Lending Club Case Study – ML38

By

Sumit Vashistha

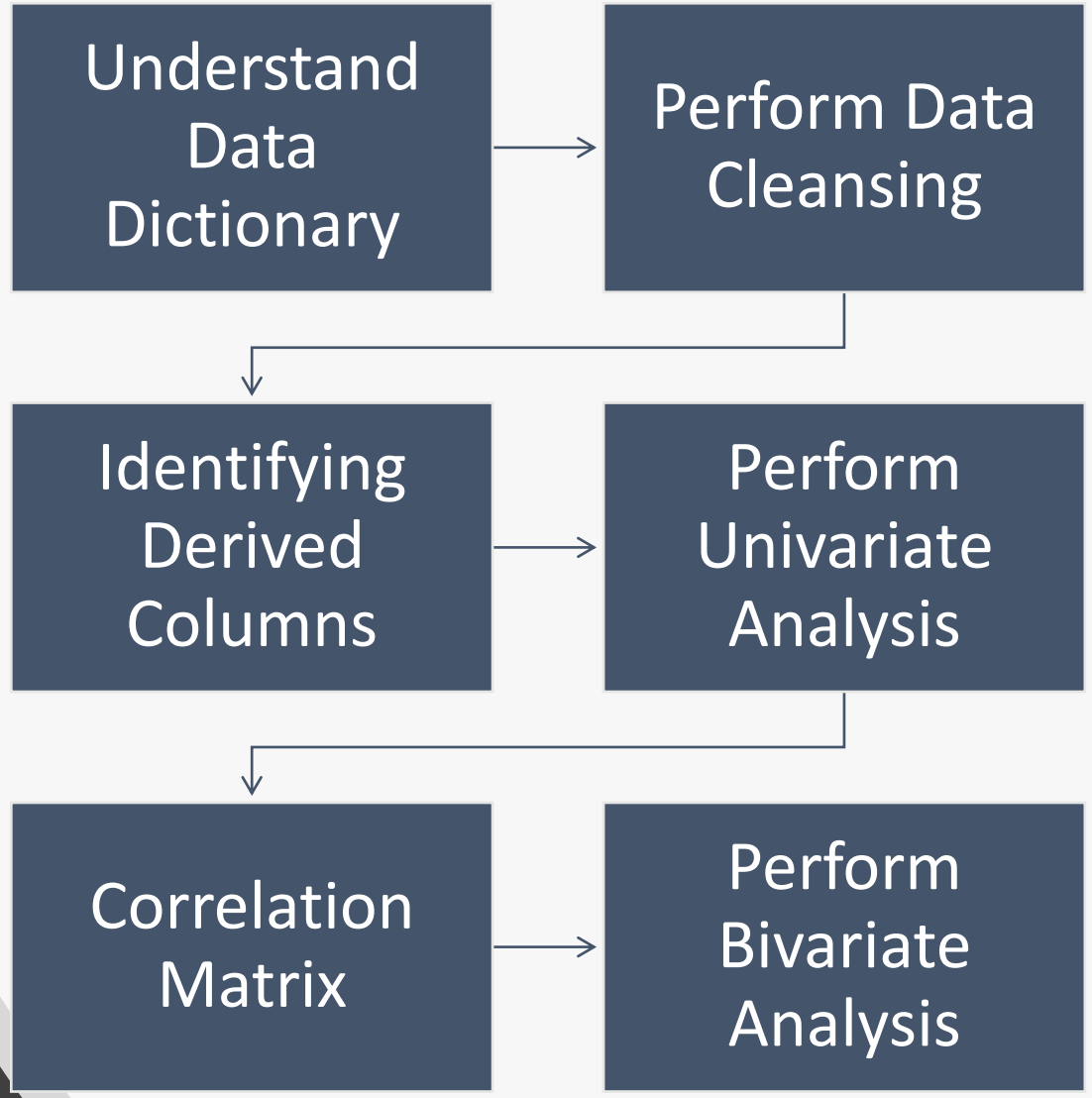
Shammi Kapoor

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

Steps followed for Analysis



Data Dictionary Understanding

- Total 116 columns provided in the data dictionary which represents the data in loan data set
- Identify some of the key columns

Attributes	Description	Reason for identifying as Key columns
Emp_length	Employment length of an individual	Can be used to understand how an individual pays of loan in beginning of his/her career as compared to having 10 years of exp
Loan_amnt	Actual Amount for which borrower applied for the loan	Can be used to 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	Can be studied if tenure has an impact on loan payment
Public_rec	Number of derogatory public records	Can be useful to identify risk status associated with a customer
pub_rec_bankruptcies	Number of public record bankruptcies	These customers can be denied loan if we have sufficient data of previous bankruptcies
Purpose	Actual need for which loan has been applied for	Studied to find where we are observing charged off like someone starting business / buying house / credit card payments
home_ownership	Current accommodation of the loan applicant	Can be studied to identify if people living in rented accomation vs own have higher chances of defaulting due to extra load.

Data Cleansing

- Identifying columns having null values
  - Around 54 columns identified to be empty
  - Dropping these columns from data frames
- Removing % characters in some of the columns int_rate and revol_util as these can be used for numeric calculations
- Changing some of the columns to numeric data type like loan_amnt, funded_amnt, int_rate, funded_amount_inv, installment, annual_inc

Identifying Derived Columns

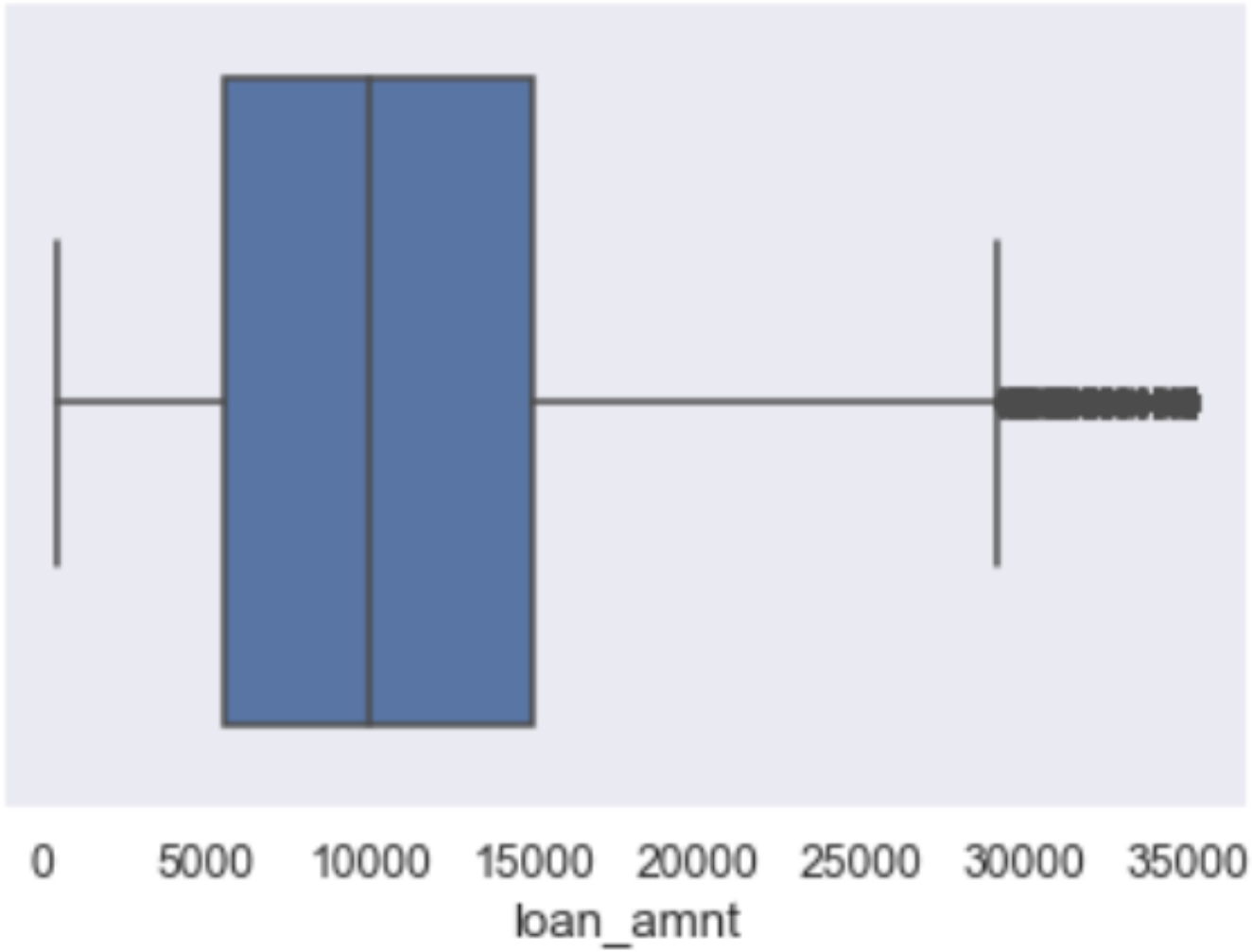
Loan Issued Date(issue_d) can be used derived two new columns like month and year

