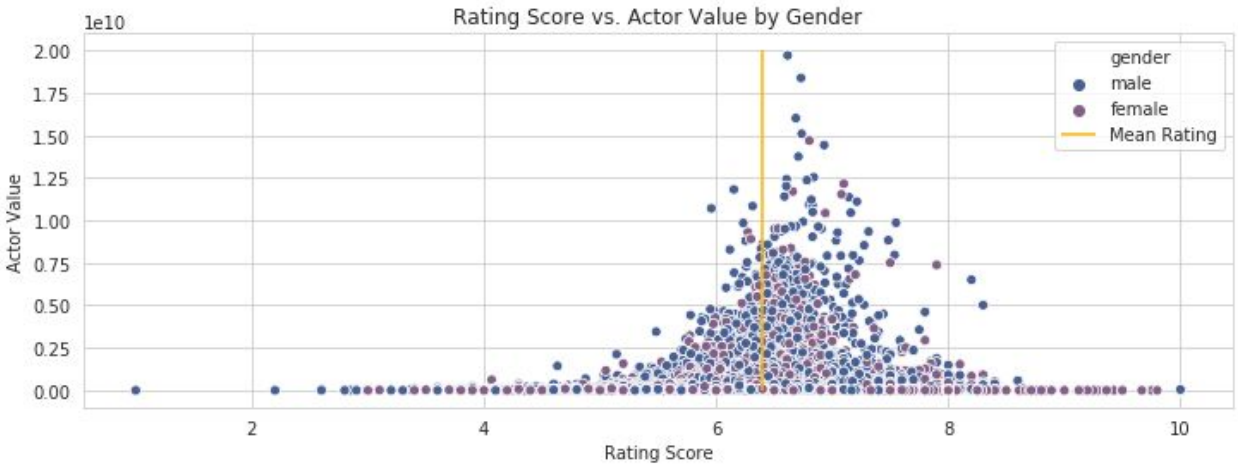
Capstone 2: Data Story

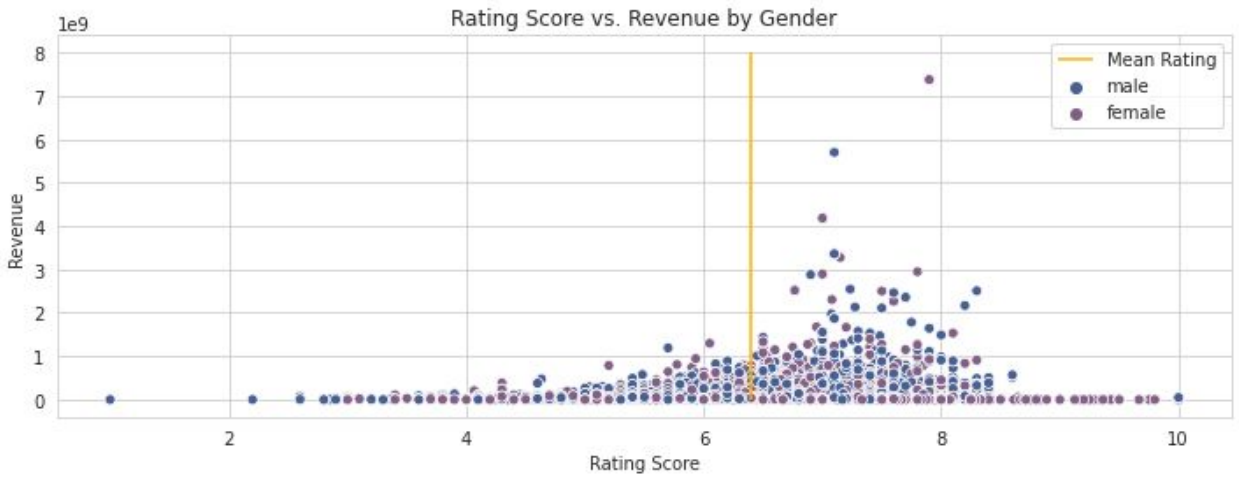
The visual analysis for the second capstone project began with revisiting the observations noted in the Data Story notebook of the first linear regression project. It was important to highlight any interpretability improvements that may have resulted from the way that the aggregation of the monetary values had been shifted, from summation to averaging, along with the removal of the outlier movies with the largest revenues. Two notebooks were created, which showed the results of applying these changes in succession. The first Data Story notebook was created with the dataset that had been aggregated, using the average monetary values for budget and revenue. The second Data Story (Optimized Target) notebook used that same aggregation method, but only after the movies with outlier revenues had been removed.

