##Problem Statement  
  
# ##Loan Smart is a lending advisory firm. Based on their client’s characteristic and needed loan amount they  
# advise them on which Financial Institution to apply for loan at. So far their recommendations have been  
# based hunches business experience. Now they are trying to leverage power of data that they have collected so  
# far.  
# They want to check whether given their client’s characteristics , they can predict how much interest rates  
# they will be offered by various financial institution. They want to run with proof of concept for this idea.  
# They have given us data collected for one such financial institution ABC Capitals Ltd.  
# What we need to do is to figure out whether using that data we can predict interest rate offered to client. We  
# have developed the problem the way you’d encounter problems in projects. You are given training data and  
# testing data, testing data doesnt have response values. We’ll eventually want to make prediction on this data  
# where the response is unknown.

## Loading Packages

library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(corrplot)

## corrplot 0.90 loaded

library(corrplot)

## Importing The Data  
  
ld\_train=read.csv("E:\\Data science\\R\\Additional Datasets (1)\\Additional Datasets\\loan\_data\_train (1).csv")  
ld\_test=read.csv("E:\\Data science\\R\\Additional Datasets (1)\\Additional Datasets\\loan\_data\_test (1).csv")

## Data Size and Structure  
  
dim(ld\_train)

## [1] 2200 15

glimpse(ld\_train)

## Rows: 2,200  
## Columns: 15  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <chr> "25000", "19750", "2100", "28000", "242~  
## $ Amount.Funded.By.Investors <chr> "25000", "19750", "2100", "28000", "174~  
## $ Interest.Rate <chr> "18.49%", "17.27%", "14.33%", "16.29%",~  
## $ Loan.Length <chr> "60 months", "60 months", "36 months", ~  
## $ Loan.Purpose <chr> "debt\_consolidation", "debt\_consolidati~  
## $ Debt.To.Income.Ratio <chr> "27.56%", "13.39%", "3.50%", "19.62%", ~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ FICO.Range <chr> "720-724", "710-714", "690-694", "710-7~  
## $ Open.CREDIT.Lines <chr> "11", "14", "13", "12", "6", "2", "5", ~  
## $ Revolving.CREDIT.Balance <chr> "15210", "19070", "893", "38194", "3106~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ Employment.Length <chr> "5 years", "4 years", "< 1 year", "10+ ~

glimpse(ld\_test)

## Rows: 300  
## Columns: 14  
## $ ID <int> 20093, 62445, 65248, 81822, 57923, 3377~  
## $ Amount.Requested <int> 5000, 18000, 7200, 7200, 22000, 6000, 9~  
## $ Amount.Funded.By.Investors <chr> "5000", "18000", "7200", "7200", "22000~  
## $ Loan.Length <chr> "60 months", "60 months", "60 months", ~  
## $ Loan.Purpose <chr> "moving", "debt\_consolidation", "debt\_c~  
## $ Debt.To.Income.Ratio <chr> "12.59%", "4.93%", "25.16%", "17.27%", ~  
## $ State <chr> "NY", "CA", "LA", "NY", "MI", "NJ", "CA~  
## $ Home.Ownership <chr> "RENT", "RENT", "MORTGAGE", "MORTGAGE",~  
## $ Monthly.Income <dbl> 4416.67, 5258.50, 3750.00, 3416.67, 608~  
## $ FICO.Range <chr> "690-694", "710-714", "750-754", "790-7~  
## $ Open.CREDIT.Lines <chr> "13", "6", "13", "14", "9", ".", "11", ~  
## $ Revolving.CREDIT.Balance <int> 7686, 11596, 7283, 4838, 20181, 1442, 8~  
## $ Inquiries.in.the.Last.6.Months <int> 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 2, 0, ~  
## $ Employment.Length <chr> "< 1 year", "10+ years", "6 years", "10~

##Getting rid of the IDs but keeping the test IDs in a vector. These are needed to compose the submission file  
test\_lebels=ld\_test$ID

