Homework 1

Due on Feb 4th W4415 Multivariate Statistical Inference

Zhang Yunyan UNI: yz2861

Q1.

R code (The explanations and results are quoted after #)

```
MD<-function(x,mu,covar){
  mahala=sqrt(t(x-mu)%*%inv(covar)%*%(x-mu))
  return(mahala)
}
#   For the input arguments,
#   x stands for a vector or matrix of data (x_1, x_2, ..., x_N)^T;
#   mu stands for the mean vector (μ_1, μ_2, ..., μ_N)^T of the input vector(matrix) x;
#   covar stands for the covariance of the input vector(matrix) x.

#  Example
#  Data collected from http://www.jennessent.com/arcview/mahalanobis_description.htm
x<-c(410,400)
mu<-c(500,500)
covar<-matrix(c(6291.55737,3754.32851,3754.32851,6280.77066),ncol = 2,nrow = 2)
MD(x,mu,covar)
#  1.348286
```

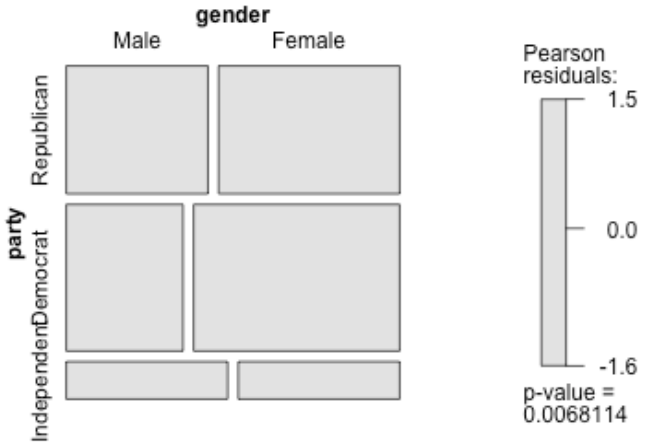
The square of Mahalanobis Distance is 1.817874, which is equal to the result on the website (Actually there is some rounding in the calculation process on the website).

Q2.

R code

```
library(vcd)
M<-as.table(rbind(c(180,170,60),c(230,300,60)))
dimnames(M)<-
      list(gender=c("Male","Female"),party=c("Republican","Democrat","Independent"))
Xsq<-chisq.test(M)
# Pearson's Chi-squared test
# data:  M
# X-squared = 9.9783, df = 2, p-value = 0.006811

Xsq$observed
#            party
# gender   Republican Democrat Independent
# Male            180     170        60
# Female          230     300        60
```

```
Xsq$expected
#             party
# gender   Republican Democrat Independent
# Male           168.1    192.7       49.2
# Female         241.9    277.3       70.8

Xsq$residuals
#             party
# gender   Republican   Democrat Independent
# Male     0.9178318 -1.6352532   1.5397181
# Female -0.7651191  1.3631728  -1.2835333

Xsq$stdres
#             party
# gender   Republican  Democrat Independent
# Male        1.555647 -2.924294    2.136849
# Female  -1.555647  2.924294   -2.136849

mosaicplot(M)
```

