Assignment 4

Erik Pak

2023-05-21

Problem 1

(20 points) Paper Review: An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. It contains a usage of Correspondence Analysis. Review the paper and evaluate their usage of the technique. In particular, address in detail the following points. You should be able to fill at least a page with your review and analysis, and each point should be answered in a complete paragraph with several sentences. If you claim something about the paper, you should be able to back it up with a quote or evidence from the paper.

Problem 1 a)

The study utilized Correspondence Analysis (CA) to determine the image of the Lake Hospital System (LHS), a multi-institutional system. By analyzing the data collected through a questionnaire survey, the study aimed to assess consumers' perception of LHS's quality and preference compared to its competitors. CA deemed appropriate for this dataset, allowing the researchers to analyze the strengths and weaknesses of each hospital's image across thirteen factors.

The findings of the study provide valuable insights for improving LHS's image. Based on the results obtained through correspondence analysis, recommendations, and implications can be made to enhance the LHS's image, including comparing its internal skill level with its external image of other healthcare providers. Finally, LHS used the insights to modify current offerings, plan new programs, and improve communication strategies. By implementing the suggested measures, LHS can effectively address the identified weaknesses and capitalize on its strengths, improving its overall image and consumer perception.

Problem 1 b)

The study contains the following categorical levels for the features:

emer: Expert emergency treatment

hart: Heart disease prevention and treatment

rehb: Rehabilitation services

CAnC: Cancer treatment

call: Call-in health information services

womn: Women's health services

lasr: Laser surgeries

outp: Outpatient services

docs: Doctors keeping up with medical advances

attn: Staff giving personal attention

shrs: Special programs for seniors

comm: Offering community programs

tech: Advanced technological equipment

Correspondence Analysis (CA) is applied in the study using categorical data to analyze the Lake Hospital System (LHS) image. The method involves converting any continuous variables into a discrete form. CA is a singular value decomposition of a matrix of chi-square distances, producing eigenvalues and eigenvectors used to calculate interpoint distances. The analysis aims to maximize the interrelationships between rows and columns of the data matrix.

Interpretation of CA involves examining eigenvalues associated with principal axes and analyzing the coordinates and contributions of the axes for rows and columns.

Overall, CA provides insights into the relationships and positioning of attributes and objects based on categorical data, aiding in the assessment and improvement of the Lake Hospital System's image.

Problem 1 c)

In the analysis, the researchers plotted the rows and columns of the contingency table regarding the first two factors of CA. The points on the map represent the attributes and objects, and their positions reflect their associations and similarities in the dataset. By examining the positions of attributes and objects on the graph, the researchers could gain insights into their relative proximity or distance from each other, indicating their interrelationships. In addition, the graphical display helped identify clusters or groupings of attributes and objects that shared similar characteristics or perceptions, including the visual representation provided by the graphs, enabling a deeper understanding of the relationships between attributes and objects.

Overall, using graphs in CA analysis allowed for a more intuitive and comprehensive understanding of the data, facilitating the identification of patterns, similarities, and associations between attributes and objects in the Lake Hospital System study.

d) 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, wheat were the results?

Problem 1 d)

This study seemed more like an exploratory analysis to determine the image of LHS and its competitors for market research after corporate reorganization. It was appropriate since it was an exploratory analysis that aimed to uncover insights and generate hypotheses rather than test specific hypotheses. The study performed Absolute Contributiontions to Variance by Rows and Columns since the table was provided with Correspondence Analysis (CA) to gain a better understanding of the most influential categories and their impact on the correspondence analysis.

The study concludes that Correspondence Analysis, along with the Absolute Contributiontions to Variance by Rows and Columns used to assess the significance and importance of the relationships captured by CA, can be valuable tools for evaluating consumer perceptions, understanding competitive positioning, and making strategic decisions in market research. Furthermore, by leveraging the insights gained from the analysis, organizations like LHS can better align their services and image with consumer expectations and preferences, ultimately improving their market performance.

Problem 1 e)

Correspondence Analysis (CA) is a powerful technique that helps LHS marketers understand and manage the multidimensional nature of hospital image. It condenses complex data into meaningful representations, allowing for easier interpretation and conclusion drawing. By visualizing relationships between factors, CA provides insights into the relative positioning of hospitals in the market. Marketers can identify patterns, clusters, and trends, enabling comparative analysis with competitors. This information helps develop targeted positioning strategies and communicate unique attributes to the target audience. CA also aids internal analysis, facilitating decisions on service expansion and market development. By aligning positioning and messaging strategies, hospitals can enhance their brand and competitive advantage.

Problem 1 e)

The study in question has several areas that could be improved. Firstly, it lacks demographic information about the participants, which is important for understanding the generalizability of the findings and potential biases. Secondly, the study would benefit from a face validity evaluation to ensure that the content accurately measures the intended variables. Additionally, checking for factor reliability is necessary to assess the consistency of the factors being measured. Lastly, it is important to determine the sample size and adequacy to ensure the reliability and representativeness of the findings.

Problem 2)

(20 points) The file "Survey.csv" contains survey responses to a questionnaire about Wikepedia pages. Each question's responses are on a 5 point likert scale. Perform an ordinal Principal Factor Analysis (exploratory) on this data addressing the following points. The questions are:

QU1: Articles in Wikipedia are reliable

QU2: Articles in Wikipedia are updated

QU3: Articles in Wikipedia are comprehensive

QU4: In my area of expertise, Wikipedia has a lower quality than other educational resources

QU5: I trust in the editing system of Wikipedia

VIS1: Wikipedia improves visibility of students' work

VIS2: It is easy to have a record of the contributions made in Wikipedia

VIS3: I cite Wikipedia in my academic papers

IM1: The use of Wikipedia is well considered among colleagues

IM2: In academia, sharing open educational resources is appreciated

IM3: My colleagues use Wikipedia

Load Libraries

```
library(tidyverse)      # Data manipulation package
library(corrplot)       # correlation plot
library(ggpubr)         # combine plots
library(psych)          # used for describe & pairs.panel plot
library(ca)             # correspondence analysis
library(polycor)        # hetcor
library(readxl)         # read excel sheet
library(MASS)           # Also contains a qda - Quadratic discriminant analysis function
library(rfUtilities)    # accuracy function
```

Problem 2a)

Pearson correlation measures linear relationships between continuous variables, while Spearman rank correlation evaluates the rankings or order of observations between variables. Kendall tau correlation assesses the ordinal association between variables, making it suitable for both continuous and ordinal data. Pearson correlation assumes normality and linearity, while Spearman and Kendall tau correlations are more robust and can handle data that violates these assumptions.

```
# import csv file
survey <- read.csv("Survey.csv")

# disply head
head(survey)
```

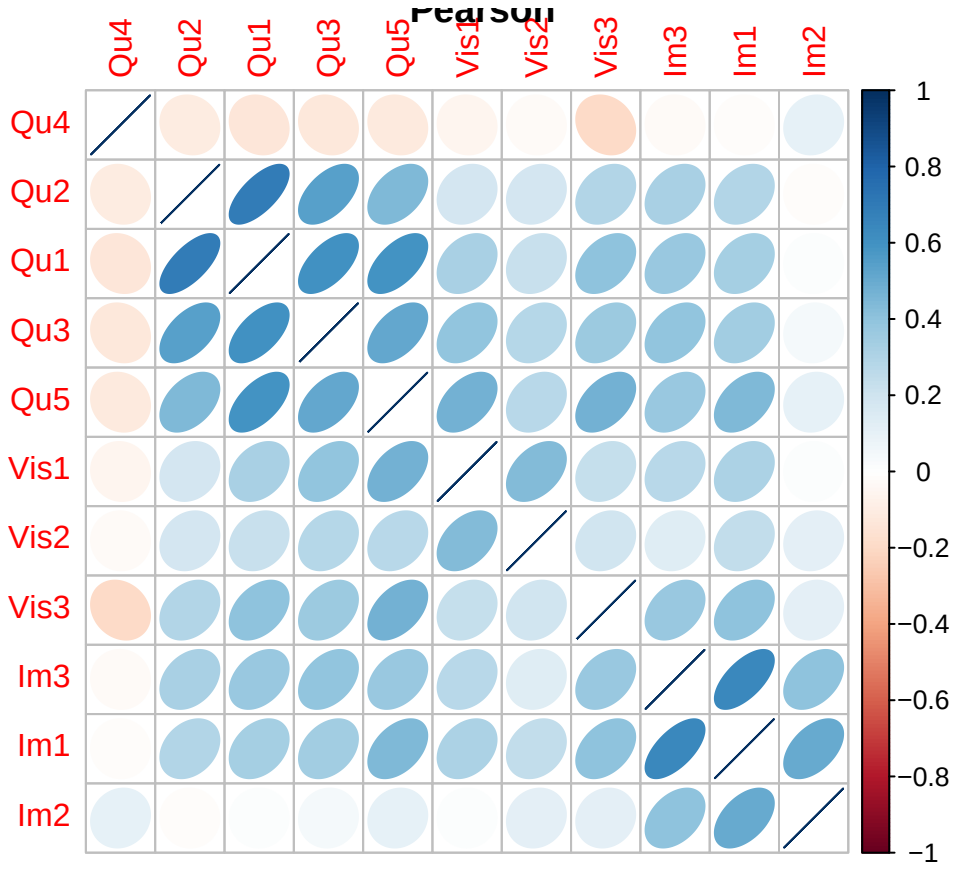
```
##   Qu1 Qu2 Qu3 Qu4 Qu5 Vis1 Vis2 Vis3 Im1 Im2 Im3
## 1   3   3   2   2   3    3    3    1   2   4   2
```

```
## 2   4   4   3   3   2    3    3    1   1   2   1
## 3   2   2   2   5   3    2    3    2   2   4   3
## 4   3   4   3   3   3    3    4    3   2   2   3
## 5   4   5   4   3   4    4    4    4   3   2   3
## 6   3   3   3   4   4    3    4    4   3   5   3
```

