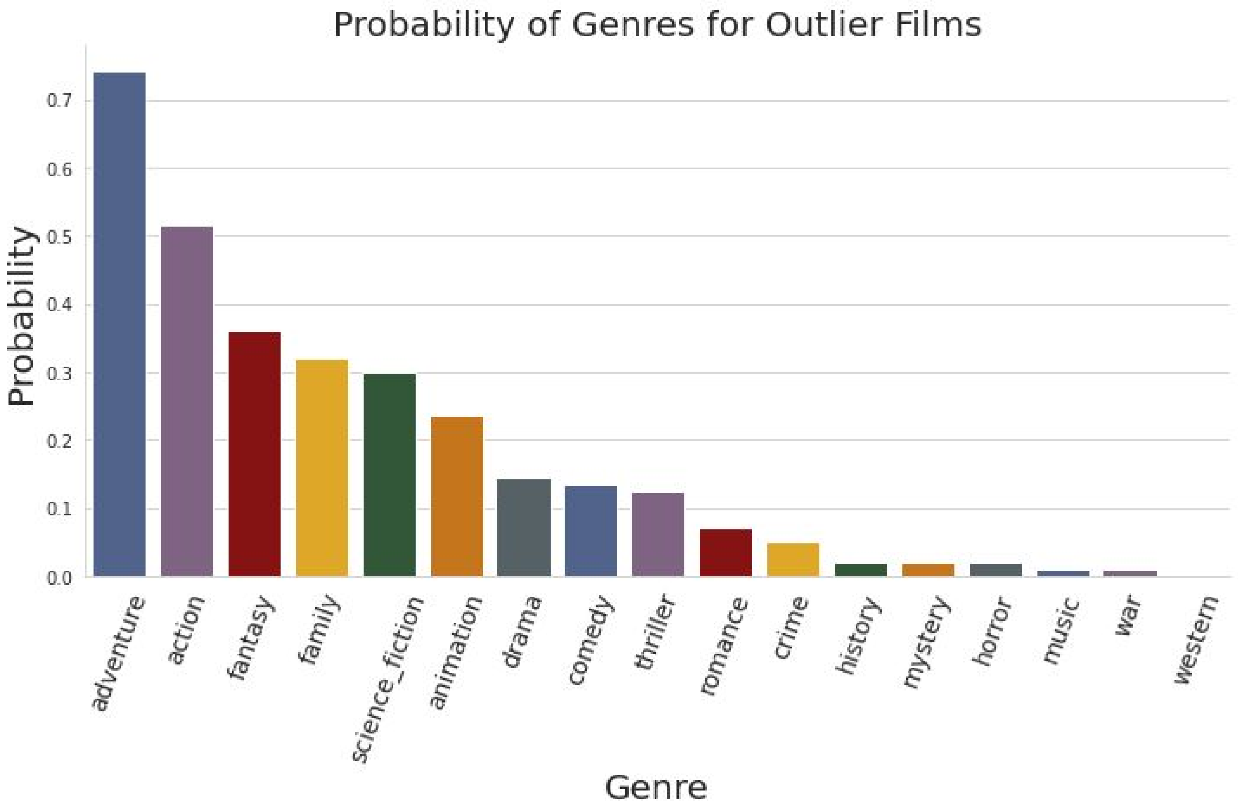
Capstone 2: Optimize Target

The other way that the datasets were different relates to the way that the data was transformed after cleaning. In the OLS Regressor notebook of the first project, the linear regression model performance was optimized by applying a variety of data modification techniques on the features, such as various regularizations, feature selection, and outlier extraction. In the Optimize Target notebook of the second capstone project, the data modifications were restricted to the target variable, as opposed to the predictors. The selection of the optimal target values used for model training was obtained by comparing model performance, after applying the techniques of log transformation, outlier extraction, and scaling on either the revenues of individual movies (before aggregating over the actors) or the average movie revenues of each actor (after aggregation). Before transformations were made, a baseline mean absolute error (MAE) of $56,799,637 was obtained for the average 10 fold cross validation (CV) test errors, using an untuned XGBoost model.

To get the optimal model, the best strategy was to modify the movie revenues before aggregation over the actors. Removing the movies with revenues greater than three standard deviations from the mean of all movie revenues proved very useful. Performing this technique, directly on the movie revenue, reduced the error by just over 20%, with a CV test MAE of $45,296,500. Taking the natural log of these values improved that error by an additional 7% to a CV test MAE of $41,215,414 for an untuned model.

