Capstone Project Milestone Report: Predicting Movie Success

Problem: The entertainment industry is a high-profile industry with movie producers often investing millions of dollars into making movies. The success of a movie can depend on several factors such as budget, cast, genre, release date, critical reception etc. Understanding the key drivers and predicting movie success can therefore be extremely useful in making key business decisions in the pre-production, production and distribution stages. For example, predicted movie success can be used to determine compensation of cast and crew, marketing as well as other aspects of the budget not yet determined.

Client: Movie producers and investors can help mitigate potential losses and investment risk by better analyzing and predicting movie success. Movie studios can also use this analysis to determine marketing, budgeting and creative decisions.

Dataset: "The Movies Dataset" on Kaggle contains data (cast, crew, plot keywords, budget, revenue, posters, release dates, IMDB rating, languages, production companies etc.) for 45,000 movies. In addition, CSV files of the 50 highest grossing actors and 35 highest grossing directors of all-time was exported from IMDB to add to the analysis.

Approach: I will use supervised learning algorithms to determine the key drivers of movie success and predict the success of a movie in 3 different ways:

a) Predicting *revenue* using a regression model,
b) Predicting *categorical movie success* (i.e. blockbuster, break-even, flop etc.) using a classification model, and
c) Predicting *return-on-investment* (i.e. profit/budget) using regression in order provide more actionable analysis for producers and investors.

Developing 3 slightly different prediction models will enable me to better assess the robustness and accuracy of each model.

List of files in dataset:

File	Source	Description	Columns included/excluded in analysis
credits.csv	The Movie Database (TMDB)	Contains cast and crew information	All columns (ID, Cast, and Crew) were included in the analysis.
movies_metadata.csv	The Movie Database (TMDB)	Contains most of the data on the movies from The Movie DataBase	Columns with extraneous information such as homepage url, poster_path, tagline etc. were excluded. See next section for list of columns included.

highest_grossing_directors.csv	IMDB	List of 35 highest grossing directors of all time	'Name' [of director]
highest_grossing_directors.csv	IMDB	List of 50 highest grossing actors of all time	'Name' [of actor]

Data Cleaning/Wrangling Steps:

Converting csv files to pandas DataFrames and retaining relevant columns:

1. The first step was to export the 'credits.csv' and 'movies_metadata.csv' files to pandas DataFrames, **credits** and **movies**, respectively.
2. The **credits** DataFrame includes 3 columns: 'id', 'cast', and 'crew', all of which were kept. Below is a list of columns kept from the **movies** DataFrame:
   'adult', 'budget', 'genres', 'id', 'original_language', 'release_date', 'revenue', 'runtime', 'status', 'title', 'vote_average', 'vote_count', 'collection'

Dealing with missing data, null values, and errors in data capture:

3. Duplicate rows were dropped from both DataFrames.
4. Next, a high-level view of the data was obtained using .info() and .describe() to note columns with null values or incorrect data types.
   a. For example, the 'budget' column in **movies** was of type 'object' instead of a numerical type, indicating the data was stored as strings. Similarly, the 'id' column values in **movies** were also stored as strings while the same movie 'id' data in the credits DataFrame was stored as integers. These discrepancies were resolved using pd.to_numeric().
   b. It was also noted that the 'budget' and 'revenue' columns in **movies** (now both of numeric type) had a large number of 0 values. Specifically, up to 75% of the values in these columns were 0s. Upon closer inspection, it was concluded that while these values were not identified as null, they did indicate missing data. Thus, the 0 values in both columns were tagged as null (np.nan) using the replace function and all rows with missing budget or revenue data were dropped. Similarly, any rows with missing values in the 'runtime', 'vote_average', and 'vote_count' columns were are also dropped.
   c. The 'belongs_to_collection' column in **movies** indicated >4000 null values, which indicated that the associated movies did not belong to a collection or franchise. In other words, these movies did not have a prequel or sequel. The data in this column were dictionaries stored as strings. First, the data was converted to type dict using .ast_literal_eval() and the collection name was extracted from the dictionary (e.g. 'Toy Story Collection' for all the Toy Story movies) into another column titled 'collection'. Any null values in this 'collection' column were replaced

with 0s to tag the movies that did not belong to a collection. The original 'belongs_to_collection' column was then dropped.

5. Once null and missing values had been appropriately handled, it was noted that the budget and revenue columns in **movies** still contained several single-digit values, which likely indicates errors in data capture. This was confirmed by cross-referencing a few of these values with their actual revenue and budget values manually. Since most movies have a budget of at least $10,000, all movies with a budget below $10,000 (n = 59) were dropped from this dataset to ensure that all erroneous 'budget' data was removed. After the application of this filter, there were still several movies with single-digit revenues. A second filter was the applied to the 'revenue' column to ensure the removal of any erroneous revenue data. Any movies with a revenue lower than $1500 (n = 57) were dropped, leaving 5,283 movies in the **movies** DataFrame.
6. All the movies in the 'adult' column of movies were categorized as 'False' so this column was dropped since it was not adding any additional insight.

Merging the two DataFrames:

7. Next, the **movies** and **credits** DataFrames were merged on the 'id' column. This added the 'cast' and 'crew' columns from the **credits** DataFrame to the **movies** DataFrame. The new merged DataFrame will henceforth be referred to as **df**.
8. The 'cast', 'crew'', and 'genres' columns in **df** were lists of dictionaries stored as strings so the data in these columns was converted to type 'list'. Empty lists were tagged as null and rows with null values in the 'cast', 'crew' and 'genres' columns were then dropped from **df**.
9. New columns for 'actor' and 'director' were created by extracting the name of first-listed actor from the 'cast' column (assumed to be the lead actor) and the name of the director from the 'director' column.
10. Since each movie is often associated with more than 1 genre, dummy genre variables were created to identify the genres associated with each movie. In other words, each genre was converted into a binary column, with a value of 1 if the movie belonged to that genre and a value of 0 if it did not.
11. The 'status' of a 4 movies was listed as 'rumored' and 'post-production' which likely doesn't include reliable metrics for budget and revenue. Therefore, only movies with a released status were included and the 'status' column was then dropped.

