LENDING CLUB CASE STUDY

SAURABH JAIN

JITENDRA SUTHAR

PROBLEM STATEMENT

- The problem statement concerns to a Finance company which offers various types of loans to urban customers.
- The problem statement is as follows

“To develop an efficient and accurate methodology to decide on the suitability of an applicant for accepting or rejecting their loan applications.”

- The implications of the processes quite significant as

➢ 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

- The challenges in developing the methodology is due to wide variety of data available for the applicants. Choosing the right data set to will allow to take a informed decision

OBJECTIVE

- The primary objective of this case study is
- Identify the correct set of data from the available data for analysis
- Applying various EDA techniques like univariant and bivariant analysis to process the data
- Provide inferences from the data to identify the criteria for approving or rejecting loans

APPROACH

- Data Loading: The Finance company has provided with historical data of the customers pertaining to their credit history and other background information. The first step to load the data to visualize the information available in the data set.
- Data Cleaning: The data provided by the finance company contains wide range of information and all the information is not relevant for the problem statement under consideration. The columns containing the irrelevant data are to dropped off. In addition, there may be row or columns containing null or duplicate values which will be required to be cleaned up. The outliners and missing value treatment will also be required.
- Data Transformation: The provided data must be converted into uniform formats to allow comparison.
- Data Analysis: The prime objective being the analysis of the provided data to derive the inferences and correlation between various attributes to develop an approach for loan sanction based on historical data. The univariant and bivariant analysis techniques are to be used to generate the information from the raw data set for analysis purpose.
- Recommendations: Based on the analysis of the historical data to provide the recommendation on the attributes to be considered while sanctioning a loan to mitigate the risk of default and to avoid false negative and rejection of suitable applicant

DATA LOADING

- Python is being used as the programming language for data cleaning and analysis purpose.
- The following file provided are used
  - loan.csv: The file contain the complete data set for the past customers
  - Data_Dictionary.xlsx: The file provides the definition of all the terms used in loan.csv file
- Only the “loan.csv” file has to be loaded for analysis purpose

DATA CLEANING

- Data Reduction
- There are total of 111 columns defining various data attributes
- There are total of 39717 rows excluding the header row. Ideally each row denotes a unique customer data.
- Out of 111 columns 20 are only relevant for data analysis and remaining will be dropped off. The data indicated in the following slide is only considered. The remaining columns are removed.
- One new column named “issued_month” is created which will include month of issued_d column which will help in future ploting.

DATA CLEANING Cont..

Column considered for analysis

addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
dti	A ratio calculated using the borrower's total monthly debt payments on the
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
purpose	A category provided by the borrower for the loan request.
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
loan_amnt	time, the credit department reduces the loan amount, then it will be reflected
loan_status	Current status of the loan
member_id	A unique LC assigned Id for the borrower member.

DATA CLEANING Cont..

- Missing Values

➢ Emp_len, emp_title, mnth_since_last_delinqu and mnth_since_last_record are the cloumns which has maximum rows with null values

The missing values were treated in following manner

Emp_len: all the records with missing values were dropped, as it can critical parameter, and replacing it with mean/median may impact amalysis.

emp_title : There are 2459 missing valueswe can safely rename those unspecified tag.

revol_util : replace NaN values with median of revol_util

DATA CLEANING Cont..

- Data Transformation
- For most of the columns the data types were consistent with the data under consideration
- For few of the columns the data types were changed to attain consistent values the data
- the specific columns are treated as follows
- We are removing 'months' string from term column.
- int_rate - removing % symbols
- issue_d - converting it date format
- earliest_cr_line - converting it date format
- Converting emp_length into categorical type.

DATA CLEANING Cont..

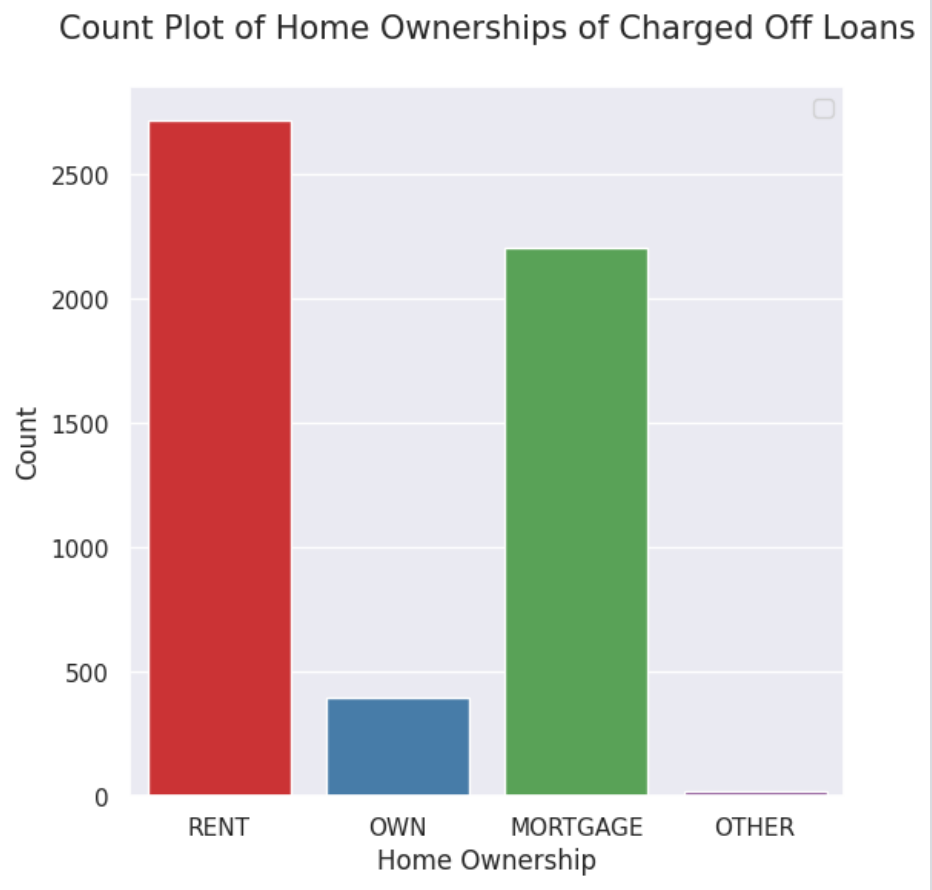
- Outliner Treatment
- Annual income, Loan amount and interest rates are taken as parameters for removing the outliners. It is found that majority of
- Annual income : A box Plot reveal the range of income in which most of the entries fits. There were very few entries above 250000 and those are considered as outliners.
- Loan amount : A box plot of the loan amount reveals that the loan value above 30000 can be considered as outliners. But those higher loan amount cause more loss to company so removing it will be unwise choice. We will anaylize them too.

