TMDB Score Prediction

**Problem Introduction**

1. Here I try to predict movie score based on the dataset which contain 28 variables. We know that before we watch a movie, it’s common to figure out what genre the movie is and who are the director and actor is. If we are a big fan of them, we tend to rate the movie with a high score. So, in this notebook I try to use these variables to predict the movie score. Please to keep in mind that I set movie score: 0-4 as awful, 4-6 as average, 6-8 as good and 8-10 as great.

The dataset includes the information of around 5000 different movies , with 28 columns: color, director\_name', 'num\_critic\_for\_reviews', 'duration',

'director\_facebook\_likes', 'actor\_3\_facebook\_likes', 'actor\_2\_name',

'actor\_1\_facebook\_likes', 'gross', 'genres', 'actor\_1\_name',

'movie\_title', 'num\_voted\_users', 'cast\_total\_facebook\_likes',

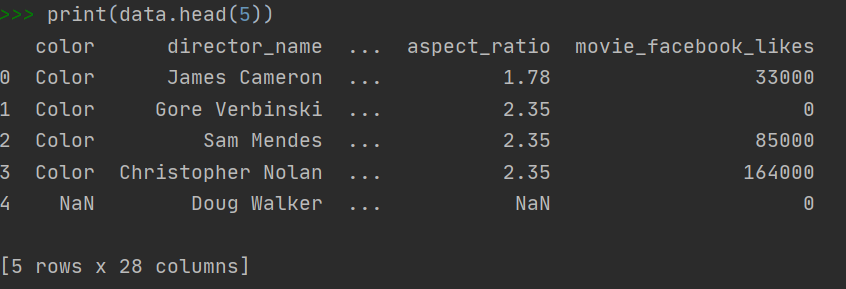
'actor\_3\_name', 'facenumber\_in\_poster', 'plot\_keywords',

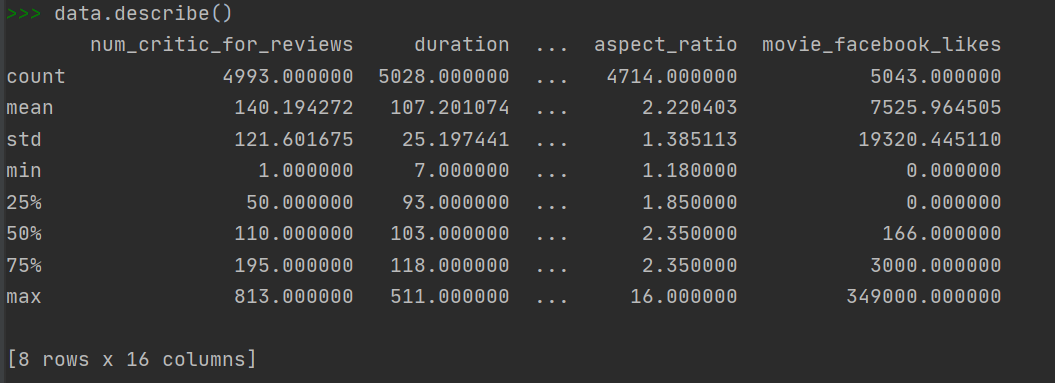
'movie\_imdb\_link', 'num\_user\_for\_reviews', 'language', 'country',

