

TMDB Movie Revenue Analysis Project



DISCOVER MOVIES TV SHOWS PEOPLE

Apps Forums Leaderboard Contribute API Support



EN

LOGIN

SIGN UP

Search for a movie, tv show, person...

Discussions

CATCH UP NOW

That's a Wrap!

READ THE 2018 RECAP

On TV



In Theaters



The Movie Database (TMDB) is a community built movie and TV database. TMDB offers extensive metadata for movies, TV shows and people - and even high resolution posters and fanart.

Movies made an estimated \$41.7 billion in 2018, and the film industry is more popular than ever. However, what movies make the most money at the box office? How much does a director matter? Or the budget?

TMDB has a public dataset with metadata on over 7000 past films from their database. TMDB has put a challenge to the public: can you try and predict the overall worldwide box office revenue for films?

Data points in the metadata includes cast, crew, plot keywords, budget, posters, release dates, languages, production companies and countries.

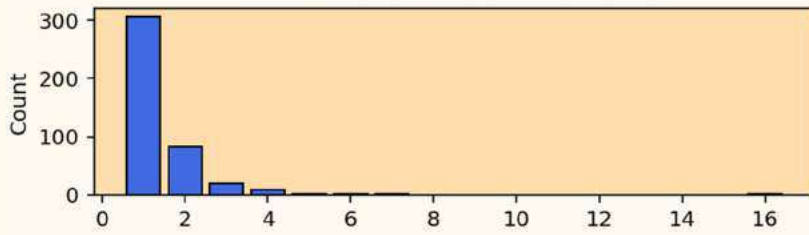
In this report, a thorough data analysis of the metadata is conducted, with the aim of extracting key findings in the data. Further, to try and predict box-office revenue for movies, a linear model is built based on statistical methods - to try and predict revenue for movies as best as possible.

The results indicate that there is a lot of important findings in the data, and that the revenue for many movies can be predicted remarkably well.

The public dataset and challenge was found on Kaggle:
<https://www.kaggle.com/c/tmdb-box-office-prediction>

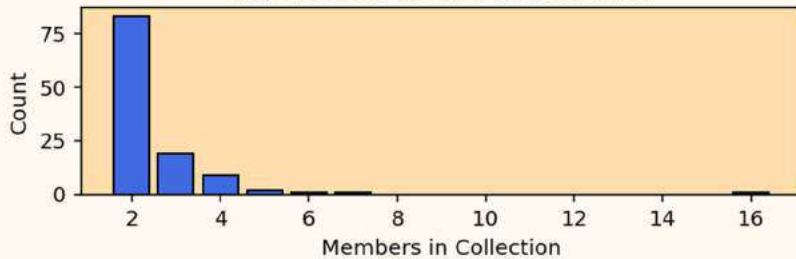
Data Analysis - TMDb Project - Amount of movies in collections, budget values, and amount of movies in genres

Amount of Movies Belonging to a Collection
At Least One Movie in Collection

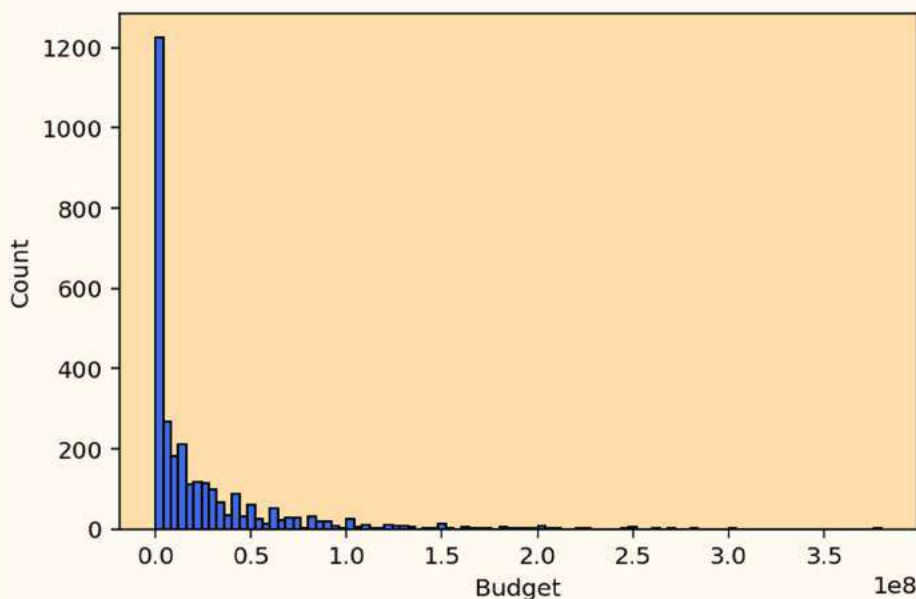


- Most collections have one movie
- Moderate amount of collections with 2-4 movies
- One collection with 16 movies (James Bond)
- Makes sense - rarely more than three movies in a collection e.g. sequel, trilogy

At Least Two Movies in Collection

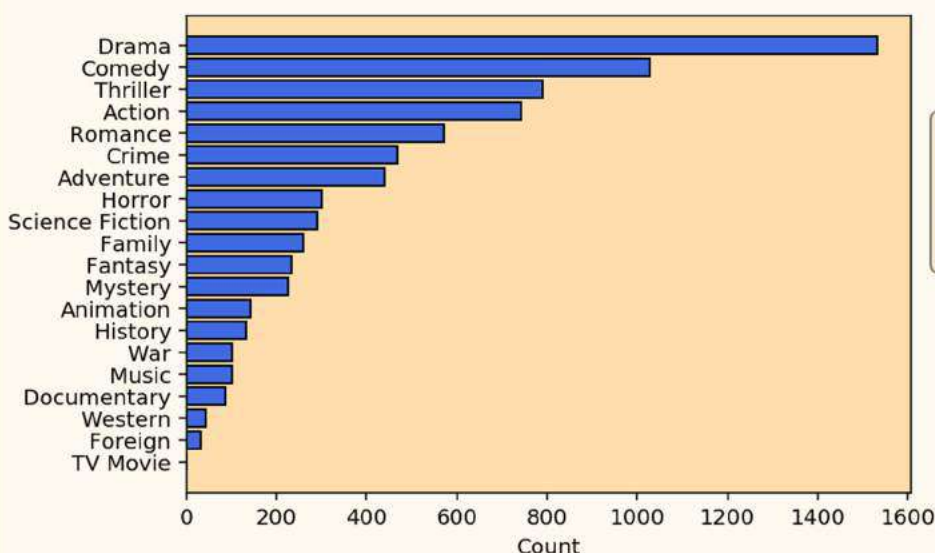


Histogram of Budget Values



- Most movies are low-budget
- Budget decreases exponentially
- A few high-budget movies - mostly likely blockbuster movies

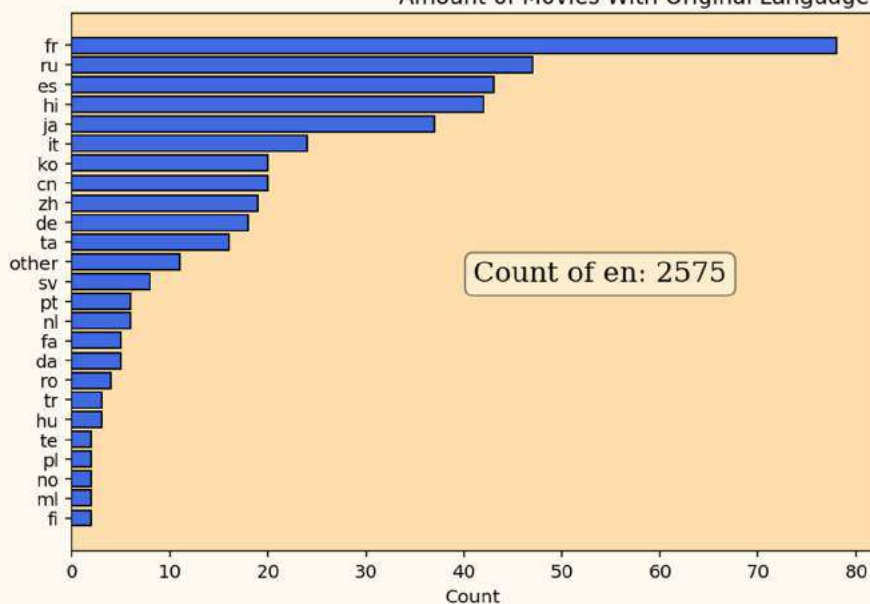
Amount of Movies Related to Each Genre



- The most popular genres are drama, comedy, thriller and action
- Relatively few music and documentary type movies

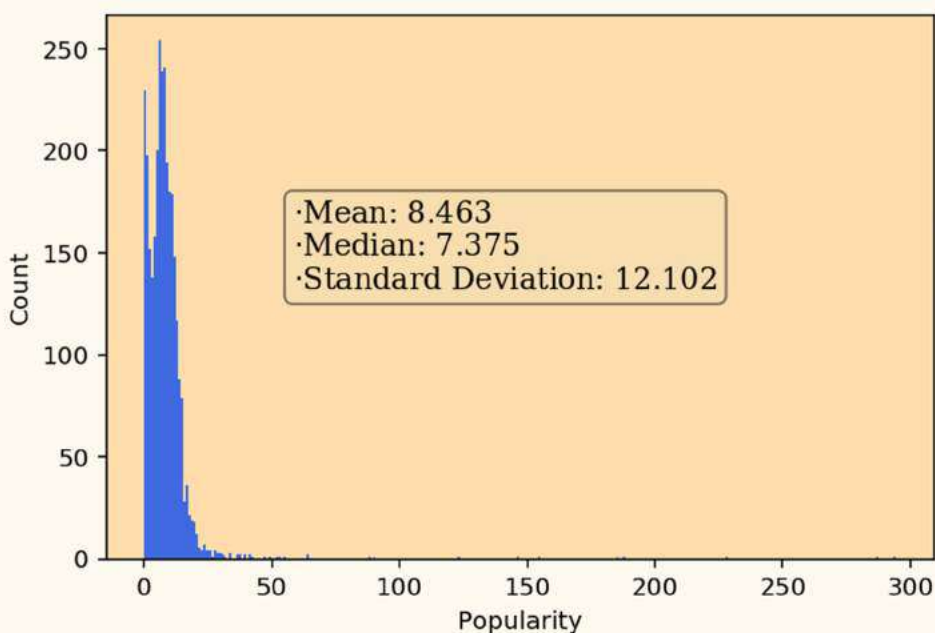
Data Analysis - TMDb Project - Original language, popularity, and production companies

Amount of Movies With Original Language



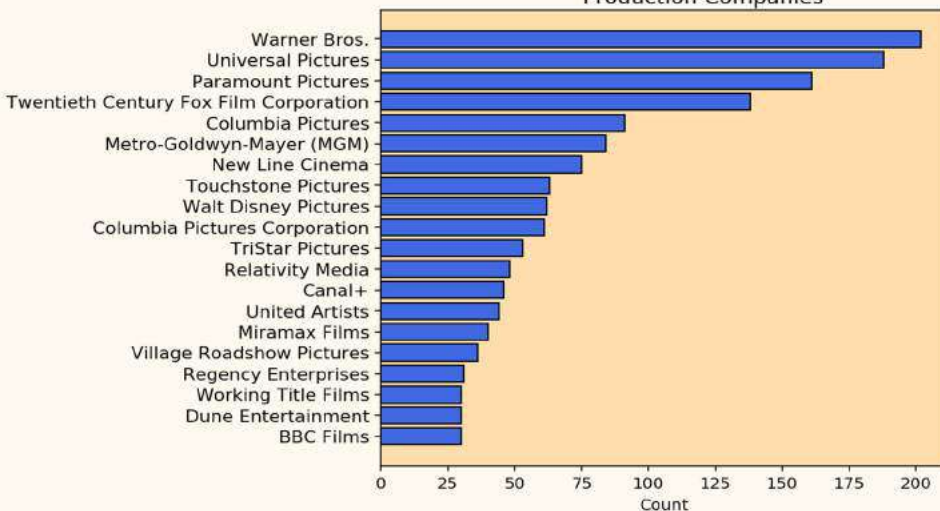
- Other: Amount of spoken languages with one related movie
- Most movies are in english
- French movies are second by a wide margin, probably related to France having a rich, historic art culture

Distribution of Popularity Values



- Most movies are moderately popular
- A few movies with huge popularity

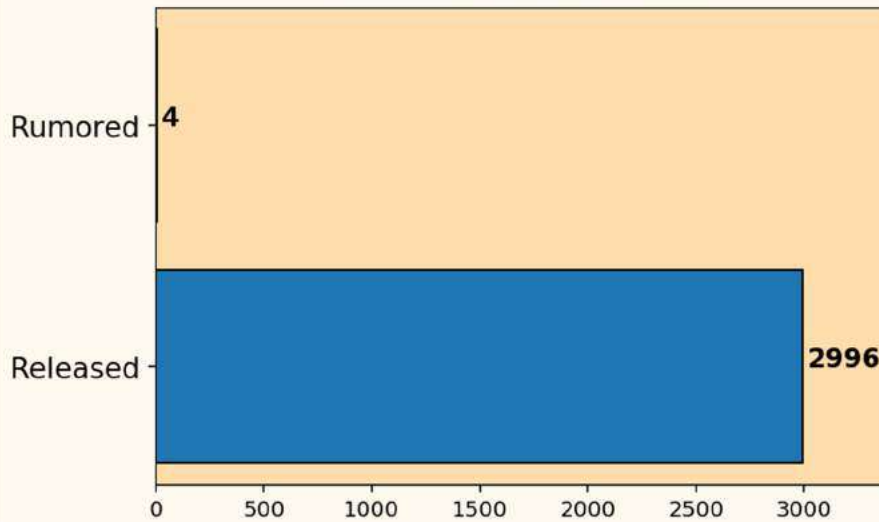
Amount of Movies Produced by Most Popular Production Companies



- Most popular production companies are American production companies, such as Warner Bros, Paramount Pictures, etc.
- A couple of European production companies, like BBC Films, Canal+, etc.

