Midterm

Name: Zixin Ouyang

3. Load the csv file into R and construct the following attributes.

## load csv. file into R:

```
Arrowsmith<-read.csv('/Users/Constance/Desktop/Arrowsmith.csv',skip = 4,
colClasses=c(rep(NA,15),rep("NULL",14)))
```

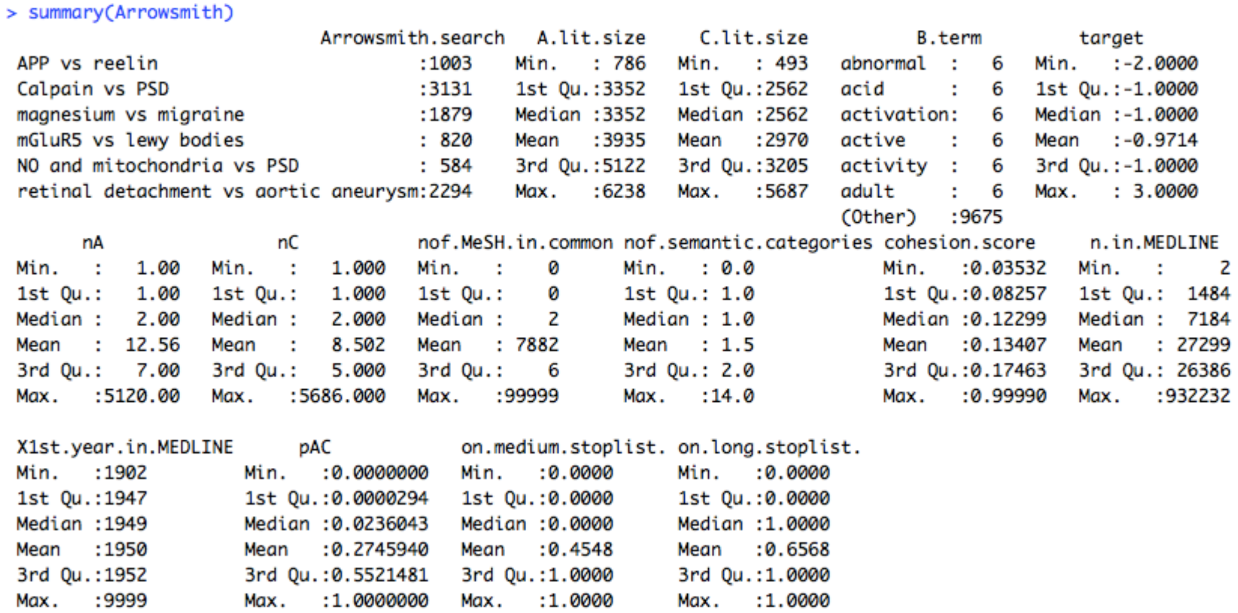
##transformation:

```
X1<-ifelse((Arrowsmith$nA>1|Arrowsmith$A.lit.size<1000) &
(Arrowsmith$nC>1|Arrowsmith$C.lit.size<1000),1,0)
X2<-ifelse(Arrowsmith$nof.MeSH.in.common>0
&Arrowsmith$nof.MeSH.in.common<99999,1,ifelse(Arrowsmith$nof.MeSH.in.common==999
99,0.5,0))
X3<-ifelse(Arrowsmith$nof.semantic.categories>0,1,0)
X4<-ifelse(Arrowsmith$cohesion.score<0.3,Arrowsmith$cohesion.score,0.3)
X5<--abs(log10(Arrowsmith$n.in.MEDLINE)-3)
X6<-pmax(pmin(Arrowsmith$X1st.year.in.MEDLINE,2005),1950)
X7<-pmin(8,-log10(Arrowsmith$pAC+0.000000001))
I1<-ifelse(Arrowsmith$Arrowsmith.search=='retinal detachment vs aortic aneurysm',1,0)
I2<-ifelse(Arrowsmith$Arrowsmith.search=='NO and mitochondria vs PSD',1,0)
I3<-ifelse(Arrowsmith$Arrowsmith.search=='mGluR5 vs lewy bodies',1,0)
I4<-ifelse(Arrowsmith$Arrowsmith.search=='magnesium vs migraine',1,0)
I5<-ifelse(Arrowsmith$Arrowsmith.search=='Calpain vs PSD',1,0)
I6<-ifelse(Arrowsmith$Arrowsmith.search=='APP vs reelin',1,0)
Y<-ifelse(Arrowsmith$target==0|Arrowsmith$target==2,1,0)
```

4.

##summary statistics before transformation:

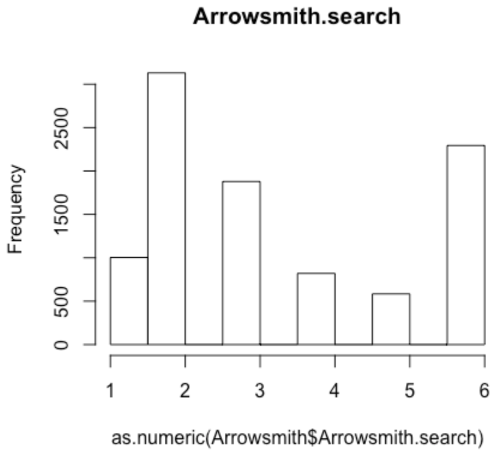
```
summary(Arrowsmith)
```

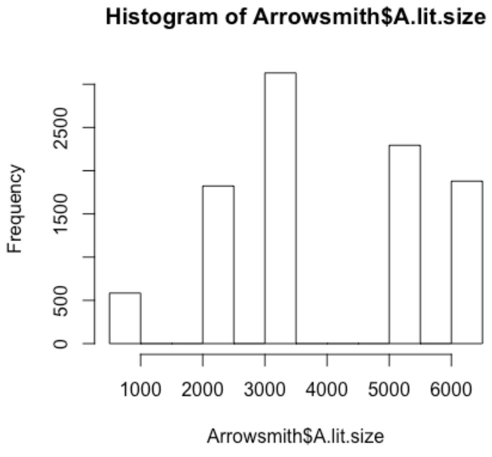


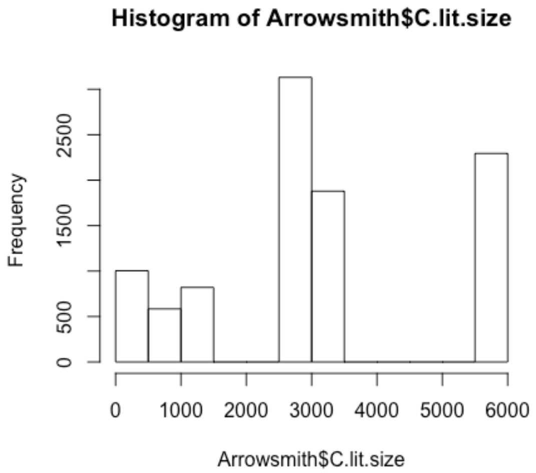
```
> summary(Arrowsmith)
                               Arrowsmith.search   A.lit.size     C.lit.size          B.term         target
 APP vs reelin                          :1003   Min.   : 786   Min.   : 493   abnormal  :   6   Min.   :-2.0000
 Calpain vs PSD                         :3131   1st Qu.:3352   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.: 26386
 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
 1st Qu.:1947         1st Qu.:0.0000294   1st Qu.:0.0000      1st Qu.:0.0000
 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
```

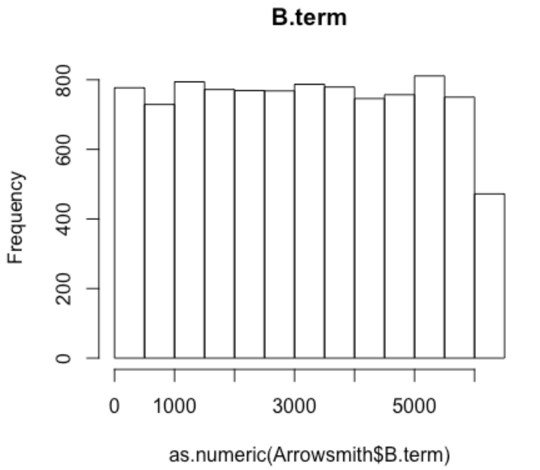
According to notation in the original xls. file, some missing values annotated by 99999 (MeSH), 0.9999 (cohesion), or 9999 (year). Thus, the maximums of nof.MeSH.in.common, cohension.score, and X1st.year.in.MEDLINE are the missing values.

