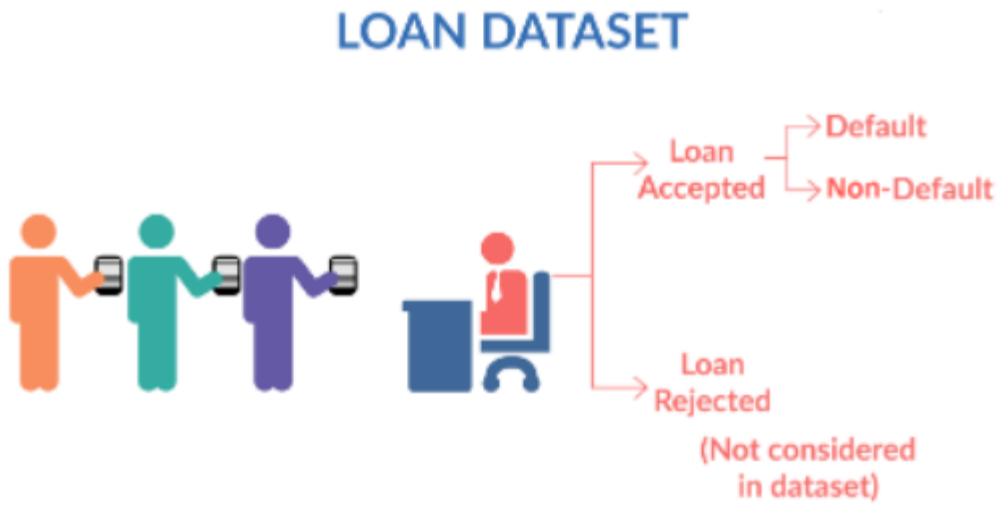
PROBLEM STATEMENT



Analyse the loan dataset on how **consumer attributes** and **loan attributes** influence the tendency of default using EDA.

In other words, to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

DATA DESCRIPTION

The data set contains 39717 rows and 111 column data

Dependent Variable: The column(variable) which is the indicator of loan default ‘loan_status’. It contains three types of values

- **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

Categorical Variable: There are 26 categorical variable(the variables containing all null values are excluded)
6 ordered categorical variables and 20 unordered categorical variables

Continuous variable: There are 28 continuous variables (the variables containing all null values are excluded)

DATA CLEANING

1. Dropped columns with all null values (57 columns are dropped)

2. Dropped columns where majority of values are null(more than 64%):, mths_since_last_record, mths_since_last_delinq

3. Dropped columns where all the values are same , hence it wont be helpful in the data analysis: policy_code, application_type, next_pymnt_d , initial_list_status, pymnt_plan, tax_liens, collections_12_mths_ex_med

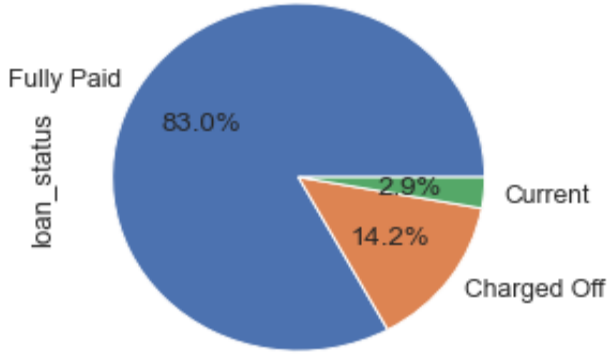
4. Dropped few more columns which wont be helpful in the data analysis:
   zip_code(another field state is available which can be used for the analysis)
   desc(long texts, similar inference can be drawn from purpose column)
   member_id and url (one unique identifier(id column) is enough for a data set)
   emp_title (more than 60 % of data is unique)
   title(similar inference can be drawn from purpose column)

5. Removed the % symbol in ‘int_rate’ and ‘revol_util’ column and coverted it to numeric so that it will be helpful in the EDA

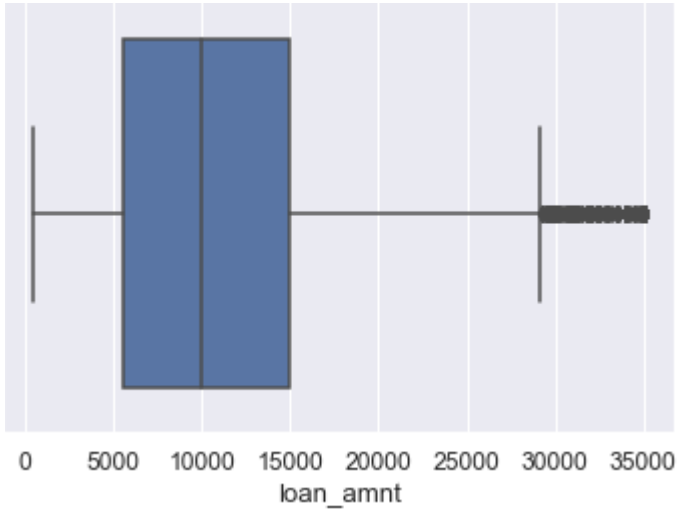
6. Replaced null values in emp_length to 0 years.(There can be a posibilty that the employees can have more experience nut not updated in the system . But there is no way to analyze that data and get a appropriate result.

UNIVARAITE ANALYSIS

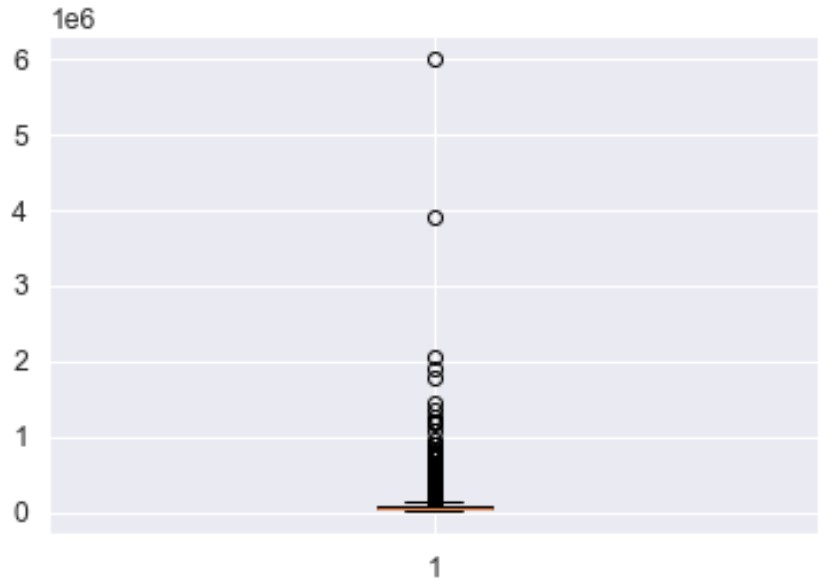
1. Loan_status: Analysing the percentage of applicants belong to each category. The Charged Off applicants comprises of 14.2% of the total dataset



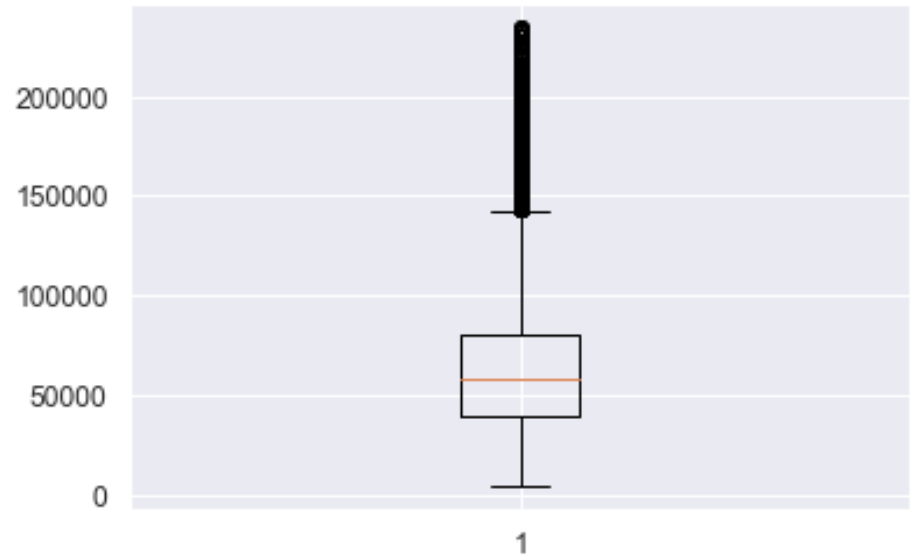
2. loan_amount: Analysing the data in loan amount to remove outliers(if required). No outliers were removed in loan_amount



