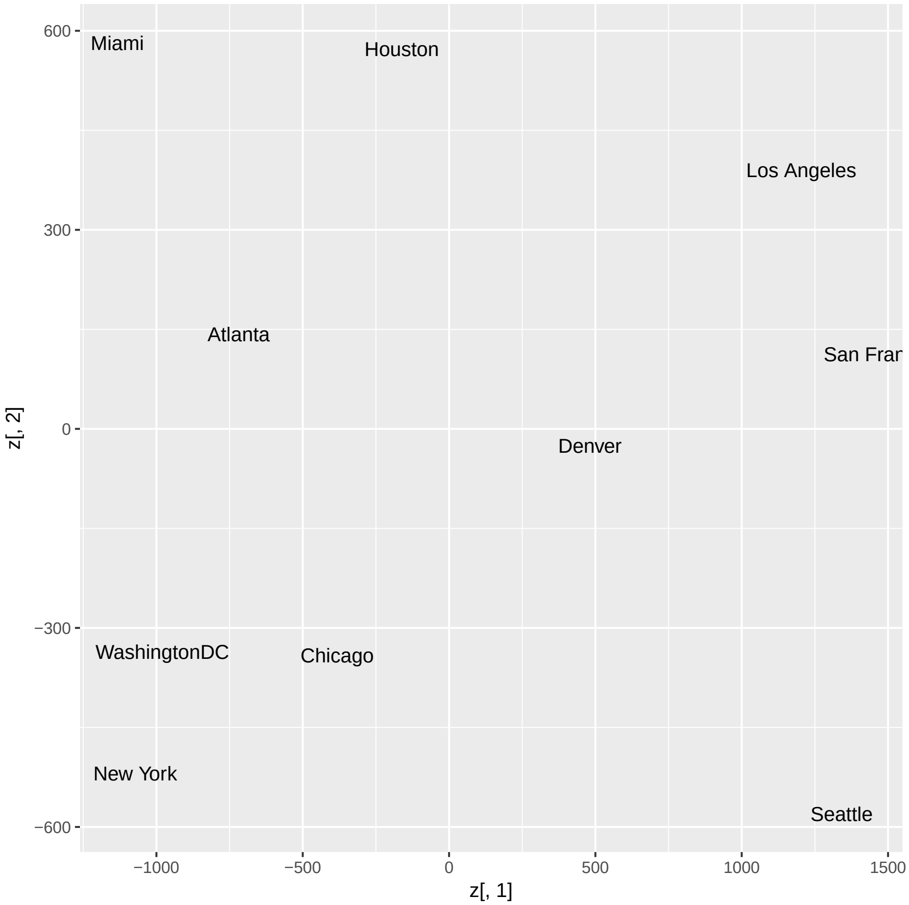
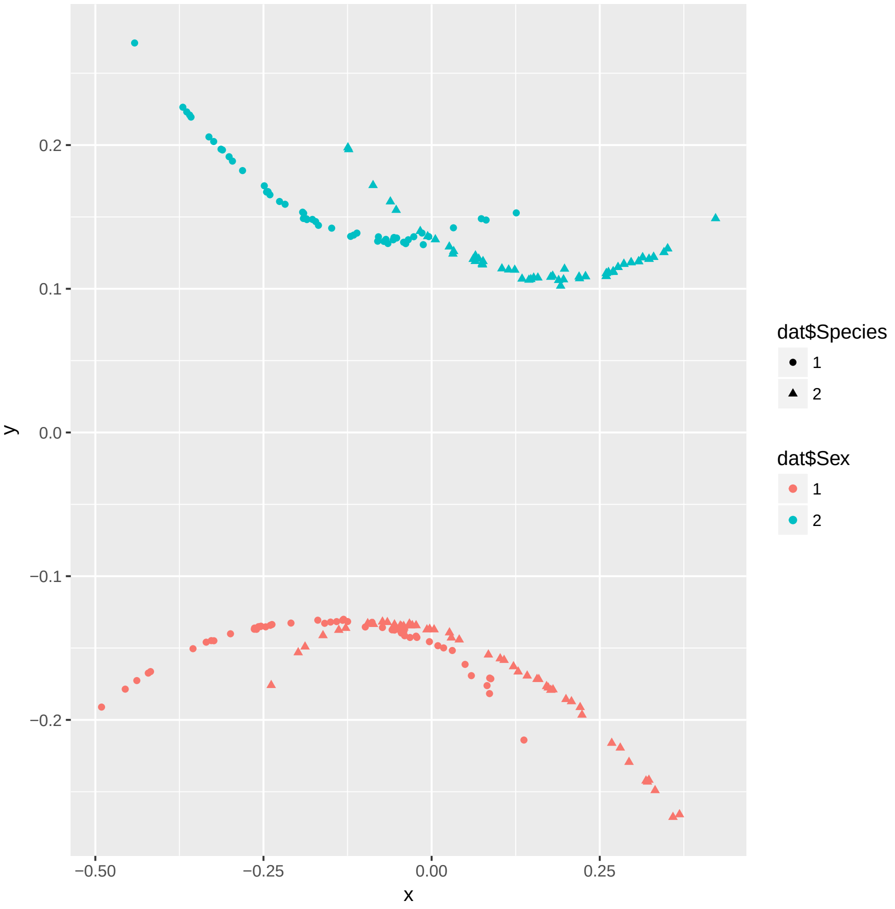
1.

