HUILONG AN

W4415 HOMEWORK 1

Q1:

Code:

```
#the input data should be n by p matrix or data.frame, where n is the number of samples#
md=function(data){
 data=as.matrix(data);
 cov=cov(data);
 mean=colMeans(data);
 print("the Mahalanobis distance of every inputs:");
 print(mahalanobis(data,mean,cov));
```

Q2:

```
sex=rep(c(0,1),c(410,590))
sex=as.factor(sex)
levels(sex)
levels(sex)=c("male","female")
prefer=c(rep(c(1,2,3),c(180,170,60)),rep(c(1,2,3),c(230,300,60)))
prefer=as.factor(prefer)
levels(prefer)=c("Republican","Democrat","Independent")
df=data.frame(sex,prefer)
fix(df)
table(df$sex,df$prefer)
```

```
> table(df$sex,df$prefer)

        Republican Democrat Independent
  male         180      170          60
  female       230      300          60
```

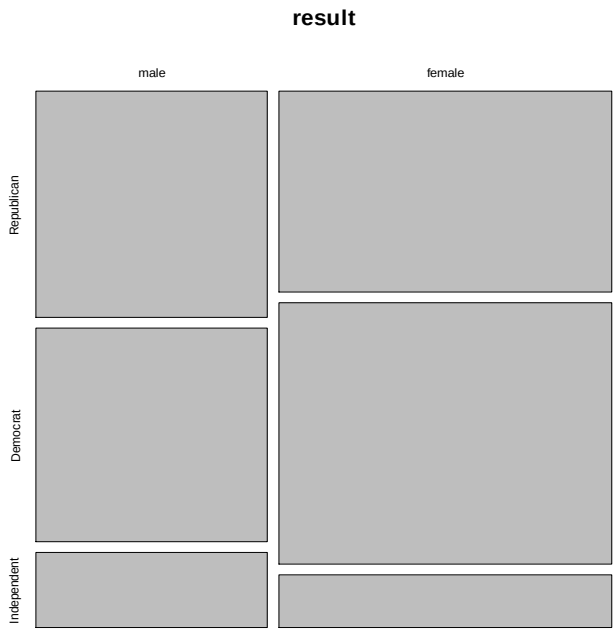
```
library(MASS)
tbl=table(df$sex,df$prefer)
chisq.test(tb1)
```

```
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 9.9783, df = 2, p-value = 0.006811
```

chisq.test(tbl)$residuals

```
> chisq.test(tbl)$residuals

         Republican   Democrat Independent
male      0.9178318 -1.6352532   1.5397181
female   -0.7651191  1.3631728  -1.2835333
```

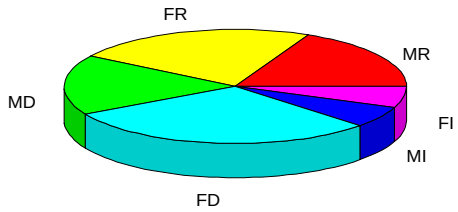
mosaicplot(tbl,main="result")

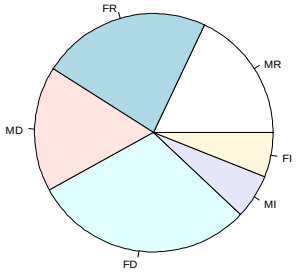


pie3D(tbl,labels=c("MR","FR","MD","FD","MI","FI"))

