upGrad
#LifeKoKaroLift

Lending Club Case Study

Submitted by: Vishal Arora

Agenda

- Problem Statement and business objectives
- Application of EDA approach
- Data understanding
- Data cleaning and manipulation
- Derive Columns for analysis
- Univariate analysis
- Bi-variate analysis
- Mutivariate analysis and correlation matrix

Problem Statement

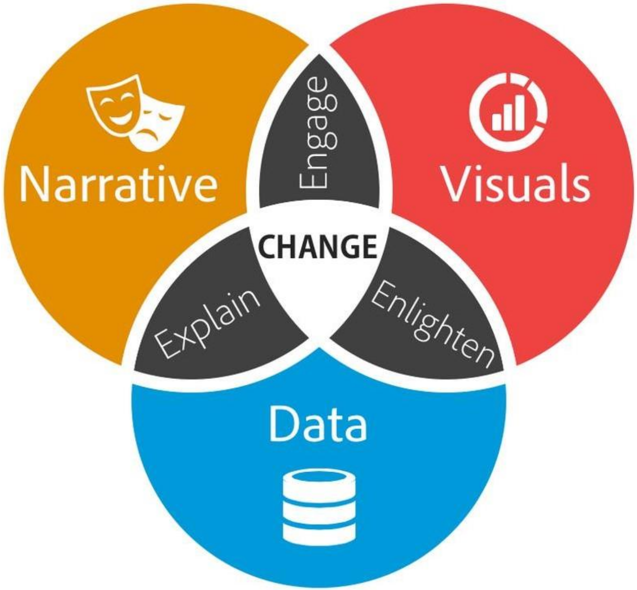
upGrad

The case study is based on a consumer finance company. The company specializes in lending different 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:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. 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

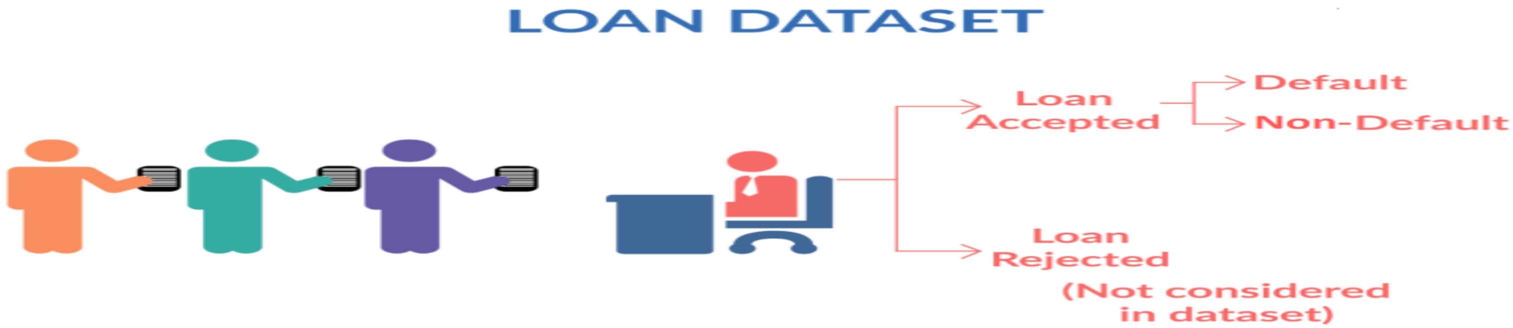
The data shared by UpGrad contains the information about past loan applicants and whether they 'defaulted' or not.

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.



Problem Statement

upGrad



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:

1.1 Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

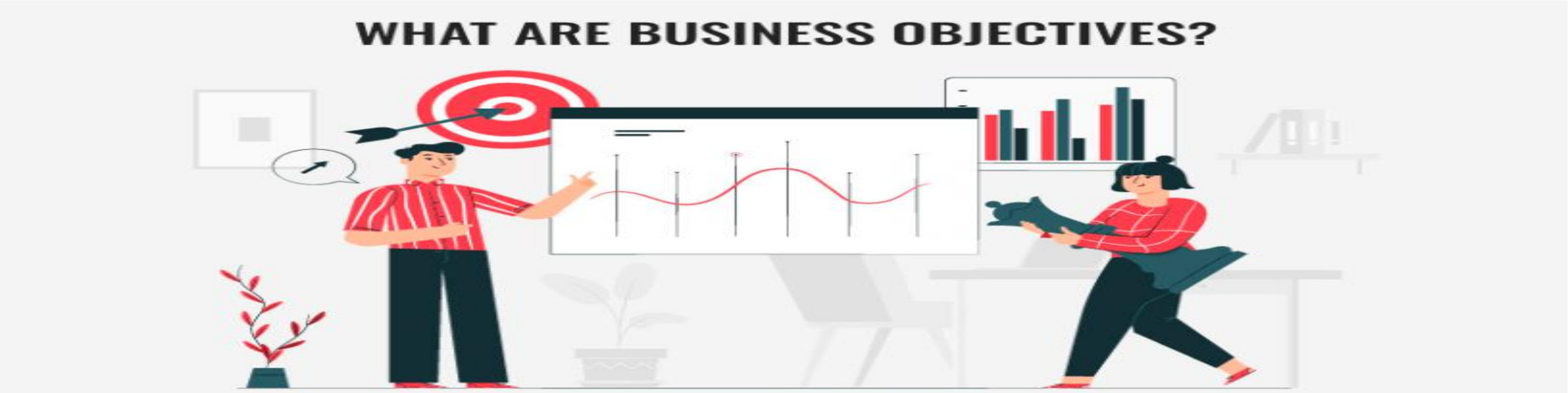
1.2 Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

1.3 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)

Business Objective

upGrad



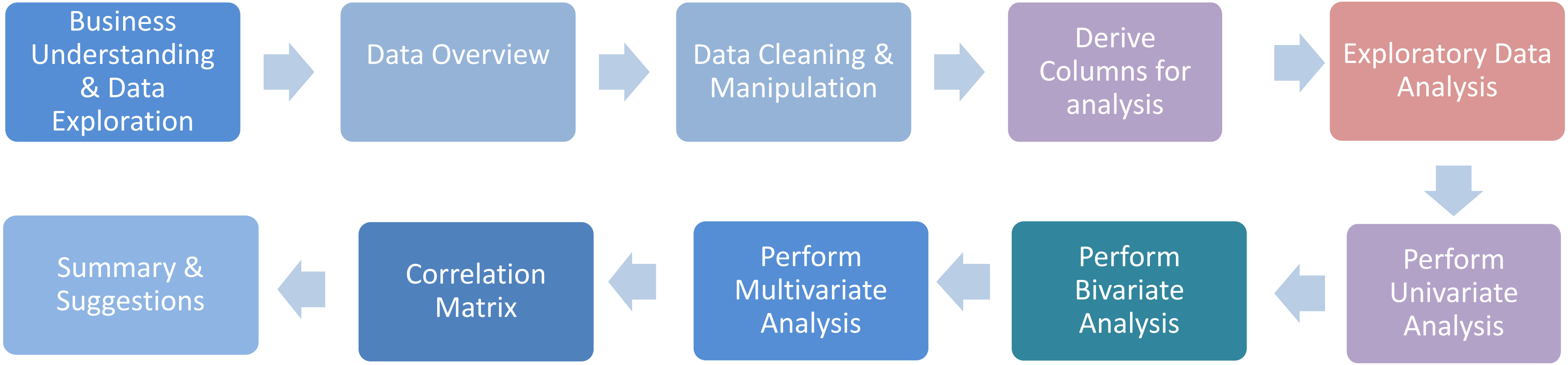
Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

The **business objective** is to identify risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

Application of EDA approach

upGrad



Business Understanding & Data Exploration

Data Overview

Data Cleaning & Manipulation

Derive Columns for analysis

Exploratory Data Analysis

Perform Univariate Analysis

Perform Bivariate Analysis

Perform Multivariate Analysis

Correlation Matrix

Summary & Suggestions

Data Overview

```
(39717, 111) <class 'pandas.core.frame.DataFrame'> RangeIndex: 39717 entries,
0 to 39716 Columns: 111 entries, id to total_il_high_credit_limit dtypes: float64(74),
int64(13), object(24) memory usage: 33.6+ MB None id int64 member_id int64
loan_amnt int64 funded_amnt int64 funded_amnt_inv float64 ... tax_liens float64
tot_hi_cred_lim float64 total_bal_ex_mort float64 total_bc_limit float64
total_il_high_credit_limit float64 Length: 111, dtype: object
```

Data Cleaning & Manipulation

Data Cleaning & Manipulation

upGrad

- ❑ There are no duplicate rows in loan dataset
- ❑ Number of empty columns 54 (There are 54 columns have all missing values)
  - ❑ Dropping sequence all columns have null values (Column from 53 to 104)
  - ❑ Drop additional columns which do not make any sense in the entire analysis
    - ❑ Reasons for each fields
    - ❑ member_id: This is unique field and does not have any relevance to our analysis
    - ❑ emp_title: This does not have relevance as there is no standard and business sense (Such as : Rydeer, Veolia Transportaton)
    - ❑ pymnt_plan: This contains only "n" and does not make any sense in analysis
    - ❑ # url: This has no business relevance

Data Cleaning & Manipulation

- ❑ desc: This is non standard remarks by data entry person which do not have any relevance in analysis
- ❑ zip_code: This contains information where first 3 digits are available and last 2 digits are xx example 860xx. We have better field next to it addr_state which we can use for analysis.
- ❑ next_pymnt_d: This fiels contains 1140 records out out of 39717 and do not add any value to our analysis
- ❑ collections_12_mths_ex_med: This contains 0 or NA and do not add any value to analysis
- ❑ mths_since_last_major_derog: This contains NA and do not add any value to analysis
- ❑ tax_liens: This contains 0 or NA and do not add any value to analysis

Data Cleaning & Manipulation

- There are no duplicate rows in loan dataset
- Number of empty columns 54 (There are 54 columns have all missing values)
  - tot_hi_cred_lim : This contains NA and do not add any value to analysis
  - total_bal_ex_mort : This contains NA and do not add any value to analysis
  - total_bc_limit : This contains NA and do not add any value to analysis
  - total_il_high_credit_limit : This contains NA and do not add any value to analysis
- Number of empty rows 0
- Drop additional columns which are not needed further in analysis and reasons for it
  - application_type: This contains only value " INDIVIDUAL" which do not add any value to analysis
  - policy_code: This contains only value "1" which do not add any value to analysis
  - initial_list_status: This contains only value "f" which do not add any value to analysis

Data Cleaning & Manipulation

- ❑ We have left finally with Rows: 39717 and Columns: 42
- ❑ pub_rec_bankruptcies is cleaned for further analysis
  - ❑ There were 697 Null values(na) which are replaced with Not Known
- ❑ Remove % symbol from interest rate column so that it can be used in calculations
  - ❑ Int_rate
  - ❑ revol_until
- ❑ convert amount columns into numeric data to find correlation among important variables.['loan_amnt','funded_amnt','int_rate','funded_amnt_inv','installment','annual_inc','dti','emp_length','total_pymnt']

We are remaining with 39717 Rows and 42 Columns for conducting analysis

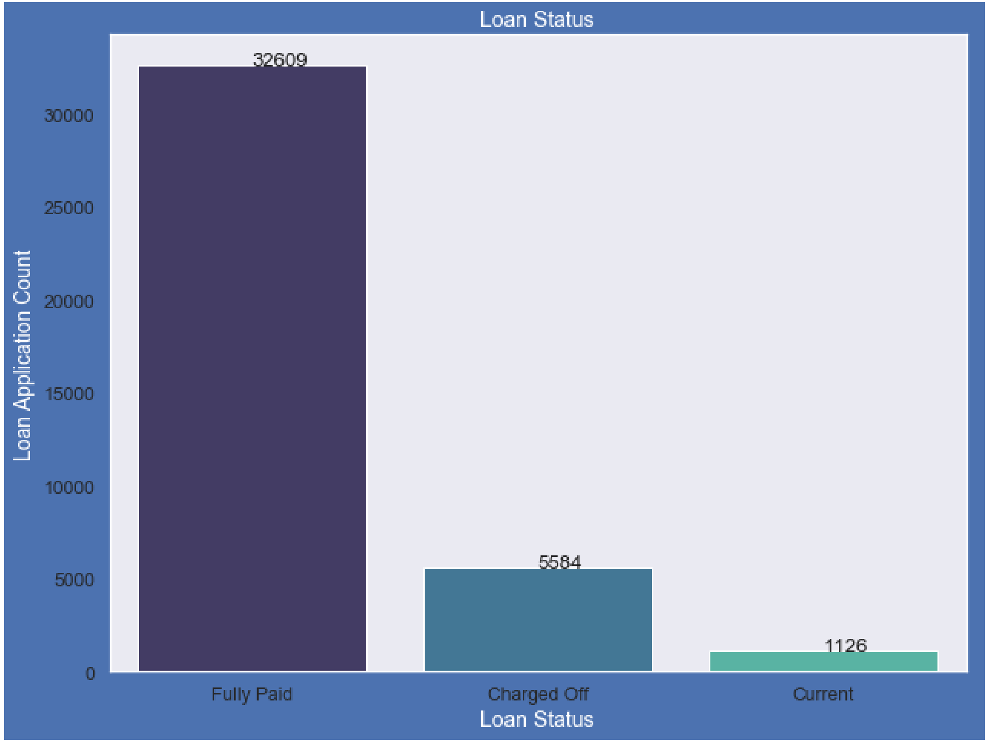
Derive Columns for Analysis

- ❑ Two new columns month and year were derived from issue date
- ❑ Categorize loan amounts into buckets which will help in analysis further in bivariate analysis
  - ❑ '0-7000', '7000-14000', '14000-21000', '21000-28000', '28000 +'
- ❑ Categorize interest rates into buckets which will help in analysis further in bivariate analysis.
  - ❑ '0-10', '10-13', '12.5-16', '16 +‘
- ❑ Categorize dti(Debt to interest ratio) into buckets for bivariate analysis.
  - ❑ '0-5', '05-10', '10-15', '15-20', '25+'

