Problem 1

[**20 pts**] Download the "kellog.dat" data file which contains data on 22 cereals from Kellog. Each cereal has 9 metric values that measure various aspects of the cereal. We are not given the meanings of these variables, but in spite of this use the data to classify the cereals:

1. Read the data into a data.frame in R. Note that the data file has two extra rows, you can ignore these with the "skip=2" parameter in read.table, or you can manually delete them. Also, you will want to put the cereal names in the row.names with "row.names=1" which indicates to use the first column as the row names.

2. Compute the distance matrix with "dist". Just treat the ordinal and binary categorical variables as metric variables (this is actually ok here because they are either interval variables), or binary variables encoded as [0, 1].

3. Run multidimensional scaling on the distance matrix with the "isoMDS" command from the MASS library. This computes MDS and provides a bit more and as its output, providing both an array of "$points" to plot and a stress value. Plot the points from c) and report the stress value. How faithfully does the plot reproduce the distances in the data according to the stress value (remember the stress value from R is actually multiplied by 100 so it is a percentage)?

4. How many clusters or groups does the data fall into? Can you identify some distinct groupings? Interpret at least two of the groupings of cereals based on their names in the data file.

5. Run an agglomerative hierarchical clustering on the dataset and plot the result as a dendogram.

6. At a level of 3 clusters in the dendogram, use the cutree(h, k=3) command to evaluate the clusters and then replot the MDS using these categories to color the data. Interpret the results.

7. (Extra Credit, 3 points) Give a practical interpretation for at least one of the two dimensions in the MDS.

Solution: (Problem 1 Source Code)

1.

```
# output
> data = read.table("kellog.dat", skip=2, row.names=1)
> head(data)
                     V2   V3     V4     V5     V6     V7     V8   V9
        V10 V11
AllBran         0.1818 0.6 0.3333 0.8125 0.6429 0.0000 0.3333 1.0
    0.9677   0
AllBranFlakes   0.0000 0.6 0.0000 0.4375 1.0000 0.0667 0.0000 1.0
    1.0000   0
AppleJacks      0.5455 0.2 0.0000 0.3906 0.0714 0.2667 0.9333 0.5
    0.0323   0
CornFlakes      0.4545 0.2 0.0000 0.9063 0.0714 0.9333 0.1333 0.0
    0.0484   0
CorPops         0.5455 0.0 0.0000 0.2813 0.0714 0.4000 0.8000 0.5
    0.0000   0
CracklinOatBran 0.5455 0.4 1.0000 0.4375 0.2857 0.2000 0.4667 1.0
    0.4516   0
```

2.

```
> kellog.x = as.matrix(data)
> kellog.dist = dist(kellog.x)
```

3.

```
> kellog.mds = isoMDS(kellog.dist)
initial  value 19.915627
iter   5 value 14.639105
iter  10 value 14.220989
final  value 14.179482
converged
> kellog.mds$stress
[1] 14.17948
```

Stress of 14.17948 indicate approximately 14% variance which is not a good fit.

4. From visualization perspective, there might be 3 patterns: first one contains the most condense part from -0.5 to 0.5, and other two parts are less than -0.5 and greater than 0.5.