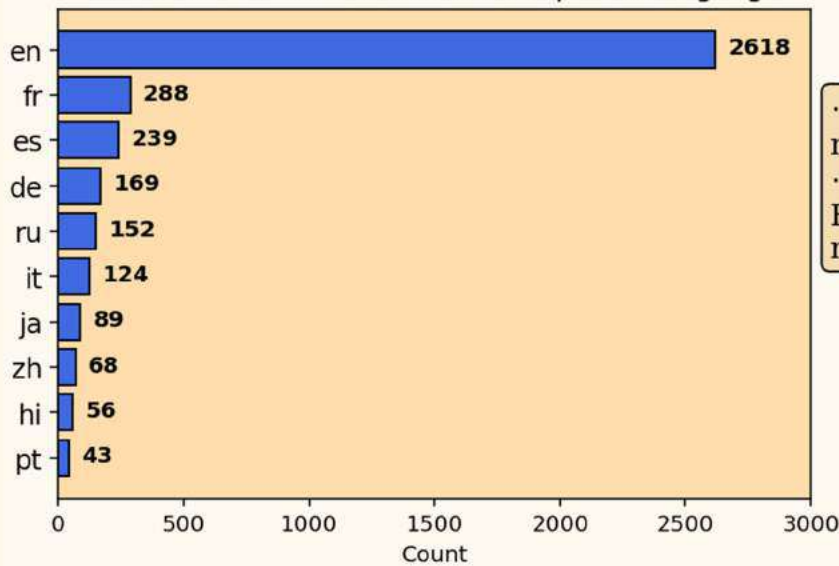
Data Analysis - TMDb Project - Status, spoken languages, and runtime

Distribution of Status Values



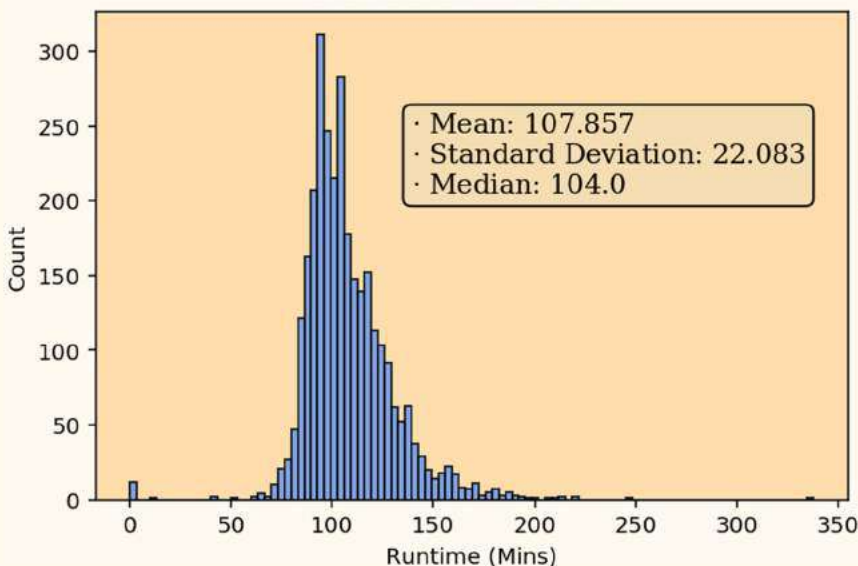
- The only two possible status values are 'Released' and 'Rumored'
- 2996 movies have the status 'Released' and 4 movies have the status 'Rumored'
- Hence, it is an easily discarded attribute

Amount of Movies with a Certain Spoken Language



- As with original language, most movies have English as a spoken language
- French, Spanish, German as top European languages makes sense, as they are major European nations

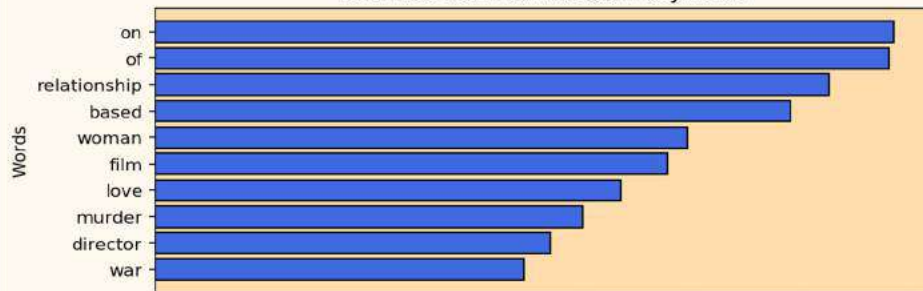
Distribution of Runtime



- Most movies have a runtime just below 2 hours
- Note that there is a moderate proportion of movies longer than 2 hours.

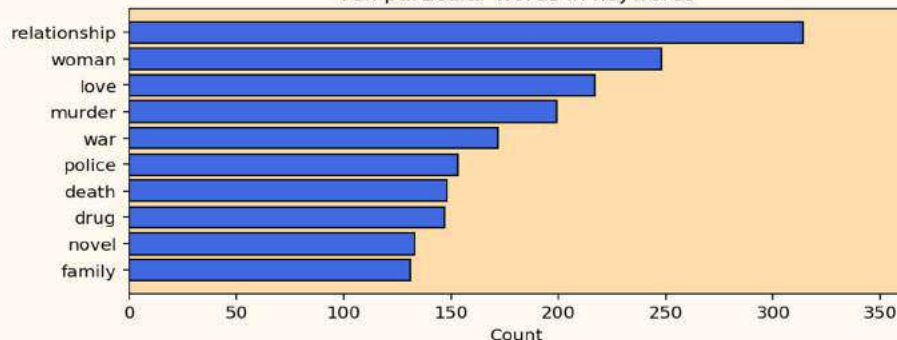
Data Analysis - TMDB Project - Keywords, titles, and taglines

Ten most common words in keywords



- The keywords are dominated by words that are very descriptive of movies
- Most common words are interesting words, with very low amount of sentence-building words
- Major takeaway is that keywords are suitable to aggregate main characteristics of movies

Ten particular words in keywords



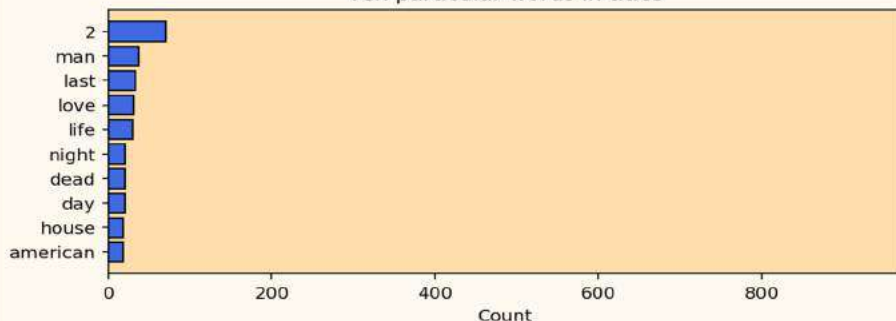
- Manually chosen words have high counts, relative the common words
- The usage of 'novel' indicates a lot of movies are associated with novel(s)
- Words like relationship, love, murder, war, death, drug, family indicates a wide range of categories of movies exists

Ten most common words in titles



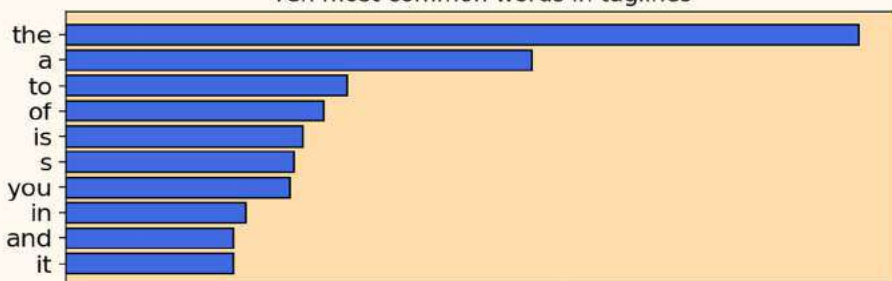
- A mix of common and interesting words in title
- 'The' is most common, by a wide margin - shows a popularity in a 'The'-structure of titles

Ten particular words in titles



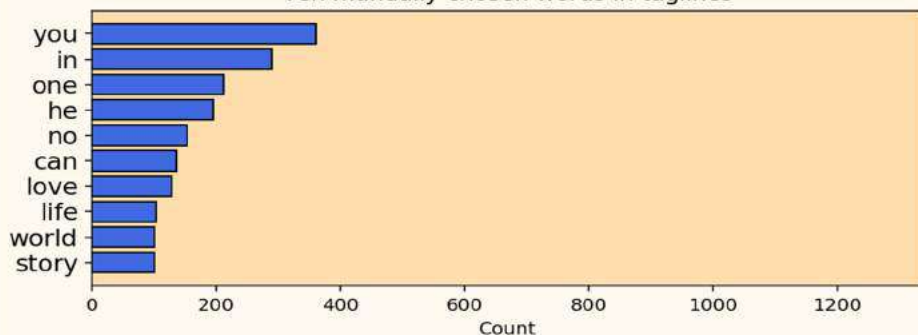
- Manually chosen words have moderately high counts - indicates that it may be possible, again, to utilize title words to aggregate main characteristics of movies.
- The usage of '2' indicates a lot of sequel movies
- Love, life, day indicates movies with positive, uplifting nature
- Night, dead indicates a prevalence of movies with a sense of danger, scary in them
- A wide reach in different possible categories of movies in title

