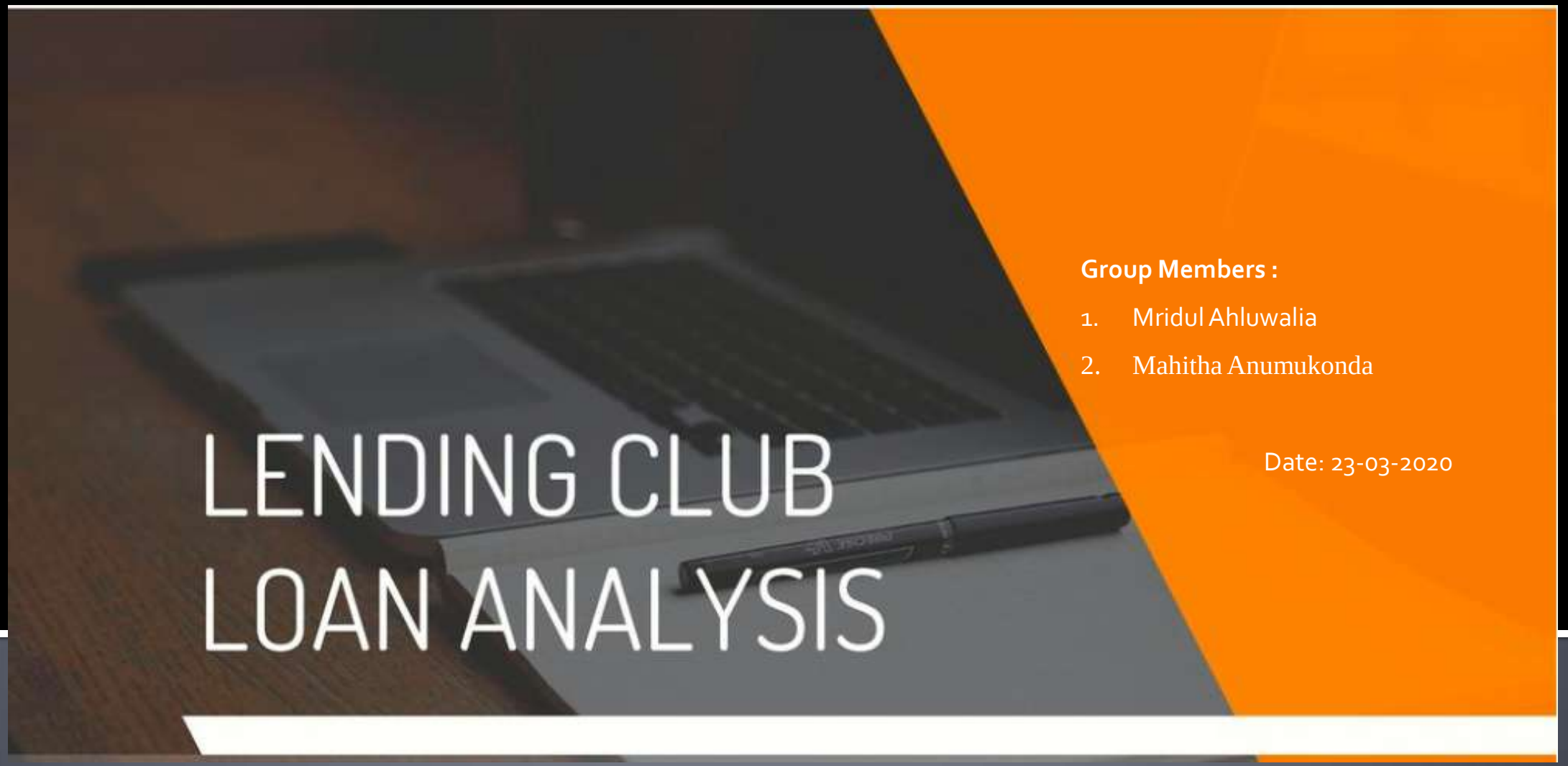


Group Members :

1. Mridul Ahluwalia
2. Mahitha Anumukonda

Date: 23-03-2020

LENDING CLUB ANALYSIS OVERVIEW

Identifying the Business Problem

Business Understanding: Lending Club is an innovative platform that facilitates peer-to-peer loans. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company

