Lending Club Case Study

An Exploratory Data Analysis

By- Neha Sengupta

Business Objective:

- A consumer finance company **Lending Club** wants to identify risky loan applicants based on historical data of applicants.

- They want to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

- The company can then use this knowledge for its portfolio and risk assessment ,which will help them to reduce sanctioning such loans & thereby cutting down the amount of credit loss.

- Identification of such Variables/Factors which are strong indicators of loan default using EDA is the aim of this case study.

Analysis Approach

Data Cleaning

- Understanding the data
- Handle missing values & remove nulls
- Fix datatype discrepancies ,Create derived metrics

Univariate Analysis

- Finding trends in distribution for numerical variables
- Finding trends in frequencies of categorical variables

Segmented Univariate

- Finding patterns within the segments/categories of data variables
- Finding patterns withing segments of derived metrics

Bivariate & Multivariate Analysis

- Analyzing 2 or more variables and identifying their relationship
- Correlation analysis

Insights and Recommendations

- Provide insights of EDA
- Identify top factors that affect default
- Recommend ways to prevent future defaults

Data Cleaning & Manipulation

We addressed various data quality issues by:

- Identifying and imputing missing values
- Removing data redundancies
- Filtering/assumptions on data based on business knowledge
- Standardising Values (fixing datatypes, date & string manipulation, binning)
- Outlier treatment
- Creating derived metrics(business driven, type driven)

Assumptions about the Data:

1. There are mainly 3 types of data present in the dataset:

- Consumer data: such as annual income, home ownership etc.
- Loan attributes: such as Rate of interest, loan term etc.
- Data/columns which are calculated after loan is approved

We will not consider the 3rd type of data since they will have no relevance to our analysis.

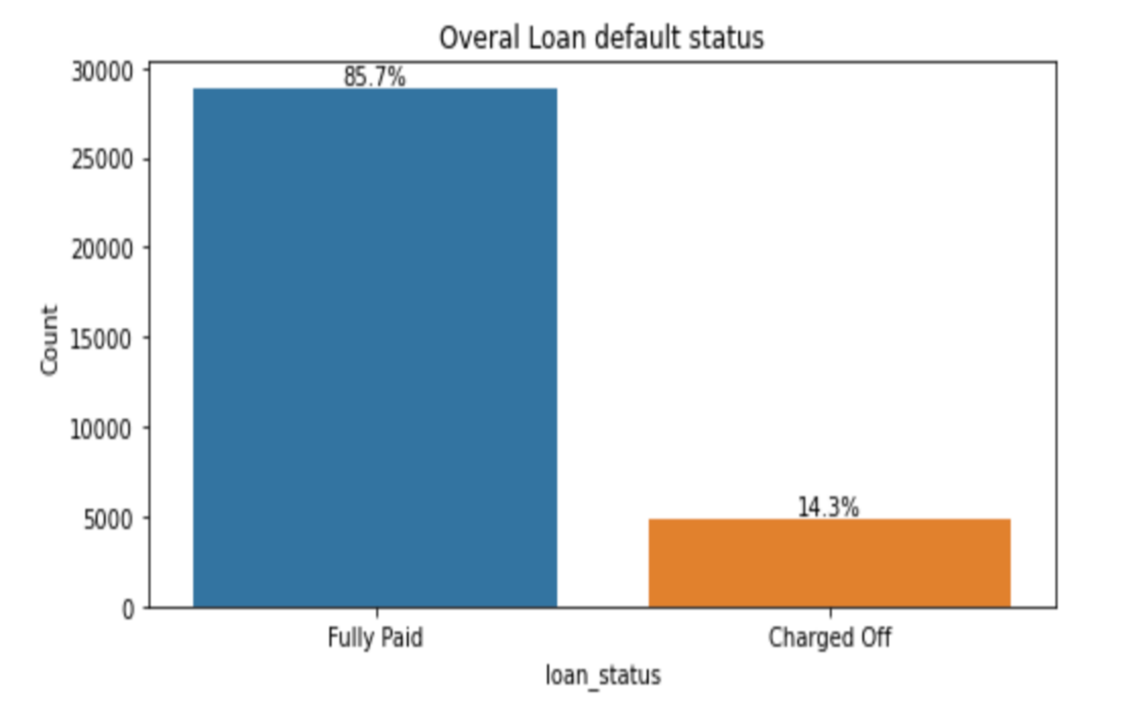
2. The target variable in our analysis is the Loan status.

- We will only consider loans with status Fully paid or Defaulted, and omit 'Current' loans

Understanding our target variable- **Loan Status**

- Most loans were fully paid compared to the charged off/defaulted ones

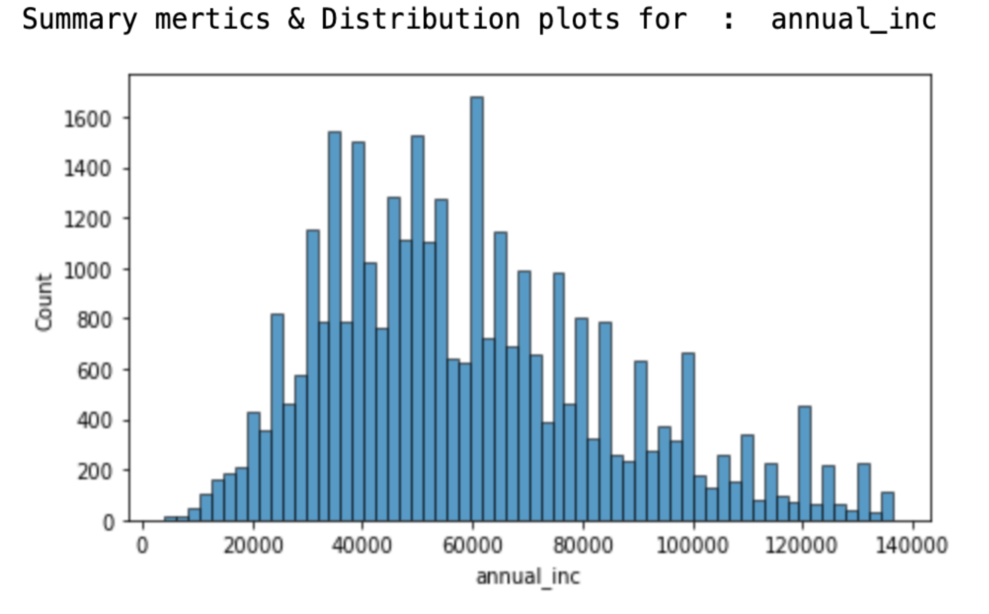
- However , the 14.3% default indicates that there is some scope of improvement in the loan approval process to further reduce the charged off loans



Univariate Analysis

Continuous Variables

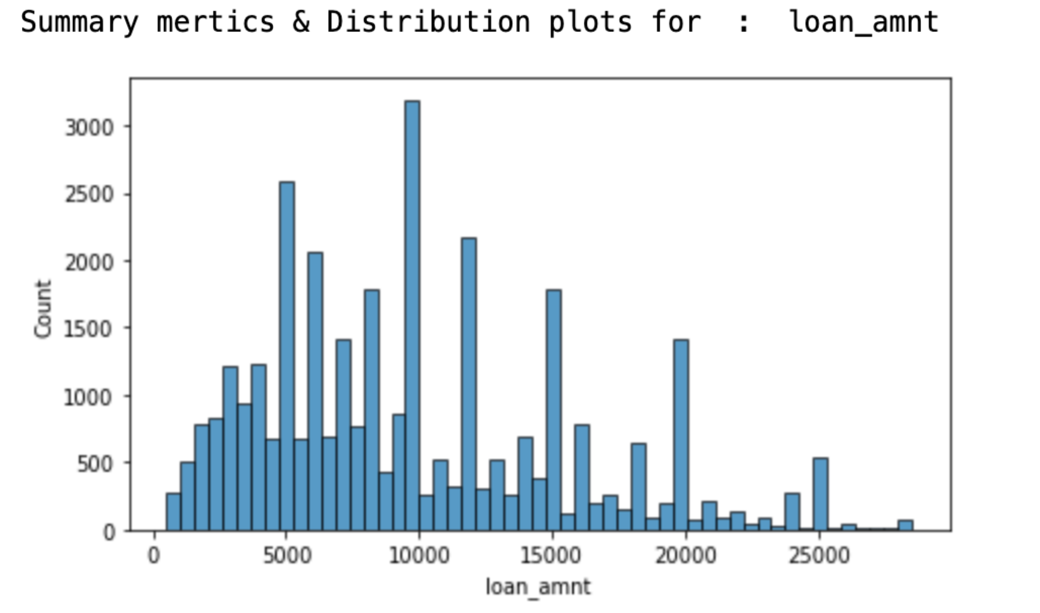
Annual Income



Summary mertics & Distribution plots for : annual_inc

- Majority of borrowers have an Annual income in the ranges of under 60K.
- This suggests lower income groups are opting for loans more often.

