MY LETTERBOXD DATA
(DSA 210 TERM PROJECT)

Muhammet Emin Çal 30782

Introduction

This project focuses on analyzing personal **Letterboxd ratings** data to identify patterns in movie preferences. The main objective is to explore how movie ratings vary based on attributes such as **genre**, **runtime**, and **release year**. By examining these aspects, the project aims to uncover trends in movie preferences and how they may have changed over time.

The analysis is based on data from the **Letterboxd platform**, where movies are rated on a 5-star scale. The project investigates factors that may influence these ratings, including whether certain genres tend to receive higher ratings, whether ratings differ across decades, and how runtime correlates with average ratings. Additionally, hypothesis testing is used to assess whether ratings differ significantly between movies released before and after a particular year.

The techniques used in this project include data cleaning, exploratory data analysis (EDA), statistical hypothesis testing, and machine learning methods. The results aim to provide insights into personal movie-watching habits and demonstrate how data analysis can reveal underlying patterns in individual preferences.

Data Collection and Preprocessing

Data Source

The dataset for this project is derived from my **Letterboxd** profile, which contains a collection of movies I have watched and rated. Letterboxd is a social platform that allows users to track movies they've seen, rate them on a 5-star scale, and categorize them by genre, release year, and other attributes. The data for this project was extracted either through the **Letterboxd API** or by exporting my ratings in CSV format from my profile. The dataset includes key attributes for each movie, such as:

- **Title**: The name of the movie
- **Genres**: One or more genres associated with the movie
- **Release Year**: The year the movie was released
- **Average Rating**: The rating I assigned to the movie on a 5-star scale
- **Runtime**: The duration of the movie in minutes
- **Director**: The director of the movie

Data Collection Process

The data was collected by exporting my movie ratings from my **Letterboxd** profile in CSV format. This format provides a convenient and structured way to access the necessary information. Once exported, the data was loaded into a Pandas DataFrame for further analysis. A few missing or incomplete values were present in the raw dataset, which required preprocessing to ensure consistency and usability during analysis.

Data Cleaning and Preprocessing

1. **Handling Missing Data**:
   Missing values were present in several columns, particularly for the **Genres**, **Average Rating**, and **Runtime** attributes. To address these issues:
   - The **Genres** column was filled with an empty list ('[]') when no genres were assigned. Additionally, genres were normalized to a standardized format using a genre mapping dictionary to ensure consistency (e.g., "Action-packed space and alien sagas" was mapped to "Action").

    - The **Average Rating** column was filled with the median rating when missing values were detected, as this was the most appropriate imputation method given the skewed distribution of ratings.
    - The **Runtime** column was similarly imputed using the median value to fill in missing runtime data.
2. **Creating Decade Categories**:
   To analyze trends over time, the **Release Year** was used to categorize movies into decades. The **Decade** column was created by dividing the release year by 10, multiplying it by 10, which resulted in the year range for each decade (e.g., 1980 for movies released between 1980 and 1989).
3. **Dropping Irrelevant Columns**:
   Certain columns that were not relevant for the analysis (such as **Watches** and other extra metadata) were dropped to streamline the dataset and focus only on the key features needed for analysis.
4. **Exploding Multi-Genre Entries**:
   Since movies can belong to multiple genres, the **Genres** column was "exploded," meaning that each movie was represented multiple times—once for each genre it was associated with. This transformation allowed for more granular analysis of ratings by individual genres.

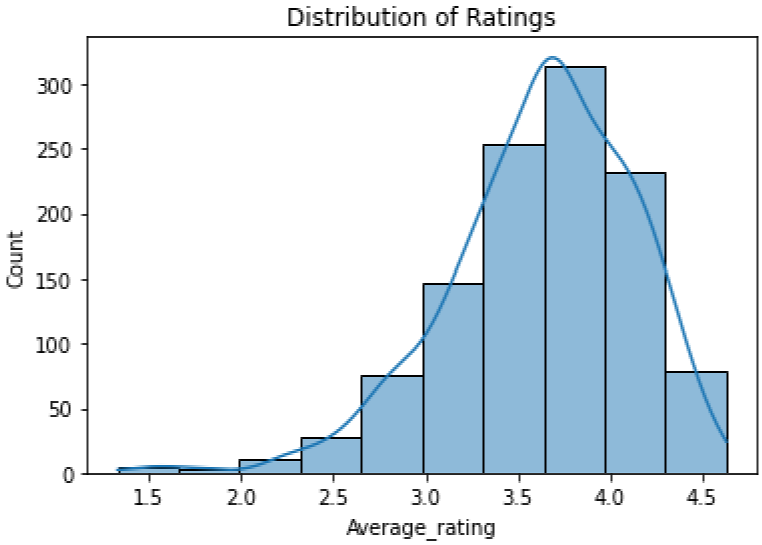
Once these preprocessing steps were completed, the data was ready for deeper exploratory analysis and visualization.

Exploratory Data Analysis (EDA)

Overview

The primary goal of the exploratory data analysis (EDA) was to identify patterns, trends, and insights from my Letterboxd ratings dataset. During this phase, various statistical and visual methods were employed to better understand the relationships between different variables such as movie genres, release years, runtime, and ratings.

Here's the distribution of ratings:



By this plot, it can be seen that I mostly enjoy what I watch.

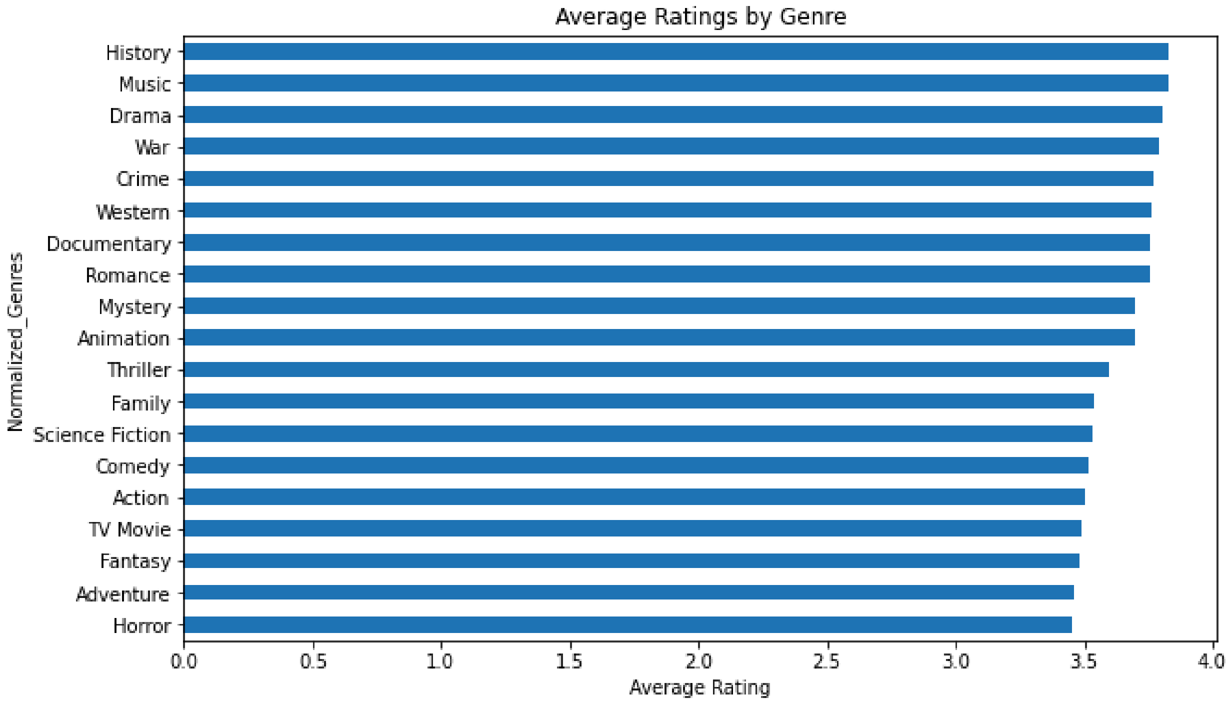
Genre Analysis

To explore how my preferences for different movie genres influenced my ratings, the following steps were performed:

Average Ratings by Genre:

The average rating for each genre was calculated by grouping the data based on Normalized_Genres. This gave insights into which genres received higher ratings on average. A horizontal bar plot was generated to visually compare the average ratings for different genres.