Identify various trends like number of loan application increase/decreases by year

Interest Rate has increased/decrease by year

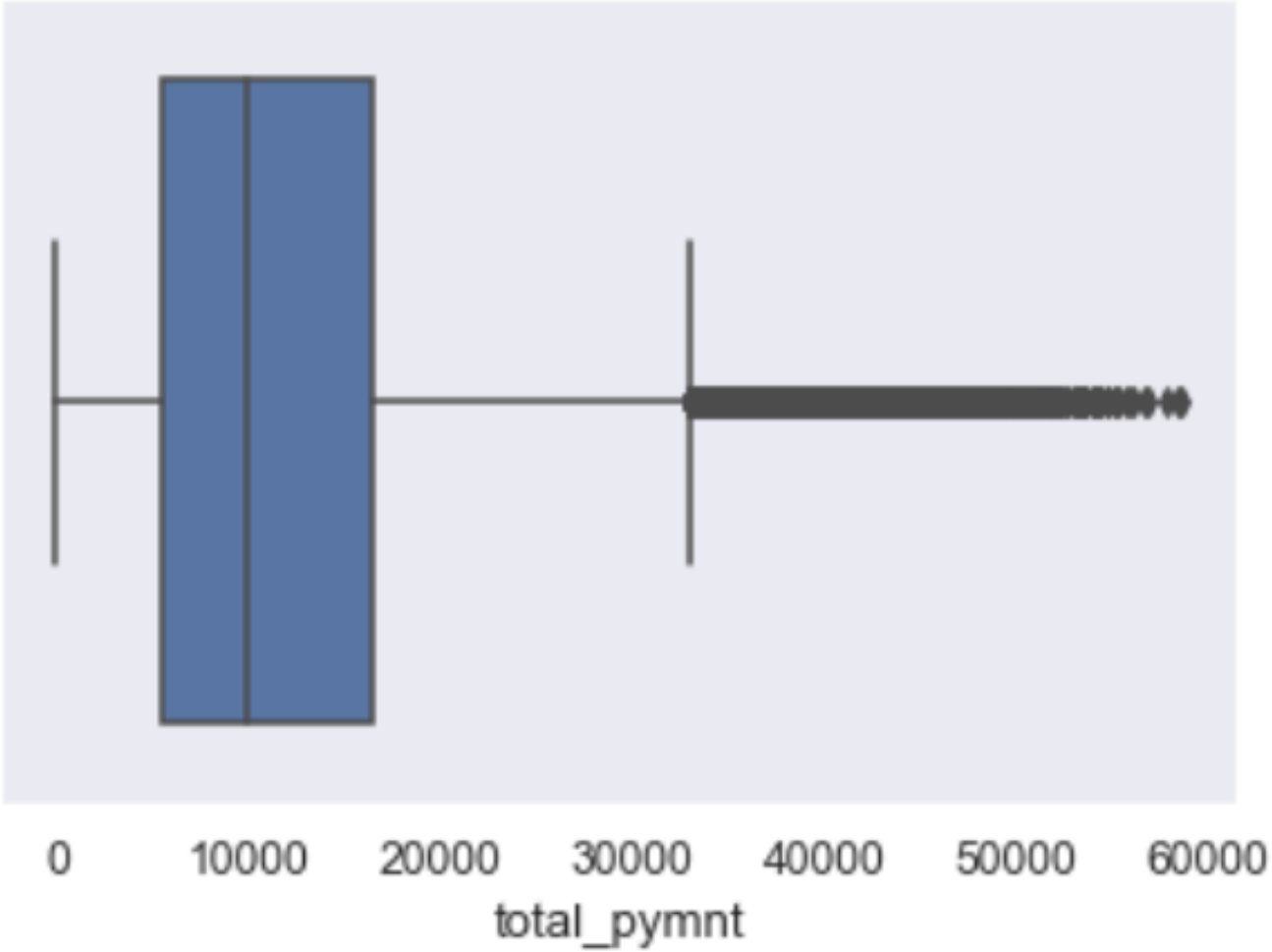
Univariate Analysis - Loan Amount

- Average Loan amount is around 11000 and 75% is around 15000$
- Most of the population applied for loan below 15000$ whereas as the maximum is 35000$.



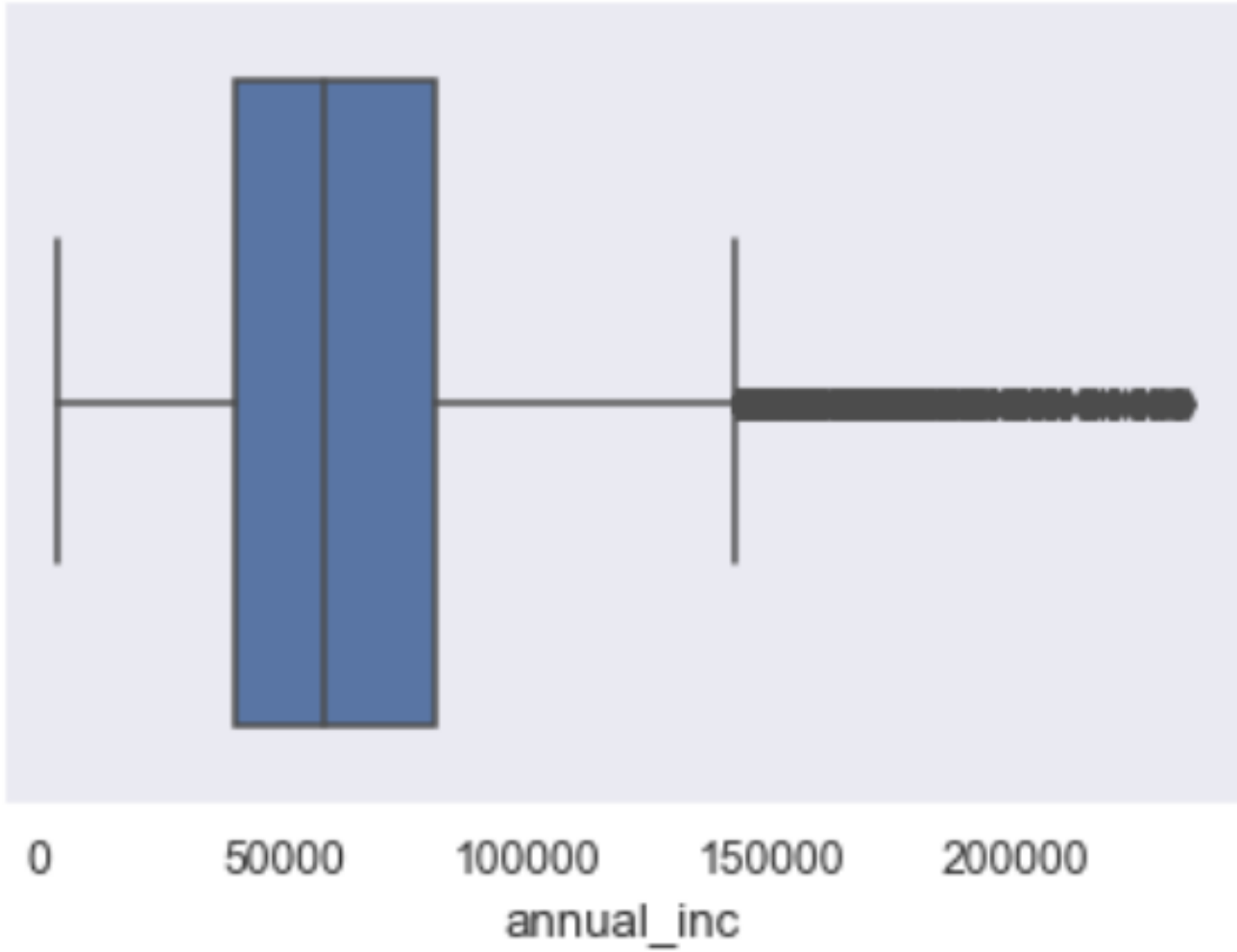
Univariate Analysis - Total Payment

- Most of the people have done payment around 12000$ which currently looks like most of the people have fully paid their loan
- 99% looks on higher side around 58000K which is higher than the max loan amount applied.
- Currently interest component looks like to be reason of higher payment but need more study to identify the reason for higher total payment for 75% above cases.