Univariant Analysis

- A Univariant Analysis was carried out to derive more insight from the data for various attributes of the data. It helps to understand the customer base of the financial company based on visualization of individual parameters of the applicants.
- The following section provides the details of results obtained from Univariant analysis

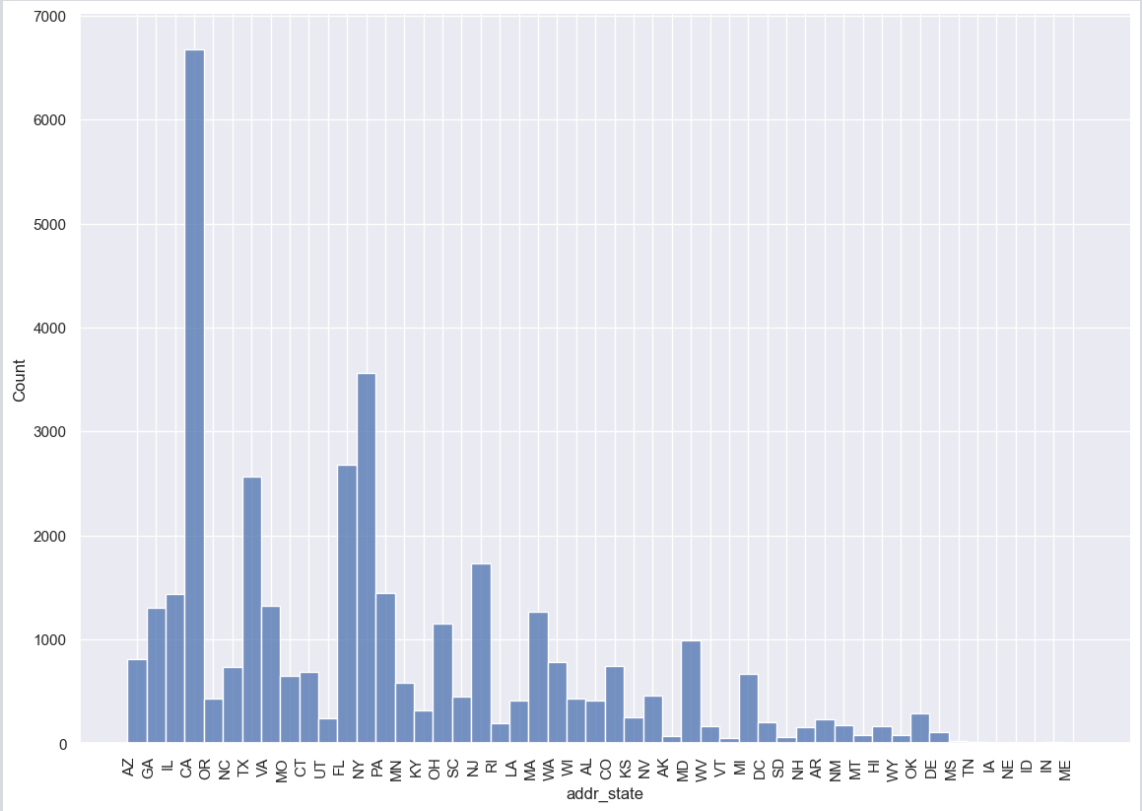
Home ownership

- home_ownership: The graph gives a trend that the most of the loans were sanctioned to the customers who does not own a home. This information will further be corelated to other varialbes to derive insights.



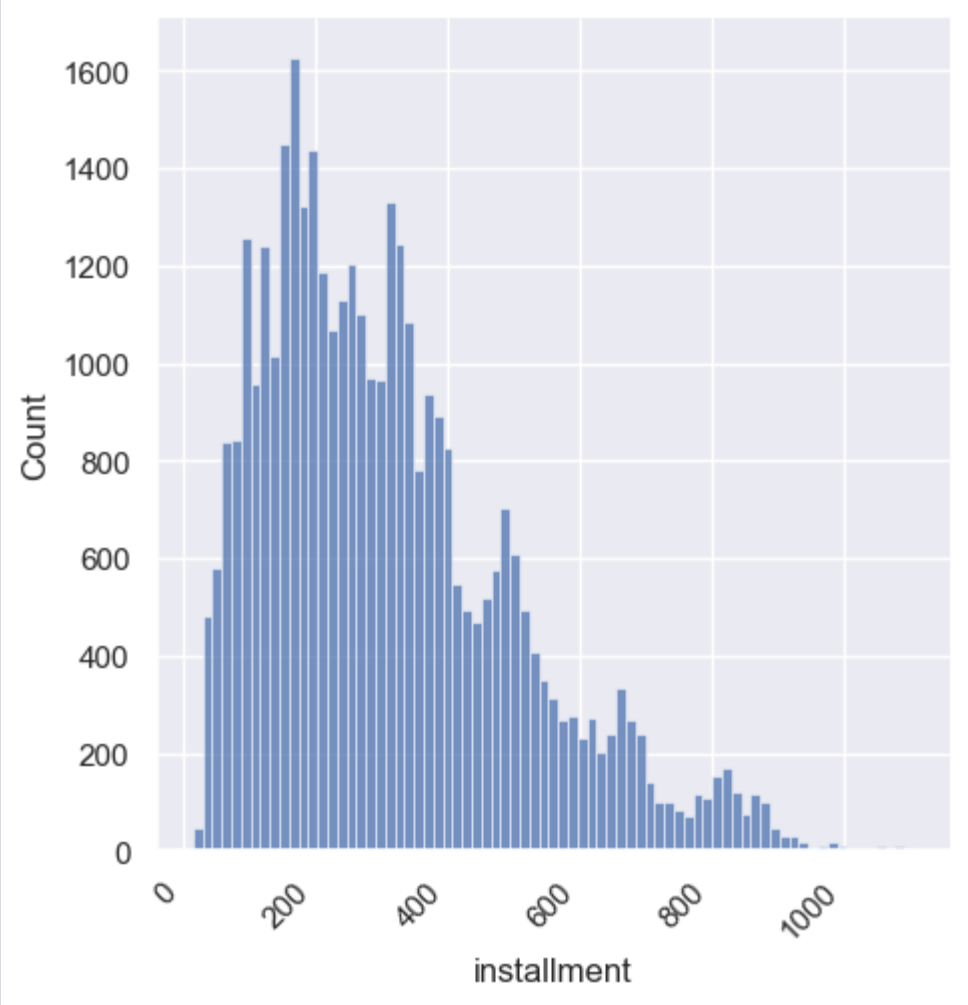
State

- addr_state: Few of the state like “CA” ,”FL” and “NY” has the higher number of applicants compared to other states.



