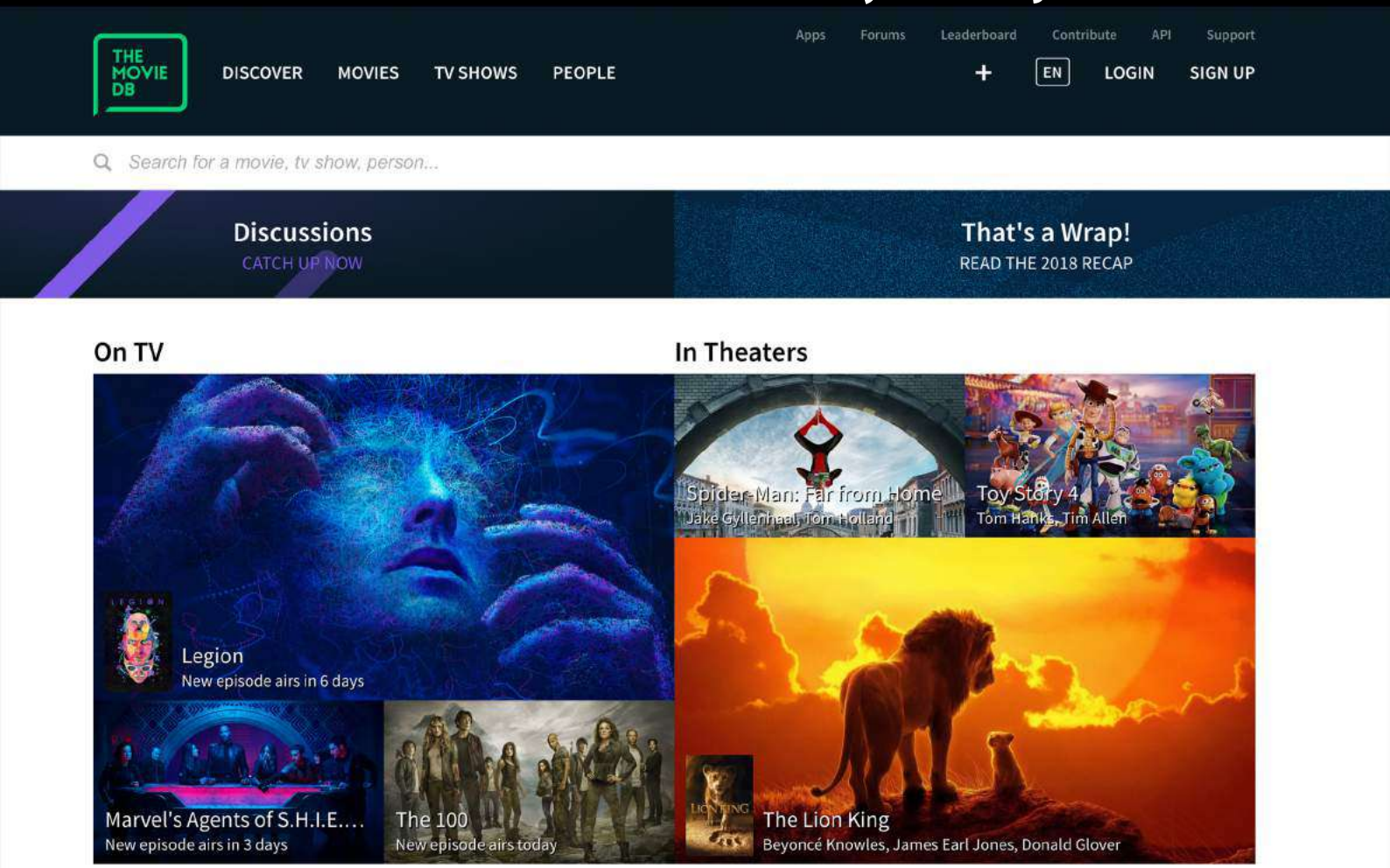
TMDB Movie Revenue Analysis Project



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?

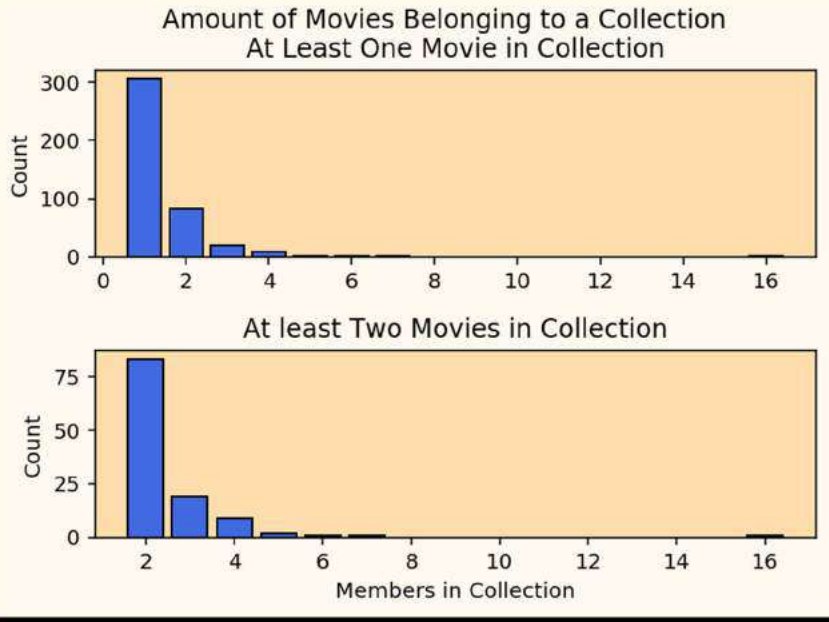
Data points in the metadata includes cast, crew, plot keywords, budget, posters, release dates, languages, production companies and countries.

In this report, a thourough 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.

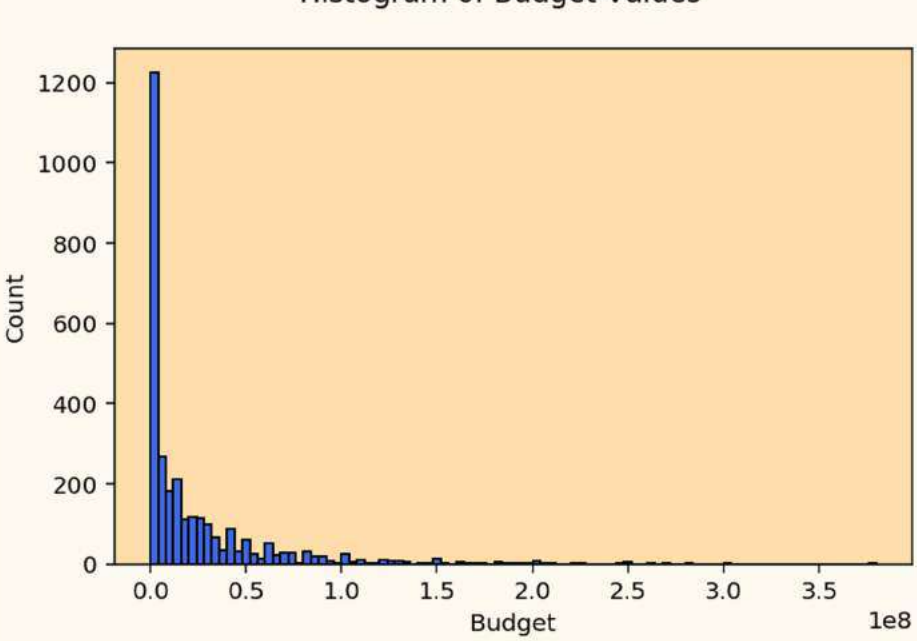
The results indicate that there is a lot of important findings in the data, and that the revenue for many movies can be predicted remarkably well.

The public dataset and challenge was found on Kaggle:
https://www.kaggle.com/c/tmdb-box-office-prediction
The code can be found at:
https://github.com/wildanwildan94/TMDB-Data-Analysis-Prediction-Inference

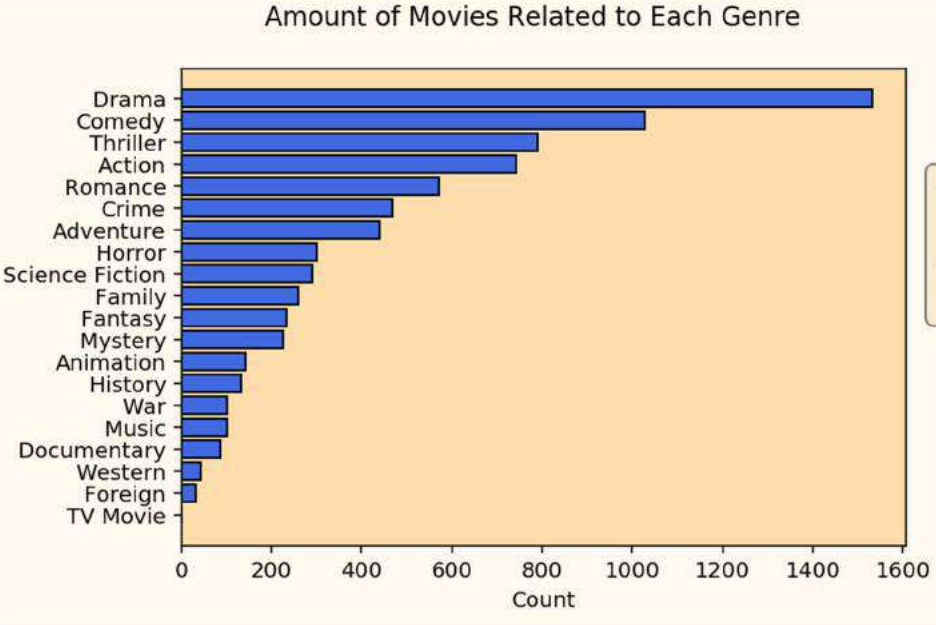
Data Analysis - TMDB Project - Amount of movies in collections, budget values, and amount of movies in genres



· 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

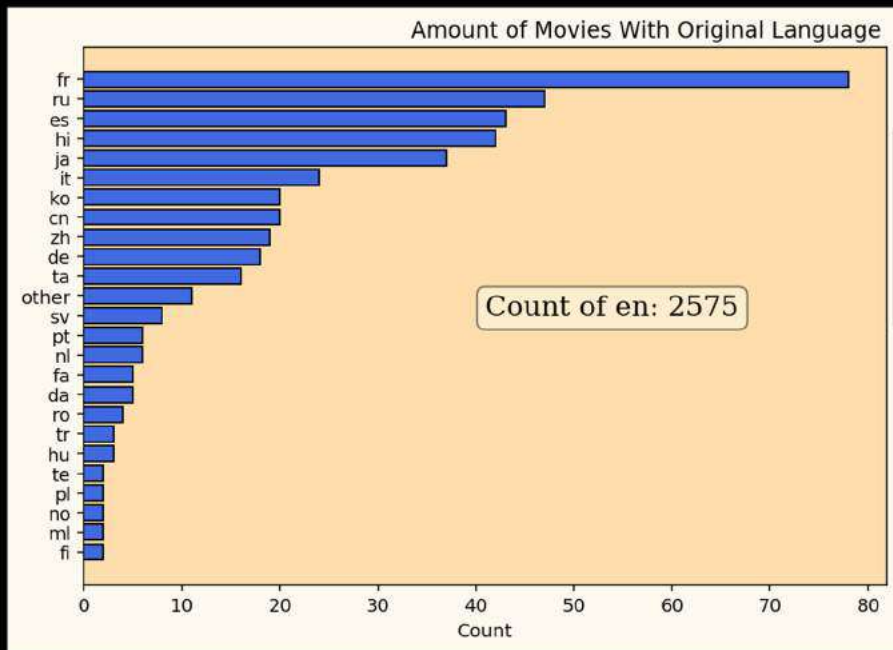


· Most movies are low-budget
· Budget decreases exponentially
· A few high-budget movies - mosty likely blockbuster movies

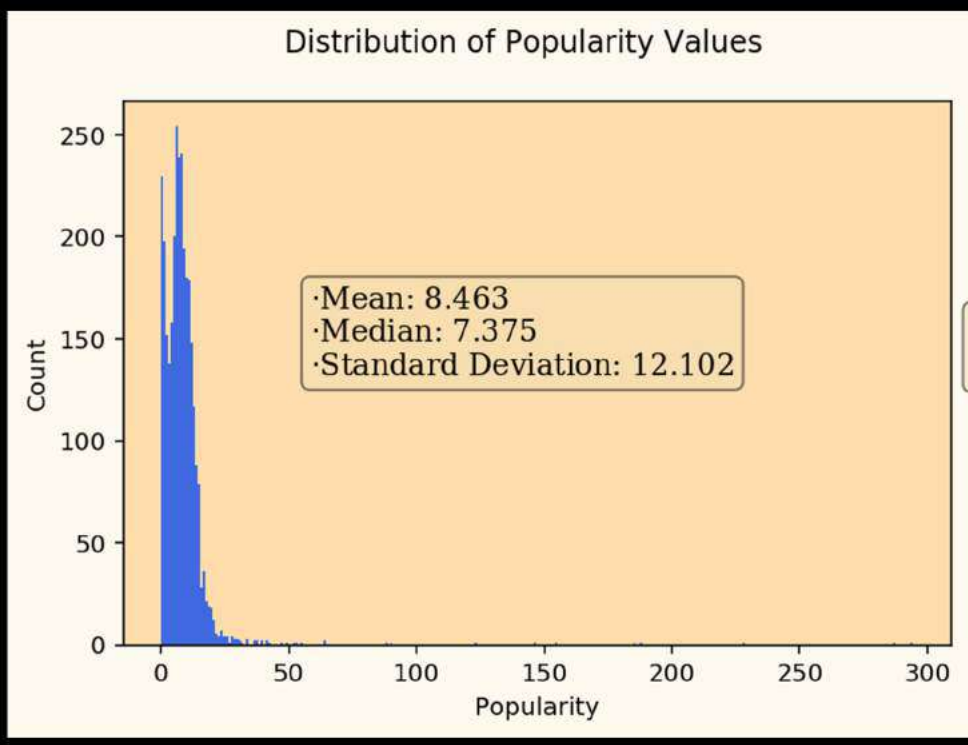


·The most popular genres are drama, comedy, thriller and action
· Relatively few music and documentary type movies

