Name: Syed Noor Razi Ali
StudentID: 2070326

DSC 424
ASSIGNMENT 4

PART 1: Paper Review

A) How are they applying CA? What variables are being analyzed and what types of categorical levels do they contain?

They use CA to categorize data from many perspectives and characteristics, as well as how services are linked to hospitals.
There are 13 variables in the study with 3 variables - emer, canc, and tech which are the categorical variables.

B) How did they use graphs from the CA in their analysis?

They used a two-dimensional graph based on CA's two major components. It shows the features and hospitals in a single graphic Interpretation of the graph is proximity among rows and columns of contiguous table (features and hospital).

C) Did they use any techniques to evaluate goodness of fit? If not, was it appropriate that they did not? How would it have helped their exposition if they had? If they did, what were the results?

The eigenvectors can be used to count the variance and reveal the underlaying structure and position of features of components. But there is nothing to evaluate goodness of fit such as hypothesis testing that value is countable or not. They will able to give the number of dimensions to display.

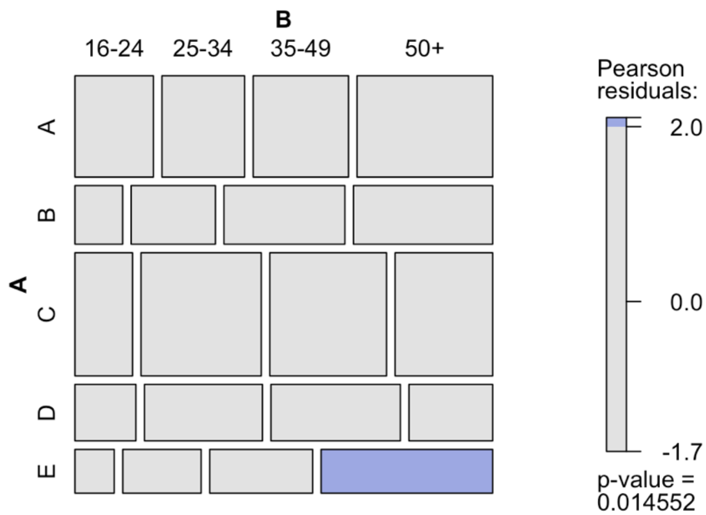
D) What conclusions does CA allow them to draw? How impactful are those conclusions? Are there any practical, actionable implications from their conclusions?

CA can be used in the healthcare system to examine the hospital for strategy development, marketing, and offering new services, according to the study. Yes, the conclusion has implications for the health-care industry in terms of analyzing patients and diseases.

Name: Syed Noor Razi Ali
StudentID: 2070326

PART 2: Correspondence Analysis

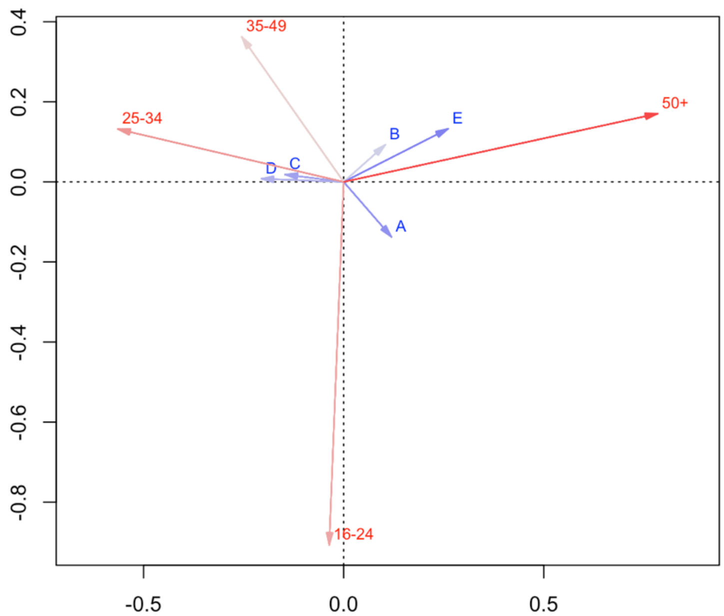
A: Create a mosaic plot using the contingency table in the csv file.



Here, from the plot we can see that the Store E with Age group of 50+ have a high correspondence.

B: Plot the correspondence analysis. Which two variables have the highest correspondence. The least?