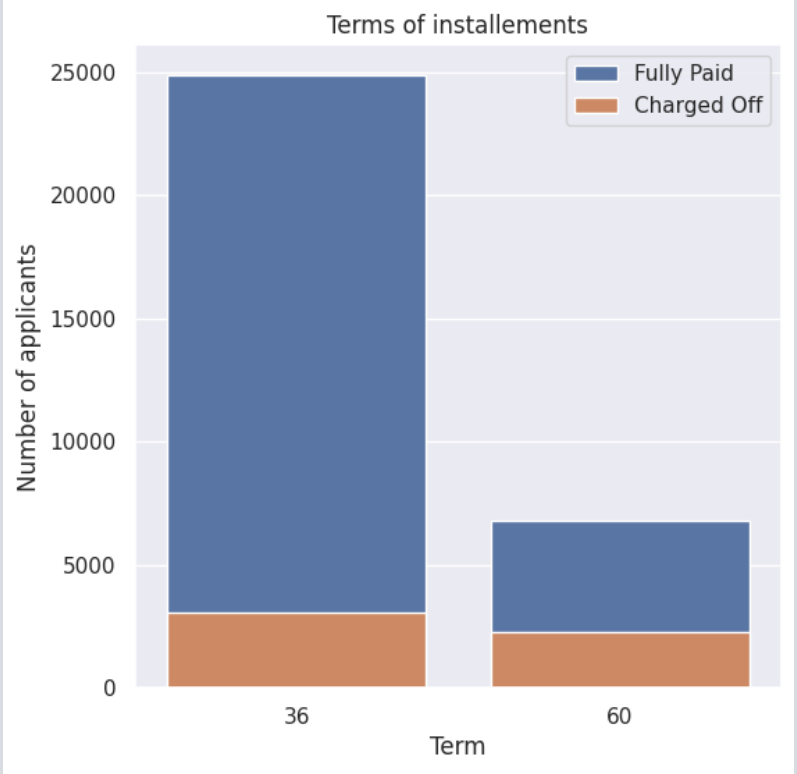
Installments

- Installment: Most of the installments are below 500 and he concentration is high in range of 200 to 400



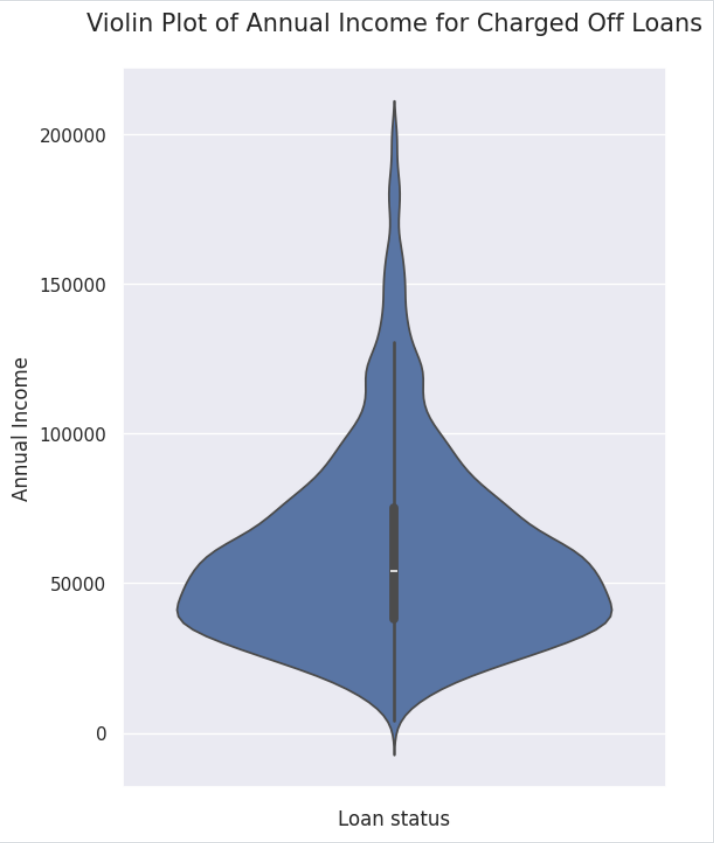
Terms of instalment

- Term: 60 months term is likely to be "Defaulted" more in proportion to 36 months term.