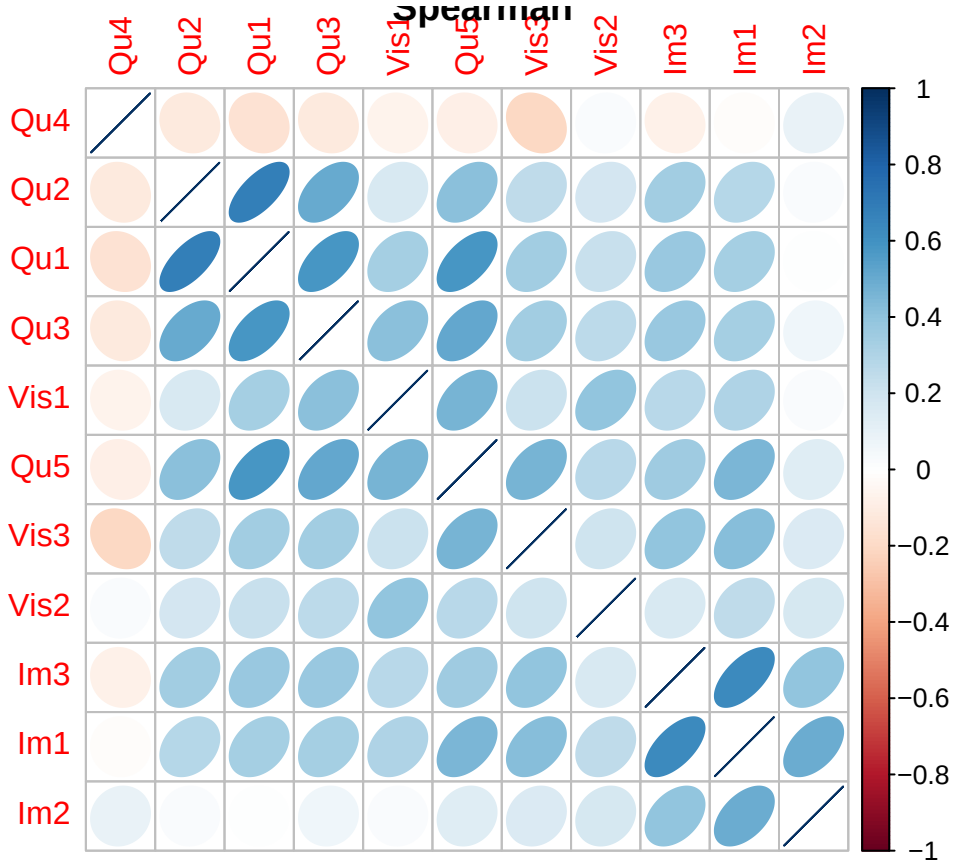
```
# summary
summary(survey)
```

```
##       Qu1             Qu2             Qu3            Qu4             Qu5
##  Min.   :2.000   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:2.00   1st Qu.:2.000   1st Qu.:2.500
##  Median :3.000   Median :4.000   Median :3.00   Median :3.000   Median :3.000
##  Mean   :3.218   Mean   :3.513   Mean   :2.95   Mean   :3.076   Mean   :3.084
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000   Max.   :5.000
##       Vis1            Vis2            Vis3           Im1             Im2
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:2.00   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :2.000   Median :2.00   Median :3.000
##  Mean   :3.067   Mean   :3.151   Mean   :2.092   Mean   :2.42   Mean   :3.303
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:3.00   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :4.00   Max.   :5.000
##       Im3
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :2.748
##  3rd Qu.:3.000
##  Max.   :5.000
```

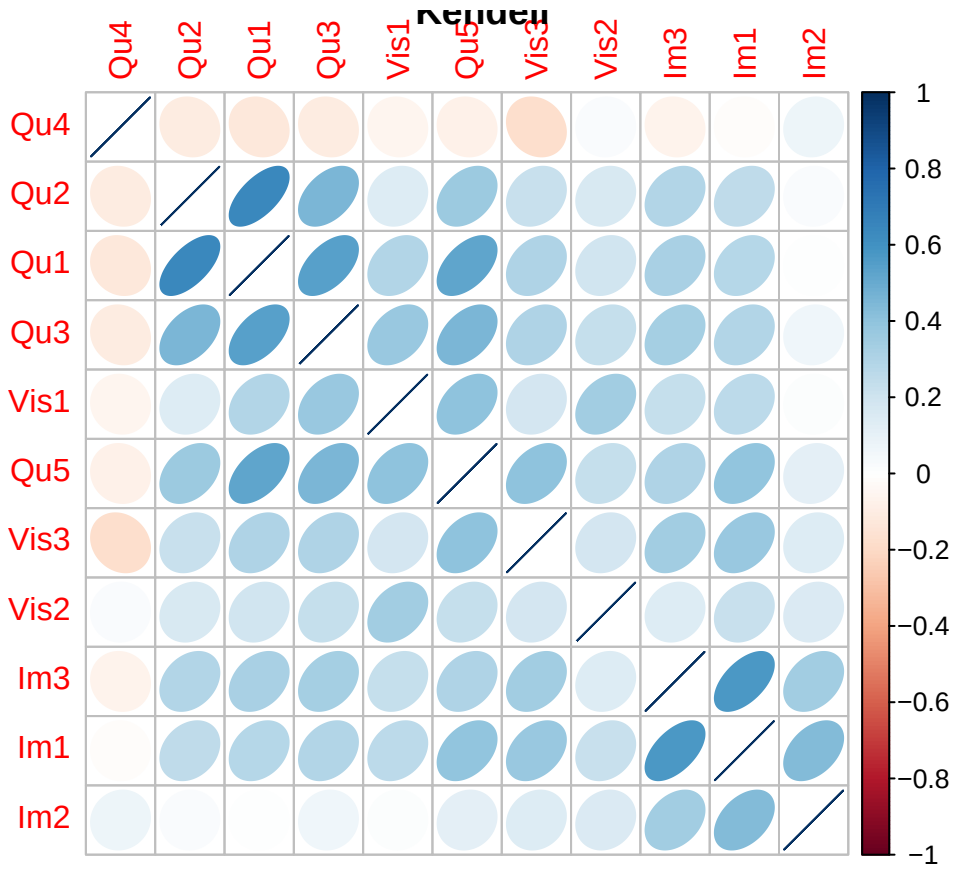
```
# compute pearson matrix
p <- cor(survey, method = "pearson")
# visualize
corrplot(p, method = "ellipse", order = "AOE", title = "Pearson")
```




```
# compute spearman matrix
s <- cor(survey, method = "spearman")
# visualize
corrplot(s, method = "ellipse", order = "AOE", title = "Spearman")
```




```
# compute kendell matrix
k <- cor(survey, method = "kendall")
# visualize
corrplot(k, method = "ellipse", order = "AOE", title = "Kendell")
```



Problem 2b)

The KMO statistic measures the sampling adequacy for factor analysis by assessing the proportion of variance in the variables that may be caused by underlying factors that range from 0 to 1. For example, a KMO value of 0.82 indicates a relatively high degree of common variance among the variables, suggesting that a reasonable amount of shared information can be effectively analyzed using factor analysis techniques.

```
# KMO sample adequacy est
KMO(survey)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = survey)
## Overall MSA =  0.82
## MSA for each item =
##  Qu1  Qu2  Qu3  Qu4  Qu5 Vis1 Vis2 Vis3  Im1  Im2  Im3
## 0.82 0.79 0.91 0.75 0.88 0.76 0.77 0.89 0.79 0.65 0.83
```

Problem 2c)

Based on the scree plot, we concluded that three components are appropriate for the factor analysis. These three components explain 80.52% the variance of the data.

```
# PCA analysis using Pearson
Ppca <- prcomp(p, scale = T)

# summary
summary(Ppca)
```

