# Lending Club Case Study

Exploratory Data Analysis (EDA)

BY:

Deepen Kumar Sahoo

Sudhir Singh

# Contents:

- Problem Statement
- Brief Data Summary
- Data Cleaning
- Data Conversion
- Treating Missing Values
- Derived Metrics
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Conclusions

# Problem Statement

- You work for a consumer finance company which is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures etc. to urban customers.
- Like most other Lending companies, approving loans of risky applicants (who are likely to default the loan) can lead to credit loss (financial loss) for the company.
- The aim of the case study is to do EDA on the data set provided and find out the driving factors (driver variables) which are strong indicators of loan default. The company can utilize the understanding (insights) for portfolio and risk assessment.

# Brief Data Summary

- There are 39717 rows and 111 columns in the data.
- Out of 111 columns, 87 are Numerical and 24 are Categorical.

# Data Cleaning

- There were no header/footer or summary rows (e.g. Total or Sub-total rows)
- There were 1140 rows where loan_status was ‘Current’; those rows were removed as they wouldnot contribute to the EDA analysis

- There were no duplicate rows
- There were 55 columns in the data having 100% Null values and hence they were dropped
- As per industry practice, we dropped 3 more fields namely desc, mths_since_last_delinq, mths_since_last_record as they were having more than 30% Null values.
- There were no such rows with all the column values as NULL
- There were 11 columns who had only single unique value (throughout entire column) so dropped them as they wont help in analysis
- Dropped the fields 'member_id' and 'url' as they had 100% unique values.
- Behavioural columns are populated after the approval of any loan application and hence their data will not be available during the loan approval process. So dropped those 22 behavioural columns.
- Dropped the fields 'title' and 'emp_title', which are having so many unique text values
- Restricted our analysis till Group level as far as the 'grading of loans' is concerned. Hence, we will kept the 'grade' column only and removed the 'sub_grade' column.

# Data Conversion

- Trimmed additional string value from ‘term’ column and converted the field from object to integer datatype
- Trimmed the ‘%’ symbol from ‘int_rate’ and converted it to integer datatype from object.
- Converted the datatype of ‘funded_amnt’ and ‘loan_amnt’ from float to int.
- Converted the ‘issue_d’ column from object to datetime format.

# Treating Missing Values

- ‘emp_length’ and ‘pub_rec_bankruptcies’ columns were having 2.67% and 1.80% of rows as null respectively. So, preferred to drop those rows as the percentage of Null values were less (less than 5%, which is a threshold as per industry practice).

# Derived Metrics

- Derived  ‘issue_year’ and ‘issue_month ’ columns from ‘issue_d’ columns which were used in further analysis.

# Outlier Treatment

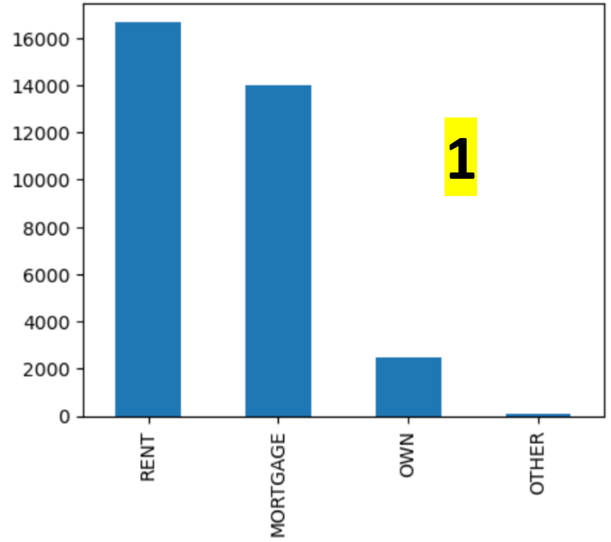
- Outliers existed for these numerical fields: 'loan_amnt', 'funded_amnt', ‘funded_amnt_inv', 'int_rate', 'installment' and ‘annual_inc'.
- Outlier treatment was done for the above fields using the Inter-Quartile Range (IQR) theory
- Observed significant changes in the box plots of above fields, before and after the outlier removal process (which constitutes the Univariate Analysis of Numerical Columns).

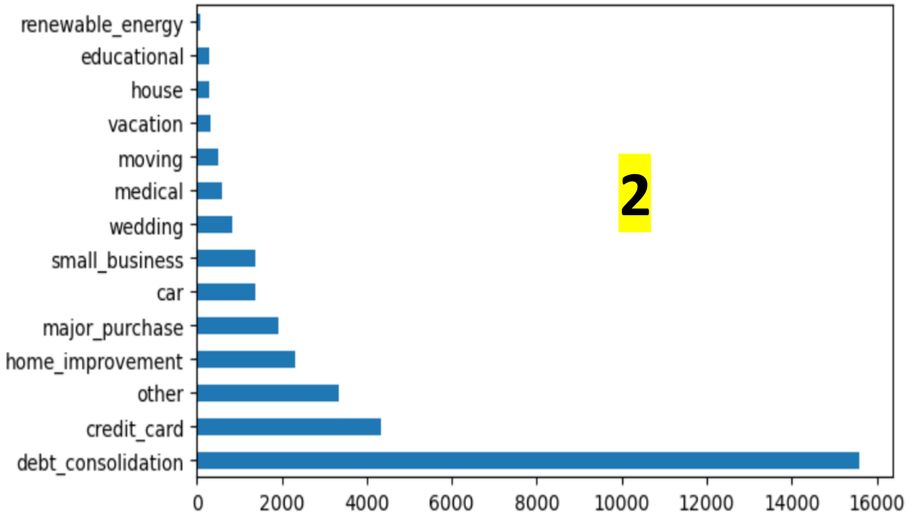
# Univariate Analysis

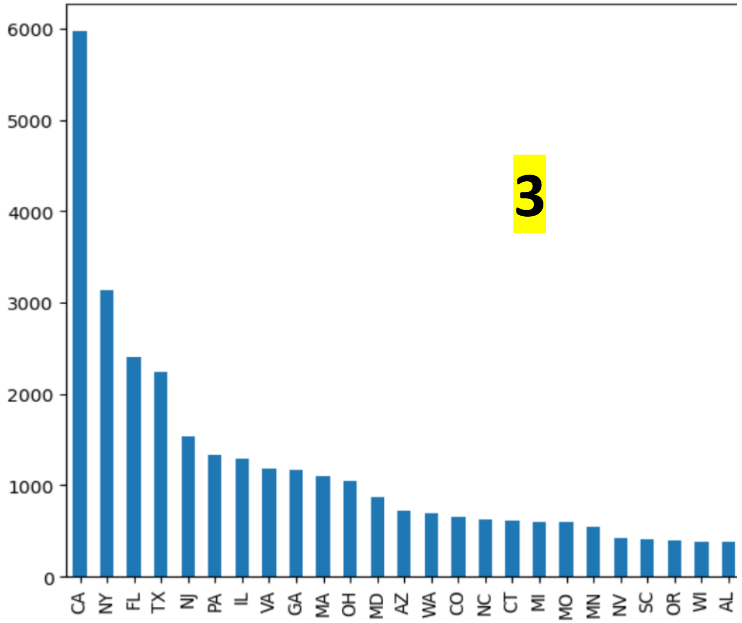
## *On Unordered & Ordered Categorical Fields*