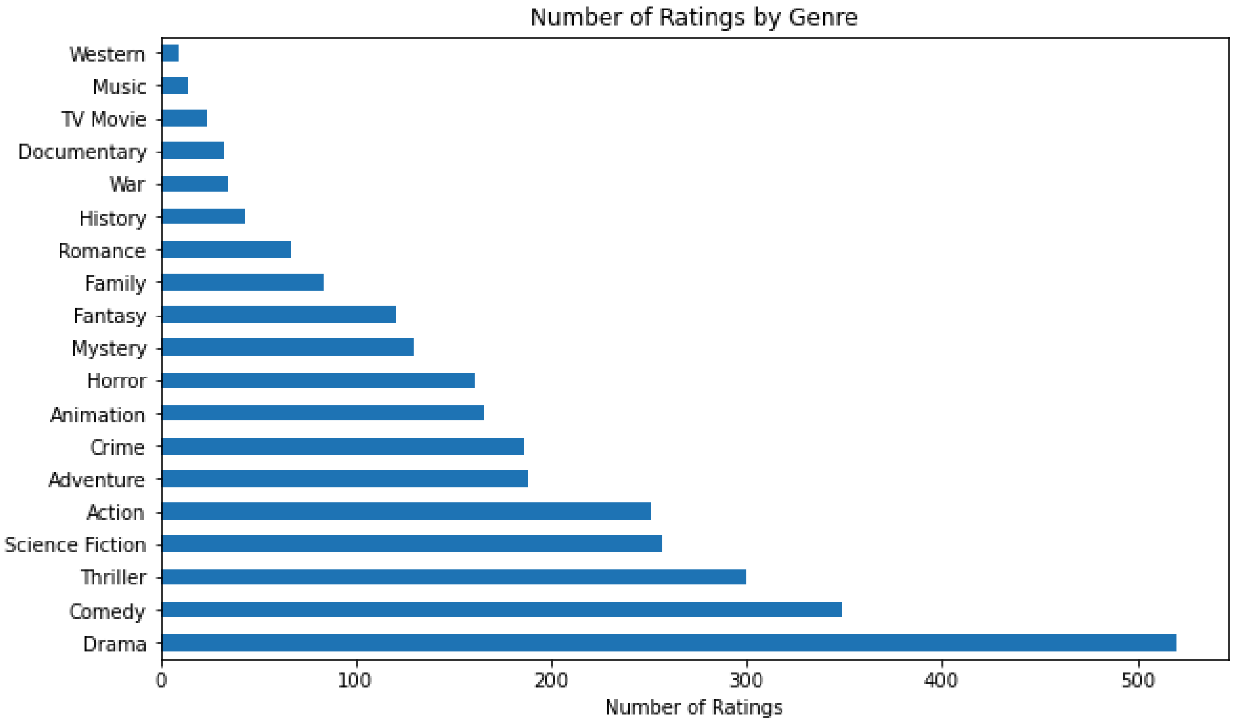
Plot:



Number of Ratings by Genre:

To understand how many movies were rated in each genre, a bar plot was generated showing the number of ratings per genre.

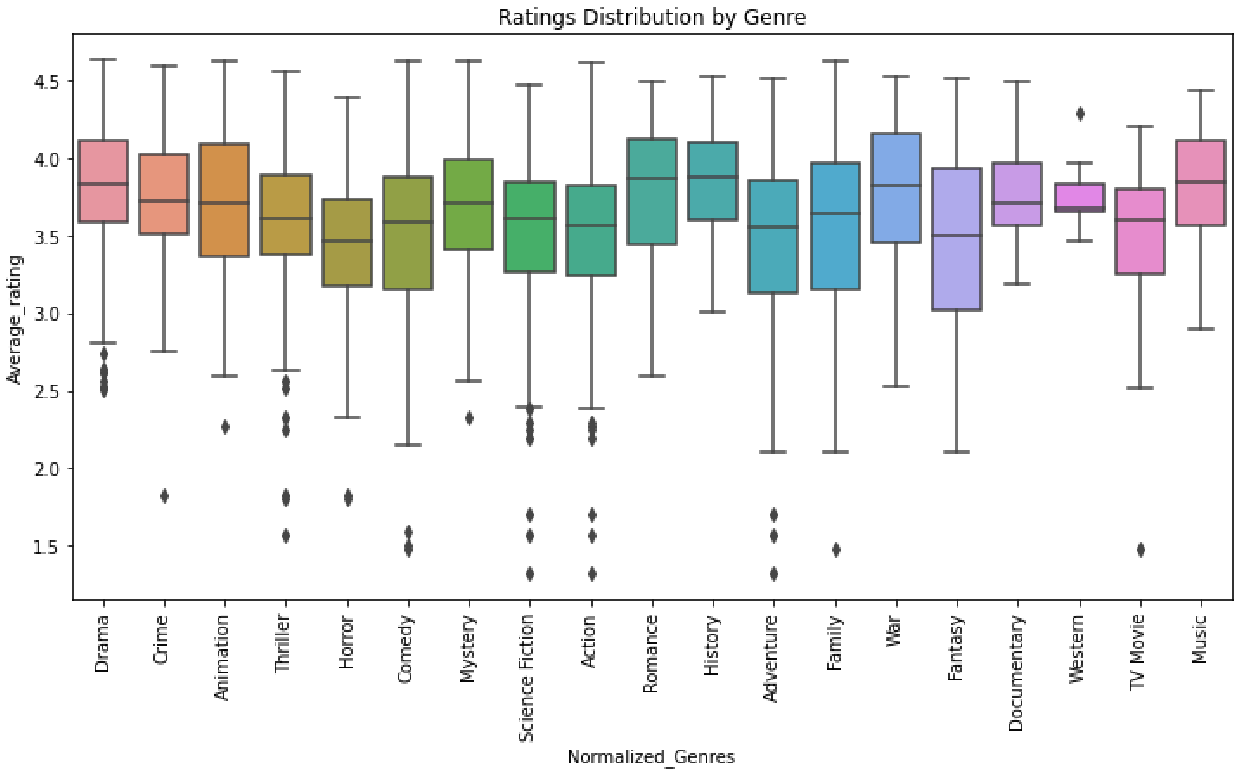
Plot:



Ratings Distribution by Genre:

A box plot was used to visualize the distribution of ratings across different genres. This allowed for a comparison of the spread and outliers in ratings within each genre, shedding light on any genres where my ratings varied more significantly.

Plot:



