Presentation Report

Lending Club Case Study

Group Members:

1. Mohit Angurala
2. Mithinti Bala Venkata Mani Suhitha

The problem statement

Organization

Lending Club organization a largest online loan marketplace, specialized in lending different categories of loans to the urban customers.

Context

Lending Club would like to know about the strong factors behind loan default and for this the company makes use of this knowledge for its portfolio as well as risk assessment.

Problem statement

To analyze the way, consumer and loan attribute affect the tendency of default. As a data scientist we are supposed to analyze the risky loan applicants.

Dataset given contains information of loan applicants issued through the period of 2007-11.

Analysis Approach

Clean Data	**Univariate Analysis**	**Segmented Univariate Analysis**	**Bivariate and Multivariant Analysis**	**Summarize Results**
Impute the missing values	Check distributions and frequencies of various numerical and categorical variables	Drop columns with null values, all random values or single	Do correlation analysis Check how two variables affect each other or a third variable	Publish insights and observations
Converting values to int, float, date time representation	Create derived variables	Analyze variables against segments of other variables	Analyze joint distributions	

Data Cleaning

Delete Columns: Deleted columns with more than 50% Null Columns

Delete Rows: For the remaining Nulls removed the rows with Null values

Duplicate data: Removed the duplicate rows and columns

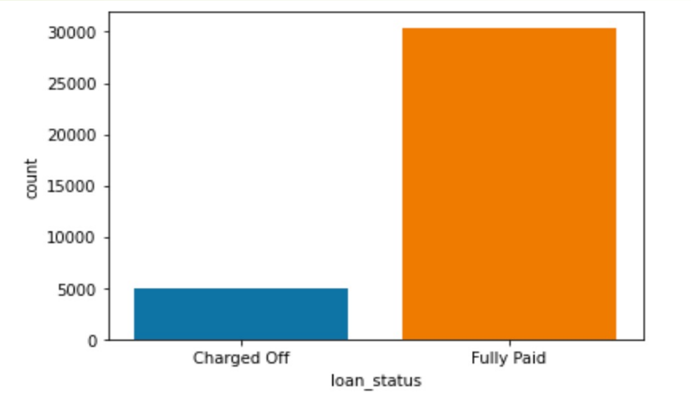
Sanitary Check: Checked whether data is logically correct and cleaned the wrong data

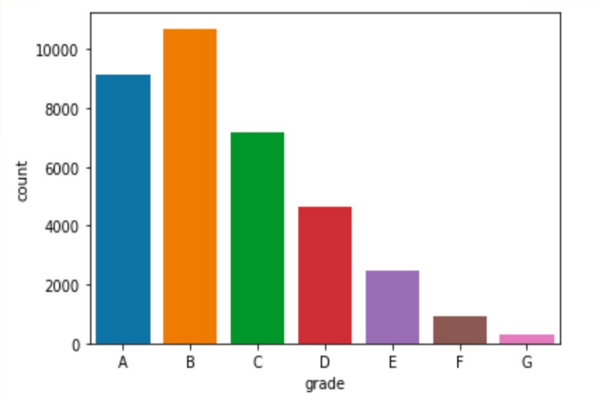
Data Preparation

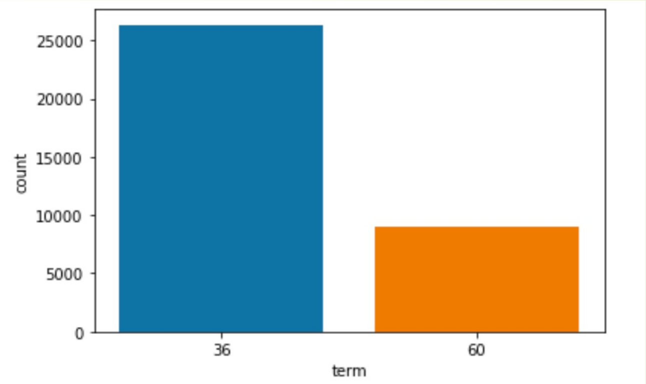
- Columns which are 100% unique columns cannot bifurcate the data so removed them.
- Columns with single value do not add any value to the analysis so removed them.
- Date columns are transformed to correct shape and extracted month and year related information into derived columns.
- Standardize the precision value of float columns.
- Modified categorical variable to standard format for ease of analysis like term and emp_length

Univariate Analysis

After Data cleansing and preparing there are 44 columns in the dataset and univariate analysis is performed all the columns. Few important results are added in the slides

