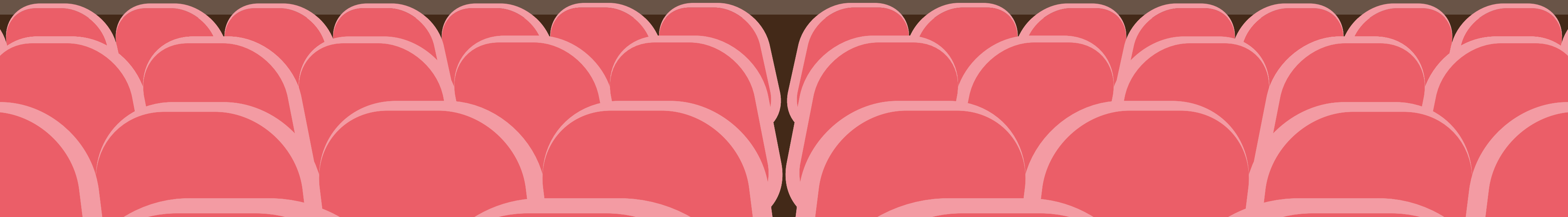
WILL YOUR MOVIE BE A HIT OR FLOP?

Jenny Wang



AGENDA

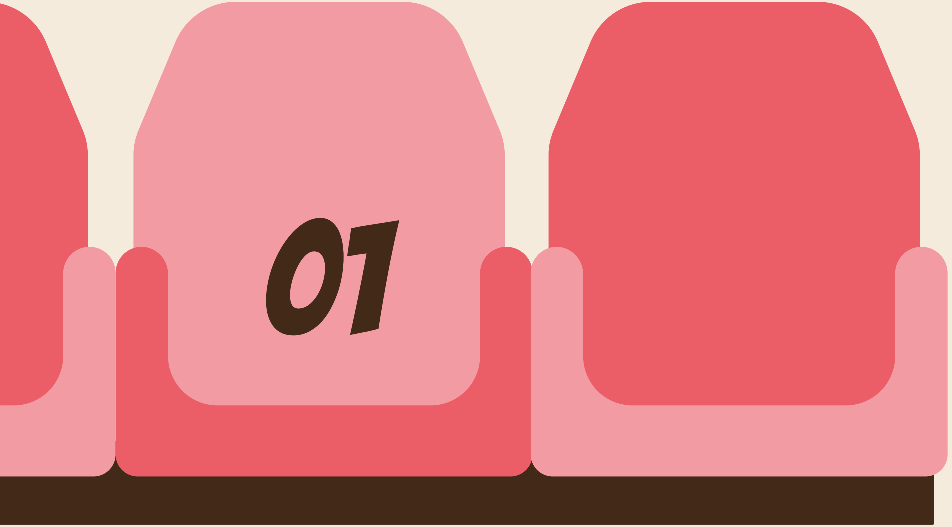
01. CONTEXT

02. GOAL

03. APPROACH

04. RESULTS

05. FUTURE WORK



CONTEXT

DOMESTIC YEARLY BOX OFFICE

$7.5 B

2000

$10.5 B

2010

$11.3 B

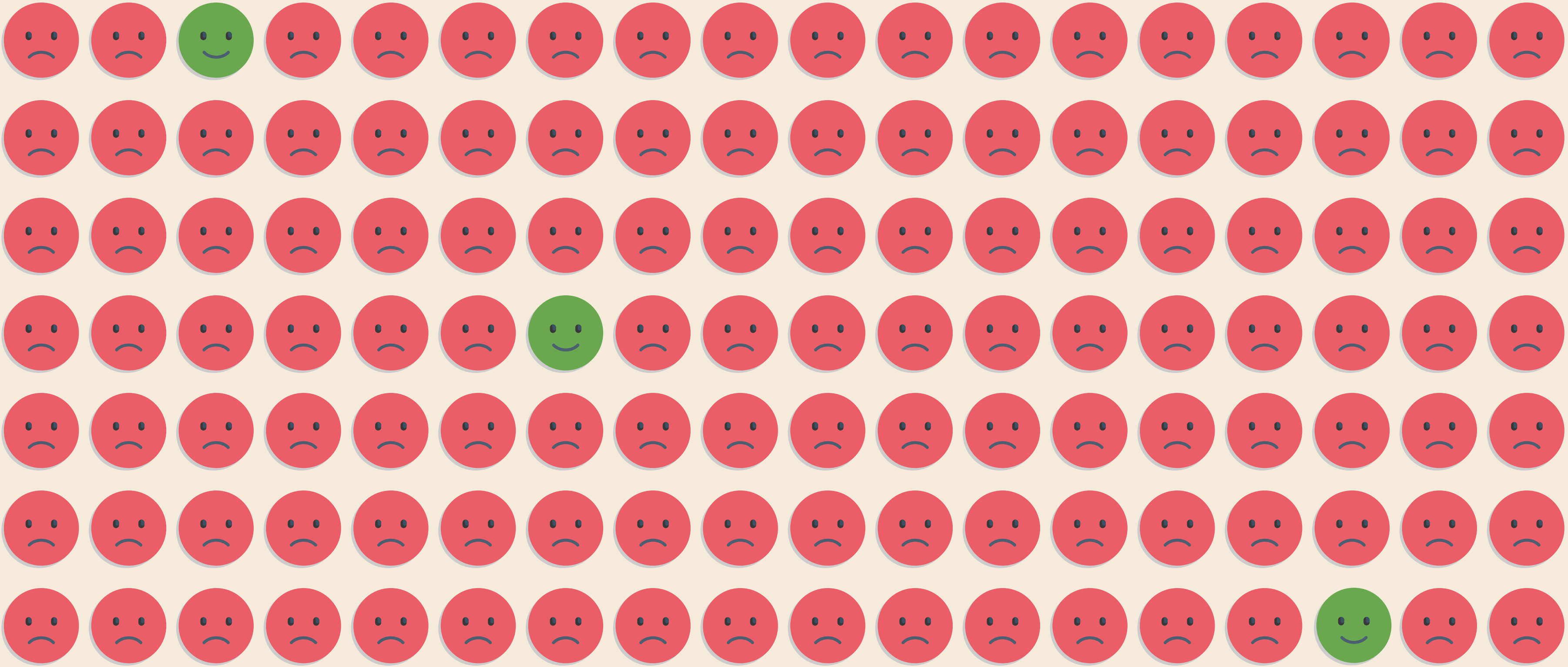
2019

WHY SHOULD YOU CARE?

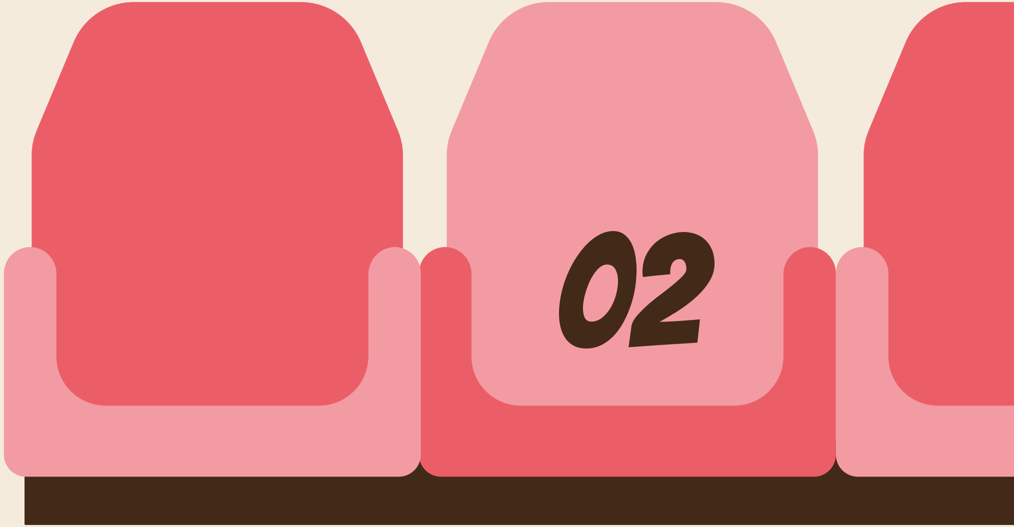
This movie is one of those that I wish I’ve never seen.

Bob Mondiabric
Film Critic, NPR

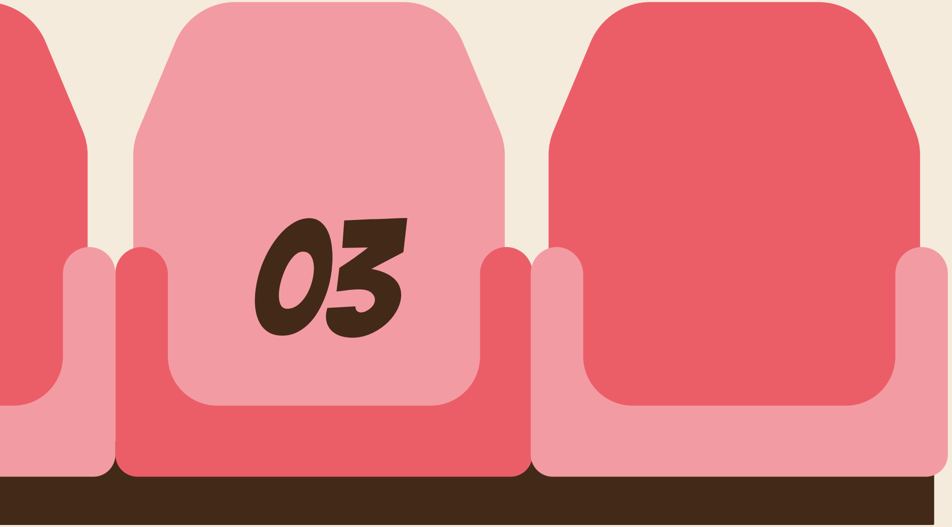
WHY SHOULD YOU CARE?