3. annual_inc: : Analysing the data in loan amount to remove outliers(if required). Outliers were removed

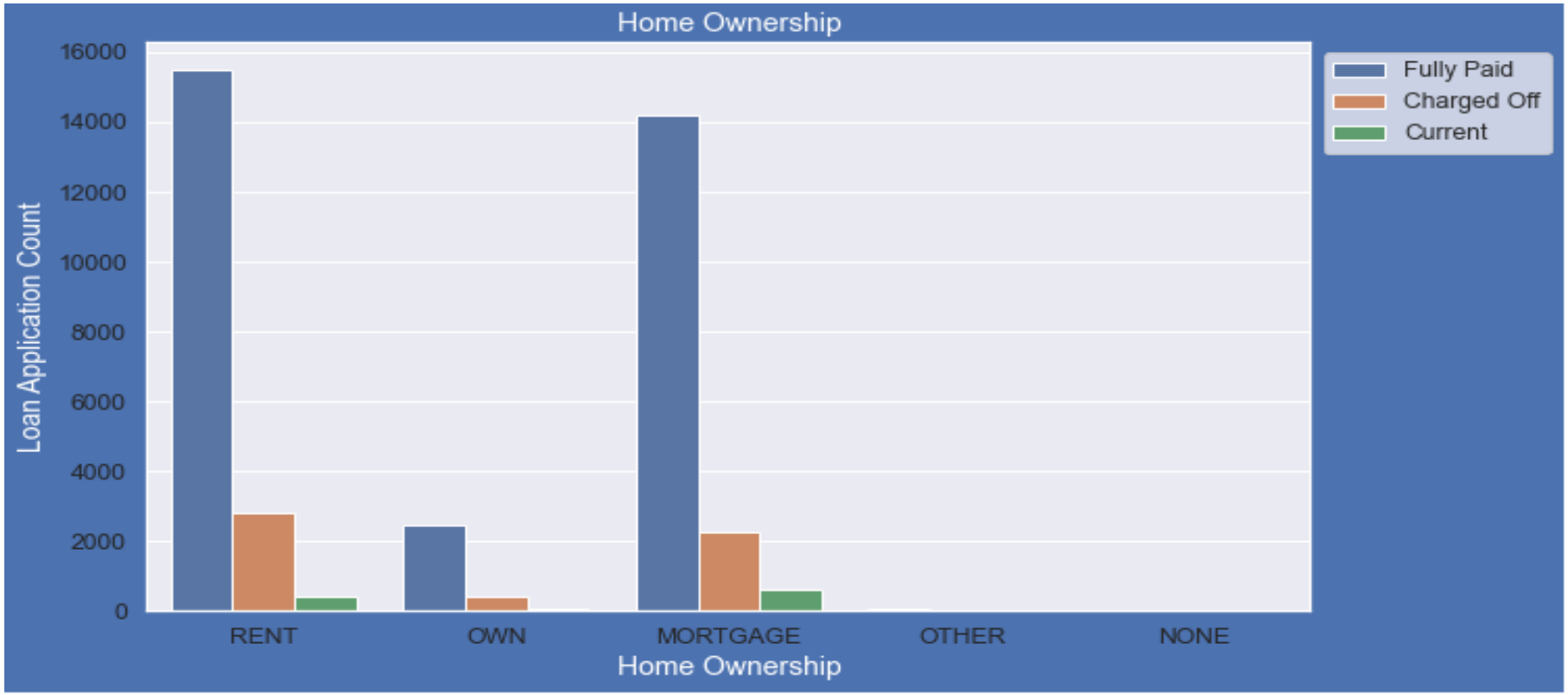


Before removal of outlier

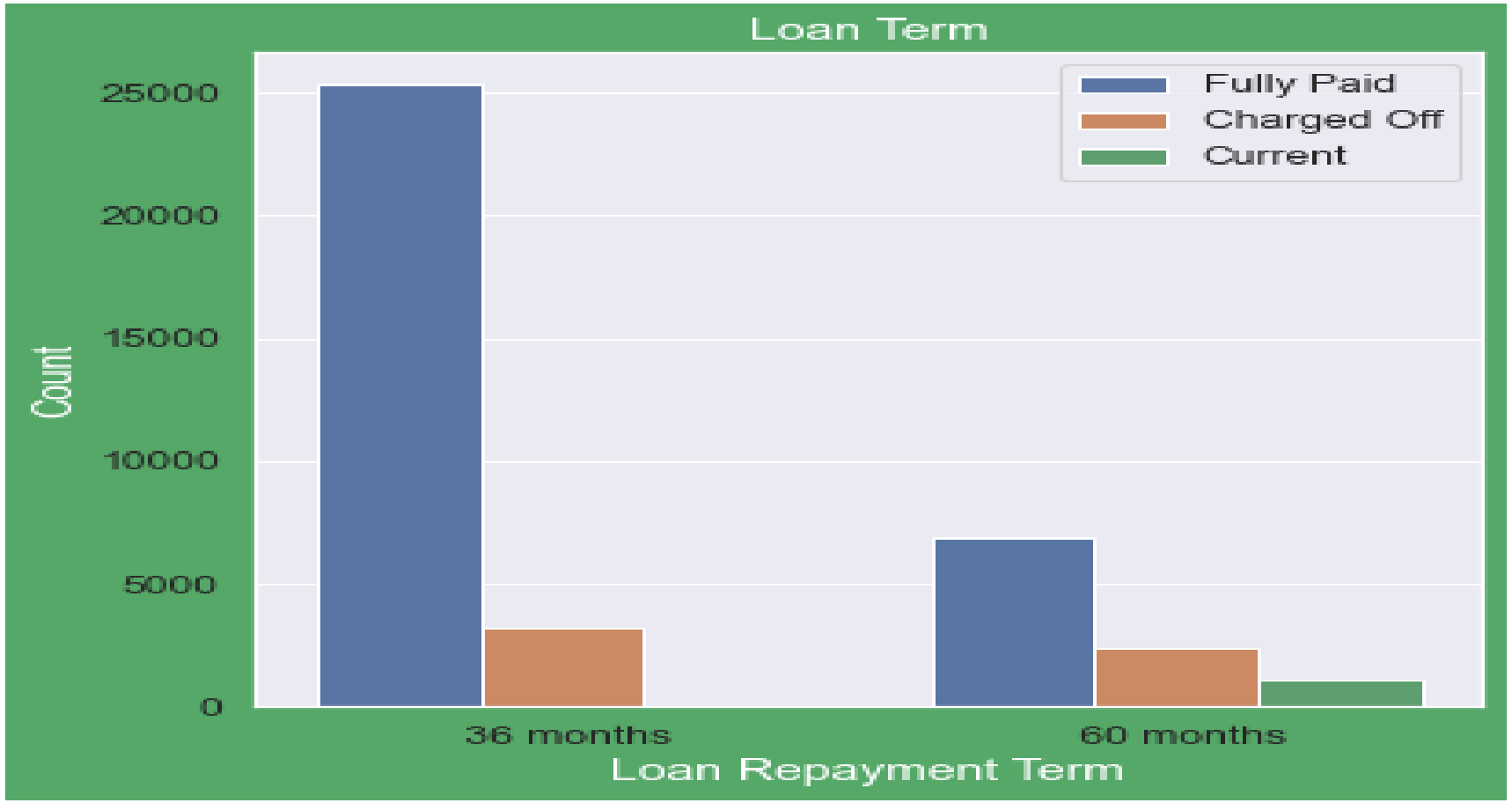


After removal of outlier

4. home_ownership(Unordered categorical variable) :We can understand than applicants having own house are more likely to pay the loan than applicants having a rented house or on mortgage



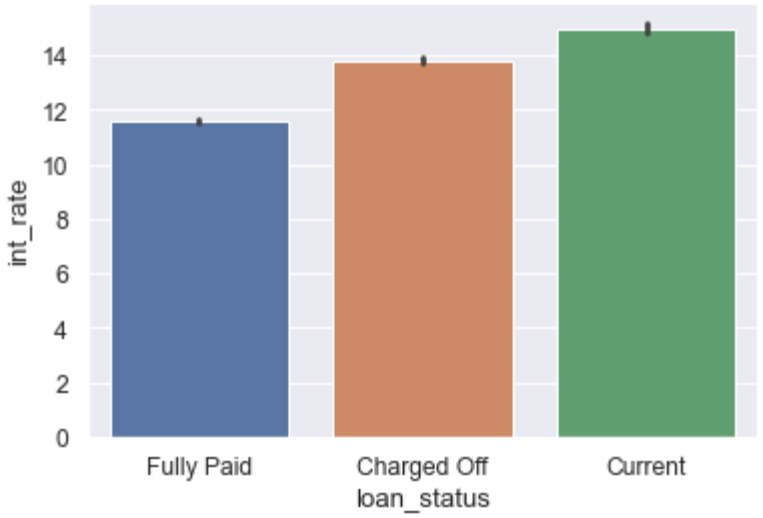
5 . term(Ordered categorical variable) : Applicants having 60 months term are more likely to be charged off, since the charged off proportion of applicants in 60 months term is more as compared to 36 month



