Homework 10: Principle Component Analysis

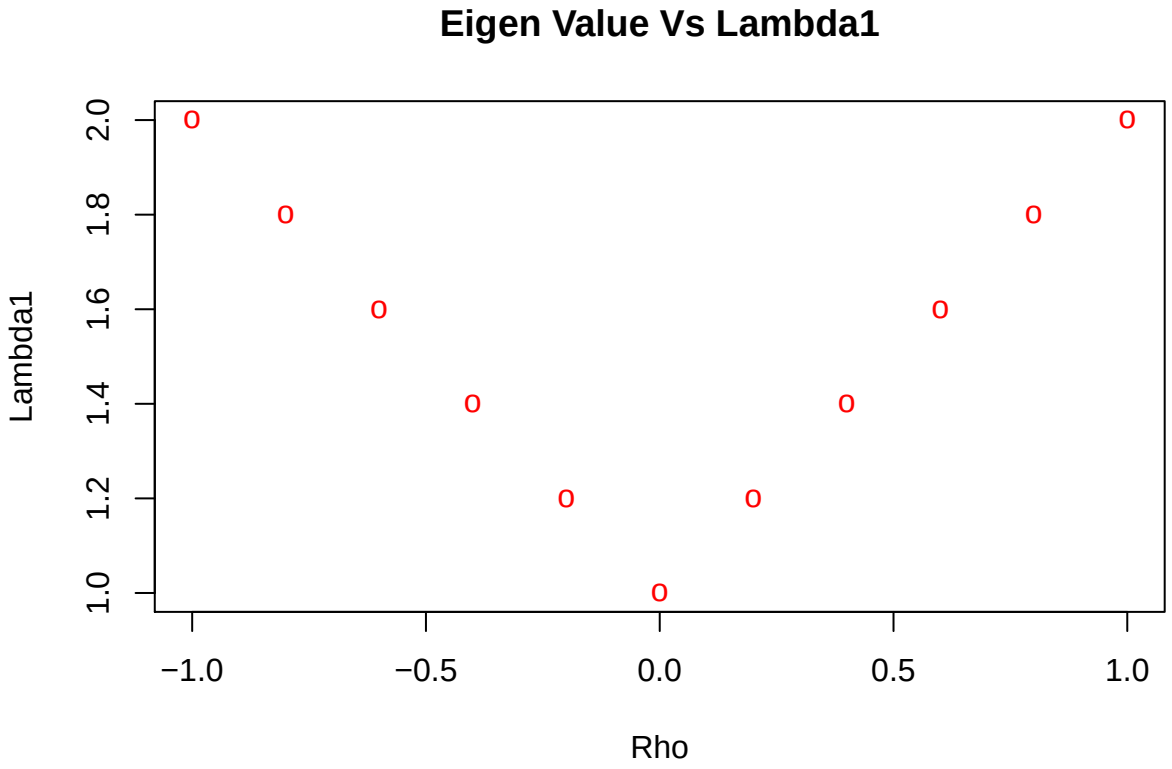
Bijesh Mishra

```
rm(list = ls())
library(MASS)
```

Q1.A: Construct a plot of ρ (-1 to 1) Vs λ_1 (First eigen value for the eigen decomposition). Label y-axis as "Lambda1". Explain when principle component analysis will be most effective at dimension reduction.

```
rhoo = seq(-1, 1, 0.2)
lambda = NULL
for (i in 1:length(rhoo)){
        sigmaaa = matrix(c(1, rhoo[i], rhoo[i], 1), 2, 2, byrow = TRUE)
        decompoo = eigen(sigmaaa)
        lambda[i] = decompoo$value[1]
}

plot(rhoo, lambda,
     xlab = "Rho", ylab = "Lambda1",
     pch = "o", col = "red",
     main = "Eigen Value Vs Lambda1")
```



Answer: Lambda1 (λ_1) is the variance of first principle component which is first eigen value and ρ is COV(X_1,X_2). When $\rho = 1$ or -1 (ie. perfectly correlated in magnitude) and Var(X_1) = 1 and Var(X_2) = 1

then we can use either X_1 or X_2 to explain maximum variation as given by principle component Z_1. So, this is the most effective condition to reduce the dimension as one of the variable inluded in PCA can explain all variability in the data.

Q1B.I Construct a scatter plot for each dataset.

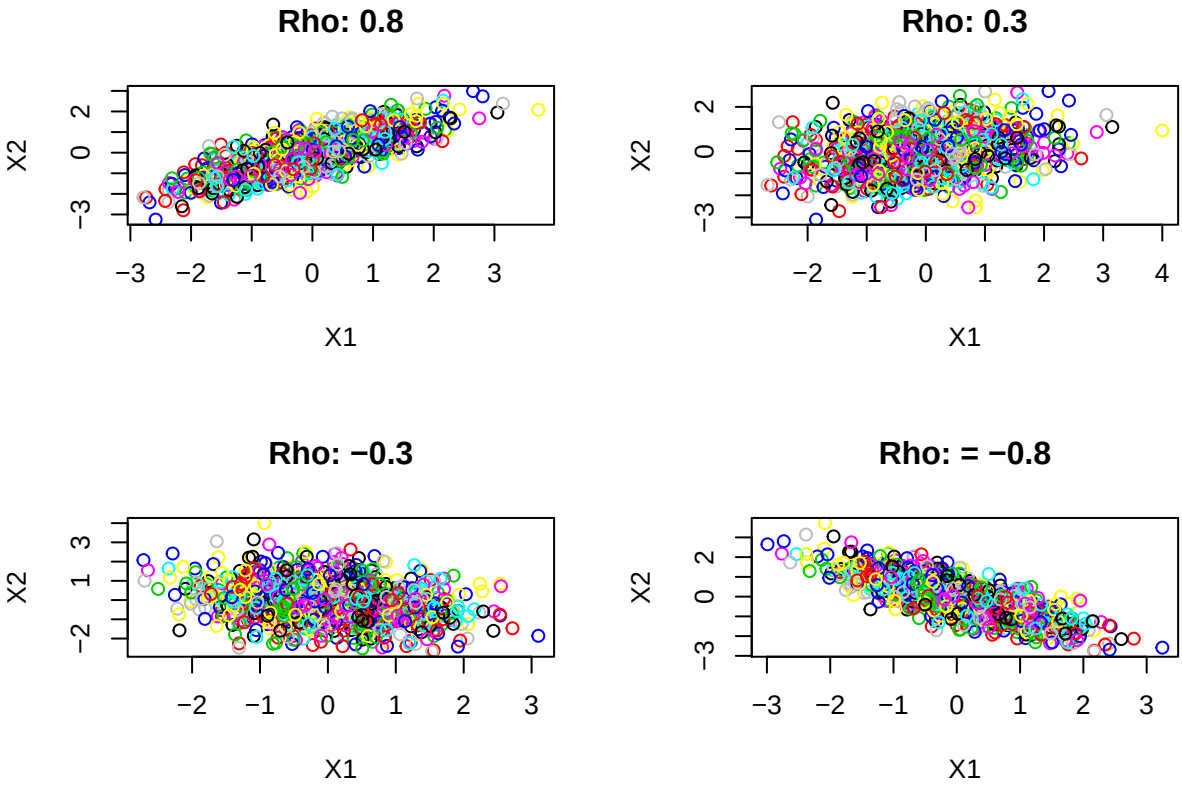
```
set.seed(999)
sigma.1b1 = matrix (c(1, 0.8, 0.8, 1), nrow = 2, ncol = 2) # Covariance matrix: 0.8
x.1b1 = mvrnorm(1000, mu = c(0,0), Sigma = sigma.1b1)

set.seed(999)
sigma.1b2 = matrix (c(1, 0.3, 0.3, 1), nrow = 2, ncol = 2) # Covariance matrix: 0.3
x.1b2 = mvrnorm(1000, mu = c(0,0), Sigma = sigma.1b2)

set.seed(999)
sigma.1b3 = matrix (c(1, -0.3, -0.3, 1), nrow = 2, ncol = 2) # Covariance matrix: -0.3
x.1b3 = mvrnorm(1000, mu = c(0,0), Sigma = sigma.1b3)

set.seed(999)
sigma.1b4 = matrix (c(1, -0.8, -0.8, 1), nrow = 2, ncol = 2) # Covariance matrix: -0.8
x.1b4 = mvrnorm(1000, mu = c(0,0), Sigma = sigma.1b4)

par(mfrow = c(2,2))
plot(x.1b1, xlab = "X1", ylab = "X2", col = c(1:1000), main = "Rho: 0.8")
plot(x.1b2, xlab = "X1", ylab = "X2", col = c(1:1000), main = "Rho: 0.3")
plot(x.1b3, xlab = "X1", ylab = "X2", col = c(1:1000), main = "Rho: -0.3")
plot(x.1b4, xlab = "X1", ylab = "X2", col = c(1:1000), main = "Rho: = -0.8")
```



