LENDING CLUB

CREDIT RISK ANALYTICS

EXPLORATORY DATA ANALYSIS

Presented by

Saikrupa Purushothaman
p.Saikrupa@gmail.com

Our understanding of requirements

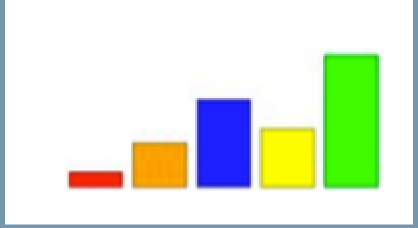
- Consumer Finance Company specializing in lending various types of loans to urban consumers
- Analyse risks in lending to a consumer using the data provided
- Two types of associated risks
  - If the applicant is likely to repay, then not approving the loan will result in the loss to the company
  - If the applicant is not likely to repay (likely to default), approving the loan will lead to financial loss to the company
- Loan data for the past years has been provided in the form of CSV files
- Objective
  - Identify patterns that indicate if an applicant is likely to default
  - Understand driving factors behind loan defaults
- Pattern identification will help the Consumer Finance Company to make decisions such as denying the loan application, lending at interest rate, reducing the loan amount etc

Data

- Past loan data provided for the years 2007-2011
- Data includes only Current (ongoing loans), Fully Paid loans and Charged Off loan details
- Rejected Loan details not provided
- Additional details related to loans and borrower attributes are available for analysis

High Level Approach

Data Cleaning



Univariate Analysis

Bivariate Analysis

Multivariate Analysis

Recommendation

Data Cleaning for Analysis

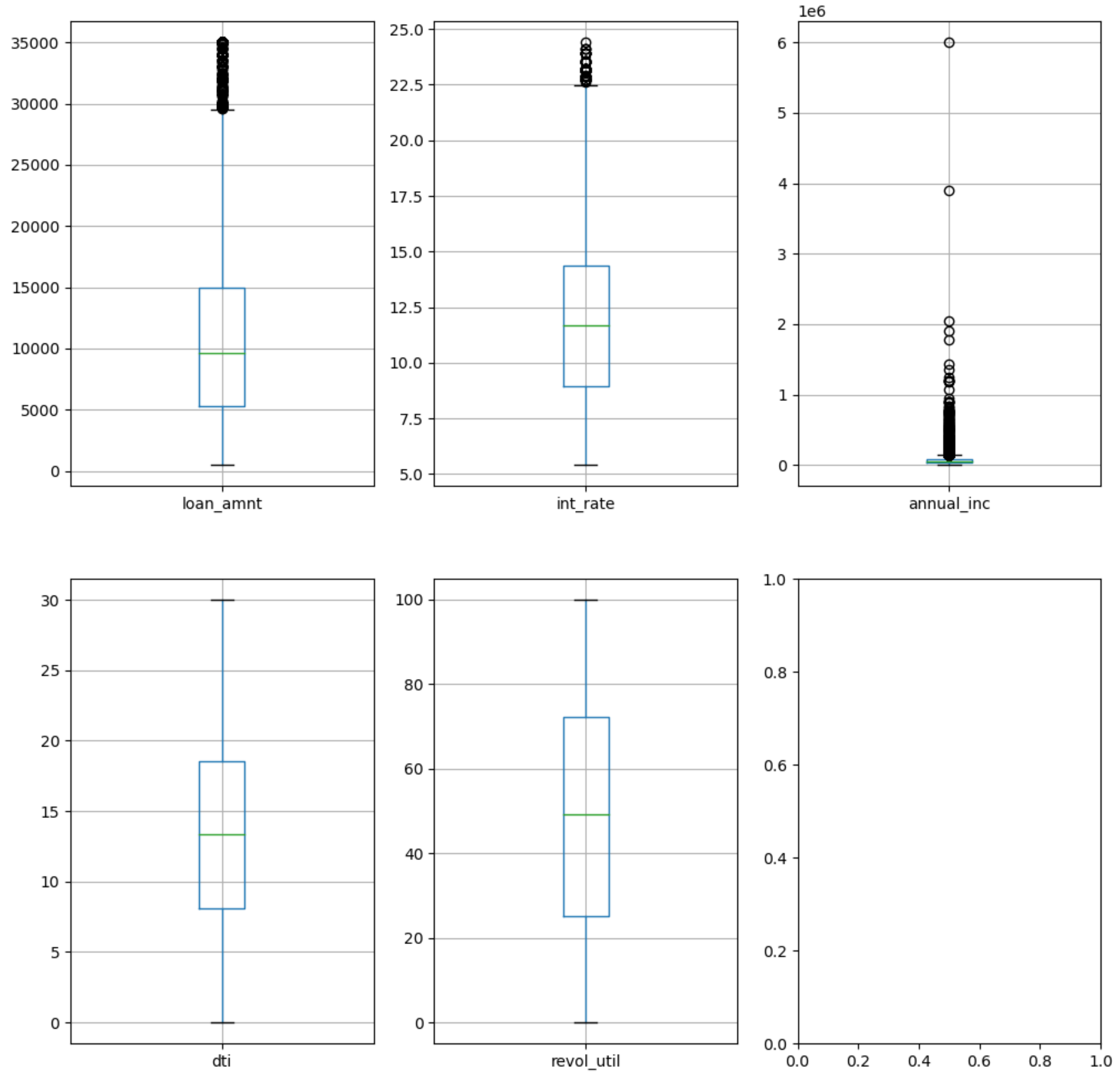
Data Cleaning Steps	Description	Comments
Fix Rows and Columns	Remove columns with NULL values in all rows	These columns are not useful for analysis
	Removing rows related to current loans	Current loans are ongoing and does not help determine patters
	Removing columns related to current loans	These are relevant to current loans and not useful for analysis
	Dropping columns based on cardinality	Columns with high number of unique values may not be used for categorization and pattern seeking
	Dropping other irrelevant columns	A few other columns not directly relevant to the pattern seeking activity has been removed
Handling Missing Values	Identify columns with NaN valurd	
	Handle NaN values for emp_length column	emp_length = NaN could be because the borrowers are students. So, retaining it.
	Handle NaN values for revol_util column	revol_util = NaN is replaced witgh 0
Remove highly correlated columns	Plot a correlation matrix	
	Identify columns that have high correlation	This is done as highly correlated columns will lead to identifying similar patterns. So, removing it to reduce the dataset columns
Standardizing Values	Standardising data types	
	Standardising precision for float data types	
	Removing outliers	Outliers are removed for annual_inc and int_rate columns
Deriving additional information	Creating additional columns for analysis	Additional columns like month and year based on issue_d

Data Cleaning for Analysis

Based on the correlation matrix, the following highly correlated columns were removed with an understanding that they may return similar results in analysis

- funded_amnt, funded_amnt_inv and installment are highly corelated with loan_amnt
- total_pymnt is highly corelated with total_pymnt_inv, loan_amnt
- pub_rec_bankruptcies is highly correlated with pub_rec
- Based on teh above explanation, the following columns will be removed
  - funded_amnt
  - funded_amnt_inv
  - installment
  - total_pymnt
  - total_pymnt_inv
  - pub_rec_bankruptcies

