Gramener Case Study

SUBMISSION

Group Name:

1. Anshika Misra
2. Madhulika Joshi
3. Sharanya Vijay
4. Piyu Chatterjee

Business Objectives

The company for which data has been provided is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

The objective of this case study is to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

The company can utilize this knowledge for its **portfolio and risk assessment.**

Steps involved

1. Data Import
2. Data Cleaning
   - Convert all data to lower case
   - Find the number of unique members for each column
   - Find and remove the column with only NA values
   - Find and remove the columns that have only a single value in the entire column
   - Find and remove the columns that have only 0's or NA's
   - Remove the columns url, desc, emp_title, title as no value add to the current analysis
   - Remove string months from loan term column
   - Remove % from int_rate and revol_util column
   - Strip of last 2 xx - zip_code
3. Data Conversion
   - Convert to factors - term, grade,sub_grade,emp_length, home_ownership, verification_status, loan_status, purpose, addr_state
   - Convert to numeric - int_rate, revol_util
   - Convert to date - issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d, last_credit_pull_d
4. Add Derived Columns
   - Extract Year and Month from issue_d, earliest_cr_line, last_pymnt_d, next_pymnt_d, last_credit_pull_d and add it to the data frame
   - Categorize and add columns for purpose, inq_last_6mths, dti into categorical data
   - Quantify values of sub_grade and loan_status

Steps involved – contd.

5. Perform Univariate analysis on Continuous and Discrete Variables
   - Characteristics of Continuous Variables
   - Plots for Discrete Variables:
     - Term, Grade, Sub Grade, Employment Length, Home Ownership, Verification Status, Loan Status, Purpose
6. Perform Bivariate analysis
7. Derive Correlations
8. Draw Conclusions

Variables used

Borrower Assessment	
int_rate	Interest Rate on the loan
grade	LC assigned loan grade. There are 7 grades from A-G with A being the safest and G being riskiest
sub_grade	LC assigned loan subgrade. Each grade is further divided into 7 sub-grades each From A1 to A7, B1 to B7… G1 to G7. A1 is the safest and G7 is the riskiest
Loan Characteristics	
purpose	A category provided by the borrower for the loan request. List of purpose car, credit_card, debt_consolidation, educational, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation, wedding
loan_amnt	The listed amount of the loan applied for by the borrower.
Borrower Characteristics	
emp_length	Employment length in years. 12 values from <1 year, 1, 2, … 9, 10, >10 years
home_ownership	The home ownership status provided by the borrower during registration. The values are: mortgage, none, other, own, rent
annual_inc	The self-reported annual income provided by the borrower during registration.
Credit History	
earliest_cr_line	The month the borrower's earliest reported credit line was opened
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
mths_since_last_record	The number of months since the last public record.
revol_bal	Total credit revolving balance
total_acc	The total number of credit lines currently in the borrower's credit file
mths_since_last_delinq	The number of months since the borrower's last delinquency.
pub_rec_bankruptcies	Number of public record bankruptcies
Borrower Indebtedness	
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

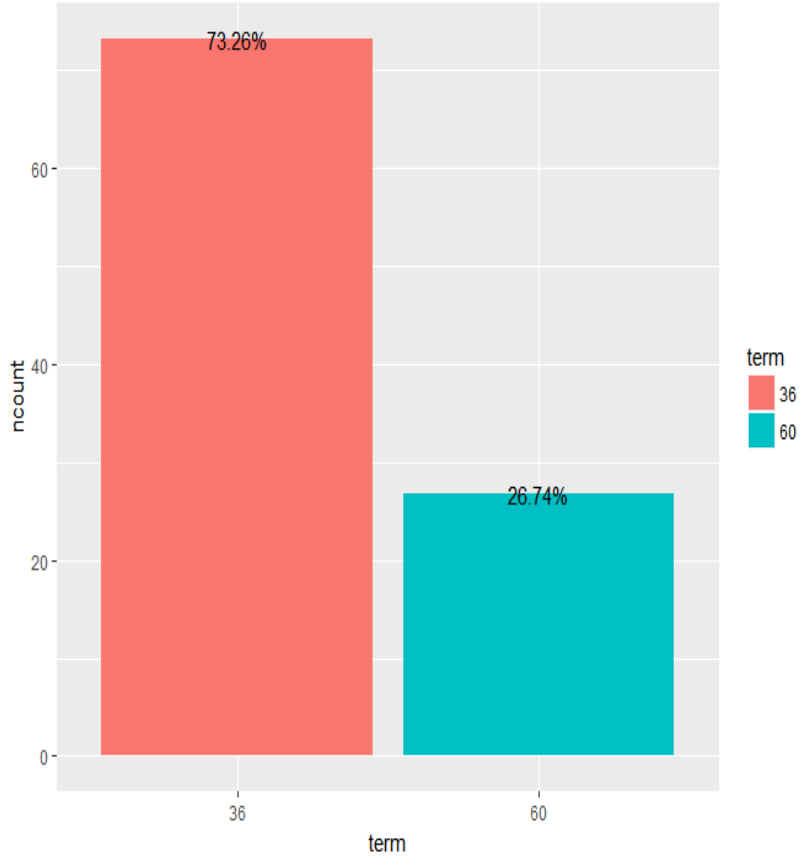
Other Varaibles	
id	A unique LC assigned ID for the loan listing.
member_id	A unique LC assigned Id for the borrower member.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
installment	The monthly payment owed by the borrower if the loan originates.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
loan_status	Current status of the loan
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
addr_state	The state provided by the borrower in the loan application
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_prncp	Principal received to date
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
recoveries	post charge off gross recovery
collection_recovery_fee	post charge off collection fee
last_pymnt_d	Last month payment was received
last_pymnt_amnt	Last total payment amount received
next_pymnt_d	Next scheduled payment date
last_credit_pull_d	The most recent month LC pulled credit for this loan

Continuous Variables – Characteristics

Continuous Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
loan_amnt	500	5500	10000	11220	15000	35000	
funded_amnt	500	5400	9600	10950	15000	35000	
funded_amnt_inv	0	5000	8975	10400	14400	35000	
int_rate	5.42	9.25	11.86	12.02	14.59	24.59	
installment	15.69	167	280.2	324.6	430.8	1305	
annual_inc	4000	40400	59000	68970	82300	6000000	
issue_d	6/1/2007	5/1/2010	2/1/2011	11/3/2010	8/1/2011	12/1/2011	
dti	0	8.17	13.4	13.32	18.6	29.99	
delinq_2yrs	0	0	0	0.1465	0	11	
earliest_cr_line	2/1/1969	12/1/1993	5/1/1998	4/11/1997	9/1/2001	12/1/2068	
inq_last_6mths	0	0	1	0.8692	1	8	
mths_since_last_delinq	0	18	34	35.9	52	120	25682
mths_since_last_record	0	22	90	69.7	104	129	36931
open_acc	2	6	9	9.294	12	44	
pub_rec	0	0	0	0.05506	0	4	
revol_bal	0	3703	8850	13380	17060	149600	
revol_util	0	25.4	49.3	48.83	72.4	99.9	50
total_acc	2	13	20	22.09	29	90	
out_prncp	0	0	0	51.23	0	6311	
out_prncp_inv	0	0	0	50.99	0	6307	
total_pymnt	0	5577	9900	12150	16530	58560	
total_pymnt_inv	0	5112	9287	11570	15800	58560	
total_rec_prncp	0	4600	8000	9793	13650	35000	
total_rec_int	0	662.2	1349	2264	2833	23560	
total_rec_late_fee	0	0	0	1.363	0	180.2	
recoveries	0	0	0	95.22	0	29620	
collection_recovery_fee	0	0	0	12.41	0	7002	
last_pymnt_d	1/1/2008	4/1/2012	4/1/2013	4/10/2013	6/1/2014	5/1/2016	71
last_pymnt_amnt	0	218.7	546.1	2679	3293	36120	
next_pymnt_d	6/1/2016	6/1/2016	6/1/2016	6/1/2016	6/1/2016	7/1/2016	38577
last_credit_pull_d	5/1/2007	6/1/2013	3/1/2015	9/7/2014	5/1/2016	5/1/2016	2
pub_rec_bankruptcies	0	0	0	0.0433	0	2	697

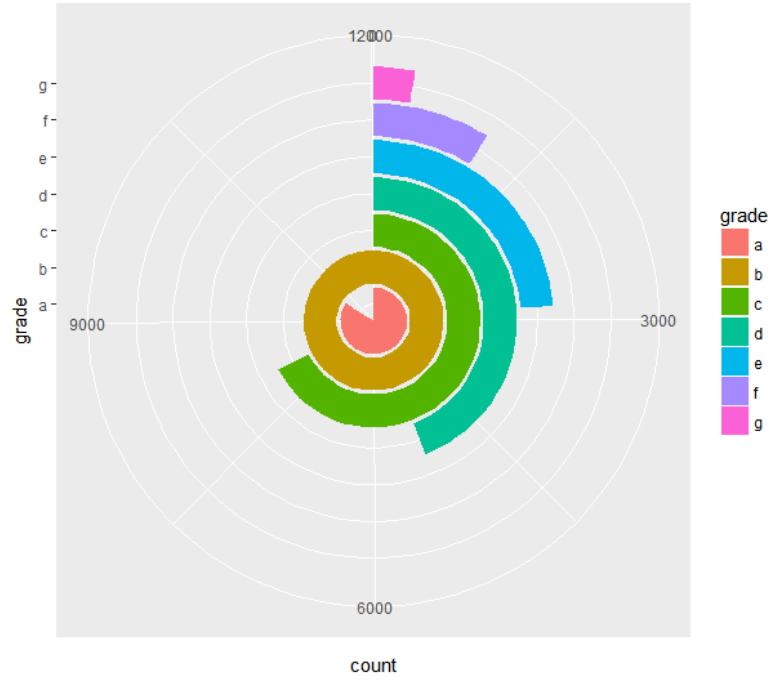
Univariate Analysis – Term and Grade

Term	Count	Percentage
36	29096	0.733
60	10621	0.267



Around ¾ of all loans are for 3 years and only ¼ are for 5 year loan period

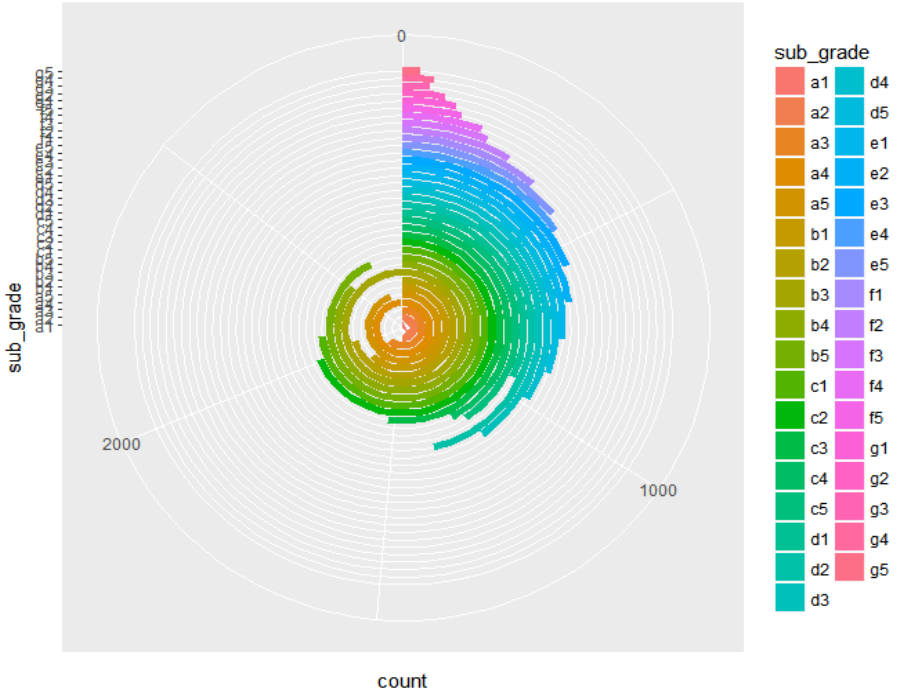
Grade	Count	Percentage
a	10085	25.39
b	12020	30.26
c	8098	20.39
d	5307	13.36
e	2842	7.16
f	1049	2.64
g	316	0.80



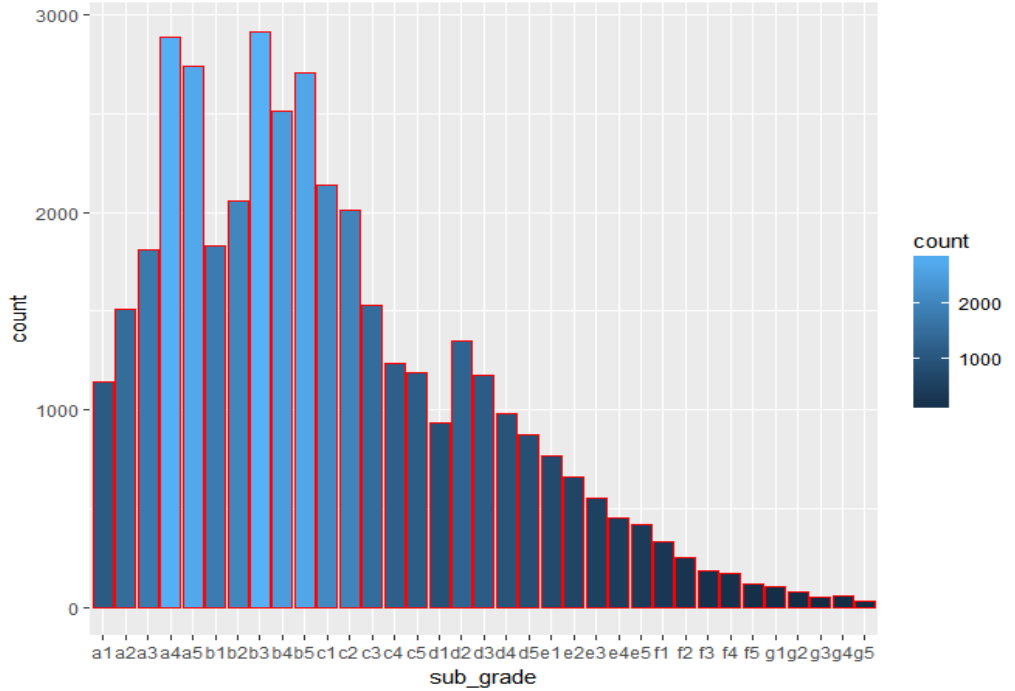
Of all loans ¾ are in Grades A, B and C and ¼ are in Grades D, E, F and G