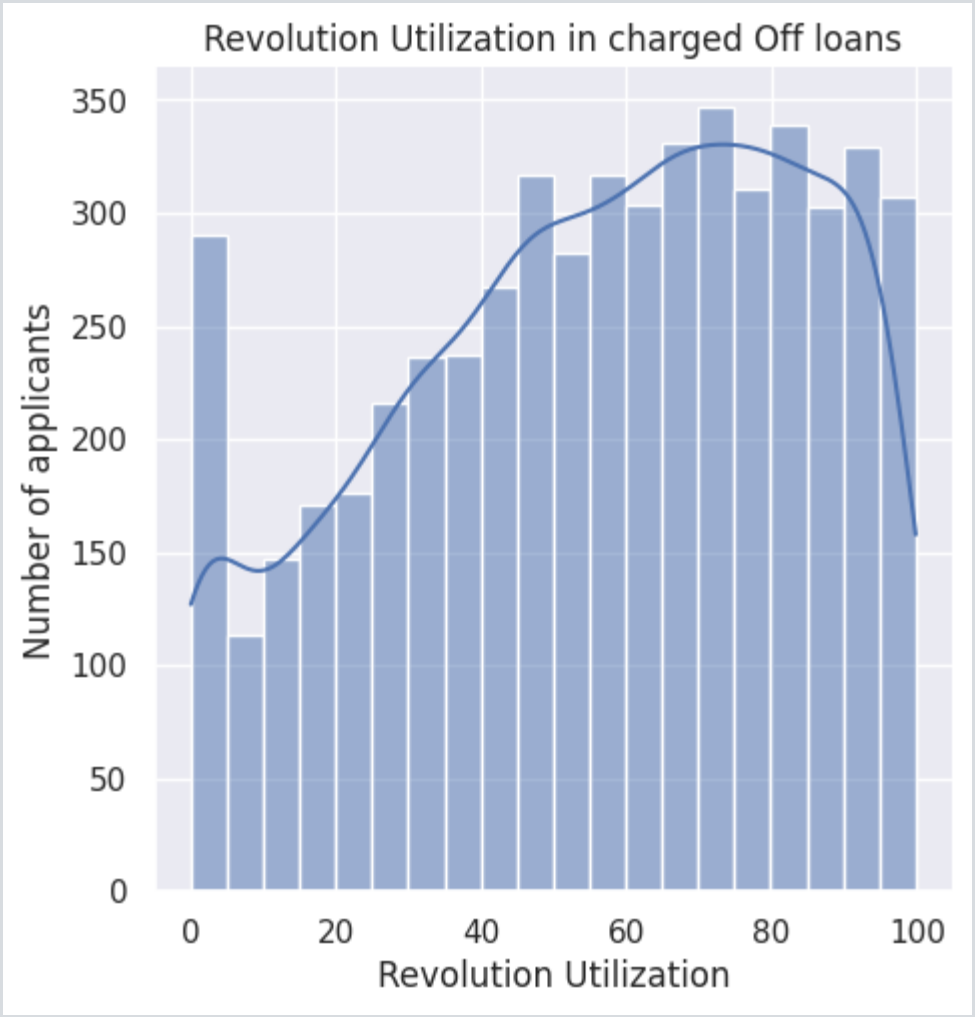
Annual Income

- Annual_income: After outliner treatments the annual income major at 40000-50000.



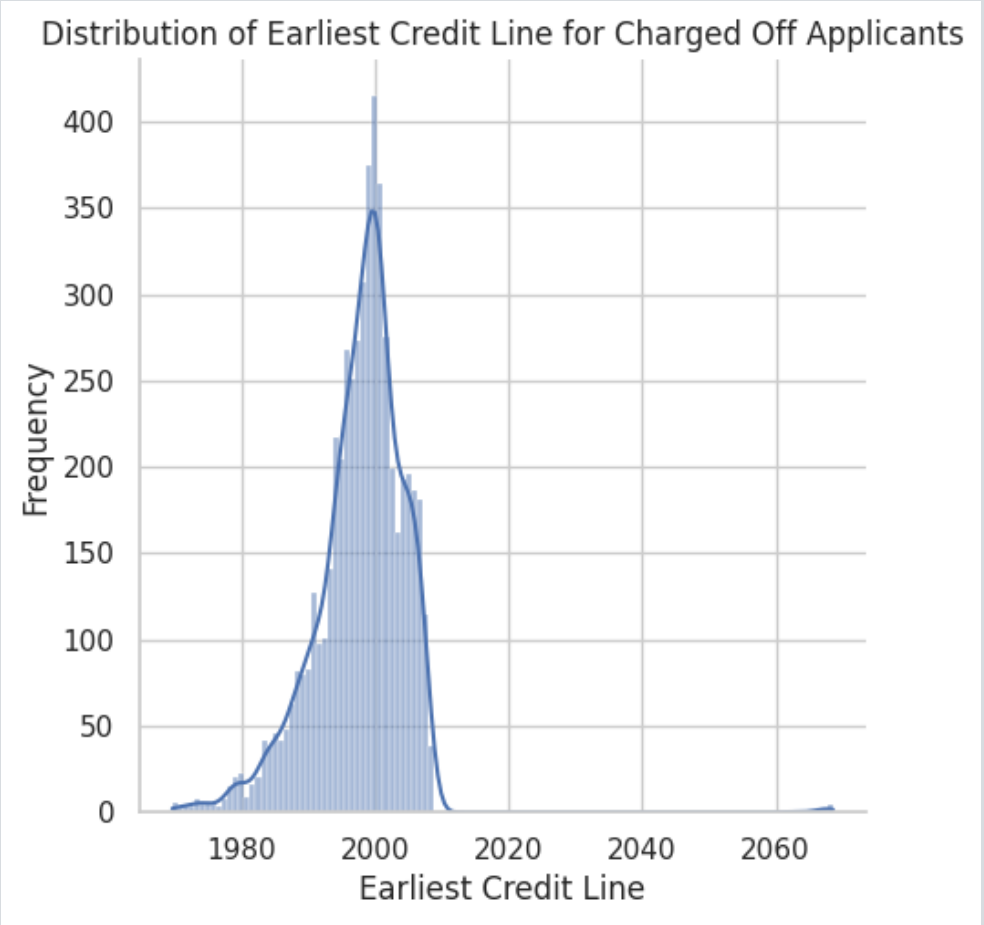
Revolution Utilization

- It is the percentage of available credit that is being used on revolving accounts.
- It's a key metric that can directly affect credit scores and shows how responsibly credit is being managed.
- **Insight**: People with higher revolution unitization tend to defaults more.



Earliest Credit line

- It is the earliest date a applicant had credit line attached to his name.
- **Insight**: People who having earliest credit line between 1990 and 2000 tend to show more Default Rate.

