Homework 06

Joseph Blubaugh
jblubau1@tamu.edu

STAT 636-720

1)

```
## a)
X1 = read.delim("T8-4.DAT", header=FALSE)
colnames(X1) = c("x1", "x2", "x3", "x4", "x5")

head(X1)
```

```
          x1         x2         x3         x4         x5
1  0.0130338 -0.0078431 -0.0031889 -0.0447693  0.0052151
2  0.0084862  0.0166886 -0.0062100  0.0119560  0.0134890
3 -0.0179153 -0.0086393  0.0100360  0.0000000 -0.0061428
4  0.0215589 -0.0034858  0.0174353 -0.0285917 -0.0069534
5  0.0108225  0.0037167 -0.0101345  0.0291900  0.0409751
6  0.0101713 -0.0121978 -0.0083768  0.0137083  0.0029895
```

```
## Sample Variance
S1 = var(X1)

## Pricipal Components
(pca.1 = prcomp(X1))
```

```
Standard deviations:
[1] 0.03698213 0.02647942 0.01593118 0.01194163 0.01090352

Rotation:
          PC1        PC2         PC3        PC4         PC5
x1 -0.2228228  0.6252260 -0.32611218  0.6627590 -0.11765952
x2 -0.3072900  0.5703900  0.24959014 -0.4140935  0.58860803
x3 -0.1548103  0.3445049  0.03763929 -0.4970499 -0.78030428
x4 -0.6389680 -0.2479475  0.64249741  0.3088689 -0.14845546
x5 -0.6509044 -0.3218478 -0.64586064 -0.2163758  0.09371777
```

```
## Multiply the data with the eigen vectors
PC.1 = as.matrix(X1) %*% eigen(S1)$vectors
head(PC.1, 5)
```

```
             [,1]         [,2]          [,3]          [,4]         [,5]
[1,] -0.025211170  0.011998822  0.0384604565 -0.0014851773 -0.003473245
[2,]  0.022477337  0.005379563 -0.0001338018  0.0025744885 -0.013159477
[3,] -0.009091407 -0.010694384 -0.0080312540 -0.0119552884  0.011384079
[4,] -0.016363352  0.026824644  0.0211235548 -0.0002609519  0.014600282
[5,]  0.047307028 -0.015030242  0.0106928576  0.0108208650 -0.008328983
```

```
## b)
## Proportion of variance explained by PCA
diag(var(PC.1)) / sum(diag(S1)); summary(pca.1)
```

```
[1] 0.52926066 0.27133298 0.09821584 0.05518400 0.04600652
```

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5
Standard deviation     0.03698 0.02648 0.01593 0.01194 0.01090
Proportion of Variance 0.52926 0.27133 0.09822 0.05518 0.04601
Cumulative Proportion  0.52926 0.80059 0.89881 0.95399 1.00000
```

In the first component all of the variables have the same sign and so it can be interpreted as a weighted average linear combination. The ammount of total variance explained by the first component is 53% with the first 3 explaining 90% of the variance in the data. Depending on the application PC4 and PC5 may not be useful since they do not explain the variance very well.

2)

a)

```
X2 = read.delim("T8-5.DAT", header=FALSE)
colnames(X2) = c("x1", "x2", "x3", "x4", "x5")

## Sample Variance
S2 = var(X2)

## PCA on Variance Matrix
(pca.2s = prcomp(X2))

Standard deviations:
[1] 10.3448177  6.2985820  2.8932449  1.6934798  0.3933104

Rotation:
            PC1         PC2         PC3         PC4          PC5
x1  0.038887287 -0.07114494  0.18789258  0.97713524 -0.057699864
x2 -0.105321969 -0.12975236 -0.96099580  0.17135181 -0.138554092
x3  0.492363944 -0.86438807  0.04579737 -0.09104368  0.004966048
x4 -0.863069865 -0.48033178  0.15318538 -0.02968577  0.006691800
x5 -0.009122262 -0.01474342 -0.12498114  0.08170118  0.988637470

summary(pca.2s)

Importance of components:
                          PC1    PC2     PC3     PC4     PC5
Standard deviation     10.345 6.2986 2.89324 1.69348 0.39331
Proportion of Variance  0.677 0.2510 0.05295 0.01814 0.00098
Cumulative Proportion   0.677 0.9279 0.98088 0.99902 1.00000

