Homework 1:

1. Use chat-GPT or a generative AI of your choice for this question. Use whichever model you prefer or have access to. Create a question that requires the use of a LEFT JOIN. The question should be similar in difficulty to question 7 or 8 (must be medium-advanced difficulty question). Type out the problem, solve it, and provide the answer.

Questions 7 & 8 – for reference

7. Calculate the total revenue for each artist and rank the top 5 artists by revenue. Use the sale_price.
8. Find the average sale price per square meter of all paintings in each museum and rank the museums by this value in descending order. The average sale price per square meter would make a great new column of a dataset. Display the first 5 museums of this new dataset, the average price per square meter, and the total number of paintings used in the calculation.

2. Use chat-GPT or a generative AI of your choice for this question. Use whichever model you prefer or have access to. Create a question that requires the use of a FULL OUTER JOIN. The question should be similar in difficulty to questions 1-8 (can be an easy question). Type out the problem, solve it, and provide the answer.

* Note: Questions 1-8 range grow in difficulty – all are SQL questions.

Homework 3:

1. Using chat-GPT (or some other generative AI), ask it what model would be best for this data.
    a. Explain if you agree with its answer.
    b. Research and attempt a different classifier (suggested by the AI). Does it perform better or worse than the logistic regression classifier? Why?

* Note: This problem extends a textbook problem: (ISLR 4.16

Homework 5:

PROBLEM SET_UP

1. (13 pts) Alzheimer's Disease Synthetic Dataset

Ref: https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset?resource=download

The data is attached to the submission folder on D2L and is available at the link above.

The target class is Diagnosis.

a. How many levels does the target have?
b. Create an 80/20 split for train/test
c. (2 pts) Build a logistic regression model (use *function* 'glm()') to predict whether a patient has Alzheimer's or not. Use only the 3 best features. Use the train/test from above.
d. (5 pts) Now, use cross-fold validation (k=5) (do NOT use (b), DO use the entire dataset) to build:
    i. Logistic regression model. Use (method='glm'). Compare the accuracy to (c).
    ii. Naïve Baye's model. Record the accuracy.
    iii. QDA. Record the accuracy.
    iv. KNN. Record the accuracy.

    Hint: use trainControl() & train()
e. You likely have an error with KNN. What is the error? Look at your answer for (a) – do you understand where the error is coming from? Explain.
f. AI PART: (You may choose to do this problem without generative AI.) Using generative AI, create a custom trainControl that fixes the error. It is very, very likely that whatever it gives you will not work, but it will be close. Try and fix the code. It does not need to work for full credit. Instead, for full-credit, comment your attempt edited from the LLM of your choice with a line-by-line explanation of how it is meant to fix the error in (e). A solution and explanation will be posted after the assignment is closed.

Homework 6:

1. Using an LLM of yourself, find a way to ensemble other than taking an average. Describe how to do it here. You do NOT need to implement it. Double check the LLM information with a quick literature search using the web.

Homework 8:

1. Find an article or webpage on survival analysis.
    a. Post the link
    b. Summarize the article
    c. Have chat-GPT or another LLM summarize the article

d. Which summary do you think better represents the article, and which summary better represents what you learned from the article?
e. Give one real-world example of a survival analysis use-case. You may need to find another article.

In-class Assignment 4:

1. Use the advertising.csv file1, answer the following questions. The target is the 'Clicked on Ad' field.
    a. What do the fields/column names mean? You may use Google/chat-GPT to answer this question. What units do the fields have?
    b. Is the target field balanced?
    c. Which fields are numerical, and which are categorical?
    d. Ask chat-GPT which two variables are most likely to be important when predicting if a user clicked on an ad. Paste the prompt. Explain if you agree with the AI's analysis.
    e. Do some basic exploratory analysis on the two variables chat-GPT gave you, or two variables of your choosing if you disagree with it.

In-class Assignment 6:

1. SMOTE stands for Synthetic Minority Over-sampling Technique.
    a. First, ask chat-GPT to explain what SMOTE is to you in three different ways. You can ask it to explain it to you from different perspectives, with various background information, etc. For example, I asked it to explain it to me like I am a grade-schooler, and I asked it to explain it to me through the context of a data scientist.
        i. How did the LLM's explanations differ?
    b. Now search for articles on SMOTE – you may find some on Medium, GeeksforGeeks, etc. Find at least 2. Read them thoroughly.
        i. Is there conflicting information between what the LLM gave you and the articles? Which do you think is correct? How can you check?
        ii. If the LLM was wrong, what happens if you attempt to correct it? Can you correct it over and over?
        iii. Even if you believe the LLM is correct, try to 'correct' it with something you know is not true. What happens? How do you know which output is correct?

In-class Assignment 12:

1. Do some basic research on any topic we have covered in class.
    a. List what you've learned in a few bullet points below. Cite your source (just a link). My lectures are included!
        i.
        ii.
        iii.
        iv.
    b. Have an LLM of your choice take the information you found and make it do the following (there is no wrong way to do this, and you can `spice up` the prompts if you want to!):
        i. Explain the information to a 5-year-old
        ii. Explain the information to a world-renowned professional data scientist
        iii. Explain the information to your best friend, pet, or sibling.
    c. Which explanation is your favorite? There's no wrong answer.

Bonus:

Some questions caused students who were overly relying on AI to mess-up. These were noted in class and through our LMS.

The first problem of HW 7 (out of 8) requires students to hand-create a neural network. The problem is broken into many steps. It was meant to match the output of the NN shown in class. Many students received predictions of all 0s, or all 1s. This was the biggest reason why (by a show of hands):

1. **You trusted a hallucinating LLM**. Chat-GPT will tell you that you only need one output node. It misses that the instructions asked for two, and that *y* is to be treated as a factor, not as numeric 0/1. It will give you one output node and tell you to compare it to 0.5, then assign labels. For some people, this works because you did not turn *y* into a factor. When I mentioned the issue, it showed me how to look for two nodes, then adjusted itself. I am not sure why a number of you didn't do that, but I am glad you did it in my class and not during a coding interview.

On the 11th In-Class Assignment (out of 13), students were asked to create autocorrelation and partial autocorrelation graphs, then comment on seasonality. Many students used the wrong graph to look for the wrong things. For example, they looked for spikes on the autocorrelation graph instead of the partial autocorrelation graph. ICA answers are given in-class and are recorded and posted shortly after lecture, so I could not figure out where they mixed these up. Chat-GPT and Claude both confused the two graphs when I asked the same question.