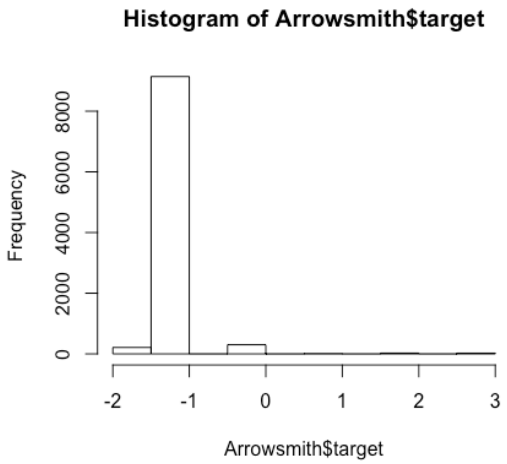
```
##histograms before transformation:
hist(as.numeric(Arrowsmith$Arrowsmith.search),main="Arrowsmith.search")
hist(Arrowsmith$A.lit.size)
hist(Arrowsmith$C.lit.size)
hist(as.numeric(Arrowsmith$B.term),main="B.term")
hist(Arrowsmith$target)
hist(Arrowsmith$nA)
hist(Arrowsmith$nC)
hist(Arrowsmith$nof.MeSH.in.common,main="nof.MeSH.in.common")
hist(Arrowsmith$nof.semantic.categories,main="semantic.categories")
hist(Arrowsmith$cohesion.score,main="cohesion.score")
hist(Arrowsmith$n.in.MEDLINE,main="n.in.MEDLINE")
hist(Arrowsmith$X1st.year.in.MEDLINE,main="X1st.year.in.MEDLINE")
hist(Arrowsmith$pAC)
hist(Arrowsmith$on.medium.stoplist.,main="on.medium.stoplist.")
hist(Arrowsmith$on.long.stoplist.,main="on.long.stoplist.")
```

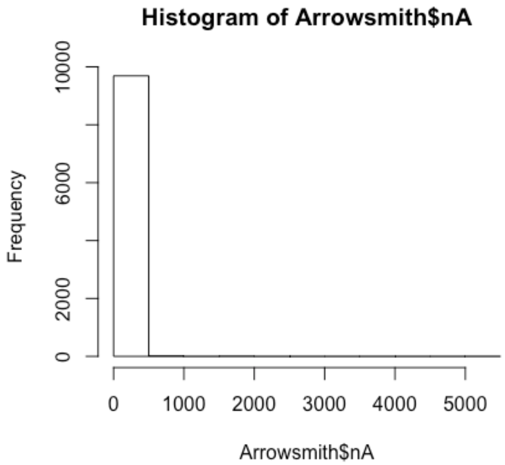


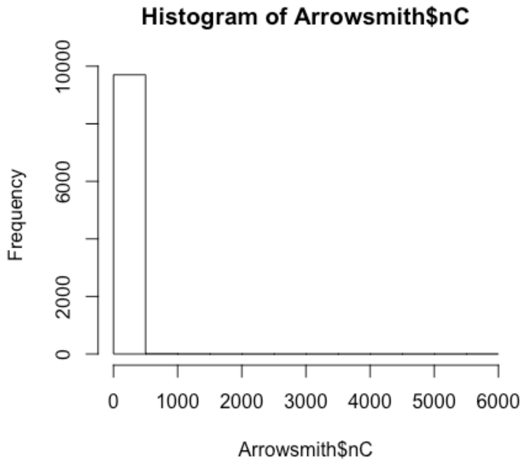


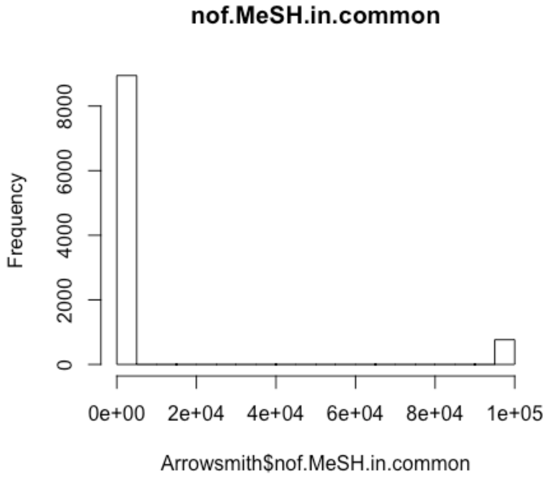

Histogram of Arrowsmith$C.lit.size
Frequency
0
500
1500
2500
0
1000
2000
3000
4000
5000
6000
Arrowsmith$C.lit.size

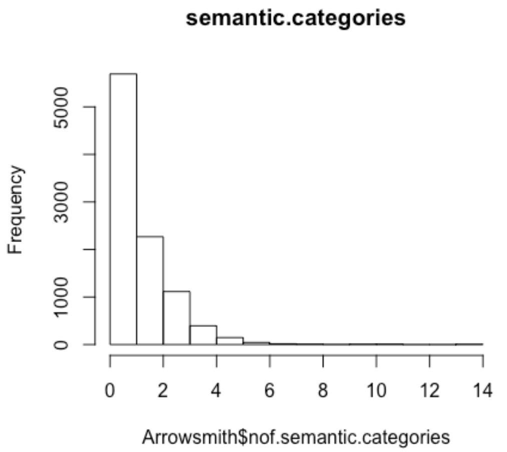

B.term
Frequency
0
200
400
600
800
0
1000
3000
5000
as.numeric(Arrowsmith$B.term)

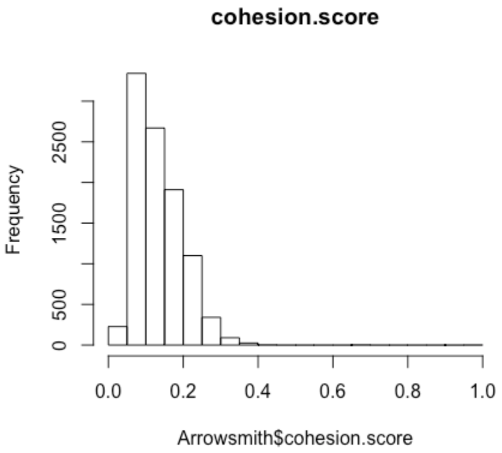

Histogram of Arrowsmith$target
Frequency
0
2000
4000
6000
8000
-2
-1
0
1
2
3
Arrowsmith$target

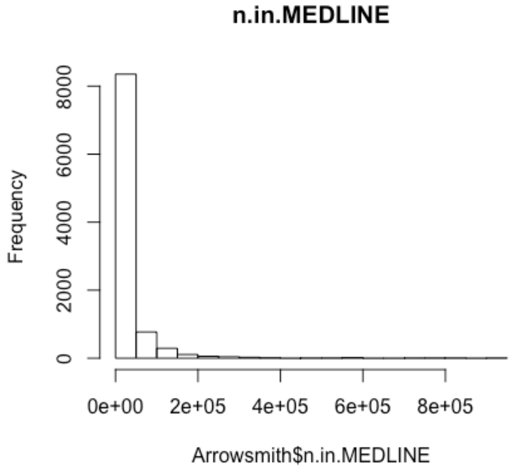

Histogram of Arrowsmith$nA
Frequency
0
2000
6000
10000
0
1000
2000
3000
4000
5000
Arrowsmith$nA

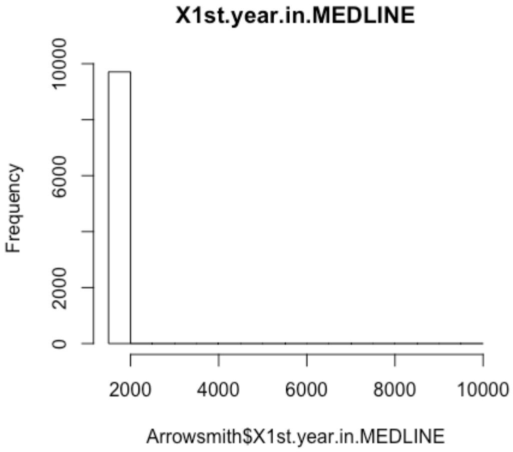

Histogram of Arrowsmith$nC
Frequency
0
2000
6000
10000
0
1000
2000
3000
4000
5000
6000
Arrowsmith$nC

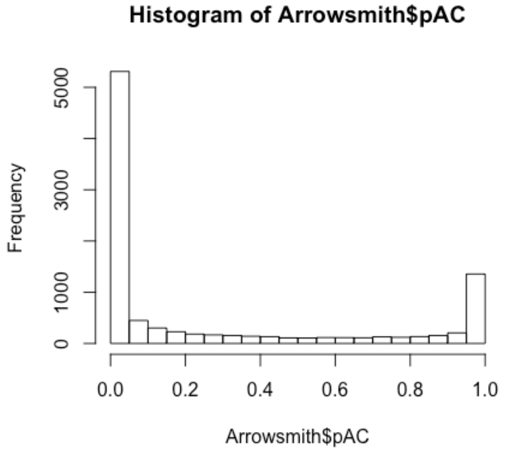

nof.MeSH.in.common
Frequency
0
2000
4000
6000
8000
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Arrowsmith$nof.MeSH.in.common

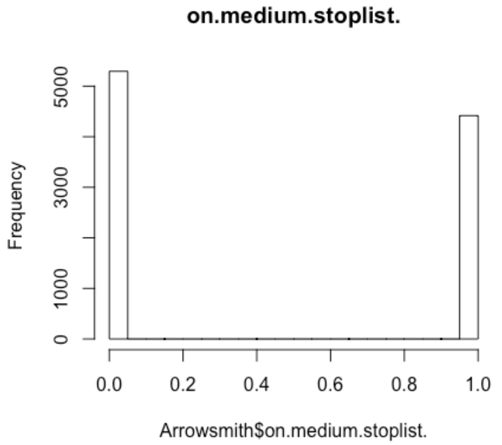

semantic.categories
Frequency
0
1000
3000
5000
0
2
4
6
8
10
12
14
Arrowsmith$nof.semantic.categories

