DSC 424: Advanced Data Analysis and Regression

Assignment 04

Name: Sidhant Thakur

Student ID: 2020181

Problem 1

a) They use CA to study that was carried out by randomly selecting participants from the area associated with the hospitals and asking them which of 13 different features they associated with the respective hospitals, to better understand patients' perceptions about the hospitals.

The type to categorical variable they contain are expert emergency treatment (emer), cancer treatment (canc), special program for senior(snrs), outpatient services (outp), women's health services (outp), laser surgery (lasr), offering community program (comm), and advanced technological equipment etc.

b) They use graph from CA in their analysis based on two principal components derived from correspondence analysis because CA scales the row and column of input data matrix. The graph provides a joint graphic representation of the features and the hospitals in two-dimensional space. They use proximity as an indication of the feature's, patients view those different hospitals is having.

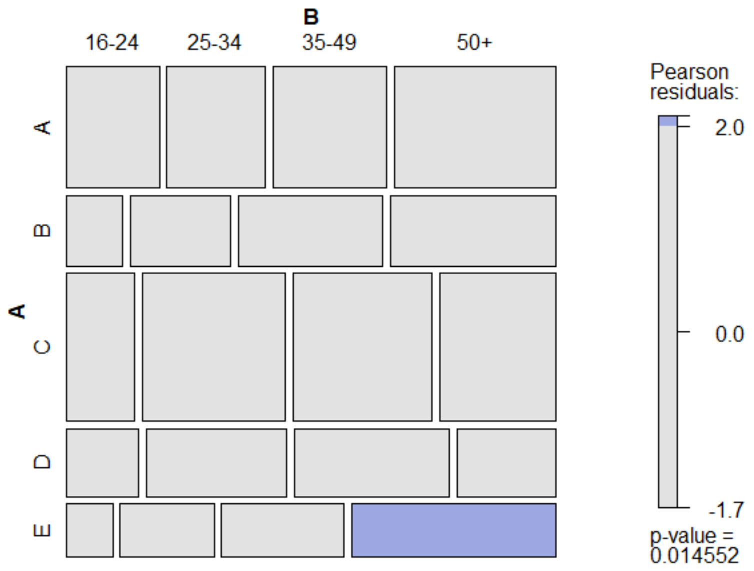
c) There is no evidence of goodness of fit, but they check the performance of the correspondence by evaluation the amount of variance explained by the factors

d) The conclusion the CA allow them to draw is that correspondence analysis enabled LHS planners to visualize their hospitals' comparative advantages and disadvantages in relation to their competitors' positions of strength and weakness. Moreover, the impactful of the conclusion is that CA is useful in determining the relative importance of each attribute. In addition, correspondence analysis programs present a graphic display of the relative positions of attribute and objects in the same low-dimensional joint space

For example, the study was able to suggest that the Cleveland Clinic is most closely associated with "cancer treatment, heart disease prevention and having advanced technological equipment. This highlights the hospital's strengths and allows management to improve on different areas of focus.

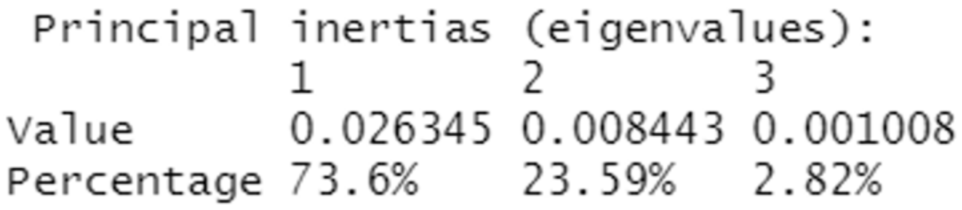
Problem 2

a) Create a mosaic plot using the contingency table in the csv file.



From, the mosaic plot it can be interpreted that store E with age profile 50+ have a high correspond

b) Plot the correspondence analysis. Which two variables have the highest correspondence. The least?

