MSDS 6306: Doing Data Science - Case Study 01

Due: June 26th, 2018 (Tuesday, One Hour Before Live Session 8)

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

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

b. The introduction needs to be written as if you are presenting the work to someone who has given you the data to analyze and wants to understand the result. Assume it's a presentation for a client. This may take some imagination of whom your client might be. If it sounds like a student presentation, that is not acceptable. You are graduate professional students, please be professional.

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.

e. 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. *Pretend the code is not visible.

f. 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!!

g. You should expand your repository with at least this RMarkdown file, 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.

h. 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 1 hour before live session 8.

During live session 8, each group will have up to 20 minutes to present their case study report. A portion of your grade will be based on the presentation. The presentation slides (if applicable), 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.

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.