Final Report: 10 Years of IMDB Data

Team Members: Andreas Lemos, Samantha Williams, Omar Kapur

Introduction:

Since we are sheltering in place, on pause, staying home or quarantined, we have been watching way more films during down time. A good film will make you laugh, cry, motivate you or let you escape into a fantasy world, but not all of them are great. And yet somehow they can still be successful in popularity and financially which lead us to take a deeper dive into a dataset of the 1000 most popular movies on IMDB. We decided to focus our questions around Revenue as that is a universal bench mark for success and this dataset already filtered the films to top 1000 most popular according to IMDB.

Research Questions:

A film could be popular and generate very little Revenue, so we wanted to discover what is the most important driver of Revenue for a film. From there we developed three tiers of analysis:

1. What factors have an immediately noticeable impact on Revenue within the data?
    a. As our collective attention spans shrink, does Revenue change with runtime?
    b. How does Metascore (Critics Score) and Rank (IMDB Voters Score) compare? How does it influence Revenue?
2. How does Revenue change by Genre?
    a. What is the most common Genre tag?
    b. Is there a Genre that is “riskier” when it comes to Revenue? Is there Genre that is the best?
3. Can a Director make a film more successful?
    a. Is there a specific Genre of film that a Director generates more Revenue?

Our Data:

The primary dataset was downloaded from Kaggle that contained 1,000 of the most popular movies according to IMDB in the last 10 years. Below is the metadata on our data set and how we interpreted some of the fields for our use.

Field Name	Description	Sample Value	Interpretation
Rank	Movie rank order (also Index)	1	Did not use this data. Could not reproduce the calculation and is how IMDB identifies the most popular films.
Title	The title of the film	Suicide Squad	Used with Revenue to sort films.
Genre	A comma-separated list of Genres used to classify the film	Action,Adventure,Fantasy	Genre tags identified and paired with Title and Revenue. Used dummies on this field.
Description	Brief one-sentence movie summary	The Rebel Alliance makes a risky move to steal the plans for the Death Star, setting up the epic saga to follow.	Used for word clouds to see the most common words However, not relevant to our research.
Director	The name of the film's Director	M. Night Shyamalan	Grouped with Revenue and Genre.
Actors	A comma-separated list of the main stars of the film	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe Saldana	Interesting field, but did not explore further in our analysis
Year	The year that the film released as an integer	2007	Filtered films by year
Runtime (Minutes)	The duration of the film in minutes	127	Grouped with Revenue
Rating	User (of IMDB) rating for the movie 0-10	8.3	Paired with Revenue and Genre
Votes	Number of votes (from IMDB Users)	232072	Explored relation to Revenue and Metascore

Revenue (Millions)	Movie Revenue in millions	368.31	Removed all rows with missing values.
Metascore	An aggregated average of critic scores, values are between 0 and 100, higher scores represent positive reviews	49	Assumption: Enough critics scored each film to make this value relevant

Initial Data Exploration and Cleaning:

Cleaning Primary Dataset - The original dataset had 1000 rows and 12 columns. Each row represents a unique film. The dataset had very few missing values, with 128 missing from the Revenue (Millions) and 64 missing from Metascore. All records with missing data in the Revenue variable were removed from the dataset. Leaving 872 rows of data to work with.

Genre - There are 20 unique Genres classified in this dataset. Films can be classified by multiple Genres. We needed to break out films with multiple Genres into single Genre and achieved this using dummy

Rank - We removed the Rank variable because we did not know how IMDB calculated this figure in the dataset and did not provide further clarification in the metadata.

Description - This field was only used to make a word cloud. Ultimately we determined that it had very little relevance to our research questions.

Director - This was a clean field that had no null or missing values.

Actors - This was a list of actors. All rows contained four actors listed with the exception of a single row that contained three actors. Upon further research into the anomaly, it appears that the film only had three actors. No other verification of actor data was conducted or required. It is assumed that the actors are listed in order of billing[1].

Year - The films in this list were released between 2006 and 2016 with 42 % of the films on this list released between 2015-2016.

[1] *Actors whose names appear first are said to have "top billing" because they are usually the principal characters in the film and have the most screen time.*

Runtime (Minutes) - This was a clean variable, with the average film runtime was 113.172 minutes.

