# CSC 424 Homework 3

3.

Source Code:

```
library(CCA)
library(psych)
library(CCP)
ds <- read.csv("/Users/Yiyang/Documents/CSC 
424/Datasets/data_marsh_cleaned_homework2.csv", sep = ",", header = TRUE)

#x: water, y: soil
X <- ds[-1, 2: 6]
x <- apply(X, 2, as.numeric)

Y <- ds[-1, 7: 9]
y <- apply(Y, 2, as.numeric)
```

1)

a.

```
#a. combine all
c <- matcor(x, y)
cross <- c$XYcor[1:5, 6:8]
round(cross, 2)

#Corr test
p <- corr.p(cross, nrow(x))
p

#Canonlical correlation
cc1 <- cc(x, y)

wilks1 <- ccaWilks(X, Y, cc1)
wilks1
round(wilks1, 2)
```

Output:

```
> wilks1 <- ccaWilks(X, Y, cc1)
> wilks1
         WilksL        F df1      df2            p
[1,] 0.6963021 4.051995  15 433.8093 6.185853e-07
[2,] 0.8179043 4.176302   8 316.0000 9.094796e-05
[3,] 0.9284064 4.087068   3 159.0000 7.921523e-03
> round(wilks1, 2)
     WilksL    F df1    df2    p
[1,]   0.70 4.05  15 433.81 0.00
[2,]   0.82 4.18   8 316.00 0.00
[3,]   0.93 4.09   3 159.00 0.01
```

b).
Source Code:
Y2 <- Y[, 2: 3]
y2 <- y[, 2: 3]
y2
cc2 <- cc(x, y2)
cc2

wilks2 <- ccaWilks(X, Y2, cc2)
wilks2
round(wilks2, 2)

Ouput:

```
> wilks2 <- ccaWilks(X, Y2, cc2)
> wilks2
        WilksL        F df1 df2            p
[1,] 0.7800982 4.177702  10 316 1.892668e-05
[2,] 0.9019218 4.322556   4 159 2.400337e-03
> round(wilks2, 2)
     WilksL    F df1 df2 p
[1,]   0.78 4.18  10 316 0
[2,]   0.90 4.32   4 159 0
```

c).
Source Code:
#c
Y3 <- data.frame(TPRSDFB = Y$TPRSDFB)
Y3
y3 <- as.matrix(y[, 3])
y3
cc3 <-cc(x, y3)
cc3

wilks3 <- ccaWilks(X, Y3, cc3)
wilks3
round(wilks3, 2)

Output:

```
> wilks3 <- ccaWilks(X, Y3, cc3)
> wilks3
       WilksL        F df1 df2          p
[1,] 0.888069 4.008028   5 159 0.00188145
> round(wilks3, 2)
     WilksL    F df1 df2 p
[1,]   0.89 4.01   5 159 0
```

d).
For the first one the correlations are 0.3855843, 0.3449978, 0.2675698.

For the second one the correlations are 0.3675203, 0.3131744.
For the third one the correlation is 0.3345609.

e).
Since all the three correlations have small p value, it means we need to reject null hypothesis, which means the correlations are not equal to zero.

2.
Output:

```
$xcoef
                 [,1]         [,2]          [,3]
MEHGSWB    0.720571333 -0.613310304   0.442819677
TURB       0.014902006  0.003947628   0.046585662
DOCSWD    -0.122898091 -0.045649299  -0.038307498
SRPRSWFB -15.972715690 77.864165952 -98.959103678
THGFSFC    0.004124619 -0.009849176  -0.009493841

$ycoef
                [,1]         [,2]         [,3]
THGSDFC  0.011415578 -0.010169482 -0.014106076
TCSDFB  -0.077556675 -0.037720634  0.072787341
TPRSDFB -0.002969355  0.002268621 -0.004222605
```

a).
Soil variates and water variables
V1 = 0.720571333MEHGSWB + 0.014902006TURB – 0.122898091DOCSWD – 15.972715690SRPRSWFB + 0.004124619THGFSFC
V2 = -0.613310304MEHGSWB + 0.003947628TURB – 0.045649299DOCSWD + 77.864165952SRPRSWFB – 0.009849176THGFSFC
V3 = 0.442819677MEHGSWB + 0.46585662TURB – 0.038307498DOCSWD – 98.959103678SRPRSWFB – 0.009493841THGFSFC

