Capstone Project - Week1-8

Amol Gote

12/18/2020

set.seed(123)  
library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ----------------------------------------------------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(corrplot)

## corrplot 0.84 loaded

library(tidyverse)  
library(ROCR)  
library(gbm)

## Loaded gbm 2.1.8

library(ROSE)

## Loaded ROSE 0.0-3

library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(parallel)   
library(tree)

## Registered S3 method overwritten by 'tree':  
## method from  
## print.tree cli

library(rpart)  
library(rpart.plot)

fundamentals\_ds <- read.csv("./data/Fundamentals\_DS.csv", na.strings=c(""," "))  
nrow(fundamentals\_ds)

## [1] 17416

### Filter the dataset with the sector assigned.

fundamentals\_ds\_sjm <- fundamentals\_ds %>%  
 filter(tic == 'SJM')  
gsector\_sjm <- head(fundamentals\_ds\_sjm$gsector, 1)  
fundamentals\_ds <- fundamentals\_ds %>%  
 filter(gsector == gsector\_sjm)  
nrow(fundamentals\_ds)

## [1] 2323

gsector\_sjm

## [1] 30

### Filter the data to restrict the dataset to contain all annually reported statements, and exclude restatements.

names(fundamentals\_ds)[names(fundamentals\_ds) == "ï..gvkey"] <- "gvkey"  
fundamentals\_ds\_filter <- fundamentals\_ds %>%  
 filter(datafmt == 'STD')   
nrow(fundamentals\_ds\_filter)

## [1] 1243

target\_company <- fundamentals\_ds\_filter %>%  
 filter(tic == 'SJM')  
target\_company\_gv\_key <- head(target\_company$gvkey, 1)  
target\_company\_gv\_key

## [1] 9777

ncol(fundamentals\_ds\_filter)

## [1] 1768

#summary(fundamentals\_ds\_filter)

summary(fundamentals\_ds\_filter$mkvalt)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.03 29.65 261.19 7270.94 2406.22 241440.44 284

### Remove all the columns which contain 25% or greater NA or null values

fundamentals\_ds\_filter\_1 <- fundamentals\_ds\_filter[ lapply( fundamentals\_ds\_filter, function(x) sum(is.na(x)) / length(x) ) < 0.25 ]  
ncol(fundamentals\_ds\_filter\_1)

## [1] 318

nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

colnames(fundamentals\_ds\_filter\_1)

## [1] "gvkey" "datadate" "fyear" "indfmt" "consol"   
## [6] "popsrc" "datafmt" "tic" "conm" "acctstd"   
## [11] "ajex" "ajp" "curcd" "curncd" "currtr"   
## [16] "final" "fyr" "ismod" "pddur" "scf"   
## [21] "src" "upd" "apdedate" "fdate" "acchg"   
## [26] "acdo" "aco" "acodo" "acominc" "acox"   
## [31] "act" "aldo" "am" "ao" "aocidergl"   
## [36] "aociother" "aocipen" "aocisecgl" "aodo" "aoloch"   
## [41] "aox" "ap" "aqc" "at" "bkvlps"   
## [46] "caps" "capx" "capxv" "ceq" "ceql"   
## [51] "ceqt" "ch" "che" "chech" "ci"   
## [56] "cibegni" "cicurr" "cidergl" "cimii" "ciother"   
## [61] "cipen" "cisecgl" "citotal" "cogs" "cshfd"   
## [66] "cshi" "csho" "cshpri" "cshr" "cstk"   
## [71] "cstkcv" "cstke" "dc" "dclo" "dcom"   
## [76] "dcpstk" "dcvsr" "dcvsub" "dcvt" "dd"   
## [81] "dd1" "dd2" "dd3" "dd4" "dd5"   
## [86] "diladj" "dilavx" "dlc" "dltis" "dlto"   
## [91] "dltp" "dltr" "dltt" "dm" "dn"   
## [96] "do" "donr" "dp" "dpact" "dpc"   
## [101] "dpvieb" "drc" "drlt" "ds" "dudd"   
## [106] "dv" "dvc" "dvp" "dvpa" "dvt"   
## [111] "ebit" "ebitda" "emp" "epsfi" "epsfx"   
## [116] "epspi" "epspx" "esopct" "esopnr" "esopr"   
## [121] "esopt" "esub" "esubc" "exre" "fatb"   
## [126] "fatc" "fatn" "fatp" "fiao" "fincf"   
## [131] "fopo" "fopox" "gdwl" "gp" "ib"   
## [136] "ibadj" "ibc" "ibcom" "ibmii" "icapt"   
## [141] "intan" "intano" "intc" "intpn" "invch"   
## [146] "invt" "itcb" "ivaco" "ivaeq" "ivao"   
## [151] "ivch" "ivncf" "ivst" "ivstch" "lco"   
## [156] "lcox" "lcoxdr" "lct" "lifr" "lo"   
## [161] "loxdr" "lse" "lt" "mib" "mibn"   
## [166] "mibt" "mii" "mrc1" "mrcta" "msa"   
## [171] "ni" "niadj" "nopi" "nopio" "np"   
## [176] "oancf" "oiadp" "oibdp" "opeps" "oprepsx"   
## [181] "pi" "pnca" "pncad" "pncaeps" "pnrsho"   
## [186] "ppegt" "ppent" "ppeveb" "prca" "prcad"   
## [191] "prcaeps" "prsho" "prstkc" "pstk" "pstkc"   
## [196] "pstkl" "pstkn" "pstkr" "pstkrv" "rdip"   
## [201] "rdipa" "rdipd" "rdipeps" "re" "rea"   
## [206] "reajo" "recch" "recco" "recd" "rect"   
## [211] "recta" "rectr" "reuna" "revt" "sale"   
## [216] "seq" "seqo" "siv" "spce" "spced"   
## [221] "spceeps" "spi" "sppe" "sppiv" "sstk"   
## [226] "stkco" "teq" "tstk" "tstkc" "tstkn"   
## [231] "tstkp" "txbco" "txbcof" "txc" "txdb"   
## [236] "txdba" "txdbca" "txdbcl" "txdc" "txdi"   
## [241] "txditc" "txfed" "txfo" "txndb" "txndba"   
## [246] "txndbl" "txndbr" "txo" "txp" "txpd"   
## [251] "txr" "txs" "txt" "txw" "wcap"   
## [256] "xacc" "xi" "xido" "xidoc" "xint"   
## [261] "xintopt" "xopr" "xoptd" "xopteps" "xrent"   
## [266] "xsga" "exchg" "costat" "fic" "naicsh"   
## [271] "sich" "cshtr\_c" "dvpsp\_c" "dvpsx\_c" "prcc\_c"   
## [276] "prch\_c" "prcl\_c" "adjex\_c" "cshtr\_f" "dvpsp\_f"   
## [281] "dvpsx\_f" "mkvalt" "prcc\_f" "prch\_f" "prcl\_f"   
## [286] "adjex\_f" "rank" "au" "auop" "auopic"   
## [291] "ceoso" "cfoso" "dpact\_fn" "rdipa\_fn" "rdipd\_fn"   
## [296] "rdipeps\_fn" "stkco\_fn" "add1" "addzip" "busdesc"   
## [301] "city" "conml" "ein" "fyrc" "ggroup"   
## [306] "gind" "gsector" "gsubind" "idbflag" "incorp"   
## [311] "loc" "naics" "phone" "priusa" "sic"   
## [316] "state" "stko" "weburl"

write.csv(fundamentals\_ds\_filter\_1, file = "data/fundamentals\_ds\_filter\_1.csv", row.names=FALSE)

### Remove all the columns which are related to general information of the company like address, phone, url etc…

#fundamentals\_ds\_filter\_1 <- fundamentals\_ds\_filter\_1[ lapply( fundamentals\_ds\_filter\_1, function(x) sum(is.na(x)) / length(x) ) < 0.25 ]  
ncol(fundamentals\_ds\_filter\_1)

## [1] 318

fundamentals\_ds\_filter\_1 <- subset(fundamentals\_ds\_filter\_1, select = -c(datadate,indfmt,curncd,consol,popsrc,conm,curcd,apdedate,fdate,add1,addzip,busdesc,  
 city,conml,weburl,phone,loc,final,fyr,acchg,fic,  
 xido,xidoc,naicsh,sich,au,auop,auopic,fyrc,ggroup,gind,gsector,gsubind,priusa))  
write.csv(fundamentals\_ds\_filter\_1, file = "data/fundamentals\_ds\_filter\_2.csv", row.names=FALSE, na="")  
ncol(fundamentals\_ds\_filter\_1)

## [1] 284

nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

### Identify columns which have values which are new to zero variance

nzv\_ds <- nearZeroVar(fundamentals\_ds\_filter\_1, saveMetrics = TRUE)  
nzv\_ds <- nzv\_ds[nzv\_ds[,"nzv"] > 0, ]  
nzv\_ds

## freqRatio percentUnique zeroVar nzv  
## datafmt 0.00000 0.08045052 TRUE TRUE  
## ajex 50.59091 3.05711987 FALSE TRUE  
## ajp 50.59091 3.05711987 FALSE TRUE  
## currtr 121.77778 5.79243765 FALSE TRUE  
## ismod 45.48000 0.16090105 FALSE TRUE  
## pddur 1239.00000 0.40225261 FALSE TRUE  
## scf 0.00000 0.08045052 TRUE TRUE  
## upd 94.61538 0.16090105 FALSE TRUE  
## acdo 1161.00000 3.13757039 FALSE TRUE  
## aldo 1187.00000 1.04585680 FALSE TRUE  
## aocisecgl 0.00000 0.08045052 TRUE TRUE  
## ciother 217.20000 7.88415125 FALSE TRUE  
## cstke 272.75000 7.88415125 FALSE TRUE  
## dcom 1177.00000 1.76991150 FALSE TRUE  
## dcvsr 276.75000 5.87288817 FALSE TRUE  
## dcvsub 291.75000 1.44810941 FALSE TRUE  
## dcvt 217.60000 6.99919549 FALSE TRUE  
## diladj 220.60000 4.82703138 FALSE TRUE  
## donr 538.50000 5.47063556 FALSE TRUE  
## drlt 561.00000 5.06838294 FALSE TRUE  
## ds 284.00000 2.73531778 FALSE TRUE  
## dudd 374.33333 5.14883347 FALSE TRUE  
## dvp 223.60000 5.06838294 FALSE TRUE  
## dvpa 1176.00000 1.20675784 FALSE TRUE  
## esopct 1168.00000 2.33306516 FALSE TRUE  
## esopnr 598.50000 0.64360418 FALSE TRUE  
## esopr 0.00000 0.08045052 TRUE TRUE  
## esopt 598.50000 0.64360418 FALSE TRUE  
## fatn 254.00000 1.20675784 FALSE TRUE  
## itcb 0.00000 0.08045052 TRUE TRUE  
## mib 383.66667 4.02252615 FALSE TRUE  
## pnrsho 208.60000 7.16009654 FALSE TRUE  
## prcad 21.23404 2.01126307 FALSE TRUE  
## prcaeps 20.72917 2.01126307 FALSE TRUE  
## prsho 393.33333 1.28720837 FALSE TRUE  
## pstk 93.90909 7.96460177 FALSE TRUE  
## pstkc 135.87500 5.55108608 FALSE TRUE  
## pstkl 92.63636 8.68865648 FALSE TRUE  
## pstkn 96.09091 6.19469027 FALSE TRUE  
## pstkr 393.33333 1.93081255 FALSE TRUE  
## pstkrv 92.63636 8.68865648 FALSE TRUE  
## rdip 1202.00000 0.16090105 FALSE TRUE  
## rdipa 1194.00000 0.16090105 FALSE TRUE  
## rdipd 1157.00000 0.16090105 FALSE TRUE  
## rdipeps 1157.00000 0.16090105 FALSE TRUE  
## rea 957.00000 5.06838294 FALSE TRUE  
## tstkp 297.00000 0.16090105 FALSE TRUE  
## txndbr 1120.00000 0.16090105 FALSE TRUE  
## txo 1106.00000 3.29847144 FALSE TRUE  
## txw 1026.00000 8.20595334 FALSE TRUE  
## xi 1202.00000 0.16090105 FALSE TRUE  
## xintopt 1025.00000 0.16090105 FALSE TRUE  
## xoptd 0.00000 0.08045052 TRUE TRUE  
## xopteps 0.00000 0.08045052 TRUE TRUE  
## adjex\_c 50.28571 2.97666935 FALSE TRUE  
## adjex\_f 50.38095 2.97666935 FALSE TRUE  
## rank 0.00000 0.08045052 TRUE TRUE  
## dpact\_fn 23.02174 0.32180209 FALSE TRUE  
## rdipa\_fn 0.00000 0.08045052 TRUE TRUE  
## rdipd\_fn 285.75000 0.16090105 FALSE TRUE  
## rdipeps\_fn 228.40000 0.16090105 FALSE TRUE  
## stkco\_fn 0.00000 0.08045052 TRUE TRUE

### Remove columns which have values which are of zero variance

nzv\_ds\_cols <- nearZeroVar(fundamentals\_ds\_filter\_1)  
fundamentals\_ds\_filter\_1 <- fundamentals\_ds\_filter\_1[, -nzv\_ds\_cols]  
write.csv(fundamentals\_ds\_filter\_1, file = "data/fundamentals\_ds\_filter\_3.csv", row.names=FALSE, na="")

ncol(fundamentals\_ds\_filter\_1)

## [1] 222

nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

### Based on manual analysis of the columns and considering the end goal, remove the variables which logically do not make sense or might have correlattion with existing set of variables.

fundamentals\_ds\_filter\_1 <- subset(fundamentals\_ds\_filter\_1, select = -c(acctstd,src,acodo,acox,  
 aox,capxv,ceql,cibegni,cicurr,cidergl,  
 cimii,cipen,cisecgl,citotal,cshfd,cshpri,dclo,dcpstk,  
 dltis,dlto,dltr,do,dp,dpvieb,drc,  
 dv,dvc,epsfx,epspx,exre,  
 fopox,ibadj,ibc,ibcom,ibmii,  
 invch,ivaco,ivao,lco,  
 lcox,lcoxdr,lct,loxdr,mibn,  
 mibt,mii,msa,niadj,np,oprepsx,pnca,ppent,ppeveb,  
 recco,rectr,sale,spced,spceeps,  
 tstkc,txbco,txbcof,txdb,  
 txdbca,txdi,  
 txditc,txndb,xopr,exchg,costat,  
 ceoso,cfoso,idbflag,naics,sic,stko))  
write.csv(fundamentals\_ds\_filter\_1, file = "data/fundamentals\_ds\_filter\_4.csv", row.names=FALSE, na="")

colnames(fundamentals\_ds\_filter\_1)

## [1] "gvkey" "fyear" "tic" "aco" "acominc" "act"   
## [7] "am" "ao" "aocidergl" "aociother" "aocipen" "aodo"   
## [13] "aoloch" "ap" "aqc" "at" "bkvlps" "caps"   
## [19] "capx" "ceq" "ceqt" "ch" "che" "chech"   
## [25] "ci" "cogs" "cshi" "csho" "cshr" "cstk"   
## [31] "cstkcv" "dc" "dd" "dd1" "dd2" "dd3"   
## [37] "dd4" "dd5" "dilavx" "dlc" "dltp" "dltt"   
## [43] "dm" "dn" "dpact" "dpc" "dvt" "ebit"   
## [49] "ebitda" "emp" "epsfi" "epspi" "esub" "esubc"   
## [55] "fatb" "fatc" "fatp" "fiao" "fincf" "fopo"   
## [61] "gdwl" "gp" "ib" "icapt" "intan" "intano"   
## [67] "intc" "intpn" "invt" "ivaeq" "ivch" "ivncf"   
## [73] "ivst" "ivstch" "lifr" "lo" "lse" "lt"   
## [79] "mrc1" "mrcta" "ni" "nopi" "nopio" "oancf"   
## [85] "oiadp" "oibdp" "opeps" "pi" "pncad" "pncaeps"   
## [91] "ppegt" "prca" "prstkc" "re" "reajo" "recch"   
## [97] "recd" "rect" "recta" "reuna" "revt" "seq"   
## [103] "seqo" "siv" "spce" "spi" "sppe" "sppiv"   
## [109] "sstk" "stkco" "teq" "tstk" "tstkn" "txc"   
## [115] "txdba" "txdbcl" "txdc" "txfed" "txfo" "txndba"   
## [121] "txndbl" "txp" "txpd" "txr" "txs" "txt"   
## [127] "wcap" "xacc" "xint" "xrent" "xsga" "cshtr\_c"   
## [133] "dvpsp\_c" "dvpsx\_c" "prcc\_c" "prch\_c" "prcl\_c" "cshtr\_f"   
## [139] "dvpsp\_f" "dvpsx\_f" "mkvalt" "prcc\_f" "prch\_f" "prcl\_f"   
## [145] "ein" "incorp" "state"

nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

#fundamentals\_ds\_filter\_1

### Filter and create a seperate dataset for restatement

#### Remove all the columns which contain 10% or greater NA or null values

fundamentals\_restmt\_ds\_filter <- fundamentals\_ds %>%  
 filter(datafmt == 'SUMM\_STD' & gsector == gsector\_sjm)  
 #filter(datafmt == 'SUMM\_STD')  
   
std\_cols <- colnames(fundamentals\_ds\_filter\_1)  
fundamentals\_restmt\_ds\_filter <- subset(fundamentals\_restmt\_ds\_filter, select = c(std\_cols))  
fundamentals\_restmt\_ds\_filter <- fundamentals\_restmt\_ds\_filter[ lapply( fundamentals\_restmt\_ds\_filter, function(x) sum(is.na(x)) / length(x) ) < 0.1 ]  
summary(fundamentals\_restmt\_ds\_filter)

## gvkey fyear tic at   
## Min. : 1239 Min. :2009 0161A : 5 Min. : 0.00   
## 1st Qu.: 10852 1st Qu.:2010 0173A : 5 1st Qu.: 19.66   
## Median : 29517 Median :2011 AOI : 5 Median : 317.63   
## Mean : 75238 Mean :2011 BF.B : 5 Mean : 5919.97   
## 3rd Qu.:162517 3rd Qu.:2012 BNNY : 5 3rd Qu.: 2767.12   
## Max. :264393 Max. :2013 CASY : 5 Max. :204751.00   
## (Other):1050 NA's :20   
## capx cogs dltt epsfi   
## Min. : 0.000 Min. : 0 Min. : 0.00 Min. : -70.300   
## 1st Qu.: 0.366 1st Qu.: 13 1st Qu.: 0.00 1st Qu.: -0.035   
## Median : 12.813 Median : 245 Median : 16.23 Median : 0.430   
## Mean : 229.376 Mean : 5040 Mean : 1469.99 Mean : 4.928   
## 3rd Qu.: 101.290 3rd Qu.: 2172 3rd Qu.: 789.47 3rd Qu.: 1.985   
## Max. :13510.000 Max. :349199 Max. :47079.00 Max. :1126.180   
## NA's :41 NA's :40 NA's :33 NA's :65   
## epspi ib ni nopi   
## Min. : -70.300 Min. :-1728.282 Min. :-1575.621 Min. :-2366.000   
## 1st Qu.: -0.040 1st Qu.: -0.682 1st Qu.: -0.693 1st Qu.: -5.813   
## Median : 0.395 Median : 11.068 Median : 10.489 Median : 0.000   
## Mean : 5.007 Mean : 424.227 Mean : 413.637 Mean : 10.569   
## 3rd Qu.: 1.960 3rd Qu.: 108.980 3rd Qu.: 110.281 3rd Qu.: 0.738   
## Max. :1126.180 Max. :16963.000 Max. :16999.000 Max. : 8234.000   
## NA's :86 NA's :27 NA's :30 NA's :44   
## pi reuna seq teq   
## Min. :-2251.837 Min. :-7883.37 Min. :-7766.00 Min. :-6274.00   
## 1st Qu.: -0.712 1st Qu.: -9.31 1st Qu.: 4.14 1st Qu.: 4.13   
## Median : 16.577 Median : 27.36 Median : 110.65 Median : 110.66   
## Mean : 577.995 Mean : 2151.53 Mean : 2123.17 Mean : 2184.02   
## 3rd Qu.: 158.523 3rd Qu.: 443.86 3rd Qu.: 924.87 3rd Qu.: 976.52   
## Max. :25662.000 Max. :80197.00 Max. :76343.00 Max. :81738.00   
## NA's :33 NA's :35 NA's :17 NA's :21   
## txt wcap xint xsga   
## Min. :-523.555 Min. :-11878.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.: 0.337 1st Qu.: 0.150 1st Qu.: 7.22   
## Median : 3.554 Median : 27.381 Median : 3.711 Median : 87.06   
## Mean : 166.817 Mean : 213.857 Mean : 89.108 Mean : 1392.23   
## 3rd Qu.: 48.900 3rd Qu.: 257.084 3rd Qu.: 61.965 3rd Qu.: 662.60   
## Max. :8105.000 Max. : 9900.000 Max. :3341.000 Max. :90920.00   
## NA's :30 NA's :64 NA's :107 NA's :73   
## dvpsp\_c dvpsx\_c dvpsp\_f dvpsx\_f   
## Min. : 0.0000 Min. : 0.0000 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 0.0000 Median : 0.0000 Median : 0.0000 Median : 0.0000   
## Mean : 0.4543 Mean : 0.4558 Mean : 0.4490 Mean : 0.4505   
## 3rd Qu.: 0.5867 3rd Qu.: 0.5950 3rd Qu.: 0.5769 3rd Qu.: 0.5825   
## Max. :21.0000 Max. :21.0000 Max. :21.0000 Max. :21.0000   
## NA's :65 NA's :65 NA's :62 NA's :62   
## ein incorp   
## 13-4306188: 5 DE :495   
## 16-0716709: 5 NV :179   
## 16-0733425: 5 FL : 43   
## 20-1266625: 5 VA : 35   
## 23-1614034: 5 CO : 32   
## (Other) :1014 (Other):241   
## NA's : 41 NA's : 55

nrow(fundamentals\_restmt\_ds\_filter)

## [1] 1080

### Add restement variables and the magnitude as pecentages

sample\_restmt\_ds\_filter <- fundamentals\_restmt\_ds\_filter #%>%  
 #filter(gvkey == 1076)  
sample\_ds\_filter <- fundamentals\_ds\_filter\_1 #%>%  
 #filter(gvkey == 1076)  
#nrow(sample\_restmt\_ds\_filter)  
#nrow(sample\_ds\_filter)  
#head(sample\_restmt\_ds\_filter)  
#head(sample\_ds\_filter)  
  
fundamentals\_ds\_filter\_1$restmt\_at <- 0  
fundamentals\_ds\_filter\_1$restmt\_at\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_capx <- 0  
fundamentals\_ds\_filter\_1$restmt\_capx\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_cogs <- 0  
fundamentals\_ds\_filter\_1$restmt\_cogs\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_dltt <- 0  
fundamentals\_ds\_filter\_1$restmt\_dltt\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_epsfi <- 0  
fundamentals\_ds\_filter\_1$restmt\_epsfi\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_epspi <- 0  
fundamentals\_ds\_filter\_1$restmt\_epspi\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_ib <- 0  
fundamentals\_ds\_filter\_1$restmt\_ib\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_ni <- 0  
fundamentals\_ds\_filter\_1$restmt\_ni\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_nopi <- 0  
fundamentals\_ds\_filter\_1$restmt\_nopi\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_pi <- 0  
fundamentals\_ds\_filter\_1$restmt\_pi\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_reuna <- 0  
fundamentals\_ds\_filter\_1$restmt\_reuna\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_seq <- 0  
fundamentals\_ds\_filter\_1$restmt\_seq\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_teq <- 0  
fundamentals\_ds\_filter\_1$restmt\_teq\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_txt <- 0  
fundamentals\_ds\_filter\_1$restmt\_txt\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_wcap <- 0  
fundamentals\_ds\_filter\_1$restmt\_wcap\_mag <- 0.0  
  
#fundamentals\_ds\_filter\_1$restmt\_ci <- 0  
#fundamentals\_ds\_filter\_1$restmt\_ci\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_xint <- 0  
fundamentals\_ds\_filter\_1$restmt\_xint\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_xsga <- 0  
fundamentals\_ds\_filter\_1$restmt\_xsga\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_dvpsp\_f <- 0  
fundamentals\_ds\_filter\_1$restmt\_dvpsp\_f\_mag <- 0.0  
  
fundamentals\_ds\_filter\_1$restmt\_dvpsx\_f <- 0  
fundamentals\_ds\_filter\_1$restmt\_dvpsx\_f\_mag <- 0.0  
  
for (row in 1:nrow(sample\_restmt\_ds\_filter)){  
 restmt\_item\_gvkey <- as.integer(sample\_restmt\_ds\_filter[row, "gvkey"])  
 restmt\_item\_fyear <- sample\_restmt\_ds\_filter[row, "fyear"]  
 restmt\_item\_at <- sample\_restmt\_ds\_filter[row, "at"]  
 restmt\_item\_capx <- sample\_restmt\_ds\_filter[row, "capx"]  
 restmt\_item\_cogs <- sample\_restmt\_ds\_filter[row, "cogs"]  
 restmt\_item\_dltt <- sample\_restmt\_ds\_filter[row, "dltt"]  
 restmt\_item\_epsfi <- sample\_restmt\_ds\_filter[row, "epsfi"]  
 restmt\_item\_epspi <- sample\_restmt\_ds\_filter[row, "epspi"]  
   
 restmt\_item\_ib <- sample\_restmt\_ds\_filter[row, "ib"]  
 restmt\_item\_ni <- sample\_restmt\_ds\_filter[row, "ni"]  
 restmt\_item\_nopi <- sample\_restmt\_ds\_filter[row, "nopi"]  
 restmt\_item\_pi <- sample\_restmt\_ds\_filter[row, "pi"]  
 restmt\_item\_reuna <- sample\_restmt\_ds\_filter[row, "reuna"]  
 restmt\_item\_seq <- sample\_restmt\_ds\_filter[row, "seq"]  
 restmt\_item\_teq <- sample\_restmt\_ds\_filter[row, "teq"]  
 restmt\_item\_txt <- sample\_restmt\_ds\_filter[row, "txt"]  
 restmt\_item\_wcap <- sample\_restmt\_ds\_filter[row, "wcap"]  
   
 restmt\_item\_xint <- sample\_restmt\_ds\_filter[row, "xint"]  
 restmt\_item\_xsga <- sample\_restmt\_ds\_filter[row, "xsga"]  
 restmt\_item\_dvpsp\_f <- sample\_restmt\_ds\_filter[row, "dvpsp\_f"]  
 restmt\_item\_dvpsx\_f <- sample\_restmt\_ds\_filter[row, "dvpsx\_f"]  
  
 row\_count <- as.integer(nrow(subset(fundamentals\_ds\_filter\_1, gvkey == restmt\_item\_gvkey & fyear == restmt\_item\_fyear)))  
   
 if (row\_count > 0){  
 fundamental\_stmt\_row <- fundamentals\_ds\_filter\_1 %>%  
 filter(gvkey == restmt\_item\_gvkey & fyear == restmt\_item\_fyear)  
  
 stmt\_item\_gvkey <- fundamental\_stmt\_row["gvkey"]  
 stmt\_item\_fyear <- fundamental\_stmt\_row["fyear"]  
 stmt\_item\_at <- fundamental\_stmt\_row["at"]  
 stmt\_item\_capx <- fundamental\_stmt\_row["capx"]  
 stmt\_item\_cogs <- fundamental\_stmt\_row["cogs"]  
 stmt\_item\_dltt <- fundamental\_stmt\_row["dltt"]  
 stmt\_item\_epsfi <- fundamental\_stmt\_row["epsfi"]  
 stmt\_item\_epspi <- fundamental\_stmt\_row["epspi"]  
 stmt\_item\_ib <- fundamental\_stmt\_row["ib"]  
 stmt\_item\_ni <- fundamental\_stmt\_row["ni"]  
 stmt\_item\_nopi <- fundamental\_stmt\_row["nopi"]  
 stmt\_item\_pi <- fundamental\_stmt\_row["pi"]  
 stmt\_item\_reuna <- fundamental\_stmt\_row["reuna"]  
 stmt\_item\_seq <- fundamental\_stmt\_row["seq"]  
 stmt\_item\_teq <- fundamental\_stmt\_row["teq"]  
 stmt\_item\_txt <- fundamental\_stmt\_row["txt"]  
 stmt\_item\_wcap <- fundamental\_stmt\_row["wcap"]  
 stmt\_item\_xint <- fundamental\_stmt\_row["xint"]  
 stmt\_item\_xsga <- fundamental\_stmt\_row["xsga"]  
 stmt\_item\_dvpsp\_f <- fundamental\_stmt\_row["dvpsp\_f"]  
 stmt\_item\_dvpsx\_f <- fundamental\_stmt\_row["dvpsx\_f"]  
  
   
  
 if (!is.na(restmt\_item\_at) & !is.na(stmt\_item\_at) & stmt\_item\_at != 0 & restmt\_item\_at != stmt\_item\_at){  
 fundamentals\_ds\_filter\_1$restmt\_at[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_at - stmt\_item\_at)/stmt\_item\_at) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_at\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_capx) & !is.na(stmt\_item\_capx) & restmt\_item\_capx != stmt\_item\_capx){  
 fundamentals\_ds\_filter\_1$restmt\_capx[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_capx == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_capx - stmt\_item\_capx)/stmt\_item\_capx) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_capx\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_cogs) & !is.na(stmt\_item\_cogs) & restmt\_item\_cogs != stmt\_item\_cogs){  
 fundamentals\_ds\_filter\_1$restmt\_cogs[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_cogs == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_cogs - stmt\_item\_cogs)/stmt\_item\_cogs) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_cogs\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_dltt) & !is.na(stmt\_item\_dltt) & restmt\_item\_dltt != stmt\_item\_dltt){  
 fundamentals\_ds\_filter\_1$restmt\_dltt[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_dltt == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_dltt - stmt\_item\_dltt)/stmt\_item\_dltt) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_dltt\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_epsfi) & !is.na(stmt\_item\_epsfi) & restmt\_item\_epsfi != stmt\_item\_epsfi){  
 fundamentals\_ds\_filter\_1$restmt\_epsfi[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_epsfi == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_epsfi - stmt\_item\_epsfi)/stmt\_item\_epsfi) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_epsfi\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_epspi) & !is.na(stmt\_item\_epspi) & restmt\_item\_epspi != stmt\_item\_epspi){  
 fundamentals\_ds\_filter\_1$restmt\_epspi[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_epspi == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_epspi - stmt\_item\_epspi)/stmt\_item\_epspi) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_epspi\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_ib) & !is.na(stmt\_item\_ib) & restmt\_item\_ib != stmt\_item\_ib){  
 fundamentals\_ds\_filter\_1$restmt\_ib[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_ib - stmt\_item\_ib)/stmt\_item\_ib) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_ib\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_ni) & !is.na(stmt\_item\_ni) & restmt\_item\_ni != stmt\_item\_ni){  
 fundamentals\_ds\_filter\_1$restmt\_ni[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_ni - stmt\_item\_ni)/stmt\_item\_ni) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_ni\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_nopi) & !is.na(stmt\_item\_nopi) & restmt\_item\_nopi != stmt\_item\_nopi){  
 fundamentals\_ds\_filter\_1$restmt\_nopi[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_nopi == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_nopi - stmt\_item\_nopi)/stmt\_item\_nopi) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_nopi\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_pi) & !is.na(stmt\_item\_pi) & restmt\_item\_pi != stmt\_item\_pi){  
 fundamentals\_ds\_filter\_1$restmt\_pi[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_pi - stmt\_item\_pi)/stmt\_item\_pi) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_pi\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_reuna) & !is.na(stmt\_item\_reuna) & restmt\_item\_reuna != stmt\_item\_reuna){  
 fundamentals\_ds\_filter\_1$restmt\_reuna[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_reuna - stmt\_item\_reuna)/stmt\_item\_reuna) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_reuna\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_seq) & !is.na(stmt\_item\_seq) & restmt\_item\_seq != stmt\_item\_seq){  
 fundamentals\_ds\_filter\_1$restmt\_seq[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_seq - stmt\_item\_seq)/stmt\_item\_seq) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_seq\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_teq) & !is.na(stmt\_item\_teq) & restmt\_item\_teq != stmt\_item\_teq){  
 fundamentals\_ds\_filter\_1$restmt\_teq[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_teq - stmt\_item\_teq)/stmt\_item\_teq) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_teq\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_txt) & !is.na(stmt\_item\_txt) & restmt\_item\_txt != stmt\_item\_txt){  
 fundamentals\_ds\_filter\_1$restmt\_txt[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 if (stmt\_item\_txt == 0.0){  
 magnitude <- 100.00  
 }  
 else{  
 magnitude <- ((restmt\_item\_txt - stmt\_item\_txt)/stmt\_item\_txt) \* 100.0  
 }  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_txt\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
  
 if (!is.na(restmt\_item\_wcap) & !is.na(stmt\_item\_wcap) & restmt\_item\_wcap != stmt\_item\_wcap){  
 fundamentals\_ds\_filter\_1$restmt\_wcap[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_wcap - stmt\_item\_wcap)/stmt\_item\_wcap) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_wcap\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
   
 if (!is.na(restmt\_item\_xint) & !is.na(stmt\_item\_xint) & stmt\_item\_xint != 0 & restmt\_item\_xint != stmt\_item\_xint){  
 fundamentals\_ds\_filter\_1$restmt\_xint[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_xint - stmt\_item\_xint)/stmt\_item\_xint) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_xint\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
   
 if (!is.na(restmt\_item\_xsga) & !is.na(stmt\_item\_xsga) & restmt\_item\_xsga != stmt\_item\_xsga){  
 fundamentals\_ds\_filter\_1$restmt\_xsga[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_xsga - stmt\_item\_xsga)/stmt\_item\_xsga) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_xsga\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
   
 if (!is.na(restmt\_item\_dvpsp\_f) & !is.na(stmt\_item\_dvpsp\_f) & restmt\_item\_dvpsp\_f != stmt\_item\_dvpsp\_f){  
 fundamentals\_ds\_filter\_1$restmt\_dvpsp\_f[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_dvpsp\_f - stmt\_item\_dvpsp\_f)/stmt\_item\_dvpsp\_f) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_dvpsp\_f\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
   
 if (!is.na(restmt\_item\_dvpsx\_f) & !is.na(stmt\_item\_dvpsx\_f) & restmt\_item\_dvpsx\_f != stmt\_item\_dvpsx\_f){  
 fundamentals\_ds\_filter\_1$restmt\_dvpsx\_f[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- 1  
 magnitude <- ((restmt\_item\_dvpsx\_f - stmt\_item\_dvpsx\_f)/stmt\_item\_dvpsx\_f) \* 100.0  
 magnitude <- as.double(round(magnitude, digits = 3))  
 fundamentals\_ds\_filter\_1$restmt\_dvpsx\_f\_mag[fundamentals\_ds\_filter\_1$gvkey == restmt\_item\_gvkey & fundamentals\_ds\_filter\_1$fyear == restmt\_item\_fyear] <- magnitude  
 }  
   
 }  
}  
#head(fundamentals\_ds\_filter\_1)  
nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

### Removing aco == NA as all those rows do not have any data

nrow(fundamentals\_ds\_filter\_1)

## [1] 1243

fundamentals\_ds\_filter\_2 <- subset(fundamentals\_ds\_filter\_1, !is.na(aco))  
nrow(fundamentals\_ds\_filter\_2)

## [1] 1205

fundamentals\_ds\_filter\_2[is.na(fundamentals\_ds\_filter\_2)] <- 0  
summary(fundamentals\_ds\_filter\_2)