Loan Amount



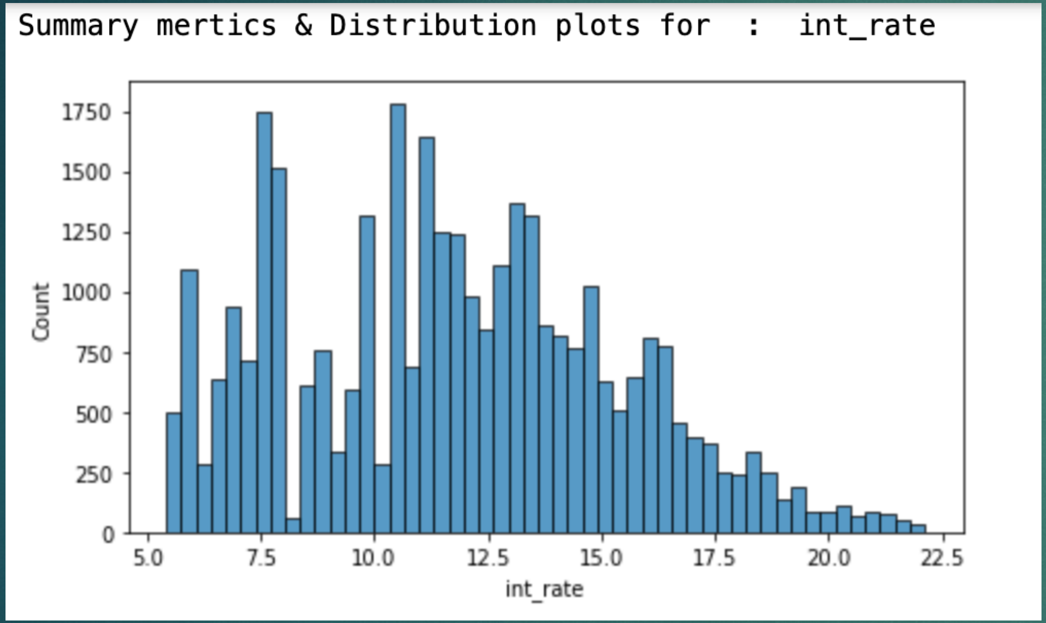
Summary mertics & Distribution plots for : loan_amnt

- Loans are mostly requested in the ranges of 5k to 10k ,with peaks at 5k and 10k
- Loan over 25K are rare

Univariate Analysis

Continuous Variables

Interest Rate

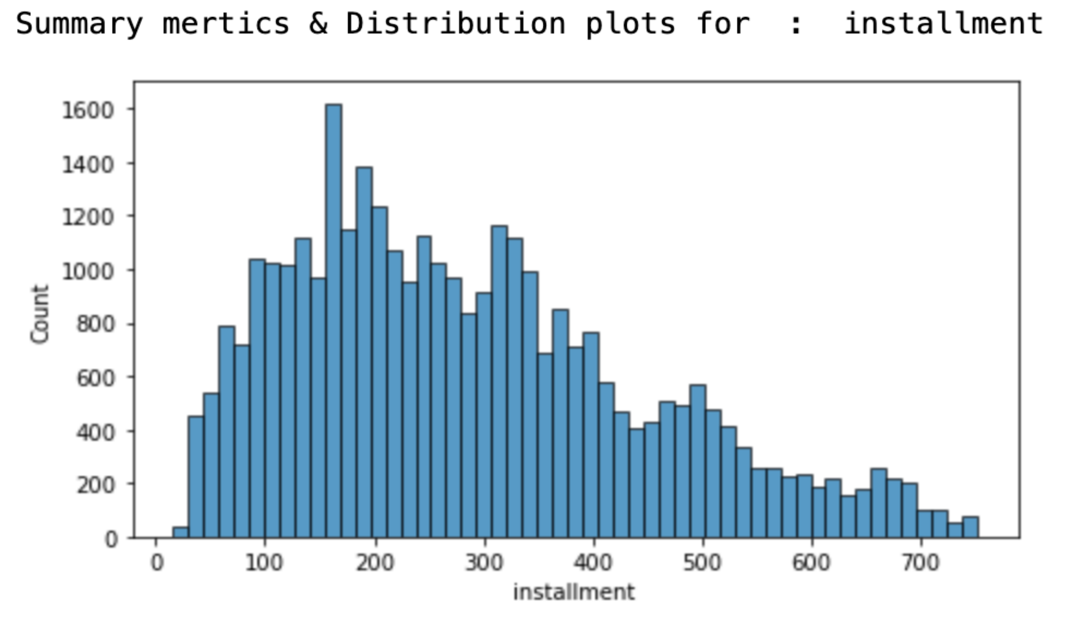


Summary mertics & Distribution plots for : int_rate

- 11% is the most popular rate of interest for loans for most borrowers

- Number of loans with an interest rate higher than 16% sees a downward trend on the overall data

Installments



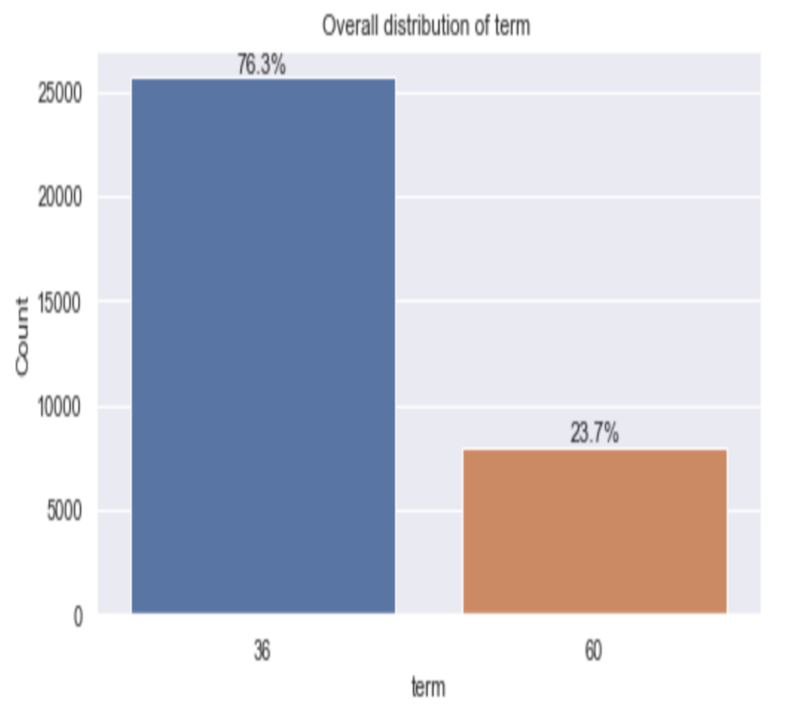
Summary mertics & Distribution plots for : installment

- Most borrowers opt for a less EMI (within n the ranges of 150 to 250)

Univariate Analysis

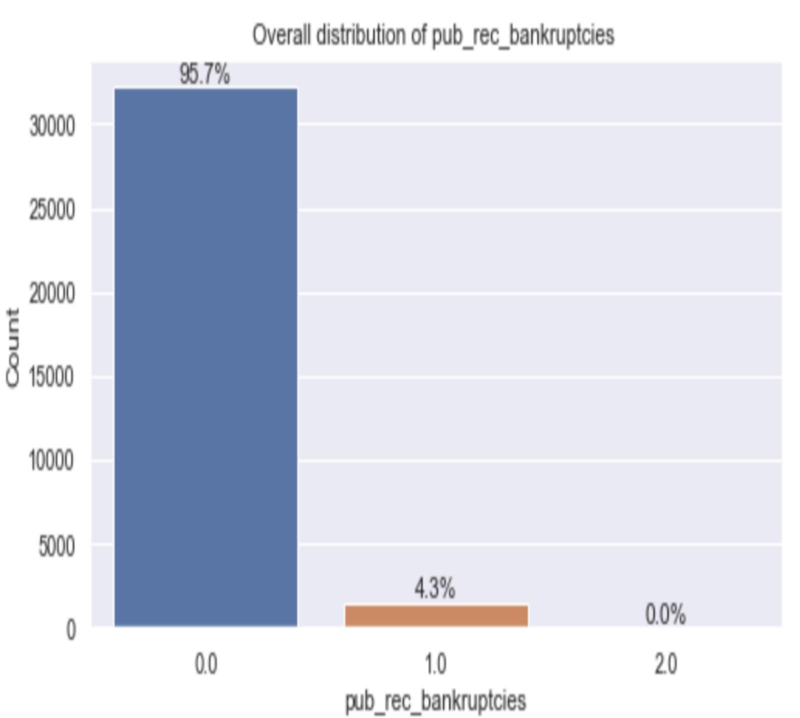
Categorical Variables

Term



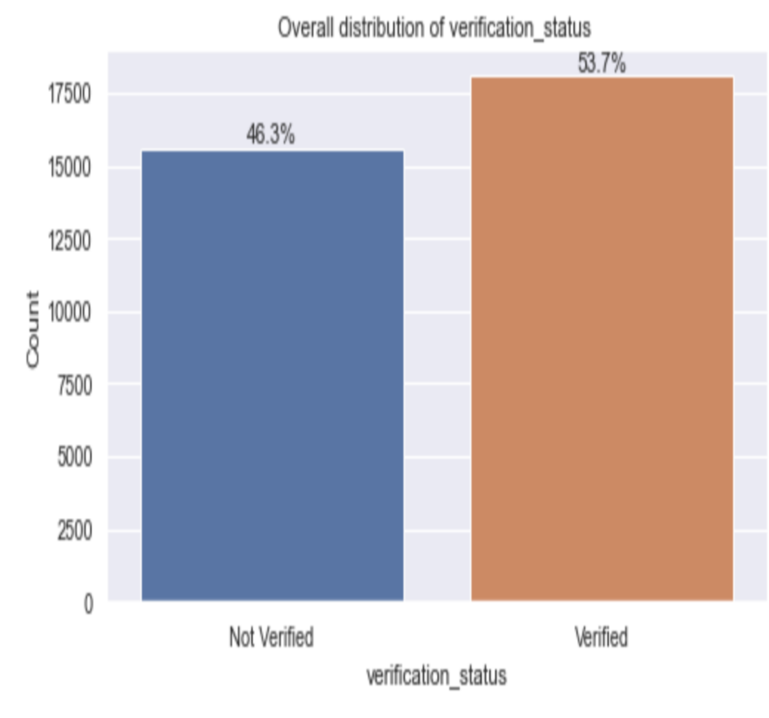
•36-month term period is the more popular option among borrowers