Univariate Analysis – Sub Grade

A	Count	B	Count	C	Count	D	Count	E	Count	F	Count	G	Count
a1	1139	b1	1830	c1	2136	d1	931	e1	763	f1	329	g1	104
a2	1508	b2	2057	c2	2011	d2	1348	e2	656	f2	249	g2	78
a3	1810	b3	2917	c3	1529	d3	1173	e3	553	f3	185	g3	48
a4	2886	b4	2512	c4	1236	d4	981	e4	454	f4	168	g4	56
a5	2742	b5	2704	c5	1186	d5	874	e5	416	f5	118	g5	30



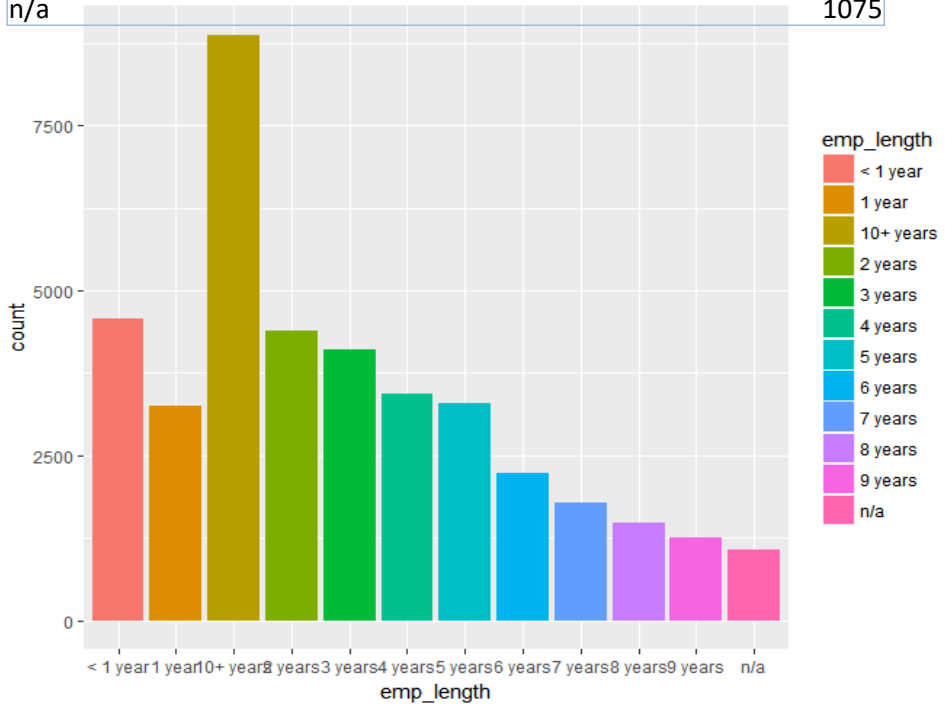
The 35 sub grades a1 through g5 are evaluated as least risky to most risky



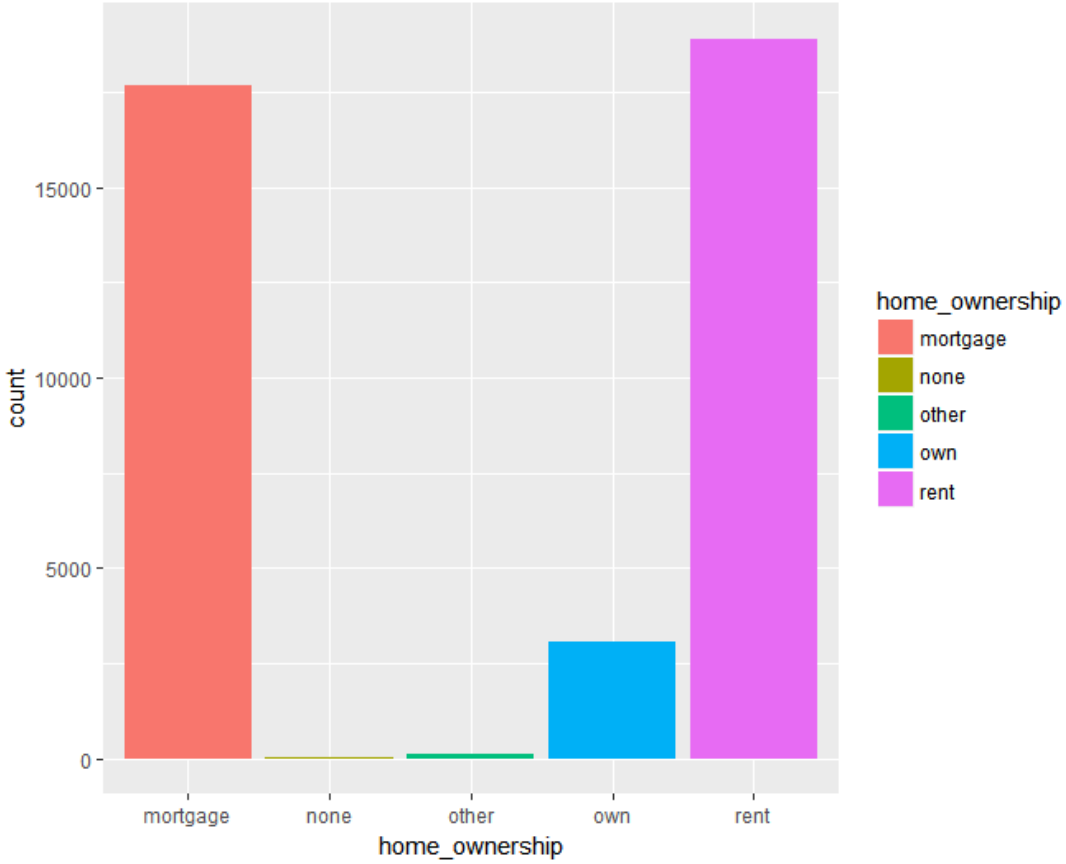
The maximum amount of loans are in B grade, followed by A and C grades.

Univariate Analysis – Emp Length & Home Ownership

emp_length	Count
<1 year	4583
1 year	3240
10+ year	8879
2 year	4388
3 year	4095
4 year	3436
5 year	3282
6 year	2229
7 year	1773
8 year	1479
9 year	1258
n/a	1075

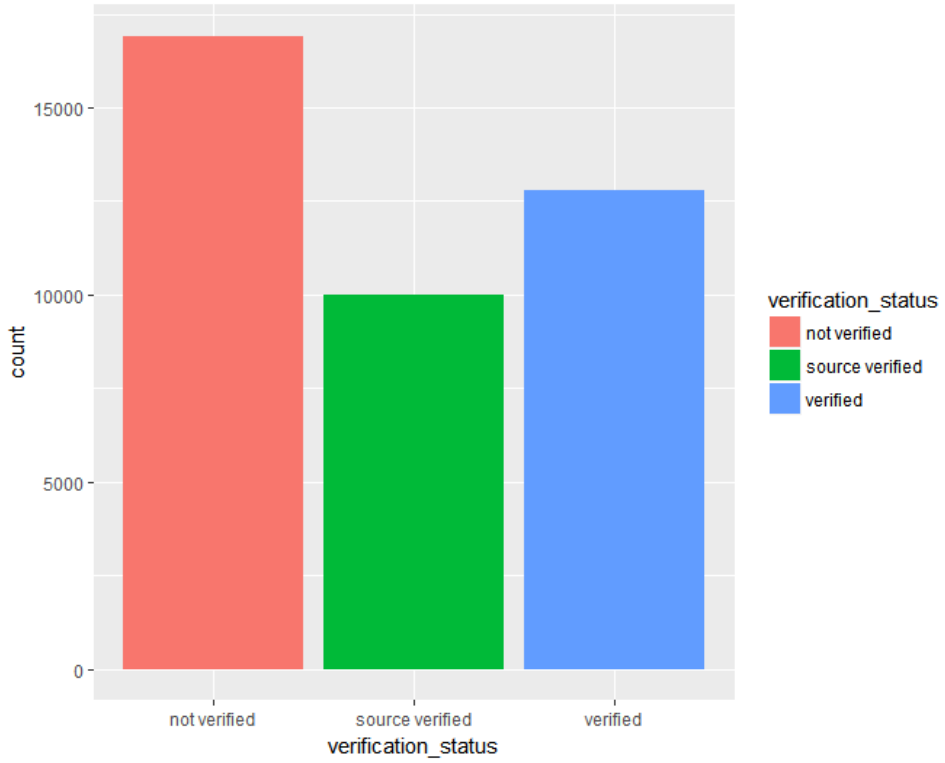


home_ownership	Count
mortgage	17659
none	3
other	98
own	3058
rent	18899



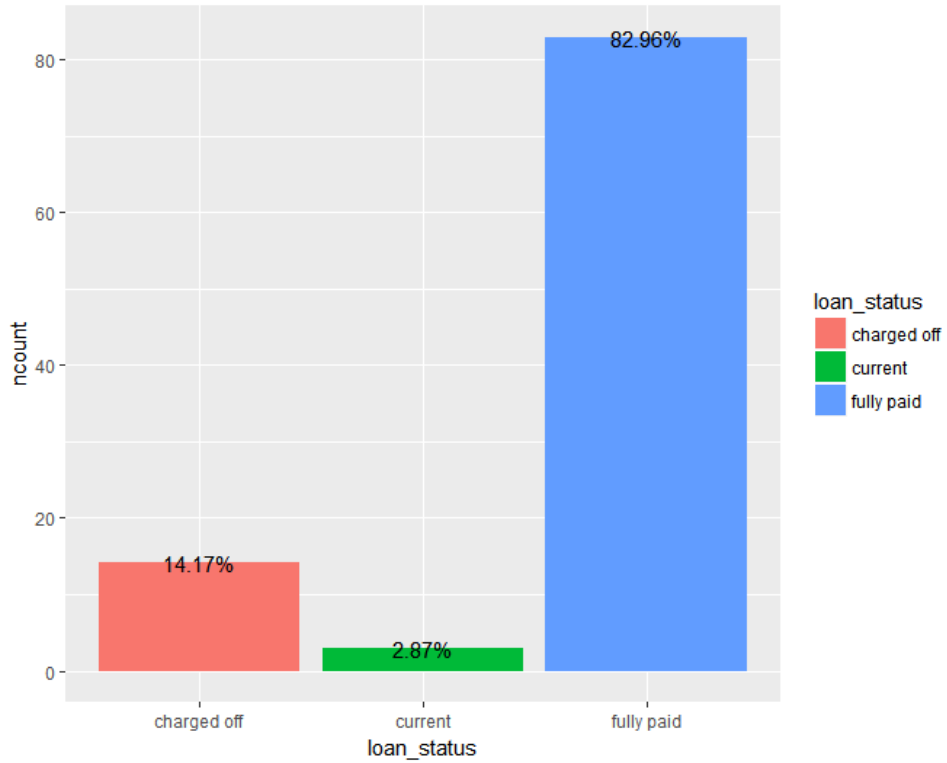
Univariate Analysis – Verification and Loan Status

verification_status	Count
not verified	16921
source verified	9987
verified	12809



43% loans are not verified and 57 & are verified

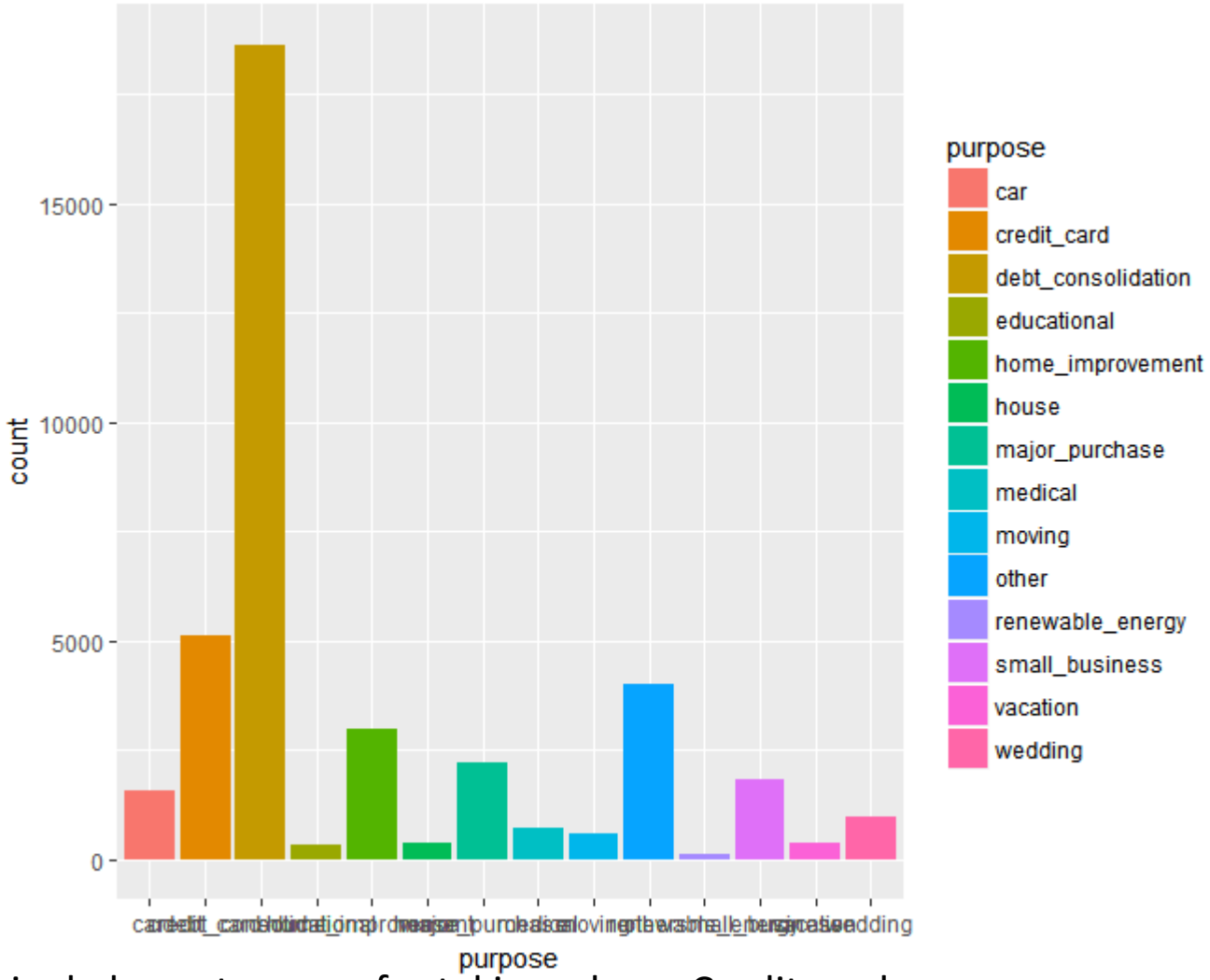
loan_status	Count	Percentage
charged off	5627	0.14167737
current	1140	0.02870307
fully paid	32950	0.82961956