After finding the best form for the target values, an analysis was performed of the outlier movies that had been removed and how those deletions changed the data that remained. There were 97 movies that were extracted from the original dataset, which consisted of 5626 films. Most of these movies were either parts of a blockbuster series or were animation films. These movies contained 541 actors. From a count of 11,693 actors in the original dataset, 83 of them were removed, as they only appeared in these outlier films. The remaining actors had their average revenues decreased to varying degrees. Mark Hamill lost nearly 98% of his average movie revenue, after the Star Wars saga was removed. Interestingly, Harrison Ford only lost 47% of his average movie revenue. Yet, he had Raiders of the Lost Ark removed, as well. This attested to the solid career he had, outside of his big blockbuster series appearances.

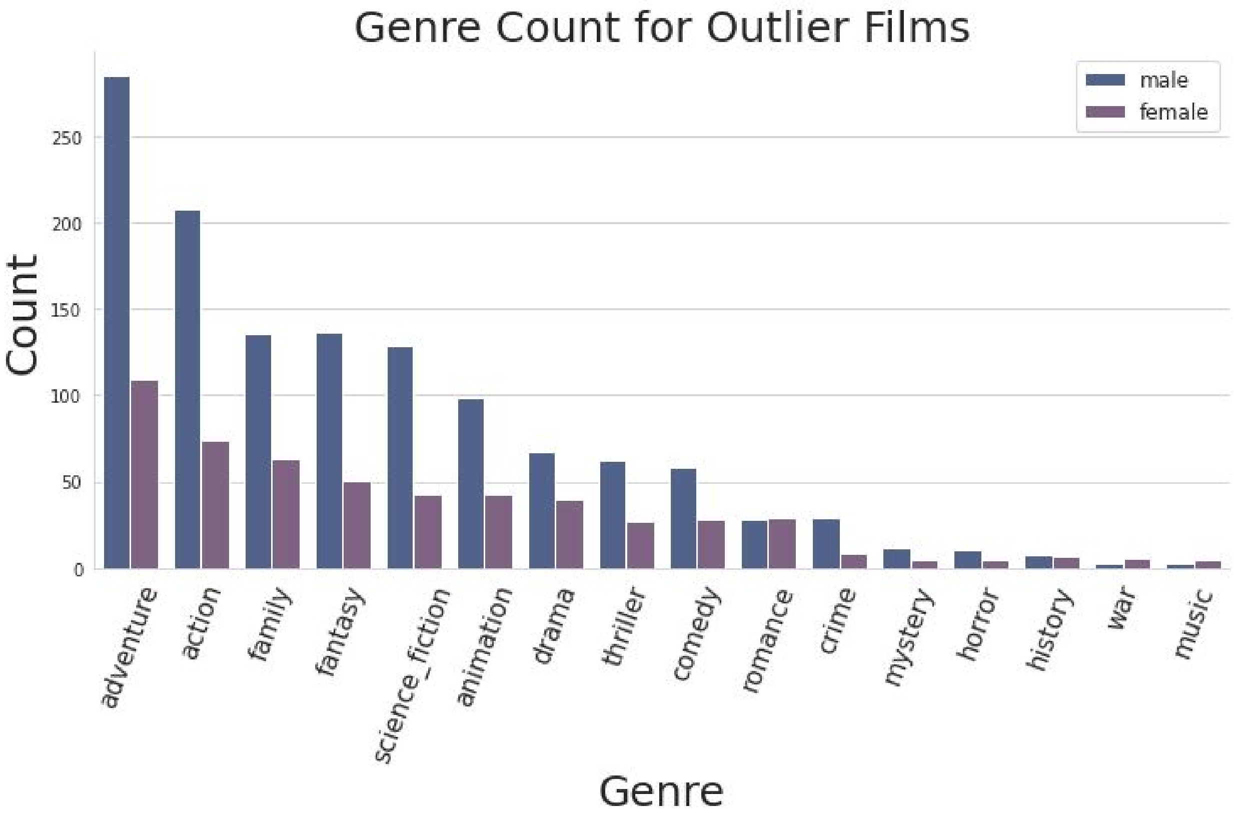
Next, the effects of removing the outlier movies shifted to how that transformation affected the distribution of the movie genres.



Adventure films comprised 74% of the outlier movies. Action was a genre description for 52% of them, as well. Fantasy, family, science fiction, and animation each made up at least 20% of the outlier movies. Many of these genres shared movies with each other. For instance, the adventure genre was included in the description of 45 out of 50 action movies, 28 out of 29 science fiction films, and 31 out of 35 fantasy releases. The family genre was named in all 23 animation films. These genre clusters blurred the lines, when defining what characteristics could be used to categorize them.

