Assignment Report-

# DATA EXTACTION AND LOAD

**Load the data and read it**

setwd("D:\\CBA\\sa2\\Assignment")

insData <- read.csv("aData.csv")

**# A Quick look on the columns and structure**

str(insData)

**# we have a total 15 variables, Company name is Categorical, rest are numerical**

**# expanse is the response variable**

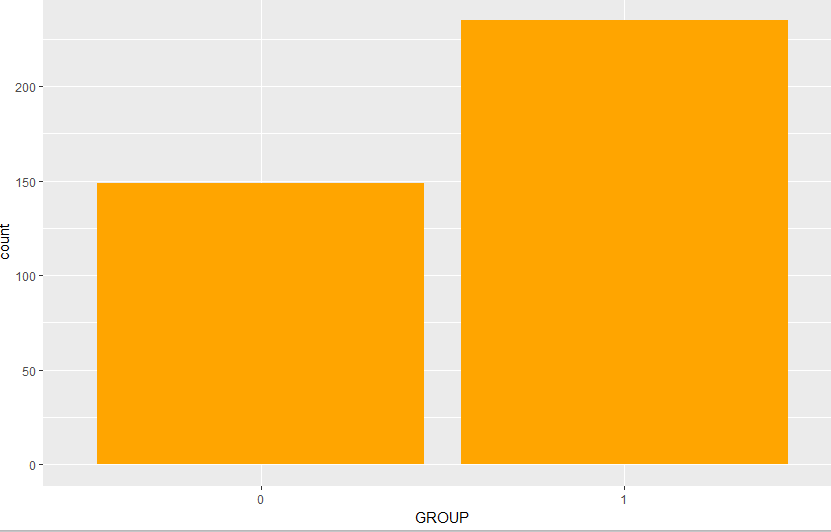
**#We can leave out the company name variable and proceed.**

insdata1 <- insData

insdata1$COMPANY\_NAME <-NULL

# EXPLORATERORY DATA ANALYSIS - UNIVARIATE

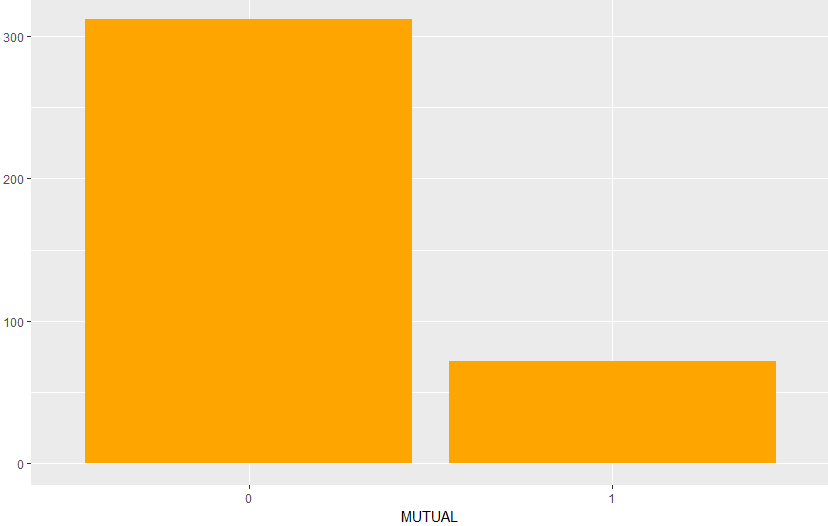
GROUP-



# total 61 % companies are affiliated making it somewhat prevalent case to find patterns for such companies.

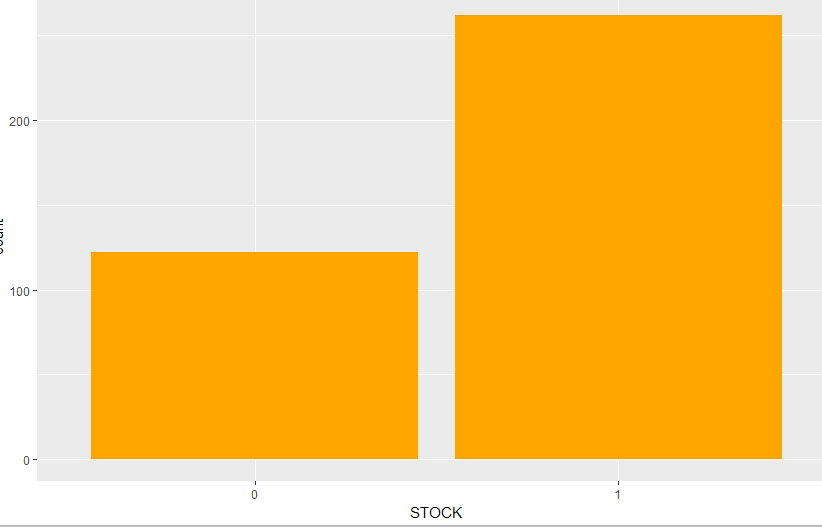
#let's do the same for remaining two categorical variables, which are mutual and stock.

**MUTUAL**



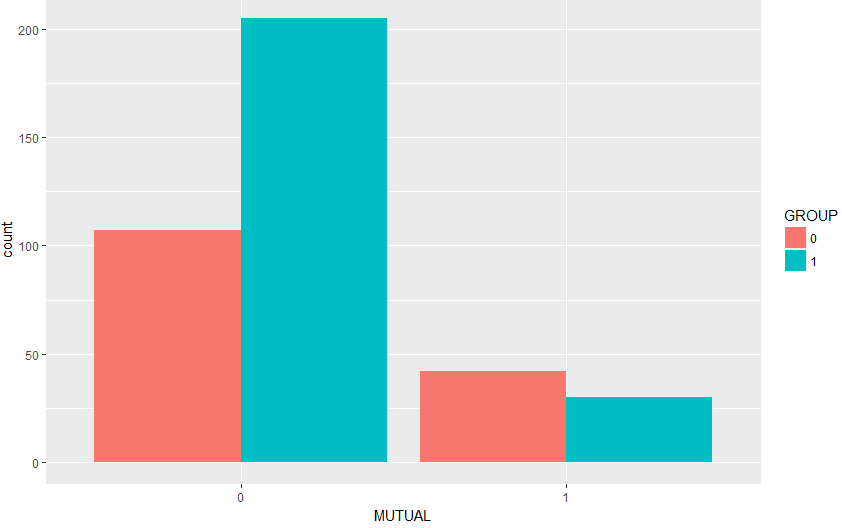
# About 81% comapnies are non Mutual, so more patterns for non mutual companies may be there .

**STOCK-**



# About 69% companies are STOCK companies, so more patterns for STOCK companies may be there.

**MUTUAL and GROUP**



# we see a contradictory fact here , For non mutual companies affiliated companies are almost double than non affiliated ones where as for

# Mutual companies the distribution is almost equal(more on affiliate actually)

# this seems like Affiliated companies are more inclined to become non Mutual

# around 27% of companies are non mutual and non affiliated

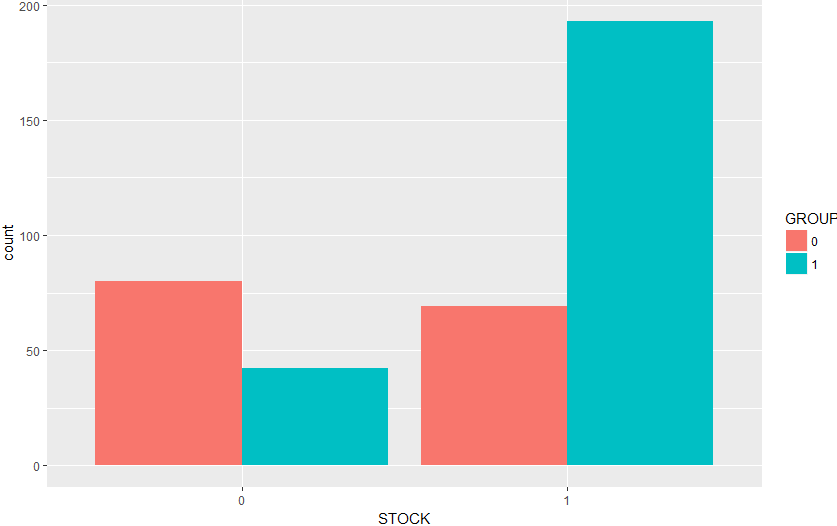
# Around 53% of companies are Non mutual and Affiliated

# Around 11 % are Mutual and Non Affiliated

# around 8 % are mutual and affiliated

# we clearly see an interaction here between these category types , perhaps we can put an interaction bw these 2.

**GROUP and STOCK**



# around 21 % companies are non stock and non affiliated

# around 11 % of companies are non stock and affiliated

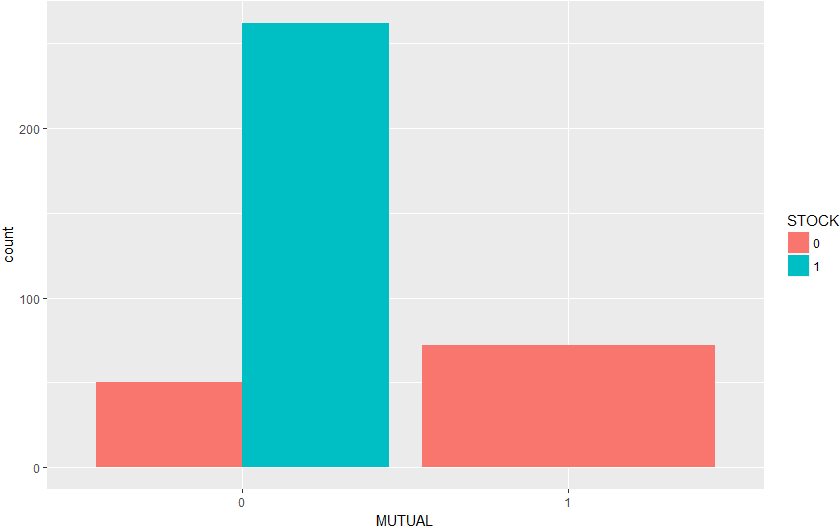
# around 18% companies are Stock and non affiliated

# around 50% companies are both stock and affiliated

# seems like Stock companies are more inclined to become affiliated(Kind of seems opposite trend since companies which are mutual seems less inclined to become affiliated)

# perhaps we can put an interaction bw these 2.

**STOCK and MUTUAL**



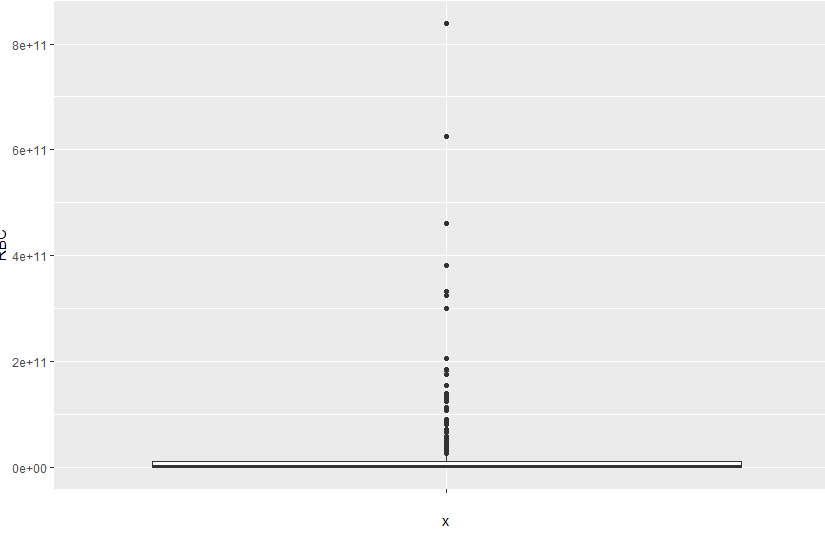
# around 13 % are non mutual and non stock(maybe individual companies)

