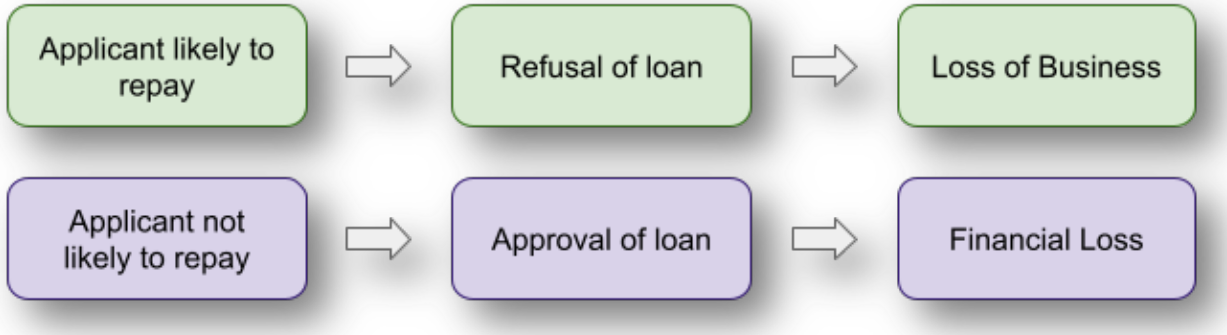
Lending Club Case Study

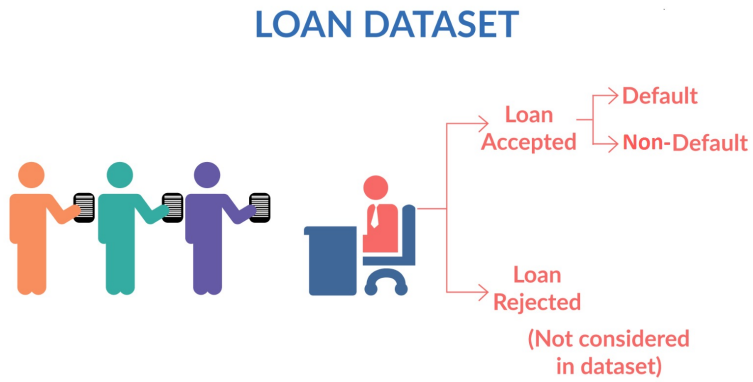
SUBMISSION

Name: Suprabhat Paul

Introduction

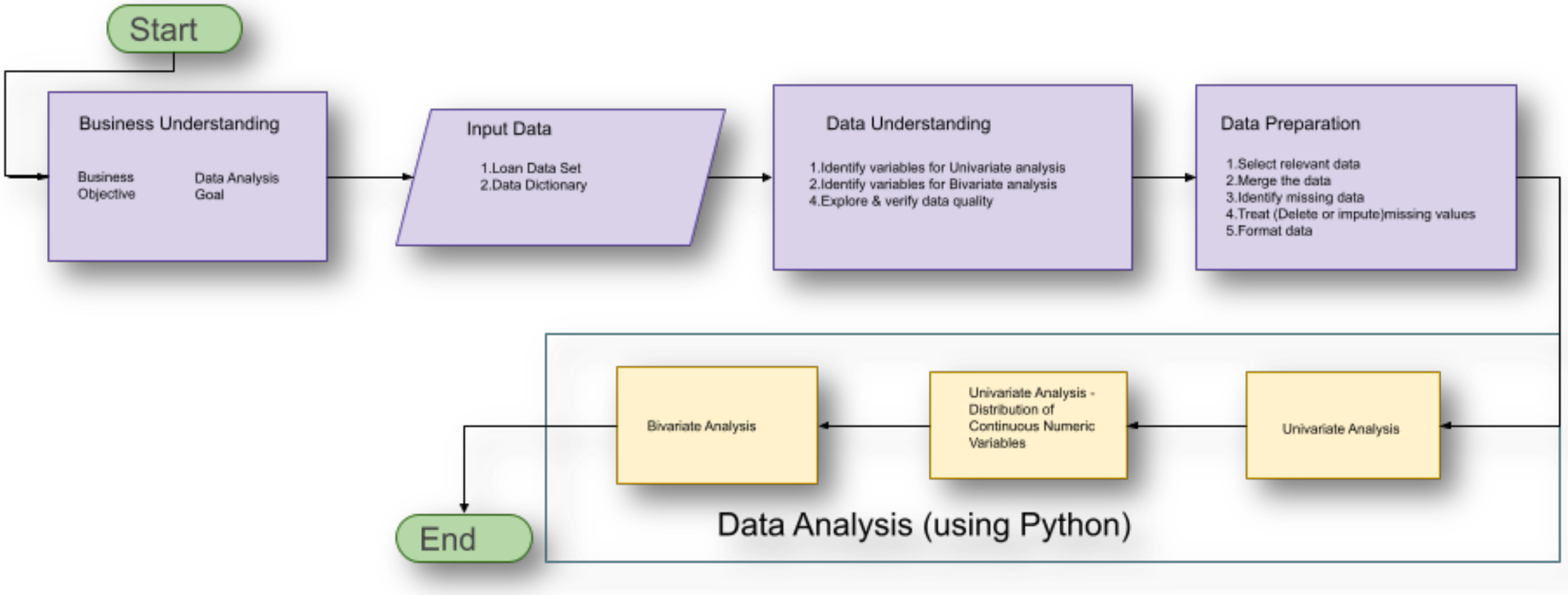
A consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision





- **Business objective:** The company wants to understand the driving factors (or driver variables) behind loan defaulter, i.e. the variables which are strong pointers of defaulter. The company can use this knowledge for its portfolio and risk/threat assessment.
  - Identification of loan applicant traits that tend to "Default" paying back.
  - Understand the "Deriving Factors" or "Driver Variables" behind loan default.
  - Scrutiny the new loan applicants portfolio and risk assignment based on historic data.
- **Business Understanding:** As a financial organisation it is always risky to lend loans, actually it largest source of fiscal loss (called credit loss). The credit loss is the quantum of money lost by the lender when the borrower refuses to pay or runs down with the amount owed. In short, borrowers who don't pay cause the largest quantum of loss to the lenders. In this case, the customer labelled as' charged-off' are the 'defaulters'.

Problem solving methodology



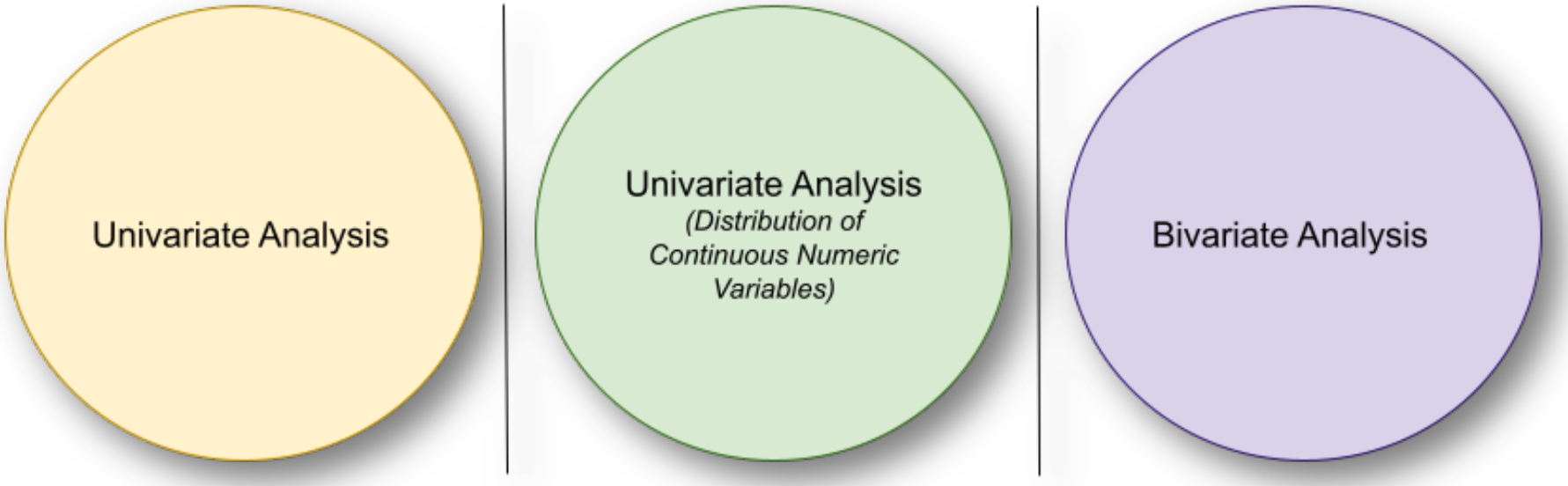
Data Cleansing - Process

- ✓ Verify the data and attributes
- ✓ Remove Columns for which all rows are NULL
- ✓ Remove columns which are non-essential for the analysis
- ✓ Check any duplicate rows exists
- ✓ Convert columns to standard datetime format
- ✓ Convert column with % (percentage) to numeric format
- ✓ Convert int amount column attributes float format
- ✓ Conversion all string attributes to Upper Case
- ✓ Re- Verify the percentage NULL Values in loan data frame
- ✓ Remove redundant Key column like member_id
- ✓ Remove high % NULL value columns
- ✓ Filter rows only having loan applications for loan status as 'Charged Off' or 'Fully Paid' for further data analysis
- ✓ Split filtered loan based on Loan Status
- ✓ Percentage rows left after Data clean up in Loan

Data Cleansing - Conclusion

- ✓ Checked and confirmed that no duplicates exists in loan data frame.
- ✓ Month and Year type string column attributes are verified & converted to Datetime objects.
- ✓ The rate columns with % symbol are removed and converted to numeric format.
- ✓ The ‘loan_amnt’ and ‘funded_amnt’ columns converted to correct float format.
- ✓ Inconsistencies in character columns are removed by converting all columns values to upper case.
- ✓ Removed the redundant member_id column and retain id column as key.
- ✓ Nonessential & high null value percentage column attributes `mths_since_last_record` and `next_pymnt_d` are dropped.

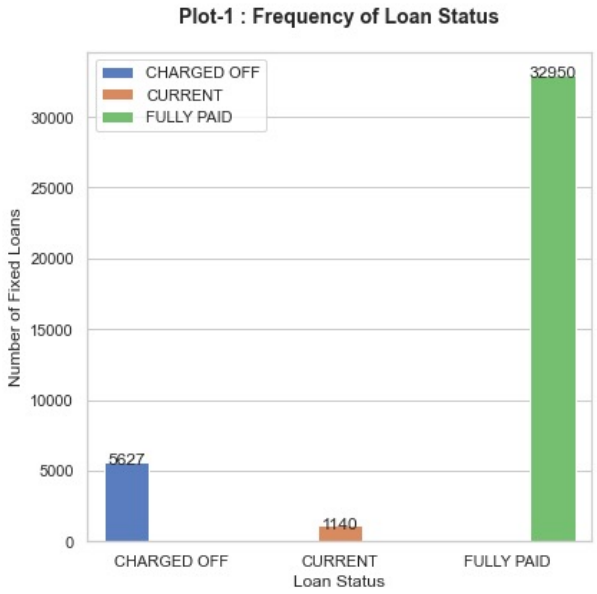
Data Analysis

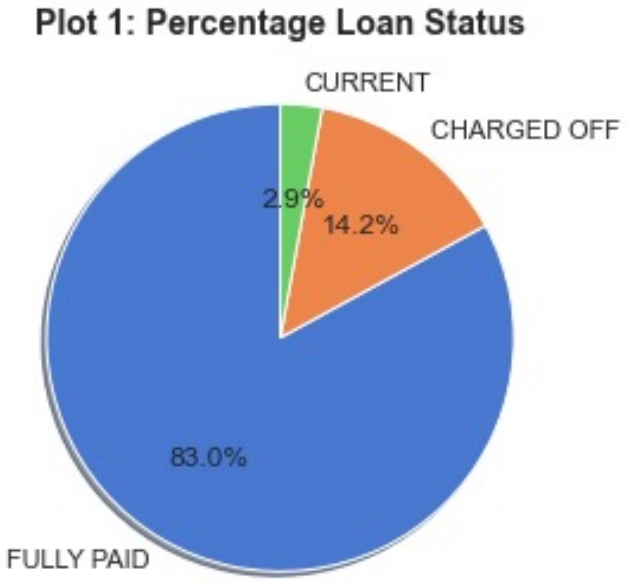


Univariate Analysis

- ✓ Frequency and Percentage Loan Status .
- ✓ Loan Term against Default Loan Applications (Charged Off Loans)
- ✓ Assigned Grades versus Number of Default Applicants (Charged off Loans).
- ✓ Assigned sub-Grades versus Number of Default Applicants (Charged off Loans).
- ✓ Employee Experience versus Number of Default Applicants (Charged off Loans).
- ✓ Home Ownership versus Number of Default Applicants (Charged off Loans).
- ✓ Verification Status versus Number of Default Applicants (Charged off Loans).
- ✓ Purpose versus Number of Default Applicants
- ✓ Resident State against Number of Default Applicants
- ✓ Resident State against Segmented All Loan Status (Charged off , Fully Paid and Current Loans)

Analyze the Frequency and Percentage Loan Status .