Rating - We looked at IMDB user Ratings and Revenue as it appeared to have the strongest correlation.

Votes - Looking at the numerical values and graphs as compared to Revenue, the min number of votes recorded by IMDB is 178, with the max number of votes 1.79 Million.

Revenue (Millions) - This variable had 128 missing values that we removed from our data set for further exploration, giving us 872 films. There were no films with a negative Revenue; however there was one film, *A Kind of Murder*, that listed Revenue as 0.0, this could be an error. The film that made the most Revenue was Star Wars Episode 7, directed by Max JJ Abrams at $936 Million. The average Revenue for a film is $82 Million.

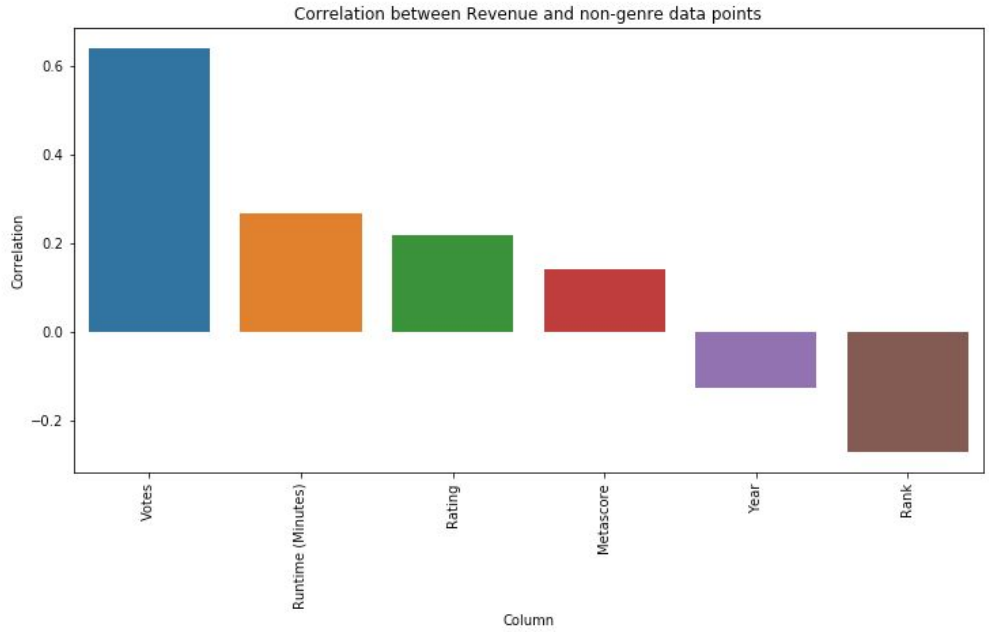
Metascore - 64 Missing data points, 30 of them also had missing values for Revenue.

Our Data Story:

A. First level of analysis:

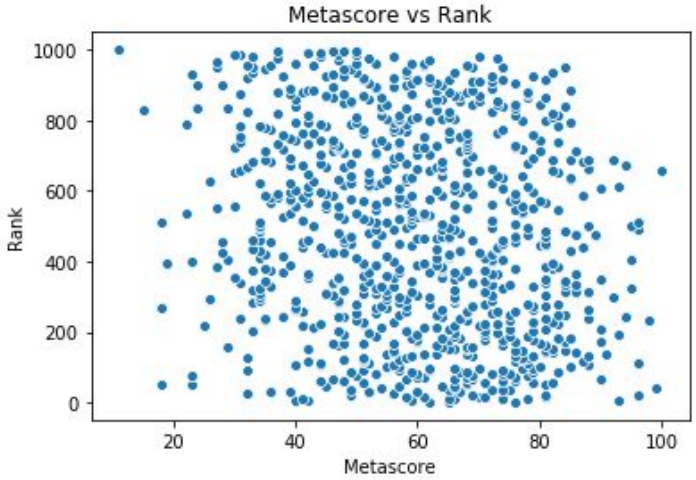
Our approach to answering this question was to look at the correlation between Revenue and different columns. We then looked into each column and relationships between columns at a deeper level.

What is the Correlation Between Revenue and other Variables (Excluding Genre)?



