MTH 325, Homework #9, Spring 2023 Name ______________________________________

Instructions: Show work or attach R code used to perform calculations (or any other technology used). Be sure to answer all parts of each problem as completely as possible, and attach work to this cover sheet with a staple.

1. Consider the data in **325homework9data.xlsx**. This is (fictitious) health data modelling blood pressure percentiles.
   a. Identify each of the out-of-bounds values in each of the independent variables. (Note: Person is not a variable. Blood Pressure is the dependent variable.) Replace these with null values.
   b. How many variables are missing in the whole data set? Would it be safe to remove them? Why or why not?
   c. Create a correlation table of the variables. Which variable does the best job predicting blood pressure if the null values are removed from that one variable?
   d. Find the equation of a simple linear regression model predicting blood pressure. Plot the data and the regression line.
   e. Create a column in your data labeling the row as missing or not.
   f. Impute the missing measurement in your independent variable using:
      i. Mean
      ii. Median
      iii. Regress from another unused variable
   g. Plot each imputation scheme on a graph and create the corresponding regression model using the original and imputed data.
   h. Compare the results of each method to dropping the missing data. Do you notice any potential issues? Discuss the possible biases that could creep into any/each of these methods.

2. The data in **325homework9data2.xlsx** is data on ice sheet dynamics in Greenland derived from satellite data. The data comprises an irregular time series. The variable of interest is the last column: delta H dynamic.
   a. Select a method of modeling the data based on date that we've discussed in class. Since the time series is irregular, it will need to be some sort of regression model.
   b. Use the model you develop to impute (create) a regular time series for the data. Use as much of the original data as you can, and use the model for the remaining values.
   c. Now use the rules of regular time series to analyze the data. Seasonal trends have been removed, but you may still want to run it through a seasonal decomposition package. How does the trend you find differ from the model you developed for the original time series? If at all?
   d. What if you impute the "missing" values with a linear trend between consecutive observations? How does this impact the model you end up developing?