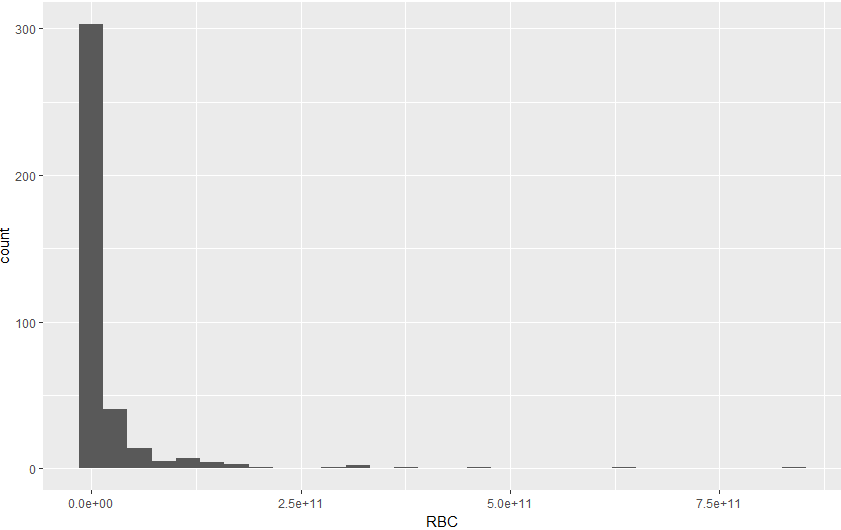
# around 68 % are non mutual and Stock

# around 19 % are Mutual and non stock

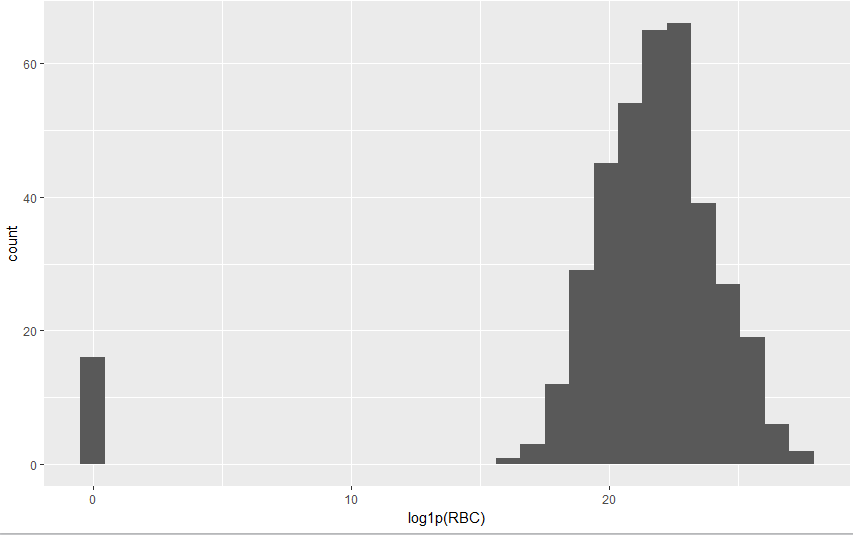
# we see a clear trend here no company is listed as mutual or stock if is present in other category(no intersection group is present)

RISK BASED CAPITAL (RBC)-





**# we can try a log trnasformation to reduce variance**



# clearly log does a better job in reducing the variance.Though we have to look for the observation at left hand .

**sum(is.null(insdata1$RBC))**

There are no missing values.

**# Analyse EEPENSES**

**summary(insdata1$EXPENSES)**

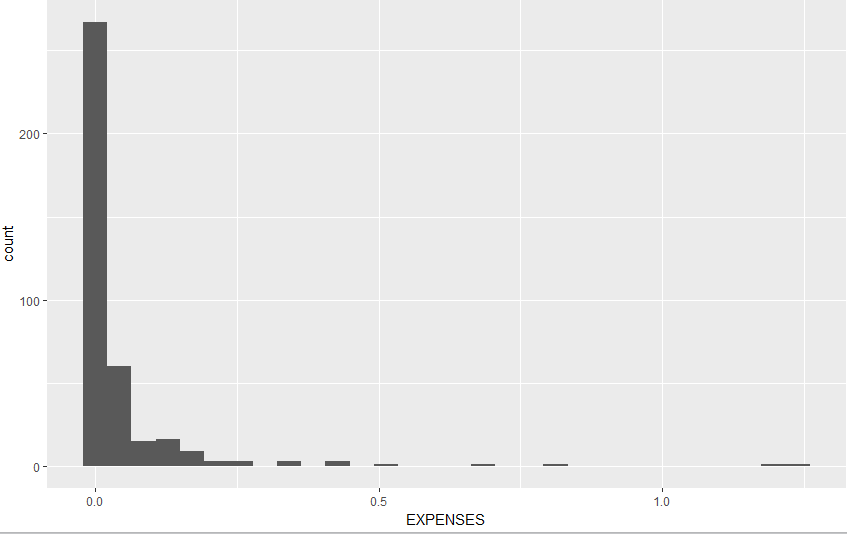
# A couple of points to note

# there are a few negative expanses (which is no possible, perhaps wrong data, we may mark these as zero or remove these transactions)

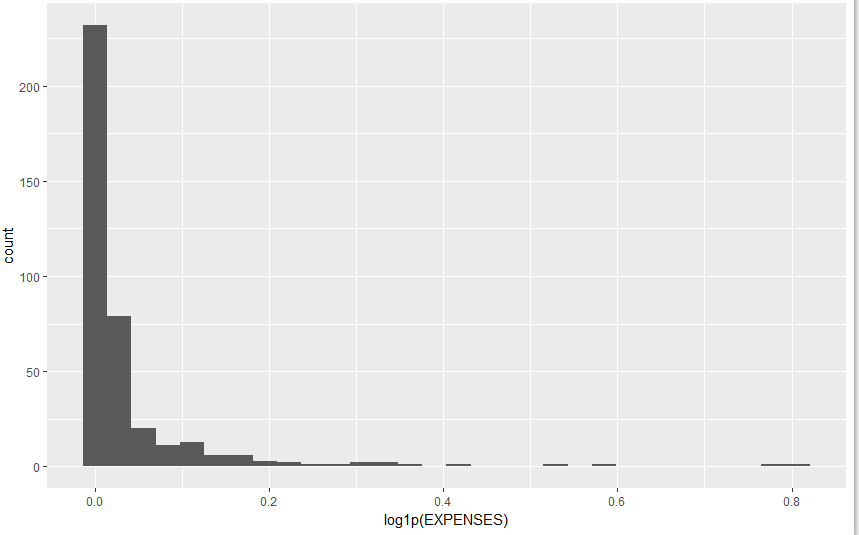
# 50 % of companies have 8540 dollars of expense, while mean is very high, even higher than the 3rd quartile and max value is 3 times the mean and 6 times the 3rd quartile

# a very right skewed distribution

# Most the companies have very less than quarter of a million dollars, we may consider the possibility of including these high end values, as if included as it is , these will shift the model towards high end.

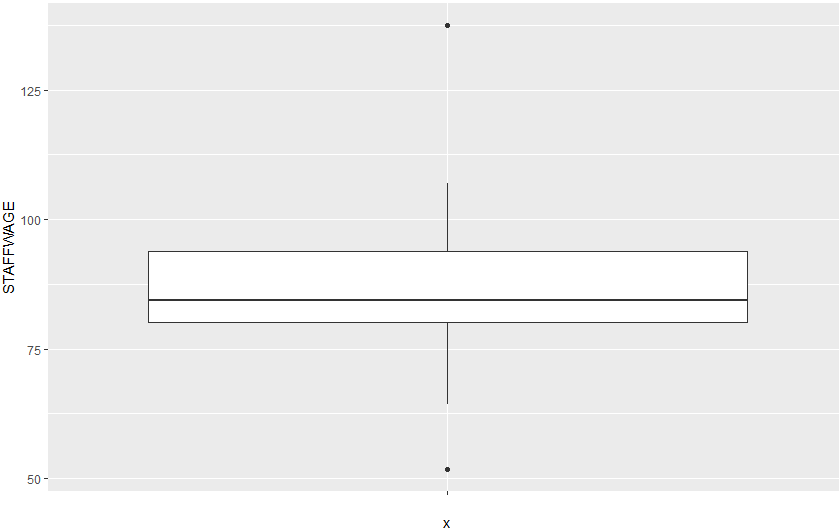


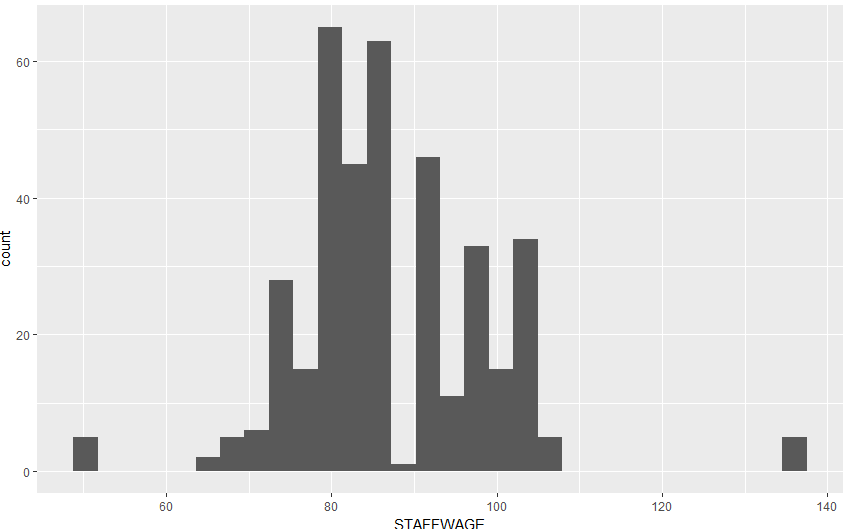
Tae a log transformation-



Even Log transformation can’t covert it into a normal distribution.

**# Analyse STAFFWAGE**





# seems like a normally distributed variable' staff salaries are fairly distributed

# Most wages are distributed between range of 43 to 107(approx.) having mean at 87 (there seems a few observations at both ends, which are worth investigating).

