Statistical Analysis of Bankruptcy

Based on Financial Indicators

Group- V: Adrija Saha, Sampurna Mondal, Shrayan Roy

Date: 12/04/2023

Data Description :

- The dataset considered here, is an Annual financial data of financially sound firms and the firms which went bankrupt after two years.
- It contains 46 observations and 5 columns, where last column is the categorical response variable. Which is 0, if the firm went **bankrupt** and 1, if the firm remains **financially sound**.
- 21 observations on bankrupt firms and 25 observations on Financially Sound firms.
- Rest of the four columns are explanatory variables.
- The explanatory variables provided in this dataset are all continuous. Their names are given by -

  1. **Ratios of cash flow to total debt (CFTD)**
  2. **Ratios of net income to total assets (NITA)**
  3. **Ratios of current assets to total liabilities (CATL)**
  4. **Ratios of current assets to net sales (CANS)**

Let's have a look at our dataset :

Show 10 entries Search:

	CFTD	NITA	CATL	CANS	y
1	-0.45	-0.41	1.09	0.45	0
2	-0.56	-0.31	1.51	0.16	0
3	0.06	0.02	1.01	0.4	0
4	-0.07	-0.09	1.45	0.26	0
5	-0.1	-0.09	1.56	0.67	0
6	-0.14	-0.07	0.71	0.28	0
7	0.04	0.01	1.5	0.71	0
8	-0.07	-0.06	1.37	0.4	0
9	0.07	-0.01	1.37	0.34	0
10	-0.14	-0.14	1.42	0.43	0

Showing 1 to 10 of 46 entries

Previous 1 2 3 4 5 Next

Understanding the meaning of Explanatory variables :

- ## Ratios of cash flow to total debt:

This ratio is a type of coverage ratio and can be used to determine how long it would take a company to repay its debt if it devoted all of its cash flow to debt repayment. More the value of the ratio more financially stable it is.

- ## Ratios of net income to total assets:

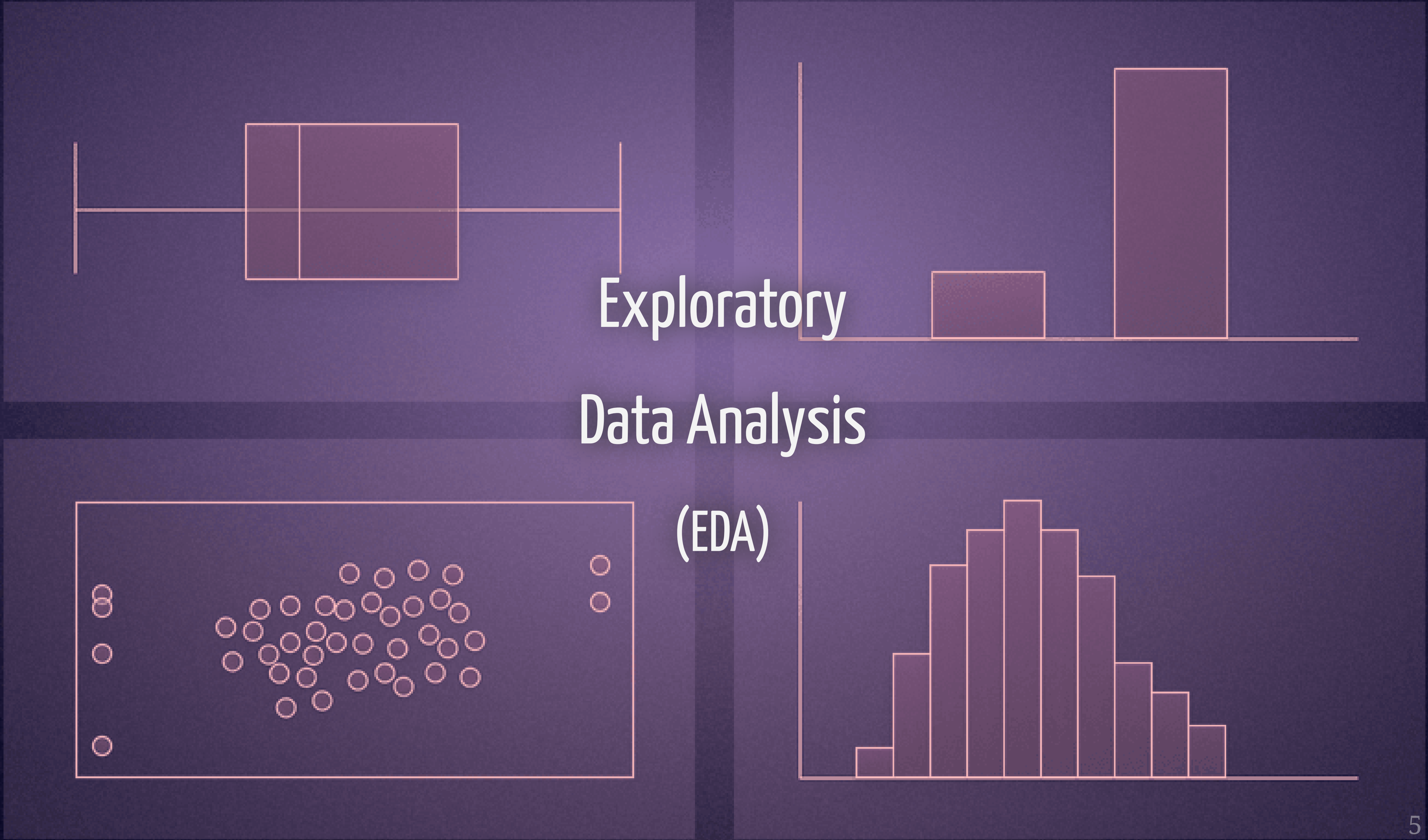
It refers to a financial ratio that indicates how profitable a company is in relation to its total assets. A higher ROA means a company is more efficient.

- ## Ratios of current assets to total liabilities:

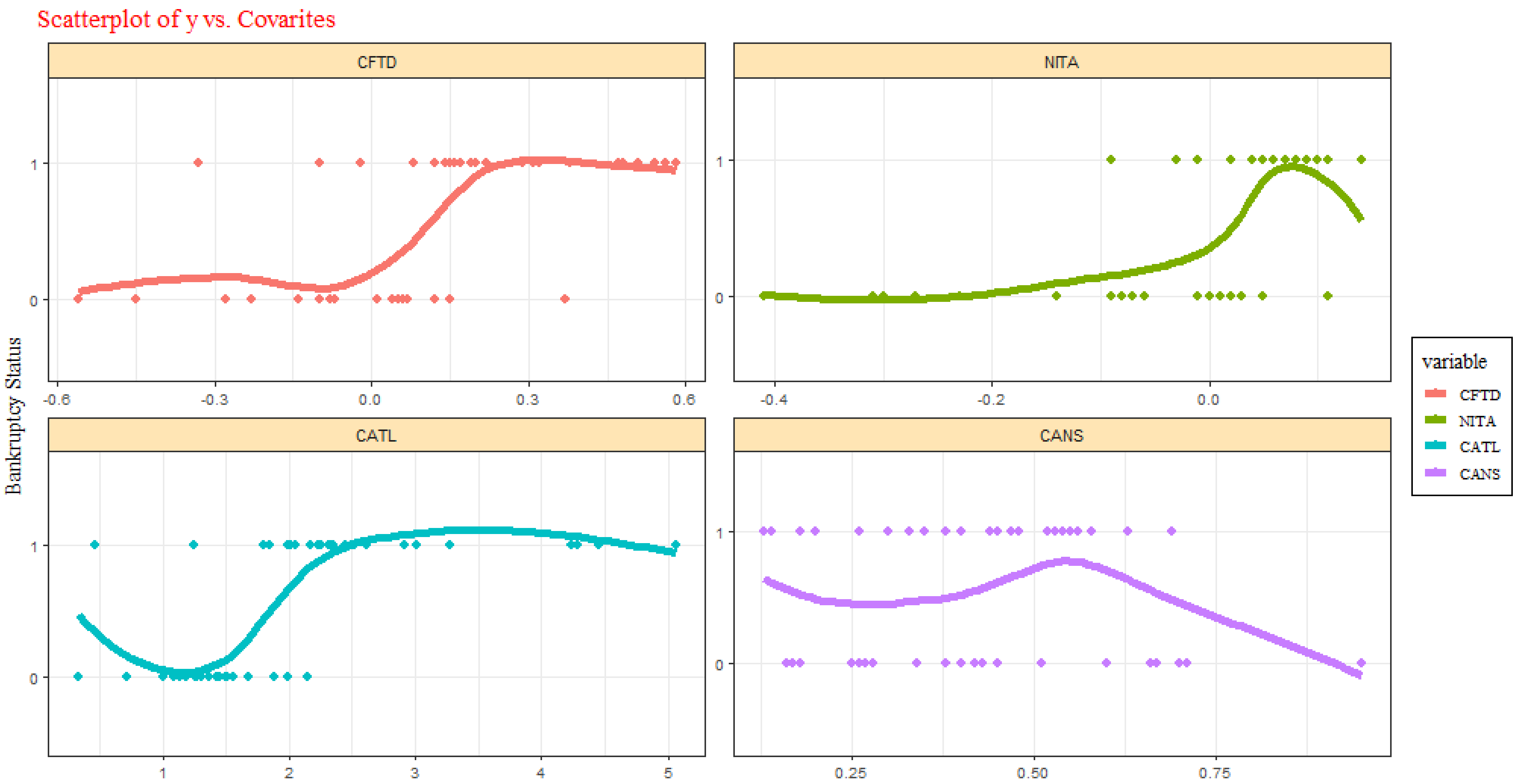
This ratio measures a company's ability to pay short-term obligations or those due within one year with its total assets.Hence, higher value of this ratio indicates less probable of being bankrupt.

- ## Ratios of current assets to net sales:

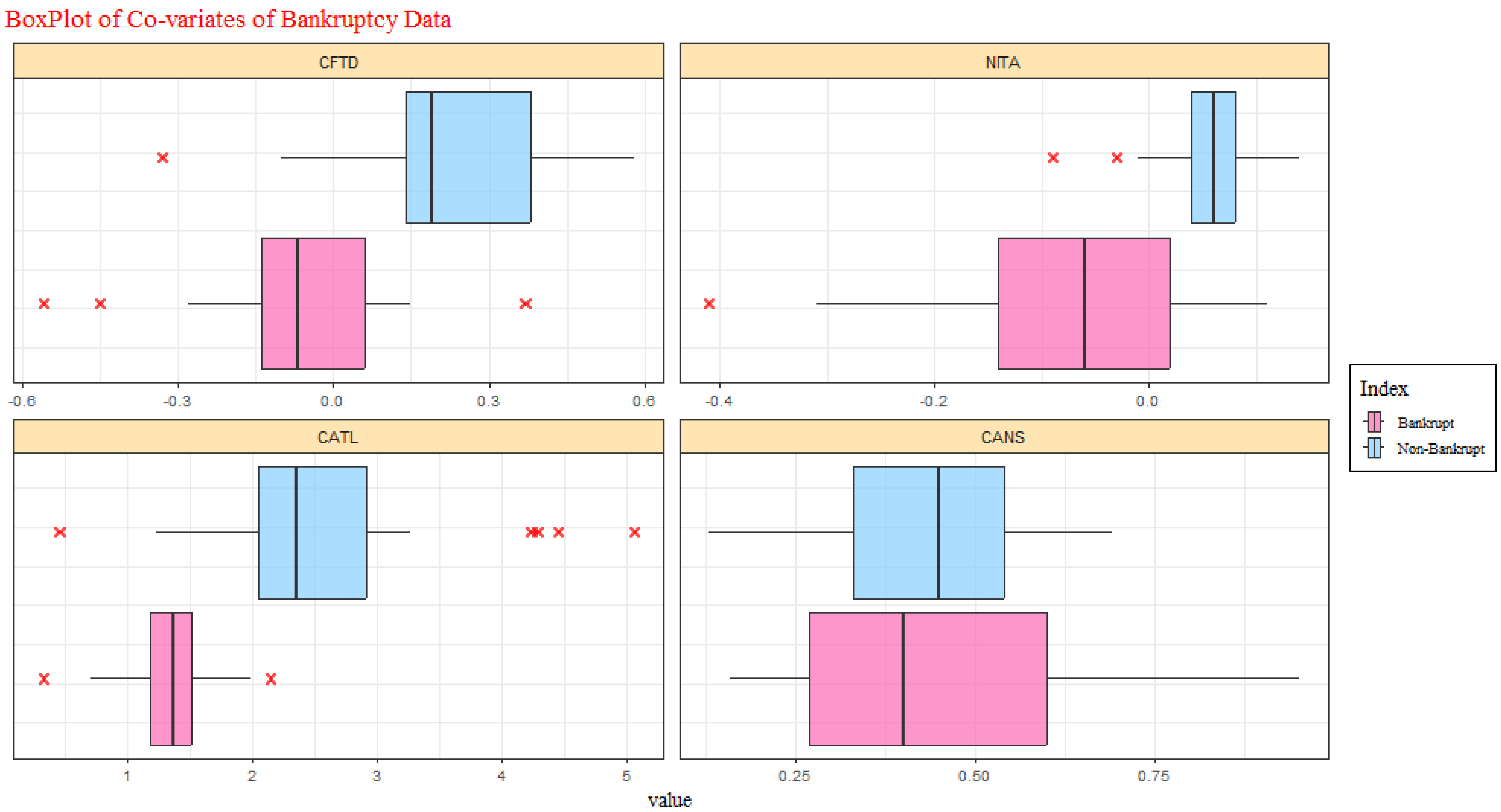
This ratio has an inverse relation with current assets turnover.A higher asset turnover ratio means a better percentage of sales.The less the amount of current assets-net sales ratio, the better the ability of the company to generate sales.


