ERDEM AKAY 30598

PROJECT REPORT

Motivation

This project stems from a personal curiosity to delve into the patterns and insights within my movie-watching experiences on Letterboxd. By scrutinizing factors such as director popularity, runtime, and the correlation between my ratings and those of other Letterboxd users, I aim to gain a deeper understanding of my cinematic preferences. The exploration of these facets not only provides valuable insights into my viewing habits but also serves as a captivating journey into the intricate interplay between personal tastes and broader audience sentiments.

Data Source

I gathered data from my Letterboxd profile, extracting information about the movies I've watched, my ratings, and additional details. By web scraping my Letterboxd film list, I created a dataset to analyze and gain insights into my movie-watching experiences and preferences.

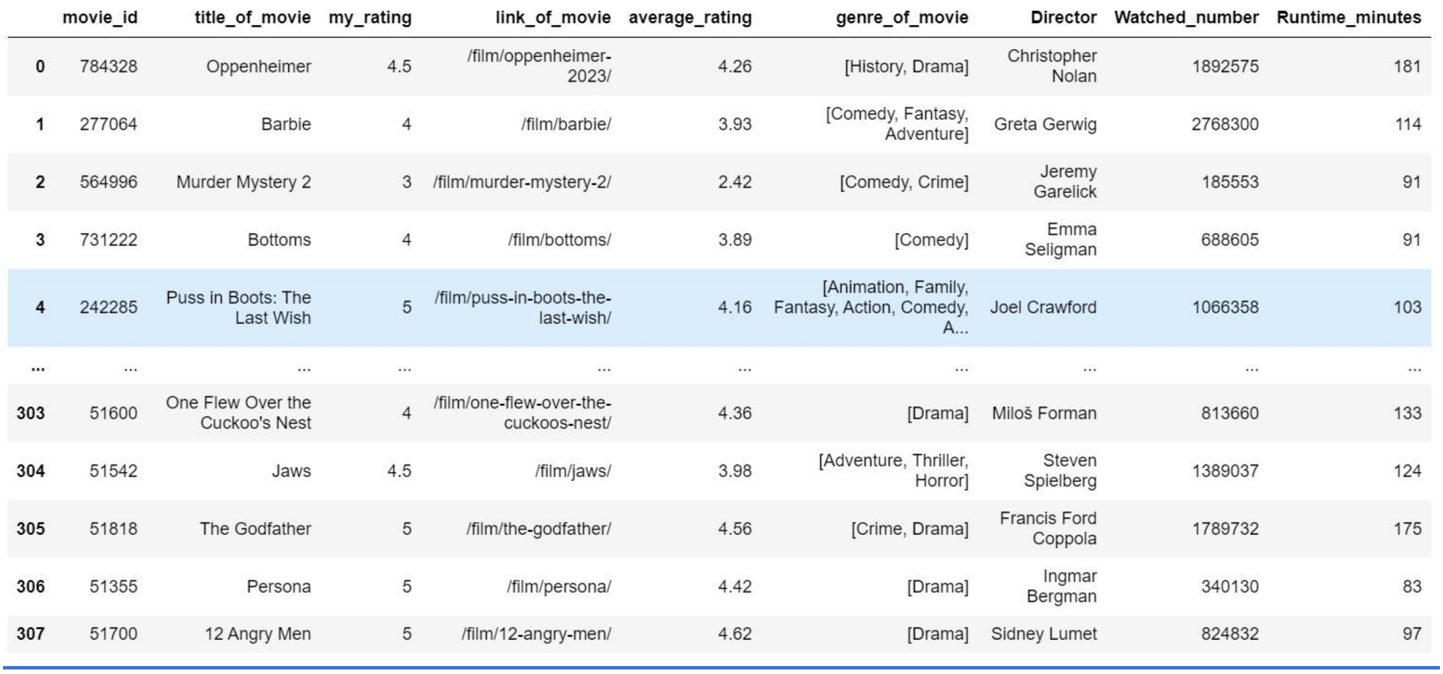
Data Analysis

- **Data Collection:**
    - Film data was collected from the Letterboxd platform, capturing information about movies watched, personal ratings, directors, and runtime.
- **Data Cleaning:**

    - Data cleaning involved handling missing values, converting data types, and ensuring the dataset's integrity.

- **Exploratory Data Analysis (EDA):**
    - Descriptive statistics, such as mean and count, were calculated to summarize the main features of the dataset.
    - Visualizations, including scatter plots, bar charts, and heatmaps, were created to explore relationships and patterns in the data.
- **Statistical Analysis :**
    - Statistical tests, such as the Pearson correlation coefficient, were used to quantify relationships between variables, like personal ratings and popular ratings.
- **Regression Analysis:**
    - Linear regression models were applied to understand the potential correlation between runtime and personal ratings.
- **Director Analysis:**
    - Directors with a significant number of watched movies were selected for further analysis. Ratings were aggregated to compare personal ratings with average ratings by other users.
- **Visualization:**
    - Various types of visualizations, including scatter plots, dot charts, and heatmaps, were utilized to present findings in a visually appealing manner.
- **Hypothesis Testing:**
    - Hypothesis tests were conducted to assess the significance of correlations and relationships within the data.
- **Machine Learning:**
    - A linear regression model was implemented to predict personal ratings based on movie runtime.

