#read the dataset  
file=readRDS("loan\_data\_1.rds")

#create frame for dataset  
credit\_frame=data.frame(file)

str(credit\_frame)

## 'data.frame': 29092 obs. of 8 variables:  
## $ loan\_status : int 0 0 0 0 0 0 1 0 1 0 ...  
## $ loan\_amnt : int 5000 2400 10000 5000 3000 12000 9000 3000 10000 1000 ...  
## $ int\_rate : num 10.6 NA 13.5 NA NA ...  
## $ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 1 5 2 3 2 2 4 ...  
## $ emp\_length : int 10 25 13 3 9 11 0 3 3 0 ...  
## $ home\_ownership: Factor w/ 4 levels "MORTGAGE","OTHER",..: 4 4 4 4 4 3 4 4 4 4 ...  
## $ annual\_inc : num 24000 12252 49200 36000 48000 ...  
## $ age : int 33 31 24 39 24 28 22 22 28 22 ...

head(credit\_frame)

## loan\_status loan\_amnt int\_rate grade emp\_length home\_ownership  
## 1 0 5000 10.65 B 10 RENT  
## 2 0 2400 NA C 25 RENT  
## 3 0 10000 13.49 C 13 RENT  
## 4 0 5000 NA A 3 RENT  
## 5 0 3000 NA E 9 RENT  
## 6 0 12000 12.69 B 11 OWN  
## annual\_inc age  
## 1 24000 33  
## 2 12252 31  
## 3 49200 24  
## 4 36000 39  
## 5 48000 24  
## 6 75000 28

sapply(credit\_frame,function(x) sum(is.na(x)))

## loan\_status loan\_amnt int\_rate grade emp\_length   
## 0 0 2776 0 809   
## home\_ownership annual\_inc age   
## 0 0 0

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.4.4

CrossTable(credit\_frame$loan\_status)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 29092   
##   
##   
## | 0 | 1 |   
## |-----------|-----------|  
## | 25865 | 3227 |   
## | 0.889 | 0.111 |   
## |-----------|-----------|  
##   
##   
##   
##

CrossTable(credit\_frame$home\_ownership)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 29092   
##   
##   
## | MORTGAGE | OTHER | OWN | RENT |   
## |-----------|-----------|-----------|-----------|  
## | 12002 | 97 | 2301 | 14692 |   
## | 0.413 | 0.003 | 0.079 | 0.505 |   
## |-----------|-----------|-----------|-----------|  
##   
##   
##   
##

CrossTable(credit\_frame$grade)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 29092   
##   
##   
## | A | B | C | D | E |   
## |-----------|-----------|-----------|-----------|-----------|  
## | 9649 | 9329 | 5748 | 3231 | 868 |   
## | 0.332 | 0.321 | 0.198 | 0.111 | 0.030 |   
## |-----------|-----------|-----------|-----------|-----------|  
##   
##   
## | F | G |   
## |-----------|-----------|  
## | 211 | 56 |   
## | 0.007 | 0.002 |   
## |-----------|-----------|  
##   
##   
##   
##

print(summary(credit\_frame$loan\_amnt))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 500 5000 8000 9594 12250 35000

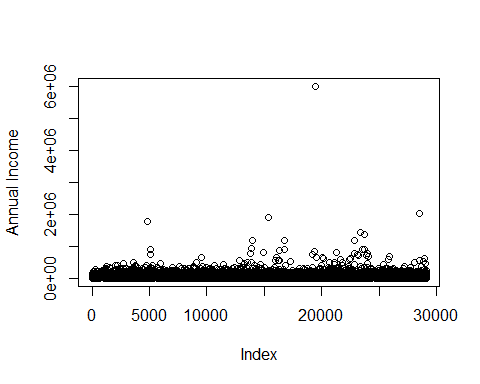
print(summary(credit\_frame$int\_rate))

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 5.42 7.90 10.99 11.00 13.47 23.22 2776

