Clustering images using Principal Component Analysis (PCA)
In this problem we will investigate how to cluster images using PCA. In this case the dimensions are the pixels of the image. In particular, we will investigate a data set consisting of images of handwritten digits 1 and 7. This is a subset of larger data set of handwritten digits which is often used to test and benchmark classification and learning algorithms (see http://yann.lecun.com/exdb/mnist/).

The data for this problem are provided in the file *digitsPCA.m*. Each image is 28x28 pixels, so can be thought of as a point in 784 dimensional space. Our objective is to reduce the dimensionality of the image space via PCA, and find a low-dimensional subspace where the images from digits 1 and 7 form separable and distinct clusters.

- Load the data into MATLAB. The matrix `Data` is 28x28x1000, and represents a stack of 1000 images, each with 28x28 pixels. The first 500 images are of the digit 1, and the last 500 images are of digit 7. To visualize an example in grayscale you can type in MATLAB: `imagesc(Data(:,:,1)); colormap gray.`
- To perform PCA on this data set you need to transform each image from a 28x28 matrix into a 784x1 column vector. The command `reshape` will be helpful. Then, construct a matrix that has each image vector as a column. The size of this matrix should be 784x1000.

1. **Center the data from each pixel by subtracting off the average pixel intensity of that pixel. Why do we do this before performing PCA? Hint: think about, in 2D, what goes wrong if you perform PCA on a cloud of data that is not centered at 0.**
2. **Using the centered data, compute the pixel covariance matrix. This is a simple matrix multiplication. What size is this matrix? What does the ij-th entry of this matrix represent?**
3. **Use the singular value decomposition to find the eigenvectors and eigenvalues of the pixel covariance matrix. What do the eigenvectors represent?**
4. **Write the pixel covariance matrix as a product of matrices from the SVD decomposition.**
5. **In two subplots, plot the eigenvalues in descending order, and the percentage of variance explained (`cumsum` may be helpful). How much variance is explained by the first two principal components?**
6. **Reshape the first 9 eigenvectors back into 28x28 matrices. Using subplot, show them as gray scale images. What do these images represent?**
7. **Rotate the mean-subtracted data onto the eigenvector basis set. Hint: this is just matrix multiplication. The first two elements of each vector in this rotated basis correspond to the projection of the data onto the first two eigenvectors. This is called the first two principal components**
8. **Then, make a scatter plot of the first principal component versus the second principal component. Use different symbols for examples belonging to the digit one**

and to the digit seven. Explain what you see. In this space, are the points corresponding to the digit 1 linearly separable from the points corresponding to the digit 7?

Extra Credit: Carry out the same analysis using only data for the digit 1. Explain what you see.