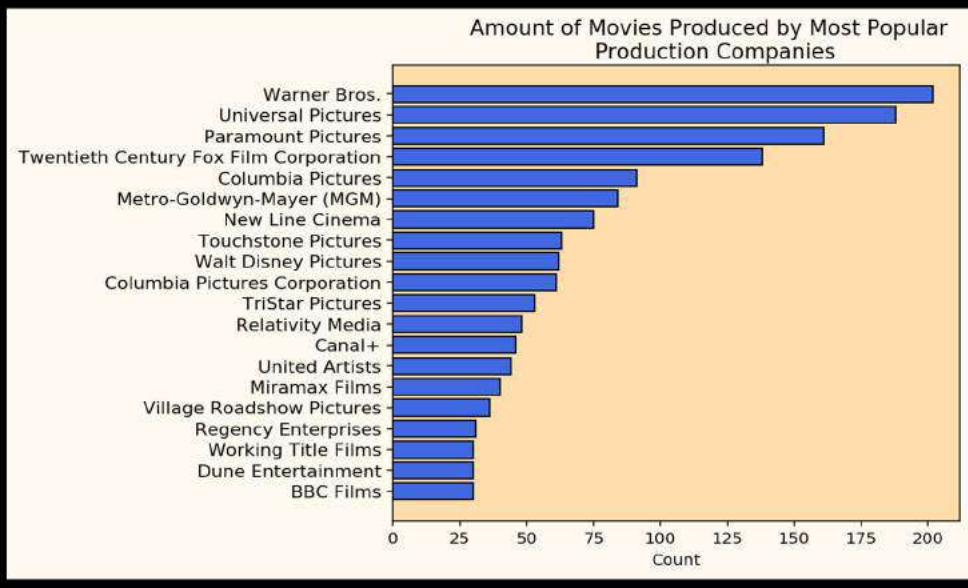
Data Analysis - TMDB Project - Original language, popularity, and production companies



·Other: Amount of spoken languagues 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

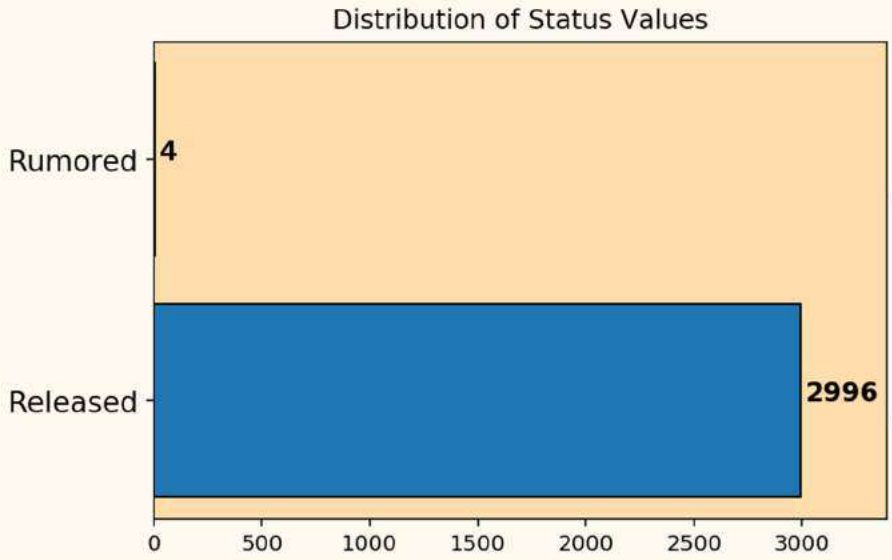


·Most movies are moderately popular
·A few movies with huge popularity

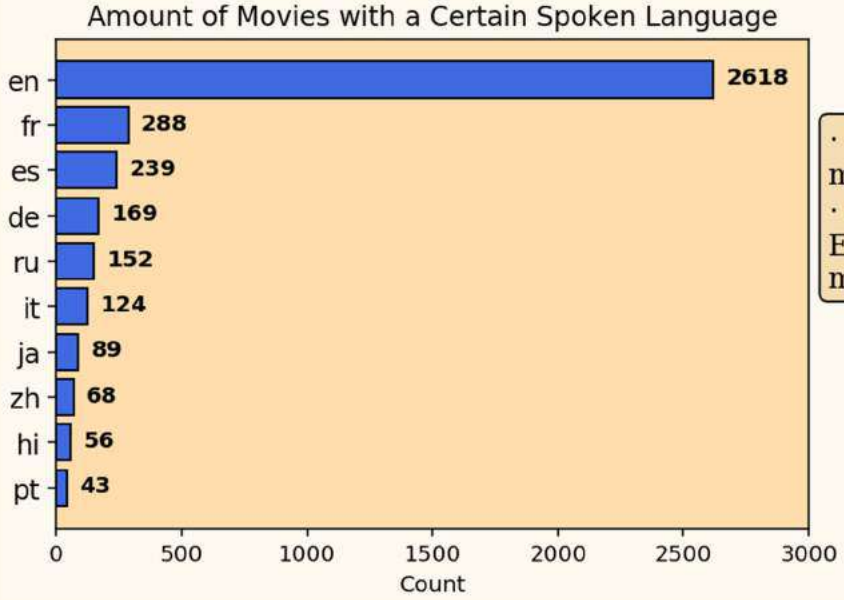


· Most popular production companies are American production companies, such as Warner Bros, Paramount Pictures, etc.
· A couple of European production companies, like BBC Films, Canal+, etc.

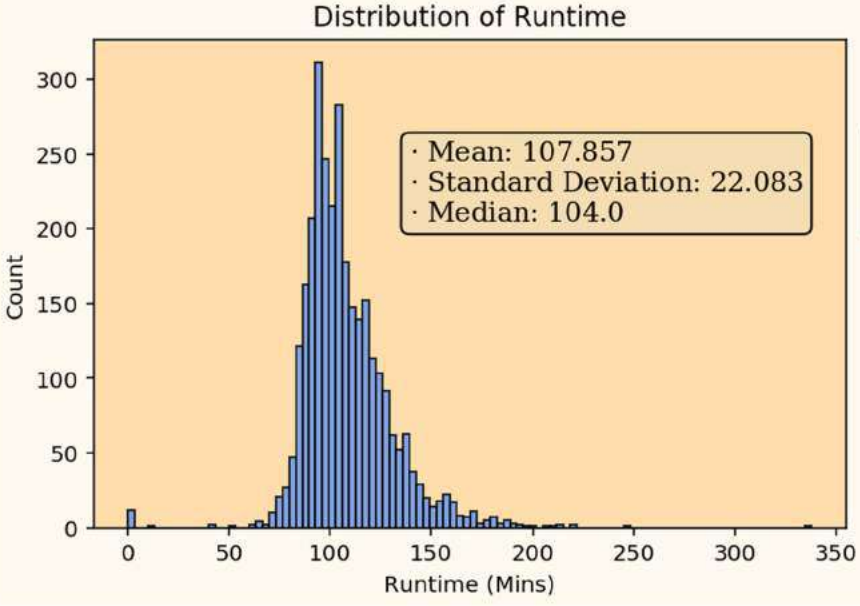
Data Analysis - TMDB Project - Status, spoken languages, and runtime



· 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 discared attribute

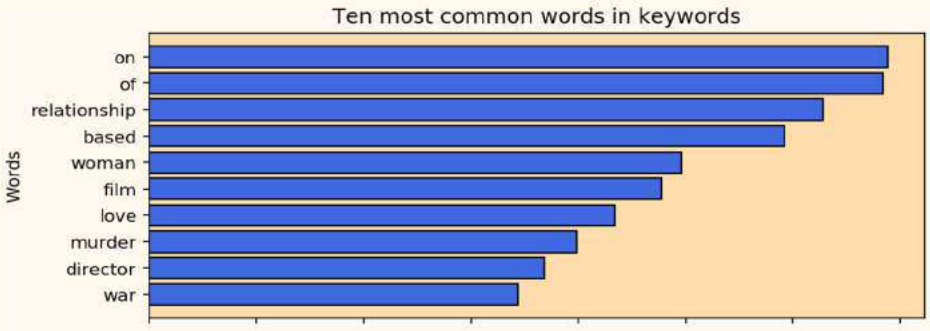


· As with original language, most movies have English as a spoken language
· French, Spanish, German as top European languages makes sense, as they are major European nations

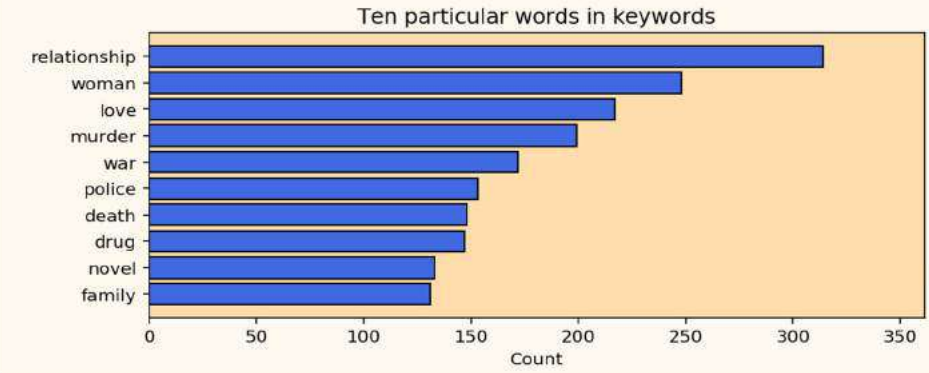


· Most movies have a runtime just below 2 hours
· Note that there is a moderate proportion of movies longer than 2 hours.

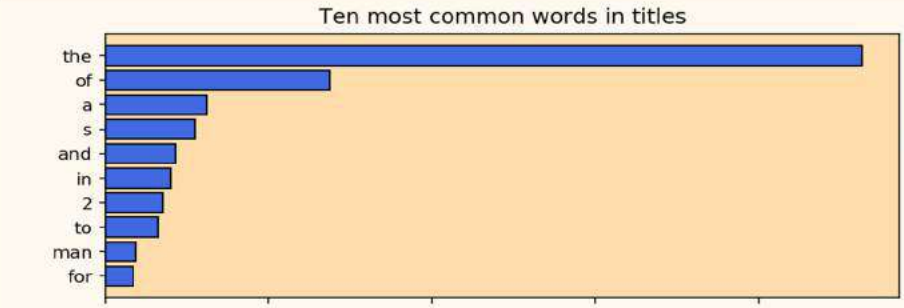
Data Analysis - TMDB Project - Keywords, titles, and taglines



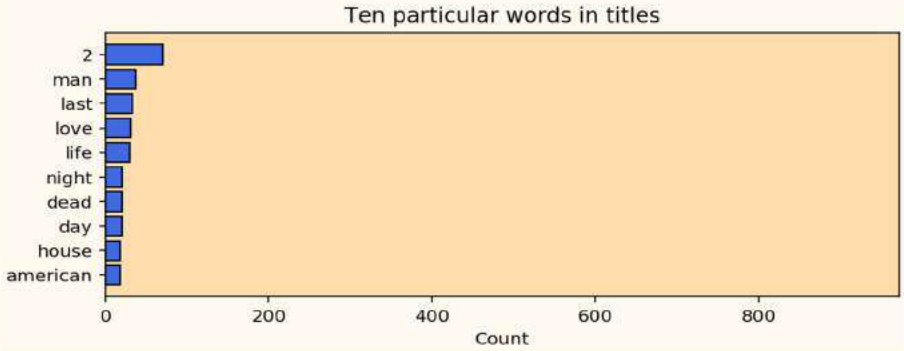
· The keywords are dominated by words that are very descriptive of movies
· Most common words are interesting words, with very low amount of sentence-building words
· Major takeaway is that keywords are suitable to aggregate main characterstics of movies



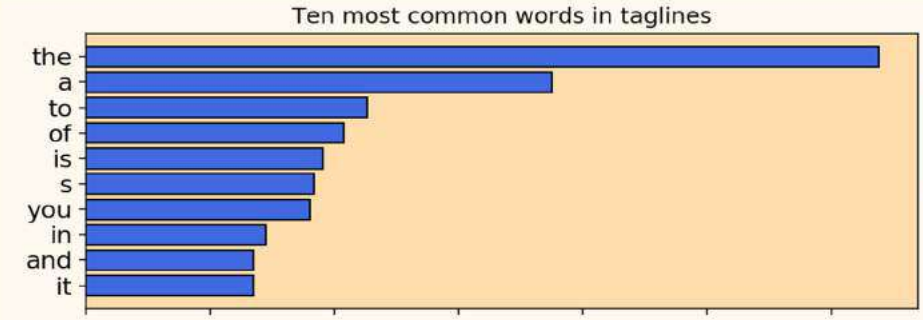
· Manually chosen words have high counts, relative the common words
· The usage of 'novel' indicates a lot of movies are associated with novel(s)
· Words like relationship, love murder, war, death, drug, family indicates a wide range of categories of movies exists