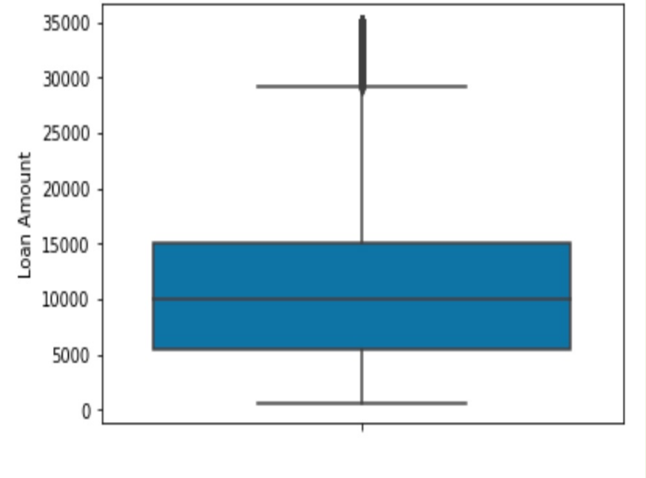




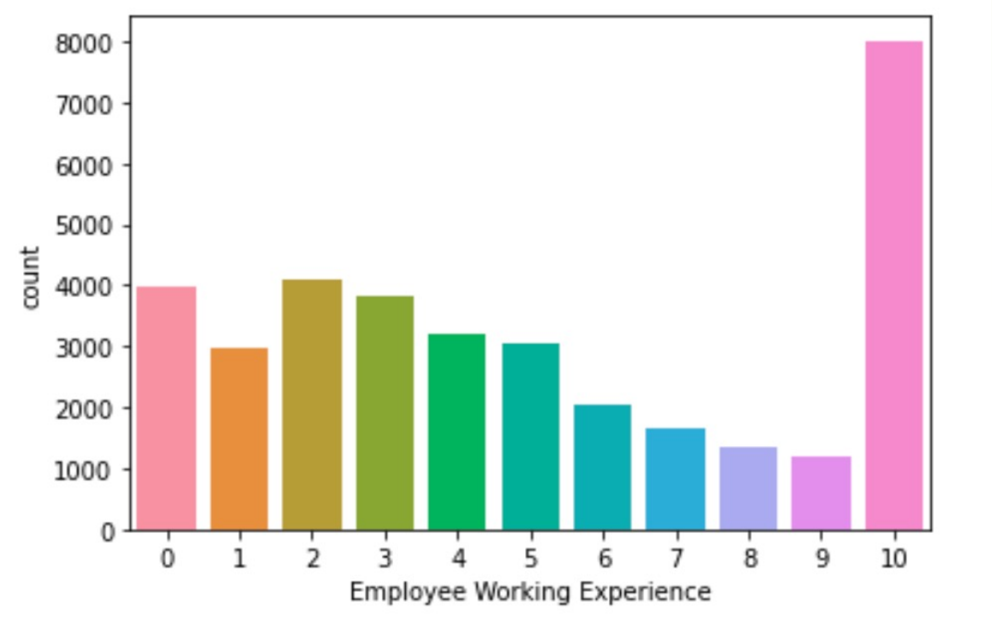


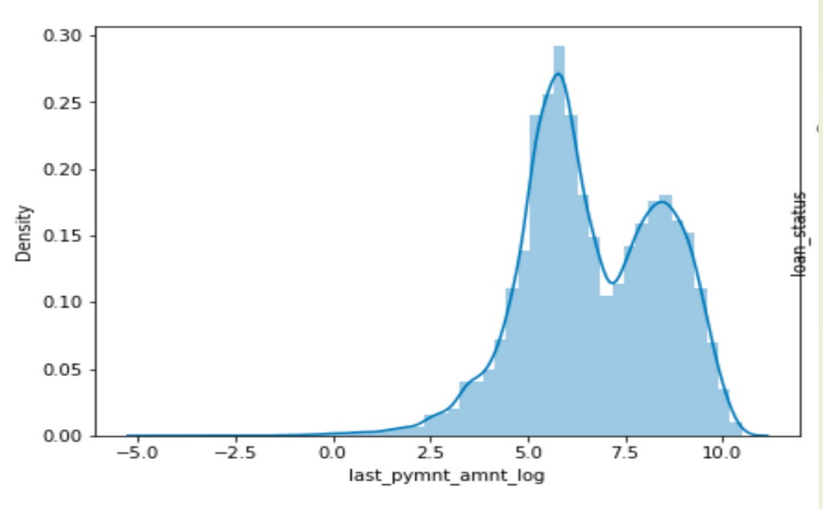
Inferences

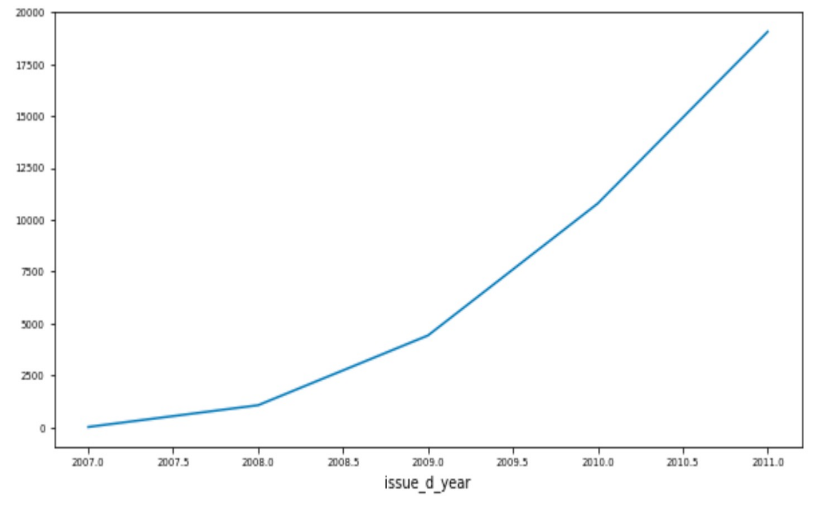
- Loan status the targeted variable is categorical and has three variable of which Fully paid and Charged Off are related to the problem and Fully paid is dominating the other leading to class imbalance.
- Grades A,B loans are opted by a greater number of borrowers.
- Loans of 36-months term are been opted most of the borrowers.
- Loan amount ranges from 500-35,000 having median at 10,000.



Univariate Analysis





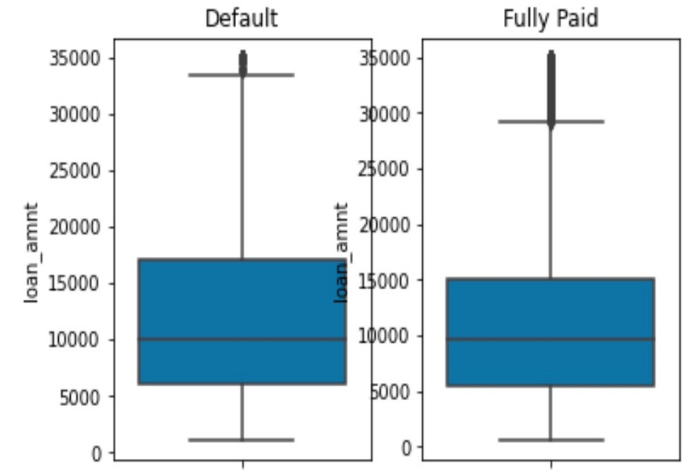


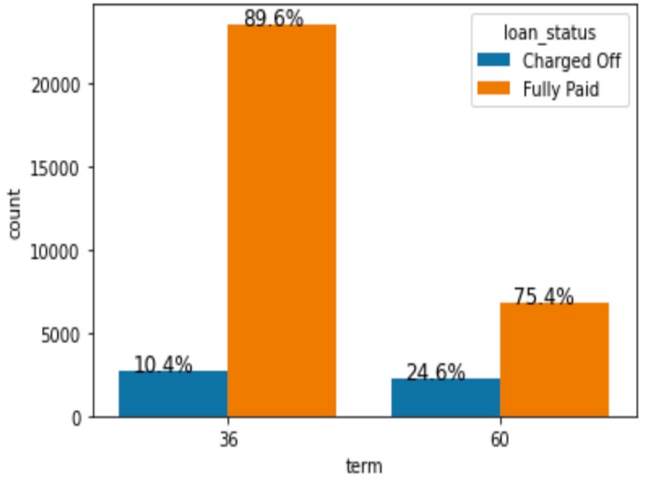
Inferences

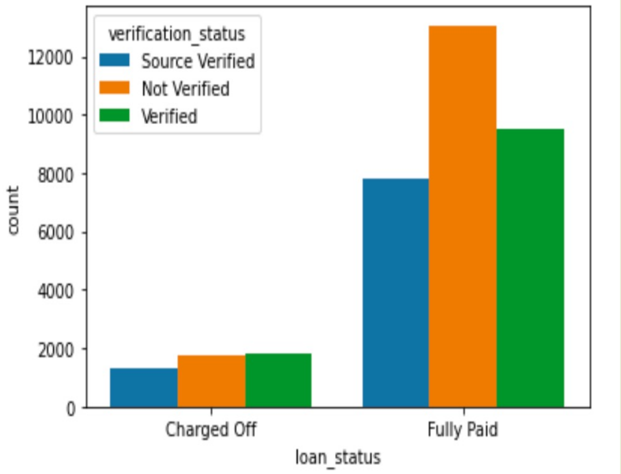
- Majority of borrowers who applied for the loan have greater than or equal to 10 years of experience.
- Last payment amount by a borrower is following rough normal distribution.
- The number of claims issued increased each year form 2007 – 2011.

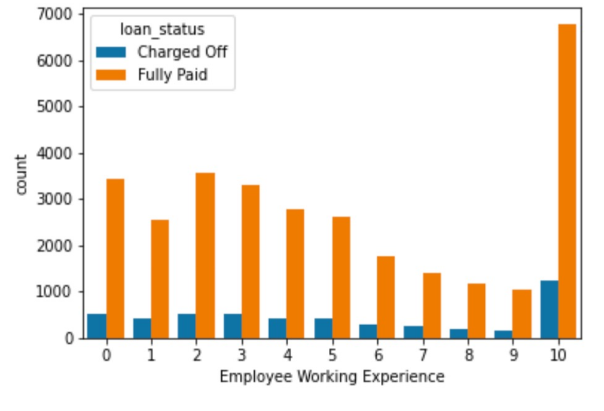
Segmented Univariate Analysis

Segmented univariate analysis is also done for almost all the relevant columns and added few important results





