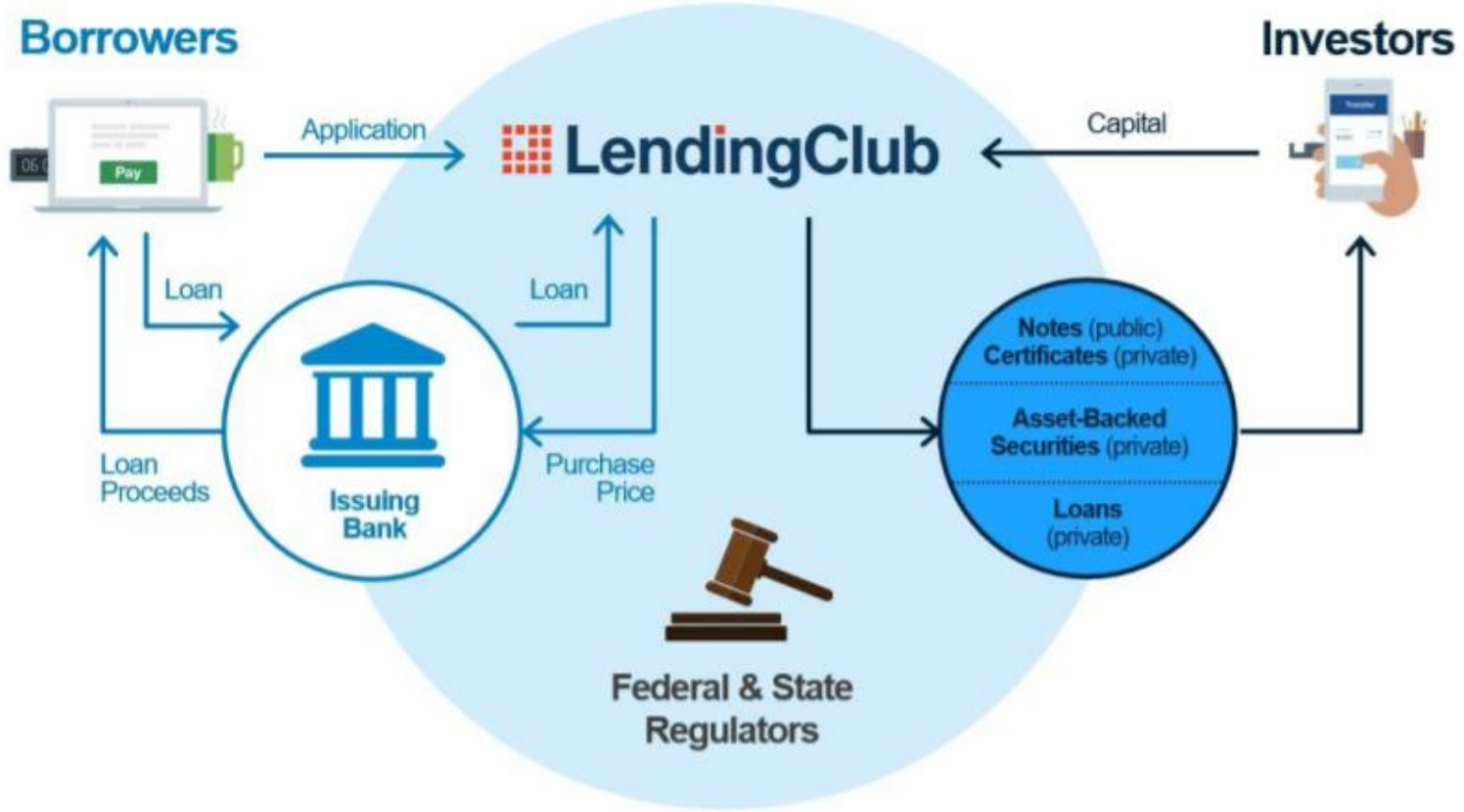
# Lending Club Case Study





**Group Members:**
**Aarushi Gupta**
**Ankit Sharma**

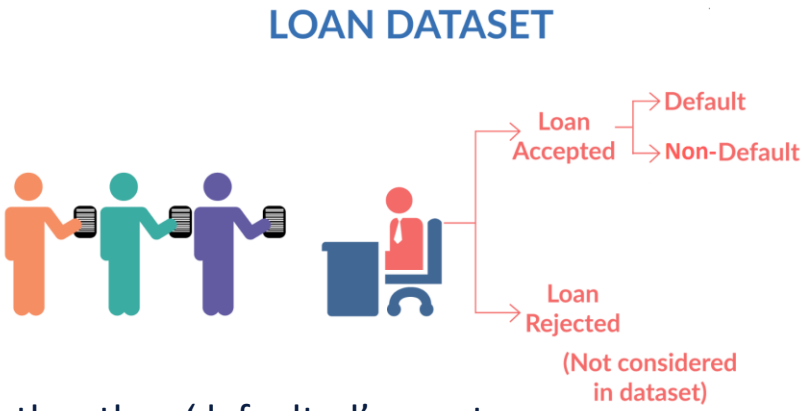
# CONTENT

- Problem statement
- Fixing Rows and Columns
- Data Preparation and Standardization
- Dealing with Missing Values
- Removing Outliers
- Univariate Analysis on Categorical Variables
- Univariate Analysis on Numerical Variables
- Segmented Univariate Analysis (Using concept of Binning)
- Correlation Metrics and Heat Map for all the variables
- Bivariate Analysis
- Recommendations on the basis of Univariate and Bivariate Analysis

# LendingClub


- **LendingClub** is a peer-to-peer lending company, headquartered in San Francisco, California.
- LendingClub enabled borrowers to create unsecured personal loans between $1,000 and $40,000.
- The standard loan period was three years. Investors were able to search and browse the loan listings on LendingClub website and select loans that they wanted to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose.
- Investors made money from the interest on these loans. LendingClub made money by charging borrowers an origination fee and investors a service fee.

# Problem Statement




- The data given contains 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. through **Exploratory Data Analysis (EDA)** . Thus, we have to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

- When a person applies for a loan, there are **two types of decisions** that could be taken by the company:
- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
  1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
  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'.
  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

- **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)

# Fixing Rows and Columns

- Loading the dataset. (There are 39,717 rows and 111 columns)

| | id | member_id | loan_amnt | funded_amnt | funded_amnt_inv | term | int_rate | installment | grade | sub_grade | ... | num_tl_90g_dpd_24m | num_tl_op_past_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1077501 | 1296599 | 5000 | 5000 | 4975 | 36 months | 10.65% | 162.87 | B | B2 | ... | NaN | |
| 1 | 1077430 | 1314167 | 2500 | 2500 | 2500 | 60 months | 15.27% | 59.83 | C | C4 | ... | NaN | |
| 2 | 1077175 | 1313524 | 2400 | 2400 | 2400 | 36 months | 15.96% | 84.33 | C | C5 | ... | NaN | |

- Finding duplicate rows and columns and fixing it.

- Checking the number of null values in the entire dataset and removing the columns which have all the null values. (39,717 rows and 54 columns)

- Removing the columns which has large percentage of null values present in it. (39,717 rows and 53 columns)

- Finding the unique values for all the columns and removing those columns in which each row has unique values and if only single value is present in the whole column. (39,717 rows and 41 columns)

- The fields that are created after a loan application is approved doesn't make sense for our analysis towards the business objective .So, will remove the columns which is after loan approval. (39,717 rows and 27 columns)

# Data Preparation and Standardization

- Checking for the data types of all the columns. (Initially all the columns are object types)

- Changing columns such as *loan_amnt, funded_amnt, funded_amnt_inv, installment, annual_inc, dti, inq_last_6mths, open_acc, pub_rec,total_acc, pub_rec_bankruptcies* into appropriate numeric types.

- Change the columns *int_rate* and *revol_util* from string to float type by first stripping % sign and then changing to numeric type.

- Changing column name term to *term_in_month.*

- Changing the date columns such as the *earliest_cr_line, issue_d, last_credit_pull_d* from string format to datetime format.

- Since loan status "Current" doesn't give any info for our analysis for approving or rejecting application, So dropping this data makes sense.

- Mapping loan status 'Fully Paid' as 0 and 'Charged_off' as 1 for our analysis

```
loan_amnt                int64
funded_amnt              int64
funded_amnt_inv        float64
term_in_months           int32
int_rate               float64
installment            float64
grade                   object
sub_grade               object
emp_length              object
home_ownership          object
annual_inc             float64
verification_status     object
issue_d                 object
loan_status              int64
purpose                 object
zip_code                object
addr_state              object
dti                    float64
earliest_cr_line        object
inq_last_6mths           int64
open_acc                 int64
pub_rec                  int64
revol_util             float64
total_acc                int64
last_credit_pull_d      object
pub_rec_bankruptcies   float64
dtype: object
```

# Dealing with Missing Values

- Exploring all the columns with null values.

```
emp_length              1033
revol_util                50
pub_rec_bankruptcies     697
dtype: int64
```

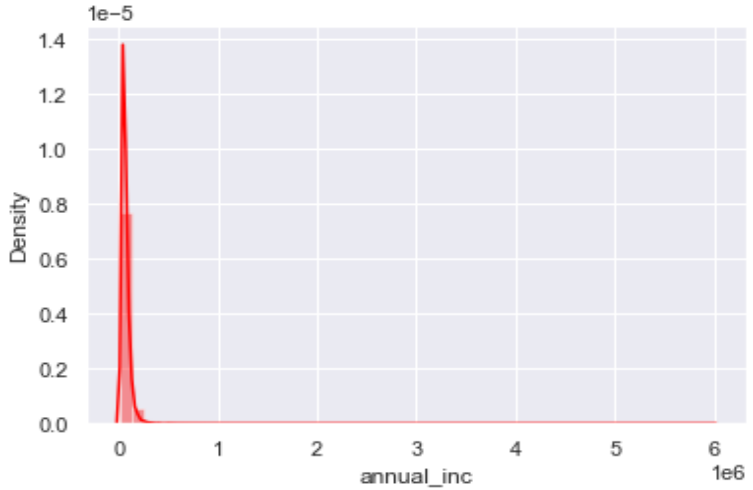
- Filling the missing values for *emp_length* with 10+years since this is the **mode** of the value and there are not much rows as compared to the entire dataset.

- Filling the missing values for *revol_util* with **median** value of the column.

