# PREDICTING IMDB STAR RATINGS

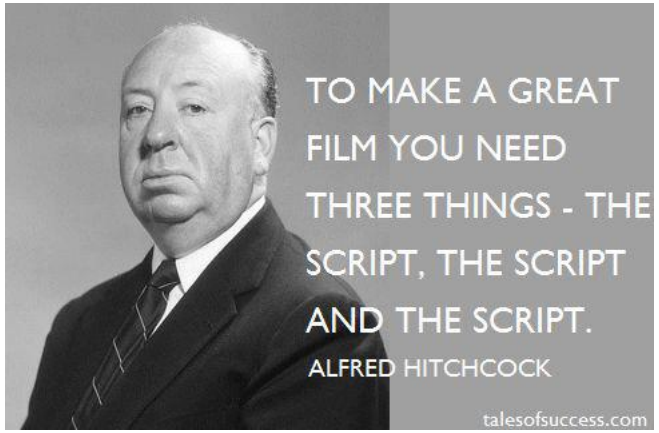
Michael Green

# 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.

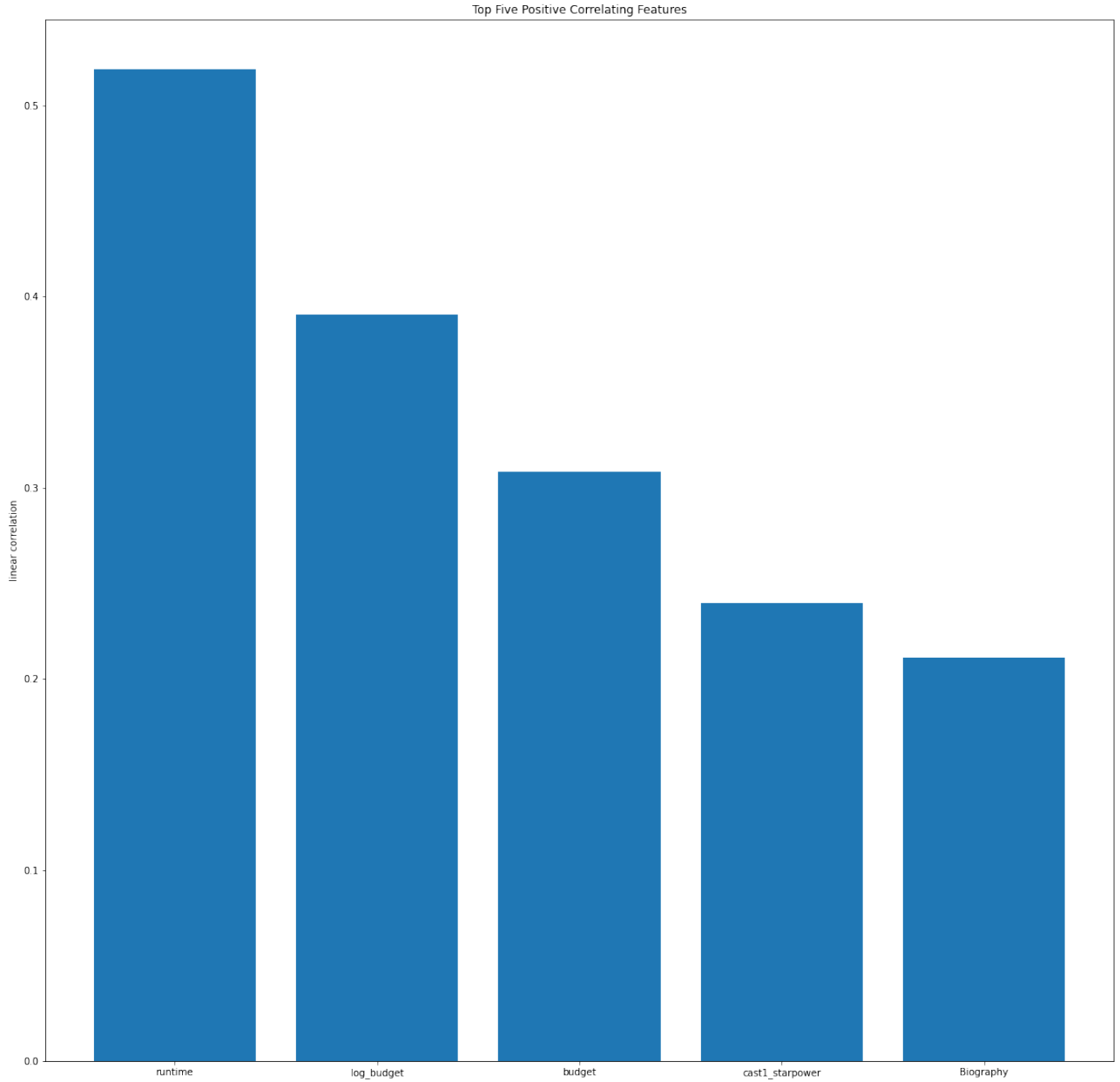
# DATA SOURCE

- 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`.

