Data Visualization [Optional Review Questions]:
Describe how you could modify a scatter plot to encode information about a categorical variable.

List two ways you can solve overplotting.

Standardization and Normalization:
What is the difference between standardization and normalization?

Given a numpy array `x` that contains numerical values, write a line of code that *standardizes* `x`.

CIFAR10 has 50000 32x32 pixel images, where each pixel is comprised of three channels for the RGB values. Say the dataset is stored as a 50000 x 32 x 32 x 3 tensor. Let's say we want to take the channelwise mean and standard deviation. Across which axes will we do this?

One-Hot Encoding:
Suppose you have a large dataset containing survey data, and one of the features is the participant's hometown. Assume that there are a very large number of unique towns. Would it be appropriate to directly apply one-hot encoding on the feature in this situation?

Now assume this data is given to you in a pandas dataframe called `survey` with a "hometown" feature. After appropriately preprocessing the feature, we want to one-hot encode it. Write a code snippet that replaces the original hometown feature with one-hot encoded features.

Processing Text Fields:
Does the following pattern fully match with the text? Ignore the quotes, they just denote that they're both strings.

- Pattern: "\w+@\w*"
- Text: "eecs@berkeley.edu"

Given that you have records that will always have the same format as the following log:
```
127.0.0.1 user-identifier frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
```
Write a line of regex that extracts the date, month, and year in separate capture groups. Assume that the record is stored in a string called `log` and `re` is already imported.

Handling Imbalanced Class Distributions
Describe one disadvantage of directly modifying the dataset to balance classes instead of using an alternative method such as changing the evaluation metric or collecting more data.

In what kinds of data/problems is SMOTE a good algorithm for generating synthetic samples? When would it not be appropriate?

Handling Null/Missing Values:
Briefly outline the two overarching methods that are used to handle null values.

Given a pandas dataframe `housing` with a feature "price", write a line of code that replaces null values with the mean price. Which one of the overarching methods is this?
When using a model such as linear regression or k nearest neighbors to impute null values, what is one concern to always look out for?

Removing Outliers (without OMP):
What is a scenario where you wouldn't want to remove outliers?

What is the main problem shared by all outlier removal methods we've provided?

Suppose you have a pandas dataframe `titanic` that contains the Titanic dataset. After plotting passenger fares on a histogram, you decide to remove all rows where the fare feature is larger than 200. Write a line of code that performs this operation.

Image Data Augmentation:
Why is it important to only apply transforms to the training set and not the validation set?

Let's say we decide to take five random crops (randomly picking between the set of corner crops and the center crop) and then a random horizontal flip as our augmenting scheme. By what factor are we increasing the effective size of the dataset?

Data Decorrelation and Whitening
Why do we decorrelate features?

What is the intuition behind using the U matrix as our linear transformation in both PCA whitening and ZCA whitening?