MOVIE RECOMMENDER ENGINE

Introduction

Recommender Systems (RS) keep growing in popularity in the world of machine learning and predictive modelling. It is ubiquitous and is used in virtually every industry. Due to its direct impact on the bottom line, RS has been embraced by most of the major corporations in the world. Wal-Mart and other supermarkets use it to predict the purchase activities of their customers. This allows for a more convenient shopping experience, which also translates into more shopping on the consumer's part and consequently more revenue to the business. Online stores such as Amazon, Ebay and Etsy use RS to recommend products to buyers. Pandora and Spotify use it to recommend music from specific genres to listeners based on their listening history. LinkedIn, a professional networking site, uses RS to recommend individuals, recruiters, and advertisers to one other. LinkedIn reported that about 50% of the total job applications and job views by individuals were as a result of their recommendation model. The other business areas which have adopted the RS approach include social networking sites (Facebook, Instagram etc.), search engines, restaurants, insurance companies, online dating among countless others.

Currently, one major area where RS benefits the vast majority of Americans is in movie recommendations. Netflix, an entertainment company, employs over 300 people, spending a staggering $150 million annually, just to generate movie recommendations for viewers based on their previous preferences and movie-watching history. For a company like Netflix, accurate movie recommendations are critical since the company gets most of its revenue from customer subscriptions. The company uses a "one free month trial" policy to reel in most of their new customers. During the registration, they ask each new customer to rate at least five movies from different genres, which provides them with customer-specific data. One major challenge is that at any given time, there are millions of movies for a customer to choose from. Therefore, the movie selection process can be a very tedious exercise for the average consumer. Often, customers will cancel their subscription after a few failed attempts at finding the right movie to watch. Therefore, to prevent loss of revenue through subscription cancellations, Netflix helps its customers to locate movies they would like to watch using their RS. Hence, there is a direct correlation between the accuracy of their RS and their income generation. It therefore came as no surprise when in 2008, Netflix organized a competition with a million-dollar price tag for the winner who will improve their existing RS by at least 10%.

Some of the business questions which we will seek to answer include the following:

Given a user's history or movie preferences, which movie is the user likely to be interested in?

Should we adopt a binary class approach (recommend or not recommend) or a multi-class approach (ratings - 1, 2, 3, 4 and 5)

If there is more than one movie to recommend, what ranking system should be used to determine the order of the movie list?

Data Wrangling Techniques - A Movie Recommender System

The birth of the Motion Picture Camera in the late 18th century gave birth to possibly the most potent form of entertainment in existence: Cinema. Movies have managed to enthral audiences ever since one second clips of racing horses emerged in the 1890s to the introduction of sound in the 1920s to the birth of colour in the 1930s to mainstream 3D Movies in the early 2010s.

Cinema had humble origins in terms of plot, direction and acting (mainly due to its extremely short duration in its early days) but since then, movie industries around the world have been blessed with creative geniuses in the form of directors, screenwriters, actors, sound designers and cinematographers. It has also spread itself into a plethora of genres ranging from romance to comedy to science fiction to horror.

As with almost every kid born in the last century, I was amazed by movies. I was addicted to it. And I've always wanted to know more about the enigmatic world of cinema. In this notebook, I will try and gain some insights using data. With us, we have a dataset of about 45000 movies with metadata collected from TMDB. Using this data, we will try and answer various questions that I've always had about movies.

Features

Adult: Indicates if the movie is X-Rated or Adult.
Belongs_to_collection: A thingified dictionary that gives information on the movie series the Particular film belongs to.
Budget: The budget of the movie in dollars.
Genres: A thingified list of dictionaries that list out all the genres associated with the movie. Homepage: The Official Homepage of the move.
Id: The ID of the move.
Imdb_id: The IMDB ID of the movie.
Original_language: The language in which the movie was originally shot in.
Original_title: The original title of the movie.
Overview: A brief blurb of the movie.
Popularity: The Popularity Score assigned by TMDB.
Production_companies: A thingified list of production companies involved with the making of the movie.
Production_countries: A thingified list of countries where the movie was shot/produced in.
Release_date: Theatrical Release Date of the movie.
Revenue: The total revenue of the movie in dollars.
Runtime: The runtime of the movie in minutes.
Spoken_languages: A thingified list of spoken languages in the film.
Tagline: The tagline of the movie.
Title: The Official Title of the movie.
Video: Indicates if there is a video present of the movie with TMDB.
Vote_average: The average rating of the movie.
Vote_count: The number of votes by users, as counted by TMDB.

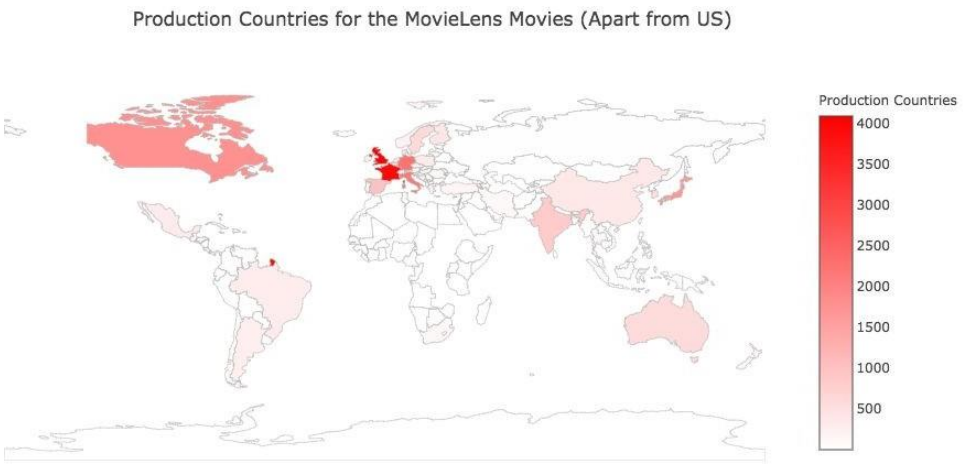
Some of the data wrangling techniques used are mentioned below as follows:

- Data Set was downloaded in 4 different subsets which were then merged together for analytical reasoning.
- Null values were searched and discarded to avoid errors on the dataset.
- Duplicate values were first highlighted, cross checked with other columns for difference and then discarded to avoid errors in statistical modeling.
- Data was then plotted using a scatter plot and a bar chart to check for outliers in ratings for different movies along with the user.