Finally we are remaining with 39717 Rows and 48 Columns for conducting analysis

Data Analysis

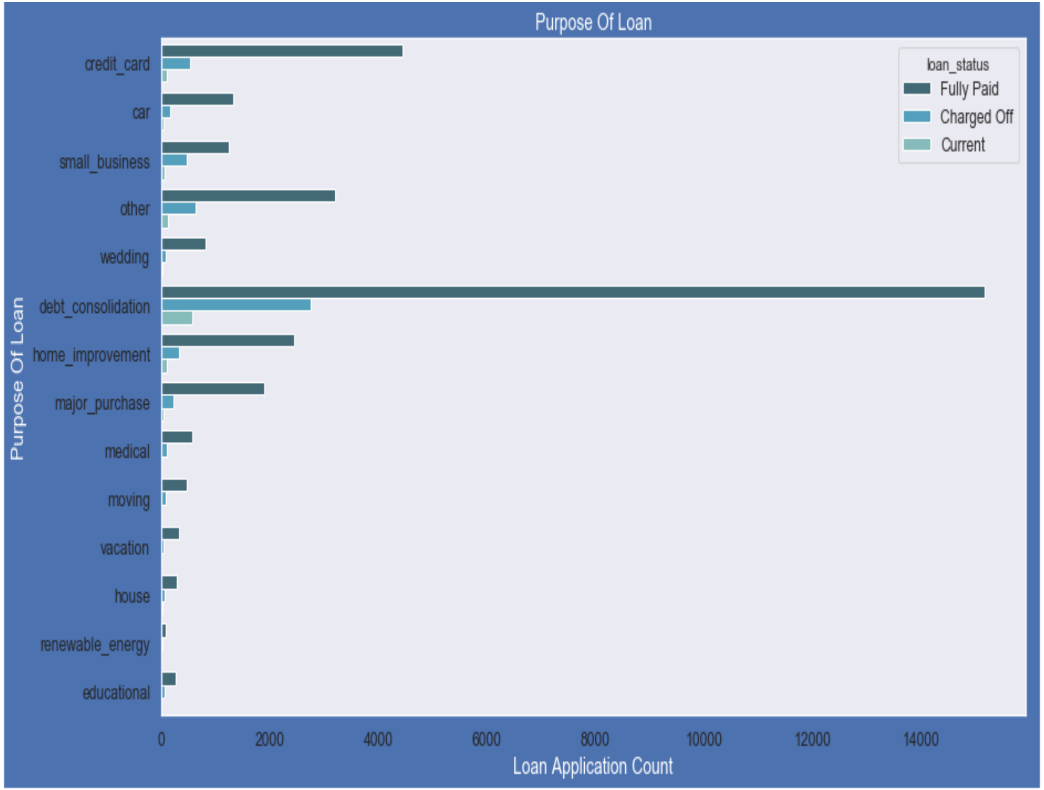
Univariate Analysis



- ❑ 14% loans were charged off out of total loan issued.
- ❑ 83% loans were fully paid out of total loan issued.

Fully Paid	82.96 %
Charged Off	14.17 %
Current	02.87 %

Univariate Analysis



Purpose of Loans

❑ Larger portion of loans were taken for the purpose of debt consolidation & paying credit card bill.

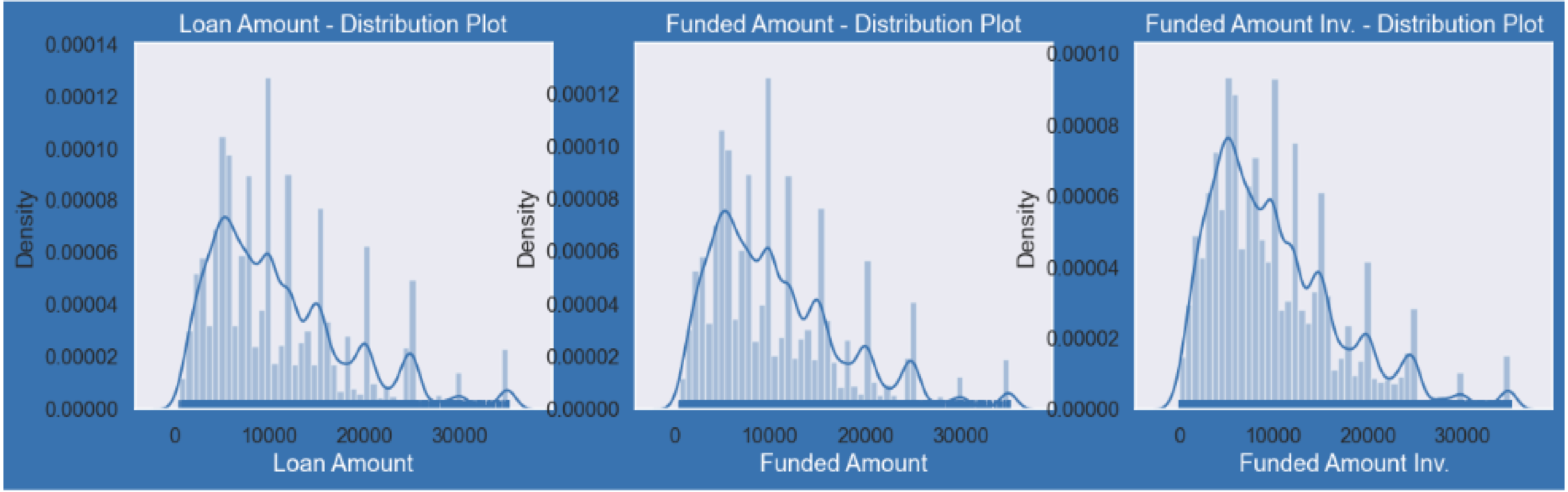
❑ Number of charged off counts also high for these loans.

Loan Purpose Percentage

debt_consolidation	46.93%
credit_card	12.92%
other	10.05%
home_improvement	7.49%
major_purchase	5.51%
small_business	4.60%

Univariate Analysis

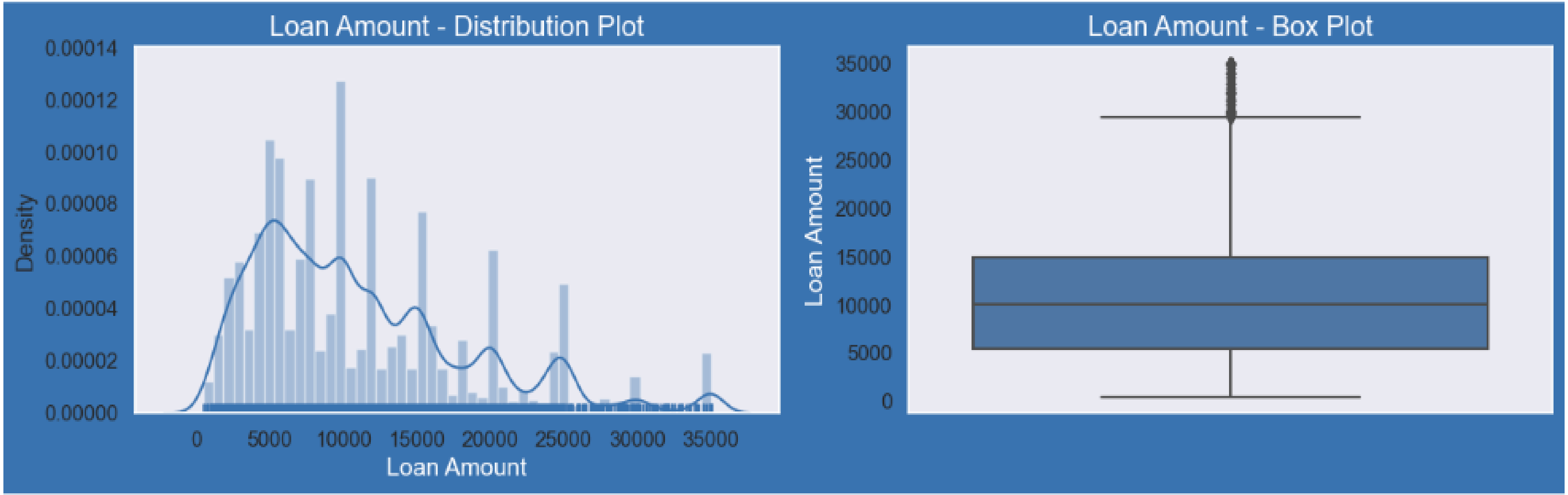
upGrad



Distribution of three loan amount fields using distribution plot.

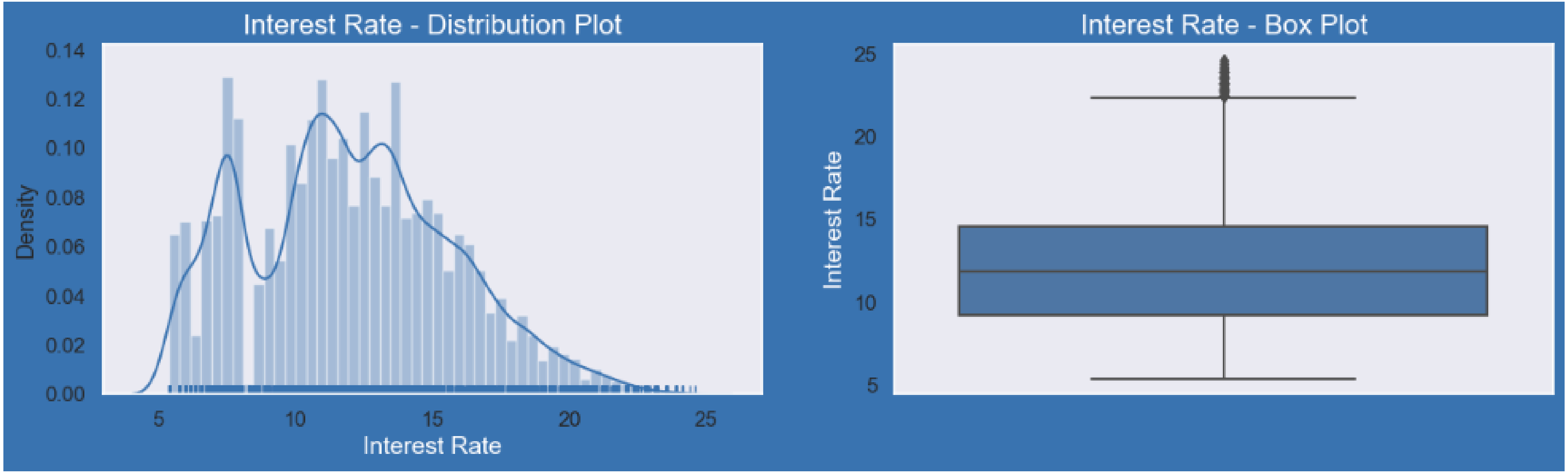
- ❑ Distribution of amounts for all three looks very much similar.
- ❑ We will work with only loan amount column for rest of our analysis.

Univariate Analysis



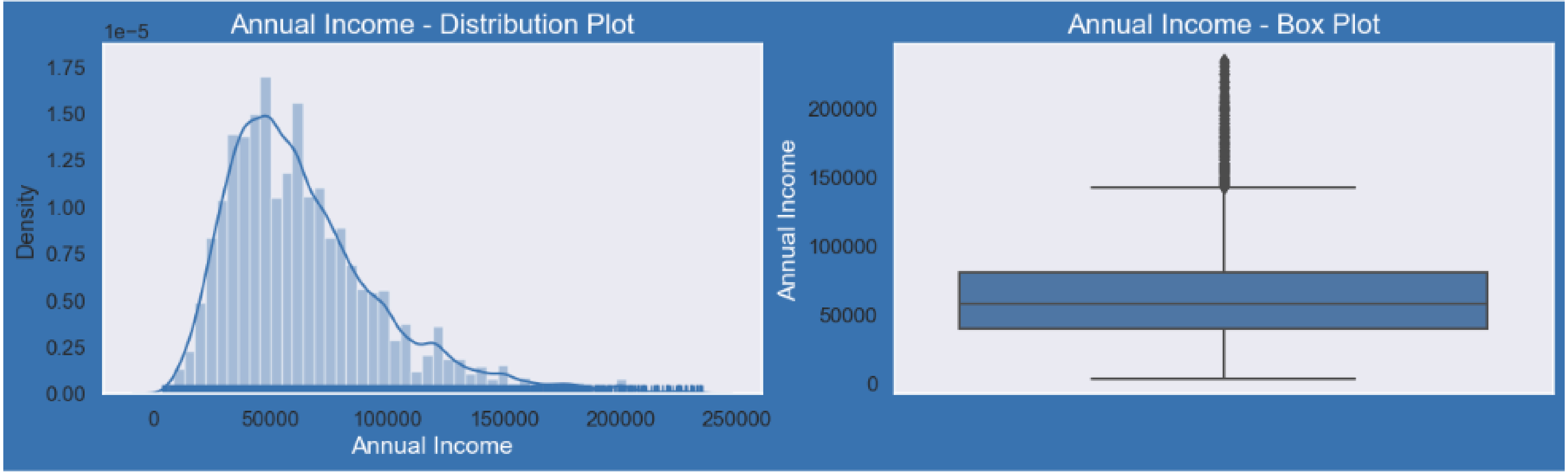
Plots show that most of the Loan amounts are in range of 5000 - 15000

