Machine learning multiple regression model to predict interest rate on loan data

By Deepak kumar singh

INDEX

TOPIC	PAGE NO.
1. Aim & Abstract	02-03
2. Understanding of data	03
3. Cleaning of data and feature engineering	03
4. Visualization	04-09
5.Sampling	10
6.Regression model	11
7. word cloud	11
. 8.Result and conclusion	12

Aim of the Project : -----

The aim of the project is to build the machine learning model to predict interest rate on the basis parameters related to loan.

Symbol used in the project

X1	Interest Rate on the loan
X2	A unique id for the loan.
X3	A unique id assigned for the borrower.
X4	Loan amount requested
X5	Loan amount funded
X6	Investor-funded portion of loan
X7	Number of payments (36 or 60)
X8	Loan grade
X9	Loan subgrade
X10	Employer or job title (self-filled)
X11	Number of years employed (0 to 10; 10 = 10 or more)
X12	Home ownership status: RENT, OWN, MORTGAGE, OTHER.
X13	Annual income of borrower
X14	Income verified, not verified, or income source was verified
X15	Date loan was issued
X16	Reason for loan provided by borrower
X17	Loan category, as provided by borrower
X18	Loan title, as provided by borrower
X19	First 3 numbers of zip code
X20	State of borrower
X21	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding
X22	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
X23	Date the borrower's earliest reported credit line was opened
X24	Number of inquiries by creditors during the past 6 months.
X25	Number of months since the borrower's last delinquency.
X26	Number of months since the last public record.
X27	Number of open credit lines in the borrower's credit file.
X28	Number of derogatory public records
X29	Total credit revolving balance
X30	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving
X31	The total number of credit lines currently in the borrower's credit file
X32	The initial listing status of the loan. Possible values are – W, F

Abstract

I have developed a machine learning model with 95.5 % accuracy on train data to predict interest rate of loan .

Steps involved to make machine learning model are

1. Understanding of data

2. cleaning of data

3.sampling of data

4. Multiple linear regression model

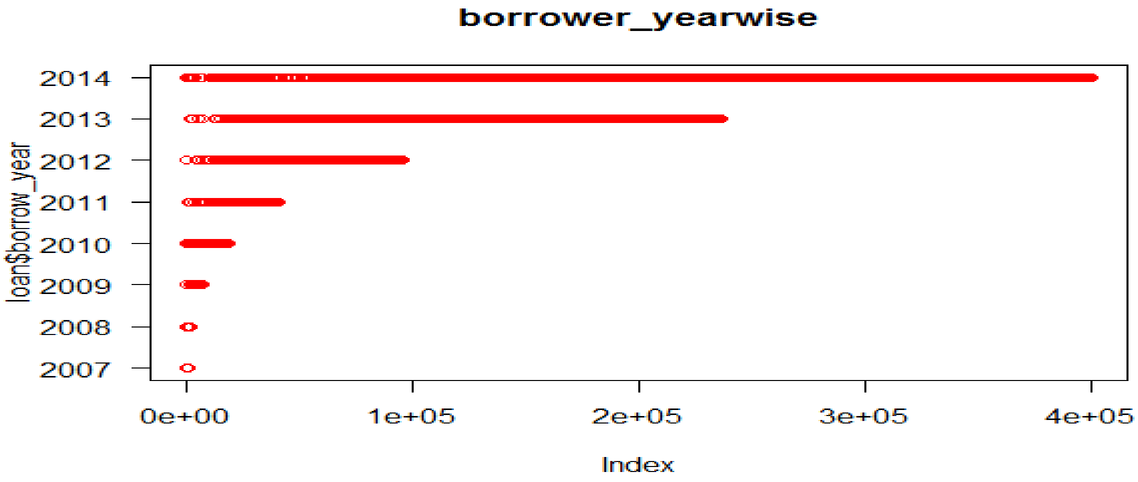
1. Understanding of data

I have two set of data one with dimension (400000, 32) that helps to build model while one with dimension (80000, 32) on which model is used to predict interest rate. Using different set of commands I got to know about its structure that helps me to process further steps .

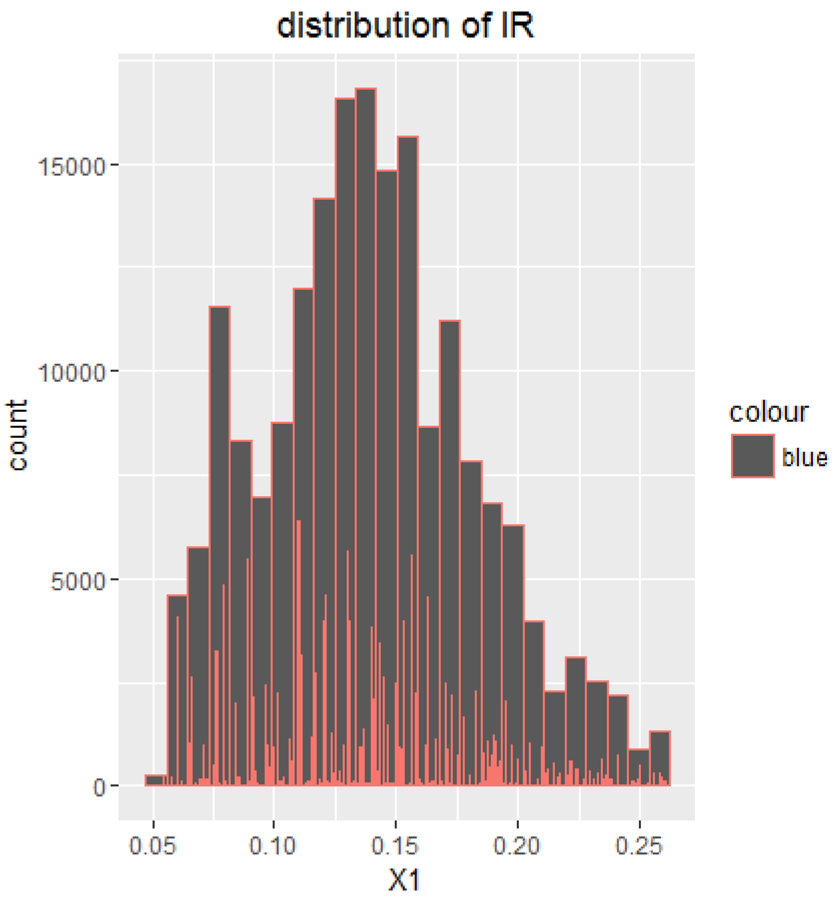
2. cleaning of data and feature engineering

