Predicting IMBD Scores

Shivani Dedhia, Akhila Pamukuntla, Nafis Chowdhury, Akshita Jain

STA 9750 Final Project

Introduction

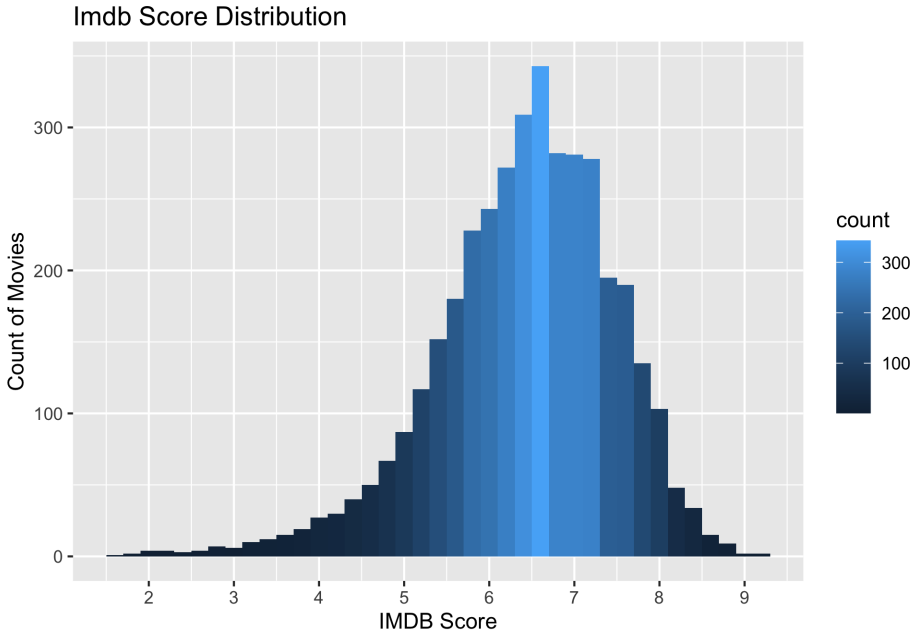
Many factors make a movie succesful and profitable. Some of the factors include budget of the movie, actor's and director's popularity, etc. We will find out which factor has the most impact on the movies success and profitability.

We have the IMDB 5000 movie dataset, which has 28 variables, 5043 movies across 100 years in 66 countries. There are many varibles in the dataset such as Director, Actors, Duration, Gross, Budget, Genres, Facebook Likes, etc.

Hypothesis

Data Exploration

After exploring the data variables such as languages, aspect ratio, imdb movie link and color were removed to reduce bias. As many of these variables had data heavily skewed on one side.Count of movies rated above 7.5 which are considered to be highly recommended. Majority of the movies are rated 7.6 with only a handful rated above 9. The highest rating recieved by one movie is 9.3. This shows there is not enough data to understand the variables that have the strongest impact on highly rated movies.

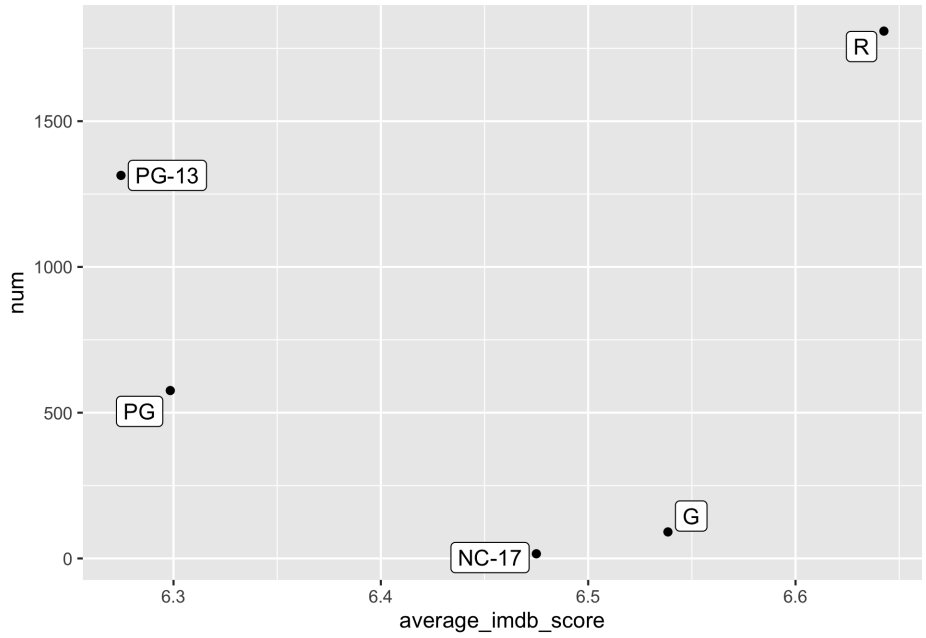


```
## # A tibble: 17 x 2
## # Groups:   imdb_score [17]
##    imdb_score     n
##         <dbl> <int>
##  1        7.6   100
##  2        7.7    90
##  3        7.8    83
##  4        8      55
##  5        7.9    52
##  6        8.1    48
##  7        8.2    24
##  8        8.3    24
##  9        8.5    19
## 10        8.4    15
## 11        8.6     8
## 12        8.7     7
## 13        8.8     5
## 14        8.9     4
## 15        9       2
## 16        9.2     1
## 17        9.3     1
```