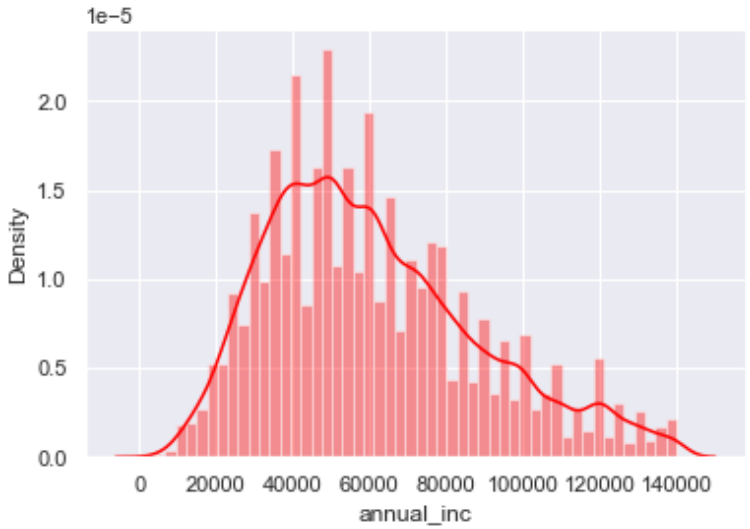
- Filling the missing values for *pub_rec_bankruptcies* with **mode** of the values.

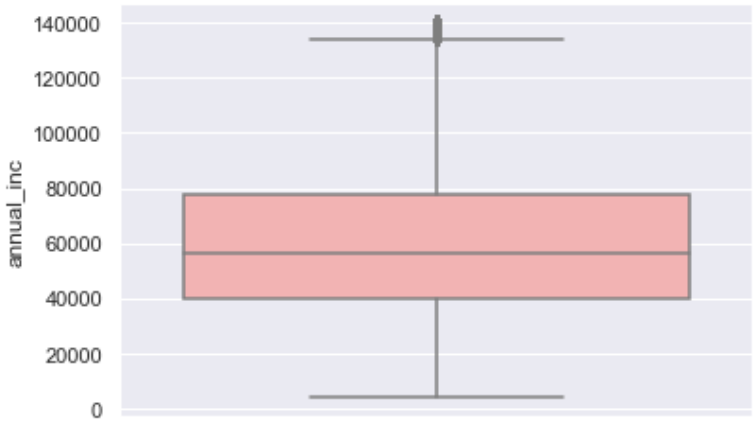
```
loan_amnt               0
funded_amnt             0
funded_amnt_inv         0
term_in_months          0
int_rate                0
installment             0
grade                   0
sub_grade               0
emp_length              0
home_ownership          0
annual_inc              0
verification_status     0
issue_d                 0
loan_status             0
purpose                 0
zip_code                0
addr_state              0
dti                     0
earliest_cr_line        0
inq_last_6mths          0
open_acc                0
pub_rec                 0
revol_util              0
total_acc               0
last_credit_pull_d      0
pub_rec_bankruptcies    0
dtype: int64
```

# Removing Outliers





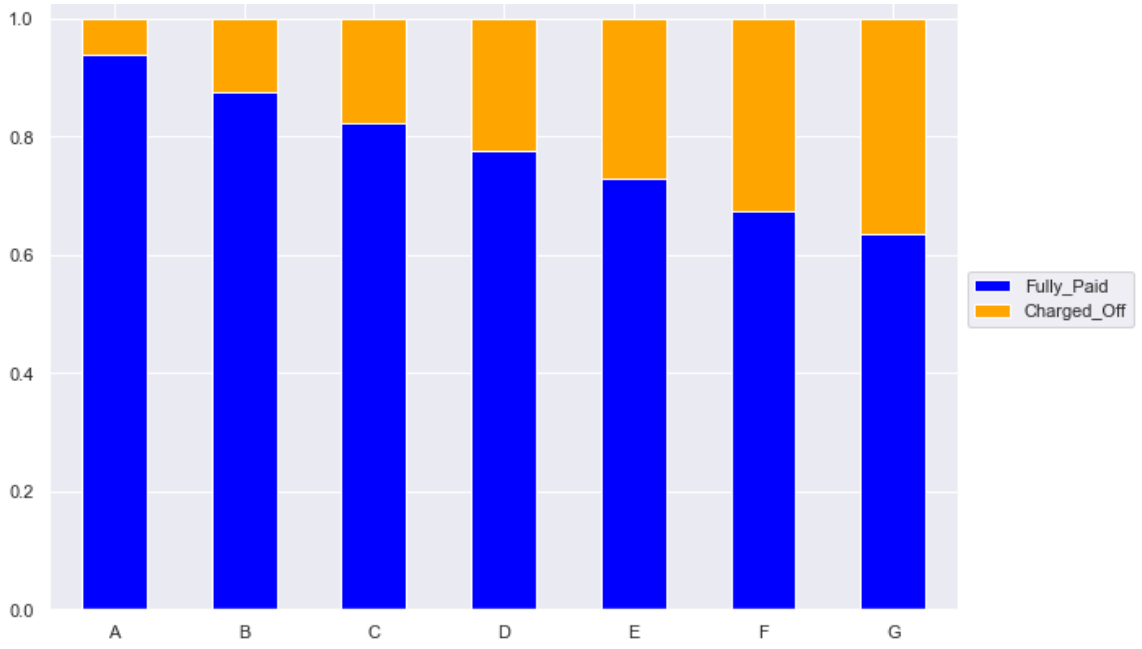



- The Annual Income has large outliers as can be seen from the top two plots(distributive plot and boxplot respectively).
- Here outliers are situated at the higher end.
- Removing them by dropping 1% or 0.5% of the problematic samples (Below graphs shows normal and symmetric distribution of the annual income after removing outliers)

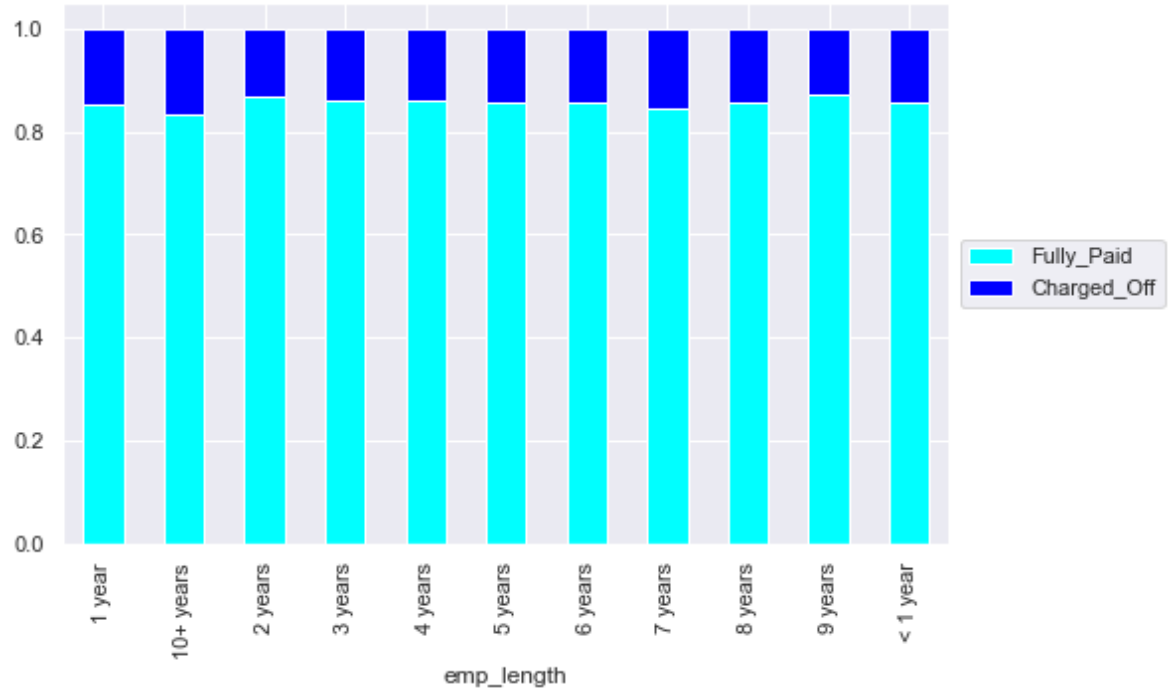
# Univariate Analysis of Categorical Variables

## Grades




- As can be seen borrowers with grade 'A' has least chances of defaulting
- The lower the grade of the borrower (here G is the lowest), more chances are for defaulting.

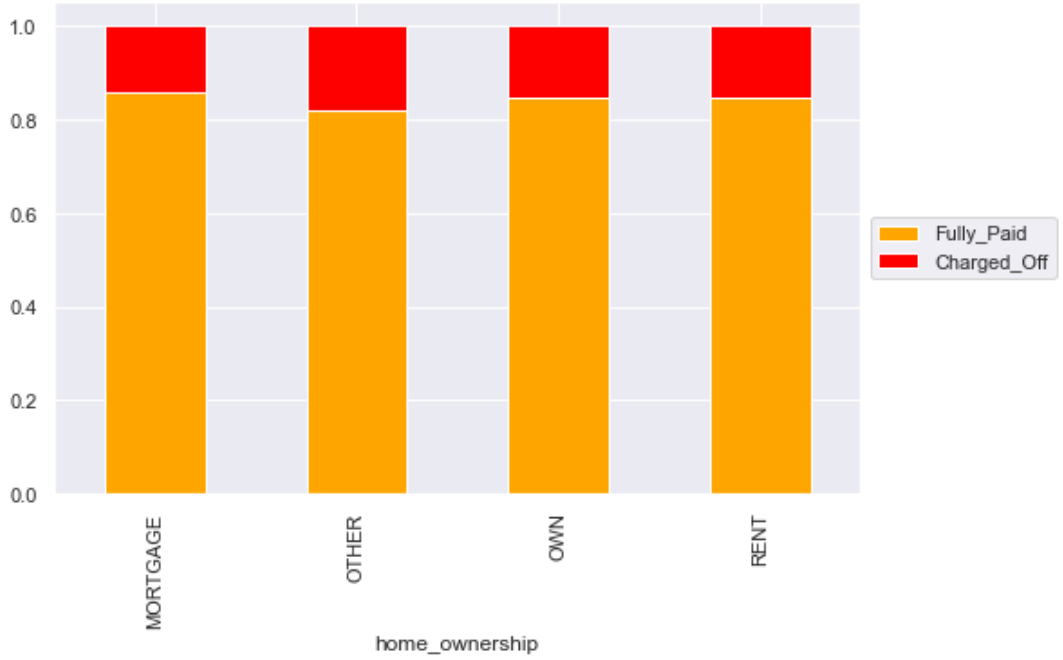
## Employment *Length*




- There is no significant effect on employment length on the borrower's default rate or not
- But the borrower with large employment length have more chances for paying back the loan.

