UDACITY DATA VISUALIZATION NANODEGREE

Project: Build a Data Storytelling Project – Final, using the movies dataset

By Krishna Nadoor

13/02/2022

VIEW THE TABLEAU STORY

Link:
https://public.tableau.com/app/profile/krishna.nadoor/viz/UdacityDataStorytelling-Final/DataStorytellingFinal

PROBLEM STATEMENT

*What are the **key features** that the **top 3%** of movies by **revenue** all **have in common**?*

SYNTHESIS

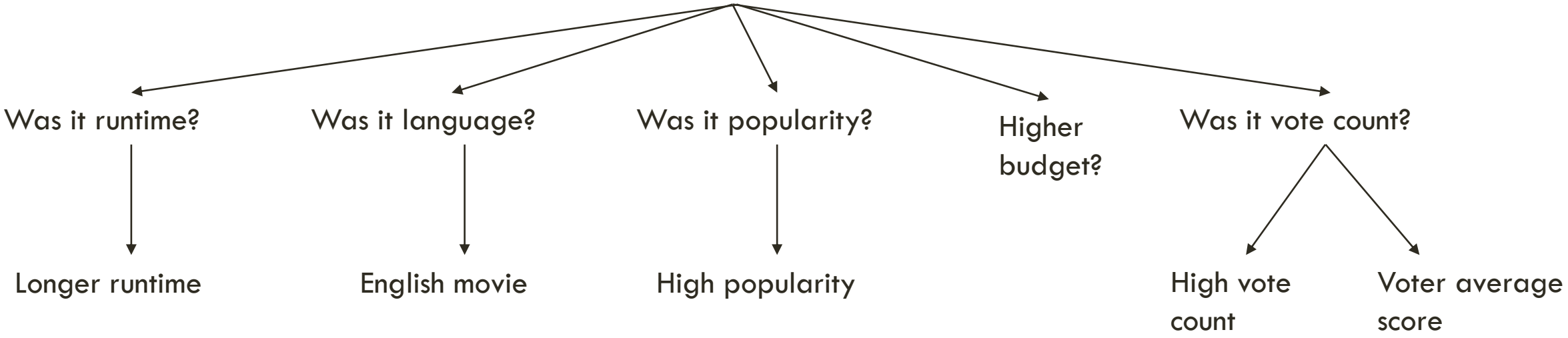
- The length of a movie is a key feature that determines the popularity and subsequent higher revenues.
- English movies have higher revenues than non-English movies, possibly due to higher budgets.
- A high vote count coupled with a high average viewer score is an indication of possible higher revenues.
- Higher film budget results in higher revenue.

OVERVIEW OF ANALYSIS

- Comparison of the playback length of movies with the movie revenues.
- Comparing the revenue of English vs Non-English movies to determine correlation between language being a factor in revenue outcome.
- A high average viewer score is not an indicator of higher revenue unless it is coupled with a high vote count.
- Comparison between budget of a film and the revenue earned.

ISSUE TREE

*What are the **key features** that the **top 3%** of movies by **revenue** all **have in common?***