## adding response column in test dataset to combine both test and train data

ld\_test$Interest.Rate=NA  
  
ld\_train$data="Train"  
ld\_test$data="Test"

ld\_all=rbind(ld\_train,ld\_test)

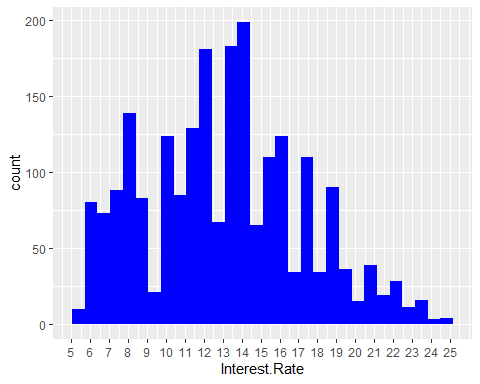
## Exploring some of the most important variables

##Response variable :IntereseRate

ld\_all=ld\_all %>% mutate(Interest.Rate=as.numeric(gsub("%","",Interest.Rate)))

ggplot(data=ld\_all[!is.na(ld\_all$Interest.Rate),],aes(x=Interest.Rate))+  
 geom\_histogram(fill="blue")+scale\_x\_continuous(breaks=seq(0,25,by=1))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



summary(ld\_all$Interest.Rate)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 5.42 10.16 13.11 13.05 15.80 24.89 300

### Data Prep

ld\_all=ld\_all %>%   
 mutate(Amount.Requested=as.numeric(Amount.Requested),  
 Debt.To.Income.Ratio=as.numeric(gsub("%","",Debt.To.Income.Ratio)),  
 Amount.Funded.By.Investors=as.numeric(Amount.Funded.By.Investors),  
 Open.CREDIT.Lines=as.numeric(Open.CREDIT.Lines),  
 Revolving.CREDIT.Balance=as.numeric(Revolving.CREDIT.Balance))

## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion  
  
## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion

glimpse(ld\_all)

## Rows: 2,500  
## Columns: 16  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <dbl> 25000, 19750, 2100, 28000, 24250, 5400,~  
## $ Amount.Funded.By.Investors <dbl> 25000.00, 19750.00, 2100.00, 28000.00, ~  
## $ Interest.Rate <dbl> 18.49, 17.27, 14.33, 16.29, 12.23, 8.90~  
## $ Loan.Length <chr> "60 months", "60 months", "36 months", ~  
## $ Loan.Purpose <chr> "debt\_consolidation", "debt\_consolidati~  
## $ Debt.To.Income.Ratio <dbl> 27.56, 13.39, 3.50, 19.62, 23.79, 6.27,~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ FICO.Range <chr> "720-724", "710-714", "690-694", "710-7~  
## $ Open.CREDIT.Lines <dbl> 11, 14, 13, 12, 6, 2, 5, 11, 24, 7, 10,~  
## $ Revolving.CREDIT.Balance <dbl> 15210, 19070, 893, 38194, 31061, 90, 84~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ Employment.Length <chr> "5 years", "4 years", "< 1 year", "10+ ~  
## $ data <chr> "Train", "Train", "Train", "Train", "Tr~

ld\_all=ld\_all %>% mutate(f1=as.numeric(substr(FICO.Range,1,3)),  
 f2=as.numeric(substr(FICO.Range,5,7)),  
 fico=0.5\*(f1+f2)) %>%   
 select(-FICO.Range,-f1,-f2)

glimpse(ld\_all)

