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

### 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

### 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

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="")

ncol(fundamentals\_ds\_filter\_1)

## [1] 147

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) %>%  
 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,csho,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,csho,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 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   
##   
## 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

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) %>%  
 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)

### 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 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   
##   
## 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"  
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) %>%  
 summarise(  
 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) %>%  
 summarise(  
 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)

### 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)  
#summary(ds\_final)

### 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]  
}

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

training\_subset\_ds\_final\_1 <- subset(training\_set, select = -c(gvkey,tic, 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(gvkey,tic, 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())  
  
  
  
training\_subset\_ds\_final <- training\_subset\_ds\_final\_1  
test\_subset\_ds\_final <- test\_subset\_ds\_final\_1  
  
  
ncol(training\_subset\_ds\_final)

## [1] 60

summary(training\_subset\_ds\_final)

## aco acominc ao aoloch   
## Min. : 0.000 Min. :-19306.57 Min. : 0.000 Min. :-465.250   
## 1st Qu.: 0.354 1st Qu.: -27.16 1st Qu.: 0.081 1st Qu.: -1.718   
## Median : 8.418 Median : 0.00 Median : 7.787 Median : 0.000   
## Mean : 175.533 Mean : -186.69 Mean : 180.074 Mean : 4.847   
## 3rd Qu.: 93.621 3rd Qu.: 0.00 3rd Qu.: 94.202 3rd Qu.: 1.499   
## Max. :4706.135 Max. : 3495.34 Max. :5132.600 Max. : 655.000   
## aqc bkvlps caps ch   
## Min. : -12.45 Min. : -82.3 Min. : -701.475 Min. : 0.000   
## 1st Qu.: 0.00 1st Qu.: 0.1 1st Qu.: 6.589 1st Qu.: 1.211   
## Median : 0.00 Median : 3.6 Median : 43.450 Median : 18.311   
## Mean : 97.38 Mean : 8306.8 Mean : 720.357 Mean : 362.988   
## 3rd Qu.: 14.22 3rd Qu.: 12.6 3rd Qu.: 466.961 3rd Qu.: 204.962   
## Max. :5559.02 Max. :1216746.9 Max. :28658.250 Max. :7382.800   
## chech cshi cstk dlc   
## Min. :-305.7500 Min. : 0.583 Min. : 0.000 Min. : 0.000   
## 1st Qu.: -0.1546 1st Qu.: 19.236 1st Qu.: 0.036 1st Qu.: 0.272   
## Median : 0.4164 Median : 53.045 Median : 0.274 Median : 4.258   
## Mean : 39.4348 Mean : 266.290 Mean : 196.281 Mean : 310.500   
## 3rd Qu.: 10.6056 3rd Qu.: 147.925 3rd Qu.: 14.294 3rd Qu.: 101.126   
## Max. :1358.0000 Max. :6253.511 Max. :7290.750 Max. :15926.126   
## dltt dvt ebit epspi   
## Min. : 0.00 Min. : -0.006 Min. : -208.760 Min. :-14.0200   
## 1st Qu.: 0.03 1st Qu.: 0.000 1st Qu.: -0.344 1st Qu.: -0.0550   
## Median : 15.04 Median : 0.000 Median : 20.195 Median : 0.2442   
## Mean : 1310.28 Mean : 203.193 Mean : 700.909 Mean : 1.7342   
## 3rd Qu.: 939.51 3rd Qu.: 47.886 3rd Qu.: 351.478 3rd Qu.: 1.8613   
## Max. :42659.60 Max. :6572.535 Max. :24345.400 Max. :230.7025   
## nopi re rect revt   
## Min. : -38.5478 Min. :-7570.29 Min. : 0.00 Min. : 0.0   
## 1st Qu.: -0.0012 1st Qu.: -11.13 1st Qu.: 1.48 1st Qu.: 16.1   
## Median : 0.1845 Median : 18.93 Median : 24.00 Median : 311.6   
## Mean : 43.2791 Mean : 1700.90 Mean : 498.48 Mean : 7926.1   
## 3rd Qu.: 4.7685 3rd Qu.: 428.38 3rd Qu.: 335.30 3rd Qu.: 3845.0   
## Max. :2224.4000 Max. :68884.60 Max. :15020.07 Max. :442511.4   
## siv sstk teq tstk   
## Min. : 0.0000 Min. : 0.0000 Min. :-2208.96 Min. : 0.000   
## 1st Qu.: 0.0000 1st Qu.: 0.0014 1st Qu.: 3.49 1st Qu.: 0.000   
## Median : 0.0000 Median : 1.0247 Median : 103.41 Median : 0.000   
## Mean : 34.0780 Mean : 37.2116 Mean : 2445.19 Mean : 589.285   
## 3rd