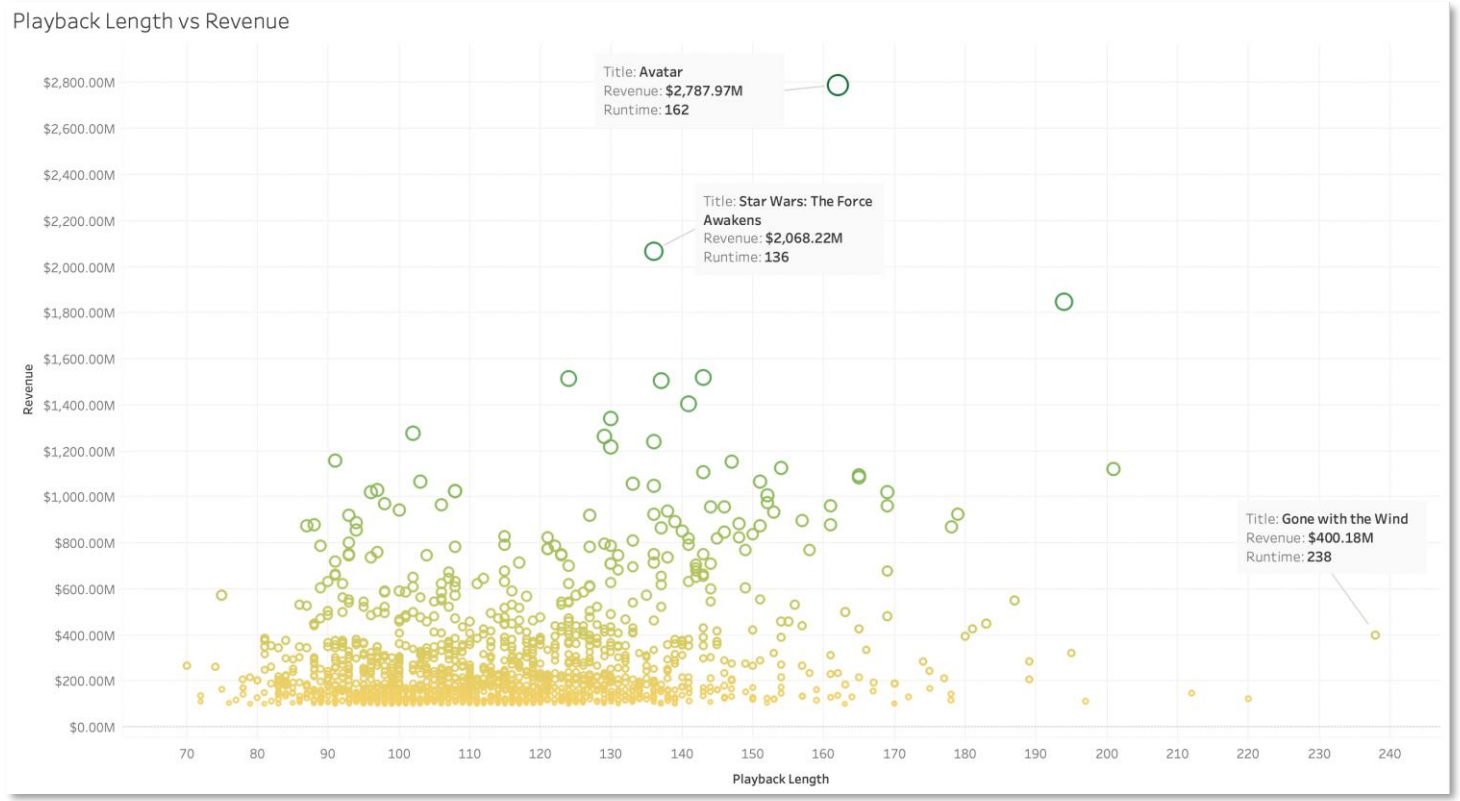


PLAYBACK LENGTH HAS IMPACT ON MOVIE REVENUE

We can see examples here of movies longer than 120 minutes having larger revenues. Such as **Avatar** and **Star Wars The Force Awakens**.

However, it can be noticed that extended runtime does not always translate to higher revenues. The movie, **Gone with the Wind,** only had revenues of $400.18M.

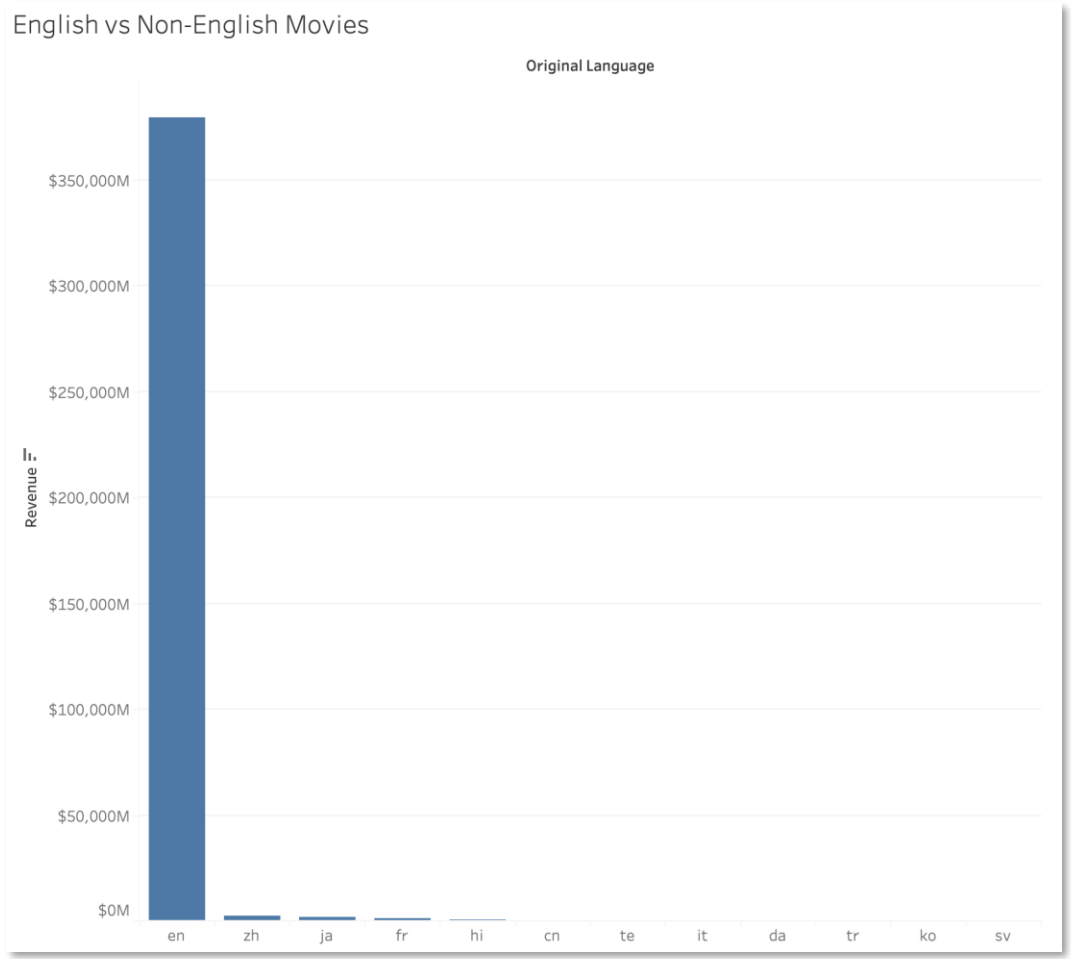
This simply means that we need to look at other features as well to learn more.