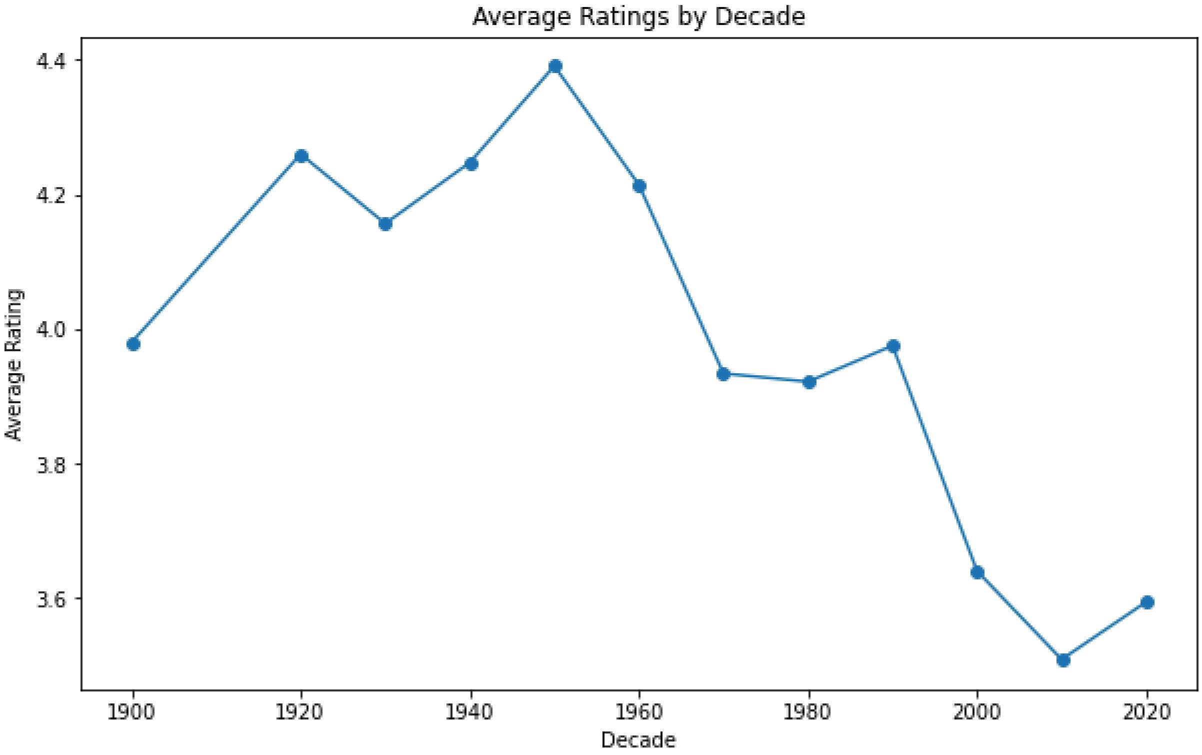
Temporal Trends

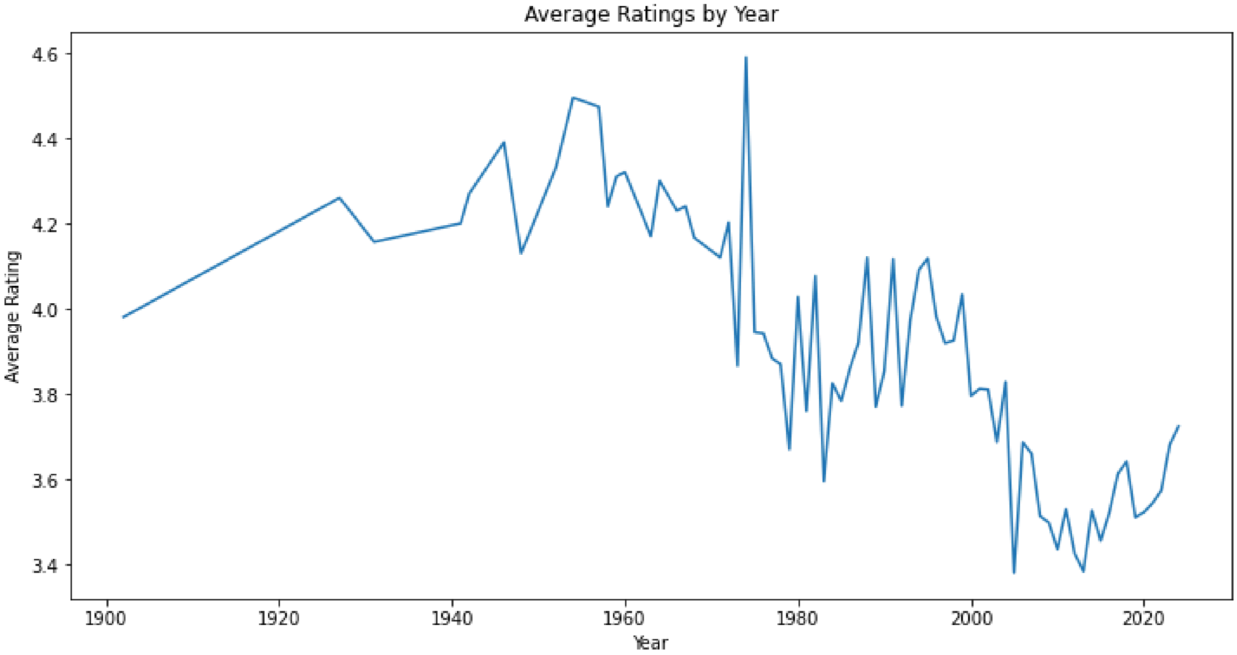
To explore how my ratings have changed over time, I examined the relationship between release year and average rating:

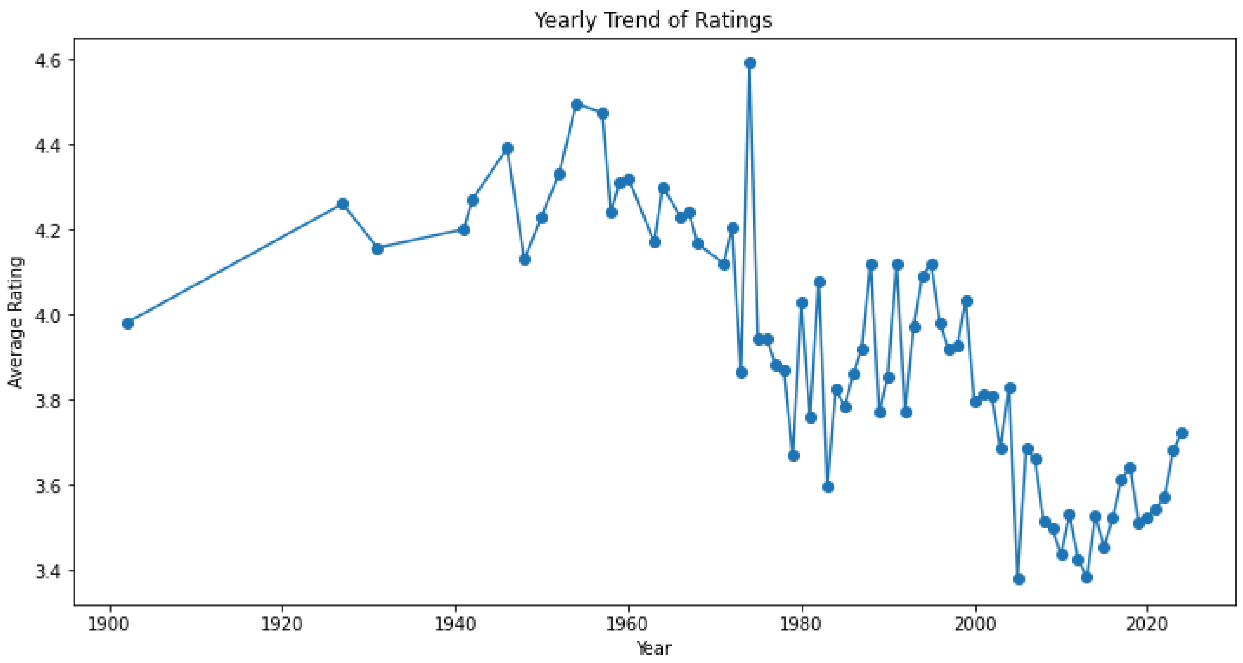
Average Ratings by Decade/Year and Yearly Trend of Ratings:

The average rating by decade and years was calculated, grouping the data by Decade and Years. A line plot was generated to visualize how my ratings have changed over the decades and years. This analysis helped identify any trends in how my movie ratings might have evolved with time.

