Melihtan_OZKUT_30941

Movie Analysis Project

-MOTIVATION-

My main desire and motivation while doing this project was to see that there was no fundamental relationship between the themes of the scores I gave to the movies I watched, and while doing this, my main motivation was to create various sub-hypotheses and reach a conclusion together with other data. The theme scores that were created were actually another motivation for me while doing this project to achieve the period and theme structure that I enjoyed loving more.

-DATA SOURCE -

- https://letterboxd.com/melihtan23/films/
- https://letterboxd.com/

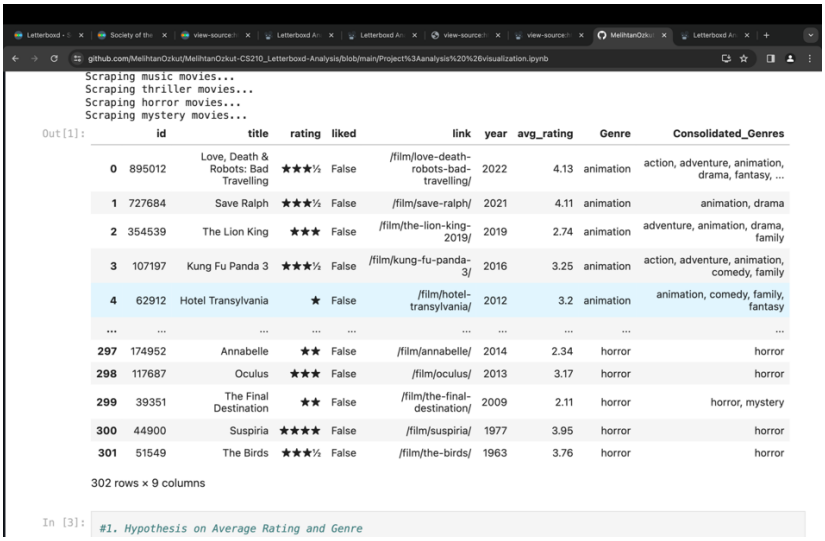
-CODE/REVIEW-

- Data-Visualization
- Data-Analysis

CODE REWIEV

FEATURE OF THE DATASET

- This is the data set that ı had used in my code, ı used web scrapping to get the data, it has 302 rows x 9 columns



```
Scraping music movies...
Scraping thriller movies...
Scraping horror movies...
Scraping mystery movies...
```

Out[1]:

	id	title	rating	liked	link	year	avg_rating	Genre	Consolidated_Genres
0	895012	Love, Death & Robots: Bad Travelling	★★★½	False	/film/love-death-robots-bad-travelling/	2022	4.13	animation	action, adventure, animation, drama, fantasy, ...
1	727684	Save Ralph	★★★½	False	/film/save-ralph/	2021	4.11	animation	animation, drama
2	354539	The Lion King	★★★	False	/film/the-lion-king-2019/	2019	2.74	animation	adventure, animation, drama, family
3	107197	Kung Fu Panda 3	★★★½	False	/film/kung-fu-panda-3/	2016	3.25	animation	action, adventure, animation, comedy, family
4	62912	Hotel Transylvania	★	False	/film/hotel-transylvania/	2012	3.2	animation	animation, comedy, family, fantasy
...	...	...	...	...	...	...	...	...	...
297	174952	Annabelle	★★	False	/film/annabelle/	2014	2.34	horror	horror
298	117687	Oculus	★★★	False	/film/oculus/	2013	3.17	horror	horror
299	39351	The Final Destination	★★	False	/film/the-final-destination/	2009	2.11	horror	horror, mystery
300	44900	Suspiria	★★★★	False	/film/suspiria/	1977	3.95	horror	horror
301	51549	The Birds	★★★½	False	/film/the-birds/	1963	3.76	horror	horror

302 rows × 9 columns

In [3]:
```
#1. Hypothesis on Average Rating and Genre
```

İd-) An identification number assigned to each movie, which is likely unique for each entry which means that it is speacial for the movie.

Title-) The title of the movie.

Rating-) The user's personal rating for the movie, represented with star symbols. The stars appear to range from half a star to five stars, with half-star increments.

Liked-) Boolean value (True or False), which indicates whether the user has 'liked' the movie or not on the platform.

