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

**Description**

The Beers dataset contains a list of 2410 US craft beers and the 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.

**Audience**

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. The company has given the data to 10 consulting companies, yours is one of them and each group gets 7 minutes to give their analysis and make it to the next round. Each group gets a maximum of 7 minutes to present.

20% of your grade will be based on the presentation.

The pitch is focused on 9 questions / items described in the next section. Beyond those general questions you may and should speculate / anticipate what may be of interest to them. This could make all the difference in their decision.

**Deliverables:**

1. **UNIT 8 Live Session: EDA Tuesday, Oct 10**

Each team will need to meet with the professor and present their EDA in Unit 8 Live Session. These are one on one meetings between each team and the professor. One scheduled time commitment in Unit 8 is to attend this 20 ish minute meeting with your professor. The rest of the time is reserved to work on your project with your partner. But we will meet at 6:30 for 15 minutes.

Your goal is to present your EDA (Answers to Questions 1,3,4,5,6,7). At this point, teams should have presentation quality slides and presentation prepared. Responses to each of the questions listed above should be prepared and addressed in this meeting.

The grade for this portion is based on the slide deck and the presentation. If the team is prepared and delivers a well-practiced presentation it should be easy to earn close to a 100% here. Both team members must present about equal time. The presentation is one of the main skills we are practicing ☺. Please make sure and see the asynch videos for Unit 8.

1. **A GitHub repository (Due Saturday October 21st 11:59pm CST)**

The GitHub repo should contain the following items and a link to the GitHub repo should be placed on a Word Doc (or PDF) and submitted to Canvas for Unit 9

1. *An RMarkdown file containing the following:*
2. 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.
3. 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.
4. Use R to address the 9 questions / items below:
5. How many breweries are present in each state?
6. Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file. (RMD only, this does not need to be included in the presentation or the deck.)
7. Address the missing values in each column. Include a short mention of if you are assuming the data to be MCAR, MAR or NMAR.
8. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.
9. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?
10. Comment on the summary statistics and distribution of the ABV variable.
11. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgment of a relationship and EXPLAIN your answer.
12. Budweiser would also like to investigate the difference with respect to IBU and ABV between IPAs (India Pale Ales) and other types of Ale (any beer with “Ale” in its name other than IPA). You decide to use KNN classification to investigate this relationship. Provide statistical evidence one way or the other. You can of course assume your audience is comfortable with percentages … KNN is very easy to understand conceptually.

In addition, while you have decided to use KNN to investigate this relationship (KNN is required) you may also feel free to supplement your response to this question with any other methods or techniques you have learned. Creativity and alternative solutions are always encouraged.

1. Knock their socks off! Find one other useful inference from the data that you feel Budweiser may be able to find value in. You must convince them why it is important and back up your conviction with appropriate statistical evidence.

**Directives on RMD File:**

1. Give clear, explicit answers to the questions. Just the code to answer the questions is not nearly 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.
2. 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!!

*2. Knit HTML file.*

*3. Codebook, Both CSV files and a ReadMe.md*

The Readme file 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.

**Codebook:** The code book is an explanation of concepts or definitions needed to know to understand the data and reproduce the analysis. This could include the machine and operating system the project was run on, a data dictionary of the data, any transformations or feature creation performed with the data, merging details, or any other pieces of general information that someone should know about the environment and data. For this project the codebook should be a Word doc or pdf or could be the ReadMe file in your GitHub repo.

*4. PPT Slides*

Described more below and should have the link to your YouTube presentation … described further below as well.)

**C. Final YouTube / Zoom Video: Due Saturday October 21st 11:59pm**

The group will need to record and upload to YouTube or Zoom a single **7-minute** or less presentation of your findings. To be clear, this is one video with both people presenting (each person should not make their own video). Each person should present roughly half of the presentation.

At this point you should know your presentation backwards and forwards. If there are too many errors in the recording, you should start over until you have a very polished presentation that does not exceed 7 minutes. I regularly need 4-7 takes to make a 7 minute video.

To record you can use Zoom or you can download Camtasia (free trial) which is a video software available at <https://www.techsmith.com/video-editor.html>. If you have a Mac you may wish to use QuickTime or of course you may choose to use your preferred screen capture software. I recommend Zoom… very easy.

The presentation slides that include a link to your video should be in your Case Study Github repo as well as on the Google Doc provided by your professor (below). The goal is to communicate the findings of the project in a clear, concise and scientific manner. Also, uploading to YouTube or Zoom is not difficult. Here is a YouTube video to help with uploading to YouTube: <https://www.youtube.com/watch?v=VtF2AgFSLAw>

Uploading to Zoom is automatic… just please be sure to not protect it with a password or please make the password available to your professor. It is preferred, if possible, to make your video public (not password protected) for a week in order to let your peers learn from you presentation. This is a huge value add to you classmates and they will add reciprocate by making theirs viewable.

Your professor will make the Google Doc link available to everyone in the class so that your peers can benefit from your work and so that you can benefit from theirs. Again, student’s presentation links will be available for a week at which time you may take your video off of YouTube / Zoom if you wish.

Google Doc Link:

<https://docs.google.com/document/d/1l_bgUf8O7mKt9OXadoGh8A3g1PWbeWA-nJGjj858KRk/edit?usp=sharing>

**Collaboration and Peer Review**

This will be a team project. We expect that all team members will do equal work and give their best to advance the knowledge of both themselves and their teammate. **This is a collaborative project, be sure and communicate early and often; mutual respect is key.**

You will be providing two peer reviews that will be submitted to 2DS in Unit 8 and Unit 9 under: ***Project 1: EDA and Peer Review*** (by Saturday, October 14th 11:59pm / Unit 8) and ***Project 1: Final Documentation, Presentation and Peer Review Assignment*** (by Saturday October, 21st 11:59pm / Unit 9). See the Rubric for detailed information on the peer review.

**Unit 9 Live Session:**

During Live Session for Unit 9, we will have our DSNOW and then the professor will be available for a live Q & A about the project / presentation. We will adjourn early, please use the extra time to work on your project one way or the other.

**Extra Value!**

The genesis of this project came from our Big Data Advisory Board here at SMU which is comprised of 30+ SVPs and CEOs from Dallas based companies such as AT&T, Microsoft and Capital One just to name a few. I have discussed this project with nearly every member of this board and 100% of these conversations have included the fact that there is not a more important skill to them than effective communication and presentation. This is a chance to practice and hone your skills. I will be providing specific feedback along the way and you will present in many (if not every) course in the program! Please let me know if you have any questions!

In addition, the idea to record the videos came from this advisory board as well. Some on the board mentioned that they saw huge value in using the video as a part of an applicant’s portfolio in the application / hiring process. In addition, many also remembered the value in watching others present as well as themselves.

You have great ideas and insights! Let’s work together to continually polish and hone your skills in delivering those high value inferences!

**Finally!**

Please use ChatGPT! A good strategy would be to do the project on your own to learn in depth the challenges and nuances of the data the methods that you apply to the data; then, work to phrase a question or questions (prompt) to give ChatGPT the request(s) that will help generate the code or materials that you are looking for. This will likely take at least a few attempts and rewrites to get your most thorough set of instructions. Finally, take note on how close and how useful ChatGPTs responses are and if it was worth your time. What types of errors did it make, what value did it bring? Please add a slide or two to briefly answer those questions … although please add those to the appendix.. you don’t have to put them in the video presenation… but you can if it works for you!