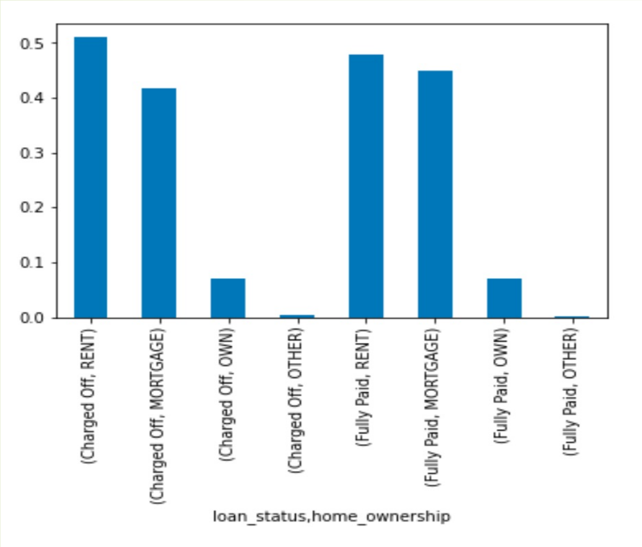


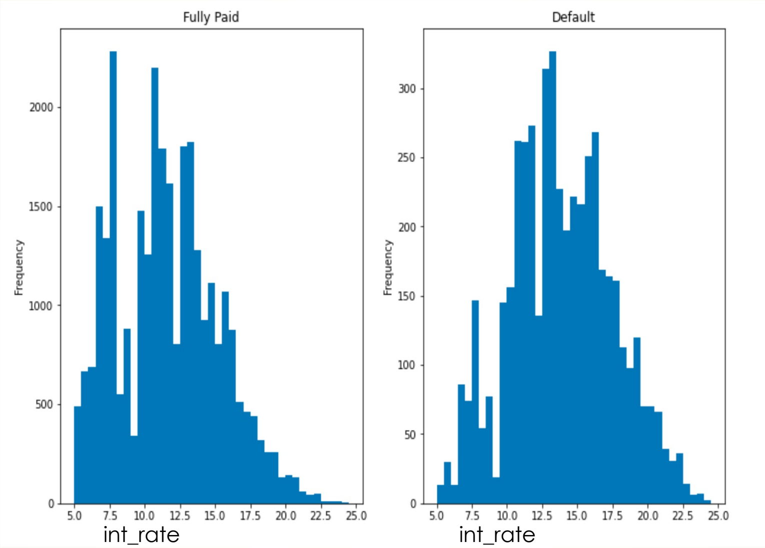


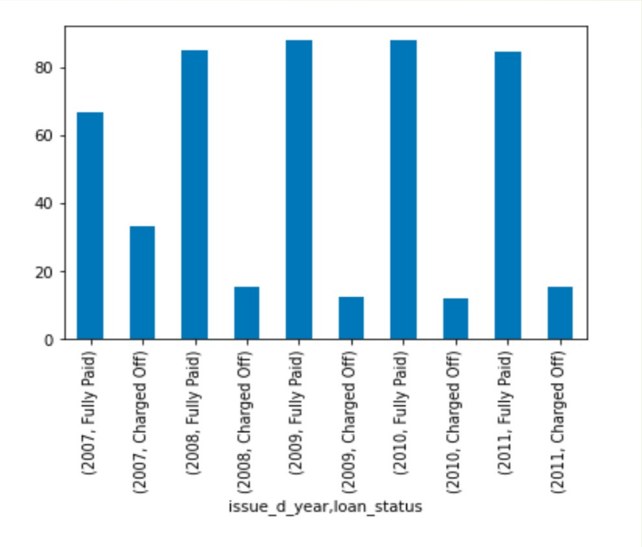
Inferences

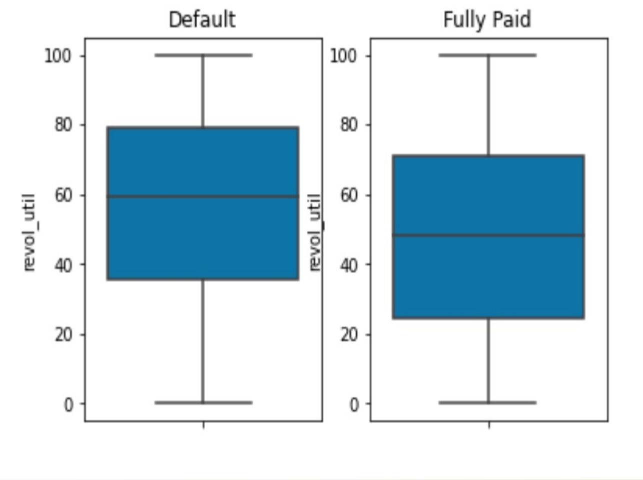
- Loan amount value is comparatively low for fully paid loans than defaulted with small range between 5,000-15,000.
- Of all 36-month term claims 90% are paid.
- Not verified claims are leading in fully paid claims and Verified claims are leading in Defaulted claims which is very contradicting with the purpose of verification.
- Employees with 10 years are leading in both categories which might be because of imbalance in the category counts.