If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

- Aim: To **identify** patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, **reducing** the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- Result: To understand **driving factors (or driver variables)** behind loan default i.e. the variables which are strong indicators of default

DATA PREPARATION & PROCESSING

Prior to data analysis, the data was reviewed, cleaned and prepared as follows:

- Removed columns that obviously had no relation to the analysis in question (E.g. Applicant ID, Employee Title etc.)
- Removed columns that had bad quality data (i.e. missing values in observations, unintelligible values etc.)
- Removed columns that had identical relationships to the analysis in question (E.g. funded_amnt and funded_amnt_inv as they are always the same as loan_amt)
- Established derived columns from existing columns to facilitate model analysis (E.g
- Issue_d was converted to issued_year where year is used for our analysis. etc.)
- Converted continuous variables to range of values to enhance interpretation of results (E.g. loan_amt, int_rate, Annual_income, revol_util, etc.)
- Though the data was of good quality, we found that we had to perform some clean-up activity.
- We then removed observations that were missing data in key variables. This, however, was very small percentage of the population.
- Finally, we converted the variable called loan status into a binary variable called "default" for use as our target /dependent variable.

CLASSIFYING VARIABLES FOR OUR ANALYSIS

Borrower dependent variables

- LOAN AMOUNT
- EMPLOYEMENT LENGTH
- HOME OWNERSHIP
- PURPOSE
- ADDRESS STATE
- TERM
- Annual Income

Lending club dependent variables

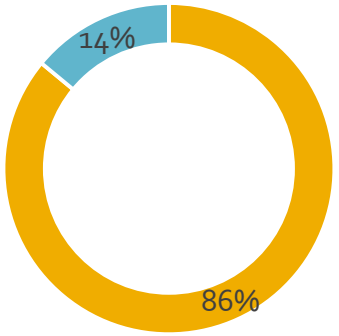
- INTEREST RATE
- GRADE
- SUB GRADE
- VERIFICATION STATUS
- LOAN ISSUED YEAR
- DTI
- PUBLIC REC BANKRUPTCIES
- REVOL UTIL

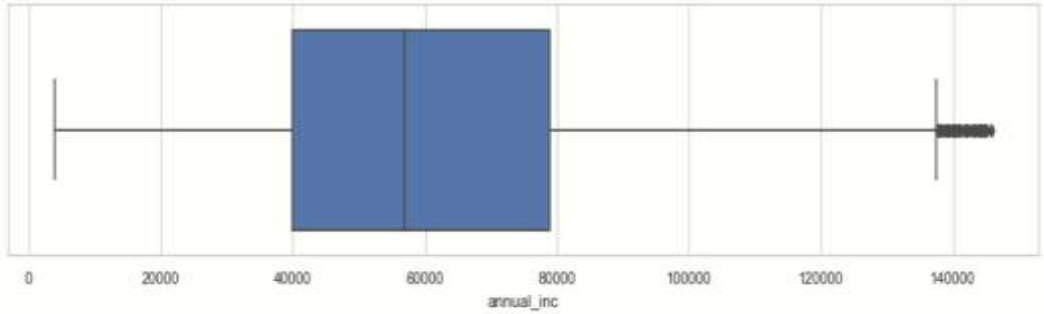
Loan Behavioural variables (Do not aid in our analysis)

acc_now_delinq
application_type
chargeoff_within_12_mths
collection_recovery_fee
collections_12_mths_ex_med
fico_range_high
fico_range_low
funded_amnt
funded_amnt_inv
inq_last_12m
inq_last_6mnths
last_credit_pull_d
last_fico_range_high
last_fico_range_low
last_pymnt_amnt
last_pymnt_d
mths_since_last_delinq
mths_since_last_record
mths_since_last_delinq
mths_since_last_record
next_pymnt_d
open_acc
out_prncp
out_prncp_inv
Recoveries
Revol_bal