Adding the 'top_actor' and top_director' features:

12. Next, the 'highest_grossing_actors.csv' and 'highest_grossing_directors.csv' files were loaded into pandas DataFrames (**top_actors, top_directors)** and each DataFrame was cross-referenced with the 'actor' and 'director' columns in **df** to create 2 new binary columns: 'top_actor' and 'top_director'. If a movie's main actor or director were in the highest grossing lists, their respective column value would be 1; otherwise, it would be 0.

Adding the release season feature:

13. The ‘release_date’ was converted to type *datetime* and each date was categorized into different release seasons and holidays - Summer, Holiday Season (Thanksgiving, Christmas, New Year’s), Valentine’s Day, Labor Day, and MLK Day). If the date did not fall into any of these aforementioned categories, just the month of the release date was extracted. This information was stored in a new **df** column: ‘release_season’.

Adding additional dependent variables for predicting success (ROI and categorical success):

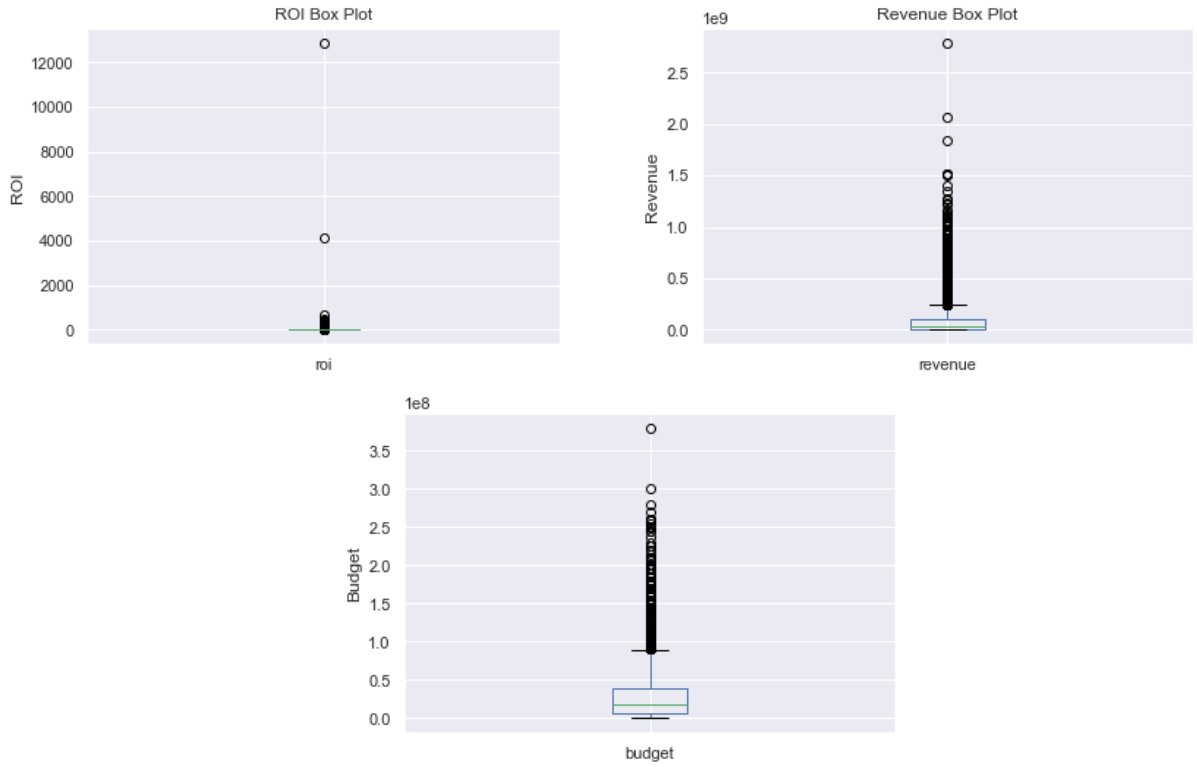
14. A new ‘roi’ column was created to calculate the Return on Investment as (revenue-budget)/budget.
15. The categorical success of a movie (blockbuster, hit, flop etc.) was determined using the ratio of revenue to budget (revenue/budget).

Ratio of Revenue to Budget	**Categorical Success**
>= 5	Blockbuster
>= 2.5 and < 5	Hit
>= 1 and < 2.5	Break-Even
>= 0.25 and < 1	Flop
< 0.25	Disaster

16. Finally, I index of df was reset and a final check was done to ensure the data looked good, was of the right type, and had no null values. The final, cleaned dataframe includes 5,262 movies.