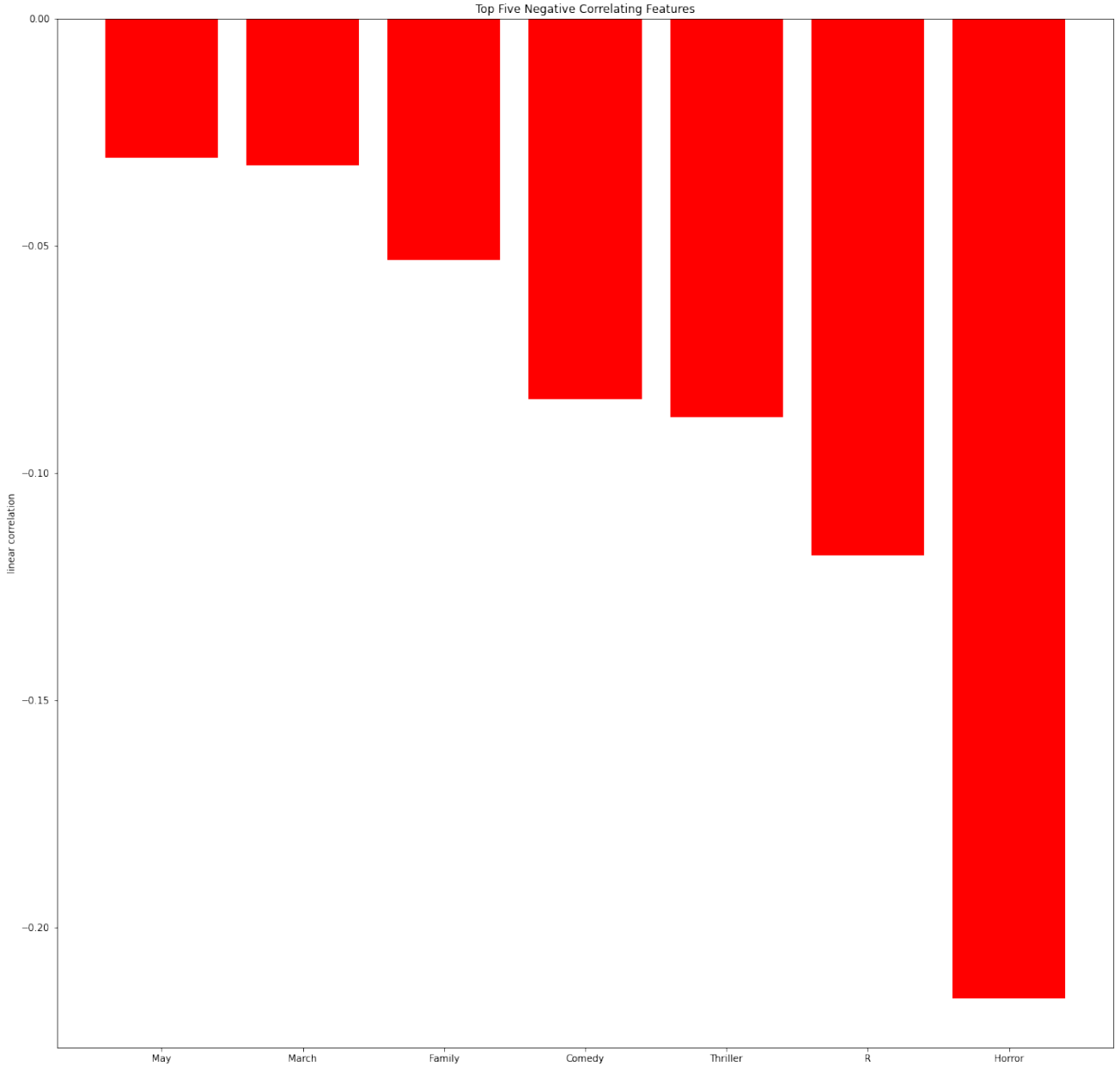
# FEATURES IN DATA SOURCE USED IN MODEL

- 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)
  - **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

```
[('runtime', 0.025214543277458278),
 ('budget', 6.807720486700231e-10),
 ('August', 0.02094246922183872),
 ('December', -0.22789909501589561),
 ('February', -0.1184072095835705),
 ('January', -0.06038472276660286),
 ('July', -0.1268907930387392),
 ('June', -0.051562439208146287),
 ('March', -0.12235181076357225),
 ('May', -0.24310040744285022),
 ('November', -0.006389011115727724),
 ('October', -0.03604633930763422),
 ('September', 0.0033420202986247424),
 ('Adventure', 0.28414389139343926),
 ('Animation', 0.8983631145800551),
 ('Biography', 0.7933746547342677),
 ('Comedy', 0.29718697754938167),
 ('Crime', 0.36724144420807003),
 ('Drama', 0.5253843742214763),
 ('Family', -0.4386695937735968),
 ('Fantasy', -0.37646970918484046),
 ('Horror', -0.03808708876117907),
 ('Music', 1.702680877663265),
 ('Musical', 2.1960197057099213),
 ('Mystery', 0.8706774268621756),
 ('Romance', 0.34865681193241155),
 ('Sci-Fi', 0.8986524328798289),
 ('Sport', 0.24860044186789365),
 ('Thriller', -0.3966856926089714),
 ('PG', -1.0397069536832078),
 ('PG-13', -0.8861972801566934),
 ('R', -0.8341237620032511),
 ('star_power', 0.1678769616927409),
 ('cast1_starpower', 0.14783901005356814),
 ('cast2_starpower', 0.032564072697596945),
 ('cast3_starpower', 0.11120695608145978),
 ('log_budget', 0.13950182412867151)]
```

```
model3.intercept_
```

