

# CS210 DATA SCIENCE COURSE PROJECT REPORT

## 1. Introduction: My Cinematic Diary Through Letterboxd Data

In this data science project, I explore my cinematic experience over the past two years using data from my Letterboxd diary — a social media site where users log, rate, and review the movies they watch. My dataset covers essential details like watch dates, film names, release years, personal ratings, general ratings, directors, and genres.

The goal is to use this data to uncover patterns and correlations that shed light on the evolution of my movie preferences. By using methods like data visualization, regression models, and correlation calculations my aim is not just to uncover patterns in my changing tastes but also to show how data science can enhance our understanding of individualized cinematic experiences.

The details of the dataset and findings can be found in the GitHub repository.

## 2. Data Collection: Scraping Insights with Selenium

To compile a comprehensive dataset for analysis, I employed Selenium, an automation tool, along with the Chrome web driver. The data extraction process initiated from the diary page, where, I navigated to each film's individual page to retrieve seven key attributes:

Watched Date: Date I Watched the Movie

Film Name: Name of The Movie







Film Year: Movie Release Date

My Rating: My Rating for the Movie on a Scale of 0 to 5


General Rating: Average of the Ratings Given by All Viewers

Director: The Director of the Movie


Genres: Genres of the Movie

WATCHED			DIARY	REVIEWS	RATING	DIARY YEAR	DECADE	GENRE	SERVICE	Sort by	WATCHED DATE	
MONTH	DAY	FILM	RELEASED	RATING	LIKE	REWATCH	REVIEW	EDIT				
FEB 2023	28	 <b>Soul</b>	2020	★★★★☆	♥							✎ ...
	26	 <b>Mary and Max</b>	2009	★★★★★	♥							✎ ...
	26	 <b>Amadeus</b>	1984	★★★★☆	♥							✎ ...
	21	 <b>WHAT DID JACK DO?</b>	2017	★★★★☆	♥							✎ ...
	20	 <b>Lady Bird</b>	2017	★★★★☆	♥							✎ ...
	04	 <b>The Fabelmans</b>	2022	★★★★☆	♥							✎ ...





1.7M 189K 562K

WHERE TO WATCH  Trailer

Google Play ... TR RENT BUY

Google Play ... US BUY

Vudu US BUY

## Soul 2020

Directed by Pete Docter

IS ALL THIS LIVING REALLY WORTH DYING FOR?

Joe Gardner is a middle school teacher with a love for jazz music. After a successful audition at the Half Note Club, he suddenly gets into an accident that separates his soul from his body and is transported to the You Seminar, a center in which souls develop and gain passions before being transported to a newborn child. Joe must enlist help from the other souls-in-training, like 22, a soul who has spent eons in the You Seminar, in order to get back to Earth.

REMOVE ADS x

CAST CREW DETAILS GENRES RELEASES

GENRES Family Drama Animation Comedy Fantasy

THEMES Humanity And The World Around Us Holiday Joy And Heartwarming Chr... Adorable Animals And Heartwarmi...

Logged Like Watchlist

Rated

★★★★☆

Show your activity

Add a review...

Log again...

Add to lists...

Go [PATRON](#) to change poster

Share

RATINGS 6.2K FANS

3.9

★★★★☆

The iterative process involved moving back and forth between the film's specific page and the diary section, ensuring a systematic collection of data for every film in my cinematic diary. Subsequently, leveraging the capabilities of the Pandas library, I transformed this raw data into a structured DataFrame. The resulting DataFrame encapsulates the nuances of each film encounter, serving as the foundation for the subsequent stages of analysis.