Name: Syed Noor Razi Ali
StudentID: 2070326



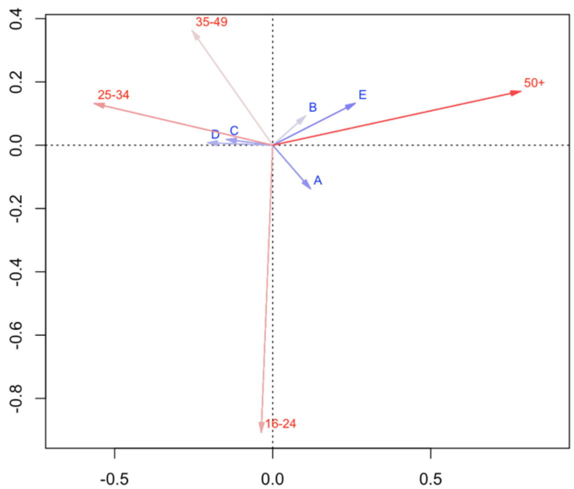
From the graph we can say that 25-34 and 35-49 have highest correspondence.
50+ and 16-24 have least correspondence.

```
> summary(fit)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.026345  73.6  73.6  ******************
 2      0.008443  23.6  97.2  ******
 3      0.001008   2.8 100.0  *
        -------- -----
 Total: 0.035797 100.0


Rows:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |    A |  264 1000  245 |  119 430 143 | -138 570 592 |
2 |    B |  153  889   93 |  104 496  63 |   93 393 155 |
3 |    C |  321  961  203 | -146 946 261 |   18  15  13 |
4 |    D |  147  966  181 | -206 965 237 |    8   1   1 |
5 |    E |  114  986  278 |  261 784 296 |  133 202 239 |

Columns:
    name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 | 1624 |  153  997  196 |  -15   5   1 | -213 992 822 |
2 | 2534 |  254  954  250 | -182 937 318 |   24  16  17 |
3 | 3549 |  286  843   93 |  -77 512  65 |   62 332 131 |
4 |   50 |  307  997  461 |  230 982 615 |   28  15  29 |
```

Dim1 accounts for 73.6% of the separation or variability of the data with eigenvalues of 0.026. Dim2 accounts for 23.6% of variability with eigenvalue of 0.008 and dim1 and dim2 explain the 97.2% of the separation.

Name: Syed Noor Razi Ali
StudentID: 2070326

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



Store A
There is most high correspondence the Age group 50+ and second highest for age group 16-24. There is least correspondence with the Age group of 25-34

Store B
There is most high correspondence the Age group 50+ and second highest for age group 35-49. There is least correspondence with the Age group of 25-34

Store C
There is most high correspondence the Age group 25-34 and second highest for age group 35-49. There is least correspondence with the Age group of 50+

Store D
There is most high correspondence the Age group 25-34 and second highest for age group 35-49. There is least correspondence with the Age group of 50+

Store E
There is most high correspondence the Age group 50+ and second highest for age group 35-49. There is least correspondence with the Age group of 25-34