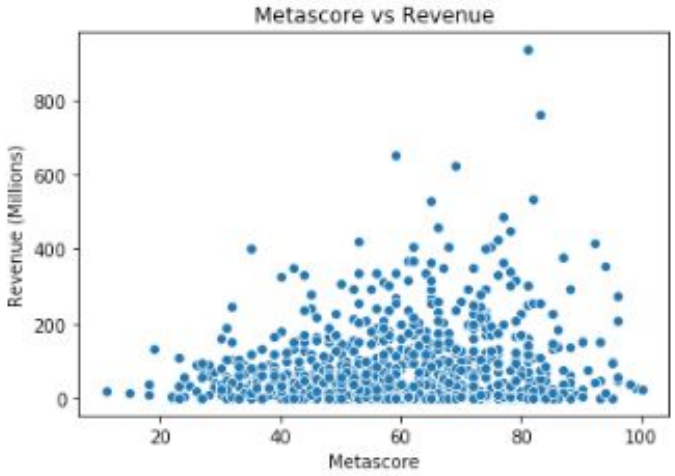
Takeaway: Runtime has a positive correlation - given the movies in our dataset, longer runtimes are correlating with more Revenue. However our data is limited to the most popular (popularity is provided by IMDB by Rank) films on IMDB for the 10 year timeframe. The year of the film's release has a negative correlation; however the data is not evenly distributed and is not uniform in the dataset. It would be helpful to have time series data on revenue. Movies that were released earlier have had more time to accrue revenue, and the IMDB methodology may be accounting for revenue in different ways. Rank is as expected since a highly rated film is ranked number 1, so a negative correlation between rank and revenue implies that as the rank improves (decreases), the Revenue increases.

How does Metascore and Rank compare?



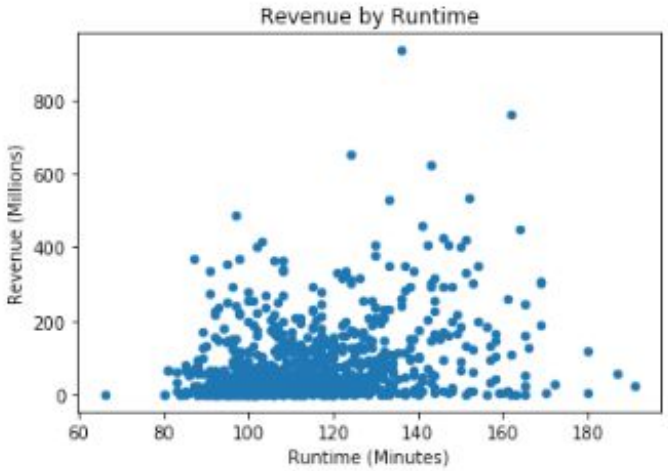
Takeaway: Metascore, composed of critics reviews of the film, vs Rank that is calculated by IMDB, provided no insights. It may be helpful to know how both of these numbers are calculated.

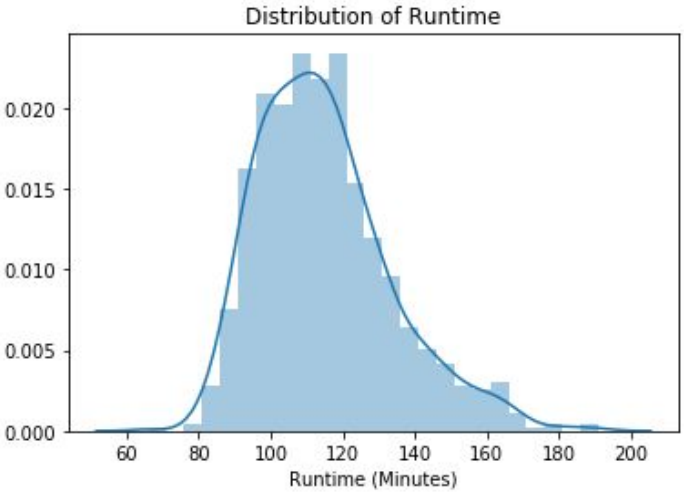
How does Metascore and Revenue relate to one another?



Takeaway: There is not a clear, strong relationship that shows up between Metascore and Revenue. There are a number of films that have high Metascore values and yet do not have high Revenue values.

