CIS 544 - FINAL PROJECT

CLASSIFYING MOVIES FROM REVENUE AND RATINGS

Ana Pacella

Background

- Data: Publicly available at Kaggle "IMDB Movies Dataset".

- Why this data/topic?

- Problem: Can we classify how good a movie is from its revenue and rating?

DATA

1

IMDB Rating

Rating of the movie at IMDB site

2

Meta Score

Score earned by the movie

2

Number of Votes

Total number of votes

3

Gross

Money earned by the movie.

4

Released Year

Year at which that movie released

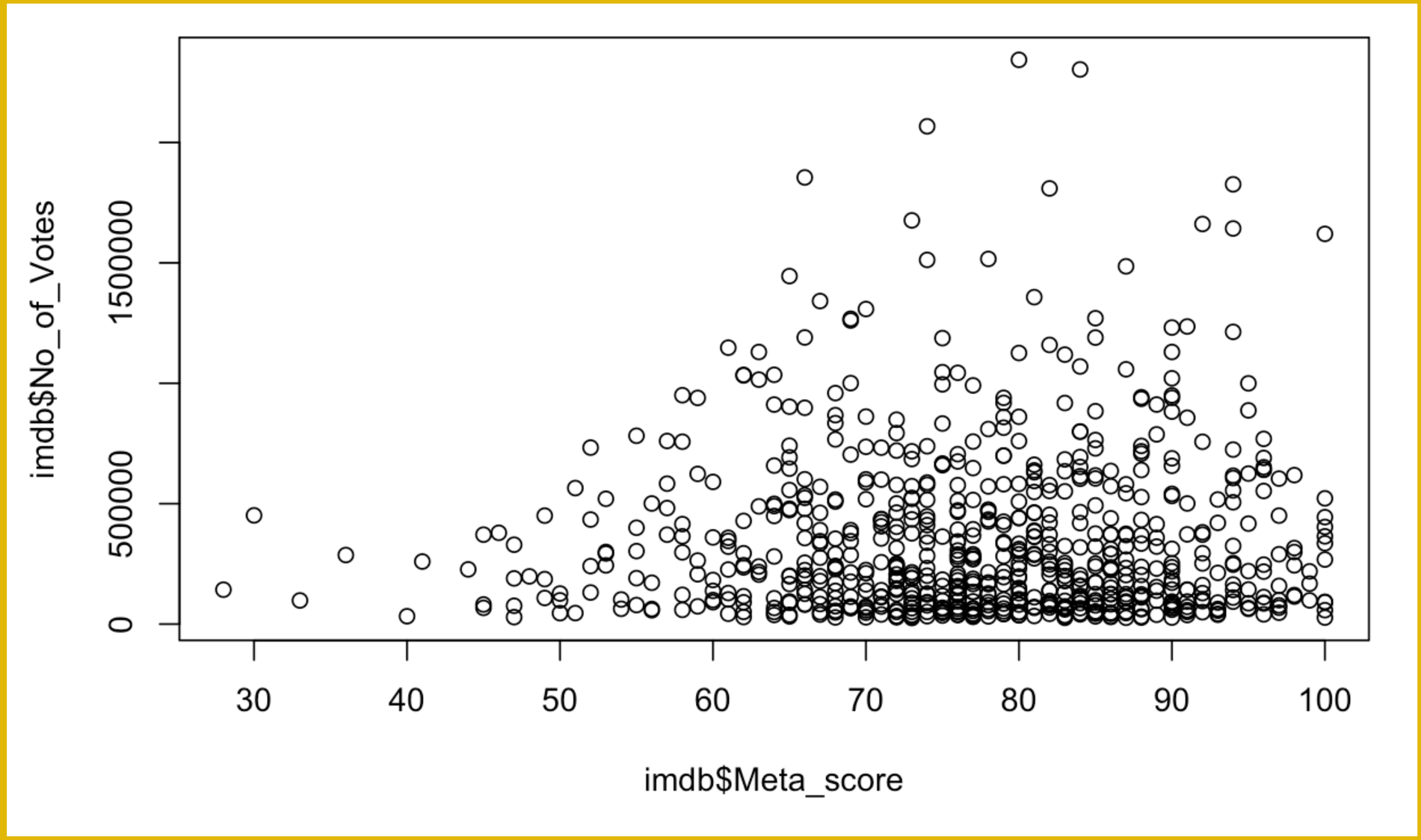
4

Run Time Min

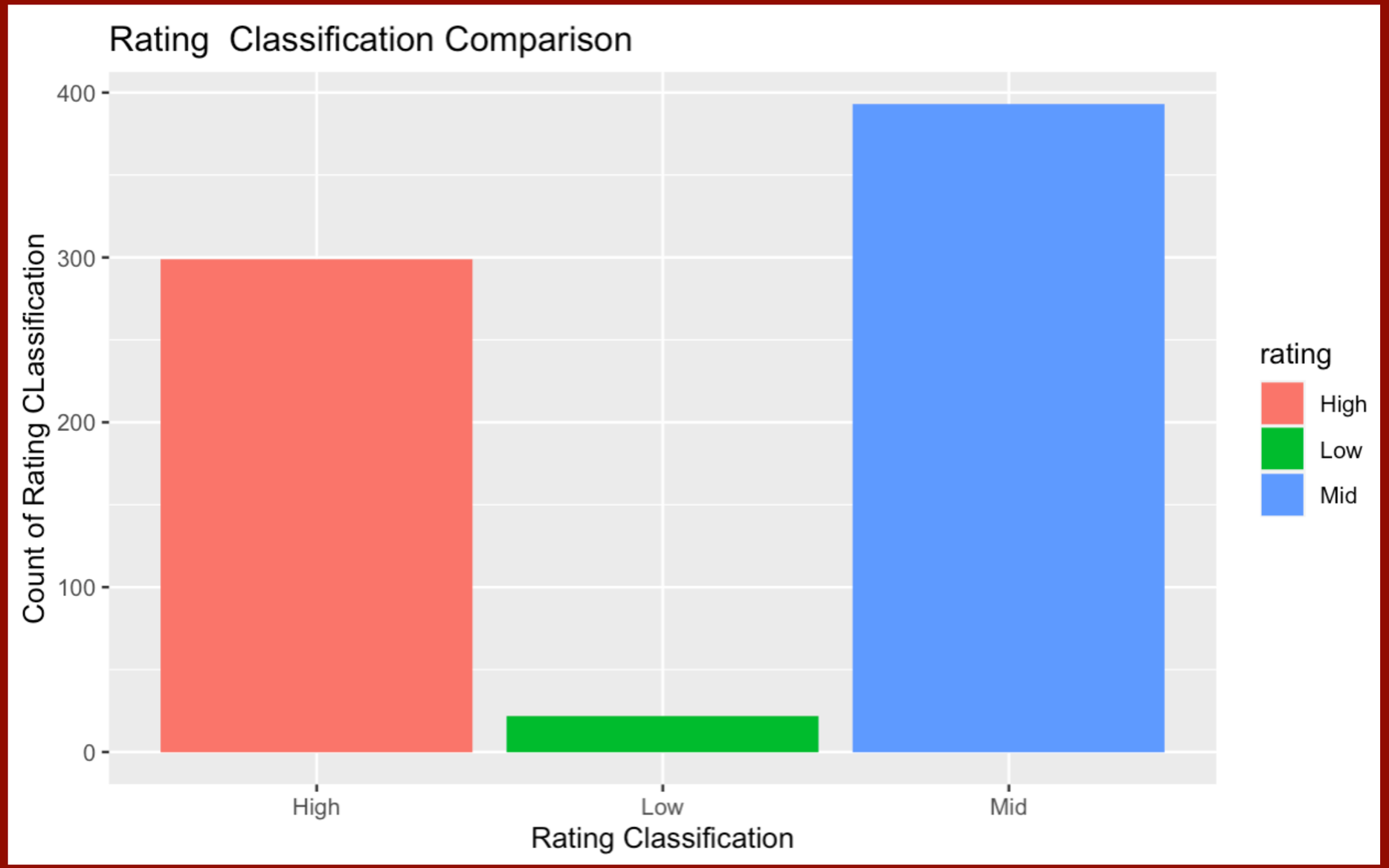
Total runtime of the movie

OTHER VARIABLES: MOVIE TITLE, GENRE, DIRECTOR, ACTORS, OVERVIEW, CLASSIFICATION...