```
require(ggplot2)

## Loading required package:  ggplot2

a1=c(0,587,1212,701,1936,604,748,2139,2182,543)
a2=c(0,0,920,940,1745,1188,713,1858,1737,597)
a3=c(0,0,0,879,831,1726,1631,949,1021,1494)
a4=c(0,0,0,0,1374,968,1420,1645,1891,1220)
a5=c(0,0,0,0,0,2339,2451,347,959,2300)
a6=c(0,0,0,0,0,0,1092,2594,2734,923)
a7=c(0,0,0,0,0,0,0,2571,2408,205)
a8=c(0,0,0,0,0,0,0,0,678,2442)
a9=c(rep(0,9),2329)
a10=rep(0,10)
matr=cbind(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10)
mat=(matr+t(matr))^2
mat=data.frame(mat)
rownames(mat)=c("Atlanta","Chicago","Denver","Houston","Los Angeles","Miami","New York","San
#Construct the data matrix
mat1=apply(mat,2,function(x) x-mean(x))
mat2=apply(mat1,1,function(x) x-mean(x))
mat_ip=-1/2*mat2
ei=eigen(mat_ip)
z=cbind(sqrt(ei$values[1])*ei$vectors[,1],sqrt(ei$values[2])*ei$vectors[,2])
qplot(x=z[,1], y=z[,2], label=rownames(mat), geom="text")
```



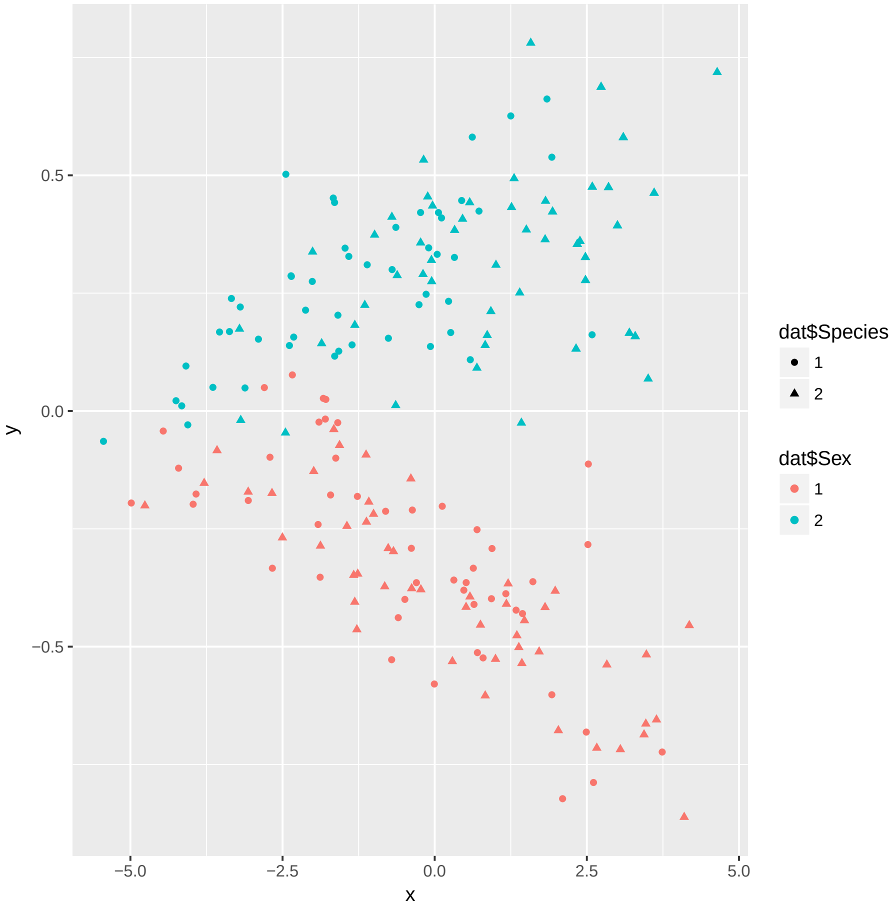
2.