Segmented Univariate Analysis





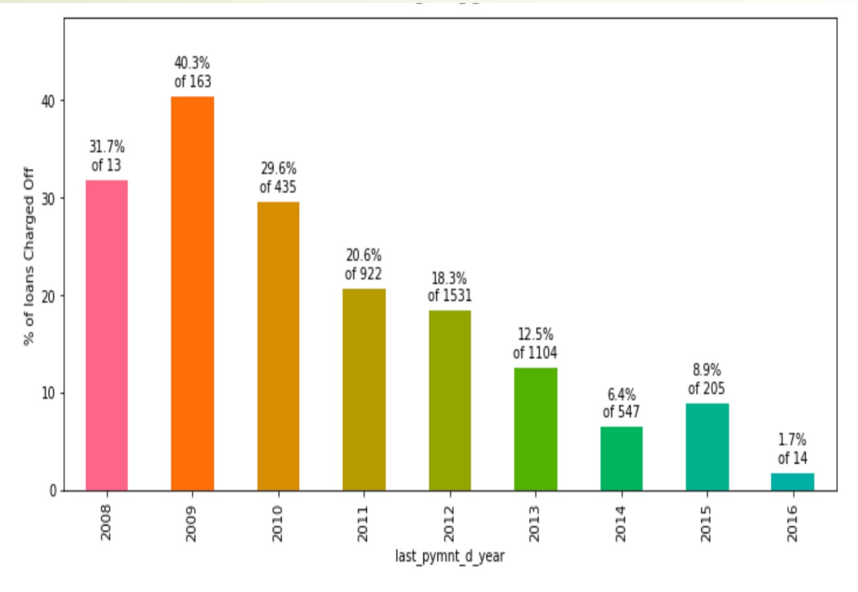


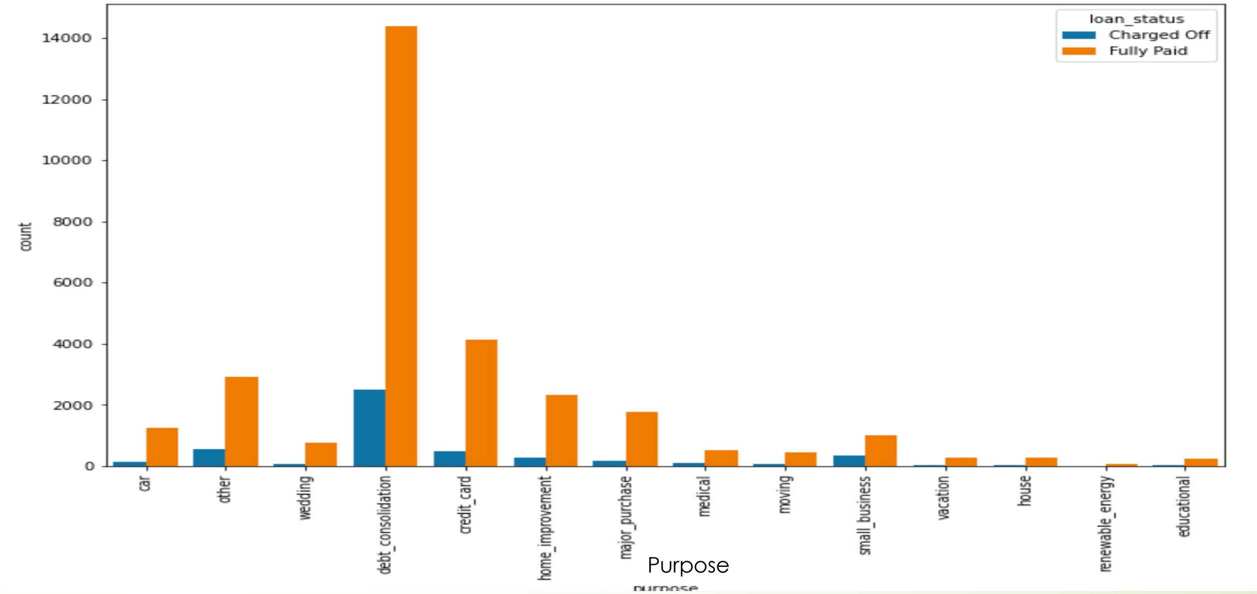


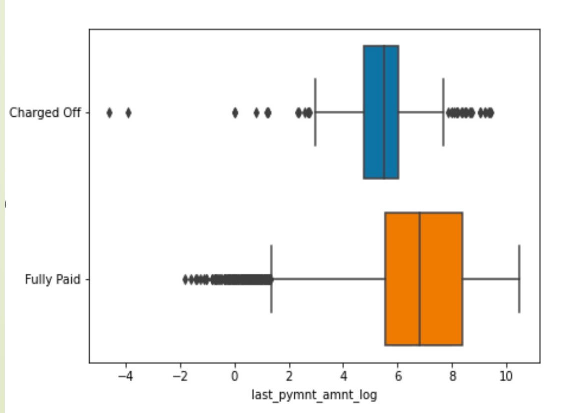
Inferences

- The borrowers with rented home are leading in both the categories.
- Fully Paid loans are having a greater number of interest rates at interest rate 7.5-8 and 10.5-11 next. Defaulted more when interest rate is around 11.00 - 15.00.
- Of all the years 2010 has more Fully paid claims. The number of loans increased for each year along with defaulted claims.
- Fully paid loans have less revol_util when compared to Defaulted by nearly 10% .

Segmented Univariate Analysis







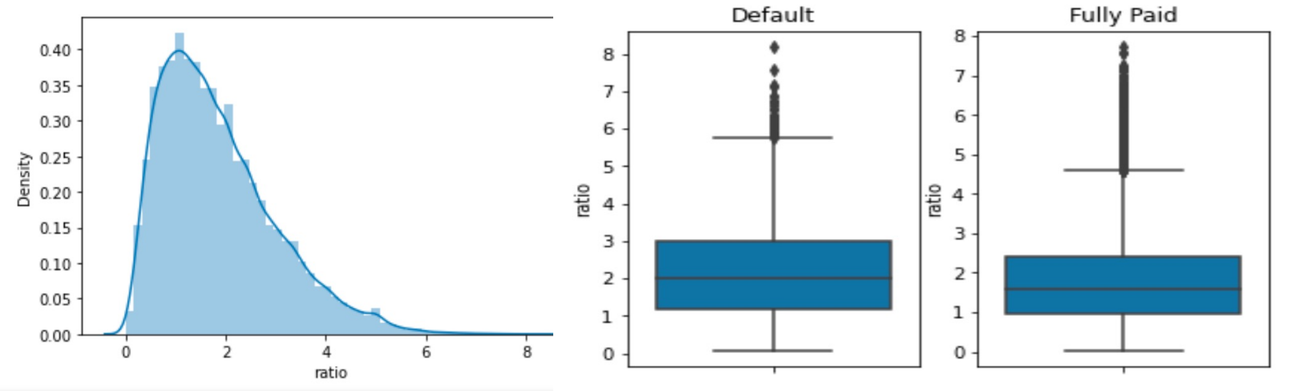
Inferences

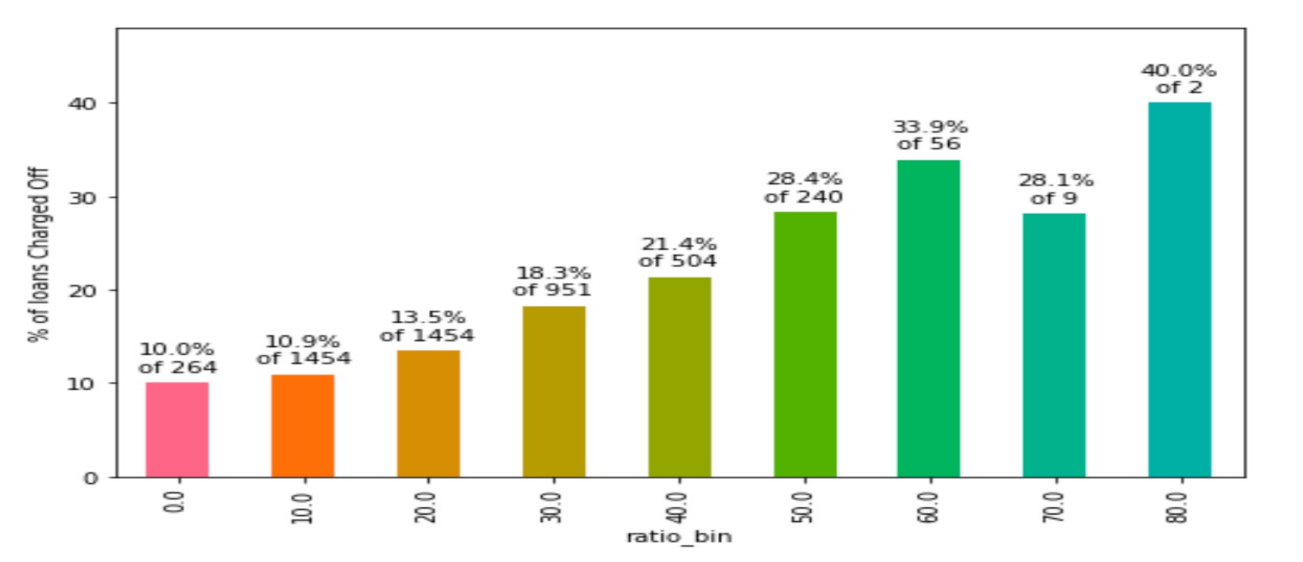
- Of all the loans the once with last payment in 2009 has more default percentage and 2016 least which is because of a smaller number of claims available else we can say 2014 the least. We can further analyse why 2009 have more defaults (if there is any business change or other economic factors).
- The loans sanctioned for the purpose of wedding and major purpose has more fully paid percentage with 90% and sanctioned for Education has nearly 16% default percentage (The count of loans given for education are very less so the percentage is showing very huge).
- Fully paid loans have greater last payment amount when compared to defaulted claims.

