Lending Club Case Study

How consumer and loan attributes influence the tendency of default

Problem Statement

Problem Statement

A consumer finance company which specialises in lending various types of loans to urban customers is trying to understand the defaulters and thus prevent bad loans.

Process : 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:

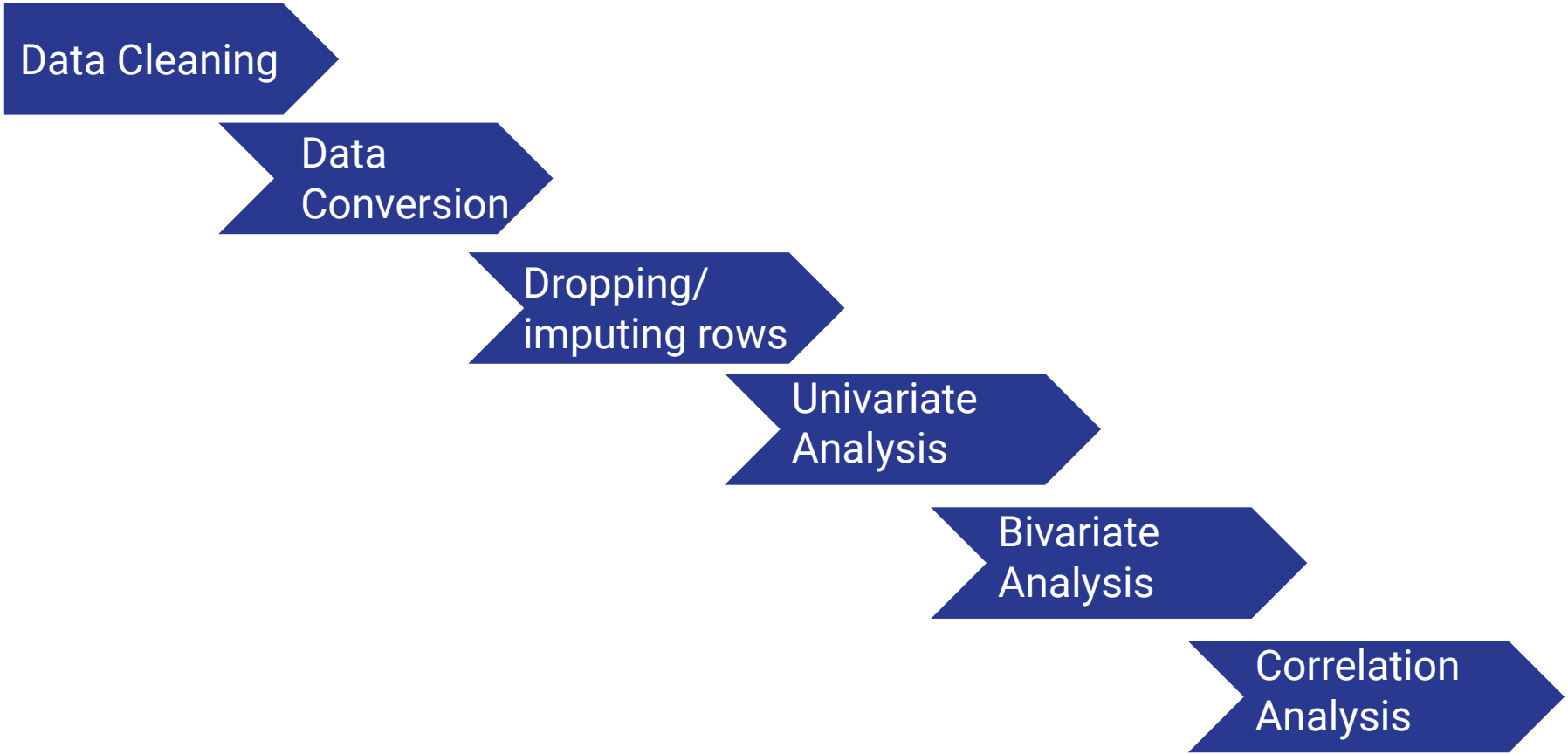
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

The data given contains information about past loan applicants and whether they 'defaulted' or not. The aim is 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.