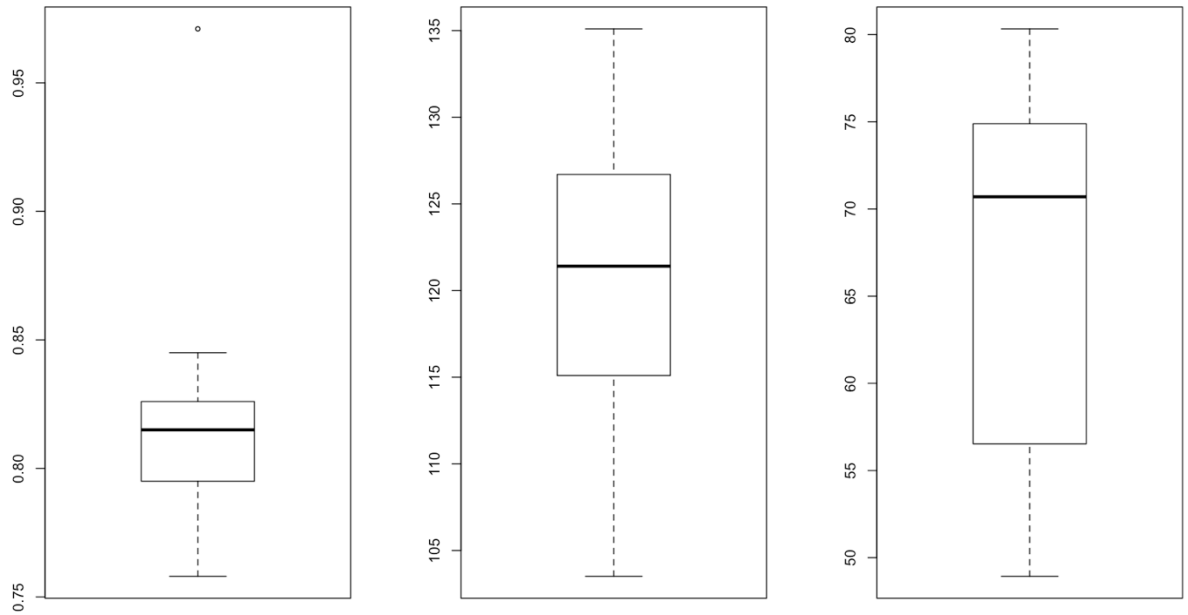


In this case, the null and alternate hypothesis are:
H_0: voting preference and gender are independent
H_1: voting preference and gender are not independent
The result of the chi-square test rejects the null hypothesis at $\alpha = 1\%$ level, as the p-value is 0.68%. We could draw a conclusion that there is a relationship between voting number and gender. (However, the mosaic plot won't show a colored cell count unless Pearson residuals are great than 2 or smaller than -2.)

Q3.
R code

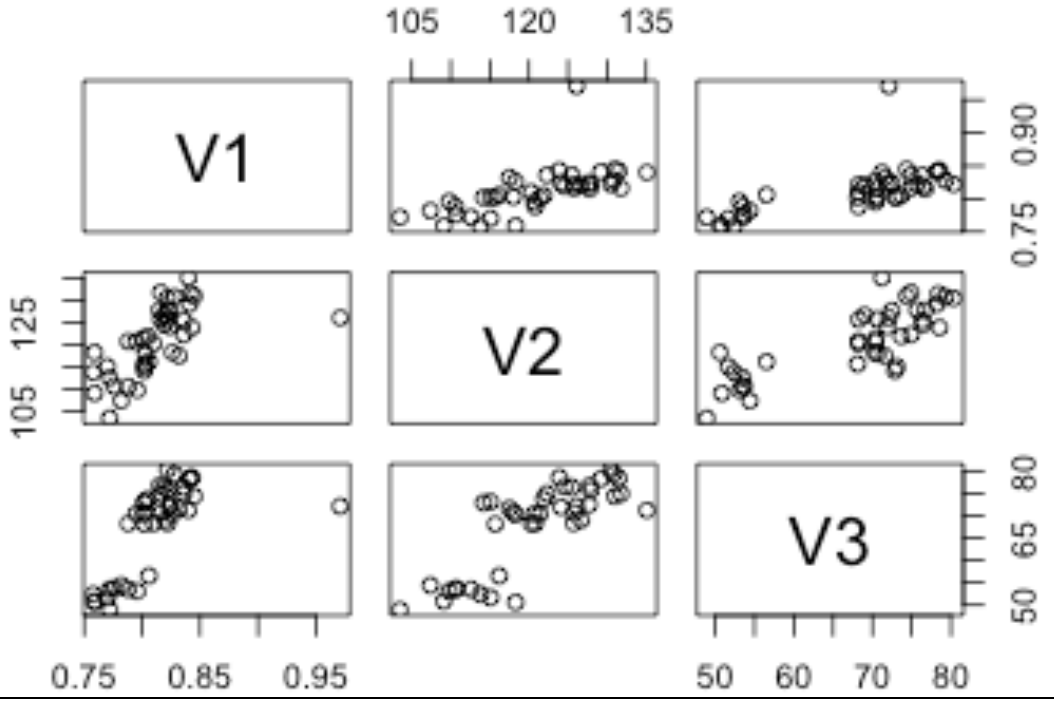
```
datahw1 <- read.table("~/datahw1.txt",header=FALSE)
```

```
library(fBasics)
basicStats(datahw1)
#                         V1         V2         V3
# nobs            41.000000  41.000000  41.000000
# NAs              0.000000   0.000000   0.000000
# Minimum          0.758000 103.510000  48.930000
# Maximum          0.971000 135.100000  80.330000
# 1. Quartile      0.795000 115.100000  56.530000
# 3. Quartile      0.826000 126.700000  74.890000
# Mean             0.811854 120.953415  67.723171
# Median           0.815000 121.410000  70.700000
# Sum             33.286000 4959.090000 2776.650000
# SE Mean          0.005554   1.202854   1.529041
# LCL Mean         0.800629 118.522356  64.632863
# UCL Mean         0.823078 123.384473  70.813479
# Variance         0.001265  59.321148  95.856667
# Stdev            0.035561   7.702022   9.790642
# Skewness         1.875342  -0.248375  -0.723352
# Kurtosis         7.405886  -0.875625  -1.030870

par(mfrow=c(1,3))
boxplot(datahw1$V1)
boxplot(datahw1$V2)
boxplot(datahw1$V3)
```

