Lending Club Case Study

By: Minh Ly + DEVESH KUMAR..

Overview

Problem Statement

You work for a consumer finance company which specialises in lending various types of loans to urban customers. 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

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    - **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Data Cleaning

01 No duplicate rows

02 There were 1140 rows with the loan status of "current" present; these have been removed since the loan status of "current" is not included in the study.

03 55 columns that had just null or blank row values and didn't contribute to the analysis have been eliminated.

04 The unique elements "url" and "member_id" have been eliminated.preserved "id" for analysis purposes in the future.

Data Cleaning

01 The text/description values for "desc" and "title," which do not participate, have been removed from the analysis.

02 Eliminating the subgroup since we were only able to analyze at the "Group" level.

03 By utilizing domain expertise, behavioral data is collected; as a result, it is not accessible for loan approval and is not involved in EDA analysis. => 21 behavioural data columns has deleted.

04 The unique elements "url" and "member_id" have been eliminated.preserved "id" for analysis purposes in the future.

Data Cleaning

01 Eight columns with values of 1 have been removed from the study due to their uniqueness.

02 Two columns that had more than 50% of the data were eliminated because they were unnecessary.

03 38577 rows and 20 columns remain

Data Conversions vs Derived Columns

- The 'term' column's excess string value has been removed and changed to an int data type.
- The string "int_rate" has been transformed to an integer. Extra '%' has been removed.
- The columns "funded_amnt" and "loan_funded_amnt" were made to float.
- The values in the columns loan_amnt, funded_amnt, funded_amnt_inv, int_rate, and dti are rounded to two decimal places.

Data Conversions vs Derived Columns

- Deriving two columns from "issue_d," "issue_year" and "issue_month," which will be used in further analysis.
- For improved analysis, the derived columns "loan_amnt_b," "annual_inc_b," "int_rate_b," and "dti_b" (multiple bucket type of data from continuous data) have been generated.

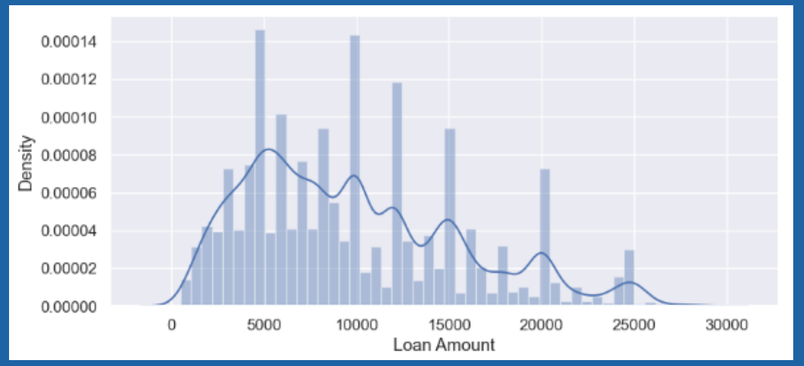
Dropping/Imputing Data

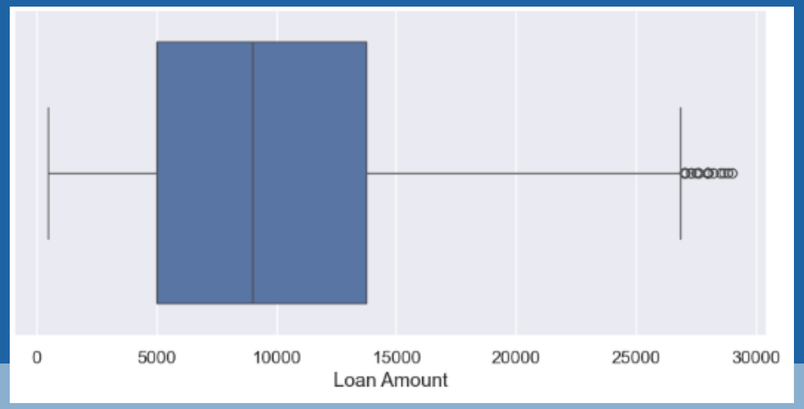
- 2.67% and 1.80% of the rows in "emp_lenght" and "pub_rec_bankruptcies" are null, which is a relatively tiny percentage of the data that may be dropped.
- 4.48% of rows were eliminated overall.
- For the numerical data sets "loan_amnt," "funded_amnt," "funded_amnt_inv," "int_rate," "installment," and "annual_inc," outliers leave.
- The quantile method has been used to address outliers in the aforementioned disciplines.

Univariate Analysis

Loan Amount

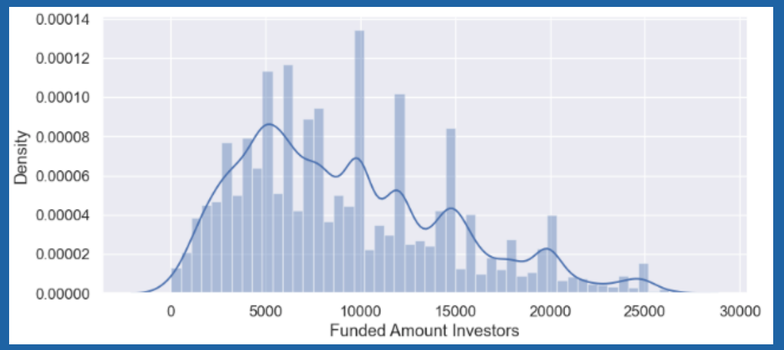
- The majority of the loan applications fell between $5,000 and $14,000.
- A maximum of $27,000 was requested for a loan.

