Question 3.1.a
Please review the included r file 3.1.a.R along with this write-up.

After loading data, I randomly sampled 85% of the full data set and assigned it to a training set. I used the remaining 15% as a test set. I used cv.kknn to perform 10-fold cross validation with a variety of ks. I manually iterated over k values using this model and the accuracy test on lines 21-26 of the attached code to determine that maximum accuracy on the cross-validation data was achieved at a k of 7. Accuracy at this k value was 85.58%.

I then initialized a kknn model using the previously determined best k of 7 and generated predictions on the test dataset. I generated a confusion matrix and computed accuracy of this model on the test set. Accuracy on the test dataset was 80.80%, lower than on the cross-validation set.

Question 3.1.b
Please review the included r file 3.1.b.R along with this write-up.

After loading data, I randomly sampled 70% of the full data set and assigned it to a training set. I split the remaining 30% half between a validation set and a test set (meaning that 15% of the full dataset was used in each of the validation and test sets). I initialized a kknn model using the training dataset for training and the validation dataset for testing. I manually iterated over k values using this model and the confusion matrix and accuracy test on lines 22-29 of the attached code to determine that maximum accuracy on validation dataset was achieved at a k of 3. Accuracy at this k value was 81.63%.

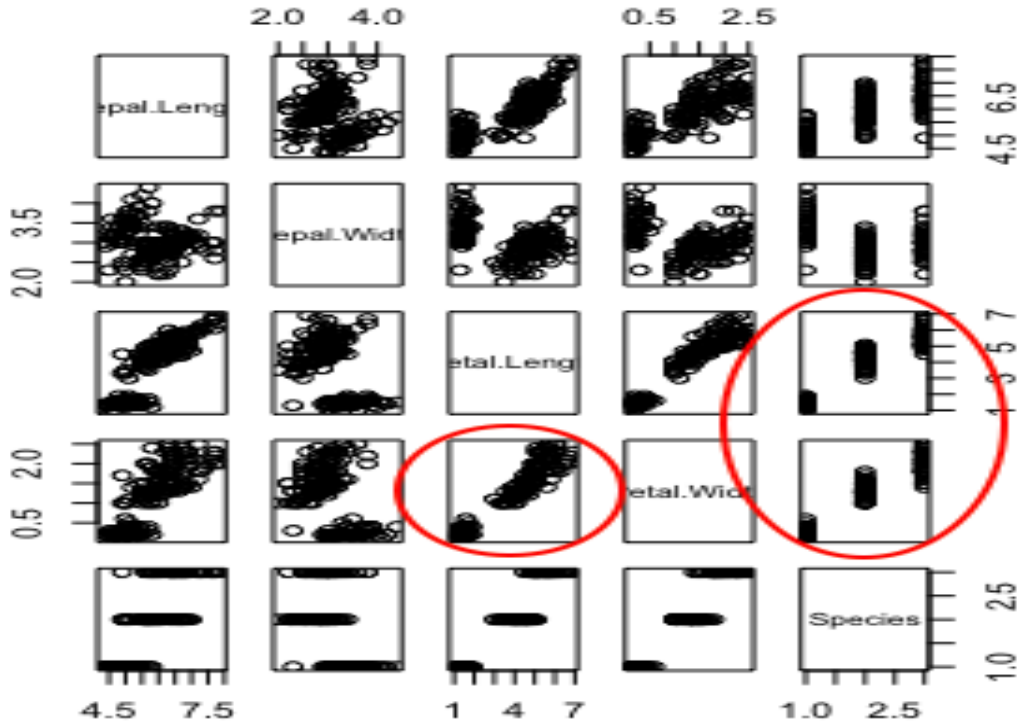
I then initialized a kknn model using the training and test datasets and previously determined best k of 7 and generated predictions on the test dataset. I generated a confusion matrix and computed accuracy of this model on the test set. Accuracy on the test dataset was 77.78%, lower than on the validation set.

It is telling that the k value was different and accuracy lower using a simple train/val/test split rather than cross validation. Given the small size of this dataset, random variance in sampling can have a major impact on results. Using cross validation helps with this on the training and validation set, but the same issue remains on the test set. This can be shown by varying the seed set for the random sample, which impacts the test accuracy.

Question 4.1
I work for a political campaign. A clustering model would be appropriate for segmenting voters into categories to determine whether they would be likely to support certain candidates or issues. Predictors could include income level, political party, age, occupation, or education levels.

Question 4.2
Please review the included R file 4.2.R along with this write-up.

First, I loaded and inspected the data. I then plotted all combos of predictors. From the plot below, it appears that Petal Length and Petal Width are likely to be the best predictors, because they both appear to have a fairly linear relationship between the predictors and the species. There is little overlap between each species in a given range of petal lengths and widths. K of 3 should be the best k value as we know that there are 3 species in the dataset.



I initialized the model and tried a variety of predictors. The best results were with k=3 and Petal Length and Petal Width as predictors, as expected. This combination provided 94.3% accuracy. I then compared the actual clusters by species with the predicted clusters from this model. The table below shows the count of actual and predicted species for each of the 150 data points.

Predicted Clusters	Actual Species		
	setosa	versicolor	virginica
setosa	50	0	0
virginia	0	2	46
versicolor	0	48	4

The graphs below show the species data points with and without predicted clusters outlined. The outlines on the second graph show the outline of the model's predicted clusters while the

points are colored by actual species. This shows that the model predicted clusters are generally very close to the actual species, correctly classifying all 50 setosa data points and 94 out of 100 of the remaining points.