The first relation observed was between the actors' average movie ratings and the target variable, the revenue of those movies, when sorted by gender.

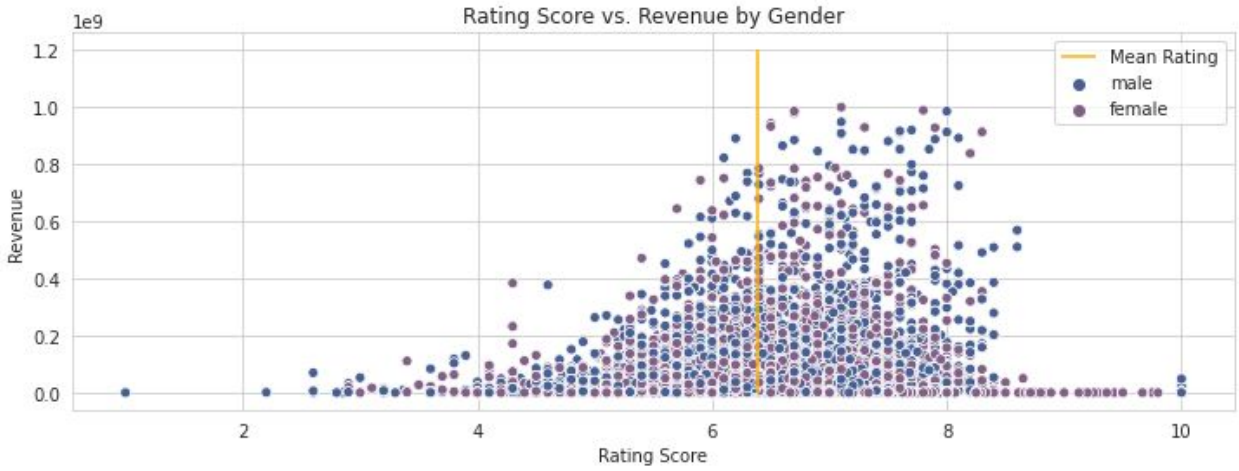


The plot for the aggregation of revenues using summation, named Actor Value in the first project, revealed gender bias in the data, as there were very few purple dots, signifying female actors, at target values above $8,000,000,000. Also, actors with higher target values counted in larger numbers at rating scores greater than the average, shown by the vertical yellow line.

After the switch to the aggregation of revenues using averaging, called revenue in the second project, was made, the genders appeared more evenly distributed at the highest revenues.

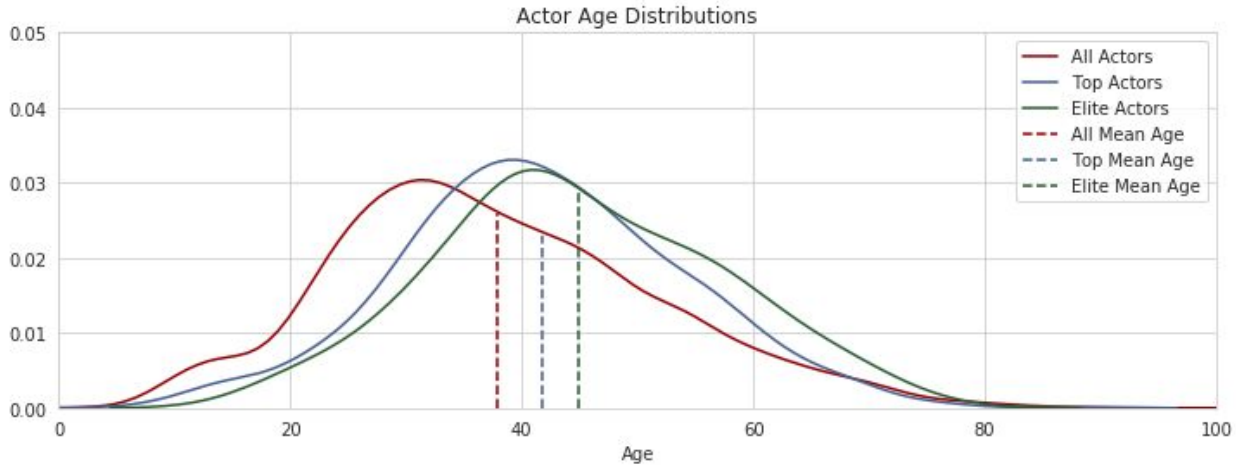


The relation between revenue and rating persisted. The Pearson correlation coefficient between the two was found to be 0.17. The key difference to notice was that the actors with the largest target value, now average movie revenue, were no longer obviously biased toward male actors. In fact, the Pearson coefficient between these variables was -0.02 for both datasets of the second project. This was the result of removing their advantage of having more movies to contribute to their revenue sum, since the careers of male actors were longer on average. Predictions made for the highest potential earners for the average movie would have been more biased toward them, when using the dataset from the first project. Removing the outlier movies did not improve this result. The genders are well distributed, and the top revenue data points are more tightly clustered.