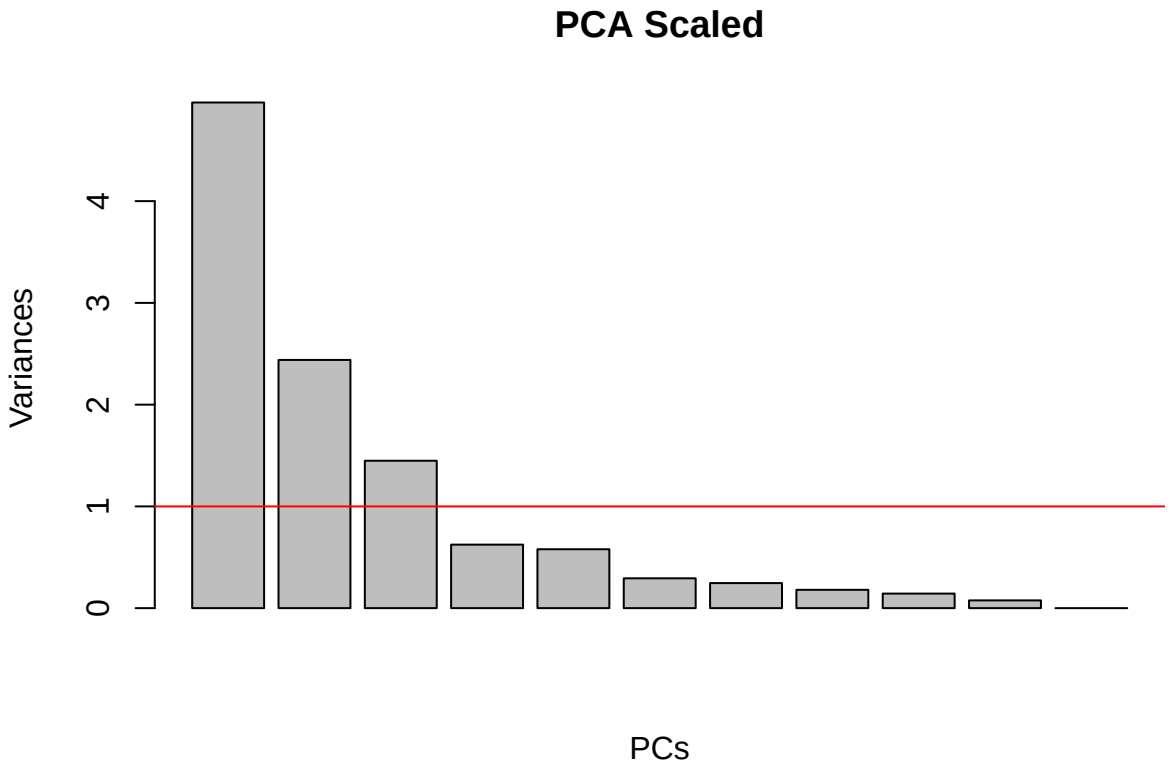
```
## Importance of components:
##                           PC1     PC2     PC3      PC4      PC5      PC6      PC7
```

```
## Standard deviation     2.2292 1.5619 1.2036 0.79016 0.76094 0.54163 0.49623
## Proportion of Variance 0.4518 0.2218 0.1317 0.05676 0.05264 0.02667 0.02239
## Cumulative Proportion  0.4518 0.6735 0.8052 0.86200 0.91464 0.94131 0.96370
##                            PC8     PC9    PC10      PC11
## Standard deviation     0.42426 0.37833 0.27609 6.092e-17
## Proportion of Variance 0.01636 0.01301 0.00693 0.000e+00
## Cumulative Proportion  0.98006 0.99307 1.00000 1.000e+00
```

```
# bar scree plot
screeplot(Ppca, npcs = 11, type = 'barplot',
          main = "PCA Scaled") + title(xlab = "PCs")
```

```
## integer(0)
```

```
abline(1, 0, col = "red")
```



Problem 2d)

A Factor analysis, to identify the latent factors or dimensions that explain the covariance structure among a set of observed variables using Spearman correlation matrix.

```
# principal spearman
sFactor <- principal(s, rotate="varimax", nfactors = 3)

# print sFactor
print(sFactor)
```

```
## Principal Components Analysis
## Call: principal(r = s, nfactors = 3, rotate = "varimax")
## Standardized loadings (pattern matrix) based upon correlation matrix
##        RC1  RC2  RC3   h2   u2 com
## Qu1   0.83 0.09 0.19 0.73 0.27 1.1
## Qu2   0.75 0.10 0.06 0.58 0.42 1.0
```

```
## Qu3   0.70 0.14 0.34 0.62 0.38 1.5
## Qu4  -0.47 0.04 0.42 0.40 0.60 2.0
## Qu5   0.63 0.27 0.38 0.61 0.39 2.1
## Vis1  0.34 0.06 0.70 0.61 0.39 1.5
## Vis2  0.11 0.14 0.75 0.59 0.41 1.1
## Vis3  0.50 0.44 0.01 0.45 0.55 2.0
## Im1   0.27 0.80 0.21 0.76 0.24 1.4
## Im2  -0.19 0.83 0.08 0.72 0.28 1.1
## Im3   0.37 0.73 0.06 0.67 0.33 1.5
##
##                        RC1  RC2  RC3
## SS loadings           2.98 2.19 1.58
## Proportion Var        0.27 0.20 0.14
## Cumulative Var        0.27 0.47 0.61
## Proportion Explained  0.44 0.32 0.23
## Cumulative Proportion 0.44 0.77 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.09
##
## Fit based upon off diagonal values = 0.93
```

Problem 2e)

We have conducted both a principal factor analysis (PFA) and a confirmatory factor analysis (CFA) to evaluate the goodness of fit of a model with three factors.

From the PFA, a root mean square of the residuals (RMSR/RMSEA) of 0.09 is greater than the recommended value of 0.05. This suggests that the model needs to fit the data better regarding the root mean square error. However, the empirical chi-square value of 106.86 with a very low probability ($p < 4.3e\text{-}12$) is desirable since it is much lower than 0.1. This discrepancy between the chi-square test and the RMSEA could be due to a need for more sample size since eleven features should have at least twenty or thirty per feature to be conservative.

To further investigate the adequacy of the model, we conducted a confirmatory factor analysis (CFA). The obtained p-value of 0.673 suggests that the null hypothesis, which states that the three factors adequately explain the covariance and fail to reject the null hypothesis. Therefore, based on the CFA results, the three factors are sufficient to explain the data.

```r
# principal factor spearman
sFactorFit <- principal(s, n.obs=nrow(survey), nfactors = 3)

# print
print(sFactorFit)
```

```
## Principal Components Analysis
## Call: principal(r = s, nfactors = 3, n.obs = nrow(survey))
## Standardized loadings (pattern matrix) based upon correlation matrix
##        RC1  RC2  RC3   h2   u2 com
## Qu1   0.83 0.09 0.19 0.73 0.27 1.1
## Qu2   0.75 0.10 0.06 0.58 0.42 1.0
## Qu3   0.70 0.14 0.34 0.62 0.38 1.5
## Qu4  -0.47 0.04 0.42 0.40 0.60 2.0
## Qu5   0.63 0.27 0.38 0.61 0.39 2.1
## Vis1  0.34 0.06 0.70 0.61 0.39 1.5
```

```
## Vis2  0.11 0.14 0.75 0.59 0.41 1.1
## Vis3  0.50 0.44 0.01 0.45 0.55 2.0
## Im1   0.27 0.80 0.21 0.76 0.24 1.4
## Im2  -0.19 0.83 0.08 0.72 0.28 1.1
## Im3   0.37 0.73 0.06 0.67 0.33 1.5
##
##                        RC1  RC2  RC3
## SS loadings           2.98 2.19 1.58
## Proportion Var        0.27 0.20 0.14
## Cumulative Var        0.27 0.47 0.61
## Proportion Explained  0.44 0.32 0.23
## Cumulative Proportion 0.44 0.77 1.00
##
## Mean item complexity =  1.5
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.09
##  with the empirical chi square  106.86  with prob <  4.3e-12
##
## Fit based upon off diagonal values = 0.93
```

```
# factor analysis
fitFA  <-  factanal(covmat=s, rotation = "varimax", n.obs=nrow(survey), factors=3)

# print
print(fitFA)
```

```
##
## Call:
## factanal(factors = 3, covmat = s, n.obs = nrow(survey), rotation = "varimax")
##
## Uniquenesses:
##   Qu1   Qu2   Qu3   Qu4   Qu5  Vis1  Vis2  Vis3   Im1   Im2   Im3
## 0.196 0.409 0.500 0.966 0.475 0.167 0.792 0.716 0.235 0.583 0.466
##
## Loadings:
##      Factor1 Factor2 Factor3
## Qu1   0.861   0.149   0.202
## Qu2   0.749   0.169
## Qu3   0.583   0.201   0.346
## Qu4  -0.180
## Qu5   0.516   0.312   0.402
## Vis1  0.164           0.895
## Vis2  0.140   0.177   0.396
## Vis3  0.315   0.394   0.171
## Im1   0.184   0.820   0.242
## Im2  -0.112   0.636
## Im3   0.275   0.650   0.189
##
##                Factor1 Factor2 Factor3
## SS loadings      2.209   1.880   1.406
## Proportion Var   0.201   0.171   0.128
## Cumulative Var   0.201   0.372   0.500
##
## Test of the hypothesis that 3 factors are sufficient.
```

```
## The chi square statistic is 21.36 on 25 degrees of freedom.
## The p-value is 0.673
```

Problem 2f)

The RC1 is influenced by QU1 (Articles in Wikipedia are reliable), QU2 (Articles in Wikipedia are updated), QU3 (Articles in Wikipedia are comprehensive), QU5 (I trust in the editing system of Wikipedia), VIS3 (I cite Wikipedia in my academic papers), and QU4 (In my area of expertise, Wikipedia has a lower quality than other educational resources) are group together but QU4 is only variable going in the opposite direction. Lastly, VIS3 is influencing RC2, and QU4 is also having an influence on RC3.

On the other hand, RC2 is mainly influenced by IM1 (The use of Wikipedia is well considered among colleagues), IM2 (In academia, sharing open educational resources is appreciated), and IM3 (My colleagues use Wikipedia). Still, near the cutoff, we have VIS3. All these factors act in the same direction, and VIS3 is roughly evenly split with RC1.

Lastly, RC3 is mostly VIS1 (Wikipedia improves visibility of students' work) and VIS2 (It is easy to have a record of the contributions made in Wikipedia), but there is QU4 near the cutoff. So all these factors act in the same direction, and QU4 is closely split with RC1 but heading in the opposite direction.