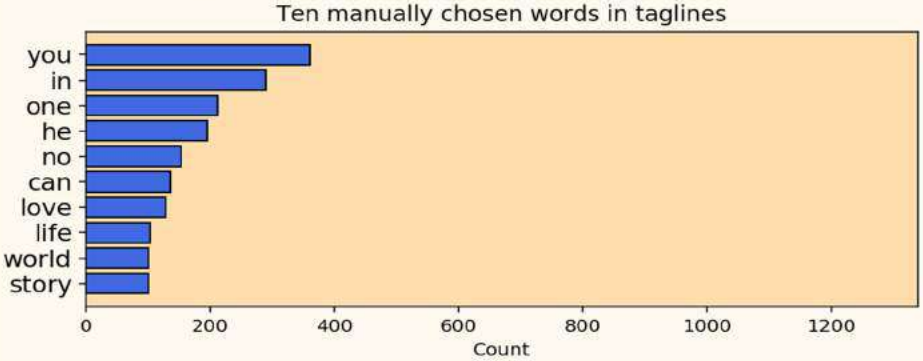
· A mix of common and interesting words in title
· 'The' is most common, by a wide margin - shows a popularity in a 'The'-structure of titles



· Manually chosen words have moderately high counts - indicates that it may be possible, again, to utilize title words to to aggregate main characterstics 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

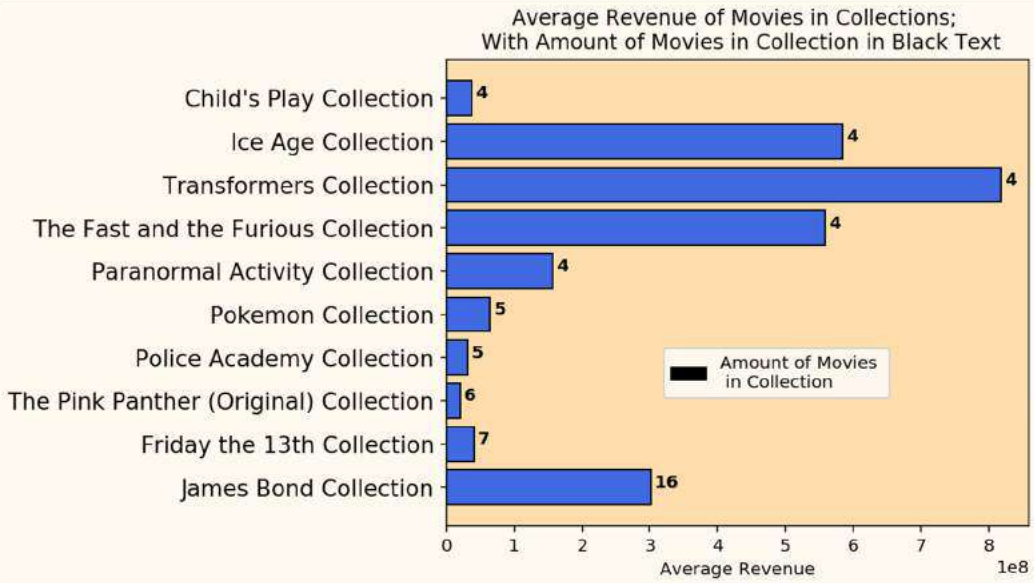


· Most common words are uninteresting, sentence-building words
· Not indicative of any prevalent characteristics of movies

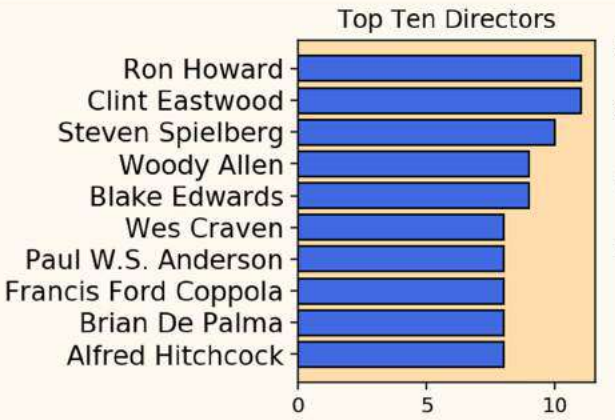


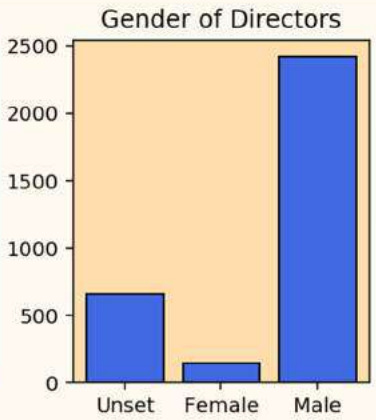
· Manually chosen words have moderately high counts - indicates that it may be possible to utilize tagline words to to aggregate main characterstics of movies.
· Love, life indicates a lot of movies associated with a positive message
· Word, story shows the adventurous associated movies are prevalent

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

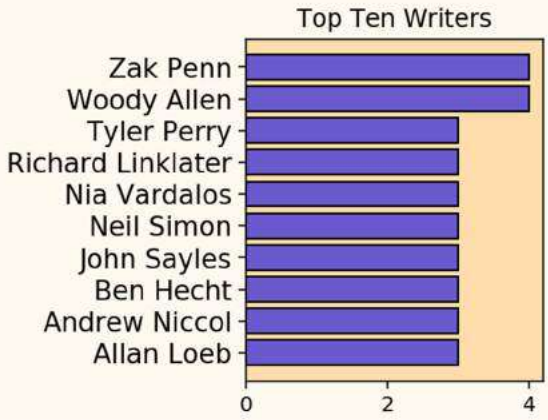


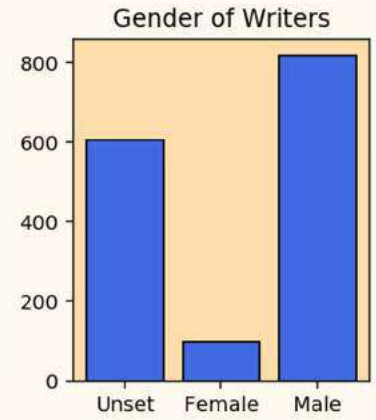
· The collections with the most movies are mainstream collections, - like Transformers and James Bond
· Most popular collections have between 4-5 movies
· The collections with the most, average revenue are Transformers, Ice Age and The Fast and the Furious, which is not a surprise



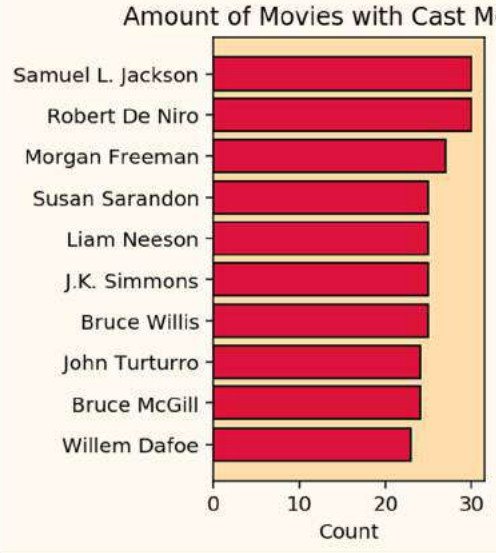


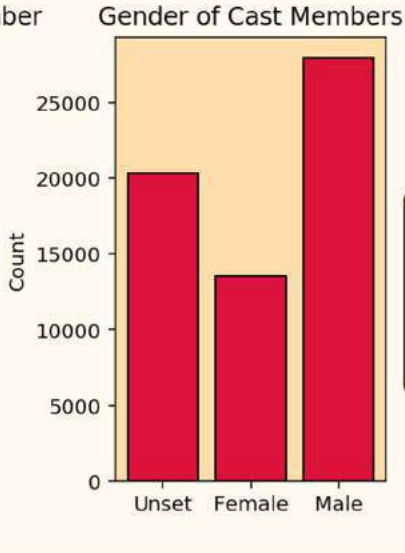
· Most popular directors are male
· A wide difference in amount of male and female directors





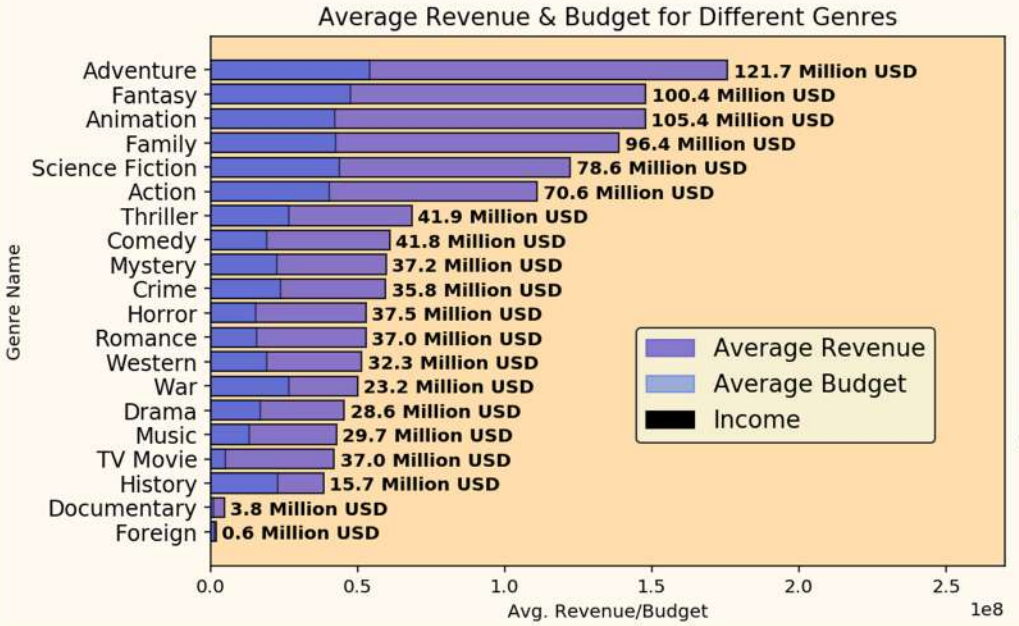
· Similar to directors, most popular writers are male
· A huge gender imbalance in amount of of male and female writers



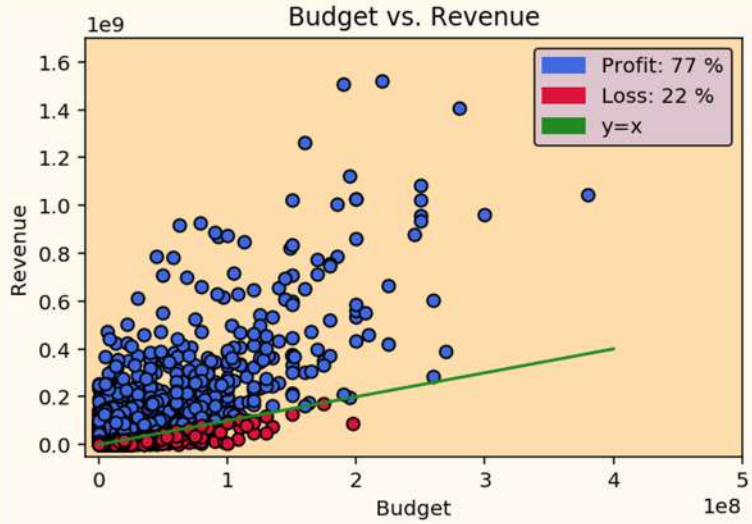


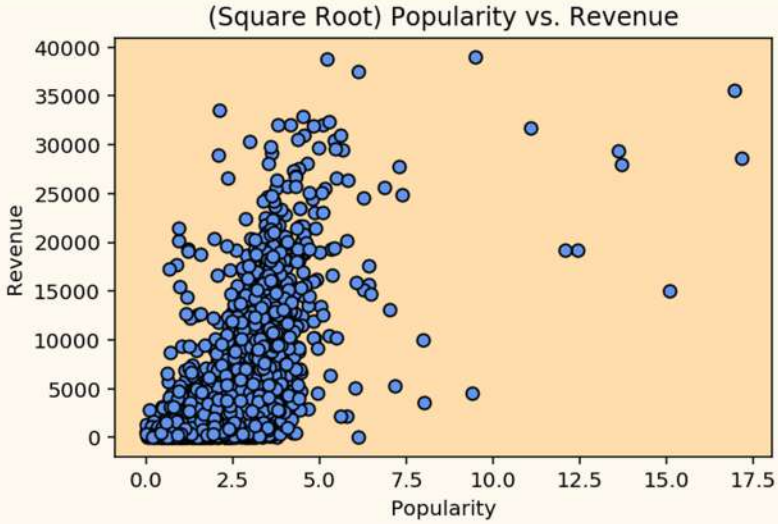
· Most popular cast members are male, with just a single female
· There is a large difference in amount of male and female cast in movies, indicating a large gender imbalance in the cast of movies

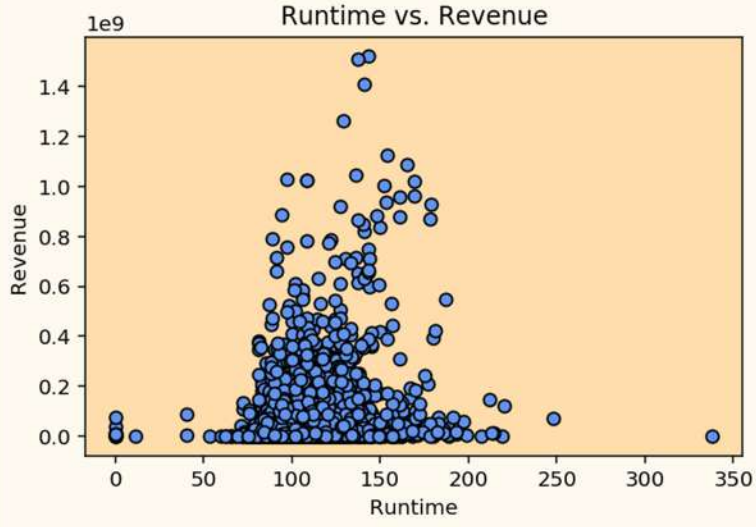
Data Analysis - TMDB Project - Genres, budget, popularity, and runtime