There are three categories for the loans:

- 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
- 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)

Approach


Data Cleaning
Data Conversion
Dropping/ imputing rows
Univariate Analysis
Bivariate Analysis
Correlation Analysis

Data Cleaning

- **Removing unnecessary rows**
  - Remove headers, trailers/summary (Not present)
  - Remove duplicates (Not present)
  - Remove rows with loans status as current as they do not contribute to the analysis
  - 2.87% rows removed
- **Removing unnecessary columns**
  - Remove columns with only null or NA values
  - Remove text or descriptive fields which cannot be used for analysis
  - Remove sub-group level data as analysis is limited to group level
  - Remove behavioral data which are captured after loan approval
  - Remove columns with the same value throughout
  - Remove columns with > 50% null values

Data Conversion

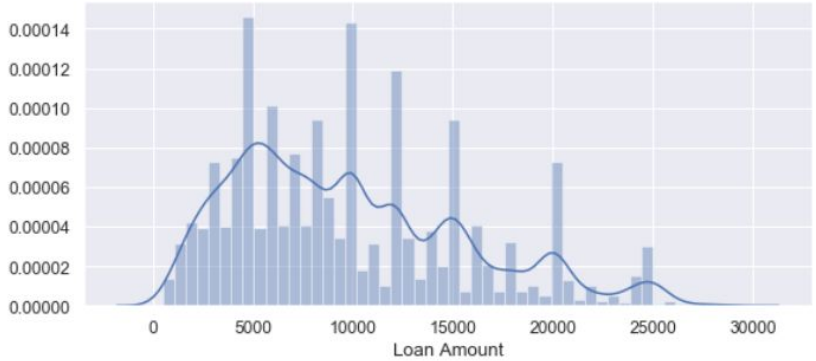
- Data type conversions to more usable format for analysis, including date type conversion
- Values are rounded off wherever needed

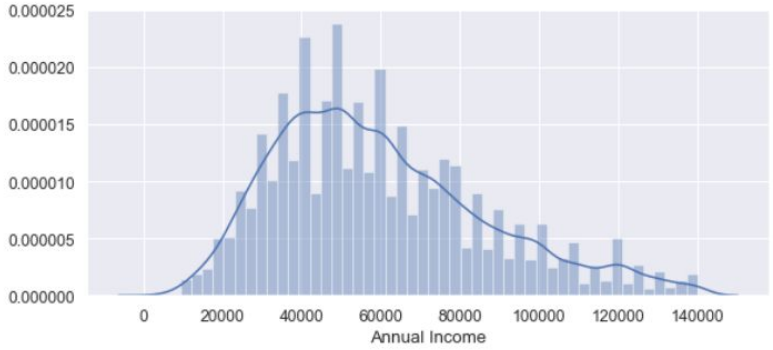
Dropping / imputing rows

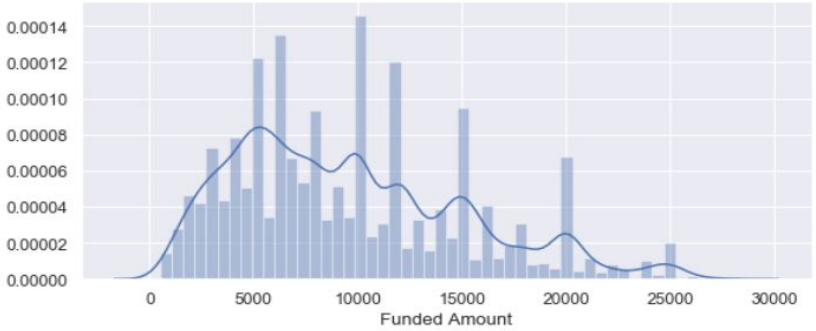
- Removing rows with null / NA values
- Getting derived columns

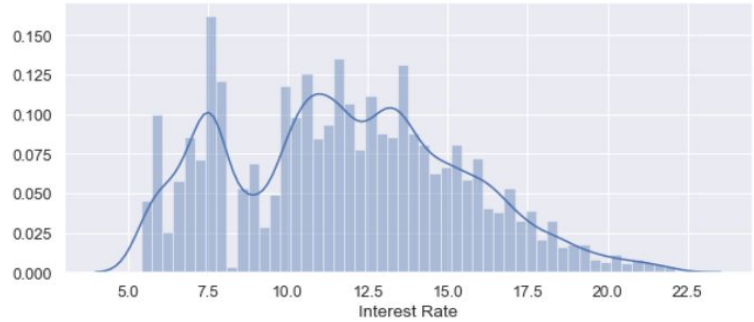
Univariate analysis

- Earlier used boxplot for various columns to identify and remove outliers from the analysis (box plots not shown in slide)





