Lab 9 (95 pts): Linear Regression
Objective: Creating Scatterplots, Calculating Correlation, and Determining the Least-Squares Regression Lines, Check Assumptions, Perform Inference

A. (95 points) the relationship between actual elapsed time and distance (Data Set: cleaned airline2008Nov)
We expect that there should be a strong relationship between the actual elapsed time and the distance. In this lab, we want to quantify this relationship. Because there are no distances more than 3000 miles in the continental United States, we will only consider flights that are 3000 miles or less. In addition, we also need to do a transformation for the data to be normal. This is complicated by the fact that we also need to have linear data. Therefore we have to transform both X and Y in the following fashion:

\[ Y' = \frac{Y^{0.3} - 1}{0.3}, X' = \frac{X^{0.9} - 1}{0.9} \]

These transformations are called BoxCox transformations. To take a power in R, you use a carat. To take a power in SAS, you use a double asterisk.

<table>
<thead>
<tr>
<th>Function</th>
<th>R</th>
<th>SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>z^2</td>
<td>z^2</td>
<td>z**2</td>
</tr>
</tbody>
</table>

SAS comments:

```
proc corr data = dataset;
   var first second;
run;
```

```
proc reg data = dataset plots(maxpoints=82000);
   ...;
```

Some of the following questions may be done by hand. If done by hand, all work needs to be shown.

1. (10 pts.) Code. Remember to create the restricted data set which consists of distances less than or equal to 3000 miles and to transform the variables.

2. (5 pts) Make a scatterplot of the data with the transformed distance on the x axis and the transformed actual elapsed time on the y axis. Please include the linear regression line on the plot.

3. (5 pts) From the scatterplot in part (2), describe the form, direction, and strength of the relationship. Identify any outliers. Is the relationship approximately linear?

4. (5 pts) Find the correlation between the transformed actual elapsed time (Y) and the transformed distance (X). Are your conclusions about the strength the same in this part as in part (3)? If they are different, provide a possible explanation for the difference.
5. (5 pts) Look at the scatterplot for these data that you made in part (2). Is the correlation a good numerical summary of the graphical display in the scatterplot? Please explain by discussing the reasons why correlation can or cannot be used.

6. (6 pts) Obtain the equation of the least-squares regression line for predicting the transformed actual elapsed time from the transformed distance. What is $r^2$ for these data?

7. (5 pts) Predict the transformed actual elapsed time when the transformed distance is 297 and calculate the residual. This part may be done by hand.
   Addendum: For 5 pts. Calculated the predicted value.
   Bonus 5: pts. Determine the residual: Hint: You will have to use the untransformed data set with Distance = 500 miles. Use the data point on 11/2 with a departure time of 13:47 and an arrival time of 15:53.

8. (5 pts) Obtain the residuals and plot them versus the transformed distance. There is no need to have a listing of the residuals. Is there anything unusual to report? If so, explain. Are the conclusions from the residual plot the same as from the scatterplot (parts 3 and 5)? If they are different, provide a possible explanation for the difference.

9. (5 pts) Do the residuals appear to be approximately Normal? Explain your answer. Be sure to include the appropriate graph(s) in your answer.

10. (5 pts) Based on your answers to parts, (2), (8), and (9), do the assumptions for the linear regression analysis appear reasonable? Explain your answer.

11. (12 pts) Construct and interpret a 99% confidence interval for the slope and the intercept. What is the significance of the result for the slope? Is the inference on the intercept of interest in this problem? Why or why not?

12. (10 pts) Is there significant evidence that transformed distance is associated with the transformed actual elapsed time at a 0.01 significance level? Please perform the 4-step process (state hypotheses, give a test statistic and $P$-value, and state your conclusion).

13. (6 pts.) How are the results from parts 11 and 12 similar? How are they different?

14. (11 pts.) Write a short paragraph in complete English sentences summarizes the results which is understandable to non-statisticians. The summary should contain the following parts: a) is the model appropriate to use, b) What is the relationship between the distance and the actual elapsed time? c) Is this situation good for prediction? d) Is there any causality in this situation? e) Can you generalize this situation to November of 2015? f) Provide a justification for not including distances over 3000 miles which does not include anything concerning making the assumptions valid.