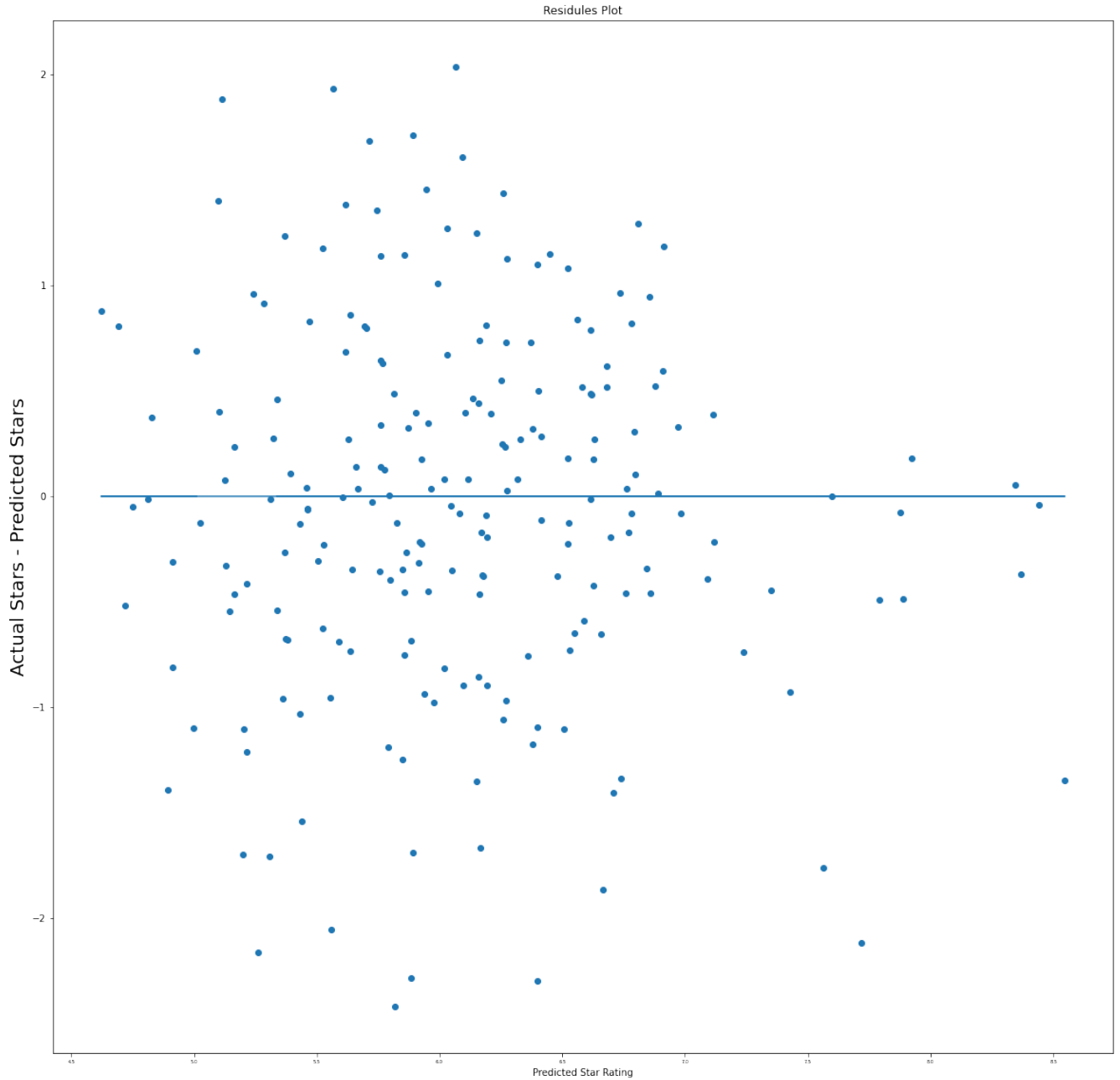
```
1.6797817451895325
```

# FINAL TEST SCORE

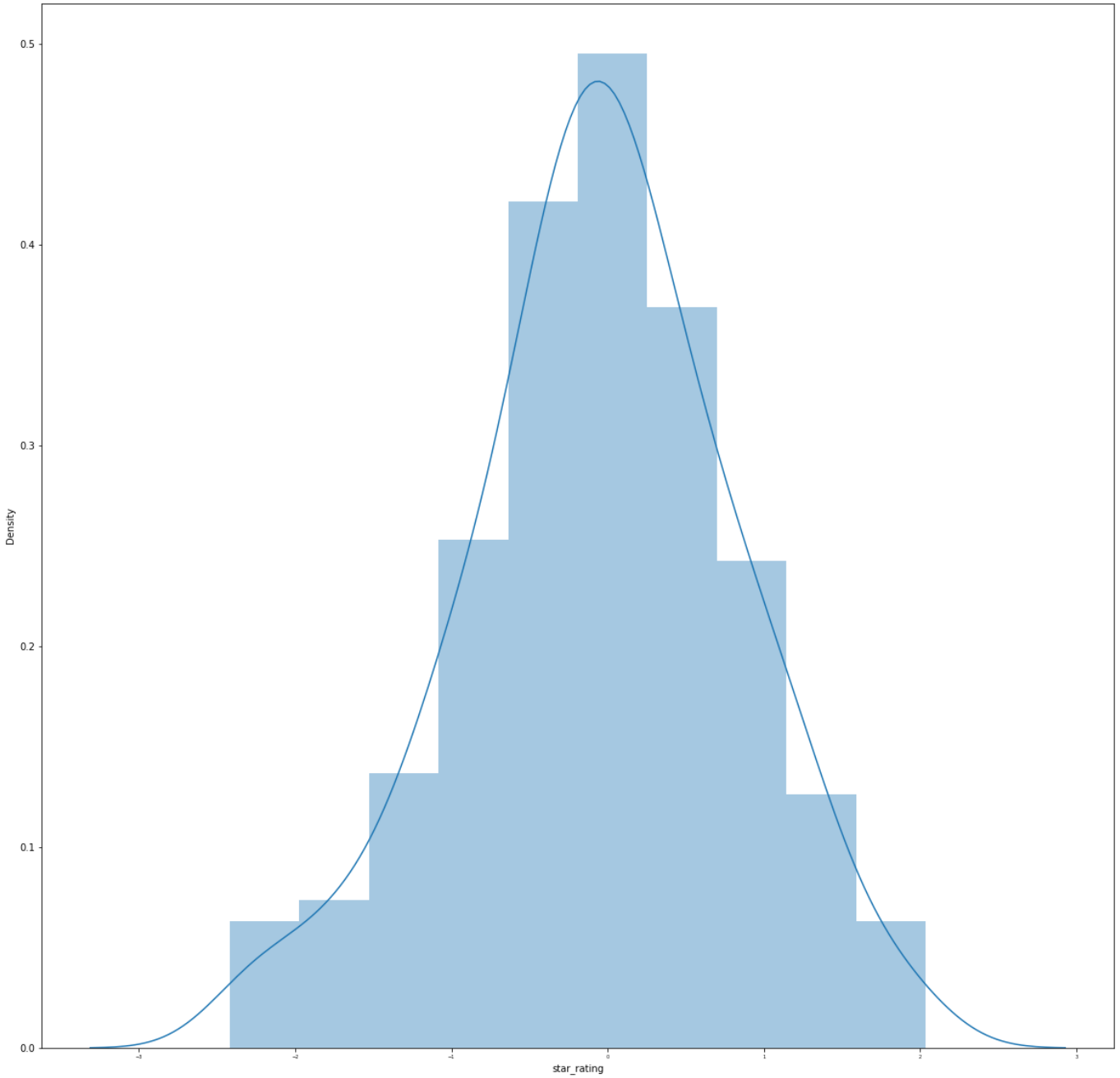
