Untitled

#For this assignment, we will be working with two different datasets. For problem # 1, we will still be working with the Penguin dataset.

```
library(palmerpenguins)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(nnet)
library(forcats)
library(knitr)
```

```
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
```

```
##                                 NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

```
penguins_df = na.omit(penguins)
summary(penguins_df)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :146   Biscoe   :163   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :123   1st Qu.:39.50   1st Qu.:15.60
##  Gentoo   :119   Torgersen: 47   Median :44.50   Median :17.30
##                                  Mean   :43.99   Mean   :17.16
##                                  3rd Qu.:48.60   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172       Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190       1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197       Median :4050                Median :2008
##  Mean   :201       Mean   :4207                Mean   :2008
##  3rd Qu.:213       3rd Qu.:4775                3rd Qu.:2009
##  Max.   :231       Max.   :6300                Max.   :2009
```

#Remove N/a from penguins dataset and remove the variable "year"

```
org_penguins_df <- penguins_df
penguins_df <- penguins_df %>%
  dplyr::select(-year) %>%
  mutate_if(is.factor, as.numeric)
penguins_df$species <- org_penguins_df$species
print(penguins_df)
```

```
## # A tibble: 333 x 7
##    species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
##    <fct>    <dbl>          <dbl>         <dbl>            <int>       <int>
##  1 Adelie       3           39.1          18.7              181        3750
##  2 Adelie       3           39.5          17.4              186        3800
##  3 Adelie       3           40.3          18                195        3250
##  4 Adelie       3           36.7          19.3              193        3450
##  5 Adelie       3           39.3          20.6              190        3650
##  6 Adelie       3           38.9          17.8              181        3625
##  7 Adelie       3           39.2          19.6              195        4675
##  8 Adelie       3           41.1          17.6              182        3200
##  9 Adelie       3           38.6          21.2              191        3800
## 10 Adelie       3           34.6          21.1              198        4400
## # ... with 323 more rows, and 1 more variable: sex <dbl>
```

```
#Normalization
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x))) }
```

```
penguins_df_norm <- as.data.frame(lapply(penguins_df[3:6], normalize))
```

```
penguins_df_norm$sex <- penguins_df$sex
penguins_df_norm$island <- penguins_df$island
penguins_df_norm$species <- penguins_df$species
head(penguins_df_norm)
```

```
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex island species
## 1      0.2545455     0.6666667         0.1525424   0.2916667   2      3  Adelie
## 2      0.2690909     0.5119048         0.2372881   0.3055556   1      3  Adelie
## 3      0.2981818     0.5833333         0.3898305   0.1527778   1      3  Adelie
## 4      0.1672727     0.7380952         0.3559322   0.2083333   1      3  Adelie
## 5      0.2618182     0.8928571         0.3050847   0.2638889   2      3  Adelie
## 6      0.2472727     0.5595238         0.1525424   0.2569444   1      3  Adelie
```

```
set.seed(1234)
ind <- sample(2, nrow(penguins_df_norm), replace=TRUE, prob=c(0.70, 0.30))
```

```
penguin.train <- penguins_df_norm[ind==1, 1:6]
# Inspect training set
head(penguin.train)
```

```
##   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex island
## 1      0.2545455     0.6666667         0.1525424   0.2916667   2      3
## 2      0.2690909     0.5119048         0.2372881   0.3055556   1      3
## 3      0.2981818     0.5833333         0.3898305   0.1527778   1      3
## 4      0.1672727     0.7380952         0.3559322   0.2083333   1      3
## 6      0.2472727     0.5595238         0.1525424   0.2569444   1      3
## 7      0.2581818     0.7738095         0.3898305   0.5486111   2      3
```

```
# Compose test set
penguin.test <- penguins_df_norm[ind==2, 1:6]
# Inspect test set
head(penguin.test)
```

```
##    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex island
## 5      0.26181818     0.8928571        0.30508475   0.2638889   2      3
## 14     0.08363636     0.6309524        0.20338983   0.1736111   1      3
## 16     0.20727273     0.6190476        0.03389831   0.1944444   1      1
## 26     0.26909091     0.4285714        0.10169492   0.1527778   1      2
## 28     0.26909091     0.5595238        0.27118644   0.1666667   1      2
## 29     0.32000000     0.6904762        0.20338983   0.3333333   2      2
```