Findings



	movie_id	title_of_movie	my_rating	link_of_movie	average_rating	genre_of_movie	Director	Watched_number	Runtime_minutes
0	784328	Oppenheimer	4.5	/film/oppenheimer-2023/	4.26	[History, Drama]	Christopher Nolan	1892575	181
1	277064	Barbie	4	/film/barbie/	3.93	[Comedy, Fantasy, Adventure]	Greta Gerwig	2768300	114
2	564996	Murder Mystery 2	3	/film/murder-mystery-2/	2.42	[Comedy, Crime]	Jeremy Garelick	185553	91
3	731222	Bottoms	4	/film/bottoms/	3.89	[Comedy]	Emma Seligman	688605	91
4	242285	Puss in Boots: The Last Wish	5	/film/puss-in-boots-the-last-wish/	4.16	[Animation, Family, Fantasy, Action, Comedy, A...	Joel Crawford	1066358	103
...	...	...	...	...	...	...	...	...	...
303	51600	One Flew Over the Cuckoo's Nest	4	/film/one-flew-over-the-cuckoos-nest/	4.36	[Drama]	Miloš Forman	813660	133
304	51542	Jaws	4.5	/film/jaws/	3.98	[Adventure, Thriller, Horror]	Steven Spielberg	1389037	124
305	51818	The Godfather	5	/film/the-godfather/	4.56	[Crime, Drama]	Francis Ford Coppola	1789732	175
306	51355	Persona	5	/film/persona/	4.42	[Drama]	Ingmar Bergman	340130	83
307	51700	12 Angry Men	5	/film/12-angry-men/	4.62	[Drama]	Sidney Lumet	824832	97

This is the dataset I created with webscraping:

- **Movie Information:**

  ID of the movie

  Title of the movie

  Link to the movie

- **Ratings:**

  Your personal rating for each movie

  Average rating of the movie (from Letterboxd users)

- **Movie Details:**

  Genre of the movie

  Director of the movie

  Duration of the movie

- **Popularity:**

  Number of users who have watched the movie

RESULTS of HYPOTHESES:

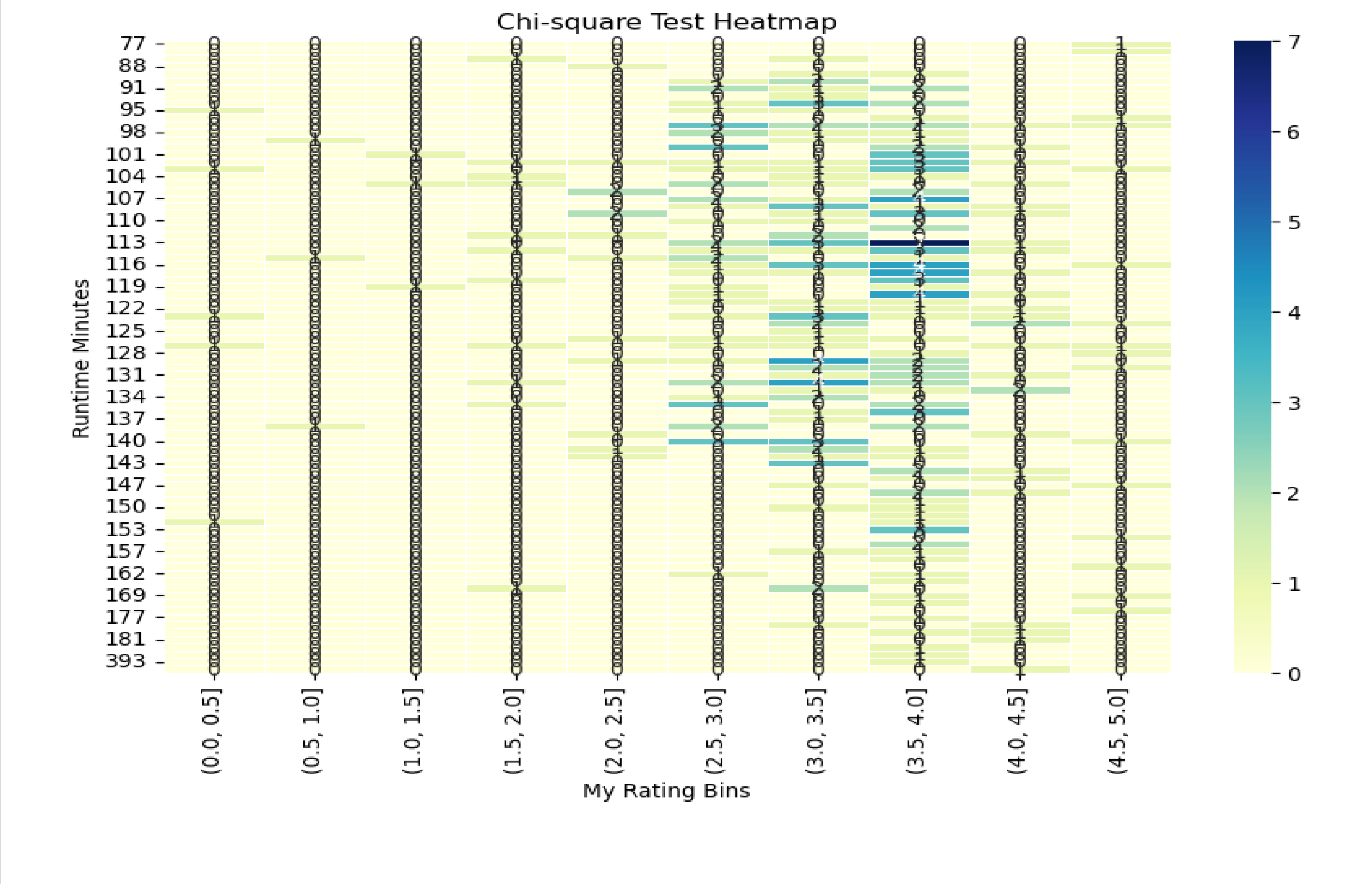
Movie Duration and My Rating

Null Hypothesis:

- There is no association between movie duration and my rating categories.