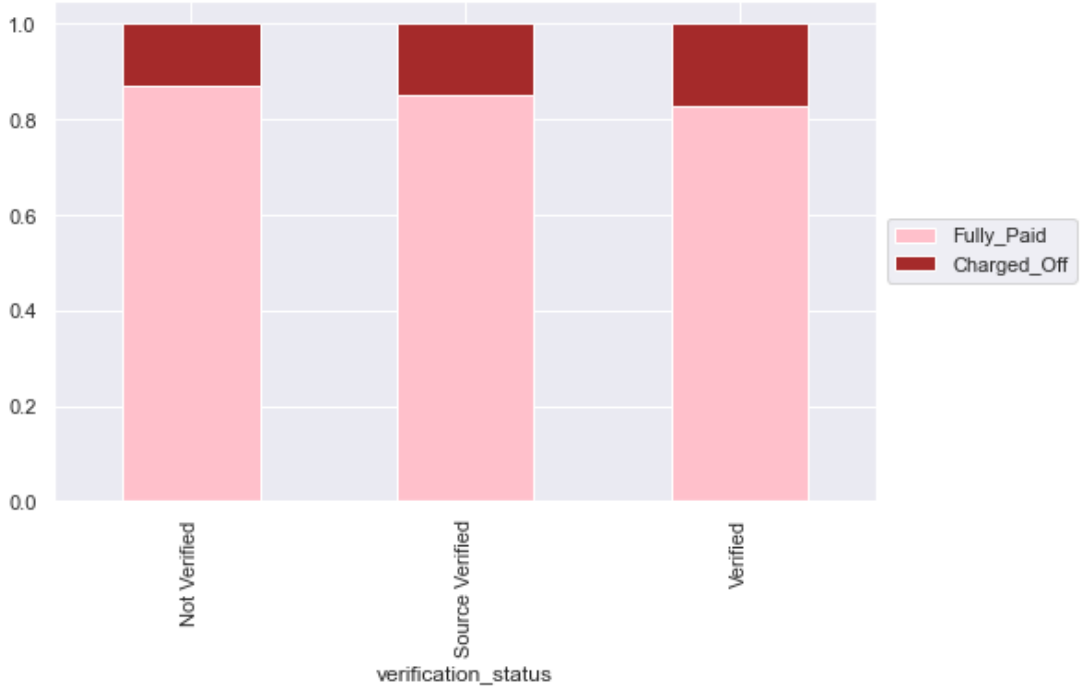
# Univariate Analysis of Categorical Variables

## Home Ownership




- Nothing much can be concluded for the home ownership since the default rate is almost similar for all the categories
- Also the 'OTHER' category is not specified which might contain more relevant information about the category of home ownership for customers that are more likely to default.

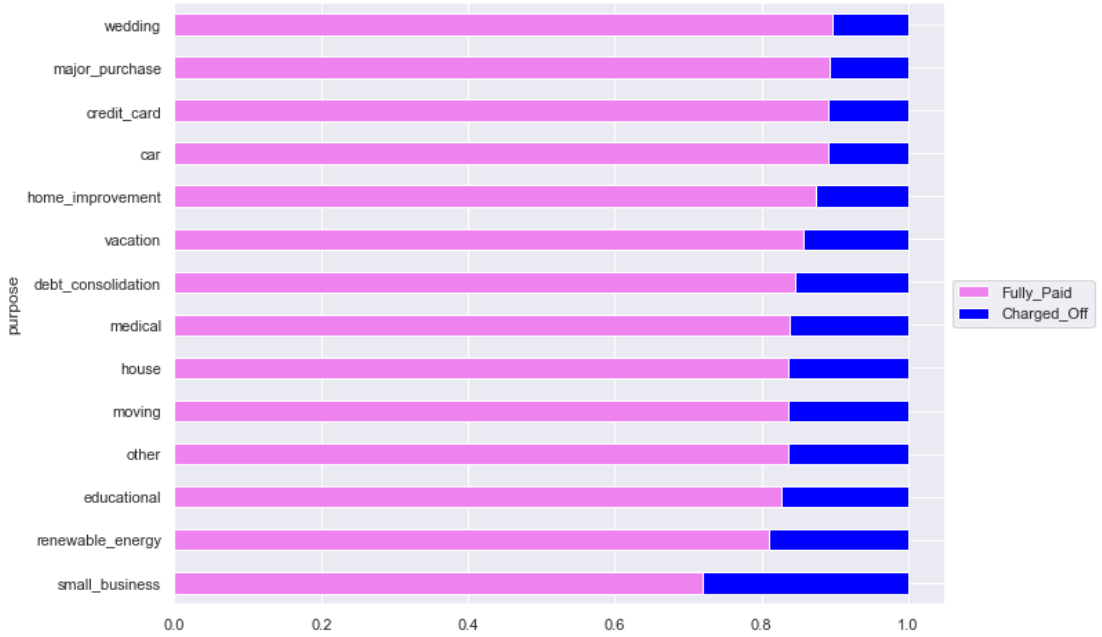
## Verification Status




- As can be seen verified customers have higher default rate
- One reason for this might be that the verification process was not done properly and hence even the verified customers have chances for defaulting.

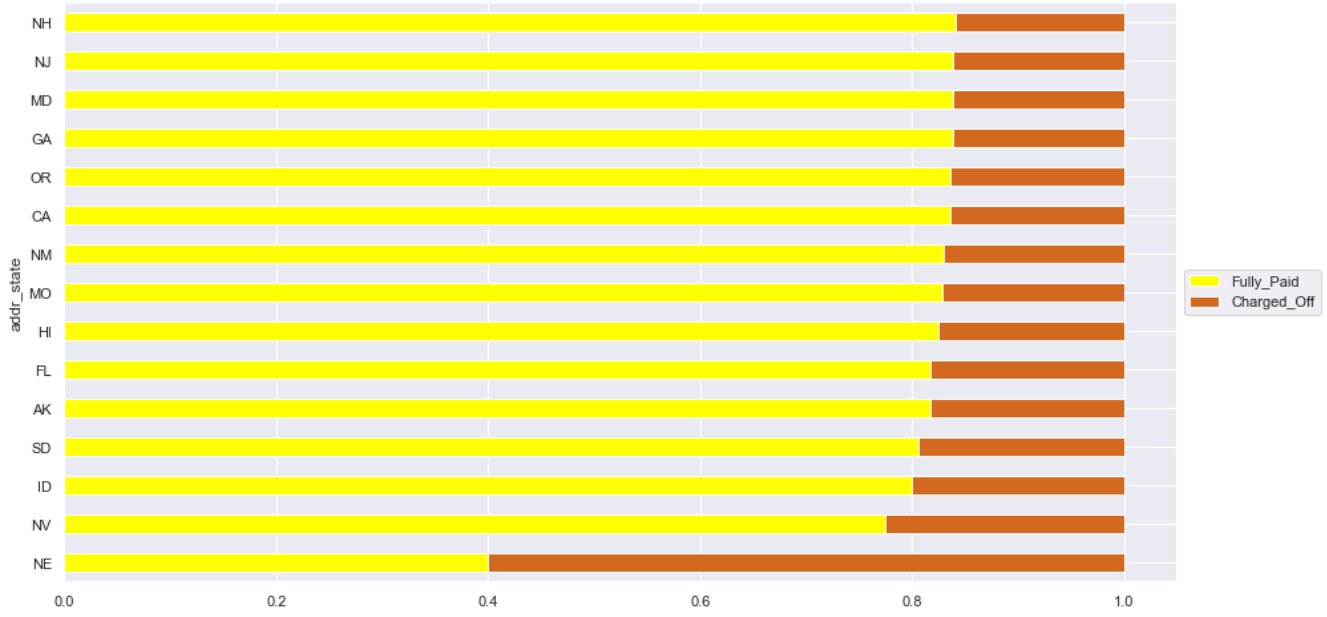
# Univariate Analysis of Categorical Variables

## Loan Purpose




- It can be seen that the borrowers who are taking loan for small businesses and renewable energies have highest default rate which is about 27% and 18% respectively.

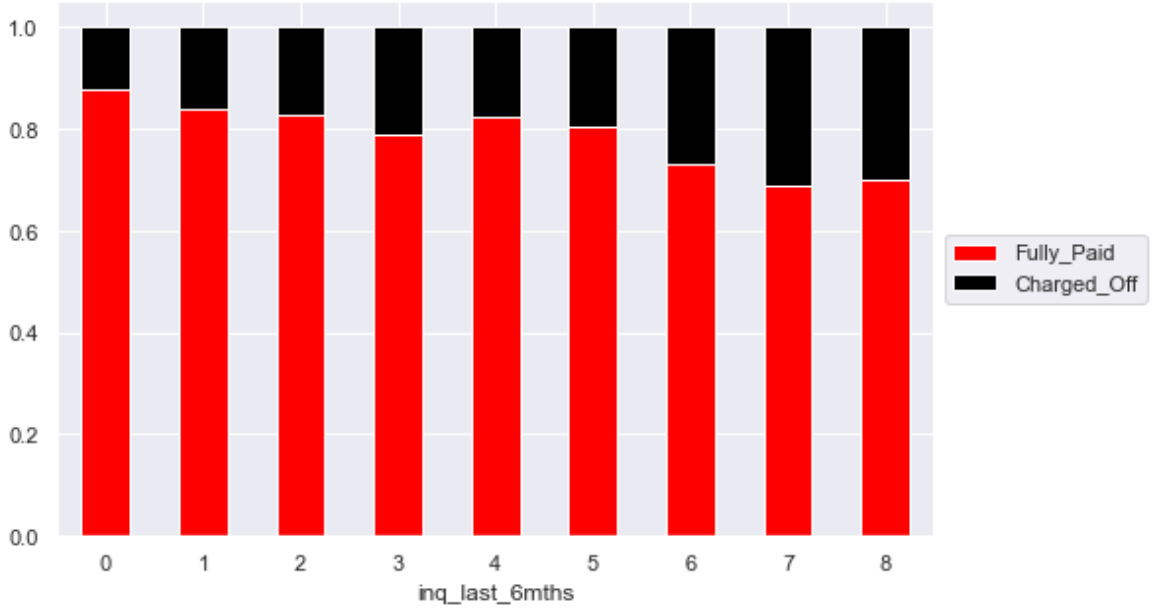
## State




- As can be seen the most risky state is **Nebraska** which accounts for 60% default rate.
- Thus, applicants belonging to this state have higher chances of default than other states