Bivariate Analysis
Loan Amount vs Annual Income

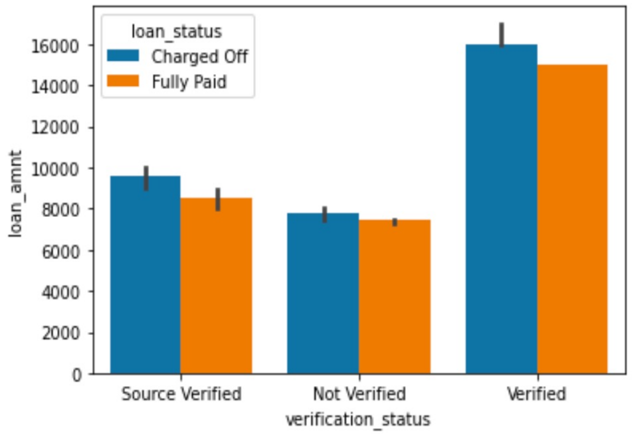
Inferences

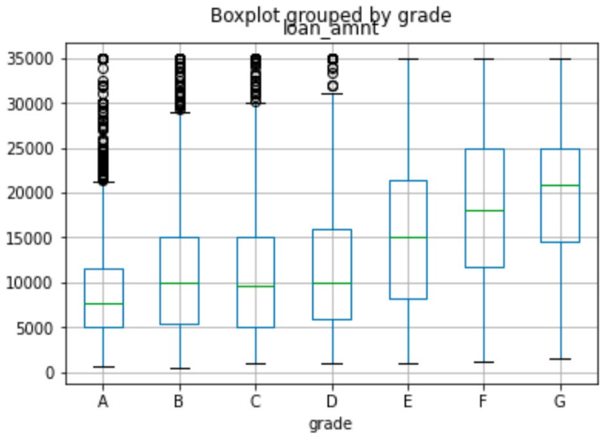
- Derived variable “Ratio” is the ratio of Loan Amount to Annual Income multiplied by 10.
- Defaulted claims have relatively high ratio(2) compared to Fully paid(1.6).
- loans with loan amount taking more part of annual income are more likely to get defaulted.

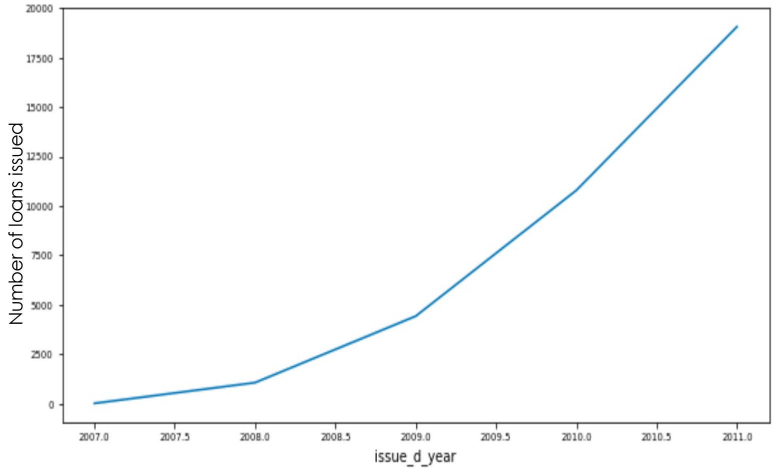


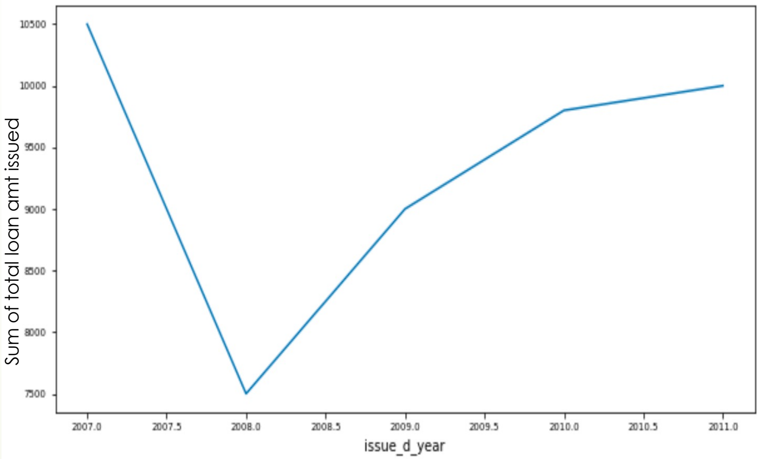


Bivariate Analysis - Loan Amount





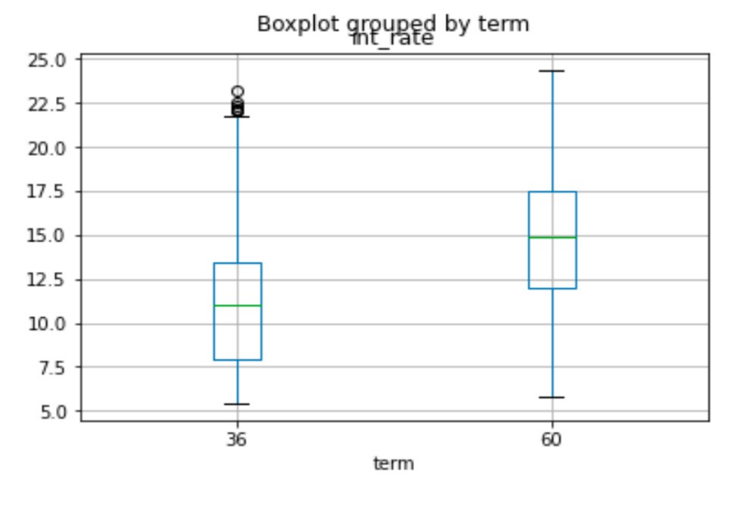


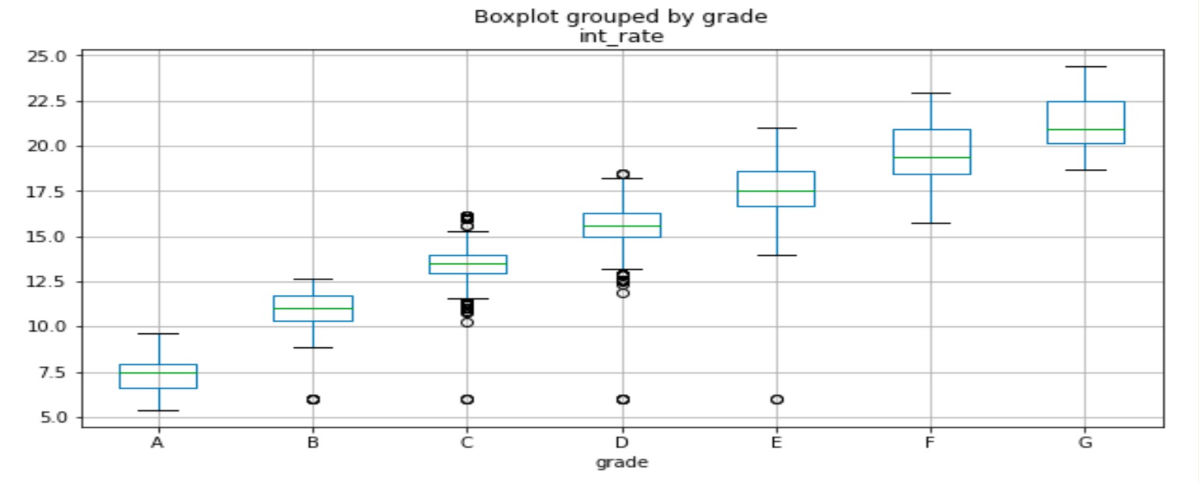


Inferences

- High loan amounts are been verified and low amount loans are either Source verified loans or Not Verified which is inline with need.
- Loan amount increases with grade of the loan.
- Even though the number of claims issued increased every year there is a huge drop in the total loan amount issued in 2008.

