**What determines the successful movies?**

**Data Source:** Kaggle-IMDB 5000 movie dataset

<https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>

We download the dataset with csv format from the link above and saved in our project folder.

**Data Cleaning:** Remove missing value rows, 3756 rows for analysis.

**Programming Language:** Python

**Modules and Packages:**

1. Pandas
2. Matplotlib.pyplot
3. Numpy
4. Seaborn
5. Scipy.stats

**Data Processing:**

We cleaned the data by deleting the rows with missing values. And as to each research question, we renamed the values of variables of our interests due to the duplicated or similar values.

**Analysis Method:**

1. Exploratory analysis

We did trend analysis with line charts and scatter plots.

1. Statistical analysis

We used bar plots, violin plots, t-test, anova analysis to find the answers to our research questions and hypotheses.

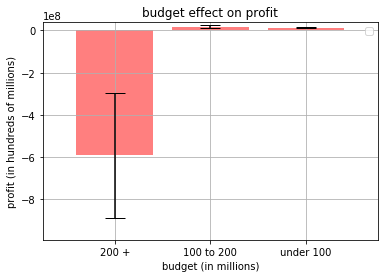
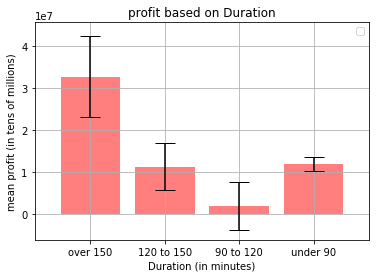
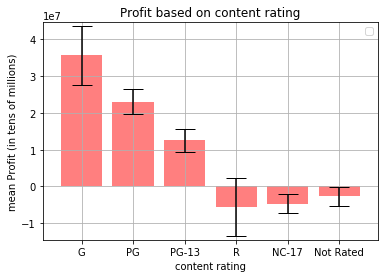
**Research questions**:

**Main Question:** What makes a successful movie?

**Sub-Questions:**

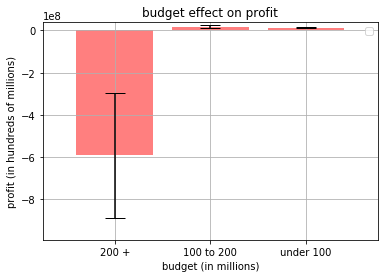
1. What controllable variables will affect profit?

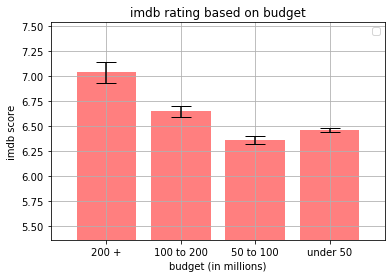
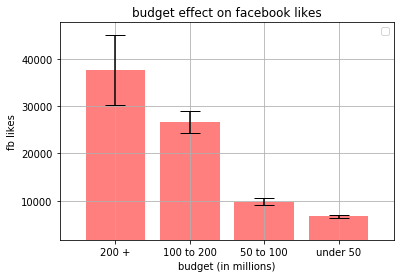
Based on the graphs shown below, movies that are available to a wider audience perform better. “G” rated movies return much higher profit than “R” rated movies, which average very poorly. Surprisingly, longer movies typically return more profit than shorter ones. 90 - 120 minute movies returning the least but under 90 minutes return as about much as 120-150 minute movies. Very high budget movies (over 200 million) appear to return the least profit, averaging a significant loss. Long, “G” rated movies with a budget under 200 million appear to be the ideal for generating a high profit.



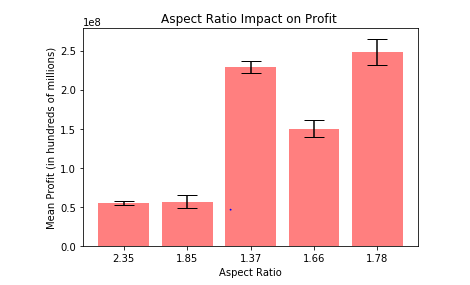
1. How does budget affect movie success?