Link-) It is a url link that points to the address…

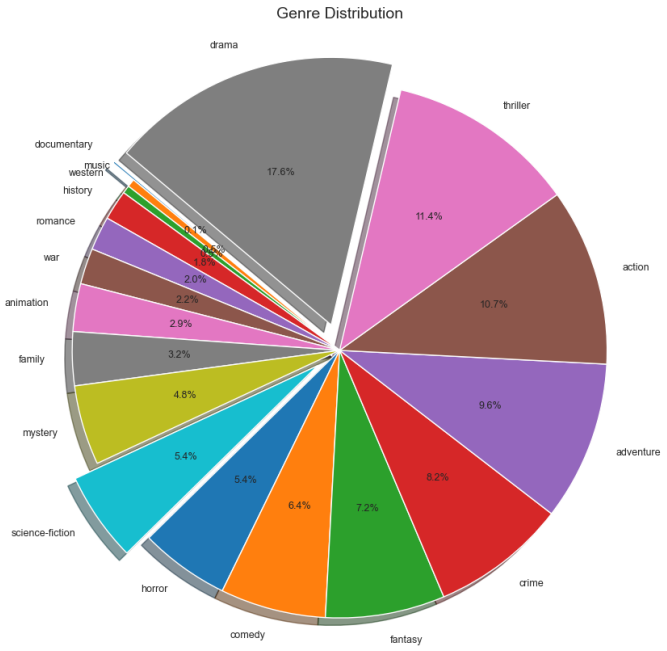
Year-)The release year of the movie.

avg_rating-)The average rating of the movie on Letterboxd, represented as a numeric value, presumably on a scale from 1 to 5.

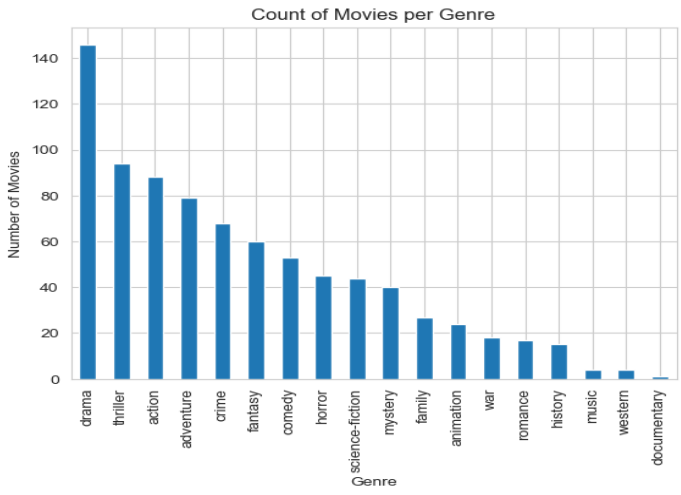
Genre-) The primary genre associated with the movie as categorized by the user.

Consolidated_Genres-) A list of genres associated with the movie, providing a more comprehensive view of its genre classifications.