· 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

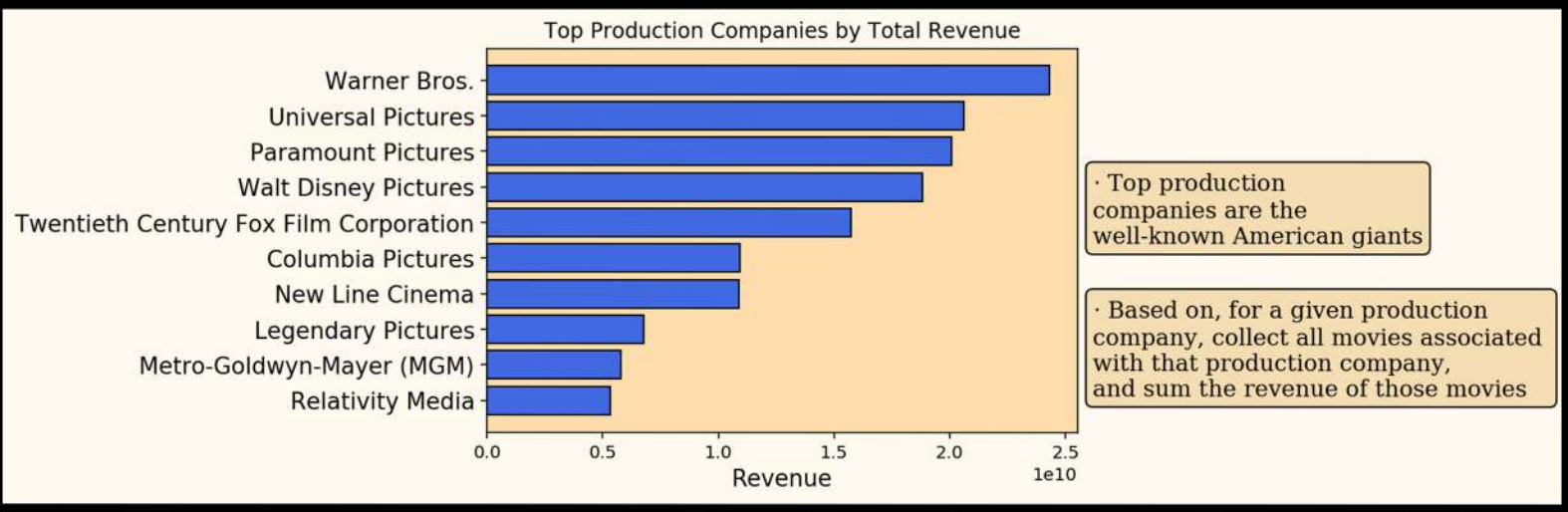






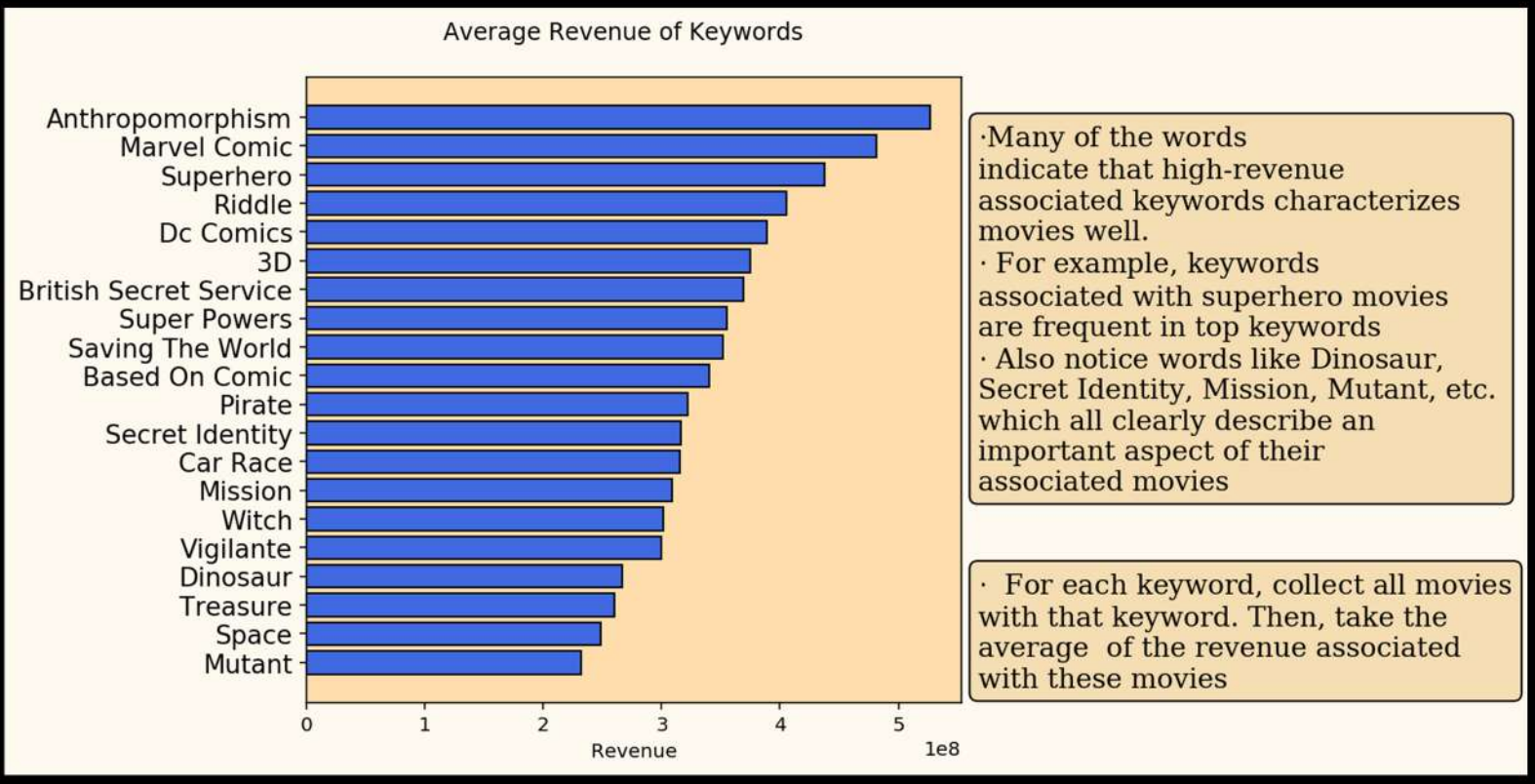
· More than 3/4 of movies are profitable - but there exists just below a quarter of movies that are at a loss
· Higher revenue seems to be associated with higher popularity, especially for moderate and higher values of revenue
· No particular dependence between runtime and revenue. Just a more diverse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

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



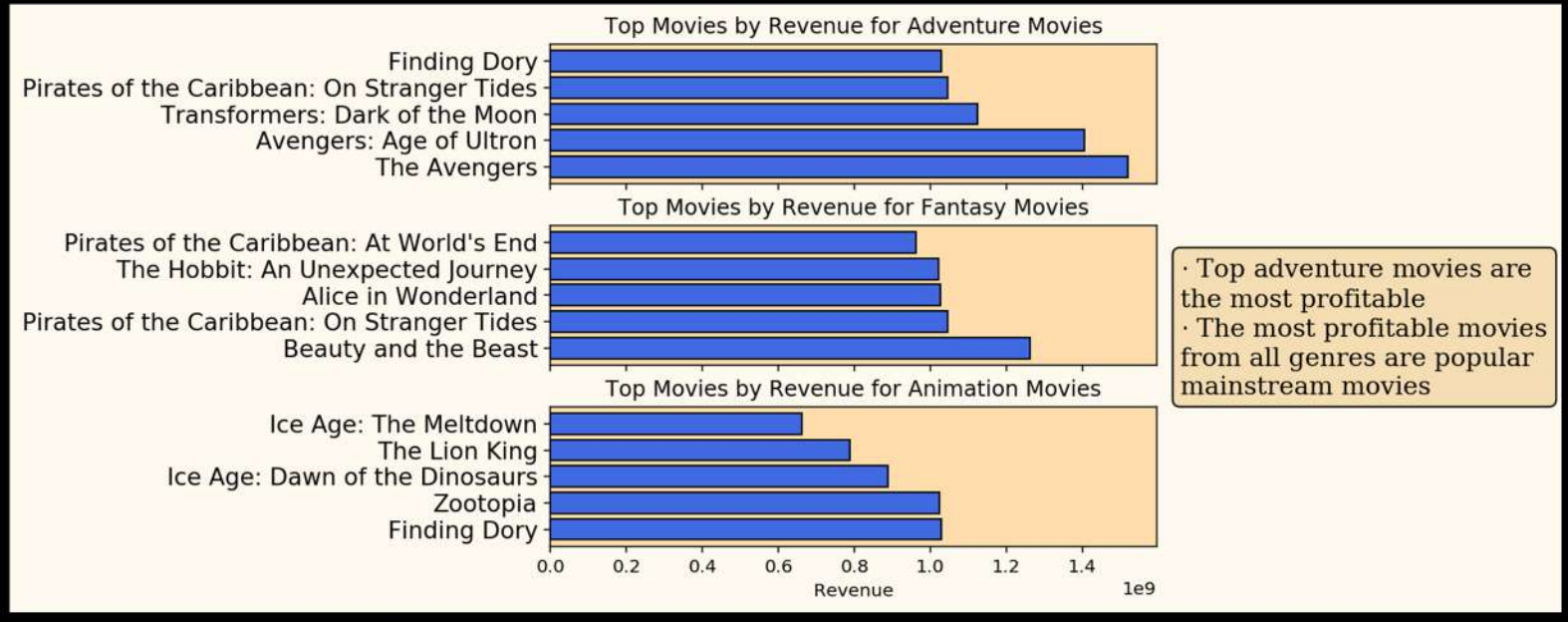
· Top production companies are the well-known American giants

· Based on, for a given production company, collect all movies associated with that production company, and sum the revenue of those movies



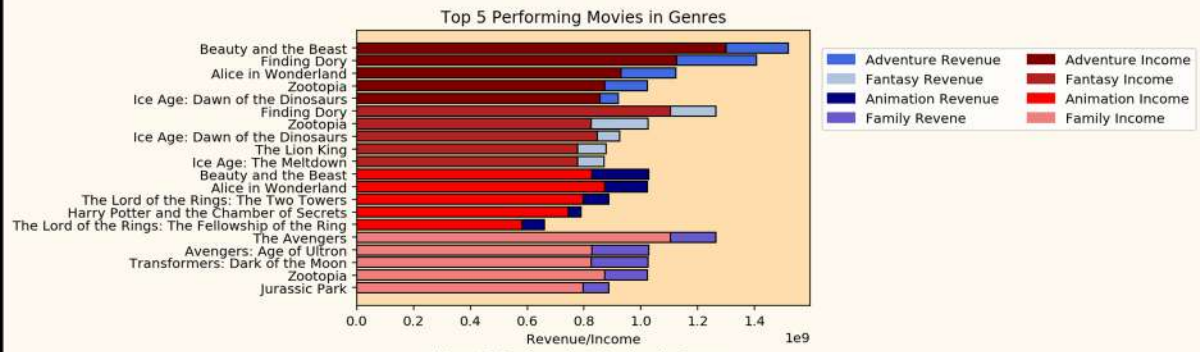
·Many of the words indicate that high-revenue associated keywords characterizes movies well.
· For example, keywords associated with superhero movies are frequent in top keywords
· Also notice words like Dinosaur, Secret Identity, Mission, Mutant, etc. which all clearly describe an important aspect of their associated movies

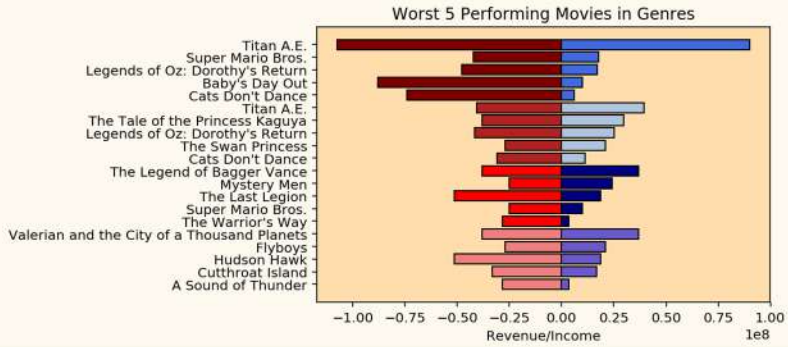
· For each keyword, collect all movies with that keyword. Then, take the average of the revenue associated with these movies



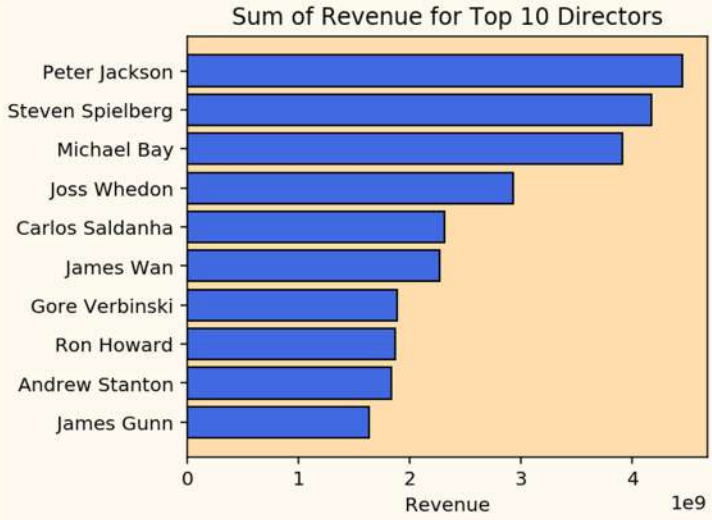
· Top adventure movies are the most profitable
· The most profitable movies from all genres are popular mainstream movies