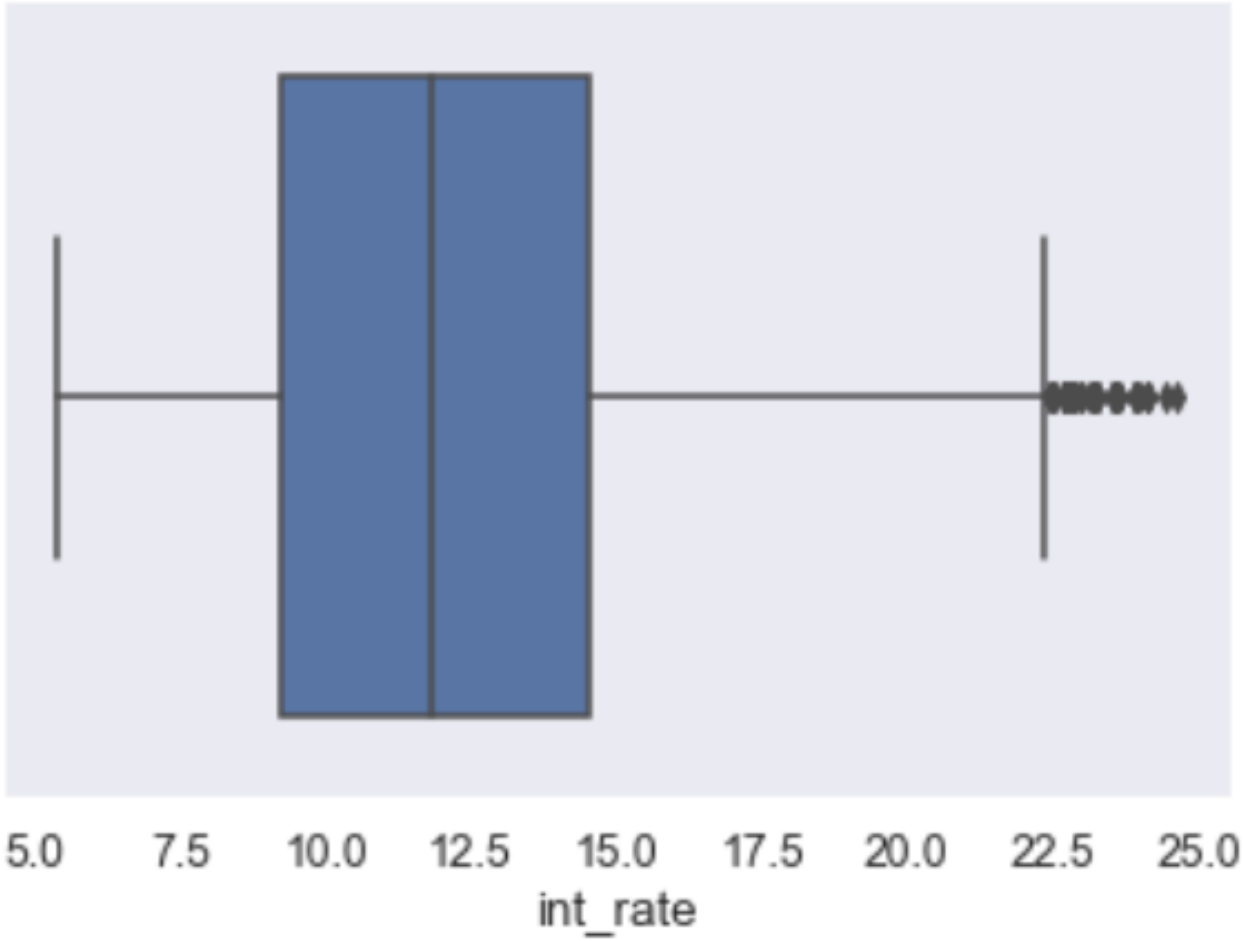
Univariate Analysis - Annual Income

- On describing the Annual Income found outlier component which needs to be removed in order to get correct picture of data.
- On removing the Outlier, the 99% also looks good and dropped from 600K to 200K
- Average annual income looks like around 65000 which is almost double to the loan amount applied in most of the cases



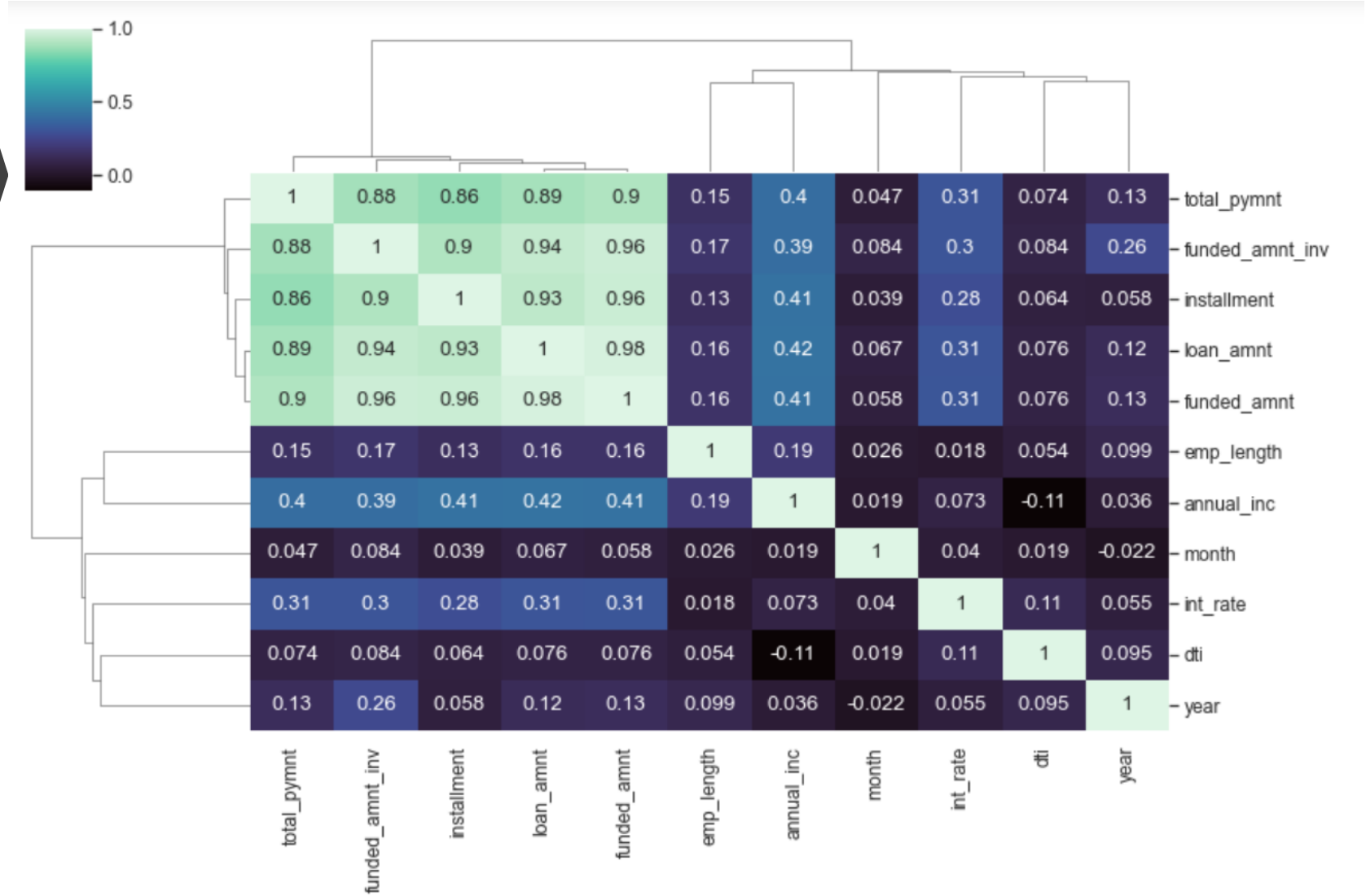
Univariate Analysis - Interest Rate

- Average interest rate falls under 12%
- Also interest rate of for 75% goes from 14% to 25%.
- Need to identify if higher loan amount leads to higher interest rate