**# For past three variables RBS, Expanses and StaffWages, we saw the distribution is not normal and contains outliers, we may do a few things.**

**# 1- replace them with mean, median, mode or max-min value of distribution.**

**# 2- Assign a lowest and highest cap and put the values in that cap(same value for a continuous variable , say 10000 for any value above 10000 will be replace by 10000 and same for 10) .**

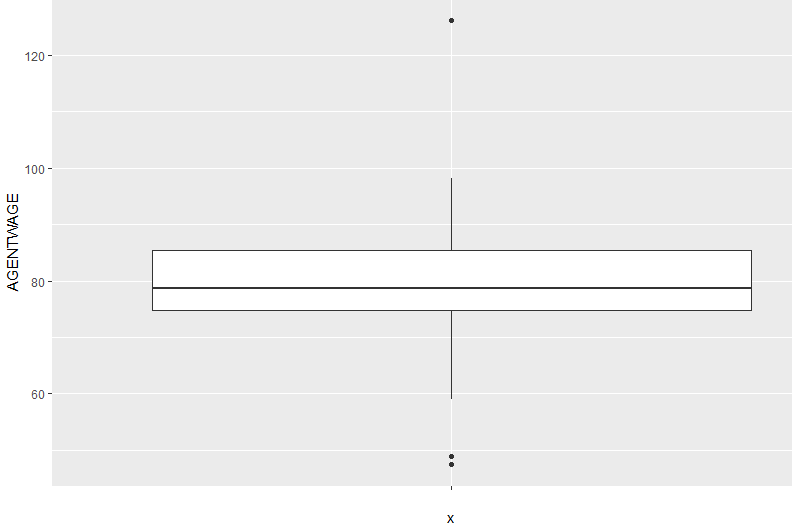
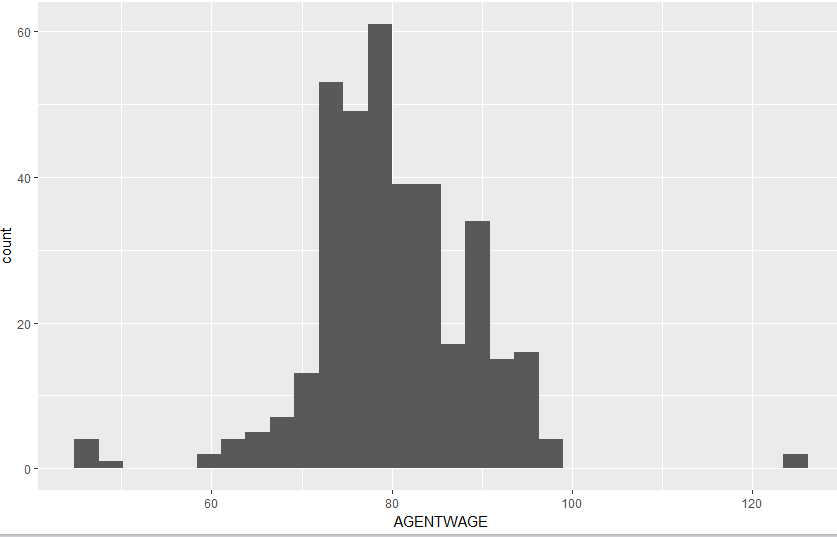
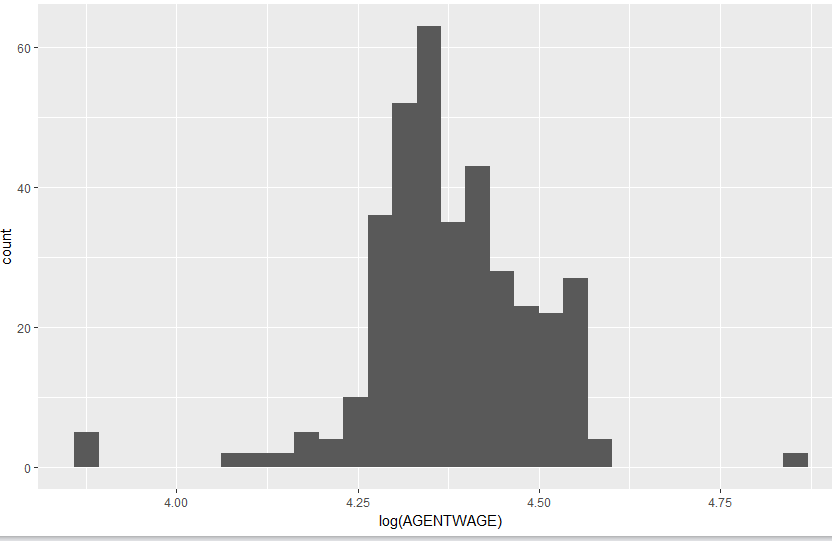
**# 3-we take the necessary X,Y transformation .**

**# Analyse AgentWage**

# It has 15 missing values

# seems like a normal distribution (mean and median are pretty close)

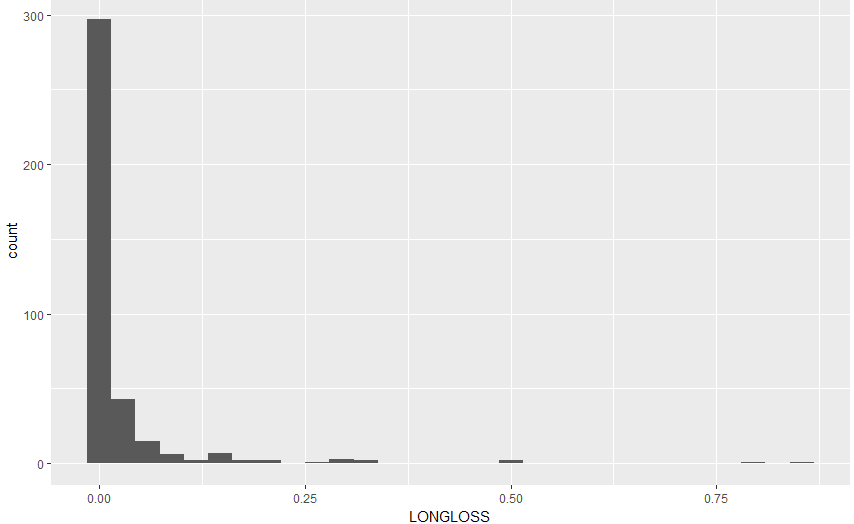
# Max value is thrice the min value (not much variation)

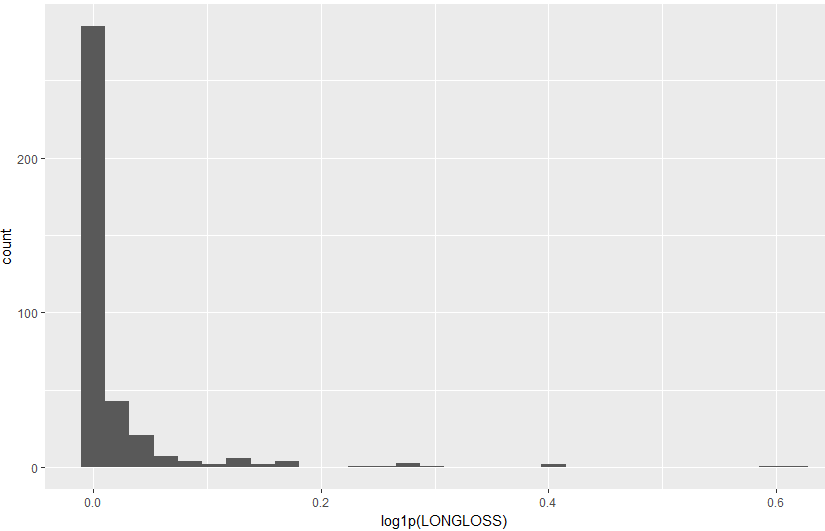
  

Log hasn’t done much srinkage . We will have to figure out the desired transformation after we analyse th regression plots.

**# analyse longloss**

# again a few values are negative, we may ask buiness if we can but here we will make these zero.





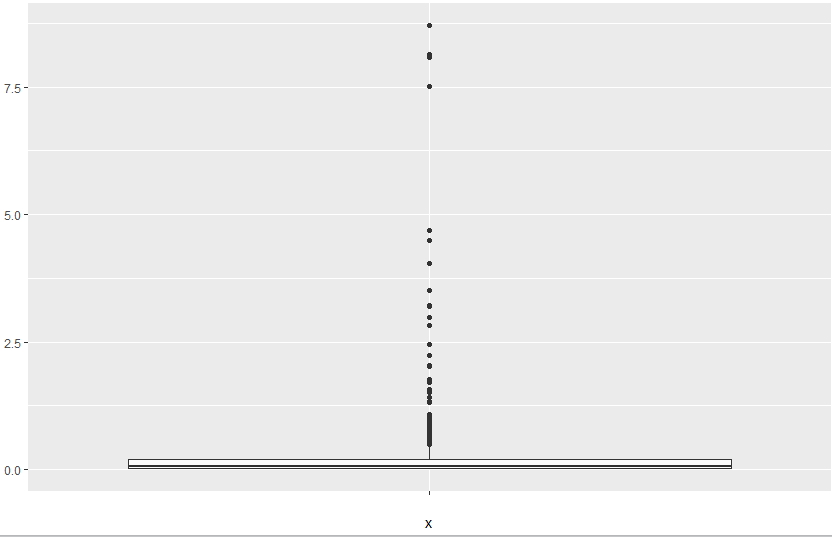
#no matter what transformation is used distribution stays skewed, perhaps the variation in values is too much to be suppressed by atransfromatio

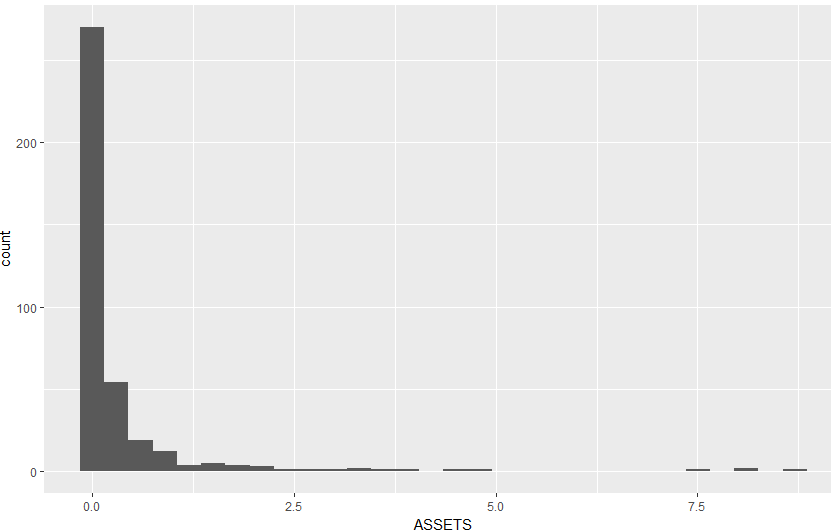
#more will be analyzed in multivariate analysis

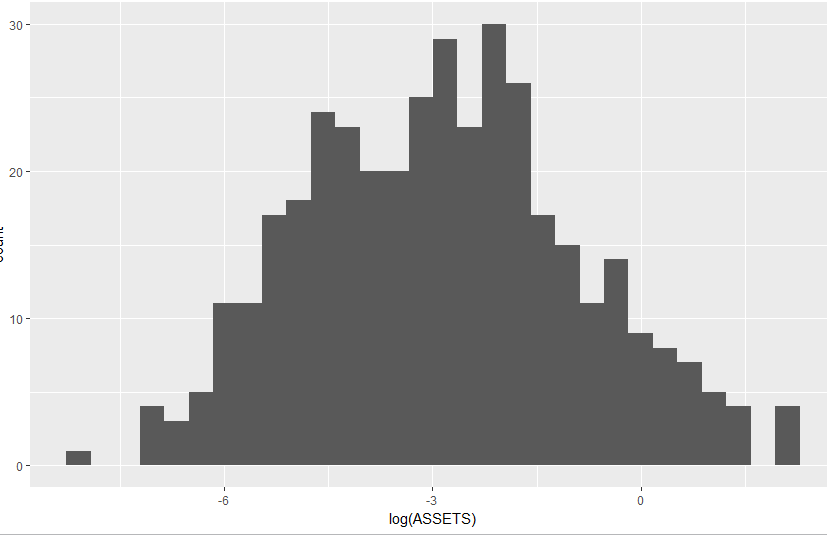
**We have similar Distribution for SHORTLOSS, GPWPERSONAL and GPWCOMM .**

**# analyse Assets**

# the distribution of assets is a very good example of how a few high observations can shift the mean and it doesn't represen the distribution correctly.