Q1B.II

```r
eigen(sigma.1b1) # 0.8
```

```
## eigen() decomposition
## $values
## [1] 1.8 0.2
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

```r
eigen(sigma.1b2) # 0.3
```

```
## eigen() decomposition
## $values
## [1] 1.3 0.7
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

```r
eigen(sigma.1b3) # -0.3
```

```
## eigen() decomposition
## $values
## [1] 1.3 0.7
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,]  0.7071068 -0.7071068
```

```r
eigen(sigma.1b4) # -0.8
```

```
## eigen() decomposition
## $values
## [1] 1.8 0.2
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,]  0.7071068 -0.7071068
```

- First PC for rho 0.8: $Z_1 = 0.707\ X_1 + 0.707\ X_2$
- Variance of Ist PC of Rho 0.8: $(\lambda_1) = 1.8$
- First PC for rho 0.3: $Z_1 = 0.707\ X_1 + 0.707\ X_2$
- Variance of Ist PC of Rho 0.3: $(\lambda_1) = 1.3$
- First PC for rho -0.3: $Z_1 = -0.707\ X_1 + 0.707\ X_2$
- Variance of Ist PC of Rho -0.3: $(\lambda_1) = 1.3$
- First PC for rho -0.8: $Z_1 = -0.707\ X_1 + 0.707\ X_2$
- Variance of Ist PC of Rho -0.8: $(\lambda_1) = 1.8$

Q1.B.III:

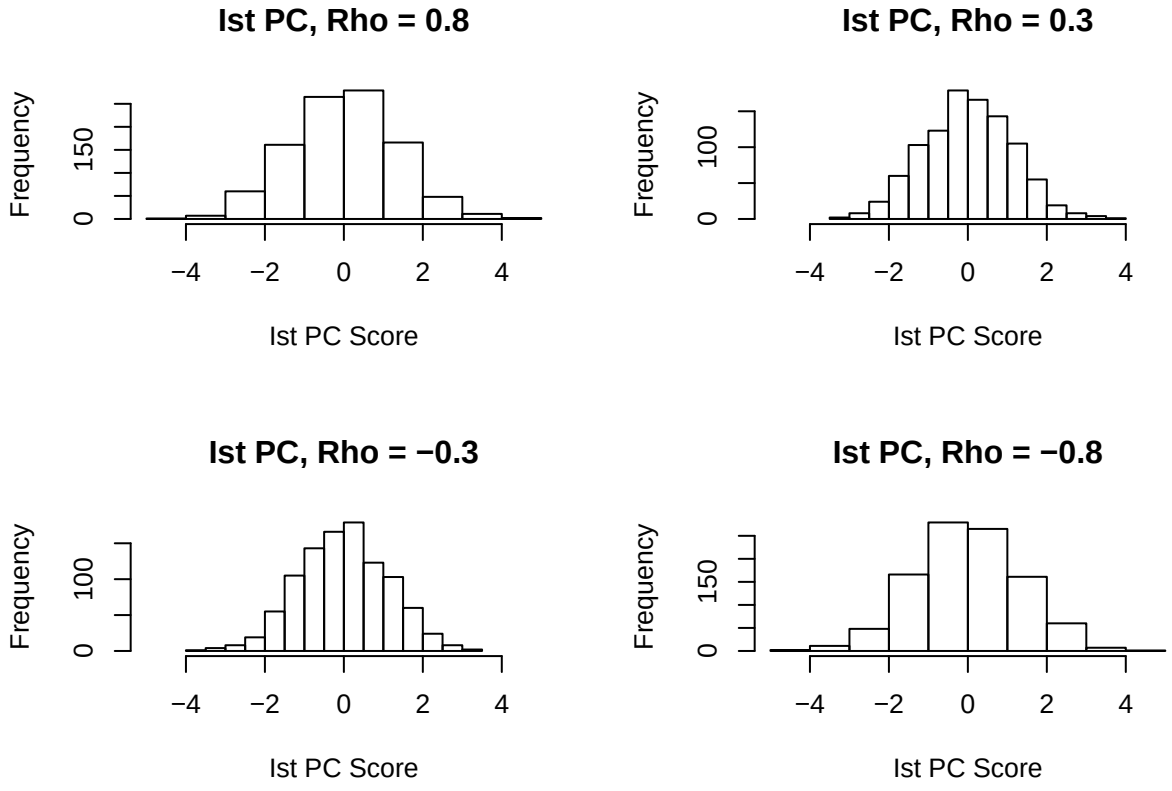
```
x.b31 = as.data.frame(x.1b1) #Rho = 0.8
pca.b31 = princomp(x.b31) #Rho = 0.8

x.b32 = as.data.frame(x.1b2) #0.3
pca.b32 = princomp(x.b32) #Rho = 0.3

x.b33 = as.data.frame(x.1b3) # -0.3
pca.b33 = princomp(x.b33) #Rho = -0.3

x.b34 = as.data.frame(x.1b4) # -0.8
pca.b34 = princomp(x.b34) #Rho = -0.8

par(mfrow = c(2,2))
hist(pca.b31$scores[,1],
     xlab = "Ist PC Score", ylab = "Frequency",
     xlim = c(-5, 5), main = "Ist PC, Rho = 0.8")
hist(pca.b32$scores[,1],
     xlab = "Ist PC Score", ylab = "Frequency",
     xlim = c(-5, 5), main = "Ist PC, Rho = 0.3")
hist(pca.b33$scores[,1],
     xlab = "Ist PC Score", ylab = "Frequency",
     xlim = c(-5, 5), main = "Ist PC, Rho = -0.3")
hist(pca.b34$scores[,1],
     xlab = "Ist PC Score", ylab = "Frequency",
     xlim = c(-5, 5), main = "Ist PC, Rho = -0.8")
```



```
pca.b31$loadings # Ist and 2nd PCs Loadings and Variance Explained for Rho = 0.8
```

```
##
## Loadings:
##    Comp.1 Comp.2
## V1  0.720  0.694
## V2  0.694 -0.720
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