No outliers were found post wrangling method and later values were checked using the .describe() function in the pandas library. Exploratory Data Analysis was done post successful wrangling.

EXPLORATORY DATA VISUALIZATION AND ANALYSIS

In this section, the various insights produced through descriptive statistics and data visualization is presented.



1. The Movies in the dataset are overwhelmingly in the English Language and shot in the United States of America.
2. Europe is also an extremely popular location with the UK, France, Germany and Italy in the top five.
3. Japan and India are the most popular Asian countries when it comes to movie production.

Franchise Movies

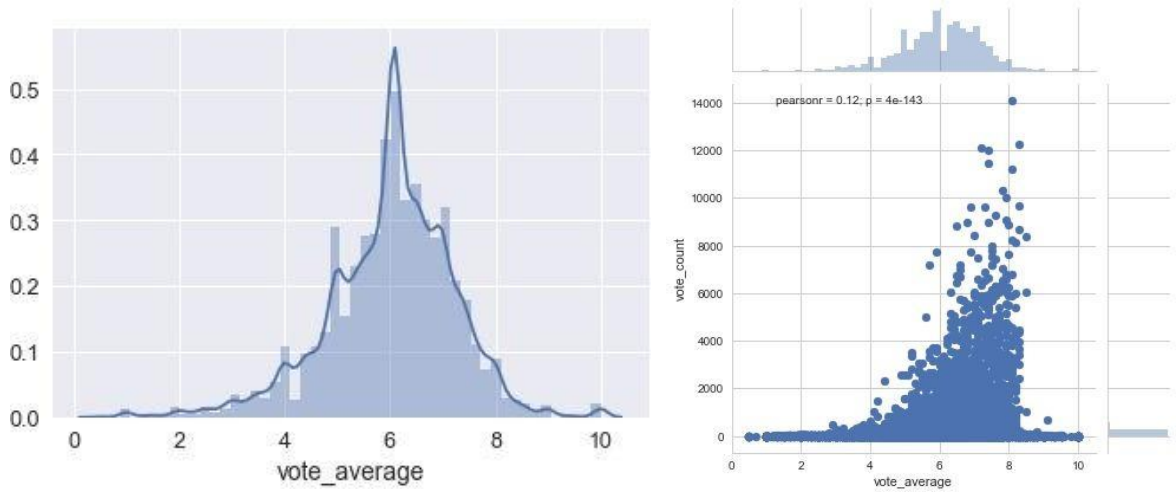
1. The Harry Potter Franchise is the most successful movie franchise raking in more than 7.707 billion dollars from 8 movies.
2. The Star Wars Movies come in a close second with a 7.403 billion dollars from 8 movies too.
3. The Avatar Collection, although just consisting of one movie at the moment, is the most successful franchise of all time with the sole movie raking in close to 3 billion dollars.
4. The James Bond Movies is the largest franchise ever with over 26 movies released. Friday the 13th and Pokémon come in at a distant second and third with 12 and 11 movies respectively.

Production Companies

Warner Bros is the highest earning production company of all time earning a staggering 63.5 billion dollars from close to 500 movies. Universal Pictures and Paramount Pictures are the second and the third highest earning companies with 55 billion dollars and 48 billion dollars in revenue respectively.

Pixar Animation Studios has produced the most successful movies, on average. This is not surprising considering the amazing array of movies that it has produced in the last few decades: Up, Finding Nemo, Inside Out, Wall-E, Ratatouille, the Toy Story Franchise, Cars Franchise, etc. Marvel Studios with an average gross of 615 million dollars comes in second with movies such as Iron Man and The Avengers under its banner.

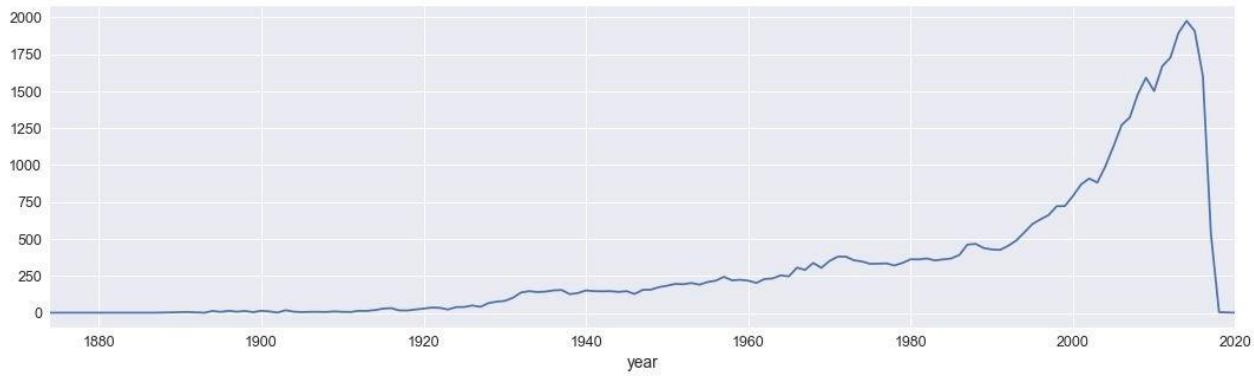
Popularity, Vote Average and Vote Count



1. Minions is the most popular movie by the TMDB Popularity Score. Wonder Woman and Beauty and the Beast, two extremely successful woman centric movies come in second and third respectively.
2. Inception and The Dark Knight, two critically acclaimed and commercially successful Christopher Nolan movies figure at the top of The Most Voted On Movies Chart.
3. The Shawshank Redemption and The Godfather are the two most critically acclaimed movies in the TMDB Database. Interestingly, they are the top 2 movies in IMDB's Top 250 Movies list too. They have a rating of over 9 on IMDB as compared to their 8.5 TMDB Scores.
4. Surprisingly, the Pearson Coefficient of the two aforementioned quantities is a measly 0.097 which suggests that there is no tangible correlation. In other words, popularity and vote average and independent quantities. It would be interesting to discover how TMDB assigns numerical popularity scores to its movies.
5. There is a very small correlation between Vote Count and Vote Average. A large number of votes on a particular movie does not necessarily imply that the movie is good.