Outlier Removal



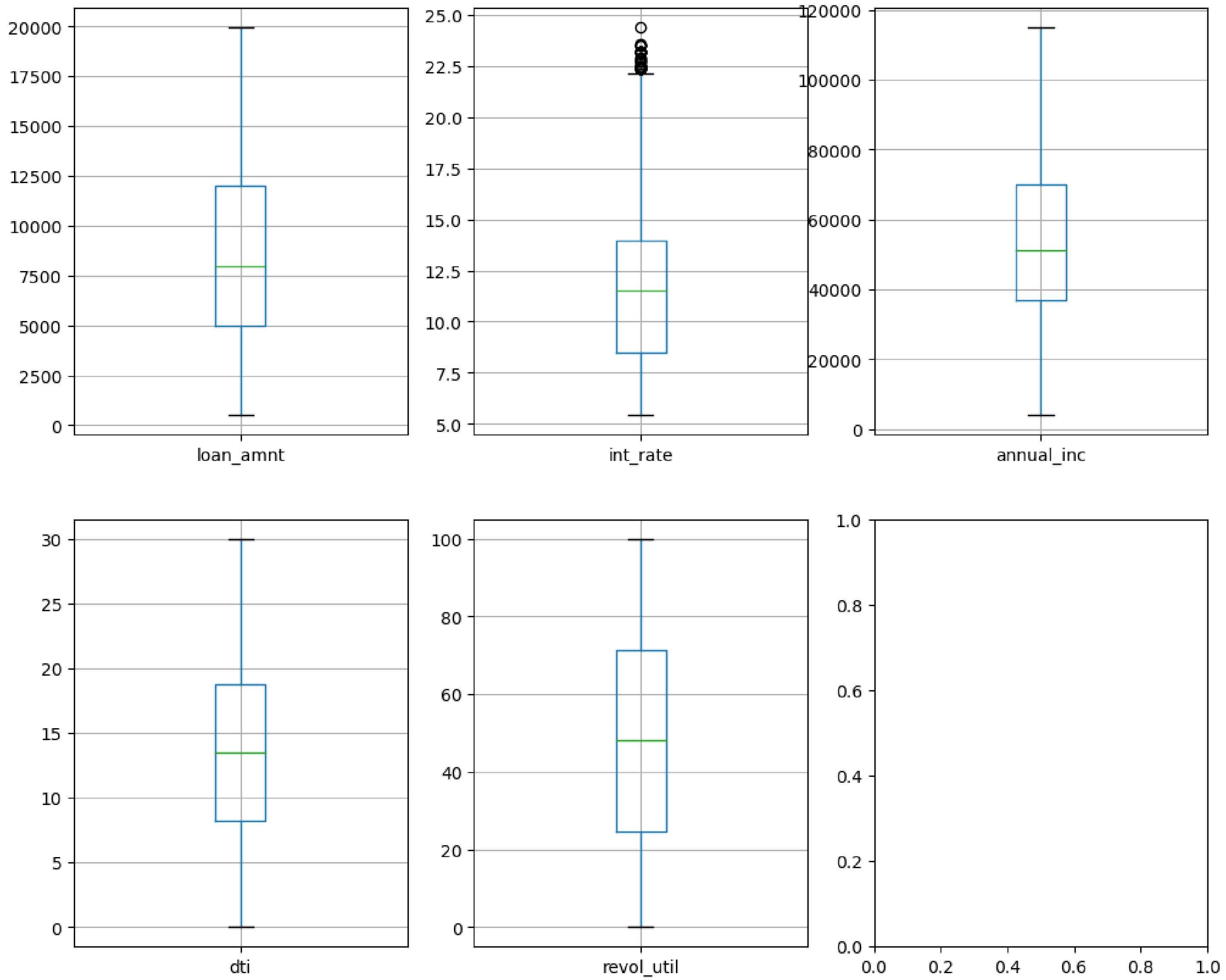
- Outliers can be seen in loan_amnt, int_rate and annual_inc columns.
- Outliers were removed for annual_inc and loan_amnt columns at 90^{th} percentile
- Outliers were not removed for int_rate column to not reduce the original row size significantly

Outlier Removal

Post removal of Outliers

in

- loan_amnt
- annual_inc



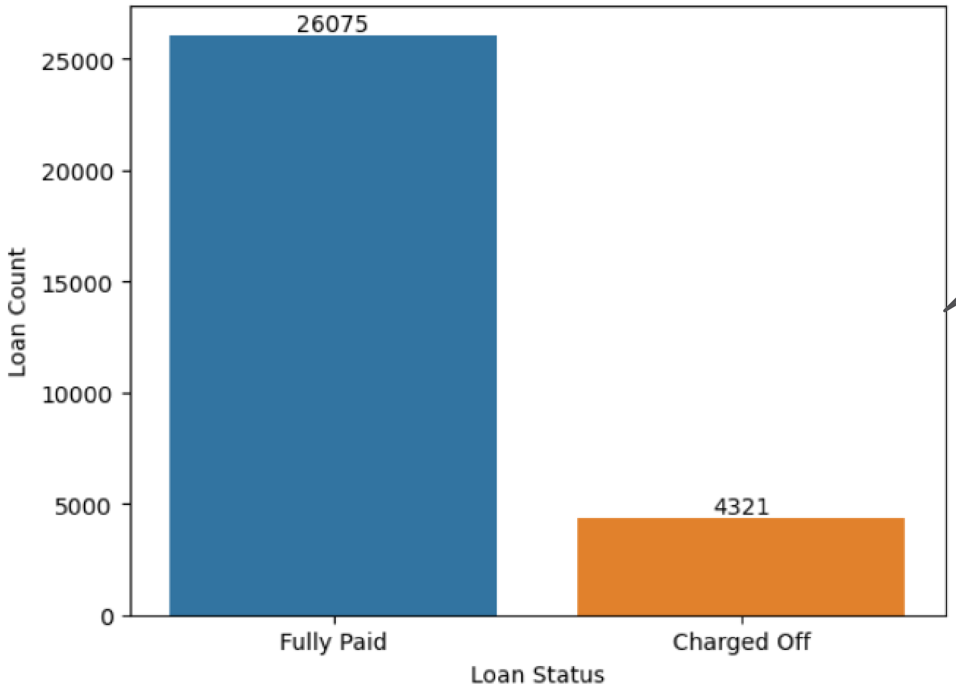
Univariate Analysis

The following was done as part of Univariate Analysis

Understand the split between fully paid and charged off (default) loans

Explore Charged Off (defaulters) data across various category and continuous variables

Univariate Analysis



~ 17% of the overall loans are being Charged Off and are considered as Default loans

