1. Making Files in R Studio:
   * “Welcome to our tutorial on using R for data analysis. By the end of this series of tutorials, you will be able to use RStudio to determine the correlation coefficient for a set of data on fiber length and fleece weight.”
   * “The first thing we need to do is create a folder where we will save all of our work.” (create a folder for the sake of the video)
   * “Now we will launch RStudio and create a new R Markdown file where we will be doing our data analysis. To do this, choose the “file” option in the top ribbon, then hover over the “New File” option and choose “R Markdown…”. Give your file a title, for this tutorial we will call our file “Biology 207 Tutorial”, but we encourage you to choose a more descriptive name that describes what analysis we are performing. We will choose our output to be an html file, then hit create file. Once you have the file open, feel free to delete all of the sample code that RStudio puts in there by default.”
2. Code Chunk 1: Reading in and Viewing Data in R Studio
   * “First we will look at how you import and view a dataset”
   * \*\*Go to the excel sheet with our dataset\*\*
   * “This is our excel sheet with the data we are going to explore. We will save this sheet in the same folder that we created our r markdown file in. Then we will import the data into R Studio.”
   * \*\*Go back to R Studio\*\*
   * “Now that the excel spreadsheet is saved in the location we need it in, we will run some code to put it in R Studio. The first line of code you see (zoom in on line 9) imports a package that we will use to view excel spreadsheets in R Studio. When you run this line of code, R Studio will import that package into your packages library on the bottom right. Always make sure that the checkmark next to your package is checked so that R Studio can utilize that package’s function.”
   * “Now that the package is installed, we will import our excel file using the read\_excel package. We need to store the file as a variable for referencing purposes later on. In this case we chose to name the variable that stores the file “FiberLengthFleeceWeight”. The arrow pointing to this variable tells R Studio that we are reading our excel file and storing the data from the file in this variable. We could have also use an equals sign here.”
   * “Viewing the information in the excel file within R Studio can be done by using the “View” function. This can be accomplished by typing (line 10), which tells R Studio to run the view function on the variable storing the information in our excel spreadsheet.”
   * “The summary function will show us some basic statistical information about our data.”
3. Code Chunk 2

* “Next, we will be installing and loading the packages we will need to complete this analysis. Installing a package is similar to adding an extension to Google Chrome. It enhances the functionality and introduces new features to R.”
* “These next three lines of code will install the necessary packages.”
* “In order to use these packages, you must load them into R. The following three lines of code will do that.”
* “The ggplot2 package is going to be used to make graphs, the cowplot package is used for the aesthetic theme of the graph, and UsingR will allow us to run statistical tests.”

1. Code Chunk 3: Checking for Normal Distributions in R
   * “Now that we have our packages installed and our data loaded into R Studio, we will begin to explore the data. An import aspect of data analysis is making sure that our data meets certain criteria to qualify as being normally distributed.”
   * “The first test we will run is a simple exploratory data analysis of the data. This will show us whether the data in our dataset appears normally distributed, and will tell us if we need to make any alterations to the data. To do this we will run (line 24). The code inside of the parentheses tells R Studio to run a simple EDA on the dataset we loaded in and specifically to explore the ‘Fiber length’ variable. This is what the $ after the file name indicates.”
   * “Our eda does appear normally distributed. Given how small our sample size is, the easiest way to tell is by looking at the Normal Q-Q plot of our data. The data appears to follow the linear trend in this model very well. We will confirm that our dataset is normally distributed using a shapiro test.”
   * “Running the shapiro test will give us some statistics about the fiber length variable. What we are looking for is a p-value that is greater than 0.05 to tell if our data is normally distributed. The p-value in this case is 0.3822, which indicates to us that the data is indeed normally distributed.”
   * “We will now run the same test on the other variable, fleece weight. The Normal Q-Q plot of this variable appears normally distributed as well. A shapiro test confirms this. The p-value is 0.1833 which is above our level of significance of 0.05.”
2. Code Chunk 4:

* “Now, we will compile the data into a scatterplot. It is best to use a scatterplot to visualize this data, because we have two numerical variables.”
* “This first line of code tells R to plot the data points in the dataset.”
* “The next line tells R which is the dependent and independent variables.”
* Geom\_point tells R to plot points for the datapoints.
* Geom\_smooth tells R to plot a linear regression line with 95% confidence interval.
* “Theme cowplot is the cowplot package we installed earlier that simply tells R what aesthetic theme to use.”
* “Finally, xlab and ylab label the x and y axes.

6. Code Chunk 5

* “Finally, in order to find the correlation coefficient, we will need to fit a linear regression line to this data.”
* “The first line of code tells R to fit this linear regression. The tilde tells R that fleece weight depends on fiber length, which matches what we have designated as the x and y variables.”
* “Summary(fit) outputs the statistical information. It gives us a variety of data, however we are particularly interested in the Adjusted R-squared. This is our correlation coefficient. As denoted here, our correlation coefficient is 0.9408, which shows that there is a significant linear relationship between fiber length and fleece weight.”