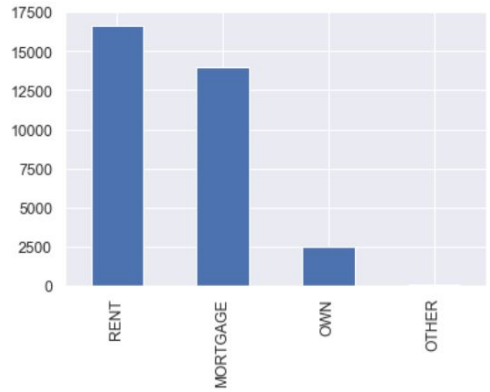




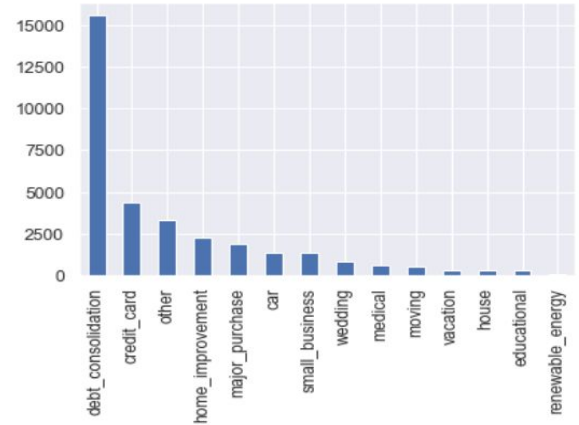
Observations from Univariate analysis

- Loan amount :
  - Mostly in the range of 5000 to 140000
  - Median loan amount is ~9000.
  - Mean loan amount: 9820.84
- Annual Income:
  - Mostly in the range of 40000 to 750000
  - Median annual income is 55000.
  - Mean annual income: 59883.28
- Funded amount:
  - Mostly in the range of 5000 to 130000
  - Median funded amount is 8500.
  - Mean funded amount: 9593.83
- Interest rate:
  - Mostly in the range of 8.9% to 14.26%
  - Median interest rate is 11.71%

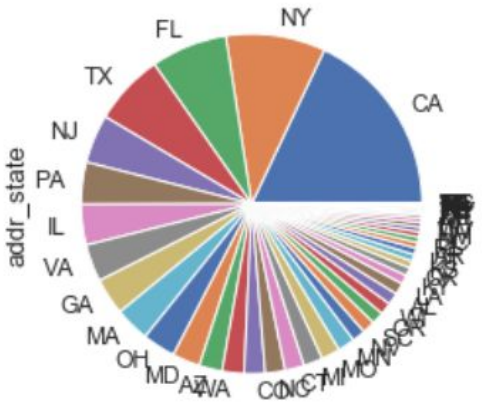
Unordered categorical variable analysis



Home Ownership



Purpose

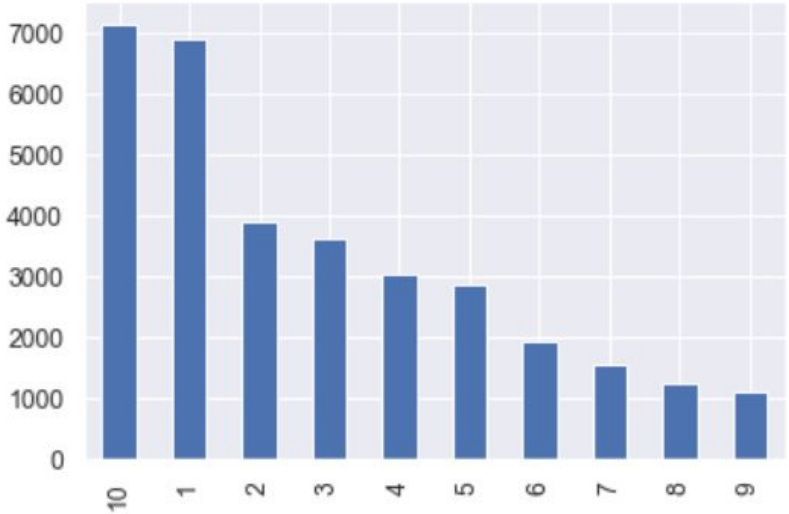


Loan address (state)

Observations:

- Most loan applicants stay on rent or have a mortgage
- Most of the loan applicants are planning to use the loan for debt_consolidations
- Most loan applicants are from CA