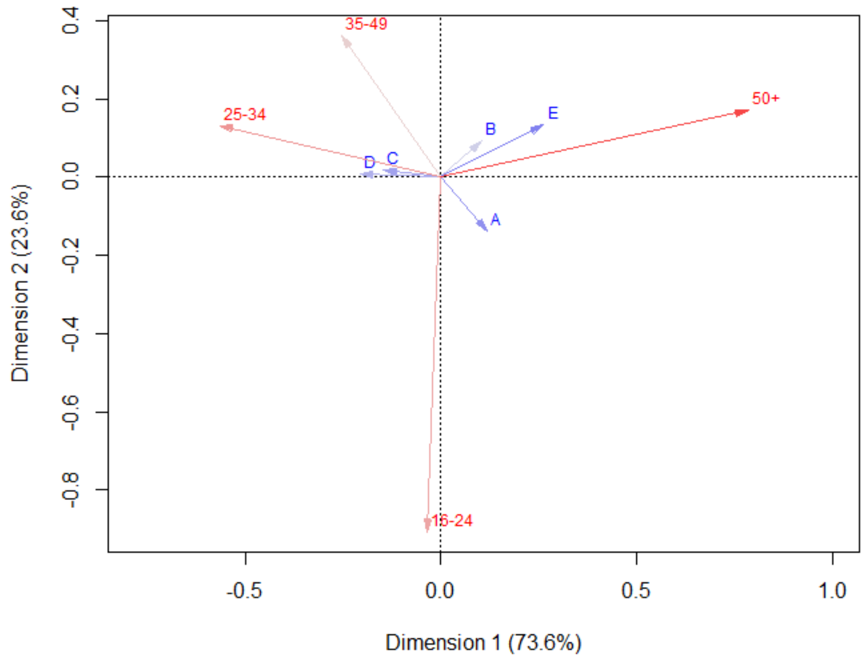


```
 Principal inertias (eigenvalues):
           1        2        3
Value      0.026345 0.008443 0.001008
Percentage 73.6%    23.59%   2.82%
```

According to the above table, the first component accounts for 73.6 percent of the variance and has an eigenvalue of 0.026.

Second, with an eigenvalue of 0.0084, accounts for 23.59 percent of the variance.

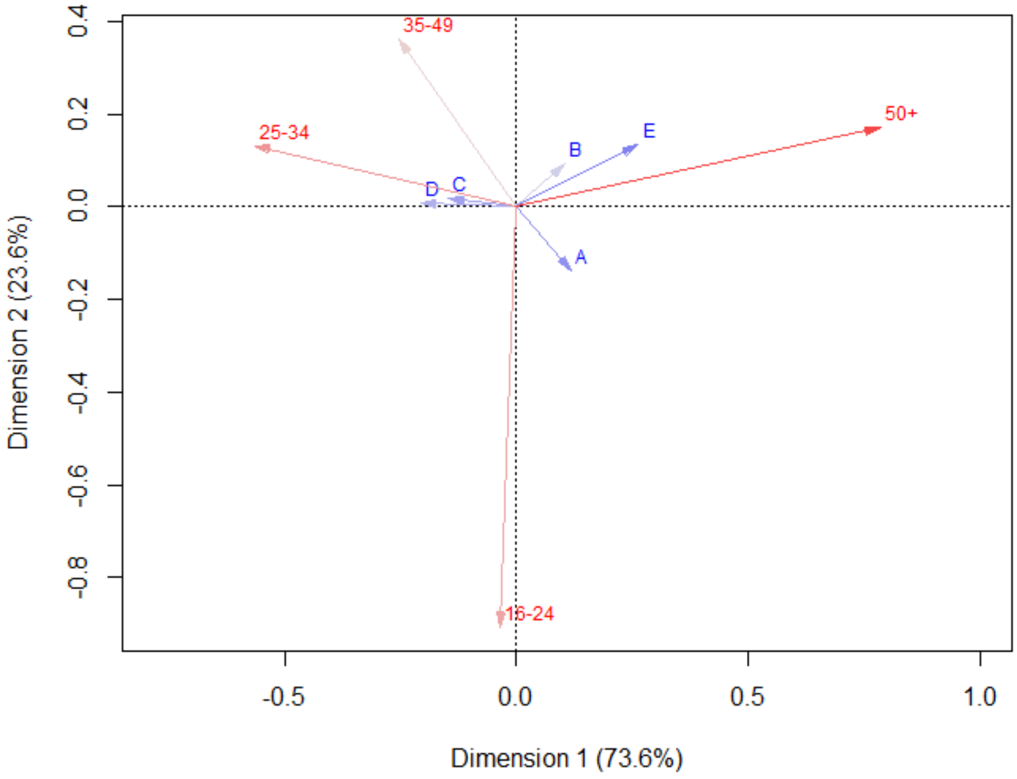
The third component is responsible for 2.82 percent of the variance and has an eigenvalue of 0.001.