```
penguin.trainLabels <- penguins_df_norm[ind==1,7]
# Inspect result
print(penguin.trainLabels)
```

```
##   [1] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##   [8] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [15] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [22] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [29] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [36] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [43] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [50] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [57] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
```

```
##  [64] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [71] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [78] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [85] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [92] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [99] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [106] Adelie    Adelie    Adelie    Adelie    Adelie    Gentoo    Gentoo
## [113] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [120] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [127] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [134] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [141] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [148] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [155] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [162] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [169] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [176] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [183] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [190] Gentoo    Gentoo    Gentoo    Gentoo    Chinstrap Chinstrap Chinstrap
## [197] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [204] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [211] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [218] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [225] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [232] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [239] Chinstrap
## Levels: Adelie Chinstrap Gentoo
```

```
# Compose `penguin` test labels
penguin.testLabels <- penguins_df_norm[ind==2, 7]
# Inspect result
print(penguin.testLabels)
```

```
##  [1] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [8] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [15] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [22] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [29] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [36] Adelie    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [43] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [50] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [57] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [64] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [71] Gentoo    Gentoo    Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [78] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [85] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [92] Chinstrap Chinstrap Chinstrap
## Levels: Adelie Chinstrap Gentoo
```

```
library(class)
NROW(penguin.trainLabels)
```

```
## [1] 239
```

So, we have 239 observations in our training data set. The square root of 239 is around 15.35, therefore we'll create two models. One with 'K' value as 15 and the other model with a 'K' value as 16.

Please use K-nearest neighbor (KNN) algorithm to predict the species variable. Please be sure to walk through the steps you took. (40 points)

#Model Evaluation

```
penguin_pred <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=1)

penguin_pred
```

```
##  [1] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
##  [8] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [15] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [22] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [29] Adelie    Adelie    Adelie    Adelie    Adelie    Adelie    Adelie
## [36] Adelie    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [43] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [50] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [57] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [64] Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo    Gentoo
## [71] Gentoo    Gentoo    Chinstrap Chinstrap Adelie    Chinstrap Chinstrap
## [78] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [85] Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap Chinstrap
## [92] Chinstrap Chinstrap Chinstrap
## Levels: Adelie Chinstrap Gentoo
```

```
# Confusiin Matrix
cm <- table(penguin.testLabels, penguin_pred)
cm
```

```
##                   penguin_pred
## penguin.testLabels Adelie Chinstrap Gentoo
##          Adelie        36         0      0
##          Chinstrap      1        21      0
##          Gentoo         0         0     36
```

So, 36 Adelie are correctly classified as Adelie.

Out of 22 Chinstrap, 21 Chinstrap are correctly classified as Chinstrap and 1 is classified as Adelie.

36 Gentoo are correctly classified as Gentoo.

```
misClassError <- mean(penguin_pred != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 3

```
penguin_pred_3 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=3)
misClassError_3 <- mean(penguin_pred_3 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_3))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 5

```
penguin_pred_5 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=5)

misClassError_5 <- mean(penguin_pred_5 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_5))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 7

```
penguin_pred_7 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=7)

misClassError_7 <- mean(penguin_pred_7 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_7))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 9

```
penguin_pred_9 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=9)

misClassError_9 <- mean(penguin_pred_9 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_9))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 11

```
penguin_pred_11 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=11)

misClassError_11 <- mean(penguin_pred_11 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_11))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 13

```
penguin_pred_13 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=13)

misClassError_13 <- mean(penguin_pred_13 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_13))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 15

```
penguin_pred_15 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=15)

misClassError_15 <- mean(penguin_pred_15 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_15))
```

```
## [1] "Accuracy = 0.98936170212766"
```

K = 16

```
penguin_pred_16 <- knn(train = penguin.train, test = penguin.test, cl = penguin.trainLabels, k=16)

misClassError_16 <- mean(penguin_pred_16 != penguin.testLabels)
print(paste('Accuracy =', 1-misClassError_16))
```

```
## [1] "Accuracy = 1"
```