1. Read the CSV file & Apply the summary command)

Dt<-read.csv(file="C:\\Users\\tejve\\Desktop\\Arrowsmith.csv",head=TRUE,sep=",",skip = 4)

attach(Dt)

summary(Dt)

The command will give all the statistical Parameters of all the variables in Data Frame. (Missing values) & Outliers (Boxplot)

```
> summary(Dt)
                               Arrowsmith.search   A.lit.size      C.lit.size          B.term           target
 APP vs reelin                          :1003     Min.   : 786   Min.   : 493   abnormal  :   6   Min.   :-2.0000
 Calpain vs PSD                         :3131     1st Qu.:3352   1st Qu.:2562   acid      :   6   1st Qu.:-1.0000
 magnesium vs migraine                  :1879     Median :3352   Median :2562   activation:   6   Median :-1.0000
 mGluR5 vs lewy bodies                  : 820     Mean   :3935   Mean   :2970   active    :   6   Mean   :-0.9714
 NO and mitochondria vs PSD             : 584     3rd Qu.:5122   3rd Qu.:3205   activity  :   6   3rd Qu.:-1.0000
 retinal detachment vs aortic aneurysm:2294     Max.   :6238   Max.   :5687   adult     :   6   Max.   : 3.0000
                                                                                (Other)   :9675
       nA                 nC            nof.MeSH.in.common nof.semantic.categories cohesion.score      n.in.MEDLINE
 Min.   :   1.00   Min.   :   1.000   Min.   :    0      Min.   : 0.0            Min.   :0.03532   Min.   :     2
 1st Qu.:   1.00   1st Qu.:   1.000   1st Qu.:    0      1st Qu.: 1.0            1st Qu.:0.08257   1st Qu.:  1484
 Median :   2.00   Median :   2.000   Median :    2      Median : 1.0            Median :0.12299   Median :  7184
 Mean   :  12.56   Mean   :   8.502   Mean   : 7882      Mean   : 1.5            Mean   :0.13407   Mean   : 27299
 3rd Qu.:   7.00   3rd Qu.:   5.000   3rd Qu.:    6      3rd Qu.: 2.0            3rd Qu.:0.17463   3rd Qu.: 26387
 Max.   :5120.00   Max.   :5686.000   Max.   :99999      Max.   :14.0            Max.   :0.99990   Max.   :932232
```

```
 X1st.year.in.MEDLINE       pAC              on.medium.stoplist. on.long.stoplist.
 Min.   :1902         Min.   :0.0000000   Min.   :0.0000      Min.   :0.0000     M
 1st Qu.:1947         1st Qu.:0.0000294   1st Qu.:0.0000      1st Qu.:0.0000     N
 Median :1949         Median :0.0236043   Median :0.0000      Median :1.0000
 Mean   :1950         Mean   :0.2745940   Mean   :0.4548      Mean   :0.6568
 3rd Qu.:1952         3rd Qu.:0.5521481   3rd Qu.:1.0000      3rd Qu.:1.0000
 Max.   :9999         Max.   :1.0000000   Max.   :1.0000      Max.   :1.0000
```

 # Summary Statistics: (Before Transformation)

#1.

#Literature sizes should be comparable.

y_equals_x <- function(x) {x}

y_equals_x_by_2 <- function(x) {x/2}

Plot literature C size vs literature A size.

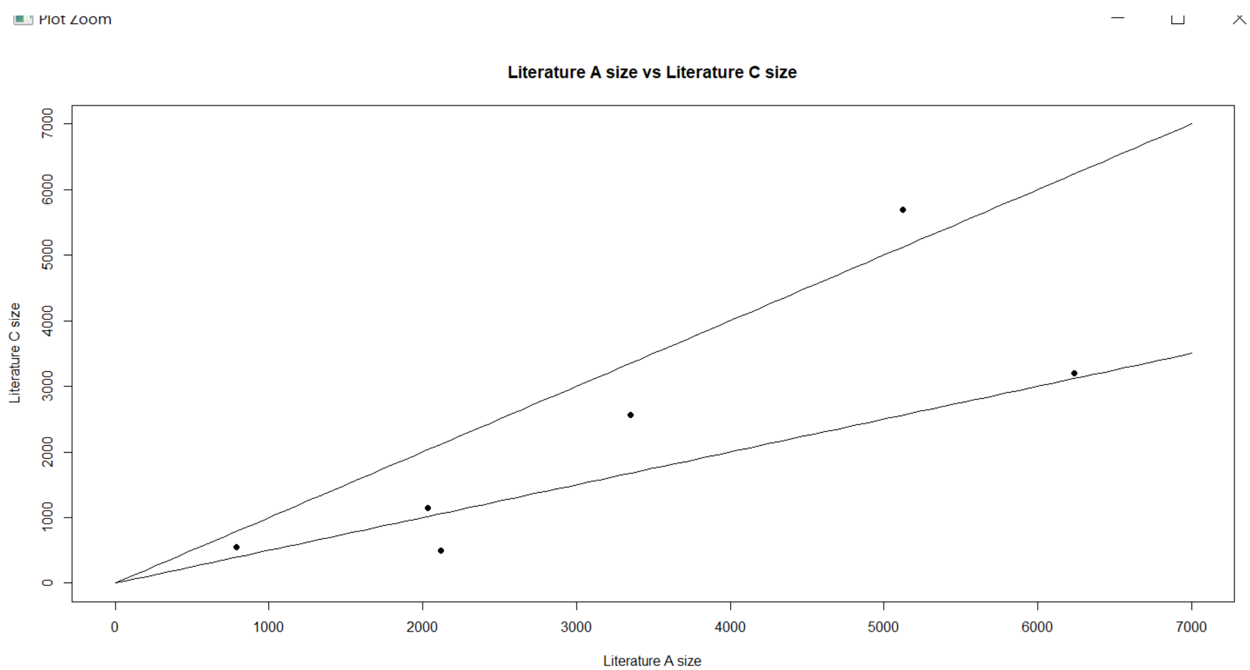
lit_A_size <- tapply(A.lit.size, Arrowsmith.search, mean) lit_C_size <- tapply(C.lit.size, Arrowsmith.search, mean)

plot(lit_A_size,lit_C_size,xlim = c(0,7000), ylim = c(0,7000), xlab = "Literature A size", ylab= "Literature C size",pch = 16, main = "Literature A size vs Literature C size")

curve(y_equals_x,from = 0, to = 7000, add = TRUE)

curve(y_equals_x_by_2,from = 0, to = 7000, add = TRUE)

#The literature (A and C) sizes are not comparable as we can see

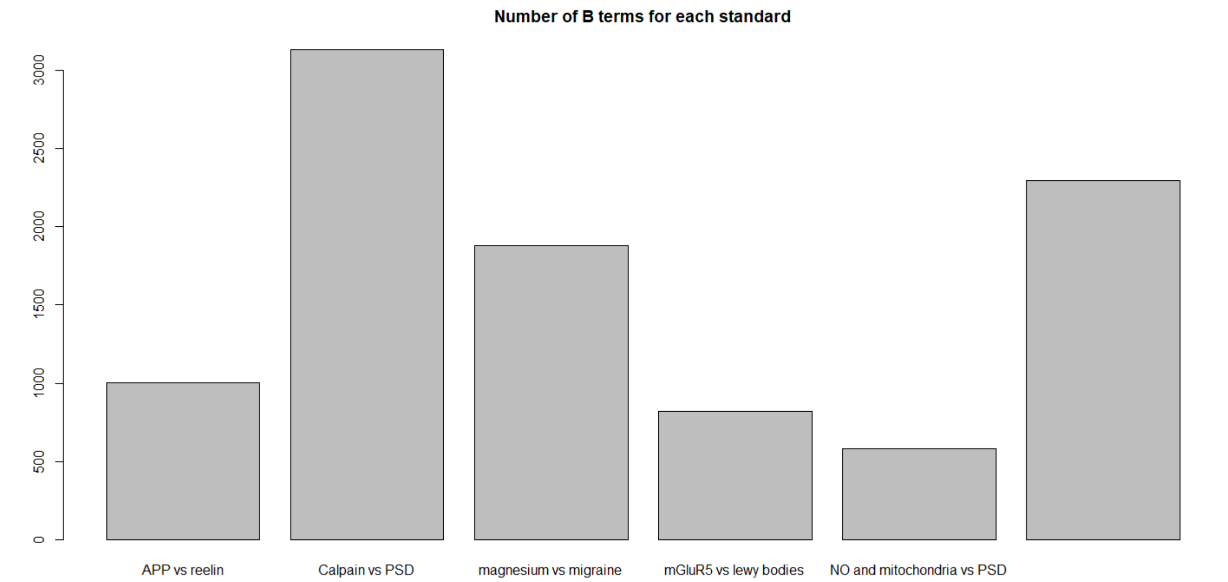


#2.