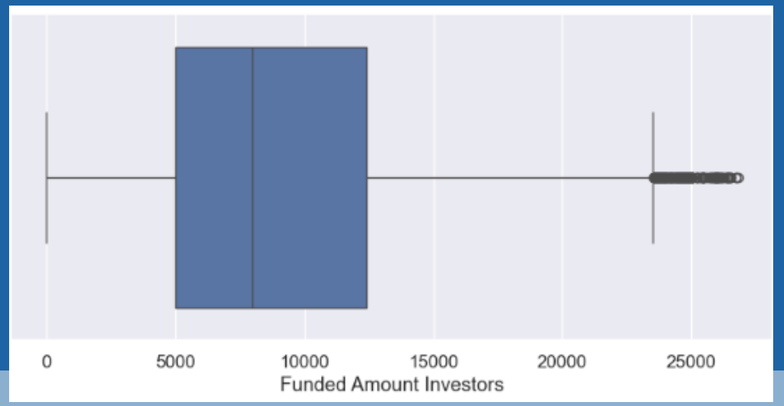




Total amount committed

- Total amount commited range from 5000 and 13000
- Average funded amount investor is: 9210

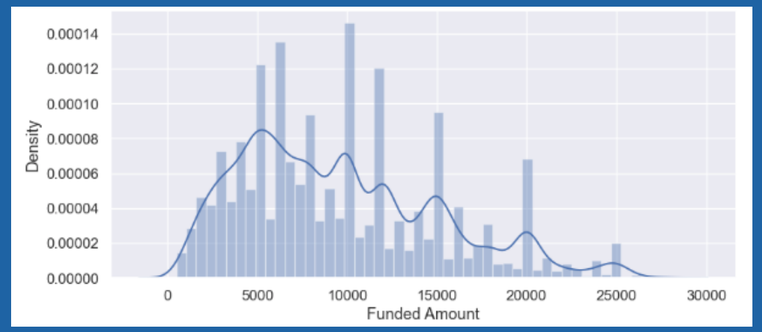


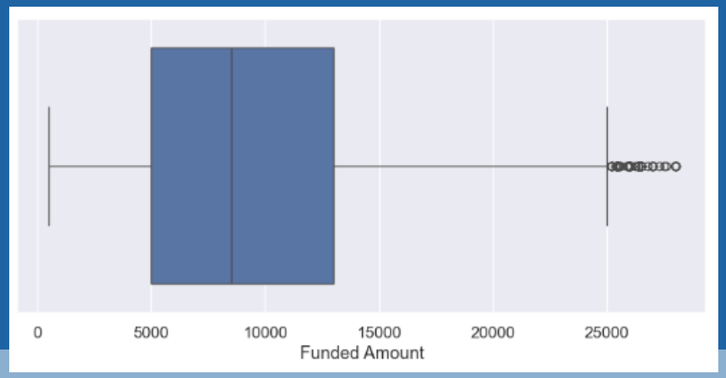


Univariate Analysis

Funded Amount

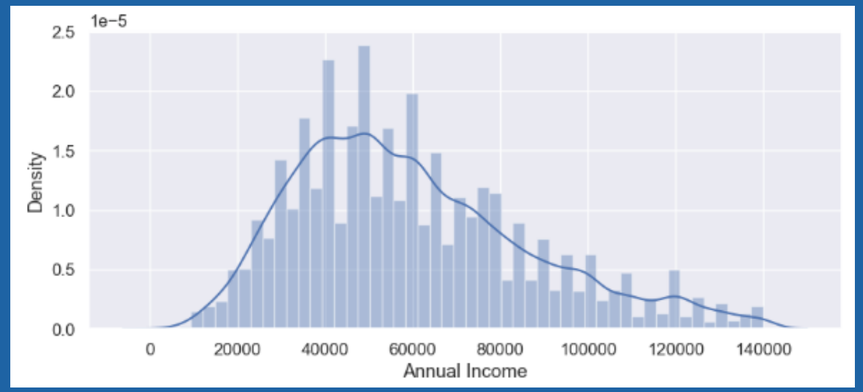
- Funded amount range from $5,000 and $13,000 or less.
- The yearly average income is: 9593

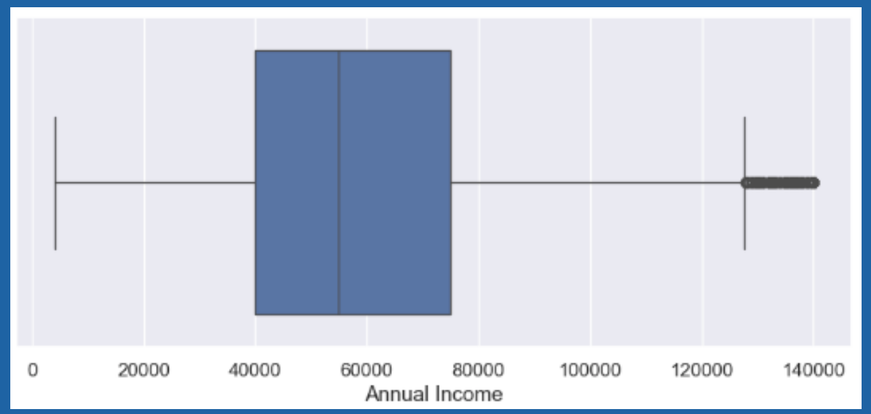




Annual Income

- Most applicants have an annual income of between $40,000 and $75,000 or less.
- The yearly average income is: 59883
- Annual Income shows left skewed normal distribution thus we can sa that the majority of burrowers have very low annual income compared to rest.

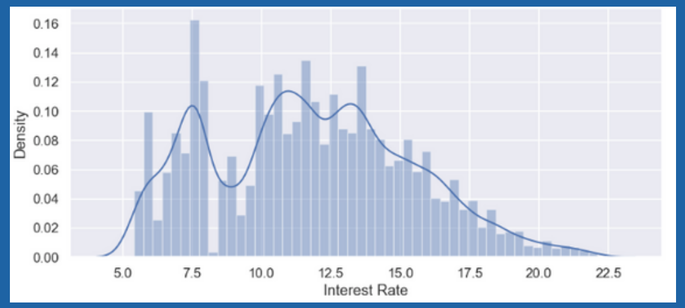


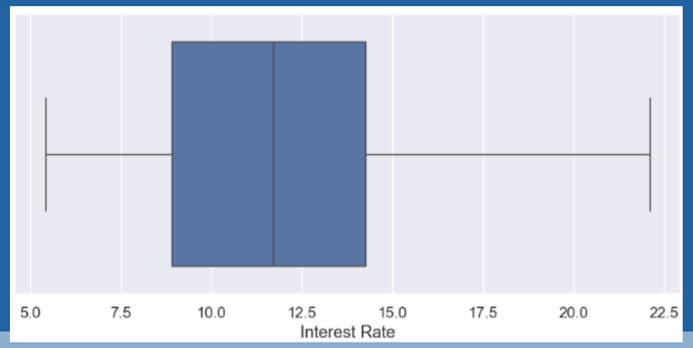


Univariate Analysis

Interest Rate

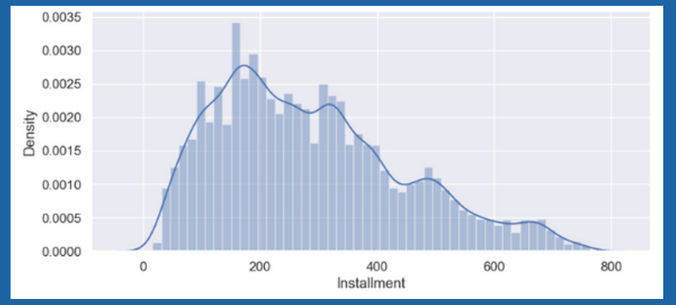
- The majority of the applicant's interest rate falls between 8% and 14%.
- Average Rate of interest of rate is 11.7 %