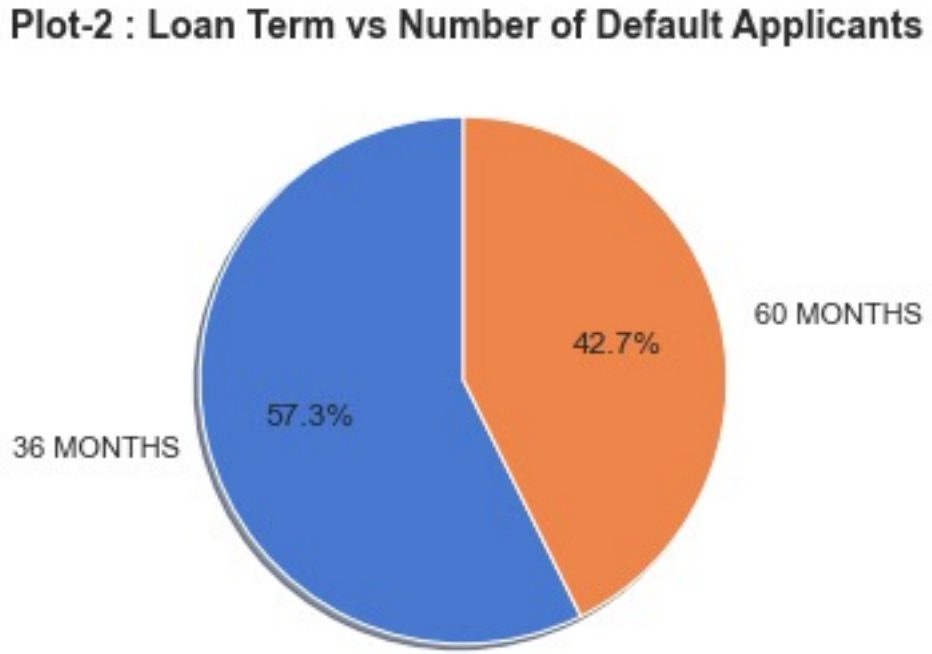




Observations

- Indicates out of 39717 loans granted to applicants 5627 loans were 'default' (charged off loans) and 1140 loans are 'current' (or payment active loans for which installments are getting paid).
- Indicates the highest percentage customers are in 'fully paid' category i.e., who fully repay their loan (the principla and interest rate) - `83.%`
- Approximately `14.2%` of customers fall in the category of 'Charged Off' i.e., they don't pay their due in time or for a long period of time (defaulted).

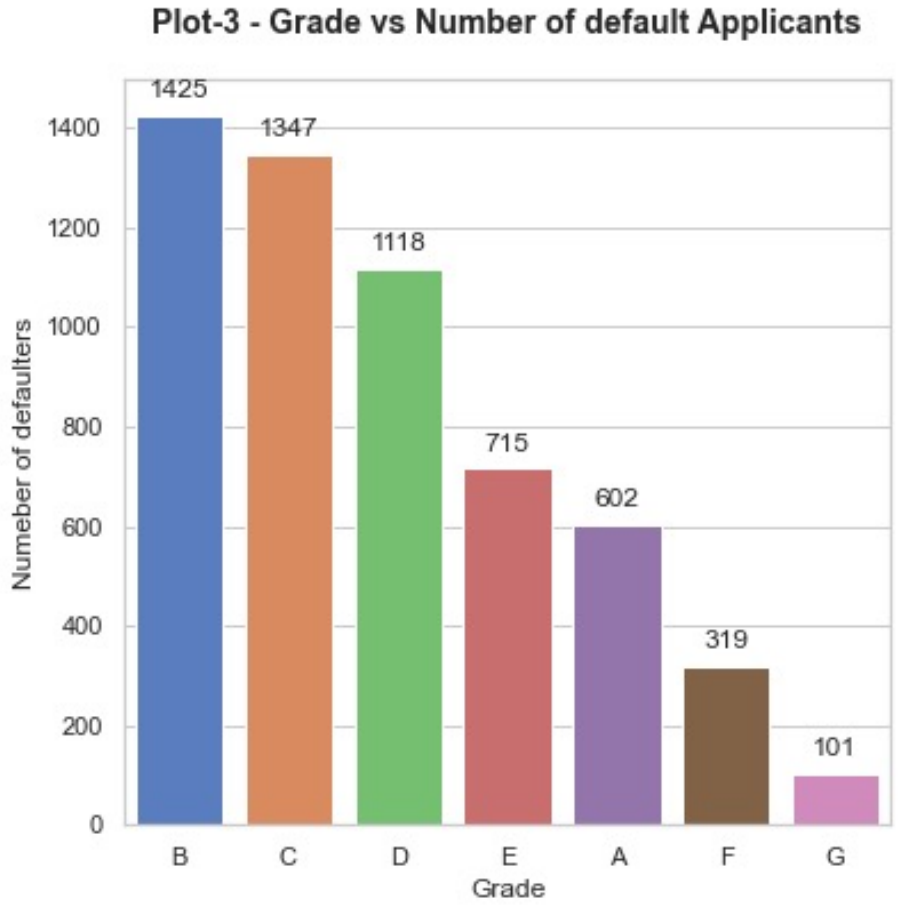
Analyze the Loan Term against Default Loan Applications (Charged Off Loans)



Observations

- The Plot-2 shows percentage default applicants are in category of payment terms in 36 months or 60 months.
- The percentage of 36 months category(`57.3%`) is higher than 60 months (`42.7%`) category.

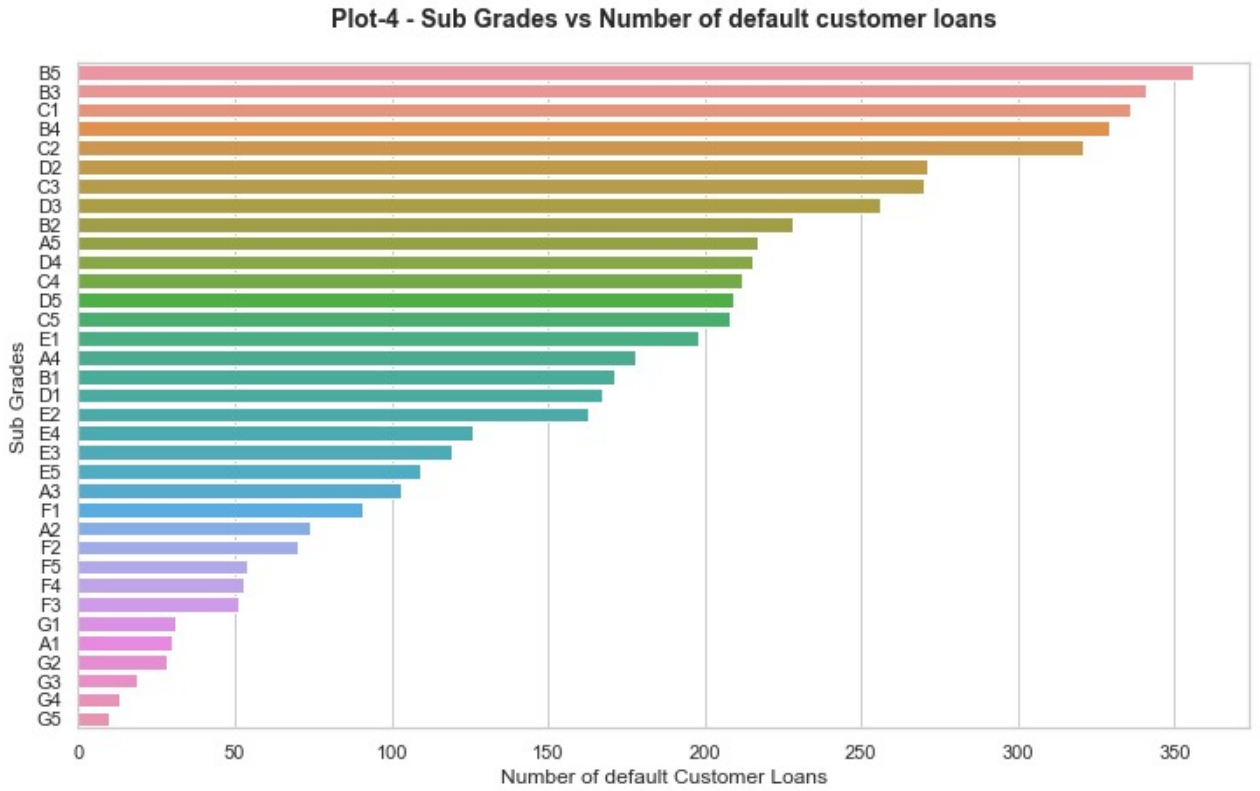
Analyze the Assigned Grades versus Number of Default Applicants (Charged off Loans).



Observations

- Indicates the highest number 'default' (charged off) customers whose loans are under LC assigned grade 'B' (1425 defaulters) and followed by 'C' (1347 defaulters).

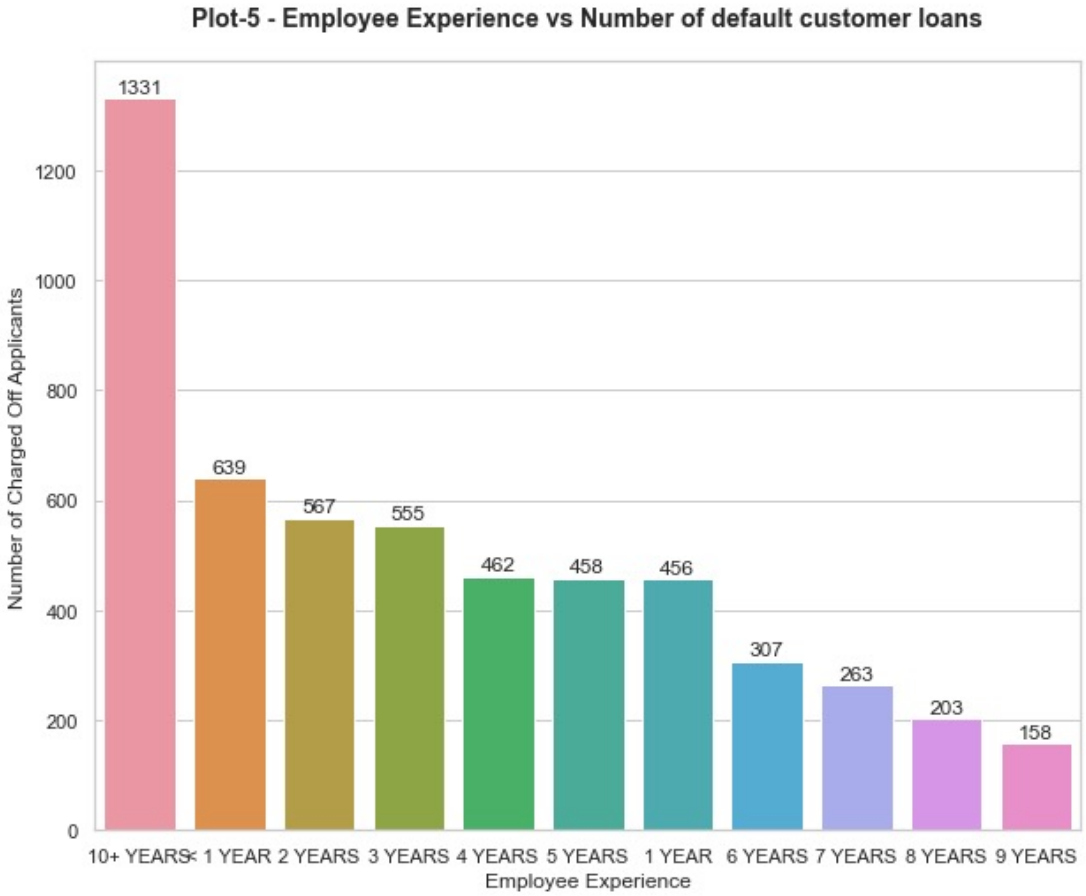
Sub Grades vs Number of default customer loans



Observations

- Indicates the highest number 'default' (charged off) customers whose loans are under LC assigned sub grade 'B5' (356) and followed by 'B3' (341).

Employee Experience vs Number of default customer loans



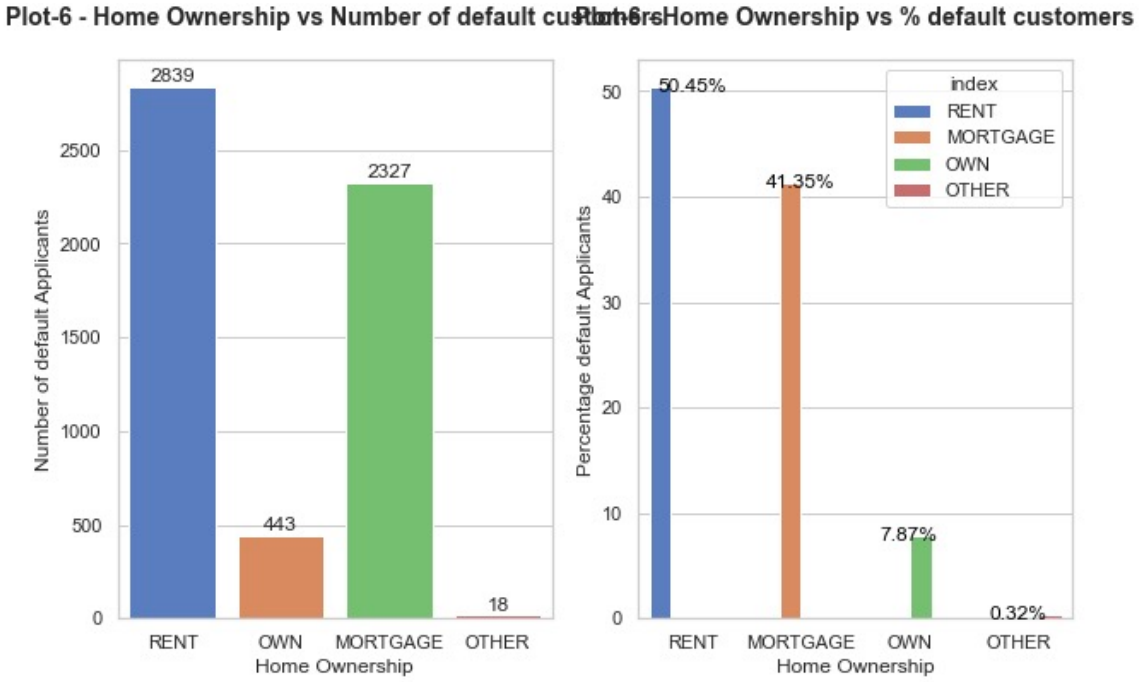
Observations

- Indicate employees with experience in the range greater than 10+ years (1331) followed by less than 1 years (639) are highest number of defaulters.