PCA.2s = as.matrix(X2) %*% pca.2s$rotation
round(cor(X2, PCA.2s), 3)

      PC1    PC2    PC3    PC4    PC5
x1  0.218 -0.243  0.295  0.898 -0.012
x2 -0.350 -0.263 -0.894  0.093 -0.018
x3  0.683 -0.730  0.018 -0.021  0.000
x4 -0.946 -0.321  0.047 -0.005  0.000
x5 -0.167 -0.165 -0.641  0.245  0.689

##confirm correlation
(eigen(S2)$vectors[1,1] * sqrt(eigen(S2)$values[1])) / sqrt(S2[1,1])
```

```
[1] 0.2182675

## PCA on Correlation Matrix
(pca.2r = prcomp(X2, center = TRUE, scale. = TRUE))

Standard deviations:
[1] 1.4113534 1.1694129 0.9296006 0.7314787 0.4912604

Rotation:
          PC1        PC2         PC3        PC4        PC5
x1  0.2625829 -0.4629936  0.78390268 -0.2169291  0.2347882
x2 -0.5933541 -0.3256442 -0.16407255  0.1446471  0.7028828
x3  0.3256978 -0.6051419 -0.22487455  0.6628689 -0.1943206
x4 -0.4792022  0.2524850  0.55070086  0.5716730 -0.2766497
x5 -0.4932213 -0.4996473 -0.06882436 -0.4072024 -0.5801162

summary(pca.2r)

Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.4114 1.1694 0.9296 0.7315 0.49126
Proportion of Variance 0.3984 0.2735 0.1728 0.1070 0.04827
Cumulative Proportion  0.3984 0.6719 0.8447 0.9517 1.00000

PCA.2r = as.matrix(X2) %*% pca.2r$rotation
round(cor(X2, PCA.2r), 3)

      PC1    PC2    PC3   PC4    PC5
x1  0.306 -0.389  0.068 0.072 -0.057
x2 -0.570  0.023  0.211 0.362  0.413
x3  0.654 -0.935 -0.569 0.453 -0.157
x4 -0.905  0.643  0.944 0.622 -0.541
x5 -0.339 -0.086  0.103 0.173  0.301
```

b)

PCA.2s 93% of the variation can be explained by the first 2 pricipal components. The first components looks primarily to be made up of the weighted differences between x3 anx x4. The second component looks to be a weighted average of x2, x3, and x4.

PCA.2r 67% of the variation can be explained by the first 2 principal components. The first component is made up of a weighted difference between (x1,x3) and (x2,x4,x5). The second component is also a weighted difference between (x1,x2,x3,5) and (x4)

I would recommend using the PCA.2s because it explains a lot more variation with fewer variables.

3)

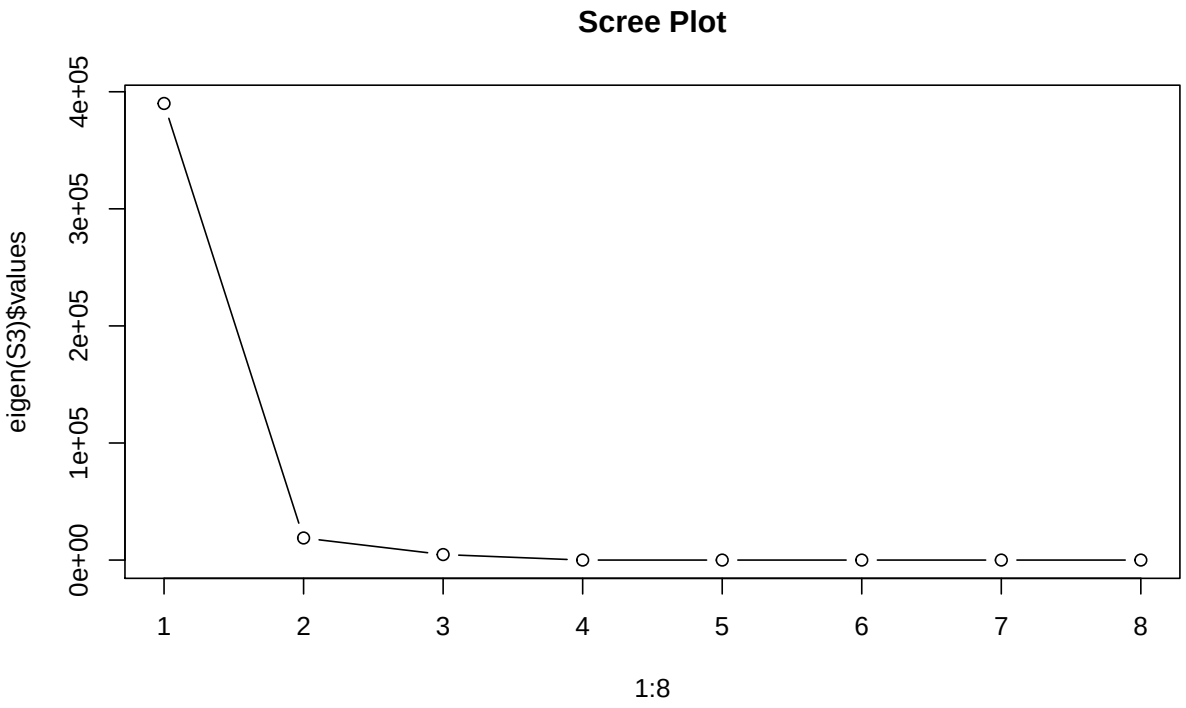
```
## a)
X3 = read.table("T1-10.DAT", quote="\"", comment.char="")
colnames(X3) = c("Breed", "SalePr", "YrHgt", "FtFrBody", "PrctFFB", "Frame",
                 "BkFat", "SaleHt", "SaleWt")