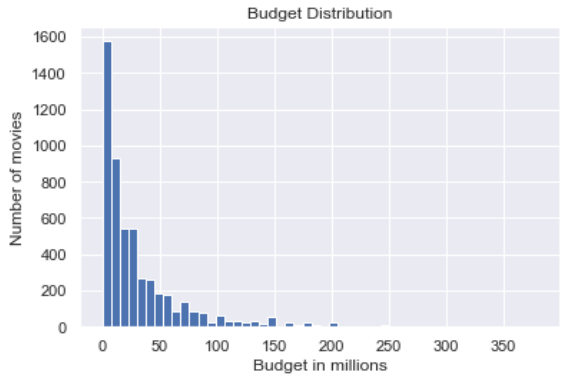
Exploratory Data Analysis

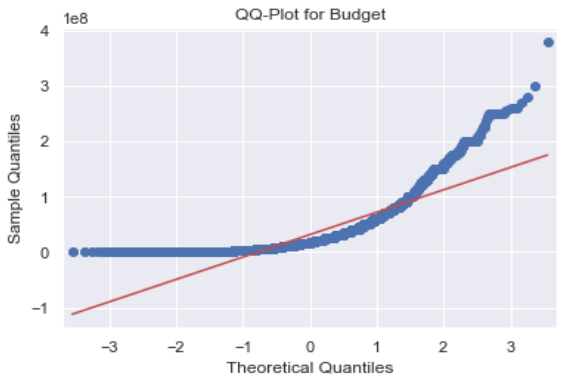
Are there any outliers in the budget, revenue, and ROI numerical data?



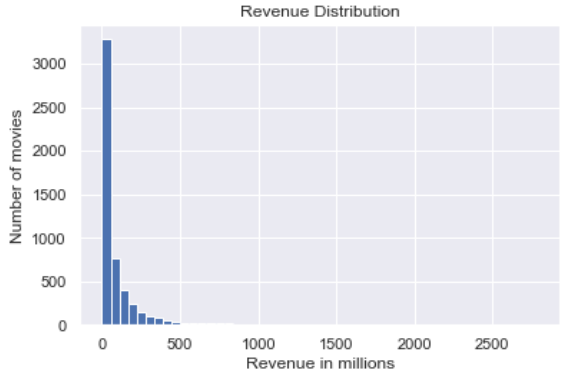
All three sets of data show outliers on the higher end, with the ROI data appearing to have some extremely strong outliers. These extreme outliers in ROI reflective of movies made on smaller budgets that performed exceedingly well at the box-office (e.g. Paranormal Activity).The revenue and budget data also show a large number of outliers on the high end (greater than ~200 M USD and greater than ~100 M, respectively), although the outliers are less prominent than those observed in the ROI data. Since movie revenues and budgets have been increasing significantly over the years (see analysis below), movies with budgets and revenues typically considered as outliers will still be taken into account in the initial predictive model.

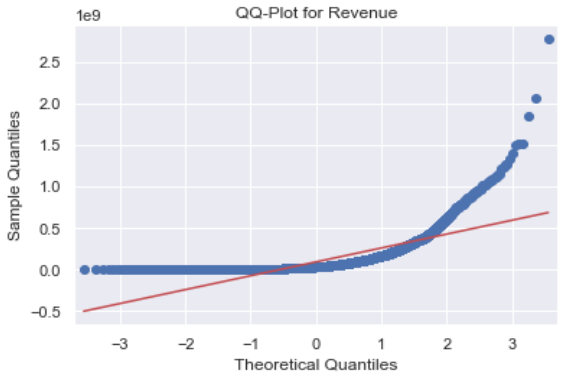
What do the distributions for budget, revenue, and ROI look like? Are the distributions normal?



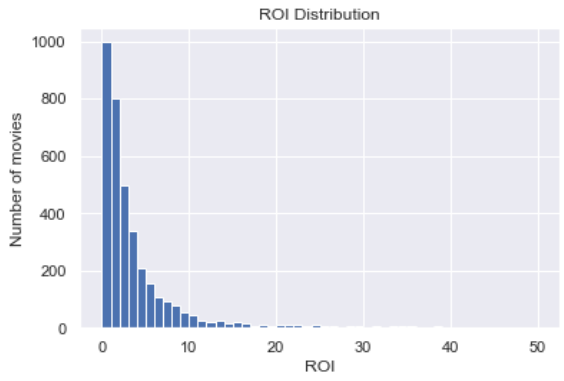


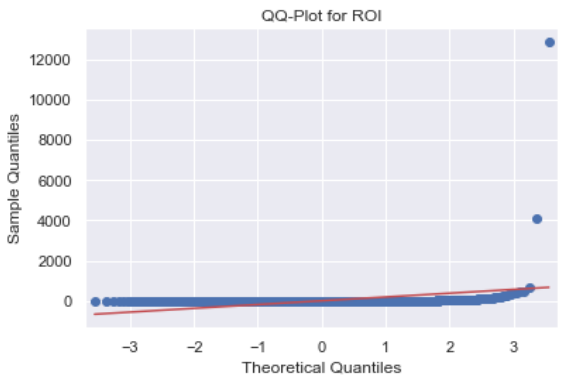
The majority of movies have a budget of less than 50 million USD, with the largest budget group being less than 10 M USD. The distribution is not normal, as confirmed by the quantile-quantile (QQ) plot.





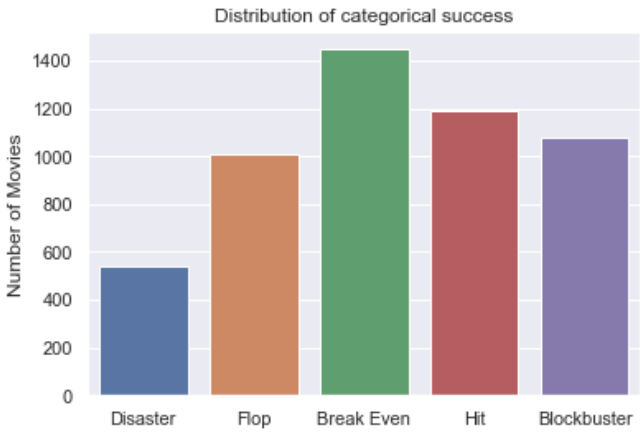
Most movies have a revenue of less than 200 million USD with the majority grossing less than 50-100 million USD. The distribution does not appear to be normal.