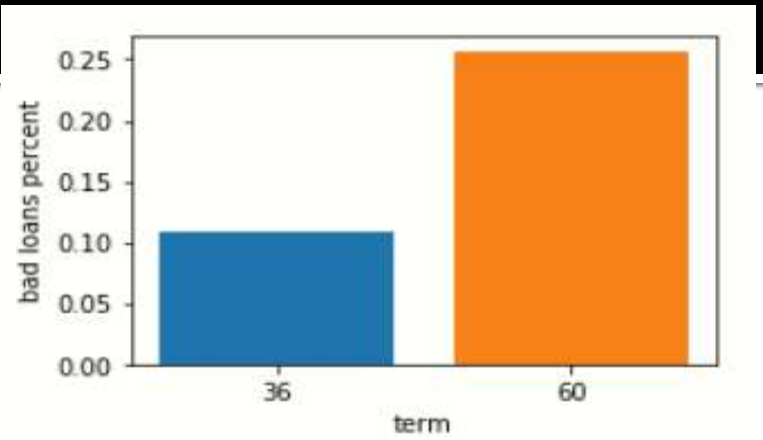
UNIVARIATE ANALYSIS - Borrower Variables against the percentage of defaulters

Target Column - Loan Status

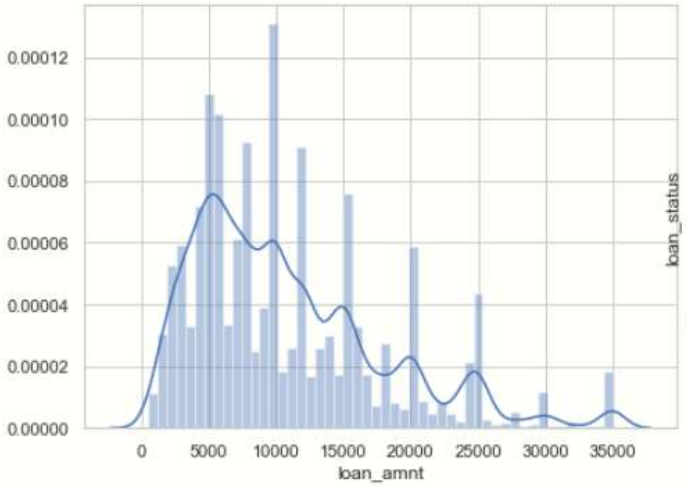


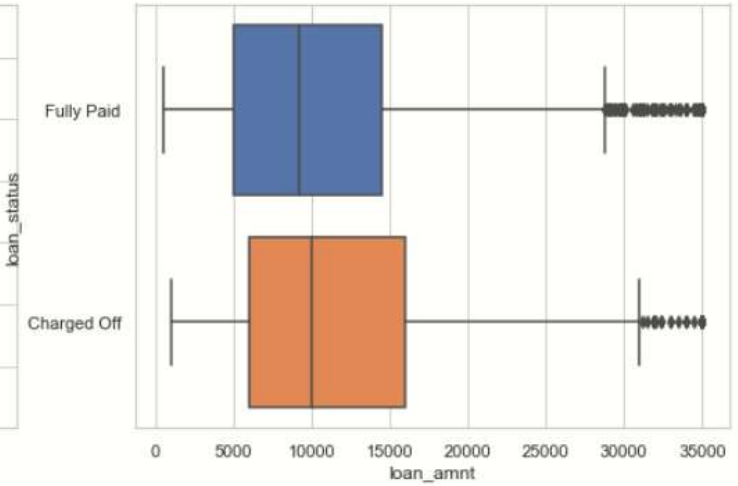


The annual income reprted by the borrowers range from min of 4,000 to max of 6,000,000. Median annual income is around 60,000 Most people have an annual income less than 115000



People who take loans for longer duration are 15% more likely to default.