Alternative Hypothesis:

- There is evidence that movie duration and my rating categories are associated.



Conclusion:

- **Fail to reject the null hypothesis**: There is no evidence that movie duration and my rating are dependent.
- This suggests that, based on the analysis, there is no statistically significant association between movie duration and your ratings.

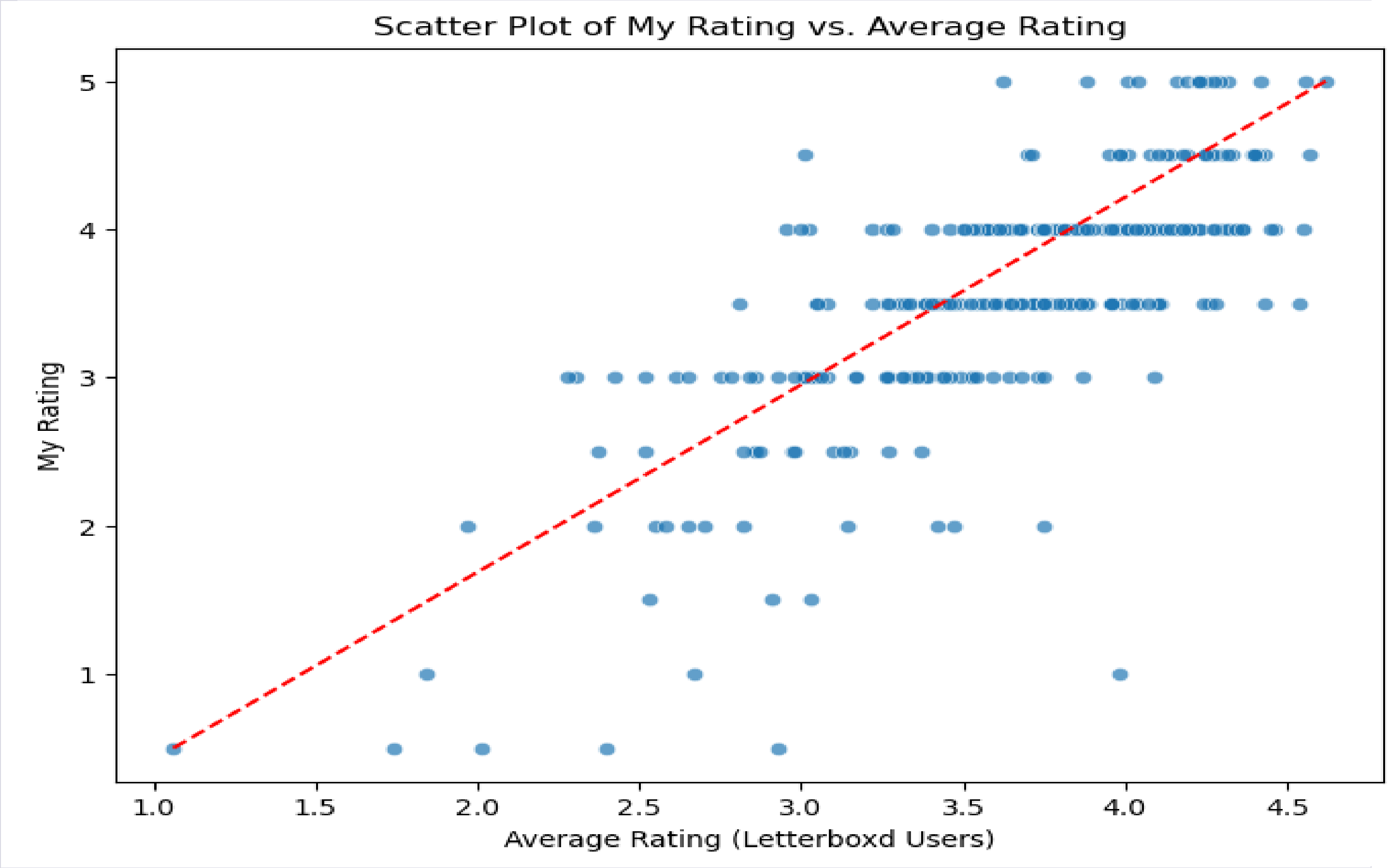
My Ratings and Average Ratings

Null Hypothesis:

- There is no correlation between my ratings and the average ratings on Letterboxd.