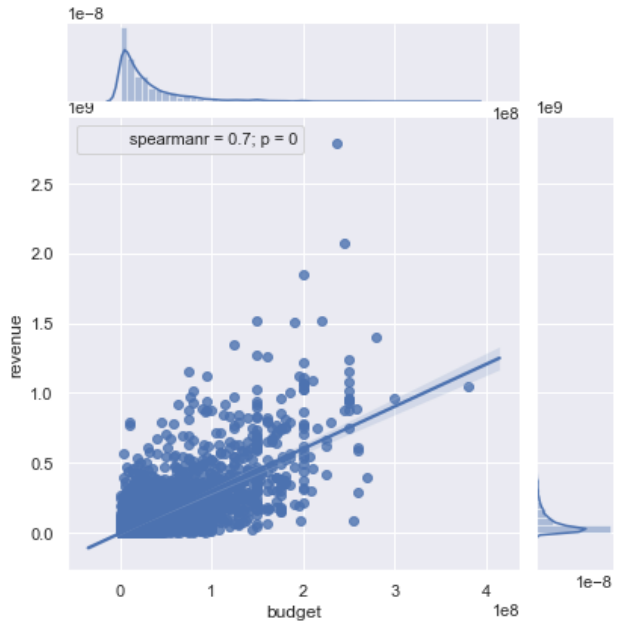
The vast majority of movies have an ROI of less than 5. The distribution is not normal.

What does the distribution of categorical success look like?



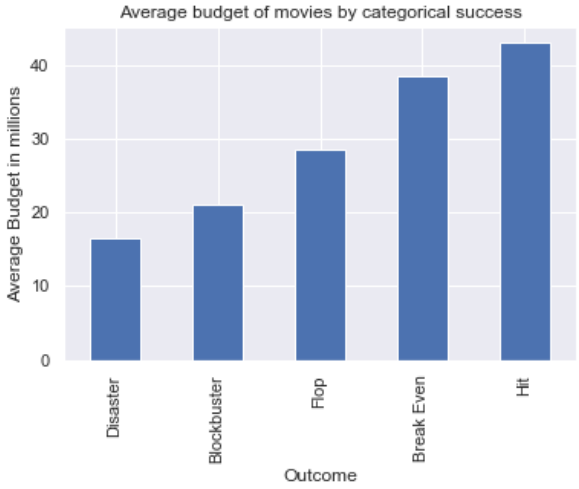
The largest category of movies manage to break-even, with hits, blockbusters, and flops having similar ratios. The smallest group of films are disasters.

What does the relationship between budget and revenue look like?



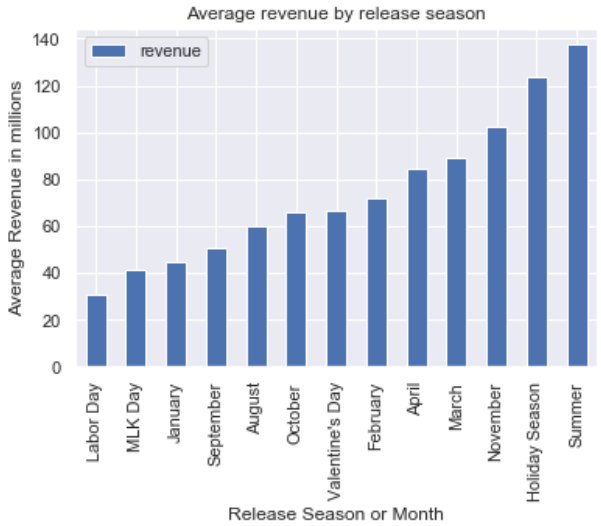
Since both the budget and revenue distributions are not normal, the Spearman's rank correlation can be used to assess the dependency of the two variables. As expected, there appears to be a strong correlation between revenue and budget. Even though the sample distributions are not normal, the Pearson's correlation was found to be quite similar to the Spearman's rank correlation (0.73 vs 0.7). Visual analysis of the scatter plot also shows more scatter at higher budgets.

What is the relationship between categorical success and budget?



The average budget is lowest for movies that are disasters or blockbusters (i.e. extremes on the success scale). This phenomenon can be interpreted in two different ways. Having too low of a budget can affect the quality, reputation, and star power of a movie, thus resulting in significantly lower revenues, which is the case for ‘disasters’. On the other hand, having a lower budget also allows the movie to recover their costs more easily and be categorized as a blockbuster (note that blockbusters are classified as movies that gross 5 times their budget). Whether a lower-budget movie is a disaster or blockbuster likely depends on other factors such as genre, star power, release season etc.

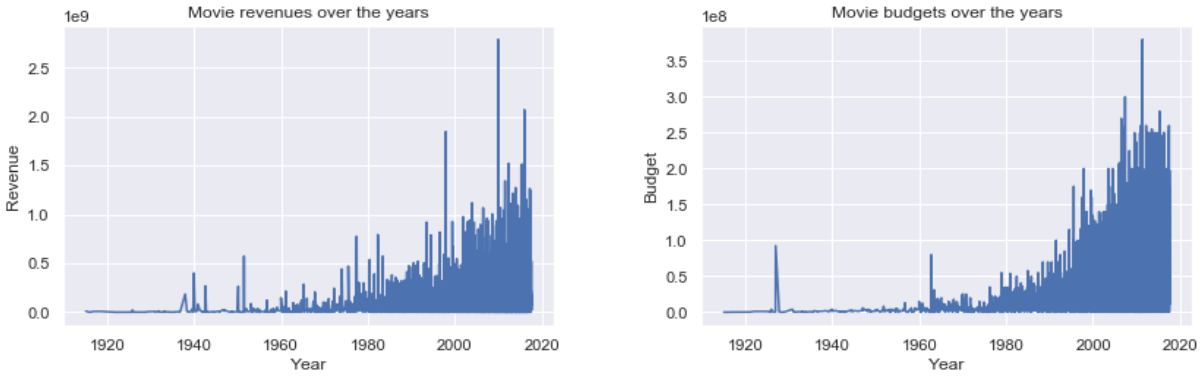
Which release season typically results in the highest movie revenues?



