CSC 424 Homework 4

2.

There are three analyses in our project, PCA, CFA and Cluster Analysis. What I do is provide ideas for group and coding for each analyst. Since our project is pretty good for the PCA and CFA, then I think these two analyses may not be evolved, however, our project has two categorical variables, then we decide to apply Cluster Analysis to numeric variables. Then I envision that there might be some interesting conclusions from Cluster Analysis. Getting some information from each cluster we get. These analyses provide me some new way to deal with large-scale data, according to the three methods we use, in my view, they have some similar points, they could find the similarity and divide the data into some small groups depends on the similarity they have. And then simplify the process of dealing with big data.

3.

Besides applying PCA and CFA in our dataset, cluster analysis is another method we apply, although there are two categorical variables in our data, which are date, and event. Then I think convert some of the numeric variables to categorical variables, such as wind speed, and wind direction. The data is separated into three clusters, according to wind direction and speed, I think some information could be gathered from those groups, similarity and dissimilarity could tell us why data could be separated like this.

4.

a).

Source Code:

```
library(MASS)
library(gdata)

dt <- read.xls("/Users/Yiyang/Documents/CSC 424/BondRating.xls", sheet = "training", header = 
TRUE)
head(dt)
names(dt) <- lapply(dt[1, ], as.character)
dt <- dt[-1,]
dt1 <- apply(dt[4: 13], 2, as.numeric)

df <- data.frame(dt[1: 3], dt1)

brLda <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + 
LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = df)
brLda

p <- predict(brLda, newdata = df[,4:13])$class
p

table(p, df$CODERTG)
```

Output:

```
Call:
lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER +
    LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = df)

Prior probabilities of groups:
        1         2         3         4         5         6         7 
0.1111111 0.1604938 0.1481481 0.1604938 0.1604938 0.1358025 0.1234568 

Group means:
     LOPMAR  LFIXCHAR    LGEARRAT    LTDCAP      LLEVER   LCASHLTD     LACIDRAT   LCURRAT LRECTURN  LASSLTD
1 -1.738889 1.6637778 -0.99555556 0.2881111  0.12388889 -0.3940000  0.059888889 0.6932222 1.943889 1.804000
2 -2.094385 1.8042308 -1.05315385 0.2641538 -0.08338462 -0.3925385 -0.003692308 0.6640769 2.266308 1.733462
3 -2.017917 1.7306667 -0.94075000 0.3034167  0.04291667 -0.4003333  0.017500000 0.6387500 2.074250 1.693417
4 -2.213923 1.3204615 -1.01200000 0.2704615 -0.02153846 -0.5720769 -0.063230769 0.7600769 2.032077 1.721769
5 -1.981846 1.7073077 -0.75800000 0.3272308  0.07430769 -0.7765385  0.137076923 0.7471538 1.950000 1.510077
6 -2.078545 0.9529091 -0.07790909 0.4812727  0.44972727 -1.4103636 -0.033181818 0.7031818 1.818182 1.103182
7 -1.783600 0.5873000  0.10860000 0.5248000  0.64370000 -1.4720000 -0.031600000 0.4642000 1.650000 0.993700

Coefficients of linear discriminants:
                LD1        LD2        LD3          LD4           LD5        LD6
LOPMAR   -0.7720156  -2.993776 -1.0902999   1.19056396   0.003079991 -1.0907388
LFIXCHAR  0.3309649  -1.032219  2.0342609  -0.17225468  -0.566130362  0.4446614
LGEARRAT  2.0228900 -13.206606  4.3603205  30.56370258  19.296973115 -8.6572293
LTDCAP   27.6725970  15.434851  1.0663233 -30.15183168   0.636947862 22.5703473
LLEVER   -5.2113899   4.540020 -5.2197916 -13.97013291 -12.485287860  4.5123115
LCASHLTD -0.8040312   3.684976 -0.6103313  -1.47884309   2.343115368  2.1285439
LACIDRAT -0.2978150  -3.360777 -0.7014467  -0.09884748   0.507853522 -0.9383520
LCURRAT  -2.0007312   2.040593 -1.1419790   1.51718949  -2.677213623  3.2930473
LRECTURN -1.1369903  -2.245231 -0.6432160   0.81809242   0.686713979 -0.9182123
LASSLTD   5.2328461 -14.461158  1.3481935  26.33072526  16.502239043 -5.7011832

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6 
0.6309 0.1209 0.1005 0.0705 0.0587 0.0186 
```

```
p    1  2  3  4  5  6  7
  1  4  1  0  0  1  1  0
  2  3  7  3  1  1  0  0
  3  0  1  6  0  1  0  2
  4  1  2  2 11  2  0  1
  5  0  2  1  1  8  1  0
  6  1  0  0  0  0  8  1
  7  0  0  0  0  0  1  6
```