Alternative Hypothesis:

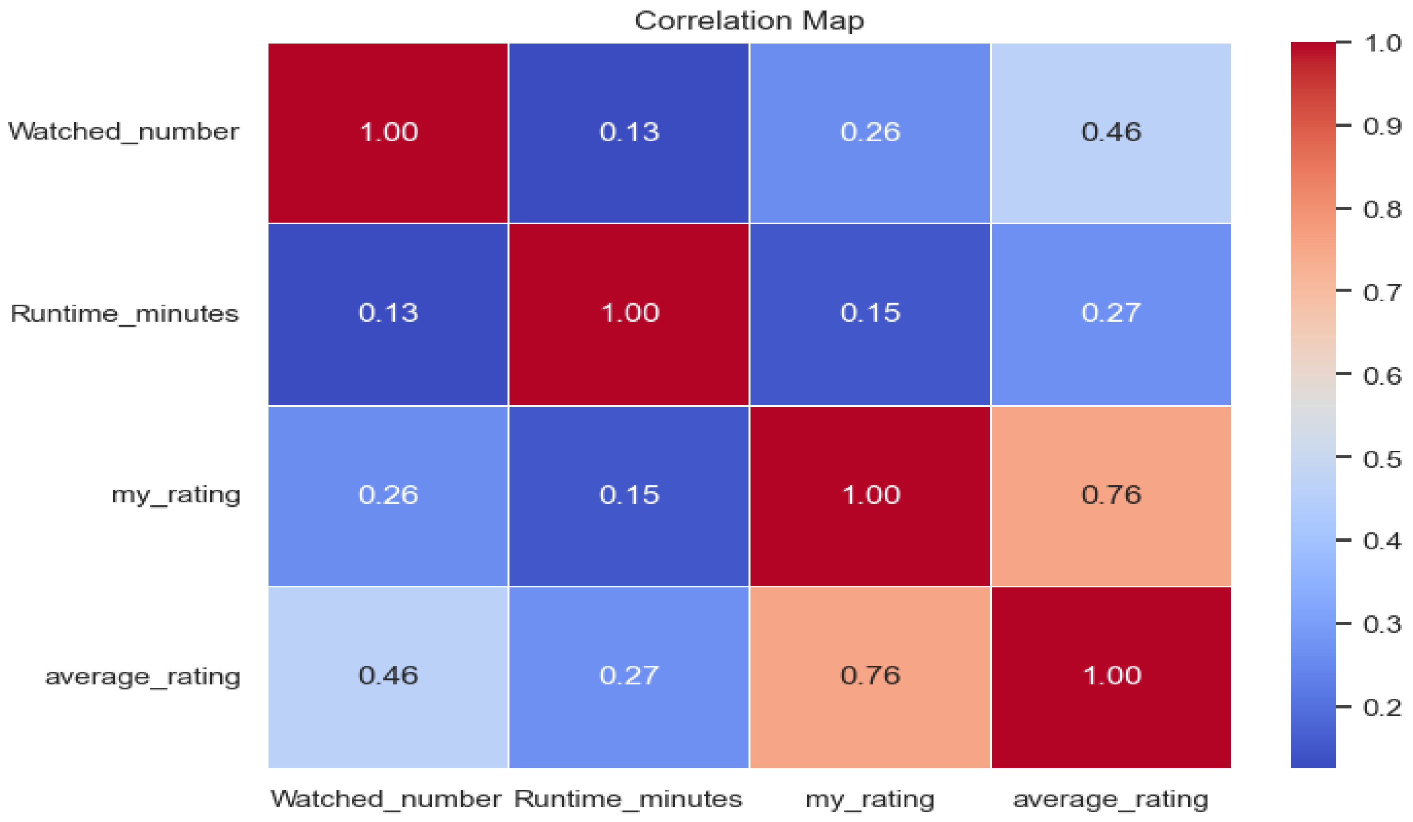
- There is a significant correlation between my ratings and the average ratings on Letterboxd.



Conclusion:

- **Reject the null hypothesis:** There is strong evidence of a correlation. This suggests that there are similarities between my ratings and Letterboxd users' ratings.

