**DSC 324/424  
Homework 4 (covers modules 5 and 7)**

**Name:** GOUTHAM SELVAKUMAR

**Due: May 17th at 11:59pm CST**

1. **Paper Review (Grads only; Undergrad extra credit)**   
   **Download the Hospital Image correspondence analysis paper and answer the following questions. You should be able to write a few sentences for each:**
2. **How are they applying CA? What variables are being analyzed and what types of categorical levels do they contain?**

* They are using the correspondence analysis in analyzing the hospital image. In the paper they defined the hospital image as the sum of beliefs, ideas and the impressions that a person holds on the hospitals which they are associated with.
* They wanted to use the correspondence analysis in understanding the relationships among the intra system competitors and hospital features and in visualizing these relationships in low dimensional space.
* The data is collected in the form of contingency table and the attributes include the hospital characteristics (or) features such as : “Cancer Treatment”, “Laser Surgery”, “Outpatient Services” etc.…, they have two levels of yes (or) no. In the paper they mentioned that rather than rating the hospitals on the evaluated scales, respondents can simply indicate with either yes (or) no.

1. **How did they use graphs from the CA in their analysis?**

* They used graphs for decision making. The graphs are useful in understanding the relationship of different hospital features and the marketers can implement these techniques for making the decisions.
* In the paper they mentioned that the ‘Lake Hospital System (LHS)’ measured the hospital image in 1986 to 1988. Initially they used the image analysis in understanding the changes occurred due to the reorganization. However, after further research the early studies were used to access the target audience perception of overall quality and preference for LHS and comparing them with their competitors. The image analysis was used to develop the communication strategies.

**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?**

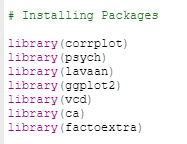
* They did not utilize any of the techniques to evaluate the goodness of fit. It is smarter to evaluate the goodness of fit and find the true positives, true negatives, false positives, and false negatives. The objective of their analysis is to group the hospitals of similar features so, it is important to evaluate the accuracy and find whether the hospitals are mapped to proper features (or) not.
* In the paper they referenced that one interviewer said: “The hospitals in Northeast Ohio have different characteristics than that mentioned in your analysis”. This demonstrated that the assessment of good fit is important for this data.

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

* The results indicate that the “Cleveland Clinic” was considered a place for the “Heart Disease Prevention and Treatment”. “Doctors who keep up with the medical advances” and “Offering community programs” account for more than half of variance.
* There are numerous implications from their conclusions. The LHS planners altered their programs and extended their market reach. They additionally developed their defense strategies to improve the product lines.
* They mentioned that corresponding analysis enabled LHS planners to visualize their hospitals advantages and disadvantages relative to competitors. It also helped them to develop a strategy to represent their system in the minds of the consumers.

**2. Correspondence Analysis**

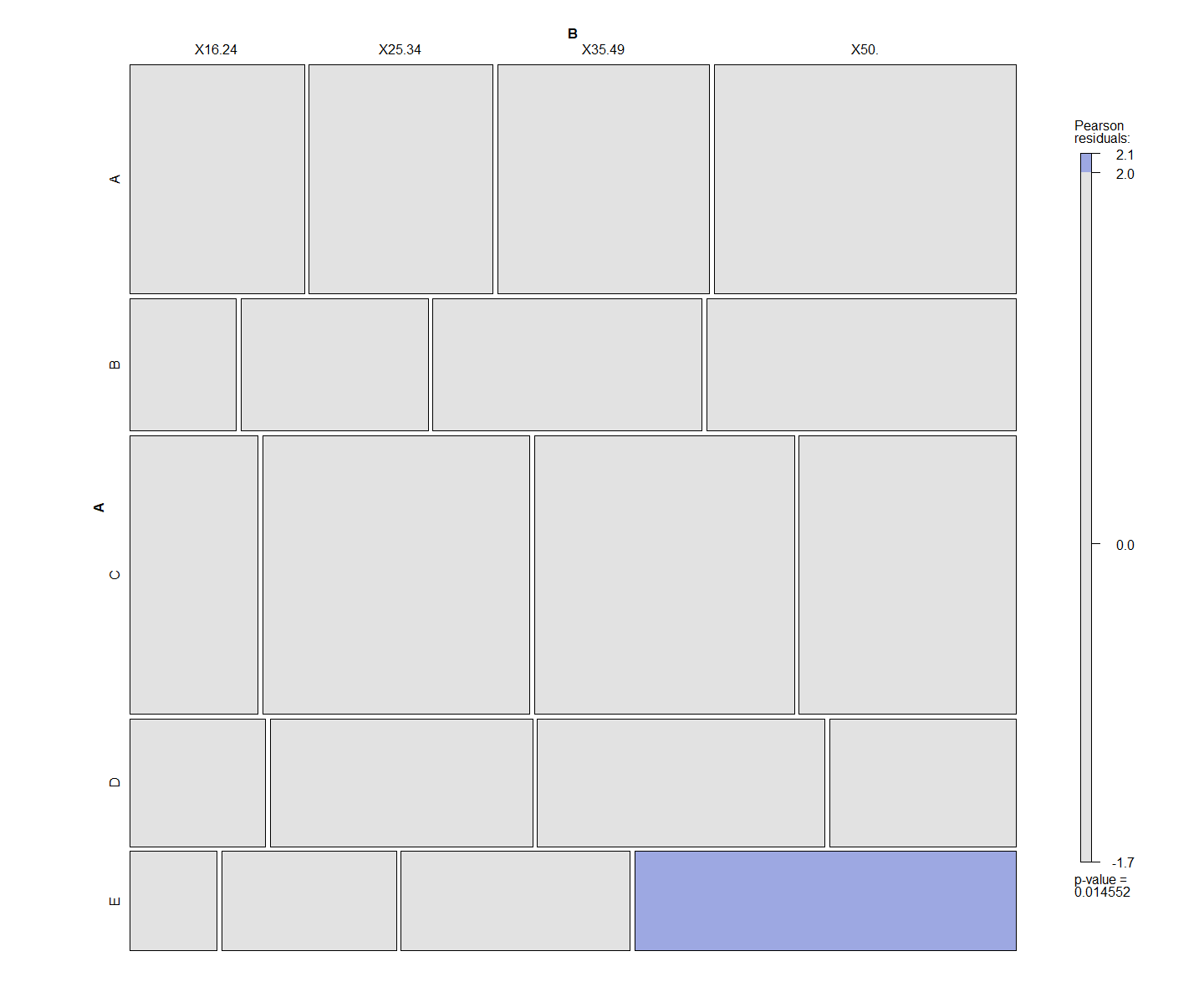
**Perform a correspondence analysis on the stores and ages data in StoresAndAges.csv. In this file you are provided with the table for the two sets of categories. Perform the following:**