#Number of B terms for each standard:

barplot(tapply(B.term,Arrowsmith.search,length),main="Number of B terms for each standard")

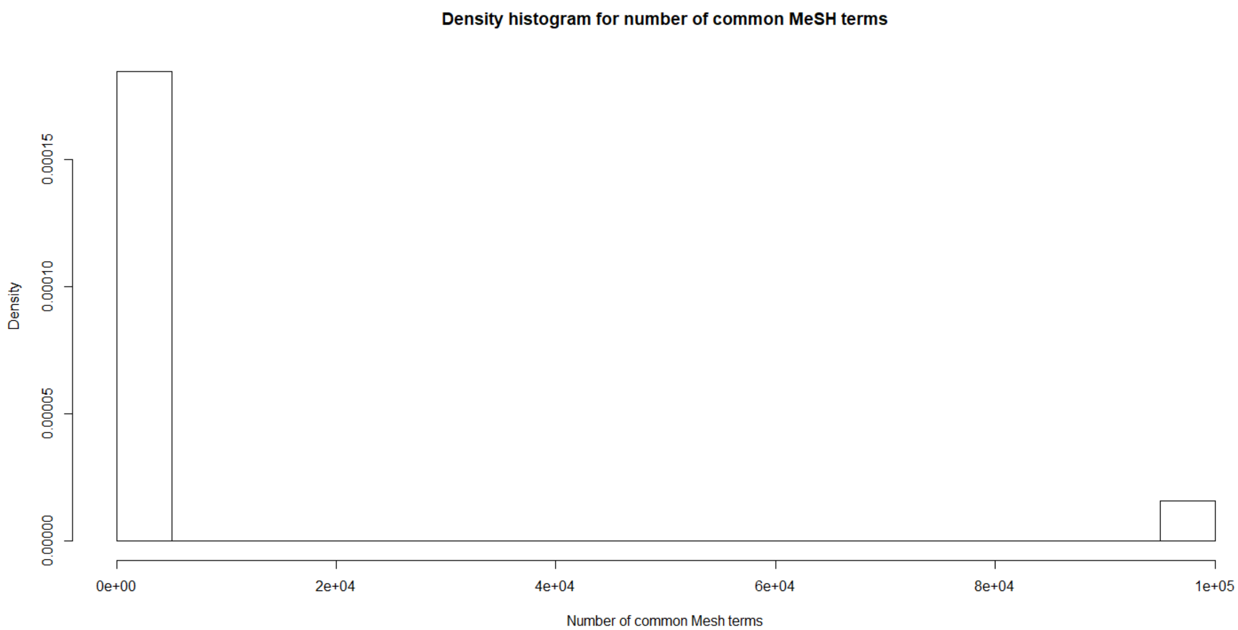
#There is quite a deep variation in number of B-terms for each standard



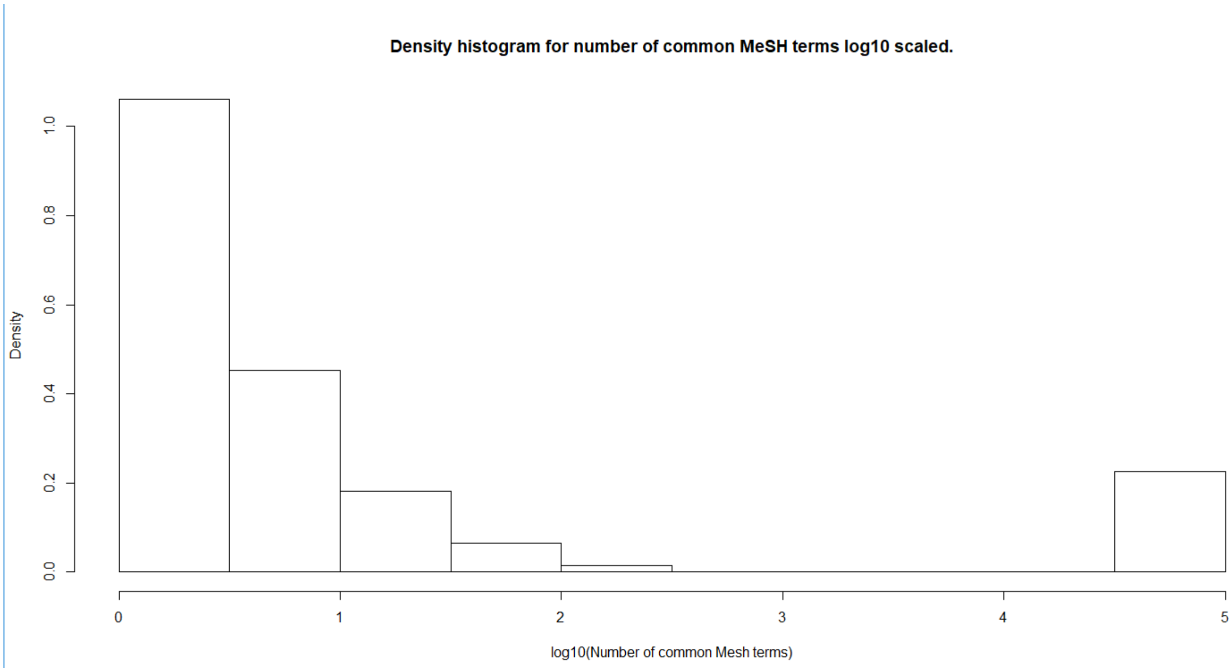
#3.

#Nof Mesh Terms in common:

hist(nof.MeSH.in.common,probability = TRUE, main = "Density histogram for number of common MeSH terms",xlab = "Number of common Mesh terms")



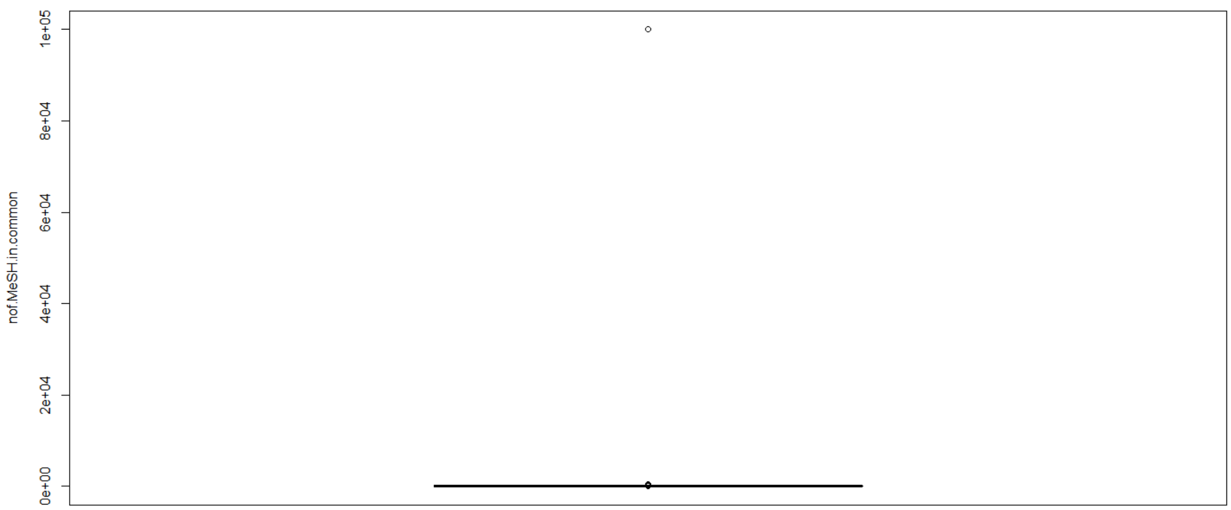
hist(log10(nof.MeSH.in.common),probability = TRUE, main = "Density histogram for number of common MeSH terms log10 scaled.",xlab = "log10(Number of common Mesh terms)")



boxplot(nof.MeSH.in.common,ylab="nof.MeSH.in.common")

#From the Box-plot we see that term has outliers.(points which outside the 1.5 times the Interquartile Range)