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

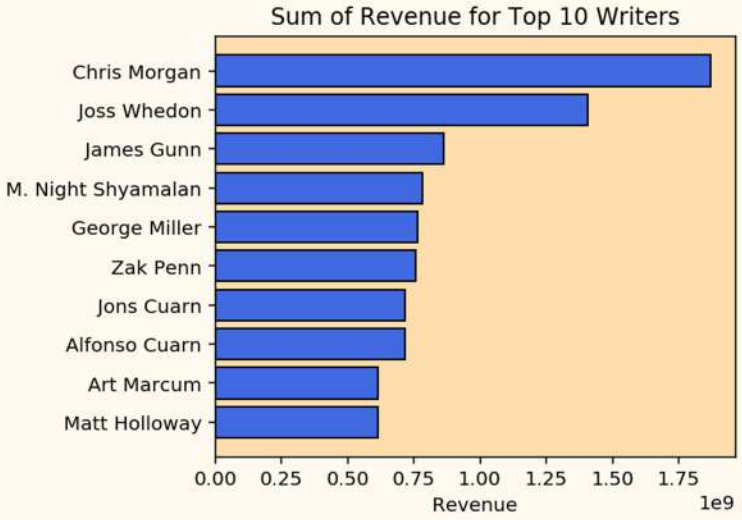




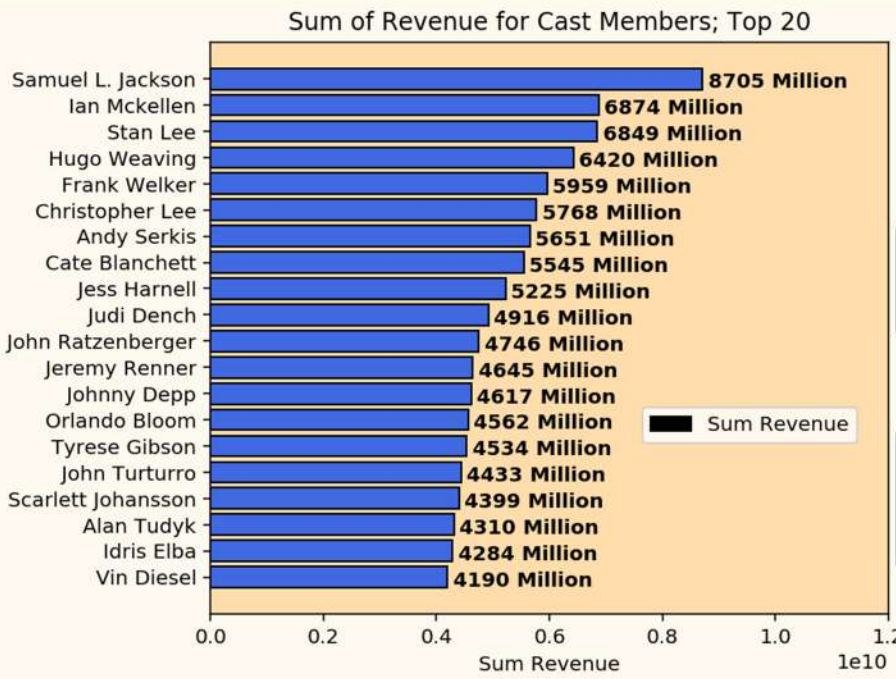
- Top performing movies are well-known blockbuster movies
- Movies from the same collection are frequent, like Lord of the Rings, and Ice Age movies
- A couple of movies appear in multiple genres, like Zootopia
- The only well-known movie among the worst performing is Titan A.E. - otherwise there are only unknown movies - as expected perhaps



- Only male directors present
- Many well-known directors present, like Peter Jackson and Michael Bay



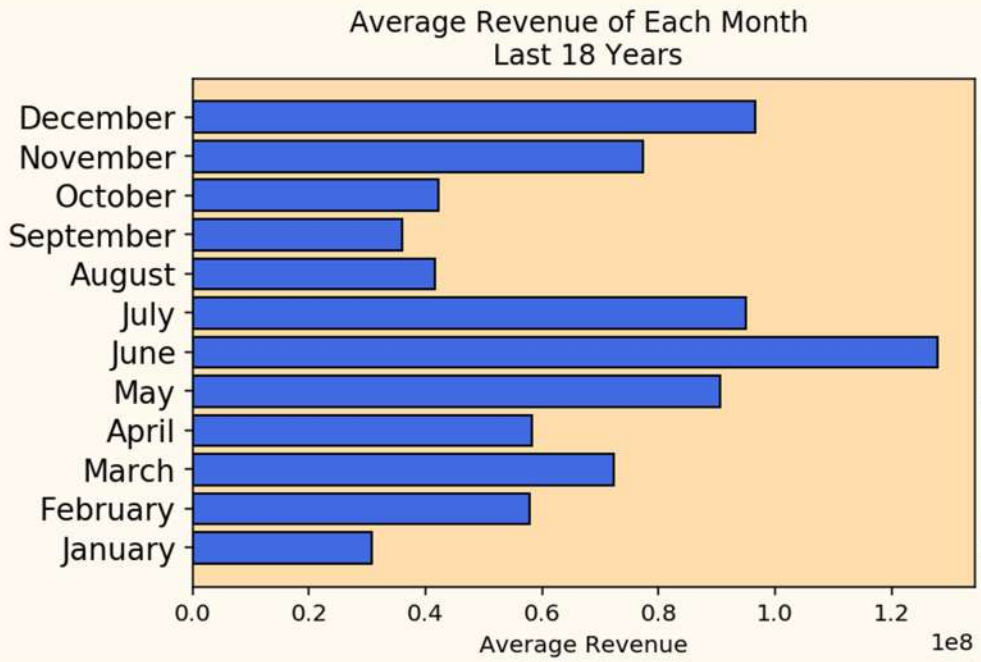
- As for directors, only male writers present
- Chris Morgan (The Fast and Furious) and Joss Whedon (Avengers) in the top



- Mostly male in the top
- A few females, like Cate Blanchett, Judi Dench

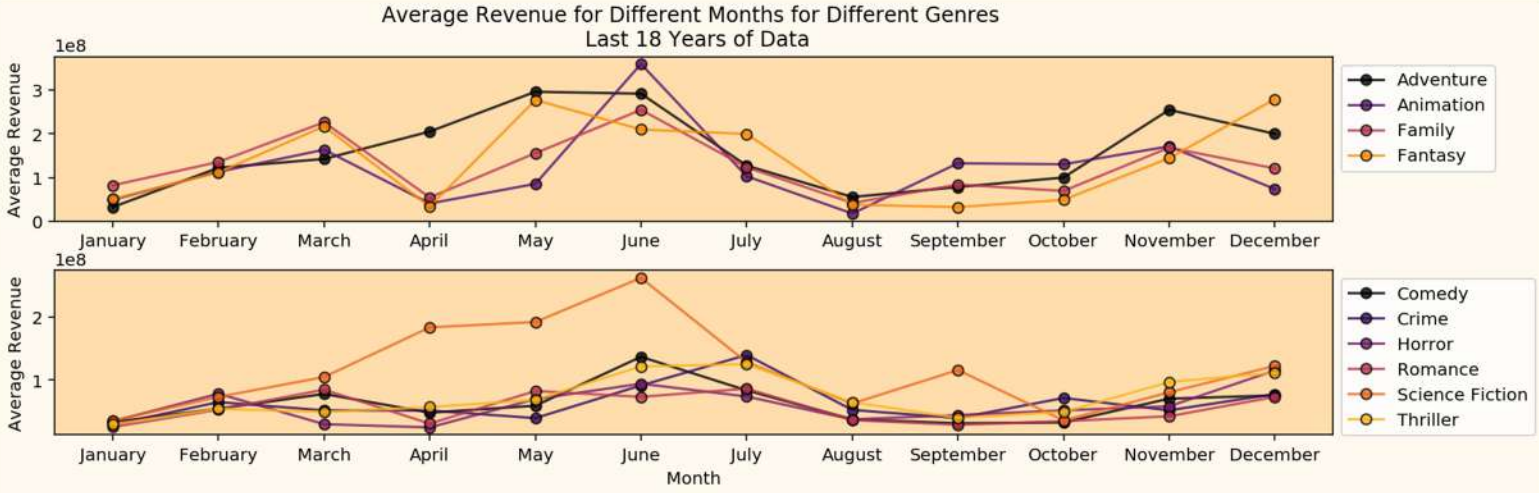
- Based on, for a given Cast Member, collect all movies associated with that cast member, and sum the revenue of those movies

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

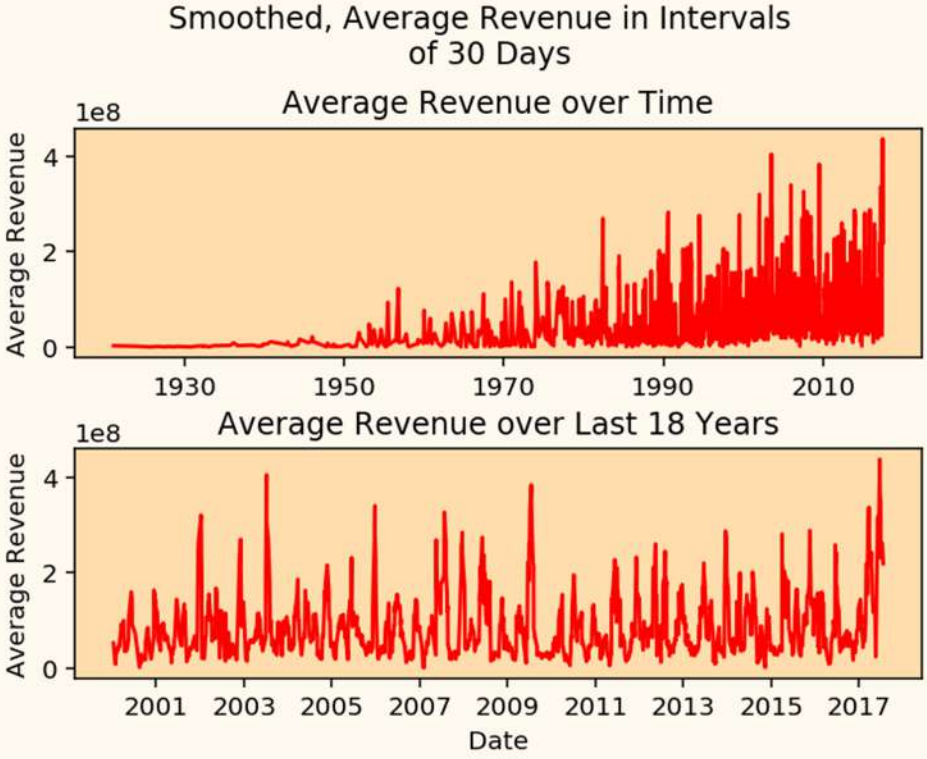


· Average revenue peaks during the summer and December
· Indicates that semester seasons are attractive for high revenue movies

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

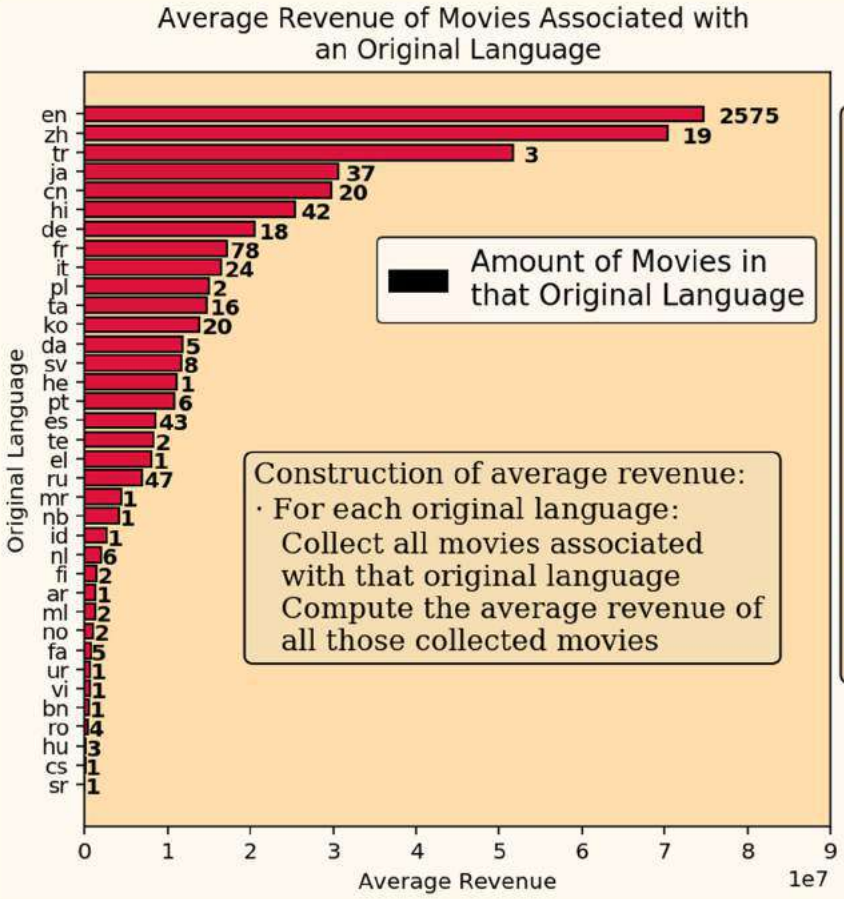


· Most genres have average revenue peaks during the semester seasons (summer & winter)
· The average revenue over months vary between different genres