data has been converted to proper usable data set by removing "%" , "$","" etc and also has been converted into proper structure of data. I have derived year and month of borrow and credit year separately in a hope to extract some valuable information like in which year and in which month ,major loan has been passed. This graph clearly depicts the number of borrower increases year wise and almost we get logarithmic increment of borrower rate with time .

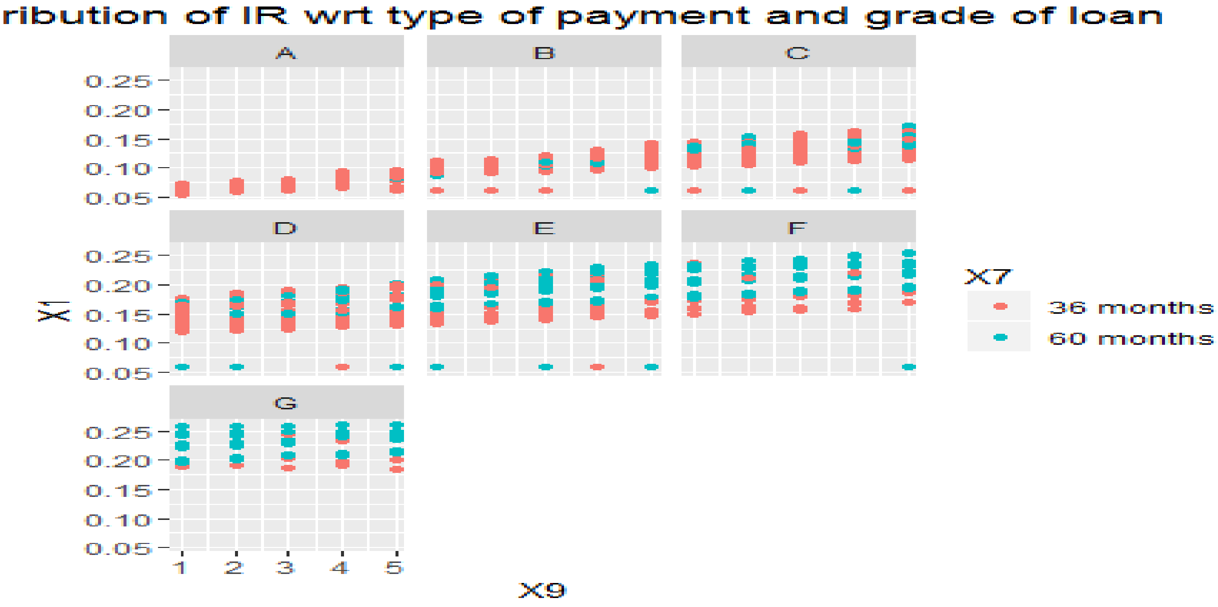
.



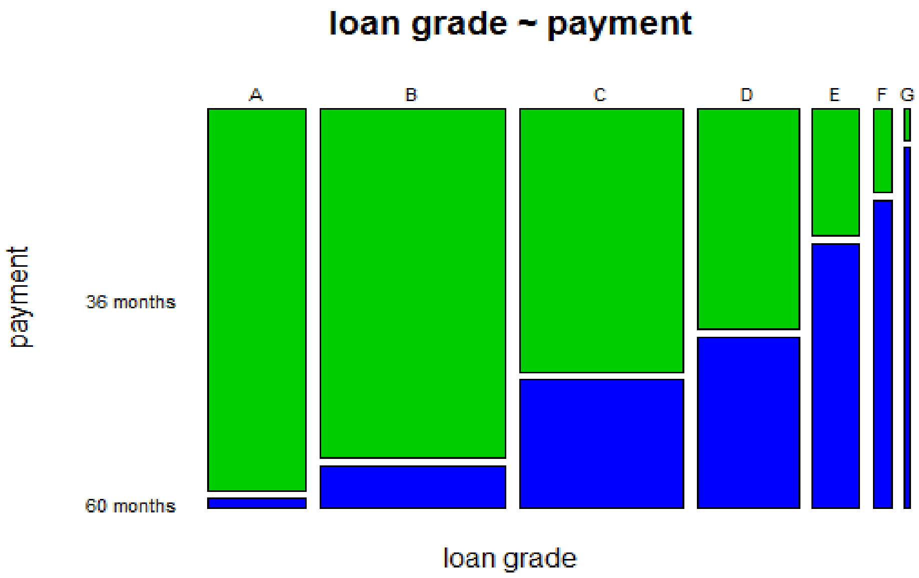
Understanding of data with help **of graph**