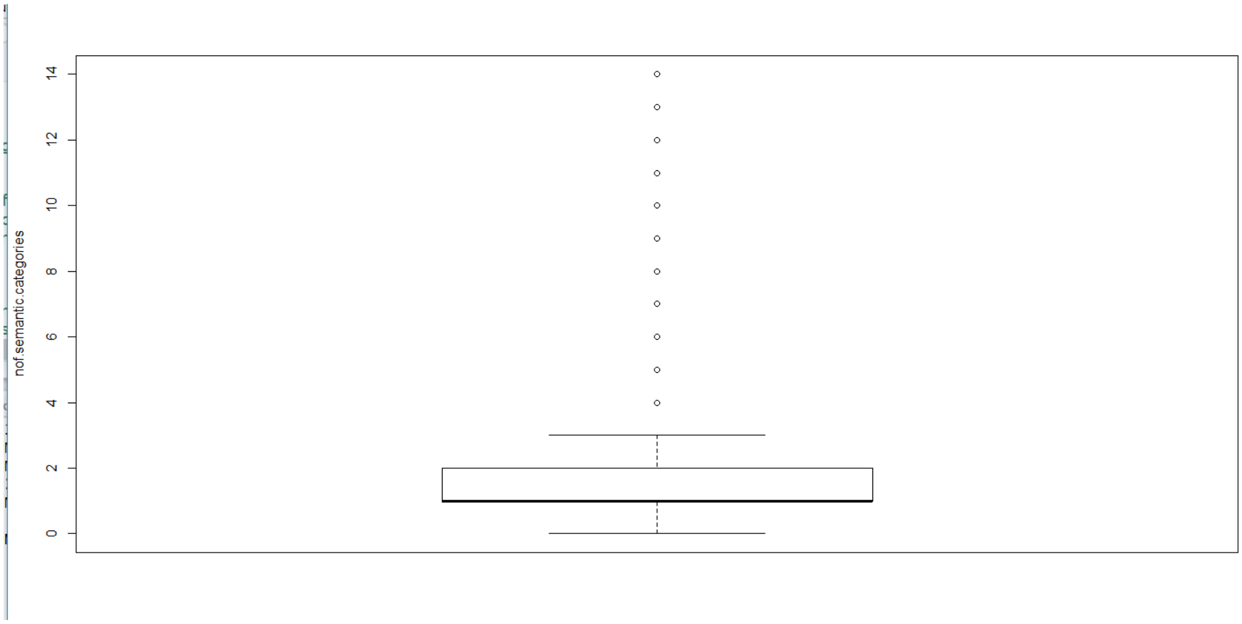
It also has missing values(99999)



#4. nof.semantic.categories

boxplot(nof.semantic.categories,ylab="nof.semantic.categories")

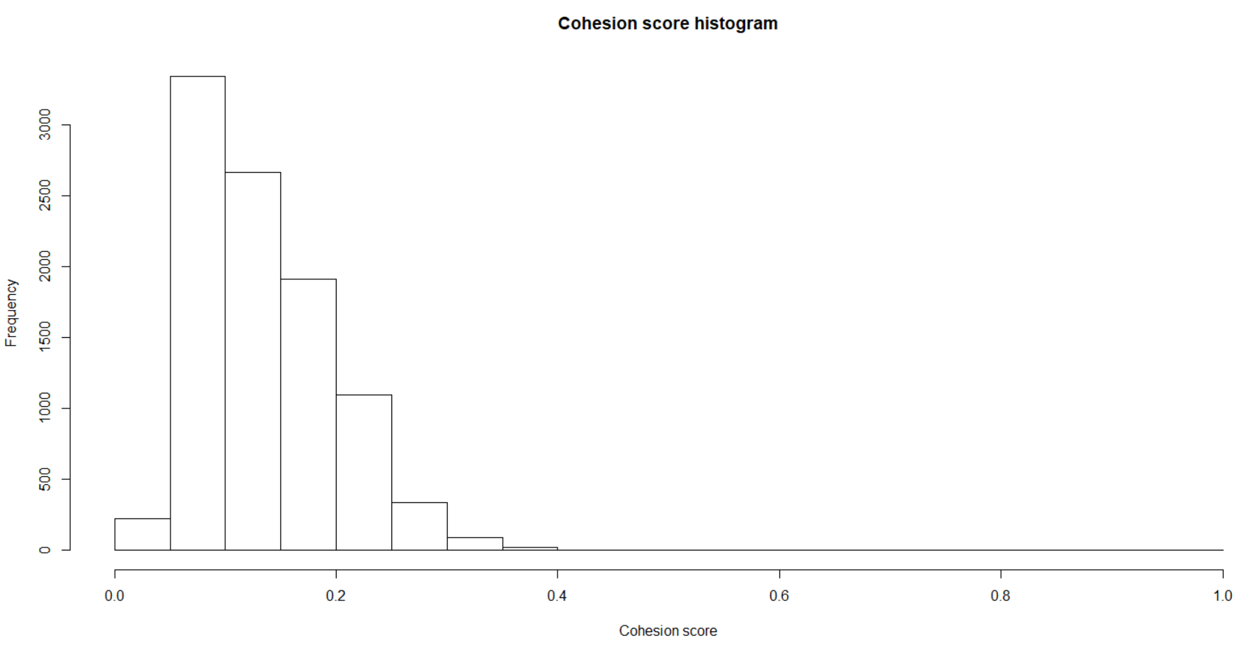
#From the boxplot The variable nof.semantic.Categories has outliers.(points which outside the 1.5 times the Interquartile Range)



5. Exploring cohesion score

hist(cohesion.score, main = "Cohesion score histogram",xlab = "Cohesion score")

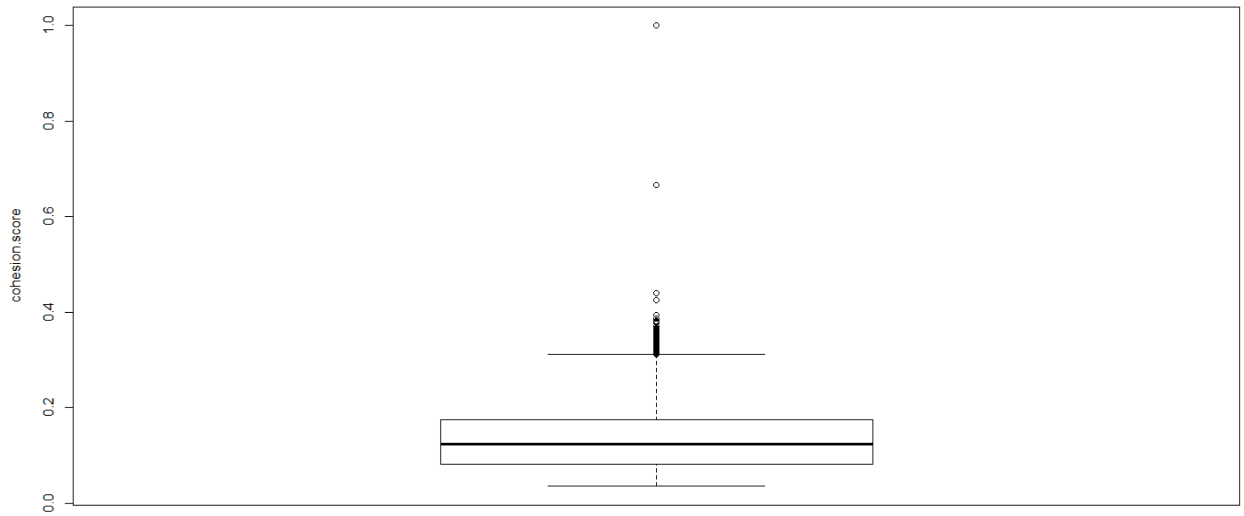
As seen in the histogram, number of values above 0.3 is very few. Hence, that might be the motive choosing 0.3 as the upper limit.



boxplot(cohesion.score,ylab="cohesion.score")

#From the boxplot we see that the variable has significant outliers.(points which outside the 1.5 times the Interquartile Range)

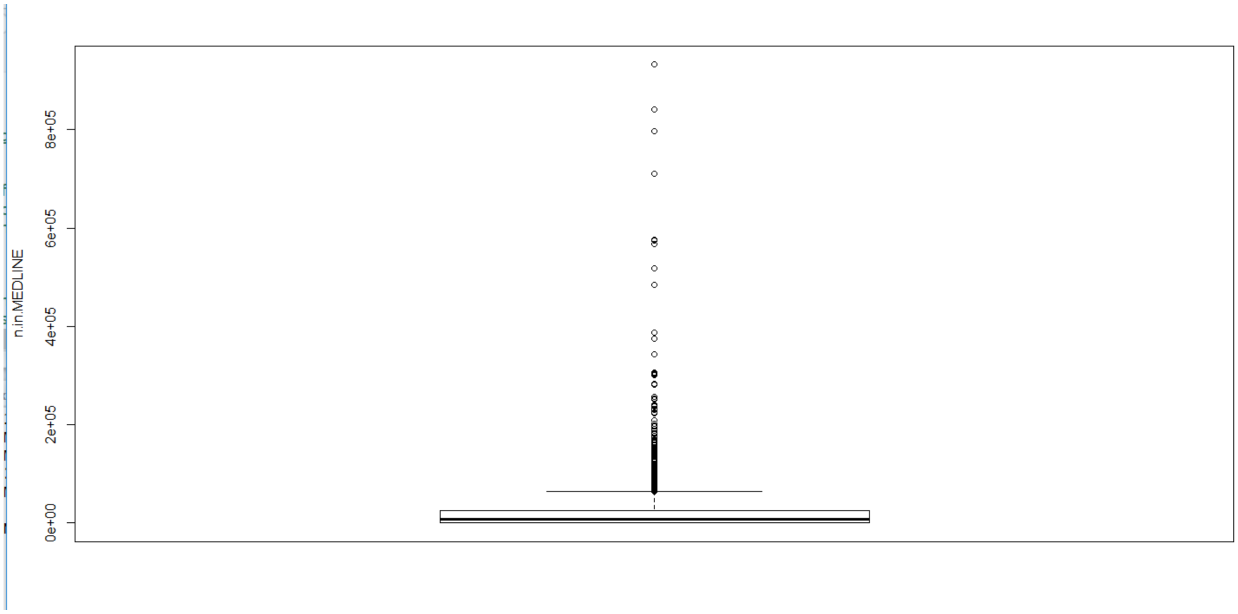
The variable cohesion.score has missing values(0.99990)