Ten most common words in taglines



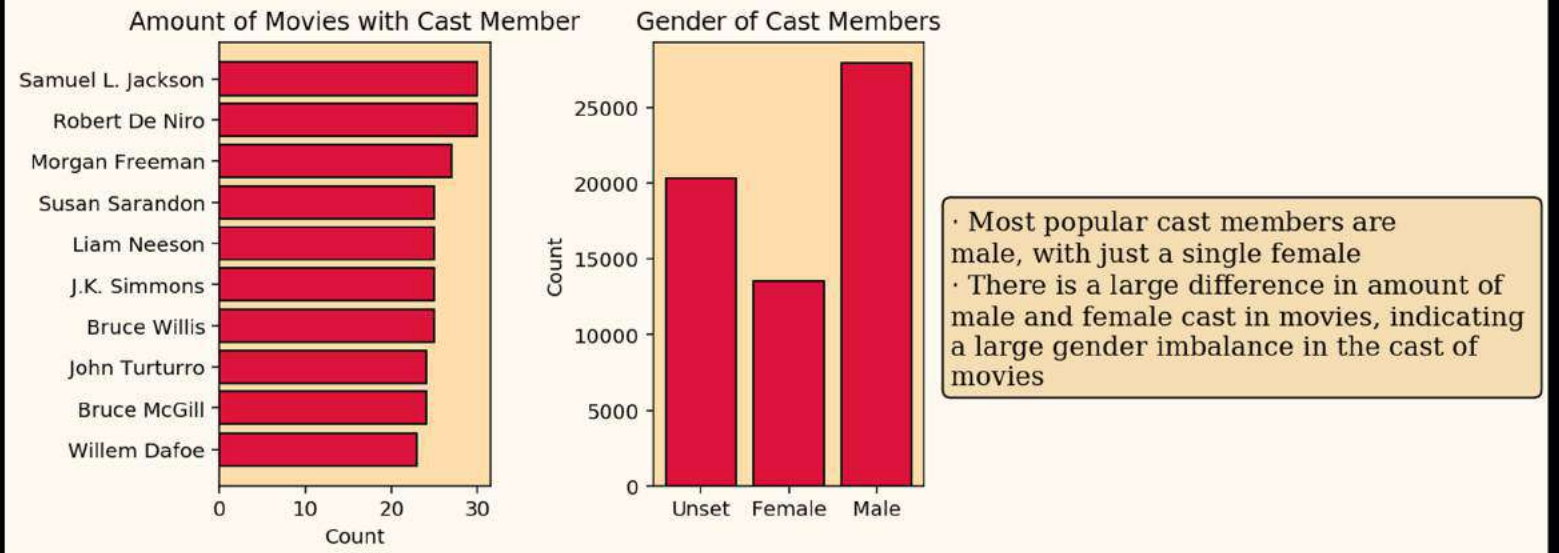
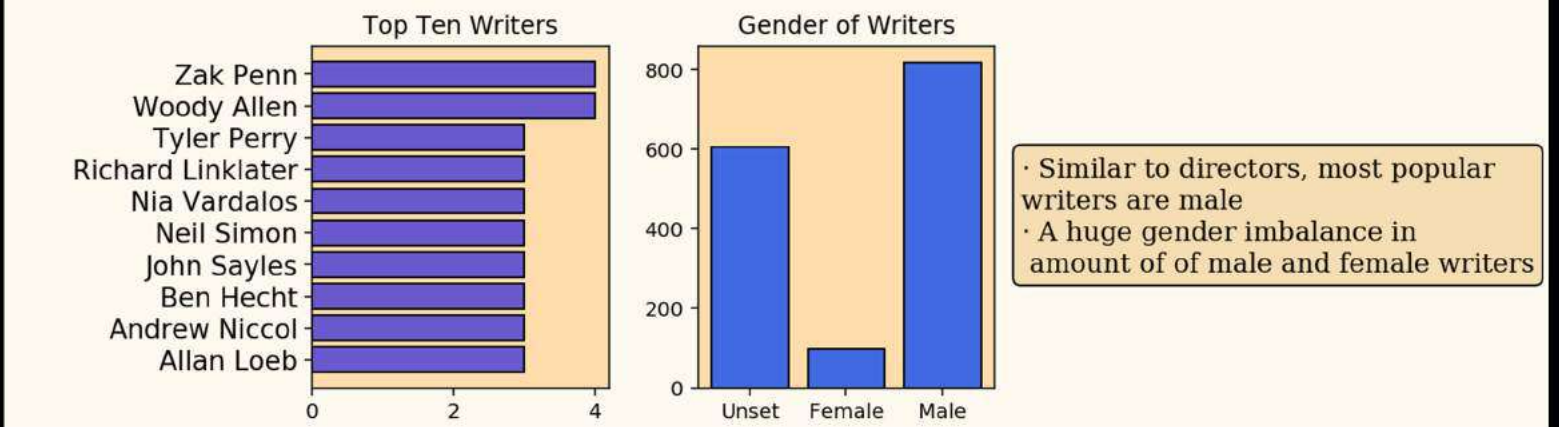
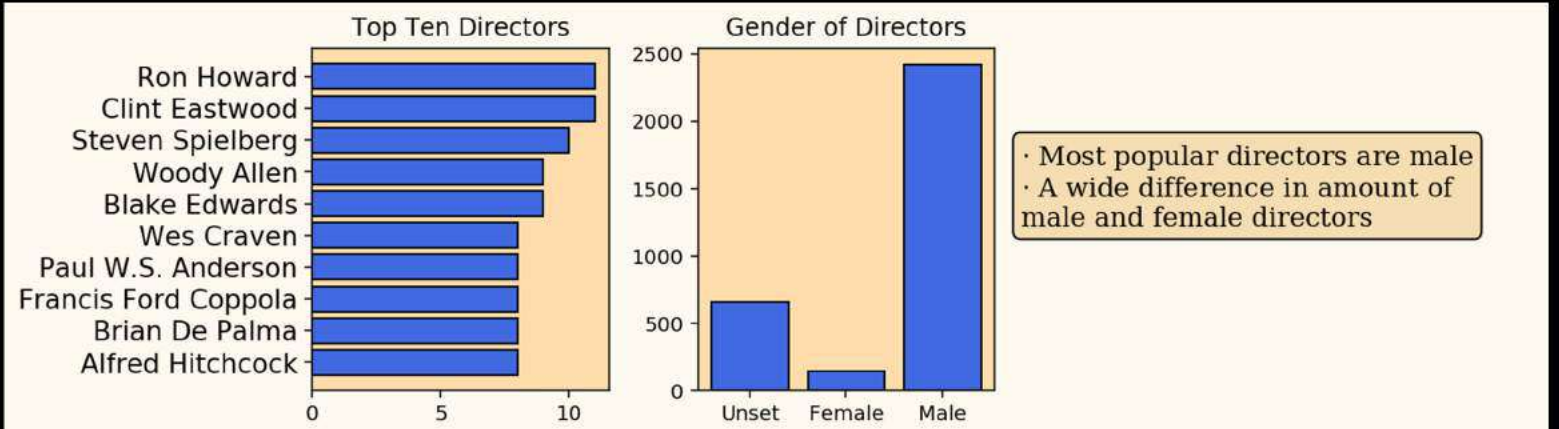
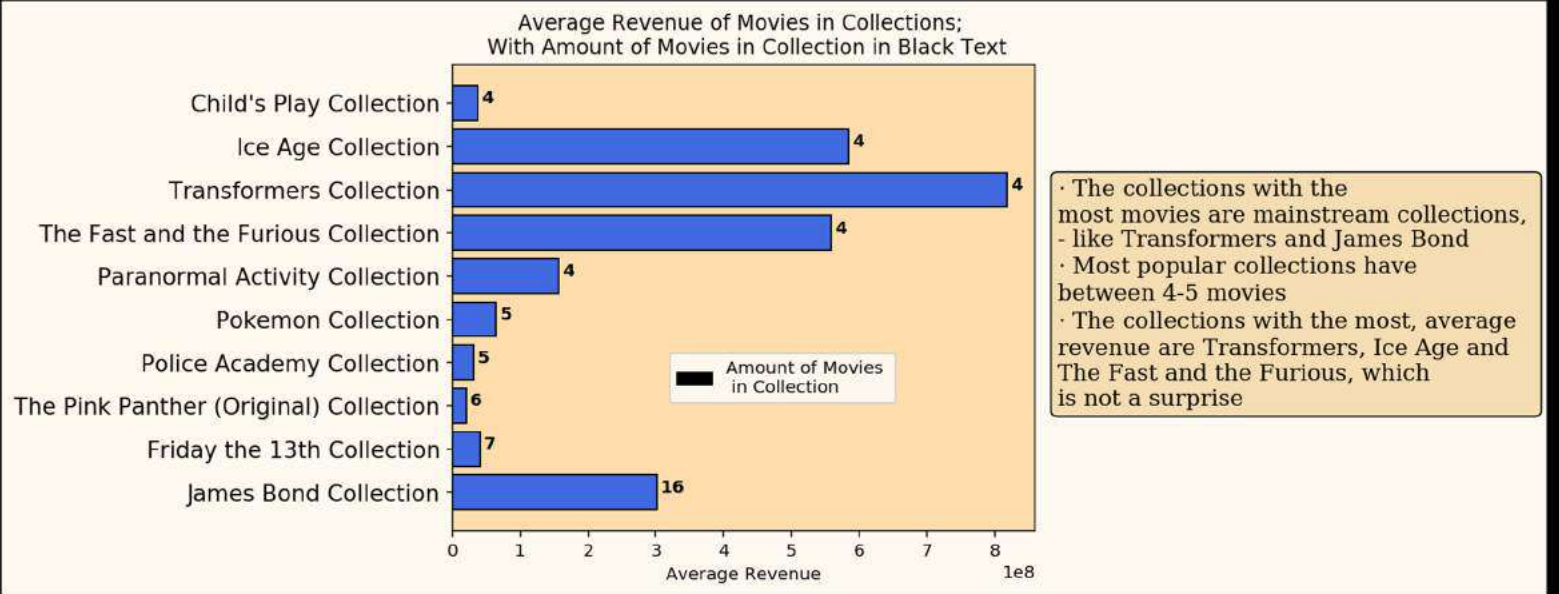
- Most common words are uninteresting, sentence-building words
- Not indicative of any prevalent characteristics of movies

Ten manually chosen words in taglines



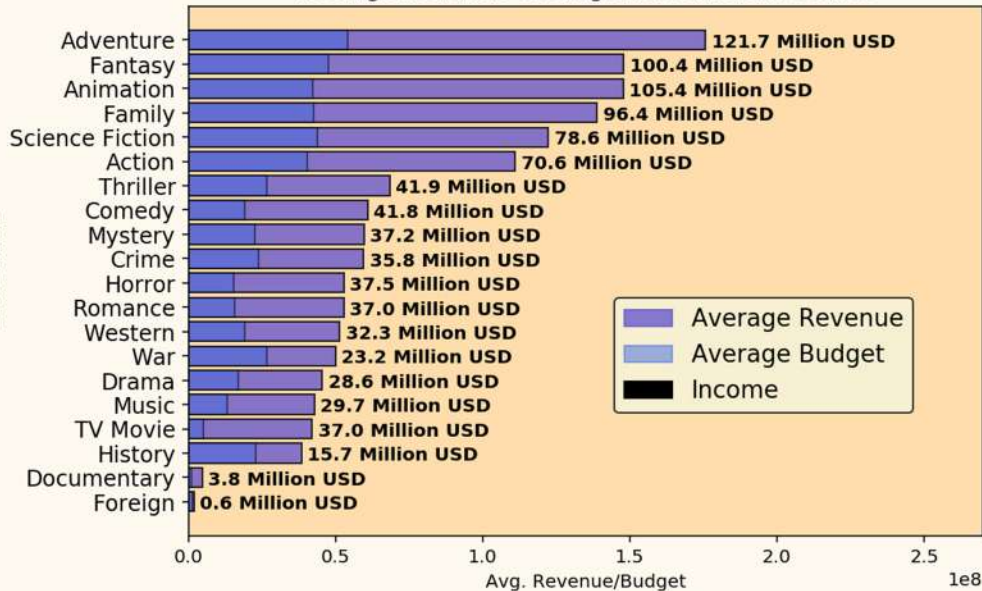
- Manually chosen words have moderately high counts - indicates that it may be possible to utilize tagline words to aggregate main characteristics of movies.
- Love, life indicates a lot of movies associated with a positive message
- Word, story shows the adventurous associated movies are prevalent

Data Analysis - TMDb Project - Collections, directors and writers, and cast



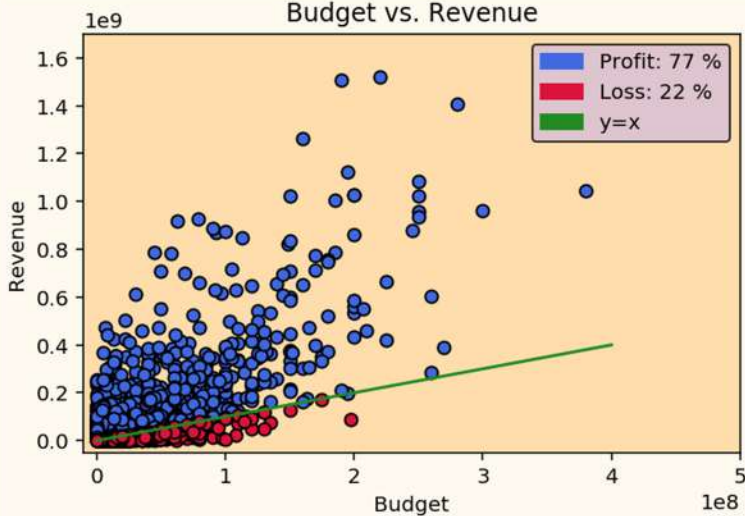
Data Analysis - TMDb Project - Genres, budget, popularity, and runtime

Average Revenue & Budget for Different Genres

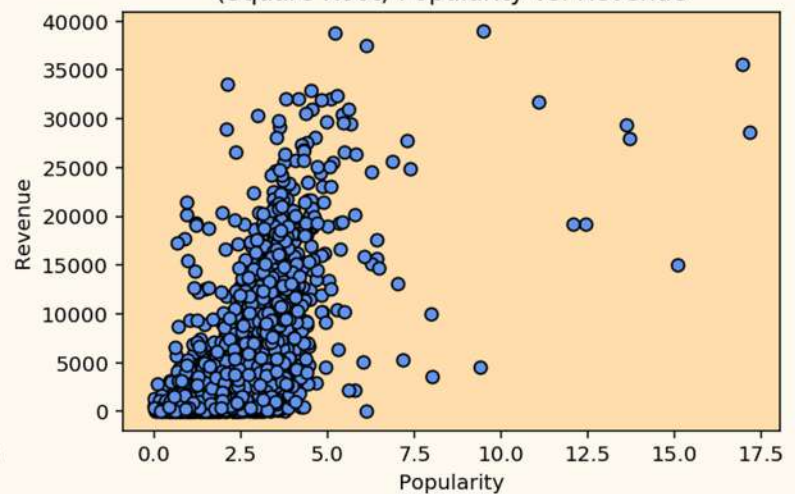


- Most profitable genres are Adventure, Fantasy, Animation, Family. Possibly because those genres attract parents & children, but also adults
- Most other genres generate much less in revenue and profit
- History-centered movies seems to perform the worst

