Movie Popularity Analysis

Aparna R. Joshi 14CO204
Isha Tarte 14CO217

Problem Statement

To investigate the extent to which a movie's average rating and thus, its popularity, can be predicted after learning the relationship between the rating and a movie's various attributes from a training set.

The entire data about any movie can be obtained from IMDB.

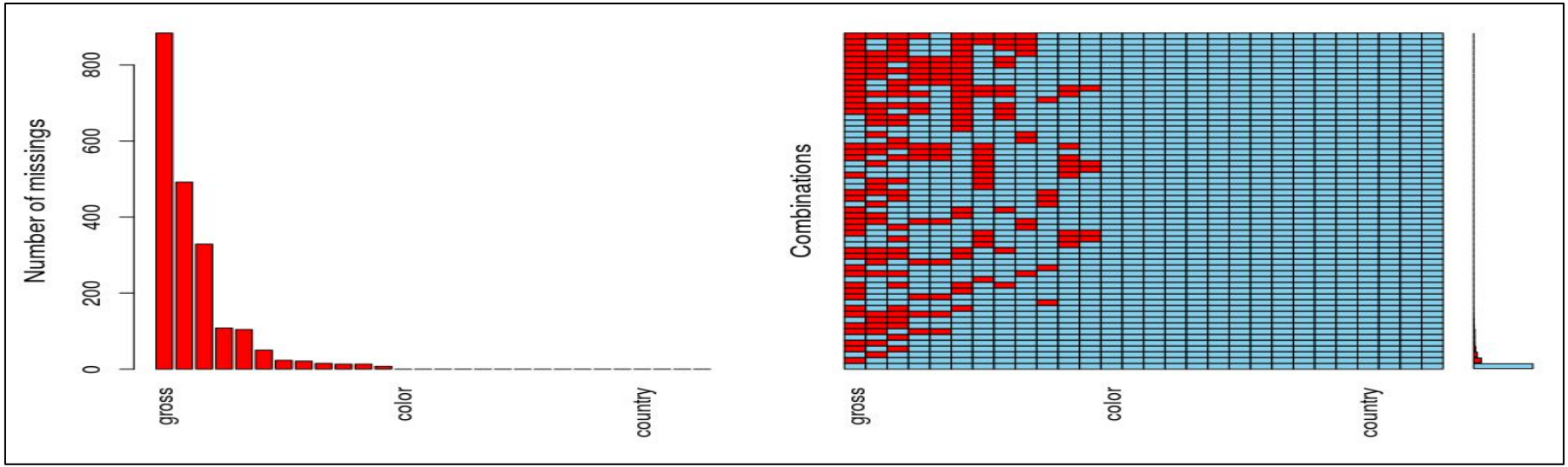
Acquiring the Dataset

We obtained the dataset by scraping the movie titles, followed by scraping the data about the movies from IMDB. For the purpose of scraping, we used the python library scrapy.

The dataset consisted information about 5000 movies and 25 attributes.