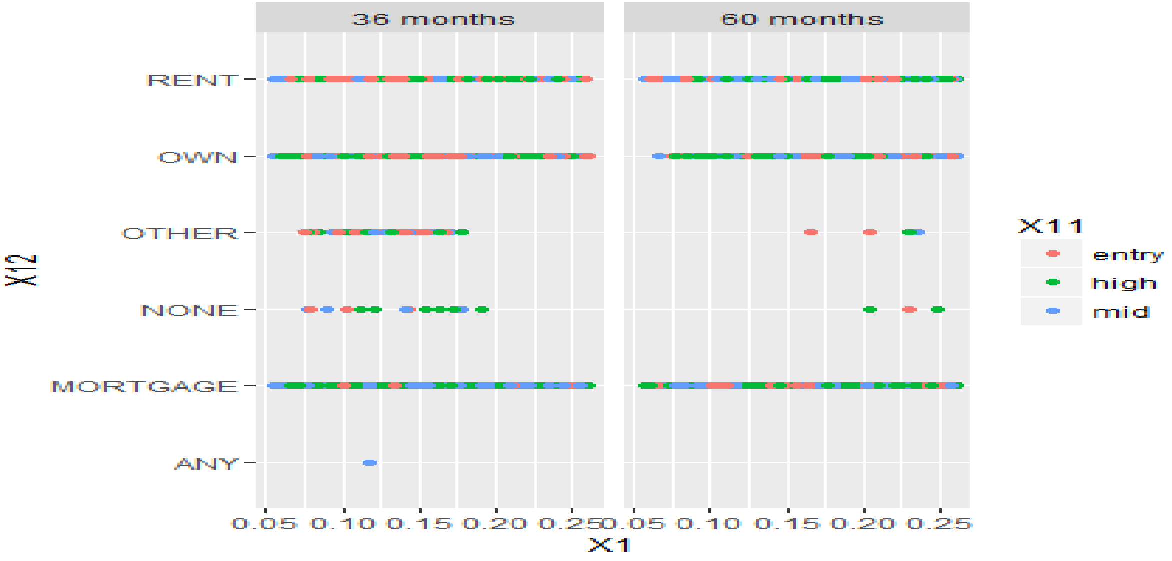
Interest Rate varies from 5 to 26 percent with mean value 15 percent . This depicted data is dependent variable ,now I make my effort to relate some independent variable with it .



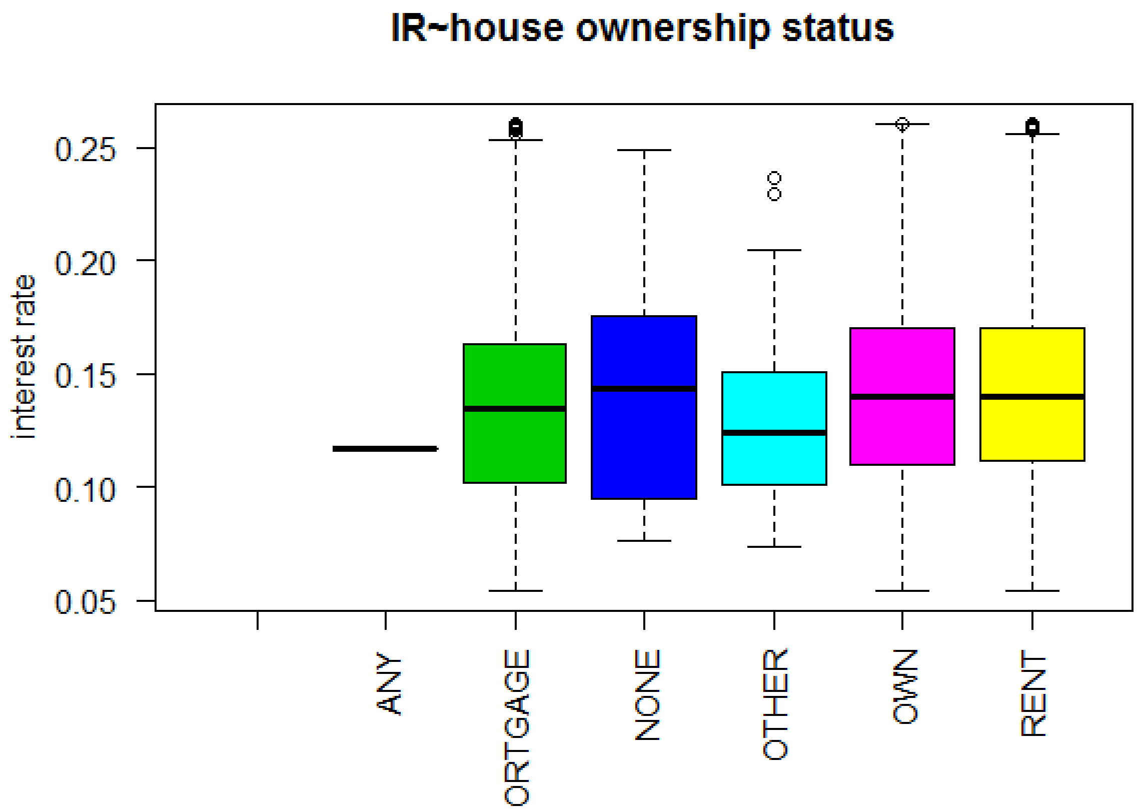
Here, interest rate on the loan depends on the choice of loan grade and choice of payment .payment with 60 months has more interest rate followed by 36 months payments in most of the cases . Also, borrower has only two choice and that choice of payment in 36 months decreases as we moves from loan grade A to G. And loan grade G has highest rate of interest while A has the least. Below graph depicts the broader view on distribution of loan grade wrt payment type.