#6. n.in.MEDLINE

boxplot(n.in.MEDLINE,ylab="n.in.MEDLINE")

From the boxplot we see the variable n.in.MEDLINE has outlier.(points which outside the 1.5 times the Interquartile Range)

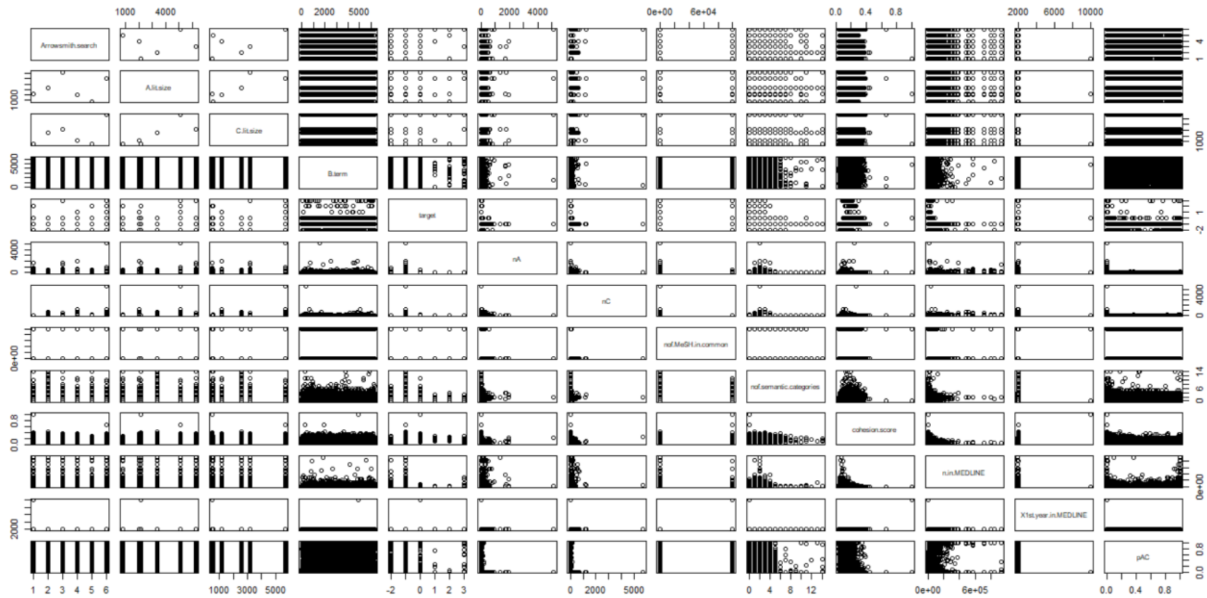


#7.

#Pairwise scatter plots.

plot(Dt[1:13])

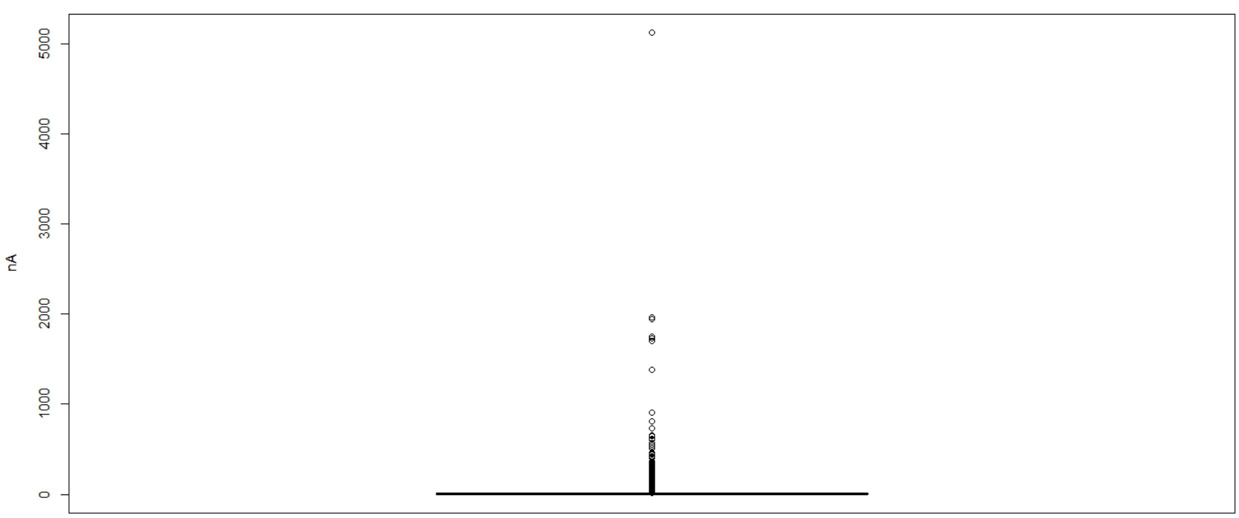
#There was no relation found among any pairs of variables, but just literature A and literature C size. However, these two are not used to calculate two different features.



#8 (Na)

boxplot(nA,ylab="nA")

From the boxplot the "nA" variable has outlier.(points which outside the 1.5 times the Interquartile Range)

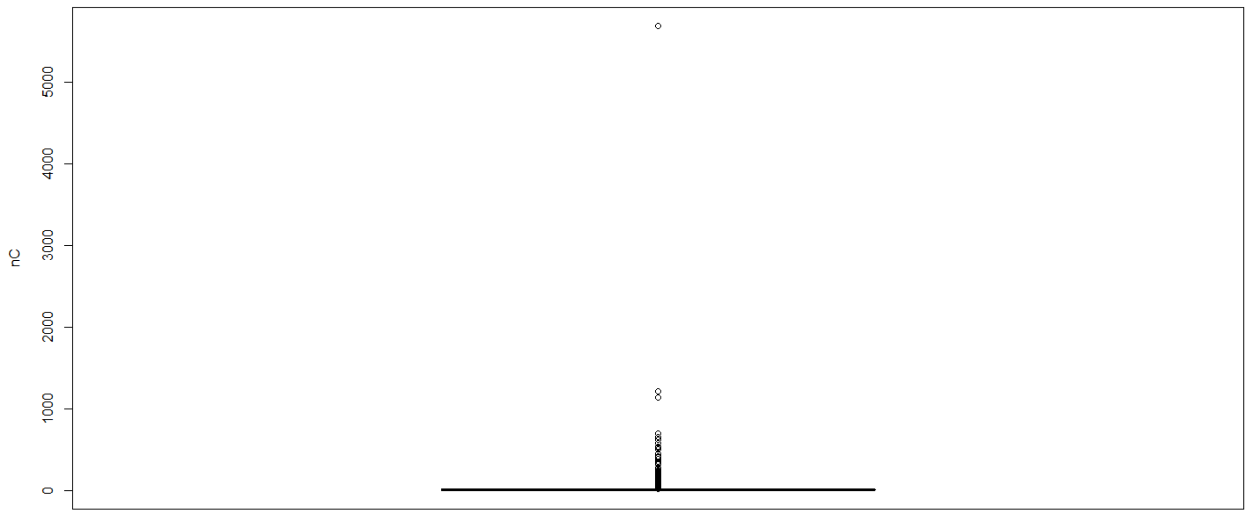


Max :1.0000000 Max :1.0000 Max :1.0000

#9 (Nc)

boxplot(nC,ylab="nC")

From the boxplot the "nC" variable has outlier.(points which outside the 1.5 times the Interquartile Range)



 # 10. A new feature. x_new

x_new <- -abs((nA/A.lit.size) -(nC/C.lit.size))

I think that none of the features capture intuition that if a B-term occurs a lot frequently in A but not in C or vice-versa, then it is less relevant. In other words, number of articles containing B-term should be in both literatures should be comparable.

The number of occurences in MEDLINE is a similar feature but that only captures that very common and very rare words in MEDLINE are less relevant. It doesn't capture the intuition that a B-term very frequent in one literature but very rare in other literature will mostly likely be irrelevant.

I am not sure if this will prove to be a useful parameter. It is even difficult to test because we are not provided with any test data. The best evaluation I can do is check for statistical significance based on the p-value

#After Transformation:

1.