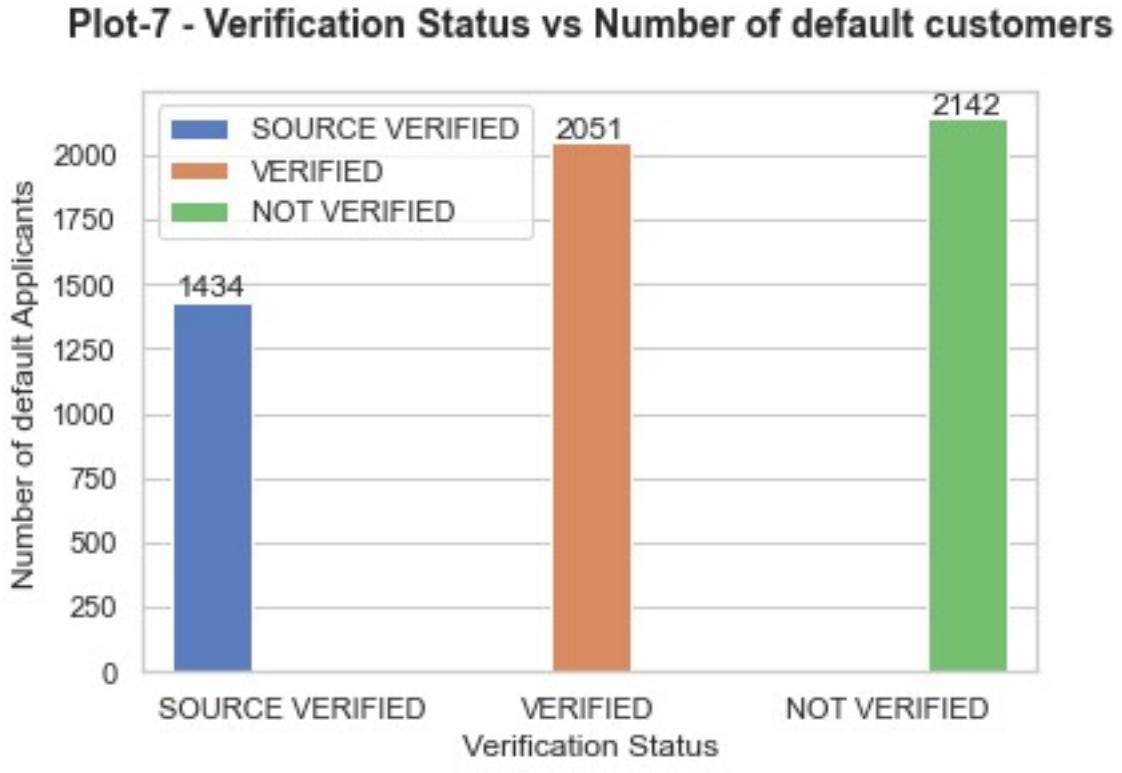
Home Ownership vs Default customers

Observations

- Indicate employees who are staying in RENT are the highest defaulters with high as 50% i.e., 2839 defaulters. out of 5627 applicants.
- So, most defaulters have rented (50.45%) or mortgaged (41.35%) homes.



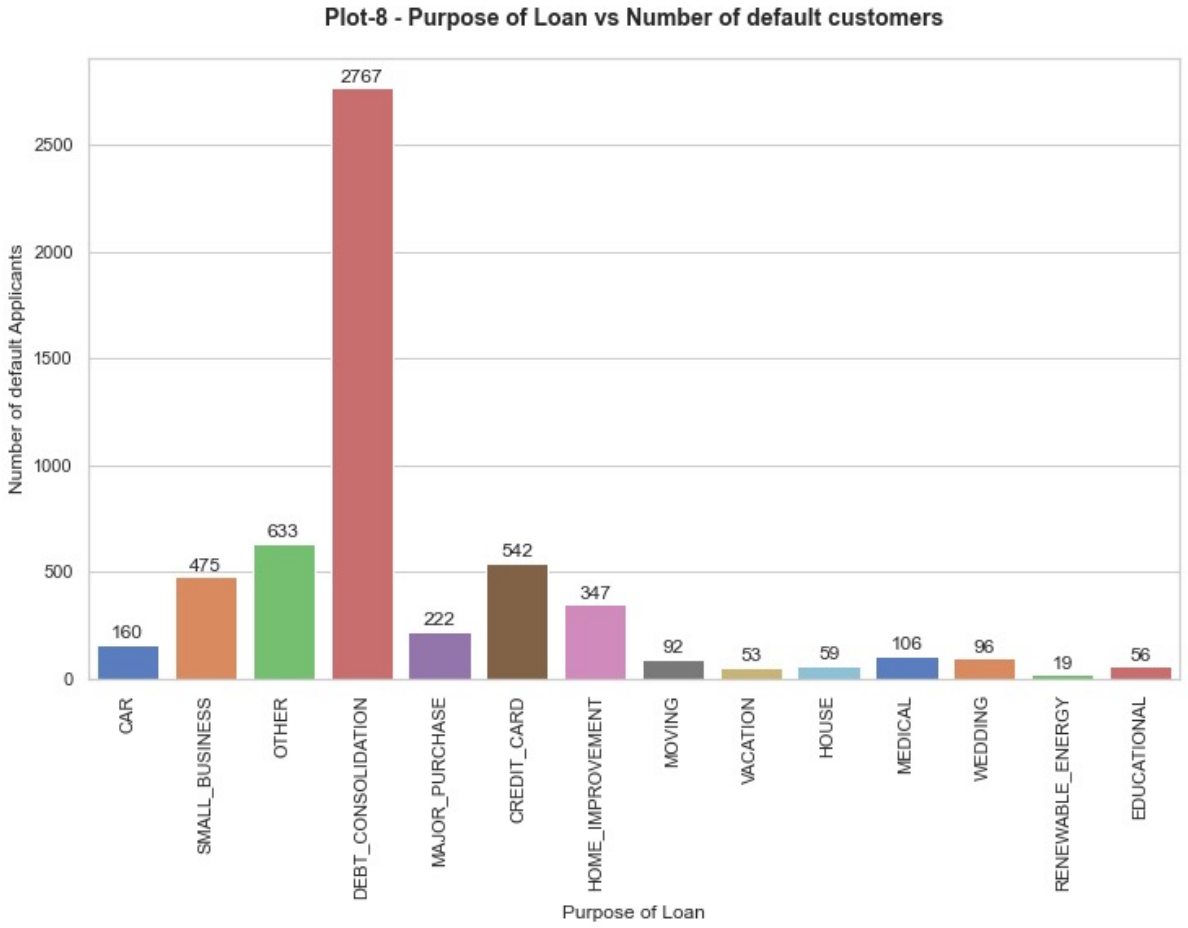
Verification Status vs Number of default customers



Observations

- Indicate employee verification status against the default loan Applicants.
- It is very important to note, many applicants' status are not verified by the bank, before lending to the 'high risky applicants'.
- The number of not verified cases are `2142`.

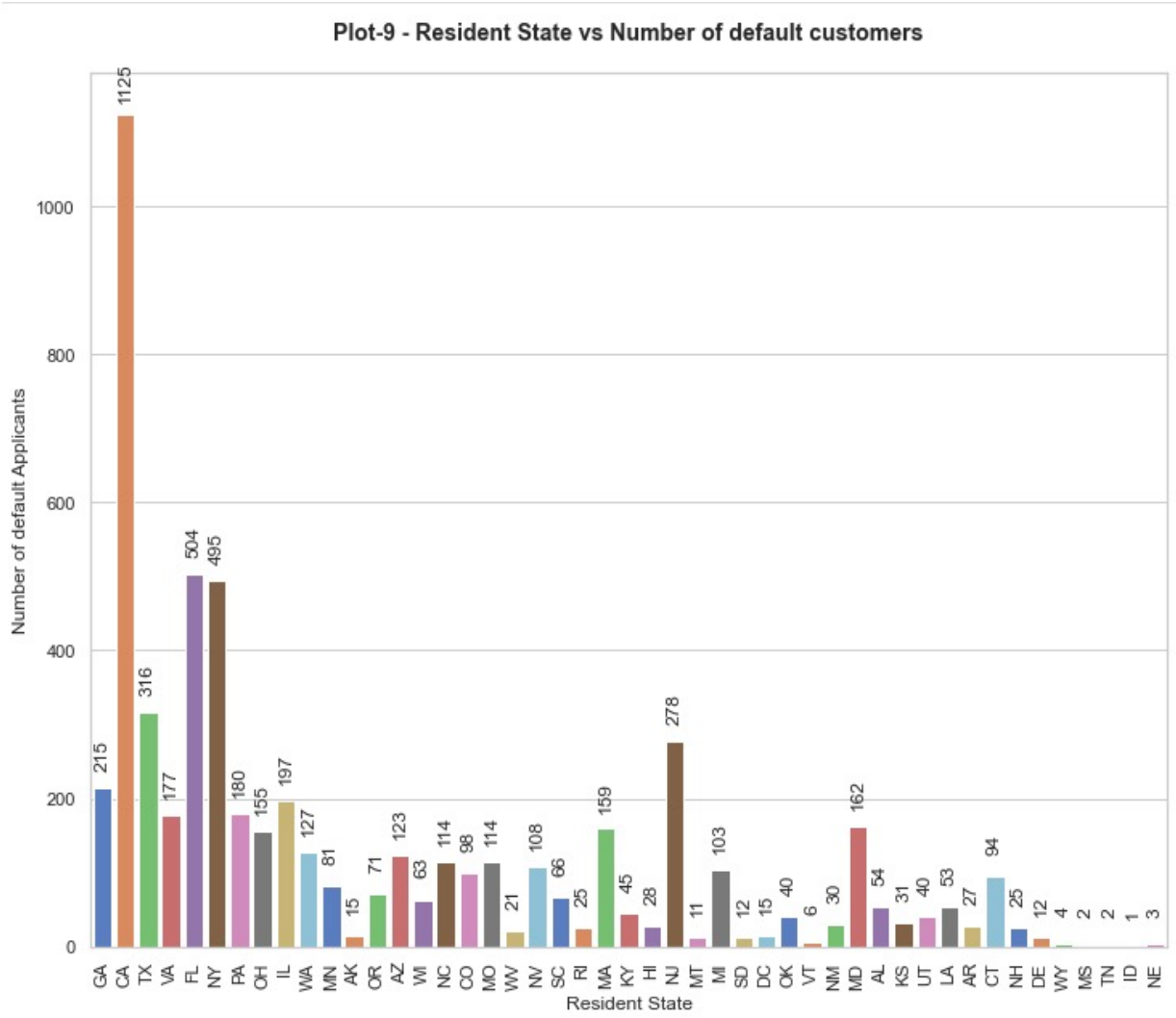
Purpose of Loan vs Number of default customers



Observations

- The above plot is for Purpose of the loan against the number of default applicants.
- It is clear, the highest number of defaulted application are against the ‘debt_consolidation’ purpose - `2767`.