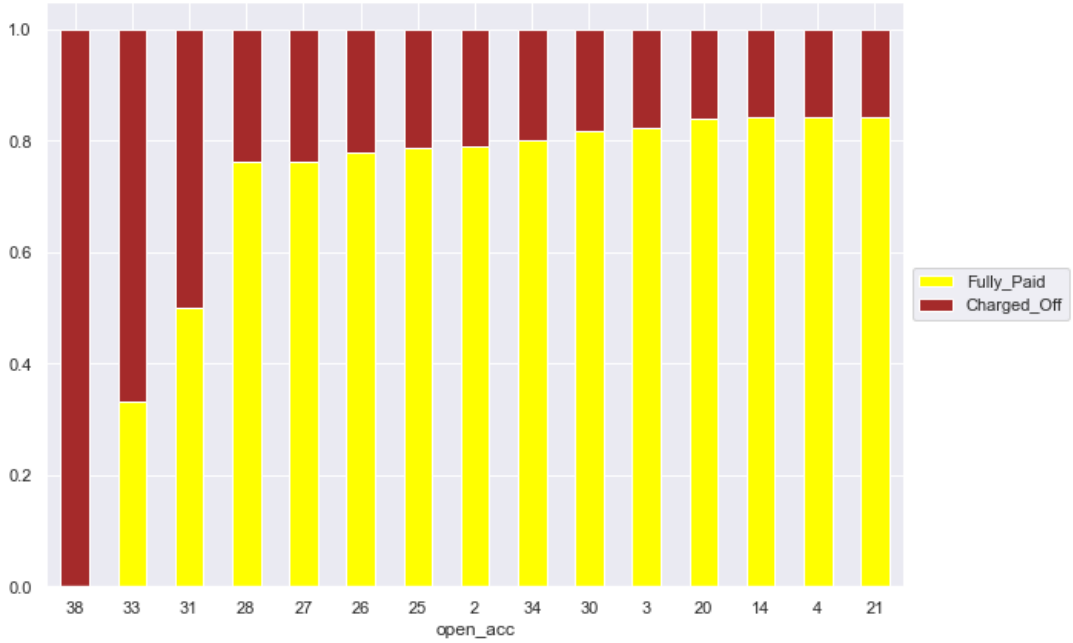
# Univariate Analysis of Categorical Variables

## Inquiries in Last 6 months




- Most borrowers which can default have 7 and 6 enquiries in the last 6 months with default rate of above 25%

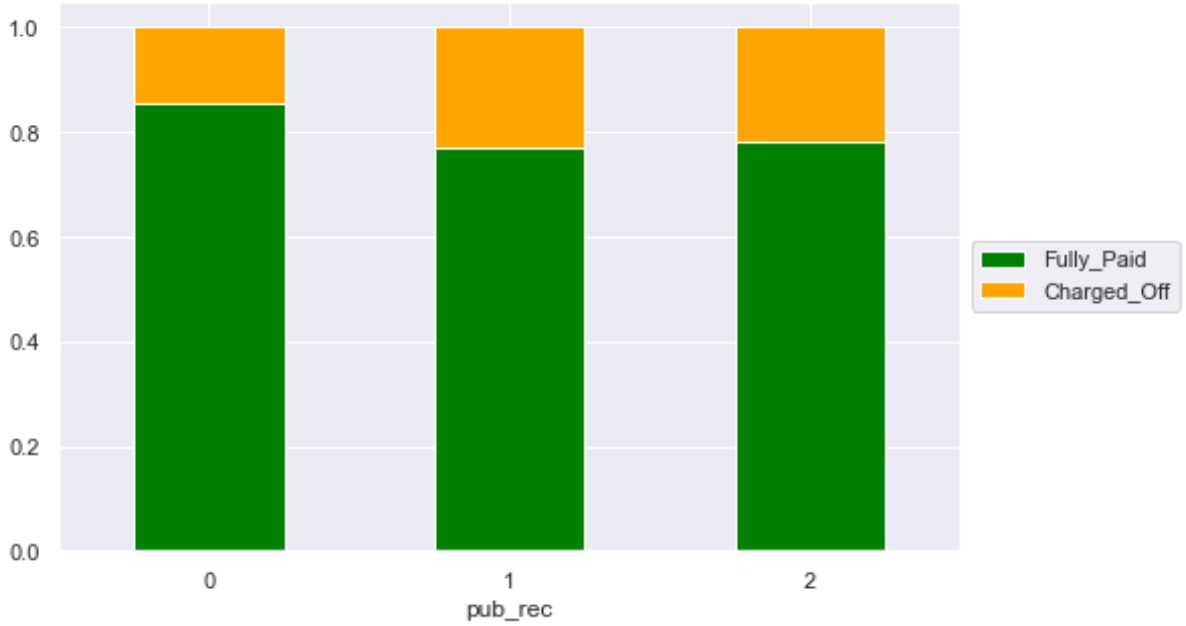
## Number of open credit lines in the borrower's credit file




- Borrowers with very large credit lines such as the borrower with 33 or 38 is least likely to pay his loan back.
- Thus, risk increases for borrowers with many credit lines open.

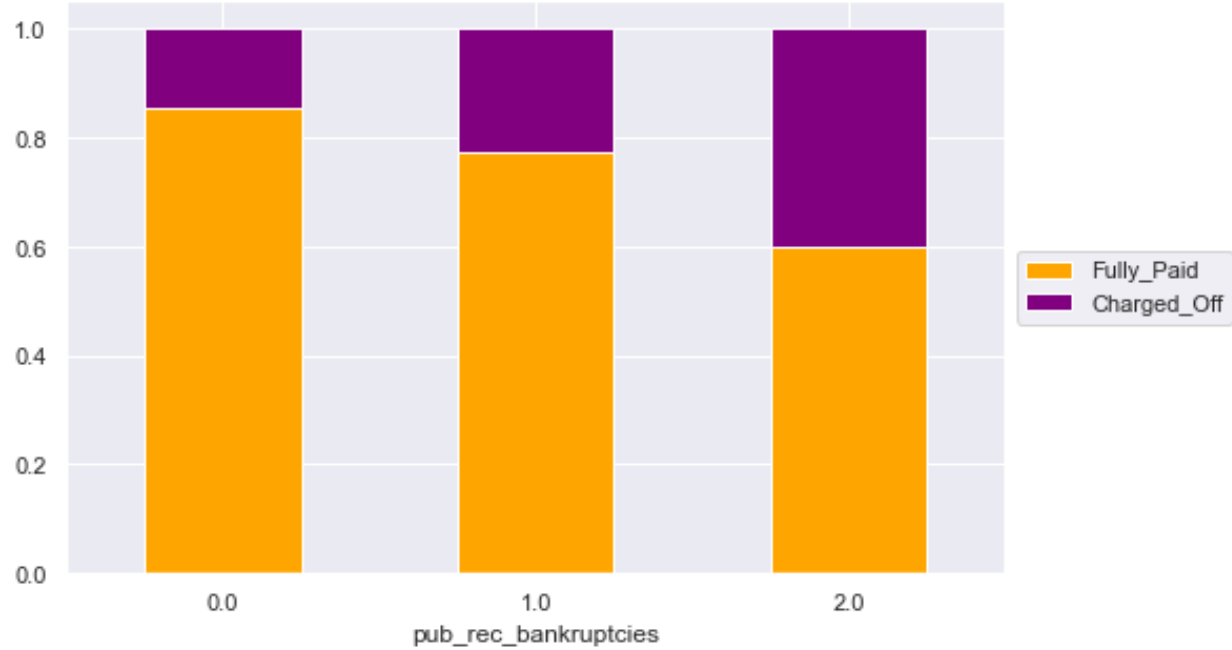
# Univariate Analysis of Categorical Variables

## Number of derogatory public records




- It can be seen that borrowers with non-zero derogatory public records are more likely to default (hence charged off).

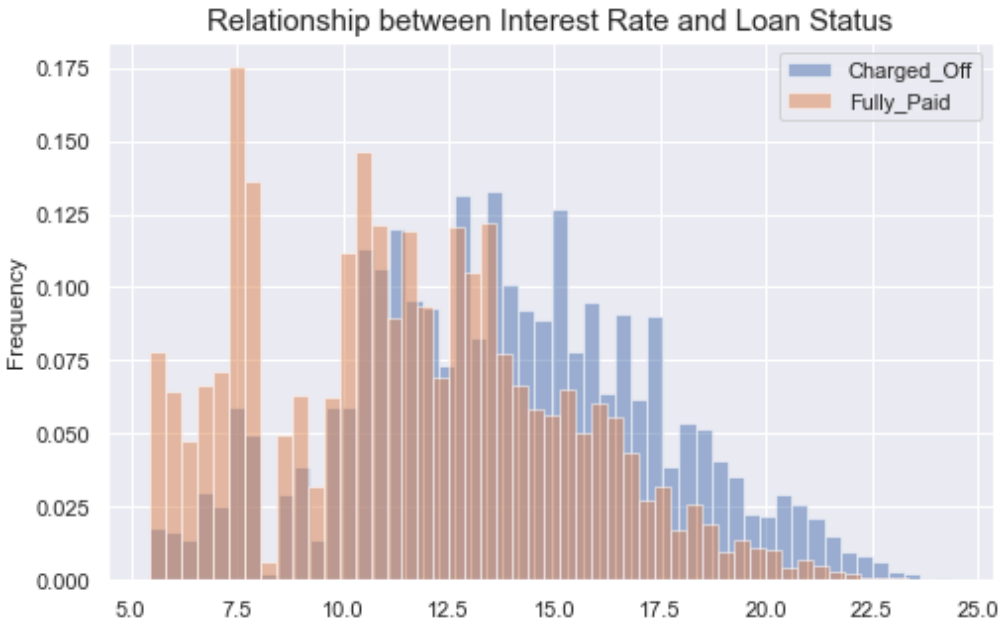
## Number of public record bankruptcies




- As can be seen more the number of public bankruptcy records, higher the chances of borrowers for defaulting the loan.