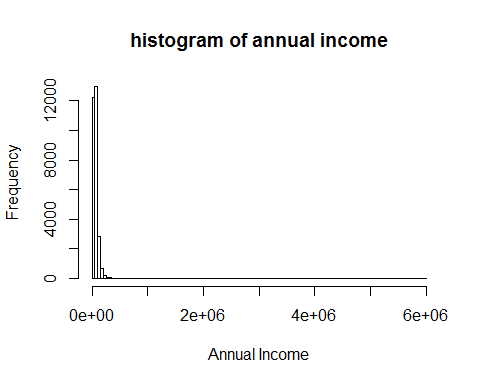
print(summary(credit\_frame$emp\_length))

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 2.000 4.000 6.145 8.000 62.000 809

plot(credit\_frame$annual\_inc,ylab="Annual Income")



hist(credit\_frame$annual\_inc,sqrt(nrow(credit\_frame)),xlab = "Annual Income",main = "histogram of annual income")

 From scatter plot extrem value is removed

index\_outlier\_annua\_inc<-which(credit\_frame$annual\_inc>3000000)  
 index\_outlier\_annua\_inc

## [1] 19486

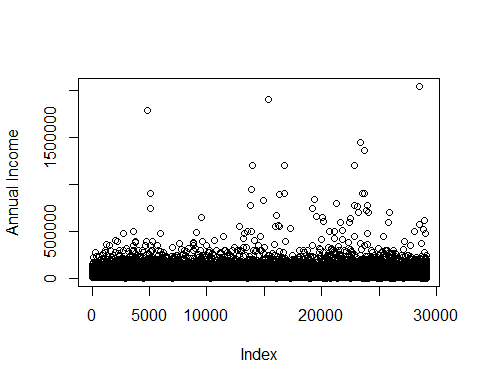
nrow(credit\_frame)

## [1] 29092

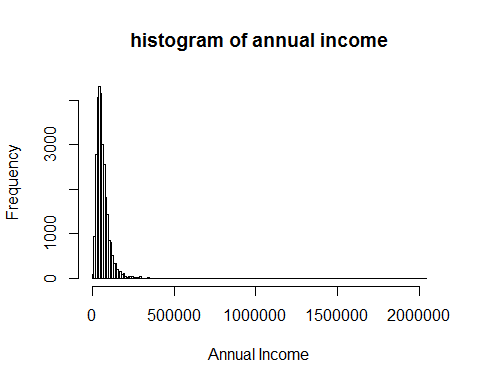
credit\_frame<-credit\_frame[-index\_outlier\_annua\_inc,]  
 nrow(credit\_frame)

## [1] 29091

plot(credit\_frame$annual\_inc,ylab="Annual Income")

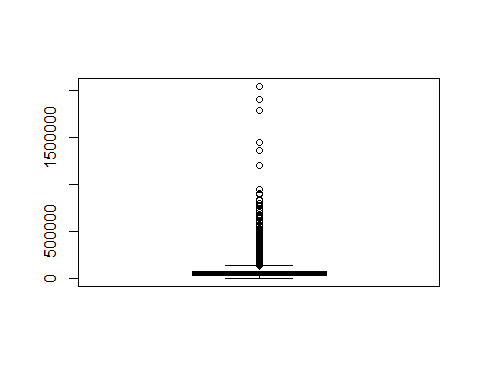


hist(credit\_frame$annual\_inc,sqrt(nrow(credit\_frame)),xlab = "Annual Income",main = "histogram of annual income")



still not look normanl

boxplot(credit\_frame$annual\_inc)



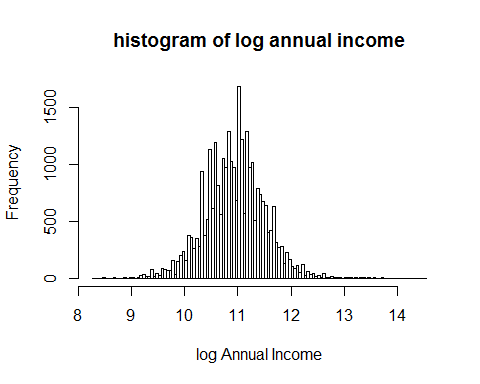
summary(credit\_frame$annual\_inc)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4000 40000 56400 66965 80000 2039784

temp\_vector<-c("log\_annual\_income")  
 credit\_frame[,temp\_vector]<-NA

credit\_frame$log\_annual\_income<-log(credit\_frame$annual\_inc)

hist(credit\_frame$log\_annual\_inc,sqrt(nrow(credit\_frame)),xlab = "log Annual Income",main = "histogram of log annual income")