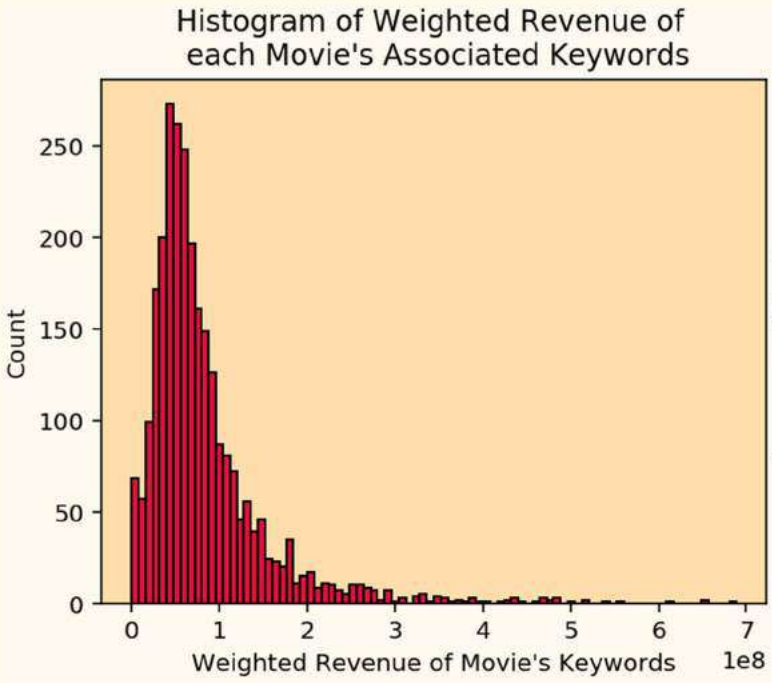


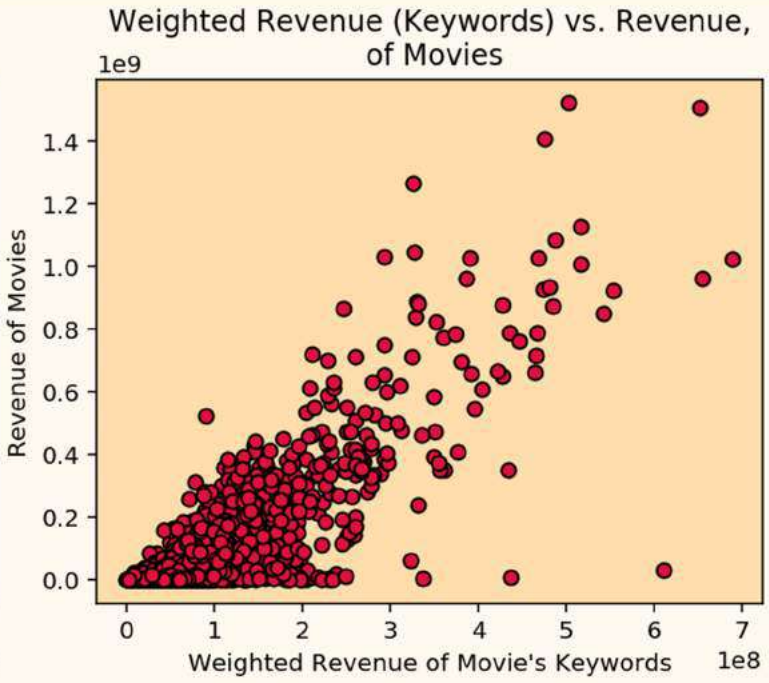
· Smoothed, average revenue increases over time - most likely due to inflation and/or better performing movies over time, and an increase in movies done over time
· For the last 18 years of average revenue, there is a clear seasonality and repetition in the average revenue over the years

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



· 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

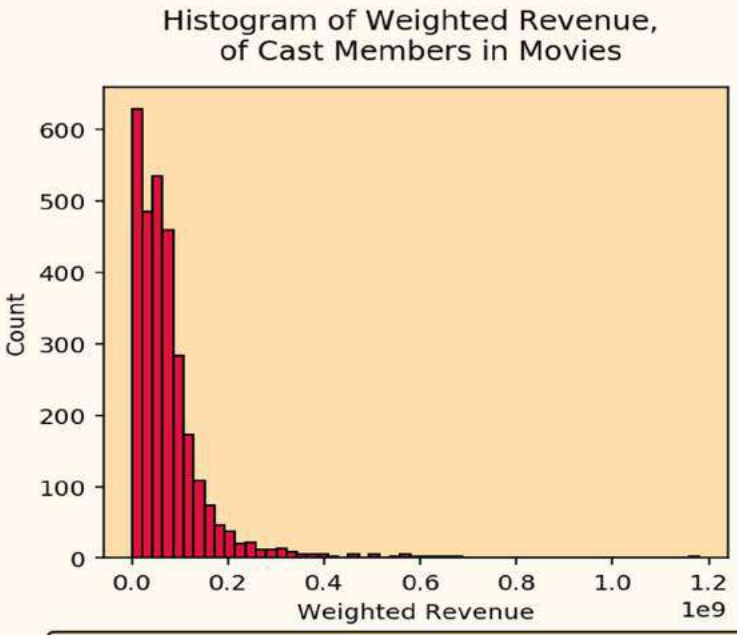


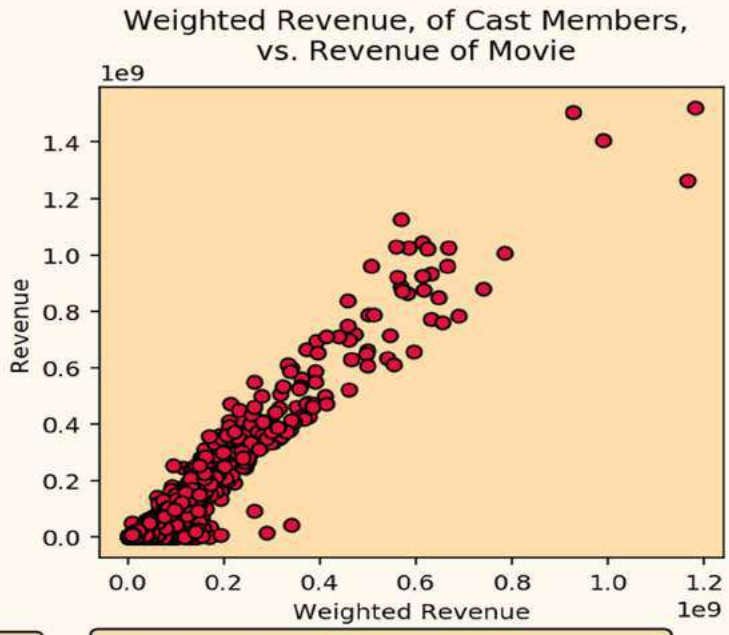


Construction of Weighted Revenue:
· (1) For each keyword:
Collect all movies associated with that keyword
Compute the average revenue of all those movies
· (2) For each movie:
Collect all keywords associated with that movie
Collect average revenue associated with each keyword, from (1)
Compute the weighted 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

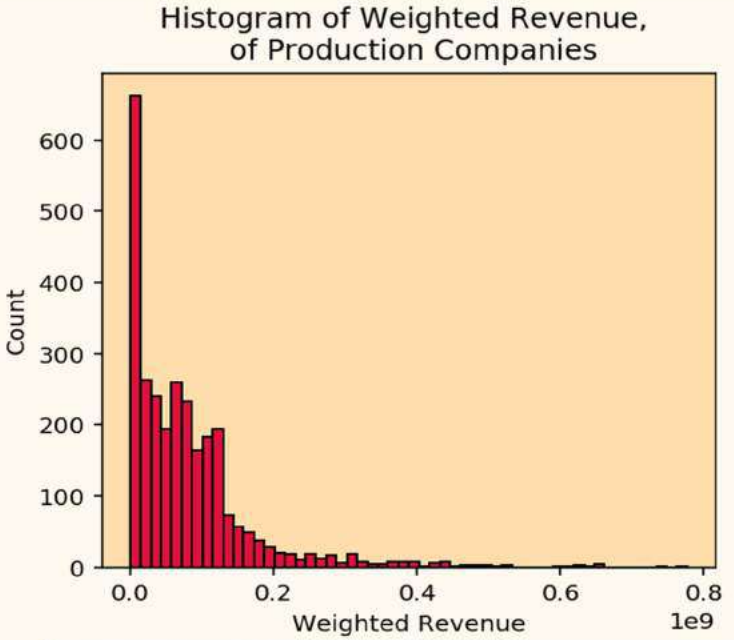
Construct Attributes & Data Analysis - TMDB Project - Cast and Production

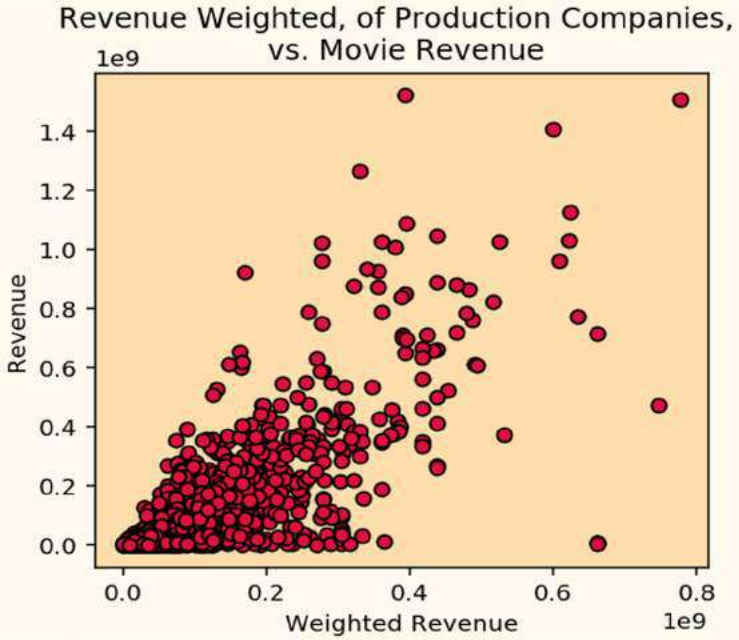




Construction of Weighted Revenue:
- (1) For each Cast Member:
  - Collect all movies associated with that cast member
  - Compute the average revenue of all those movies associated with that cast member
- (2) For each movie:
  - Collect all cast members(s) associated with that movie
  - Collect average 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

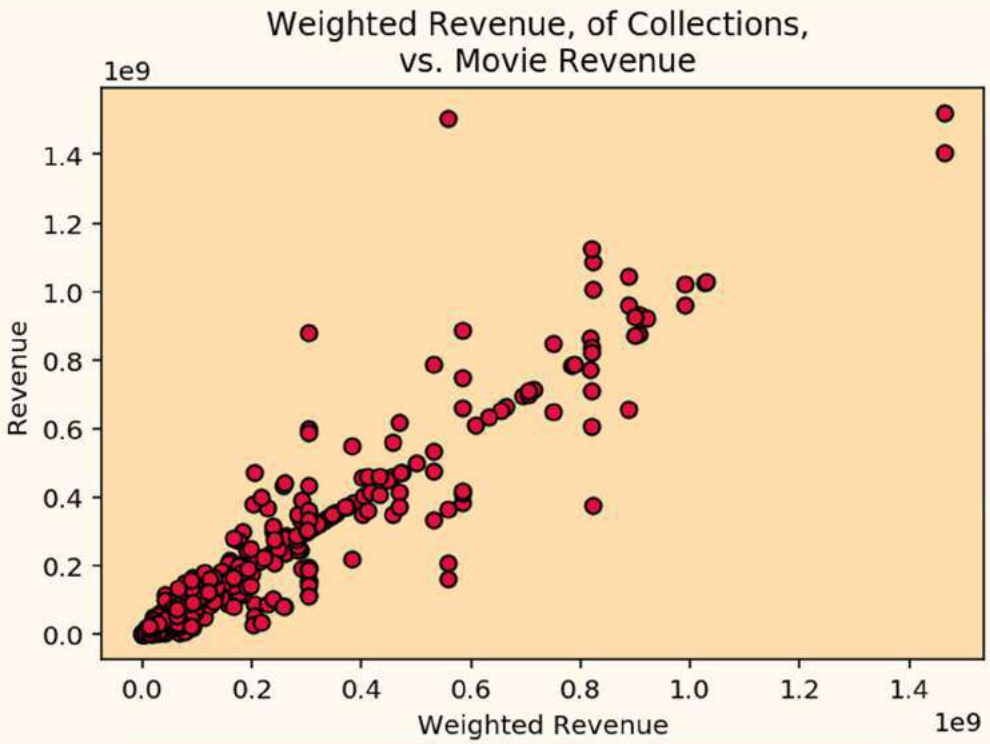




Construction of Weighted Revenue:
- (1) For each Production Company:
  - Collect all movies associated with that Production Company
  - Compute the average revenue of all those movies
- (2) For each movie:
  - Collect all Production Companies associated with that movie
  - Collect average revenue associated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