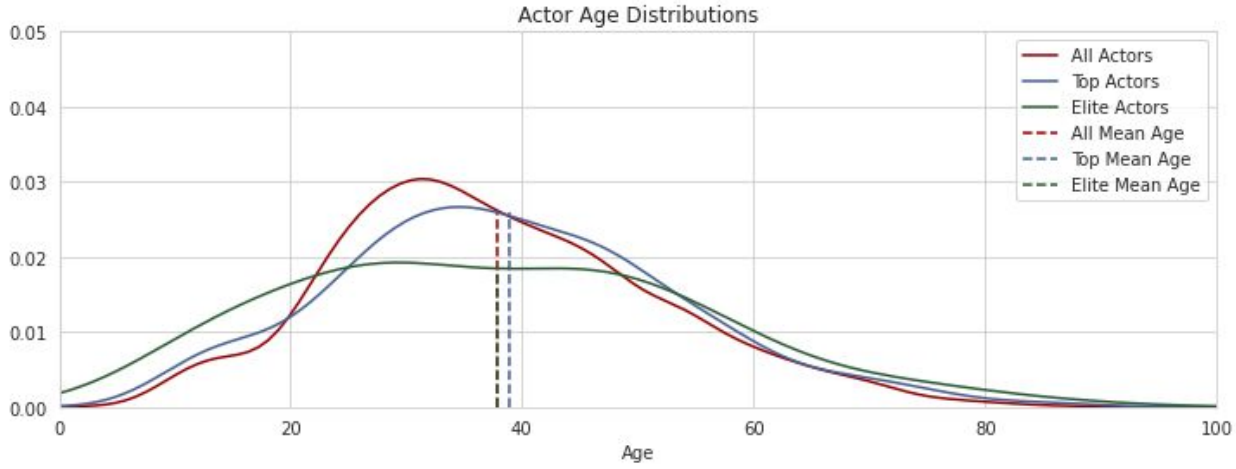


The analysis continued after the actors were grouped by various tiers, based on the average revenues of their films. The first tier included all actors. The second group, named the top tier, contained only those actors with average movie revenues that were in the top 10%, overall. The final tier consisted of actors whose average movie revenues made up the top 1% of all actors. These were the elite tier of actors. Note that the actors in the elite tier were counted in the top

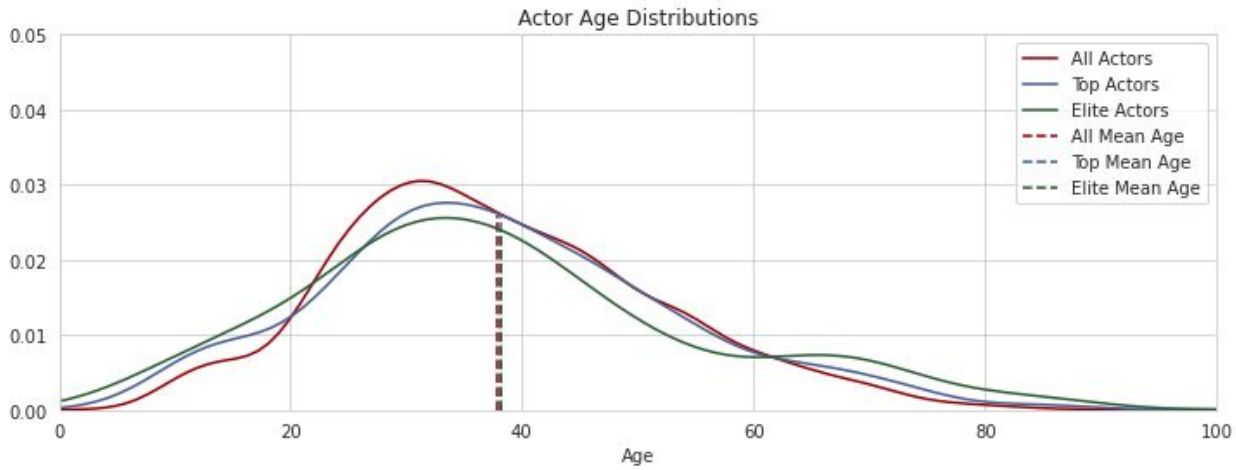
tier, as well. At this point in the notebook, the various predictor value distributions were analyzed with respect to the target variable. The changes in the actor age distribution were noteworthy.