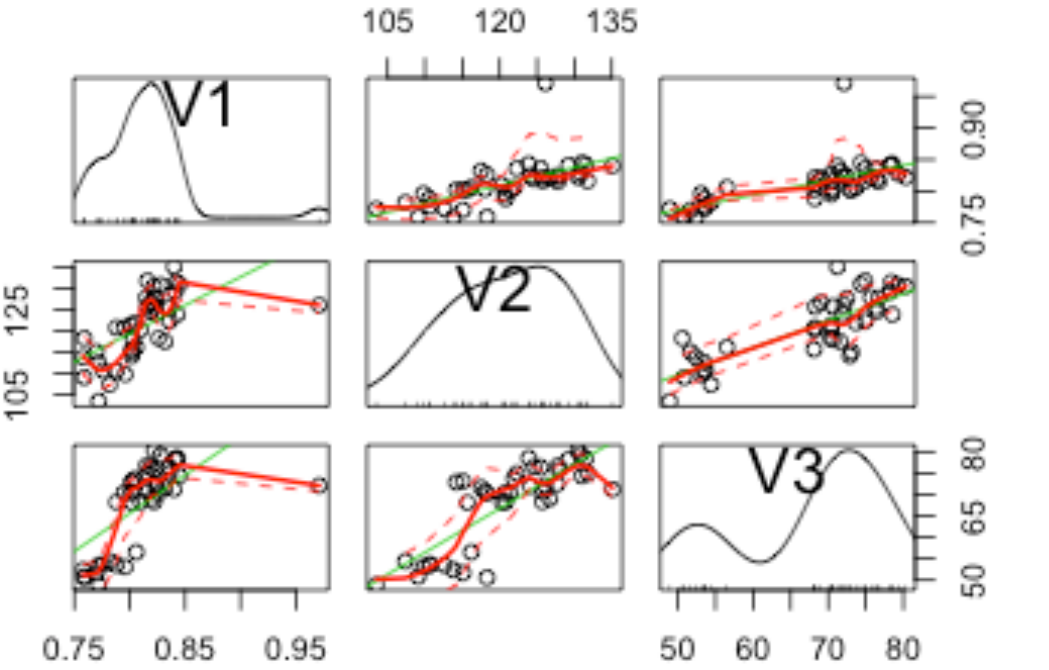


The boxplots (from left to right is V1, V2, V3 from datahw1 respectively) show that there is an outlier in V1.

```
pairs(~V1+V2+V3, data=datahw1)
```

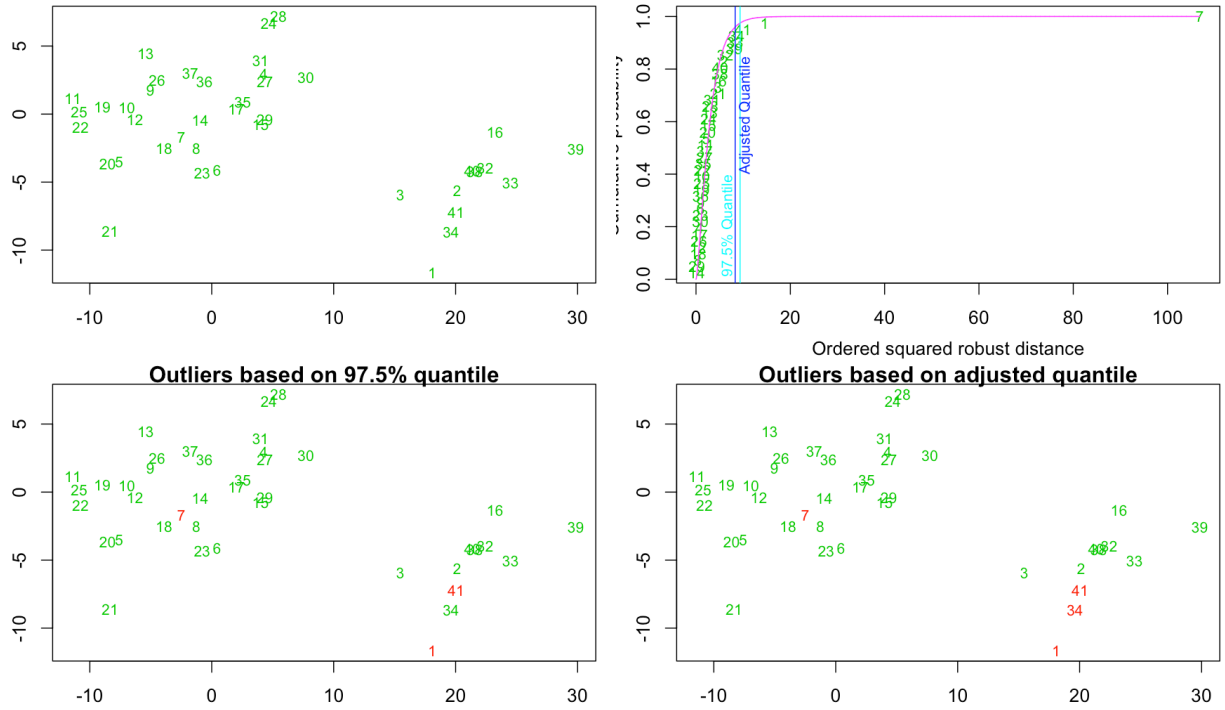



```
library(car)
scatterplotMatrix(datahw1)
```



