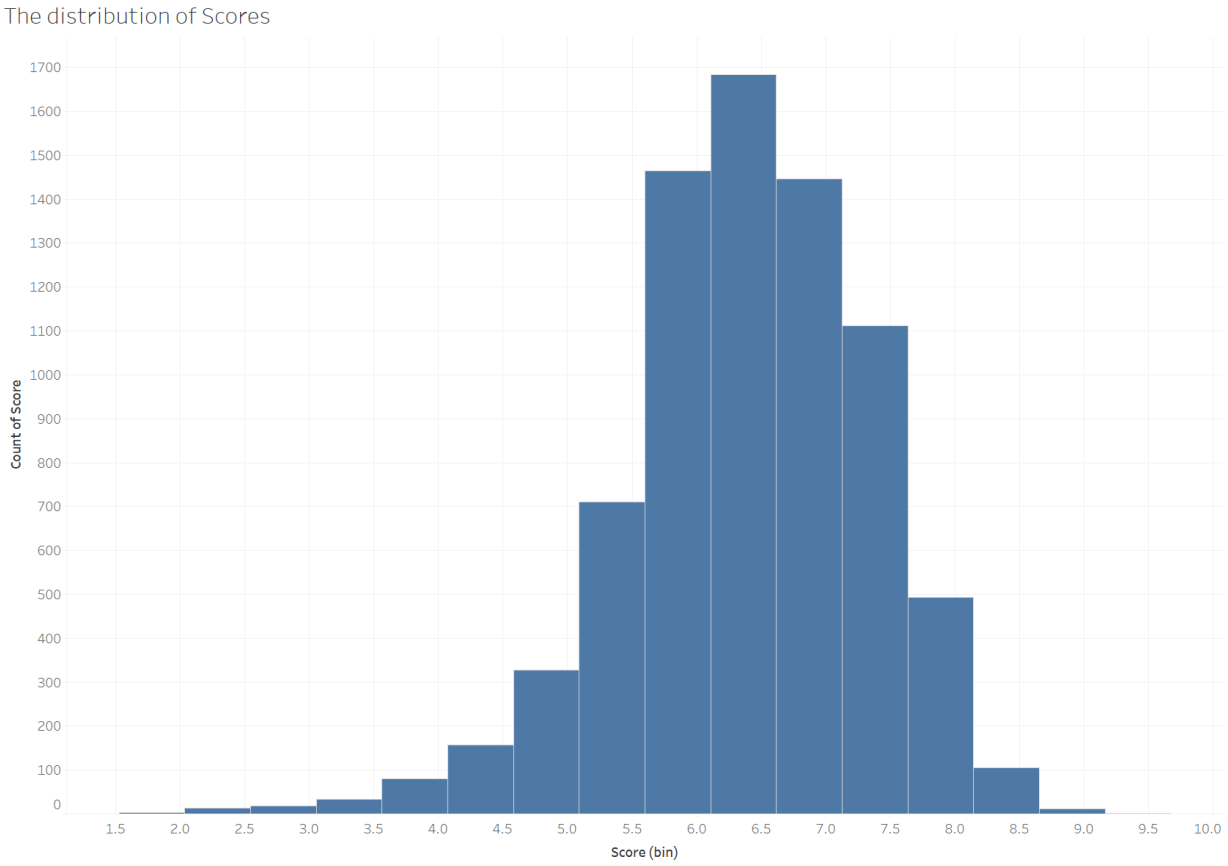
The process of scoring movies on the internet seems to be a mystery in both the movie industry and to the general public. A score is typically a 1 to 10 star review that users of the internet can leave on a movie scoring website after watching a movie. The scores displayed on movie information websites are typically averages of every user's score. It might seem that these subjective and biased ratings left by many audience members would not have much of an impact on a single movie.

Therefore, the question is: What components of a movie or a score are associated with the score of a movie?

Some components of a movie that this project will focus on include duration of movies, country of origin, different genres of movies, movie budget and gross, etc. The dataset was found at https://www.kaggle.com/danielgrijalvas/movies, and was precleaned movie dataset retrieved/scraped from the IMDb website each year with 200 entries. The data was not further filtered or subseted (since it was already in precleaned fashion) and entries were filtered for each visualization as specified (except for dropping nulls when necessary). This report is on the more exploratory side, looking at different aspects of the dataset while keeping one variable to compare to (in reference to the okcupid example).

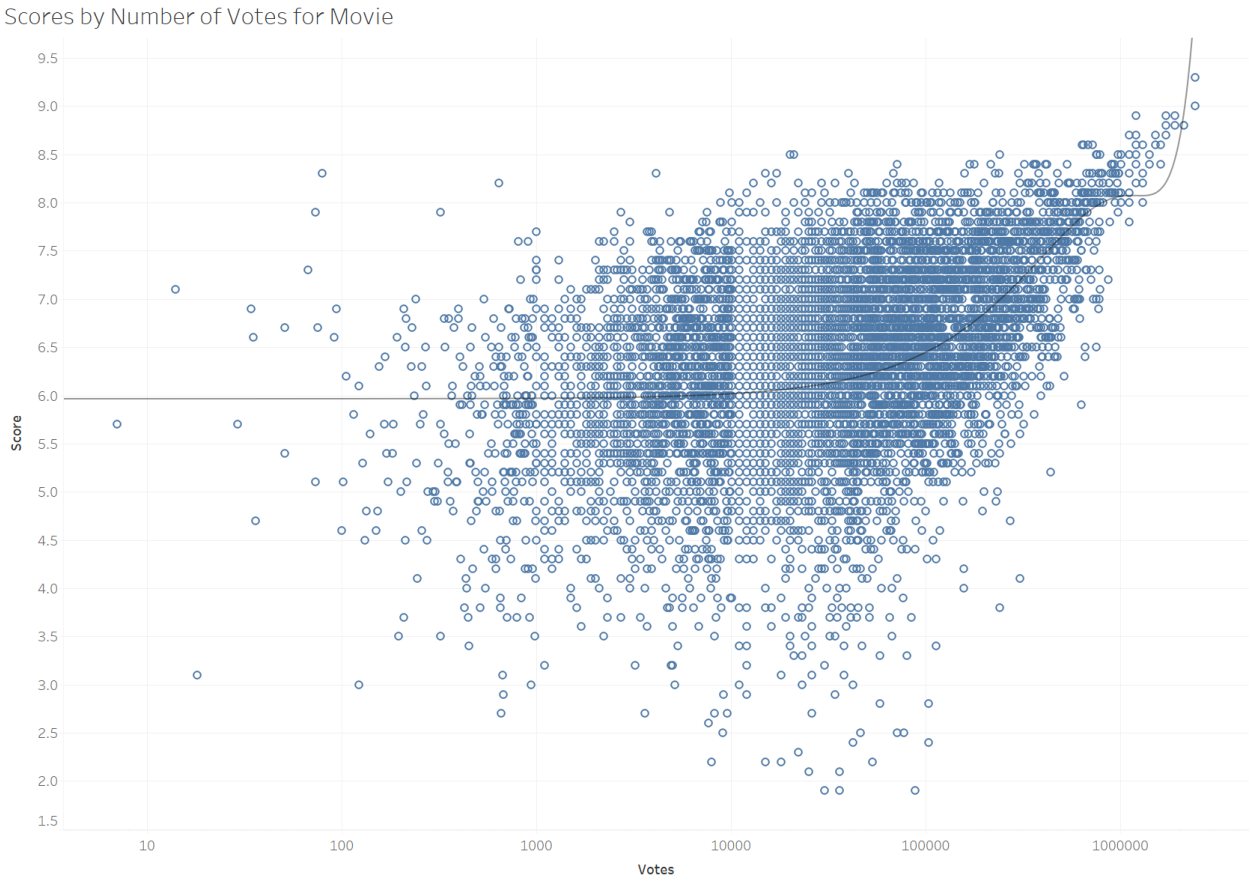
Scores:



The trend of count of Score for Score (bin).

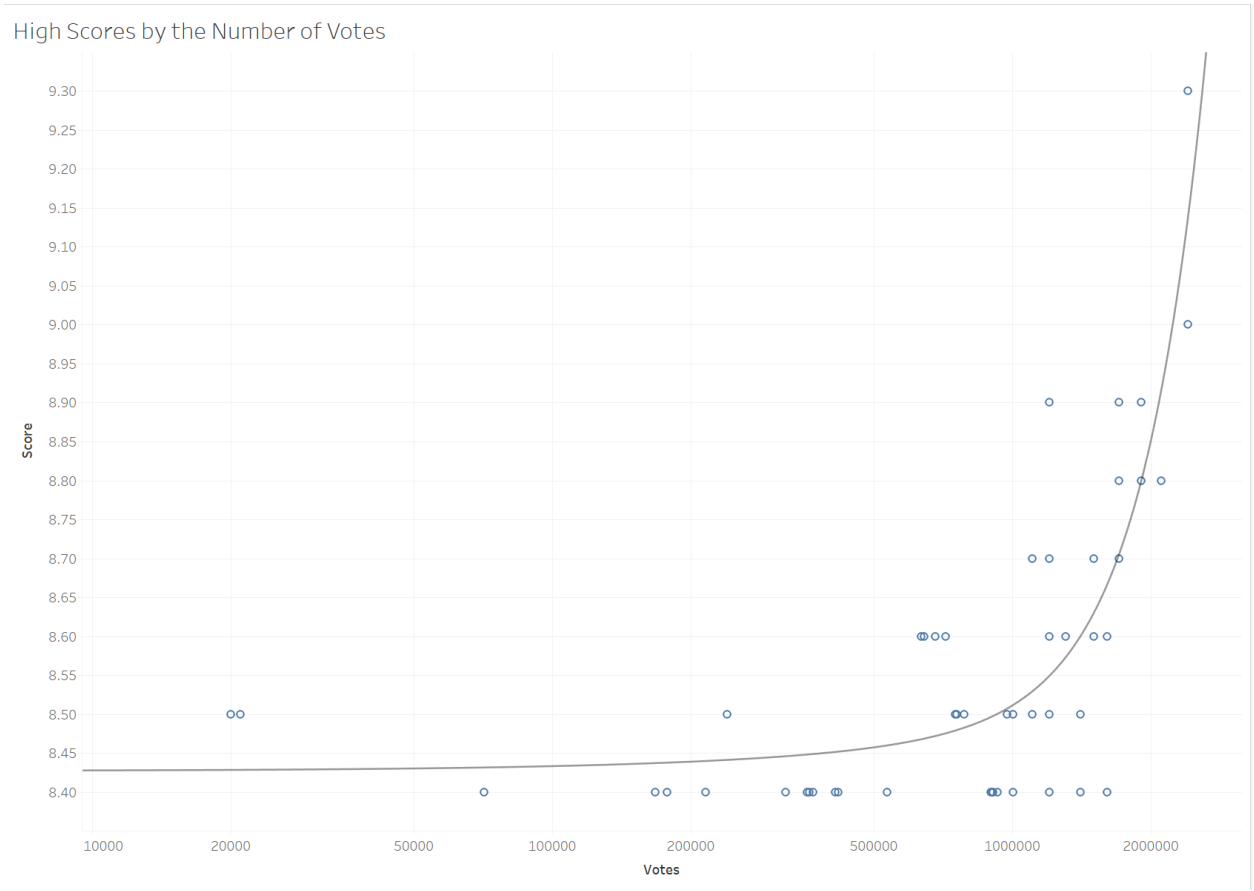
In this first visualization, this is the distribution (by count) of all 6280 movies over 40 years by their scores. In this distribution, it is clear that the average score for movies is not the expected average at 50% or a 5.0 score, rather in between 6.0 to 6.5 (the distribution of score counts seem to be skewed to the left). There are not many movies that are scored poorly (from 1.0 to 3.5) and neither are there many movies from 9.0 to 10.0 range either. There seems to be a majority of movies in the range between 5.0 to 8.0.

Scores and Votes:



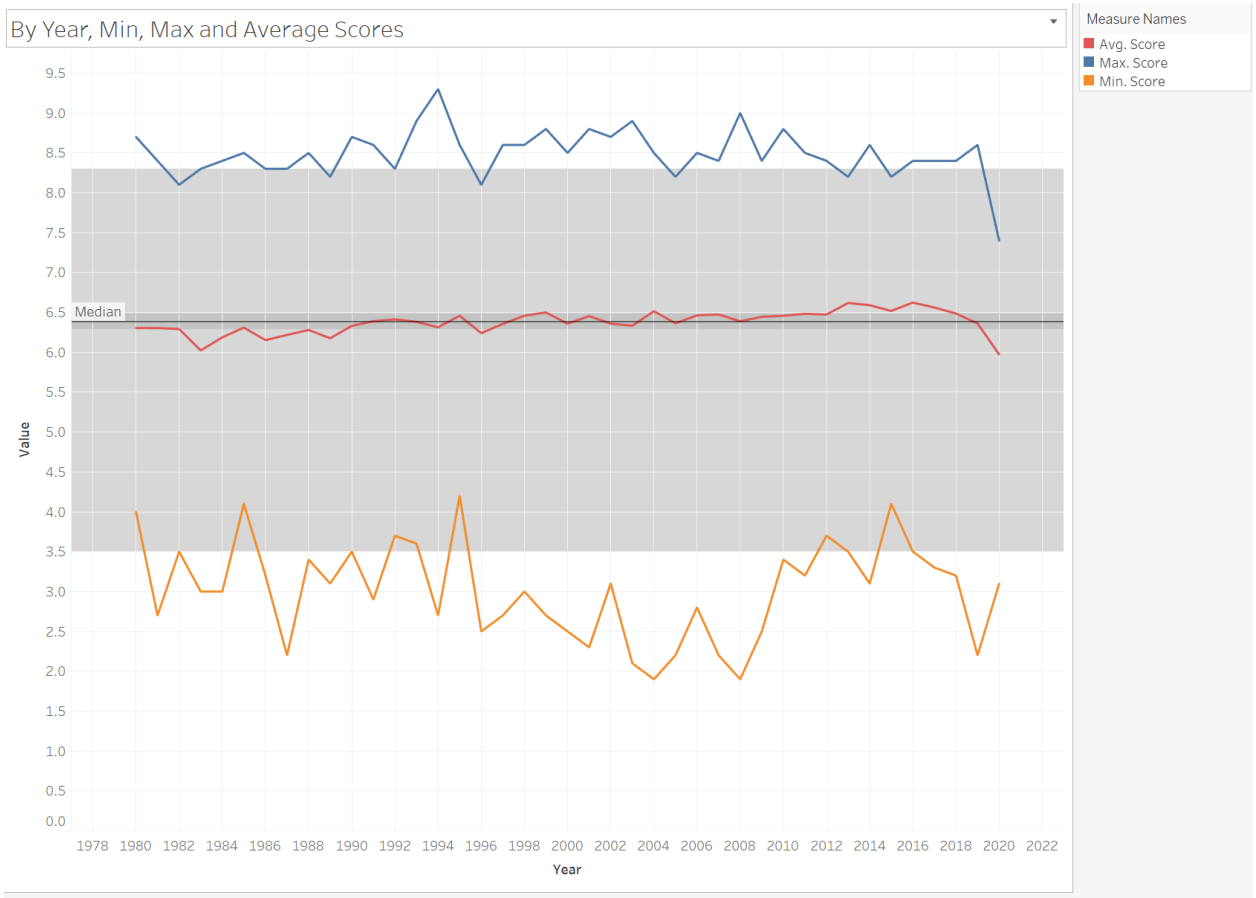
Votes vs. Score.

In analyzing the score, it is important to factor in how many people voted for each of these movies, especially to see if there is a trend of overwhelming consensus for highly and lowly scored movies. Using a trend line, there is nothing significant enough to report (hard a small r value for the trend line, likely because most of the movies have a very similar number of votes ranging from 10000 to 100000, so it is hard to find any relationship). However, for some of the extremely well scored movies (8.0+ scored), they seemed to get a lot more votes than other movies when looking at just the visualization itself. (the visualization is log-scaled). It would probably be interesting to zoom into the higher ranged scores to investigate if this visual trend can be confirmed.



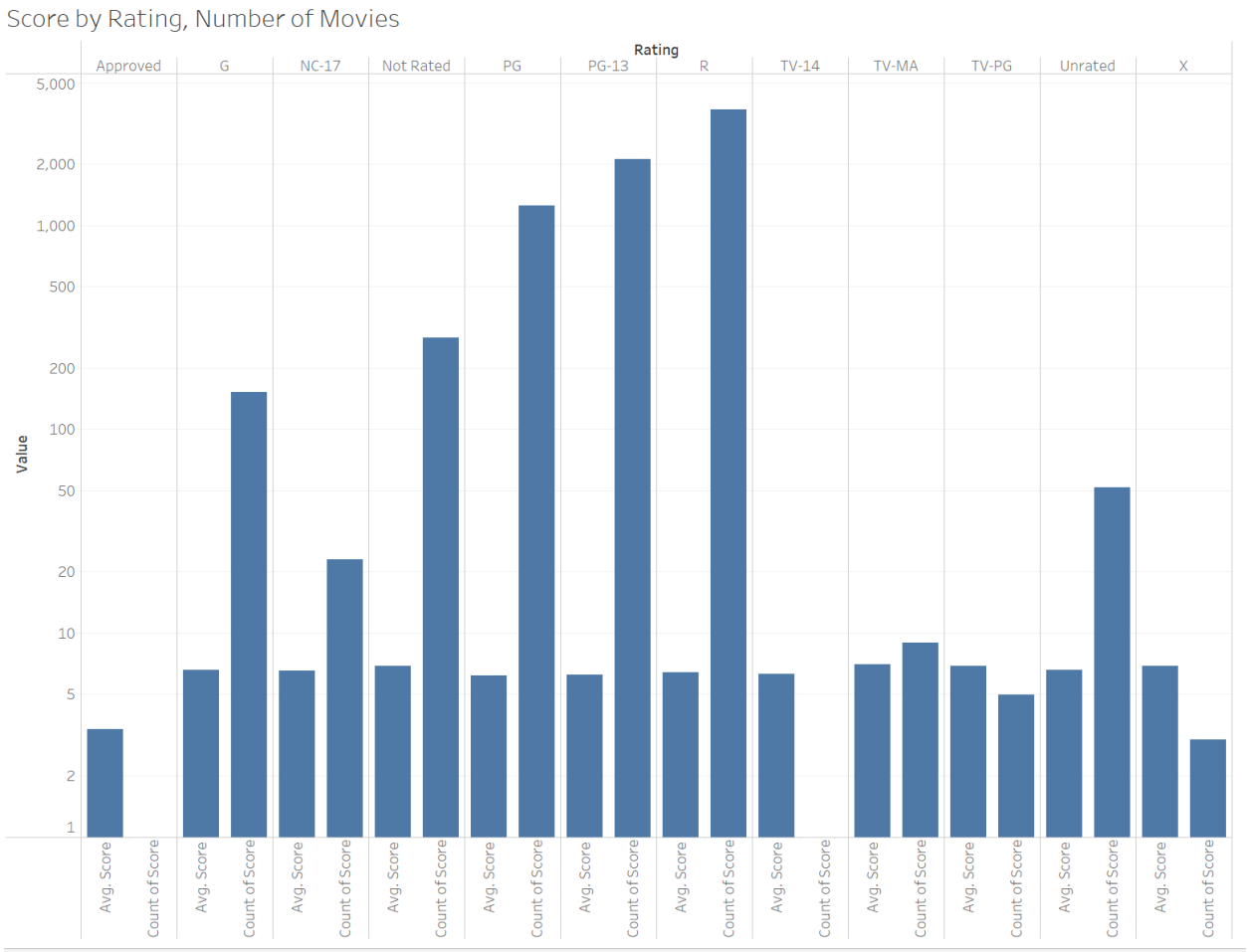
When zooming into the data to focus on the best movies, there seems to be a more prominent relationship between the score and voting and higher scored movies do in fact receive more votes (significantly higher r value for the trend line), as seen by the previous visualization and confirming the pattern seen in the visualization above.