Univariate Analysis



Plot show that most of the Interest Rates on loans are in range of 10% - 15%

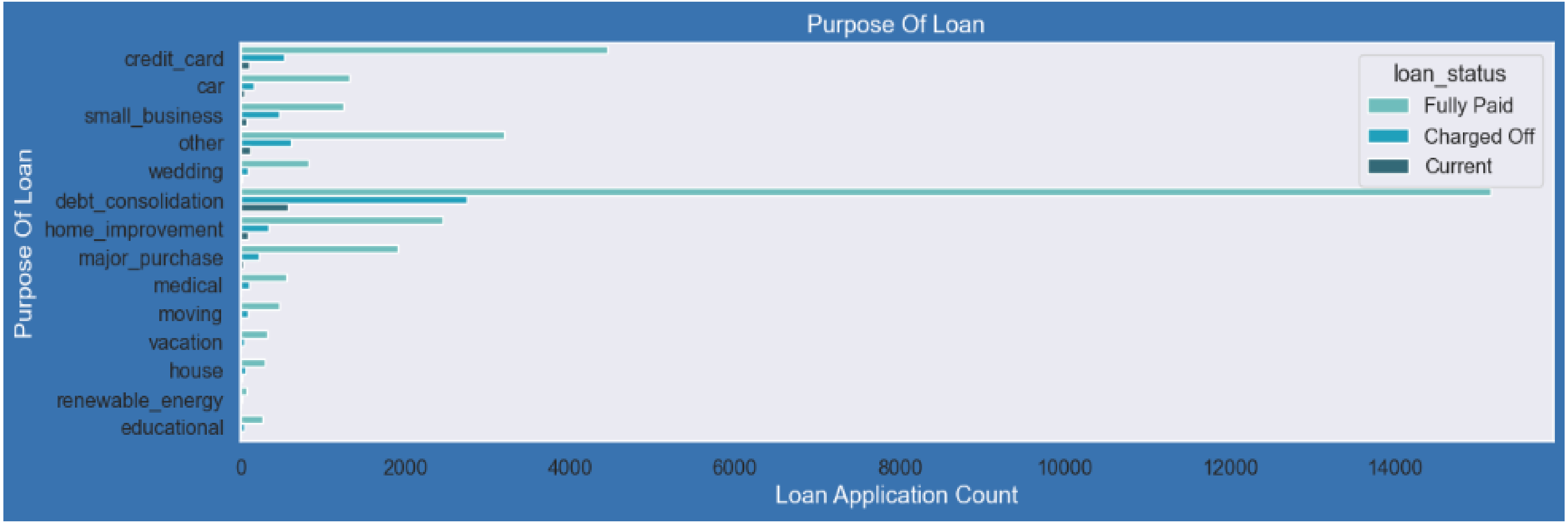
Univariate Analysis



Borrower's Annual incomes are in range of 40000- 80000

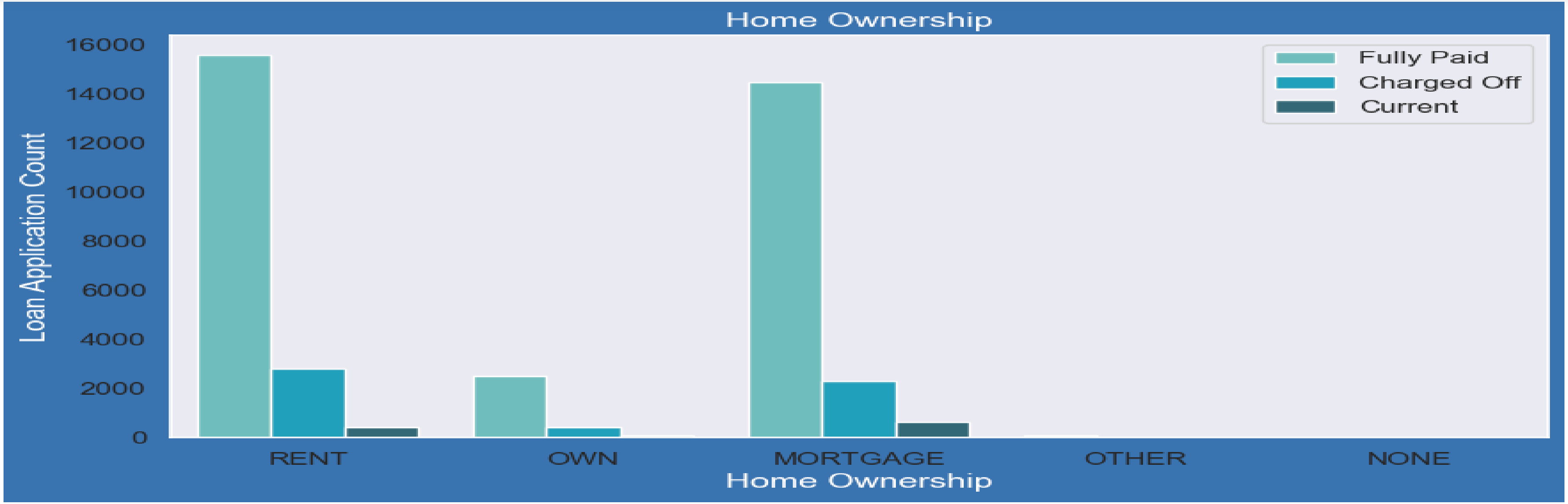
Univariate Analysis

upGrad



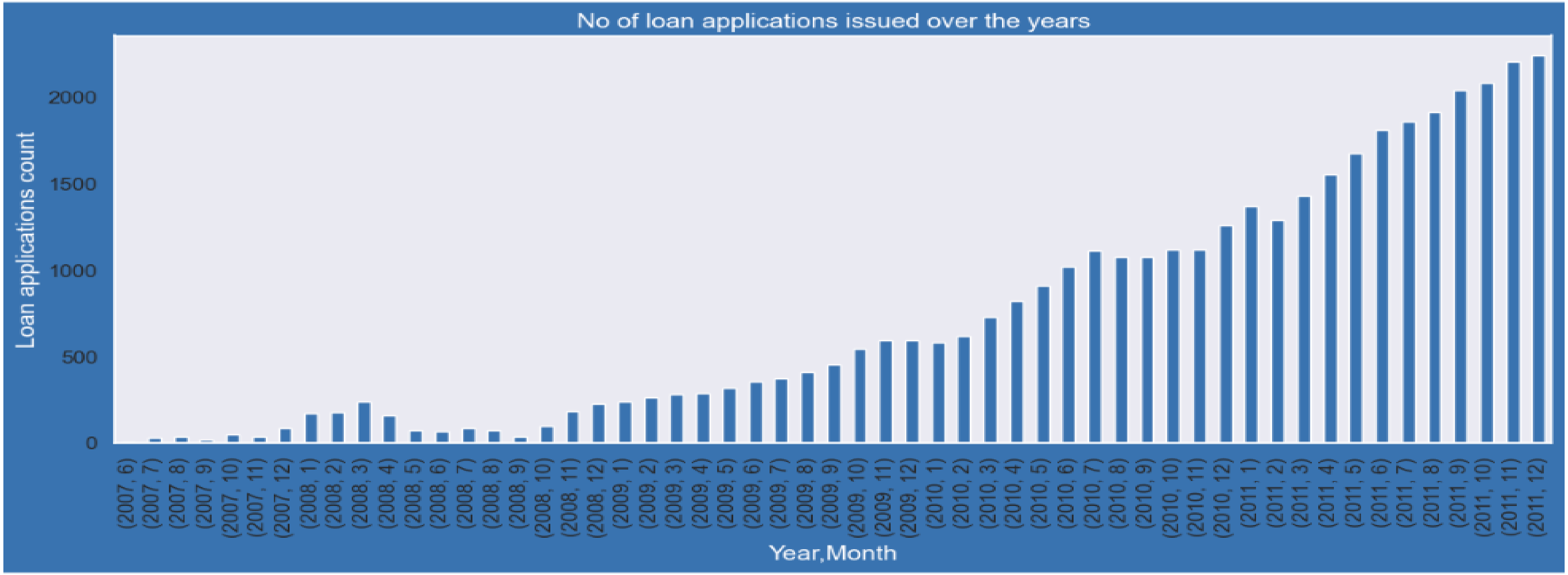
- ❑ Plot shows that most of the loans were taken for the purpose of debt consolidation & paying credit card bill.
- ❑ Number of charged off count also high too for these loans.

Univariate Analysis



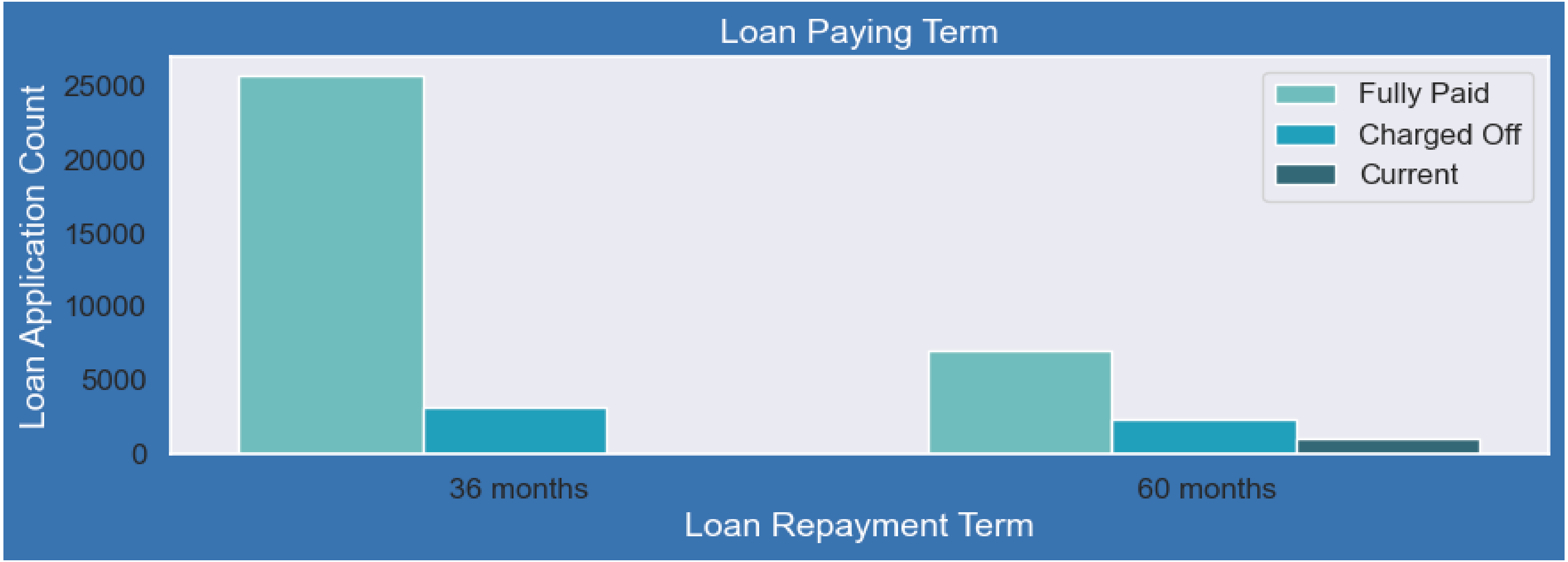
Loan application counts are high for rented home owner and default is also in similar

Univariate Analysis



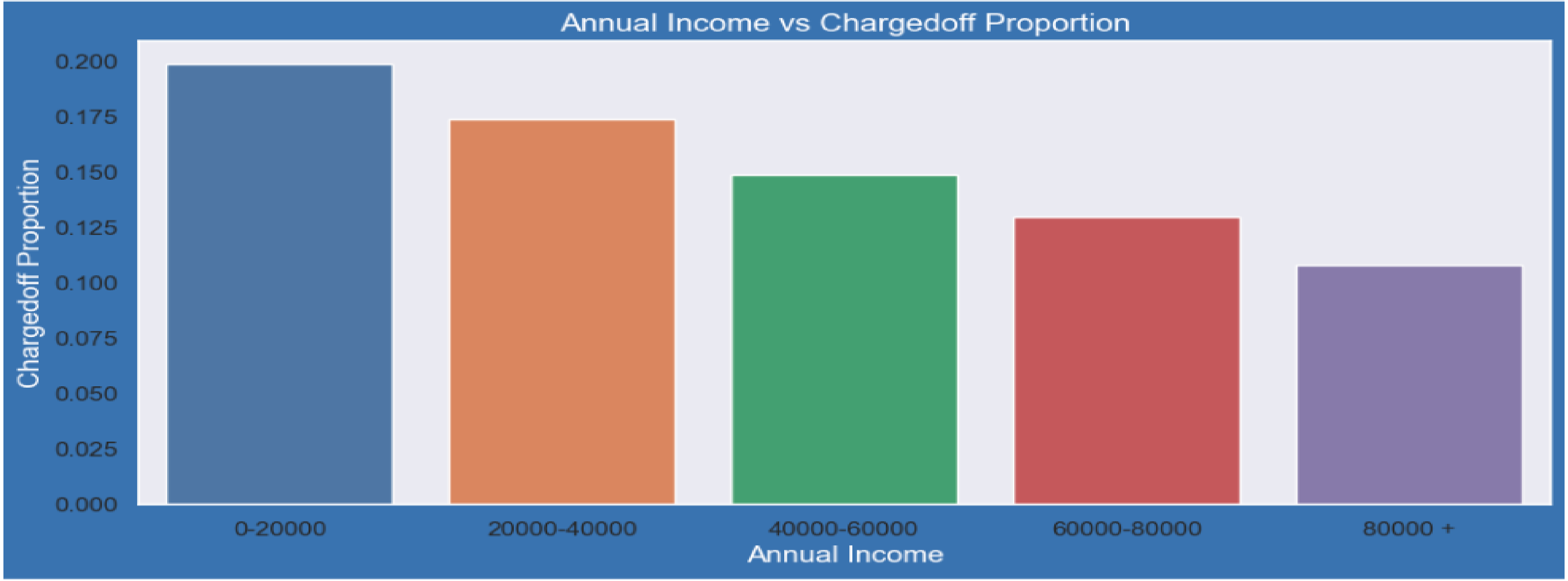
- ❑ **Loan application is increasing every passing year.**
- ❑ **Increase in number of loan applications are adding more to number of charged off applications.**
- ❑ **Number of loans issued in 2008(May-October) got dipped, may be due to Recession.**