GOAL



WHAT SHOULD YOU AIM FOR YOUR NEXT MOVIE?



03

APPROACH

1

Web Scraping

Gathered information of **1,000 movies** from IMDb website

2

Data Scrubbing

Handle outliers and null values

3

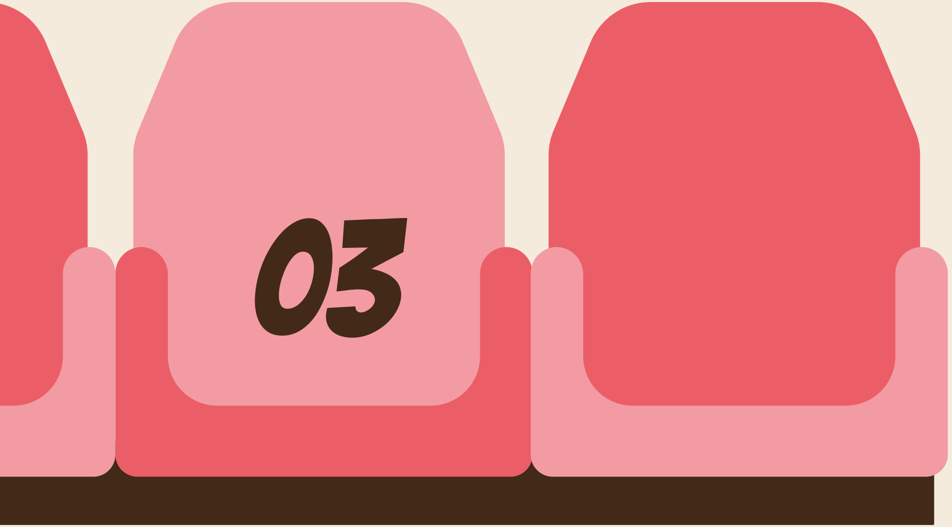
Data Exploring

Inspect correlation and perform feature engineering

4

Data Modeling

Linear regression and train/validate model



RESULTS

INDEPENDENT VARIABLES IN THE MODEL

Runtime

Genre (Drama, Action, Thriller, Horror)

Number of Reviews

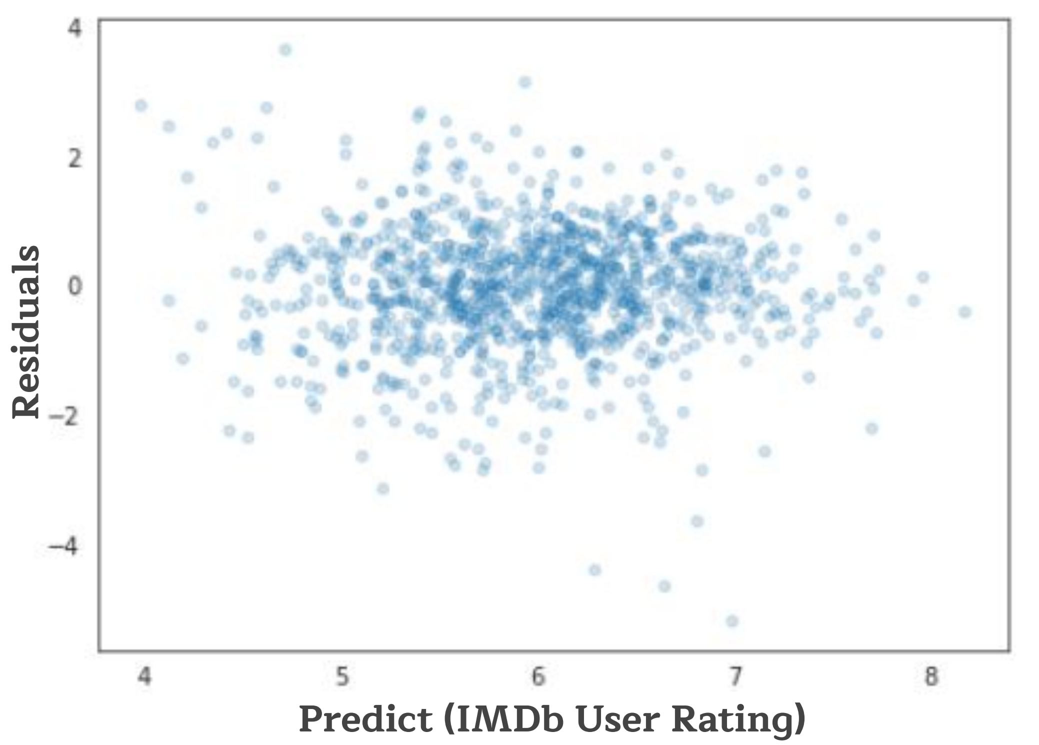
HOW DOES OUR PREDICTION TURN OUT?