In conclusion, RC1 is related chiefly to the usability of Wikipedia, RC2 is more in tune with users, especially in academia, and RC3 is mainly associated with the recognition achieved.

```r
# print loadings
print(sFactorFit$loadings, cutoff = 0.4, sort = T, order = T)
```

```
##
## Loadings:
##      RC1    RC2    RC3
## Qu1   0.829
## Qu2   0.751
## Qu3   0.696
## Qu5   0.629
## Vis3  0.501  0.442
## Im1          0.799
## Im2          0.825
## Im3          0.730
## Vis1                0.704
## Vis2                0.745
## Qu4  -0.471         0.421
##
##                  RC1   RC2   RC3
## SS loadings    2.977 2.185 1.582
## Proportion Var 0.271 0.199 0.144
## Cumulative Var 0.271 0.469 0.613
```

Problem 2g)

There are many consistencies in the associations between variables, and there are also slight differences in the specific loadings and the variance explained by each component. These differences may arise due to the analysis approach.

```r
# make copy from survey
dsOrdered <- survey

# ordered factors in the new dataframe
for (i in 1:(ncol(dsOrdered)))
  dsOrdered[, i] = factor(dsOrdered[, i], levels = c(1, 2, 3, 4, 5), ordered = T)
```

```r
# summary of new dataframe
summary(dsOrdered)
```

```
##  Qu1    Qu2    Qu3    Qu4    Qu5    Vis1   Vis2   Vis3   Im1    Im2    Im3
##  1: 0   1: 1   1: 2   1: 3   1: 7   1: 5   1: 4   1:50   1:16   1: 3   1:10
##  2:22   2:11   2:31   2:34   2:23   2:28   2:18   2:34   2:52   2:30   2:39
##  3:51   3:42   3:59   3:44   3:47   3:46   3:58   3:13   3:36   3:28   3:44
##  4:44   4:56   4:25   4:27   4:37   4:34   4:34   4:18   4:15   4:44   4:23
##  5: 2   5: 9   5: 2   5:11   5: 5   5: 6   5: 5   5: 4   5: 0   5:14   5: 3
```

```r
# hetcor
h <- hetcor(dsOrdered) %>% suppressWarnings()

# principal factor
hCor = principal(r = h$correlations, n.obs=nrow(survey), nfactors = 3, rotate="varimax")

# print
print(hCor)
```

```
## Principal Components Analysis
## Call: principal(r = h$correlations, nfactors = 3, rotate = "varimax",
##     n.obs = nrow(survey))
## Standardized loadings (pattern matrix) based upon correlation matrix
##        RC1  RC2  RC3   h2   u2 com
## Qu1   0.86 0.14 0.23 0.81 0.19 1.2
## Qu2   0.81 0.09 0.09 0.68 0.32 1.0
## Qu3   0.74 0.16 0.36 0.70 0.30 1.5
## Qu4  -0.42 0.19 0.16 0.24 0.76 1.7
## Qu5   0.65 0.27 0.44 0.69 0.31 2.1
## Vis1  0.22 0.10 0.82 0.74 0.26 1.2
## Vis2  0.06 0.08 0.82 0.69 0.31 1.0
## Vis3  0.57 0.37 0.12 0.48 0.52 1.8
## Im1   0.29 0.82 0.24 0.82 0.18 1.4
## Im2  -0.19 0.86 0.00 0.77 0.23 1.1
## Im3   0.37 0.77 0.10 0.74 0.26 1.5
##
##                        RC1  RC2  RC3
## SS loadings           3.17 2.32 1.85
## Proportion Var        0.29 0.21 0.17
## Cumulative Var        0.29 0.50 0.67
## Proportion Explained  0.43 0.32 0.25
## Cumulative Proportion 0.43 0.75 1.00
##
## Mean item complexity =  1.4
## Test of the hypothesis that 3 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.08
##  with the empirical chi square  84.96  with prob <  1.9e-08
##
## Fit based upon off diagonal values = 0.96
```

```r
# print loadings
print(hCor$loadings, cutoff = 0.4, sort = T, order = T)
```

```
##
```

```
## Loadings:
##      RC1    RC2    RC3
## Qu1   0.857
## Qu2   0.815
## Qu3   0.741
## Qu5   0.649         0.440
## Vis3  0.571
## Im1          0.823
## Im2          0.856
## Im3          0.772
## Vis1                0.823
## Vis2                0.822
## Qu4  -0.419
##
##                  RC1   RC2   RC3
## SS loadings    3.175 2.319 1.847
## Proportion Var 0.289 0.211 0.168
## Cumulative Var 0.289 0.499 0.667
```

Problem 3

(20 points) Perform a correspondence analysis on the stores and ages data in StoresAndAges.csv. In this file you are provided with the table for the two sets of categories. In particular perform the following

Problem 3 a)

In a mosaic plot, the width of the rectangles represents the proportion of cases within each category of one variable. In contrast, the height represents the proportion of cases within each category of the other variable. The area of each rectangle is proportional to the joint frequency or proportion of cases in each combination of categories.

```r
# import csv file
storeNage <- read.csv("StoresAndAges.csv")

# display
glimpse(storeNage)
```

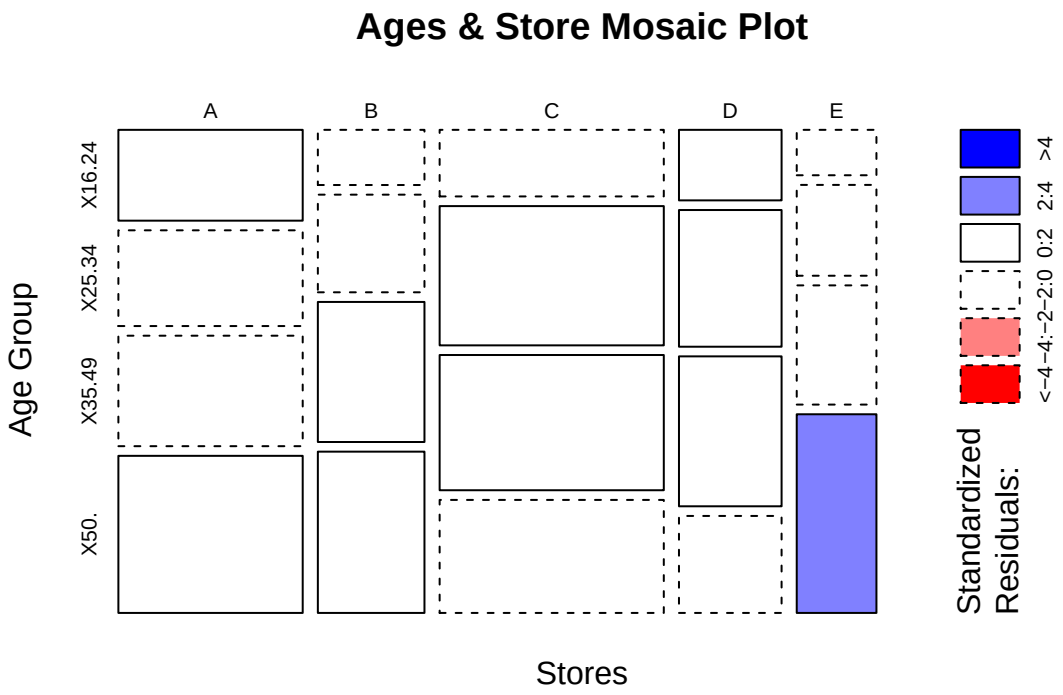
```
## Rows: 5
## Columns: 5
## $ X      <chr> "A", "B", "C", "D", "E"
## $ X16.24 <int> 37, 13, 33, 16, 8
## $ X25.34 <int> 39, 23, 69, 31, 16
## $ X35.49 <int> 45, 33, 67, 34, 21
## $ X50.   <int> 64, 38, 56, 22, 35
```

```r
# use names and column names instead of an index number
row.names(storeNage) <-  storeNage$X

# remove first column
storeNage <- storeNage[,-1]

# mosaic plot
mosaicplot(storeNage, main = "Ages & Store Mosaic Plot",
           # sub = "Product Colors by Stores",
           xlab = "Stores",
           ylab = "Age Group",
           # border = "chocolate",
```

```
          shade = TRUE)
```



Problem 3 b)

Visualization of the associations between categories using a biplot from the correspondence analysis. This biplot helps gain insights into the relationships, patterns, and dependencies among the variables by visually identifying clusters or groupings of categories.

