Homework 1

You may form small groups (e.g. of up to four people) to work on this assignment, but you must write up all solutions by yourself. List your study partners for the homework on the first page, or "none" if you had no partners. **Start each problem on a new page**, and be sure to clearly label where each problem and subproblem begins. All problems must be submitted in order (all of P1 before P2, etc.). No late homeworks will be accepted. This is not out of a desire to be harsh, but rather out of fairness to all students in this large course.

Question 1 (True or False)
Provide brief explanations for your answers.
(a) Correlation implies causation.
(b) Data scientists do most of their work in Python and are unlikely to use other tools.
(c) Most data scientists spend the majority of their time developing new models.
(d) The use of historical data to make decisions about the future can reinforce historical biases.
(e) All data science investigations start with an existing dataset.
(f) Linear regression is an example of an unsupervised model.
(g) Linear regression can be used to confidently make predictions inside the range of data on which it was trained on (assume that linear regression works well as a model for this data).

Question 2
You recently learned about Google's new system for detecting breast cancer in mammograms (https://www.nature.com/articles/s41586-019-1799-6). The system was trained on a large dataset of annotated mammogram images from the UK and the USA, most of which were acquired on devices made by Hologic. The paper shows that the model can be trained on the UK dataset and still perform well on the USA dataset. Your friend finds the work exciting, and would like to use Google's pre-trained model to detect breast cancer in Brazil. Is this a good idea? Why or why not?

Question 3
Your friend working at UCLA housing has been given the task of determining how students feel about the new dorms. Your friend wants to accomplish this by scraping Twitter and news article comment sections for tweets containing keywords and hashtags related to the new dorm and then running them through a model that does sentiment analysis. This model contains an algorithm that says whether the text exhibits positive, neutral, or negative sentiment. What are some of the issues, if any, with what your friend proposes?

Question 4

You would like to see if you can predict the probability that a given student will take a particular class in the upcoming quarter.

(a) What are some features you would try to gather to investigate this problem (e.g. student's year in school, professor teaching the course)?

(b) How would you formulate your labels? (Think about ordinal/categorical/numerical encoding, min/max/possible values, etc.)

(c) How could you source/obtain/gather the above data?

Problem 5

In Project 1 you learned about imputing data, the step a data scientist must take to deal with missing or null values in a dataset.

(a) List four different strategies you could reasonably use to address null values. For each, clarify what the advantages and disadvantages to it are. Additionally, for each strategy, speculate on what sorts of datasets it would likely be the most effective approach, and what sorts of data would it be inadvisable for.

(b) Outliers can also be problematic points in a dataset. Which of the strategies you mentioned in part (a) would also work for outliers, and why?

Problem 6

A jar contains 3 red, 2 yellow, and 1 blue marble. Aside from color, the marbles are indistinguishable. Two marbles are drawn at random without replacement from the jar. Let X represent the number of red marbles drawn and Y represent the number of blue marbles drawn.

(a) What is $P(X = 1)$?

(b) What is $P(X = 0, Y = 1)$?

(c) What is $P(X = 1 \mid Y = 1)$?

Problem 7

Evaluate the following features and determine if you would one hot encode them. Justify your response:

(a) Walking pace, measured in miles per hour.

(b) Class name (freshman, sophomore, junior, senior).

(c) City in the US (part of an address).

(d) If someone has diabetes or not.

(e) Laptop brand.

(f) Interaction term of how much sleep one got the previous night (numerical value, in hours) and test difficulty (3 possible values: easy, medium, hard) when trying to predict test score.