Recorded Bankruptcies



•Majority of applicants did not have any records of bankruptcies

Income Verification

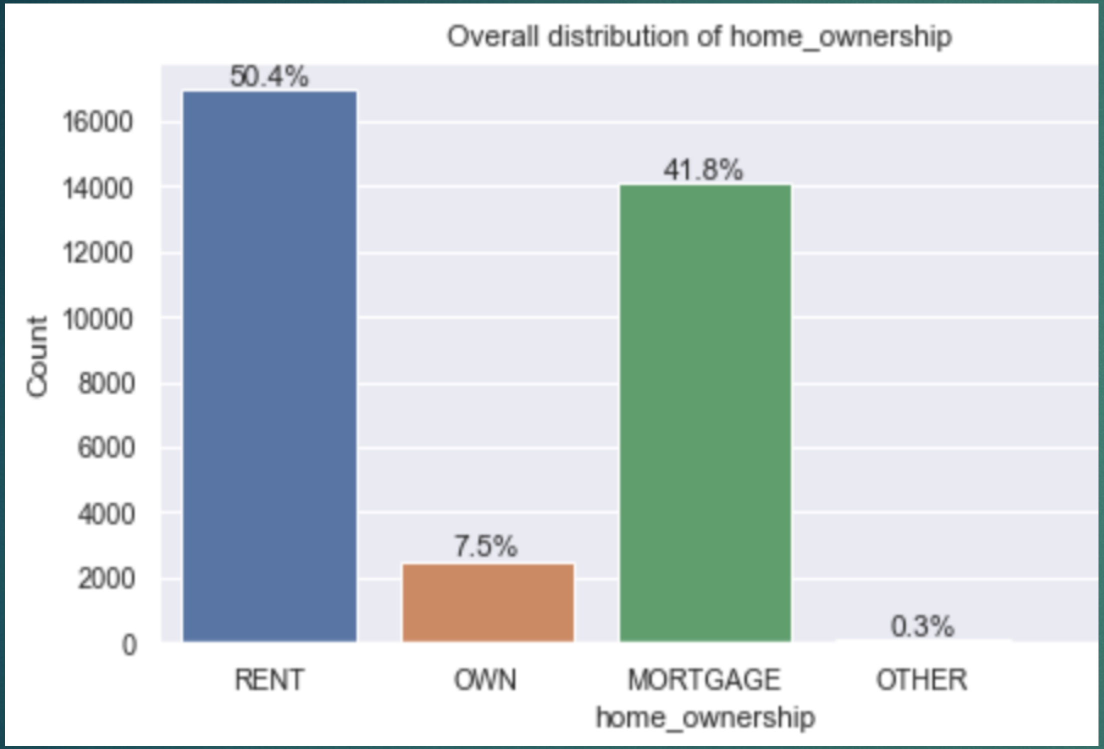


•Over 50% of the applications had a verified source of income

Univariate Analysis

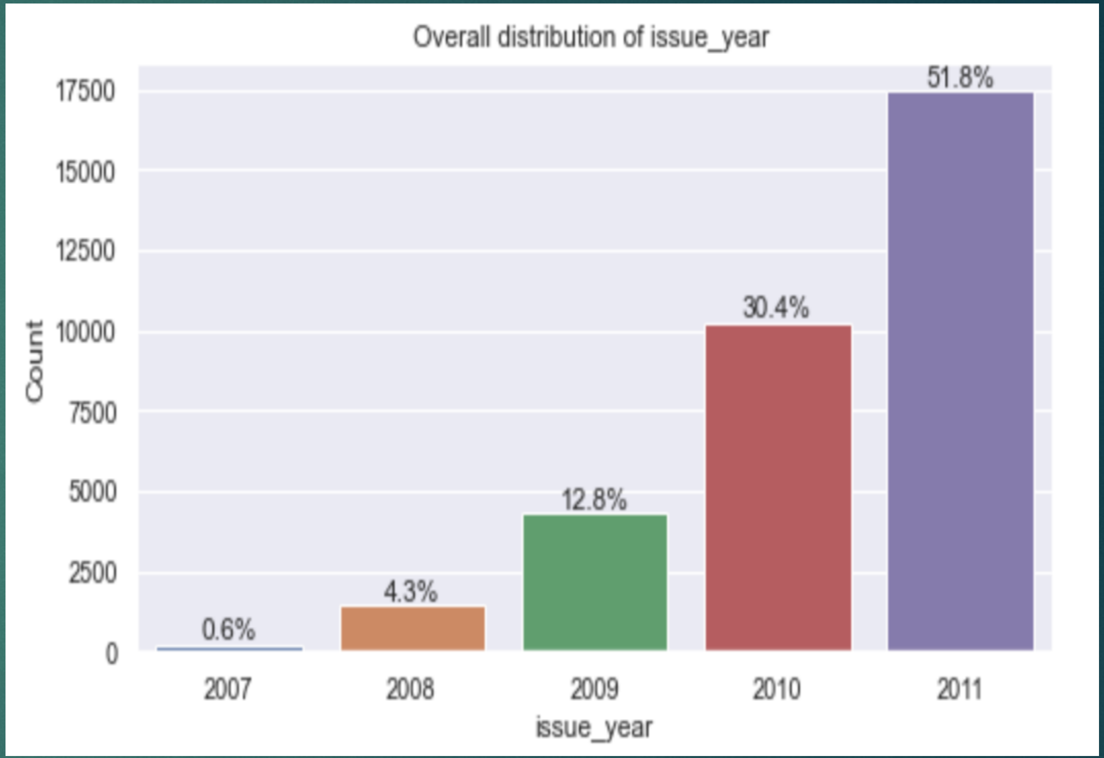
Categorical Variables

Home Ownership



•Majority of borrowers either have rented homes or have a mortgage

Issue Year

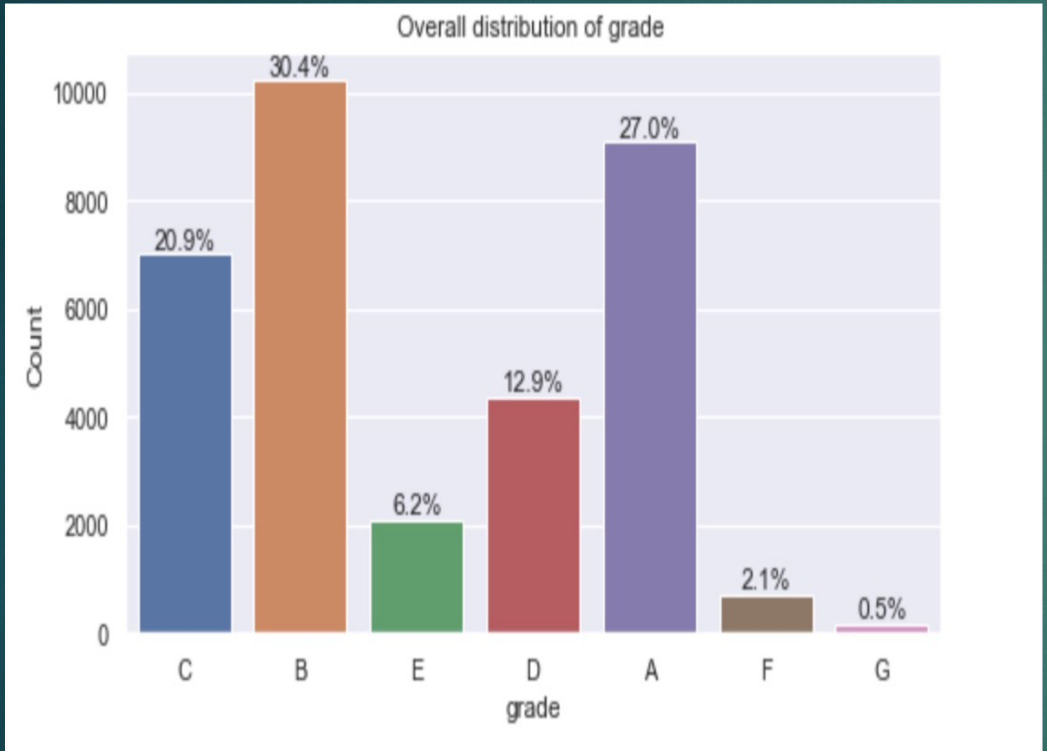


The number of loans being issued by the company have been increasing drastically over the years