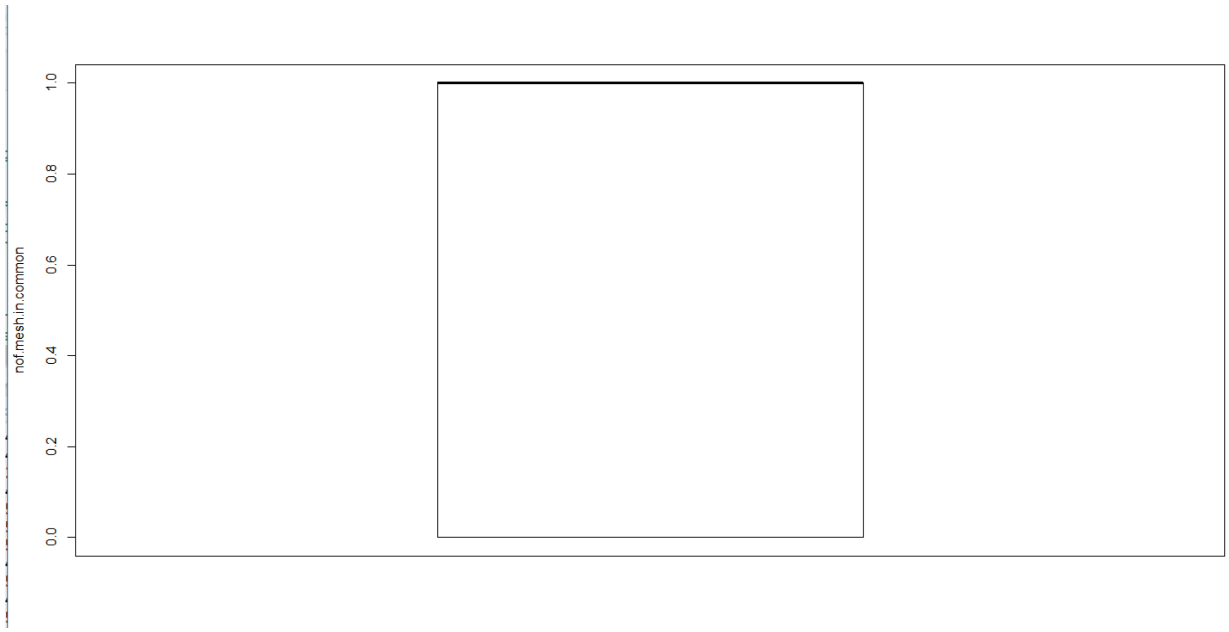
#X1 = 1 if (nA > 1 or A-lit size < 1000) and (nC > 1 or C-lit size < 1000), 0 otherwise

```
X1<-ifelse((nA > 1 | A.lit.size < 1000) & (nC > 1 | C.lit.size < 1000),1,0)
```

2. **Nof Mesh**

X2 = 1 if nof MeSH > 0 and < 99999, 0.5 if nof MeSH = 99999, 0 otherwise

```
X2<-c()
for (i in 1:length(nof.MeSH.in.common)) {
if (nof.MeSH.in.common[i] >0 & nof.MeSH.in.common[i] < 99999) {
  X2<-append(X2,1)
} else if (nof.MeSH.in.common[i] == 99999) {
  X2<-append(X2,0.5)
} else {  X2<-append(X2,0) }
}
# The variable nof.mesh.in.common has missing values before transformation,after transformation(X2)
there are no missing values.
boxplot(X2,ylab="nof.mesh.in.common")
#From the boxplot we see that there are no outliers.(points which outside the 1.5 times the
Interquartile Range)
```



3. **Nof Semantic Categories**

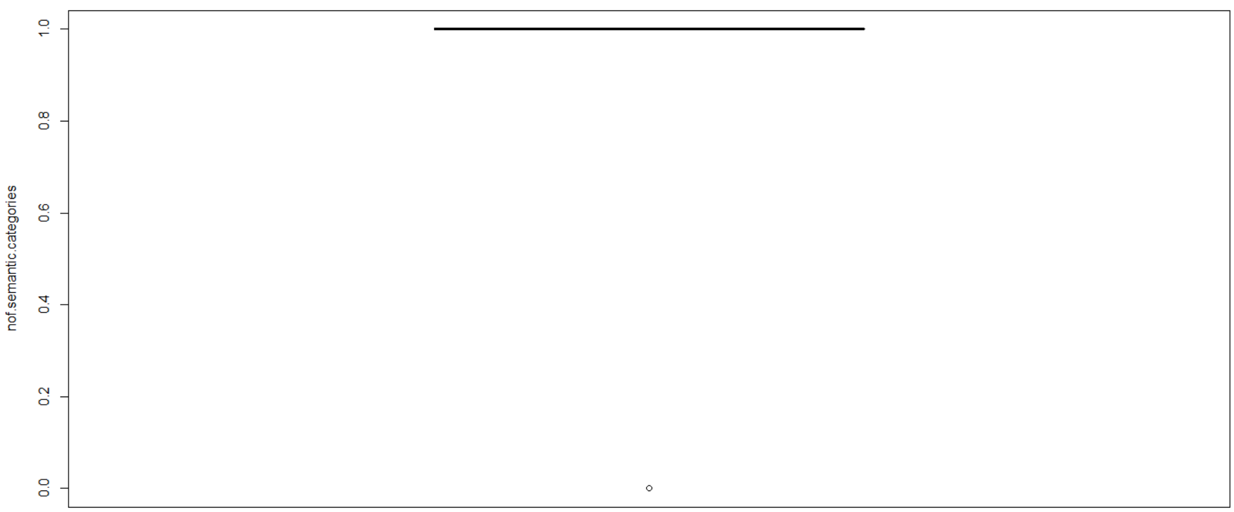
#X3 = 1 if nof semantic categories > 0, 0 otherwise

X3<-ifelse(nof.semantic.categories>0,1,0)

The variable nof.semantic.categories have outliers before transformation but after transformation it does not have.

boxplot(X3,ylab="nof.semantic.categories")

#From the boxplot we see that there are outliers.(points which outside the 1.5 times the Interquartile Range)



4. Cohesion score

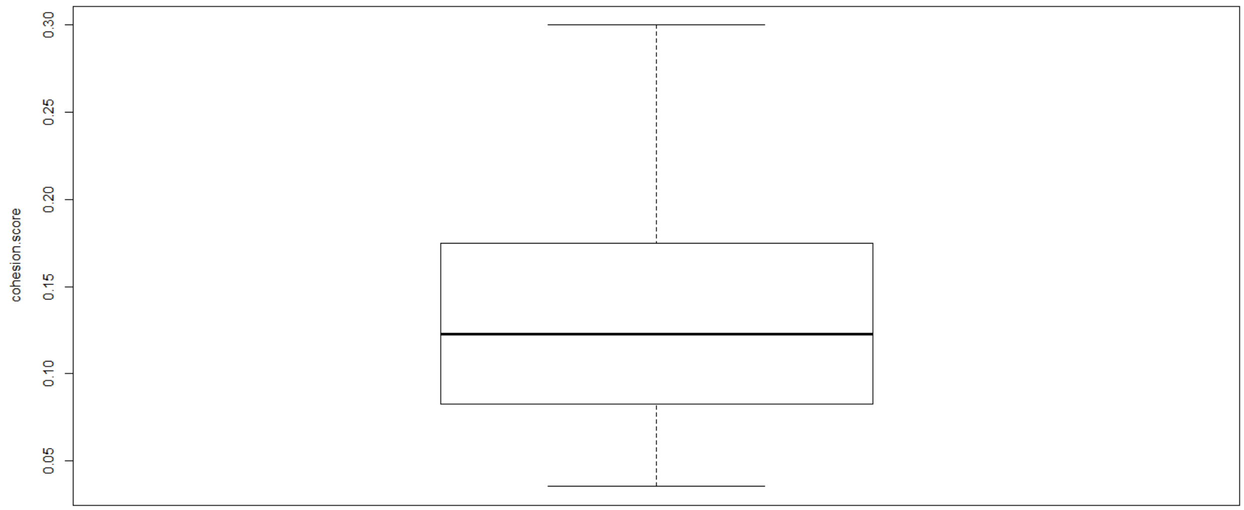
#X4 = cohesion score if cohesion score < 0.3, 0.3 otherwise

X4<-ifelse(cohesion.score<0.3,cohesion.score,0.3)

##The variable cohesion.score has missing values before transformation but after transformation there are no missing values.

boxplot(X4,ylab="cohesion.score")

#From the box plot we see that it does not have outliers.(points which outside the 1.5 times the Interquartile Range)



5.

