Keiland Pullen DSC 424 | Fall 2021
Homework 5

1.)

a.)

```
setwd("C:/Users/Home/Desktop/DePaul/DSC-424-AdvancedDataAnalysis/Week-9/Homework")
library(MASS)
cereal = read.table("kellog.dat", skip = 2)
View(cereal)
```

b.)

```
#compute distance matrix
cereal.dist = dist(cereal)
cereal.dist
```

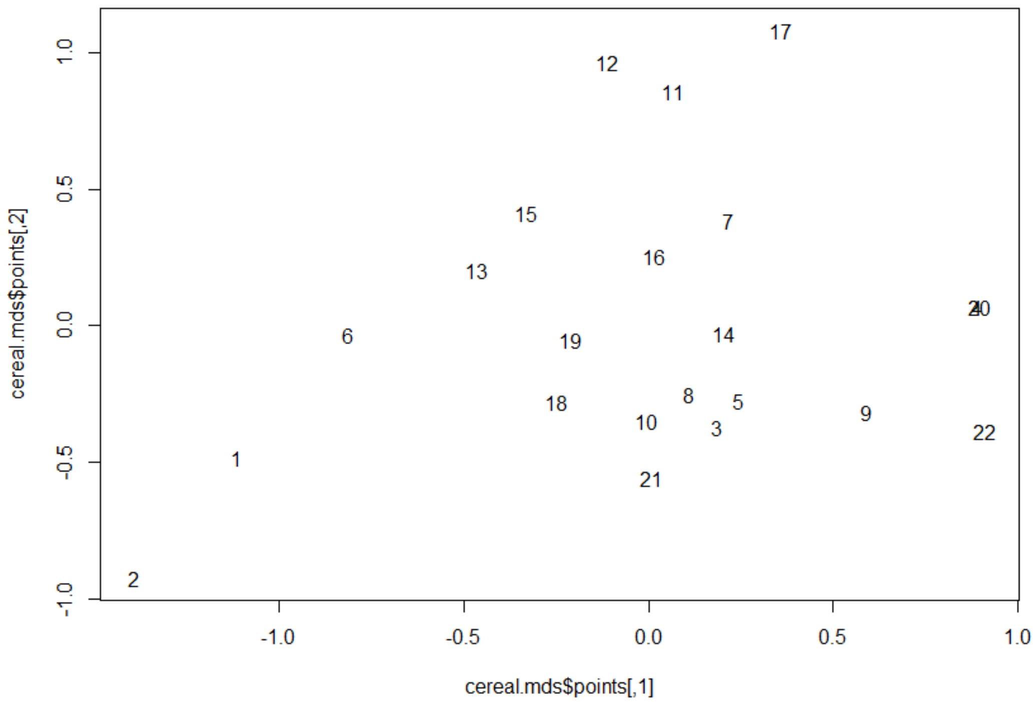
```
> cereal.dist
          1         2         3         4         5         6         7         8         9        10        11        12
2  0.7627275
3  1.6463561 1.9394628
4  1.9444096 2.1296111 1.3303107
5  1.7475826 1.9748142 0.3120582 1.2498644
6  1.1532788 1.6194217 1.3819230 1.8541782 1.4208764
7  1.6614254 1.8495372 1.2050965 1.0802540 1.1002253 1.4582195
8  1.5828006 1.9369682 0.3564785 1.3340118 0.4526876 1.1113783 1.2132153
9  1.8903267 2.1547375 0.6839052 0.8813549 0.6442168 1.6823317 1.3041942 0.7519969
10 1.5047556 1.5674432 0.7635235 1.2865107 0.7035502 1.3113703 1.1154849 0.8010485 1.0270650
11 1.7706922 2.0224332 1.4206019 1.6277485 1.3720403 1.4325133 1.1753745 1.3500563 1.5987727 1.3966045
12 1.8535042 2.1336687 1.4889333 1.7197877 1.4755779 1.5032312 1.2687261 1.4320505 1.6872791 1.4678737 0.4831629
13 1.5262955 1.9662591 1.1994278 1.7441181 1.2392626 0.8784647 1.2632319 1.0329963 1.5210744 1.2626172 1.3690600 1.1996439
14 1.5729053 1.9348745 0.6195444 0.9810702 0.5970248 1.1136701 0.8801664 0.4665635 0.6956775 0.8406089 1.2091919 1.2802362
15 1.4906280 1.8816847 1.3608276 1.4419108 1.3413498 0.9381718 0.9109614 1.1923498 1.5222873 1.2958157 1.2395704 1.1371733
16 1.3162681 1.4303223 1.1765664 1.1869528 1.1161669 1.2817517 0.4688042 1.1807495 1.3664081 0.8922996 1.1910016 1.3094208
17 1.8594132 2.0846170 1.6812740 1.5048492 1.6646404 1.8238031 1.1260445 1.6871029 1.7416150 1.6692397 0.7124123 0.8482888
18 1.1024588 1.5384141 0.9544808 1.3585306 1.0538959 1.0511956 1.2966841 0.8798771 1.0987421 1.0006810 1.4580800 1.3927434
19 1.4690907 1.5197314 0.9742168 1.5286724 0.8580850 1.2526313 0.9247498 0.9990176 1.3511064 0.5870351 1.2739387 1.4045927
20 2.0142501 2.2233598 1.3253066 0.1565389 1.2445369 1.8805868 1.0786322 1.3326933 0.8737835 1.3121001 1.6296738 1.6917097
21 1.6581126 1.9922904 0.4245868 1.5674182 0.5452639 1.1805261 1.4310786 0.2696816 0.9087276 0.8525140 1.4818790 1.5543900
22 1.7697264 1.9650416 1.3484555 0.9399196 1.4281344 1.7646503 1.3908348 1.3557161 1.2051478 1.1797844 1.7646667 1.7654580
          13        14        15        16        17        18        19        20        21
2
3
4
5
6
7
8
9
10
11
12
13
14 0.9576267
15 0.5945245 0.9088726
16 1.2614938 0.9434419 0.9631612
17 1.7127350 1.4364040 1.4248343 1.1919981
18 0.8520390 0.8338824 0.9784496 1.1590402 1.6615256
19 1.2302789 0.9947183 1.2129570 0.6978062 1.5726904 1.1923494
20 1.7062215 0.9641492 1.4157408 1.2261135 1.5129642 1.3598561 1.5525249
21 1.1211845 0.7220695 1.3772198 1.3632255 1.8755148 0.9887393 1.0582829 1.5663008
22 1.7097949 1.1474113 1.5346291 1.2774942 1.6647760 1.3052600 1.5804658 0.9637717 1.5239306
```