Based on the graphs shown below, budget appears to have a statistically significant effect on whether or not a film is successful (based on the three types of success we have defined). Higher budget films appear to perform consistently better with imdb reviews and with gaining facebook likes, however, particularly high budget films have poor return on investment, averaging significant loss. 100 to 200 million appearing to be the best budget for remaining profitable and getting high facebook likes and imdb scores.



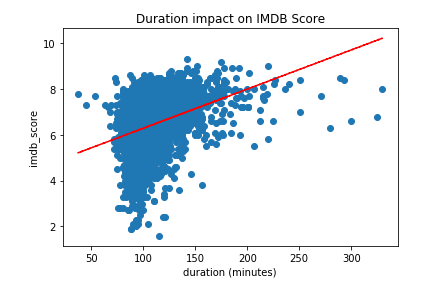


1. Does the Aspect Ratio chosen lead to a higher profit?

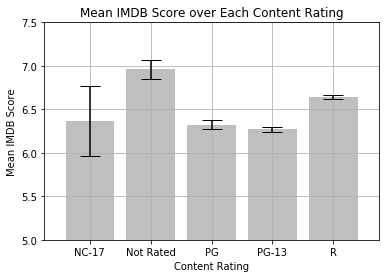


The graph indicates that the majority of directors use the 2.35 and 1.85 which are wider screens. There are more movies in these populations. The aspect ratios that seem to have rendered a higher have significantly lower populations and does not seem to tell an accurate story. The most profitable aspect ratios only had between 30-50 movies in the populations. They might have had a couple of movies in their population that did very well, but since there is a small number of movies, that data can be skewed. We would want to focus on the most popular aspect ratios which are the 2.35 and 1.85. They have almost the same population and rendered almost the same average profit. The 1.85 aspect ratio rendered a tad bit more profit than its competing ratio, but there is also a small visual difference between the two aspects. Picking one of these aspects will be more desirable because they are a better representation of the data as a whole.

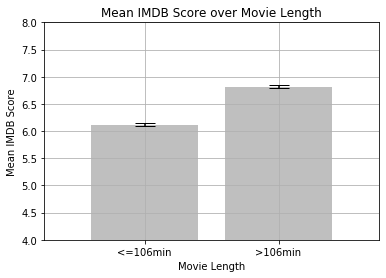
1. Does the length a movie impact the IMDB score?



There is a general increase in the IMDB score as the movie duration increases. Movies with a duration greater than 90 minutes on average have a higher IMDB score. Longer movies do not render a better/higher IMDB score. Most movies are between 75 and 160 minutes.

1. Is there any difference of IMDB scores among all content ratings? Yes  
     
   ANOVA: F-Statistic=29.37, P-value=4.42e-24

P-value is less than 0.05, so the mean IMDB scores are not identical for each content rating. Not Rated movies have higher mean IMDB score( 6.9) compared to rest, and R movies have higher ratings compared to PG and PG-13.

1. Is there any difference of IMDB scores between long and short movies? Yes  
     
   T-test for IMDB score:

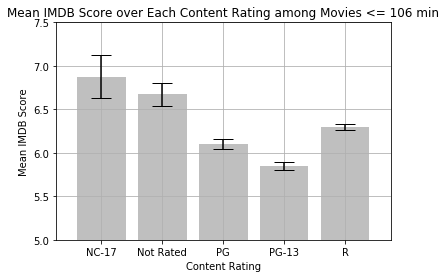
Statistic=-21.56

P-value=3.72e-97

The length 106 mins is defined by using 50% percentile.

The mean IMDB score of longer movies(6.82) is significantly higher than shorter movies(6.12).

1. Is there any difference of IMDB scores among all content ratings stratified by

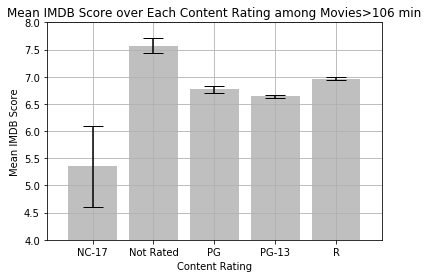
movie length? Yes  


ANOVA:

F-Statistic=18.23

P-value=1.07e-14

In short movies, NC-17 and Not Rated movies have higher mean IMDB scores than the rest, and R is higher than PG, and PG is higher than PG-13.



