MSDS 6306: Doing Data Science - Case Study 01

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

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 7 minutes of their time. 20% of your grade will be based on the presentation.

They have hired you to answer the 7 questions listed below and beyond those general questions you may speculate / anticipate what may be of interest to them.

Deliverables:

A. A GitHub repository (Due Saturday Oct 26th 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 2DS for Unit 8.

The final repo which will be checked after 11:59pm CST Oct 26th should contain:

1. *An RMarkdown file containing the following:*

    a. 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.

    b. 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.

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

Analysis 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. (RMD only, this does not need to be included in the presentation or the deck.)

3. Address the missing values 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. Comment on the summary statistics and distribution of the ABV variable.

7. 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.

8. 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.

9. 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:

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!!

2. Knit HTML file.
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!

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.

4. PPT Slides
Described more below and should have the link to your You Tube presentation … described further below as well.)

B. UNIT 8 Live Session: EDA

Each team will need to meet with the professor and present their EDA in Unit 8. It is up to the teams and the professor when and how to set up these meetings. They may be during the schedule live session time, but given time constraints, some teams will need to schedule times outside of their scheduled live session time. With that said, these are one on one meetings between each team and the professor. Your only scheduled time commitment in Unit 8 is to attend this 10ish minute meeting with your professor. The rest of the time is reserved to work on your project with your partner.
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.

B. Unit 9 Live Session: Q & A.

During Live Session for Unit 9, the professor will be available for a live Q & A about the project / presentation. Attendance at this live session is not required (attendance is optional). NOTE ABOUT POWERPOINT … You may use the same powerpoint or develop them separately. I would imagine that even if you develop the powerpoint together that each student's final powerpoint will be a little different just based on individual presentation style. Everyone has their own unique style and delivery.

C. Final YouTube Video

Each team member will need to record and upload to YouTube a **7-minute** or less presentation of your findings. At this point you should know your presentation backwards and forwards. If you trip up too much in your recording, you should start over until you have a very polished presentation that does not exceed 7 minutes.

To record you can download Camtasia (free trial) which is a video software available at https://www.techsmith.com/video-editor.html or use your preferred screen capture software (like QuickTime if you have a Mac.) The presentation slides that include a link to your video should be in the Case Study Github repo as well as on the Google Doc provided by your professor. The goal is to communicate the findings of the project in a clear, concise and scientific manner. Also, uploading to YouTube is not difficult. Here is a YouTube video to help: https://www.youtube.com/watch?v=VtF2AgFSLAw 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. Student's presentation links will be available for a week at which time you may take your video off of YouTube if you wish.

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. All members will need to push, add, commit, and pull to GitHub. **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 Oct 19 11:59pm / Unit 8) and ***Project 1: Final Documentation, Presentation and Peer Review Assignment*** (by Saturday Oct 26 11:59pm / Unit 9) . See the Rubric for detailed information on the peer review.