Homework 11

For this homework you will create an R Markdown document that outputs to a PDF and you will upload both the .Rmd and .pdf files to wolfware. Be sure to put your name in the title of the document.

The purpose of this homework is to get practice fitting kNN models using the `caret` package.

The article here gives a great example of selecting the number of neighbors to use with the `caret` package. They use repeated 10 fold cross-validation - there is of course variability in the prediction by cross-validation, repated CV gives a more stable prediction but takes a lot of computational time!

I'd like you to read through that article and then apply the same methods to the `titanicTrainData` data set from our notes/video on kNN (code below). Use the `tuneGrid` option on `train` to look at all neighbor values from 2 to 30.

Once you've settled on the appropriate number of neighbors, then predict and compare your misclassification rate on the `titanicTestData` data set.

Note: you don't need to partition the data into training and test sets using `createDataPartition` as we will do that ourselves. Also, `caret` will do the standardization of your variables for you so we won't recreate that part of the data set!

Explain your steps as you go through the relevant parts (but don't copy and paste their explanations).

Relevant code to create your data sets:

```r
#read data/clean it
titanicData <- read_csv("../datasets/titanic.csv")
titanicData <- filter(titanicData, !is.na(survived) & !is.na(fare) & !is.na(age))
titanicData$survived <- as.factor(titanicData$survived)

set.seed(1)
train <- sample(1:nrow(titanicData), size = nrow(titanicData)*0.8)
test <- dplyr::setdiff(1:nrow(titanicData), train)

titanicDataTrain <- titanicData[train, ]
titanicDataTest <- titanicData[test, ]
```