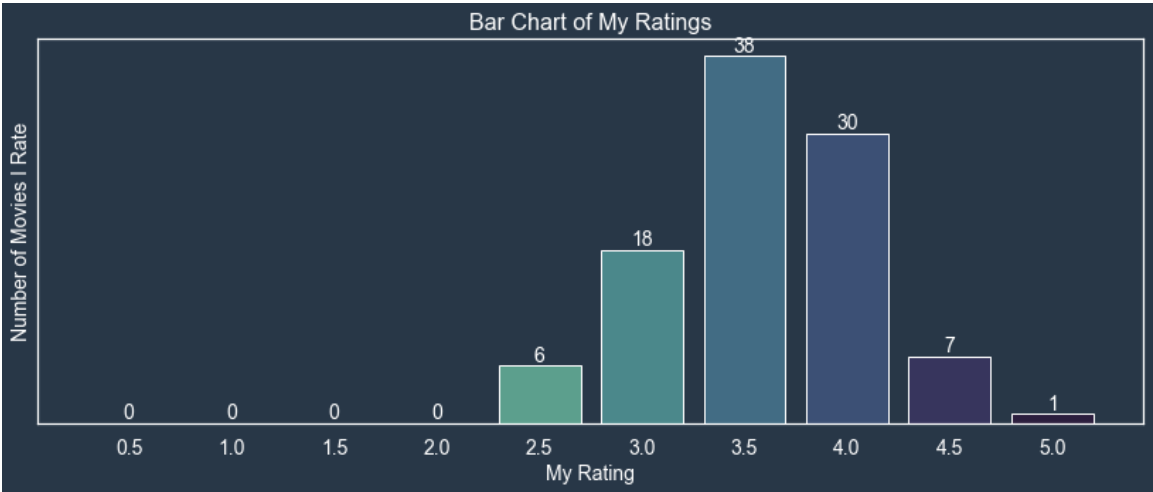
	Watched Date	Film Name	Film Year	My Rating	General Rating	Director	Genres	Month
0	2023-12-25	In Bruges	2008		4.1	Martin McDonagh	['Drama', 'Comedy', 'Crime']	2023-12
1	2023-12-25	The Nightmare Before Christmas	1993	3.5	3.9	Henry Selick	['Family', 'Animation', 'Fantasy']	2023-12
2	2023-12-16	The Matrix	1999	5.0	4.2	Lilly Wachowski	['Science Fiction', 'Action']	2023-12
3	2023-12-11	The Hunger Games	2012		3.6	Gary Ross	['Fantasy', 'Science Fiction', 'Adventure']	2023-12
4	2023-12-10	The Hunger Games: The Ballad of Songbirds & Sn...	2023	3.5	3.6	Frands Lawrence	['Science Fiction', 'Drama', 'Action']	2023-12

The final step involved storing this DataFrame in CSV format, providing a tangible and accessible record of my movie-watching experiences. The structured dataset now stands ready for exploration, allowing for analysis and correlations to be uncovered in the upcoming phases of this data science project.

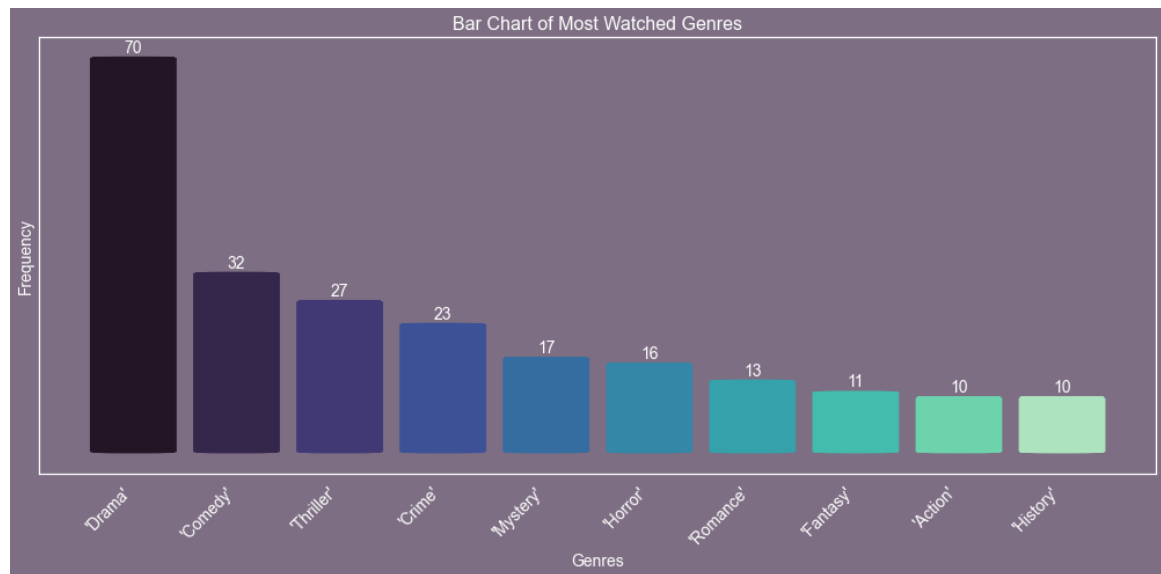
### 3. Exploratory Data Analysis: Understanding Movie Preferences