```
pca.b32$loadings # Ist and 2nd PCs Loadings and Variance Explained for Rho = 0.8
```

```
##
## Loadings:
##    Comp.1 Comp.2
## V1  0.766  0.643
## V2  0.643 -0.766
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

```
pca.b33$loadings # Ist and 2nd PCs Loadings and Variance Explained for Rho = 0.8
```

```
##
## Loadings:
##    Comp.1 Comp.2
## V1  0.643  0.766
## V2 -0.766  0.643
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

```
pca.b34$loadings # Ist and 2nd PCs Loadings and Variance Explained for Rho = 0.8
```

```
##
## Loadings:
##    Comp.1 Comp.2
## V1  0.694  0.720
## V2 -0.720  0.694
##
##                Comp.1 Comp.2
## SS loadings       1.0    1.0
## Proportion Var    0.5    0.5
## Cumulative Var    0.5    1.0
```

```
pca.b31$sdev^2 # Variances of Ist and 2nd PCs for Rho = 0.8
```

```
##   Comp.1   Comp.2
## 1.727879 0.208079
```

```
pca.b32$sdev^2 # Variances of Ist and 2nd PCs for Rho = 0.3
```

```
##    Comp.1    Comp.2
## 1.2514788 0.7262011
```

```
pca.b33$sdev^2 # Variances of Ist and 2nd PCs for Rho = -0.3
```

```
##    Comp.1    Comp.2
## 1.2514788 0.7262011
```

```
pca.b34$sdev^2 # Variances of Ist and 2nd PCs for Rho = -0.8
```

```
##   Comp.1   Comp.2
## 1.727879 0.208079
```

- Variance of Ist PC of Rho 0.8: $(\lambda_1) = 1.73$
- Variance of Ist PC of Rho 0.3: $(\lambda_1) = 1.25$
- Variance of Ist PC of Rho -0.3: $(\lambda_1) = 1.25$
- Variance of Ist PC of Rho -0.8: $(\lambda_1) = 1.73$

Variances of Ist PC here are very close but are smaller compared to variances of Ist PC from above.

Q1.B.IV:

```
sigma.1b2
```

```
##      [,1] [,2]
## [1,]  1.0  0.3
## [2,]  0.3  1.0
```

```
eigen(sigma.1b2)
```

```
## eigen() decomposition
## $values
## [1] 1.3 0.7
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

```
sigma.1b3
```

```
##      [,1] [,2]
## [1,]  1.0 -0.3
## [2,] -0.3  1.0
```

```
eigen(sigma.1b3)
```

```
## eigen() decomposition
## $values
## [1] 1.3 0.7
##
## $vectors
##            [,1]       [,2]
## [1,] -0.7071068 -0.7071068
## [2,]  0.7071068 -0.7071068
```

- Answer: When correlation is 0.8 or 0.3 between X_1 and X_2, we obtain positive sign in first PC for both X_1 and X_2. This is because of positive correlation between X_1 and X_2. But when correlation is -0.3 or

-0.8 between X_1 and X_2, we obtian negative sign for X_1 and positive sign for X_2. This is because of negative correlation between X_1 and X_2.

Q2:

```
sigma.q2.1 = matrix (c(1, 0.5, 0.5, 1), nrow = 2, ncol = 2) # COV Matrix; rho11 = 1
decomp.q2.1  = eigen(sigma.q2.1)
decomp.q2.1
```

```
## eigen() decomposition
## $values
## [1] 1.5 0.5
##
## $vectors
##           [,1]       [,2]
## [1,] 0.7071068 -0.7071068
## [2,] 0.7071068  0.7071068
```

```
sigma.q2.2 = matrix (c(5, 0.5, 0.5, 1), nrow = 2, ncol = 2) # COV Matrix; rho11 = 5
decomp.q2.2  = eigen(sigma.q2.2)
decomp.q2.2
```

```
## eigen() decomposition
## $values
## [1] 5.0615528 0.9384472
##
## $vectors
##            [,1]       [,2]
## [1,] -0.9925076  0.1221833
## [2,] -0.1221833 -0.9925076
```

```
sigma.q2.3 = matrix (c(10, 0.5, 0.5, 1), nrow = 2, ncol = 2) # COV Matrix; rho11 = 10
decomp.q2.3  = eigen(sigma.q2.3)
decomp.q2.3
```

```
## eigen() decomposition
## $values
## [1] 10.0276926  0.9723074
##
## $vectors
##             [,1]        [,2]
## [1,] -0.99846976  0.05530039
## [2,] -0.05530039 -0.99846976
```

```
sigma.q2.4 = matrix (c(100, 0.5, 0.5, 1), nrow = 2, ncol = 2) # COV Matrix; rho11 = 100
decomp.q2.4  = eigen(sigma.q2.4)
decomp.q2.4
```

```
## eigen() decomposition
## $values
## [1] 100.0025252   0.9974748
##
## $vectors
##              [,1]         [,2]
## [1,] -0.999987247  0.005050312
## [2,] -0.005050312 -0.999987247
```

- First Principle component for $\sigma_{11} = 1$: $Z_1 = 0.707\ X_1 + 0.707\ X_2$

- First Principle component for $\sigma_{11} = 5$: Z_1 = -0.993 X_1 -0.122 X_2
- First Principle component for $\sigma_{11} = 10$: Z_1 = -0.999 X_1 - 0.553 X_2
- First Principle component for $\sigma_{11} = 100$: Z_1 = -0.999 X_1 -0.005 X_2