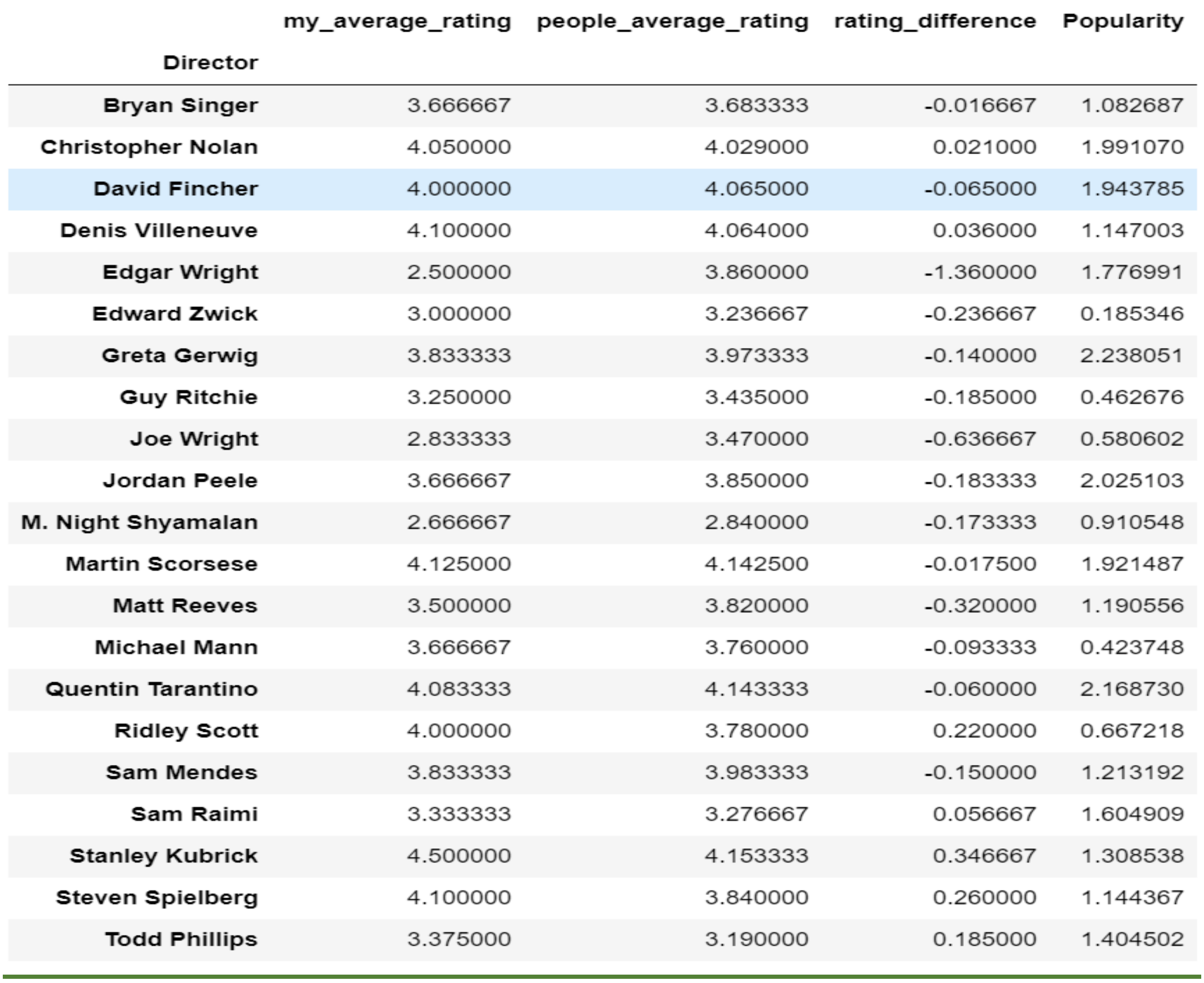
We can combine the previous results in one correlation heatmap.



As you can see there is a more correlation between my rating and average rating (0.76) rather than my rating and duration of movies (0.15).The order or correlations with my rating in decreasing order: Average rating, watched number, duration.

