PREDICTING IMDB STAR RATINGS

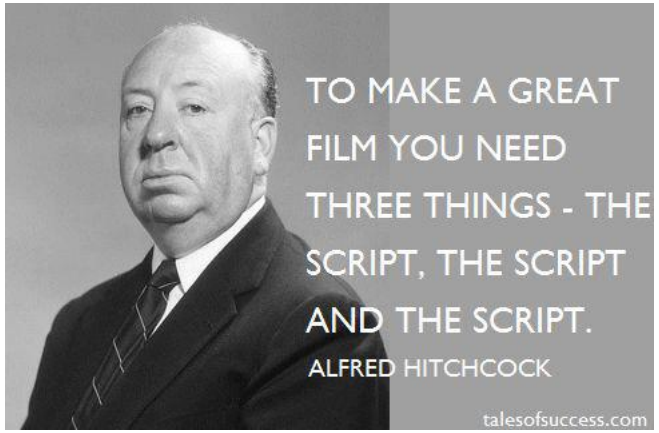
Michael Green

TOPICS

- Description of the Problem
- Data Source
- Features In Data Source Used In Model
- Synthesized Features
- Top 5 Positive and Negative Correlation Features
- Linear Equation with Coefficients
- Final Test Score
- Residual Plot, Distribution of Error, QQ Plot
- Future Work
- Appendix

DESCRIPTION OF THE PROBLEM

- There are many ideas of how to make a good movie:



Cinema is a matter of what's in the frame and what's out.
— Martin Scorsese

DESCRIPTION OF THE PROBLEM

- The problem is that these are not easily quantifiable.
- Many content creation companies (like Netflix) want be able to make movie making/buy decisions faster and with greater efficacy and over a larger volume of movie proposals.
- This project develops a model that uses as inputs features of movies that can be *well known on or before its release date* to *predict* if the movie will be well liked by viewers.

DESCRIPTION OF THE PROBLEM

- This model could then assist executives by doing a quick first pass scoring of large volumes of movie proposals.
- This would enable movie executives to then focus valuable deep analysis time on already quantitatively promising movies.