Now , I am interested to know the hidden information from level of experience.

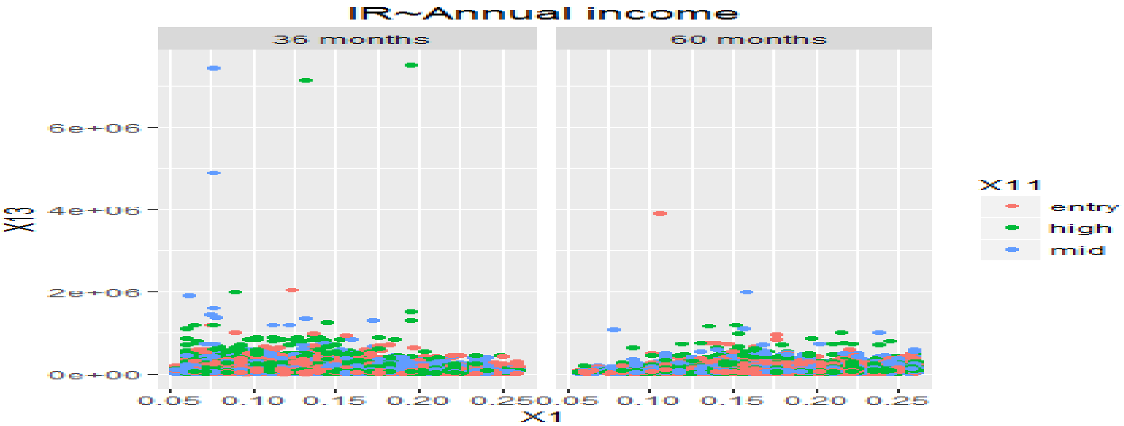


This graph clearly depicts that house status Rent, own and mortgage are highly distributed over IR and ANY has only one type of loan payment with interest rate 14% while other and none are more distributed on 36 months payments compare to 60 months payment. The level of experience does not affect interest rate on the loan .



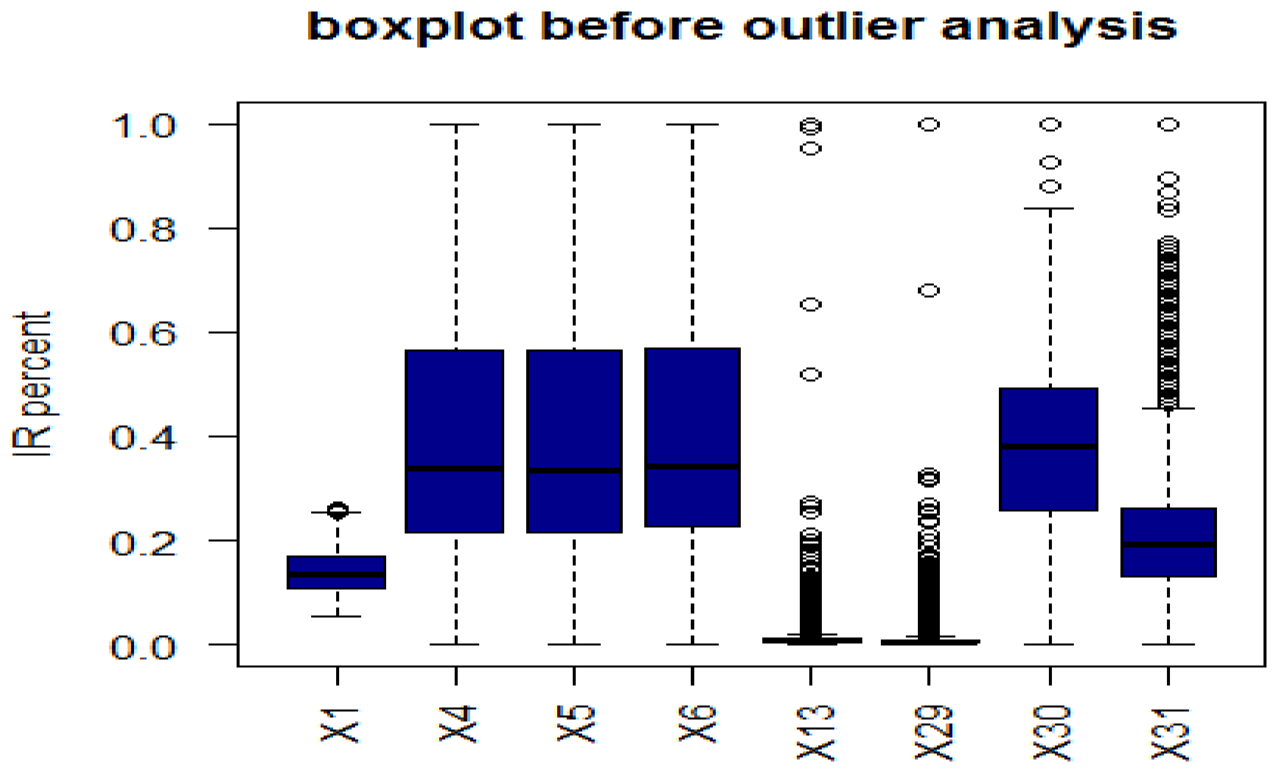
Borrower with None as house status pay the highest median rate of interest while , with status OWN and RENT pay the same median rate of interest rate.