Movies released in the summer make the most money on average followed by the holiday season. Holidays like Labor Day and MLK Day don't appear to have much of an impact. Outside

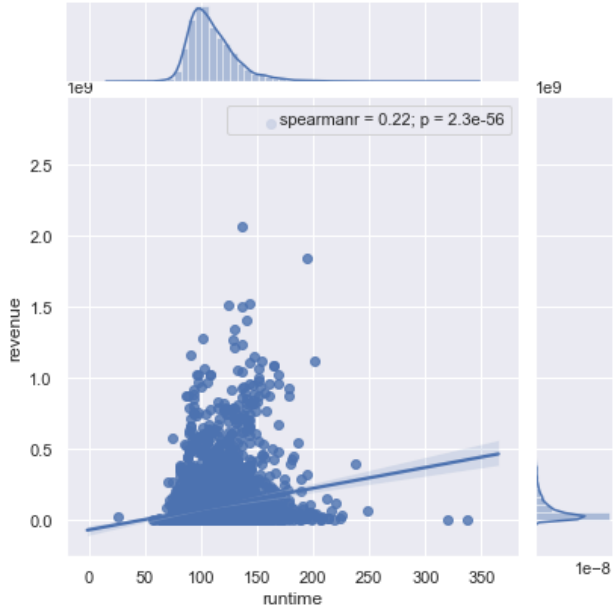
of the holidays and summer, movies that have grossed the highest are typically released in November, March and April.

How has the revenue and budget of movies changed over the years?



Overall, both movie revenues and movie budgets have been steadily increasing since the 1980s, with the budgets seeing a steeper increase since the 2000s. Keeping this in mind, it may be important to keep the outliers in the budget and revenue data since they could be representative of the general trend of higher budgets and revenues.

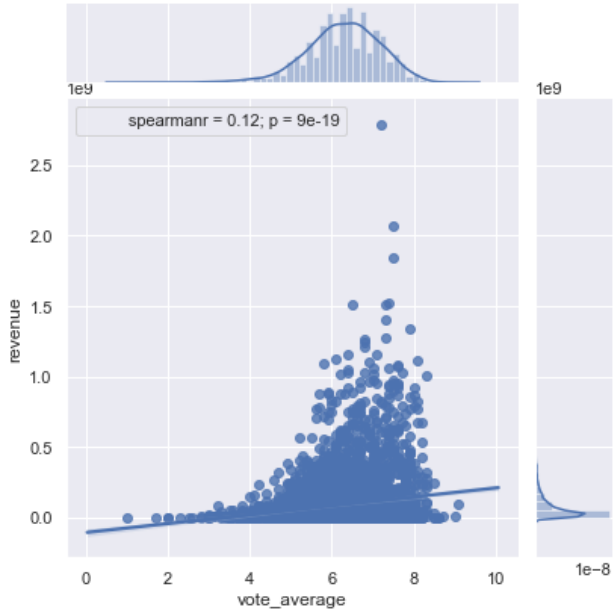
Is there a correlation between runtime and revenue?



Overall, there appears to be a weak correlation between runtime and revenue, based on the Spearman's Rank correlation. Since the average runtime of a movie in our sample is 110 minutes, we can also look at the correlation between shorter movies (less than 2 hours long) and longer movies (more than 2 hours long) with respect to revenue, separately. The Spearman's rank correlation for shorter movies is 0.156 and is statistically significant, indicating

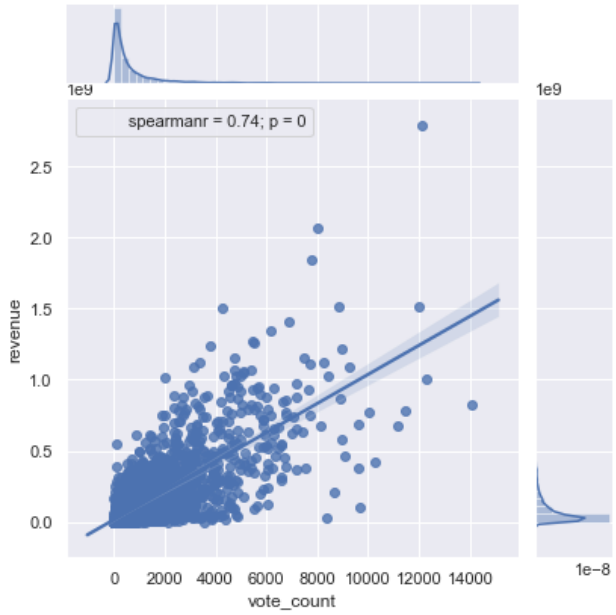
a weak correlation between runtime and revenue for shorter movies. However, the correlation coefficient for longer movies is -0.036 and statistically insignificant. In other words, there does not appear to be a significant correlation for movies longer than 2 hours.

Is there a relationship between the average vote rating on The Movie Database (TMDB) and revenue?