The age groups 25-34 and 35-49 have the greatest correspondence.

Age groups 50+ and 16-24 have the least correspondence.

c) With each store, create an age profile for the store. Which customer ages are most highly and least highly represented?



Store A

The customer of Age group of 16-24have the highest correspondence, after that Age group of 50+ have the highest correspondence, after that Age group of 35-49 have the highest correspondence.

The customer of Age group of 25-34 have the least correspondence.

Store B

The customer of Age group of 50+ have the highest corresponds, after that Age group of 35-49 have the highest corresponds, after that Age group of 25-34 have the highest correspondence.

The customer of Age group of 16-24 have the least correspondence.

Store C

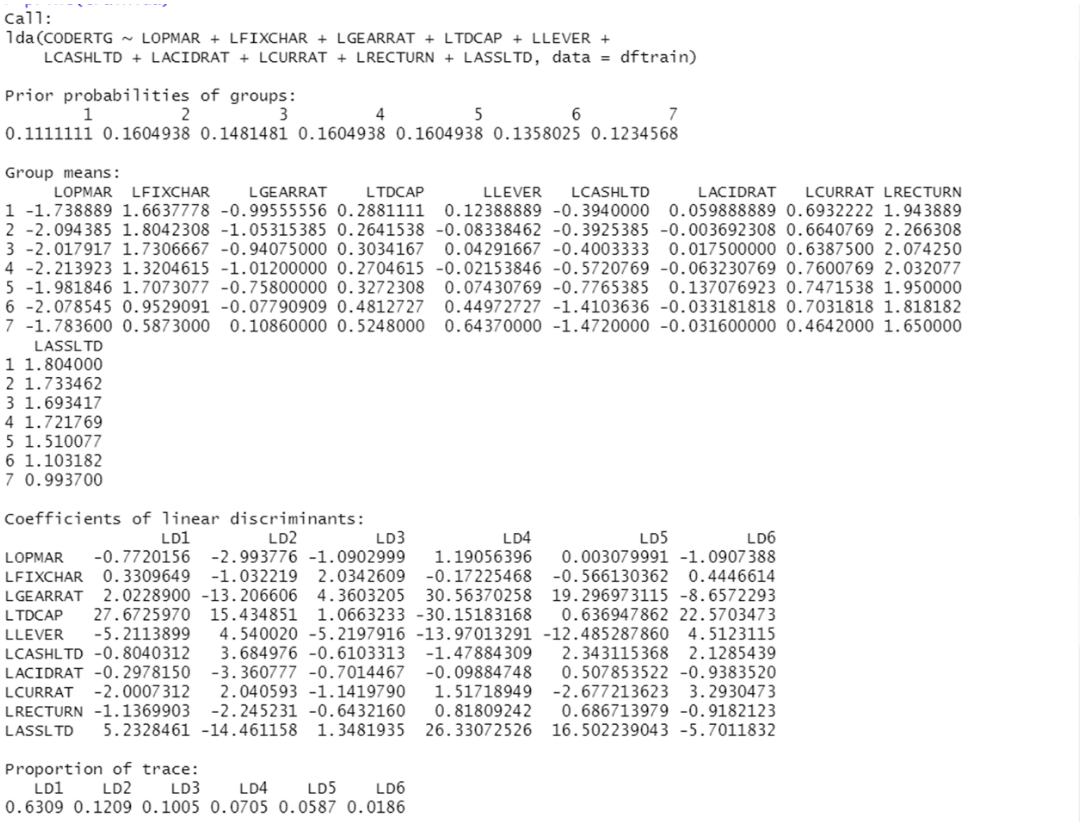
The customer of Age group of 25-34 have the highest correspondence, after that Age group of 35-49 have the highest correspondence, after that Age group of 16-24 have the highest correspondence.

The customer of Age group of 50+ have the least correspondence.

