Analyzing a Successful Movie – Part 2

By Yuming Tsang, Nalu Zou, Jerome Orille

Summary of research questions

- Which countries make the most successful movies?
    - Countries with top 5 average ratings:
        - Federal Republic of Yugoslavia: 8.1
        - Republic of Macedonia: 8.1
        - Iran: 7.92
        - Brazil: 7.78
        - Switzerland: 7.77
    - Countries with top 5 highest ratings
        - USA: 9.3
        - New Zealand: 8.8
        - Brazil: 8.7
        - India: 8.7
        - France: 8.6
- What factors contribute to the success of a movie?
    - Revenue has the greatest correlation with budget
    - Profit has the greatest correlation with budget
    - Ratings have the greatest correlation with runtime
- What qualities do crew members of successful movies have?
    - Top three most frequent types of crew members are:
        - Normal crew: 337 people
        - Art: 189 people
        - Visual Effects: 168 people
    - The three movies with the most total crew members are:
        - Jurassic World: 435 total crew members
        - The Dark Knight Rises: 217 total crew members
        - Avatar: 153 total crew members
- What is the optimal depth for a decision tree to predict the success of a movie?
    - The mean squared error varies wildly over various depths, leading to very little correlation. The wild variance in mean squared error over the various depths is unexpected.

Motivation and background

Producing a movie is a very costly process that involves collaboration of actors, editors, and producers. Therefore, it is important to be able to predict if a movie will be successful before investing time and money into its production. For this project, we wanted to analyze traits of movies to see what makes a movie good. We want to find correlation between features of a movie and movies that are successful. Successful is a broad term. We analyzed many features that can consider a movie successful, such as gross income or the average rating of the movie.

Datasets

https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv

This is our first dataset. It is called 'The Movies Dataset.' It was compiled by Rounak Banik on November 9th, 2017. The movies collected are from the Full MovieLens Dataset, and it only features movies released on or before July 2017.

The dataset features metadata about movies, which include crew members, producers, budget, revenue, reviews, etc. This dataset is mainly used for its data on crew members.

https://www.kaggle.com/danielgrijalvas/movies

This dataset, called "Movie Industry," is also from Kaggle. It was compiled by Daniel Grijalva two years ago and contains more simple data than compared to the previous dataset. There are 6820 movies in this dataset from the years 1986 to 2016. Each movie has 15 attributes such as the budget, country of origin, director, genre revenue, rating, and IMDb score.

https://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/

This dataset is a simple map of countries. This is used for the comparison of movies produced by country. This gives us the geometries to enable mapping.

The last dataset, which was from the IMDb database, was not used. We felt like we had sufficient data to work with from the two datasets above. Also, replicating results would be different in the future, because the database updates itself daily. This would make testing from day to day difficult.

Methodology

Our analysis will start off by analyzing movies in countries. The datasets provided to us show us where the movies are produced. We can take advantage of this by checking out which countries had the highest ratings and highest average ratings. This data can be visualized with two maps of the world. One map will show the highest rating a movie produced by a country has received. Another map will show the average movie rating per country. In other words, the maps will compare how movies are rated in each country.

We would go on to compare factors of a movie that make movies successful. Some factors we would use in our analysis would include revenue, profit (gross income), and ratings. We would test these features against three labels: budget, year of production, and runtime. We chose these features as they are numerical and would be easy to compare with each other. Plots would be made between the three features and three labels each, resulting in a total of nine plots. All the results would be summarized and combined in a correlation matrix, where the correlations of all the data can be easily compared.

After that, we will look at crew members from movies. We would find the top ten movies by profit in descending order. We would check to see how many crew members

worked on the movies in total. This would be done visually with a bar graph, showing the movies with the total amount of crew members per movie. Then, we would compare the departments of crew members. This would be done by grouping the members by department. We can visualize this with a bar graph, showing the department with the total number of crew members in that department throughout the top ten movies. All in all, the first graph would describe movies and their total amount of crew members, while the second graph would describe types of crew members and how much of each type throughout the top ten movies.

Finally, learning from these patterns on how successful movies may be would then be used to help predict whether or not a future movie would be successful. Features used to predict such a problem would include genre, producers, movie runtime and target audience. What we are trying to predict is the gross revenue. A movie having high gross revenue would be considered successful in our eyes. By knowing what movies are 'good' in advance, we would know if they are worth watching.