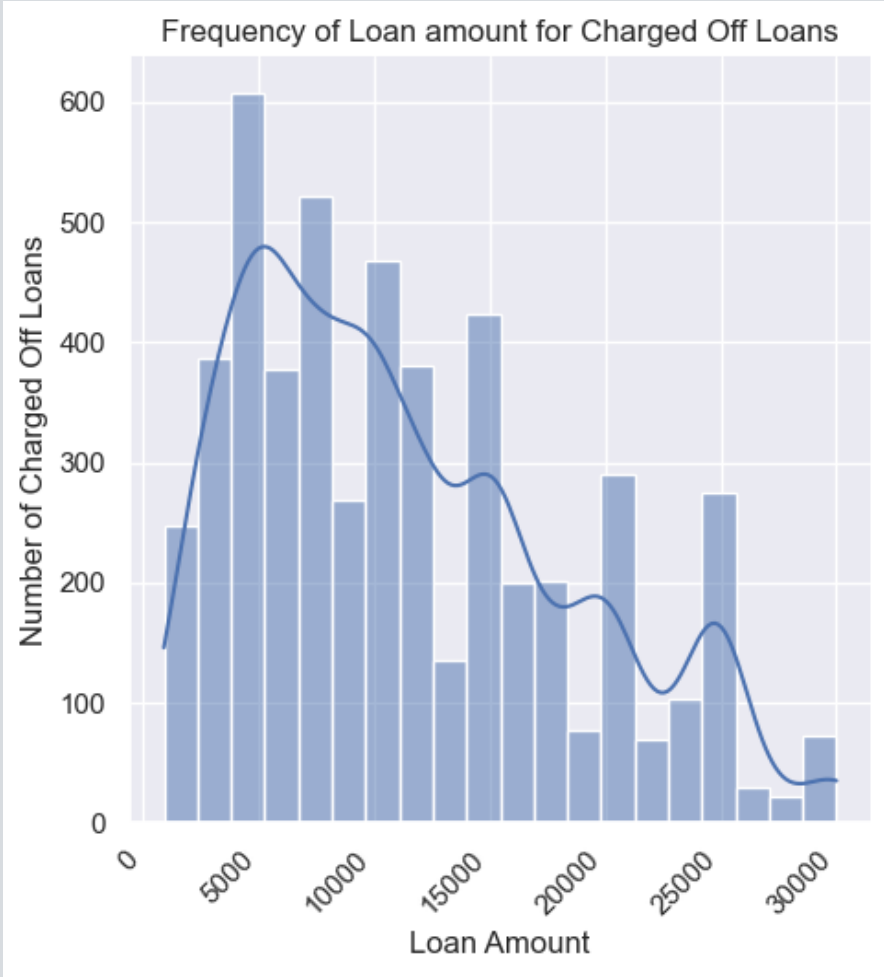


Bivariant Analysis

- A Bivariant Analysis was carried out to establish a correlation between various attributes to the status off loan. This analysis is important to understand the category for which the charge off was higher and to derive the conclusion
- The following section provides the details of results obtained from Univariant analysis

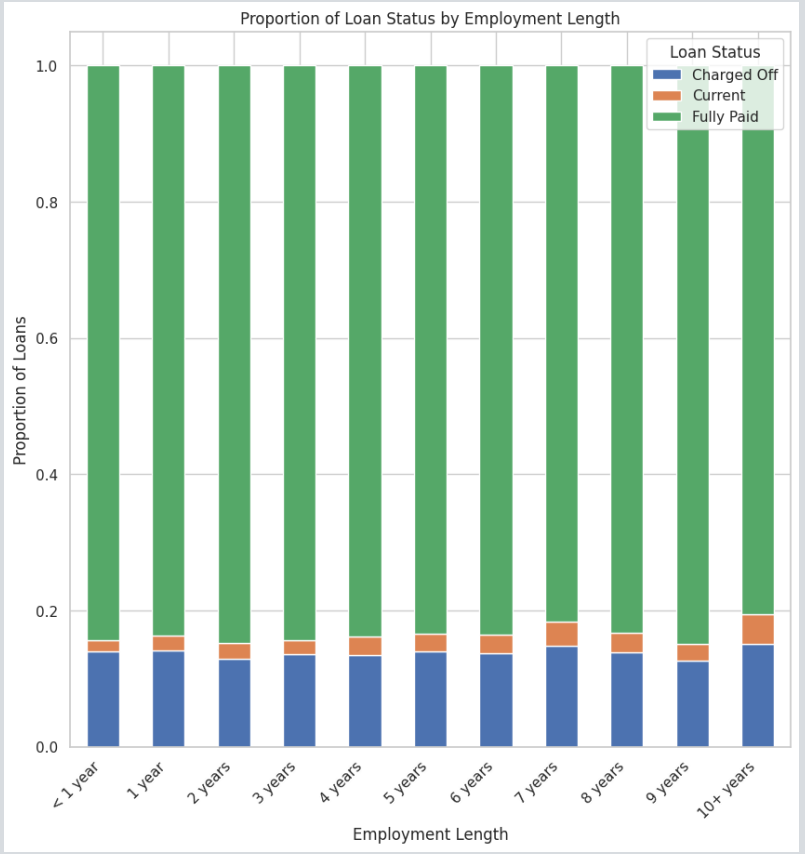
Loan Amount Freq.

- Charged off loans w.r.t. loan Amount: The data indicates that there is decreasing trend in number of charged off loans as the loan amount increases. The maximum charged off loans are for amount within range of 5000-10000.

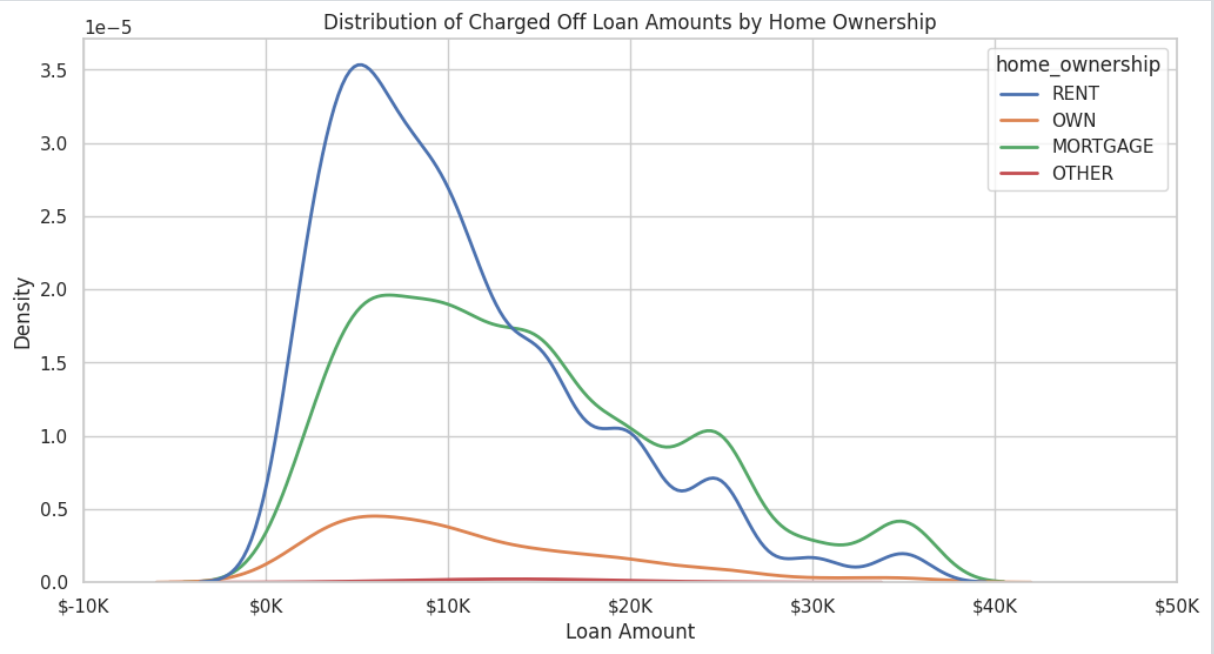


Employee Experience

- This plot shows that proportion of applicants in each employment length bar marked as charged off is pretty consistent.
- **Insights**: Employment length(Experience) alone can't be a deciding factor in lending loans



