# Prep Exercise (PE08) Lining up our models (Linear Modeling)

### General Instructions

1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.

General Directions:

* Before attempting the Prep Exercise, read through Chapter 16 of *An Introduction to Data Science* and execute the code provided within the readings to familiarize yourself with the code and output.
* Getting Started: The chapter on linear models (“Lining Up Our Models”) introduces linear predictive modeling using the tool known as multiple regression. The term “multiple regression” has an odd history, dating back to an early scientific observation of a phenomenon called “regression to the mean.” These days, multiple regression is just an interesting name for using a simple linear modeling technique to measure the connection between one or more predictor variables and an outcome variable. In this week’s homework, you will use an open data set to predict “spring fawn count” from the three other variables in the data set. The term “fawn” refers to a newly-born antelope.

### Prep Exercise

1. **In a paragraph or two, explain the concept of linear modeling and its uses within data analysis.**

Linear modeling is finding the best fit line for a set of data points that represents how a dependent variable varies based on an independent variable. X-axis has independent variable and Y-axis has dependent variable. The regression line can be represented by y=b+mx, where b is y-intercept and m is slope.

Linear modeling is helpful in data analysis. Given a dataset, linear model can be used to determine how one or more variable(s) is connected to another variable. It shows how an independent variable affects a dependent variable.

1. **Getting Ready: Read data from a URL directly into a data frame, the data at the end of the URL is an excel file.**
   1. Read in data from the following URL into a data frame  
      <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/excel/mlr01.xls>
   2. Use the ‘download.file’ command to download the excel spreadsheet, and store it in a temporary local file
   3. Use the read\_excel command from the readxl library (you might need to install and library readxl). Store the data in the dataframe ‘df’.
2. **Examine your dataframe using the View() and str() functions.**
   1. Verify that there are 4 columns within your dataframe, use the following URL to identify what each column represents. List the column names and their respective representations below.   
      <http://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html>