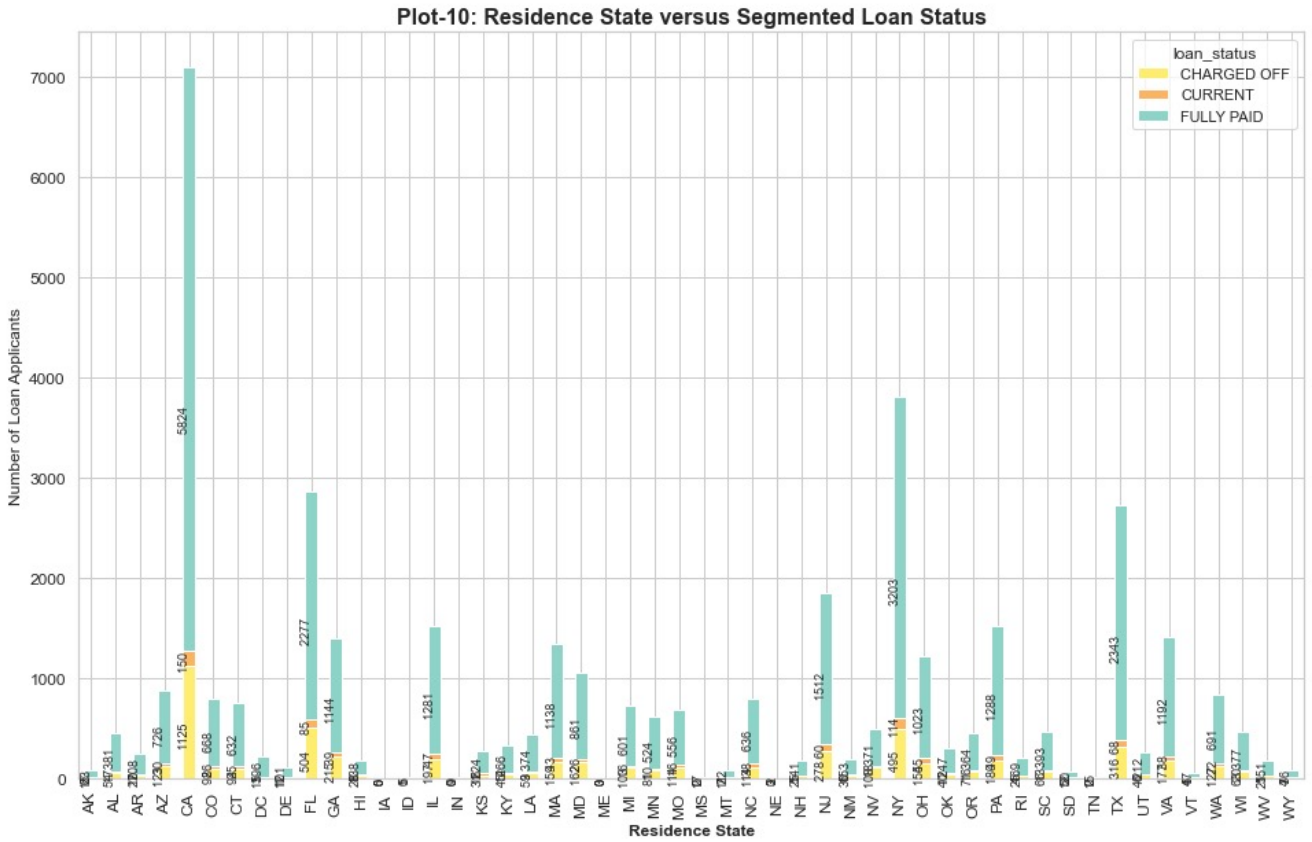
Resident State vs Number of default customers



Observations

- Indicate the number of charged off customer in each of the resident state.
- It is clear, CA (California) State has the greatest number of cases -`1125`, which are marked charged off.

Residence State vs Segmented Loan Status



Observations

- Indicate the number of charged off and Fully Paid customers in each of the resident state.
- It is clear, CA (California) State has most number of cases of defaulted loans -`1125` of all loans which is 16% of the total loan Applicants in CA.

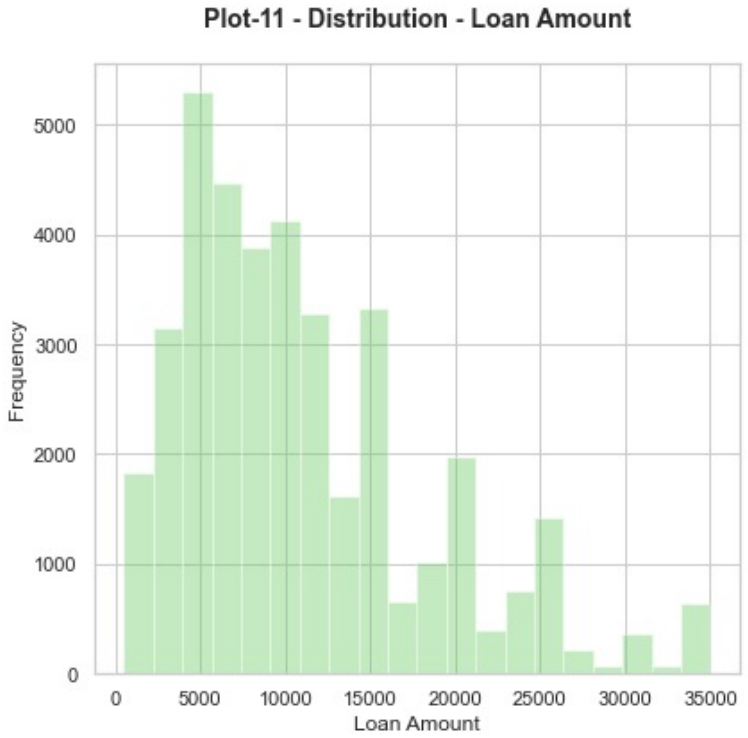
Univariate Analysis - Distribution of Continuous Numeric Variables

- ✓ Loan Amount Analysis
- ✓ Funded Amount Analysis
- ✓ Funded Amount Invested Analysis
- ✓ Loan Issue Date (Year) Analysis
- ✓ LC pulled credit year for the loan Analysis
- ✓ Last Credit Payment Year Analysis for the loan Analysis

Loan Amount Analysis

Observations

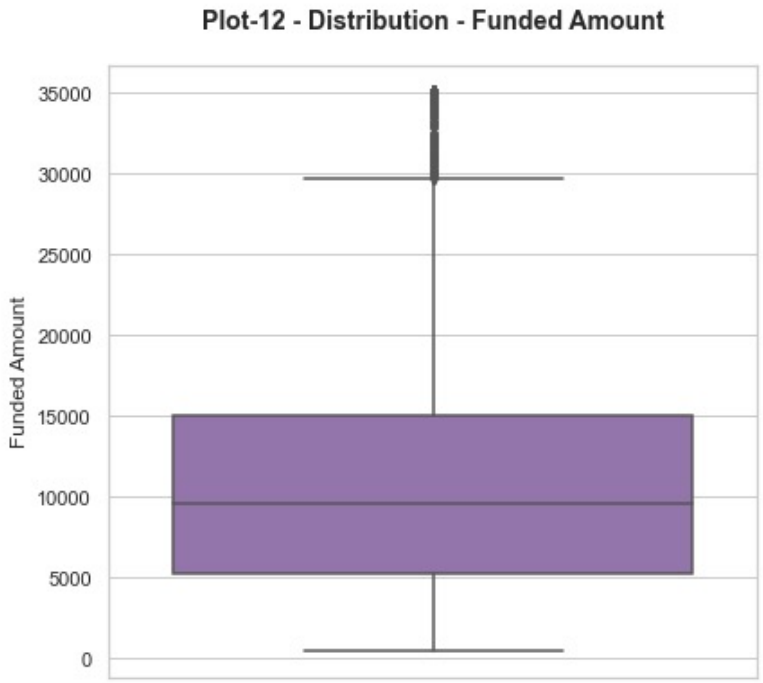
- Indicate distribution of loan amount applicants. It is gaussian distribution but there are outliers at the end.
- The mean of `$11,047.03` and 25% and 75% quartiles of `$5,300.00` and `$15,000.00` respectively.

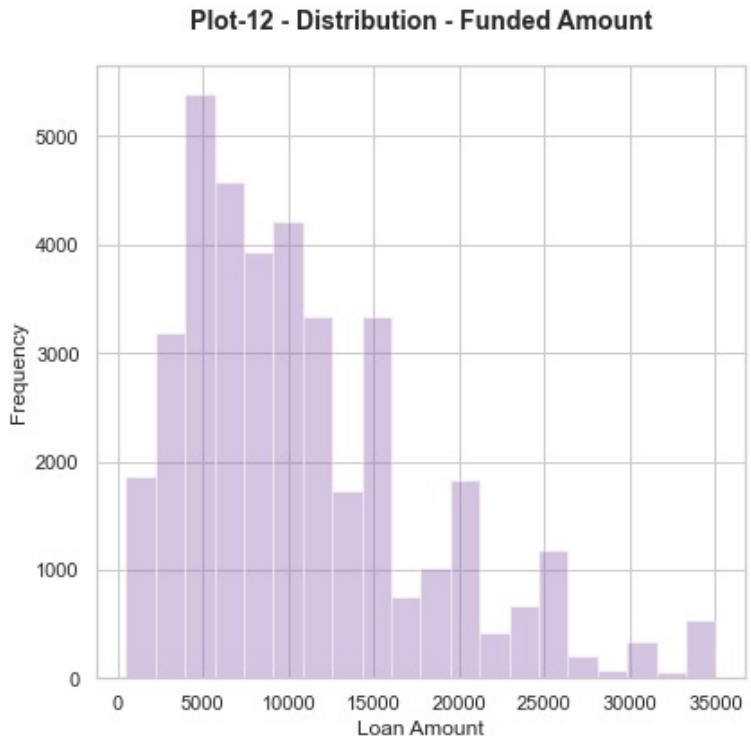


Distribution - Funded Amount

Observations

- Indicate distribution of funded amount across loan applicants. It has almost normal distribution with mean of `$10,784.06` and 25% and 75% quartiles of `$5,200.00` and `$15,000.00` respectively.

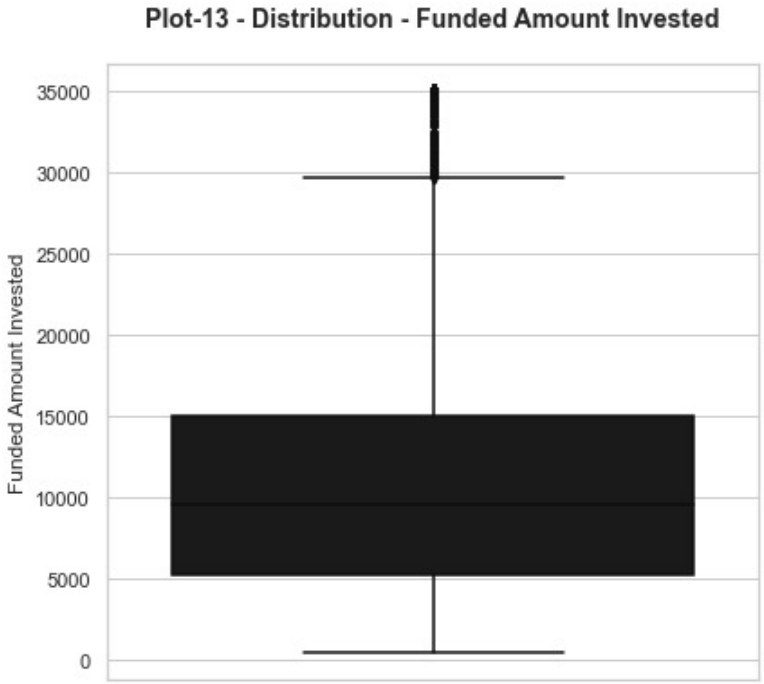


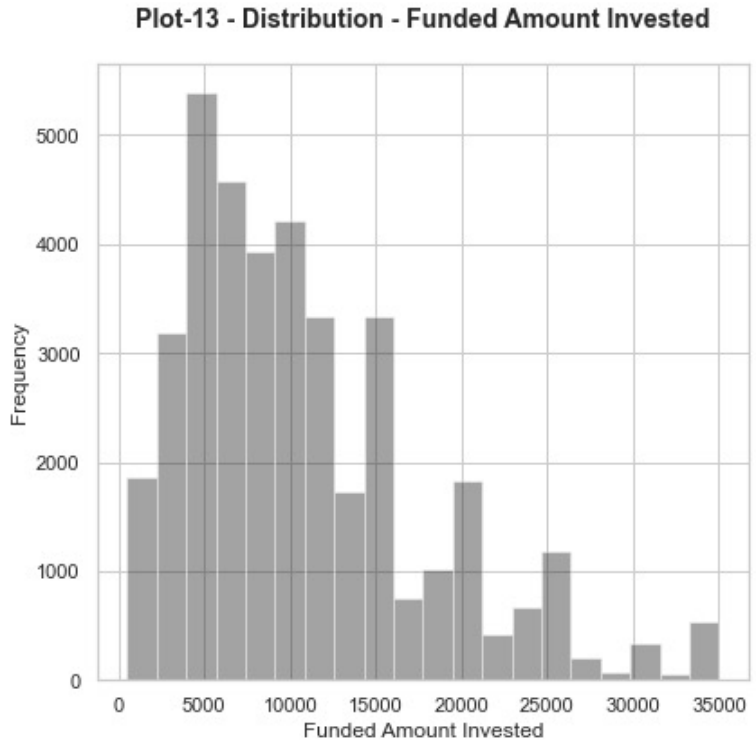


Distribution - Funded Amount Invested

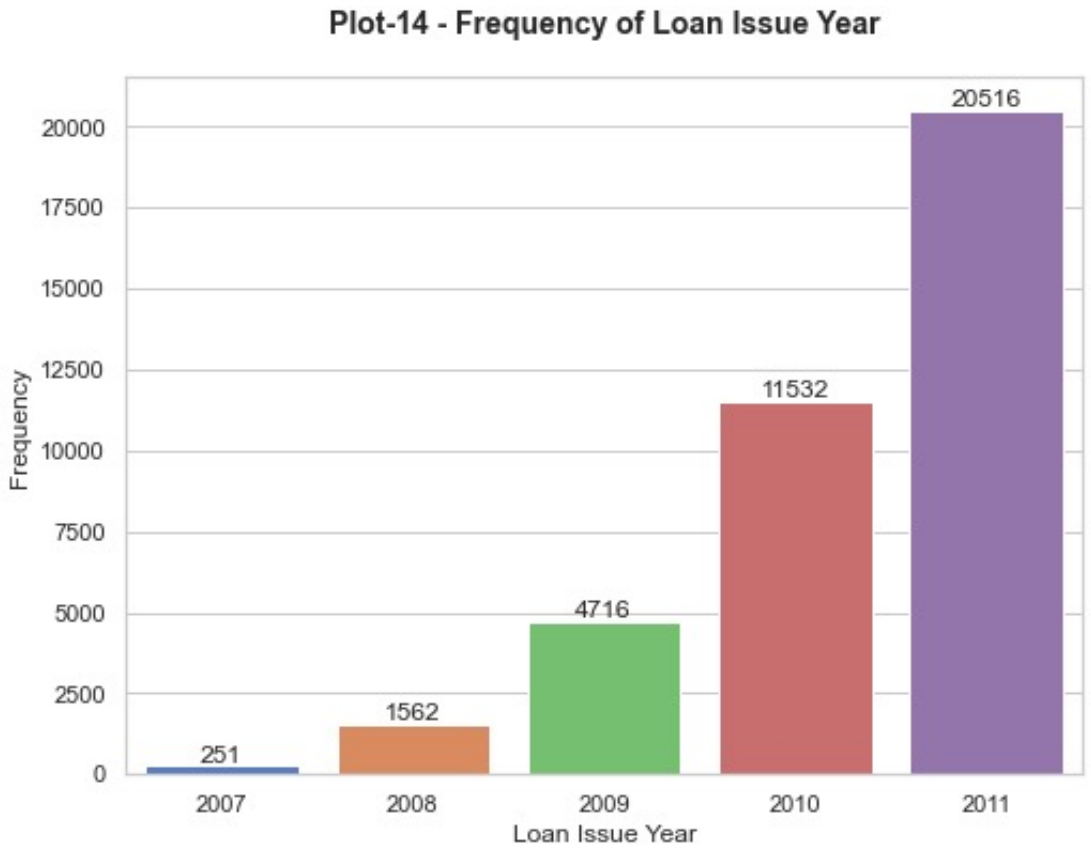
Observations

- Indicate distribution of funded amount Invested across loans. It has almost normal distribution with mean of `$10,222.48`.
- Fund invested are higher in the range of `$5000 - $10000` range.