c.)

```
#run multidimensional scaling
cereal.mds = isoMDS(cereal.dist)

#get stress value
cereal.mds$stress

#Plot MDS
plot(cereal.mds$points, type="n")
text(cereal.mds$points, labels = as.character(1:nrow(cereal)))
```

```
> cereal.mds = isoMDS(cereal.dist)
initial  value 19.915627
iter   5 value 14.620451
iter  10 value 14.224381
iter  10 value 14.220757
iter  10 value 14.218694
final  value 14.218694
converged

> #get stress value
> cereal.mds$stress
[1] 14.21869
```

Or .14%



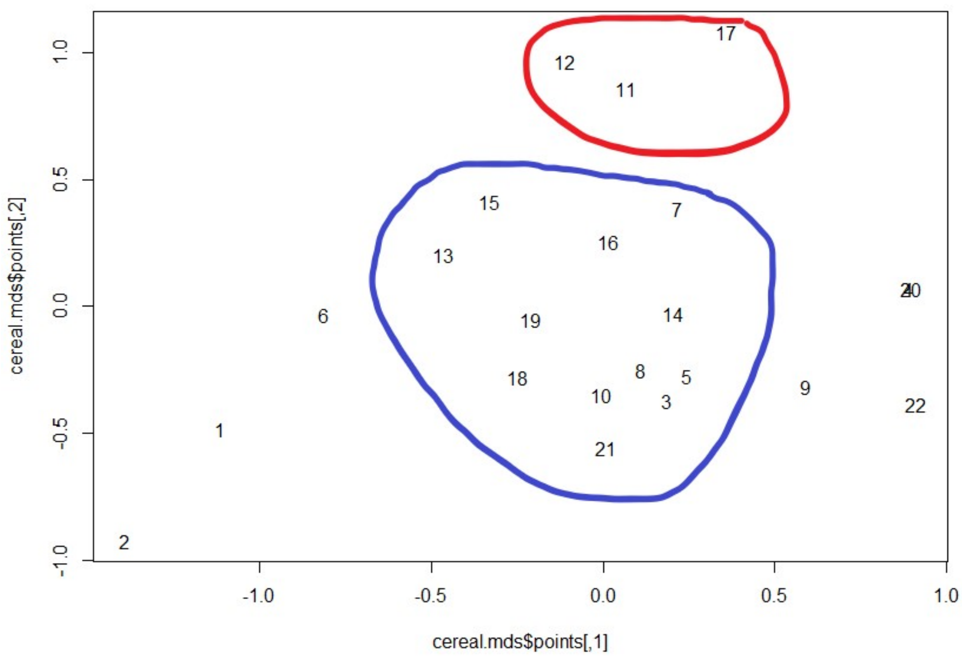
With a stress value of .14, which is considered a good measure, the plot does a decent job at reproducing the distances.