## gvkey fyear tic aco   
## Min. : 1239 Min. :2009 0161A : 5 Min. : 0.000   
## 1st Qu.: 11178 1st Qu.:2010 0173A : 5 1st Qu.: 0.688   
## Median : 30651 Median :2011 AOI : 5 Median : 14.229   
## Mean : 78271 Mean :2011 BF.B : 5 Mean : 216.805   
## 3rd Qu.:163887 3rd Qu.:2012 BNNY : 5 3rd Qu.: 126.500   
## Max. :277487 Max. :2013 CAG : 5 Max. :6593.000   
## (Other):1175   
## acominc act am ao   
## Min. :-23363.657 Min. : 0.00 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -43.616 1st Qu.: 14.54 1st Qu.: 0.000 1st Qu.: 0.251   
## Median : -0.009 Median : 209.70 Median : 0.184 Median : 13.293   
## Mean : -223.117 Mean : 2239.24 Mean : 22.033 Mean : 231.775   
## 3rd Qu.: 0.000 3rd Qu.: 1673.25 3rd Qu.: 6.200 3rd Qu.: 125.966   
## Max. : 5241.118 Max. :61185.00 Max. :736.211 Max. :6847.000   
##   
## aocidergl aociother aocipen aodo   
## Min. :-5300.00 Min. :-7685.00 Min. :-4296.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: -15.90 1st Qu.: 0.124   
## Median : 0.00 Median : 0.00 Median : 0.00 Median : 9.919   
## Mean : -10.23 Mean : -16.07 Mean : -108.57 Mean : 217.871   
## 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 117.008   
## Max. : 455.00 Max. : 1576.49 Max. : 85.61 Max. :6847.000   
##   
## aoloch ap aqc at   
## Min. :-1738.853 Min. : 0.00 Min. : -684.417 Min. : 0.0   
## 1st Qu.: -3.612 1st Qu.: 2.07 1st Qu.: 0.000 1st Qu.: 32.4   
## Median : 0.000 Median : 39.08 Median : 0.000 Median : 565.0   
## Mean : 7.200 Mean : 776.93 Mean : 122.530 Mean : 7580.3   
## 3rd Qu.: 4.886 3rd Qu.: 310.32 3rd Qu.: 1.723 3rd Qu.: 4400.0   
## Max. : 2141.000 Max. :38080.00 Max. :17538.000 Max. :204751.0   
##   
## bkvlps caps capx ceq   
## Min. :-142340 Min. : -782.34 Min. : 0.000 Min. :-7766.00   
## 1st Qu.: 0 1st Qu.: 5.38 1st Qu.: 0.529 1st Qu.: 7.42   
## Median : 5 Median : 50.10 Median : 19.535 Median : 166.72   
## Mean : 9513 Mean : 1101.50 Mean : 288.775 Mean : 2790.13   
## 3rd Qu.: 13 3rd Qu.: 539.48 3rd Qu.: 157.216 3rd Qu.: 1485.40   
## Max. :4231100 Max. :63538.00 Max. :13510.000 Max. :76343.00   
##   
## ceqt ch che   
## Min. :-48900.00 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -4.50 1st Qu.: 1.722 1st Qu.: 2.421   
## Median : 20.38 Median : 26.881 Median : 38.884   
## Mean : 118.05 Mean : 449.550 Mean : 561.182   
## 3rd Qu.: 257.78 3rd Qu.: 242.818 3rd Qu.: 307.335   
## Max. : 56745.00 Max. :12803.000 Max. :20268.000   
##   
## chech ci cogs cshi   
## Min. :-4361.000 Min. :-1645.04 Min. : 0.0 Min. : 0.00   
## 1st Qu.: -2.195 1st Qu.: -0.43 1st Qu.: 17.9 1st Qu.: 15.51   
## Median : 0.274 Median : 16.03 Median : 407.3 Median : 49.95   
## Mean : 49.167 Mean : 562.03 Mean : 6143.1 Mean : 297.18   
## 3rd Qu.: 23.018 3rd Qu.: 220.28 3rd Qu.: 3390.7 3rd Qu.: 163.73   
## Max. : 4295.100 Max. :32850.89 Max. :349199.0 Max. :15664.33   
##   
## csho cshr cstk cstkcv   
## Min. : 0.00 Min. : 0.000 Min. : 0.000 Min. : 0.0000   
## 1st Qu.: 15.70 1st Qu.: 0.026 1st Qu.: 0.026 1st Qu.: 0.0010   
## Median : 49.33 Median : 0.371 Median : 0.413 Median : 0.0100   
## Mean : 266.38 Mean : 14.241 Mean : 218.565 Mean : 0.6711   
## 3rd Qu.: 157.79 3rd Qu.: 3.481 3rd Qu.: 39.000 3rd Qu.: 0.3200   
## Max. :15662.93 Max. :2311.000 Max. :24144.697 Max. :80.3400   
##   
## dc dd dd1 dd2   
## Min. : 0.000 Min. : 0.0 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 0.00   
## Median : 0.000 Median : 0.0 Median : 1.304 Median : 0.21   
## Mean : 5.573 Mean : 391.5 Mean : 200.709 Mean : 145.27   
## 3rd Qu.: 0.000 3rd Qu.: 0.0 3rd Qu.: 34.376 3rd Qu.: 19.00   
## Max. :901.000 Max. :40526.0 Max. :7846.000 Max. :5748.00   
##   
## dd3 dd4 dd5 dilavx   
## Min. : 0.000 Min. : 0.00 Min. : 0.000 Min. :-1579.237   
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: -0.354   
## Median : 0.031 Median : 0.00 Median : 0.000 Median : 14.425   
## Mean : 139.008 Mean : 119.12 Mean : 139.030 Mean : 540.741   
## 3rd Qu.: 17.921 3rd Qu.: 10.04 3rd Qu.: 8.526 3rd Qu.: 221.645   
## Max. :5658.000 Max. :5247.00 Max. :5971.641 Max. :16999.000   
##   
## dlc dltp dltt dm   
## Min. : 0.000 Min. :-199.212 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.248 1st Qu.: 0.000 1st Qu.: 0.02 1st Qu.: 0.000   
## Median : 6.107 Median : 0.000 Median : 42.74 Median : 0.637   
## Mean : 435.065 Mean : 149.923 Mean : 1736.06 Mean : 177.689   
## 3rd Qu.: 111.090 3rd Qu.: 9.112 3rd Qu.: 1216.60 3rd Qu.: 61.622   
## Max. :20281.813 Max. :5629.040 Max. :47079.00 Max. :4413.000   
##   
## dn dpact dpc dvt   
## Min. : 0 Min. : 0.00 Min. : 0.000 Min. : -0.457   
## 1st Qu.: 0 1st Qu.: 4.36 1st Qu.: 0.507 1st Qu.: 0.000   
## Median : 0 Median : 93.46 Median : 11.751 Median : 0.000   
## Mean : 1089 Mean : 1548.38 Mean : 202.612 Mean : 261.726   
## 3rd Qu.: 400 3rd Qu.: 939.05 3rd Qu.: 114.538 3rd Qu.: 61.738   
## Max. :45073 Max. :60771.00 Max. :8870.000 Max. :7358.491   
##   
## ebit ebitda emp epsfi   
## Min. : -348.830 Min. : -150.53 Min. : 0.000 Min. : -18.340   
## 1st Qu.: 0.108 1st Qu.: 1.18 1st Qu.: 0.076 1st Qu.: -0.010   
## Median : 41.259 Median : 63.07 Median : 1.797 Median : 0.500   
## Mean : 890.581 Mean : 1105.73 Mean : 30.407 Mean : 4.659   
## 3rd Qu.: 471.209 3rd Qu.: 634.12 3rd Qu.: 14.800 3rd Qu.: 2.020   
## Max. :26027.000 Max. :34528.00 Max. :2200.000 Max. :1126.180   
##   
## epspi esub esubc fatb   
## Min. : -18.340 Min. : -35.0 Min. :-1078.02 Min. : 0.00   
## 1st Qu.: -0.010 1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.: 0.00   
## Median : 0.500 Median : 0.0 Median : 0.00 Median : 4.54   
## Mean : 4.677 Mean : 30.2 Mean : -13.12 Mean : 840.15   
## 3rd Qu.: 2.030 3rd Qu.: 0.0 3rd Qu.: 0.00 3rd Qu.: 239.91   
## Max. :1126.180 Max. :1419.6 Max. : 100.21 Max. :95488.00   
##   
## fatc fatp fiao fincf   
## Min. : 0.00 Min. : 0.000 Min. :-9494.08 Min. :-27546.163   
## 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: -5.60 1st Qu.: -134.000   
## Median : 0.00 Median : 0.347 Median : 0.00 Median : -1.583   
## Mean : 85.38 Mean : 217.919 Mean : -50.73 Mean : -390.433   
## 3rd Qu.: 21.64 3rd Qu.: 49.619 3rd Qu.: 0.00 3rd Qu.: 2.296   
## Max. :5828.00 Max. :26184.000 Max. :10337.10 Max. : 4188.000   
##   
## fopo gdwl gp ib   
## Min. :-5386.000 Min. : 0.00 Min. : -49.55 Min. :-1579.237   
## 1st Qu.: 0.053 1st Qu.: 0.00 1st Qu.: 11.33 1st Qu.: -0.353   
## Median : 2.511 Median : 8.32 Median : 194.64 Median : 17.409   
## Mean : 84.244 Mean : 1717.86 Mean : 2958.37 Mean : 549.144   
## 3rd Qu.: 29.462 3rd Qu.: 535.00 3rd Qu.: 1543.17 3rd Qu.: 243.376   
## Max. : 2526.000 Max. :69927.00 Max. :125060.00 Max. :16999.000   
##   
## icapt intan intano intc   
## Min. : -647.66 Min. : 0.00 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 20.63 1st Qu.: 0.38 1st Qu.: 0.01 1st Qu.: 0.000   
## Median : 381.19 Median : 31.93 Median : 9.11 Median : 0.000   
## Mean : 4694.26 Mean : 2716.80 Mean : 998.94 Mean : 1.713   
## 3rd Qu.: 2881.55 3rd Qu.: 1012.93 3rd Qu.: 332.00 3rd Qu.: 0.000   
## Max. :127389.00 Max. :99265.00 Max. :32620.00 Max. :110.000   
##   
## intpn invt ivaeq ivch   
## Min. : -0.046 Min. : 0.00 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 0.024 1st Qu.: 4.12 1st Qu.: 0.0 1st Qu.: 0.00   
## Median : 1.925 Median : 71.76 Median : 0.0 Median : 0.00   
## Mean : 92.475 Mean : 890.93 Mean : 222.3 Mean : 69.64   
## 3rd Qu.: 67.000 3rd Qu.: 648.95 3rd Qu.: 0.9 3rd Qu.: 0.00   
## Max. :2612.000 Max. :44858.00 Max. :13830.9 Max. :14782.00   
##   
## ivncf ivst ivstch lifr   
## Min. :-16609.000 Min. : 0.000 Min. :-6702.000 Min. : -87.00   
## 1st Qu.: -195.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.00   
## Median : -24.559 Median : 0.000 Median : 0.000 Median : 0.00   
## Mean : -370.668 Mean : 101.460 Mean : -3.398 Mean : 29.28   
## 3rd Qu.: -0.342 3rd Qu.: 1.753 3rd Qu.: 0.000 3rd Qu.: 0.00   
## Max. : 15528.872 Max. :9854.000 Max. : 6707.000 Max. :2100.00   
##   
## lo lse lt mrc1   
## Min. :-8821.23 Min. : 0.0 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 32.4 1st Qu.: 11.47 1st Qu.: 0.097   
## Median : 10.99 Median : 565.0 Median : 227.58 Median : 4.200   
## Mean : 543.28 Mean : 7580.3 Mean : 4610.42 Mean : 74.821   
## 3rd Qu.: 287.15 3rd Qu.: 4400.0 3rd Qu.: 3149.50 3rd Qu.: 41.407   
## Max. :19714.69 Max. :204751.0 Max. :121921.00 Max. :2536.000   
##   
## mrcta ni nopi nopio   
## Min. : 0.000 Min. :-1575.62 Min. :-686.000 Min. :-686.000   
## 1st Qu.: 0.000 1st Qu.: -0.36 1st Qu.: 0.000 1st Qu.: -0.004   
## Median : 2.459 Median : 17.43 Median : 0.262 Median : 0.075   
## Mean : 331.950 Mean : 586.58 Mean : 57.630 Mean : 46.064   
## 3rd Qu.: 57.463 3rd Qu.: 246.64 3rd Qu.: 8.000 3rd Qu.: 4.000   
## Max. :25428.000 Max. :36538.58 Max. :2377.000 Max. :2196.000   
##   
## oancf oiadp oibdp opeps   
## Min. :-2435.00 Min. : -348.830 Min. : -150.53 Min. : -11.330   
## 1st Qu.: 0.00 1st Qu.: 0.108 1st Qu.: 1.18 1st Qu.: 0.000   
## Median : 34.25 Median : 41.259 Median : 63.07 Median : 0.600   
## Mean : 818.08 Mean : 890.581 Mean : 1105.73 Mean : 4.693   
## 3rd Qu.: 394.71 3rd Qu.: 471.209 3rd Qu.: 634.12 3rd Qu.: 2.140   
## Max. :26249.00 Max. :26027.000 Max. :34528.00 Max. :1126.180   
##   
## pi pncad pncaeps ppegt   
## Min. :-2052.598 Min. :-3.6800 Min. :-3.6800 Min. : 0.00   
## 1st Qu.: -0.309 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 11.03   
## Median : 25.386 Median : 0.0000 Median : 0.0000 Median : 237.26   
## Mean : 796.445 Mean : 0.1224 Mean : 0.1226 Mean : 3502.95   
## 3rd Qu.: 340.990 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 2046.90   
## Max. :25737.000 Max. :54.5500 Max. :54.5500 Max. :178678.00   
##   
## prca prstkc re   
## Min. :-261.3000 Min. : 0.000 Min. :-29020.54   
## 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: -7.75   
## Median : 0.0000 Median : 0.000 Median : 39.15   
## Mean : -0.7143 Mean : 210.080 Mean : 2240.66   
## 3rd Qu.: 0.0000 3rd Qu.: 8.481 3rd Qu.: 781.40   
## Max. : 95.5500 Max. :14776.000 Max. : 73570.00   
##   
## reajo recch recd rect   
## Min. :-28991.49 Min. :-2882.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -26.71 1st Qu.: -9.045 1st Qu.: 0.000 1st Qu.: 2.419   
## Median : 0.00 Median : -0.243 Median : 0.195 Median : 48.531   
## Mean : -92.22 Mean : -23.466 Mean : 12.795 Mean : 585.366   
## 3rd Qu.: 0.00 3rd Qu.: 0.063 3rd Qu.: 4.000 3rd Qu.: 427.000   
## Max. : 10590.65 Max. : 1121.637 Max. :620.109 Max. :15764.063   
##   
## recta reuna revt seq   
## Min. :-23372.642 Min. :-7883.37 Min. : 0.0 Min. :-7766.0   
## 1st Qu.: -0.225 1st Qu.: -7.21 1st Qu.: 34.6 1st Qu.: 10.1   
## Median : 0.000 Median : 40.20 Median : 654.1 Median : 180.7   
## Mean : -107.092 Mean : 2356.02 Mean : 9101.4 Mean : 2852.3   
## 3rd Qu.: 0.200 3rd Qu.: 767.05 3rd Qu.: 4825.3 3rd Qu.: 1501.3   
## Max. : 5631.000 Max. :80197.00 Max. :474259.0 Max. :76343.0   
##   
## seqo siv spce spi   
## Min. :-30165.16 Min. : 0.00 Min. : -688.784 Min. :-2628.00   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: -0.186 1st Qu.: -13.92   
## Median : 0.00 Median : 0.00 Median : 16.625 Median : 0.00   
## Mean : 42.27 Mean : 65.66 Mean : 538.722 Mean : -39.93   
## 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 239.164 3rd Qu.: 0.00   
## Max. : 11023.33 Max. :12791.00 Max. :16999.000 Max. : 6523.00   
##   
## sppe sppiv sstk stkco   
## Min. : 0.000 Min. :-26151.137 Min. : -1.831 Min. : -2.446   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 0.000 Median : 0.000 Median : 0.115 Median : 0.980   
## Mean : 12.176 Mean : -37.163 Mean : 41.794 Mean : 19.963   
## 3rd Qu.: 1.254 3rd Qu.: 0.001 3rd Qu.: 7.560 3rd Qu.: 12.000   
## Max. :1002.000 Max. : 1409.479 Max. :1750.000 Max. :453.000   
##   
## teq tstk tstkn txc   
## Min. :-6274.00 Min. : -1.45 Min. : 0.000 Min. :-247.200   
## 1st Qu.: 10.37 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 185.46 Median : 0.00 Median : 0.000 Median : 0.068   
## Mean : 2948.62 Mean : 1031.13 Mean : 31.347 Mean : 152.731   
## 3rd Qu.: 1538.49 3rd Qu.: 14.53 3rd Qu.: 1.933 3rd Qu.: 30.277   
## Max. :81738.00 Max. :71966.00 Max. :2638.000 Max. :8619.000   
##   
## txdba txdbcl txdc txfed   
## Min. : 0.00 Min. : 0.000 Min. :-929.000 Min. :-489.00   
## 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: -0.067 1st Qu.: 0.00   
## Median : 0.00 Median : 0.000 Median : 0.000 Median : 0.00   
## Mean : 63.63 Mean : 5.507 Mean : 9.010 Mean : 85.17   
## 3rd Qu.: 9.20 3rd Qu.: 0.000 3rd Qu.: 3.515 3rd Qu.: 14.12   
## Max. :3170.95 Max. :469.000 Max. :1050.000 Max. :6377.00   
##   
## txfo txndba txndbl txp   
## Min. :-688.66 Min. : 0.00 Min. : 0.00 Min. : -0.252   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.00 Median : 10.11 Median : 13.24 Median : 0.000   
## Mean : 62.96 Mean : 251.72 Mean : 489.69 Mean : 50.853   
## 3rd Qu.: 4.00 3rd Qu.: 141.46 3rd Qu.: 249.48 3rd Qu.: 6.420   
## Max. :3855.00 Max. :6450.00 Max. :15376.00 Max. :2211.000   
##   
## txpd txr txs txt   
## Min. :-115.974 Min. : 0.00 Min. :-58.000 Min. :-456.811   
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 3.335 Median : 0.00 Median : 0.000 Median : 6.439   
## Mean : 198.654 Mean : 13.58 Mean : 12.208 Mean : 223.756   
## 3rd Qu.: 56.992 3rd Qu.: 0.00 3rd Qu.: 2.336 3rd Qu.: 86.628   
## Max. :8641.000 Max. :1292.68 Max. :743.000 Max. :8105.000   
##   
## wcap xacc xint xrent   
## Min. :-11878.000 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.257 1st Qu.: 0.293 1st Qu.: 0.174 1st Qu.: 0.080   
## Median : 39.370 Median : 12.423 Median : 4.871 Median : 3.753   
## Mean : 262.231 Mean : 433.890 Mean : 111.833 Mean : 83.986   
## 3rd Qu.: 347.414 3rd Qu.: 168.000 3rd Qu.: 88.811 3rd Qu.: 45.200   
## Max. : 14286.000 Max. :18202.000 Max. :3341.000 Max. :2800.000   
##   
## xsga cshtr\_c dvpsp\_c dvpsx\_c   
## Min. : 0.00 Min. :0.000e+00 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.: 7.42 1st Qu.:1.871e+06 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 106.32 Median :1.698e+07 Median : 0.0000 Median : 0.0000   
## Mean : 1845.66 Mean :2.287e+08 Mean : 0.4928 Mean : 0.4959   
## 3rd Qu.: 868.97 3rd Qu.:1.430e+08 3rd Qu.: 0.6397 3rd Qu.: 0.6400   
## Max. :90920.00 Max. :5.728e+09 Max. :21.0000 Max. :21.0000   
##   
## prcc\_c prch\_c prcl\_c cshtr\_f   
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. :0.000e+00   
## 1st Qu.: 0.70 1st Qu.: 1.80 1st Qu.: 0.33 1st Qu.:1.778e+06   
## Median : 9.64 Median : 14.25 Median : 6.59 Median :1.672e+07   
## Mean : 29.67 Mean : 34.17 Mean : 22.30 Mean :2.295e+08   
## 3rd Qu.: 34.75 3rd Qu.: 39.71 3rd Qu.: 26.21 3rd Qu.:1.398e+08   
## Max. :2794.97 Max. :2948.24 Max. :2500.00 Max. :6.052e+09   
##   
## dvpsp\_f dvpsx\_f mkvalt prcc\_f   
## Min. : 0.0000 Min. : 0.0000 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 2.30 1st Qu.: 0.79   
## Median : 0.0000 Median : 0.0000 Median : 71.92 Median : 9.86   
## Mean : 0.4891 Mean : 0.4921 Mean : 5786.58 Mean : 29.48   
## 3rd Qu.: 0.6218 3rd Qu.: 0.6397 3rd Qu.: 1404.11 3rd Qu.: 34.49   
## Max. :21.0000 Max. :21.0000 Max. :241440.44 Max. :2794.97   
##   
## prch\_f prcl\_f ein incorp   
## Min. : 0.00 Min. : 0.00 13-4306188: 5 DE :487   
## 1st Qu.: 1.90 1st Qu.: 0.35 16-0716709: 5 NV :166   
## Median : 14.50 Median : 6.80 16-0733425: 5 FL : 43   
## Mean : 34.07 Mean : 22.18 20-1266625: 5 VA : 35   
## 3rd Qu.: 39.59 3rd Qu.: 26.10 23-1614034: 5 CO : 32   
## Max. :2948.24 Max. :2500.00 (Other) :999 (Other):240   
## NA's :181 NA's :202   
## state restmt\_at restmt\_at\_mag restmt\_capx   
## CA :140 Min. :0.00000 Min. :-93.6120 Min. :0.00000   
## NY : 99 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.00000   
## FL : 72 Median :0.00000 Median : 0.0000 Median :0.00000   
## IL : 65 Mean :0.07801 Mean : 0.5079 Mean :0.03983   
## NJ : 62 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.00000   
## (Other):511 Max. :1.00000 Max. :729.9070 Max. :1.00000   
## NA's :256   
## restmt\_capx\_mag restmt\_cogs restmt\_cogs\_mag restmt\_dltt   
## Min. : -100.00 Min. :0.0000 Min. : -100.00 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00 Median :0.0000 Median : 0.00 Median :0.00000   
## Mean : 22.87 Mean :0.2788 Mean : 32.33 Mean :0.03154   
## 3rd Qu.: 0.00 3rd Qu.:1.0000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :28133.57 Max. :1.0000 Max. :31225.00 Max. :1.00000   
##   
## restmt\_dltt\_mag restmt\_epsfi restmt\_epsfi\_mag restmt\_epspi   
## Min. :-100.0000 Min. :0.000 Min. : -165.8 Min. :0.000   
## 1st Qu.: 0.0000 1st Qu.:0.000 1st Qu.: 0.0 1st Qu.:0.000   
## Median : 0.0000 Median :0.000 Median : 0.0 Median :0.000   
## Mean : 0.2049 Mean :0.112 Mean : 396.6 Mean :0.112   
## 3rd Qu.: 0.0000 3rd Qu.:0.000 3rd Qu.: 0.0 3rd Qu.:0.000   
## Max. : 399.2380 Max. :1.000 Max. :198666.7 Max. :1.000   
##   
## restmt\_epspi\_mag restmt\_ib restmt\_ib\_mag restmt\_ni   
## Min. : -163.3 Min. :0.0000 Min. :-313.738 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.0000 Median : 0.000 Median :0.00000   
## Mean : 399.2 Mean :0.1029 Mean : 5.186 Mean :0.04481   
## 3rd Qu.: 0.0 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :198666.7 Max. :1.0000 Max. :8051.351 Max. :1.00000   
##   
## restmt\_ni\_mag restmt\_nopi restmt\_nopi\_mag restmt\_pi   
## Min. :-168.75 Min. :0.0000 Min. :-1868600.0 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.: -51.6 1st Qu.:0.00000   
## Median : 0.00 Median :1.0000 Median : 0.0 Median :0.00000   
## Mean : 7.04 Mean :0.5718 Mean : -1923.0 Mean :0.09876   
## 3rd Qu.: 0.00 3rd Qu.:1.0000 3rd Qu.: 0.0 3rd Qu.:0.00000   
## Max. :8051.35 Max. :1.0000 Max. : 274013.7 Max. :1.00000   
##   
## restmt\_pi\_mag restmt\_reuna restmt\_reuna\_mag restmt\_seq   
## Min. :-8243.033 Min. :0.00000 Min. :-9841.348 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.000 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : -1.457 Mean :0.06639 Mean : -2.913 Mean :0.07884   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. : 8051.351 Max. :1.00000 Max. :12545.000 Max. :1.00000   
##   
## restmt\_seq\_mag restmt\_teq restmt\_teq\_mag restmt\_txt   
## Min. : -388.04 Min. :0.00000 Min. : -375.93 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00 Median :0.00000 Median : 0.00 Median :0.00000   
## Mean : 46.03 Mean :0.07469 Mean : 46.61 Mean :0.06473   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :37620.00 Max. :1.00000 Max. :37620.00 Max. :1.00000   
##   
## restmt\_txt\_mag restmt\_wcap restmt\_wcap\_mag restmt\_xint   
## Min. :-8437.071 Min. :0.00000 Min. :-130.3750 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median : 0.0000 Median :0.0000   
## Mean : -8.021 Mean :0.07137 Mean : 0.5482 Mean :0.1228   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.0000   
## Max. : 361.538 Max. :1.00000 Max. : 825.0000 Max. :1.0000   
##   
## restmt\_xint\_mag restmt\_xsga restmt\_xsga\_mag restmt\_dvpsp\_f  
## Min. :-100.000 Min. :0.0000 Min. :-100.000 Min. :0   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0   
## Median : 0.000 Median :0.0000 Median : 0.000 Median :0   
## Mean : 1.988 Mean :0.1461 Mean : 4.145 Mean :0   
## 3rd Qu.: 0.000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0   
## Max. :3814.830 Max. :1.0000 Max. :5651.351 Max. :0   
##   
## restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f restmt\_dvpsx\_f\_mag  
## Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0   
##

colnames(fundamentals\_ds\_filter\_2)

## [1] "gvkey" "fyear" "tic"   
## [4] "aco" "acominc" "act"   
## [7] "am" "ao" "aocidergl"   
## [10] "aociother" "aocipen" "aodo"   
## [13] "aoloch" "ap" "aqc"   
## [16] "at" "bkvlps" "caps"   
## [19] "capx" "ceq" "ceqt"   
## [22] "ch" "che" "chech"   
## [25] "ci" "cogs" "cshi"   
## [28] "csho" "cshr" "cstk"   
## [31] "cstkcv" "dc" "dd"   
## [34] "dd1" "dd2" "dd3"   
## [37] "dd4" "dd5" "dilavx"   
## [40] "dlc" "dltp" "dltt"   
## [43] "dm" "dn" "dpact"   
## [46] "dpc" "dvt" "ebit"   
## [49] "ebitda" "emp" "epsfi"   
## [52] "epspi" "esub" "esubc"   
## [55] "fatb" "fatc" "fatp"   
## [58] "fiao" "fincf" "fopo"   
## [61] "gdwl" "gp" "ib"   
## [64] "icapt" "intan" "intano"   
## [67] "intc" "intpn" "invt"   
## [70] "ivaeq" "ivch" "ivncf"   
## [73] "ivst" "ivstch" "lifr"   
## [76] "lo" "lse" "lt"   
## [79] "mrc1" "mrcta" "ni"   
## [82] "nopi" "nopio" "oancf"   
## [85] "oiadp" "oibdp" "opeps"   
## [88] "pi" "pncad" "pncaeps"   
## [91] "ppegt" "prca" "prstkc"   
## [94] "re" "reajo" "recch"   
## [97] "recd" "rect" "recta"   
## [100] "reuna" "revt" "seq"   
## [103] "seqo" "siv" "spce"   
## [106] "spi" "sppe" "sppiv"   
## [109] "sstk" "stkco" "teq"   
## [112] "tstk" "tstkn" "txc"   
## [115] "txdba" "txdbcl" "txdc"   
## [118] "txfed" "txfo" "txndba"   
## [121] "txndbl" "txp" "txpd"   
## [124] "txr" "txs" "txt"   
## [127] "wcap" "xacc" "xint"   
## [130] "xrent" "xsga" "cshtr\_c"   
## [133] "dvpsp\_c" "dvpsx\_c" "prcc\_c"   
## [136] "prch\_c" "prcl\_c" "cshtr\_f"   
## [139] "dvpsp\_f" "dvpsx\_f" "mkvalt"   
## [142] "prcc\_f" "prch\_f" "prcl\_f"   
## [145] "ein" "incorp" "state"   
## [148] "restmt\_at" "restmt\_at\_mag" "restmt\_capx"   
## [151] "restmt\_capx\_mag" "restmt\_cogs" "restmt\_cogs\_mag"   
## [154] "restmt\_dltt" "restmt\_dltt\_mag" "restmt\_epsfi"   
## [157] "restmt\_epsfi\_mag" "restmt\_epspi" "restmt\_epspi\_mag"   
## [160] "restmt\_ib" "restmt\_ib\_mag" "restmt\_ni"   
## [163] "restmt\_ni\_mag" "restmt\_nopi" "restmt\_nopi\_mag"   
## [166] "restmt\_pi" "restmt\_pi\_mag" "restmt\_reuna"   
## [169] "restmt\_reuna\_mag" "restmt\_seq" "restmt\_seq\_mag"   
## [172] "restmt\_teq" "restmt\_teq\_mag" "restmt\_txt"   
## [175] "restmt\_txt\_mag" "restmt\_wcap" "restmt\_wcap\_mag"   
## [178] "restmt\_xint" "restmt\_xint\_mag" "restmt\_xsga"   
## [181] "restmt\_xsga\_mag" "restmt\_dvpsp\_f" "restmt\_dvpsp\_f\_mag"  
## [184] "restmt\_dvpsx\_f" "restmt\_dvpsx\_f\_mag"

nrow(fundamentals\_ds\_filter\_2)

## [1] 1205

### The dataset contains company wise data across financial year, this dataset needs to be consolidate to single ror for each of the company

### So group the dataset by gvkey and summarize all variables

final\_ds\_initial <- fundamentals\_ds\_filter\_2 %>%  
 group\_by(gvkey,tic) %>%  
 dplyr::summarize(  
 aco = mean(aco),  
 acominc = mean(acominc),  
 act = mean(act),  
 ao = mean(ao),  
 aocidergl = mean(aocidergl),  
 aocipen = mean(aocipen),  
 aodo = mean(aodo),  
 aoloch = mean(aoloch),  
 ap = mean(ap),  
 aqc = mean(aqc),  
 at = mean(at),  
 bkvlps = mean(bkvlps),  
 caps = mean(caps),  
 capx = mean(capx),  
 ceq = mean(ceq),  
 ceqt = mean(ceqt),  
 ch = mean(ch),  
 che = mean(che),  
 chech = mean(chech),  
 ci = mean(ci),  
 cogs = mean(cogs),  
 cshi = mean(cshi),  
 csho = mean(csho),  
 cstk = mean(cstk),  
 cstkcv = mean(cstkcv),  
 dd1 = mean(dd1),  
 dilavx = mean(dilavx),  
 dlc = mean(dlc),  
 dltt = mean(dltt),  
 dm = mean(dm),  
 dn = mean(dn),  
 dpact = mean(dpact),  
 dpc = mean(dpc),  
 dvt = mean(dvt),  
 ebit = mean(ebit),  
 ebitda = mean(ebitda),  
 epsfi = mean(epsfi),  
 epspi = mean(epspi),  
 fiao = mean(fiao),  
 fincf = mean(fincf),  
 fopo = mean(fopo),  
 gdwl = mean(gdwl),  
 gp = mean(gp),  
 ib = mean(ib),  
 icapt = mean(icapt),  
 intan = mean(intan),  
 intano = mean(intano),  
 invt = mean(invt),  
 ivch = mean(ivch),  
 ivncf = mean(ivncf),  
 ivst = mean(ivst),  
 lo = mean(lo),  
 lse = mean(lse),  
 lt = mean(lt),  
 ni = mean(ni),  
 nopi = mean(nopi),  
 nopio = mean(nopio),  
 oancf = mean(oancf),  
 oiadp = mean(oiadp),  
 oibdp = mean(oibdp),  
 opeps = mean(opeps),  
 pi = mean(pi),  
 ppegt = mean(ppegt),  
 re = mean(re),  
 reajo = mean(reajo),  
 rect = mean(rect),  
 recta = mean(recta),  
 reuna = mean(reuna),  
 revt = mean(revt),  
 seq = mean( seq ),  
 siv = mean( siv ),  
 spce = mean(spce),  
 spi = mean(spi),  
 sppiv = mean(sppiv),  
 sstk = mean(sstk),  
 teq = mean(teq),  
 tstk = mean(tstk),  
 tstkn = mean(tstkn),  
 txp = mean(txp),  
 txr = mean(txr),  
 txt = mean(txt),  
 wcap = mean(wcap),  
 xint = mean(xint),  
 restmt\_at = mean(restmt\_at),  
 restmt\_at\_mag = mean(restmt\_at\_mag),  
 restmt\_capx = mean(restmt\_capx),  
 restmt\_capx\_mag = mean(restmt\_capx\_mag),  
 restmt\_cogs = mean(restmt\_cogs),  
 restmt\_cogs\_mag = mean(restmt\_cogs\_mag),  
 restmt\_dltt = mean(restmt\_dltt),  
 restmt\_dltt\_mag = mean(restmt\_dltt\_mag),  
 restmt\_epsfi = mean(restmt\_epsfi),  
 restmt\_epsfi\_mag = mean(restmt\_epsfi\_mag),  
 restmt\_epspi = mean(restmt\_epspi),  
 restmt\_epspi\_mag = mean(restmt\_epspi\_mag),  
 restmt\_ib = mean(restmt\_ib),  
 restmt\_ib\_mag = mean(restmt\_ib\_mag),  
 restmt\_ni = mean(restmt\_ni),  
 restmt\_ni\_mag = mean(restmt\_ni\_mag),  
 restmt\_nopi = mean(restmt\_nopi),  
 restmt\_nopi\_mag = mean(restmt\_nopi\_mag),  
 restmt\_pi = mean(restmt\_pi),  
 restmt\_pi\_mag = mean(restmt\_pi\_mag),  
 restmt\_reuna = mean(restmt\_reuna),  
 restmt\_reuna\_mag = mean(restmt\_reuna\_mag),  
 restmt\_seq = mean(restmt\_seq),  
 restmt\_seq\_mag = mean(restmt\_seq\_mag),  
 restmt\_teq = mean(restmt\_teq),  
 restmt\_teq\_mag = mean(restmt\_teq\_mag),  
 restmt\_txt = mean(restmt\_txt),  
 restmt\_txt\_mag = mean(restmt\_txt\_mag),  
 restmt\_wcap = mean(restmt\_wcap),  
 restmt\_wcap\_mag = mean(restmt\_wcap\_mag),  
   
 restmt\_xint = mean(restmt\_xint),  
 restmt\_xint\_mag = mean(restmt\_xint\_mag),  
   
 restmt\_xsga = mean(restmt\_xsga),  
 restmt\_xsga\_mag = mean(restmt\_xsga\_mag),  
   
 restmt\_dvpsp\_f = mean(restmt\_dvpsp\_f),  
 restmt\_dvpsp\_f\_mag = mean(restmt\_dvpsp\_f\_mag),  
   
 restmt\_dvpsx\_f = mean(restmt\_dvpsx\_f),  
 restmt\_dvpsx\_f\_mag = mean(restmt\_dvpsx\_f\_mag)  
   
 )

## `summarise()` regrouping output by 'gvkey' (override with `.groups` argument)

summary(final\_ds\_initial)