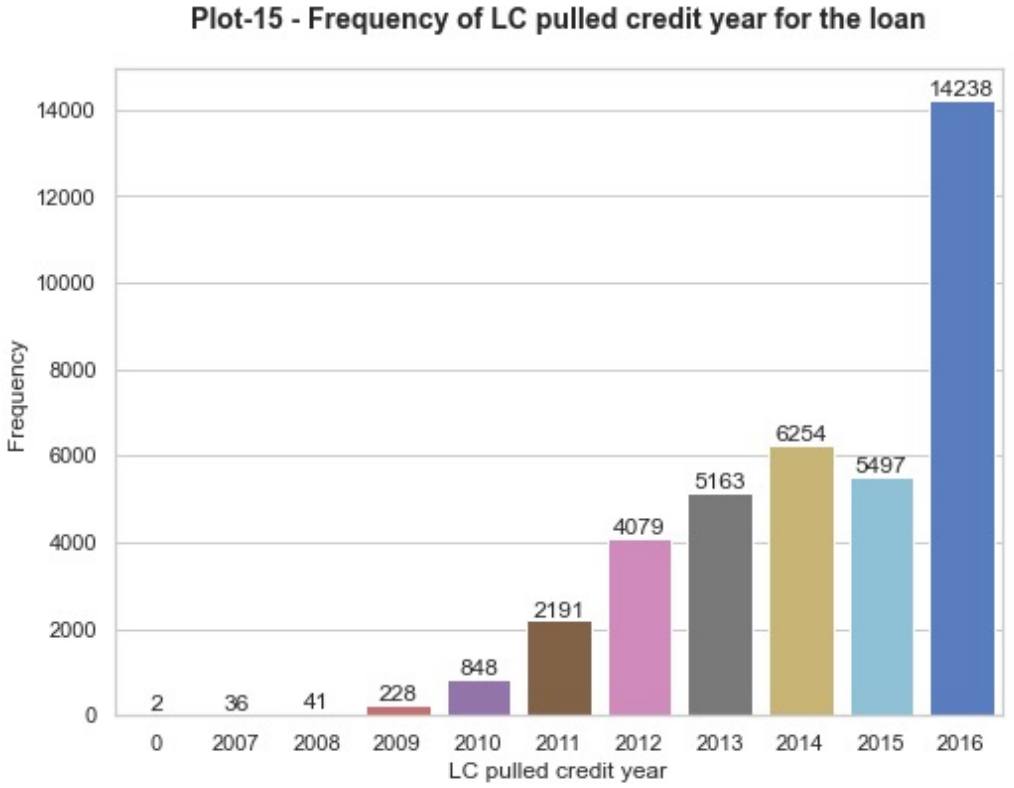
Frequency of Loan Issue Year



Observations

- Indicate the majority of the loans were issued during year 2011.
- The number of loans issued during 2011 is `20516`

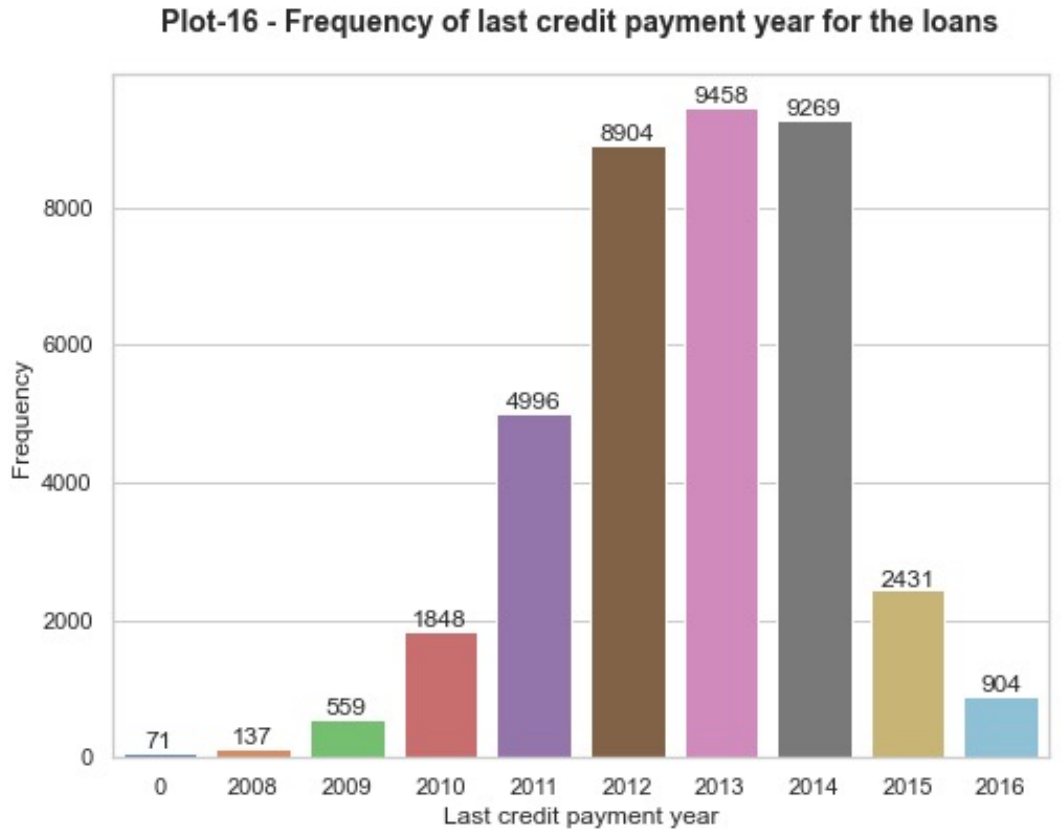
Frequency of LC pulled credit year for the loan



Observations

- Indicate most of the loans were pulled credit for the loans in the year 2011.
- The number of pulled credit for the loans during 2011 is `14328`

Frequency of last credit payment year for the loans



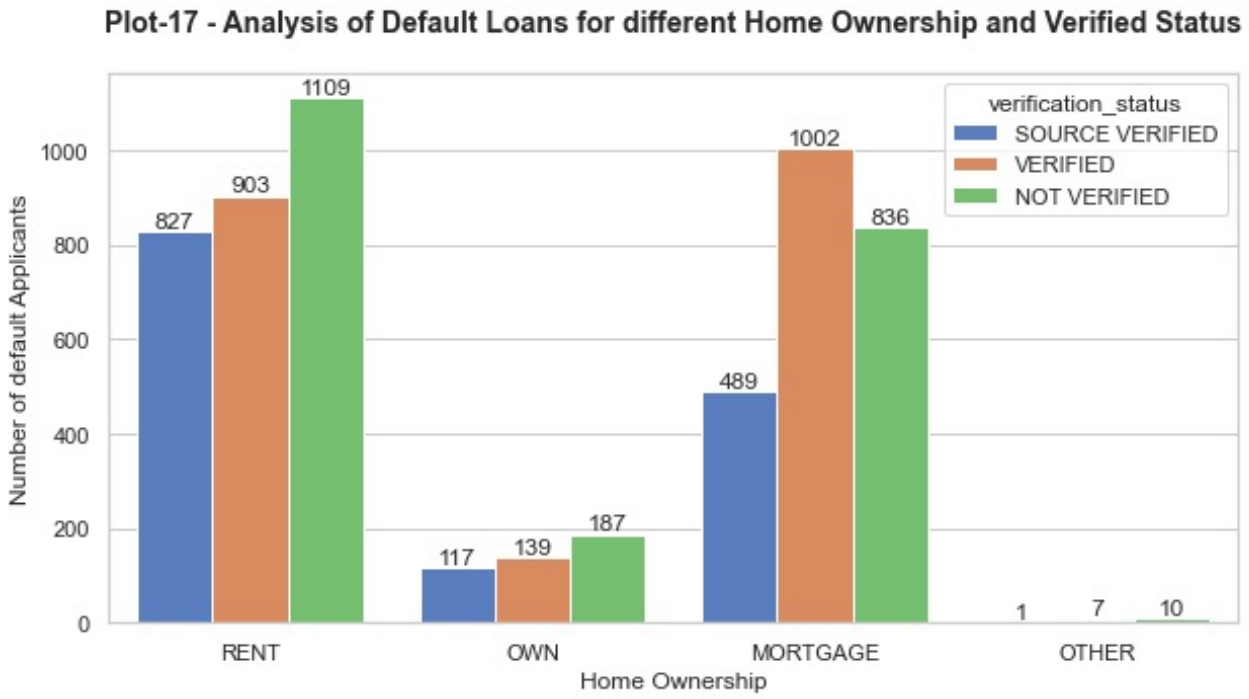
Observations

- Indicate the highest number credit payment happened for the loans during the year 2013.
- The number of credit payments for the loans during 2013 is `9458`

Bivariate Analysis

- ✓ Analysis of Default Loans for different Home Ownership and Verified Status.
- ✓ Analysis of Annual Income Group Categories against Number of Default Loans
- ✓ Analysis of Annual Income Group for term against Number of Default Loans
- ✓ Analysis of Annual Income Group for the Verification Status against Number of default Loans
- ✓ Analysis of Interest Rate Bin Category versus Term loans for default loans.
- ✓ Analysis of Interest Rate Category and Employee Service length.
- ✓ Correlation between Loan Amount and Funded Amount for the Purpose of Loan.
- ✓ Analysis of years of experience vs House residence vs Number of defaults.
- ✓ Analysis of Annual Income vs Number of defaults
- ✓ Analysis of Annual Income vs Number of public record bankruptcies
- ✓ Bivariate Correlation Analysis..