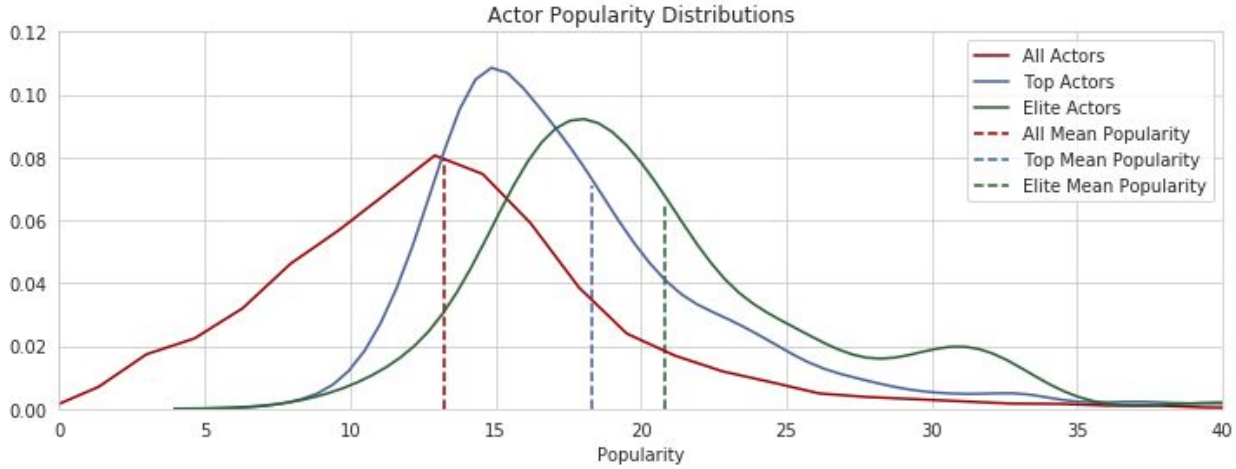
The average actor age increased with the progression to the upper actor tiers, when observing the data from the first project. These values started to converge, after the aggregation method was changed from summation to averaging.



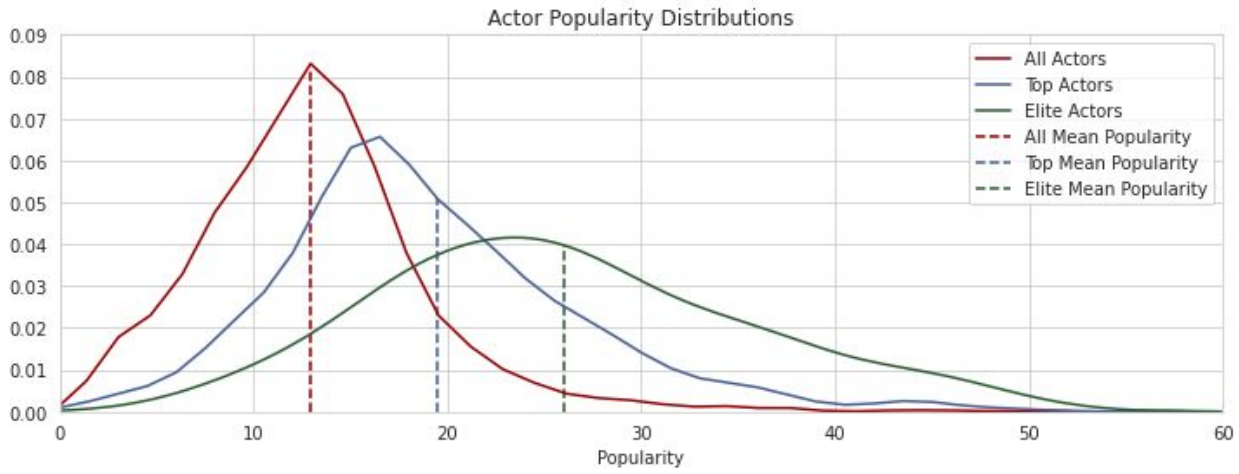
The effect of removing the films with the highest revenues completed this convergence.



The movie popularity rating was a TMDb proprietary metric, based on recent traffic pertaining to a movie on their website.