'content\_rating', 'budget', 'title\_year', 'actor\_2\_facebook\_likes',

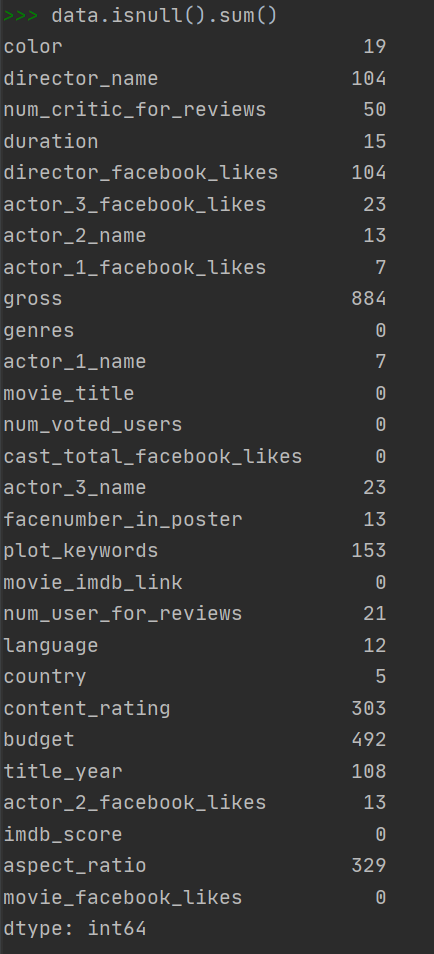
'imdb\_score', 'aspect\_ratio', 'movie\_facebook\_likes',’Revenue”