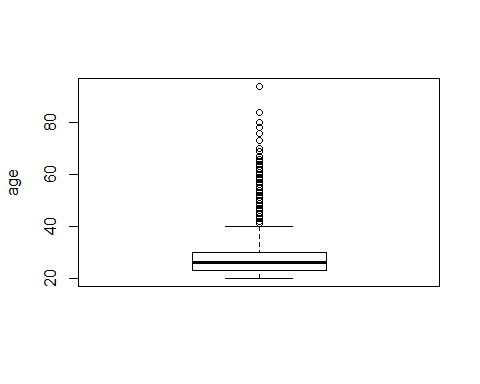


summary(credit\_frame$log\_annual\_income)

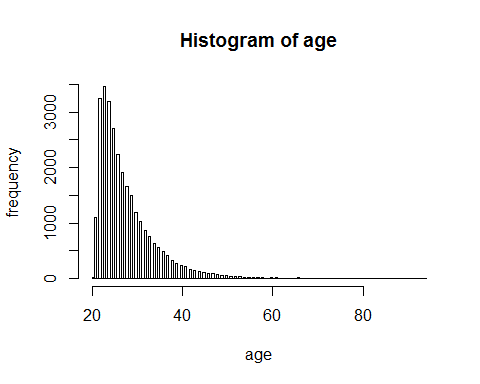
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8.294 10.597 10.940 10.942 11.290 14.528

Looks normal distribution acepted

boxplot(credit\_frame$age,ylab="age")



hist(credit\_frame$age,sqrt(nrow(credit\_frame)),ylab = "frequency",xlab = "age",main = "Histogram of age")



summary(credit\_frame$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.0 23.0 26.0 27.7 30.0 94.0

Looks constant mean and median So distribution is quet normal Interest rate has some missing values so fill it with mean

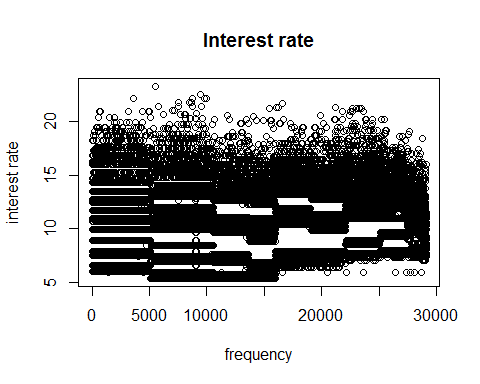
summary(credit\_frame$int\_rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 5.42 7.90 10.99 11.00 13.47 23.22 2776

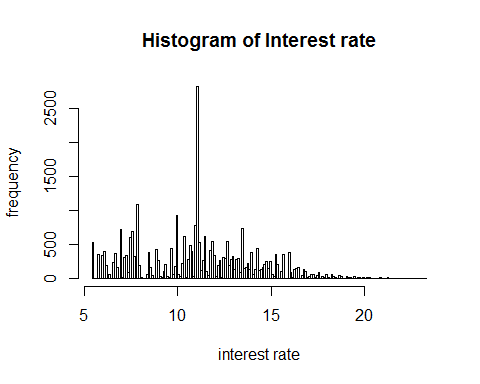
index\_int\_rate\_NA<-which(is.na(credit\_frame$int\_rate))  
credit\_frame$int\_rate[index\_int\_rate\_NA]<-mean(credit\_frame$int\_rate,na.rm = TRUE)  
summary(credit\_frame$int\_rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.42 8.49 11.00 11.00 13.11 23.22

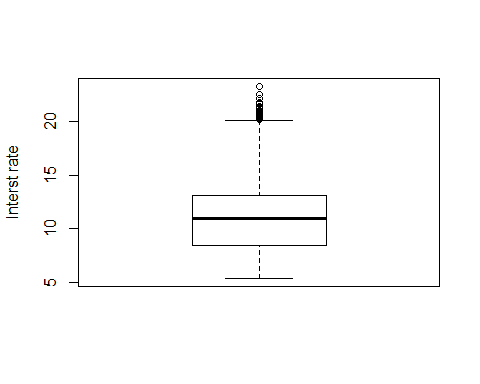
plot(credit\_frame$int\_rate,xlab = "frequency",ylab = "interest rate",main = "Interest rate")



hist(credit\_frame$int\_rate,sqrt(nrow(credit\_frame)),ylab = "frequency",xlab = "interest rate",main = "Histogram of Interest rate")



boxplot(credit\_frame$int\_rate,ylab="Interst rate")



summary(credit\_frame$int\_rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.42 8.49 11.00 11.00 13.11 23.22