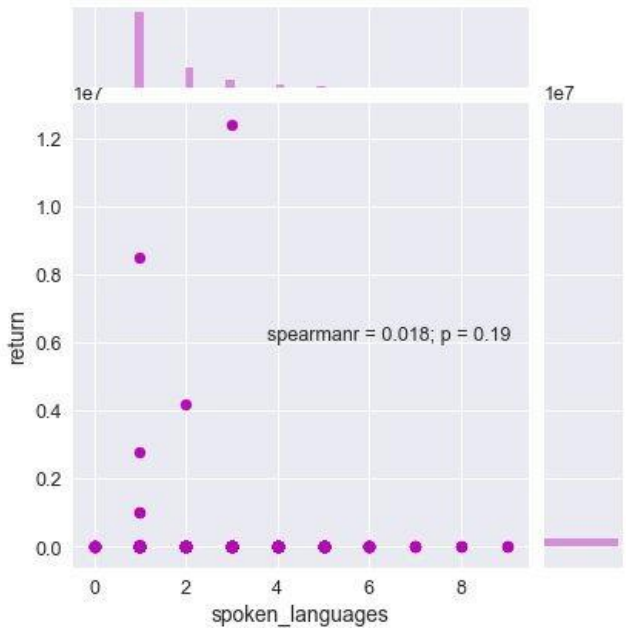
Release Day

Friday is clearly the most popular day for movie releases. This is understandable considering the fact that it usually denotes the beginning of the weekend. Sunday and Monday are the least popular days and this can be attributed to the same aforementioned reason.



The oldest movie, Passage of Venus, was a series of photographs of the transit of the planet Venus across the Sun in 1874. They were taken in Japan by the French astronomer Pierre Janssen using his 'photographic revolver'. This is also the oldest movie on both IMDB and TMDB.

Spoken Languages

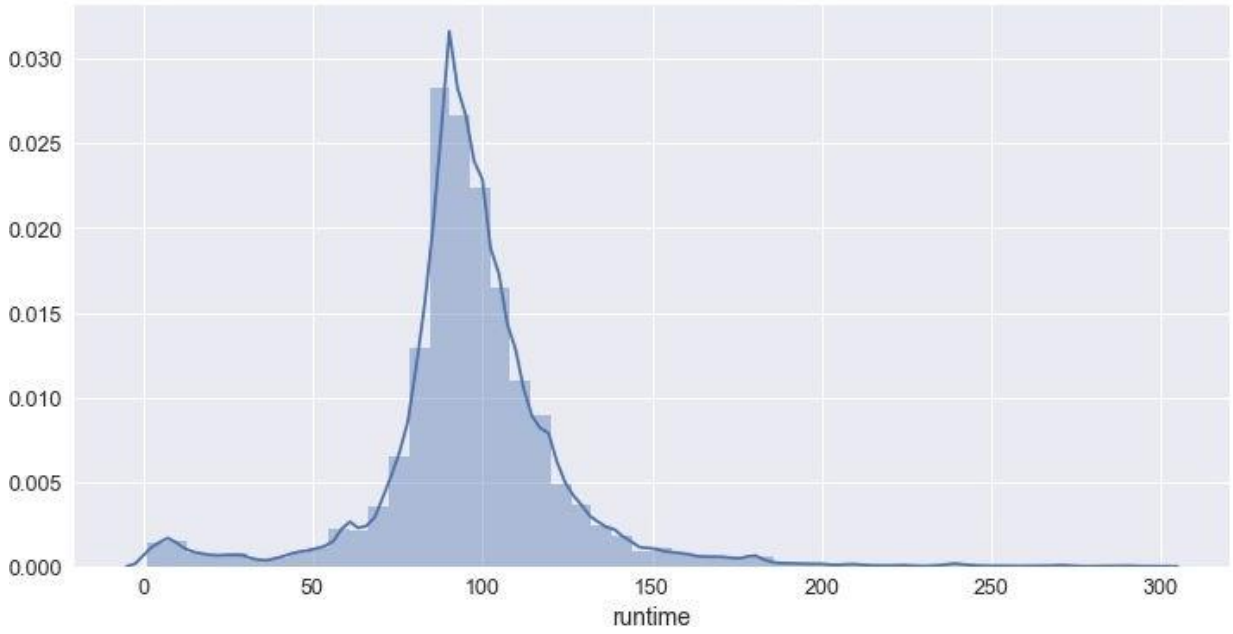


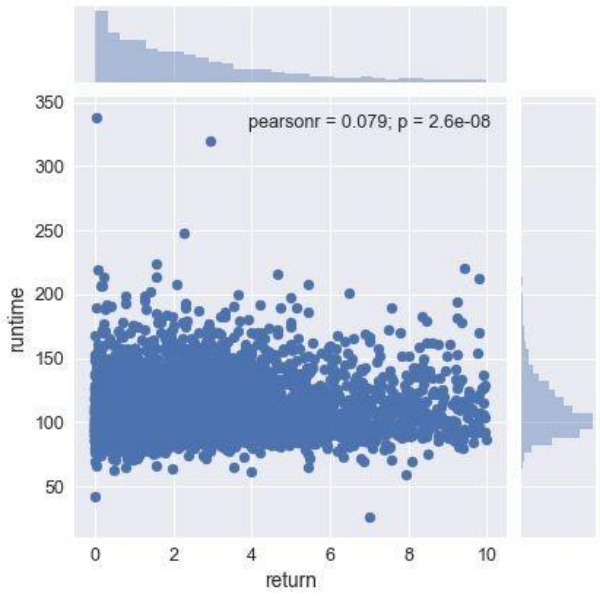
The movie with the greatest number of languages, Visions of Europe is actually a collection of 25 short films by 25 different European directors. This explains the sheer diversity of the movie in terms of language.

There is no correlation between the number of languages and returns of a movie.

Runtime

The average length of a movie is about 1 hour and 30 minutes. The longest movie on record in this dataset is a staggering 1256 minutes (or 20 hours) long.