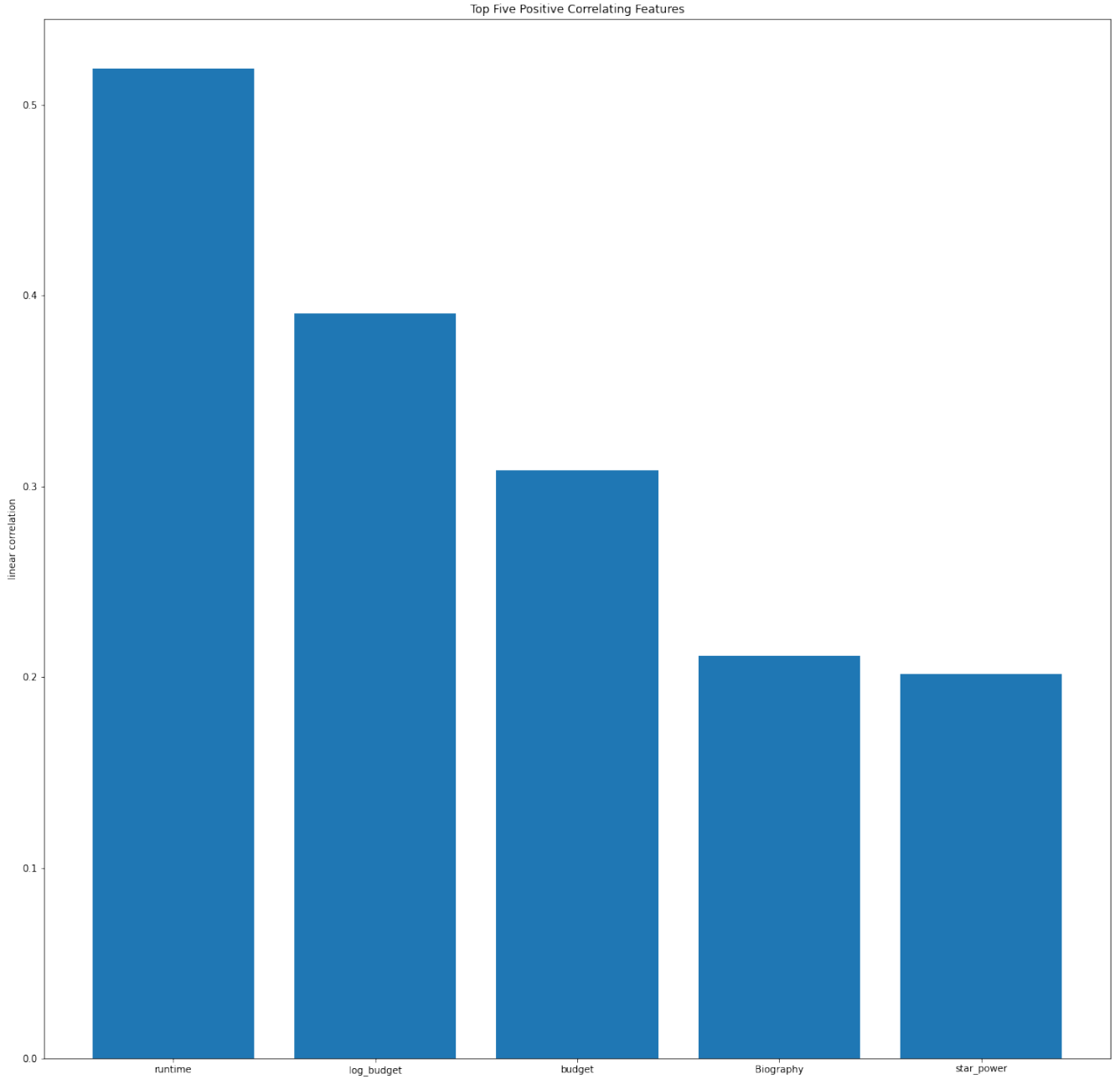
DATA SOURCE

- The starting data set are 4000 movies released in the US with an MPAA rating between January 1, 2010 and December 31, 2019.
- This original data set was then cleaned to 2127 movies that had all required data fields defined.
- The data was scraped from imdb.com using Python `requests` and `BeautifulSoup`.