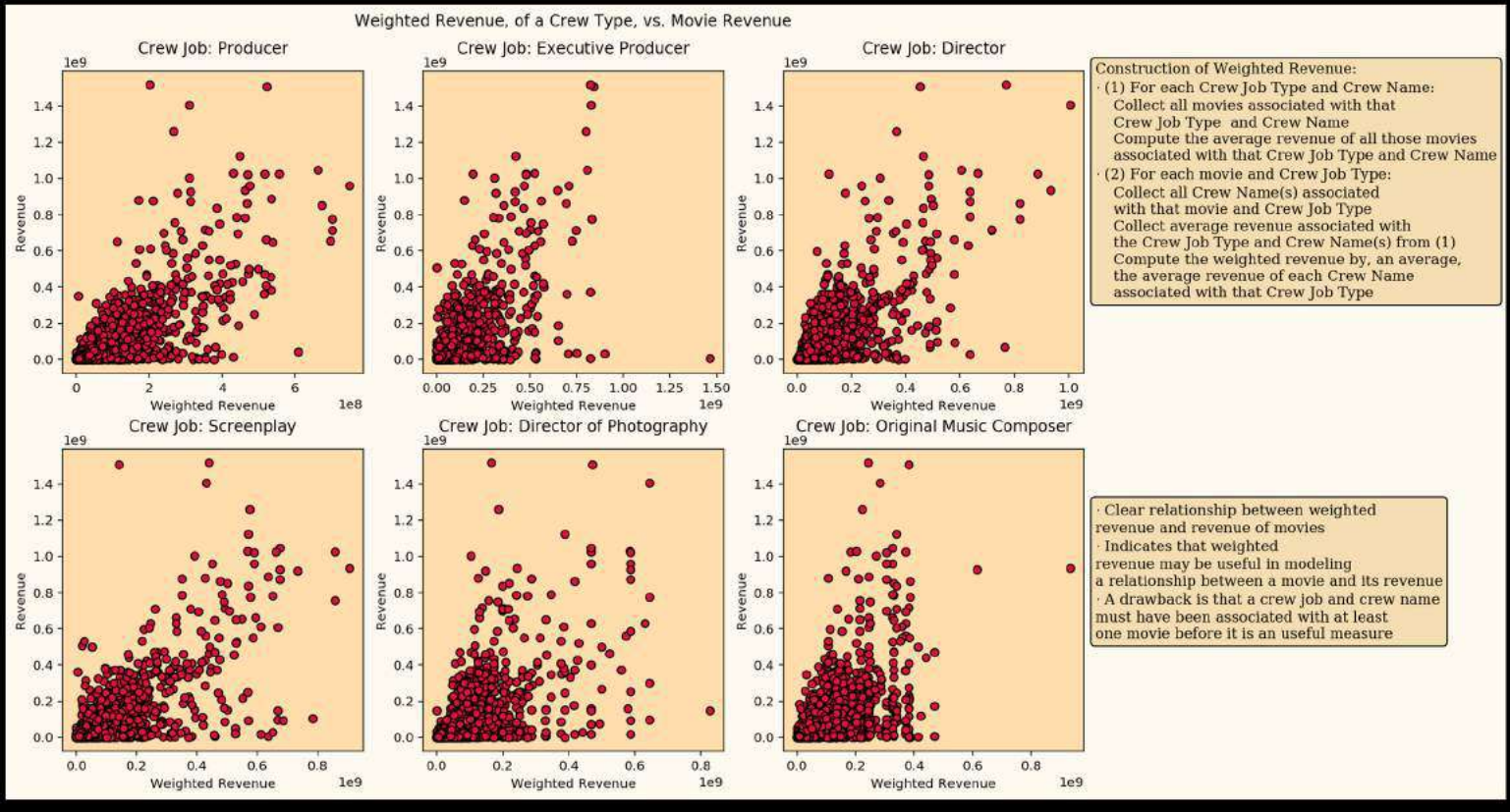
- Clear relationship between weighted revenue and revenue of movies
- Indicates that weighted revenue may be useful modeling a relationship between a movie and its revenue
- A drawback is that it resembles the relationship between weighted revenue for keywords and movie revenue

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



· 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

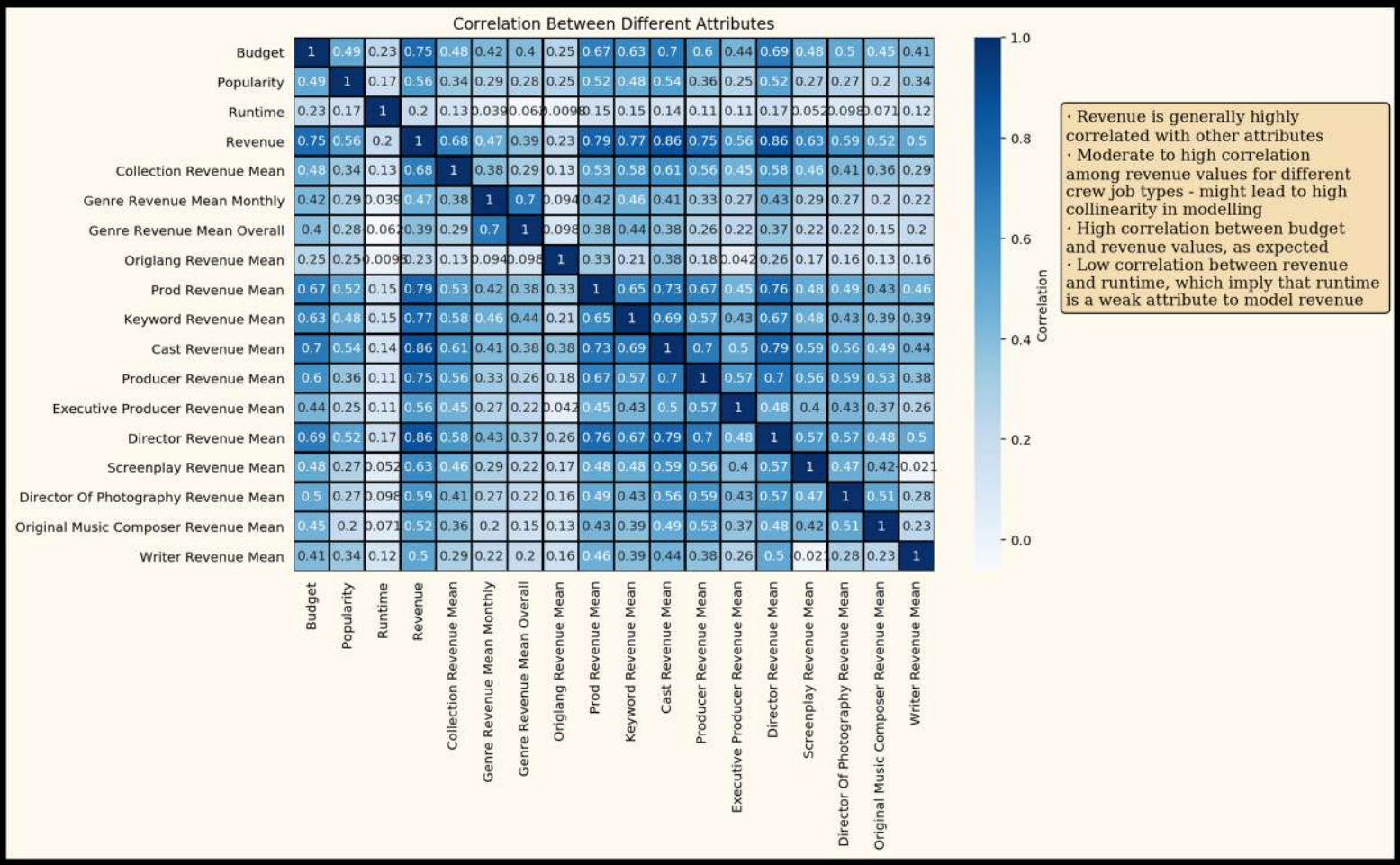
Construction of Weighted Revenue:
· (1) For each Collection:
  Collect all movies associated with that Collection
  Compute the average revenue of all those movies associated with that Collection
· (2) For each movie:
  Collect all Collection(s) associated with that movie
  Collect average revenue associated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



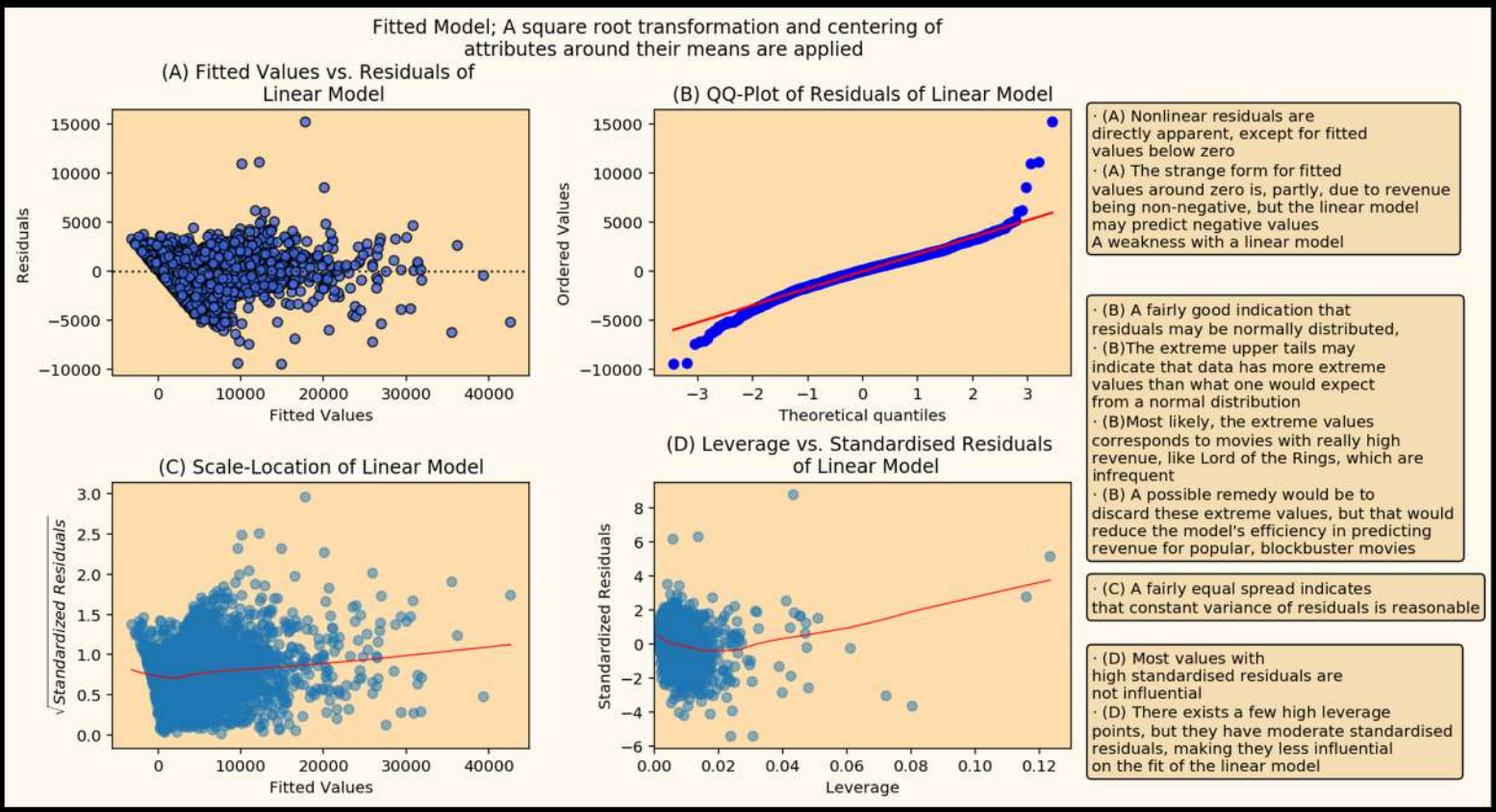
Construction of Weighted Revenue:
· (1) For each Crew Job Type and Crew Name:
  Collect all movies associated with that Crew Job Type and Crew Name
  Compute the average revenue of all those movies associated with that Crew Job Type and Crew Name
· (2) For each movie and Crew Job Type:
  Collect all 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 adequcy of the linear model



· Revenue is generally highly correlated with other attributes
· Moderate to high correlation among revenue values for different crew job types - might lead to high collinearity in modelling
· High correlation between budget and revenue values, as expected
· Low correlation between revenue and runtime, which imply that runtime is a weak attribute to model revenue



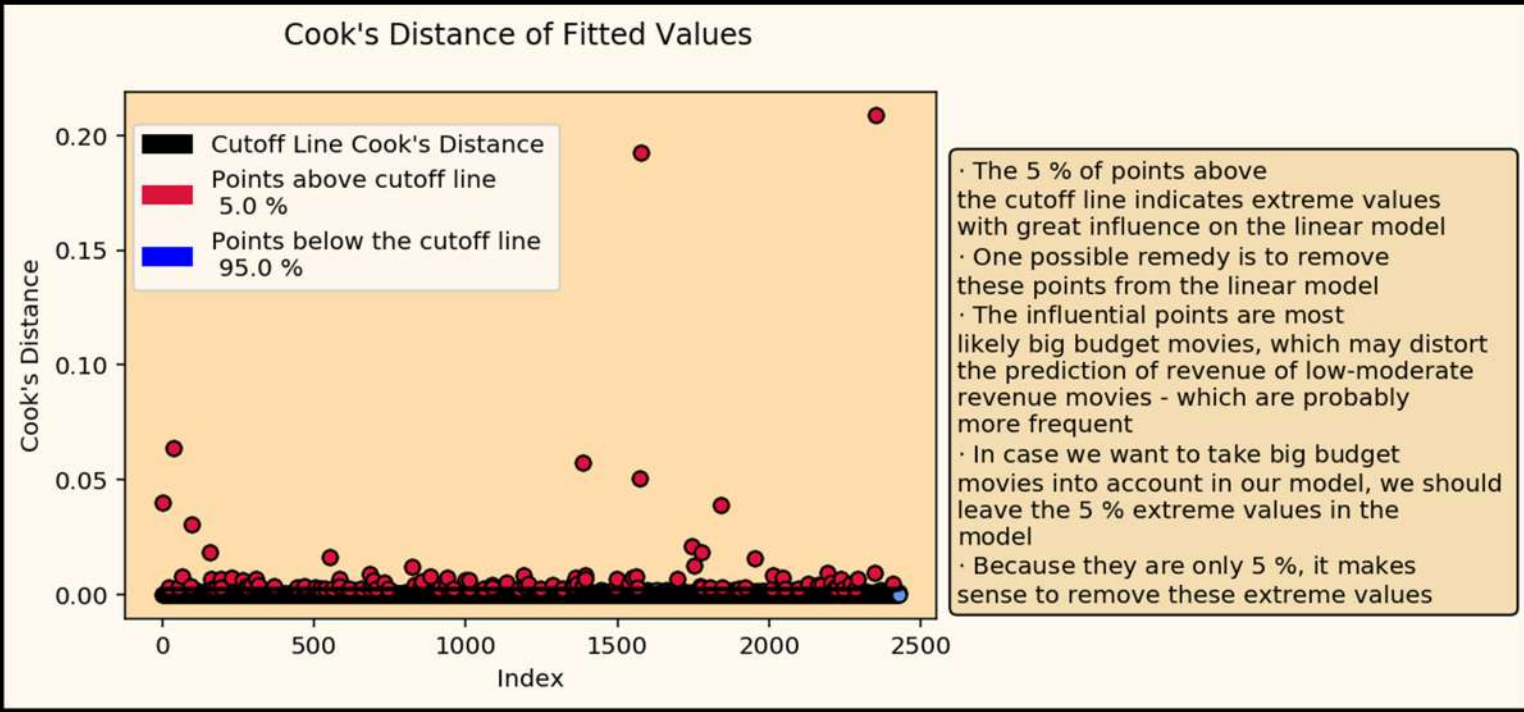
· (A) Nonlinear residuals are directly apparent, except for fitted values below zero
· (A) The strange form for fitted values around zero is, partly, due to revenue being non-negative, but the linear model may predict negative values
A weakness with a linear model