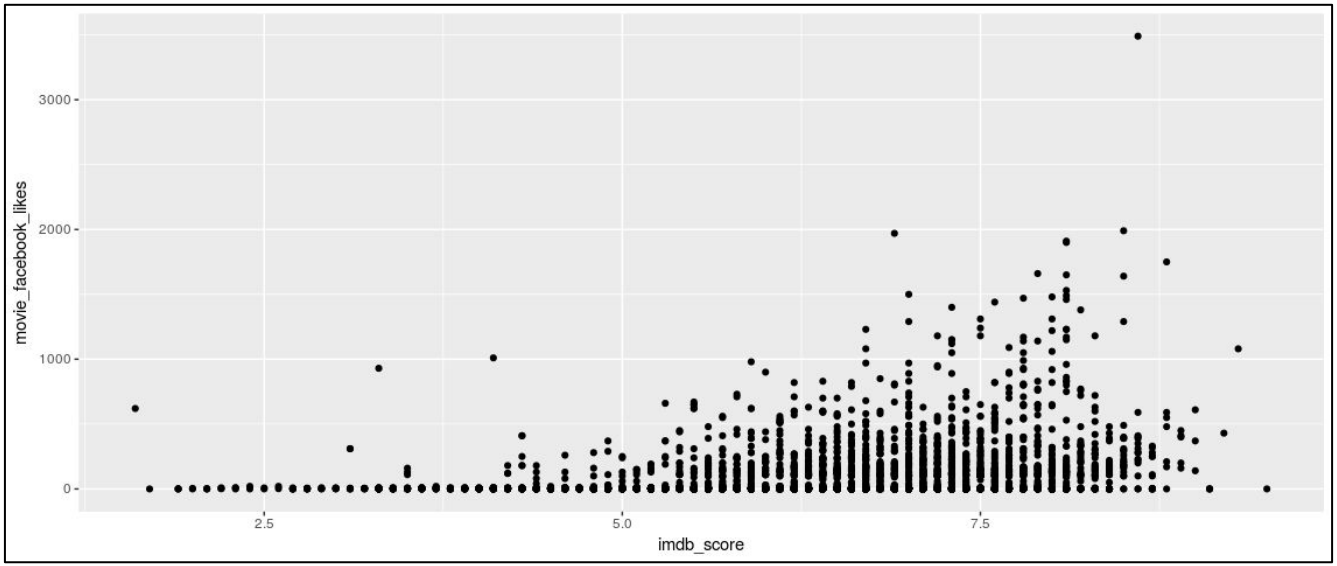
Data Analysis

We investigate the missingness of our dataset, sort the attributes in the decreasing order of number of missing values, and do not consider these as “good attributes”. “Gross” has highest missing values at 884.

