**MSDS 6306: Doing Data Science - Case Study 01**

**Due: Saturday, March 2nd 2019 11:59pm.**

Submit the link to the GitHub repository via the space provided for in the Case Study 01 page in 2DS.

**Description**

The Beers dataset contains a list of 2410 US craft beers and Breweries dataset contains 558 US breweries. The datasets descriptions are as follows.

**Beers.csv:**

Name: Name of the beer.

Beer\_ID: Unique identifier of the beer.

ABV: Alcohol by volume of the beer.

IBU: International Bitterness Units of the beer.

Brewery\_ID: Brewery id associated with the beer.

Style: Style of the beer.

Ounces: Ounces of beer.

**Breweries.csv:**

Brew\_ID: Unique identifier of the brewery.

Name: Name of the brewery.

City: City where the brewery is located.

State: U.S. State where the brewery is located.

**Instructions**

Deliverable: A GitHub repository with an RMarkdown file containing the following:

1. Introduction to the project. The introduction should not reference a project. No part of this should be informal.

b. The introduction needs to be written as if you are presenting the work to the CEO and CFO of Budweiser (your client) and that they only have had one class in statistics. If it sounds like a student presentation, that is not acceptable. You may assume that the CEO and CFO gave you the data and gave you the directive to report any interesting finding that you may uncover through your analysis.

c. Briefly explain the purpose of the code. The explanations should appear as a sentence or two before or after the code chunk. Even though you will not be hiding the code chunks (so that I can see the code), you need to assume that the client can’t see them.

d. Use R to code answers concerning the seven questions below.

i. Give clear, explicit answers to the questions. Just the code to answer the questions is not enough, even if the code is correct and gives the correct answer. You must state the answer in a complete sentence outside the code chunk.

ii. Conclusion to the project. Summarize your findings from this exercise. The file must be readable in GitHub. In other words, don’t forget to keep the md file!!

iii. In fact, you will also upload the knit html file to GitHub as well. This will allow for plots and tables to supplement your answers (part e) to the 7 questions below. You are already making an Rmd file, we should take advantage of it and knit a nice presentation of the project!

iv. You should expand your repository with at least this RMarkdown file, the knit html file, the codebook, the two CSV files, and a Readme file that describes the purpose of the project and codebook. The repo can be structured however you like, but it should make sense and be easily navigated.

v. This will be a team project. I expect that all team members will do equal work. All members will need to push, add, commit, and pull to GitHub (GitHub tracks commits!). This is a collaborative project, so take it seriously and plan with your teammates. The due date for submission is Saturday 11:59pm. Each team member will need to record and upload to YouTube a **5-minute** presentation. To do this you can download Jing which is a free video software available at <https://www.techsmith.com/jing-tool.html> or use your preferred screen capture software (like QuickTime if you have a Mac). You can assume that your audience is the CEO and CFO of Budweiser (your client) and that they only have had one class in statistics and have indicated that you cannot take more than 5 minutes of their time. 20% of your grade will be based on the presentation. The presentation slides that include a link to your video should be in the Case Study Github repo before the start of the session. The goal is to communicate the findings of the project in a clear, concise and scientific manner. I will make the link available to everyone in the class so that your peers can benefit from your work and so that you can benefit from theirs. The links will be available for a week at which time you may take your video off of YouTube (or screencast.com if using Jing) if you wish. Finally, include the link in your RMarkdown file.

vi. There is no asynchronous material for Units 7 and 8 although I will be available during those weeks to meet with each team individually. I will be available Tuesday Feb19 and Thursday Feb 21 and will post a sign-up sheet for times on the wall. **Meeting at least once during this time is required.** The activities for the Live Session on Thursday Feb 28th are described next.

vii. On Thursday February 28th (Live Session) you will each give your presentation in person to an audience of one (me). This will not be for a letter grade rather a completion grade: full credit if it is done and no credit if it is not done. The idea is that you will gain live presentation practice and will also get feedback that should be addressed in the final 5-minute video of your presentation (I recommend not shooting the final video until after Thursday’s Live Session.) You will sign up for a 10 minute presentation time in which you will present your project live and I will provide you written and/or verbal feedback.

**Questions**

1. How many breweries are present in each state?

2. Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file.

3. Report the number of NA's in each column.

4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.

5. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?

6. Summary statistics for the ABV variable.

7. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot.

You are welcome to use the ggplot2 library for graphs. Please ignore missing values in your analysis. Make your best judgment of a relationship and EXPLAIN your answer.