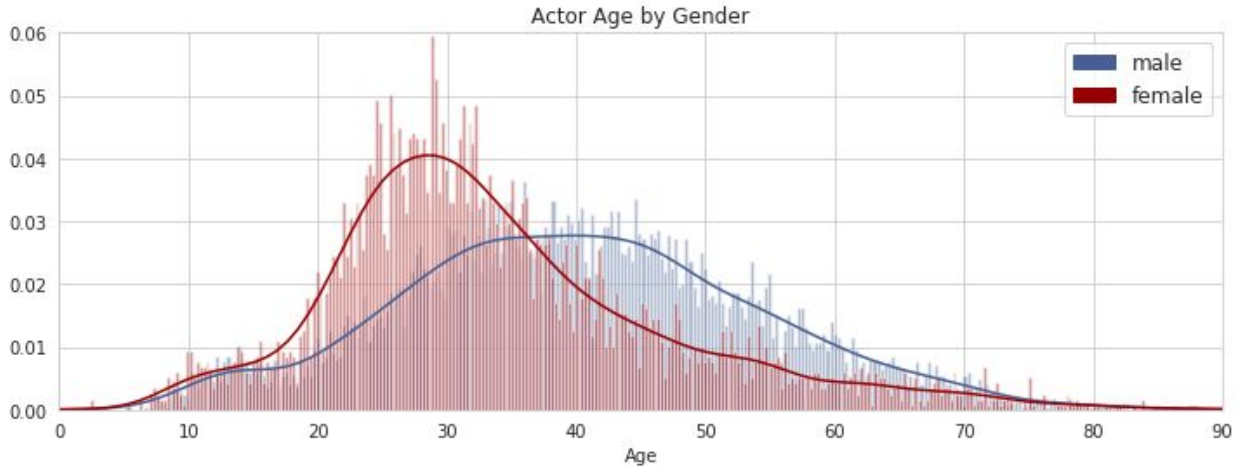
Capstone 1: Statistical Analysis

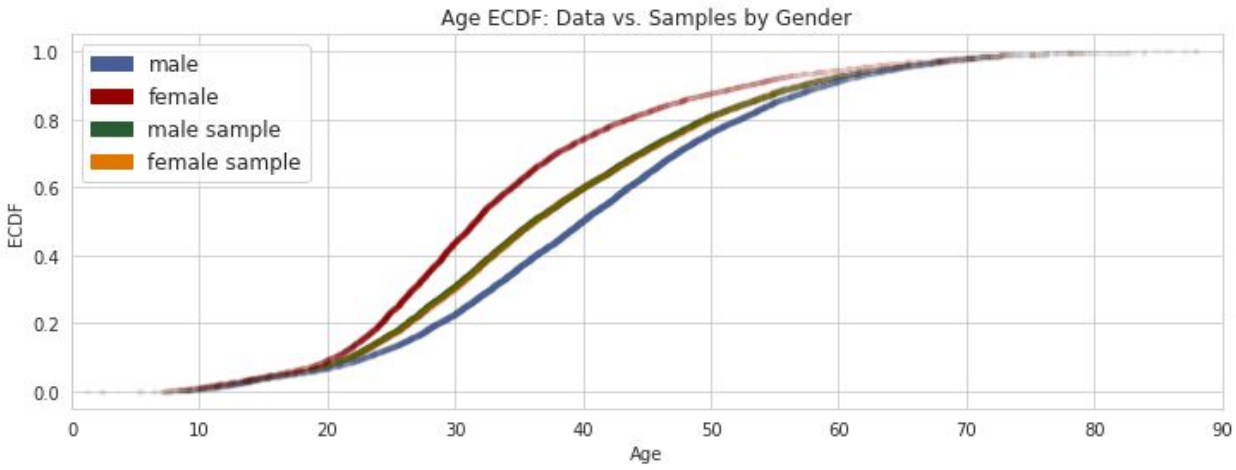
For the statistical analysis of the dataset for this project, I used several different techniques to perform hypothesis testing. I compared distributions of the mean between independent variables and between the target variable and one of the predictors. I applied hypothesis testing to evaluate a correlation coefficient, as well.

