Lending Club case study

Exploratory Data Analysis

Balaji Venkatraman S

1. Problem Statement
2. Objectives
3. Data Summary
4. Dat Cleaning
5. Data Conversions
6. Univariate Analysis
7. Bivariate Analysis
8. Conclusions

Problem Statement

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,
    a. Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
    b. 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'.
    c. 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)

Objective

1. Identify variables that provide strong indicators of potential loan default thus helping Lending club to decide approval/rejection of loan.
2. Identification of such applicants using EDA is the aim of this case study
3. Perform EDA operations like.
    a. Data Cleaning
    b. Univariate Analysis
    c. Segmented Analysis
    d. Bivariate Analysis
    e. Derived Metrics

Data Summary

- The loan data has 391717 rows
- The loan data has 111 columns.
- Hence the shape of the data is (39717, 111)

Data Cleaning

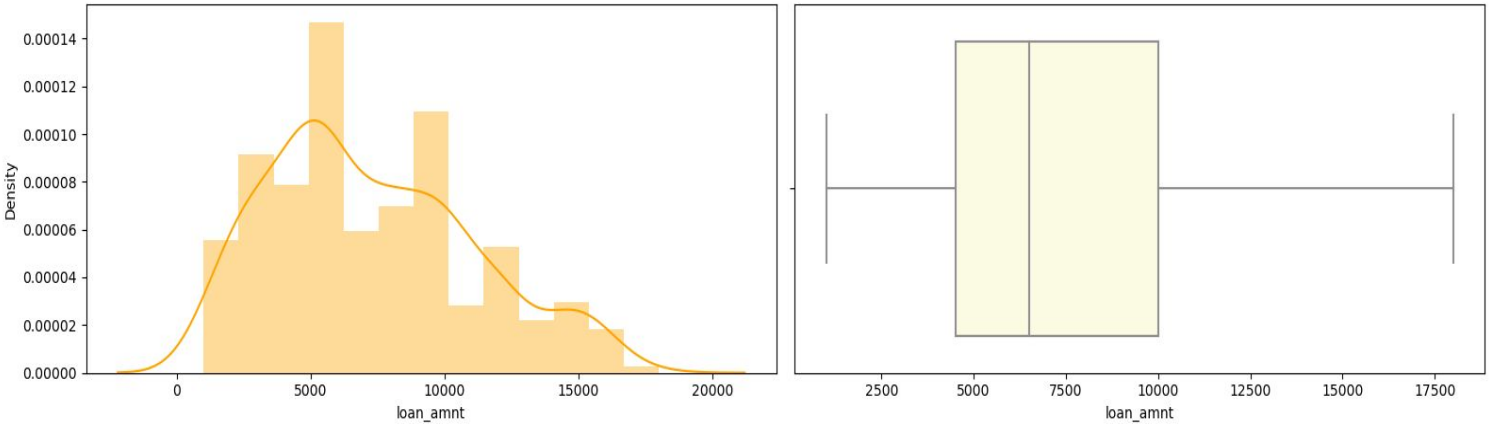
- No duplicate rows found
- Delete loan_status=’current’ since ‘current’ status does not contribute in the data analysis
- There are 55 columns are NA, hence they are removed
- Dropping sub_grade column since do not contribute data analysis
- Remove text based columns like text, title
- After the data cleaning the shape of the data frame is (38577, 53)

Data Conversions

- The `term` string value is converted to int for data analysis
- Unwanted ‘%’ character from ‘int_rate’ has been removed in order to work with data analysis
- The ‘emp_length’ has unnecessary characters like space and <. Hence they have been trimmed
- ‘Issue_date’ is converted into data format
- Created three new column issue_year, issue_day, issue_month from issue_date field
- Three buckets have been created annual_inc_bucket, int_rate_bucket, loan_amt_bucket
- Unwanted outliers have beed removed