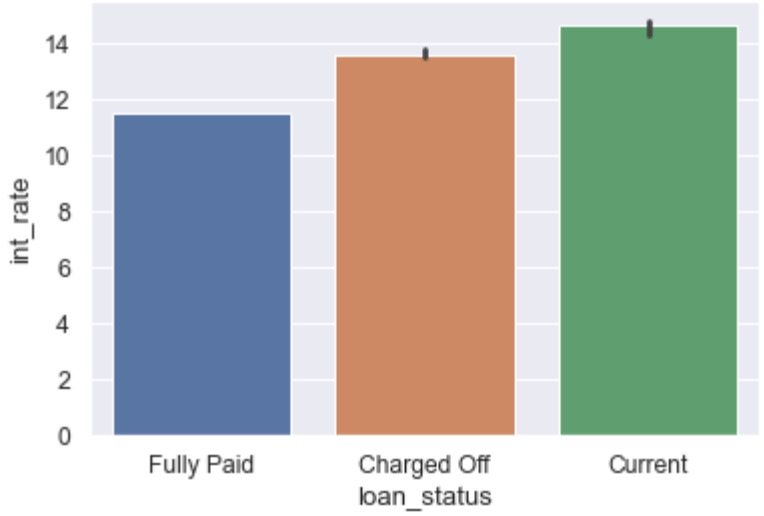
SEGMENTED UNIVARIATE ANALYSIS

1. int_rate(Continuous Variable) : Analyzing the mean and Median value interest rate on loan_status

It can be inferred that , applicants with high interest rate are more likely to be Charged Off



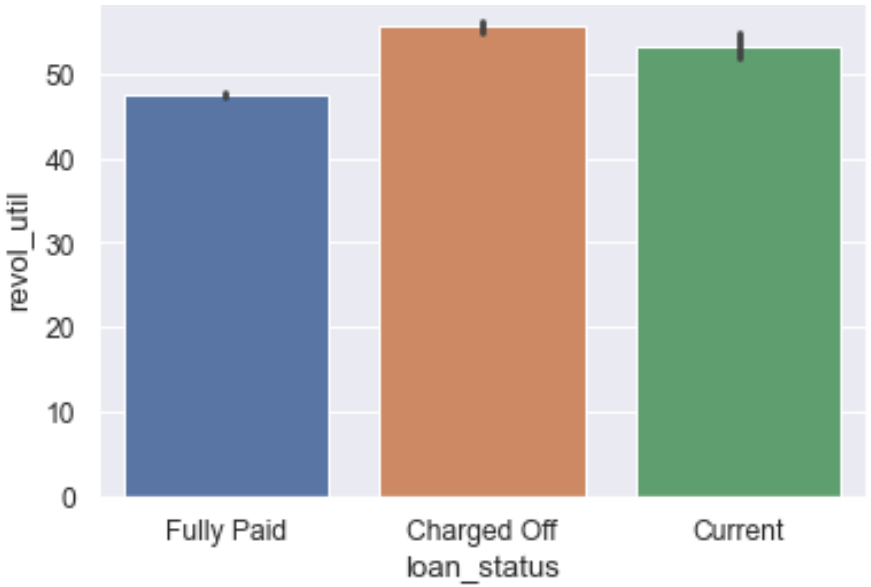
Mean : Inc_rate vs loan_status



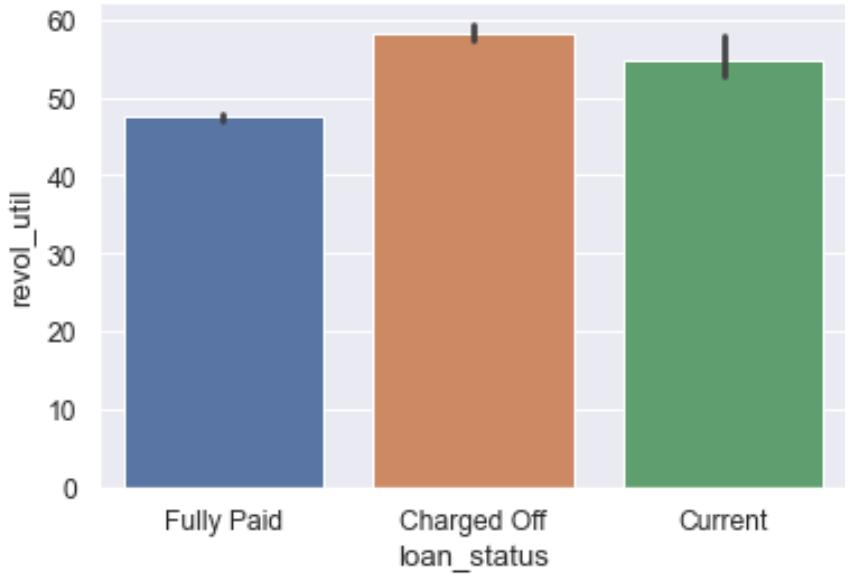
Median : Inc_rate vs loan_status

3. revol_util(Continuous Variable) : Analyzing the mean and Median value revolving utilization rate on loan_status

It can be inferred that , application with high revolving utilization rate are more likely to be Charged off