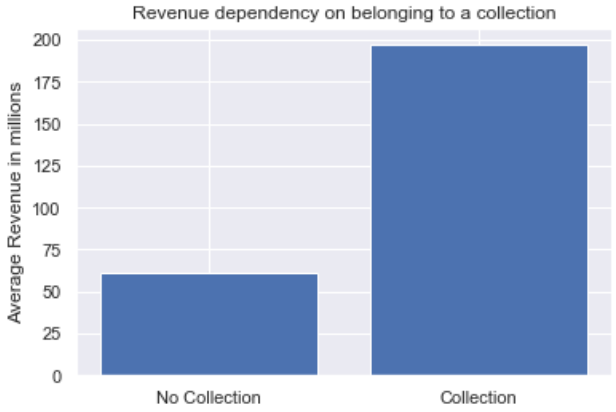
Overall, the Spearman's rank correlation coefficient is 0.12, indicating a weak positive correlation between the average vote rating and revenue. Visual analysis of the scatter plot suggests that movies with an average vote rating of 5 or less seem to be more closely associated with lower revenues. In this region (n = 433 movies), the Spearman's rank correlation coefficient is 0.195. The majority of movies, however, have an average vote rating above 5 (n=4829) but the relationship between revenue and rating appears to be even weaker in this region. This is confirmed by the Spearman's rank correlation coefficient which is 0.059, indicating a weaker correlation for movies with higher ratings.

Is there a relationship between the number of votes on The Movie Database (TMDB) and revenue?



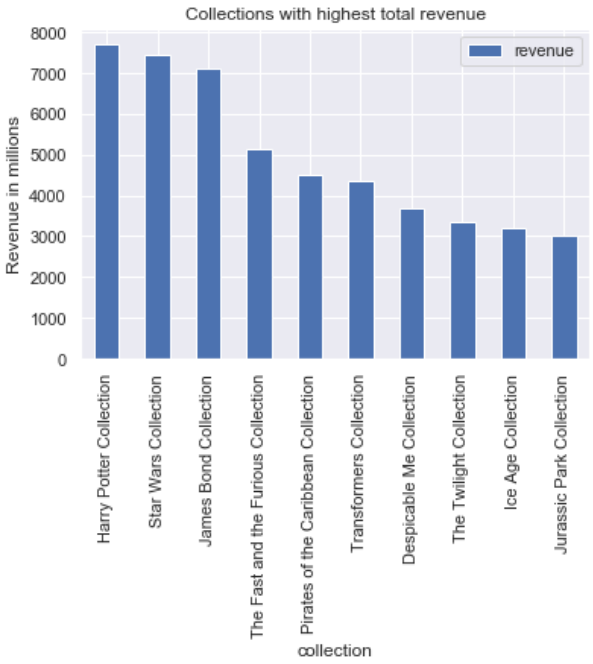
Interestingly, there is a much stronger positive correlation (coefficient = 0.74) between the number of votes and revenue, with higher vote counts on TMDB typically being associated with higher revenue. This indicates that the number of votes might be a better predictor than the average vote rating, indicating that people tend to vote more for movies that perform better at the box office. Visual analysis of the scatter plot shows significantly more scatter for vote counts greater than ~7000.

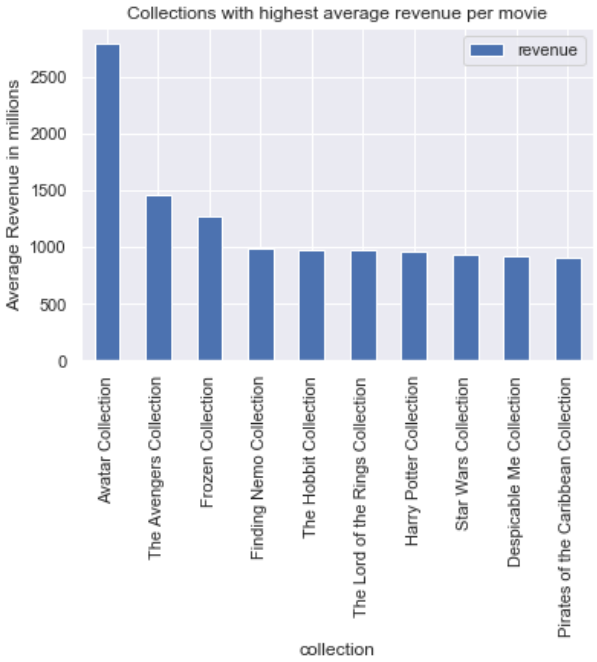
Do movies that belong to a collection typically gross higher than movies that do not? Is there a statistically significant difference in mean revenue between movies that belong to a collection and movies that don't?



Movies that belong to a collection (i.e. franchise) gross, on average, almost 4 times higher than movies that do not belong to a collection. Statistically, this difference in mean revenue is significant as observed by the t-test, which resulted in a p-value close to 0 indicating that the means for these 2 groups are not equal.

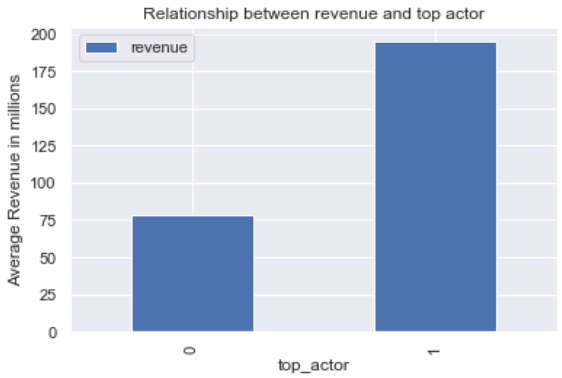
What are the top 10 collections with the highest total revenue and highest average revenue?