Now lets take a look at distribution of IR wrt annual income



This graph entails about random distribution of IR over annual income and experience. Random distribution means no relation among interest rate and annual income of borrower.

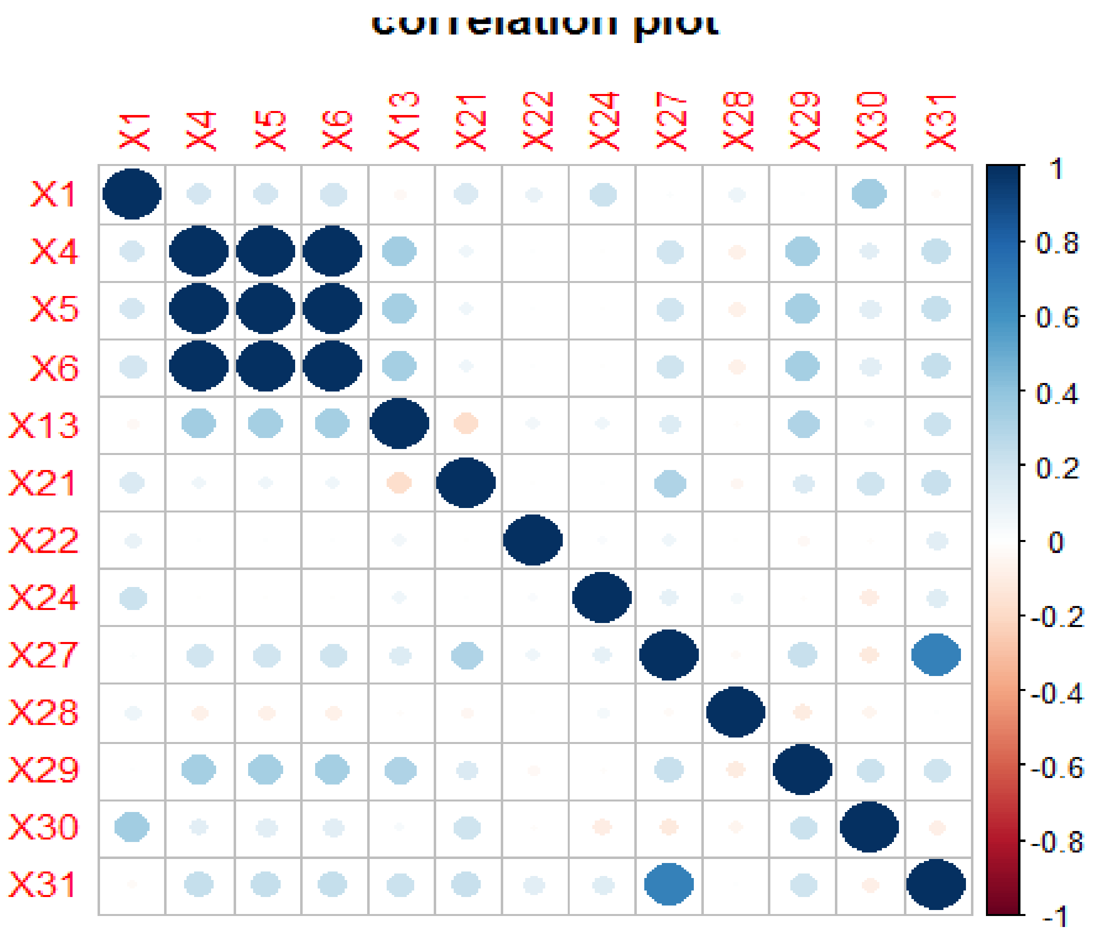
OUTLIER ANALYSIS



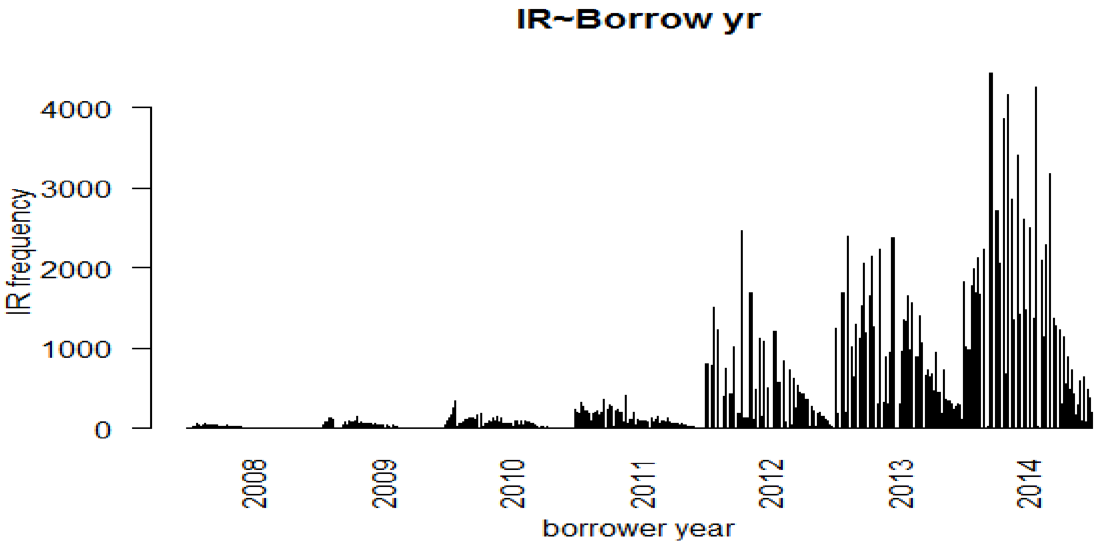
Here , from box plot it is quite clear that X4,X5 AND X6 are highly correlated but some variable like X13,X29 ,X31 have some outliers .so I have to remove outliers in X13,X29 and X31.

