CSE 416: Introduction to Machine Learning
Assignment #1
Instructor: Pemi Nguyen
due: Friday, April 8, 2022, 11:59 pm.

Instructions:

- **Answers:** Please provide detailed yet concise explanations for your work.
- **Turn-in:** Do not write your name on your pages (your Gradescope account will identify you to us) and do not include a copy of the exercise's question in what you turn in. You must use Gradescope to upload your written homework solutions. You will submit a single PDF file containing your solutions to all the exercises in the homework. Each numbered homework question must be answered on its own page (or pages). You must follow the Gradescope prompts that have you link exercise numbers to your pages. **We do not accept handwritten solutions**, so please typeset them, such as Microsoft Word or LaTeX editor.

1. (a) Explain the basic differences between train set, validation set and test set.
   (b) Describe the procedure to train multiple models with different complexities and how to choose the best one, using the approach where we use only one validation set.
2. Please determine whether the following practices are good or not. If not, please explain why.
   (a) Amal has a housing price dataset of 1 million examples. He uses the first 20% of the dataset as the test set, the next 20% as the validation set, and the remainder as the train set.
   (b) After training a classifier model using 50 columns in a dataset and seeing a training error of 1%, Zeynep evaluates it on a different dataset containing unseen data and gets an error of 50%. She decides to add more columns and gather their values in order to get a better model.
   (c) Aric, a data scientist, claims he can train a model that can predict perfectly (zero test error) the height of a person by gathering various biological statistics about them.
   (d) Pemi intended to use a k-fold cross-validation with $k = 100$. However, due to limited budget and computing resources, he reduces the number of folds k to 10.
3. For the coding notebook assignment, which linear regression model (the one using basic features versus the one using advanced features) do you think would perform better in the future? Report the numbers you had and provide a brief explanation.
4. **Extra credit**: What is the issue with the two linear regression models from the coding assignment when we use the column "zip code"? Propose a solution to overcome it. (Hint: It has to do with the data type.)