The first hypothesis test was performed to examine the difference between the distribution of ages with respect to male and female actors. After separating the data by gender, I observed their distributions.



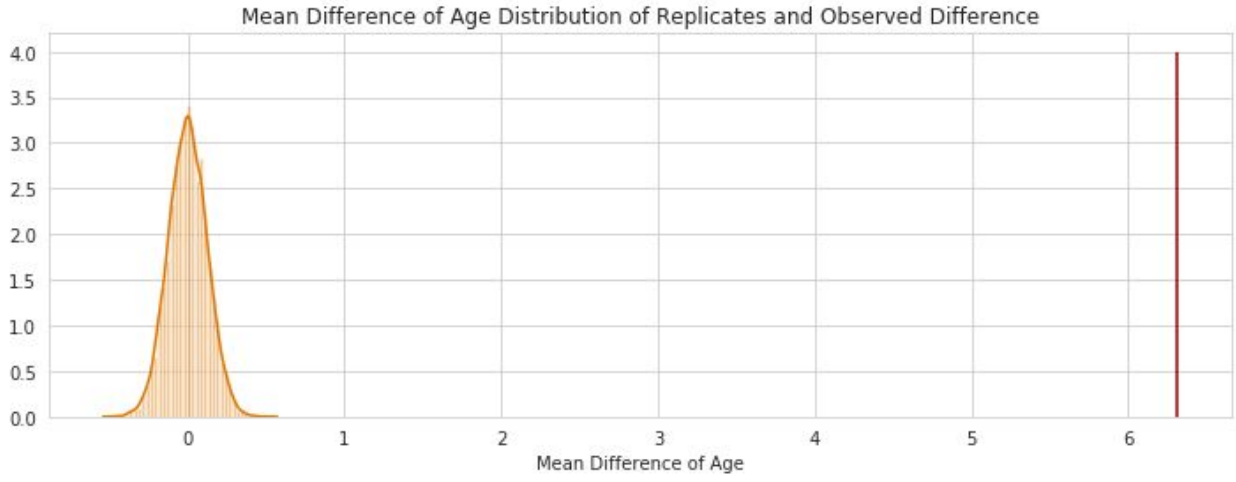
Observationally, it appeared that the average ages of female actors was younger than that of the male actors at the times of the actors' movie releases. Also, there seemed to be a short plateau for the teenage actors, indicating a lack of movie appearances for their age group. This plateau was more apparent for the male actors.

Next, I sought to break the structure of the individual distributions by combining them and reconstructing new samples of each gender through a random selection of the combined data. This selection was done without replacement, ensuring that each actor would be represented in one of the new datasets, while no actors would be observed more than once. Then, I wanted to see the likelihood of observing the statistical similarities between the original data and the new null distributions. First, I generated 10,000 permutation samples of each set that were randomly taken from the combined data. These new samples were of the same sizes as the original samples, respectively. Then, I plotted the Empirical Cumulative Distribution Functions (ECDF) for each and compared them to those of the observed data.



The curves showed a distinct gap between the genders of the observed data, while the curves of the samples overlapped. This indicated that the distributions between the genders were not the same.