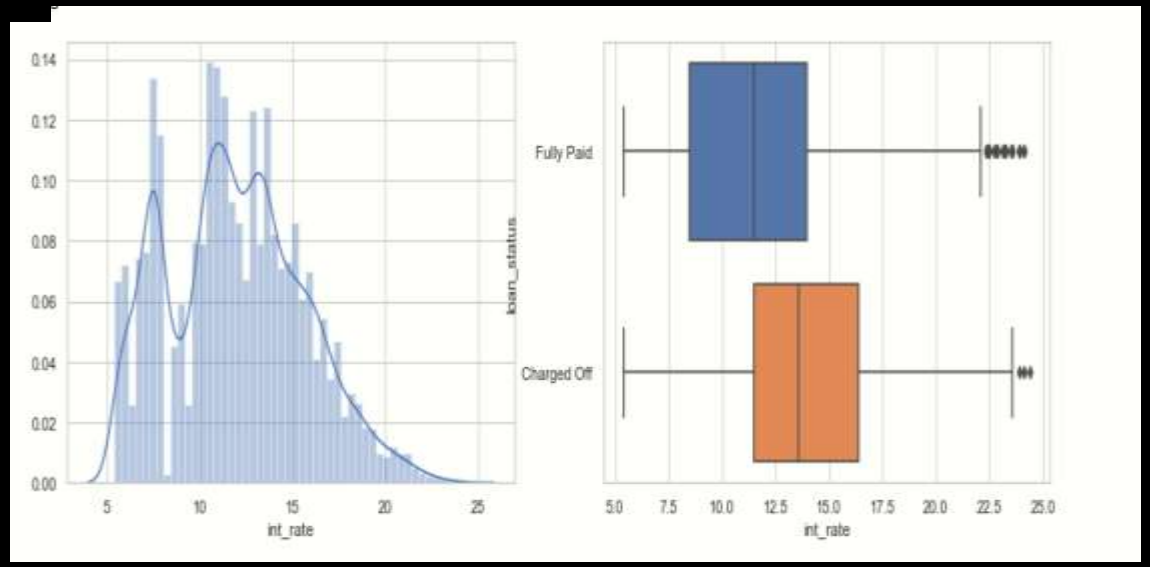


loan_status	count	mean
Charged Off	5016.0	12012.848884
Fully Paid	30017.0	10600.740414

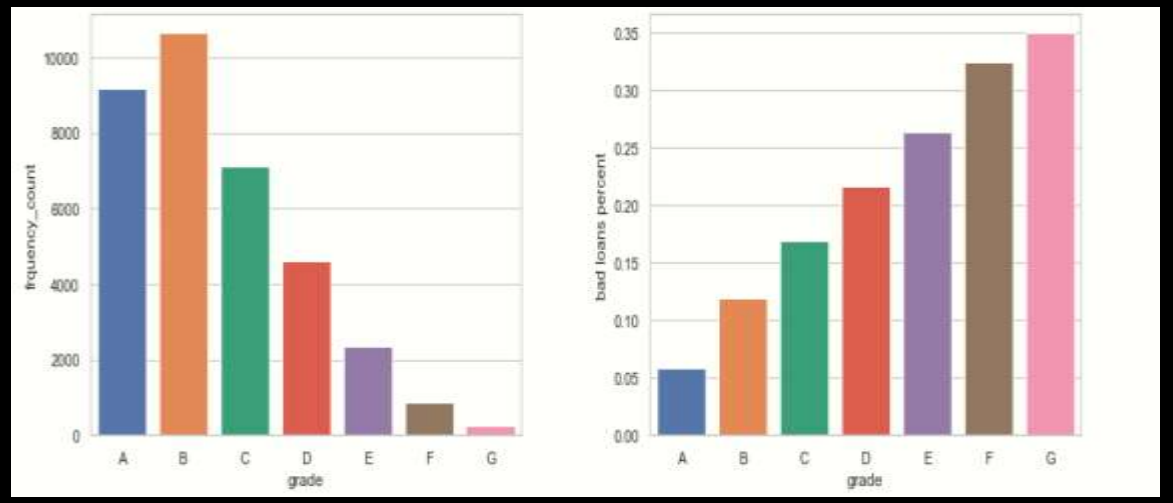
The applied loan amount distribution is slightly right-skewed with mean greater than the median. Most of the loans granted are below 15000 (75 percentile value) Funding amounts see a spike around each 5000 boundary. Charged off loans are shifted towards higher average loan amount request.