Univariate Analysis

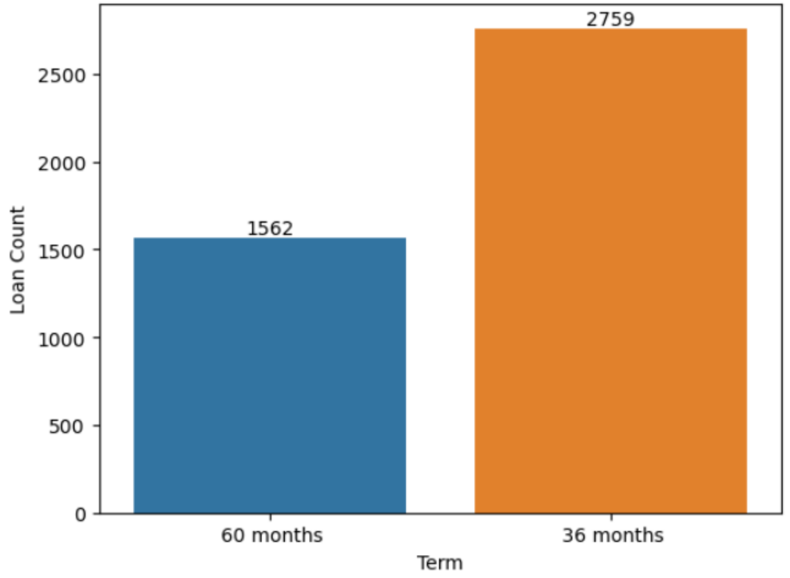
Univariate Analysis

- Considering that the objective of the exercise is to identify patterns in default loan data, the univariate analysis is aimed at the "**Default Loan**" **subset**
- This subset is derived filtering the loan_status column in the dataset to **"Charged Off"**
- Univariate Analysis was done on the following categorical and continuous variables

Unordered Categorical Variables	Ordered Categorical Variables	Continuous Variables
1. term 2. home_ownership 3. purpose 4. verification_status	1. grade 2. emp_length (work experience) 3. ssue_yearmonth	1. loan_amnt 2. int_rate 3. annual_inc 4. dti 5. revol_util

Univariate Analysis – Category Variables

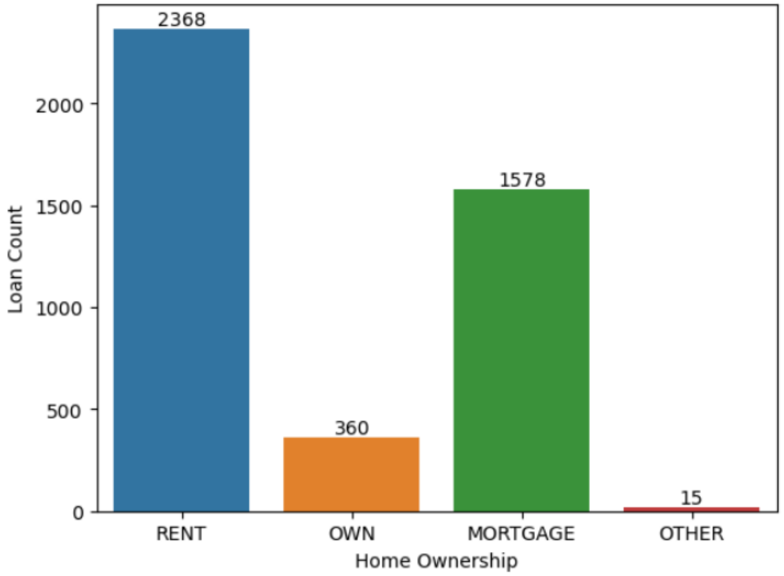
Count & % split by term



36 months	63.85096
60 months	36.14904

Borrowers whose repayment term is shorter (36 months) default more than the ones with longer repayment term.

Count & % split by home_ownership

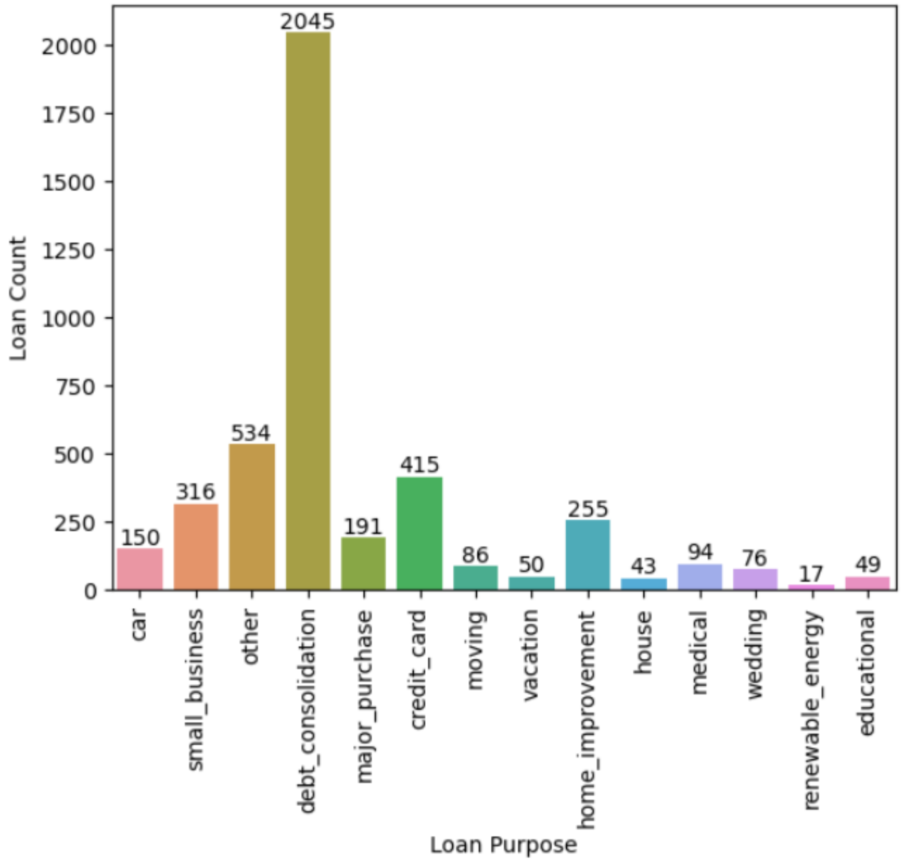


RENT	54.80213
MORTGAGE	36.51932
OWN	8.33140
OTHER	0.34714

Percentage of borrowers who have own home and fail to pay the loan is much lesser than the percentage of borrowers who are either on rent or mortgage

Univariate Analysis – Category Variables

<u>Count & % split by Purpose of the loan</u>

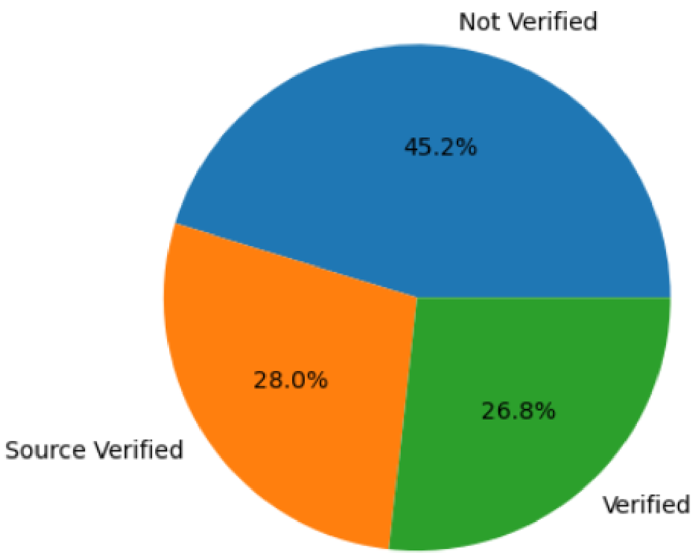


purpose	
debt_consolidation	47.32701
other	12.35825
credit_card	9.60426
small_business	7.31312
home_improvement	5.90141
major_purchase	4.42027
car	3.47142
medical	2.17542
moving	1.99028
wedding	1.75885
vacation	1.15714
educational	1.13400
house	0.99514
renewable_energy	0.39343

About ~47% of the borrowers who borrow for the purpose of consolidating debts have defaulted.

