Take Home Challenge:

In this situation we are given 3 data sets relating to ticketing. One data frame has statistics about a given ticket such as who gave the ticket, age of the ticket receiver, etc. Another dataframe gives the ward that is related to the ticketing event. The last data frame gives the status of the ticketing event and how much the ticket is for. With this data we have to determine: The mean ticket price between men and women, if certain officers give large tickets than others, and if certain wards give larger tickets than others. We are then supposed to create a model that predicts the status of a ticket.

After cleaning up the data (the court data dataframe had many input errors), many of the questions can be solved with statistics. The tests used are T test and ANOVA. T test looks at the means, standard deviations, and number of samples between two groups and sees if there is a difference between the two groups. If the p value from the test is below 0.05 we can consider it significant. If it is not, then we can not say the two groups are statistically different. ANOVA is a similar test however it can be performed with multiple groups. The t-test between the ticket price means of men and women was 5.7E-13 which means there is a statistical difference between the two groups. Performing anova on the different officers and different wards failed to prove there is a statistical difference inside the respective groups.

For modeling, I converted the officers to dummies. I was going to perform the same to the wards however keeping the wards as is helped the model (not a great practice but still.)I converted men and women into a binary set and converted status into intergers just labeling them 1,2,3,4. I created new interaction terms but the strongest interaction term was between age and ticket amount. I created new features such as ‘large\_ticket’ which checked if the ticket size was abobe 100. The other features I created were related to age. I used a gradient boosting classifier because they have been successful for me in the past. I put it in a over vs rest classification and got 0.465 as the test score