****

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

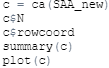
Created the mosaic plot of stores and ages. Ages are represented on X-axis and stores are represented on Y-axis. From the plot we can infer that the stores A and C have the maximum observations and store E has the minimum observations, this is indicated from the size of the plots. In store E people of age above 50 have high Pearson residuals. The R-code and the mosaic plot is shown below,

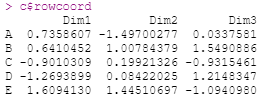




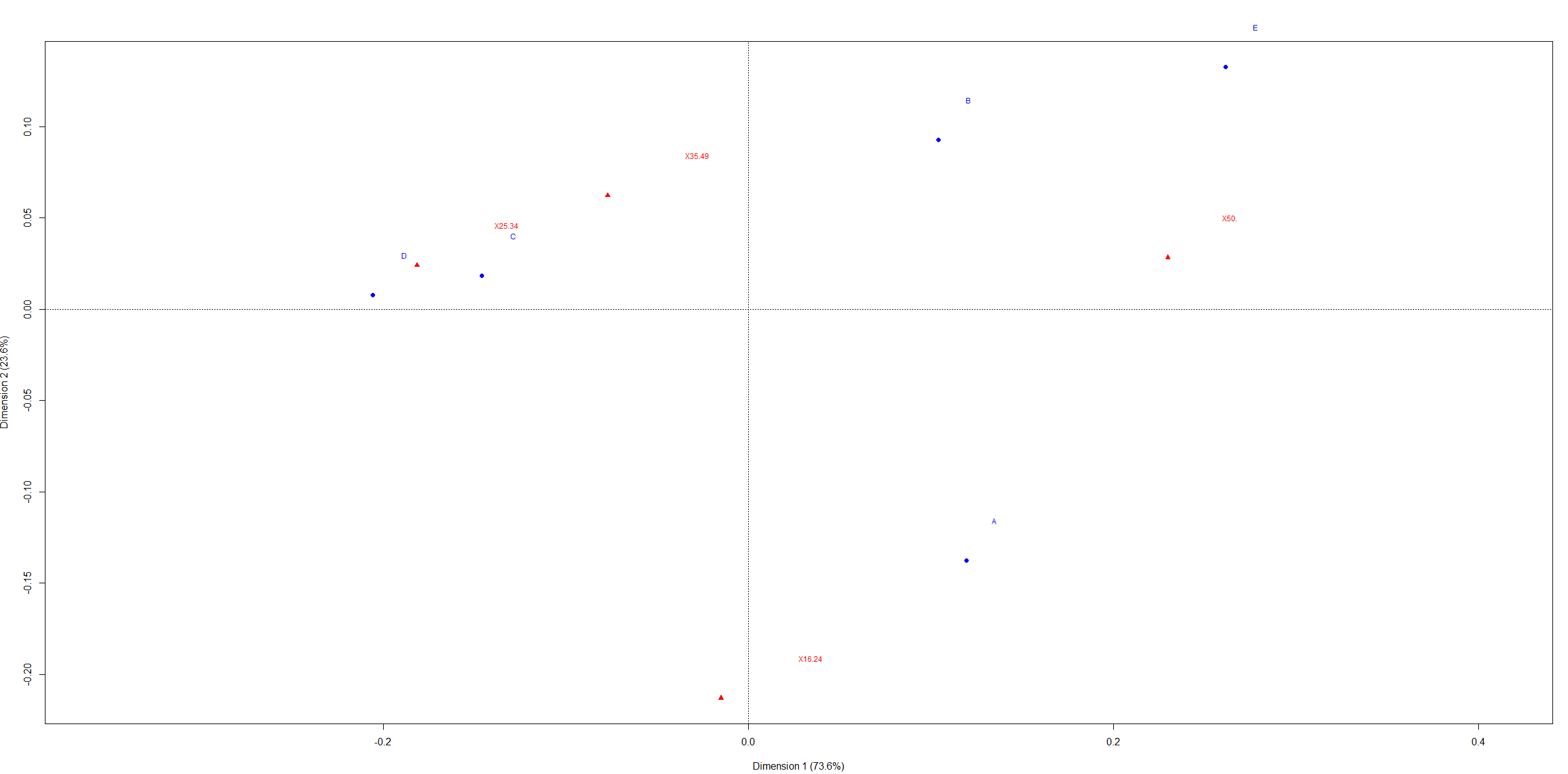
1. **Plot the correspondence analysis. Which two variables have the highest correspondence? The least?**

Performed the corresponding analysis for the data. The R-code and the output are shown below,





In Dim1 stores A, B and E have positive coordinates, C and D have negative coordinated, and we can also infer that E is far away from zero line in Dim1. In Dim2 except store A, all stores have positive coordinates. In Dim3 stores A, B and D have positive coordinates, C and E have negative coordinates. The graph representing these points are shown below,