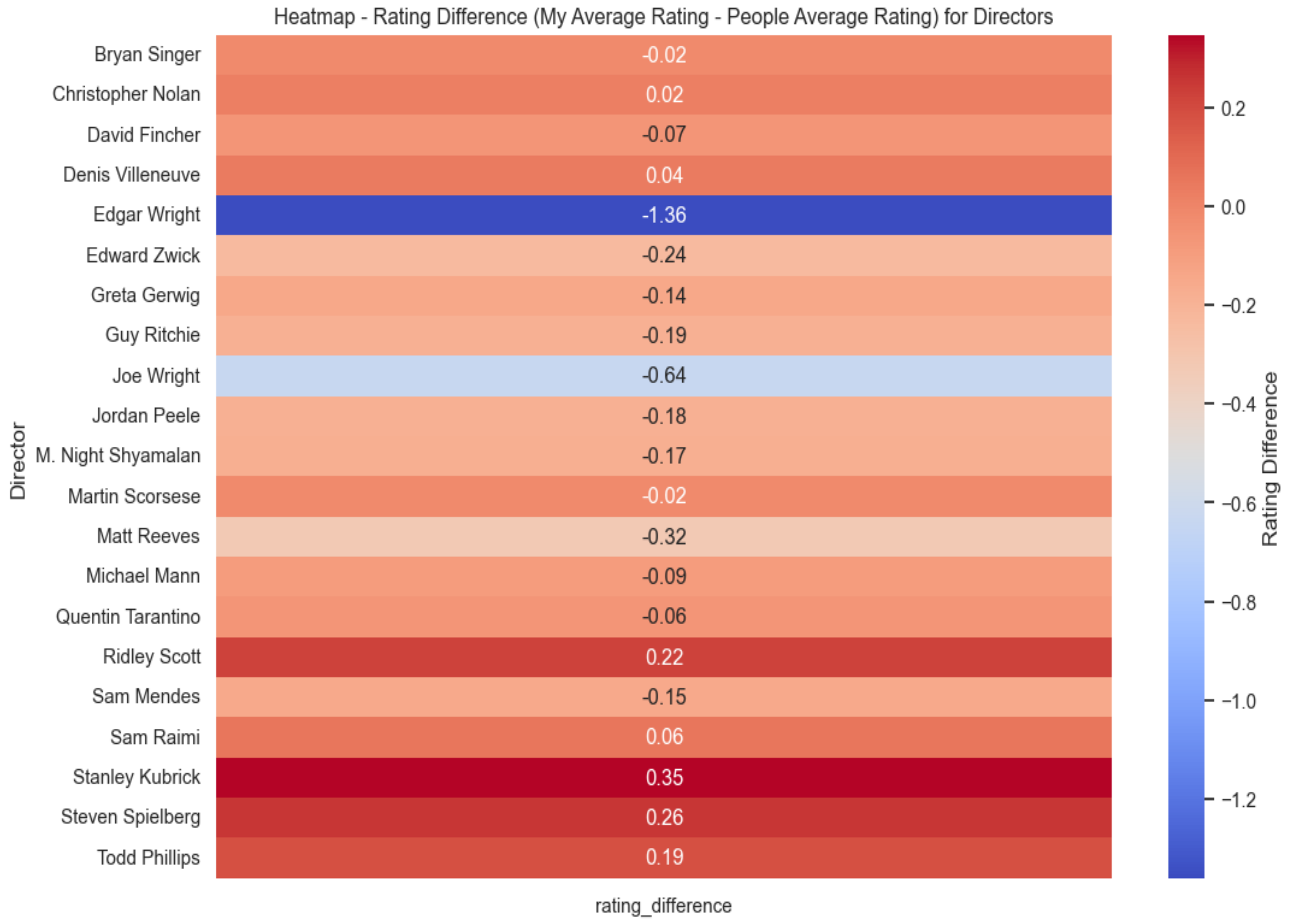
Director Dataset

After examining some hypotheses obtained from my movie dataset, I created another set called "director dataset" which includes my average rating for directors with respect to the movies that I've watched. Then I added average ratings of Letterboxd users. For better understanding I've created another column called rating difference which equals [my_average_rating – people_average_rating]. Also I added another column called "Popularity". To do that I found the average of watched number of movies for every director.



	my_average_rating	people_average_rating	rating_difference	Popularity
Director				
Bryan Singer	3.666667	3.683333	-0.016667	1.082687
Christopher Nolan	4.050000	4.029000	0.021000	1.991070
David Fincher	4.000000	4.065000	-0.065000	1.943785
Denis Villeneuve	4.100000	4.064000	0.036000	1.147003
Edgar Wright	2.500000	3.860000	-1.360000	1.776991
Edward Zwick	3.000000	3.236667	-0.236667	0.185346
Greta Gerwig	3.833333	3.973333	-0.140000	2.238051
Guy Ritchie	3.250000	3.435000	-0.185000	0.462676
Joe Wright	2.833333	3.470000	-0.636667	0.580602
Jordan Peele	3.666667	3.850000	-0.183333	2.025103
M. Night Shyamalan	2.666667	2.840000	-0.173333	0.910548
Martin Scorsese	4.125000	4.142500	-0.017500	1.921487
Matt Reeves	3.500000	3.820000	-0.320000	1.190556
Michael Mann	3.666667	3.760000	-0.093333	0.423748
Quentin Tarantino	4.083333	4.143333	-0.060000	2.168730
Ridley Scott	4.000000	3.780000	0.220000	0.667218
Sam Mendes	3.833333	3.983333	-0.150000	1.213192
Sam Raimi	3.333333	3.276667	0.056667	1.604909
Stanley Kubrick	4.500000	4.153333	0.346667	1.308538
Steven Spielberg	4.100000	3.840000	0.260000	1.144367
Todd Phillips	3.375000	3.190000	0.185000	1.404502

I decided to find out which directors I like or dislike more compared to Letterboxd users.



As observed in heatmap, I like the movies of Stanley Kubrick, Steven Spielberg, Ridley, etc. more than Letterboxd users. On the other hand, I have different opinions about Edgar Wright and Joe Wright with Letterboxd community.