Univariate Analysis



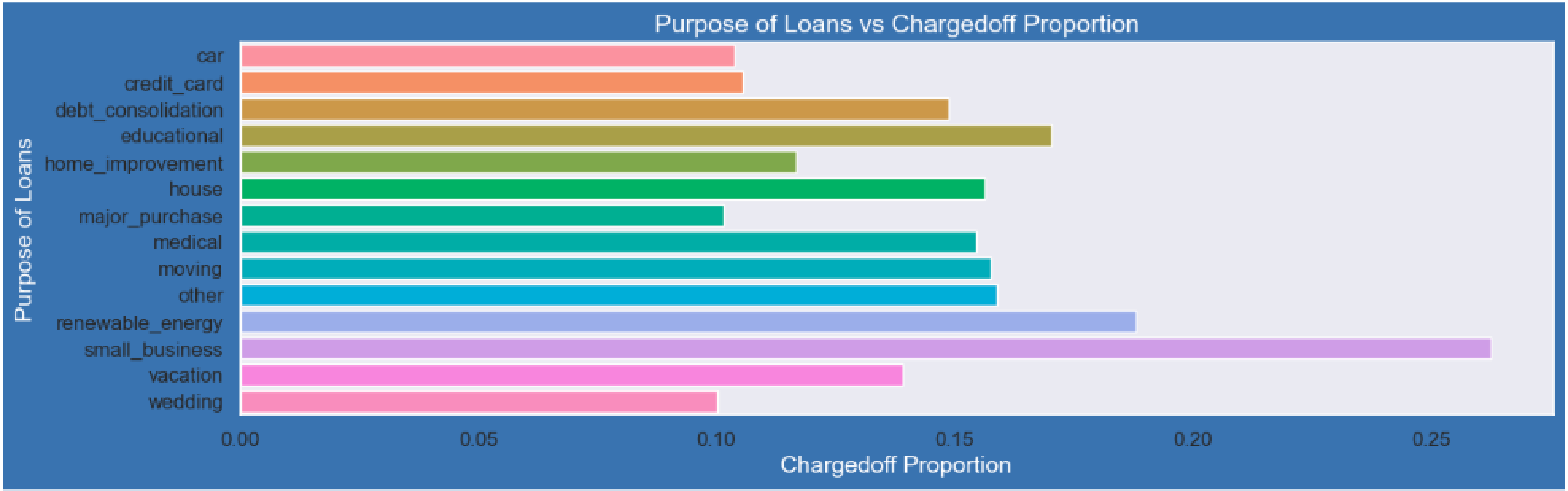
- ❑ **Loan with short tenure are less risky for getting defaulted as compared to loan for long period**
- ❑ **Ratio of fully paid to charge Off ratio is better for shorter period and hence less risky**

Bivariate Analysis



- ❑ 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.

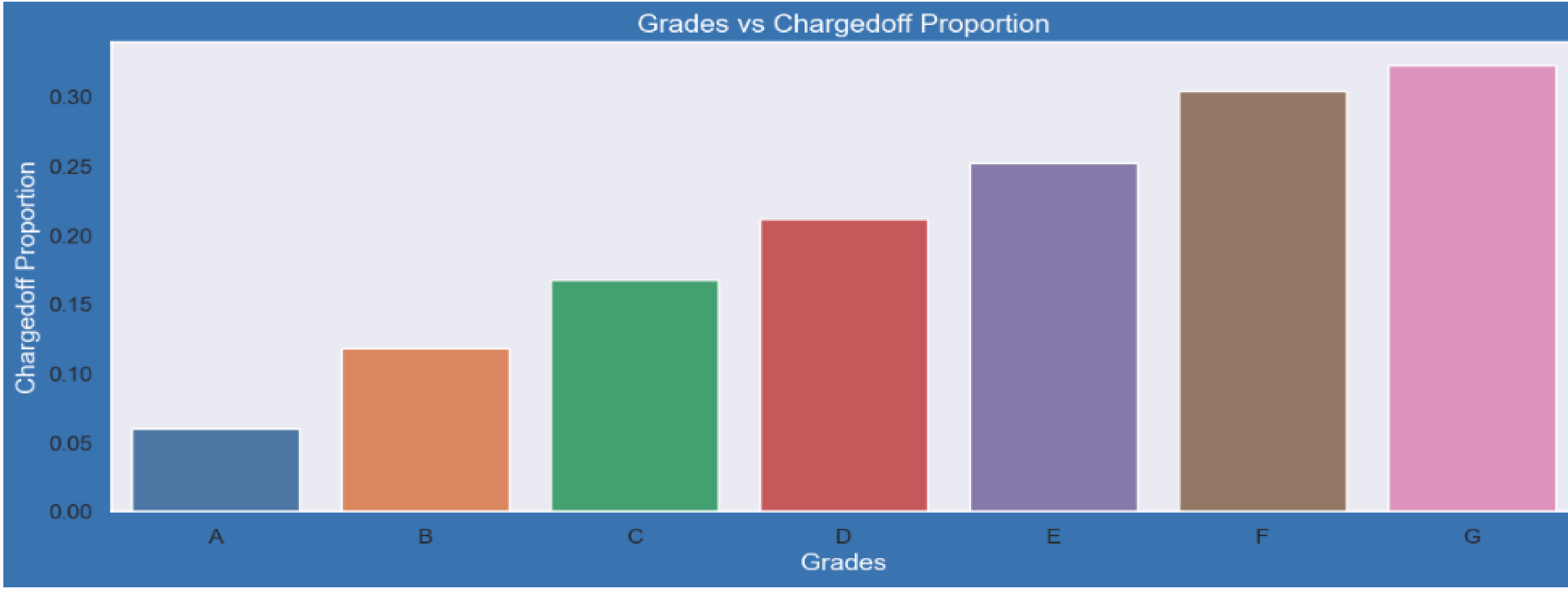
Bivariate Analysis



- ❑ Small Business applicants have high chances of getting charged off.
- ❑ Renewable_energy where charged off proportion is higher as compare to other categories.
- ❑ Wedding and Car loans are low charged off ratio

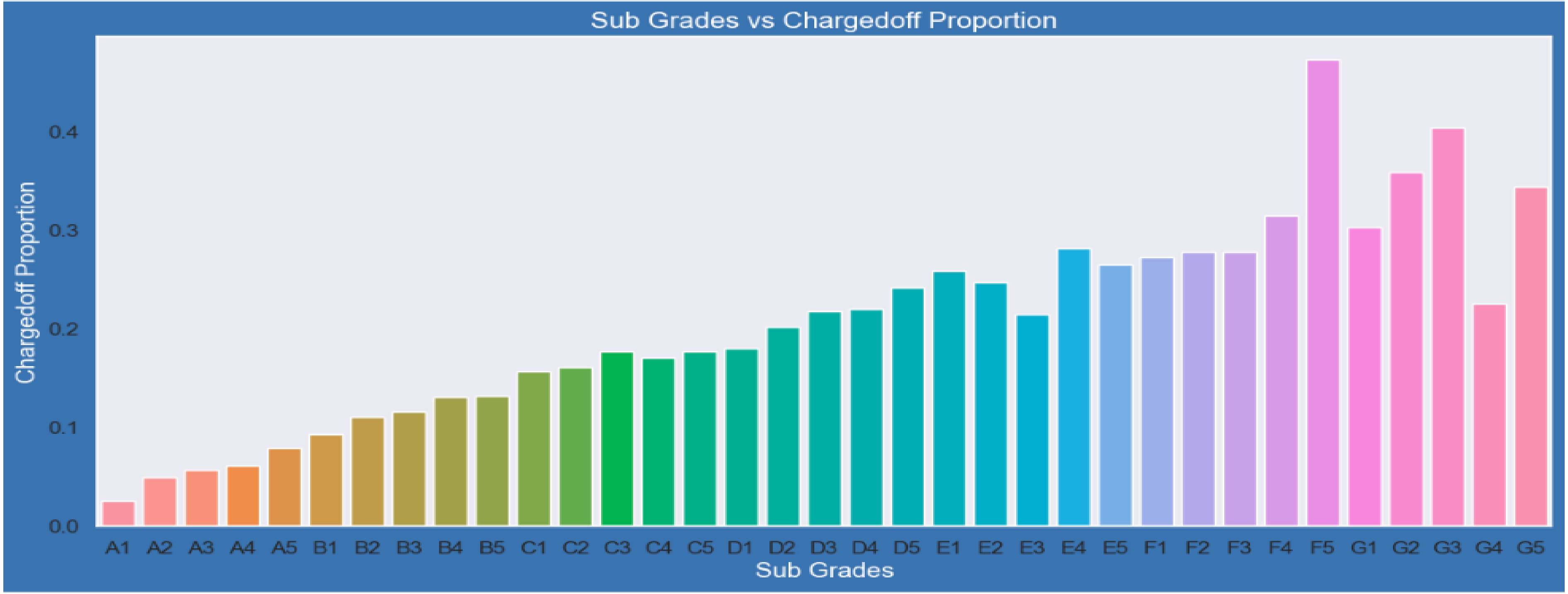
Bivariate Analysis

upGrad



- ❑ Grade "A" has very less chances of charged off.
- ❑ Grade "F" and "G" have very high chances of charged off.
- ❑ Chances of charged of is increasing with grade moving from "A" towards "G"

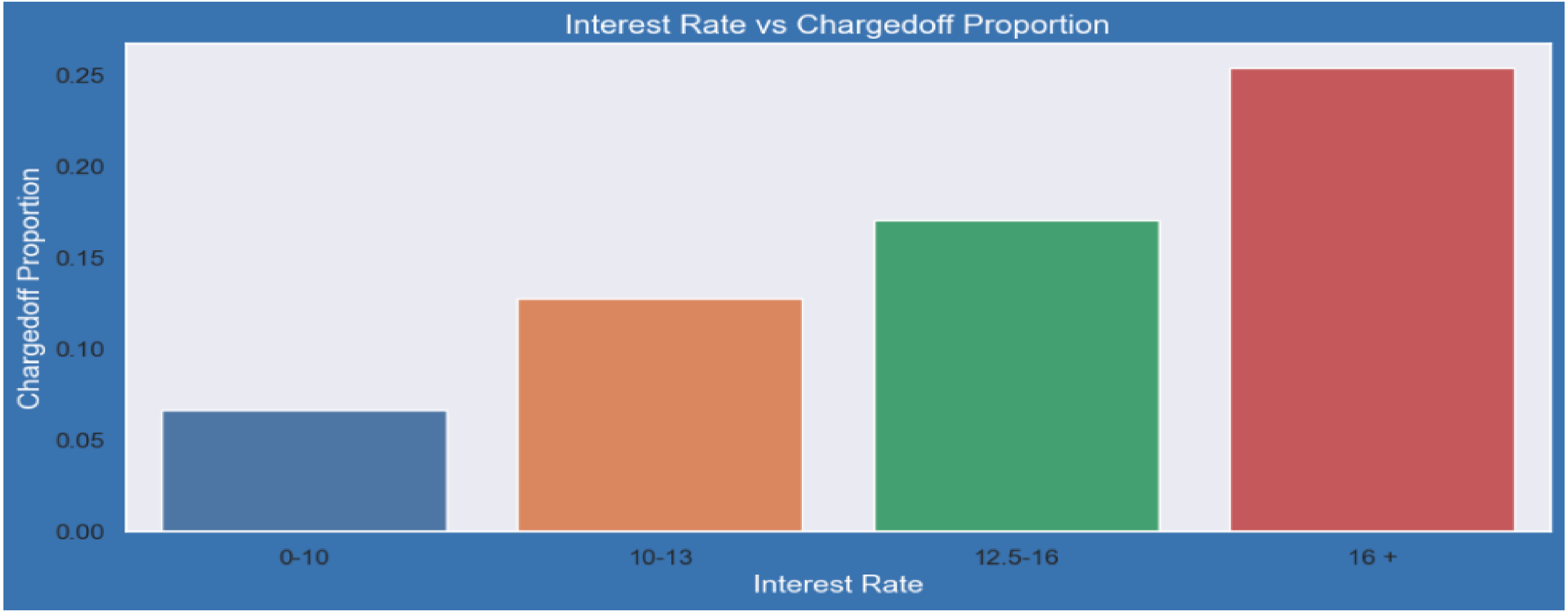
Bivariate Analysis



- ❑ sub Grades of "A" has very less chances of charged off.
- ❑ sub Grades of "F" and "G" have very high chances of charged off.
- ❑ proportion of charged off is increasing with sub grades moving from sub grades of "A" towards sub grades of "G"

Bivariate Analysis

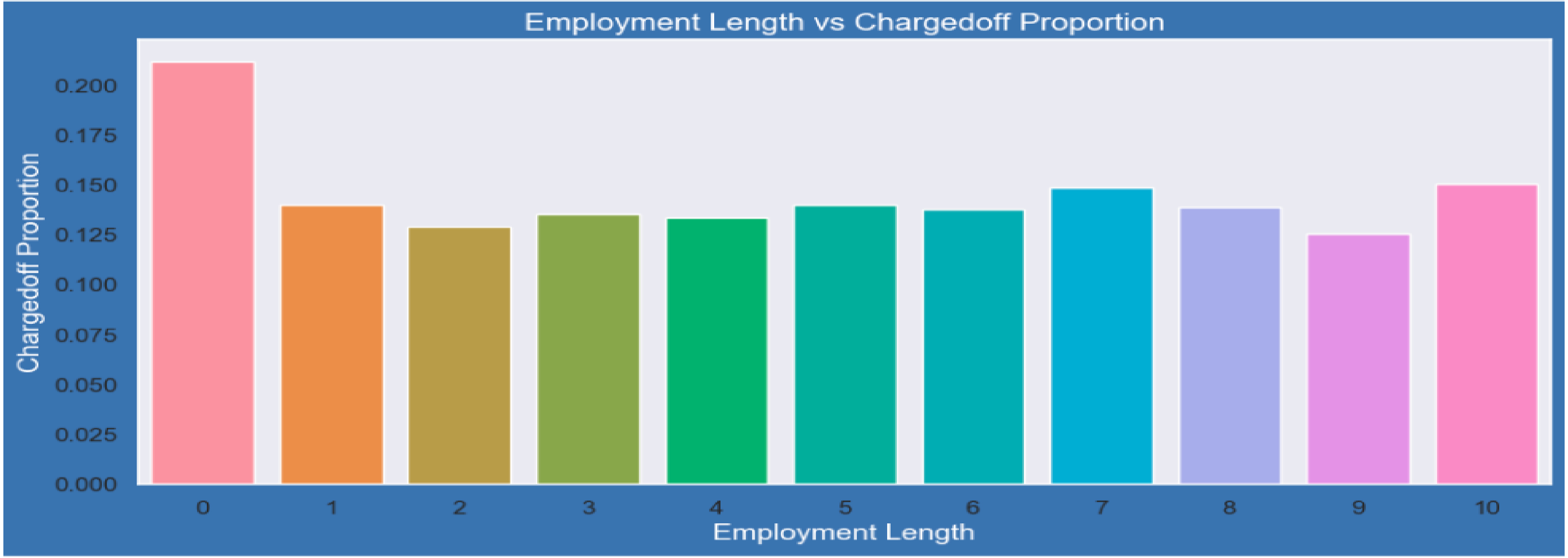
upGrad



- ❑ 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 intrest rates.
- ❑ Charged off proportion is increasing with higher interest rates.