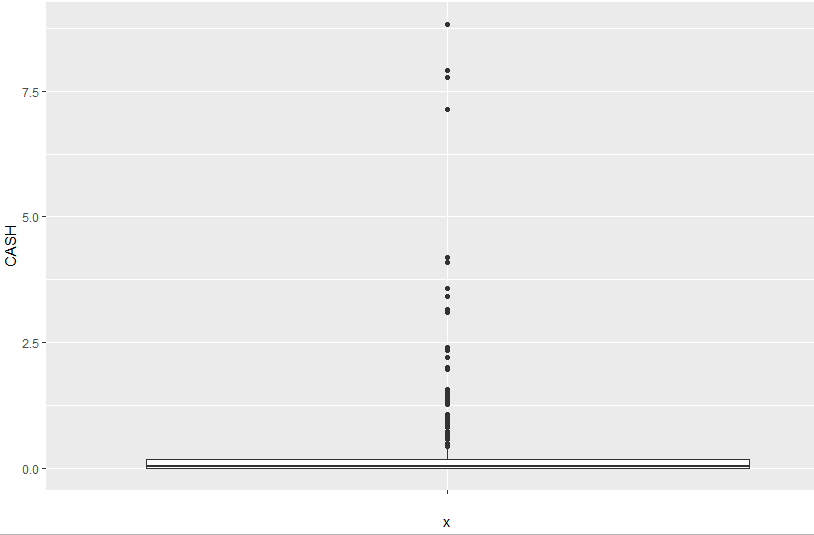
# once we take the log of the distribution , it somewhat shows a normally distributed shape

# we took the log in order reduce the variability.

**# We are just visualizing the variables differently, we will take the necessary transformation once we do the variable selection**

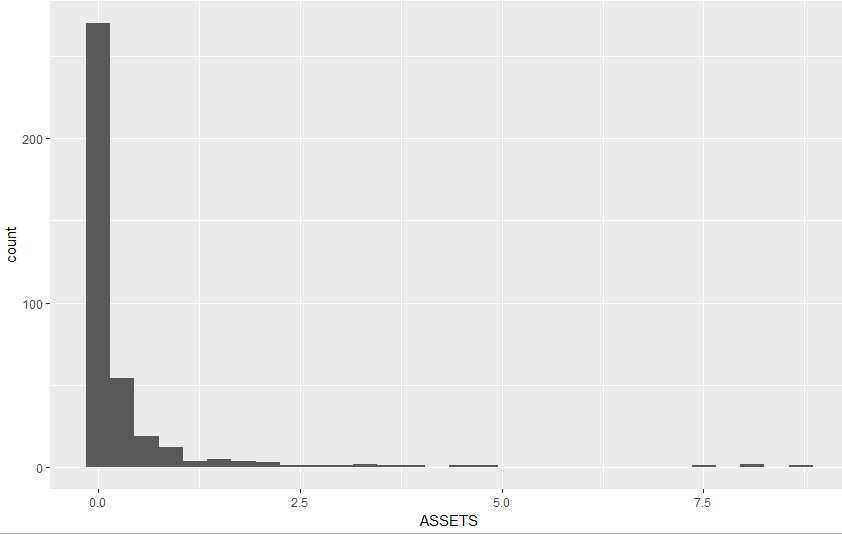
**# taking a feel of data is necessary for a good model.**

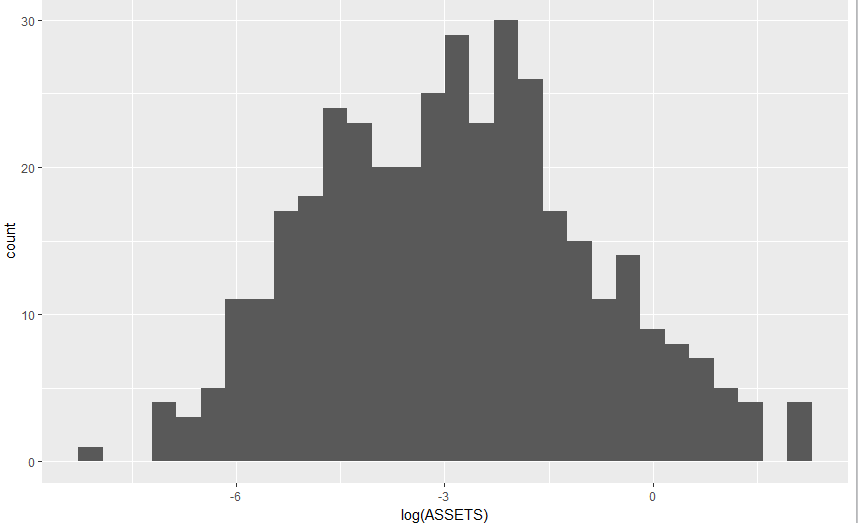
**#analyze CASH**



# A highly right skewed distribution, 75 % of total companies have cash lower than .19 million but the maximum is 9 mils, which is forcing mean to go towards right and making it skewed.

# If we think about such companies, we may have this situation in real world as well, but clearly we can't model both type using one model. We may think about capping the large value with a given constant or creating a different model for values above certain threshold.

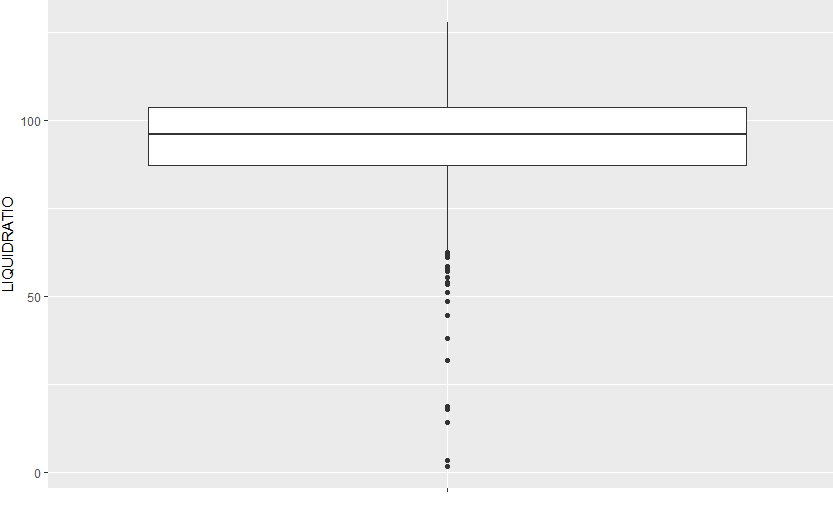


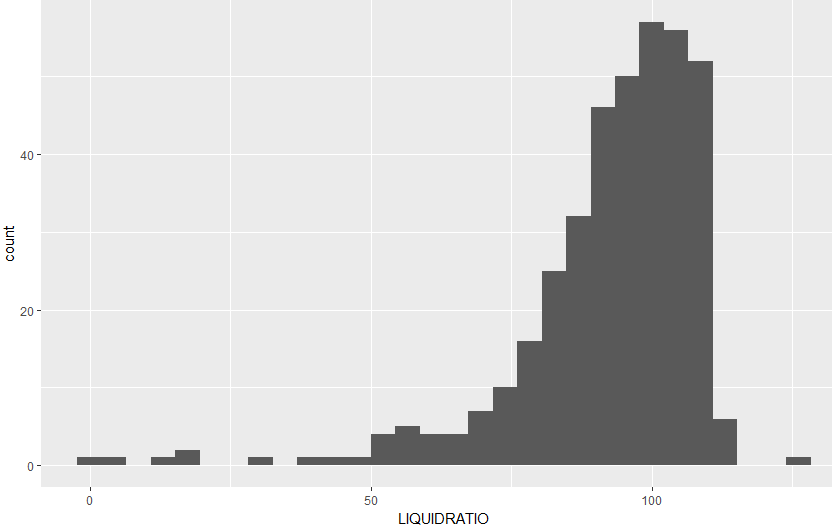


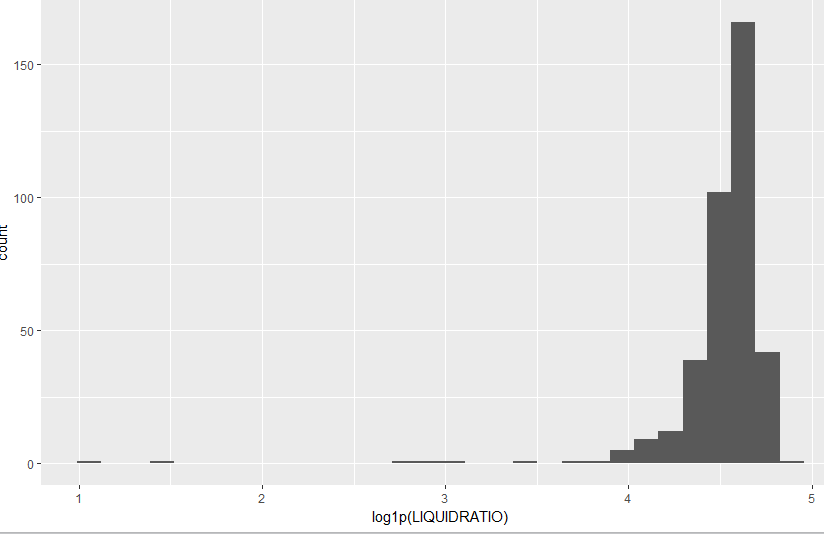
Log transformation did a good job in reducing variance.

**A Point to note is that distribution is almost identical to ASSETS which may cause multiCollinarity if both are associated with EXPENSE in a strong way.**

**#Analyse LIQUID RATIO**



# a left skewed distribution having a somewhat normal behaviour after the first quartile 



Even log couldn’t suppress all the variation.

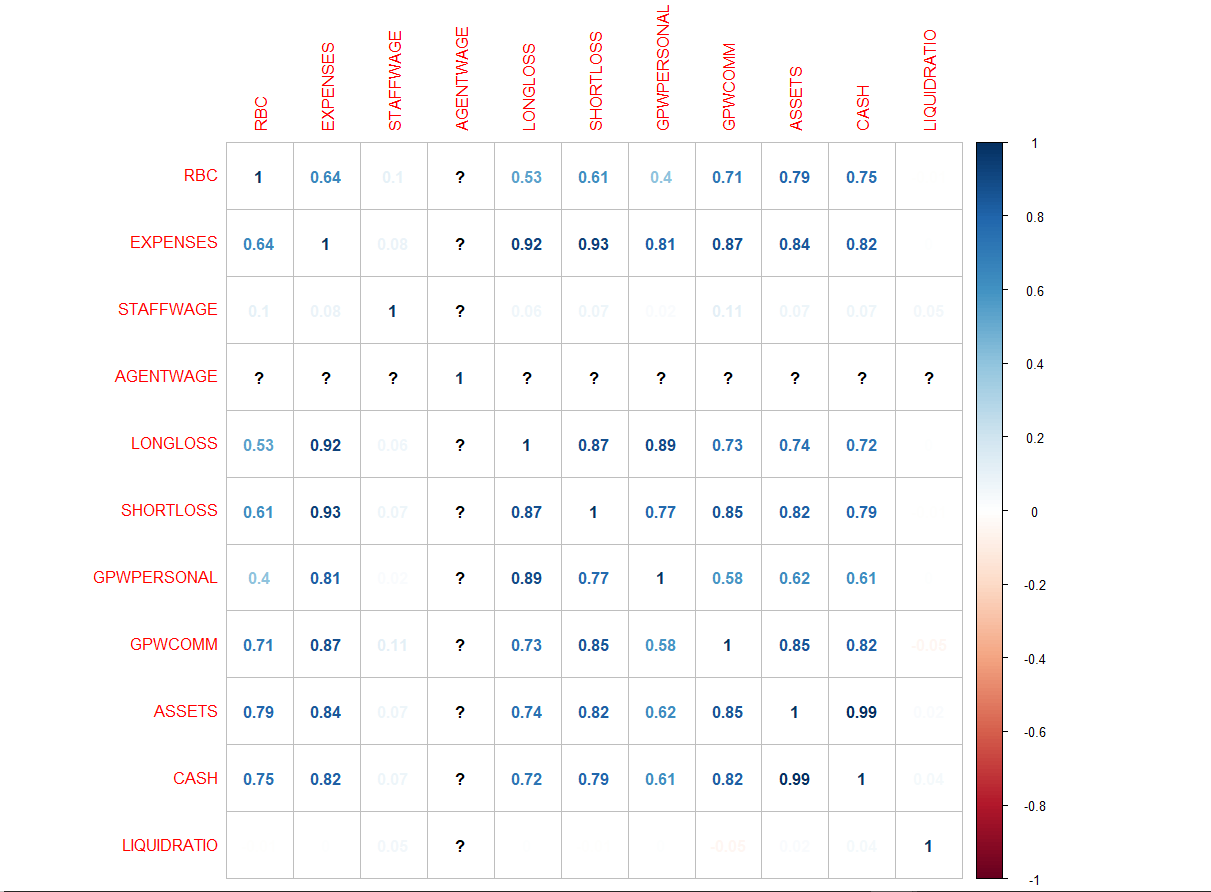
# MULTIVARIATE ANALYSIS

**# Lets do some bivariate analysis but before that we can follow certain rules to help us reduce the combinations to explore**