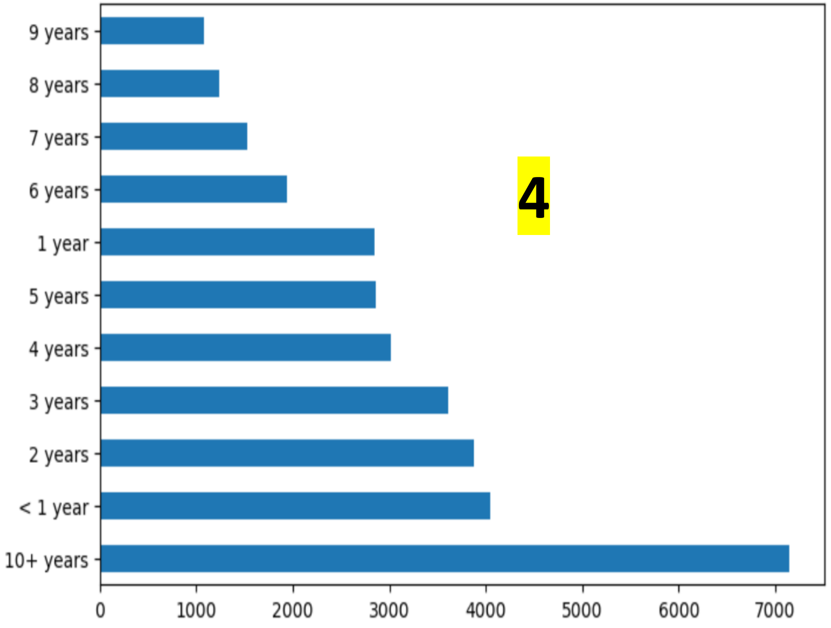
Observations:

1. Majority of loan applicants are either living on Rent or on Mortgage.
2. Most of the loan applicants have applied for loans for debt consolidation purposes.
3. Most of the Loan applicants are from State 'CA'.
4. Most of the applicants are having 10+ years of Experience.





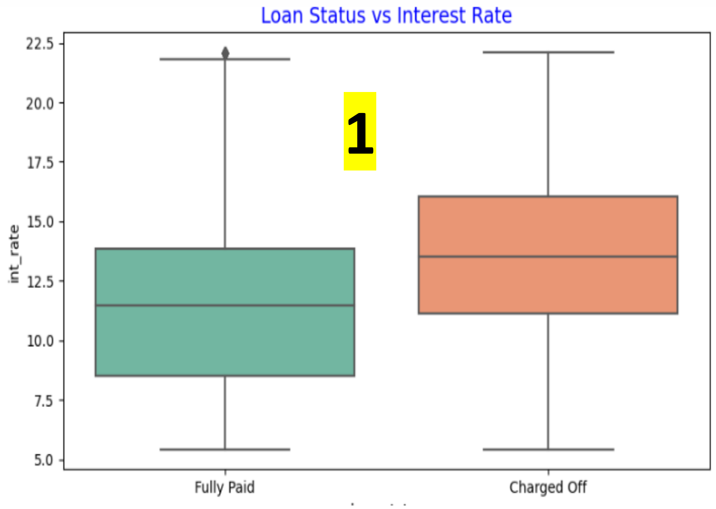


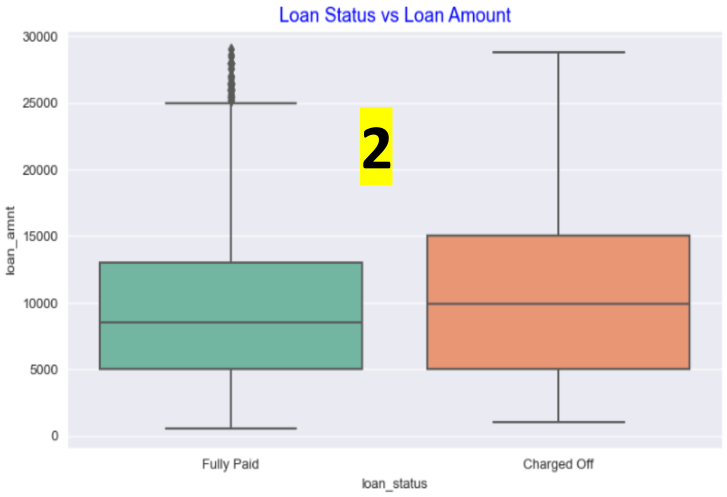


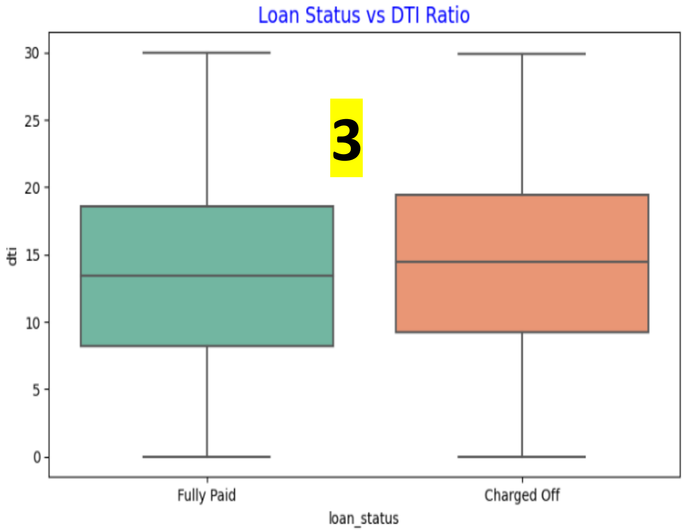
# Segmented Univariate Analysis

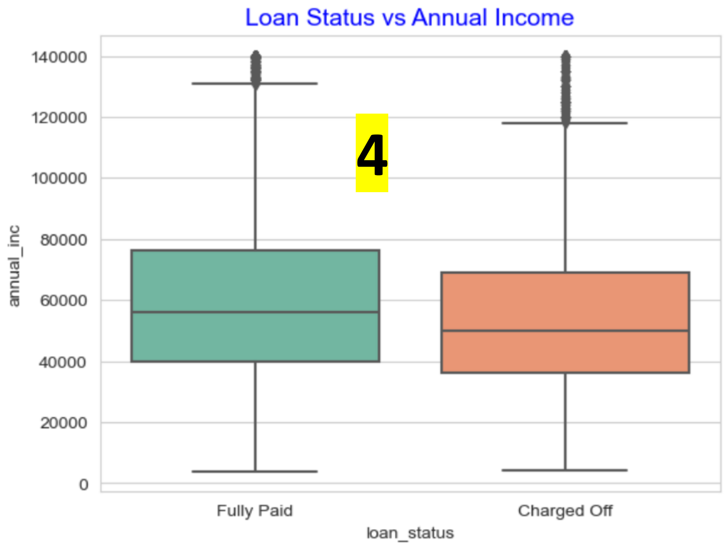
Observations:

1. More is the interest rate, more are the cases of charged-off.
2. Loan applications having higher loan amount request defaulted more.
3. Customers with higher dti have more chance of default.
4. Borrowers who are having lower annual income are more prone to default the loan.