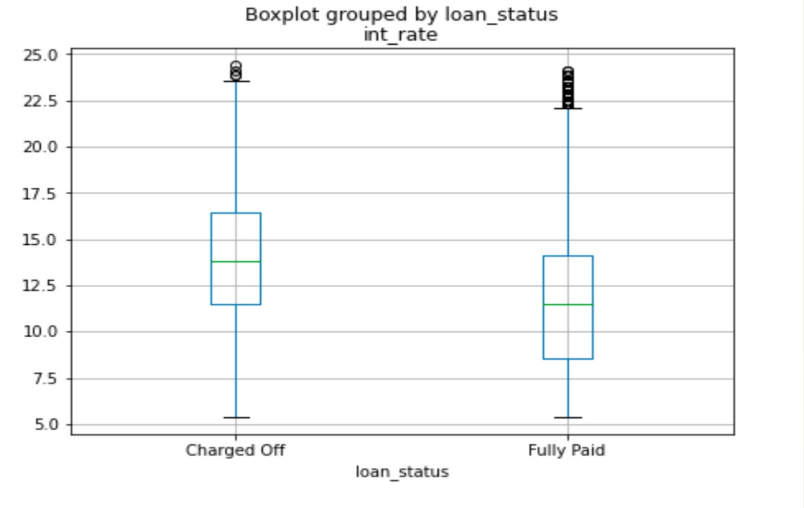
Bivariate Analysis - Interest Rate



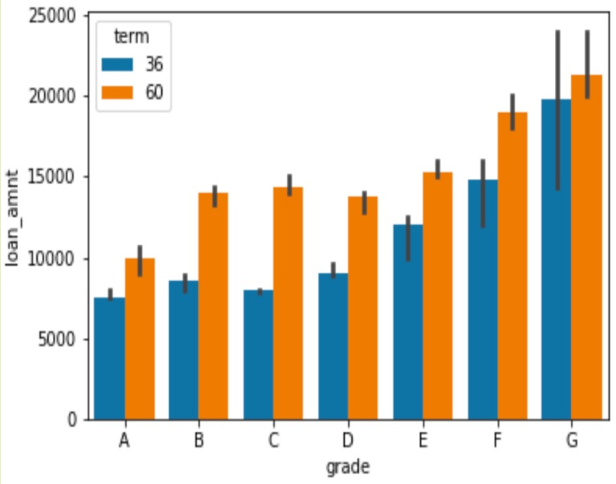


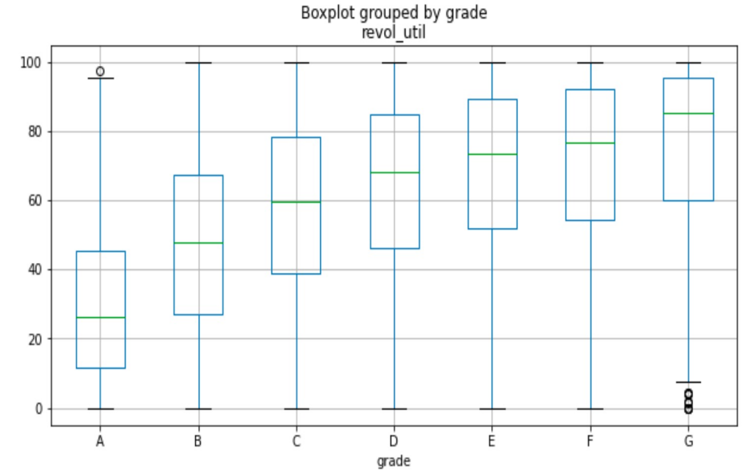
Inferences

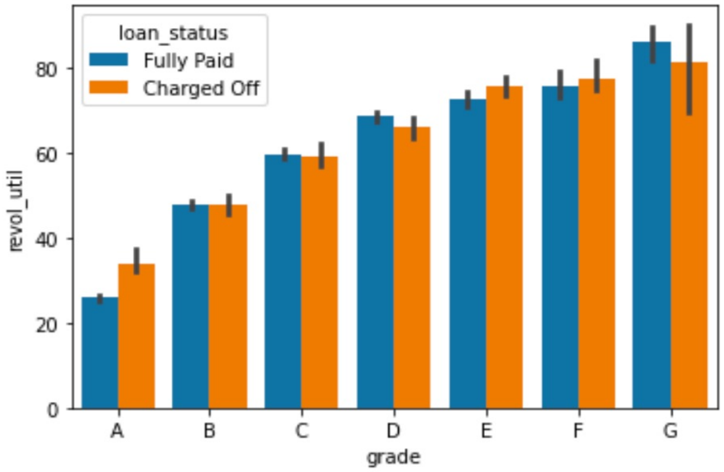
- 36-months term has less interest rates compared to 60-month term interest rates.
- As the grade increases the interest rates also increased.
- Fully paid loans have relatively less interest rate than defaulted claims.
- So, terms, grade and interest rate correlated with Loan status.