head(X3)
```

```
  Breed SalePr YrHgt FtFrBody PrctFFB Frame BkFat SaleHt SaleWt
1     1   2200  51.0     1128    70.9     7  0.25   54.8   1720
2     1   2250  51.9     1108    72.1     7  0.25   55.3   1575
3     1   1625  49.9     1011    71.6     6  0.15   53.1   1410
4     1   4600  53.1      993    68.9     8  0.35   56.4   1595
5     1   2150  51.2      996    68.6     7  0.25   55.0   1488
6     1   1225  49.2      985    71.4     6  0.15   51.4   1500
```

```
## Sample variance
S3 = var(X3[, 2:9])

## pricipal components
## the first 3 PC explain 86% of the variation
pca.3r = prcomp(X3[, 2:9], center = TRUE, scale. = TRUE)
summary(pca.3r)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.059 1.2929 0.9914 0.65723 0.55302 0.41823 0.38209
Proportion of Variance 0.530 0.2089 0.1229 0.05399 0.03823 0.02186 0.01825
Cumulative Proportion  0.530 0.7390 0.8618 0.91584 0.95407 0.97593 0.99418
                           PC8
Standard deviation     0.21580
Proportion of Variance 0.00582
Cumulative Proportion  1.00000
```

```
## b) The scree plot significantly levels off after 2 components however
##    the first 2 components only account for 74% of the data so I would
##    include the 3rd component since it will get you to 86%
plot(x = 1:8, y = eigen(S3)$values, type = "b", main = "Scree Plot")
```

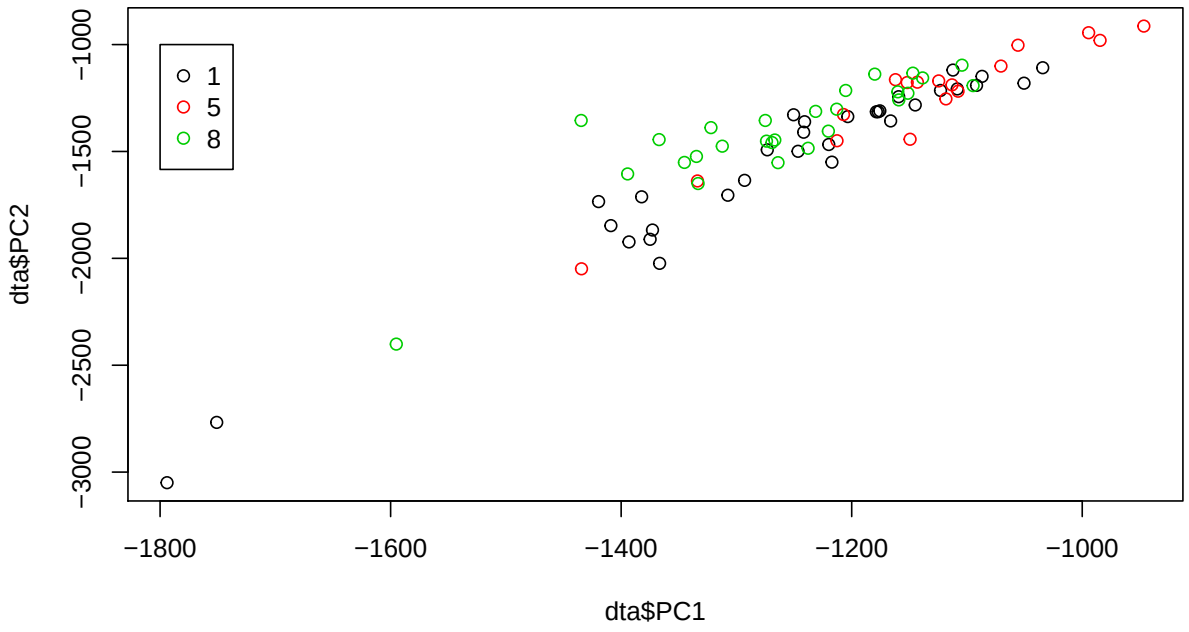