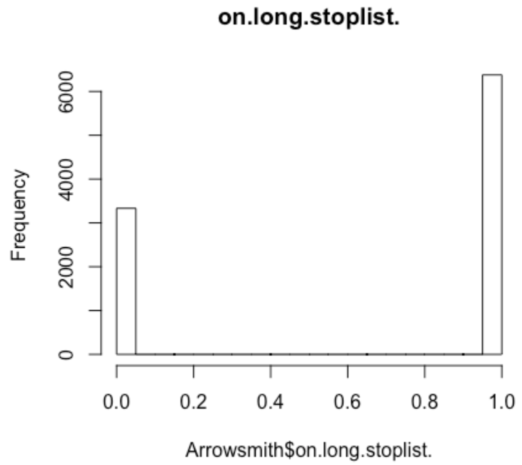

cohesion.score
Frequency
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
Arrowsmith$cohesion.score


n.in.MEDLINE
Frequency
0
2000
4000
6000
8000
0e+00
2e+05
4e+05
6e+05
8e+05
Arrowsmith$n.in.MEDLINE


X1st.year.in.MEDLINE
Frequency
0
2000
6000
10000
2000
4000
6000
8000
10000
Arrowsmith$X1st.year.in.MEDLINE


Histogram of Arrowsmith$pAC
Frequency
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Arrowsmith$pAC

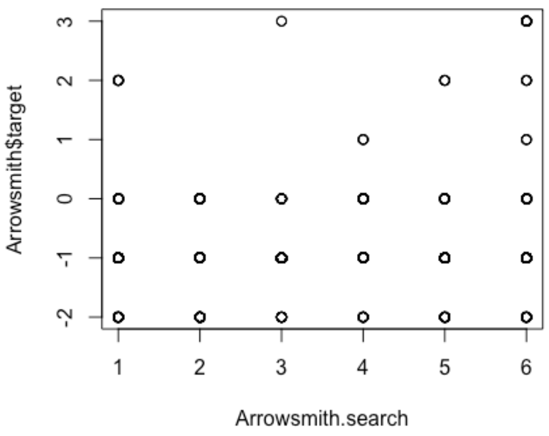

on.medium.stoplist.
Frequency
0
1000
3000
5000
0.0
0.2
0.4
0.6
0.8
1.0
Arrowsmith$on.medium.stoplist.

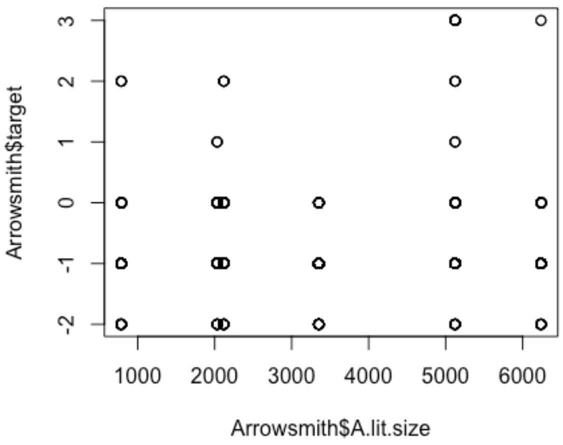


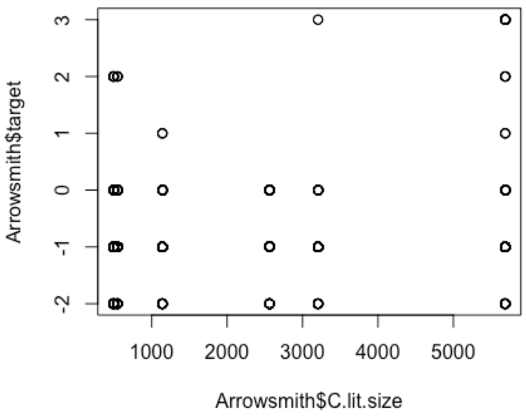
According to the histogram of target: most target is -1, a little is -2 or 0.
The histograms of sematic.categroies. and n in MEDLINE are right skewed, so there are outliers in the data.
There are also outliers in the histogram of cohesion.score.

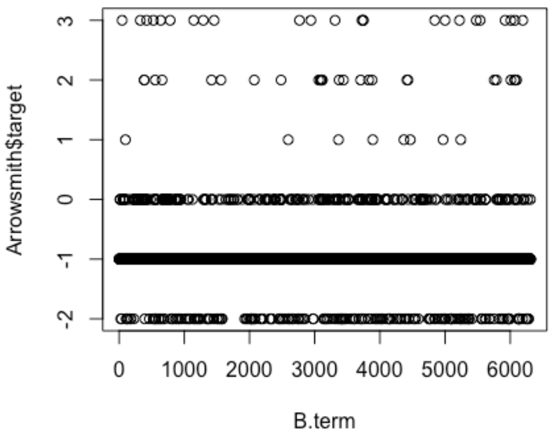
##pairwise scatter plots before transformation:

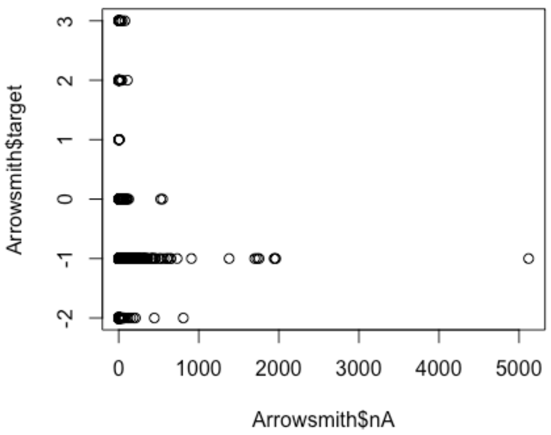
```
plot(as.numeric(Arrowsmith$Arrowsmith.search),Arrowsmith$target,xlab="Arrowsmith.search")
plot(Arrowsmith$A.lit.size,Arrowsmith$target)
plot(Arrowsmith$C.lit.size,Arrowsmith$target)
plot(as.numeric(Arrowsmith$B.term),Arrowsmith$target,xlab="B.term")
plot(Arrowsmith$nA,Arrowsmith$target)
plot(Arrowsmith$nC,Arrowsmith$target)
plot(Arrowsmith$nof.MeSH.in.common,Arrowsmith$target)
plot(Arrowsmith$nof.semantic.categories,Arrowsmith$target)
plot(Arrowsmith$cohesion.score,Arrowsmith$target)
plot(Arrowsmith$n.in.MEDLINE,Arrowsmith$target)
plot(Arrowsmith$X1st.year.in.MEDLINE,Arrowsmith$target)
plot(Arrowsmith$pAC,Arrowsmith$target)
```

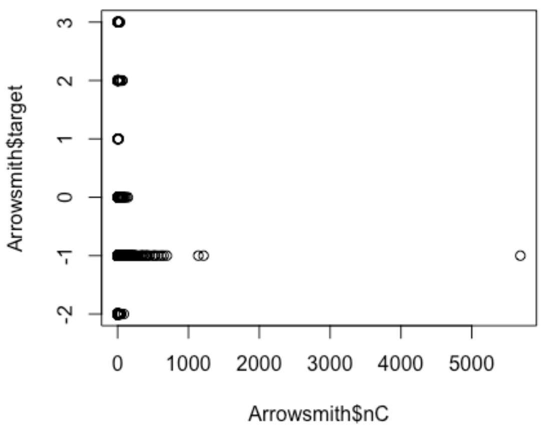

Arrowsmith$target
3
2
1
0
-1
-2
1
2
3
4
5
6
Arrowsmith.search

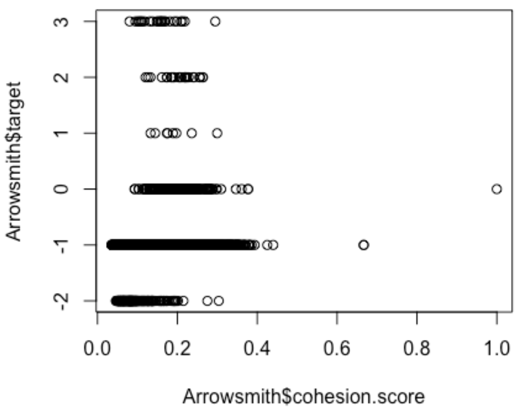

Arrowsmith$target
3
2
1
0
-1
-2
1000
2000
3000
4000
5000
6000
Arrowsmith$A.lit.size