Home ownership VS Loan Amount

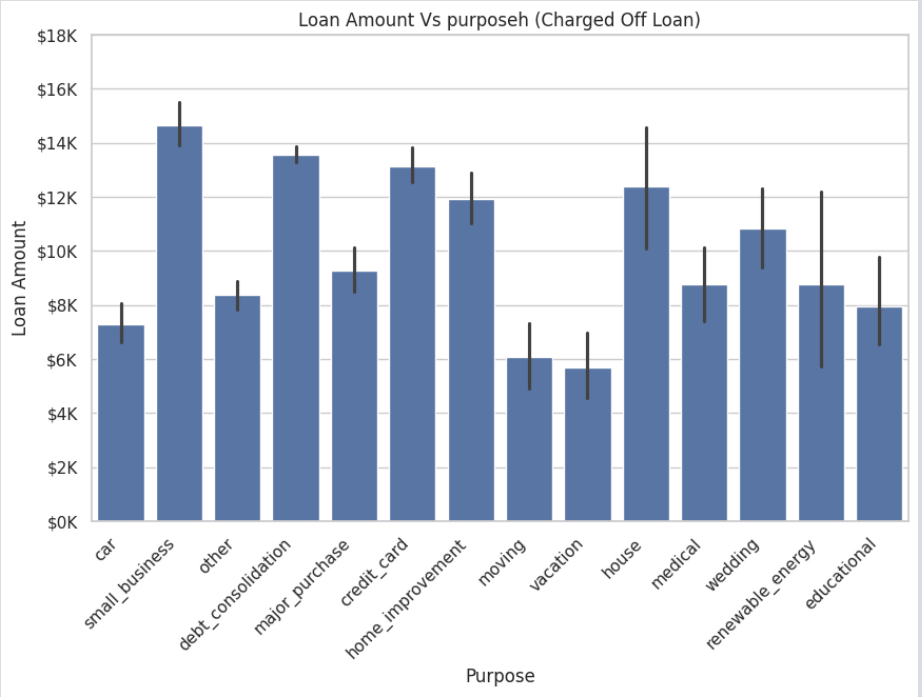


- Insight:
  - There are more people who are on RENT and taking loan of around $5k and being marked as "Charged Off"
  - After Loan amount of $15K there are more people with MORTGAGE and being marked as 'Charged Off'

Loan Amnt Vs purpose

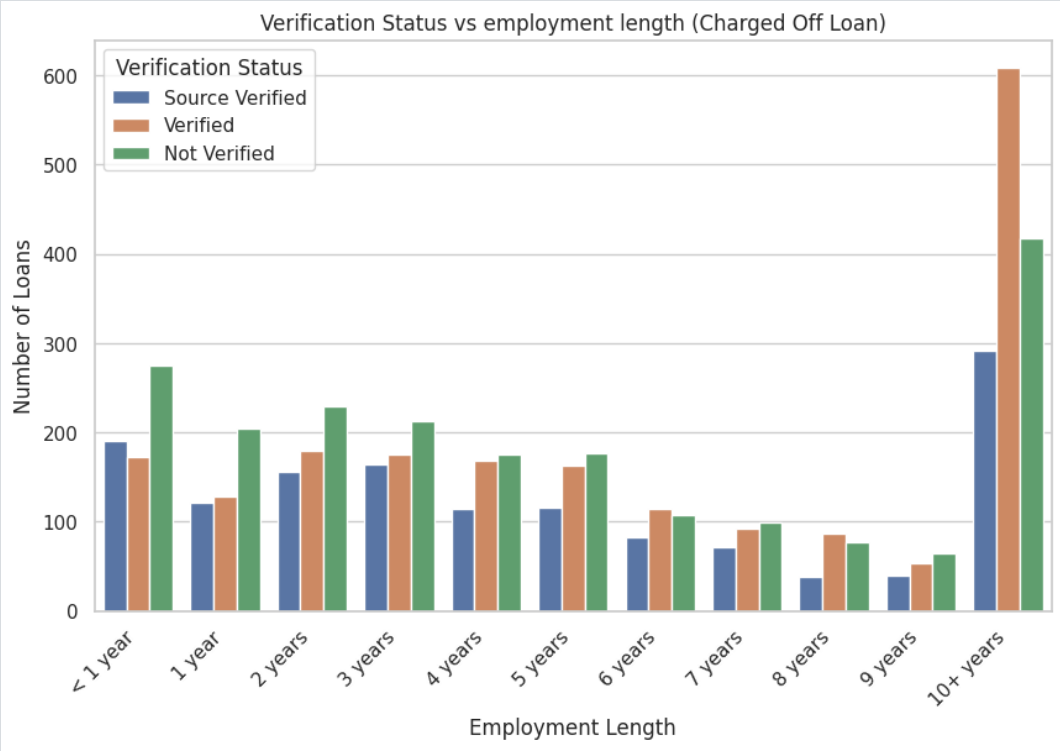
- Insight:

1. Loan of higher amount given to Small Business tends to be marked as 'charged Off'
2. This trend is followed by debt consolidation and credit card.