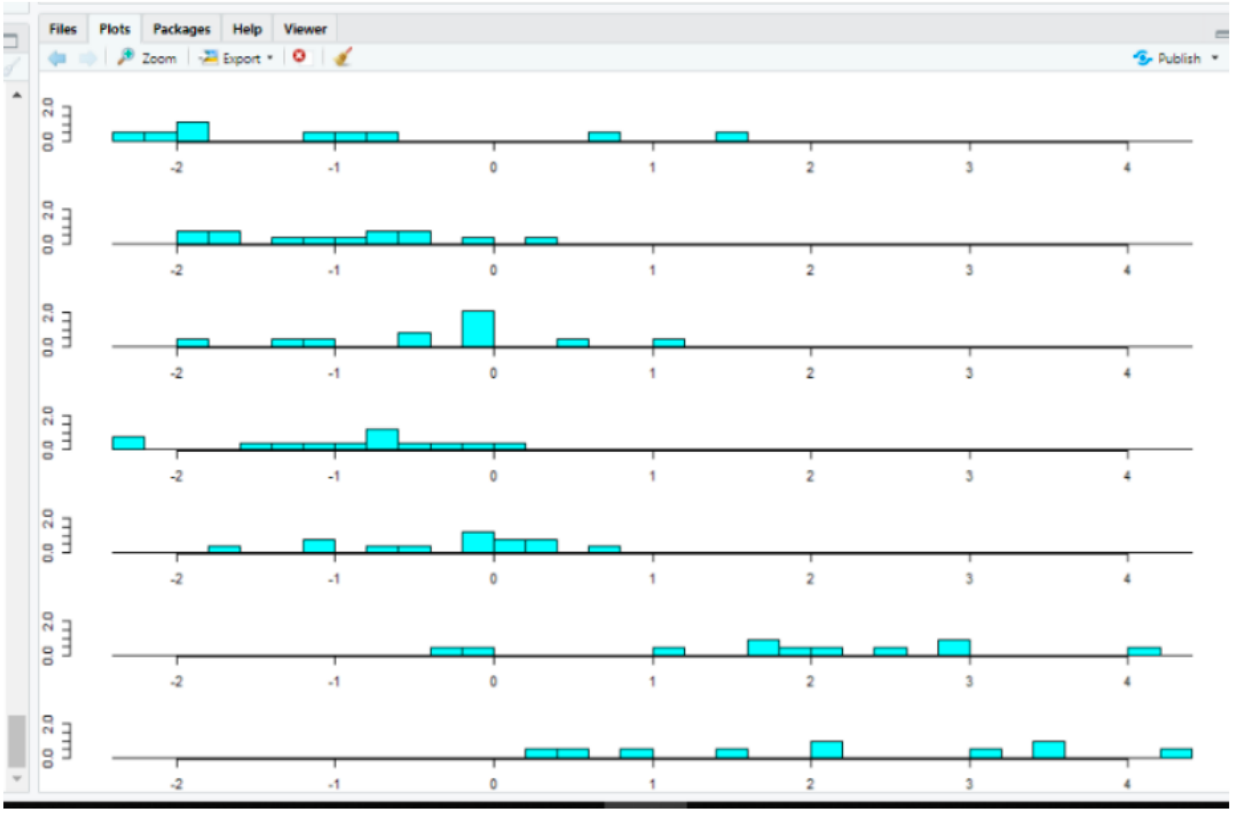
Store D
The customer of Age group of 25-34 have the highest correspondence, after that Age group of 35-49 have the highest corresponds, after that Age group of 16-24 have the highest correspondence.
The customer of Age group of 50+ have the least correspondence.
Store E
The customer of Age group of 50+ have the highest correspondence, after that Age group of 35-49 have the highest correspondence, after that Age group of 16-24 have the highest correspondence.
The customer of Age group of 25-34 have the least correspondence.