Qu.: 0.4626 3rd Qu.: 13.1662 3rd Qu.: 1124.14 3rd Qu.: 8.367   
## Max. :1622.0000 Max. :1057.0000 Max. :76602.80 Max. :25036.250   
## wcap restmt\_at\_mag restmt\_capx\_mag restmt\_cogs\_mag   
## Min. :-8236.800 Min. :-1.011750 Min. :-22.7162 Min. :-50.0000   
## 1st Qu.: -0.011 1st Qu.: 0.000000 1st Qu.: 0.0000 1st Qu.: 0.0000   
## Median : 23.347 Median : 0.000000 Median : 0.0000 Median : 0.0000   
## Mean : 266.289 Mean : 0.001964 Mean : -0.1053 Mean : 0.4418   
## 3rd Qu.: 297.249 3rd Qu.: 0.000000 3rd Qu.: 0.0000 3rd Qu.: 0.0000   
## Max. :12261.750 Max. : 1.275750 Max. : 8.3335 Max. :100.0000   
## restmt\_epspi\_mag restmt\_ni\_mag restmt\_nopi\_mag restmt\_reuna\_mag   
## Min. : -50.0 Min. : -6.611 Min. :-1868600.0 Min. : -85.20   
## 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: -78.6 1st Qu.: 0.00   
## Median : 0.0 Median : 0.000 Median : 0.0 Median : 0.00   
## Mean : 422.8 Mean : 10.733 Mean : -7953.8 Mean : 16.89   
## 3rd Qu.: 0.0 3rd Qu.: 0.000 3rd Qu.: 14.5 3rd Qu.: 0.00   
## Max. :77081.7 Max. :2683.890 Max. : 68865.1 Max. :4181.70   
## restmt\_teq\_mag restmt\_txt\_mag restmt\_wcap\_mag restmt\_xint\_mag   
## Min. : -105.39 Min. :-88.7704 Min. : -3.149 Min. :-62.735   
## 1st Qu.: 0.00 1st Qu.: 0.0000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 0.00 Median : 0.0000 Median : 0.000 Median : 0.000   
## Mean : 49.55 Mean : -0.9757 Mean : 1.635 Mean : -1.039   
## 3rd Qu.: 0.00 3rd Qu.: 0.0000 3rd Qu.: 0.000 3rd Qu.: 0.000   
## Max. :12541.75 Max. : 17.8622 Max. :412.500 Max. : 0.562   
## cshtrd\_m prccd\_m prcod\_m trfd\_m   
## Min. : 0 Min. : 0.0018 Min. : 0.0362 Min. : 1.000   
## 1st Qu.: 18102 1st Qu.: 1.1236 1st Qu.: 3.2153 1st Qu.: 1.063   
## Median : 113102 Median : 8.7208 Median : 11.3416 Median : 1.231   
## Mean : 871492 Mean : 32.5869 Mean : 40.3436 Mean : 3.013   
## 3rd Qu.: 614661 3rd Qu.: 30.9156 3rd Qu.: 34.9277 3rd Qu.: 1.844   
## Max. :13129451 Max. :2217.8253 Max. :2217.8208 Max. :218.416   
## trfm\_m pe\_ratio wc\_ratio de\_ratio   
## Min. : 1.000 Min. : -1220 Min. : 0.0000 Min. :-60.4827   
## 1st Qu.: 1.000 1st Qu.: -5 1st Qu.: 0.3720 1st Qu.: 0.2062   
## Median : 1.000 Median : 12 Median : 0.7177 Median : 0.7397   
## Mean : 2.015 Mean : 502470 Mean : 1.2751 Mean : 0.8107   
## 3rd Qu.: 1.494 3rd Qu.: 22 3rd Qu.: 1.3550 3rd Qu.: 1.4753   
## Max. :34.527 Max. :76903023 Max. :12.4471 Max. : 80.2000   
## roe\_ratio sp\_rating\_target restmt\_at\_target restmt\_capx\_target  
## Min. :-192.20000 Min. :0.1226 Min. :0.1278 Min. :0.1341   
## 1st Qu.: 0.00089 1st Qu.:0.1226 1st Qu.:0.1278 1st Qu.:0.1341   
## Median : 0.11124 Median :0.1226 Median :0.1278 Median :0.1341   
## Mean : -0.67740 Mean :0.1292 Mean :0.1292 Mean :0.1343   
## 3rd Qu.: 0.25495 3rd Qu.:0.1343 3rd Qu.:0.1278 3rd Qu.:0.1341   
## Max. : 10.15006 Max. :0.1483 Max. :0.1509 Max. :0.1395   
## restmt\_cogs\_target restmt\_dltt\_target restmt\_epspi\_target restmt\_ib\_target  
## Min. :0.1311 Min. :0.1354 Min. :0.1258 Min. :0.1250   
## 1st Qu.:0.1311 1st Qu.:0.1354 1st Qu.:0.1258 1st Qu.:0.1250   
## Median :0.1311 Median :0.1354 Median :0.1258 Median :0.1250   
## Mean :0.1344 Mean :0.1354 Mean :0.1292 Mean :0.1283   
## 3rd Qu.:0.1424 3rd Qu.:0.1354 3rd Qu.:0.1258 3rd Qu.:0.1250   
## Max. :0.1424 Max. :0.1373 Max. :0.1530 Max. :0.1547   
## restmt\_ni\_target restmt\_nopi\_target restmt\_reuna\_target restmt\_teq\_target  
## Min. :0.1263 Min. :0.1279 Min. :0.1286 Min. :0.1273   
## 1st Qu.:0.1263 1st Qu.:0.1279 1st Qu.:0.1286 1st Qu.:0.1273   
## Median :0.1263 Median :0.1429 Median :0.1286 Median :0.1273   
## Mean :0.1273 Mean :0.1372 Mean :0.1301 Mean :0.1294   
## 3rd Qu.:0.1263 3rd Qu.:0.1429 3rd Qu.:0.1286 3rd Qu.:0.1273   
## Max. :0.1544 Max. :0.1429 Max. :0.1493 Max. :0.1513   
## restmt\_txt\_target restmt\_wcap\_target restmt\_xint\_target litigated  
## Min. :0.1266 Min. :0.1324 Min. :0.1303 0:216   
## 1st Qu.:0.1266 1st Qu.:0.1324 1st Qu.:0.1303 1: 34   
## Median :0.1266 Median :0.1324 Median :0.1303   
## Mean :0.1285 Mean :0.1328 Mean :0.1323   
## 3rd Qu.:0.1266 3rd Qu.:0.1324 3rd Qu.:0.1303   
## Max. :0.1530 Max. :0.1429 Max. :0.1454