Before diving into testing hypotheses, it'd be helpful to take a closer look at the data. Here are some simple stats and charts that give a good initial sense:

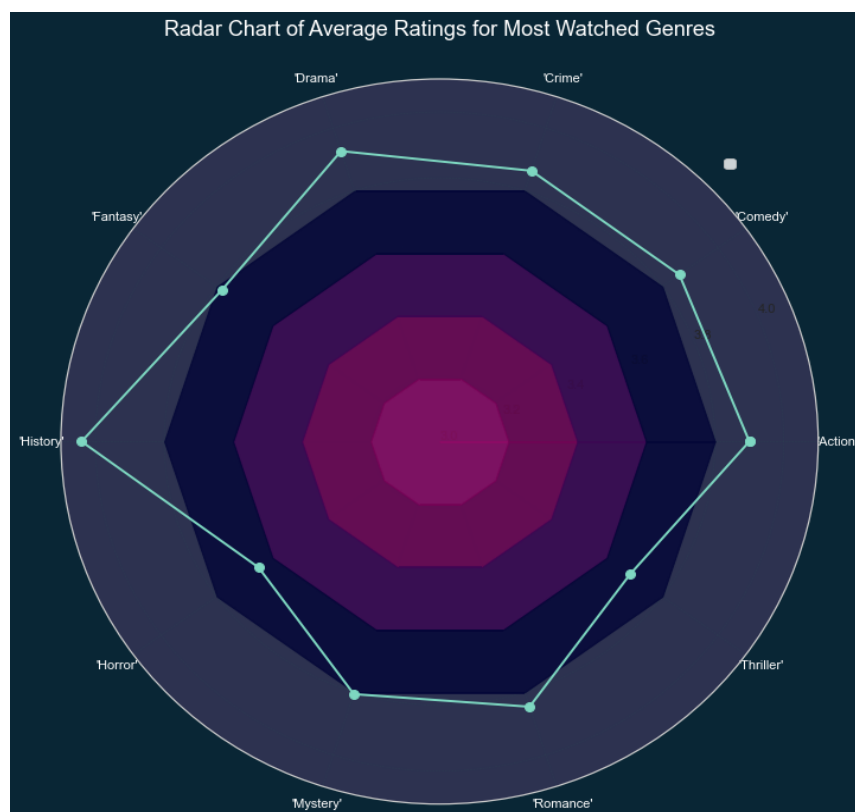
- a. **Movie Ratings Overview (Bar Chart):** Bar chart showing the frequency of my ratings to movies for each rating, on a scale from 0 to 5.



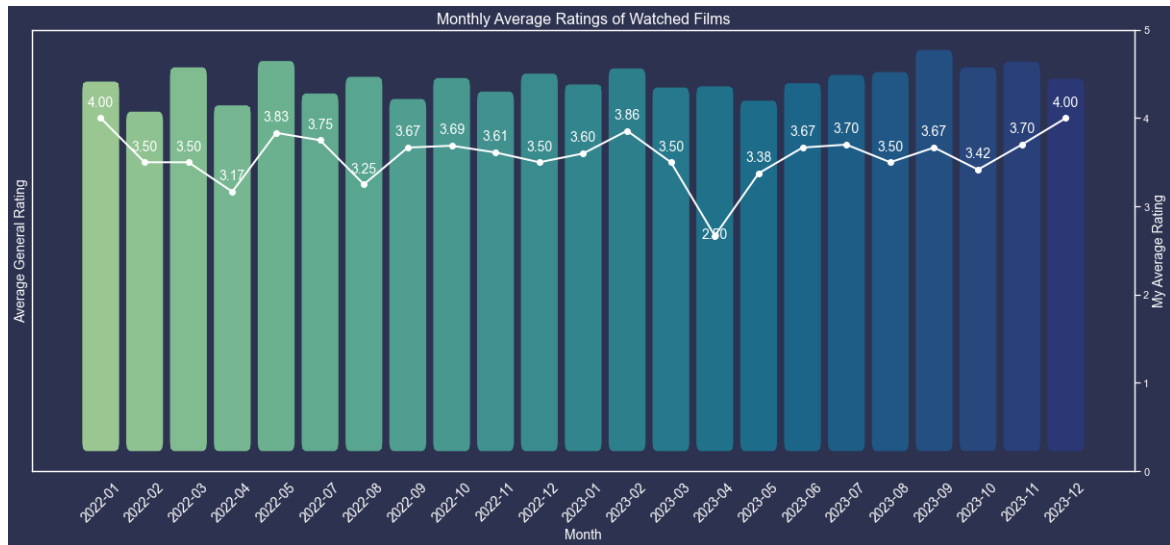
- b. **Most Watched Movie Genres (Bar Chart):** A chart showing which movie genres I watch the most. This helps pinpoint my preferred genres easily.