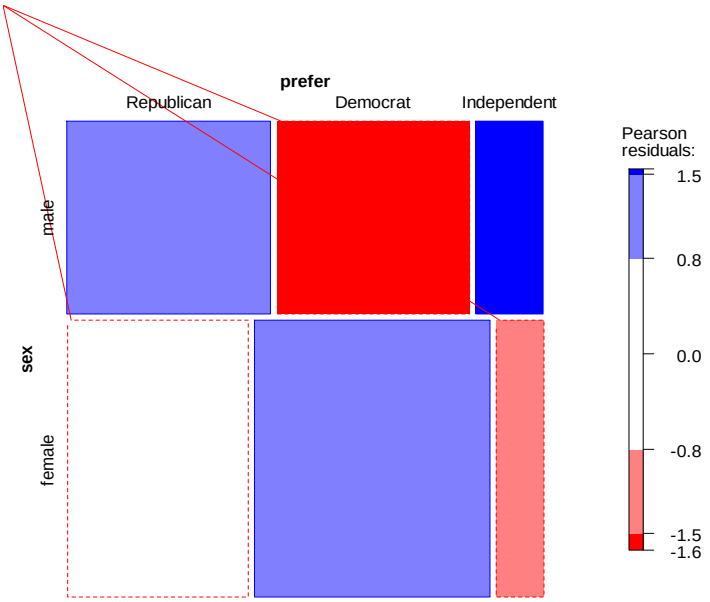
library(plotrix)

pie3D(tbl,labels=c("MR","FR","MD","FD","MI","FI"))





mosaic(~sex+prefer,data=df,shade=TRUE,gp=shading_Friendly,gp_args=list(interpolate=c(0.8,1.5)))



From the pie and mosaic plot, we can see the distribution of each groups.

As we can see the result of Chi-square test, the P-value is 0.0068, which indicates that we should conclude the alternative hypothesis. So the independece between sex and voting preference is not statistically significant at the significance level coefficient less more than 0.0068.

And we can see from the mosatic plot (which displaies the Pearson residuals) that the red cell and purple cells make the large part contribution to misfit, which means they show relatively obvious dependency between sex and preference.

Q3:

The first step should be loading data:

getwd()

setwd("c:/users/an/desktop")

data=read.table("datahw1.txt",header=T)

fix(data)

names(data)=c("x1","x2","x3")

The we can have a quick look at some basic properities of these data:

summary(data)

```
> summary(data)
       x1              x2              x3
 Min.   :0.7580   Min.   :103.5   Min.   :48.93
 1st Qu.:0.7957   1st Qu.:115.1   1st Qu.:65.22
 Median :0.8155   Median :121.7   Median :70.95
 Mean   :0.8132   Mean   :121.0   Mean   :68.15
 3rd Qu.:0.8265   3rd Qu.:127.0   3rd Qu.:74.90
 Max.   :0.9710   Max.   :135.1   Max.   :80.33
> 
```

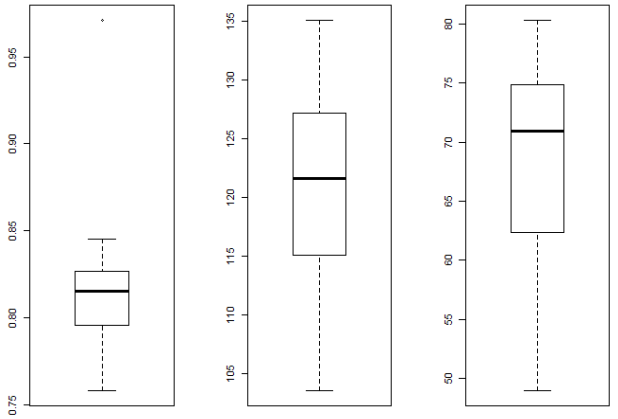
Then a boxplot:

par(mfrow=c(1,3))

boxplot(data$x1)

boxplot(data$x2)

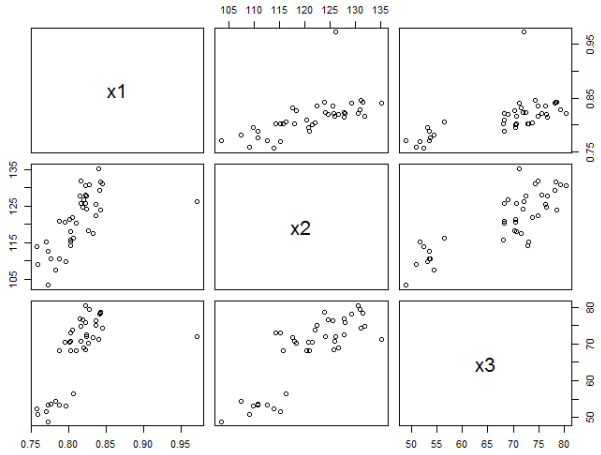
boxplot(data$x3)



The distribtion of these three variables are regular, but the fitst and third variables are not so evenly distributed. It might casue some problems, i.e, the data might not evenly distributed in the space.