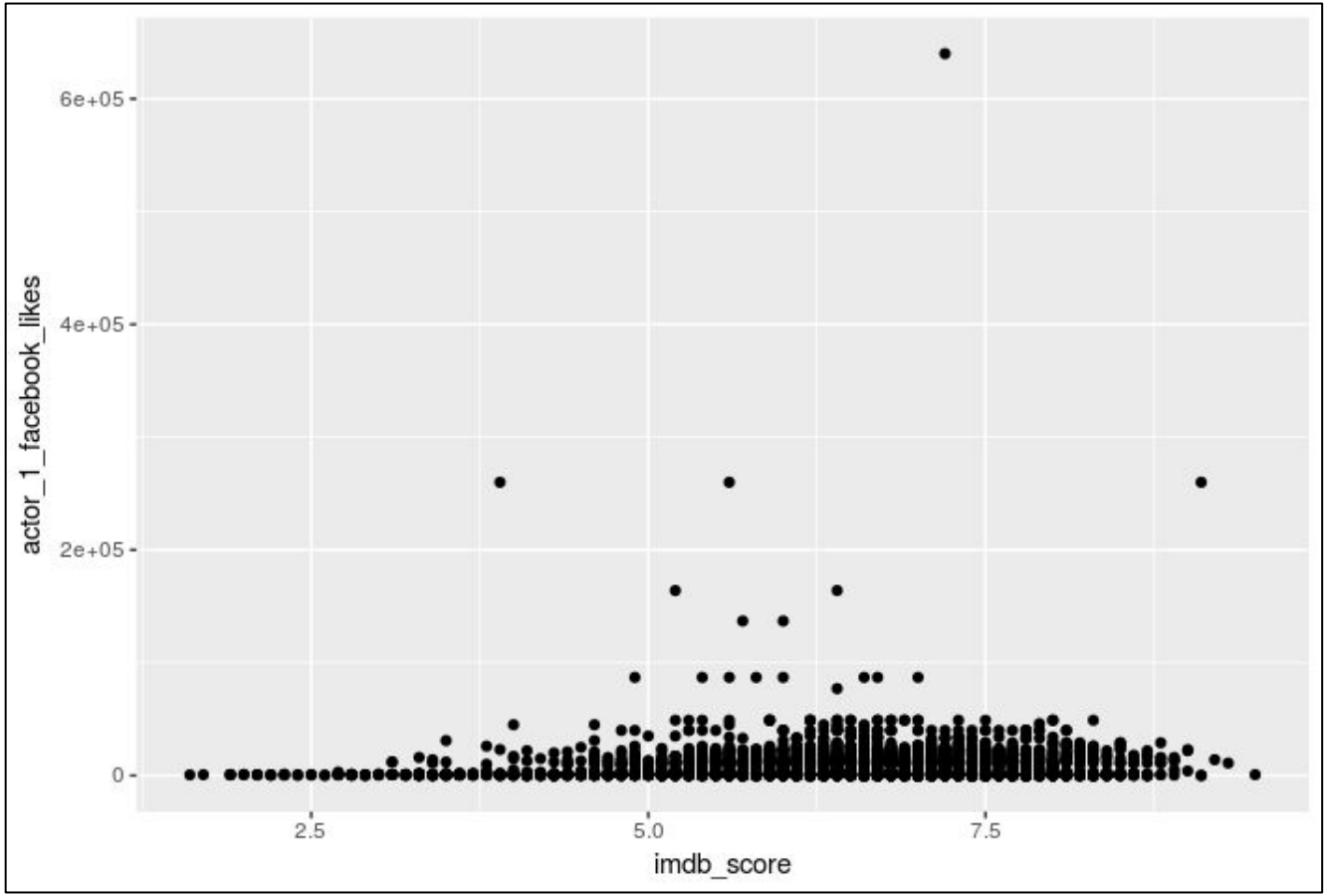


Data Analysis

The dataset consists of the attribute “IMDB score” which we train our model on. We plotted IMDB score VS other attributes and determined how IMDB score changes with change in other attributes.



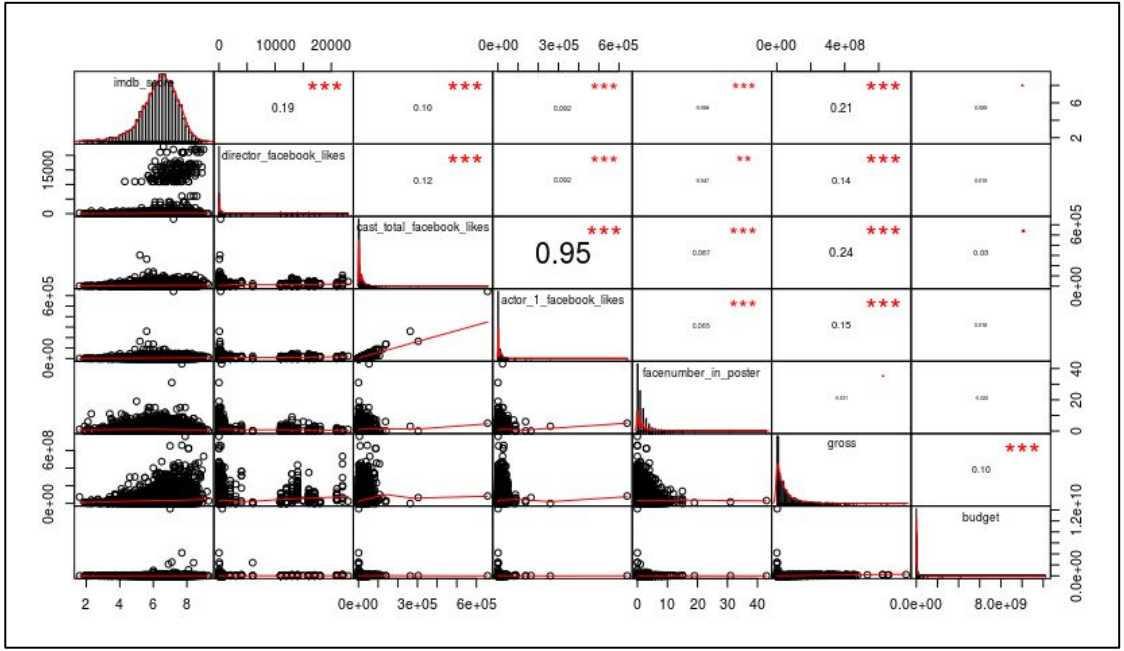
Data Analysis



Data Analysis

Further, we investigate the pairwise attribute relationships and how strongly two attributes are correlated.