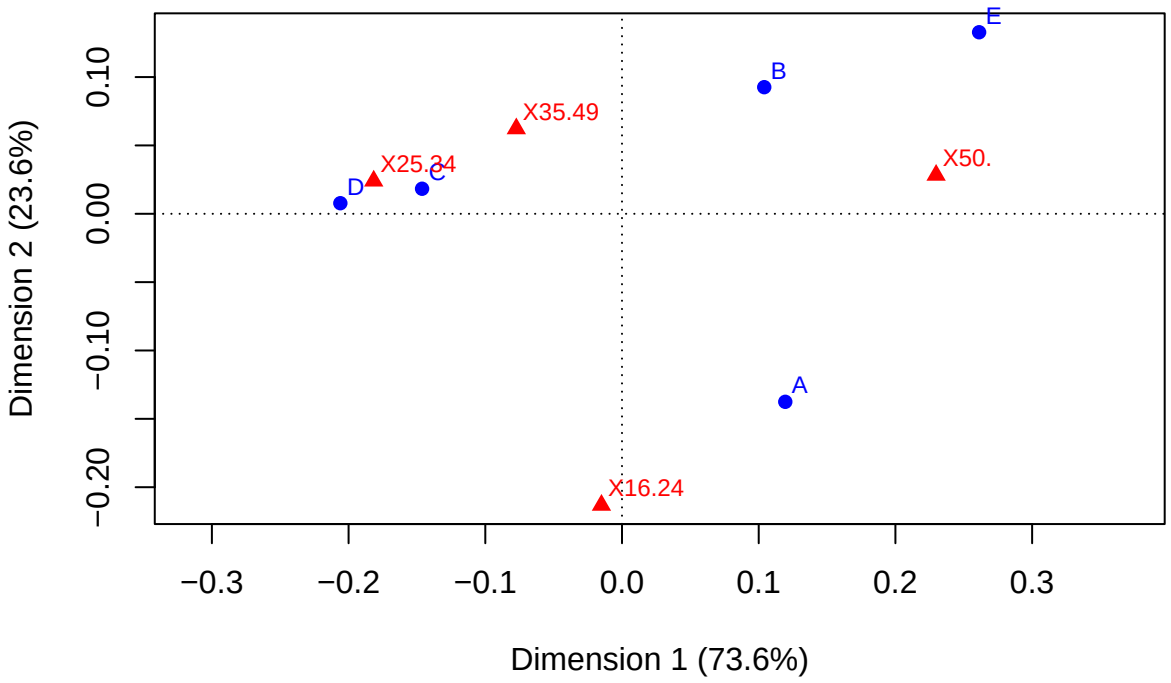
```
# correspondence analysis
fitca = ca(storeNage)

# display the fitca
fitca
```

```
##
##  Principal inertias (eigenvalues):
##            1        2        3
## Value      0.026345 0.008443 0.001008
## Percentage 73.6%    23.59%   2.82%
##
##
##  Rows:
##                 A        B         C         D        E
## Mass     0.264286 0.152857  0.321429  0.147143 0.114286
## ChiDist  0.182175 0.147723  0.150329  0.209761 0.295091
## Inertia  0.008771 0.003336  0.007264  0.006474 0.009952
## Dim. 1   0.735861 0.641045 -0.901031 -1.269390 1.609413
## Dim. 2  -1.497003 1.007844  0.199213  0.084220 1.445107
##
##
##  Columns:
##            X16.24    X25.34    X35.49     X50.
## Mass     0.152857  0.254286  0.285714 0.307143
```

```
## ChiDist  0.214019  0.187589  0.108086 0.231842
## Inertia  0.007002  0.008948  0.003338 0.016509
## Dim. 1  -0.092617 -1.118973 -0.476316 1.415583
## Dim. 2  -2.319500  0.261412  0.677569 0.307634
```

```
# plot
plot(fitca)
```



Problem 3 c)

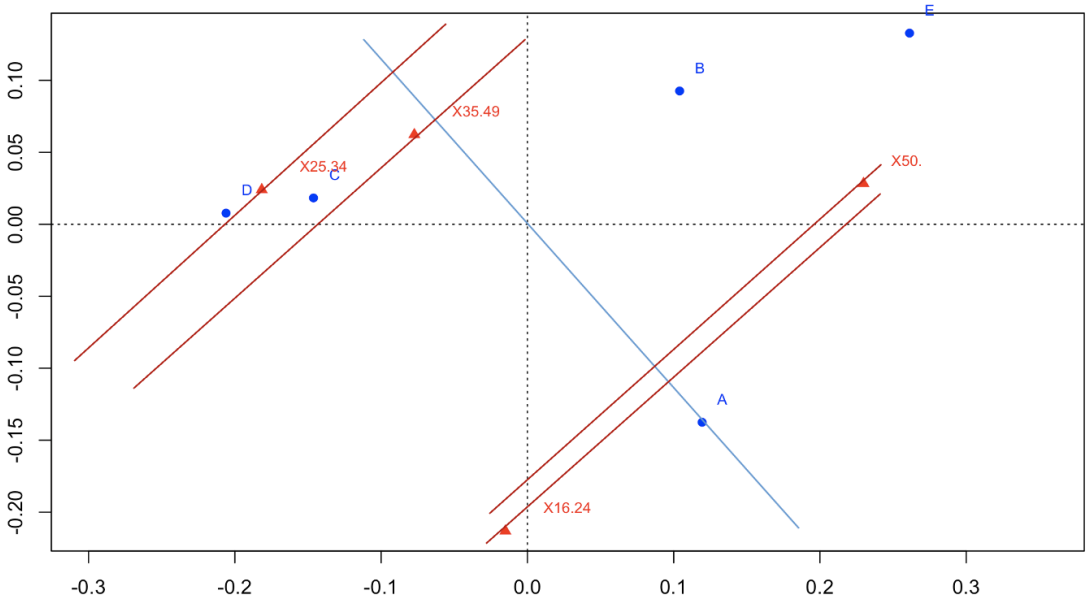
`Store A`: The most highly represented features are `X16.24` and `X50.`, indicating that this store caters to parents shopping for school incldluing themselves and caters to both age groups. The least highly represented feature is `X25.34`.

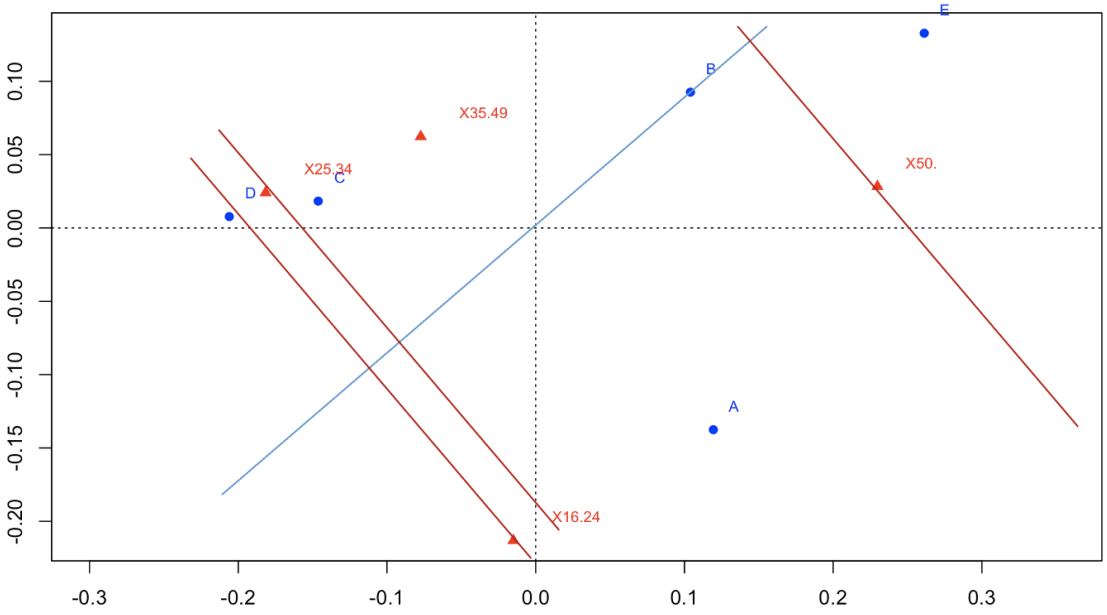
`Store B`: The most highly represented feature is `X50.`, while the least highly represented feature is `X16.24`. This suggests that the store is geared specifically towards the older age group represented by `X50.`.

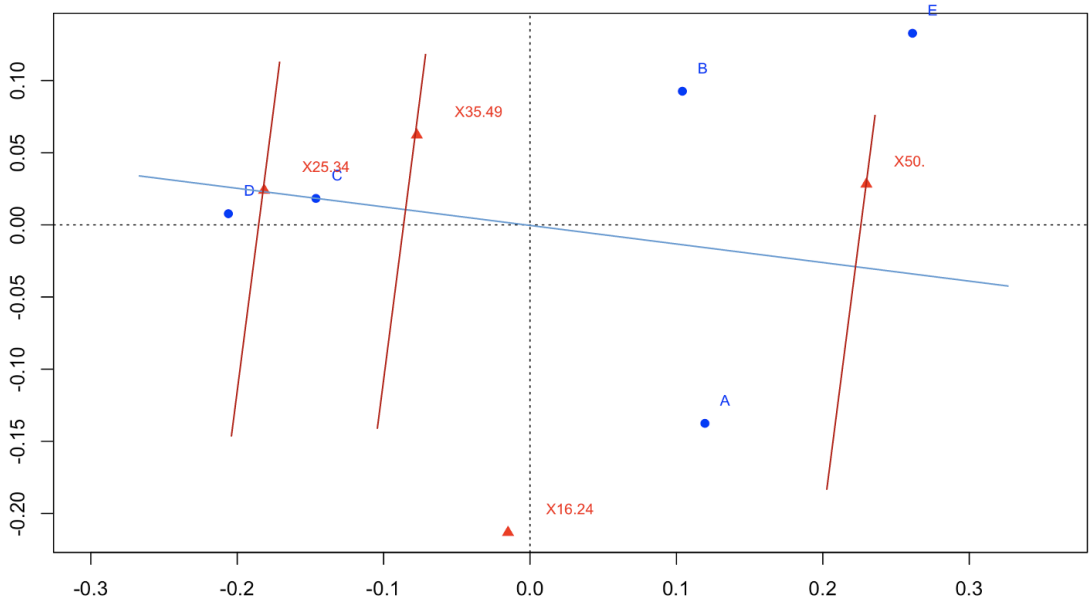
`Store C`: The most highly represented features are `X25.34` and `X35.49`, indicating that this store caters to young and middle-aged adult groups. The least highly represented feature is `X50.`.