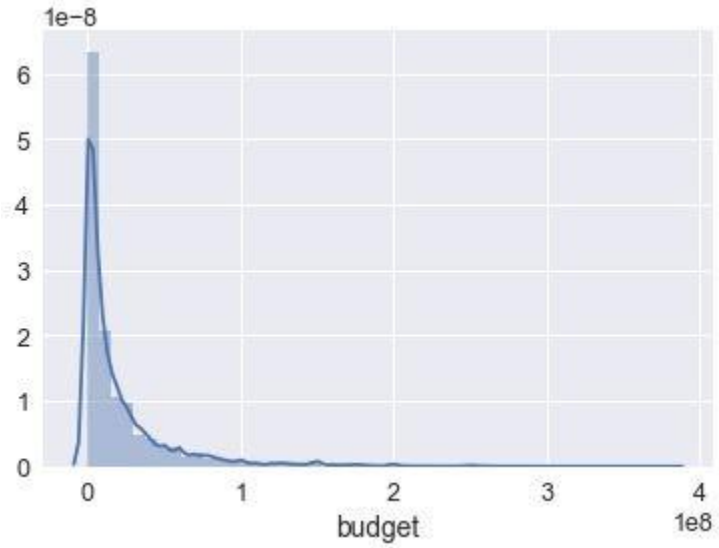




There seems to be no relationship between runtime and return. The duration of a movie is independent of its success.

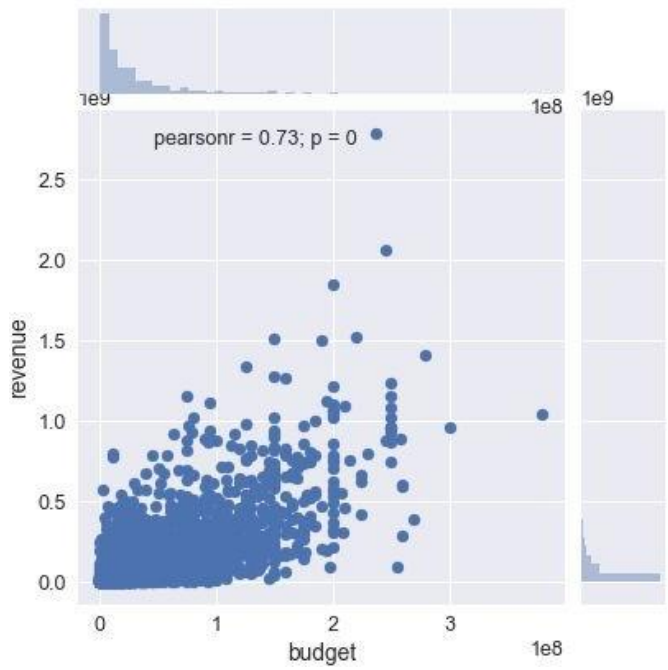
We notice that films started hitting the 60 minute mark as early as 1914. Starting 1924, films started having the traditional 90 minute duration and has remained more or less constant ever since.

Budget



The distribution of movie budgets shows an exponential decay. More than 75% of the movies have a budget smaller than 25 million dollars.

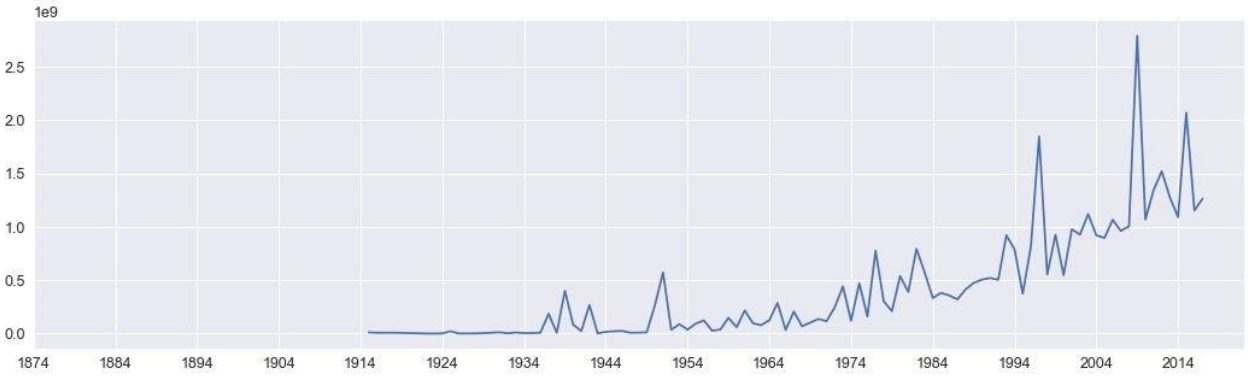
Two Pirates of the Caribbean films occupy the top spots in this list with a staggering budget of over 300 million dollars. All the top 10 most expensive films made a profit on their investment except for The Lone Ranger which managed to recoup less than 35% of its investment, taking in a paltry 90 million dollars on a 255-million-dollar budget.



The pearson r value of 0.73 between the two quantities indicates a very strong correlation.