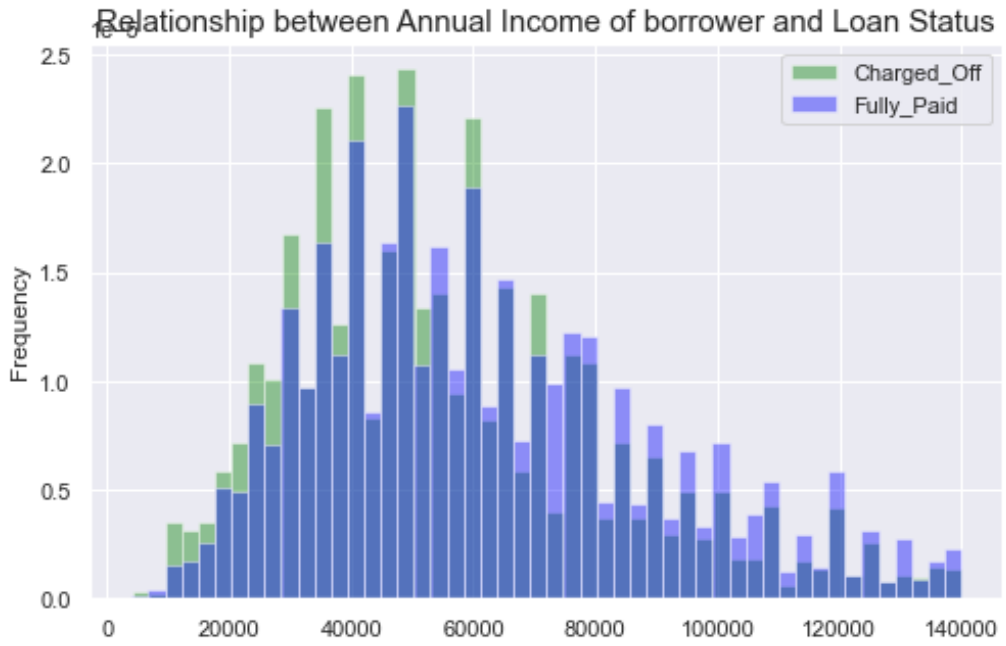
# Univariate Analysis of Numerical Variables

## Interest Rate




- Comparing for Fully Paid and Charged Off loan applicants, it can be seen that when the interest rate is greater than or equal to 11.5%, the loan default rate is more than twice that when the interest rate is below 11.5%

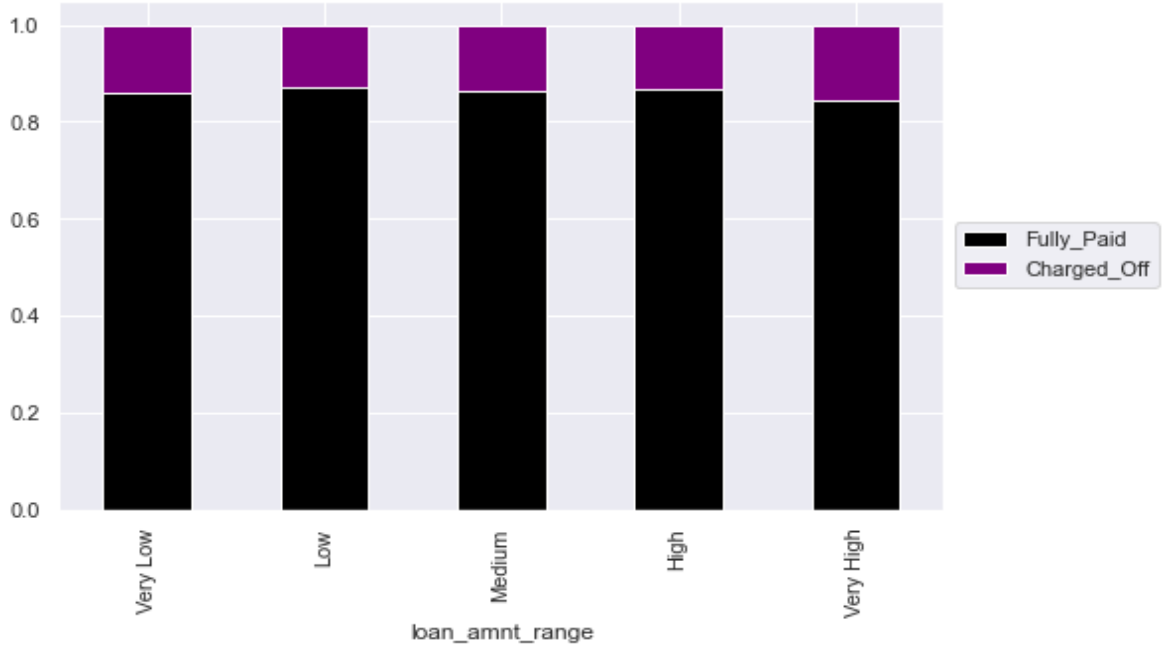
## Annual Income




- The borrowers with high annual income has lesser chances of being charged off and hence lesser chances of defaulting the loan

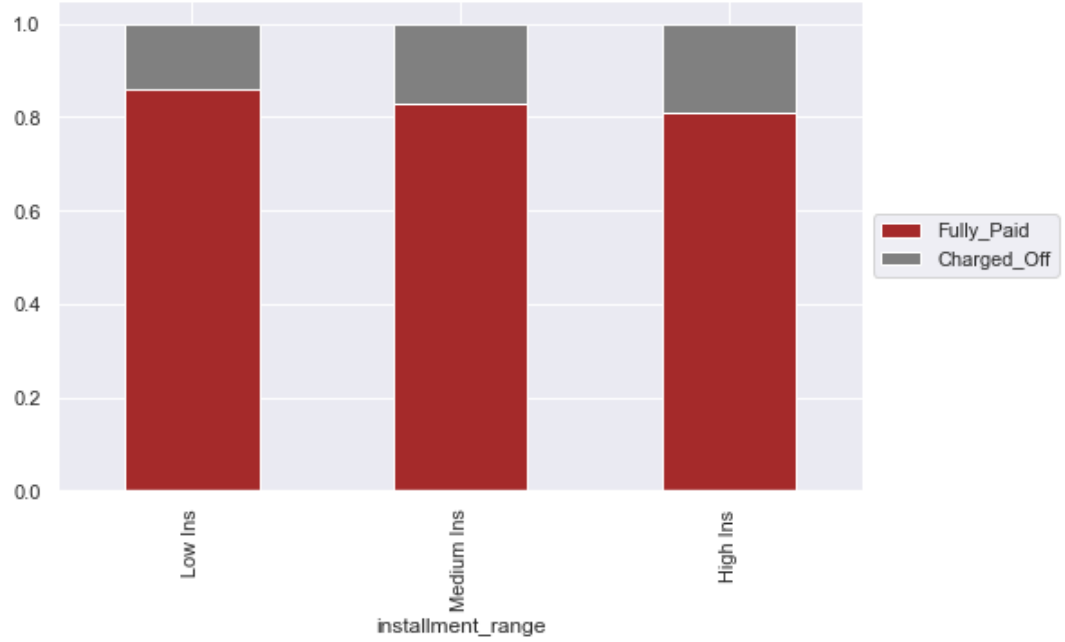
# Segmented Univariate Analysis on Numerical Variables (Using the concept of Binning)

## Loan Amount Range




- It can be seen that higher the loan amount higher are the chances of the borrower to default.

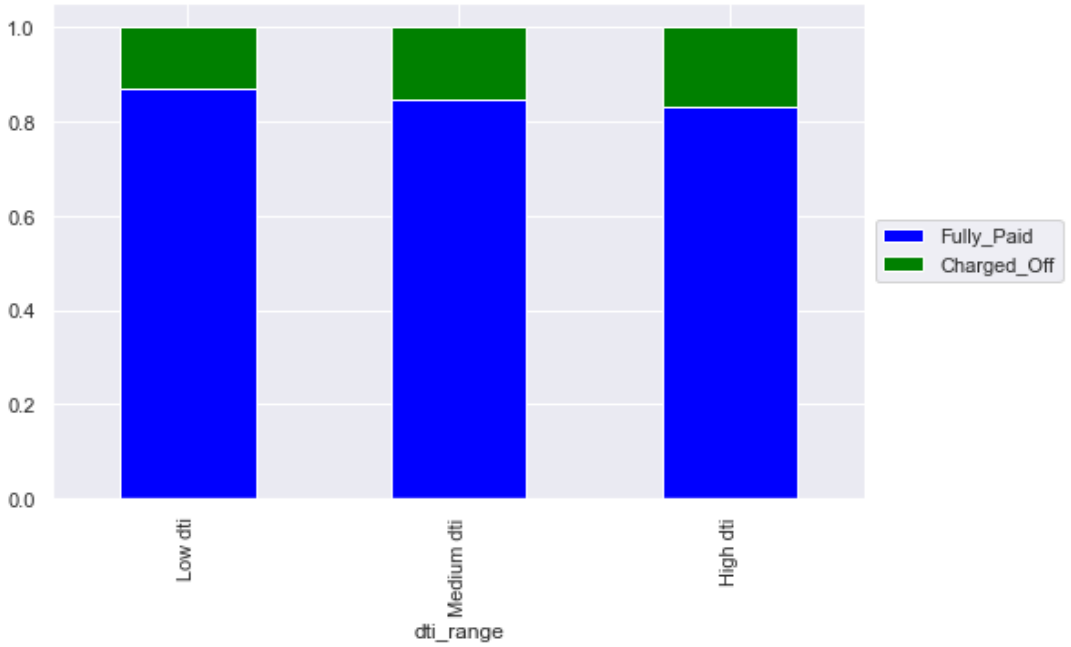
## Loan Installment Range




- It can be seen high installments increases loan default rate.