14% loans are charged off whereas 83% are fully paid

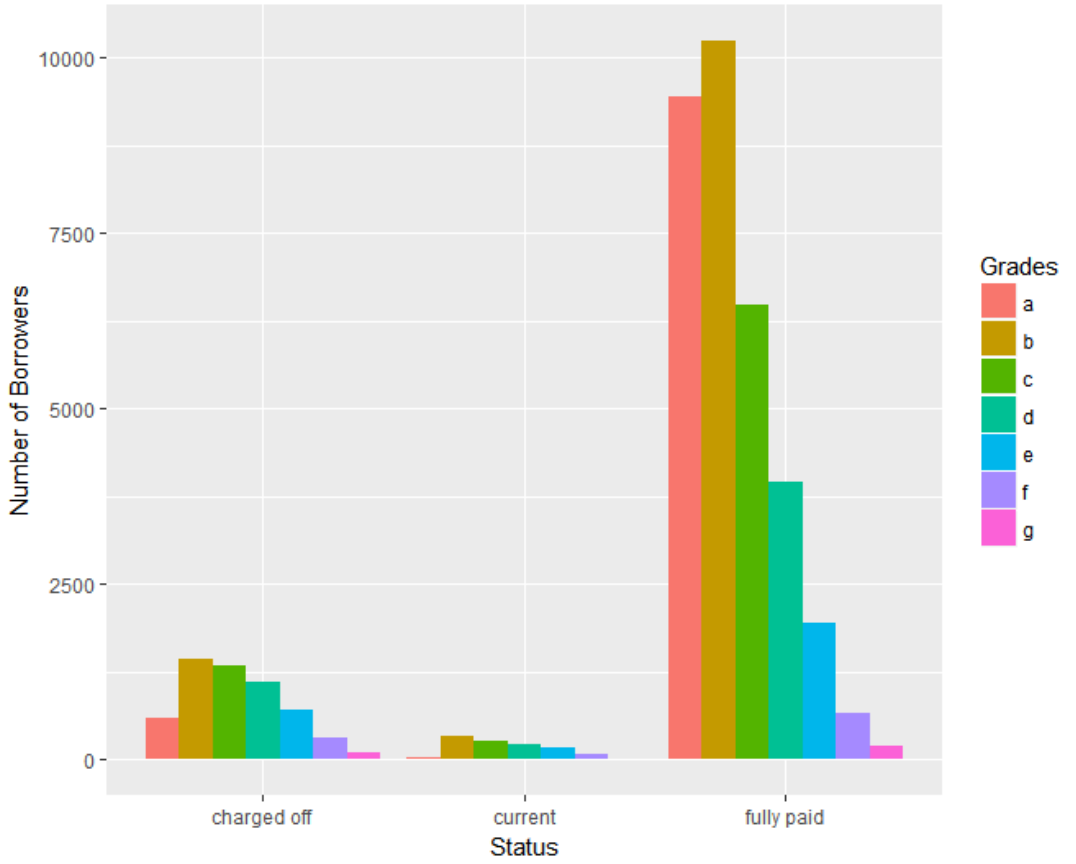
Univariate Analysis – Purpose

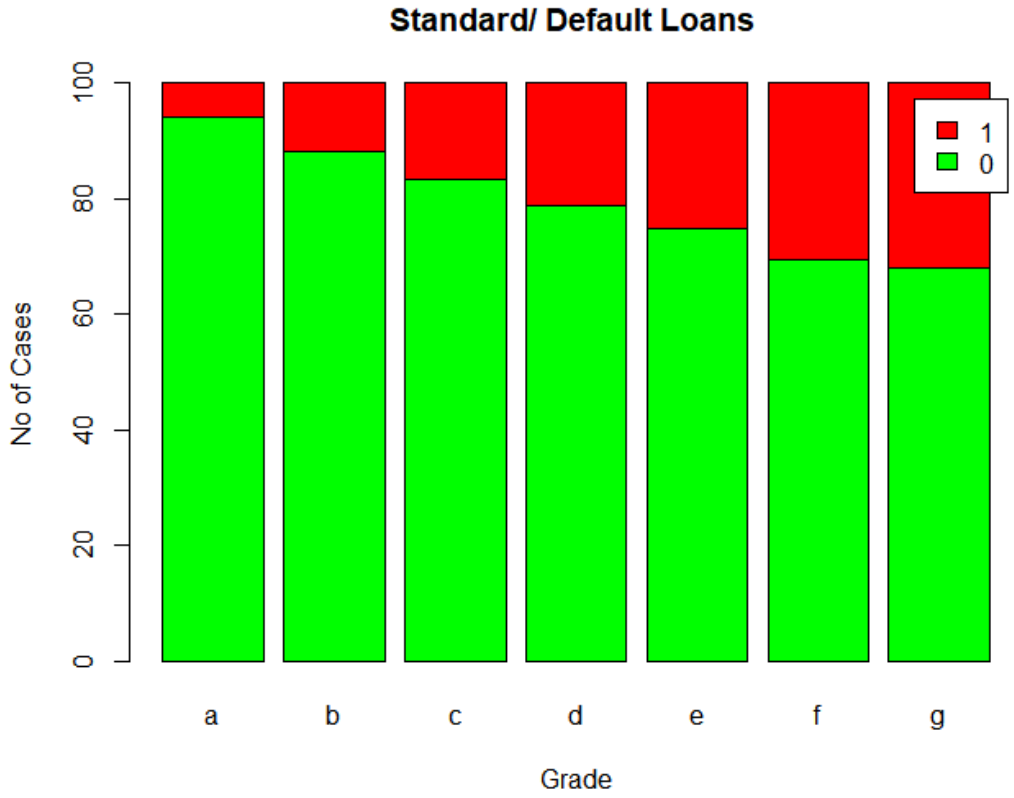
purpose	Count
car	1549
credit_card	5130
debt_consolidation	18641
educational	325
home_improvement	2976
house	381
major_purchase	2187
medical	693
moving	583
other	3993
renewable_energy	103
small_business	1828
vacation	381
wedding	947



Debt Consolidation (47% is the single largest reason for taking a loan. Credit cards are second ranking at 13%

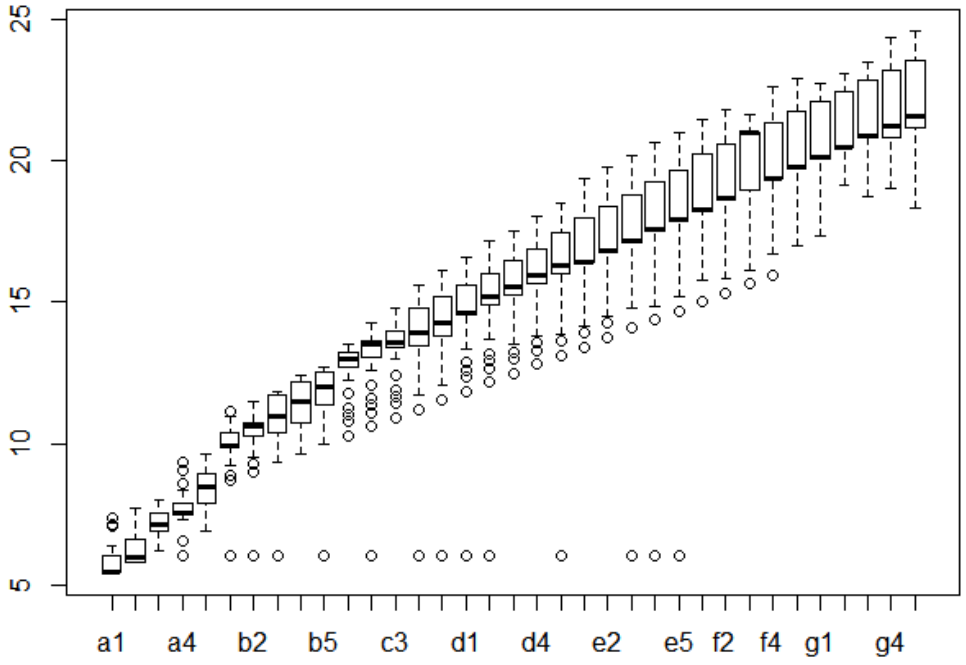
Bivariate Analysis – Loan Status Vs Grades

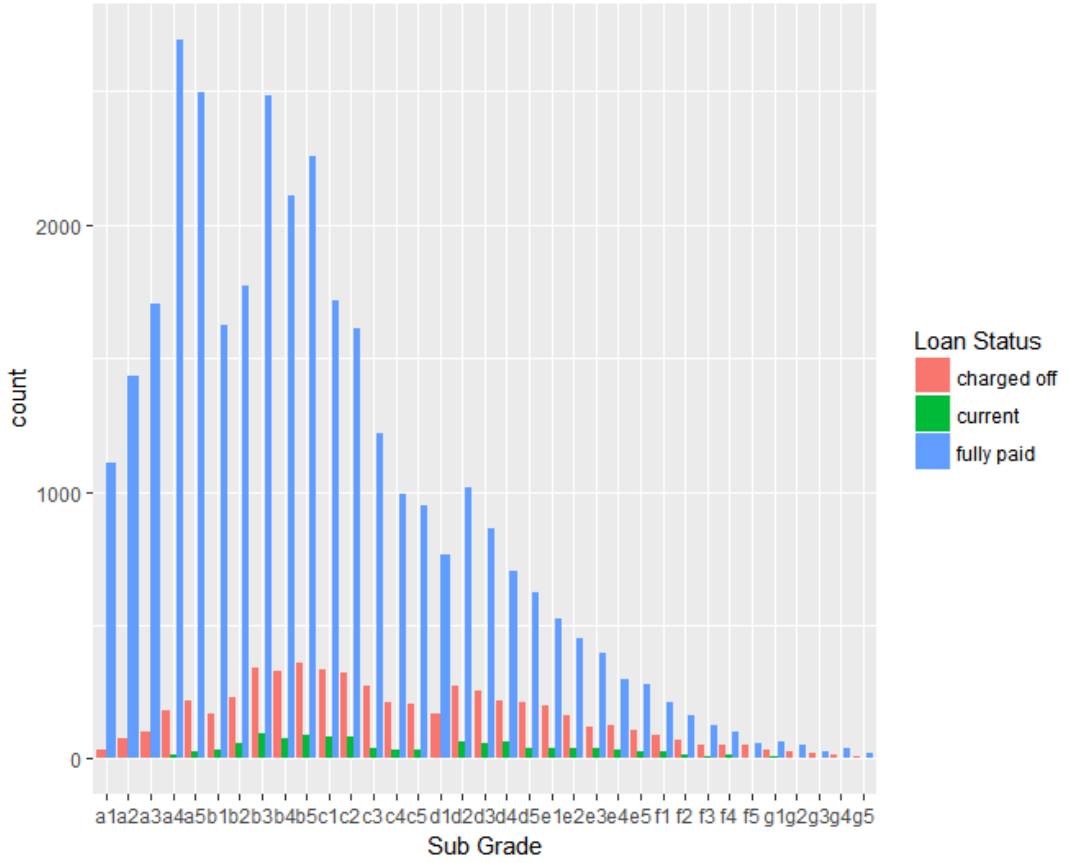




The maximum number of defaults are for B, C and D grades, but percentage wise the defaults increase from A (5%) to G(25%). So we can conclude that the risk increases from A to G.

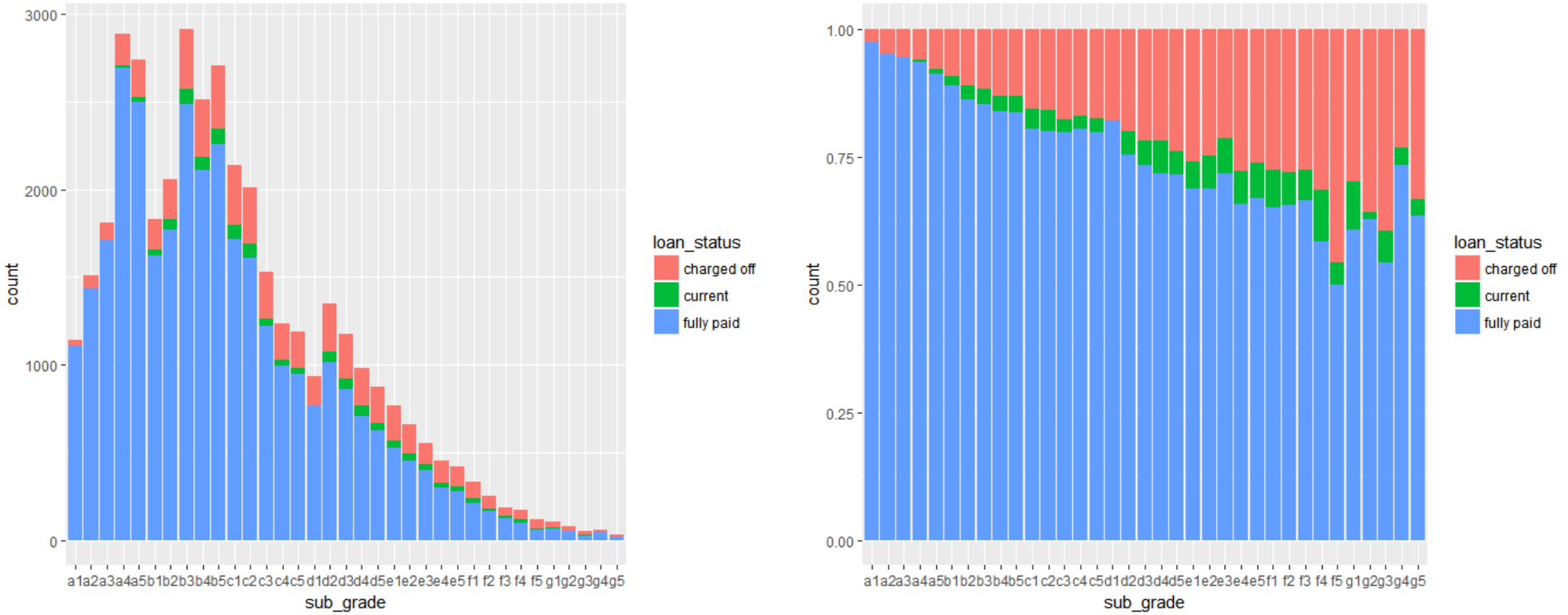
Bivariate Analysis – Loan Status Vs Sub Grades