After applying LDA, I find that there are plenty of companies in each level they should not be, however, the level 4 which means BAA level is the best one, there are 11 companies exactly BAA level, there is only 1 company actually should be AA and BA level in BAA level.

b).

Source Code:

library(MASS)
library(gdata)

dt <- read.xls("/Users/Yiyang/Documents/CSC 424/BondRating.xls", sheet = "validation", header = TRUE)
head(dt)
names(dt) <- lapply(dt[1,], as.character)
dt <- dt[-1,]
dt1 <- apply(dt[4: 13], 2, as.numeric)

df <- data.frame(dt[1: 3], dt1)

brLda <- lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER + LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = df)
brLda

p <- predict(brLda, newdata = df[,4:13])
p

table(p, df$CODERTG)

Output:

```
Call:
lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER +
    LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = df)

Prior probabilities of groups:
        1         2         3         4         5         6         7
0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571

Group means:
   LOPMAR LFIXCHAR LGEARRAT LTDCAP  LLEVER LCASHLTD LACIDRAT LCURRAT LRECTURN LASSLTD
1 -1.7115   1.2570  -1.2950 0.2200  0.1045   0.0775   -0.038  0.6050   1.5555  2.1095
2 -1.7595   1.3895  -0.8285 0.3095  0.0525  -0.5820   -0.151  0.4915   1.9725  1.5840
3 -1.7390   2.2890  -1.8435 0.2045 -0.4615   0.3220    0.096  0.8895   1.9620  2.3625
4 -2.0750   0.8125  -1.0790 0.2530 -0.0750  -0.4920   -0.467  0.6315   2.2220  1.7525
5 -2.1440   1.5530  -1.0440 0.2605 -0.1340  -0.5590    0.008  0.6475   2.1930  1.6740
6 -2.3700   0.9170  -0.0330 0.4900  0.3950  -1.5985   -0.244  0.7755   1.9855  1.0140
7 -1.8125   0.2815  -0.0375 0.4890  0.3355  -1.3485   -0.151  0.1900   2.1310  0.9390

Coefficients of linear discriminants:
                LD1        LD2        LD3        LD4        LD5        LD6
LOPMAR   -2.69120927  1.5293389 -1.2026581 -0.1541835 -0.8582843 -1.6587618
LFIXCHAR  0.01650485 -1.4655423  0.4252668  1.0702619  1.5550367 -0.5075890
LGEARRAT  0.42984985 -1.1265425  0.4333757  0.5991812  0.7036927 -0.1831204
LTDCAP   -9.30916052  3.6096960 -1.2511995  1.8919072 -5.1330331 -2.0303814
LLEVER    1.89143444  2.6101123 -4.5476300  2.0071943 -0.6385808 -0.2136826
LCASHLTD -0.22607207  0.4307177 -0.3604615  0.2167872  0.1257777 -0.1415353
LACIDRAT -6.82442048  4.1009148  0.3525305  2.2692405 -2.0724229  1.6638000
LCURRAT   5.46079189  4.8637491  0.1367320  1.5815373 -2.6709414 -1.0739454
LRECTURN -2.78633528  0.3970923  0.3304194 -0.3107691 -1.6483496 -0.4048196
LASSLTD  -0.37738469  2.1209424 -1.2296289 -0.5837193 -1.2200023  0.2157965

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5    LD6
0.4849 0.2924 0.1060 0.0961 0.0128 0.0077
```

```
p     1 2 3 4 5 6 7
  1   2 0 0 0 0 0 0
  2   0 1 0 0 0 0 0
  3   0 0 2 0 0 0 0
  4   0 0 0 2 0 0 0
  5   0 0 0 0 2 0 0
  6   0 0 0 0 0 2 0
  7   0 1 0 0 0 0 2
```