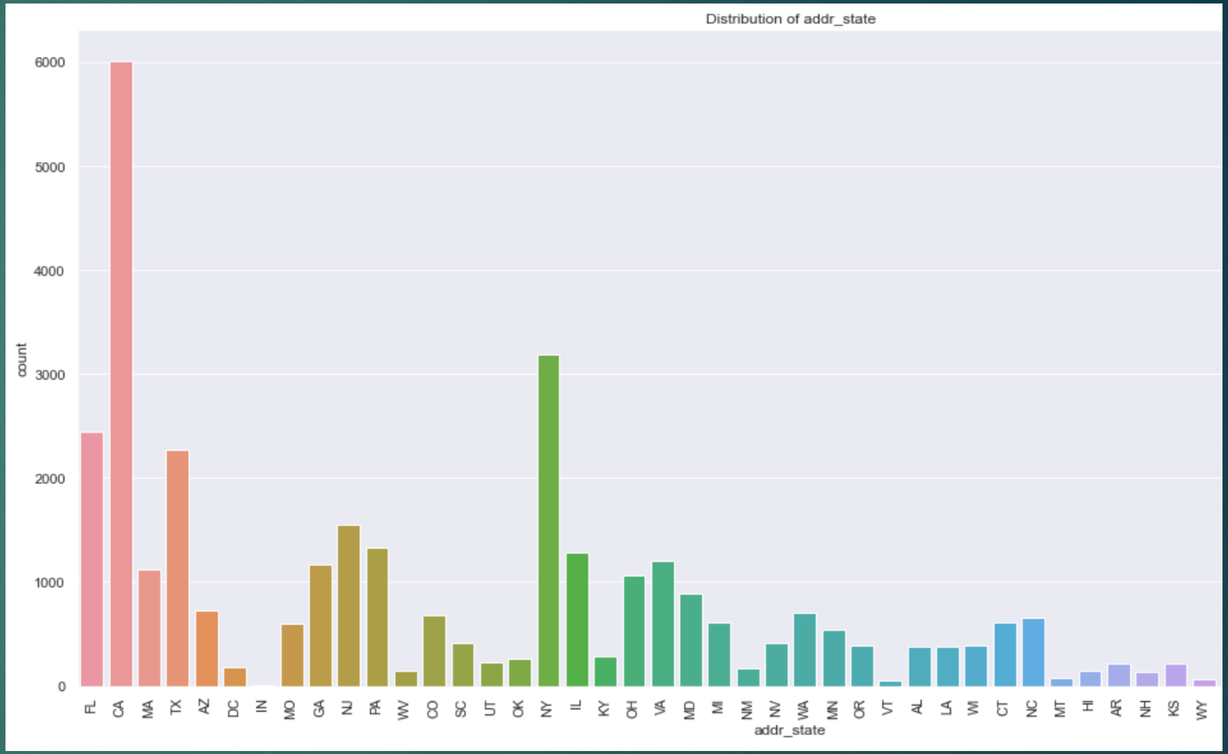
Univariate Analysis

Categorical Variables

Loan Grade



Address State of Borrower



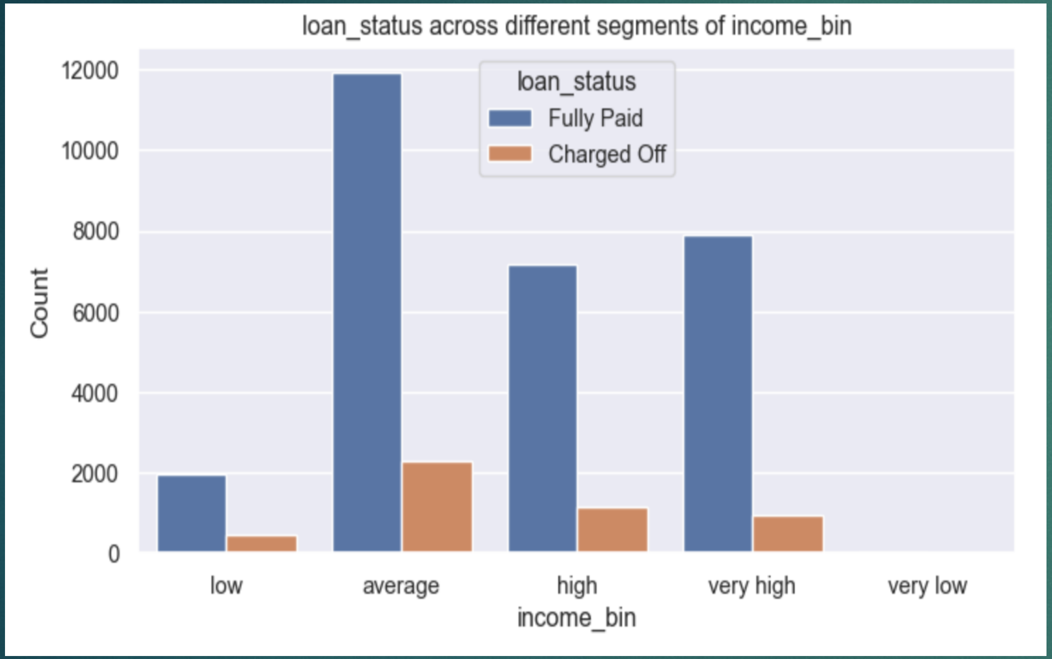
Loans classifying as Grade A and B are more in number.
The approval based on loan grading seems accurate.

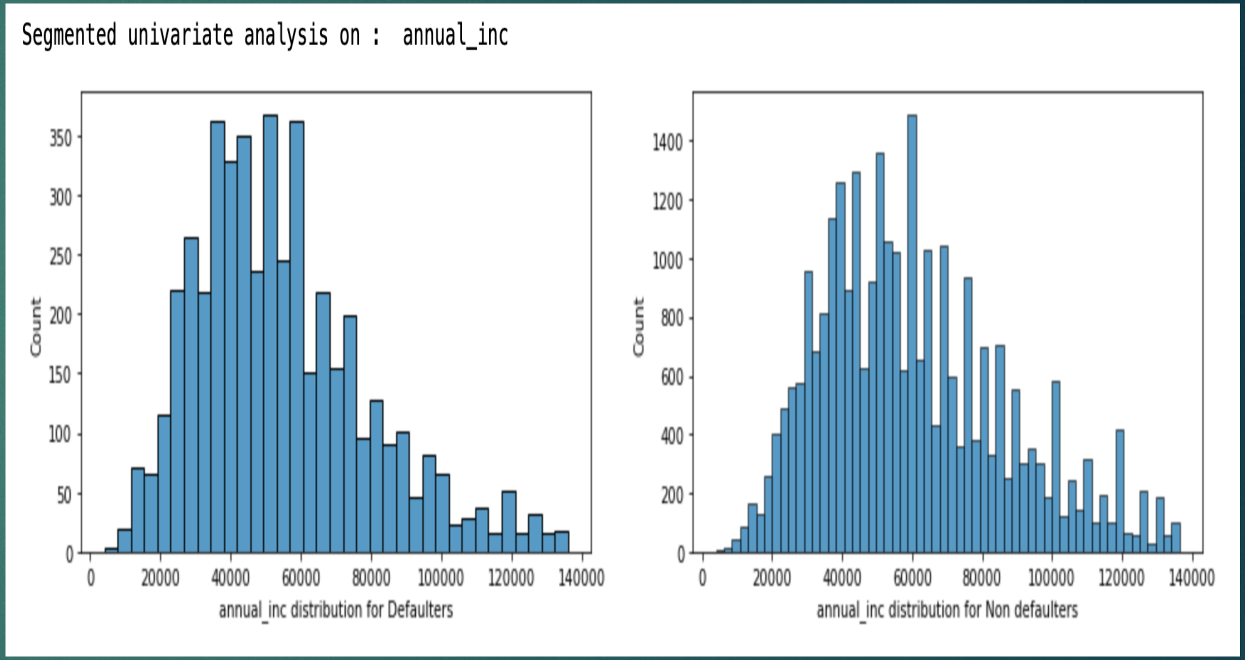
Highest numbers of borrowers are from urban states like CA followed by NY, FL and TX.

Segmented univariate Analysis

Derived Metrics → Comparing loan status with derived metrics

Annual Income vs Loan Default





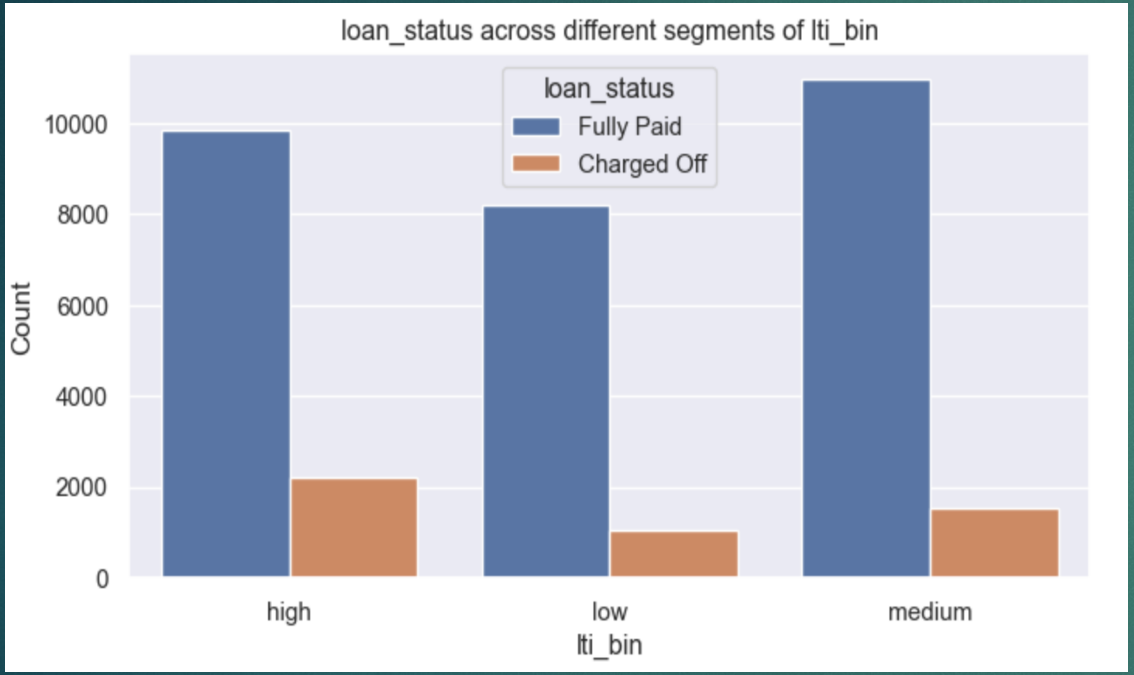
```
Segmented univariate analysis on :  annual_inc
```