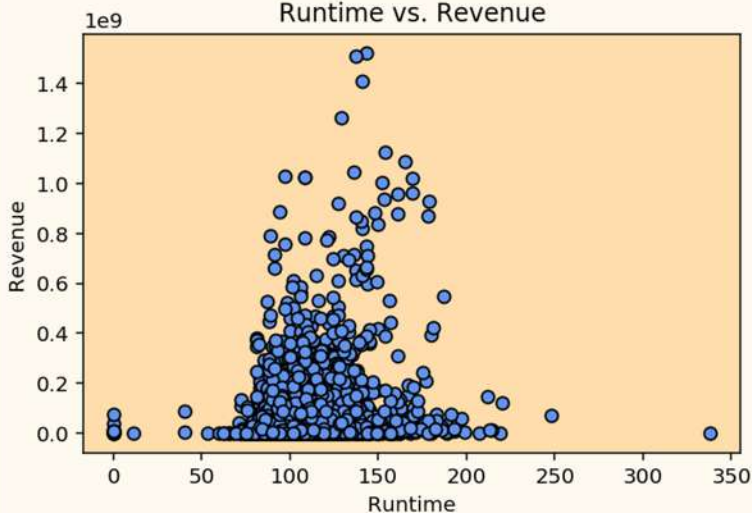
Budget vs. Revenue



(Square Root) Popularity vs. Revenue

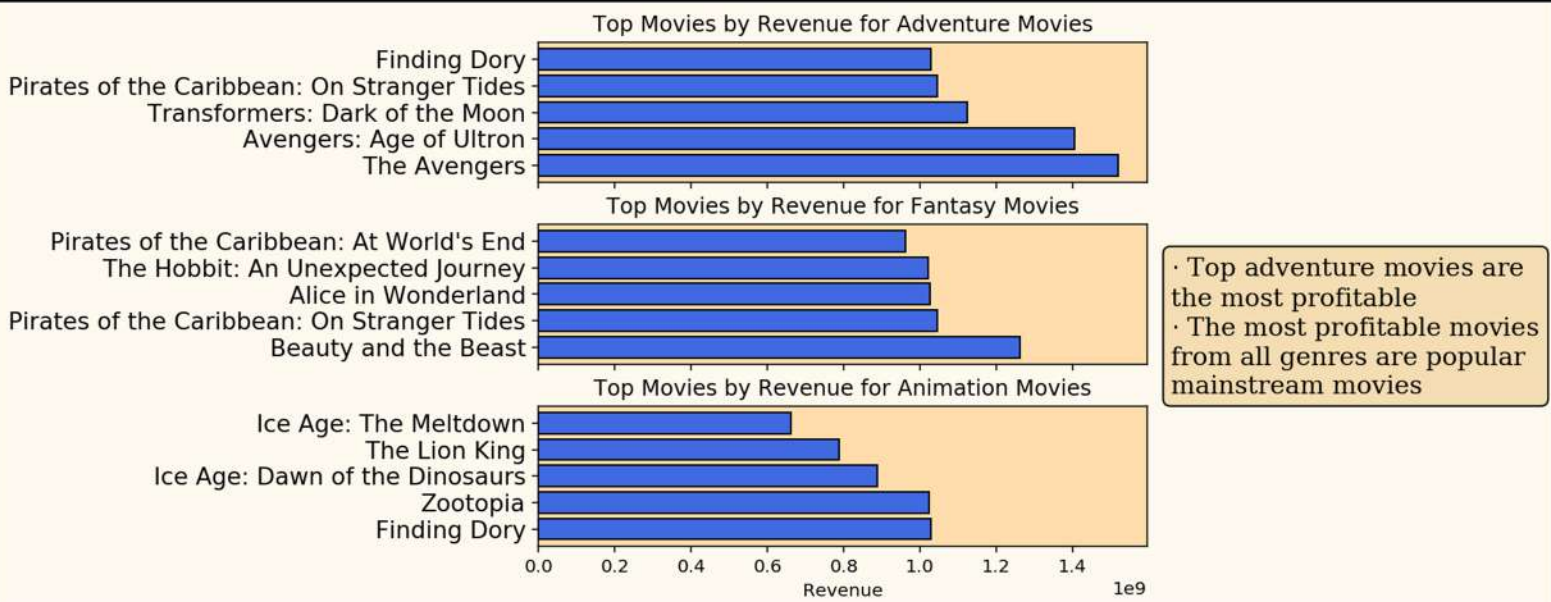
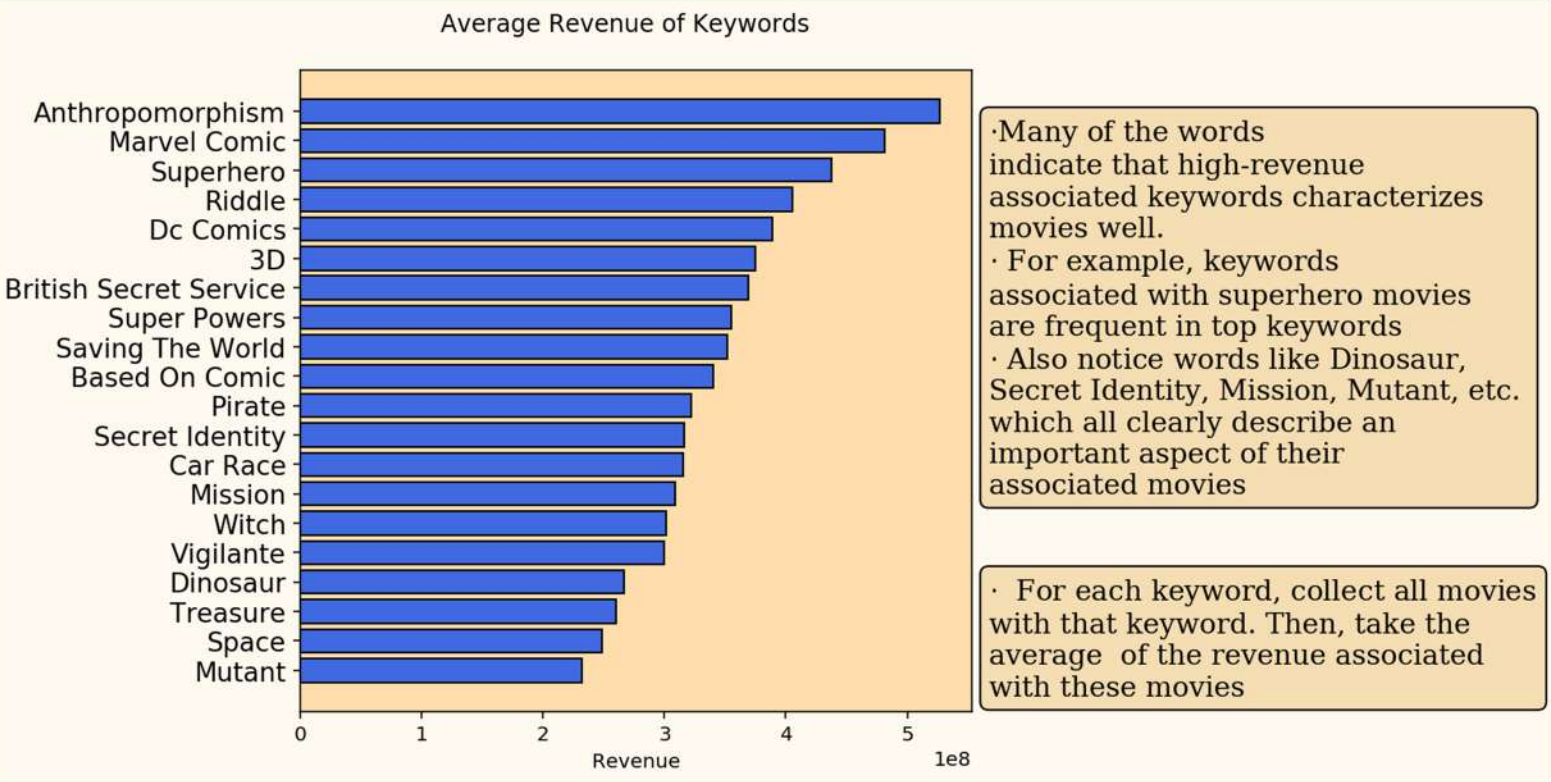
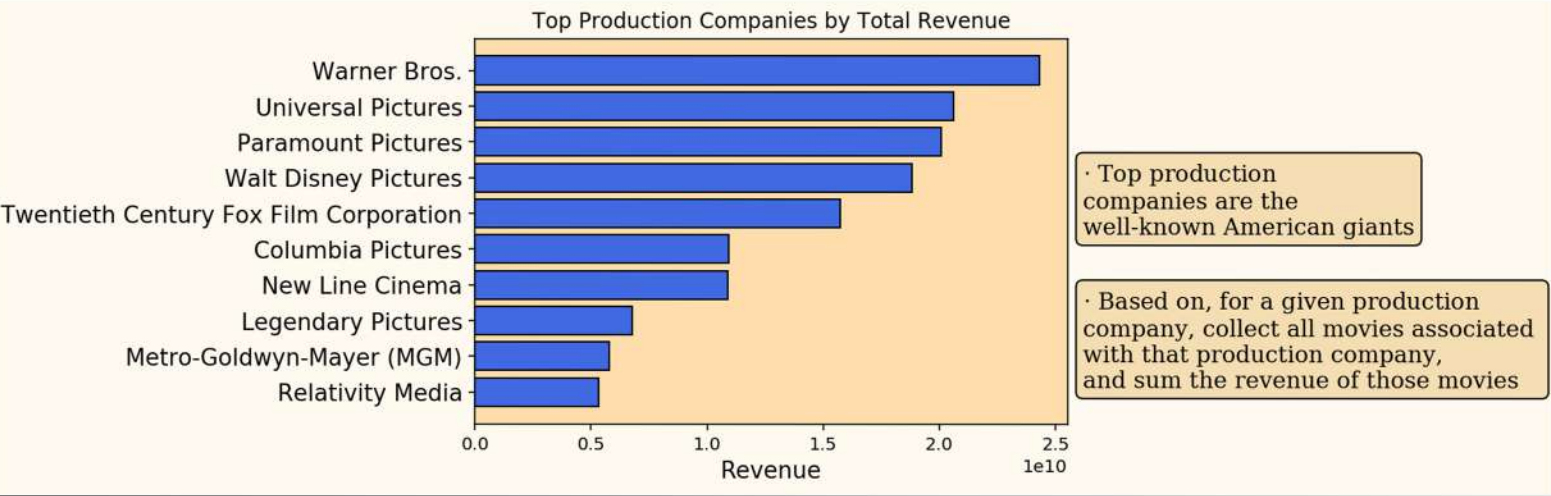