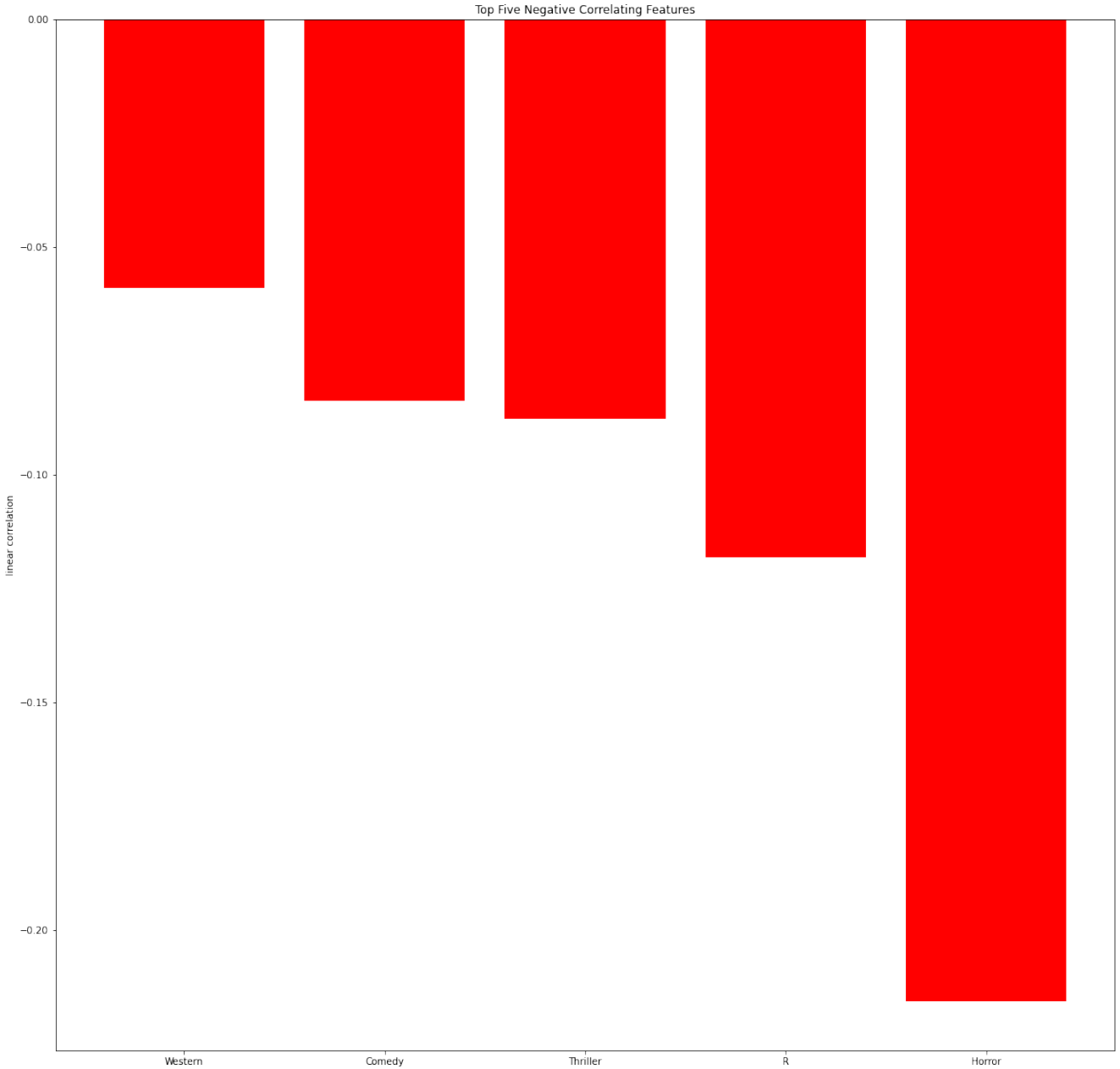
FEATURES IN DATA SOURCE USED IN MODEL

- The following features were used in the model:
  - Runtime: length of the movie in minutes
  - Budget: Amount of money it took to make the movie (in US $)
  - Release month of the movie: One-hot encoded (12 categories)
  - Genre of the movie: One-hot encoded (17 categories)
  - MPAA Rating of the Movie: (4 categories)
  - One-hot encoded value of the month the movie was released (12 categories encoded
  - One-hot encoded value of the genre of the movie (17 categories encoded)
  - **Director Star Power**: 1 point for each Best Director Award for earned or nominated by the director of said movie
  - **Cast 1 Star Power**: 1 point for each Best Actor or Best Actress Award earned or nominated by the 1st cast member listed for the movie
  - **Cast 2 Star Power**: 1 point for each Best Actor or Best Actress Award earned or nominated by the 2nd cast member listed for the movie
  - **Cast 3 Star Power**: 1 point for each Best Actor or Best Actress Award earned or nominated by the 3rd cast member listed for the movie
  - **Log(Budget):** Logarithm of the Budget

