Lending Club Case Study – ML38

By

Sumit Vashistha – Group Facilitator

Shammi Kapoor

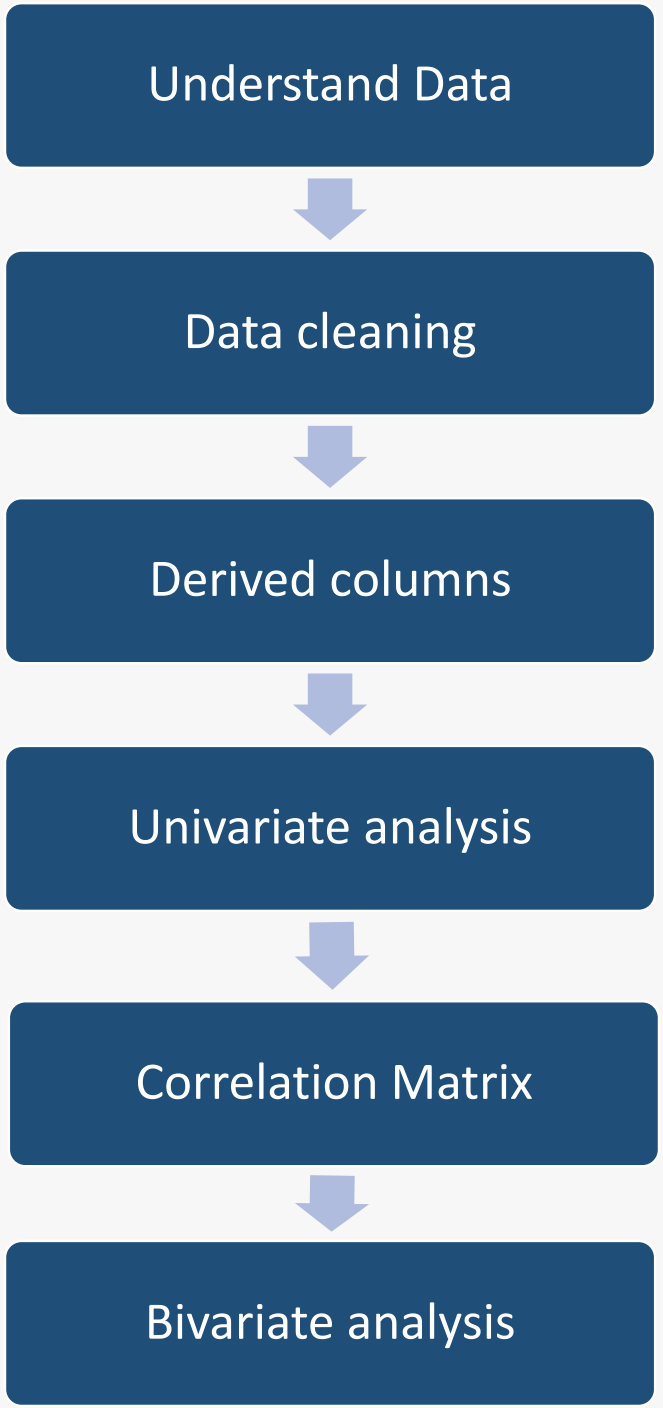
Problem Statement

You work for a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must decide 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

Apply EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default based on the provided loan data set

Exploratory Data Analysis



Understand Data

- Total Attributes =116
- Attributes used for analysis = 45
- Drop unused attributes from dataframe.

Attributes	Description	Reason for identifying as Key Attribute
Emp_length	Employment length of an individual	Individual pay loan in beginning of his/her career as compared to having 10 years of experience.
Loan_amnt	Actual Amount for which borrower applied for the loan	Identify how loan amount can have impact on payment + changes in interest rate with higher amount
Loan_status	A key column which identifies loan has been full paid off/ charged off	Other metrics can revolve around this to identify in which all cases loan has been fully paid off/charged off
Int_rate	Rate of interest applicable on the loan amount	An important metric to understand whether higher rate means high chances of charged off
Installment	Monthly payment applicable on the approved loan	Monthly installment has an impact on loan payment
Public_rec	Number of derogatory public records	Identify risk status associated with a customer
pub_rec_bankruptcies	Number of public record bankruptcies	Customers can be denied loan if we have sufficient data of previous bankruptcies
Purpose	Actual need for which loan has been applied for	Identify the purpose of loan to understand the impact on loan payment
home_ownership	Current accommodation of the loan applicant	Identify if people living in rented accommodation vs owned house have higher chances of defaulting due to high debt to income.

Data Cleaning

➢ Fix rows and columns

- Duplicate row = 0
- Empty row = 0
- Identify and drop columns having null values = 54

➢ Data correction

- Remove % characters in the columns to use it for numeric calculations
  - int_rate
  - revol_util

➢ Standardising

- Convert columns to numeric data type
  - loan_amnt
  - emp_length
  - funded_amnt
  - int_rate
  - total_pymnt
  - funded_amount_inv
  - installment
  - annual_inc,
  - dti

Derived Columns

Derived columns from Loan Issued Date(issue_d)

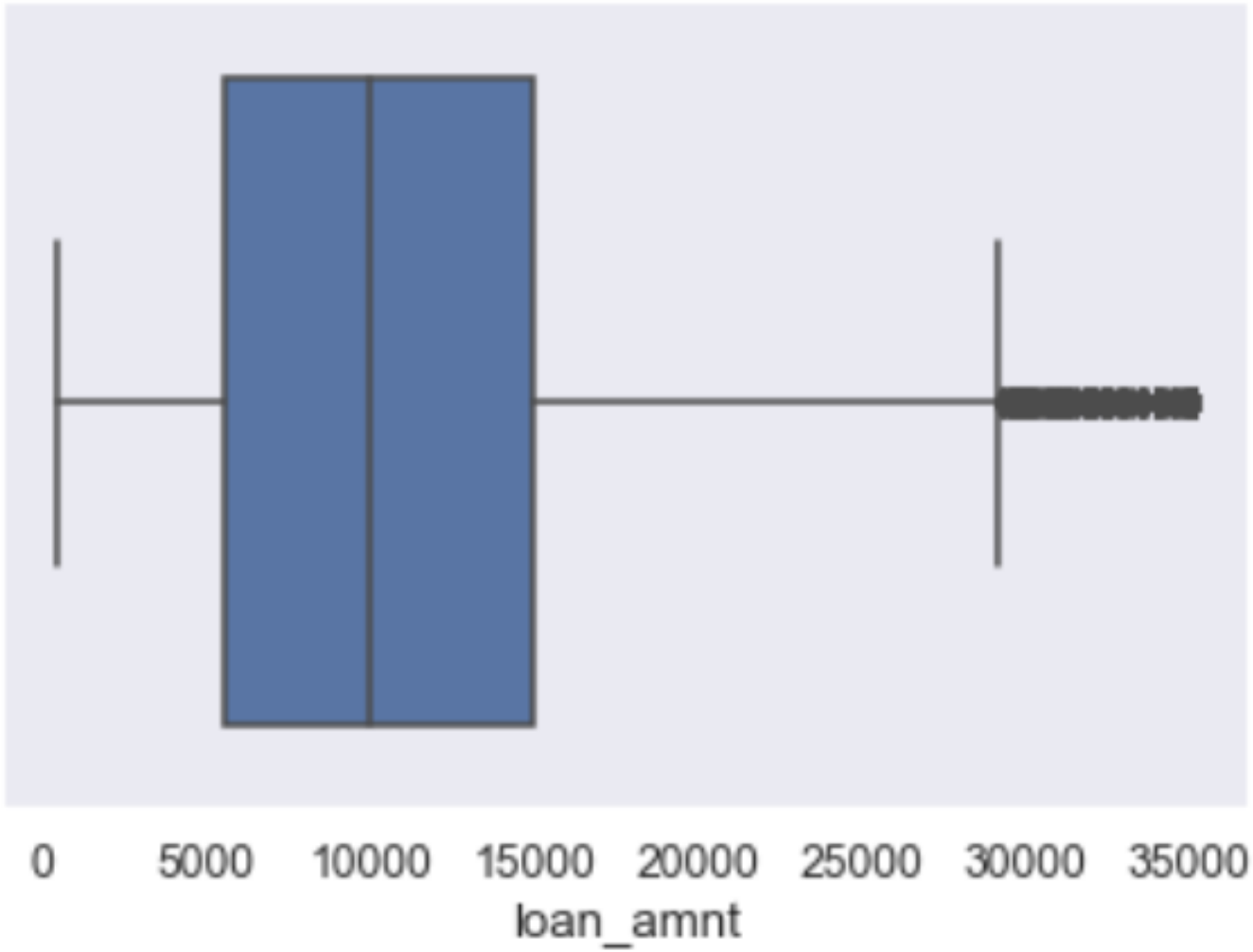
- Month
- Year

Bucketing various attributes to understand the impact with other attributes.