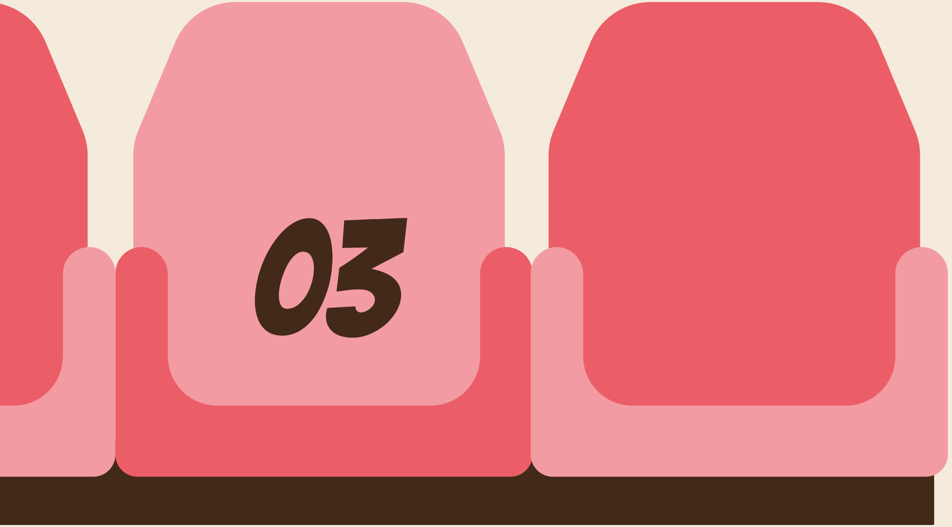
0.34

R^2

0.68

Mean Absolute Error





03

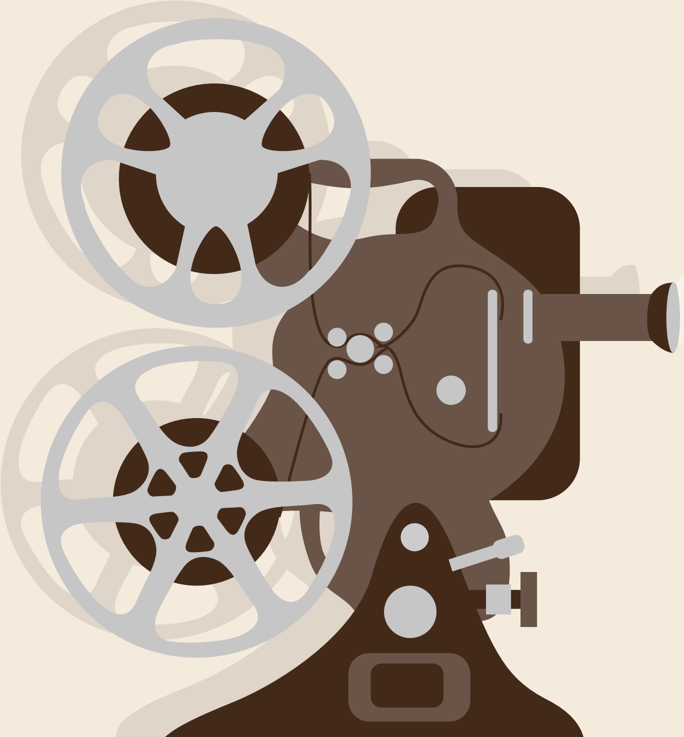
FUTURE WORK

FEATURES TO INCLUDE IN THE MODEL

Numbers of Nominations and Awards

Demographics of actors/actresses

Critics Score

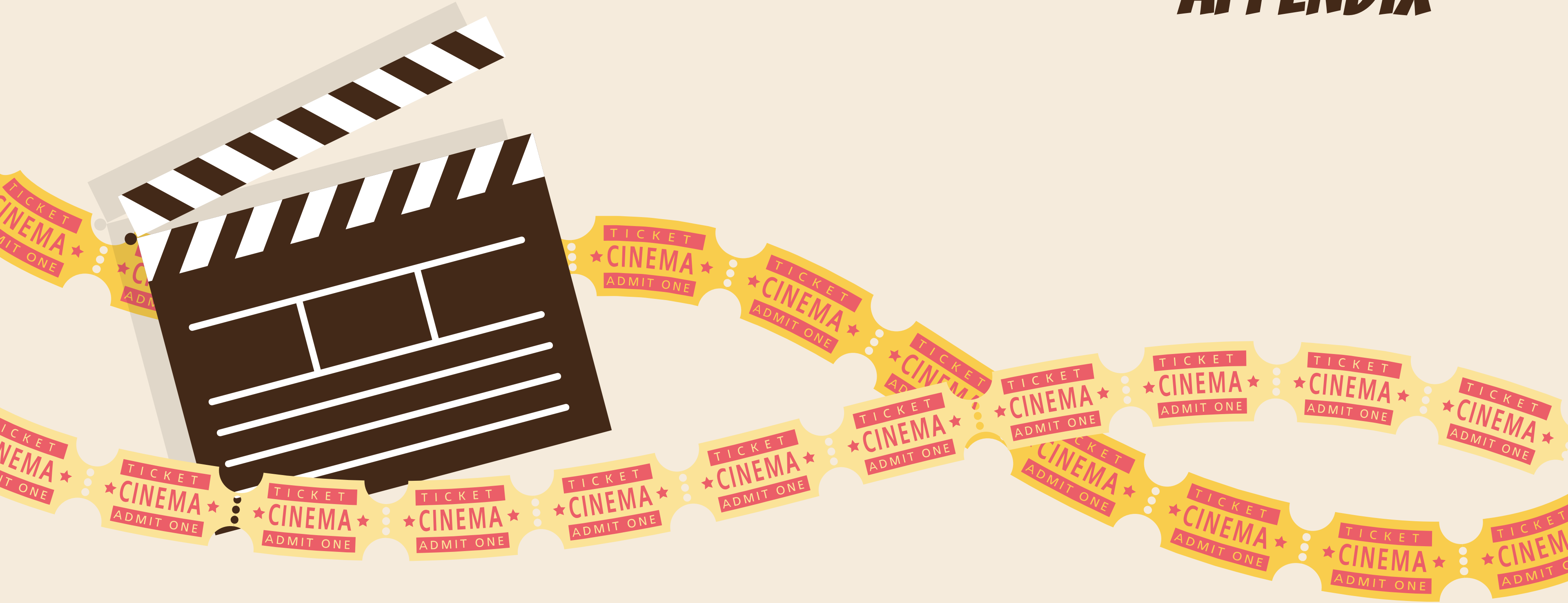


THANK YOU

ms.jcwang@gmail.com
+1 650 476 5524
alohajenny.github.io

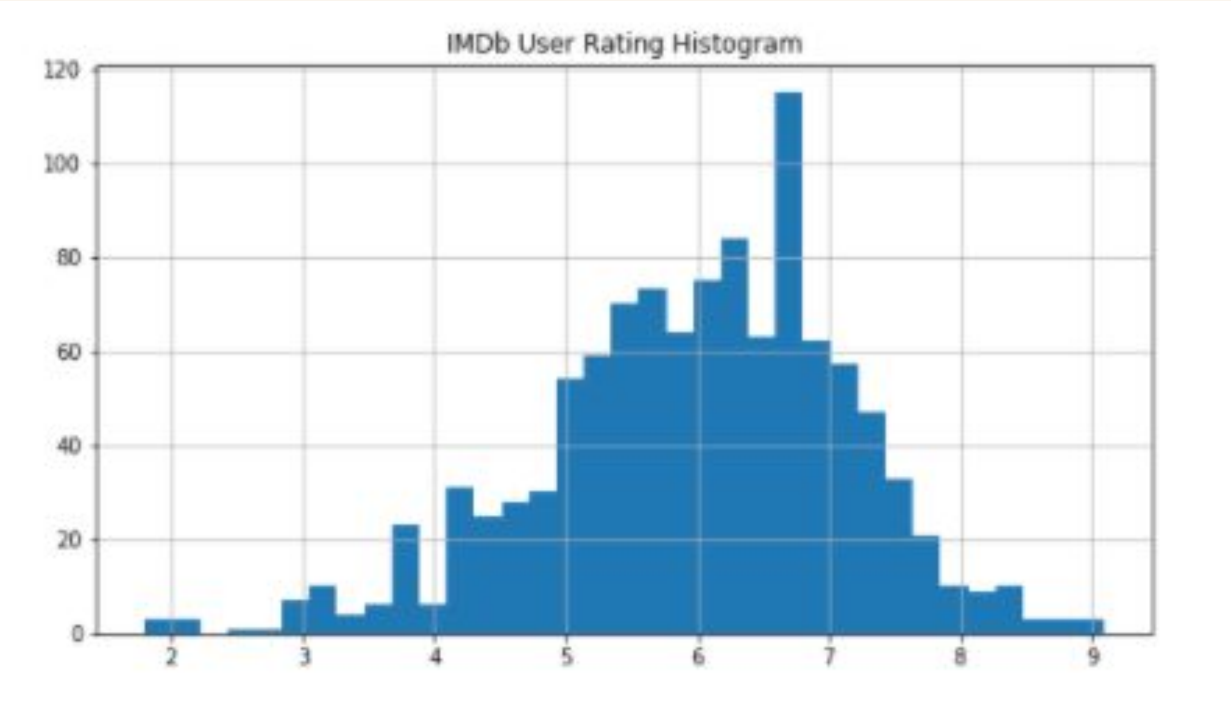
CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.