Scores and Year of Movie Released:



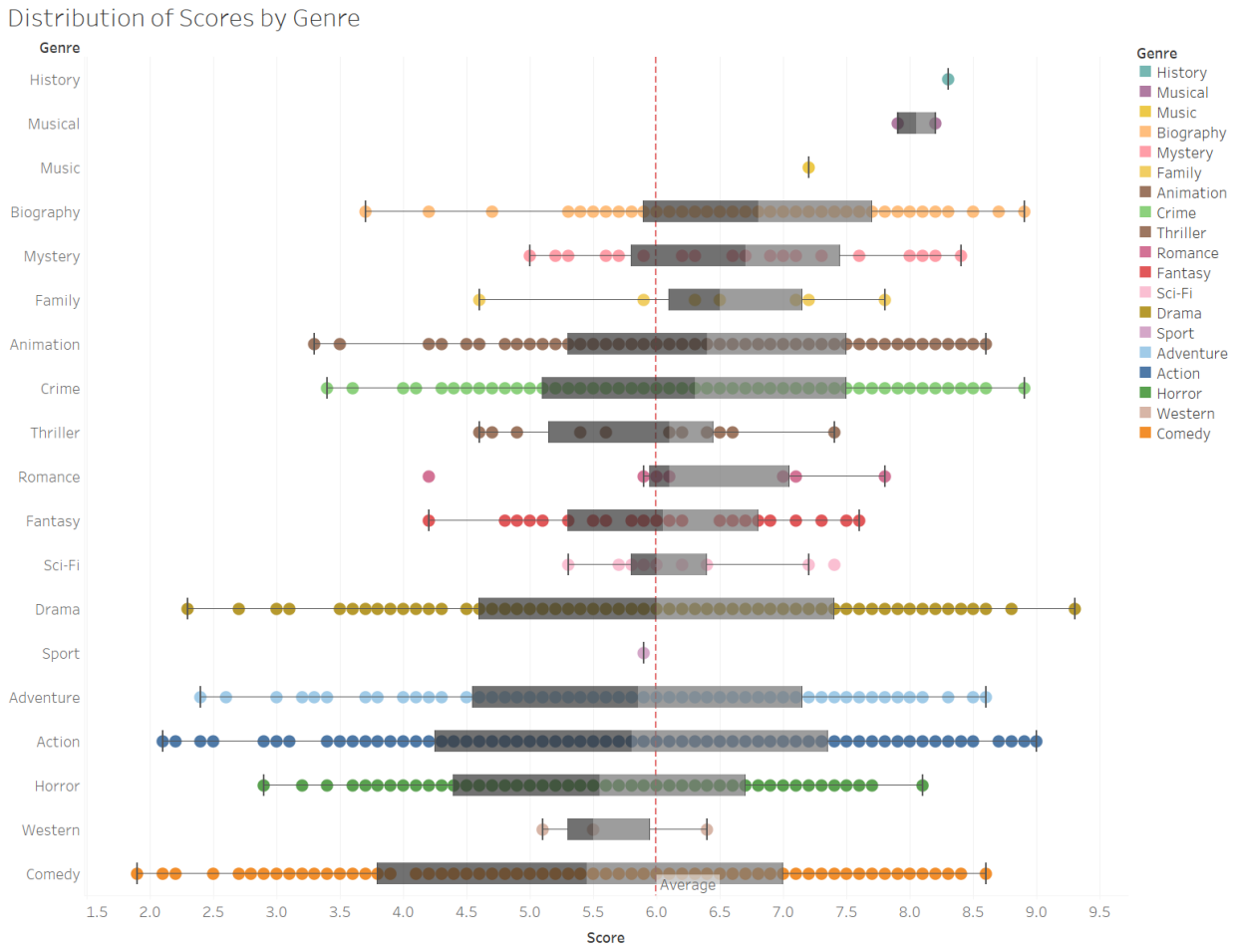
Another factor that could be influencing score could be the year that a movie was released, especially considering how the quality of films have increased over the years. Graphing the Average, Minimum and Maximum scores to see if there are trends (ups and downs due to movies being older/bad movie years). However, nothing consistent seems to exist and throughout the years, the average, min and max scores are relatively in the same range (same with the distribution of ratings), meaning that year probably does not influence the score too significantly. This could also be due to the movies per year being limited to 200 per year (and the selection of movies not being random) but nothing substantial can be concluded.

Score and Movie Rating:



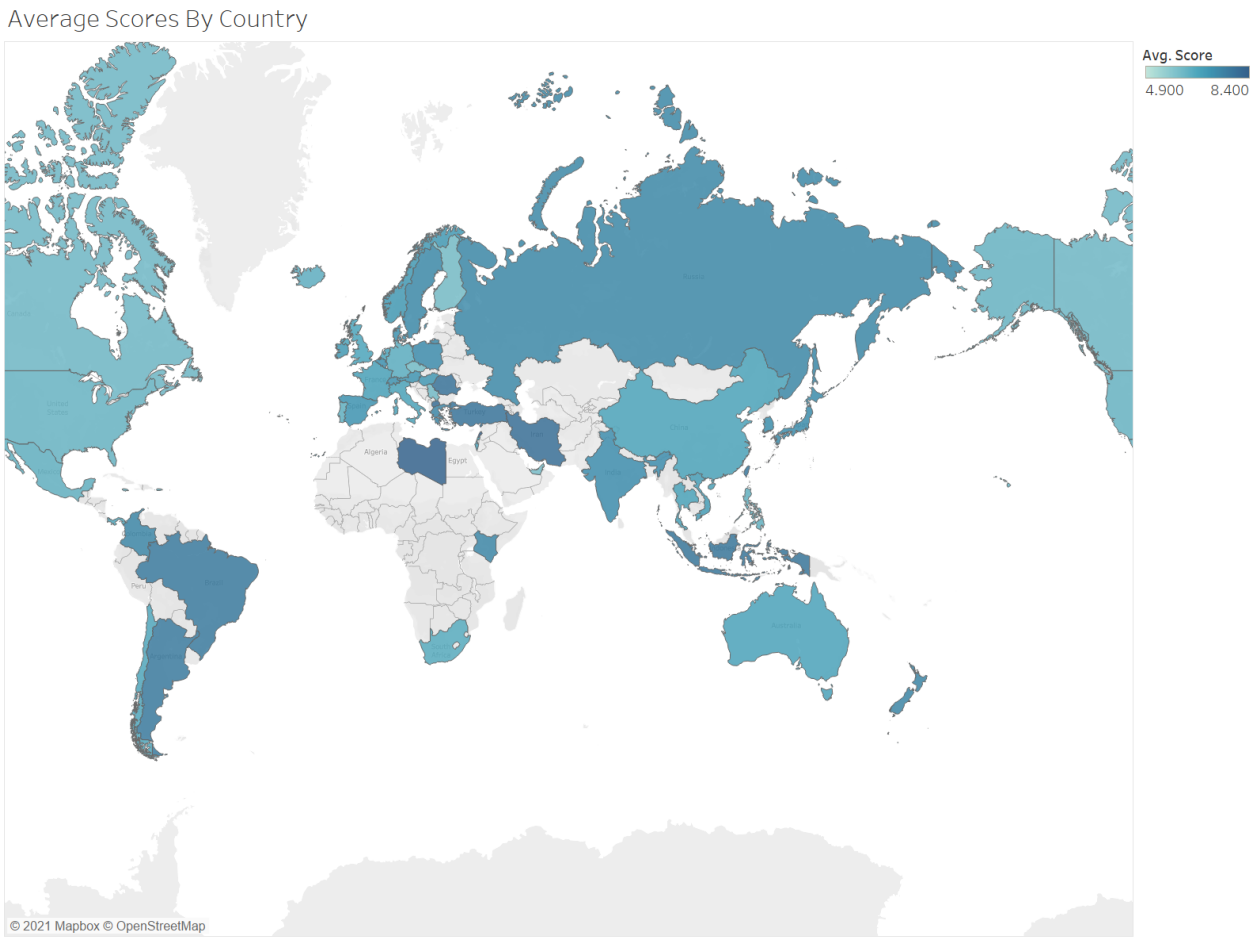
The next variable to look at is the rating of the movie to see if one of these rating categories such as family (PG) movies score better than others. There were some trends initially such as the average of approved movies scores being lower than other categories. However, when looking at the total counts/# of movies in these categories, it is hard to make this as a definitive conclusion since there are not a lot of movies within this category. Overall, the average scores between different ratings were fairly consistent with the average of the score (being around 6.5), so this does not seems to be a determining factor on the score.