### Scale the variables.

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

## [1] 1

num\_var\_end\_index

## [1] 59

target\_var\_index

## [1] 60

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

## aco acominc ao aoloch   
## Min. :-0.3300 Min. :-14.4762 Min. :-0.3699 Min. :-4.75393   
## 1st Qu.:-0.3294 1st Qu.: 0.1208 1st Qu.:-0.3698 1st Qu.:-0.06639   
## Median :-0.3142 Median : 0.1413 Median :-0.3539 Median :-0.04901   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.:-0.1540 3rd Qu.: 0.1413 3rd Qu.:-0.1764 3rd Qu.:-0.03386   
## Max. : 8.5182 Max. : 2.7878 Max. :10.1746 Max. : 6.57477   
## aqc bkvlps caps ch   
## Min. :-0.2704 Min. :-0.08967 Min. :-0.6157 Min. :-0.3744   
## 1st Qu.:-0.2398 1st Qu.:-0.08879 1st Qu.:-0.3091 1st Qu.:-0.3732   
## Median :-0.2398 Median :-0.08875 Median :-0.2931 Median :-0.3555   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2047 3rd Qu.:-0.08865 3rd Qu.:-0.1097 3rd Qu.:-0.1630   
## Max. :13.4481 Max. :12.91671 Max. :12.0989 Max. : 7.2411   
## chech cshi cstk dlc   
## Min. :-2.2327 Min. :-0.3969 Min. :-0.2394 Min. :-0.2438   
## 1st Qu.:-0.2561 1st Qu.:-0.3691 1st Qu.:-0.2393 1st Qu.:-0.2435   
## Median :-0.2524 Median :-0.3185 Median :-0.2390 Median :-0.2404   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1865 3rd Qu.:-0.1768 3rd Qu.:-0.2219 3rd Qu.:-0.1644   
## Max. : 8.5286 Max. : 8.9438 Max. : 8.6522 Max. :12.2588   
## dltt dvt ebit epspi   
## Min. :-0.34979 Min. :-0.2891 Min. :-0.4083 Min. :-1.073529   
## 1st Qu.:-0.34978 1st Qu.:-0.2891 1st Qu.:-0.3148 1st Qu.:-0.121922   
## Median :-0.34578 Median :-0.2891 Median :-0.3055 Median :-0.101530   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.000000   
## 3rd Qu.:-0.09898 3rd Qu.:-0.2210 3rd Qu.:-0.1568 3rd Qu.: 0.008656   
## Max. :11.03852 Max. : 9.0634 Max. :10.6127 Max. :15.602429   
## nopi re rect revt   
## Min. :-0.4232 Min. :-1.5004 Min. :-0.3543 Min. :-0.2502   
## 1st Qu.:-0.2238 1st Qu.:-0.2771 1st Qu.:-0.3533 1st Qu.:-0.2497   
## Median :-0.2229 Median :-0.2722 Median :-0.3373 Median :-0.2404   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.1992 3rd Qu.:-0.2059 3rd Qu.:-0.1160 3rd Qu.:-0.1288   
## Max. :11.2794 Max. :10.8724 Max. :10.3218 Max. :13.7205   
## siv sstk teq tstk   
## Min. :-0.2109 Min. :-0.3277 Min. :-0.5989 Min. :-0.2345   
## 1st Qu.:-0.2109 1st Qu.:-0.3277 1st Qu.:-0.3142 1st Qu.:-0.2345   
## Median :-0.2109 Median :-0.3187 Median :-0.3013 Median :-0.2345   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.2080 3rd Qu.:-0.2117 3rd Qu.:-0.1700 3rd Qu.:-0.2312   
## Max. : 9.8269 Max. : 8.9801 Max. : 9.5425 Max. : 9.7295   
## wcap restmt\_at\_mag restmt\_capx\_mag restmt\_cogs\_mag   
## Min. :-6.65259 Min. :-7.81318 Min. :-13.71602 Min. :-4.26046   
## 1st Qu.:-0.20835 1st Qu.:-0.01514 1st Qu.: 0.06387 1st Qu.:-0.03732   
## Median :-0.19007 Median :-0.01514 Median : 0.06387 Median :-0.03732   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.02422 3rd Qu.:-0.01514 3rd Qu.: 0.06387 3rd Qu.:-0.03732   
## Max. : 9.38493 Max. : 9.81768 Max. : 5.11904 Max. : 8.40895   
## restmt\_epspi\_mag restmt\_ni\_mag restmt\_nopi\_mag restmt\_reuna\_mag   
## Min. :-0.09321 Min. :-0.10218 Min. :-15.70772 Min. :-0.38575   
## 1st Qu.:-0.08335 1st Qu.:-0.06323 1st Qu.: 0.06648 1st Qu.:-0.06382   
## Median :-0.08335 Median :-0.06323 