Data 403: Project 1

Project Description

Overview

In this project, you will study liquor sales at stores in the state of Iowa. You will target your work towards two different clients:

Client A: Booze 'R' Us - A company that owns a large number of liquor stores in Iowa. They have asked you for help projecting their sales in the coming year, so they can decide whether they will have enough income to expand their operations.

Client B: Drinking Excess Alcohol is Dangerous (DEAD) - A nonprofit group seeking to make alcohol culture safer in Iowa. They have hired you to analyze patterns in sales to help them understand what factors drive higher or lower alcohol purchases.

The primary learning goals of this project are:

- Perform feature engineering and simple data integration tasks to prepare datasets for machine learning.
- Implement linear regression model fitting procedures and tailor your implementation to work with sufficiently large datasets
- Evaluate prediction accuracy of different models using validation metrics.
- Tailor your model selection and presentation approach depending on clients' interests; specifically, focusing on either *prediction* or *inference*.
- Consider ethical and legal responsibilities in your machine learning process and final recommendations.

Data

Data on alcohol sales in Iowa is collected by the state government here; https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy.

There are several complex data decisions you need to make for this project:

1. *What data will you pull from that website to answer your clients' questions?*
   It is likely infeasible (and probably unnecessary) to use every single dataset available. Also, you may want to consider developing your model approach on small subsets of data, before fitting the final model on a larger set.

2. *How will you clean your data?* This is real data; some entries will be missing or typoed or different across years. It will be impossible to perfectly clean the data, so you'll need to determine what cleaning is most

important. You will also need to think carefully about missing data, and how to handle it in an unbiased and ethical way.
3. *What new features will you create?* You are not limited to using only the columns currently in the data you download - you can combine and modify them into new measurements as much as you see fit.

In this project, you are **not** required to use any external data beyond what is supplied at the link.

There are also no restrictions on how you pull the data from the website. Using the API is nice, of course, but manually downloading the csvs or similar is also okay for this project.

Coding

In this project, your final analysis must be performed using **custom functions only**.

That is, you may not use any model fitting functions from R or python, such as `LinearRegression()` or `linear_reg()`. You also may not use any metric computing functions, like `r2_score()` or `rsq()`, or any cross-validation shortcuts like `cross_val_score()` or `fit_resamples()`.

You **may**, however, use built-in functions for all of your data cleaning and preparation. You may also use any functions you want during your code and model development process; for example, to check that your own functions come up with approximately correct results.

Group roles

One of the key elements of this projects is to plan out your analysis choices and steps before you start typing on a keyboard. There are many of you in each group, and also lots of work to get done; you'll need to find a way to make sure you are all on the same page, and then split up the work.

Your first task will be to assign each person a meta-role in the group.
This article provides a good list of possible roles, although you may use anything you find online or make up your own. Importantly, these are *organizational* roles; nobody should have the role of "python coder" or "report writer", even if you do end up dividing the work that way.

Deliverables

There are four deadlines for this project. All deadlines for

Only one person per group needs to submit artifacts to Canvas; when in doubt, the randomly assigned group leader is responsible.

All deadlines are **11:59pm**, except for the presentation, which takes place in class; slides must be uploaded before class.

Data Preparation: Friday, October 6th

You will turn in to Canvas a 1-2 page description of the data collection, data cleaning, and feature engineering choices you have made; as well as the code that you are using to perform these data transformations. (Do *not* turn in any raw datasets.)

This is a checkpoint to make sure you are on track; it is not graded on quality. The document can be bullet point and casual form, and the code does not need to be prepared for public consumption - we just want to see that it exists.

Project Proposal : Friday, October 13th

You will turn in **two** 1-3 page documents (including tables and visuals). These are project proposals for the two clients. Each proposal should include:

- A brief description of your plan for what data you will use and how you will wrangle it.
- A brief description of the models that you plan to fit, and how you will perform model selection and validation.
- A vague promise about the type of conclusions the clients can expect from you in your final analysis.

These should be professionally formatted and include supporting links, visuals, and/or tables or summaries.

Think of it this way: Suppose the client has paid you a preliminary fee, but they are waiting to decide if they should hire you for the full project. This document should convince them.

You may use technical terms, but do not include anything overly detailed or complex - imagine this document will go to both executives (who do not really know data science) and to an analytics department (who do know data science, but perhaps not as well as you).

This deliverable will be graded on **professionalism**, **clarity**, **quality of the plan**, and **client-focused appraoch**.

Codebook: Wednesday, October 18th

You will submit a bundle of reproducible code and documentation thereof. This can be in any format(s): scripts, notebooks, vignettes, etc.

The only requirements are:

- Code must be R, python, or a combination thereof.
- Code must be clean and well-commented.
- There must be fully runnable scripts and documents that perform the analysis you use in your final report and presentation.

You should *not* include any raw datasets - but it should be clear what data you used and from where, so that a reader could in principle recreate the analysis from scratch.

This deliverable will be graded on **correct model implementation**, **correct model selection**, and **code documentation clarity**.

Final Presentation: Friday, October 20th

You will also give an oral presentation in class, presenting your final results to your "clients".

Each of the two presentations should be no more than 5 minutes long. Your goal here is to impress the client, given them a high-level sense of your analysis, and give them the major takeway conclusions.

Every member of your group must speak; however, you do not need everyone to speak in *both* of the two presentations.

Slides **MUST** be turned in to Canvas before the start of class.

This deliverable will be graded on **quality of slides**, **speaking style**, and **impressiveness to client**.

Final Reports: Wednesday, October 25th

Your main graded item is your two written reports, one per client.

Each report needs to:

- Introduce the problem and the data
- Briefly summarize and justify the data preparation choices.
- Describe your model selection and validation process in detail.
- Provide a final model summary and justification.
- Provide a takeaway message to the client for how to use your analytical results.
- Include throughout your report, or as its own section, a discussion of your ethical concerns and recommendations.

There is not page minimum for this, but a good target is approximately 3 pages of written content per report, plus approximately 3 pages worth of tables or visuals.

This deliverable will be graded on **clarity and professionalism**, **correctness of analysis description**, **correct interpretation of results**, and **quality of meeting client needs**.