- Annual income
- Loan amount
- Interest rate
- Debt to income

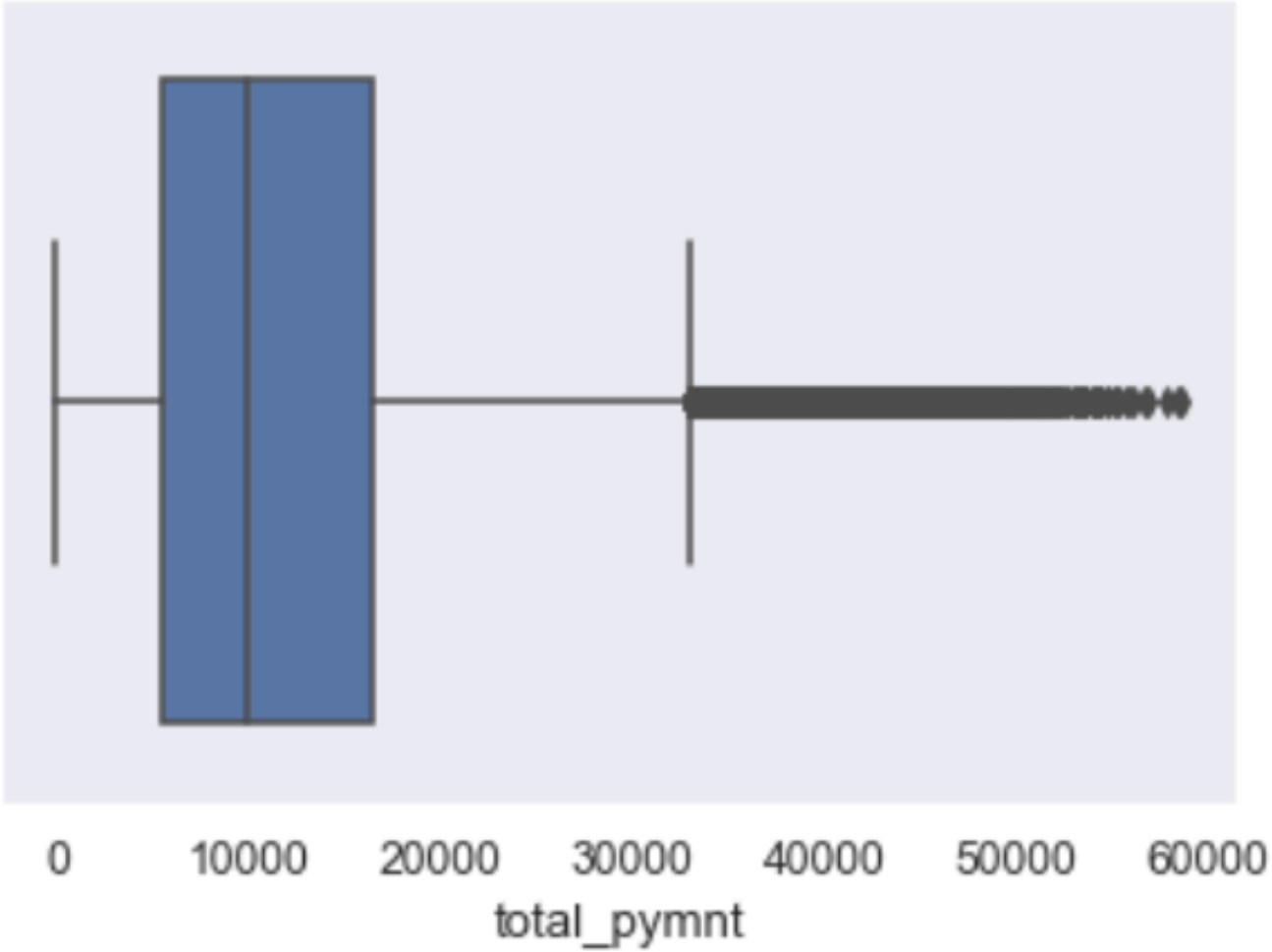
Univariate Analysis-loan amount

- Average Loan amount is around 11000
- 50% of loan amount is between ~5000$ to ~15000$.
- ~75% of people applied for loan below 15000$ whereas as the maximum is 35000$.



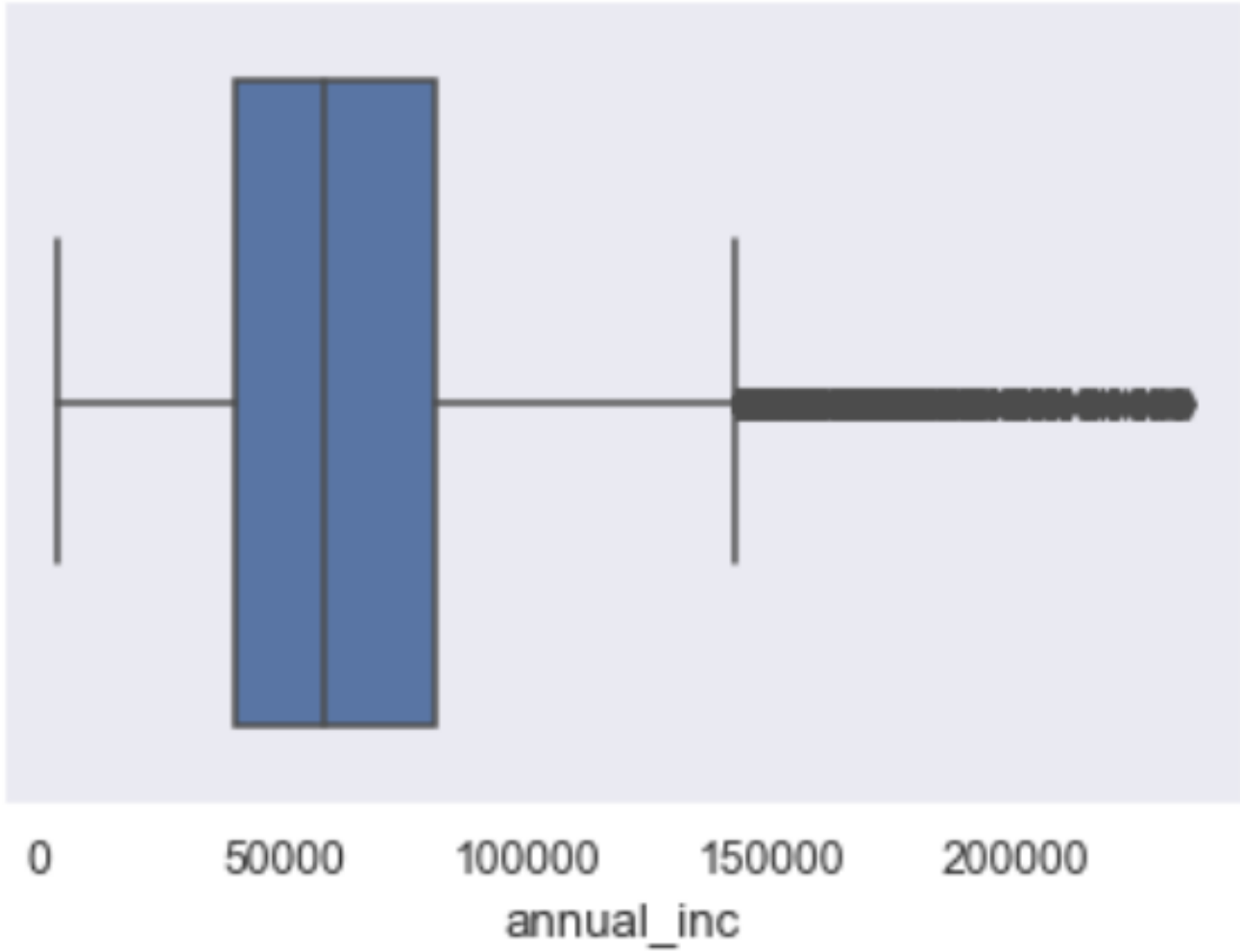
Univariate Analysis - Total Payment

- 75% is below 20000$
- IQR is 10957.5$
- 99% looks on higher side around 58000K which is higher than the max loan amount applied.