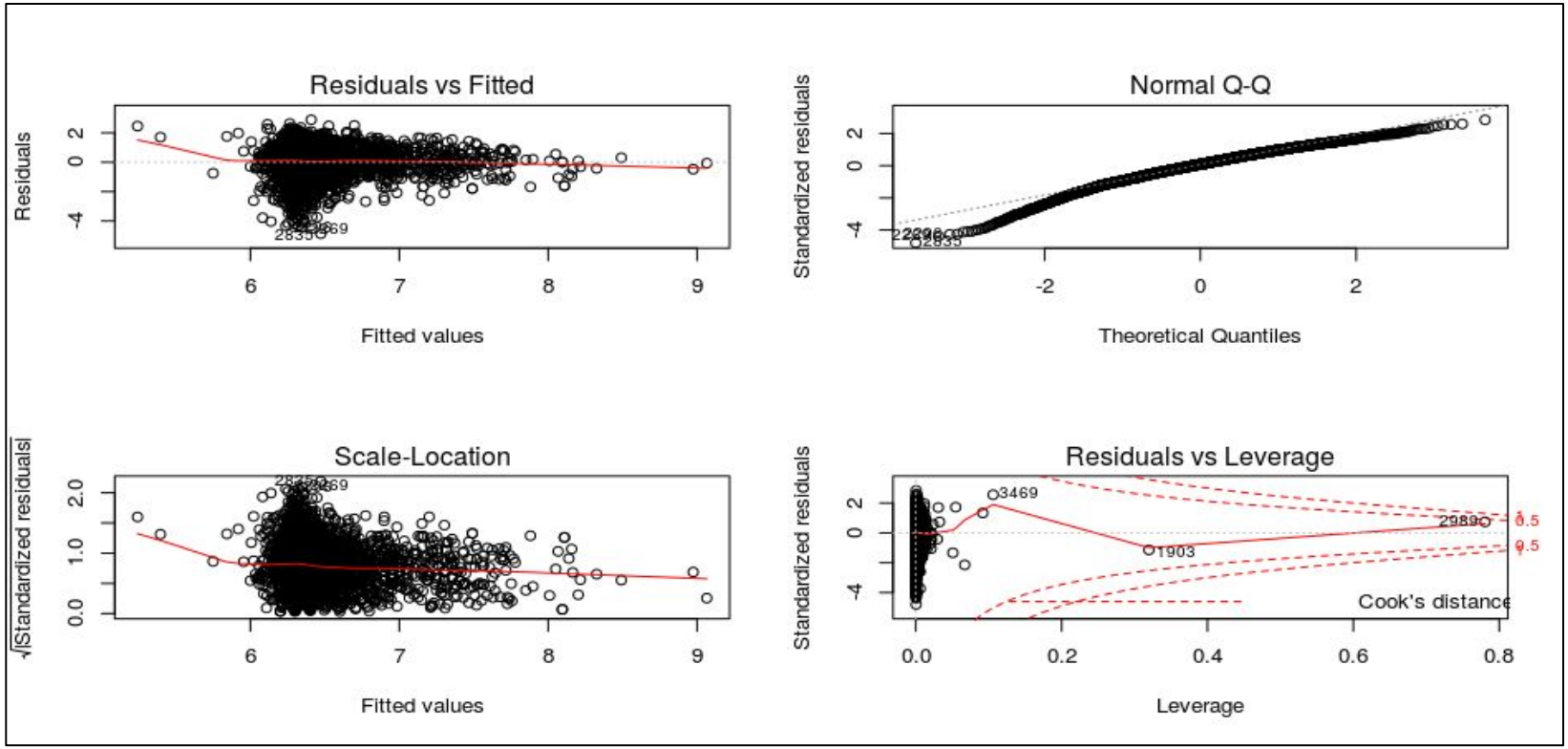
We determine that cast_total_facebook_likes is strongly correlated with actor_1_facebook_likes, and can cause multicollinearity.



Modeling our Data

We select few good attributes for linear regression. We eliminate cast_total_facebook_likes since it is highly correlated with actor_1_facebook_likes, and attributes as movie_facebook_likes and num_voted_users since these will not be available when a movie is released. Also we ignore the categorical data. The error obtained on using Linear Model is 1.016.