Correlation Matrix

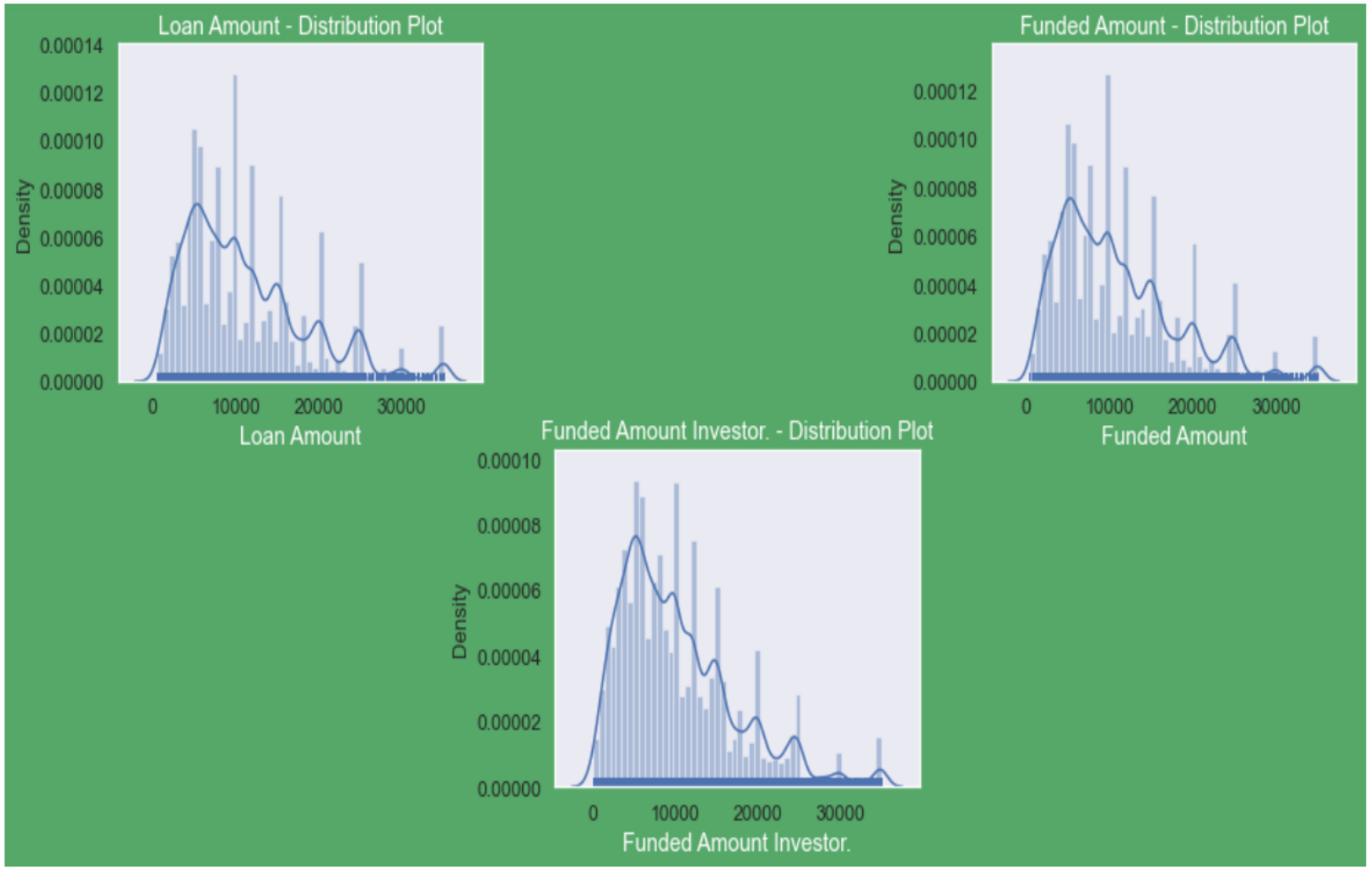
- Loan Amount, Investor Amount and Funding Amount are strongly correlated
- Annual Income and DTI is negatively correlated
- Positive Correlation between Annual Income and Employment Years means income increase with experience



Univariate Analysis - Three loan amount fields using distribution plot

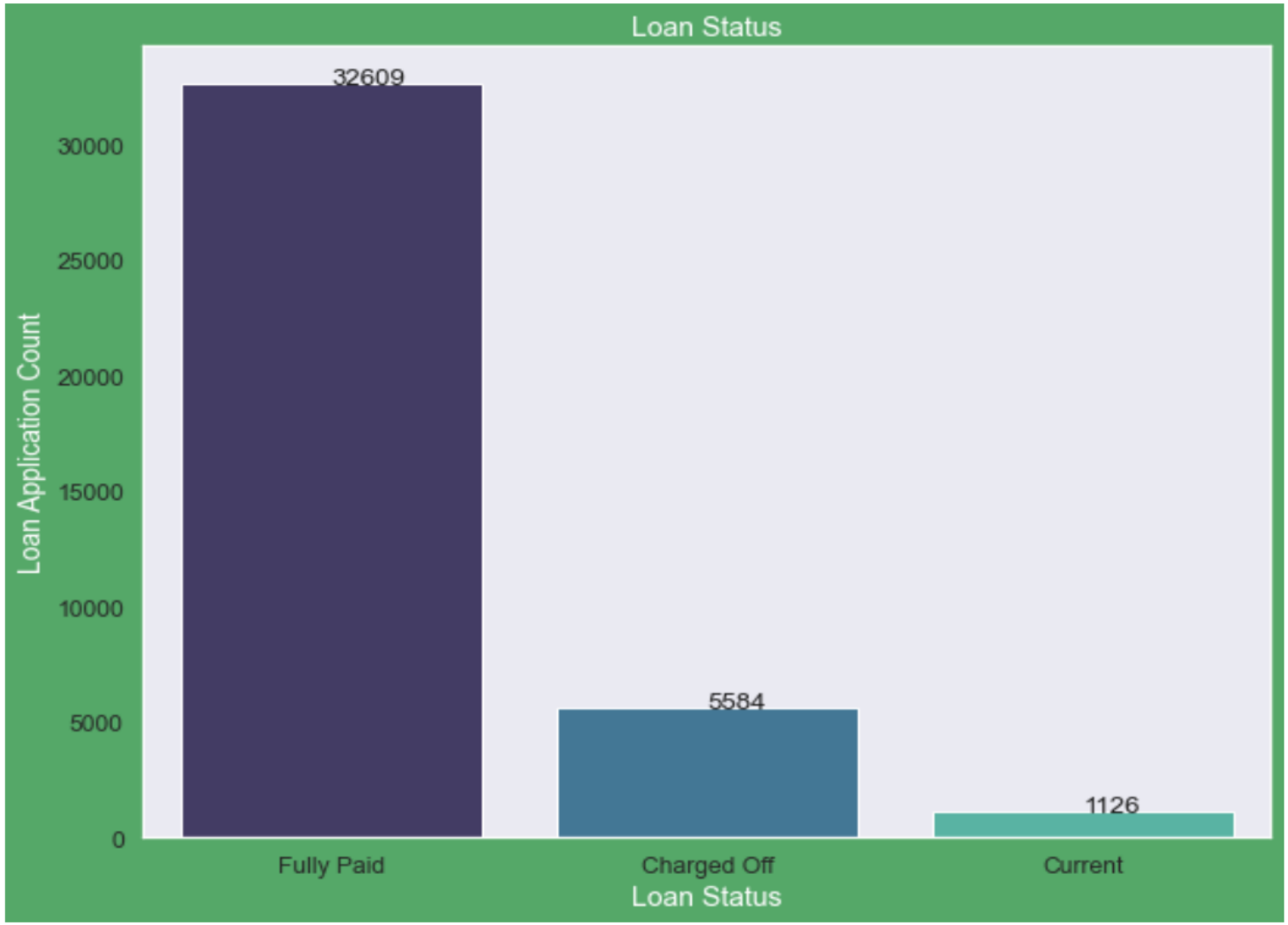
- Distribution of three loan different amounts viz Loan Amount, Funded Amount and Funded Amount Investor look very similar