TOP 5 POSITIVE AND NEGATIVE CORRELATION FEATURES



TOP 5 POSITIVE AND NEGATIVE CORRELATION FEATURES



LINEAR EQUATION WITH COEFFICIENTS

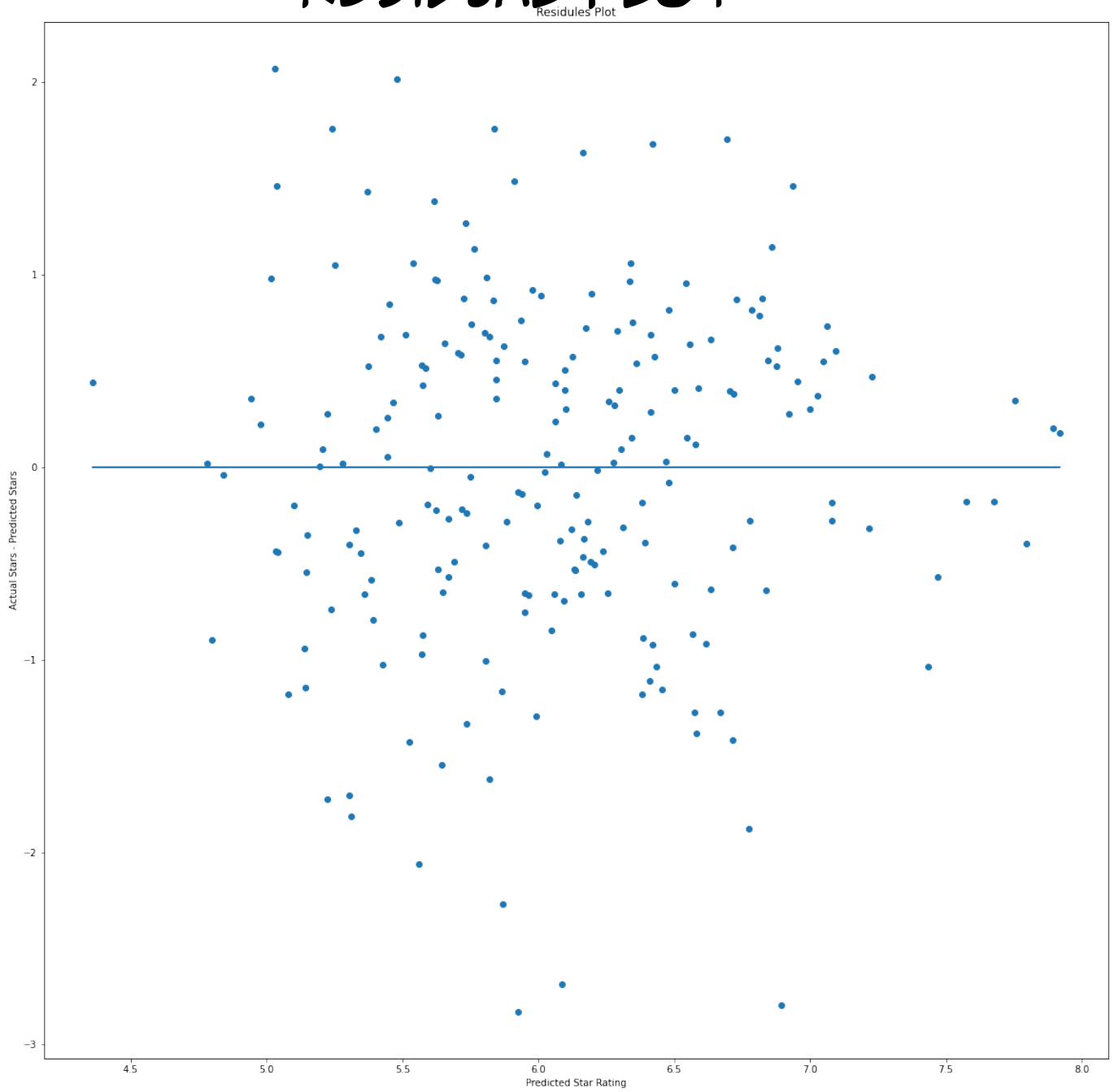
$$
\begin{aligned}
y = {} & 0.0258\text{runtime} \\
& + 1.878\text{budget} \\
& + 0.0449\text{August} \\
& - 0.1492\text{December} \\
& 0.1206\text{February} \\
& - 0.0363\text{January} \\
& -0.0672\text{July} \\
& - 0.0296\text{June} \\
& - 0.0638\text{March} \\
& - 0.1692\text{May} \\
& + 0.06582\text{November} \\
& + 0.0294\text{October} \\
& + 0.005259\text{September} \\
& + 0.3251\text{Adventure} \\
& + 0.8829\text{Animation} \\
& + 0.7957\text{Biography} \\
& + 0.2936\text{Comedy} \\
& + 0.3785\text{Crime} \\
& + 0.5407\text{Drama} \\
& - 0.3144\text{Family} \\
& - 0.42562\text{Fantasy} \\
& - 0.08323\text{Horror} \\
& + 1.6462\text{Music} \\
& + 2.198\text{Musical} \\
& + 1.003\text{Myster} \\
& + 0.45469\text{Romance} \\
& + 0.9918\text{Sci-Fi} \\
& + 1.93200\text{Sport} \\
& - 0.53621\text{Thriller} \\
& - 1.778\text{Wester} \\
& - 1.146\text{PG} \\
& - 0.977\text{PG-13} \\
& - 0.9303\text{R} + 0.18929\text{star\_power} + 0.10356*\text{cast1\_starpower} + 0*(\text{cast2\_starpower} + \text{cast3\_starpower}) + 0.14988\text{budget} \\
& + 1.5468
\end{aligned}
$$