Exploring Data
& Data Visualizations

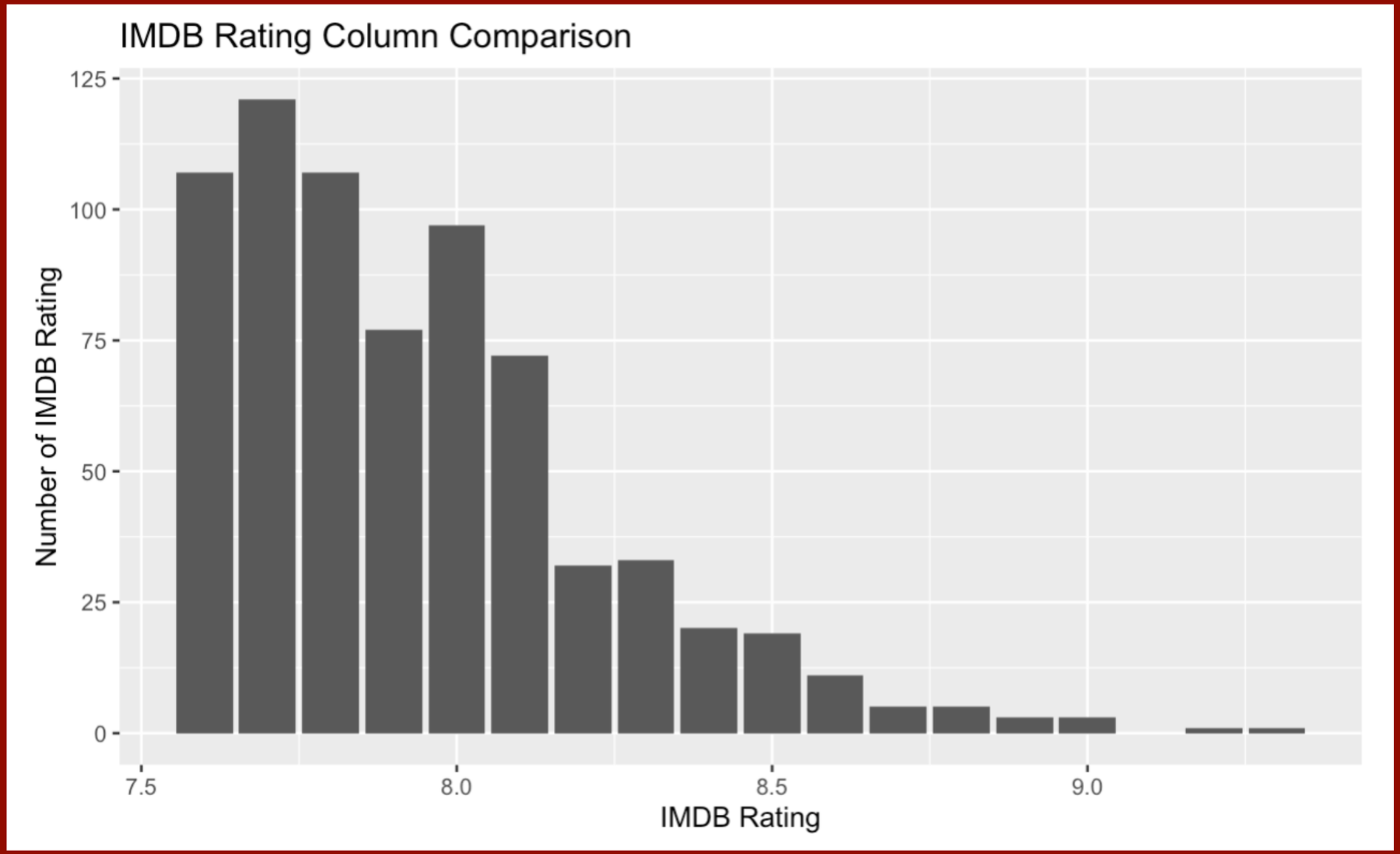


PLOTTING TWO RELEVANT VARIABLES FROM THE DATA TO SEE IF THEY HAVE ANY RELATION.