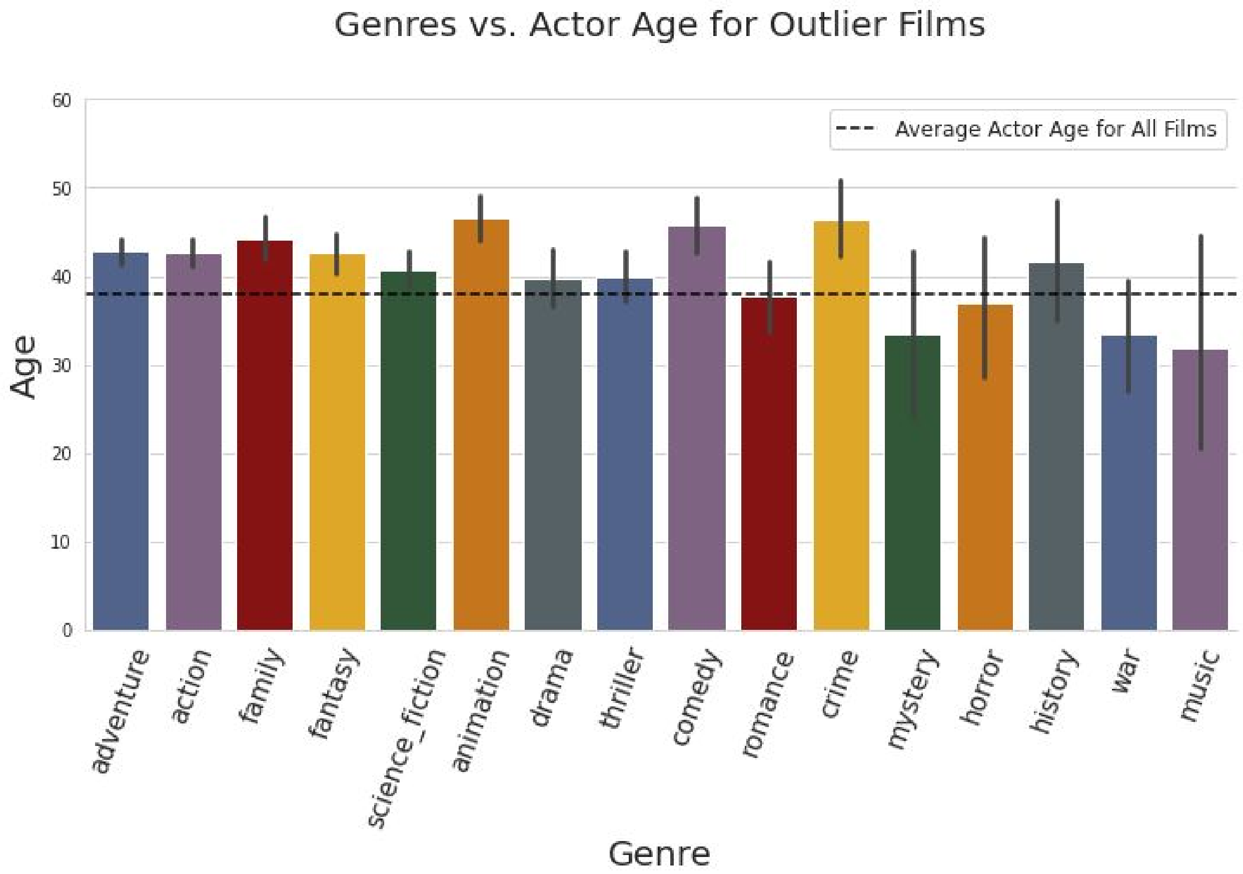
After the genres were inspected, the focus was turned to uncovering potential data biases, with respect to the socioeconomic variables. First, an analysis of the gender distribution was performed. Male actors made up almost 70% of the removed outlier data. They accounted for 61% of the full dataset. In these removed movies, the male actors had a total of 600 billion dollars in average revenue values, while the female actors had less than half that figure. Although, the percent revenue reduction, after outlier extraction, was 14% higher for the female actors. This was possible, because their count was less. In the end, the effect of many more male actors, losing a slightly smaller percent of average revenue, pulled the gender bias closer to equality.

A brief inspection of the gender distribution was performed, with respect to the genres.



The top 6 genres from the outlier movies were comprised of male actors by at least a 2:1 ratio. After these movies were taken out of the dataset, the respective gender biases in these movie categories were decreased.

The other bias that was considered was that of actor age. The actors in the outlier movies were divided into two groups, defined by the average age of all actors in the original dataset. The actors who were older than average made up 59% of those in the outlier movies. The average actor age increased from nearly 38 to just over 42 years.



This time, the top 9 genres all had average actor ages that were above average, compared to the full dataset. Both the percent average revenue change, after outlier extraction, and the sum of the average revenues in the outlier dataset were similar for each age group. This time, the effects of the removal of outlier movies was less impactful on the bias.