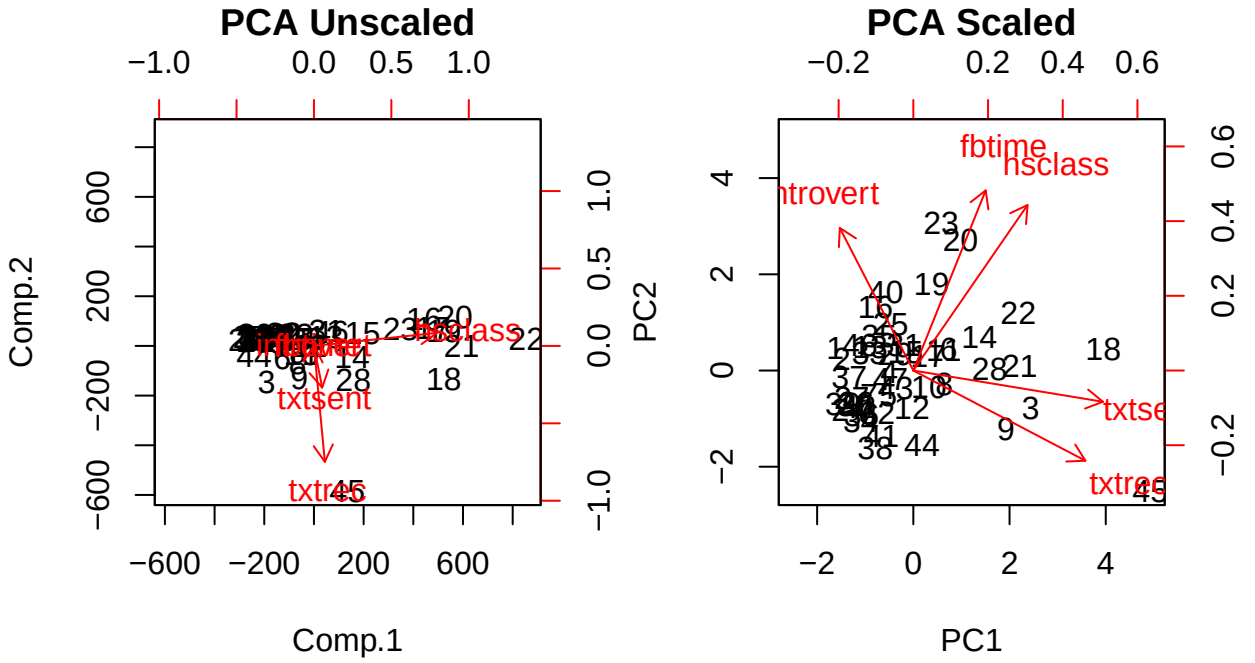
Comparing four Ist PCs above, we can see the coefficient of X_1 increasing in magnitude as σ_{11} increases from 1 to 100. At 100, coefficient of X_2 is almost zero and that of X_1 is almost 1 in magnitude. So, we need to scale the data i.e. run PCA on correlation matrix not on covariance matrix. Also, as the value of σ_{11} increases, the variance explained by first principle component increases in overall PCA. Also, the variable X_1 appears to be most important in the analysis compared to any other variable which is most likely due to increase in scale. For example, the variance explanined by X_1 and X_2 was equal in Z_1 (first PC) when σ_{11} = 1 and variance of Z_1 = 1.5 and Z_2 = 0.5. When σ_{11} = 100 (increase in scale) and variance of Z_1 = 100 and Z_2 = 0.99 which means almost all variance is explained by first PC due to increased scale of X_1.

Q3:

```
library(readxl) #read excel.
# studentdata2019 = read_excel("Dropbox//OSU/PhD//SemVISp2021//STAT5063ML//Data//StudentData2019.xlsx")
studentdata2019 = read_excel("//Users//bmishra//Dropbox//OSU//PhD//SemVISp2021//STAT5063ML//Data//Stude
q3.data = as.data.frame(cbind(studentdata2019[,3:6],
                              studentdata2019[,9]))
q3.data = setNames(q3.data,
                   tolower(names(q3.data))) #lower case names.
attach(q3.data, pos = 2L, warn.conflicts = F)
# names(q3.data)
# View(q3.data)
```

Q3.A:

```
par(mfrow = c(1,2))
q3a.unsc = princomp(q3.data, cor = FALSE) #PCA Unscaled.
biplot(q3a.unsc, scale = 0, main = "PCA Unscaled")
q3a.sc = prcomp(q3.data, scale = T) #PCA Scaled.
biplot(q3a.sc, scale = 0, main = "PCA Scaled")
```



```
as.data.frame(cbind(var(q3.data$hsclass),
                    var(q3.data$txtsent),
                    var(q3.data$txtrec),
                    var(q3.data$fbtime),
                    var(q3.data$introvert)))
```

```
##      V1       V2       V3       V4       V5
## 1 78448 2413.782 10317.63 987.4283 4.477567
```

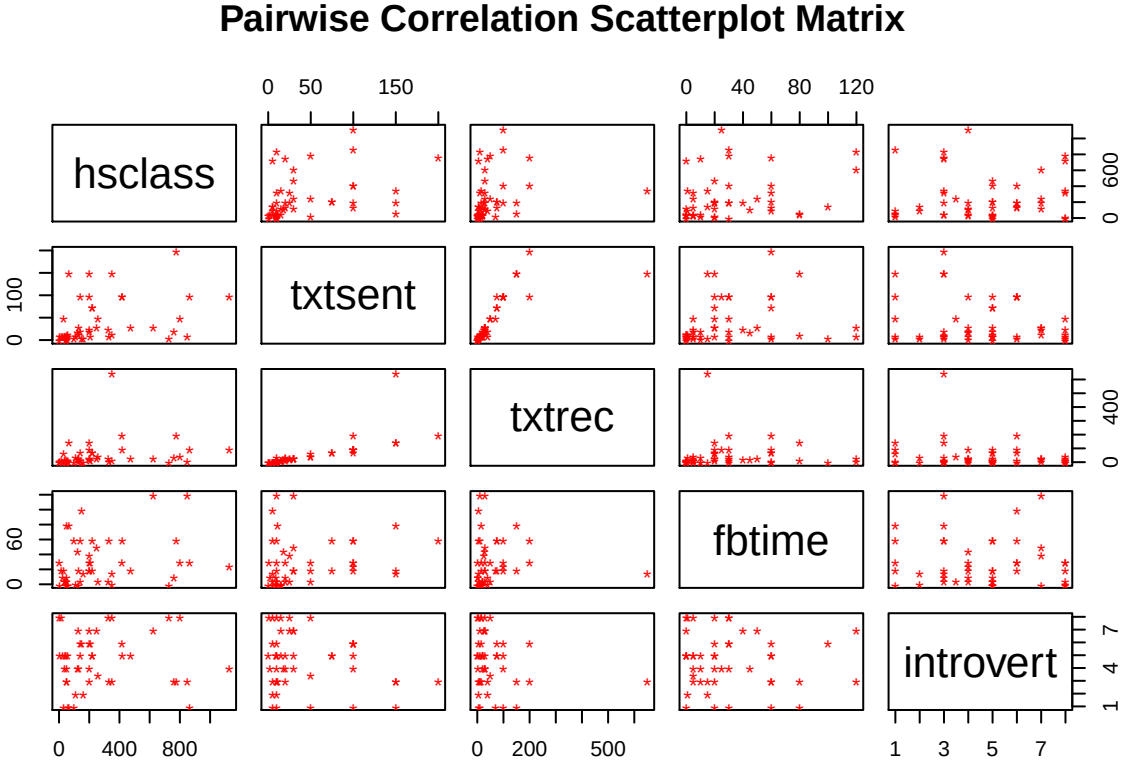
The variable do not have same unit of measurement or scale. If we perform PCA on unstandarized data, the first principle component loading vector will have a very large loading for a variable with highest variance. In out data, txtrec has highest variance and thus the first principle component loading vector will have a very large loading for txtrec which we can see in the figure above as well.

