Comp Neuro, HW1
Due Weds, 9/18 by 5PM

Question 1. Models, features, and dimensionality. Scan through a few recent issues of the journals Nature Neuroscience, and Neuron, and find a few papers that seem interesting to you from their titles and abstracts. Then, look these and try to find examples of a **regression task**, and a **classification task**. *This might take a while, especially if you do it thoughtfully with the intention of really trying to learn/understand*. Answer the following questions:

For the regression task:

a) What is the overall scientific/experimental context (i.e. what's the point of the study). Two or three sentences is fine. Describe it in plain language (using minimal jargon and technical terminology). Provide me with a link to the study.
b) What is the specific regression task, and what relationship is it trying to establish?
c) Describe:
    a. The feature(s), and their dimensionality, and how you determined this
    b. The target(s), and their dimensionality, and how you determined this
    c. The sample size, n, and how you determined this

For the classification task:

a) What is the overall scientific/experimental context (i.e. what's the point of the study). Two or three sentences is fine. Describe it in plain language (using minimal jargon and technical terminology). Provide me with a link to the study.
b) What is the specific classification task, and what relationship is it trying to accomplish?
c) Describe:
    a. The feature(s), and their dimensionality, and how you determined this
    b. The target(s), and their dimensionality, and how you determined this
    c. The sample size, n, and how you determined this

Question 2. Exploratory data analysis with Pandas. Download the Allen institute data cell types data (as was done in the class notebook), and use your pandas knowledge to answer the following questions:

a) Based on the distribution of cellular resting potentials (Vrest) does there seem to be evidence of multiple cell types? Show any plots that are relevant, and briefly explain your reasoning. Note that I'm asking you to reason *based on what the data show*, not what you believe to be true beforehand (i.e. that there are of course different types of cells)
b) Make a plot of resting potential vs. latency to first spike. Does the result surprise you? Would you say the data are strongly correlated? Explain.
c) Explain why a scatterplot matrix is always square.
d) Write a groupby statement that lets me see the median values of all measurements of all cells, broken down by brain region. How many brain regions are there?
e) Write a conditional statement that shows me ***only*** cells from the right hemisphere, and for which the resting potential is greater than -71 mV.

f) Use seaborn's displot function to show histograms of the average inter-spike interval (ISI), broken down by disease state. Describe the histograms and interpret any differences. What differences are meaningful, and what might be artifacts of small sample size, etc? Interpret the data scientifically in a sentence or two.
g) Why does it make no sense to compute a correlation between two **categorical** variables?
h) Make a new dataframe from the original dataset, consisting of only the columns with numerical/ordinal (i.e. not categorical) variables. Then, make a correlation heatmap of the data. Why is it symmetric?
i) Based on what you observe in the heatmap, which variables are most positively correlated, and which variables are most negatively correlated? Briefly explain. Does this jive with your intuition?

Question 3. Models, loss, and correlation:

There are two friends who love data. One is name John Training, and the other is named Steve Testing. They also love computing errors (who doesn't!), and as a convenient shorthand, we call the errors that they produce "Training error" and "Testing error"

In a first experiment, John trains a model on some data relating time spent writing code (x values), and coding proficiency (y values) described below. He fits a linear relationship to these data (y = ax + b), and he obtains a=5 and b = 1. John then computes his model's loss (as mean squared error, or MSE). Populate the following table, and compute the mean-squared error as well. Show your work

# hours coding/week (x)	Measured coding proficiency (/100) (y)	Model fit values	Model residual (error)
0	0.21		
2	9.1		
4	20.1		
6	28.4		
8	39.2		
10	50.3		

Next, Steve collects his own data on coding time (x) and coding proficiency (y). He obtains the following data. He tells John that if he really wants to see if his model is any good, he should calculate the loss on his (Steve's) dataset, not his own (John's).

# hours coding/week (x)	Measured coding proficiency (/100) (y)
0	-.76
2	9.93
4	21.4
6	29.8
8	41.0
10	53.4

a) Is Steve's reasoning sound? Explain briefly.
b) What is the correlation between John's data and Steve's data? Show how you calculated this.
c) What is the (approximate) twentieth percentile of Steve's data? Describe your reasoning process.
d) Draw a histogram of John's data, with two bins. Describe your reasoning process.
e) What is the training error vs. the testing error for John's model? From this calculation, is there a strong indication of overfitting? Explain and show all work/calculations.

Suppose God now descends from Heaven, and says: "Behold! I never told you this before, but 4 billion years ago I designed the universe such that there would be a precise relationship between time spent writing code (x) and coding proficiency (y) is y = x^2 + 7."

f) Presuming we can trust God, would you say that John's model is over-fit, or under-fit? Would you say it's a biased model? Describe why (strong arguments will be quantitative and include calculations and sketches)