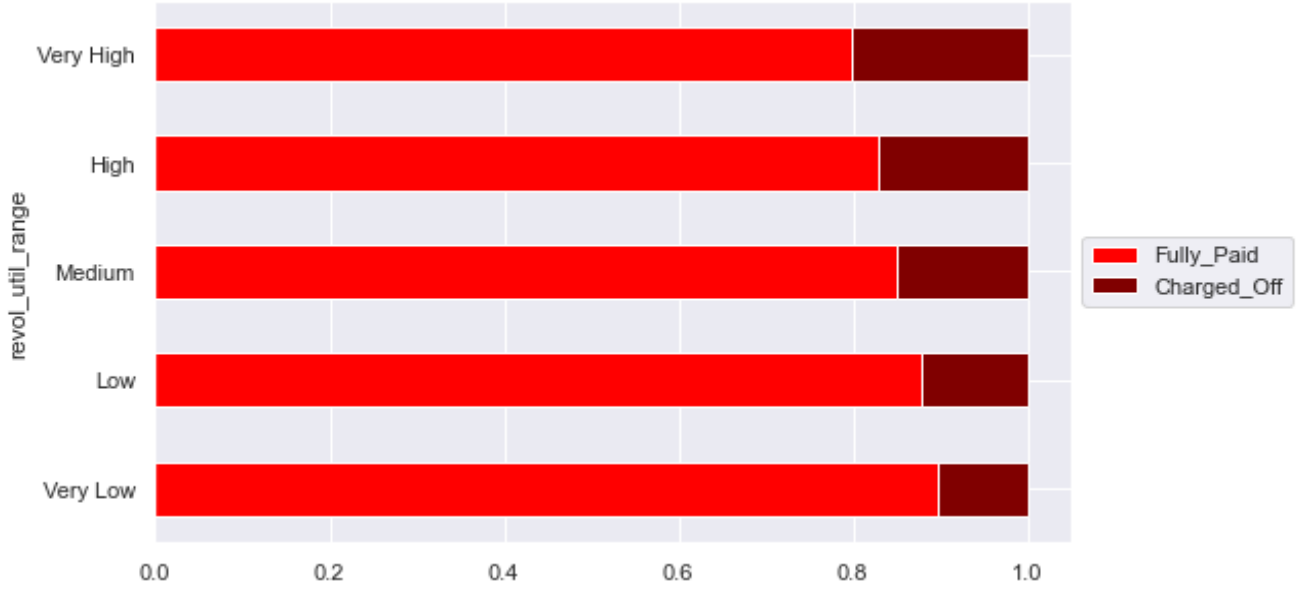
# Segmented Univariate Analysis on Numerical Variables (Using the concept of Binning)

## Debt to Income Ratio Range



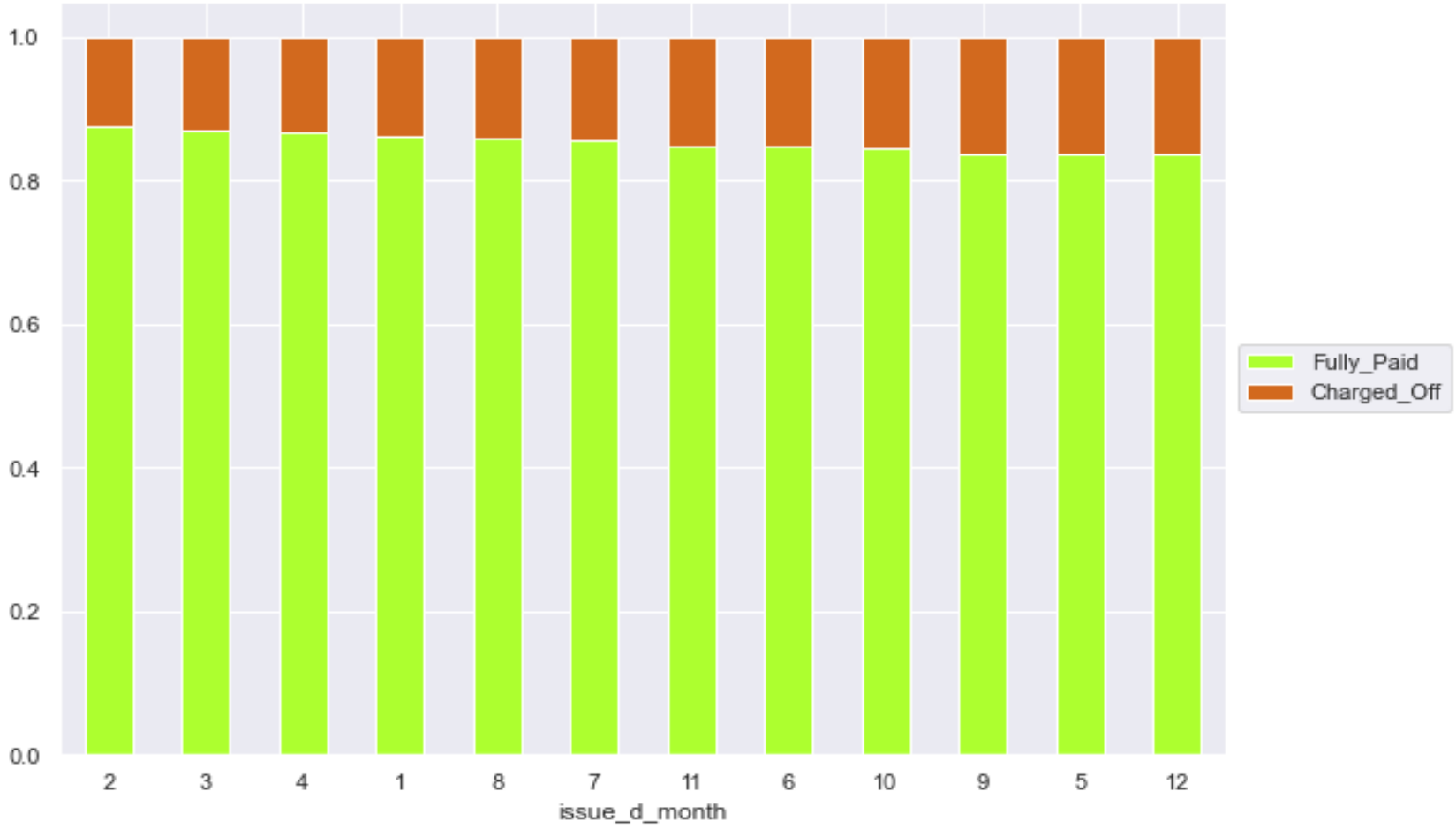

- It can be seen that larger the debt to income ratio, higher the chances of loan default.
- Lesser the debt to income ratio, better are the chances of the borrower to fully pay the loan

## Number of Revolving Credit Line



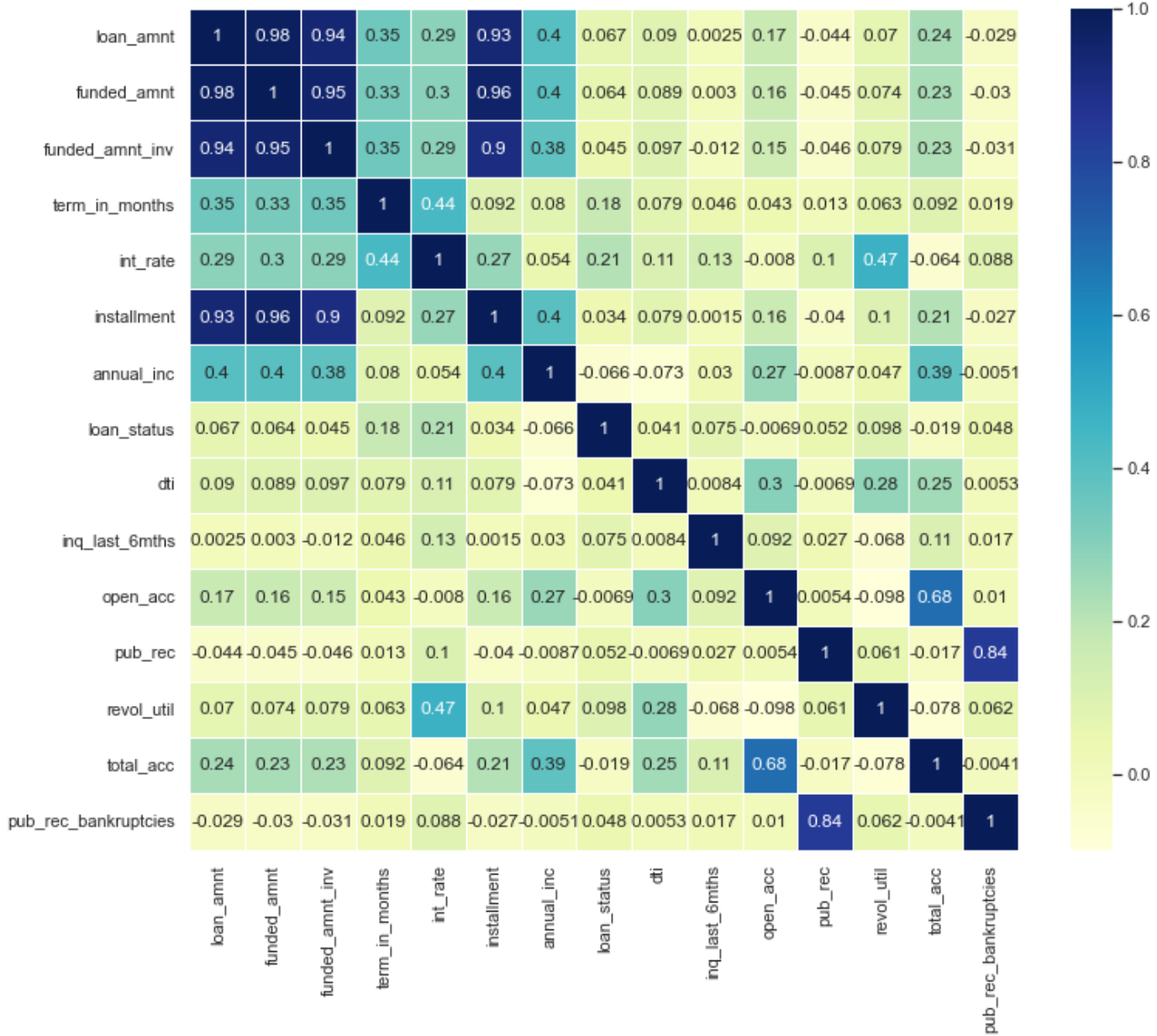

- It is clearly visible that revolving line utilization rate has a large impact on the default percentage. When this increases, the charged off percentage rises.

# Derived Metrics




- The late months (December) of an year indicated the high possibility of defaulting due to Christmas and other US festivals
- May is also another one, which is during the summer break in US where people love to travel.