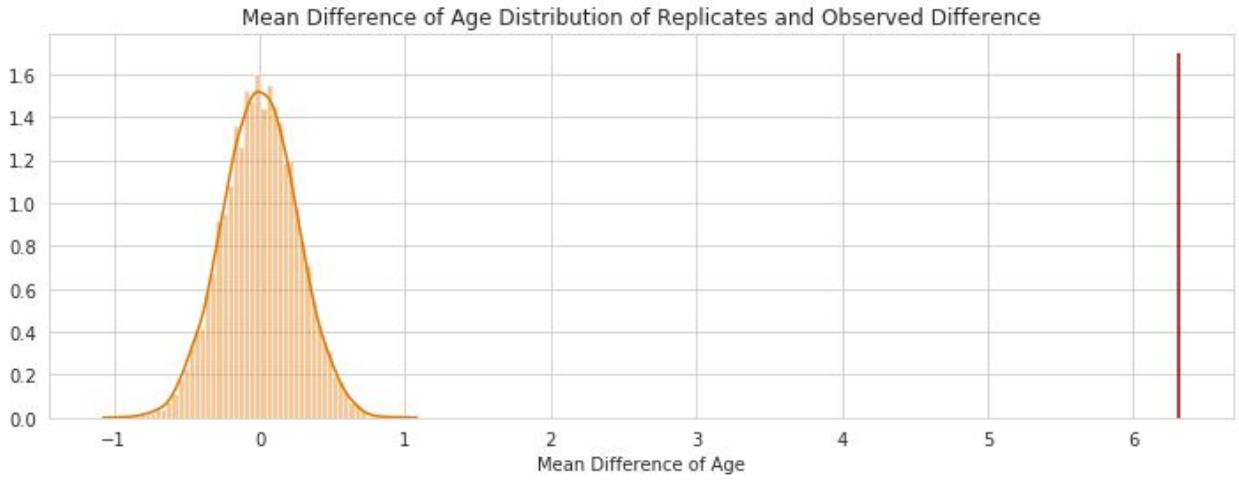
Next, I compared the difference of the mean age between the two genders with that of the 10,000 permutation replicates already generated. The mean age gap seen in the observed genders was 6.3 years. The distribution of the mean age difference of the randomly shuffled samples was plotted for comparison.



As was observed graphically, the mean age difference was found to be far beyond the distribution of values computed from the permutation samples. The p-value was calculated as zero. This meant that the probability was extremely low of observing a difference of the mean ages by chance as extreme as the one observed in the true data.

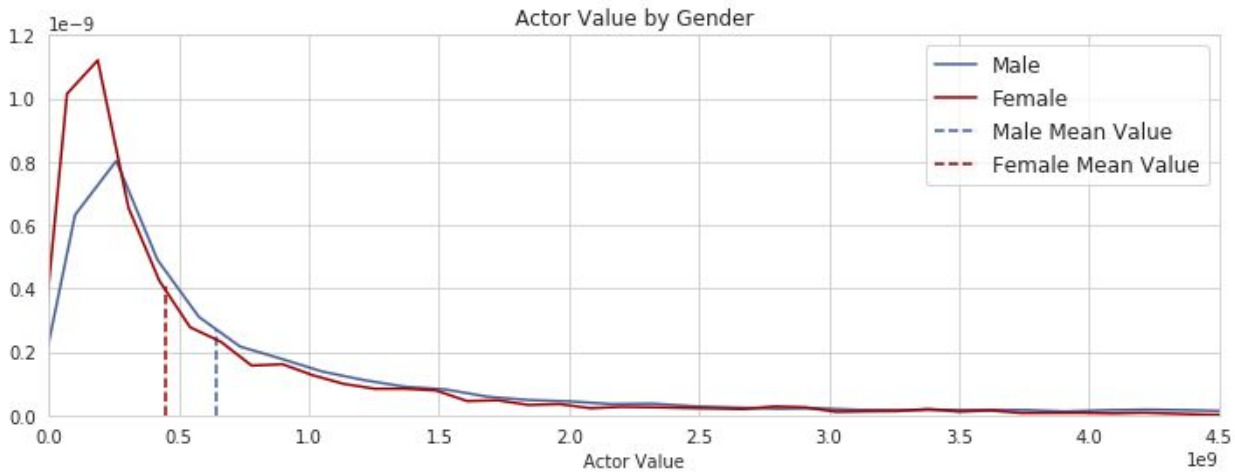
For the second hypothesis test, I examined the difference between the mean age of the two genders, without considering their distribution. First, I took the individual gender datasets and

shifted the ages of each actor within them to coincide with the mean age of all of the actors. Then, I created 10,000 randomly generated bootstrap replicates of each shifted gender dataset through selection with replacement. This type of selection allowed some actors to be observed multiple times in a new sample, while completely excluding other actors. By doing this, a collection of 10,000 new datasets for each gender was simulated, without the expense of collecting actual data. Then, I calculated the difference of the mean ages of the simulated datasets to compare their distribution to the mean of the original, unshifted data.