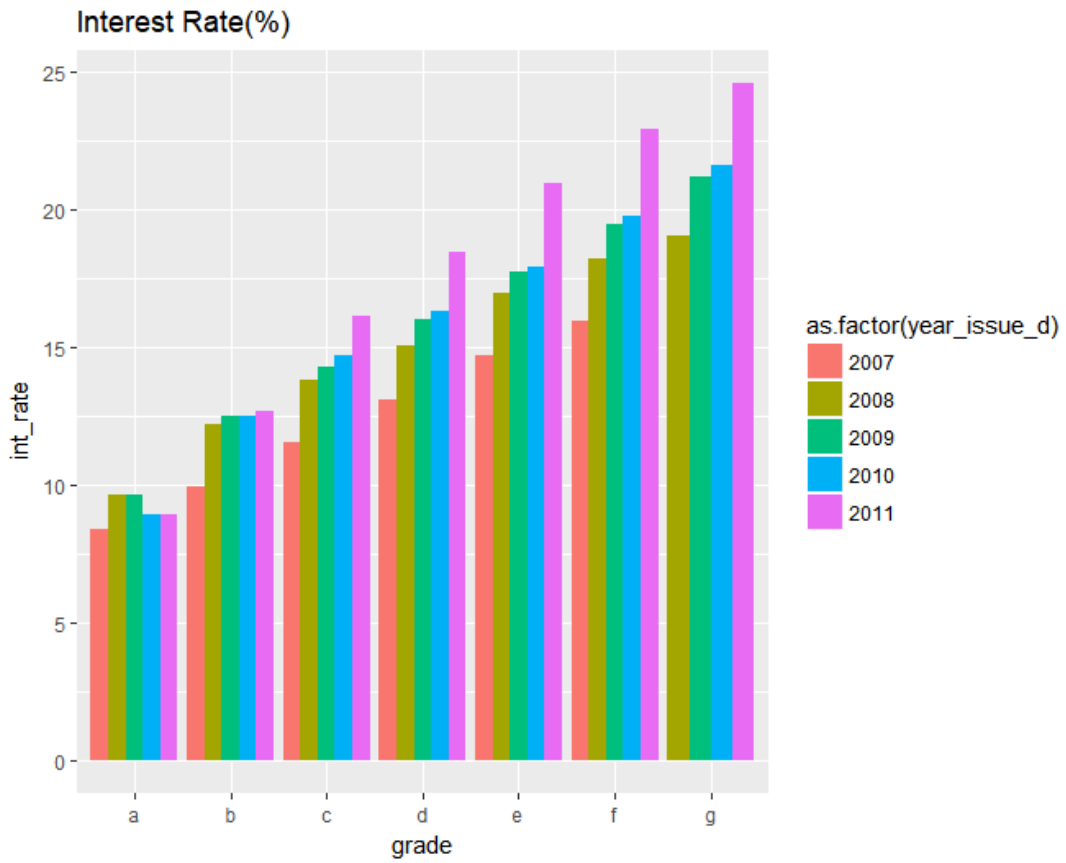
Interest Rate has a direct Correlation with Sub Grades. There is a steady increase in interest rates from A1 through G5.

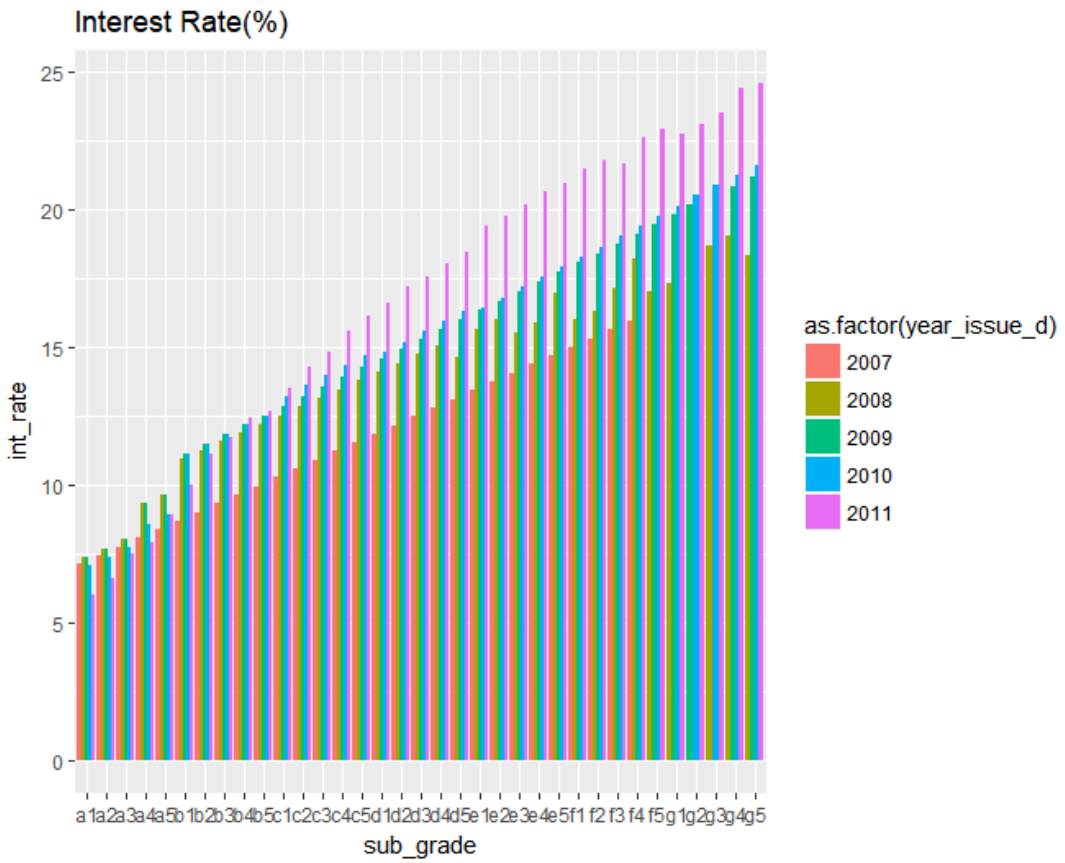
Bivariate Analysis – Loan Status Vs Sub Grades



The absolute number of defaults are high for subgrades in B and C grades. But percentage wise the defaults show a steady increase from A1 to G5

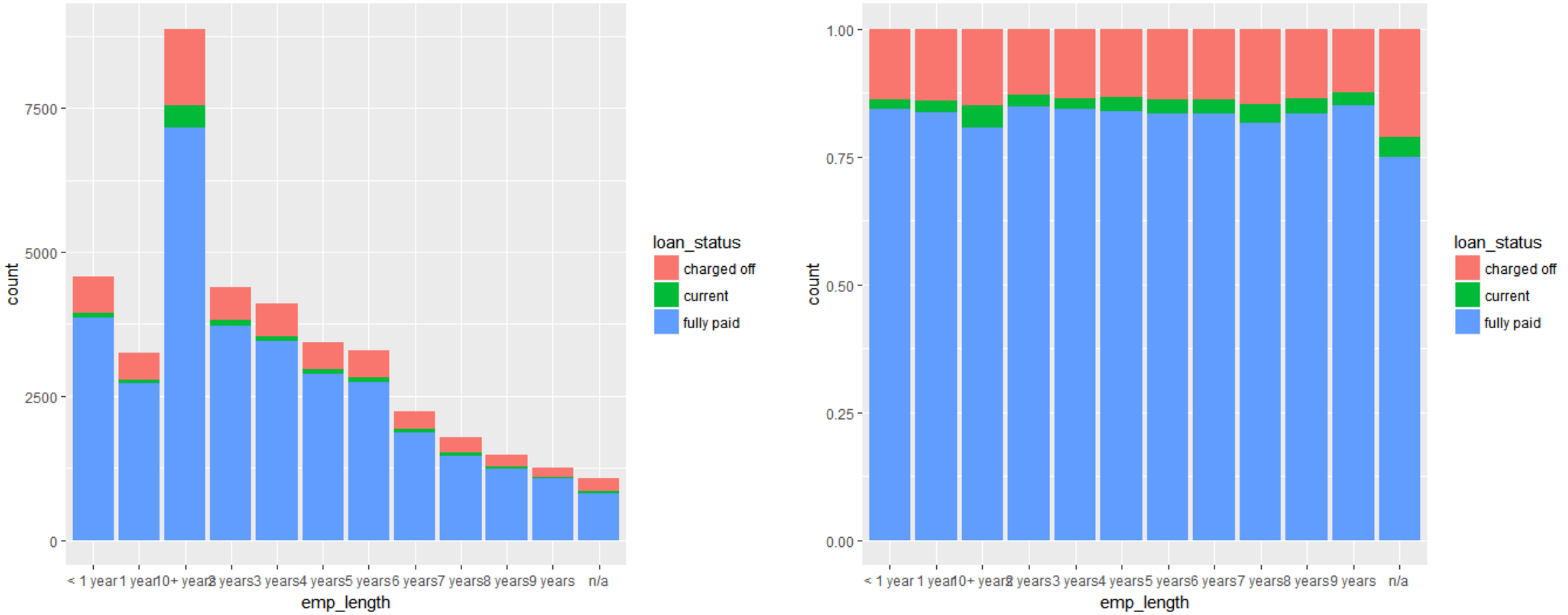
Bivariate Analysis – Grades & Sub Grades Vs Interest Rate across Years





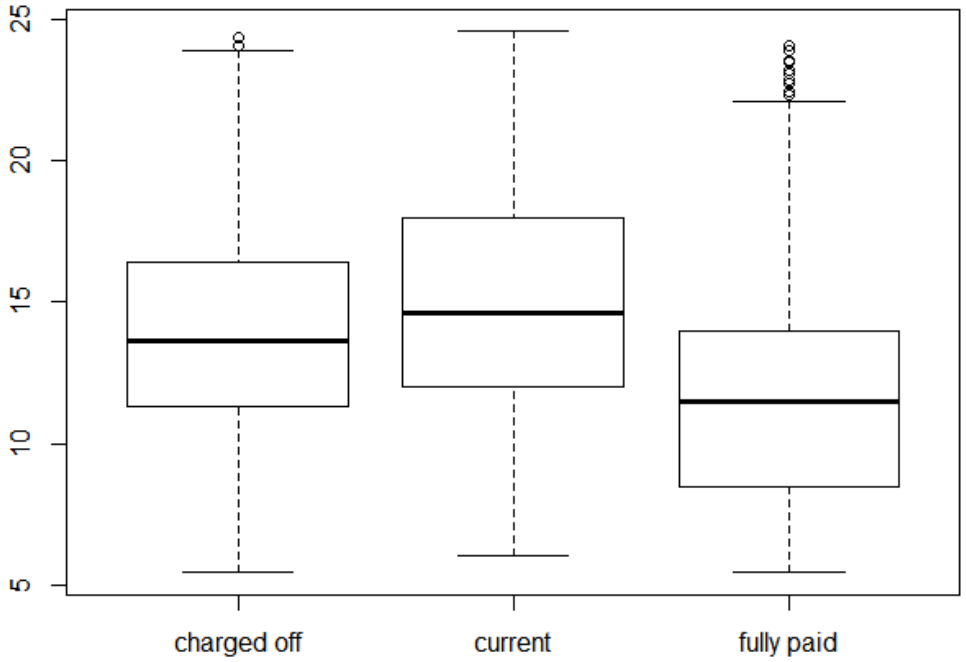
The interest rates across Grades and sub grades show an increasing trend across years with the highest rise seen in 2011 especially across D, E, F and G grades and their respective sub grades

Bivariate Analysis – Loan Status Vs Emp Length



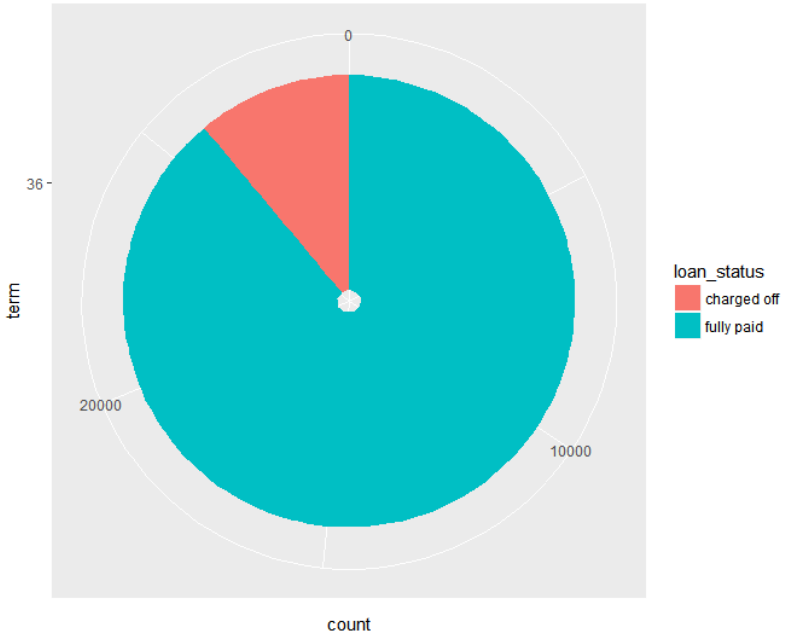
The absolute number of defaults are high for people with 10+ years of experience. But percentage wise the defaults are maximum for N/A experience. If the emp length is the duration with the current employer, then those in N/A category may be unemployed and hence the high rate of default.