d.)

There appears to be at least 2 distinct clusters.

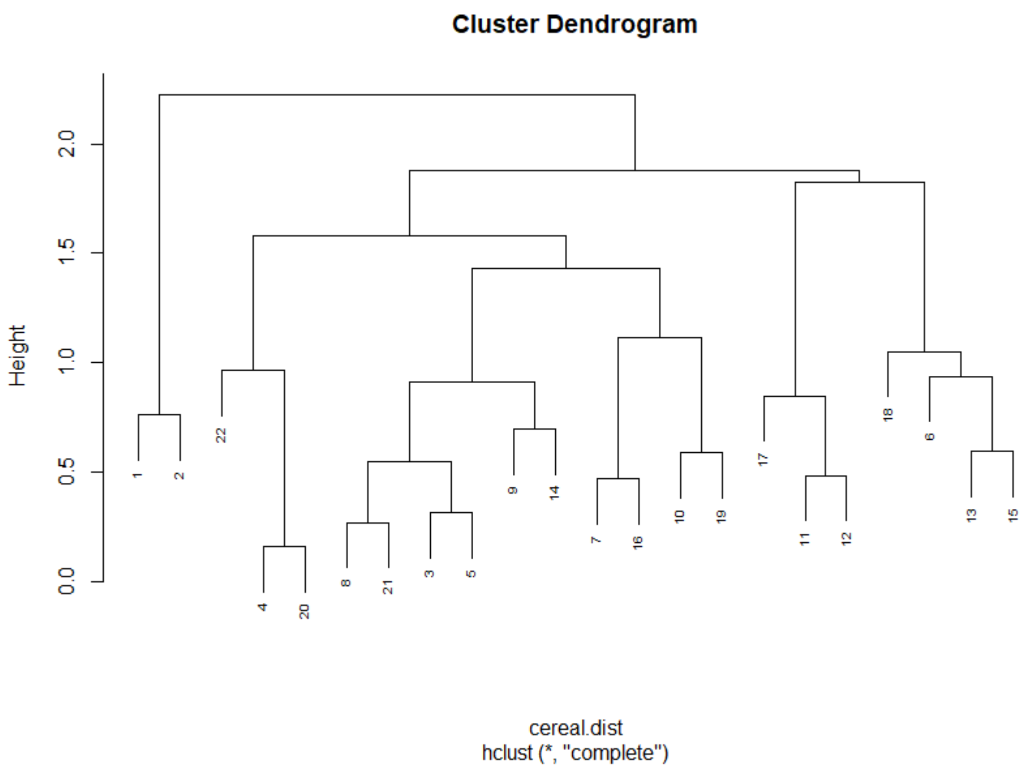
In the red cluster below, it appears that the cereals are “JustRight”, “JustRightFruitNut” and “Product19”.

In the blue cluster below, the cereals are “AppleJacks”, “CornPops”, “Crispix”, “FrootLoops”, “FrostedMiniWheats”, “Mueslix”, “Nut&Honey”, “NutriGrain”, “NutFeast”, “RaisinBran”, “RaisinWheats” and “HoneySmacks”.