After applying LDA on validation sheet, I find that there are almost all the companies in the level they should be. But in level 2 which means AA level, there is only one company in level 7 the riskiest level C.

c).

It depends, if the actual level is higher than the misclassification error, it is kind of good for the companies who borrow the bond, however, it is bad for the company who lend the bond; if the actual level is lower than the misclassification error, vise versa.

5.

a.

kdf <- read.table("/Users/Yiyang/Documents/CSC 424/kellog.dat", header = FALSE, skip = 2)
head(kdf)

```
              V1     V2  V3     V4     V5     V6     V7     V8  V9    V10 V11
1         AllBran 0.1818 0.6 0.3333 0.8125 0.6429 0.0000 0.3333 1.0 0.9677   0
2   AllBranFlakes 0.0000 0.6 0.0000 0.4375 1.0000 0.0667 0.0000 1.0 1.0000   0
3      AppleJacks 0.5455 0.2 0.0000 0.3906 0.0714 0.2667 0.9333 0.5 0.0323   0
4      CornFlakes 0.4545 0.2 0.0000 0.9063 0.0714 0.9333 0.1333 0.0 0.0484   0
5         CorPops 0.5455 0.0 0.0000 0.2813 0.0714 0.4000 0.8000 0.5 0.0000   0
6 CracklinOatBran 0.5455 0.4 1.0000 0.4375 0.2857 0.2000 0.4667 1.0 0.4516   0
```

b.

d <- dist(kdf[, 2: 11])
d

	1	2	3	4	5	6	7	8	9	10	11	12
2	0.7272322											
3	1.5697390	1.8492052										
4	1.8539219	2.0305045	1.2684015									
5	1.6662546	1.8829115	0.2975358	1.1916989								
6	1.0996082	1.5440580	1.3176119	1.7678896	1.3547525							
7	1.5841069	1.7634645	1.1490144	1.0299818	1.0490237	1.3903578						
8	1.5091412	1.8468267	0.3398889	1.2719303	0.4316207	1.0596577	1.1567554					
9	1.8023558	2.0544616	0.6520781	0.8403389	0.6142367	1.6040403	1.2435004	0.7170009				
10	1.4347282	1.4944984	0.7279911	1.2266398	0.6708088	1.2503425	1.0635731	0.7637698	0.9792681			
11	1.6882887	1.9283144	1.3544907	1.5519973	1.3081891	1.3658479	1.1206756	1.2872281	1.5243700	1.3316102		
12	1.7672470	2.0343733	1.4196422	1.6397532	1.4069083	1.4332747	1.2096828	1.3654066	1.6087575	1.3995626	0.4606777	
13	1.4552656	1.8747545	1.1436096	1.6629513	1.1815905	0.8375832	1.2044443	0.9849233	1.4502875	1.2038583	1.3053475	1.1438155
14	1.4997064	1.8448304	0.5907124	0.9354137	0.5692408	1.0618428	0.8392057	0.4448509	0.6633024	0.8014891	1.1529193	1.2206573
15	1.4212580	1.7941160	1.2974982	1.3748080	1.2789268	0.8945117	0.8685676	1.1368609	1.4514440	1.2355117	1.1818840	1.0842522
16	1.2550124	1.3637588	1.1218121	1.1317151	1.0642234	1.2221023	0.4469872	1.1258005	1.3028190	0.8507743	1.1355754	1.2484837
17	1.7728810	1.9876043	1.6030319	1.4348174	1.5871723	1.7389280	1.0736413	1.6085895	1.6605647	1.5915576	0.6792585	0.8088116
18	1.0511532	1.4668203	0.9100617	1.2953081	1.0048503	1.0022757	1.2363397	0.8389299	1.0476095	0.9541119	1.3902247	1.3279287
19	1.4007230	1.4490070	0.9288793	1.4575319	0.8181519	1.1943371	0.8817143	0.9525259	1.2882294	0.5597161	1.2146529	1.3392266
20	1.9205121	2.1198904	1.2636303	0.1492539	1.1866194	1.7930692	1.0284355	1.2706732	0.8331199	1.2510384	1.5538330	1.6129819
21	1.5809483	1.8995743	0.4048276	1.4944746	0.5198887	1.1255875	1.3644799	0.2571313	0.8664377	0.8128402	1.4129162	1.4820527
22	1.6873679	1.8735937	1.2857019	0.8961782	1.3616727	1.6825281	1.3261089	1.2926246	1.1490633	1.1248803	1.6825437	1.6832981