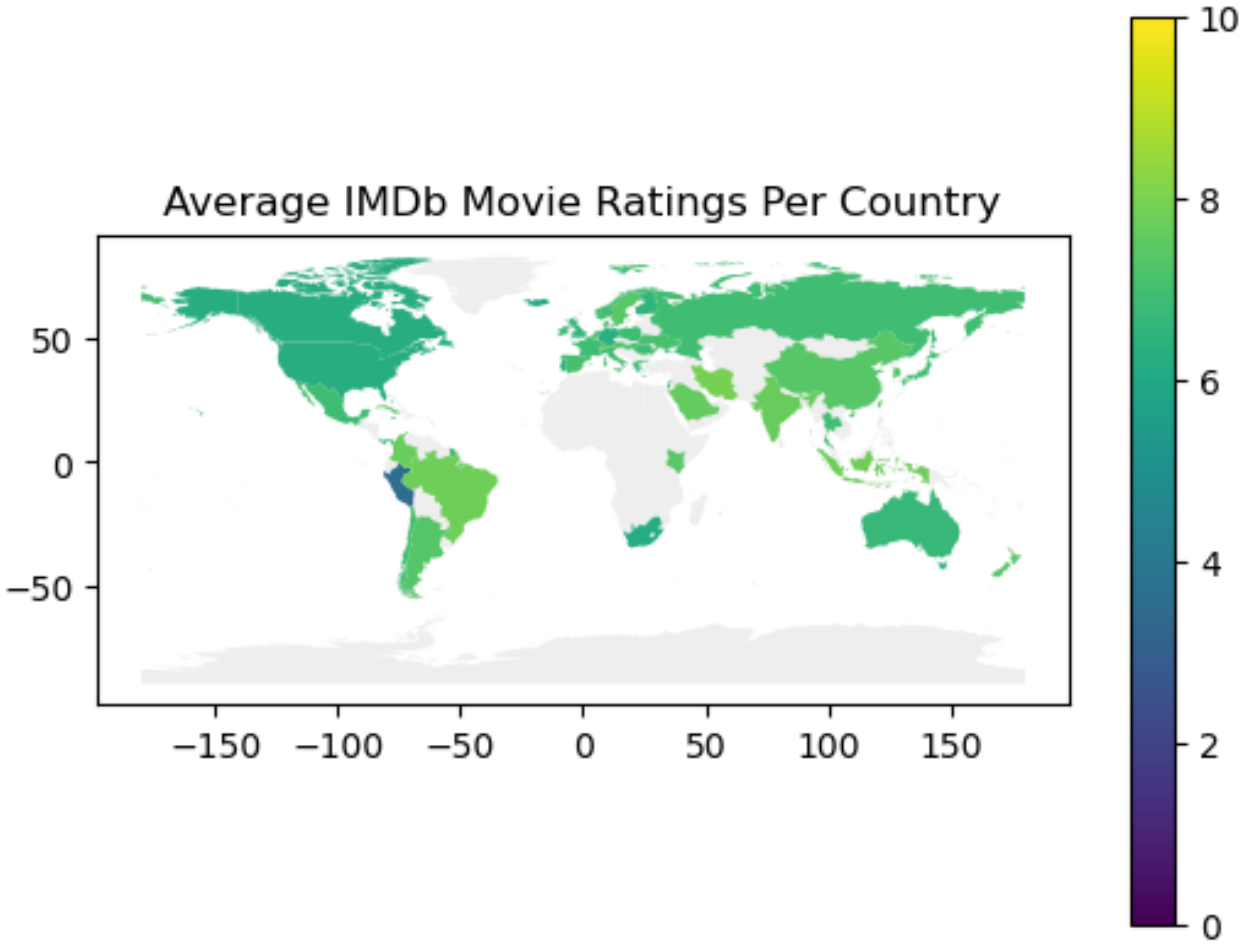
Results

- **General Notes for Reproducing/Testing**
  - Each research question has its own directory
  - .csv files related to each question are in the same directory

- **What countries make the most successful movies?**

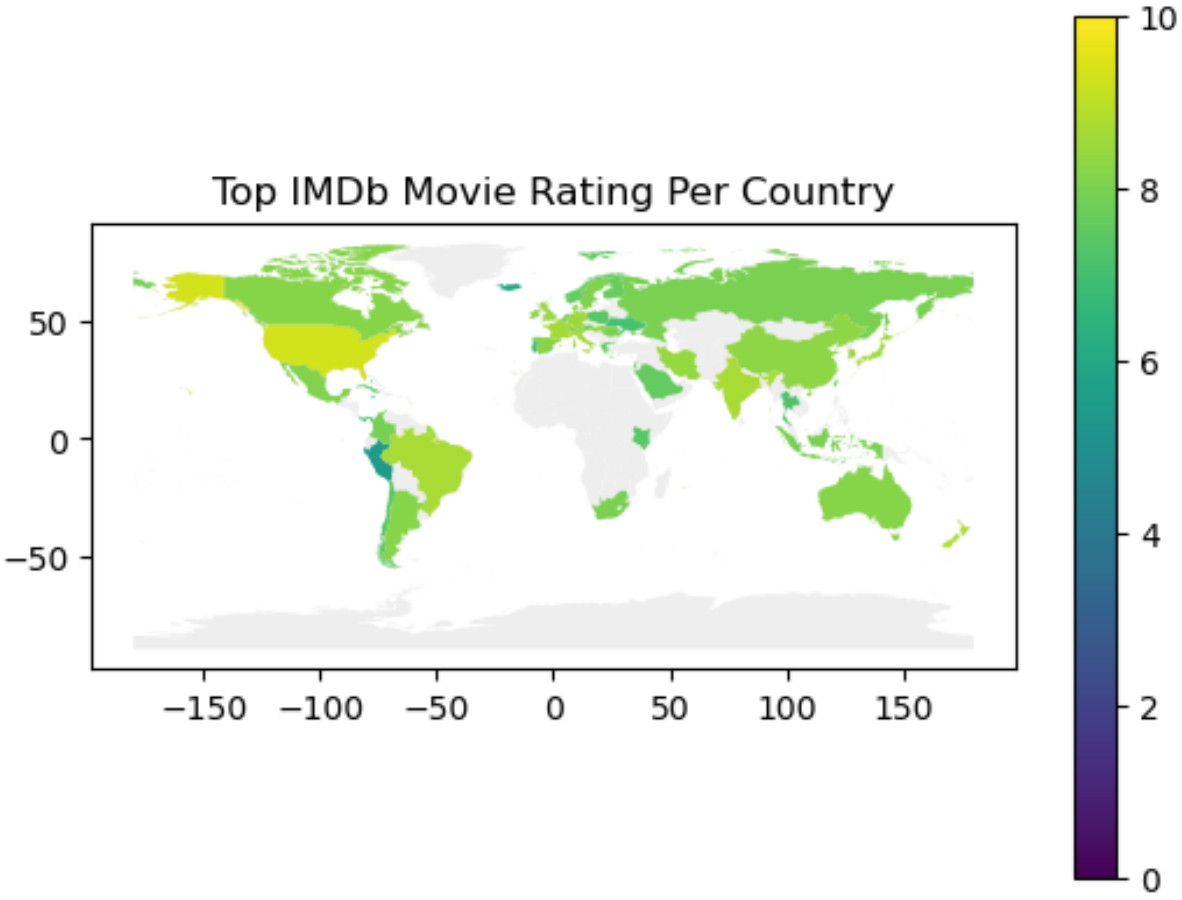
The data used for this problem was derived from the 'Movie Industry' dataset from Kaggle and the 'Countries' dataset from Natural Earth. By joining these two datasets together, we were able to turn the pandas dataframe of movies into a geopandas dataframe and represent our data on a map. We will be focusing on each movie's IMDb rating, which is a float from 0 to 10, and country of origin. Each country will be mapped and the IMDb rating will be represented by color according to a legend that has hues corresponding to numbers from 0 to 10.