Plots:


Average Ratings by Decade
Average Rating
Decade
4.4
4.2
4.0
3.8
3.6
1900
1920
1940
1960
1980
2000
2020

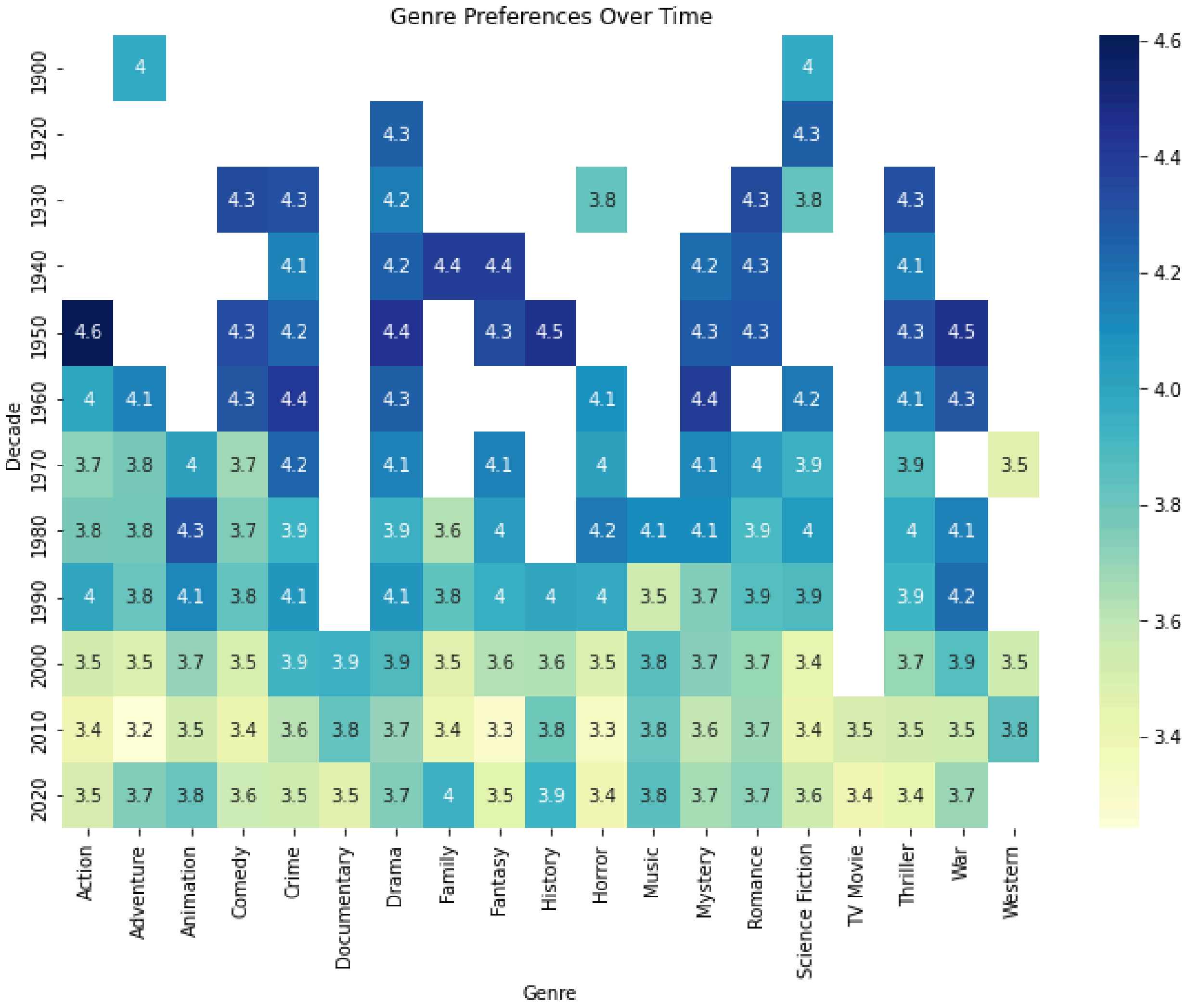

Average Ratings by Year
Average Rating
Year
4.6
4.4
4.2
4.0
3.8
3.6
3.4
1900
1920
1940
1960
1980
2000
2020

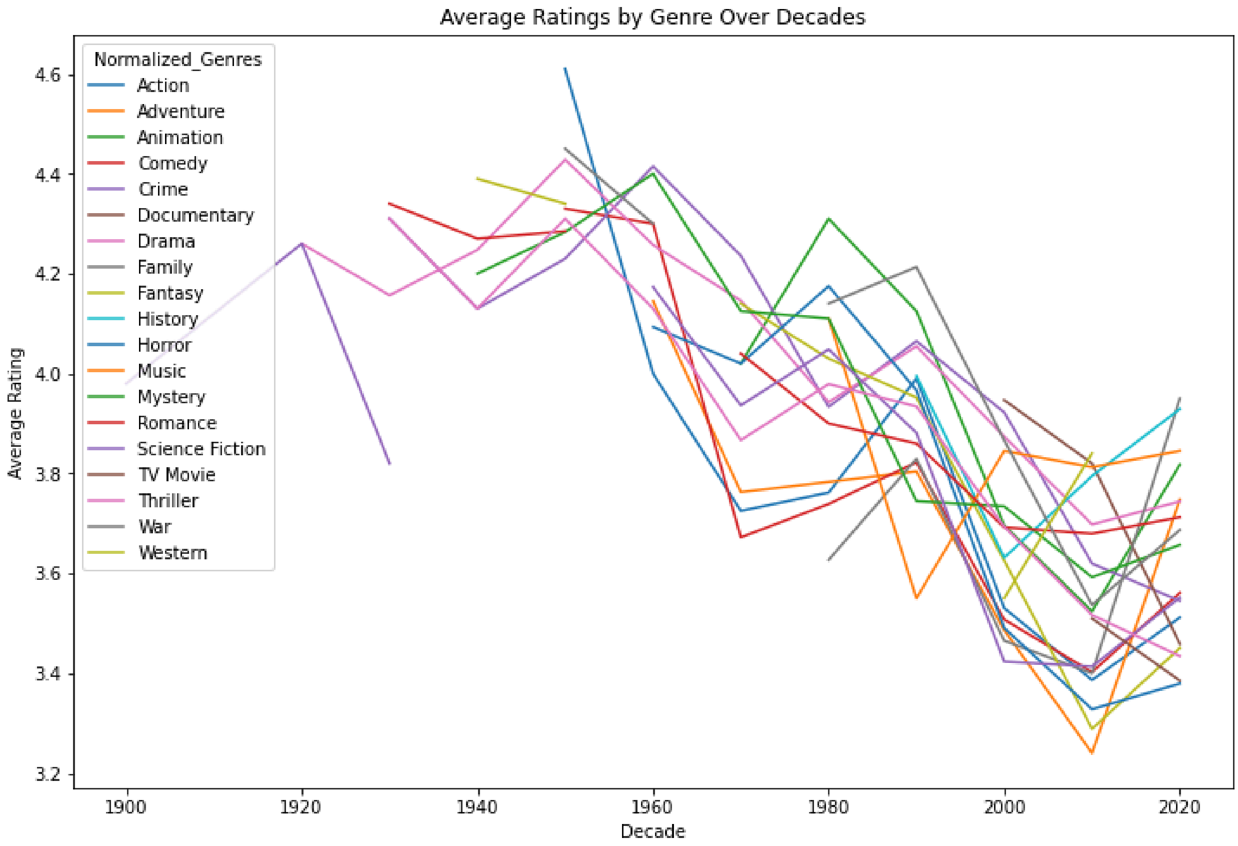


Decade and Genre Interaction:

By analyzing how genres performed across different decades, I could assess whether my genre preferences had shifted over time. A heatmap of average ratings for genres by decade was created to reveal trends in movie preferences over time. Also, another plot to show the relation between average rating, decade and genre together.

Plots:


Genre Preferences Over Time
Decade
1900
4
4
1920
4.3
4.3
1930
4.3
4.3
4.2
3.8
4.3
3.8
4.3
1940
4.1
4.2
4.4
4.4
4.2
4.3
4.1
1950
4.6
4.3
4.2
4.4
4.3
4.5
4.3
4.3
4.3
4.5
1960
4
4.1
4.3
4.4
4.3
4.1
4.4
4.2
4.1
4.3
1970
3.7
3.8
4
3.7
4.2
4.1
4.1
4
4.1
4
3.9
3.9
3.5
1980
3.8
3.8
4.3
3.7
3.9
3.9
3.6
4
4.2
4.1
4.1
3.9
4
4
4.1
1990
4
3.8
4.1
3.8
4.1
4.1
3.8
4
4
4
3.5
3.7
3.9
3.9
3.9
4.2
2000
3.5
3.5
3.7
3.5
3.9
3.9
3.9
3.5
3.6
3.6
3.5
3.8
3.7
3.7
3.4
3.7
3.9
3.5
2010
3.4
3.2
3.5
3.4
3.6
3.8
3.7
3.4
3.3
3.8
3.3
3.8
3.6
3.7
3.4
3.5
3.5
3.5
3.8
2020
3.5
3.7
3.8
3.6
3.5
3.5
3.7
4
3.5
3.9
3.4
3.8
3.7
3.7
3.6
3.4
3.4
3.7
Action
Adventure
Animation
Comedy
Crime
Documentary
Drama
Family
Fantasy
History
Horror
Music
Mystery
Romance
Science Fiction
TV Movie
Thriller
War
Western
Genre
4.6
4.4
4.2
4.0
3.8
3.6
3.4