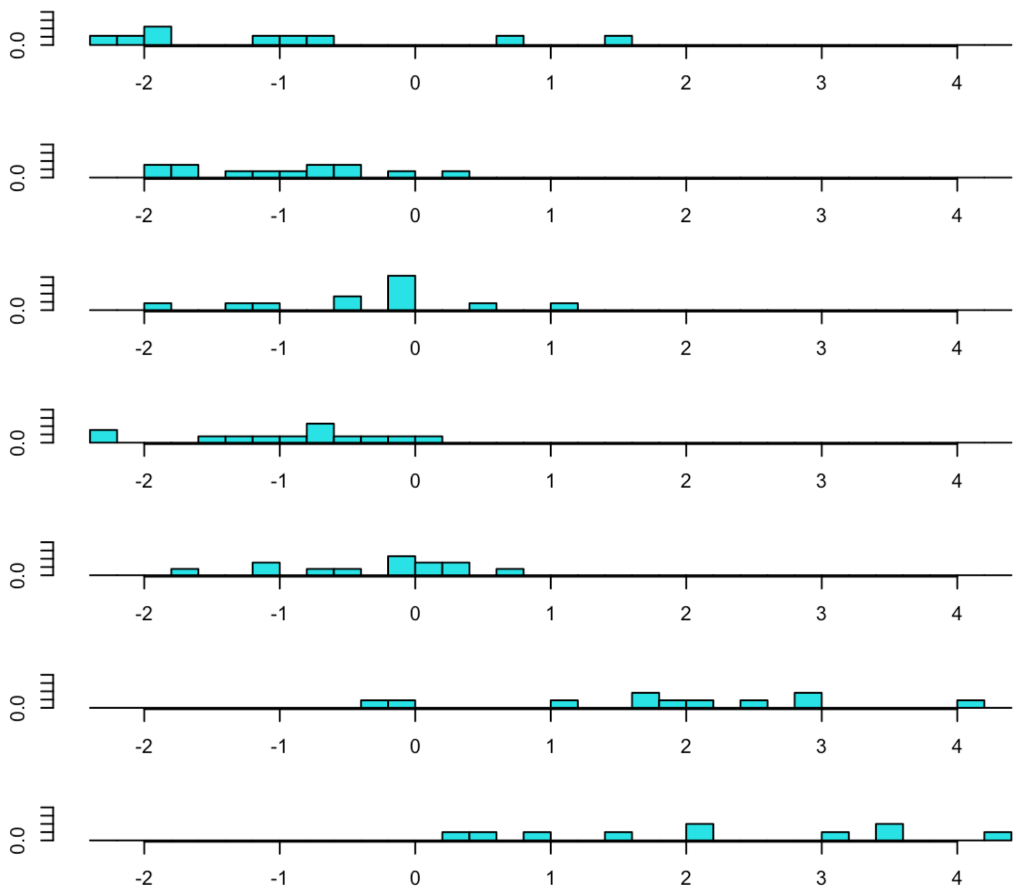
Name: Syed Noor Razi Ali
StudentID: 2070326

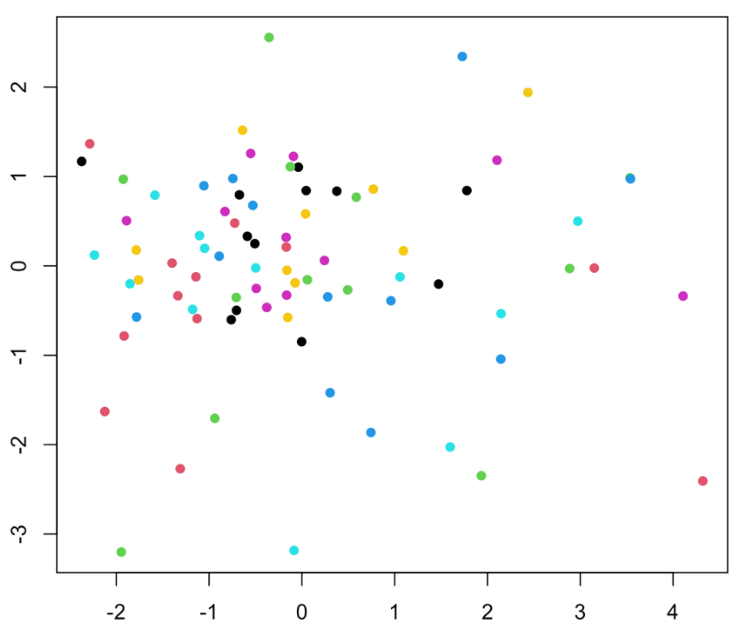
Part 3: Linear Discriminant Analysis

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 = TrainData)

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 
>
```

Name: Syed Noor Razi Ali
StudentID: 2070326





Name: Syed Noor Razi Ali
StudentID: 2070326

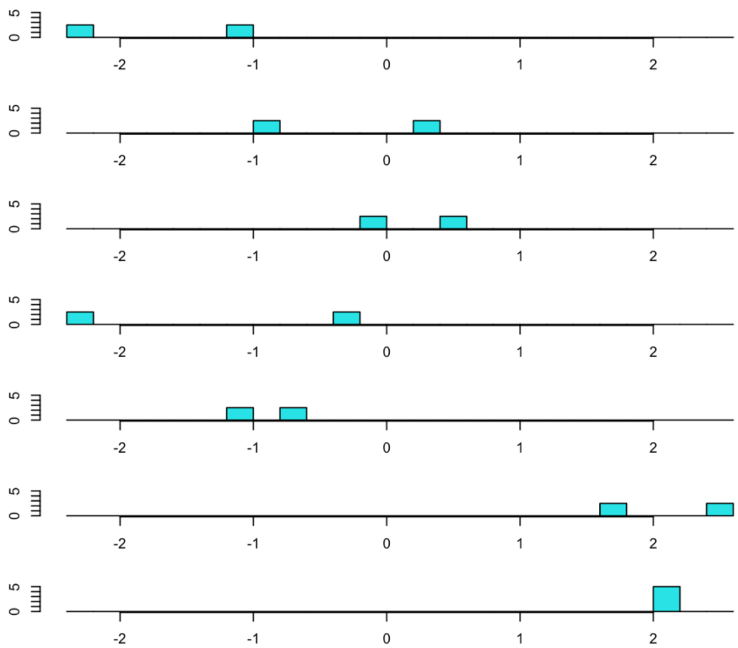
```
> # Compute a confusion matrix
> table(BondRatingTrain$CODERTG