Median : 0.06715 Median :-0.06382   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.08335 3rd Qu.:-0.06323 3rd Qu.: 0.06727 3rd Qu.:-0.06382   
## Max. :15.11354 Max. :15.74804 Max. : 0.64851 Max. :15.73665   
## restmt\_teq\_mag restmt\_txt\_mag restmt\_wcap\_mag restmt\_xint\_mag   
## Min. :-0.19531 Min. :-10.4681 Min. :-0.18338 Min. :-11.0797   
## 1st Qu.:-0.06246 1st Qu.: 0.1163 1st Qu.:-0.06268 1st Qu.: 0.1866   
## Median :-0.06246 Median : 0.1163 Median :-0.06268 Median : 0.1866   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.:-0.06246 3rd Qu.: 0.1163 3rd Qu.:-0.06268 3rd Qu.: 0.1866   
## Max. :15.74736 Max. : 2.2461 Max. :15.74706 Max. : 0.2875   
## cshtrd\_m prccd\_m prcod\_m trfd\_m   
## Min. :-0.4395 Min. :-0.21984 Min. :-0.23136 Min. :-0.13470   
## 1st Qu.:-0.4303 1st Qu.:-0.21227 1st Qu.:-0.21311 1st Qu.:-0.13046   
## Median :-0.3824 Median :-0.16102 Median :-0.16647 Median :-0.11926   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.1295 3rd Qu.:-0.01128 3rd Qu.:-0.03109 3rd Qu.:-0.07823   
## Max. : 6.1812 Max. :14.74326 Max. :12.49833 Max. :14.41506   
## trfm\_m pe\_ratio wc\_ratio de\_ratio   
## Min. :-0.3013 Min. :-0.08923 Min. :-0.75168 Min. :-8.88781   
## 1st Qu.:-0.3013 1st Qu.:-0.08902 1st Qu.:-0.53238 1st Qu.:-0.08766   
## Median :-0.3013 Median :-0.08901 Median :-0.32862 Median :-0.01030   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.:-0.1547 3rd Qu.:-0.08901 3rd Qu.: 0.04709 3rd Qu.: 0.09636   
## Max. : 9.6518 Max. :13.53480 Max. : 6.58569 Max. :11.51178   
## roe\_ratio sp\_rating\_target restmt\_at\_target restmt\_capx\_target  
## Min. :-15.46568 Min. :-0.6293 Min. :-0.2521 Min. :-0.1815   
## 1st Qu.: 0.05477 1st Qu.:-0.6293 1st Qu.:-0.2521 1st Qu.:-0.1815   
## Median : 0.06368 Median :-0.6293 Median :-0.2521 Median :-0.1815   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.07529 3rd Qu.: 0.4867 3rd Qu.:-0.2521 3rd Qu.:-0.1815   
## Max. : 0.87433 Max. : 1.8185 Max. : 3.9502 Max. : 5.4890   
## restmt\_cogs\_target restmt\_dltt\_target restmt\_epspi\_target restmt\_ib\_target   
## Min. :-0.6409 Min. :-0.1426 Min. :-0.3755 Min. :-0.3544   
## 1st Qu.:-0.6409 1st Qu.:-0.1426 1st Qu.:-0.3755 1st Qu.:-0.3544   
## Median :-0.6409 Median :-0.1426 Median :-0.3755 Median :-0.3544   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 1.5540 3rd Qu.:-0.1426 3rd Qu.:-0.3755 3rd Qu.:-0.3544   
## Max. : 1.5540 Max. : 6.9860 Max. : 2.6526 Max. : 2.8101   
## restmt\_ni\_target restmt\_nopi\_target restmt\_reuna\_target restmt\_teq\_target  
## Min. :-0.1929 Min. :-1.2857 Min. :-0.278 Min. :-0.3022   
## 1st Qu.:-0.1929 1st Qu.:-1.2857 1st Qu.:-0.278 1st Qu.:-0.3022   
## Median :-0.1929 Median : 0.7747 Median :-0.278 Median :-0.3022   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.000 Mean : 0.0000   
## 3rd Qu.:-0.1929 3rd Qu.: 0.7747 3rd Qu.:-0.278 3rd Qu.:-0.3022   
## Max. : 5.1644 Max. : 0.7747 Max. : 3.583 Max. : 3.2956   
## restmt\_txt\_target restmt\_wcap\_target restmt\_xint\_target litigated  
## Min. :-0.278 Min. :-0.1929 Min. :-0.3824 No :216   
## 1st Qu.:-0.278 1st Qu.:-0.1929 1st Qu.:-0.3824 Yes: 34   
## Median :-0.278 Median :-0.1929 Median :-0.3824   
## Mean : 0.000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.:-0.278 3rd Qu.:-0.1929 3rd Qu.:-0.3824   
## Max. : 3.583 Max. : 5.1644 Max. : 2.6049

