LENDING CLUB CASE STUDY

Deepak Mohanta
Mahak Porwal

SLIDE SUMMARY

DATA UNDERSTANDING

The objective of this section is to identify and report data quality issues and interpret and report the meanings of the variables as required

DATA CLEANING

The objective of this section is to ensure data quality, convert data into suitable formats as needed to facilitate effective data analysis and processing

DATA MANIPULATION

The objective of this section is to convert data into suitable formats as needed and perform accurate manipulation of strings and dates to facilitate effective data analysis and processing

DATA ANALYSIS

This section deals with data analysis using various methods

DATA ANALYSIS : BIVARIATE ANALYSIS

The objective of this section is to conduct bivariate analysis and identify combinations of driver variables relevant to the business objective

DATA ANALYSIS : SEGMENTED ANALYSIS

RECOMMENDATIONS

DATA UNDERSTANDING

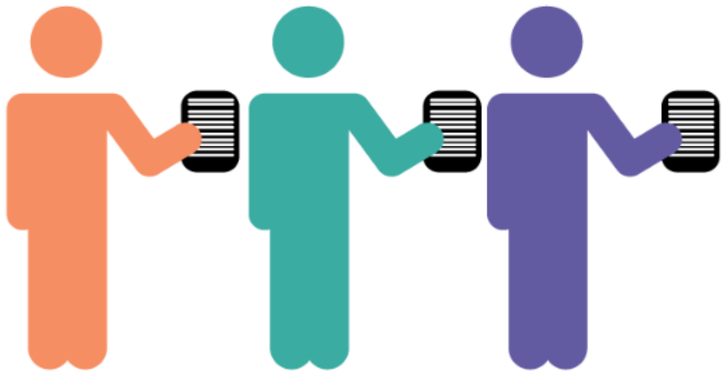
The objective of this section is to identify and report data quality issues and interpret and report the meanings of the variables as required

BUSINESS OBJECTIVE & DATA UNDERSTANDING

- We are given a loan dataset from year 2007 to 2011 covering the details of all loan applicants.
- The objective is to identity the loan or customer attributes which could become or are a driving factor in predicting whether a loan application can become defaulter.
- All the attributes mentioned in the loan dataset can be understood with the help of Data Dictionary.
- The Loan Status attribute would be our pivot henceforth. It is categorized as Fully Paid, Current and Charged Off. Here "Charged Off" delegates to defaulters who would have either refused to pay or ran away with the money owed.

BUSINESS OBJECTIVE & DATA UNDERSTANDING

LOAN DATASET



Loan Accepted → Default

Loan Accepted → Non-Default

Loan Rejected

(Not considered in dataset)

DATA QUALITY ISSUES OBSERVED

For this assignment, we have diligently conducted a comprehensive data quality assessment using a well-structured checklist. We systematically evaluated each column, applying a consistent set of methods and criteria. Our meticulous process involved a thorough examination of each column's data quality, and our observations and findings have been meticulously documented on the subsequent slides for your review.

Method	Data quality issue checked
Fix rows and columns	There are no incorrect rows
	There are no summary rows
	There are no extra rows
	There are no missing column names
	There are no inconsistent column names
	Unnecessary columns are present and contains only one cell value like 'NA', '1', '0' or. These columns are listed out in the next section and excluded from the analysis.
	There were no columns containing Multiple data values like city, state and town etc.
	All the columns have unique identifier
	There no misaligned columns. All columns are perfectly aligned.

DATA QUALITY ISSUES OBSERVED

Method	Data quality issue checked
Missing Values	There are columns with missing values represented as 'Blank' & 'NA'. These columns have been identified and listed down in the next section. For few columns appropiate imputation has been done.
	Partial missing values
Standardise Numbers	No column contain non-standard units.
	There are no values with varying scales
	Yes there are overprecision. All these columns have been identified in the next section. To avoid over precision, I've restricted the cell values to whole integers. This helps in comprehension.
	Outliars has been identified for selected columns and have been removed not to impact the analysis in data cleaning section.
Standardise Text	No extra characters were found in any of the column data.
	No different cases of same words e.g. lower and upper cases of the same word present
	Non-standard formats were not found.
Fix Invalid Values	No encoding issues were found.
	Incorrect data types for columns have been identified and fixed in the subsequent section.

DATA CLEANING

The objective of this section is to ensure data quality, convert data into suitable formats as needed to facilitate effective data analysis and processing

SUMMARY DATA CLEANING

- Missing Data Identification:
  - Uncovered rows or columns with missing values.
  - Discovered rows or columns with over 50% of their values as null.
- Uniformity Observation:
  - Identified columns with uniform values across all rows.
  - Recognized columns that provide no analytical value and can be excluded from further analysis.
- Redundant and Unique Columns:
  - Detected columns with redundant or unique information, such as "title," "url," "desc," etc., which may not contribute to the analysis goals.