**Overview of dataset**





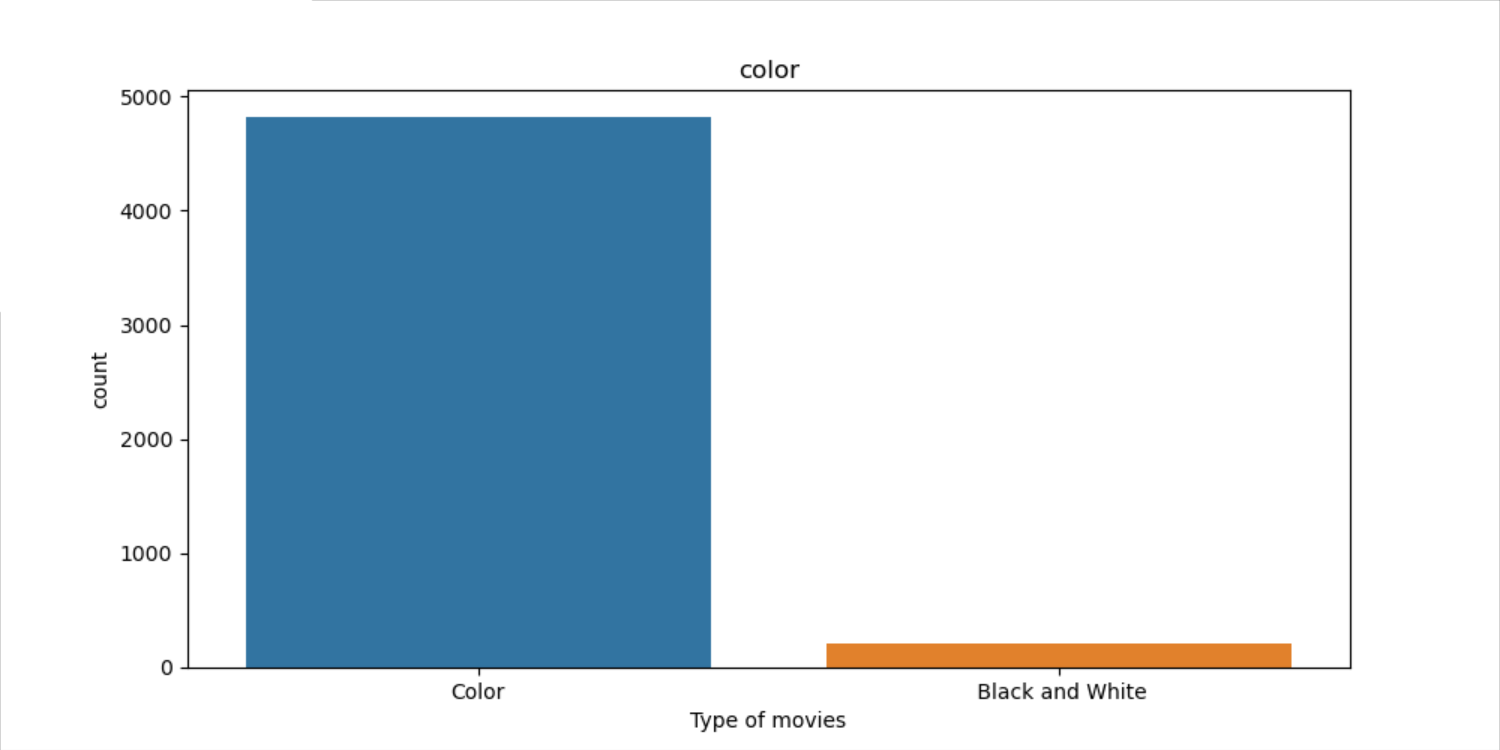
**Missing value**



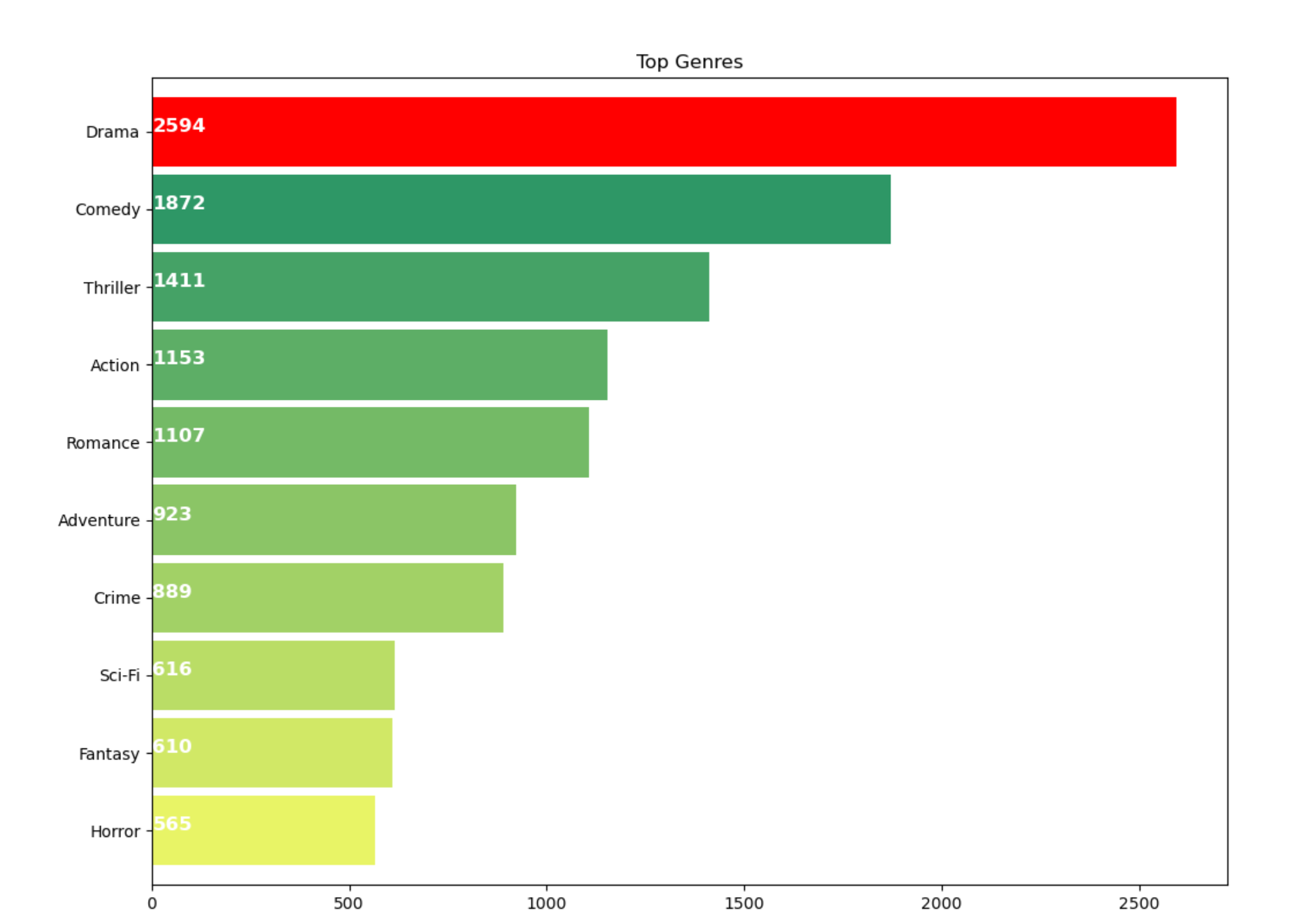
**EDA**

Color movie number VS. Black and White movies number

The majority movies are color movie, black and white movies only have a small percentage.