Bivariate Analysis

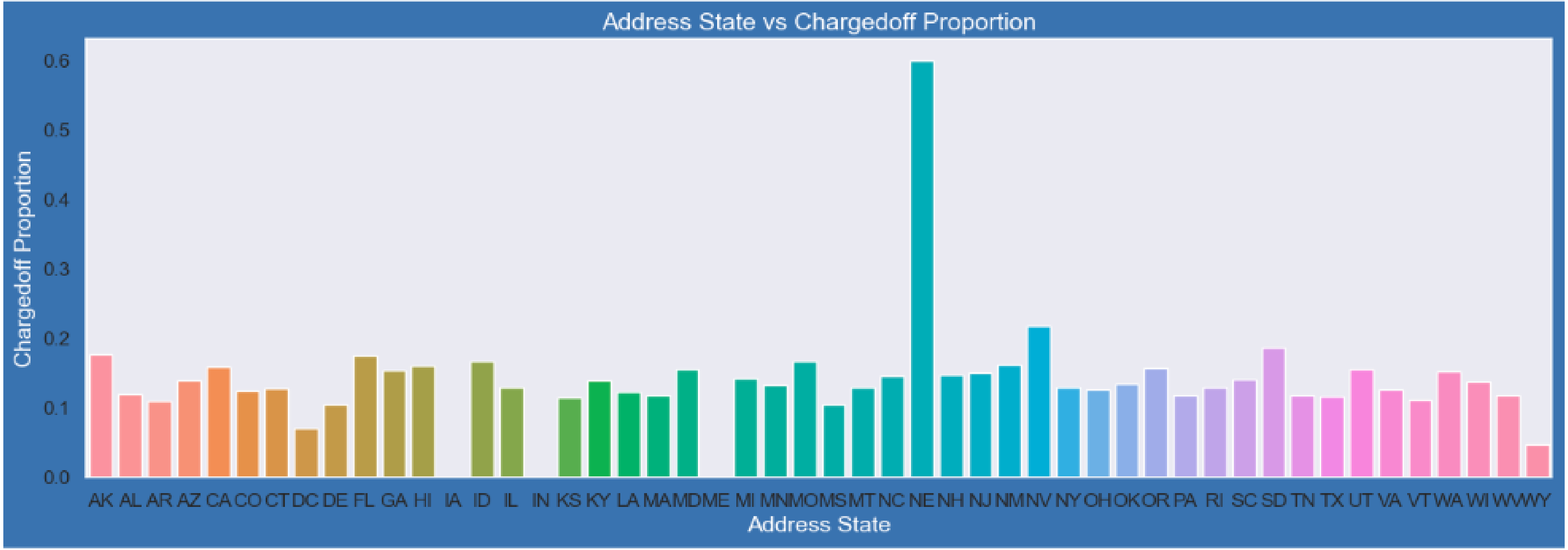
upGrad



- ❑ Those who are not working or have less than 1 year of work experience have high chances of getting charged off.
- ❑ It makes sense as with less or no experience they don't have source of income to repay loan.
- ❑ Rest of the applicants have more or less same chnaces of getting charged off.

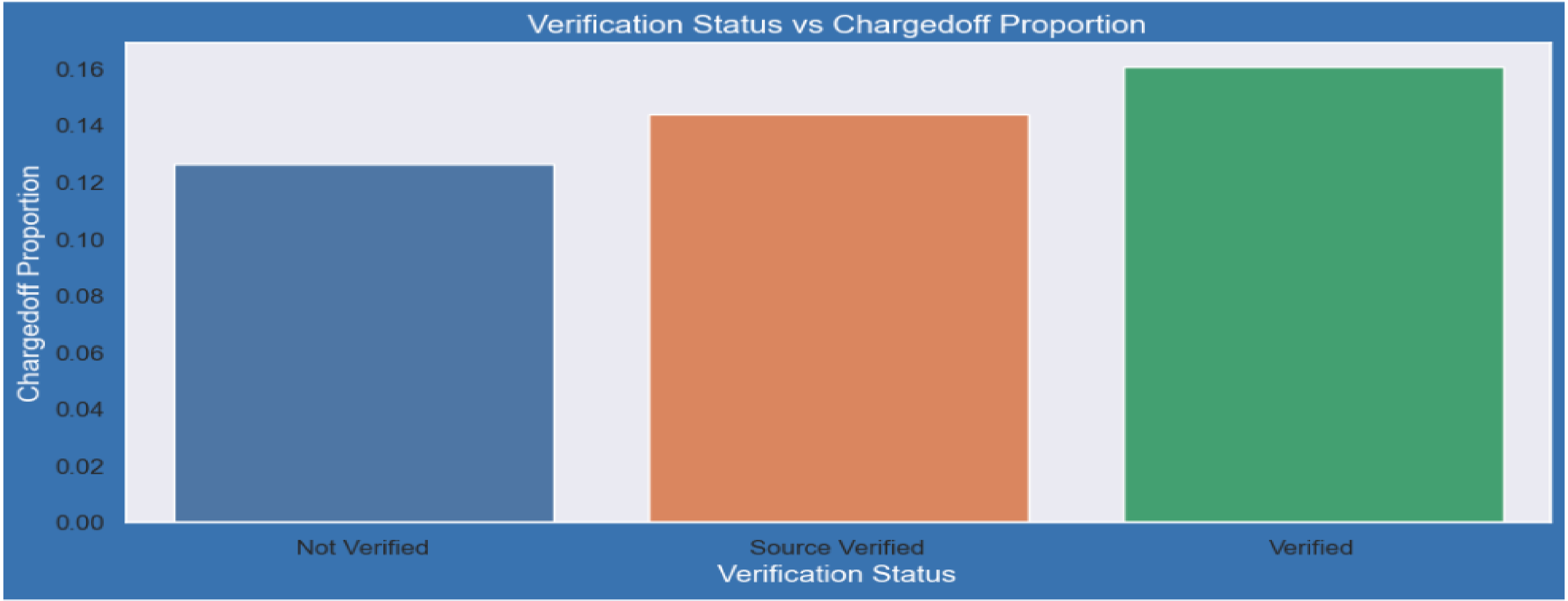
Bivariate Analysis

upGrad



- ❑ States NE has very high chances of charged off but number of applications are too low to make any decisions.
- ❑ NV,CA and FL states shows good number of charged offs in good number of applications.

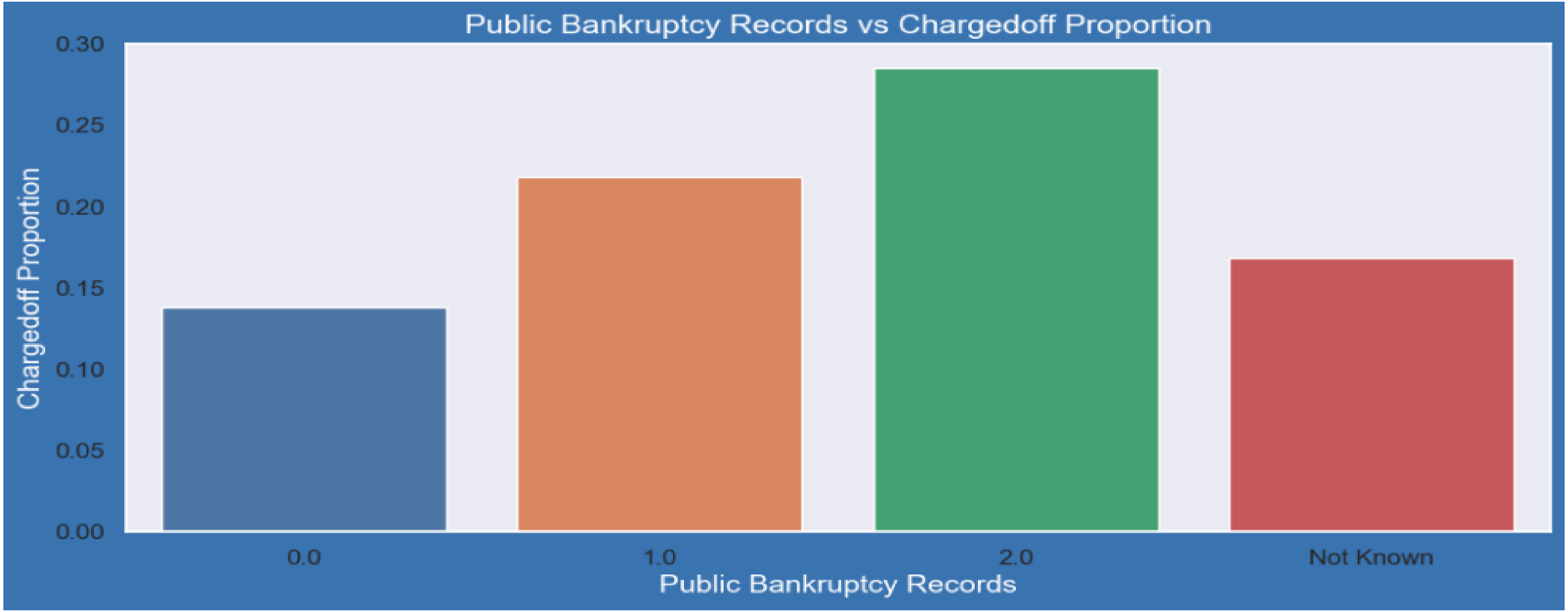
Bivariate Analysis



- ❑ **There is not much difference in charged off proportion.**
- ❑ **This variable doesn't provide any insights for charged off.**

Bivariate Analysis

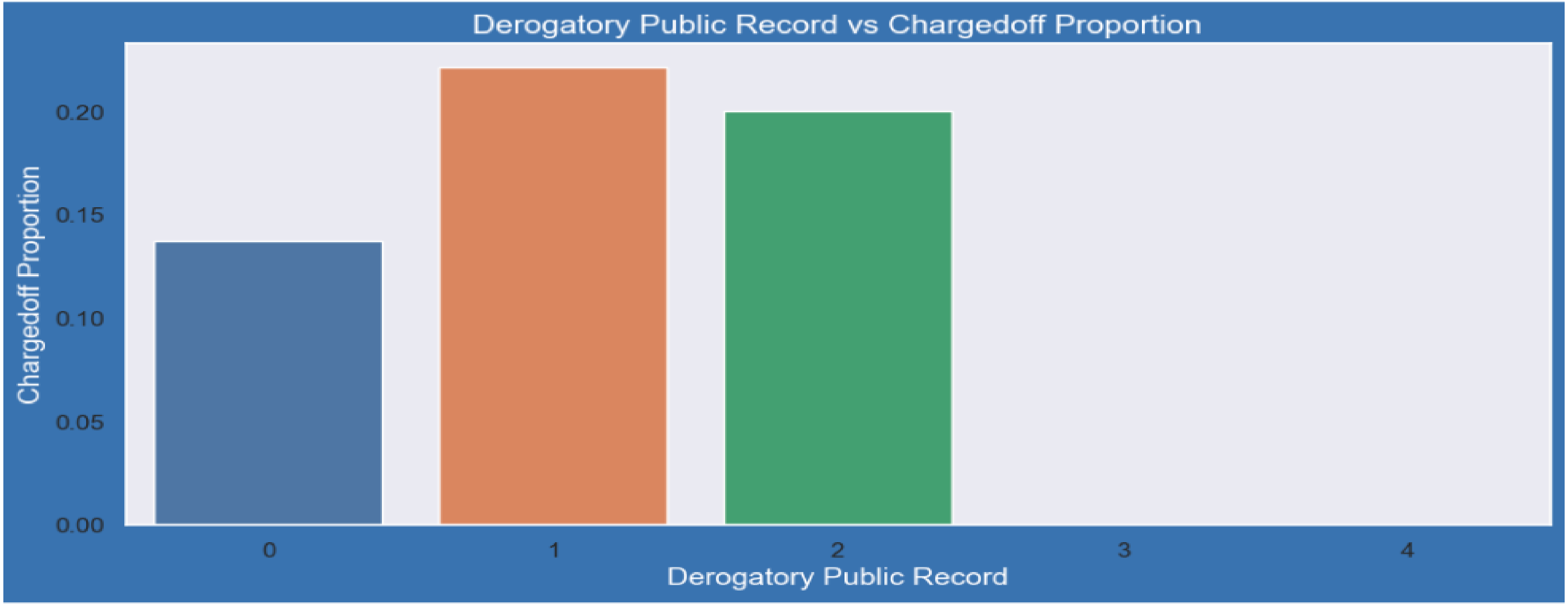
upGrad



- ❑ 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.
- ❑ Not known is the column for which we don't have any information abount borrower.
- ❑ This also makes sense that who has defaulted before has more chances of dafaulting in future as well.