Univariate Analysis – Category Variables

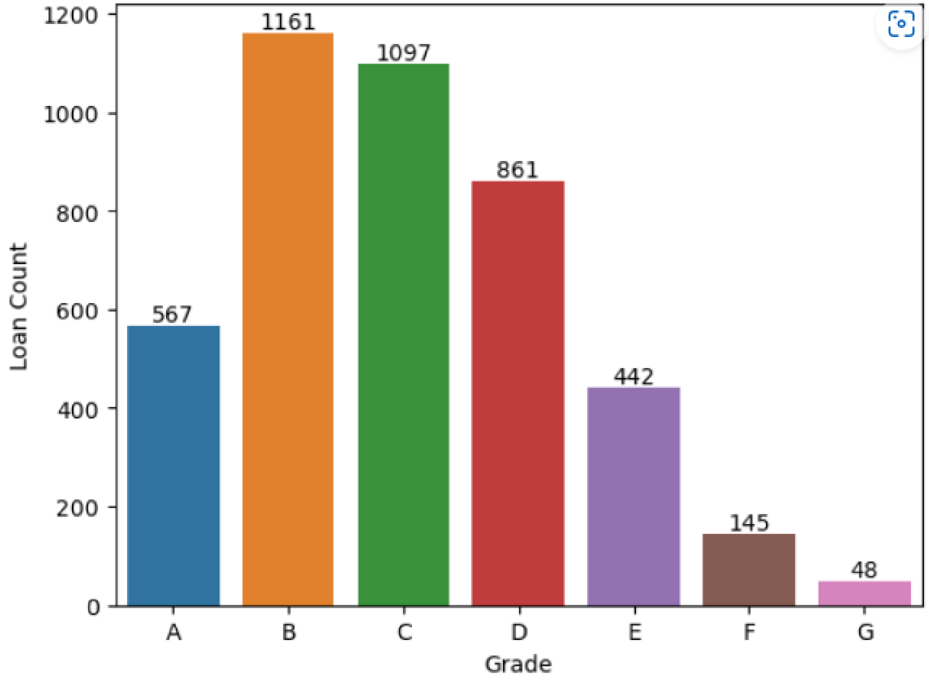
<u>% split by Income Verification Status</u>



~43% of borrowers whose income was not verified had defaulted

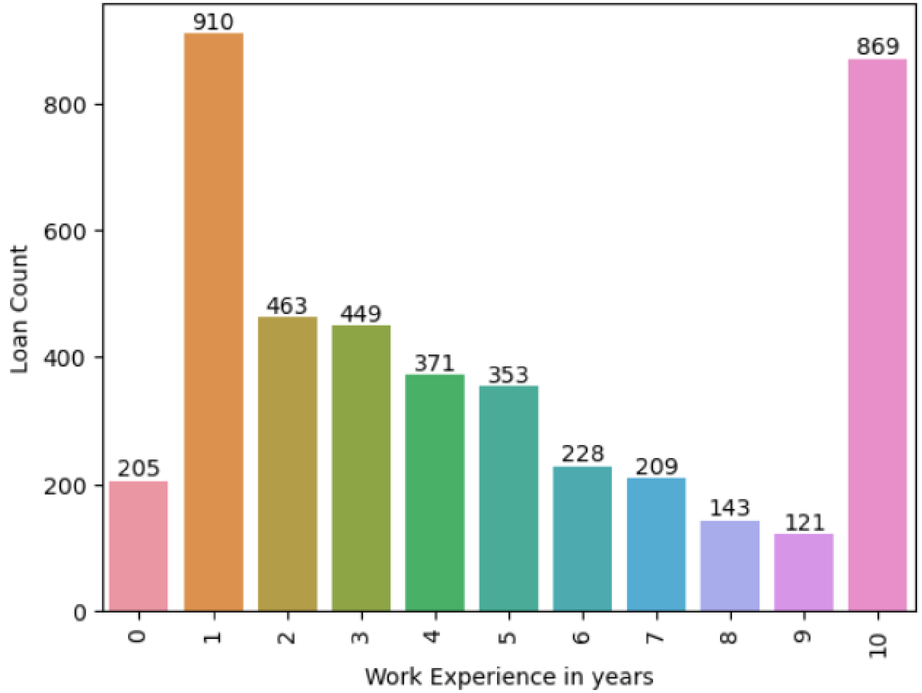
Univariate Analysis – Category Variables

Count by Grade



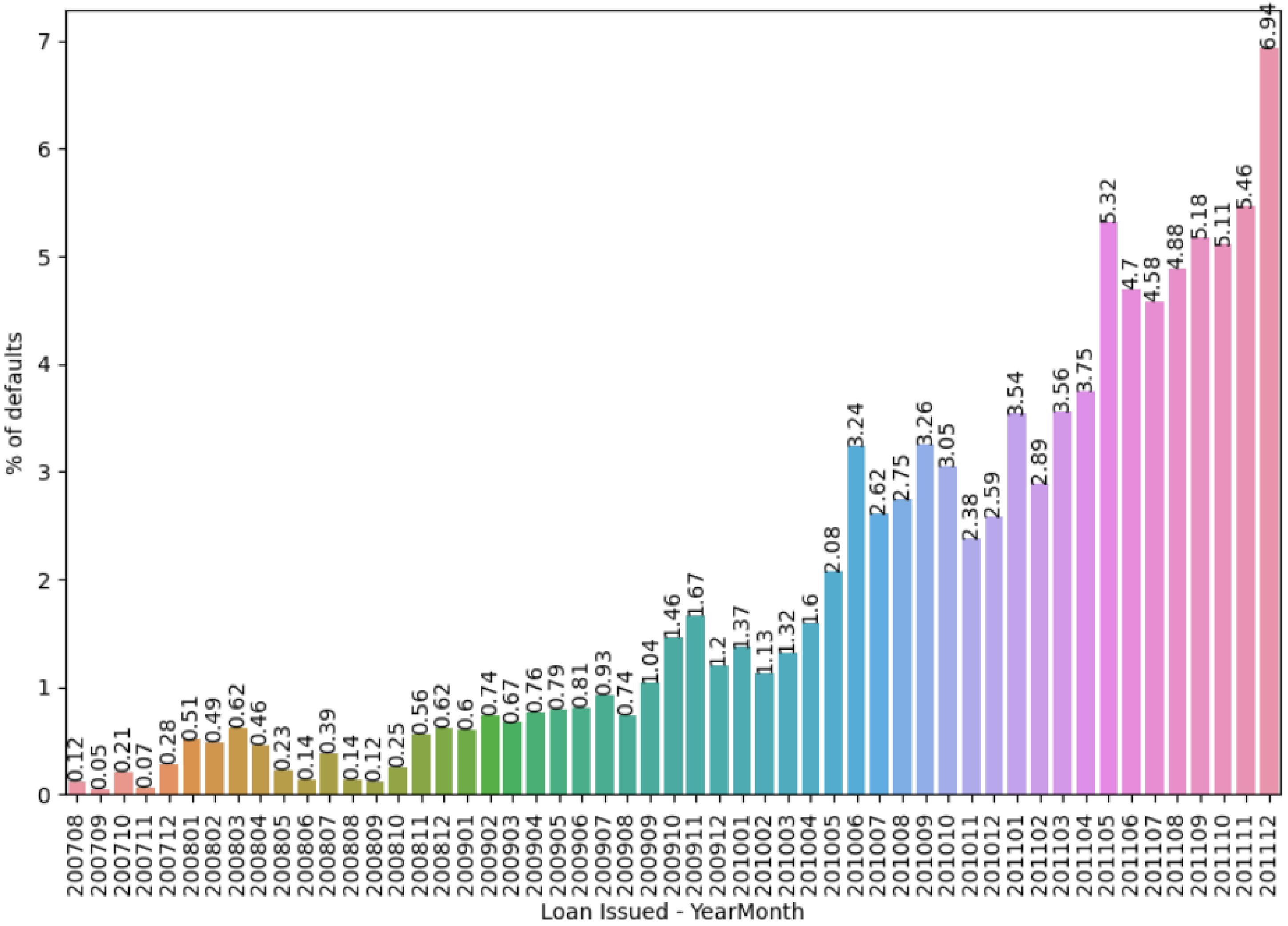
Categories B, C and D are the top categories which has more number of loan defaulters