Q3.B:

```
round(as.data.frame(cor(q3.data)),3) #correlation matrix.
```

```
##           hsclass txtsent txtrec fbtime introvert
## hsclass     1.000   0.352  0.206  0.255     0.043
## txtsent     0.352   1.000  0.726  0.192    -0.221
## txtrec      0.206   0.726  1.000  0.023    -0.196
## fbtime      0.255   0.192  0.023  1.000    -0.068
## introvert   0.043  -0.221 -0.196 -0.068     1.000
```

```
plot(q3.data, pch = "*", col = "red", main = "Pairwise Correlation Scatterplot Matrix")
```



I call correlation coefficient below 0.8 as moderate for this assignment. Text sent (txtsent) and text received (txtrec) have positive and moderate correlation as seen in the chart as well as correlatin matrix. The slope of scatterplot is increasing which shows positive correlation. This is further supported by correlation coefficient of 0.726 between txtrec and txtsent in the correlation matrix.

Q3.C.I:

```
# q3c.pca1 = prcomp(q3.data, scale = T) #PCA Scaled.
q3c.pca = princomp(q3.data, cor = TRUE) #PCA Scaled.
q3c.pca # Call, Standard Deviation of PCAS, # Variables, # Obs
```

```
## Call:
## princomp(x = q3.data, cor = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## 1.4246218 1.0657942 0.9733684 0.8006747 0.4959934
##
##  5  variables and  47 observations.
```

```
# names(q3c.pca)
```

```
# q3c.pca = princomp(q3.data, cor = TRUE) #PCA Scaled.
q3c.pca$loadings #PCA loadings.
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## hsclass    0.382  0.554  0.183  0.706  0.123
## txtsent    0.635 -0.105  0.143 -0.170 -0.733
## txtrec     0.576 -0.302  0.277 -0.261  0.657
## fbtime     0.242  0.603 -0.581 -0.475  0.121
## introvert -0.246  0.477  0.729 -0.422
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

Ist PC: $Z_1 = 0.382\text{hsclass} + 0.635\text{txtsent} + 0.576\text{txtrec} + 0.242\text{fbtime} - 0.246\text{introvert}$

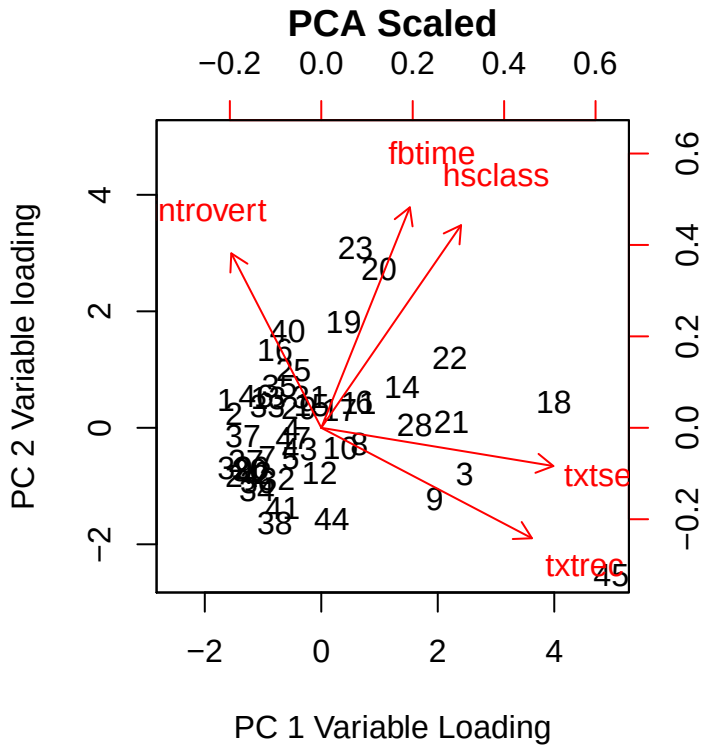
Q3.C.I: text sent and text received. In correlation matrix, txtsent and txtrec have positive correlation with coefficient of 0.726 which is the highest among any pairwise correlation. Also, in the first PC these two viariables have highest vector loadings.

Q3.C.II: In the pairwise correlation matrix, the pairwise correlations between HSClass, TxtSent, TxtRec, FbTime are positively correlated with each other but introvert is negatively correlated with txtsent, txtrec, fbtime. In Ist PC factor loading above, introvert has negative factor loading and all other variables have positive loadings.

Q3.C.III: I expect a person from a large high school who does a lot of texting and facebooking to have positive PC score. The loading of these hsclass, txtsent, txtrec and facbooking are positively loaded in Ist PC.

Q3.D:

```
par(mfrow = c(1,1))
biplot(q3c.pca, scale = 0, main = "PCA Scaled",
       ylab = "PC 2 Variable loading",
       xlab = "PC 1 Variable Loading")
```



Q3.D.I: The arrows point in the same direction and angle is small so they are strongly associated.

The arrows are plotting Ist and 2nd PC loading for each variable. For example, The arrow for fbtime is showing the direction of loading towards point (0.242, 0.603) which denote fbtime location on plot. The point is obtained from scores of fbtime on I^{st} and 2^{nd} PCs. These values are plotted on scales on top for PC_1 and on the right for PC_2. The origin of arrow is (0,0). The diection of arrow shows direction of variability from the origin.

Q3.D.II:: The texting arrows are pointed in the opposite direction to the introversion arrow. This tells us that texting variables are similar variables but introversion is opposite variable to texting variables. This further can be understood from factor loading and correlation matrix. In correlation matrix, txtsend and txtrev have positive sign and are positively correlated with each other. Also, in factor loadings, these two variables have same direction and almost same magnitue. Howeve, the introversion has negative correlation with both txtrec and txtsent in the correlation matrix as well as opposite sign in Ist PC which explains texting behavior. This is as expected.

Q3.D.III:

```
score45 = q3c.pca$scores[45,2] # Scores for 45
score45 # Scores for 45
```

```
##    Comp.2
## -2.527434
```

```
q3c.pca$loading # loadings
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## hsclass    0.382  0.554  0.183  0.706  0.123
## txtsent    0.635 -0.105  0.143 -0.170 -0.733
## txtrec     0.576 -0.302  0.277 -0.261  0.657
## fbtime     0.242  0.603 -0.581 -0.475  0.121
```

```
## introvert -0.246  0.477  0.729 -0.422
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

Fb was below average.

The sum of loading for second PC is positive. But the scores of 45th sudent is negative. The negative value tells us that the values for variables are below average. The 2^{nd} PC is not explaining texting behavior but might be explaining rest of the three variables . Also note that if we want to explain more variance, we may retain three components and let third component explain introvert behavior.

Q3.E: Center and scale data and report 5 Z-score. $Z_{txtsent}$ = (# of text sent - average text sent)/standard deviation of txt sent.

```
#scaled and centered data of all variable.
sc.q3.data = scale(q3.data, center = TRUE, scale = TRUE)
z.hsclass = (hsclass - mean(hsclass))/sd(hsclass)
z.txtsent = (txtsent - mean(txtsent))/sd(txtsent)
z.txtrec = (txtrec - mean(txtrec))/sd(txtrec)
z.fbtime = (fbtime - mean(fbtime))/sd(fbtime)
z.introvert = (introvert - mean(introvert))/sd(introvert)
scale.man = as.data.frame(cbind(z.hsclass, z.txtsent,
                                z.txtrec, z.fbtime, z.introvert))
q3e.45score = scale.man[45,] #5 Z-score for 45th Student.
q3e.45score # 5 Z-scores for 45th Student.
```

```
##    z.hsclass z.txtsent z.txtrec   z.fbtime z.introvert
## 45  0.272485  2.195207 5.837799 -0.4922478  -0.7591508
```

For txtsent and txtrec I am above average, for hsclass, I am about average and for fbtime and introvert, I am below average.