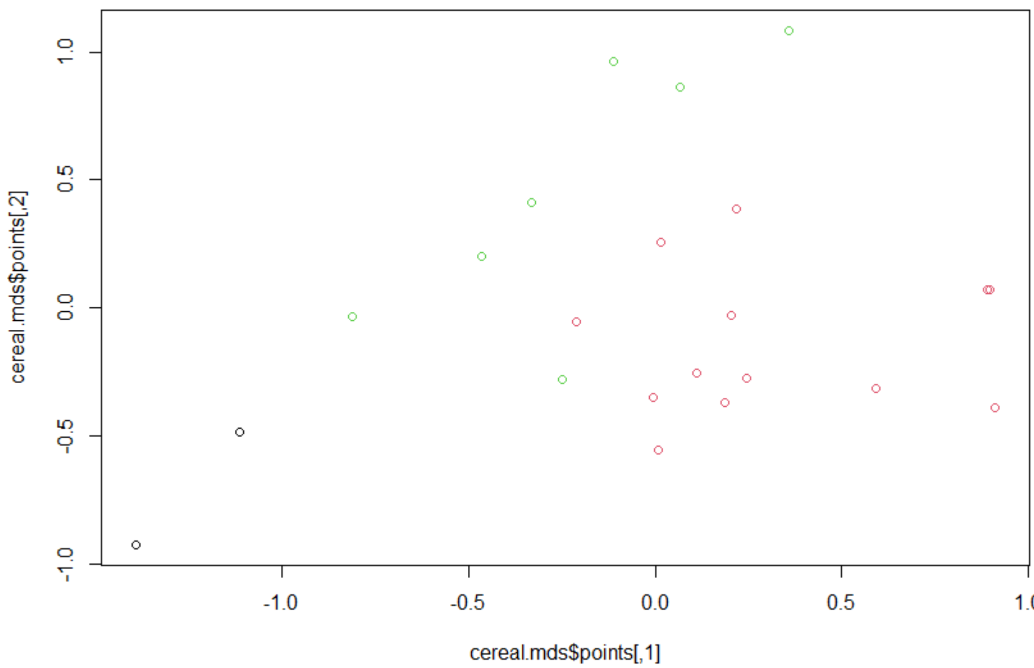
e.)

```
#Agglomerative Hierarchial clustering
clusterCereal = hclust(cereal.dist)

#Dendrogram
plot(clusterCereal, cex = 0.6, hand = -1)
```



f.)

```
#cutree method
clusterCut = cutree(clusterCereal, k=3)
head(clusterCut)

#plot mds using the colors from cutree
plot(cereal.mds$points, col=clusterCut)
```



The result here is 3 clusters. The black/blue points make up a small cluster which is composed of "AllBran" and "AllBranFlakes". This cluster appeared to be a set of outliers in the earlier plot. There are 2 points that lie on/near (-.2, 0) that are "RaisinBran" (green) and "RaisinWheats" (red). Both are in different clusters, but originally, they were thought to have been in the same cluster. Points on the far right were included in the red cluster, as they were originally thought to be outliers. Similarly, 3 points at the top are in the green cluster and they were thought to have been in their own cluster.

g.)

In some cases it appears that cereals containing similar ingredients (e.g. sugars, nuts, oats, etc) are grouped together and then combined into respective clusters.

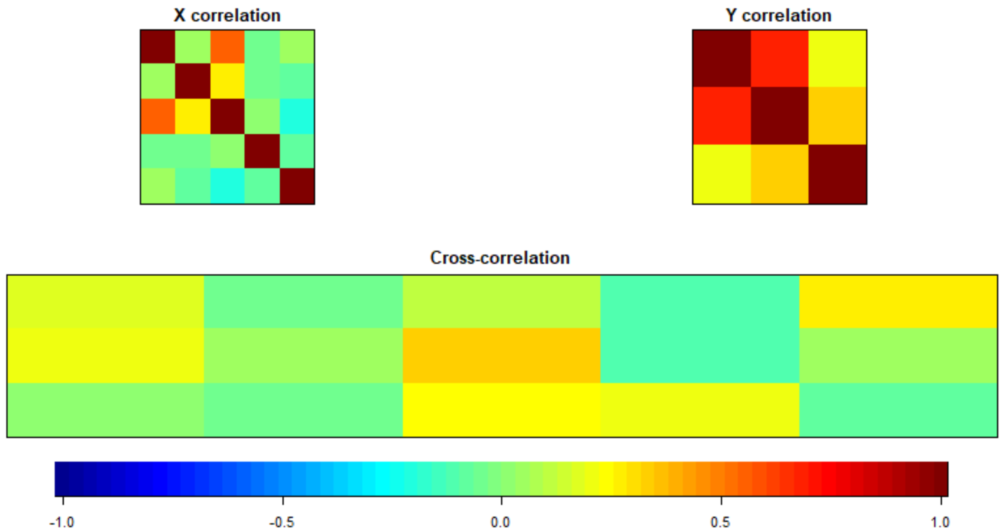
Keiland Pullen DSC 424 | Fall 2021
Homework 5