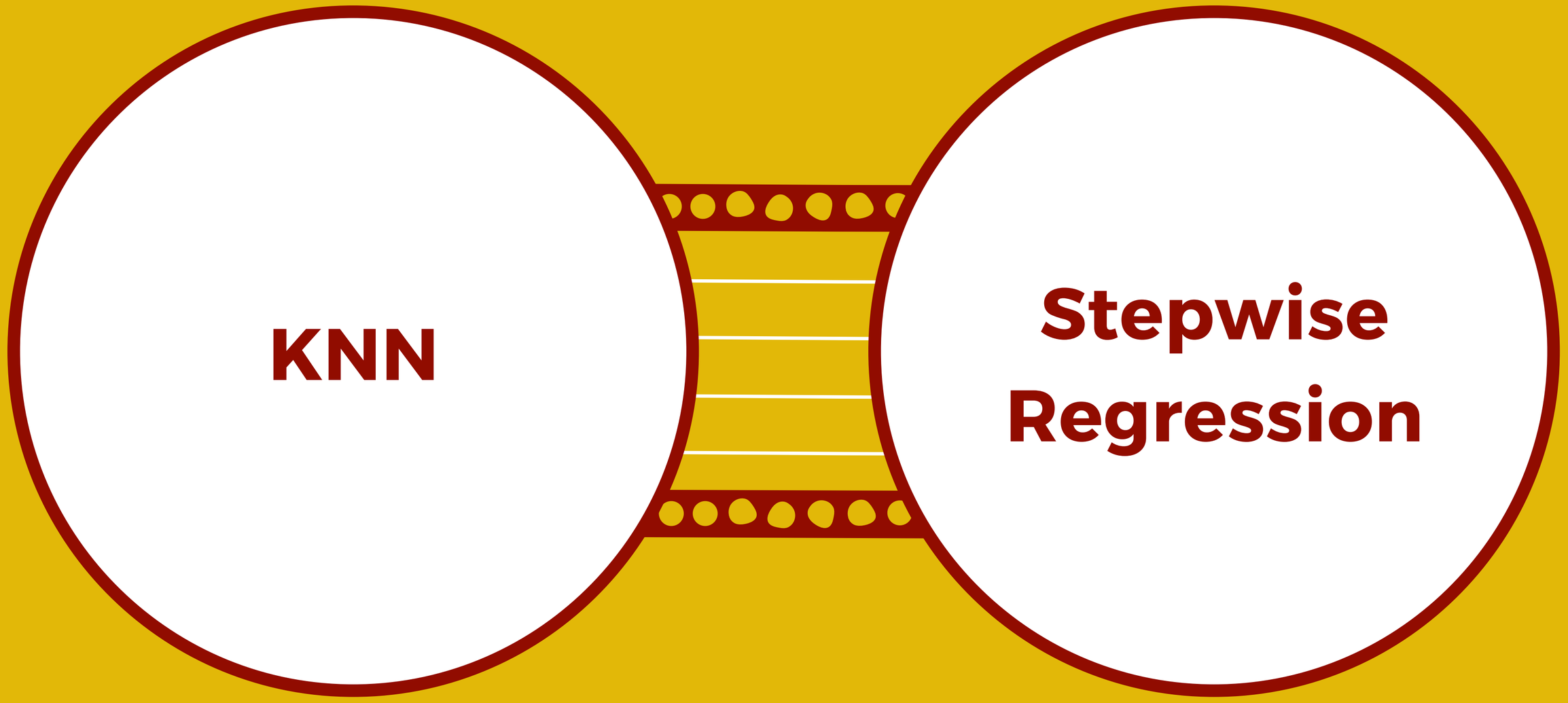
Low: <50
Mid: <80
High: all other

MOST OF THE MOVIES ARE CLASSIFIED WITH MID RATING, FOLLOWED BY HIGH RATING AND FINALLY VERY FEW WITH LOW RATING.



WE CAN SEE A COMPARISON BETWEEN THE IMDB RATING COLUMN WITH HOW MANY MOVIES HAD A SPECIFIC RATING, OR IN OTHER WORDS, HOW REPETITIVE THE RATINGS WERE.