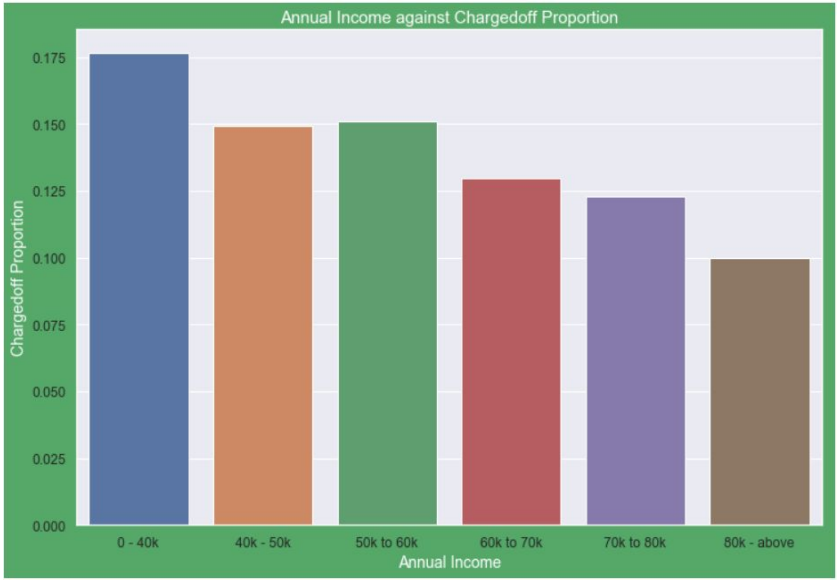
Ordered categorical variable analysis

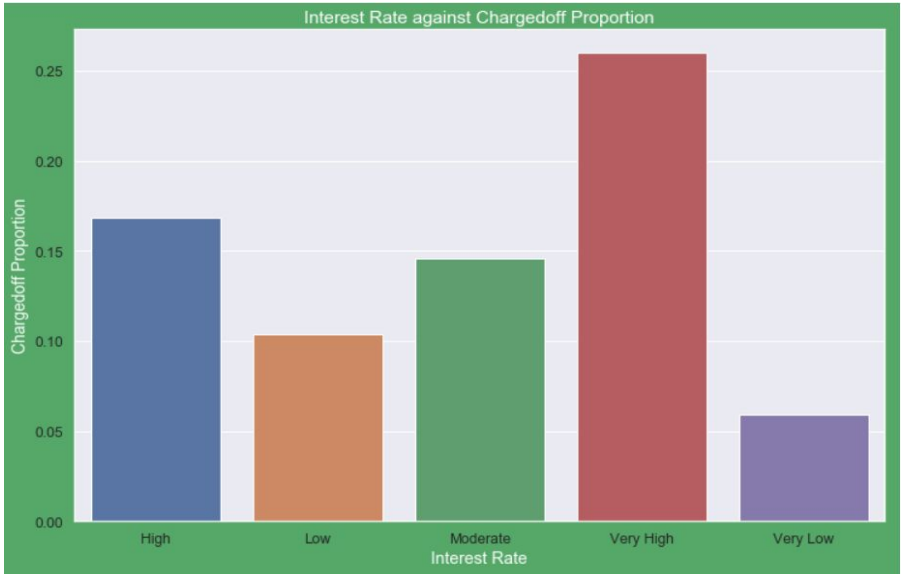


Employee length

Observation : Most loan applicants have either 10 or more years of experience or have ~1 year of experience

Bivariate analysis





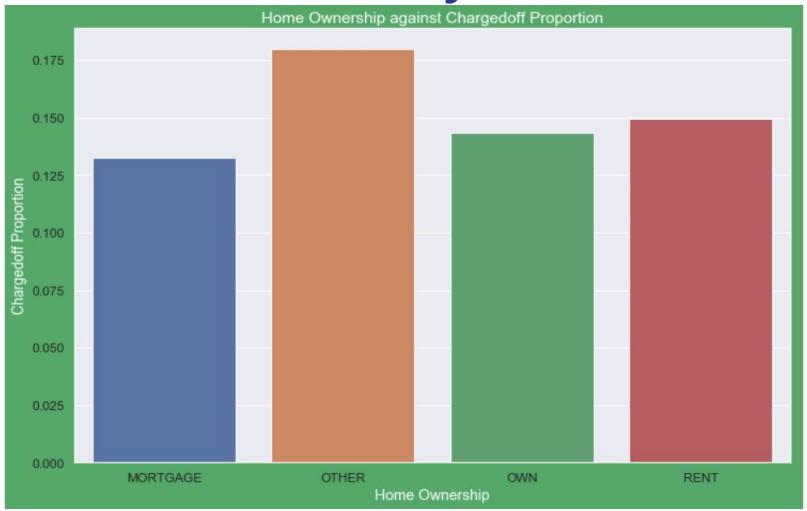
Observations:

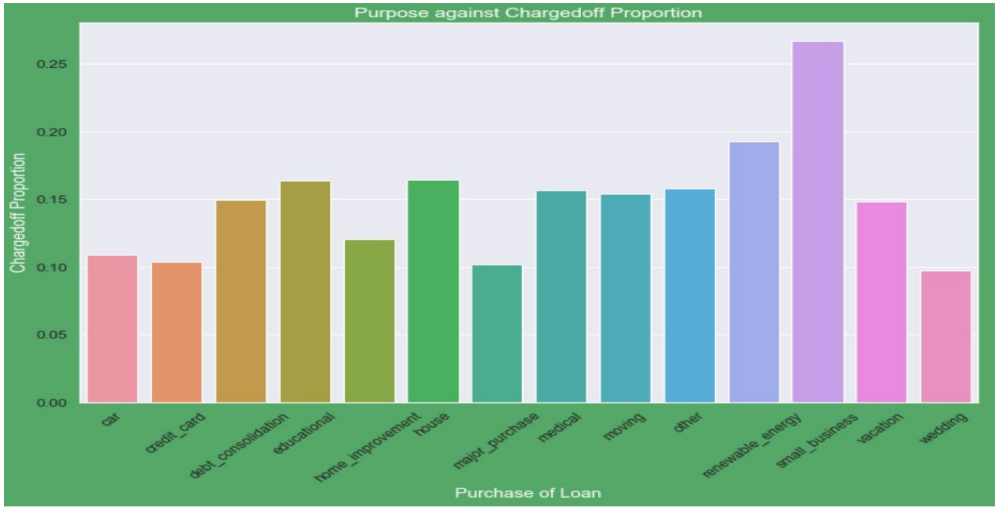
- Income range 0-20000 has highest chances of being charged off
- Income range 80000+ has least chances of being charged off
- With increase in annual income, charged off proportion reduces

Observations:

- Loans with interest rate less than 10% or very low has very less chances of charged off. Interest rates are starting from minimum 5 %
- Interest rate more than 16% or very high has good chances of charged off as compared to other interest rates.
- Charged off proportion is increasing with higher interest rates

Bivariate analysis





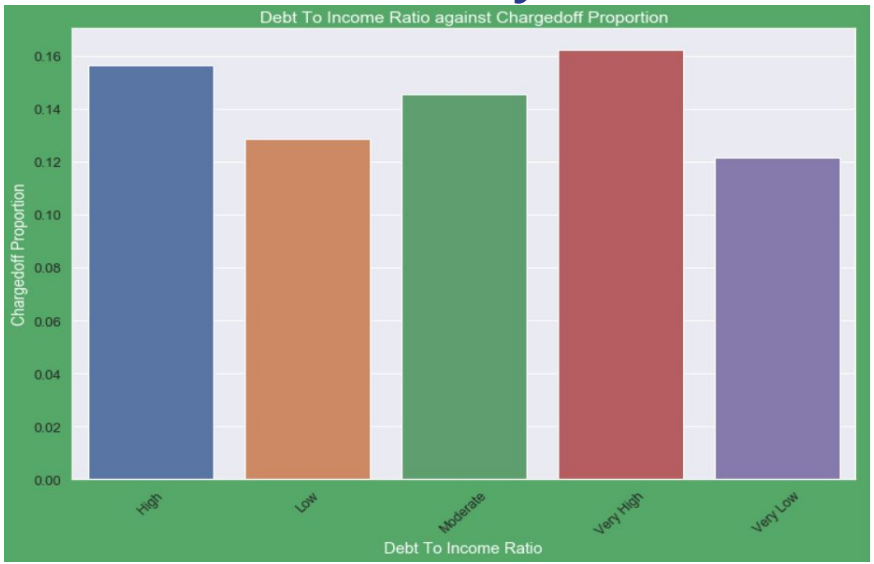
Observations:

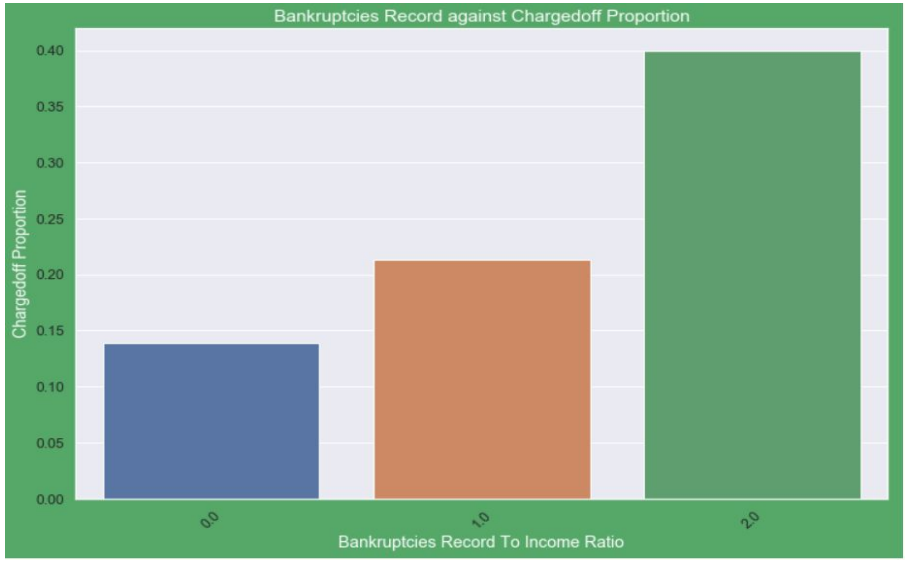
- Charged off proportion is least for mortgages