**Thunder Basin Antelope Study data for each year**

**The dataframe has 8 observations and 4 attributes which are of numeric type**

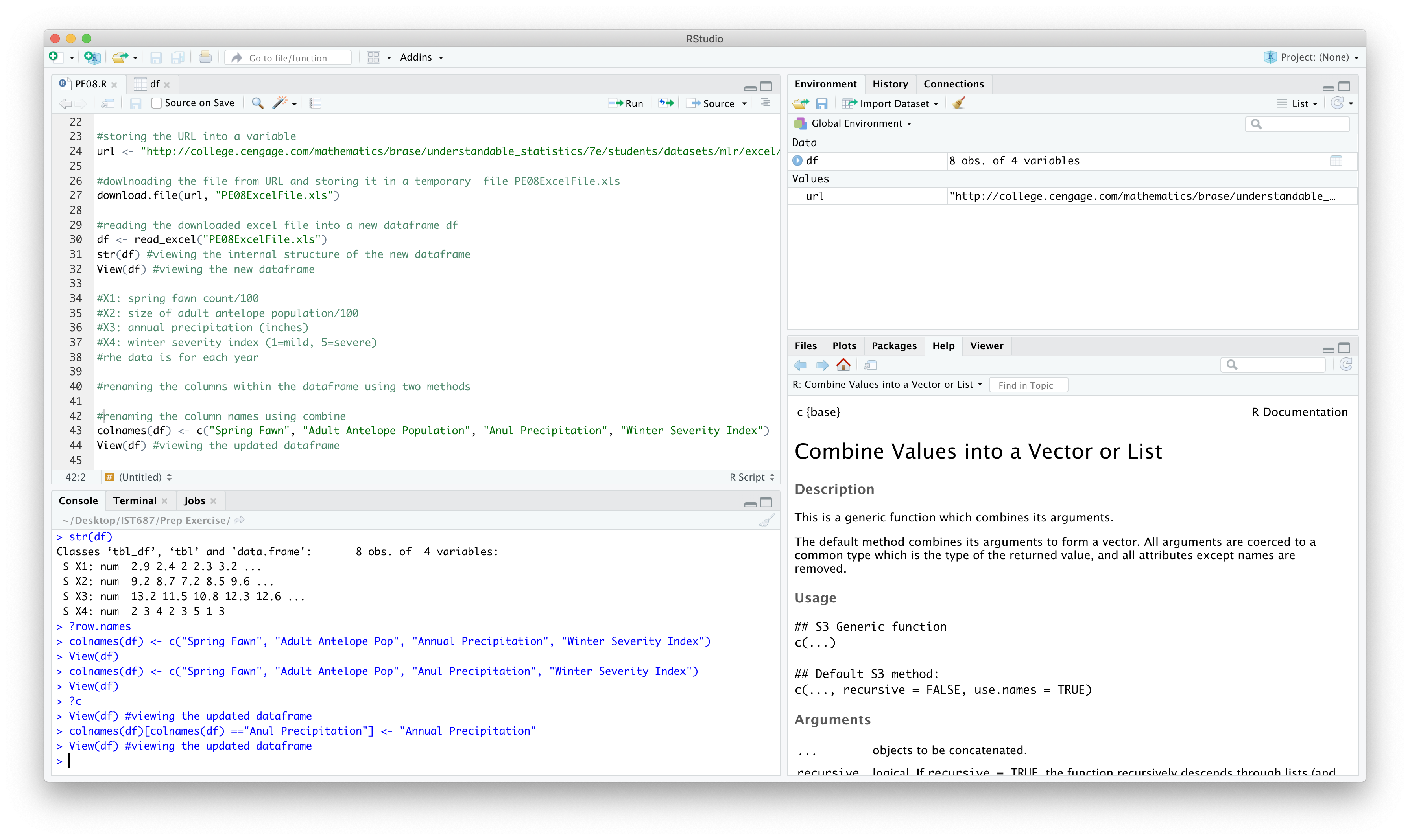
**X1: spring fawn count/100**

**X2: size of adult antelope population/100**

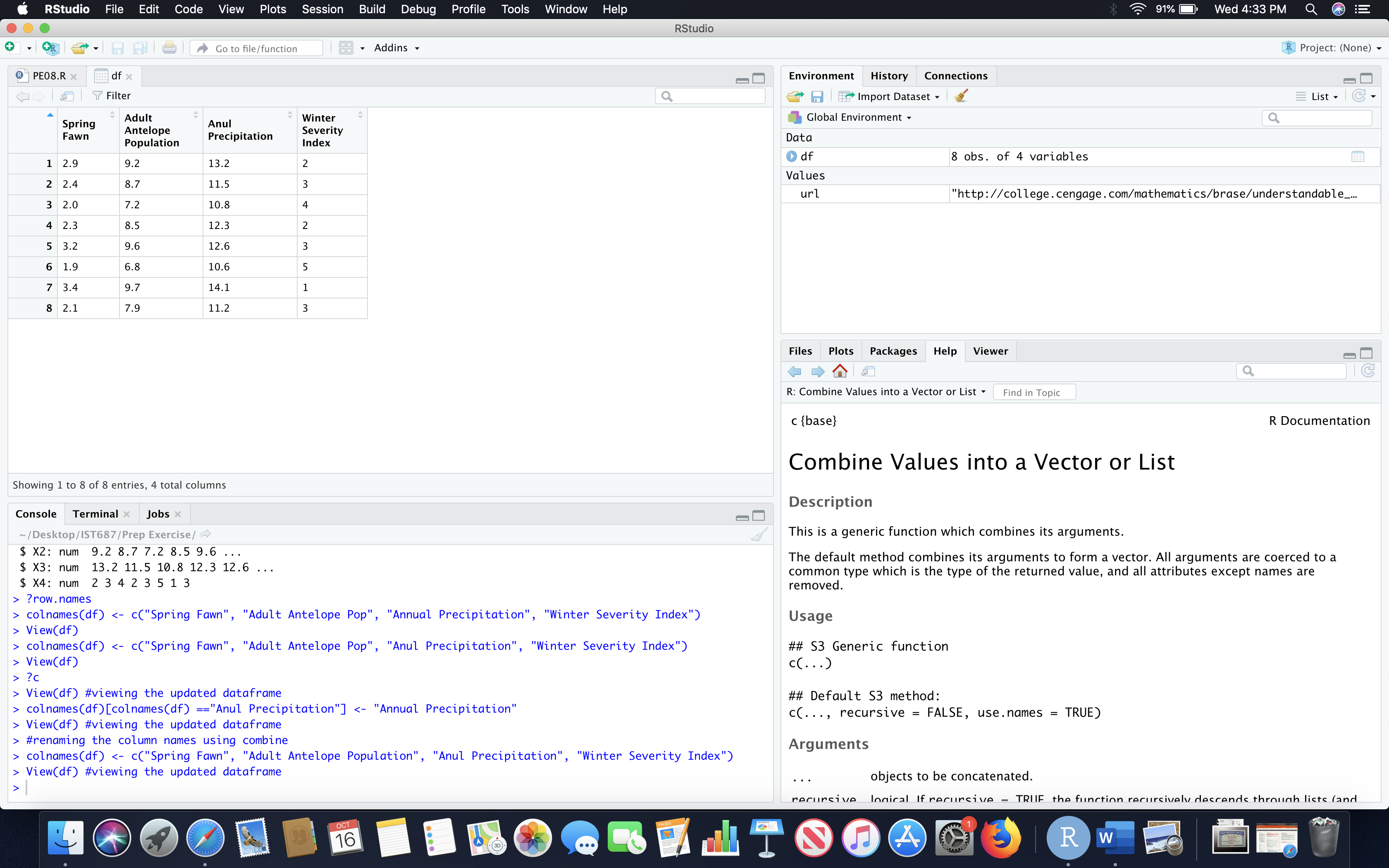
**X3: annual precipitation (inches)**

**X4: winter severity index (1=mild, 5=severe)**

1. **Renaming the columns within the dataframe using two methods.**
   1. Rename the columns within the dataframe by renaming all the columns at once. Purposely misspell the third column name, we will fix this later. If you are stuck, refer to earlier PEs and Homeworks. Place a screen shot below that shows the code that you used for this step.

****

* 1. Execute the View(df) command and paste a screenshot of the results below.

****

* 1. Rename the third column that was misspelled in the prior step by executing the code below. Explain the differences between the second approach and the first (replacing the “misspelled column name” and “correct column name” with the appropriate strings. What are the benefits of specifying the column name to be changed?