V1 = 0.011415578THGSDFC – 0.077556675TCSDFB – 0.002969355TPRSDFB
V2 = -0.010169482THGSDFC – 0.037720634TCSDFB + 0.0022686621TPRSDFB
V3 = -0.014106076THGSDFC + 0.072787341TCSDFB – 0.004222605TRPSDFB

b).

```
$scores$corr.X.xscores
               [,1]        [,2]        [,3]
MEHGSWB  -0.2138288 -0.54424426  0.05580913
TURB     -0.1207027 -0.03435814  0.49853147
DOCSWD   -0.8920181 -0.39006177  0.02464817
SRPRSWFB -0.1719363  0.58138401 -0.63983875
THGFSFC   0.4914315 -0.62009828 -0.52589688

$scores$corr.Y.xscores
                [,1]        [,2]        [,3]
THGSDFC -0.003665011 -0.30485575 -0.12523874
TCSDFB  -0.246423901 -0.26504660  0.00980968
TPRSDFB -0.275332457  0.05094524 -0.18310544
```

```
$scores$corr.X.yscores
                [,1]        [,2]         [,3]
MEHGSWB  -0.08244902 -0.18776307  0.014932836
TURB     -0.04654108 -0.01185348  0.133391950
DOCSWD   -0.34394820 -0.13457045  0.006595106
SRPRSWFB -0.06629592  0.20057620 -0.171201505
THGFSFC   0.18948827 -0.21393254 -0.140714106

$scores$corr.Y.yscores
                [,1]       [,2]        [,3]
THGSDFC -0.009505083 -0.8836455 -0.46806012
TCSDFB  -0.639092107 -0.7682559  0.03666214
TPRSDFB -0.714065477  0.1476683 -0.68432782
```

Soil variates and soil variables
V1 = -0.009505083THGSDFC – 0.0639092107TCSDFB – 0.714065477TPRSDFB
V2 = -0.8836455THGSDFC – 0.7682559TCSDFB + 0.1476683TPRSDFB
V3 = -0.46806012THGSDFC + 0.03666214TCSDFB - 0.68432782TPRSDFB

Water variates and water variables

V1 = -0.2138288MEHGSWB – 0.1207027TURB – 0.8920181DOCSWD – 0.1719363SRPRSWFB + 0.4914315THGFSFC
V2 = -0.54424426MEHGSWB – 0.03435814TURB – 0.39006177DOCSWD – 0.58138401SPRSWFB – 0.62009828THGFSFC
V3 = 0.05580913MEHGSWB + 0.49853147TURB + 0.02464817DOCSWD – 0.63983875SRPRSWFB – 0.52589688THGFSFC

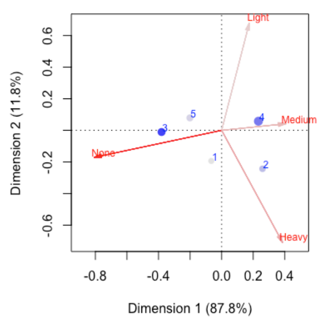
c).
The correlations between Soil variates and water variables are similar with the one between soil variates and water variables.

4.
Source Code:
smoking <- read.csv("/Users/Yiyang/Documents/CSC 424/Smoking.csv", sep = ',', header = TRUE)
colnames(smoking) <- c("Staff Group", "None", "Light", "Medium", "Heavy", "Row Total")
cSmoking = ca(smoking[1:5, 2:5])
plot(cSmoking, mass = T, contrib = "absolute", map = "rowgreen", arrows = c(F, T))
a).
Output:

```
 Principal inertias (eigenvalues):
           1        2        3
Value      0.074759 0.010017 0.000414
Percentage 87.76%   11.76%   0.49%


 Rows:
                1         2         3        4         5
Mass     0.056995  0.093264  0.264249 0.455959  0.129534
ChiDist  0.216559  0.356921  0.380779 0.240025  0.216169
Inertia  0.002673  0.011881  0.038314 0.026269  0.006053
Dim. 1  -0.240539  0.947105 -1.391973 0.851989 -0.735456
Dim. 2  -1.935708 -2.430958 -0.106508 0.576944  0.788435


 Columns:
             None    Light   Medium     Heavy
Mass     0.316062 0.233161 0.321244  0.129534
ChiDist  0.394490 0.173996 0.198127  0.355109
Inertia  0.049186 0.007059 0.012610  0.016335
Dim. 1  -1.438471 0.363746 0.718017  1.074445
Dim. 2  -0.304659 1.409433 0.073528 -1.975960
```




b).
The patterns on the plot are present the staff group. The pattern 4 is the biggest one and pattern 3 is the darkest one. Pattern 4 means this group has the most people, pattern 3 mean this group has the most possibility on none smoking. And the perpendicular distance from patterns to the vector means the smoking tendency of the group.

c).
First two eigenvectors account for 99.51% of the inertia. From the result, only one eigenvector get to 80 of the inertia. It is easy to plot the data with on dimensions.