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

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

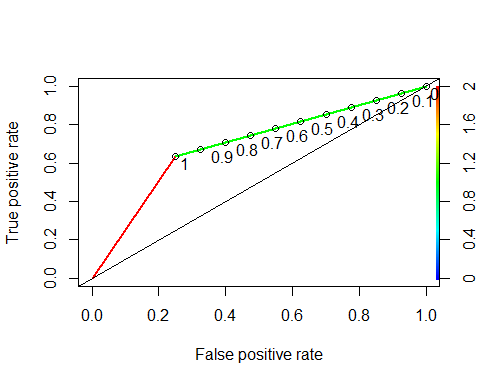
##   
## No Yes   
## 216 34

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)  
# 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)$data  
##### Sampling both (Over)  
#training\_subset\_ds\_final <- ovun.sample(litigated ~., data = training\_subset\_ds\_final, method = "over", N = no\_count\*2)$data  
  
##### Sampling both (Under)  
#training\_subset\_ds\_final <- ovun.sample(litigated ~., data = training\_subset\_ds\_final, method = "under", N = yes\_count\*2)$data  
  
table(training\_subset\_ds\_final$litigated)

##   
## No Yes   
## 159 91

glm\_control <- trainControl(  
 method = "cv",  
 number = 10,  
 summaryFunction = twoClassSummary,  
 classProbs = TRUE   
)  
set.seed(123)  
glm\_model <- train(litigated ~  
 .  
 ,   
 data = training\_subset\_ds\_final, method = "glm", family = "binomial", trControl = glm\_control)  
class\_probabilities <- predict(glm\_model, 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\_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 <- glm\_rocr\_auc@y.values[[1]]  
 glm\_auc

## [1] 0.6931818

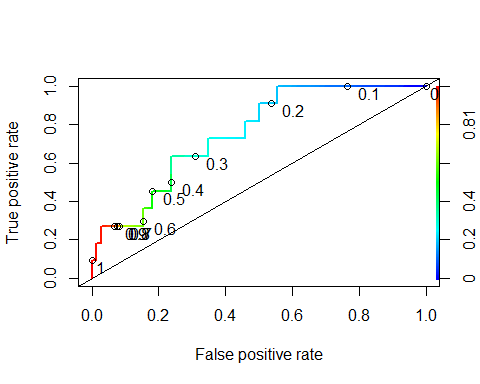
set.seed(123)  
glm\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glm\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glm\_confusion\_matrix <- confusionMatrix(factor(glm\_class\_probabilities\_litigated), factor(glm\_litigated), positive = "Yes")  
glm\_confusion\_matrix

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 54 4  
## Yes 18 7  
##   
## Accuracy : 0.7349   
## 95% CI : (0.6266, 0.8258)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.999626   
##   
## Kappa : 0.251   
##   
## Mcnemar's Test P-Value : 0.005578   
##   
## Sensitivity : 0.63636   
## Specificity : 0.75000   
## Pos Pred Value : 0.28000   
## Neg Pred Value : 0.93103   
## Prevalence : 0.13253   
## Detection Rate : 0.08434   
## Detection Prevalence : 0.30120   
## Balanced Accuracy : 0.69318   
##   
## 'Positive' Class : Yes   
##

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

glmnet\_control <- trainControl(  
 method = "cv",  
 number = 10,  
 summaryFunction = twoClassSummary,  
 classProbs = TRUE   
)  
set.seed(123)  
glmnet\_model <- train(litigated ~ ., data = training\_subset\_ds\_final, method = "glmnet", family = "binomial", trControl = glmnet\_control)  
class\_probabilities <- predict(glmnet\_model, 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\_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] 0.7537879