**# Such as seeing the correlation matrix, we can explore the variables which are having high correlation**

**# We can see the distribution of two continuous variables with regards to a third categorical variable to see the distribution for multiple categories.**

**Correlation matrix for Numerical variables.**



# From the plot we observe a good correlation between RBC and Expenses (.64) worth considering for partial correlation and scatter plot between them

# RBC and LONGLOSS

# RBC and SHORTLOSS

# RBC and GPWPCOMM

# RBC and assets

# RBC and CASH

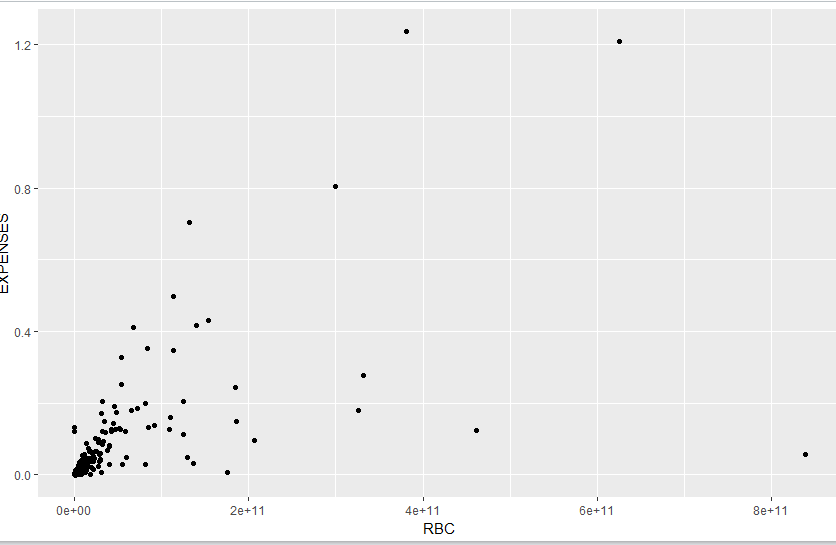
#EXPENSE is Correlated with LONGLOSS,SHORTLOSS,GPWPCOMM,,CASH, RBC,GPWPPERSONNAL

# LONGLOSS and SHORTLOSS are highly correlated. so is LonglOSS and GPWPPERSONNAL.

# The same behaviour is for most of the predictors.

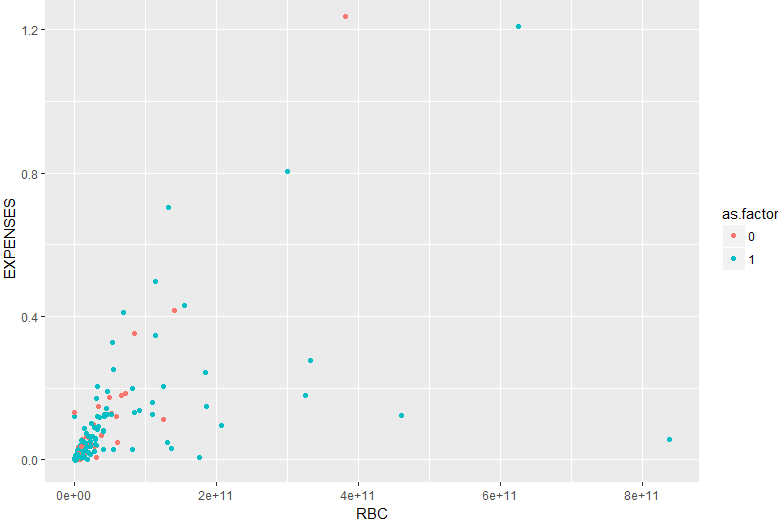
**RBC and EXPENSES**

# since we don't know the units of RBC , we will consider if the units of RBC is on thousands.



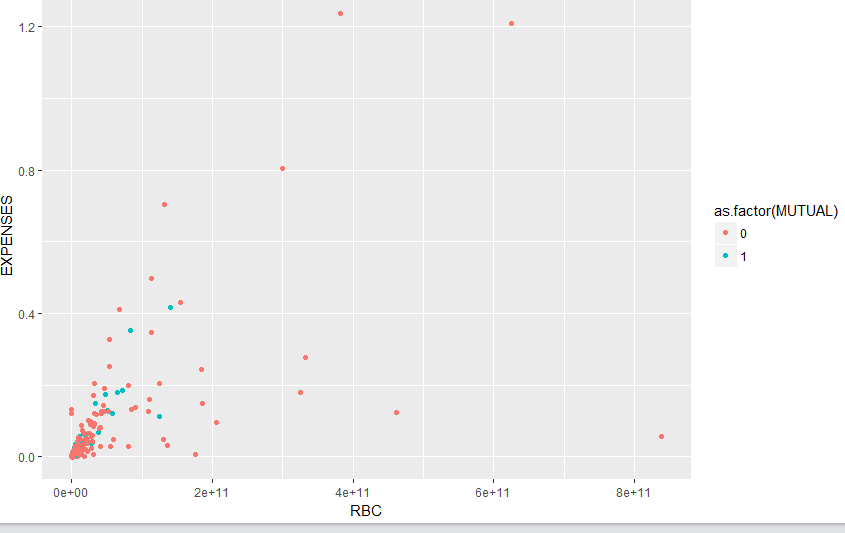
# as we suspected from corr matrix , there is clear increasing trend of expense with increasing risk based Capital. We see the variation increases as RBC increases, there are few observations for which this trend doesn't follow, we need to investigate.

**#does the company category play some role here ?**



# we see the variation is for companies which are type STOCK, NON STOCK companies follow the increasing trend with less variance

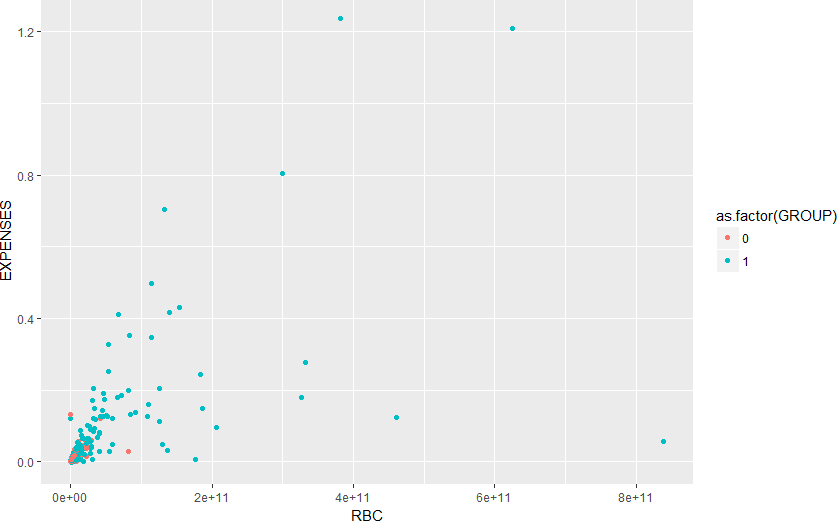
# For non STOCK companies expense can be predicted fore accurately using risk based capital



# as observed earlier MUTUAL companies exhibit opposite behaviour than Stock companies.

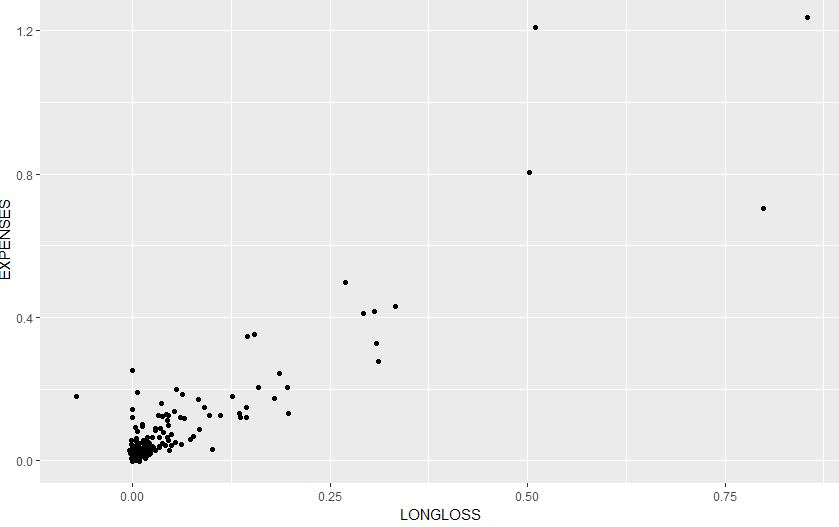
# e.g. where STOCK companies have more variation in terms of expense, MUTUAL companies have less variation.

# upon a closer LOOK we can think whether this is binary problem(a company is either stock or Mutual with a few exception)



# Again an identical plot, Showing affiliated as a dominant category

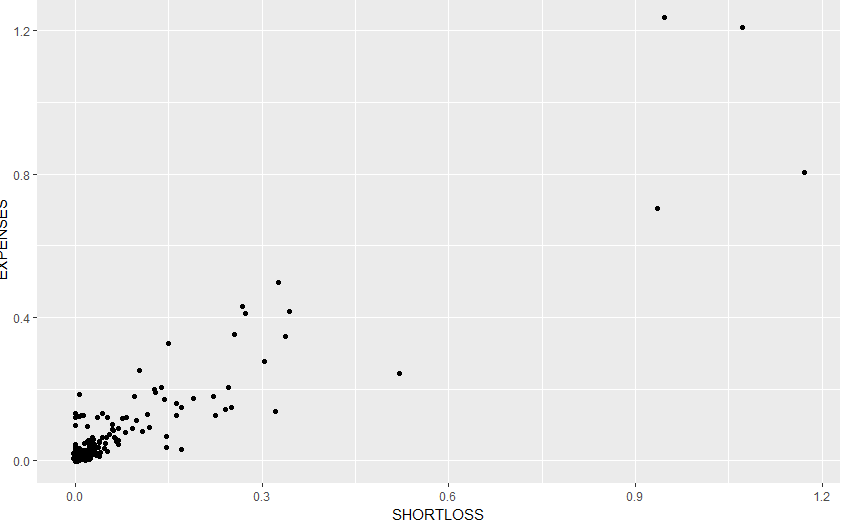
**EXPENSES and LONGLOSS**



# as expected from a correlation of .88

# More the loss , more the expense

**EXPENSES and SHORTLOSS**



# as expected from a correlation of .89

# More the loss , more the expense . slighly less variation thn Longloss

# we have a situation here, both of these variables are explaining the variation in expense quite effectively

# we may end up with collinearity problem, we need to look at vif scores for a more accurate model.

# we have similar scenario for remaining variables as well, an experimental approach with different variables will be more effective to identify most suitable set of predictors

**let's Look at partial correlation to observe the individual effect on expense**

**#ONLY AGENTWAGE has missing values.**

I would go with median to replace the missing value since there are a few extreme values in AGNETWAGE

# this shows a bit different picture

# 1. RBC doesn't have a .64 correlation with expanse. Instead it only has .09 partial correlation

# 2. long loss and short loss were highly correlated to Expense with values .92 and .93, , partially they are having .45 and .29 while asset is having negative correlation with the expense

