**MSDS 6306: Doing Data Science**

**Live session Unit 9 assignment**

**Due: Monday, July 8 at 11:59pm**

**Submission**

**ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT.** Use RMarkdown to create this file. Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, **please enable {r, echo=TRUE} so your code is visible.**

**Questions**

**Background:** Brewmeisters in Colorado and Texas have teamed up to analyze the relationship between ABV and IBU in each of their states. Use the data sets from the project to help them in their analysis. There three main questions of interest are 1) Is there a significant linear relationship between ABV (response) and IBU (explanatory), 2) Is this relationship different between beers in Colorado and Texas and 3) Is there a significant quadratic component in this relationship for either Colorado or Texas or both?

**A. Clean an prepare the data:**

1. Create column for brewery ID that is common to both datasets similar to what you did in the project. So we can merge!

2. Merge the beer and brewery data into a single dataframe.

3. Clean the State Column … get rid of extraneous white space.

4. Create One Dataset that has only Colorado and Texas beers and no IBU NAs … name it “beerCOTX”

5. Order beerCOTX by IBU (ascending) ... this will be important later in graphing

**B. Create an initial plots of the data**

6. Plot ABV v. IBU for both Colorado and Texas (two separate plots) … use ggplot and facets.

**C. Model the data**

7. For each state, fit a simple linear regression model () to assess the relationship between ABV and IBU. Use the regular plot function in base R (not ggplot) to create a scatter plot with the regression line superimposed on the plot. Again, this should be done for each state.

8. Address the assumptions of the regression model. You may assume the data are independent (even if this is a stretch.): 1. There is a normal distribution of the ABV for fixed values of IBU. 2. These normal distributions have equal standard deviations. 3. The means of these normal distributions have a linear relationship with IBU. 4. Independence (you may assume this one to be true without defense.)

**D. Gain inference from the model**

9. Make sure and print the parameter estimate table. Interpret the slope of the regression model. You should have one sentence for each interpretation. In addition, answer the question: Is there evidence that the relationship between ABV and IBU is significantly different for Texas and Colorado beers? For now, this is a judgement call.

10. Provide a confidence interval for each slope (from each state). Provide a sentence that interprets each slope (for each state) but this time include the confidence interval in your interpretation. See the Unit 9 6371 slides for an example of how to write the interpretation of the confidence interval. If you are not in 6371 and have not had it, ask a friend in the class to see the slides and discuss how to move forward. This is also covered in the Bridge Course. In short, the confidence interval contains the ***plausible*** values of the parameter (the slope in this case) given the data you observed. Given this new information, answer this question: Is there significant evidence that he relationship between ABV and IBU is significantly different for Texas and Colorado beers? This question is less subjective now and has a clear answer based on the plausible values for the parameters.

**E. Compare two competing models: External Cross Validation**

11. Using the beerCOTX dataframe, add a column to the data that is the square of the IBU column. Call it IBU2. Print the head of the dataframe with the new column.

12. For each state, create a training and test set from the data (60%/40% split respectively). Print a summary of each new data frame… there should be four: TrainingCO, TestCO, TrainingTX, TestTX.

13. Brewmeisters are curious if the relationship between ABV and IBU is purely linear or if there is evidence of a quadratic component as well. To test this we would like to compare two models:

Use the ASE loss function and external cross validation to provide evidence as to which model is more appropriate. Your analysis should include the average squared error (ASE) for the test set from Colorado and Texas. Your analysis should also include a clear discussion, using the ASEs, as to which model you feel is more appropriate.

Here is the predicted ABV for the ith beer, is the actual ABV of the ith beer and n is the sample size.

BONUS: Is there another method that you know of that will provide inference as to the significance of the squared IBU term? Please describe your thoughts and provide relevant statistics. Does this inference agree with the result of your cross validation?

**Reminder**

To complete this assignment, please submit **one** RMarkdown and matching HTML file by the deadline. Please submit all files at the same time; only one submission is granted.

Good luck!