We created two maps for this problem. The first shows the average IMDb ratings for movies per country.



From looking at this map, we can see that the countries in light green have the highest average movie ratings. If we sort these countries by highest average, we see that The Federal Republic of Yugoslavia and Republic of Macedonia both have the highest average rating of 8.1 followed by Iran with a rating of 7.92, Indonesia with 7.8, and Brazil with a rating of 7.78. These results are surprising because we expected the USA to be in the top 5 because of its Hollywood movie industry, but it is not. However, this can be explained because the average rating takes into account all movies including the ones that are rated poorly so countries we might expect to rate higher will be dragged down by the movies that did not perform well, especially if there are a larger amount of them. This makes sense because even though many successful movies were made in Hollywood, there were also many that were also many that failed.

The next figure we created shows the top IMDb movie rating per country:



From this plot, we can clearly see that the USA has the highest rated movie because it is the only country in yellow-green. If we sort these countries by highest ratings, we see that the USA with a rating of 9.3 is followed by New Zealand (8.8), Brazil (8.7), India (8.7) and France (8.6). These results are more similar to our expectations because it only takes into account the top movie produced by each country.

When comparing these two plots, we can see that higher top ratings do not necessarily mean that a country has a higher average rating. However, it may imply that making a movie in one of the countries that have high ratings and rating means gives a greater chance of creating a successful movie. This may be based on other factors within the country, including the number and quality of resources and people available to work on a movie. From our top 5 lists, Brazil is the only country to appear on both, meaning that based on our data, Brazil is the country that is most consistently producing highly rated movies.

Reproducing the Results - Countries:

1. Download ‘movies.csv’ from ‘Movie Industry’ and ‘countries’ folder from Natural Earth.

    a. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
    b. Natural Earth: https://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/
    c. Extract contents of 'countries' into same directory
2. Type 'python top_countries.py' into terminal
3. Open 'top-ratings.png' and 'top-movie.png' to view plots
4. View what is printed in terminal to access the figures generated for top 5 average ratings and top 5 highest ratings.

Testing - Countries:

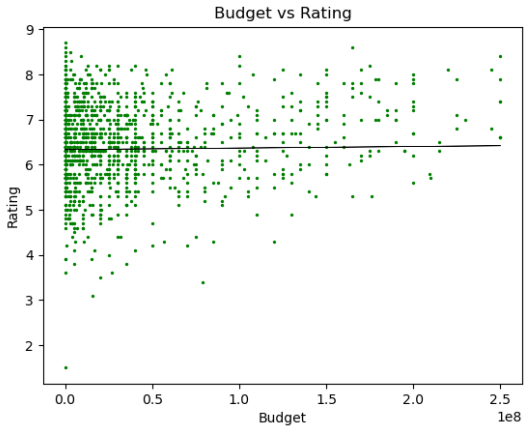
1. Download 'movies.csv' from 'Movie Industry' and 'countries' folder from Natural Earth.
    a. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
    b. Natural Earth: https://www.naturalearthdata.com/downloads/50m-cultural-vectors/50m-admin-0-countries-2/
    c. Extract contents of 'countries' into same directory
2. Type 'python test_top_countries.py' into terminal
3. View in terminal whether or not tests pass
    b. Test file creates a small dataframe and tests the functions 'sort_movies,' 'average_ratings_per_country,' and 'top_five_per_country' from 'top_countries.py.'
    c. It then compares a manually prepared expected Series with the received Series to test if they are equal using .equals(). The Series represent the countries with the top 5 average movie ratings and top 5 highest ratings. If they are all equal, the tests pass.

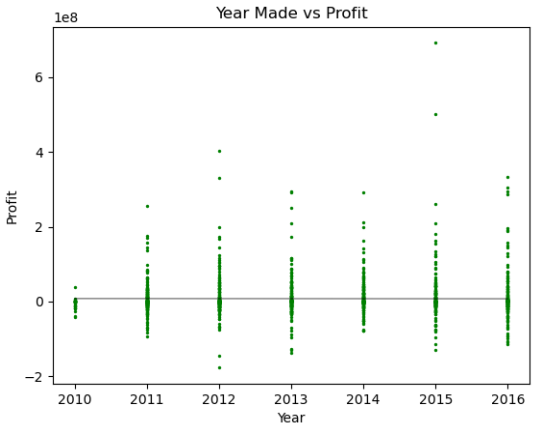
- **What factors contribute to the success of a movie?**