## gvkey tic aco acominc   
## Min. : 1239 0161A : 1 Min. : 0.000 Min. :-19306.57   
## 1st Qu.: 12107 0170A : 1 1st Qu.: 0.447 1st Qu.: -30.39   
## Median : 61311 0171A : 1 Median : 8.858 Median : 0.00   
## Mean : 83018 0173A : 1 Mean : 188.577 Mean : -194.14   
## 3rd Qu.:165694 0270B : 1 3rd Qu.: 94.290 3rd Qu.: 0.00   
## Max. :277487 0563B : 1 Max. :4760.750 Max. : 3495.34   
## (Other):342   
## act ao aocidergl aocipen   
## Min. : 0.00 Min. : 0.000 Min. :-2207.250 Min. :-2803.25   
## 1st Qu.: 10.27 1st Qu.: 0.145 1st Qu.: 0.000 1st Qu.: -10.48   
## Median : 115.71 Median : 8.322 Median : 0.000 Median : 0.00   
## Mean : 1918.97 Mean : 200.854 Mean : -9.098 Mean : -91.21   
## 3rd Qu.: 1225.80 3rd Qu.: 93.865 3rd Qu.: 0.000 3rd Qu.: 0.00   
## Max. :55264.80 Max. :5330.250 Max. : 119.000 Max. : 30.75   
##   
## aodo aoloch ap aqc   
## Min. : 0.000 Min. :-667.500 Min. : 0.00 Min. : -12.45   
## 1st Qu.: 0.070 1st Qu.: -1.851 1st Qu.: 1.17 1st Qu.: 0.00   
## Median : 6.675 Median : 0.000 Median : 18.13 Median : 0.00   
## Mean : 188.752 Mean : 6.660 Mean : 660.17 Mean : 105.33   
## 3rd Qu.: 91.195 3rd Qu.: 1.530 3rd Qu.: 241.09 3rd Qu.: 14.95   
## Max. :5330.250 Max. : 744.000 Max. :35222.20 Max. :5559.02   
##   
## at bkvlps caps capx   
## Min. : 0.00 Min. :-130515.0 Min. : -701.48 Min. : 0.000   
## 1st Qu.: 20.64 1st Qu.: 0.1 1st Qu.: 5.25 1st Qu.: 0.343   
## Median : 283.75 Median : 3.8 Median : 37.25 Median : 12.573   
## Mean : 6489.03 Mean : 11681.8 Mean : 951.28 Mean : 242.490   
## 3rd Qu.: 3172.25 3rd Qu.: 12.5 3rd Qu.: 363.48 3rd Qu.: 119.642   
## Max. :190526.20 Max. :1881687.0 Max. :62705.25 Max. :12881.200   
##   
## ceq ceqt ch che   
## Min. :-2342.49 Min. :-40530.25 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 3.36 1st Qu.: -3.63 1st Qu.: 1.399 1st Qu.: 1.496   
## Median : 105.08 Median : 13.54 Median : 21.299 Median : 26.194   
## Mean : 2388.45 Mean : 54.96 Mean : 389.436 Mean : 487.347   
## 3rd Qu.: 1033.62 3rd Qu.: 189.37 3rd Qu.: 179.819 3rd Qu.: 217.474   
## Max. :72640.80 Max. : 53931.40 Max. :10044.000 Max. :15547.750   
##   
## chech ci cogs cshi   
## Min. :-305.7500 Min. : -722.617 Min. : 0.0 Min. : 0.0   
## 1st Qu.: -0.1368 1st Qu.: -1.477 1st Qu.: 12.1 1st Qu.: 16.2   
## Median : 0.5806 Median : 9.139 Median : 216.7 Median : 49.8   
## Mean : 42.2654 Mean : 475.730 Mean : 5116.2 Mean : 267.3   
## 3rd Qu.: 10.6559 3rd Qu.: 130.370 3rd Qu.: 2471.9 3rd Qu.: 145.1   
## Max. :1543.0000 Max. :16365.200 Max. :325065.8 Max. :6253.5   
##   
## csho cstk cstkcv dd1   
## Min. : 0.00 Min. : 0.000 Min. : 0.0000 Min. : 0.000   
## 1st Qu.: 16.32 1st Qu.: 0.026 1st Qu.: 0.0010 1st Qu.: 0.000   
## Median : 49.41 Median : 0.248 Median : 0.0100 Median : 1.451   
## Mean : 240.67 Mean : 191.237 Mean : 0.6068 Mean : 169.076   
## 3rd Qu.: 142.42 3rd Qu.: 21.315 3rd Qu.: 0.2500 3rd Qu.: 39.533   
## Max. :6252.56 Max. :7290.750 Max. :20.8642 Max. :5428.500   
##   
## dilavx dlc dltt dm   
## Min. : -738.263 Min. : 0.000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: -1.320 1st Qu.: 0.302 1st Qu.: 0.16 1st Qu.: 0.000   
## Median : 4.824 Median : 5.257 Median : 17.03 Median : 1.218   
## Mean : 460.750 Mean : 374.668 Mean : 1477.91 Mean : 157.728   
## 3rd Qu.: 125.014 3rd Qu.: 99.945 3rd Qu.: 902.29 3rd Qu.: 62.483   
## Max. :15690.400 Max. :15926.126 Max. :42659.60 Max. :3900.400   
##   
## dn dpact dpc dvt   
## Min. : 0.0 Min. : 0.00 Min. : 0.000 Min. : -0.006   
## 1st Qu.: 0.0 1st Qu.: 2.02 1st Qu.: 0.319 1st Qu.: 0.000   
## Median : 0.0 Median : 52.40 Median : 7.505 Median : 0.000   
## Mean : 902.6 Mean : 1303.11 Mean : 170.666 Mean : 226.759   
## 3rd Qu.: 252.3 3rd Qu.: 768.89 3rd Qu.: 95.563 3rd Qu.: 35.368   
## Max. :42561.8 Max. :50449.80 Max. :8059.800 Max. :6572.535   
##   
## ebit ebitda epsfi epspi   
## Min. : -208.760 Min. : -33.68 Min. :-14.0200 Min. :-14.0200   
## 1st Qu.: -0.369 1st Qu.: 0.02 1st Qu.: -0.0512 1st Qu.: -0.0512   
## Median : 23.871 Median : 31.53 Median : 0.2288 Median : 0.2362   
## Mean : 761.483 Mean : 940.88 Mean : 3.9554 Mean : 3.9707   
## 3rd Qu.: 345.869 3rd Qu.: 441.54 3rd Qu.: 1.8338 3rd Qu.: 1.8638   
## Max. :24345.400 Max. :32405.20 Max. :881.6400 Max. :881.6400   
##   
## fiao fincf fopo   
## Min. :-3427.000 Min. :-11533.200 Min. :-389.5000   
## 1st Qu.: -7.661 1st Qu.: -50.075 1st Qu.: 0.1661   
## Median : -0.047 Median : -0.005 Median : 2.0777   
## Mean : -45.439 Mean : -332.514 Mean : 74.1085   
## 3rd Qu.: 0.000 3rd Qu.: 4.030 3rd Qu.: 23.4436   
## Max. : 1800.250 Max. : 824.184 Max. :1979.4552   
##   
## gdwl gp ib icapt   
## Min. : 0.00 Min. : -3.19 Min. : -727.025 Min. : -23.14   
## 1st Qu.: 0.00 1st Qu.: 6.98 1st Qu.: -1.320 1st Qu.: 10.74   
## Median : 1.95 Median : 102.86 Median : 6.421 Median : 192.63   
## Mean : 1492.38 Mean : 2502.66 Mean : 467.872 Mean : 4010.77   
## 3rd Qu.: 390.49 3rd Qu.: 1238.36 3rd Qu.: 136.674 3rd Qu.: 2098.22   
## Max. :56373.25 Max. :117445.60 Max. :15690.400 Max. :119888.20   
##   
## intan intano invt ivch   
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.17 1st Qu.: 0.04 1st Qu.: 2.19 1st Qu.: 0.00   
## Median : 18.61 Median : 7.21 Median : 37.47 Median : 0.00   
## Mean : 2370.53 Mean : 878.15 Mean : 745.30 Mean : 60.27   
## 3rd Qu.: 715.80 3rd Qu.: 217.88 3rd Qu.: 464.24 3rd Qu.: 0.27   
## Max. :86837.75 Max. :31704.00 Max. :39770.60 Max. :4366.68   
##   
## ivncf ivst lo lse   
## Min. :-13066.20 Min. : 0.000 Min. : -128.941 Min. : 0.00   
## 1st Qu.: -176.92 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 20.64   
## Median : -19.36 Median : 0.000 Median : 5.334 Median : 283.75   
## Mean : -314.10 Mean : 88.668 Mean : 475.282 Mean : 6489.03   
## 3rd Qu.: -0.24 3rd Qu.: 2.429 3rd Qu.: 203.087 3rd Qu.: 3172.25   
## Max. : 985.75 Max. :5503.750 Max. :14517.069 Max. :190526.20   
##   
## lt ni nopi   
## Min. : 0.02 Min. : -737.537 Min. :-230.2500   
## 1st Qu.: 8.05 1st Qu.: -1.661 1st Qu.: 0.0000   
## Median : 108.02 Median : 6.217 Median : 0.1979   
## Mean : 3948.06 Mean : 499.241 Mean : 47.9687   
## 3rd Qu.: 2056.19 3rd Qu.: 125.392 3rd Qu.: 4.8163   
## Max. :113297.60 Max. :17374.318 Max. :2224.4000   
##   
## nopio oancf oiadp oibdp   
## Min. :-230.2500 Min. : -61.444 Min. : -208.760 Min. : -33.68   
## 1st Qu.: -0.0111 1st Qu.: -0.142 1st Qu.: -0.369 1st Qu.: 0.02   
## Median : 0.0664 Median : 20.797 Median : 23.871 Median : 31.53   
## Mean : 37.9242 Mean : 696.038 Mean : 761.483 Mean : 940.88   
## 3rd Qu.: 2.5212 3rd Qu.: 282.996 3rd Qu.: 345.869 3rd Qu.: 441.54   
## Max. :2054.4000 Max. :24599.000 Max. :24345.400 Max. :32405.20   
##   
## opeps pi ppegt re   
## Min. : -9.8200 Min. : -739.921 Min. : 0.00 Min. :-7570.29   
## 1st Qu.: -0.0350 1st Qu.: -1.287 1st Qu.: 5.51 1st Qu.: -9.35   
## Median : 0.2971 Median : 11.654 Median : 146.66 Median : 19.16   
## Mean : 3.9977 Mean : 678.381 Mean : 2919.87 Mean : 1909.92   
## 3rd Qu.: 1.8725 3rd Qu.: 215.671 3rd Qu.: 1576.40 3rd Qu.: 441.34   
## Max. :856.8325 Max. :24079.000 Max. :161869.20 Max. :68884.60   
##   
## reajo rect recta reuna   
## Min. :-7860.75 Min. : 0.000 Min. :-19466.259 Min. :-7527.73   
## 1st Qu.: -19.79 1st Qu.: 1.636 1st Qu.: -0.169 1st Qu.: -9.13   
## Median : 0.00 Median : 28.478 Median : 0.000 Median : 21.03   
## Mean : -78.00 Mean : 510.810 Mean : -96.126 Mean : 2011.34   
## 3rd Qu.: 0.00 3rd Qu.: 312.676 3rd Qu.: 0.060 3rd Qu.: 435.07   
## Max. : 7171.53 Max. :15020.067 Max. : 1946.250 Max. :72710.50   
##   
## revt seq siv spce   
## Min. : 0.0 Min. :-2208.96 Min. : 0.000 Min. : -600.364   
## 1st Qu.: 22.4 1st Qu.: 4.15 1st Qu.: 0.000 1st Qu.: -1.121   
## Median : 333.1 Median : 106.70 Median : 0.000 Median : 6.643   
## Mean : 7618.9 Mean : 2442.71 Mean : 56.788 Mean : 459.367   
## 3rd Qu.: 3826.2 3rd Qu.: 1091.19 3rd Qu.: 0.512 3rd Qu.: 138.243   
## Max. :442511.4 Max. :72640.80 Max. :4366.827 Max. :15690.400   
##   
## spi sppiv sstk teq   
## Min. :-921.2962 Min. :-6191.874 Min. : 0.0000 Min. :-2208.96   
## 1st Qu.: -15.8642 1st Qu.: -0.046 1st Qu.: 0.0006 1st Qu.: 4.19   
## Median : -0.4417 Median : 0.000 Median : 1.2209 Median : 106.70   
## Mean : -35.2648 Mean : -32.516 Mean : 36.8987 Mean : 2524.10   
## 3rd Qu.: 0.0000 3rd Qu.: 0.007 3rd Qu.: 14.5968 3rd Qu.: 1095.88   
## Max. :1115.5000 Max. : 27.017 Max. :1513.0000 Max. :76602.80   
##   
## tstk tstkn txp txr   
## Min. : 0.0 Min. : 0.000 Min. : -0.252 Min. : 0.0000   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.0000   
## Median : 0.0 Median : 0.000 Median : 0.049 Median : 0.0000   
## Mean : 889.8 Mean : 27.024 Mean : 44.423 Mean : 11.9381   
## 3rd Qu.: 8.5 3rd Qu.: 1.282 3rd Qu.: 5.603 3rd Qu.: 0.0758   
## Max. :67539.2 Max. :1923.500 Max. :1469.476 Max. :1150.7513   
##   
## txt wcap xint restmt\_at   
## Min. : -76.388 Min. :-8236.800 Min. : 0.0000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.: -0.011 1st Qu.: 0.1635 1st Qu.:0.0000   
## Median : 3.772 Median : 25.438 Median : 2.2978 Median :0.0000   
## Mean : 189.846 Mean : 228.503 Mean : 95.8065 Mean :0.0694   
## 3rd Qu.: 63.206 3rd Qu.: 279.322 3rd Qu.: 66.5683 3rd Qu.:0.0000   
## Max. :7749.600 Max. :12261.750 Max. :2859.7500 Max. :0.8000   
##   
## restmt\_at\_mag restmt\_capx restmt\_capx\_mag restmt\_cogs   
## Min. :-23.4030 Min. :0.00000 Min. : -28.79 Min. :0.000   
## 1st Qu.: 0.0000 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.:0.000   
## Median : 0.0000 Median :0.00000 Median : 0.00 Median :0.000   
## Mean : 0.4492 Mean :0.03654 Mean : 19.82 Mean :0.249   
## 3rd Qu.: 0.0000 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.:0.500   
## Max. :182.4888 Max. :0.75000 Max. :7033.39 Max. :1.000   
##   
## restmt\_cogs\_mag restmt\_dltt restmt\_dltt\_mag restmt\_epsfi   
## Min. : -50.000 Min. :0.00000 Min. :-26.9567 Min. :0.0000   
## 1st Qu.: -0.001 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median : 0.0000 Median :0.0000   
## Mean : 29.001 Mean :0.02998 Mean : 0.1961 Mean :0.1051   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.0000   
## Max. :9299.359 Max. :1.00000 Max. :100.9780 Max. :1.0000   
##   
## restmt\_epsfi\_mag restmt\_epspi restmt\_epspi\_mag restmt\_ib   
## Min. : -50.05 Min. :0.0000 Min. : -50.0 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.: 0.0 1st Qu.:0.00000   
## Median : 0.00 Median :0.0000 Median : 0.0 Median :0.00000   
## Mean : 345.40 Mean :0.1056 Mean : 347.6 Mean :0.09334   
## 3rd Qu.: 0.00 3rd Qu.:0.0000 3rd Qu.: 0.0 3rd Qu.:0.00000   
## Max. :77081.67 Max. :1.0000 Max. :77081.7 Max. :0.80000   
##   
## restmt\_ib\_mag restmt\_ni restmt\_ni\_mag restmt\_nopi   
## Min. :-121.766 Min. :0.00000 Min. : -42.188 Min. :0.000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.000   
## Median : 0.000 Median :0.00000 Median : 0.000 Median :0.600   
## Mean : 6.455 Mean :0.04128 Mean : 8.054 Mean :0.553   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:1.000   
## Max. :2683.890 Max. :0.80000 Max. :2683.890 Max. :1.000   
##   
## restmt\_nopi\_mag restmt\_pi restmt\_pi\_mag restmt\_reuna   
## Min. :-1868600.0 Min. :0.00000 Min. :-2747.678 Min. :0.00000   
## 1st Qu.: -118.3 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : -5830.0 Mean :0.09004 Mean : -1.358 Mean :0.06303   
## 3rd Qu.: 48.9 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. : 68865.1 Max. :0.80000 Max. : 2683.890 Max. :1.00000   
##   
## restmt\_reuna\_mag restmt\_seq restmt\_seq\_mag restmt\_teq   
## Min. :-2461.679 Min. :0.00000 Min. : -105.4 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.0 1st Qu.:0.00000   
## Median : 0.000 Median :0.00000 Median : 0.0 Median :0.00000   
## Mean : -0.805 Mean :0.08501 Mean : 48.8 Mean :0.07926   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.0 3rd Qu.:0.00000   
## Max. : 4181.704 Max. :1.00000 Max. :12541.8 Max. :1.00000   
##   
## restmt\_teq\_mag restmt\_txt restmt\_txt\_mag restmt\_wcap   
## Min. : -105.4 Min. :0.00000 Min. :-2109.268 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : 49.3 Mean :0.05627 Mean : -6.922 Mean :0.07002   
## 3rd Qu.: 0.0 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :12541.8 Max. :0.80000 Max. : 47.318 Max. :1.00000   
##   
## restmt\_wcap\_mag restmt\_xint restmt\_xint\_mag restmt\_xsga   
## Min. :-43.249 Min. :0.0000 Min. :-62.735 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.000 Median :0.0000 Median : 0.000 Median :0.0000   
## Mean : 1.006 Mean :0.1069 Mean : 1.676 Mean :0.1318   
## 3rd Qu.: 0.000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.2500   
## Max. :412.500 Max. :1.0000 Max. :953.707 Max. :1.0000   
##   
## restmt\_xsga\_mag restmt\_dvpsp\_f restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f  
## Min. : -50.000 Min. :0 Min. :0 Min. :0   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median : 0.000 Median :0 Median :0 Median :0   
## Mean : 4.863 Mean :0 Mean :0 Mean :0   
## 3rd Qu.: 0.000 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :1884.021 Max. :0 Max. :0 Max. :0   
##   
## restmt\_dvpsx\_f\_mag  
## Min. :0   
## 1st Qu.:0   
## Median :0   
## Mean :0   
## 3rd Qu.:0   
## Max. :0   
##

nrow(final\_ds\_initial)

## [1] 348

### Adjust the restement values, in case if the restate percentage is greater than 50% then mark restatement as 1 or else mark as 0 For 0/false restatement magnitude is marked as 0.

final\_ds\_initial\_1 <- final\_ds\_initial   
  
for (row in 1:nrow(final\_ds\_initial\_1)){  
 row\_item\_gvkey <- as.integer(final\_ds\_initial\_1[row, "gvkey"])  
   
 restmt\_at <- final\_ds\_initial\_1[row, "restmt\_at"]  
 restmt\_at\_mag <- final\_ds\_initial\_1[row, "restmt\_at\_mag"]  
 if (restmt\_at >= 0.5){  
 restmt\_at <- 1  
 restmt\_at\_mag <- as.double(restmt\_at\_mag)  
 }  
 else{  
 restmt\_at <- 0  
 restmt\_at\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_at[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_at  
 final\_ds\_initial\_1$restmt\_at\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_at\_mag  
   
 restmt\_capx <- final\_ds\_initial\_1[row, "restmt\_capx"]  
 restmt\_capx\_mag <- final\_ds\_initial\_1[row, "restmt\_capx\_mag"]  
 if (restmt\_capx >= 0.5){  
 restmt\_capx <- 1  
 restmt\_capx\_mag <- as.double(restmt\_capx\_mag)  
 }  
 else{  
 restmt\_capx <- 0  
 restmt\_capx\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_capx[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_capx  
 final\_ds\_initial\_1$restmt\_capx\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_capx\_mag  
   
 restmt\_cogs <- final\_ds\_initial\_1[row, "restmt\_cogs"]  
 restmt\_cogs\_mag <- final\_ds\_initial\_1[row, "restmt\_cogs\_mag"]  
 if (restmt\_cogs >= 0.5){  
 restmt\_cogs <- 1  
 restmt\_cogs\_mag <- as.double(restmt\_cogs\_mag)  
 }  
 else{  
 restmt\_cogs <- 0  
 restmt\_cogs\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_cogs[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_cogs  
 final\_ds\_initial\_1$restmt\_cogs\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_cogs\_mag  
   
 restmt\_dltt <- final\_ds\_initial\_1[row, "restmt\_dltt"]  
 restmt\_dltt\_mag <- final\_ds\_initial\_1[row, "restmt\_dltt\_mag"]  
 if (restmt\_dltt >= 0.5){  
 restmt\_dltt <- 1  
 restmt\_dltt\_mag <- as.double(restmt\_dltt\_mag)  
 }  
 else{  
 restmt\_dltt <- 0  
 restmt\_dltt\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_dltt[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dltt  
 final\_ds\_initial\_1$restmt\_dltt\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dltt\_mag  
   
   
 restmt\_epsfi <- final\_ds\_initial\_1[row, "restmt\_epsfi"]  
 restmt\_epsfi\_mag <- final\_ds\_initial\_1[row, "restmt\_epsfi\_mag"]  
 if (restmt\_epsfi >= 0.5){  
 restmt\_epsfi <- 1  
 restmt\_epsfi\_mag <- as.double(restmt\_epsfi\_mag)  
 }  
 else{  
 restmt\_epsfi <- 0  
 restmt\_epsfi\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_epsfi[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_epsfi  
 final\_ds\_initial\_1$restmt\_epsfi\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_epsfi\_mag  
   
   
 restmt\_epspi <- final\_ds\_initial\_1[row, "restmt\_epspi"]  
 restmt\_epspi\_mag <- final\_ds\_initial\_1[row, "restmt\_epspi\_mag"]  
 if (restmt\_epspi >= 0.5){  
 restmt\_epspi <- 1  
 restmt\_epspi\_mag <- as.double(restmt\_epspi\_mag)  
 }  
 else{  
 restmt\_epspi <- 0  
 restmt\_epspi\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_epspi[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_epspi  
 final\_ds\_initial\_1$restmt\_epspi\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_epspi\_mag  
   
 restmt\_ib <- final\_ds\_initial\_1[row, "restmt\_ib"]  
 restmt\_ib\_mag <- final\_ds\_initial\_1[row, "restmt\_ib\_mag"]  
 if (restmt\_ib >= 0.5){  
 restmt\_ib <- 1  
 restmt\_ib\_mag <- as.double(restmt\_ib\_mag)  
 }  
 else{  
 restmt\_ib <- 0  
 restmt\_ib\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_ib[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_ib  
 final\_ds\_initial\_1$restmt\_ib\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_ib\_mag  
   
 restmt\_ni <- final\_ds\_initial\_1[row, "restmt\_ni"]  
 restmt\_ni\_mag <- final\_ds\_initial\_1[row, "restmt\_ni\_mag"]  
 if (restmt\_ni >= 0.5){  
 restmt\_ni <- 1  
 restmt\_ni\_mag <- as.double(restmt\_ni\_mag)  
 }  
 else{  
 restmt\_ni <- 0  
 restmt\_ni\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_ni[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_ni  
 final\_ds\_initial\_1$restmt\_ni\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_ni\_mag  
   
 restmt\_nopi <- final\_ds\_initial\_1[row, "restmt\_nopi"]  
 restmt\_nopi\_mag <- final\_ds\_initial\_1[row, "restmt\_nopi\_mag"]  
 if (restmt\_nopi >= 0.5){  
 restmt\_nopi <- 1  
 restmt\_nopi\_mag <- as.double(restmt\_nopi\_mag)  
 }  
 else{  
 restmt\_nopi <- 0  
 restmt\_nopi\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_nopi[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_nopi  
 final\_ds\_initial\_1$restmt\_nopi\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_nopi\_mag  
   
 restmt\_pi <- final\_ds\_initial\_1[row, "restmt\_pi"]  
 restmt\_pi\_mag <- final\_ds\_initial\_1[row, "restmt\_pi\_mag"]  
 if (restmt\_pi >= 0.5){  
 restmt\_pi <- 1  
 restmt\_pi\_mag <- as.double(restmt\_pi\_mag)  
 }  
 else{  
 restmt\_pi <- 0  
 restmt\_pi\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_pi[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_pi  
 final\_ds\_initial\_1$restmt\_pi\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_pi\_mag  
   
 restmt\_reuna <- final\_ds\_initial\_1[row, "restmt\_reuna"]  
 restmt\_reuna\_mag <- final\_ds\_initial\_1[row, "restmt\_reuna\_mag"]  
 if (restmt\_reuna >= 0.5){  
 restmt\_reuna <- 1  
 restmt\_reuna\_mag <- as.double(restmt\_reuna\_mag)  
 }  
 else{  
 restmt\_reuna <- 0  
 restmt\_reuna\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_reuna[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_reuna  
 final\_ds\_initial\_1$restmt\_reuna\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_reuna\_mag  
   
 restmt\_seq <- final\_ds\_initial\_1[row, "restmt\_seq"]  
 restmt\_seq\_mag <- final\_ds\_initial\_1[row, "restmt\_seq\_mag"]  
 if (restmt\_seq >= 0.5){  
 restmt\_seq <- 1  
 restmt\_seq\_mag <- as.double(restmt\_seq\_mag)  
 }  
 else{  
 restmt\_seq <- 0  
 restmt\_seq\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_seq[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_seq  
 final\_ds\_initial\_1$restmt\_seq\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_seq\_mag  
   
 restmt\_teq <- final\_ds\_initial\_1[row, "restmt\_teq"]  
 restmt\_teq\_mag <- final\_ds\_initial\_1[row, "restmt\_teq\_mag"]  
 if (restmt\_teq >= 0.5){  
 restmt\_teq <- 1  
 restmt\_teq\_mag <- as.double(restmt\_teq\_mag)  
 }  
 else{  
 restmt\_teq <- 0  
 restmt\_teq\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_teq[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_teq  
 final\_ds\_initial\_1$restmt\_teq\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_teq\_mag  
   
 restmt\_txt <- final\_ds\_initial\_1[row, "restmt\_txt"]  
 restmt\_txt\_mag <- final\_ds\_initial\_1[row, "restmt\_txt\_mag"]  
 if (restmt\_txt >= 0.5){  
 restmt\_txt <- 1  
 restmt\_txt\_mag <- as.double(restmt\_txt\_mag)  
 }  
 else{  
 restmt\_txt <- 0  
 restmt\_txt\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_txt[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_txt  
 final\_ds\_initial\_1$restmt\_txt\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_txt\_mag  
   
 restmt\_wcap <- final\_ds\_initial\_1[row, "restmt\_wcap"]  
 restmt\_wcap\_mag <- final\_ds\_initial\_1[row, "restmt\_wcap\_mag"]  
 if (restmt\_wcap >= 0.5){  
 restmt\_wcap <- 1  
 restmt\_wcap\_mag <- as.double(restmt\_wcap\_mag)  
 }  
 else{  
 restmt\_wcap <- 0  
 restmt\_wcap\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_wcap[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_wcap  
 final\_ds\_initial\_1$restmt\_wcap\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_wcap\_mag  
   
 restmt\_xint <- final\_ds\_initial\_1[row, "restmt\_xint"]  
 restmt\_xint\_mag <- final\_ds\_initial\_1[row, "restmt\_xint\_mag"]  
 if (restmt\_xint >= 0.5){  
 restmt\_xint <- 1  
 restmt\_xint\_mag <- as.double(restmt\_xint\_mag)  
 }  
 else{  
 restmt\_xint <- 0  
 restmt\_xint\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_xint[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_xint  
 final\_ds\_initial\_1$restmt\_xint\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_xint\_mag  
   
 restmt\_xsga <- final\_ds\_initial\_1[row, "restmt\_xsga"]  
 restmt\_xsga\_mag <- final\_ds\_initial\_1[row, "restmt\_xsga\_mag"]  
 if (restmt\_xsga >= 0.5){  
 restmt\_xsga <- 1  
 restmt\_xsga\_mag <- as.double(restmt\_xsga\_mag)  
 }  
 else{  
 restmt\_xsga <- 0  
 restmt\_xsga\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_xsga[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_xsga  
 final\_ds\_initial\_1$restmt\_xsga\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_xsga\_mag  
   
 restmt\_dvpsp\_f <- final\_ds\_initial\_1[row, "restmt\_dvpsp\_f"]  
 restmt\_dvpsp\_f\_mag <- final\_ds\_initial\_1[row, "restmt\_dvpsp\_f\_mag"]  
 if (restmt\_dvpsp\_f >= 0.5){  
 restmt\_dvpsp\_f <- 1  
 restmt\_dvpsp\_f\_mag <- as.double(restmt\_dvpsp\_f\_mag)  
 }  
 else{  
 restmt\_dvpsp\_f <- 0  
 restmt\_dvpsp\_f\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_dvpsp\_f[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dvpsp\_f  
 final\_ds\_initial\_1$restmt\_dvpsp\_f\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dvpsp\_f\_mag  
   
 restmt\_dvpsx\_f <- final\_ds\_initial\_1[row, "restmt\_dvpsx\_f"]  
 restmt\_dvpsx\_f\_mag <- final\_ds\_initial\_1[row, "restmt\_dvpsx\_f\_mag"]  
 if (restmt\_dvpsx\_f >= 0.5){  
 restmt\_dvpsx\_f <- 1  
 restmt\_dvpsx\_f\_mag <- as.double(restmt\_dvpsx\_f\_mag)  
 }  
 else{  
 restmt\_dvpsx\_f <- 0  
 restmt\_dvpsx\_f\_mag <- 0.0  
 }  
 final\_ds\_initial\_1$restmt\_dvpsx\_f[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dvpsx\_f  
 final\_ds\_initial\_1$restmt\_dvpsx\_f\_mag[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- restmt\_dvpsx\_f\_mag  
}  
  
restmt\_var\_ds <- subset(final\_ds\_initial\_1, select = c(gvkey,  
 restmt\_at,restmt\_at\_mag,  
 restmt\_capx,restmt\_capx\_mag,  
 restmt\_cogs, restmt\_cogs\_mag,  
 restmt\_dltt, restmt\_dltt\_mag,  
 restmt\_epsfi, restmt\_epsfi\_mag,  
 restmt\_epspi, restmt\_epspi\_mag,  
 restmt\_ib, restmt\_ib\_mag,  
 restmt\_ni, restmt\_ni\_mag,  
 restmt\_nopi, restmt\_nopi\_mag,  
 restmt\_pi, restmt\_pi\_mag,  
 restmt\_reuna, restmt\_reuna\_mag,  
 restmt\_seq, restmt\_seq\_mag,  
 restmt\_teq, restmt\_teq\_mag,  
 restmt\_txt, restmt\_txt\_mag,  
 restmt\_wcap, restmt\_wcap\_mag,  
   
 restmt\_xint, restmt\_xint\_mag,  
 restmt\_xsga, restmt\_xsga\_mag,  
 restmt\_dvpsp\_f, restmt\_dvpsp\_f\_mag,  
 restmt\_dvpsx\_f, restmt\_dvpsx\_f\_mag  
 ))  
  
summary(restmt\_var\_ds)

## gvkey restmt\_at restmt\_at\_mag restmt\_capx   
## Min. : 1239 Min. :0.00000 Min. : -1.4907 Min. :0.00000   
## 1st Qu.: 12107 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.00000   
## Median : 61311 Median :0.00000 Median : 0.0000 Median :0.00000   
## Mean : 83018 Mean :0.06322 Mean : 0.5283 Mean :0.02874   
## 3rd Qu.:165694 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.00000   
## Max. :277487 Max. :1.00000 Max. :182.4888 Max. :1.00000   
## restmt\_capx\_mag restmt\_cogs restmt\_cogs\_mag restmt\_dltt   
## Min. :-22.71625 Min. :0.0000 Min. : -50.00 Min. :0.00000   
## 1st Qu.: 0.00000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00000 Median :0.0000 Median : 0.00 Median :0.00000   
## Mean : -0.06393 Mean :0.3046 Mean : 29.37 Mean :0.01724   
## 3rd Qu.: 0.00000 3rd Qu.:1.0000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. : 8.33350 Max. :1.0000 Max. :9299.36 Max. :1.00000   
## restmt\_dltt\_mag restmt\_epsfi restmt\_epsfi\_mag restmt\_epspi   
## Min. :-26.9567 Min. :0.0000 Min. : -50.05 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.00 Median :0.0000   
## Mean : 0.1794 Mean :0.1293 Mean : 344.76 Mean :0.1351   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.00 3rd Qu.:0.0000   
## Max. :100.9780 Max. :1.0000 Max. :77081.67 Max. :1.0000   
## restmt\_epspi\_mag restmt\_ib restmt\_ib\_mag restmt\_ni   
## Min. : -50.0 Min. :0.0000 Min. :-121.766 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.0000 Median : 0.000 Median :0.00000   
## Mean : 346.9 Mean :0.1121 Mean : 6.836 Mean :0.04598   
## 3rd Qu.: 0.0 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :77081.7 Max. :1.0000 Max. :2683.890 Max. :1.00000   
## restmt\_ni\_mag restmt\_nopi restmt\_nopi\_mag restmt\_pi   
## Min. : -9.801 Min. :0.0000 Min. :-1868600.0 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: -92.7 1st Qu.:0.0000   
## Median : 0.000 Median :1.0000 Median : 0.0 Median :0.0000   
## Mean : 8.153 Mean :0.6322 Mean : -5817.4 Mean :0.0977   
## 3rd Qu.: 0.000 3rd Qu.:1.0000 3rd Qu.: 21.5 3rd Qu.:0.0000   
## Max. :2683.890 Max. :1.0000 Max. : 68865.1 Max. :1.0000   
## restmt\_pi\_mag restmt\_reuna restmt\_reuna\_mag restmt\_seq   
## Min. :-135.28 Min. :0.00000 Min. :-2461.679 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.00 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : 6.98 Mean :0.07759 Mean : 5.001 Mean :0.09483   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :2683.89 Max. :1.00000 Max. : 4181.704 Max. :1.00000   
## restmt\_seq\_mag restmt\_teq restmt\_teq\_mag restmt\_txt   
## Min. : -105.4 Min. :0.00000 Min. : -105.39 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.0 Median :0.00000 Median : 0.00 Median :0.00000   
## Mean : 49.1 Mean :0.08621 Mean : 49.71 Mean :0.07184   
## 3rd Qu.: 0.0 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :12541.8 Max. :1.00000 Max. :12541.75 Max. :1.00000   
## restmt\_txt\_mag restmt\_wcap restmt\_wcap\_mag restmt\_xint   
## Min. :-88.7704 Min. :0.0000 Min. :-43.249 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0.0000   
## Mean : -0.7665 Mean :0.0431 Mean : 1.043 Mean :0.1178   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.0000   
## Max. : 47.3182 Max. :1.0000 Max. :412.500 Max. :1.0000   
## restmt\_xint\_mag restmt\_xsga restmt\_xsga\_mag restmt\_dvpsp\_f  
## Min. :-62.7347 Min. :0.0000 Min. : -50.000 Min. :0   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0   
## Mean : -0.8803 Mean :0.1552 Mean : 4.984 Mean :0   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0   
## Max. : 0.5620 Max. :1.0000 Max. :1884.021 Max. :0   
## restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f restmt\_dvpsx\_f\_mag  
## Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0

final\_ds\_initial\_2 <- final\_ds\_initial\_1  
summary(final\_ds\_initial\_2)

## gvkey tic aco acominc   
## Min. : 1239 0161A : 1 Min. : 0.000 Min. :-19306.57   
## 1st Qu.: 12107 0170A : 1 1st Qu.: 0.447 1st Qu.: -30.39   
## Median : 61311 0171A : 1 Median : 8.858 Median : 0.00   
## Mean : 83018 0173A : 1 Mean : 188.577 Mean : -194.14   
## 3rd Qu.:165694 0270B : 1 3rd Qu.: 94.290 3rd Qu.: 0.00   
## Max. :277487 0563B : 1 Max. :4760.750 Max. : 3495.34   
## (Other):342   
## act ao aocidergl aocipen   
## Min. : 0.00 Min. : 0.000 Min. :-2207.250 Min. :-2803.25   
## 1st Qu.: 10.27 1st Qu.: 0.145 1st Qu.: 0.000 1st Qu.: -10.48   
## Median : 115.71 Median : 8.322 Median : 0.000 Median : 0.00   
## Mean : 1918.97 Mean : 200.854 Mean : -9.098 Mean : -91.21   
## 3rd Qu.: 1225.80 3rd Qu.: 93.865 3rd Qu.: 0.000 3rd Qu.: 0.00   
## Max. :55264.80 Max. :5330.250 Max. : 119.000 Max. : 30.75   
##   
## aodo aoloch ap aqc   
## Min. : 0.000 Min. :-667.500 Min. : 0.00 Min. : -12.45   
## 1st Qu.: 0.070 1st Qu.: -1.851 1st Qu.: 1.17 1st Qu.: 0.00   
## Median : 6.675 Median : 0.000 Median : 18.13 Median : 0.00   
## Mean : 188.752 Mean : 6.660 Mean : 660.17 Mean : 105.33   
## 3rd Qu.: 91.195 3rd Qu.: 1.530 3rd Qu.: 241.09 3rd Qu.: 14.95   
## Max. :5330.250 Max. : 744.000 Max. :35222.20 Max. :5559.02   
##   
## at bkvlps caps capx   
## Min. : 0.00 Min. :-130515.0 Min. : -701.48 Min. : 0.000   
## 1st Qu.: 20.64 1st Qu.: 0.1 1st Qu.: 5.25 1st Qu.: 0.343   
## Median : 283.75 Median : 3.8 Median : 37.25 Median : 12.573   
## Mean : 6489.03 Mean : 11681.8 Mean : 951.28 Mean : 242.490   
## 3rd Qu.: 3172.25 3rd Qu.: 12.5 3rd Qu.: 363.48 3rd Qu.: 119.642   
## Max. :190526.20 Max. :1881687.0 Max. :62705.25 Max. :12881.200   
##   
## ceq ceqt ch che   
## Min. :-2342.49 Min. :-40530.25 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 3.36 1st Qu.: -3.63 1st Qu.: 1.399 1st Qu.: 1.496   
## Median : 105.08 Median : 13.54 Median : 21.299 Median : 26.194   
## Mean : 2388.45 Mean : 54.96 Mean : 389.436 Mean : 487.347   
## 3rd Qu.: 1033.62 3rd Qu.: 189.37 3rd Qu.: 179.819 3rd Qu.: 217.474   
## Max. :72640.80 Max. : 53931.40 Max. :10044.000 Max. :15547.750   
##   
## chech ci cogs cshi   
## Min. :-305.7500 Min. : -722.617 Min. : 0.0 Min. : 0.0   
## 1st Qu.: -0.1368 1st Qu.: -1.477 1st Qu.: 12.1 1st Qu.: 16.2   
## Median : 0.5806 Median : 9.139 Median : 216.7 Median : 49.8   
## Mean : 42.2654 Mean : 475.730 Mean : 5116.2 Mean : 267.3   
## 3rd Qu.: 10.6559 3rd Qu.: 130.370 3rd Qu.: 2471.9 3rd Qu.: 145.1   
## Max. :1543.0000 Max. :16365.200 Max. :325065.8 Max. :6253.5   
##   
## csho cstk cstkcv dd1   
## Min. : 0.00 Min. : 0.000 Min. : 0.0000 Min. : 0.000   
## 1st Qu.: 16.32 1st Qu.: 0.026 1st Qu.: 0.0010 1st Qu.: 0.000   
## Median : 49.41 Median : 0.248 Median : 0.0100 Median : 1.451   
## Mean : 240.67 Mean : 191.237 Mean : 0.6068 Mean : 169.076   
## 3rd Qu.: 142.42 3rd Qu.: 21.315 3rd Qu.: 0.2500 3rd Qu.: 39.533   
## Max. :6252.56 Max. :7290.750 Max. :20.8642 Max. :5428.500   
##   
## dilavx dlc dltt dm   
## Min. : -738.263 Min. : 0.000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: -1.320 1st Qu.: 0.302 1st Qu.: 0.16 1st Qu.: 0.000   
## Median : 4.824 Median : 5.257 Median : 17.03 Median : 1.218   
## Mean : 460.750 Mean : 374.668 Mean : 1477.91 Mean : 157.728   
## 3rd Qu.: 125.014 3rd Qu.: 99.945 3rd Qu.: 902.29 3rd Qu.: 62.483   
## Max. :15690.400 Max. :15926.126 Max. :42659.60 Max. :3900.400   
##   
## dn dpact dpc dvt   
## Min. : 0.0 Min. : 0.00 Min. : 0.000 Min. : -0.006   
## 1st Qu.: 0.0 1st Qu.: 2.02 1st Qu.: 0.319 1st Qu.: 0.000   
## Median : 0.0 Median : 52.40 Median : 7.505 Median : 0.000   
## Mean : 902.6 Mean : 1303.11 Mean : 170.666 Mean : 226.759   
## 3rd Qu.: 252.3 3rd Qu.: 768.89 3rd Qu.: 95.563 3rd Qu.: 35.368   
## Max. :42561.8 Max. :50449.80 Max. :8059.800 Max. :6572.535   
##   
## ebit ebitda epsfi epspi   
## Min. : -208.760 Min. : -33.68 Min. :-14.0200 Min. :-14.0200   
## 1st Qu.: -0.369 1st Qu.: 0.02 1st Qu.: -0.0512 1st Qu.: -0.0512   
## Median : 23.871 Median : 31.53 Median : 0.2288 Median : 0.2362   
## Mean : 761.483 Mean : 940.88 Mean : 3.9554 Mean : 3.9707   
## 3rd Qu.: 345.869 3rd Qu.: 441.54 3rd Qu.: 1.8338 3rd Qu.: 1.8638   
## Max. :24345.400 Max. :32405.20 Max. :881.6400 Max. :881.6400   
##   
## fiao fincf fopo   
## Min. :-3427.000 Min. :-11533.200 Min. :-389.5000   
## 1st Qu.: -7.661 1st Qu.: -50.075 1st Qu.: 0.1661   
## Median : -0.047 Median : -0.005 Median : 2.0777   
## Mean : -45.439 Mean : -332.514 Mean : 74.1085   
## 3rd Qu.: 0.000 3rd Qu.: 4.030 3rd Qu.: 23.4436   
## Max. : 1800.250 Max. : 824.184 Max. :1979.4552   
##   
## gdwl gp ib icapt   
## Min. : 0.00 Min. : -3.19 Min. : -727.025 Min. : -23.14   
## 1st Qu.: 0.00 1st Qu.: 6.98 1st Qu.: -1.320 1st Qu.: 10.74   
## Median : 1.95 Median : 102.86 Median : 6.421 Median : 192.63   
## Mean : 1492.38 Mean : 2502.66 Mean : 467.872 Mean : 4010.77   
## 3rd Qu.: 390.49 3rd Qu.: 1238.36 3rd Qu.: 136.674 3rd Qu.: 2098.22   
## Max. :56373.25 Max. :117445.60 Max. :15690.400 Max. :119888.20   
##   
## intan intano invt ivch   
## Min. : 0.00 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.17 1st Qu.: 0.04 1st Qu.: 2.19 1st Qu.: 0.00   
## Median : 18.61 Median : 7.21 Median : 37.47 Median : 0.00   
## Mean : 2370.53 Mean : 878.15 Mean : 745.30 Mean : 60.27   
## 3rd Qu.: 715.80 3rd Qu.: 217.88 3rd Qu.: 464.24 3rd Qu.: 0.27   
## Max. :86837.75 Max. :31704.00 Max. :39770.60 Max. :4366.68   
##   
## ivncf ivst lo lse   
## Min. :-13066.20 Min. : 0.000 Min. : -128.941 Min. : 0.00   
## 1st Qu.: -176.92 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 20.64   
## Median : -19.36 Median : 0.000 Median : 5.334 Median : 283.75   
## Mean : -314.10 Mean : 88.668 Mean : 475.282 Mean : 6489.03   
## 3rd Qu.: -0.24 3rd Qu.: 2.429 3rd Qu.: 203.087 3rd Qu.: 3172.25   
## Max. : 985.75 Max. :5503.750 Max. :14517.069 Max. :190526.20   
##   
## lt ni nopi   
## Min. : 0.02 Min. : -737.537 Min. :-230.2500   
## 1st Qu.: 8.05 1st Qu.: -1.661 1st Qu.: 0.0000   
## Median : 108.02 Median : 6.217 Median : 0.1979   
## Mean : 3948.06 Mean : 499.241 Mean : 47.9687   
## 3rd Qu.: 2056.19 3rd Qu.: 125.392 3rd Qu.: 4.8163   
## Max. :113297.60 Max. :17374.318 Max. :2224.4000   
##   
## nopio oancf oiadp oibdp   
## Min. :-230.2500 Min. : -61.444 Min. : -208.760 Min. : -33.68   
## 1st Qu.: -0.0111 1st Qu.: -0.142 1st Qu.: -0.369 1st Qu.: 0.02   
## Median : 0.0664 Median : 20.797 Median : 23.871 Median : 31.53   
## Mean : 37.9242 Mean : 696.038 Mean : 761.483 Mean : 940.88   
## 3rd Qu.: 2.5212 3rd Qu.: 282.996 3rd Qu.: 345.869 3rd Qu.: 441.54   
## Max. :2054.4000 Max. :24599.000 Max. :24345.400 Max. :32405.20   
##   
## opeps pi ppegt re   
## Min. : -9.8200 Min. : -739.921 Min. : 0.00 Min. :-7570.29   
## 1st Qu.: -0.0350 1st Qu.: -1.287 1st Qu.: 5.51 1st Qu.: -9.35   
## Median : 0.2971 Median : 11.654 Median : 146.66 Median : 19.16   
## Mean : 3.9977 Mean : 678.381 Mean : 2919.87 Mean : 1909.92   
## 3rd Qu.: 1.8725 3rd Qu.: 215.671 3rd Qu.: 1576.40 3rd Qu.: 441.34   
## Max. :856.8325 Max. :24079.000 Max. :161869.20 Max. :68884.60   
##   
## reajo rect recta reuna   
## Min. :-7860.75 Min. : 0.000 Min. :-19466.259 Min. :-7527.73   
## 1st Qu.: -19.79 1st Qu.: 1.636 1st Qu.: -0.169 1st Qu.: -9.13   
## Median : 0.00 Median : 28.478 Median : 0.000 Median : 21.03   
## Mean : -78.00 Mean : 510.810 Mean : -96.126 Mean : 2011.34   
## 3rd Qu.: 0.00 3rd Qu.: 312.676 3rd Qu.: 0.060 3rd Qu.: 435.07   
## Max. : 7171.53 Max. :15020.067 Max. : 1946.250 Max. :72710.50   
##   
## revt seq siv spce   
## Min. : 0.0 Min. :-2208.96 Min. : 0.000 Min. : -600.364   
## 1st Qu.: 22.4 1st Qu.: 4.15 1st Qu.: 0.000 1st Qu.: -1.121   
## Median : 333.1 Median : 106.70 Median : 0.000 Median : 6.643   
## Mean : 7618.9 Mean : 2442.71 Mean : 56.788 Mean : 459.367   
## 3rd Qu.: 3826.2 3rd Qu.: 1091.19 3rd Qu.: 0.512 3rd Qu.: 138.243   
## Max. :442511.4 Max. :72640.80 Max. :4366.827 Max. :15690.400   
##   
## spi sppiv sstk teq   
## Min. :-921.2962 Min. :-6191.874 Min. : 0.0000 Min. :-2208.96   
## 1st Qu.: -15.8642 1st Qu.: -0.046 1st Qu.: 0.0006 1st Qu.: 4.19   
## Median : -0.4417 Median : 0.000 Median : 1.2209 Median : 106.70   
## Mean : -35.2648 Mean : -32.516 Mean : 36.8987 Mean : 2524.10   
## 3rd Qu.: 0.0000 3rd Qu.: 0.007 3rd Qu.: 14.5968 3rd Qu.: 1095.88   
## Max. :1115.5000 Max. : 27.017 Max. :1513.0000 Max. :76602.80   
##   
## tstk tstkn txp txr   
## Min. : 0.0 Min. : 0.000 Min. : -0.252 Min. : 0.0000   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.0000   
## Median : 0.0 Median : 0.000 Median : 0.049 Median : 0.0000   
## Mean : 889.8 Mean : 27.024 Mean : 44.423 Mean : 11.9381   
## 3rd Qu.: 8.5 3rd Qu.: 1.282 3rd Qu.: 5.603 3rd Qu.: 0.0758   
## Max. :67539.2 Max. :1923.500 Max. :1469.476 Max. :1150.7513   
##   
## txt wcap xint restmt\_at   
## Min. : -76.388 Min. :-8236.800 Min. : 0.0000 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.: -0.011 1st Qu.: 0.1635 1st Qu.:0.00000   
## Median : 3.772 Median : 25.438 Median : 2.2978 Median :0.00000   
## Mean : 189.846 Mean : 228.503 Mean : 95.8065 Mean :0.06322   
## 3rd Qu.: 63.206 3rd Qu.: 279.322 3rd Qu.: 66.5683 3rd Qu.:0.00000   
## Max. :7749.600 Max. :12261.750 Max. :2859.7500 Max. :1.00000   
##   
## restmt\_at\_mag restmt\_capx restmt\_capx\_mag restmt\_cogs   
## Min. : -1.4907 Min. :0.00000 Min. :-22.71625 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.00000 1st Qu.: 0.00000 1st Qu.:0.0000   
## Median : 0.0000 Median :0.00000 Median : 0.00000 Median :0.0000   
## Mean : 0.5283 Mean :0.02874 Mean : -0.06393 Mean :0.3046   
## 3rd Qu.: 0.0000 3rd Qu.:0.00000 3rd Qu.: 0.00000 3rd Qu.:1.0000   
## Max. :182.4888 Max. :1.00000 Max. : 8.33350 Max. :1.0000   
##   
## restmt\_cogs\_mag restmt\_dltt restmt\_dltt\_mag restmt\_epsfi   
## Min. : -50.00 Min. :0.00000 Min. :-26.9567 Min. :0.0000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.00 Median :0.00000 Median : 0.0000 Median :0.0000   
## Mean : 29.37 Mean :0.01724 Mean : 0.1794 Mean :0.1293   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.0000   
## Max. :9299.36 Max. :1.00000 Max. :100.9780 Max. :1.0000   
##   
## restmt\_epsfi\_mag restmt\_epspi restmt\_epspi\_mag restmt\_ib   
## Min. : -50.05 Min. :0.0000 Min. : -50.0 Min. :0.0000   
## 1st Qu.: 0.00 1st Qu.:0.0000 1st Qu.: 0.0 1st Qu.:0.0000   
## Median : 0.00 Median :0.0000 Median : 0.0 Median :0.0000   
## Mean : 344.76 Mean :0.1351 Mean : 346.9 Mean :0.1121   
## 3rd Qu.: 0.00 3rd Qu.:0.0000 3rd Qu.: 0.0 3rd Qu.:0.0000   
## Max. :77081.67 Max. :1.0000 Max. :77081.7 Max. :1.0000   
##   
## restmt\_ib\_mag restmt\_ni restmt\_ni\_mag restmt\_nopi   
## Min. :-121.766 Min. :0.00000 Min. : -9.801 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.000 Median :0.00000 Median : 0.000 Median :1.0000   
## Mean : 6.836 Mean :0.04598 Mean : 8.153 Mean :0.6322   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:1.0000   
## Max. :2683.890 Max. :1.00000 Max. :2683.890 Max. :1.0000   
##   
## restmt\_nopi\_mag restmt\_pi restmt\_pi\_mag restmt\_reuna   
## Min. :-1868600.0 Min. :0.0000 Min. :-135.28 Min. :0.00000   
## 1st Qu.: -92.7 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.0 Median :0.0000 Median : 0.00 Median :0.00000   
## Mean : -5817.4 Mean :0.0977 Mean : 6.98 Mean :0.07759   
## 3rd Qu.: 21.5 3rd Qu.:0.0000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. : 68865.1 Max. :1.0000 Max. :2683.89 Max. :1.00000   
##   
## restmt\_reuna\_mag restmt\_seq restmt\_seq\_mag restmt\_teq   
## Min. :-2461.679 Min. :0.00000 Min. : -105.4 Min. :0.00000   
## 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.0 1st Qu.:0.00000   
## Median : 0.000 Median :0.00000 Median : 0.0 Median :0.00000   
## Mean : 5.001 Mean :0.09483 Mean : 49.1 Mean :0.08621   
## 3rd Qu.: 0.000 3rd Qu.:0.00000 3rd Qu.: 0.0 3rd Qu.:0.00000   
## Max. : 4181.704 Max. :1.00000 Max. :12541.8 Max. :1.00000   
##   
## restmt\_teq\_mag restmt\_txt restmt\_txt\_mag restmt\_wcap   
## Min. : -105.39 Min. :0.00000 Min. :-88.7704 Min. :0.0000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.00 Median :0.00000 Median : 0.0000 Median :0.0000   
## Mean : 49.71 Mean :0.07184 Mean : -0.7665 Mean :0.0431   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.0000   
## Max. :12541.75 Max. :1.00000 Max. : 47.3182 Max. :1.0000   
##   
## restmt\_wcap\_mag restmt\_xint restmt\_xint\_mag restmt\_xsga   
## Min. :-43.249 Min. :0.0000 Min. :-62.7347 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Median : 0.000 Median :0.0000 Median : 0.0000 Median :0.0000   
## Mean : 1.043 Mean :0.1178 Mean : -0.8803 Mean :0.1552   
## 3rd Qu.: 0.000 3rd Qu.:0.0000 3rd Qu.: 0.0000 3rd Qu.:0.0000   
## Max. :412.500 Max. :1.0000 Max. : 0.5620 Max. :1.0000   
##   
## restmt\_xsga\_mag restmt\_dvpsp\_f restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f  
## Min. : -50.000 Min. :0 Min. :0 Min. :0   
## 1st Qu.: 0.000 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median : 0.000 Median :0 Median :0 Median :0   
## Mean : 4.984 Mean :0 Mean :0 Mean :0   
## 3rd Qu.: 0.000 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :1884.021 Max. :0 Max. :0 Max. :0   
##   
## restmt\_dvpsx\_f\_mag  
## Min. :0   
## 1st Qu.:0   
## Median :0   
## Mean :0   
## 3rd Qu.:0   
## Max. :0   
##