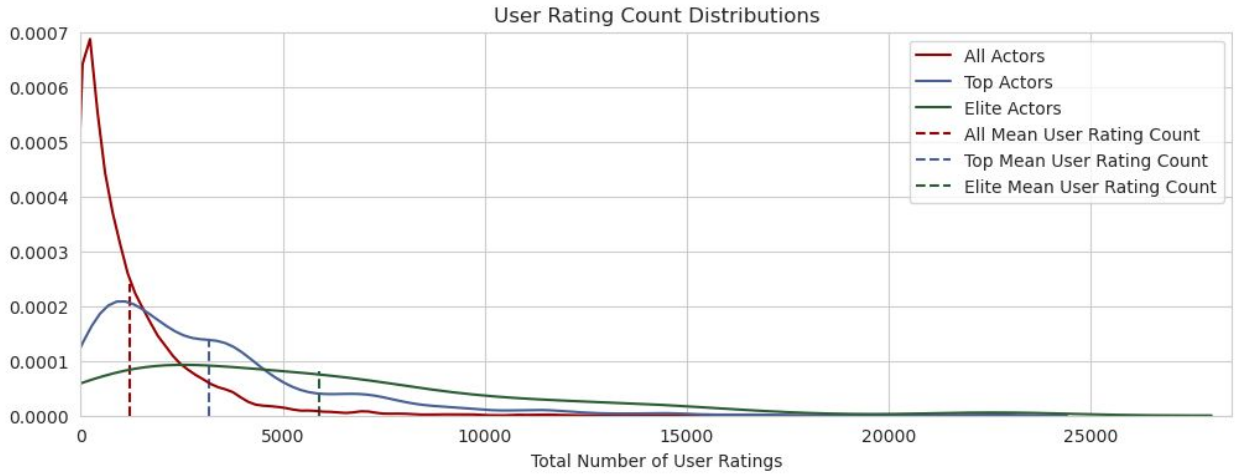
This value increased as the tiers progressed toward the elite actors.



By the time the final dataset was reached, the popularity values for the top and elite actors had shifted higher. This was the result of removing a lot of the older blockbuster series and Disney animated films, which would not have as much current interest, as the more recent releases would have.

Both the average movie rating, and the total number of users who rated a movie, grew when observing the progression through the tiers of actors, which were derived from increasing thresholds of the actors' movie revenues. This trend was present in every form of the data.

The user ratings for these films directly correspond to which movies had the highest ticket sales at the box office. The second mode, at the highest ratings in the elite actor distribution, indicated that some of those actors elicited strong tribal support from their admirers. This is the idea of star power, or the ability of actors with high name recognition to be able to drive ticket sales.