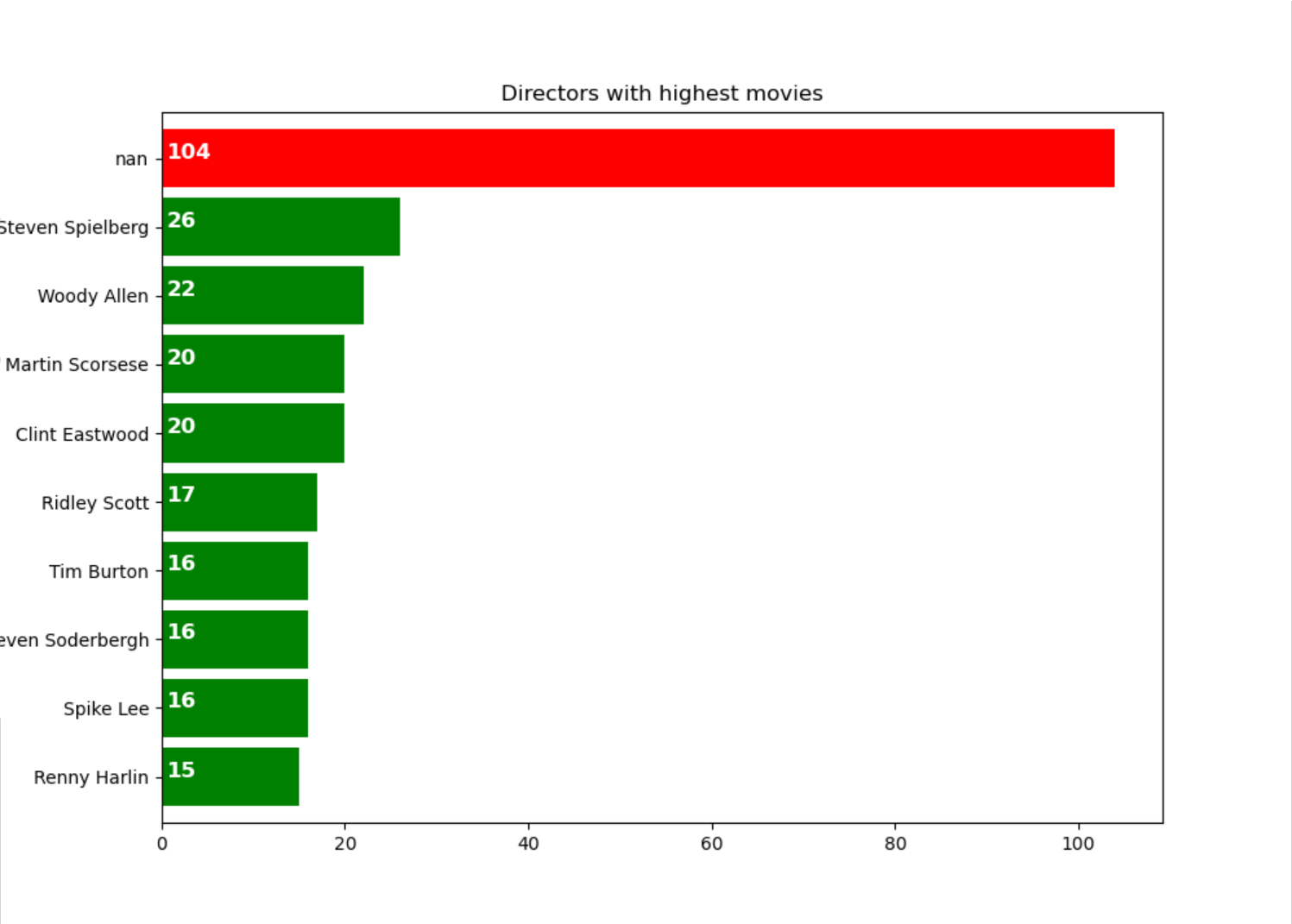
**Number of movies for each genre**



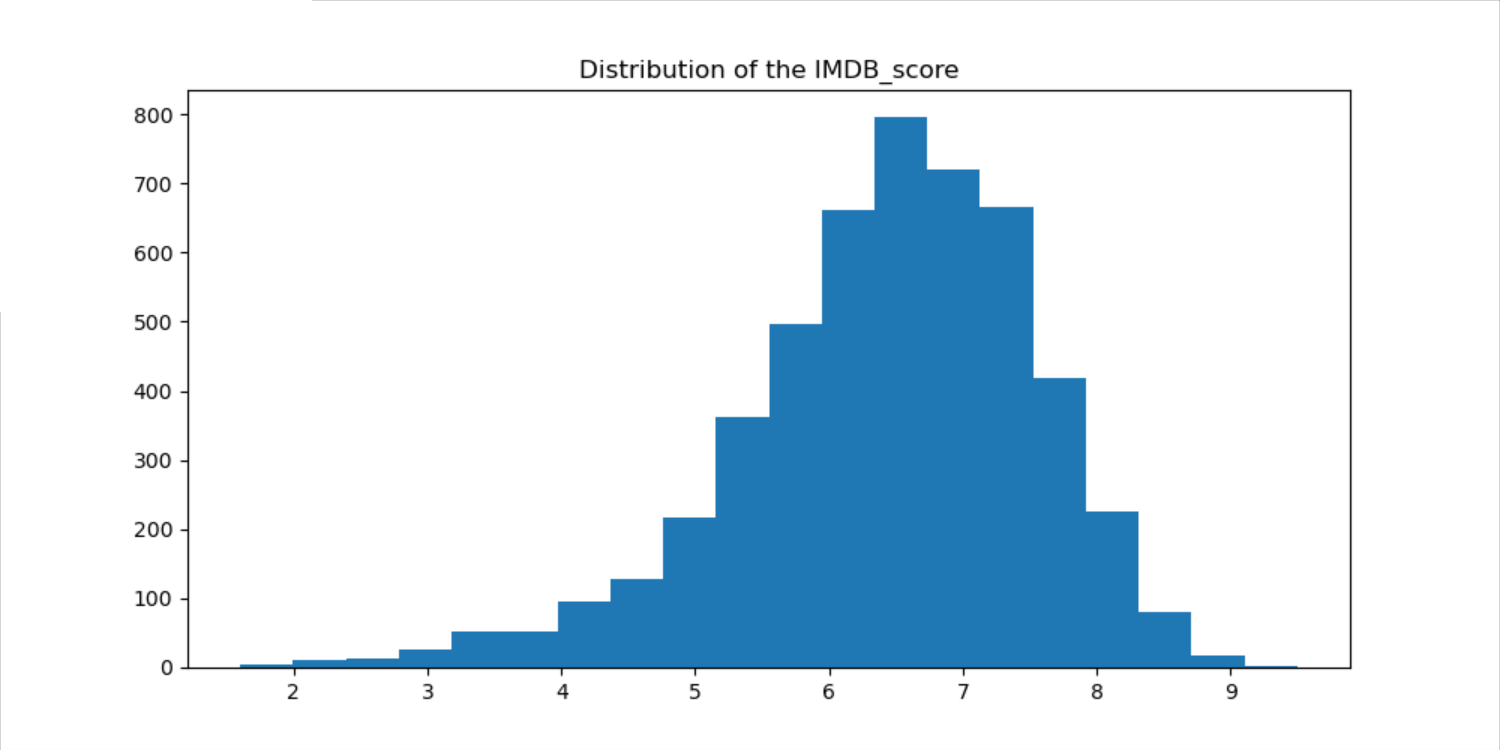
Top 5 genres are drama, comedy, thriller, action, romance