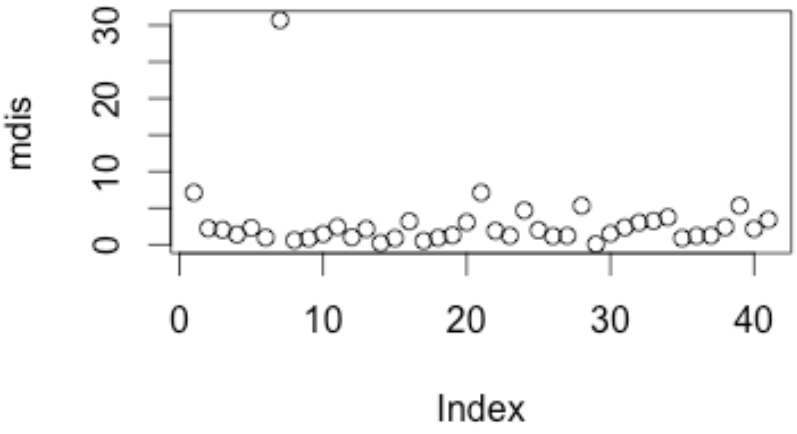
From the scatterplot matrices displayed above, there is an outlier which has an extremely large value on variable "V1", because from the V1~V2 picture (1,2) (which is on the 1st row and 2nd column on the 3x3 plot), the point shows a great distance to the mass of other points.
Also, when the outlier is excluded, there is relationship between V1, V2, V3, from the picture (1,2), (1,3) and (2,3).

```
library(mvoutlier)
aq.plot(datahw1)
```



The adjusted quantile plots for outliers indicate that there are three (1, 7, 41) for 97.5% quantile, and four(1, 7, 34, 41) for adjusted quantile. In this case, the 97.5% and adjusted quantile is quite close.

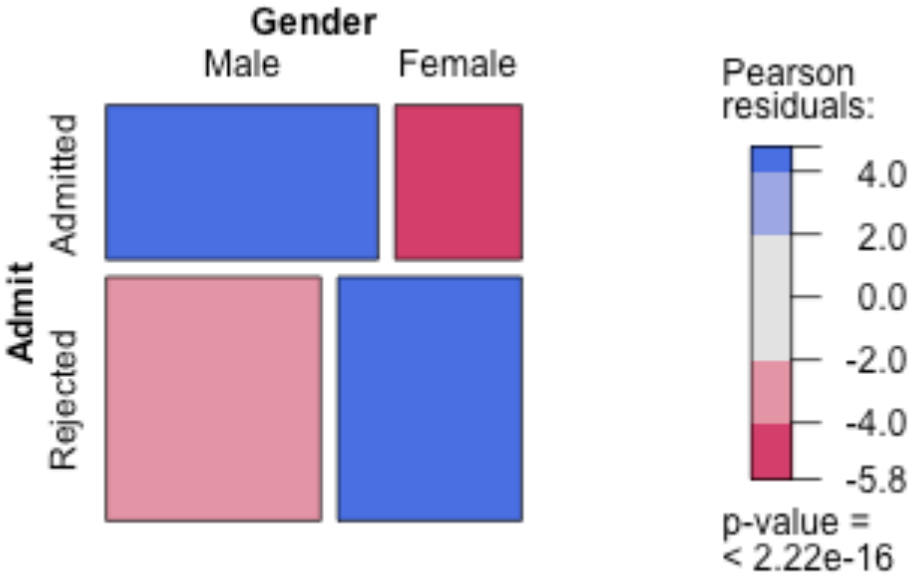
```
mdis<-mahalanobis(datahw1, colMeans(datahw1), cov(datahw1))
plot(mdis)
```



Also, the mahalanobis distance plot suggested that the 7th element in the dataset "datahw1" is an outlier.
An examination of the original data backed up the exploratory data analysis and visualization.