ENGLISH MOVIES HAVE HIGHER REVENUE THAN NON-ENGLISH MOVIES

We can see that an overwhelming amount of movies that contributed to total revenues has been English movies.

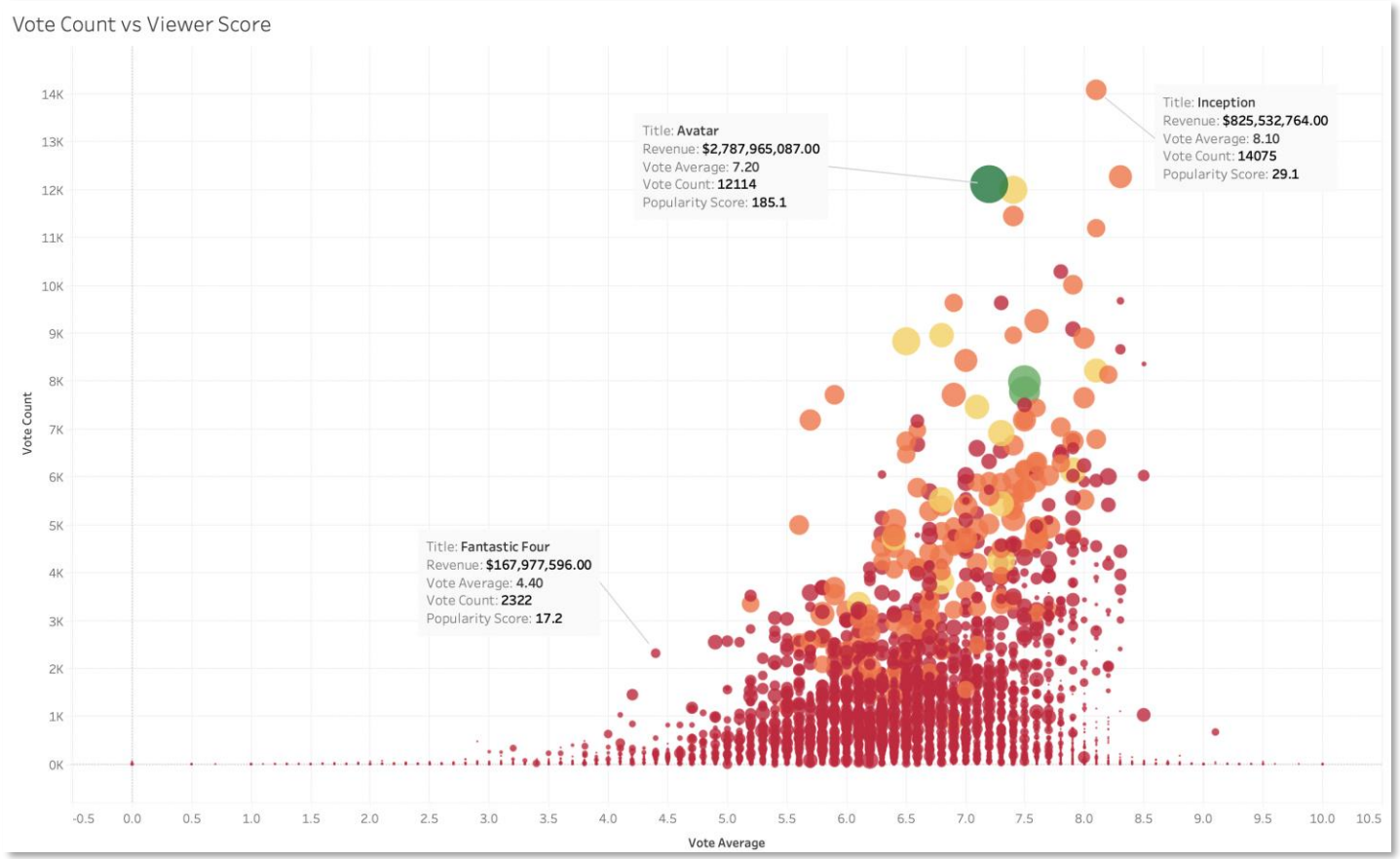
This indicates that there is a far greater amount of English movies with higher revenues compared with non-English films.