Score and Genre of Movie:



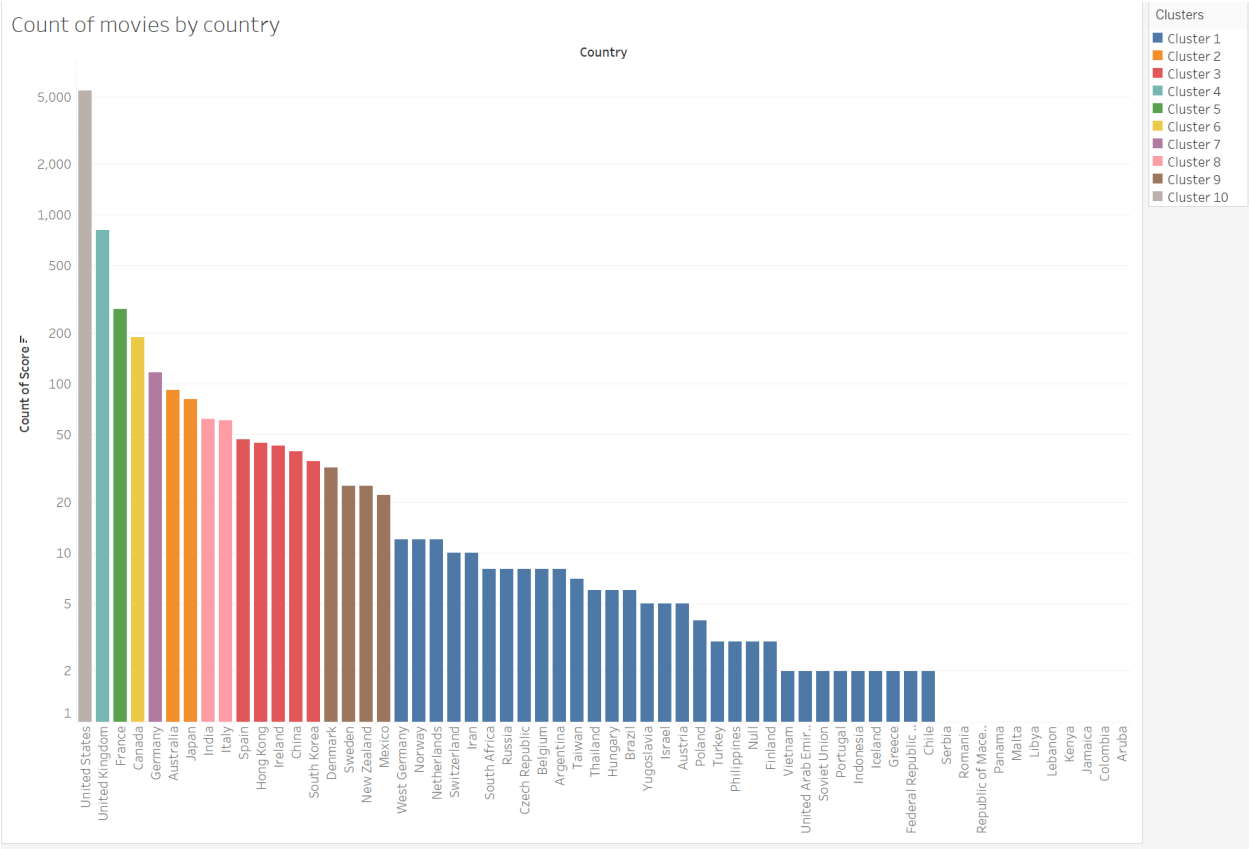
The next aspect to explore is genre alongside the scores as some subset of movies (such as oscar movies) are more liked by certain audiences. This visualization displays all genres distributions of scores in boxplot form. At first glance it seems that history, music and musicals are much better than all other genres. However it seems to be the case that there are not a lot of movies to evaluate in those categories. Though in general, movies of more serious genres such as biographies or mysteries seem to do well in the score department. Whereas comedies and dramas seem to be varied but also produce many movies. It would be interesting to explore the intricacies of genres of movies along with scores with more data such as gender and age breakdowns (wasn't available in this dataset).

Score and Country:



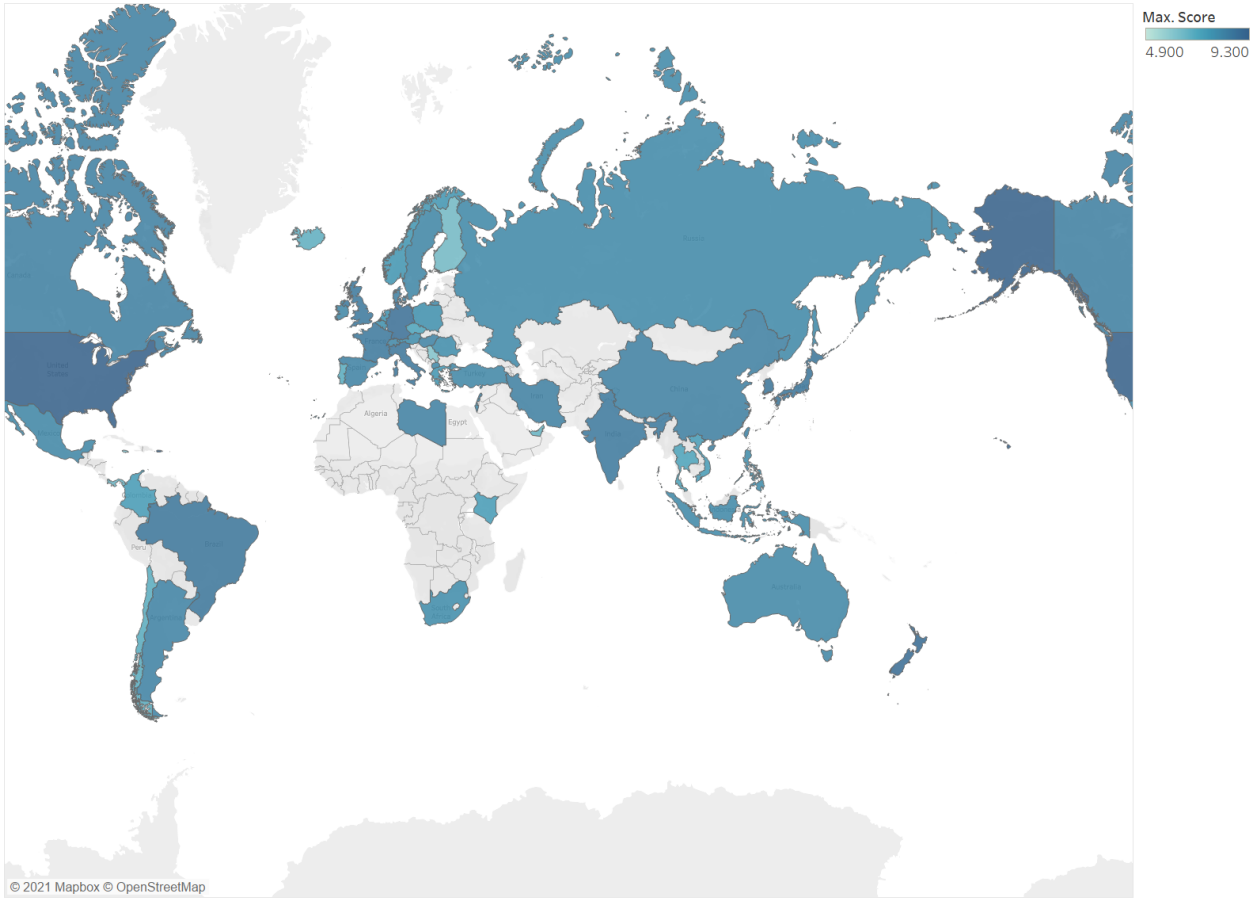
Map based on Longitude (generated) and Latitude (generated). Color shows average of Score. Details are shown for Country.

The next factor to evaluate is the country of origin of certain movies in terms of their average score. One would think that developed countries such as the US and European countries would perform better than other countries since they produce more movies than other countries. However, that does not seem to be the case with certain countries in South America and Africa that seem to have higher averages.



However, on further investigation, clustering the countries based on the count of movies indicates that many of the higher averages that some of the countries had were due to the lack of the number of countries. The US and UK had significantly more movies than other countries (as shown by the clustering), so it might be better to evaluate trends using the maximum scores or some alternative methods.

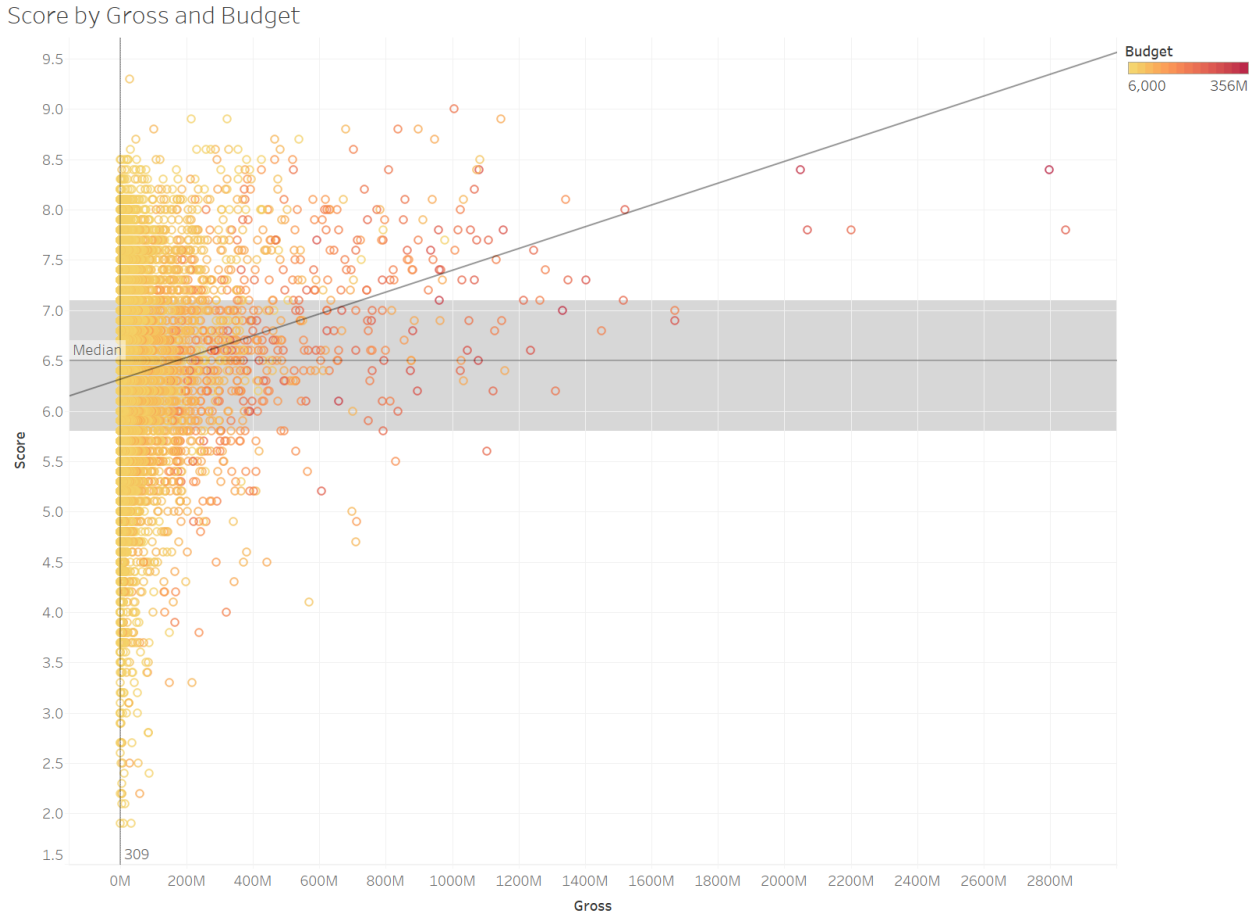
Maximum Scores by Country



Map based on Longitude (generated) and Latitude (generated). Color shows maximum of Score. Details are shown for Country.