Count by by work experience years



Borrowers with 0-1 year of experience and > 10 years of experience are the major defaulters

Univariate Analysis – Category Variables



Count by Issue Year Month

The percentage of defaults have steadily increased over time

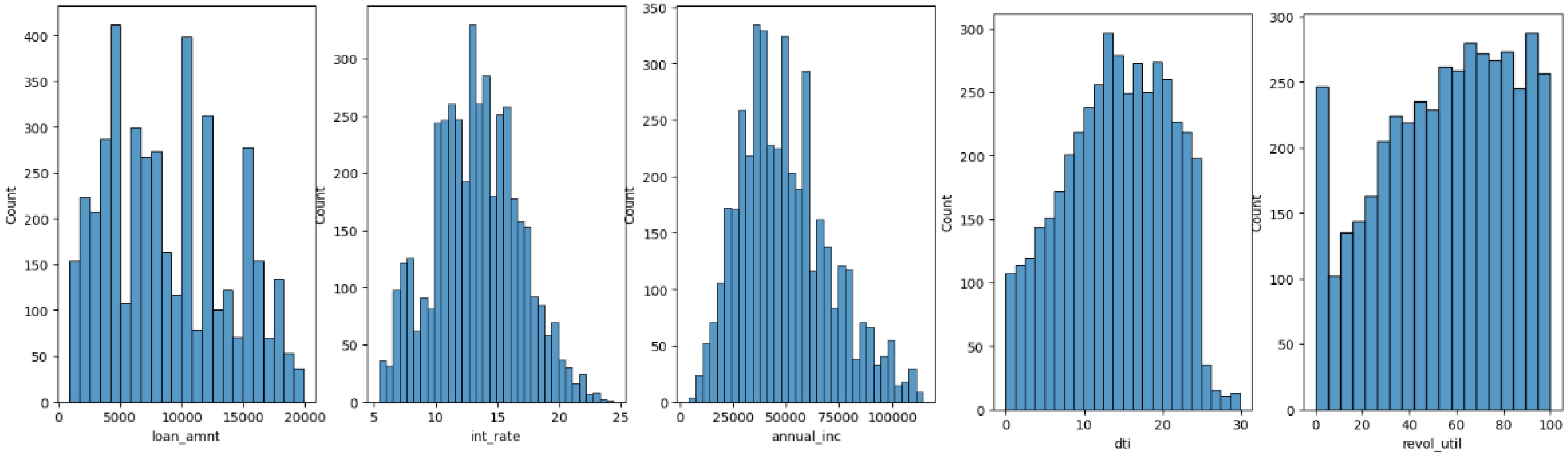
Univariate Analysis – Continuous Variables

Summary Metrics

	loan_amnt	int_rate	emp_length	annual_inc	dti	pub_rec	revol_util	issue_month
count	4321.00000	4321.00000	4321.00000	4321.00000	4321.00000	4321.00000	4306.00000	4321.00000
mean	8825.92571	13.28869	4.67160	49592.74338	14.05289	0.09327	54.96898	7.24416
std	4782.42624	3.48832	3.45490	21298.54879	6.62019	0.29791	28.01859	3.36765
min	900.00000	5.42000	0.00000	4080.00000	0.00000	0.00000	0.00000	1.00000
25%	5000.00000	10.99000	1.00000	34560.00000	9.08000	0.00000	33.40000	5.00000
50%	8000.00000	13.23000	4.00000	46800.00000	14.33000	0.00000	57.70000	8.00000
75%	12000.00000	15.65000	8.00000	62000.00000	19.46000	0.00000	78.50000	10.00000
max	19900.00000	24.40000	10.00000	114600.00000	29.85000	2.00000	99.90000	12.00000

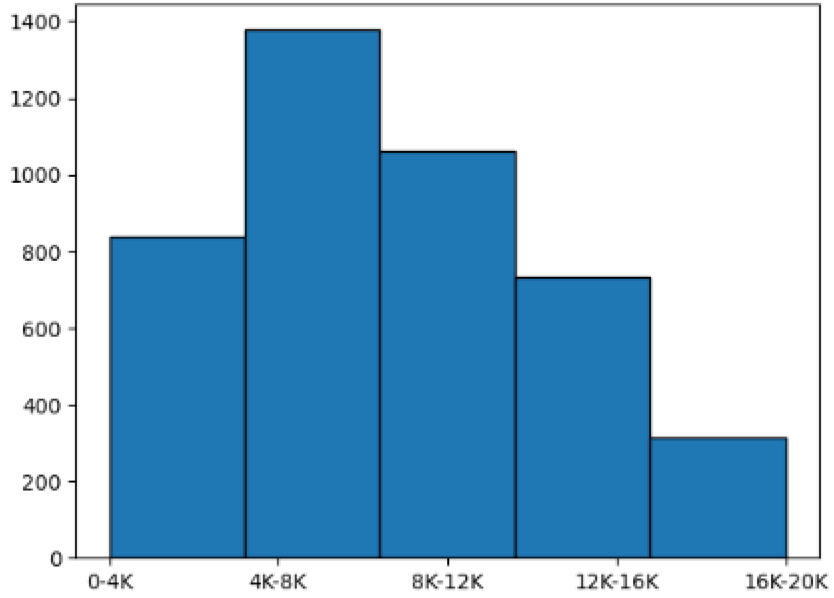
Univariate Analysis – Continuous Variables