Q4.
(1)
R code

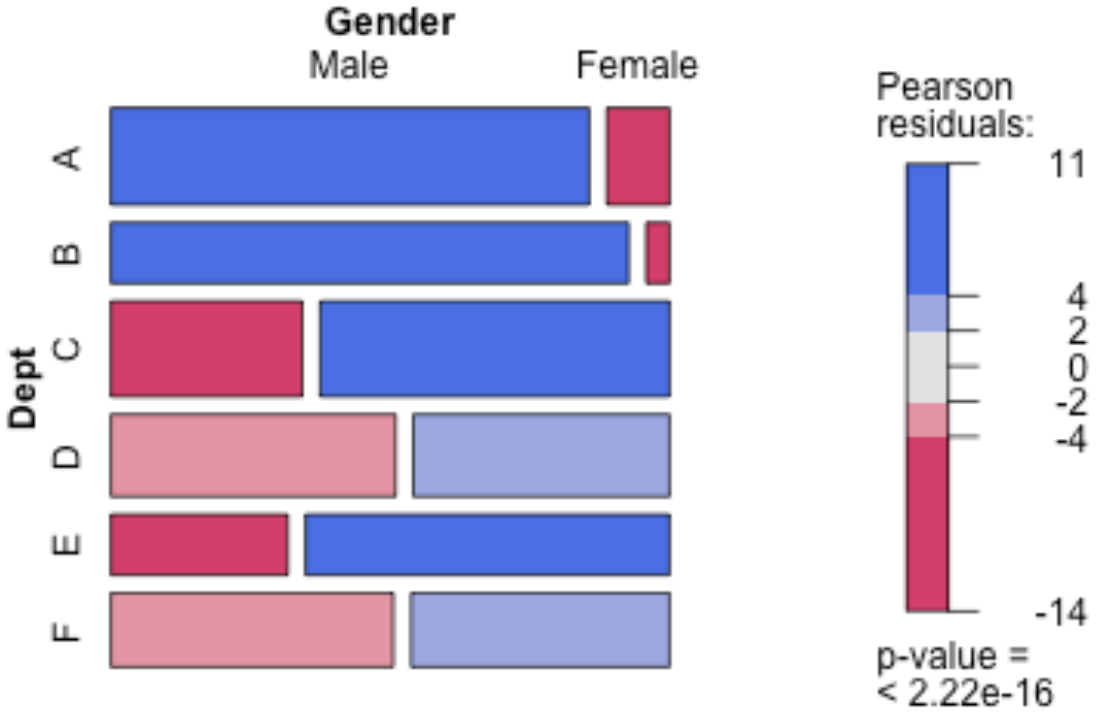
```
?UCBAdmissions
data("UCBAdmissions")
head(UCBAdmissions)
# [1] 512 313  89  19 353 207
colnames(UCBAdmissions)
# [1] "Male"   "Female"
rownames(UCBAdmissions)
# [1] "Admitted" "Rejected"
View(UCBAdmissions)
addf=as.data.frame(UCBAdmissions)
tab=xtabs(Freq~.,data=addf)
structable(~Admit+Gender,data=addf)
#          Gender Male Female
# Admit
# Admitted      1198    557
# Rejected      1493   1278
chisq.test(structable(~Admit+Gender,data=addf))
#          Pearson's Chi-squared test with Yates' continuity correction
# data:  structable(~Admit + Gender, data = addf)
# X-squared = 91.61, df = 1, p-value < 2.2e-16

mosaic(~Admit+Gender,data=addf,shade=TRUE)
```



The mosaic plot suggested that there is a significant difference between the admission decision on male and female applicants, as the observed admitted male applicants are surprisingly larger than

expected, while the observed admitted female applicants are surprisingly smaller than expected. In the case of rejected applicants, the result is inverse.
So there is dependency between gender and admission.
(2)
R code