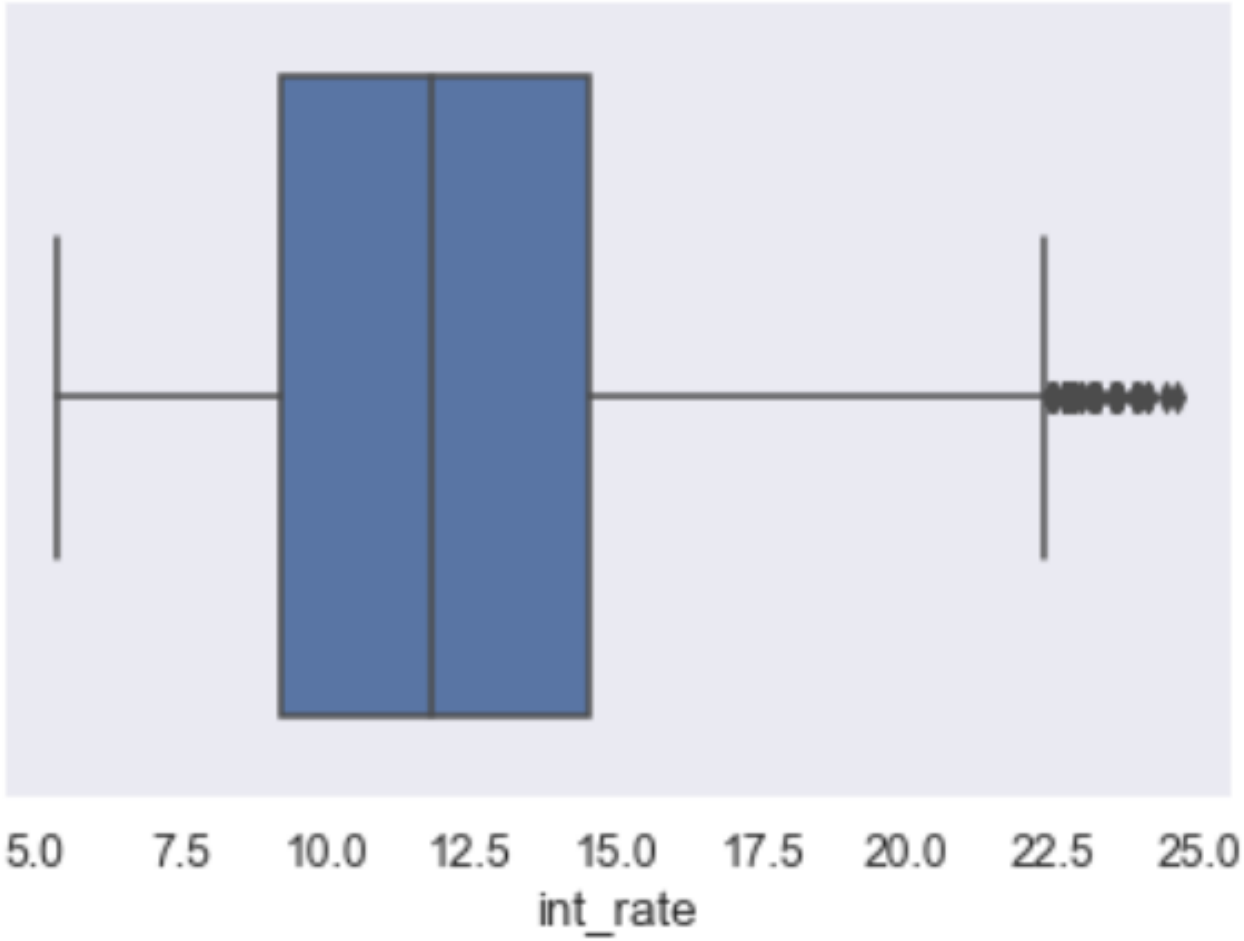
Univariate Analysis - Annual Income

- Remove Outlier component.
- Excluding Outlier, the 99% looks good and dropped from 600K to 200K
- Average annual income= ~65000 which is almost double to the loan amount applied in most of the cases.



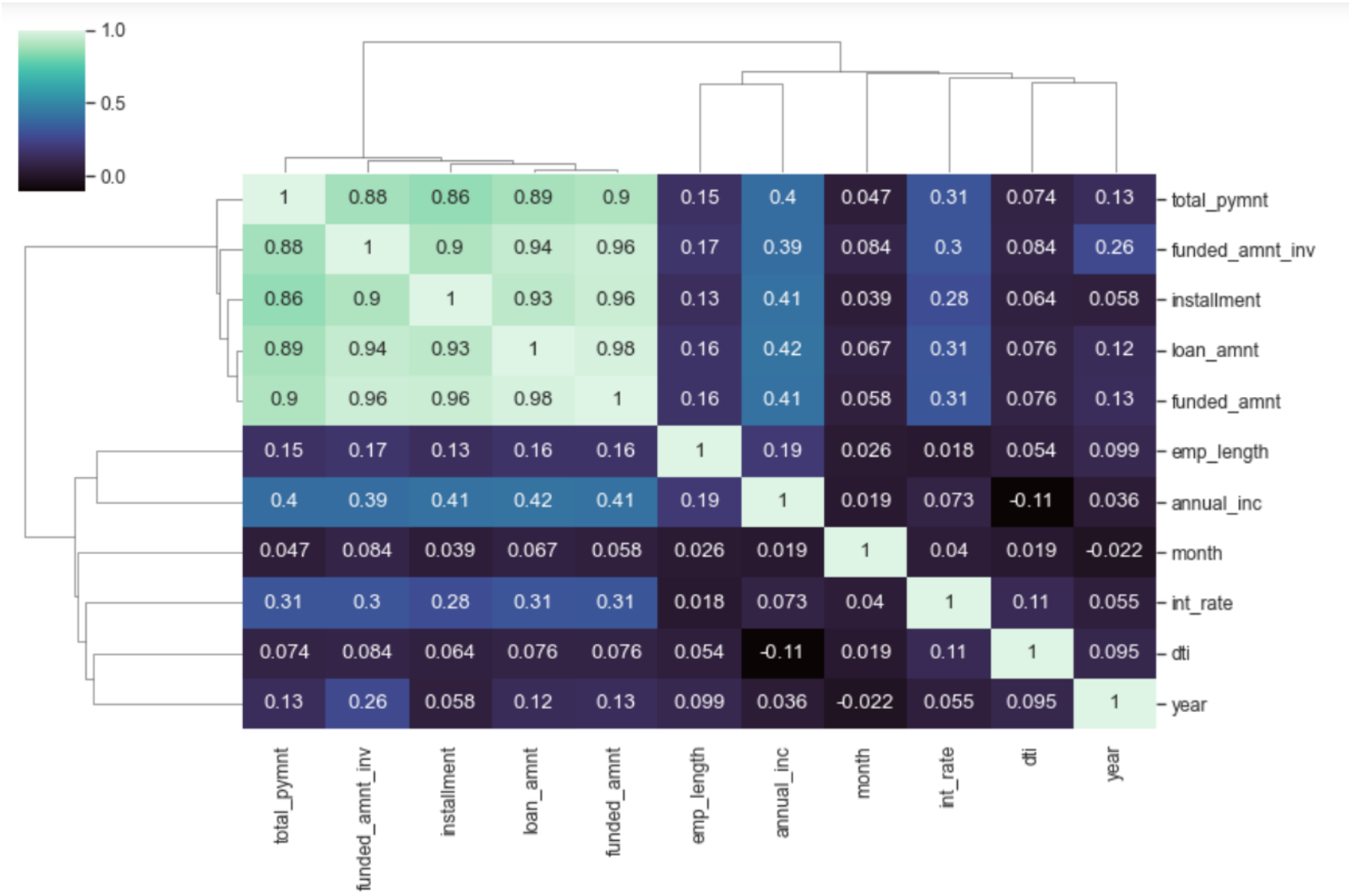
Univariate Analysis - Interest Rate

- Average interest rate falls under 12%
- 75% of interest rate falls under ~15%
- Need to identify if higher loan amount leads to higher interest rate