Average Imdb Score by Content Rating

Average score for movies by content rating is below 6.6 which is considered a poor imdb score. As per this distribution content rating does not have a strong impact on the imdb score.

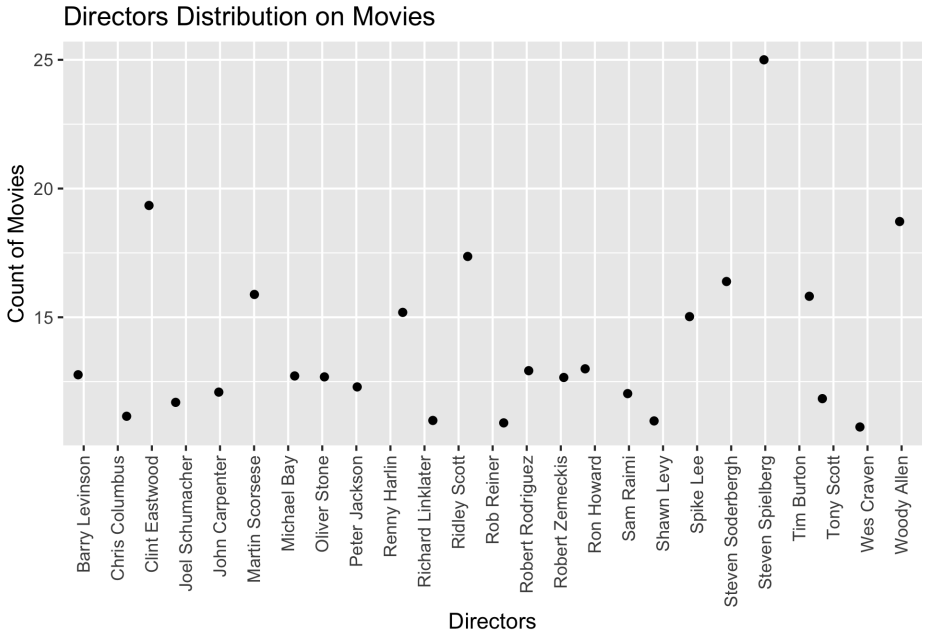
```
## `summarise()` ungrouping output (override with `.groups` argument)
```



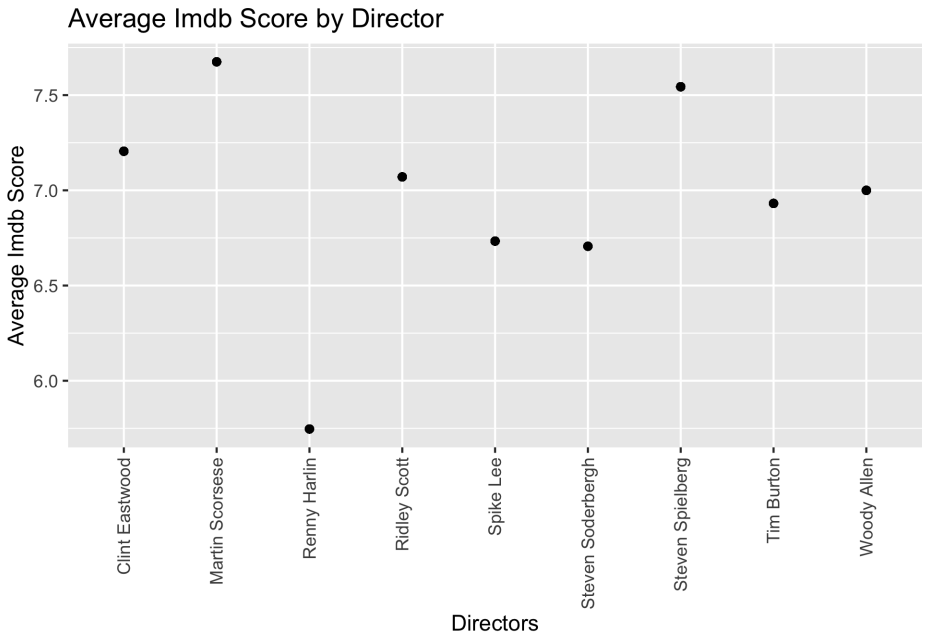
Understanding the distribution of directors and their effect on IMDB score