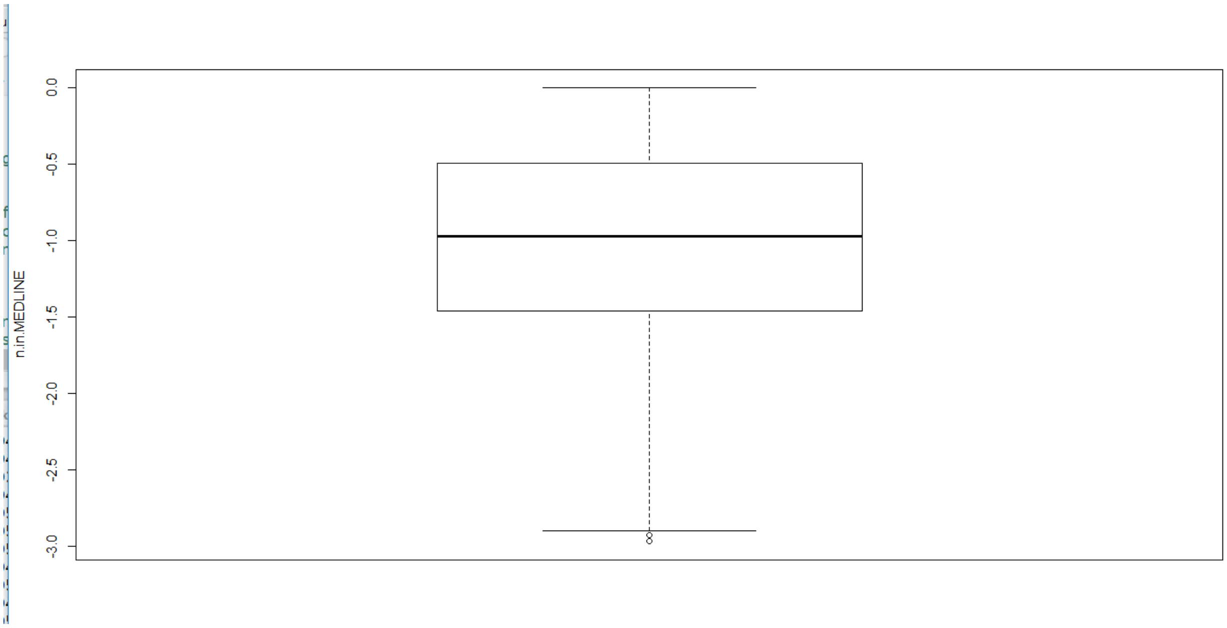
N in MEDLINE

#X5 = -|log10(n in MEDLINE) – 3|

X5<--abs(log10(n.in.MEDLINE)-3)

boxplot(X5,ylab="n.in.MEDLINE")

#From the box plot we see that n.in.MEDILINE has outliers.(points which outside the 1.5 times the Interquartile Range)



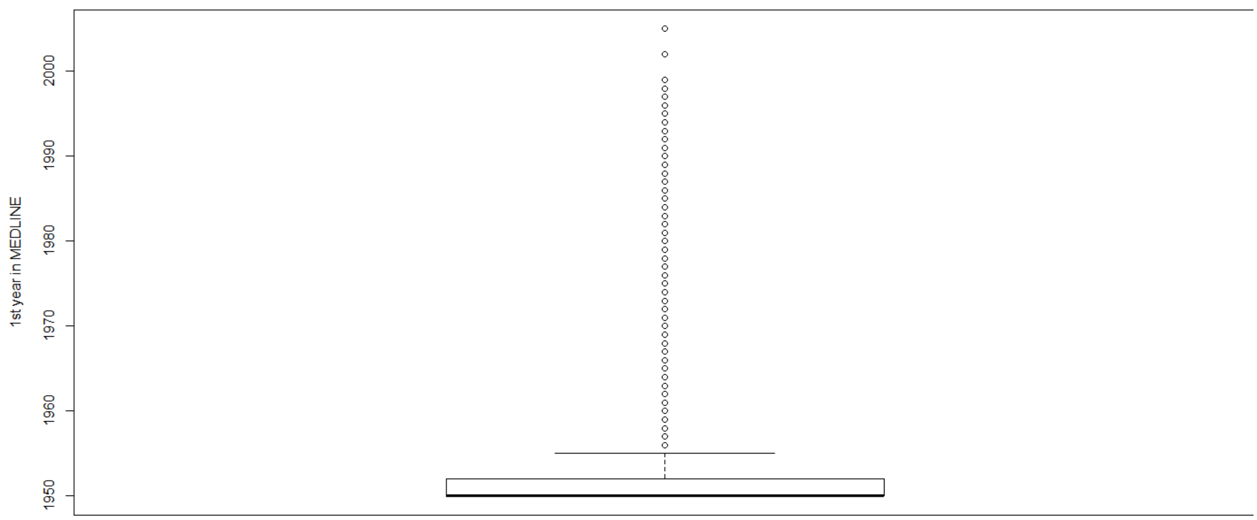
6. 1st year in MEDLINE

#X6 = max(min(1st year in MEDLINE,2005),1950)

X6<- pmax(pmin(X1st.year.in.MEDLINE,2005),1950)

boxplot(X6,ylab="1st year in MEDLINE")

#From the boxplot we see that it has too many outliers.(points which outside the 1.5 times the Interquartile Range)



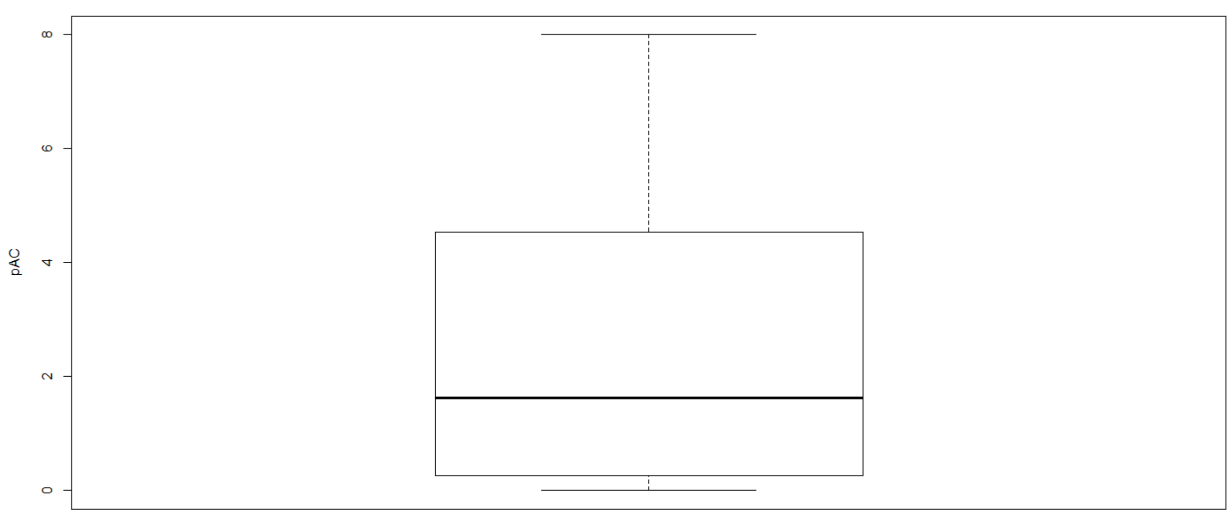
7.pAC

#X7 = min(8,-log10(pAC+0.000000001))

X7<-pmin(8,-log10(pAC+0.000000001))

boxplot(X7,ylab="pAC")

#From the box plot we see that X7 does not have any outliers.(points which outside the 1.5 times the Interquartile Range)



8.

The following are research literatures (A and C)

```
#I1 = 1 if Arrowsmith search = ‘retinal detachment’, 0 otherwise
```

```
I1<-ifelse(Arrowsmith.search=='retinal detachment vs aortic aneurysm',1,0)
I2<-ifelse(Arrowsmith.search=='NO and mitochondria vs PSD',1,0)
I3<-ifelse(Arrowsmith.search=='mGluR5 vs lewy bodies',1,0)
I4<-ifelse(Arrowsmith.search=='magnesium vs migraine',1,0)
I5<-ifelse(Arrowsmith.search=='Calpain vs PSD',1,0)
I6<-ifelse(Arrowsmith.search=='APP vs reelin',1,0)
```

9.