<u>Understanding the distribution of continuous variables</u>



Segmented Univariate Analysis

Analyse Loan Amount by binning



Most defaults are in the range of 4K-8K and 8K-12K loan amounts

Analyse Debt to Income ratio by binning

```
11-15     1086
16-20     1044
21-25      893
6-10       748
0-5        460
25-30       63
Name: bin_dti,
```

The Debt to Income ratio of 11-20 has the highest number of defaults

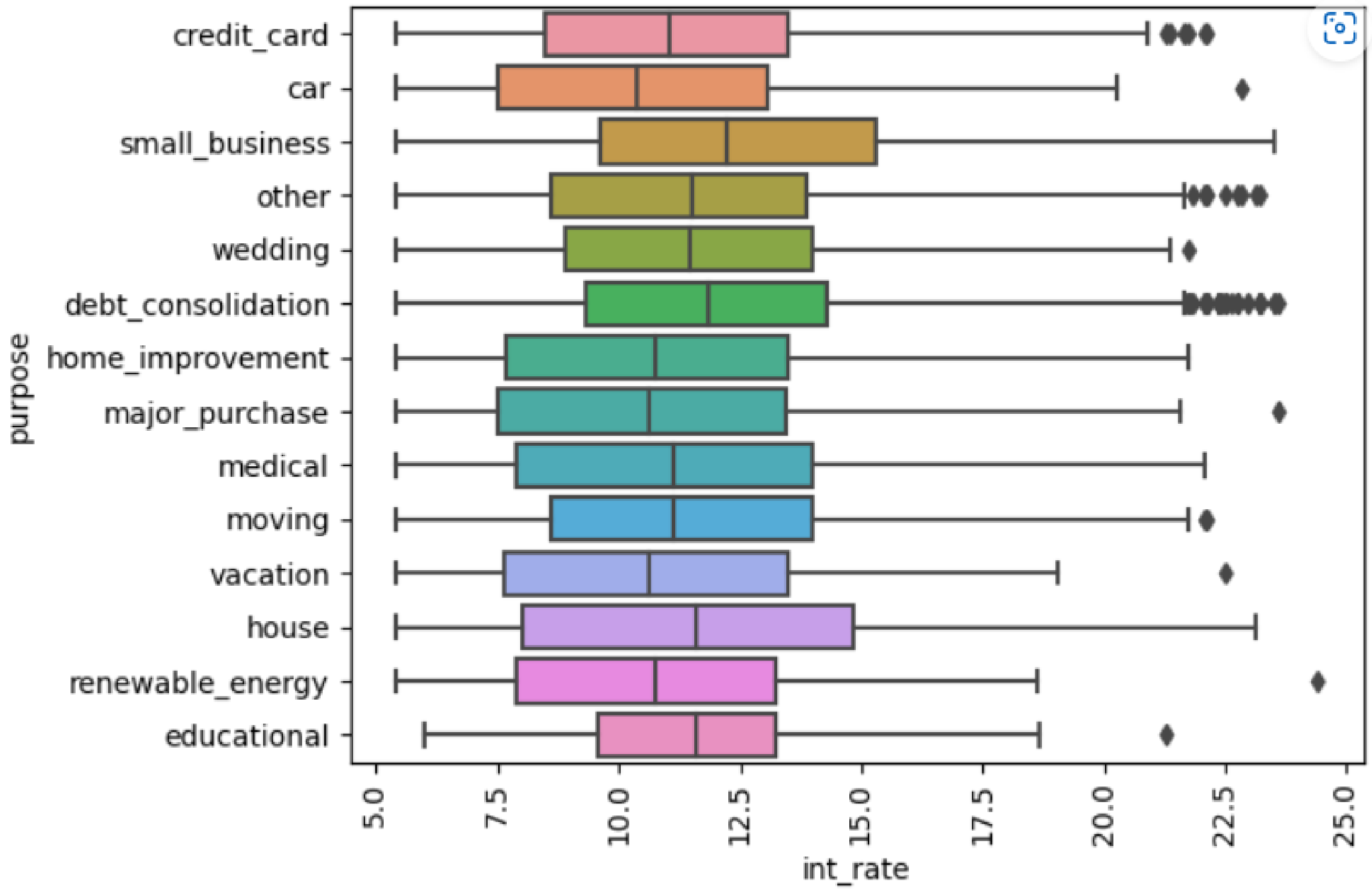
Bivariate Analysis

Bivariate Analysis

- Bivariate Analysis was done on the overall data to understand how 2 or more variables compare
- Following analysis were done
  - Analyse how interest rates vary by purpose
  - Analyse the percentage of default loan amount across various categorical variables liks grade, home ownership etc
  - Analyse how Debt to Income indicator compare to default loan amounts

Bivariate Analysis

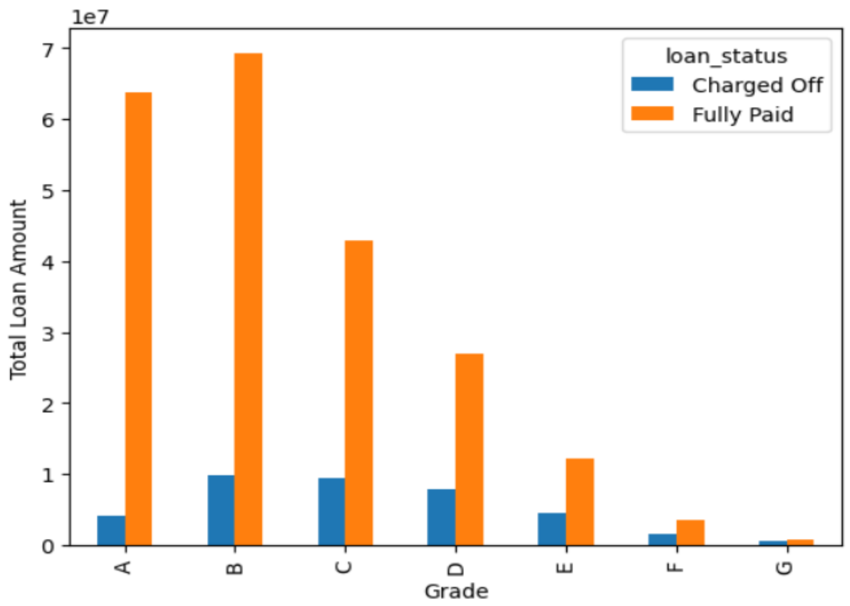
Analysing Purpose and Interest Rate using summary metrics