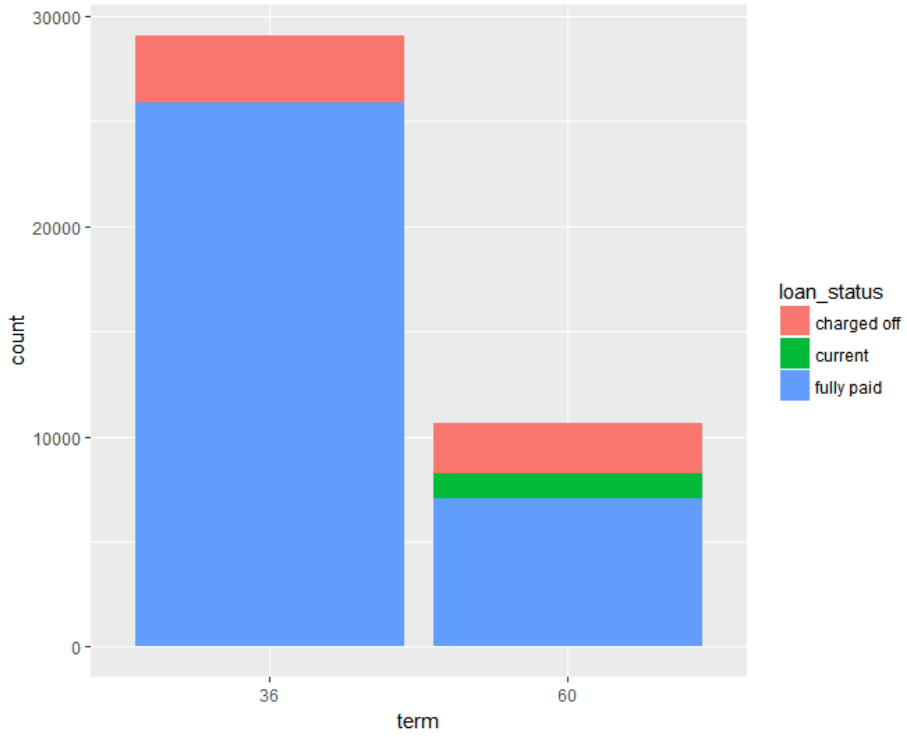
Bivariate Analysis – Loan Status Vs Interest Rate

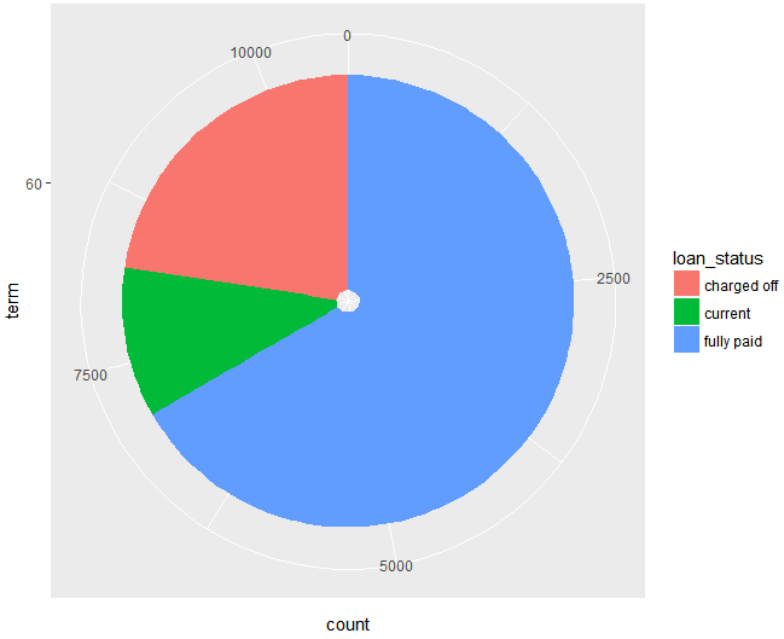


The interest rate is significantly higher for the defaulters compared to the loans that were fully paid. The loans that are currently running have the highest interest rate and that is because all these loans started in 2011 when the interest rates were significantly higher

Bivariate Analysis – Loan Status Vs Term

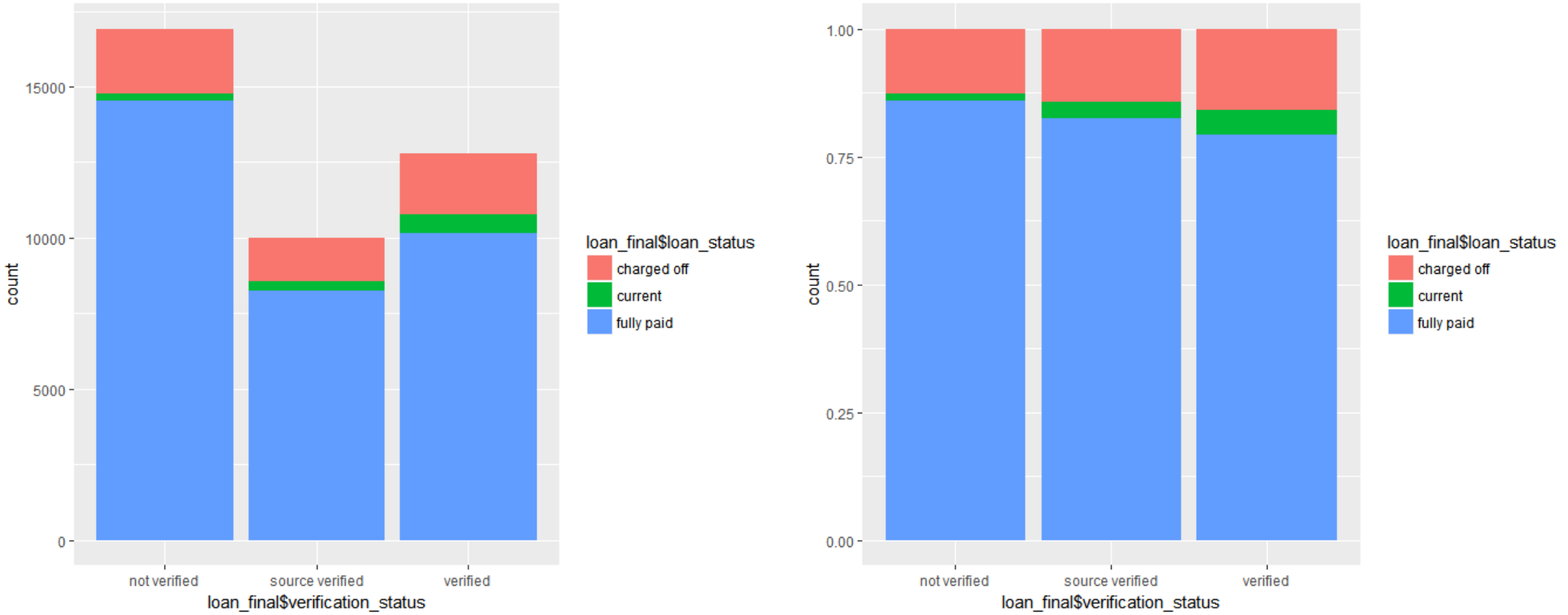






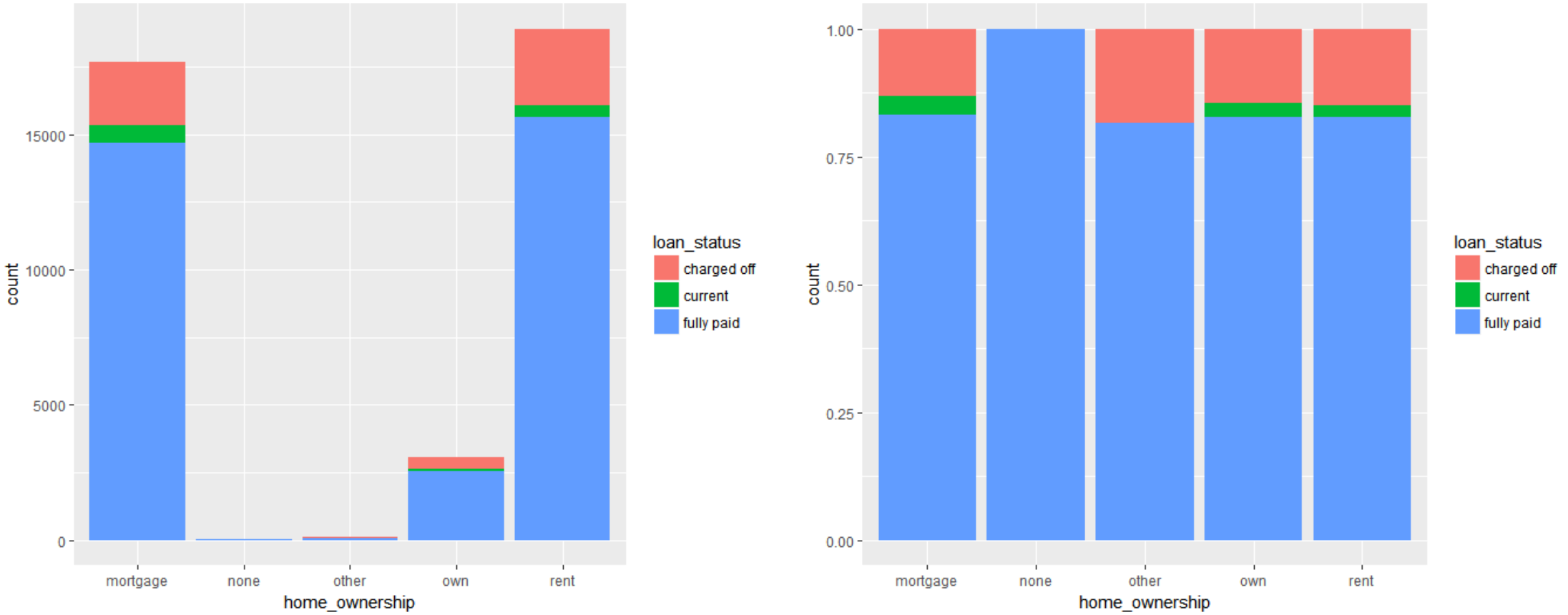
Loan with 5 year term are having higher instances of default (~25%) as compared to 3 year term(~10%). But in absolute numbers there are higher defaults for a 3 year term

Bivariate Analysis – Loan Status Vs Verification



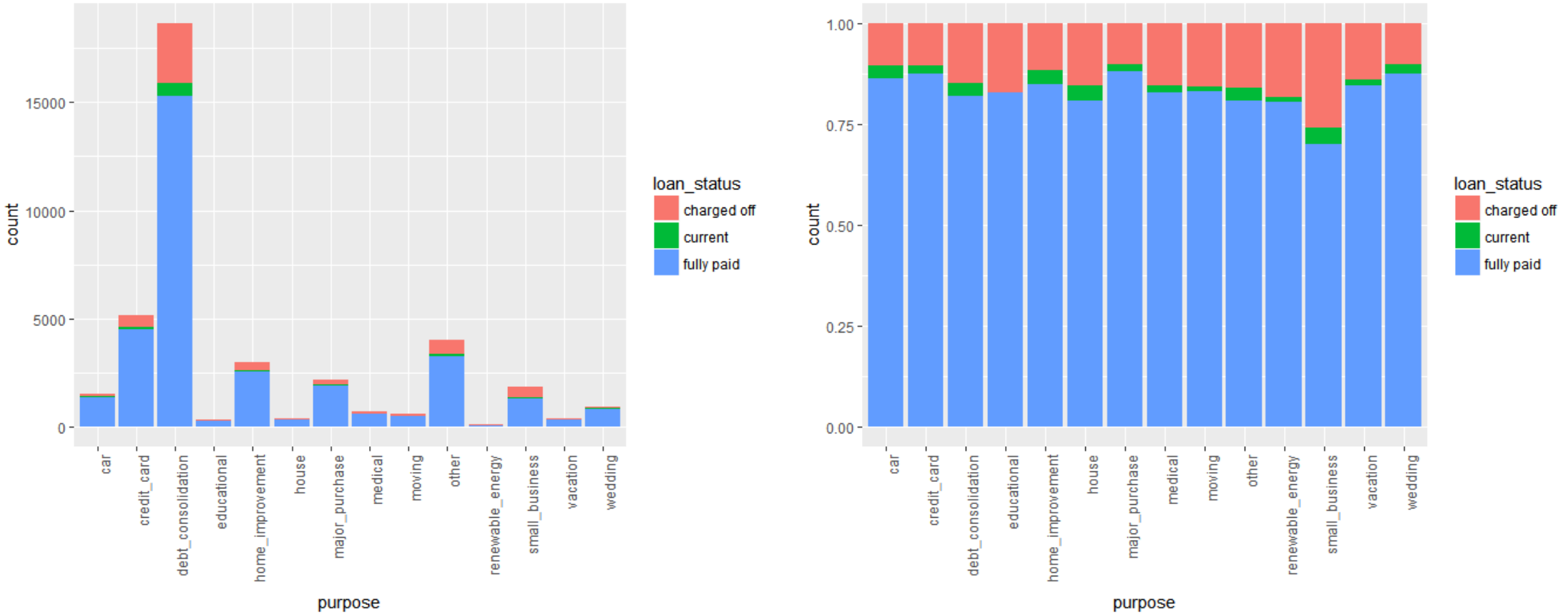
The verified loans have slightly more defaults that the unverified loans.