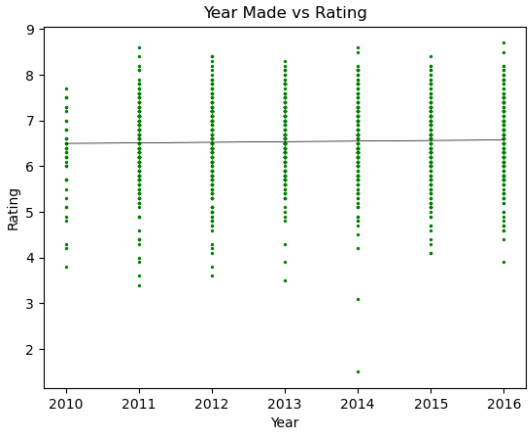
For this problem, we tested three features against three labels that determine the 'success' of a movie, again using the movies.csv dataset from Kaggle. We decided to measure 'success' by the revenue, profit, and ratings of each movie. The profit was calculated by subtracting the budget from revenue. The features that were tested were budget, year of production, and runtime, because these were the numerical data available in our dataset. We used a linear regressor for each feature vs label to try to identify what kind of relationships they had, if any, and created a correlation matrix to more holistically display our findings.

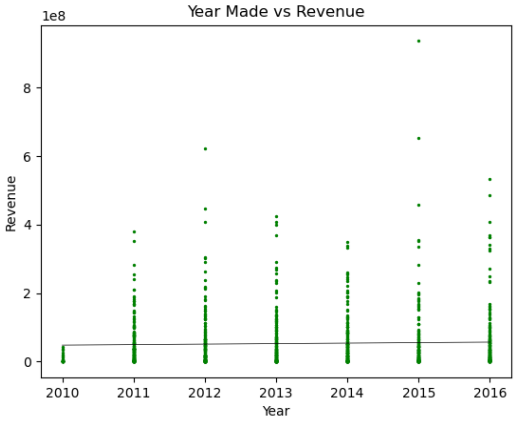
Since we generated many graphs from the linear regressor, we will classify them into three groups: no correlation, slight correlation, and high correlation. These results will be discussed and then summarized at the end with our correlation matrix.

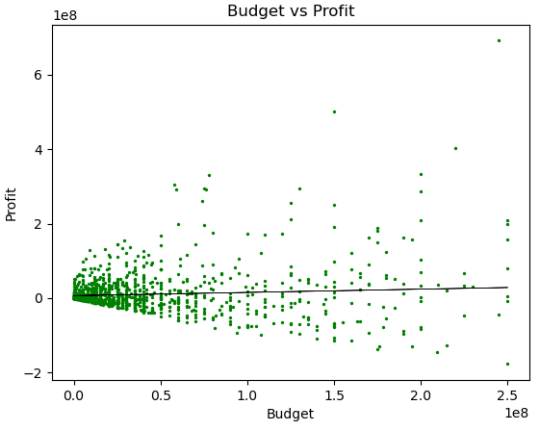
No correlation:

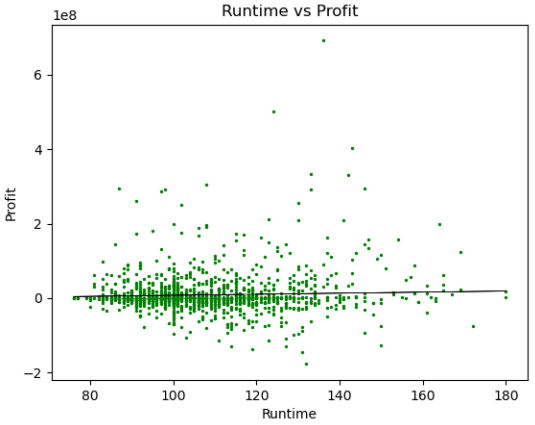








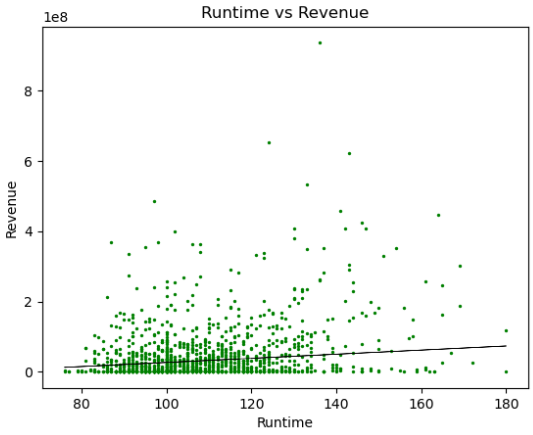




After examining these graphs, we can see that the lines are almost completely flat. This leads us to believe that the budget, runtime, and year are predicted to have no impact on its paired label in each graph. All three graphs that examine the feature of the year made show up in this group, which can lead us to believe that what year a movie is

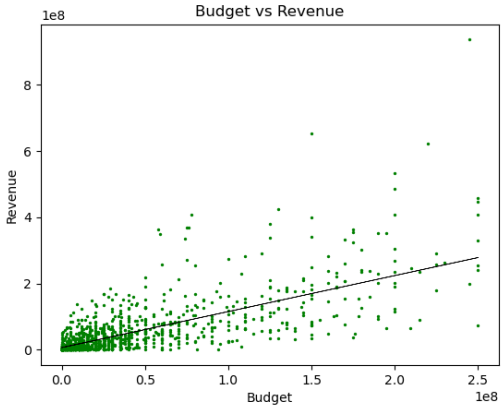
made has no impact on its success. This is interesting because we originally thought that some years would yield higher profits because of fluctuations in the economy reflected in the movie industry, but we discovered that this type of trend would be better analyzed with a line graph and more available years in the dataset.