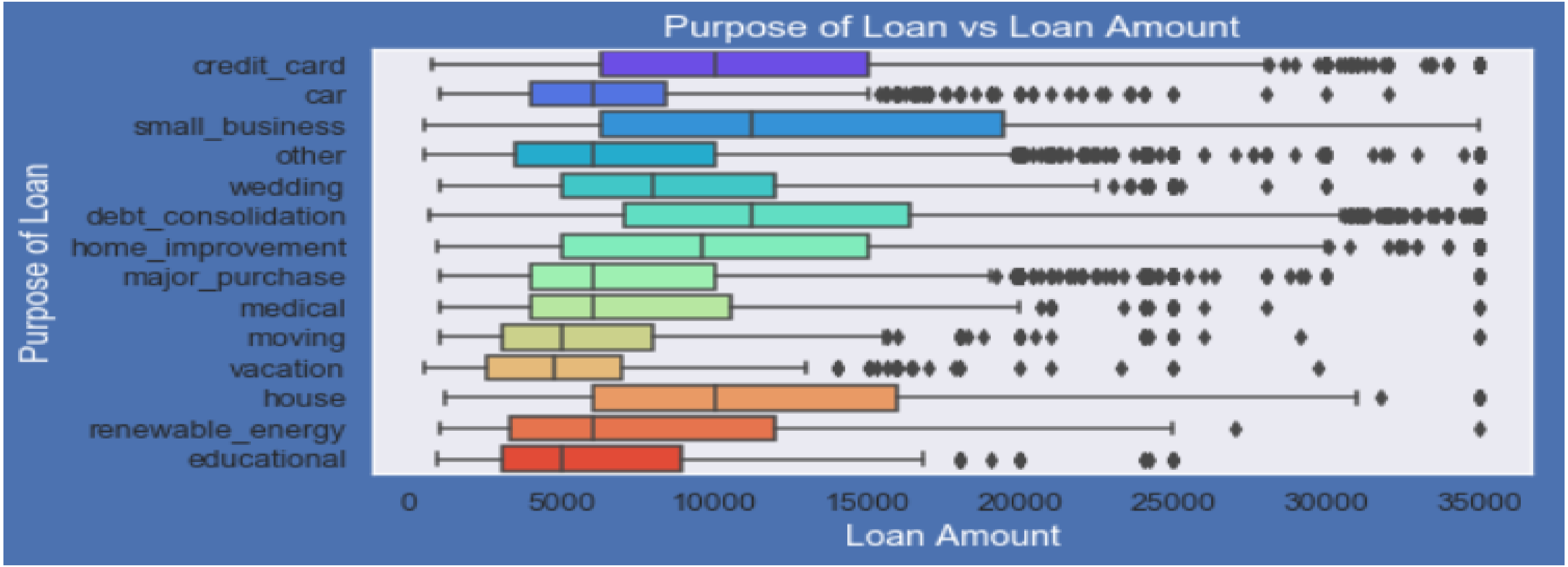
Bivariate Analysis

upGrad



- ❑ A derogatory item is an entry that may be considered negative by lenders because it indicates risk
- ❑ Ability to qualify for credit or other services. Public records and collections are derogatory items
- ❑ They reflect financial obligations that were not paid as agreed.
- ❑ Those have pub_rec value 1 or 2 have charged off chances higher than who have no Derogatory Public Record.
- ❑ pub_rec count 3-4 has less numbers so cannot reach on any conclusions.

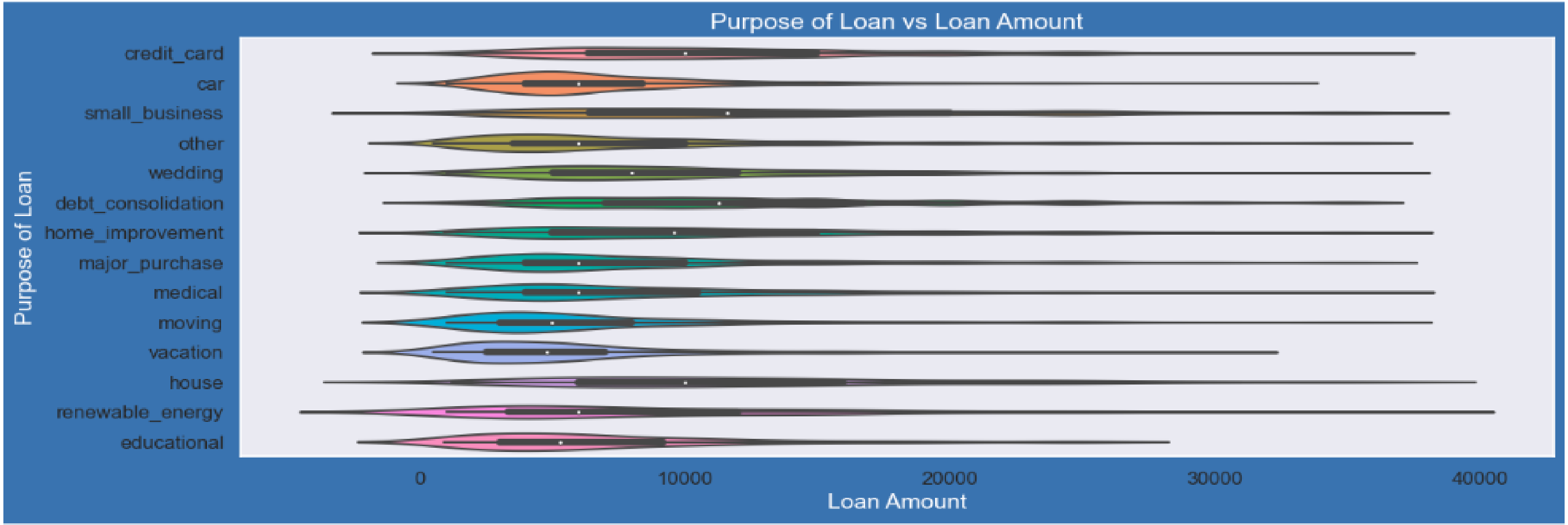
Bivariate Analysis



Purpose of Loan v/s Amount applied for Loan

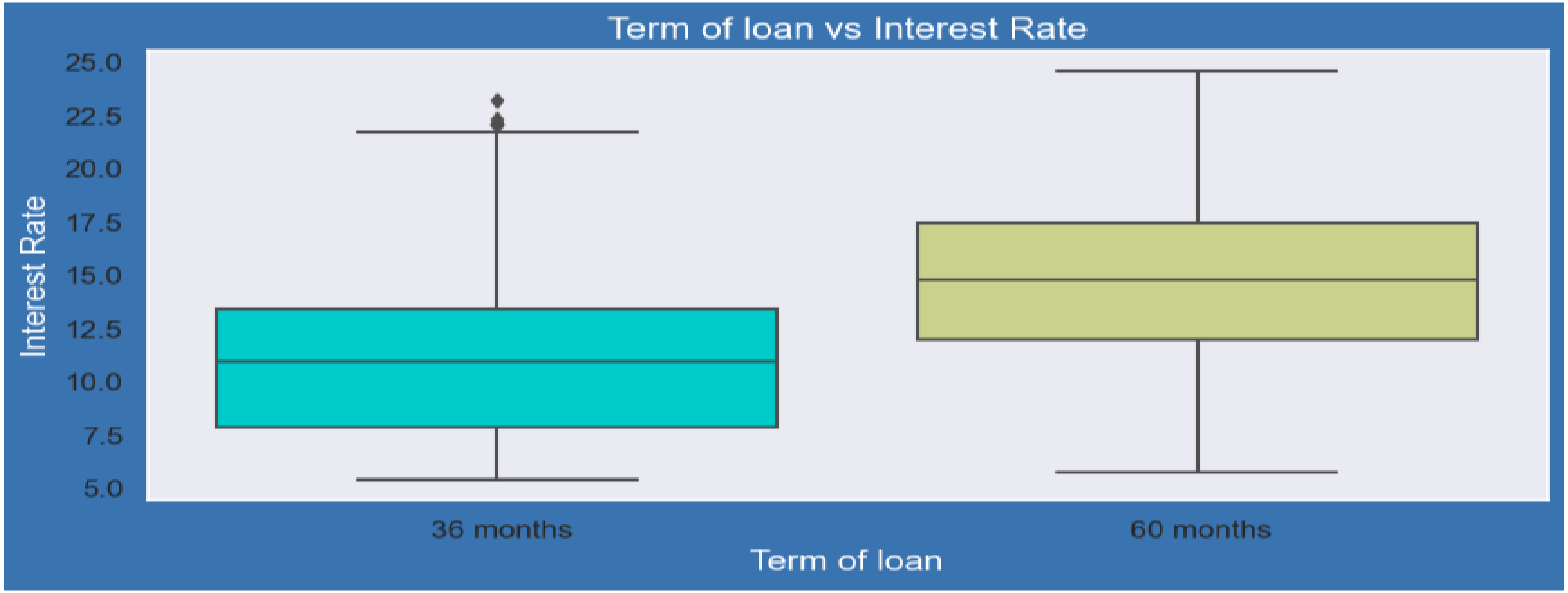
- ❑ Most of the loan amounts are big for small business purpose among all purposes.
- ❑ While the Debt consolidation is second and Credit card is third.

Bivariate Analysis



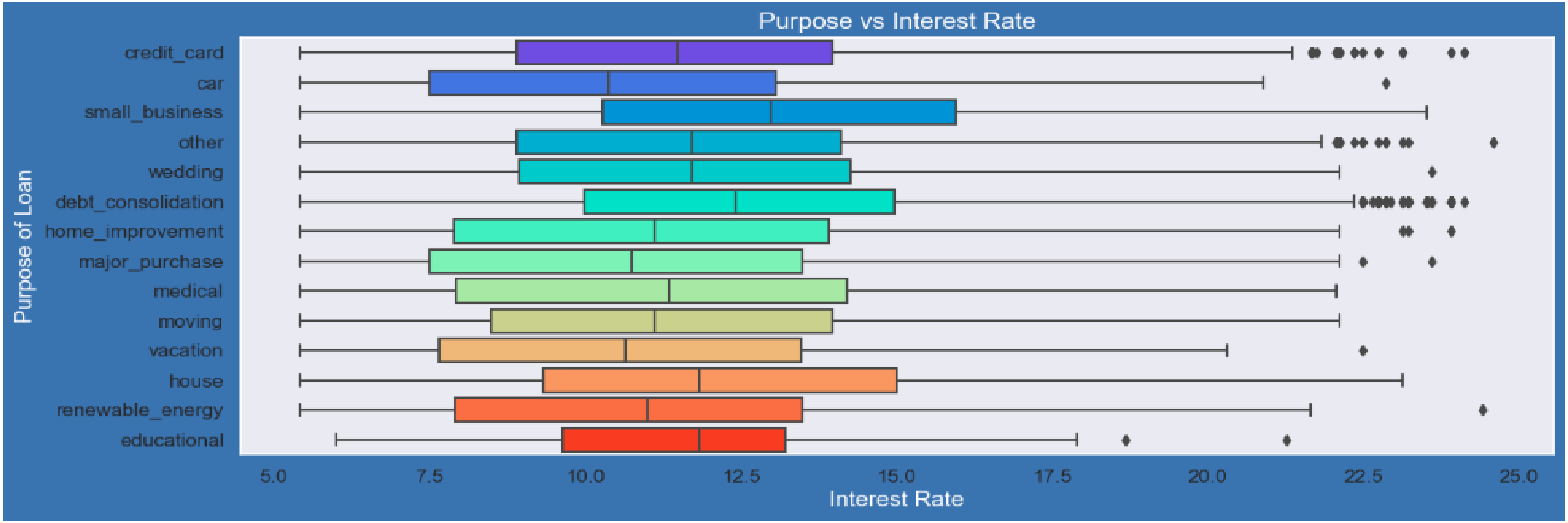
❑ **Loan taken for small business purpose, Debt consolidation and Credit card are somewhat evenly distributed as compare to loan taken for other purposes.**