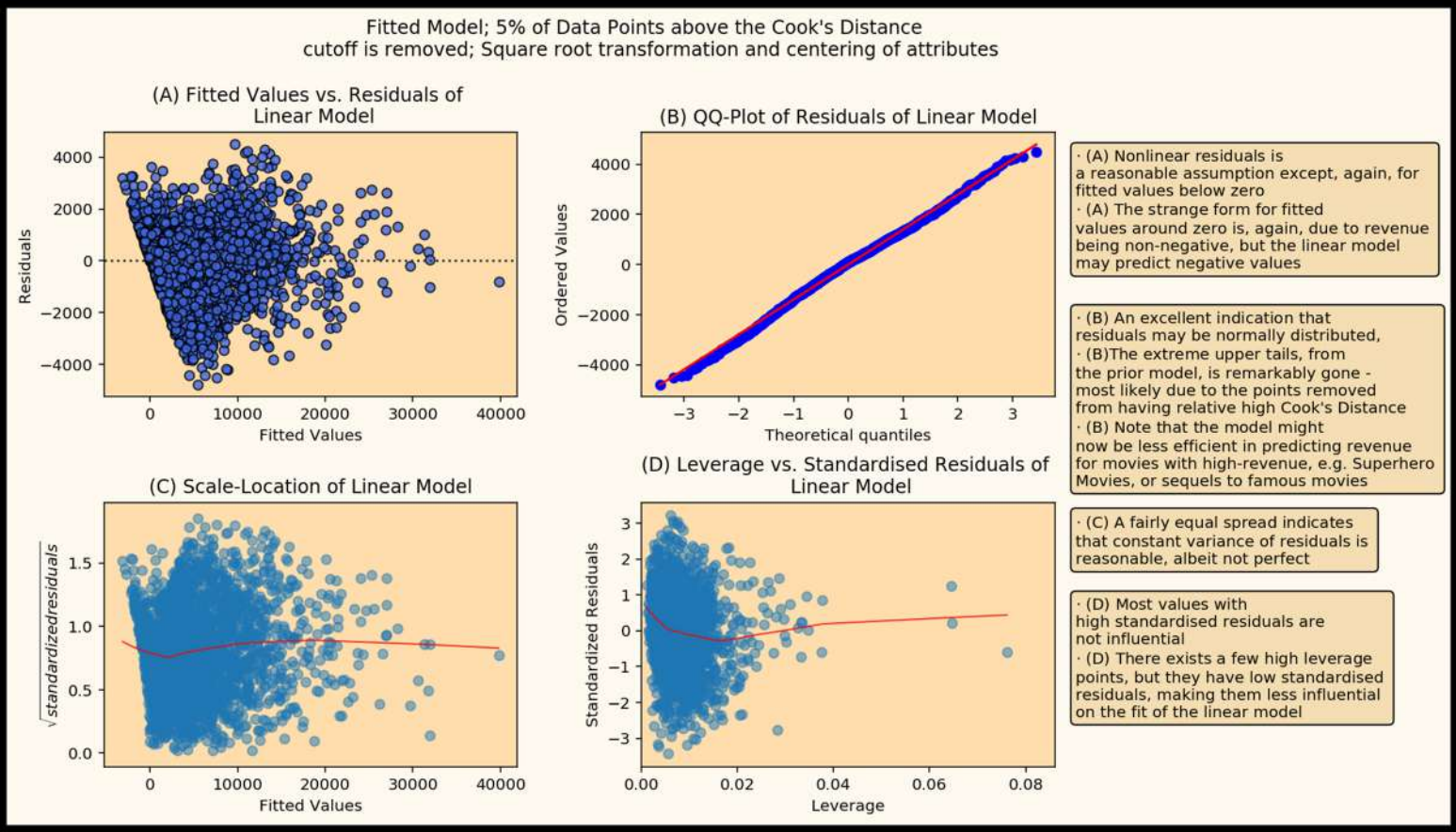
· (B) A fairly good indication that residuals may be normally distributed,
· (B)The extreme upper tails may indicate that data has more extreme values than what one would expect from a normal distribution
· (B)Most likely, the extreme values corresponds to movies with really high revenue, like Lord of the Rings, which are infrequent
· (B) A possible remedy would be to discard these extreme values, but that would reduce the model's efficiency in predicting revenue for popular, blockbuster movies

· (C) A fairly equal spread indicates that constant variance of residuals is reasonable

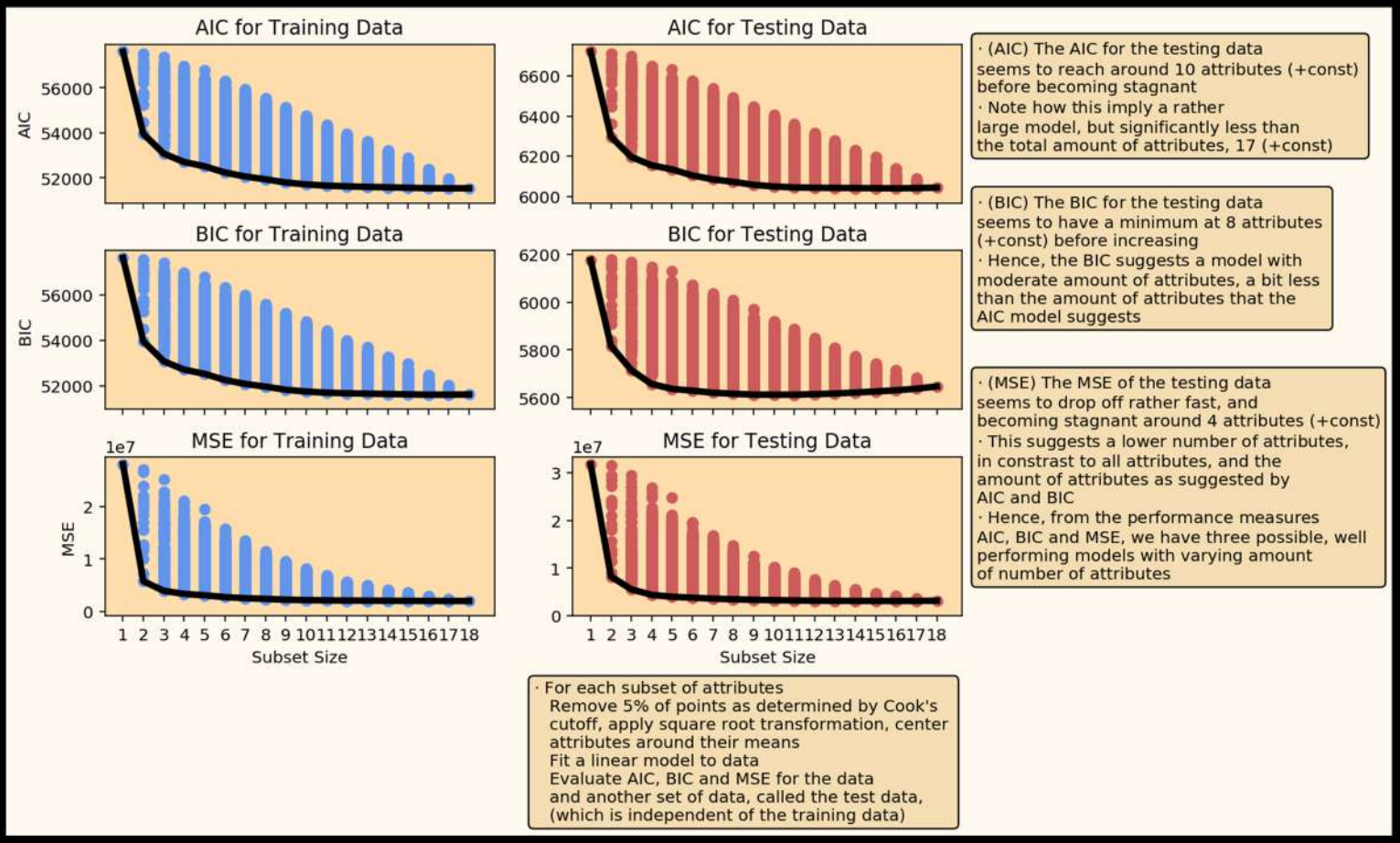
· (D) Most values with high standardised residuals are not influential
· (D) There exists a few high leverage points, but they have moderate standardised residuals, making they less influential on the fit of the linear model

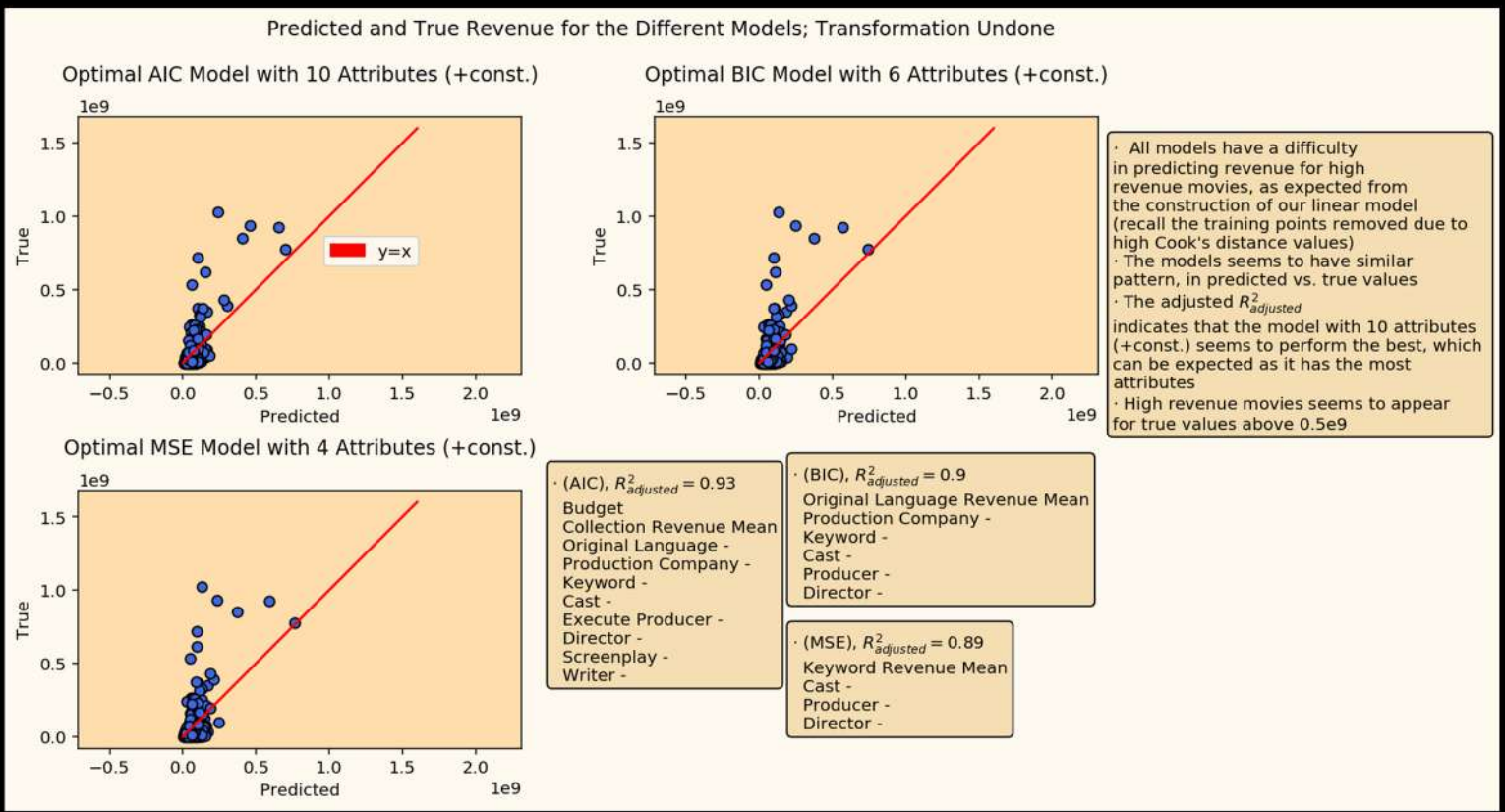
Building a Linear Model & Prediction Analysis - TMDB Project - Cook's distance analyze of fitted values, and refit of linear model with extreme influential data points removed from dataset





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





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