- Loans are mostly taken by individuals whose income is under 60K
- Most Defaulters also have an income in the range of 35k to 55k
- This medium income group of applicants seems to be risky

Segmented univariate Analysis

Derived Metrics → Comparing loan default with derived metrics

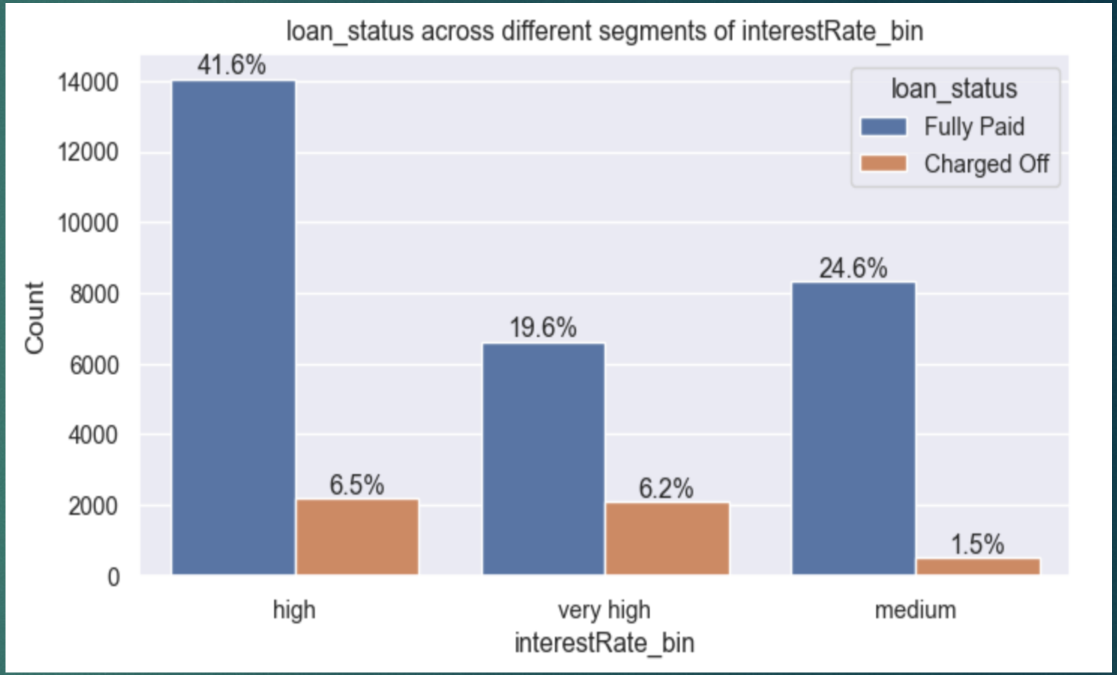
Loan-to-income Ratio Vs Loan Default



•The borrowers with high to medium ***loan to income ratio*** tend to default more

•Which suggests higher defaulters are requesting for loan amount

Rate of interest vs Loan Default



Most defaulters had loans with 13.5% interest rate

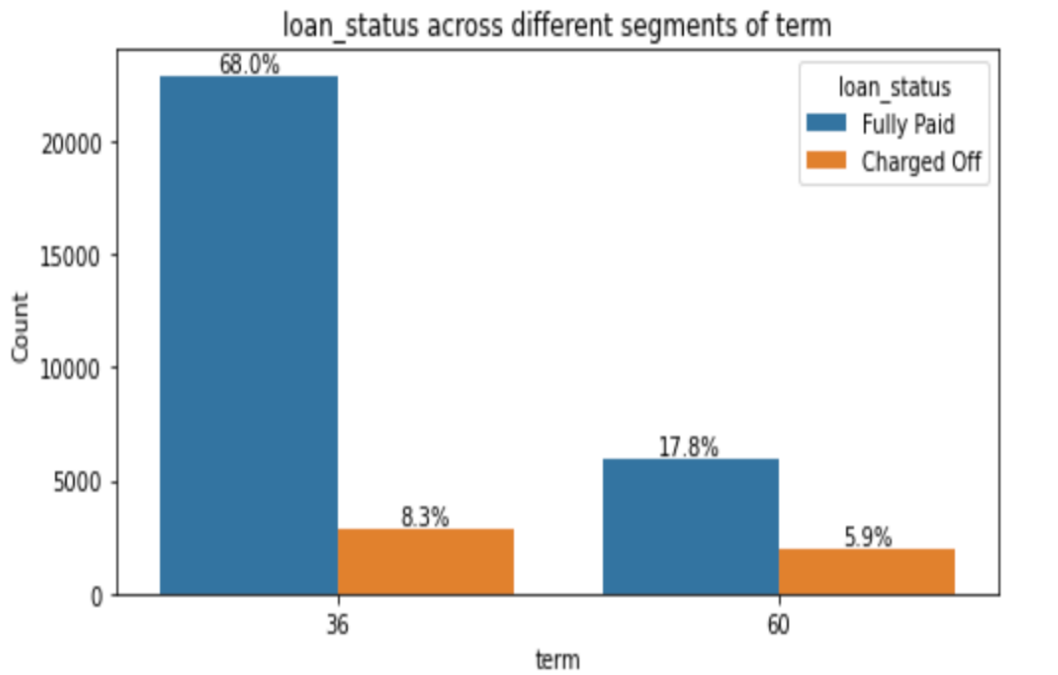
Loans with high(9-14%) or very high rate(above 14%) of interest, led to a greater number of defaults

Defaulters were prone to take loans with higher interest rate

Segmented univariate Analysis

Categorical Variables ➡ Loan Default across different segments of categorical variables

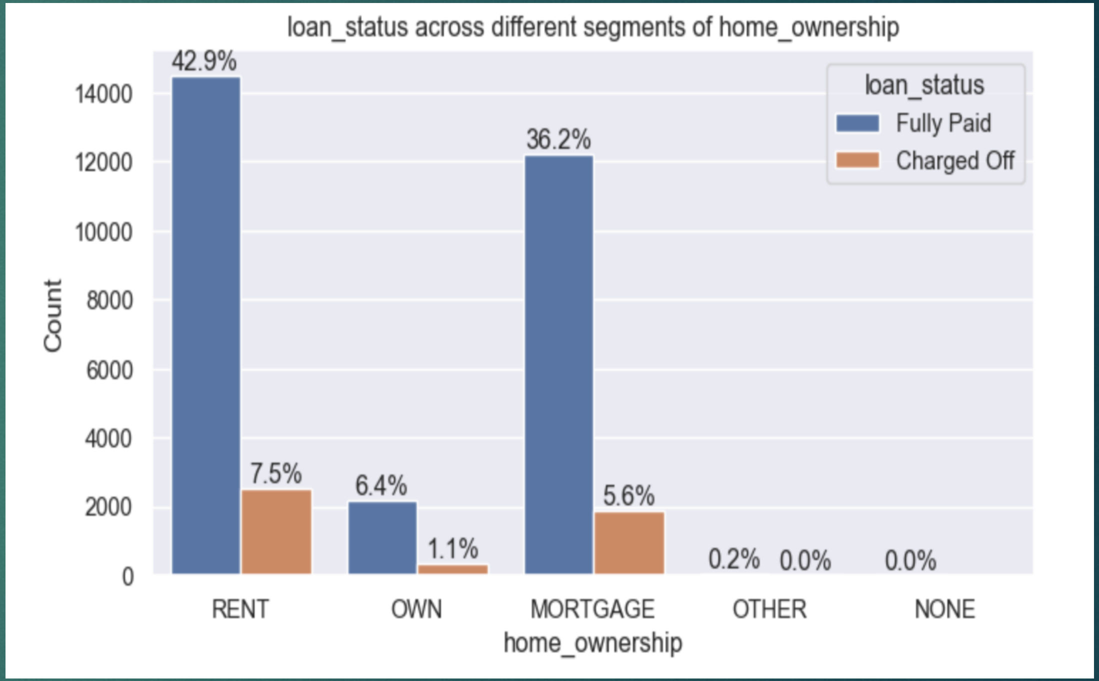
Term Vs Loan Default



- Around 25% of borrowers who opted for 60 months term ,defaulted. There were lesser defaults from the 36 months category

