Data Visualization:
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?

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?

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"

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.

What kind 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.

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?

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?