Then we can check the correlation between each variables:

```
> cor(data)
          x1        x2        x3
x1 1.0000000 0.6206762 0.6223968
x2 0.6206762 1.0000000 0.8277573
x3 0.6223968 0.8277573 1.0000000
```

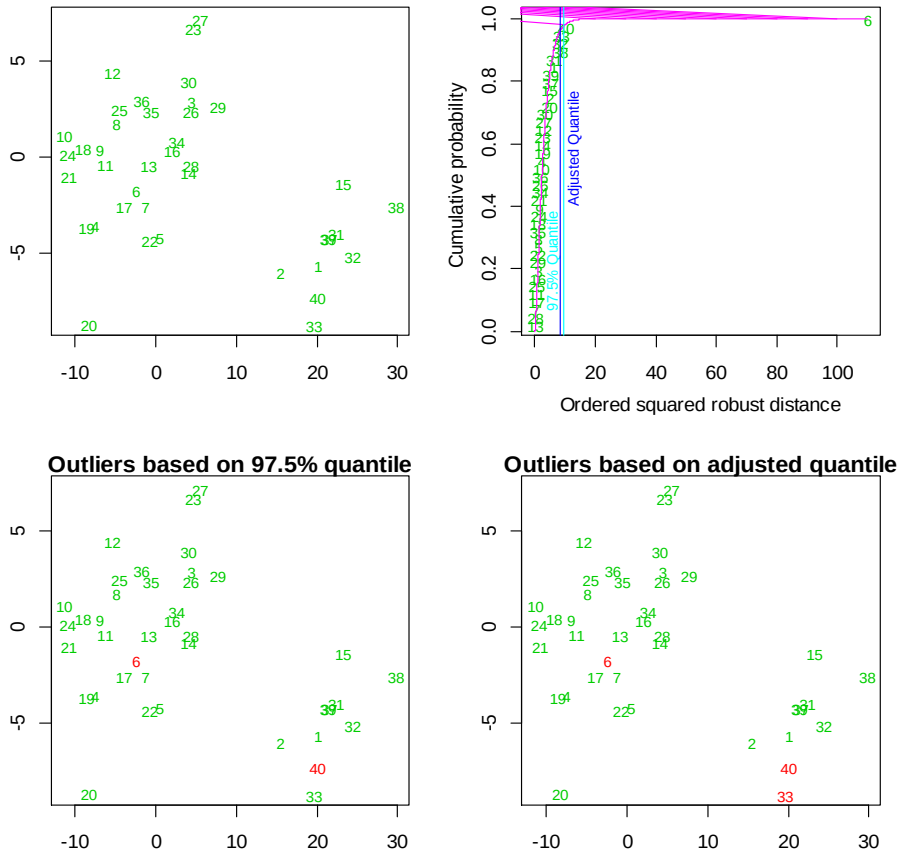


We can find that they all highly linearly and positively correlated between each other.

Then we can have a look at if there is some outliers.

Let us try first Mahalanobis Distance.

aq.plot(data)



We can find that 6th,33th, 40th might be the outliers.

However, the claim that square Mahalanobis distance is a Chi-square distribution is based on the assumption that the variables are jointly normally distributed. If these variables are normally distributed, then every variables should be normally distributed. Let us do some test on the normality of these variables.

shapiro.test(data$x1)