Univariate Analysis

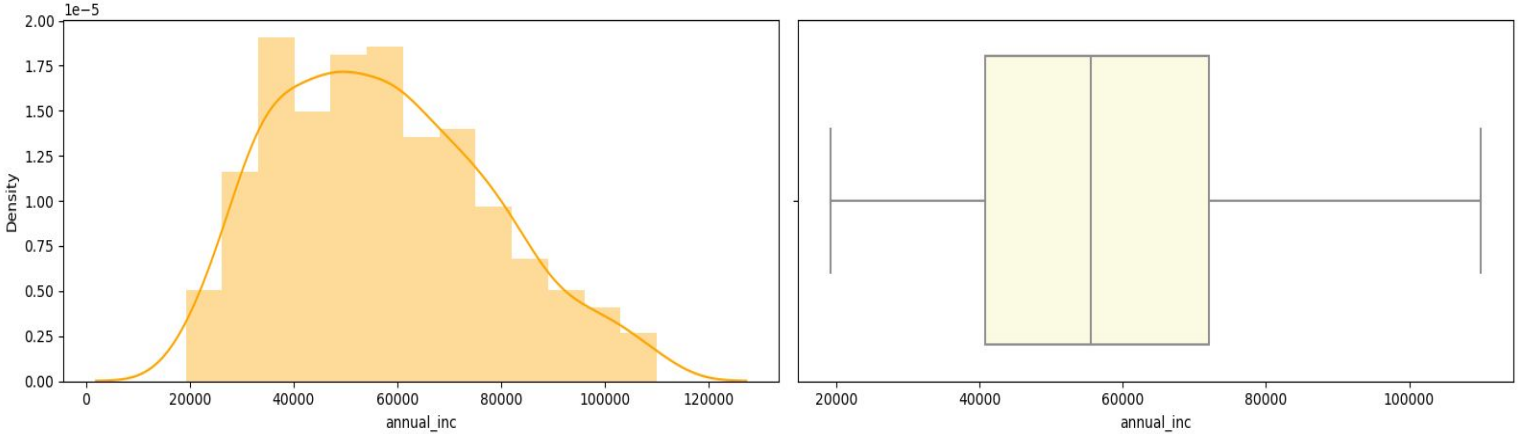
Loan Amount



Inference

1. Minimum percentile of loan applied is 5k
2. Maximum percentile of loan applied is 22k
3. 5-14k is range in which most loans are applied