Interest rates are higher for small business, house, debt_consolidation

Bivariate Analysis

Analyse charged off loan amount as a propotion of total loan amount by Grade

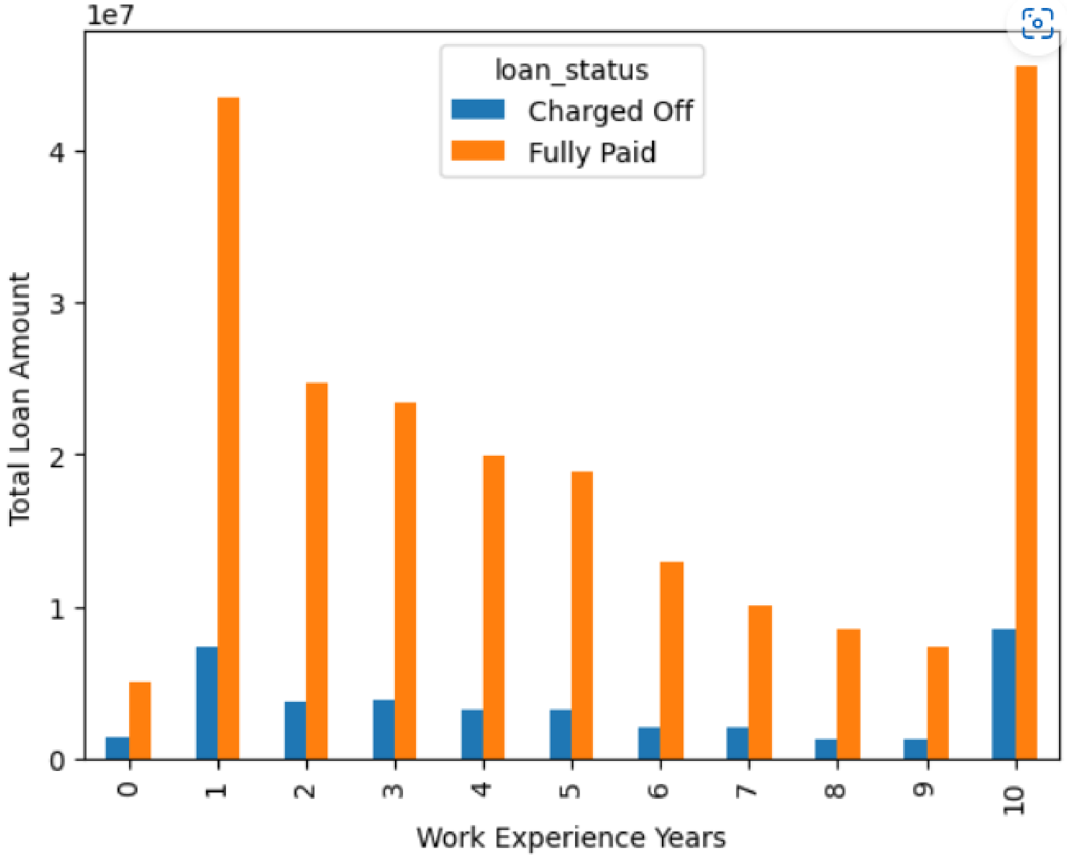


loan_status grade	Charged Off	Fully Paid	chargedoff_percent
A	4135500	63745850	6.09000
B	9905950	69293975	12.51000
C	9361250	42924925	17.90000
D	7940500	27058950	22.69000
E	4604025	12205375	27.39000
F	1656400	3622175	31.38000
G	533200	698925	43.27000

Grades G, F and E have the highest percentage of Charged Off loans

Bivariate Analysis

Analyse charged off loan amount as a propotion of total loan amount by Work Experience

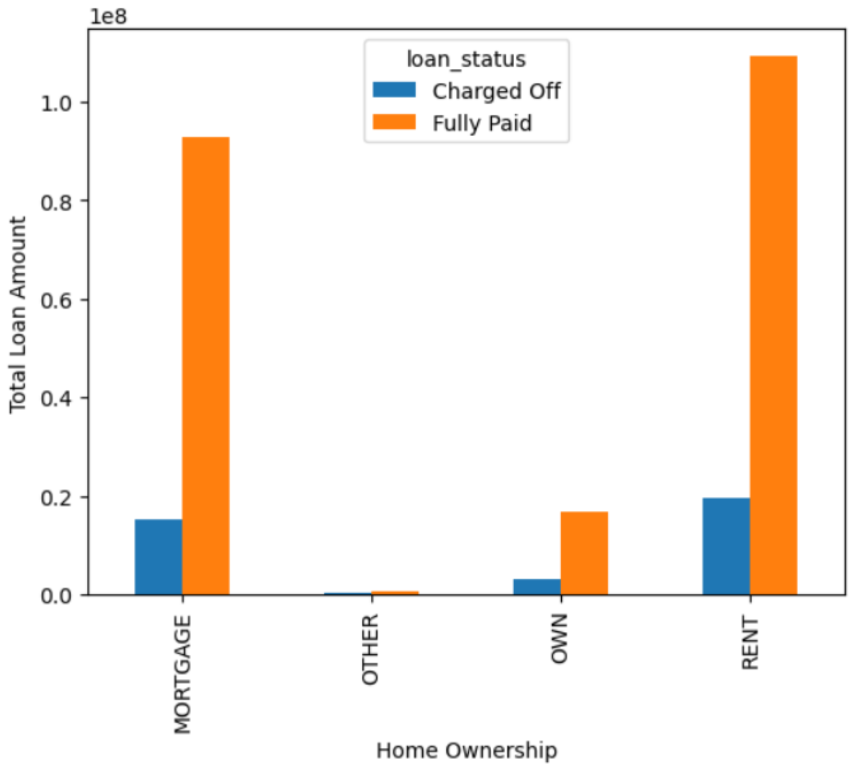


emp_length	Charged Off	Fully Paid	chargedoff_percent
0	1436925	4988025	22.36000
1	7355975	43441275	14.48000
2	3783075	24636325	13.31000
3	3904650	23337150	14.33000
4	3200425	19910575	13.85000
5	3226650	18814200	14.64000
6	2108125	12911500	14.04000
7	2075200	10122150	17.01000
8	1333025	8511000	13.54000
9	1245425	7362450	14.47000
10	8467350	45515525	15.69000

Higher percentage of defaulters are in the work experience range of 0-1 years or 10+ years

Bivariate Analysis

Analyse charged off loan amount as a propotion of total loan amount by Home Ownership