Mean: revol_util vs loan_status

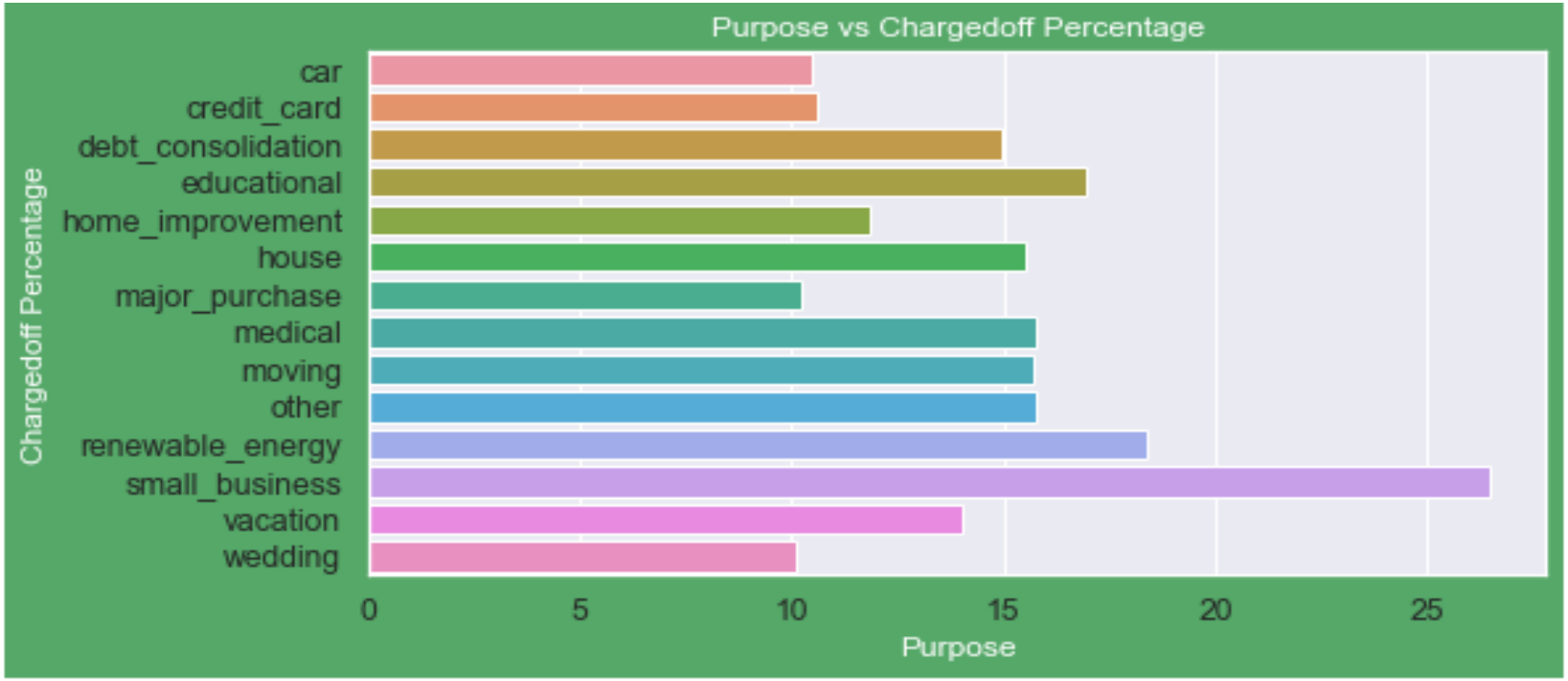


Median : revol_util vs loan_status

BIVARIATE ANALYSIS

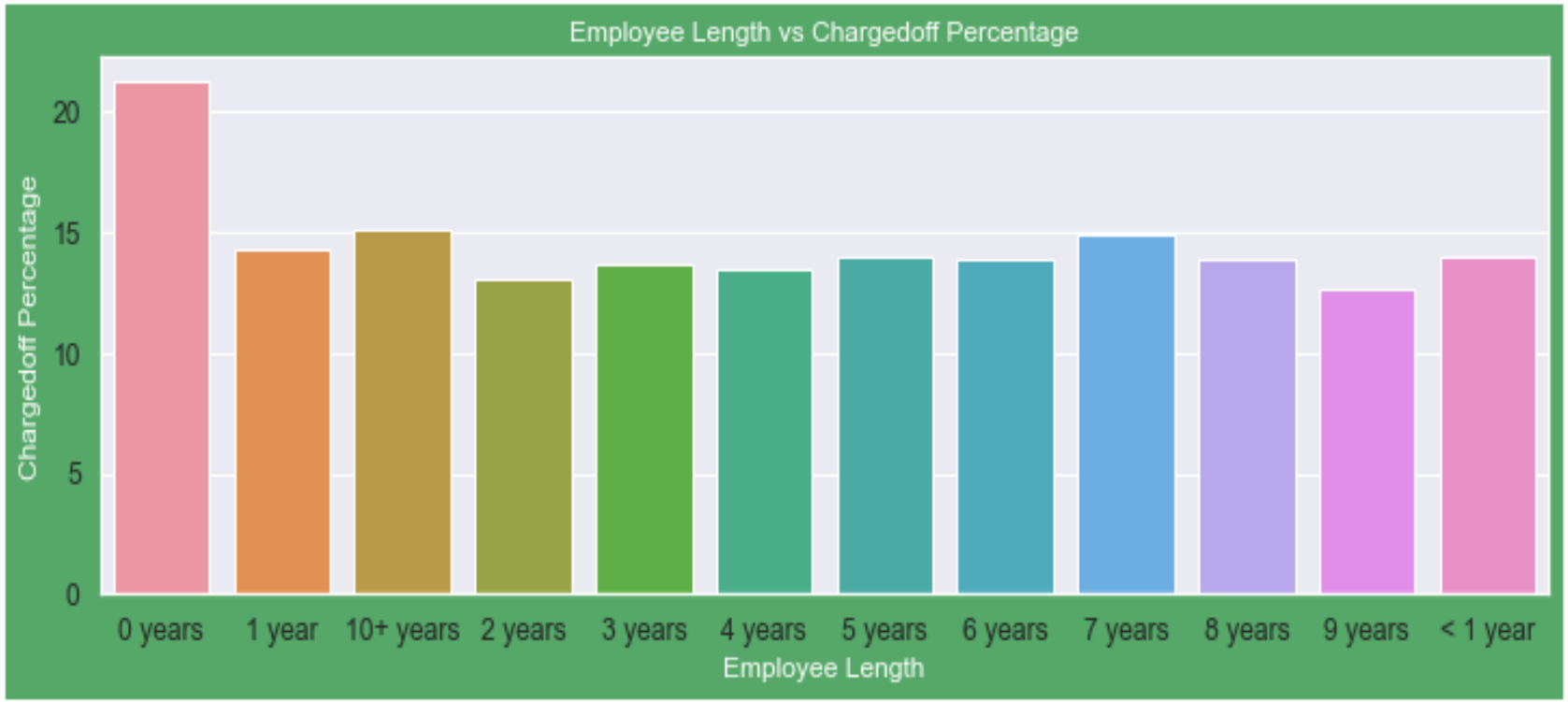
1. purpose : Analysing the impact of purpose of loan on percentage of Charged off applicants

It can be inferred that loan opted for "Small Business" are more likely to be Charged off



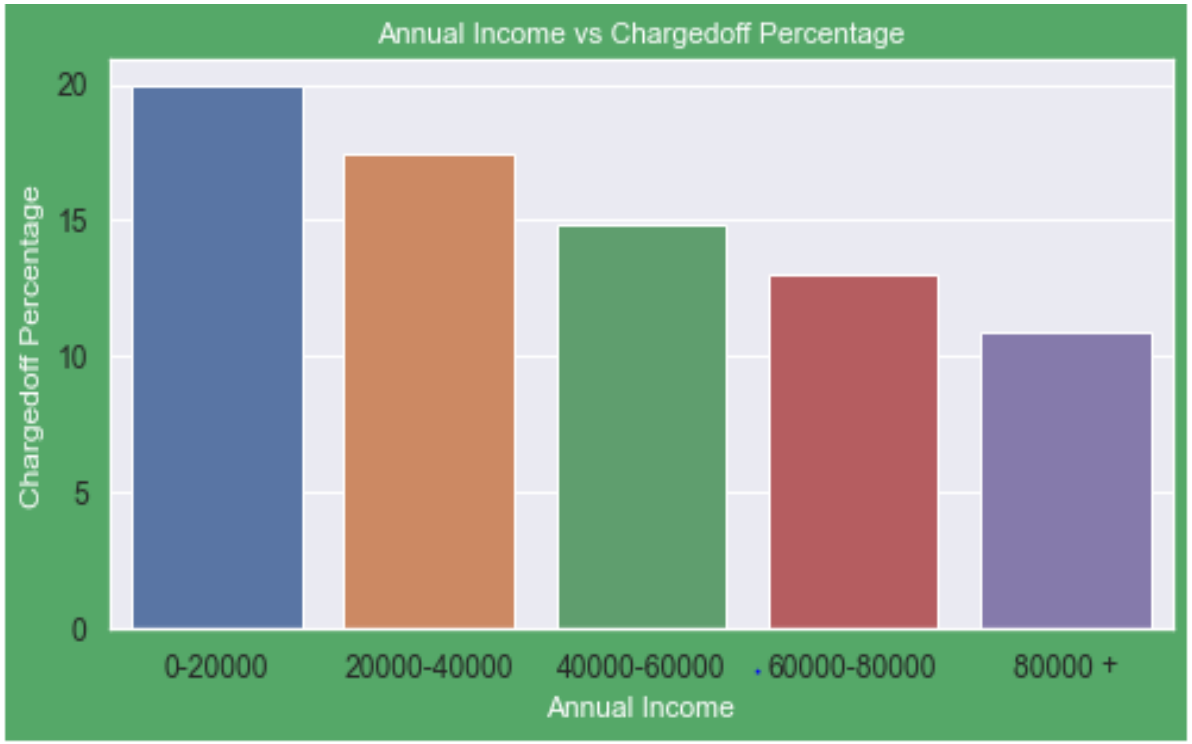
2 .emp_length: Analysing the impact of Employment length on percentage of Charged off applicants

It can be inferred that applicants with less experience or no experience are more likely to be Charged off