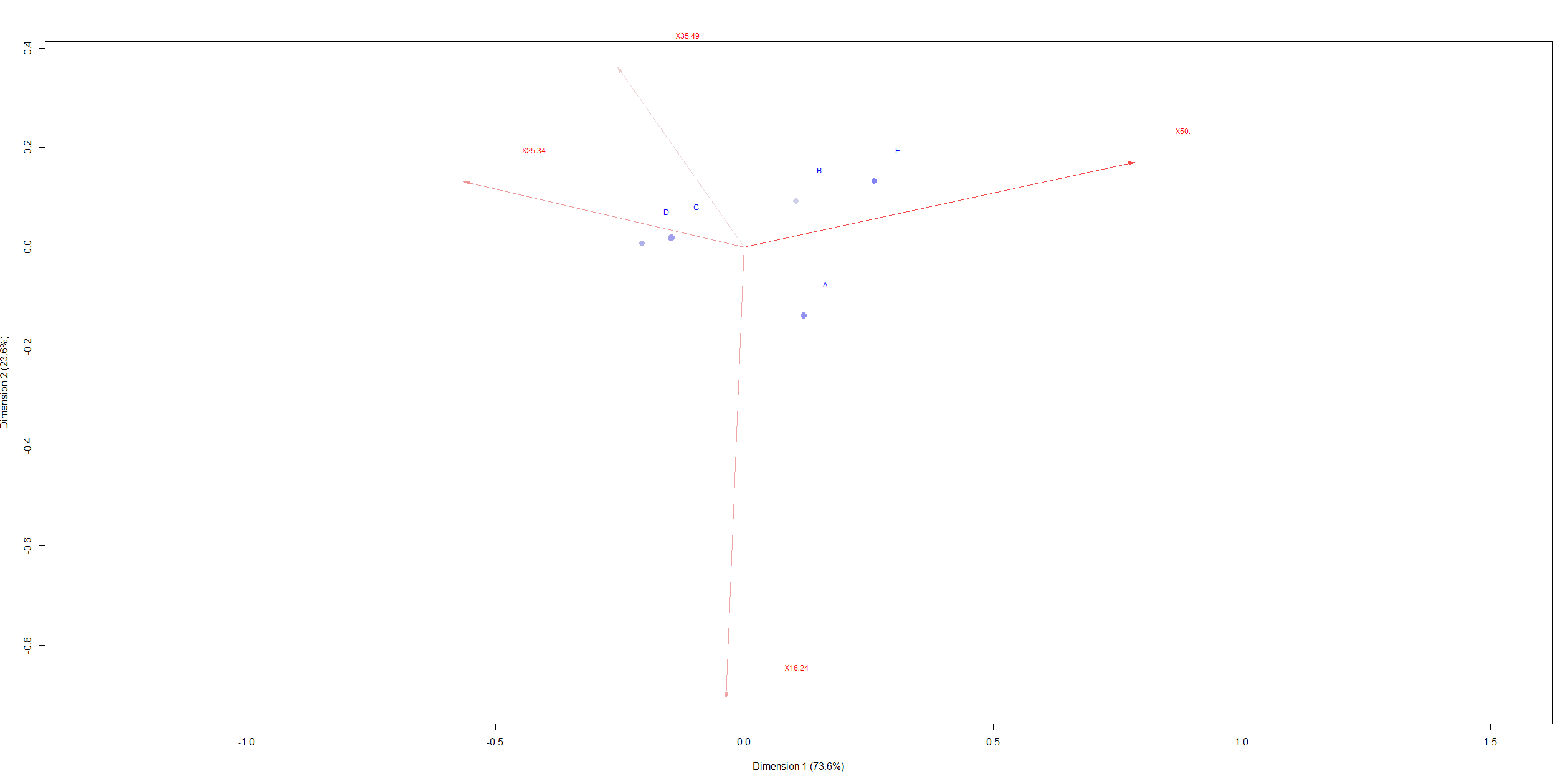


1. **With each store, create an age profile for the store. Which customer ages are most highly and least highly represented?**

For easy understanding of data I used the arrow that will compare the relative frequencies to the texts.

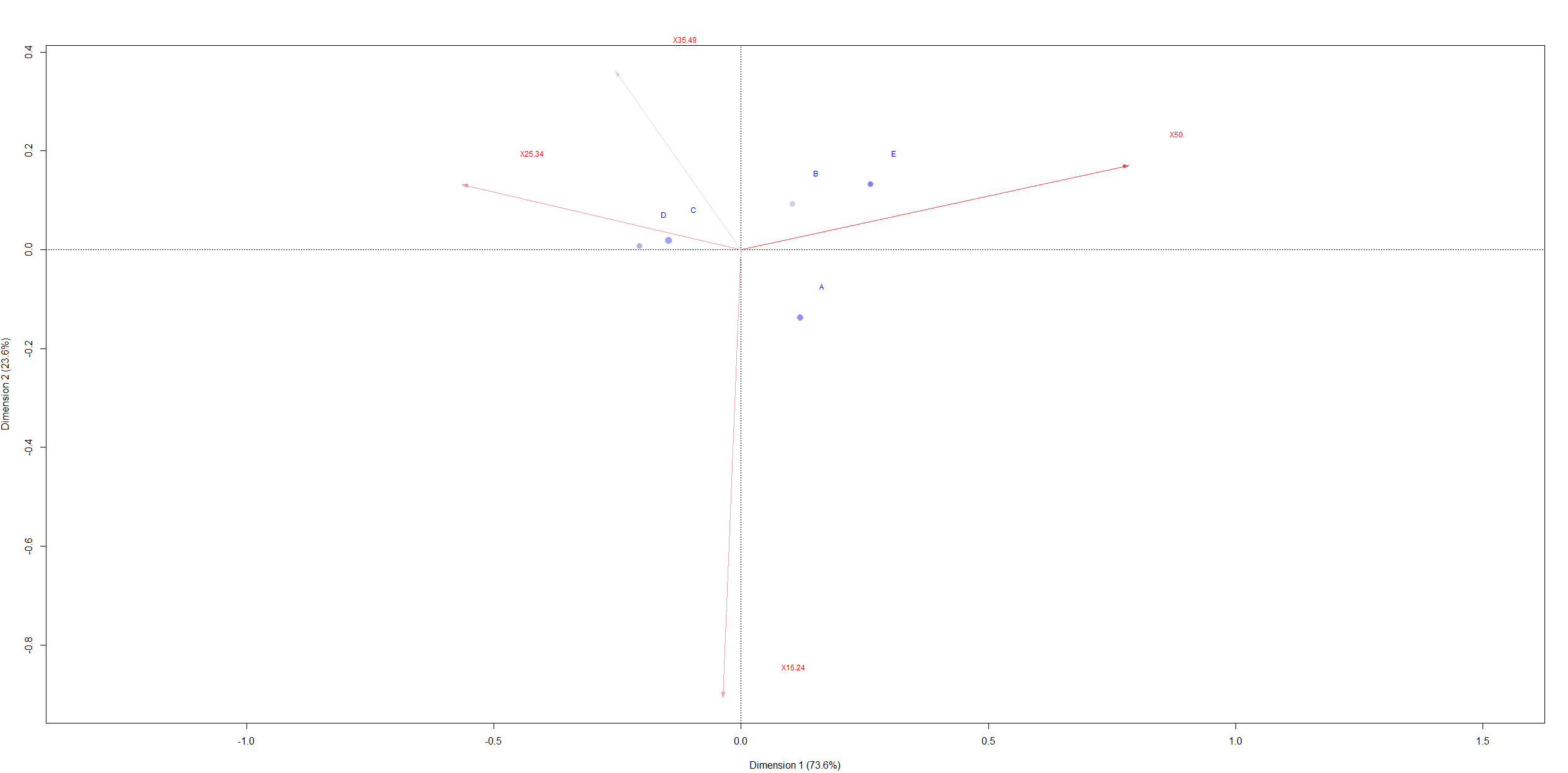
**Age profile for Store A:**

Here, acute angle corresponds to more often and obtuse angle corresponds to less often. From the graph, we can see that customer ages of 50 and 35-49 are highly represented, ages of 25-34 and 16-24 are least represented. So, the order of ages is **16.24 < 25.34 < 35.49 < 50**.