## Rows: 2,500  
## Columns: 16  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <dbl> 25000, 19750, 2100, 28000, 24250, 5400,~  
## $ Amount.Funded.By.Investors <dbl> 25000.00, 19750.00, 2100.00, 28000.00, ~  
## $ Interest.Rate <dbl> 18.49, 17.27, 14.33, 16.29, 12.23, 8.90~  
## $ Loan.Length <chr> "60 months", "60 months", "36 months", ~  
## $ Loan.Purpose <chr> "debt\_consolidation", "debt\_consolidati~  
## $ Debt.To.Income.Ratio <dbl> 27.56, 13.39, 3.50, 19.62, 23.79, 6.27,~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ Open.CREDIT.Lines <dbl> 11, 14, 13, 12, 6, 2, 5, 11, 24, 7, 10,~  
## $ Revolving.CREDIT.Balance <dbl> 15210, 19070, 893, 38194, 31061, 90, 84~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ Employment.Length <chr> "5 years", "4 years", "< 1 year", "10+ ~  
## $ data <chr> "Train", "Train", "Train", "Train", "Tr~  
## $ fico <dbl> 722, 712, 692, 712, 732, 787, 667, 737,~

table(ld\_all$Employment.Length)

##   
## . < 1 year 1 year 10+ years 2 years 3 years 4 years 5 years   
## 2 249 177 653 243 235 191 202   
## 6 years 7 years 8 years 9 years n/a   
## 163 127 108 72 77

ld\_all %>%   
 group\_by(Employment.Length) %>%   
 summarise(avg\_IR=mean(Interest.Rate,na.rm = T))

## # A tibble: 14 x 2  
## Employment.Length avg\_IR  
## <chr> <dbl>  
## 1 . 12.7  
## 2 < 1 year 12.9  
## 3 1 year 12.5  
## 4 10+ years 13.3  
## 5 2 years 12.9  
## 6 3 years 12.7  
## 7 4 years 13.0  
## 8 5 years 13.6  
## 9 6 years 13.3  
## 10 7 years 13.1  
## 11 8 years 13.0  
## 12 9 years 12.9  
## 13 n/a 12.5  
## 14 <NA> NaN

ld\_all=ld\_all %>% mutate(el=ifelse(substr(Employment.Length,1,2)=="10",10,Employment.Length),  
 el=ifelse(substr(Employment.Length,1,1)=="<",1,el),  
 el=gsub("years","",el),  
 el=gsub("year","",el),  
 el=as.numeric(el)) %>%   
 select(-Employment.Length)

## Warning in mask$eval\_all\_mutate(quo): NAs introduced by coercion

glimpse(ld\_all)

## Rows: 2,500  
## Columns: 16  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <dbl> 25000, 19750, 2100, 28000, 24250, 5400,~  
## $ Amount.Funded.By.Investors <dbl> 25000.00, 19750.00, 2100.00, 28000.00, ~  
## $ Interest.Rate <dbl> 18.49, 17.27, 14.33, 16.29, 12.23, 8.90~  
## $ Loan.Length <chr> "60 months", "60 months", "36 months", ~  
## $ Loan.Purpose <chr> "debt\_consolidation", "debt\_consolidati~  
## $ Debt.To.Income.Ratio <dbl> 27.56, 13.39, 3.50, 19.62, 23.79, 6.27,~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ Open.CREDIT.Lines <dbl> 11, 14, 13, 12, 6, 2, 5, 11, 24, 7, 10,~  
## $ Revolving.CREDIT.Balance <dbl> 15210, 19070, 893, 38194, 31061, 90, 84~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ data <chr> "Train", "Train", "Train", "Train", "Tr~  
## $ fico <dbl> 722, 712, 692, 712, 732, 787, 667, 737,~  
## $ el <dbl> 5, 4, 1, 10, 10, NA, 2, 1, NA, NA, 10, ~

table(ld\_all$Loan.Length)

##   
## . 36 months 60 months   
## 1 1950 548

ld\_all =ld\_all %>% mutate(ll\_36=as.numeric(Loan.Length=="36 months")) %>%   
 select(-Loan.Length)