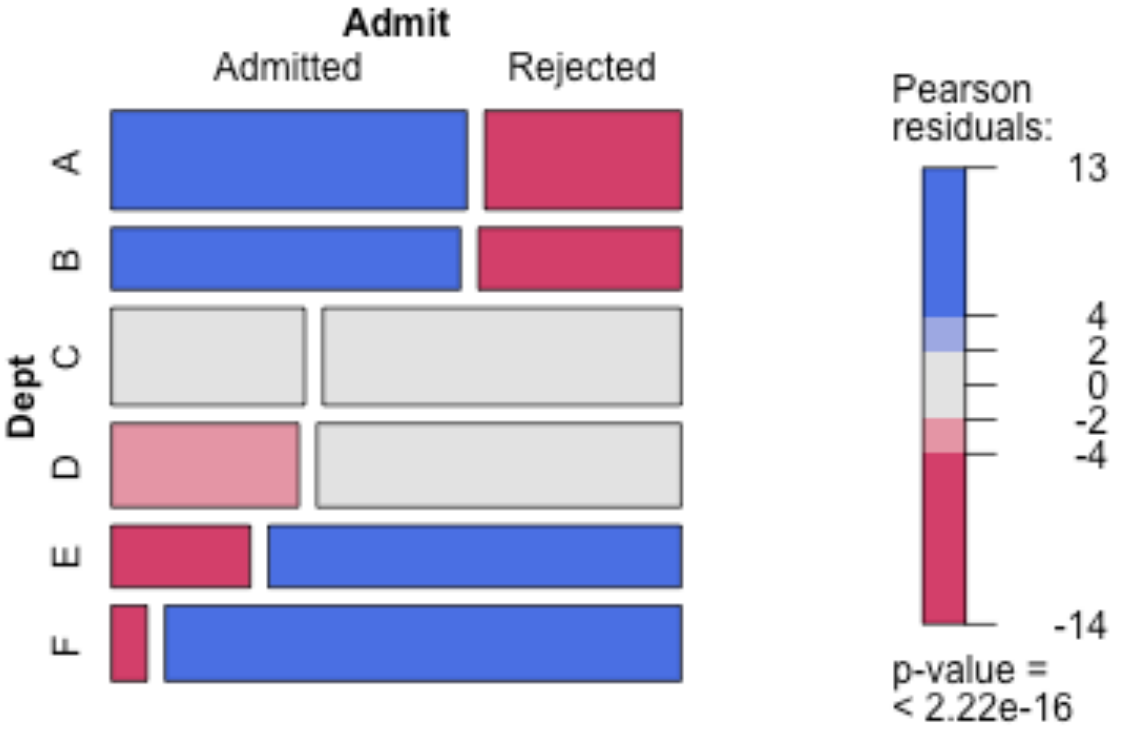
```
structable(~Dept+Gender,data=addf)
# Gender Male Female
# Dept
# A         825    108
# B         560     25
# C         325    593
# D         417    375
# E         191    393
# F         373    341
mosaic(~Dept+Gender,data=addf,shade=TRUE)
```



From the mosaic plot, male applicants are more likely to apply for Department A and B, while female ones would like to apply for Department C, D, E and F. The difference is significant.
(3)
R code

```
structable(~Dept+Admit,data=addf)
#    Admit Admitted Rejected
# Dept
# A          601     332
# B          370     215
# C          322     596
```

```
# D           269     523
# E           147     437
# F            46     668
chisq.test(structable(~Dept+Gender,data=addf))
#         Pearson's Chi-squared test
# data:  structable(~Dept + Gender, data = addf)
# X-squared = 1068.4, df = 5, p-value < 2.2e-16
mosaic(~Dept+Admit,data=addf,shade=TRUE)
```



The E and especially F Departments are the most competitive one, while it would be easier to get into Department A and B, for the Pearson residuals of Department A and B admitted is large and rejected is small, in the case of Department E and F the result is inverse.

(4)

Together with (2) and (3), I find that the departments that male applicants would like to apply for are less competitive, so they have more chances to get admission.

It seems that there is a relationship between gender and admission when the selectiveness of departments is not taken into account. However, after analyzing the admission rate within each department, there is no significant discrimination between genders (as is shown in (5) below).

(5)