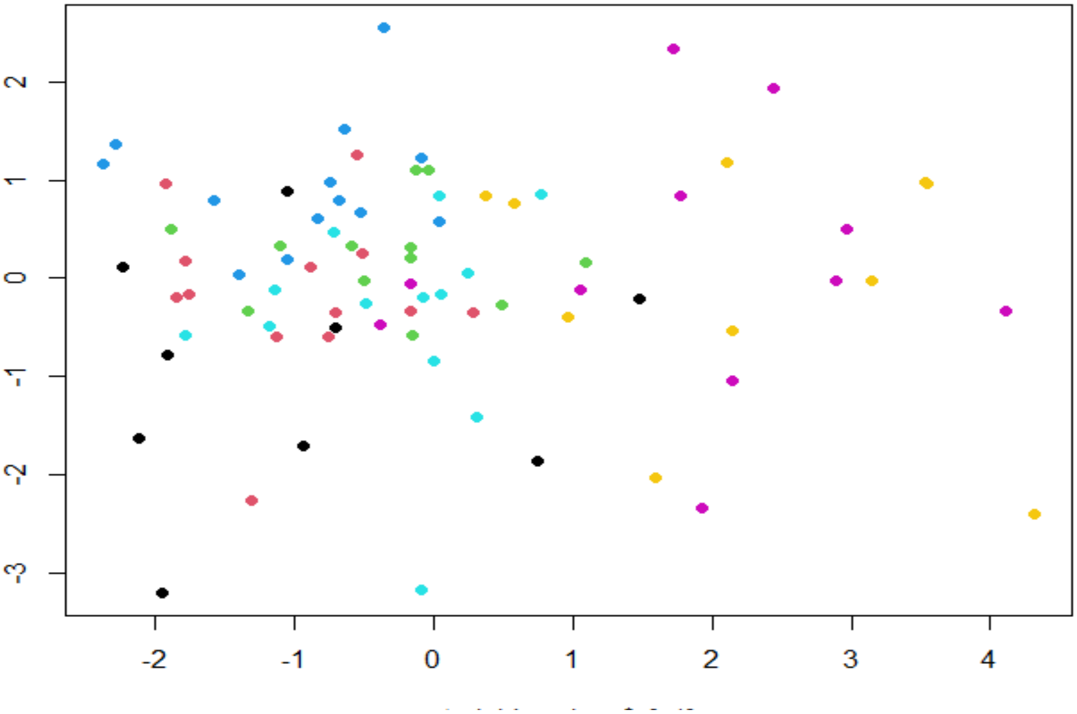
Problem 3

a) **What is the performance of the classifier on the training data? Notice that there is order in the class variables (i.e., AAA is better than AA, which is better than A…).**