Directors have been grouped by the number of movies directed. The data has been filtered to only show movies directed above 10 and below 50 remove any anomalies in the data. Directors with more movies could have a higher fan following, credibility and success rate possibly leading to a higher imdb score.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



Only a handful of directors have over 15 movies directed in this data set. Steven Spielberg is the only director to have ~ 24 movies directed. The chart below shows average imdb score for directors with 15 or more directed movies. The imdb score is above 5.5 for directors with more than 15 movies. Most directors have recieved a higher imdb score that shows the number of movies directed has a slight impact on the imdb score.

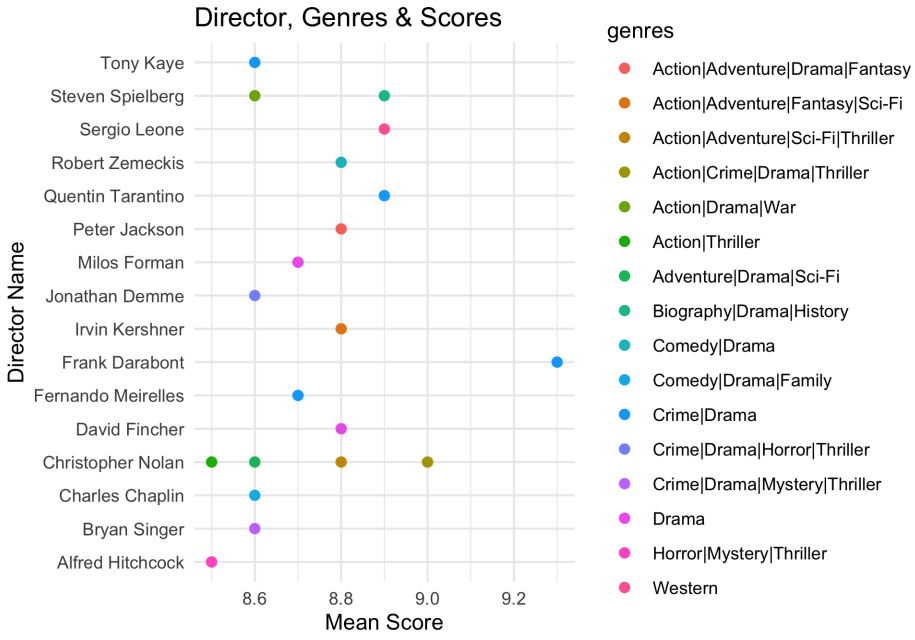


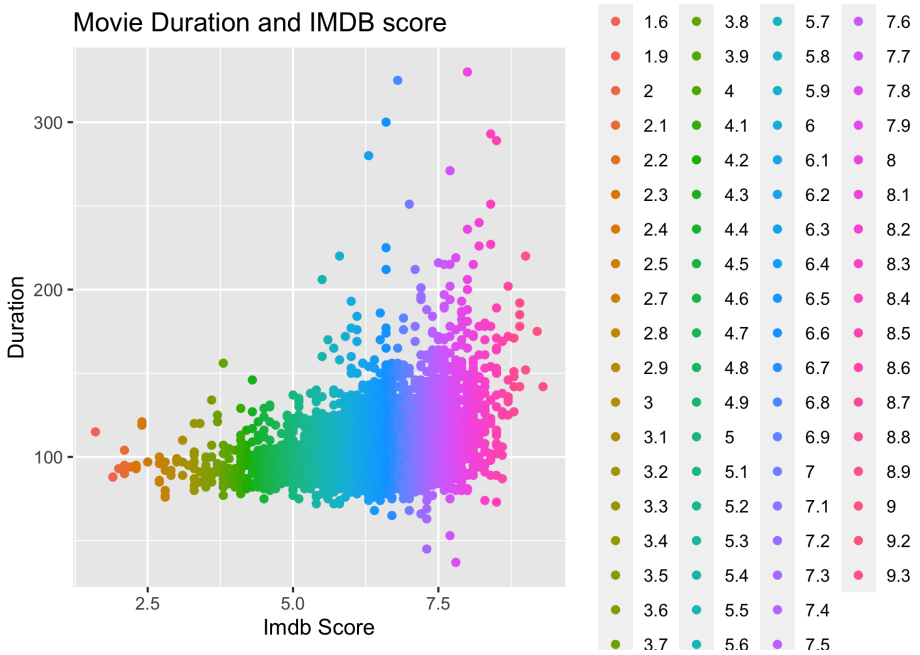
Impact of movie duration for a higher imdb score