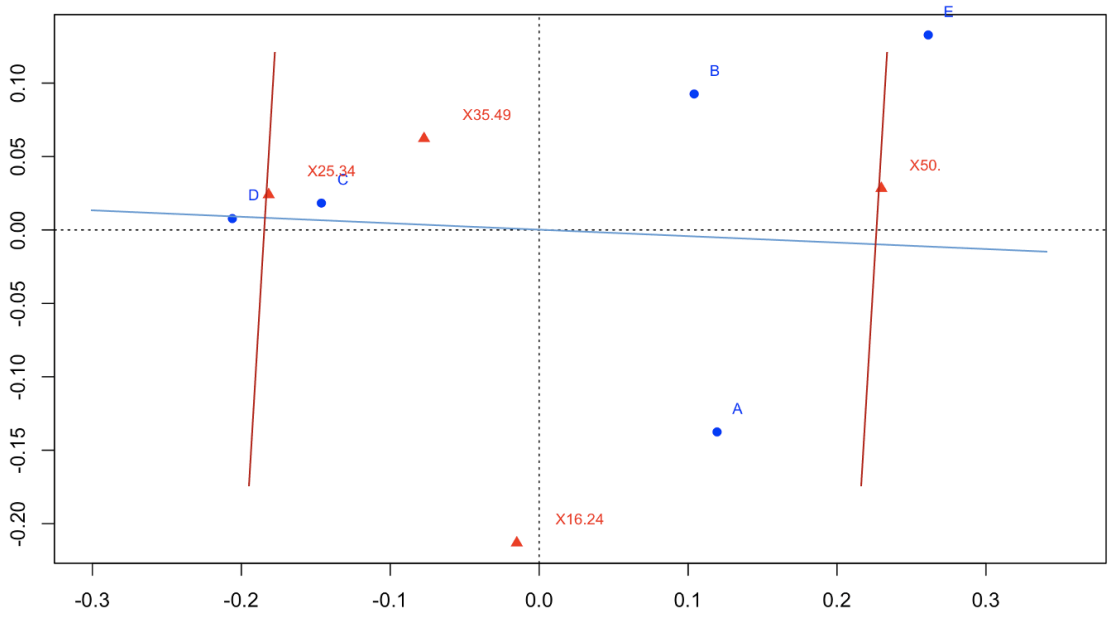
`Store D`: The most highly represented feature is `X25.34`, and the least highly represented feature is `X50.`. This suggests that the store primarily caters to a younger adult group, potentially targeting starting and mid-level age groups.

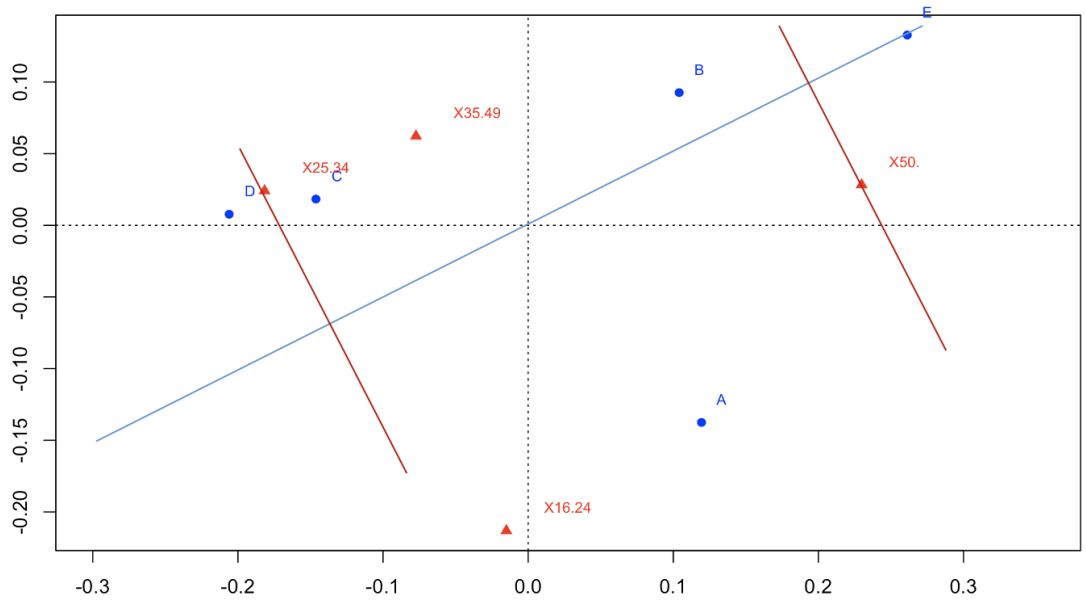
`Store E`: The most highly represented feature is `X50.`, while the least highly represented feature is `X25.34`. This indicates that the store is geared specifically towards the older age group represented by `X50.`.


E
B
X35.49
X25.34
D
C
X50.
A
X16.24
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3


E
B
X35.49
X25.34
C
D
X50.
A
X16.24
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3


E
B
X35.49
X25.34
C
D
X50.
A
X16.24
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3


E
B
X35.49
X25.34
C
D
X50.
A
X16.24
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3



Problem 3 d)

It seems that there are distinct shopping habits between the age group `X50.` and the rest of the dataset. Specifically, store `B` cater exclusively to the `X50.` age group. In contrast, store `A` appeals to both the `X50` and `X16.24` age groups and has a high representation of the feature `X16.24`, which can be associated with children living at home. While stores `D` cater mostly to the age groups `X25.34` and `X35.49` but stores `C` to `X35.49`.

From this information, we can infer that store `A's` customer base likely consists of parents and children who are potentially still living at home. Therefore, it appears that store `A` has positioned itself to cater specifically to the needs and preferences of this demographic, possibly offering products or services that are popular among parents and children still living at home.

On the other hand, store `B` caters to the age group `X50.` exclusively, suggesting that its customer base is well-established professionally and possibly has no children living in the household.

However, store `D` caters to the age groups `X25.34` and `X35.49`, suggesting that its customer base includes professionally active individuals in their prime years.

Lastly, store `C` targets explicitly the `X35.49` age group, further emphasizing its focus on customers who are professionally established and likely in their prime working years.

These observations indicate that each store has strategically positioned itself to cater to specific demographics based on age groups and corresponding shopping or spending habits.

Problem 3 e)

The first two eigenvectors in Correspondence Analysis explain a significant portion (97.2%) of the inertia or variability in the categorical data. These two eigenvectors are sufficient to capture more than 80% of the total inertia. Visualizing the data in a two-dimensional plot is a common and effective way to explore and interpret the relationships between categories in Correspondence Analysis.

```
# correspondence analysis summary
summary(fitca)
```

```
##
## 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 | X162 |  153  997  196 |  -15   5   1 | -213 992 822 |
## 2 | X253 |  254  954  250 | -182 937 318 |   24  16  17 |
## 3 | X354 |  286  843   93 |  -77 512  65 |   62 332 131 |
## 4 |  X50 |  307  997  461 |  230 982 615 |   28  15  29 |
```

Problem 4

(20 points): A common application of Discriminant Analysis is the classification of bonds into various bond rating classes. These ratings are intended to reflect the risk of the bond and influence the cost of borrowing for companies that issue bonds. Various financial ratios culled from annual reports are often used to help determine a company's bond rating. The Excel spreadsheet BondRating.xls (XLS) contains two sheets named Training data and Validation data. These are data from a sample of 95 companies selected from COMPUSTAT financial data tapes. The company bonds have been classified by Moody's Bond Ratings (1980) into seven classes of risk ranging from AAA, the safest, to C, the most risky. The data include ten financial variables for each company.

These are:

LOPMAR: Logarithm of the operating margin,

LFIXMAR: Logarithm of the pretax fixed charge coverage

LTDCAP: Long-term debt to capitalization

LGERRAT: Logarithm of total long-term debt to total equity

LLEVER: Logarithm of the leverage

LCASHLTD: Logarithm of the cash flow to long-term debt

LACIDRAT: Logarithm of the acid test ratio

LCURRAT: Logarithm of the current assets to current liabilities

LRECTURN: Logarithm of the receivable turnover

LASSLTD: Logarithm of the net tangible assets to long-term debt

The data are divided into 81 observations in the Training data sheet and 14 observations in the Validation data sheet. The bond ratings have been coded into numbers in the column with the title CODERTG, with AAA coded as 1, AA as 2, etc. Develop a Linear Discriminant Analysis model to classify the bonds in the Validation data sheet.

Problem 4 a)

The training dataset `Accuracy` is 61.72%. But, we should consider `Producers accuracy`, also known as recall or sensitivity, as a metric used to evaluate the accuracy of positive predictions in a classification task that quantifies how well a model identifies positive instances correctly.

```
# import train xlx sheet to a dataframe
bondTrain <- data.frame(read_excel("BondRating.xls", skip = 2, sheet = "training"))

# import test xlx sheet to a dataframe
bondTest  <- data.frame(read_excel("BondRating.xls", skip = 2, sheet = "validation"))

# display head for train & test
head(bondTrain)
```

```
##   OBS RATING CODERTG LOPMAR LFIXCHAR LGEARRAT LTDCAP LLEVER LCASHLTD LACIDRAT
## 1   1    AAA       1 -1.663    0.749   -0.491  0.378  0.160   -1.225    0.433
## 2   2    AAA       1 -2.382    0.814    0.147  0.534  1.188   -1.552   -1.008
## 3   3    AAA       1 -1.401    2.561   -1.797  0.142 -0.531    0.496    0.314
## 4   4    AAA       1 -2.040    2.514   -1.528  0.178 -0.325    0.019    0.149
## 5   5    AAA       1 -1.360    2.432   -1.118  0.246 -0.085   -0.083    0.033
## 6   6    AAA       1 -1.687    2.891   -1.637  0.162  0.025    0.183   -0.051
##   LCURRAT LRECTURN LASSLTD
## 1   1.120    1.629   1.277
## 2   0.553    2.415   1.357
## 3   1.014    1.728   2.273
## 4   0.773    2.612   2.070
## 5   0.344    1.854   1.772
## 6   0.328    2.197   2.361
```