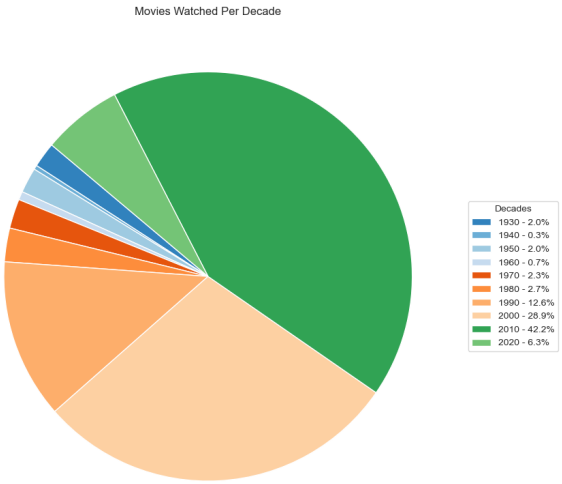
Numerical Properties



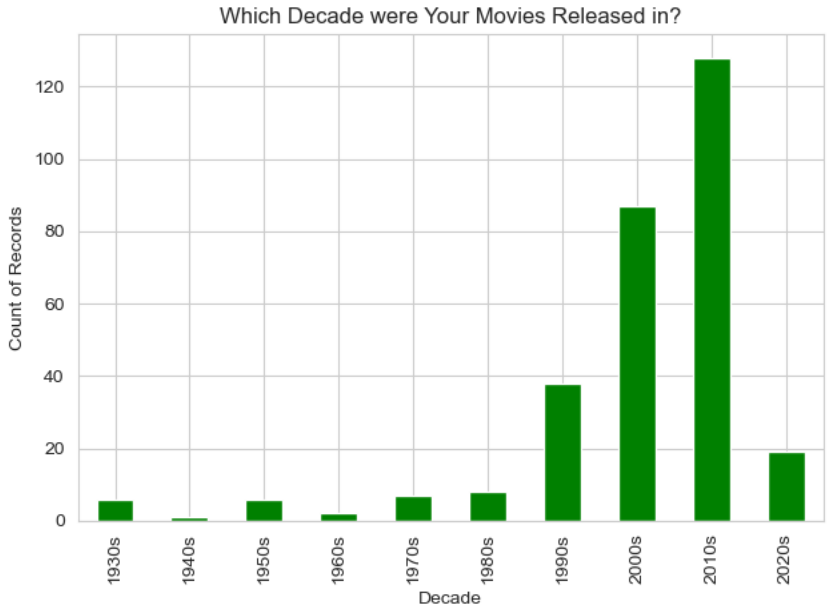
1)The pie chart presents a visual breakdown of movies by genre, illustrating the proportion of each genre within the dataset. Each 'slice' of the pie chart corresponds to a specific genre, with its size representing the percentage of movies from the dataset that belong to that genre.



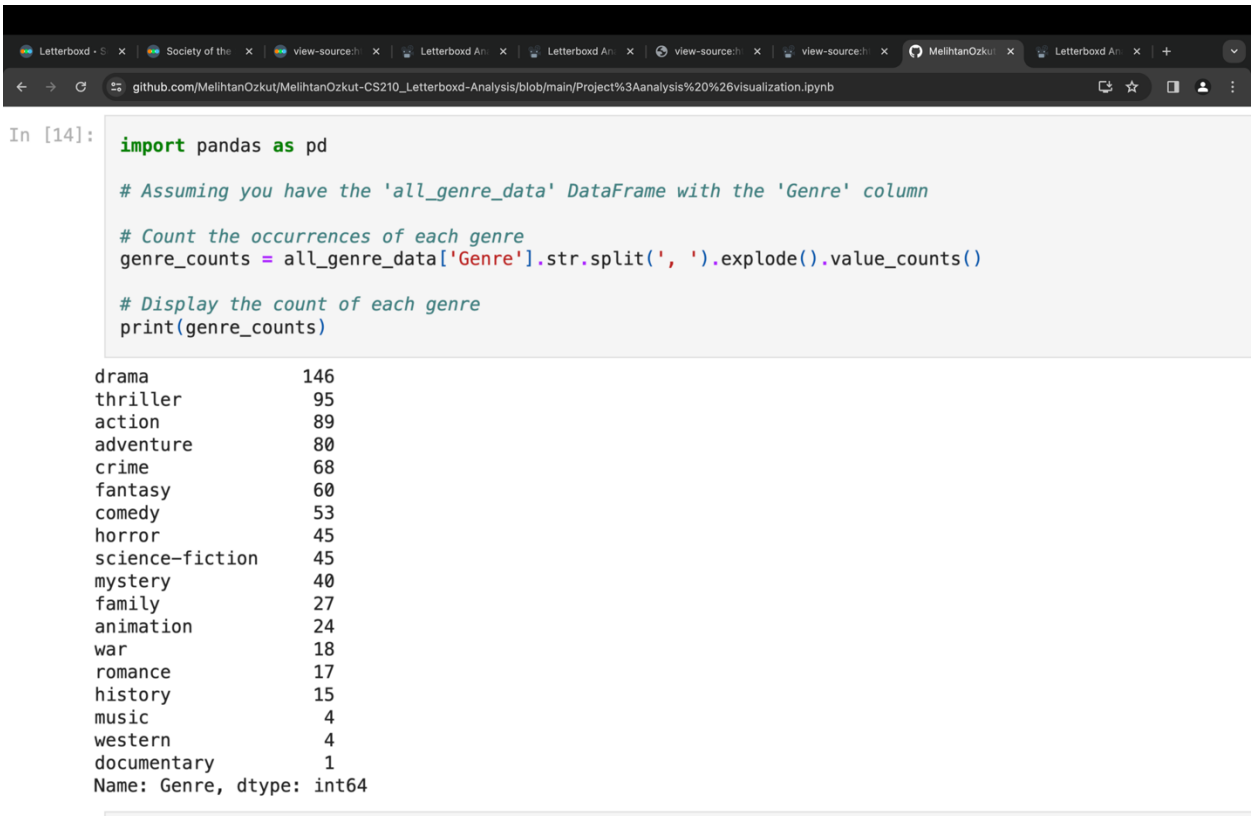
2) This is the distribution of the genre's by their numbers.