nrow(final\_ds\_initial\_2)

## [1] 348

write.csv(final\_ds\_initial\_2, file = "data/final\_ds\_initial\_2.csv", row.names=FALSE, na="")

### This first step where all the correlated variables are idenfied and then removed. This will reduce the Collinearity.

cor\_matrix\_ds <- subset(final\_ds\_initial\_2, select = -c(gvkey,tic, aodo,seq,ivch,nopio,spce,reuna,dilavx,ebitda,cshi,epsfi,   
 ib,pi,  
 oiadp,oibdp,gdwl))  
cor\_matrix <- cor(cor\_matrix\_ds)  
cor\_matrix %>%  
 as.data.frame() %>%  
 mutate(var1 = rownames(.)) %>%  
 gather(var2, value, -var1) %>%  
 arrange(desc(value)) %>%  
 group\_by(value) %>%  
 filter(row\_number()==1)

## # A tibble: 5,052 x 3  
## # Groups: value [5,052]  
## var1 var2 value  
## <chr> <chr> <dbl>  
## 1 aco aco 1   
## 2 restmt\_epspi\_mag restmt\_epsfi\_mag 1.00   
## 3 restmt\_teq\_mag restmt\_seq\_mag 1.00   
## 4 opeps epspi 1.00   
## 5 restmt\_pi\_mag restmt\_ib\_mag 0.998  
## 6 restmt\_ni\_mag restmt\_ib\_mag 0.998  
## 7 restmt\_xsga\_mag restmt\_ni\_mag 0.997  
## 8 restmt\_xsga\_mag restmt\_pi\_mag 0.997  
## 9 restmt\_pi\_mag restmt\_ni\_mag 0.997  
## 10 restmt\_xsga\_mag restmt\_ib\_mag 0.997  
## # ... with 5,042 more rows

#corrplot(cor\_matrix, method = "ellipse")  
#ncol(cor\_matrix\_ds)

fundamentals\_final\_ds <- subset(final\_ds\_initial\_2, select = -c(aodo,seq,ivch,nopio,spce,reuna,dilavx,ebitda,cshi,epsfi,   
 ib,pi,  
 oiadp,oibdp,gdwl))  
summary(fundamentals\_final\_ds)

## gvkey tic aco acominc   
## Min. : 1239 0161A : 1 Min. : 0.000 Min. :-19306.57   
## 1st Qu.: 12107 0170A : 1 1st Qu.: 0.447 1st Qu.: -30.39   
## Median : 61311 0171A : 1 Median : 8.858 Median : 0.00   
## Mean : 83018 0173A : 1 Mean : 188.577 Mean : -194.14   
## 3rd Qu.:165694 0270B : 1 3rd Qu.: 94.290 3rd Qu.: 0.00   
## Max. :277487 0563B : 1 Max. :4760.750 Max. : 3495.34   
## (Other):342   
## act ao aocidergl aocipen   
## Min. : 0.00 Min. : 0.000 Min. :-2207.250 Min. :-2803.25   
## 1st Qu.: 10.27 1st Qu.: 0.145 1st Qu.: 0.000 1st Qu.: -10.48   
## Median : 115.71 Median : 8.322 Median : 0.000 Median : 0.00   
## Mean : 1918.97 Mean : 200.854 Mean : -9.098 Mean : -91.21   
## 3rd Qu.: 1225.80 3rd Qu.: 93.865 3rd Qu.: 0.000 3rd Qu.: 0.00   
## Max. :55264.80 Max. :5330.250 Max. : 119.000 Max. : 30.75   
##   
## aoloch ap aqc at   
## Min. :-667.500 Min. : 0.00 Min. : -12.45 Min. : 0.00   
## 1st Qu.: -1.851 1st Qu.: 1.17 1st Qu.: 0.00 1st Qu.: 20.64   
## Median : 0.000 Median : 18.13 Median : 0.00 Median : 283.75   
## Mean : 6.660 Mean : 660.17 Mean : 105.33 Mean : 6489.03   
## 3rd Qu.: 1.530 3rd Qu.: 241.09 3rd Qu.: 14.95 3rd Qu.: 3172.25   
## Max. : 744.000 Max. :35222.20 Max. :5559.02 Max. :190526.20   
##   
## bkvlps caps capx ceq   
## Min. :-130515.0 Min. : -701.48 Min. : 0.000 Min. :-2342.49   
## 1st Qu.: 0.1 1st Qu.: 5.25 1st Qu.: 0.343 1st Qu.: 3.36   
## Median : 3.8 Median : 37.25 Median : 12.573 Median : 105.08   
## Mean : 11681.8 Mean : 951.28 Mean : 242.490 Mean : 2388.45   
## 3rd Qu.: 12.5 3rd Qu.: 363.48 3rd Qu.: 119.642 3rd Qu.: 1033.62   
## Max. :1881687.0 Max. :62705.25 Max. :12881.200 Max. :72640.80   
##   
## ceqt ch che   
## Min. :-40530.25 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -3.63 1st Qu.: 1.399 1st Qu.: 1.496   
## Median : 13.54 Median : 21.299 Median : 26.194   
## Mean : 54.96 Mean : 389.436 Mean : 487.347   
## 3rd Qu.: 189.37 3rd Qu.: 179.819 3rd Qu.: 217.474   
## Max. : 53931.40 Max. :10044.000 Max. :15547.750   
##   
## chech ci cogs csho   
## Min. :-305.7500 Min. : -722.617 Min. : 0.0 Min. : 0.00   
## 1st Qu.: -0.1368 1st Qu.: -1.477 1st Qu.: 12.1 1st Qu.: 16.32   
## Median : 0.5806 Median : 9.139 Median : 216.7 Median : 49.41   
## Mean : 42.2654 Mean : 475.730 Mean : 5116.2 Mean : 240.67   
## 3rd Qu.: 10.6559 3rd Qu.: 130.370 3rd Qu.: 2471.9 3rd Qu.: 142.42   
## Max. :1543.0000 Max. :16365.200 Max. :325065.8 Max. :6252.56   
##   
## cstk cstkcv dd1 dlc   
## Min. : 0.000 Min. : 0.0000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.026 1st Qu.: 0.0010 1st Qu.: 0.000 1st Qu.: 0.302   
## Median : 0.248 Median : 0.0100 Median : 1.451 Median : 5.257   
## Mean : 191.237 Mean : 0.6068 Mean : 169.076 Mean : 374.668   
## 3rd Qu.: 21.315 3rd Qu.: 0.2500 3rd Qu.: 39.533 3rd Qu.: 99.945   
## Max. :7290.750 Max. :20.8642 Max. :5428.500 Max. :15926.126   
##   
## dltt dm dn dpact   
## Min. : 0.00 Min. : 0.000 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 0.16 1st Qu.: 0.000 1st Qu.: 0.0 1st Qu.: 2.02   
## Median : 17.03 Median : 1.218 Median : 0.0 Median : 52.40   
## Mean : 1477.91 Mean : 157.728 Mean : 902.6 Mean : 1303.11   
## 3rd Qu.: 902.29 3rd Qu.: 62.483 3rd Qu.: 252.3 3rd Qu.: 768.89   
## Max. :42659.60 Max. :3900.400 Max. :42561.8 Max. :50449.80   
##   
## dpc dvt ebit epspi   
## Min. : 0.000 Min. : -0.006 Min. : -208.760 Min. :-14.0200   
## 1st Qu.: 0.319 1st Qu.: 0.000 1st Qu.: -0.369 1st Qu.: -0.0512   
## Median : 7.505 Median : 0.000 Median : 23.871 Median : 0.2362   
## Mean : 170.666 Mean : 226.759 Mean : 761.483 Mean : 3.9707   
## 3rd Qu.: 95.563 3rd Qu.: 35.368 3rd Qu.: 345.869 3rd Qu.: 1.8638   
## Max. :8059.800 Max. :6572.535 Max. :24345.400 Max. :881.6400   
##   
## fiao fincf fopo   
## Min. :-3427.000 Min. :-11533.200 Min. :-389.5000   
## 1st Qu.: -7.661 1st Qu.: -50.075 1st Qu.: 0.1661   
## Median : -0.047 Median : -0.005 Median : 2.0777   
## Mean : -45.439 Mean : -332.514 Mean : 74.1085   
## 3rd Qu.: 0.000 3rd Qu.: 4.030 3rd Qu.: 23.4436   
## Max. : 1800.250 Max. : 824.184 Max. :1979.4552   
##   
## gp icapt intan intano   
## Min. : -3.19 Min. : -23.14 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 6.98 1st Qu.: 10.74 1st Qu.: 0.17 1st Qu.: 0.04   
## Median : 102.86 Median : 192.63 Median : 18.61 Median : 7.21   
## Mean : 2502.66 Mean : 4010.77 Mean : 2370.53 Mean : 878.15   
## 3rd Qu.: 1238.36 3rd Qu.: 2098.22 3rd Qu.: 715.80 3rd Qu.: 217.88   
## Max. :117445.60 Max. :119888.20 Max. :86837.75 Max. :31704.00   
##   
## invt ivncf ivst lo   
## Min. : 0.00 Min. :-13066.20 Min. : 0.000 Min. : -128.941   
## 1st Qu.: 2.19 1st Qu.: -176.92 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 37.47 Median : -19.36 Median : 0.000 Median : 5.334   
## Mean : 745.30 Mean : -314.10 Mean : 88.668 Mean : 475.282   
## 3rd Qu.: 464.24 3rd Qu.: -0.24 3rd Qu.: 2.429 3rd Qu.: 203.087   
## Max. :39770.60 Max. : 985.75 Max. :5503.750 Max. :14517.069   
##   
## lse lt ni   
## Min. : 0.00 Min. : 0.02 Min. : -737.537   
## 1st Qu.: 20.64 1st Qu.: 8.05 1st Qu.: -1.661   
## Median : 283.75 Median : 108.02 Median : 6.217   
## Mean : 6489.03 Mean : 3948.06 Mean : 499.241   
## 3rd Qu.: 3172.25 3rd Qu.: 2056.19 3rd Qu.: 125.392   
## Max. :190526.20 Max. :113297.60 Max. :17374.318   
##   
## nopi oancf opeps ppegt   
## Min. :-230.2500 Min. : -61.444 Min. : -9.8200 Min. : 0.00   
## 1st Qu.: 0.0000 1st Qu.: -0.142 1st Qu.: -0.0350 1st Qu.: 5.51   
## Median : 0.1979 Median : 20.797 Median : 0.2971 Median : 146.66   
## Mean : 47.9687 Mean : 696.038 Mean : 3.9977 Mean : 2919.87   
## 3rd Qu.: 4.8163 3rd Qu.: 282.996 3rd Qu.: 1.8725 3rd Qu.: 1576.40   
## Max. :2224.4000 Max. :24599.000 Max. :856.8325 Max. :161869.20   
##   
## re reajo rect recta   
## Min. :-7570.29 Min. :-7860.75 Min. : 0.000 Min. :-19466.259   
## 1st Qu.: -9.35 1st Qu.: -19.79 1st Qu.: 1.636 1st Qu.: -0.169   
## Median : 19.16 Median : 0.00 Median : 28.478 Median : 0.000   
## Mean : 1909.92 Mean : -78.00 Mean : 510.810 Mean : -96.126   
## 3rd Qu.: 441.34 3rd Qu.: 0.00 3rd Qu.: 312.676 3rd Qu.: 0.060   
## Max. :68884.60 Max. : 7171.53 Max. :15020.067 Max. : 1946.250   
##   
## revt siv spi sppiv   
## Min. : 0.0 Min. : 0.000 Min. :-921.2962 Min. :-6191.874   
## 1st Qu.: 22.4 1st Qu.: 0.000 1st Qu.: -15.8642 1st Qu.: -0.046   
## Median : 333.1 Median : 0.000 Median : -0.4417 Median : 0.000   
## Mean : 7618.9 Mean : 56.788 Mean : -35.2648 Mean : -32.516   
## 3rd Qu.: 3826.2 3rd Qu.: 0.512 3rd Qu.: 0.0000 3rd Qu.: 0.007   
## Max. :442511.4 Max. :4366.827 Max. :1115.5000 Max. : 27.017   
##   
## sstk teq tstk tstkn   
## Min. : 0.0000 Min. :-2208.96 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 0.0006 1st Qu.: 4.19 1st Qu.: 0.0 1st Qu.: 0.000   
## Median : 1.2209 Median : 106.70 Median : 0.0 Median : 0.000   
## Mean : 36.8987 Mean : 2524.10 Mean : 889.8 Mean : 27.024   
## 3rd Qu.: 14.5968 3rd Qu.: 1095.88 3rd Qu.: 8.5 3rd Qu.: 1.282   
## Max. :1513.0000 Max. :76602.80 Max. :67539.2 Max. :1923.500   
##   
## txp txr txt wcap   
## Min. : -0.252 Min. : 0.0000 Min. : -76.388 Min. :-8236.800   
## 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: -0.011   
## Median : 0.049 Median : 0.0000 Median : 3.772 Median : 25.438   
## Mean : 44.423 Mean : 11.9381 Mean : 189.846 Mean : 228.503   
## 3rd Qu.: 5.603 3rd Qu.: 0.0758 3rd Qu.: 63.206 3rd Qu.: 279.322   
## Max. :1469.476 Max. :1150.7513 Max. :7749.600 Max. :12261.750   
##   
## xint restmt\_at restmt\_at\_mag restmt\_capx   
## Min. : 0.0000 Min. :0.00000 Min. : -1.4907 Min. :0.00000   
## 1st Qu.: 0.1635 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.00000   
## Median : 2.2978 Median :0.00000 Median : 0.0000 Median :0.00000   
## Mean : 95.8065 Mean :0.06322 Mean : 0.5283 Mean :0.02874   
## 3rd Qu.: 66.5683 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.00000   
## Max. :2859.7500 Max. :1.00000 Max. :182.4888 Max. :1.00000   
##   
## restmt\_capx\_mag restmt\_cogs restmt\_cogs\_mag restmt\_dltt   
## Min. :-22.71625 Min. :0.0000 Min. : -50.00 Min. :0.00000   
## 1st Qu.: 0.00000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00000 Median :0.0000 Median : 0.00 Median :0.00000   
## Mean : -0.06393 Mean :0.3046 Mean : 29.37 Mean :0.01724   
## 3rd Qu.: 0.00000 3rd Qu.:1.0000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. : 8.33350 Max. :1.0000 Max. :9299.36 Max. :1.00000   
##   
## restmt\_dltt\_mag restmt\_epsfi restmt\_epsfi\_mag restmt\_epspi   
## Min. :-26.9567 Min. :0.0000 Min. : -50.05 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.00 Median :0.0000   
## Mean : 0.1794 Mean :0.1293 Mean : 344.76 Mean :0.1351   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.00 3rd Qu.:0.0000   
## Max. :100.9780 Max. :1.0000 Max. :77081.67 Max. :1.0000   
##   
## restmt\_epspi\_mag restmt\_ib restmt\_ib\_mag restmt\_ni   
## Min. : -50.0 Min. :0.0000 Min. :-121.766 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.0000 Median : 0.000 Median :0.00000   
## Mean : 346.9 Mean :0.1121 Mean : 6.836 Mean :0.04598   
## 3rd Qu.: 0.0 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :77081.7 Max. :1.0000 Max. :2683.890 Max. :1.00000   
##   
## restmt\_ni\_mag restmt\_nopi restmt\_nopi\_mag restmt\_pi   
## Min. : -9.801 Min. :0.0000 Min. :-1868600.0 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: -92.7 1st Qu.:0.0000   
## Median : 0.000 Median :1.0000 Median : 0.0 Median :0.0000   
## Mean : 8.153 Mean :0.6322 Mean : -5817.4 Mean :0.0977   
## 3rd Qu.: 0.000 3rd Qu.:1.0000 3rd Qu.: 21.5 3rd Qu.:0.0000   
## Max. :2683.890 Max. :1.0000 Max. : 68865.1 Max. :1.0000   
##   
## restmt\_pi\_mag restmt\_reuna restmt\_reuna\_mag restmt\_seq   
## Min. :-135.28 Min. :0.00000 Min. :-2461.679 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.00 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : 6.98 Mean :0.07759 Mean : 5.001 Mean :0.09483   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :2683.89 Max. :1.00000 Max. : 4181.704 Max. :1.00000   
##   
## restmt\_seq\_mag restmt\_teq restmt\_teq\_mag restmt\_txt   
## Min. : -105.4 Min. :0.00000 Min. : -105.39 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.0 Median :0.00000 Median : 0.00 Median :0.00000   
## Mean : 49.1 Mean :0.08621 Mean : 49.71 Mean :0.07184   
## 3rd Qu.: 0.0 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :12541.8 Max. :1.00000 Max. :12541.75 Max. :1.00000   
##   
## restmt\_txt\_mag restmt\_wcap restmt\_wcap\_mag restmt\_xint   
## Min. :-88.7704 Min. :0.0000 Min. :-43.249 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0.0000   
## Mean : -0.7665 Mean :0.0431 Mean : 1.043 Mean :0.1178   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.0000   
## Max. : 47.3182 Max. :1.0000 Max. :412.500 Max. :1.0000   
##   
## restmt\_xint\_mag restmt\_xsga restmt\_xsga\_mag restmt\_dvpsp\_f  
## Min. :-62.7347 Min. :0.0000 Min. : -50.000 Min. :0   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0   
## Mean : -0.8803 Mean :0.1552 Mean : 4.984 Mean :0   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0   
## Max. : 0.5620 Max. :1.0000 Max. :1884.021 Max. :0   
##   
## restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f restmt\_dvpsx\_f\_mag  
## Min. :0 Min. :0 Min. :0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0   
## Median :0 Median :0 Median :0   
## Mean :0 Mean :0 Mean :0   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0   
## Max. :0 Max. :0 Max. :0   
##

nrow(fundamentals\_final\_ds)

## [1] 348

### Loan stocks data file

stocks\_init\_ds <- read.csv("./data/Stocks\_DS.csv", na.strings=c(""," "))  
nrow(stocks\_init\_ds)

## [1] 4187047

### From Stocks daat file choose only certain variables

### prccd ==> Price - Close - Daily

### prchd ==> Price - High - Daily

### prcld ==> Price - Low - Daily

### prcod ==> Price - Open - Daily

### trfd ==> Daily Total Return Factor

### cshtrd ==> Trading Volume - Daily

names(stocks\_init\_ds)[names(stocks\_init\_ds) == "ï..gvkey"] <- "gvkey"  
stocks\_init\_limited\_cols <- subset(stocks\_init\_ds, select = c(gvkey,cshtrd,prccd,prchd,prcld,prcod,trfd))  
stocks\_init\_limited\_cols <- stocks\_init\_limited\_cols[!is.na(stocks\_init\_limited\_cols$cshtrd)&!is.na(stocks\_init\_limited\_cols$prccd)  
 &!is.na(stocks\_init\_limited\_cols$prchd)&!is.na(stocks\_init\_limited\_cols$prcld)  
 &!is.na(stocks\_init\_limited\_cols$trfd),]  
stocks\_init\_limited\_cols$prcod[is.na(stocks\_init\_limited\_cols$prcod)] <- (stocks\_init\_limited\_cols$prchd + stocks\_init\_limited\_cols$prcld)/2  
  
stocks\_grouped\_data <- stocks\_init\_limited\_cols %>%  
 group\_by(gvkey) %>%  
 dplyr::summarize(  
 cshtrd\_m = mean(cshtrd),  
 prccd\_m = mean(prccd),  
 prchd\_m = mean(prchd),  
 prcld\_m = mean(prcld),  
 prcod\_m = mean(prcod),  
 trfd\_m = mean(trfd)  
 )

## `summarise()` ungrouping output (override with `.groups` argument)

ncol(stocks\_init\_ds)

## [1] 76

### Merge the stocks dataset with final dataset.

#fundamental\_stocks\_data <- fundamentals\_final\_ds %>%  
# inner\_join(stocks\_grouped\_data, by = 'gvkey')  
#summary(fundamental\_stocks\_data)  
  
  
for (row in 1:nrow(fundamentals\_final\_ds)){  
 row\_item\_gvkey <- as.integer(fundamentals\_final\_ds[row, "gvkey"])  
   
 specific\_stock <- stocks\_grouped\_data %>%  
 filter(gvkey == row\_item\_gvkey)   
   
 if (nrow(specific\_stock) > 0){  
 specific\_stock <- head(specific\_stock, 1)  
 fundamentals\_final\_ds$cshtrd\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$cshtrd\_m  
 fundamentals\_final\_ds$prccd\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$prccd\_m  
 fundamentals\_final\_ds$prchd\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$prchd\_m  
 fundamentals\_final\_ds$prcld\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$prcld\_m  
 fundamentals\_final\_ds$prcod\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$prcod\_m  
 fundamentals\_final\_ds$trfd\_m[final\_ds\_initial\_1$gvkey == row\_item\_gvkey] <- specific\_stock$trfd\_m  
 }  
}  
summary(fundamentals\_final\_ds)

## gvkey tic aco acominc   
## Min. : 1239 0161A : 1 Min. : 0.000 Min. :-19306.57   
## 1st Qu.: 12107 0170A : 1 1st Qu.: 0.447 1st Qu.: -30.39   
## Median : 61311 0171A : 1 Median : 8.858 Median : 0.00   
## Mean : 83018 0173A : 1 Mean : 188.577 Mean : -194.14   
## 3rd Qu.:165694 0270B : 1 3rd Qu.: 94.290 3rd Qu.: 0.00   
## Max. :277487 0563B : 1 Max. :4760.750 Max. : 3495.34   
## (Other):342   
## act ao aocidergl aocipen   
## Min. : 0.00 Min. : 0.000 Min. :-2207.250 Min. :-2803.25   
## 1st Qu.: 10.27 1st Qu.: 0.145 1st Qu.: 0.000 1st Qu.: -10.48   
## Median : 115.71 Median : 8.322 Median : 0.000 Median : 0.00   
## Mean : 1918.97 Mean : 200.854 Mean : -9.098 Mean : -91.21   
## 3rd Qu.: 1225.80 3rd Qu.: 93.865 3rd Qu.: 0.000 3rd Qu.: 0.00   
## Max. :55264.80 Max. :5330.250 Max. : 119.000 Max. : 30.75   
##   
## aoloch ap aqc at   
## Min. :-667.500 Min. : 0.00 Min. : -12.45 Min. : 0.00   
## 1st Qu.: -1.851 1st Qu.: 1.17 1st Qu.: 0.00 1st Qu.: 20.64   
## Median : 0.000 Median : 18.13 Median : 0.00 Median : 283.75   
## Mean : 6.660 Mean : 660.17 Mean : 105.33 Mean : 6489.03   
## 3rd Qu.: 1.530 3rd Qu.: 241.09 3rd Qu.: 14.95 3rd Qu.: 3172.25   
## Max. : 744.000 Max. :35222.20 Max. :5559.02 Max. :190526.20   
##   
## bkvlps caps capx ceq   
## Min. :-130515.0 Min. : -701.48 Min. : 0.000 Min. :-2342.49   
## 1st Qu.: 0.1 1st Qu.: 5.25 1st Qu.: 0.343 1st Qu.: 3.36   
## Median : 3.8 Median : 37.25 Median : 12.573 Median : 105.08   
## Mean : 11681.8 Mean : 951.28 Mean : 242.490 Mean : 2388.45   
## 3rd Qu.: 12.5 3rd Qu.: 363.48 3rd Qu.: 119.642 3rd Qu.: 1033.62   
## Max. :1881687.0 Max. :62705.25 Max. :12881.200 Max. :72640.80   
##   
## ceqt ch che   
## Min. :-40530.25 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -3.63 1st Qu.: 1.399 1st Qu.: 1.496   
## Median : 13.54 Median : 21.299 Median : 26.194   
## Mean : 54.96 Mean : 389.436 Mean : 487.347   
## 3rd Qu.: 189.37 3rd Qu.: 179.819 3rd Qu.: 217.474   
## Max. : 53931.40 Max. :10044.000 Max. :15547.750   
##   
## chech ci cogs csho   
## Min. :-305.7500 Min. : -722.617 Min. : 0.0 Min. : 0.00   
## 1st Qu.: -0.1368 1st Qu.: -1.477 1st Qu.: 12.1 1st Qu.: 16.32   
## Median : 0.5806 Median : 9.139 Median : 216.7 Median : 49.41   
## Mean : 42.2654 Mean : 475.730 Mean : 5116.2 Mean : 240.67   
## 3rd Qu.: 10.6559 3rd Qu.: 130.370 3rd Qu.: 2471.9 3rd Qu.: 142.42   
## Max. :1543.0000 Max. :16365.200 Max. :325065.8 Max. :6252.56   
##   
## cstk cstkcv dd1 dlc   
## Min. : 0.000 Min. : 0.0000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.026 1st Qu.: 0.0010 1st Qu.: 0.000 1st Qu.: 0.302   
## Median : 0.248 Median : 0.0100 Median : 1.451 Median : 5.257   
## Mean : 191.237 Mean : 0.6068 Mean : 169.076 Mean : 374.668   
## 3rd Qu.: 21.315 3rd Qu.: 0.2500 3rd Qu.: 39.533 3rd Qu.: 99.945   
## Max. :7290.750 Max. :20.8642 Max. :5428.500 Max. :15926.126   
##   
## dltt dm dn dpact   
## Min. : 0.00 Min. : 0.000 Min. : 0.0 Min. : 0.00   
## 1st Qu.: 0.16 1st Qu.: 0.000 1st Qu.: 0.0 1st Qu.: 2.02   
## Median : 17.03 Median : 1.218 Median : 0.0 Median : 52.40   
## Mean : 1477.91 Mean : 157.728 Mean : 902.6 Mean : 1303.11   
## 3rd Qu.: 902.29 3rd Qu.: 62.483 3rd Qu.: 252.3 3rd Qu.: 768.89   
## Max. :42659.60 Max. :3900.400 Max. :42561.8 Max. :50449.80   
##   
## dpc dvt ebit epspi   
## Min. : 0.000 Min. : -0.006 Min. : -208.760 Min. :-14.0200   
## 1st Qu.: 0.319 1st Qu.: 0.000 1st Qu.: -0.369 1st Qu.: -0.0512   
## Median : 7.505 Median : 0.000 Median : 23.871 Median : 0.2362   
## Mean : 170.666 Mean : 226.759 Mean : 761.483 Mean : 3.9707   
## 3rd Qu.: 95.563 3rd Qu.: 35.368 3rd Qu.: 345.869 3rd Qu.: 1.8638   
## Max. :8059.800 Max. :6572.535 Max. :24345.400 Max. :881.6400   
##   
## fiao fincf fopo   
## Min. :-3427.000 Min. :-11533.200 Min. :-389.5000   
## 1st Qu.: -7.661 1st Qu.: -50.075 1st Qu.: 0.1661   
## Median : -0.047 Median : -0.005 Median : 2.0777   
## Mean : -45.439 Mean : -332.514 Mean : 74.1085   
## 3rd Qu.: 0.000 3rd Qu.: 4.030 3rd Qu.: 23.4436   
## Max. : 1800.250 Max. : 824.184 Max. :1979.4552   
##   
## gp icapt intan intano   
## Min. : -3.19 Min. : -23.14 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 6.98 1st Qu.: 10.74 1st Qu.: 0.17 1st Qu.: 0.04   
## Median : 102.86 Median : 192.63 Median : 18.61 Median : 7.21   
## Mean : 2502.66 Mean : 4010.77 Mean : 2370.53 Mean : 878.15   
## 3rd Qu.: 1238.36 3rd Qu.: 2098.22 3rd Qu.: 715.80 3rd Qu.: 217.88   
## Max. :117445.60 Max. :119888.20 Max. :86837.75 Max. :31704.00   
##   
## invt ivncf ivst lo   
## Min. : 0.00 Min. :-13066.20 Min. : 0.000 Min. : -128.941   
## 1st Qu.: 2.19 1st Qu.: -176.92 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 37.47 Median : -19.36 Median : 0.000 Median : 5.334   
## Mean : 745.30 Mean : -314.10 Mean : 88.668 Mean : 475.282   
## 3rd Qu.: 464.24 3rd Qu.: -0.24 3rd Qu.: 2.429 3rd Qu.: 203.087   
## Max. :39770.60 Max. : 985.75 Max. :5503.750 Max. :14517.069   
##   
## lse lt ni   
## Min. : 0.00 Min. : 0.02 Min. : -737.537   
## 1st Qu.: 20.64 1st Qu.: 8.05 1st Qu.: -1.661   
## Median : 283.75 Median : 108.02 Median : 6.217   
## Mean : 6489.03 Mean : 3948.06 Mean : 499.241   
## 3rd Qu.: 3172.25 3rd Qu.: 2056.19 3rd Qu.: 125.392   
## Max. :190526.20 Max. :113297.60 Max. :17374.318   
##   
## nopi oancf opeps ppegt   
## Min. :-230.2500 Min. : -61.444 Min. : -9.8200 Min. : 0.00   
## 1st Qu.: 0.0000 1st Qu.: -0.142 1st Qu.: -0.0350 1st Qu.: 5.51   
## Median : 0.1979 Median : 20.797 Median : 0.2971 Median : 146.66   
## Mean : 47.9687 Mean : 696.038 Mean : 3.9977 Mean : 2919.87   
## 3rd Qu.: 4.8163 3rd Qu.: 282.996 3rd Qu.: 1.8725 3rd Qu.: 1576.40   
## Max. :2224.4000 Max. :24599.000 Max. :856.8325 Max. :161869.20   
##   
## re reajo rect recta   
## Min. :-7570.29 Min. :-7860.75 Min. : 0.000 Min. :-19466.259   
## 1st Qu.: -9.35 1st Qu.: -19.79 1st Qu.: 1.636 1st Qu.: -0.169   
## Median : 19.16 Median : 0.00 Median : 28.478 Median : 0.000   
## Mean : 1909.92 Mean : -78.00 Mean : 510.810 Mean : -96.126   
## 3rd Qu.: 441.34 3rd Qu.: 0.00 3rd Qu.: 312.676 3rd Qu.: 0.060   
## Max. :68884.60 Max. : 7171.53 Max. :15020.067 Max. : 1946.250   
##   
## revt siv spi sppiv   
## Min. : 0.0 Min. : 0.000 Min. :-921.2962 Min. :-6191.874   
## 1st Qu.: 22.4 1st Qu.: 0.000 1st Qu.: -15.8642 1st Qu.: -0.046   
## Median : 333.1 Median : 0.000 Median : -0.4417 Median : 0.000   
## Mean : 7618.9 Mean : 56.788 Mean : -35.2648 Mean : -32.516   
## 3rd Qu.: 3826.2 3rd Qu.: 0.512 3rd Qu.: 0.0000 3rd Qu.: 0.007   
## Max. :442511.4 Max. :4366.827 Max. :1115.5000 Max. : 27.017   
##   
## sstk teq tstk tstkn   
## Min. : 0.0000 Min. :-2208.96 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 0.0006 1st Qu.: 4.19 1st Qu.: 0.0 1st Qu.: 0.000   
## Median : 1.2209 Median : 106.70 Median : 0.0 Median : 0.000   
## Mean : 36.8987 Mean : 2524.10 Mean : 889.8 Mean : 27.024   
## 3rd Qu.: 14.5968 3rd Qu.: 1095.88 3rd Qu.: 8.5 3rd Qu.: 1.282   
## Max. :1513.0000 Max. :76602.80 Max. :67539.2 Max. :1923.500   
##   
## txp txr txt wcap   
## Min. : -0.252 Min. : 0.0000 Min. : -76.388 Min. :-8236.800   
## 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: -0.011   
## Median : 0.049 Median : 0.0000 Median : 3.772 Median : 25.438   
## Mean : 44.423 Mean : 11.9381 Mean : 189.846 Mean : 228.503   
## 3rd Qu.: 5.603 3rd Qu.: 0.0758 3rd Qu.: 63.206 3rd Qu.: 279.322   
## Max. :1469.476 Max. :1150.7513 Max. :7749.600 Max. :12261.750   
##   
## xint restmt\_at restmt\_at\_mag restmt\_capx   
## Min. : 0.0000 Min. :0.00000 Min. : -1.4907 Min. :0.00000   
## 1st Qu.: 0.1635 1st Qu.:0.00000 1st Qu.: 0.0000 1st Qu.:0.00000   
## Median : 2.2978 Median :0.00000 Median : 0.0000 Median :0.00000   
## Mean : 95.8065 Mean :0.06322 Mean : 0.5283 Mean :0.02874   
## 3rd Qu.: 66.5683 3rd Qu.:0.00000 3rd Qu.: 0.0000 3rd Qu.:0.00000   
## Max. :2859.7500 Max. :1.00000 Max. :182.4888 Max. :1.00000   
##   
## restmt\_capx\_mag restmt\_cogs restmt\_cogs\_mag restmt\_dltt   
## Min. :-22.71625 Min. :0.0000 Min. : -50.00 Min. :0.00000   
## 1st Qu.: 0.00000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.00000 Median :0.0000 Median : 0.00 Median :0.00000   
## Mean : -0.06393 Mean :0.3046 Mean : 29.37 Mean :0.01724   
## 3rd Qu.: 0.00000 3rd Qu.:1.0000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. : 8.33350 Max. :1.0000 Max. :9299.36 Max. :1.00000   
##   
## restmt\_dltt\_mag restmt\_epsfi restmt\_epsfi\_mag restmt\_epspi   
## Min. :-26.9567 Min. :0.0000 Min. : -50.05 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.00 Median :0.0000   
## Mean : 0.1794 Mean :0.1293 Mean : 344.76 Mean :0.1351   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.00 3rd Qu.:0.0000   
## Max. :100.9780 Max. :1.0000 Max. :77081.67 Max. :1.0000   
##   
## restmt\_epspi\_mag restmt\_ib restmt\_ib\_mag restmt\_ni   
## Min. : -50.0 Min. :0.0000 Min. :-121.766 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.0 Median :0.0000 Median : 0.000 Median :0.00000   
## Mean : 346.9 Mean :0.1121 Mean : 6.836 Mean :0.04598   
## 3rd Qu.: 0.0 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :77081.7 Max. :1.0000 Max. :2683.890 Max. :1.00000   
##   
## restmt\_ni\_mag restmt\_nopi restmt\_nopi\_mag restmt\_pi   
## Min. : -9.801 Min. :0.0000 Min. :-1868600.0 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: -92.7 1st Qu.:0.0000   
## Median : 0.000 Median :1.0000 Median : 0.0 Median :0.0000   
## Mean : 8.153 Mean :0.6322 Mean : -5817.4 Mean :0.0977   
## 3rd Qu.: 0.000 3rd Qu.:1.0000 3rd Qu.: 21.5 3rd Qu.:0.0000   
## Max. :2683.890 Max. :1.0000 Max. : 68865.1 Max. :1.0000   
##   
## restmt\_pi\_mag restmt\_reuna restmt\_reuna\_mag restmt\_seq   
## Min. :-135.28 Min. :0.00000 Min. :-2461.679 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.00000 1st Qu.: 0.000 1st Qu.:0.00000   
## Median : 0.00 Median :0.00000 Median : 0.000 Median :0.00000   
## Mean : 6.98 Mean :0.07759 Mean : 5.001 Mean :0.09483   
## 3rd Qu.: 0.00 3rd Qu.:0.00000 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :2683.89 Max. :1.00000 Max. : 4181.704 Max. :1.00000   
##   
## restmt\_seq\_mag restmt\_teq restmt\_teq\_mag restmt\_txt   
## Min. : -105.4 Min. :0.00000 Min. : -105.39 Min. :0.00000   
## 1st Qu.: 0.0 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.:0.00000   
## Median : 0.0 Median :0.00000 Median : 0.00 Median :0.00000   
## Mean : 49.1 Mean :0.08621 Mean : 49.71 Mean :0.07184   
## 3rd Qu.: 0.0 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.:0.00000   
## Max. :12541.8 Max. :1.00000 Max. :12541.75 Max. :1.00000   
##   
## restmt\_txt\_mag restmt\_wcap restmt\_wcap\_mag restmt\_xint   
## Min. :-88.7704 Min. :0.0000 Min. :-43.249 Min. :0.0000   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0.0000   
## Mean : -0.7665 Mean :0.0431 Mean : 1.043 Mean :0.1178   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0.0000   
## Max. : 47.3182 Max. :1.0000 Max. :412.500 Max. :1.0000   
##   
## restmt\_xint\_mag restmt\_xsga restmt\_xsga\_mag restmt\_dvpsp\_f  
## Min. :-62.7347 Min. :0.0000 Min. : -50.000 Min. :0   
## 1st Qu.: 0.0000 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.:0   
## Median : 0.0000 Median :0.0000 Median : 0.000 Median :0   
## Mean : -0.8803 Mean :0.1552 Mean : 4.984 Mean :0   
## 3rd Qu.: 0.0000 3rd Qu.:0.0000 3rd Qu.: 0.000 3rd Qu.:0   
## Max. : 0.5620 Max. :1.0000 Max. :1884.021 Max. :0   
##   
## restmt\_dvpsp\_f\_mag restmt\_dvpsx\_f restmt\_dvpsx\_f\_mag cshtrd\_m   
## Min. :0 Min. :0 Min. :0 Min. : 0   
## 1st Qu.:0 1st Qu.:0 1st Qu.:0 1st Qu.: 18833   
## Median :0 Median :0 Median :0 Median : 116999   
## Mean :0 Mean :0 Mean :0 Mean : 907451   
## 3rd Qu.:0 3rd Qu.:0 3rd Qu.:0 3rd Qu.: 614817   
## Max. :0 Max. :0 Max. :0 Max. :13129451   
## NA's :15   
## prccd\_m prchd\_m prcld\_m   
## Min. : 0.0018 Min. : 0.0019 Min. : 0.0016   
## 1st Qu.: 1.1706 1st Qu.: 1.3154 1st Qu.: 1.1278   
## Median : 9.0173 Median : 9.2103 Median : 8.8327   
## Mean : 30.6045 Mean : 31.2905 Mean : 29.9726   
## 3rd Qu.: 32.9362 3rd Qu.: 33.1643 3rd Qu.: 32.6779   
## Max. :2217.8253 Max. :2250.8331 Max. :2183.9480   
## NA's :15 NA's :15 NA's :15   
## prcod\_m trfd\_m   
## Min. : 0.0362 Min. : 1.000   
## 1st Qu.: 3.3712 1st Qu.: 1.063   
## Median : 12.3601 Median : 1.225   
## Mean : 36.4436 Mean : 2.668   
## 3rd Qu.: 36.3218 3rd Qu.: 1.856   
## Max. :2217.8208 Max. :218.416   
## NA's :15 NA's :15