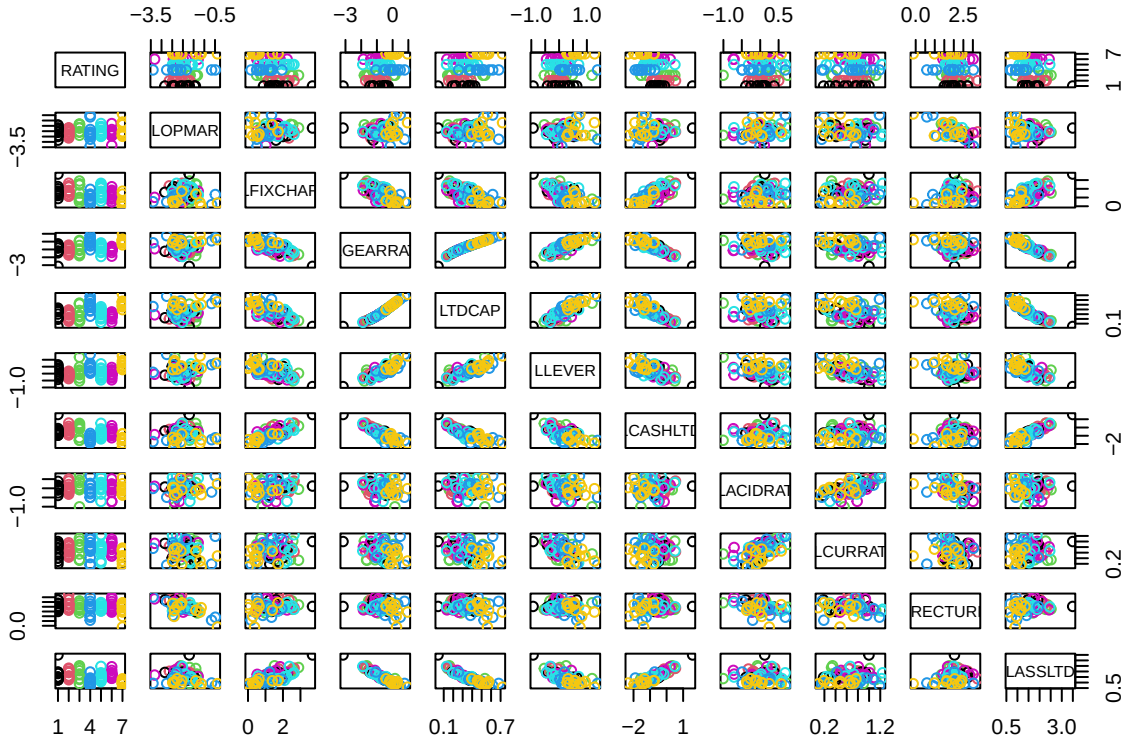
```
head(bondTest)
```

```
##   OBS RATING CODERTG LOPMAR LFIXCHAR LGEARRAT LTDCAP LLEVER LCASHLTD LACIDRAT
## 1   8    AAA       1 -1.323    0.998   -0.936  0.281 -0.042   -0.187    0.001
## 2   9    AAA       1 -2.100    1.516   -1.654  0.159  0.251    0.342   -0.077
## 3  23     AA       2 -1.743    1.626   -1.207  0.230 -0.066   -0.266   -0.229
## 4  24     AA       2 -1.776    1.153   -0.450  0.389  0.171   -0.898   -0.073
## 5  37      A       3 -1.704    3.691   -3.155  0.040 -0.936    1.573    0.122
## 6  38      A       3 -1.774    0.887   -0.532  0.369  0.013   -0.929    0.070
##   LCURRAT LRECTURN LASSLTD
## 1   0.863    1.349   1.704
## 2   0.347    1.762   2.515
## 3   0.543    1.718   1.917
## 4   0.440    2.227   1.251
## 5   0.998    2.033   3.493
## 6   0.781    1.891   1.232
```

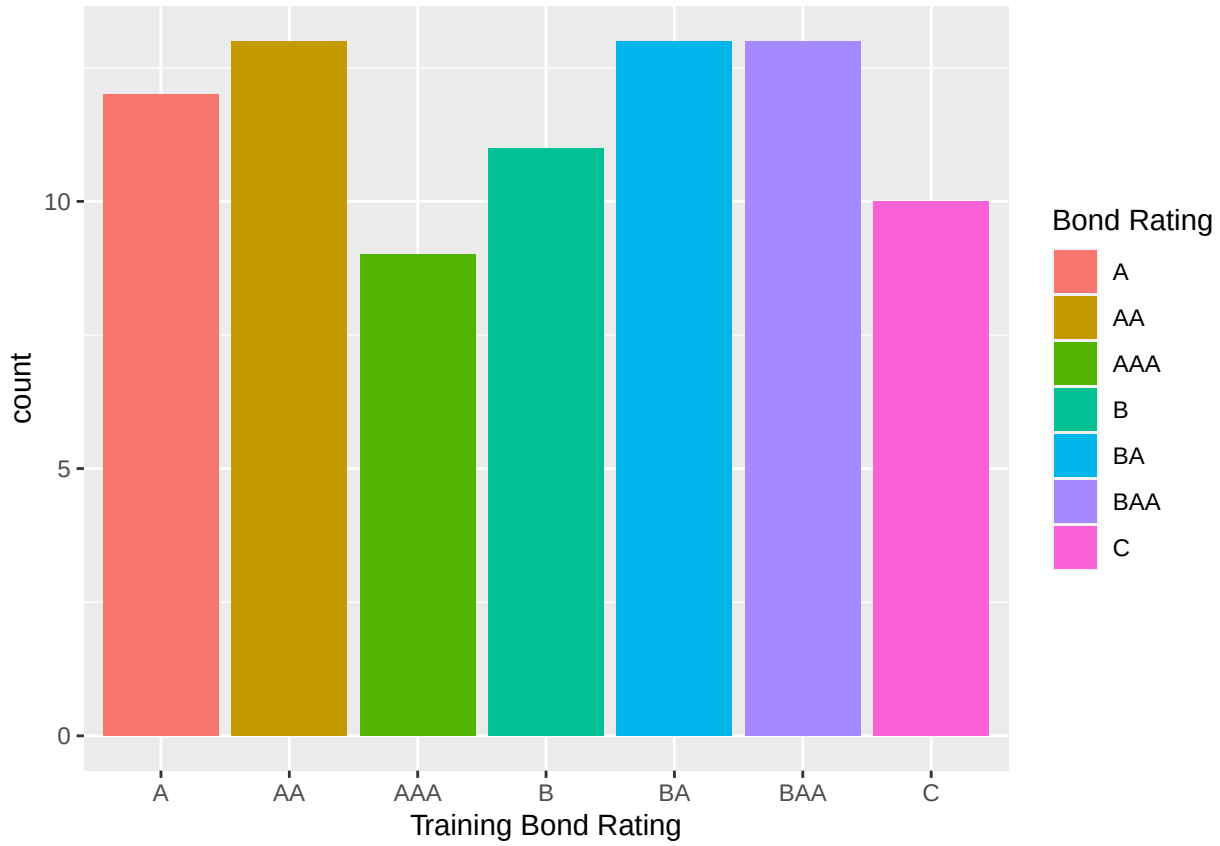
```
# remove first two columns from train & test
bondTrain <- bondTrain %>% dplyr::select(-c(OBS, CODERTG))
bondTest  <- bondTest %>% dplyr::select(-c(OBS, CODERTG))

# convert rating to factor
bondTrain$RATING <- factor(bondTrain$RATING)
bondTest$RATING  <- factor(bondTest$RATING)

# plot
plot(bondTrain, col=bondTrain$RATING)
```