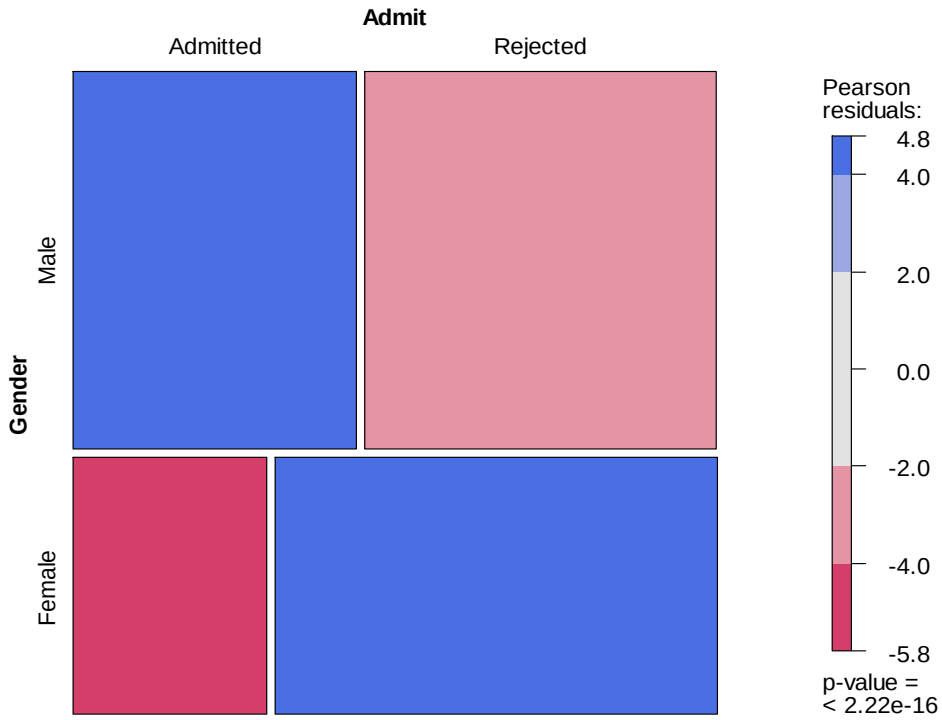
```
> shapiro.test(data$x1)

        Shapiro-Wilk normality test

data:  data$x1
W = 0.81259, p-value = 1.253e-05
```

The P-value here suggests to reject null hypothesis. So, the normality of X1 does not hold, which means the assumption that MD is Chi-square distribution does not hold. So, this result might not be very precise.

Q4:

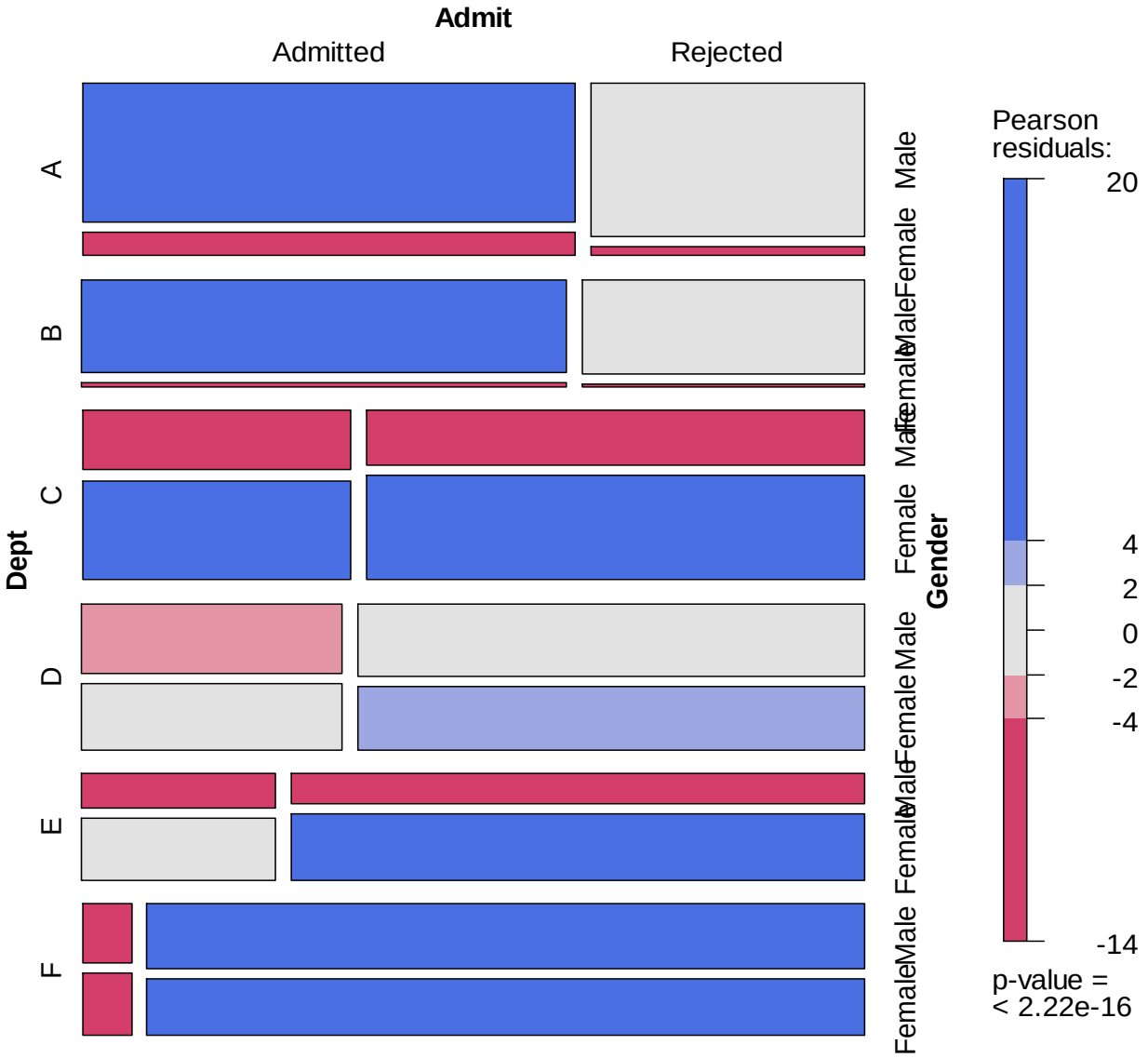
(1)

mosaic(~Gender+Admit,data=data,shade=T)