2.)

```
library(CCA)

ds = read.csv("data_marsh_cleaned.csv")
View(ds)
head(ds)

# Extract X and Y variants
water <- ds[,2:6]
soil <- ds[,7:9]

c = matcor(water, soil)
c
```

```
> c = matcor(water, soil)
> c
$Xcor
            MEHGSWB        TURB      DOCSWD    SRPRSWFB     THGFSFC
MEHGSWB   1.00000000  0.04286195  0.53653344 -0.05729504  0.04523356
TURB      0.04286195  1.00000000  0.26262016 -0.03127880 -0.08426556
DOCSWD    0.53653344  0.26262016  1.00000000  0.01784706 -0.20284406
SRPRSWFB -0.05729504 -0.03127880  0.01784706  1.00000000 -0.08581679
THGFSFC   0.04523356 -0.08426556 -0.20284406 -0.08581679  1.00000000

$Ycor
          THGSDFC    TCSDFB   TPRSDFB
THGSDFC 1.0000000 0.6677804 0.1966074
TCSDFB  0.6677804 1.0000000 0.3178176
TPRSDFB 0.1966074 0.3178176 1.0000000

$XYcor
            MEHGSWB        TURB      DOCSWD    SRPRSWFB     THGFSFC     THGSDFC      TCSDFB
MEHGSWB   1.00000000  0.04286195  0.53653344 -0.05729504  0.04523356  0.15971021  0.19749008
TURB      0.04286195  1.00000000  0.26262016 -0.03127880 -0.08426556 -0.05151880  0.04374098
DOCSWD    0.53653344  0.26262016  1.00000000  0.01784706 -0.20284406  0.11909492  0.32344092
SRPRSWFB -0.05729504 -0.03127880  0.01784706  1.00000000 -0.08581679 -0.09647552 -0.11800127
THGFSFC   0.04523356 -0.08426556 -0.20284406 -0.08581679  1.00000000  0.25310209  0.03809560
THGSDFC   0.15971021 -0.05151880  0.11909492 -0.09647552  0.25310209  1.00000000  0.66778043
TCSDFB    0.19749008  0.04374098  0.32344092 -0.11800127  0.03809560  0.66778043  1.00000000
TPRSDFB   0.02092839 -0.05980083  0.22121653  0.19411633 -0.07060351  0.19660738  0.31781764
            TPRSDFB
MEHGSWB   0.02092839
TURB     -0.05980083
DOCSWD    0.22121653
SRPRSWFB  0.19411633
THGFSFC  -0.07060351
THGSDFC   0.19660738
TCSDFB    0.31781764
TPRSDFB   1.00000000
```

img.matcor(c, type=2)