- 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



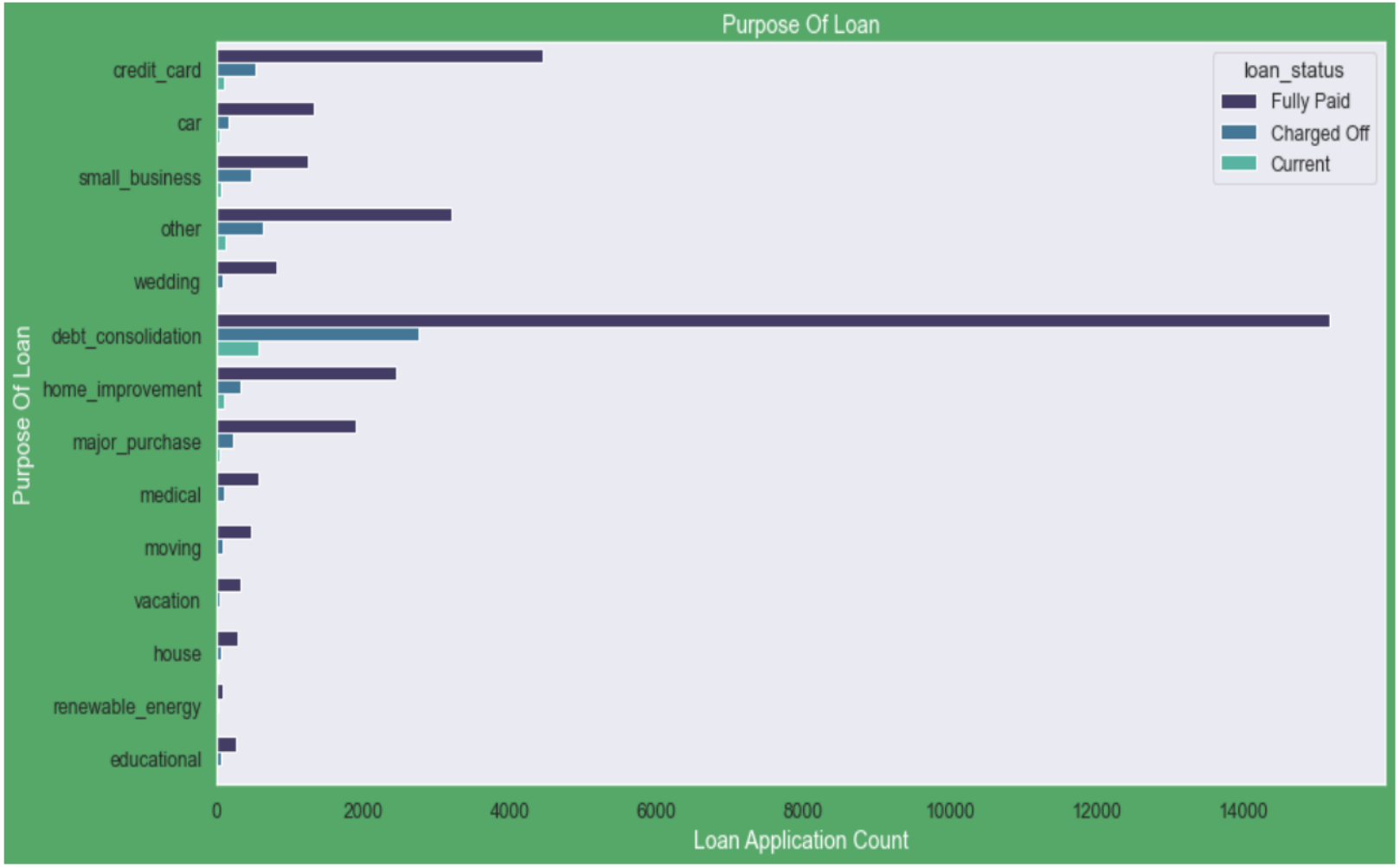
Univariate Analysis Unordered Categorical Variables Loan Status

- This plot shows that close to 14% loans were charged off out of total loan issued



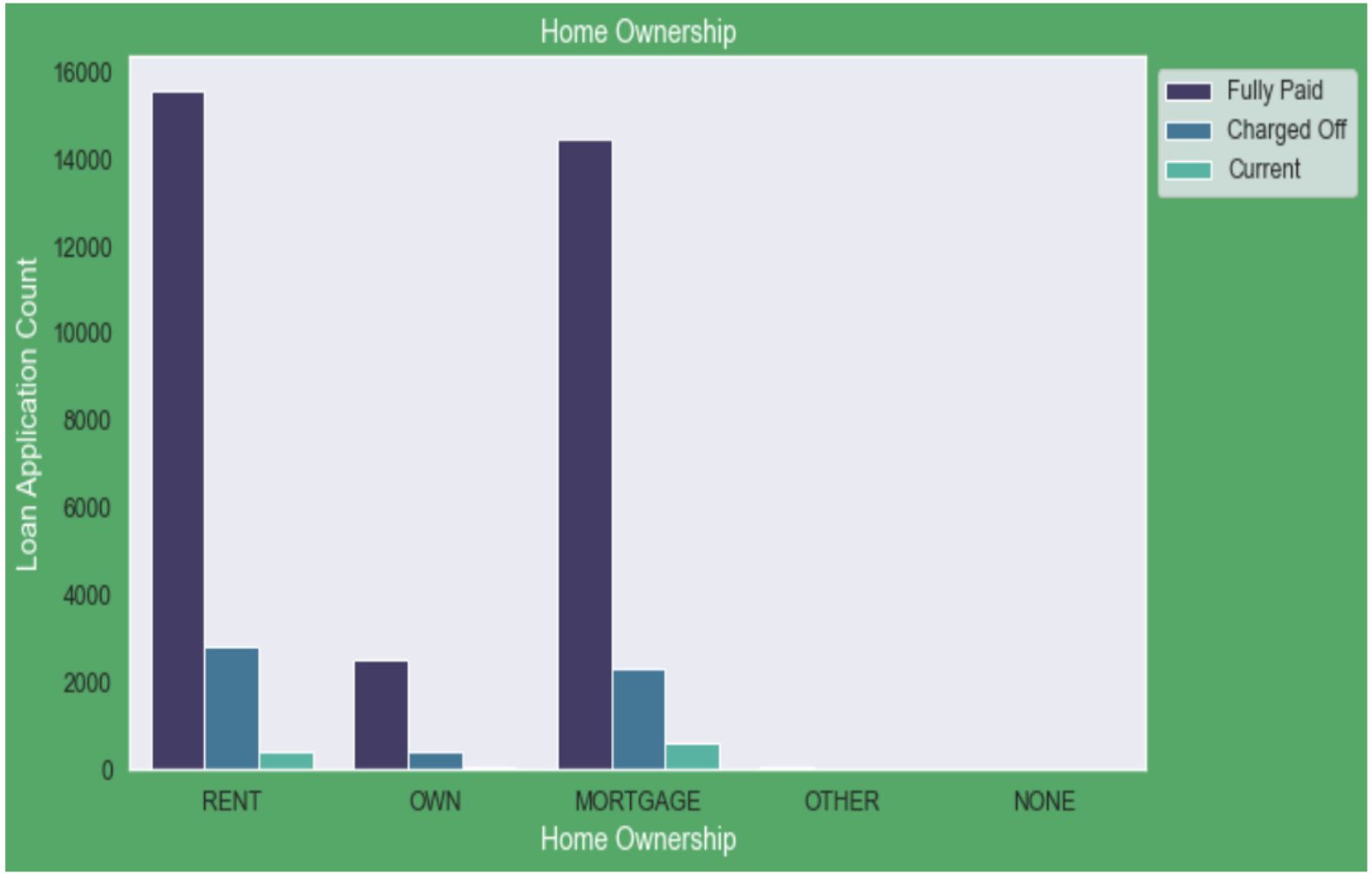
Univariate Analysis - Unordered Categorical Variables - Purpose Of Loan

- This plot shows that most of the loans were taken for the purpose of debt consolidation & paying credit card bill.
- Number of charged off count also high too for these loans.