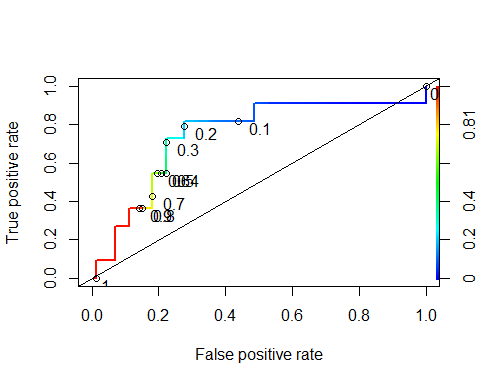
set.seed(123)  
glmnet\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
glmnet\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
glmnet\_confusion\_matrix <- confusionMatrix(factor(glmnet\_class\_probabilities\_litigated), 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 <- glmnet\_confusion\_matrix$overall["Accuracy"]

lda\_control <- trainControl(  
 method = "cv",  
 number = 10  
)  
set.seed(123)  
lda\_model <- train(litigated ~  
 .,   
 data = training\_subset\_ds\_final, method = "lda", family = "binomial", trControl = lda\_control)  
class\_probabilities <- predict(lda\_model, 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\_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 <- lda\_rocr\_auc@y.values[[1]]  
 lda\_auc

## [1] 0.7424242

set.seed(123)  
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

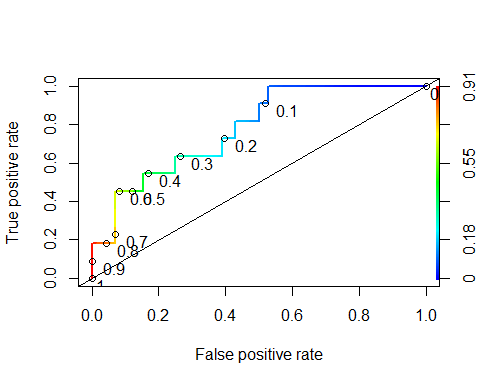
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 59 6  
## Yes 13 5  
##   
## Accuracy : 0.7711   
## 95% CI : (0.6658, 0.8562)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9948   
##   
## Kappa : 0.2158   
##   
## Mcnemar's Test P-Value : 0.1687   
##   
## 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   
##

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

gbm\_control <- trainControl(method = "repeatedcv", number = 10, repeats = 10)  
set.seed(123)  
gbm\_model <- train(litigated~.,   
 data=training\_subset\_ds\_final,   
 method = "gbm",  
 trControl = gbm\_control,  
 verbose = FALSE)

class\_probabilities <- predict(gbm\_model, 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\_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 <- gbm\_rocr\_auc@y.values[[1]]  
gbm\_auc

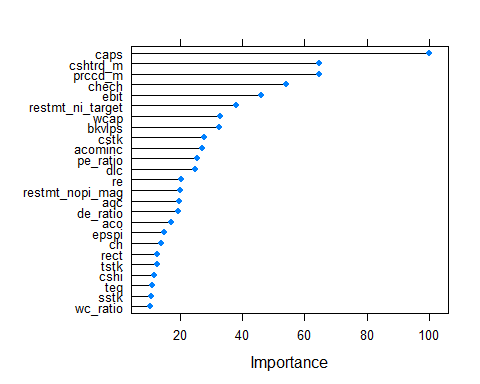
## [1] 0.7765152

set.seed(123)  
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

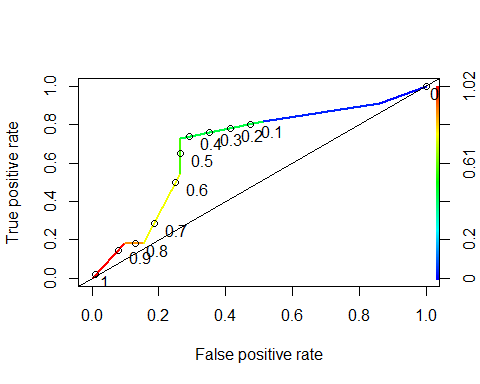
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 64 6  
## Yes 8 5  
##   
## Accuracy : 0.8313   
## 95% CI : (0.7332, 0.9046)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.8698   
##   
## Kappa : 0.3189   
##   
## Mcnemar's Test P-Value : 0.7893   
##   
## 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   
##

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

gbm\_varImp <- varImp(gbm\_model, type = 2)  
plot(gbm\_varImp, top = 25)



set.seed(123)  
tree\_model <- rpart(litigated ~ ., data = training\_subset\_ds\_final, method = "class")  
class\_probabilities <- predict(tree\_model, 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\_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

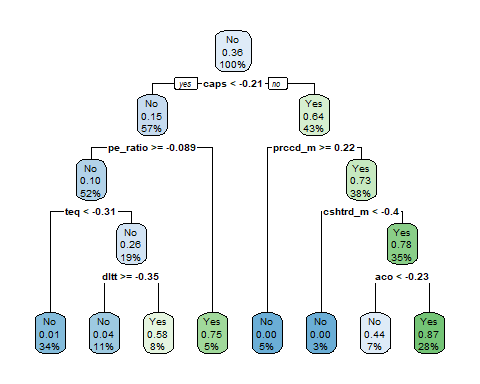
## [1] 0.6849747

set.seed(123)  
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 5  
## Yes 19 6  
##   
## Accuracy : 0.7108   
## 95% CI : (0.6009, 0.8052)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.999951   
##   
## Kappa : 0.1829   
##   
## Mcnemar's Test P-Value : 0.007963   
##   
## Sensitivity : 0.54545   
## Specificity : 0.73611   
## Pos Pred Value : 0.24000   
## Neg Pred Value : 0.91379   
## Prevalence : 0.13253   
## Detection Rate : 0.07229   
## Detection Prevalence : 0.30120   
## Balanced Accuracy : 0.64078   
##   
## 'Positive' Class : Yes   
##