ccWaterSoil = cc(water, soil)
ccWaterSoil
ccWaterSoil$cor

```
> ccWaterSoil = cc(water, soil)
> ccWaterSoil
$cor
[1] 0.3855843 0.3449978 0.2675698

$names
$names$Xnames
[1] "MEHGSWB"  "TURB"     "DOCSWD"   "SRPRSWFB" "THGFSFC"

$names$Ynames
[1] "THGSDFC" "TCSDFB"  "TPRSDFB"

$names$ind.names
  [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"  "13"  "14"  "15"
 [16] "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"  "25"  "26"  "27"  "28"  "29"  "30"
 [31] "31"  "32"  "33"  "34"  "35"  "36"  "37"  "38"  "39"  "40"  "41"  "42"  "43"  "44"  "45"
 [46] "46"  "47"  "48"  "49"  "50"  "51"  "52"  "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"
 [61] "61"  "62"  "63"  "64"  "65"  "66"  "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"
 [76] "76"  "77"  "78"  "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"  "89"  "90"
 [91] "91"  "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"  "100" "101" "102" "103" "104" "105"
[106] "106" "107" "108" "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
[121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132" "133" "134" "135"
[136] "136" "137" "138" "139" "140" "141" "142" "143" "144" "145" "146" "147" "148" "149" "150"
[151] "151" "152" "153" "154" "155" "156" "157" "158" "159" "160" "161" "162" "163" "164" "165"


$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
```

```
$scores
$scores$xscores

              [,1]        [,2]         [,3]
  [1,]  0.378474893 -1.34275741  0.091304960
  [2,] -2.955037356 -1.83427792 -0.006391163
  [3,] -0.304931015  0.61468256  0.994801518
  [4,]  0.092196483 -0.64505832  2.947765556
  [5,] -1.005565768  0.28762140 -0.767578314
.
.
.
 [160,] -0.056843095  5.26414818 -5.563288985
[161,]  1.248325001 -1.81663808  1.031730095
[162,] -0.128575729 -0.10925157 -0.290797443
[163,] -0.086338654  0.63011904  0.618678416
[164,] -2.851030507  3.98454594 -4.344331810
[165,] -0.100973038 -0.75227011  0.325480541

$scores$yscores
              [,1]        [,2]         [,3]
  [1,] -0.9072638981  0.40113158  0.611479132
  [2,]  0.6627857665  0.72791164  0.293402122
  [3,] -0.4819644023  0.98698512  0.918959599
  [4,] -0.3763663414 -0.25384499  0.150492514
  [5,]  0.9905749426  0.73047289  0.811079939
.
.
.
 [160,]  1.2225619991  0.97574032 -0.549642672
[161,]  0.0005162369 -0.99411820 -1.105455974
[162,]  0.4623219604 -1.73484232  0.283570811
[163,] -0.5108755501 -0.53101654 -0.315700072
[164,] -2.5679362436  2.67906942 -4.349825784
[165,] -0.3482924839 -0.91719932  1.071530616

$scores$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


> ccWaterSoil$cor
[1] 0.3855843 0.3449978 0.2675698
```

Keiland Pullen DSC 424 | Fall 2021
Homework 5

```
> wilksWaterSoil = ccaWilks(water, soil, ccWaterSoil)
> round(wilksWaterSoil)
     WilksL F df1 df2 p
[1,]      1 4  15 434 0
[2,]      1 4   8 316 0
[3,]      1 4   3 159 0
```

1.)

a.)

Statistic = $(1 – 0.3855843^2) + (1 - 0.3449978^2) + (1 – 0.2675698^2) = 2.6607076677$
df = 164
p-value = 0

b.)

Statistic = $(1 - 0.3449978^2) + (1 – 0.2675698^2) = 1.8093829201$
df = 164
p-value = 0

c.)

Statistic = $(1 – 0.2675698^2) = 0.9284064021$
df = 164
p-value = 0

d.)

```
ccWaterSoil = cc(water, soil)
ccWaterSoil$cor
```

```
> ccWaterSoil$cor
[1] 0.3855843 0.3449978 0.2675698
```

The canonical correlation for Water is: 0.3855843
The canonical correlation for Soil is: 0.3449978
The canonical correlation for both is: 0.2675698

It seems that the correlations between the 3 are not significant.

2.

a.)

U (water) = 0.720571333 + 0.014902006 – 0.122898091 – 15.972715690 + 0.004124619
V (soil) = 0.011415578 – 0.077556675 – 0.002969355

b.)

Soil Correlations:

$Ycor

	THGSDFC	TCSDFB	TPRSDFB
THGSDFC	1.0000000	0.6677804	0.1966074
TCSDFB	0.6677804	1.0000000	0.3178176
TPRSDFB	0.1966074	0.3178176	1.0000000

Water Correlations:

$Xcor

	MEHGSWB	TURB	DOCSWD	SRPRSWFB	THGFSFC
MEHGSWB	1.00000000	0.04286195	0.53653344	-0.05729504	0.04523356
TURB	0.04286195	1.00000000	0.26262016	-0.03127880	-0.08426556
DOCSWD	0.53653344	0.26262016	1.00000000	0.01784706	-0.20284406
SRPRSWFB	-0.05729504	-0.03127880	0.01784706	1.00000000	-0.08581679
THGFSFC	0.04523356	-0.08426556	-0.20284406	-0.08581679	1.00000000

It appears that the soil groups are more correlated to each other than the water groups.