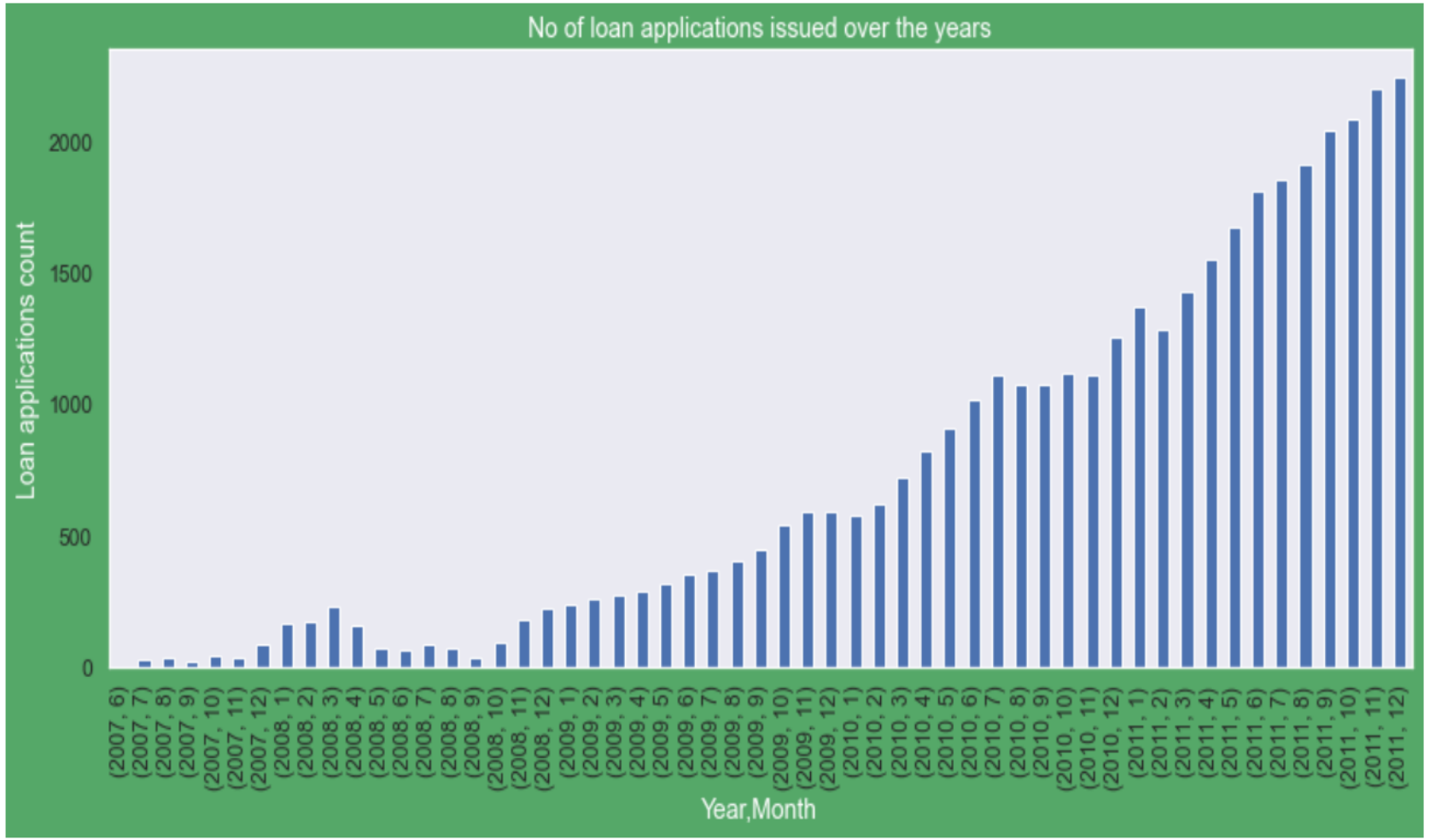
Univariate Analysis - Unordered Categorical Variables - Home Ownership

- This plot shows that most of them living in rented home or mortgazed their home.
- Applicant numbers are high from these categories so charged off is high too.



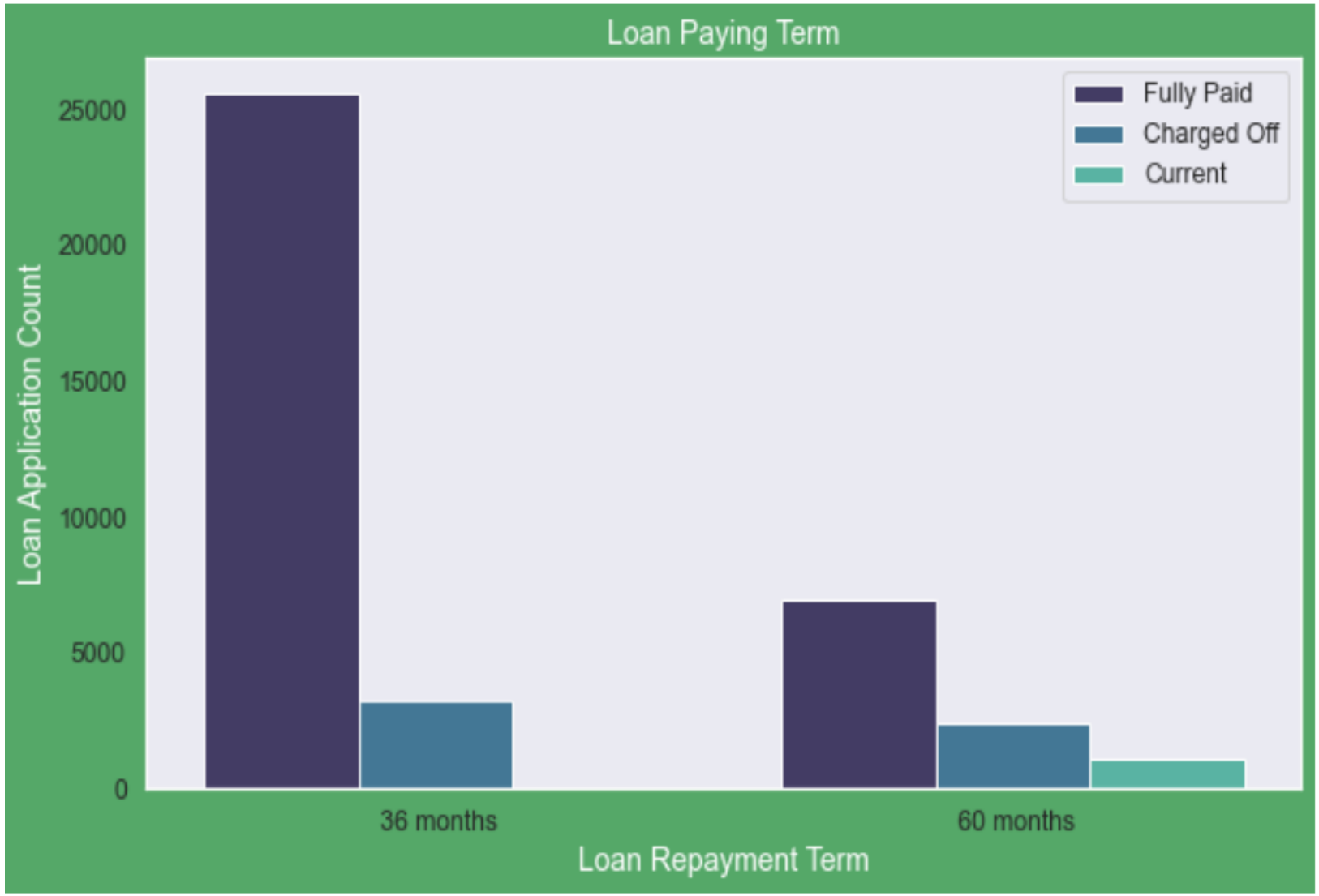
Derived Column - Ordered Categorical Variables

- By looking at the plot we can see count of loan application has increasing with every passing year.
- Also increase in number of loan applications are adding more to number of charged off applications.
- Loans issued in 2008(May-October) had a sharp fall.