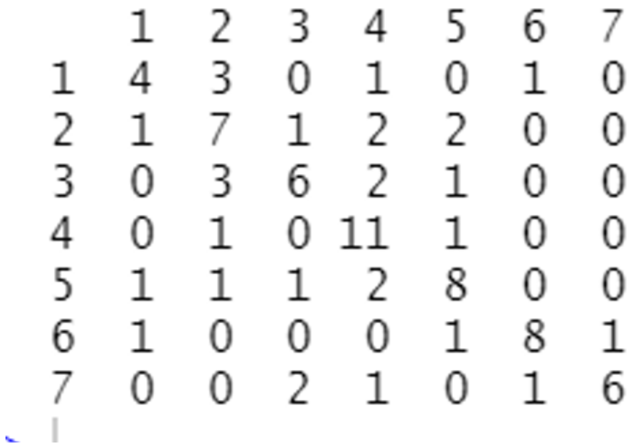
```
Call:
lda(CODERTG ~ LOPMAR + LFIXCHAR + LGEARRAT + LTDCAP + LLEVER +
    LCASHLTD + LACIDRAT + LCURRAT + LRECTURN + LASSLTD, data = dftrain)

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


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
2.0
0.0
-2
-1
0
1
2
3
4


2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4

Plotting the histogram as shown above we see that there is a lot of overlap and poor separation of the groups. When we plot the histogram, we see that there is a lot of overlap and poor group separation. Plotting the results shows little improvement and the separation in groups still appears to be overlapping significantly.

Confusion matrix:



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

We can look at the confusion matrix with the true values on the rows and the predicted values on the columns and gain some valuable insight.

According to the confusion matrix, 4 points for **AAA** are correctly classified, while three points are incorrectly classified.

For **AA**, 7 points are properly identified, while 8 points are wrongly categorized.

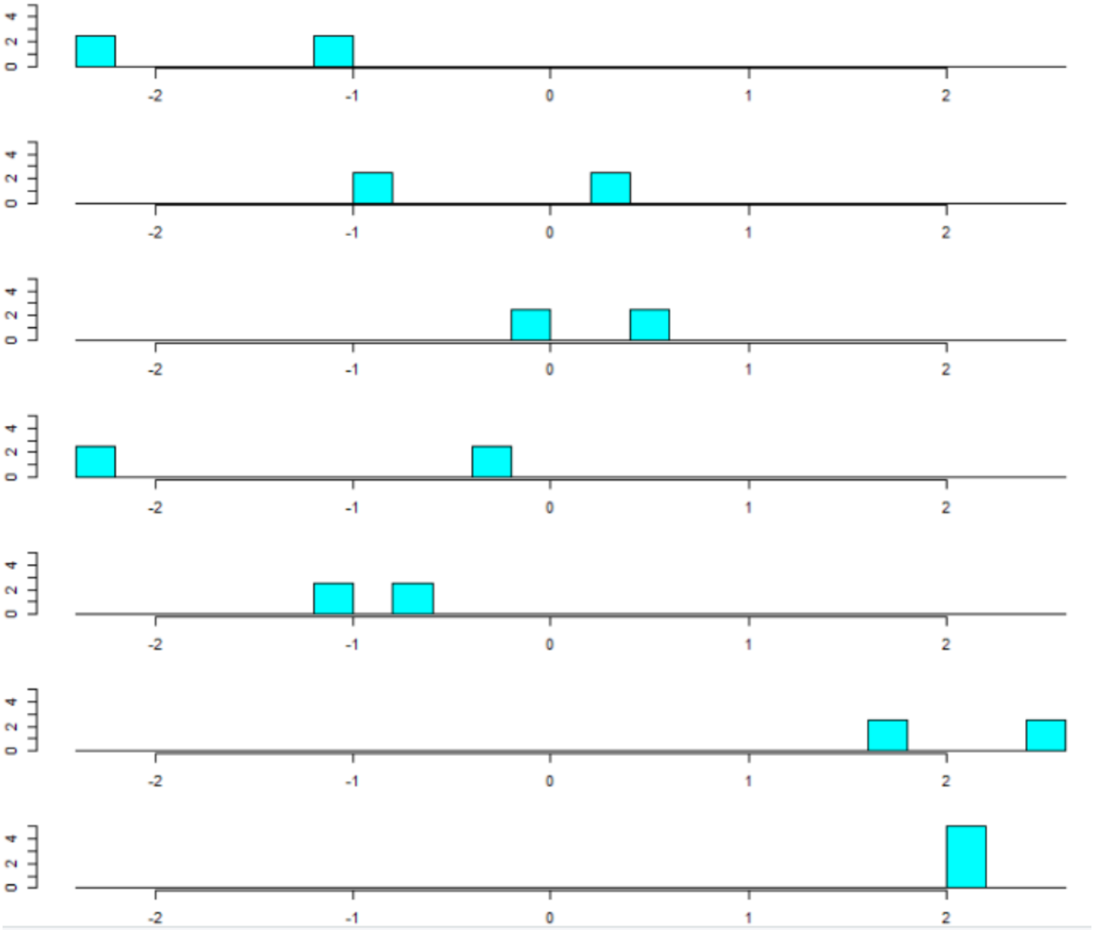
For **A**, 6 points are correctly identified while 4 points are wrongly categorized.

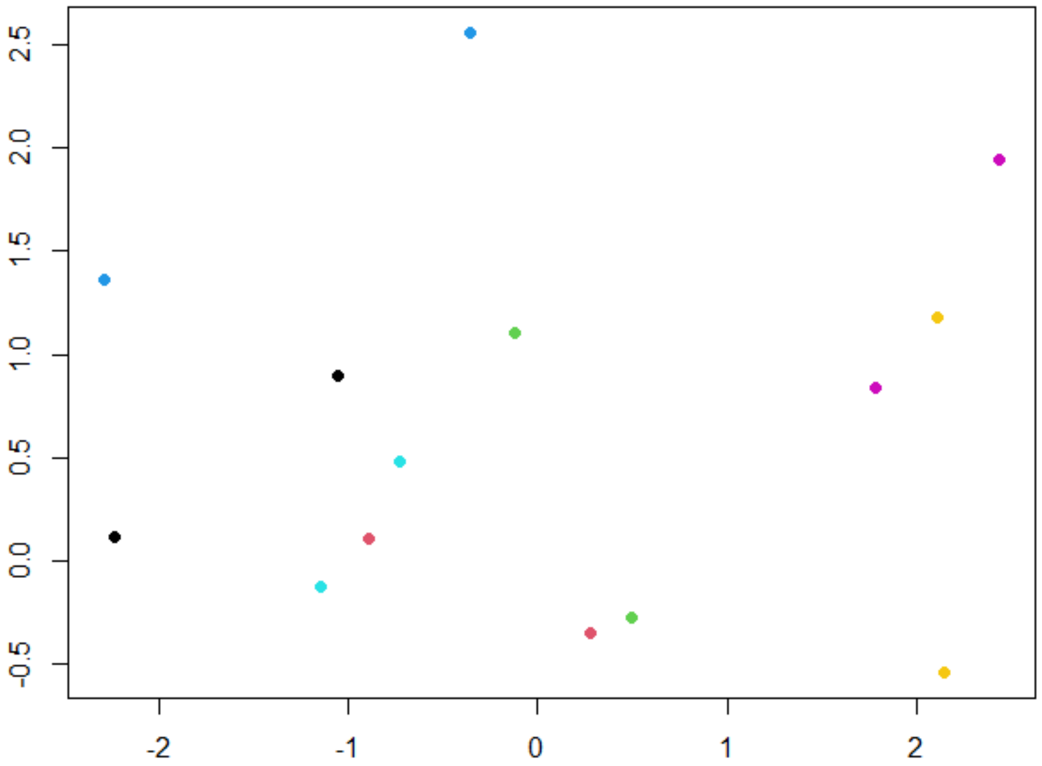
For **BAA**, 11 points are categorized correctly, whereas 8 points are misclassified. For **BA**, 8 points are assigned to the right class, while 5 points are incorrectly assigned.