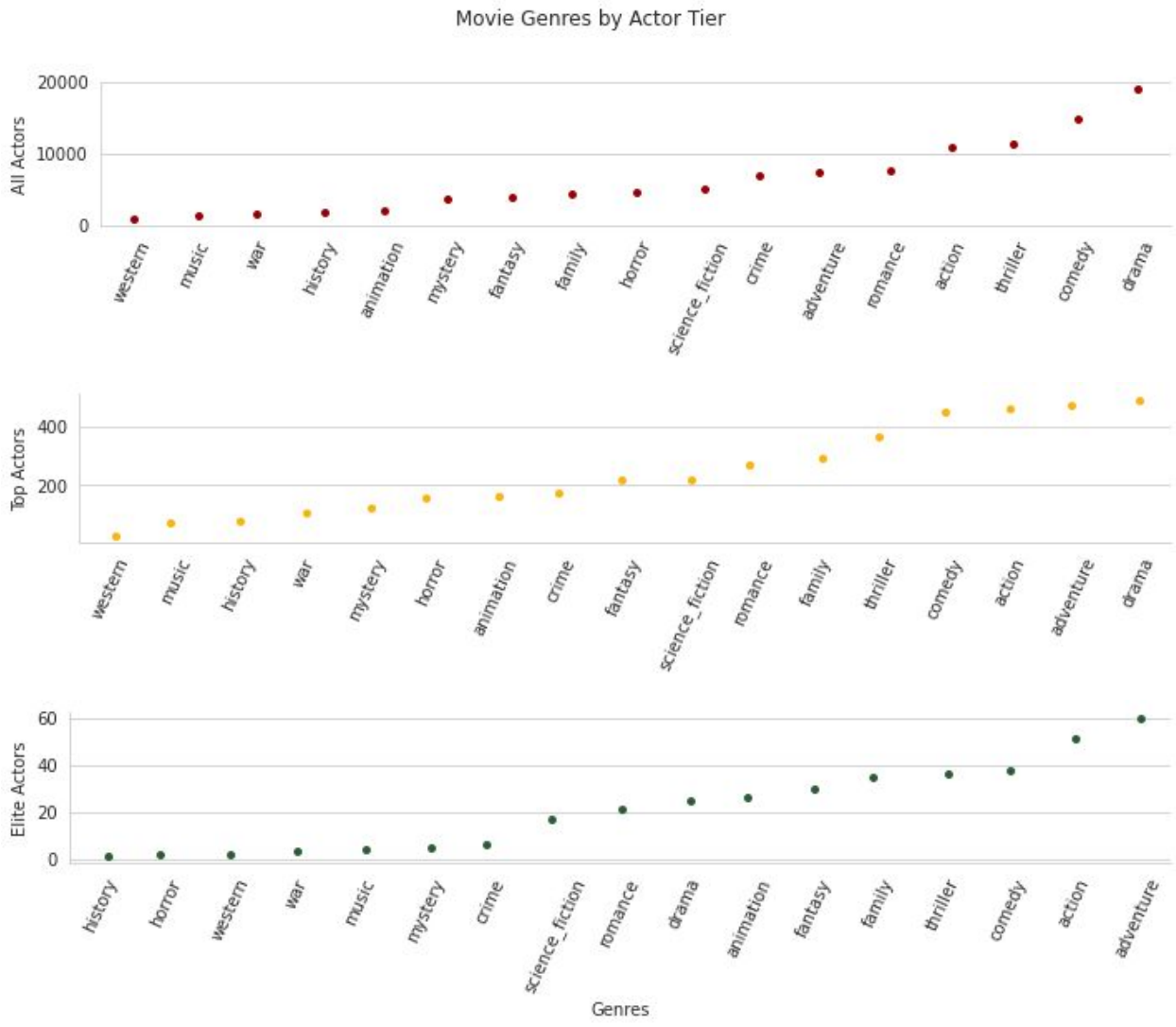


The elite actors dominated the rating traffic on the TMDb website, as well. Their average number was almost 6,000 individual ratings, which was more than 5 times the number of ratings belonging to the average actor.

Interestingly, before the removal of movie outliers, the Pearson correlation coefficients of several predictors flipped signs, when progressing through the actor tiers. The predictor, budget, went from 0.34 to - 0.32. Popularity changed from 0.33 to -0.2, and the number of movie ratings went from 0.44 to -0.26. Once the data was transformed to its final form, this behavior disappeared. It was possible that the observations for actors who were mostly, or exclusively, in movies with large revenues did not produce the same signal, as the observations of other actors. Dropping these outlier movies may have brought the entire dataset into better cohesion. This would have the effect of giving the models a simpler training set to digest, producing more accurate predictions.

In reference to the full dataset, one of the predictor correlations with the target variable was significantly strengthened, after the removal of the outlier movies. The Pearson coefficient between budget and revenue went from 0.34 to 0.5. The average revenue dropped by a factor of 60, while the average budget only fell by a factor of 10. This suggested that the outlier movies had revenues that were inflated for their budgets, compared to the rest of the movies. Given that budget is the feature most closely correlated with the target, removing the outlier movies should produce a simpler dataset for model training.

The final analysis that was revisited from the linear regression project concerned the distribution of the movie genres among the 3 tiers of actors.