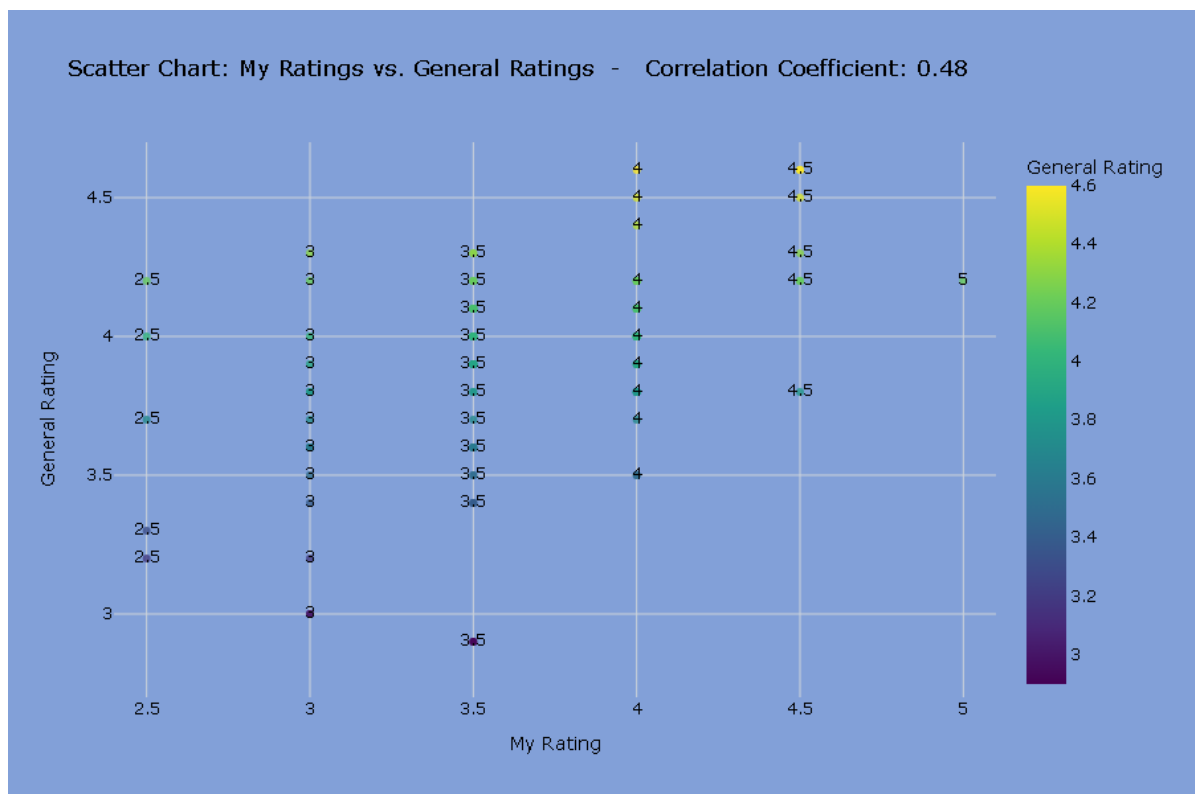
- c. **Average Ratings for Each Genre (Radar Chart):** To find my favourite genre, I showcase the average ratings for each genre. This gives a straightforward view of which genres consistently get higher ratings from me.



- d. **Monthly Rating Trends (Line and Histogram Chart):** Using a line chart, I show how my movie ratings and overall ratings change monthly. This can reveal any patterns in my movie-watching habits over time.



- e. **Scatter Chart: My Ratings vs. General Ratings (Scatter Chart):** Scatter chart showing my rating vs the general average rating for each movie. This chart is useful for showing the compatibility of movie tastes between me and other users on Letterboxd.