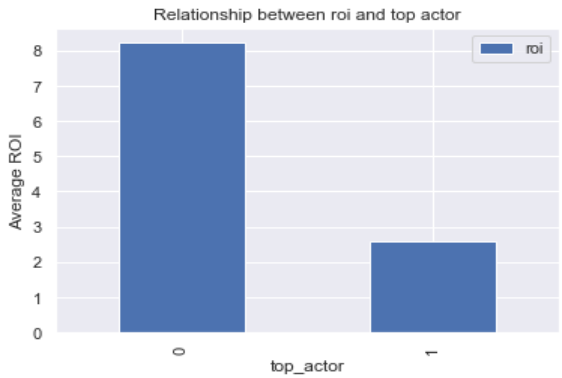




The Harry Potter, Star Wars, and James Bond collections have the highest total revenue. The Avatar collection have the highest average movie revenue, followed by the Avengers and Frozen. However, it should be noted that some collections (such as Avatar and Frozen) currently only have one movie released so the average revenue is equal to the revenue of the associated movie. Sequels are currently in the works for these movies, which is why they have been categorized as collections.

Does having a top actor as the lead have an impact on movie revenue and/or ROI? Are these observed differences in average revenue/ROI between movies with top actors and movies without top actors statistically significant?





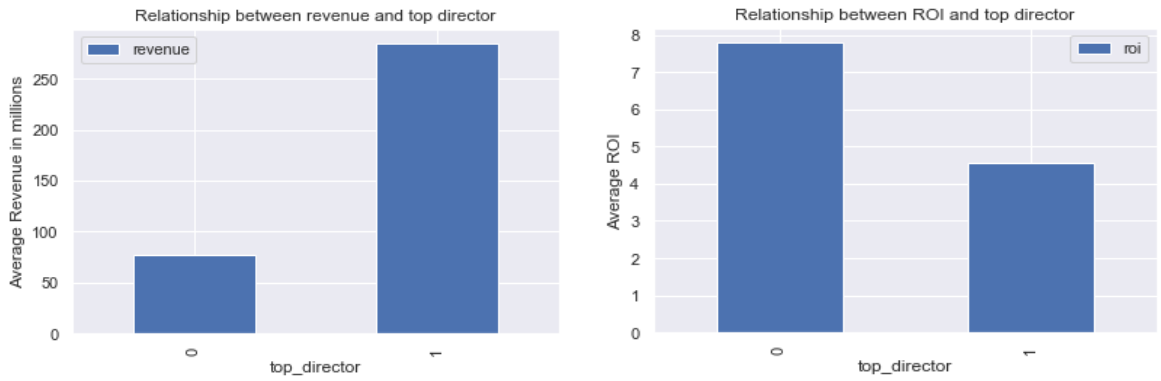
A top actor is defined as any actor appearing on highest grossing actors list on IMDB. While having a top actor as a lead does result in a much higher revenue on average (~200M vs 50M USD), it also

results in a lower ROI (~75% decrease), likely due to the higher compensation top actors demand. The difference in revenue is statistically significant, as confirmed by the t-test and bootstrapping, both of which resulted in a p-value close to 0. To validate this analysis even further, the non-parametric Mann-Whitney U test was also used since the revenue data does not follow a normal distribution. This test also concluded that the revenue distributions for both samples (movies with top actors and movies without top actors) are statistically different.

Determining statistical significance for the difference in ROI was trickier due to the presence of more extreme outliers in the ROI data. The t-test failed to reject the hypothesis that the mean ROI for movies without top actors as leads is equal to the mean ROI for movies with top actors at leads. On the other hand, the bootstrap approach did reject this hypothesis and concluded that the mean ROIs are **not** equal for the two samples. Since the ROI data has some extreme outliers, it is likely that the robustness and accuracy of the t-test was significantly compromised (it performs best for normal data with little to no outliers). The bootstrapping approach is likely to be more accurate in in this case.

However, due to the presence of extreme outliers, the mean is probably not the best test statistic to use for comparison since it is heavily influenced by outliers. It might be more beneficial to compare the distributions using non-parametric tests, such as the Mann-Whitney U test. The Mann-Whitney U test resulted in a p-value of 2.46e-5 and concluded that the ROI distributions for both samples (movies with top actors as leads and movies without top actors as leads) are, in fact, statistically different.

Does having a top director have an impact on movie revenue and/or ROI? Are the observed differences in average revenue/ROI between movies with top directors and movies without top directors statistically significant?

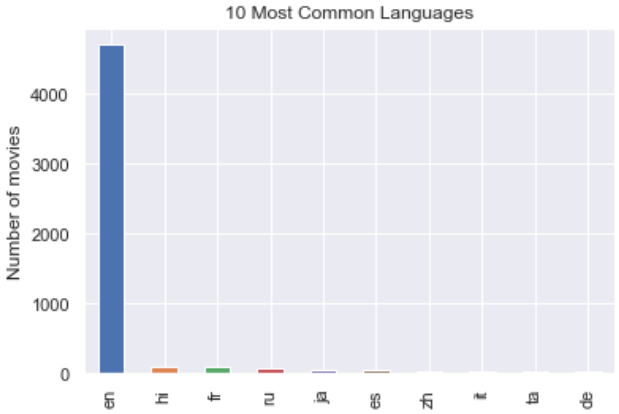


A top director is defined as any director appearing on highest grossing actors list on IMDB. Similar to hiring a top actor, a top director also results in a much higher average revenue (~280M vs 75M USD). However, while the ROI is lower with a top director, it still almost twice as large as the ROI with a top actor. This could indicate that hiring a top director can result in higher revenues on lower budgets (i.e. higher ROIs), compared to hiring a top actor.