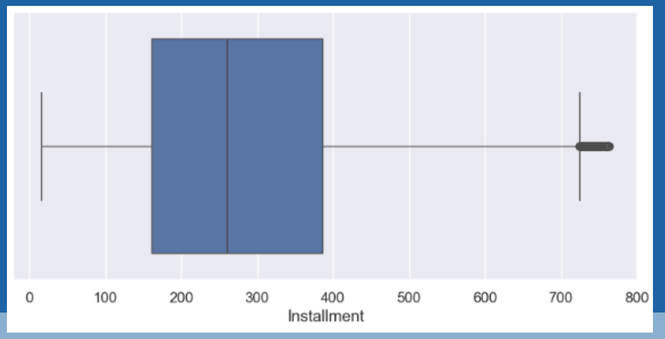




Installment

- The distribution is right-skewed, with most of the values clustered around 200 to 400.
- Average installment: 286

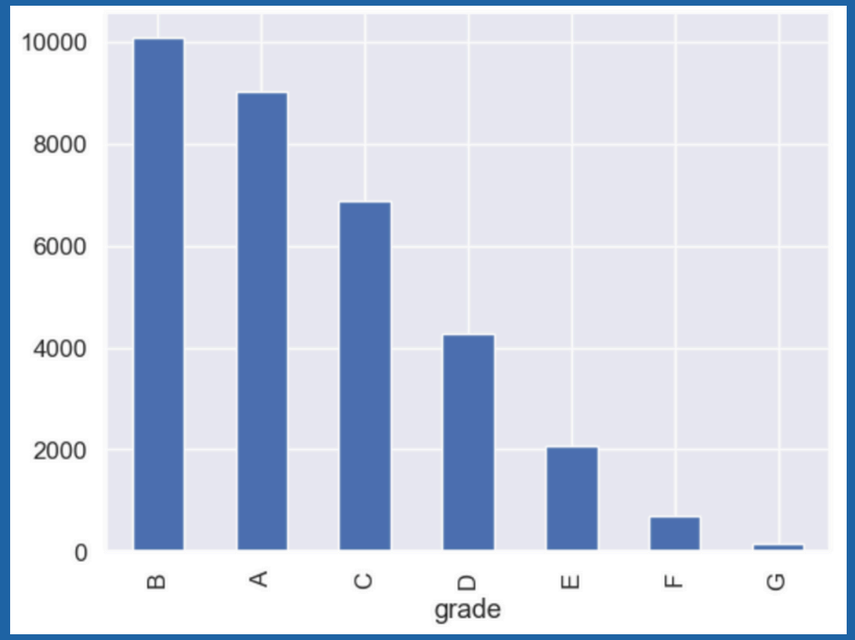




Segmented Univariate Analysis

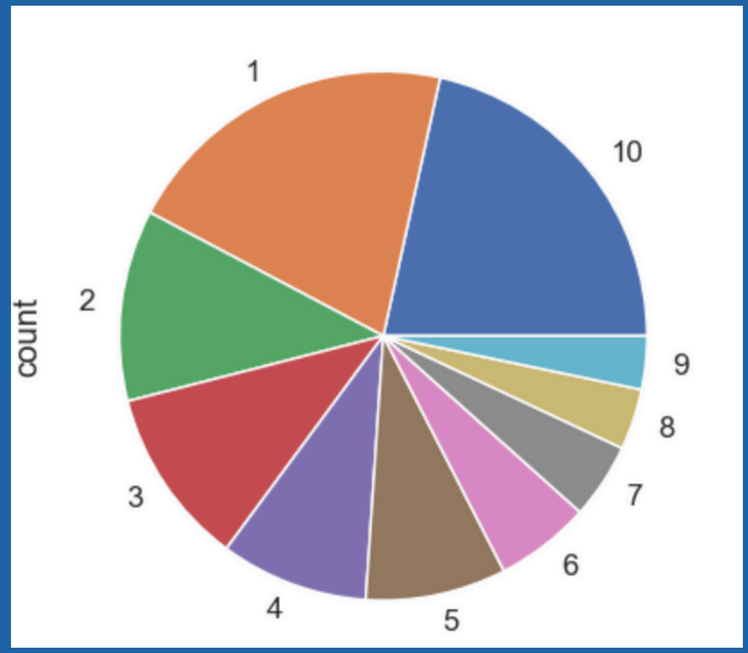
Grade

- Majority of people are in grade B and A



Employment Length

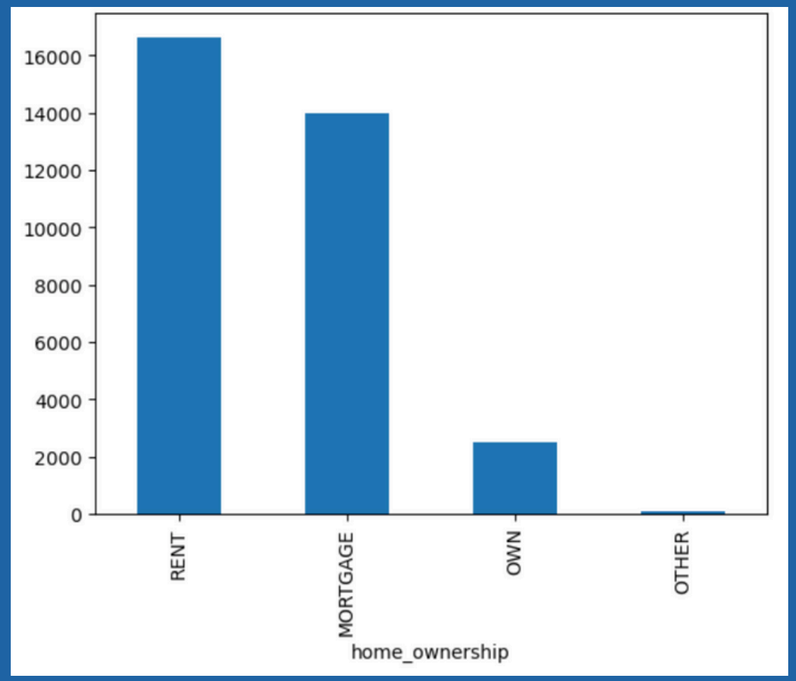
- Majority of borrowsers have working experience greater than 10 years.