Runtime vs. Revenue

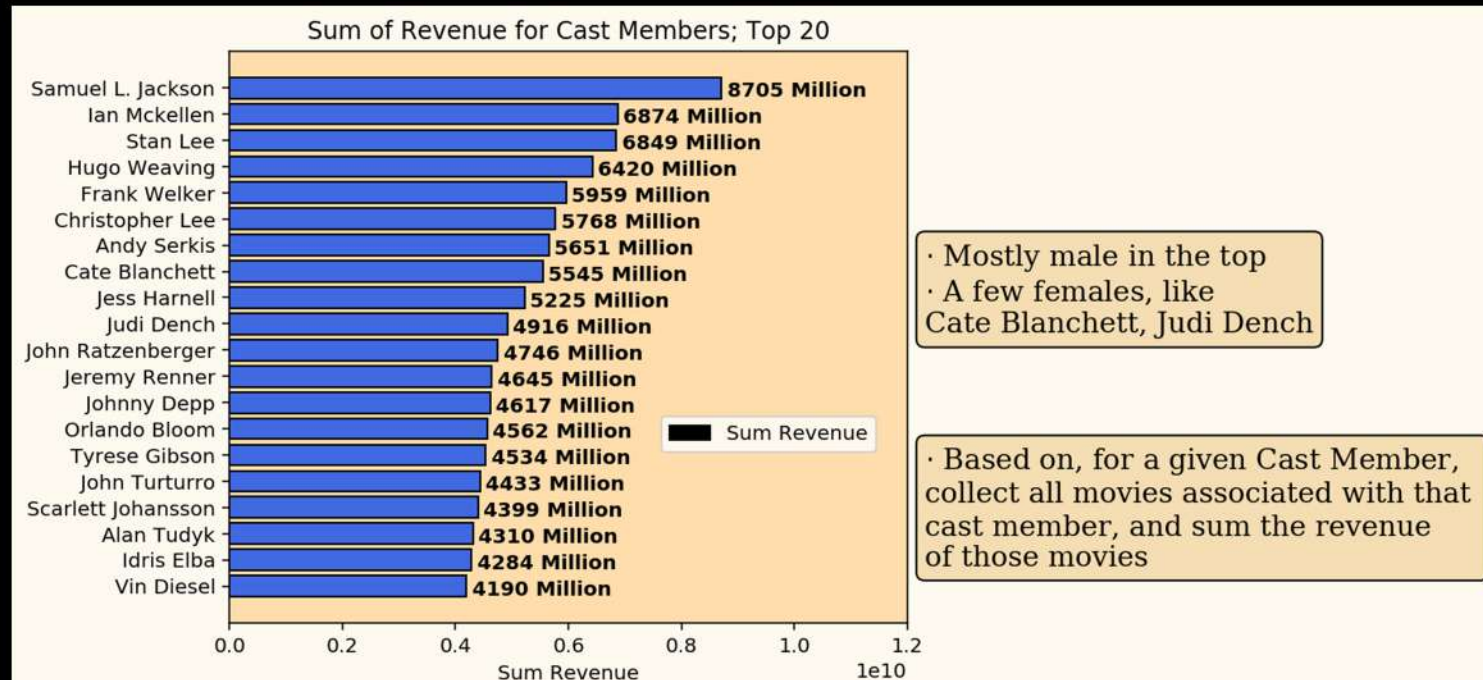
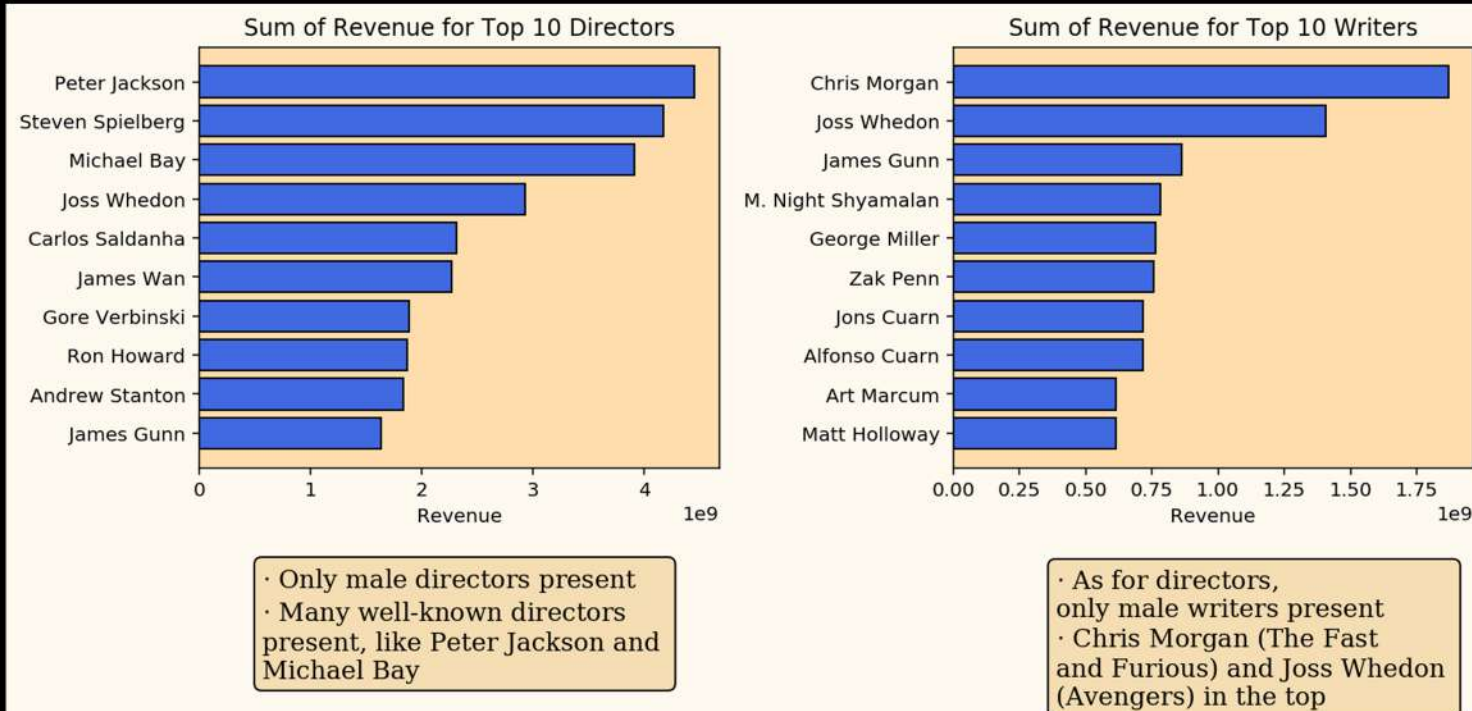
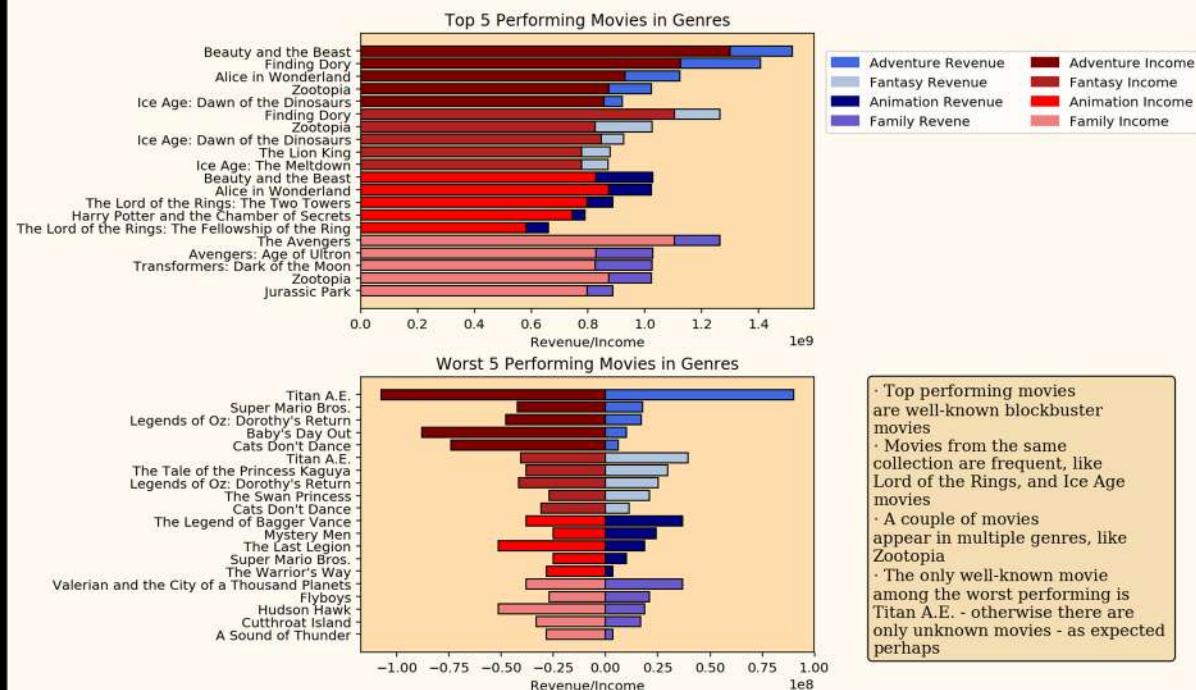


- More than 3/4 of movies are profitable - but there exists just below a quarter of movies that are at a loss
- Higher revenue seems to be associated with higher popularity, especially for moderate and higher values of revenue
- No particular dependence between runtime and revenue. Just a more diverse range of revenue values for movies with just below 2hrs of runtime

Data Analysis - TMDb Project - Production companies, keywords, and movies in different genres

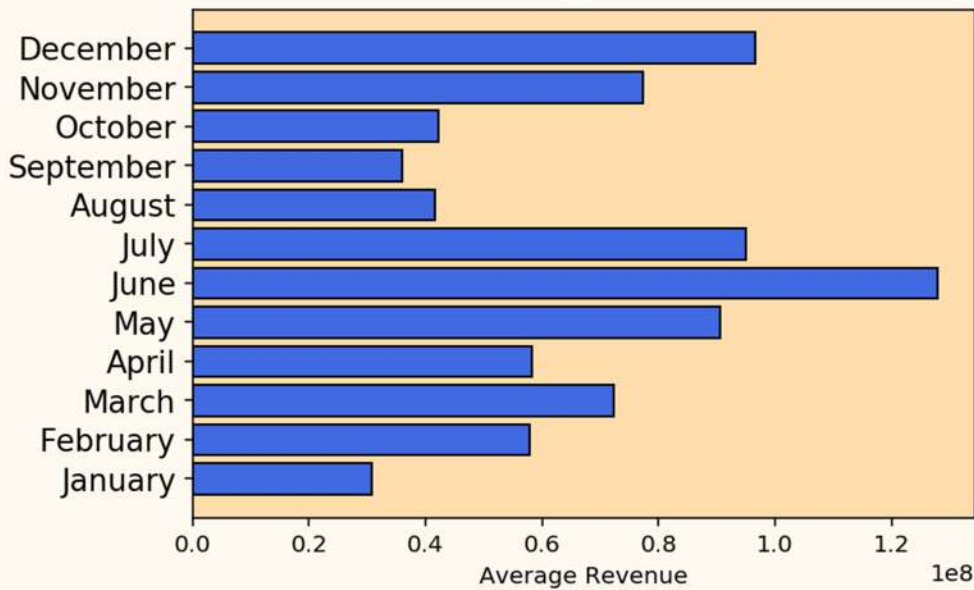


Data Analysis - TMDB Project - Genres, directors & writers, and cast



Data Analysis - TMDb Project - Revenue in months and genres, and time

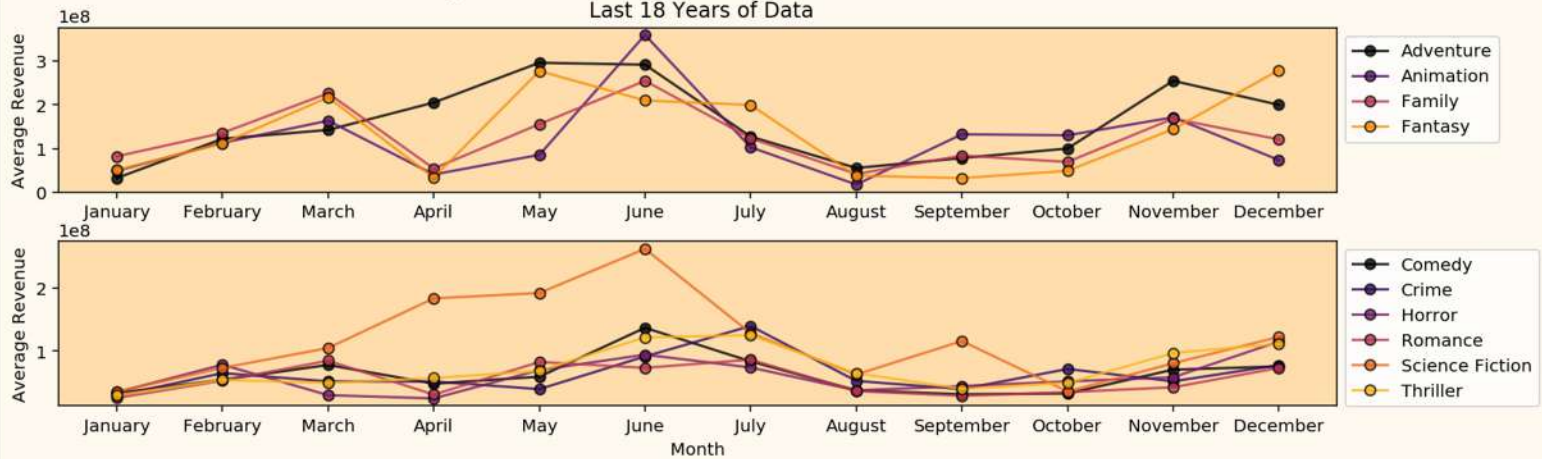
Average Revenue of Each Month
Last 18 Years



- Average revenue peaks during the summer and December
- Indicates that semester seasons are attractive for high revenue movies

- Based on, for a given month, collect all movies released in that month, and take the average of the revenue of the movies

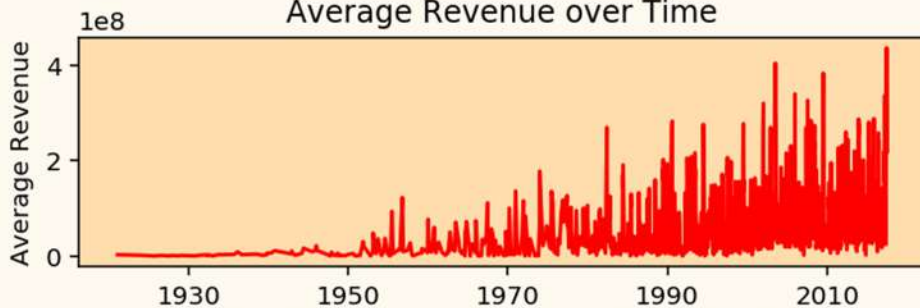
Average Revenue for Different Months for Different Genres
Last 18 Years of Data



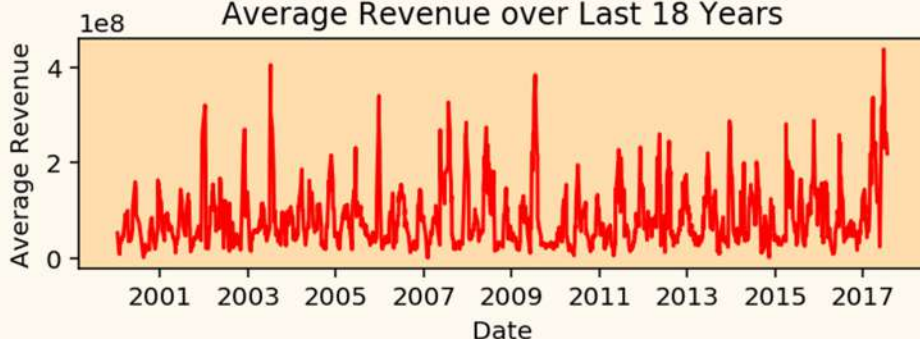
- Most genres have average revenue peaks during the semester seasons (summer & winter)
- The average revenue over months vary between different genres

