Assignment 3

Seongming Lee (01247436), Yuxuan Luo (01376247) and Nina Hauser (01418616)

Question 1

To calculate principe components, we first need to transform the faces data(A=4096 x 400) to covariance matrix(400 x 400), and then compute eigenvalue and eigenvector from the covariance matrix. And each priciple component, which represents the significance of each face, can be calculated by eigenvectors of the covariance matrix multiplying the face matrix. Hence, each face ϕ_i in the training set can be represented as a linear combination of K eigenvectors, formulated as $\phi_i = \mu + \sum_{k=1}^{k} \lambda_{1:k} v_{1:k}$ where μ is the average face, v is the eigenvectors(eigenfaces), and λ is eigenvalue. As K keeps increasing to add up more principle components, the explained variance ratio($\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i}$) is steadliy rising up to 1, which means faces are getting closer to their real faces.

```
## With k =  3
## Explained Variance Ratio:  0.4706023
```

```
## With k =  10
## Explained Variance Ratio:  0.6605544
```

```
## With k =  25
## Explained Variance Ratio:  0.7952196
```

```
## With k =  50
## Explained Variance Ratio:  0.8742574
```

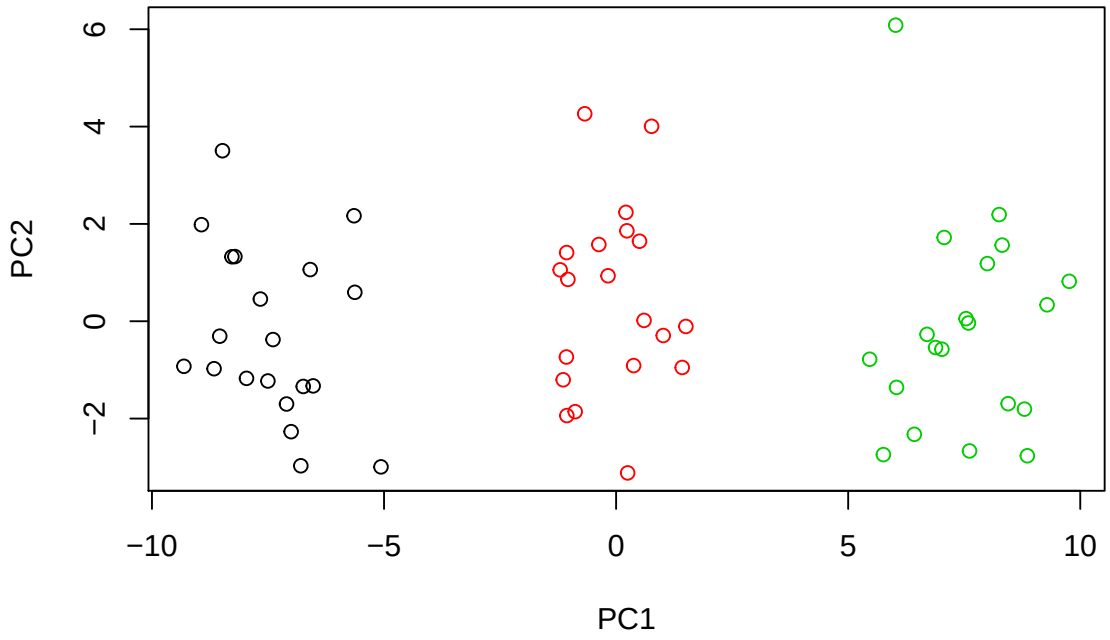
Question 2

(a) Generate a simulated data set (60*50).

```
x <- rbind(matrix(rnorm(20*50, mean = 1), nrow = 20),
           matrix(rnorm(20*50, mean=2), nrow = 20),
           matrix(rnorm(20*50, mean=3), nrow = 20))
print(dim(x))
```

```
## [1] 60 50
```

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors.



(c) The K-means obtained clusters are perfectly assigned to the true class labels.

```
##    true_class
##       1  2  3
##   1 20  0  0
##   2  0  0 20
##   3  0 20  0
```

(d) One whole class is assigned to a wrong class, while other classes are classified correctly. This is reasonable since the third cluster is generated with a mean shift of 3, while the mean shift for the second cluster is 2 and for the first class is 1.

```
##    true_class
##       1  2  3
##   1 20 20  0
##   2  0  0 20
```

(e) K-means clustering correctly classifys two clusters, but the other class is divided into two classes.

```
##    true_class
##       1  2  3
##   1 20  0  0
##   2  0  0 10
##   3  0 20  0
##   4  0  0 10
```

(f) Same result with (b), as the first two principal component score vectors carries enough information to cluster correctly.

```
##    true_class
##       1  2  3
##   1 20  0  0
##   2  0  0 20
##   3  0 20  0
```

(g) Same with (b), scaling does not change the results, since original data is generated by rnorm() with mean shifts and standard deviation of 1, and scaling is simply transforming original data into data with mean of 0 and standard deviation of 1.

```
##    true_class
##       1  2  3
##   1 20  0  0
##   2  0  0 20
##   3  0 20  0
```