Loan Status vs Interest Rate


Loan Status vs Loan Amount

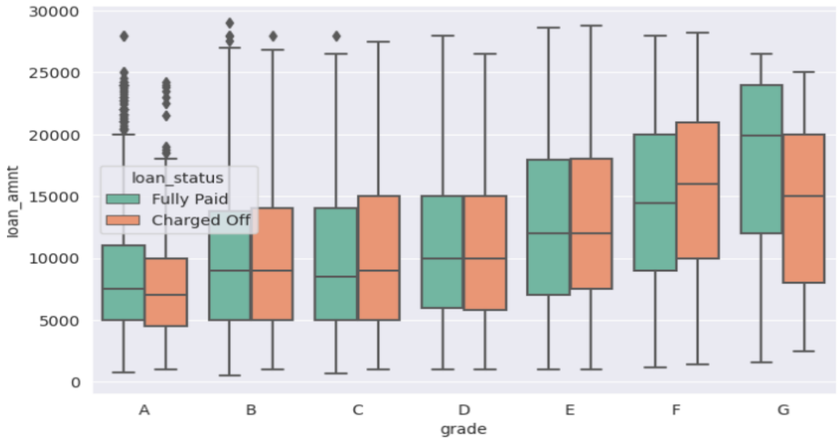

Loan Status vs DTI Ratio


Loan Status vs Annual Income

# Bivariate Analysis (Numerical Vs Categorical)

## I. 'grade' field

1

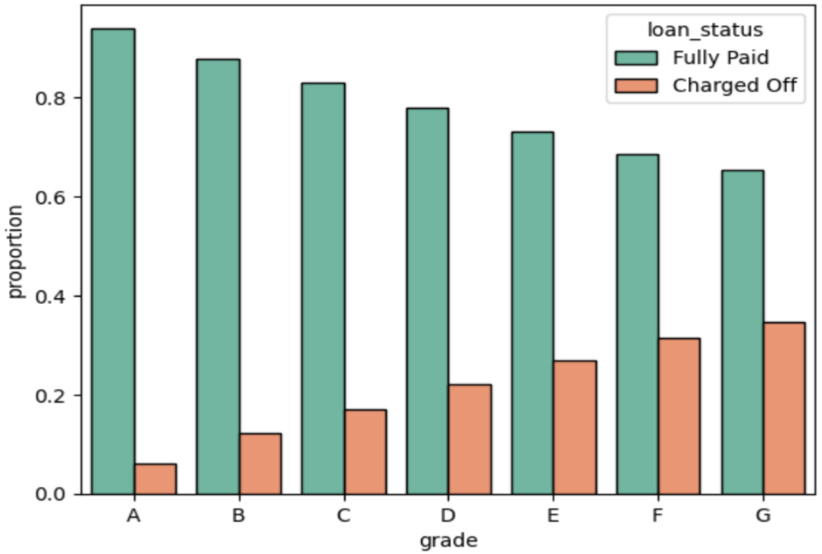


Observations:

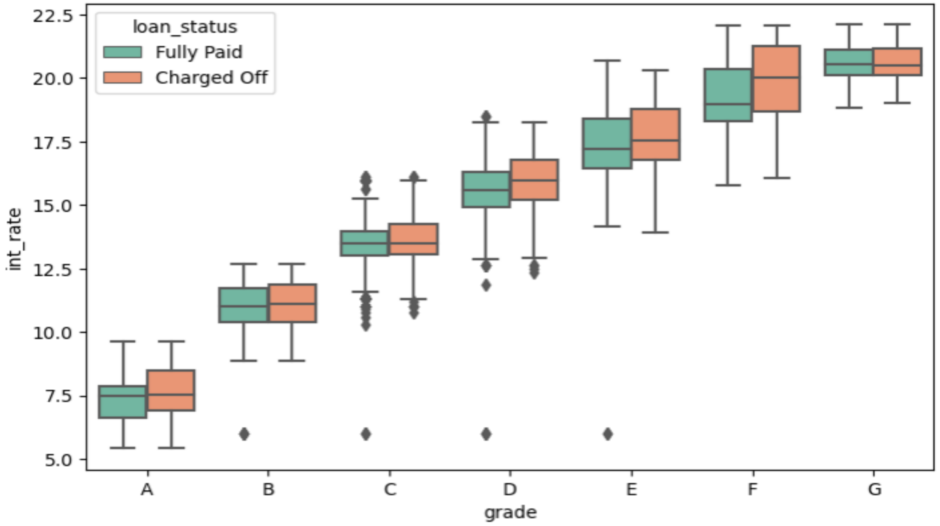
1. Customers who took lower graded loans (like 'F', 'G' etc) with higher loan amount, have more chances of being charged-off.
2. As the loan grade decreases (A to G), the proportion of loan default increases (A to G).
3. With decrease in loan grade, the loans suffer a gradual increase in the interest rate, and the borrowers become more likely to default the loan.