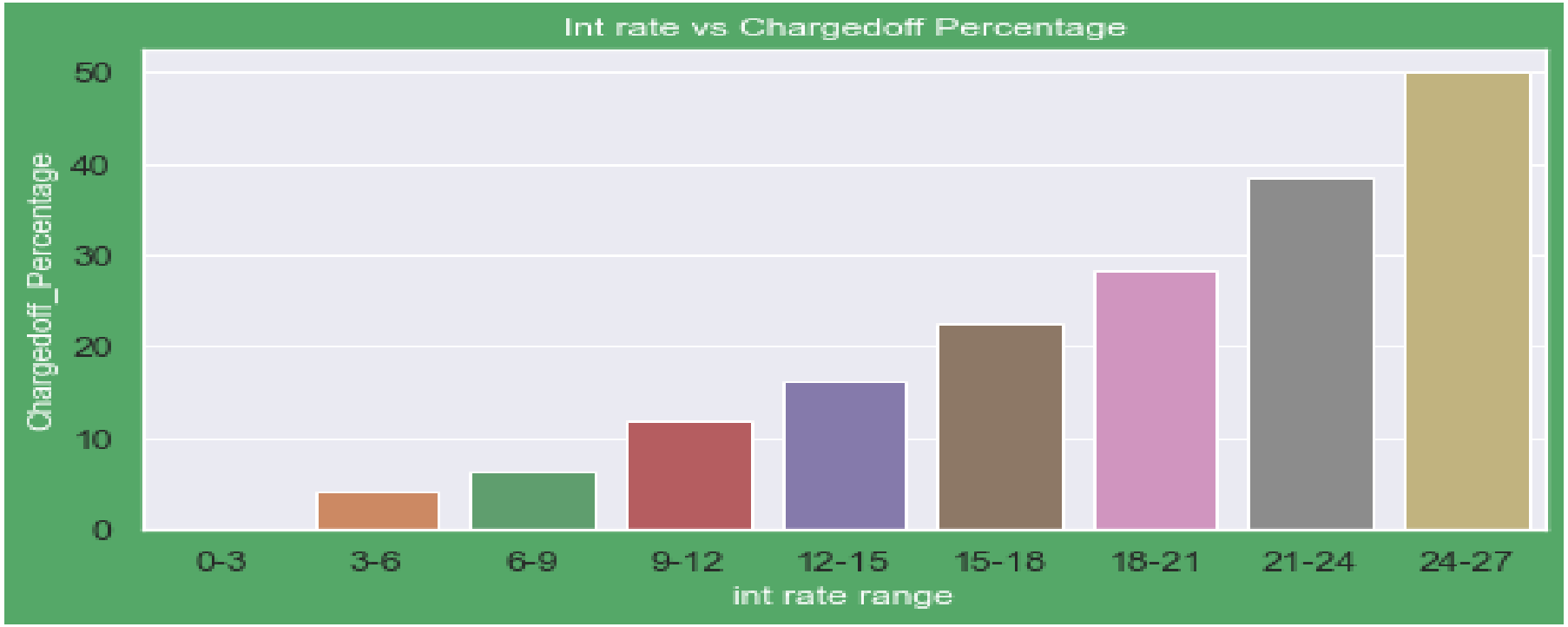
3 .annual_inc : Analysing the impact of annual income of loan on percentage of Charged off applicants

It can be inferred that applicants with low annual income are more likely to be Charged off



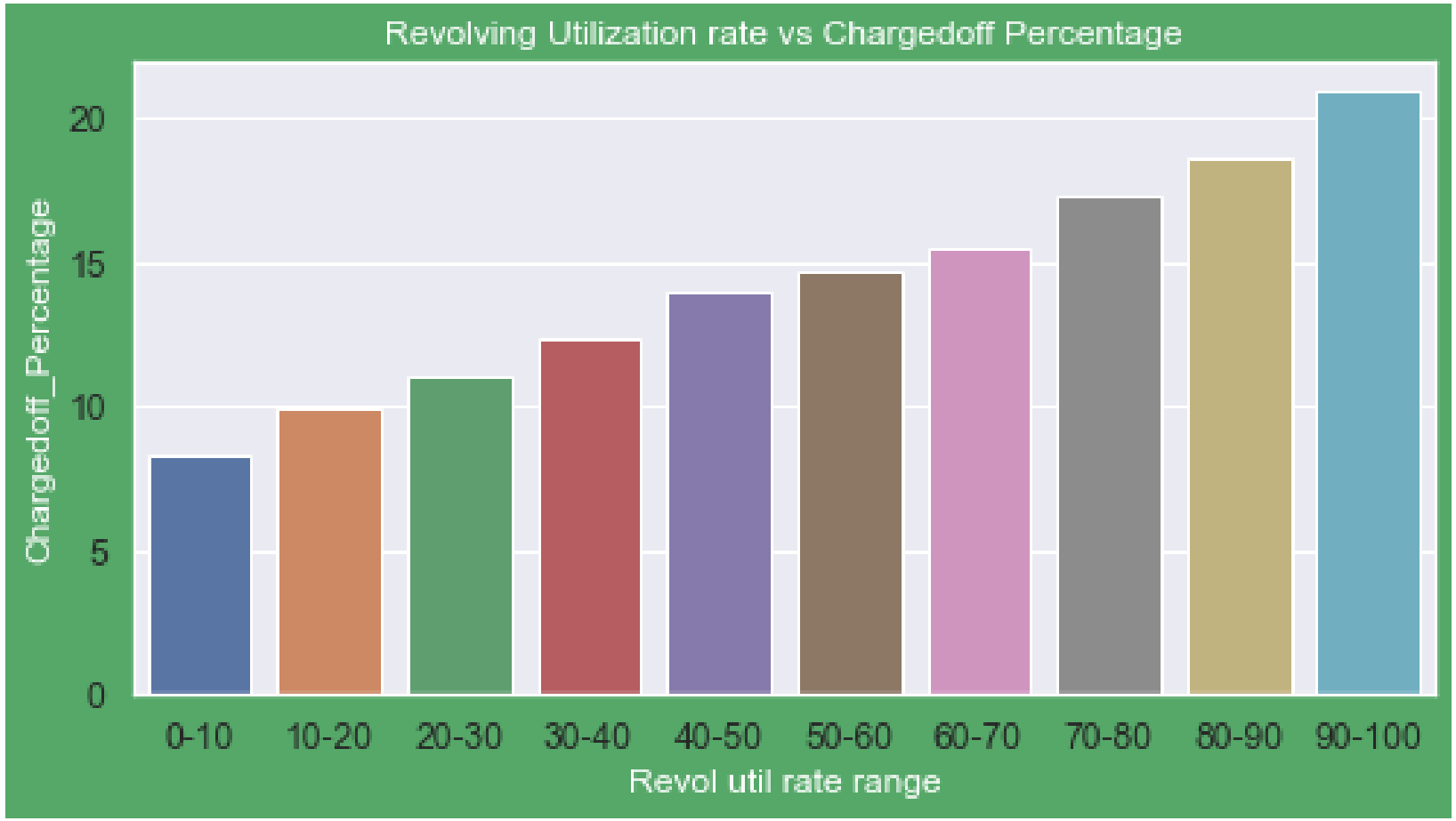
4 .int_rate : Analysing the impact of int_rate on Charged off percentage of applicants

It can be inferred that the applicants paying high interest rates are more likely to be Charged off



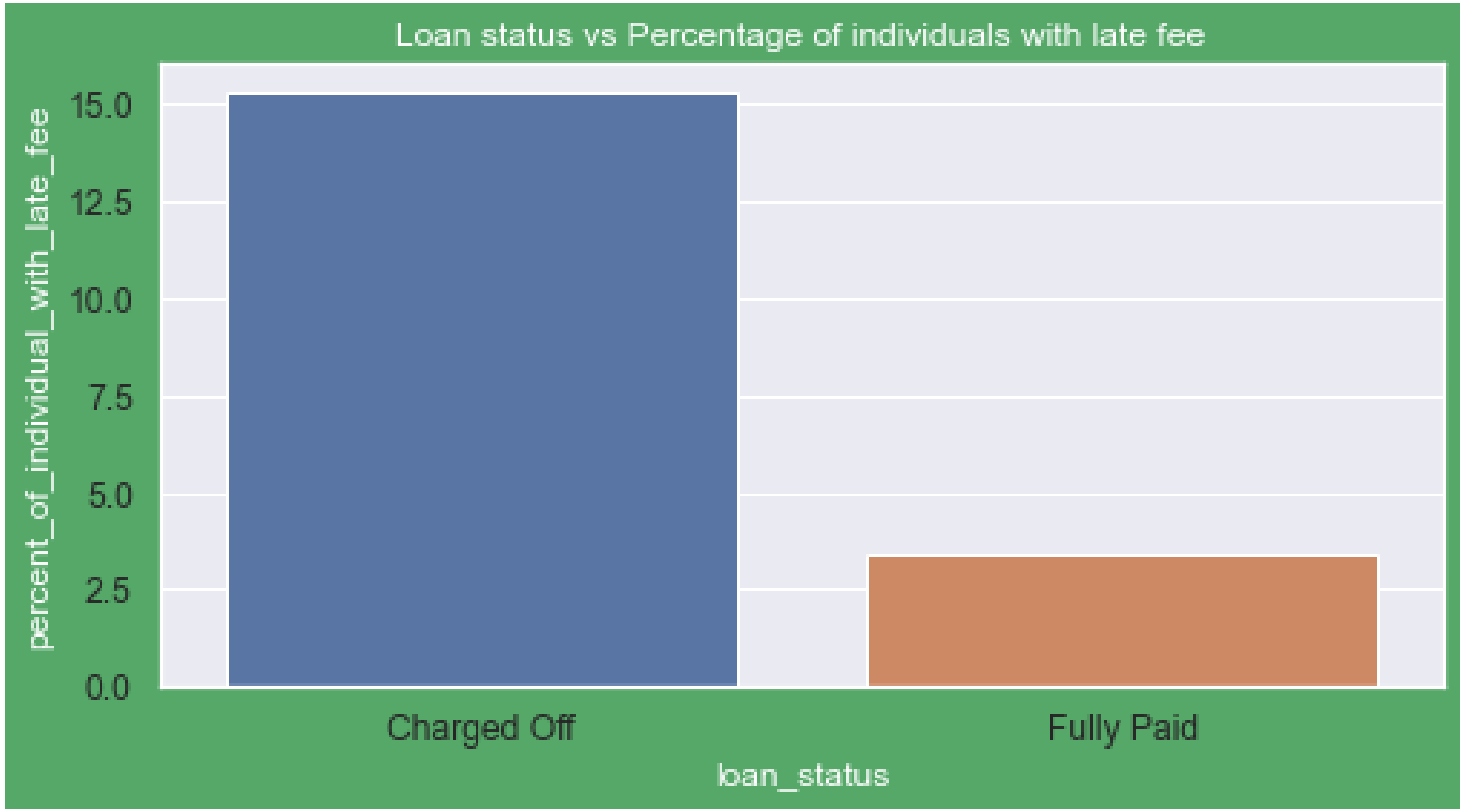
5 .revol_util: Analysing the impact of revol_util rate on Charged off percentage of applicants