UNIVARIATE ANALYSIS

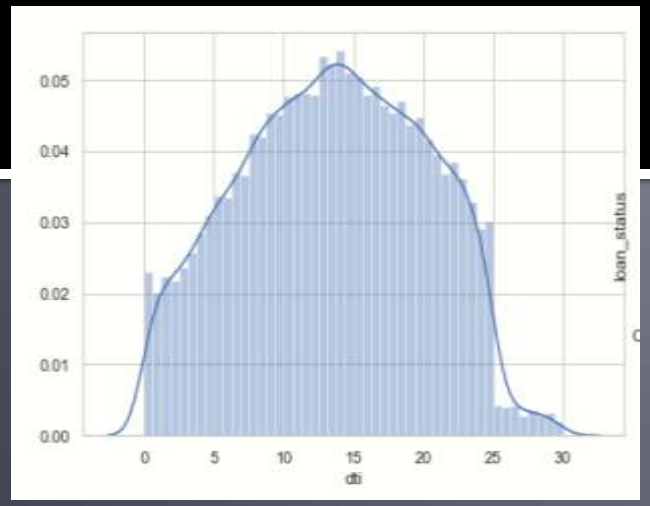
- Loan Variables against the percentage of defaulters



The interest rate for Charged Off loans appear to be higher than for Fully paid. This is naturally expected. As, the risk increases the rate of interest imposed on the loan also increases. Let's analyze this more.

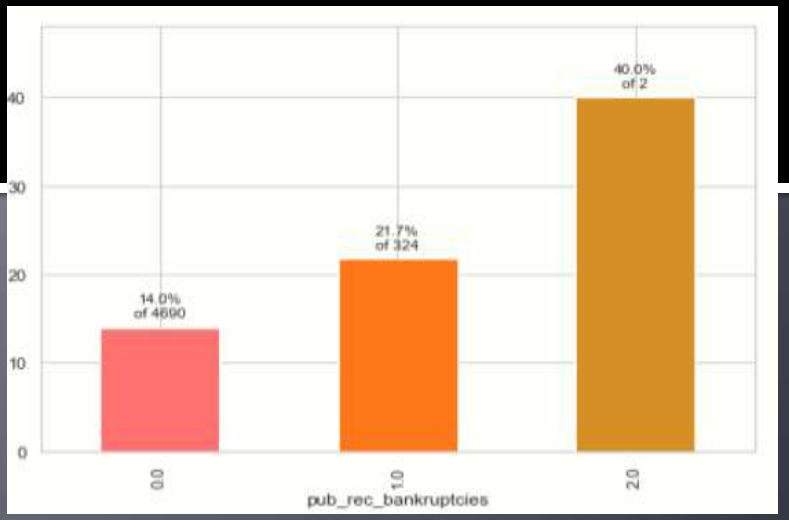


Grade A and B loans are safe. The percentages in full dataset are much higher than percentages in Charged Off loans. Grade D, E, F, G loans are less safe. Lending Clubs grading system is working well.



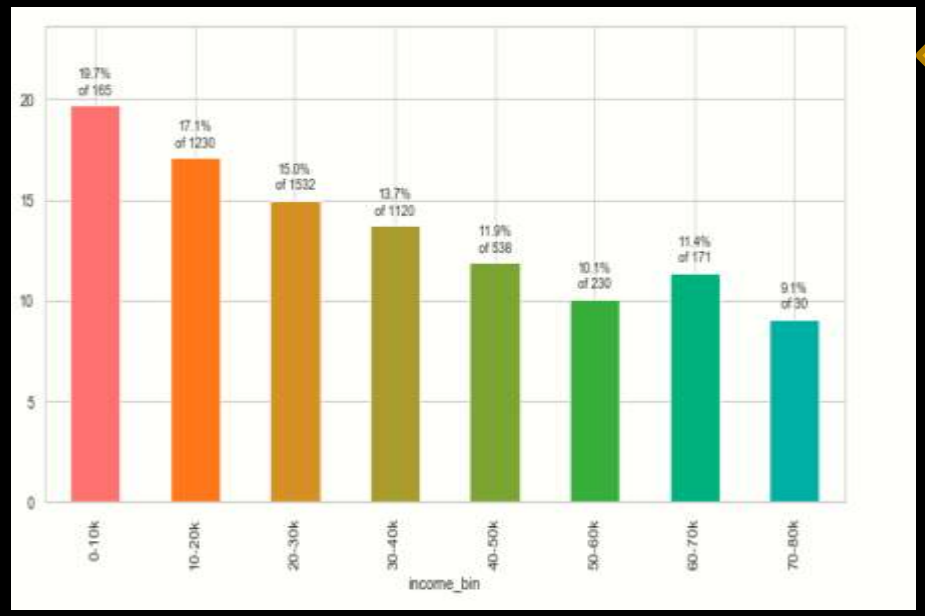

```
count    35033.000000
mean        13.502731
std          6.638062
min          0.000000
25%          8.430000
50%         13.640000
75%         18.740000
max         29.990000
Name: dti, dtype: float64
```