Smoothed, Average Revenue in Intervals
of 30 Days

Average Revenue over Time



Average Revenue over Last 18 Years

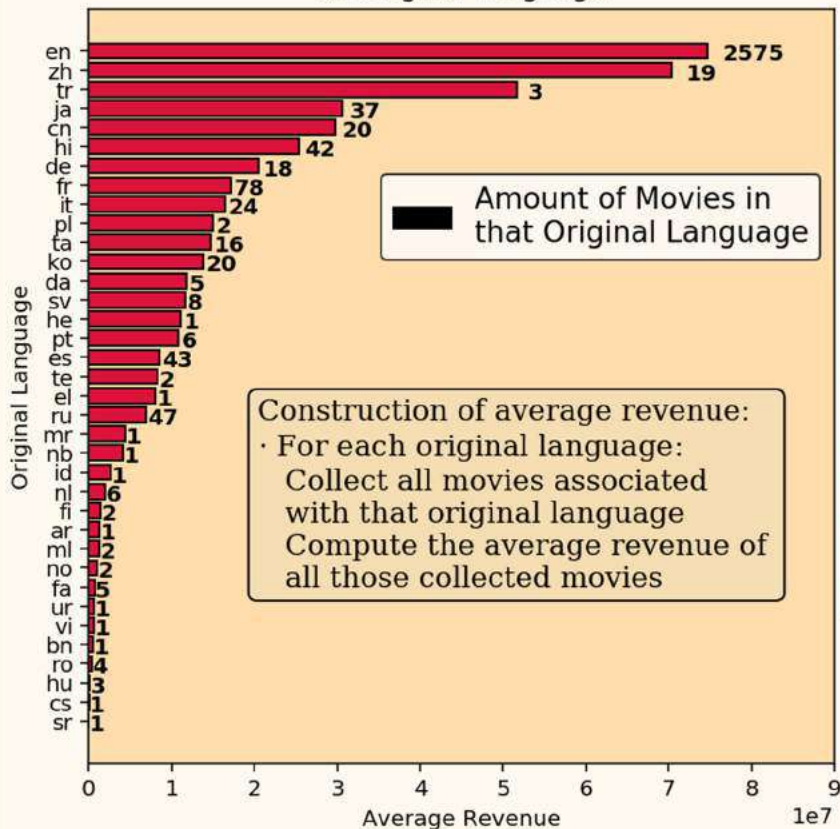


- Smoothed, average revenue increases over time - most likely due to inflation and/or better performing movies over time, and an increase in movies done over time

- For the last 18 years of average revenue, there is a clear seasonality and repetition in the average revenue over the years

Construct Attributes & Data Analysis - TMDb Project - Weighted original language and keywords revenue

Average Revenue of Movies Associated with an Original Language

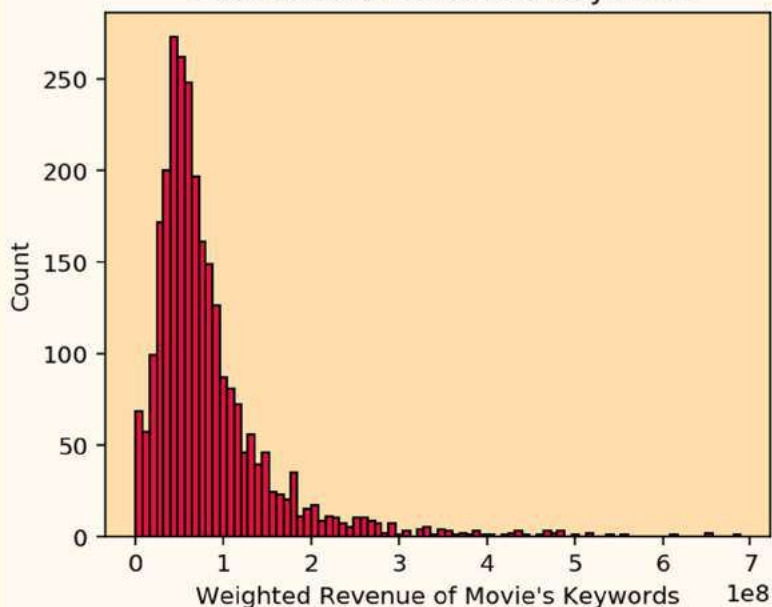


Construction of average revenue:

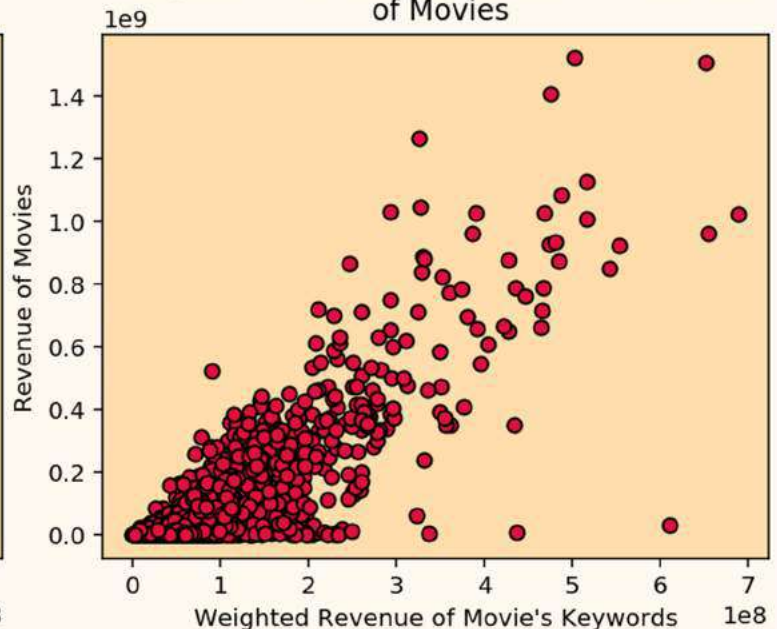
- For each original language:
 - Collect all movies associated with that original language
 - Compute the average revenue of all those collected movies

- Seems to be a clear relationship between revenue and different original languages
- English and Chinese seems to be related to higher revenue movies - makes sense as both USA/Britain and China are major developed countries, with a large middle-class
- The average revenue, for some original language, can give a rough idea of what a general movie, in that original language, might yield in box-office
- There is a clear relationship between average revenue and population size of countries, as can be seen by the German, Italian, French, Japan and Hindu original languages

Histogram of Weighted Revenue of each Movie's Associated Keywords



Weighted Revenue (Keywords) vs. Revenue, of Movies



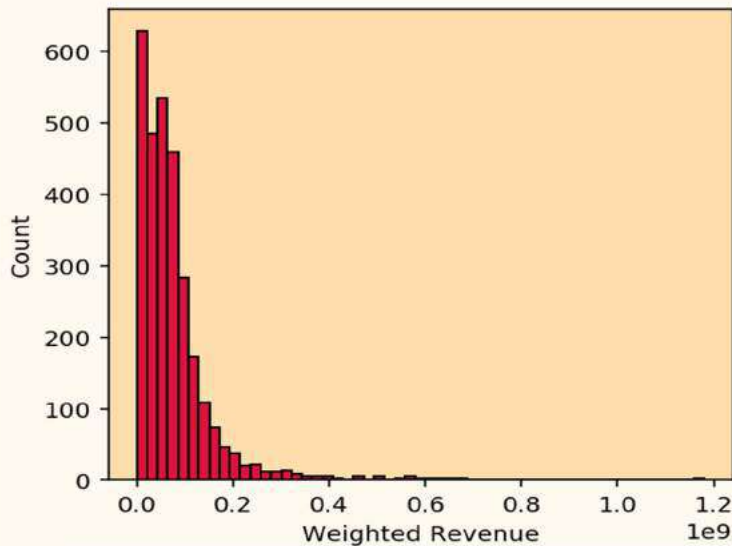
Construction of Weighted Revenue:

- (1) For each keyword:
 - Collect all movies associated with that keyword
 - Compute the average revenue of all those movies
- (2) For each movie:
 - Collect all keywords associated with that movie
 - Collect average revenue associated with each keyword, from (1)
 - Compute the weighted revenue by, an average, the average revenue of each keyword associated with the movie

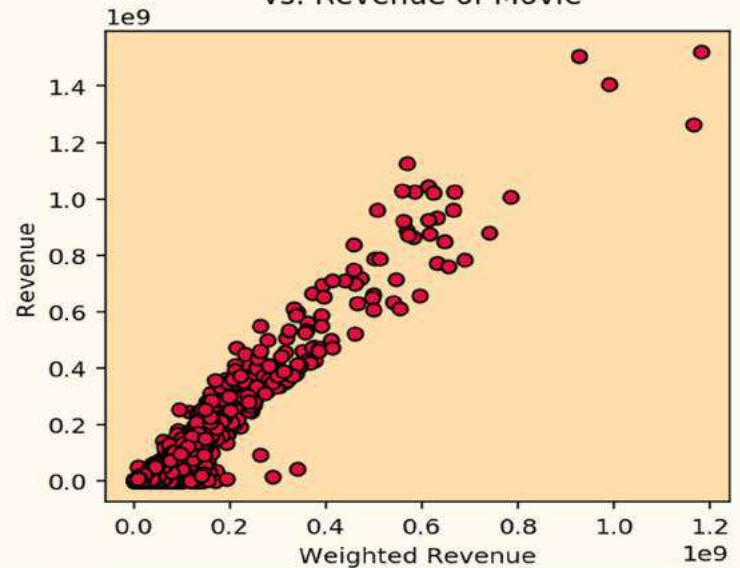
- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful modeling a relationship between a movie and its revenue

Construct Attributes & Data Analysis - TMDb Project - Cast and Production

Histogram of Weighted Revenue, of Cast Members in Movies



Weighted Revenue, of Cast Members, vs. Revenue of Movie

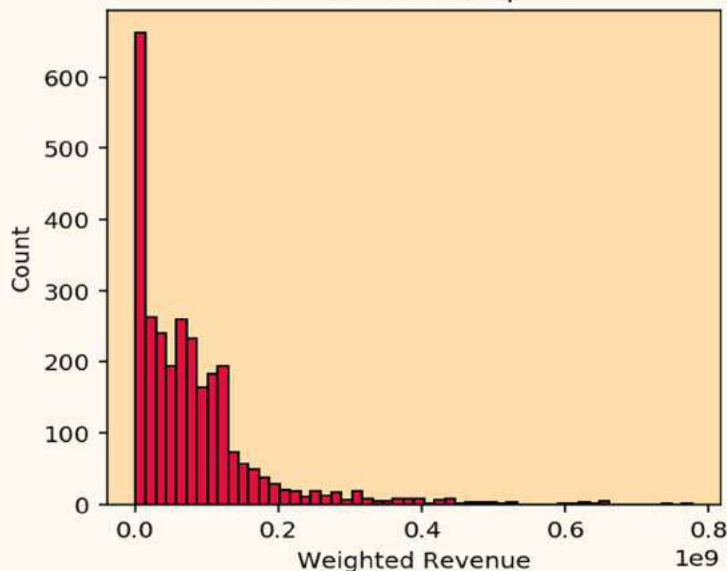


Construction of Weighted Revenue:

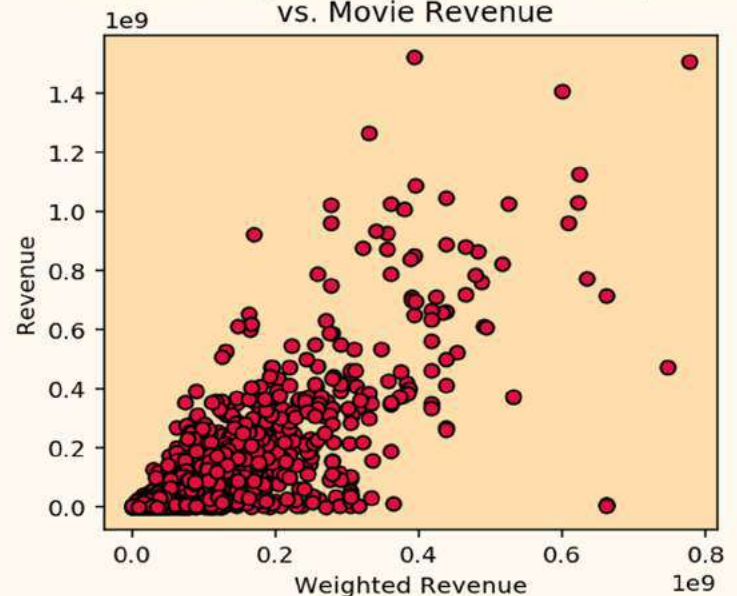
- (1) For each Cast Member:
Collect all movies associated with that cast member
Compute the average revenue of all those movies associated with that cast member
- (2) For each movie:
Collect all cast member(s) associated with that movie
Collect average revenue associated with each cast member from (1)
Compute the weighted revenue by, an average, the average revenue of each cast member associated with the movie

- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful modeling a relationship between a movie and its revenue
- A drawback is that it requires an actor(ress) to have been in at least one movie before it is an useful measure

Histogram of Weighted Revenue, of Production Companies



Revenue Weighted, of Production Companies, vs. Movie Revenue



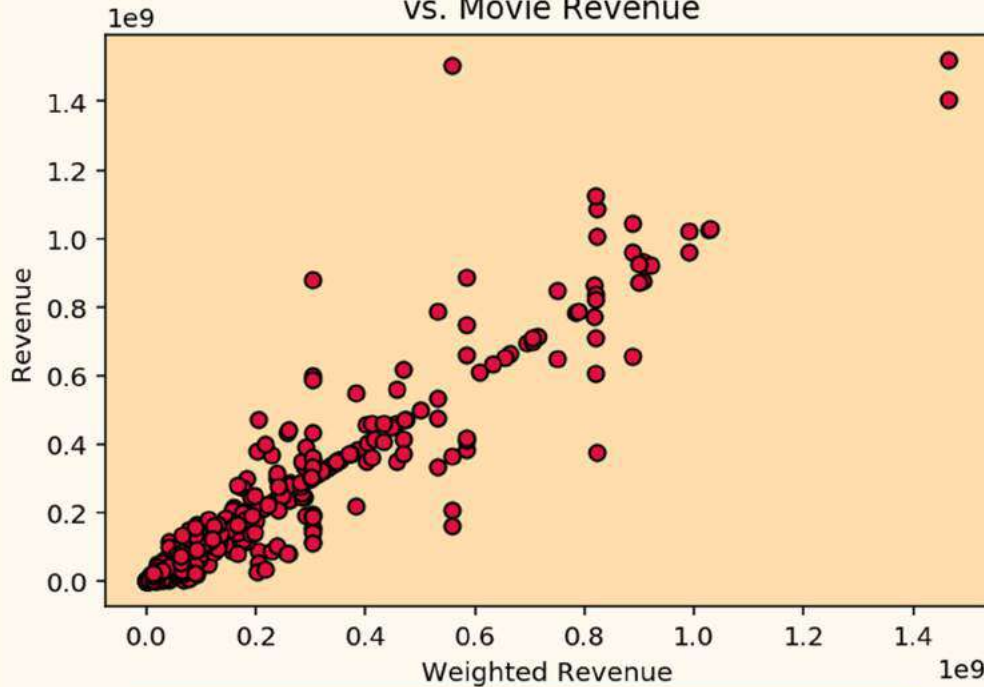
Construction of Weighted Revenue:

- (1) For each Production Company:
Collect all movies associated with that Production Company
Compute the average revenue of all those movies
- (2) For each movie:
Collect all Production Companies associated with that movie
Collect average revenue associated with each Production Company from (1)
Compute the weighted revenue by, an average, the average revenue of each Production Company associated with the movie

- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful modeling a relationship between a movie and its revenue
- A drawback is that it resembles the relationship between weighted revenue for keywords and movie revenue

Construct Attributes & Data Analysis - TMDB Project -Collections and Crew

Weighted Revenue, of Collections, vs. Movie Revenue

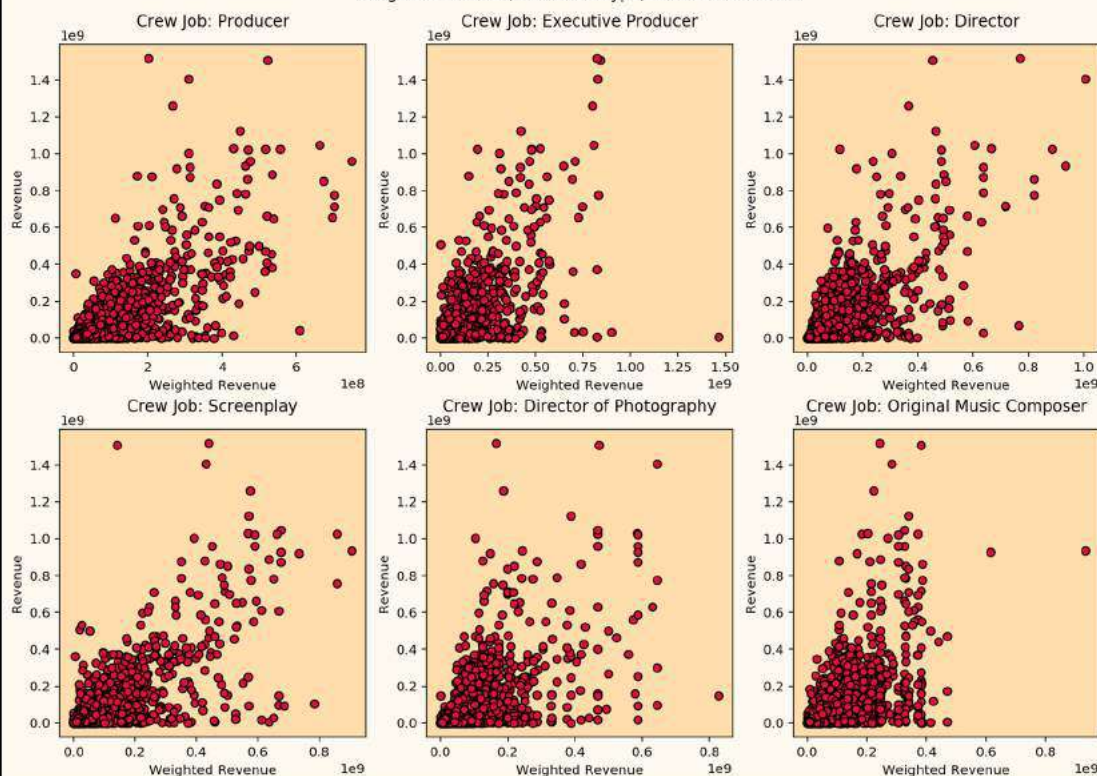


- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful modeling a relationship between a movie and its revenue
- A drawback is that most movies only belong to 1 collection
- On the other hand, it may be a useful measure for e.g. sequels

Construction of Weighted Revenue:

- (1) For each Collection:
Collect all movies associated with that Collection
Compute the average revenue of all those movies associated with that Collection
- (2) For each movie:
Collect all Collection(s) associated with that movie
Collect average revenue associated with each collection(s) from (1)
Compute the weighted revenue by, an average, the average revenue of each Collection associated with the movie

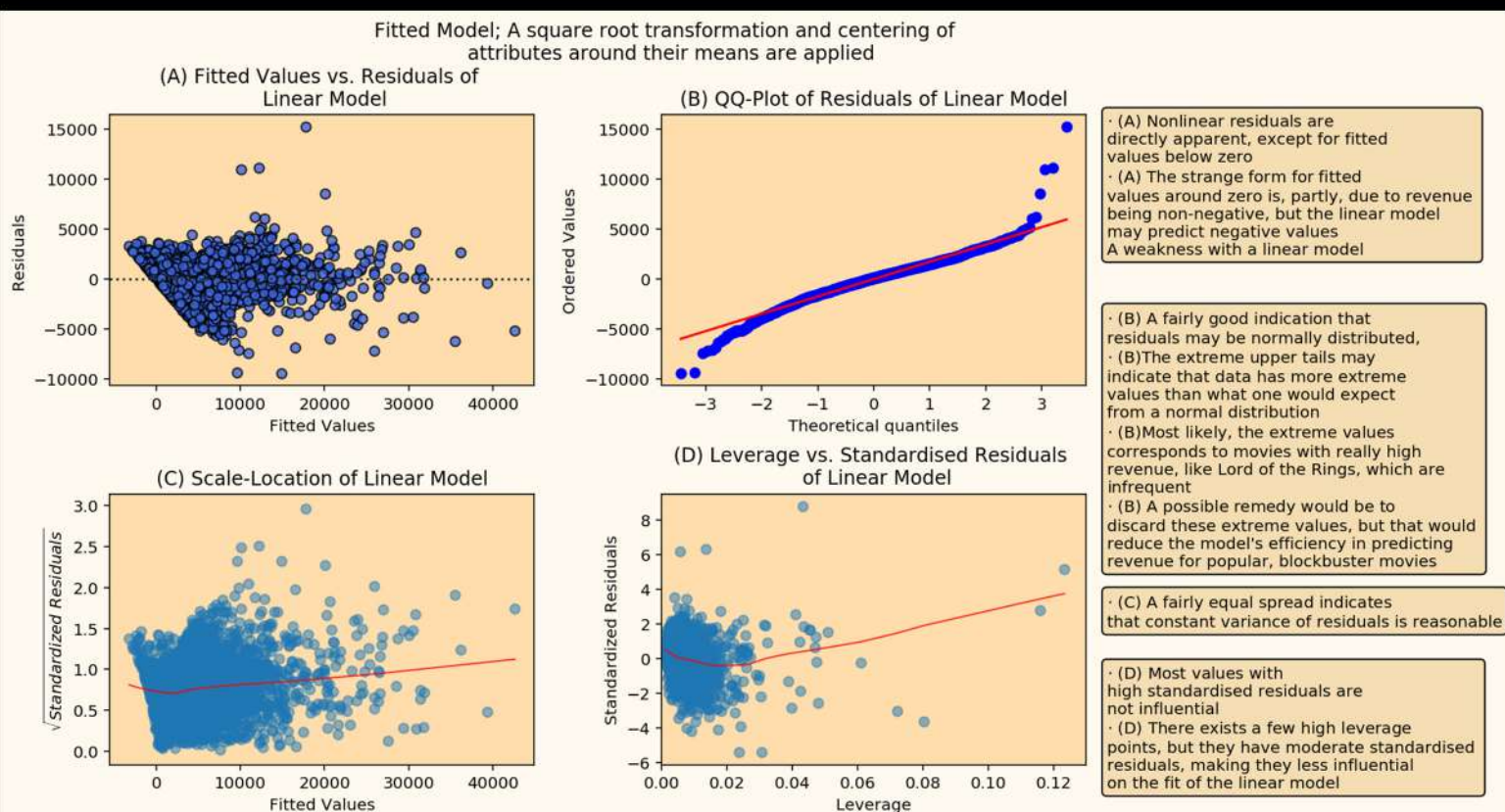
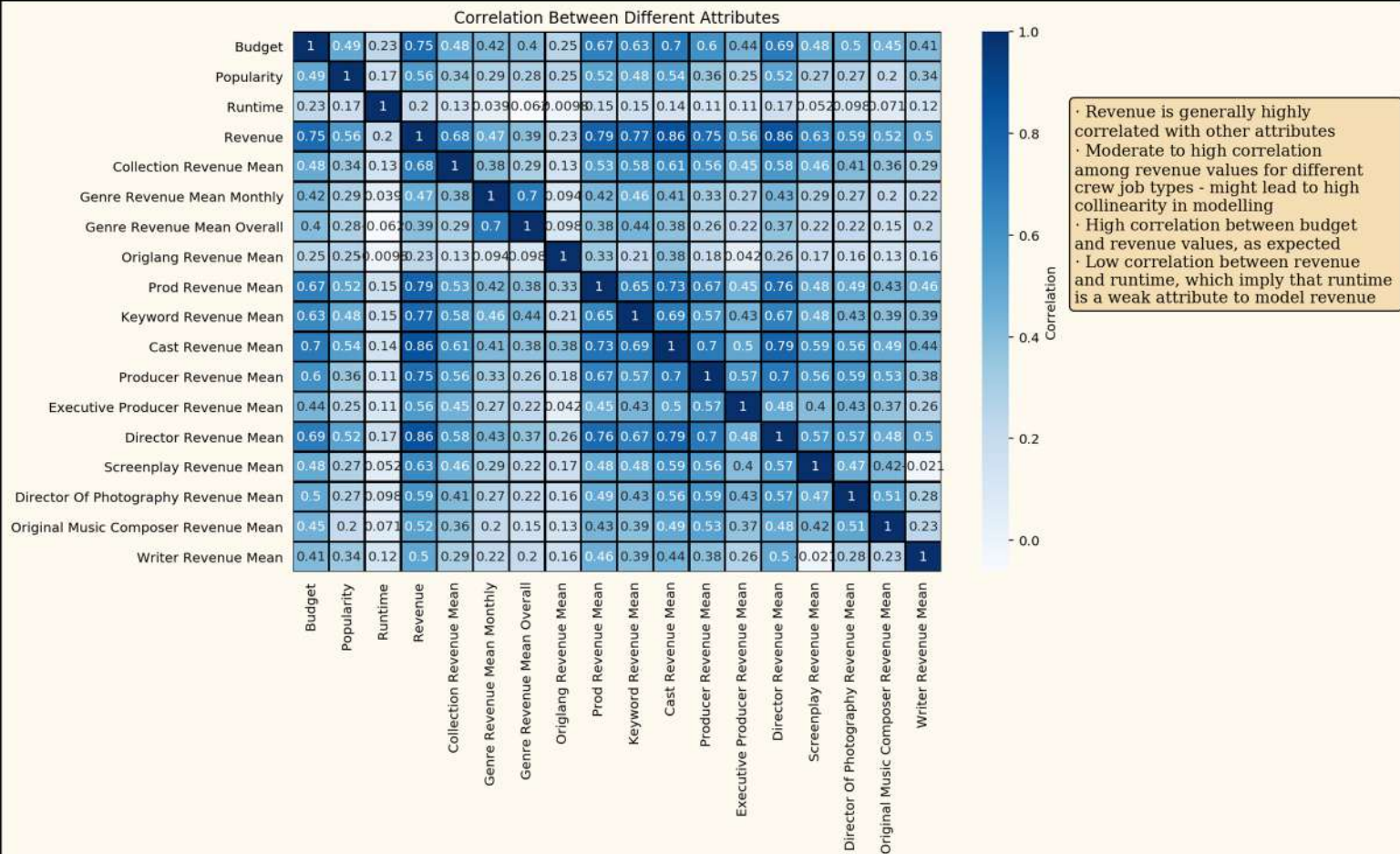
Weighted Revenue, of a Crew Type, vs. Movie Revenue



- ## Construction of Weighted Revenue:
- (1) For each Crew Job Type and Crew Name:
Collect all movies associated with that Crew Job Type and Crew Name
Compute the average revenue of all those movies associated with that Crew Job Type and Crew Name
 - (2) For each movie and Crew Job Type:
Collect all Crew Name(s) associated with that movie and Crew Job Type
Collect average revenue associated with the Crew Job Type and Crew Name(s) from (1)
Compute the weighted revenue by, an average, the average revenue of each Crew Name associated with that Crew Job Type