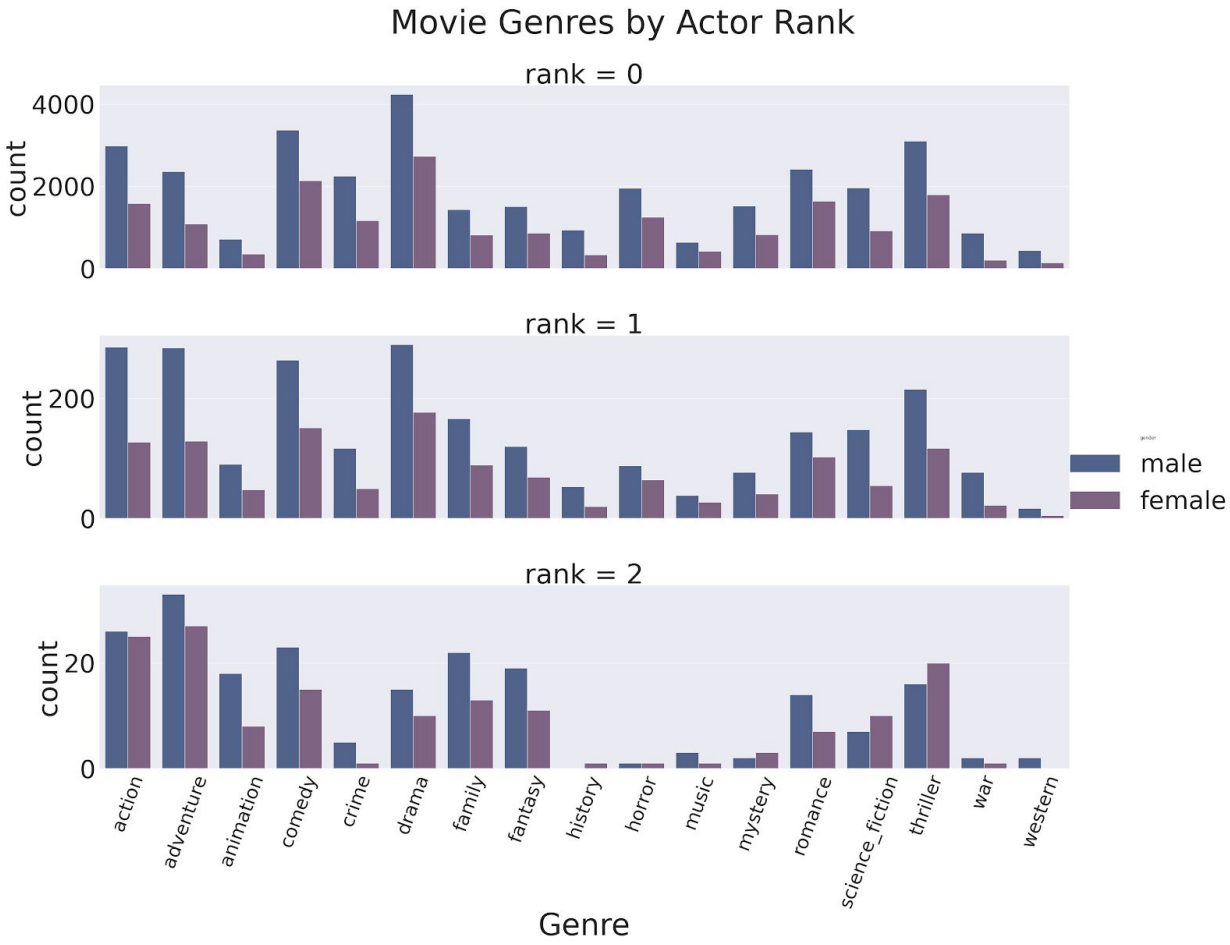


The drama genre dropped significantly in importance when observing the elite actors, which did not happen when using the dataset in the previous project. Although, comedies did have a higher count within that tier, this time around. Adventure movies began to rally for actors in the top tier, then went on to take the lead among the elite tier of actors. The genres with the lowest counts remained near the bottom, regardless of the form of the dataset.

In order to bring a more practical feel to the project, a deeper dig into the actual movies and actors behind the data was performed. The breakdown of the elite actors was observed, as they appeared in the various genres plotted above. In the first Data Story notebook of the second project, these observations were made, before outlier movies were eliminated from the dataset. In fact, the observation of the particular movies and actors behind the numbers, inspired that change, in order to eliminate the biases that were revealed. For instance, most of the appearances of elite actors who were in multiple adventure movies were acting in major film series, such as Star Wars, Harry Potter, and Lord of the Rings, among others. After the movies with the largest revenues were eliminated, what was left were mostly James Bond movies for