Q3.F: Report PC score and verify manually by computing using answer above, and the loading from the first PC. Interpret your PC.

```
q3f.scores = q3c.pca$scores #PC Scores.
q3f.scores[1] #Report PC Scores.
```

```
## [1] -1.63719
```

```
q3f.score = scale.man[1,] #5 Z-score for 45th Student from manual computation.
q3f.score # 5 Z-scores for 45th Student.
```

```
##    z.hsclass  z.txtsent   z.txtrec    z.fbtime z.introvert
## 1 -0.9735622 -0.8375478 -0.5515222 -0.01489608    1.603769
```

```
q3f.Ist.pc = q3c.pca$loadings[,1] #Ist PC loading. This is common for all students.
q3f.Ist.pc
```

```
##    hsclass    txtsent     txtrec     fbtime  introvert
##  0.3822469  0.6346995  0.5760710  0.2421575 -0.2460605
```

```
Z1.q3f = (q3f.score[1]*q3f.Ist.pc[1]) +
  (q3f.score[2]*q3f.Ist.pc[2]) +
  (q3f.score[3]*q3f.Ist.pc[3]) +
  (q3f.score[4]*q3f.Ist.pc[4]) +
```

```
  (q3f.score[5]*q3f.Ist.pc[5])
as.data.frame(cbind(Z1.q3f, q3f.scores[1])) #these scores should be equal.
```

```
##   z.hsclass q3f.scores[1]
## 1  -1.61968      -1.63719
```

This gives first PC loading for first observation. First PC is correlated with texting behavior but for the selected case, it loaded negatively in the Ist PC. So, we can expect that the person has lower values in texting related behaviour.

Q3.G: How many PCs would you need to retain 80% or more of the variances?

```
summary(q3c.pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.4246218 1.0657942 0.9733684 0.8006747 0.49599342
## Proportion of Variance 0.4059094 0.2271835 0.1894892 0.1282160 0.04920189
## Cumulative Proportion  0.4059094 0.6330929 0.8225821 0.9507981 1.00000000
```

To retain 80% or more variability, we need to retain at least 3 components as cumulative proportion of variance explained by up to 3 PCs is 82.26%.

Q4.: Cluster Analysis on the Social Media Data:

Q4.a.I:

```
set.seed(1)
q4.data = scale(q3.data, center = TRUE, scale = TRUE)  #Centered & Scaled data.
q4.kfit = kmeans(q4.data, # data
                 centers = 2, # Clusters K = 2
                 iter.max = 20, # maximum iteration.
                 nstart = 50) # # of random starting points.
q4.kfit
```

```
## K-means clustering with 2 clusters of sizes 14, 33
##
## Cluster means:
##      hsclass   txtsent     txtrec    fbtime   introvert
## 1  0.8220616  1.199312  0.8380110  0.689766 -0.11778688
## 2 -0.3487534 -0.508799 -0.3555198 -0.292628  0.04997019
##
## Clustering vector:
##  [1] 2 2 1 2 2 1 2 1 1 2 1 2 2 1 2 2 2 1 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2
## [39] 2 2 2 2 2 2 1 2 2
##
## Within cluster sum of squares by cluster:
## [1] 92.59185 71.48757
##  (between_SS / total_SS =  28.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
centroids = q4.kfit$centers
centroids #centroids for cluster 1 and cluster 2 for each variable.
```

```
##      hsclass   txtsent     txtrec    fbtime   introvert
## 1  0.8220616  1.199312  0.8380110  0.689766 -0.11778688
## 2 -0.3487534 -0.508799 -0.3555198 -0.292628  0.04997019
```

```
Within.Cluster = q4.kfit$withinss
Within.Cluster # Within cluster sum of squares (cluster 1 and cluster 2)
```

```
## [1] 92.59185 71.48757
```

```
Total.variability = (q4.kfit$betweenss/q4.kfit$totss)*100
Total.variability # total variability explained by cluster assignment.
```

```
## [1] 28.66112
```

```
total.SS = q4.kfit$totss
total.SS # Total sum of squares.
```

```
## [1] 230
```

```
Total.Within.SS = q4.kfit$tot.withinss
Total.Within.SS # Total within sum of squares.
```

```
## [1] 164.0794
```

```
Between.SS = q4.kfit$betweenss
Between.SS # Between Sum of Squares.
```

```
## [1] 65.92058
```

Q4.a.II:

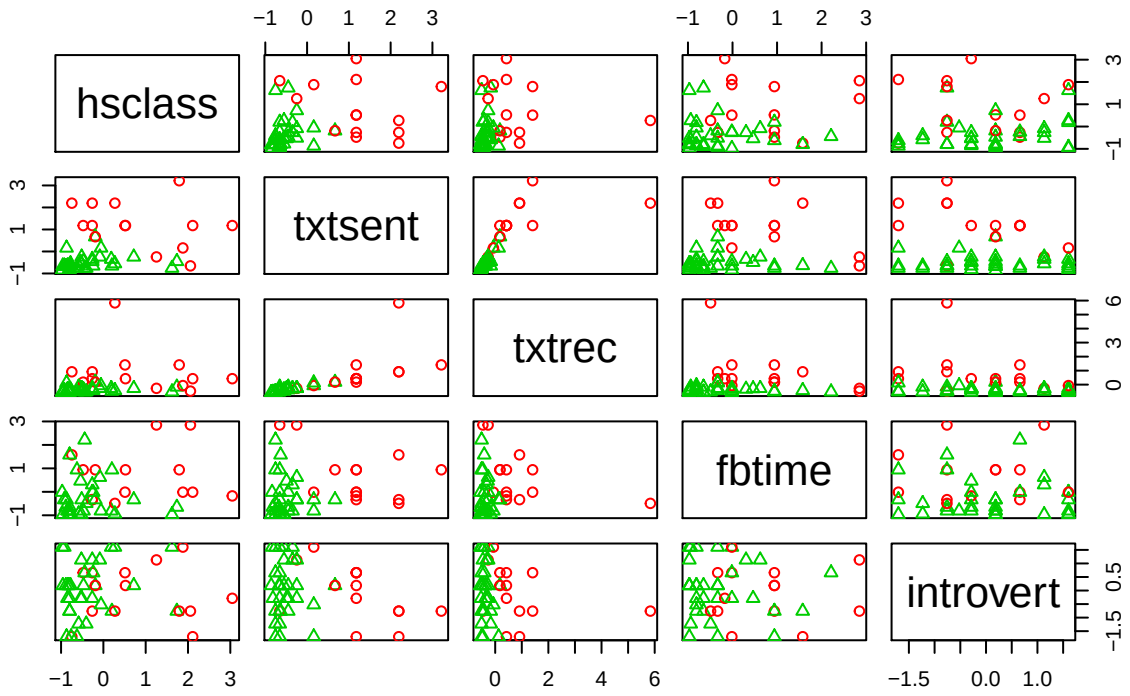
```
centroids = q4.kfit$centers
centroids #centroids for cluster 1 and cluster 2 for each variable.
```

```
##      hsclass   txtsent     txtrec    fbtime   introvert
## 1  0.8220616  1.199312  0.8380110  0.689766 -0.11778688
## 2 -0.3487534 -0.508799 -0.3555198 -0.292628  0.04997019
```