VOTE COUNT AND VIEWER SCORE DETERMINE REVENUE

Looking at the data we see that movies with higher revenues (indicated by bubble size and colour) typically have very high vote counts (indicating social proof) coupled with a high voter average rating.

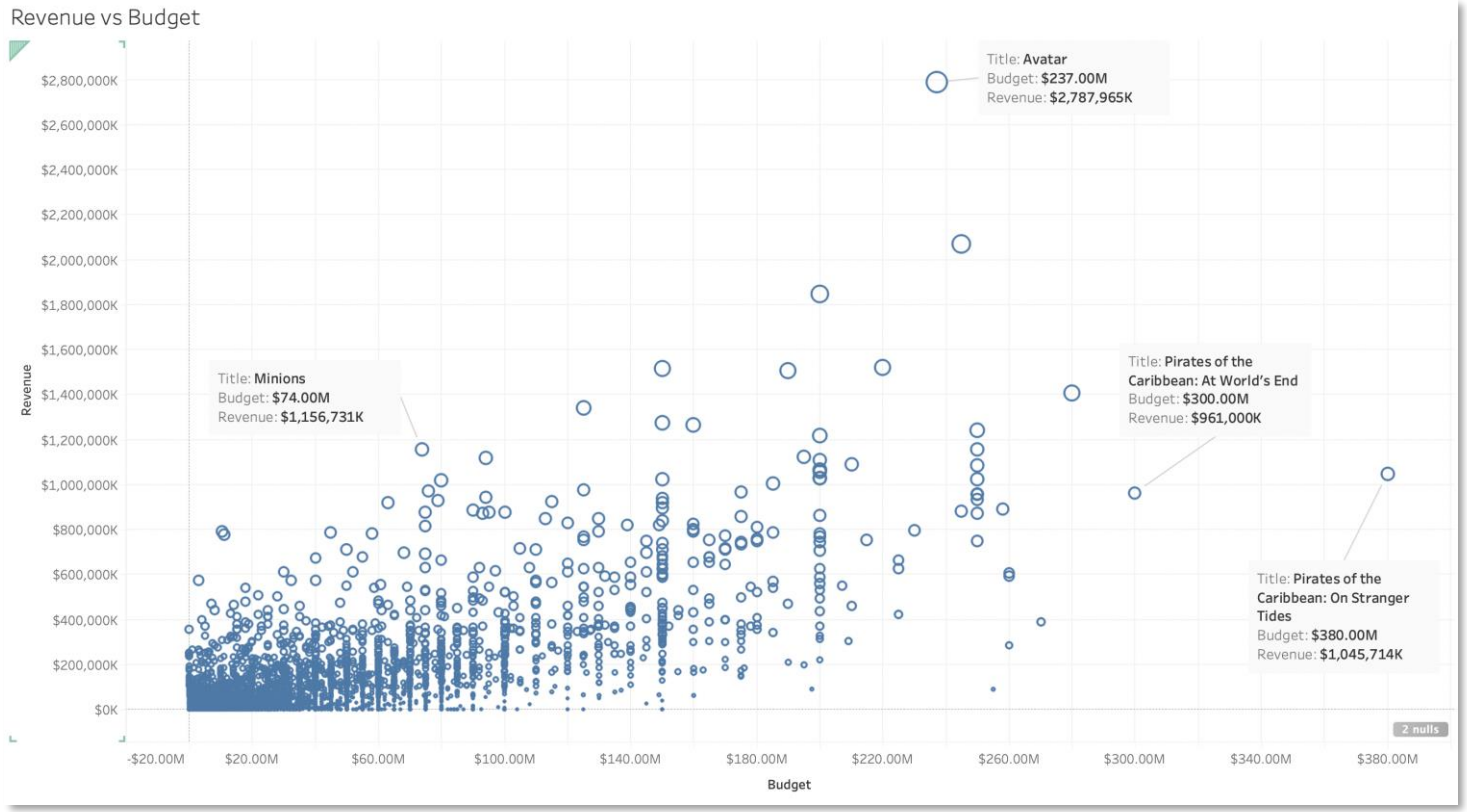
In addition to this, it was found that a movie's popularity score is also a key indicator of high revenues, **Avatar** has a popularity score of 185.1 – in contrast, **Inception**, with higher voter average and vote counts had a much lower revenue ($2.8B vs $825M) due to a comparatively poor popularity score of only 29.1



LARGER MOVIE BUDGET RESULTS IN HIGHER REVENUES

This hypothesis is valid for quite a few movies, such as **Avatar**, with a budget of around $237M, revenue was almost $2.8 billion, therefore a factor of 12.