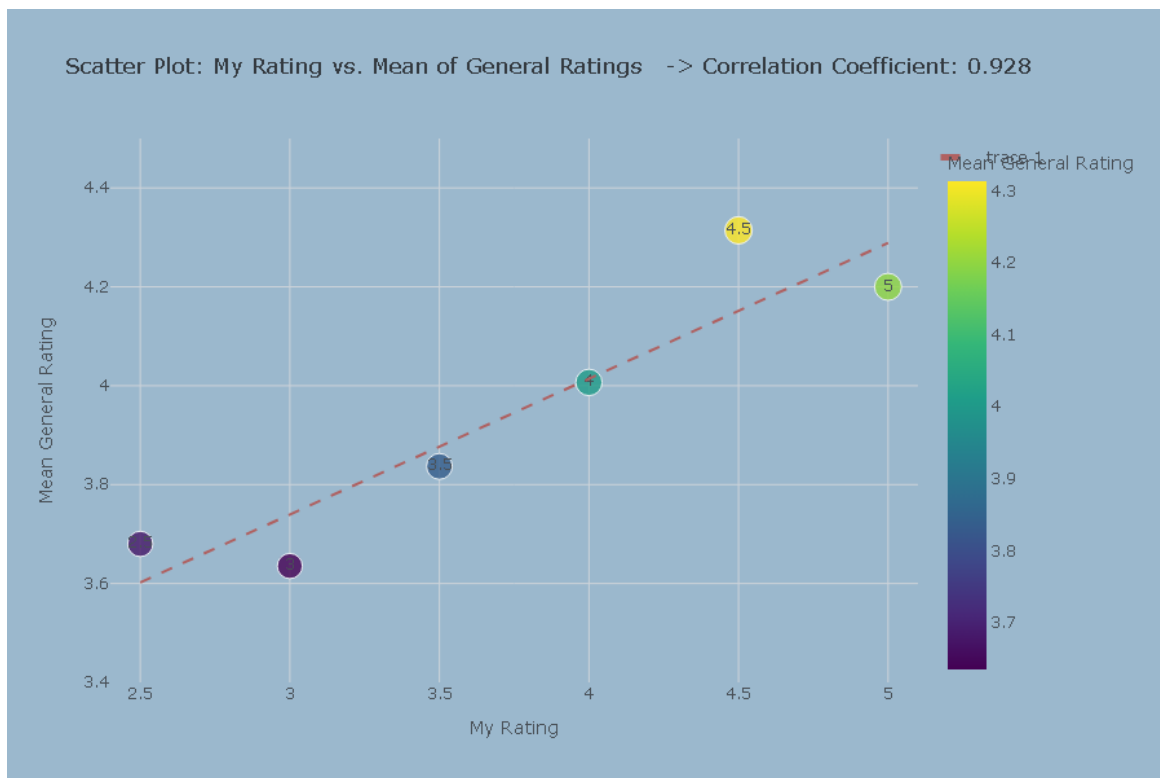
#### 4. Hypothesis Testing: Exploring Changes in Film Preferences

Over the past two years, I've been diving into the world of films, wondering how my taste and critical judgment might have changed. People often think that as you learn more about a specific field, you become pickier and harder to impress. This idea is backed up by a study called "The Impact of Music on Human Development and Well-Being," which shows that experts in a certain field tend to develop more critical views in that area.