Bivariate Analysis – Loan Status Vs Home Ownership



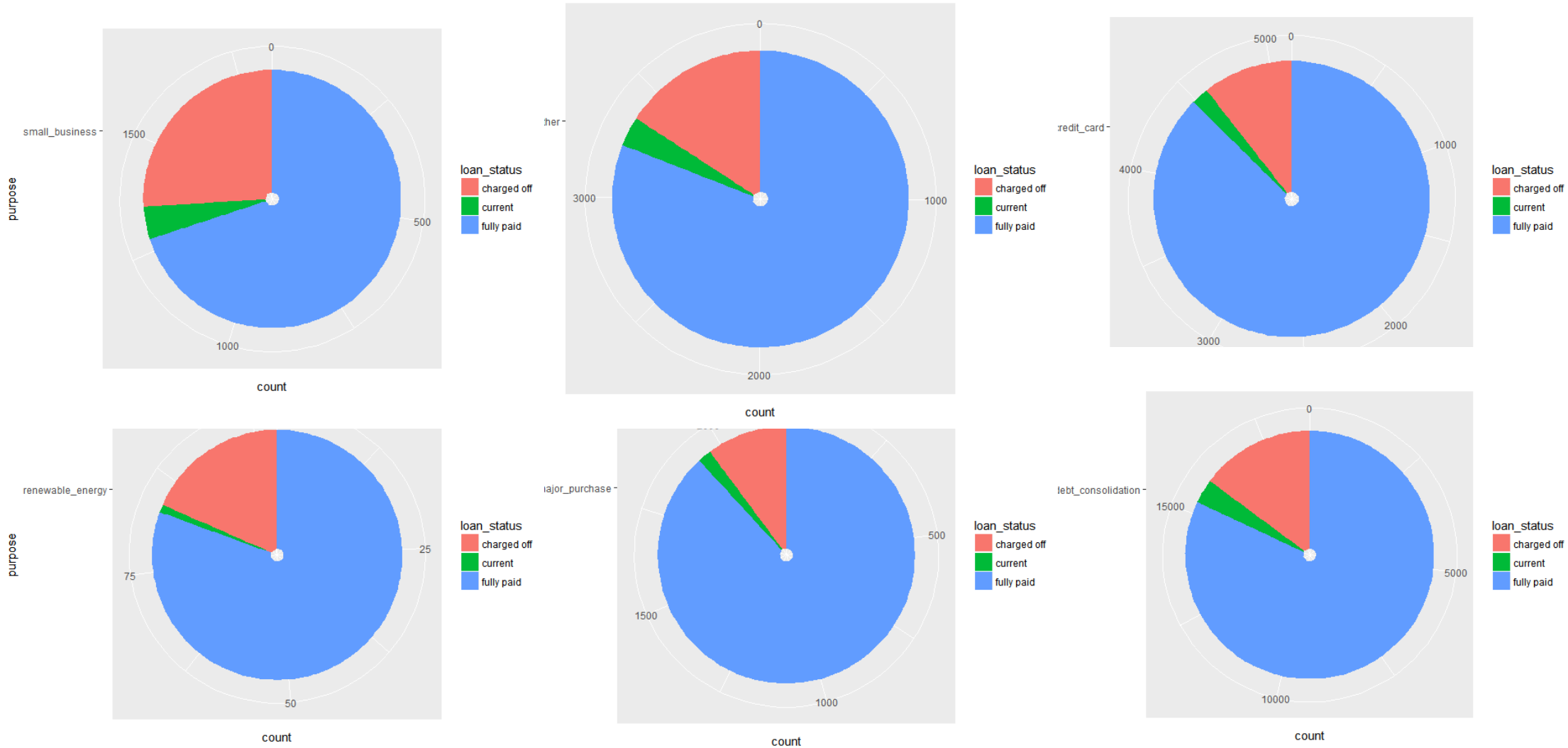
The loan defaults are highest for members having mortgage or paying rent. Compared to those who own a house, the members who live in rented or mortgaged accommodations seek loan in significantly higher number. This extra debt may be the reason for seeking out the loan.

Bivariate Analysis – Loan Status Vs Purpose



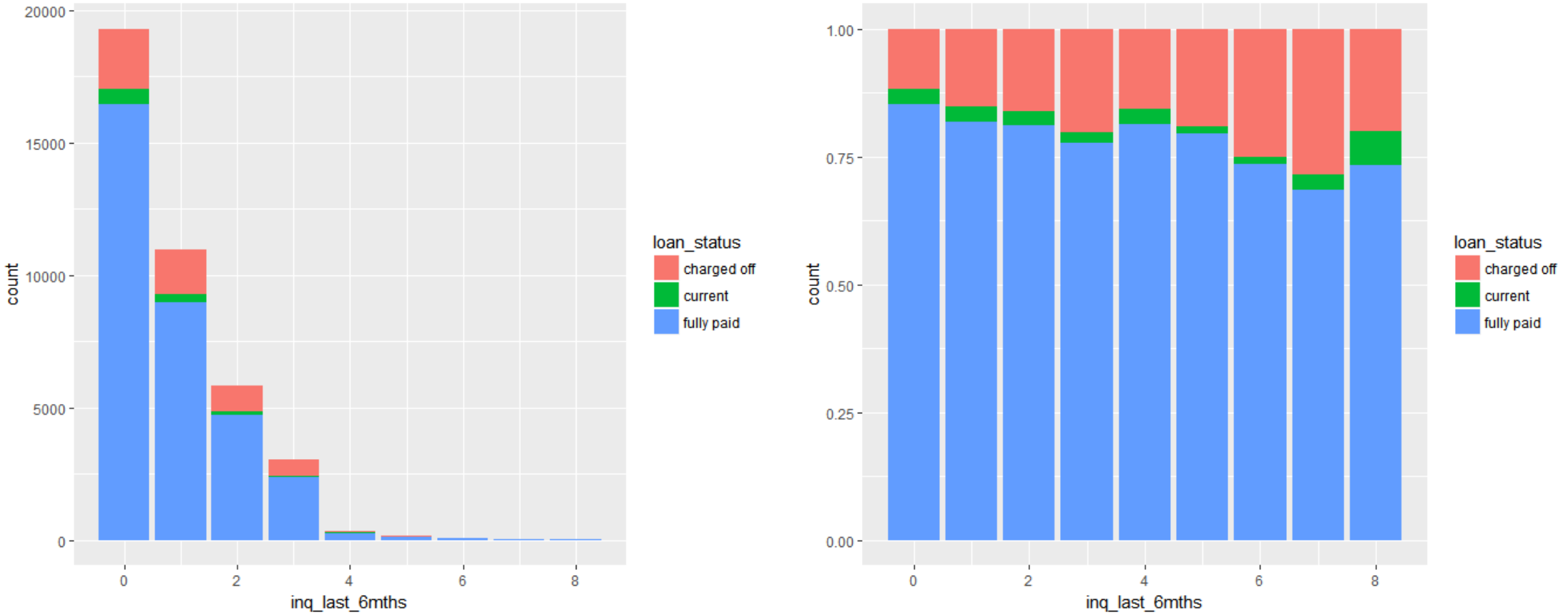
Debt Consolidation is the single largest cause for taking a loan, but those who take loans for small businesses default the most. This is because the payment of the loan is dependent on the success of the business and the closure rate of small businesses is much higher in the initial years.

Bivariate Analysis – Loan Status Vs Purpose



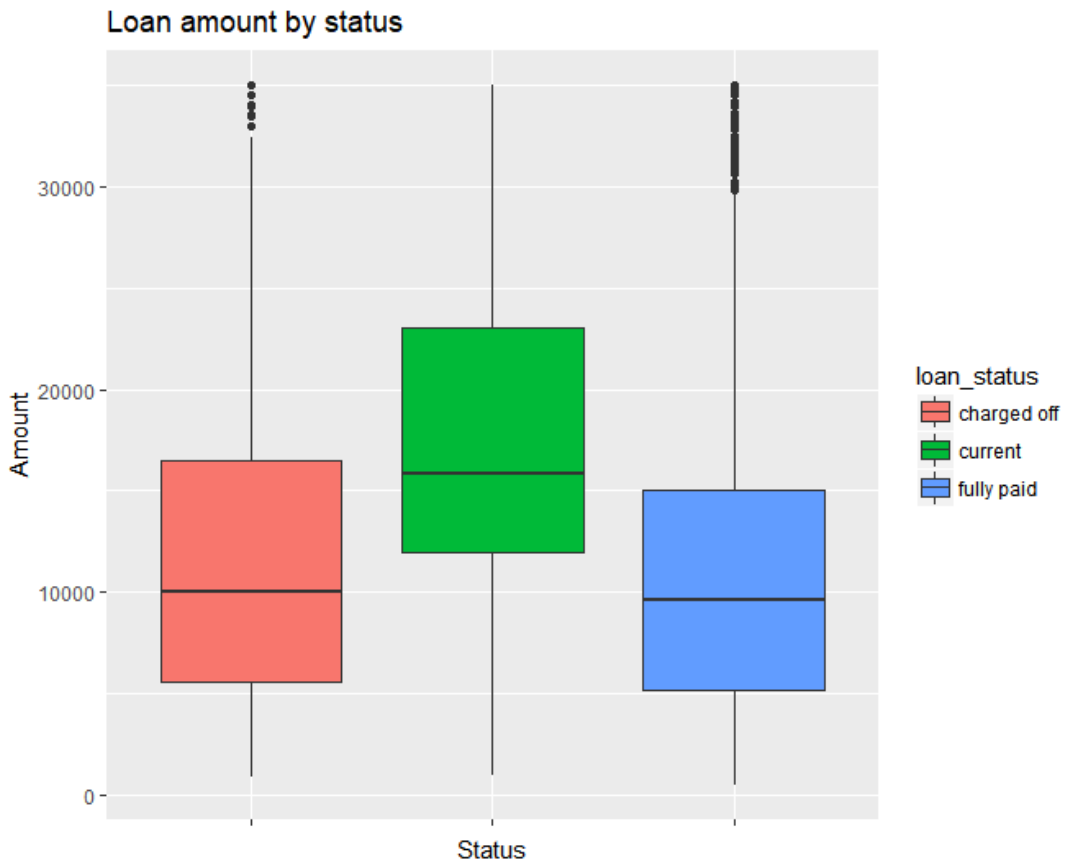
Loans status plotted against the purpose for taking the loan

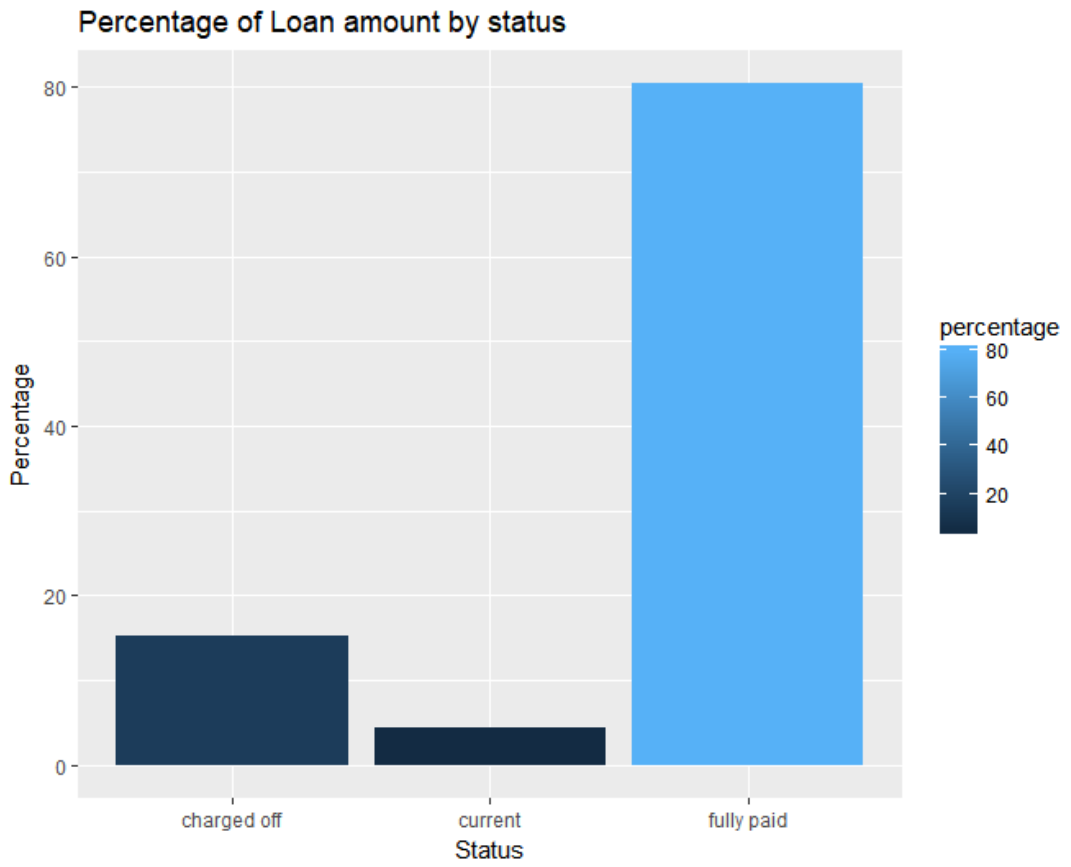
Bivariate Analysis – Loan Status Vs Inq Last 6 Months



As the number of inquiries in past 6 months increase for a person, the chances for default increase. For 6, 7 and 8 inquiries, the defaulters are 25%, 28% and 20% respectively.

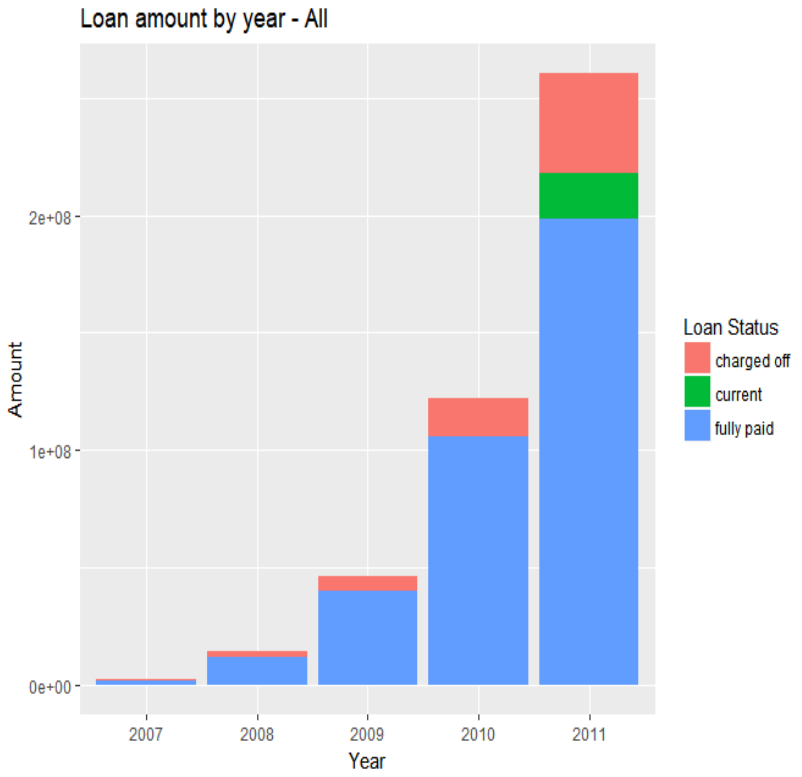
Bivariate Analysis – Loan Status Vs Loan Amt

