I worked on the Kaggle Titanic Competition. The problem that we are exploring in this assignment is to find out the best model for predicting which passengers survived the sinking of the Titanic ship. The vessel had passengers of different genders, ages, means, familial structure, and origin and these are all potential features to use in the model.

My solution is the model the titanic survival for these features using the sex of the passengers, their class, and the amount they spent on fare. To do this, I used 3 different levels of grouping to find the conditions to best incorporate in the model. The sex of passenger was simply binary: male or female. This was the most distinctive of the features as a stark number of females survived in comparison to male (74% vs 19%). Therefore, this would make the strongest top level of grouping. The next level was their class aboard the vessel. This was next most significant figure as indicated by how the proportion of females surviving in class 1 and class 2 was over 90% while it was only about 50% in class 3. The final level was the fare each passenger spent. While this is highly correlated to the class the passenger was in, it still offered a small improvement over the sex – class model. The way this field was represented was to bucket the fare prices in 2 different sets. One for tickets under $20 and the other for tickets above $20. This was because females that spend over this amount inexplicably had a much lower chance of survival (less than 33%).

This model was about a 0.5% improvement over the baseline model that was sex based. Accuracy went up from around 77.5% to about 78%. Other models that were attempted on the training set included adding a further grouping level for the number of parents/children each passenger had and the origin dock of each passenger. In the first case, for example, it was found third-class females that paid above $20 did not have a lower than 50% chance of survival if they had no parents/children on board in the training set. However, this condition was not reflected in the test data set as it did not yield any accuracy improvement.

I used Excel in formulating these models since it made grouping and aggregating along different features/dimensions very simple. I appended the fare grouping of over and under $20 to discretize these values to just 0 and 1 (low and high). I then used pivot tables to group by sex, class, and price group and aggregate the average along these groupings. The items with a greater than 50% survival rate, I decided to model as survived and the items with less, I decided to model as not having survived. As mentioned, since I found that females had a lower than 50% chance of survival in third class if their tickets were above $20, I added this as a condition to the otherwise sex based model that outputs that all females survive and all males do not.