Let's check correlation among variables

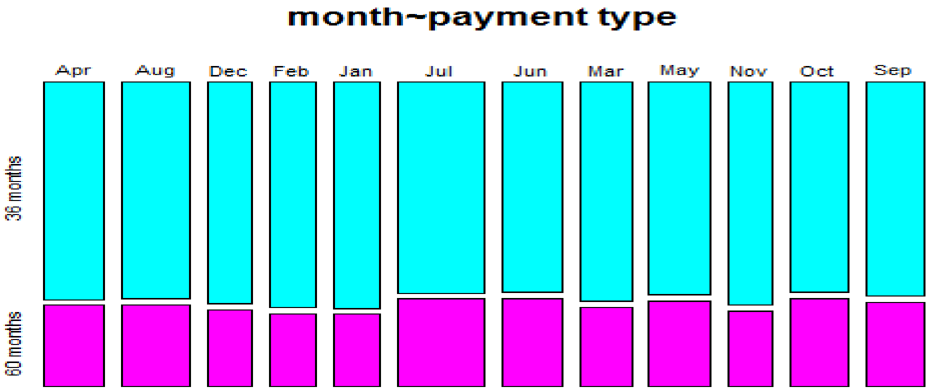
This correlation plot gives a clear view that X4, X5 , X6 are highly co-related so I have dropped two of them except X4 as all affects the rate of Interest rate on same way.



<u>Now let’s see the trend of loan IR   wrt year & month</u>

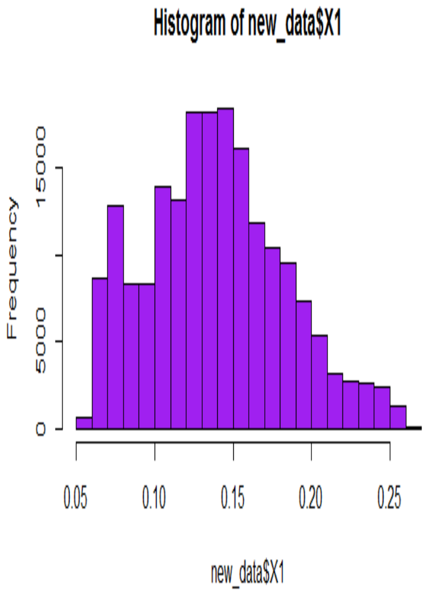


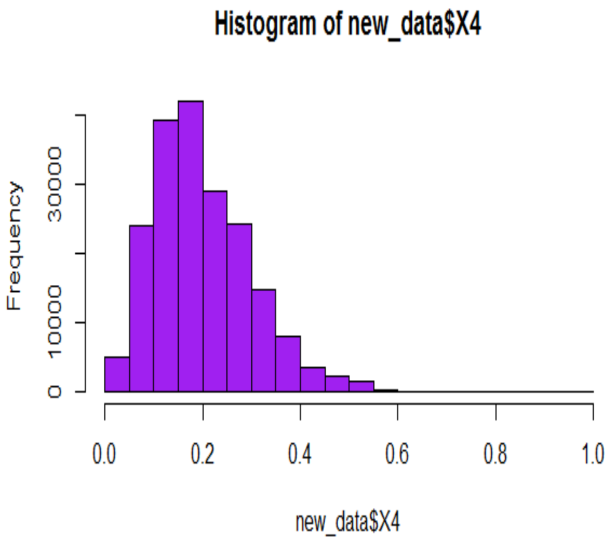
People choice towards loan increases year wise and also in 2008 no. of borrower is very less but there is drastic change in year 2012 towards loan .