- ***This indicates that a 60-month loan tenure is riskier.***

Home ownership vs Loan Default

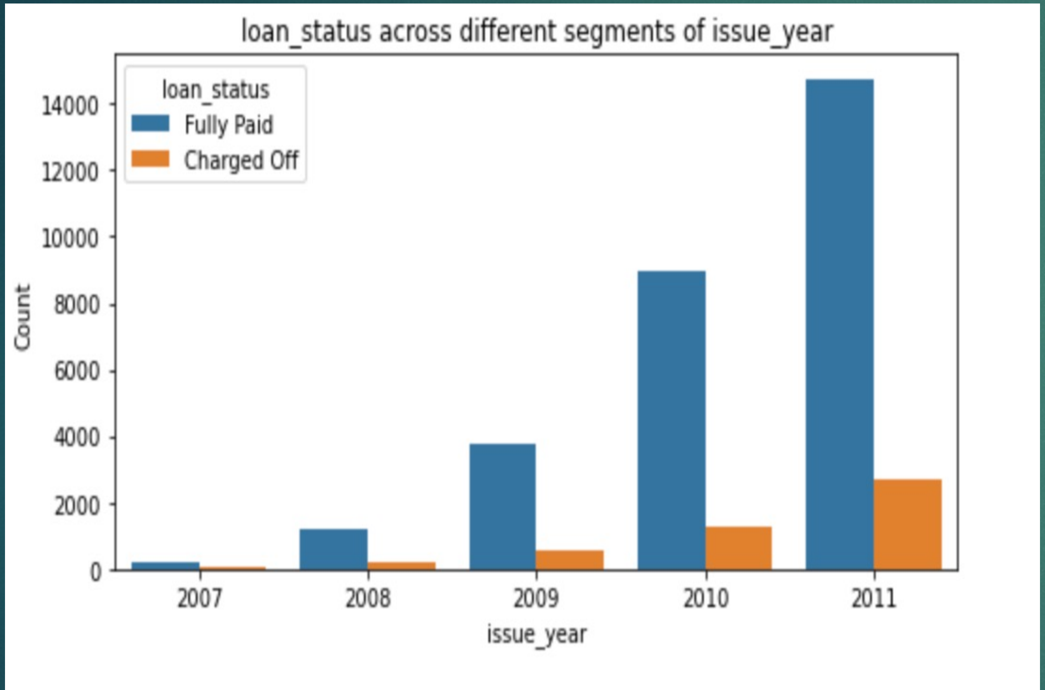


Borrowers who either have rented accommodation or have a mortgage collectively had higher default rates

Segmented univariate Analysis

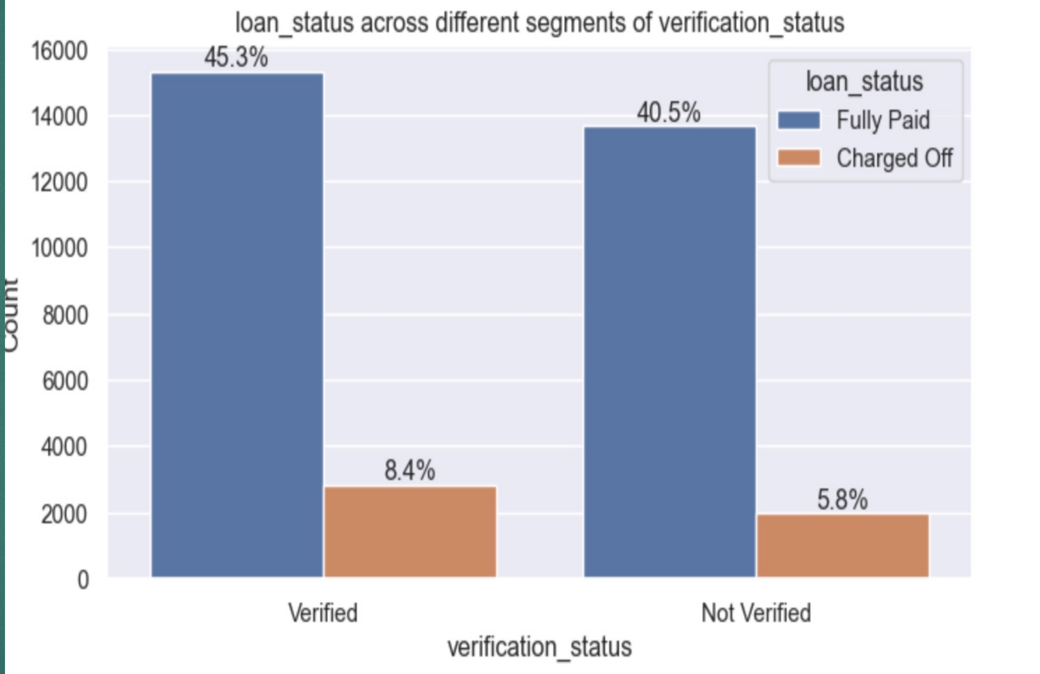
Categorical Variables → Identifying loopholes in some business processes

Issue Year Vs Loan Defaults



2011 saw the highest number of defaulters as well as highest number of approved loans

Verification Status vs Loan Defaults



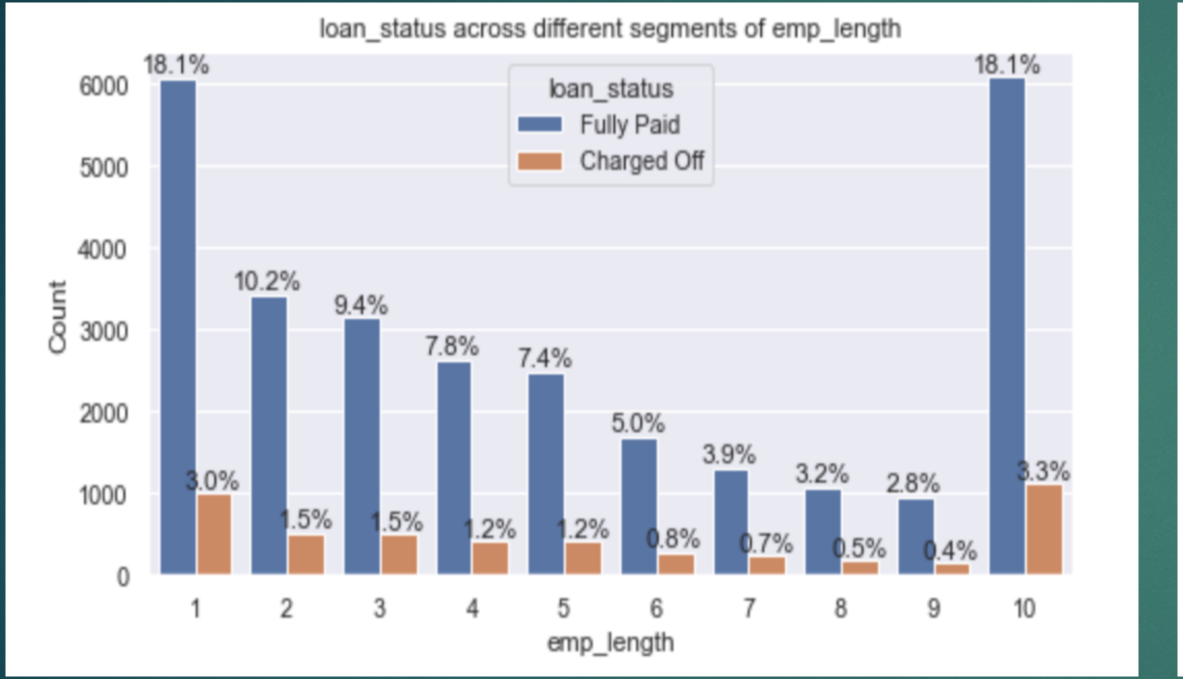
Unexpectedly, borrowers with income source as "verified" defaulted more compared to the ones without verification

Both these points indicate that over the years, the verification process used by the company may have become faulty and is leading to higher approvals of risky applicants.

Segmented univariate Analysis

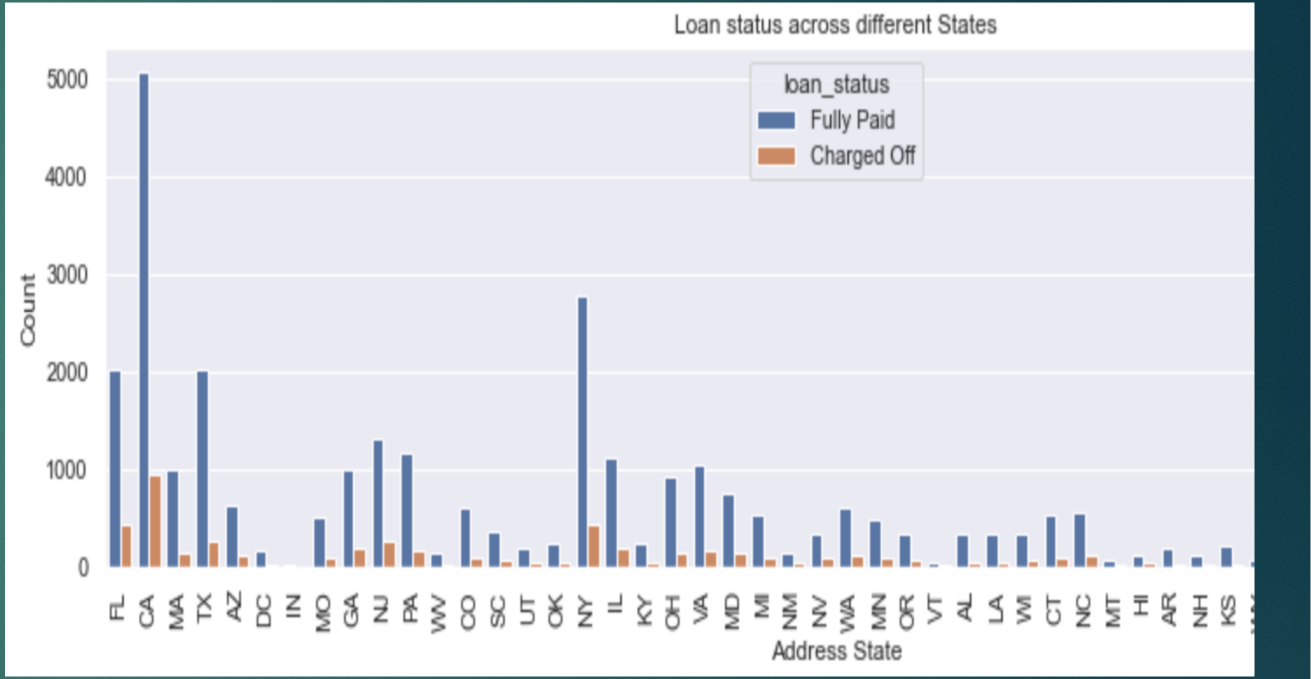
Categorical Variables → Highlighting some interesting trends in the defaulting applicants

Applicant's Work Exp Vs Loan Defaults



- Applicants with 10+ years or 1/ < 1 year experience are likely to default most.
- They are also the top borrowers.