2

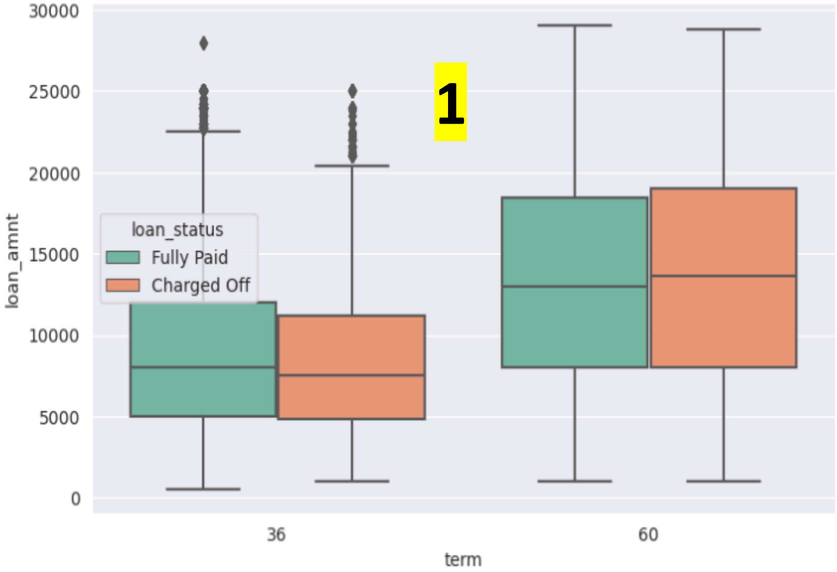
Grade vs Loan status

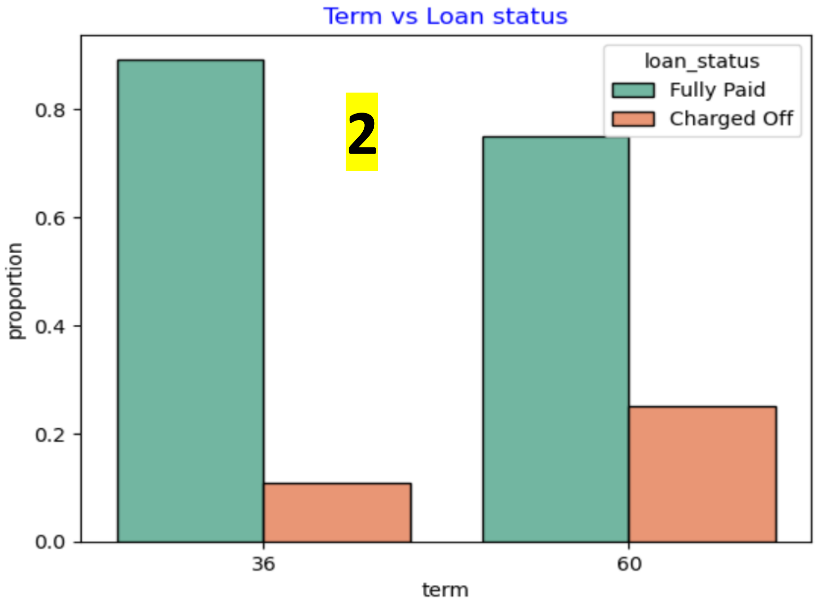


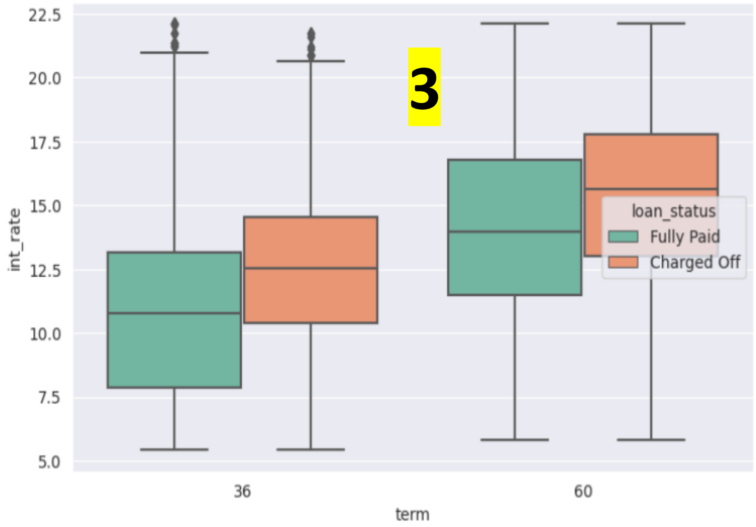
3

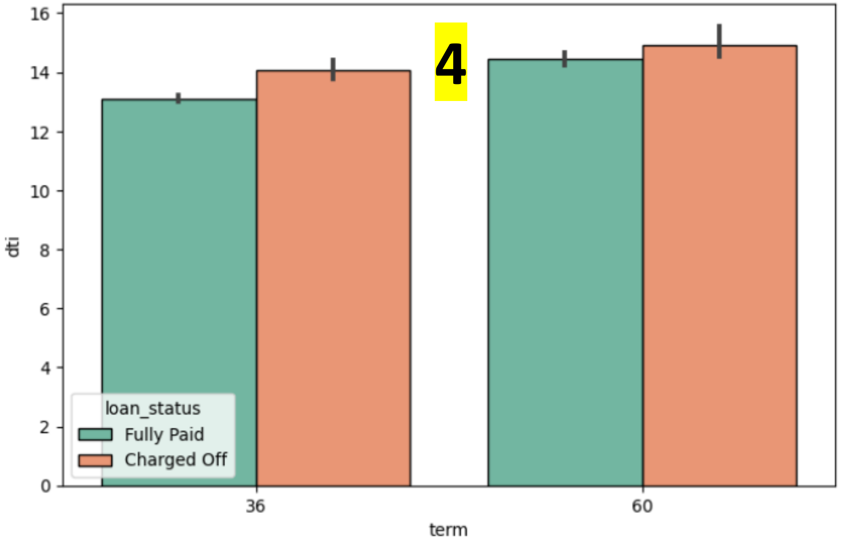


# II. ‘term’ field





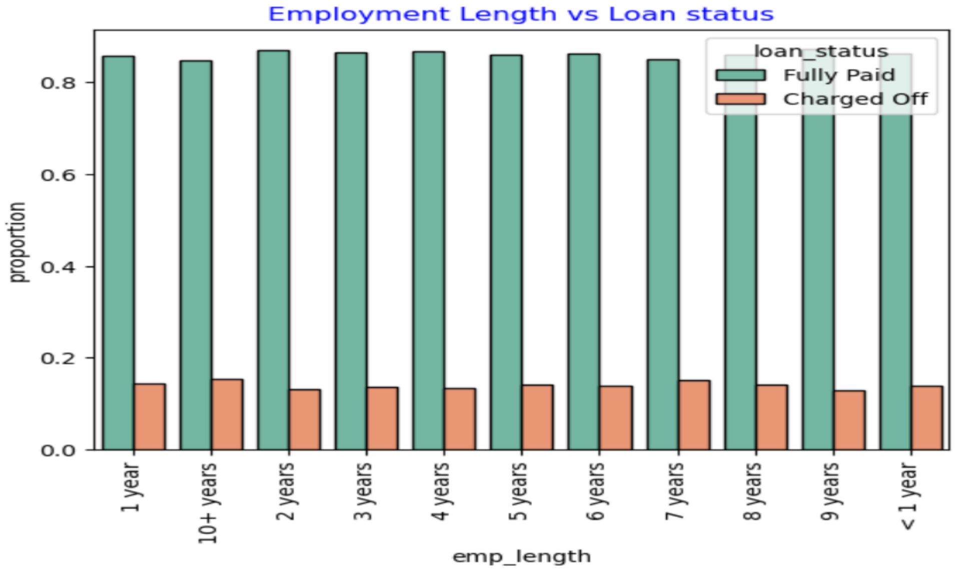




## Observations:

1. Looks like, Loan amount is not a decider for defaults in both the sub-categories of 'term'.
2. More proportion of '60 months tenured' loans defaulted as compared to proportion of ‘36 months tenured’ loans
3. High interest loans of ‘60 months’ tenure are more prone to being charged-off.
4. With rise in DTI, charged-off cases are comparatively higher than fully-paid cases, for both 36 months & 60 months tenured loans.