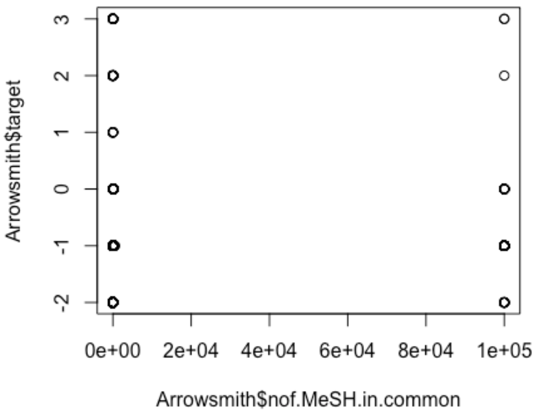

Arrowsmith$target
3
2
1
0
-1
-2
1000
2000
3000
4000
5000
Arrowsmith$C.lit.size

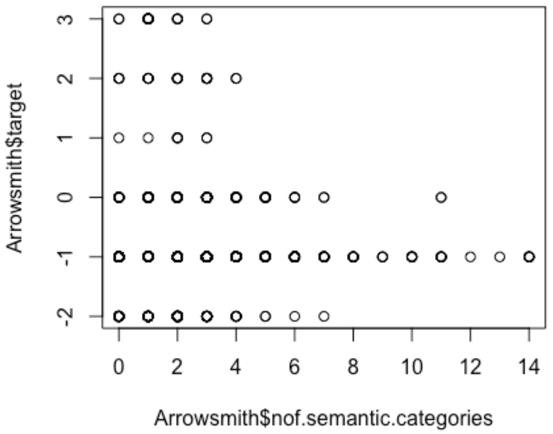

Arrowsmith$target
3
2
1
0
-1
-2
0
1000
2000
3000
4000
5000
6000
B.term

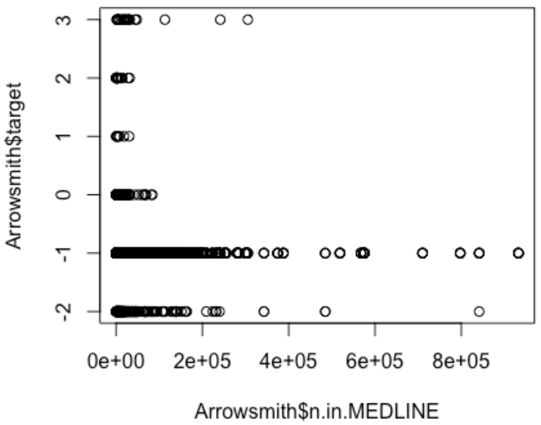

Arrowsmith$target
3
2
1
0
-1
-2
0
1000
2000
3000
4000
5000
Arrowsmith$nA

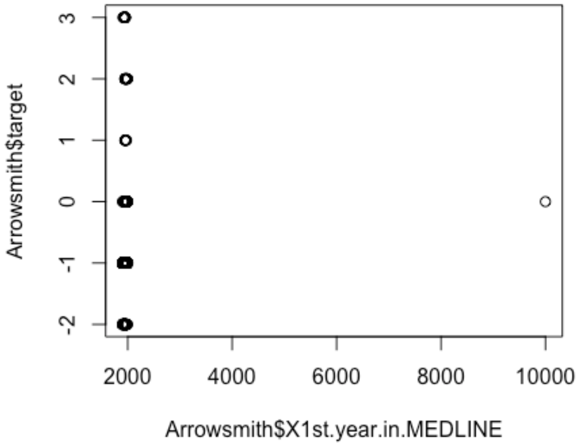

Arrowsmith$target
3
2
1
0
-1
-2
0
1000
2000
3000
4000
5000
Arrowsmith$nC

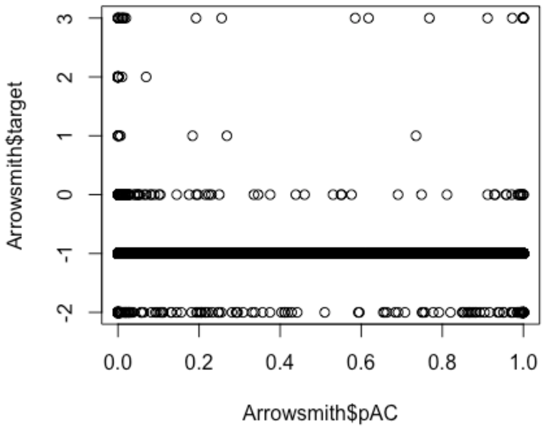

Arrowsmith$target
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Arrowsmith$cohesion.score


Arrowsmith$target
3
2
1
0
-1
-2
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Arrowsmith$nof.MeSH.in.common


Arrowsmith$target
3
2
1
0
-1
-2
0
2
4
6
8
10
12
14
Arrowsmith$nof.semantic.categories


Arrowsmith$target
3
2
1
0
-1
-2
0e+00
2e+05
4e+05
6e+05
8e+05
Arrowsmith$n.in.MEDLINE

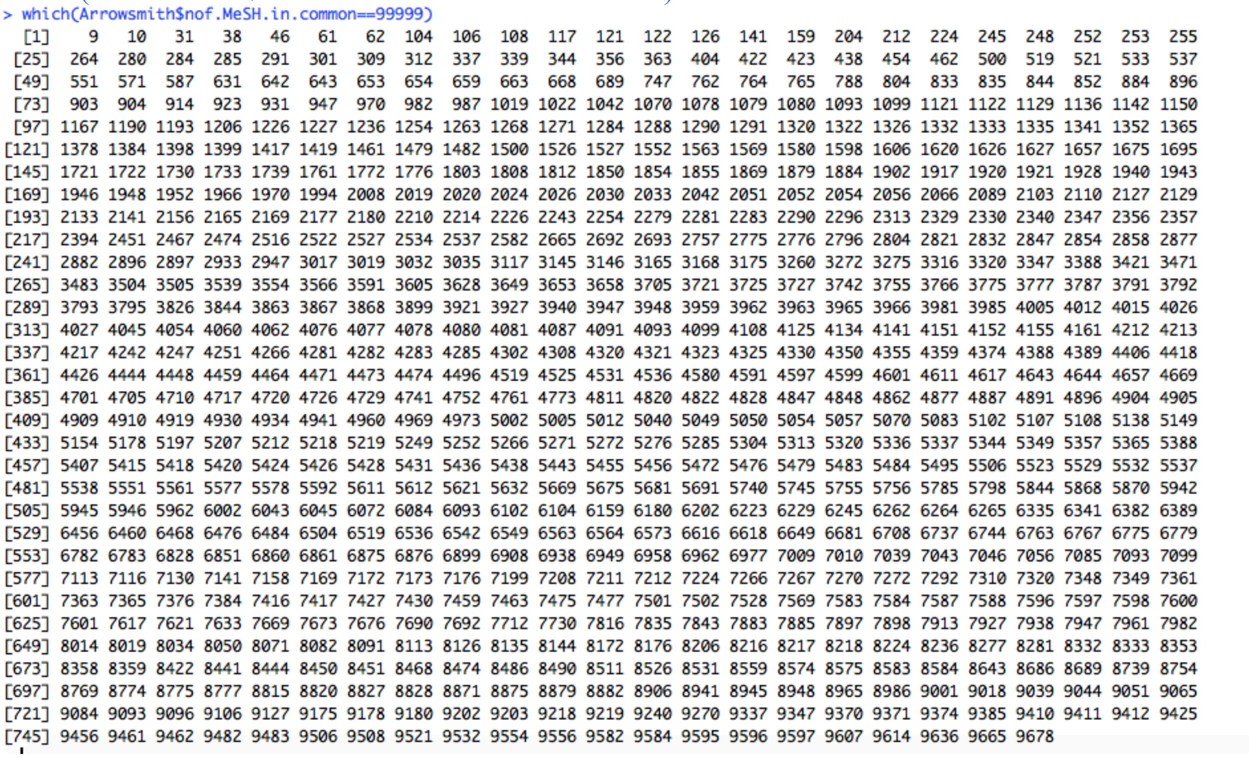

Arrowsmith$target
3
2
1
0
-1
-2
2000
4000
6000
8000
10000
Arrowsmith$X1st.year.in.MEDLINE


Arrowsmith$target
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Arrowsmith$pAC

There are outliers in the plot of nA vs target, nC vs target, cohesion.socre. vs target and n in MEDLINE vs target.

