CLUSTERING

Here we will use both hierarchical clustering and kmeans clustering to cluster the penDigits dataset. We'll use a subset of this data, using only the digits 1, 4, and 8. Even in that subset, we'll downsample the data so that we can draw hierarchical clustering diagrams that are easier to view. We'll keep the digits in order to help with the visualization as well.

```
> load("PenDigits.Rdata")
> pendig = rbind(train,test)
> subset1 = pendig[pendig$digit==1 | pendig$digit==4 | pendig$digit==8, ]
> subset1 = subset1[order(subset1$digit),]
> randomIndex=sample(c(T,F), size=3342, replace=T, prob=c(.1,.9))
> PDsubset = subset1[randomIndex,1:16]
> PDsubsetdigit=subset1[randomIndex,"digit"]
> PDsubsetdigit=droplevels(PDsubsetdigit) #Drops empty levels of the factor digit.
```

Hierarchical clustering is done with the hclust() function from base R.

```
> PDdist=dist(PDsubset)
> single = hclust(PDdist, method="single")
> plot(single)
```

k-means clustering can be done with the built in kmeans() function.

```
> clust=kmeans(PDsubset, 3)
```

We can see how well the digits clustered together by examining a table of digit by cluster number:

```
> table(PDsubsetdigit,clust$cluster)
```

```
PDsubsetdigit   1   2   3
            1   0   0 107
            4   0 123   0
            8  94   0  19
```

Clustering did well to separate the 1's and the 4's but confused some of the 8's with 1's.