Observations:

- Applicants using the loan for small business are more prone to loan defaults

Bivariate analysis





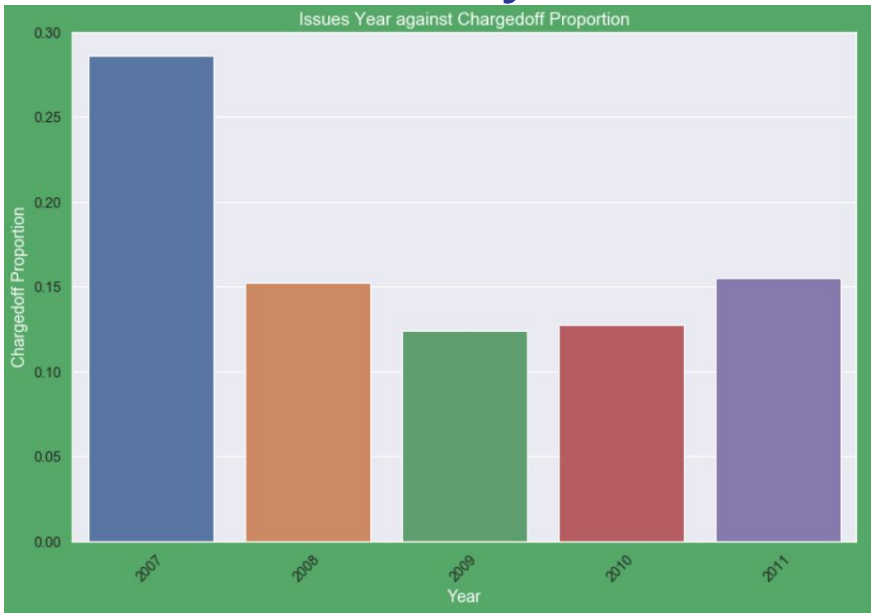
Observations:

- Higher the Debt to Income ratio, higher is the chances of loan default

Observations:

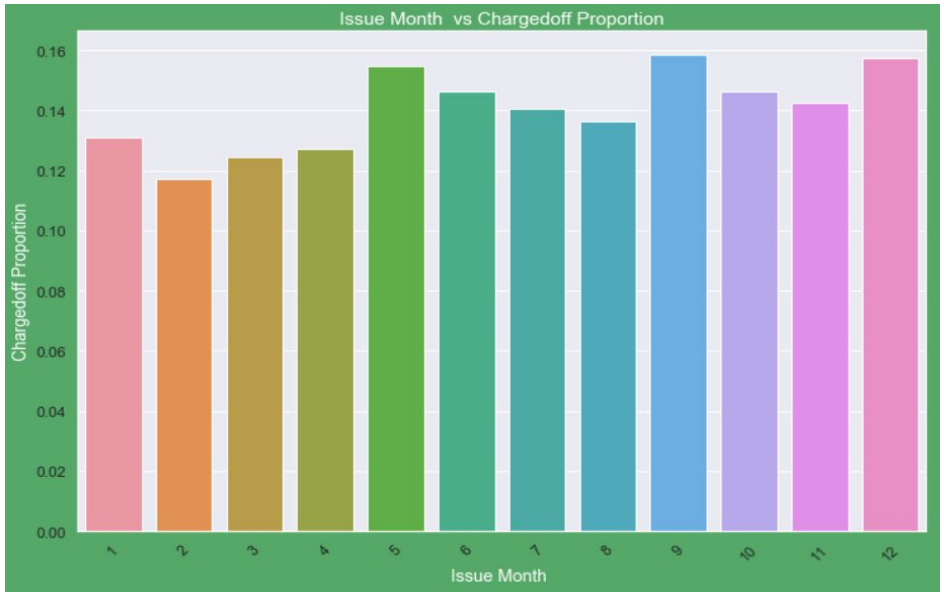
- Applicants with bankruptcy record as 2.0 have higher loan defaults

Bivariate analysis



Observations:

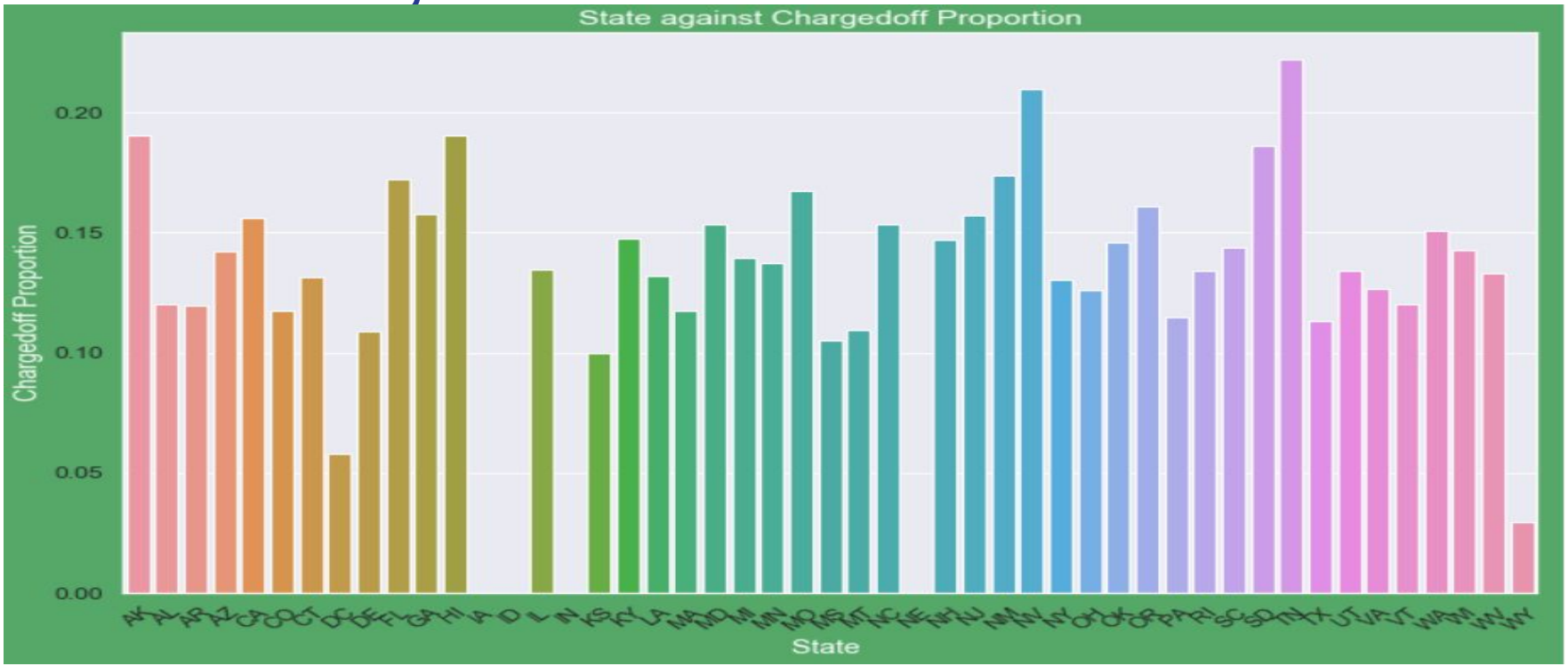
- 2007 has highest number of loan defaults



Observations:

- For applicants where loan has been issued in May, September or December, loan defaults is highest
- For applicants where loan has been issued in February, loan defaults is lowest

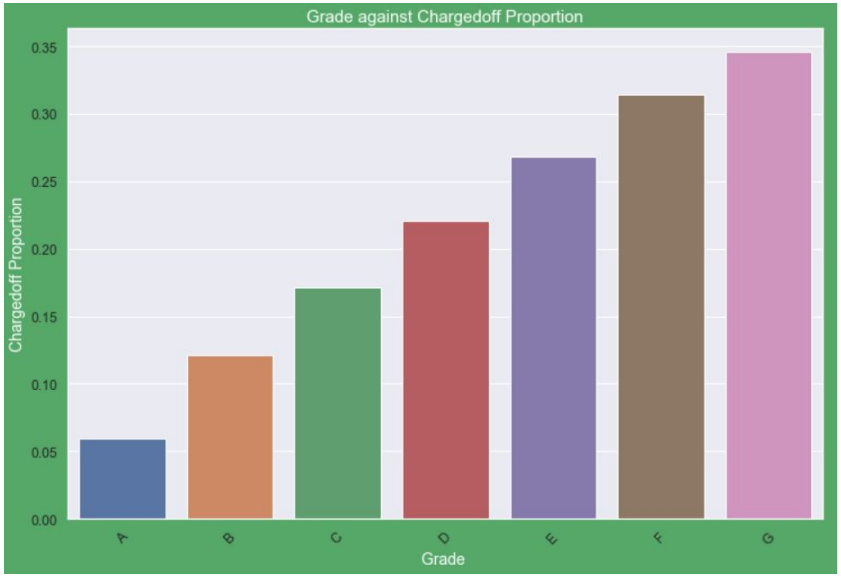
Bivariate analysis



Observations:

- TN and NV have highest number of loan defaults

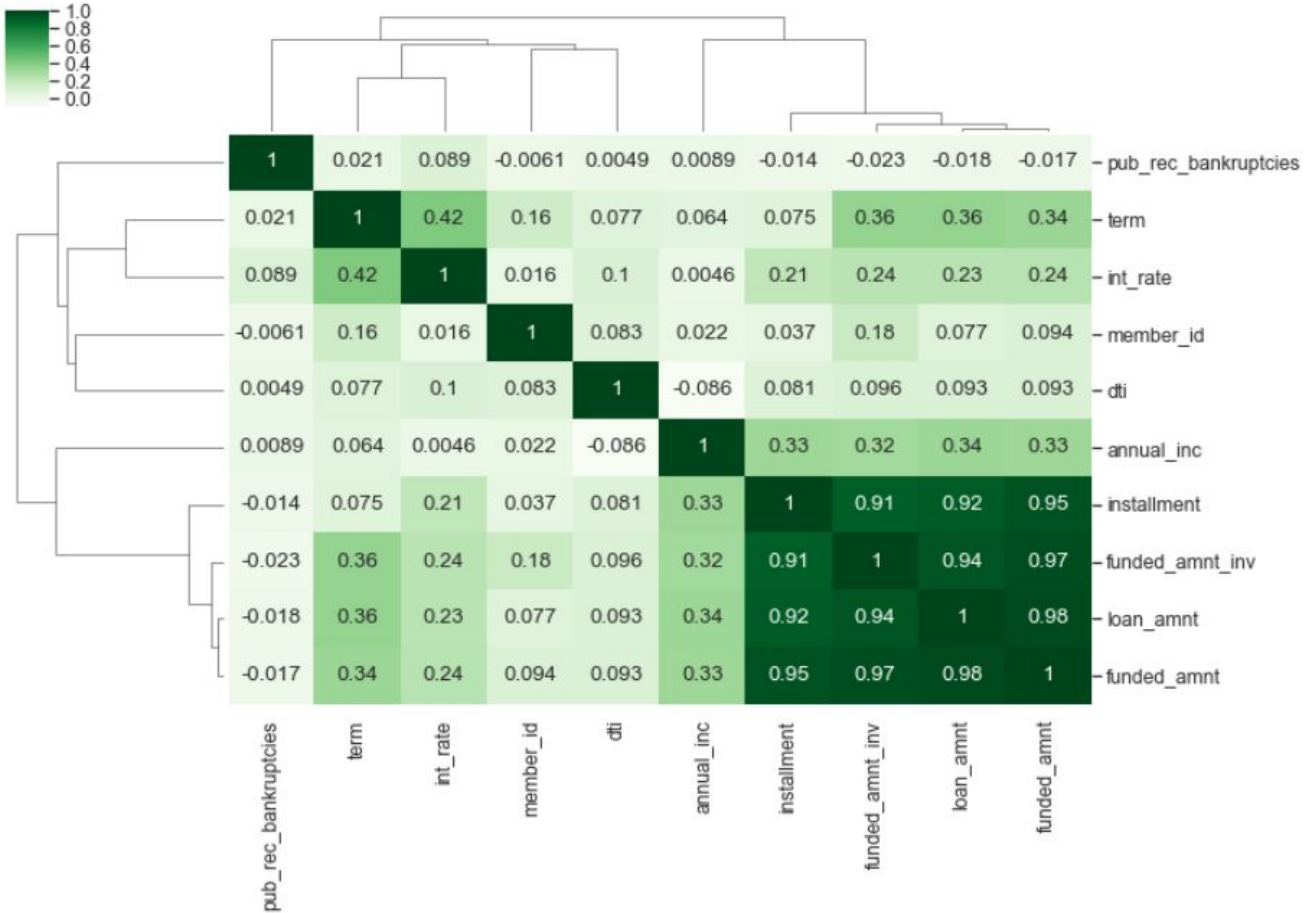
Bivariate analysis



Observations:

- As the grade moves from A to G, the loan defaults also increase

Correlation analysis



Observations from Correlation analysis

- Term has a strong correlation with loan amount and interest rate
- Annual income has strong correlation with loan_amount
- Annual income has negative correlation with debt to income ratio
- Loan amount has negative correlation with public record bankruptcies
- The total amount committed to that loan at a point in time, the total amount committed by investors for that loan at that point, the monthly payment owed by the borrower if the loan originates and the amount of loan are highly correlated

Recommendations

- What works well:
  - Applicants with lower debt to interest ratio have higher chances of paying back the loan
  - Applicants with mortgage have higher chances of paying back the loan
  - Applicants with grade A have higher changes of paying back the loan
  - Applicants with the term 36 months have less defaulters

- What does not work well:
  - Applicants with small businesses are more prone to be defaulters
  - Applicants with bankruptcy record as 2.0 have higher loan defaults

Thank You