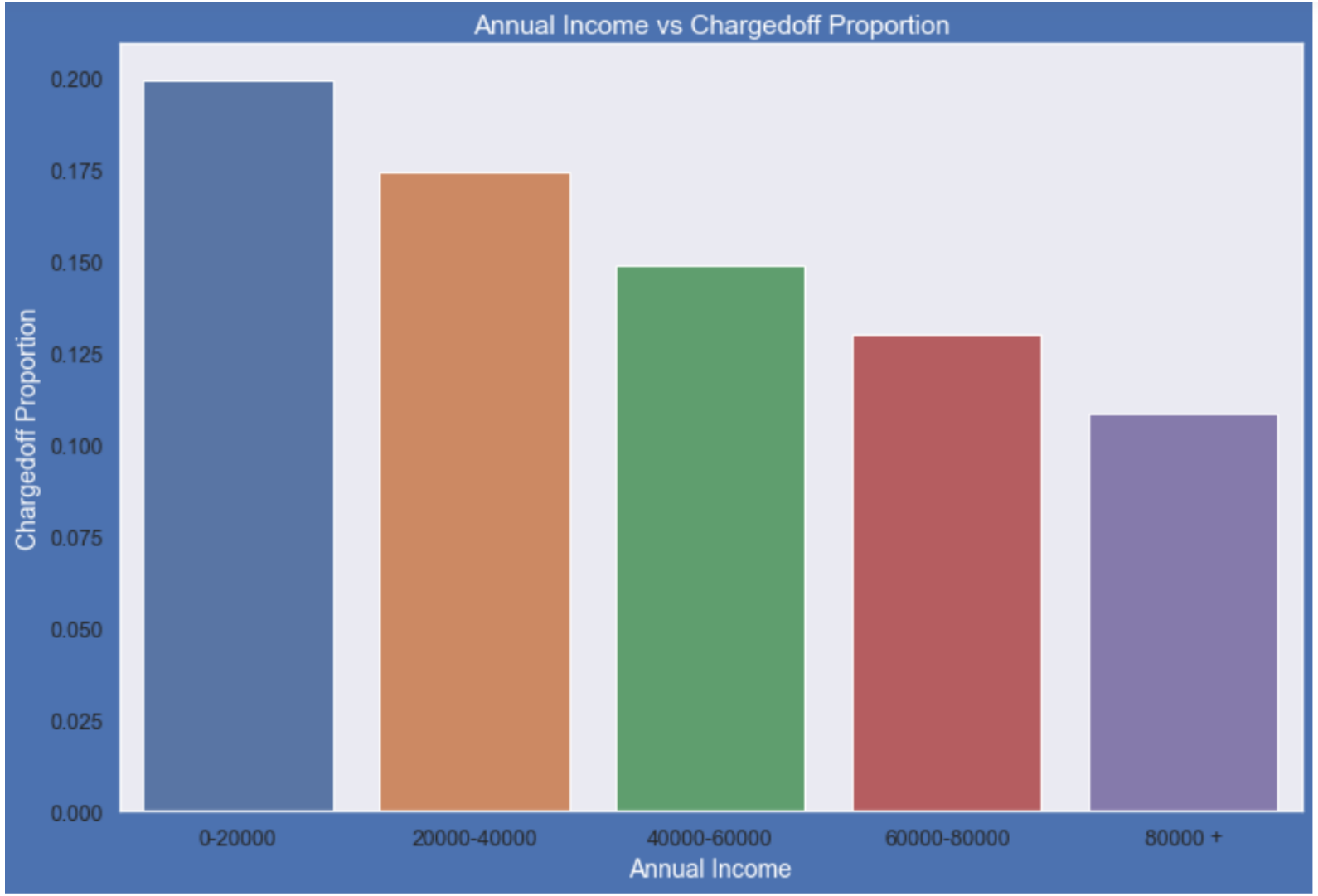
Univariate Analysis - Ordered Categorical Variables- Loan Paying Term

- This plot shows that those who had taken loan to repay in 60 months had more % of number of applicants getting charged off as compared to applicants who had taken loan for 36 months.



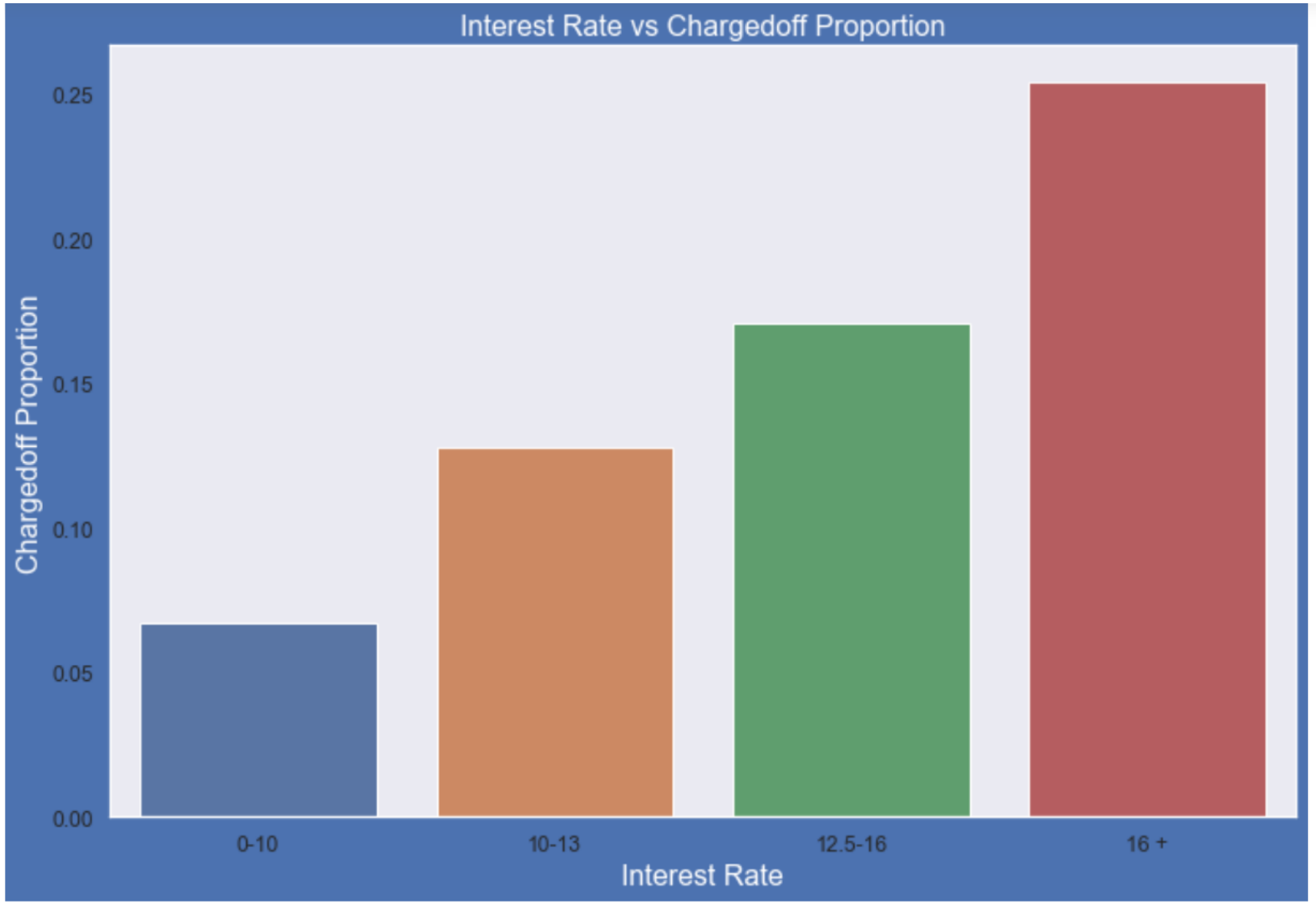
Bivariate Analysis on annual income against Chargedoff_Proportion

- 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.