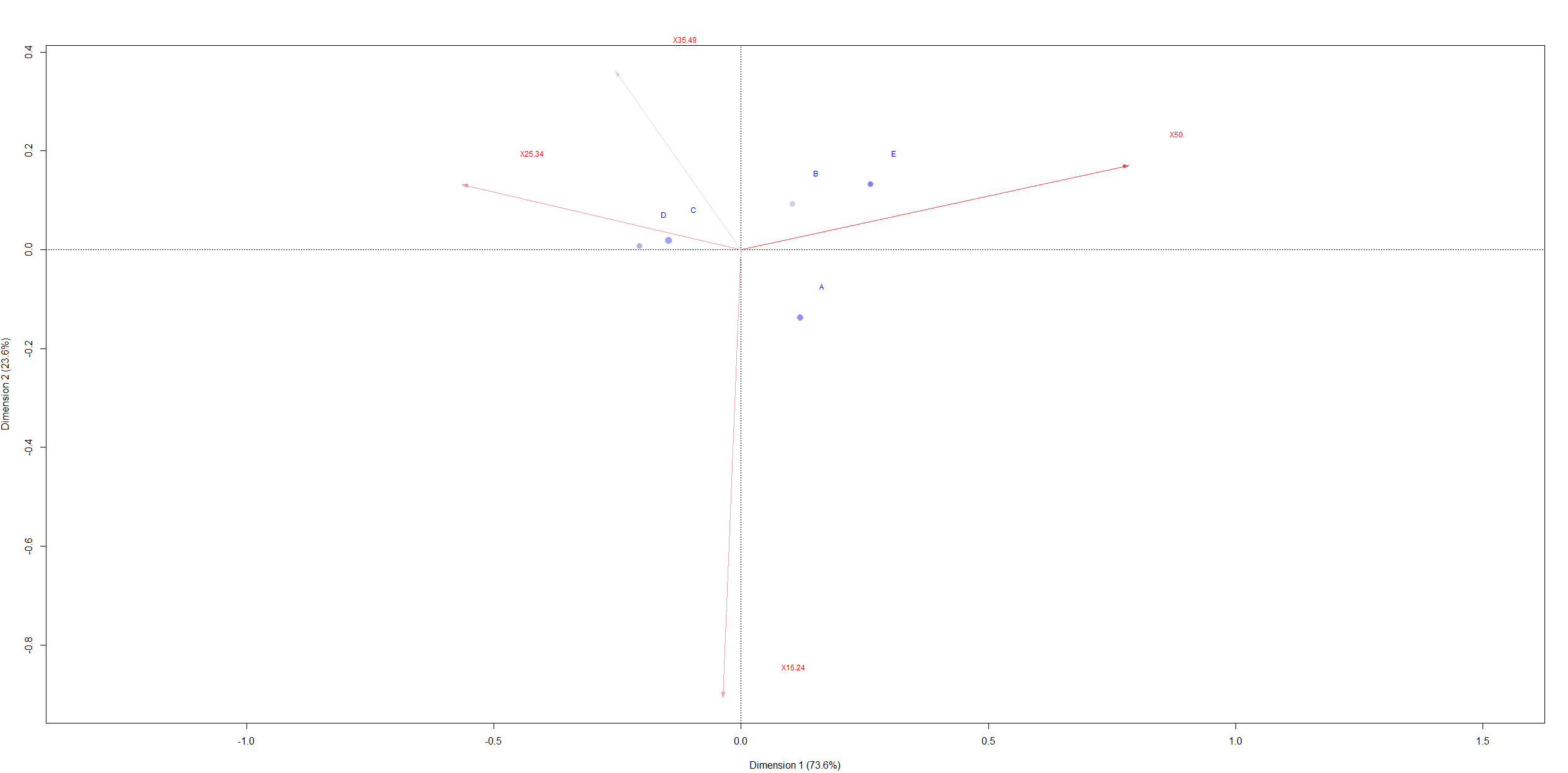
**Age profile for store B:**

For store B, customers with ages of 50+ are highly represented. Ages of 35-49 are next highly represented, ages of 16-24are least represented. The order of ages in **16.24 < 25.34 < 35.49 < 50.**



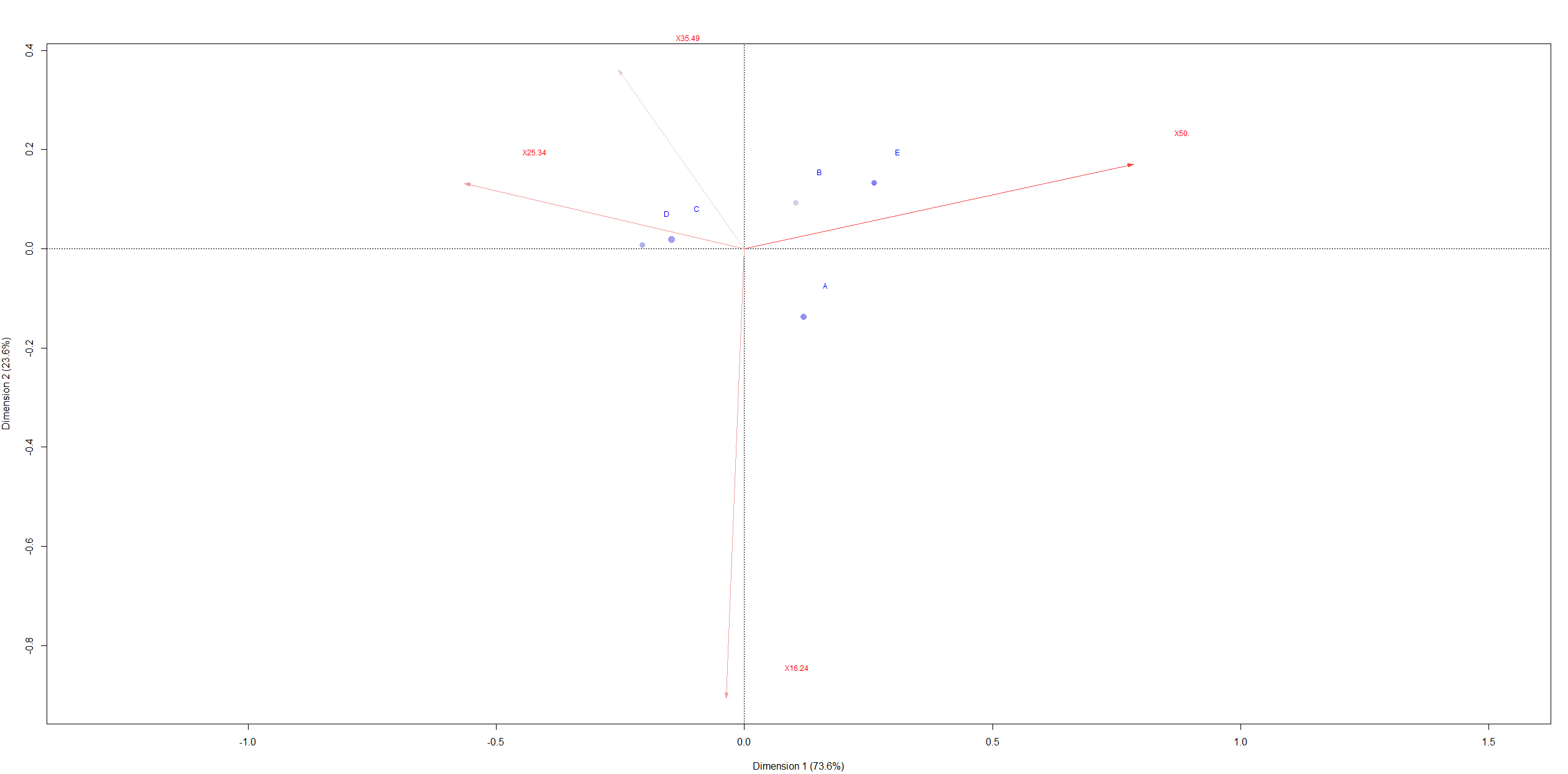
**Age profile for Store C:**

For Store C, ages of 25-34 and 35-49 are highly represented and ages of 16-24 are least represented. The order of ages is **16.24 < 50 < 35.49 < 25.34.**