The data also shows that movies don't need to have large budgets to get high revenues. For example, **Minions** had a budget of $74M, however revenues reached just over $1.1B. A factor of around 15.6. Financially, it performed better than Avatar at a budget to revenue ratio standpoint.



LIMITATIONS & BIASES

Limitations:

- Dirty data (i.e., genres, production countries fields, budget etc.)

Biases:

Data Collection:

- *Missing variable bias*, we haven't included all the variables that can influence the analysis such as the consideration of genres and country of origin.

Data Processing:

- *Outliers* (i.e., movies with high runtime but no revenue, voter average scores greater than 10 etc.)
- *Missingness* of values (i.e., no budget recorded for some movies, missing user scores etc.)

Data Insights:

- *Confirmation bias* from low vote counts (i.e., 2 votes) coupled with a high average viewer score (i.e., 8) would give the false impression that a movie is good to watch even though overall revenue may be quite low.

NEXT STEPS

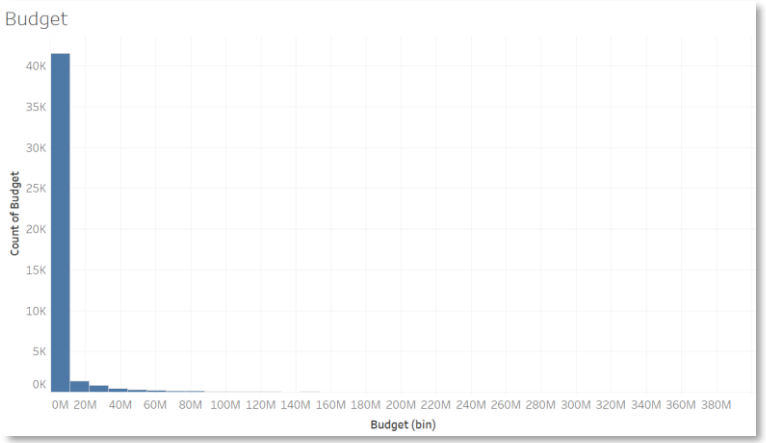
- Remove rows that have missing data (i.e., revenue, budget)
- Perform data normalization:
  - Split out the genres
  - Split out the country from the production countries field
- Remove outliers
- Repeat analysis to consider other features that indicate revenue performance such as the movie genre and the country the movie was filmed in.

APPENDIX

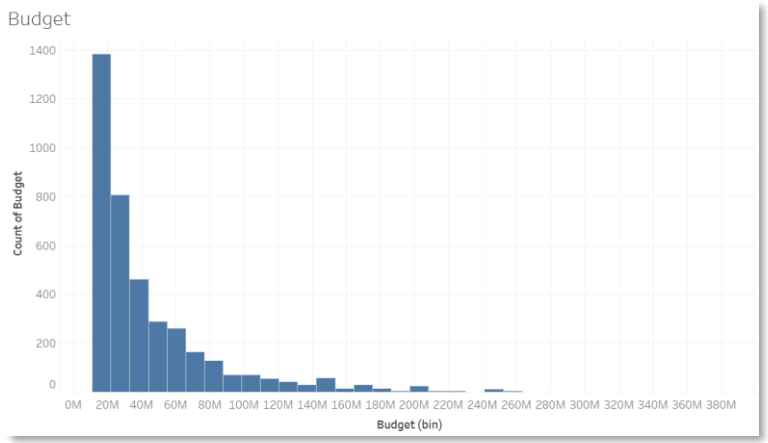
ANALYSIS OF NUMERICAL VARIABLE DISTRIBUTIONS

DATA ANALYSIS OF BUDGET

Summary Stats	Value after removal of outliers
Min	$11,000,000
Q1	$18,000,000
Q2 (median)	$30,000,000
Mean	$44,487,957.92
Q3	$55,000,000
max	$380,000