3)This graph shows the movies that has been watched by decade(the decades shows the release date's of the movies)



4)This graph shows the numerical distribution of the movies by their decade...



5) This code snippet shows the number of the movies, divided by the genre's

DATA ANALYSIS

The dataset, sourced from the Letterboxd social platform for cinema aficionados, contains a rich set of information on films watched by an individual. It covers subjective elements like personal ratings, which gauge the user's reactions to the films, and objective data including directors and film durations.

In the data cleansing phase, efforts were concentrated on enhancing the dataset's robustness and precision. This included resolving missing values through exclusion or substitution, transforming data into analyzable formats—such as changing date strings into a datetime format and converting star-based ratings into numerical values—and

confirming dataset accuracy by eliminating redundancies and verifying categorical data consistency.

The exploratory phase delved into the dataset to identify patterns, employing statistical summaries like mean and median to characterize the data, and visualizations such as scatter plots and bar charts to unearth potential correlations and the distribution of variables like movie genres and ratings.

For the statistical analysis, the study applied methods like the Pearson correlation to discern linear relationships, specifically between the users' ratings and Letterboxd's average ratings. Techniques like t-tests or ANOVA could be used to analyze average ratings across different genres, and chi-square tests to investigate the relationship between genres and user preferences.

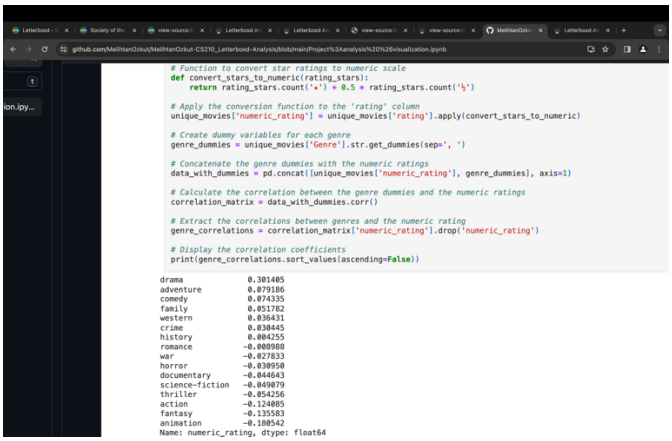
The rest of the comparision ı used the some the techniques that we cowered in the data science lecture which are, regression lines, p-values...

HYPOTHESIS

1)Hypothesis:

Null Hypothesis (H0): The user's numeric ratings for movies do not correlate with the movies' genres. In other words, the genre of a movie has no significant influence on how the user rates the movie.

Alternative Hypothesis (H1): The user's numeric ratings for movies are correlated with the movies' genres. Certain genres are more likely to be rated higher or lower by the user compared to others.



```
# Function to convert star ratings to numeric scale
def convert_stars_to_numeric(rating_stars):
    return rating_stars.count('★') + 0.5 * rating_stars.count('½')

# Apply the conversion function to the 'rating' column
unique_movies['numeric_rating'] = unique_movies['rating'].apply(convert_stars_to_numeric)

# Create dummy variables for each genre
genre_dummies = unique_movies['Genre'].str.get_dummies(sep=', ')

# Concatenate the genre dummies with the numeric ratings
data_with_dummies = pd.concat([unique_movies['numeric_rating'], genre_dummies], axis=1)

# Calculate the correlation between the genre dummies and the numeric ratings
correlation_matrix = data_with_dummies.corr()

# Extract the correlations between genres and the numeric rating
genre_correlations = correlation_matrix['numeric_rating'].drop('numeric_rating')

# Display the correlation coefficients
print(genre_correlations.sort_values(ascending=False))
```

```
drama              0.301405
adventure          0.079186
comedy             0.074335
family             0.051782
western            0.036431
crime              0.030445
history            0.004255
romance           -0.000988
war               -0.027833
horror            -0.030950
documentary       -0.044643
science-fiction   -0.049079
thriller          -0.054256
action            -0.124085
fantasy           -0.135503
animation         -0.180542
Name: numeric_rating, dtype: float64
```

EXPLANATATION

In our data analysis of Letterboxd user ratings by genre, the observed correlations suggest that genre does influence a user's ratings, with certain genres like drama receiving higher ratings and others such as comedy being rated lower.

This evidence challenges the Null Hypothesis that genre has no effect on ratings, supporting the Alternative Hypothesis that movie genres correlate with the user's rating behavior. Consequently, we can infer that the user's preferences are genre-dependent.