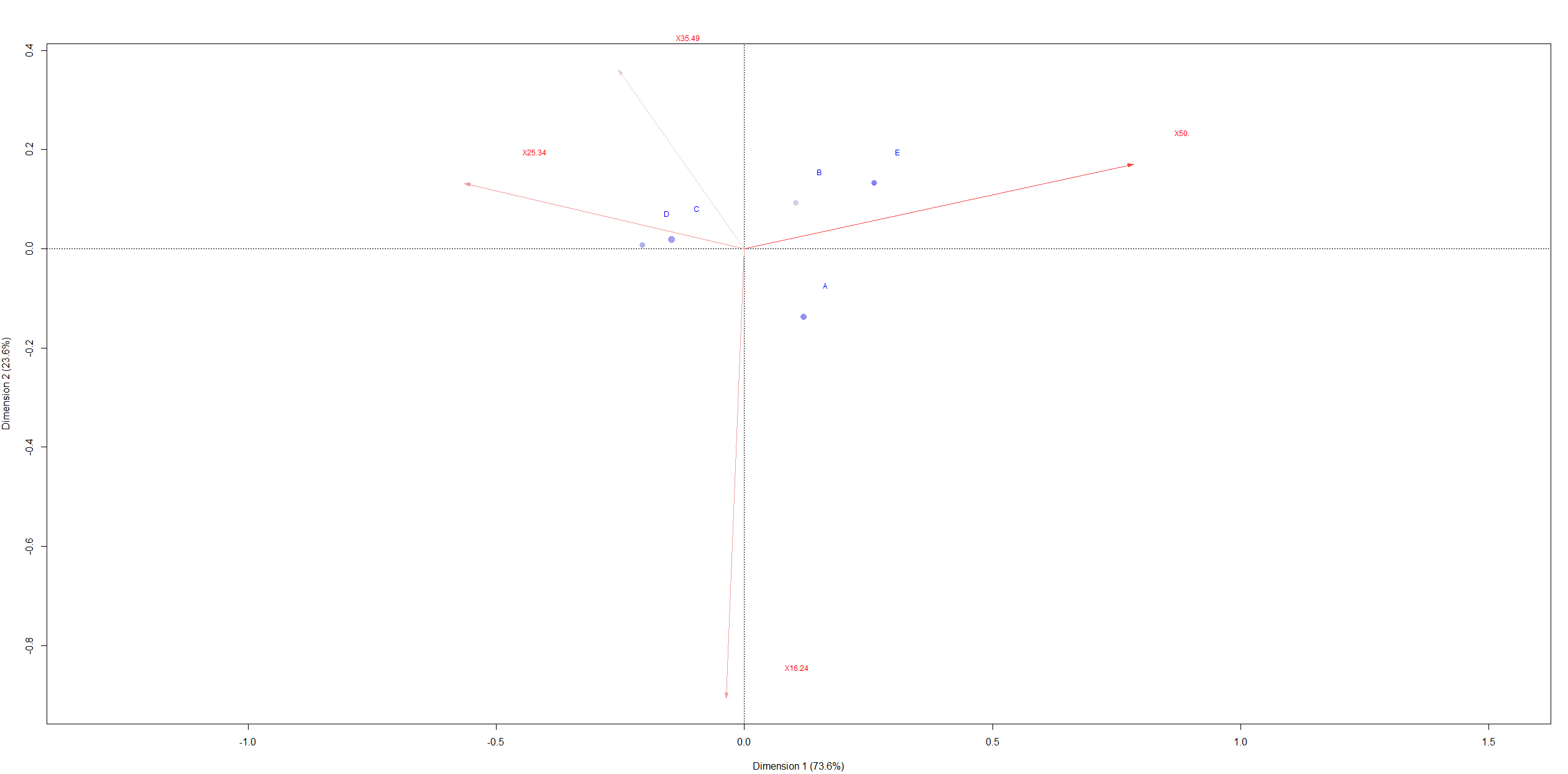
**Age profile for store D:**

For store D, ages of 25-34 and 35-39 are highly represented and ages of 50+ and 16-24 are least represented. The order of ages is **16.24 < 50 < 35.49 < 25.34.**

****

**Age profile for store E:**

For store E, ages of 50+ are highly represented and ages of 16-24 are least represented. The order of ages is **16.24 < 25.34 < 35.49 < 50.**



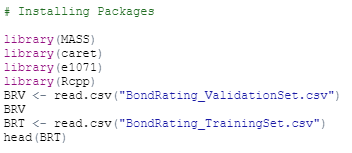
**3. Linear Discriminant Analysis**

**A common application of Linear 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.**

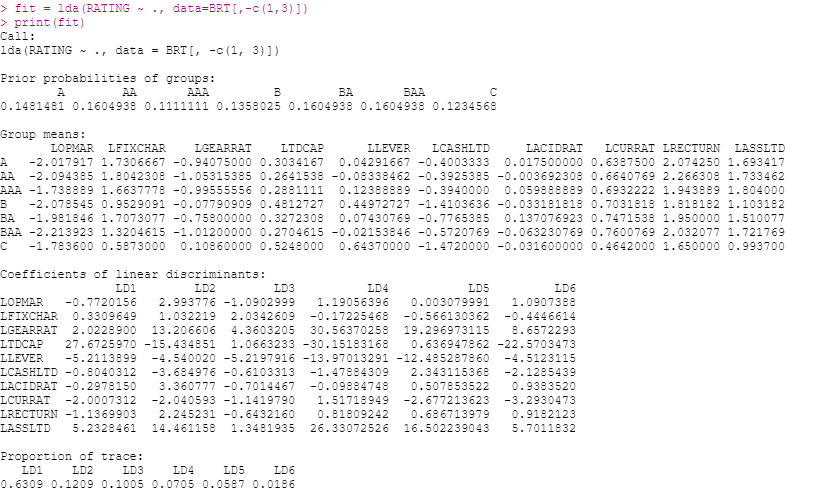
**The data are divided into 81 observations in the Training data sheet and 14 observations in the Validation data sheet. The bond ratings have been coded into numbers in the column with the title CODERTG, with AAA coded as 1, AA as 2, etc. Develop a Linear Discriminant Analysis model to classify the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

****

I splitted the BondRating.xls dataset into Training and Test Set individually and saved them. So, used one for Training Set and Test Set.