**Directors name with highest movies**

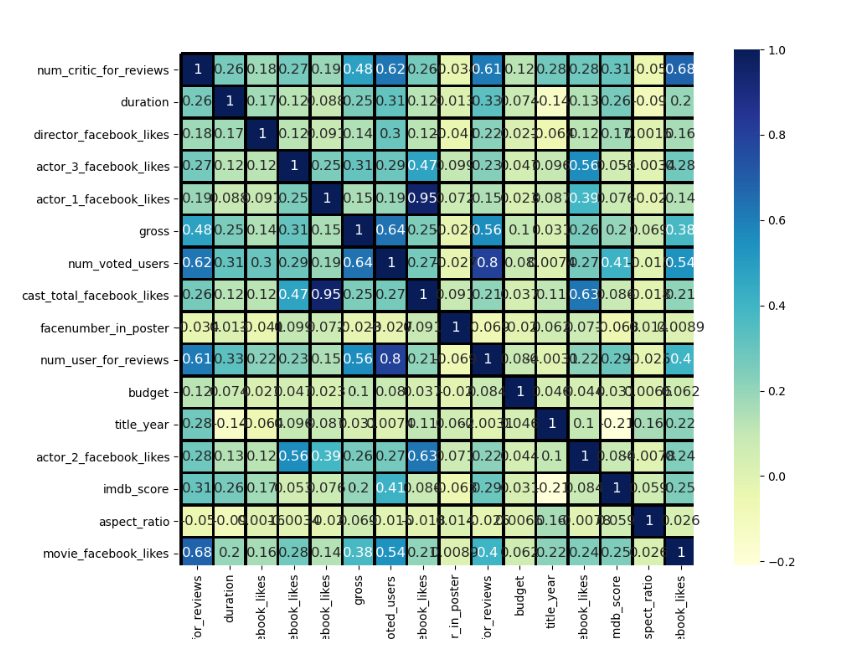
It seems that Steven Spielberg is the most productive director in this dataset. Despite that, there are 104 movies we don’t know who are their director



**Distribution of IMDB\_score**



**Correlations between columns**



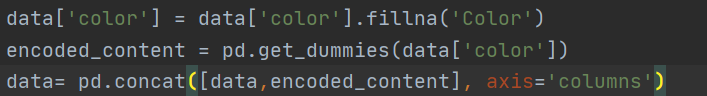
(1) There is a high correlation among those variables: cast\_total\_facebook\_likes, actor\_1\_facebook\_likes, actor\_2\_facebook\_likes, actor\_3\_facebook\_likes.

(2)Also,num\_critic\_for\_reviews,num\_voted\_users,num\_user\_for\_reviews,movie\_facebook\_likes and gross have a high correlation

**Data cleaning**

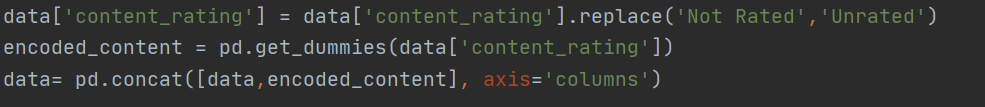
1.Color(19 missing values)

since there is a column that provide movie link on imdb, so it’s easily to find out that all the movie with missing value on this column are color. So I just impute those blanks with color and create dummy variables



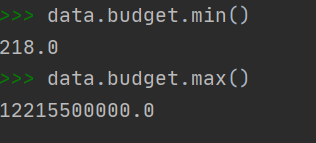
1. content\_rating

There are many values called Not Rated in these columns, which have the same meaning with Unrated. So I just recode it. And then create dummy variable.