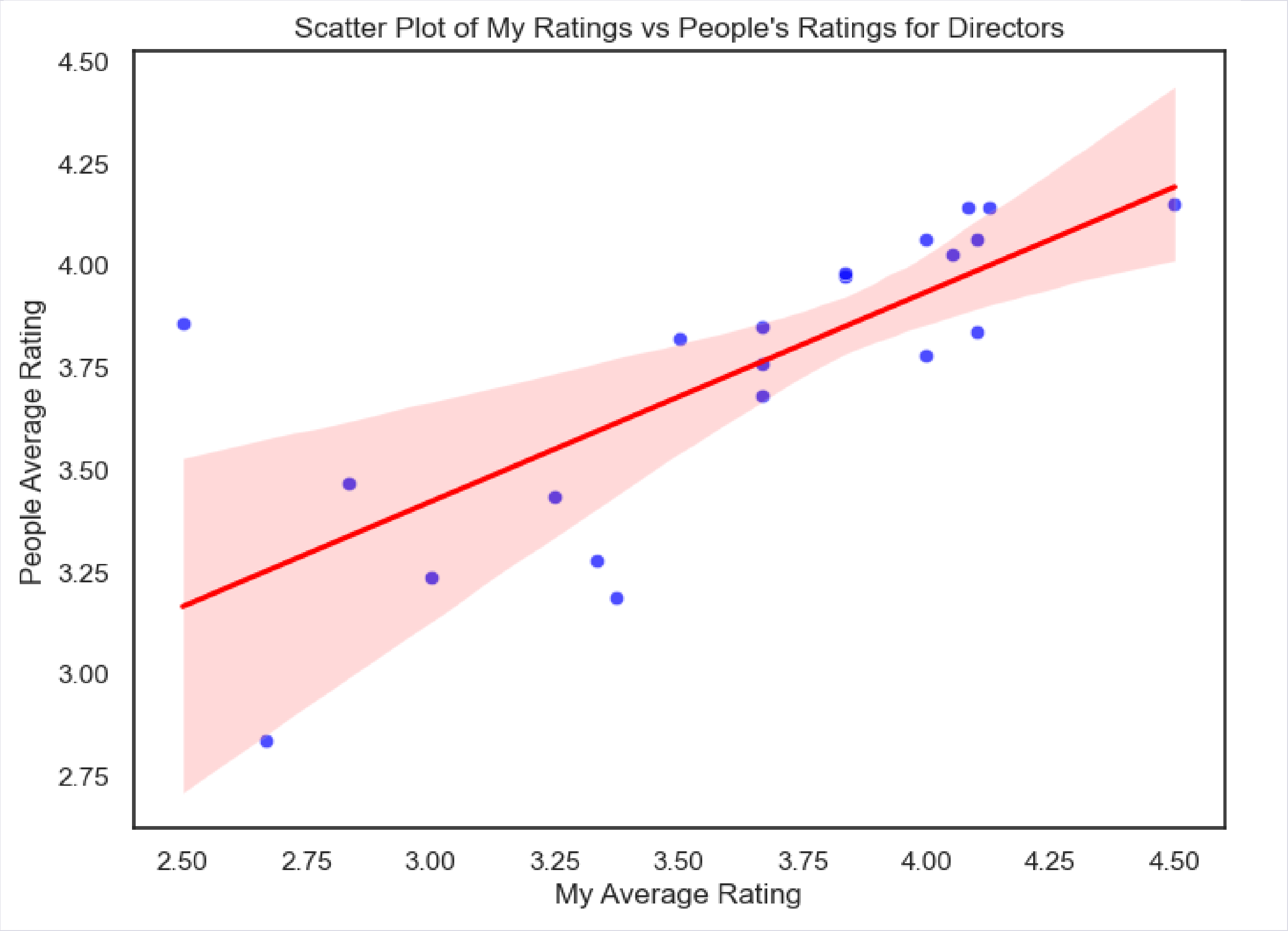
My Ratings and Average Ratings for Directors

Null Hypothesis:

- There is no correlation between my average ratings and the average ratings of Letterboxd users for directors

Alternative Hypothesis:

- There is a significant correlation between my average ratings and the average ratings of Letterboxd users for directors.



Conclusion:

We **reject the null hypothesis** as there is strong evidence of a correlation. This suggests that there are similarities between my ratings and Letterboxd users' average ratings for directors.

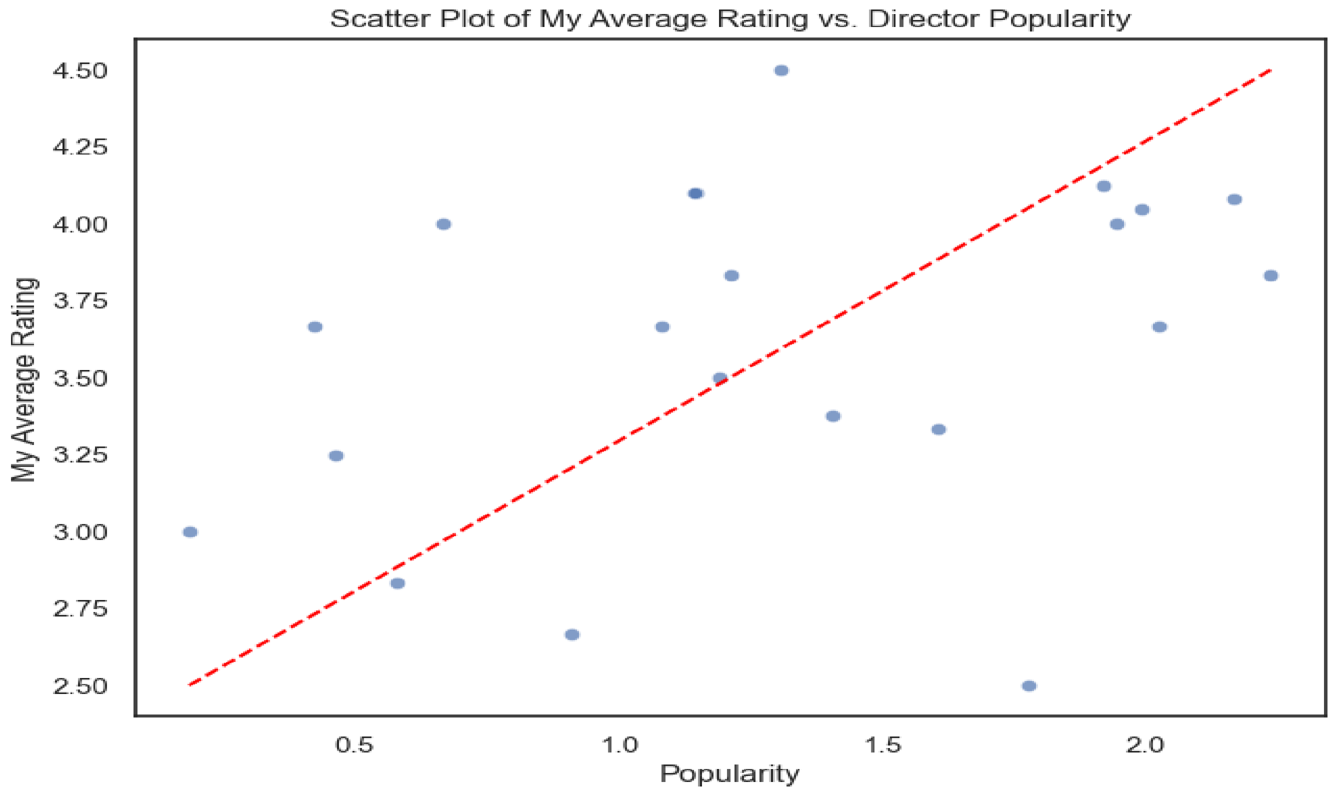
My Ratings for Directors and Their Popularity