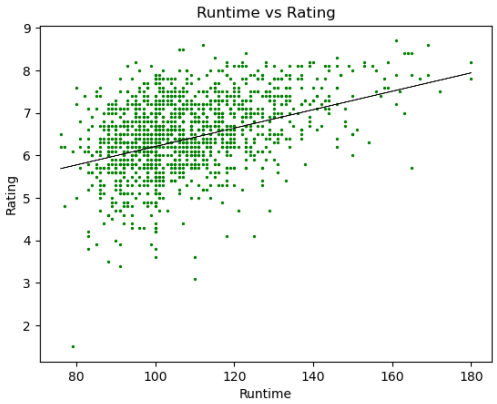
Slight correlation:



There is only one graph included in this group because although the trend line is not flat, it is not extremely strong and the dot formation is still very sporadic. The results of this graph show us that as runtime increases, revenue is also predicted to increase slightly. This leads us to believe that longer films generally have slightly higher chances of generating more revenue than shorter films. Due to this, runtime may play a small part in determining the success of a movie.

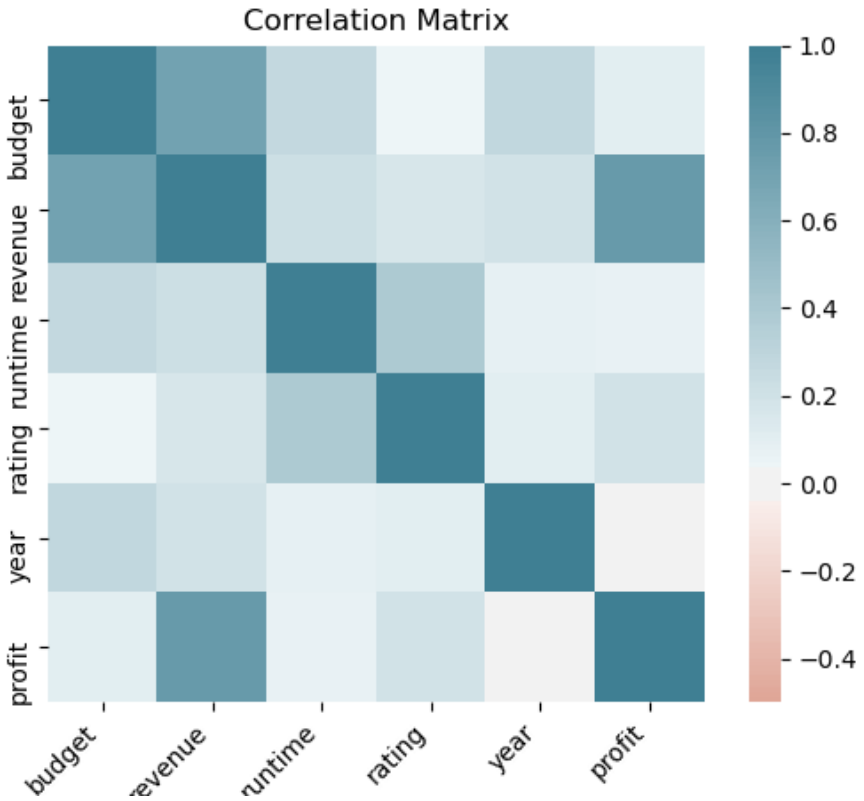
High correlation:





These graphs demonstrate that there is a positive correlation between budget and revenue and runtime and ratings. Looking first at the first graph, we see that most of the dots are following this pattern of a positive correlation and the trend line is much steeper than our previous graphs. Our linear regressor model predicted higher revenues for movies with higher budgets. This gives us strong evidence that having a larger budget for a movie will yield higher revenue. This result makes sense because qualified and professional crew members will cost more to hire and generally lead to higher quality films that do well in the box office. The second graph displays a positive correlation between runtime and ratings, meaning that our linear regressor model predicted higher ratings for movies with longer runtimes. The dots all seem to generally follow this trend, although they are quite spread out vertically. This trend suggests that creating a movie that is longer will increase the chances that it will be more highly rated. We originally did not think there would be much of a relationship between these two factors, but it could be explained by the fact that longer movies have the resources necessary to make the movie that long so the producers may be more experienced, therefore creating movies that yield higher ratings.

Correlation Matrix:



Our correlation matrix plots every factor against every factor and the saturation of the hues display the correlation between the factors. A darker blue hue represents a strong positive correlation, while an orange hue represents a negative correlation. Since none of our previous graphs revealed negative correlations, we made the legend range from -0.5 to 1 rather than -1 to 1 to make it clear that there are no negative correlations but still have enough contrast between positive correlations.

We will analyze this matrix by looking along the rows for each factor of success and seeing the factor it most strongly correlates with that is not itself or another factor of success. The three factors of success are revenue, profit, and rating. After looking at the plot, we can see that revenue has the bluest hue when paired with budget, and the same goes for profit with budget and rating with runtime. With these results, we can conclude that higher budgets and runtime increase the potential success of a movie. Since our budget and revenue patch on the correlation matrix have the darkest hue of these three and our budget vs revenue graph seems to have the strongest trend, we can conclude that of these factors, budget is the biggest determinant in the success of a movie.