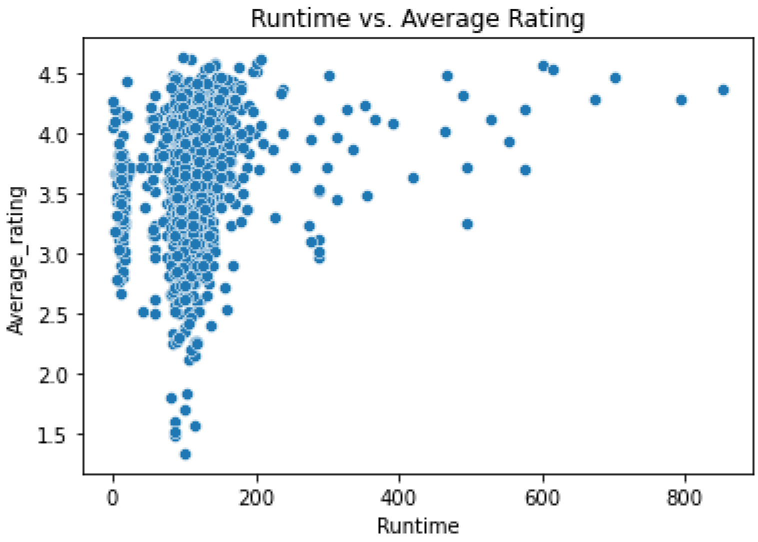
Runtime Analysis

To investigate whether the length of a movie influenced my ratings, I performed the following analysis:

Runtime vs. Rating:

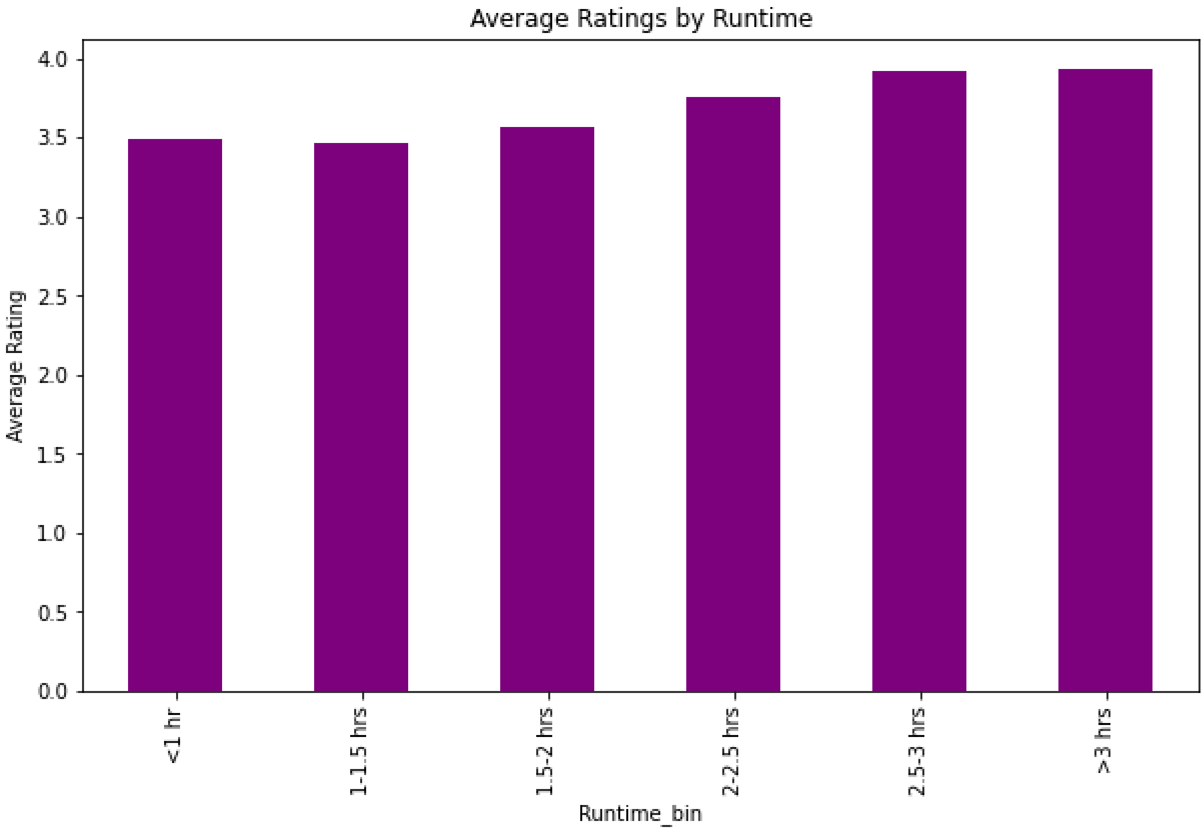
A scatter plot was created to visualize the relationship between movie runtime and average rating. This helped to explore if there was any correlation between how long a movie was and how highly I rated it.

Plot:



Average Ratings by Runtime Categories:

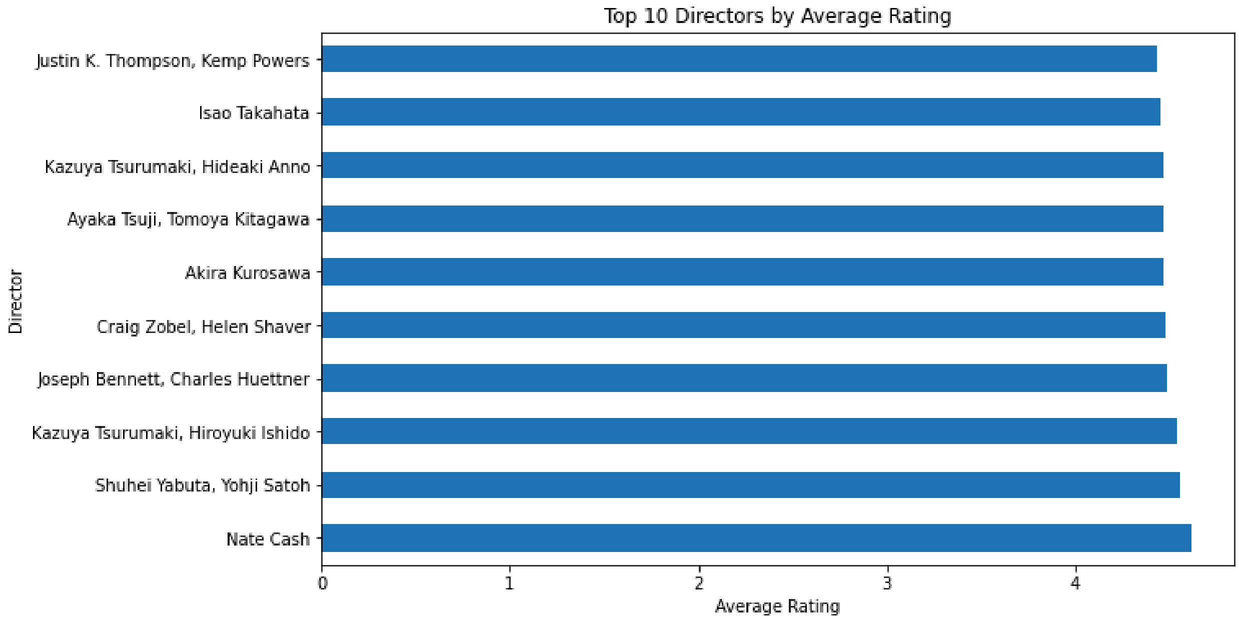
The movies were grouped into different runtime categories (e.g., <1 hr, 1-1.5 hrs, 1.5-2 hrs) and the average rating for each category was calculated. A bar plot was created to show how my ratings varied across different runtime categories.