Exploratory
Data Analysis
(EDA)

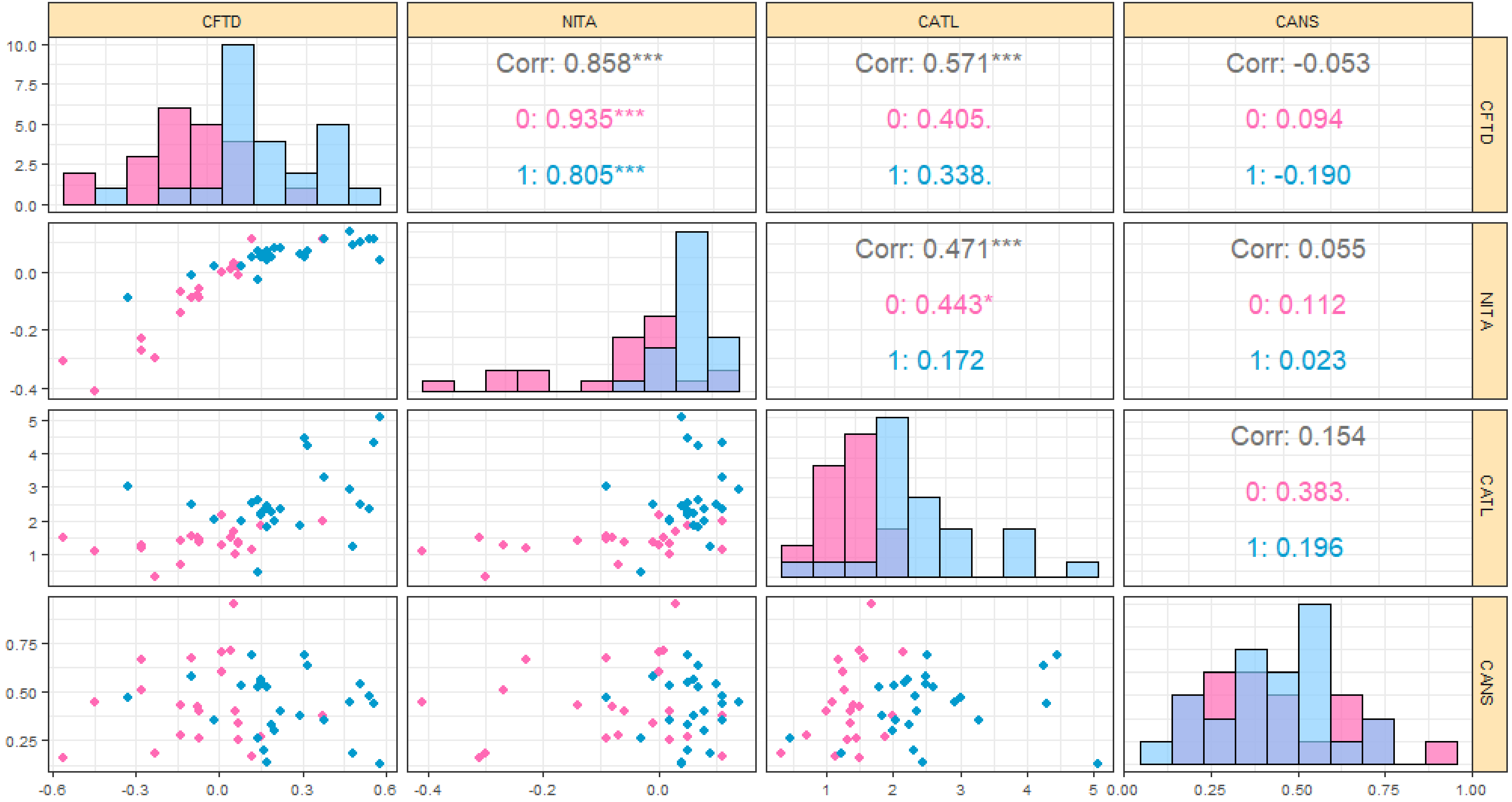
Exploratory Data Analysis: Scatterplot



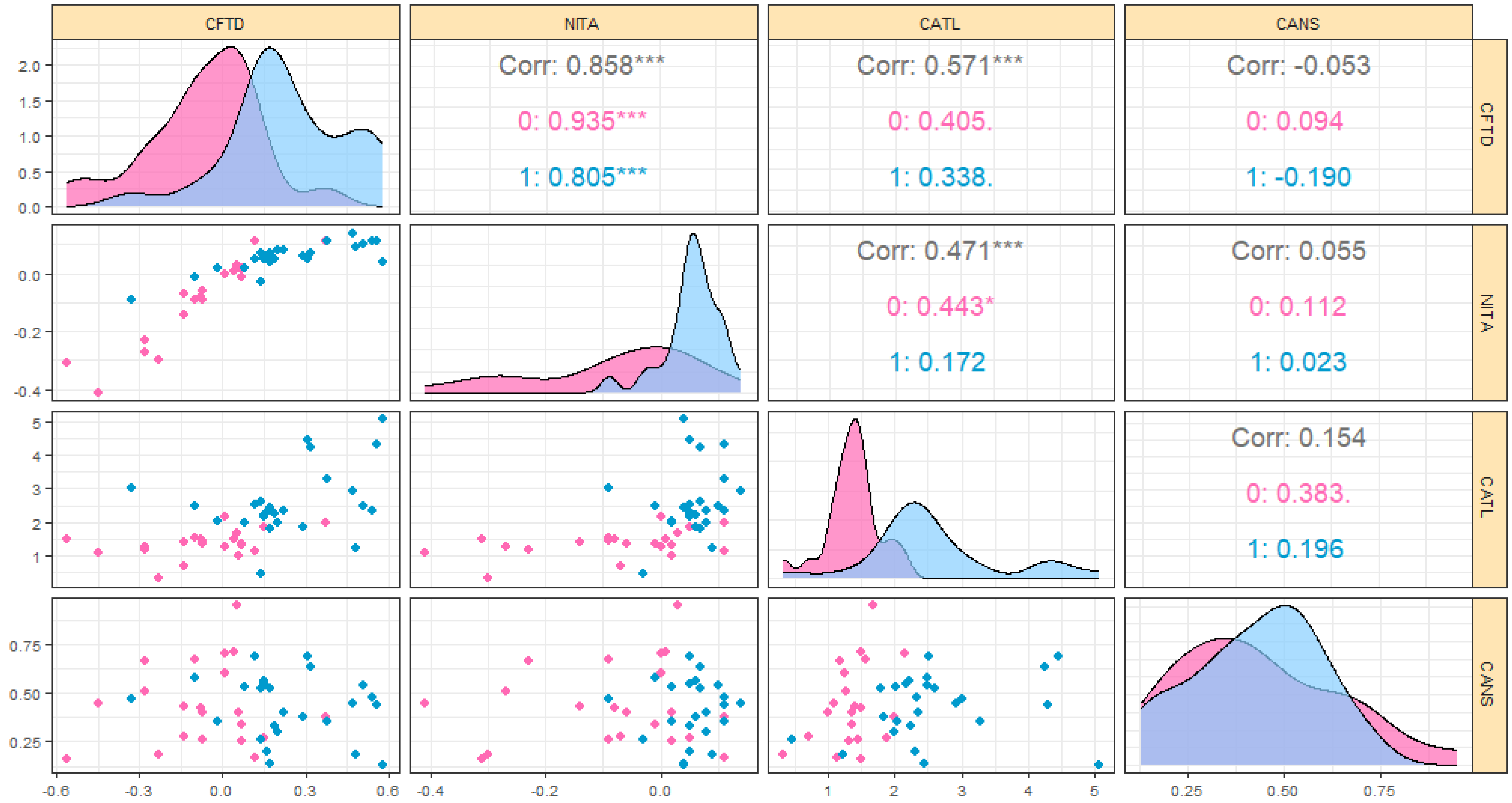
Exploratory Data Analysis: Boxplot



Exploratory Data Analysis: Pairwise Comparison

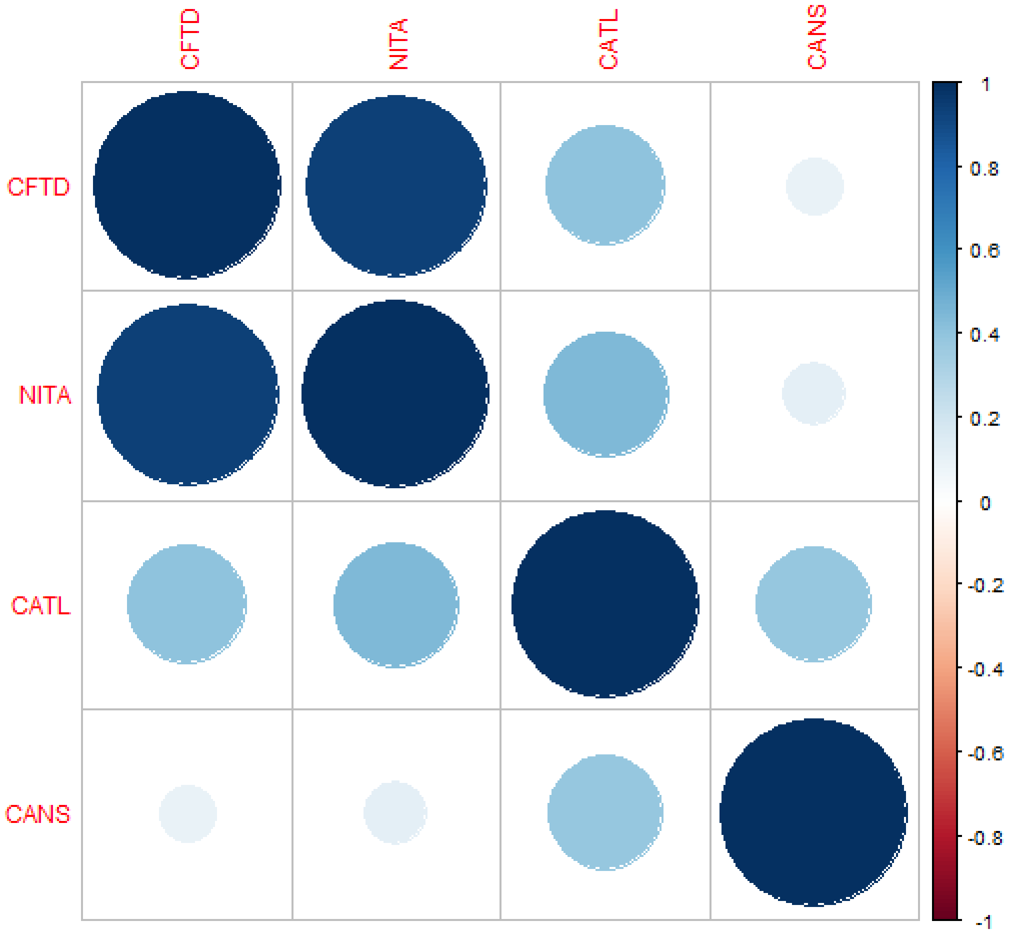


Exploratory Data Analysis: Pairwise Comparison



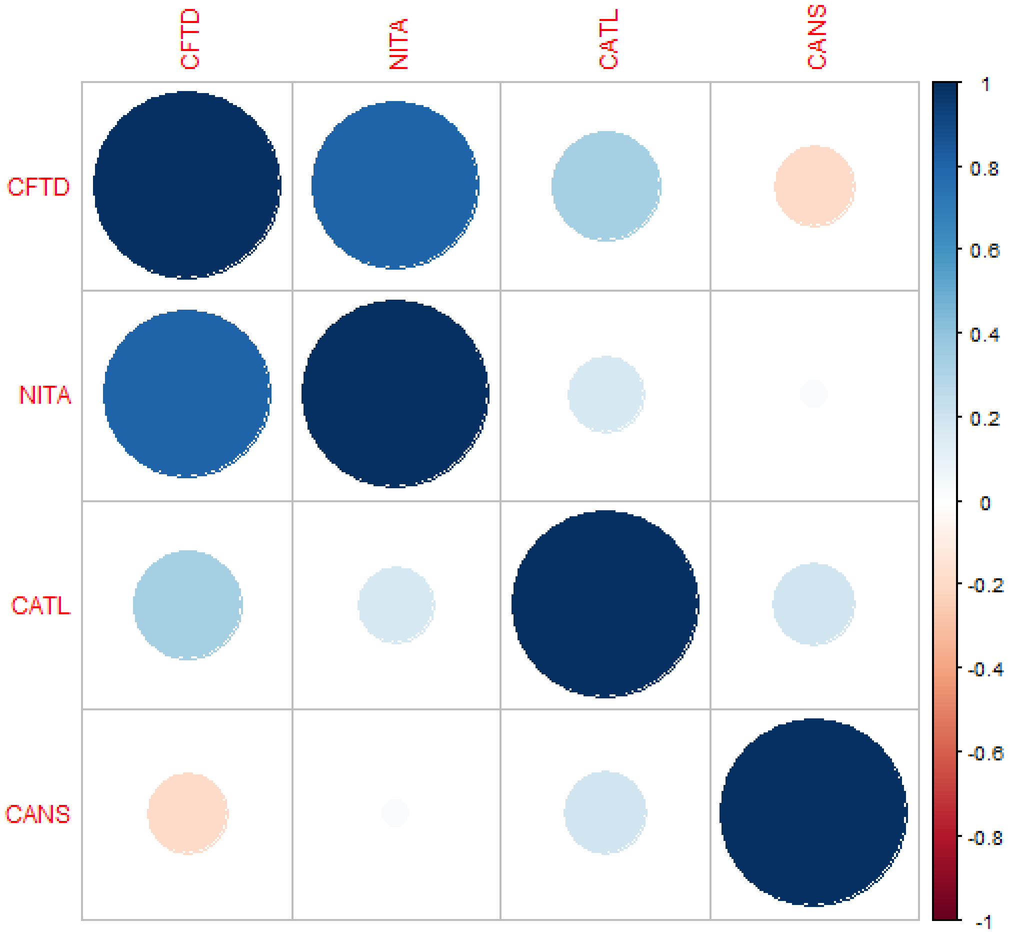
Exploratory Data Analysis: Correlation Plot

- For Bankrupt Firms



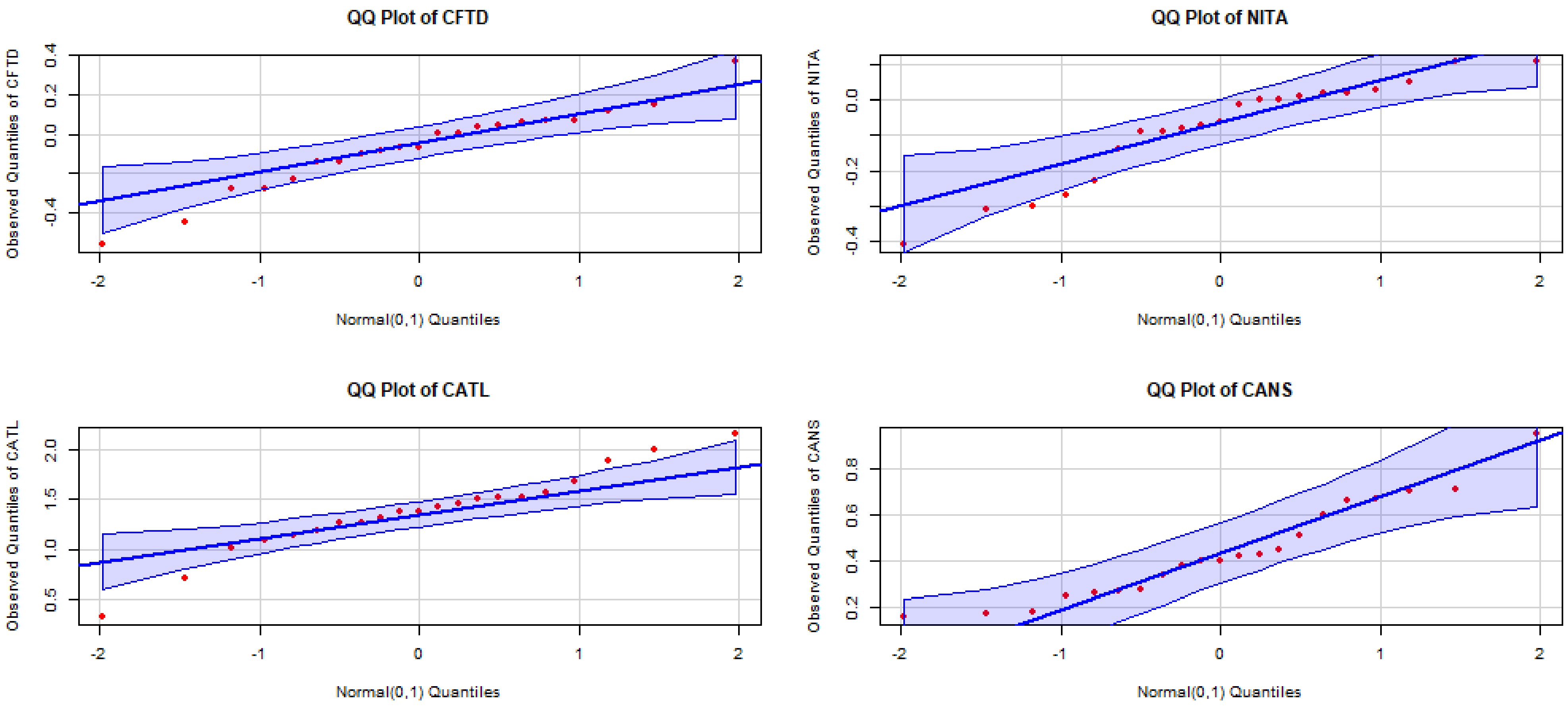
Exploratory Data Analysis: Correlation Plot