Segmented Univariate Analysis

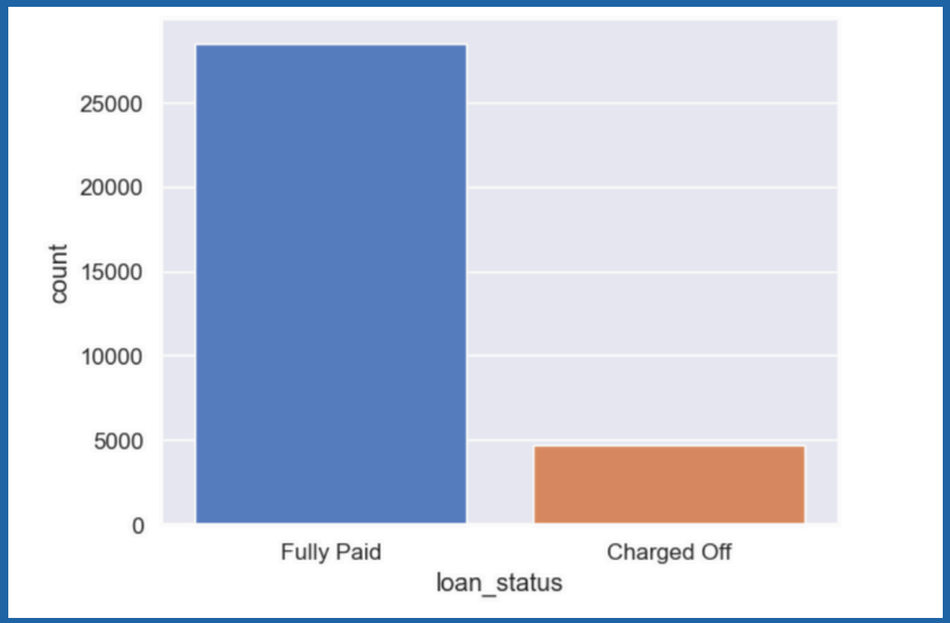
Home Ownership

- Major people don't own a property, and are on rent or on mortgage.



Loan Status

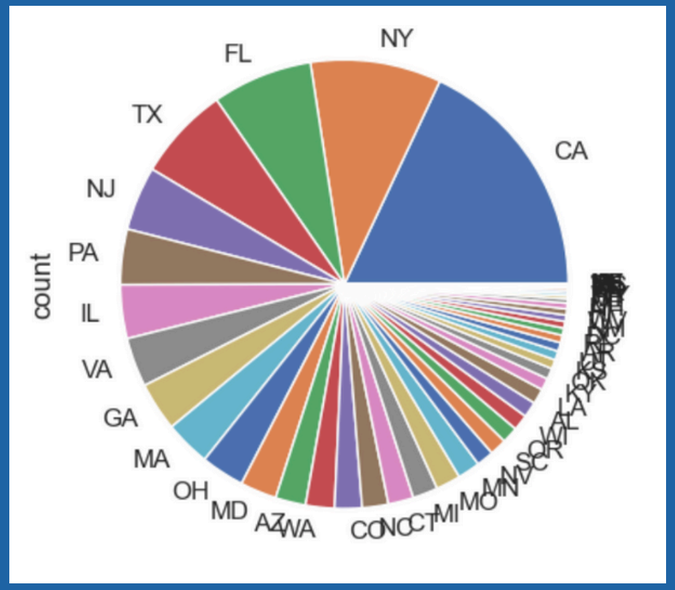
- Majority of candidates fully paid the loan, one-sixth of them charged off.