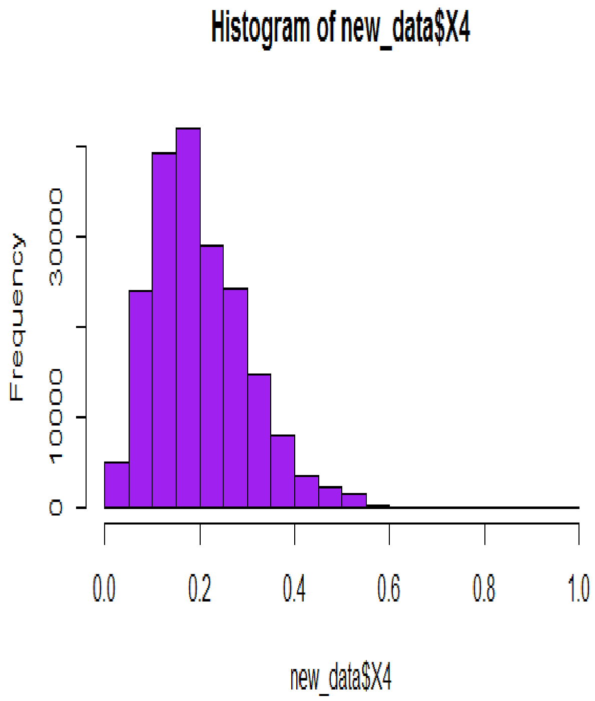


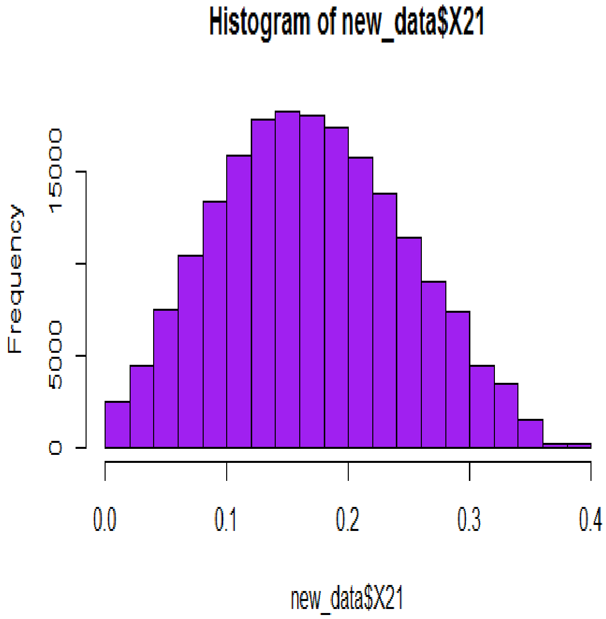
Borrower's choice of payment does not affected by month. Also choice of payment is almost constant throughout the year irrespective of month.

<u>Sampling technique method:</u>







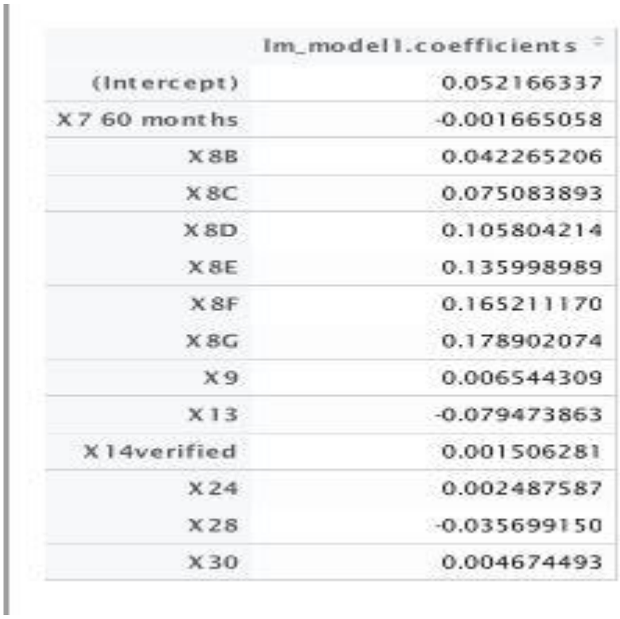


As approximately these graphs follow binomial distribution curve.so I have taken mean value while sampling ,train and test data.

I have taken 5100 sample data out of 193262 to keep mean of sample data equal to 0.1392(approximately equal to whole data) .Again, using trial and error method ,I have selected 67% of sample data as train and remaining as test with mean value 0.1397 & 0.1401 and p value = 2.2 e-16.

4. **Multiple regression model:**

I have used multiple linear regression model and sort out that X7 ,X8,X9,X13,X14,X24,X28,X30,borrow_year variable are more important factor to predict interest rate .As I have taken care of Pr value and relation with graph to sort out best suited variable . I have taken P r >0.05 is suitable for null hypothesis. I have got 96.2% accuracy but Despite being borrow _year being important vector , I have to drop this variable as it is not suitable to predict Interest rate of holding data. And finally my accuracy goes down to **95.5%.**



	lm_model1.coefficients
(Intercept)	0.052166337
X7 60 months	-0.001665058
X8B	0.042265206
X8C	0.075083893
X8D	0.105804214
X8E	0.135998989
X8F	0.165211170
X8G	0.178902074
X9	0.006544309
X13	-0.079473863
X14verified	0.001506281
X24	0.002487587
X28	-0.035699150
X30	0.004674493

This table shows the coefficient and intercept of different variable .

Variable X8 plays a crucial role to predict IR on loan data. Also these variable are positively related with dependent variable except X7_60months, X13 and X28.