Analysis of Default Loans for different Home Ownership and Verified Status

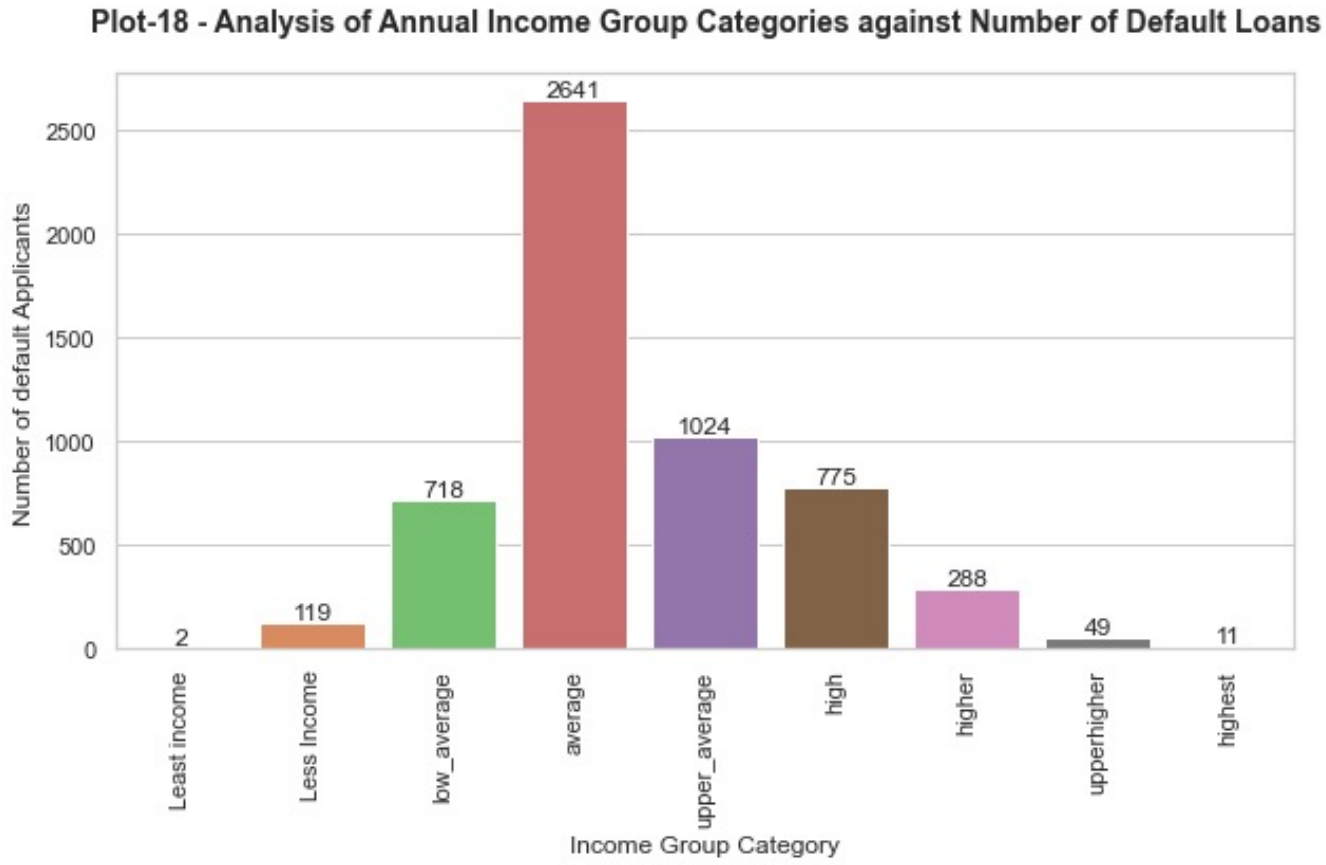


Observations

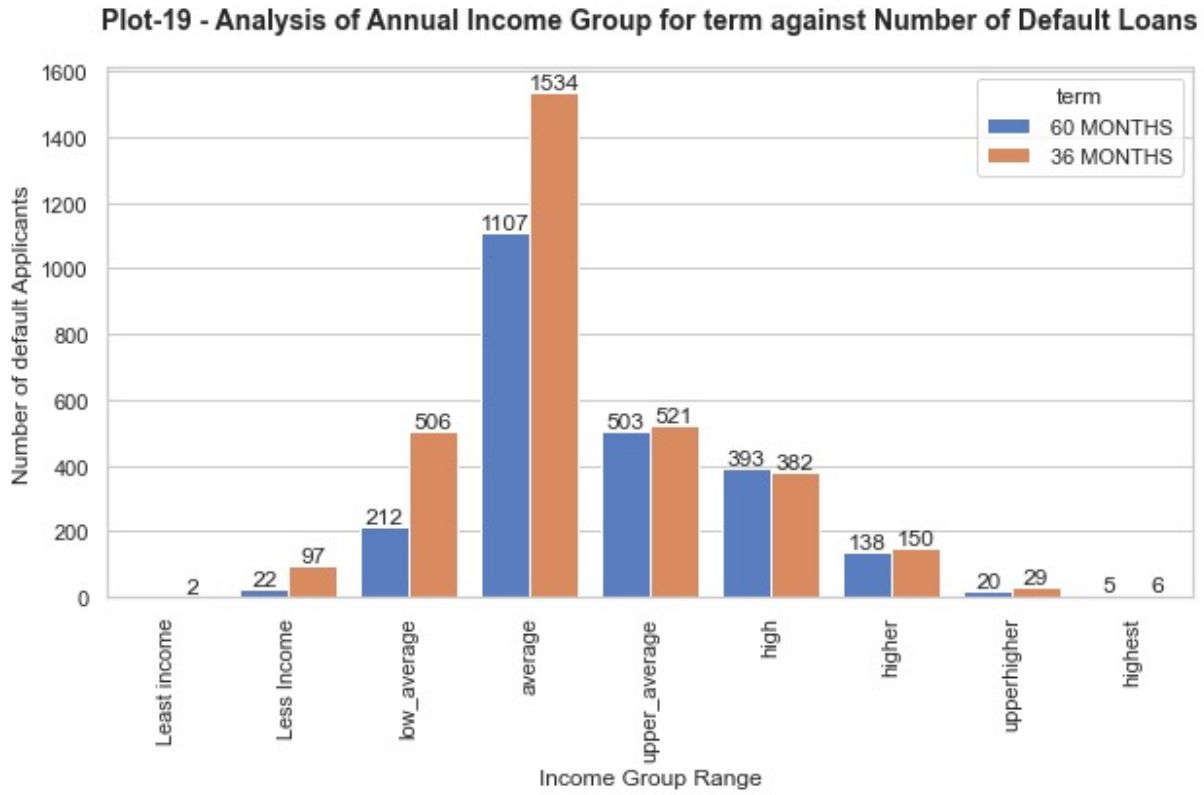
- Indicate the majority defaulters in the in the RENT category.
- In that, around `1109` cases in which source was `not verified` indicating the huge risk.

Analysis of Annual Income Group Categories against Number of Default Loans

- **Observations**
- Indicate Maximum defaulters are in annual income range of `30000 - 60000`.



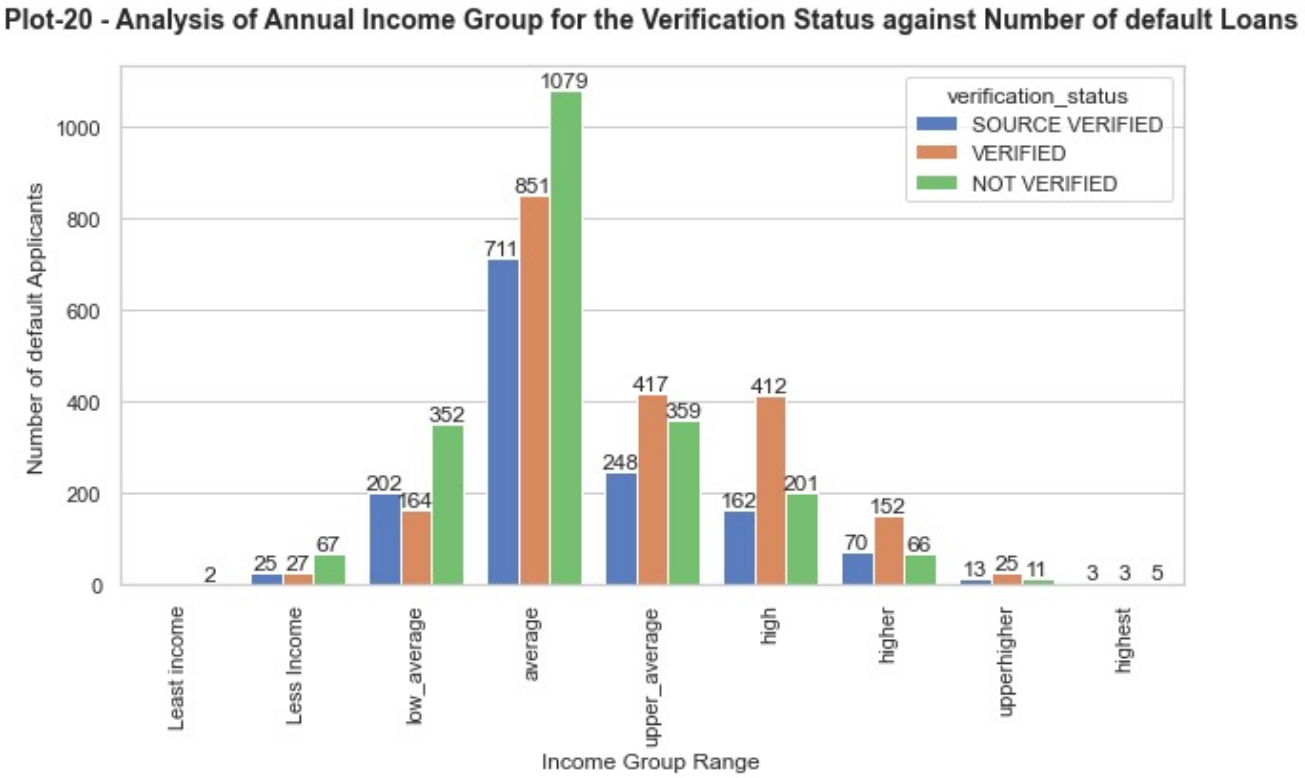
Analysis of Annual Income Group for term against Number of Default Loans



Observations

- Indicate Maximum defaulters in the term of 36 and 60 months are in the annual income range of `30000 - 60000`.

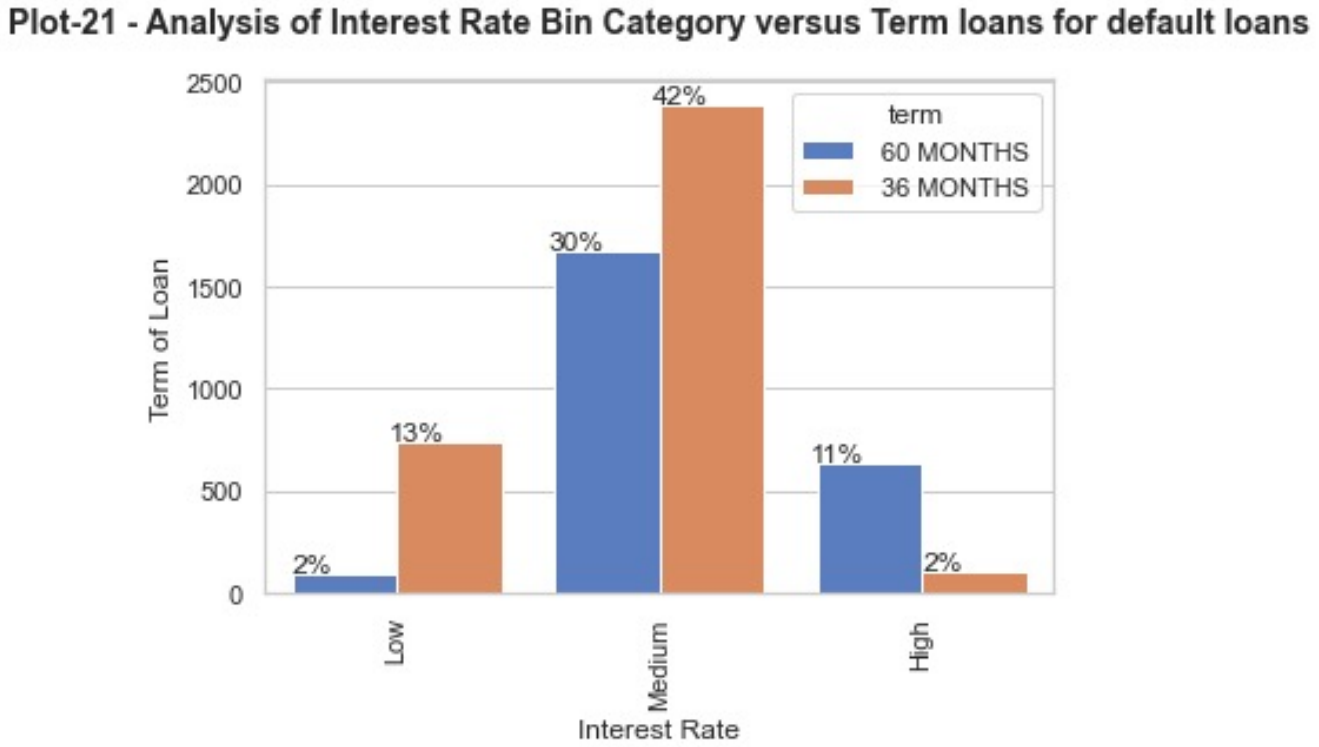
Analysis of Annual Income Group for the Verification Status against Number of default Loans



Observations

- Indicate Maximum defaulters are either not verified in the annual income range of 30000 - 60000.
- This is a huge risk.