Bivariate Analysis - Grade







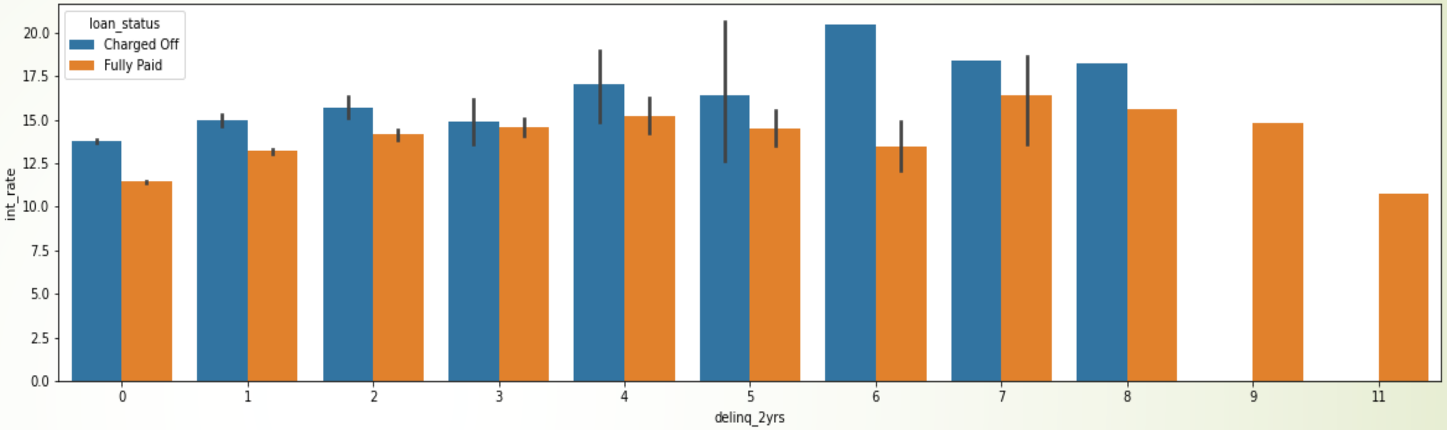
Inferences

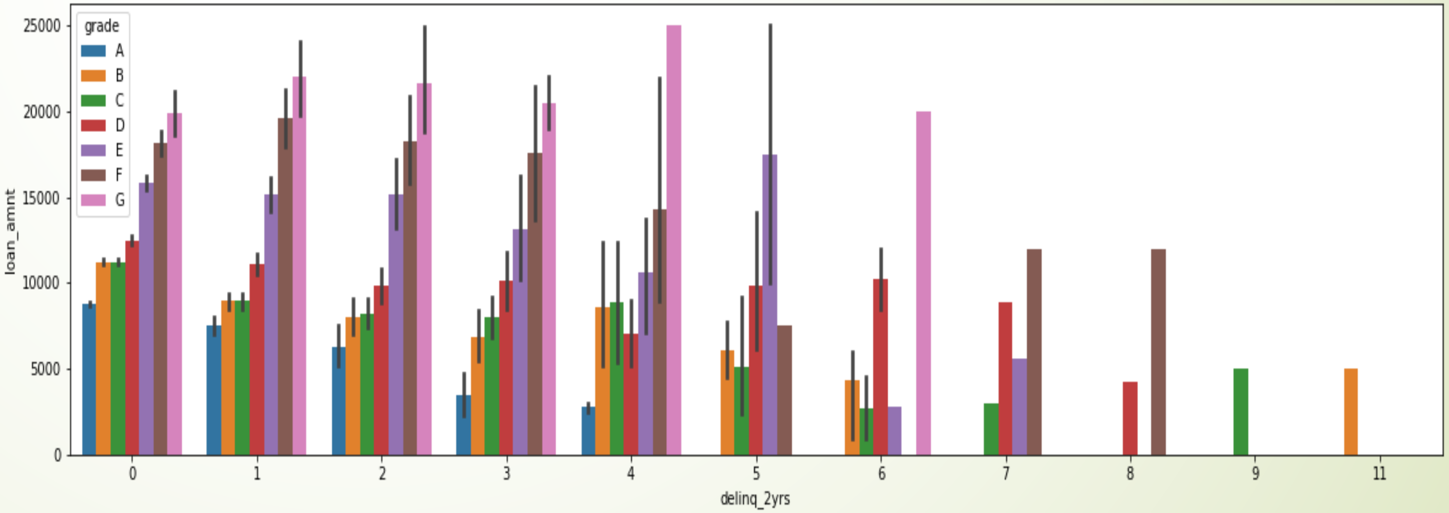
- As the grade increases the loan amount increases for the 36 and 60 months terms.
- As the grade increases the revolving utilization also increases.
- For fully paid claims the revolving utilization increases with grade and for defaulted claims there is a drop in the revolving utilization in higher grade claims.
- Revolving utilisation rate is more for high grades as to take huge amount of loan with high interest there must have been a huge need which is making the person to utilise the loan more than the lower grade loans.
- For lower grades (less interest) there is more utilisation of credit for Defaulted claims as the amount is less than other grades and so the interest rate people are tending more to utilise the loan not thinking of defaulting in the higher grades as the amount and interest increases the tendency to use off more credit is reducing

Bivariate Analysis - delinq_2yrs

Inferences

- It can be observed that the people with more deling_2yr value has more interest rate which is good as the person is assumed to be riskier by in the data it is also evident that high interest rates is having high loan amount as well. The high deling_2yr values have high Charge off percentage which may lead to more losses. Then the company can give low amounts with low interest for the people with more deling_2yr.
- So, we need to avoid giving high grade loans for the people with more deling_2yr value.

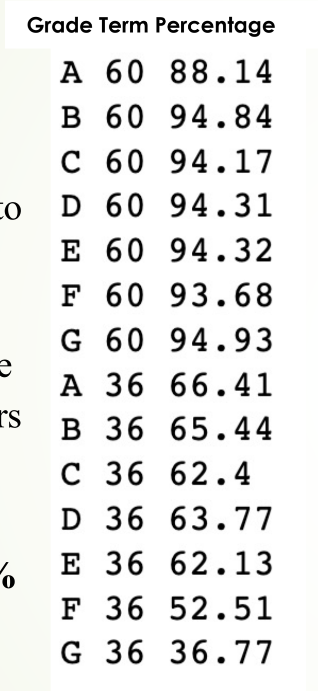




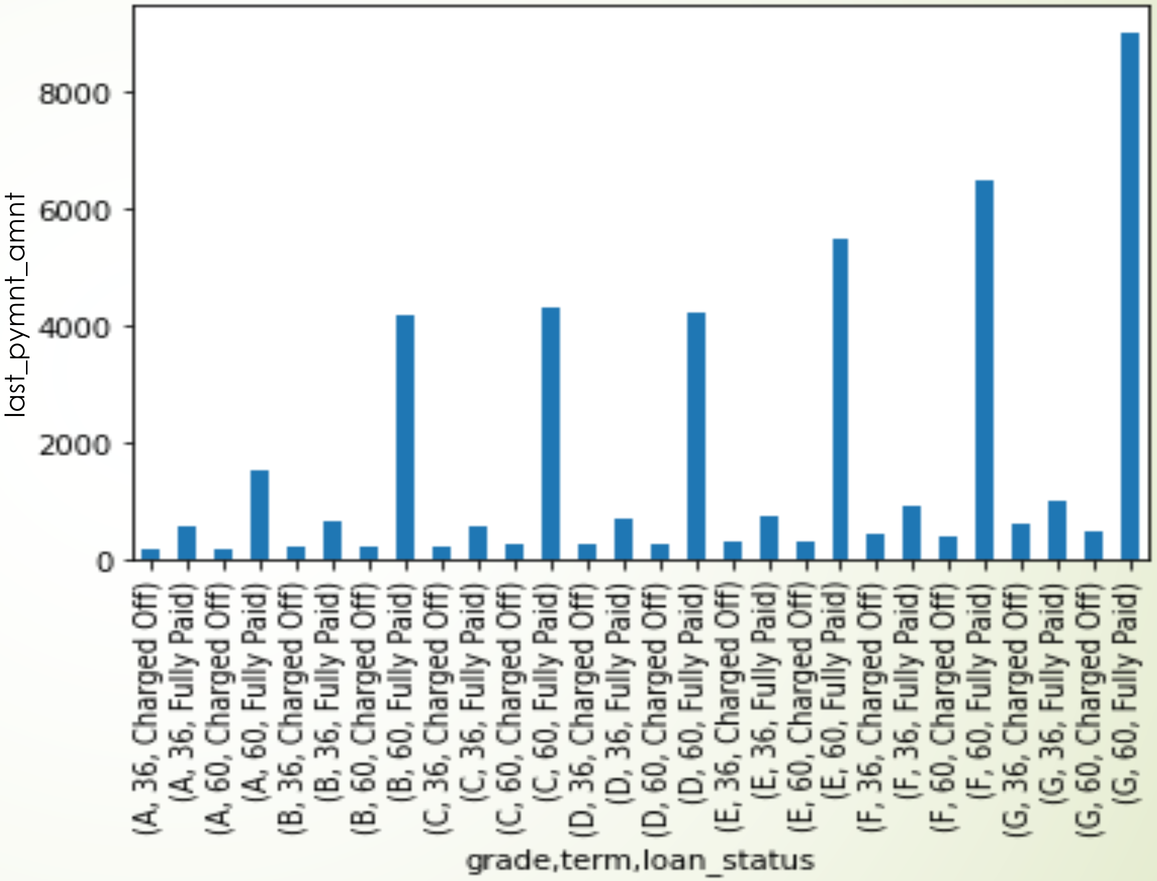
Bivariate Analysis - Last payment amount

Inferences

- When we categorise the people according to grade(as it is related to loan amount)
- The defaulted borrowers of 36 months term are paying on average **62%** less than Fully paid borrowers of same grade and term.
- For the borrowers with 60 months term the defaulters are paying **94%** less last payment than Fully paid borrowers.