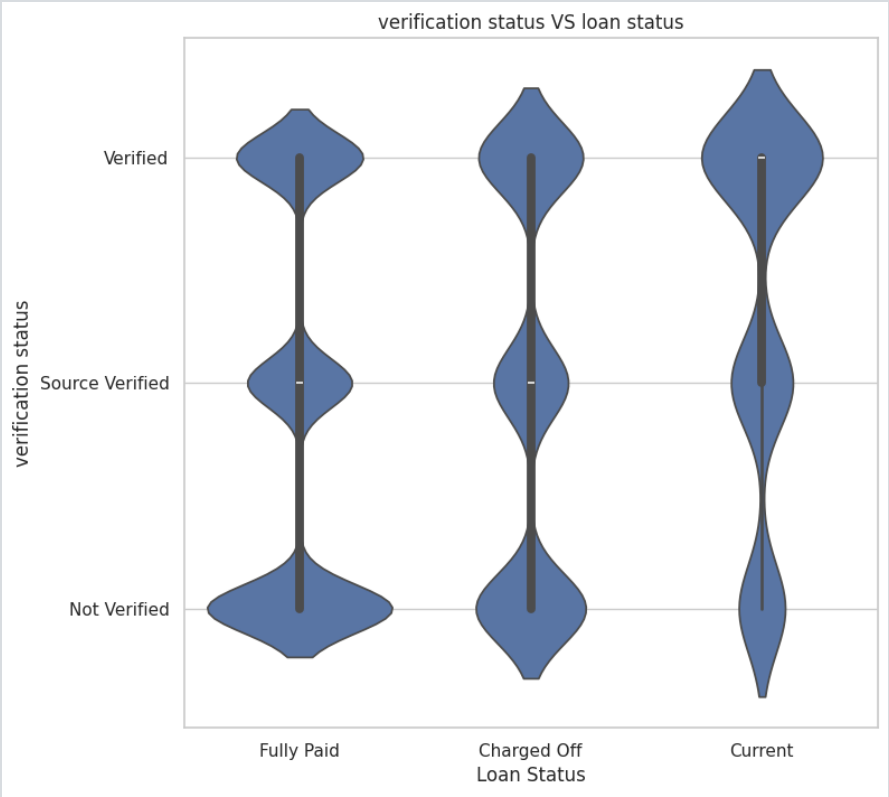


Importance of Verification

- **Insight**:
  1. There are more Not Verified applicant with experience of less then 3-4 years and marked as 'charged off'.
  2. Except less than one year experience, rest all categories are having less count in Source Verified, So to minimize credit loss company should do more source verification.
- **Company should lessen the number of un-verified applicant's acceptance rate.**



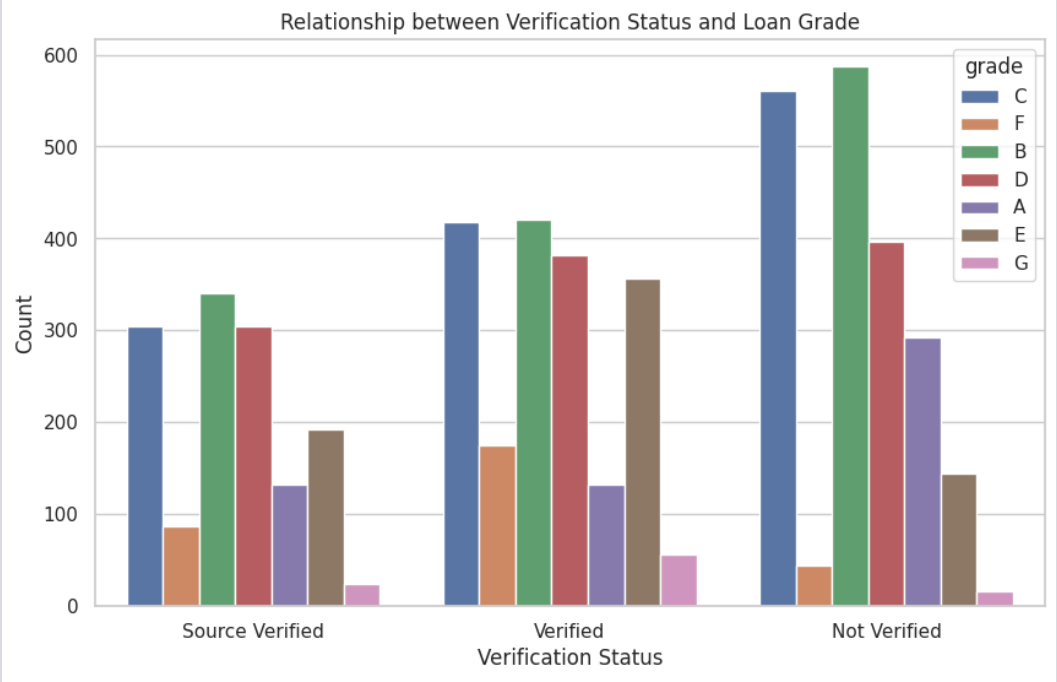
Verification Vs Loan status



- As we have seen in previous slide, here too we can see how source verified are less in charged off category.

Grade-wise verification status

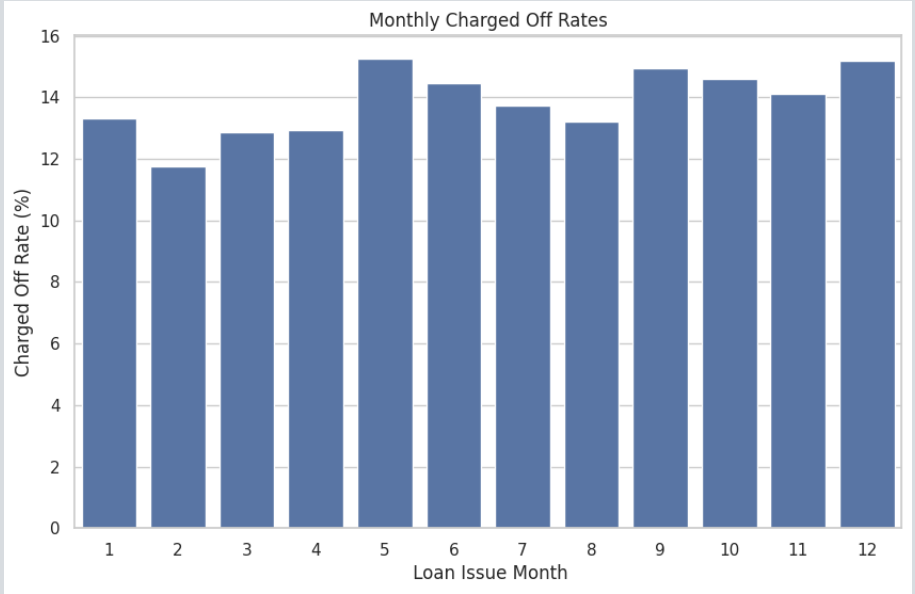
- **Insight:**
- People with grade B & C are trusted by company without any verification done, and they are the one who are being defaulted more.
- Company shouldn't blindly trust anyone’s grade and lend them loan.