For **B**, 8 points are assigned to the right class, while 2 points are incorrectly assigned.

For **C**, 6 points are categorized as accurate, while 1 point is misclassified.

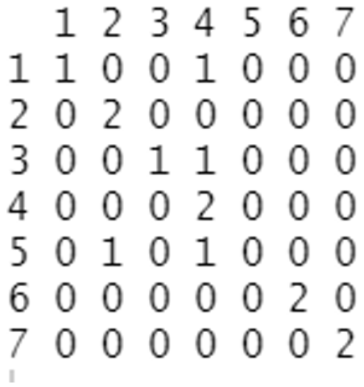
b) What is the performance of the classifier on the validation data?





Next, we can examine the classifier's performance on the validation set to see how well it can divide the data into relevant groupings. A histogram of the scatter of each group reveals that groups 1,4,5,6, and 7 have good separation. However, when we look at the plots, we can see that there is little to no separation between the groups.

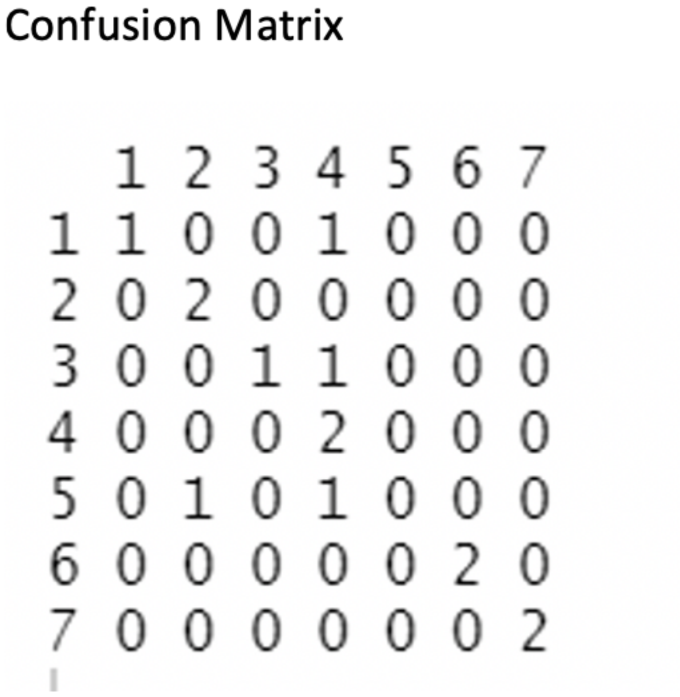
Confusion Matrix



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

I find that there are almost all the companies in the level they should be. But in level 2 which mean **AA**, 2 points are correctly classified and 1 incorrectly classified, moreover for level 4, which mean **BAA**, 2 points are correctly classified, while 3 are incorrectly.

c)Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?



Confusion Matrix

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

The worst type of mistake would appear to be classifying a bond as not risky when it was extremely risky, as this would mean investments could be lost when the underlying asset was thought to be stable. This is opposed to accidently classifying a safe investment as risky and simply forgoing the investment or assuming less risk than originally believed.