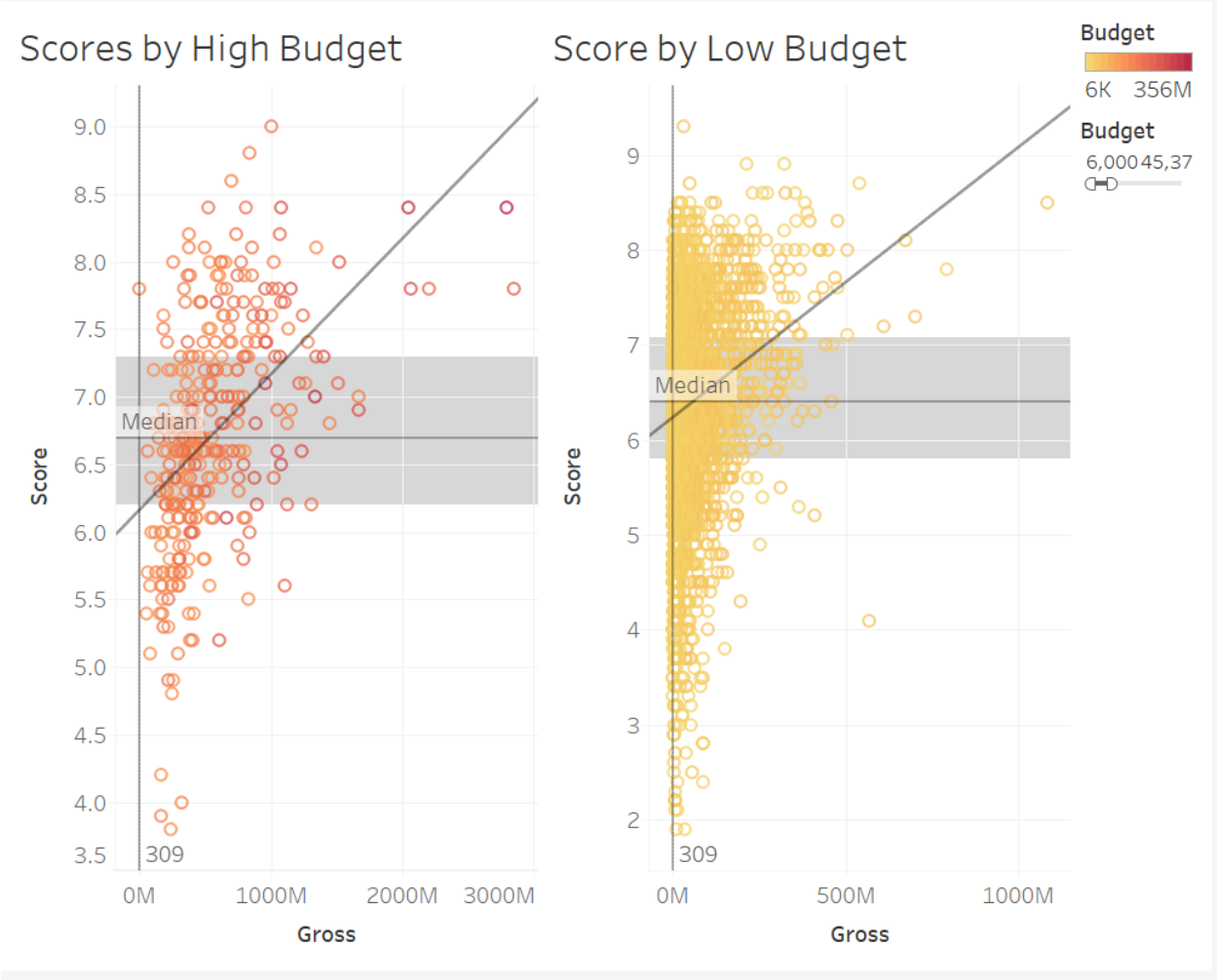
When looking at the maximum scores, it seems that all countries have the potential to produce highly rated movies, not just the countries highlighted by the average score map. The US, in fact produces some of the highest scored movies (since they have so many films created there). It could be the case that the movies that are found on the movie website from less prolific movie producing countries that they either smaller but better overall pool of movies or the badly performing movies have been excluded from the website (explanation of the trends present).

Score and Gross & Budget:



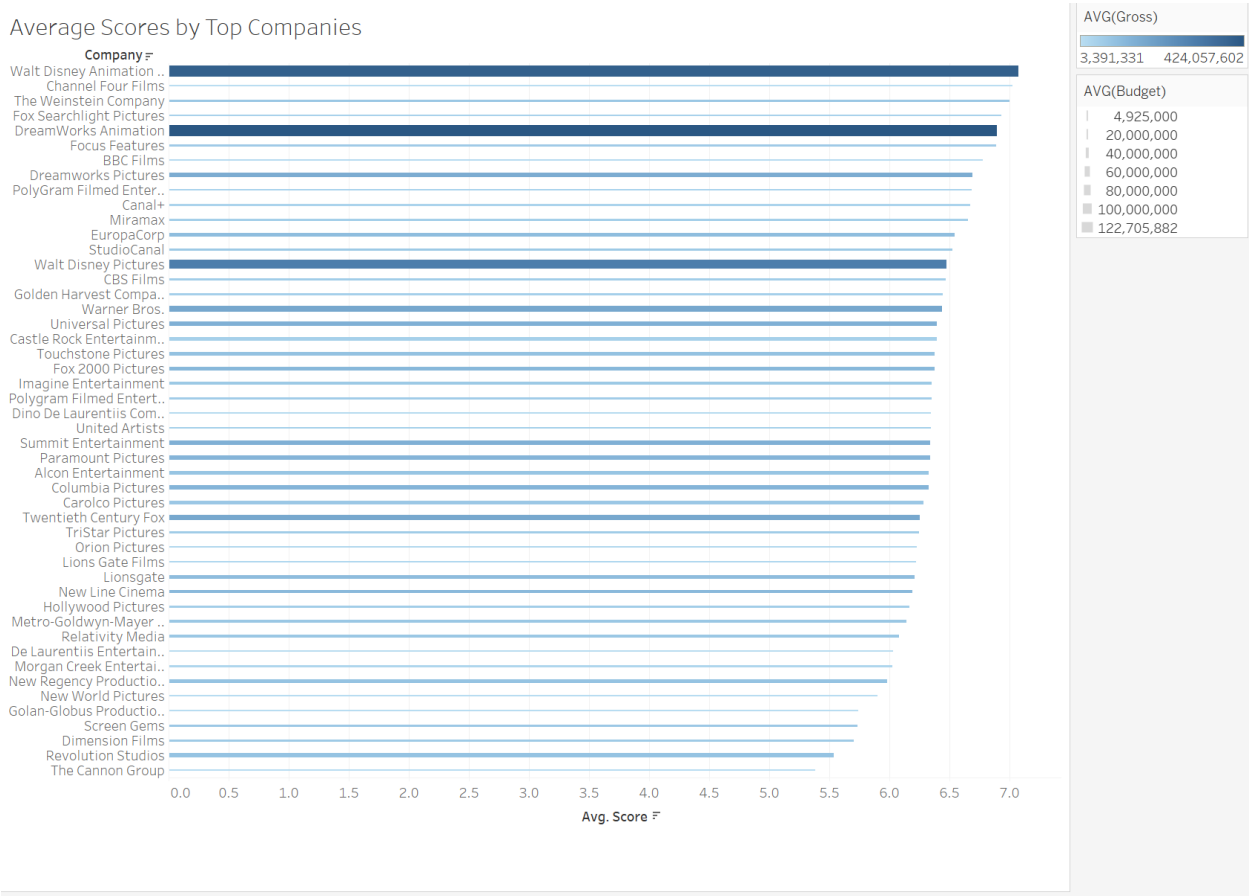
Gross vs. Score. Color shows sum of Budget.