DATA CLEANING

- The following list of columns contains only ‘NA’ as cell value.
- These columns are not included in the current analysis.

dti_joint	avg_cur_bal	num_rev_tl_bal_gt_0
verification_status_joint	bc_open_to_buy	num_sats
tot_coll_amt	bc_util	num_tl_120dpd_2m
tot_cur_bal	mo_sin_old_il_acct	num_tl_30dpd
open_acc_6m	mo_sin_old_rev_tl_op	num_tl_90g_dpd_24m
open_il_6m	mo_sin_rcnt_rev_tl_op	num_tl_op_past_12m
open_il_12m	mo_sin_rcnt_tl	pct_tl_nvr_dlq
open_il_24m	mort_acc	percent_bc_gt_75
mths_since_rcnt_il	mths_since_recent_bc	tot_hi_cred_lim
total_bal_il	mths_since_recent_bc_dlq	total_bal_ex_mort
il_util	mths_since_recent_inq	total_bc_limit
open_rv_12m	mths_since_recent_revol_delinq	total_il_high_credit_limit
open_rv_24m	num_accts_ever_120_pd	annual_inc_joint
max_bal_bc	num_actv_bc_tl	mths_since_last_major_derog
all_util	num_actv_rev_tl	
total_rev_hi_lim	num_bc_sats	
inq_fi	num_bc_tl	
total_cu_tl	num_il_tl	
inq_last_12m	num_op_rev_tl	
acc_open_past_24mths	num_rev_accts	

DATA CLEANING

- The following columns are not included in the analysis.
- Few of the columns contained very insignificant number of NA values
- On the other hand, other columns contain unstructured text. To keep the analysis simple, we have not included this column in our analysis.
- There are four columns which contains only single value which were also not included in the current analysis

Columns Removed	Reason
chargeoff_within_12_mths	56 NA values insignificant as compared to 39k other '0' values
tax_liens	39 NA values insignificant as compared to 39k other '0' values
collections_12_mths_ex_med	56 NA values insignificant as compared to the other '0' values
next_pymnt_d	1140 '16-June' values insignificant as compared to the other 39k 'Blank' values
desc	contains unstructured text
emp_title	contains unstructured text
title	contains unstructured text
initial_list_status	contains only single value 'f'
application_type	contains only single value 'single'
policy_code	contains only single value '1'
pymnt_plan	contains only single value 'n'

DATA MANIPULATION

The objective of this section is to convert data into suitable formats as needed and perform accurate manipulation of strings and dates to facilitate effective data analysis and processing

DATA MANIPULATION

- Extracted year and month information from the existing date column.
- Why is this required ? This is required to check whether there is a significant influence of year and month data on the default rate.

Extracting year & month information from a single column	Created New Columns
last_credit_pull_d	last_credit_pull_d_year
	last_credit_pull_d_month
last_pymnt_d	last_pymnt_d_year
	last_pymnt_d_month
earliest_cr_line	earliest_cr_line_year
	earliest_cr_line_month

- Removed % from Interest Rate(int_rate) and Revolving Utilization(revol_util).
- Manupulated emp_length to have just digits and filled the blank values with 0.
- Filled pub_rec_bankruptcies column blank values with Not Known.

DERIVED COLUMNS

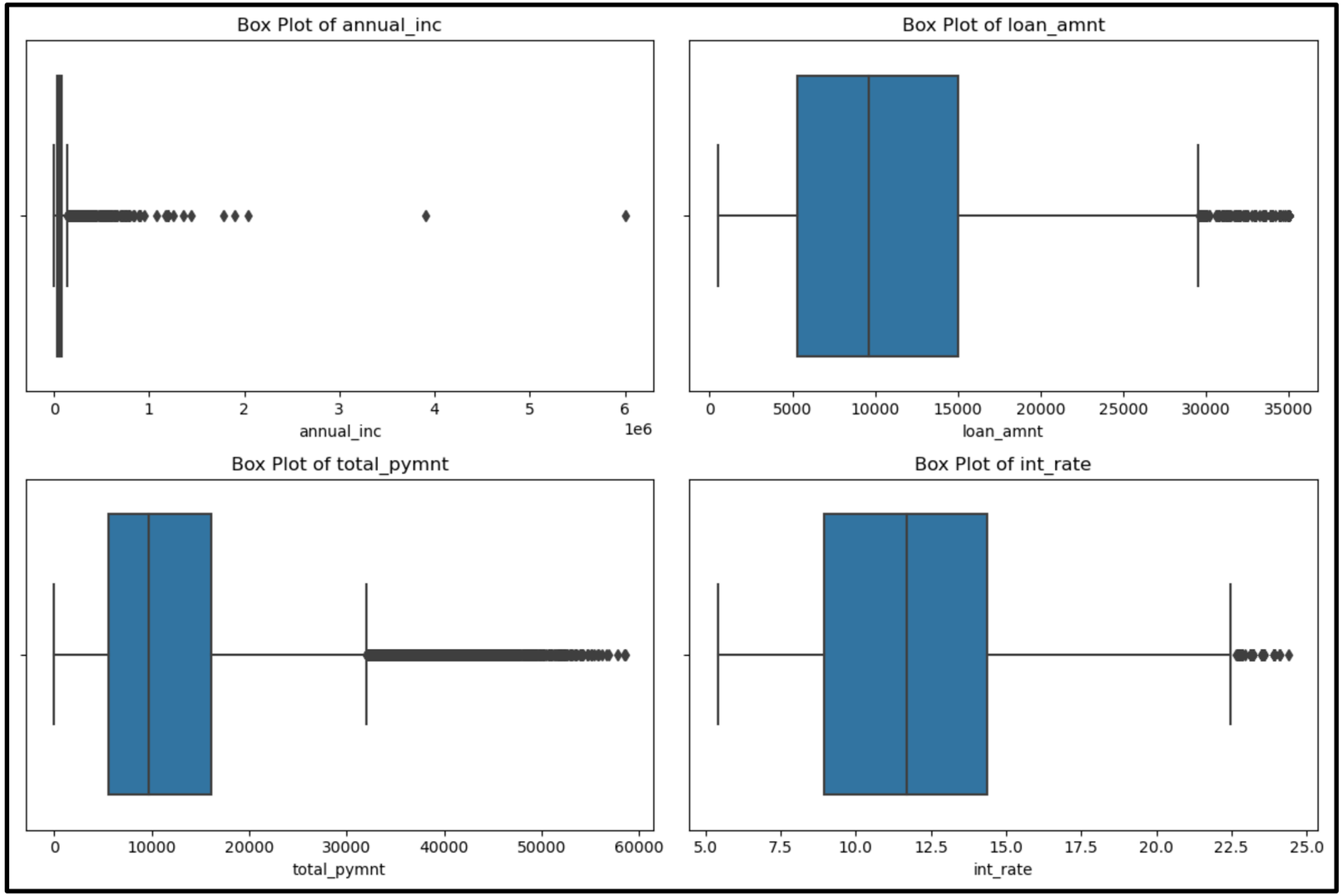
- Extracted month and year date from the following columns
  - last_credit_pull_d
  - last_pymnt_d
  - earliest_cr_line
  - issue_d
- Created Bins for the following columns
  - loan_amnt
  - annual_inc
  - int_rate
  - dti_categories