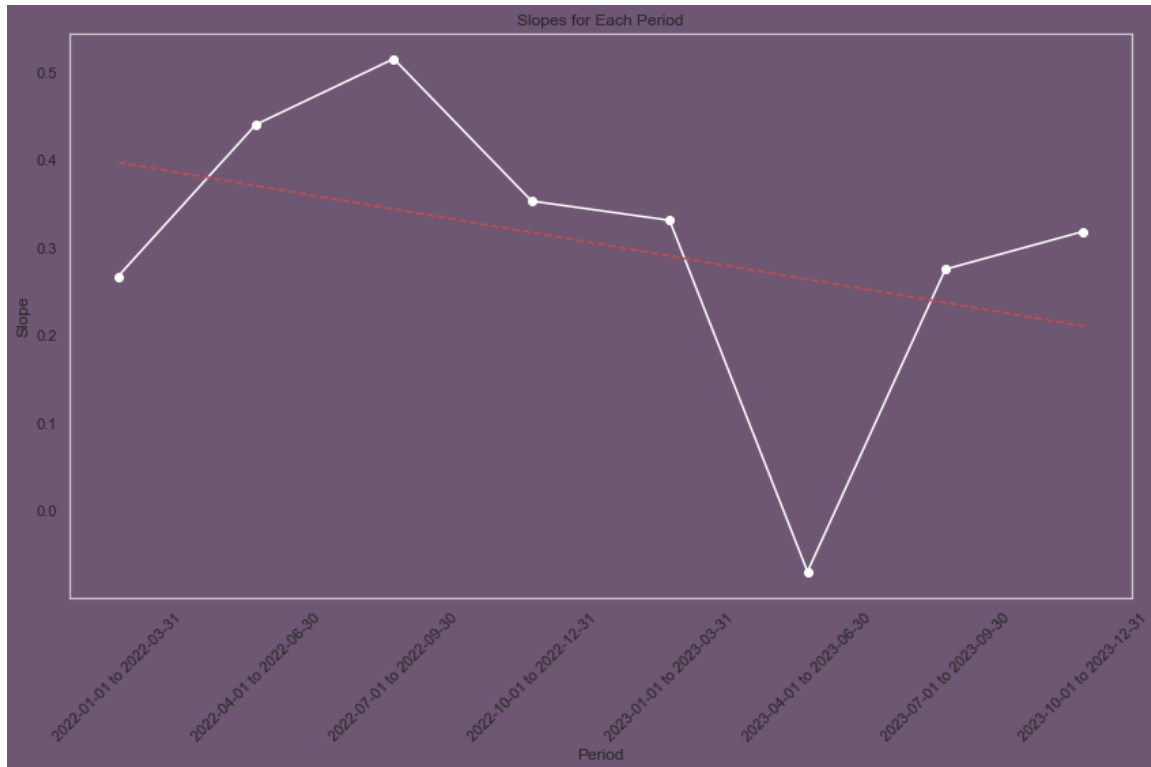
So, I decided to figure out how my liking threshold for movies has changed using a simple calculation. My hypothesis is that there hasn't been a big shift in how I appreciate films.

Since I watch both good and not-so-good movies, just looking at how my ratings changed over time wouldn't make much sense. Therefore, the most reasonable idea was, to do the calculation by comparing it with the average rating people gave to the movies. For this, I averaged the overall scores of those films for each rating value I could give to the films. For example, I calculated the average of the general ratings of the 18 films that I gave 3 points. In this way, I calculated the average rating of people gave, for each rating value.

The result of the correlation I found is 0.928. So that it is confirmed that the rating I gave to the films and the value given by other users were in a certain consistency.



The trendline has a slope of 0.27. I divided the data frame into 8 periods (3 months) and made the same calculation for every period, observed the slopes. If the slope is greater than 0.27, it means that I thought more critical for that period. If it is smaller than 0.27 it means that I approached positively.