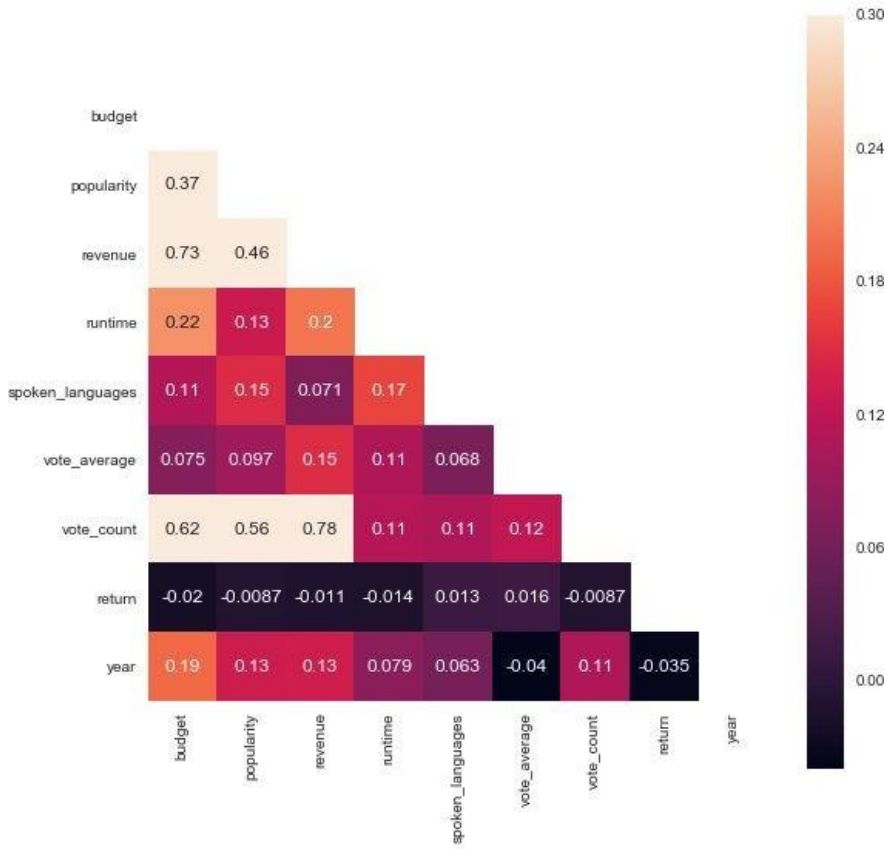
Revenue

The mean gross of a movie is 68.7 million dollars whereas the median gross is much lower at 16.8 million dollars, suggesting the skewed nature of revenue. The lowest revenue generated by a movie is just 1 dollar whereas the highest grossing movie of all time has raked in an astonishing 278 billion dollars.

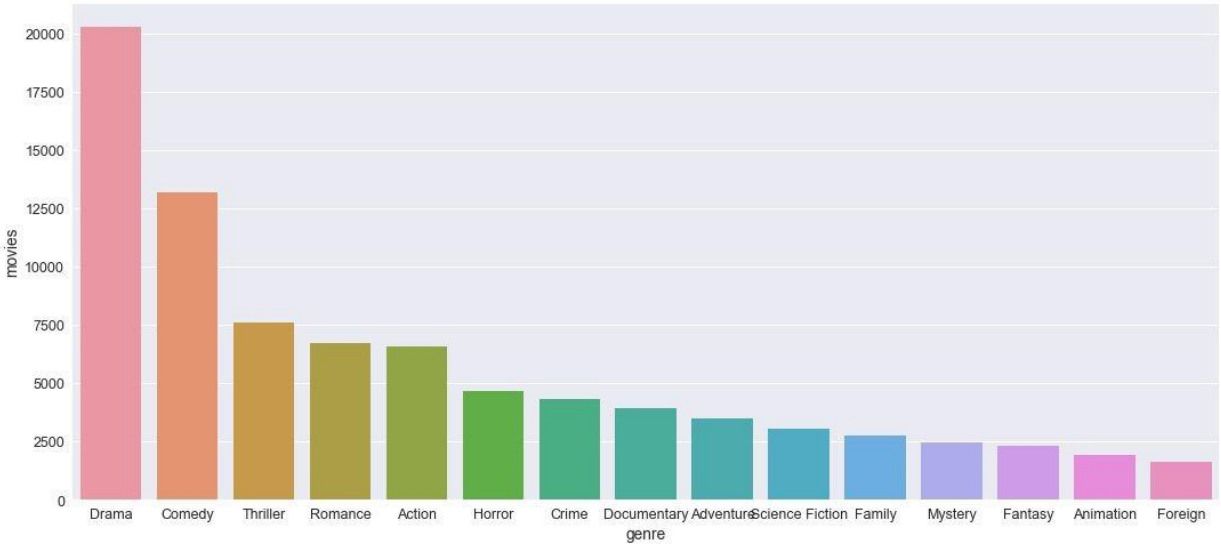


As can be seen from the figure, the maximum gross has steadily risen over the years. The world of movies broke the 1-billion-dollar mark in 1997 with the release of Titanic. It took another 12 years to break the 2-billion-dollar mark with Avatar. Both these movies were directed by James Cameron.

