movies Get More Watches?



**Appendix C**: Do longer movies get more watches?



**Appendix D**: Correlation between Watches and Likes

This chart cannot be rotated.

**Appendix E**: Average likes by Original Language