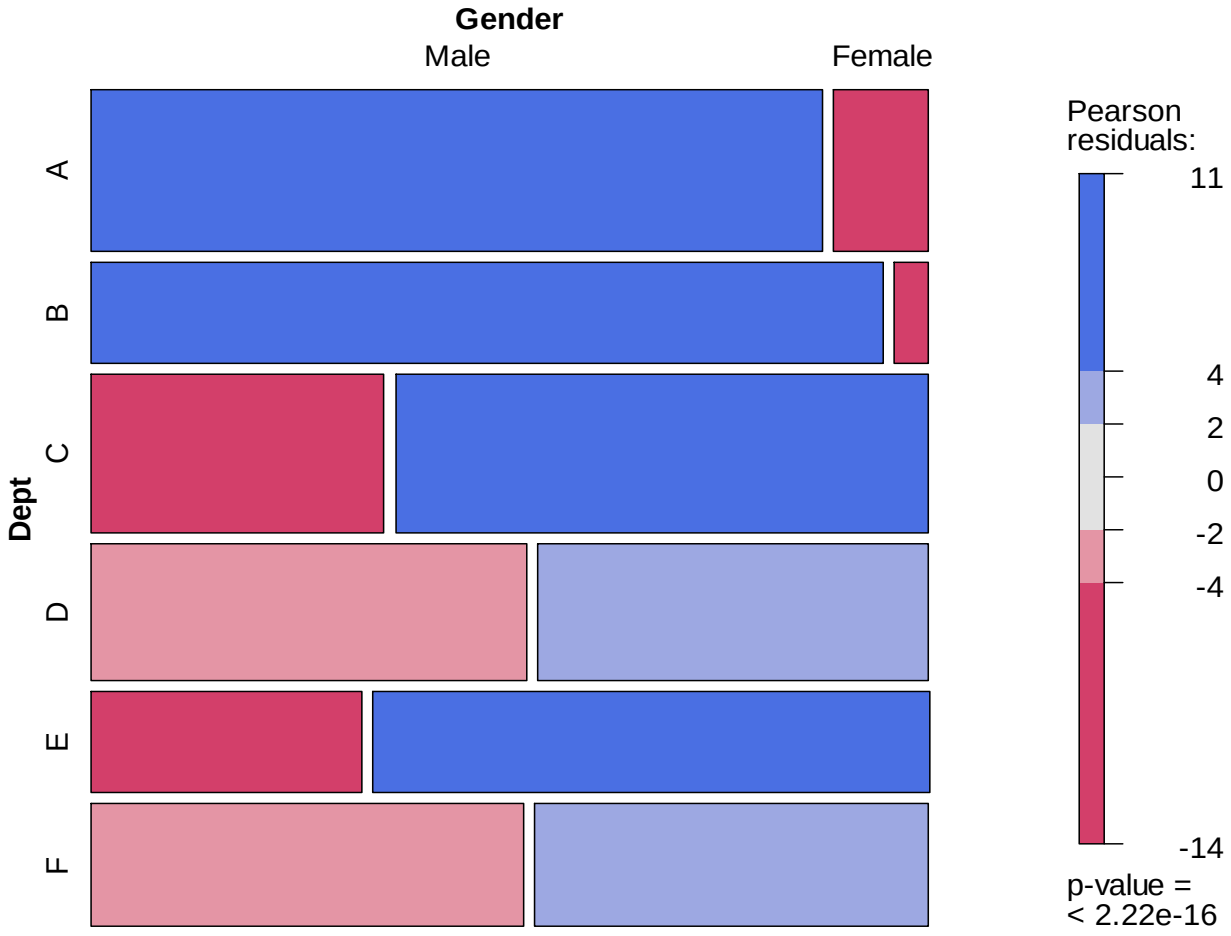


The p-value here is extremly smally, so we have overwhelming evidence show the gender and admit are not independent.

(2)