### Corelation with Response Variable

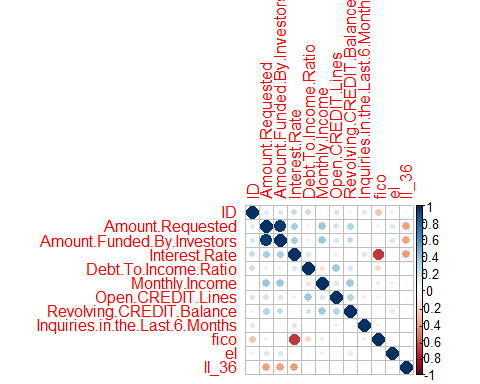
numVar=ld\_all[which(sapply(ld\_all,is.numeric))]  
numVarNames=names(numVar)  
numVarNames

## [1] "ID" "Amount.Requested"   
## [3] "Amount.Funded.By.Investors" "Interest.Rate"   
## [5] "Debt.To.Income.Ratio" "Monthly.Income"   
## [7] "Open.CREDIT.Lines" "Revolving.CREDIT.Balance"   
## [9] "Inquiries.in.the.Last.6.Months" "fico"   
## [11] "el" "ll\_36"

cat(" There are ",length(numVarNames)," Numeric Variables")

## There are 12 Numeric Variables

cor\_numVar=cor(numVar, use="pairwise.complete.obs")  
corrplot(cor\_numVar)



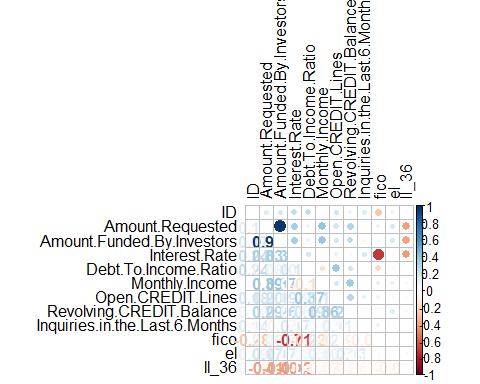
cor\_sorted=as.matrix(sort(cor\_numVar[,"Interest.Rate"],decreasing = TRUE))  
cor\_sorted

## [,1]  
## Interest.Rate 1.000000000  
## Amount.Funded.By.Investors 0.335370627  
## Amount.Requested 0.330000558  
## ID 0.244658605  
## Debt.To.Income.Ratio 0.177861105  
## Inquiries.in.the.Last.6.Months 0.168890022  
## Open.CREDIT.Lines 0.091181493  
## Revolving.CREDIT.Balance 0.061473148  
## el 0.045090097  
## Monthly.Income 0.008227301  
## ll\_36 -0.424350562  
## fico -0.708393891

cor\_high=names(which(apply(cor\_sorted, 1, function(x) abs(x)>0.5)))  
#select only high corelations  
  
cor\_high

## [1] "Interest.Rate" "fico"

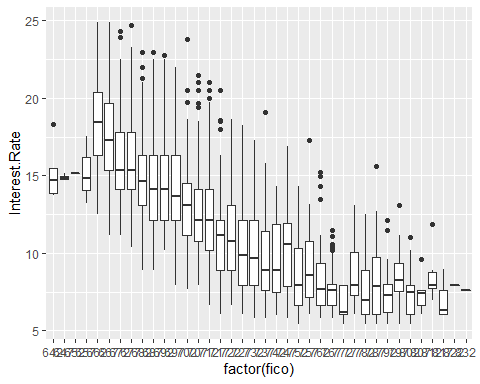
corrplot.mixed(cor\_numVar, tl.col="black", tl.pos = "lt")



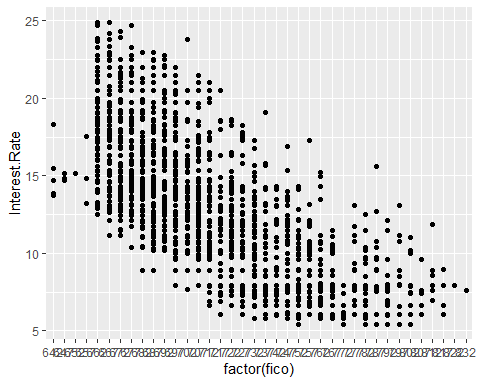
# The variable Amount\_Funded\_By\_Investors happens to have high correlation with Interest.Rate [our target],  
# but we should still drop this variables because it contains information which will not be available at the time  
# when we need to use this model.