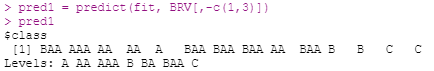
1. **What is the performance of the classifier on the training data? Notice that there is order in the class variables (i.e., AAA is better than AA, which is better than A,…).**

Performed linear discriminant analysis on training set to predict the “RATING”. Firstly, excluded the OBS and CODERTG columns and fitted the model. The R code and outputs are shown below,



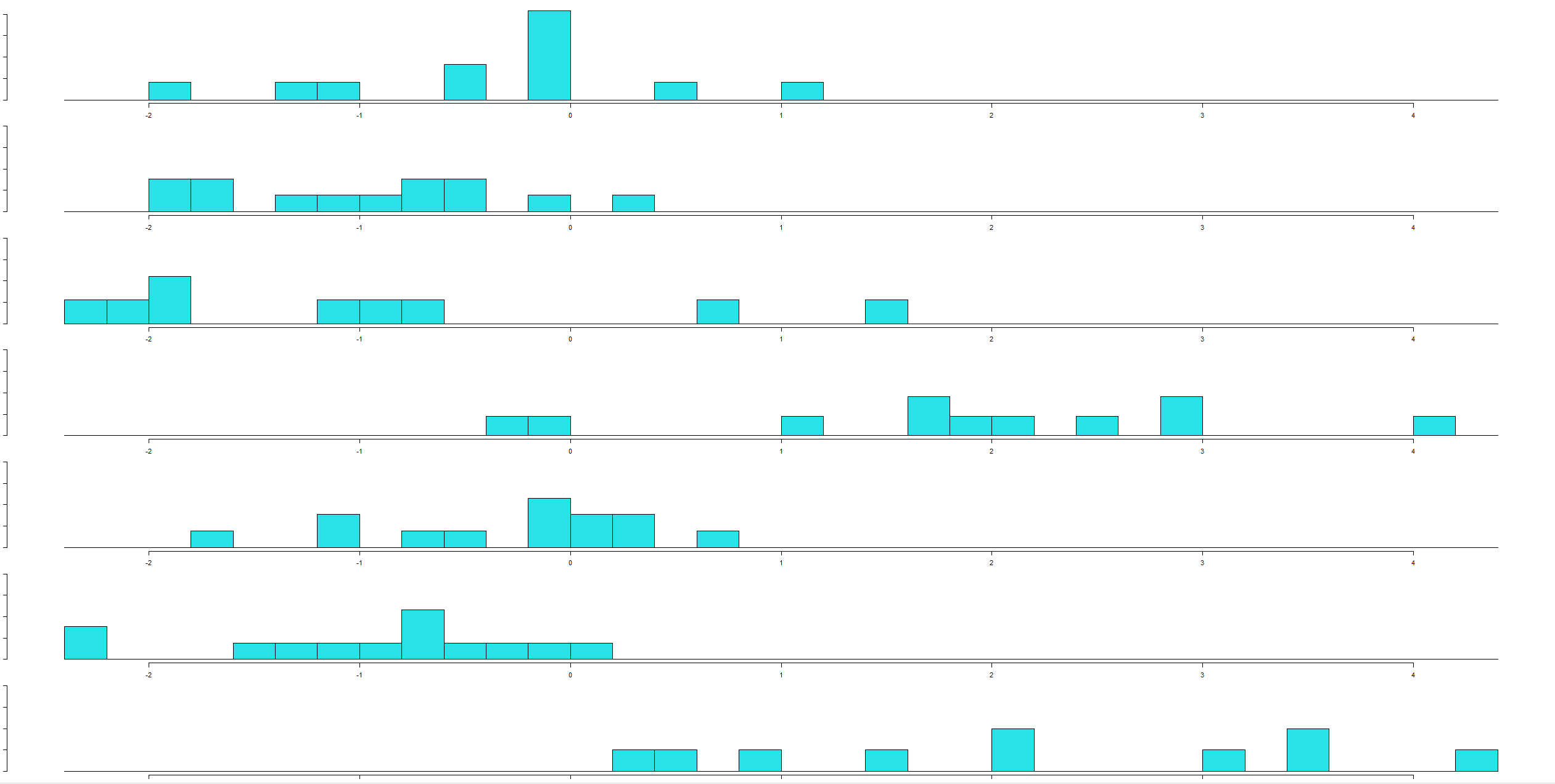
First component traced about 63% of data and second component traced about 12% of data. In LD1 LTDCAP and LLEVER has the highest and lowest coefficients. In LD2 LASSLTD and LTDCAP has the highest and lowest coefficients.

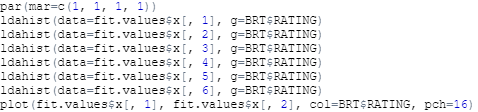
Based on the prior probabilities of groups if the observation is 0.1481 it will fall under the “A” category. If it is 0.1604 it will fall under the “AA” category, furthermore for “0.111” will fall under the “AAA” category respectively and the rest of them in their category accordingly.

