Let's say we work at a credit card company and we're trying to figure out if people are going to pay their bills on time. We have everyone's purchases, split into four main categories: groceries, dining out, utilities, and entertainment. What are some ways you might use KNN to create this model? What aspects of KNN would be useful? Write up your thoughts in submit a link below.

If we have information on whether or not the customer paid their bill at the end of the month (this will be our classification variable), we can use the other four dimensions to calculate the Euclidian distances in a univariate and/or multivariate way. For example, we can look at one dimension (such as entertainment spend) in isolation that might provide information on its own about whether or not a customer is likely to pay their bill. Customers with very little money to spend on entertainment might be more likely to miss their payment, for example, and so we could use the nearest neighbors on a simple line to make a classification.

Of course, we would also need to examine how these category spend variables perform when we combine 2-4 dimensions to calculate the distance to the nearest neighbors. Perhaps there is significant clustering among payers/non-payers when we add dining spend as a dimension to our existing entertainment-spend based model. Ideally, we should examine each possible combination of these 4 categories in 1, 2, 3, and 4 dimensional distance calculations and test output accuracy.

Perhaps we could also employ PCA on this analysis to reduce dimensionality and model complexity. For example, entertainment spend and dining out spend might show significant covariance, allowing us to drop one of these variables from our input data for the KNN classifier.