APPENDIX

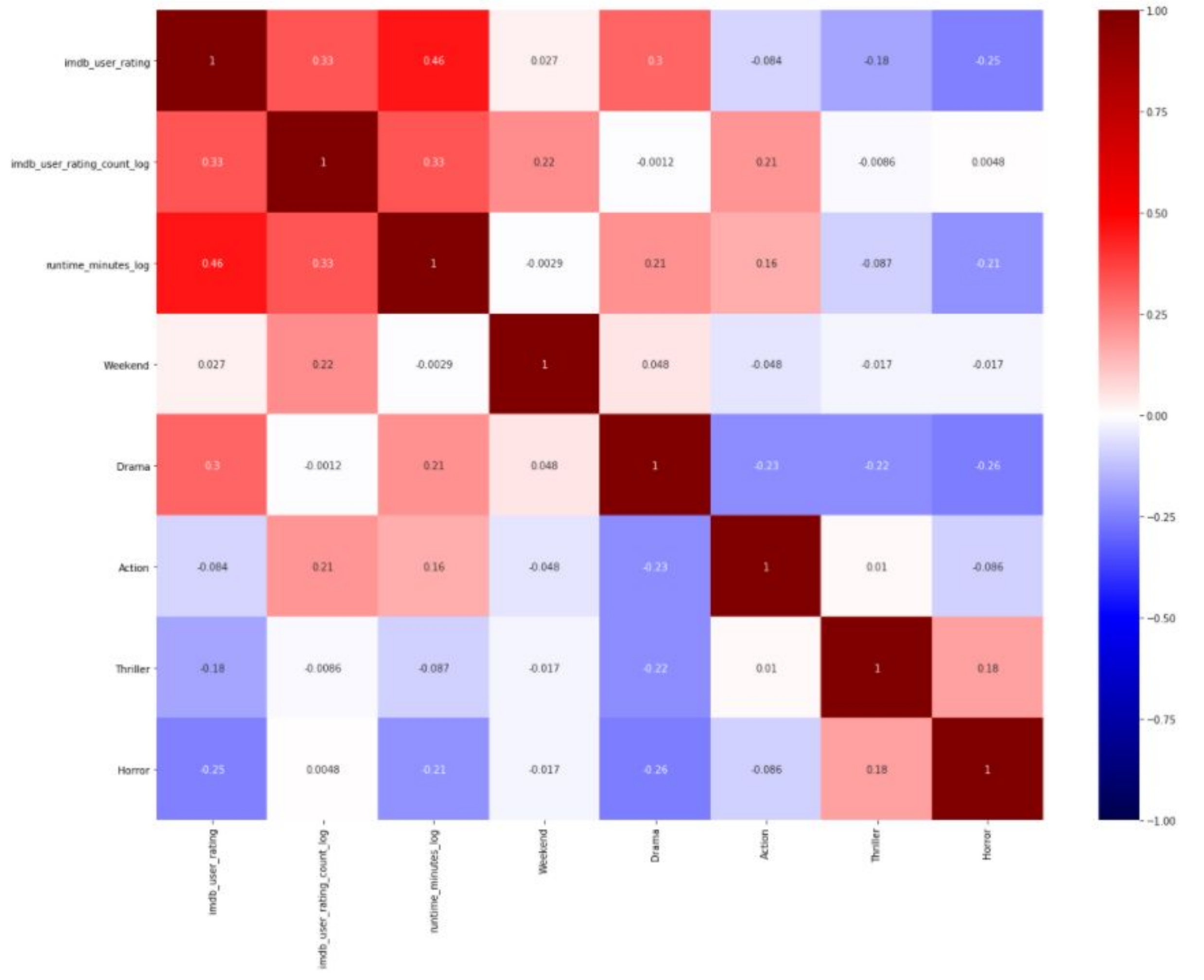


APPENDIX

HISTOGRAM OF DEPENDENT VARIABLE



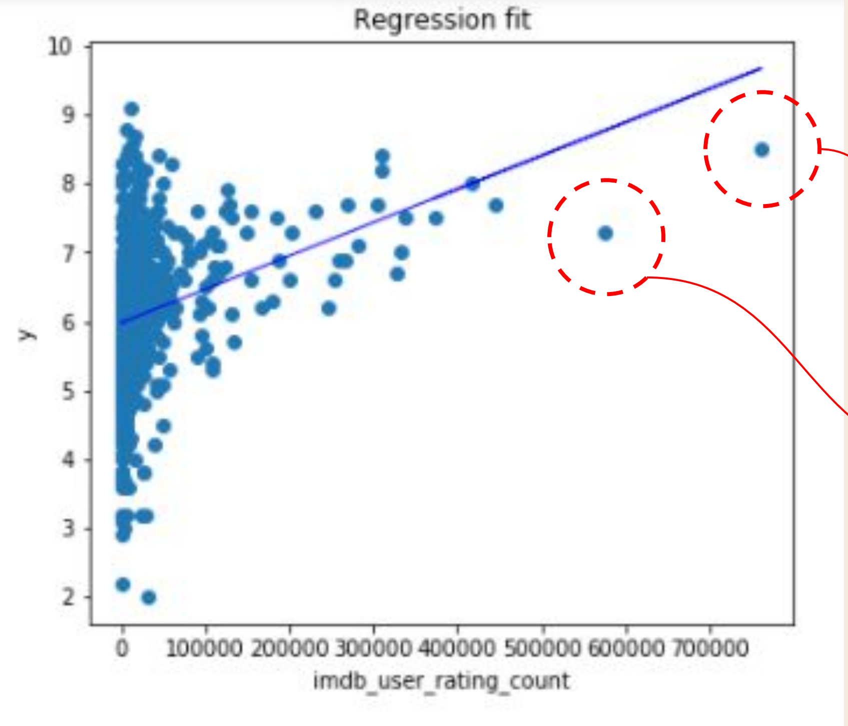
NO CORRELATION BETWEEN INDEPENDENT VARIABLES IN THE MODEL



	imdb_user_rating	imdb_user_rating_count_log	runtime_minutes_log	Weekend	Drama	Action	Thriller	Horror
imdb_user_rating	1	0.33	0.46	0.027	0.3	-0.084	-0.18	-0.25
imdb_user_rating_count_log	0.33	1	0.33	0.22	-0.0012	0.21	-0.0086	0.0048
runtime_minutes_log	0.46	0.33	1	-0.0029	0.21	0.16	-0.087	-0.21
Weekend	0.027	0.22	-0.0029	1	0.048	-0.048	-0.017	-0.017
Drama	0.3	-0.0012	0.21	0.048	1	-0.23	-0.22	-0.26
Action	-0.084	0.21	0.16	-0.048	-0.23	1	0.01	-0.086
Thriller	-0.18	-0.0086	-0.087	-0.017	-0.22	0.01	1	0.18
Horror	-0.25	0.0048	-0.21	-0.017	-0.26	-0.086	0.18	1

APPENDIX

REMOVED OUTLIERS: > 500,000 REVIEWS?

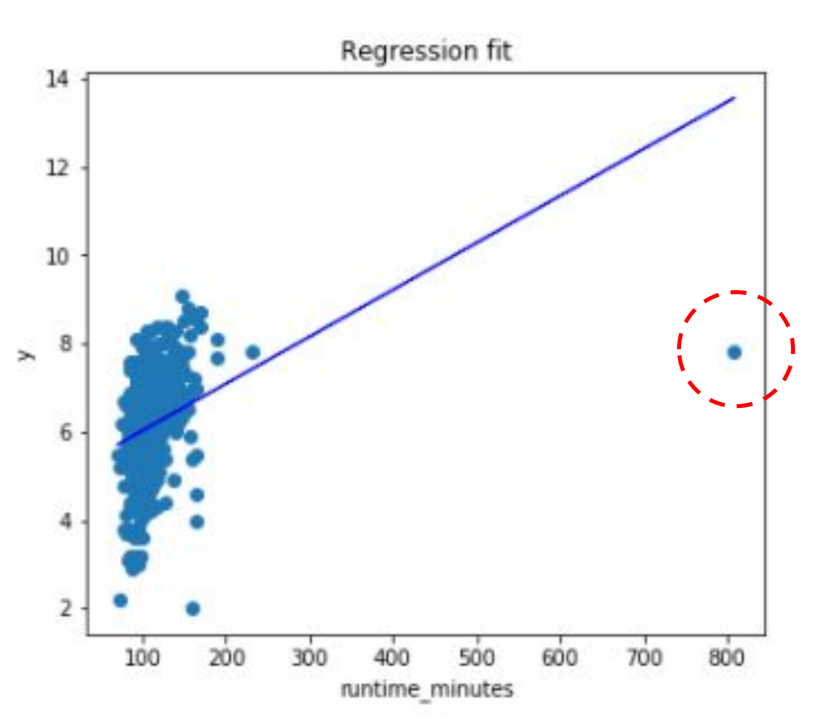


Median of Number of Reviews: 2,645

- **Avengers: Infinity War**
  Number of reviews: 761,632
  Box Office: 2.05 billion USD
  Awards: 44 wins & 72 nominations

- **Black Panther**
  Number of reviews: 573,738
  Box Office: 1.34 billion USD
  Awards: 112 wins & 265 nominations

REMOVED OUTLIER: 13-HOUR FILM?

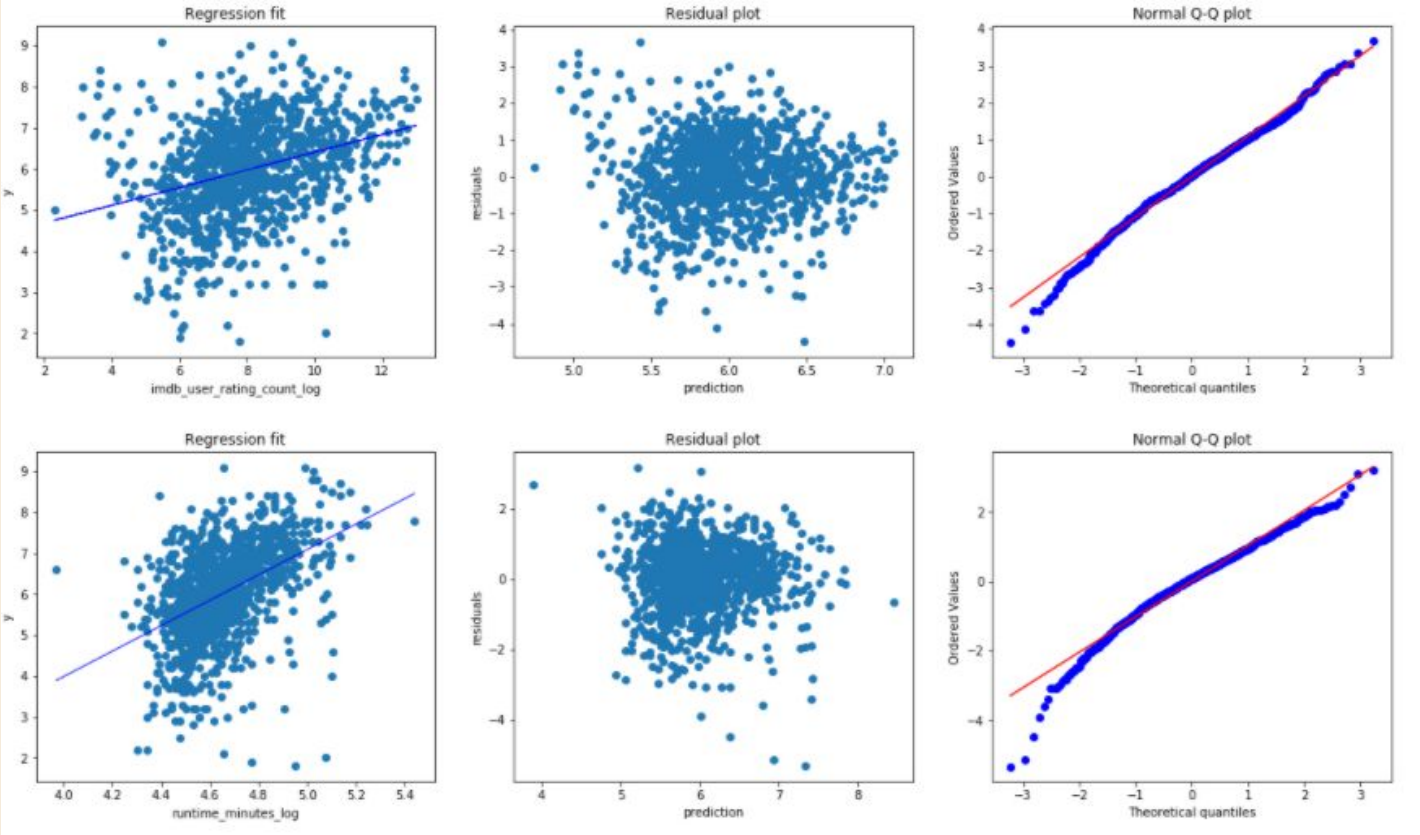


La Flor (English: The Flower) is a 2018 Argentine film written and directed by Mariano Llinás. With a length of 808 minutes excluding intermissions, it is the **longest film in the history of Argentine cinema**.