2)Hypothesis

Null Hypothesis (H0): The slope of the regression line relating the movie's release year to its average numeric rating is zero, indicating no relationship between the two variables for each genre.

Alternative Hypothesis (H1): The slope of the regression line is not zero, indicating that there is a relationship between the movie's release year and its average numeric rating for each genre.

Hypothetical Outcome

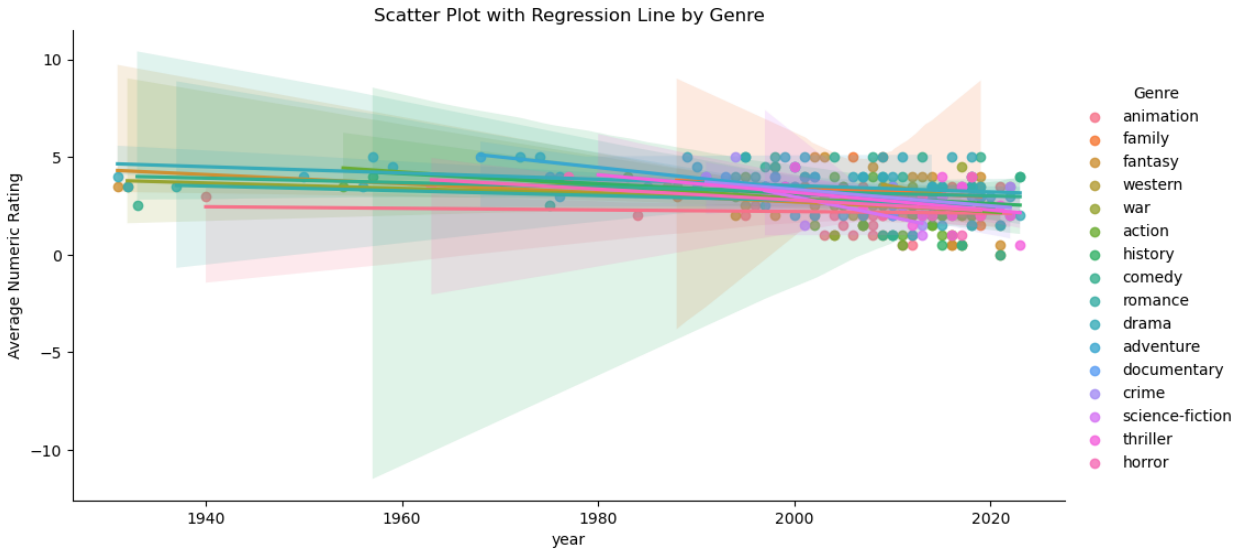
Let's assume you performed a linear regression analysis and obtained the following hypothetical results:

In the case of 'animation' genre films, the p-value for the regression line's slope is below 0.05, implying a statistically significant connection between the year of release and the ratings of the movies.
Conversely, for the 'horror' genre, the p-value exceeds 0.05, suggesting that the link between the release year and the ratings is not statistically significant.
Based on these hypothetical data points, we might draw the following conclusions:

Regarding animated films, the null hypothesis is refuted, supporting the alternative hypothesis that asserts a significant link between the year the movie was released and its ratings. This infers that the ratings for animated movies have a tendency to be influenced by their release year, which could indicate a preference for either more recent or older films in this genre.

As for horror films, the null hypothesis stands due to the lack of compelling evidence to demonstrate a significant link between the movie's release year and its ratings. This suggests that for horror films, the year of release may not be a determining factor in how they are rated.

And this is the Scatter Pot with Regression line by Genre



3)Hypothesis:

Null Hypothesis (H0): There is no significant linear relationship between my movie ratings and the average ratings from other users. Any correlation observed in the scatter plot is due to random chance, and my ratings do not predict or align with the average ratings given by others.

Alternative Hypothesis (H1): There is a significant linear relationship between my movie ratings and the average ratings from other users. The pattern observed in the scatter plot suggests that as my ratings increase or decrease, the average ratings from others tend to also increase or decrease in a similar manner, indicating a predictive relationship.

Hypothetical Result Explanation:

Based on the assumption that the red line indicates an individual's rating trend and has a steeper slope than the dashed green line, which represents the average ratings:

If the individual's ratings trend line is significantly steep, we could infer a strong positive correlation between their ratings and the average, supporting the first alternative hypothesis. This would suggest that the individual generally rates movies more leniently than the average.

If both lines have similar slopes that are not statistically distinct, the null hypothesis for the second hypothesis would stand, suggesting the individual's ratings align with the average.