-> People having atleast 2 bankruptcies record are likely to default.
Lower grade people are much likely to default.
More analysis is required on this.



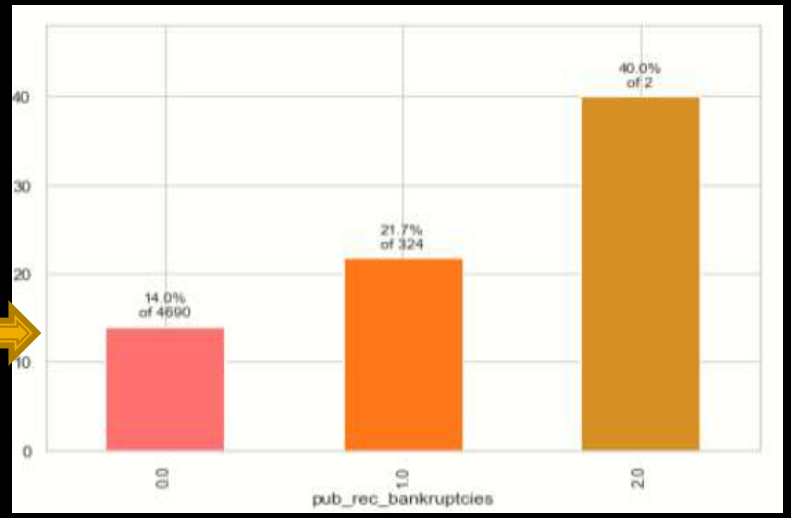
Segmented Univariate Analysis & Derived Metrics

Segmented Univariate Analysis: Performed this on annual income, interest rate, loan amount, public record bankruptcies. Below are such plots for annual income & pub_rec_bankruptcies.



Annual income for various segments is compared against bad loans percentage.

Bankruptcies for 3 segments (0,1,2) is compared against the bad loans percentage.



Derived metrics:

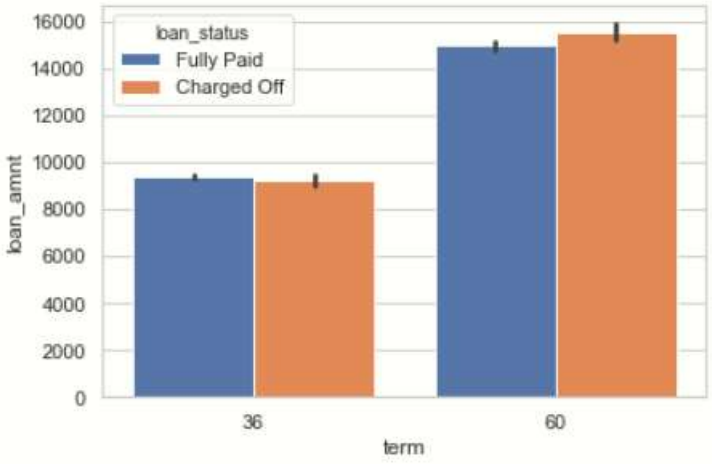
Data Derived Variable - ratio of loan amount to annual income is calculated and As long as loan amount is less than 20% of annual income, defaults are low.

Type Derived Variable - The issue_d & earliest_cr_line columns are considered. These columns are date columns, converted to **datetime** format and year has been extracted for analysis.