- For Financial sound Firms



Checking Normality in Bankrupt Firms:

Q-Q Plot | Shapiro Wilk Test



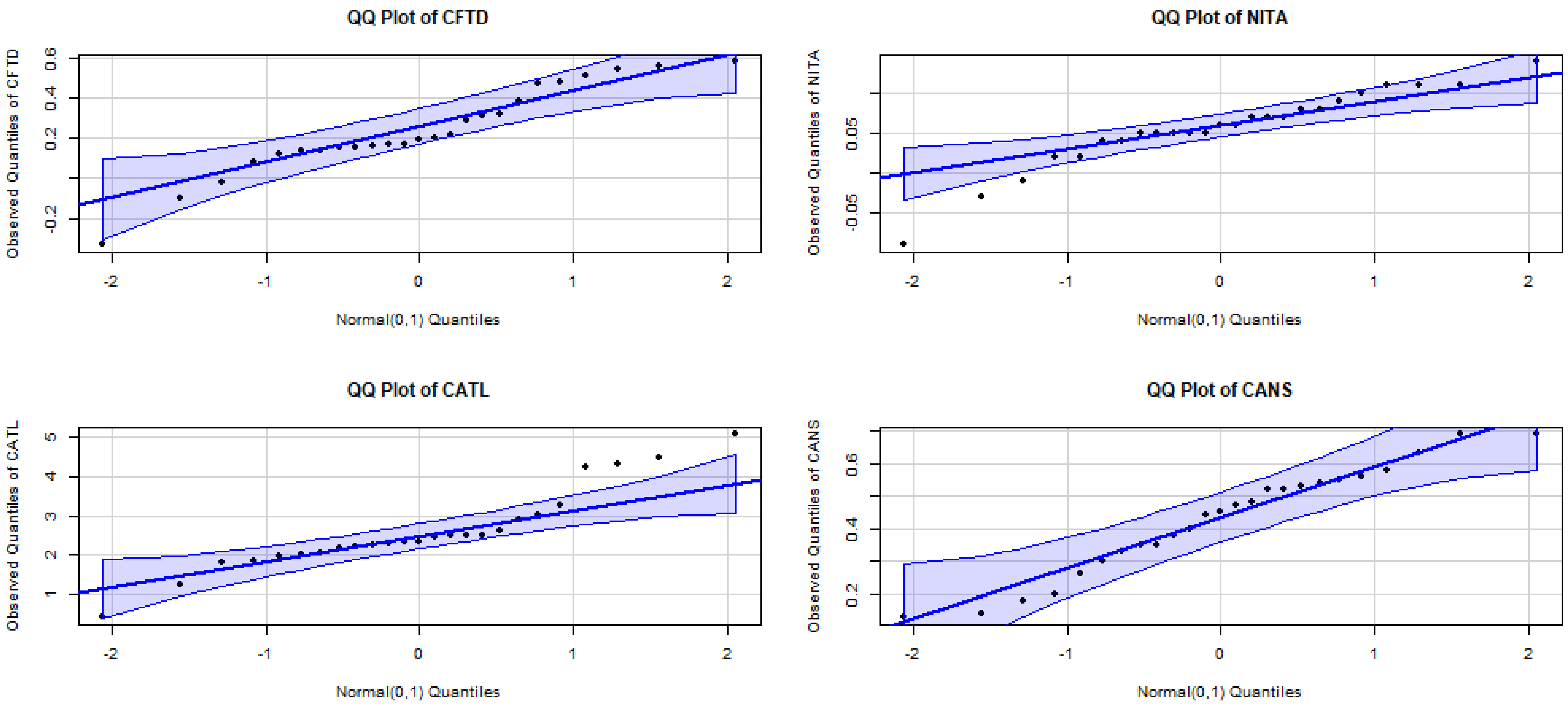
Checking Normality in Bankrupt Firms:

Q-Q Plot | Shapiro Wilk Test

	Value of Test Statistic	p-Value	Decision
CFTD	0.95817	0.48000	Accept
NITA	0.91084	0.05706	Accept
CATL	0.95948	0.50570	Accept
CANS	0.93724	0.19210	Accept

Checking Normality in Financially sound Firms:

Q-Q Plot | Shapiro Wilk Test

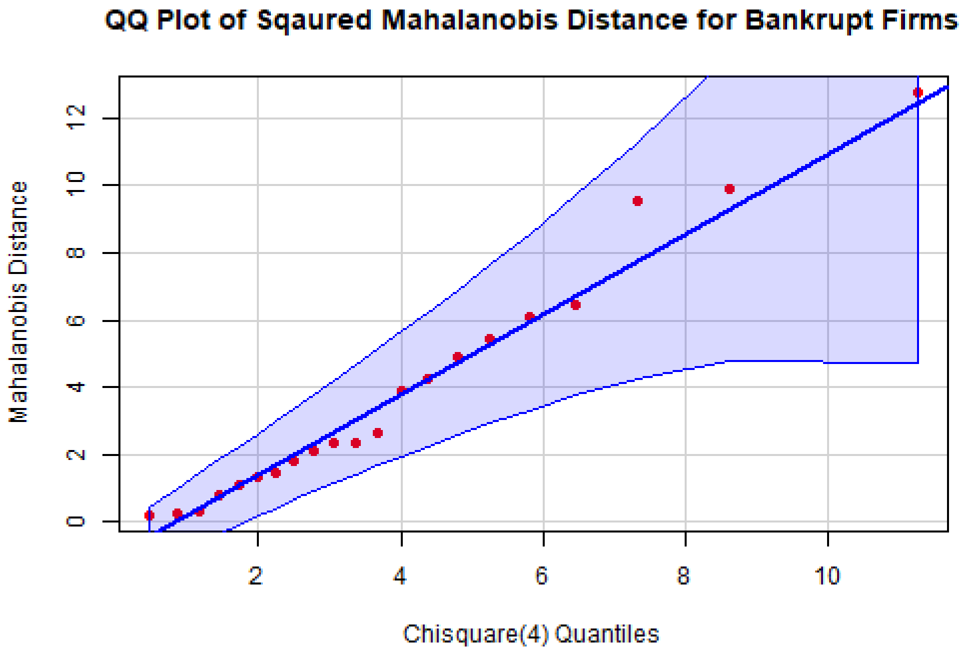


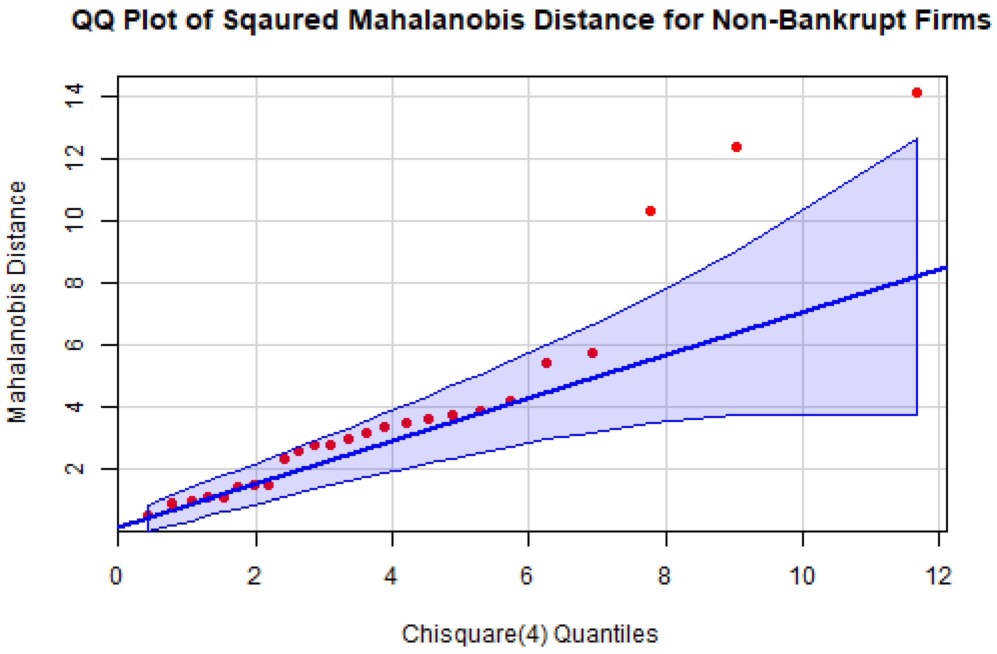
Checking Normality in Financially sound Firms:

Q-Q Plot | **Shapiro Wilk Test**

	Value of Test Statistic	p-Value	Decision
CFTD	0.94170	0.16200	Accept
NITA	0.92382	0.06265	Accept
CATL	***0.90742***	***0.02671***	***Reject***
CANS	0.96139	0.44290	Accept

Chi-Square Plot for checking Multivariate Normality

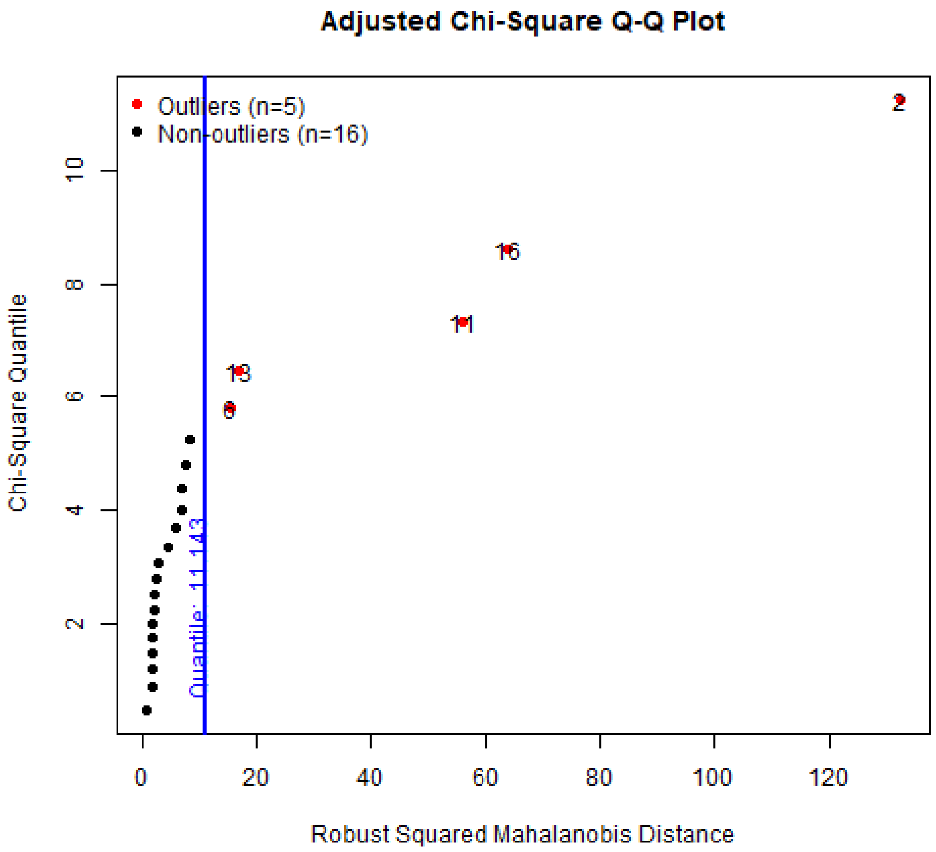




Royston Test : A test of Multivariate Normality :

- Royston’s test uses the Shapiro-Wilk/Shapiro-Francia statistic to test multivariate normality. If kurtosis of the data is greater than 3, then it uses the Shapiro-Francia test for leptokurtic distributions, otherwise it uses the Shapiro-Wilk test for platykurtic distributions.
- It's implementation is available in **MVN** package **R**.
- For more details see MVN: An R Package for Assessing Multivariate Normality

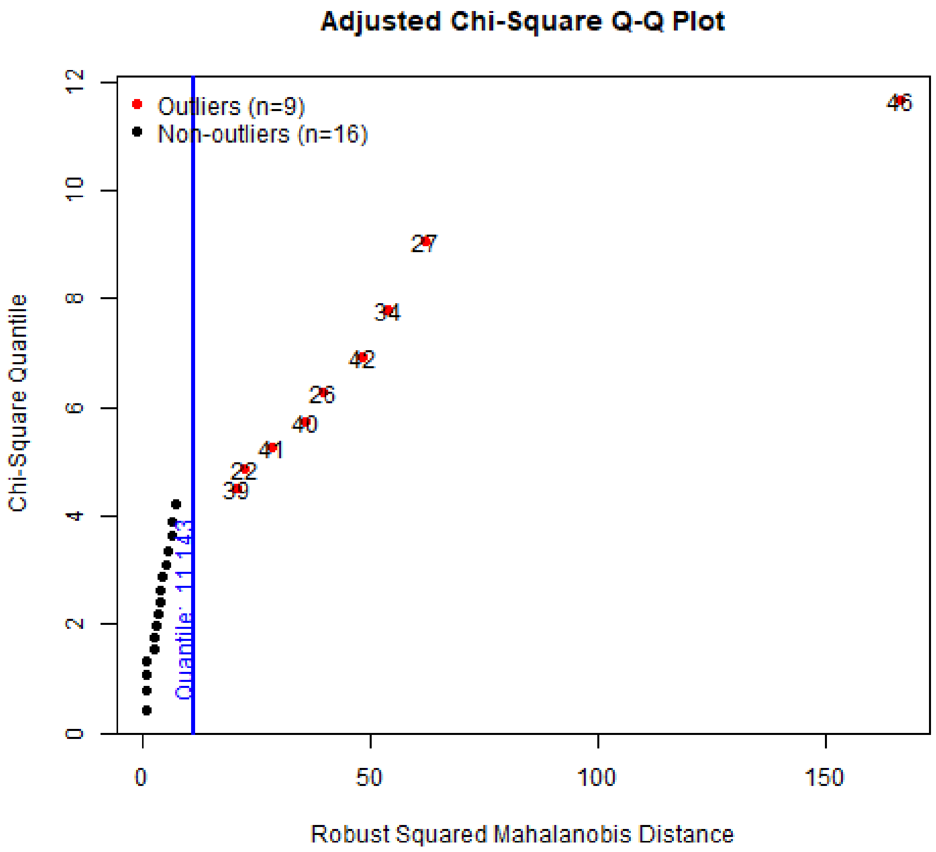
Test for Multivariate Normality : Royston Test (Bankrupt Firms):




```
MVN::mvn(data = My.data0[,-5], mvnTest = "roy
         desc = FALSE,showOutliers = TRUE,mul
```

```
##       Test        H  p value MVN
## 1 Royston 6.04872 0.129197 YES

##    Observation Mahalanobis Distance Outlier
## 2            2              132.443    TRUE
## 16          16               63.885    TRUE
## 11          11               55.893    TRUE
## 13          13               16.726    TRUE
## 6            6               15.691    TRUE
```