Monthly Charged Off Rate

- **Insight:**

1. Month of may shows most Charged Off rate of about 15.31%
2. December stands in this race with 15.17%

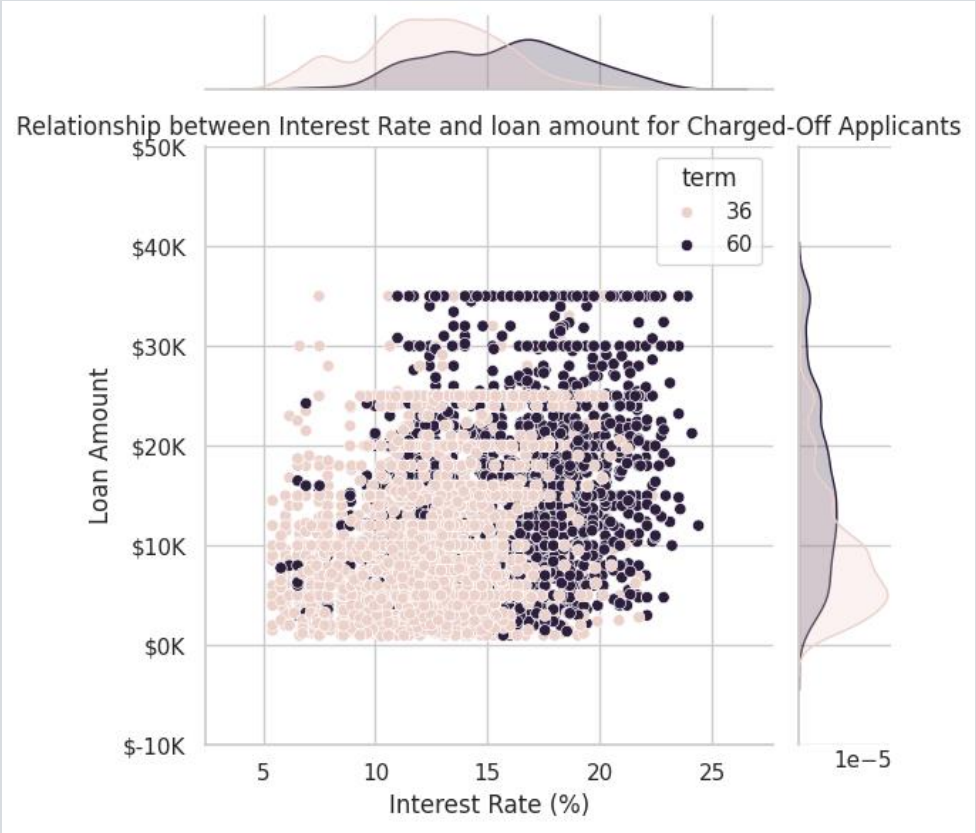


Multi-variate Analysis

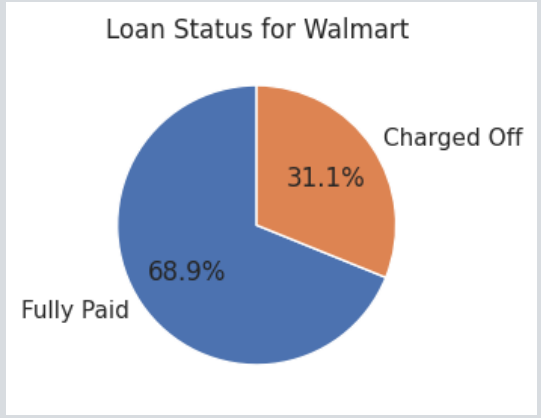
Relationship between Interest Rate and loan amount for Charged-Off Applicants

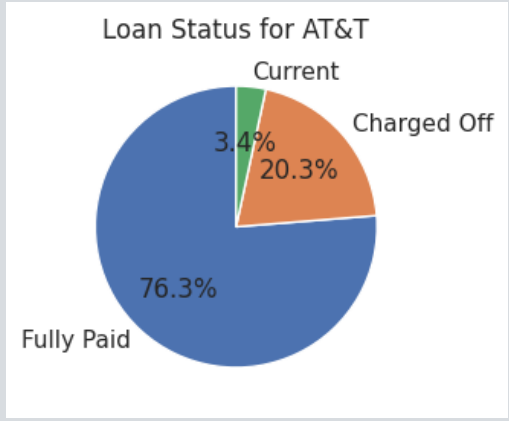
Insight:

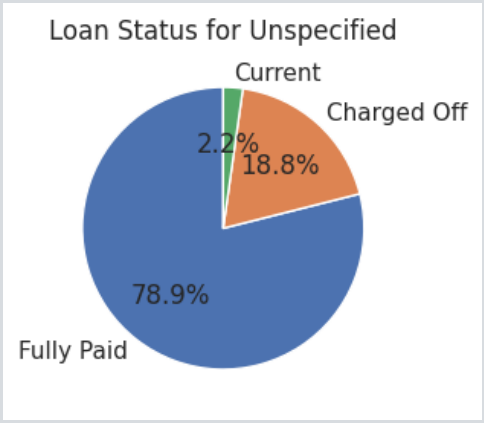
1. People with 60 term and having higher interest Rate than 10% tend to be defaulted more.
2. 60 months term having loan more between 10k-20k and having 35k are tend to be defaulted more.
3. Poeple having 36 months term having loan amount less than 10k tend to be defaulted more.



Loan Status of Top 3 employee Title







Insight:

- These are top 3 employee title which have highest number of default rates.

Conclusion

- Longer loan terms (60 months) and higher interest rates increase the likelihood of defaults.
- Non-homeowners and loans for small businesses or debt consolidation show higher default risks.
- Loan amounts between $5,000 and $10,000 have the highest charge-off rates.
- Unverified applicants and those with grades B & C default more frequently.
- Revolving credit utilization is a key indicator of default risk.
- Employment length alone is not a reliable factor for predicting defaults.
- Seasonal spikes in defaults are observed in May and December.
- Recommendations include enhancing source verification, tightening approval criteria for high-risk loans, and implementing targeted strategies based on applicant profiles.