From distribution and summary it looks normal and acepted

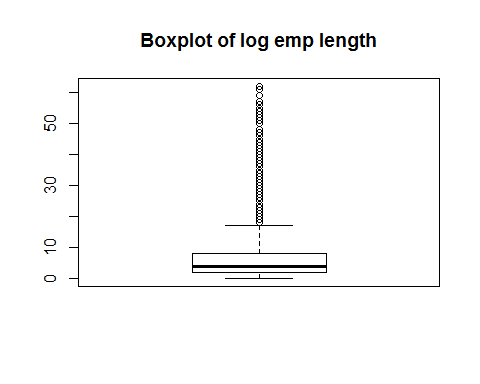
summary(credit\_frame$emp\_length)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 2.000 4.000 6.145 8.000 62.000 809

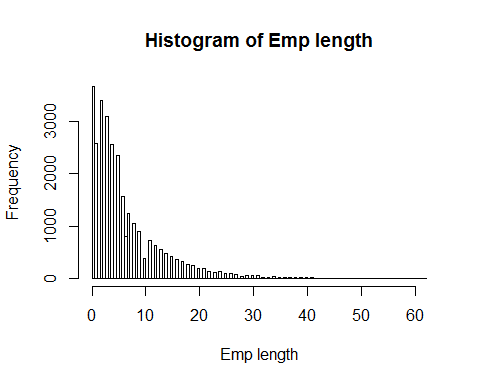
index\_emp\_length\_NA<-which(is.na(credit\_frame$emp\_length))  
credit\_frame$emp\_length[index\_emp\_length\_NA]<-mean(credit\_frame$emp\_length,na.rm = TRUE)  
summary(credit\_frame$emp\_length)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 4.000 6.145 8.000 62.000

boxplot(credit\_frame$emp\_length,main="Boxplot of log emp length")



hist(credit\_frame$emp\_length,sqrt(nrow(credit\_frame)),ylab = "Frequency",xlab = "Emp length",main = "Histogram of Emp length")

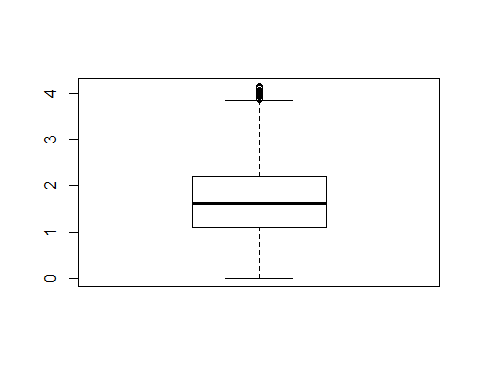
 Emp length has extrem values so preprocess it

credit\_frame$emp\_length<-credit\_frame$emp\_length+1

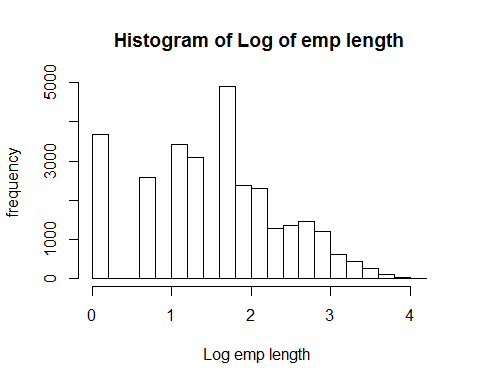
temp\_vector<-c("log\_emp\_length")  
 credit\_frame[,temp\_vector]<-NA

credit\_frame$log\_emp\_length<-log(credit\_frame$emp\_length)