Segmented Univariate Analysis

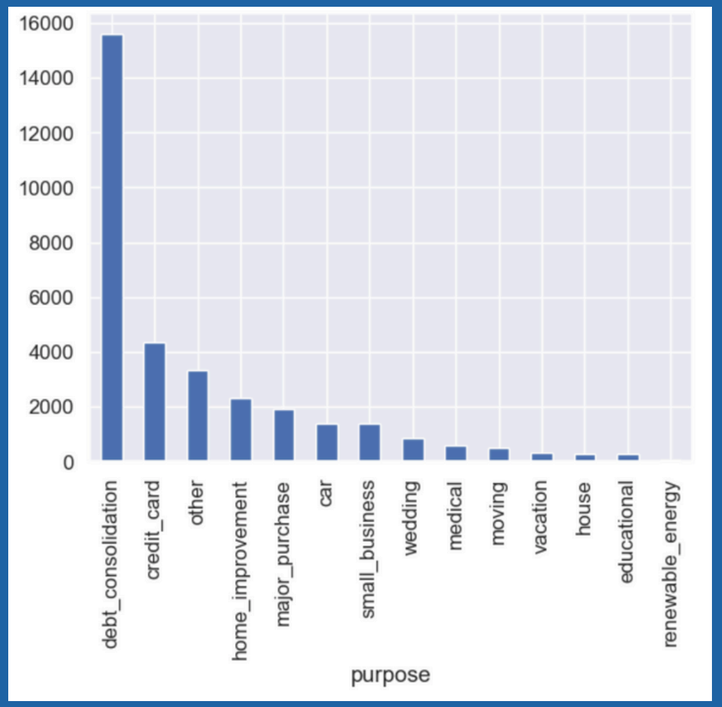
State

- Most of the loan applicants are from California



Purpose

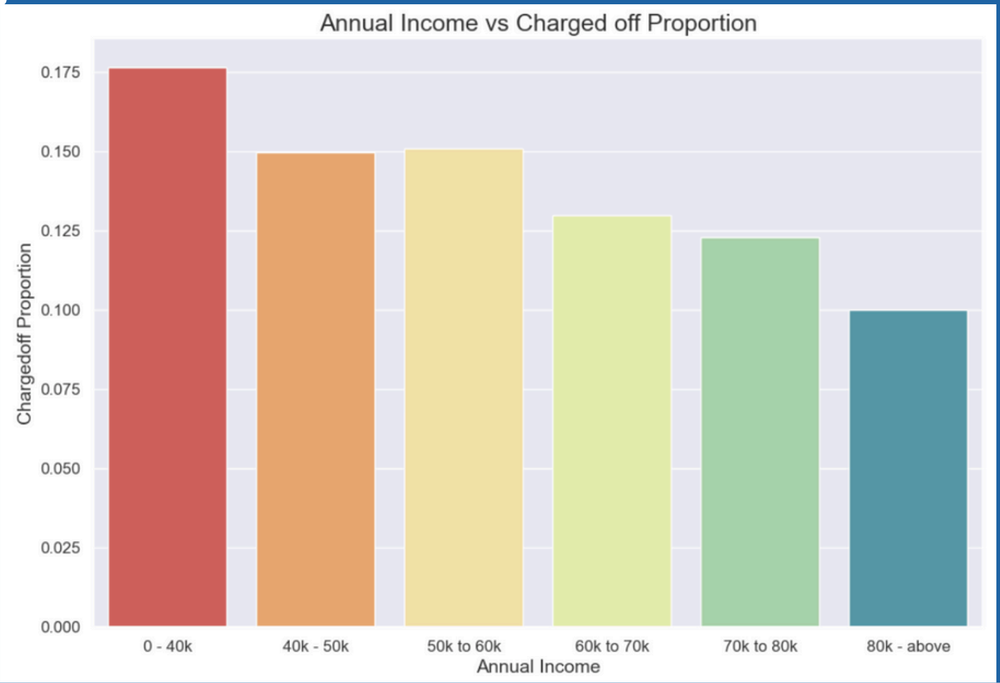
- Most of the loan applicants are for debt_consolidations.



Bivariate Analysis

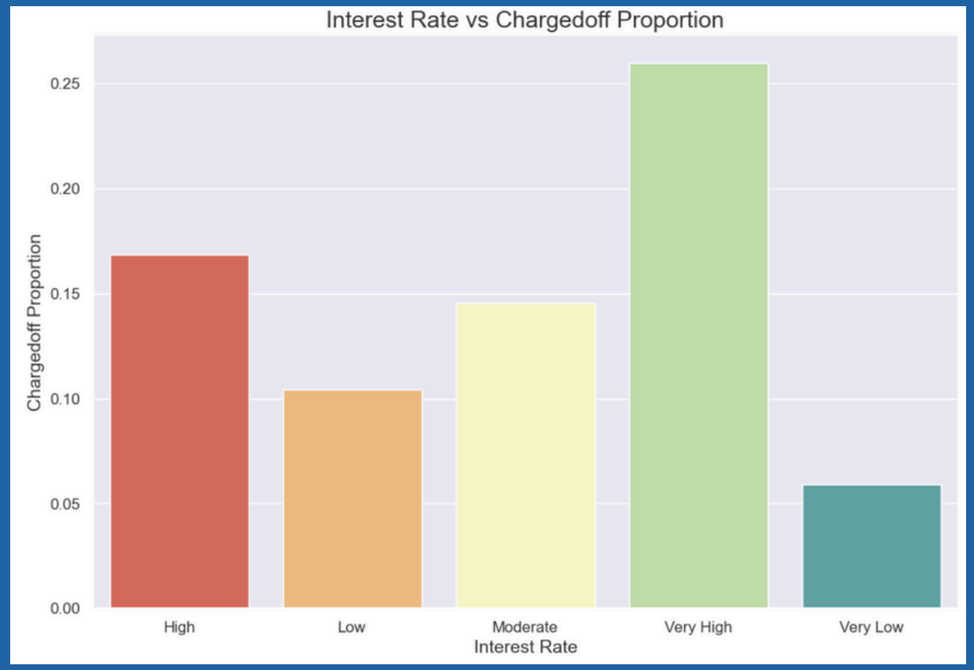
Chargeoff_Proportion vs Annual Income

- Income ranges above $80,000 have a lower likelihood of being charged off.
- Income ranges between $0-$20,000 have a higher likelihood of being charged off.
- As annual income increases, the proportion of charged-off loans decreases.



Chargeoff_Proportion vs Interest Rate

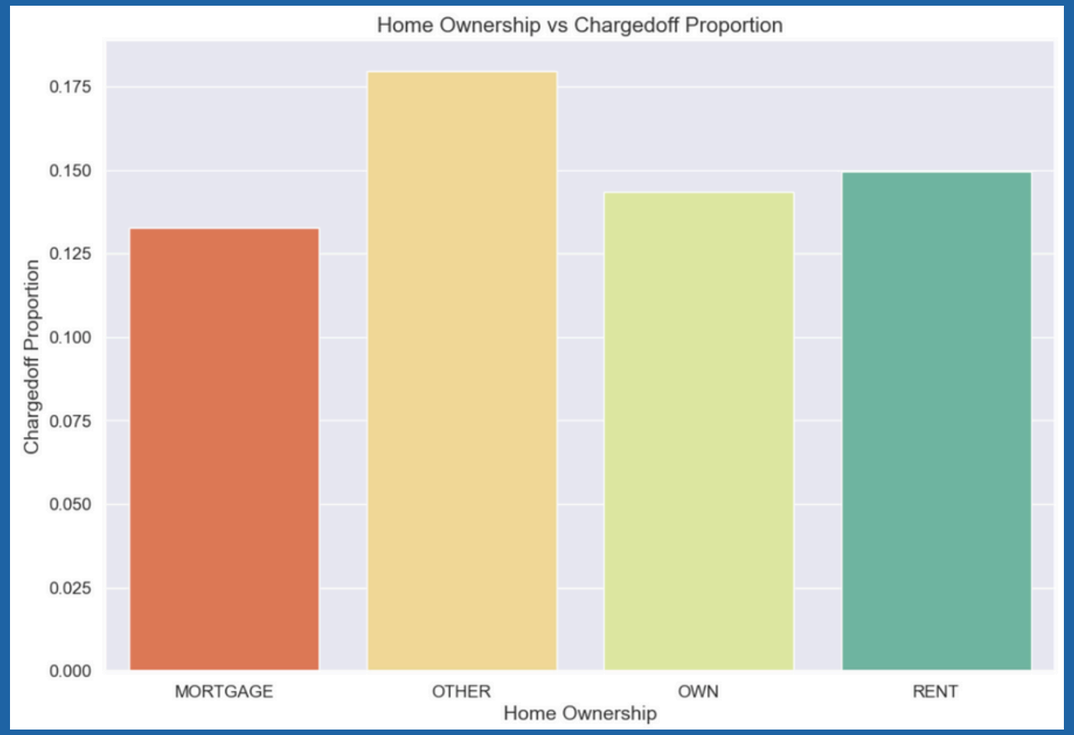
- Interest rates below 10% show a minimal likelihood of being charged off, with rates starting as low as 5%.
- Interest rates above 16% have a higher likelihood of being charged off compared to other interest rate categories.
- The proportion of charged-off loans increases as interest rates rise.