What does Revenue and Runtime look like? Is there a runtime that always brings in larger Revenue dollars?





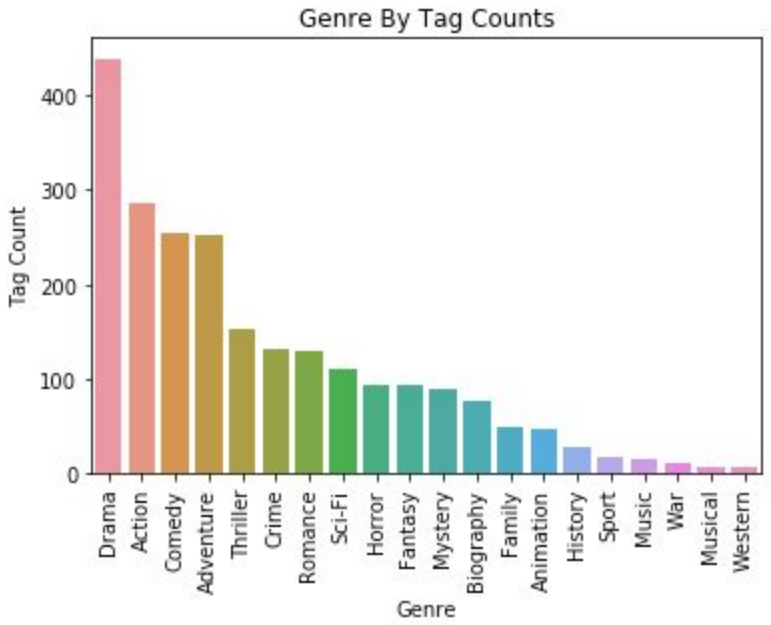
Takeaway: The Runtime variable has a relatively normal distribution, with most films being within roughly 90-140 minutes in length. This makes sense from a domain perspective as movies that are too short may leave viewers unsatisfied, and movies that are too long may lose their attention. However based on the previous correlation chart, Runtime does have a positive correlation with Revenue, suggesting that while there are

both upper and lower bounds, some longer movies do certainly make more (these may be blockbuster movies).

B. Second level of analysis:

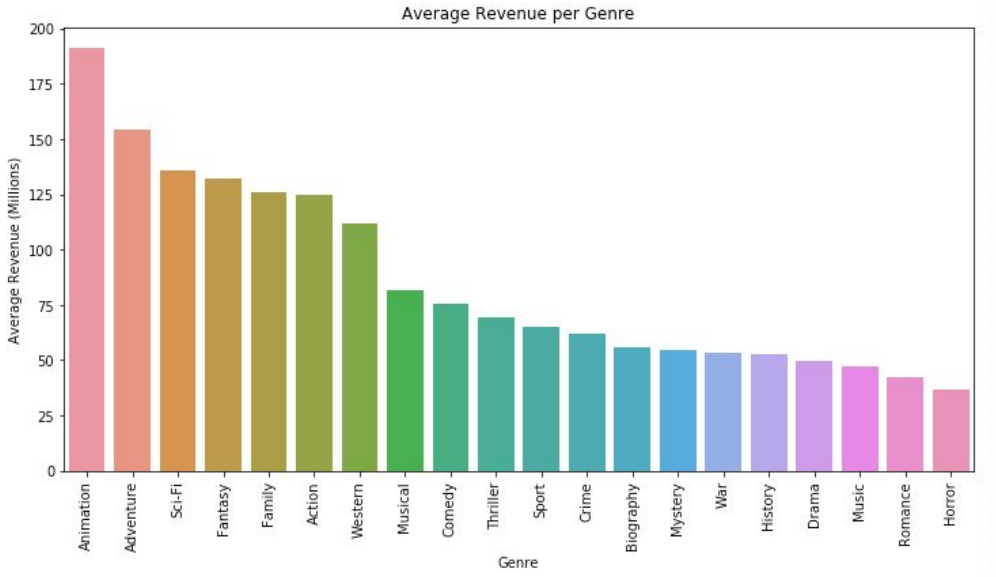
There was a challenge when analysing this column as there were multiple genres listed per film. We used dummy variables to break out the Genre tags. As a result there may be a higher error in the analysis by genre of this dataset. We also looked at the original Genre classifications as compared to Revenue.