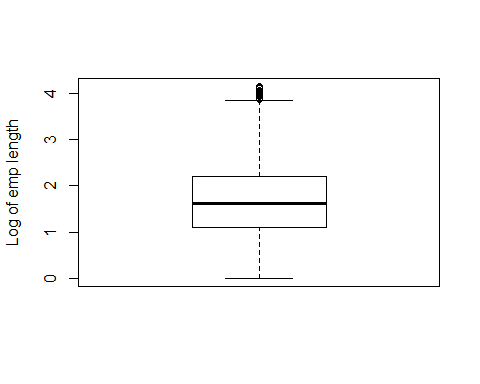
boxplot(credit\_frame$log\_emp\_length)



hist(credit\_frame$log\_emp\_length,ylab = "frequency",xlab ="Log emp length ", main = "Histogram of Log of emp length")



boxplot(credit\_frame$log\_emp\_length,ylab="Log of emp length")

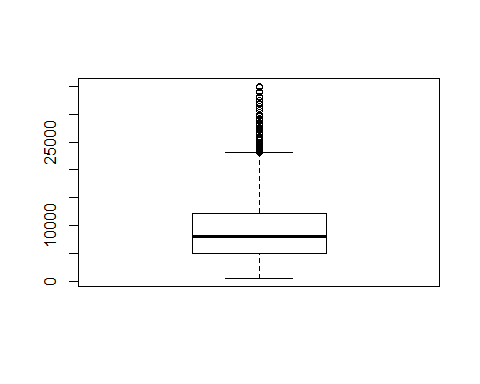


summary(credit\_frame$log\_emp\_length)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.099 1.609 1.591 2.197 4.143

Looks normal because normal mean and median

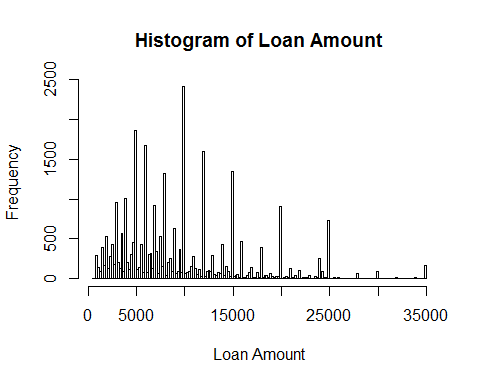
boxplot(credit\_frame$loan\_amnt)



summary(credit\_frame$loan\_amnt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 500 5000 8000 9594 12250 35000

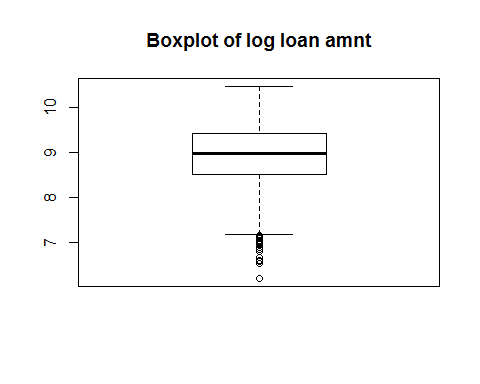
hist(credit\_frame$loan\_amnt,sqrt(nrow(credit\_frame)),ylab = "Frequency",xlab = "Loan Amount", main = "Histogram of Loan Amount")



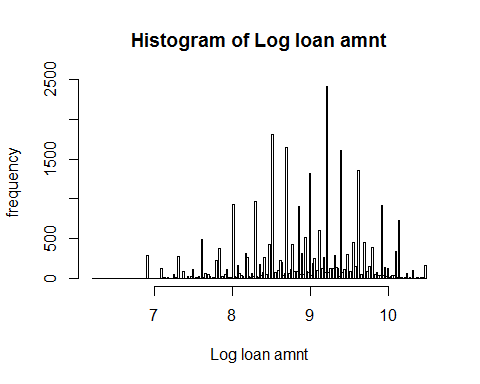
temp\_vector<-c("log\_loan\_amnt")  
 credit\_frame[,temp\_vector]<-NA

credit\_frame$log\_loan\_amnt<-log(credit\_frame$loan\_amnt)

boxplot(credit\_frame$log\_loan\_amnt,main="Boxplot of log loan amnt")



hist(credit\_frame$log\_loan\_amnt,sqrt(nrow(credit\_frame)),ylab = "frequency",xlab = "Log loan amnt",main = "Histogram of Log loan amnt")



summary(credit\_frame$log\_loan\_amnt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.215 8.517 8.987 8.942 9.413 10.463

