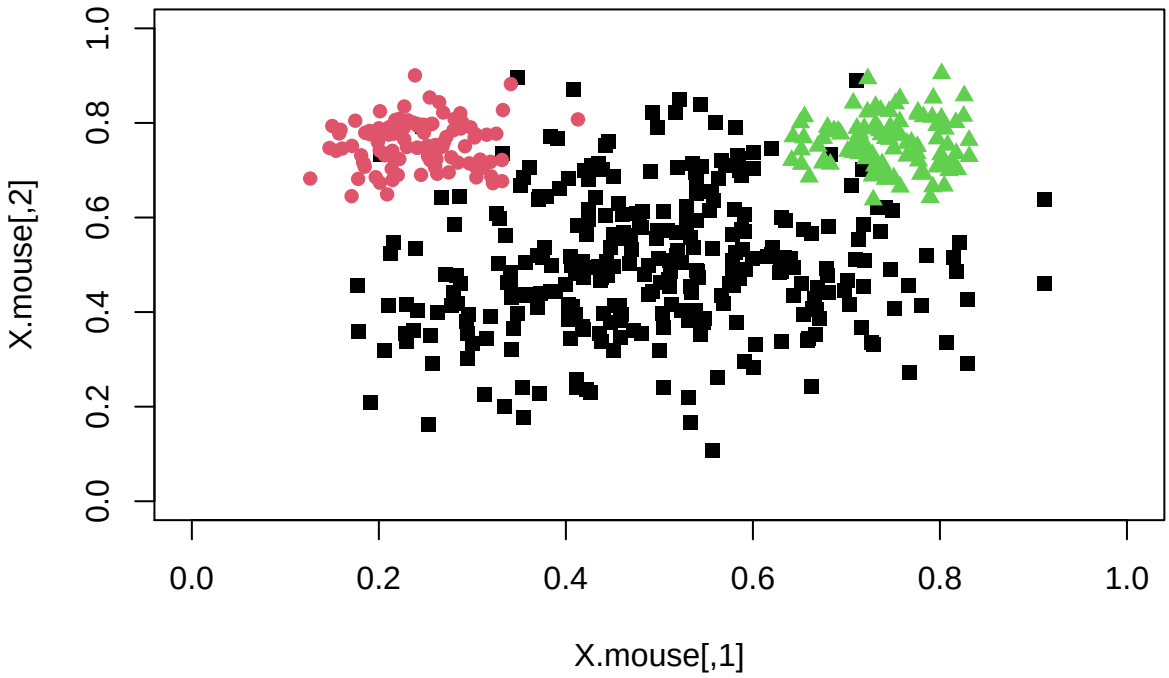
Problem 6.1

Chen Bo Calvin Zhang

09/11/2020

First generate the data.

```
# generate the "mouse" data
library("mnormt")

# class one (head)
n1 = 300
mu1 = c(0.5,0.5)
Sigma1 = matrix(c(0.15^2, 0, 0, 0.15^2), 2)
x1 = rmnorm(n1,mean=mu1,var=Sigma1)

# class two (left ear)
n2=100
mu2 = c(0.25,0.75)
Sigma2 = matrix(c(0.05^2, 0, 0, 0.05^2), 2)
x2 = rmnorm(n2,mean=mu2,var=Sigma2)

# class three (right ear)
n3 = 100
mu3 = c(0.75,0.75)
Sigma3 = matrix(c(0.05^2, 0, 0, 0.05^2), 2)
x3 = rmnorm(n3,mean=mu3,var=Sigma3)

# put all data together in one matrix
X.mouse = rbind(x1,x2,x3)
L.mouse = factor(c( rep("Head", n1), rep("Left Ear", n2), rep("Right Ear", n3) ))

plot(X.mouse, col=as.integer(L.mouse),
     pch=as.integer(L.mouse)+14, ylim=c(0,1), xlim=c(0,1))
```



Now, perform K-means with K = 3.