How many Genres by tag counts?



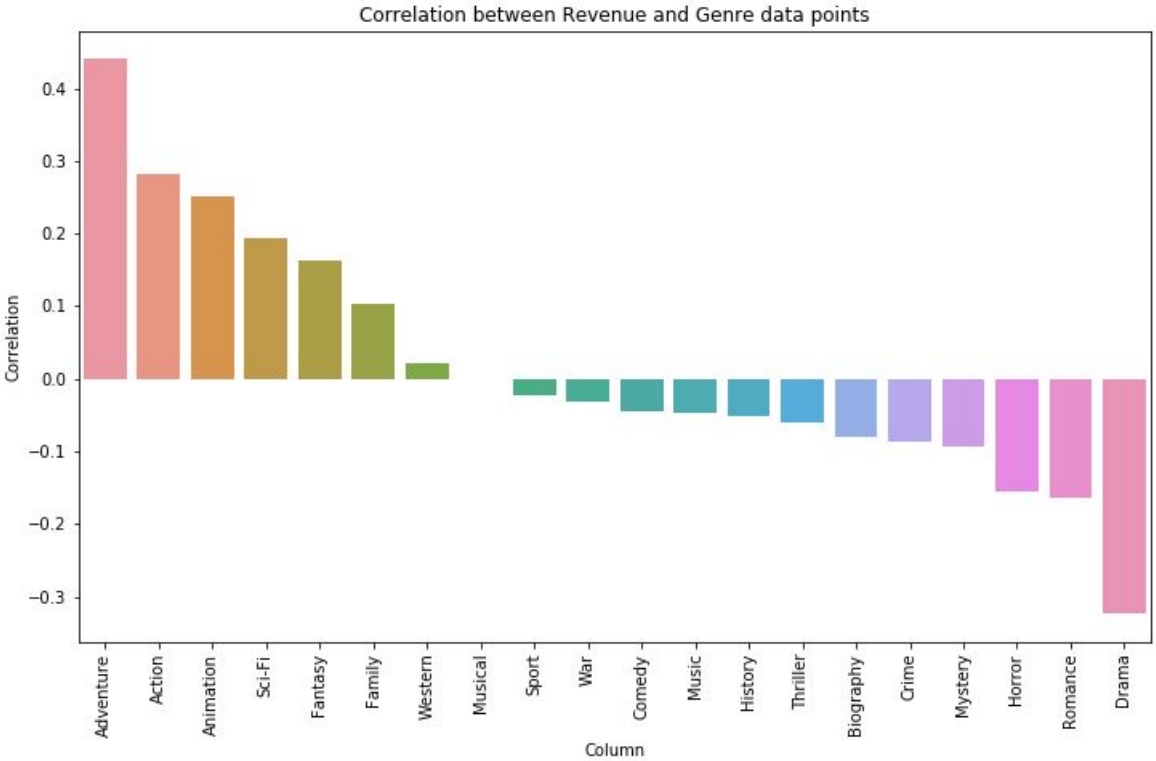
Takeaway: Drama is the most frequent tag among the films in our dataset. We were curious what films are action and not adventure and discovered that 148 films are all action and not tagged with adventure, 104 films are tagged with adventure and not action and 155 films have both tags.

What is the mean Revenue by Genre?