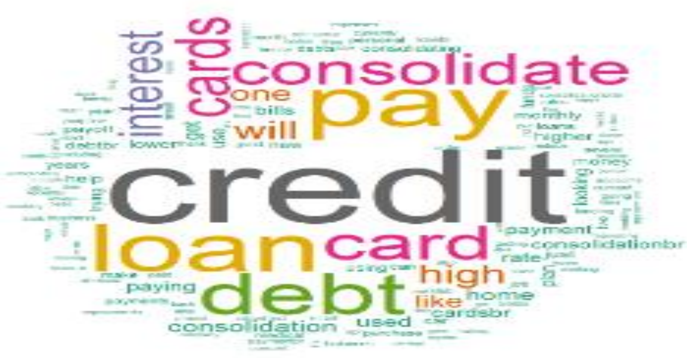
Multiple regression formula :

Loan1$X1 = 0.052 - 0.00166*X7_60Months +0.042*X8_B +0.075* X8_C +0.105* X8_D +0.135* X8_E +0.165* X8_F +0.178* X8_G+0.006*X9 -0.0794*X13+0.0015* X14_Verified +0.002*X24-0.035*X28+0.0046*X30

5. Reason behind loan

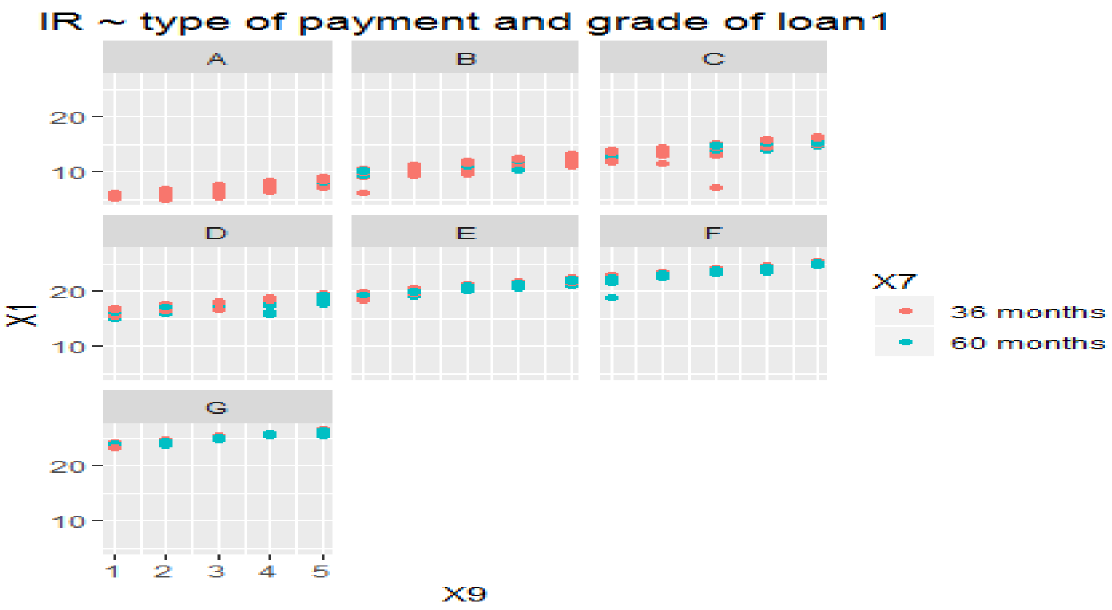
   I have mined the text data to find out reason behind loan. first I have chosen the word between 10 to 100 as most of the explained reason fall in between them .Then building word cloud with clean text clearly show that the main reason to take loan is to consolidate credit card debt and to pay other

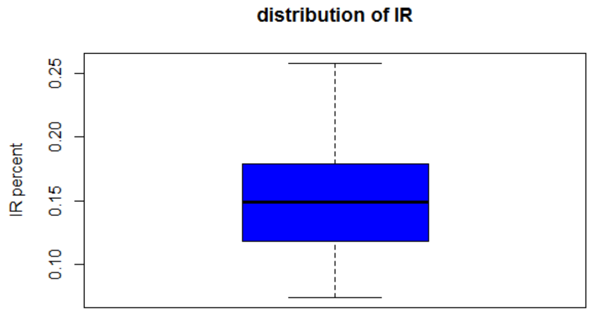
Loan.also it is clear that very less people take loan for education .people alredy in debt prefer to take loan .

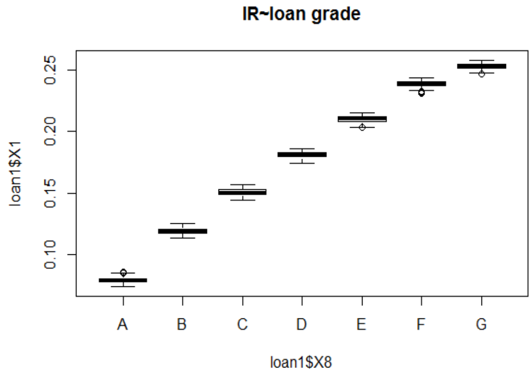


RESULT

this regression model gives IR on loan with min. 4.88% ,max. =26.67% , mean =13.92%.







This graph clearly shows that interest rate increases with loan grade as A has low value of Interest rate while G has the highest. also low subgrade has highest rate of interest and choice of payment changes as we move from Grade A-G.

conclusion :

Number of payment , loan grade and subgrade ,annual income,income verification ,number of enquiry,no. of deregotary public records and revolving line utilization rate are factor to determine interest rate on loan data. All the factor shows positive effect on loan interest except number of payment,annual income of borrower, no. of derogatory public records.

Number of payment:- 36 months of payment option has less interest rate as compare to 60 months payment option.

Loan grade:- interest rate increases as loan grade shifts from A - G. A has the least while G loan grade has the highest interest rate.

Level of experience does not affect IR on loan data.

And more person are getting habituated to take loan with time and has been observed a drastic change in year 2012.Also people already in debt prefer to take loan either to consolidate credit card debt or to pay other loan with high interest.