State of domicile vs Loan Defaults



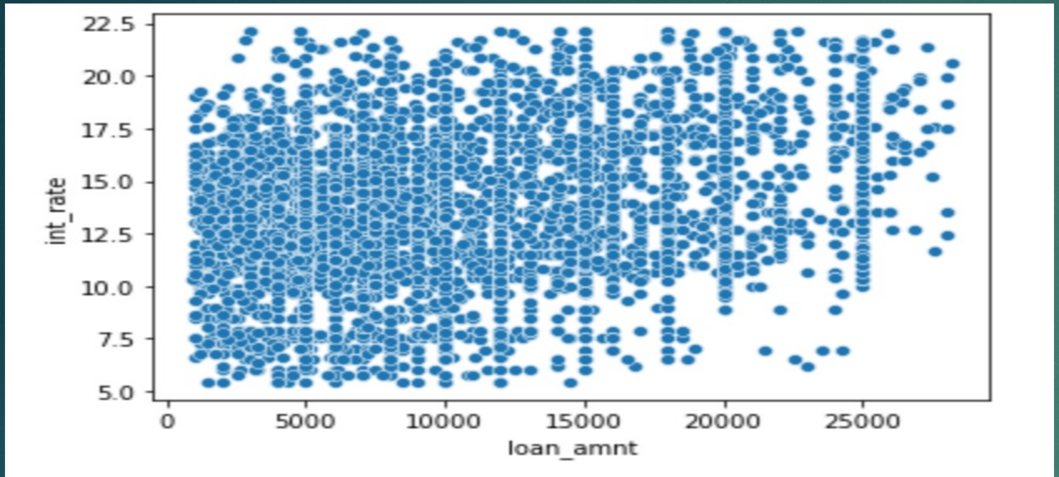
- CA state had the highest number of defaults
- This was followed by FL and NY
- Most borrowers are also from the same set of large urban states

These trends indicate a consumeristic culture among borrowers. It is important to verify their Debt-to-income ratio accurately to avoid defaults.

Bi-Variate Analysis

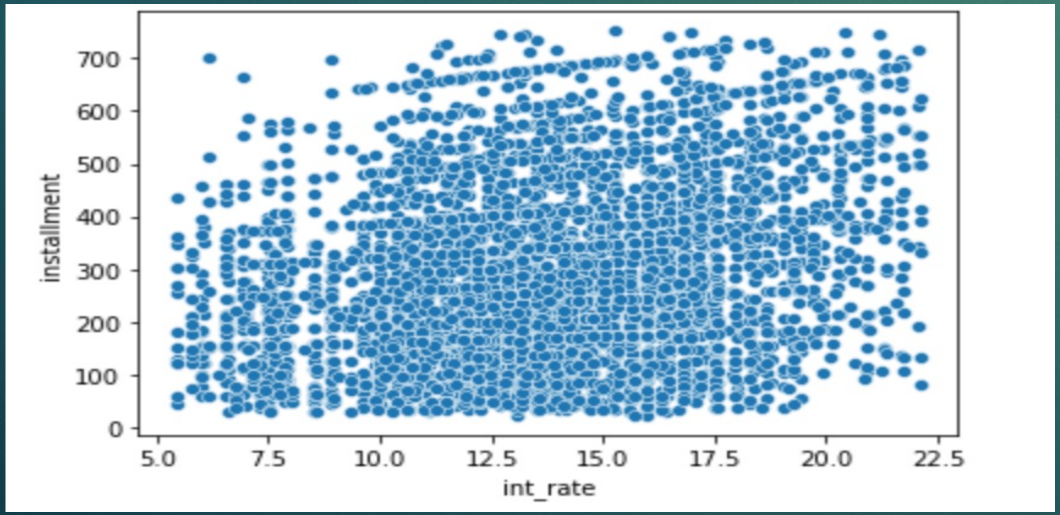
Numerical vs Numerical

Loam amount vs Rate of Interest



The scatter plot of loan amount vs interest rate among defaulters suggests higher the loan amount, higher will be the interest rate.

Rate of interest vs Installment



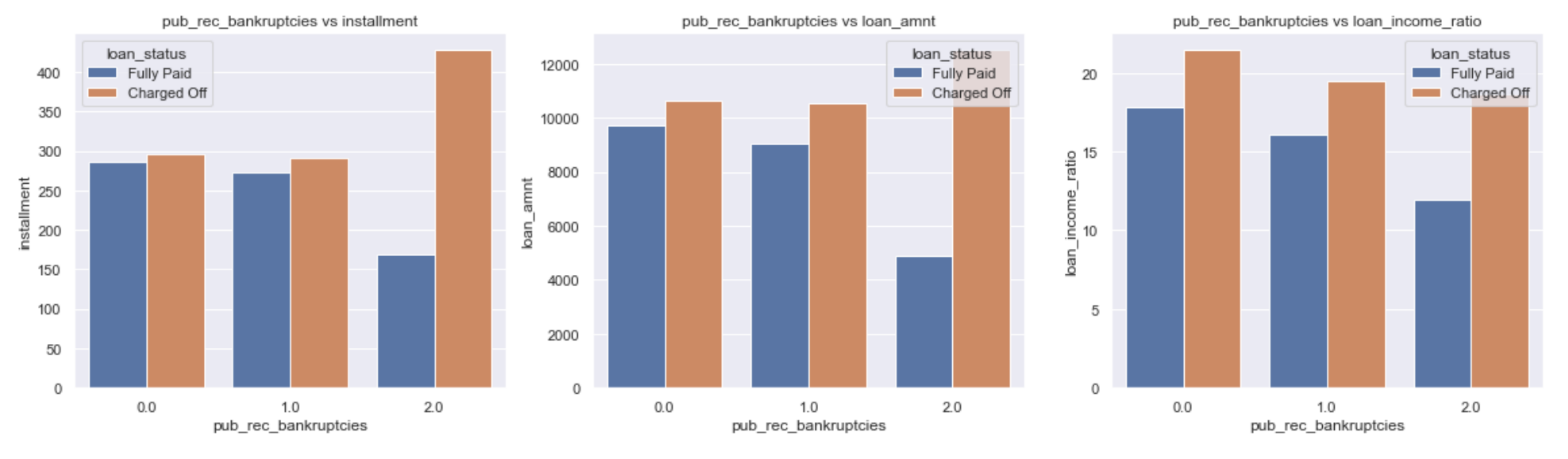
Similarly, the scatter plot of interest rate vs installment among defaulters suggests that loans with higher interest rate have higher EMIs .

Since Higher rate of interest leads to higher default, we can say that defaulters are also paying very high EMIs.

Bi-Variate Analysis

Categorical vs Numerical

- Public recorded bankruptcies vs Loan amount
- Public recorded bankruptcies vs Installment
- Public recorded bankruptcies vs Loan to income ratio



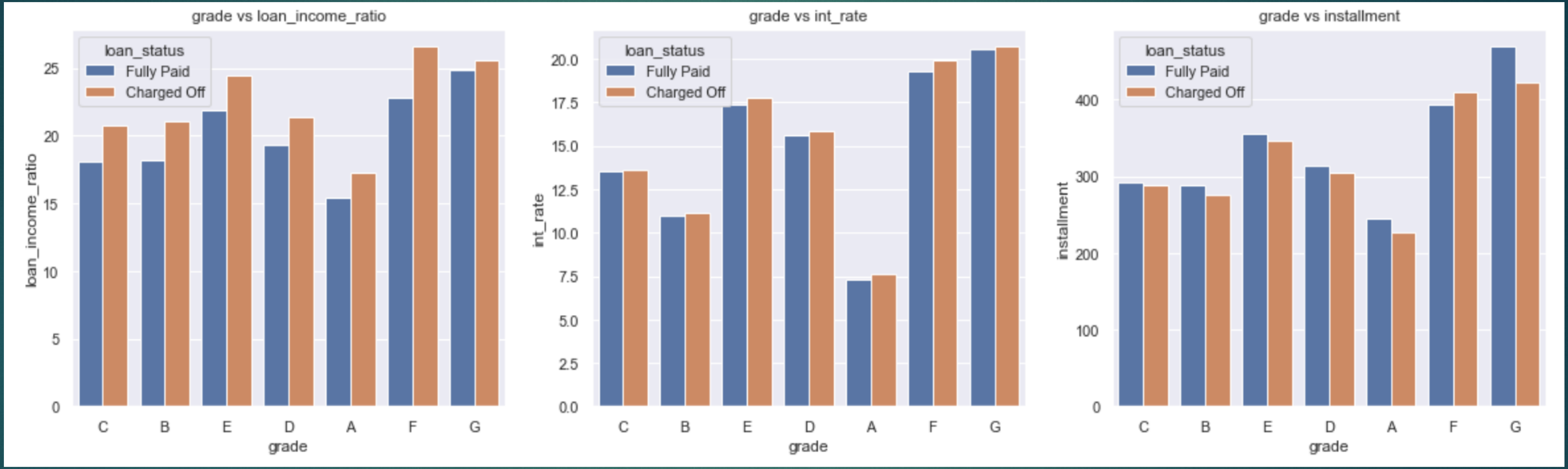
Having a non-zero public recorded bankruptcies among borrowers is a clear indicator of default.