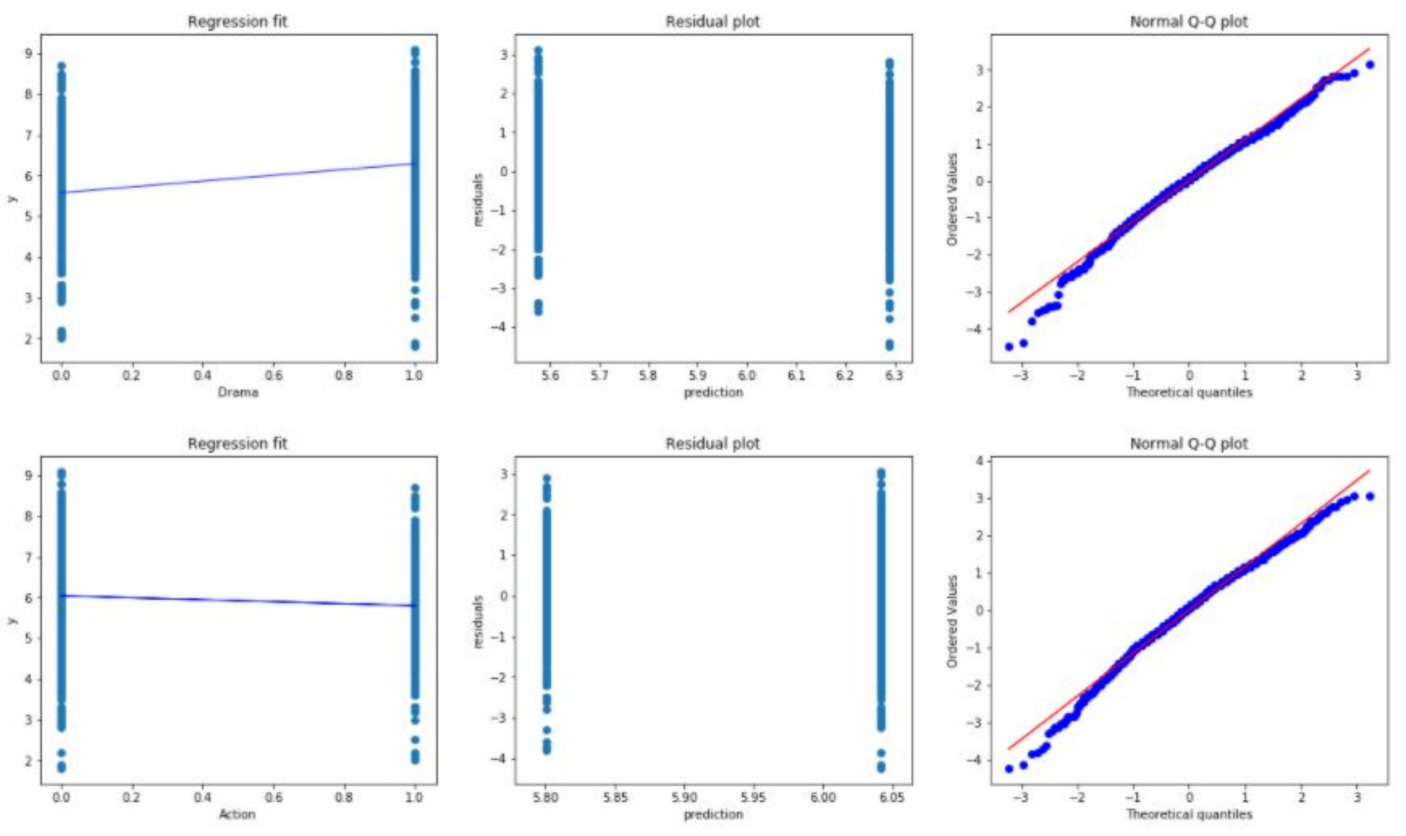
The median of movie runtime in this dataset is 103 minutes. La Flor is 21 standard deviations from the mean.

APPENDIX

REGRESSION FIT, RESIDUAL PLOT, Q-Q PLOT

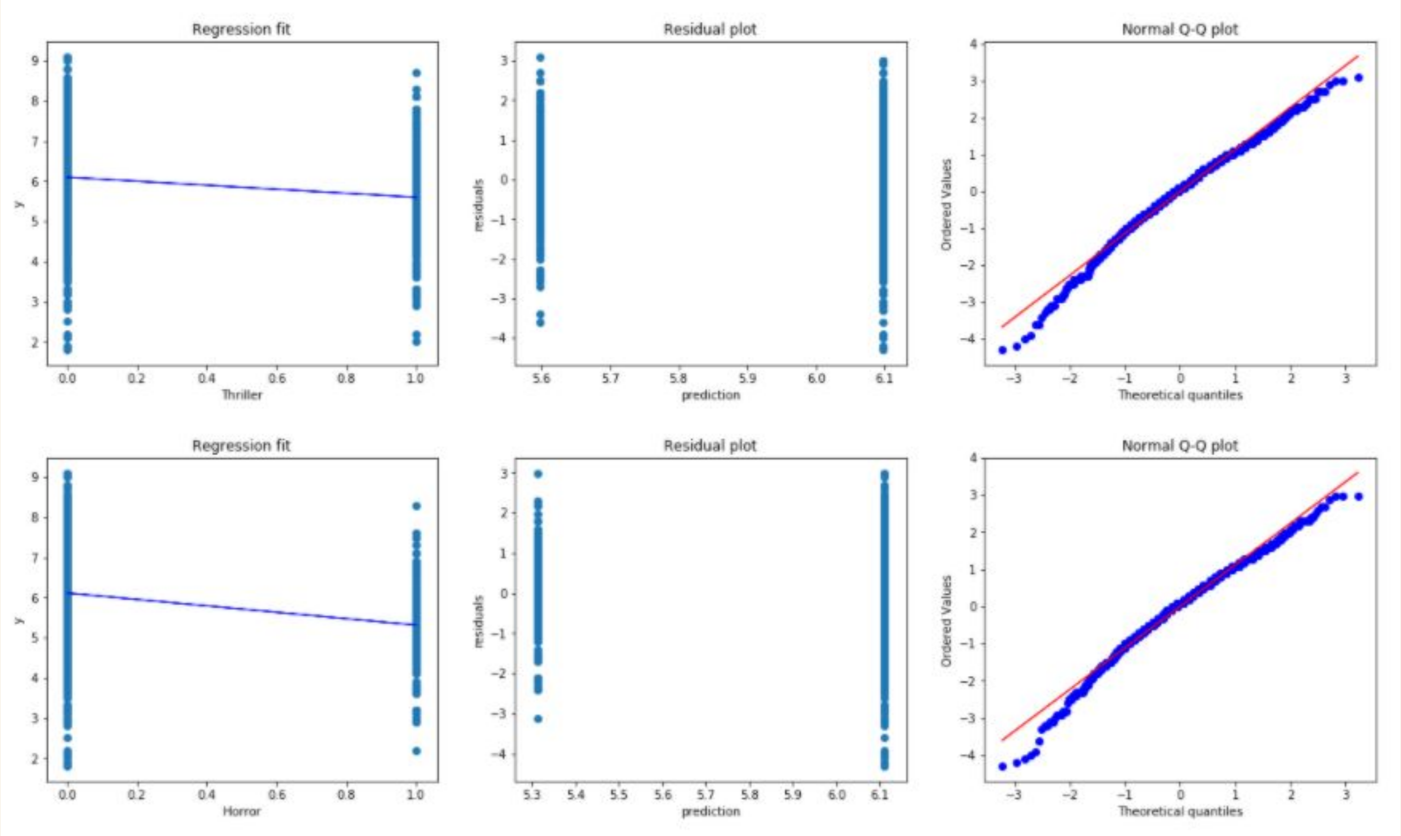


REGRESSION FIT, RESIDUAL PLOT, Q-Q PLOT

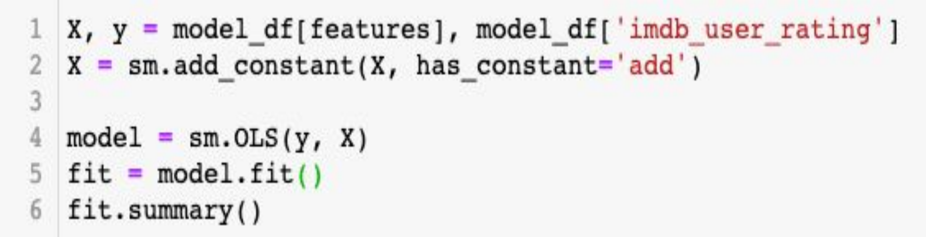


APPENDIX

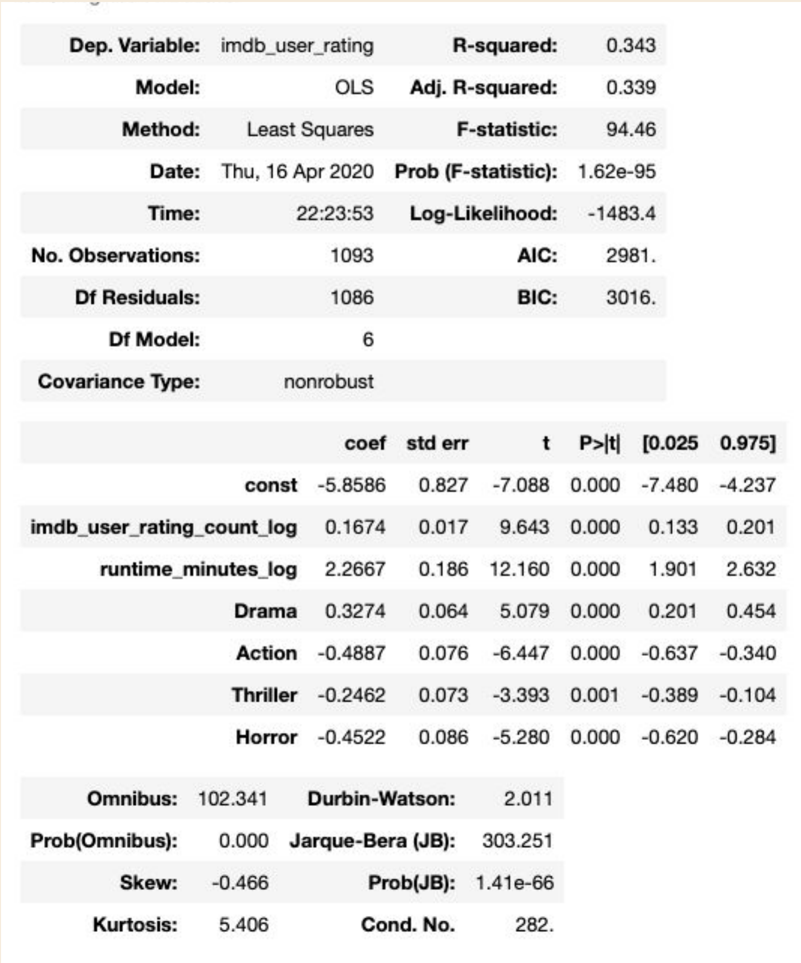
REGRESSION FIT, RESIDUAL PLOT, Q-Q PLOT



MODEL SUMMARY STATISTICS (NO SCALER)



```python
X, y = model_df[features], model_df['imdb_user_rating']
X = sm.add_constant(X, has_constant='add')

model = sm.OLS(y, X)
fit = model.fit()
fit.summary()
```