	13	14	15	16	17	18	19	20	21
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14	0.9130613								
15	0.5668569	0.8665760							
16	1.2027871	0.8995366	0.9183382						
17	1.6330287	1.3695575	1.3585262	1.1365256					
18	0.8123873	0.7950756	0.9329151	1.1051015	1.5842025				
19	1.1730249	0.9484267	1.1565091	0.6653321	1.4995015	1.1368606			
20	1.6268183	0.9192802	1.3498559	1.1690533	1.4425548	1.2965719	1.4802744		
21	1.0690075	0.6884662	1.3131275	1.2997845	1.7882332	0.9427259	1.0090331	1.4934093	
22	1.6302255	1.0940138	1.4632114	1.2180429	1.5873017	1.2445166	1.5069150	0.9189203	1.4530108

c.

fit <- cmdscale(d, eig = TRUE, k = 2)

fit

```
$points
              [,1]         [,2]
 [1,] -1.03418269 -0.245530932
 [2,] -1.19326919 -0.279718754
 [3,]  0.28708586 -0.426579763
 [4,]  0.72242702  0.084885505
 [5,]  0.35805654 -0.357843276
 [6,] -0.64765067 -0.170378924
 [7,]  0.13566780  0.386797628
 [8,]  0.20301598 -0.404308778
 [9,]  0.69269615 -0.396851417
[10,]  0.03261563 -0.338750546
[11,] -0.08945545  0.796019209
[12,] -0.09113710  0.848345562
[13,] -0.28636447 -0.006340016
[14,]  0.29751912 -0.124652605
[15,] -0.20797764  0.296956461
[16,] -0.16929999  0.221151430
[17,]  0.04091922  1.087443450
[18,] -0.20592492 -0.297227127
[19,] -0.21601741 -0.113423789
[20,]  0.79149376  0.113552457
[21,]  0.14493948 -0.560781727
[22,]  0.43484295 -0.112764048
```

```
$eig
 [1]  5.371962e+00  4.199940e+00  3.432574e+00  1.817835e+00  1.276243e+00  6.582689e-01  4.837697e-01  3.257599e-01
 [9]  5.385893e-02  1.349159e-02  5.133927e-16  3.518514e-16  3.486236e-16  3.159873e-16  2.945104e-16  1.010249e-16
[17] -7.556221e-17 -9.771501e-17 -1.667971e-16 -3.043729e-16 -3.837904e-16 -8.459685e-16

$x
NULL

$ac
[1] 0

$GOF
[1] 0.5428186 0.5428186
```