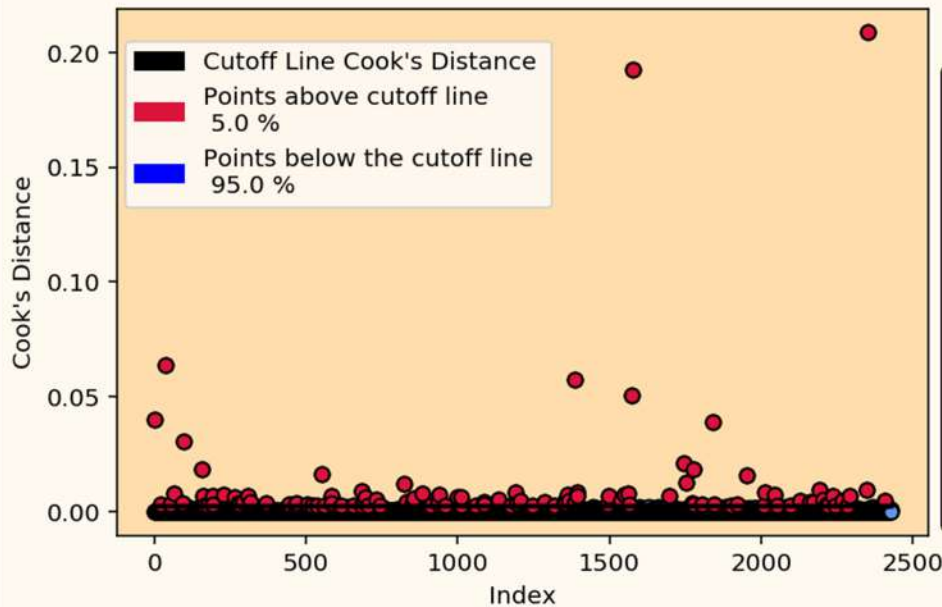
- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful in modeling a relationship between a movie and its revenue
- A drawback is that a crew job and crew name must have been associated with at least one movie before it is an useful measure

Building a Linear Model & Prediction Analysis - TMDb Project - Initial analyze of attributes with a correlation matrix, and a initial fitted model with diagnostic plots for analyzing adequacy of the linear model



Building a Linear Model & Prediction Analysis - TMDb Project - Cook's distance analyze of fitted values, and refit of linear model with extreme influential data points removed from dataset

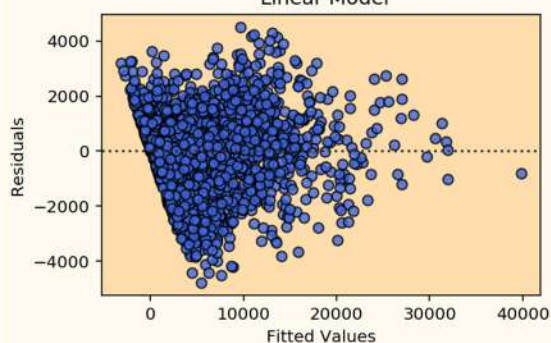
Cook's Distance of Fitted Values



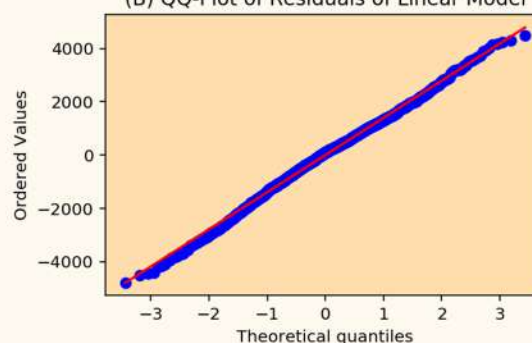
- The 5 % of points above the cutoff line indicates extreme values with great influence on the linear model
- One possible remedy is to remove these points from the linear model
- The influential points are most likely big budget movies, which may distort the prediction of revenue of low-moderate revenue movies - which are probably more frequent
- In case we want to take big budget movies into account in our model, we should leave the 5 % extreme values in the model
- Because they are only 5 %, it makes sense to remove these extreme values

Fitted Model; 5% of Data Points above the Cook's Distance cutoff is removed; Square root transformation and centering of attributes

(A) Fitted Values vs. Residuals of Linear Model



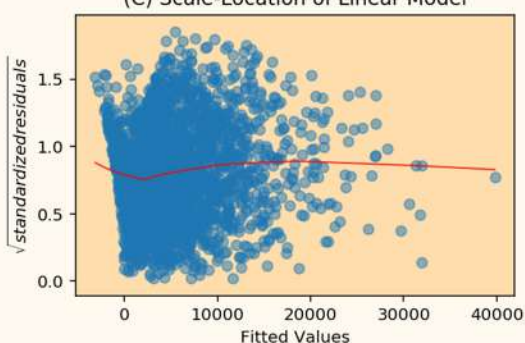
(B) QQ-Plot of Residuals of Linear Model



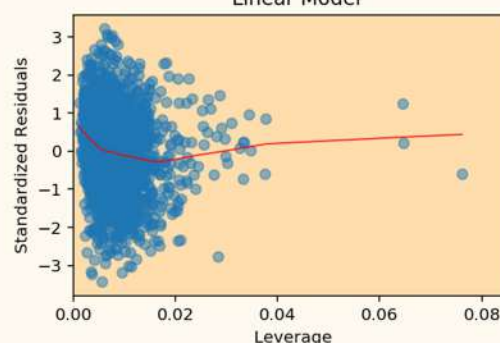
- (A) Nonlinear residuals is a reasonable assumption except, again, for fitted values below zero
- (A) The strange form for fitted values around zero is, again, due to revenue being non-negative, but the linear model may predict negative values

- (B) An excellent indication that residuals may be normally distributed,
- (B) The extreme upper tails, from the prior model, is remarkably gone - most likely due to the points removed from having relative high Cook's Distance
- (B) Note that the model might now be less efficient in predicting revenue for movies with high-revenue, e.g. Superhero Movies, or sequels to famous movies

(C) Scale-Location of Linear Model



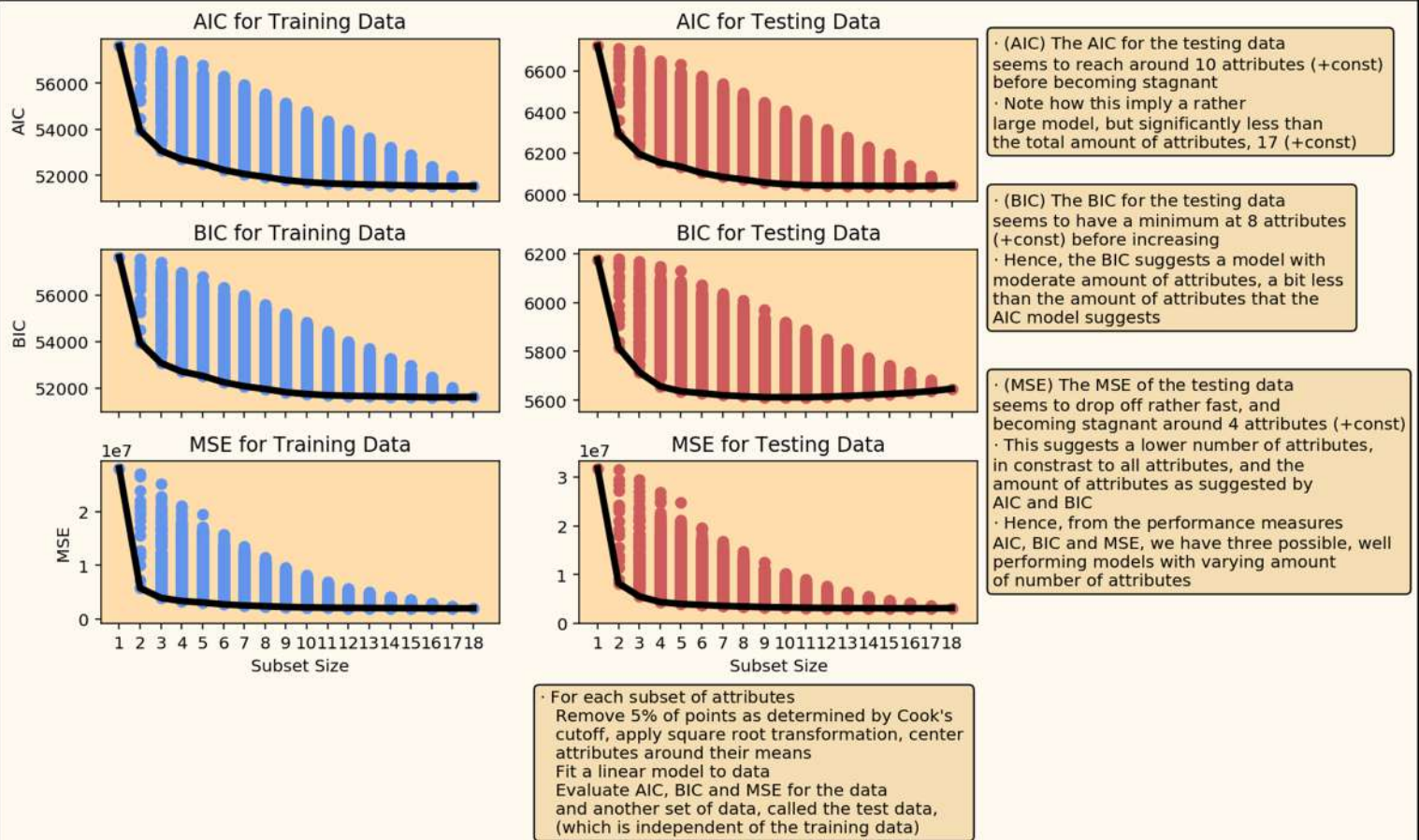
(D) Leverage vs. Standardised Residuals of Linear Model



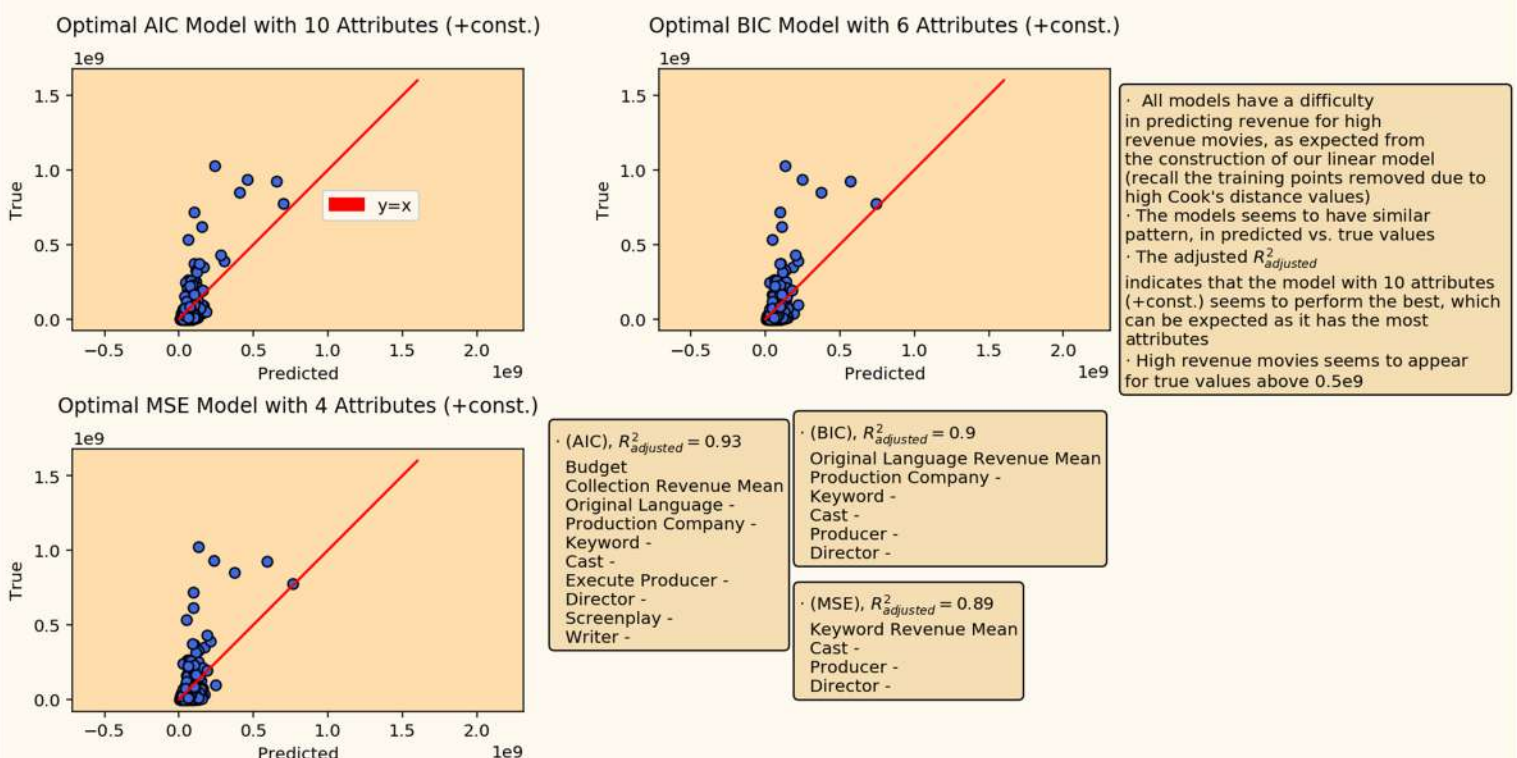
- (C) A fairly equal spread indicates that constant variance of residuals is reasonable, albeit not perfect

- (D) Most values with high standardised residuals are not influential
- (D) There exists a few high leverage points, but they have low standardised residuals, making them less influential on the fit of the linear model

Building a Linear Model & Prediction Analysis - TMDb Project - Model selection with usage of AIC, BIC and mean squared error on training and testing data, and evaluation of three selected models



Predicted and True Revenue for the Different Models; Transformation Undone



Building a Linear Model & Prediction Analysis - TMDb Project - Evaluation of selected AIC model