nrow(fundamentals\_final\_ds)

## [1] 348

fundamental\_stocks\_data <- fundamentals\_final\_ds[!is.na(fundamentals\_final\_ds$cshtrd\_m) &!is.na(fundamentals\_final\_ds$prccd\_m)  
 &!is.na(fundamentals\_final\_ds$prchd\_m) &!is.na(fundamentals\_final\_ds$prcld\_m)  
 &!is.na(fundamentals\_final\_ds$prcod\_m) &!is.na(fundamentals\_final\_ds$trfd\_m),]

# cor\_matrix\_ds <- subset(fundamental\_stocks\_data, select = -c(gvkey,tic))  
# cor\_matrix <- cor(cor\_matrix\_ds)  
# cor\_matrix %>%  
# as.data.frame() %>%  
# mutate(var1 = rownames(.)) %>%  
# gather(var2, value, -var1) %>%  
# arrange(desc(value)) %>%  
# group\_by(value) %>%  
# filter(row\_number()==1)

fundamental\_stocks\_data\_final <- fundamental\_stocks\_data  
#summary(fundamental\_stocks\_data\_final)  
nrow(fundamental\_stocks\_data\_final)

## [1] 333

### Load securities dataset

securities\_init\_ds <- read.csv("./data/Securities\_DS.csv", na.strings=c(""," "))  
names(securities\_init\_ds)[names(securities\_init\_ds) == "ï..gvkey"] <- "gvkey"  
ncol(securities\_init\_ds)

## [1] 55

securities\_init\_ds\_1 <- subset(securities\_init\_ds, select = -c(iid,isalrt,primiss,ajexm,  
 spgim,spiim,spmim,cheqvm,curcddvm,dvpsxm,  
 sphcusip,sphiid, sphmid,sphname,sphsec,sphtic,sphvg,sph100,  
 cyear,mkvalincl,exchg,tpci,city,  
 conml,costat,ggroup,gind, gsubind,loc,naics,sic,state, curcdm,   
 navm,adrrm,rawpm,rawxm,cshoq,csfsm,  
 datadate,tic,conm,cmth  
 ))  
  
  
#summary(securities\_init\_ds\_1)

### Choose following variables from the securities dataset

### trfm ==> Monthly Total Return Factor

### dvrate ==> Dividend Rate - Monthly

fund\_stock\_securities\_ds <- fundamental\_stocks\_data\_final   
securities\_init\_ds\_2 <- securities\_init\_ds\_1 %>%  
 filter(!is.na(trfm) & !is.na(trt1m)) %>%  
 group\_by(gvkey) %>%  
 dplyr::summarize(  
 trfm\_m = mean(trfm)  
 )  
  
fund\_stock\_securities\_ds$trfm\_m <- NA  
for (row in 1:nrow(fund\_stock\_securities\_ds)){  
 row\_item\_gvkey <- as.integer(fund\_stock\_securities\_ds[row, "gvkey"])  
 specific\_security <- securities\_init\_ds\_2 %>%  
 filter(gvkey == row\_item\_gvkey)   
 if (nrow(specific\_security) > 0){  
 security\_row <- head(specific\_security, 1)  
 trfm\_m <- as.numeric(security\_row$trfm\_m)   
 fund\_stock\_securities\_ds$trfm\_m[fund\_stock\_securities\_ds$gvkey == row\_item\_gvkey] <- trfm\_m  
 }  
}  
  
  
  
securities\_init\_ds\_3 <- securities\_init\_ds\_1 %>%  
 filter(!is.na(dvrate)) %>%  
 group\_by(gvkey) %>%  
 dplyr::summarize(  
 dvrate\_m = mean(dvrate)  
 )  
  
fund\_stock\_securities\_ds$dvrate\_m <- NA  
for (row in 1:nrow(fund\_stock\_securities\_ds)){  
 row\_item\_gvkey <- as.integer(fund\_stock\_securities\_ds[row, "gvkey"])  
 specific\_security <- securities\_init\_ds\_3 %>%  
 filter(gvkey == row\_item\_gvkey)   
 if (nrow(specific\_security) > 0){  
 security\_row <- head(specific\_security, 1)  
 dvrate\_m <- as.numeric(security\_row$dvrate\_m)   
 fund\_stock\_securities\_ds$dvrate\_m[fund\_stock\_securities\_ds$gvkey == row\_item\_gvkey] <- dvrate\_m  
 }  
}  
  
  
#summary(fund\_stock\_securities\_ds)

### Load ratings dataset.

### Load this variable splticrm ==> S&P Domestic Long Term Issuer Credit Rating

### splticrm - This is categorical variable with unique ratings, each of the rating are given numeric values.

### Highest rating gets high numeric values and as rating decreases the numerica value assigned decreases.

ratings\_init\_ds <- read.csv("./data/Ratings\_DS.csv", na.strings=c("", " "))  
names(ratings\_init\_ds)[names(ratings\_init\_ds) == "ï..gvkey"] <- "gvkey"  
ratings\_init\_ds$datadate <- as.Date(ratings\_init\_ds$datadate, "%m/%d/%Y")  
ratings\_init\_ds$splticrm = factor(ratings\_init\_ds$splticrm, levels=c(levels(ratings\_init\_ds$splticrm), "NR"))  
ratings\_init\_ds$splticrm[is.na(ratings\_init\_ds$splticrm)] = "NR"  
  
ratings\_init\_ds$splticrm\_num\_value <- 0  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "AAA"] <- 100  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "AA"] <- 90  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "AA-"] <- 85  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "A+"] <- 80  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "A"] <- 75  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "A-"] <- 70  
  
  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BBB+"] <- 65  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BBB"] <- 60  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BBB-"] <- 55  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BB+"] <- 50  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BB"] <- 45  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "BB-"] <- 40  
  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "B+"] <- 35  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "B"] <- 30  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "B-"] <- 25  
  
  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "CCC+"] <- 20  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "CCC"] <- 19  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "CCC-"] <- 18  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "CC"] <- 17  
  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "D"] <- 10  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "SD"] <- 10  
ratings\_init\_ds$splticrm\_num\_value[ratings\_init\_ds$splticrm == "NR"] <- 0  
ratings\_init\_ds$splticrm\_num\_value <- factor(ratings\_init\_ds$splticrm\_num\_value)  
#levels(ratings\_init\_ds$splticrm)  
#str(ratings\_init\_ds)

### Merge the ratings dataset and fundamentals dataset, by assigning the rating from ratings dataset to each company using gvkey

### Along with rating, additional variable is added which indicates whether the rating has

### 1. Increased

### 2. Decreased

### 3. NOCHANGE

fund\_stock\_securities\_rating\_ds <- fund\_stock\_securities\_ds #%>%  
 #filter(gvkey == 1078)  
fund\_stock\_securities\_rating\_ds$sp\_rating <- "NOTRATED"   
rated\_companies <- ratings\_init\_ds %>%  
 filter(splticrm != "NR")  
  
for (row in 1:nrow(fund\_stock\_securities\_rating\_ds)){  
 row\_item\_gvkey <- as.integer(fund\_stock\_securities\_rating\_ds[row, "gvkey"])  
  
 specific\_rating <- rated\_companies %>%  
 filter(gvkey == row\_item\_gvkey) %>%  
 arrange(datadate)  
 if (nrow(specific\_rating) > 0){  
 first\_row <- head(specific\_rating, 1)  
 last\_row <- tail(specific\_rating, 1)  
 start\_value <- as.integer(first\_row$splticrm\_num\_value)  
 end\_value <- as.integer(last\_row$splticrm\_num\_value)  
 if (start\_value == end\_value){  
 fund\_stock\_securities\_rating\_ds$sp\_rating[fund\_stock\_securities\_rating\_ds$gvkey == row\_item\_gvkey] <- "NoCHANGE"  
 }else if (start\_value < end\_value){  
 fund\_stock\_securities\_rating\_ds$sp\_rating[fund\_stock\_securities\_rating\_ds$gvkey == row\_item\_gvkey] <- "INCREASED"  
 }else if (start\_value > end\_value){  
 fund\_stock\_securities\_rating\_ds$sp\_rating[fund\_stock\_securities\_rating\_ds$gvkey == row\_item\_gvkey] <- "DECREASED"  
 }  
 }  
}  
fund\_stock\_securities\_rating\_ds$sp\_rating <- factor(fund\_stock\_securities\_rating\_ds$sp\_rating)  
#summary(fund\_stock\_securities\_rating\_ds)

### Load Sca filing dataset for the target variable

sca\_fillings\_ds <- read.csv("./data/SCA\_Filings\_and\_Settlements.csv", na.strings=c(""," "))  
sca\_fillings\_ds$SettlementAmount = gsub("\\$", "", sca\_fillings\_ds$SettlementAmount)  
sca\_fillings\_ds$SettlementAmount = as.numeric(gsub("\\,", "", sca\_fillings\_ds$SettlementAmount))  
#summary(sca\_fillings\_ds)  
ncol(sca\_fillings\_ds)

## [1] 6

### Add the target variable litigated to the dataset under analysis.

### From each of the conany record identify if has any sca filing from the sca fileing dataset,if entry exists then mark for that company lititgated = true or else if the record ### does not exist in sca filing for that particular comapny then mark litigate attribute as false.

### Also in case of litigation indentify if there is any settlement amount and add same to main dataset. In case if multiple settlement amount exists for particular company then ### take the maximum settlement amount.

fund\_stock\_securities\_rating\_ds$litigated <- 0  
fund\_stock\_securities\_rating\_ds$litigation\_settlement <- NA  
fund\_stock\_securities\_rating\_ds\_1 <- fund\_stock\_securities\_rating\_ds   
  
for (row in 1:nrow(fund\_stock\_securities\_rating\_ds\_1)){  
 row\_item\_tic <- lapply(fund\_stock\_securities\_rating\_ds\_1[row, "tic"], as.character)  
 row\_item\_gvkey <- as.integer(fund\_stock\_securities\_rating\_ds[row, "gvkey"])  
 specific\_sca\_filings <- sca\_fillings\_ds %>%  
 filter(Ticker == row\_item\_tic)   
 if (nrow(specific\_sca\_filings) > 0){  
 fund\_stock\_securities\_rating\_ds$litigated[fund\_stock\_securities\_rating\_ds$gvkey == row\_item\_gvkey] <- 1  
 specific\_sca\_filings\_max <- specific\_sca\_filings %>%  
 filter(!is.na(SettlementAmount)) %>%  
 arrange(SettlementAmount)  
   
 if (nrow(specific\_sca\_filings\_max) > 0){  
 last\_row <- tail(specific\_sca\_filings\_max, 1)  
 settlement\_amount <- as.numeric(last\_row$SettlementAmount)  
 fund\_stock\_securities\_rating\_ds$litigation\_settlement[fund\_stock\_securities\_rating\_ds$gvkey == row\_item\_gvkey] <- settlement\_amount  
 }  
 }  
}  
fund\_stock\_securities\_rating\_ds$litigated <- as.factor(fund\_stock\_securities\_rating\_ds$litigated)  
#summary(fund\_stock\_securities\_rating\_ds)

#colnames(fund\_stock\_securities\_rating\_ds)

### Based on manual review. idetify the columns that need to be removed from the final dataset.

fund\_stock\_securities\_rating\_ds\_final <- subset(fund\_stock\_securities\_rating\_ds, select = -c(aocidergl, aocipen, ceqt, cstkcv, dd1,  
 dpc,icapt, intan, intano, ivncf, ivst,  
 lo, lse, opeps, reajo, recta,  
 spi, tstkn, txp, txr,  
 restmt\_epsfi\_mag, restmt\_epsfi,  
 restmt\_pi, restmt\_pi\_mag,  
 restmt\_seq, restmt\_seq\_mag,  
 restmt\_xsga, restmt\_xsga\_mag,  
 restmt\_dvpsp\_f, restmt\_dvpsp\_f\_mag,   
 restmt\_dvpsx\_f, restmt\_dvpsx\_f\_mag  
 ))  
#summary(fund\_stock\_securities\_rating\_ds\_final)

# cor\_matrix\_ds <- subset(fund\_stock\_securities\_rating\_ds\_final, select = -c(gvkey,tic, sp\_rating, litigated))  
# cor\_matrix <- cor(cor\_matrix\_ds)  
# cor\_matrix %>%  
# as.data.frame() %>%  
# mutate(var1 = rownames(.)) %>%  
# gather(var2, value, -var1) %>%  
# arrange(desc(value)) %>%  
# group\_by(value) %>%  
# filter(row\_number() == 1)

# DE ==> Debt to equity ratio

# wc ==> Working capital ratio

# pe ==> Pricing to earning ratio

# ROE ==> Return on Equity

fund\_stock\_securities\_rating\_ds\_final$epspi[fund\_stock\_securities\_rating\_ds\_final$epspi == 0] <- 0.000001  
fund\_stock\_securities\_rating\_ds\_final$lt[fund\_stock\_securities\_rating\_ds\_final$lt == 0] <- 0.000001  
fund\_stock\_securities\_rating\_ds\_final$teq[fund\_stock\_securities\_rating\_ds\_final$teq == 0] <- 0.000001  
fund\_stock\_securities\_rating\_ds\_final$pe\_ratio <- fund\_stock\_securities\_rating\_ds\_final$prccd\_m/fund\_stock\_securities\_rating\_ds\_final$epspi  
fund\_stock\_securities\_rating\_ds\_final$wc\_ratio <- fund\_stock\_securities\_rating\_ds\_final$act/fund\_stock\_securities\_rating\_ds\_final$lt  
fund\_stock\_securities\_rating\_ds\_final$de\_ratio <- fund\_stock\_securities\_rating\_ds\_final$lt/fund\_stock\_securities\_rating\_ds\_final$teq  
fund\_stock\_securities\_rating\_ds\_final$roe\_ratio <- fund\_stock\_securities\_rating\_ds\_final$ni/fund\_stock\_securities\_rating\_ds\_final$teq  
#temp <- subset(fund\_stock\_securities\_rating\_ds\_final, select = c(gvkey,tic, epspi, prccd\_m,pe\_ratio, act, lt,wc\_ratio, teq,de\_ratio, ni, roe\_ratio))  
#head(temp)  
trfm\_median <- median(as.numeric(fund\_stock\_securities\_rating\_ds\_final$trfm\_m),na.rm=TRUE)  
fund\_stock\_securities\_rating\_ds\_final$trfm\_m[is.na(fund\_stock\_securities\_rating\_ds\_final$trfm\_m)] <- trfm\_median

### From the final dataset identify correlated variables

cor\_matrix\_ds <- subset(fund\_stock\_securities\_rating\_ds\_final, select = -c(gvkey,tic, sp\_rating, litigated))  
cor\_matrix <- cor(cor\_matrix\_ds)  
cor\_matrix %>%  
 as.data.frame() %>%  
 mutate(var1 = rownames(.)) %>%  
 gather(var2, value, -var1) %>%  
 arrange(desc(value)) %>%  
 group\_by(value) %>%  
 filter(row\_number() == 1)

## # A tibble: 3,572 x 3  
## # Groups: value [3,572]  
## var1 var2 value  
## <chr> <chr> <dbl>  
## 1 aco aco 1   
## 2 prchd\_m prccd\_m 1.00   
## 3 prcld\_m prccd\_m 1.00   
## 4 prcld\_m prchd\_m 0.998  
## 5 restmt\_ni\_mag restmt\_ib\_mag 0.998  
## 6 teq ceq 0.996  
## 7 ni ci 0.994  
## 8 oancf ebit 0.988  
## 9 revt cogs 0.988  
## 10 ppegt capx 0.988  
## # ... with 3,562 more rows

#### Remove Highly correlated variables, in general if correlation value is > 0.9 and check the correlation matrix values

cor\_matrix\_ds <- subset(fund\_stock\_securities\_rating\_ds\_final, select = -c(gvkey,tic, sp\_rating, litigated, prchd\_m, prcld\_m, ni, restmt\_ib\_mag, ceq, oancf, cogs, ppegt,  
 lt, ci, restmt\_dltt\_mag, invt, che, ap, at, xint, gp, act, txt, capx,  
 dm, dn, dpact,fiao,fincf,sppiv,fopo)) #latest removed variables  
cor\_matrix <- cor(cor\_matrix\_ds)  
cor\_matrix %>%  
 as.data.frame() %>%  
 mutate(var1 = rownames(.)) %>%  
 gather(var2, value, -var1) %>%  
 arrange(desc(value)) %>%  
 group\_by(value) %>%  
 filter(row\_number() == 1)

## # A tibble: 1,655 x 3  
## # Groups: value [1,655]  
## var1 var2 value  
## <chr> <chr> <dbl>  
## 1 aco aco 1   
## 2 prccd\_m epspi 0.945  
## 3 ebit dvt 0.945  
## 4 restmt\_teq\_mag restmt\_ni\_mag 0.934  
## 5 teq ebit 0.892  
## 6 ebit dltt 0.892  
## 7 teq ao 0.879  
## 8 prcod\_m prccd\_m 0.879  
## 9 ebit ao 0.874  
## 10 re ebit 0.859  
## # ... with 1,645 more rows

### Remove Highly correlated variables from above analysis.

fund\_stock\_securities\_rating\_ds\_final\_10 <- subset(fund\_stock\_securities\_rating\_ds\_final, select = -c(prchd\_m, prcld\_m, ni, restmt\_ib\_mag,   
 ceq, oancf, cogs, ppegt,  
 lt, ci, restmt\_dltt\_mag, invt, che,   
 ap, at, xint, gp, act, txt, capx,  
 dm, dn, dpact,fiao,fincf,sppiv,fopo)) #latest removed variables  
#nrow(fund\_stock\_securities\_rating\_ds\_final)  
#summary(fund\_stock\_securities\_rating\_ds\_final)

### Convert all categorical variables to factor

ds\_final <- fund\_stock\_securities\_rating\_ds\_final\_10  
ds\_final <- subset(ds\_final, select = -c(dvrate\_m)) #, litigation\_settlement))  
ds\_final$restmt\_at <- as.factor(ds\_final$restmt\_at)  
ds\_final$restmt\_capx <- as.factor(ds\_final$restmt\_capx)  
ds\_final$restmt\_cogs <- as.factor(ds\_final$restmt\_cogs)  
ds\_final$restmt\_dltt <- as.factor(ds\_final$restmt\_dltt)  
ds\_final$restmt\_epspi <- as.factor(ds\_final$restmt\_epspi)  
ds\_final$restmt\_ib <- as.factor(ds\_final$restmt\_ib)  
ds\_final$restmt\_ni <- as.factor(ds\_final$restmt\_ni)  
ds\_final$restmt\_nopi <- as.factor(ds\_final$restmt\_nopi)  
ds\_final$restmt\_reuna <- as.factor(ds\_final$restmt\_reuna)  
ds\_final$restmt\_teq <- as.factor(ds\_final$restmt\_teq)  
ds\_final$restmt\_txt <- as.factor(ds\_final$restmt\_txt)  
ds\_final$restmt\_wcap <- as.factor(ds\_final$restmt\_wcap)  
ds\_final$restmt\_xint <- as.factor(ds\_final$restmt\_xint)  
ds\_final$sp\_rating <- as.factor(ds\_final$sp\_rating)  
ds\_final$litigated <- as.factor(ds\_final$litigated)

### Perform target ecoding to all the categorical variables

# split the data into training and (held-out) test sets  
training\_ind <- createDataPartition(ds\_final$litigated,  
p = 0.75,  
list = FALSE,  
times = 1)  
training\_set <- ds\_final[training\_ind, ]  
test\_set <- ds\_final[-training\_ind, ]  
  
nrow(training\_set)

## [1] 250

nrow(test\_set)

## [1] 83

threshold <- 250  
#head(training\_set$litigated)  
  
threshold <- 250  
target\_enc\_train <- function(variable, level) {  
 training\_set$litigated <- as.numeric(as.vector(training\_set$litigated))  
 train\_avg\_target <- colMeans(training\_set[, "litigated"])  
 if (nrow(training\_set[training\_set[, variable]==level, ])==0) {  
 return(train\_avg\_target)  
 } else {  
 level\_num\_obs <- nrow(training\_set[training\_set[, variable]==level,])  
 level\_avg\_target <- colMeans(training\_set[training\_set[, variable]==level, "litigated"])  
 return((level\_num\_obs\*level\_avg\_target+threshold\*train\_avg\_target)/(level\_num\_obs+threshold))  
 }  
}  
  
sp\_rating\_target <- mapply(target\_enc\_train, variable = "sp\_rating", level = levels(training\_set$sp\_rating), USE.NAMES = FALSE)  
names(sp\_rating\_target) <- levels(training\_set$sp\_rating)  
training\_set$sp\_rating\_target <- 0  
for (level in levels(training\_set$sp\_rating)) {  
 training\_set[training\_set[, "sp\_rating"]==level, "sp\_rating\_target"] <- sp\_rating\_target[level]  
}  
  
test\_set$sp\_rating\_target <- 0  
for (level in levels(training\_set$sp\_rating)) {  
 test\_set[test\_set[, "sp\_rating"]==level, "sp\_rating\_target"] <- sp\_rating\_target[level]  
}  
  
  
restmt\_at\_target <- mapply(target\_enc\_train, variable = "restmt\_at", level = levels(training\_set$restmt\_at), USE.NAMES = FALSE)  
names(restmt\_at\_target) <- levels(training\_set$restmt\_at)  
training\_set$restmt\_at\_target <- 0  
for (level in levels(training\_set$restmt\_at)) {  
 training\_set[training\_set[, "restmt\_at"]==level, "restmt\_at\_target"] <- restmt\_at\_target[level]  
}  
  
test\_set$restmt\_at\_target <- 0  
for (level in levels(training\_set$restmt\_at)) {  
 test\_set[test\_set[, "restmt\_at"]==level, "restmt\_at\_target"] <- restmt\_at\_target[level]  
}  
  
restmt\_capx\_target <- mapply(target\_enc\_train, variable = "restmt\_capx", level = levels(training\_set$restmt\_capx), USE.NAMES = FALSE)  
names(restmt\_capx\_target) <- levels(training\_set$restmt\_capx)  
training\_set$restmt\_capx\_target <- 0  
for (level in levels(training\_set$restmt\_capx)) {  
 training\_set[training\_set[, "restmt\_capx"]==level, "restmt\_capx\_target"] <- restmt\_capx\_target[level]  
}  
  
test\_set$restmt\_capx\_target <- 0  
for (level in levels(training\_set$restmt\_capx)) {  
 test\_set[test\_set[, "restmt\_capx"]==level, "restmt\_capx\_target"] <- restmt\_capx\_target[level]  
}  
  
restmt\_cogs\_target <- mapply(target\_enc\_train, variable = "restmt\_cogs", level = levels(training\_set$restmt\_cogs), USE.NAMES = FALSE)  
names(restmt\_cogs\_target) <- levels(training\_set$restmt\_cogs)  
training\_set$restmt\_cogs\_target <- 0  
for (level in levels(training\_set$restmt\_cogs)) {  
 training\_set[training\_set[, "restmt\_cogs"]==level, "restmt\_cogs\_target"] <- restmt\_cogs\_target[level]  
}  
  
test\_set$restmt\_cogs\_target <- 0  
for (level in levels(training\_set$restmt\_cogs)) {  
 test\_set[test\_set[, "restmt\_cogs"]==level, "restmt\_cogs\_target"] <- restmt\_cogs\_target[level]  
}  
  
restmt\_dltt\_target <- mapply(target\_enc\_train, variable = "restmt\_dltt", level = levels(training\_set$restmt\_dltt), USE.NAMES = FALSE)  
names(restmt\_dltt\_target) <- levels(training\_set$restmt\_dltt)  
training\_set$restmt\_dltt\_target <- 0  
for (level in levels(training\_set$restmt\_dltt)) {  
 training\_set[training\_set[, "restmt\_dltt"]==level, "restmt\_dltt\_target"] <- restmt\_dltt\_target[level]  
}  
  
test\_set$restmt\_dltt\_target <- 0  
for (level in levels(training\_set$restmt\_dltt)) {  
 test\_set[test\_set[, "restmt\_dltt"]==level, "restmt\_dltt\_target"] <- restmt\_dltt\_target[level]  
}  
  
restmt\_epspi\_target <- mapply(target\_enc\_train, variable = "restmt\_epspi", level = levels(training\_set$restmt\_epspi), USE.NAMES = FALSE)  
names(restmt\_epspi\_target) <- levels(training\_set$restmt\_epspi)  
training\_set$restmt\_epspi\_target <- 0  
for (level in levels(training\_set$restmt\_epspi)) {  
 training\_set[training\_set[, "restmt\_epspi"]==level, "restmt\_epspi\_target"] <- restmt\_epspi\_target[level]  
}  
  
test\_set$restmt\_epspi\_target <- 0  
for (level in levels(training\_set$restmt\_epspi)) {  
 test\_set[test\_set[, "restmt\_epspi"]==level, "restmt\_epspi\_target"] <- restmt\_epspi\_target[level]  
}  
  
restmt\_ib\_target <- mapply(target\_enc\_train, variable = "restmt\_ib", level = levels(training\_set$restmt\_ib), USE.NAMES = FALSE)  
names(restmt\_ib\_target) <- levels(training\_set$restmt\_ib)  
training\_set$restmt\_ib\_target <- 0  
for (level in levels(training\_set$restmt\_ib)) {  
 training\_set[training\_set[, "restmt\_ib"]==level, "restmt\_ib\_target"] <- restmt\_ib\_target[level]  
}  
  
test\_set$restmt\_ib\_target <- 0  
for (level in levels(training\_set$restmt\_ib)) {  
 test\_set[test\_set[, "restmt\_ib"]==level, "restmt\_ib\_target"] <- restmt\_ib\_target[level]  
}  
  
  
restmt\_ni\_target <- mapply(target\_enc\_train, variable = "restmt\_ni", level = levels(training\_set$restmt\_ni), USE.NAMES = FALSE)  
names(restmt\_ni\_target) <- levels(training\_set$restmt\_ni)  
training\_set$restmt\_ni\_target <- 0  
for (level in levels(training\_set$restmt\_ni)) {  
 training\_set[training\_set[, "restmt\_ni"]==level, "restmt\_ni\_target"] <- restmt\_ni\_target[level]  
}  
  
test\_set$restmt\_ni\_target <- 0  
for (level in levels(training\_set$restmt\_ni)) {  
 test\_set[test\_set[, "restmt\_ni"]==level, "restmt\_ni\_target"] <- restmt\_ni\_target[level]  
}  
  
restmt\_nopi\_target <- mapply(target\_enc\_train, variable = "restmt\_nopi", level = levels(training\_set$restmt\_nopi), USE.NAMES = FALSE)  
names(restmt\_nopi\_target) <- levels(training\_set$restmt\_nopi)  
training\_set$restmt\_nopi\_target <- 0  
for (level in levels(training\_set$restmt\_nopi)) {  
 training\_set[training\_set[, "restmt\_nopi"]==level, "restmt\_nopi\_target"] <- restmt\_nopi\_target[level]  
}  
  
test\_set$restmt\_nopi\_target <- 0  
for (level in levels(training\_set$restmt\_nopi)) {  
 test\_set[test\_set[, "restmt\_nopi"]==level, "restmt\_nopi\_target"] <- restmt\_nopi\_target[level]  
}  
  
  
restmt\_reuna\_target <- mapply(target\_enc\_train, variable = "restmt\_reuna", level = levels(training\_set$restmt\_reuna), USE.NAMES = FALSE)  
names(restmt\_reuna\_target) <- levels(training\_set$restmt\_reuna)  
training\_set$restmt\_reuna\_target <- 0  
for (level in levels(training\_set$restmt\_reuna)) {  
 training\_set[training\_set[, "restmt\_reuna"]==level, "restmt\_reuna\_target"] <- restmt\_reuna\_target[level]  
}  
  
test\_set$restmt\_reuna\_target <- 0  
for (level in levels(training\_set$restmt\_reuna)) {  
 test\_set[test\_set[, "restmt\_reuna"]==level, "restmt\_reuna\_target"] <- restmt\_reuna\_target[level]  
}  
  
restmt\_teq\_target <- mapply(target\_enc\_train, variable = "restmt\_teq", level = levels(training\_set$restmt\_teq), USE.NAMES = FALSE)  
names(restmt\_teq\_target) <- levels(training\_set$restmt\_teq)  
training\_set$restmt\_teq\_target <- 0  
for (level in levels(training\_set$restmt\_teq)) {  
 training\_set[training\_set[, "restmt\_teq"]==level, "restmt\_teq\_target"] <- restmt\_teq\_target[level]  
}  
  
test\_set$restmt\_teq\_target <- 0  
for (level in levels(training\_set$restmt\_teq)) {  
 test\_set[test\_set[, "restmt\_teq"]==level, "restmt\_teq\_target"] <- restmt\_teq\_target[level]  
}  
  
restmt\_txt\_target <- mapply(target\_enc\_train, variable = "restmt\_txt", level = levels(training\_set$restmt\_txt), USE.NAMES = FALSE)  
names(restmt\_txt\_target) <- levels(training\_set$restmt\_txt)  
training\_set$restmt\_txt\_target <- 0  
for (level in levels(training\_set$restmt\_txt)) {  
 training\_set[training\_set[, "restmt\_txt"]==level, "restmt\_txt\_target"] <- restmt\_txt\_target[level]  
}  
  
test\_set$restmt\_txt\_target <- 0  
for (level in levels(training\_set$restmt\_txt)) {  
 test\_set[test\_set[, "restmt\_txt"]==level, "restmt\_txt\_target"] <- restmt\_txt\_target[level]  
}  
  
restmt\_wcap\_target <- mapply(target\_enc\_train, variable = "restmt\_wcap", level = levels(training\_set$restmt\_wcap), USE.NAMES = FALSE)  
names(restmt\_wcap\_target) <- levels(training\_set$restmt\_wcap)  
training\_set$restmt\_wcap\_target <- 0  
for (level in levels(training\_set$restmt\_wcap)) {  
 training\_set[training\_set[, "restmt\_wcap"]==level, "restmt\_wcap\_target"] <- restmt\_wcap\_target[level]  
}  
  
test\_set$restmt\_wcap\_target <- 0  
for (level in levels(training\_set$restmt\_wcap)) {  
 test\_set[test\_set[, "restmt\_wcap"]==level, "restmt\_wcap\_target"] <- restmt\_wcap\_target[level]  
}  
  
restmt\_xint\_target <- mapply(target\_enc\_train, variable = "restmt\_xint", level = levels(training\_set$restmt\_xint), USE.NAMES = FALSE)  
names(restmt\_xint\_target) <- levels(training\_set$restmt\_xint)  
training\_set$restmt\_xint\_target <- 0  
for (level in levels(training\_set$restmt\_xint)) {  
 training\_set[training\_set[, "restmt\_xint"]==level, "restmt\_xint\_target"] <- restmt\_xint\_target[level]  
}  
  
test\_set$restmt\_xint\_target <- 0  
for (level in levels(training\_set$restmt\_xint)) {  
 test\_set[test\_set[, "restmt\_xint"]==level, "restmt\_xint\_target"] <- restmt\_xint\_target[level]  
}

nrow(training\_set)