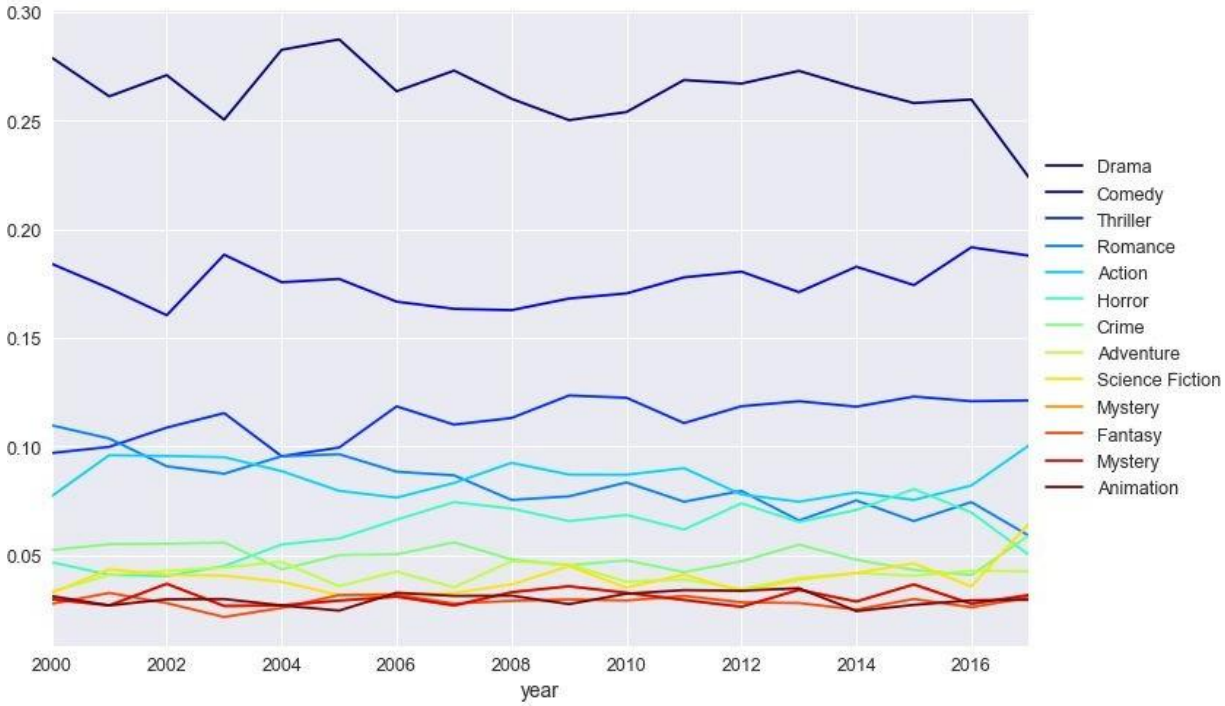
Correlation Matrix



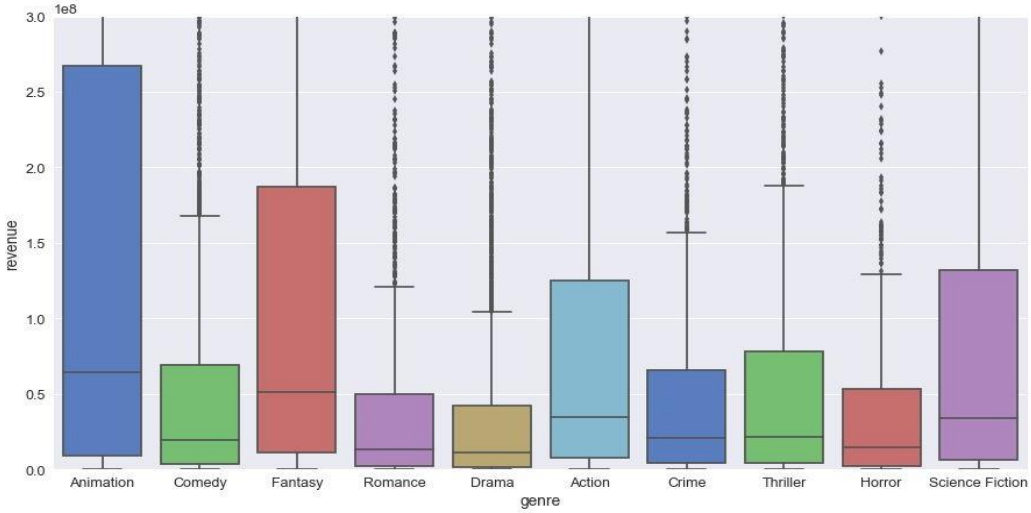
Genres



Drama is the most commonly occurring genre with almost half the movies identifying itself as a drama film. Comedy comes in at a distant second with 25% of the movies having adequate doses of humor. Other major genres represented in the top 10 are Action, Horror, Crime, Mystery, Science Fiction, Animation and Fantasy.

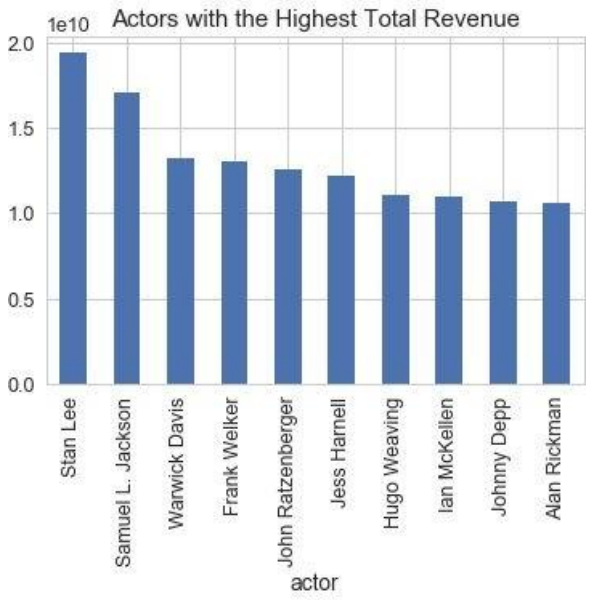


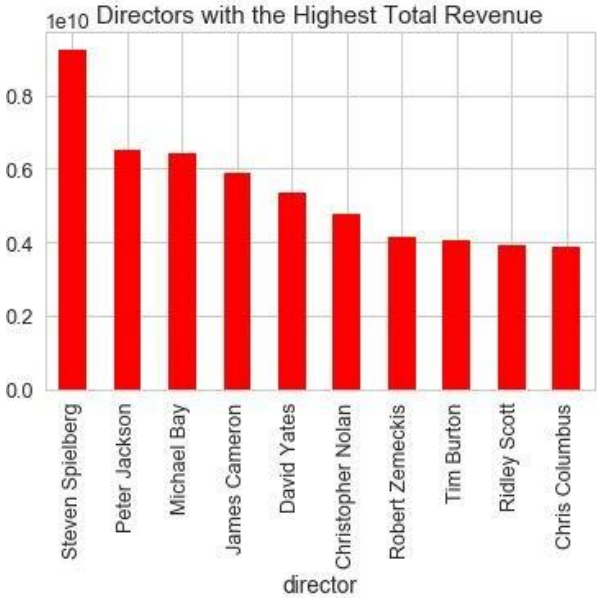
The proportion of movies of each genre has remained fairly constant since the beginning of this century except for Drama. The proportion of drama films has fallen by over 5%. Thriller movies have enjoyed a slight increase in their share.

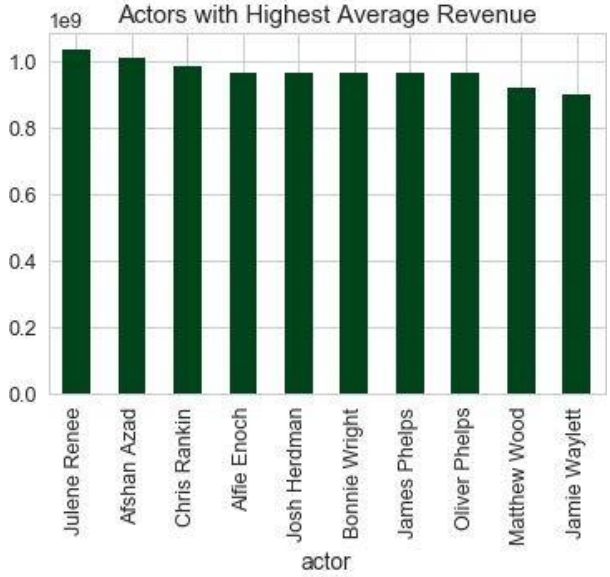


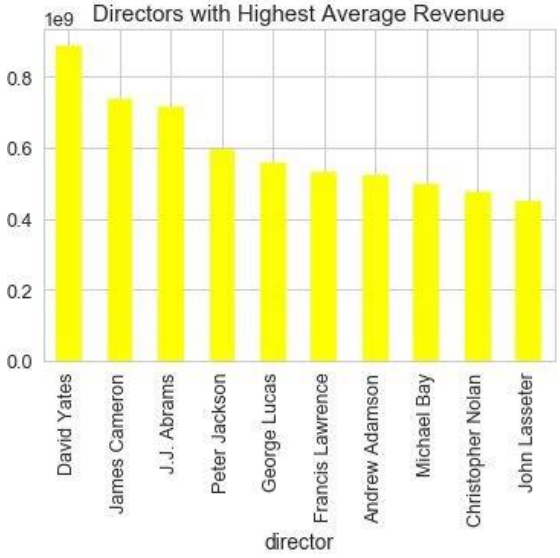
Animation movies has the largest 25-75 range as well as the median revenue among all the genres plotted. Fantasy and Science Fiction have the second and third highest median revenue respectively.