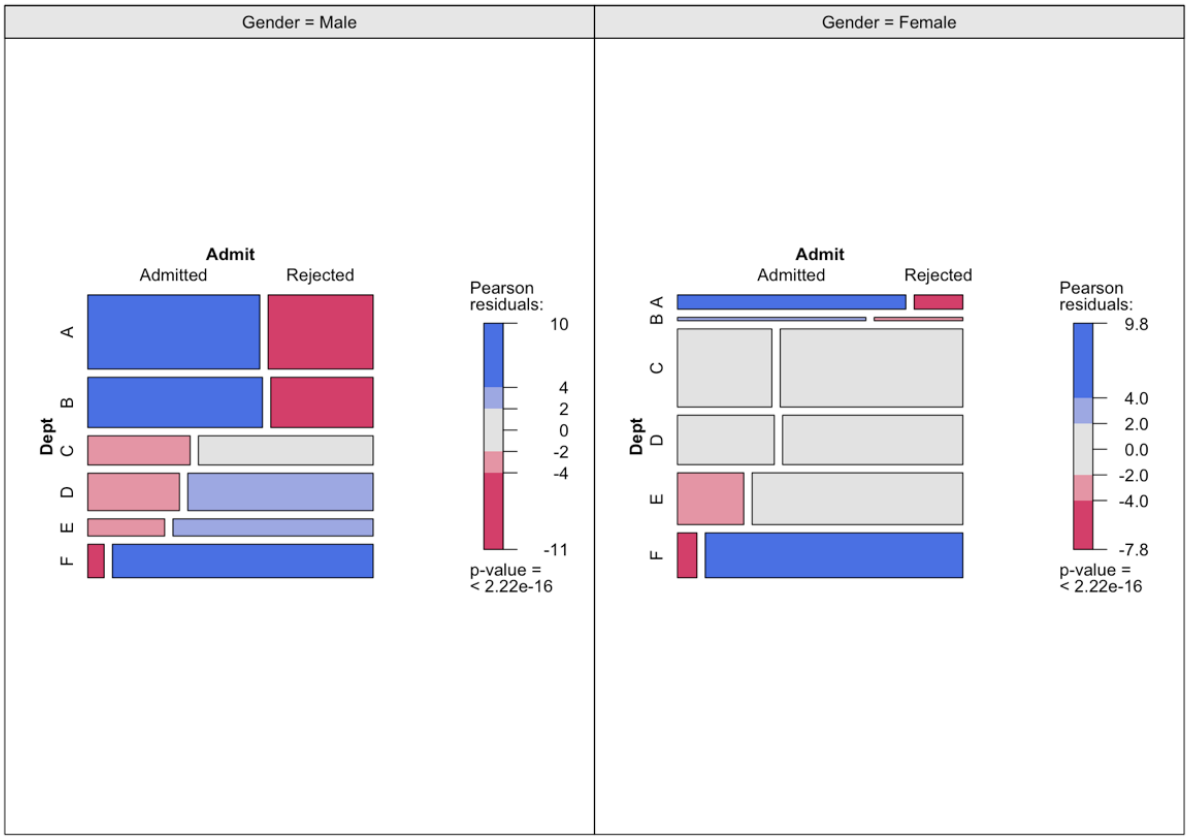
R code

```
flat<-structable(~Dept+Gender+Admit,data=addf)
mat1<-matrix(flat,ncol = 2,nrow = 12)
sumup<-structable(~Dept+Gender,data=addf)
mat2<-cbind(as.vector(sumup),as.vector(sumup))
mat1/mat2
```

```
#              Admitted       Rejected
# A  Male     0.62060606 0.3793939
# B  Male     0.63035714 0.3696429
# C  Male     0.36923077 0.6307692
# D  Male     0.33093525 0.6690647
# E  Male     0.27748691 0.7225131
# F  Male     0.05898123 0.9410188
# A Female  0.82407407 0.1759259
# B Female  0.68000000 0.3200000
# C Female  0.34064081 0.6593592
# D Female  0.34933333 0.6506667
# E Female  0.23918575 0.7608142
# F Female  0.07038123 0.9296188
```

The proportions above show that in the same department, the admission rate is almost equal between genders.

```
cotabplot(~Dept+Admit | Gender,addf,shade=TRUE)
```



The gender bias within the departments is not significant. As Department C and E like to admit more male applicants than female ones. In the other departments, the admission rates for women are higher, however.

(6)

i.

R code

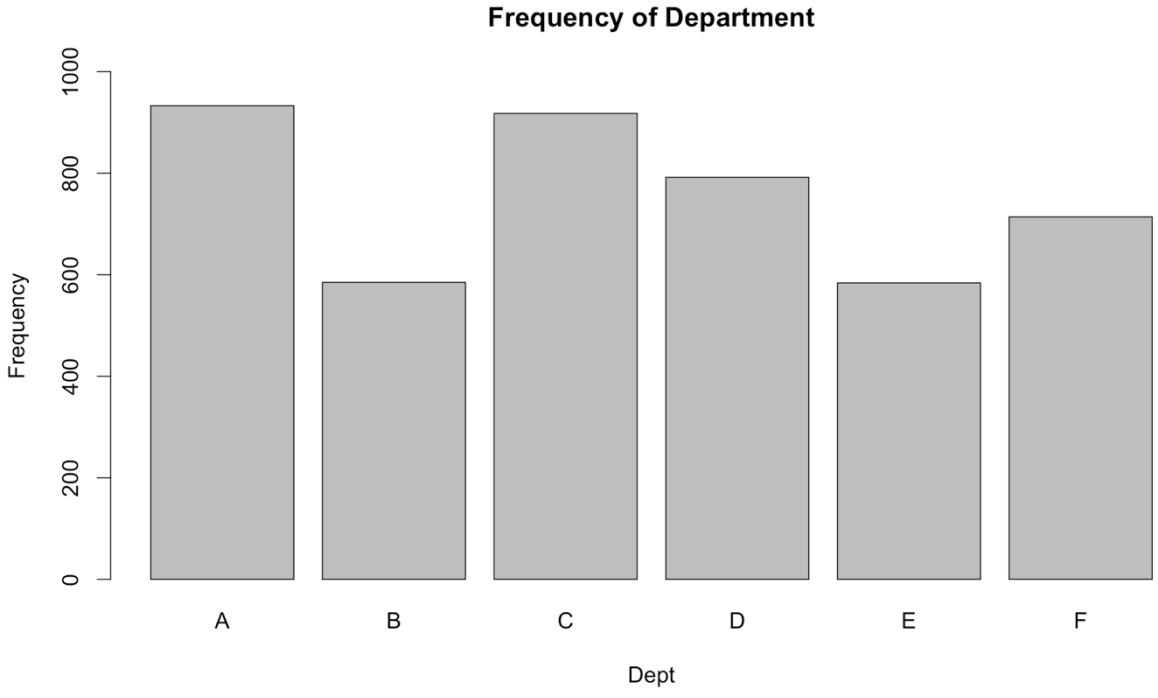
```
structable(~Gender+Admit,data=UCBAdmissions)
#   Admit Admitted Rejected
# Gender
# Male          1198     1493
# Female         557     1278
structable(~Gender,data=addf)
# Gender
# Male    2691
# Female  1835
```

The number of male applicants is about 147% of female ones.
The admission rate for male is 1198/2691=44.5%, for female which is 557/1835=30.4%, and the admission rate of male applicants is about 147% of female ones.

ii.