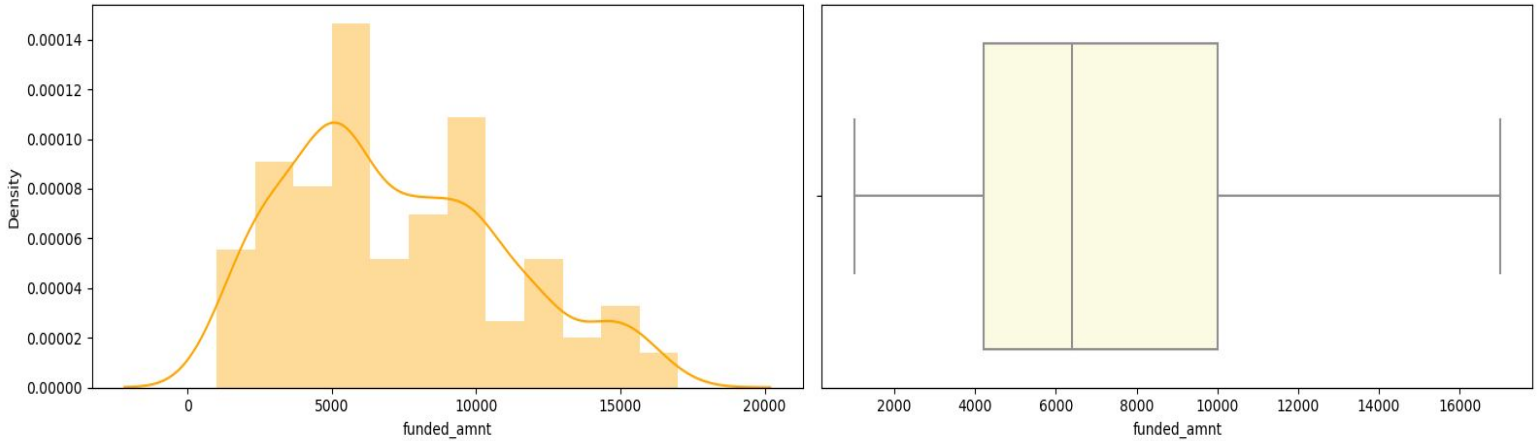
Annual Income



Inference

1. The Average annual income is 58700 which is 58k
2. Minimum annual income is 4k and maximum is 165k

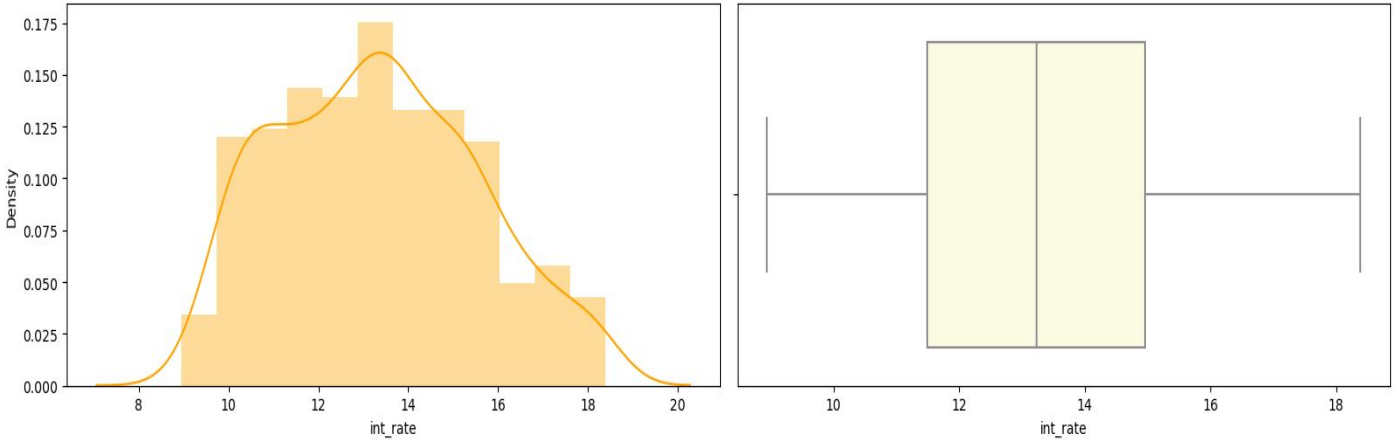
Funded Amount



Inference

1. The average funded amount is 8609
2. Maximum funded amount is 17k and minimum is 1k

Interest rate

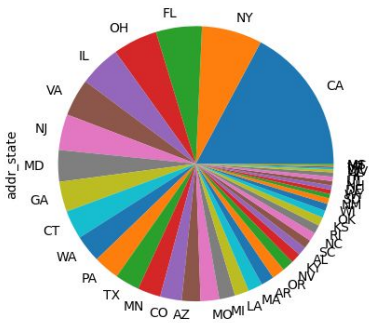


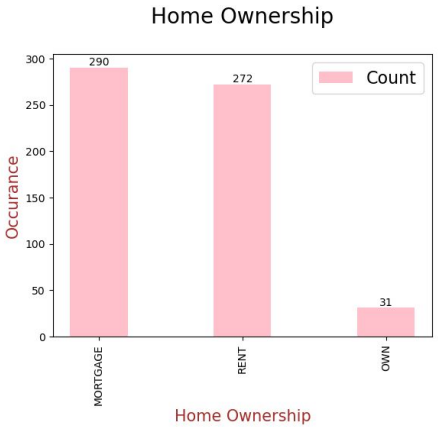
Inference

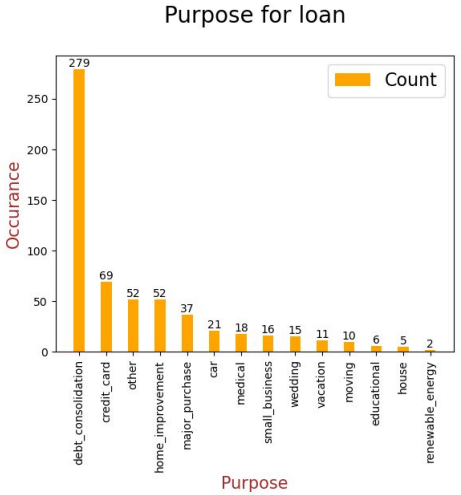
1. The maximum interest rate is 17.44%
2. The Average interest rate is 11.49%
3. Maximum users have the interest rate between 11% to 15%