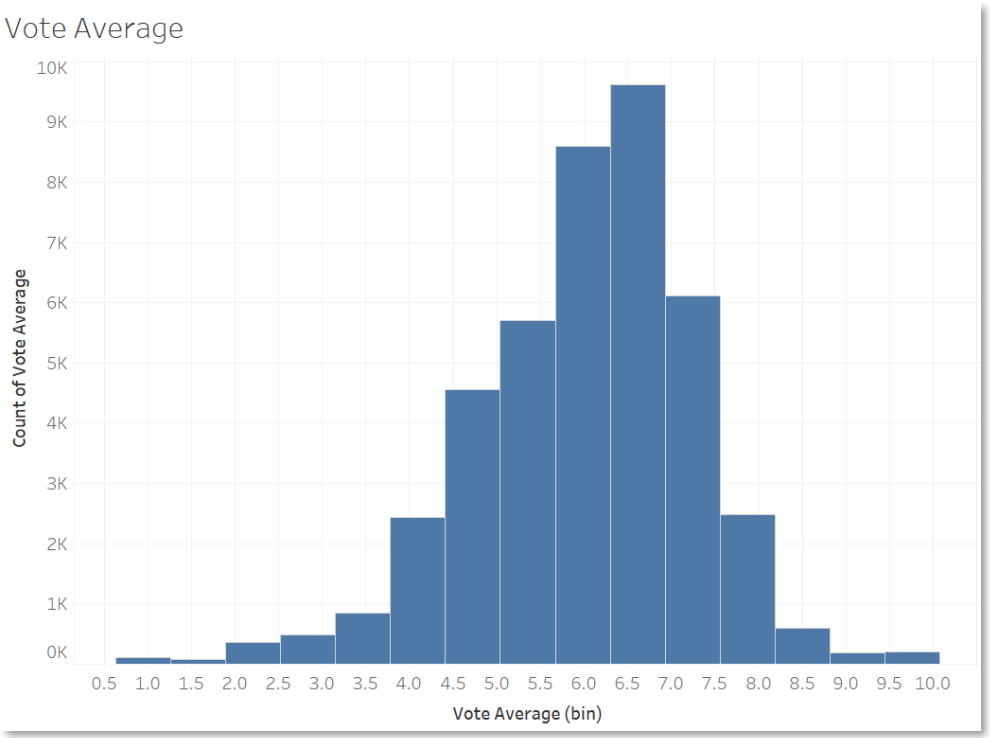
With outliers of budget being present.



With outliers removed, so histogram only shows data that has budget amounts.

DATA ANALYSIS OF VOTE AVERAGE

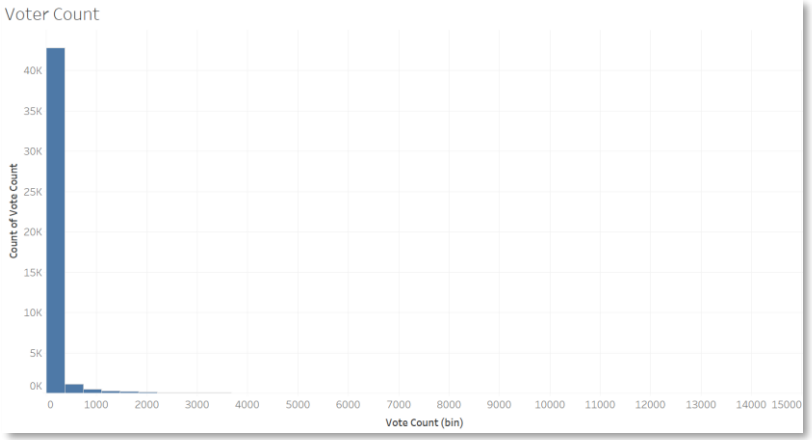
Summary Stats	Value after removal of outliers
Min	0.70
Q1	5.30
Q2 (median)	6.10
Mean	6.017
Q3	6.90
max	10



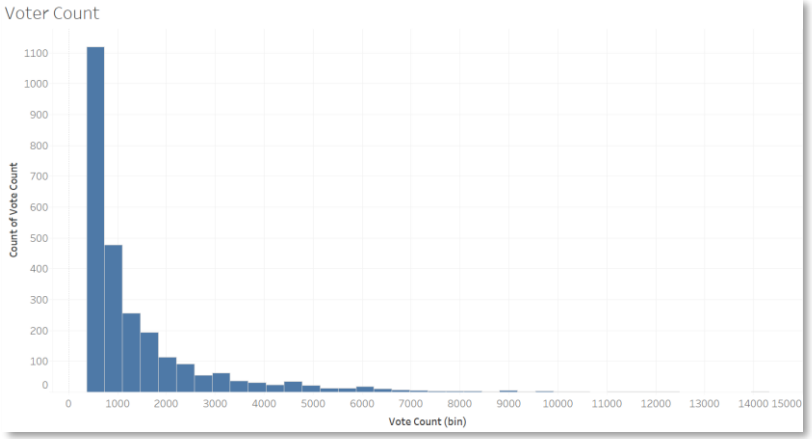
Here we see a normal distribution indicating around 68.2% of data is centred around voter scores being between 6 to 6.5, outliers exist on the far left however not in great quantity that it effects the data. Ratings greater than 10 have been removed.