# Correlation Metrics and Heat Map for all the variables



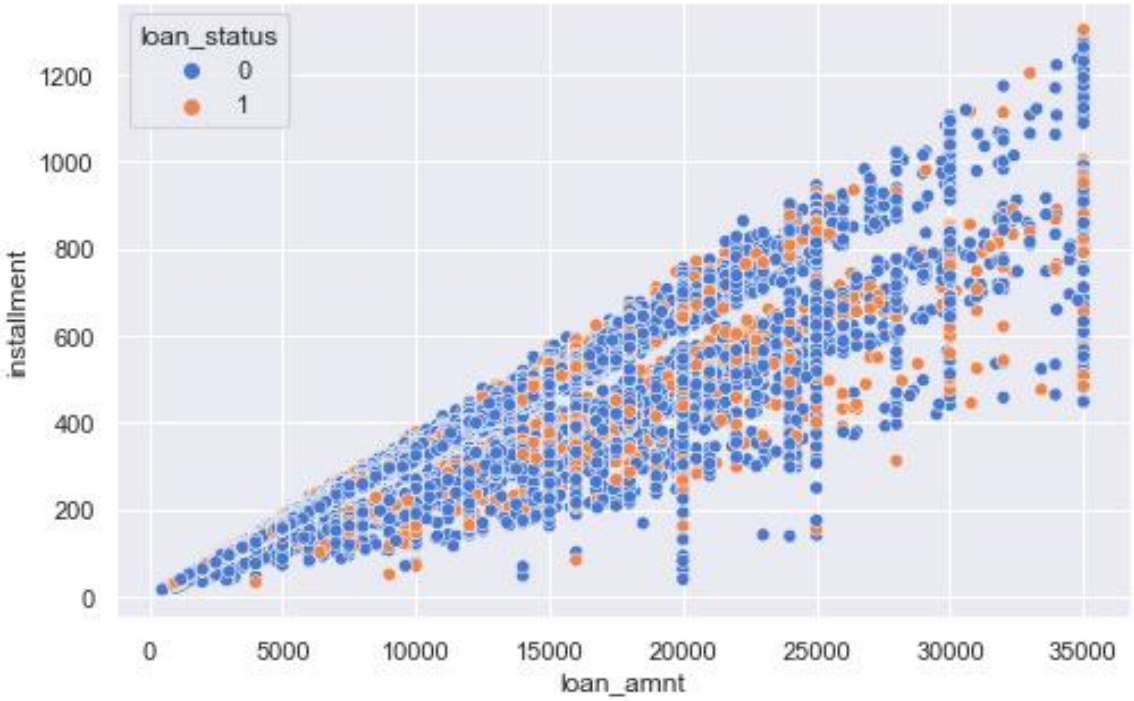

Using Correlation Metrics and Heat Map

It can be seen that there is strong relationship between:

- Loan amount and Installment
- Number of derogatory public record and number of public record bankruptcies
- Number of open credit lines in the borrower's credit files and the total number of credit lines currently in borrower's account
- Interest rate and revolving line utilization rate

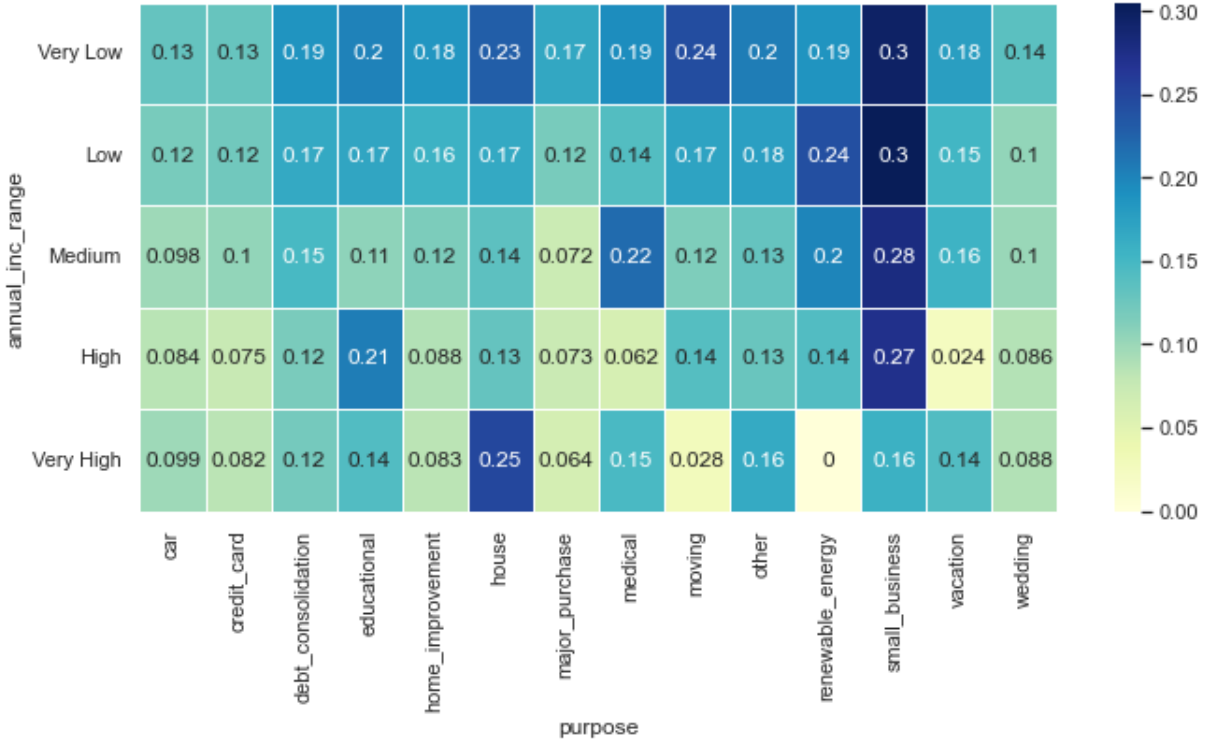
# Bivariate Analysis

## Relationship between loan amount and installment




- There is a strong relationship between loan amount and installment.
- Higher the loan amount applied for, higher will be installments for the borrower.

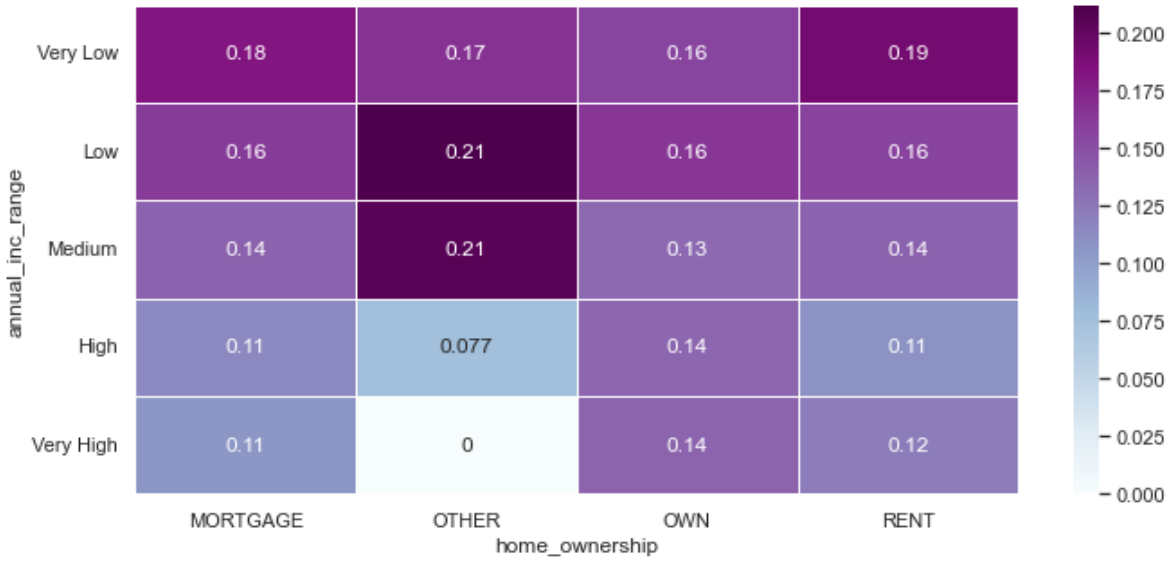
## Relationship between Annual income of applicant and purpose of taking the loan




- People with high annual incomes apply loan for the purpose of small business, house, renewable energy
- People with low income usually need loan for car, debt consolidation, educational purposes

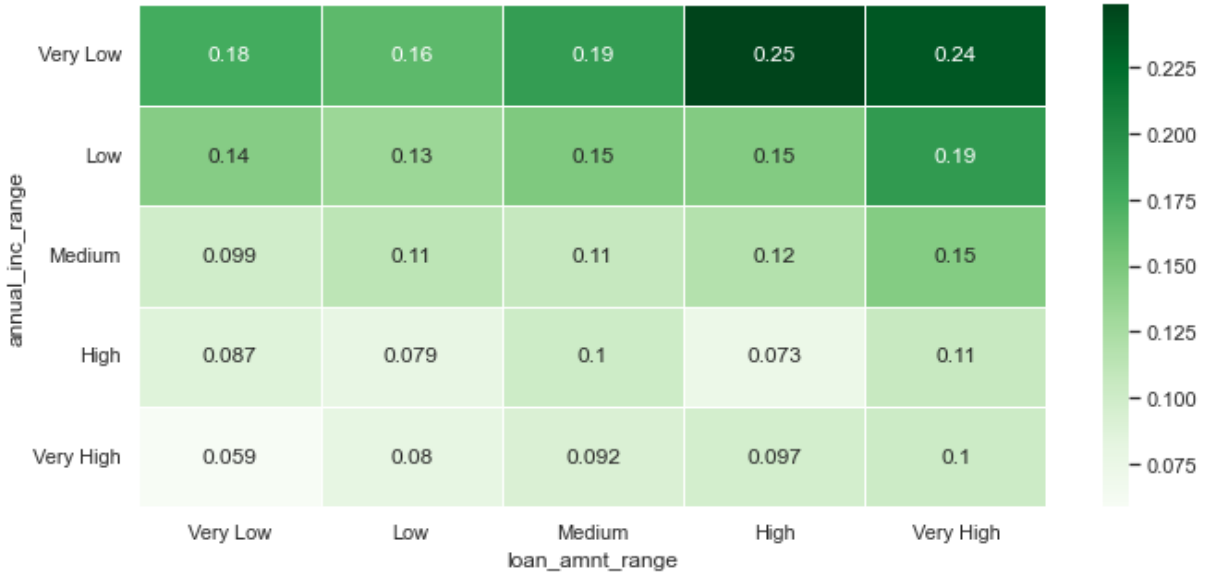
# Bivariate Analysis

## Relationship between Annual Income and Home Ownership




- As can be seen people with high annual income have home ownership as mortgage,
- Borrowers with low annual income has home ownership as rent
- And the other reasons which are not clearly mentioned also affect the default rate greatly