## III. ‘emp_length’ field

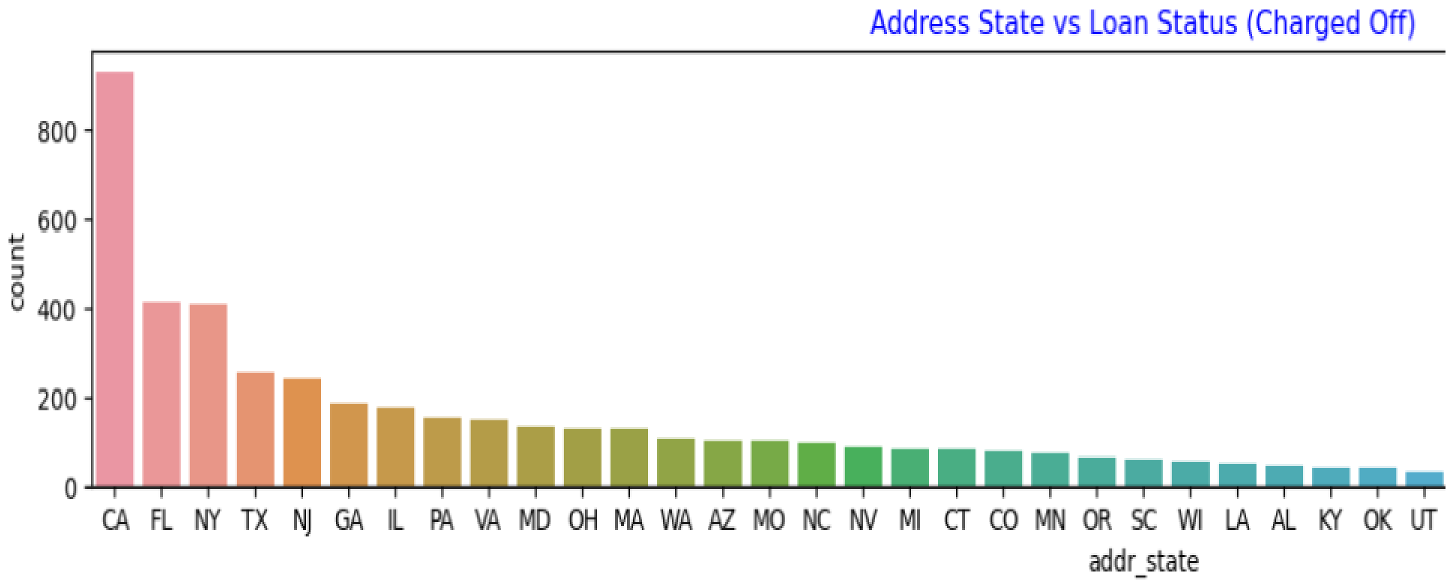


1

Observation:

1. Employment length is not a decider for loan defaults.

## IV. ‘addr_state’ field



2

Observation:

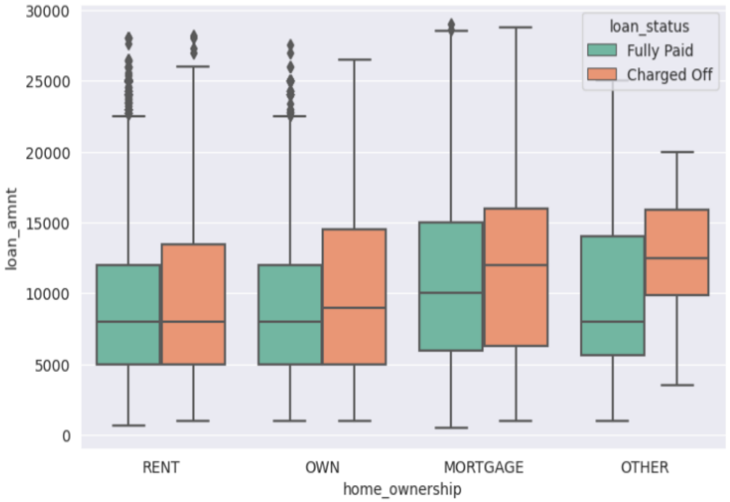
2. Majority of applicants who defaulted the loan belong to states 'CA','FL' and 'NY'.

# V. ‘home_ownership’ field

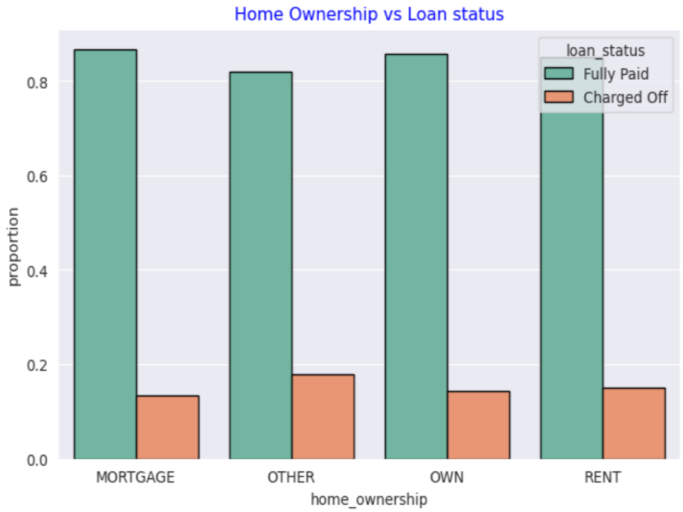
Observations:

1. Borrowers from ‘OTHER’ and ‘MORTGAGE’ home ownership status, taking higher loan amount have defaulted more.
2. There is slightly high proportion of defaults for 'OTHER' sub-category.
3. Irrespective of Home ownership status, when the rate of interest is high, the charged-off rate is also high.
4. Applicants from 'OTHER' home ownership sub-category have lesser dti compared to other sub-categories.