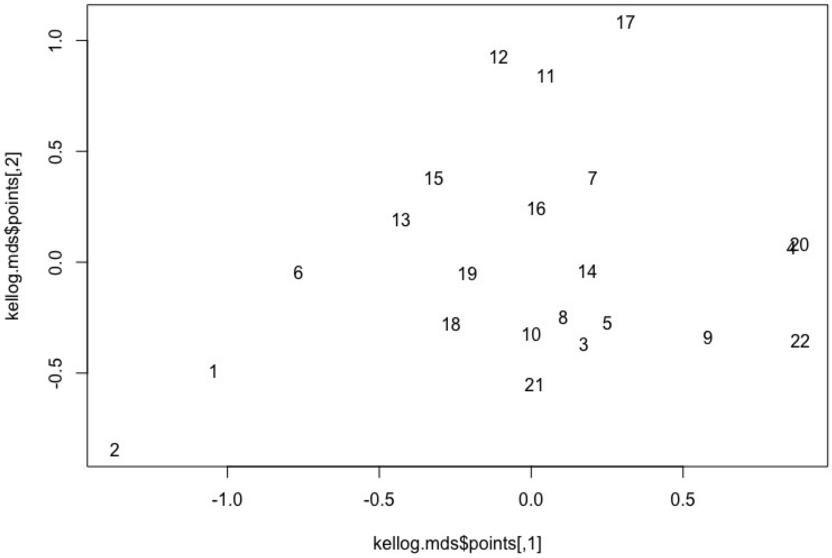

Figure 1: Multidimensional Scaling Plot

5.

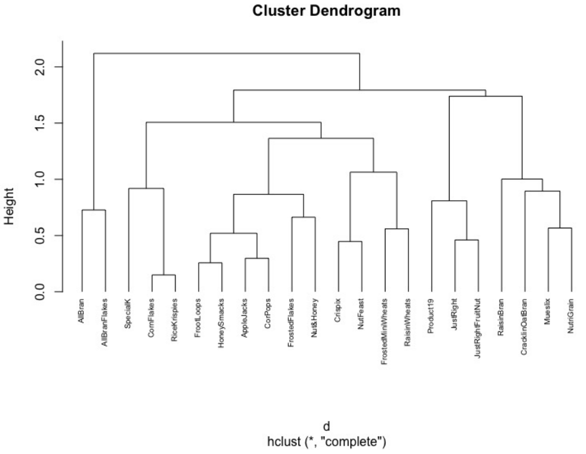

Figure 2: Agglomerative Hierarchical Clustering Dendrogram

6. If we using agglomerative hierarchical clustering, we can see part 2 and 3 are mostly overlap with each other, turns out a bad cluster method. Hence I switch the process from bottom-up to top-down (divisive hierarchical clustering), it gives much better result (Figure 4).