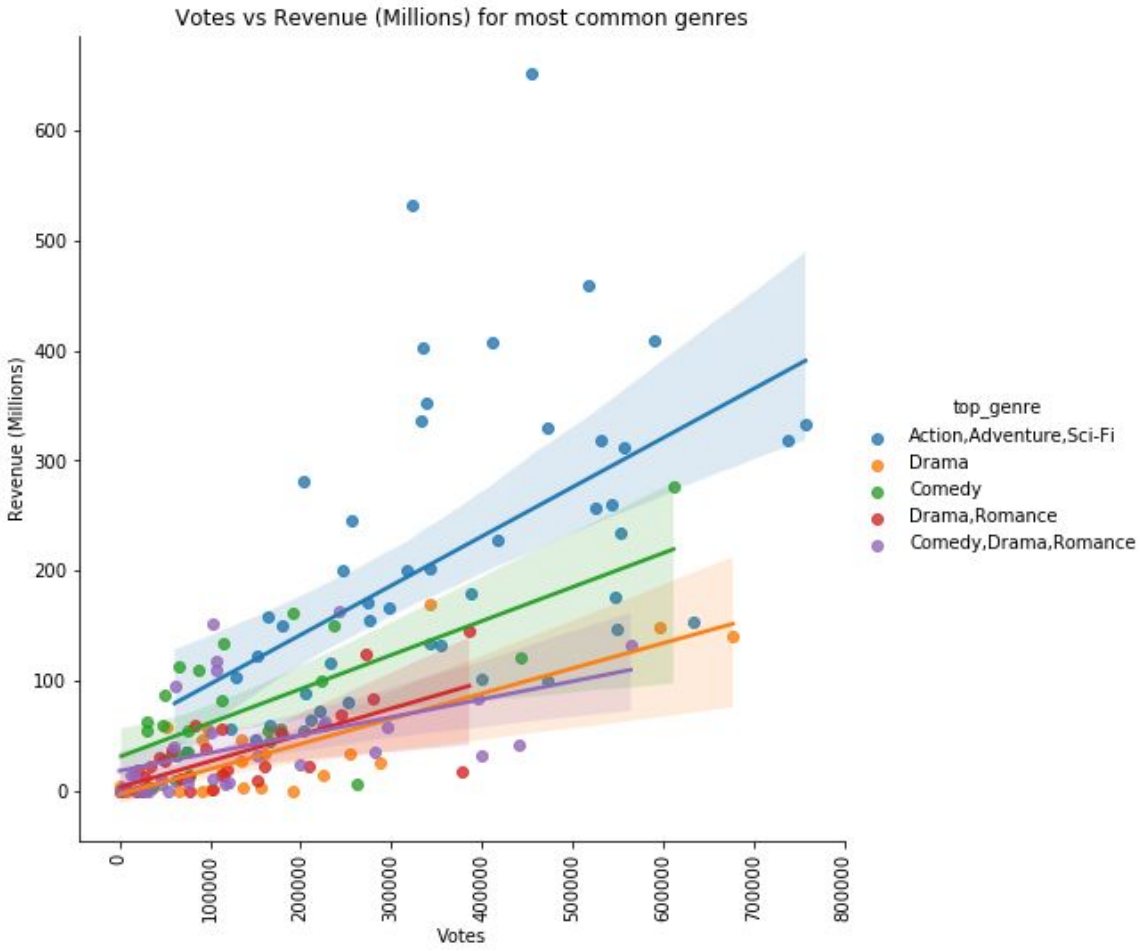
Takeaway: The Horror films that made this list had the lowest amount of Revenue generated on average $36 Million. The Genre Animation generated the most Revenue at $191 Million on average.

What is the correlation between Genre and Revenue?



Takeaway: Adventure, Action, and Animation genres have the strongest positive correlation to Revenue. Meanwhile, Drama, Romance, and Horror have the strongest negative correlation to Revenue. It is likely that there were some films with multiple Genre tags in addition to drama that did not generate as much Revenue.

What Genre has the most Votes by Revenue?