Test for Multivariate Normality : Royston Test (Financial Sound Firms):




```
MVN::mvn(data = My.data1[,-5], mvnTest = "roy
         desc = FALSE,showOutliers = TRUE,mul
```

```
##       Test        H    p value MVN
## 1 Royston 12.45531 0.01239924  NO

##    Observation Mahalanobis Distance Outlier
## 46          46              166.133    TRUE
## 27          27               62.055    TRUE
## 34          34               53.992    TRUE
## 42          42               48.329    TRUE
## 26          26               39.645    TRUE
## 40          40               35.802    TRUE
## 41          41               28.260    TRUE
## 22          22               22.397    TRUE
## 39          39               20.710    TRUE
```

A short Note on Robust Mahalanobis Distance :

- Classical Mahalanobis distance is used as a method of detecting outliers.
- But it involves estimate of mean vector and variance-covariance matrix.So, affected by outliers !
- So, a robust method is used to find estimate of mean vector and variance-covariance matrix. Depending upon choice of estimator, we will get different Robust Mahalanobis Distance
- **R** uses adjusted quantile method based Mahalanobis Distance.
- For more details : Selcuk Korkmaz, Dincer Goksuluk and Gokmen Zararsiz : MVN: An R Package for Assessing Multivariate Normality

Discriminant Analysis

Checking Multivariate Normality dropping variables :

- Three Variables at a time!

Royston Test | Chi-Square Plot for CFTD,NITA & CANS

Here only we will check dropping CATL.Since, taking CATL disturbs univariate normality in 2nd population.

- For Bankrupt Firms

```
MVN::mvn(My.data[My.data$y == 0,-c(3,5)],mvnTest = "royston")$multivariateNormality
```

```
##       Test        H   p value MVN
## 1 Royston 4.823069 0.1128417 YES
```

- For Financial sound Firms

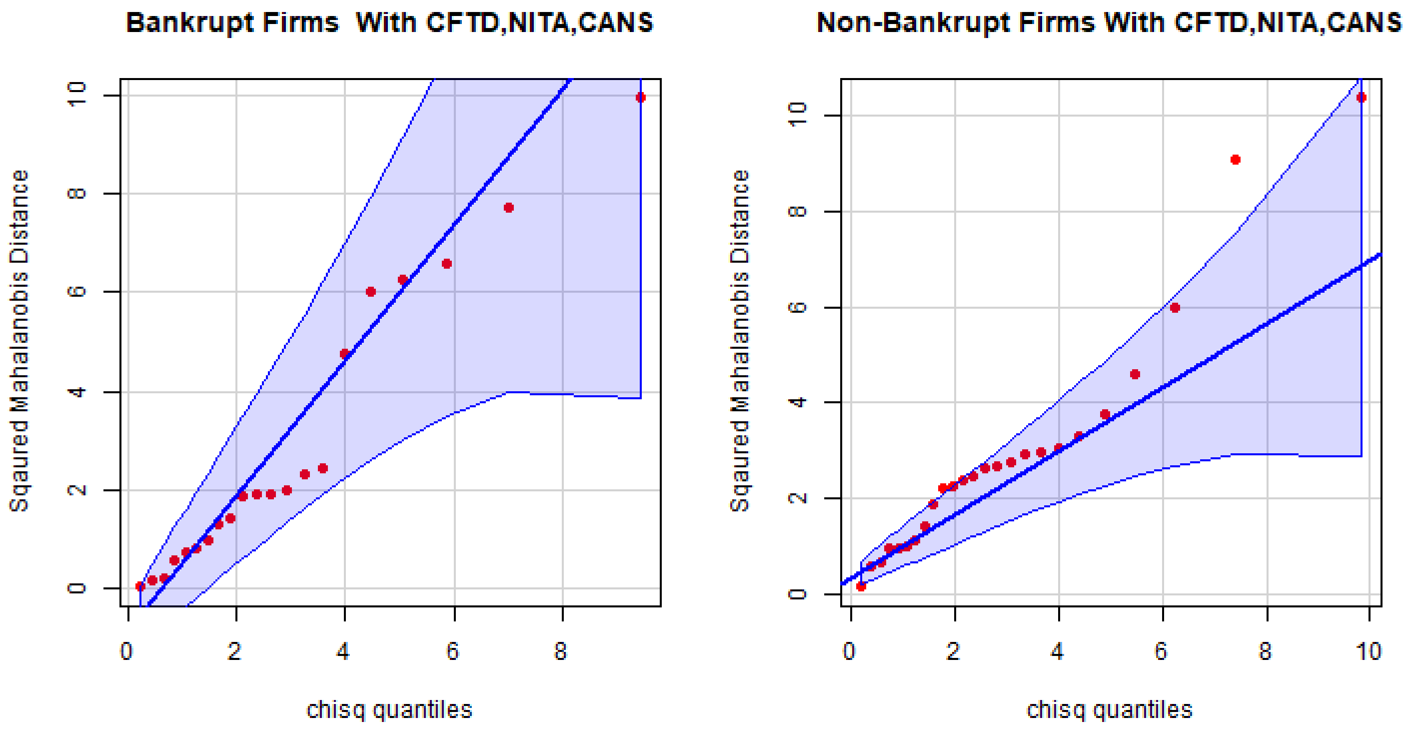
```
MVN::mvn(My.data[My.data$y == 1,-c(3,5)],mvnTest = "royston")$multivariateNormality
```

```
##       Test        H    p value MVN
## 1 Royston 6.856861 0.06696829 YES
```

Checking Multivariate Normality dropping variables :

- Three Variables at a time!

Royston Test | Chi-Square Plot for CFTD,NITA & CANS



Analysis with CFTD, NITA & CANS:

Box-M Test QDA Performance

```
heplots::boxM(as.matrix(My.data[,-c(3,5)]) ~ as.factor(y),data = My.data)
```

```
##
##      Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 46.237, df = 6, p-value = 2.655e-08
```

Analysis with CFTD, NITA & CANS:

Box-M Test | QDA | Performance

```
qda(My.data[,-c(3,5)],My.data$y)
```

```
## Call:
## qda(My.data[, -c(3, 5)], My.data$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD        NITA     CANS
## 0 -0.06904762 -0.08142857 0.437619
## 1  0.23520000  0.05560000 0.426800
```

Analysis with CFTD, NITA & CANS:

Box-M Test | QDA | **Performance**

Training Set Performance

```
table(Actual = My.data[,5], Predicted = predict(lda(My.data[,-c(3,5)],My.data$y))$class)
```

```
##        Predicted
## Actual  0  1
##       0 13  8
##       1  3 22
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], Qda_Model.3$class)
```

```
## [1] 0.2173913
```

Checking Multivariate Normality dropping variables :

- Two Variables at a time !

Royston Test | Chi-Square Plots for CFTD & CANS | Chi-Square Plots for NITA & CANS

Here only we will not include CATL anywhere.Since, taking CATL disturbs univariate normality in 2nd population.

For Bankrupt firms

	Variables Included	Test statistic	p-Value	Decision
1	CFTD,NITA	3.151806	0.1168098	Accept
3	***CFTD,CANS***	***2.737765***	***0.2543941***	***Accept***
5	***NITA,CANS***	***5.322533***	***0.0698239***	***Accept***

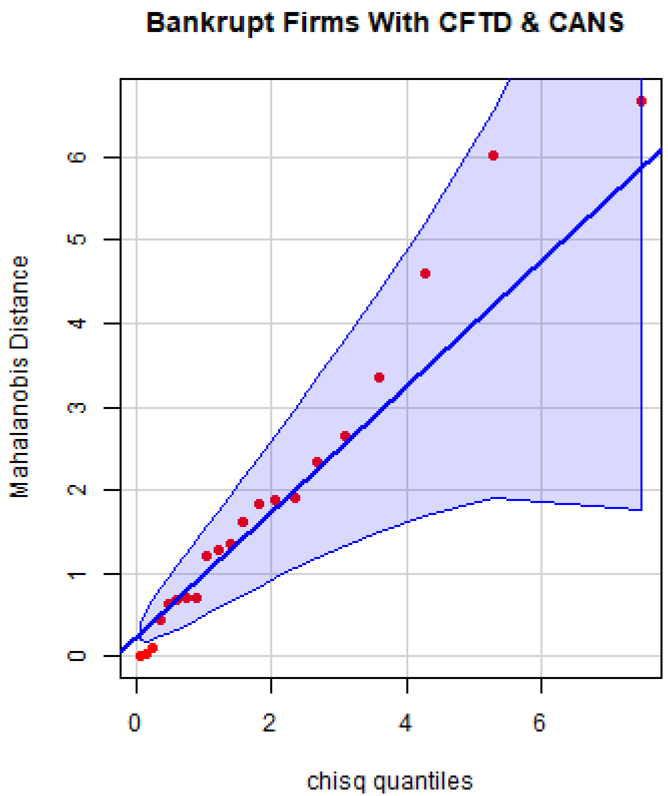
For Financially sound firms

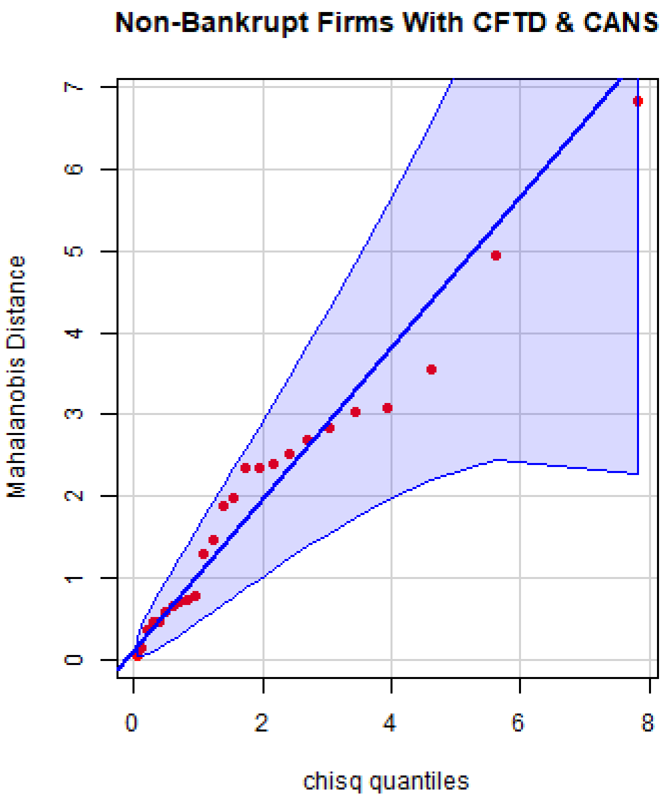
	Variables Included	Test statistic	p-Value	Decision
2	CFTD,NITA	6.122089	0.0392588	Reject
4	***CFTD,CANS***	***2.735482***	***0.2545416***	***Accept***
6	***NITA,CANS***	***5.146921***	***0.0762711***	***Accept***

Checking Multivariate Normality dropping variables :

- Two Variables at a time !

Royston Test | Chi-Square Plots for CFTD & CANS | Chi-Square Plots for NITA & CANS

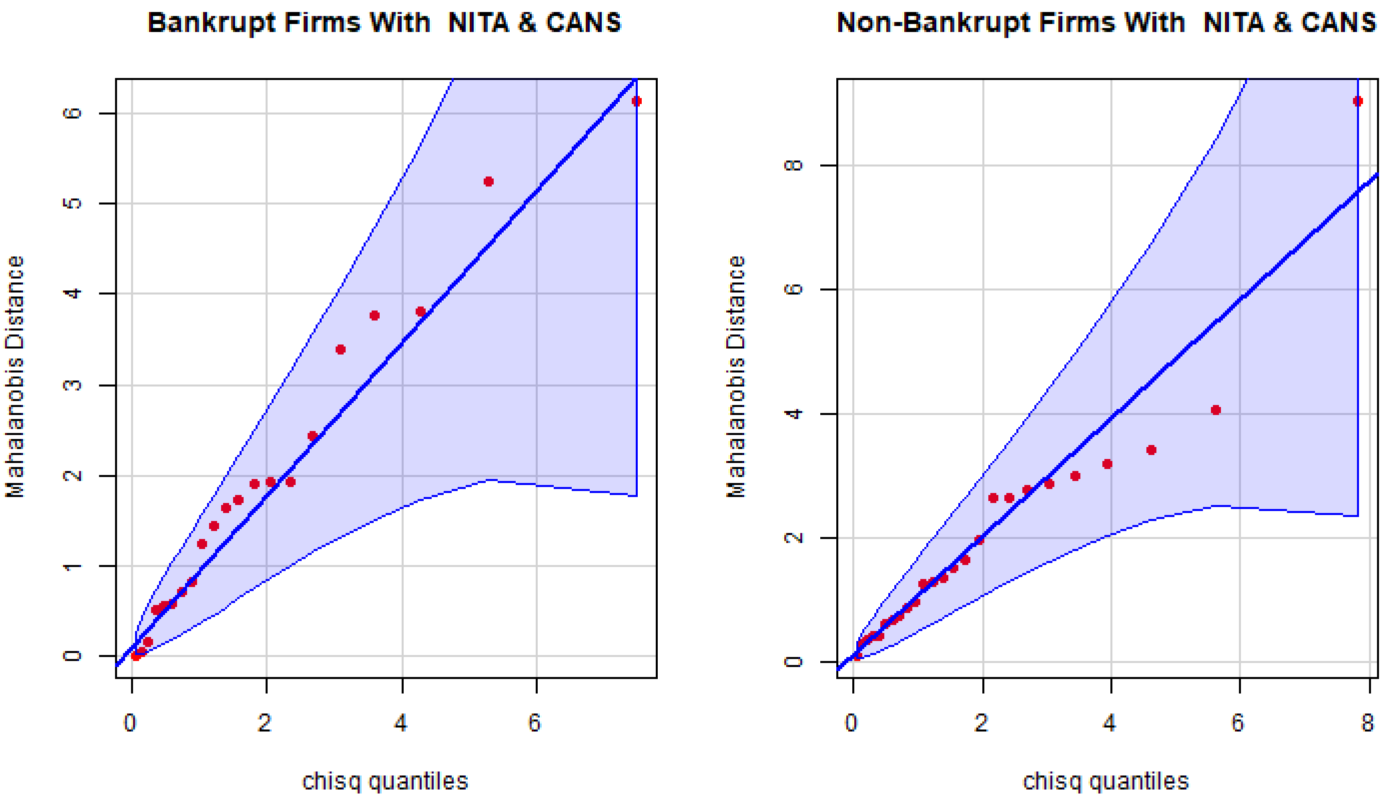




Checking Multivariate Normality dropping variables :

- Two Variables at a time !