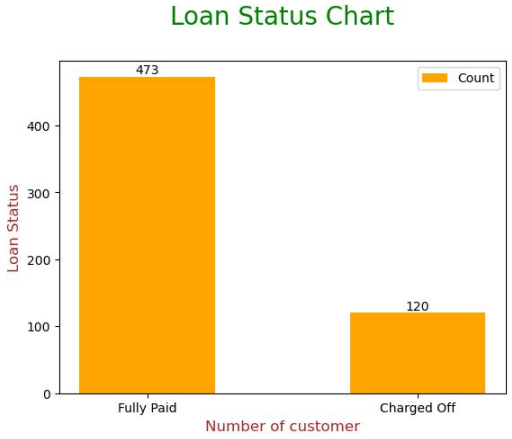
Final Observations

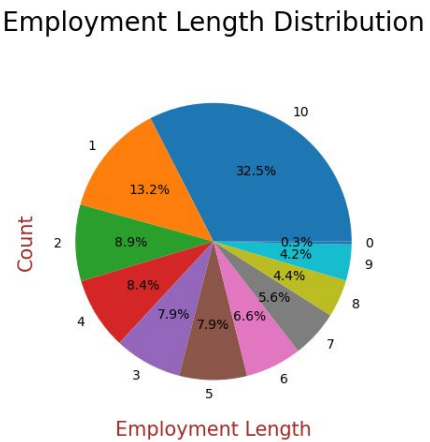
1. Most of applicants are for debt_consolidation
2. Most of the applications are in RENT and MORTGAGE
3. Most of the customers are already paid off
4. Most of the applicants are with 10+ years of experience
5. Most of the applicants are from CA





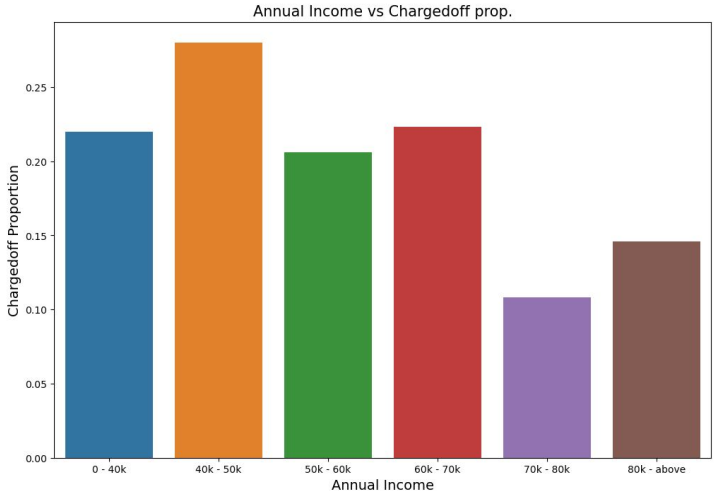






Bivariate Analysis

Annual Income vs Charged off.