##looking for missing value before transformation:

```
which(Arrowsmith$nof.MeSH.in.common==99999)
```



```
> which(Arrowsmith$nof.MeSH.in.common==99999)
  [1]    9   10   31   38   46   61   62  104  106  108  117  121  122  126  141  159  204  212  224  245  248  252  253  255
 [25]  264  280  284  285  291  301  309  312  337  339  344  356  363  404  422  423  438  454  462  500  519  521  533  537
 [49]  551  571  587  631  642  643  653  654  659  663  668  689  747  762  764  765  788  804  833  835  844  852  884  896
 [73]  903  904  914  923  931  947  970  982  987 1019 1022 1042 1070 1078 1079 1080 1093 1099 1121 1122 1129 1136 1142 1150
 [97] 1167 1190 1193 1206 1226 1227 1236 1254 1263 1268 1271 1284 1288 1290 1291 1320 1322 1326 1332 1333 1335 1341 1352 1365
[121] 1378 1384 1398 1399 1417 1419 1461 1479 1482 1500 1526 1527 1552 1563 1569 1580 1598 1606 1620 1626 1627 1657 1675 1695
[145] 1721 1722 1730 1733 1739 1761 1772 1776 1803 1808 1812 1850 1854 1855 1869 1879 1884 1902 1917 1920 1921 1928 1940 1943
[169] 1946 1948 1952 1966 1970 1994 2008 2019 2020 2024 2026 2030 2033 2042 2051 2052 2054 2056 2066 2089 2103 2110 2127 2129
[193] 2133 2141 2156 2165 2169 2177 2180 2210 2214 2226 2243 2254 2279 2281 2283 2290 2296 2313 2329 2330 2340 2347 2356 2357
[217] 2394 2451 2467 2474 2516 2522 2527 2534 2537 2582 2665 2692 2693 2757 2775 2776 2796 2804 2821 2832 2847 2854 2858 2877
[241] 2882 2896 2897 2933 2947 3017 3019 3032 3035 3117 3145 3146 3165 3168 3175 3260 3272 3275 3316 3320 3347 3388 3421 3471
[265] 3483 3504 3505 3539 3554 3566 3591 3605 3628 3649 3653 3658 3705 3721 3725 3727 3742 3755 3766 3775 3777 3787 3791 3792
[289] 3793 3795 3826 3844 3863 3867 3868 3899 3921 3927 3940 3947 3948 3959 3962 3963 3965 3966 3981 3985 4005 4012 4015 4026
[313] 4027 4045 4054 4060 4062 4076 4077 4078 4080 4081 4087 4091 4093 4099 4108 4125 4134 4141 4151 4152 4155 4161 4212 4213
[337] 4217 4242 4247 4251 4266 4281 4282 4283 4285 4302 4308 4320 4321 4323 4325 4330 4350 4355 4359 4374 4388 4389 4406 4418
[361] 4426 4444 4448 4459 4464 4471 4473 4474 4496 4519 4525 4531 4536 4580 4591 4597 4599 4601 4611 4617 4643 4644 4657 4669
[385] 4701 4705 4710 4717 4720 4726 4729 4741 4752 4761 4773 4811 4820 4822 4828 4847 4848 4862 4877 4887 4891 4896 4904 4905
[409] 4909 4910 4919 4930 4934 4941 4960 4969 4973 5002 5005 5012 5040 5049 5050 5054 5057 5070 5083 5102 5107 5108 5138 5149
[433] 5154 5178 5197 5207 5212 5218 5219 5249 5252 5266 5271 5272 5276 5285 5304 5313 5320 5336 5337 5344 5349 5357 5365 5388
[457] 5407 5415 5418 5420 5424 5426 5428 5431 5436 5438 5443 5455 5456 5472 5476 5479 5483 5484 5495 5506 5523 5529 5532 5537
[481] 5538 5551 5561 5577 5578 5592 5611 5612 5621 5632 5669 5675 5681 5691 5740 5745 5755 5756 5785 5798 5844 5868 5870 5942
[505] 5945 5946 5962 6002 6043 6045 6072 6084 6093 6102 6104 6159 6180 6202 6223 6229 6245 6262 6264 6265 6335 6341 6382 6389
[529] 6456 6460 6468 6476 6484 6504 6519 6536 6542 6549 6563 6564 6573 6616 6618 6649 6681 6708 6737 6744 6763 6767 6775 6779
[553] 6782 6783 6828 6851 6860 6861 6875 6876 6899 6908 6938 6949 6958 6962 6977 7009 7010 7039 7043 7046 7056 7085 7093 7099
[577] 7113 7116 7130 7141 7158 7169 7172 7173 7176 7199 7208 7211 7212 7224 7266 7267 7270 7272 7292 7310 7320 7348 7349 7361
[601] 7363 7365 7376 7384 7416 7417 7427 7430 7459 7463 7475 7477 7501 7502 7528 7569 7583 7584 7587 7588 7596 7597 7598 7600
[625] 7601 7617 7621 7633 7669 7673 7676 7690 7692 7712 7730 7816 7835 7843 7883 7885 7897 7898 7913 7927 7938 7947 7961 7982
[649] 8014 8019 8034 8050 8071 8082 8091 8113 8126 8135 8144 8172 8176 8206 8216 8217 8218 8224 8236 8277 8281 8332 8333 8353
[673] 8358 8359 8422 8441 8444 8450 8451 8468 8474 8486 8490 8511 8526 8531 8559 8574 8575 8583 8584 8643 8686 8689 8739 8754
[697] 8769 8774 8775 8777 8815 8820 8827 8828 8871 8875 8879 8882 8906 8941 8945 8948 8965 8986 9001 9018 9039 9044 9051 9065
[721] 9084 9093 9096 9106 9127 9175 9178 9180 9202 9203 9218 9219 9240 9270 9337 9347 9370 9371 9374 9385 9410 9411 9412 9425
[745] 9456 9461 9462 9482 9483 9506 9508 9521 9532 9554 9556 9582 9584 9595 9596 9597 9607 9614 9636 9665 9678
```

```
which(Arrowsmith$cohesion.score==0.9999)
[1] 9506
which(Arrowsmith$X1st.year.in.MEDLINE==9999)
[1] 9506
```

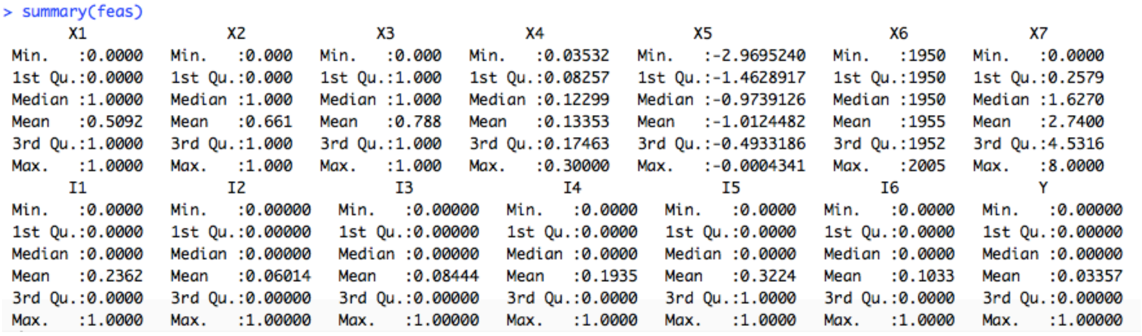
We can get the numbers and locations of missing values using the which( )function. There are 765 missing values in nof.MeSH.in.common, and there is one missing value in cohesion.score and X1st.year.in.MEDLINE.

## construct a data frame using X1,X2,X3,X4,X5,X6,X7,I1,I2,I3,I4,I5,I6,I7,I8 as variables:

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

##summary statistics after transformation:

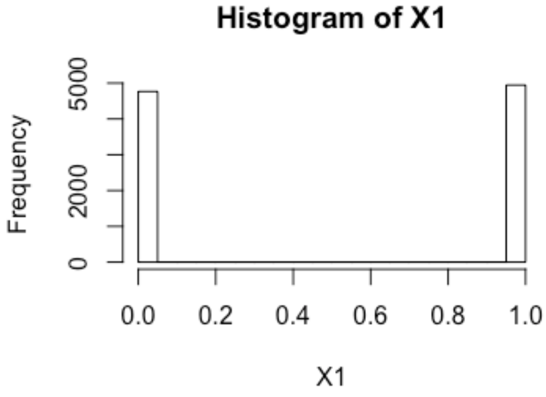
```
summary(feas)
```

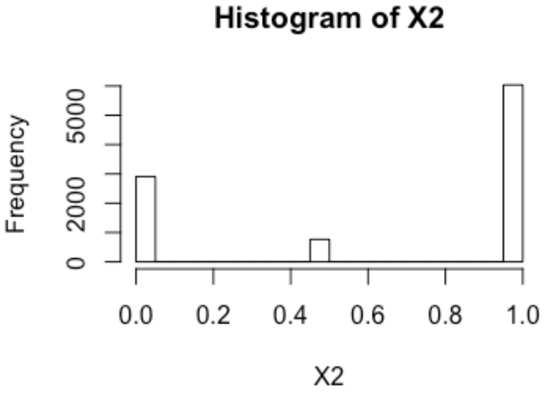


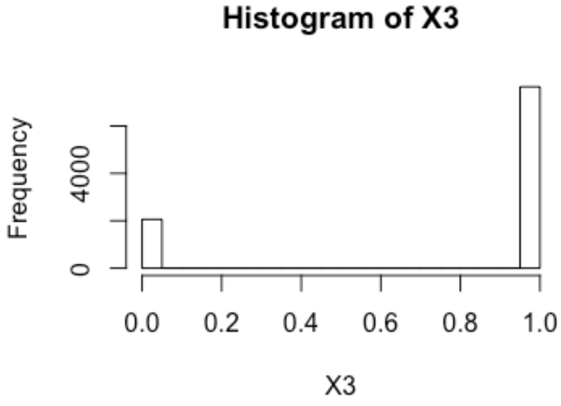
```
> summary(feas)
       X1               X2              X3              X4                X5                  X6             X7        
 Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.03532   Min.   :-2.9695240   Min.   :1950   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.08257   1st Qu.:-1.4628917   1st Qu.:1950   1st Qu.:0.2579  
 Median :1.0000   Median :1.000   Median :1.000   Median :0.12299   Median :-0.9739126   Median :1950   Median :1.6270  
 Mean   :0.5092   Mean   :0.661   Mean   :0.788   Mean   :0.13353   Mean   :-1.0124482   Mean   :1955   Mean   :2.7400  
 3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.17463   3rd Qu.:-0.4933186   3rd Qu.:1952   3rd Qu.:4.5316  
 Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :0.30000   Max.   :-0.0004341   Max.   :2005   Max.   :8.0000  
       I1               I2                I3                I4               I5               I6               Y          
 Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000  
 Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000  
 Mean   :0.2362   Mean   :0.06014   Mean   :0.08444   Mean   :0.1935   Mean   :0.3224   Mean   :0.1033   Mean   :0.03357  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000  
```

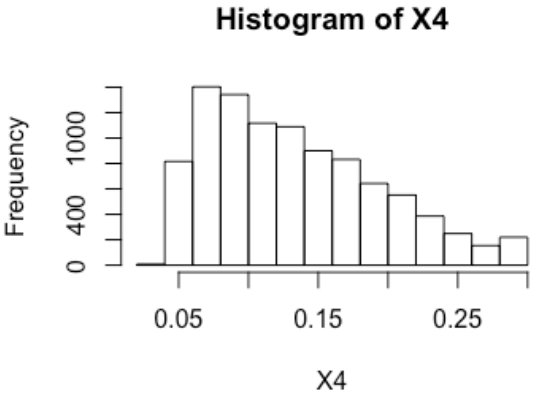
After transformation, the maximums of some variables are no longer the missing values.