DATA ANALYSIS OF VOTER COUNT

Summary Stats	Value after removal of outliers
Min	367
Q1	531
Q2 (median)	856
Mean	1,428.28
Q3	1,641
max	14,075



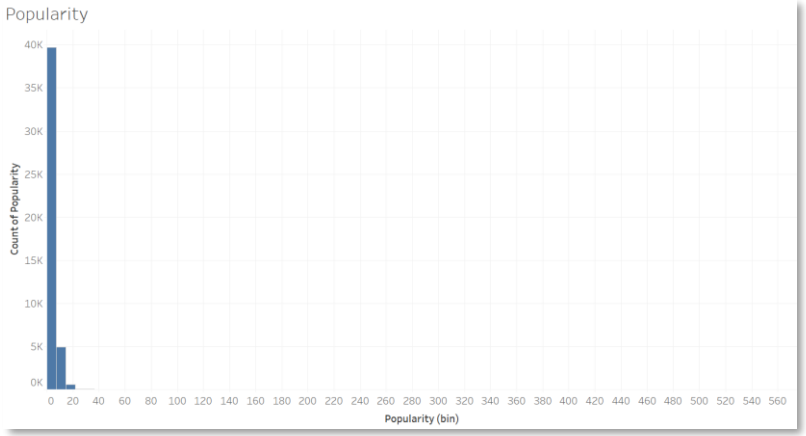
With outliers of voter count being present.



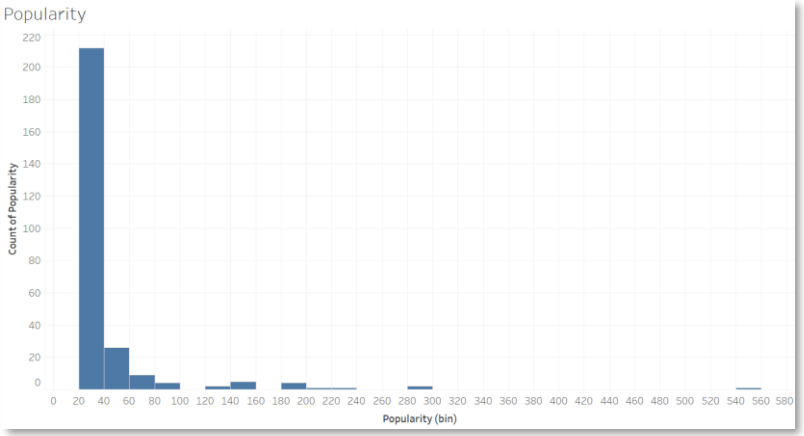
With outliers removed, so histogram only shows data that has voter count data. Distribution is right skewed, indicating majority of movies have a very low vote count compared to others that have counts in the thousands.

DATA ANALYSIS OF POPULARITY

Summary Stats	Value after removal of outliers
Min	7.30
Q1	8.60
Q2 (median)	10.20
Mean	12.10
Q3	12.60
max	547.50



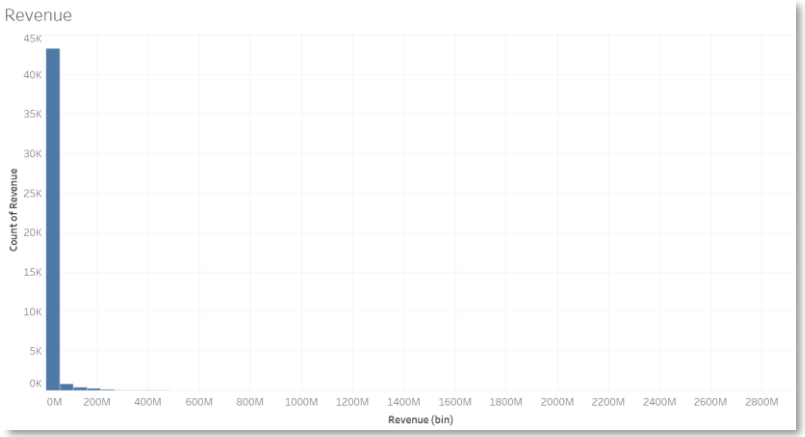
With outliers of popularity being present gives the false impression that nearly all movies have a popularity of 0.