Linear Regression



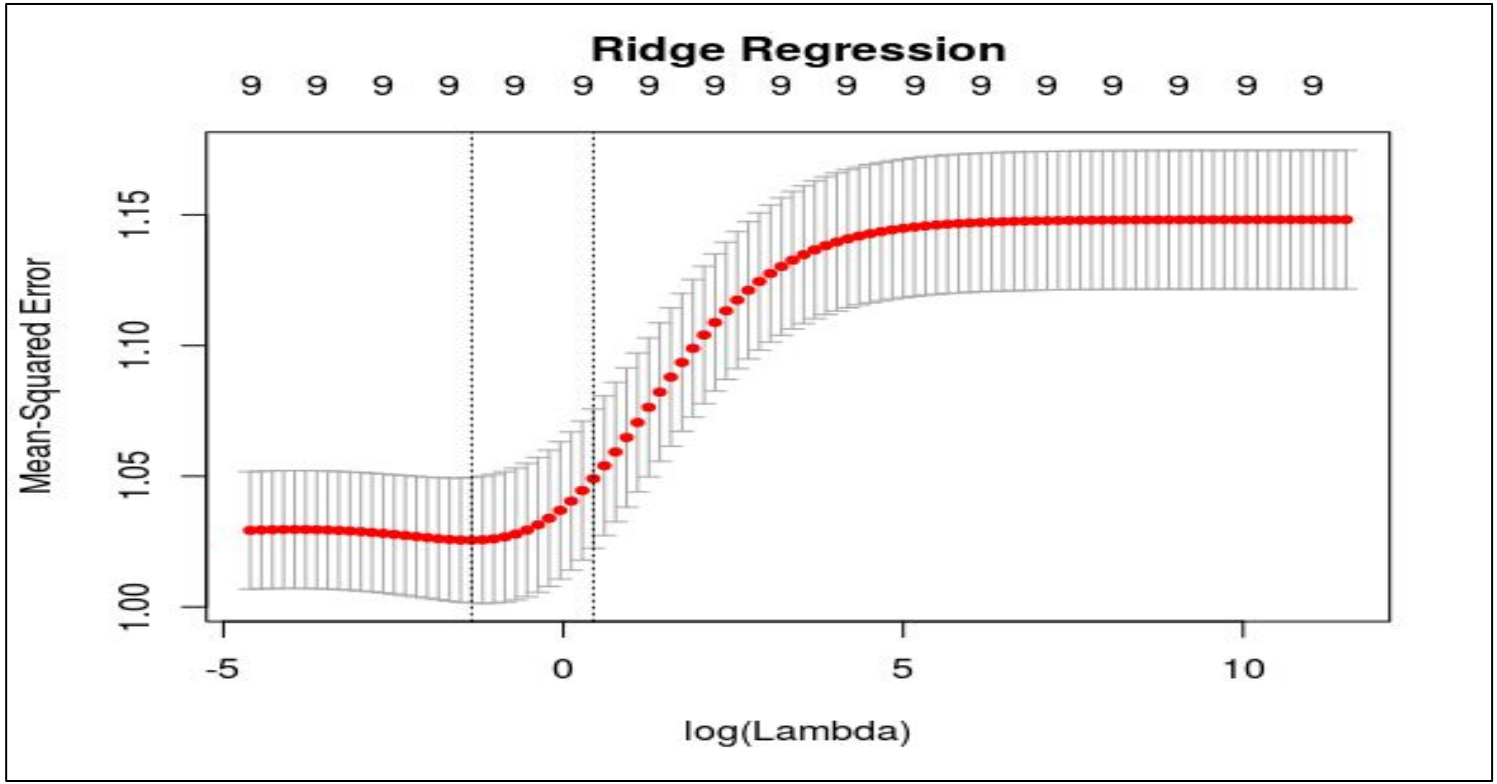
Ridge Regression

Since ridge regression can overcome multicollinearity, we choose 10 attributes for training.

The model involves automated calculation of lambda values and coefficients based on each lambda value for the test data (70:30 ratio).

We perform 10 fold cross validation to estimate the mean squared error (MSE). MSE obtained is 0.9332368. The Min_Max Accuracy obtained was 89.13%.