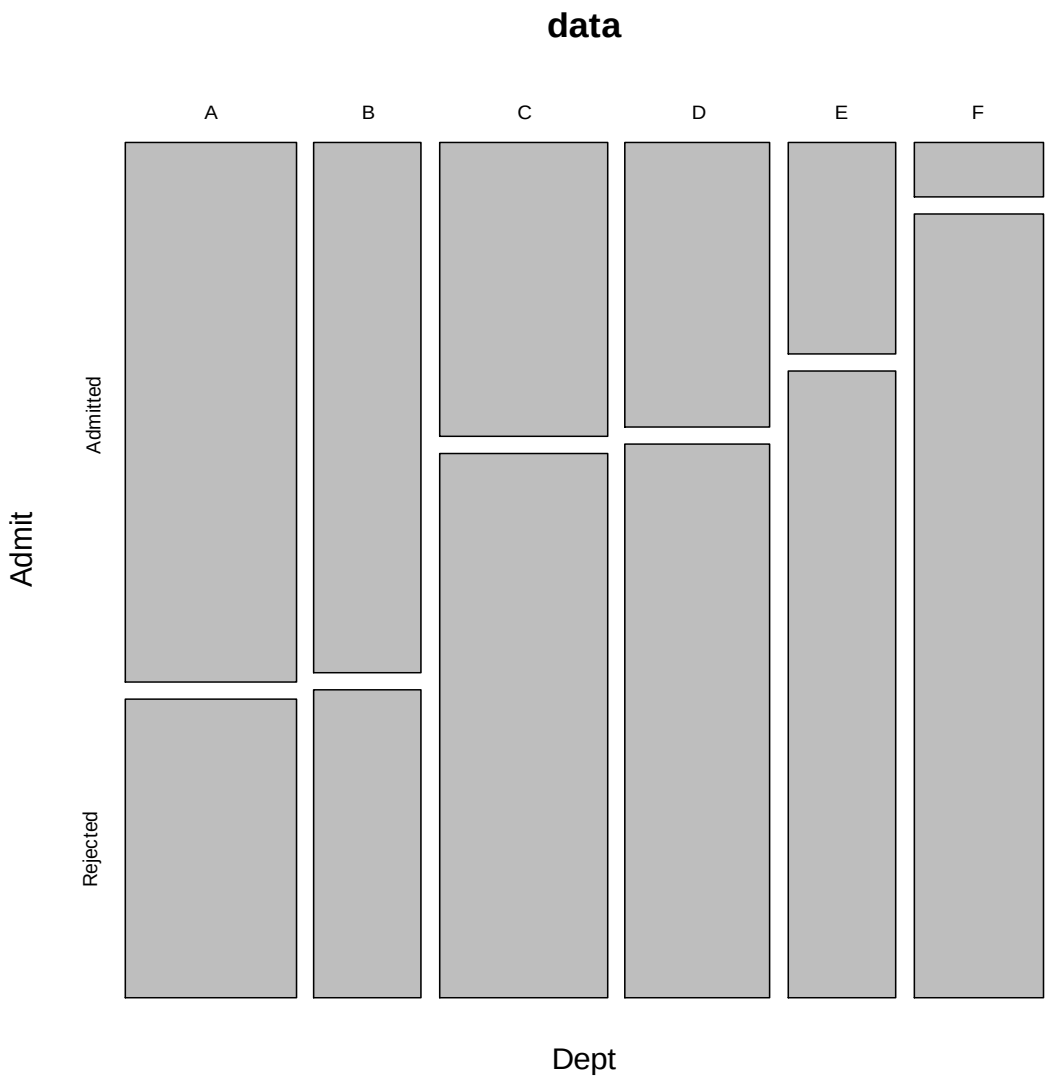


From the p-value, we can say that gender, dept, and Admit are not mutually independent. Actually, we learned from (1) that gender and admit is not independent.



We can learn from the P-value that there is overwhelming evidence shows dept and gender are not independent. The Pearson residuals are all relatively large than expected.

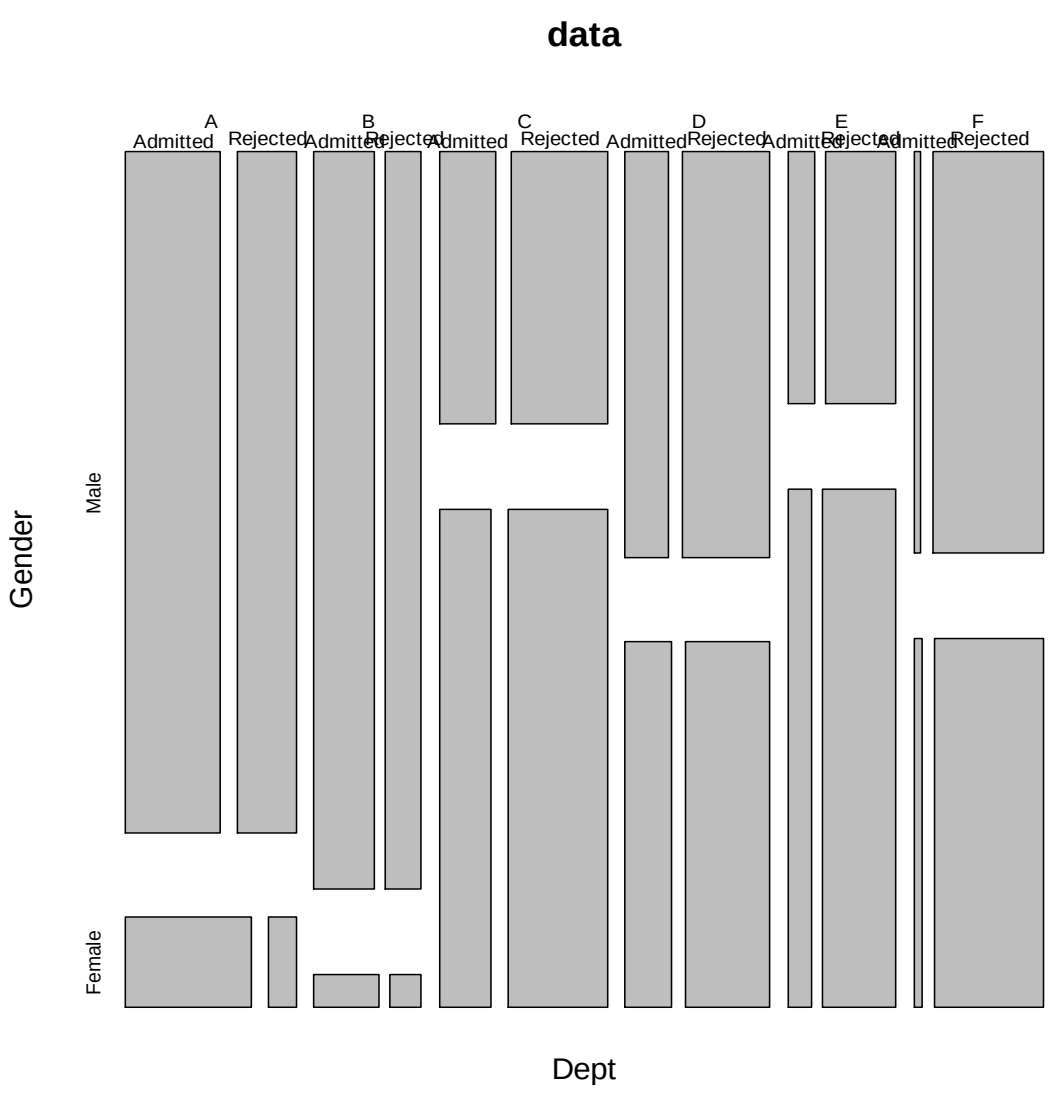
(3)