# 3. cash and rbf are having a .55 correlation

# 4. staff wage and agent wage are highly correlated

# 5. assets and CSH are having a perfect liner relationship

# 6. CASH is having a negative correlation with RBC.

**# next question which comes to mind is how do we approach the problem**

**# best way is writing down the problem**

**# Explain the expense as the function of given variables.**

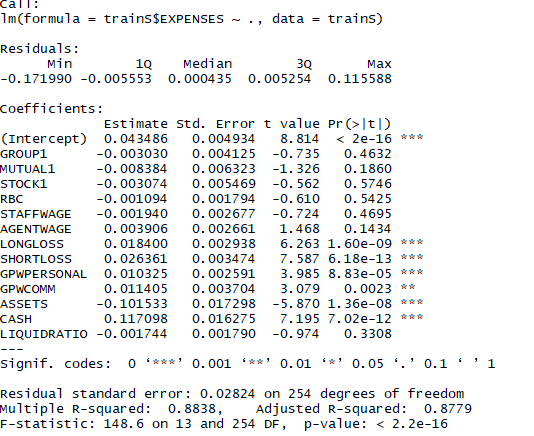
**#We are doing a few data edits for higher values as we believe that these higher values will only affect the parameters of model.**

**#Scaling all the numerical predictors.**

**# performing train test split( train set -268 rows(70 % of original data set) test set- 116 rows.**

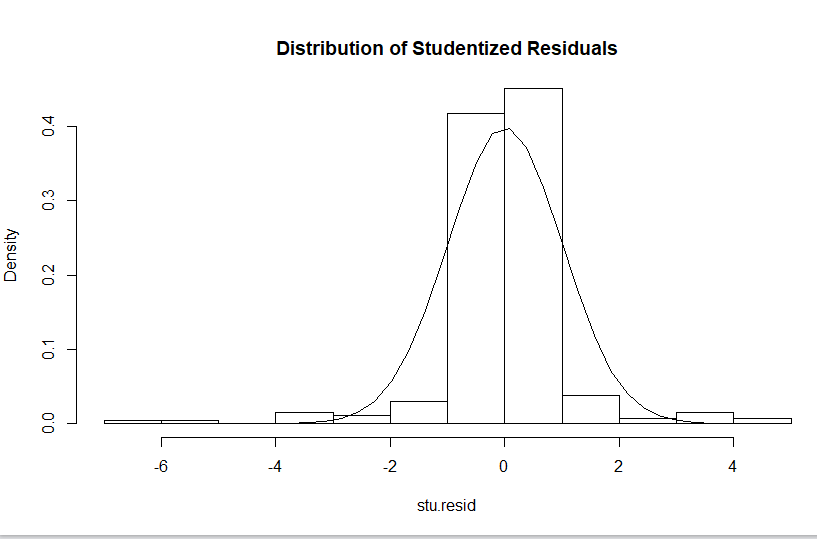
**Base Model-**

**model1<- lm(trainS$EXPENSES~. ,data = trainS)**



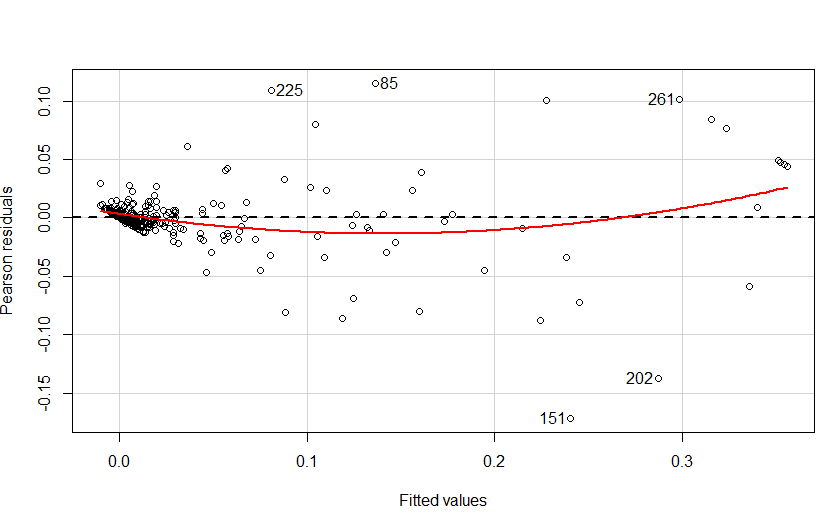
**RMSE OF the model on testSet is** 0.01971216

**Many variables are not considered. Let’s look ate residual vs fitted part.**



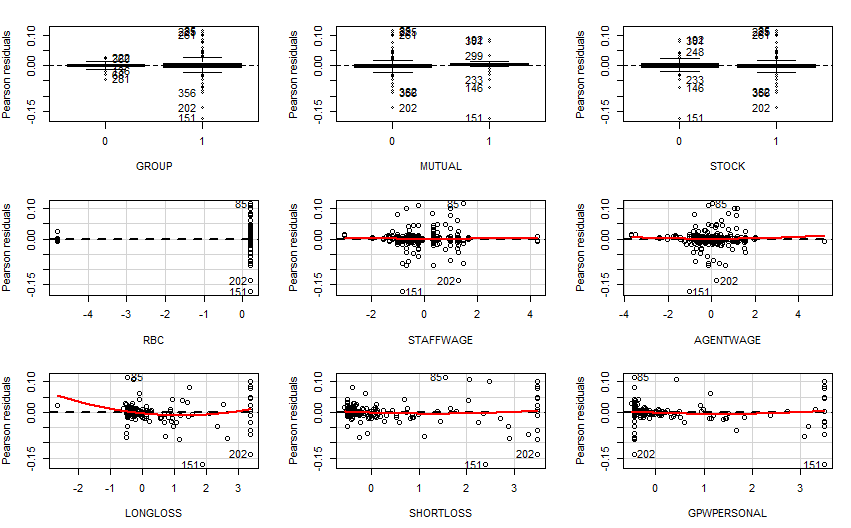
From the plot we see that there are certain large errors at left side and right side.

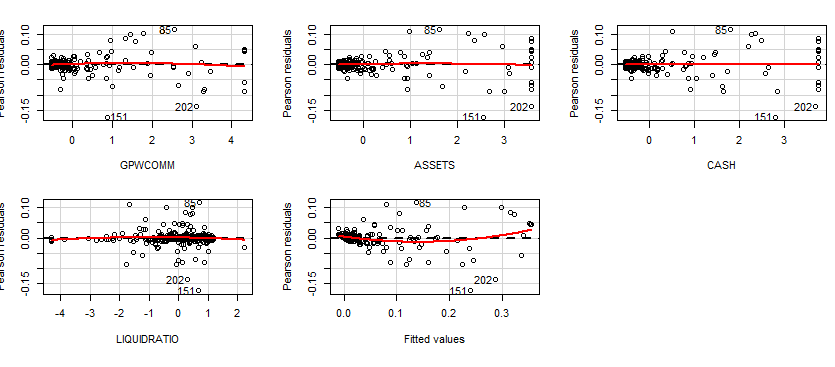
**Let’s look at the variance of residuals**



**There seems to be a trend (a quadratic one) .In middle and at end , model is being pulled to downwards and upwards.**

**# Let's look for residuals and regressors plots'**





**We See a few patterns in residuals**

# LONGLOSS- Outlier influence

# SHORTLOSS - skewed towards right

# GPWPERSONAL - skewed to right

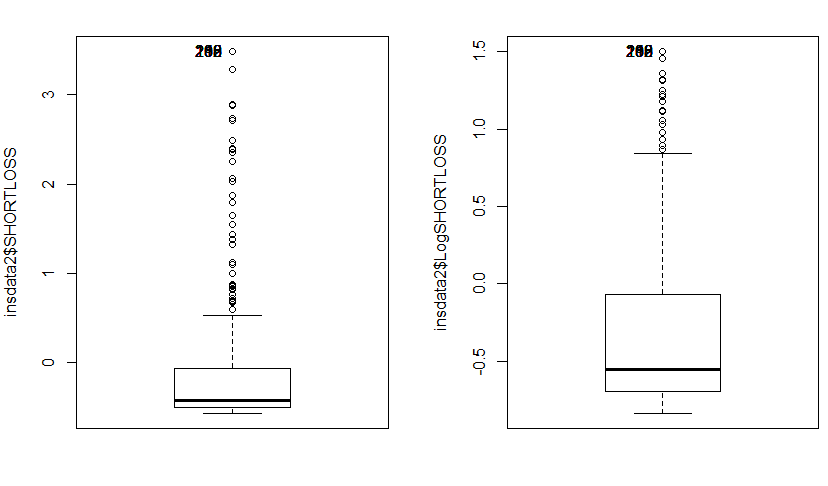
# GPWPCOMM - skewed to right

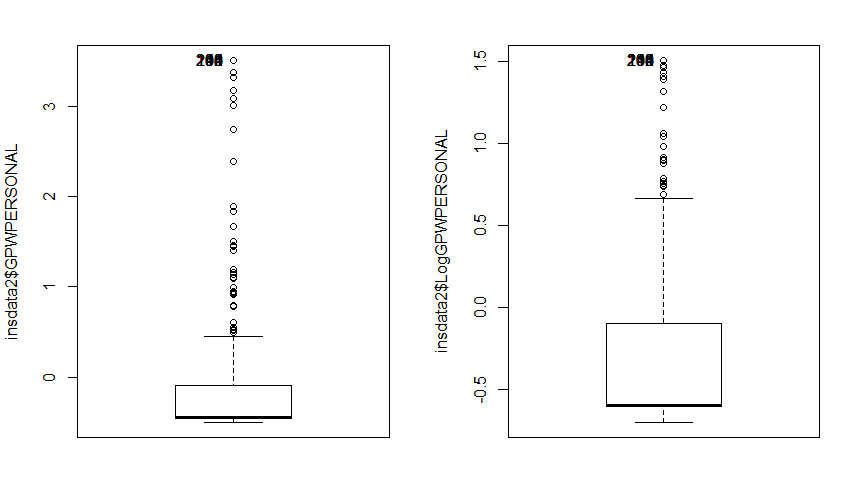
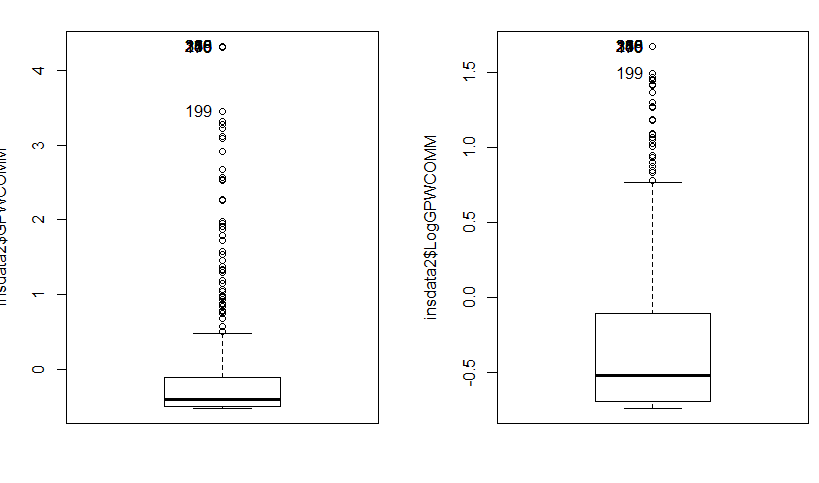
# Assets- Skewed to right