Correlation Matrix

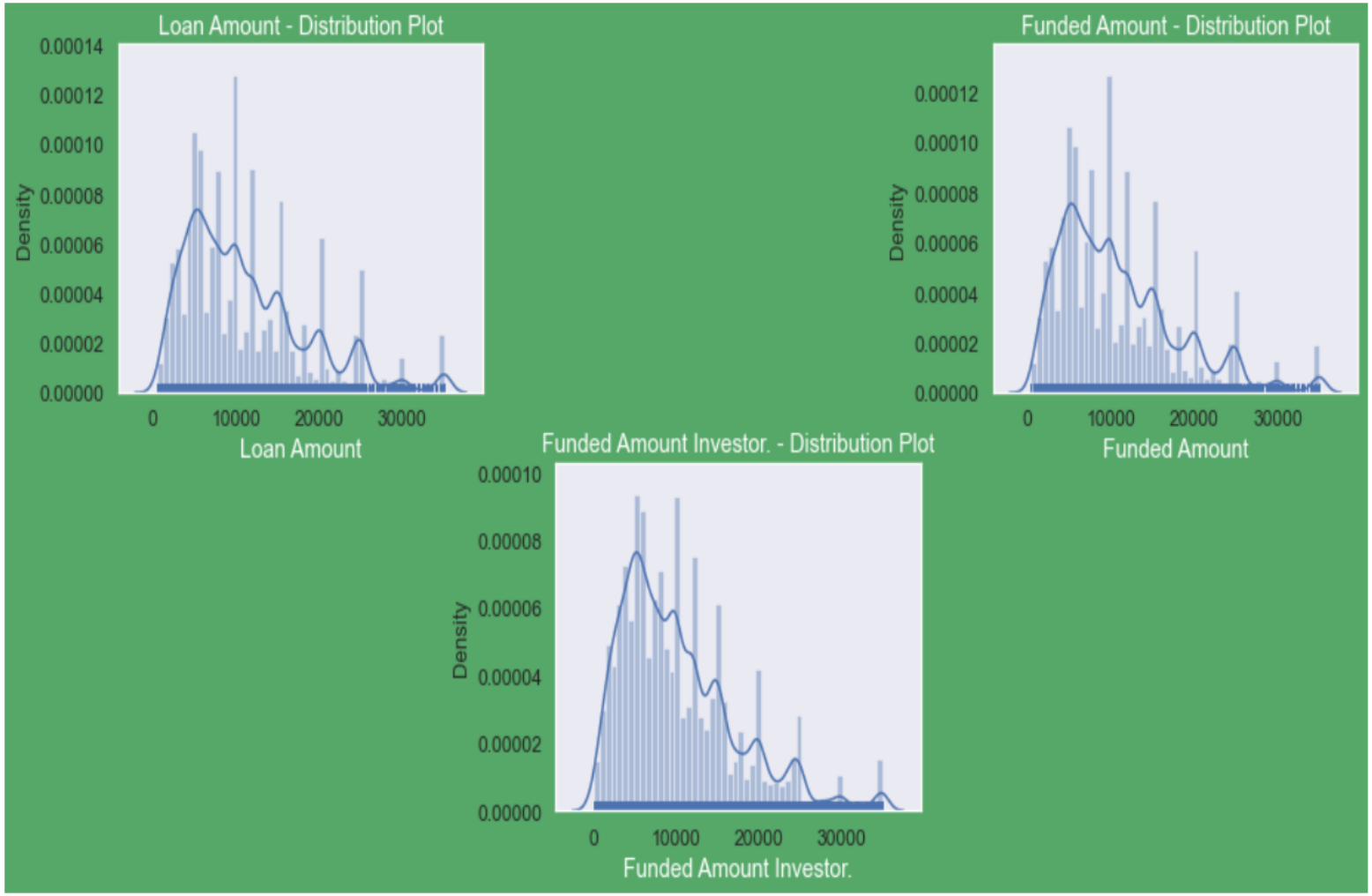
- Loan Amount, Investor Amount and Funding Amount are strongly correlated
- Annual Income and DTI is negatively correlated
- Positive Correlation between Annual Income and Employment Years means income increase with experience



Univariate Analysis - Three loan amount fields using distribution plot

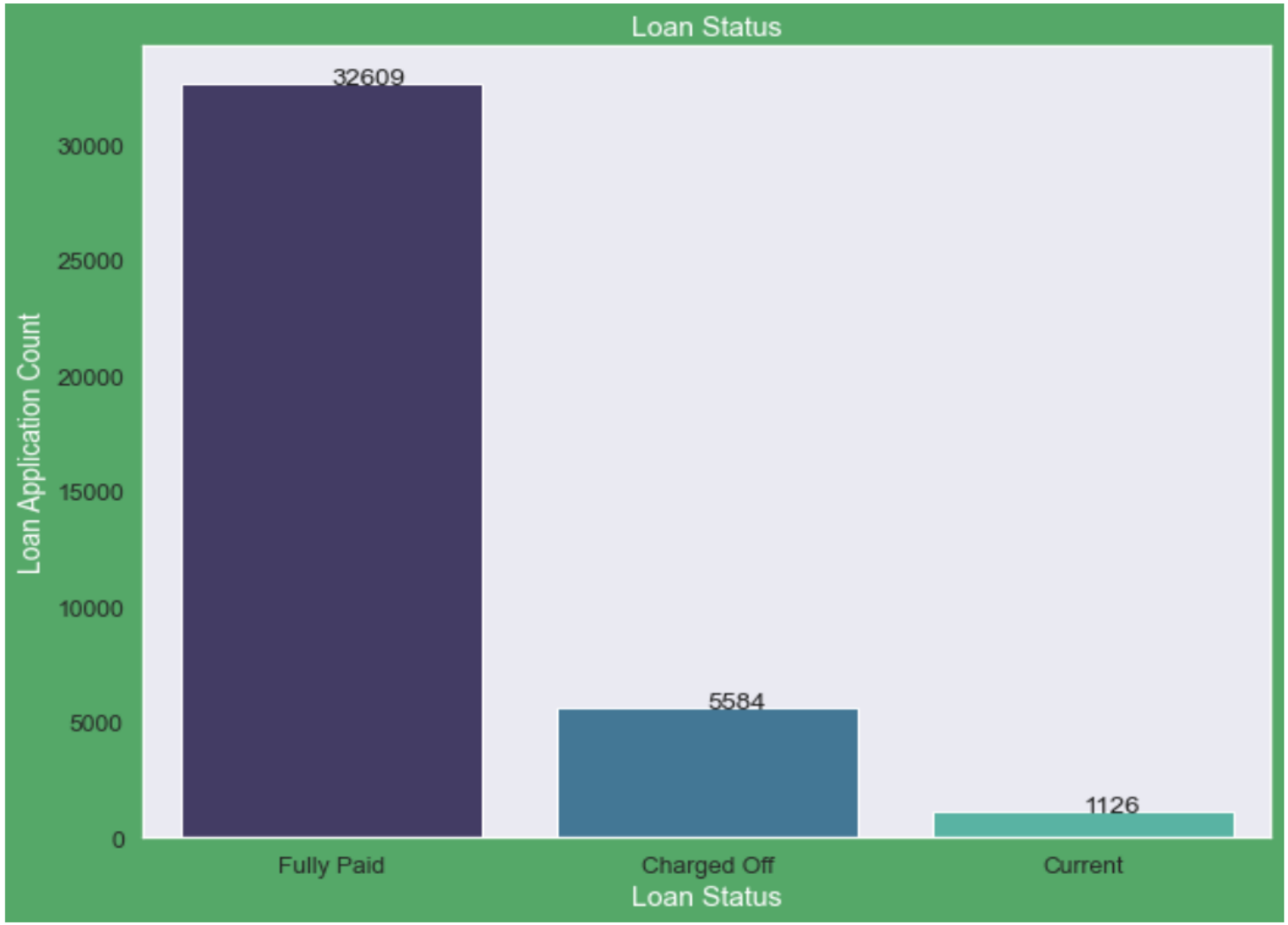
- Distribution of Loan Amount, Funded Amount and Funded Amount Investor look very similar

- This means we don’t need to go ahead with all the three for further analysis and choose to go with instead loan amount only