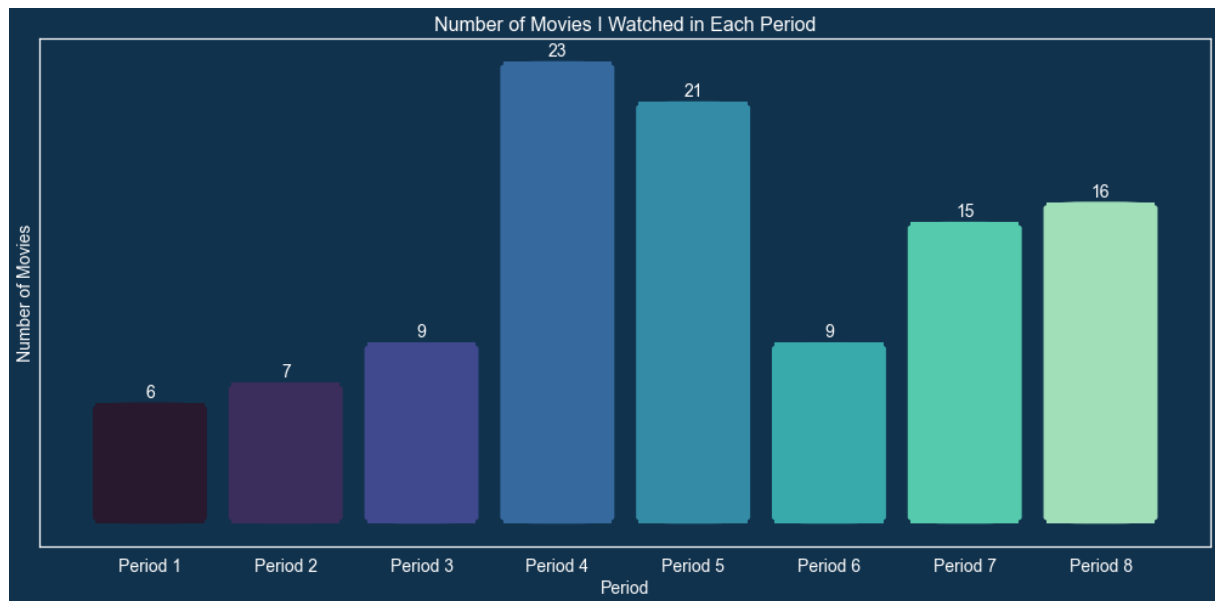


As can be seen in the chart above, the trendline of slopes for each period has a decreasing trend. It shows that my liking threshold did not decrease.

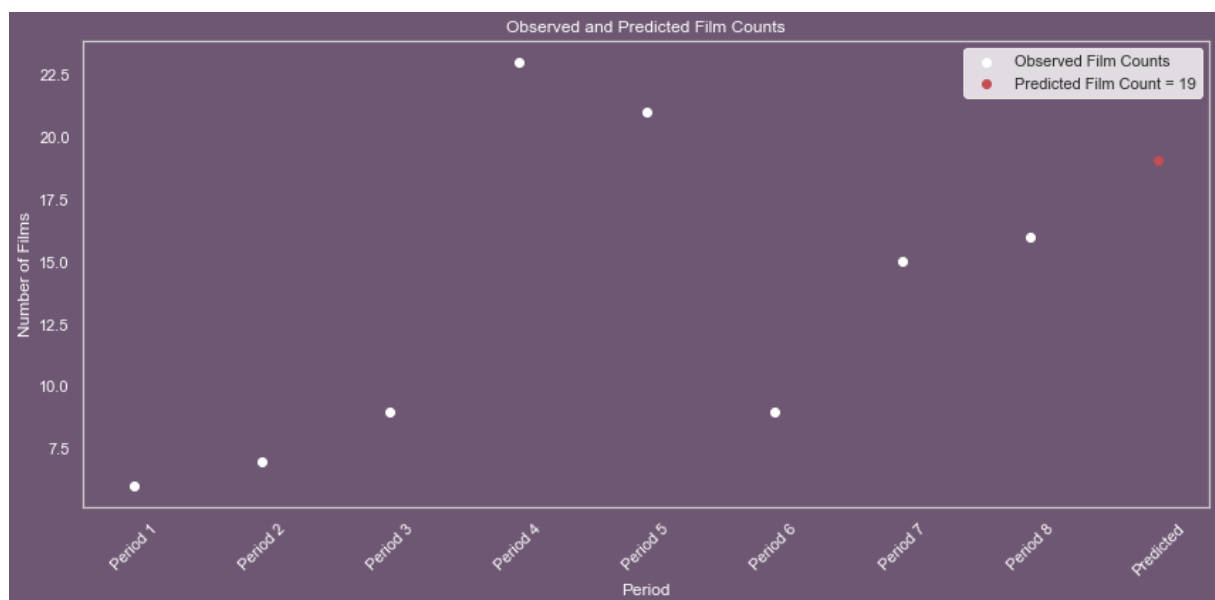
As a conclusion, after 2 years, based on the general ratings, I can say that my liking threshold still has not decreased.

## 5. Linear Regression

After examining my cinema data, I was curious about how my movie-watching habits evolved. I broke down my two-year dataset into three-month intervals, showing the movie count for each period with a bar chart.



The next step involved using a linear regression model to predict the number of movies I might watch in the upcoming period. The computer learns from past data to find a line that fits the pattern of movie-watching over different periods. This line helps predict how many movies I might watch in the next three months. By looking at trends, the model estimates future movie-watching behavior in a simple, numerical way.



As a result, in this case, it suggests I'll likely watch around 19 movies in the upcoming period.



## 6. Conclusion

By using data science, I could see how my film preferences changed over two years. From creating charts to predicting future movie counts with linear regression, this project showed how data helps understand personal cinema experiences. Visualizing with EDA showed me many details that I did not expect about my taste and experience in cinema. In summary, this project highlights how using data can give insights into personal interests, making it easier to understand habits and preferences in the world of cinema.