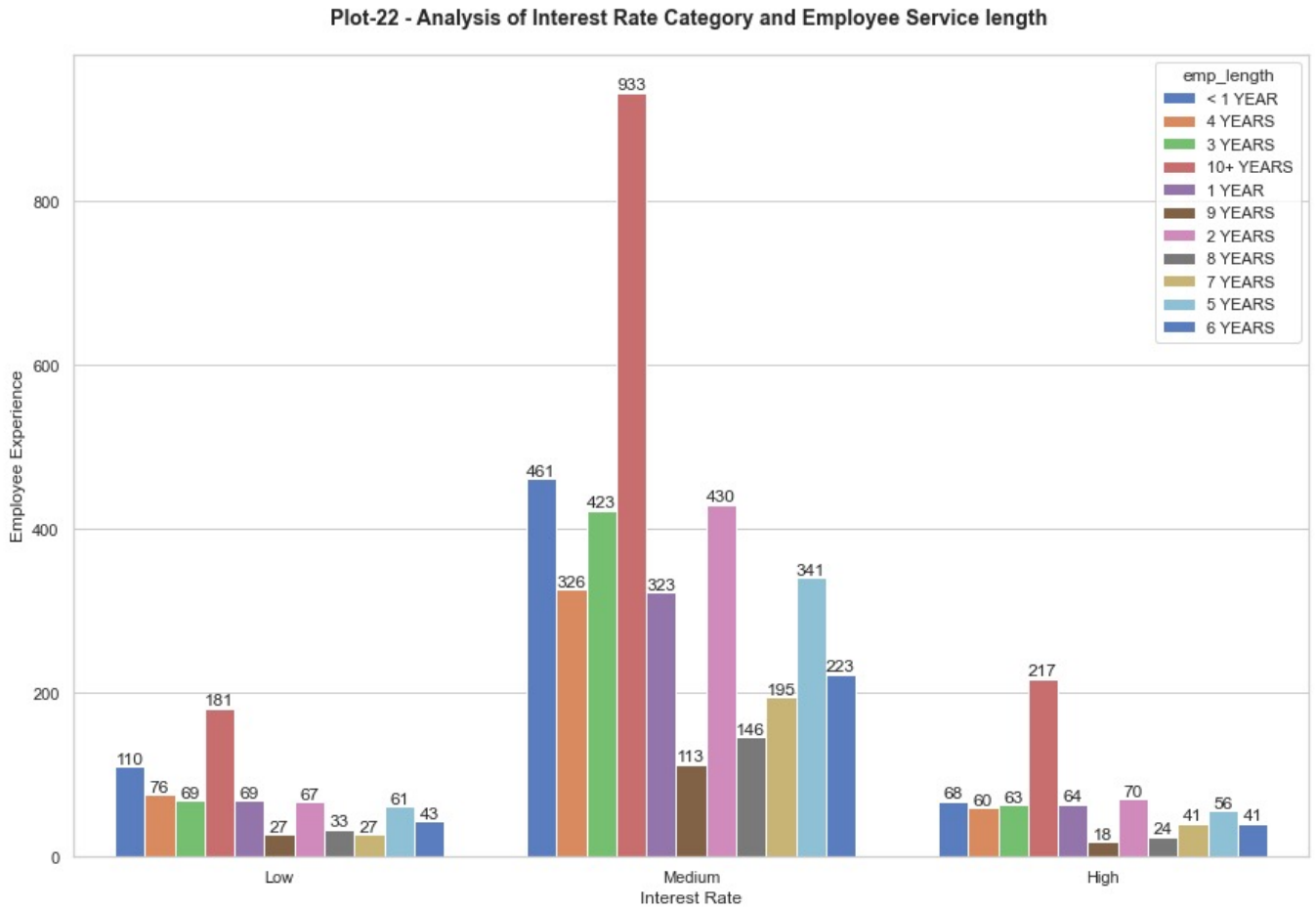
Analysis of Interest Rate Bin Category versus Term loans for default loans



Observations

- Indicate Maximum defaulters are in Medium Interest rate i.e., between `10 to 18%`.
- Out of 72% , `42% have taken loan term as 36 Months` and `30% as 60 Months`.

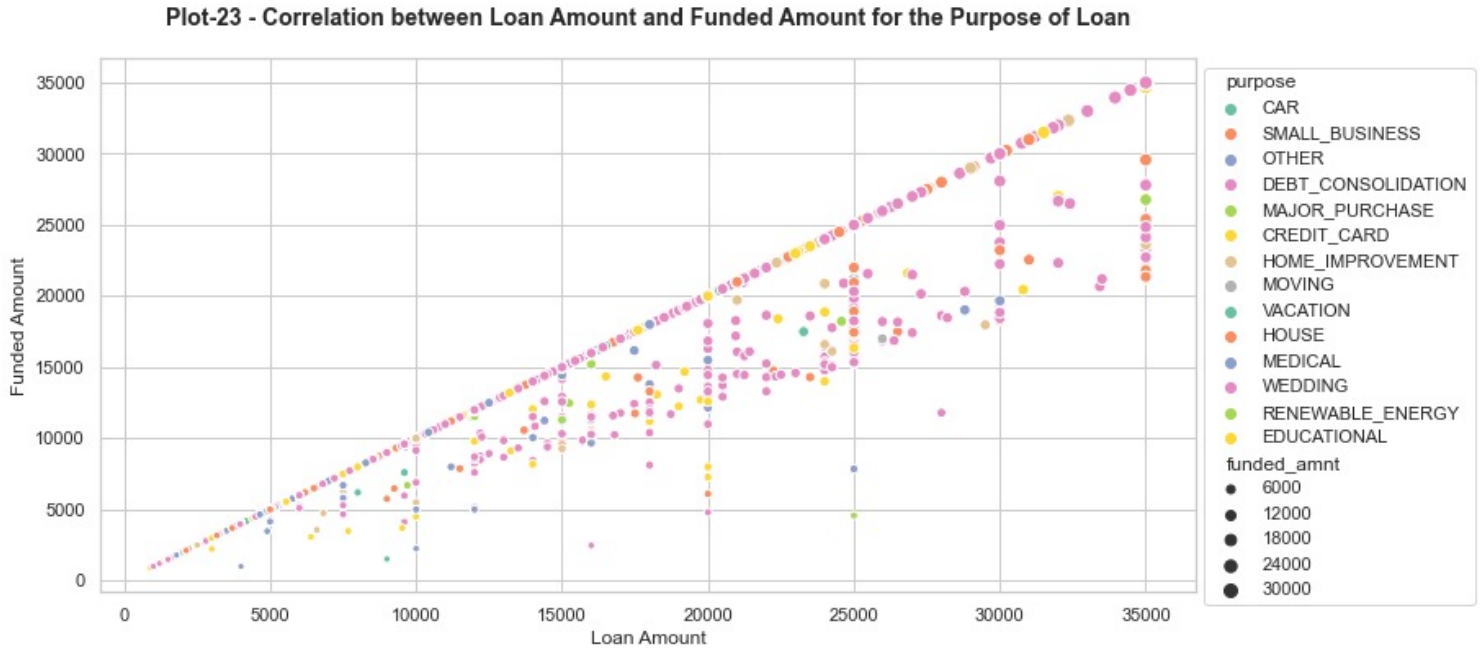
Analysis of Interest Rate Category and Employee Service length



Observations

- Indicate Employees having service more than 10+ years have taken the Most Medium Interest rate Loans

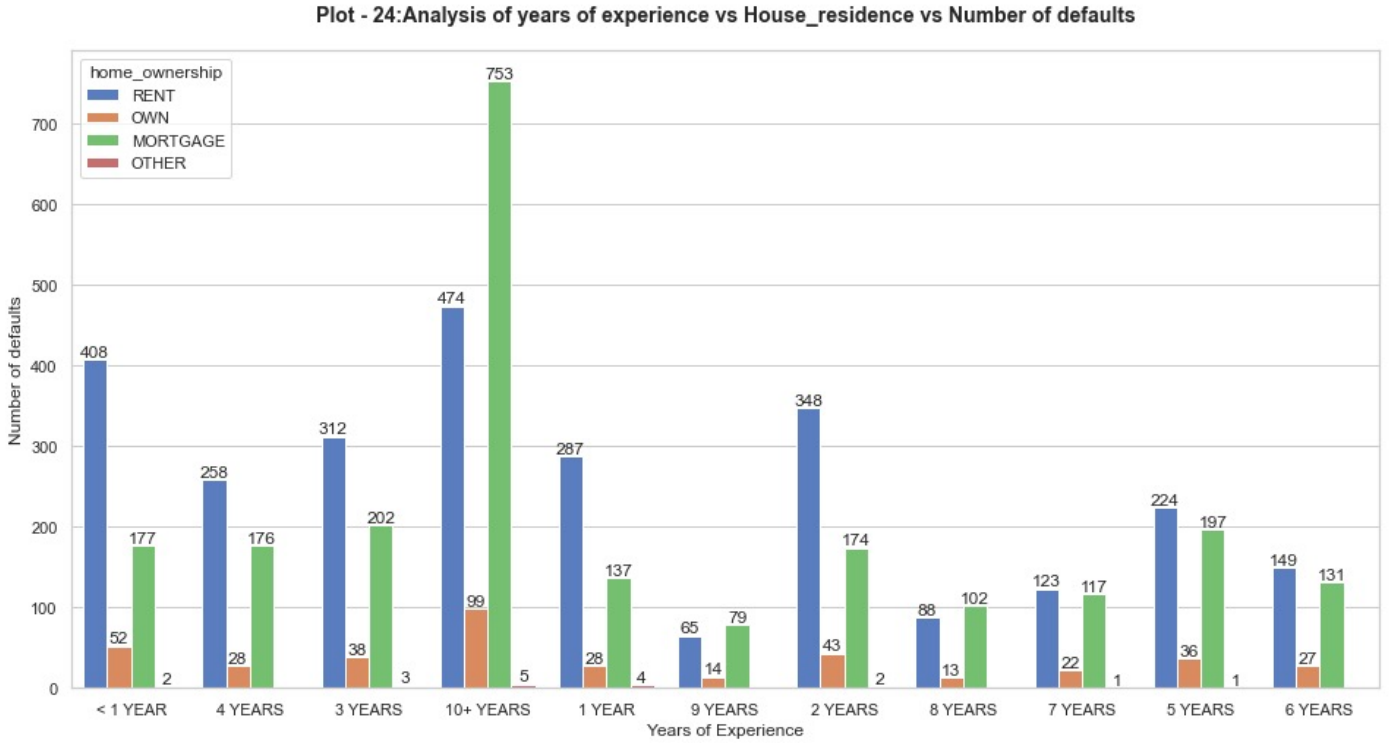
Correlation between Loan Amount and Funded Amount for the Purpose of Loan



Observations

- Indicate Positive correlation between `loan_amn`t and `funded amount` and it is linear.
- It clearly shows that with Increase in Loan Amount the funded amount increases
- `Maximum funded Amount` is for `Debt Consolidation` indicated by circle marked larger and Pink color.

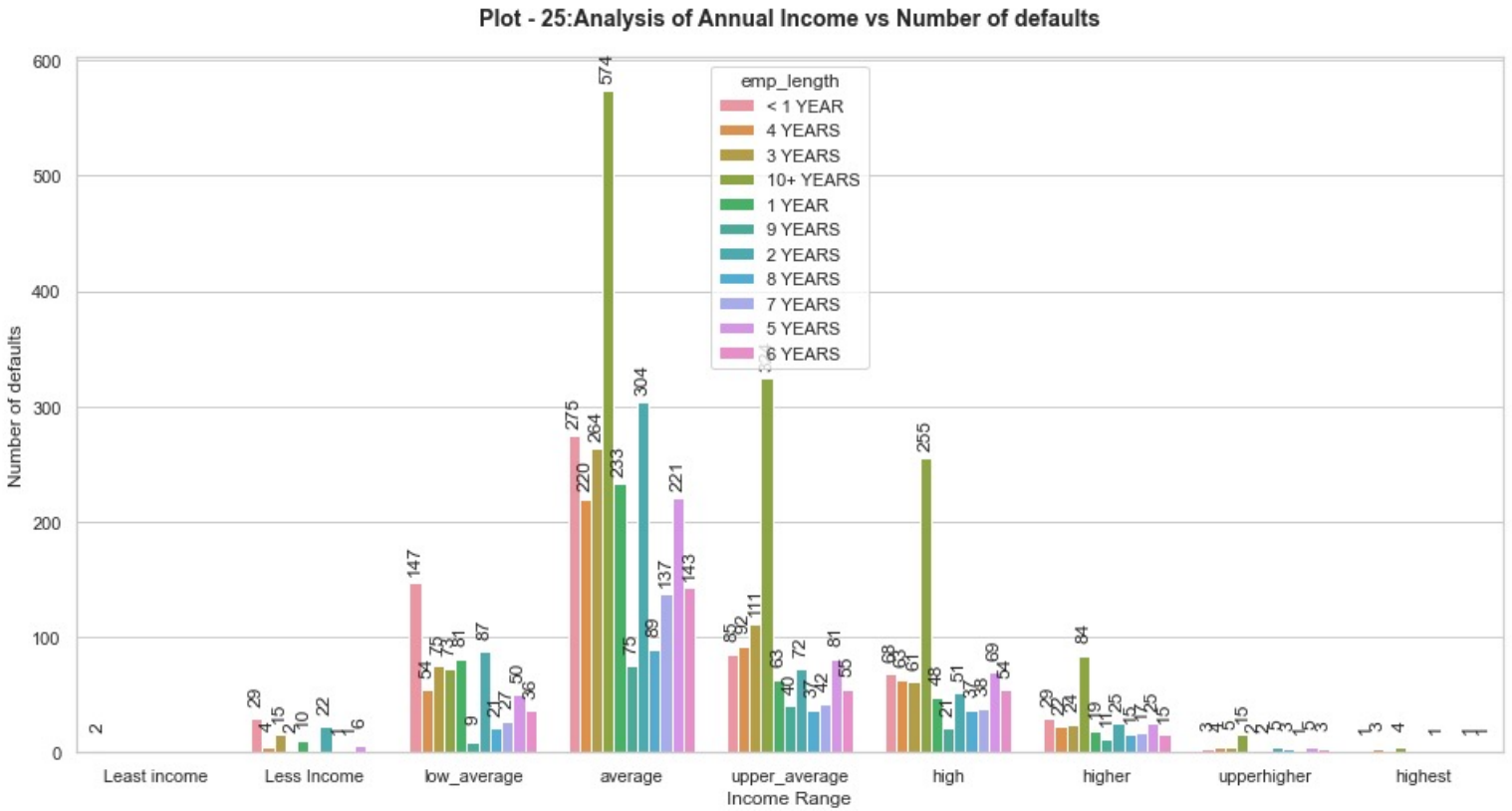
Analysis of years of experience vs House residence vs Number of defaults



Observations

- Indicate applicants who dont own a House after 10 or more years of Experience have a high default rate.