```
## c)
pca.3r

Standard deviations:
[1] 2.0592064 1.2928577 0.9914294 0.6572314 0.5530197 0.4182267 0.3820873
[8] 0.2157955

Rotation:
                PC1          PC2         PC3         PC4         PC5
SalePr   -0.1921089 -0.561675555  0.42133057 -0.17792148 -0.62750027
YrHgt    -0.4491949 -0.006648805  0.28494631  0.00285735  0.32564034
FtFrBody -0.3970224  0.076930360 -0.43727228 -0.28918480  0.08918088
PrctFFB  -0.3286091  0.415619778 -0.23222110 -0.47352003 -0.45383165
Frame    -0.4371693 -0.067289983  0.29910806 -0.11687743  0.34633145
BkFat     0.1581482 -0.609884173 -0.31859984 -0.52563785  0.34327966
SaleHt   -0.4508356 -0.055692911  0.02721363  0.26629293  0.10001146
SaleWt   -0.2762729 -0.355288527 -0.55176117  0.54737491 -0.19620575
                PC6         PC7         PC8
SalePr   -0.2031253  0.04772161 -0.03384373
YrHgt     0.1635363  0.04523054 -0.76298464
FtFrBody -0.7077314  0.22758184 -0.03145767
PrctFFB   0.4630018 -0.14419352 -0.00166139
Frame     0.2490221  0.40715483  0.59455576
```

```
BkFat     0.2250509 -0.23654900 -0.02802110
SaleHt   -0.1331998 -0.79669648  0.24397400
SaleWt    0.2938358  0.25813914 -0.04372241
```

```
## PC1 could be viewed as a weighted difference between all variables and
##     BkFat, or BkFat could small enough that we can drop it and say we
##     have a weighted average between all variables
##
## PC2 is a weighted difference between (SalePrc, BkFat, SaleWt) and (PrctFB)
##     after we drop the variables near 0
```

```
## d)
PC.3 = as.matrix(X3[, 2:9]) %*% pca.3r$rotation
dta = data.frame(BREED = X3$Breed, PC.3[, 1:2])

plot(x = dta$PC1, y = dta$PC2, col = factor(dta$BREED))
legend(x = -1800, y = -1000, unique(dta$BREED), col = 1:3, pch = 1)
```



```
## You can definetely see that breed 5 is difference from breed 8. And Breed 1
## has a lot of variation. The 3 outliers are in terms of price.  These are
## the top 3 priced bulls. The summary statistics indicate that they are in the
## upper percentiles (except for the fat measurements) compared to their same
## breed.