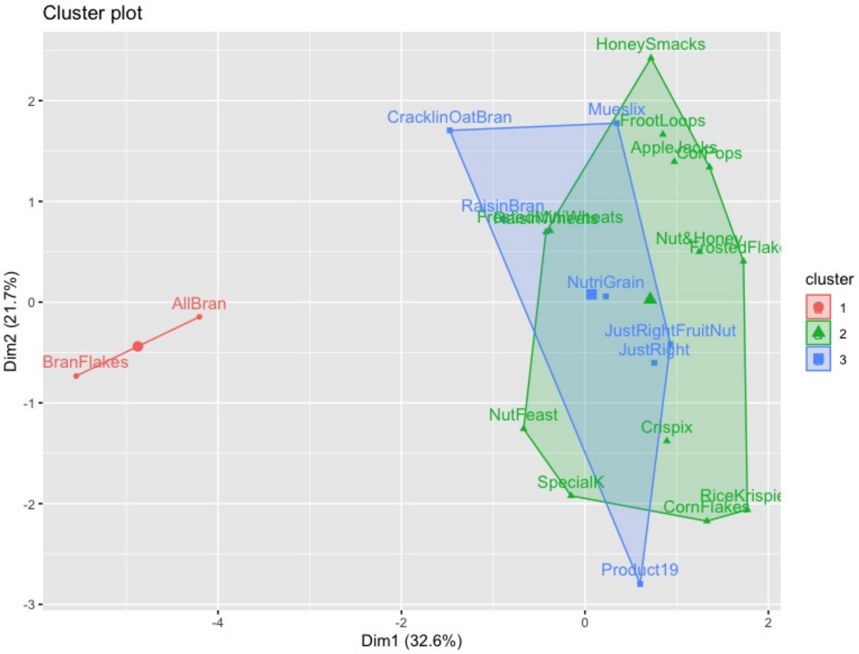

Figure 3: Agglomerative Hierarchical Clustering Plot

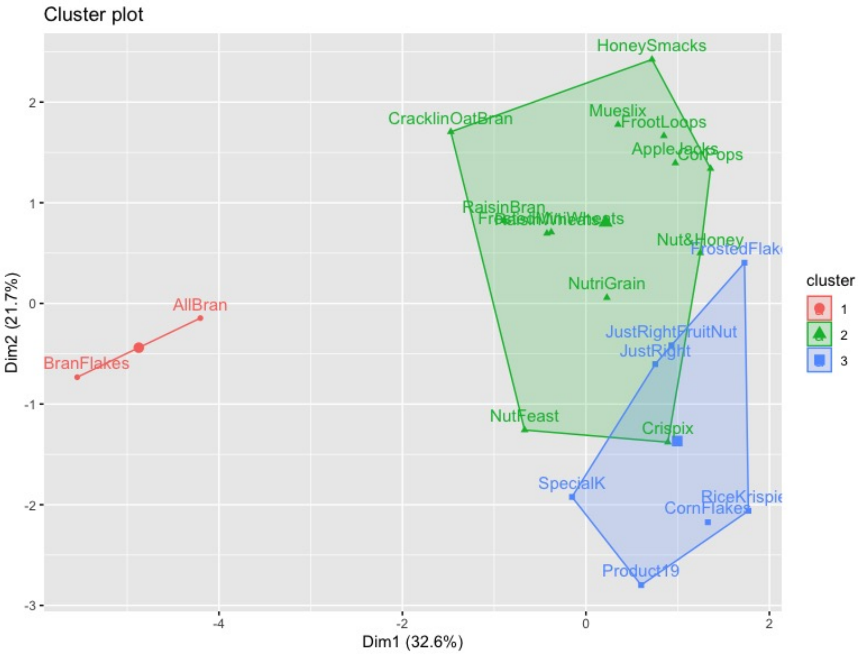

Figure 4: Divisive Hierarchical Clustering Plot

Problem 2

[10 pts] Perform a canonical correlation analysis, describing the relationships between the soil and water variables using the data1 found in data_marsh_cleaned.csv.

1. Answer the following questions regarding the canonical correlations. (Note that a, b and c can all be done directly from the output of canonical correlation)

   (a) Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.

   (b) Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value.

   (c) Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value.

   (d) Present the three canonical correlations and list any conclusions that you can draw.

   Answer the following questions regarding the canonical variates.

   (a) Give the formulae for the first canonical variate for the soil and water variables.

   (b) Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables and use these to interpret the variates (do this as best as you can. Even with a lack of domain knowledge you should be able to describe the relationship in more general terms given the variables involved and the correlations.)

Solution:

1.

```
> round(wilksSoil, 2)
# output
      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
```

The first variance and all other variances followed has information, so we want to include the first variance. Second one has p-value of 0 which can reject the null hypothesis, so it is sufficient to capture the correlation, same as third one, it has p-value of 0.01 which is significantly different from zero as well. Thus we want include all three variances.

2.

```
> round(-loadingsSoil$corr.X.xscores, 2)
          [,1]  [,2]  [,3]
MEHGSWB   0.21  0.54 -0.06
TURB      0.12  0.03 -0.50
DOCSWD    0.89  0.39 -0.02
SRPRSWFB  0.17 -0.58  0.64
THGFSFC  -0.49  0.62  0.53
> round(-loadingsSoil$corr.Y.yscores, 2)
        [,1]  [,2]  [,3]
THGSDFC 0.01  0.88  0.47
TCSDFB  0.64  0.77 -0.04
TPRSDFB 0.71 -0.15  0.68
```

In terms of evaluate the overall level of correlation, we can see that it is improved by overall improvement in all of the attributes except THGFSFC. If we are concentrating on overall performance, we want to focus on DOCSWD since it has much higher correlation than others. For second column, it shows that MEHGSWB and THGFSFC have the most impact on THGSDFC and TCSDFB.