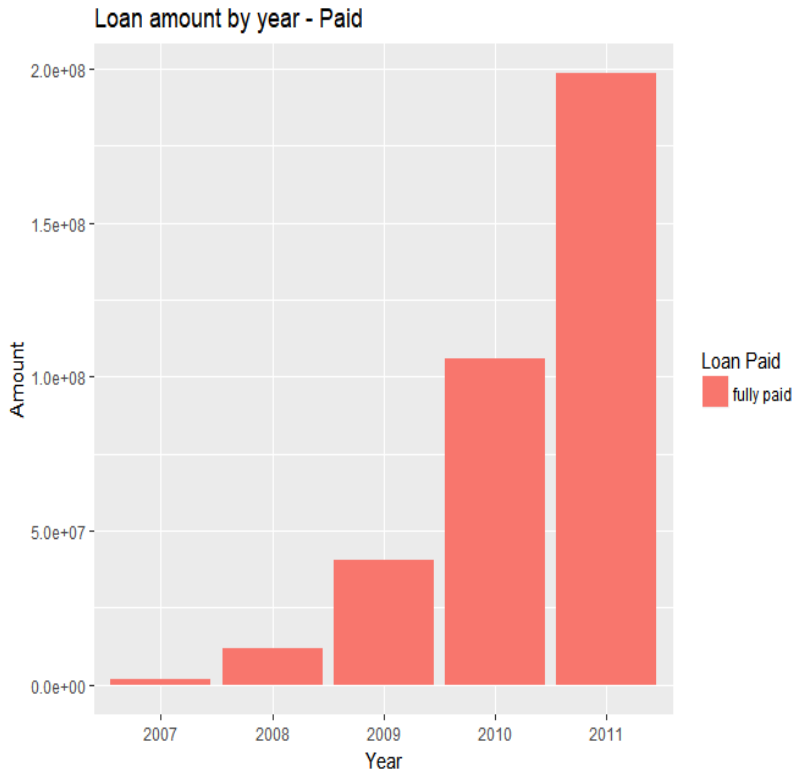


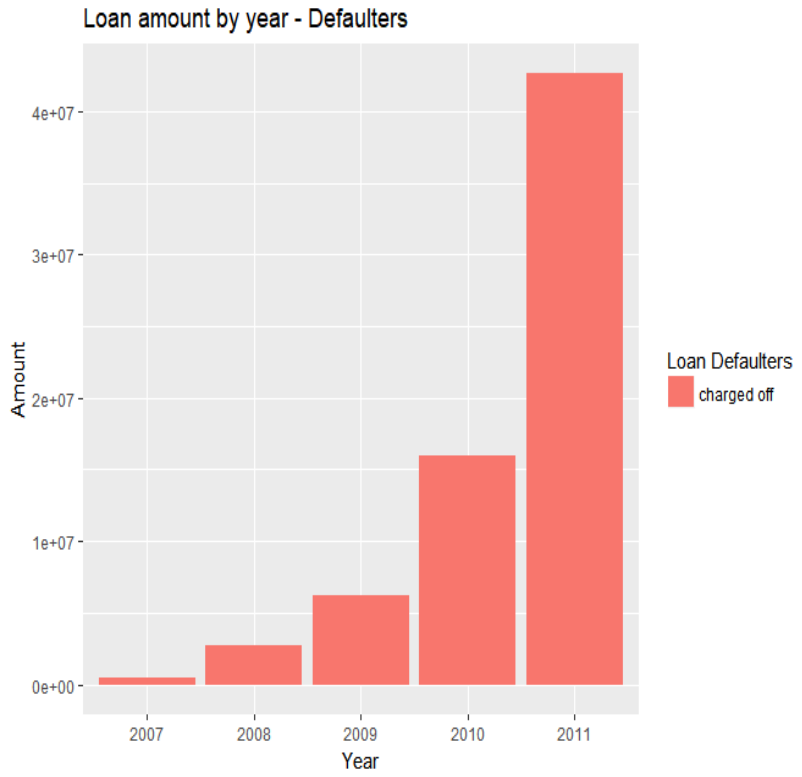


Around 80% of the loans taken by loan amount are paid off. The ones that are charged off have slightly higher loan amount compared to the ones paid off.

Bivariate Analysis – Loan Status Vs Loan Amount across Years

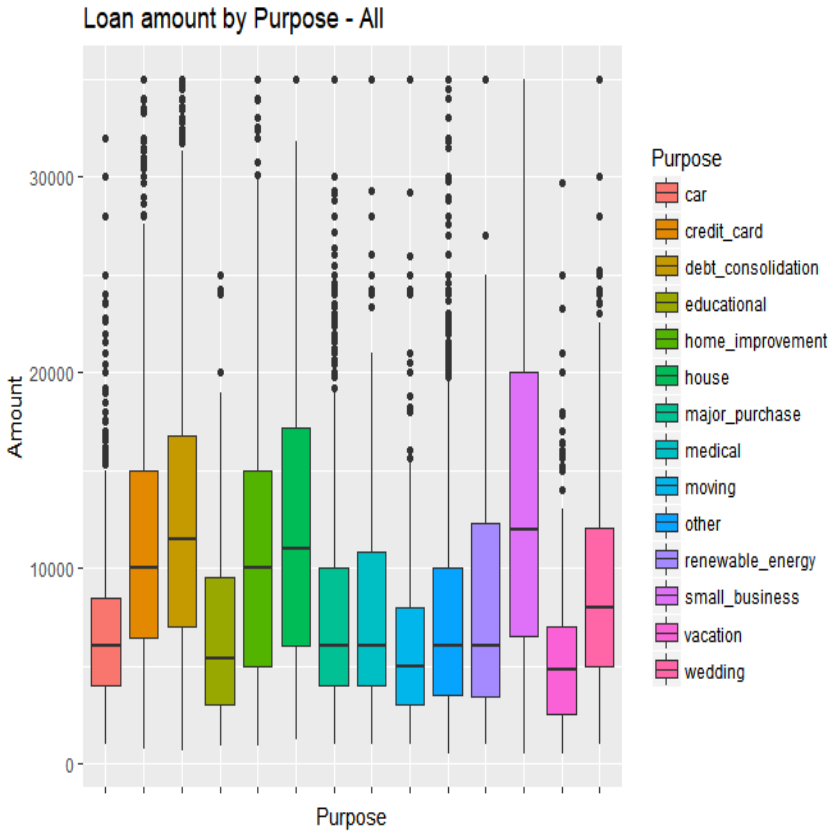


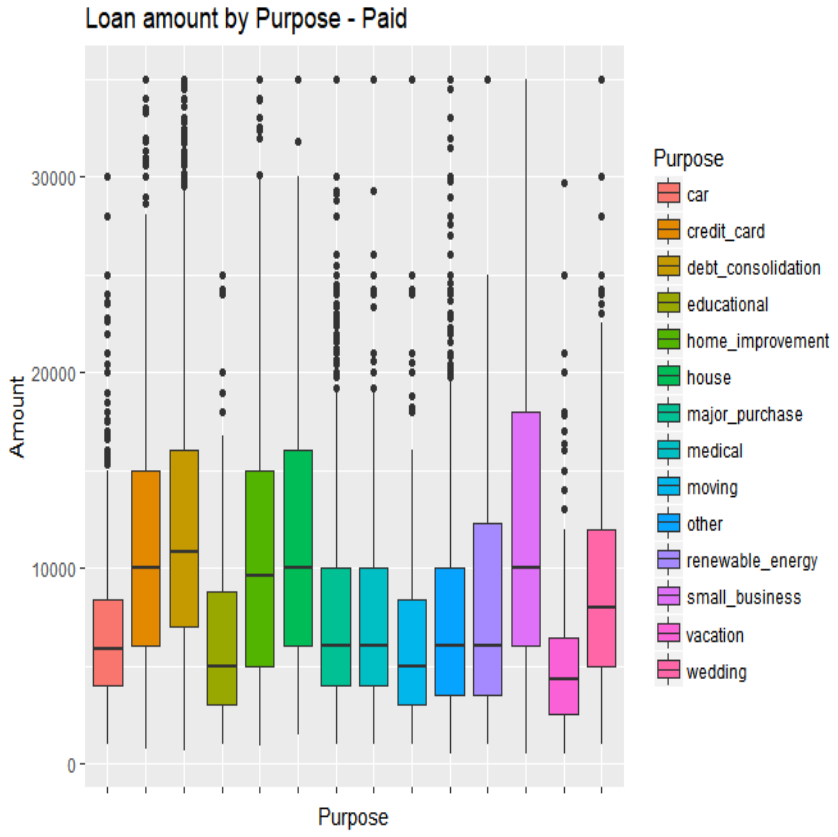


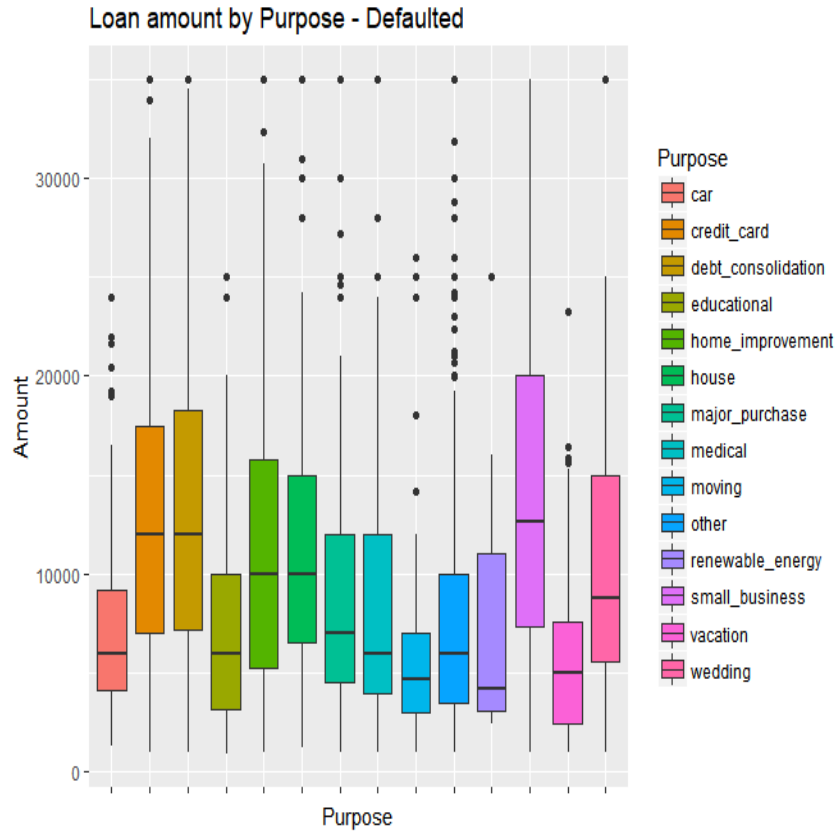


The number of defaulters in increasing year on year with the highest being in 2011. The jump in defaults in 2011 is significantly higher.

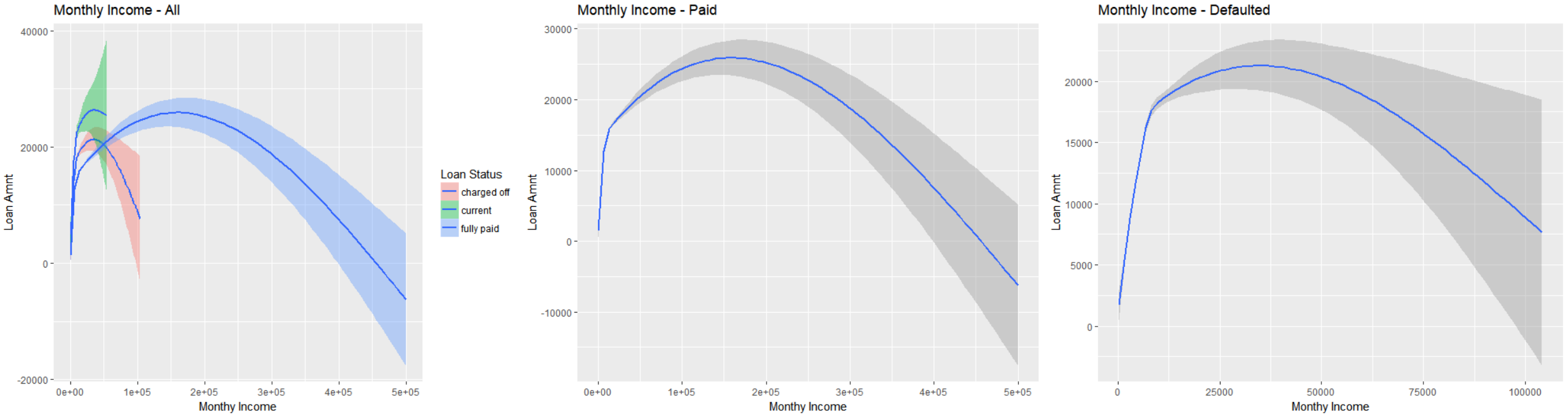
Bivariate Analysis – Loan Amount Vs Purpose







Bivariate Analysis –Loan Amount Vs Monthly Income



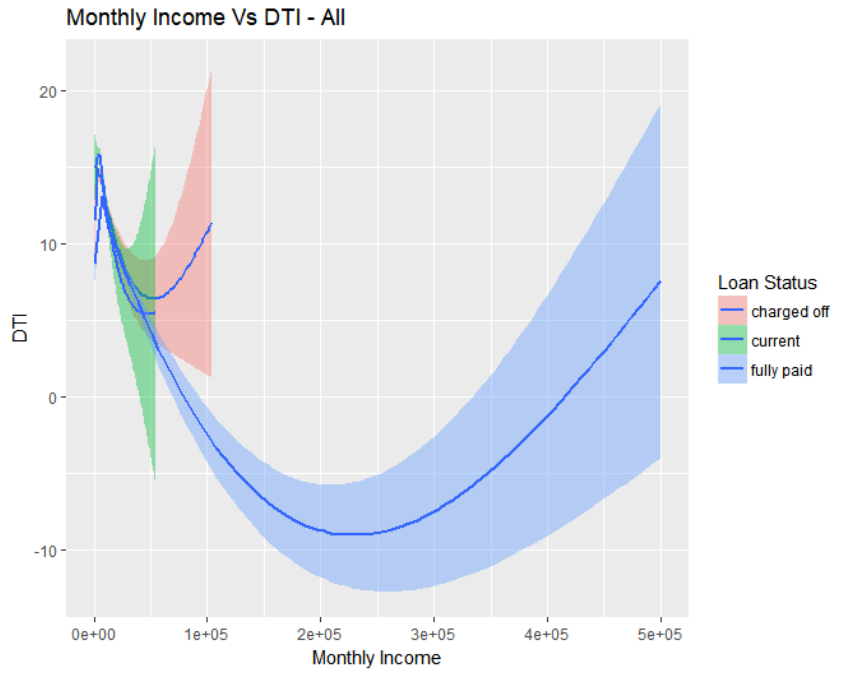
Correlation between loan_amnt and int_rate is positive (0.31) for all loans. Thus loan is increasing with interest rate.