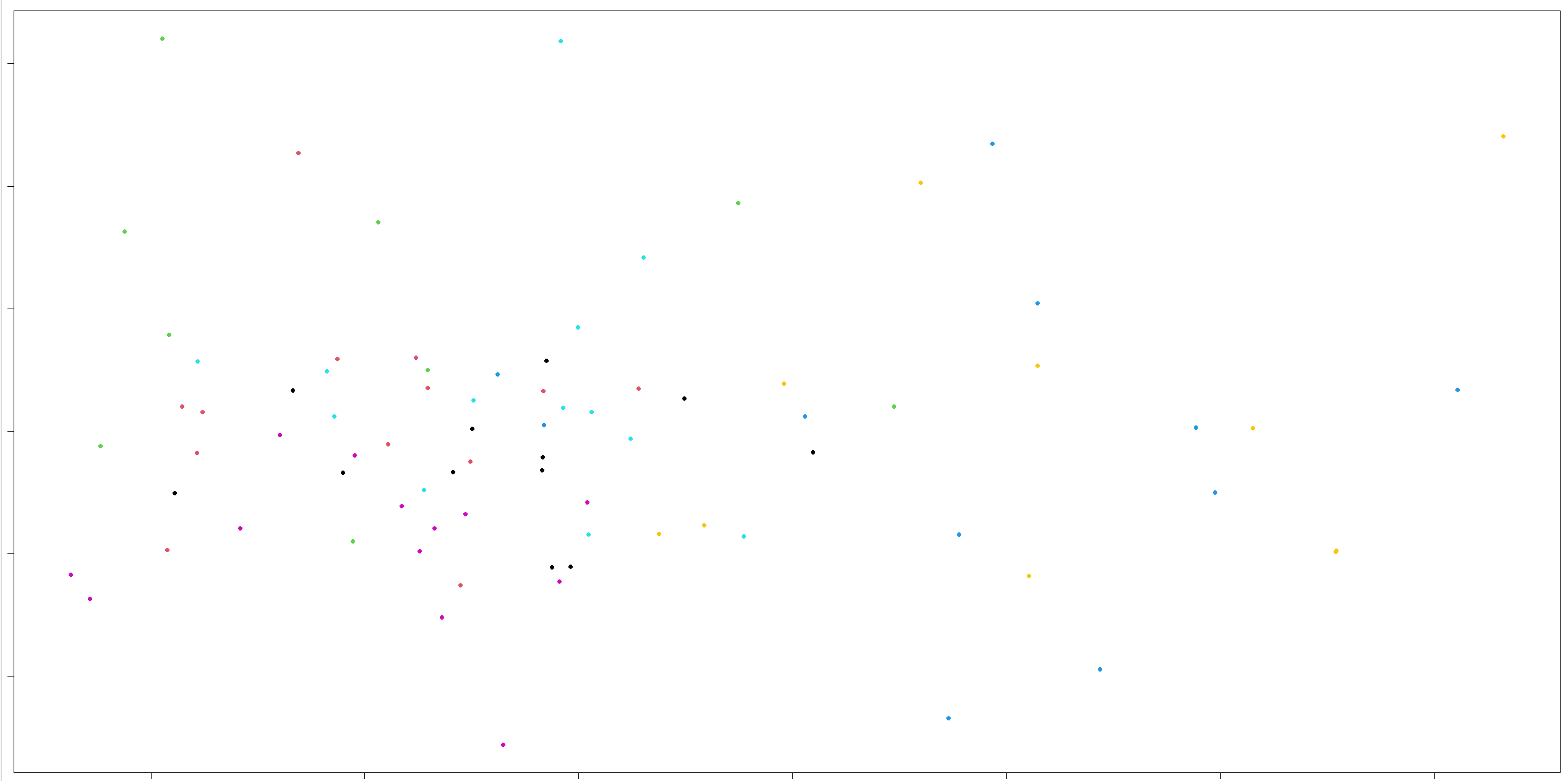




The outputs that are given above indicate that the predicted values and the observed values for “RATINGS” variable. There are seven ratings in the data so, to look at the separation we need to plot the seven histograms. Below, I will be showing the best separation that I can get. We can see that there is a separation between the last two ratings (BAA and C) and rest of them are mixed.

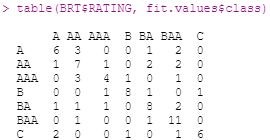


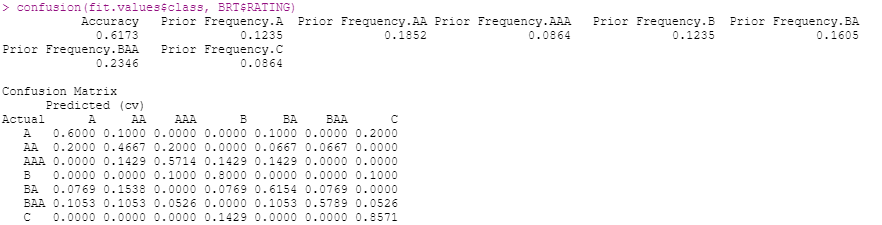




The above scatter plot also depicts that the same output as histograms. Only blue and yellow are separated from others and they are not completely separated.

Next, I computed the confusion matrix and table for the fitted values. From the table we can infer that 6 A’s are correctly predicted and one “A” is wrongly predicted as “AA” and as “BA” and two “A’s” are wrongly predicted as “C”.

