Problem 2:

a)

```
> set.seed(2)
> x <- matrix(rnorm(20 * 3 * 50, mean = 0, sd = 0.001), ncol = 50)
> x[1:20, 2] <- 1
> x[21:40, 1] <- 2
> x[21:40, 2] <- 2
> x[41:60, 1] <- 1
> true.labels <- c(rep(1, 20), rep(2, 20), rep(3, 20))
```

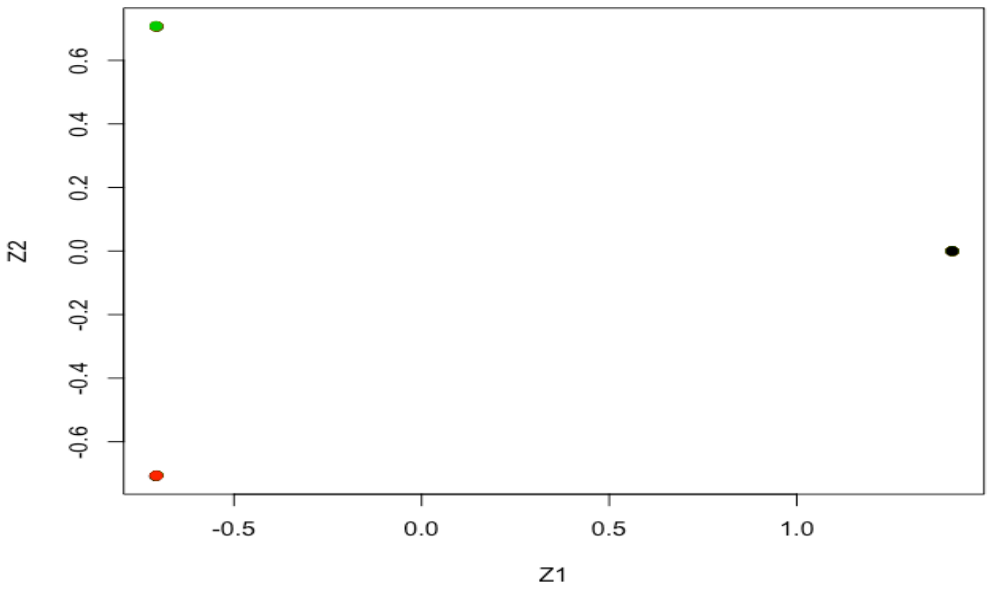
Generated a simulated data set with 20 observations in each of the three classes (i.e. 60 observations total), and 50 variables.