The next component to compare to score is the budget and the gross of the movies. It seems intuitive that higher budget and gross would indicate a better scored movie. Though there is no trend correlation between the data, it does seem like there is some relationship between high budget and gross in general leading to better overall scores when it is more varied for lower budget and gross movies. It would be interesting to compare budget and gross with other variables to see other interesting relationships.

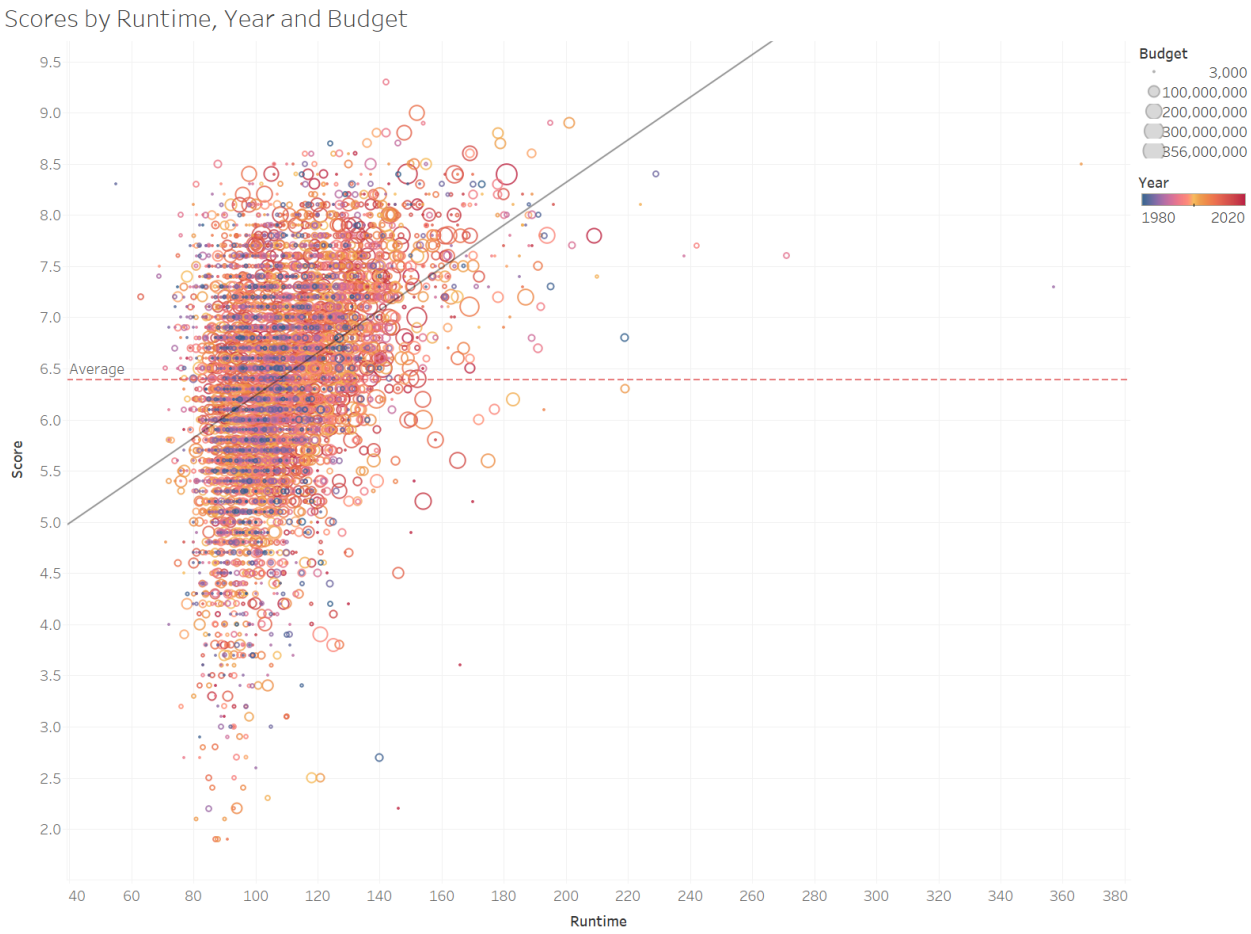


This visualization compares the high budget and low budget films and where they typically fall. There are a lot more low budget movies than high budget movies, however, in general, high budget movies seem to generally have a higher likelihood of receiving higher scores rather than very bad scores (in the 2-3 star range) and seem to also receive more profits/gross.

Scores and Movie Company:



The next attribute to analyze scores is by the companies that produce more than 15 movies by gross and movie because some studios are more popular than others (such as disney). There does seem to be a somewhat significant difference between certain movie companies than others and that can be due to the gross and budget differences between studios. For example, bigger and well known companies such as Walt Disney Animation seemed to be spending more and making more money on average along with receiving higher scores from the public in comparison to other lesser known companies with lesser budgets and grosses. So score of a movie can be somewhat dependent on what company they were made by, especially if made by popular companies (whereas mid tier companies have mid of the road results).



Runtime vs. Score. Color shows details about Year. Size shows sum of Budget.

In the same line of reasoning, runtime also could be impactful on the score of a movie, with longer movies getting worse scores? However, there doesn't seem to be such a trend present, with most runtimes having varied results and movies having fairly similar results. When overlaying with year and budget, nothing significant can be determined other than movies in the past had smaller budgets and shorter runtimes (not relevant to the questions overall).

Some general conclusions from this report:

- Genre of movies has some impact on if the movie will have a higher chance of being well-regarded by audiences (history, biography vs. drama, comedy)
- Gross and Budget being higher for a movie mostly ensure that the movie will typically have a score above the average.
- The company that a movie is produced at can have an influence on the score of the movie.
- A movie's year of release, rating and runtime does not seem to affect the movie's score
- All countries are capable of producing well-scored movies.
- **In terms of making a movie that scored well on movie websites, the genre, the company and the budget and gross seem to be very important factors.**

Next Steps/Issues to resolve:

- Not all aspects of movies were looked at in detail such as director, stars, writers
- Not all aspects of voting breakdowns were found in this dataset such as gender and age of the voters.
- Genre, Gross, Budget and Companies seemed to be the biggest factors influencing the score and should be explored further.
- Should use a more all-encompassing dataset of movies rather than a smaller sampling of movies.