Models Used

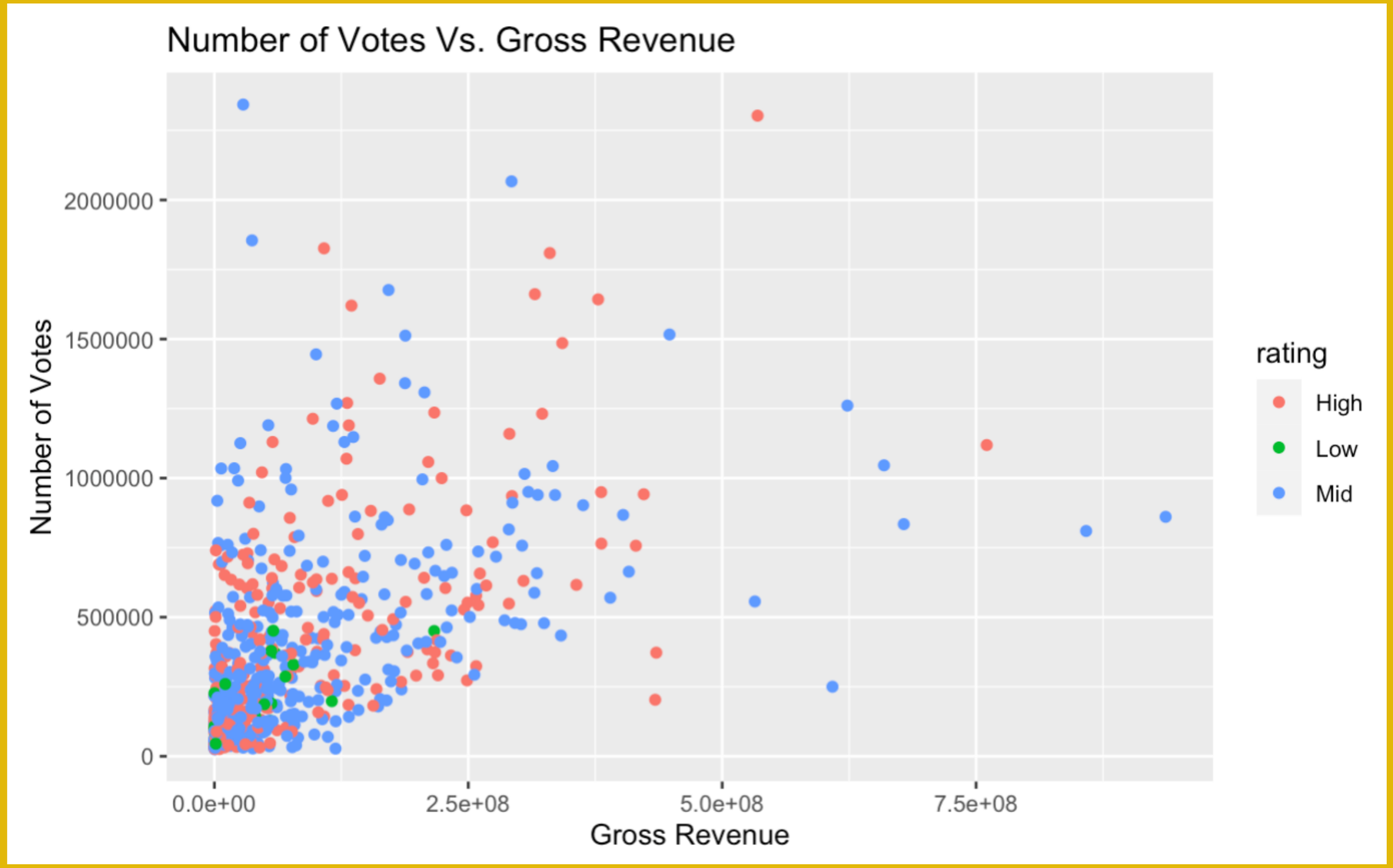


KNN

NUMBER OF VOTES VS. GROSS REVENUE

CONFUSION ACCURACY IS: 47%

We could say that there is a positive relation between the number of votes a movie got (popularity) and its gross revenue.