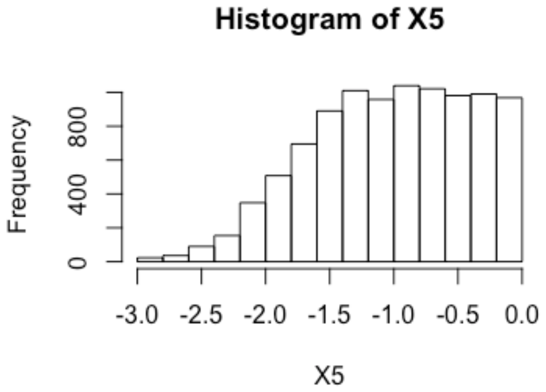
```
##histograms after transformation:
hist(X1)
hist(X2)
hist(X3)
hist(X4)
hist(X5)
hist(X6)
hist(X7)
hist(I1)
hist(I2)
hist(I3)
hist(I4)
hist(I5)
hist(I6)
hist(Y)
```

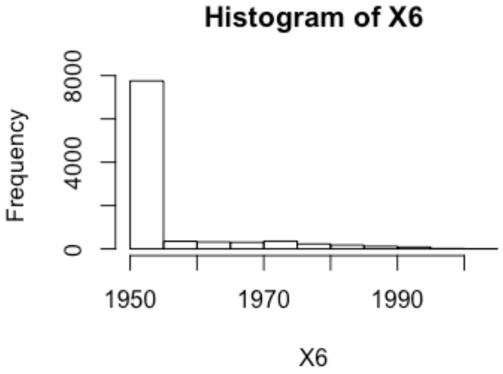


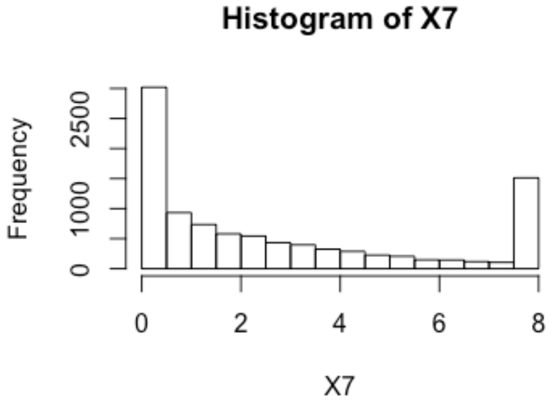


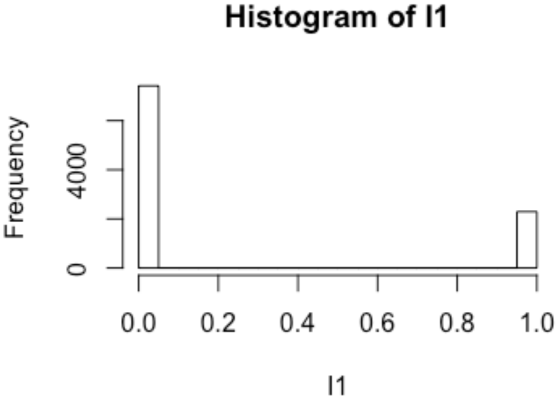


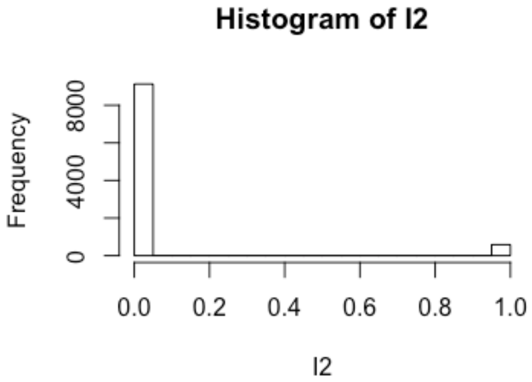


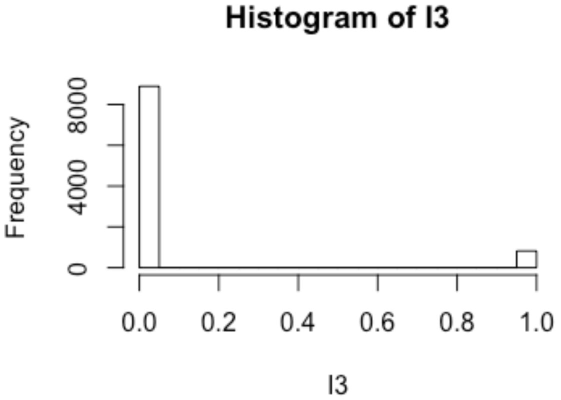

Histogram of X5
Frequency
0
400
800
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
X5