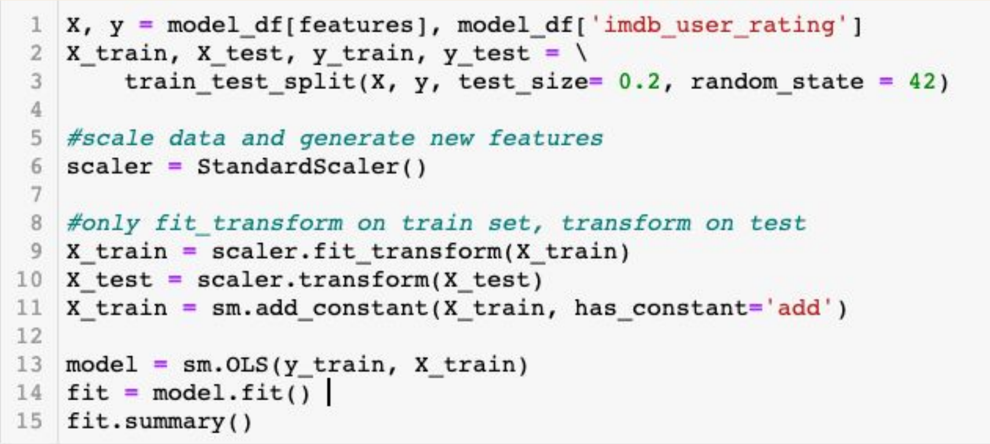
Dep. Variable:	imdb_user_rating	**R-squared:**	0.343
Model:	OLS	**Adj. R-squared:**	0.339
Method:	Least Squares	**F-statistic:**	94.46
Date:	Thu, 16 Apr 2020	**Prob (F-statistic):**	1.62e-95
Time:	22:23:53	**Log-Likelihood:**	-1483.4
No. Observations:	1093	**AIC:**	2981.
Df Residuals:	1086	**BIC:**	3016.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-5.8586	0.827	-7.088	0.000	-7.480	-4.237
imdb_user_rating_count_log	0.1674	0.017	9.643	0.000	0.133	0.201
runtime_minutes_log	2.2667	0.186	12.160	0.000	1.901	2.632
Drama	0.3274	0.064	5.079	0.000	0.201	0.454
Action	-0.4887	0.076	-6.447	0.000	-0.637	-0.340
Thriller	-0.2462	0.073	-3.393	0.001	-0.389	-0.104
Horror	-0.4522	0.086	-5.280	0.000	-0.620	-0.284

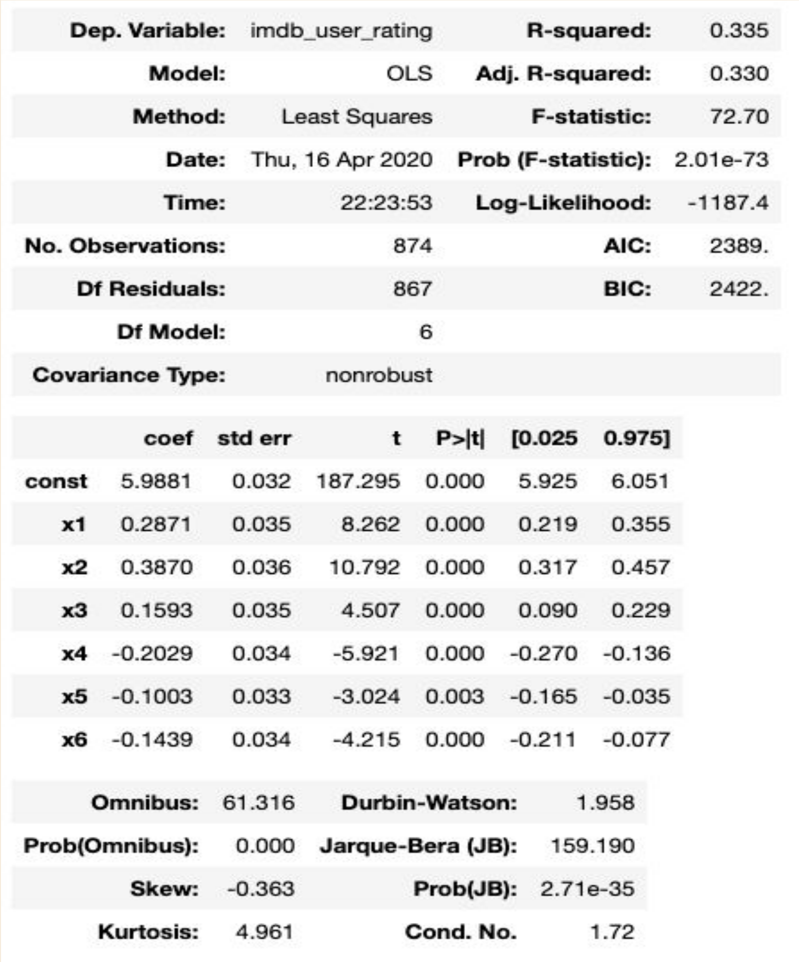
Omnibus:	102.341	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	303.251
Skew:	-0.466	**Prob(JB):**	1.41e-66
Kurtosis:	5.406	**Cond. No.**	282.

APPENDIX

MODEL SUMMARY STATISTICS (SCALER)

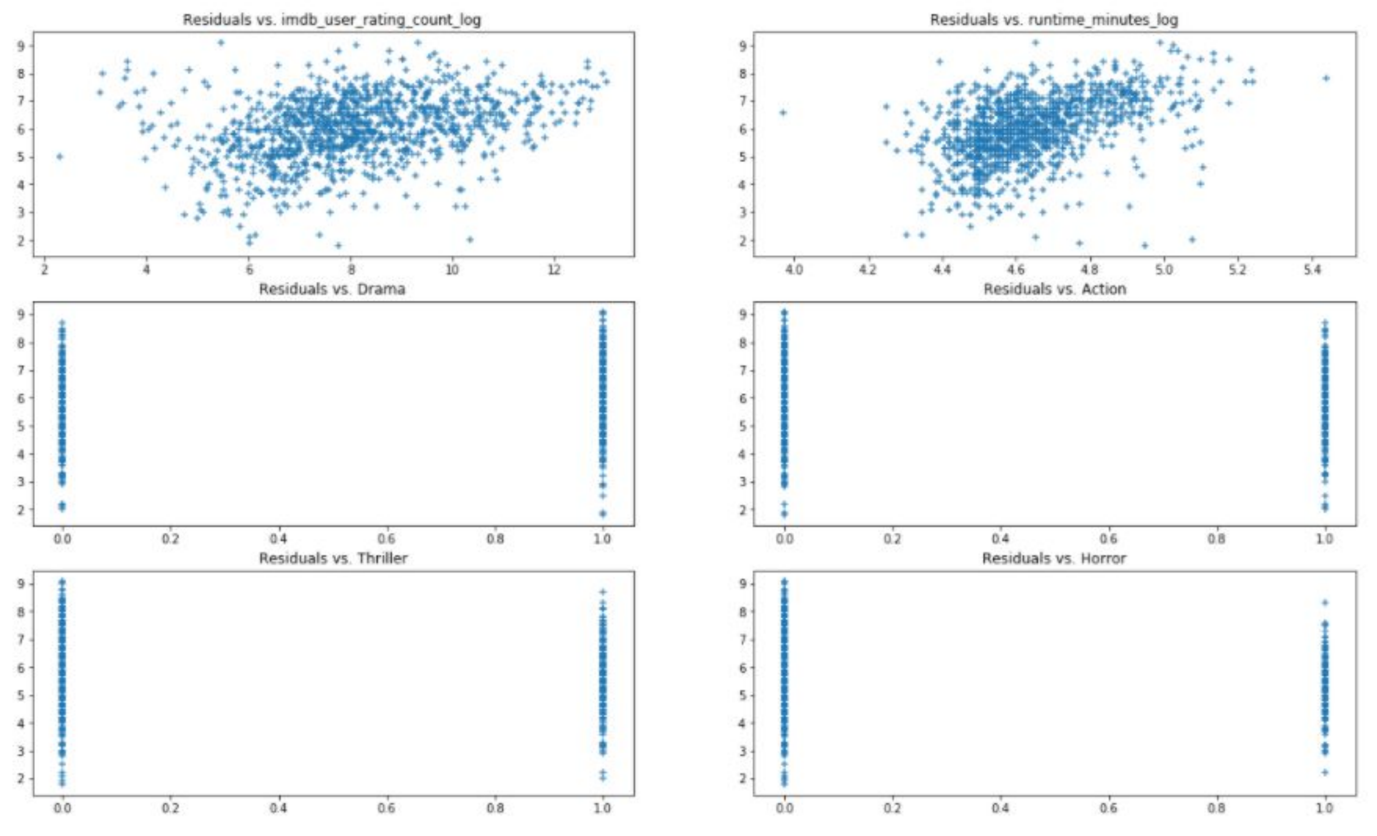


```python
X, y = model_df[features], model_df['imdb_user_rating']
X_train, X_test, y_train, y_test = \
    train_test_split(X, y, test_size= 0.2, random_state = 42)

#scale data and generate new features
scaler = StandardScaler()

#only fit_transform on train set, transform on test
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
X_train = sm.add_constant(X_train, has_constant='add')

model = sm.OLS(y_train, X_train)
fit = model.fit() |
fit.summary()
```