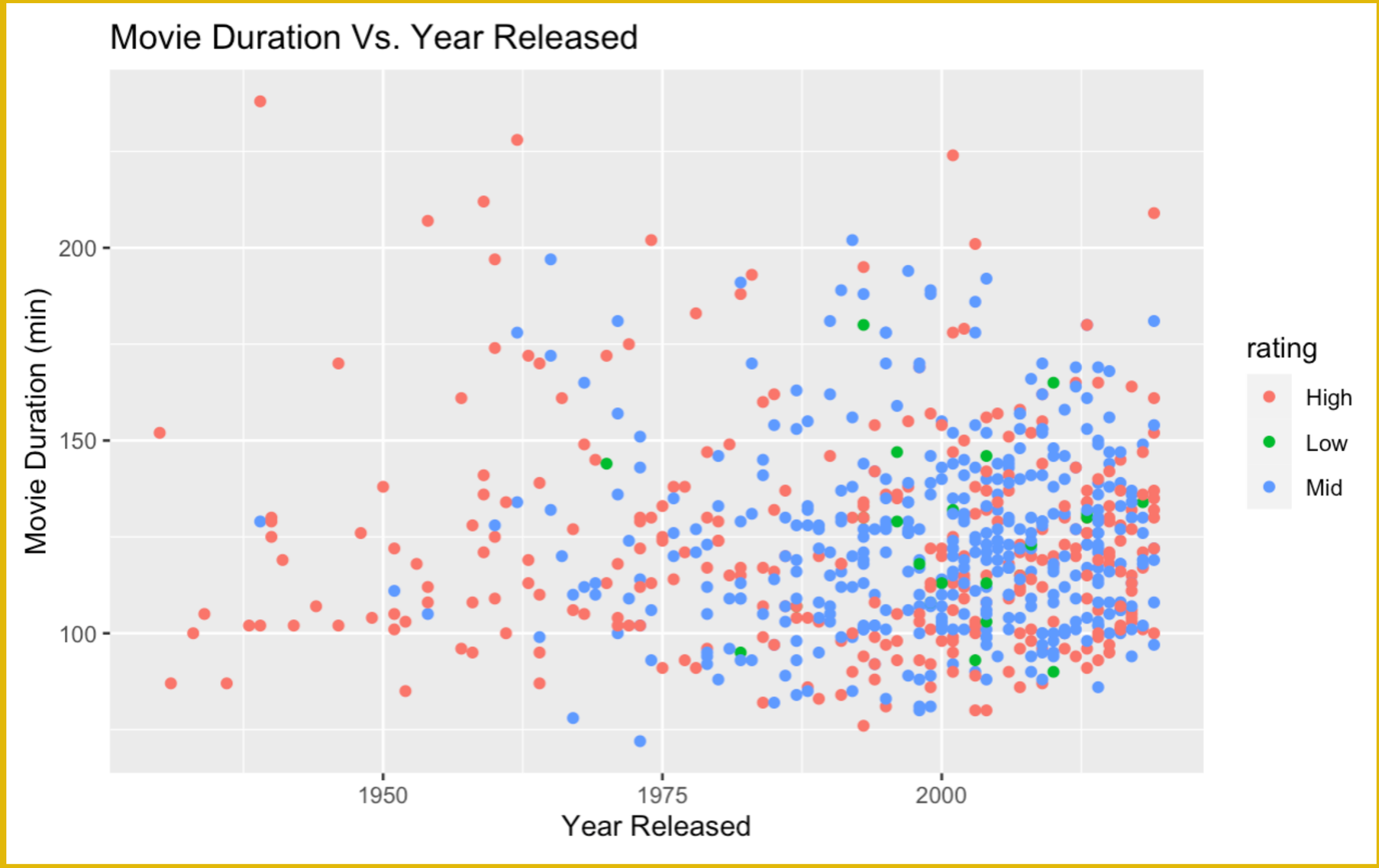


KNN

MOVIE DURATION VS. YEAR RELEASED

CONFUSION ACCURACY IS: 59%

I would say there is not really any relation between the duration of a movie, the year released, or its rating class.

