In the previous video, we observed that Age and FrancePopulation are highly correlated.

But what is correlation?

Correlation measures the linear relationship between two variables and is a number between -1 and +1.

A correlation of +1 means a perfect positive linear relationship.

A correlation of -1 means a perfect negative linear relationship.

In the middle of these two extremes is a correlation of 0, which means that there is no linear relationship between the two variables.

When we say that two variables are highly correlated, we mean that the absolute value of the correlation is close to 1.

Let's look at some examples.

This plot graphs WinterRain on the x-axis and Price on the y-axis.

By visually inspecting the plot, it's hard to detect any linear relationship.

But it turns out that the correlation between WinterRain and Price is 0.14.

So there's a slight positive relationship between these two variables.

This plot has HarvestRain on the x-axis and the Average Growing Season Temperature on the y-axis.

It's again hard to visually see a linear relationship between the two variables, and it turns out that the correlation is -0.06, even closer to 0 than before.

This plot shows Age of the Wine on the x-axis and the Population of France on the y-axis.

We can easily see that there's a strong negative linear relationship between these two variables.

This makes sense, since the population of France has increased with time.

If we compute the correlation, we get -0.99.

So these two variables are indeed highly correlated.

Let's compute some correlations in R.

We can compute the correlation between a pair of variables in R by using the cor function.

Let's compute the correlation between WinterRain and Price.

So we type cor, and then in parentheses we give the names of the two variables, WinterRain and Price.

If we hit Enter, this function tells us that the correlation between the two variables is 0.137.

Let's look at another example.

This time we'll compute the correlation between Age and FrancePopulation.

So again, we use the cor function, but this time we give as the two variables Age and FrancePopulation.

As we saw earlier, Age and FrancePopulation are highly correlated with a correlation of -0.99.

We can also compute the correlation between all pairs of variables in our data set using the cor function.

To do so, we just type cor, and then our data set name wine.

The output is a grid of numbers with the rows and columns labeled with the variables in our data set.

To find the correlation between two variables, you just need to find the row for one of them and the column for the other.

For example, we can find the column for Age and then go down to the row for FrancePopulation to see the number that we just computed.

Or we could check if WinterRain is highly correlated with any other independent variables by looking at the WinterRain column.

So how does this information help us understand our linear regression model?

We've confirmed that Age and FrancePopulation are definitely highly correlated.

So we do have multicollinearity problems in our model that uses all of the available independent variables.

Keep in mind that multicollinearity refers to the situation when two independent variables are highly correlated.

A high correlation between an independent variable and the dependent variable is a good thing since we're trying to predict the dependent variable using the independent variables.

Now due to the possibility of multicollinearity, you always want to remove the insignificant variables one at a time.

Let's see what would have happened if we had removed both Age and FrancePopulation, since they were both insignificant in our model that used all of the independent variables.

We'll call this new model model5, and again, we'll use the lm function to predict Price using as independent variables AGST, HarvestRain, and WinterRain.

Again, we'll use the data set wine.

If we take a look at the summary of this new model and look at the Coefficients table, we can see that AverageGrowingSeasonTemperature and HarvestRain are very significant, and WinterRain is almost significant.

So this model looks pretty good, but if we look at our R-squared, we can see that it dropped to 0.75.

The model that includes Age has an R-squared of 0.83.

So if we had removed Age and FrancePopulation at the same time, we would have missed a significant variable, and the R-squared of our final model would have been lower.

So why didn't we keep FrancePopulation instead of Age?

Well, we expect Age to be significant.

Older wines are typically more expensive, so Age makes more intuitive sense in our model.

Multicollinearity reminds us that coefficients are only interpretable in the presence of other variables being used.

High correlations can even cause coefficients to have an unintuitive sign.

We'll see an example of this in the next lecture.

So we fixed the multicollinearity issue caused by Age and FrancePopulation.

Do we have any other highly-correlated independent variables?

There is no definitive cut-off value for what makes a correlation too high.

But typically, a correlation greater than 0.7 or less than -0.7 is cause for concern.

If you look back at all of the correlations we computed for our data set, you can see that it doesn't look like we have any other highly-correlated independent variables.

So we'll stick with model4 for the rest of this lecture, the model that uses AGST, HarvestRain, WinterRain, and Age as the independent variables.