- These borrowers are paying higher EMIs (which inversely indicates high rate of interest of 14-16%)
- Are having a high loan to income ratio (payback capability is low)
- Are applying for a higher loan amount leading to defaults

Bi-Variate Analysis

Categorical vs Numerical

- Grade vs Loan to income ratio
- Grade vs Interest Rate
- Grade vs Loan to installment



A loan with lower grades like F and G are a clear indicator of default. Such borrowers
•Are paying a higher interest rate (around 14-16%)
•Are paying a very high instalments
•Are having a high loan to income ratio (payback capability is low)

Multivariate Analysis

Positive Correlations: When 1 variable increases the other increases

EMI , loan amount and loan-to income ratio Higher the loan amount, higher will be EMI which might lead to default in payments. When the loan amount is high, the LTI ratio will also increase so payback power reduces leading to defaults.

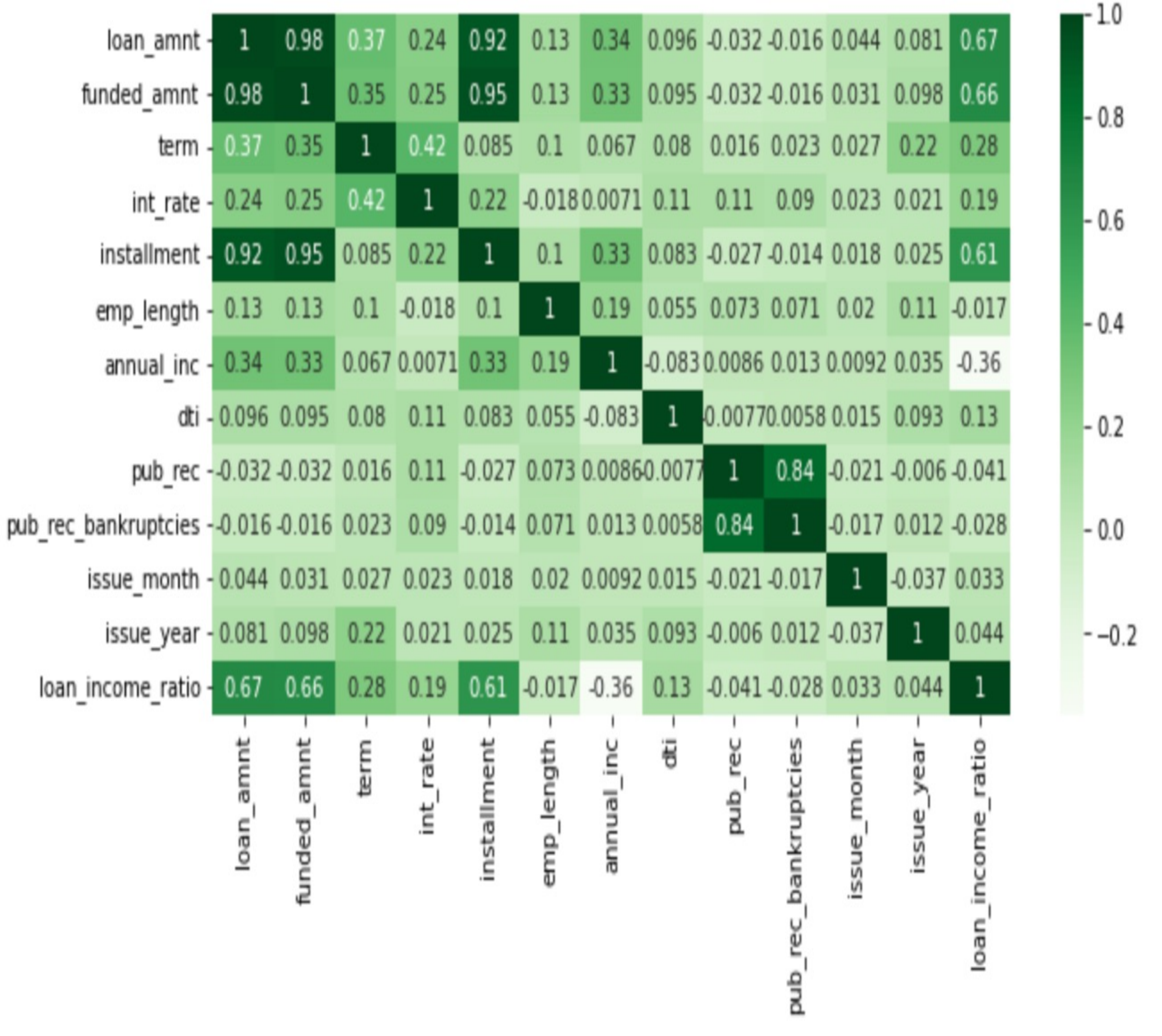
Public recorded bankruptcies and Derogatory records Higher the number of Public recorded bankruptcies would lead to higher number of derogatory public records. These borrowers have a history of defaulting and might be considered risky applicants

Term and interest rate, indicating interest rate increases with longer terms

Negative correlations: When 1 variable increases while the other decreases

Public rec bankruptcies and loan amount Which makes sense as there are less chances of loan being approved if there are a greater number of derogatory records/bankruptcies

Annual income and loan-income-ratio A high loan-income-ratio indicates a lower income and therefore making the borrower a riskier applicant



	loan_amnt	funded_amnt	term	int_rate	installment	emp_length	annual_inc	dti	pub_rec	pub_rec_bankruptcies	issue_month	issue_year	loan_income_ratio
loan_amnt	1	0.98	0.37	0.24	0.92	0.13	0.34	0.096	-0.032	-0.016	0.044	0.081	0.67
funded_amnt	0.98	1	0.35	0.25	0.95	0.13	0.33	0.095	-0.032	-0.016	0.031	0.098	0.66
term	0.37	0.35	1	0.42	0.085	0.1	0.067	0.08	0.016	0.023	0.027	0.22	0.28
int_rate	0.24	0.25	0.42	1	0.22	-0.018	0.0071	0.11	0.11	0.09	0.023	0.021	0.19
installment	0.92	0.95	0.085	0.22	1	0.1	0.33	0.083	-0.027	-0.014	0.018	0.025	0.61
emp_length	0.13	0.13	0.1	-0.018	0.1	1	0.19	0.055	0.073	0.071	0.02	0.11	-0.017
annual_inc	0.34	0.33	0.067	0.0071	0.33	0.19	1	-0.083	0.0086	0.013	0.0092	0.035	-0.36
dti	0.096	0.095	0.08	0.11	0.083	0.055	-0.083	1	-0.0077	0.0058	0.015	0.093	0.13
pub_rec	-0.032	-0.032	0.016	0.11	-0.027	0.073	0.0086	-0.0077	1	0.84	-0.021	-0.006	-0.041
pub_rec_bankruptcies	-0.016	-0.016	0.023	0.09	-0.014	0.071	0.013	0.0058	0.84	1	-0.017	0.012	-0.028
issue_month	0.044	0.031	0.027	0.023	0.018	0.02	0.0092	0.015	-0.021	-0.017	1	-0.037	0.033
issue_year	0.081	0.098	0.22	0.021	0.025	0.11	0.035	0.093	-0.006	0.012	-0.037	1	0.044
loan_income_ratio	0.67	0.66	0.28	0.19	0.61	-0.017	-0.36	0.13	-0.041	-0.028	0.033	0.044	1

Summary

Defaulters list

Risky Consumer attributes :

- **Annual income** : Annual income in the ranges of 35k to 60k see higher defaults.
- **Loan to income ratio** : Borrowers with high loan to income ratio tend to default more as payback power is less
- **Home ownership** : Borrowers who either have rented accommodation or have a mortgage collectively had higher default rates
- **State**: Borrowers from large urban states are prone to defaulting
- **Public recorded bankruptcies** : Having a non-zero public recorded bankruptcies is a strong indicator of default

Risky Loan attributes:

- **Loan amount** : Higher the loan amount, higher the chances of default.
- **Rate of interest**: Interest rate of 13.5% and above lead to higher defaults.
- **Term**: A 60-month loan tenure is riskier than 36 months.
- **EMI**: Higher EMIs lead to higher chances of defaulting
- **Verification Status** : Unexpectedly, borrowers with income source as "verified" defaulted more compared to the ones without verification. This indicates that the verification procedure followed by the company may not be accurate
- **Grade**: Loan Grades F & G are good indicators of identifying defaults.

Some Recommendations:

- TO DO :
  - Encourage a shorter term for loans
  - Consider a non-zero public recorded bankruptcies for loan approval
  - Re visit the income verification process and identify loopholes if any
- STOP:
  - Restrict loans given to individuals with high loan to amount ratio
  - Restrict high interest loans given to urban population
- CONTINUE:
  - The process for risk categorization via **Loan Grades** as they have proved to be good indicators for identifying defaults