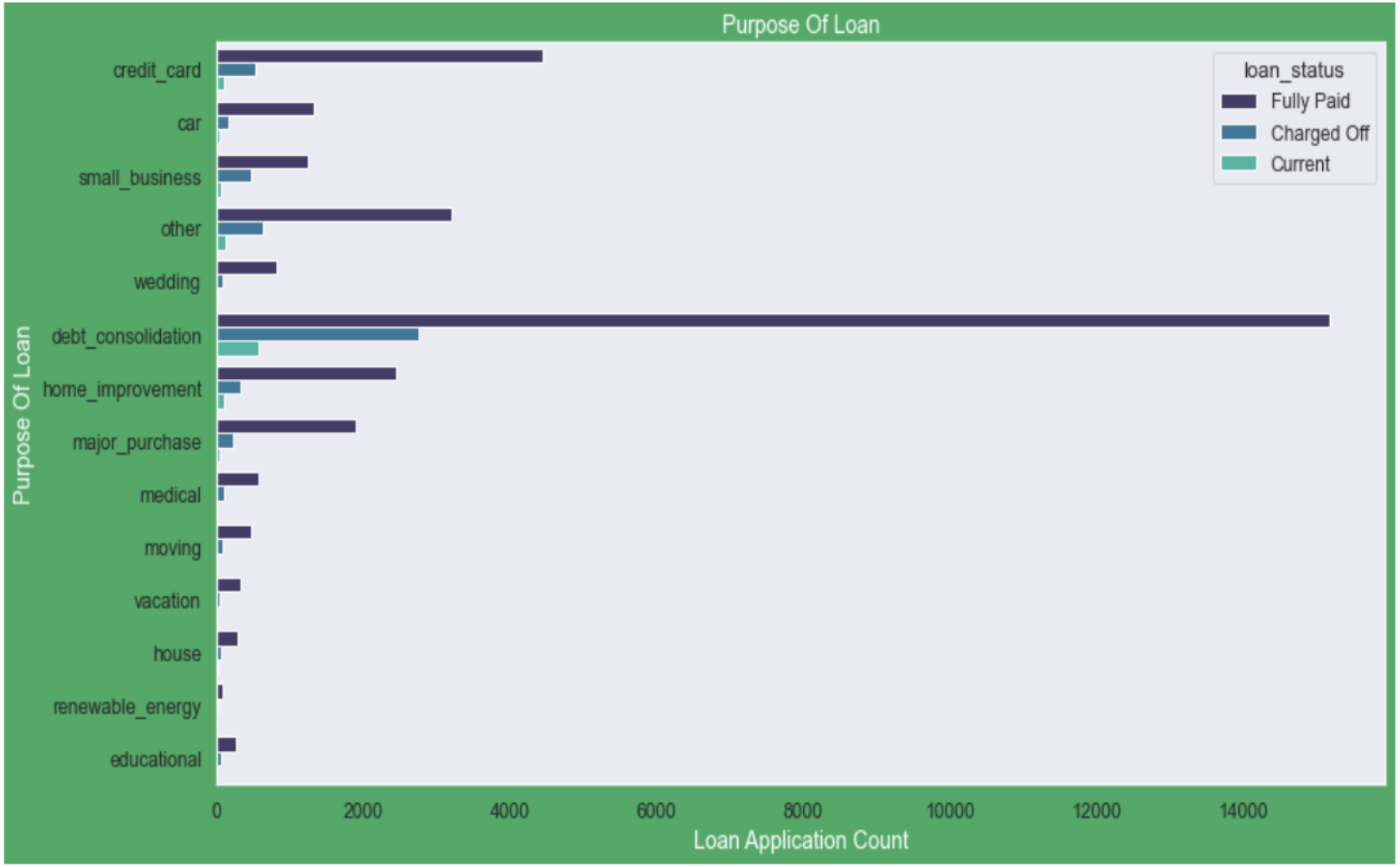
Univariate Analysis Unordered Categorical Variables Loan Status

- ~14% loans were charged off.
- ~83% loans are fully paid.



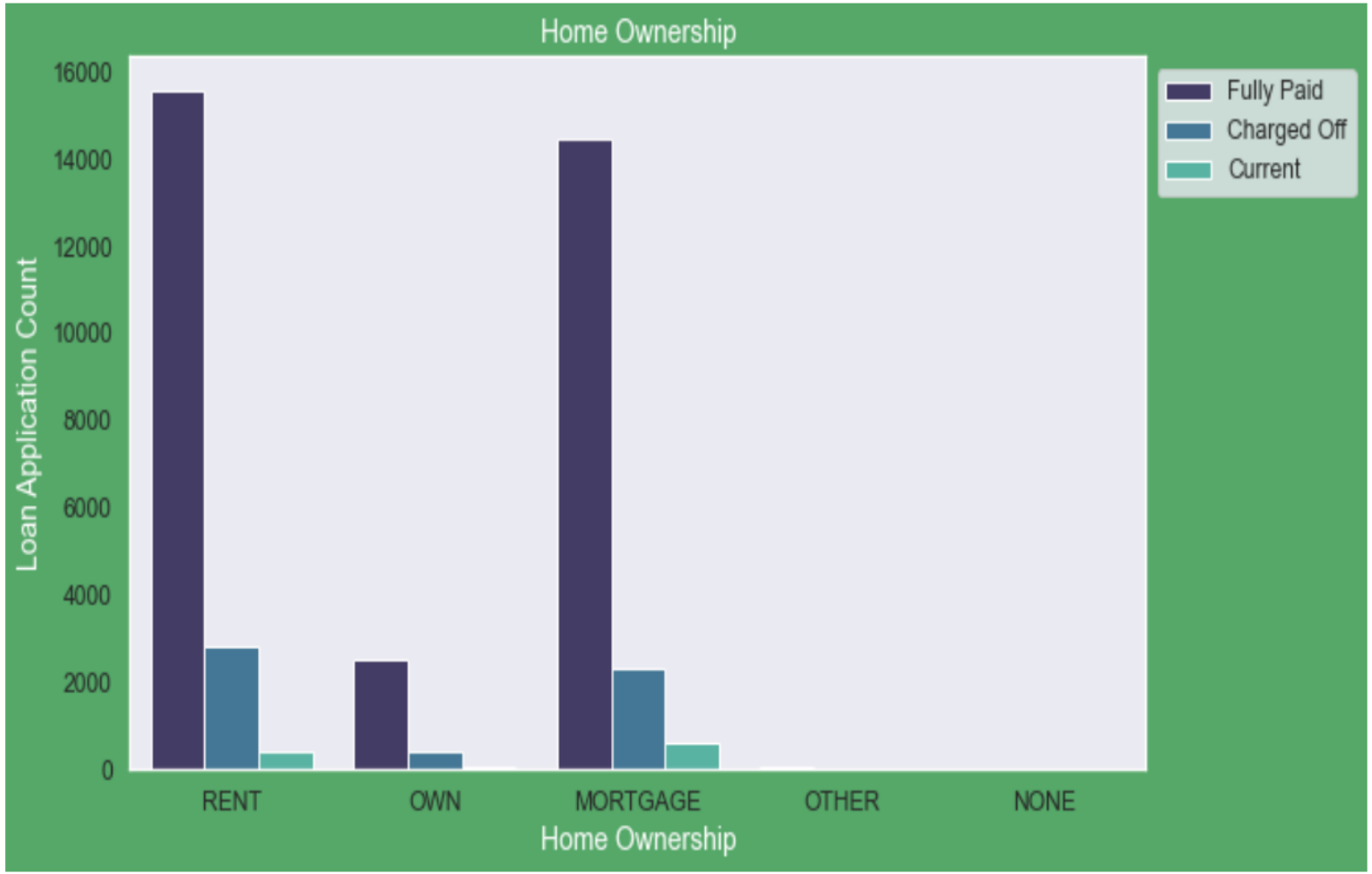
Univariate Analysis - Unordered Categorical Variables - Purpose Of Loan

- Most of the loans were taken for debt consolidation & paying credit card bill.
- Number of charged off count also high too for these loans.