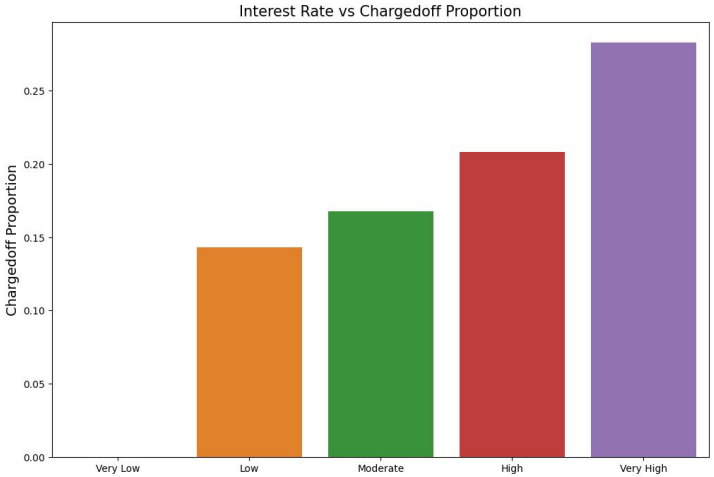


Inference

1. Income range 70-80k has less chances of charged off.

2. Income range 40-50k has high chances of charged off.

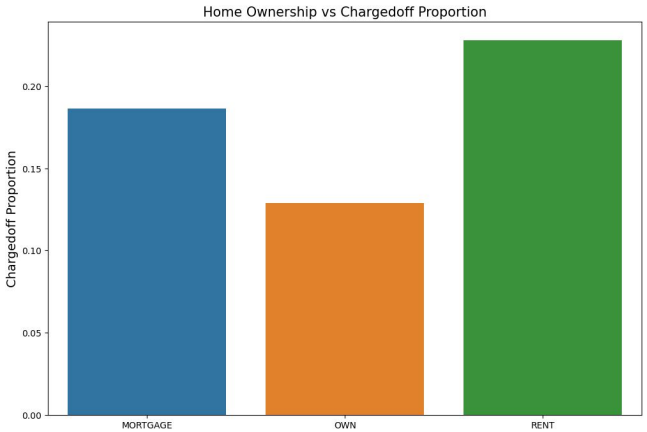
Interest rate vs Charged off



Inference

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

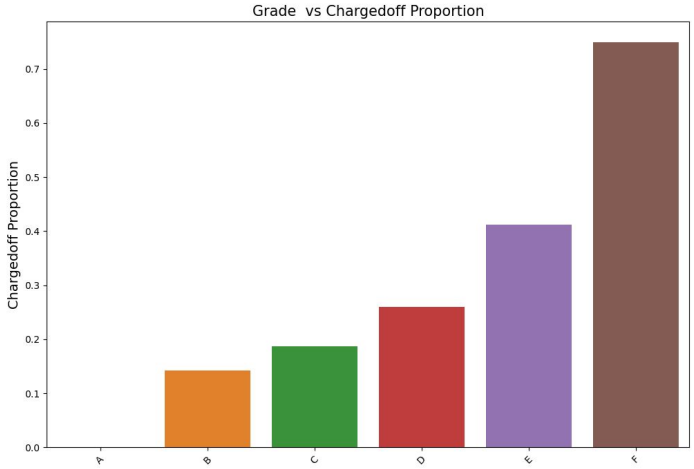
Home Ownership vs Charged off



Inference

1. RENT ownership has highest charged off
2. Those who are not owning the home is having high chances of loan defaults.

Grade vs Charged off

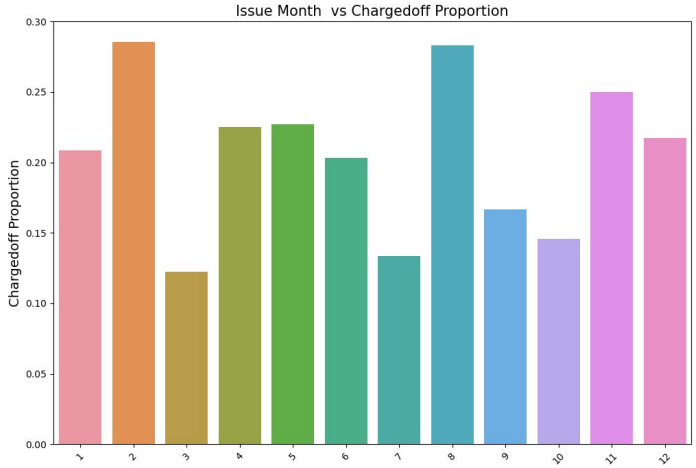


Inference

1. The Loan applicants with loan Grade F is having highest Loan Defaults.

2. The Loan applicants with loan A is having lowest Loan Defaults.

Issue Month vs Charged off

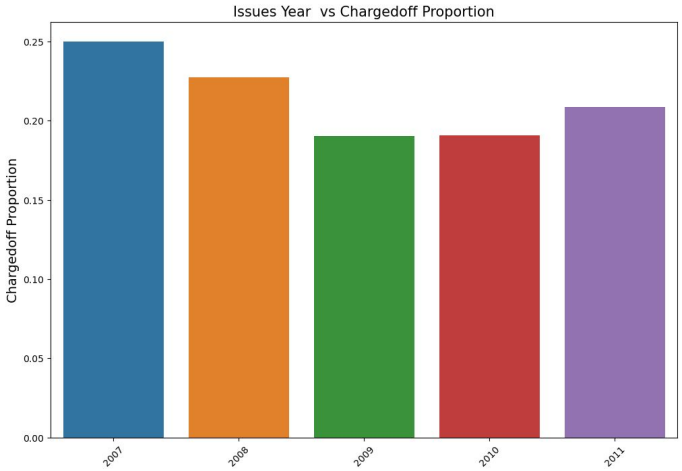


Inference

1. Those loan has been issued in Feb, Aug and Nov is having high number of loan defaults

2. Those loan has been issued in month of March is having low number of loan defaults

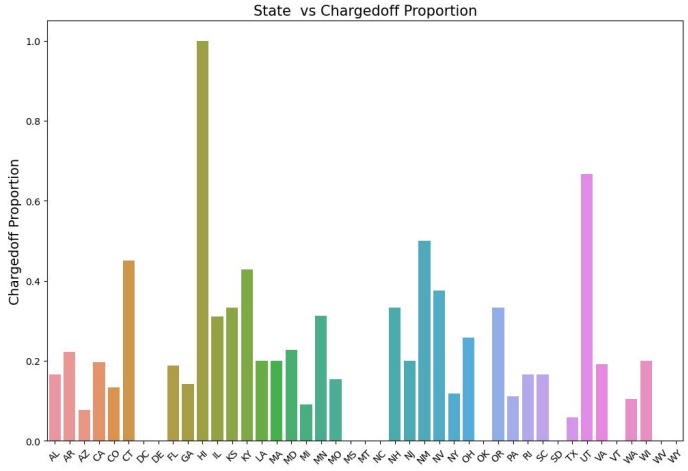
Issue Year vs Charged off