## [1] 250

nrow(test\_set)

## [1] 83

target\_company\_row <- training\_set %>%   
 filter(gvkey == target\_company\_gv\_key)  
  
if (nrow(target\_company\_row) > 0) {  
 training\_set <- subset(training\_set, gvkey != target\_company\_gv\_key)  
} else {  
   
 target\_company\_row <- test\_set %>%   
 filter(gvkey == target\_company\_gv\_key)  
 test\_set <- subset(test\_set, gvkey != target\_company\_gv\_key)  
}  
  
nrow(training\_set)

## [1] 249

nrow(test\_set)

## [1] 83

target\_company\_row

## # A tibble: 1 x 77  
## # Groups: gvkey [1]  
## gvkey tic aco acominc ao aoloch aqc bkvlps caps ch chech csho  
## <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 9777 SJM 93.3 -45.8 111. -7.08 168. 47.1 4262. 249. -60.6 110.  
## # ... with 65 more variables: cstk <dbl>, dlc <dbl>, dltt <dbl>, dvt <dbl>,  
## # ebit <dbl>, epspi <dbl>, nopi <dbl>, re <dbl>, rect <dbl>, revt <dbl>,  
## # siv <dbl>, sstk <dbl>, teq <dbl>, tstk <dbl>, wcap <dbl>, restmt\_at <fct>,  
## # restmt\_at\_mag <dbl>, restmt\_capx <fct>, restmt\_capx\_mag <dbl>,  
## # restmt\_cogs <fct>, restmt\_cogs\_mag <dbl>, restmt\_dltt <fct>,  
## # restmt\_epspi <fct>, restmt\_epspi\_mag <dbl>, restmt\_ib <fct>,  
## # restmt\_ni <fct>, restmt\_ni\_mag <dbl>, restmt\_nopi <fct>,  
## # restmt\_nopi\_mag <dbl>, restmt\_reuna <fct>, restmt\_reuna\_mag <dbl>,  
## # restmt\_teq <fct>, restmt\_teq\_mag <dbl>, restmt\_txt <fct>,  
## # restmt\_txt\_mag <dbl>, restmt\_wcap <fct>, restmt\_wcap\_mag <dbl>,  
## # restmt\_xint <fct>, restmt\_xint\_mag <dbl>, cshtrd\_m <dbl>, prccd\_m <dbl>,  
## # prcod\_m <dbl>, trfd\_m <dbl>, trfm\_m <dbl>, sp\_rating <fct>,  
## # litigated <fct>, litigation\_settlement <dbl>, pe\_ratio <dbl>,  
## # wc\_ratio <dbl>, de\_ratio <dbl>, roe\_ratio <dbl>, sp\_rating\_target <dbl>,  
## # restmt\_at\_target <dbl>, restmt\_capx\_target <dbl>, restmt\_cogs\_target <dbl>,  
## # restmt\_dltt\_target <dbl>, restmt\_epspi\_target <dbl>,  
## # restmt\_ib\_target <dbl>, restmt\_ni\_target <dbl>, restmt\_nopi\_target <dbl>,  
## # restmt\_reuna\_target <dbl>, restmt\_teq\_target <dbl>,  
## # restmt\_txt\_target <dbl>, restmt\_wcap\_target <dbl>, restmt\_xint\_target <dbl>

### Remove the restatement variables which have near zero variance

training\_subset\_ds\_final\_1 <- subset(training\_set, select = -c(sp\_rating,  
 restmt\_at,  
 restmt\_capx,  
 restmt\_cogs,  
 restmt\_dltt,  
 restmt\_epspi,  
 restmt\_ib,  
 restmt\_ni,  
 restmt\_nopi,  
 restmt\_reuna,  
 restmt\_teq,  
 restmt\_txt,  
 restmt\_wcap,  
 restmt\_xint))  
  
  
test\_subset\_ds\_final\_1 <- subset(test\_set, select = -c(sp\_rating,  
 restmt\_at,  
 restmt\_capx,  
 restmt\_cogs,  
 restmt\_dltt,  
 restmt\_epspi,  
 restmt\_ib,  
 restmt\_ni,  
 restmt\_nopi,  
 restmt\_reuna,  
 restmt\_teq,  
 restmt\_txt,  
 restmt\_wcap,  
 restmt\_xint))  
  
target\_company\_row\_1 <- subset(target\_company\_row, select = -c(sp\_rating,  
 restmt\_at,  
 restmt\_capx,  
 restmt\_cogs,  
 restmt\_dltt,  
 restmt\_epspi,  
 restmt\_ib,  
 restmt\_ni,  
 restmt\_nopi,  
 restmt\_reuna,  
 restmt\_teq,  
 restmt\_txt,  
 restmt\_wcap,  
 restmt\_xint))  
  
  
training\_subset\_ds\_final\_1 <- training\_subset\_ds\_final\_1 %>%  
 relocate(litigated, .after = last\_col())  
  
test\_subset\_ds\_final\_1 <- test\_subset\_ds\_final\_1 %>%  
 relocate(litigated, .after = last\_col())  
  
target\_company\_row\_final <- target\_company\_row\_1 %>%  
 relocate(litigated, .after = last\_col())  
  
  
training\_subset\_ds\_final\_1 <- training\_subset\_ds\_final\_1 %>%  
 relocate(litigation\_settlement , .after = last\_col())  
  
test\_subset\_ds\_final\_1 <- test\_subset\_ds\_final\_1 %>%  
 relocate(litigation\_settlement, .after = last\_col())  
  
target\_company\_row\_final <- target\_company\_row\_final %>%  
 relocate(litigation\_settlement, .after = last\_col())  
  
training\_subset\_ds\_final <- training\_subset\_ds\_final\_1  
test\_subset\_ds\_final <- test\_subset\_ds\_final\_1  
  
ncol(training\_subset\_ds\_final)

## [1] 63

summary(training\_subset\_ds\_final)

## gvkey tic aco acominc   
## Min. : 1266 0161A : 1 Min. : 0.000 Min. :-19306.57   
## 1st Qu.: 12575 0170A : 1 1st Qu.: 0.351 1st Qu.: -24.15   
## Median : 31392 0173A : 1 Median : 8.138 Median : 0.00   
## Mean : 80970 3AHII : 1 Mean : 175.864 Mean : -187.26   
## 3rd Qu.:164700 3AIRW : 1 3rd Qu.: 93.742 3rd Qu.: 0.00   
## Max. :264393 3CAGZ : 1 Max. :4706.135 Max. : 3495.34   
## (Other):243   
## ao aoloch aqc bkvlps   
## Min. : 0.000 Min. :-465.250 Min. : -12.45 Min. : -82.3   
## 1st Qu.: 0.075 1st Qu.: -1.647 1st Qu.: 0.00 1st Qu.: 0.1   
## Median : 7.711 Median : 0.000 Median : 0.00 Median : 3.6   
## Mean : 180.350 Mean : 4.895 Mean : 97.10 Mean : 8339.9   
## 3rd Qu.: 91.925 3rd Qu.: 1.500 3rd Qu.: 12.32 3rd Qu.: 12.6   
## Max. :5132.600 Max. : 655.000 Max. :5559.02 Max. :1216746.9   
##   
## caps ch chech csho   
## Min. : -701.475 Min. : 0.000 Min. :-305.7500 Min. : 0.52   
## 1st Qu.: 6.498 1st Qu.: 1.202 1st Qu.: -0.1535 1st Qu.: 18.88   
## Median : 42.415 Median : 18.018 Median : 0.4822 Median : 52.49   
## Mean : 706.132 Mean : 363.447 Mean : 39.8367 Mean : 249.12   
## 3rd Qu.: 412.843 3rd Qu.: 181.018 3rd Qu.: 10.6358 3rd Qu.: 144.58   
## Max. :28658.250 Max. :7382.800 Max. :1358.0000 Max. :6252.56   
##   
## cstk dlc dltt dvt   
## Min. : 0.000 Min. : 0.000 Min. : 0.00 Min. : -0.006   
## 1st Qu.: 0.035 1st Qu.: 0.269 1st Qu.: 0.03 1st Qu.: 0.000   
## Median : 0.264 Median : 4.004 Median : 14.73 Median : 0.000   
## Mean : 196.959 Mean : 311.379 Mean : 1309.06 Mean : 203.163   
## 3rd Qu.: 12.908 3rd Qu.: 101.250 3rd Qu.: 929.11 3rd Qu.: 46.679   
## Max. :7290.750 Max. :15926.126 Max. :42659.60 Max. :6572.535   
##   
## ebit epspi nopi re   
## Min. : -208.760 Min. :-14.0200 Min. : -38.5478 Min. :-7570.29   
## 1st Qu.: -0.366 1st Qu.: -0.0550 1st Qu.: -0.0013 1st Qu.: -11.17   
## Median : 17.853 Median : 0.2425 Median : 0.1790 Median : 18.87   
## Mean : 700.042 Mean : 1.7230 Mean : 43.4463 Mean : 1704.11   
## 3rd Qu.: 351.466 3rd Qu.: 1.8200 3rd Qu.: 4.8362 3rd Qu.: 411.00   
## Max. :24345.400 Max. :230.7025 Max. :2224.4000 Max. :68884.60   
##   
## rect revt siv sstk   
## Min. : 0.000 Min. : 0.0 Min. : 0.0000 Min. : 0.0000   
## 1st Qu.: 1.479 1st Qu.: 16.1 1st Qu.: 0.0000 1st Qu.: 0.0012   
## Median : 23.143 Median : 308.6 Median : 0.0000 Median : 0.9955   
## Mean : 499.237 Mean : 7936.7 Mean : 34.1581 Mean : 37.3398   
## 3rd Qu.: 343.470 3rd Qu.: 3793.0 3rd Qu.: 0.3503 3rd Qu.: 13.3750   
## Max. :15020.067 Max. :442511.4 Max. :1622.0000 Max. :1057.0000   
##   
## teq tstk wcap restmt\_at\_mag   
## Min. :-2208.96 Min. : 0.000 Min. :-8236.800 Min. :-1.011750   
## 1st Qu.: 3.38 1st Qu.: 0.000 1st Qu.: -0.014 1st Qu.: 0.000000   
## Median : 103.36 Median : 0.000 Median : 22.957 Median : 0.000000   
## Mean : 2434.16 Mean : 591.651 Mean : 263.686 Mean : 0.001972   
## 3rd Qu.: 1118.67 3rd Qu.: 8.386 3rd Qu.: 288.837 3rd Qu.: 0.000000   
## Max. :76602.80 Max. :25036.250 Max. :12261.750 Max. : 1.275750   
##   
## restmt\_capx\_mag restmt\_cogs\_mag restmt\_epspi\_mag restmt\_ni\_mag   
## Min. :-22.7162 Min. :-50.0000 Min. : -50.0 Min. : -6.611   
## 1st Qu.: 0.0000 1st Qu.: 0.0000 1st Qu.: 0.0 1st Qu.: 0.000   
## Median : 0.0000 Median : 0.0000 Median : 0.0 Median : 0.000   
## Mean : -0.1057 Mean : 0.4436 Mean : 424.5 Mean : 10.776   
## 3rd Qu.: 0.0000 3rd Qu.: 0.0000 3rd Qu.: 0.0 3rd Qu.: 0.000   
## Max. : 8.3335 Max. :100.0000 Max. :77081.7 Max. :2683.890   
##   
## restmt\_nopi\_mag restmt\_reuna\_mag restmt\_teq\_mag restmt\_txt\_mag   
## Min. :-1868600.0 Min. : -85.20 Min. : -105.39 Min. :-88.7704   
## 1st Qu.: -77.3 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.0000   
## Median : 0.0 Median : 0.00 Median : 0.00 Median : 0.0000   
## Mean : -7977.9 Mean : 16.96 Mean : 49.75 Mean : -0.9797   
## 3rd Qu.: 15.1 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.: 0.0000   
## Max. : 68865.1 Max. :4181.70 Max. :12541.75 Max. : 17.8622   
##   
## restmt\_wcap\_mag restmt\_xint\_mag cshtrd\_m prccd\_m   
## Min. : -3.149 Min. :-62.735 Min. : 0 Min. : 0.0018   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 18025 1st Qu.: 1.1177   
## Median : 0.000 Median : 0.000 Median : 111926 Median : 8.6464   
## Mean : 1.642 Mean : -1.035 Mean : 872188 Mean : 32.3903   
## 3rd Qu.: 0.000 3rd Qu.: 0.000 3rd Qu.: 614194 3rd Qu.: 30.4042   
## Max. :412.500 Max. : 0.562 Max. :13129451 Max. :2217.8253   
##   
## prcod\_m trfd\_m trfm\_m pe\_ratio   
## Min. : 0.0362 Min. : 1.000 Min. : 1.000 Min. : -1220   
## 1st Qu.: 3.1852 1st Qu.: 1.063 1st Qu.: 1.000 1st Qu.: -5   
## Median : 11.3376 Median : 1.227 Median : 1.000 Median : 12   
## Mean : 40.1783 Mean : 3.017 Mean : 2.007 Mean : 504488   
## 3rd Qu.: 34.8010 3rd Qu.: 1.827 3rd Qu.: 1.475 3rd Qu.: 22   
## Max. :2217.8208 Max. :218.416 Max. :34.527 Max. :76903023   
##   
## wc\_ratio de\_ratio roe\_ratio sp\_rating\_target  
## Min. : 0.0000 Min. :-60.4827 Min. :-192.20000 Min. :0.1226   
## 1st Qu.: 0.3719 1st Qu.: 0.2035 1st Qu.: -0.00117 1st Qu.:0.1226   
## Median : 0.7205 Median : 0.7431 Median : 0.11160 Median :0.1226   
## Mean : 1.2785 Mean : 0.8113 Mean : -0.68051 Mean :0.1292   
## 3rd Qu.: 1.3596 3rd Qu.: 1.4790 3rd Qu.: 0.25771 3rd Qu.:0.1343   
## Max. :12.4471 Max. : 80.2000 Max. : 10.15006 Max. :0.1483   
##   
## restmt\_at\_target restmt\_capx\_target restmt\_cogs\_target restmt\_dltt\_target  
## Min. :0.1278 Min. :0.1341 Min. :0.1311 Min. :0.1354   
## 1st Qu.:0.1278 1st Qu.:0.1341 1st Qu.:0.1311 1st Qu.:0.1354   
## Median :0.1278 Median :0.1341 Median :0.1311 Median :0.1354   
## Mean :0.1292 Mean :0.1343 Mean :0.1345 Mean :0.1354   
## 3rd Qu.:0.1278 3rd Qu.:0.1341 3rd Qu.:0.1424 3rd Qu.:0.1354   
## Max. :0.1509 Max. :0.1395 Max. :0.1424 Max. :0.1373   
##   
## restmt\_epspi\_target restmt\_ib\_target restmt\_ni\_target restmt\_nopi\_target  
## Min. :0.1258 Min. :0.1250 Min. :0.1263 Min. :0.1279   
## 1st Qu.:0.1258 1st Qu.:0.1250 1st Qu.:0.1263 1st Qu.:0.1279   
## Median :0.1258 Median :0.1250 Median :0.1263 Median :0.1429   
## Mean :0.1292 Mean :0.1283 Mean :0.1273 Mean :0.1372   
## 3rd Qu.:0.1258 3rd Qu.:0.1250 3rd Qu.:0.1263 3rd Qu.:0.1429   
## Max. :0.1530 Max. :0.1547 Max. :0.1544 Max. :0.1429   
##   
## restmt\_reuna\_target restmt\_teq\_target restmt\_txt\_target restmt\_wcap\_target  
## Min. :0.1286 Min. :0.1273 Min. :0.1266 Min. :0.1324   
## 1st Qu.:0.1286 1st Qu.:0.1273 1st Qu.:0.1266 1st Qu.:0.1324   
## Median :0.1286 Median :0.1273 Median :0.1266 Median :0.1324   
## Mean :0.1301 Mean :0.1294 Mean :0.1285 Mean :0.1328   
## 3rd Qu.:0.1286 3rd Qu.:0.1273 3rd Qu.:0.1266 3rd Qu.:0.1324   
## Max. :0.1493 Max. :0.1513 Max. :0.1530 Max. :0.1429   
##   
## restmt\_xint\_target litigated litigation\_settlement  
## Min. :0.1303 0:215 Min. : 1500000   
## 1st Qu.:0.1303 1: 34 1st Qu.: 3362500   
## Median :0.1303 Median : 6375000   
## Mean :0.1322 Mean :10993750   
## 3rd Qu.:0.1303 3rd Qu.: 9562500   
## Max. :0.1454 Max. :47500000   
## NA's :241

target\_company\_row\_final

## # A tibble: 1 x 63  
## # Groups: gvkey [1]  
## gvkey tic aco acominc ao aoloch aqc bkvlps caps ch chech csho  
## <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 9777 SJM 93.3 -45.8 111. -7.08 168. 47.1 4262. 249. -60.6 110.  
## # ... with 51 more variables: cstk <dbl>, dlc <dbl>, dltt <dbl>, dvt <dbl>,  
## # ebit <dbl>, epspi <dbl>, nopi <dbl>, re <dbl>, rect <dbl>, revt <dbl>,  
## # siv <dbl>, sstk <dbl>, teq <dbl>, tstk <dbl>, wcap <dbl>,  
## # restmt\_at\_mag <dbl>, restmt\_capx\_mag <dbl>, restmt\_cogs\_mag <dbl>,  
## # restmt\_epspi\_mag <dbl>, restmt\_ni\_mag <dbl>, restmt\_nopi\_mag <dbl>,  
## # restmt\_reuna\_mag <dbl>, restmt\_teq\_mag <dbl>, restmt\_txt\_mag <dbl>,  
## # restmt\_wcap\_mag <dbl>, restmt\_xint\_mag <dbl>, cshtrd\_m <dbl>,  
## # prccd\_m <dbl>, prcod\_m <dbl>, trfd\_m <dbl>, trfm\_m <dbl>, pe\_ratio <dbl>,  
## # wc\_ratio <dbl>, de\_ratio <dbl>, roe\_ratio <dbl>, sp\_rating\_target <dbl>,  
## # restmt\_at\_target <dbl>, restmt\_capx\_target <dbl>, restmt\_cogs\_target <dbl>,  
## # restmt\_dltt\_target <dbl>, restmt\_epspi\_target <dbl>,  
## # restmt\_ib\_target <dbl>, restmt\_ni\_target <dbl>, restmt\_nopi\_target <dbl>,  
## # restmt\_reuna\_target <dbl>, restmt\_teq\_target <dbl>,  
## # restmt\_txt\_target <dbl>, restmt\_wcap\_target <dbl>,  
## # restmt\_xint\_target <dbl>, litigated <fct>, litigation\_settlement <dbl>

training\_subset\_ds\_final\_settlement <- training\_subset\_ds\_final  
test\_subset\_ds\_final\_settlement <- test\_subset\_ds\_final

### Scale the variables.

num\_var\_start\_index <- 3  
num\_var\_end\_index <- ncol(training\_subset\_ds\_final) - 2  
target\_var\_index <- ncol(training\_subset\_ds\_final) - 1  
  
num\_var\_start\_index

## [1] 3

num\_var\_end\_index

## [1] 61

target\_var\_index

## [1] 62

test\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index] <- scale(test\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index],   
 center = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, mean),   
 scale = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, sd))  
  
#target\_company\_row\_final  
target\_company\_row\_final[, num\_var\_start\_index:num\_var\_end\_index] = scale(target\_company\_row\_final[, num\_var\_start\_index:num\_var\_end\_index],   
 center = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, mean),   
 scale = apply(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index], 2, sd))  
  
training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index] <- scale(training\_subset\_ds\_final[, num\_var\_start\_index:num\_var\_end\_index])  
  
levels(training\_subset\_ds\_final$litigated)[levels(training\_subset\_ds\_final$litigated) == 1] <- "Yes"  
levels(training\_subset\_ds\_final$litigated)[levels(training\_subset\_ds\_final$litigated) == 0] <- "No"  
  
target\_company\_row\_final

## # A tibble: 1 x 63  
## # Groups: gvkey [1]  
## gvkey tic aco acominc ao aoloch aqc bkvlps caps ch chech  
## <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 9777 SJM -0.155 0.107 -0.142 -0.121 0.174 -0.0885 1.54 -0.118 -0.649  
## # ... with 52 more variables: csho <dbl>, cstk <dbl>, dlc <dbl>, dltt <dbl>,  
## # dvt <dbl>, ebit <dbl>, epspi <dbl>, nopi <dbl>, re <dbl>, rect <dbl>,  
## # revt <dbl>, siv <dbl>, sstk <dbl>, teq <dbl>, tstk <dbl>, wcap <dbl>,  
## # restmt\_at\_mag <dbl>, restmt\_capx\_mag <dbl>, restmt\_cogs\_mag <dbl>,  
## # restmt\_epspi\_mag <dbl>, restmt\_ni\_mag <dbl>, restmt\_nopi\_mag <dbl>,  
## # restmt\_reuna\_mag <dbl>, restmt\_teq\_mag <dbl>, restmt\_txt\_mag <dbl>,  
## # restmt\_wcap\_mag <dbl>, restmt\_xint\_mag <dbl>, cshtrd\_m <dbl>,  
## # prccd\_m <dbl>, prcod\_m <dbl>, trfd\_m <dbl>, trfm\_m <dbl>, pe\_ratio <dbl>,  
## # wc\_ratio <dbl>, de\_ratio <dbl>, roe\_ratio <dbl>, sp\_rating\_target <dbl>,  
## # restmt\_at\_target <dbl>, restmt\_capx\_target <dbl>, restmt\_cogs\_target <dbl>,  
## # restmt\_dltt\_target <dbl>, restmt\_epspi\_target <dbl>,  
## # restmt\_ib\_target <dbl>, restmt\_ni\_target <dbl>, restmt\_nopi\_target <dbl>,  
## # restmt\_reuna\_target <dbl>, restmt\_teq\_target <dbl>,  
## # restmt\_txt\_target <dbl>, restmt\_wcap\_target <dbl>,  
## # restmt\_xint\_target <dbl>, litigated <fct>, litigation\_settlement <dbl>

summary(training\_subset\_ds\_final)

## gvkey tic aco acominc   
## Min. : 1266 0161A : 1 Min. :-0.3300 Min. :-14.4470   
## 1st Qu.: 12575 0170A : 1 1st Qu.:-0.3293 1st Qu.: 0.1232   
## Median : 31392 0173A : 1 Median :-0.3147 Median : 0.1415   
## Mean : 80970 3AHII : 1 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:164700 3AIRW : 1 3rd Qu.:-0.1541 3rd Qu.: 0.1415   
## Max. :264393 3CAGZ : 1 Max. : 8.5009 Max. : 2.7827   
## (Other):243   
## ao aoloch aqc bkvlps   
## Min. :-0.3698 Min. :-4.74499 Min. :-0.2692 Min. :-0.08984   
## 1st Qu.:-0.3696 1st Qu.:-0.06602 1st Qu.:-0.2386 1st Qu.:-0.08897   
## Median :-0.3540 Median :-0.04940 Median :-0.2386 Median :-0.08893   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.:-0.1813 3rd Qu.:-0.03426 3rd Qu.:-0.2083 3rd Qu.:-0.08883   
## Max. :10.1539 Max. : 6.56127 Max. :13.4226 Max. :12.89060   
##   
## caps ch chech csho   
## Min. :-0.6113 Min. :-0.3742 Min. :-2.2327 Min. :-0.3866   
## 1st Qu.:-0.3038 1st Qu.:-0.3729 1st Qu.:-0.2584 1st Qu.:-0.3580   
## Median :-0.2882 Median :-0.3556 Median :-0.2543 Median :-0.3058   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1274 3rd Qu.:-0.1878 3rd Qu.:-0.1887 3rd Qu.:-0.1626   
## Max. :12.1385 Max. : 7.2263 Max. : 8.5161 Max. : 9.3353   
##   
## cstk dlc dltt dvt   
## Min. :-0.2397 Min. :-0.2440 Min. :-0.3488 Min. :-0.2885   
## 1st Qu.:-0.2397 1st Qu.:-0.2438 1st Qu.:-0.3488 1st Qu.:-0.2885   
## Median :-0.2394 Median :-0.2408 Median :-0.3448 Median :-0.2885   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2240 3rd Qu.:-0.1646 3rd Qu.:-0.1012 3rd Qu.:-0.2222   
## Max. : 8.6348 Max. :12.2342 Max. :11.0168 Max. : 9.0453   
##   
## ebit epspi nopi re   
## Min. :-0.4071 Min. :-1.070685 Min. :-0.4232 Min. :-1.4979   
## 1st Qu.:-0.3137 1st Qu.:-0.120920 1st Qu.:-0.2243 1st Qu.:-0.2770   
## Median :-0.3056 Median :-0.100687 Median :-0.2233 Median :-0.2722   
## Mean : 0.0000 Mean : 0.000000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1561 3rd Qu.: 0.006599 3rd Qu.:-0.1993 3rd Qu.:-0.2089   
## Max. :10.5920 Max. :15.572979 Max. :11.2569 Max. :10.8504   
##   
## rect revt siv sstk   
## Min. :-0.3542 Min. :-0.2501 Min. :-0.2110 Min. :-0.3282   
## 1st Qu.:-0.3531 1st Qu.:-0.2496 1st Qu.:-0.2110 1st Qu.:-0.3282   
## Median :-0.3377 Median :-0.2403 Median :-0.2110 Median :-0.3195   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1105 3rd Qu.:-0.1306 3rd Qu.:-0.2088 3rd Qu.:-0.2106   
## Max. :10.3009 Max. :13.6928 Max. : 9.8069 Max. : 8.9623   
##   
## teq tstk wcap restmt\_at\_mag   
## Min. :-0.5964 Min. :-0.2350 Min. :-6.64063 Min. :-7.79754   
## 1st Qu.:-0.3122 1st Qu.:-0.2350 1st Qu.:-0.20600 1st Qu.:-0.01517   
## Median :-0.2994 Median :-0.2350 Median :-0.18806 Median :-0.01517   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.1690 3rd Qu.:-0.2317 3rd Qu.: 0.01965 3rd Qu.:-0.01517   
## Max. : 9.5272 Max. : 9.7101 Max. : 9.37296 Max. : 9.79789   
##   
## restmt\_capx\_mag restmt\_cogs\_mag restmt\_epspi\_mag restmt\_ni\_mag   
## Min. :-13.68831 Min. :-4.25205 Min. :-0.09336 Min. :-0.10223   
## 1st Qu.: 0.06399 1st Qu.:-0.03739 1st Qu.:-0.08352 1st Qu.:-0.06336   
## Median : 0.06399 Median :-0.03739 Median :-0.08352 Median :-0.06336   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.06399 3rd Qu.:-0.03739 3rd Qu.:-0.08352 3rd Qu.:-0.06336   
## Max. : 5.10905 Max. : 8.39193 Max. :15.08304 Max. :15.71626   
##   
## restmt\_nopi\_mag restmt\_reuna\_mag restmt\_teq\_mag restmt\_txt\_mag   
## Min. :-15.67603 Min. :-0.38523 Min. :-0.19517 Min. :-10.4469   
## 1st Qu.: 0.06656 1st Qu.:-0.06395 1st Qu.:-0.06259 1st Qu.: 0.1166   
## Median : 0.06721 Median :-0.06395 Median :-0.06259 Median : 0.1166   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.06734 3rd Qu.:-0.06395 3rd Qu.:-0.06259 3rd Qu.: 0.1166   
## Max. : 0.64741 Max. :15.70489 Max. :15.71558 Max. : 2.2421   
##   
## restmt\_wcap\_mag restmt\_xint\_mag cshtrd\_m prccd\_m   
## Min. :-0.18326 Min. :-11.0587 Min. :-0.4389 Min. :-0.21813   
## 1st Qu.:-0.06281 1st Qu.: 0.1856 1st Qu.:-0.4299 1st Qu.:-0.21061   
## Median :-0.06281 Median : 0.1856 Median :-0.3826 Median :-0.15991   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.:-0.06281 3rd Qu.: 0.1856 3rd Qu.:-0.1298 3rd Qu.:-0.01338   
## Max. :15.71528 Max. : 0.2863 Max. : 6.1685 Max. :14.71819   
##   
## prcod\_m trfd\_m trfm\_m pe\_ratio   
## Min. :-0.22997 Min. :-0.1347 Min. :-0.2986 Min. :-0.08941   
## 1st Qu.:-0.21193 1st Qu.:-0.1305 1st Qu.:-0.2986 1st Qu.:-0.08920   
## Median :-0.16523 Median :-0.1195 Median :-0.2986 Median :-0.08919   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.:-0.03081 3rd Qu.:-0.0795 3rd Qu.:-0.1578 3rd Qu.:-0.08919   
## Max. :12.47556 Max. :14.3859 Max. : 9.6410 Max. :13.50746   
##   
## wc\_ratio de\_ratio roe\_ratio sp\_rating\_target   
## Min. :-0.75253 Min. :-8.870033 Min. :-15.43447 Min. :-0.6249   
## 1st Qu.:-0.53360 1st Qu.:-0.087951 1st Qu.: 0.05475 1st Qu.:-0.6249   
## Median :-0.32842 Median :-0.009862 Median : 0.06384 Median :-0.6249   
## Mean : 0.00000 Mean : 0.000000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.04772 3rd Qu.: 0.096630 3rd Qu.: 0.07561 3rd Qu.: 0.4963   
## Max. : 6.57371 Max. :11.488569 Max. : 0.87283 Max. : 1.8344   
##   
## restmt\_at\_target restmt\_capx\_target restmt\_cogs\_target restmt\_dltt\_target  
## Min. :-0.2527 Min. :-0.1818 Min. :-0.6427 Min. :-0.1429   
## 1st Qu.:-0.2527 1st Qu.:-0.1818 1st Qu.:-0.6427 1st Qu.:-0.1429   
## Median :-0.2527 Median :-0.1818 Median :-0.6427 Median :-0.1429   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2527 3rd Qu.:-0.1818 3rd Qu.: 1.5496 3rd Qu.:-0.1429   
## Max. : 3.9417 Max. : 5.4776 Max. : 1.5496 Max. : 6.9717   
##   
## restmt\_epspi\_target restmt\_ib\_target restmt\_ni\_target restmt\_nopi\_target  
## Min. :-0.3763 Min. :-0.3552 Min. :-0.1933 Min. :-1.2815   
## 1st Qu.:-0.3763 1st Qu.:-0.3552 1st Qu.:-0.1933 1st Qu.:-1.2815   
## Median :-0.3763 Median :-0.3552 Median :-0.1933 Median : 0.7772   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.3763 3rd Qu.:-0.3552 3rd Qu.:-0.1933 3rd Qu.: 0.7772   
## Max. : 2.6465 Max. : 2.8038 Max. : 5.1536 Max. : 0.7772   
##   
## restmt\_reuna\_target restmt\_teq\_target restmt\_txt\_target restmt\_wcap\_target  
## Min. :-0.2786 Min. :-0.3029 Min. :-0.2786 Min. :-0.1933   
## 1st Qu.:-0.2786 1st Qu.:-0.3029 1st Qu.:-0.2786 1st Qu.:-0.1933   
## Median :-0.2786 Median :-0.3029 Median :-0.2786 Median :-0.1933   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2786 3rd Qu.:-0.3029 3rd Qu.:-0.2786 3rd Qu.:-0.1933   
## Max. : 3.5752 Max. : 3.2884 Max. : 3.5752 Max. : 5.1536   
##   
## restmt\_xint\_target litigated litigation\_settlement  
## Min. :-0.3763 No :215 Min. : 1500000   
## 1st Qu.:-0.3763 Yes: 34 1st Qu.: 3362500   
## Median :-0.3763 Median : 6375000   
## Mean : 0.0000 Mean :10993750   
## 3rd Qu.:-0.3763 3rd Qu.: 9562500   
## Max. : 2.6465 Max. :47500000   
## NA's :241

training\_subset\_ds\_final\_orig <- training\_subset\_ds\_final

colnames(training\_subset\_ds\_final)

## [1] "gvkey" "tic" "aco"   
## [4] "acominc" "ao" "aoloch"   
## [7] "aqc" "bkvlps" "caps"   
## [10] "ch" "chech" "csho"   
## [13] "cstk" "dlc" "dltt"   
## [16] "dvt" "ebit" "epspi"   
## [19] "nopi" "re" "rect"   
## [22] "revt" "siv" "sstk"   
## [25] "teq" "tstk" "wcap"   
## [28] "restmt\_at\_mag" "restmt\_capx\_mag" "restmt\_cogs\_mag"   
## [31] "restmt\_epspi\_mag" "restmt\_ni\_mag" "restmt\_nopi\_mag"   
## [34] "restmt\_reuna\_mag" "restmt\_teq\_mag" "restmt\_txt\_mag"   
## [37] "restmt\_wcap\_mag" "restmt\_xint\_mag" "cshtrd\_m"   
## [40] "prccd\_m" "prcod\_m" "trfd\_m"   
## [43] "trfm\_m" "pe\_ratio" "wc\_ratio"   
## [46] "de\_ratio" "roe\_ratio" "sp\_rating\_target"   
## [49] "restmt\_at\_target" "restmt\_capx\_target" "restmt\_cogs\_target"   
## [52] "restmt\_dltt\_target" "restmt\_epspi\_target" "restmt\_ib\_target"   
## [55] "restmt\_ni\_target" "restmt\_nopi\_target" "restmt\_reuna\_target"   
## [58] "restmt\_teq\_target" "restmt\_txt\_target" "restmt\_wcap\_target"   
## [61] "restmt\_xint\_target" "litigated" "litigation\_settlement"

training\_subset\_ds\_final <- training\_subset\_ds\_final\_orig  
table(training\_subset\_ds\_final$litigated)

##   
## No Yes   
## 215 34

training\_subset\_ds\_final\_no <- training\_subset\_ds\_final %>%  
 filter(litigated == 'No')  
no\_count <- nrow(training\_subset\_ds\_final\_no)  
training\_subset\_ds\_final\_yes <- training\_subset\_ds\_final %>%  
 filter(litigated == 'Yes')  
yes\_count <- nrow(training\_subset\_ds\_final\_yes)  
#training\_subset\_ds\_final %>%  
# arrange(gvkey)  
  
table(training\_subset\_ds\_final$litigated)

##   
## No Yes   
## 215 34

# Sampling both (Over and under combination)  
training\_subset\_ds\_final <- ovun.sample(litigated ~   
 .,   
 data = training\_subset\_ds\_final, method = "both", N = no\_count + yes\_count, p = 0.35, seed = 222, na.action = na.pass)$data  
  
##### Sampling both (Over)  
#training\_subset\_ds\_final <- ovun.sample(litigated ~., data = training\_subset\_ds\_final, method = "over", N = no\_count\*2)$data  
  
##### Sampling both (Under)  
#training\_subset\_ds\_final <- ovun.sample(litigated ~., data = training\_subset\_ds\_final, method = "under", N = yes\_count\*2)$data  
  
##### Manual sample adjustment  
# litigated\_no <- training\_subset\_ds\_final %>%   
# filter(litigated == 'No')  
#   
# litigated\_yes <- training\_subset\_ds\_final %>%   
# filter(litigated == 'Yes')  
#   
# litigated\_no\_sample\_index <- sample(nrow(litigated\_no), 97)  
# litigated\_no\_sample <- litigated\_no[litigated\_no\_sample\_index, ]  
# #litigated\_no\_sample  
#   
# training\_subset\_ds\_final <- bind\_rows(litigated\_yes, litigated\_no\_sample)   
#   
# rows <- sample(nrow(training\_subset\_ds\_final))  
# training\_subset\_ds\_final <- training\_subset\_ds\_final[rows, ]  
  
table(training\_subset\_ds\_final$litigated)

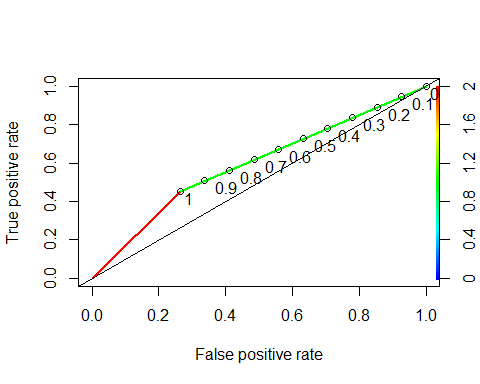
##   
## No Yes   
## 158 91

#summary(training\_subset\_ds\_final)  
#training\_subset\_ds\_final

glm\_control <- trainControl(  
 method = "cv",  
 number = 10,  
 summaryFunction = twoClassSummary,  
 classProbs = TRUE   
)  
set.seed(123)  
glm\_model\_1 <- train(litigated ~  
 aco + acominc + ao + aoloch + aqc + bkvlps + caps +   
 ch + chech + csho + cstk + dlc + dltt + dvt +  
 ebit + epspi + nopi + re + rect + revt + siv +  
 sstk + teq + tstk + wcap +  
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_target +   
 restmt\_capx\_target +   
 restmt\_dltt\_target +   
 restmt\_wcap\_target +   
 restmt\_teq\_target + restmt\_teq\_mag +  
 restmt\_cogs\_target + restmt\_cogs\_mag +  
 restmt\_nopi\_target + restmt\_nopi\_mag  
 ,   
 data = training\_subset\_ds\_final, method = "glm", family = "binomial", trControl = glm\_control)  
class\_probabilities <- predict(glm\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glm\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glm\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glm\_confusion\_matrix <- confusionMatrix(factor(glm\_class\_probabilities\_litigated), factor(glm\_litigated), positive = "Yes")  
glm\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 53 6  
## Yes 19 5  
##   
## Accuracy : 0.6988   
## 95% CI : (0.5882, 0.7947)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 1.0000   
##   
## Kappa : 0.1271   
##   
## Mcnemar's Test P-Value : 0.0164   
##   
## Sensitivity : 0.45455   
## Specificity : 0.73611   
## Pos Pred Value : 0.20833   
## Neg Pred Value : 0.89831   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.28916   
## Balanced Accuracy : 0.59533   
##   
## 'Positive' Class : Yes   
##

glm\_accuracy\_1 <- glm\_confusion\_matrix$overall["Accuracy"]  
glm\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
glm\_rocr\_roc <- performance(glm\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(glm\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



glm\_rocr\_auc <- performance(glm\_rocr\_pred, measure = "auc")  
glm\_auc\_1 <- glm\_rocr\_auc@y.values[[1]]  
glm\_auc\_1