Cast and Crew









REGRESSION: PREDICTING MOVIE REVENUES

Predicting Movie Revenues is an extremely popular problem in Machine Learning which has created a huge amount of literature. Most of the models proposed in these papers use far more potent features than what we possess at the moment. These include Facebook Page Likes, Information on Tweets about the Movie, YouTube Trailer Reaction (Views, Likes, Dislikes, etc.), Movie Rating (MPCAA, CBIFC) among many others.

To compensate for the lack of these features, we are going to cheat a little. We will be using TMDB's Popularity Score and Vote Average as our features in our model to assign a numerical value to popularity. However, it must be kept in mind that these metrics will not be available when predicting movie revenues in the real world, when the movie has not been released yet.

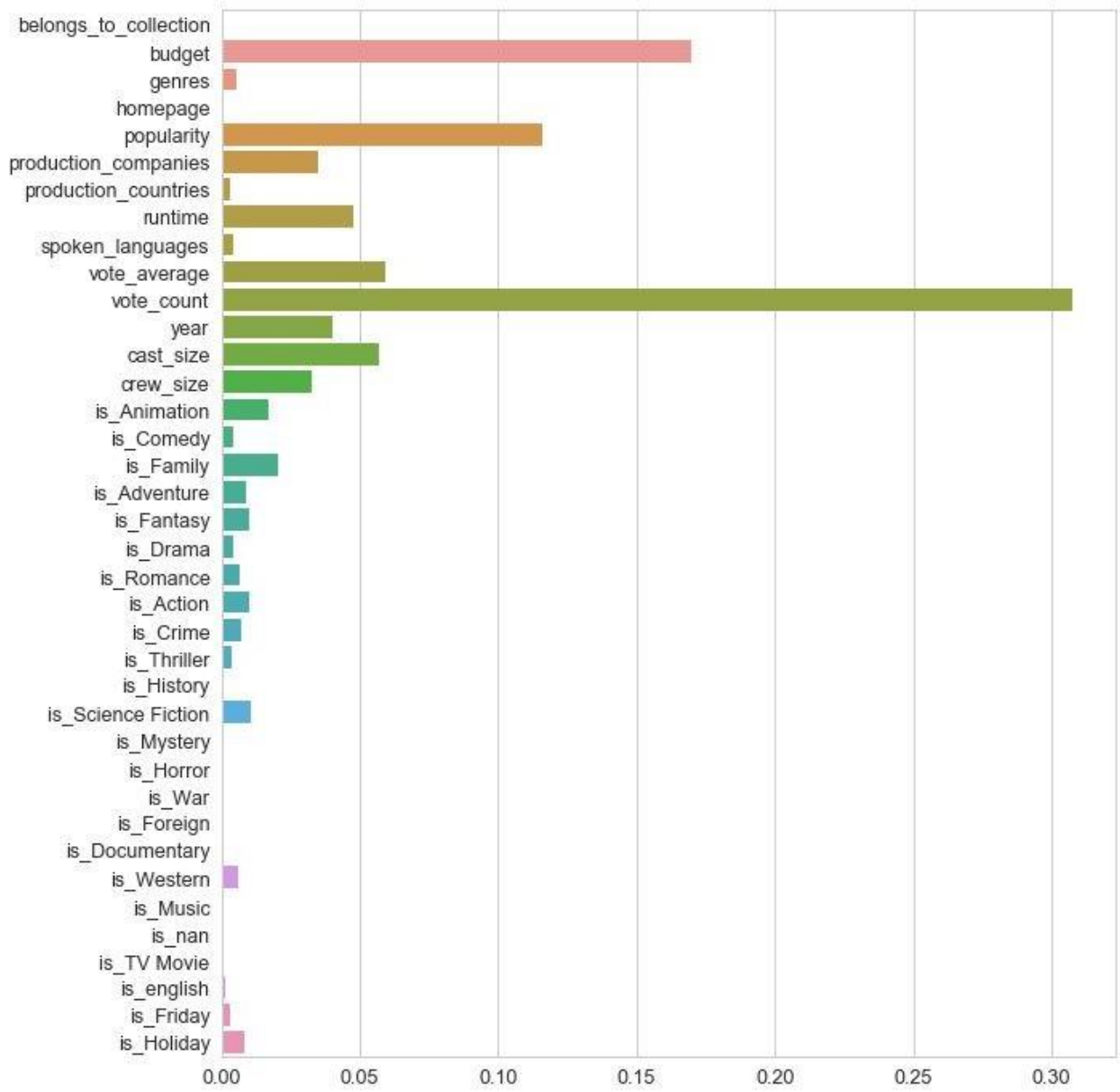
Feature Engineering

1. belongs_to_collection will be turned into a Boolean variable. 1 indicates a movie is a part of collection whereas 0 indicates it is not.
2. Genres will be converted into number of genres.
3. Homepage will be converted into a Boolean variable that will indicate if a movie has a homepage or not.
4. original_language will be replaced by a feature called is_foreign to denote if a particular film is in English or a Foreign Language.
5. production_companies will be replaced with just the number of production companies collaborating to make the movie.
6. production_countries will be replaced with the number of countries the film was shot in.
7. Day will be converted into a binary feature to indicate if the film was released on a Friday.
8. Month will be converted into a variable that indicates if the month was a holiday season.