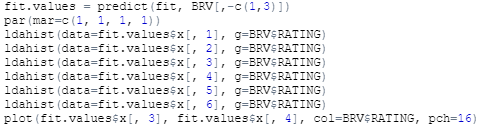


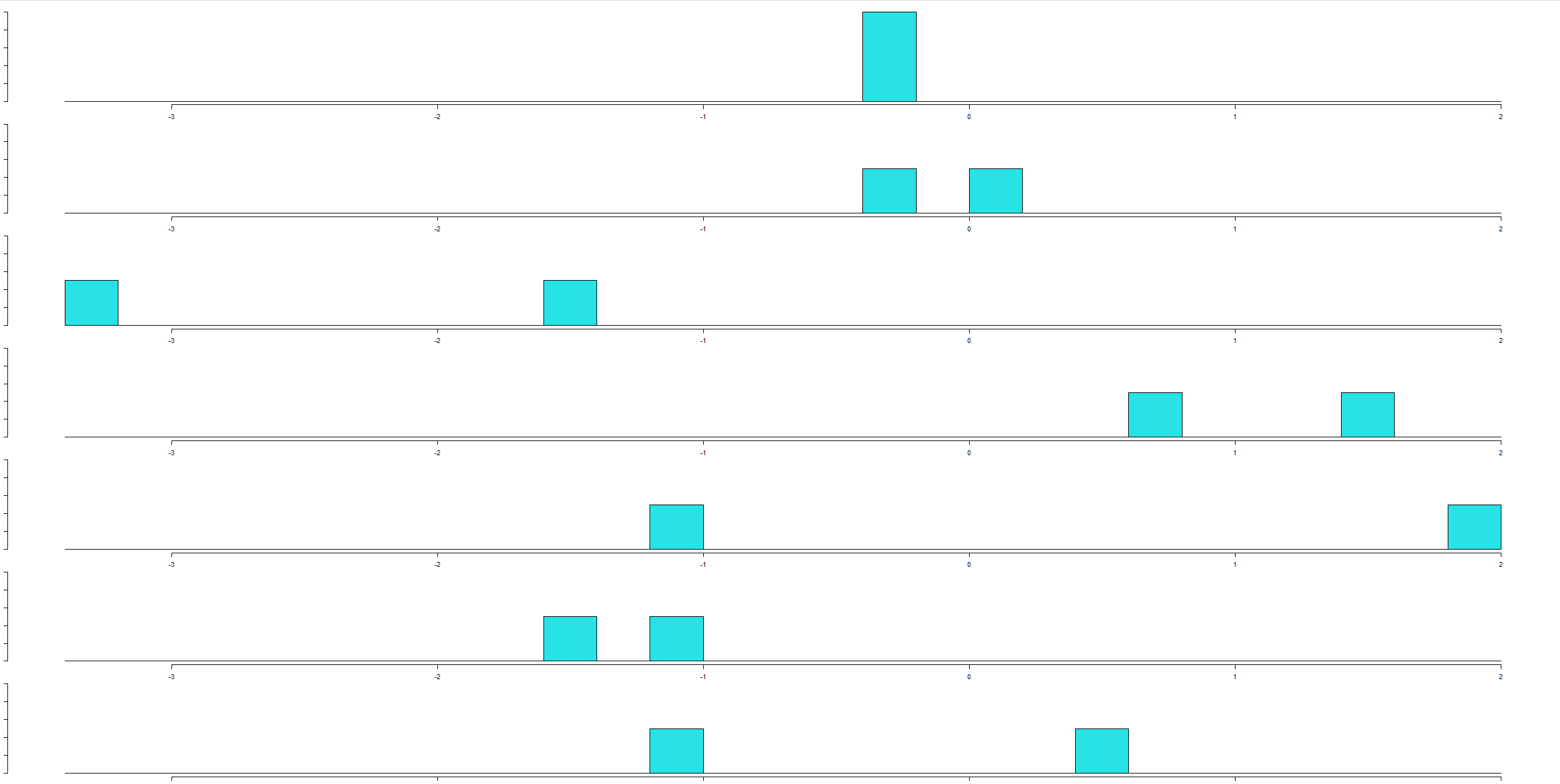


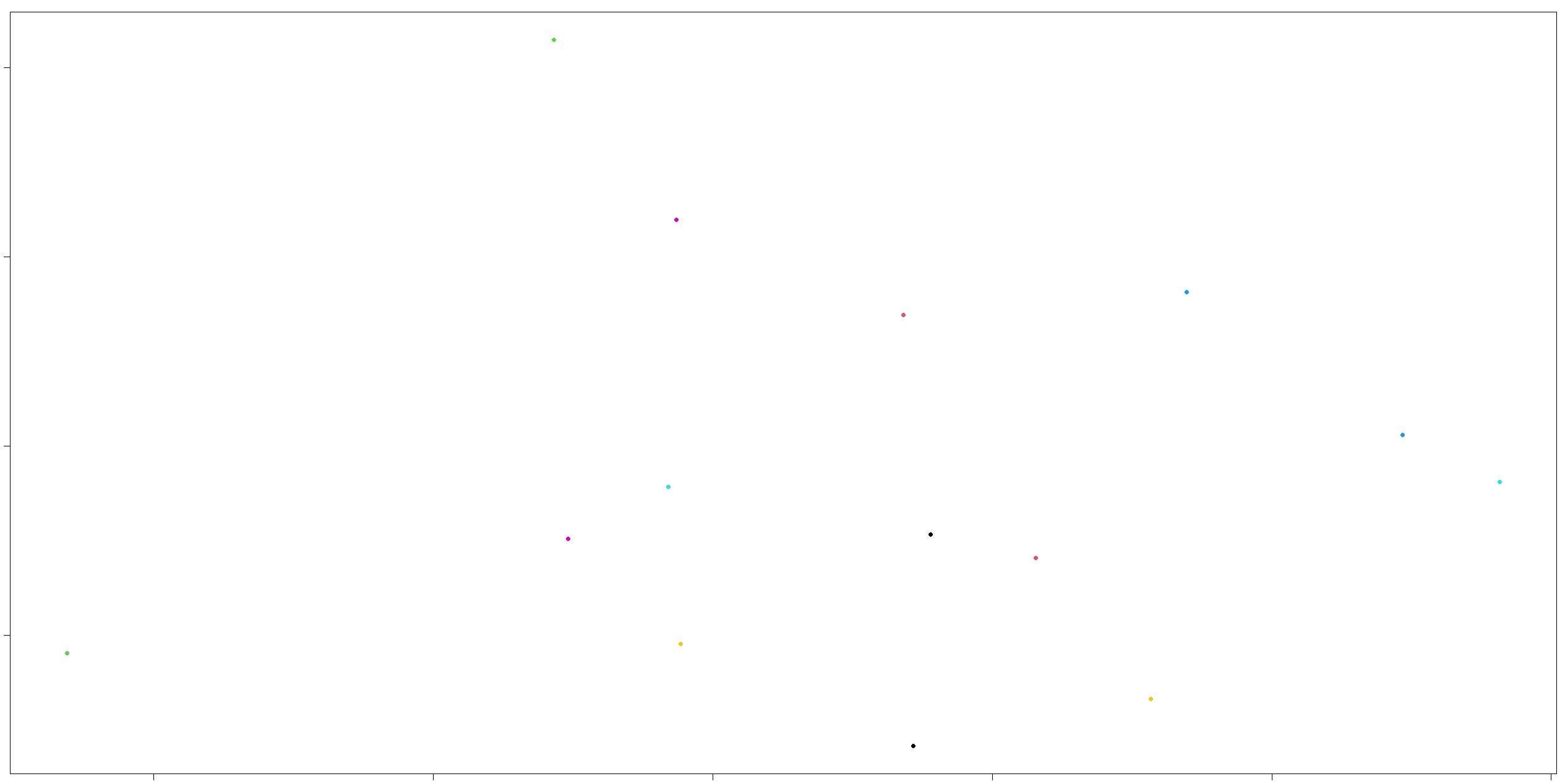
The accuracy achieved by the classifier on the training set is to be at 61.73%.

1. **What is the performance of the classifier on the validation data?**

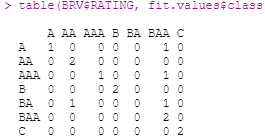
Used the fitted model to predict the values on the validation set and plotted the histograms to check the separation between the ratings. Below I showed the plot of the best separation that can be achieved.



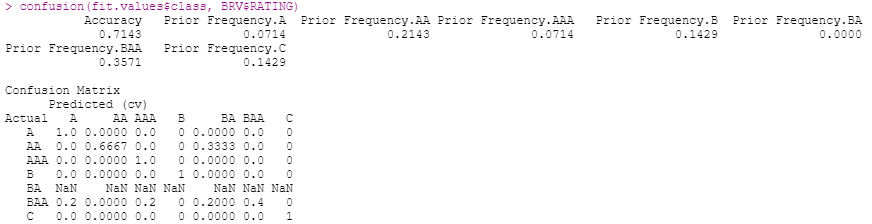




From the histogram and scatter plot we can predict that the A is separated from AAA, B, and BAA. Then, AA is separated from AAA, B, and BAA. Furthermore, AAA is separated from B and C.



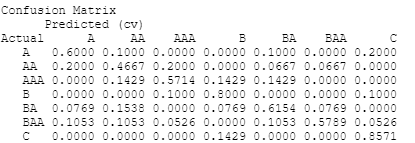
From the above we can infer that ratings A, AAA, B, and C are correctly predicted. The fitted model did not predict any of the BA ratings in Validation Set. One AA is wrongly predicted as BA. One BAA is wrongly predicted as AAA and BA and A.



Moreover, the model did not predict any BA ratings, we got NaN values. 71.43% of accuracy was achieved on the Validation Set. Therefore, the fitted model performed well on the Validation Set than the Training Set.

1. **Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?**

Yes, certain misclassification errors can be worse than the others, this can be measured by computing the confusion matrix. The confusion matrix gives the misclassification rates of the model. For example, the confusion matrix of training set is shown below,



For bond rating of “A” 20% are wrongly predicted as risky “C”, in this case misclassification error is more. 40% of bond rating of “AA” is wrongly predicted as “A” and “AA” (20% each). They are misclassified between extremely safe and moderately safe, here the misclassification error is reduced.