R code

```
require(graphics)
Depart<-as.data.frame(structable(~Dept,data=UCBAdmissions))
barplot(Depart$Freq,names.arg=Depart$Dept,
        xlab="Dept",ylab="Frequency",ylim=c(0,1000),main="Frequency of Department")
```



The most popular departments are A and C, while the least ones are B and E.

Q5

(1) X_1 and X_2 are independent, because the Σ_{12}, Σ_{21} elements are 0.
(2) X_1 and X_3 are not independent, because the Σ_{13}, Σ_{31} elements are not 0.
(3) X_2 and X_3 are independent, because the Σ_{23}, Σ_{32} elements are 0.
(4) (X_1, X_3) and X_2 are independent, because X_1 and X_2, X_3 and X_2 are independent respectively.
(5) X_1 and $X_1 + 3X_2 - 5X_3$ are not independent, because to linear transformation A= $\begin{bmatrix}1 & 0 & 0\\ 1 & 3 & -5\end{bmatrix}$, $A\Sigma A^T = \begin{bmatrix}1 & 0 & 0\\ 1 & 3 & -5\end{bmatrix}\begin{bmatrix}4 & 0 & -1\\ 0 & 5 & 0\\ -1 & 0 & 2\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 3\\ 0 & -5\end{bmatrix} = \begin{bmatrix}4 & 9\\ 9 & 109\end{bmatrix}$, obviously the Σ_{12}, Σ_{21} elements are not 0.

Q6

Given $X_2,\ldots,X_p$, the density function for X1 is $p(x_1|x_{2,\ldots,}x_p) = \frac{p(x_1,x_2,\ldots,x_p)}{p(x_2,\ldots,x_p)}$.

First we divide the original Σ into a matrix $\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$

And $\Omega = \Sigma^{-1} = \frac{1}{(\Sigma_{11}\Sigma_{22}-\Sigma_{12}\Sigma_{21})}\begin{bmatrix}\Sigma_{22} & -\Sigma_{12}\\ -\Sigma_{21} & \Sigma_{11}\end{bmatrix}$, so $\Omega_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Let $\eta_1 = x_1 + T\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}, \eta_2 = \begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}$, where $T = -\Sigma_{12}\Sigma_{22}^{-1}$

$$E\left[\left(\eta_1 - E(\eta_1)\right)\left(\eta_2 - E(\eta_2)\right)^T\right]$$

$$= E\left[\left(x_1 + T\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix} - E(x_1) - TE\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}\right)\right)\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix} - E\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}\right)\right)^T\right]$$

$$= E\left[T\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix} - E\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}\right)\right)\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix} - E\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix}\right)\right)^T\right]$$

$$+ E\left[\left(x_1 - E(x_1)\right)\left(\begin{bmatrix}x_2\\ \cdots\\ x_p\end{bmatrix} - E\left(\begin{bmatrix}x_2\\ \cdots\\ \square_p\end{bmatrix}\right)\right)^T\right] = \mathrm{T}\Sigma_{22} + \Sigma_{21} = -\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} + \Sigma_{12}$$

$$= 0$$

So η_1, η_2 are independent. And because

$\eta = \begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \begin{bmatrix}1 & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ \cdots\\ x_p\end{bmatrix} = DX$ is a linear transformation of X

Because $p_\eta(y_1, y_2) = p(x_1, x_{2,\ldots,}x_p)|D|$, and $|D| = 1$, and η_1, η_2 are independent,
$p(x_1, x_{2,\ldots,}x_p) = p_\eta(y_1, y_2) = p_{\eta_1}(y_1)p_{\eta_2}(y_2)$

So the conditional density function is $p(x_1|x_{2,\ldots,}x_p) = \frac{p(x_1,x_2,\ldots,x_p)}{p(x_2,\ldots,x_p)} = \frac{p_{\eta_1}(y_1)p_{\eta_2}(y_2)}{p_{\eta_2}(y_2)} = p_{\eta_1}(y_1)$

Because $\eta = DX$ has a distribution N(0,$D\Sigma D^T$), Var(y_1)=$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Hence, $Var(x_1|x_{2,\ldots,}x_p) = \Omega_{11}^{-1}$