FINAL TEST SCORE

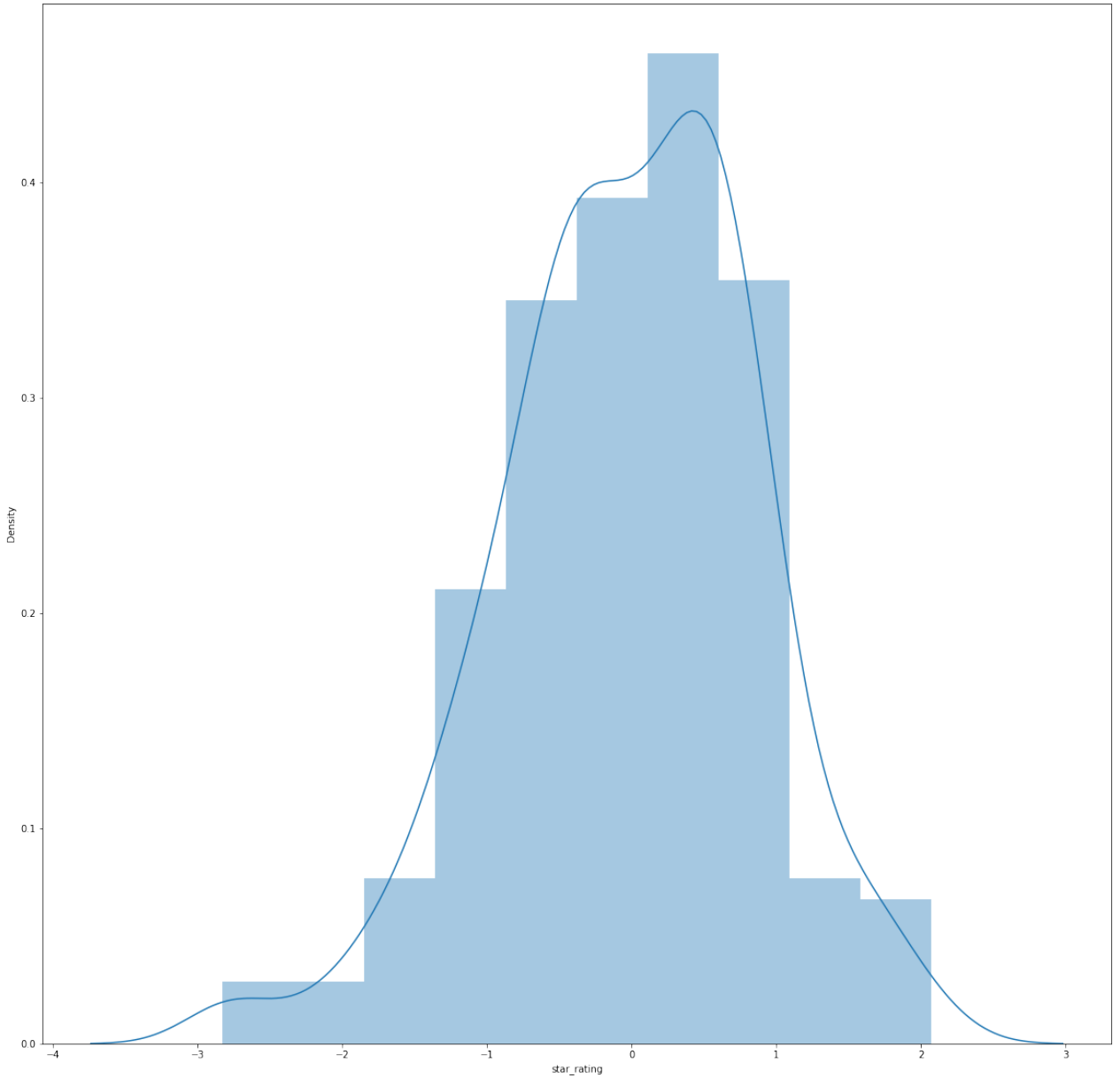
$$R^2 = 0.363$$

RESIDUAL PLOT

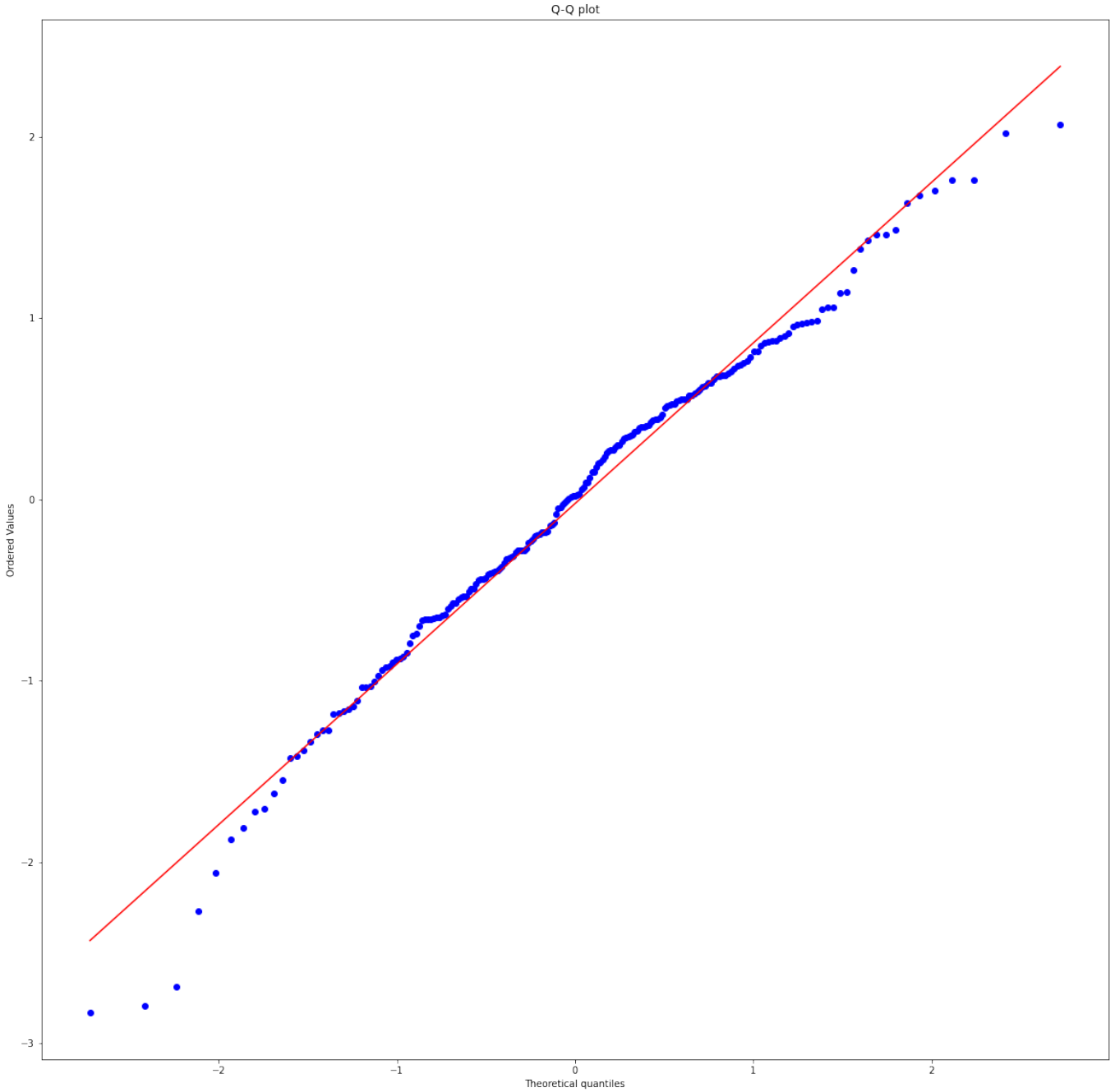
Mean of the Error = -0.02



DISTRIBUTION OF ERROR



QQ PLOT


Q-Q plot
Ordered Values
Theoretical quantiles
2
1
0
−1
−2
−3
−2
−1
0
1
2

FUTURE WORK

- Scrape more data. There are over 50,000 movies that have been released over the last 10 years.
- Collect more data on the cast and crew: How many producers on the movie have been awarded Best Picture, how many writers have been awarded for Best Adapted Screen Play.
- Dynamic Duos, Trios, Quartets, Etc.: Anecdotally there are a lot of movies where you see the same (director, cast) combinations. Quantify the level of influence the existence of these tuples have on the star rating of the movie.

APPENDIX