Let's now see how the baseline method used by D2Hawkeye would perform on this data set.

The baseline method would predict that the cost bucket for a patient in 2009 will be the same as it was in 2008.

So let's create a classification matrix to compute the accuracy for the baseline method on the test set.

So we'll use the table function, where the actual outcomes are ClaimsTest$bucket2009, and our predictions are ClaimsTest$bucket2008.

The accuracy is the sum of the diagonal, the observations that were classified correctly, divided by the total number of observations in our test set.

So we want to add up 110138 + 10721 + 2774 + 1539 + 104.

And we want to divide by the total number of observations in this table, or the number of rows in ClaimsTest.

So the accuracy of the baseline method is 0.68.

Now how about the penalty error?

To compute this, we need to first create a penalty matrix in R. Keep in mind that we'll put the actual outcomes on the left, and the predicted outcomes on the top.

So we'll call it PenaltyMatrix, which will be equal to a matrix object in R.

And then we need to give the numbers that should fill up the matrix: 0, 1, 2, 3, 4.

That'll be the first row.

And then 2, 0, 1, 2, 3.

That'll be the second row.

4, 2, 0, 1, 2 for the third row.

6, 4, 2, 0, 1 for the fourth row.

And finally, 8, 6, 4, 2, 0 for the fifth row.

And then after the parentheses, type a comma, and then byrow = TRUE, and then add nrow = 5.

Close the parentheses, and hit Enter.

So what did we just create?

Type PenaltyMatrix and hit Enter.

So with the previous command, we filled up our matrix row by row.

The actual outcomes are on the left, and the predicted outcomes are on the top.

So as we saw in the slides, the worst outcomes are when we predict a low cost bucket, but the actual outcome is a high cost bucket.

We still give ourselves a penalty when we predict a high cost bucket and it's actually a low cost bucket, but it's not as bad.

So now to compute the penalty error of the baseline method, we can multiply our classification matrix by the penalty matrix.

So go ahead and hit the Up arrow to get back to where you created the classification matrix with the table function.

And we're going to surround the entire table function by as.matrix to convert it to a matrix so that we can multiply it by our penalty matrix.

So now at the end, close the parentheses and then multiply by PenaltyMatrix and hit Enter.

So what this does is it takes each number in the classification matrix and multiplies it by the corresponding number in the penalty matrix.

So now to compute the penalty error, we just need to sum it up and divide by the number of observations in our test set.

So scroll up once, and then we'll just surround our entire previous command by the sum function.

And we'll divide by the number of rows in ClaimsTest and hit Enter.

So the penalty error for the baseline method is 0.74.

In the next video, our goal will be to create a CART model that has an accuracy higher than 68% and a penalty error lower than 0.74.