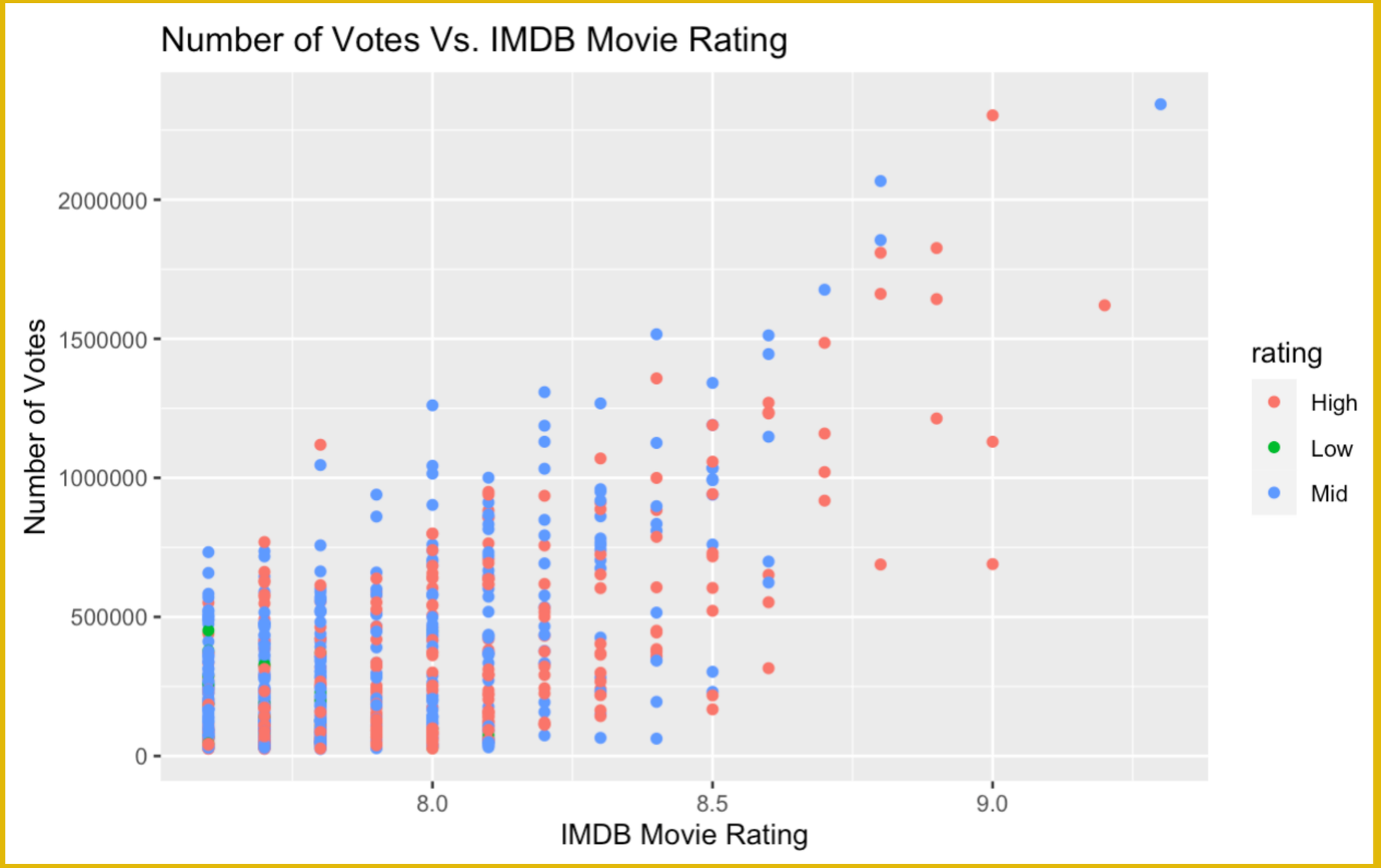


KNN

NUMBER OF VOTES VS. IMDB MOVIE RATING

CONFUSION ACCURACY IS: 51%

We could say there is a positive correlation between the Number of Votes a movie got and its IMDB Movie Rating as well as for its rating class.



STEPWISE REGRESSION

LINEAR REGRESSION RMSE: 10.82127

STEPWISE FOWARD RMSE: 10.82127

STEPWISE BACKWARD RMSE: 10.82127

STEPWISE BOTH RMSE: 10.92213

THANK YOU!