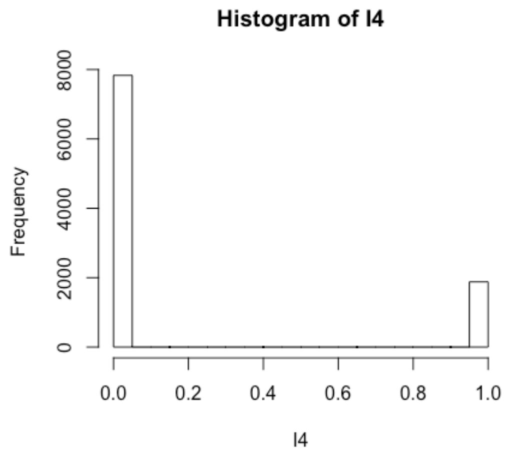

Histogram of X6
Frequency
0
4000
8000
1950
1970
1990
X6

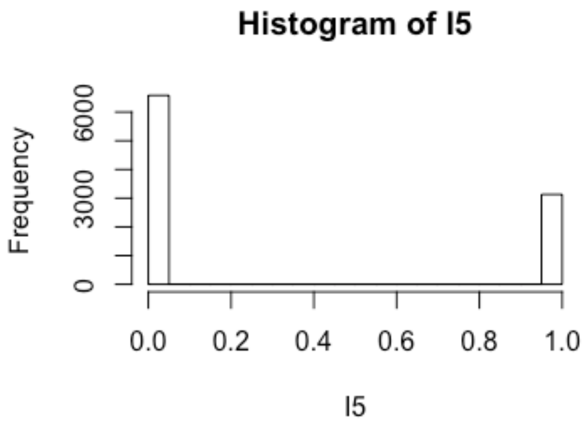

Histogram of X7
Frequency
0
1000
2500
0
2
4
6
8
X7

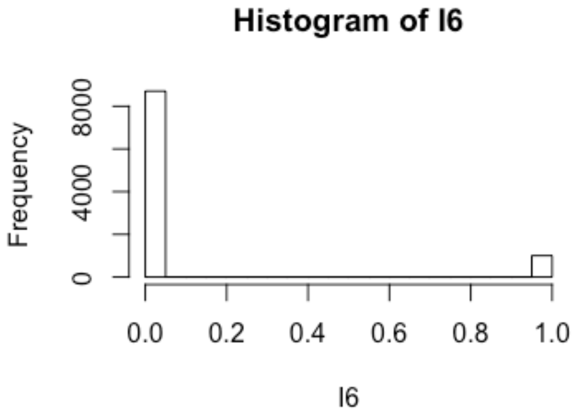

Histogram of I1
Frequency
0
4000
0.0
0.2
0.4
0.6
0.8
1.0
I1

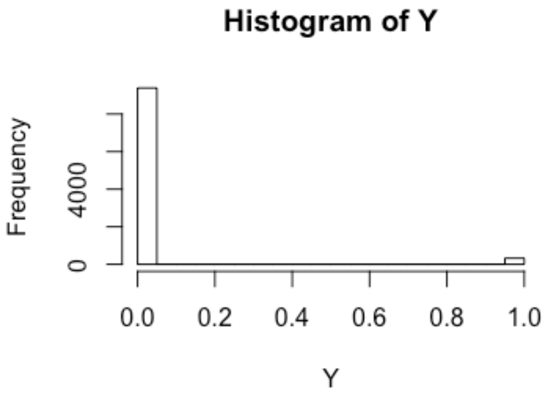

Histogram of I2
Frequency
0
4000
8000
0.0
0.2
0.4
0.6
0.8
1.0
I2


Histogram of I3
Frequency
0
4000
8000
0.0
0.2
0.4
0.6
0.8
1.0
I3



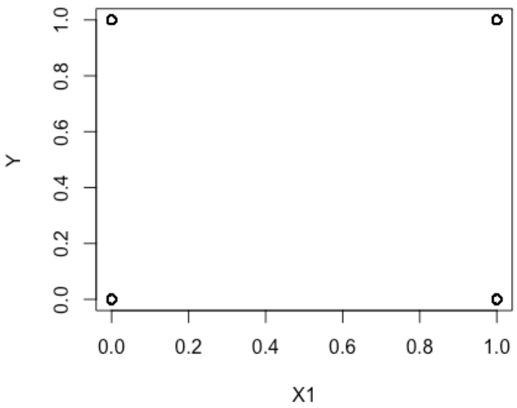


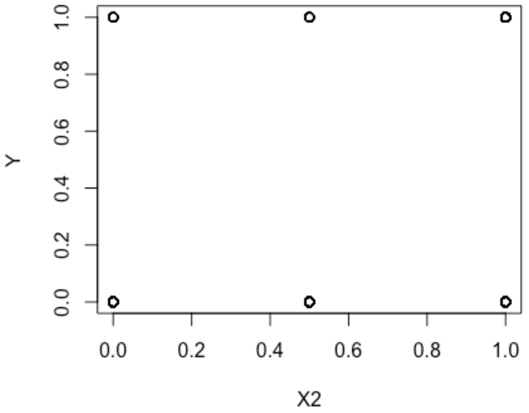


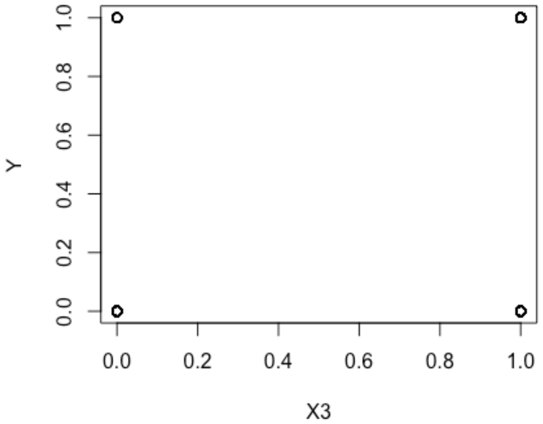


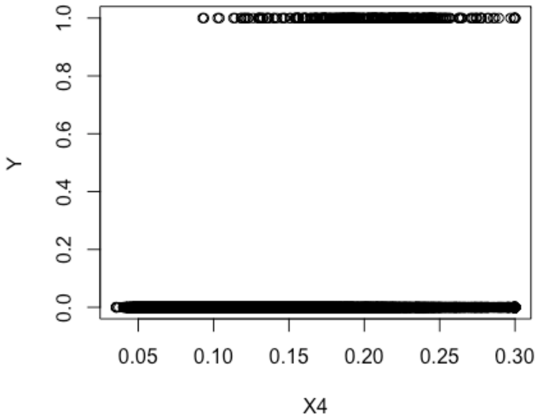
The histogram of X5 is left skewed and the histogram of X6 is right skewed, so there are outliers in the data.

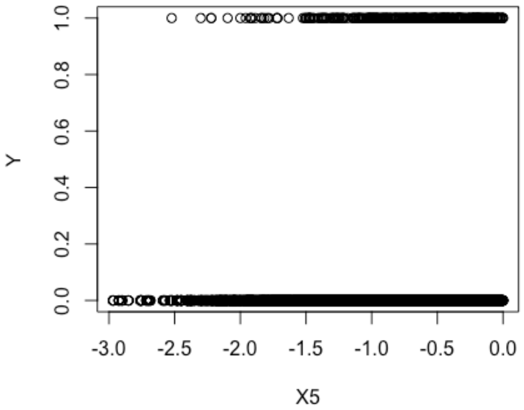
```
##pairwise scatter plots after transformation:
plot(X1,Y)
plot(X2,Y)
plot(X3,Y)
plot(X4,Y)
plot(X5,Y)
plot(X6,Y)
plot(X7,Y)
```

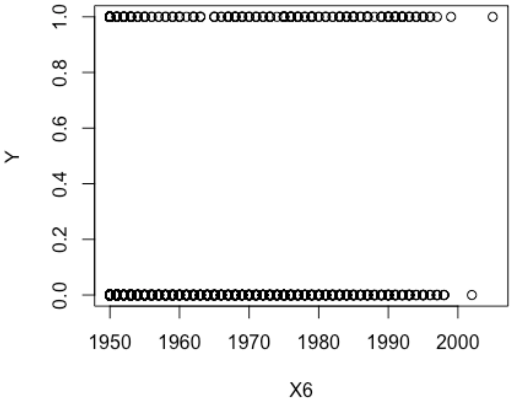

Y
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
X1

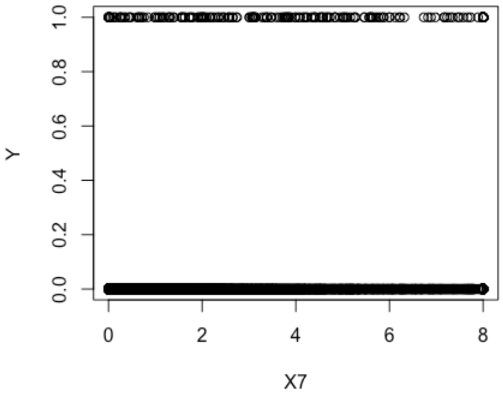

Y
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
X2


Y
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
X3


Y
0.0
0.2
0.4
0.6
0.8
1.0
0.05
0.10
0.15
0.20
0.25
0.30
X4

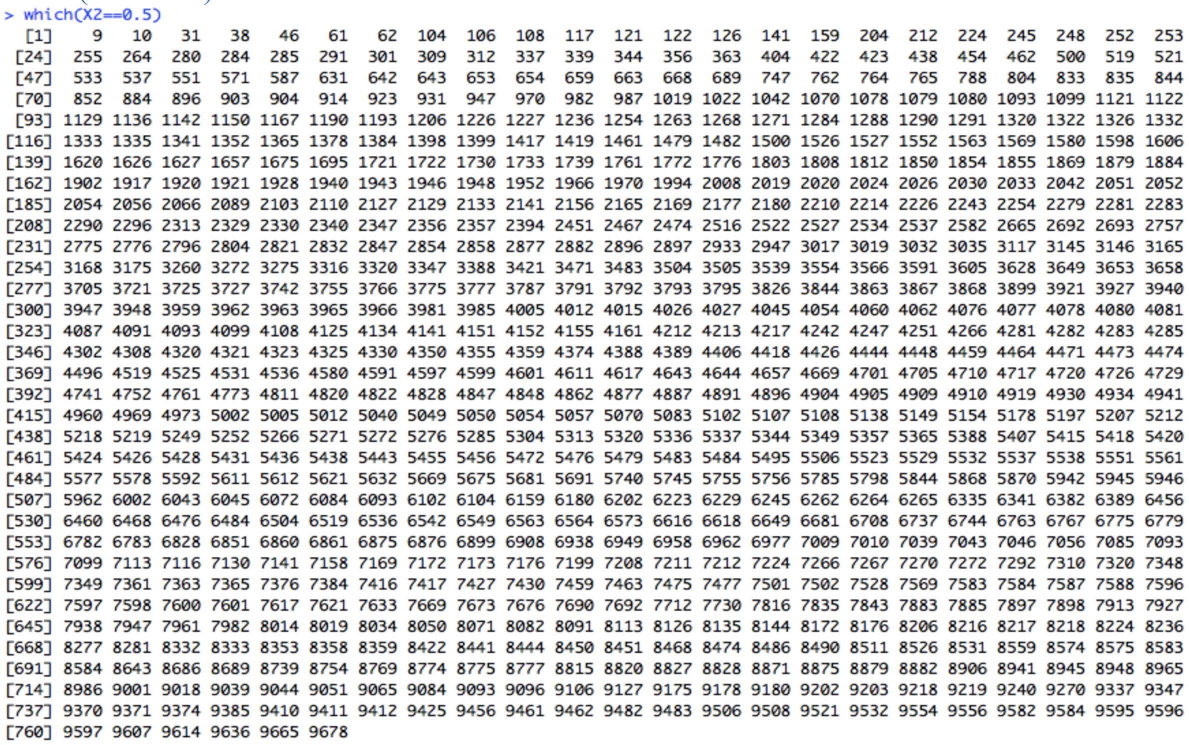

Y
0.0
0.2
0.4
0.6
0.8
1.0
-3.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
X5


Y
0.0
0.2
0.4
0.6
0.8
1.0
1950
1960
1970
1980
1990
2000
X6



There are outliers in the plot of X5 vs Y.

