Lab 9 (95 pts + 5 pts. BONUS): 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

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 unexpected long delays can skew this relationship, we want to consider only the flights with an Arrival Delay of less than 60 minutes and a combined TaxiIn plus TaxiOut of less than 60 minutes. In addition, we only wish to consider flights within the continental United States. The longest flight in our data set that meets that requirement is a 2,704 mile flight from Boston to San Francisco, so we will only consider flights that are 2,704 miles or less. (There will still be some flights left in the data set that are shorter than that but still outside the continental US, but we won't worry about that.)

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 2704 miles, with Arrival Delays of less than 60 minutes, and with combined TaxiIn plus TaxiOut times of less than 60 minutes.
2. (5 pts) Make a scatterplot of the data with the distance on the *x* axis and the 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 actual elapsed time (Y) and the 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 actual elapsed time from the distance. What is r^2 for these data?
7. (5 pts) Predict the actual elapsed time when the distance is 297, and calculate the residual. This part may be done by hand.

   Bonus: 5 pts.: Determine the residual. Use the data point on 11/7 from MSP to MKE with a departure time of 15:29.
8. (5 pts) Obtain the residuals and plot them versus the 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 distance is associated with actual elapsed time at a 0.01 significance level? Please perform the 4-step process (identify the variables, 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 summarizing 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) In addition, provide a justification for not including distances over 2704 miles which does not include anything concerning making the assumptions valid.