Model Validation

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

Before beginning the assignment tasks, read-in the "bike_cleaned.csv" file into a data frame called "bike". This is the same data that you used in the Module 2 Multiple Linear Regression and Special Issues assignment. As we did in that assignment you should convert "dteday" from a character variable to a date variable. Convert the remaining character variables to factors. You can do this one variable at a time or use a "mutate_if". Finally, convert the "hr" variable into a factor.

Question 1: Split the data into training and testing sets. Your training set should have 70% of the data. Use a random number (set.seed) of 1234. Your split should be stratified by the "count" variable.

How many rows of data are in the training set? I know it's probably a bit annoying to keep answering this question about the number of rows, but it's helpful to be able to validate that your split code is correct before proceeding :)

Question 2 Stratifying the split by the "count" variable serves what purpose?

A. Stratifying by "count" ensures that unusual values for "count" are eliminated
B. Stratifying by "count" ensures that "count" is similarly represented in both the training and testing sets
C. Stratifying by "count" ensures that the training set contains the "count" variable
D. None of the above

Question 3: Build a linear regression model (using the training set) to predict "count" using the variables "season", "mnth", "hr", "holiday", and "weekday", "temp", and "weathersit".

What is the adjusted R-squared value (to four digits) of the resulting model?

Question 4: Use the predict functions to make predictions (using your model from Question 3) on the *training* set. **Hint: Be sure to store the predictions in an object, perhaps named "predict_train" or similar.** Develop a histogram of the predictions (Hint: The predictions are likely stored in a variable called ".pred" in your predictions object).

Select the statements below that are likely true about the distribution of predictions?

A. The maximum number of rides predicted for an hour is around 600 B. The average number of rides predicted per hour is around 450
C. Some predictions for the number of rides in an hour are negative
D. None of these statements are true

Question 5: Determine the performance of your model on the testing set.

What is the R-squared value (to four decimal places) of your model on the testing set? REMINDER: DO NOT build a model on the testing set. Use your model that was developed on the training set.