Business Derived Variable – The interest rate, dti, loan amount , annual income are considered to be business derived as we have binned them to high, low, very high, medium values (Though we had decided the range here, it usually depends upon the business).

BIVARIATE ANALYSIS

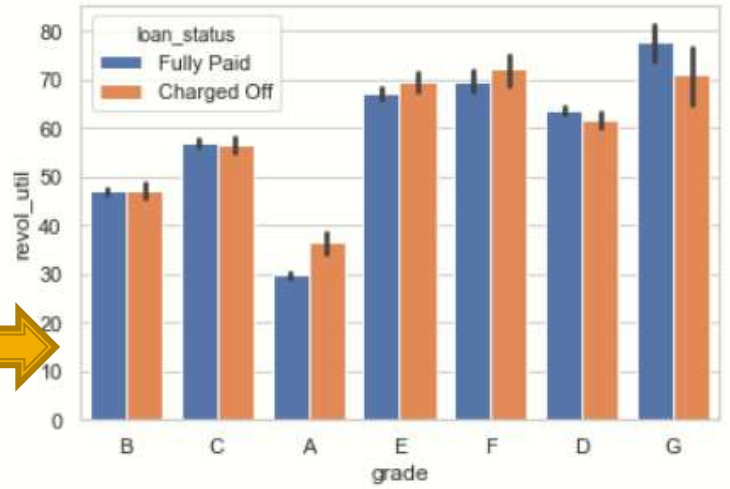
Loan amount vs loan status vs Term



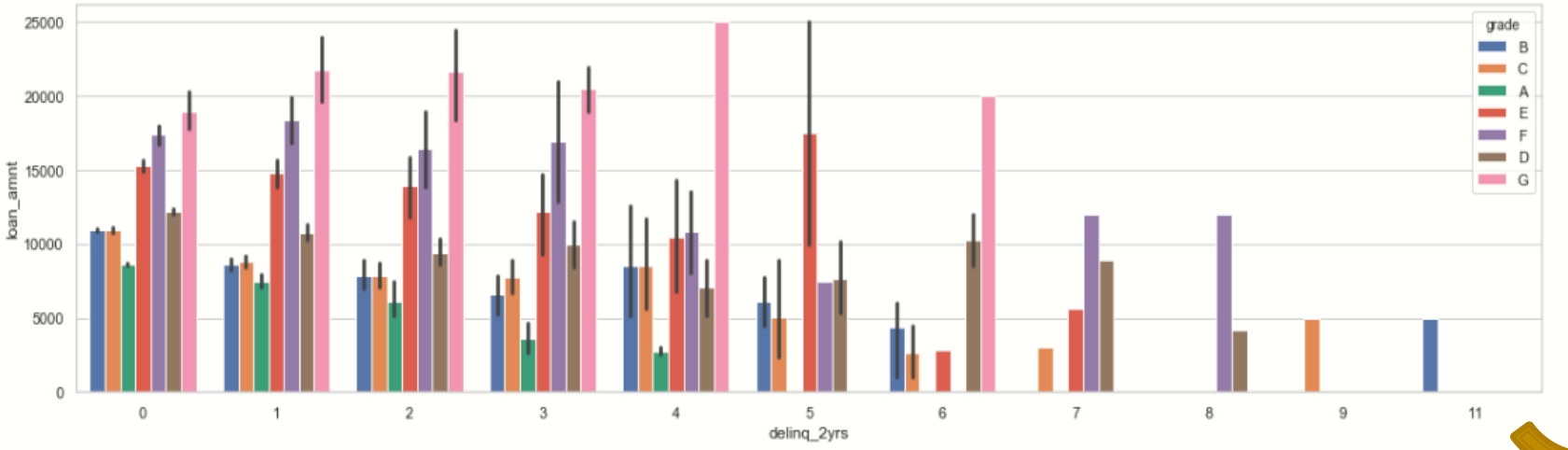
Higher loan amount are associated with longer terms and see higher Charge Offs.