Bivariate Analysis



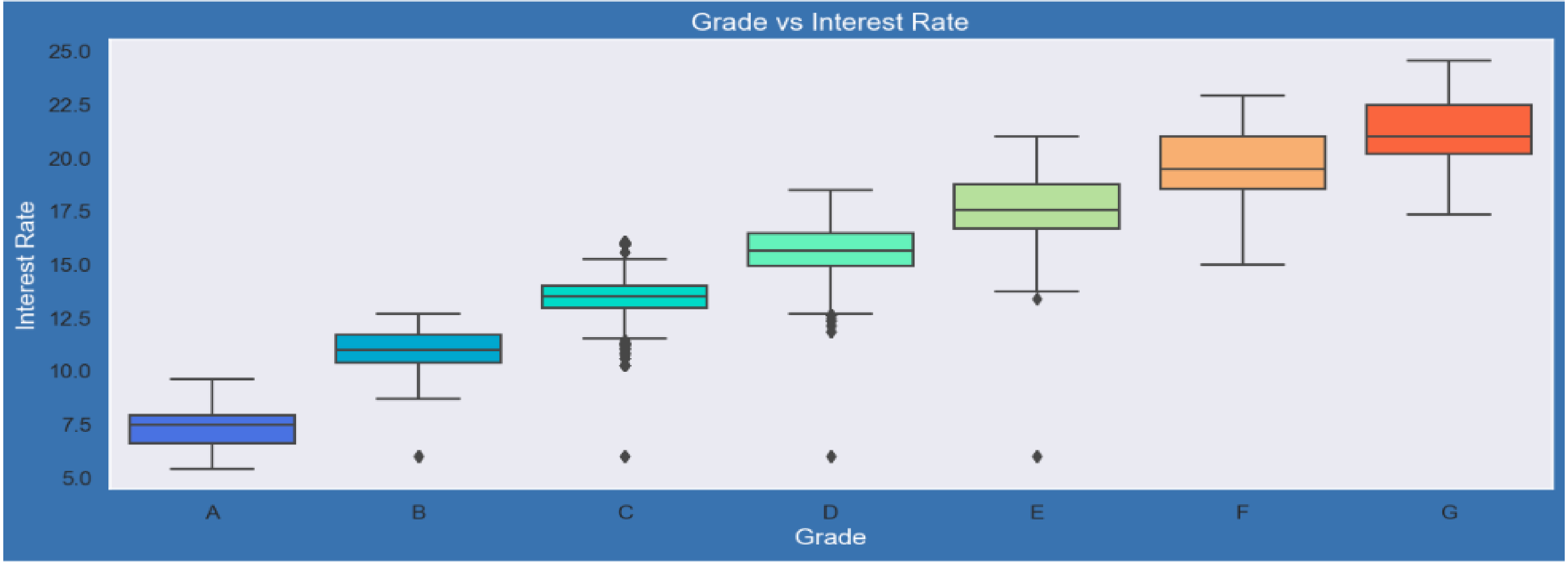
- ❑ **It is clear that avearge intrest rate is higher for 60 months loan term.**
- ❑ **Most of the loans issued for longer term had higher interest rates for re-payment.**

Bivariate Analysis



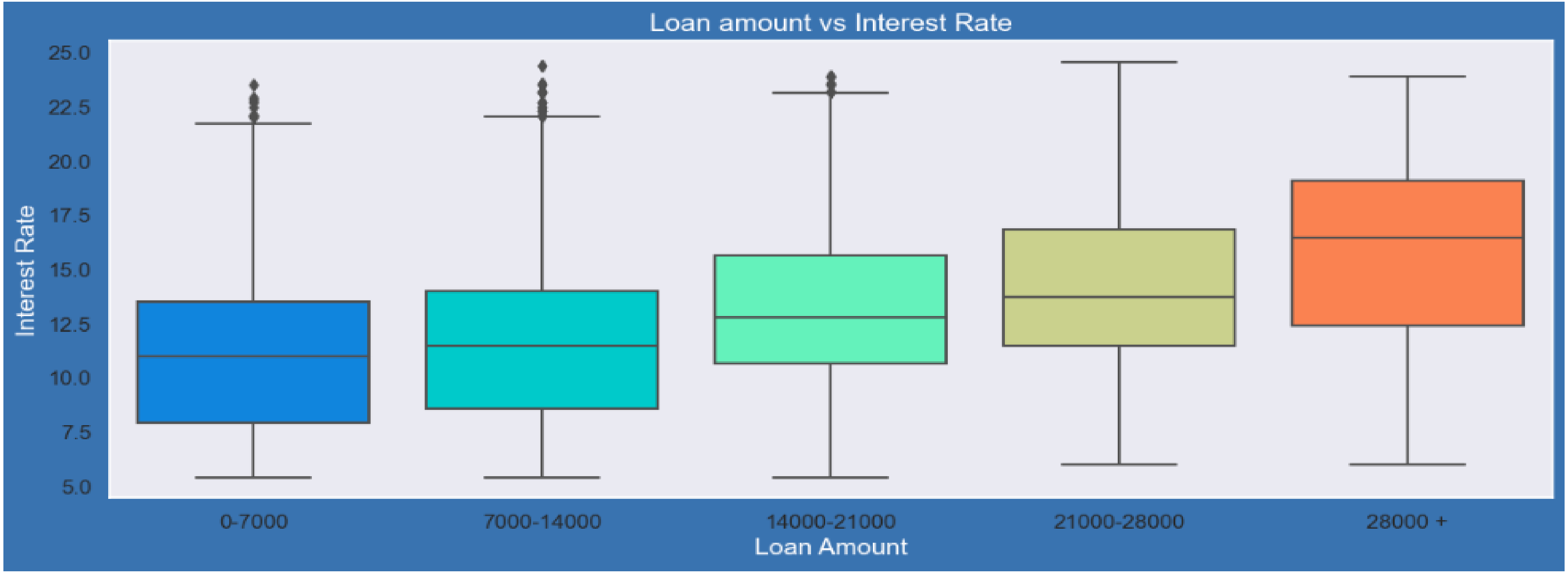
- ❑ **It is clear that average interest rate is highest for small business purpose.**
- ❑ **Loans taken for small business purposes had to repay the loan with more interest rate as compared to other.**
- ❑ **Debt consolidation is 2nd where borrowers had to pay more interest rate.**

Bivariate Analysis



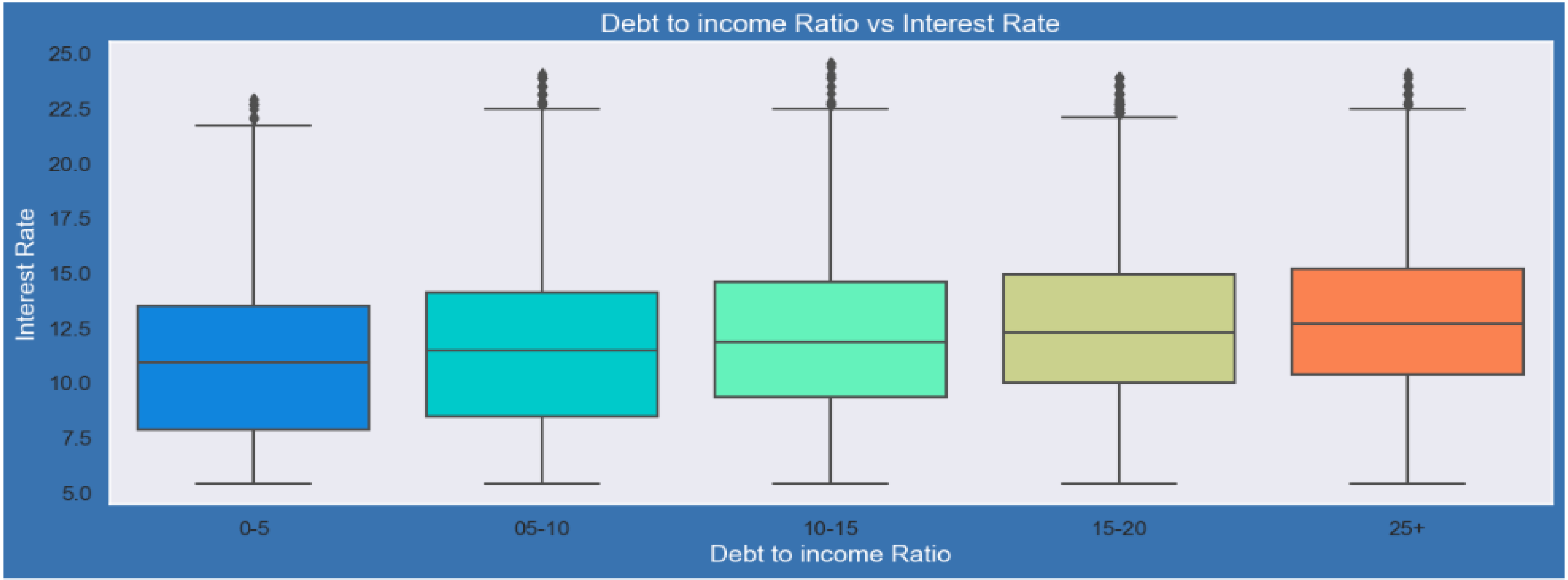
- ❑ A-grade is a top letter grade for a lender to assign to a borrower.
- ❑ The higher the borrower's credit grade, The lower the interest rate offered to that borrower on a loan.
- ❑ It is clear that interest rate is increasing with grades moving from A to F.

Bivariate Analysis



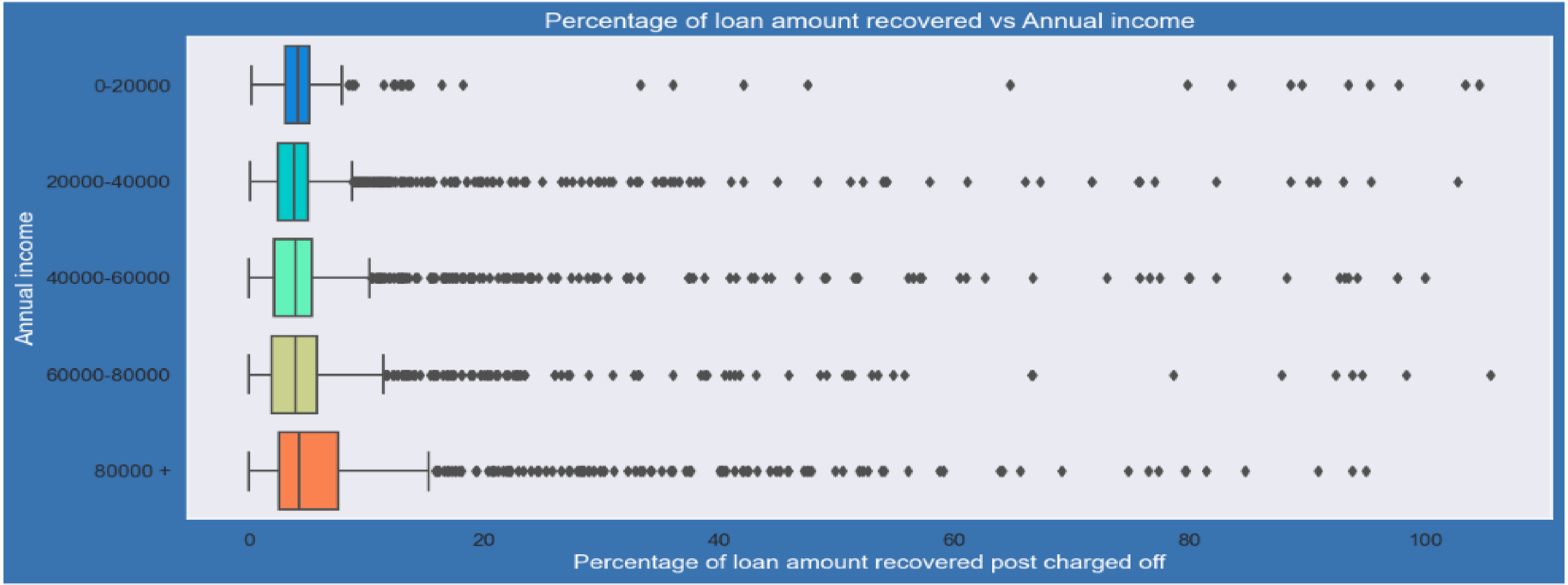
- ❑ It is clear that interest rate is increasing with loan amount increase.
- ❑ When loan amount is more it is taken for longer loan term, we saw earlier that longer the loan term more the interest rate.

Bivariate Analysis



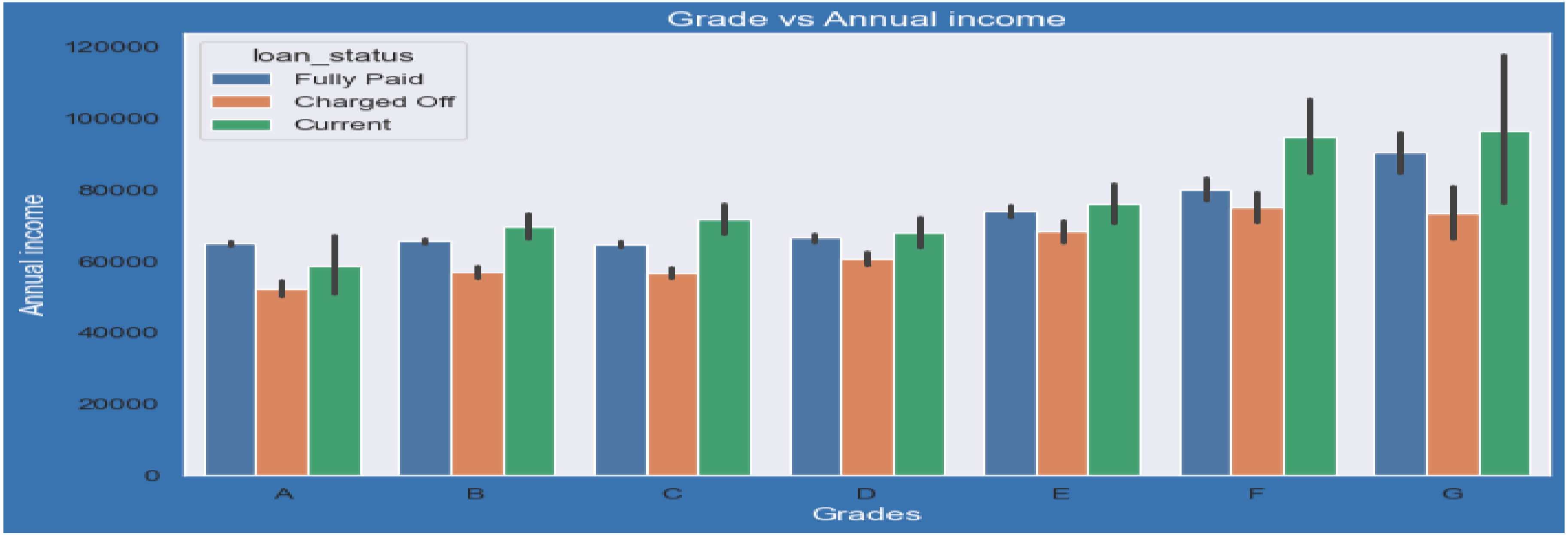
- ❑ **If your DTI is low enough you may get a lower interest rate.**
- ❑ **Plot shows no significant variation but there is slight increase in interest rate with increase in DTI.**