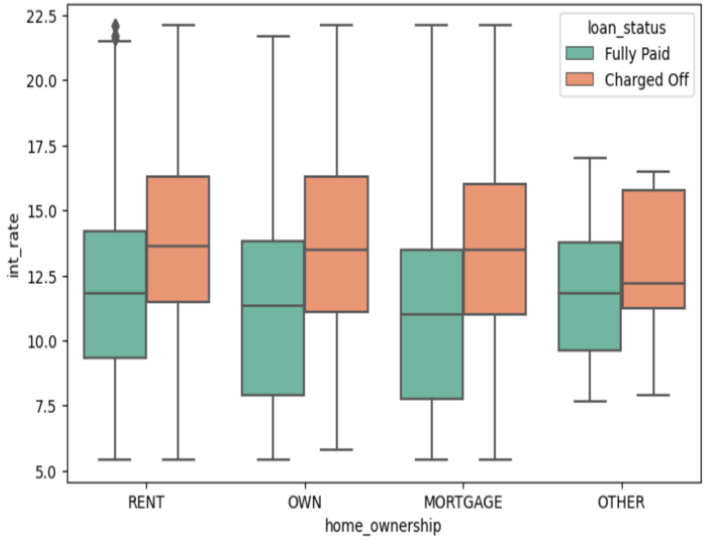
1



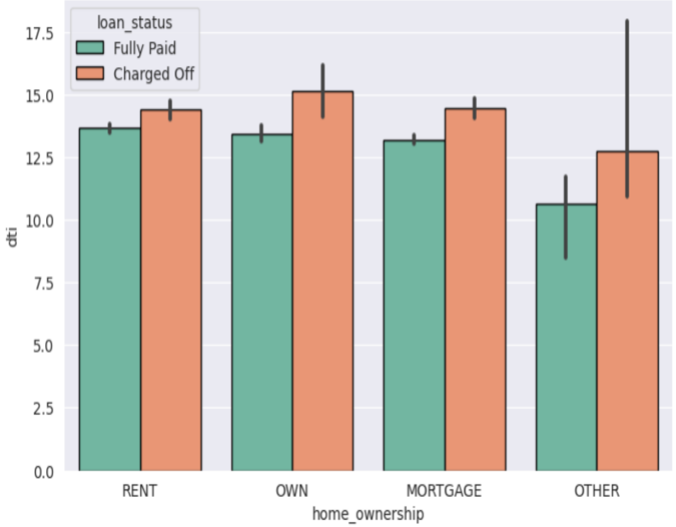
2



3



4

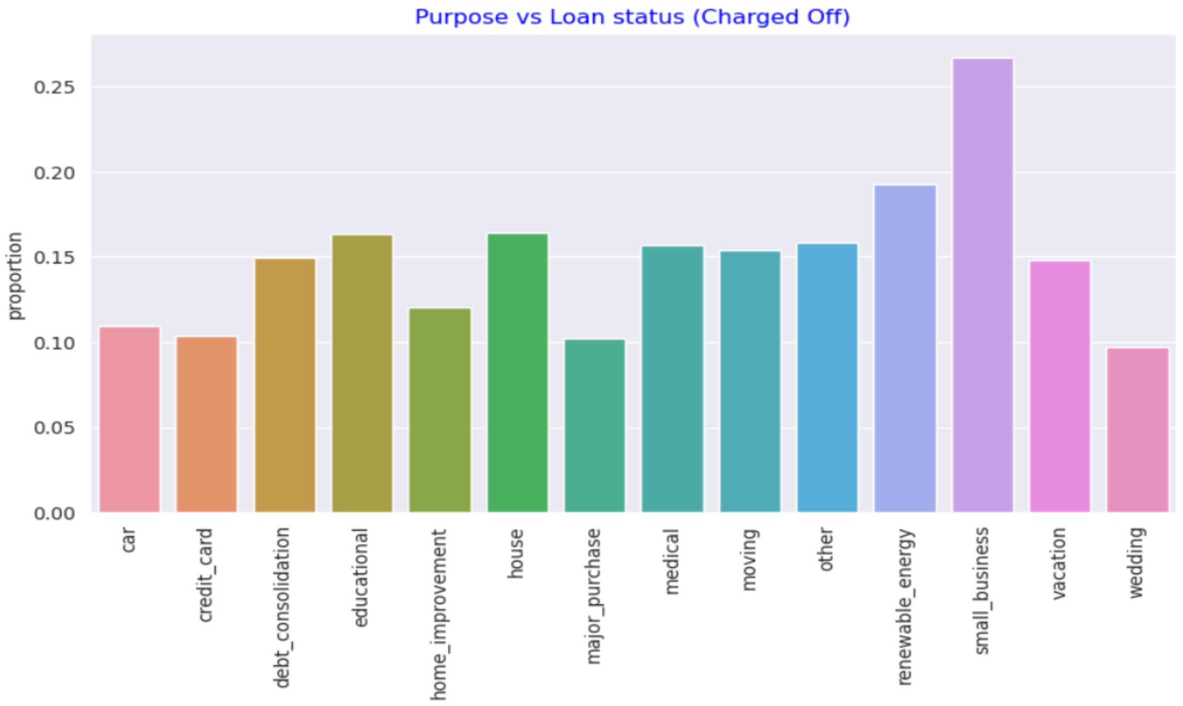


# VI. ‘purpose’ field

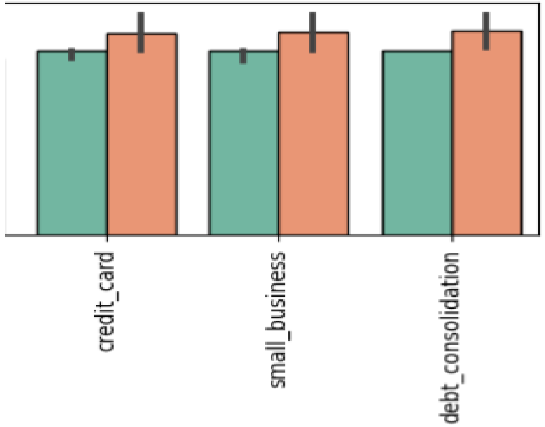
Observations:

1. Small Business purpose loans have the highest charged-off proportion.
2. Applicants requesting higher loan amounts for debt_consolidation, credit_card and small_business purposes have defaulted more.
3. Home loans with higher interest rate have defaulted the most.