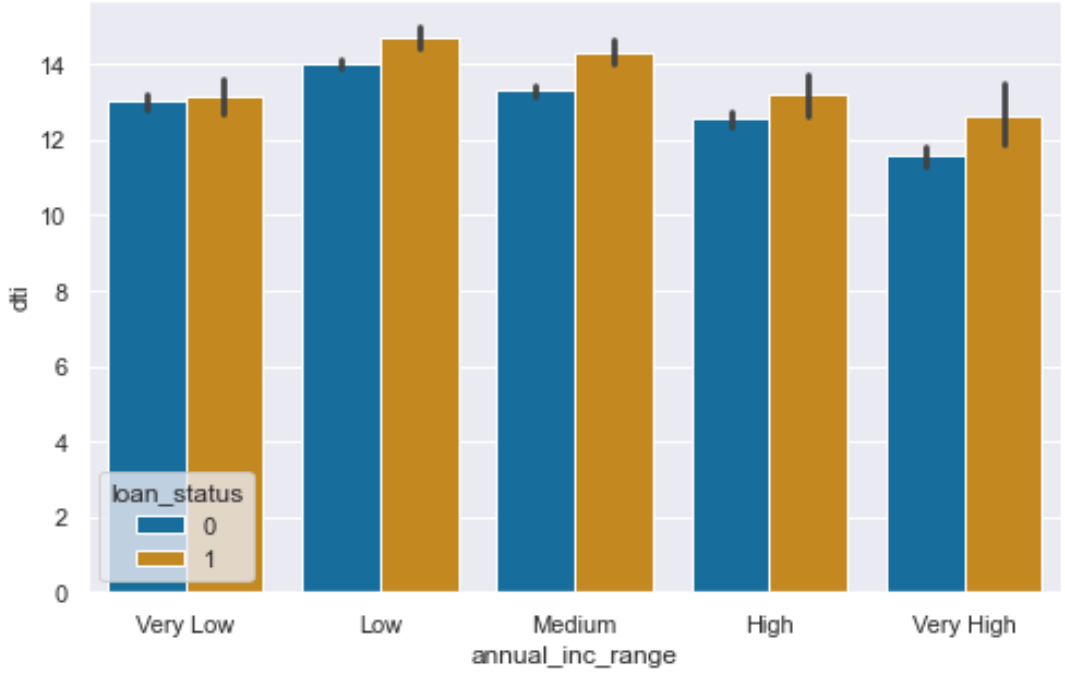
## Relationship between Annual Income and Loan Amount




- People with low annual income need high loan amount which is obvious from the fact that if someone earns less, that person will need more money(loan) from the banks

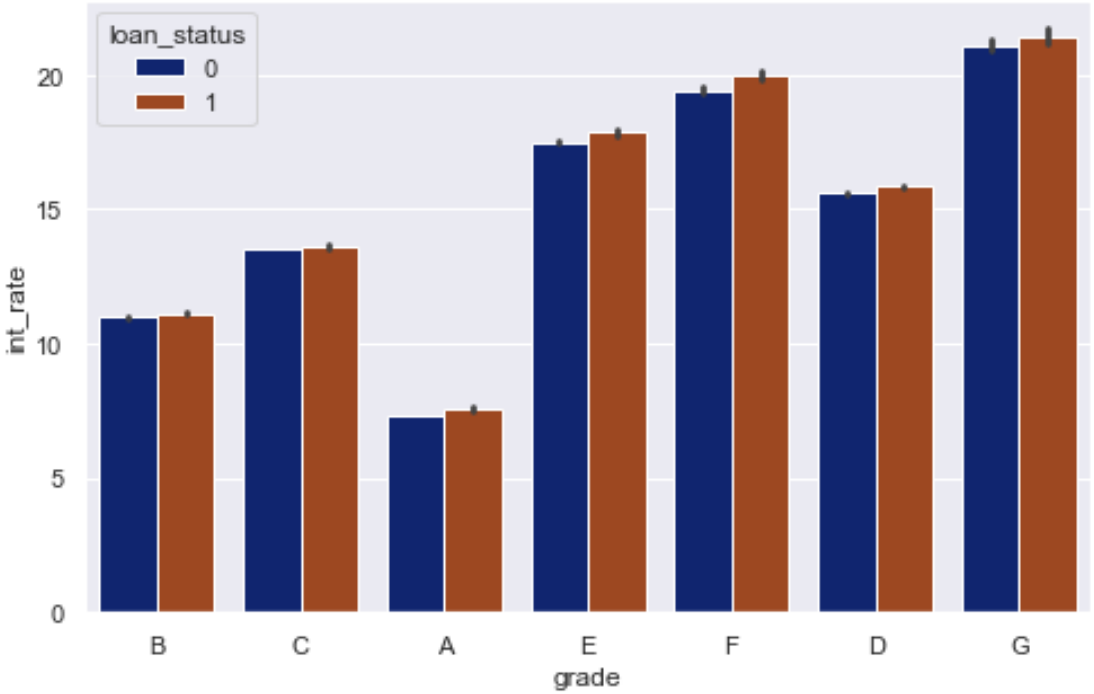
# Bivariate Analysis

## Relationship between Annual Income and Debt to Income Ratio




- People with low annual income have large debt to income ratio since the debt is more as compared to his income
- Hence, there default rate is higher

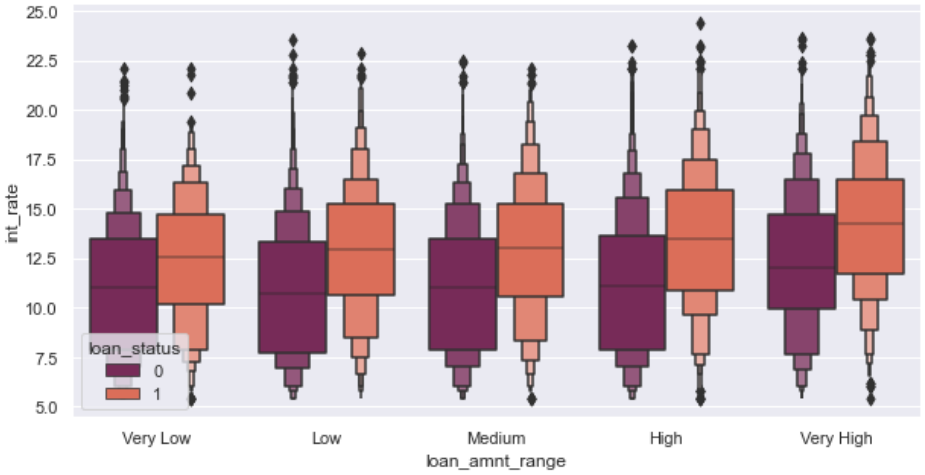
## Relationship between Annual Income and Borrower's Grade




- It can clearly be seen that G, F grade borrowers will get loan but for high interest rate and chances of default are also higher for such applicants

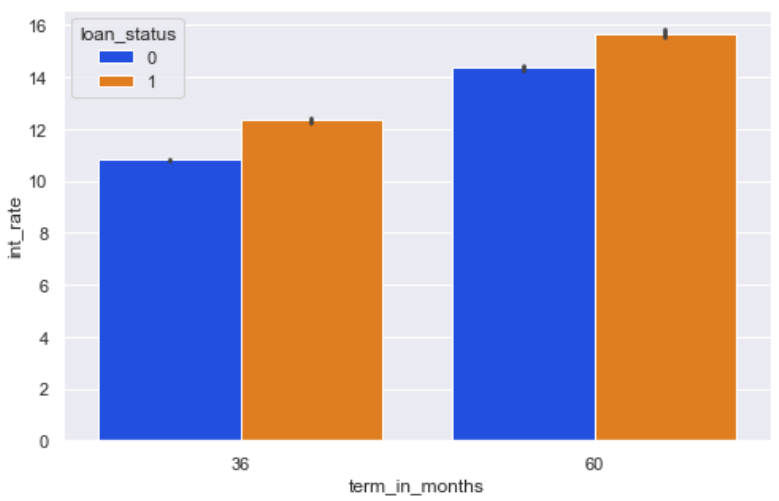
# Bivariate Analysis

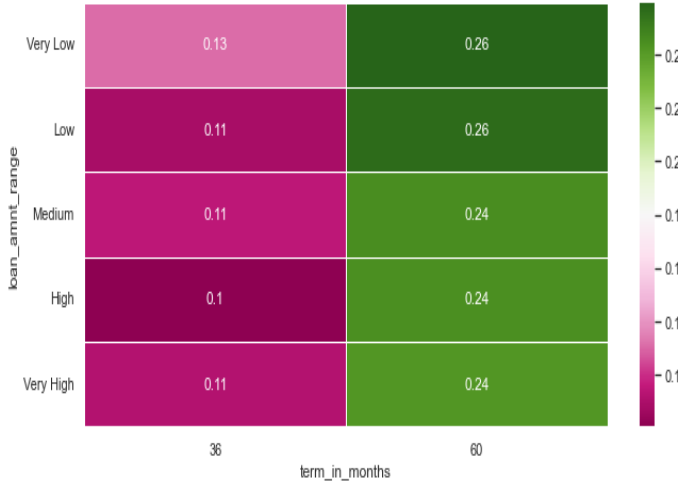
## Relationship between Loan Amount and Interest Rate




- Borrowers applying for very high loan amount will have more terms(terms in months) and thus their interest rate will increase leading to increase in default rate

## Relationship between Loan amount, Term in months and Interest Rate




- Borrowers with term of 60 months whatever loan amount he has applied for have much greater chances of default
- The reason as can be seen from the bar plot might be that as the term increases the interest rate also increases with time

# RECOMMENDATIONS

## UNIVARIATE ANALYSIS

The below analysis type of customer could be charged off due to univariate analysis:

1: The chances of Grade 'G' customer to be charged off is high 36 %

2: The chances of Sub-Grade 'F5' customer to be charged off is high 51 %

3: Most of the people with small business are having high chances to get charged off 27%

4: Most risky states from analysis is 'NE' with almost 60% get charged off

5: Most borrowers which can be charged off have 7 and 6 enquiries in the last 6 months with default rate of above 25%

6: Borrowers with very large credit lines such as '33 or 38' is most likely to get charged off

7: Borrowers with number of payments in' 60 months' are more likely to default. 'Hence are charged off'

8: Borrowers with higher interest rate than '11.5%' could be possibly highly  get charged off

9: Borrowers whose annual income is less than '70000' are most likely to takes loan and also most likely to get charged off

10: Most of people who likely to miss there loan in month of 'December' and 'may' likely to get charged off

# RECOMMENDATIONS

## BIVARIATE ANALYSIS

1. Applicants that apply for larger loan amount will have larger installments.

2. More chances of loan default are for the applicants with one or more public derogatory and public bankruptcies.

3. Applicants with very low to low annual income 3k-50k takes loan for buying a car or for educational purposes while applicants with annual income as high as 100k-160k apply loan for buying or building a house.

4. Applicants who have taken a loan for small business and the loan amount is greater than 10k.

5. Applicants with low annual income of around 10k-50k need high loan amount 12k-40k which is obvious from the fact that if someone earns less, that person will need more money(loan) from the banks.

6. Applicants with low annual income have high debt to income ratio and thus their default rate is higher.

7. Applicants with grades of F or G have higher default rates and they are given loan for higher interest rates of around 20%.

8. Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of more that 20%.

9. Applicants with terms of 60 months are more likely to default since the interest rate increases for such applicants.