    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
>
```

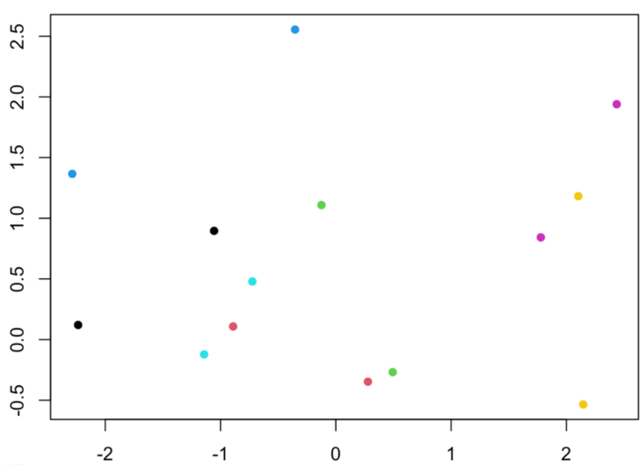
```
> source("Confusion.R")
> confusion(Trainlda.values$class, BondRatingTrain$CODERTG)
         Accuracy Prior Frequency.1 Prior Frequency.2 Prior Frequency.3 Prior Frequency.4 Prior Frequency.5
           0.6173            0.0864            0.1852            0.1235            0.2346            0.1605
Prior Frequency.6 Prior Frequency.7
           0.1235            0.0864

Confusion Matrix
      Predicted (cv)
Actual      1      2      3      4      5      6      7
     1 0.5714 0.1429 0.0000 0.0000 0.1429 0.1429 0.0000
     2 0.2000 0.4667 0.2000 0.0667 0.0667 0.0000 0.0000
     3 0.0000 0.1000 0.6000 0.0000 0.1000 0.0000 0.2000
     4 0.0526 0.1053 0.1053 0.5789 0.1053 0.0000 0.0526
     5 0.0000 0.1538 0.0769 0.0769 0.6154 0.0769 0.0000
     6 0.1000 0.0000 0.0000 0.0000 0.0000 0.8000 0.1000
     7 0.0000 0.0000 0.0000 0.0000 0.0000 0.1429 0.8571
>
```

We can deduce from the scatterplot and histogram that the data is scattered in general; there is no pattern or group in the graph, and studying the confusion matrix, where we can see the prediction error, we can deduce that certain organizations are present on several levels. In the matrix, we can see the percentage value of the actual and anticipated values. Only level C predicts 0.87 percent correctness, with a 0.14 percent BAA level misclassification.

B: What is the performance of the classifier on the validation data?

Name: Syed Noor Razi Ali
StudentID: 2070326






```
> # Compute a confusion matrix
> table(validationData$CODERTG

   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
>
```

When we apply classification to Validation data, we observe that most companies now fall into the category to which they belong, but there are no good scatters across groups as seen in the scatterplot. The level C, BAA, AA, and BA are well classified.

Name: Syed Noor Razi Ali
StudentID: 2070326

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

According to the confusion matrix, 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.
The True positive values are on the diagonals and misclassification points are on the off diagonals.
The confusion matrix indicates that 1 point for (1,1) is accurately identified, with no misclassification mistakes.
For (2,2), two points are correctly detected, but one is incorrectly classified.
1 point is accurately detected for (3,3) with no misclassification mistakes.
For (4,4), 2 points are properly identified, while 3 points are incorrectly classified.
For (6,6), the proper class receives 2 points while the erroneous class receives 0 points.
For (7,7), 2 points are correct, while 0 points are incorrect