Royston Test | Chi-Square Plots for CFTD & CANS | **Chi-Square Plots for NITA & CANS**



Analysis with CFTD & CANS :

Box-M Test and MANOVA | LDA | Performance

```r
heplots::boxM(as.matrix(My.data[,-c(2,3,5)]) ~ as.factor(y),data = My.data)
```

```
##
##      Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 2.3791, df = 3, p-value = 0.4975
```

```r
model.manova <- manova(cbind(CFTD,CANS)~y,data = My.data)
summary(model.manova)
```

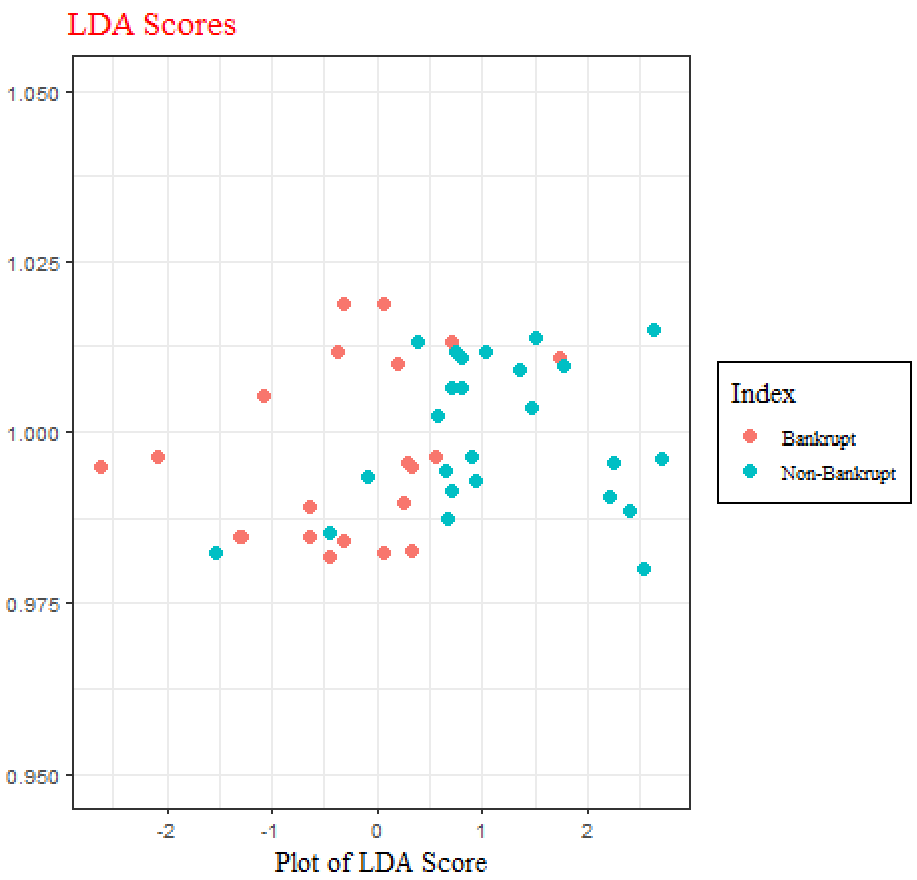
```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## y           1 0.34432    11.29      2     43 0.0001145 ***
## Residuals 44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analysis with CFTD & CANS :

Box-M Test and MANOVA | **LDA** | Performance

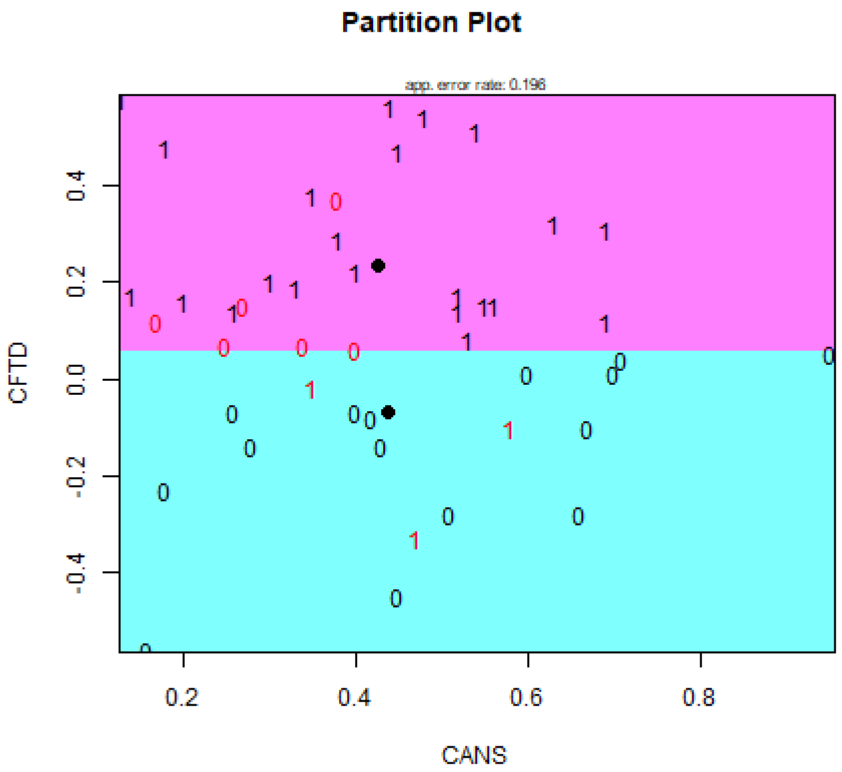
```
lda(My.data[,-c(2,3,5)],My.data$y)
```

```
## Call:
## lda(My.data[, -c(2, 3, 5)], My.data$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD     CANS
## 0 -0.06904762 0.437619
## 1  0.23520000 0.426800
##
## Coefficients of linear discriminants:
##              LD1
## CFTD 4.67736451
## CANS 0.01965838
```



Analysis with CFTD & CANS :

Box-M Test and MANOVA | LDA | **Performance**



Training Set Performance

```
table(Actual = My.data[,5], Predicted = predi
```

```
##         Predicted
## Actual   0  1
##        0 15  6
##        1  3 22
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], Lda_Model.1$class)
```

```
## [1] 0.2391304
```

Analysis with NITA & CANS:

Box-M Test QDA Performance

```r
heplots::boxM(as.matrix(My.data[,-c(1,3,5)]) ~ as.factor(y),data = My.data)
```

```
##
##     Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 23.435, df = 3, p-value = 3.277e-05
```

Analysis with NITA & CANS:

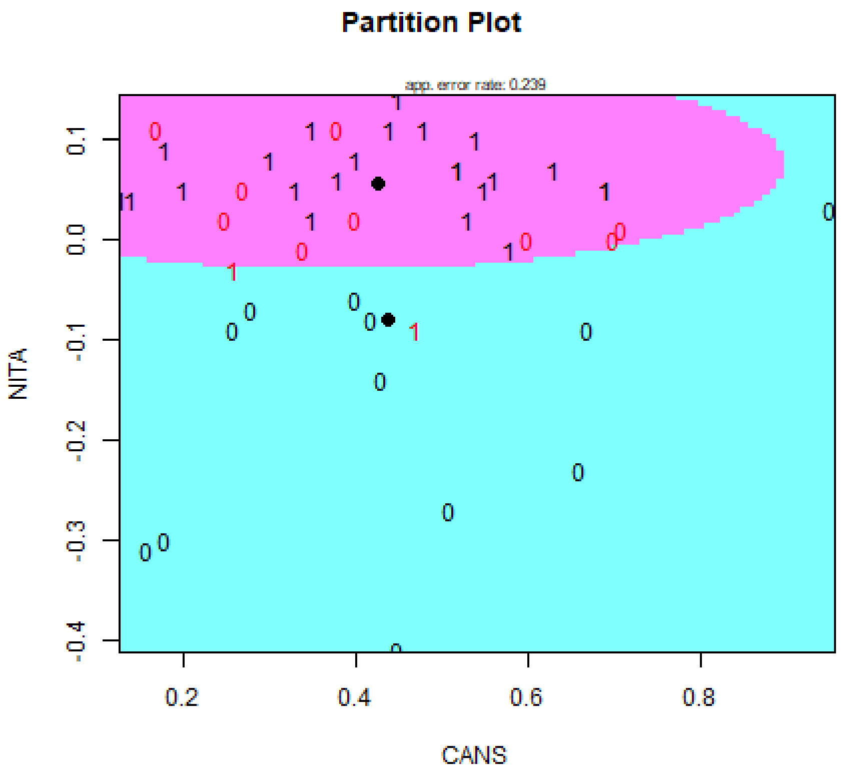
Box-M Test | QDA | Performance

```
qda(My.data[,-c(1,3,5)],My.data$y)
```

```
## Call:
## qda(My.data[, -c(1, 3, 5)], My.data$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          NITA     CANS
## 0 -0.08142857 0.437619
## 1  0.05560000 0.426800
```

Analysis with NITA & CANS:

Box-M Test | QDA | **Performance**



Training Set Performance

```
table(Actual = My.data[,5], Predicted = predi
```

```
##        Predicted
## Actual  0  1
##       0 12  9
##       1  2 23
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], Qda_Model.2$class)
```

```
## [1] 0.2608696
```

Transformation for Multivariate Normality :

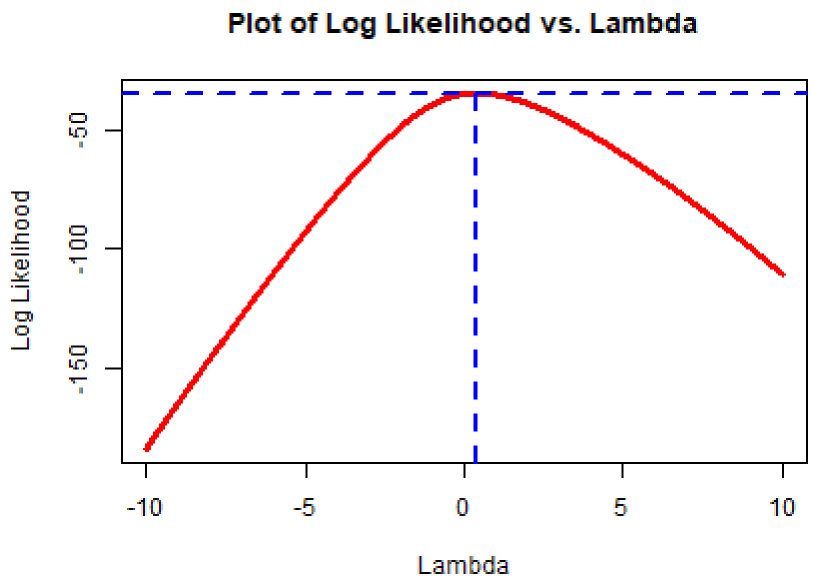
- **Box-Cox Transformation** is a commonly used transformation for normality.
- But, Applicability of this is restricted to positive valued variables only.
- Yeo-Johnson :A New Family of Power Transformations to Improve Normality or Symmetry suggested a generalized Box-cox transformation. Which is defined as -

$$\psi(y,\lambda) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & y \geq 0, \lambda \neq 0 \\ log(y+1), & y \geq 0, \lambda = 0 \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & y < 0, \lambda \neq 2 \\ -log(-y+1), & y < 0, \lambda = 2 \end{cases}$$

- We will use this to transformation to achieve normality of third variable.
- To obtain Optimal λ, we will use likelihood based approach.

Transforming CATL for Financially sound firms:

Finding Optimum lambda | Test for Multivariate Normality




```
##   Optimal Lambda Likelihood Value
##          0.41000        -34.83718
```

Transforming CATL for Financially sound firms:

Finding Optimum lambda | **Test for Multivariate Normality**

- After transformation:

```
MVN::mvn(data = My.data_trans0[My.data_trans0$y == 1,-5], mvnTest = "royston",
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 11.19854 0.02137175  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   CFTD       0.9417    0.1620    YES
## 2 Shapiro-Wilk   NITA       0.9238    0.0626    YES
## 3 Shapiro-Wilk   CATL       0.9256    0.0688    YES
## 4 Shapiro-Wilk   CANS       0.9614    0.4429    YES
```

Multivariate Normality Rejected !

Finding Optimum λ based on joint likelihood:

- Applying Yeo-Johnson family of power transformation is yielding univariate normality. But, we are not getting multivariate normality.
- Instead we could find the log likelihood (mentioned in Yeo-Johnson Paper) of two populations separately for same λ. And then maximize the sum of the log-likelihood as a function of λ.
- So,we will maximize -

$$l_{n_1,n_2}(\lambda|X_1, X_2) = l_{n_1}(\lambda|X_1) + l_{n_2}(\lambda|X_2)$$

as a function of λ.

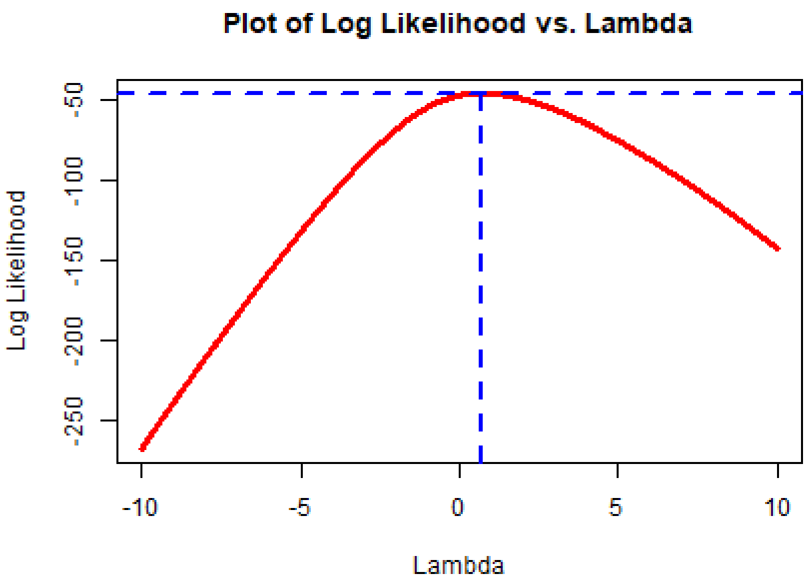
- Then, we will get same lambda for both the population.

Transforming CATL maximing joint likelihood:

```
g.boxcox<- function(data0,data1,lambda.seq){

  #used to calculate joint likelihood
  lbc.mv <- function(lambda.1){
  l.gbc(data0,lambda.1)+l.gbc(data1,lambda.1)
  #l.gbc calculates likelihood based on Yeo-Johnson
  }

  #plot of likelihood vs. lambda graph...
  plot(lambda.seq,vapply(lambda.seq, FUN = lbc.mv, FUN.VALUE = 2),
       col = "red",type = "l",xlab = "Lambda",ylab = "Log Likelihood",
       main = "Plot of Log Likelihood vs. Lambda",lwd = 3)

  #adding reference line...
  abline(h = max(vapply(lambda.seq, FUN = lbc.mv, FUN.VALUE = 2)) - 0.5,lty = 2,col = "blue",lwd
  abline(v = lambda.seq[which.max(vapply(lambda.seq, FUN = lbc.mv, FUN.VALUE = 2))],lty = 2,col =

  #Printing the value of optimal lambda...
print(c("Optimal Lambda" = lambda.seq[which.max(vapply(lambda.seq, FUN = lbc.mv, FUN.VALUE = 2))]
          "Likelihood Value" = max(vapply(lambda.seq, FUN = lbc.mv, FUN.VALUE = 2))))

}
```

Transforming CATL by maximizing joint likelihood:

Finding Optimum lambda | For Bankrupt Firms | For Financially sound Firms




```
##   Optimal Lambda Likelihood Value
##          0.72000        -45.66093
```

Transforming CATL by maximizing joint likelihood:

Finding Optimum lambda | **For Bankrupt Firms** | For Financially sound Firms

- After transformation:

```
MVN::mvn(data = My.data_trans2[My.data_trans2$y == 0,-5], mvnTest = "royston",
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 6.668277 0.09935673 YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   CFTD       0.9582    0.4800    YES
## 2 Shapiro-Wilk   NITA       0.9108    0.0571    YES
## 3 Shapiro-Wilk   CATL       0.9487    0.3222    YES
## 4 Shapiro-Wilk   CANS       0.9372    0.1921    YES
```

Multivariate Normality Accepted !

Transforming CATL by maximizing joint likelihood:

Finding Optimum lambda | For Bankrupt Firms | **For Financially sound Firms**

- After transformation:

```
MVN::mvn(data = My.data_trans2[My.data_trans2$y == 1,-5], mvnTest = "royston",
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 11.46395 0.0190663  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   CFTD      0.9417    0.1620    YES
## 2 Shapiro-Wilk   NITA      0.9238    0.0626    YES
## 3 Shapiro-Wilk   CATL      0.9205    0.0527    YES
## 4 Shapiro-Wilk   CANS      0.9614    0.4429    YES
```

Multivariate Normality Rejected !

Multivariate version of Yeo-Johnson family of Transformation:

- Univariate transformation is not helping much !
- **Solution**: Multivariate version of **Yeo-Johnson Transformation**.
- Implementation is available in **R**, **powerTransform** function of car package.
- Let's try to implement this transformation on Financially Sound Firms first.

Using Multivariate Yeo-Johnson transformation :

Finding Optimum lambda | For Financially sound Firms | For Bankrupt Firms | Chisqaure Plot

```
lambda.2 <- car::powerTransform(as.matrix(My.data1[,-5]),family = "yjPower")
lambda.2
```

```
## Estimated transformation parameters
##      CFTD      NITA      CATL      CANS
## 1.2504445 5.3233515 0.6919768 1.7079405
```

Using Multivariate Yeo-Johnson transformation :

Finding Optimum lambda | **For Financially sound Firms** | For Bankrupt Firms | Chisqaure Plot

- After transformation:

```
MVN::mvn(data = My.data_trans4[My.data_trans4$y == 1,-5], mvnTest = "royston",
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 6.981046 0.1261358 YES
##
## $univariateNormality
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk CFTD_Trans    0.9454    0.1970    YES
## 2 Shapiro-Wilk NITA_Trans    0.9714    0.6809    YES
## 3 Shapiro-Wilk CATL_Trans    0.9214    0.0552    YES
## 4 Shapiro-Wilk CANS_Trans    0.9663    0.5539    YES
```

Multivariate Normality Accepted !

Using Multivariate Yeo-Johnson transformation :

Finding Optimum lambda | For Financially sound Firms | **For Bankrupt Firms** | Chisqaure Plot

- After transformation:

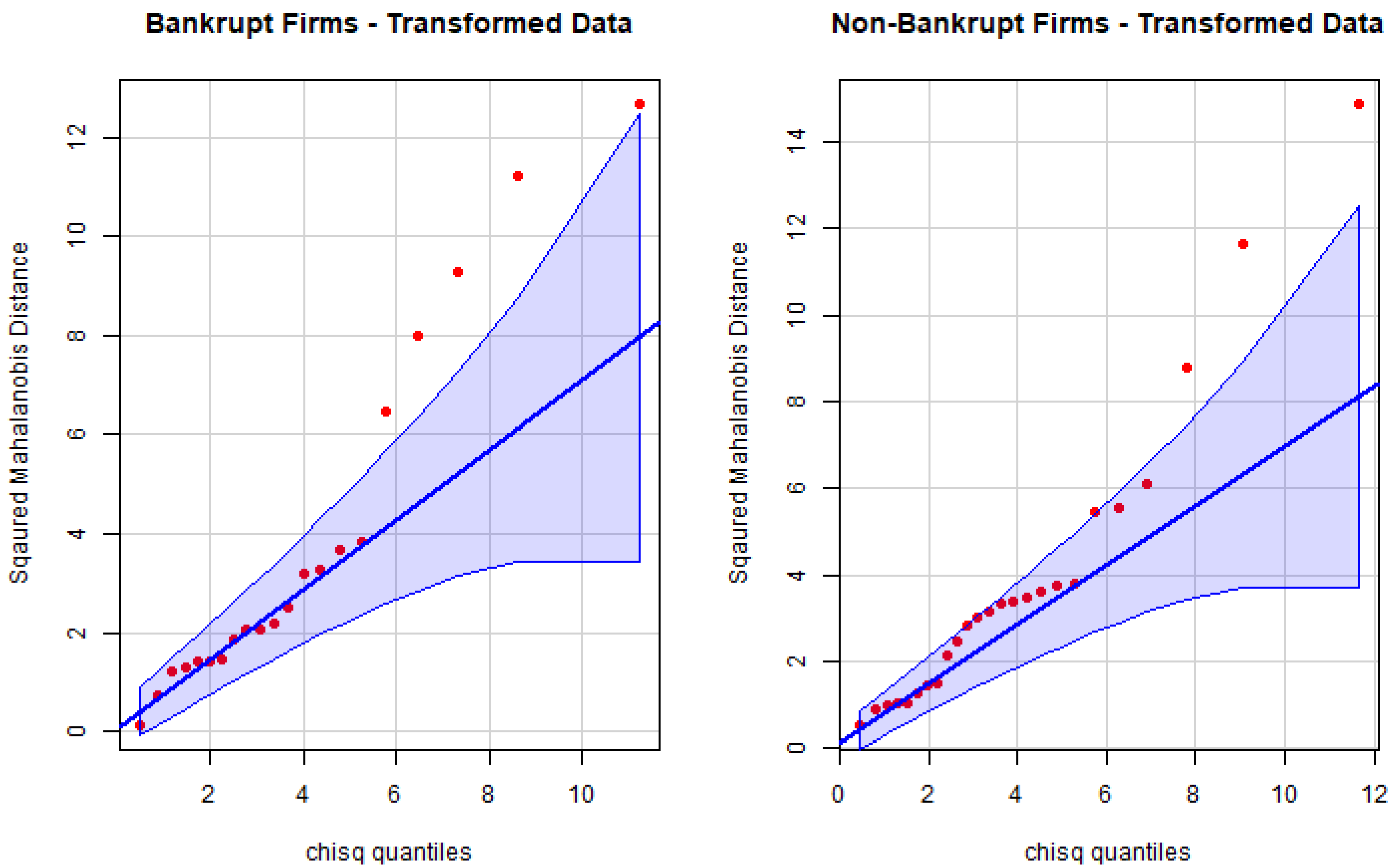
```
MVN::mvn(data = My.data_trans4[My.data_trans4$y == 0,-5], mvnTest = "royston",
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test       H  p value MVN
## 1 Royston 4.99441 0.200548 YES
##
## $univariateNormality
##           Test   Variable Statistic   p value Normality
## 1 Shapiro-Wilk CFTD_Trans    0.9637    0.5927    YES
## 2 Shapiro-Wilk NITA_Trans    0.9571    0.4596    YES
## 3 Shapiro-Wilk CATL_Trans    0.9474    0.3045    YES
## 4 Shapiro-Wilk CANS_Trans    0.9186    0.0813    YES
```

Multivariate Normality Accepted !

Using Multivariate Yeo-Johnson transformation :

Finding Optimum lambda | For Financially sound Firms | For Bankrupt Firms | **Chisqaure Plot**



Analysis Using All Transformed Variables:

Box-M Test QDA Performance

```r
heplots::boxM(as.matrix(My.data_trans4[,-5]) ~ as.factor(y),data = My.data_trans4)
```

```
##
##     Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 35.987, df = 10, p-value = 8.462e-05
```

Analysis Using All Transformed Variables:

Box-M Test | QDA | Performance

```
qda(My.data_trans4[,-5],My.data_trans4$y)
```

```
## Call:
## qda(My.data_trans4[, -5], My.data_trans4$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##    CFTD_Trans  NITA_Trans CATL_Trans CANS_Trans
## 0 -0.06391799 -0.04291975   1.169570  0.5162760
## 1  0.24660291  0.06830228   2.028308  0.4970221
```

Analysis Using All Transformed Variables:

Box-M Test | QDA | **Performance**

Training Set Performance

```
table(Actual = My.data[,5], Predicted = predict(qda(My.data_trans4[,-5],My.data_trans4$y))$class)
```

```
##        Predicted
## Actual  0  1
##       0 19  2
##       1  1 24
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], Qda_Model.4$class)
```

```
## [1] 0.1521739
```

Less than the all previous cases !

Is their any better possible classifier with less variables :

- Already discussed some discriminant rules after dropping variables.
- Less number of variables in a model is always good!
- Unless and until we are sacrificing much on misclassification error rate.
- Now, let us see after transformation how are the performances of some other rules.
- Here, we will judge based on Leave-one out cross-validated estimate of actual error rate.
- First, let's drop one Variable at a time!

Transforming CFTD,NITA,CATL :

Finding Optimum lambda | For Bankrupt Firms | For Financially sound Firms

```
lambda.6 <- car::powerTransform(as.matrix(My.data1[,-c(4,5)]),family = "yjPower")
lambda.6
```

```
## Estimated transformation parameters
##      CFTD      NITA      CATL
## 1.1222010 5.4300750 0.6541604
```

Transforming CFTD,NITA,CATL :

Finding Optimum lambda | **For Bankrupt Firms** | For Financially sound Firms

- After transformation:

```
MVN::mvn(with(My.data[My.data$y == 0,],yjPower(cbind(CFTD,NITA,CATL),coef(lambda.6))), mvnTest =
          univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 2.705383 0.3148556 YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk CFTD^1.12    0.9611    0.5390    YES
## 2 Shapiro-Wilk NITA^5.43    0.9566    0.4510    YES
## 3 Shapiro-Wilk CATL^0.65    0.9457    0.2812    YES
```

Multivariate Normality Accepted !

Transforming CFTD,NITA,CATL :

Finding Optimum lambda | For Bankrupt Firms | **For Financially sound Firms**

- After transformation:

```
MVN::mvn(with(My.data[My.data$y == 1,],yjPower(cbind(CFTD,NITA,CATL),coef(lambda.6))), mvnTest =
         univariateTest = "SW", desc = FALSE)
```

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 6.616813 0.07627823 YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk CFTD^1.12    0.9438    0.1809    YES
## 2 Shapiro-Wilk NITA^5.43    0.9720    0.6952    YES
## 3 Shapiro-Wilk CATL^0.65    0.9225    0.0585    YES
```

Multivariate Normality Accepted !

Analysis taking CFTD, NITA, CATL(after transformation) :

Box-M Test | QDA | Performance

```
heplots::boxM(as.matrix(My.data_trans6[,c(1,2,3)]) ~ as.factor(y),data = My.data_trans6)
```

```
##
##      Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 27.858, df = 6, p-value = 9.994e-05
```

Analysis taking CFTD, NITA, CATL(after transformation) :

Box-M Test | QDA | Performance

```
qda(My.data_trans6[,c(1,2,3)], My.data_trans6[,5])
```

```
## Call:
## qda(My.data_trans6[, c(1, 2, 3)], My.data_trans6[, 5])
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD        NITA     CATL
## 0 -0.06652109 -0.04223014 1.147983
## 1  0.24068187  0.06865672 1.970263
```

Analysis taking CFTD, NITA, CATL(after transformation) :

Box-M Test QDA **Performance**

Training Set Performance

```
table(Actual = My.data_trans6[,5], Predicted = predict(qda(My.data_trans6[,c(1,2,3)], My.data_tra
```

```
##         Predicted
## Actual   0  1
##        0 17  4
##        1  2 23
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], Qda_Model.6$class)
```

```
## [1] 0.1521739
```

Same as taking all four variables !

Transforming CFTD,CATL,CANS:

To bring multivariate normality, we will use optimum λ =0.72 for transforming CATL.

For Bankrupt Firms | For Financially sound Firms

- After transformation:

```
MVN::mvn(My.data_trans2[My.data_trans2$y == 0,-c(2,5)],mvnTest = "royston",univariateTest = "SW",
         desc = F)
```

```
## $multivariateNormality
##      Test        H   p value MVN
## 1 Royston 4.605265 0.2059711 YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   CFTD       0.9582    0.4800    YES
## 2 Shapiro-Wilk   CATL       0.9487    0.3222    YES
## 3 Shapiro-Wilk   CANS       0.9372    0.1921    YES
```

Multivariate Normality Accepted !

Transforming CFTD,CATL,CANS:

To bring multivariate normality, we will use optimum λ =0.72 for transforming CATL.

For Bankrupt Firms | For Financially sound Firms

- After transformation:

```
MVN::mvn(My.data_trans2[My.data_trans2$y == 1,-c(2,5)],mvnTest = "royston",univariateTest = "SW",
         desc = F)
```

```
## $multivariateNormality
##      Test        H    p value MVN
## 1 Royston 7.449324 0.05910793 YES
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk   CFTD       0.9417    0.1620    YES
## 2 Shapiro-Wilk   CATL       0.9205    0.0527    YES
## 3 Shapiro-Wilk   CANS       0.9614    0.4429    YES
```

Multivariate Normality Accepted !