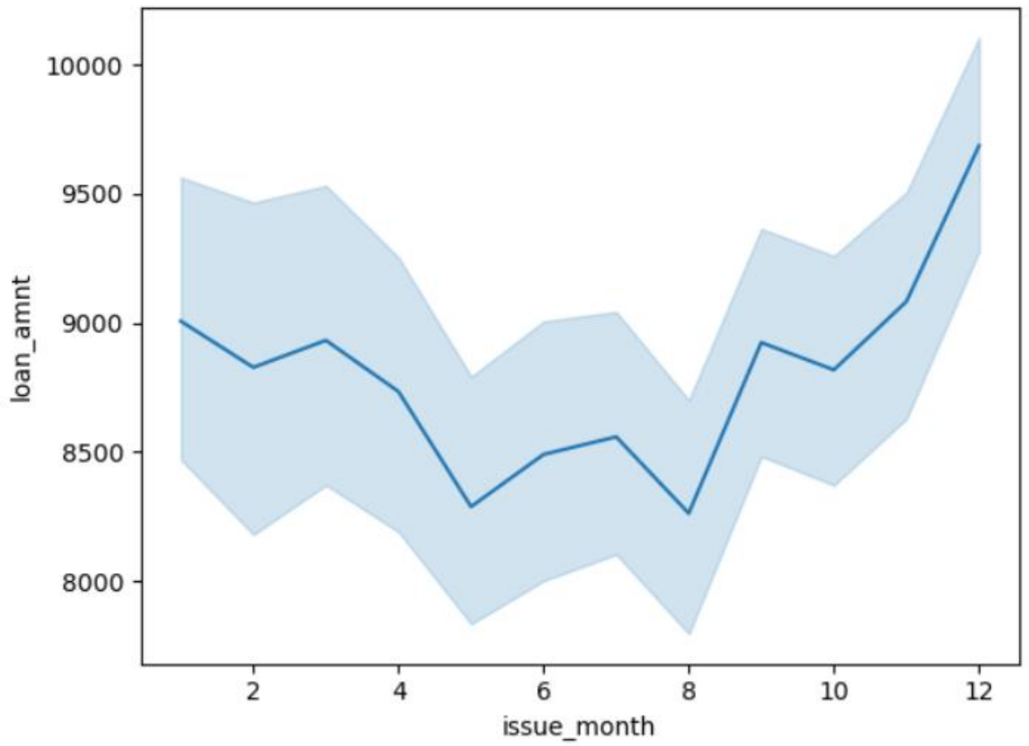


loan_status	Charged Off	Fully Paid	chargedoff_percent
home_ownership			
MORTGAGE	15345200.00000	92763100.00000	14.19000
OTHER	182450.00000	632025.00000	22.40000
OWN	3008275.00000	16798950.00000	15.19000
RENT	19600900.00000	109342100.00000	15.20000

Though the overall loan amount against "Other" is small, the percentage of default loan amount is higher. For other categories like Own, Rent and Mortgage, the % of default loan is similar.

Bivariate Analysis

<u>Analyse the relationship between default loan amount and issue months</u>

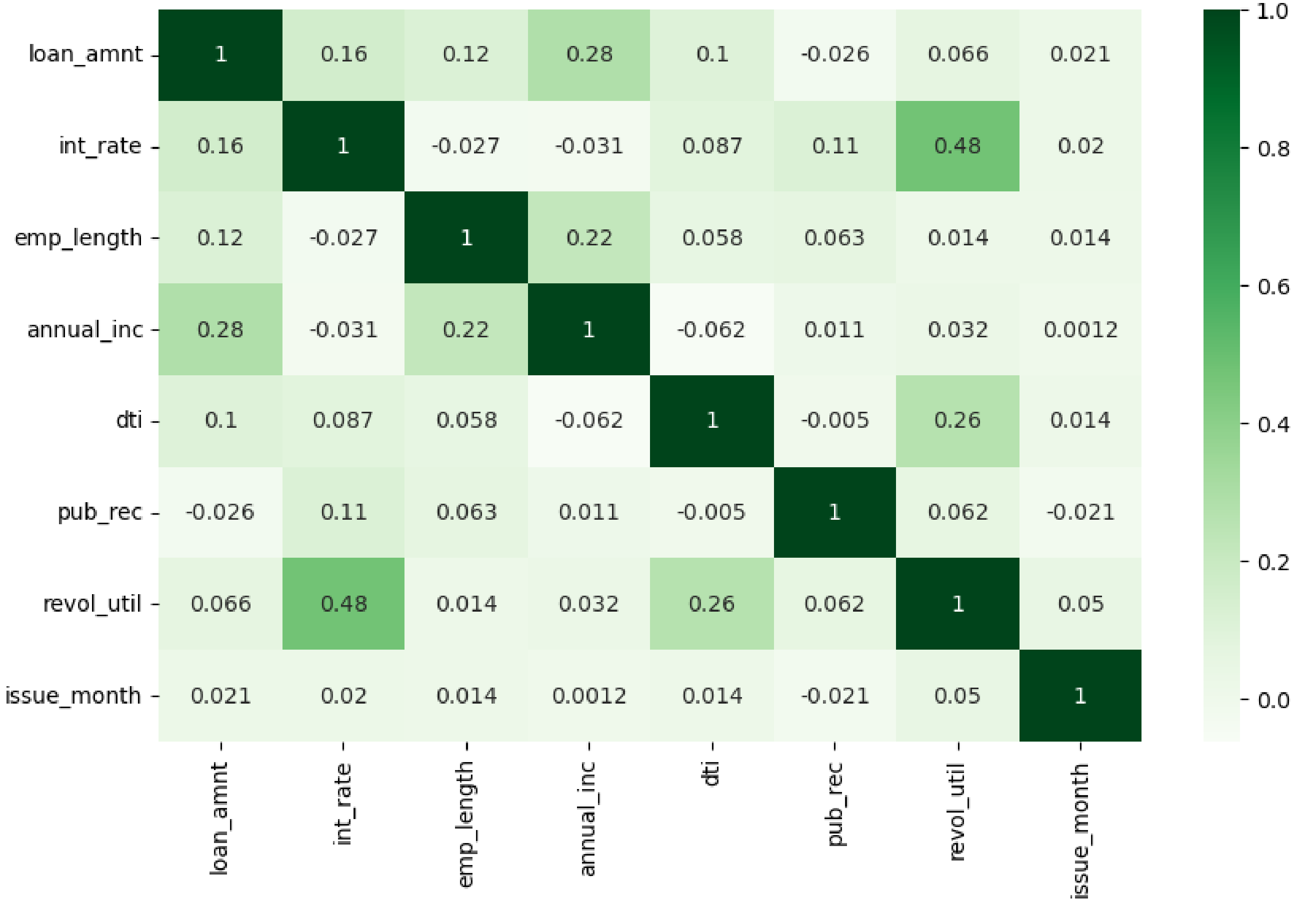


Defaults on high on loan amounts that were issued in the month of January and December across the years

Multivariate Analysis

Multivariate Analysis

Correlation Matrix of continuous variables



No strong correlation between any two variables. Interest Rate and Revolving Line Utilization Rate are positively correlated.

Recommendations

Recommendations

Major Driving Factors for loan defaults

- **Home Ownership** – Borrowers whose home ownership values are rent or mortgage have a higher risk of defaulting loans.
- **Purpose** – If the purpose of the loan is debt consolidation, there is a higher probability of defaulting loans
- **Income Verification** – If the income is not verified, higher the risk of default
- **Work Experience** – Borrowers with work experience range between 0-1 and 10+ years have had a higher default loan amount value.
- **Issue Month** – Loans issued during the months of December, January have higher default loan amounts.
- **Loan Amounts** – Borrowers who took loans for amounts in the range of 4K-12K have defaulted more.

Recommendations

Major Driving Factors for loan sanctions

- **Interest Rate & Purpose** – In the past, higher interest loans had been sanctioned for purposes like small business, house, debt consolidation. The Finance company can look at loan applicants for these purposes for higher interest rates and hence higher income.
- **Grade** – Applicants from Grade A have less defaulted loans. Company can look to sanction loans to applicants in Grade A

Thank you

Saikrupa Purushothaman

p.saikrupa@gmail.com