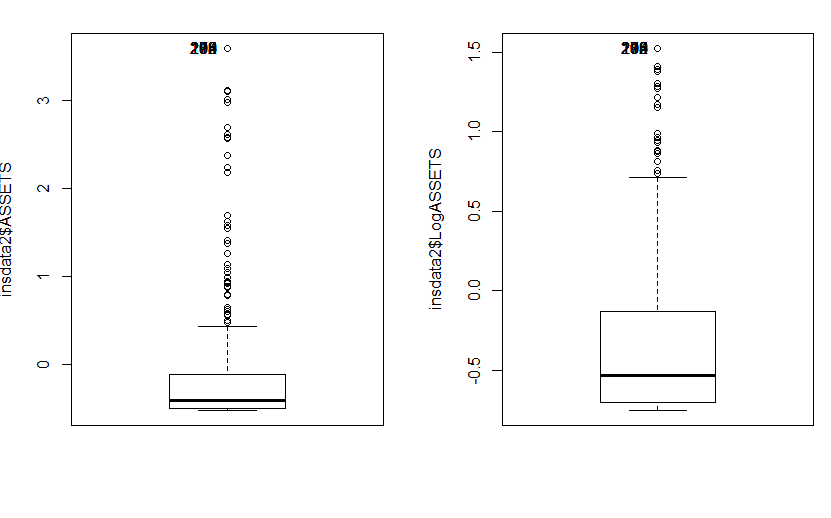
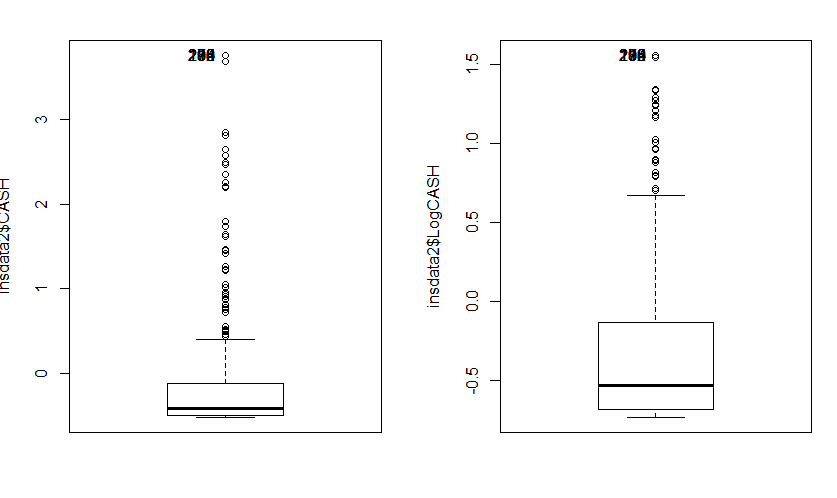
# CASH SKEWED to Right