Bivariate Analysis

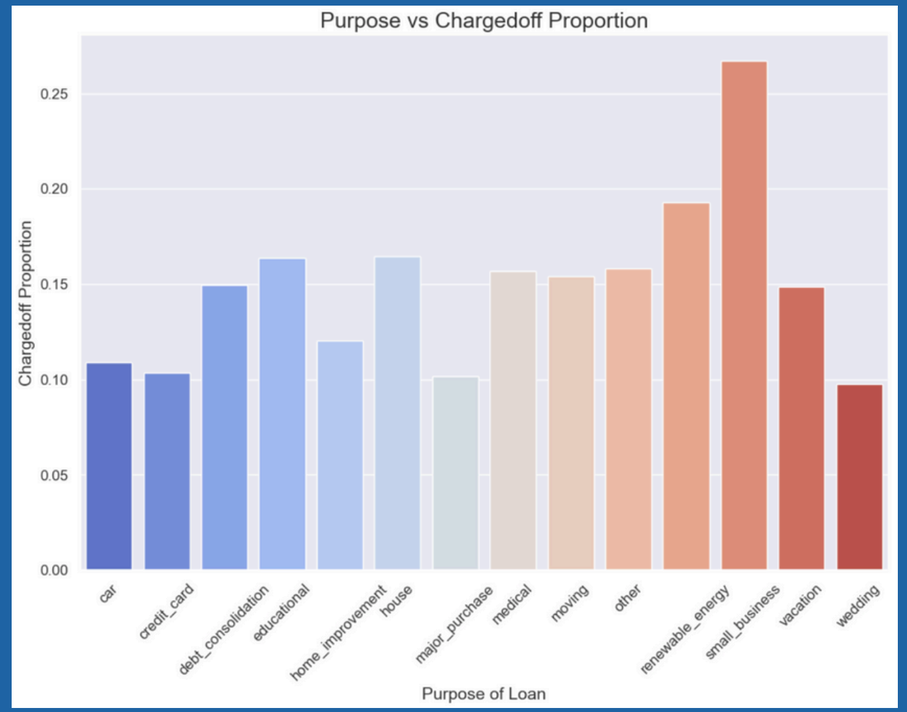
Chargeoff_Proportion vs Home Ownership

- Those who are not owning the home is having high chances of loan defaults.



Chargeoff_Proportion vs Purpose

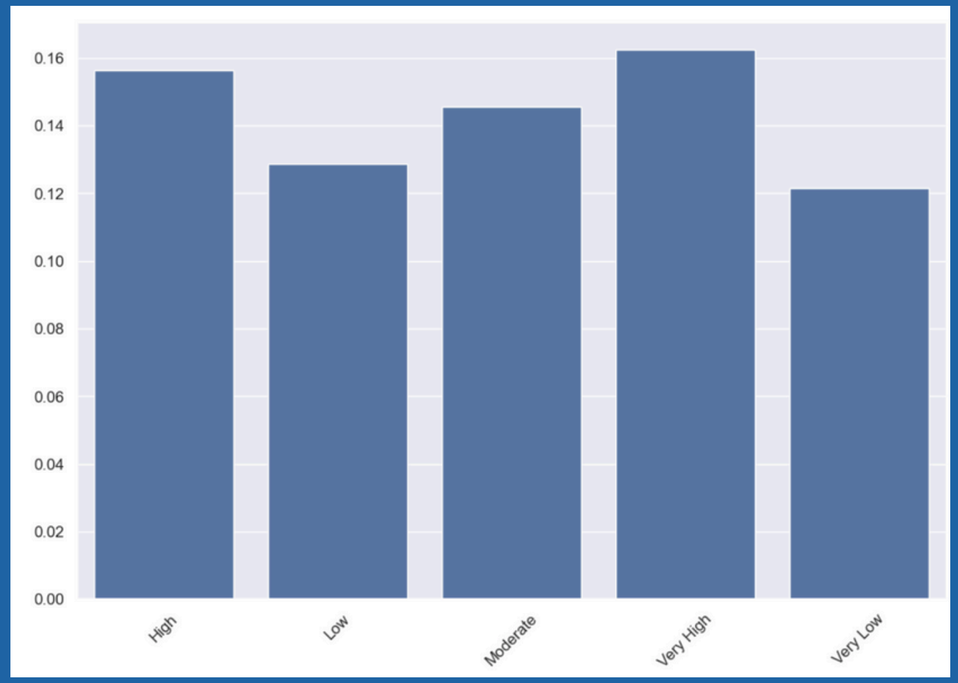
- Applicants with a home loan tend to have a lower likelihood of loan defaults.
- Applicants with loans for small businesses are more likely to default on their loans.