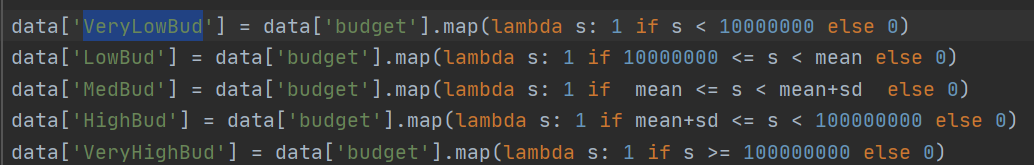


1. budget

there is a huge span from 218 to 12215500000.



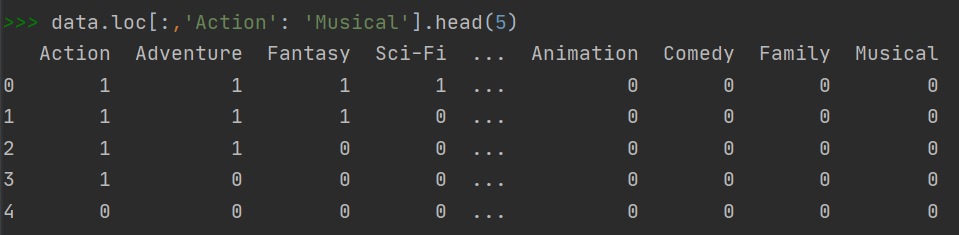
So I split it and create new variables for the purpose of modeling



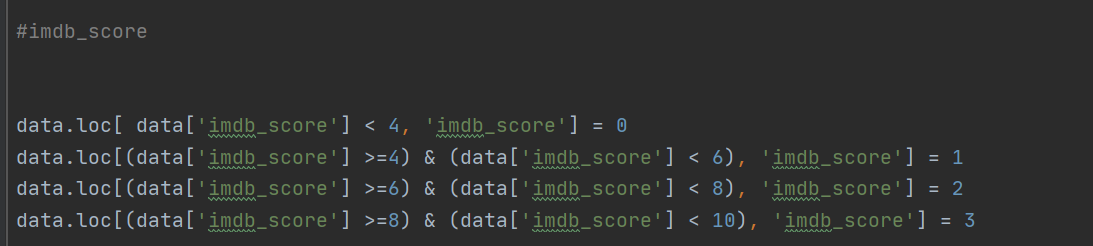
1. Genres / director name/actor name

Since there is a possible that one movie can fall into several different genres, so I create new variables in the name of genres

It looks like this, and it the same with director name, actor name that I do



1. For the target variable imdb\_score, I make it as ordinal that score<4 as 0, 4-6 as 1,6-8 as 2 and 8-10 as 3 which represent awful, average , good and great.



At last, I dropped those variables



**Model**

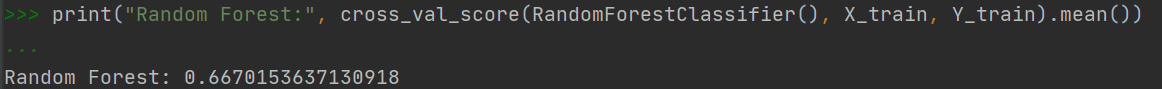
There are 5 models, Decision tree, Logistic regression, Random forest, SVC and GaussianNB

here are the results

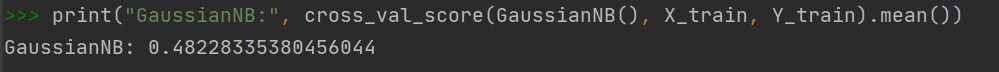
here is model performance that measured by cross validation method











We can find that Random Forest ,which got the highest score among five models ,is our best model.

Here is the cross validation scores of five different models,