Here's my favourite 10 directors by average rating:

1. **Nate Cash** (Over the Garden Wall) - 4.62
2. **Shuhei Yabuta, Yohji Satoh** (Vinland Saga) - 4.56
3. **Kazuya Tsurumaki, Hiroyuki Ishido** (End of Evangelion) - 4.54
4. **Joseph Bennett, Charles Huettner** (Scavengers Reign) - 4.49
5. **Craig Zobel, Helen Shaver** (The Penguin) - 4.48
6. **Akira Kurosawa** (Seven Samurai, High and Low, Rashomon) - 4.476667
7. **Ayaka Tsuji, Tomoya Kitagawa** (Sousou no Frieren) - 4.47
8. **Kazuya Tsurumaki, Hideaki Anno** (Evangelion) - 4.47
9. **Isao Takahata** (Grave of the Fireflies) - 4.46
10. **Justin K. Thompson, Kemp Powers** (Spider-Man: Into the Spider-Verse) - 4.44

Plot:



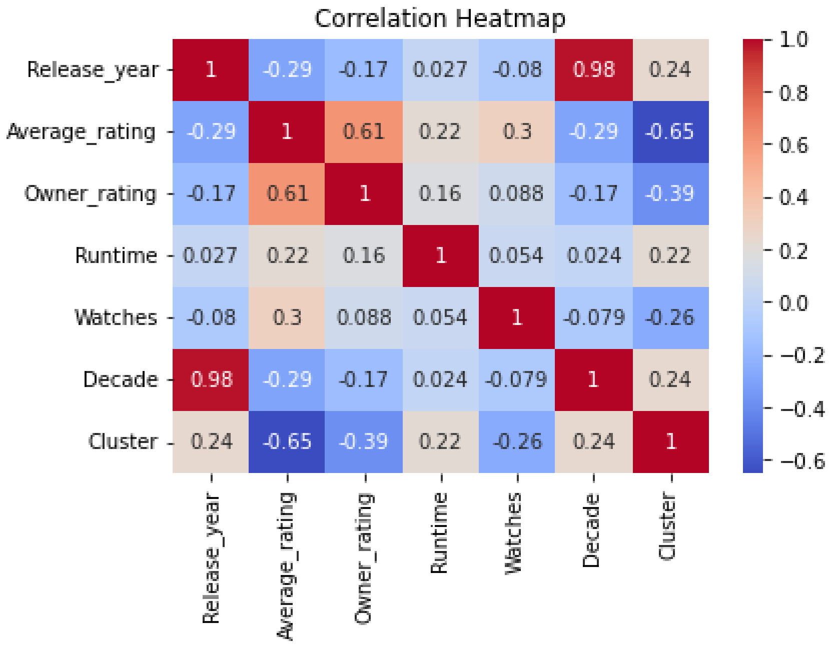
Correlation and Clustering

To gain deeper insights into the relationships between numerical features:

Correlation Matrix:

A heatmap was generated to display the correlation between various numerical features (e.g., Runtime, Average Rating, Release Year). This helped identify which features were most strongly correlated with one another.

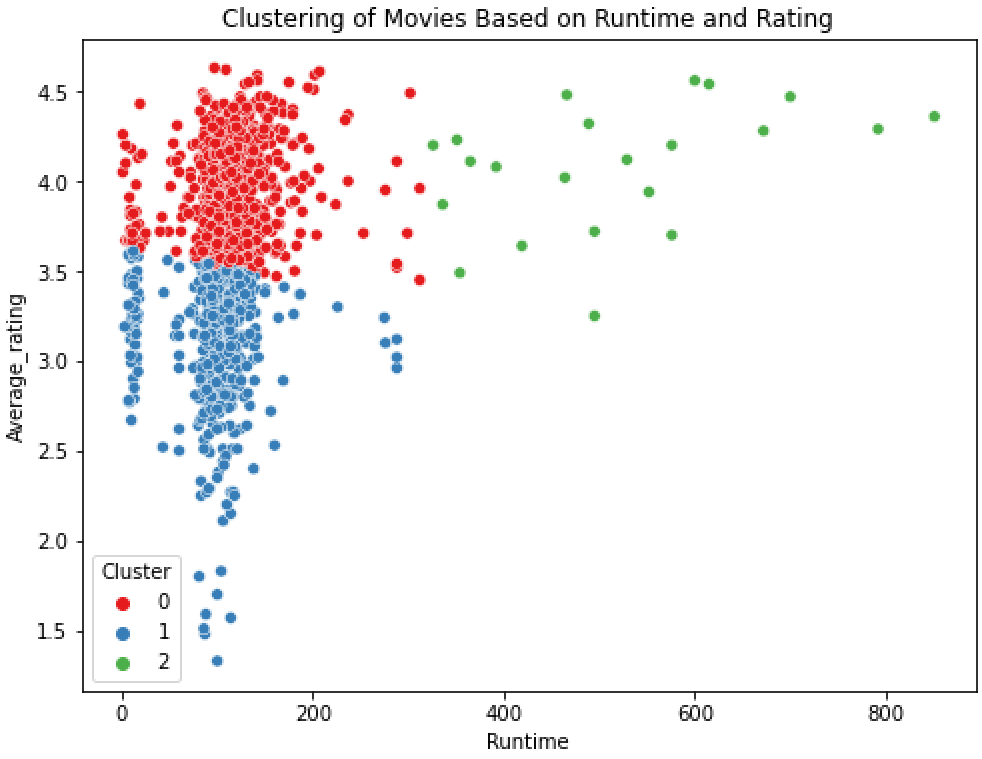
Plot:



Clustering of Movies:

To understand if there were distinct clusters of movies based on Runtime and Average Rating, K-Means clustering was applied. A scatter plot was generated to visualize the clustering results, showing how movies were grouped based on these two features.

Plot:



Statistical Analysis and Hypothesis Testing

Overview

Statistical analysis was performed to derive meaningful insights from the data, examining patterns and relationships between various features, such as movie genres, runtime, and release years. In addition, hypothesis testing was conducted to assess whether certain changes in my ratings were statistically significant.

T-test: Comparing Ratings Before and After 2000

One of the key statistical tests conducted was a **t-test** to compare the average ratings of movies released before and after the year 2000. The purpose of this test was to determine if there was a significant difference in how I rated movies from these two time periods.

Null and Alternative Hypothesis:

- **Null Hypothesis (H_0)**: There is no significant difference in the average ratings of movies released before and after 2000.
- **Alternative Hypothesis (H_a)**: There is a significant difference in the average ratings of movies released before and after 2000.

The following steps were carried out:

1. The ratings for movies released before 2000 were separated from those released in 2000 or later.
2. A two-sample t-test was performed to assess the difference in average ratings for the two groups.

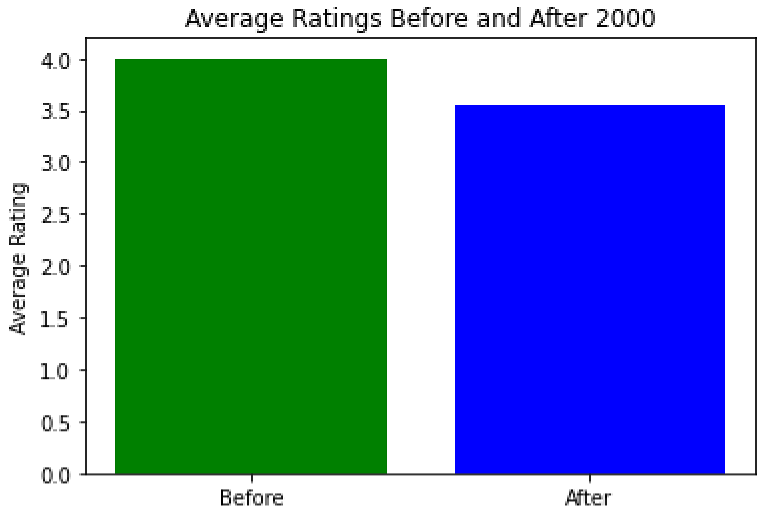
Results:

The calculated t-statistic and p-value were as follows:

- **T-statistic**: 11.241557228363192
- **P-value**: 6.9985769175922644e-28

Since the p-value is **less than 0.05**, the null hypothesis was rejected, Null hypothesis was rejected, indicating a significant difference in ratings before and after 2000.