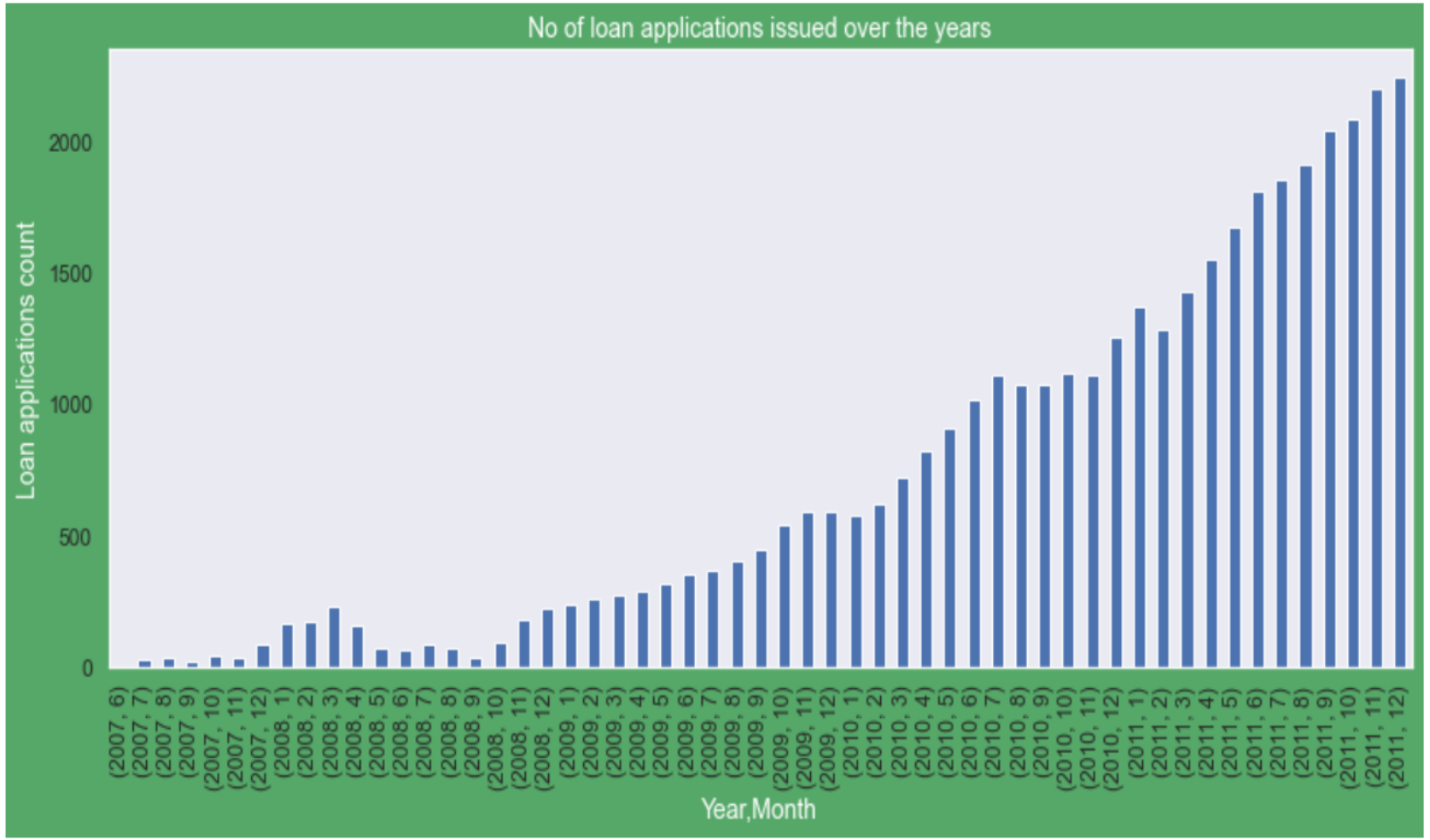
Univariate Analysis - Unordered Categorical Variables - Home Ownership

- Most of people living in rented house or have mortgaged their home.
- Applicant numbers are high from these categories so charged off is high too.



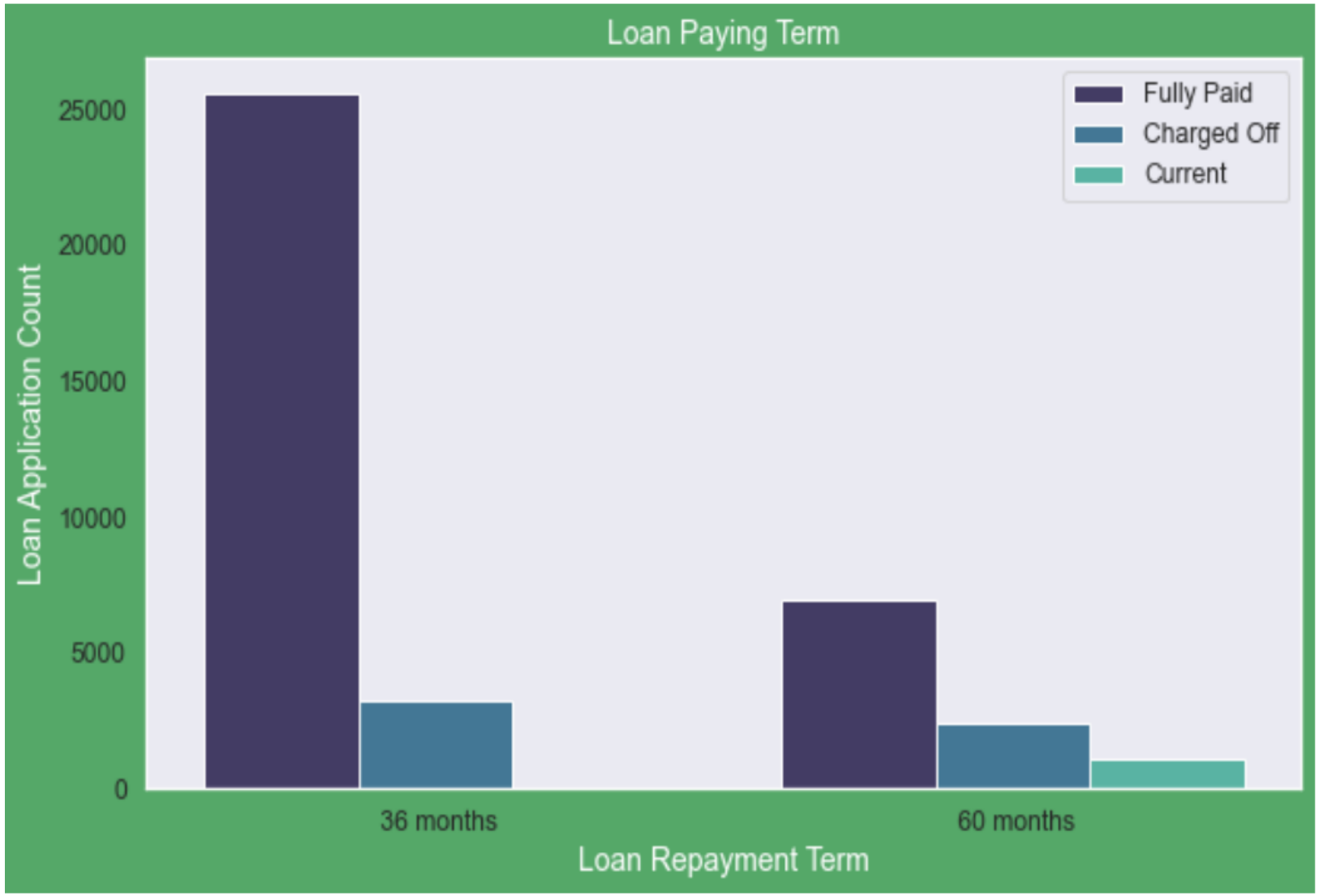
Univariate Analysis - Ordered Categorical Variables

- Count of loan application has increasing with every passing year.
- Loans issued in 2008(May-October) had a sharp fall.