# LIQUID RATIO - SKEWED To Left

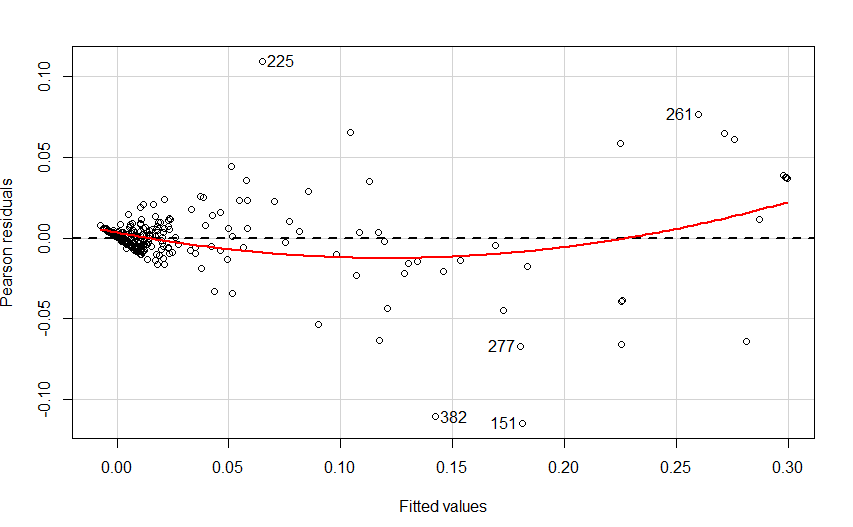
New Variables-

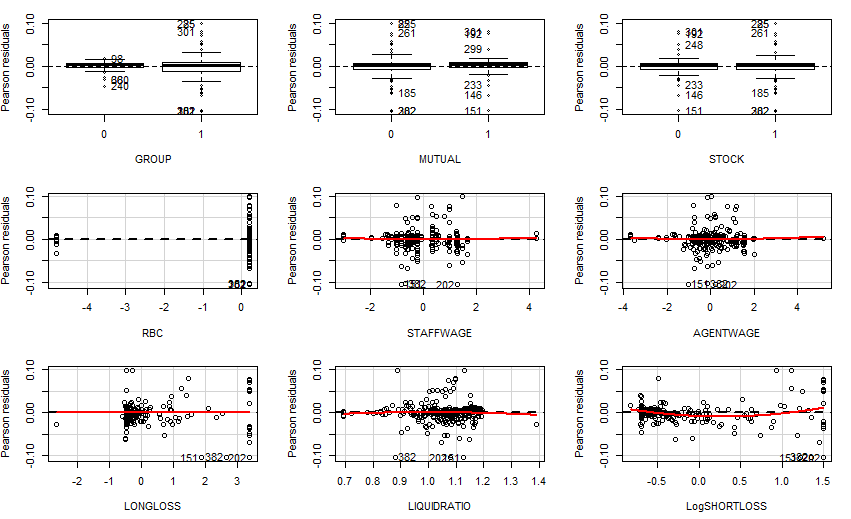


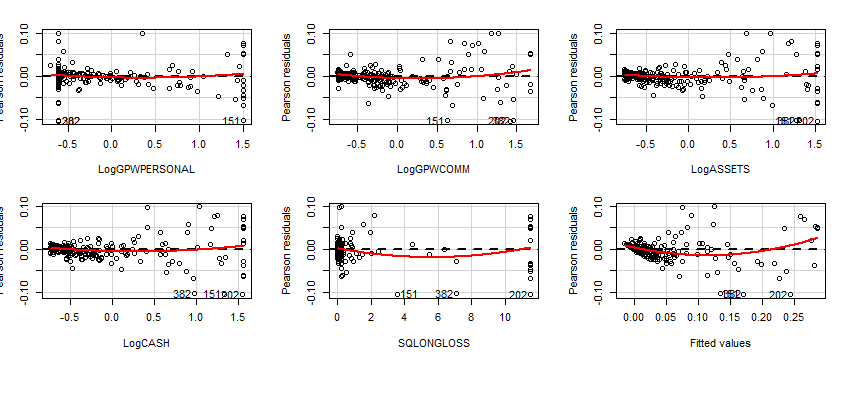




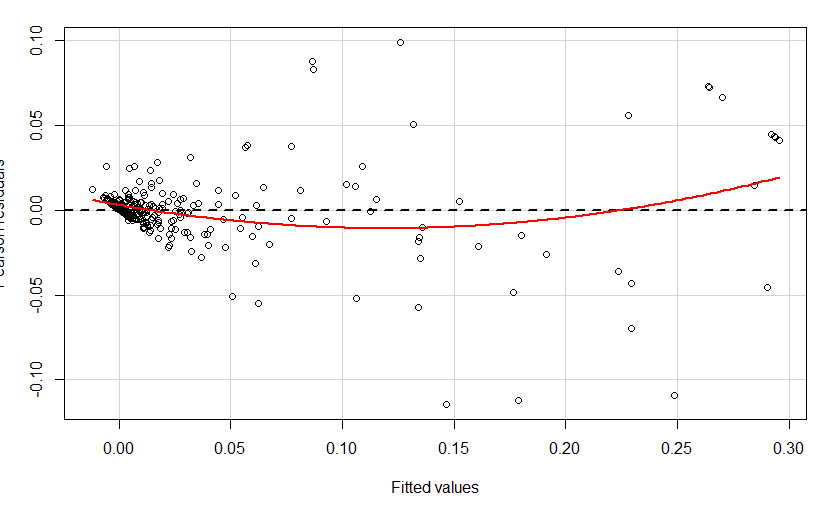
**Model result in second iteration-**

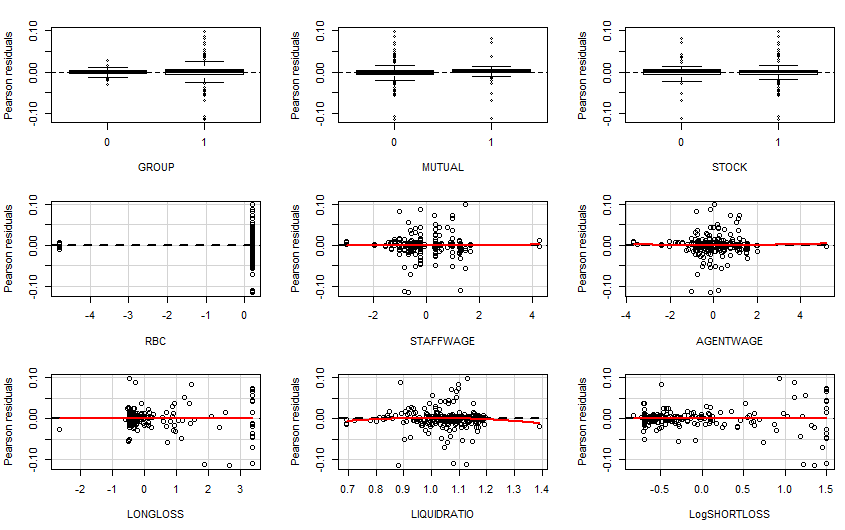


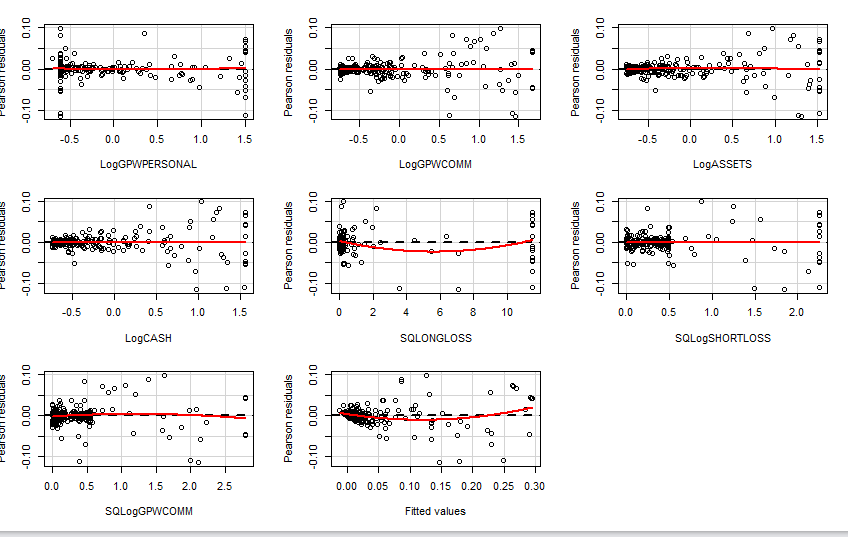




**#model give three observation as influential(151,225,382), let's try the same model without thses observations**

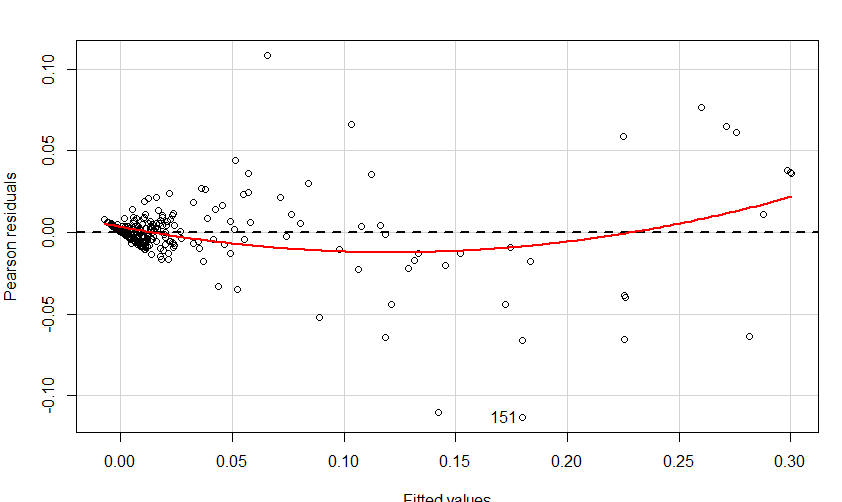


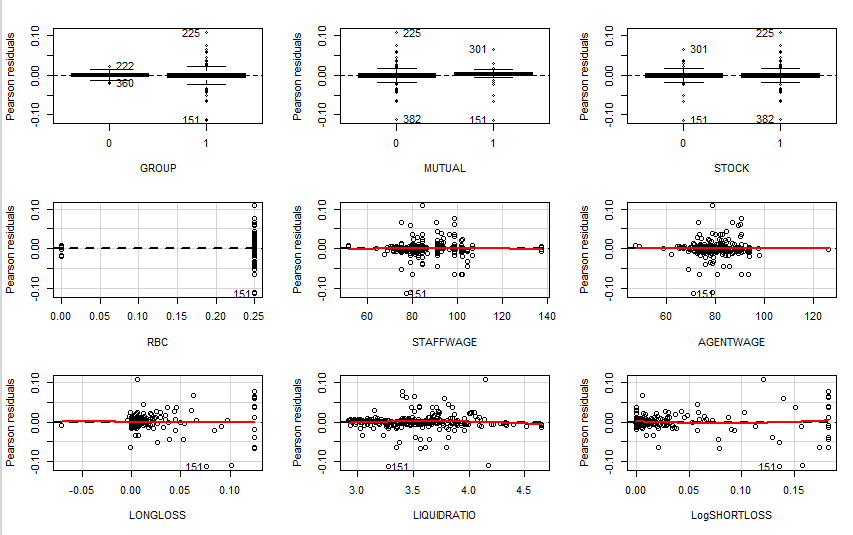


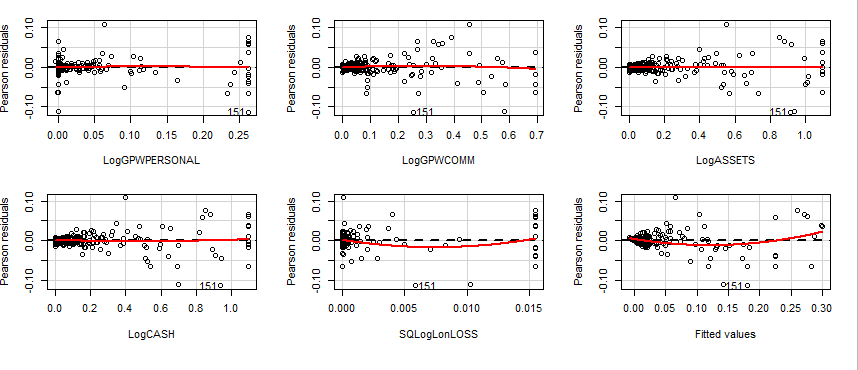


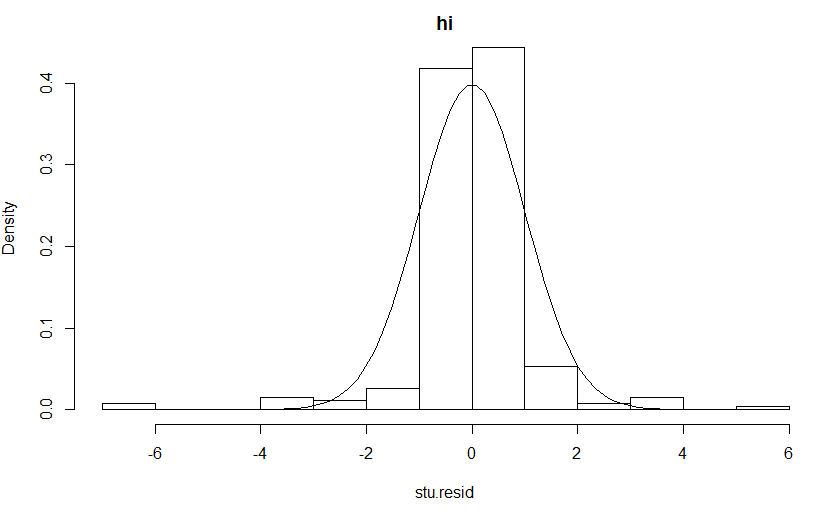
**There is a skewness in SQLONGLOSS(right skewed), we can take log transformation.**

**Model iteration 4**



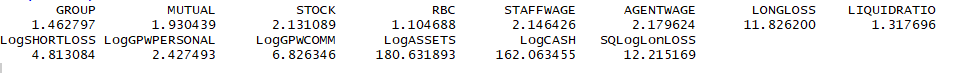






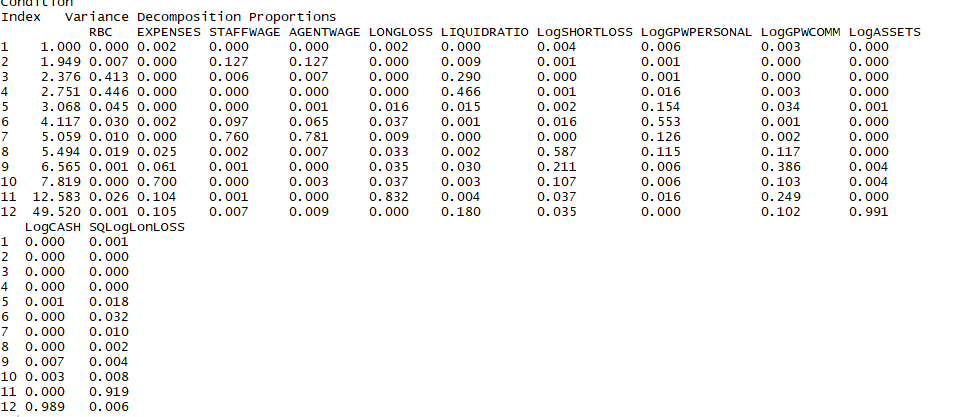
**There appears to be no problem here**

# Looking for collinearity



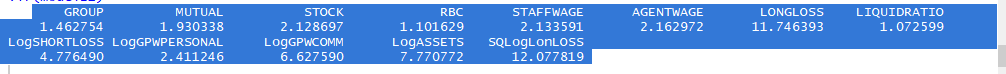
**#VIF for LONGCASH and LOGASSET is very high**

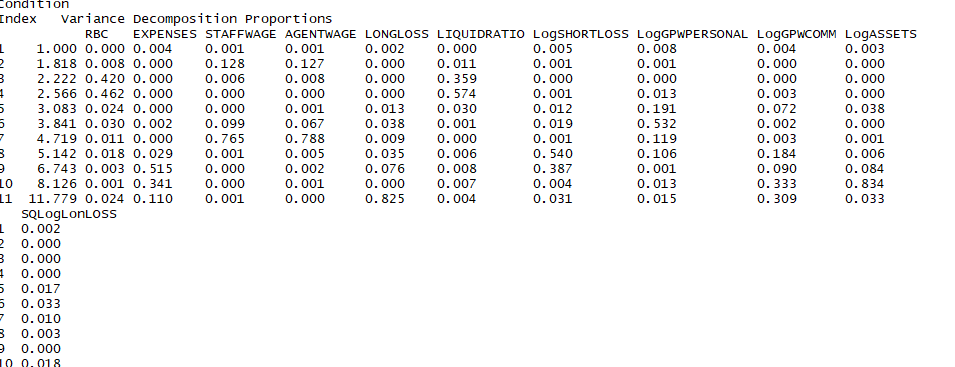
**colldiag( trainS[-c(151,175,225,235),-c(1,2,3,9,10,11,12,13,20)],center = TRUE)**



The high VIFs and Conditional Indices are here for LOGASSET and LOGCASH. We can try with one of them.

After removing LOGCASH from model





We see no major problems here, high conditional indices are may be because of the transformations done.

# BEST SUBSET SELECTION

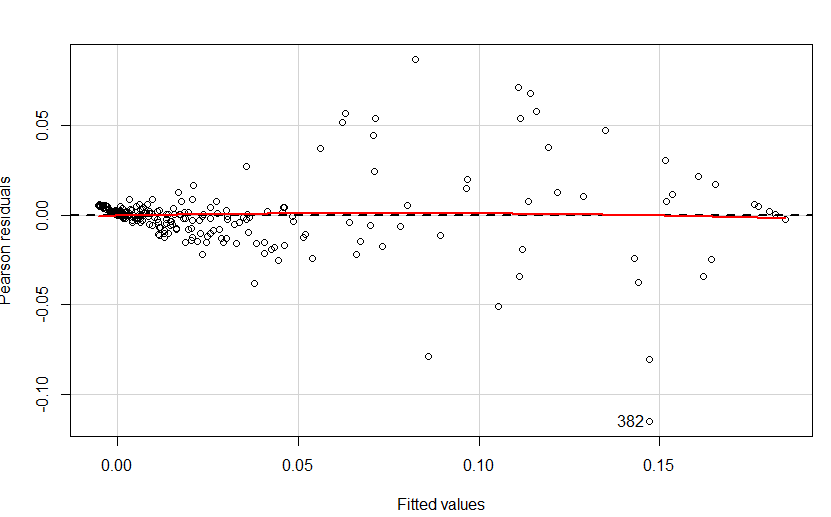
Best model given is

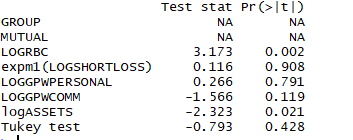
log1p(trainS$EXPENSES) ~ LogSHORTLOSS + LogGPWPERSONAL + LogGPWCOMM +

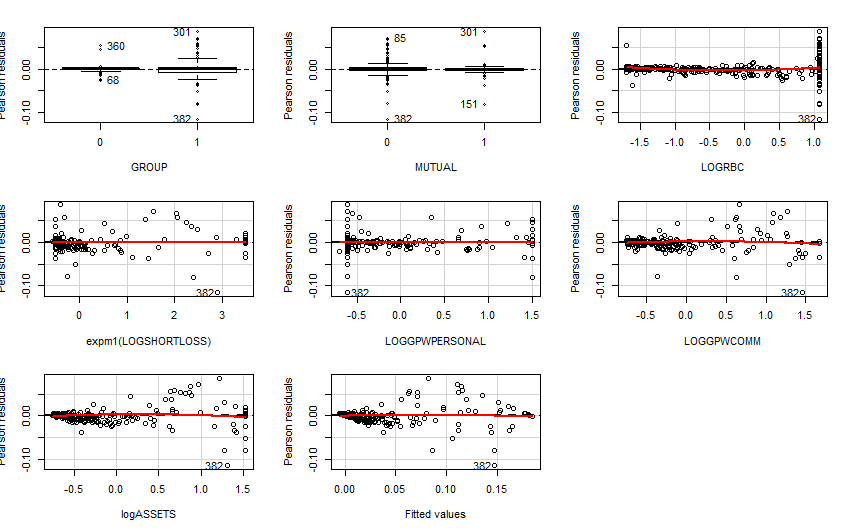
LogASSETS + SQLogLonLOSS

Model5 <- log1p(trainS1$EXPENSES) ~ GROUP + MUTUAL + LOGRBC +

expm1(LOGSHORTLOSS) + LOGGPWPERSONAL + LOGGPWCOMM + logASSETS







**No visible patterns for residual( a straight line).**