Lab 9 (100 points + 20 points BONUS): Linear Regression Objective: Creating Scatterplots, Calculating Correlation, Determining the Least-Squares Regression Lines, Checking Assumptions, and Performing Inference

A. (10 points) Online Prelab

B. (90 points) The relationship between the amount of Average Test Score (TestScore) and Median Income of each County (MedianIncome). In this lab, we want to explore the relationship between the county average test score and median income. In particular, we are interested in the relationship that holds for typical counties, excluding those that are very wealthy or very poor. Due to the phenomenon of "diminishing returns," the full range of data is better modeled by a curve -- not a straight line -- which is outside the scope of STAT 350. There are many ways we could choose to restrict the data to "typical counties," but we will proceed by taking only the values of MedianIncome that lie in the interval $(m - 2s, m + 2s)$, where m is the sample average and s is the sample standard deviation. Please note the interval is an *open interval,* meaning the values $m - 2s$ and $m + 2s$ will be excluded as well. *You will need to restrict the dataset to ensure you obtain the correct answers.*

1. (5 points) Code. **Please clearly indicate the code you used to restrict the dataset, and use the restricted dataset throughout the lab. If you have to manually put in the numbers for the average and standard deviation of the MedianIncome in the code, please provide the relevant output in this question.**
2. (5 points) Make a scatterplot of the data with MedianIncome on the x-axis and TestScore on the y-axis. Please include the linear regression line on the plot.
3. (5 points) From the scatterplot in part (2), describe the form, direction, and strength of the relationship. Also, identify any outliers and influential points. Is the relationship approximately linear?
4. (5 points) Find the correlation between MedianIncome and TestScore. 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 points) Look at the scatterplot for these data that you made in part (2). Is the correlation a good numerical summary of the pattern in the scatterplot? Please explain by discussing the reasons why in general correlation can or cannot be used to describe relationships.
6. (5 points) Obtain and report the estimated equation of the linear regression line for predicting TestScore from the MedianIncome. Also report R^2.
7. (5 points) Plot the residuals (Y) versus MedianIncome (X). Is there anything unusual to report? If so, please explain. Are the conclusions from the residual plot the same as from the scatterplot in part (3)? If they are different, provide a possible explanation for the difference.
8. (6 points) Do the residuals appear to be approximately Normal? Explain your answer. Be sure to include the appropriate graph(s) in your answer.

9. (6 points) Based on your answers to parts (3), (7), and (8), are the assumptions for the linear regression analysis reasonable? Explain your answer.
10. (12 points) Construct and interpret a 99% confidence interval for the slope and the intercept. What does the result of the slope mean (what is usually being tested)? Is the inference on the intercept of interest in this situation? Why or why not?
11. (15 points) Is there significant evidence that MedianIncome is associated with TestScore at a 0.01 significance level? Please perform the four-step hypothesis test.
12. (6 points) How are the results from parts (10) and (11) similar? How are they different?
13. (10 points) Write a short paragraph in complete English sentences summarizing the results. The summary should contain the following parts: a) Is the model appropriate to use? b) What is the relationship between MedianIncome and TestScore? c) Is there any causality in this situation? If so, what causes what? Please explain your answer.
14. (20 points) BONUS. Calculate and interpret the 99% *confidence interval for the mean response* and *99% prediction interval*, both at MedianIncome (x) = 30,000.
    a) (5 points) Code. Even if the code is given in part (1), please repeat it here.
    b) (5 points) Find and interpret the confidence interval for the mean response for y at MedianIncome = 30,000.
    c) (5 points) Find and interpret the prediction interval at MedianIncome = 30,000.
    d) (5 points) In general, how is the confidence interval for the mean response different from the prediction interval? Use the results of b) and c) to justify your answer.