Univariate Analysis - Ordered Categorical Variables- Loan Paying Term

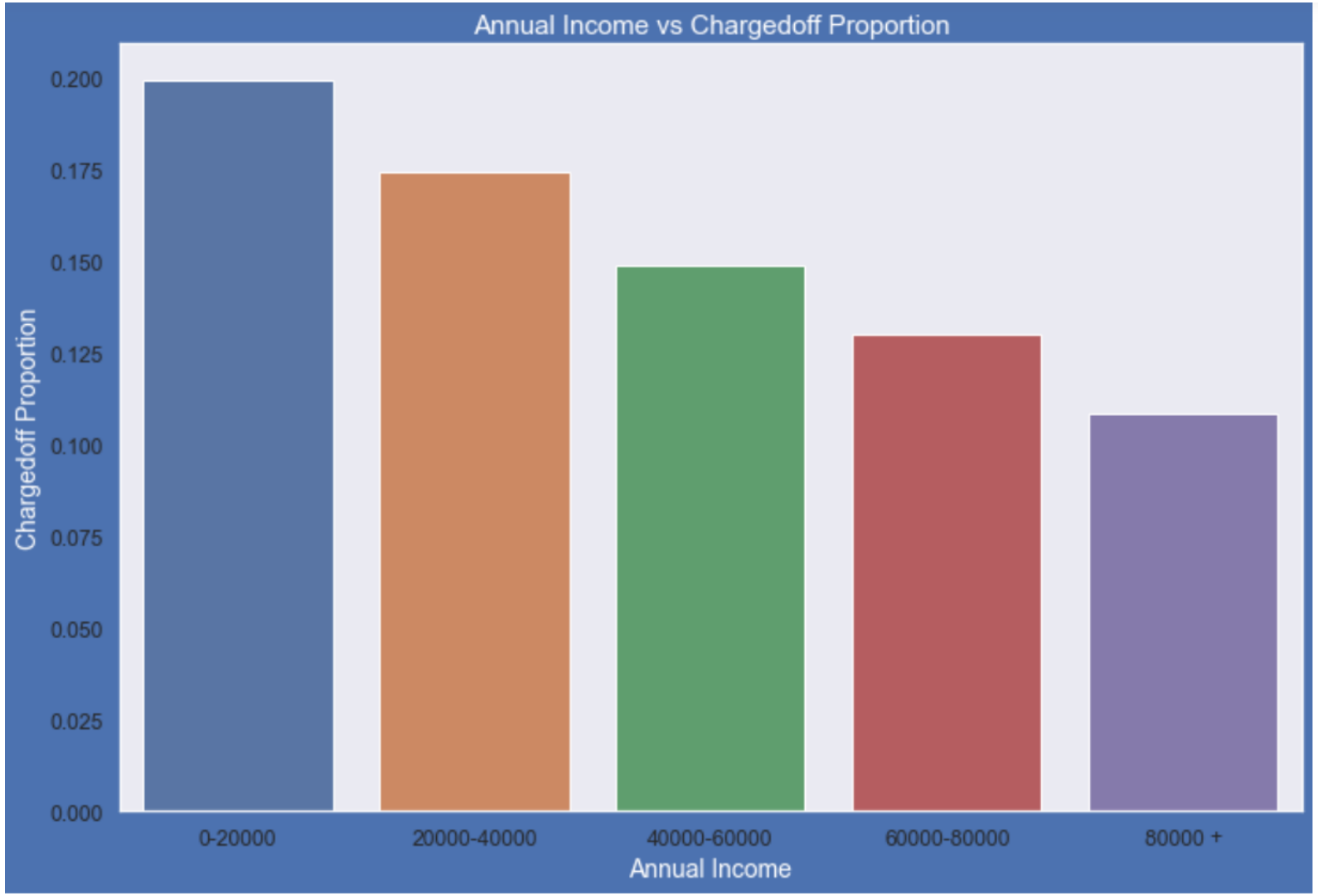
- % of Charged off is higher in 60 month terms.
- % of fully paid is higher in 36 month.



Bivariate Analysis on annual income against Chargedoff_Proportion

Annual Income	Charged Off Proportion
80000+	low
<20000	high

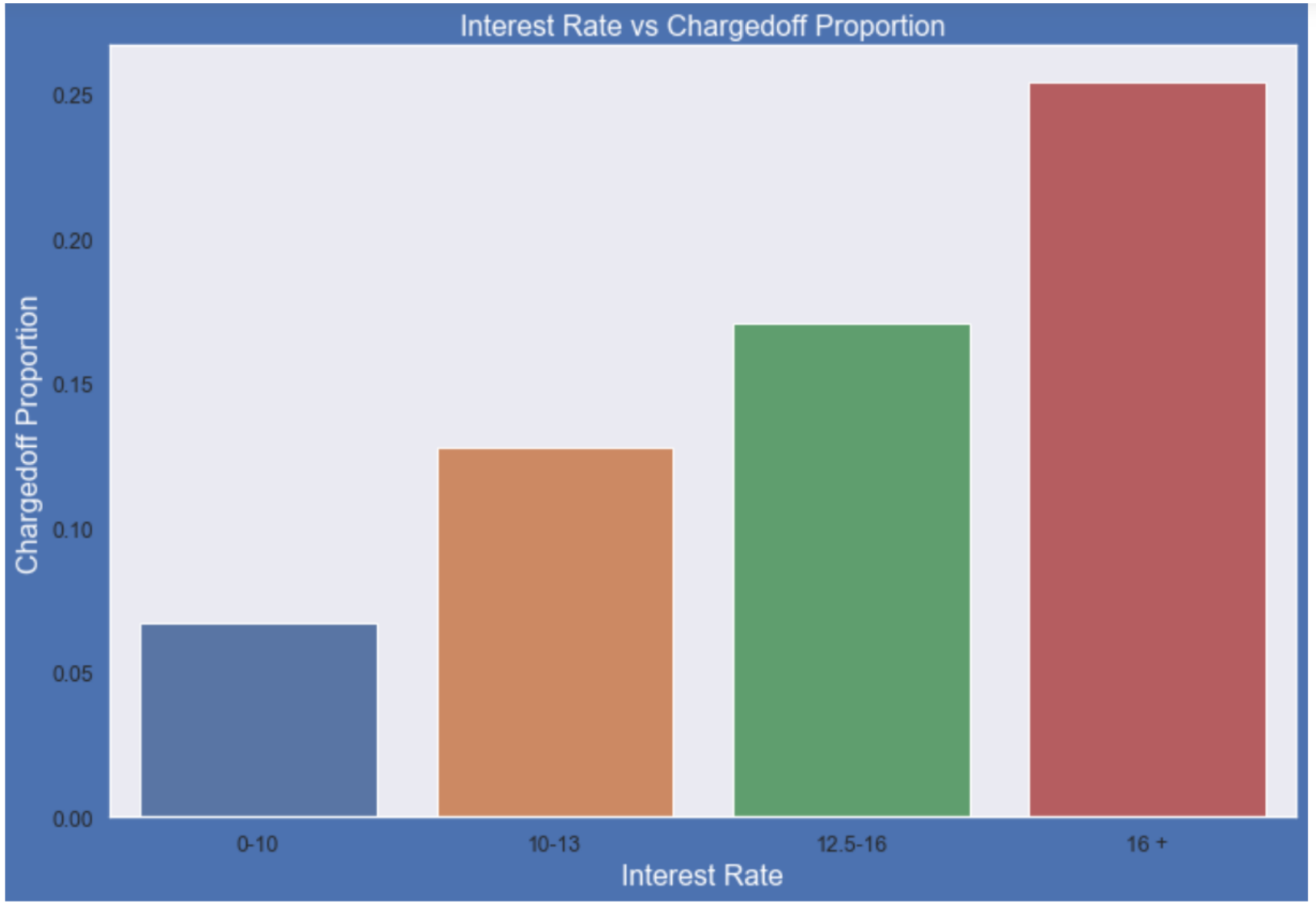
- Income range 80000+ has less chances of charged off.
- Income range 0-20000 has high chances of charged off.
- With increase in annual income charged off proportion got decreased.