Similar to the analysis for top actors, the Mann-Whitney U test can be used to compare the differences in revenue and ROI since both samples follow a non-normal distribution. The test concluded that the revenues for both samples (movies with top directors and movies without) are

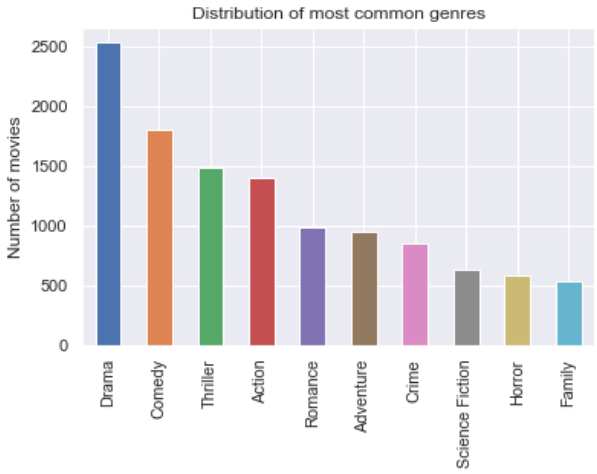
from different population distributions (note that the t-test and bootstrapping approach also reached the same conclusion). The same conclusion was reached for the ROI data. Therefore, there is statistical difference between the two distributions, indicating that revenue/ROI values for movies with top directors are statistically likely to be different form the revenue/ROI values for movies without top directors.

What are the 10 most common languages used in this dataset?



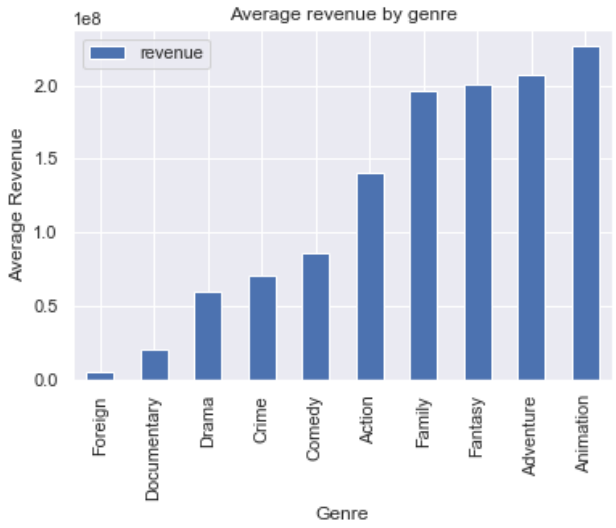
The vast majority of movies in this dataset are made in English (n=4711) followed by Hindi (n=96) and French (n=86).

What is the most common genre in this dataset?

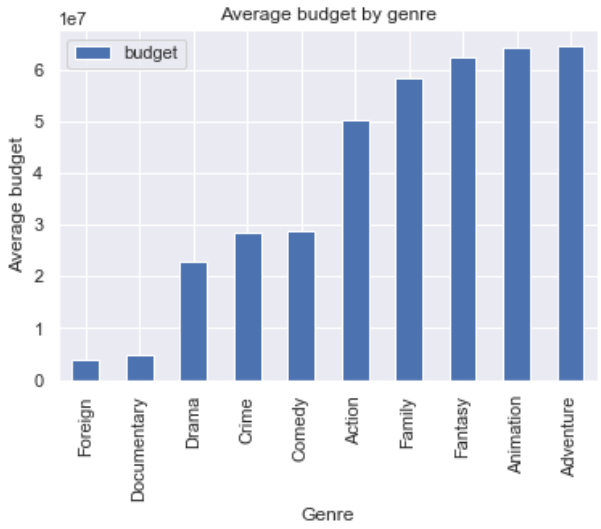


The most common genre is drama (~2500) followed by comedies, thrillers, and action movies.

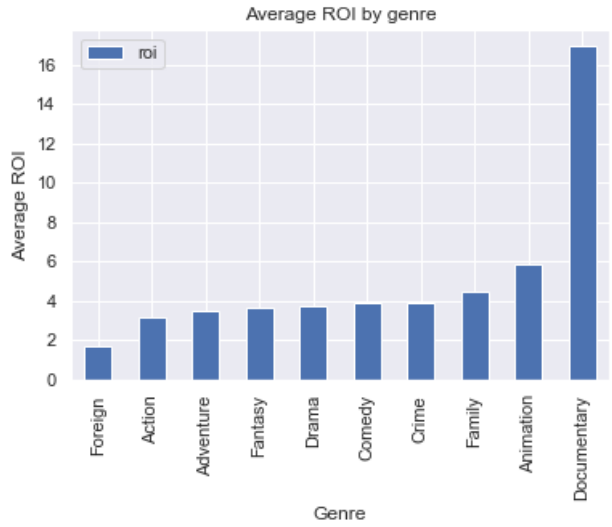
What genres are associated with the highest revenues, budgets, and ROIs?



Animated movies appear to have the highest average revenue followed closely by family, fantasy, and adventure. Foreign films, documentaries and dramas appear to have the lowest average revenues.



Adventure and animated movies appear to have the highest budgets, followed closely by fantasy and family. In comparison with the previous plot, it seems that movie genres with higher budgets typically result in higher revenues, which was also seen with the budget-revenue correlation plot earlier.



Documentary films have the highest ROI by far, followed by animated movies. Most other genres (except for foreign) result in similar ROIs on average.

Is there a relationship between genre and categorical success?

The chi square test of independence resulted in a p-value close to 0, thus rejecting the null hypothesis of independence between genre and categorical success. It can, therefore, be concluded that there is a statistically significant correlation between the two variables.