It can be inferred that the applicants paying high revolving utilization rate more likely to be Charged off



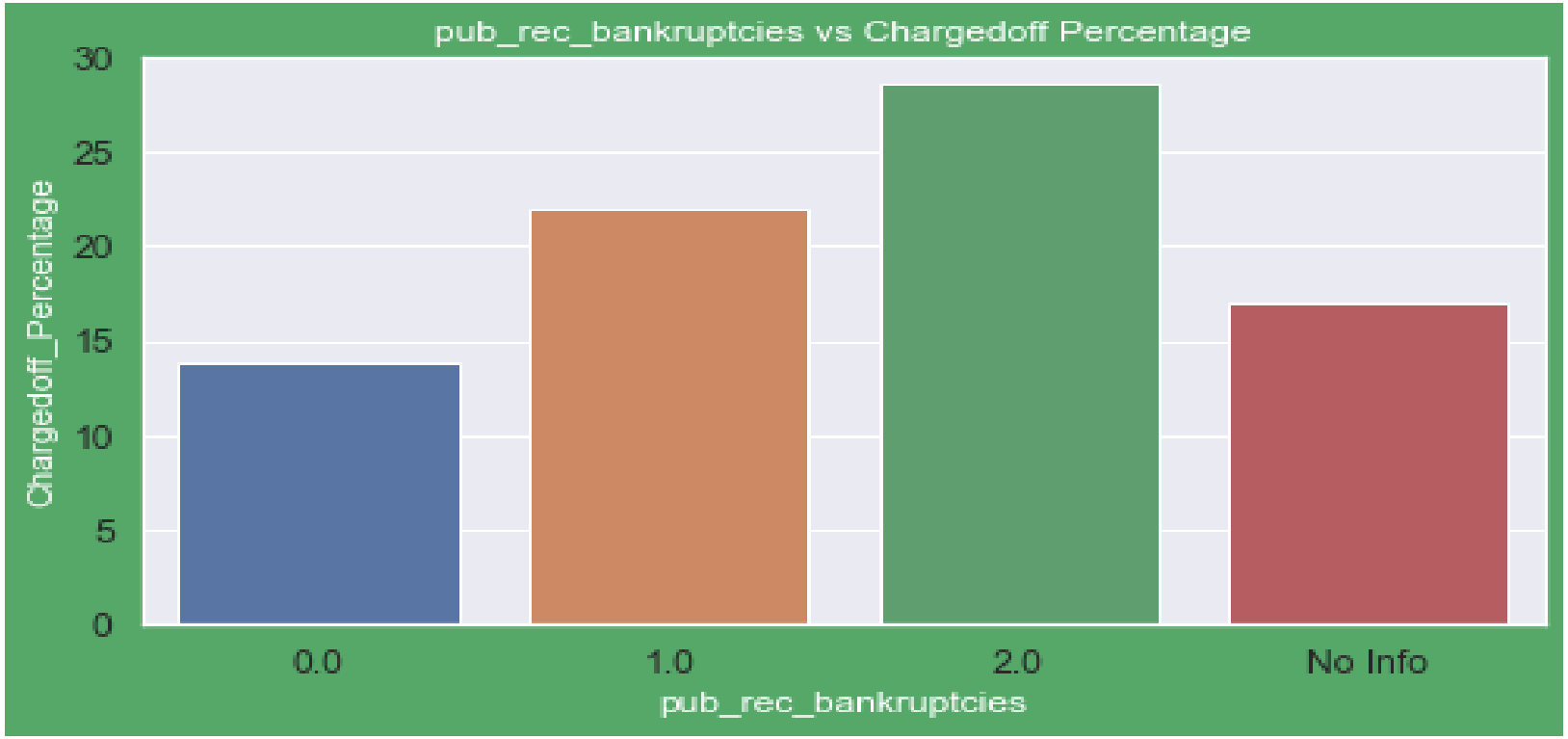
6 . late_fee: Analysing the impact of late_fee on loan status

It can be inferred that percentage of applicants having late fee is more for Charged off(around 15%) as compared to Fully paid (around 3%)



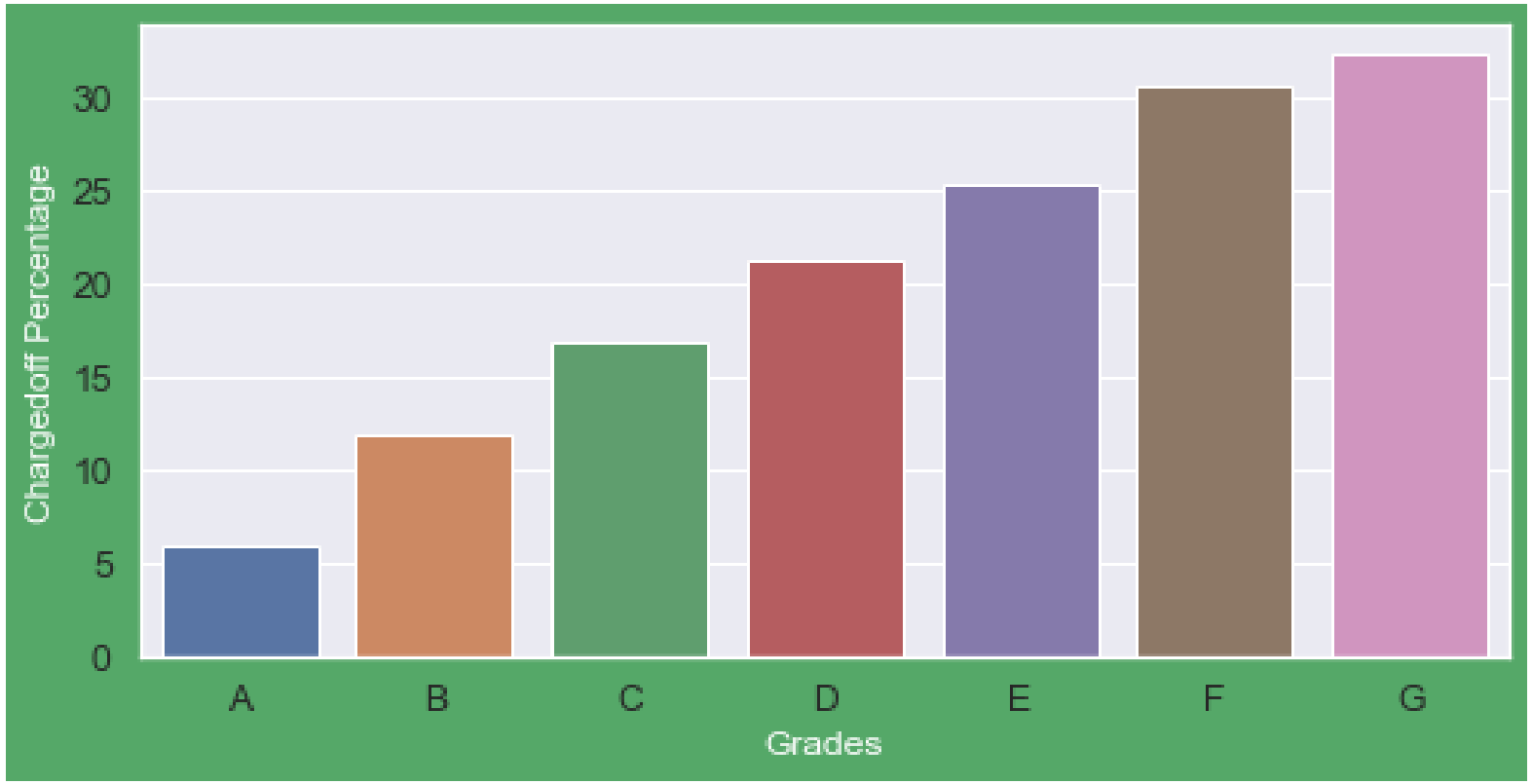
7 .pub_rec_bankruptcies : Analysing the impact of public record bankruptcies of loan on Charged off percentage of applicants

