Andrew Renninger
Homework Three

S1:
- Load EuroSat data
- Visualize few images to check the multi-channel images and labels
- Q: What is the number of images in this dataset: **2700**
- Q: What is the number of labels in this dataset: **5 unique labels ('Forest', 'Highway', 'Industrial', 'Residential', 'River')**
- Q: What is the size of each image in this dataset: **64 x 64 (by 13 bands)**

S2:
- Create data matrixes X and y as follows:
  - Define L1: a list of labels for class 0 (example: 5 and 8)
  - Define L2: a list of labels for class 1 (example: 2, 6 and 7)
  - Define CH: a list of image channels used (example: 7)
  - Select all images for labels in L1 and L2
  - Extract only channels in CH from each selected image and flatten them to a feature matrix X
  - Create the labels y (with binary labels 0 and 1)
- Q: What is the size of X: **1100 x 16384 (though this will be unique to the classes chosen)**
- Q: What is the size of y: **1100**

S3:
- Using X and y create a split dataset with 60% training and 40% testing data with similar distributions for the two classes
- Train a linear SVM classifier on the training data and run it on the testing data
- Draw the ROC for the classification
- Q: What is the training accuracy? **100 (overfit with more features than cases)**
- Q: What is the testing accuracy? **89.5**
- Q: What is the AUC? **99 / 96 / 97 / 98 / 98 / 99 across six folds**

S4:
- Repeat steps S2 and S3 to create a classifier that classifies Industrial vs Residential
- Q: What is the training accuracy? **100**
- Q: What is the testing accuracy? **81.5**
- Q: What is the AUC? **87 / 87 / 83 / 87 / 89 / 87 across six folds**

S5:
- Repeat steps S4 using all image channels as input
- Q: What is the training accuracy? **92.6**
- Q: What is the testing accuracy? **82.5**
- Q: What is the AUC? **89 / 86 / 87 / 90 / 90 / 91 across six folds**

S6:
- Repeat S4 with leave-10%-out cross validation
- Q: What is the training accuracy? **90.5**
- Q: What is the testing accuracy? **84.3**
- Q: What is the AUC? **90 mean across 10 folds**

(I did this section a little differently to experiment and I did not set a seed so the results will vary.)

S7:
- Repeat S2 and S3 to implement a binary classifier between "Residential + Industrial" vs "Others"
- Q: What is the training accuracy? **100**
- Q: What is the testing accuracy? **75.11**
- Q: What is the AUC? **83 mean across 6 folds**

S8:
- Q: Which pair of classes is the one with the highest classification accuracy? **Forest vs Highway**
- Q: Which pair of classes is the one with the lowest classification accuracy? **Industrial vs Residential (without PCA)**

Also included, A neural network did better, though—even with fewer components.