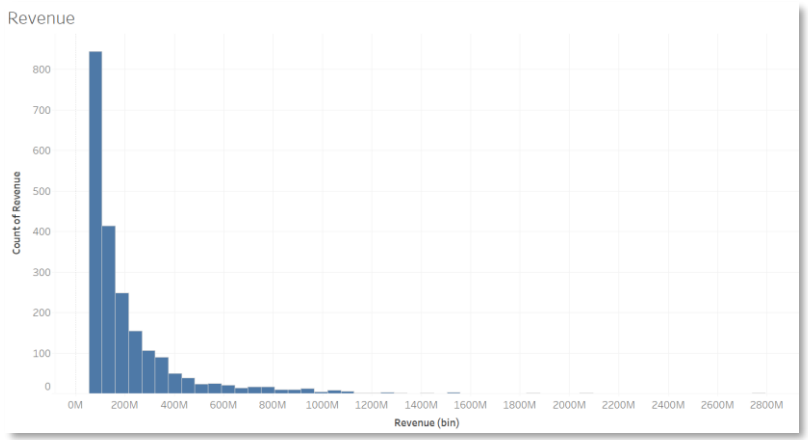
With outliers removed and increasing the bin size to 20, the right skewed histogram now shows data for which popularity exists. The graph indicates that majority of the movies have a popularity between 0 to 40. Only one movie has a high popularity score of 547.5.

DATA ANALYSIS OF REVENUE

Summary Stats	Value after removal of outliers
Min	$53,825,515
Q1	$83,538,436.50
Q2 (median)	$130,457,118
Mean	$208,846,339.42
Q3	$242,590,614.25
max	$2,787,965,087



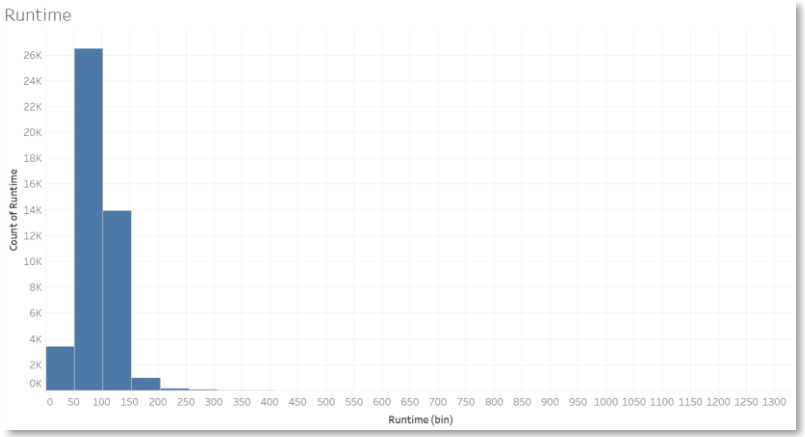
With outliers of revenue being present gives the false impression that nearly all movies have a revenue of $0.



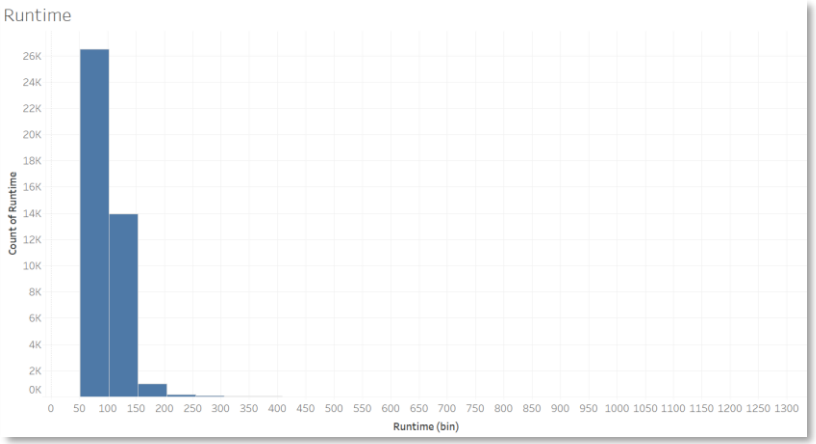
With outliers removed we see this right skewed distribution has majority of the movies with revenues ranging up to only $54M, which is not bad considering minimum budget as per prior histogram indicated it to be up to $11M. We can see one movie on the far right has a budget of $2.74B. This indicates that there are very few AAA movies being made compared to B grade movies.

DATA ANALYSIS OF RUNTIME

Summary Stats	Value after removal of outliers (minutes)
Min	51
Q1	88
Q2 (median)	96
Mean	100.97
Q3	108
max	1,256



Even though outliers with runtime of 0 minutes are present, the frequency of the outliers is not enough to skew the graph.



With outliers removed we see this right skewed distribution has majority of the movies having a runtime of up to 102 minutes, or just over 1.5 hours. There are shows here with a very large runtime, for example Centennial has a runtime of 1,256 minutes – but this is a TV mini series and not a one-off movie. Removal of these records won't change the shape of the distribution.