The average imdb score is shown compared to the duration of the movie. The duration of the movie increases as the imdb score increases. Duration of the movie is not a strong determinator of the imdb score but has a slight impact on the score. The number of movies rated above 7 increases as the movie duration increases.

Top 20 movies grouped by genres with the highest imdb score.

```
## `summarise()` regrouping output by 'director_name' (override with `.groups` argument)
```

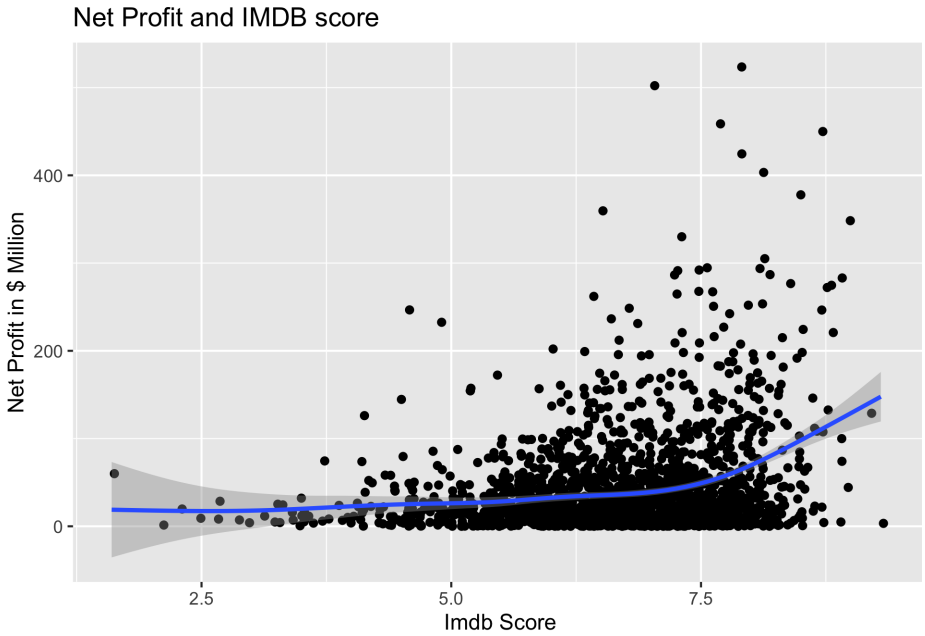




Impact of net profit on imdb score.

Movies with a net profit over 200 million has a higher imdb rating. Net profit is one of the strong indicators of a higher imdb score. As the trend below shows higher net profits translates to a higher rating. It could be assumed that the viewrship for movies with higher net profits was higher thus recieved a higher movie rating.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Training the model

```
##
## Call:
## lm(formula = imdb_score ~ duration + director_facebook_likes +
##     num_critic_for_reviews + budget, data = imdb_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.184 -0.542  0.065  0.660  2.588
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.64e+00   8.61e-02   53.94  < 2e-16 ***
## duration                  1.29e-02   7.94e-04   16.24  < 2e-16 ***
## director_facebook_likes   3.19e-05   5.72e-06    5.57  2.8e-08 ***
## num_critic_for_reviews    2.34e-03   1.44e-04   16.25  < 2e-16 ***
## budget                   -1.01e-10   7.12e-11   -1.42     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.93 on 3039 degrees of freedom
## Multiple R-squared:  0.211,  Adjusted R-squared:  0.21
## F-statistic:  203 on 4 and 3039 DF,  p-value: <2e-16
```