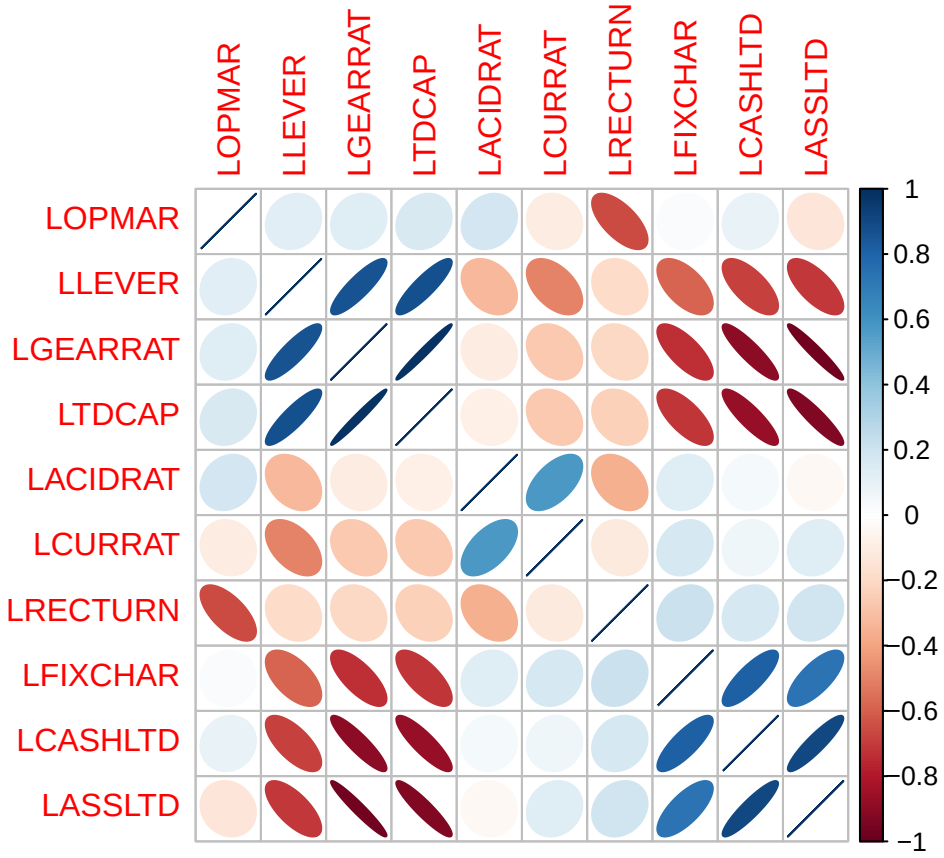

```
# bond rating distribution plot for train
ggplot(bondTrain, aes(RATING, after_stat(count), fill = RATING)) +
  geom_bar()  + labs(x = "Training Bond Rating", fill = "Bond Rating")
```




```
# summary
summary(bondTrain)
```

```
##  RATING       LOPMAR           LFIXCHAR        LGEARRAT           LTDCAP
##  A  :12   Min.   :-3.399   Min.   :0.000   Min.   :-3.1550   Min.   :0.0400
##  AA :13   1st Qu.:-2.382   1st Qu.:0.847   1st Qu.:-1.1220   1st Qu.:0.2440
##  AAA: 9   Median :-2.100   Median :1.385   Median :-0.6950   Median :0.3320
##  B  :11   Mean   :-2.004   Mean   :1.419   Mean   :-0.7003   Mean   :0.3454
##  BA :13   3rd Qu.:-1.694   3rd Qu.:1.807   3rd Qu.:-0.2870   3rd Qu.:0.4280
##  BAA:13   Max.   :-0.384   Max.   :3.691   Max.   : 0.9490   Max.   :0.7210
##  C  :10
##      LLEVER           LCASHLTD          LACIDRAT          LCURRAT
##  Min.   :-0.9360   Min.   :-2.3270   Min.   :-1.0080   Min.   :0.0880
##  1st Qu.:-0.1010   1st Qu.:-1.2420   1st Qu.:-0.1890   1st Qu.:0.4940
##  Median : 0.1280   Median :-0.7600   Median : 0.0110   Median :0.7050
##  Mean   : 0.1558   Mean   :-0.7558   Mean   : 0.0121   Mean   :0.6729
##  3rd Qu.: 0.3810   3rd Qu.:-0.3300   3rd Qu.: 0.2040   3rd Qu.:0.8460
##  Max.   : 1.3860   Max.   : 1.5730   Max.   : 0.7450   Max.   :1.2430
##
##     LRECTURN         LASSLTD
##  Min.   :-0.134   Min.   :0.582
##  1st Qu.: 1.706   1st Qu.:1.232
##  Median : 1.949   Median :1.466
##  Mean   : 1.977   Mean   :1.521
##  3rd Qu.: 2.295   3rd Qu.:1.790
##  Max.   : 3.252   Max.   :3.493
##
```

```
# correlation
corrplot(cor(bondTrain[, -1], method = "pearson"), method = "ellipse",
         order = "hclust")
```




```
# Linear Discriminant Analysis
fitLDA = lda(RATING ~ ., data = bondTrain)

# print output
print(fitLDA)
```

```
## Call:
## lda(RATING ~ ., data = bondTrain)
##
## Prior probabilities of groups:
##         A         AA        AAA         B        BA       BAA         C
## 0.1481481 0.1604938 0.1111111 0.1358025 0.1604938 0.1604938 0.1234568
##
## Group means:
##        LOPMAR  LFIXCHAR    LGEARRAT    LTDCAP      LLEVER   LCASHLTD
## A   -2.017917 1.7306667 -0.94075000 0.3034167  0.04291667 -0.4003333
## AA  -2.094385 1.8042308 -1.05315385 0.2641538 -0.08338462 -0.3925385
## AAA -1.738889 1.6637778 -0.99555556 0.2881111  0.12388889 -0.3940000
## B   -2.078545 0.9529091 -0.07790909 0.4812727  0.44972727 -1.4103636
## BA  -1.981846 1.7073077 -0.75800000 0.3272308  0.07430769 -0.7765385
## BAA -2.213923 1.3204615 -1.01200000 0.2704615 -0.02153846 -0.5720769
## C   -1.783600 0.5873000  0.10860000 0.5248000  0.64370000 -1.4720000
##         LACIDRAT   LCURRAT LRECTURN  LASSLTD
```

```
## A    0.017500000 0.6387500 2.074250 1.693417
## AA  -0.003692308 0.6640769 2.266308 1.733462
## AAA  0.059888889 0.6932222 1.943889 1.804000
## B   -0.033181818 0.7031818 1.818182 1.103182
## BA   0.137076923 0.7471538 1.950000 1.510077
## BAA -0.063230769 0.7600769 2.032077 1.721769
## C   -0.031600000 0.4642000 1.650000 0.993700
##
## Coefficients of linear discriminants:
##                  LD1        LD2        LD3          LD4           LD5
## LOPMAR     0.7720156  -2.993776  1.0902999  -1.19056396   0.003079991
## LFIXCHAR  -0.3309649  -1.032219 -2.0342609   0.17225468  -0.566130362
## LGEARRAT  -2.0228900 -13.206606 -4.3603205 -30.56370258  19.296973115
## LTDCAP   -27.6725970  15.434851 -1.0663233  30.15183168   0.636947862
## LLEVER     5.2113899   4.540020  5.2197916  13.97013291 -12.485287860
## LCASHLTD   0.8040312   3.684976  0.6103313   1.47884309   2.343115368
## LACIDRAT   0.2978150  -3.360777  0.7014467   0.09884748   0.507853522
## LCURRAT    2.0007312   2.040593  1.1419790  -1.51718949  -2.677213623
## LRECTURN   1.1369903  -2.245231  0.6432160  -0.81809242   0.686713979
## LASSLTD   -5.2328461 -14.461158 -1.3481935 -26.33072526  16.502239043
##                 LD6
## LOPMAR   -1.0907388
## LFIXCHAR  0.4446614
## LGEARRAT -8.6572293
## LTDCAP   22.5703473
## LLEVER    4.5123115
## LCASHLTD  2.1285439
## LACIDRAT -0.9383520
## LCURRAT   3.2930473
## LRECTURN -0.9182123
## LASSLTD  -5.7011832
##
## Proportion of trace:
##    LD1    LD2    LD3    LD4    LD5    LD6
## 0.6309 0.1209 0.1005 0.0705 0.0587 0.0186
```

```r
# prediction train
predTrain = predict(fitLDA, bondTrain)

# predictor information
accuracy(bondTrain$RATING, predTrain$class)
```

```
## Accuracy (PCC): 61.7283950617284%
##
## Cohen's Kappa: 0.5506
##
## Users accuracy:
##    A   AA  AAA    B   BA  BAA    C
## 60.0 46.7 57.1 80.0 61.5 57.9 85.7
##
##
## Producers accuracy:
##    A   AA  AAA    B   BA  BAA    C
## 50.0 53.8 44.4 72.7 61.5 84.6 60.0
##
```

```
##
## Confusion matrix
##      y
## x      A AA AAA  B BA BAA  C
##   A    6  3   0  0  1   2  0
##   AA   1  7   1  0  2   2  0
##   AAA  0  3   4  1  0   1  0
##   B    0  0   1  8  1   0  1
##   BA   1  1   1  0  8   2  0
##   BAA  0  1   0  0  1  11  0
##   C    2  0   0  1  0   1  6
```

Problem 4 b)

The validation/test dataset `Accuracy` is 71.43% which suggests that the model is generalizing well and performing well on unseen data since it is higher than the training dataset. However, it is equally important to assess the `Producers accuracy` to determine how well the model correctly identifies positive instances. A high `Producers accuracy` indicates that the model effectively captures positive cases and minimizes false negatives, and in this scenario, correctly identifying positive instances is of significant importance.

```
# prediction test
predTest = predict(fitLDA, bondTest)

# predictor information
accuracy(bondTest$RATING, predTest$class)
```

```
## Accuracy (PCC): 71.4285714285714%
##
## Cohen's Kappa: 0.6667
##
## Users accuracy:
##     A    AA   AAA     B    BA   BAA     C
## 100.0  66.7 100.0 100.0   NaN  40.0 100.0
##
##
## Producers accuracy:
##   A  AA AAA   B  BA BAA   C
##  50 100  50 100   0 100 100
##
##
## Confusion matrix
##      y
## x     A AA AAA B BA BAA C
##   A   1  0   0 0  0   1 0
##   AA  0  2   0 0  0   0 0
##   AAA 0  0   1 0  0   1 0
##   B   0  0   0 2  0   0 0
##   BA  0  1   0 0  0   1 0
##   BAA 0  0   0 0  0   2 0
##   C   0  0   0 0  0   0 2
```

Problem 4 c)

Yes, misclassification of bond ratings can have a substantial financial impact on investors. For example, suppose a bond is inaccurately rated higher than its actual creditworthiness. In that case, investors may be led to believe it is a safer investment than it genuinely is can result in investors unknowingly taking on higher default risk and potentially experiencing financial losses if the issuer fails to meet its obligations.

Conversely, if a bond is misclassified with a lower rating than its actual creditworthiness, investors may be deterred from investing in what could be a relatively safe bond. This can lead to missed investment opportunities, and the bond issuer may need help to attract investors. It may need to offer higher interest rates to compensate for the perceived risk, increasing its borrowing costs, and will not ensure that issuers can access capital at fair terms.

We should use multiple metrics like the F1 score combining precision and recall into a single metric, providing a balance between the two. It is calculated as the harmonic mean of precision and recall and is useful when considering false positives and negatives. Also, it can be helpful to display the ROC curve to visualize the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) by utilizing the One-vs-Rest scheme compares each class against all the others. Finally, a confusion matrix provides a more detailed analysis of the classification results. It presents the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) counts. From these values, you can calculate various performance metrics if desired.