```
require(cluster)

## Loading required package:  cluster

dat=read.table("crabs.txt",header = T)
dat=cbind(dat[,1:2],scale(dat[,3:7]))
dat[,1]=factor(dat[,1])
dat[,2]=factor(dat[,2]) #Import the data, and set appropriate type to the variadbles.
mds = cmdscale(daisy(dat[,1:7], metric = "gower"), k=2)
colnames(mds)=c("x","y")
ggplot(data=as.data.frame(mds), aes(x=x, y=y, color=dat$Sex, shape=dat$Species)) +geom_point
```



We do the MDS first: the data is the same as homework 1, we don't include the basic analysis here again, and because of the problem of units, we need to scale those quantative variables so that they can make similar effects to the final results. For the choice of variables, I think this should depend on the choice of distances' measure. Firstly, I use the Gower's distance here, because we have two variables' type: quantative and binary. From the plot, we can see with MDS different sexes are clearly separated, and different species have a small part of overlapping, which means some crabs from different species have similar overall quality; in other word, the species cannot be well discriminated.

```
mds = cmdscale(dist(dat[,3:7]), k=2)
colnames(mds)=c("x","y")
ggplot(data=as.data.frame(mds), aes(x=x, y=y, color=dat$Sex, shape=dat$Species)) +geom_point
```



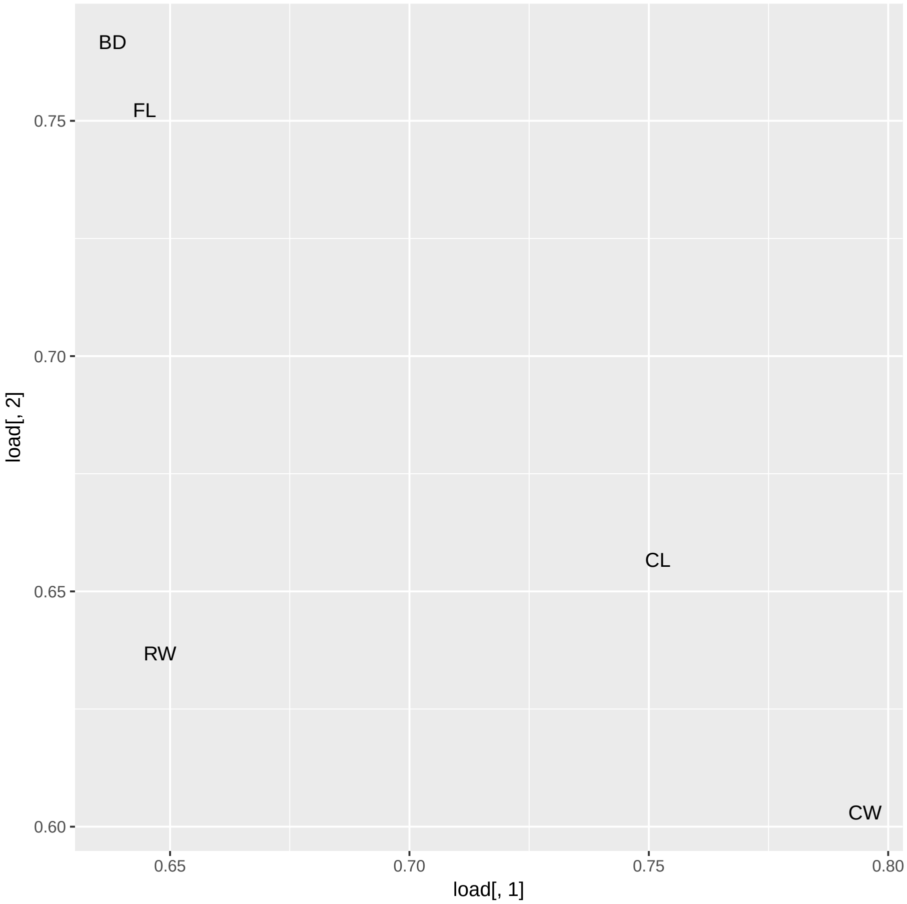
Then we consider dropping the first two binary variables, and use the regular p2 euclidean distance for the remaining quantative variables. We can see without two binary variables, the points representing crabs' dimensions are more dispersive, but I believe some information has dropped compared to the last one. And as we know, the classical MDS has the same projection as PCA, so the plot is the same and conclusions are the same: male and female are clearly divided into two different parts, and species are totally mixed up, which is different with things shown by last plot.