$$R^2 = 0.3839$$

# RESIDUAL PLOT

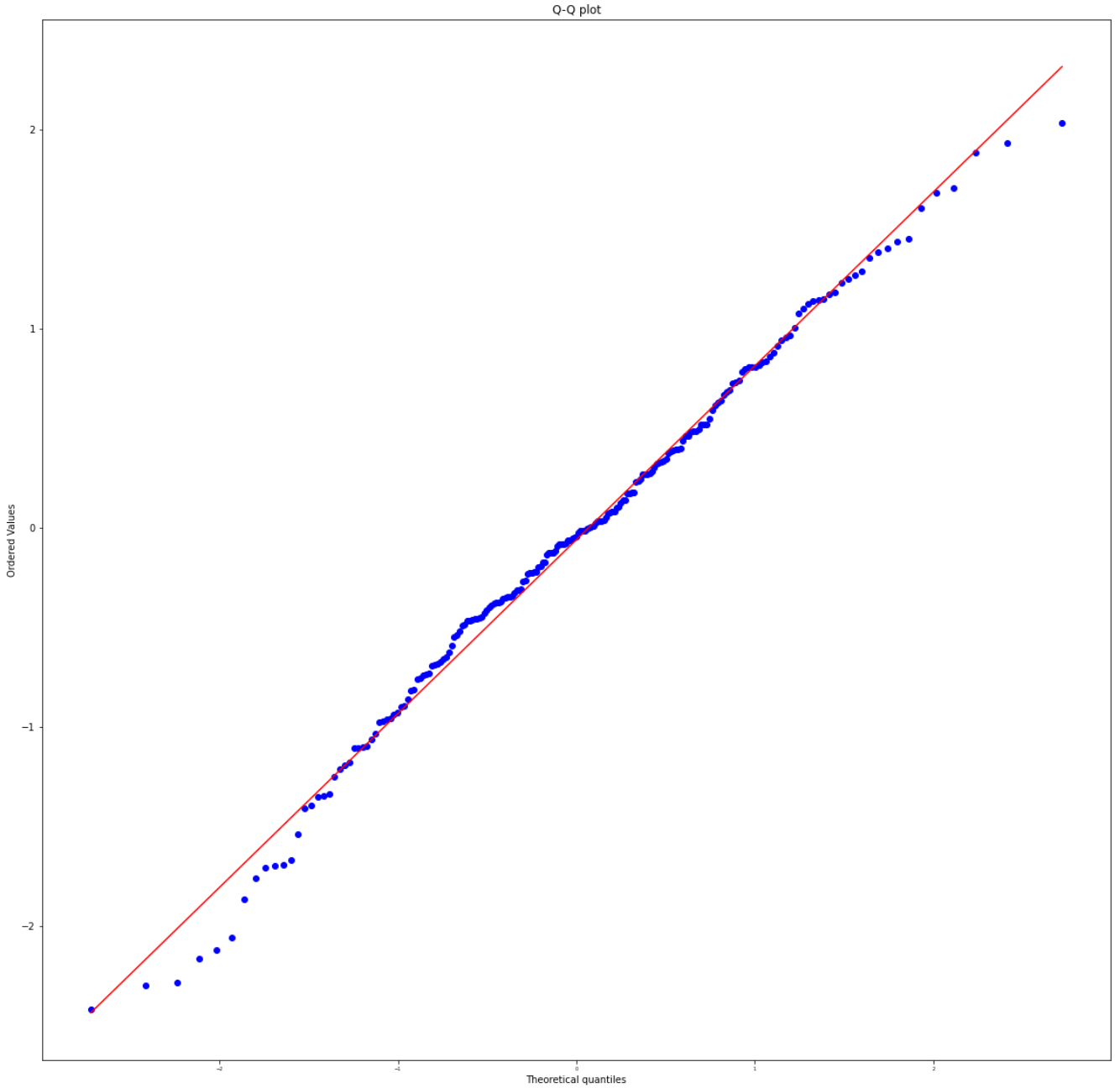
Mean of the Error = -0.05



# DISTRIBUTION OF ERROR



# QQ PLOT



# 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