To provide a clearer visual representation of the differences, a bar plot of the average ratings before and after 2000 was created:



Average Rating Before 2000: 4.000114285714287

Average Rating After 2000: 3.5576446280991765

Cohen's d: Effect Size Calculation

To quantify the size of the difference between ratings before and after 2000, **Cohen's d** was calculated. Cohen's d measures the standardized difference between the two means and is used to determine whether the difference is large enough to be of practical significance.

The formula for Cohen's d is: d = (μ1 - μ2)/σpooled

Where:

- μ1 and μ2 are the means of the two groups (before and after 2000).
- σpooled is the pooled standard deviation of both groups.

The calculated **Cohen's d** was -1.0199157054955752, which provides an interpretation of the effect size:

- **Small effect**: d<0.2
- **Medium effect**: 0.2≤d<0.5
- **Large effect**: d≥0.5

So in terms of Cohen’s d (absolute value), effect size is large. It suggests a huge practical impact.

Correlation Analysis: Relationship Between Runtime and Rating

To explore whether there was a significant relationship between **runtime** and **average rating**, the **Pearson correlation coefficient** was computed. This statistical measure quantifies the linear relationship between two continuous variables.

Results:

The correlation between **Runtime** and **Average Rating** was calculated as 0.21832461311900542. A correlation coefficient close to 1 or -1 would indicate a strong linear relationship, while a coefficient close to 0 suggests little to no linear relationship. So there’s no significant relationship. (The plot can be seen in EDA part)

Regression Analysis

To further explore the relationship between movie features (**Runtime**, **Release Year** and **Decade**) and **Average Rating**, a **linear regression model** was built. This model aimed to predict the average rating based on these features.

The results from the regression model (using **OLS regression**) provided the following coefficient estimates:

- **Runtime**: 0.001574
- **Release Year**: -0.006991
- **Decade**: -0.003636

Based on these coefficients, we see that as release year/decade increases average rating falls. On the contrary, bigger runtime leads to higher average rating.

The **p-values** for the coefficients were also checked to determine the statistical significance of each feature. Features with a p-value less than 0.05 were considered statistically significant.

Here are the p-values of coefficients:

const (Intercept)	2.469133e-34
Runtime	5.683820e-16
Release_year	1.494989e-01
Decade	4.443021e-01

This means that the intercept (this means that even when all predictors are zero, the model's baseline prediction (average rating) is statistically valid) and runtime is highly significant because their p-values are smaller than 0.05.

P-values of release year and decade are bigger than 0.05 shows that they are not statistically significant. For release year, while the coefficient suggests a negative relationship, we cannot confidently assert that this effect is not due to random variation. Additionally, for decade, there is no strong evidence to suggest that the decade a movie was released has a meaningful impact on the average rating after accounting for other factors.

Model Performance:

The regression model's **R-squared** value was 0.46852845644912466, which explains the proportion of variance in the **Average Rating** that can be explained by the features in the model. Since the scale is 0.5-5 we can say that it’s acceptable via having doubts.

This section presented the statistical analysis and hypothesis testing conducted during the project. It provided a detailed examination of the differences in ratings before and after 2000, explored correlations between runtime and ratings, and tested for normality and statistical significance. The results from these tests helped provide a deeper understanding of the factors influencing my movie ratings.

Machine Learning and Clustering

Overview

In this project, machine learning techniques were employed to gain deeper insights into the data, model relationships between features, and explore potential patterns. Specifically, I applied clustering to group movies based on runtime and average rating, as well as built predictive models for average movie ratings. These techniques provided a more nuanced understanding of how various movie attributes might influence ratings.

Model-Based Analysis

To predict ratings and explore feature importance, several machine learning models were applied:

Linear Regression:

A linear regression model was used to predict Average Rating based on various features such as Runtime and Release Year. The model coefficients were examined to determine the impact of each feature on the predicted rating.

Random Forest Regressor:

A Random Forest model was trained to predict Average Rating, and the feature importances were visualized to identify which features had the most influence on the rating predictions..

K-Means Clustering

To explore patterns within the data, I performed **K-Means clustering**, a popular unsupervised machine learning algorithm. This method is used to partition the data into distinct clusters based on the similarity of features—in this case, **Runtime** and **Average Rating**.

Steps:

1. **Data Preprocessing**: The features selected for clustering were **Runtime** and **Average Rating**. To standardize these features, I used the **StandardScaler**, which scales the features to have a mean of 0 and a standard deviation of 1. This step is crucial for K-Means, as it ensures that both features contribute equally to the distance calculations.
2. **Clustering**: Using the scaled data, I applied the **K-Means algorithm** with 3 clusters (chosen based on trial and error and understanding the underlying data structure). The algorithm assigns each data point (movie) to one of the clusters based on the proximity of its features to the cluster centroids.
3. **Cluster Assignment**: Each movie was assigned to a cluster, and the resulting clusters were visualized in a scatter plot, where the color of each point corresponds to its cluster.

Results:

The scatter plot (can be seen in the EDA part) shows the clustering of movies based on their **Runtime** and **Average Rating**. It illustrates how the movies were grouped into three clusters, which likely represent different types of movies:

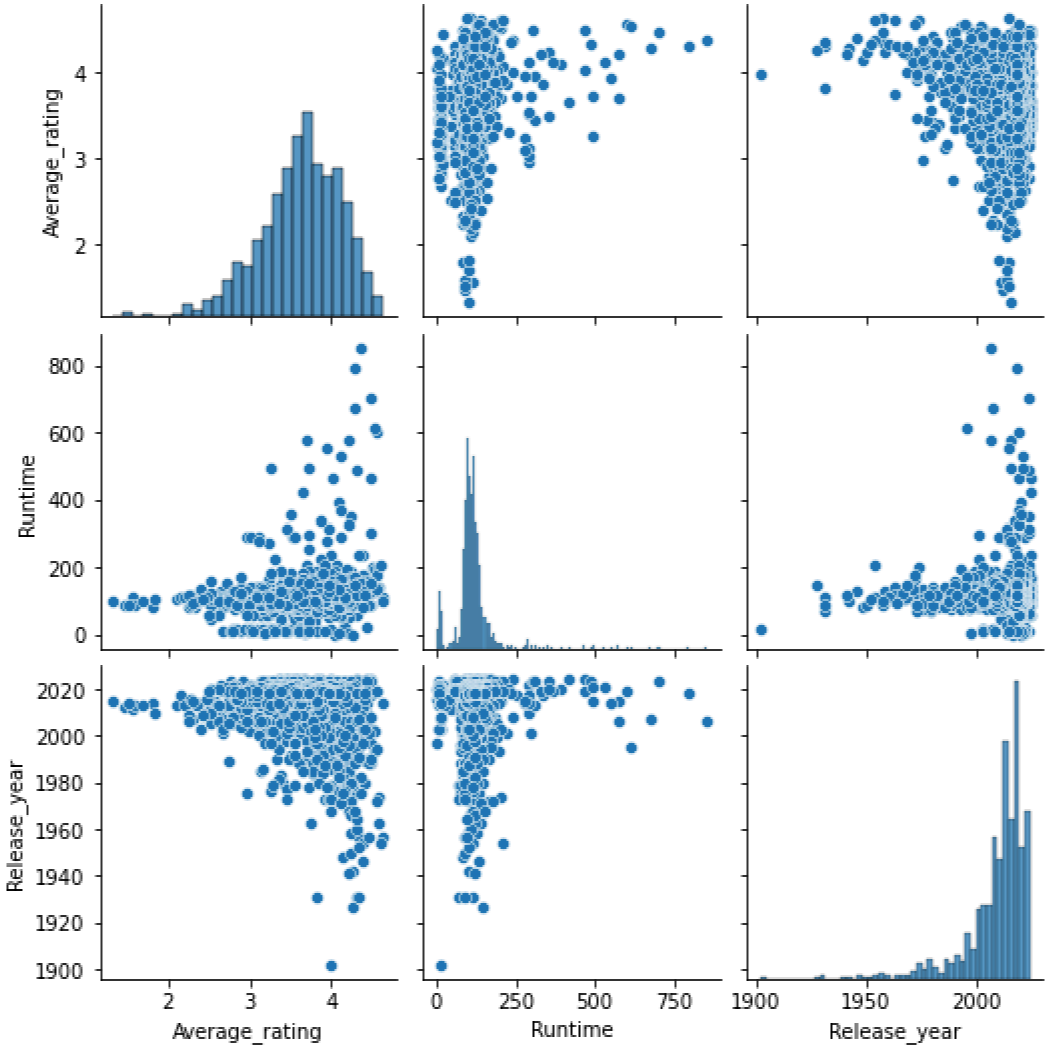
From the plot, it is evident that:

- One cluster tends to represent **shorter films** with **lower ratings**.
- Another cluster contains **medium-length films** with **moderate ratings**.
- The third cluster contains **longer films** with **higher ratings**.