It can be inferred that applicants with more public record bankruptcies are more likely to be Charged off



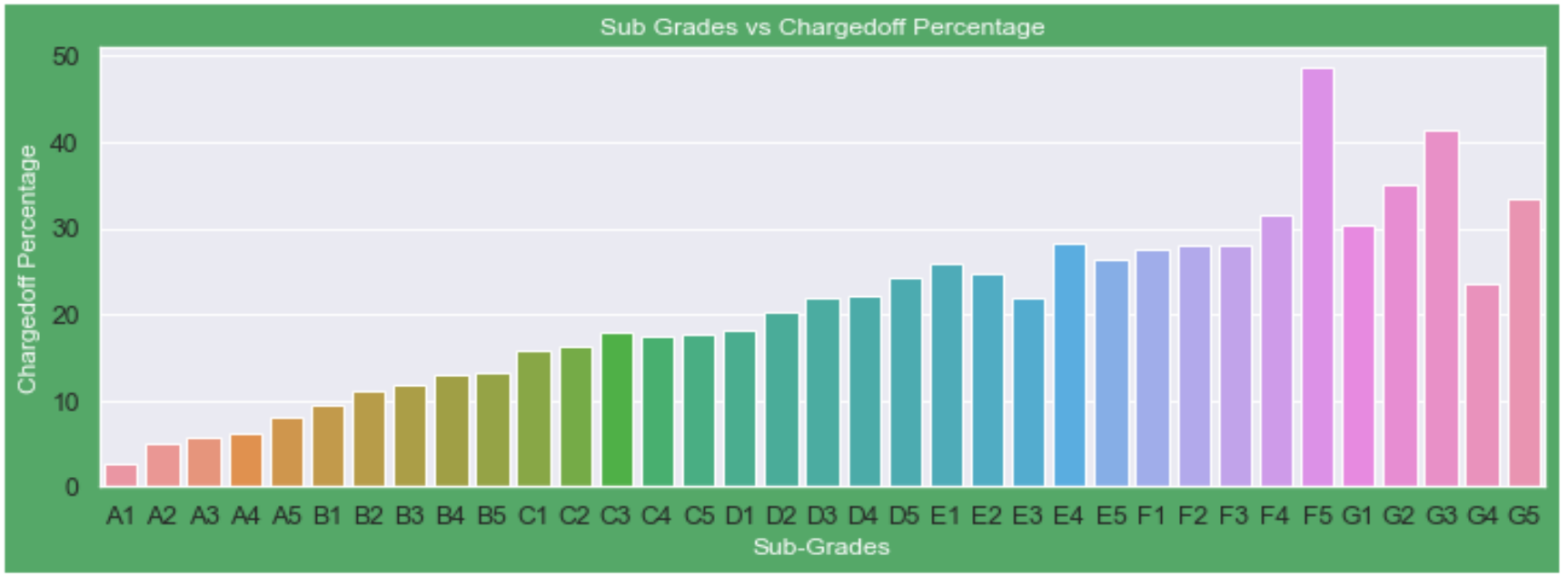
8. grade : Analysing the grade on Charged off percentage of applicants

It can be inferred that majority of the Charged off applicants belong to Grade F and G , and there is a increase in charged off percentage with each grade



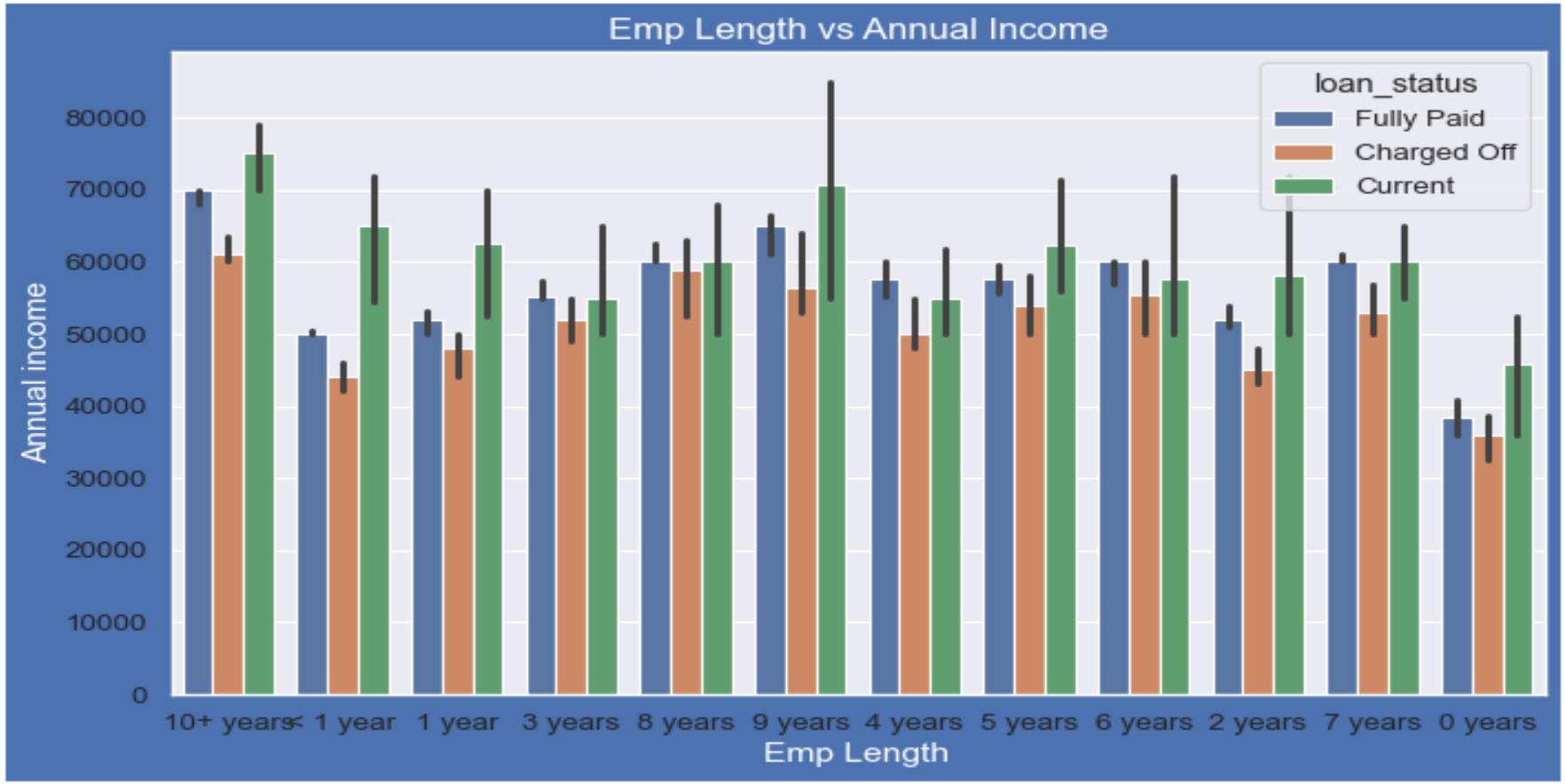
9. subgrade : Analysing the impact of subgrade on Charged off percentage of applicants

It can be inferred that majority of the Charged off applicants belong to subgrades of F and G.