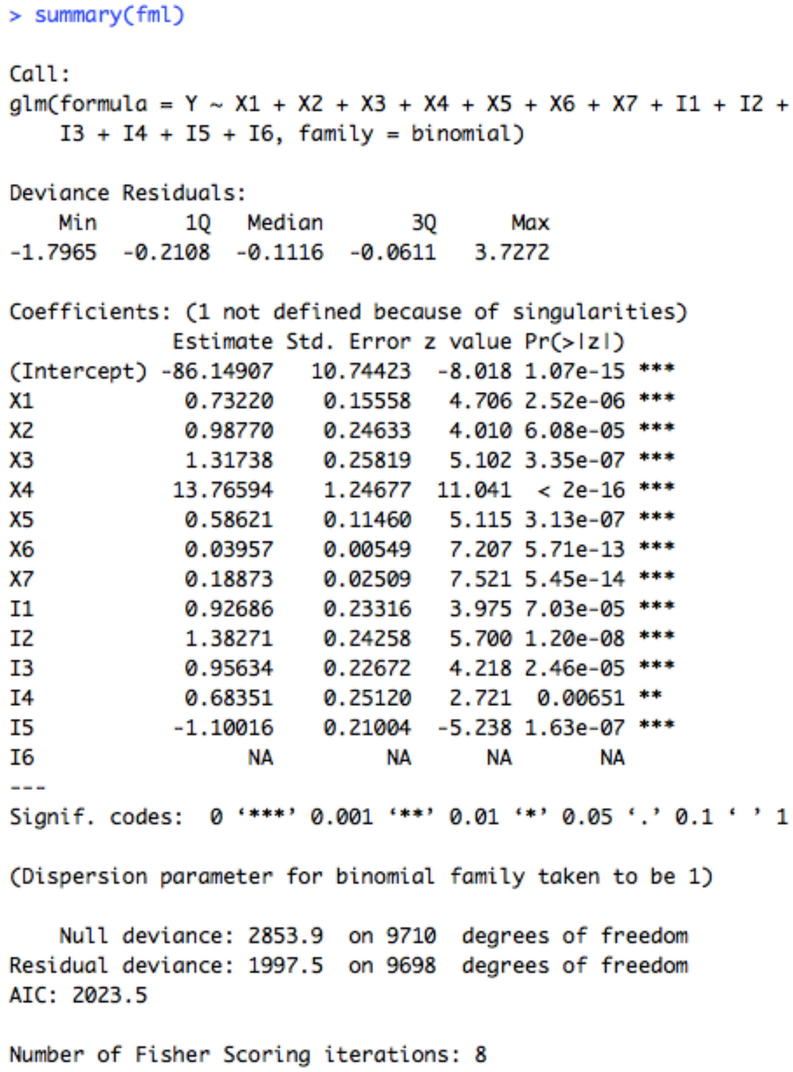
##looking for missing value after transformation:
which(X2==0.5)



```
> which(X2==0.5)
  [1]    9   10   31   38   46   61   62  104  106  108  117  121  122  126  141  159  204  212  224  245  248  252  253
 [24]  255  264  280  284  285  291  301  309  312  337  339  344  356  363  404  422  423  438  454  462  500  519  521
 [47]  533  537  551  571  587  631  642  643  653  654  659  663  668  689  747  762  764  765  788  804  833  835  844
 [70]  852  884  896  903  904  914  923  931  947  970  982  987 1019 1022 1042 1070 1078 1079 1080 1093 1099 1121 1122
 [93] 1129 1136 1142 1150 1167 1190 1193 1206 1226 1227 1236 1254 1263 1268 1271 1284 1288 1290 1291 1320 1322 1326 1332
[116] 1333 1335 1341 1352 1365 1378 1384 1398 1399 1417 1419 1461 1479 1482 1500 1526 1527 1552 1563 1569 1580 1598 1606
[139] 1620 1626 1627 1657 1675 1695 1721 1722 1730 1733 1739 1761 1772 1776 1803 1808 1812 1850 1854 1855 1869 1879 1884
[162] 1902 1917 1920 1921 1928 1940 1943 1946 1948 1952 1966 1970 1994 2008 2019 2020 2024 2026 2030 2033 2042 2051 2052
[185] 2054 2056 2066 2089 2103 2110 2127 2129 2133 2141 2156 2165 2169 2177 2180 2210 2214 2226 2243 2254 2279 2281 2283
[208] 2290 2296 2313 2329 2330 2340 2347 2356 2357 2394 2451 2467 2474 2516 2522 2527 2534 2537 2582 2665 2692 2693 2757
[231] 2775 2776 2796 2804 2821 2832 2847 2854 2858 2877 2882 2896 2897 2933 2947 3017 3019 3032 3035 3117 3145 3146 3165
[254] 3168 3175 3260 3272 3275 3316 3320 3347 3388 3421 3471 3483 3504 3505 3539 3554 3566 3591 3605 3628 3649 3653 3658
[277] 3705 3721 3725 3727 3742 3755 3766 3775 3777 3787 3791 3792 3793 3795 3826 3844 3863 3867 3868 3899 3921 3927 3940
[300] 3947 3948 3959 3962 3963 3965 3966 3981 3985 4005 4012 4015 4026 4027 4045 4054 4060 4062 4076 4077 4078 4080 4081
[323] 4087 4091 4093 4099 4108 4125 4134 4141 4151 4152 4155 4161 4212 4213 4217 4242 4247 4251 4266 4281 4282 4283 4285
[346] 4302 4308 4320 4321 4323 4325 4330 4350 4355 4359 4374 4388 4389 4406 4418 4426 4444 4448 4459 4464 4471 4473 4474
[369] 4496 4519 4525 4531 4536 4580 4591 4597 4599 4601 4611 4617 4643 4644 4657 4669 4701 4705 4710 4717 4720 4726 4729
[392] 4741 4752 4761 4773 4811 4820 4822 4828 4847 4848 4862 4877 4887 4891 4896 4904 4905 4909 4910 4919 4930 4934 4941
[415] 4960 4969 4973 5002 5005 5012 5040 5049 5050 5054 5057 5070 5083 5102 5107 5108 5138 5149 5154 5178 5197 5207 5212
[438] 5218 5219 5249 5252 5266 5271 5272 5276 5285 5304 5313 5320 5336 5337 5344 5349 5357 5365 5388 5407 5415 5418 5420
[461] 5424 5426 5428 5431 5436 5438 5443 5455 5456 5472 5476 5479 5483 5484 5495 5506 5523 5529 5532 5537 5538 5551 5561
[484] 5577 5578 5592 5611 5612 5621 5632 5669 5675 5681 5691 5740 5745 5755 5756 5785 5798 5844 5868 5870 5942 5945 5946
[507] 5962 6002 6043 6045 6072 6084 6093 6102 6104 6159 6180 6202 6223 6229 6245 6262 6264 6265 6335 6341 6382 6389 6456
[530] 6460 6468 6476 6484 6504 6519 6536 6542 6549 6563 6564 6573 6616 6618 6649 6681 6708 6737 6744 6763 6767 6775 6779
[553] 6782 6783 6828 6851 6860 6861 6875 6876 6899 6908 6938 6949 6958 6962 6977 7009 7010 7039 7043 7046 7056 7085 7093
[576] 7099 7113 7116 7130 7141 7158 7169 7172 7173 7176 7199 7208 7211 7212 7224 7266 7267 7270 7272 7292 7310 7320 7348
[599] 7349 7361 7363 7365 7376 7384 7416 7417 7427 7430 7459 7463 7475 7477 7501 7502 7528 7569 7583 7584 7587 7588 7596
[622] 7597 7598 7600 7601 7617 7621 7633 7669 7673 7676 7690 7692 7712 7730 7816 7835 7843 7883 7885 7897 7898 7913 7927
[645] 7938 7947 7961 7982 8014 8019 8034 8050 8071 8082 8091 8113 8126 8135 8144 8172 8176 8206 8216 8217 8218 8224 8236
[668] 8277 8281 8332 8333 8353 8358 8359 8422 8441 8444 8450 8451 8468 8474 8486 8490 8511 8526 8531 8559 8574 8575 8583
[691] 8584 8643 8686 8689 8739 8754 8769 8774 8775 8777 8815 8820 8827 8828 8871 8875 8879 8882 8906 8941 8945 8948 8965
[714] 8986 9001 9018 9039 9044 9051 9065 9084 9093 9096 9106 9127 9175 9178 9180 9202 9203 9218 9219 9240 9270 9337 9347
[737] 9370 9371 9374 9385 9410 9411 9412 9425 9456 9461 9462 9482 9483 9506 9508 9521 9532 9554 9556 9582 9584 9595 9596
[760] 9597 9607 9614 9636 9665 9678
```

Due to transformation, we can only find missing values in X2 using the which( ) function.

```
##logistic regression model:
fml<-glm(Y~X1+X2+X3+X4+X5+X6+X7+I1+I2+I3+I4+I5+I6, family=binomial)
summary(fml)
```



```
> summary(fml)

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

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

Number of Fisher Scoring iterations: 8
```

B-term score: $y=0.73x_1+0.99x_2+1.32x_3+13.8x_4+0.59x_5+0.040x_6+0.19x_7$

Through calling the summary of the logistic regression model, we can get the statistical significance of each parameter.

The validity of assumptions in this binomial logistic regression model:
a. The dependent variable should consist of two categorical, independent (unrelated) groups. The two categories of the dependent variable need to be mutually exclusive and exhaustive. That assumption is valid because Y =1 or 0 and Y cannot be equal to 1 at the same time it is equal to 0.
b. There are two or more independent variables, which should be measured at the continuous or nominal level.

That assumption is valid. There are seven independent variables in the data, and they are all measured at the nominal level because of the normalization.
c. The observations are independent.
That assumption is valid. There are 9711 observations in the model, and they are independent.
d. The data must not show multicollinearity, which occurs when you have two or more independent variables that are highly correlated with each other.
That assumption is not very valid. Some variables are not independent. For example, whether the B-term first appeared recently within MEDLINE as a whole would have a relationship with the frequency of B-term within MEDLINE as a whole.
e. There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.
That assumption is valid. We had first checked that the individual features satisfied a linear model and did not show strong interactive features, and we had executed the transformation before construct the logistic regression model.
f. There should be no significant outliers, high leverage points or highly influential points.
The assumption is valid. Even though there are few outliners, but they are not significant or highly influential.

x_1, x_2, x_3, x_4, x_5, x_6, x_7 *are the feature values of B-term score, and their estimate are the corresponding weights. The weight captures the effect of each feature regardless of the nature of the B-terms. Each unit change of* x_1 *will increase y by 0.73, each unit change of* x_2 *will increase y by 0.99, each unit change of x3 will increase y by 1.32, each unit change of* x_4 *will increase y by 13.8, each unit change of* x_5 *will increase y by 0.59, each unit change of* x_6 *will increase y by 0.0040, and each unit change of* x_7 *will increase y by 0.19.*
I have followed the method on the paper, so my parameter estimates are the same as the ones reported.

5.
The features of B-term score had been normalized and scored, which simplified the calculation in modeling fitting and thus made the process easy.
There are 765 missing values in the nof.MeSH.in.common. 0.5 had assigned to X2 if none of the articles in AB (or BC) have assigned MeSH terms. It was a convenient method to handle missing value.
The 1400 word stoplist feature was excluded from the final multi-dimensional model. That made the process easy, since the stoplist feature was redundant and made some marked relevant terms scored poorly.