These insights can help understand how runtime may influence ratings or how different genres might align with particular runtime categories.

Pairwise Relationships and Visualizations

To gain additional insights into the relationships between variables, I visualized the pairwise relationships between different features using a **pair plot**. The pair plot displayed correlations between **Average Rating**, **Runtime**, and **Release Year**. The plots allowed me to see if there were any clear relationships or patterns that could inform future analyses or feature selections for predictive modeling:



Predictive Modeling

In addition to clustering, I built **predictive models** to forecast the **Average Rating** based on other features, such as **Runtime**, **Release Year**, and **Decade**. This modeling process involved using both **Linear Regression** and **Random Forest Regressor**, which are common approaches for regression tasks.

Linear Regression

Linear regression was used to model the relationship between movie features and ratings. The **Ordinary Least Squares (OLS)** method was employed to fit the model.

Model Setup:

- **Independent Variables**: Runtime, Release Year, Decade
- **Dependent Variable**: Average Rating

The regression coefficients indicated the relationship between each independent variable and the dependent variable. For instance:

- **Runtime**: A positive or negative coefficient would suggest how runtime influences the rating.
- **Release Year**: This could reflect a trend in ratings over time.

Model Performance:

The **R-squared** value from the linear regression model was 0.149, indicating how much of the variance in ratings was explained by the model. Which is relatively low.

Random Forest Regressor

A **Random Forest Regressor** was also trained on the data to assess how well the model could predict movie ratings. Random forests are an ensemble learning method that builds multiple decision trees and aggregates their predictions.

Model Setup:

- **Independent Variables**: Runtime, Release Year, Decade, and other relevant features

- **Dependent Variable**: Average Rating

Model Performance:

The **feature importance** chart revealed which features had the greatest influence on predicting movie ratings. Features such as **Runtime** and **Release Year** had high importance, while other features, like **Decade**, contributed less to the model’s performance.

Results:

The model’s **mean squared error (MSE)** and **RMSE** were calculated to evaluate performance:

- **MSE**: 0.06307782684055058
- **RMSE**: 0.2511529948866837

These values provide an understanding of how well the model fits the data. The lower the RMSE, the better the model’s predictions. On the contrary to the linear regression model in terms of RMSE, random forest regressor does a better job, almost half of the linear regression model’s RMSE.

Conclusion of Machine Learning Analysis

Both clustering and predictive modeling provided valuable insights into the relationships between movie attributes and ratings:

- **Clustering** helped to identify distinct groups of movies based on runtime and rating, offering insights into how these features interact.
- **Predictive models**, especially the Random Forest, showed how well certain features could predict ratings, with **Runtime** and **Release Year** emerging as the most important predictors.

These findings open up opportunities for deeper analysis, such as exploring more complex relationships or introducing additional features (e.g., genre, director) into the models to improve prediction accuracy.

This section provides an overview of the clustering and machine learning models implemented in the project, offering insights into how runtime and release year influence ratings and identifying important features for predictive modeling.

Findings, Insights, Limitations and Future Work

Key Findings

Based on the data analysis and machine learning models implemented in this project, several interesting findings and insights emerged:

1. **Genres and Ratings**:
   - Different movie genres showed varying average ratings. **Action** and **Horror** genres, for example, exhibited relatively low ratings, while genres like **Science Fiction** and **Thriller** had higher ratings on average.
   - The variation in average ratings by genre suggests that I may have differing expectations or preferences depending on the genre, and this could be influenced by factors such as personal taste, cultural relevance, or the prominence of particular films within each genre.
   - The analysis may be influenced by sample bias, as genres with higher sample sizes (e.g., favorite genres) include a broader range of films, while less-favored genres likely contain only highly recommended titles. This discrepancy may artificially elevate the average ratings of underrepresented genres and lower those of overrepresented ones.
2. **Decades and Movie Ratings**:
   - Movies from the **1980s** and **1990s** tended to have higher average ratings, while movies from the **2010s** and **2020s** exhibited a decline in average ratings.
   - This trend could be indicative of shifting movie-making standards, evolving preferences, or changes in critical reception over the years.

It might also reflect broader societal and cultural changes influencing movie production and consumption.

3. **Runtime and Rating Relationship**:
   - A positive relationship between **movie runtime** and **average rating** was observed, particularly in longer films. This suggests that longer films tend to have higher ratings, potentially due to the scope of their storylines or the depth of character development.
   - However, there were outliers, with some short films receiving high ratings, indicating that runtime alone is not the sole determinant of a movie's quality.
4. **Cluster Insights**:
   - K-Means clustering revealed three distinct clusters based on **runtime** and **average rating**:
     - **Cluster 1**: Short films with lower ratings
     - **Cluster 2**: Medium-length films with moderate ratings
     - **Cluster 3**: Longer films with higher ratings
   - This clustering suggests that movie length may be correlated with the perceived quality of a film.
5. **Trends Over Time**:
   - There is a visible trend in the data showing how **movie ratings** have evolved over time. Ratings seemed to peak in earlier decades, with a gradual decline in more recent years.
   - This trend might point to changes in movie production quality, audience expectations, or the influence of digital platforms, which have altered how movies are distributed and consumed.
6. **Impact of Release Year on Ratings**:
   - The **T-test** conducted on the ratings before and after the year **2000** suggested a statistically significant difference in ratings. Movies released after 2000 generally had lower ratings than those released before.
   - This finding could indicate shifts in the film industry, where earlier films may have had more critical acclaim or larger budgets, while

recent movies may be facing a more competitive, crowded market or differing tastes from modern audiences.

7. **Director Influence on Ratings**:
   - Certain **directors** consistently produced films with higher average ratings.
   - This reinforces the idea that directorial style, reputation, and previous successes can significantly influence a film's reception.

Insights for Future Analysis

1. **Exploring Genre-Related Factors**:
   - Further analysis could be done to examine how specific **sub-genres** within broad categories like **Action** or **Horror** influence ratings. For example, are **psychological thrillers** rated higher than **slasher films**? Understanding these nuances can provide deeper insights into genre preferences.
2. **Incorporating Additional Features**:
   - Future models could incorporate additional features such as **director**, **actor**, and **budget** to improve predictive accuracy. It would be interesting to see if these factors play a significant role in rating predictions.
   - Additionally, incorporating **audience reviews** or **critic ratings** could give a more comprehensive view of a movie's reception.
3. **Sentiment Analysis on Reviews**:
   - To supplement the numerical data, **sentiment analysis** on user reviews could provide valuable insights. Understanding whether a movie’s positive or negative reception aligns with its numerical rating would help validate or challenge some of the conclusions derived from ratings alone.
4. **Longitudinal Analysis**:
   - A longitudinal analysis tracking **rating trends** over several decades would provide a clearer picture of how tastes and industry standards

have evolved. Comparing these trends with societal or historical events might offer explanations for some of the observed patterns.

5. **Impact of Digital Platforms**:
    - The rise of digital platforms like **Netflix**, **Amazon Prime**, and **Disney+** has transformed the way movies are distributed. Future research could investigate whether movies released through these platforms tend to have different ratings compared to those released in theaters.

Conclusion

In summary, the analysis and machine learning models applied in this project have provided valuable insights into the relationship between various movie features (such as genre, runtime, and release year) and their ratings. The findings suggest that factors like movie length, director, and genre can significantly impact the way movies are rated. However, there are still many avenues for further research, including exploring the influence of additional features and expanding the analysis to incorporate more complex datasets.

These insights lay the foundation for future exploration into the factors that contribute to movie ratings and help us better understand the evolution of movie preferences over time.

References

Letterboxd. (n.d.). *Letterboxd: Movie Ratings and Reviews*. Retrieved from https://letterboxd.com/