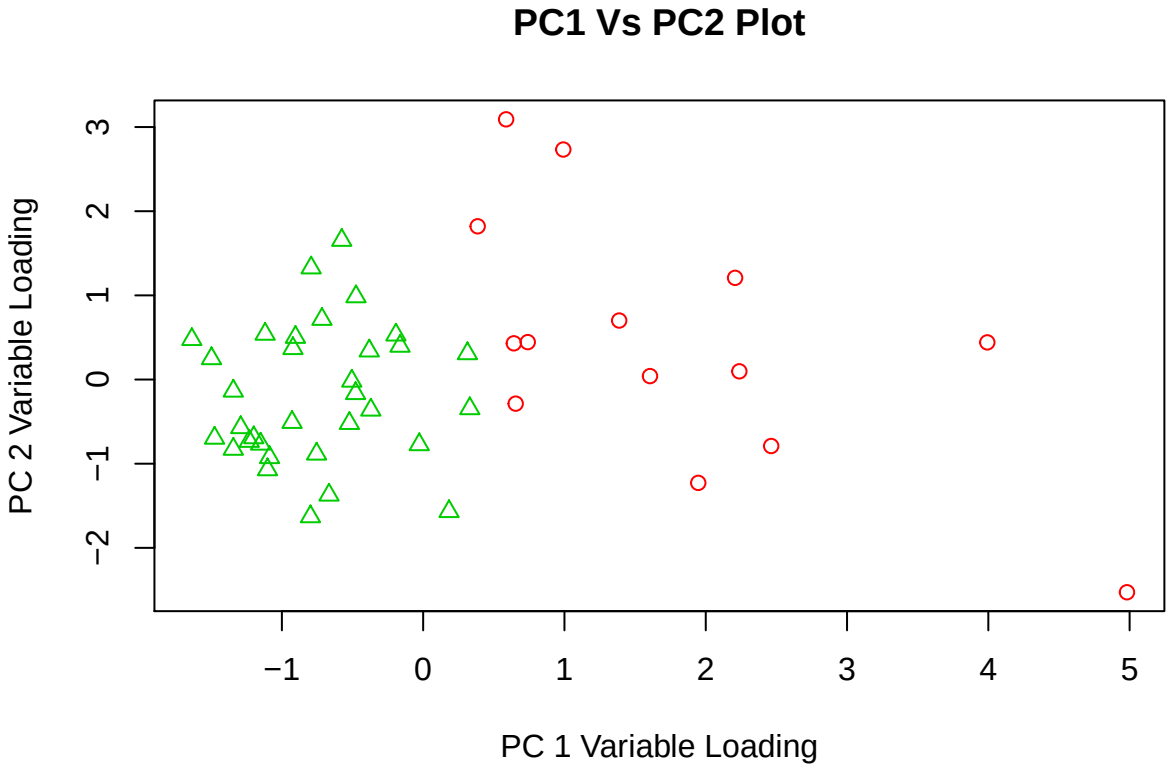
In cluster 1 here, the loading is defined by txt sent and txt received as these two variables has highest loadings which is same as Ist PC of PCA defined by loading of txtsent and txtrec. The cluster 1 has higher centroid value compared to cluster 2 for all variables except introvert. The introvert has higher centroid value for cluster 2 and lower centroid value for cluster 1. As seen in the first PC earlier, hsclass, txtsent, txtrec, fbtime loaded in the first principle component positively and intorvert loaded negatively in Ist PC which is also reflected in the K-mean clustering.

Q4.a.III: Use pairs function to plot all pairwise plots on centered and scaled data.

```
pairs(q4.data,
      col = q4.kfit$cluster + 1,
      pch = q4.kfit$cluster,
      main = "Pairwise Plot")
```




```
plot(q3c.pca$scores[,1:2],
     col = q4.kfit$cluster + 1,
     pch = q4.kfit$cluster,
     xlab = "PC 1 Variable Loading",
     ylab = "PC 2 Variable Loading",
     main = "PC1 Vs PC2 Plot")
```



The bivariate plot depicits across cluster variation the best. In the scatter plot red circles and green triangles are mixed with each other and hard to understand/differentiate the trend. However, in the PC bivariate plot, we can clearly divide the plot into two clusters based on the color and shape of plot characteristics. This is because PCs explain large variations in the data using less number of uncorrelated variables (PCs) unlike scatterplot which uses five variables to explain the variance in the data.

Q4.b: Get dendrogram for agglomerative cluster analysis for complete, average, single and centroid linkage:

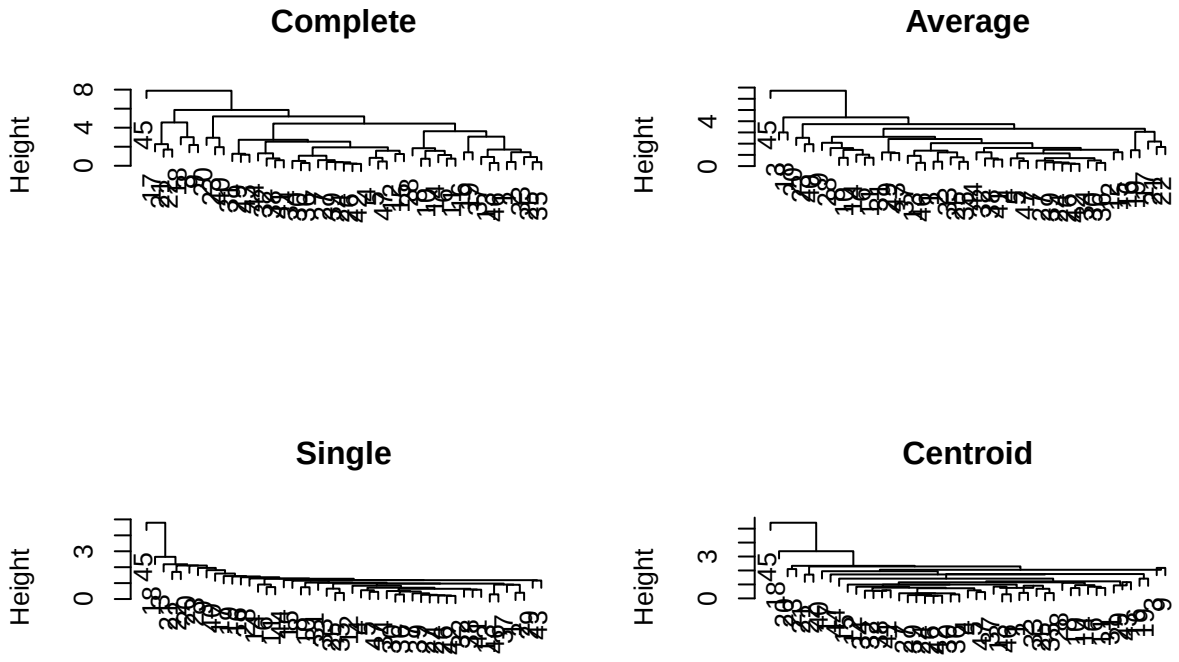
```
# q4.data = scale(q3.data, center = TRUE, scale = TRUE)
par(mfrow = c(2,2))
q4b.complete = hclust(dist(q4.data),
                      method = "complete")
plot(q4b.complete,
     xlab = "", sub = " ",
     main = "Complete")

q4b.average = hclust(dist(q4.data),
                      method = "average")
plot(q4b.average,
     xlab = "", sub = " ",
     main = "Average")

q4b.single = hclust(dist(q4.data),
                      method = "single")
plot(q4b.single,
     xlab = "", sub = " ",
     main = "Single")

q4b.centroid = hclust(dist(q4.data),
                      method = "centroid")
```

```
plot(q4b.centroid,
     xlab = "", sub = " ",
     main = "Centroid")
```