```
cor(dat[,3:7])#variables should be quantitative

##            FL        RW        CL        CW        BD
## FL 1.0000000 0.9161758 0.9791077 0.9659242 0.9878567
## RW 0.9161758 1.0000000 0.8984834 0.9038372 0.8990104
## CL 0.9791077 0.8984834 1.0000000 0.9951980 0.9838880
## CW 0.9659242 0.9038372 0.9951980 1.0000000 0.9693672
## BD 0.9878567 0.8990104 0.9838880 0.9693672 1.0000000
```

```
fa= factanal(factors=3, covmat=cov(dat[,3:7]))
```

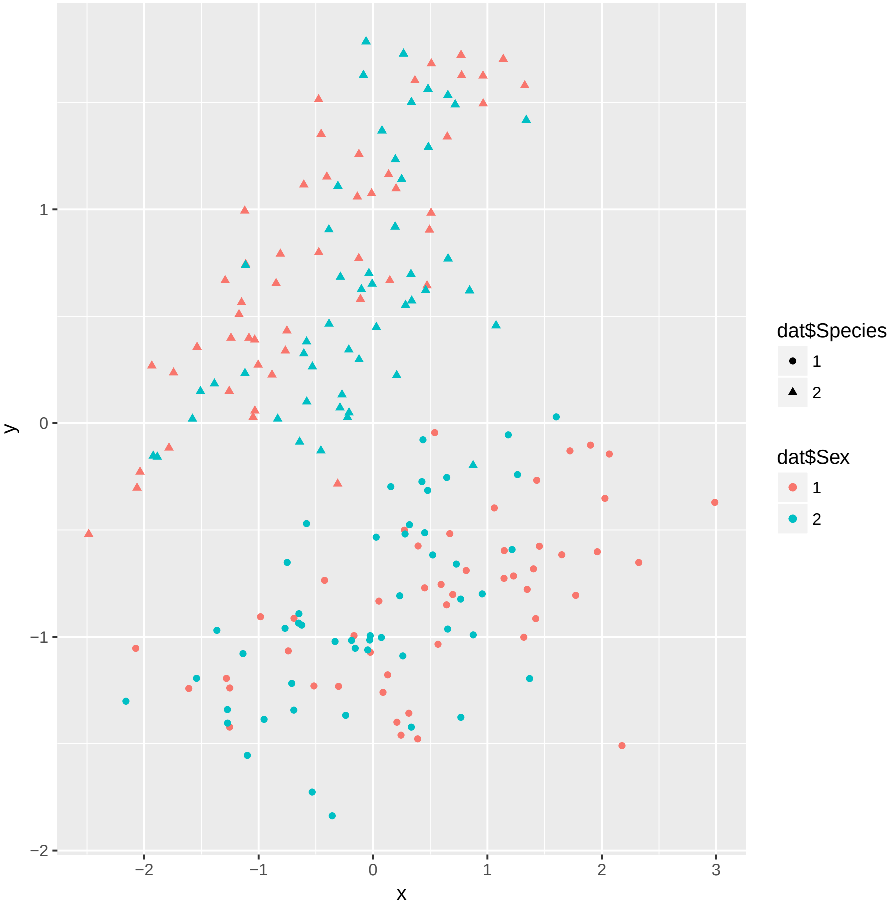
```
## Error in factanal(factors = 3, covmat = cov(dat[, 3:7])):  3 factors
are too many for 5 variables
```

```
fa1= factanal(factors=2, covmat=cov(dat[,3:7]))
fa2= factanal(factors=2, covmat=cov(dat[,3:7]),rotation = "promax")
load=fa1$loadings[,1:2]
qplot(x=load[,1], y=load[,2], label=rownames(load), geom="text")
```



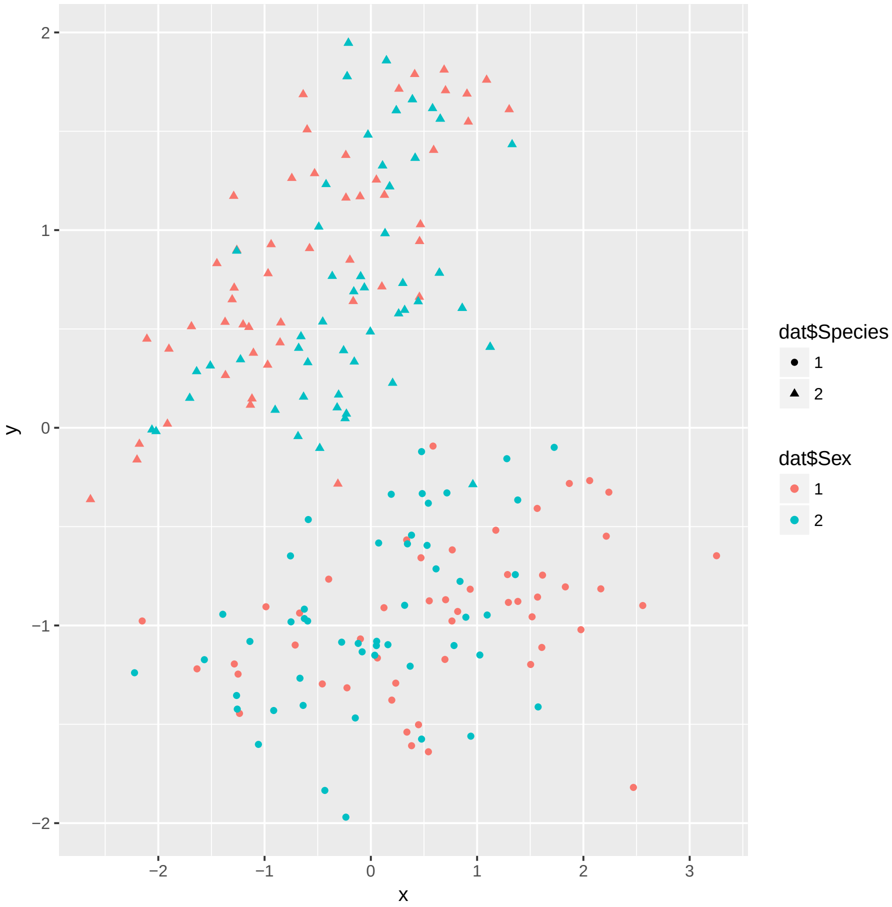
Secondly, we do the FA. And as we know FA can only be applied to quantative variables, so we drop the first two binary variables. And the correlations between variables are close enough to do FA. After trying, factor numbers can be 1 or 2, we choose 2 so that we can have more convincible plot. And we find both two rotation methods don't have useful interpretaion, so we just use the default varimax. From the plot of loadings for different variables, we find BD and FL have similar underlying factors, and others are not so similar with each other.

```
fa_scores = factanal(x=dat[,3:7], factors=2, scores='regression')
scores=fa_scores$scores