lets create some new features

CrossTable(credit\_frame$grade,credit\_frame$loan\_status,prop.r = TRUE,prop.c = FALSE,prop.t = FALSE,prop.chisq = FALSE)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Row Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 29091   
##   
##   
## | credit\_frame$loan\_status   
## credit\_frame$grade | 0 | 1 | Row Total |   
## -------------------|-----------|-----------|-----------|  
## A | 9084 | 565 | 9649 |   
## | 0.941 | 0.059 | 0.332 |   
## -------------------|-----------|-----------|-----------|  
## B | 8344 | 985 | 9329 |   
## | 0.894 | 0.106 | 0.321 |   
## -------------------|-----------|-----------|-----------|  
## C | 4903 | 844 | 5747 |   
## | 0.853 | 0.147 | 0.198 |   
## -------------------|-----------|-----------|-----------|  
## D | 2651 | 580 | 3231 |   
## | 0.820 | 0.180 | 0.111 |   
## -------------------|-----------|-----------|-----------|  
## E | 692 | 176 | 868 |   
## | 0.797 | 0.203 | 0.030 |   
## -------------------|-----------|-----------|-----------|  
## F | 155 | 56 | 211 |   
## | 0.735 | 0.265 | 0.007 |   
## -------------------|-----------|-----------|-----------|  
## G | 35 | 21 | 56 |   
## | 0.625 | 0.375 | 0.002 |   
## -------------------|-----------|-----------|-----------|  
## Column Total | 25864 | 3227 | 29091 |   
## -------------------|-----------|-----------|-----------|  
##   
##

Accordingly from above observation the loan status depends on the grade of the customer

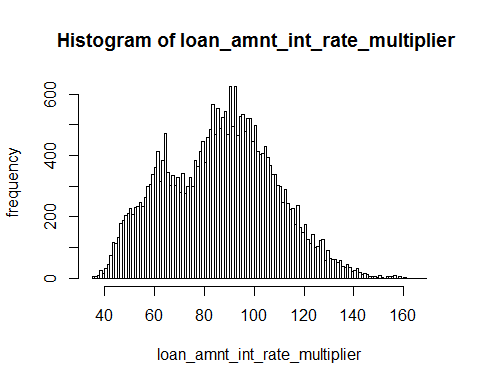
temp\_vector<-c("grade\_multiplier")  
 credit\_frame[,temp\_vector]<-NA

credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='A']<-(0.941))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='B']<-(0.894))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='C']<-(0.853))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='D']<-(0.820))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='E']<-(0.797))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='F']<-(0.735))  
credit\_frame<-within(credit\_frame,grade\_multiplier[grade=='G']<-(0.625))

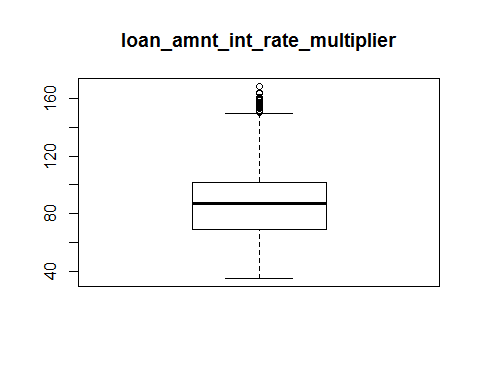
temp\_vector<-c("loan\_amnt\_int\_rate\_multiplier")  
 credit\_frame[,temp\_vector]<-NA

credit\_frame$loan\_amnt\_int\_rate\_multiplier<-((credit\_frame$int\_rate)\*(credit\_frame$log\_loan\_amnt)\*(credit\_frame$grade\_multiplier))

hist(credit\_frame$loan\_amnt\_int\_rate\_multiplier,sqrt(nrow(credit\_frame)),ylab = "frequency",xlab = "loan\_amnt\_int\_rate\_multiplier",main = "Histogram of loan\_amnt\_int\_rate\_multiplier ")



boxplot(credit\_frame$loan\_amnt\_int\_rate\_multiplier,main="loan\_amnt\_int\_rate\_multiplier")



summary(credit\_frame$loan\_amnt\_int\_rate\_multiplier)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 35.23 69.34 87.35 86.42 101.56 168.33