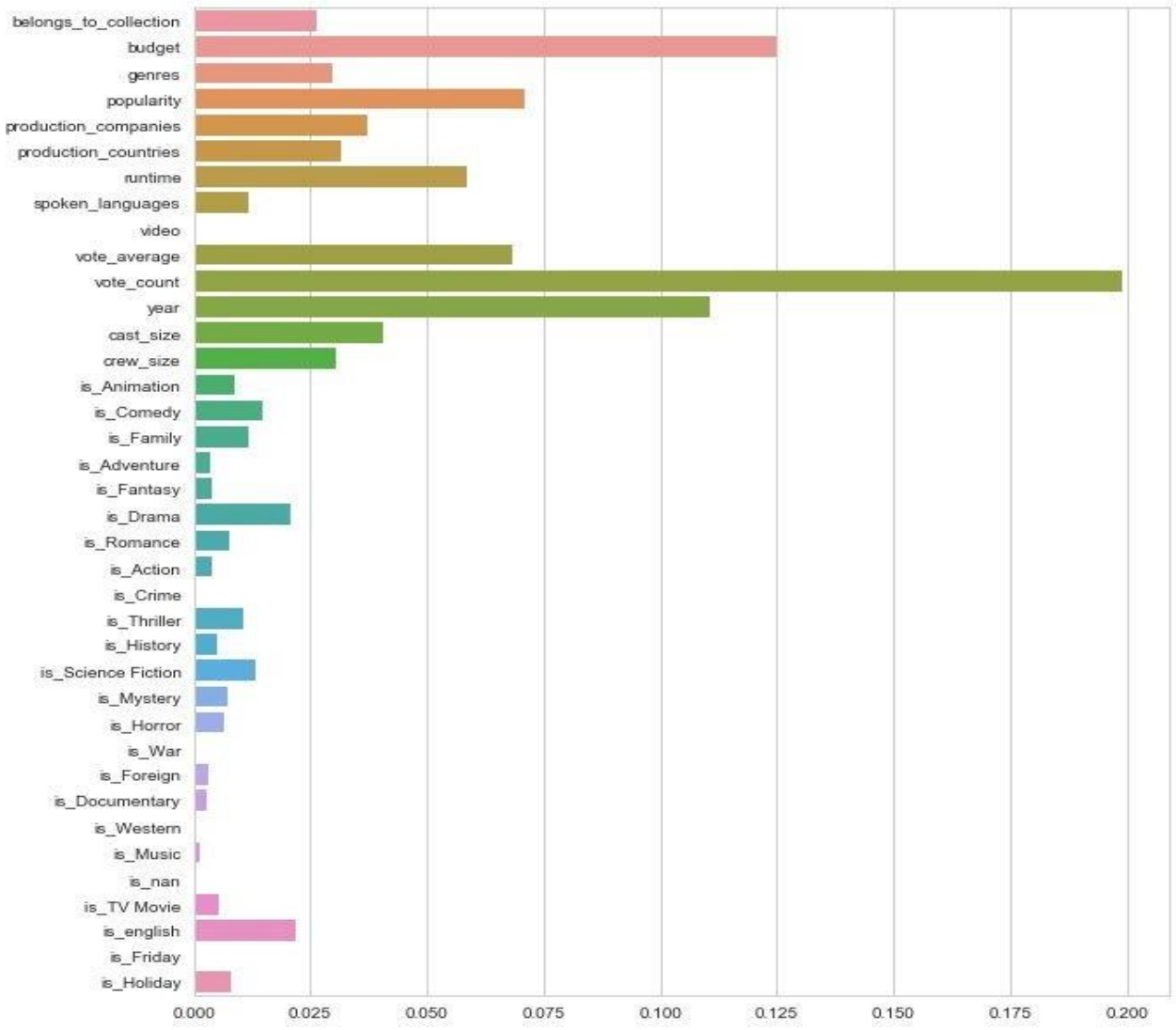
Model

The model that I choose for regression is the Gradient Boosting Regression. The Coefficient of Determination Score obtained by the regressor was 0.78

Feature Importance



We notice that vote_count, a feature we cheated with, is the most important feature to our Gradient Boosting Model. This goes on to show the importance of popularity metrics in determining the revenue of a movie. Budget was the second most important feature followed by Popularity (Literally, a popularity metric) and Crew Size.



CLASSIFICATION: PREDICTING MOVIE SUCCESS

The Classification model uses the same Feature Engineering steps as those followed by the Regression Model built in the previous section.

Model

The model that I choose for classification is the Gradient Boosting Classifier. The model showcased an accuracy of 80% with unseen test cases.

With this, we will conclude our discussion on the classification model and move on to the main part.

RECOMMENDATION ENGINE

The next step was to build a classifier to train the data on and then test its performance against the test data. With all the feature engineering already done in the previous step, applying machine learning was a fairly concise step.

Content Based Recommender

My approach to building the recommender was extremely hacky. What I did was create a metadata dump for every movie which consisted of genres, director, main actors and keywords. I then used a Count vectorizer to create a count matrix. I then calculated the cosine similarities and returned movies that are most similar.

I also added a mechanism to remove bad movies and return movies which are popular and have had a good critical response.

I took the top 25 movies based on similarity scores and calculate the vote of the 60th percentile movie. Then, using this as the value of m , I calculated the weighted rating of each movie using IMDB's formula like I did with the Simple Recommender.

```
In [27]: get_recommendations('The Dark Knight').head(10)

Out[27]: 7931                       The Dark Knight Rises
         132                              Batman Forever
         1113                             Batman Returns
         8227    Batman: The Dark Knight Returns, Part 2
         7565                Batman: Under the Red Hood
         524                                      Batman
         7901                          Batman: Year One
         2579              Batman: Mask of the Phantasm
         2696                                       JFK
         8165    Batman: The Dark Knight Returns, Part 1
         Name: title, dtype: object
```

CONCLUSION

This report highlighted the processes of data wrangling, inferential statistics, data visualization, feature engineering and predictive modelling performed on the Movies Dataset. All the results and insights gained as part of these processes were also highlighted. With these insights, a Gradient Boosting Regressor and Classifier were built to predict Movie Revenue and Success respectively with a Score of 0.78 and 0.8 respectively.

In addition, a recommendation engine was built based on similar idea and algorithm:

1. Content Based Recommender: We built two content-based engines; one that took movie overview and taglines as input and the other which took metadata such as cast, crew, genre and keywords to come up with predictions. We also devised a simple filter to give greater preference to movies with more votes and higher ratings.

The code associated with this report is available at:
Monish_Sareen/movie_recommender_engine