Inference

1. Year 2007 is highest loan defaults.
2. 2009 is having lowest loan defaults.

State vs Charged off



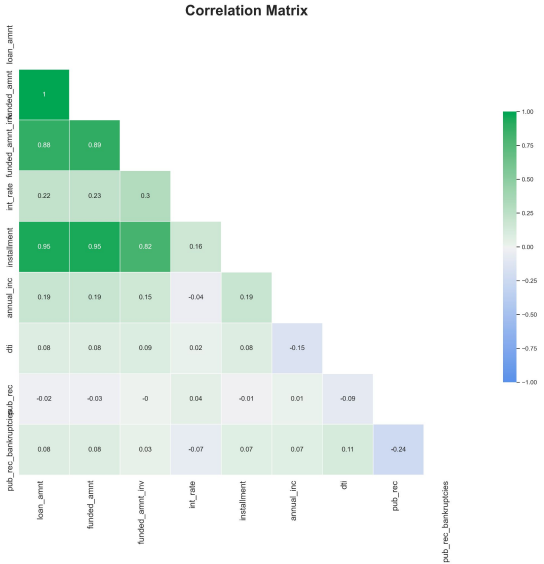
Inference

1. HI States is holding highest number of loan defaults.

2. TX is having low number of loan defaults

Correlation Analysis

Correlations



Inference

1. term has a strong correlation with loan amount
2. term has a strong correlation with interest rate
3. annual income has a strong correlation with loan_amount

Conclusions

- Income range 40-50k has high chances of charged off.
- Rent and Mortgage applicants are high percentages of loan defaulter
- Small business have high chances of loan defaults
- HI States is holding highest number of loan defaults
- Those loan has been issued in month of March is having low number of loan defaults
- 2009 is having lowest loan defaults and 2007 is having highest loan defaults
- The Loan applicants with loan Grade F is having highest Loan Defaults.
- Those loan has been issued in Feb, Aug and Nov is having high number of loan defaults

Thanks