Reproducing the results – Factors of Successful Movies:

1. Download ‘movies.csv’ from ‘Movie Industry’
    a. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
2. Type ‘python success-factors.py’ into terminal
3. Open these produced images to view graphs:
    a. 'Budget vs Revenue.png'
    b. 'Budget vs Rating.png'
    c. 'Budget vs Profit.png'
    d. 'Year vs Revenue.png'
    e. 'Year vs Rating.png'
    f. 'Year vs Profit.png'
    g. 'Runtime vs Revenue.png'
    h. 'Runtime vs Rating.png'
    i. 'Runtime vs Profit.png'
4. Open 'Correlation Matrix.png' to view correlation matrix

Testing – Factors of Successful Movies:

1. Download ‘movies.csv’ from ‘Movie Industry’
    a. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
2. Type ‘python success-factors.py’ into terminal
3. View average root mean squared error for 300 iterations of our linear regressor for each comparison of factors.
    a. For our linear regression model, we calculated the root mean squared error to see how our model performed on our test sets as compared to the training sets. We ran it many times with different inputs to ensure consistency.
    b. Our algorithm seems to be most accurate when comparing budget vs rating, year vs rating, and runtime vs rating, as these root mean square error values are all around 1, meaning it can make reasonably good predictions when comparing these factors.
    c. However, our algorithm does not seem to be a good fit for comparing the other variables, as the root mean squared error for all the other factors are around 40,000,000 when run multiple times. This may be due to the extreme variability of our dataset which makes it difficult to predict accurate results.

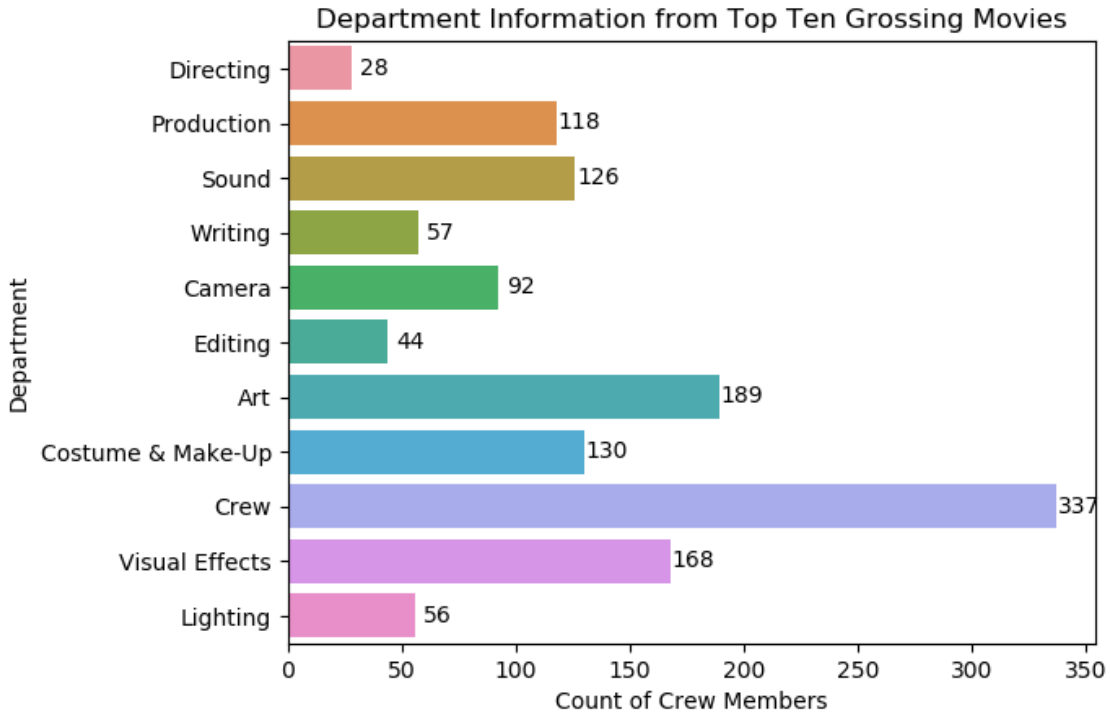
- **What qualities do crew members of successful movies have?**

The data for this question was joined by the two datasets from Kaggle: ‘The Movies Dataset’ and ‘Movie Industry.’ Only the data from the top ten grossing movies from ‘Movie Industry’ were used in these graphs. ‘The Movies Dataset’ provided data on crew members, and was primarily the reason for the joining of the two datasets.

For this question, two bar graphs were made. The first bar graph dealt with counting the total number of crew members per crew member type. A crew member type was

defined by a crew member's department. Departments include art, music, etc. All the crew members were a part of one of the top ten grossing movies, though the movies were not significant in this graph. This graph aims to compare which types of crew members were most prevalent THROUGHOUT the top ten grossing movies.

Total Number of Crew Members per Department:

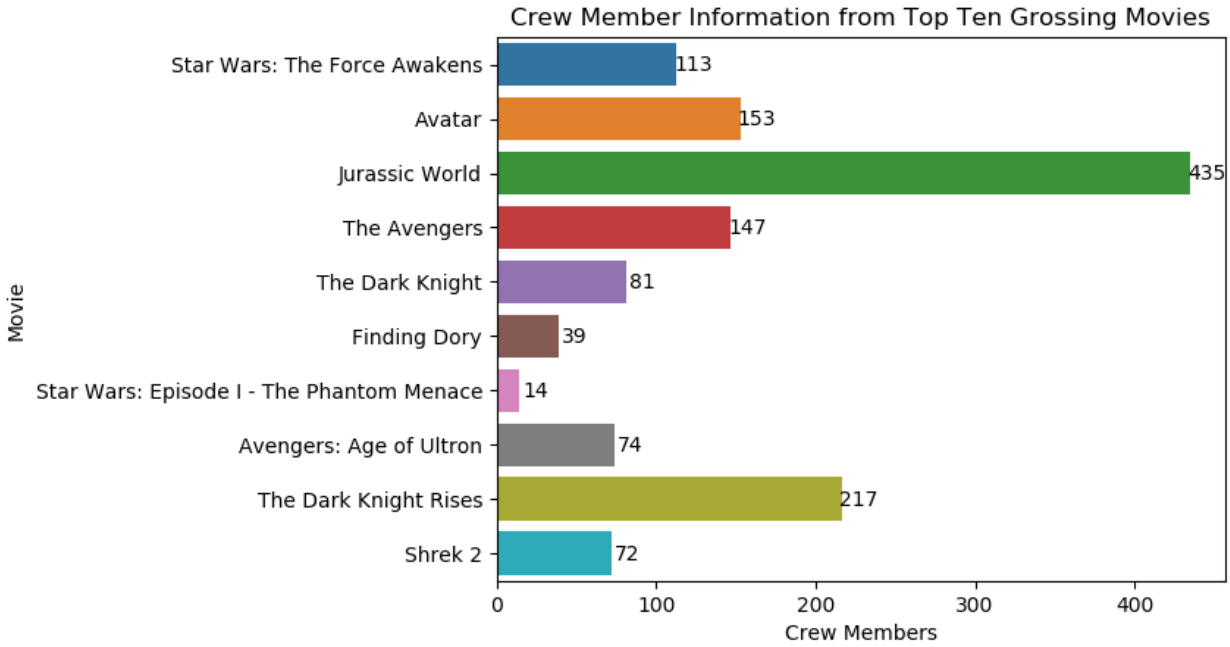


In the dataset, it never explicitly stated what 'crew' meant in the 'credits.csv' file. We interpreted 'crew' as people who were uncategorized for some reason. Looking past that, we can see that some notable counts of types include art and visual effects, with counts of 189 people and 168 people respectively. It would make sense to assume that the more people there is for the role, the more important that role has in making the movie. However, this is not necessarily true. For example, directors are VERY important in the direction of a movie, yet there are only 28 total directors from the top ten grossing movies.

There are some interesting occurrences that we would not have expected from this graph. We expected sound to be one of the highest counts. The top ten grossing movies that we included had movies like Avengers and Star Wars Episode 1, which had very memorable soundtracks. Also, we found it strange that there were 28 total directors from 10 movies… That would mean that there would be an average of 3 directors per movie, rounded up.

The second bar graph dealt with the total number of crew members per movie. Again, the data was from the top ten grossing movies. This time however, the movie has significance in the comparison of crew members. This graph aims to show which movies had the most crew members.

Total Crew Members per Movie:



In this graph, we can see that Jurassic World had the most crew members by far with 435 total members. This is followed by The Dark Knight Rises with 217 members and Avatar with 153 members. This was not apparent in the graph, but the movies are sorted in descending order by gross income. Star Wars: The Force Awakens had the highest gross revenue, yet it was nowhere near the highest total number of crew members in the top ten grossing movies. This is surprising to me, because we would think that the more crew members working a movie, the more polished it would be and more money it would make as a result. A very strange result we can see from this graph is Star Wars: Episode 1 having only 14 total crew members. A quick Google search tells us that this is obviously not correct.

When computing this data, I did notice a glaring flaw with 'The Movies Dataset.' We noticed that the ID values of the 'credits.csv' file had dates in the ID values rather than numbers. This required some additional code in removing the rows with incorrect ID values. We did the computations in 'csv_maker.py.' Perhaps the source of 'The Movies Dataset' had some faulty data within. This could potentially be the reason for interesting values, such as the 28 directors or the 14 crew members working on Star Wars: Episode 1.

Reproducing the Results - Crew Members

1. Download 'credits.csv' and 'movies_metadata.csv' from 'The Movies Dataset,' and 'movies.csv' from 'Movie Industry'
    a. The Movies Dataset: https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv
    b. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
2. Run 'crew_dept_info.py' in the terminal

    a. This generates two bar graphs
    b. This also generates a smaller .csv file called ‘crew_info.csv’ that is used for testing
3. Open the files ‘department_graph.png’ and ‘crew_graph.png’ to see the results

Testing - Crew Members

1. Download ‘credits.csv’ and ‘movies_metadata.csv’ from ‘The Movies Dataset,’ and ‘movies.csv’ from ‘Movie Industry’
    a. The Movies Dataset: https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv
    b. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
2. Run ‘crew_dept_info.py’ for the needed ‘crew_info.csv’ file
3. Run ‘test_crew_full.py’ in the terminal
    a. The testing file samples ‘crew_info.csv,’ using only one row chosen at random from the top ten grossing movies
    b. ‘test_util.py’ is used in this file for the ‘assert_equals’ method
    c. The row represents one movie; this testing file checks to see if the totaled number of crew members from each department matches up with the total number of crew members of the movie