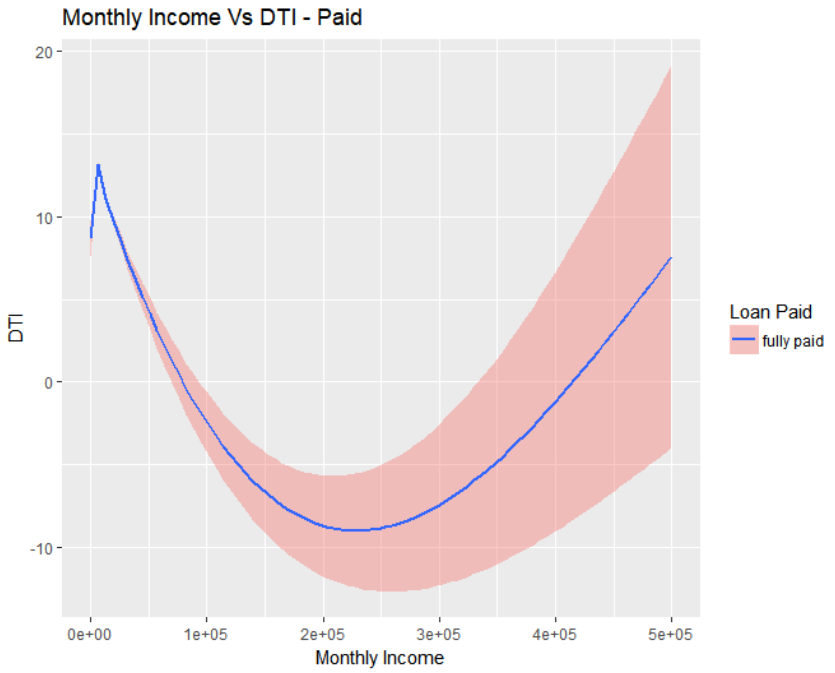
Correlation between loan_amnt and int_rate is higher (0.35) for defaulted loans.

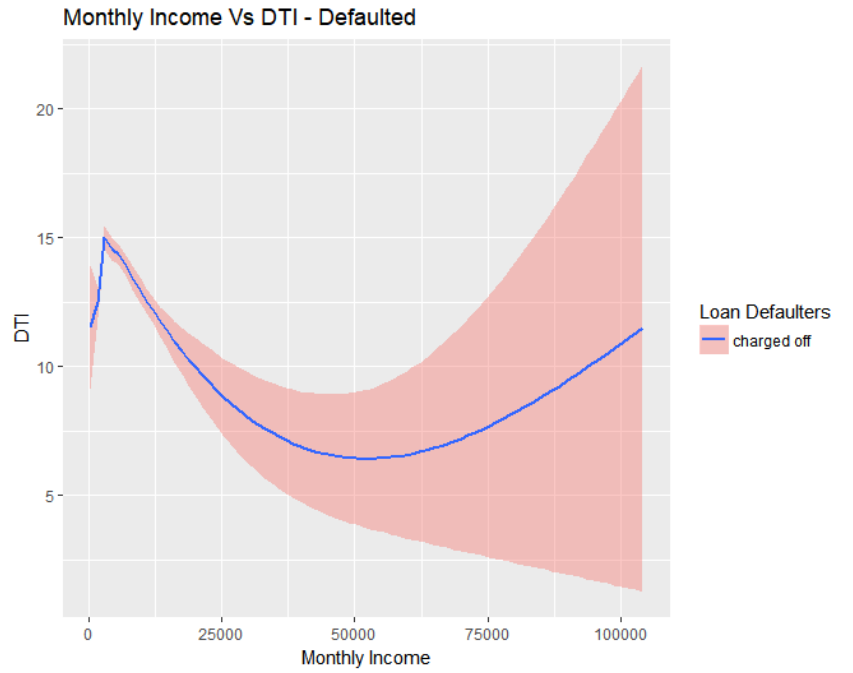
Correlation between loan_amnt and int_rate is lower (0.29) for paid loans.

Thus the rate of interest increase for paid loans is lesser than that for the defaulted loans.

Bivariate Analysis – DTI Vs Monthly Income

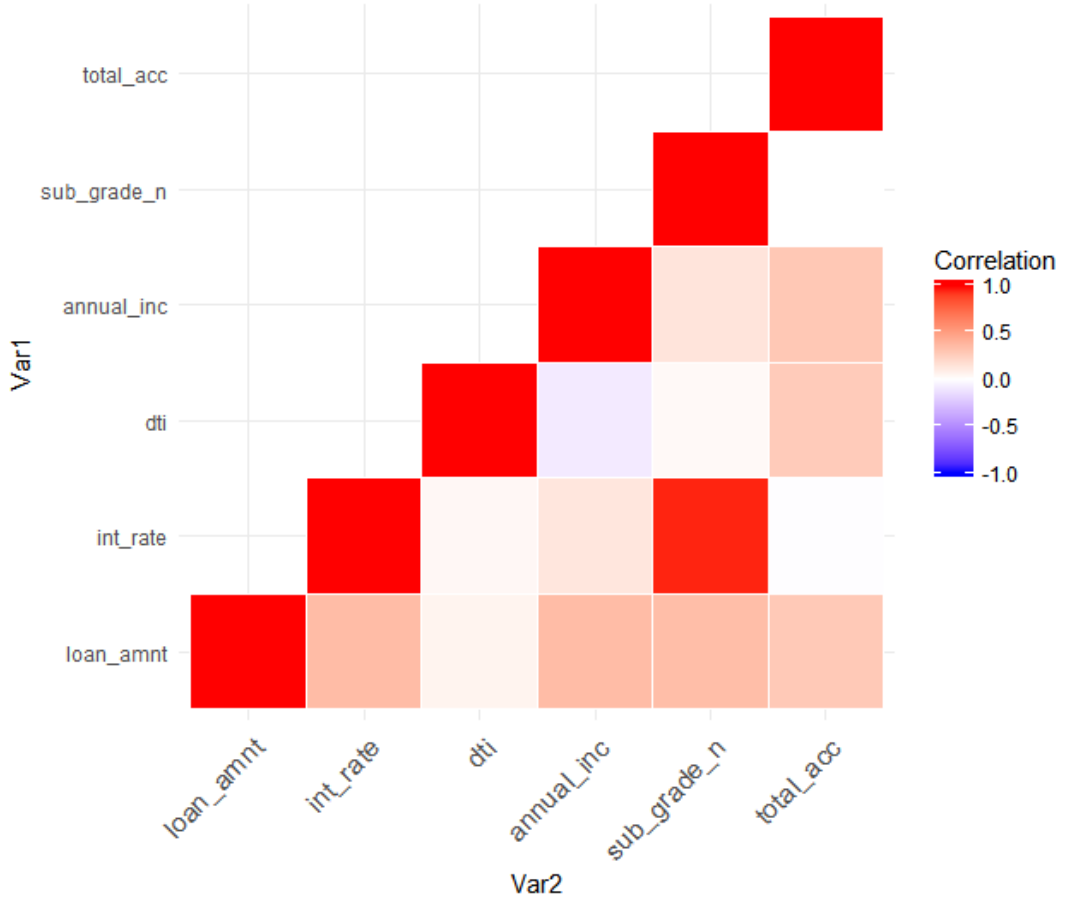




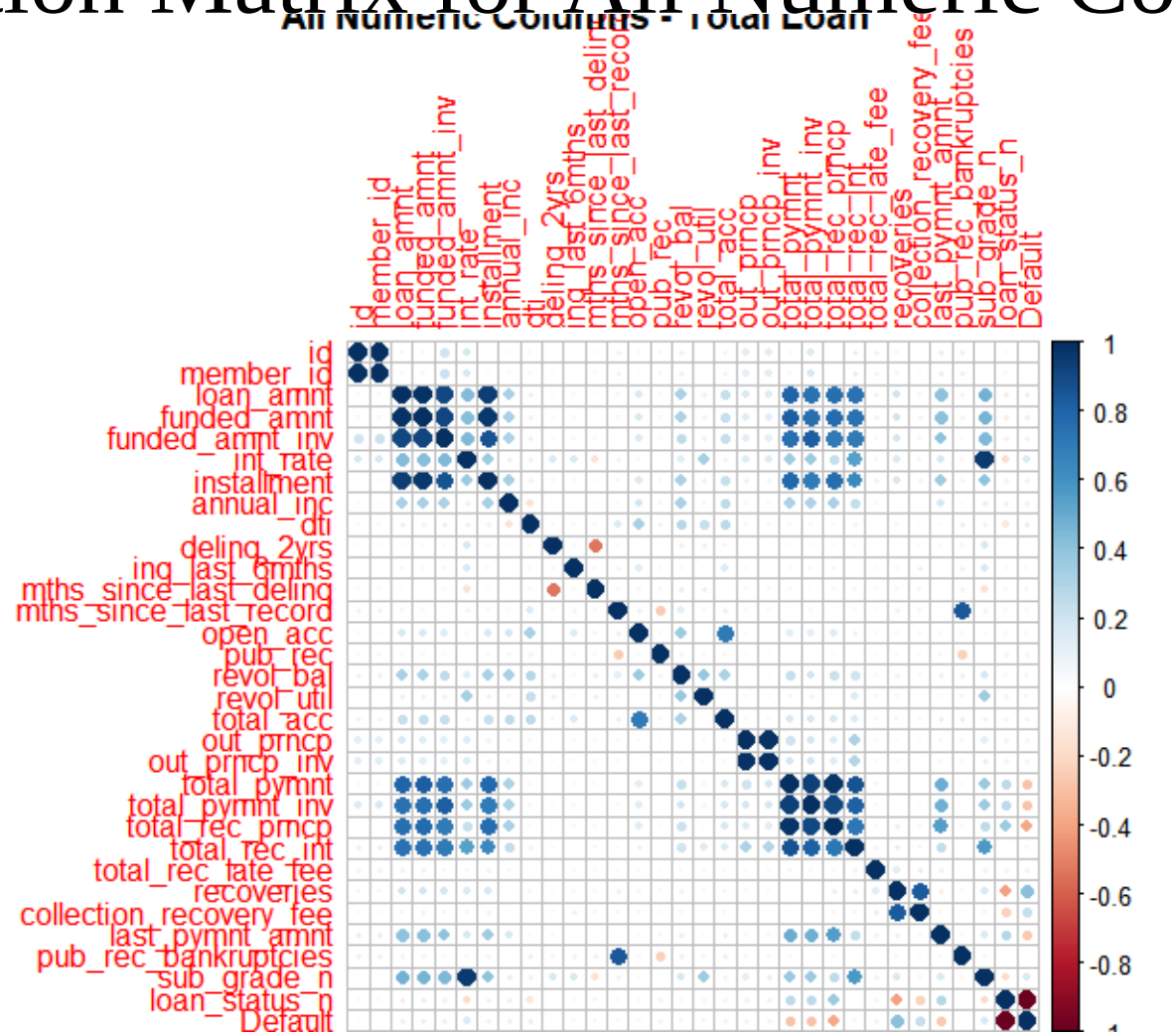


DTI is more for less monthly income, It increases till it reaches its peak for a significantly low monthly income and then it decreases with increasing annual income, finally goes to medium level of dti with max annual income. The defaulters have significantly higher DTI as compared to those who have paid off.

Correlation Matrix



Correlation Matrix for All Numeric Columns



The Correlation matrix for all numeric columns for all the loans. The blue colored dots show positive correlation with the darker dots having the higher correlation.

Conclusions

- The maximum number of defaults are for B, C and D grades, but percentage wise the defaults increase from A (5%) to G(25%). So we can conclude that the risk increases from A to G.
- Sub-grades defined from A1 to G5 are clear indicators of increasing interest and increasing loan amount. Percentage of default increase for the grades E, F, G as compared to A,B,C and D.
- Most of the loans are sanctioned for A.B and C grades and across all the grades, and nearly 50% loans are sanctioned for debt consolidation.
- The interest rates across Grades and sub grades show an increasing trend across years with the highest rise seen in 2011 especially across D, E, F and G grades and their respective sub grades
- The absolute number of defaults are high for people with 10+ years of experience. But percentage wise the defaults are maximum for N/A experience. If the employment length is the duration with the current employer, then those in N/A category may be unemployed and hence the high rate of default.
- The interest rate is significantly higher for the defaulters compared to the loans that were fully paid. The loans that are currently running have the highest interest rate and that is because all these loans started in 2011 when the interest rates were significantly higher
- Loan with 5 year term are having higher instances of default (~25%) as compared to 3 year term(~10%). But in absolute numbers there are higher defaults for a 3 year term
- The loan defaults are highest for members having mortgage or paying rent. Compared to those who own a house, the members who live in rented or mortgaged accommodations seek loan in significantly higher number. This extra debt may be the reason for seeking out the loan.
- Debt Consolidation is the single largest cause for taking a loan, but those who take loans for small businesses default the most. This is because the payment of the loan is dependent on the success of the business and the closure rate of small businesses is much higher in the initial years. Small_business (26%), renewable_energy(18%) and educational(17%) are top divisions for defaulters
- As the number of inquiries in past 6 months increase for a person, the chances for default increase. For 6, 7 and 8 inquiries, the defaulters are 25%, 28% and 20% respectively.
- Correlation between loan_amnt and int_rate is positive (0.31) for all loans. Thus interest rate is increasing with loan amount. Correlation between loan_amnt and int_rate is higher (0.35) for defaulted loans. Correlation between loan_amnt and int_rate is lower (0.29) for paid loans. Thus the rate of interest increase for paid loans is lesser than that for the defaulted loans.
- **The Driver Variables for Loan Default are: Purpose, Interest Rate, Loan Amount, Inq since last 6 month, DTI, Home_ownership**
- Grade/ Sub Grade is derived from the driver variables and hence not considered driver variables themselves

References

- https://www.lendingclub.com/
- https://www.lendingclub.com/info/prospectus.action
- https://www.econstor.eu/bitstream/10419/148902/1/875128831.pdf
- http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139427