#The output (In logistic Regression)

```
#Y = 1 if target = 0 or 2, 0 otherwise
```

```
Y<-ifelse(target==0 | target==2 ,1,0)
```

10. Combining all the variables above:

```
new_frame<data.frame(X1=X1,X2=X2,X3=X3,X4=X4,X5=X5,X6=X6,X7=X7,I1=I1,I2=I2,I3=I3,I4=I4,I5=I5,I6=I6,Y=Y)
summary(new_frame)
```

```
> summary(new_frame)
       X1               X2              X3              X4                X5                  X6
 Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.03532   Min.   :-2.9695240   Min.   :1950
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.08257   1st Qu.:-1.4628917   1st Qu.:1950
 Median :1.0000   Median :1.000   Median :1.000   Median :0.12299   Median :-0.9739126   Median :1950
 Mean   :0.5092   Mean   :0.661   Mean   :0.788   Mean   :0.13353   Mean   :-1.0124482   Mean   :1955
 3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.17463   3rd Qu.:-0.4933186   3rd Qu.:1952
 Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :0.30000   Max.   :-0.0004341   Max.   :2005
       X7               I1               I2                I3                I4               I5
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.2579   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :1.6270   Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000
 Mean   :2.7400   Mean   :0.2362   Mean   :0.06014   Mean   :0.08444   Mean   :0.1935   Mean   :0.3224
 3rd Qu.:4.5316   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000
 Max.   :8.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

```
       I6                Y
 Min.   :0.0000   Min.   :0.00000
 1st Qu.:0.0000   1st Qu.:0.00000
 Median :0.0000   Median :0.00000
 Mean   :0.1033   Mean   :0.03357
 3rd Qu.:0.0000   3rd Qu.:0.00000
 Max.   :1.0000   Max.   :1.00000
```

#Logistic Regression

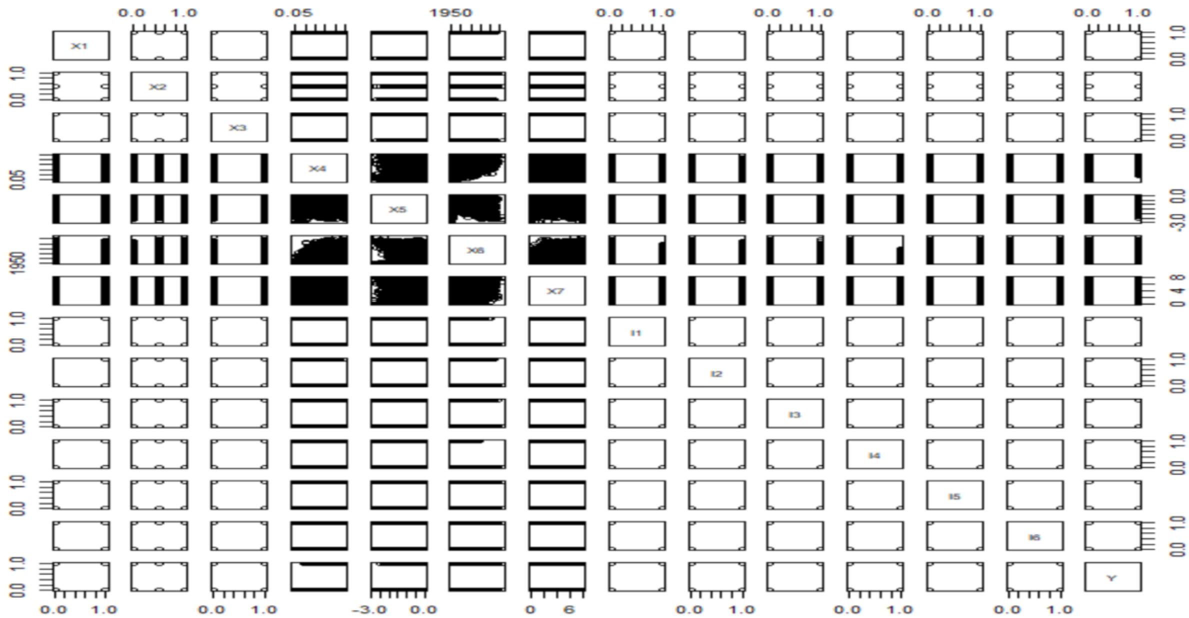
1.

Checking for assumptions:

Independence: As sample size is quite large, assuming that all observations are independent from each other.

Multicollinearity: I checked for corelation by pairs command (Scatter plot). There was none found .Hence, it can be concluded that there is no multicollinearity.

plot(new_frame[1:14])



2.

 # Weights' interpretation:

All weights are interpreted as how much would log(odds) change with one unit change in that feature provided everything else is constant.

For binary features, weight is interpreted as how much log(odds) will change if that feature is true.

3.

 # Model evalation techniques used:

1. Check p-values of all estimates.

2. Check null deviance of the model

4.

fit_model<glm(formula=Y~X1+X2+X3+X4+X5+X6+X7+I1+I2+I3+I4+I5+I6,family=binomial,data=new_frame)

summary(fit_model)

#I saw from the summary(fit_model) command ,all the weights related to my feature (x1 to x7) are related to the model described in Table S2 (in the supplemental data file) in the research paper.

```
Call:
glm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + I1 + I2 +
    I3 + I4 + I5 + I6, family = binomial, data = new_frame)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-1.7965  -0.2108  -0.1116  -0.0611   3.7272

