Simple Linear Regression and Correlation

In this assignment you will complete a variety of tasks related to correlation and simple linear regression.

Good habits I strongly recommend creating a new RStudio project for every assignment and for each lecture as you follow-along. Using a good directory structure will make it much easier for you to find your work later.

Libraries: For this assignment you will need the following libraries: tidyverse, tidymodels, GGally, and lmtest.

Read-in the "airquality" data set (a default R dataset) as a dataframe called "air". To do this use the code below:

```
air = airquality
```

Details concerning this dataset can be found here: http://rpubs.com/Nitika/linearRegression_Airquality.

Question 1 How many rows are in the "air" dataframe?

Question 2 How many columns are in the "air" dataframe?

Question 3 True/False: There is missing data in "Ozone" variable in the dataframe.

Question 4 Which variable is most likely to be the response (Y) variable?

A. Ozone
B. Solar.R
C. Wind
D. Temp
E. Month
F. Day

We have three approaches that we can typically select from to deal with missing data:

1. Delete the rows with missing data
2. Delete the columns with missing data
3. Impute (i.e., estimate or guess) values to replace the missing values

Here we'll choose to delete rows with any missing data. Use the code below to apply the "drop_na" function to the "air" dataframe. The resulting dataframe will be called "air2". You will use this dataframe for the remainder of the assignment.

```
air2 = air %>% drop_na()
```

Question 5 How many rows remain in this new (air2) data frame?

Question 6 How many columns remain in this new (air2) data frame?

Use the "ggpairs" function to develop a visualization of the relationships in this dataset and to show correlation values for the combinations of variables.

Then use the "ggcorr" function to develop a correlation matrix for the variables. Hint: Use "label = TRUE" in the "ggcorr" function to show the correlation values.

Question 7 Which variable is most strongly correlated with the "Ozone" variable?

A. Solar.R
B. Wind
C. Temp
D. Month
E. Day

Question 8 Which variable is least strongly correlated with the "Ozone" variable?

A. Solar.R
B. Wind
C. Temp
D. Month
E. Day

Question 9 Plot "Temp" (x axis) versus "Ozone" (y axis) using the "ggplot" function. Choose an appropriate chart type. Which statement best describes the relationship between "Temp" and "Ozone"?

A. As Temp increases, Ozone decreases
B. As Temp increases there is no noticeable change in Ozone
C. As Temp increases, Ozone increases

Use Tidymodels to create a linear regression model using "Temp" to predict "Ozone". You miss wish to call your model fit "lm_fit".

Question 10 What is the slope of this regression model (to four decimal places)?

Question 11 what is the R-squared value of this model (not Adjusted R-squared) (to three decimal places)?

Question 12 Is the "Temp" variables significant in the model?

Question 13 Use the code below to generate 95% confidence intervals for the coefficients. Note that you may need to change "lm_fit" to the name of your model fit if you used a different name.

True/False: A 95% confidence interval for the slope coefficient does not contain zero.

```
confint(lm_fit$fit$fit$fit)
```

Question 14: Using your linear regression model with "Temp" to predict "Ozone", what is the predicted "Ozone" value when "Temp" is equal to 80 (to two decimal places)?

Question 15 Perform appropriate model diagnostics to verify whether or not the model appears to meet the four linear regression model assumptions.

True/False: There is no evidence of non-independent (autocorrelated) residuals.