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

rpart.plot(tree\_model)

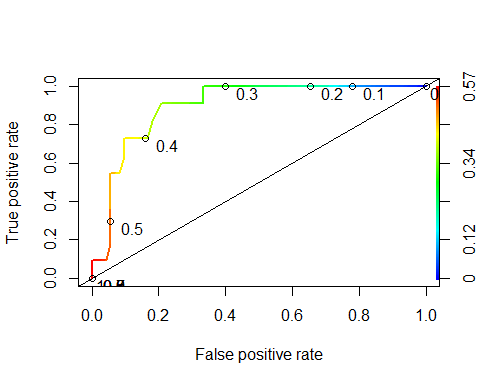


rf\_x <- training\_subset\_ds\_final[,num\_var\_start\_index:num\_var\_end\_index]  
rf\_y <- training\_subset\_ds\_final[,target\_var\_index]  
recommended\_mtry <- sqrt(ncol(rf\_x))  
rfGrid <- expand.grid(mtry=recommended\_mtry)

rfControl <- trainControl(method='repeatedcv', number=10, repeats=3)  
set.seed(123)  
rf\_model <- train(litigated~.,   
 data=training\_subset\_ds\_final,   
 method='rf',   
 metric='Accuracy',   
 tuneGrid=rfGrid,   
 trControl=rfControl)

class\_probabilities <- predict(rf\_model, 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\_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 <- rf\_rocr\_auc@y.values[[1]]  
rf\_auc

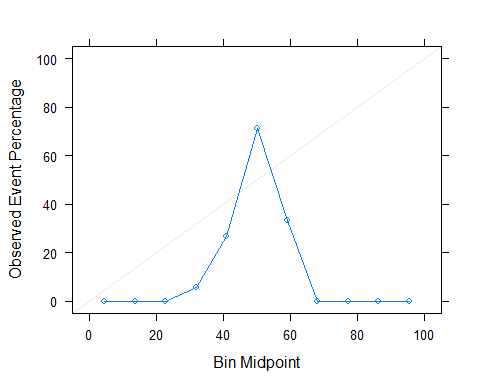
## [1] 0.8945707

set.seed(123)  
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

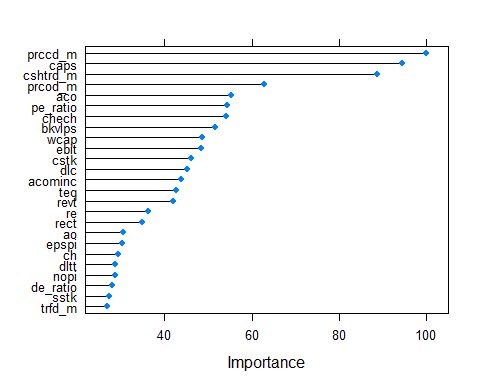
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 68 8  
## Yes 4 3  
##   
## Accuracy : 0.8554   
## 95% CI : (0.7611, 0.923)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.6969   
##   
## Kappa : 0.2567   
##   
## Mcnemar's Test P-Value : 0.3865   
##   
## Sensitivity : 0.27273   
## Specificity : 0.94444   
## Pos Pred Value : 0.42857   
## Neg Pred Value : 0.89474   
## Prevalence : 0.13253   
## Detection Rate : 0.03614   
## Detection Prevalence : 0.08434   
## Balanced Accuracy : 0.60859   
##   
## 'Positive' Class : Yes   
##

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

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



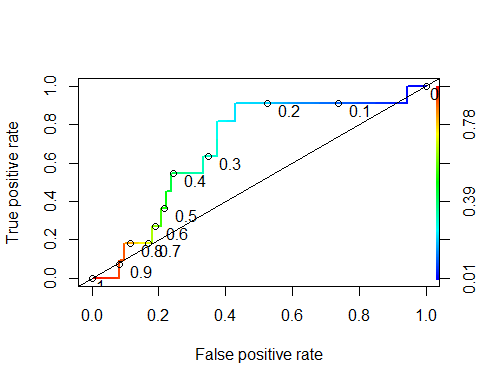
rf\_varImp <- varImp(rf\_model, type = 2)  
plot(rf\_varImp, top = 25)



rf\_varImp

## rf variable importance  
##   
## only 20 most important variables shown (out of 59)  
##   
## Overall  
## prccd\_m 100.00  
## caps 94.45  
## cshtrd\_m 88.73  
## prcod\_m 62.84  
## aco 55.25  
## pe\_ratio 54.31  
## chech 54.18  
## bkvlps 51.74  
## wcap 48.62  
## ebit 48.51  
## cstk 46.18  
## dlc 45.23  
## acominc 43.75  
## teq 42.64  
## revt 42.11  
## re 36.21  
## rect 34.86  
## ao 30.58  
## epspi 30.27  
## ch 29.37