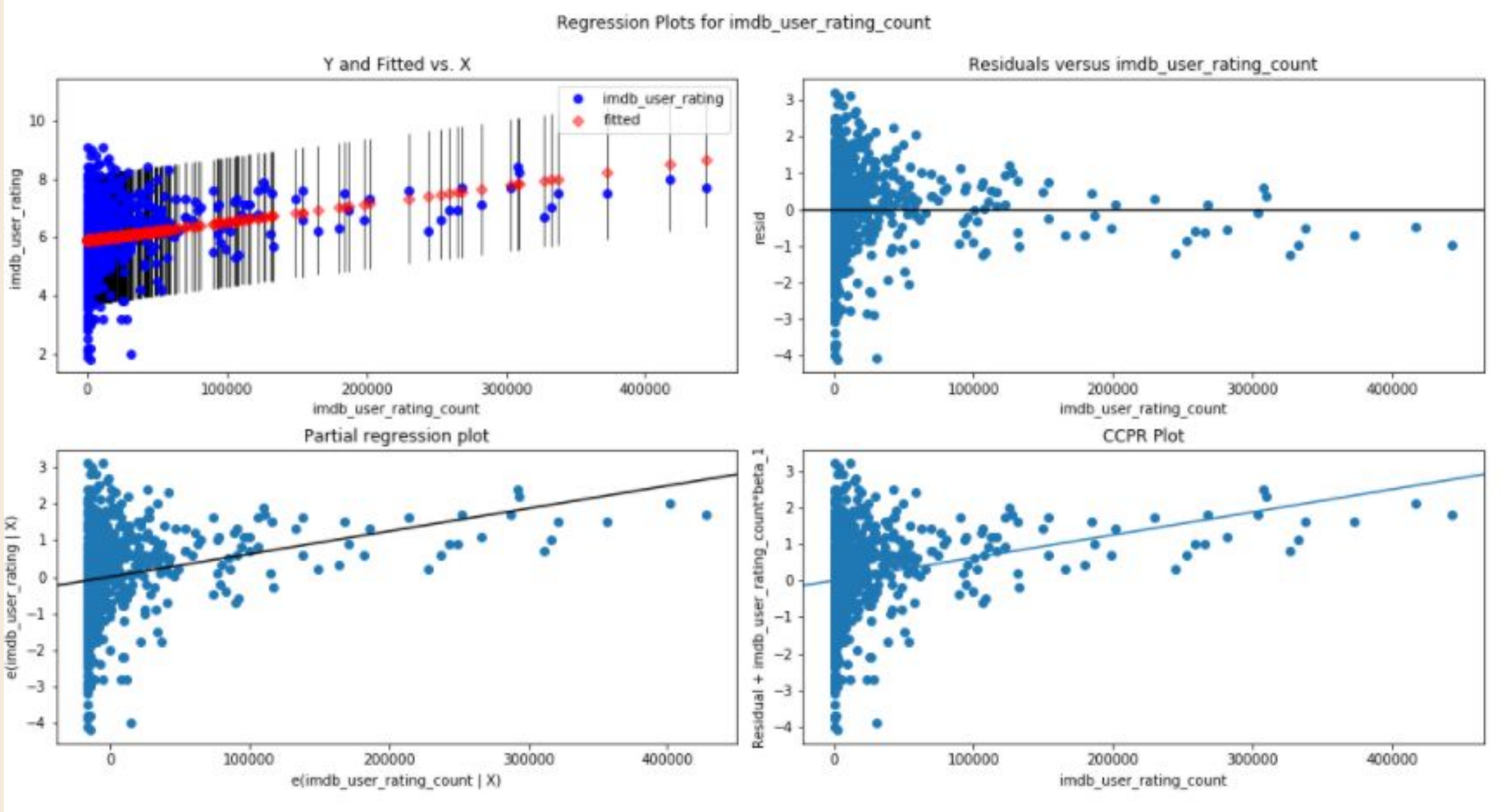
```
      Dep. Variable:  imdb_user_rating          R-squared:     0.335
             Model:               OLS      Adj. R-squared:     0.330
            Method:     Least Squares         F-statistic:     72.70
              Date:  Thu, 16 Apr 2020  Prob (F-statistic):  2.01e-73
              Time:          22:23:53      Log-Likelihood:   -1187.4
 No. Observations:               874                 AIC:     2389.
      Df Residuals:              867                 BIC:     2422.
          Df Model:                6
  Covariance Type:         nonrobust

          coef   std err        t   P>|t|   [0.025   0.975]
const   5.9881     0.032  187.295   0.000    5.925    6.051
   x1   0.2871     0.035    8.262   0.000    0.219    0.355
   x2   0.3870     0.036   10.792   0.000    0.317    0.457
   x3   0.1593     0.035    4.507   0.000    0.090    0.229
   x4  -0.2029     0.034   -5.921   0.000   -0.270   -0.136
   x5  -0.1003     0.033   -3.024   0.003   -0.165   -0.035
   x6  -0.1439     0.034   -4.215   0.000   -0.211   -0.077

       Omnibus:  61.316     Durbin-Watson:     1.958
Prob(Omnibus):    0.000  Jarque-Bera (JB):   159.190
          Skew:  -0.363          Prob(JB):  2.71e-35
      Kurtosis:   4.961         Cond. No.       1.72
```