revol_util and grade(and therefore int_rate) are correlated in some way. The revol_util is positivly correlated to the grade. As the grade goes from A to E the revol_util also increases. This may be because higer loan amounts are accosiated with higher grades.

revol_util Vs grade Vs Loan Status



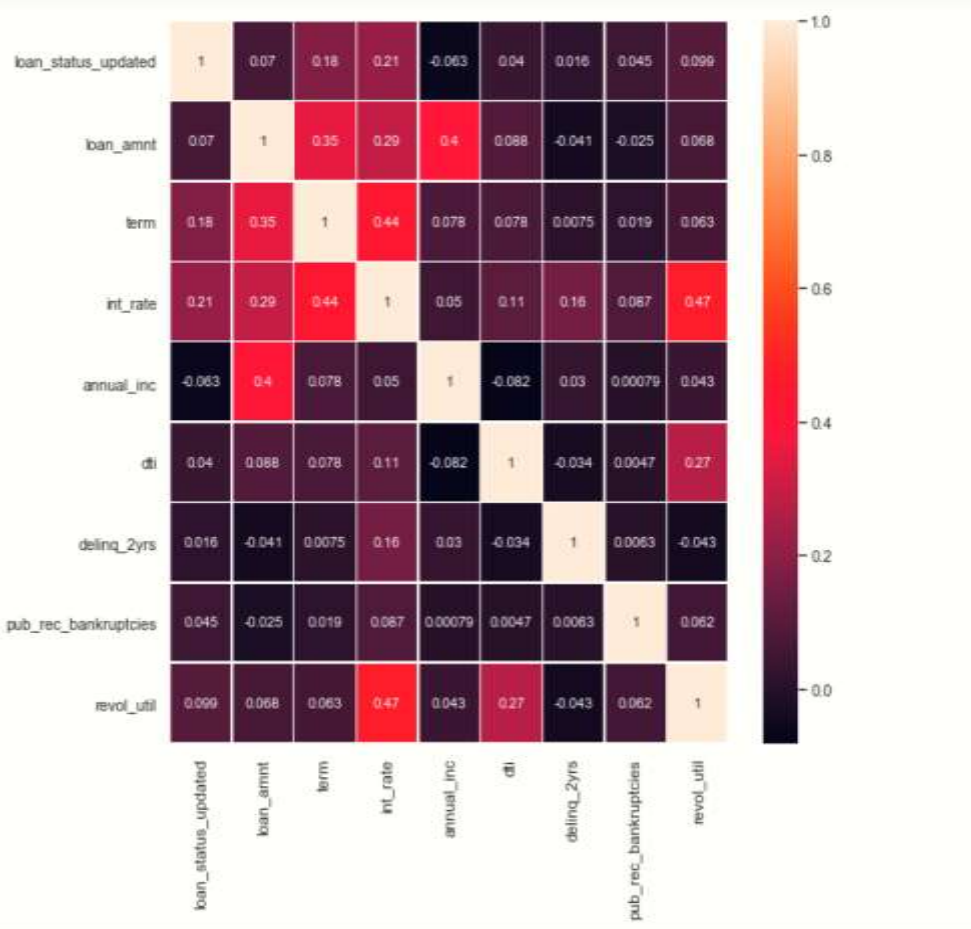
delinq_2yr VS loan amount VS grade



In genral, intrest rate offered inceases with the number of deliquency of the borrower.

CONCLUSION

Correlation factor among variables:



Below are list of factors which are contributing towards default risk

- Grade/Sub Grade (as we move from grade A to G probability of default increases)
- Interest rate (positive correlation with probability of default)
- Term (positive correlation with probability of default)
- Issued year (negative correlation with probability of default)
- Loan Amount (positive correlation with default probability)
- Annual Income (negative correlation with default probability)
- Delinq_2yr (positive correlation with default probability)

Recommendations

- The lower is the grade (i.ie., towards G) the higher is the percentage of defaulters. Hence interest rate to be charged is more for lower grades.
- Choose loan with 36 months period as 60 months term have more percentage of defaulters.
- The higher the loan amount, the higher the likelihood of default Choose loans that $9000 or less.
- Lower the annual income, higher is the dti .So choose dti less than 20%.