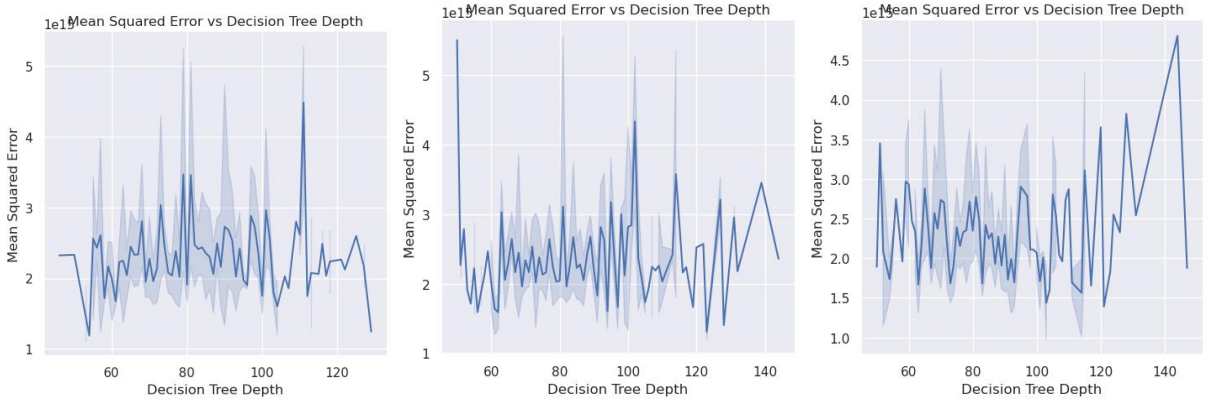
- **What is the optimal depth for training a DecisionTreeRegressor to predict gross movie revenue?**

A DecisionTreeRegressor model was fitted to the Kaggle dataset in order to predict the gross revenue of a movie given the following features:

- The movie's total production budget
- The name of the production company. This value is originally a string, but was converted to a float dummy value.
- The production country
- Name of the director
- Name of the movie genre
- The movie's rating
- Total runtime
- Imdb user rating
- Name of the lead actor
- Name of the script writer
- Year that the move was released.

A test size of 0.2 was used so that there the model could be verified for its accuracy on unlabeled data. The program constructed and trained 300 DecisionTreeRegressors using the data, using the "best" option for splitting. This was chosen because there were so many features, so only the most relevant features should be used to split the data. Each DecisionTree was trained on a random sample of 1000 elements from the data frame. This was to lower the time it took to train each model and added variance to each training set. A Pandas Dataframe was constructed with mean squared error and model depth as columns. The depth of each tree was plotted on the X axis, and its mean square error was plotted against the y axis. The graph is shown below. The line

graph was constructed by seaborn relplot. A line graph was used because the accuracy score vs model depth represents a continuous function. Each time the program was run, the graph would vary slightly due to the random nature of constructing DecisionTreeRegressors. However, the overall patterns of the graphs remained the same throughout. Here are graphs generated from 3 runs:



The graph yields the following observations:

- There is not a very clear correlation between the mean squared error and the depth of the model. This is unexpected, since one would expect that increasing the ability for the model to fit the data should lower its mean squared error across all parts of the graph. By our expectation, the graph should've been a downward sloping graph. An explanation for this unexpected behavior is that increasing the depth of the tree may cause it to over fit the data. By overfitting the data, the model is built too closely to the training data, and noticing correlations that don't actually exist in the real world. This causes the model to lose accuracy when testing against the unlabeled data. Therefore, increasing the depth of a tree may increase its accuracy by fitting the training data better, but may also reduce accuracy by overfitting the data. This causes the unexpected observation that depth is not clearly correlated to mean squared error.
- Sometimes, the mean squared error has a sudden, massive fluctuation at a particular value compared to its neighbors. It is unexpected that the model would lose so much accuracy compared to its neighbors, since the depths are similar. In this explanation, we will consider 125 to be the layer value at which a massive fluctuation occurs. One possible explanation for the sudden loss of accuracy near the 125th layer is that the model splits itself in a completely different way for each layer count. Since a decision tree regressor uses different splitting values for each new model built, it may be that all trees with 125 layers were split with similar values at all layers, while all trees with 124 or 126 layers were split in a completely different way. Thus, since trees at 125 layers are completely different from trees at 124 or 126 layers, it could be that those trees at 125 happen to be split in such a way that the model is fitted to the training data in a very inaccurate way, while trees at 124 and 126 are fitted to the data in a very accurate way.

Reproducing the Results – Predicting Gross Revenue

1. Download ‘movies.csv’ from ‘Movie Industry’

    a. Movie Industry: https://www.kaggle.com/danielgrijalvas/movies
2. Type "python movie-success-predictor.py" into the terminal
3. Open "movie-success-prediction.png" to view the graph
4. Run the graph multiple times if you wish to observe multiple iterations. Please note that the graphs won't perfectly match with those shown above, but the general correlation is the same.

Testing – Predicting Gross Revenue

1. The program for constructing the machine learning algorithm was first tested by constructing a single model to see if it predicts values in a reasonable range. Then, the mean squared error was added to make sure that it gives a reasonable value for the error. The errors shown in the graph are very large due to the nature of the mean squared error equation. The mean squared error scales based on the values of the inputs, so it is expected that the errors will have a higher scale.
2. The graphs were tested by running it multiple times to ensure that the observations were not a one-time phenomena.

Work Plan Evaluation

Our work plan predictions were fairly accurate and we were able to finish all of our tasks ahead of time. The dates were set at reasonable intervals so that we had enough time to finish our tasks while also being able to have room before the deadline. We also used the Ed workspace, which was a good decision since it allowed us to collaborate online without going through the process of setting up a git repository.

Collaboration

We did the project work on our own. We did not ask for any help on the assignment.

Conclusion

Most of the data we worked with had some satisfactory results. Examples include finding the top five ratings on average each country produced and finding the most popular departments that crew worked in. However, at the same time, some weird results were produced. An example is the 14 lone crew members working on Star Wars: Episode I. We learned that big datasets often have faulty results because of processing large amounts of data. This is a lesson to research more on the datasets and where they spawned from. Despite the faults, we learned a lot about movies and what makes them successful. Our data visualizations and machine learning models that we created all serve to help us understand successful movies and what drives them to become what they are.