Now, as we start to think about building regression models with this data set, we need to consider the possibility that there is multicollinearity within the independent variables.

And there's a good reason to suspect that there would be multicollinearity amongst the variables, because in some sense, they're all measuring the same thing, which is how strong the Republican candidate is performing in the particular state.

So while normally, we would run the correlation function on the training set, in this case, it doesn't work.

It says, x must be numeric.

And if we go back and look at the structure of the training set, it jumps out why we're getting this issue.

It's because we're trying to take the correlations of the names of states, which doesn't make any sense.

So to compute the correlation, we're going to want to take the correlation amongst just the independent variables that we're going to be using to predict, and we can also add in the dependent variable to this correlation matrix.

So I'll take cor of the training set but just limit it to the independent variables-- Rasmussen, SurveyUSA, PropR, and DiffCount.

And then also, we'll add in the dependent variable, Republican.

So there we go.

We're seeing a lot of big values here.

For instance, SurveyUSA and Rasmussen are independent variables that have a correlation of 0.94, which is very, very large and something that would be concerning.

It means that probably combining them together isn't going to do much to produce a working regression model.

So let's first consider the case where we want to build a logistic regression model with just one variable.

So in this case, it stands to reason that the variable we'd want to add would be the one that is most highly correlated with the outcome, Republican.

So if we read the bottom row, which is the correlation of each variable to Republican, we see that PropR is probably the best candidate to include in our single-variable model, because it's so highly correlated, meaning it's going to do a good job of predicting the Republican status.

So let's build a model.

We can call it mod1.

So we'll call the glm function, predicting Republican, using PropR alone.

As always, we'll pass along the data to train with as our training set.

And because we have logistic regression, we need family = "binomial".

And we can take a look at this model using the summary function.

And we can see that it looks pretty nice in terms of its significance and the sign of the coefficients.

We have a lot of stars over here.

PropR is the proportion of the polls that said the Republican won.

We see that that has a very high coefficient in terms of predicting that the Republican will win in the state, which makes a lot of sense.

And we'll note down that the AIC measuring the strength of the model is 19.8.

So this seems like a very reasonable model.

Let's see how it does in terms of actually predicting the Republican outcome on the training set.

So first, we want to compute the predictions, the predicted probabilities that the Republican is going to win on the training set.

So we'll create a vector called pred1, prediction one, then we'll call the predict function.

We'll pass it our model one.

And we're not going to pass it newdata, because we're just making predictions on the training set right now.

We're not looking at test set predictions.

But we do need to pass it type = "response" to get probabilities out as the predictions.

And now, we want to see how well it's doing.

So if we used a threshold of 0.5, where we said if the probability is at least 1/2, we're going to predict Republican, otherwise, we'll predict Democrat.

Let's see how that would do on the training set.

So we'll want to use the table function and look at the training set Republican value against the logical of whether pred1 is greater than or equal to 0.5.

So here, the rows, as usual, are the outcome -- 1 is Republican, 0 is Democrat.

And the columns-- TRUE means that we predicted Republican, FALSE means we predicted Democrat.

So we see that on the training set, this model with one variable as a prediction makes four mistakes, which is just about the same as our smart baseline model.

So now, let's see if we can improve on this performance by adding in another variable.

So if we go back up to our correlations here, we're going to be searching, since there's so much multicollinearity, we might be searching for a pair of variables that has a relatively lower correlation with each other, because they might kind of work together to improve the prediction overall of the Republican outcome.

If two variables are highly, highly correlated, they're less likely to improve predictions together, since they're so similar in their correlation structure.

So it looks like, just looking at this top left four by four matrix, which is the correlations between all the independent variables, basically the least correlated pairs of variables are either Rasmussen and DiffCount, or SurveyUSA and DiffCount.

So the idea would be to try out one of these pairs in our two-variable model.

So we'll go ahead and try out SurveyUSA and DiffCount together in our second model.

So to save ourselves some typing, we can hit up a few times until we get to the model definition for model one.

And then we can just change the variables.

In this case, we're now using SurveyUSA plus DiffCount.

We'll also need to remember to change the name of our model from mod1 to mod2.

And now, just like before, we're going to want to compute out our predictions.

So we'll say pred2 is equal to the predict of our model 2, again, with type = "response", because we need to get those probabilities.

Again, we're not passing in newdata.

This is a training set prediction.

And finally, we can use the up arrows to see how our second model's predictions are doing at predicting the Republican outcome in the training set.

And we can see that we made one less mistake.

We made three mistakes instead of four on the training set-- so a little better than the smart baseline but nothing too impressive.

And the last thing we're going to want to do is to actually look at the model and see if it makes sense.

So we can run summary of our model two.

And we can see that there are some things that are pluses.

For instance, the AIC has a smaller value, which suggests a stronger model.

And the estimates have, again, the sign we would expect.

So SurveyUSA and DiffCount both have positive coefficients in predicting if the Republican wins the state, which makes sense.

But a weakness of this model is that neither of these variables has a significance of a star or better, which means that they are less significant statistically.

So there are definitely some strengths and weaknesses between the two-variable and the one-variable model.

We'll go ahead and use the two-variable model when we make our predictions on the testing set.