Bivariate Analysis



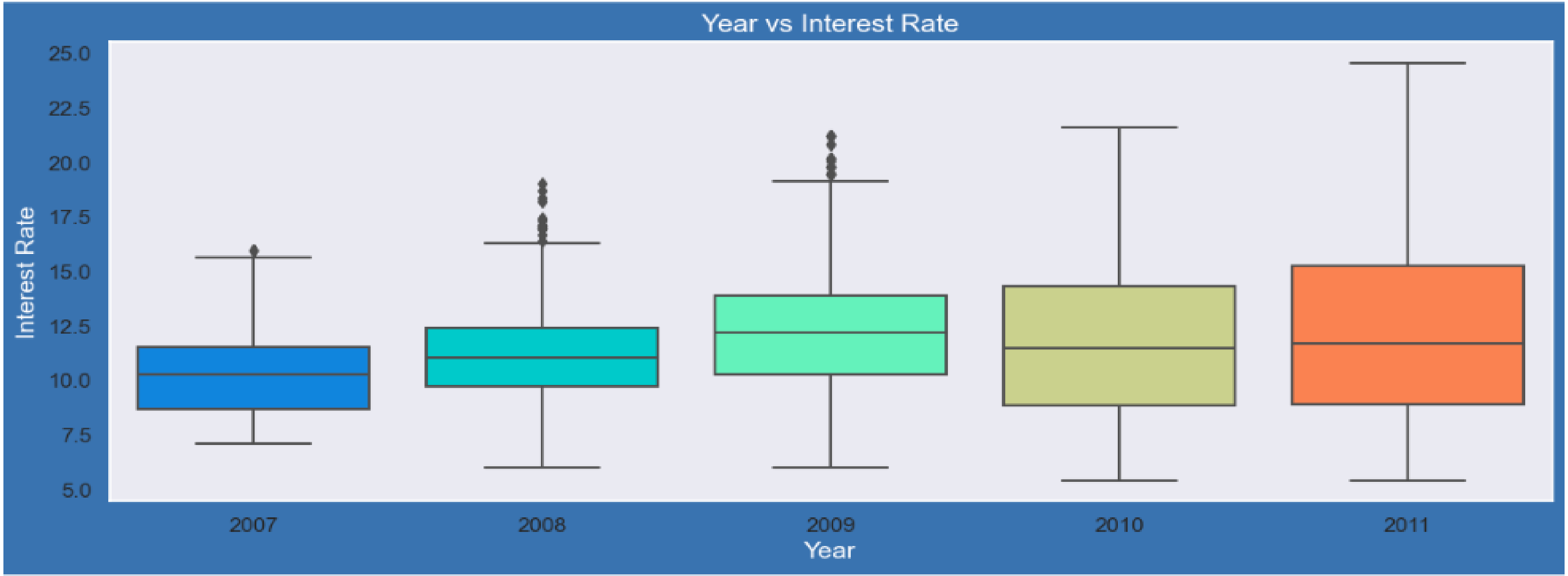
- ❑ **Higher percentage of loan amount is recovered when annual income is high.**
- ❑ **Plot shows no significant variation but there is slight increase in recovery percentage with increase in annual income.**

Bivariate Analysis



- ❑ **From this we can conclude that the ones getting 'charged off' have lower annual incomes than the ones who'paid fully' for each and every grade (i.e. at same interest range)**

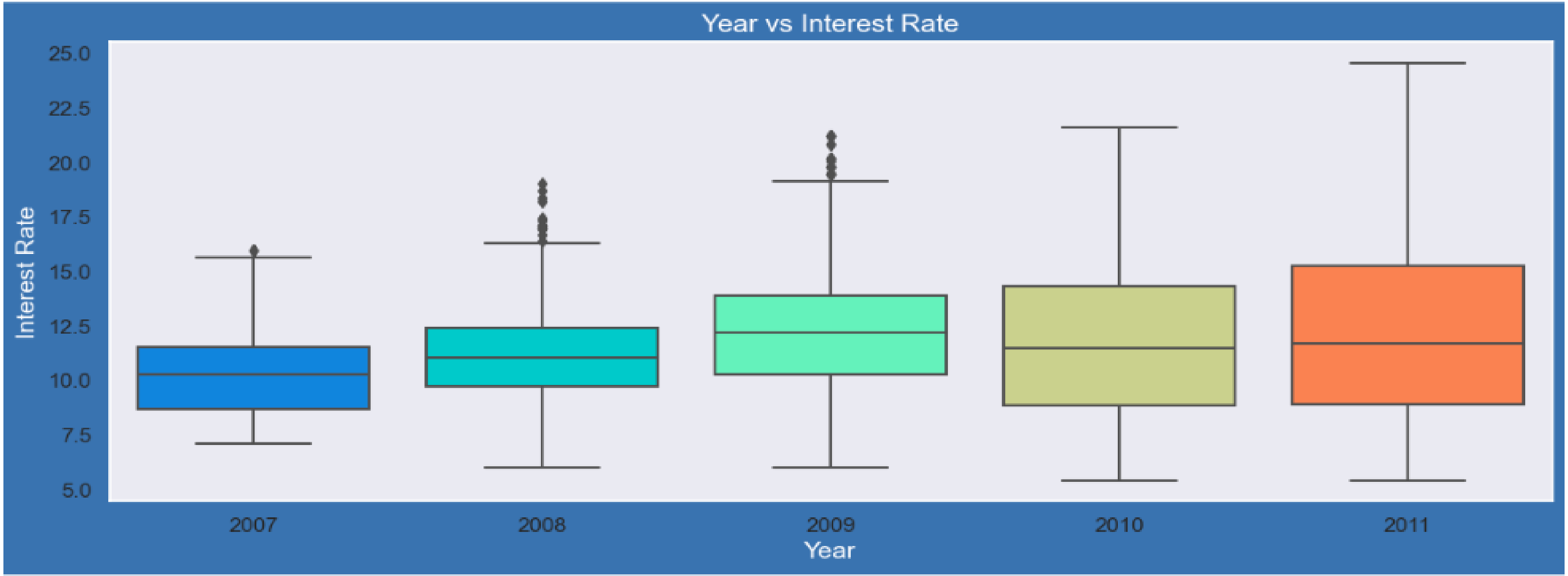
Bivariate Analysis



- ❑ **Plot shows interest rate is increasing slowly with increase in year.**

Bivariate Analysis

upGrad

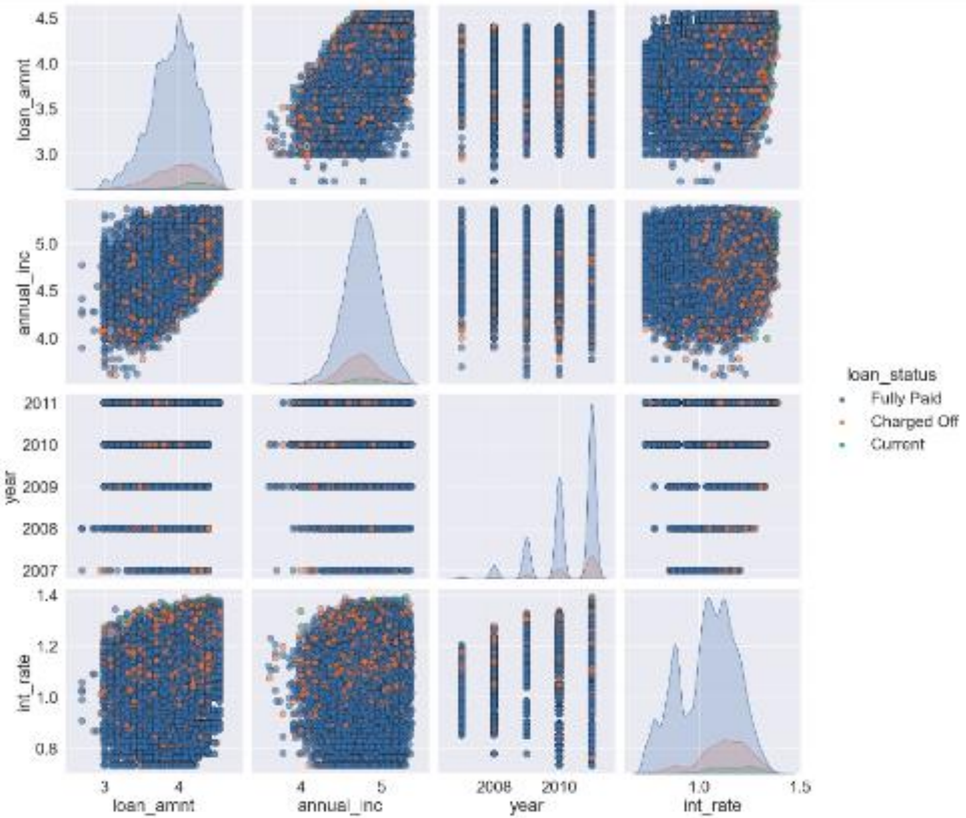


❑ **Plot shows intrest rate is increasing slowly with increase in year.**

Bivariate Analysis - Correlation Matrix-Quantitative Variables(Refer worksheet for diagram)

- ❑ **Observation is that Loan amount, investor amount, funding amount are strongly correlated.**
- ❑ **Annual income with DTI(Debt-to-income ratio) is negatively correlated. That means when annual income is low DTI is high & vice versa.**
- ❑ **Debt income ratio is the percentage of a consumer's monthly gross income that goes toward paying debts.**
- ❑ **Positive correlation between annual income and employment years.**

Multivariate Analysis: Pair plot

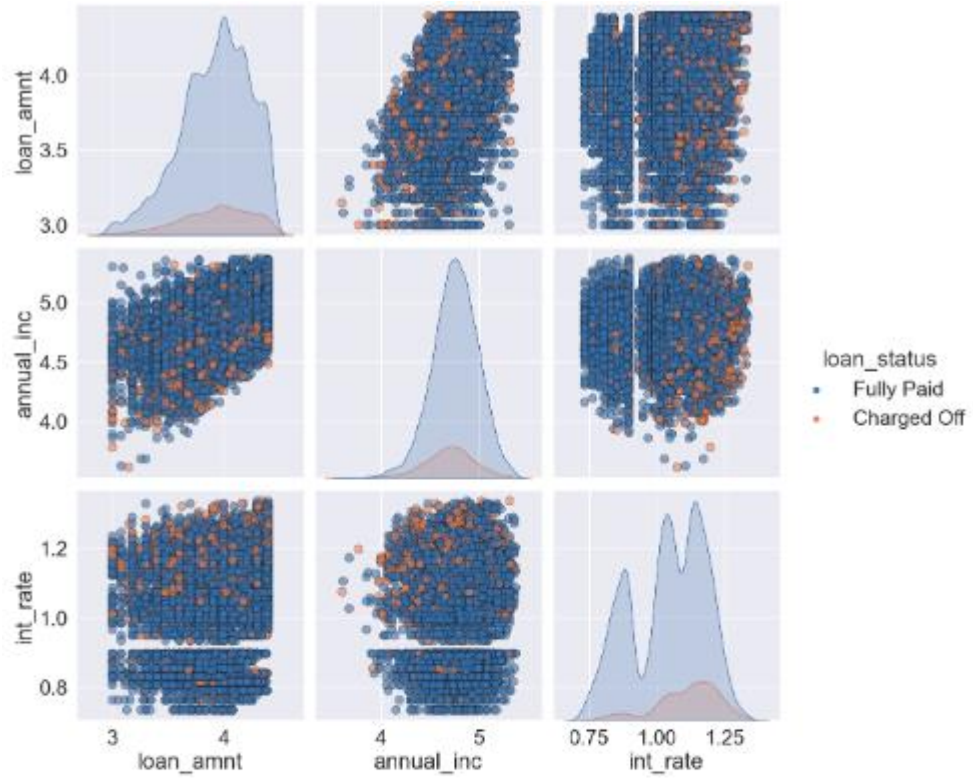


Observing Loan Amount, Annual Income, Year when loan was taken, Interest Rate at which loan was taken to each other :

- ❑ Higher the interest rate , Higher charged off ratio
- ❑ Higher the annual income, Higher the loan amount slightly.
- ❑ Increase in number of charged off with increase in year.
- ❑ Interest rate is increasing with loan amount increase

Multivariate Analysis: Pair plot

upGrad



Observing Loan Amount, Annual Income, Year when loan was taken, Interest Rate at which loan was taken to each other :

- ❑ Higher the interest rate , Higher charged off ratio
- ❑ Higher the annual income, Higher the loan amount slightly.
- ❑ Increase in number of charged off with increase in year.
- ❑ Interest rate is increasing with loan amount increase

Summary

Based on the Analysis done on the Variables, we conclude the below mentioned points

- ❑ Small Business Applicants have high chances of getting charged off.
- ❑ Charged off proportion increases with grades moving from “A” towards “G”.
- ❑ Charged off proportion increases as Interest Rate Increases.
- ❑ Higher the public bankruptcy record greater the charged-off proportion.
- ❑ The loan amounts are bigger on average for small business purpose among all purposes of Loan.
- ❑ Those who already have Derogatory Public Records have higher charged off chances than others.
- ❑ Average interest rate is considerably higher for 60 months loan term than 36 months.
- ❑ Ones getting charged off have lower annual incomes than the ones who has fully paid for each and every grade.

Recommendation

❑ Loans for Small Business Applicants should be checked properly.

❑ Loan approval should be avoided for those who already have Derogatory Public Records.

❑ Loan approval should be avoided for those who already have Public Bankruptcy Records.

❑ Loan approval for Low quality loans should be avoided or given for smaller loan repayment term.

❑ Lower annual income applicants should be avoided for big loan amounts with higher interest Rates.

❑ Loan approval should be avoided for applicants who doesn't have a source of income.

upGrad

#LifeKoKaroLift

Thank You!