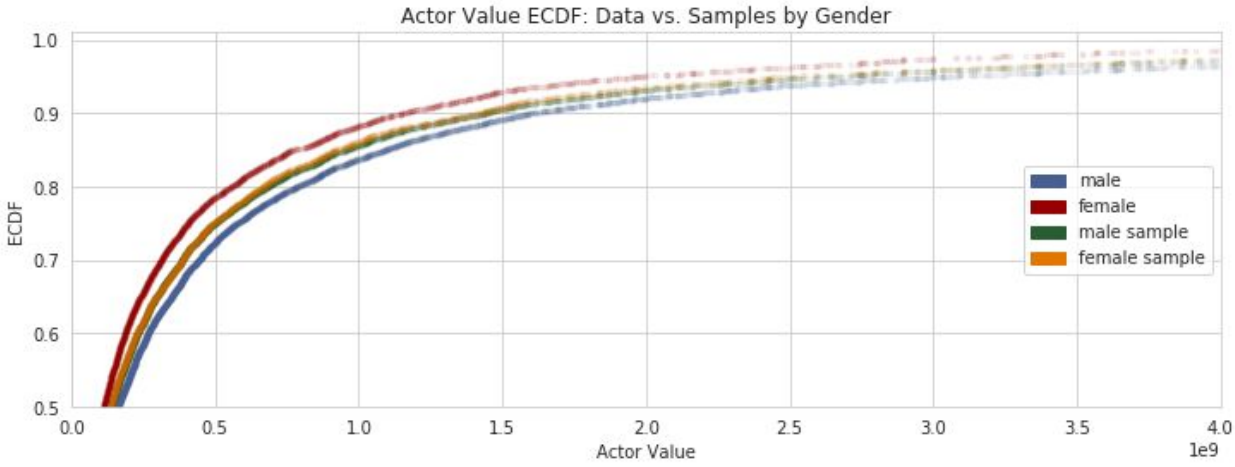
As would be expected, the variance of the difference between the new sample distributions was greater than it was when samples were created with replacement. Yet, the mean difference of age from the true data was still far away from the mean of the new distribution. The p-value was given as zero, once again. This confirmed that regardless of whether or not the two sets came from the same distribution, it was highly unlikely that the difference in mean age between the genders was due to chance.

The next hypothesis test focused on the target variable, actor value. This target represented the total of revenue generated by all of the films over each actor's career. The goal of this test was to determine if there was a significant difference between genders with respect to the target values. First, I plotted the target variable distributions by gender.



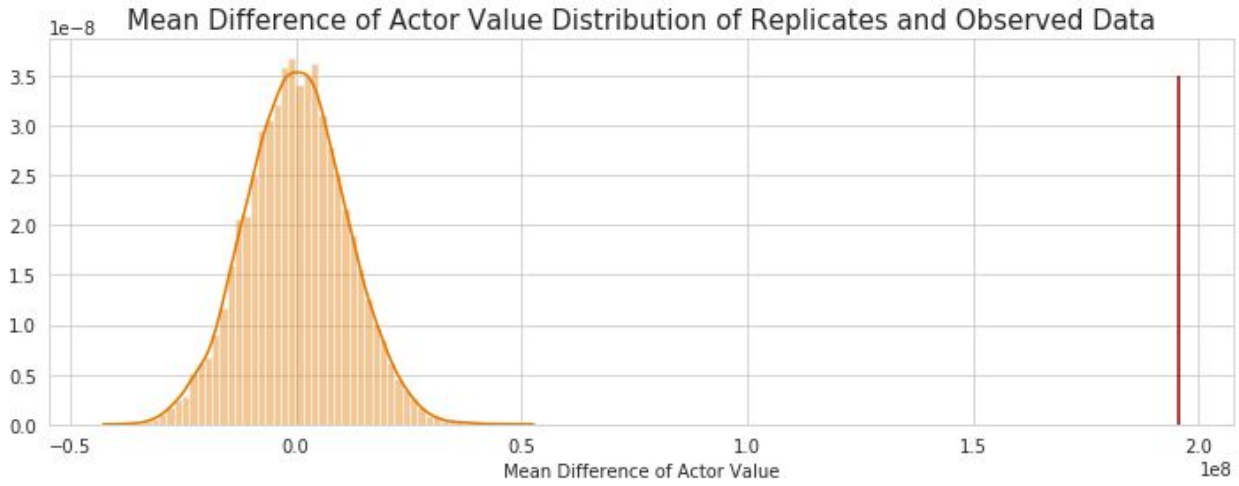
The distribution shapes appeared similar, although the female actors were clustered lower in the target values, while the distribution of the male actors peaked at a greater value.