## outliers
outliers = order(dta$PC1)[1:3]
X3[outliers, ]
```

```
   Breed SalePr YrHgt FtFrBody PrctFFB Frame BkFat SaleHt SaleWt
4      1   4600  53.1      993    68.9     8  0.35   56.4   1595
8      1   4000  51.5     1060    69.3     7  0.30   55.6   1765
58     8   3450  54.8     1039    70.6     8  0.10   58.7   1600
```

```
## Breed 1
summary(subset(X3, Breed == 1, 2:9))
```

```
     SalePr         YrHgt          FtFrBody        PrctFFB
 Min.   :1225   Min.   :47.60   Min.   : 841.0   Min.   :64.90
 1st Qu.:1481   1st Qu.:48.98   1st Qu.: 923.0   1st Qu.:68.35
 Median :1638   Median :49.75   Median : 973.5   Median :69.60
 Mean   :1956   Mean   :49.89   Mean   : 969.2   Mean   :69.81
 3rd Qu.:2250   3rd Qu.:51.00   3rd Qu.: 997.2   3rd Qu.:71.33
 Max.   :4600   Max.   :53.10   Max.   :1128.0   Max.   :75.80
     Frame           BkFat          SaleHt          SaleWt
 Min.   :5.000   Min.   :0.15   Min.   :51.20   Min.   :1325
 1st Qu.:5.750   1st Qu.:0.15   1st Qu.:52.83   1st Qu.:1478
 Median :6.000   Median :0.25   Median :53.35   Median :1526
 Mean   :6.062   Mean   :0.25   Mean   :53.55   Mean   :1552
 3rd Qu.:7.000   3rd Qu.:0.30   3rd Qu.:54.62   3rd Qu.:1648
 Max.   :8.000   Max.   :0.50   Max.   :56.40   Max.   :1842
```

```
## Breed 8
summary(subset(X3, Breed == 8, 2:9))
```

```
     SalePr         YrHgt          FtFrBody        PrctFFB
 Min.   :1200   Min.   :49.80   Min.   : 928.0   Min.   :70.60
 1st Qu.:1450   1st Qu.:51.30   1st Qu.: 994.5   1st Qu.:71.10
 Median :1550   Median :52.30   Median :1040.0   Median :74.00
 Mean   :1678   Mean   :52.14   Mean   :1062.8   Mean   :73.63
 3rd Qu.:1838   3rd Qu.:52.95   3rd Qu.:1089.0   3rd Qu.:74.90
 Max.   :3450   Max.   :54.80   Max.   :1383.0   Max.   :81.40
     Frame           BkFat           SaleHt          SaleWt
```

```
 Min.   :6.000   Min.   :0.1000   Min.   :53.90   Min.   :1375
 1st Qu.:7.000   1st Qu.:0.1000   1st Qu.:55.35   1st Qu.:1498
 Median :7.000   Median :0.1000   Median :55.80   Median :1595
 Mean   :7.074   Mean   :0.1222   Mean   :56.03   Mean   :1593
 3rd Qu.:7.500   3rd Qu.:0.1500   3rd Qu.:56.80   3rd Qu.:1670
 Max.   :8.000   Max.   :0.2000   Max.   :59.60   Max.   :1904
```

```
## Breed 5
summary(subset(X3, Breed == 5, 2:9))
```

```
     SalePr         YrHgt          FtFrBody        PrctFFB
 Min.   : 975   Min.   :47.20   Min.   : 843.0   Min.   :65.30
 1st Qu.:1225   1st Qu.:48.60   1st Qu.: 913.0   1st Qu.:67.10
 Median :1325   Median :49.00   Median : 934.0   Median :68.20
 Mean   :1443   Mean   :49.15   Mean   : 940.1   Mean   :68.54
 3rd Qu.:1500   3rd Qu.:49.90   3rd Qu.: 998.0   3rd Qu.:69.80
 Max.   :2750   Max.   :51.00   Max.   :1056.0   Max.   :72.90
     Frame           BkFat           SaleHt          SaleWt
 Min.   :5.000   Min.   :0.1500   Min.   :49.40   Min.   :1285
 1st Qu.:5.000   1st Qu.:0.1500   1st Qu.:51.50   1st Qu.:1410
 Median :6.000   Median :0.2000   Median :52.30   Median :1520
 Mean   :5.588   Mean   :0.2147   Mean   :52.19   Mean   :1502
 3rd Qu.:6.000   3rd Qu.:0.2500   3rd Qu.:52.90   3rd Qu.:1550
 Max.   :7.000   Max.   :0.3500   Max.   :54.40   Max.   :1735
```