Bivariate Analysis

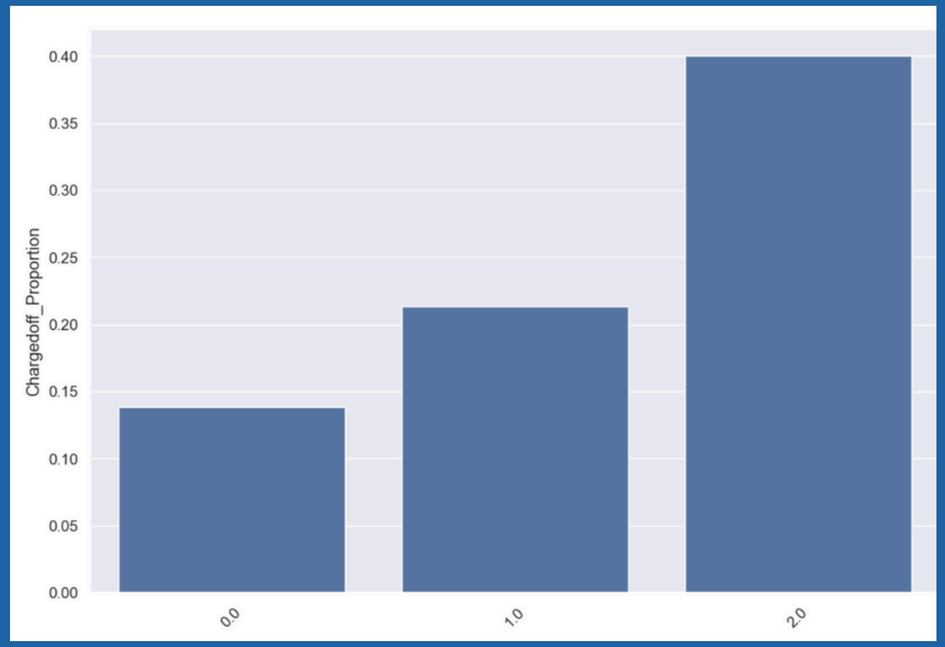
Chargeoff_Proportion vs DTI

- A high DTI (Debt-to-Income) ratio is associated with a greater risk of loan defaults.
- A lower DTI ratio corresponds to a reduced likelihood of loan defaults.



Chargeoff_Proportion vs Bankruptcies Record

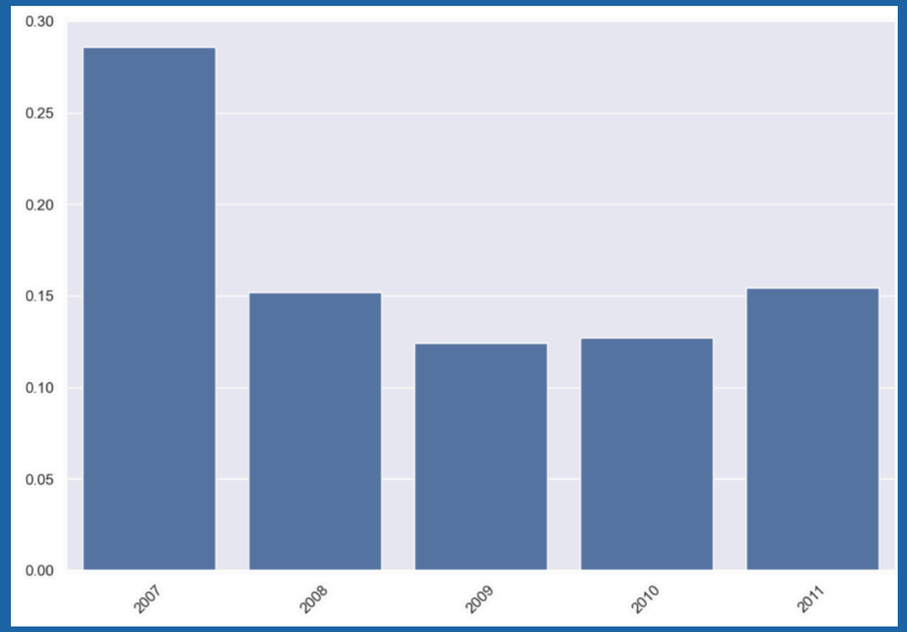
- A bankruptcy record of 2 significantly increases the likelihood of loan defaults.
- A bankruptcy record of 0 has minimal impact on loan defaults.
- Fewer bankruptcies correlate with a lower risk of loan default.



Bivariate Analysis

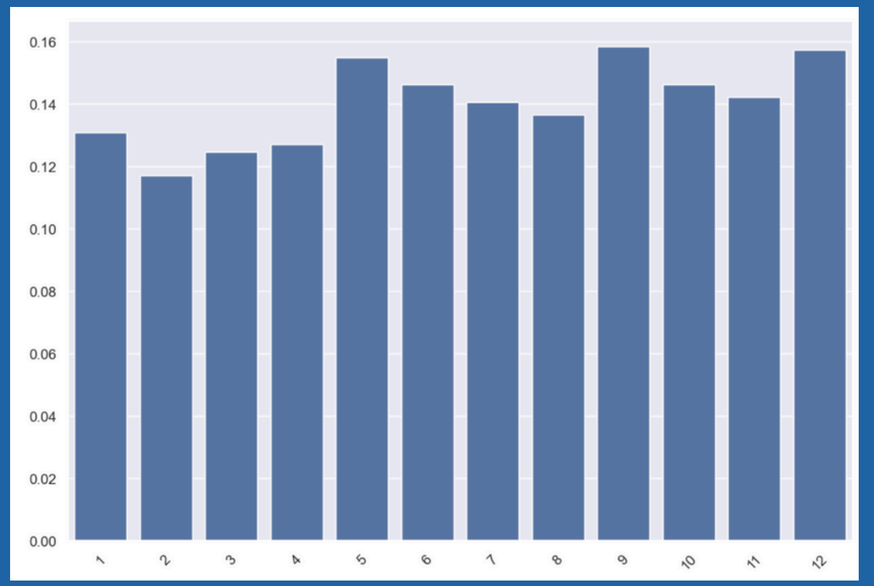
Chargeoff_Proportion vs Issues Year

- Year 2007 is highest loan defaults, 2009 is the lowest.