## [1] 0.5953283

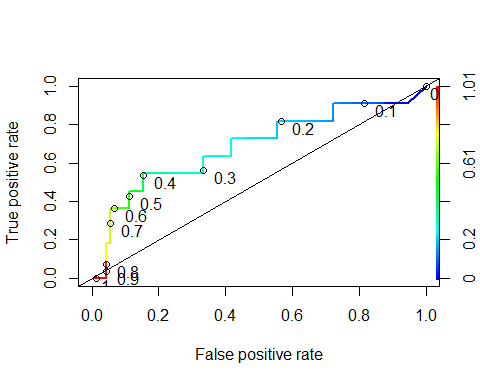
summary(glm\_model\_1)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -8.49 0.00 0.00 0.00 8.49   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.131e+15 4.605e+07 -111431553 <2e-16 \*\*\*  
## aco 4.043e+15 4.460e+07 90648900 <2e-16 \*\*\*  
## acominc 3.220e+15 2.348e+07 137178692 <2e-16 \*\*\*  
## ao -5.428e+14 3.610e+07 -15035192 <2e-16 \*\*\*  
## aoloch 7.484e+14 1.134e+07 65985858 <2e-16 \*\*\*  
## aqc -1.799e+15 1.886e+07 -95415693 <2e-16 \*\*\*  
## bkvlps -4.083e+16 4.532e+08 -90093108 <2e-16 \*\*\*  
## caps -5.965e+14 2.821e+07 -21147286 <2e-16 \*\*\*  
## ch 2.603e+15 2.856e+07 91136017 <2e-16 \*\*\*  
## chech -4.722e+14 1.055e+07 -44767007 <2e-16 \*\*\*  
## csho -8.314e+14 1.753e+07 -47427961 <2e-16 \*\*\*  
## cstk 5.620e+14 1.347e+07 41728260 <2e-16 \*\*\*  
## dlc -1.026e+16 6.144e+07 -166928123 <2e-16 \*\*\*  
## dltt 5.309e+15 4.711e+07 112700161 <2e-16 \*\*\*  
## dvt -7.486e+15 7.284e+07 -102767882 <2e-16 \*\*\*  
## ebit 3.307e+15 7.892e+07 41904311 <2e-16 \*\*\*  
## epspi -3.465e+14 2.924e+07 -11848284 <2e-16 \*\*\*  
## nopi -1.091e+15 3.578e+07 -30508757 <2e-16 \*\*\*  
## re 1.722e+15 9.209e+07 18697194 <2e-16 \*\*\*  
## rect -4.393e+14 3.454e+07 -12718259 <2e-16 \*\*\*  
## revt -3.625e+14 3.580e+07 -10125245 <2e-16 \*\*\*  
## siv 1.275e+15 1.655e+07 77072451 <2e-16 \*\*\*  
## sstk 8.848e+14 9.141e+06 96797074 <2e-16 \*\*\*  
## teq -7.777e+14 1.127e+08 -6900627 <2e-16 \*\*\*  
## tstk -1.571e+12 4.632e+07 -33925 <2e-16 \*\*\*  
## wcap -1.513e+15 2.238e+07 -67587267 <2e-16 \*\*\*  
## cshtrd\_m -2.089e+14 7.604e+06 -27467564 <2e-16 \*\*\*  
## prccd\_m 2.320e+14 3.865e+07 6003860 <2e-16 \*\*\*  
## prcod\_m 5.635e+13 1.431e+07 3936974 <2e-16 \*\*\*  
## trfd\_m 8.241e+13 3.554e+06 23187236 <2e-16 \*\*\*  
## trfm\_m -3.220e+14 1.715e+07 -18778428 <2e-16 \*\*\*  
## pe\_ratio 1.959e+13 8.171e+06 2397887 <2e-16 \*\*\*  
## wc\_ratio 2.331e+14 5.077e+06 45920525 <2e-16 \*\*\*  
## de\_ratio 4.476e+14 6.794e+06 65883443 <2e-16 \*\*\*  
## roe\_ratio 5.034e+14 6.450e+06 78034543 <2e-16 \*\*\*  
## sp\_rating\_target -9.879e+13 8.245e+06 -11982371 <2e-16 \*\*\*  
## restmt\_at\_target 4.130e+14 8.428e+06 49001365 <2e-16 \*\*\*  
## restmt\_capx\_target 2.637e+14 6.838e+06 38570407 <2e-16 \*\*\*  
## restmt\_dltt\_target 6.535e+13 6.252e+06 10452290 <2e-16 \*\*\*  
## restmt\_wcap\_target -8.275e+13 4.911e+06 -16848970 <2e-16 \*\*\*  
## restmt\_teq\_target -3.090e+14 8.789e+06 -35155034 <2e-16 \*\*\*  
## restmt\_teq\_mag 8.047e+14 3.556e+08 2262754 <2e-16 \*\*\*  
## restmt\_cogs\_target -1.795e+14 5.858e+06 -30637310 <2e-16 \*\*\*  
## restmt\_cogs\_mag -6.047e+13 5.223e+06 -11577591 <2e-16 \*\*\*  
## restmt\_nopi\_target -7.571e+13 5.641e+06 -13421319 <2e-16 \*\*\*  
## restmt\_nopi\_mag 2.500e+14 4.314e+06 57938573 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 326.94 on 248 degrees of freedom  
## Residual deviance: 3532.28 on 203 degrees of freedom  
## AIC: 3624.3  
##   
## Number of Fisher Scoring iterations: 19

set.seed(123)  
glm\_model\_2 <- train(litigated ~  
 acominc +   
 bkvlps +   
 caps +   
 dlc +   
 dvt +   
 nopi +   
 re +   
 rect +   
 revt +   
 tstk +   
 wcap +   
 prccd\_m +   
 prcod\_m +   
 trfm\_m +   
 pe\_ratio +   
 wc\_ratio +   
 sp\_rating\_target +   
 restmt\_at\_target +   
 restmt\_capx\_target +   
 restmt\_dltt\_target +   
 restmt\_cogs\_target +   
 restmt\_nopi\_target +   
 restmt\_nopi\_mag  
 ,   
 data = training\_subset\_ds\_final, method = "glm", family = "binomial", trControl = glm\_control)  
class\_probabilities <- predict(glm\_model\_2, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glm\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glm\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glm\_confusion\_matrix <- confusionMatrix(factor(glm\_class\_probabilities\_litigated), factor(glm\_litigated), positive = "Yes")  
glm\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 64 7  
## Yes 8 4  
##   
## Accuracy : 0.8193   
## 95% CI : (0.7195, 0.8952)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9225   
##   
## Kappa : 0.2432   
##   
## Mcnemar's Test P-Value : 1.0000   
##   
## Sensitivity : 0.36364   
## Specificity : 0.88889   
## Pos Pred Value : 0.33333   
## Neg Pred Value : 0.90141   
## Prevalence : 0.13253   
## Detection Rate : 0.04819   
## Detection Prevalence : 0.14458   
## Balanced Accuracy : 0.62626   
##   
## 'Positive' Class : Yes   
##

glm\_accuracy\_2 <- glm\_confusion\_matrix$overall["Accuracy"]  
glm\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
glm\_rocr\_roc <- performance(glm\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(glm\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



glm\_rocr\_auc <- performance(glm\_rocr\_pred, measure = "auc")  
glm\_auc\_2 <- glm\_rocr\_auc@y.values[[1]]  
glm\_auc\_2

## [1] 0.6856061

summary(glm\_model\_2)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5378 -0.6855 -0.3338 0.7187 2.3352   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.251e+03 7.944e+02 -1.574 0.11541   
## acominc 1.485e+00 1.565e+00 0.949 0.34279   
## bkvlps -3.087e+03 1.396e+03 -2.212 0.02699 \*   
## caps 1.166e-01 7.493e-01 0.156 0.87634   
## dlc -5.101e+00 2.907e+00 -1.755 0.07930 .   
## dvt -5.324e+00 2.054e+00 -2.591 0.00956 \*\*  
## nopi 2.268e+00 9.762e-01 2.324 0.02014 \*   
## re -1.076e-01 1.728e+00 -0.062 0.95037   
## rect -4.942e-02 1.618e+00 -0.031 0.97563   
## revt 3.644e+00 1.592e+00 2.289 0.02210 \*   
## tstk 1.540e+00 9.650e-01 1.596 0.11046   
## wcap -5.575e-02 7.998e-01 -0.070 0.94443   
## prccd\_m 7.919e+00 5.656e+00 1.400 0.16148   
## prcod\_m -5.030e+00 6.412e+00 -0.785 0.43272   
## trfm\_m -5.185e-02 4.802e-01 -0.108 0.91400   
## pe\_ratio -1.092e+04 8.644e+03 -1.264 0.20632   
## wc\_ratio 1.574e-01 1.789e-01 0.880 0.37906   
## sp\_rating\_target 7.720e-01 2.800e-01 2.757 0.00583 \*\*  
## restmt\_at\_target 4.130e-01 2.188e-01 1.888 0.05904 .   
## restmt\_capx\_target 5.081e-01 2.200e-01 2.309 0.02094 \*   
## restmt\_dltt\_target -1.908e-01 2.505e-01 -0.762 0.44614   
## restmt\_cogs\_target -4.161e-01 2.336e-01 -1.782 0.07482 .   
## restmt\_nopi\_target -2.037e-01 2.187e-01 -0.931 0.35171   
## restmt\_nopi\_mag 4.413e+00 7.017e+00 0.629 0.52941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 326.94 on 248 degrees of freedom  
## Residual deviance: 206.35 on 225 degrees of freedom  
## AIC: 254.35  
##   
## Number of Fisher Scoring iterations: 18

set.seed(123)  
glm\_model\_3 <- train(litigated ~  
 dvt +   
 nopi +   
 revt +   
 tstk +   
 prccd\_m +   
 trfm\_m +   
 pe\_ratio +   
 wc\_ratio +   
 sp\_rating\_target  
 ,   
 data = training\_subset\_ds\_final, method = "glm", family = "binomial", trControl = glm\_control)  
class\_probabilities <- predict(glm\_model\_3, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glm\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glm\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glm\_confusion\_matrix <- confusionMatrix(factor(glm\_class\_probabilities\_litigated), factor(glm\_litigated), positive = "Yes")  
glm\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 64 6  
## Yes 8 5  
##   
## Accuracy : 0.8313   
## 95% CI : (0.7332, 0.9046)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.8698   
##   
## Kappa : 0.3189   
##   
## Mcnemar's Test P-Value : 0.7893   
##   
## Sensitivity : 0.45455   
## Specificity : 0.88889   
## Pos Pred Value : 0.38462   
## Neg Pred Value : 0.91429   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.15663   
## Balanced Accuracy : 0.67172   
##   
## 'Positive' Class : Yes   
##

glm\_accuracy\_3 <- glm\_confusion\_matrix$overall["Accuracy"]  
glm\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
glm\_rocr\_roc <- performance(glm\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(glm\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



glm\_rocr\_auc <- performance(glm\_rocr\_pred, measure = "auc")  
glm\_auc\_3 <- glm\_rocr\_auc@y.values[[1]]  
glm\_auc\_3

## [1] 0.7537879

summary(glm\_model\_3)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8622 -0.7650 -0.6942 0.9230 2.5164   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -264.6458 491.3706 -0.539 0.59017   
## dvt -3.7066 1.3471 -2.752 0.00593 \*\*   
## nopi 0.6730 0.4909 1.371 0.17044   
## revt 1.9624 0.7308 2.685 0.00725 \*\*   
## tstk 0.7940 0.4234 1.875 0.06074 .   
## prccd\_m -0.1059 0.2135 -0.496 0.61971   
## trfm\_m -0.3501 0.4549 -0.770 0.44158   
## pe\_ratio -2956.1840 5508.9445 -0.537 0.59153   
## wc\_ratio 0.2332 0.1596 1.462 0.14385   
## sp\_rating\_target 0.8367 0.2019 4.145 3.4e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 326.94 on 248 degrees of freedom  
## Residual deviance: 251.56 on 239 degrees of freedom  
## AIC: 271.56  
##   
## Number of Fisher Scoring iterations: 18

glmnet\_control <- trainControl(  
 method = "cv",  
 number = 10,  
 summaryFunction = twoClassSummary,  
 classProbs = TRUE   
)  
set.seed(123)  
glmnet\_model\_1 <- train(litigated ~   
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data = training\_subset\_ds\_final, method = "glmnet", family = "binomial", trControl = glmnet\_control)  
class\_probabilities <- predict(glmnet\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glmnet\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glmnet\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glmnet\_confusion\_matrix <- confusionMatrix(factor(glmnet\_class\_probabilities\_litigated), factor(glmnet\_litigated), positive = "Yes")  
glmnet\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 59 6  
## Yes 13 5  
##   
## Accuracy : 0.7711   
## 95% CI : (0.6658, 0.8562)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9948   
##   
## Kappa : 0.2158   
##   
## Mcnemar's Test P-Value : 0.1687   
##   
## Sensitivity : 0.45455   
## Specificity : 0.81944   
## Pos Pred Value : 0.27778   
## Neg Pred Value : 0.90769   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.21687   
## Balanced Accuracy : 0.63699   
##   
## 'Positive' Class : Yes   
##

glmnet\_accuracy\_1 <- glmnet\_confusion\_matrix$overall["Accuracy"]

set.seed(123)  
glmnet\_model\_2 <- train(litigated ~   
 aco + acominc + ao + aoloch + aqc + bkvlps + caps +   
 ch + chech + csho + cstk + dlc + dltt + dvt +  
 ebit + epspi + nopi + re + rect + revt + siv +  
 sstk + teq + tstk + wcap +  
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_target +   
 restmt\_capx\_target +   
 restmt\_dltt\_target +   
 restmt\_wcap\_target +   
 restmt\_teq\_target + restmt\_teq\_mag +  
 restmt\_cogs\_target + restmt\_cogs\_mag +  
 restmt\_nopi\_target + restmt\_nopi\_mag,   
 data = training\_subset\_ds\_final, method = "glmnet", family = "binomial", trControl = glmnet\_control)  
class\_probabilities <- predict(glmnet\_model\_2, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glmnet\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glmnet\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glmnet\_confusion\_matrix <- confusionMatrix(factor(glmnet\_class\_probabilities\_litigated), factor(glmnet\_litigated), positive = "Yes")  
glmnet\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 60 8  
## Yes 12 3  
##   
## Accuracy : 0.759   
## 95% CI : (0.6527, 0.8462)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9977   
##   
## Kappa : 0.0919   
##   
## Mcnemar's Test P-Value : 0.5023   
##   
## Sensitivity : 0.27273   
## Specificity : 0.83333   
## Pos Pred Value : 0.20000   
## Neg Pred Value : 0.88235   
## Prevalence : 0.13253   
## Detection Rate : 0.03614   
## Detection Prevalence : 0.18072   
## Balanced Accuracy : 0.55303   
##   
## 'Positive' Class : Yes   
##

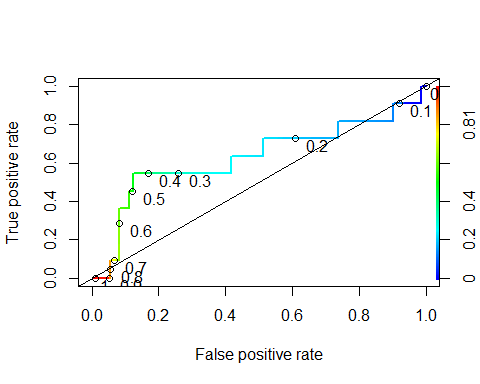
glmnet\_accuracy\_2 <- glmnet\_confusion\_matrix$overall["Accuracy"]

set.seed(123)  
glmnet\_model\_3 <- train(litigated ~   
 bkvlps +   
 dlc +   
 dvt +   
 nopi +   
 rect +   
 revt +   
 tstk +   
 wcap +   
 prccd\_m +   
 prcod\_m +   
 pe\_ratio +   
 wc\_ratio +   
 sp\_rating\_target +   
 restmt\_at\_target +   
 restmt\_capx\_target +   
 restmt\_nopi\_target,   
 data = training\_subset\_ds\_final, method = "glmnet", family = "binomial", trControl = glmnet\_control)  
class\_probabilities <- predict(glmnet\_model\_3, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
glmnet\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glmnet\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glmnet\_confusion\_matrix <- confusionMatrix(factor(glmnet\_class\_probabilities\_litigated), factor(glmnet\_litigated), positive = "Yes")  
glmnet\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 64 6  
## Yes 8 5  
##   
## Accuracy : 0.8313   
## 95% CI : (0.7332, 0.9046)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.8698   
##   
## Kappa : 0.3189   
##   
## Mcnemar's Test P-Value : 0.7893   
##   
## Sensitivity : 0.45455   
## Specificity : 0.88889   
## Pos Pred Value : 0.38462   
## Neg Pred Value : 0.91429   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.15663   
## Balanced Accuracy : 0.67172   
##   
## 'Positive' Class : Yes   
##

glmnet\_accuracy\_3 <- glmnet\_confusion\_matrix$overall["Accuracy"]

glmnet\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
glmnet\_rocr\_roc <- performance(glmnet\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(glmnet\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



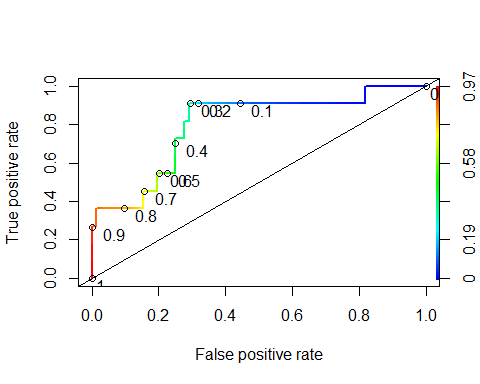
glmnet\_rocr\_auc <- performance(glmnet\_rocr\_pred, measure = "auc")  
glmnet\_auc <- glmnet\_rocr\_auc@y.values[[1]]  
glmnet\_auc

## [1] 0.6275253

gbm\_control <- trainControl(method = "repeatedcv", number = 10, repeats = 10)  
set.seed(123)  
gbm\_model\_1 <- train(litigated~  
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data=training\_subset\_ds\_final,   
 method = "gbm",  
 trControl = gbm\_control,  
 verbose = FALSE)  
  
class\_probabilities <- predict(gbm\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
gbm\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
gbm\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
gbm\_confusion\_matrix <- confusionMatrix(factor(gbm\_class\_probabilities\_litigated), factor(gbm\_litigated), positive = "Yes")  
gbm\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 56 5  
## Yes 16 6  
##   
## Accuracy : 0.747   
## 95% CI : (0.6396, 0.8361)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9990   
##   
## Kappa : 0.2271   
##   
## Mcnemar's Test P-Value : 0.0291   
##   
## Sensitivity : 0.54545   
## Specificity : 0.77778   
## Pos Pred Value : 0.27273   
## Neg Pred Value : 0.91803   
## Prevalence : 0.13253   
## Detection Rate : 0.07229   
## Detection Prevalence : 0.26506   
## Balanced Accuracy : 0.66162   
##   
## 'Positive' Class : Yes   
##

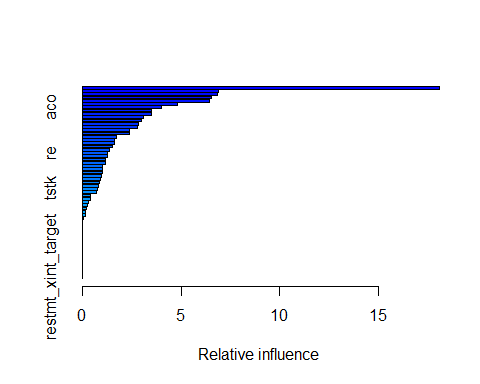
gbm\_accuracy\_1 <- gbm\_confusion\_matrix$overall["Accuracy"]  
gbm\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
gbm\_rocr\_roc <- performance(gbm\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(gbm\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



gbm\_rocr\_auc <- performance(gbm\_rocr\_pred, measure = "auc")  
gbm\_auc\_1 <- gbm\_rocr\_auc@y.values[[1]]  
gbm\_auc\_1

## [1] 0.7954545

summary(gbm\_model\_1)

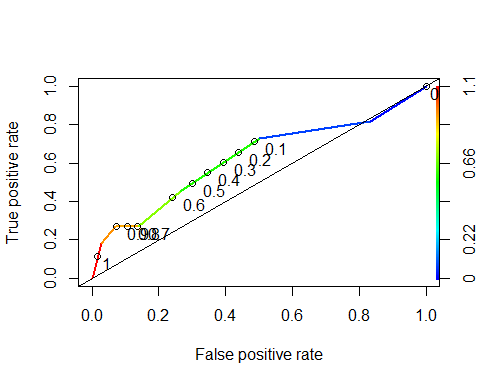


## var rel.inf  
## caps caps 18.07831944  
## dlc dlc 6.88627029  
## cshtrd\_m cshtrd\_m 6.84867012  
## prccd\_m prccd\_m 6.53401449  
## ebit ebit 6.41413120  
## cstk cstk 4.80156804  
## aco aco 4.00091708  
## prcod\_m prcod\_m 3.52031428  
## acominc acominc 3.48222060  
## wcap wcap 3.07182895  
## bkvlps bkvlps 2.99672348  
## pe\_ratio pe\_ratio 2.83105359  
## restmt\_nopi\_mag restmt\_nopi\_mag 2.78159259  
## wc\_ratio wc\_ratio 2.41058341  
## chech chech 2.37753171  
## trfd\_m trfd\_m 1.72339810  
## restmt\_ni\_target restmt\_ni\_target 1.63082527  
## ao ao 1.61775626  
## teq teq 1.54141432  
## aqc aqc 1.38748239  
## re re 1.28733137  
## restmt\_epspi\_target restmt\_epspi\_target 1.26962616  
## nopi nopi 1.19159279  
## roe\_ratio roe\_ratio 1.18979540  
## ch ch 1.02505827  
## siv siv 1.02127206  
## de\_ratio de\_ratio 0.99702012  
## sstk sstk 0.97114918  
## dltt dltt 0.93250197  
## trfm\_m trfm\_m 0.88367793  
## revt revt 0.80302419  
## tstk tstk 0.79369172  
## aoloch aoloch 0.72953898  
## epspi epspi 0.42120511  
## sp\_rating\_target sp\_rating\_target 0.39767194  
## dvt dvt 0.31475530  
## csho csho 0.25636091  
## restmt\_txt\_target restmt\_txt\_target 0.22033979  
## restmt\_epspi\_mag restmt\_epspi\_mag 0.16102464  
## restmt\_teq\_mag restmt\_teq\_mag 0.14313209  
## restmt\_cogs\_target restmt\_cogs\_target 0.04490493  
## restmt\_ib\_target restmt\_ib\_target 0.00870954  
## rect rect 0.00000000  
## restmt\_at\_mag restmt\_at\_mag 0.00000000  
## restmt\_capx\_mag restmt\_capx\_mag 0.00000000  
## restmt\_cogs\_mag restmt\_cogs\_mag 0.00000000  
## restmt\_ni\_mag restmt\_ni\_mag 0.00000000  
## restmt\_reuna\_mag restmt\_reuna\_mag 0.00000000  
## restmt\_txt\_mag restmt\_txt\_mag 0.00000000  
## restmt\_wcap\_mag restmt\_wcap\_mag 0.00000000  
## restmt\_xint\_mag restmt\_xint\_mag 0.00000000  
## restmt\_at\_target restmt\_at\_target 0.00000000  
## restmt\_capx\_target restmt\_capx\_target 0.00000000  
## restmt\_dltt\_target restmt\_dltt\_target 0.00000000  
## restmt\_nopi\_target restmt\_nopi\_target 0.00000000  
## restmt\_reuna\_target restmt\_reuna\_target 0.00000000  
## restmt\_teq\_target restmt\_teq\_target 0.00000000  
## restmt\_wcap\_target restmt\_wcap\_target 0.00000000  
## restmt\_xint\_target restmt\_xint\_target 0.00000000

set.seed(123)  
tree\_model\_1 <- rpart(litigated ~   
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data = training\_subset\_ds\_final, method = "class")  
class\_probabilities <- predict(tree\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- 1 - class\_probabilities[1: nrow(test\_subset\_ds\_final)]  
tree\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
  
tree\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
tree\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
tree\_confusion\_matrix <- confusionMatrix(factor(tree\_class\_probabilities\_litigated), factor(tree\_litigated), positive = "Yes")  
tree\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 53 6  
## Yes 19 5  
##   
## Accuracy : 0.6988   
## 95% CI : (0.5882, 0.7947)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 1.0000   
##   
## Kappa : 0.1271   
##   
## Mcnemar's Test P-Value : 0.0164   
##   
## Sensitivity : 0.45455   
## Specificity : 0.73611   
## Pos Pred Value : 0.20833   
## Neg Pred Value : 0.89831   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.28916   
## Balanced Accuracy : 0.59533   
##   
## 'Positive' Class : Yes   
##

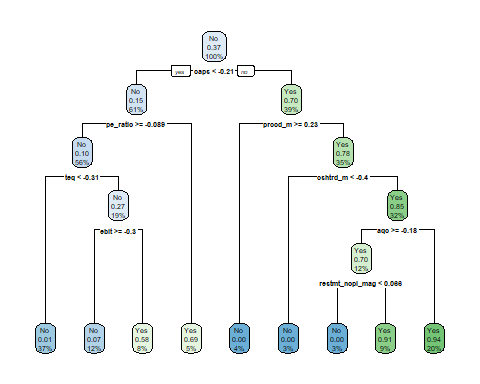
tree\_accuracy\_1 <- tree\_confusion\_matrix$overall["Accuracy"]  
  
tree\_rocr\_roc <- performance(tree\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(tree\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



tree\_rocr\_auc <- performance(tree\_rocr\_pred, measure = "auc")  
tree\_auc <- tree\_rocr\_auc@y.values[[1]]  
tree\_auc

## [1] 0.625

rpart.plot(tree\_model\_1)



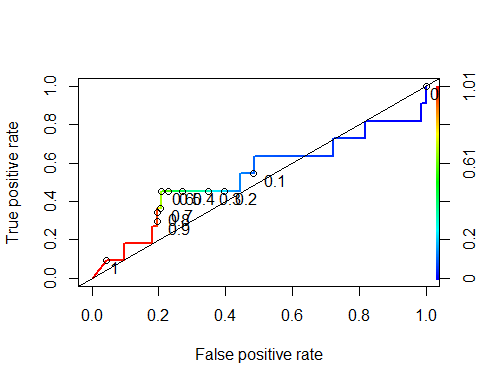
lda\_control <- trainControl(  
 method = "cv",  
 number = 10  
)  
set.seed(123)  
lda\_model\_1 <- train(litigated ~  
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data = training\_subset\_ds\_final, method = "lda", family = "binomial", trControl = lda\_control)  
class\_probabilities <- predict(lda\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_onehot <- class\_probabilities$Yes  
lda\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
lda\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
lda\_confusion\_matrix <- confusionMatrix(factor(lda\_class\_probabilities\_litigated), factor(lda\_litigated), positive = "Yes")  
lda\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 53 6  
## Yes 19 5  
##   
## Accuracy : 0.6988   
## 95% CI : (0.5882, 0.7947)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 1.0000   
##   
## Kappa : 0.1271   
##   
## Mcnemar's Test P-Value : 0.0164   
##   
## Sensitivity : 0.45455   
## Specificity : 0.73611   
## Pos Pred Value : 0.20833   
## Neg Pred Value : 0.89831   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.28916   
## Balanced Accuracy : 0.59533   
##   
## 'Positive' Class : Yes   
##

lda\_accuracy\_1 <- lda\_confusion\_matrix$overall["Accuracy"]  
lda\_accuracy\_1

## Accuracy   
## 0.6987952

lda\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_onehot, test\_subset\_ds\_final$litigated)  
lda\_rocr\_roc <- performance(lda\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(lda\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



lda\_rocr\_auc <- performance(lda\_rocr\_pred, measure = "auc")  
lda\_auc\_1 <- lda\_rocr\_auc@y.values[[1]]  
lda\_auc\_1

## [1] 0.5309343

rf\_x <- training\_subset\_ds\_final[,num\_var\_start\_index:num\_var\_end\_index]  
rf\_y <- training\_subset\_ds\_final[,target\_var\_index]  
recommended\_mtry <- sqrt(ncol(rf\_x))  
rfGrid <- expand.grid(mtry=recommended\_mtry)

rfControl <- trainControl(method='repeatedcv', number=10, repeats=3)  
set.seed(123)  
rf\_model\_1 <- train(litigated~  
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data=training\_subset\_ds\_final,   
 method='rf',   
 metric='Accuracy',   
 tuneGrid=rfGrid,   
 trControl=rfControl,   
 ntree = 100  
 )  
  
class\_probabilities <- predict(rf\_model\_1, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
  
rf\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
rf\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
rf\_confusion\_matrix <- confusionMatrix(factor(rf\_class\_probabilities\_litigated), factor(rf\_litigated), positive = "Yes")  
rf\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 67 7  
## Yes 5 4  
##   
## Accuracy : 0.8554   
## 95% CI : (0.7611, 0.923)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.6969   
##   
## Kappa : 0.3187   
##   
## Mcnemar's Test P-Value : 0.7728   
##   
## Sensitivity : 0.36364   
## Specificity : 0.93056   
## Pos Pred Value : 0.44444   
## Neg Pred Value : 0.90541   
## Prevalence : 0.13253   
## Detection Rate : 0.04819   
## Detection Prevalence : 0.10843   
## Balanced Accuracy : 0.64710   
##   
## 'Positive' Class : Yes   
##

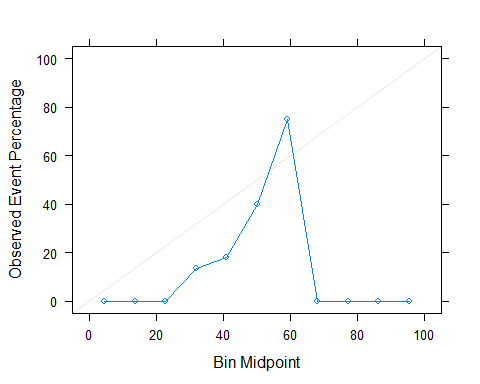
rf\_accuracy\_1 <- rf\_confusion\_matrix$overall["Accuracy"]  
rf\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
rf\_rocr\_roc <- performance(rf\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rf\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



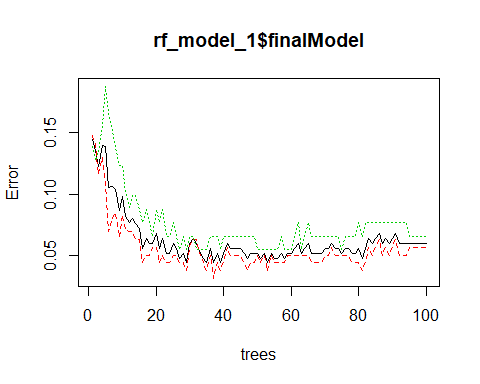
rf\_rocr\_auc <- performance(rf\_rocr\_pred, measure = "auc")  
rf\_auc\_1 <- rf\_rocr\_auc@y.values[[1]]  
rf\_auc\_1

## [1] 0.8863636

calibration\_curve <- calibration(litigated ~ class\_probabilities\_litigated,  
 data = test\_subset\_ds\_final,  
 class = 1)  
plot(calibration\_curve)



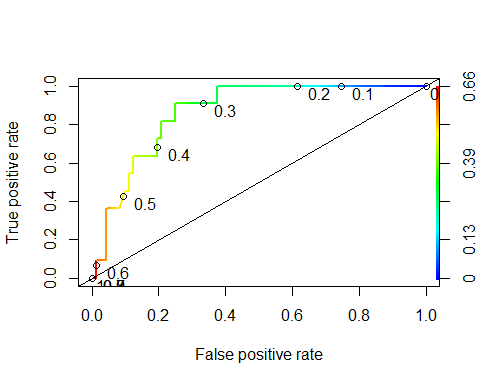
plot(rf\_model\_1$finalModel)



set.seed(123)  
rf\_model\_2 <- train(litigated~  
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data=training\_subset\_ds\_final,   
 method='rf',   
 metric='Accuracy',   
 tuneGrid=rfGrid,   
 trControl=rfControl,   
 ntree = 300  
 )  
  
class\_probabilities <- predict(rf\_model\_2, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
  
rf\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
rf\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
rf\_confusion\_matrix <- confusionMatrix(factor(rf\_class\_probabilities\_litigated), factor(rf\_litigated), positive = "Yes")  
rf\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 66 7  
## Yes 6 4  
##   
## Accuracy : 0.8434   
## 95% CI : (0.7471, 0.9139)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.795   
##   
## Kappa : 0.2915   
##   
## Mcnemar's Test P-Value : 1.000   
##   
## Sensitivity : 0.36364   
## Specificity : 0.91667   
## Pos Pred Value : 0.40000   
## Neg Pred Value : 0.90411   
## Prevalence : 0.13253   
## Detection Rate : 0.04819   
## Detection Prevalence : 0.12048   
## Balanced Accuracy : 0.64015   
##   
## 'Positive' Class : Yes   
##

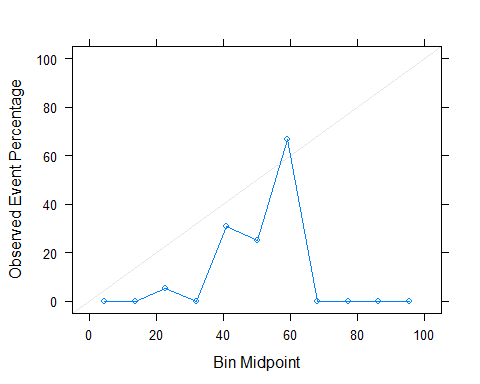
rf\_accuracy\_2 <- rf\_confusion\_matrix$overall["Accuracy"]  
rf\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
rf\_rocr\_roc <- performance(rf\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rf\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



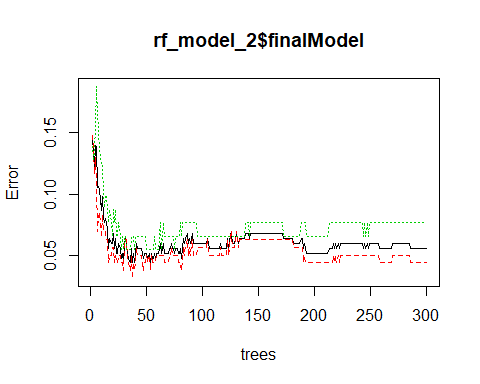
rf\_rocr\_auc <- performance(rf\_rocr\_pred, measure = "auc")  
rf\_auc\_2 <- rf\_rocr\_auc@y.values[[1]]  
rf\_auc\_2

## [1] 0.8642677

calibration\_curve <- calibration(litigated ~ class\_probabilities\_litigated,  
 data = test\_subset\_ds\_final,  
 class = 1)  
plot(calibration\_curve)



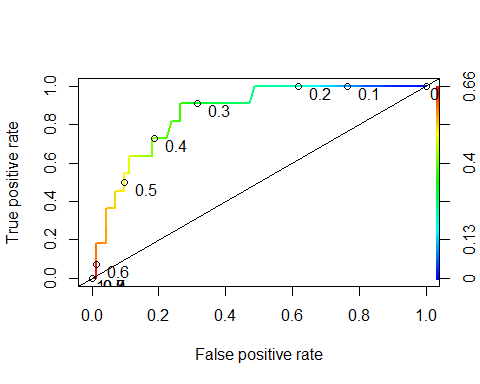
plot(rf\_model\_2$finalModel)



set.seed(123)  
rf\_model\_3 <- train(litigated~  
 aco + acominc + ao + aoloch + aqc +  
 bkvlps + caps + ch + chech + csho + cstk + dlc +  
 dltt + dvt + ebit + epspi + nopi + re + rect +  
 revt + siv + sstk + teq + tstk + wcap +   
 cshtrd\_m + prccd\_m + prcod\_m + trfd\_m + trfm\_m + pe\_ratio +  
 wc\_ratio + de\_ratio + roe\_ratio + sp\_rating\_target +  
 restmt\_at\_mag + restmt\_capx\_mag + restmt\_cogs\_mag +   
 restmt\_epspi\_mag + restmt\_ni\_mag + restmt\_nopi\_mag +   
 restmt\_reuna\_mag + restmt\_teq\_mag + restmt\_txt\_mag +   
 restmt\_wcap\_mag + restmt\_xint\_mag +   
 restmt\_at\_target + restmt\_capx\_target + restmt\_cogs\_target +   
 restmt\_dltt\_target + restmt\_epspi\_target + restmt\_ib\_target +   
 restmt\_ni\_target + restmt\_nopi\_target + restmt\_reuna\_target +   
 restmt\_teq\_target + restmt\_txt\_target + restmt\_wcap\_target +   
 restmt\_xint\_target,   
 data=training\_subset\_ds\_final,   
 method='rf',   
 metric='Accuracy',   
 tuneGrid=rfGrid,   
 trControl=rfControl,   
 ntree = 500  
 )  
  
class\_probabilities <- predict(rf\_model\_3, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
  
rf\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
rf\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
rf\_confusion\_matrix <- confusionMatrix(factor(rf\_class\_probabilities\_litigated), factor(rf\_litigated), positive = "Yes")  
rf\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 65 6  
## Yes 7 5  
##   
## Accuracy : 0.8434   
## 95% CI : (0.7471, 0.9139)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.795   
##   
## Kappa : 0.3441   
##   
## Mcnemar's Test P-Value : 1.000   
##   
## Sensitivity : 0.45455   
## Specificity : 0.90278   
## Pos Pred Value : 0.41667   
## Neg Pred Value : 0.91549   
## Prevalence : 0.13253   
## Detection Rate : 0.06024   
## Detection Prevalence : 0.14458   
## Balanced Accuracy : 0.67866   
##   
## 'Positive' Class : Yes   
##

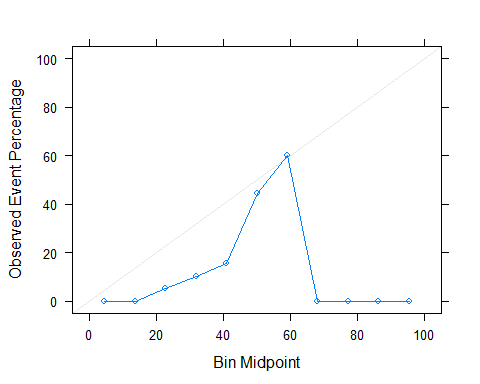
rf\_accuracy\_3 <- rf\_confusion\_matrix$overall["Accuracy"]  
rf\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
rf\_rocr\_roc <- performance(rf\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rf\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



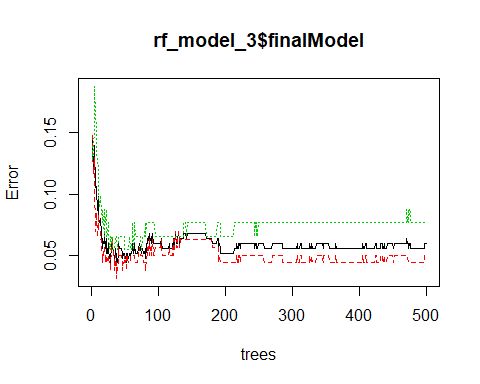
rf\_rocr\_auc <- performance(rf\_rocr\_pred, measure = "auc")  
rf\_auc\_3 <- rf\_rocr\_auc@y.values[[1]]  
rf\_auc\_3