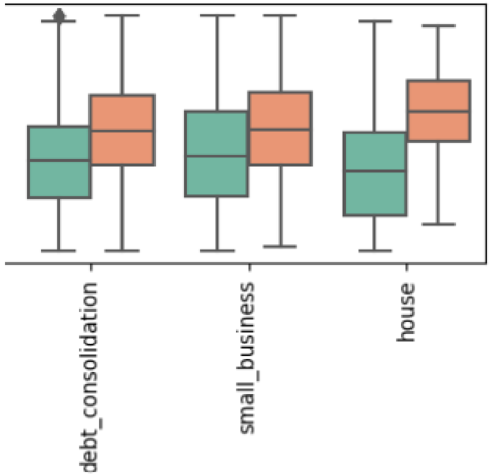
1



2

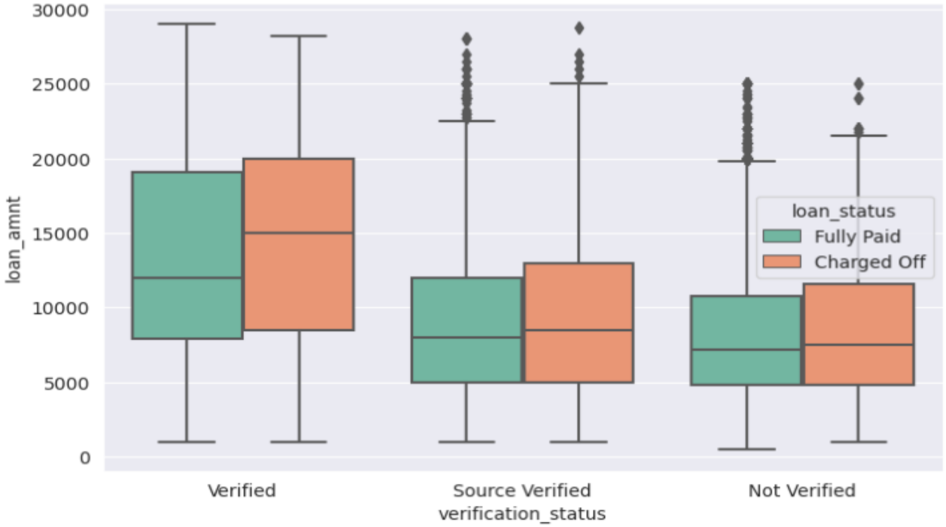


3



# VII. ‘verification_status’ field

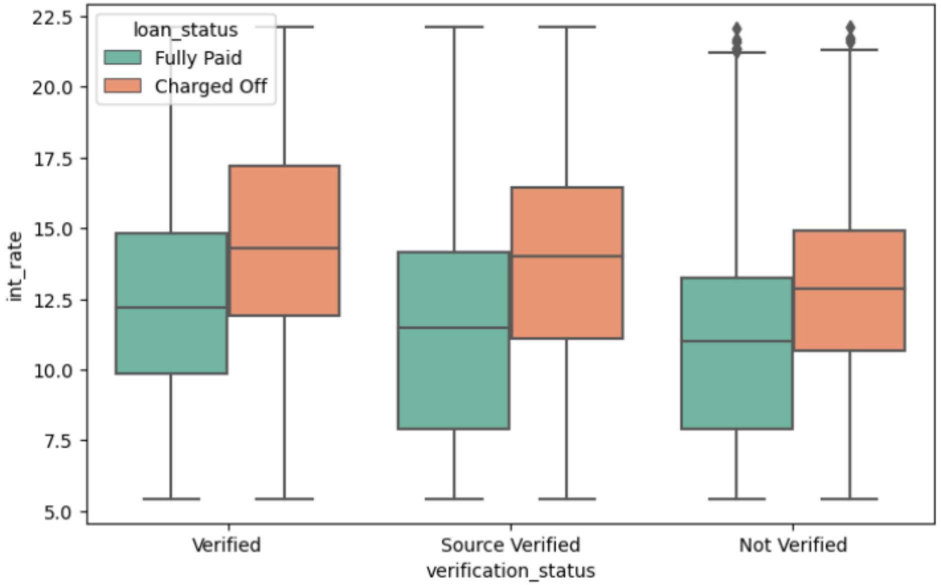
1



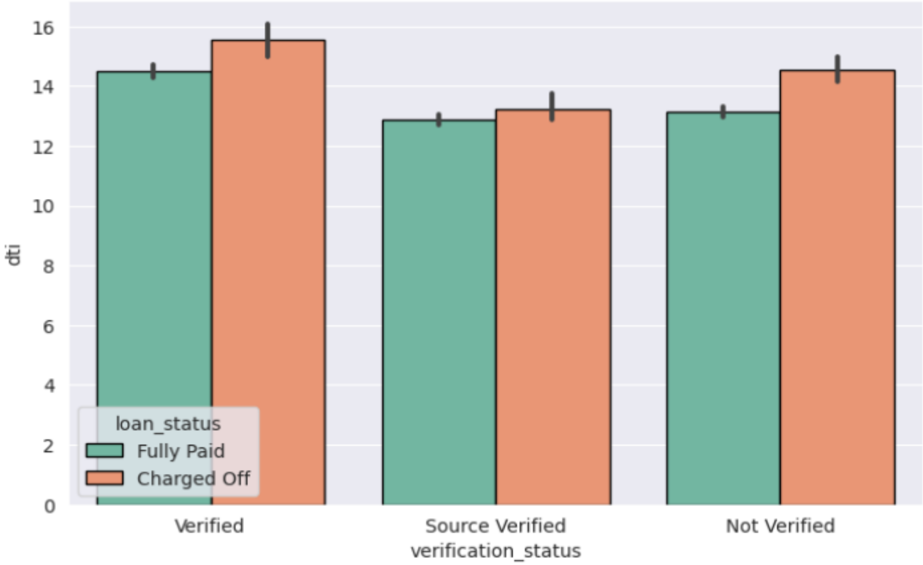
Observations:

1. For all the segments of verification_status, higher the loan amount, higher are the chances of default.
2. Irrespective of verification_status, more the interest rate, more are the chances of default.
3. For all the segments of verification_status, higher the dti ratio, higher are the chances of defaulting the loan.

2

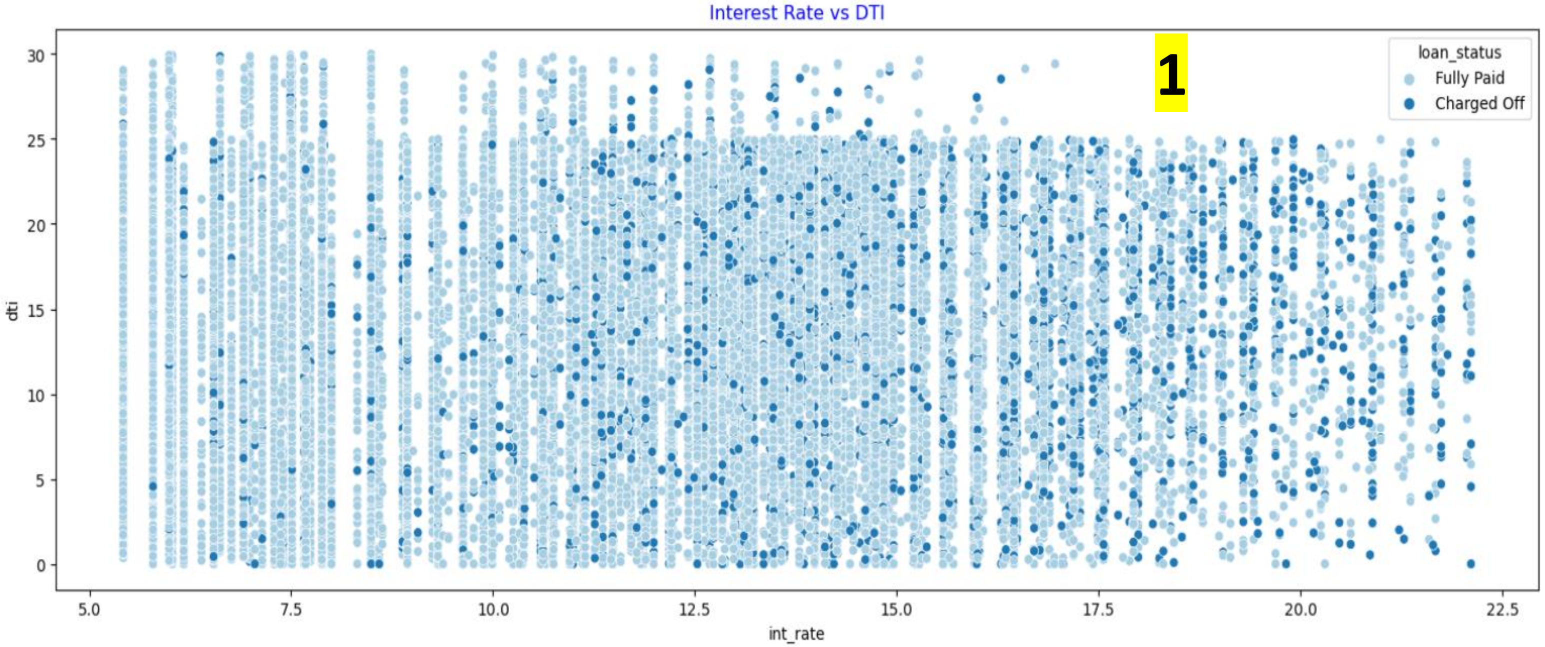


3



# Bivariate Analysis (Both fields NUMERICAL)

## I. Interest Rate vs DTI



Observation:

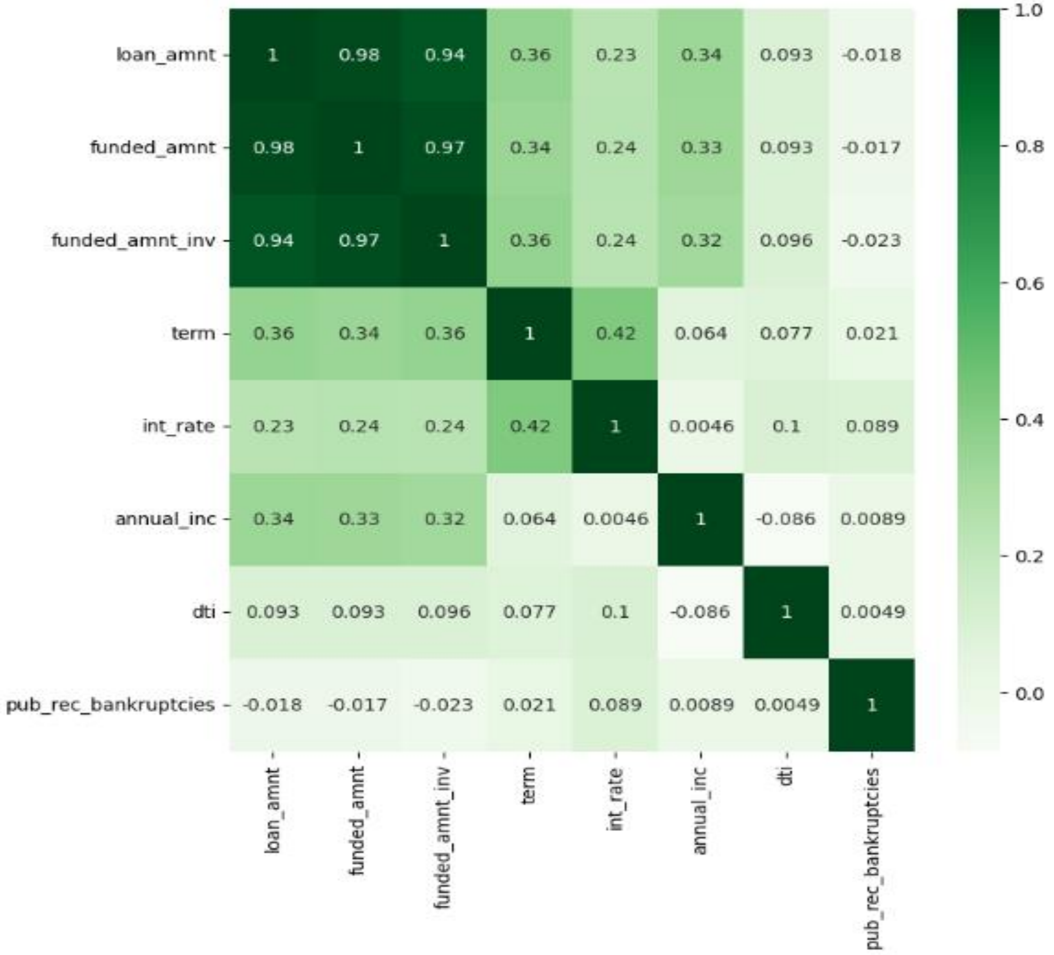
1. The data points are very much distributed across the plot; but irrespective of DTI, for higher interest loans, charged-off rates are also high.

For plots ‘Loan Amount vs Annual income’, ‘Loan Amount vs Interest Rate’ and ’Loan Amount vs DTI’ , it was observed that the data points were pretty much spread across the plots and no specific pattern was found.

# Multivariate Analysis

## Correlation Matrix and Heatmap

Correlation between Numerical Columns



|  | loan_amnt | funded_amnt | funded_amnt_inv | term | int_rate | annual_inc | dti | pub_rec_bankruptcies |
|---|---|---|---|---|---|---|---|---|
| loan_amnt | 1 | 0.98 | 0.94 | 0.36 | 0.23 | 0.34 | 0.093 | -0.018 |
| funded_amnt | 0.98 | 1 | 0.97 | 0.34 | 0.24 | 0.33 | 0.093 | -0.017 |
| funded_amnt_inv | 0.94 | 0.97 | 1 | 0.36 | 0.24 | 0.32 | 0.096 | -0.023 |
| term | 0.36 | 0.34 | 0.36 | 1 | 0.42 | 0.064 | 0.077 | 0.021 |
| int_rate | 0.23 | 0.24 | 0.24 | 0.42 | 1 | 0.0046 | 0.1 | 0.089 |
| annual_inc | 0.34 | 0.33 | 0.32 | 0.064 | 0.0046 | 1 | -0.086 | 0.0089 |
| dti | 0.093 | 0.093 | 0.096 | 0.077 | 0.1 | -0.086 | 1 | 0.0049 |
| pub_rec_bankruptcies | -0.018 | -0.017 | -0.023 | 0.021 | 0.089 | 0.0089 | 0.0049 | 1 |

Observations:

1. Strong Correlation: ‘funded_amnt’ has strong correlation with ‘loan_amnt’ & ‘funded_amnt_inv’. ‘loan_amnt’ has strong correlation with ‘funded_amnt_inv’.
2. Moderate Correlation: ‘term’ has moderate correlation with ‘int_rate’, ’loan_amnt’ & ’funded_amnt_inv’
3. Negative Correlation: ‘pub_rec_bankruptcies’ has negative correlation with ’loan_amnt’ & ‘funded_amnt’. ‘dti’ negative correlation with ‘annual_inc’.

# Conclusions:

- The company should minimize high interest loans of ‘60 months’ tenure as they are more prone to being charged-off.
- Borrowers from ‘OTHER’ and ‘MORTGAGE’ home ownership status, taking higher loan amount have defaulted more. The company should keep this in mind while approving loans from such type of borrowers.
- Majority of applicants who defaulted the loan belong to states 'CA', 'FL' and 'NY'. So, the company should reduce the number of loan sanctions to the borrowers, who are from these states, to cut down the amount of credit loss.
- Loan grades are a good metric for detecting defaulters. As the loan grade decreases (A to G), the proportion of loan default increases (A to G). The company should be careful while approving lower graded loans and should do more examination of borrowers applying for such lower graded loans.
- Small Business purpose loans have the highest charged-off proportion. Company should reduce issuing loans to the borrowers who are taking loans for such purpose.
- For all the segments of verification_status, higher the dti ratio, higher are the chances of defaulting the loan.