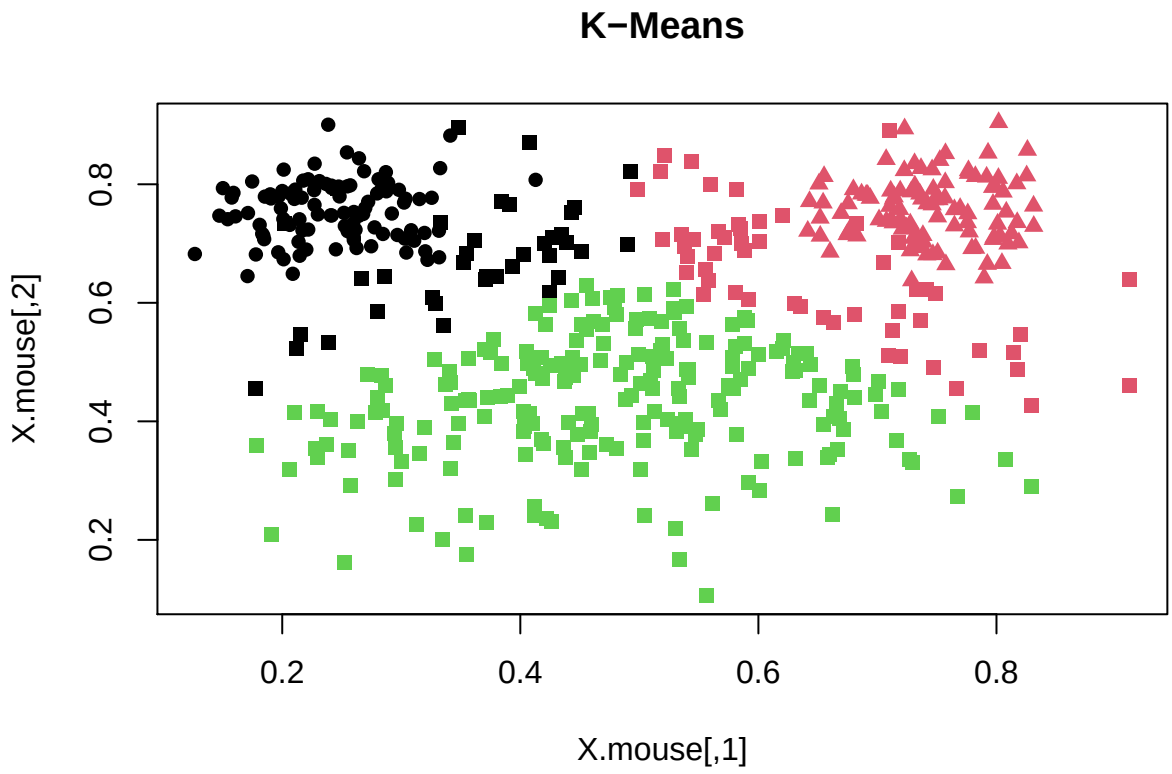
```
kmeans.out3 = kmeans(X.mouse, centers = 3)
kmeans.out3
```

```
## K-means clustering with 3 clusters of sizes 138, 153, 209
##
## Cluster means:
##        [,1]      [,2]
## 1 0.2761190 0.7337529
## 2 0.7114315 0.7193387
## 3 0.4806691 0.4318485
##
## Clustering vector:
##   [1] 3 3 2 3 3 3 3 3 3 1 1 3 3 3 3 3 2 2 2 3 3 3 1 2 3 3 3 3 1 1 3 2 3 2 3 2 1
##  [38] 2 2 3 3 3 2 3 1 3 2 3 3 1 1 3 3 3 2 3 3 3 2 3 3 2 3 3 3 3 2 3 3 3 3 3 2 3
##  [75] 3 1 3 2 1 3 3 3 3 3 1 3 3 3 3 3 3 3 2 3 1 3 2 2 3 3 3 3 3 2 3 3 3 2 3 3 3
## [112] 3 3 2 3 3 3 2 3 3 3 3 2 3 3 2 3 2 2 3 3 3 3 3 1 3 2 3 3 2 3 1 3 3 3 2 1 2
## [149] 1 2 3 3 3 3 1 3 1 3 3 3 3 3 3 1 3 3 3 1 3 2 3 3 3 3 3 3 2 3 2 1 3 3 2 3 3
## [186] 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 1 3 3 3 2 1 3 2 3 3 2 3 3 3 3 2 3 3 2
## [223] 3 1 2 3 1 3 3 2 3 3 3 1 3 2 1 3 1 3 3 3 1 3 3 3 3 2 2 3 3 3 1 1 3 3 1 2 3
## [260] 3 3 3 3 3 3 3 1 3 3 3 1 3 3 3 1 3 3 3 3 2 2 3 1 3 3 3 3 3 1 3 3 3 3 3 3 3
## [297] 3 3 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
## [408] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

```
## [445] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [482] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1] 1.636527 2.439395 5.900844
##  (between_SS / total_SS =  71.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

And plot the clusters obtained with K-means.

```
plot(X.mouse, col = kmeans.out3$cluster, pch = as.integer(L.mouse) + 14, main = "K-Means")
```



Compare the predicted clusters with the original ones.

```
table(L.mouse, kmeans.out3$cluster)
```

```
##
## L.mouse       1   2   3
##   Head       38  53 209
##   Left Ear  100   0   0
##   Right Ear   0 100   0
```

Let us apply K-means with varying values of K and check the between group and within group variation.

```
between_var = numeric(10)
within_var = numeric(10)

for (k in 1:10)
{
  kmeans.out = kmeans(X.mouse, centers = k)
  between_var[k] = kmeans.out$betweenss
  within_var[k] = kmeans.out$tot.withinss
}
```

Lastly, let us plot the variations.

```
plot(1:10, between_var, ylim=c(0, 34), type = "b", xlab = "K",
     ylab = "Variation", main = "K-means Mouse Data")
points(1:10, within_var, type = "b", col = 2, pch = 2)
legend("right", c("Between SS (explained)", "Within SS (unexplained)"),
       col=c(1,2), pch=c(1,2))
```