Null Hypothesis:

- There is no correlation between my average rating for a director and the popularity of that director.

Alternative Hypothesis:

- There is a significant correlation between my average rating for a director and the popularity of that director.

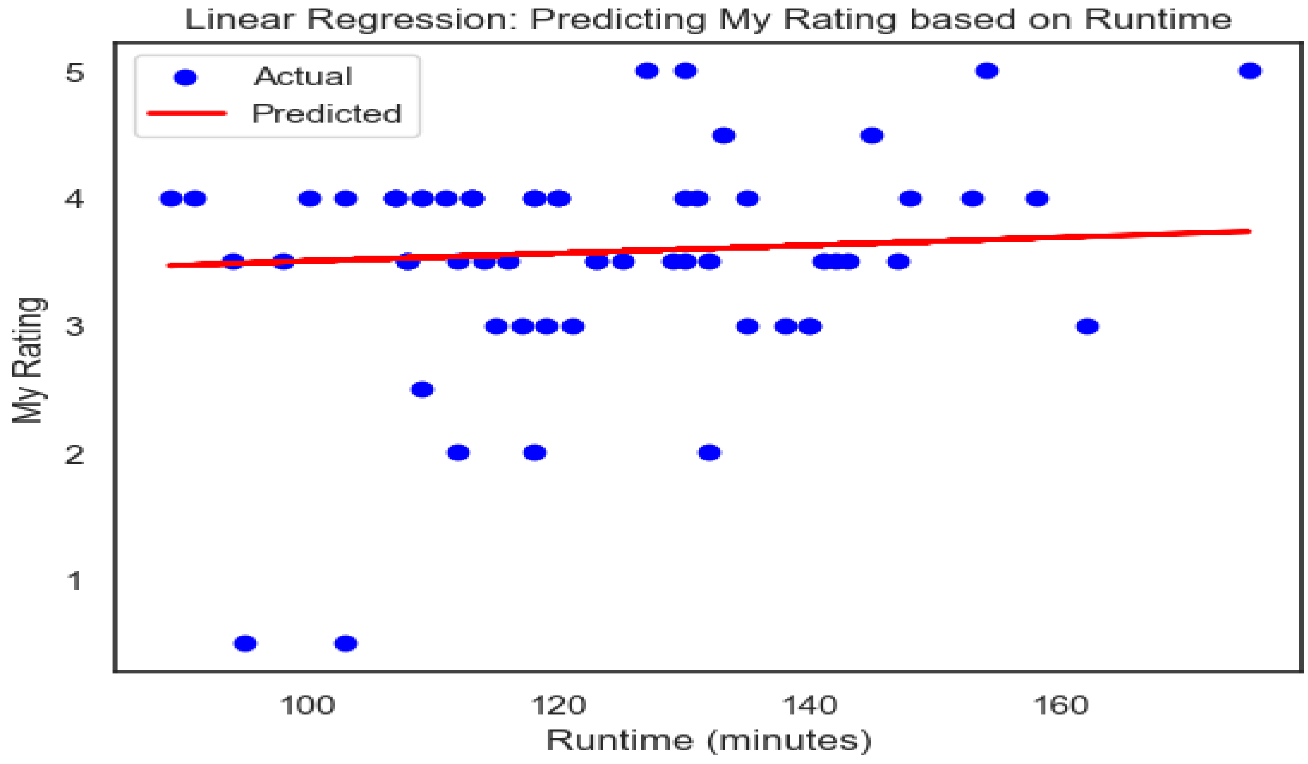


Conclusion:

We **fail to reject the null hypothesis** as there is no evidence of a correlation. This suggests that my average rating for a director may not be significantly correlated with their popularity.

Machine Learning Model That Predicts My Rating based on Runtime

Mean Squared Error: 0.6980655808446911



Limitations

- ***Sample Size for Directors:*** The number of movies I have watched for each director may vary significantly. For instance, I have watched ten movies for Nolan, whereas there are only three for Poole. This may lead to bias in my rating.

- ***Less Feature:*** I could incorporate more features into my project such as genres, release year etc.

Future Work

- ***Collaborative Filtering:*** Implement collaborative filtering techniques to recommend movies based on your preferences and compare them against your actual viewing patterns.

- ***Machine Learning Models:*** Experiment with machine learning models for predicting your ratings based on various features. This could uncover hidden patterns and improve the accuracy of predictions.

- ***Temporal Analysis:*** Analyze changes in your preferences over time. Are there certain periods where you tend to rate movies higher or lower? This could be influenced by personal experiences or external factors.