ld\_all$Amount.Funded.By.Investors=NULL

ggplot(data=ld\_all[!is.na(ld\_all$Interest.Rate),],aes(x=factor(fico),y=Interest.Rate))+geom\_boxplot()



ggplot(data=ld\_all[!is.na(ld\_all$Interest.Rate),],aes(x=factor(fico),y=Interest.Rate))+geom\_point()



### Handling categorical Variable

table(ld\_all$Loan.Purpose)

##   
## car credit\_card debt\_consolidation educational   
## 50 444 1307 15   
## home\_improvement house major\_purchase medical   
## 152 20 101 30   
## moving other renewable\_energy small\_business   
## 29 200 4 87   
## vacation wedding   
## 21 39

round(tapply(ld\_all$Interest.Rate,ld\_all$Loan.Purpose,mean,na.rm=T))

## car credit\_card debt\_consolidation educational   
## 11 13 14 10   
## home\_improvement house major\_purchase medical   
## 12 14 11 11   
## moving other renewable\_energy small\_business   
## 14 13 8 13   
## vacation wedding   
## 12 12

# we’ll combine categories into new one which have similar reponse rate as per the table obtained abvoe and  
# make dummies for them as well while we are at it.

ld\_all=ld\_all %>%  
mutate(lp\_10=as.numeric(Loan.Purpose=='educational'),  
lp\_11=as.numeric(Loan.Purpose %in% c("major\_purchase","medical","car")),  
lp\_12=as.numeric(Loan.Purpose %in% c("vacation","wedding","home\_improvement")),  
lp\_13=as.numeric(Loan.Purpose %in% c("other","small\_business","credit\_card")),  
lp\_14=as.numeric(Loan.Purpose %in% c("debt\_consolidation","house","moving"))) %>%  
select(-Loan.Purpose)

glimpse(ld\_all)

## Rows: 2,500  
## Columns: 19  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <dbl> 25000, 19750, 2100, 28000, 24250, 5400,~  
## $ Interest.Rate <dbl> 18.49, 17.27, 14.33, 16.29, 12.23, 8.90~  
## $ Debt.To.Income.Ratio <dbl> 27.56, 13.39, 3.50, 19.62, 23.79, 6.27,~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ Open.CREDIT.Lines <dbl> 11, 14, 13, 12, 6, 2, 5, 11, 24, 7, 10,~  
## $ Revolving.CREDIT.Balance <dbl> 15210, 19070, 893, 38194, 31061, 90, 84~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ data <chr> "Train", "Train", "Train", "Train", "Tr~  
## $ fico <dbl> 722, 712, 692, 712, 732, 787, 667, 737,~  
## $ el <dbl> 5, 4, 1, 10, 10, NA, 2, 1, NA, NA, 10, ~  
## $ ll\_36 <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, ~  
## $ lp\_10 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ lp\_11 <dbl> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~  
## $ lp\_12 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~  
## $ lp\_13 <dbl> 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, ~  
## $ lp\_14 <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, ~

## Custom Fuction for creating dummy variables  
  
CreateDummies=function(data,var,freq\_cutoff=0){  
t=table(data[,var])  
t=t[t>freq\_cutoff]  
t=sort(t)  
categories=names(t)[-1]  
for( cat in categories){  
name=paste(var,cat,sep="\_")  
name=gsub(" ","",name)  
name=gsub("-","\_",name)  
name=gsub("\\?","Q",name)  
name=gsub("<","LT\_",name)  
name=gsub("\\+","",name)  
data[,name]=as.numeric(data[,var]==cat)  
}  
data[,var]=NULL  
return(data)  
}