nnGrid <- expand.grid(size = 8:10, decay = 0.2)  
nnControl <- trainControl(method = "repeatedcv",  
 repeats = 5,  
 number=10,  
 classProbs = TRUE)  
set.seed(123)  
nn\_model <- train(litigated~.,   
 data=training\_subset\_ds\_final,  
 method = "nnet",  
 tuneGrid = nnGrid,  
 trControl = nnControl,  
 importance = TRUE,  
 trace = FALSE,  
 MaxNWts = 1500)  
  
class\_probabilities <- predict(nn\_model, 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  
  
nn\_rocr\_pred <- prediction(test\_subset\_ds\_final$class\_probabilities\_litigated, test\_subset\_ds\_final$litigated)  
nn\_rocr\_roc <- performance(nn\_rocr\_pred, measure = "tpr", x.measure = "fpr")  
plot(nn\_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)



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

## [1] 0.6830808

set.seed(123)  
nn\_litigated <- ifelse(test\_subset\_ds\_final$litigated == 1, "Yes", "No")  
nn\_class\_probabilities\_litigated <- ifelse(test\_subset\_ds\_final$class\_probabilities\_litigated > 0.50, "Yes", "No")  
nn\_confusion\_matrix <- confusionMatrix(factor(nn\_class\_probabilities\_litigated), factor(nn\_litigated), positive = "Yes")  
nn\_confusion\_matrix

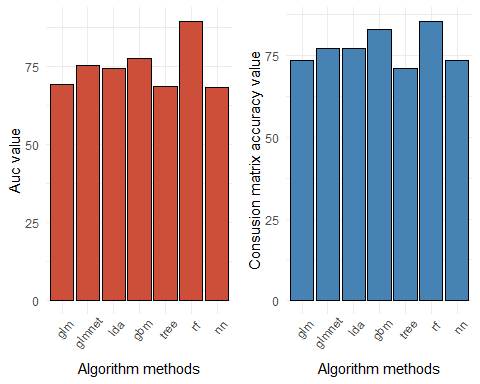
## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 57 7  
## Yes 15 4  
##   
## Accuracy : 0.7349   
## 95% CI : (0.6266, 0.8258)  
## No Information Rate : 0.8675   
## P-Value [Acc > NIR] : 0.9996   
##   
## Kappa : 0.1187   
##   
## Mcnemar's Test P-Value : 0.1356   
##   
## Sensitivity : 0.36364   
## Specificity : 0.79167   
## Pos Pred Value : 0.21053   
## Neg Pred Value : 0.89062   
## Prevalence : 0.13253   
## Detection Rate : 0.04819   
## Detection Prevalence : 0.22892   
## Balanced Accuracy : 0.57765   
##   
## 'Positive' Class : Yes   
##

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

auc\_df <- data.frame("algorithm\_method" = c("glm"), "auc" = c(glm\_auc) \* 100, "accuracy" = c(glm\_accuracy) \* 100)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("glmnet"), "auc" = c(glmnet\_auc) \* 100, "accuracy" = c(glmnet\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("lda"), "auc" = c(lda\_auc) \* 100, "accuracy" = c(lda\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("gbm"), "auc" = c(gbm\_auc) \* 100, "accuracy" = c(gbm\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("tree"), "auc" = c(tree\_auc) \* 100, "accuracy" = c(tree\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("rf"), "auc" = c(rf\_auc) \* 100, "accuracy" = c(rf\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df\_temp <- data.frame("algorithm\_method" = c("nn"), "auc" = c(nn\_auc) \* 100, "accuracy" = c(nn\_accuracy) \* 100)  
auc\_df <- rbind(auc\_df,auc\_df\_temp)  
  
auc\_df <- auc\_df %>%   
 arrange(desc(auc))  
  
auc\_df

## algorithm\_method auc accuracy  
## Accuracy5 rf 89.45707 85.54217  
## Accuracy3 gbm 77.65152 83.13253  
## Accuracy1 glmnet 75.37879 77.10843  
## Accuracy2 lda 74.24242 77.10843  
## Accuracy glm 69.31818 73.49398  
## Accuracy4 tree 68.49747 71.08434  
## Accuracy6 nn 68.30808 73.49398

auc\_plot <- ggplot(data=auc\_df, aes(x=algorithm\_method, y=auc)) +  
 geom\_bar(colour="black", stat="identity", fill="tomato3") +  
 ylab("Auc value") +  
 xlab("Algorithm methods") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))  
  
accuracy\_plot <- ggplot(data=auc\_df, aes(x=algorithm\_method, y=accuracy)) +  
 geom\_bar(colour="black", stat="identity", fill = "steelblue") +  
 ylab("Consusion matrix accuracy value") +  
 xlab("Algorithm methods") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle =50, hjust=0.75))  
  
grid.arrange(auc\_plot, accuracy\_plot,ncol=2)



nrow(training\_subset\_ds\_final)

## [1] 250

nrow(test\_subset\_ds\_final)

## [1] 83

ncol(training\_subset\_ds\_final)

## [1] 60

colnames(training\_subset\_ds\_final)

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