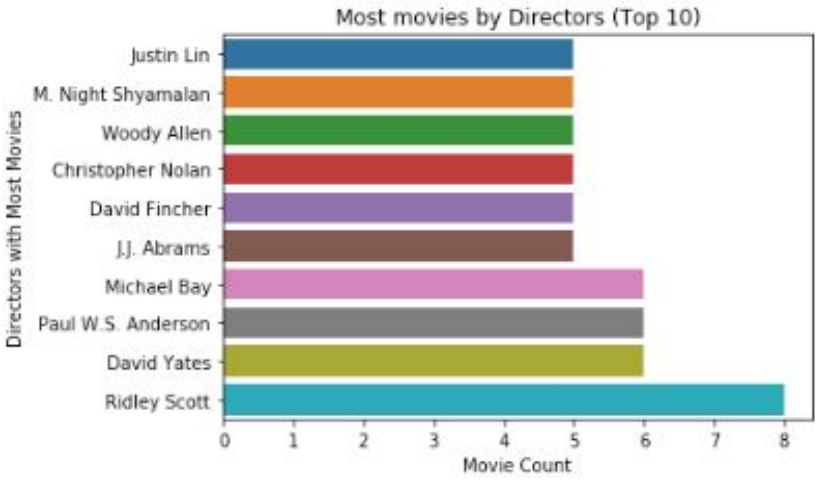


Takeaway: We were curious about Revenue and original Genre categories. We found that the grouped genre of Action, Adventure, Sci-Fi is the most popular among IMDB voters and generates more Revenue than any film with a drama tag. It appears that the more Votes a film has on IMDB, the more Revenue generated across the top Genres.

C. Third level of analysis:

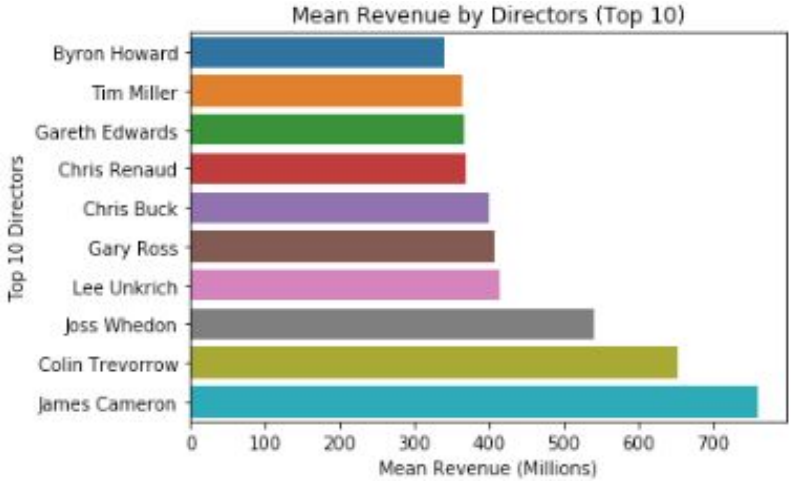
There are 540 unique Directors in this dataset, with most only having one film listed within the 10 year dataset. As a result, we were not able to get much insight as to the impact that the Director may have on film Revenue.

Who are the top 10 Directors with the most number of films?



Takeaway: Most directors only had one film they directed for the dataset time period. However, there were some directors that had more than five movies listed. In general, this did not translate into more revenue for these directors over the years.

Who are the top 10 Directors by Revenue?



Takeaway: It was interesting to note that of all the Directors that generated the most Revenue each only had one film in the dataset.

Conclusion:

Most Influential Factors:

Votes and Runtime were the most important factors for determining Revenue, within the dataset we analyzed. Votes had a positive correlation to Revenue, which made sense to us - a film that has more votes likely has had more people watch it, and will have made more money.

We also saw that some genres clearly outperformed others for Revenue earned. Action, Adventure, and Animation tended to do really well financially, while Horror, Drama, and others did poorly. Even when we accounted for the number of votes that a film received, we still saw trends that showed some genres outperforming others. We also had disparities in the number of genres represented in the dataset, but saw that some genres, such as Animation, had fewer films but still performed well.

Challenges:

Limiting the scope of our analysis to not include statistical inference was a challenge as we were curious. However a bigger obstacle was that we used only one dataset that limited our analysis options. We were unable to find a dataset readily available that would have been useful for our analysis without investing a large amount of time scraping the information from the web. While we enjoyed the exploration of this data set we found some of the more complex plots we wanted to create in python difficult.

Next Steps:

Our analysis was primarily based on Revenue. It shows that Drama films, for example, don't tend to do so well compared to Animations and Action movies. However, if we also had Cost data for each film, it could turn out that the Drama genre would do much better if Profit were used.

Furthermore, data on Month and Year of Release for the films would have been a nice element to have. Combined with Genre, it would most likely lead to some seasonality aspect and insight.

It is not clear what timeframe the Revenue data encompasses. Is it just for the release weekend, or for the entire time range of the data. Having time-series data of Revenue would allow us to ensure a more accurate comparison in our analysis.

Finally, a better categorization of genres would be helpful in comparing the different genres. A comma-separated value ends up giving us too many value combinations to analysis, making it more difficult to provide insights.