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -86.14907   10.74423  -8.018 1.07e-15 ***
X1            0.73220    0.15558   4.706 2.52e-06 ***
X2            0.98770    0.24633   4.010 6.08e-05 ***
X3            1.31738    0.25819   5.102 3.35e-07 ***
X4           13.76594    1.24677  11.041  < 2e-16 ***
X5            0.58621    0.11460   5.115 3.13e-07 ***
X6            0.03957    0.00549   7.207 5.71e-13 ***
X7            0.18873    0.02509   7.521 5.45e-14 ***
I1            0.92686    0.23316   3.975 7.03e-05 ***
I2            1.38271    0.24258   5.700 1.20e-08 ***
I3            0.95634    0.22672   4.218 2.46e-05 ***
I4            0.68351    0.25120   2.721  0.00651 **
I5           -1.10016    0.21004  -5.238 1.63e-07 ***
I6                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2853.9  on 9710  degrees of freedom
Residual deviance: 1997.5  on 9698  degrees of freedom
AIC: 2023.5
```

#The null deviance decrease from 2853.9 to 1997.5 ,so the model is better fitted by logistic regression.

#I saw from the summary(fit_model) command ,all the weights related to my feature (x1 to x7) are related to the model described in Table S2 (in the supplemental data file) in the research paper.

All the P -values of the features(estimate) (X1 to X7) is less than 0.05 ,so the model is statistically significant .There is a relationship between output (Y) and the features (X1 to X7).

The estimate(Weight) of X1 is "0.73220"->binary feature, weight is interpreted as how much log(odds) will change if that feature is true

#The estimate (Weight) of X2 is "0.98770"->log(odds) change with one unit change in that feature provided everything else is constant.

#The estimate (Weight) of X3 is "1,31738"->binary feature, weight is interpreted as how much log(odds) will change if that feature is true

#The estimate (Weight) of X4 is "13.76594"->log(odds) change with one unit change in that feature provided everything else is constant.

#The estimate (Weight) of X5 is "0.58621"->log(odds) change with one unit change in that feature provided everything else is constant.

#The estimate (Weight) of X6 is "0.03957"->log(odds) change with one unit change in that feature provided everything else is constant.

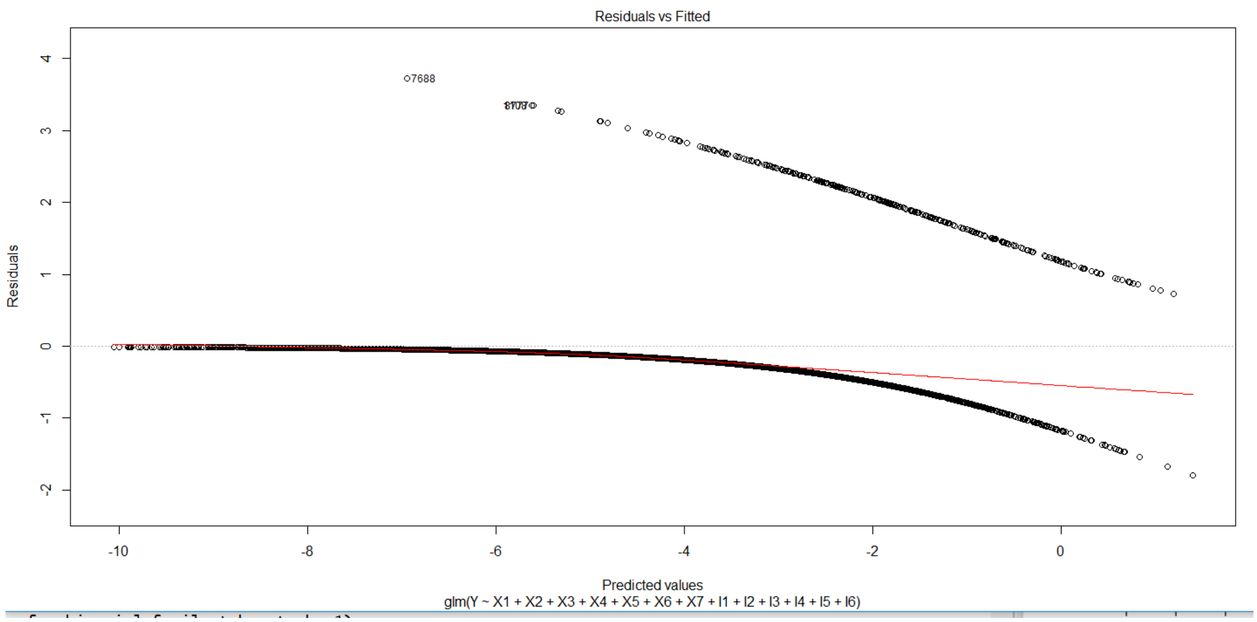
#The estimate (weight) of X7 is "0.18873"->log(odds) change with one unit change in that feature provided everything else is constant.

#The intercept is (-86.14907) ,it means when the features predicting the output has zero value , then we will get this output(log(odds))

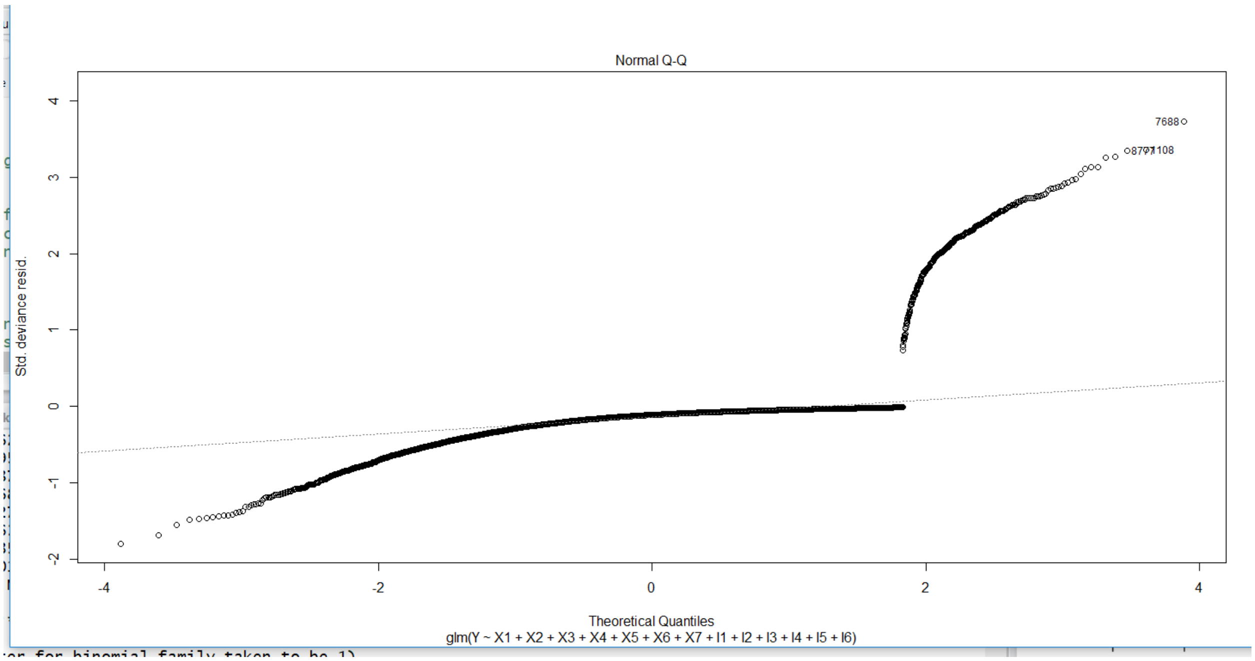
6. **Logistic Regression Plots**

plot(fit_model)

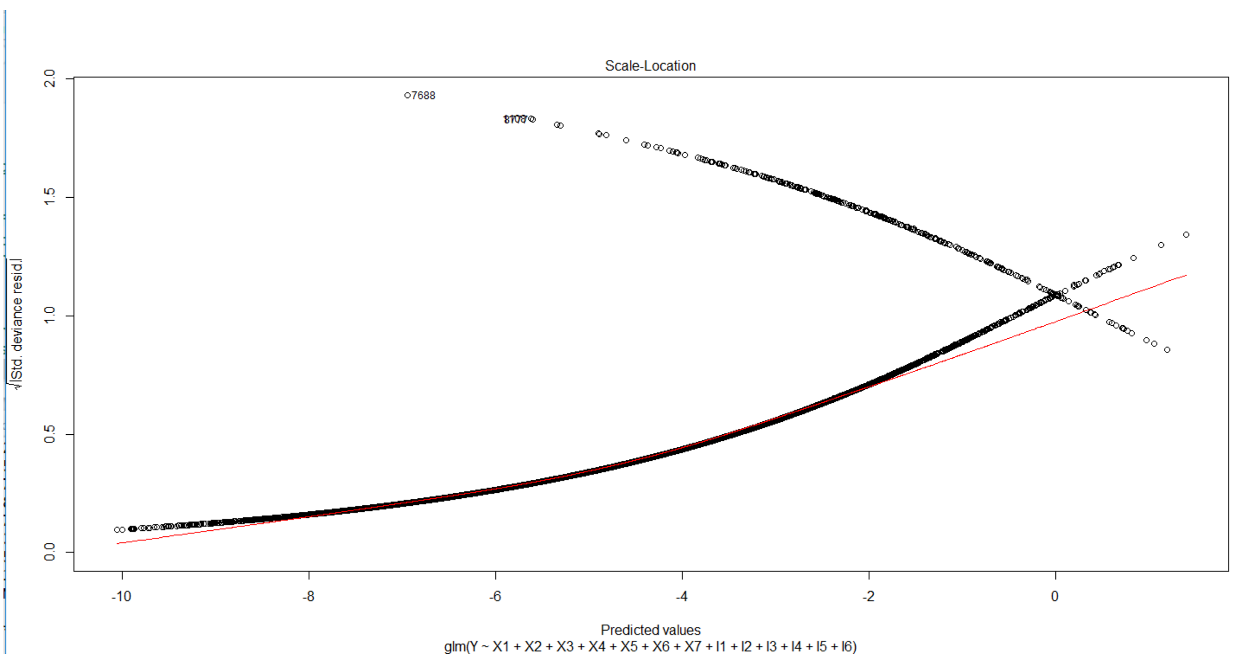
#Residual vs Fitted model



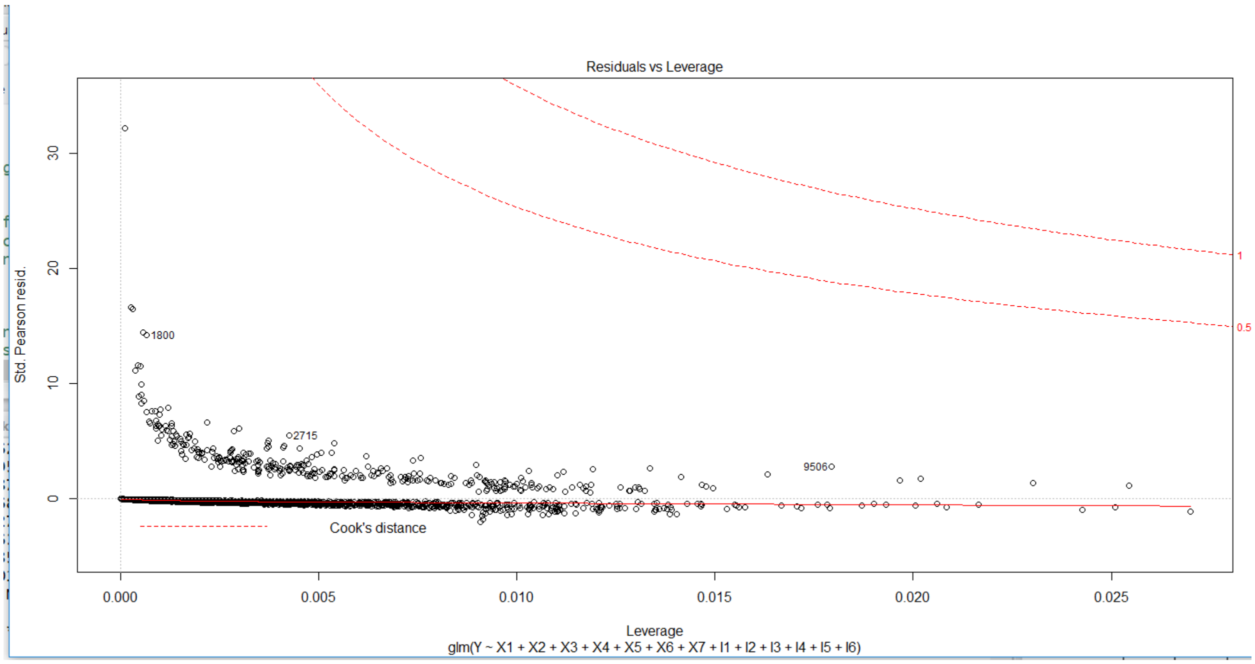
Normal Q-Q Plot



Above Plot (**Square Root of Y-axis: So that outliers are clearly visible)**



Residual vs Leverage points



Residuals vs Leverage
Std. Pearson resid.
30
20
10
0
1800
2715
9506
Cook's distance
1
0.5
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
glm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + I1 + I2 + I3 + I4 + I5 + I6)