## [1] 0.8598485

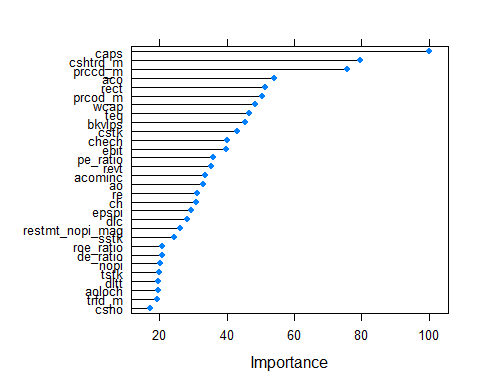
calibration\_curve <- calibration(litigated ~ class\_probabilities\_litigated,  
 data = test\_subset\_ds\_final,  
 class = 1)  
plot(calibration\_curve)



plot(rf\_model\_3$finalModel)



rf\_varImp <- varImp(rf\_model\_3, type = 2)  
plot(rf\_varImp, top = 30)



var\_imp\_ds <- rf\_varImp$importance  
var\_imp\_ds <- var\_imp\_ds %>%   
 as.data.frame() %>%  
 rownames\_to\_column() %>%  
 arrange(desc(Overall))  
var\_imp\_ds

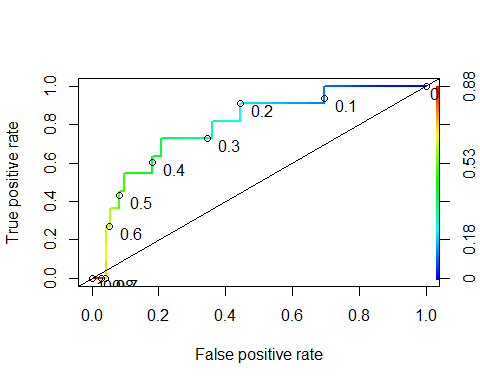
## rowname Overall  
## 1 caps 100.0000000  
## 2 cshtrd\_m 79.4194040  
## 3 prccd\_m 75.6368393  
## 4 aco 54.0231931  
## 5 rect 51.3476987  
## 6 prcod\_m 50.5521603  
## 7 wcap 48.2702510  
## 8 teq 46.5031948  
## 9 bkvlps 45.4259259  
## 10 cstk 43.1477802  
## 11 chech 39.9661312  
## 12 ebit 39.9171856  
## 13 pe\_ratio 35.9751181  
## 14 revt 35.2311891  
## 15 acominc 33.5397407  
## 16 ao 33.0993100  
## 17 re 31.1518851  
## 18 ch 30.9108924  
## 19 epspi 29.3656036  
## 20 dlc 28.3540475  
## 21 restmt\_nopi\_mag 26.2759188  
## 22 sstk 24.4141120  
## 23 roe\_ratio 20.9414244  
## 24 de\_ratio 20.8365221  
## 25 nopi 20.2377197  
## 26 tstk 20.0359429  
## 27 dltt 19.6315370  
## 28 aoloch 19.6117649  
## 29 trfd\_m 19.4111230  
## 30 csho 17.3033236  
## 31 aqc 17.0805341  
## 32 wc\_ratio 17.0044080  
## 33 restmt\_ni\_target 14.5559931  
## 34 restmt\_txt\_target 13.4750466  
## 35 trfm\_m 11.7773451  
## 36 dvt 10.7226866  
## 37 restmt\_epspi\_mag 8.0844478  
## 38 siv 7.7162546  
## 39 restmt\_cogs\_mag 7.4181810  
## 40 sp\_rating\_target 7.3408214  
## 41 restmt\_teq\_mag 5.8596958  
## 42 restmt\_ib\_target 5.8312712  
## 43 restmt\_nopi\_target 4.8757243  
## 44 restmt\_epspi\_target 4.7699079  
## 45 restmt\_teq\_target 4.5993146  
## 46 restmt\_at\_mag 3.8817843  
## 47 restmt\_xint\_mag 3.5342492  
## 48 restmt\_txt\_mag 3.4468243  
## 49 restmt\_ni\_mag 3.1899483  
## 50 restmt\_at\_target 2.5844875  
## 51 restmt\_reuna\_mag 2.3211074  
## 52 restmt\_cogs\_target 1.5612726  
## 53 restmt\_capx\_mag 1.3239253  
## 54 restmt\_capx\_target 0.8517064  
## 55 restmt\_wcap\_mag 0.7616944  
## 56 restmt\_reuna\_target 0.7064857  
## 57 restmt\_xint\_target 0.2319740  
## 58 restmt\_dltt\_target 0.1785672  
## 59 restmt\_wcap\_target 0.0000000

write.csv(var\_imp\_ds, file = "data/var\_imp\_ds.csv", row.names=FALSE)

set.seed(123)  
rf\_model\_4 <- train(litigated ~  
 caps +  
 cshtrd\_m +  
 prccd\_m +   
 aco +  
 rect +  
 prcod\_m +  
 wcap +  
 teq +  
 bkvlps +  
 cstk +  
 chech +  
 ebit +  
 pe\_ratio +  
 revt +  
 acominc +  
 ao +  
 re +  
 ch +  
 epspi +  
 dlc  
 ,   
 data=training\_subset\_ds\_final,   
 method='rf',   
 metric='Accuracy',   
 tuneGrid=rfGrid,   
 trControl=rfControl,   
 ntree = 300)  
  
class\_probabilities <- predict(rf\_model\_4, newdata = test\_subset\_ds\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
test\_subset\_ds\_final$class\_probabilities\_litigated <- class\_probabilities$Yes  
  
rf\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
rf\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
rf\_confusion\_matrix <- confusionMatrix(factor(rf\_class\_probabilities\_litigated), factor(rf\_litigated), positive = "Yes")  
rf\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 66 7  
## Yes 6 4  
##   
## Accuracy : 0.8434   
## 95% CI : (0.7471, 0.9139)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.795   
##   
## Kappa : 0.2915   
##   
## Mcnemar's Test P-Value : 1.000   
##   
## Sensitivity : 0.36364   
## Specificity : 0.91667   
## Pos Pred Value : 0.40000   
## Neg Pred Value : 0.90411   
## Prevalence : 0.13253   
## Detection Rate : 0.04819   
## Detection Prevalence : 0.12048   
## Balanced Accuracy : 0.64015   
##   
## 'Positive' Class : Yes   
##

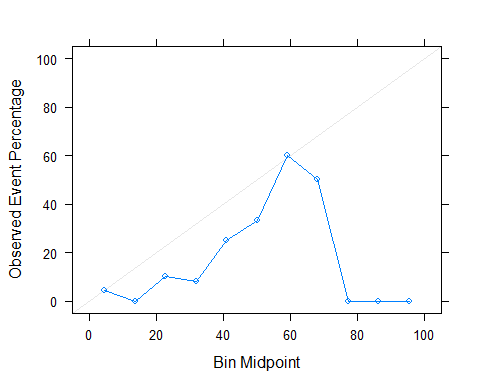
rf\_accuracy\_4 <- rf\_confusion\_matrix$overall["Accuracy"]  
  
  
rf\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
rf\_rocr\_roc <- performance(rf\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(rf\_rocr\_roc,  
colorize = TRUE,  
print.cutoffs.at = seq(0, 1, by = 0.1),  
text.adj = c(-0.5, 1),  
lwd = 2)  
abline(a = 0, b = 1)



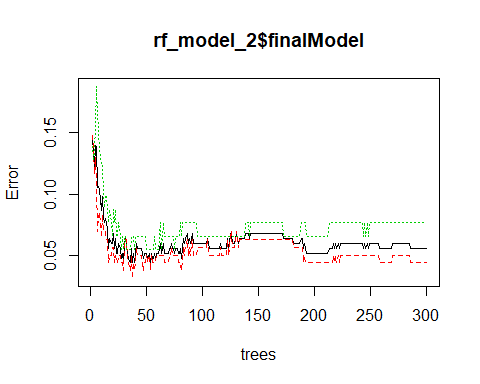
rf\_rocr\_auc <- performance(rf\_rocr\_pred, measure = "auc")  
rf\_auc\_4 <- rf\_rocr\_auc@y.values[[1]]  
rf\_auc\_4

## [1] 0.7954545

calibration\_curve <- calibration(litigated ~ class\_probabilities\_litigated,  
 data = test\_subset\_ds\_final,  
 class = 1)  
plot(calibration\_curve)



plot(rf\_model\_2$finalModel)



auc\_df <- data.frame("algorithm\_method" = c("glm\_1"), "auc" = c(glm\_auc\_1) \* 100, "accuracy" = c(glm\_accuracy\_1) \* 100, "notes" = c("All variables"))  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glm\_2"), "auc" = c(glm\_auc\_2) \* 100, "accuracy" = c(glm\_accuracy\_2) \* 100,   
 "notes" = c("First Set of Signigficant variables variables"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glm\_3"), "auc" = c(glm\_auc\_3) \* 100, "accuracy" = c(glm\_accuracy\_3) \* 100,  
 "notes" = c("Second set of signigficant variables variables"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glmnet\_1"), "auc" = c(glmnet\_auc) \* 100, "accuracy" = c(glmnet\_accuracy\_1) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glmnet\_1"), "auc" = c(glmnet\_auc) \* 100, "accuracy" = c(glmnet\_accuracy\_2) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glmnet\_1"), "auc" = c(glmnet\_auc) \* 100, "accuracy" = c(glmnet\_accuracy\_3) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
   
auc\_df\_temp <- data.frame("algorithm\_method" = c("lda\_1"), "auc" = c(lda\_auc\_1) \* 100, "accuracy" = c(lda\_accuracy\_1) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("gbm\_1"), "auc" = c(gbm\_auc\_1) \* 100, "accuracy" = c(gbm\_accuracy\_1) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("tree\_1"), "auc" = c(tree\_auc) \* 100, "accuracy" = c(tree\_accuracy\_1) \* 100,"notes" = c("Default"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
   
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_1"), "auc" = c(rf\_auc\_1) \* 100, "accuracy" = c(rf\_accuracy\_1) \* 100,"notes" = c("Trees = 100"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_2"), "auc" = c(rf\_auc\_2) \* 100, "accuracy" = c(rf\_accuracy\_2) \* 100,"notes" = c("Trees = 300"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_3"), "auc" = c(rf\_auc\_3) \* 100, "accuracy" = c(rf\_accuracy\_3) \* 100,"notes" = c("Trees = 500"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf\_4"), "auc" = c(rf\_auc\_4) \* 100, "accuracy" = c(rf\_accuracy\_4) \* 100,"notes" = c("Trees = 500 and Only important variables"))  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df <- auc\_df %>%   
 arrange(desc(auc))  
  
auc\_df

## algorithm\_method auc accuracy  
## Accuracy9 rf\_1 88.63636 85.54217  
## Accuracy10 rf\_2 86.42677 84.33735  
## Accuracy11 rf\_3 85.98485 84.33735  
## Accuracy12 rf\_4 79.54545 84.33735  
## Accuracy7 gbm\_1 79.54545 74.69880  
## Accuracy2 glm\_3 75.37879 83.13253  
## Accuracy1 glm\_2 68.56061 81.92771  
## Accuracy3 glmnet\_1 62.75253 77.10843  
## Accuracy4 glmnet\_1 62.75253 75.90361  
## Accuracy5 glmnet\_1 62.75253 83.13253  
## Accuracy8 tree\_1 62.50000 69.87952  
## Accuracy glm\_1 59.53283 69.87952  
## Accuracy6 lda\_1 53.09343 69.87952  
## notes  
## Accuracy9 Trees = 100  
## Accuracy10 Trees = 300  
## Accuracy11 Trees = 500  
## Accuracy12 Trees = 500 and Only important variables  
## Accuracy7 Default  
## Accuracy2 Second set of signigficant variables variables  
## Accuracy1 First Set of Signigficant variables variables  
## Accuracy3 Default  
## Accuracy4 Default  
## Accuracy5 Default  
## Accuracy8 Default  
## Accuracy All variables  
## Accuracy6 Default

target\_company\_row\_final

## # A tibble: 1 x 63  
## # Groups: gvkey [1]  
## gvkey tic aco acominc ao aoloch aqc bkvlps caps ch chech  
## <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 9777 SJM -0.155 0.107 -0.142 -0.121 0.174 -0.0885 1.54 -0.118 -0.649  
## # ... with 52 more variables: csho <dbl>, cstk <dbl>, dlc <dbl>, dltt <dbl>,  
## # dvt <dbl>, ebit <dbl>, epspi <dbl>, nopi <dbl>, re <dbl>, rect <dbl>,  
## # revt <dbl>, siv <dbl>, sstk <dbl>, teq <dbl>, tstk <dbl>, wcap <dbl>,  
## # restmt\_at\_mag <dbl>, restmt\_capx\_mag <dbl>, restmt\_cogs\_mag <dbl>,  
## # restmt\_epspi\_mag <dbl>, restmt\_ni\_mag <dbl>, restmt\_nopi\_mag <dbl>,  
## # restmt\_reuna\_mag <dbl>, restmt\_teq\_mag <dbl>, restmt\_txt\_mag <dbl>,  
## # restmt\_wcap\_mag <dbl>, restmt\_xint\_mag <dbl>, cshtrd\_m <dbl>,  
## # prccd\_m <dbl>, prcod\_m <dbl>, trfd\_m <dbl>, trfm\_m <dbl>, pe\_ratio <dbl>,  
## # wc\_ratio <dbl>, de\_ratio <dbl>, roe\_ratio <dbl>, sp\_rating\_target <dbl>,  
## # restmt\_at\_target <dbl>, restmt\_capx\_target <dbl>, restmt\_cogs\_target <dbl>,  
## # restmt\_dltt\_target <dbl>, restmt\_epspi\_target <dbl>,  
## # restmt\_ib\_target <dbl>, restmt\_ni\_target <dbl>, restmt\_nopi\_target <dbl>,  
## # restmt\_reuna\_target <dbl>, restmt\_teq\_target <dbl>,  
## # restmt\_txt\_target <dbl>, restmt\_wcap\_target <dbl>,  
## # restmt\_xint\_target <dbl>, litigated <fct>, litigation\_settlement <dbl>

class\_probabilities\_prediction <- predict(rf\_model\_4, newdata = target\_company\_row\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
class\_probabilities\_prediction

## No Yes  
## 1 0.7433333 0.2566667

litigated\_probablity <- class\_probabilities\_prediction$Yes  
rf\_litigated <- ifelse(litigated\_probablity > 0.50, "Yes", "No")  
rf\_actual\_litigated\_num <- as.numeric(as.character(target\_company\_row\_final$litigated))  
rf\_actual\_litigated <- ifelse(rf\_actual\_litigated\_num == 1, "Yes", "No")  
rf\_litigated

## [1] "No"

rf\_actual\_litigated

## [1] "No"

class\_probabilities\_prediction <- predict(gbm\_model\_1, newdata = target\_company\_row\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
litigated\_probablity <- class\_probabilities\_prediction$Yes  
gbm\_litigated <- ifelse(litigated\_probablity > 0.50, "Yes", "No")  
gbm\_actual\_litigated\_num <- as.numeric(as.character(target\_company\_row\_final$litigated))  
gbm\_actual\_litigated <- ifelse(rf\_actual\_litigated\_num == 1, "Yes", "No")  
gbm\_litigated

## [1] "No"

gbm\_actual\_litigated

## [1] "No"

class\_probabilities\_prediction <- predict(glm\_model\_3, newdata = target\_company\_row\_final[, -1\*c(target\_var\_index:target\_var\_index)], type = "prob")  
litigated\_probablity <- class\_probabilities\_prediction$Yes  
glm\_litigated <- ifelse(litigated\_probablity > 0.50, "Yes", "No")  
glm\_actual\_litigated\_num <- as.numeric(as.character(target\_company\_row\_final$litigated))  
glm\_actual\_litigated <- ifelse(rf\_actual\_litigated\_num == 1, "Yes", "No")  
glm\_litigated

## [1] "No"

glm\_actual\_litigated

## [1] "No"

nrow(training\_subset\_ds\_final)

## [1] 249

#summary(training\_subset\_ds\_final$litigation\_settlement)  
#training\_litigated <- training\_subset\_ds\_final %>%  
# filter(!is.na(litigation\_settlement))  
#training\_litigated

nrow(test\_subset\_ds\_final)

## [1] 83

summary(test\_subset\_ds\_final)

## gvkey tic aco acominc   
## Min. : 1239 3NUIN : 1 Min. :-0.33000 Min. :-4.90380   
## 1st Qu.: 11506 3RCGP : 1 1st Qu.:-0.32852 1st Qu.: 0.12167   
## Median : 61940 3WTER : 1 Median :-0.31428 Median : 0.14150   
## Mean : 86318 9919B : 1 Mean : 0.09162 Mean :-0.02124   
## 3rd Qu.:165809 ACCA : 1 3rd Qu.:-0.15477 3rd Qu.: 0.14177   
## Max. :277487 ACV : 1 Max. : 8.60339 Max. : 0.73844   
## (Other):77   
## ao aoloch aqc bkvlps   
## Min. :-0.3698 Min. :-6.786225 Min. :-0.2391 Min. :-0.08911   
## 1st Qu.:-0.3690 1st Qu.:-0.070216 1st Qu.:-0.2386 1st Qu.:-0.08896   
## Median :-0.3433 Median :-0.049435 Median :-0.2386 Median :-0.08891   
## Mean : 0.1641 Mean : 0.009719 Mean : 0.1159 Mean :-0.08889   
## 3rd Qu.:-0.1812 3rd Qu.:-0.037898 3rd Qu.:-0.1918 3rd Qu.:-0.08884   
## Max. :10.5592 Max. : 7.459511 Max. :11.8448 Max. :-0.08851   
##   
## caps ch chech csho   
## Min. :-0.3891 Min. :-0.3742 Min. :-1.3414 Min. :-0.38738   
## 1st Qu.:-0.3049 1st Qu.:-0.3715 1st Qu.:-0.2577 1st Qu.:-0.36267   
## Median :-0.2948 Median :-0.3307 Median :-0.2514 Median :-0.31557   
## Mean : 0.4641 Mean : 0.1257 Mean : 0.1235 Mean :-0.04013   
## 3rd Qu.:-0.1789 3rd Qu.:-0.2205 3rd Qu.:-0.1662 3rd Qu.:-0.17047   
## Max. :26.9237 Max. : 9.9659 Max. : 9.7113 Max. : 4.83193   
##   
## cstk dlc dltt dvt   
## Min. :-0.239744 Min. :-0.2440 Min. :-0.3488 Min. :-0.2885   
## 1st Qu.:-0.239711 1st Qu.:-0.2434 1st Qu.:-0.3484 1st Qu.:-0.2885   
## Median :-0.239272 Median :-0.2385 Median :-0.3422 Median :-0.2859   
## Mean :-0.001669 Mean : 0.2062 Mean : 0.1889 Mean : 0.1430   
## 3rd Qu.:-0.198235 3rd Qu.:-0.1684 3rd Qu.:-0.1447 3rd Qu.:-0.2380   
## Max. : 4.639207 Max. :11.5257 Max. :10.0940 Max. : 8.1907   
##   
## ebit epspi nopi re   
## Min. :-0.3855 Min. :-0.67418 Min. :-1.4127 Min. :-0.6285   
## 1st Qu.:-0.3137 1st Qu.:-0.12047 1st Qu.:-0.2242 1st Qu.:-0.2760   
## Median :-0.3016 Median :-0.08776 Median :-0.2230 Median :-0.2665   
## Mean : 0.1187 Mean :-0.06063 Mean : 0.1099 Mean : 0.1468   
## 3rd Qu.:-0.1558 3rd Qu.: 0.01153 3rd Qu.:-0.2020 3rd Qu.:-0.1601   
## Max. : 6.8299 Max. : 0.33152 Max. : 8.4431 Max. :10.3899   
##   
## rect revt siv sstk   
## Min. :-0.35415 Min. :-0.2501 Min. :-0.2110 Min. :-0.32820   
## 1st Qu.:-0.35301 1st Qu.:-0.2489 1st Qu.:-0.2110 1st Qu.:-0.32784   
## Median :-0.32642 Median :-0.2385 Median :-0.2110 Median :-0.31146   
## Mean : 0.02663 Mean :-0.0382 Mean : 0.6190 Mean : 0.02719   
## 3rd Qu.:-0.13780 3rd Qu.:-0.1317 3rd Qu.:-0.2072 3rd Qu.:-0.16962   
## Max. : 3.97877 Max. : 2.5505 Max. :26.7597 Max. :12.97037   
##   
## teq tstk wcap restmt\_at\_mag   
## Min. :-0.49183 Min. :-0.2350 Min. :-4.08605 Min. : -11.4814   
## 1st Qu.:-0.31113 1st Qu.:-0.2350 1st Qu.:-0.20612 1st Qu.: -0.0152   
## Median :-0.29670 Median :-0.2350 Median :-0.18126 Median : -0.0152   
## Mean : 0.06301 Mean : 0.5155 Mean :-0.07864 Mean : 16.9783   
## 3rd Qu.:-0.19796 3rd Qu.:-0.2296 3rd Qu.:-0.04445 3rd Qu.: -0.0152   
## Max. : 8.10690 Max. :26.5936 Max. : 3.03926 Max. :1403.6870   
##   
## restmt\_capx\_mag restmt\_cogs\_mag restmt\_epspi\_mag restmt\_ni\_mag   
## Min. :-0.97047 Min. : -2.4941 Min. :-0.09336 Min. :-0.12098   
## 1st Qu.: 0.06399 1st Qu.: -0.0374 1st Qu.:-0.08352 1st Qu.:-0.06336   
## Median : 0.06399 Median : -0.0374 Median :-0.08352 Median :-0.06336   
## Mean : 0.09371 Mean : 10.2334 Mean :-0.04787 Mean :-0.05245   
## 3rd Qu.: 0.06399 3rd Qu.: -0.0374 3rd Qu.:-0.08352 3rd Qu.:-0.06336   
## Max. : 3.56514 Max. :783.8351 Max. : 2.78327 Max. : 0.82112   
##   
## restmt\_nopi\_mag restmt\_reuna\_mag restmt\_teq\_mag restmt\_txt\_mag   
## Min. :-0.64972 Min. :-9.34673 Min. :-0.09428 Min. :-5.98501   
## 1st Qu.: 0.06557 1st Qu.:-0.06395 1st Qu.:-0.06259 1st Qu.: 0.11658   
## Median : 0.06721 Median :-0.06395 Median :-0.06259 Median : 0.11658   
## Mean : 0.06376 Mean :-0.17673 Mean : 0.01185 Mean : 0.01606   
## 3rd Qu.: 0.06770 3rd Qu.:-0.06395 3rd Qu.:-0.06259 3rd Qu.: 0.11658   
## Max. : 0.47474 Max. :-0.06266 Max. : 5.89963 Max. : 1.49019   
##   
## restmt\_wcap\_mag restmt\_xint\_mag cshtrd\_m prccd\_m   
## Min. :-1.71706 Min. :-2.40275 Min. :-0.43893 Min. :-0.21807   
## 1st Qu.:-0.06281 1st Qu.: 0.18559 1st Qu.:-0.42366 1st Qu.:-0.20636   
## Median :-0.06281 Median : 0.18559 Median :-0.35338 Median :-0.13412   
## Mean :-0.08391 Mean : 0.08985 Mean : 0.07225 Mean :-0.05224   
## 3rd Qu.:-0.06281 3rd Qu.: 0.18559 3rd Qu.:-0.13466 3rd Qu.: 0.05481   
## Max. : 0.14913 Max. : 0.18559 Max. : 5.35907 Max. : 1.09582   
##   
## prcod\_m trfd\_m trfm\_m pe\_ratio   
## Min. :-0.228955 Min. :-0.13469 Min. :-0.29865 Min. :-0.08922   
## 1st Qu.:-0.203907 1st Qu.:-0.13052 1st Qu.:-0.29865 1st Qu.:-0.08919   
## Median :-0.139275 Median :-0.12091 Median :-0.28600 Median :-0.08919   
## Mean :-0.088692 Mean :-0.09278 Mean : 0.02889 Mean :-0.02596   
## 3rd Qu.: 0.009519 3rd Qu.:-0.06942 3rd Qu.:-0.09338 3rd Qu.:-0.08919   
## Max. : 0.328594 Max. : 0.34720 Max. : 7.38589 Max. : 5.13804   
##   
## wc\_ratio de\_ratio roe\_ratio sp\_rating\_target   
## Min. :-0.751606 Min. : -4.43827 Min. :-2.81320 Min. :-0.62493   
## 1st Qu.:-0.519491 1st Qu.: -0.06525 1st Qu.: 0.05101 1st Qu.:-0.62493   
## Median :-0.326251 Median : 0.01136 Median : 0.06219 Median :-0.62493   
## Mean : 0.057698 Mean : 3.91328 Mean : 0.31813 Mean :-0.02906   
## 3rd Qu.: 0.002698 3rd Qu.: 0.17527 3rd Qu.: 0.07079 3rd Qu.: 0.49628   
## Max. : 8.664975 Max. :314.19896 Max. :22.95235 Max. : 1.83438   
##   
## restmt\_at\_target restmt\_capx\_target restmt\_cogs\_target restmt\_dltt\_target  
## Min. :-0.2527 Min. :-0.18183 Min. :-0.64273 Min. :-0.14286   
## 1st Qu.:-0.2527 1st Qu.:-0.18183 1st Qu.:-0.64273 1st Qu.:-0.14286   
## Median :-0.2527 Median :-0.18183 Median :-0.64273 Median :-0.14286   
## Mean : 0.1011 Mean :-0.04546 Mean : 0.09685 Mean :-0.05715   
## 3rd Qu.:-0.2527 3rd Qu.:-0.18183 3rd Qu.: 1.54961 3rd Qu.:-0.14286   
## Max. : 3.9417 Max. : 5.47759 Max. : 1.54961 Max. : 6.97166   
##   
## restmt\_epspi\_target restmt\_ib\_target restmt\_ni\_target restmt\_nopi\_target  
## Min. :-0.3763 Min. :-0.35523 Min. :-0.1933 Min. :-1.2815   
## 1st Qu.:-0.3763 1st Qu.:-0.35523 1st Qu.:-0.1933 1st Qu.:-1.2815   
## Median :-0.3763 Median :-0.35523 Median :-0.1933 Median : 0.7772   
## Mean : 0.2064 Mean :-0.01269 Mean : 0.2577 Mean : 0.1571   
## 3rd Qu.:-0.3763 3rd Qu.:-0.35523 3rd Qu.:-0.1933 3rd Qu.: 0.7772   
## Max. : 2.6465 Max. : 2.80378 Max. : 5.1536 Max. : 0.7772   
##   
## restmt\_reuna\_target restmt\_teq\_target restmt\_txt\_target restmt\_wcap\_target  
## Min. :-0.2786 Min. :-0.30288 Min. :-0.2786 Min. :-0.1933   
## 1st Qu.:-0.2786 1st Qu.:-0.30288 1st Qu.:-0.2786 1st Qu.:-0.1933   
## Median :-0.2786 Median :-0.30288 Median :-0.2786 Median :-0.1933   
## Mean : 0.1393 Mean : 0.08654 Mean : 0.0000 Mean : 0.1933   
## 3rd Qu.:-0.2786 3rd Qu.:-0.30288 3rd Qu.:-0.2786 3rd Qu.:-0.1933   
## Max. : 3.5752 Max. : 3.28839 Max. : 3.5752 Max. : 5.1536   
##   
## restmt\_xint\_target litigated litigation\_settlement  
## Min. :-0.3763 0:72 Min. : 2700000   
## 1st Qu.:-0.3763 1:11 1st Qu.: 4775000   
## Median :-0.3763 Median : 6850000   
## Mean :-0.1214 Mean : 6850000   
## 3rd Qu.:-0.3763 3rd Qu.: 8925000   
## Max. : 2.6465 Max. :11000000   
## NA's :81   
## class\_probabilities\_litigated class\_probabilities\_onehot  
## Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.08333 1st Qu.:0.04209   
## Median :0.20000 Median :0.09622   
## Mean :0.25454 Mean :0.31689   
## 3rd Qu.:0.38500 3rd Qu.:0.50571   
## Max. :0.87000 Max. :1.00000   
##

library(Rmisc)

## Warning: package 'Rmisc' was built under R version 3.6.3

## Loading required package: plyr

## Warning: package 'plyr' was built under R version 3.6.3

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following object is masked from 'package:purrr':  
##   
## compact

settlement\_training\_ds <- training\_subset\_ds\_final\_settlement %>%  
 filter(!is.na(litigation\_settlement))  
  
settlement\_test\_ds <- test\_subset\_ds\_final\_settlement %>%  
 filter(!is.na(litigation\_settlement))  
  
settlement\_ds <- rbind(settlement\_training\_ds,settlement\_test\_ds)  
settlement\_ds

## # A tibble: 10 x 63  
## # Groups: gvkey [10]  
## gvkey tic aco acominc ao aoloch aqc bkvlps caps ch  
## <int> <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 3505 TAP 126. 31. 301. -84.0 580. 43.8 3623. 841.   
## 2 5301 GAPTQ 39.9 -77.4 143. -42.2 0 -16.0 505. 303.   
## 3 7691 NAFC 23.5 -13.8 10.2 -1.25 29.1 29.5 116. 0.965  
## 4 12785 PPC 98.2 -46.0 105. 31.4 0 4.21 1545. 181.   
## 5 24316 MNST 52.3 -0.106 65.4 -10.5 0 7.60 268. 287.   
## 6 112968 MTEX 5.17 -0.789 7.55 0.881 0 4.71 42.4 18.6   
## 7 149379 SPU 0.568 13.9 14.3 -2.43 0 5.59 59.2 63.7   
## 8 178795 GRO 116. -18.7 11.9 1.82 0.693 3.44 356. 34.6   
## 9 114959 YUII 0.338 3.75 43.3 1.18 0 5.32 115. 34.5   
## 10 163627 DMND 26.8 6.25 25.7 -1.20 154. 14.6 322. 4.48   
## # ... with 53 more variables: chech <dbl>, csho <dbl>, cstk <dbl>, dlc <dbl>,  
## # dltt <dbl>, dvt <dbl>, ebit <dbl>, epspi <dbl>, nopi <dbl>, re <dbl>,  
## # rect <dbl>, revt <dbl>, siv <dbl>, sstk <dbl>, teq <dbl>, tstk <dbl>,  
## # wcap <dbl>, restmt\_at\_mag <dbl>, restmt\_capx\_mag <dbl>,  
## # restmt\_cogs\_mag <dbl>, restmt\_epspi\_mag <dbl>, restmt\_ni\_mag <dbl>,  
## # restmt\_nopi\_mag <dbl>, restmt\_reuna\_mag <dbl>, restmt\_teq\_mag <dbl>,  
## # restmt\_txt\_mag <dbl>, restmt\_wcap\_mag <dbl>, restmt\_xint\_mag <dbl>,  
## # cshtrd\_m <dbl>, prccd\_m <dbl>, prcod\_m <dbl>, trfd\_m <dbl>, trfm\_m <dbl>,  
## # pe\_ratio <dbl>, wc\_ratio <dbl>, de\_ratio <dbl>, roe\_ratio <dbl>,  
## # sp\_rating\_target <dbl>, restmt\_at\_target <dbl>, restmt\_capx\_target <dbl>,  
## # restmt\_cogs\_target <dbl>, restmt\_dltt\_target <dbl>,  
## # restmt\_epspi\_target <dbl>, restmt\_ib\_target <dbl>, restmt\_ni\_target <dbl>,  
## # restmt\_nopi\_target <dbl>, restmt\_reuna\_target <dbl>,  
## # restmt\_teq\_target <dbl>, restmt\_txt\_target <dbl>, restmt\_wcap\_target <dbl>,  
## # restmt\_xint\_target <dbl>, litigated <fct>, litigation\_settlement <dbl>

settlement\_ds\_subset <- subset(settlement\_ds, select = c(gvkey,tic, csho, prccd\_m, prcod\_m,litigated,litigation\_settlement))  
settlement\_ds\_subset$stk\_price\_m <- (settlement\_ds\_subset$prccd\_m + settlement\_ds\_subset$prcod\_m)/2  
settlement\_ds\_subset$market\_cap <- settlement\_ds\_subset$stk\_price\_m \* settlement\_ds\_subset$csho \* 1000000  
settlement\_ds\_subset$settlement\_ratio <- as.numeric(settlement\_ds\_subset$litigation\_settlement/settlement\_ds\_subset$market\_cap)  
settlement\_ds\_subset

## # A tibble: 10 x 10  
## # Groups: gvkey [10]  
## gvkey tic csho prccd\_m prcod\_m litigated litigation\_sett~ stk\_price\_m  
## <int> <fct> <dbl> <dbl> <dbl> <fct> <dbl> <dbl>  
## 1 3505 TAP 183. 47.2 40.0 1 6000000 43.6   
## 2 5301 GAPTQ 54.9 6.99 6.99 1 9000000 6.99  
## 3 7691 NAFC 12.2 29.5 29.5 1 6750000 29.5   
## 4 12785 PPC 237. 9.68 9.67 1 1500000 9.68  
## 5 24316 MNST 127. 61.2 61.1 1 47500000 61.1   
## 6 112968 MTEX 14.6 6.96 6.85 1 11250000 6.91  
## 7 149379 SPU 26.2 2.93 2.94 1 2200000 2.93  
## 8 178795 GRO 55.4 1.31 1.32 1 3750000 1.31  
## 9 114959 YUII 20.2 2.97 4.42 1 2700000 3.70  
## 10 163627 DMND 22.1 35.1 35.1 1 11000000 35.1   
## # ... with 2 more variables: market\_cap <dbl>, settlement\_ratio <dbl>

#CI(settlement\_ds\_subset$litigation\_settlement, ci=0.90)

settlement\_gvkey <- settlement\_ds\_subset$gvkey  
settlement\_gvkey

## [1] 3505 5301 7691 12785 24316 112968 149379 178795 114959 163627

fundamentals\_ds\_settlement <- fundamentals\_ds\_filter %>%  
 filter(gvkey %in% settlement\_gvkey)  
  
fundamentals\_ds\_settlement\_1 <- subset(fundamentals\_ds\_settlement, select = c(gvkey,tic, mkvalt))  
  
fundamentals\_ds\_settlement\_2 <- fundamentals\_ds\_settlement\_1 %>%  
 filter(!is.na(mkvalt))   
  
fundamentals\_ds\_settlement\_3 <- fundamentals\_ds\_settlement\_2 %>%  
 group\_by(gvkey) %>%  
 dplyr::summarize(  
 market\_value = mean(mkvalt)  
 )

## `summarise()` ungrouping output (override with `.groups` argument)

settlement\_ds\_subset$market\_value <- 0  
for (row in 1:nrow(settlement\_ds\_subset)){  
 settlement\_item\_gvkey <- as.integer(settlement\_ds\_subset[row, "gvkey"])  
 fundamental\_settlement\_row <- fundamentals\_ds\_settlement\_3 %>%  
 filter(gvkey == settlement\_item\_gvkey)  
 market\_value <- as.numeric(fundamental\_settlement\_row["market\_value"])  
 settlement\_ds\_subset$market\_value[settlement\_ds\_subset$gvkey == settlement\_item\_gvkey] <- market\_value \* 1000000  
}  
  
settlement\_ds\_subset$market\_value[settlement\_ds\_subset$tic == 'GRO'] <- as.numeric(72720449)  
  
settlement\_ds\_subset$settlement\_ratio\_2 <- as.numeric(settlement\_ds\_subset$litigation\_settlement/settlement\_ds\_subset$market\_value)  
  
settlement\_ds\_subset

## # A tibble: 10 x 12  
## # Groups: gvkey [10]  
## gvkey tic csho prccd\_m prcod\_m litigated litigation\_sett~ stk\_price\_m  
## <int> <fct> <dbl> <dbl> <dbl> <fct> <dbl> <dbl>  
## 1 3505 TAP 183. 47.2 40.0 1 6000000 43.6   
## 2 5301 GAPTQ 54.9 6.99 6.99 1 9000000 6.99  
## 3 7691 NAFC 12.2 29.5 29.5 1 6750000 29.5   
## 4 12785 PPC 237. 9.68 9.67 1 1500000 9.68  
## 5 24316 MNST 127. 61.2 61.1 1 47500000 61.1   
## 6 112968 MTEX 14.6 6.96 6.85 1 11250000 6.91  
## 7 149379 SPU 26.2 2.93 2.94 1 2200000 2.93  
## 8 178795 GRO 55.4 1.31 1.32 1 3750000 1.31  
## 9 114959 YUII 20.2 2.97 4.42 1 2700000 3.70  
## 10 163627 DMND 22.1 35.1 35.1 1 11000000 35.1   
## # ... with 4 more variables: market\_cap <dbl>, settlement\_ratio <dbl>,  
## # market\_value <dbl>, settlement\_ratio\_2 <dbl>

#CI(settlement\_ds\_subset$litigation\_settlement, ci=0.95)  
settlement\_ratio\_mean <- mean(settlement\_ds\_subset$settlement\_ratio\_2)  
settlement\_ratio\_medain <- median(settlement\_ds\_subset$settlement\_ratio\_2)  
settlement\_ratio\_mean

## [1] 0.05577476

settlement\_ratio\_medain

## [1] 0.01638755

target\_company\_ds\_settlement <- fundamentals\_ds\_filter %>%  
 filter(tic == 'SJM')  
target\_company\_market\_value <- mean(target\_company\_ds\_settlement$mkvalt) \* 1000000  
target\_company\_settlement\_amt <- target\_company\_market\_value \* settlement\_ratio\_medain  
target\_company\_market\_value

## [1] 9090440580

settlement\_ratio\_medain

## [1] 0.01638755

target\_company\_settlement\_amt

## [1] 148970013

settlement\_1 <- settlement\_ds\_subset$litigation\_settlement  
#inference(settlement\_1, est = "mean", type = "ci", method = "theoretical")  
std\_error <- function(x) sd(x)/sqrt(length(x))  
std\_error\_settlement\_amout <- std\_error(settlement\_1)  
std\_error\_settlement\_amout

## [1] 4296698

settlement\_amt\_high <- target\_company\_settlement\_amt + (1.96 \* std\_error\_settlement\_amout)  
settlement\_amt\_low <- target\_company\_settlement\_amt - (1.96 \* std\_error\_settlement\_amout)   
settlement\_amt\_high

## [1] 157391541

settlement\_amt\_low

## [1] 140548485

settlement\_1\_sd = sd(settlement\_1)  
settlement\_1\_mean = mean(settlement\_1)  
settlement\_1\_sd

## [1] 13587352

settlement\_1\_mean

## [1] 10165000

z\_value <- 1.96  
n <- 11  
#inference(settlement\_1, est = "mean", type = "ci", method = "theoretical")  
#inference(settlement\_1, est = "median", type = "ci", method = "simulation")