Analysis taking CFTD, CATL, CANS(after transformation) :

Box-M Test   QDA   Performance

```
heplots::boxM(as.matrix(My.data_trans2[,-c(2,5)]) ~ as.factor(y),data = My.data_trans2)
```

```
##
##      Box's M-test for Homogeneity of Covariance Matrices
##
## data:  Y
## Chi-Sq (approx.) = 16.082, df = 6, p-value = 0.01332
```

Analysis taking CFTD, CATL, CANS(after transformation) :

Box-M Test | QDA | Performance

```
qda(My.data_trans2[,-c(2,5)],My.data_trans2$y)
```

```
## Call:
## qda(My.data_trans2[, -c(2, 5)], My.data_trans2$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD     CATL     CANS
## 0 -0.06904762 1.185911 0.437619
## 1  0.23520000 2.072756 0.426800
```

Analysis taking CFTD, CATL, CANS(after transformation) :

Box-M Test QDA **Performance**

Training Set Performance

```
table(Actual = My.data_trans2[,5], Predicted = predict(qda(My.data_trans2[,-c(2,5)],My.data_trans
```

```
##        Predicted
## Actual  0  1
##       0 19  2
##       1  1 24
```

AER Estimate (Cross Validated)

```
aer(My.data_trans2[,5], Qda_Model.7$class)
```

```
## [1] 0.1304348
```

Even, Better than including all four transformed variables !

Results of other case :

- When we tried to transform NITA, CATL, CANS, neither Univariate nor Multivariate transformations help!

Table of results of taking 2 variables at a time: (Which were not discussed previously)

Subsets	Transformation	Box-M p-Value	QDA CV AER estimate
CFTD & NITA	Multivariate	0.001785	0.2391304
CFTD & CATL	Not possible	NA	NA
NITA & CATL	Multivariate	0.015010	0.1304348
CATL & CANS	Univariate	0.002709	0.1521739

Only transformed NITA & transformed CATL is producing the lowest estimate of AER amongst all!

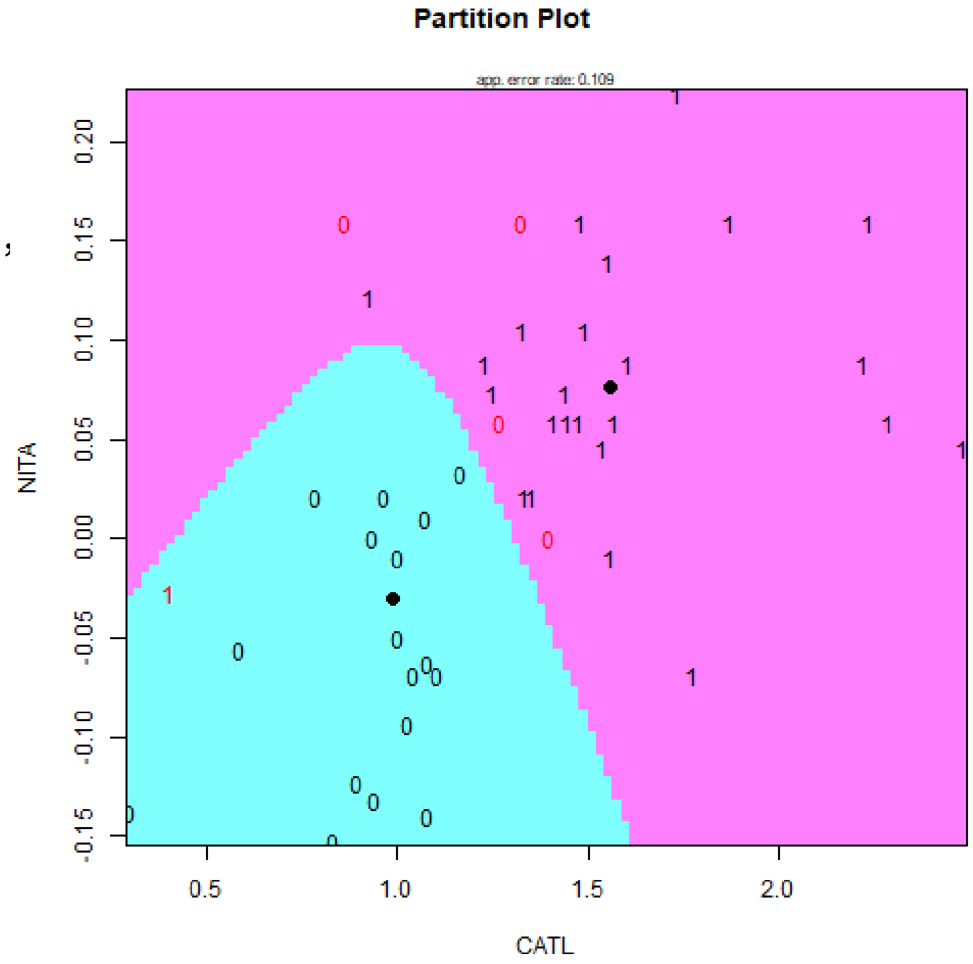
Optimum λ for this transformation is

```
## Estimated transformation parameters
##      NITA      CATL
## 7.5873497 0.3379119
```

Analysis Using Transformed NITA and Transformed CATL:

```
qda(My.data_trans51[,c(2,3)], My.data_trans51
```

```
## Call:
## qda(My.data_trans51[, c(2, 3)], My.data_trans51[,
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          NITA      CATL
## 0 -0.02994461 0.9864808
## 1  0.07632512 1.5606233
```

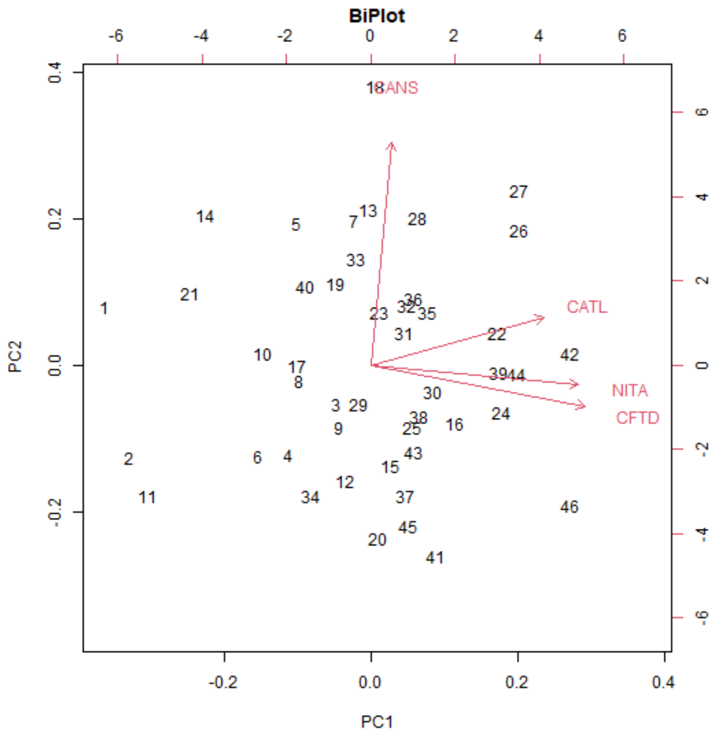


Principal Component Analysis :

PCA   Plots

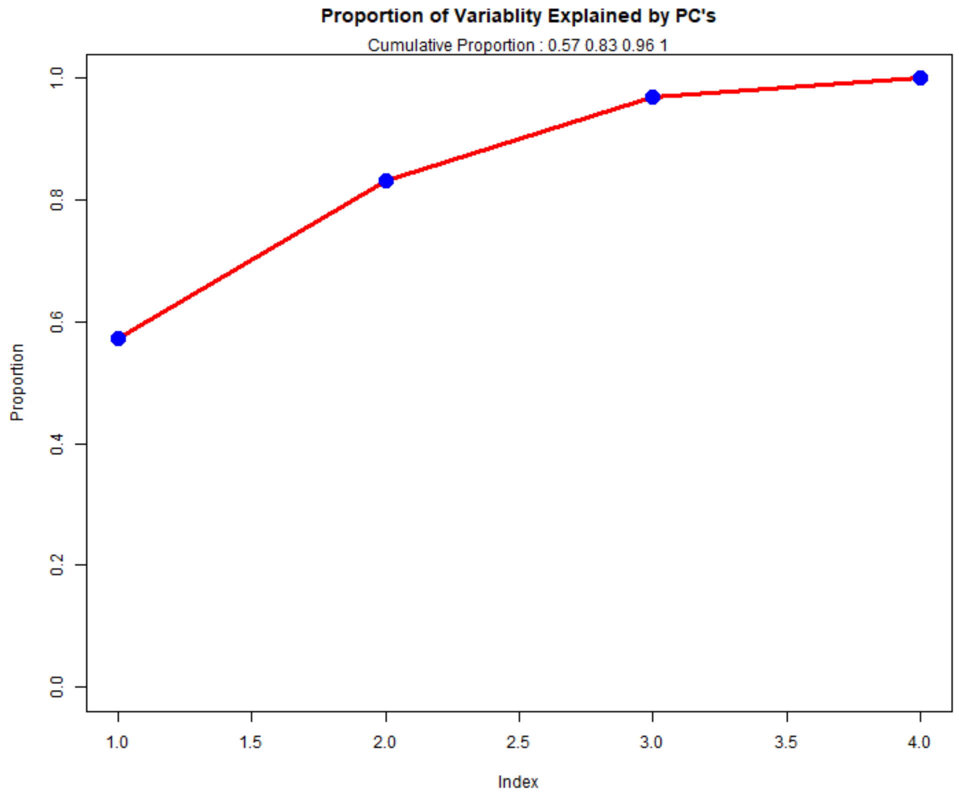
```
pca <- prcomp(My.data[,-5],scale = T)
pca
```

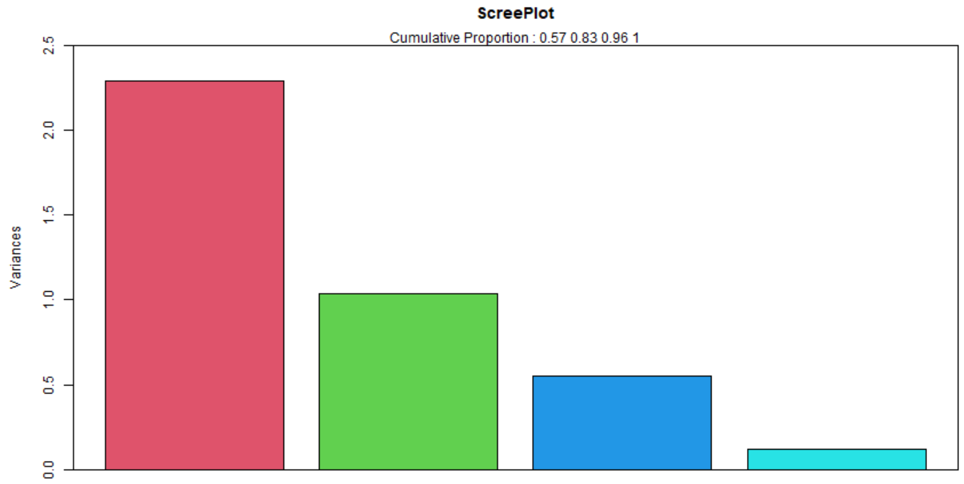
```
## Standard deviations (1, .., p=4):
## [1] 1.5121409 1.0187432 0.7437780 0.3498378
##
## Rotation (n x k) = (4 x 4):
##              PC1         PC2        PC3        PC4
## CFTD 0.62014111 -0.17691691  0.1967033 -0.7385345
## NITA 0.59989827 -0.08361003  0.4630444  0.6470868
## CATL 0.50193544  0.20371413 -0.8266216  0.1525063
## CANS 0.06006574  0.95927594  0.2521796 -0.1121929
```



Principal Component Analysis :

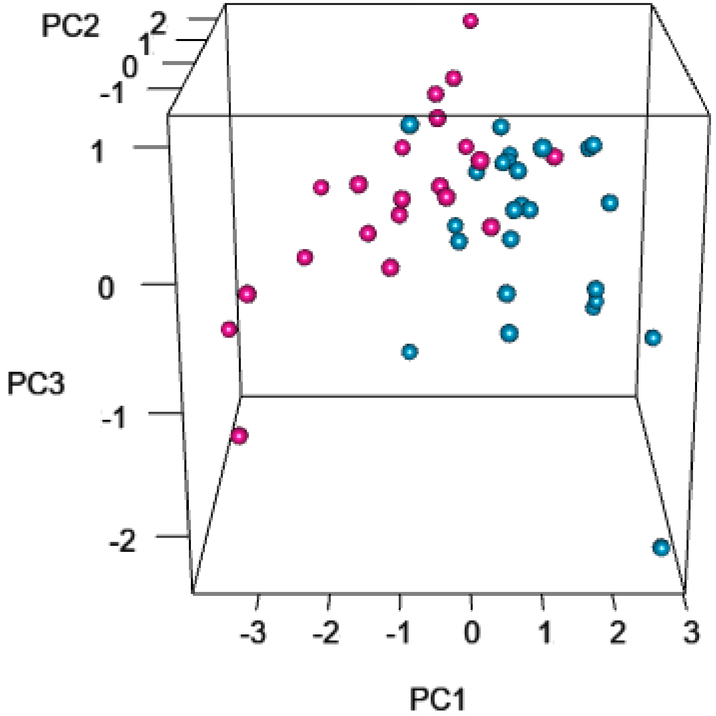
PCA Plots






```
## [1] Cum Prop. Explained: 0.57,0.83,0.96,1
```

3D Plot of first three Principal Components :



LDA & QDA based on all original variables :

LDA | Performance | QDA | Performance

```
lda(My.data[,-5],My.data$y)
```

```
## Call:
## lda(My.data[, -5], My.data$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD        NITA     CATL     CANS
## 0 -0.06904762 -0.08142857 1.366667 0.437619
## 1  0.23520000  0.05560000 2.593600 0.426800
##
## Coefficients of linear discriminants:
##              LD1
```

LDA & QDA based on all original variables :