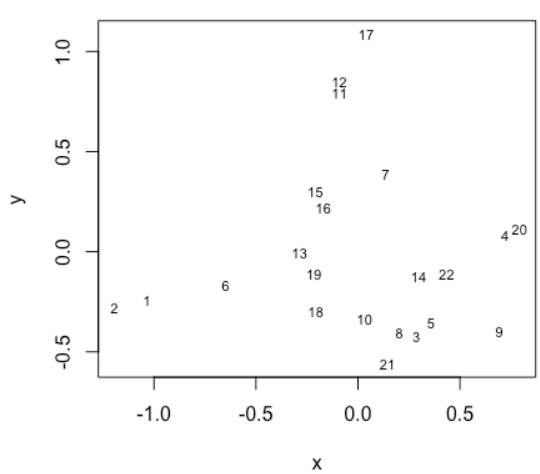
d.

x <- fit$points[, 1]

y <- fit$points[, 2]

plot(x, y, type="n")

text(x, y, labels = row.names(kdf), cex=.7)



e.

I would divide into five groups, (2, 1, 6), (7, 15 ,16, 13, 19 ,18), (4, 20), (17, 12, 11), (10, 21, 8, 3, 5, 9, 14, 22)

f.
Five groups, the distinct group is (4, 20) from the plot.

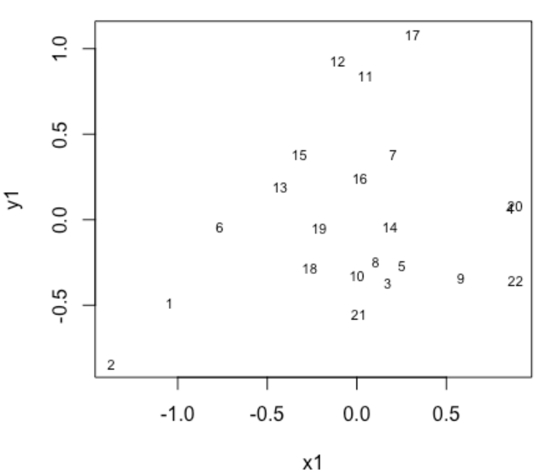
g.
Group (2, 1, 6), the names of them are “All Bran”, “All Bran Flakes”, “Cracklin Oat Bran”, their names all have Bran, which means they have similar ingredients, then they fall into a group. Group (17, 12, 11), the names of them are “Just Right”, “Just Right Fruit Nut”, “Product 19”, although only the first two have the similar names, comparing the variables, these three products also have some similar ingredients.

h.
I find variable V9 is a good dimension, after analysis, I would infer that, V9 should be the fat of the Kellogg production.

i.
mds <- isoMDS(d)
mds
x1 <- mds$points[, 1]
y1 <- mds$points[, 2]
plot(mds$points, type = "n")
text(x1, y1, labels = row.names(kdf), cex=.7)

```
$points
              [,1]        [,2]
 [1,] -1.0456570471 -0.48862265
 [2,] -1.3709102019 -0.84438640
 [3,]  0.1727535899 -0.36939825
 [4,]  0.8580549470  0.06548128
 [5,]  0.2507671664 -0.27152298
 [6,] -0.7661716477 -0.04267063
 [7,]  0.2022623264  0.38124818
 [8,]  0.1051078269 -0.24650621
 [9,]  0.5816347128 -0.33907330
[10,]  0.0007362656 -0.32447482
[11,]  0.0488667545  0.84134516
[12,] -0.1072958170  0.92942832
[13,] -0.4287958947  0.19476285
[14,]  0.1847247383 -0.04162486
[15,] -0.3207596002  0.38297593
[16,]  0.0169327872  0.24468830
[17,]  0.3107729356  1.08348337
[18,] -0.2635717661 -0.27966222
[19,] -0.2103007422 -0.05065814
[20,]  0.8836460596  0.08187876
[21,]  0.0112503568 -0.55230444
[22,]  0.8859522497 -0.35438724

$stress
[1] 14.17948
```



Comparing the two plot, I find all the patterns are still in the similar position with the plot in d. However, all the patterns are moving along with the vector <-0.5, 0.5>. Pattern 4 and 20 are almost coincident.