Bivariate Analysis on interest rate against Chargedoff_Proportion

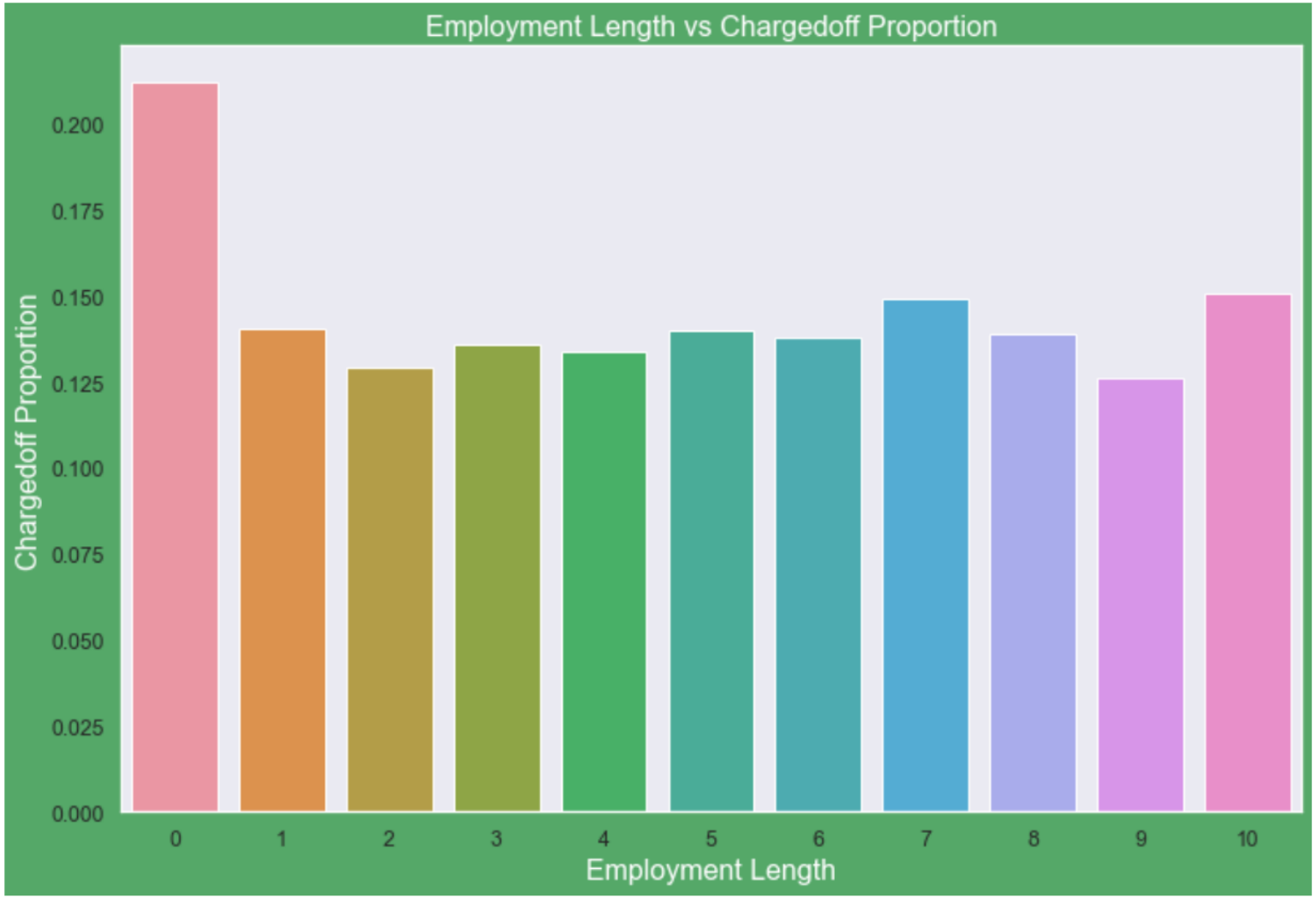
Interest Rate	Charged Off Proportion
16 +	high
0-10	Low

- Charged off proportion is increasing with higher interest rates.
- High interest loan having high chance to be charged off
- Low interest loan having low chance to be charged off



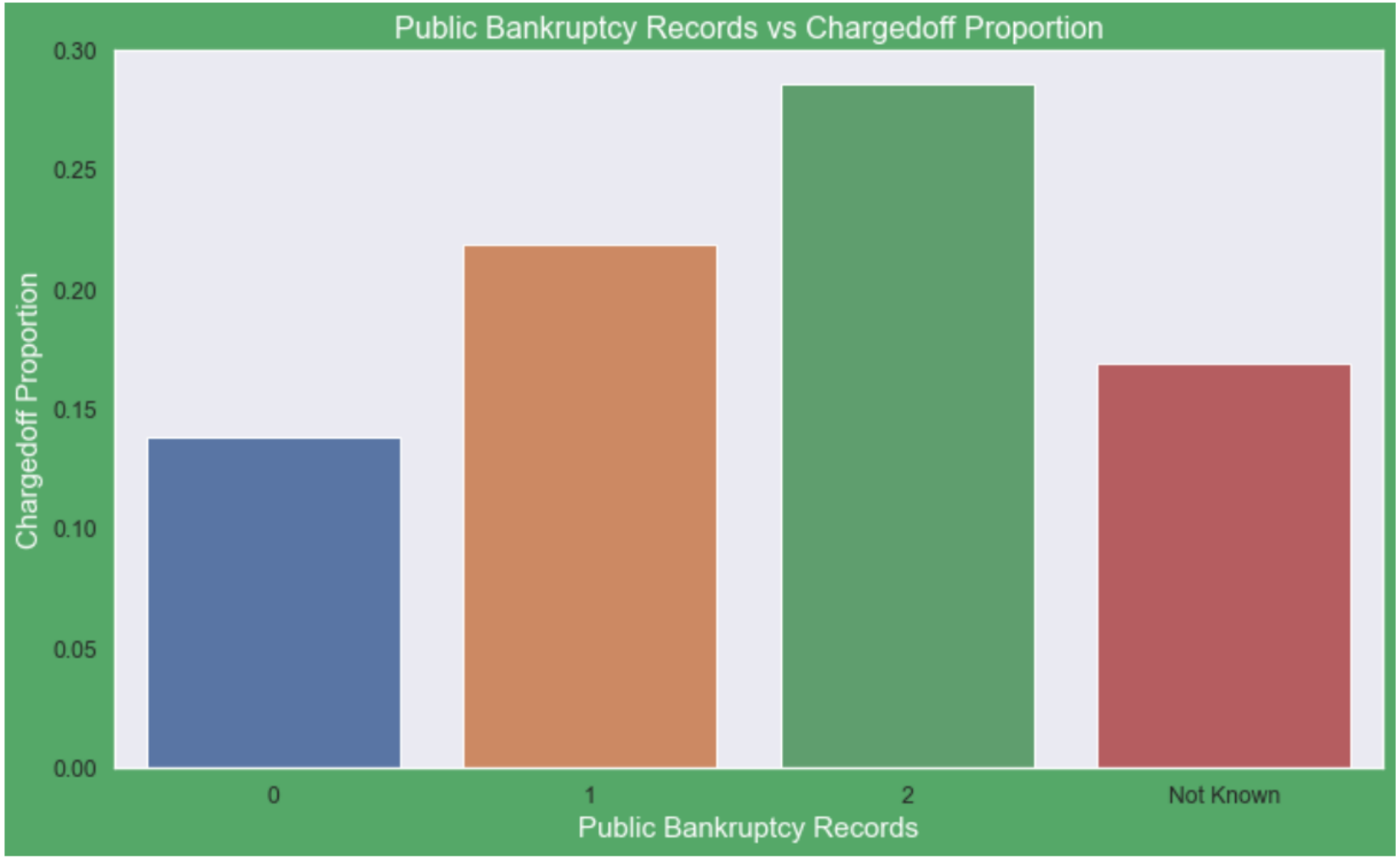
Bivariate Analysis on employment length against Chargedoff_Proportion

- Those who are not working or have less than 1 year of work experience have high chances of getting charged off.
- Rest of the applicants have almost same chances of getting charged off.



Bivariate Analysis on pub_rec_bankruptcies against Chargedoff_Proportion

- Public record of bankruptcies is 1, have charged off proportion higher than no record of bankruptcies.
- Public record of bankruptcies is 2 has even higher charged off proportion but those numbers are not significant to decide.
- Overall, who has defaulted in past will have more chances of defaulting in the future also.



Bivariate Analysis some more facts

- Purpose of loan vs Loan amount
  - Median,95th percentile,75th percentile of loan amount is highest for loan taken for small business purpose among all purposes.
  - Debt consolidation is second and Credit card comes 3rd.
  - Small business is not having outliers.

- Term of loan vs Interest Rate
  - Average interest rate is higher for 60 months loan term
  - Most of the loans issued for longer term had higher interest rates for repayment

Conclusion

- Interest rate is increasing with loan amount increase
- A-grade is a top letter grade for a lender to assign to a borrower.
- Higher percentage of loan amount is recovered when annual income is high.
- The ones getting 'charged off' have lower annual incomes than the ones who 'paid fully' for each grade.
- There is positive trend of increase in interest rate for loan with every passing year from 2007 to 2011.
- Public Bankruptcies record can have impact on lending .
- There was fall in loan application in 2008 may be due to global recession.