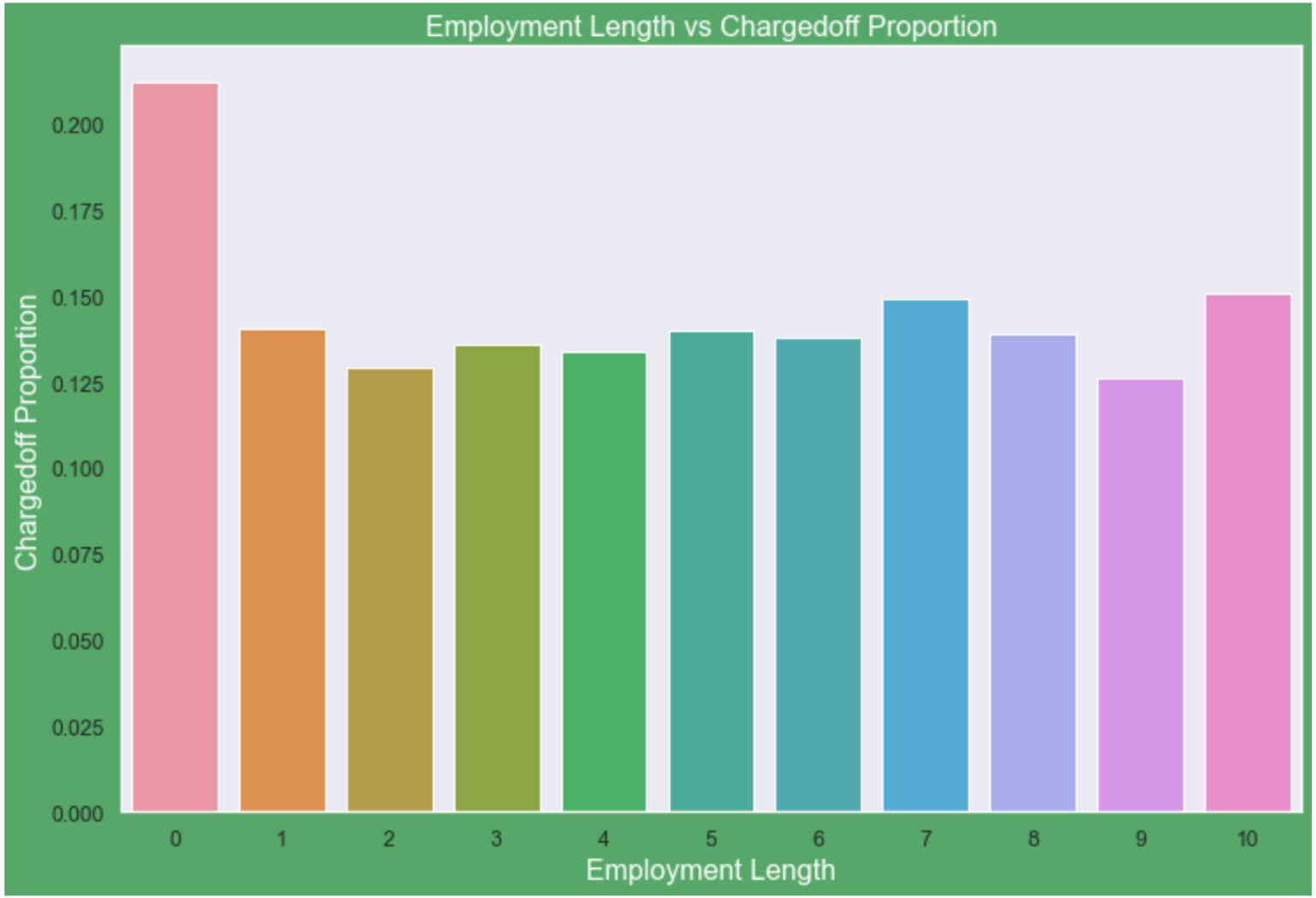
Bivariate Analysis on interest rate against Chargedoff_Proportion

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



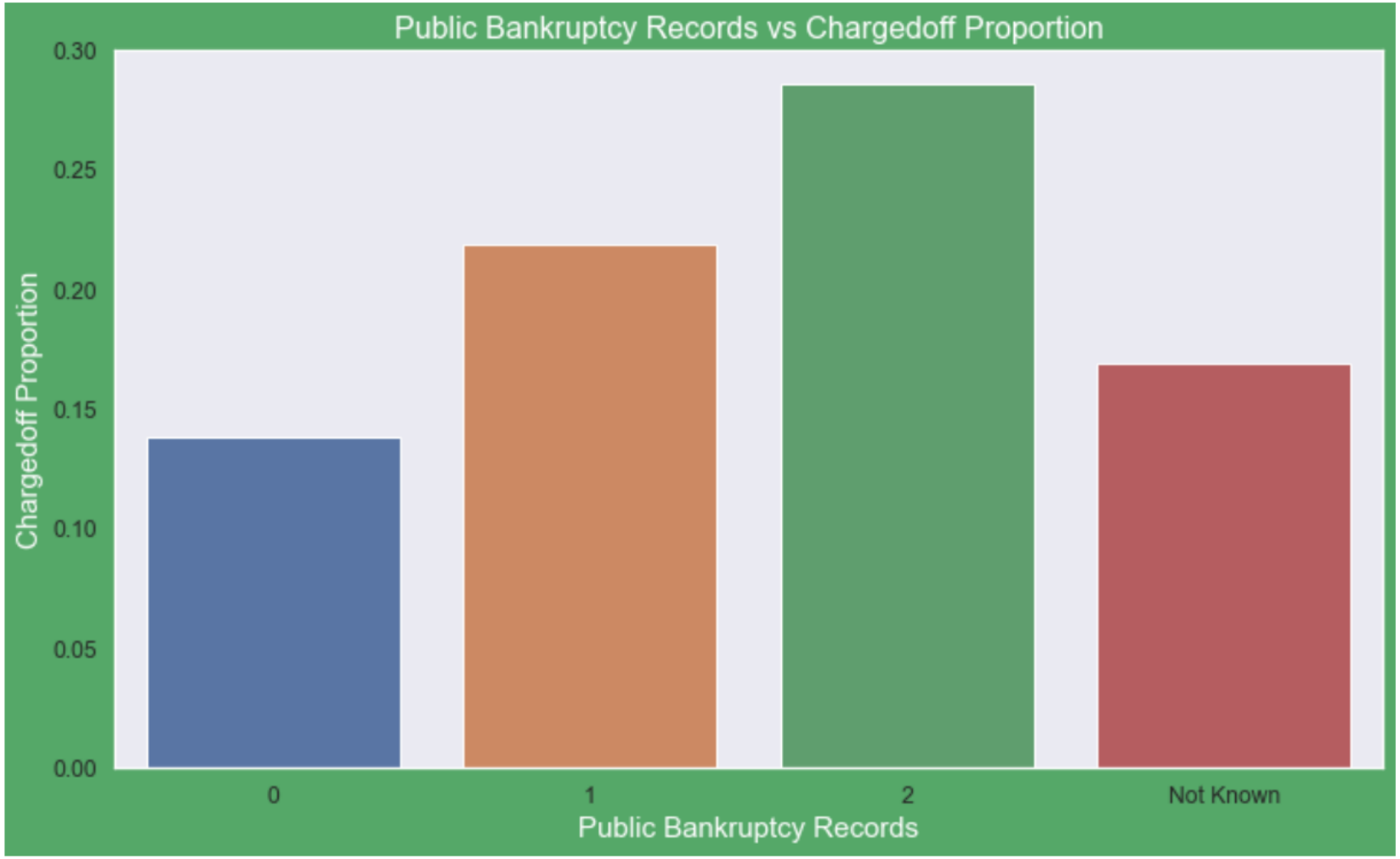
Bivariate Analysis on employment length against Chargedoff_Proportion

- 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 same chances of getting charged off.



Bivariate Analysis on pub_rec_bankruptcies against Chargedoff_Proportion

- Those who already have pub_rec_bankruptcies value 1, have charged off proportion higher than who have no pub_rec_bankruptcies.
- pub_rec_bankruptcies count 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.

- 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.