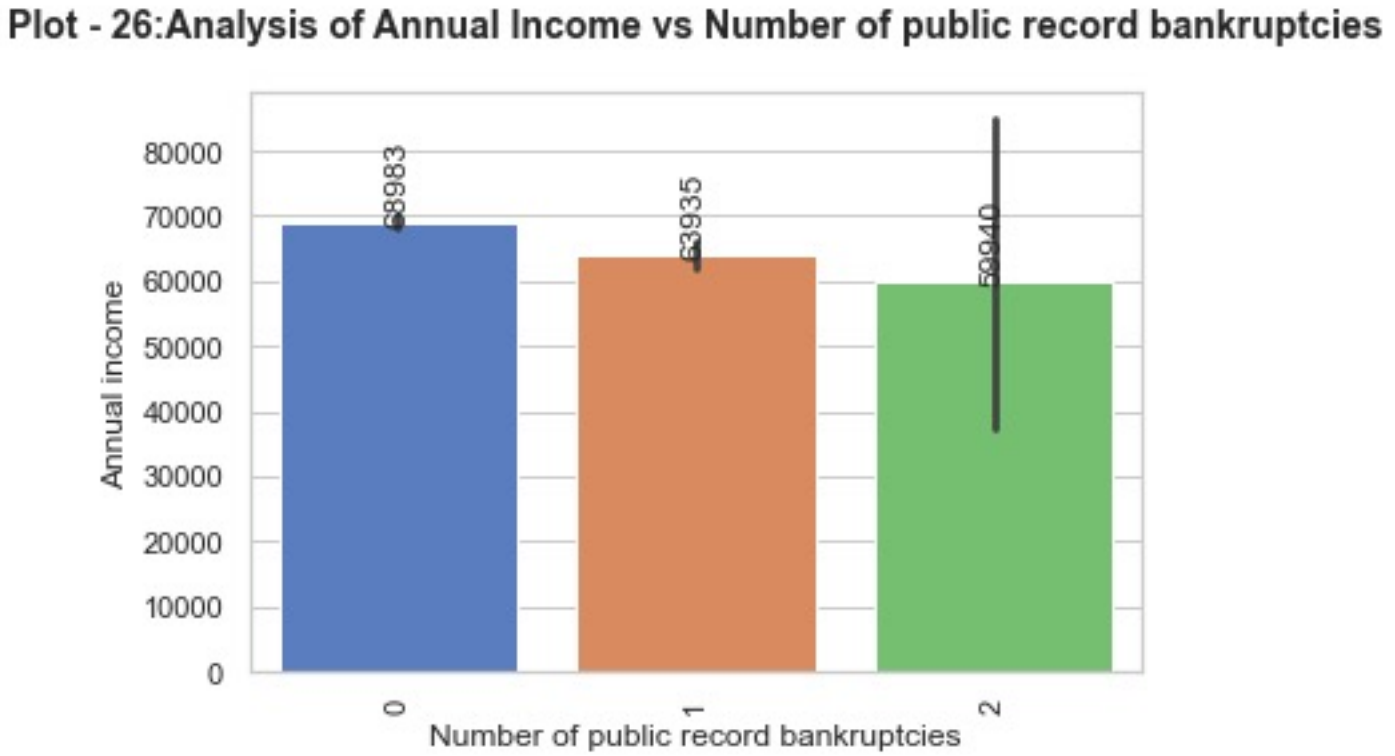
Analysis of Annual Income vs Number of defaults



Observations

- Indicate employee in the Average income zone(30000 - 60000) having experience >= 10 years are peak Defaulters.

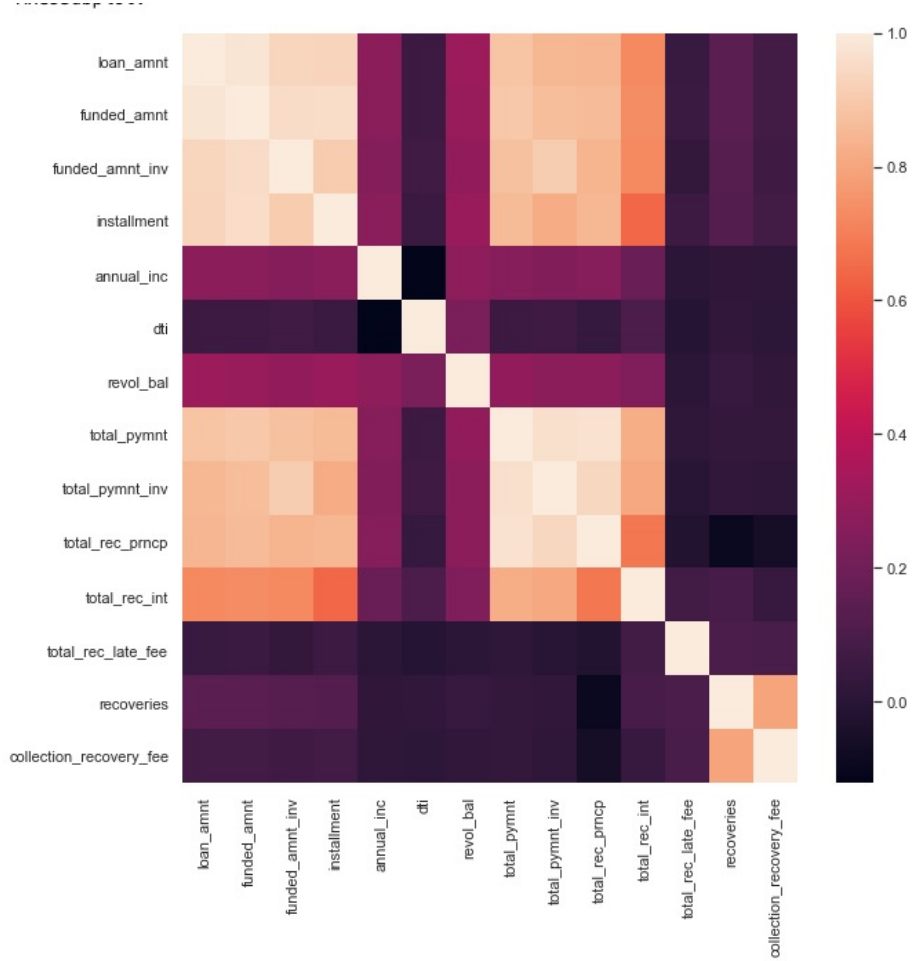
Analysis of Annual Income vs Number of public record bankruptcies



Observations

- Indicate employee in the Average income zone(30000 - 60000) having experience >= 10 years are peak Defaulters.

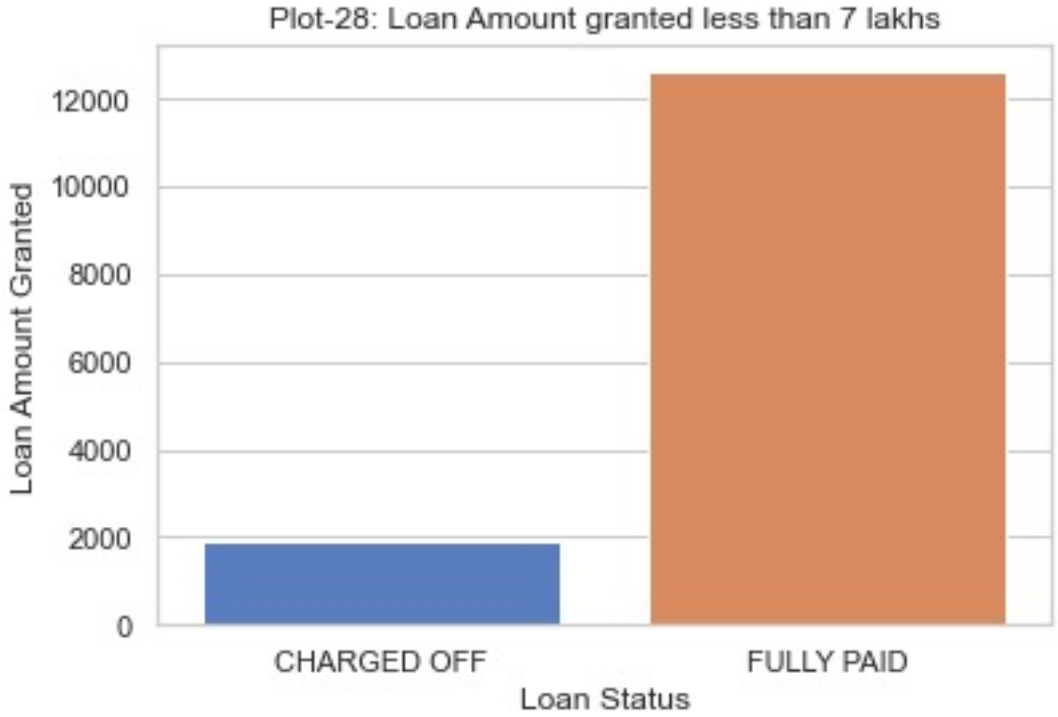
Bivariate Correlation Analysis.



Observations

- Indicate From the heat map have picked the Topmost Correlated Variables
- Loan Amount , Funded Amount
- Instalment, Funded Amount Invested
- Total Payment, Total Payment Invested
- pub_rec,pub_rec_bankruptcies. (Deleted in above graph)

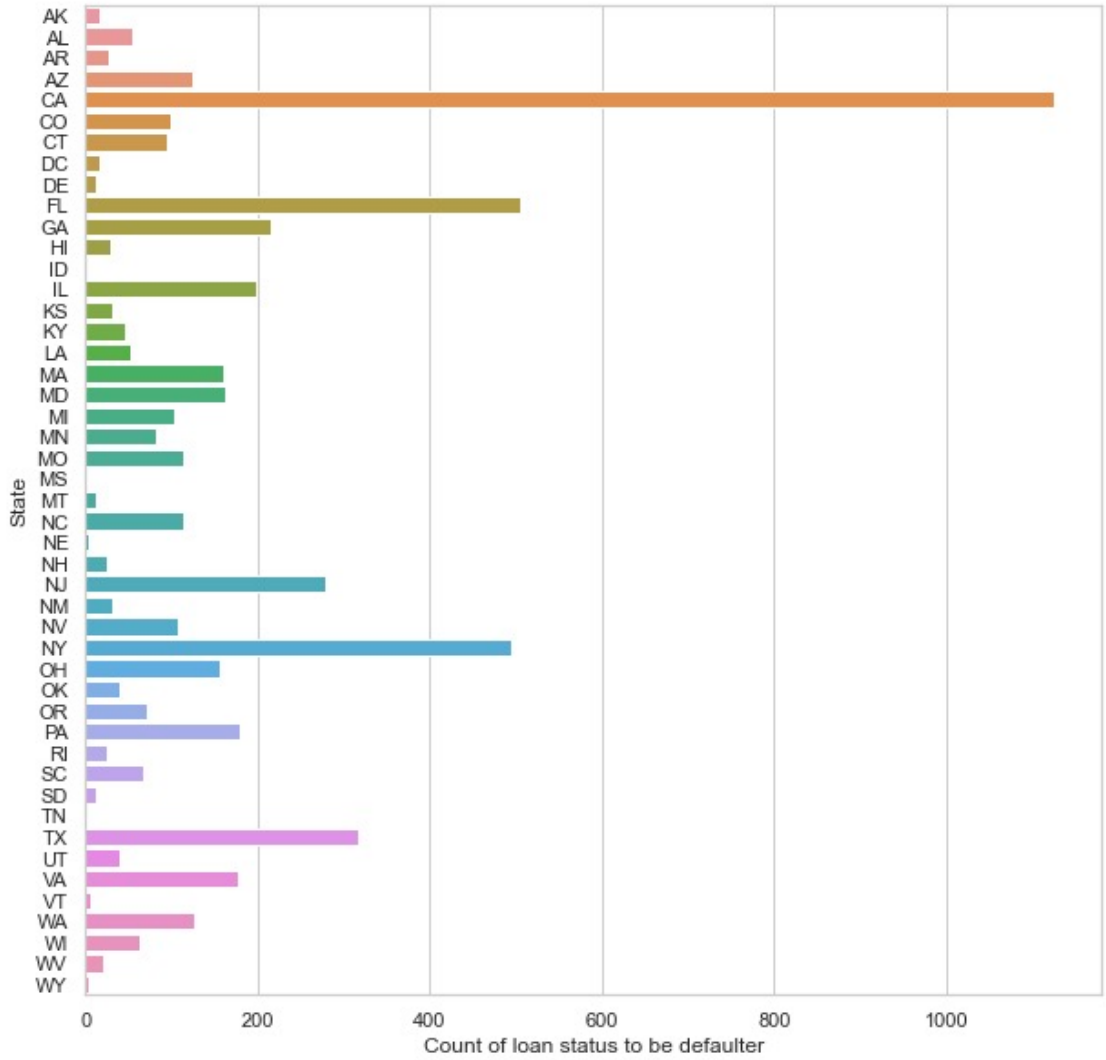
Loan Amount granted less than 7 lakhs



Observations

Funded amount is left skewed, most of the loan amount given < 7L

State wise Defaulter



Observations

The applicant belonging to CA state, need more scrutiny must be done, as tendency to default is high

Conclusions

- Grading system is working as expected, the Low-grade loans have high tendency to default.
- People with records of bankruptcies should be denied loan.
- Higher interest rate loan has more defaulter. Always need to check the background of applicant thoroughly whenever interest rate is high.
- Small businesses are particularly risky, therefore should be cautiously sanctioned.
- Individuals who have more open credit lines are less likely default, therefore number of open credit lines is a good indicator of financial safety.
- When the purpose is debt consolidation, more scrutiny of applicant as has high tendency to default.
- People with records of bankruptcies should be denied loan. individuals may be better strategy than approving their loan with higher interest.
- The applicant belonging to CA state, need more scrutiny must be done, as tendency to default is high