b) After performing PCA on the 60 observations:

```
> summary(pca)
Importance of components:
                          PC1    PC2      PC3      PC4      PC5      PC6      PC7      PC8      PC9     PC10     PC11     PC12
Standard deviation     1.008 0.5821 0.001731 0.001673 0.001648 0.001582 0.001543 0.001497 0.001474 0.001411 0.001393 0.001335
Proportion of Variance 0.750 0.2499 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
Cumulative Proportion  0.750 1.0000 0.999970 0.999970 0.999970 0.999970 0.999980 0.999980 0.999980 0.999980 0.999980 0.999980
                          PC13     PC14     PC15     PC16    PC17     PC18     PC19     PC20     PC21      PC22      PC23
Standard deviation     0.001297 0.001257 0.001244 0.001226 0.00116 0.001118 0.001091 0.001021 0.001012 0.0009849 0.0009378
Proportion of Variance 0.000000 0.000000 0.000000 0.000000 0.00000 0.000000 0.000000 0.000000 0.000000 0.0000000 0.0000000
Cumulative Proportion  0.999980 0.999990 0.999990 0.999990 0.99999 0.999990 0.999990 0.999990 0.999990 0.9999900 0.9999900
                          PC24      PC25      PC26      PC27     PC28      PC29      PC30      PC31      PC32      PC33
Standard deviation     0.0009316 0.0009081 0.0008668 0.0008228 0.000801 0.0007486 0.0007124 0.0006966 0.0006733 0.0006323
Proportion of Variance 0.0000000 0.0000000 0.0000000 0.0000000 0.000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
Cumulative Proportion  0.9999900 0.9999900 1.0000000 1.0000000 1.000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
                          PC34      PC35      PC36      PC37      PC38      PC39      PC40      PC41      PC42      PC43
Standard deviation     0.0005909 0.0005654 0.0005381 0.0005325 0.0004756 0.0004476 0.0004261 0.0003914 0.0003774 0.0003144
Proportion of Variance 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
Cumulative Proportion  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
                          PC44      PC45      PC46      PC47      PC48     PC49      PC50
Standard deviation     0.0002964 0.0002732 0.0002495 0.0001915 0.0001466 0.000129 7.787e-05
Proportion of Variance 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.000000 0.000e+00
Cumulative Proportion  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.000000 1.000e+00
```

```
> pca$x[,1:2]
             PC1           PC2
 [1,] -0.7079228 -7.076535e-01
 [2,] -0.7071573 -7.068897e-01
 [3,] -0.7061651 -7.058937e-01
 [4,] -0.7080866 -7.078204e-01
 [5,] -0.7073449 -7.070720e-01
 [6,] -0.7071940 -7.069282e-01
 [7,] -0.7067857 -7.065196e-01
 [8,] -0.7074565 -7.071901e-01
 [9,] -0.7058831 -7.056148e-01
[10,] -0.7073860 -7.071167e-01
[11,] -0.7069927 -7.067245e-01
[12,] -0.7065928 -7.063271e-01
[13,] -0.7075651 -7.072968e-01
[14,] -0.7080215 -7.077547e-01
[15,] -0.7060274 -7.057605e-01
[16,] -0.7089206 -7.086542e-01
[17,] -0.7066636 -7.063975e-01
[18,] -0.7072613 -7.069955e-01
[19,] -0.7065710 -7.063041e-01
[20,] -0.7069825 -7.067125e-01
[21,]  1.4140879 -7.088995e-05
[22,]  1.4140874 -6.893673e-05
[23,]  1.4140878 -7.309105e-05
[24,]  1.4140860 -6.934807e-05
[25,]  1.4140859 -7.045088e-05
[26,]  1.4140878 -6.983408e-05
[27,]  1.4140866 -7.447988e-05
[28,]  1.4140855 -6.910345e-05
[29,]  1.4140863 -7.290071e-05
[30,]  1.4140863 -7.054766e-05
[31,]  1.4140874 -6.989590e-05
[32,]  1.4140872 -7.065443e-05
[33,]  1.4140855 -7.013529e-05
[34,]  1.4140874 -7.005627e-05
[35,]  1.4140877 -7.296213e-05
[36,]  1.4140863 -6.608230e-05
[37,]  1.4140894 -7.285675e-05
[38,]  1.4140867 -6.920393e-05
[39,]  1.4140872 -6.744255e-05
[40,]  1.4140861 -6.949980e-05
[41,] -0.7064221  7.064353e-01
[42,] -0.7069992  7.070155e-01
[43,] -0.7074064  7.074178e-01
[44,] -0.7077105  7.077255e-01
[45,] -0.7077918  7.078046e-01
[46,] -0.7057358  7.057486e-01
[47,] -0.7065186  7.065305e-01
[48,] -0.7057616  7.057767e-01
[49,] -0.7074790  7.074930e-01
[50,] -0.7074294  7.074432e-01
[51,] -0.7079055  7.079200e-01
[52,] -0.7073606  7.073721e-01
[53,] -0.7068480  7.068625e-01
[54,] -0.7062227  7.062363e-01
[55,] -0.7067821  7.067967e-01
[56,] -0.7068591  7.068727e-01
[57,] -0.7063127  7.063250e-01
[58,] -0.7063711  7.063849e-01
[59,] -0.7071057  7.071214e-01
[60,] -0.7077361  7.077523e-01
```

- Separated three classes amongst two dimensions.



c) K-means clustering where k=3:

```
> km.out <- kmeans(x, 3, nstart = 20)
> table(true.labels, km.out$cluster)

true.labels  1  2  3
           1 20  0  0
           2  0 20  0
           3  0  0 20
```

After K-means clustering of the observations with K=3, we can see that clusters obtained in K-means clustering are perfectly clustered compared to the true class labels.

d) K-means clustering where k=2:

```
> km.out <- kmeans(x, 2, nstart = 20)
> table(true.labels, km.out$cluster)

true.labels  1  2
           1 20  0
           2  0 20
           3 20  0
```

After K-means clustering of the observations with K=2, we can see that all observations of the one of the three clusters obtained in K-means clustering are absorbed in one of the two clusters.

e) K-means clustering where k=4:

```
> km.out <- kmeans(x, 4, nstart = 20)
> table(true.labels, km.out$cluster)

true.labels  1  2  3  4
           1  9  0  0 11
           2  0 20  0  0
           3  0  0 20  0
```

After K-means clustering of the observations with K=4, we can see the first cluster gets split into two different clusters.

f) Performing K-means clustering with k=3 on the first two principal component core vectors:

```
> km <- kmeans(pca$x[, 1:2], 3, nstart = 20)
> table(true.labels, km$cluster)

true.labels  1  2  3
           1  0  0 20
           2  0 20  0
           3 20  0  0
```

After K-means clustering of the observations with K=3 on the first two principal component core vectors, we can see that clusters obtained in K-means clustering are perfectly clustered compared to the true class labels.

g) Performing K-means clustering with k=3 on the data after scaling each variable to have standard deviation one:

```
> km.out <- kmeans(scale(x), 3, nstart = 20)
> table(true.labels, km.out$cluster)

true.labels  1  2  3
           1  8  2 10
           2  0 19  1
           3 11  1  8
```

When clustering on the scaled data we can see that we get worse results than with unscaled data, as the distance between the observations is affected by scaling.