Chargeoff_Proportion vs Issue Month

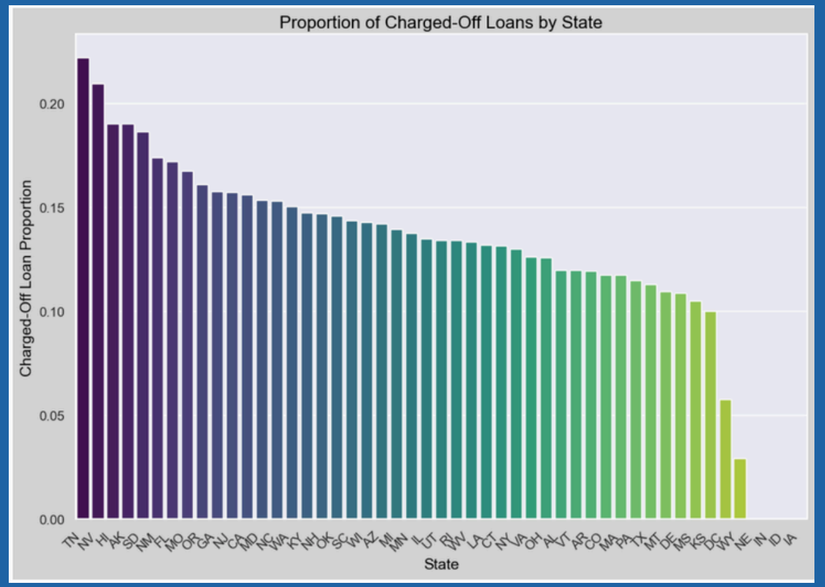
- Loans issued in May, September, and December show a high number of defaults.
- Loans issued in February also exhibit a significant number of defaults.
- Most loan defaults occur among applicants whose loans were approved between September and December.



Bivariate Analysis

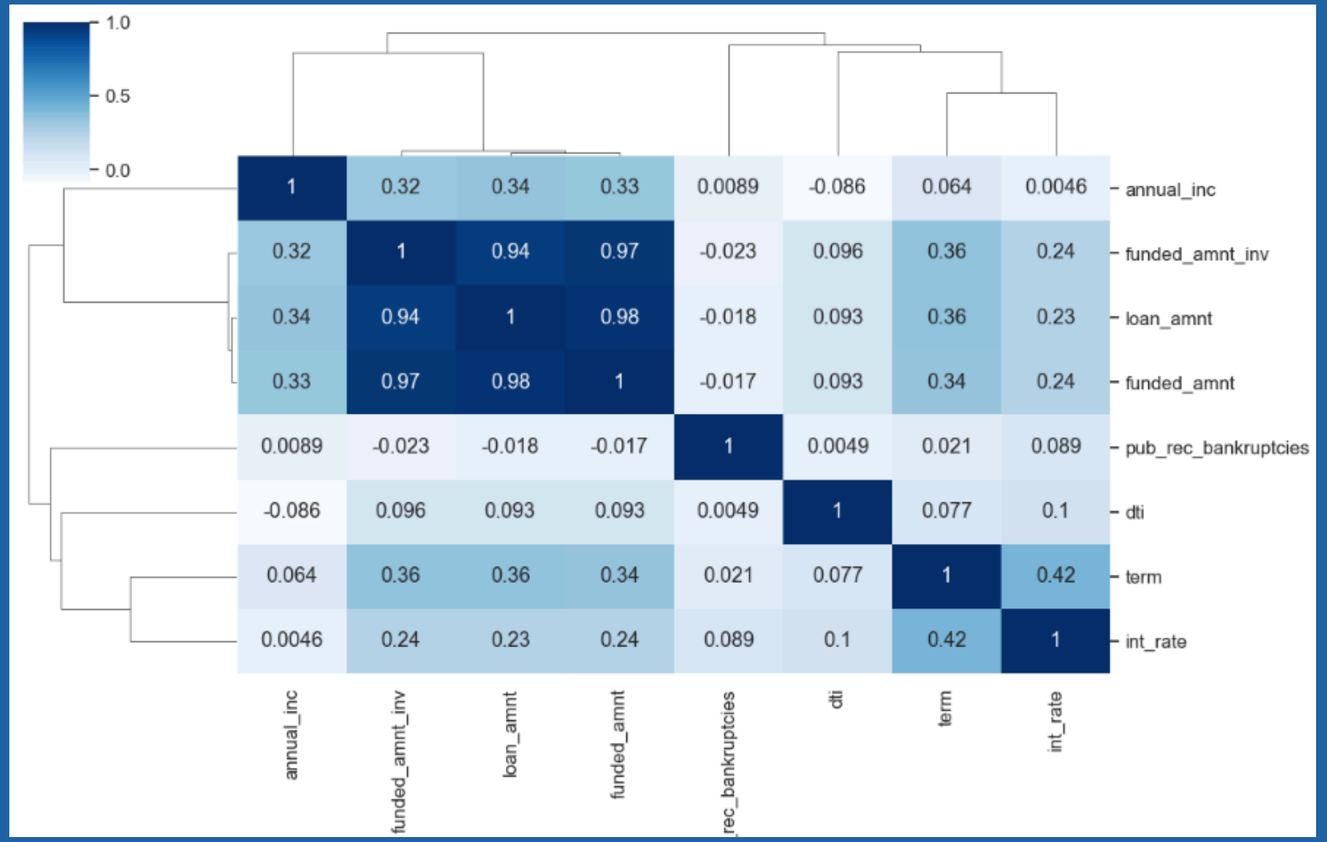
Charge off vs State

- Tennessee (TN) has the highest number of loan defaults.
- California (CA) has a lower number of loan defaults.



Correlation

- Loan amount is negatively correlated with public record bankruptcies.
- Annual income shows a negative correlation with debt-to-income ratio (DTI).
- Loan term has a strong correlation with loan amount.
- Loan term is also strongly correlated with interest rate.
- Annual income is strongly correlated with loan amount.



Conclusion & Recommendation

- The income range of 0–40,000 has a significant likelihood of being charged off.
- Compared to interest rates in other categories, an interest rate greater than 16% has a good likelihood of being charged off.
- The likelihood of a loan defaulter is higher for those without a home.
- There is a significant risk of loan default for candidates who have small business loans.
- A high DTI number indicates a significant default risk.
- The likelihood of loan defaults increases with the number of bankruptcies.
- The state with the most loan defaults rate is TN.
- The highest rate of loan defaults is among loan applicants with a grade of G.

Rec 01

Accept the loan for people with income above 50,000

Rec 02

Careful with TN state since it has the highest charge off rate.

THANK YOU!