ANOVA:

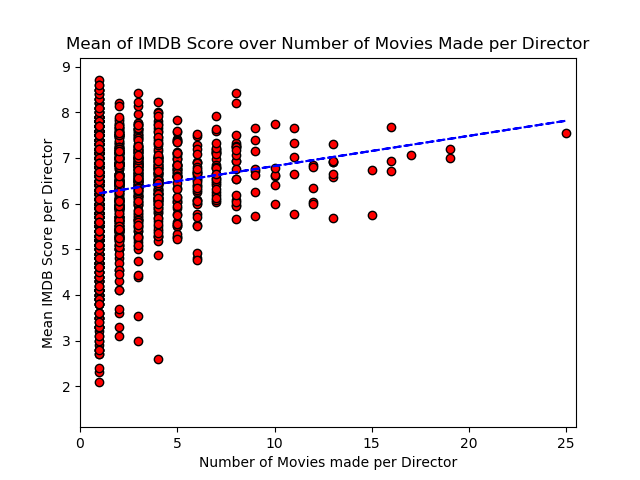
F-Statistic:18.94

P-value=2.75e-15

In long movies, Not rated movies have the highest IMDB score. Than R is higher than PG, PG-13 and NC-17, PG and PG-13 is higher than NC-17.

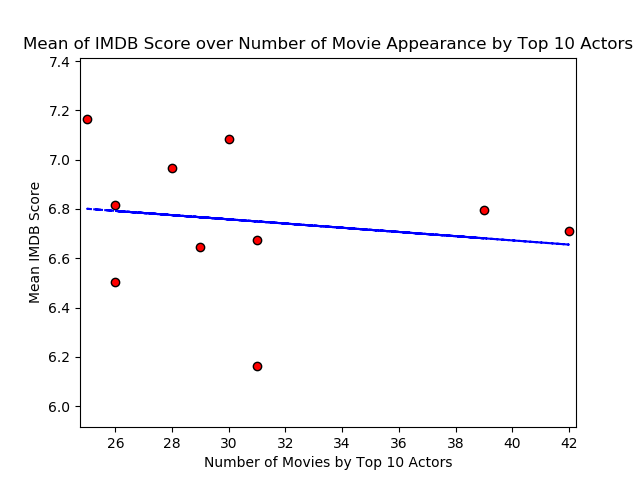
The mean IMDB scores are not identical for each content rating among longer movies, and it showed the similar pattern among shorter movies.

1. Does an increase in the quantity of movies made, increase a director’s mean IMDB scores?



The chart above shows a weak positive correlation between the number of movies made by a director and their mean IMDB score.

1. Does hiring top actors guarantee high IMDB scores?



The chart above show a weak negative, nearly horizontal correlation between the number of movies featuring top actors and their mean IMDB score

**Conclusions**:

Duration and content rating both appear to have an effect on profit. Longer movies with g or pg ratings are the most successful

Higher budgets result in slightly higher imdb scores and much higher facebook likes, however, particularly high budgets (over 200 million) result in a loss of profit on average.

The mean IMDB scores are not identical for each content rating.

The mean IMDB score of longer movies(6.82) is significantly higher than shorter movies(6.12).

The mean IMDB scores are not identical for each content rating among longer movies, and it showed the similar pattern among shorter movies.

There is a positive relationship between the number of movies a director has made and how successful those movies are. This may suggest that as directors create more movies over time there is a higher chance of achieving above average success.

There is a weak negative to zero relationship between top actors featured in movies and the average success of their movies. This may suggest that the choice in actor does not have as strong an influence as initially thought. Additionally, top actors will have a higher quantity of movie successes and movie failures as their quantity of movies filmed increases. Hence, explaining why their mean IMDB score average very similar to each other.

**Reference**:

1) Trendline for Mean IMDB Score over Years was taken from the stackoverflow answer below:

https://stackoverflow.com/questions/26447191/how-to-add-trendline-in-python-matplotlib-dot-scatter-graphs

2) Number of Movie by Year was taken from below link:

https://www.kaggle.com/aninda123/imdb-movie-analysis