Grade	Term	Percentage
A	60	88.14
B	60	94.84
C	60	94.17
D	60	94.31
E	60	94.32
F	60	93.68
G	60	94.93
A	36	66.41
B	36	65.44
C	36	62.4
D	36	63.77
E	36	62.13
F	36	52.51
G	36	36.77



CORRELATION

Inferences

High correlations:

Loan amount - Instalments

open accounts - total accounts

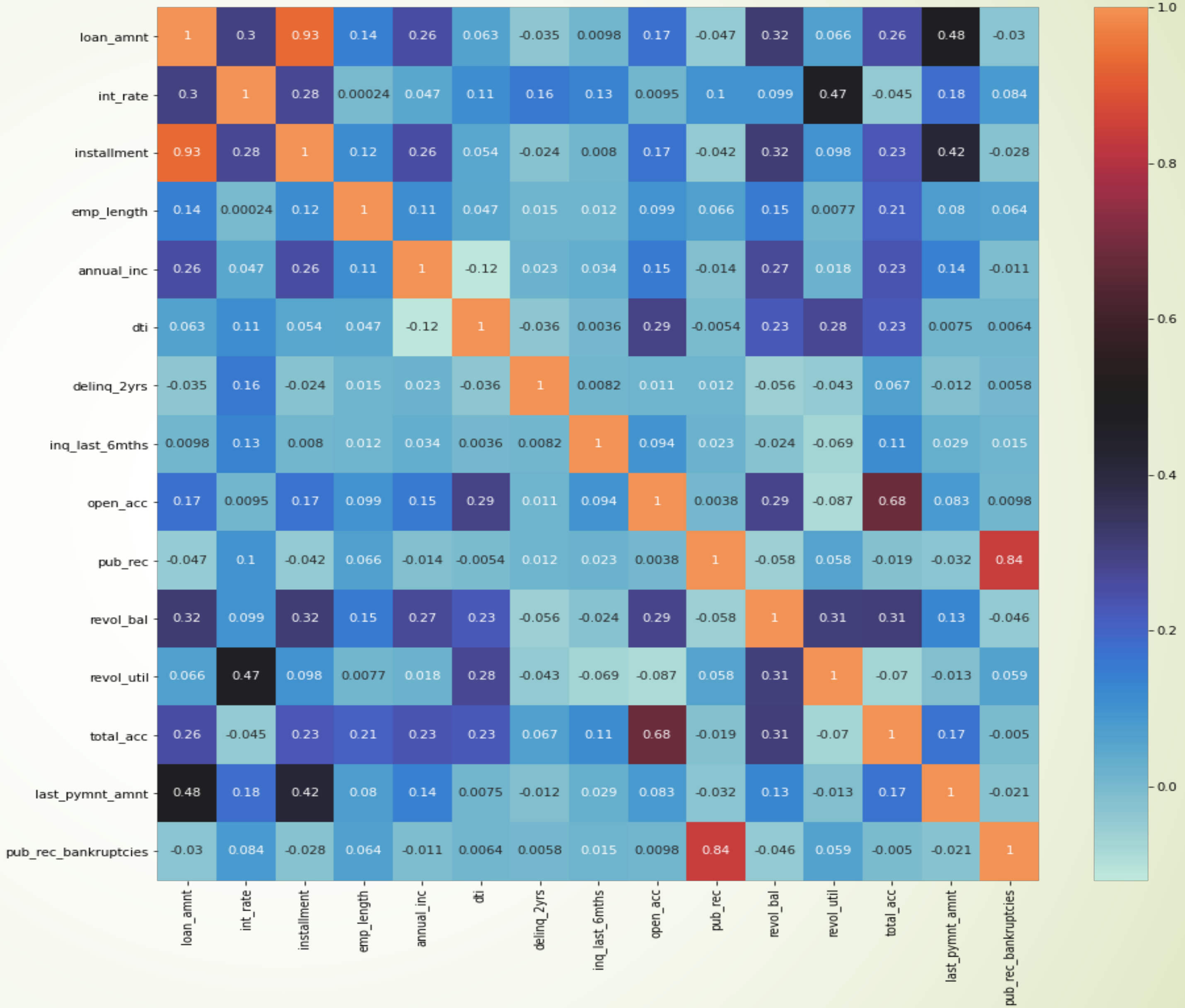
Number of derogatory public records - public record bankruptcies

loan amount - last payment amount

interest amount - Revolving line utilization rate

Instalments - last payment amount

There are no new interpretations from the heat map. All are obvious once.



	loan_amnt	int_rate	installment	emp_length	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	last_pymnt_amnt	pub_rec_bankruptcies
loan_amnt	1	0.3	0.93	0.14	0.26	0.063	-0.035	0.0098	0.17	-0.047	0.32	0.066	0.26	0.48	-0.03
int_rate	0.3	1	0.28	0.00024	0.047	0.11	0.16	0.13	0.0095	0.1	0.099	0.47	-0.045	0.18	0.084
installment	0.93	0.28	1	0.12	0.26	0.054	-0.024	0.008	0.17	-0.042	0.32	0.098	0.23	0.42	-0.028
emp_length	0.14	0.00024	0.12	1	0.11	0.047	0.015	0.012	0.099	0.066	0.15	0.0077	0.21	0.08	0.064
annual_inc	0.26	0.047	0.26	0.11	1	-0.12	0.023	0.034	0.15	-0.014	0.27	0.018	0.23	0.14	-0.011
dti	0.063	0.11	0.054	0.047	-0.12	1	-0.036	0.0036	0.29	-0.0054	0.23	0.28	0.23	0.0075	0.0064
delinq_2yrs	-0.035	0.16	-0.024	0.015	0.023	-0.036	1	0.0082	0.011	0.012	-0.056	-0.043	0.067	-0.012	0.0058
inq_last_6mths	0.0098	0.13	0.008	0.012	0.034	0.0036	0.0082	1	0.094	0.023	-0.024	-0.069	0.11	0.029	0.015
open_acc	0.17	0.0095	0.17	0.099	0.15	0.29	0.011	0.094	1	0.0038	0.29	-0.087	0.68	0.083	0.0098
pub_rec	-0.047	0.1	-0.042	0.066	-0.014	-0.0054	0.012	0.023	0.0038	1	-0.058	0.058	-0.019	-0.032	0.84
revol_bal	0.32	0.099	0.32	0.15	0.27	0.23	-0.056	-0.024	0.29	-0.058	1	0.31	0.31	0.13	-0.046
revol_util	0.066	0.47	0.098	0.0077	0.018	0.28	-0.043	-0.069	-0.087	0.058	0.31	1	-0.07	-0.013	0.059
total_acc	0.26	-0.045	0.23	0.21	0.23	0.23	0.067	0.11	0.68	-0.019	0.31	-0.07	1	0.17	-0.005
last_pymnt_amnt	0.48	0.18	0.42	0.08	0.14	0.0075	-0.012	0.029	0.083	-0.032	0.13	-0.013	0.17	1	-0.021
pub_rec_bankruptcies	-0.03	0.084	-0.028	0.064	-0.011	0.0064	0.0058	0.015	0.0098	0.84	-0.046	0.059	-0.005	-0.021	1

Conclusion

- High grade, interest claims are are likely to get defaulted than low grade and interest loans. So, they are risky to give to bad profile borrowers.
- When there is a borrower who opted for 36-month term and if their payment reduces by 60% then the borrower is like to get defaulted from next month.
- If a borrower who opted for 60-month term’s payment reduces by 94% then the borrower is likely to get defaulted from the next month so necessary precautions can be taken to stop defaulting.
- Its risky to give high grade loans to the borrowers with more than 5 delinquents in past 2 years as he/she is more likely to get defaulted instead can give low grade loans with 36-months terms which are more likely to get fully paid.
- If there is less revolving utilization of loan for high grade loans, then the borrower is more likely to get defaulted.
- For low grade loans if there is high revolving utilization compared to the same group borrower then the borrower is more likely to get defaulted.
- Verification process is not stopping any defaulting of loans so necessary actions are to be taken to strengthen it.
- Borrowers with high percentage of Loan Amount to Annual Income ration are risky.
- The loans sanctioned for the purpose of wedding and major purpose has more fully paid percentage with 90% and sanctioned for Education has nearly 16% default percentage.