RESIDUAL PLOTS FOR INDEPENDENT VARIABLES



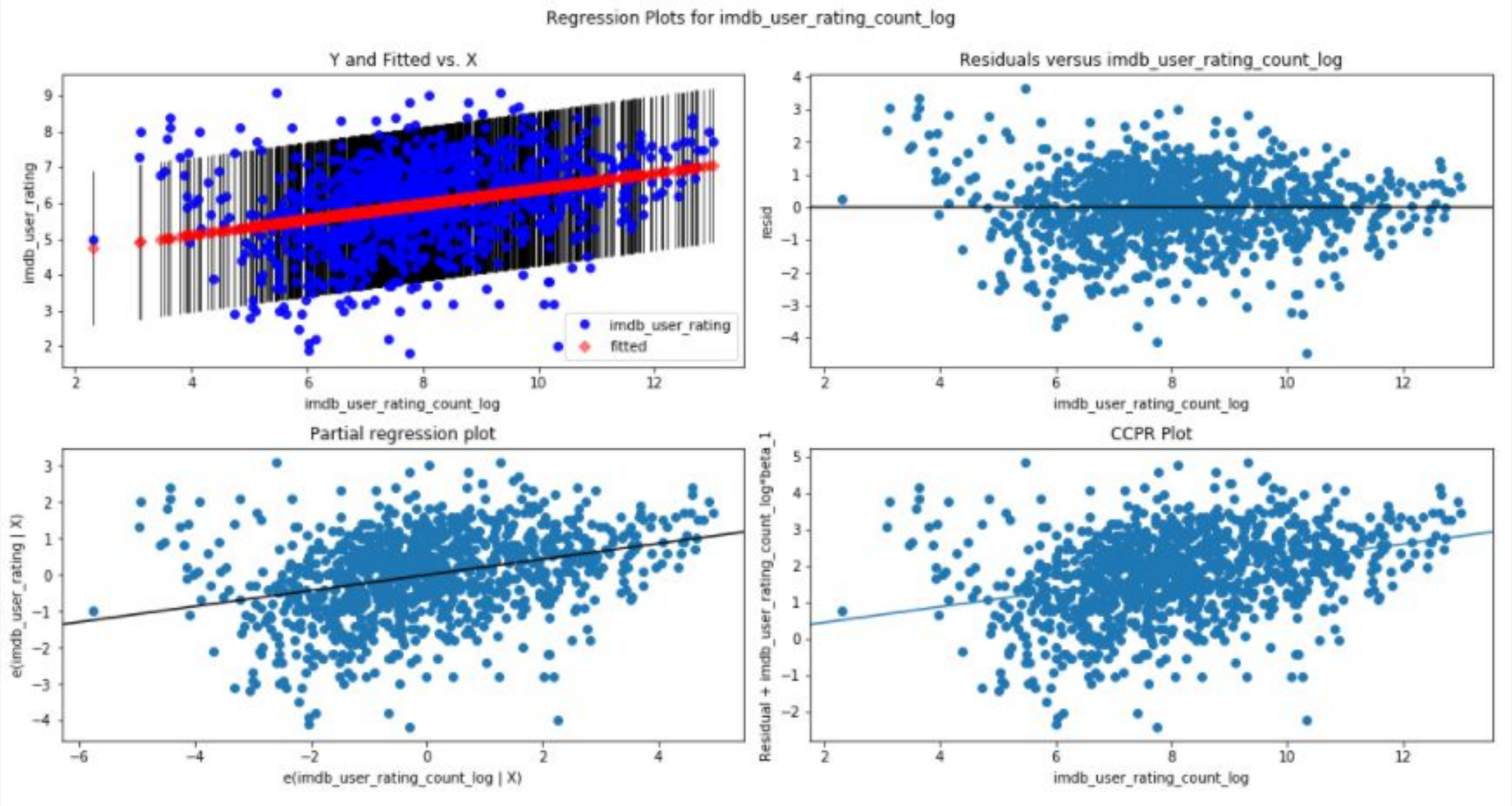
APPENDIX

IMDB_USER_RATING_COUNT: WITHOUT LOG TRANSFORMATION

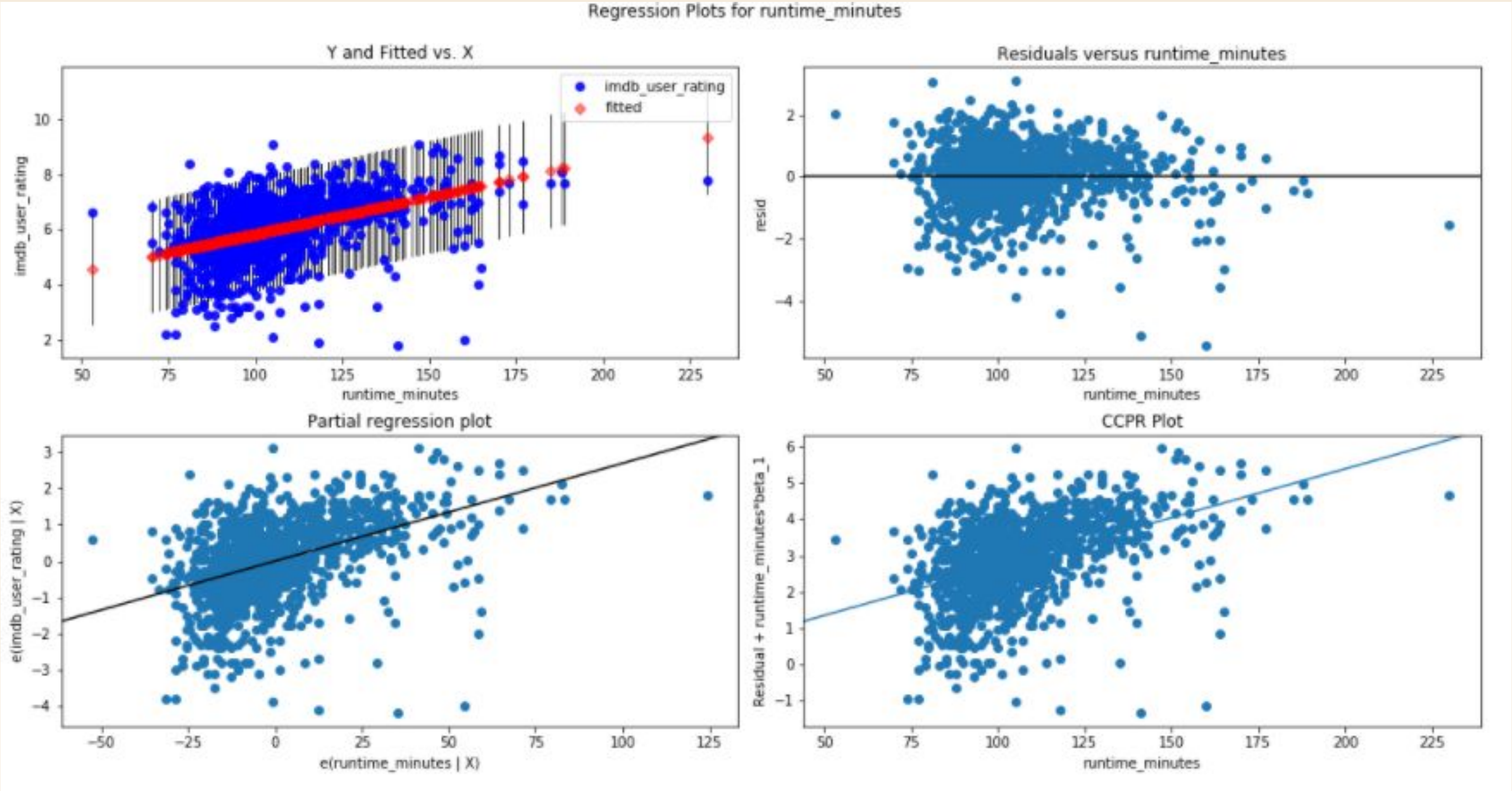


APPENDIX

IMDB_USER_RATING_COUNT: WITH LOG TRANSFORMATION

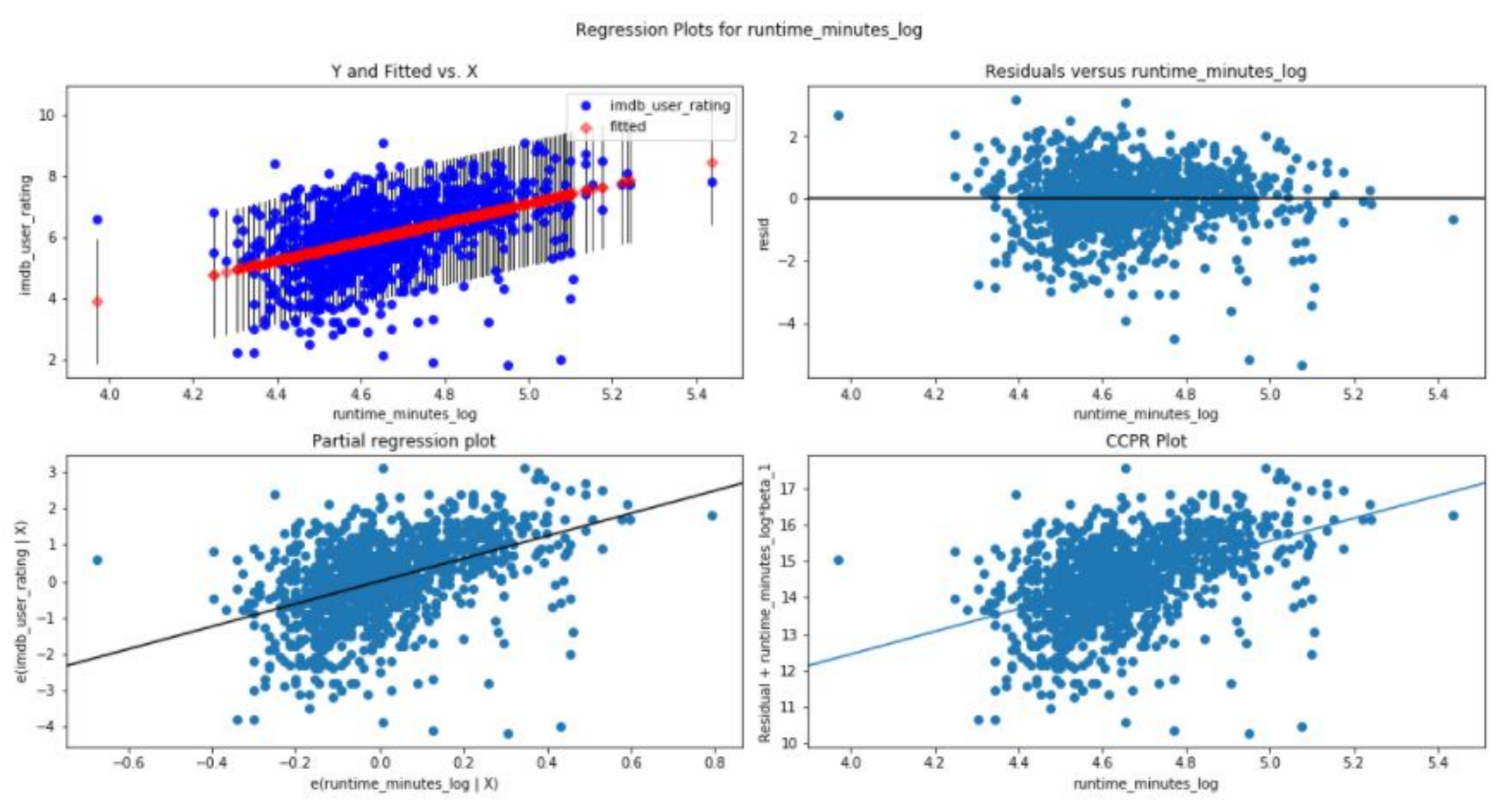


RUNTIME_MINUTES: WITHOUT LOG TRANSFORMATION



APPENDIX

RUNTIME_MINUTES: WITH LOG TRANSFORMATION



DETAILED RESULTS

- For a 1% increase in number of IMDb reviews, the IMDb user rating changes by 0.09%
- For a 1% increase in movie runtime, the IMDb user rating changes by 14.12%
- If this movie includes "Drama" as one of the genres, the IMDb user rating changes by 94.75%
- If this movie includes "Action" as one of the genres, the IMDb user rating changes by -70.28%
- If this movie includes "Thriller" as one of the genres, the IMDb user rating changes by -45.6%
- If this movie includes "Horror" as one of the genres, the IMDb user rating changes by -72.17%

This suggests that try not to have your movie runtime too short, ensure that your film has the "Drama" factor, and lastly avoid focusing too much elements in "Action", "Horror", or "Thriller".

APPENDIX

DETAILED RESULTS

- Train-Test-Split R^2 score (test score): 0.3726
- Train-Test-Split R^2 score (train score): 0.3347
- Mean Absolute Error: 0.6884
- Mean Squared Error: 0.8759
- Root Mean Squared Error: 0.9359

Feature coefficient results:

+ imdb_user_rating_count_log : 0.16
+ runtime_minutes_log : 2.27
+ Drama : 0.32
+ Action : -0.50
+ Thriller : -0.25
+ Horror : -0.41