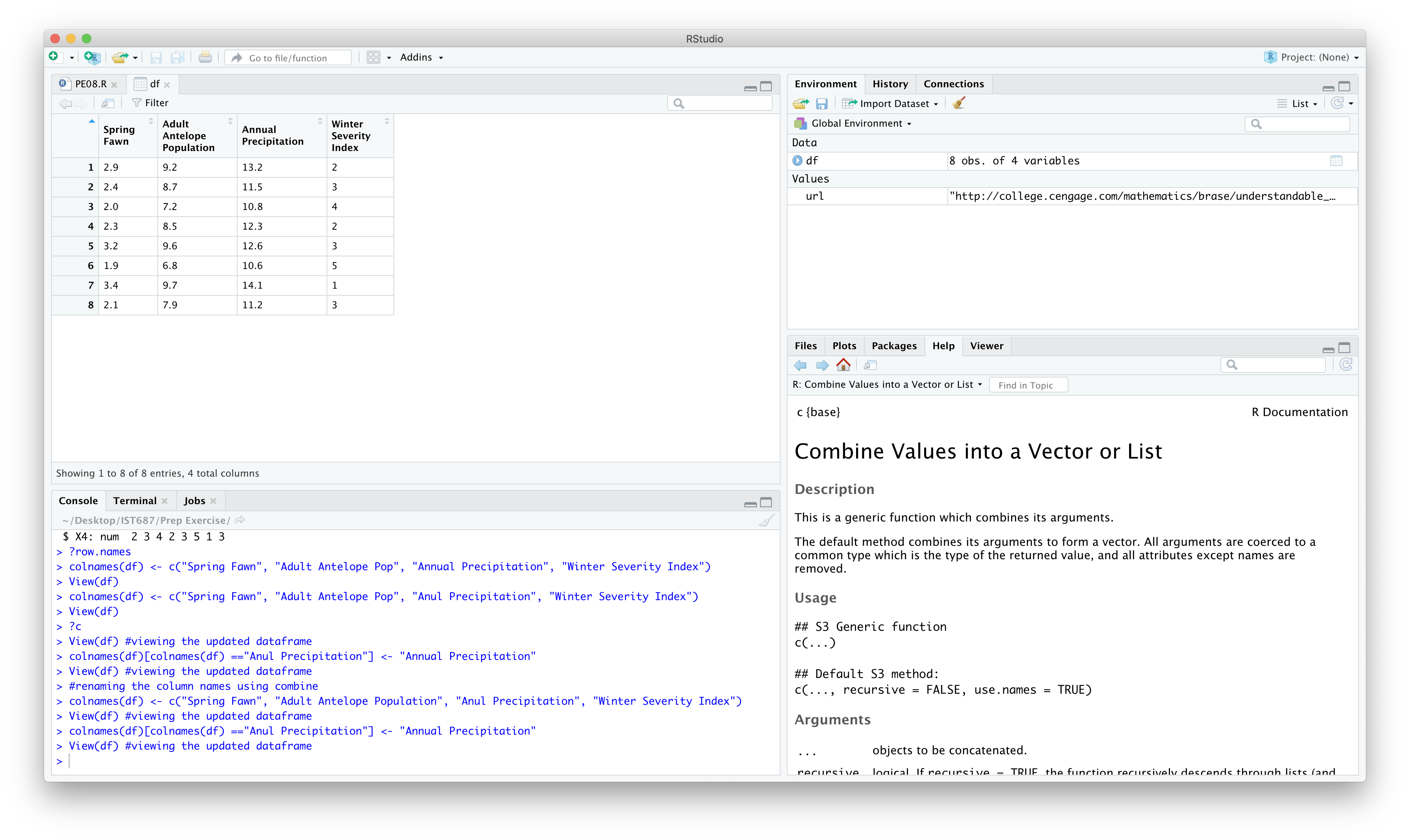
**colnames(df)[colnames(df)=="misspelled column name"] <- "correct column name"**

**colnames(df)[colnames(df)=="Anul Precipitation"] <- "Annual Precipitation"**

**Consider the example above for the second approach. Within the df dataframe, it fetches the column with name “Anul Precipitation” and renames the column as “Annual Precipitation”. Whereas in the first approach, a set of strings are combined and assigned to the column names of df dataframe. The drawback of this approach is that if the number of arguments within the combine command is less than the number of columns in the dataframe, then the latter columns will not have column name.**

**The benefits of specifying the column name to be changed is that we will know which column we are renaming and avoid having a column unnamed.**

1. **Verify that your changes are present within the dataframe using the View command.**
   1. Attach a screenshot of your correctly named dataframe below.

****

1. **Define a bivariate plot (i.e., explain what is a bivariate plot) and then explain how you would create a bivariate plot within R.**

Bivariate plot shows the relation between two variables within a dataset. A bivariate plot can be created in R using the plot() function. This function takes a minimum of two arguments.

plot(x, y)

where x is the coordinate values for x-axis,

y is the coordinate values for y-axis

1. **Explain the significance of the R-Squared value of a regression model.**

R-Squared value of a regression model determines how well independent variable(s) can predict the value of dependent variable. R-Squared value of 1 means that dependent variable is perfectly predicted by independent variable(s), while a value of 0 means independent variable(s) do not really predict the value of dependent variable.

1. **List any additional resources you used here.**
2. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*