Ridge Regression



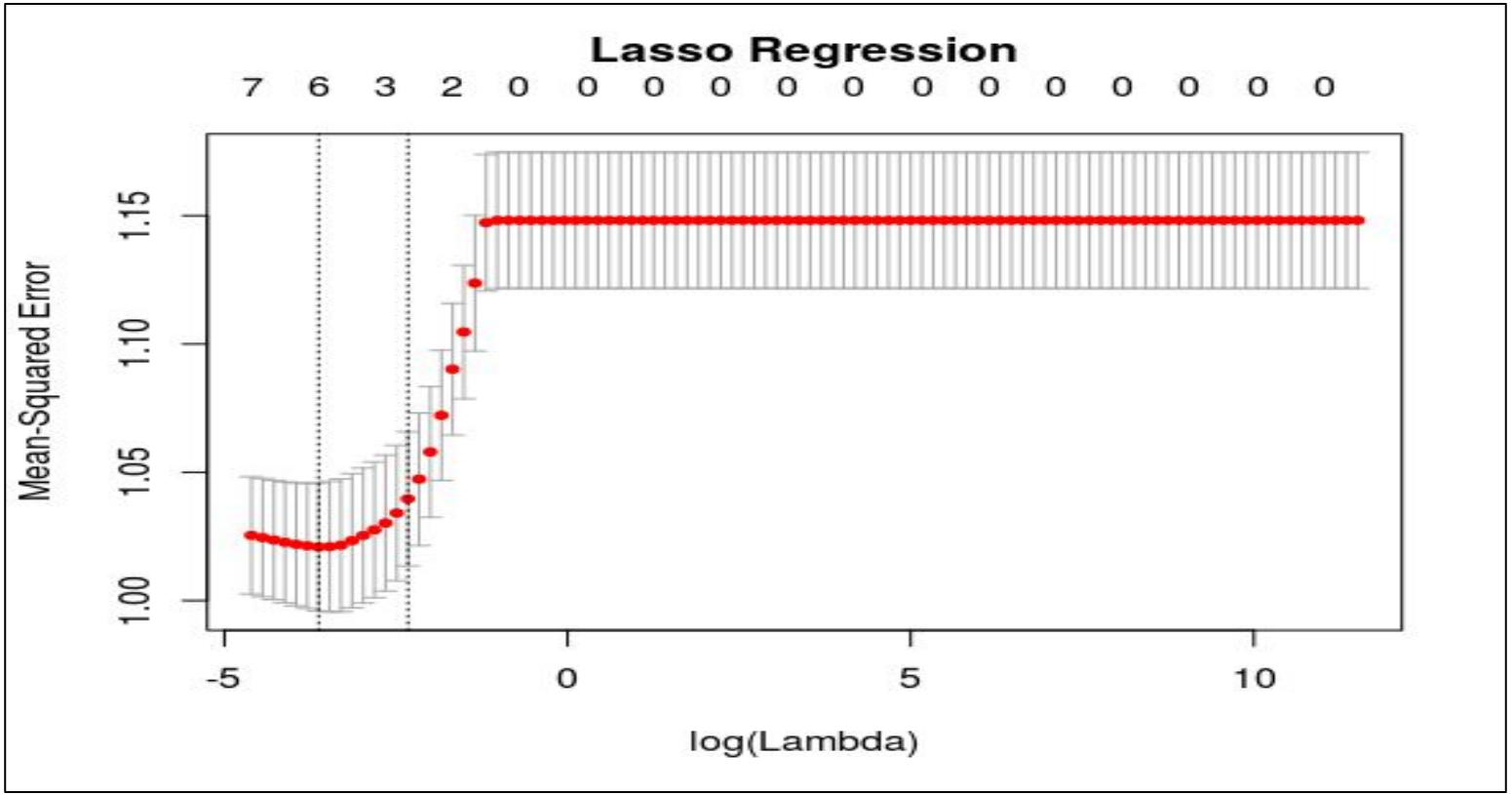
Lasso Regression

We tested our model by Lasso regression as well. This yielded an MSE of 0.9368678.

The only difference between Ridge and Lasso is the shape of the constraint imposed on the data values to keep the variance in bounds.

The Min_Max Accuracy obtained was 89.11%.

Lasso Regression



Conclusion

Obtaining an **integrated dataset** from internet sources, we were able to identify the most **influential attributes** for movie popularity analysis using **correlation**.

We divide our scraped data of about 5000 movies with **relevant features** into training and testing sets in a ratio of 70:30.

We initially performed **multivariate linear regression** on the data and obtained an error of 1.016. Following that, we performed **Ridge** and **Lasso regression** on the data which resulted in considerably lesser error.

References

- Armstrong, Nick, and Kevin Yoon. Movie rating prediction. Technical Report, Carnegie Mellon University, 2008.
- Augustine, Achal, and Manas Pathak. User rating prediction for movies. Technical report, University of Texas at Austin, 2008.