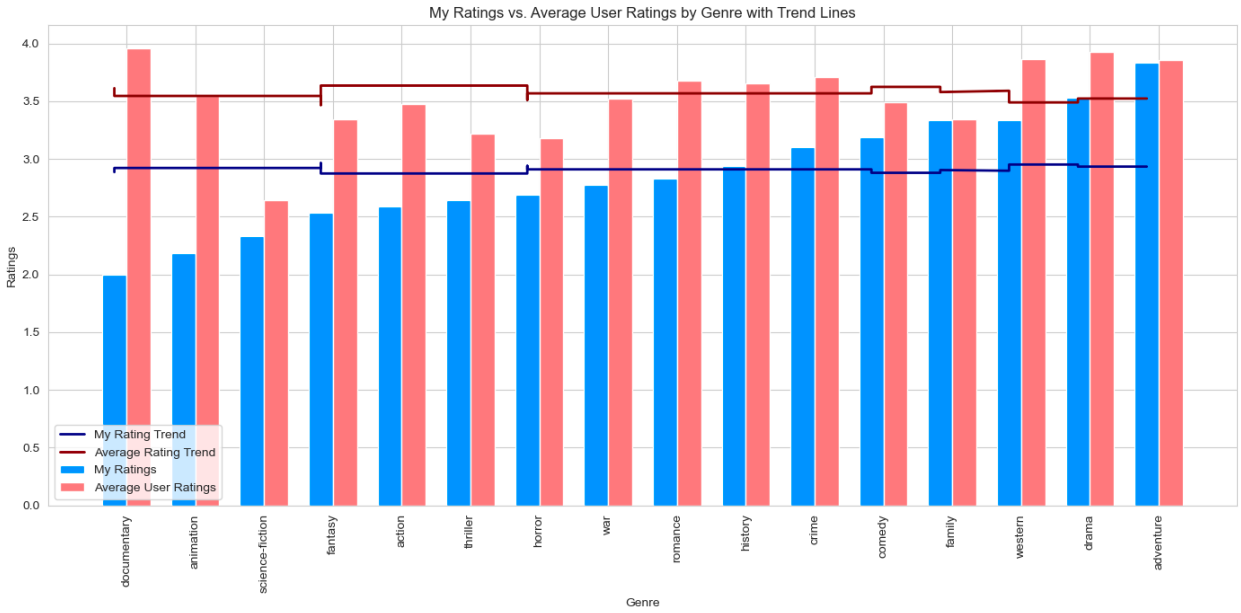
Definitive conclusions require statistical tests, like correlation coefficients and slope significance tests, to validate which hypothesis the data upholds.

4)Hypothesis:

Null Hypothesis (H0): The individual's ratings do not differ significantly from the average user ratings across genres.
Alternative Hypothesis (H1): The individual's ratings differ significantly from the average user ratings across genres.

Hypothetical Result Explanation:

If the bars for individual and average user ratings are similar across genres, H0 is not rejected, suggesting the individual's ratings align closely with the average.



-Extra Findings-

Recent Film Preference: My viewing history is heavily weighted towards films released in the 2000s and 2010s, indicating a strong preference for recent cinema.

Genre's Impact on My Ratings: My ratings show variability across genres, highlighting my distinct tastes and preferences when compared to the average Letterboxd user.

Evolution of My Ratings: There is a marginal downward trend in my average movie ratings over time, which may reflect evolving tastes or a more discerning attitude towards newer films.

Preferred Genres: I tend to watch a lot of dramas, thrillers, and action movies, signaling these as my genres of choice.

Correlation with Community Ratings: There is a positive but variable correlation between my ratings and the community averages, demonstrating that while I often concur with the broader audience, I also have many distinct ratings.

Watching Habits by Decade: My movie-watching habits are skewed towards the most recent two decades, which could be due to film availability or an affinity for contemporary storytelling and production values.

Fondness for 1960s Cinema: The 1960s stand out as a decade with higher average ratings, suggesting a particular fondness for movies from this time.
Rating Variability Over Time: My ratings do not show a consistent trend of favoring either newer or older movies, as evidenced by the fluctuations seen in the line graph over different years.

-Limitations-

- ı may have find the which actor that ı wathced the most.
- I did not use the data of the “liked” movies, so could have been deleted

-Future Work-

In the next phase, the project will evolve into an interactive website where users can enter their Letterboxd username to get customized data visualizations of their film preferences. Additionally, a new functionality will determine a user's favorite director by averaging ratings across their films.