NAcol=which(colSums(is.na(ld\_all))>0)  
sort(colSums(sapply(ld\_all[NAcol],is.na)),decreasing = T)

## Interest.Rate el   
## 300 80   
## Open.CREDIT.Lines Amount.Requested   
## 9 5   
## Revolving.CREDIT.Balance Monthly.Income   
## 5 3   
## Inquiries.in.the.Last.6.Months ID   
## 3 1   
## Debt.To.Income.Ratio State   
## 1 1   
## Home.Ownership ll\_36   
## 1 1   
## lp\_10   
## 1

#It doesnt make sense to keep an observation with a missing ID, we’ll filter that  
ld\_all=ld\_all[!is.na(ld\_all$ID),]

glimpse(ld\_all)

## Rows: 2,499  
## Columns: 19  
## $ ID <int> 79542, 75473, 67265, 80167, 17240, 3273~  
## $ Amount.Requested <dbl> 25000, 19750, 2100, 28000, 24250, 5400,~  
## $ Interest.Rate <dbl> 18.49, 17.27, 14.33, 16.29, 12.23, 8.90~  
## $ Debt.To.Income.Ratio <dbl> 27.56, 13.39, 3.50, 19.62, 23.79, 6.27,~  
## $ State <chr> "VA", "NY", "LA", "NV", "OH", "OH", "CA~  
## $ Home.Ownership <chr> "MORTGAGE", "MORTGAGE", "OWN", "MORTGAG~  
## $ Monthly.Income <dbl> 8606.56, 6737.50, 1000.00, 7083.33, 583~  
## $ Open.CREDIT.Lines <dbl> 11, 14, 13, 12, 6, 2, 5, 11, 24, 7, 10,~  
## $ Revolving.CREDIT.Balance <dbl> 15210, 19070, 893, 38194, 31061, 90, 84~  
## $ Inquiries.in.the.Last.6.Months <int> 3, 3, 1, 1, 2, 2, 0, 1, 0, 1, 0, 0, 0, ~  
## $ data <chr> "Train", "Train", "Train", "Train", "Tr~  
## $ fico <dbl> 722, 712, 692, 712, 732, 787, 667, 737,~  
## $ el <dbl> 5, 4, 1, 10, 10, NA, 2, 1, NA, NA, 10, ~  
## $ ll\_36 <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, ~  
## $ lp\_10 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~  
## $ lp\_11 <dbl> 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~  
## $ lp\_12 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~  
## $ lp\_13 <dbl> 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, ~  
## $ lp\_14 <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, ~

sort(table(ld\_all$State),decreasing = T)

##   
## CA NY TX FL IL GA PA NJ VA MA OH MD NC CO WA CT AZ MI AL MN   
## 433 255 174 169 101 97 96 94 78 73 71 68 64 61 58 50 46 45 38 38   
## MO NV OR SC WI KY LA KS OK UT NH RI WV AR NM HI AK DC DE MT   
## 33 32 30 27 26 23 22 21 21 16 15 15 14 13 13 12 11 11 8 7   
## VT SD WY IN . IA MS   
## 5 4 4 3 1 1 1

for(col in c("State","Home.Ownership")){  
ld\_all=CreateDummies(ld\_all,col,100)  
}

for(col in names(ld\_all)){  
if(sum(is.na(ld\_all[,col]))>0 & !(col %in% c("ID","data","Interest.Rate"))){  
ld\_all[is.na(ld\_all[,col]),col]=mean(ld\_all[ld\_all$data=="Train",col],na.rm=T)  
}  
}

ld\_train=ld\_all %>% filter(data=='train') %>% select(-data)  
ld\_test=ld\_all %>% filter(data=='test') %>% select(-data,-Interest.Rate)

set.seed(2)  
s=sample(1:nrow(ld\_train),0.7\*nrow(ld\_train))  
ld\_train1=ld\_train[s,]  
ld\_train2=ld\_train[-s,]