T

We can learn from the area of every cell that department A and C have the most applicants, while people who apply F department seems the most hard to get admissions.

(4)



We learned from (2) that gender and departments are not independent, which means some departments are just attractive to a certain gender students. And we also learned from (3) that the number of applicants varies a lot among different departments. If we combine these two result and also have look at the plot above, we can conclude that some departments attract a certain gender students and they actually admit this certain gender students while some other departments, like F, just treat everyone equally.

(5)

Actually, yes. The departments, like A,B favors male students, while departments like C,E much favors female students.

(6)

I think there have some latent variables, like the size of the department and popularity, actually have some influence on these three variables, which might also have some influence on their relationships. Maybe we can see the relationship among these three variables much clear if we exclude the influence of these latent variables.

Q5

We can see from the covariance matrix sigma that the covariance of x1 and x2, x2 and x3 are 0. Because "independent" and "uncorrelated" are equivalent in normal distribution. So, x1 and x2 are independent, so are x2 and x3. But x1 and x3 are not independent.

Because x2 is independent from x1 and x3, so (x1,x3) and x2 are independent.

$$\begin{pmatrix} x1 \\ x1+3x2-5x3 \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & -5 \end{pmatrix}*\begin{pmatrix} x1 \\ x2 \\ x3 \end{pmatrix}$$

So $\begin{pmatrix} x1 \\ x1+3x2-5x3 \end{pmatrix}$~ Normal(Au+A*sigma*transpose(A)), where A=$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & -5 \end{pmatrix}$

A*sigma*transpose(A)=$\begin{pmatrix} 4 & 9 \\ 9 & 109 \end{pmatrix}$, so they are not independent.

Q6.

f(x1,x2,x3…xp) $\propto$ exp(-0.5*$(x-u)^{-1}\Sigma(x-u)$) (1)

and f(x1|x2,x3,…xp) is also normal density function. (The property of joint normal distribution).

f(x1|x2,x3,…xp) $\propto$ exp(-0.5*(x1-u1)^2/ σ^2), where σ^2 is the conditional variance of x1 given x2,x3,…xp. And u is the conditional expectation of x1 given x2,x3,x4,…xp.

Notice that we can determine the value of the conditional variance of x1 by the term of x1^2. And we know that the only part of (1) that the form of x1^2 is $\Omega_{11} * x1\text{^}2$. So, by comparing the form of x1^2/σ^2 , the conditional variance of x1 given x2,x3…xp is $1/\Omega_{11}$.