Again, I drew 10,000 permutation samples for each gender and plotted all four ECDFs. I was looking for an obvious difference between the distribution of the original data and the distribution of the samples generated through combining and shuffling the two genders.



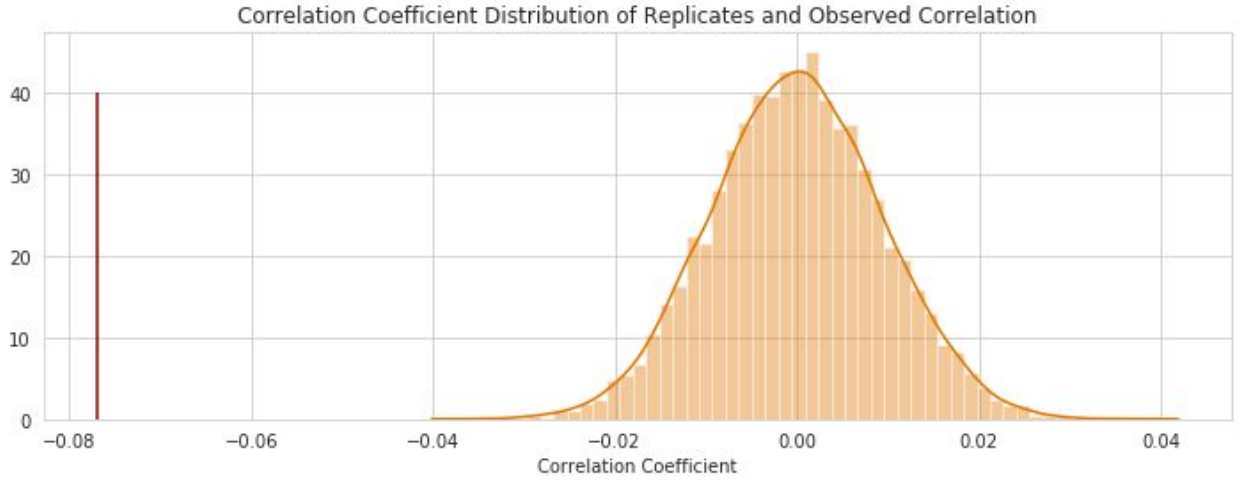
The two curves lined up tightly for the sample data, showing no gender bias when samples were obtained randomly. By observing the true data, it could be seen that a greater portion of the female actors were registered at lower target values than the proportion of male actors. Visually, the two original datasets appear distinctly distributed.

As before, I compared the distribution of the mean difference between the genders from the samples generated to the true average calculated from the data.



The true mean difference in career movie revenue was $195,749,566.18. As could be seen from the plot, this value was far from the distribution obtained from the permutation datasets. The p-value was calculated as zero, again. This meant that we could be extremely confident that there was a real difference between the distribution of the mean target values when comparing the genders in the original data.

Finally, I applied hypothesis testing to examine the correlation coefficient between the target variable and one of the predictor variables. Once again, I chose to examine gender. I permuted the values of each gender dataset and generated 10,000 permutation replicates of their correlation coefficients. This was the same technique I used when I analyzed the mean difference in ages using bootstrap replicates that were generated by selection with replacement.



The true Pearson correlation coefficient of -0.076862 was not in the neighborhood of those found in the sample distribution. The p-value for this test was zero. This meant that we could be very confident that the actor's gender and their career movie revenues were negatively correlated.