LDA | **Performance** | QDA | Performance

Training Set Performance

```
table(Actual = My.data[,5], Predicted = predi
```

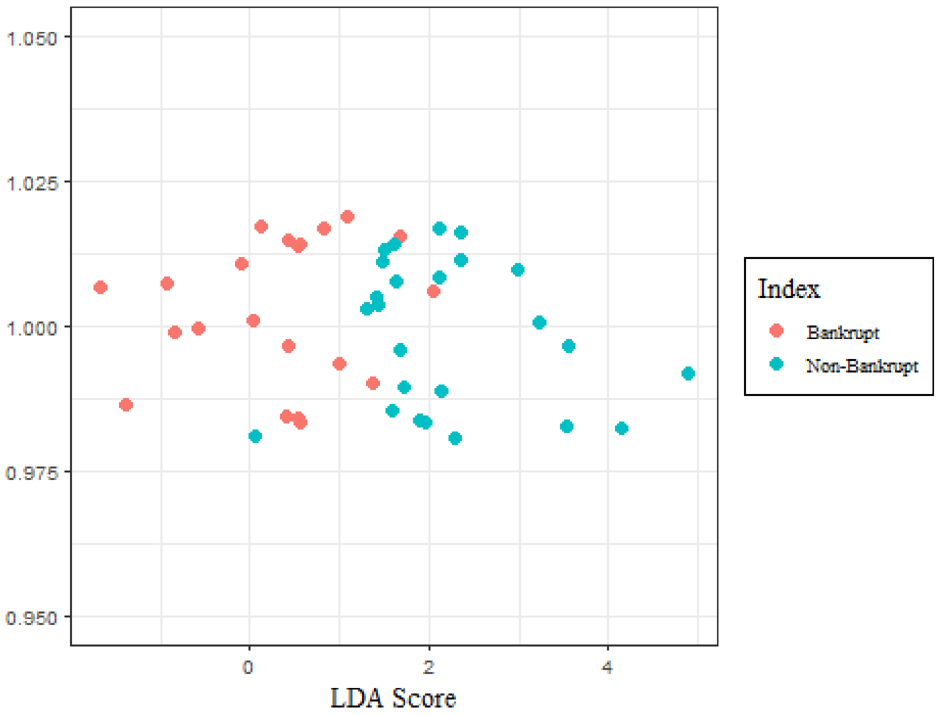
```
##        Predicted
## Actual  0  1
##       0 18  3
##       1  1 24
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], lda_Model.8$class)
```

```
## [1] 0.1304348
```

Plot of LDA Scores



LDA & QDA based on all original variables :

LDA Performance QDA Performance

```
qda(My.data[,-5],My.data$y)
```

```
## Call:
## qda(My.data[, -5], My.data$y)
##
## Prior probabilities of groups:
##         0         1
## 0.4565217 0.5434783
##
## Group means:
##          CFTD        NITA     CATL     CANS
## 0 -0.06904762 -0.08142857 1.366667 0.437619
## 1  0.23520000  0.05560000 2.593600 0.426800
```

LDA & QDA based on all original variables :

LDA | Performance | QDA | **Performance**

Training Set Performance

```
table(Actual = My.data[,5], Predicted = predict(qda(My.data[,-5],My.data$y))$class)
```

```
##         Predicted
## Actual   0  1
##        0 19  2
##        1  1 24
```

AER Estimate (Cross Validated)

```
aer(My.data[,5], qda_Model.8$class)
```

```
## [1] 0.1086957
```

Best till now!

Factor Analysis

Factor Analysis :

- It is used to identify the underlying structure or patterns in a set of variables and to reduce their complexity into a smaller number of factors or components.
- But to proceed with factor analysis, we need to first test whether the variables are actually related. i.e, whether the Correlation matrix of the variables is an Identity matrix.
- For that we will use **Bartlett Test of Sphericity**. The hypothesis is H_0: $R = I$ vs. H_1 : Not H_0 Where, R is the population correlation matrix.
- The test statistic is given by -

$$-log(det(R^*))\frac{(N-1-(2p+5))}{6}$$

Where, R^* is the sample correlation matrix. N is the sample size, and p is the number of variables.
It has asymptotic χ^2 distribution with d.f $\frac{p(p-1)}{2}$. It is sensitive to deviation from normality.

Factor Analysis :

Bartlett's Test | Principal Component Method | Maximum Likelihood Method | FA Diagram

- Bartlett's Test of Sphericity !

```
cortest.bartlett(My.data_trans4[,-5])
```

```
## $chisq
## [1] 88.4141
##
## $p.value
## [1] 6.467218e-17
##
## $df
## [1] 6
```

Thus, Bartlett's test is rejected !

Factor Analysis :

Bartlett's Test | **Principal Component Method** | Maximum Likelihood Method | FA Diagram

- Using Principal Component Method & Varimax Rotation :

```
fc <- fa((My.data_trans4[,-5]), nfactors = 2,rotate = "varimax",fm = "pa")
fc$loadings
```

```
##
## Loadings:
##              PA1    PA2
## CFTD_Trans  1.035 -0.122
## NITA_Trans  0.860
## CATL_Trans  0.598  0.343
## CANS_Trans         0.496
##
##                  PA1   PA2
## SS loadings    2.169 0.379
## Proportion Var 0.542 0.095
## Cumulative Var 0.542 0.637
```

Factor Analysis :

Bartlett's Test | Principal Component Method | **Maximum Likelihood Method** | FA Diagram

- Using Maximum Likelihood Method & Varimax Rotation :

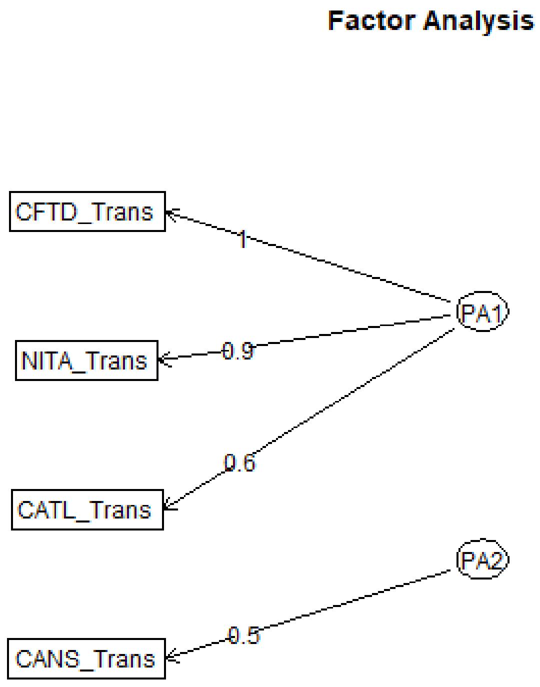
```
fc_n <- fa((My.data_trans4[,-5]), nfactors = 2,rotate = "varimax", fm = "ml")
fc_n$loadings
```

```
##
## Loadings:
##             ML1    ML2
## CFTD_Trans  0.993
## NITA_Trans  0.893
## CATL_Trans  0.598  0.152
## CANS_Trans         0.997
##
##                  ML1   ML2
## SS loadings    2.143 1.026
## Proportion Var 0.536 0.257
## Cumulative Var 0.536 0.792
```

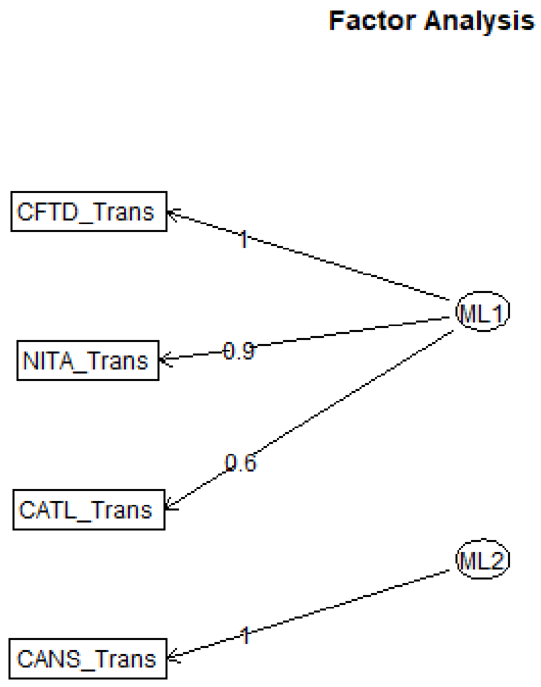
Factor Analysis :

Bartlett's Test | Principal Component Method | Maximum Likelihood Method | **FA Diagram**

For Principal Component Method :



For Maximum Likelihood Method :



Rotation Does not Change Fitted-Matrix :

Fitted-Matrix Graphical Illustration

Fitted Matrix with no rotation :

```
##            CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
## CFTD_Trans      1.000      0.889      0.579     -0.063
## NITA_Trans      0.889      1.000      0.531      0.008
## CATL_Trans      0.579      0.531      1.000      0.170
## CANS_Trans     -0.063      0.008      0.170      1.000
```

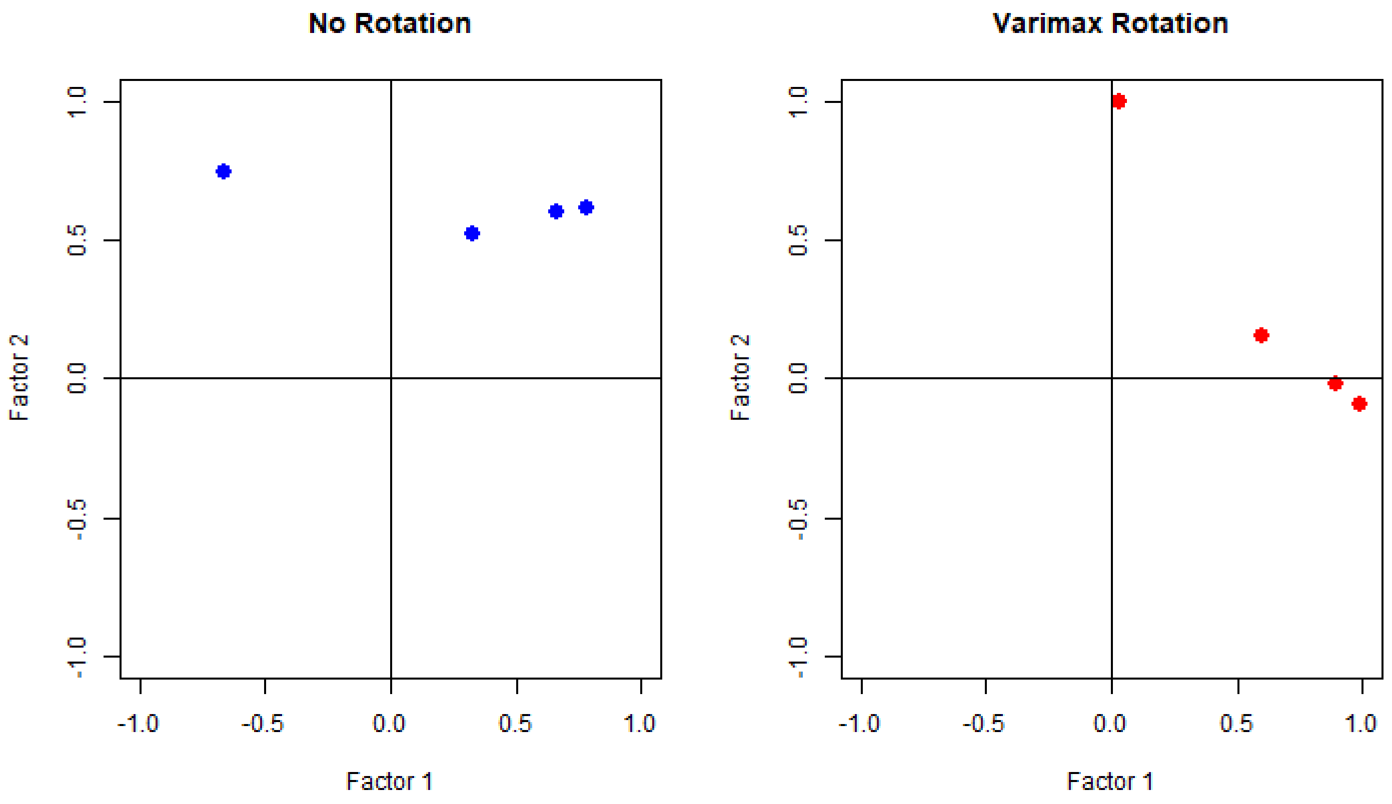
Fitted Matrix with Varimax rotation :

```
##            CFTD_Trans NITA_Trans CATL_Trans CANS_Trans
## CFTD_Trans      1.000      0.889      0.579     -0.063
## NITA_Trans      0.889      1.000      0.531      0.008
## CATL_Trans      0.579      0.531      1.000      0.170
## CANS_Trans     -0.063      0.008      0.170      1.000
```

Exactly Same !

Rotation Does not Change Fitted-Matrix :

Fitted-Matrix | Graphical Illustration



Further Exploration

Logistic regression :

- In LDA, QDA, we assume that **X** has mixture gaussian distribution and groupwise it has multivariate normal distribution.

- But in Logistic regression, we assume $X_{p\times 1}$ to be non-stochastic and we model

$$P_r(Y = 1|x_1, x_2, \ldots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

  Where, $\beta_0, \beta_1, \ldots, \beta_p$ are the parameters of the model.

Fitting Logistic Regression Model :

Fitted Model | Model Evaluation

```
Logistic_Model.10 <- glm(y ~.,data = My.data,family = binomial(link = "logit"))
summary(Logistic_Model.10)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial(link = "logit"), data = My.data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.30416  -0.44545   0.00725   0.49102   2.62396
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.320      2.366  -2.248  0.02459 *
## CFTD           7.138      6.002   1.189  0.23433
## NITA          -3.703     13.670  -0.271  0.78647
```

Fitting Logistic Regression Model :

Fitted Model | Model Evaluation

Taking 0.5 as threshold value !

Training Set Performance

```
table(Actual= My.data$y,Predicted=ifelse(predict.glm(Logistic_Model.10,type = "response") > 0.5,1
```

```
##         Predicted
## Actual   0  1
##       0 18  3
##       1  1 24
```

Error Rate Estimate (Cross Validated)

```
## [1] 0.1086957
```

Again, we are getting 10.86% Error Rate estimate !

Profile Analysis:

- Profile Analysis is a multivariate data analysis technique that is applicable to situations in which p treatments are administrated to two or more groups of subjects.
- The question of equality of mean vectors is divided into several specific questions such as

  1.Are the population profiles parallel?

  2.Are they coincident? (Assuming they are parallel)

  3.Are the profiles level? (Assuming they are coincident)

- **Assumptions:**

  > - The test scores should have a multivariate normal distribution.

  - We can transform the data to retain multivariate normality

  > - Homogeneity of the variance covariance matrix of test scores.

  - Box-M Test rejected homogeneity assumption.

  > - **So,We cannot perform Profile Analysis here !!!**

Summary :

- From EDA we have seen that, CFTD, NITA and CATL are well separating bankrupt firms from financially sound firms. From Factor analysis, we have got that these three are contributing to the first factor and CANS is contributing to the second factor.

- Also from EDA, we have seen that CFTD and NITA are very highly correlated.

- Plotting first three principal components, we visualized that the data is well separated, so we applied LDA or QDA even without multivariate normality.

- Finally, we have seen QDA to the original data and Logistic regression are yielding lowest AER(estimated)(11% approx.).

- Further, if we only take transformed NITA and CATL, then also we are not sacrificing much on AER(estimated)(13% approx.).

Thank You