colnames(scores)=c("x","y")
ggplot(data=as.data.frame(scores), aes(x=x, y=y, color=dat$Sex, shape=dat$Species)) +geom_po
```



```
fa_scores1 = factanal(x=dat[,3:7], factors=2, scores="Bartlett")
scores1=fa_scores1$scores
colnames(scores1)=c("x","y")
factanal(factors=2, x=dat[,3:7])

##
## Call:
## factanal(x = dat[, 3:7], factors = 2)
##
```

```
## Uniquenesses:
##    FL    RW    CL    CW    BD
## 0.018 0.175 0.005 0.005 0.005
##
## Loadings:
##    Factor1 Factor2
## FL 0.645   0.752
## RW 0.648   0.637
## CL 0.752   0.657
## CW 0.795   0.603
## BD 0.638   0.767
##
##                Factor1 Factor2
## SS loadings      2.440   2.355
## Proportion Var   0.488   0.471
## Cumulative Var   0.488   0.959
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 131.76 on 1 degree of freedom.
## The p-value is 1.69e-30

ggplot(data=as.data.frame(scores1), aes(x=x, y=y, color=dat$Sex, shape=dat$Species)) +geom_p
```



We use two ways to get the FA scores, and final plots show they are pretty the same, just a slightly difference in scale. And a interesting thing is that the FA divides two different species, instead of sexes shown in the plots of PCA (MDS): I guess FA may be able to pick out features that are different with PCA (classical enclidean MDS); but the main reason is that from above p value, we know the FA fits the data very badly, so it cannot even show out the differences between sexes. It's inappropriate to use FA for this dataset. So the previous conclusion about variables by FA is also doubtable.