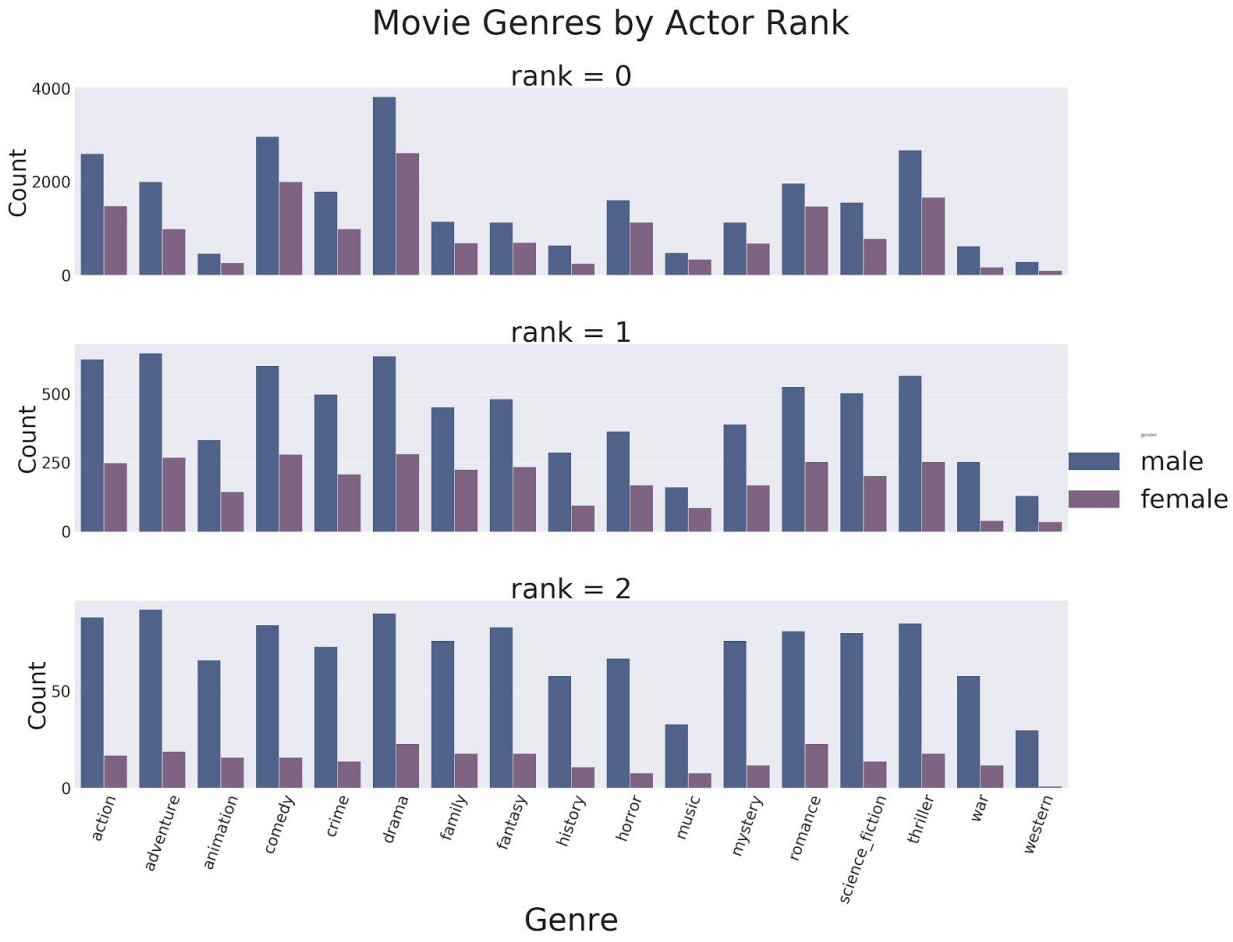
actors with multiple adventure film appearances. Though technically still part of a film series, considering the regular turnover of actors in Bond flicks throughout the years, it is fair to say that those films were a good tradeoff for Harry Potter movies, which had a more regular cast throughout its series. The same could be said for family movies. What was a list of mostly Disney movies from the 50s and 60s became one with a broader selection, including a more recent Disney release, a Pixar film, and many other movies that were not in the animation genre, as well. The diversity of the movies and actors became richer, after the outlier films were removed.

As in the first project, the actors were ranked in their separate tiers. This time, the elite actors were excluded from the top tier. This gave rank 2 for elite actors, rank 1 for only the actors with average revenues between the top 10 and 1%, and rank 0 for those actors with revenues below the top 10%. Then, the counts of the genre appearances by each actor rank were plotted, while each genre was separated by gender. Note that the elite actors were not included in rank 1.



While this plot didn't show much that was different from the previous one, with respect to the genre distribution between the different ranks, it did show the male actors outnumbering the female actors in almost every genre at every tier. This was not so unusual, as female actors comprised just under 40% of all the actors in the dataset. The exceptions were found only in the

rank 2 distribution. Thriller and science fiction were the genres most strongly represented by female actors in that rank. This observation was contrasted with the plot from the first project.



The plot from the first project showed a very strong bias toward elite actors being male. In contrast, the distribution between the genders became more representative of the dataset, after the target values were optimized. In fact, the female actors became over-represented among the rank 2 actors, after the outlier films were removed and monetary values were averaged. In the dataset from the first project, the female actors comprised just over 20% of the elite actors, where they represented over 43% of the elite actors in the optimized dataset. Recall, just under 40% of the actors were female in either dataset. The final choice of data representation increased model performance, while it reduced the gender bias of the data, something that should be ensured gets checked, when processing socioeconomic data.