9. Annual Income vs Employment length with loan status.

It can be inferred that the annual income of Charged off applicants are lower than fully paid irrespective of the employment length percentage of applicants having late fee is more for Charged off as compared to Fully paid

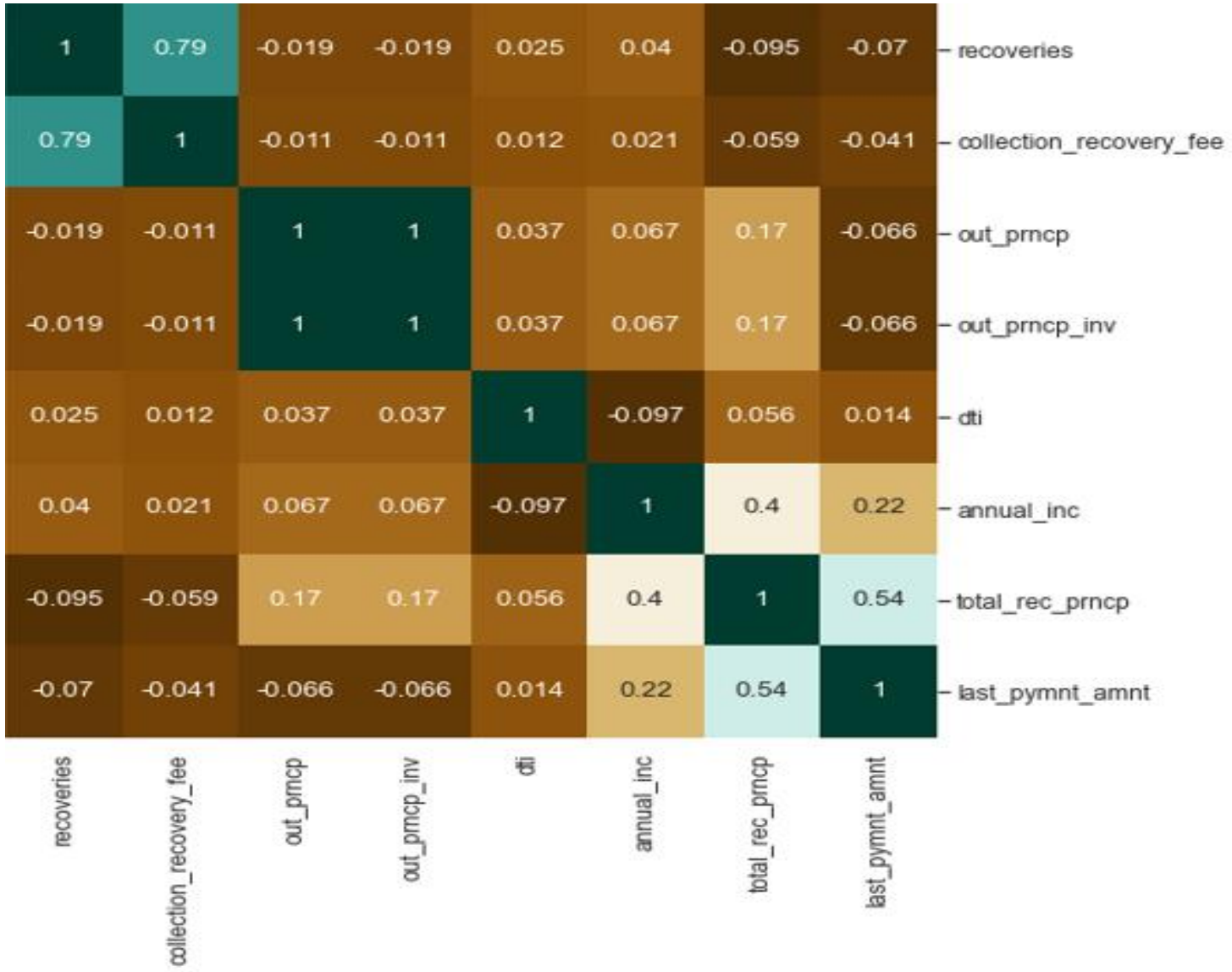


CORRELATION MATRIX

1. annual_inc and dti are highly correlated(negatively)

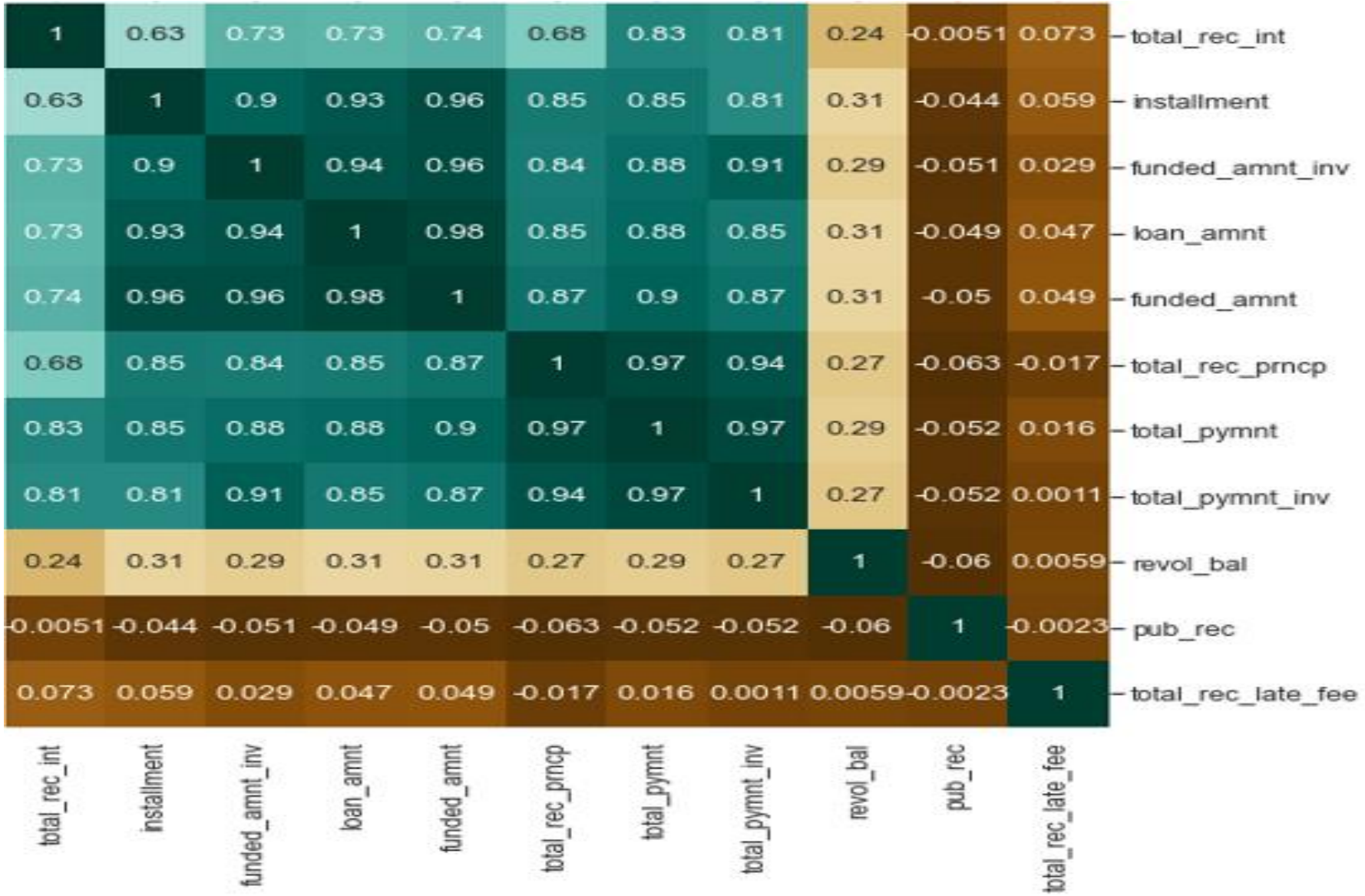
2. recoveries and is highly correlated with total_rec_prncp (negatively)

3. total_rec_prncp is negatively correlated with recoveries and collection_recovery_ fee



1. revol_bal is negatively correlated with pub_rec

2. total_rec_int,installment, funded_amnt_inv, funded_amnt, loan_amnt, total_rec_prncp, total_payment, total_payment_inv are positively correlated with each other and negatively correlated with pub_rec and total_rec_late_fee



RECOMENDATIONS

Below are the major driving factors for loan default as per my analysis:

1. Lower Annual income (default percentage is maximum for annual income <20000)
2. Purpose of loan is “Small Business”
3. High Interest Rate (>21%)
4. Term of loan is 60 months
5. Home ownership is ‘Rent’ or ‘Mortgage’
6. Revolving utilization rate above 90%
7. Employment term <=1
8. Public record bankruptcies >=2
9. Applicants paying late fee
10. Grade G and F applicants (subgrades of F and G)

If the company would not choose to provide loans or take necessary actions for applicants satisfying multiple conditions from the above list then there might be a chance of reducing defaulter applicants