As usual, we will start by reading in our data and looking at it in the R console.

So we can create a data frame called polling using the read.csv function for our PollingData.csv file.

And we can take a look at its structure with the str command.

And what we can see is that as expected, we have a state and a year variable for each observation, as well as some polling data and the outcome variable, Republican.

So something we notice right off the bat is that even though there are 50 states and three election years, so we would expect 150 observations, we actually only have 145 observations in the data frame.

So using the table function, we can look at the breakdown of the polling data frame's Year variable.

And what we see is that while in the 2004 and 2008 elections, all 50 states have data reported, in 2012, only 45 of the 50 states have data.

And actually, what happened here is that pollsters were so sure about the five missing states that they didn't perform any polls in the months leading up to the 2012 election.

So since these states are particularly easy to predict, we feel pretty comfortable moving forward, making predictions just for the 45 remaining states.

So the second thing that we notice is that there are these NA values, which signify missing data.

So to get a handle on just how many values are missing, we can use our summary function on the polling data frame.

And what we see is that while for the majority of our variables, there's actually no missing data, we see that for the Rasmussen polling data and also for the SurveyUSA polling data, there are a decent number of missing values.

So let's take a look at just how we can handle this missing data.

There are a number of simple approaches to dealing with missing data.

One would be to delete observations that are missing at least one variable value.

Unfortunately, in this case, that would result in throwing away more than 50% of the observations.

And further, we want to be able to make predictions for all states, not just for the ones that report all of their variable values.

Another observation would be to remove the variables that have missing values, in this case, the Rasmussen and SurveyUSA variables.

However, we expect Rasmussen and SurveyUSA to be qualitatively different from aggregate variables, such as DiffCount and PropR, so we want to retain them in our data set.

A third approach would be to fill the missing data points with average values.

So for Rasmussen and SurveyUSA, the average value for a poll would be very close to zero across all the times with it reported, which is roughly a tie between the Democrat and Republican candidate.

However, if PropR is very close to one or zero, we would expect the Rasmussen or SurveyUSA values that are currently missing to be positive or negative, respectively.

This leads to a more complicated approach called multiple imputation in which we fill in the missing values based on the non-missing values for an observation.

So for instance, if the Rasmussen variable is reported and is very negative, then the missing SurveyUSA value would likely be filled in as a negative value as well.

Just like in the sample.split function, multiple runs of multiple imputation will in general result in different missing values being filled in based on the random seed that is set.

Although multiple imputation is in general a mathematically sophisticated approach, we can use it rather easily through pre-existing R libraries.

We will use the Multiple Imputation by Chained Equations, or mice package.

So just like we did in lecture with the ROCR package, we're going to install and then load a new package, the mice package.

So we run install.packages, and we pass it mice, which is the name of the package we want to install.

So you have to select a mirror near you for the installation, and hopefully everything will go smoothly and you'll get the package mice installed.

So after it's installed, we still need to load it so that we can actually use it, so we do that with the library command.

If you have to use it in the future, all you'll have to do is run library instead of installing and then running library.

So for our multiple imputation to be useful, we have to be able to find out the values of our missing variables without using the outcome of Republican.

So, what we're going to do here is we're going to limit our data frame to just the four polling related variables before we actually perform multiple imputation.

So we're going to create a new data frame called simple, and that's just going to be our original polling data frame limited to Rasmussen, SurveyUSA, PropR, and DiffCount.

We can take a look at the simple data frame using the summary command.

What we can see is that we haven't done anything fancy yet.

We still have our missing values.

All that's changed is now we have a smaller number of variables in total.

So again, multiple imputation, if you ran it twice, you would get different values that were filled in.

So, to make sure that everybody following along gets the same results from imputation, we're going to set the random seed to a value.

It doesn't really matter what value we pick, so we'll just pick my favorite number, 144.

And now we're ready to do imputation, which is just one line.

So we're going to create a new data frame called imputed, and we're going to use the function complete, called on the function mice, called on simple.

So the output here shows us that five rounds of imputation have been run, and now all of the variables have been filled in.

So there's no more missing values, and we can see that using the summary function on imputed.

So Rasmussen and SurveyUSA both have no more of those NA or missing values.

So the last step in this imputation process is to actually copy the Rasmussen and SurveyUSA variables back into our original polling data frame, which has all the variables for the problem.

And we can do that with two simple assignments.

So we'll just copy over to polling Rasmussen, the value from the imputed data frame, and then we'll do the same for the SurveyUSA variable.

And we'll use one final check using summary on the final polling data frame.

And as we can see, Rasmussen and SurveyUSA are no longer missing values.