Q4.b.45: Remove 45th Observation and Redo PCA, Kmean and Hierarchial:

```
q4.45.data = rbind(q3.data[1:44, ], q3.data[46:47,])
q4.45 = scale(q4.45.data, center = TRUE, scale = TRUE)

q4.45.pca = princomp(q4.45, cor = TRUE) #PCA Scaled.
q4.45.pca$loadings
```

```
##
## Loadings:
##           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## hsclass    0.376  0.532         0.759
## txtsent    0.616 -0.124 -0.230 -0.217  0.711
## txtrec     0.608 -0.151 -0.277 -0.194 -0.702
## fbtime     0.274  0.381  0.783 -0.407
## introvert -0.186  0.730 -0.508 -0.417
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

```
summary(q4.45.pca)
```

```
## Importance of components:
```

```
##                            Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.5212649 1.0339400 0.9599592 0.8020633 0.22780236
## Proportion of Variance 0.4628494 0.2138064 0.1843043 0.1286611 0.01037878
## Cumulative Proportion  0.4628494 0.6766558 0.8609601 0.9896212 1.00000000
```

```
q4.45.kfit = kmeans(q4.45, # data
                    centers = 2, # Clusters K = 2
                    iter.max = 20, # maximum iteration.
                    nstart = 50) # # of random starting points.
q4.45.kfit
```

```
## K-means clustering with 2 clusters of sizes 14, 32
##
## Cluster means:
##      hsclass    txtsent     txtrec     fbtime   introvert
## 1  0.7867942  1.1911064  1.1246963  0.6847296 -0.06647517
## 2 -0.3442225 -0.5211091 -0.4920546 -0.2995692  0.02908288
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  2  2  1  2  2  1  2  1  1  1  1  2  2  1  2  2  2  1  1  1  1  1  1  2  2  2
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 46 47
##  2  1  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] 77.74714 71.26096
##  (between_SS / total_SS =  33.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

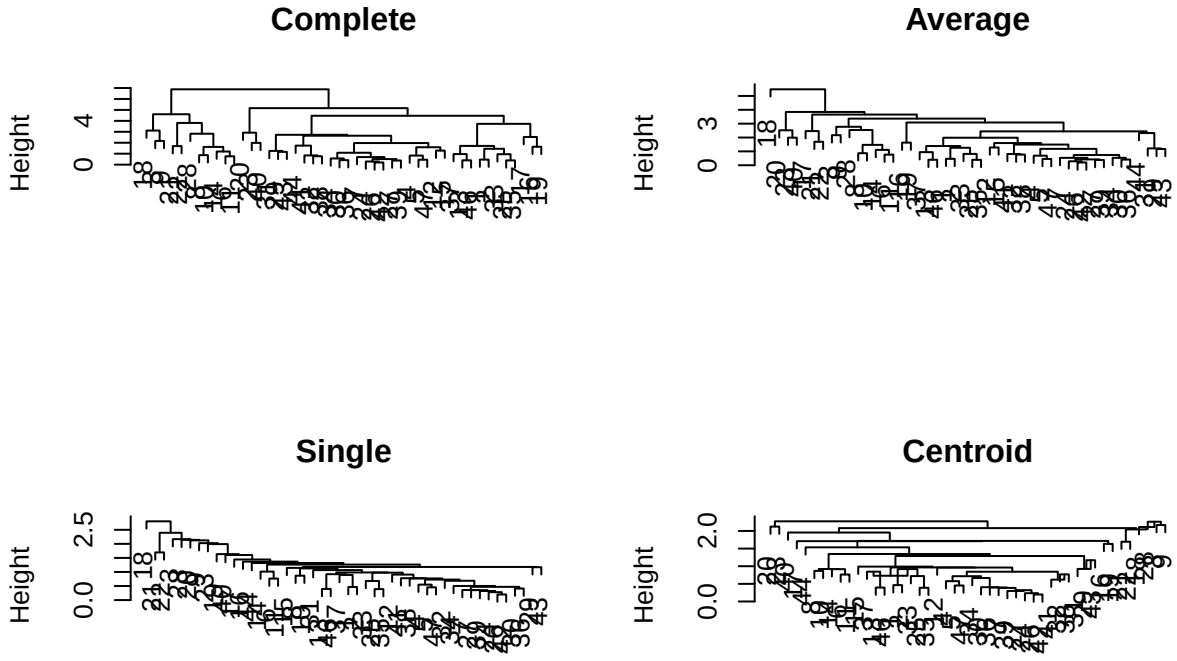
```
par(mfrow = c(2,2))
q4.45.complete = hclust(dist(q4.45),
                        method = "complete")
plot(q4.45.complete,
     xlab = "", sub = " ",
     main = "Complete")

q4.45.average = hclust(dist(q4.45),
                       method = "average")
plot(q4.45.average,
     xlab = "", sub = " ",
     main = "Average")

q4.45.single = hclust(dist(q4.45),
                      method = "single")
plot(q4.45.single,
     xlab = "", sub = " ",
     main = "Single")

q4.45.centroid = hclust(dist(q4.45),
                        method = "centroid")
plot(q4.45.centroid,
```

```
     xlab = "", sub = " ",
     main = "Centroid")
```



Yes. Removing 45th Student from the observation can change the result in PCA, K-mean clusering and Hierarchial clustering. But depends upon which level of change we are considering, change might be significant or non-significant. For example, removing 45th student did not change the how each variaable loads in first and second factors but the variance explained by both increased. However, loadings of variables in PC3, PC4 and PC5 changed. In K-mean clustering, Total proportion of variation attributed to cluster assignment increased and cluster means also change when student 45 is removed from the dataset. Clustering as shown by dendrogram using various method also changed.

EXTRA CREDIT PROBLEM FROM Q1 IN NEXT PAGE

Remarks after grading: Extra credit does not make sense and no points was obtained. So, discard extra credit solution.