Extracting year, month information from a single column	Created New Columns
last_credit_pull_d	last_credit_pull_d_year
	last_credit_pull_d_month
last_pymnt_d	last_pymnt_d_year
	last_pymnt_d_month
earliest_cr_line	earliest_cr_line_year
	earliest_cr_line_month
issue_d	issue_d_year
	issue_d_month

DATA MANIPULATION : OUTLIER IDENTIFICATION

- List of columns for which the outliers analysis was done
  - annual_inc
  - loan_amnt
  - total_pymnt
  - int_rate
- Step – 1: Plot box plots to see if outliers exist or not
- Step – 2: If outliers exist then remove those outliers

- Observation/s and insight/s
  - The box plot clearly shows that there are outliers present in the the data sets



DATA ANALYSIS

This section deals with data analysis using various methods

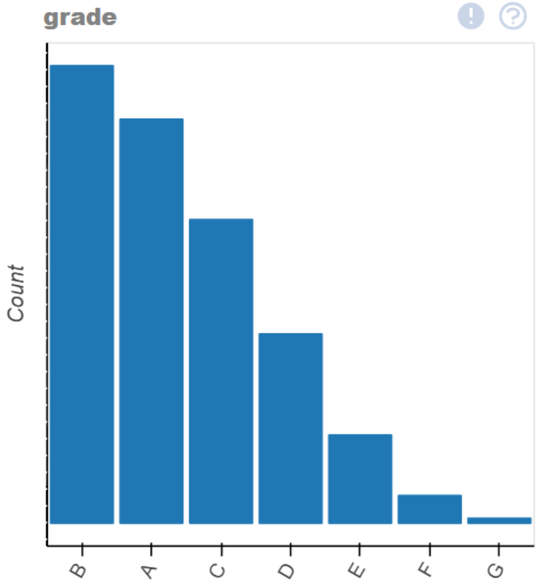
DATA ANALYSIS : UNIVARIATE & SEGMENTED UNIVARIATE

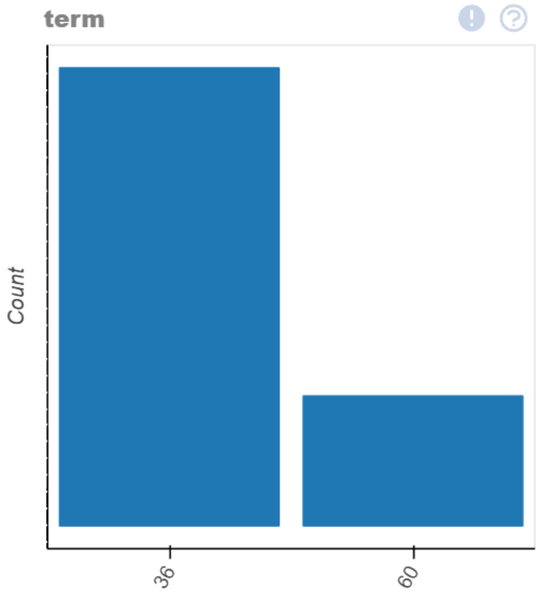
The objective of this section is to conduct univariate and segmented univariate analyses and identify 5 critical driver variables

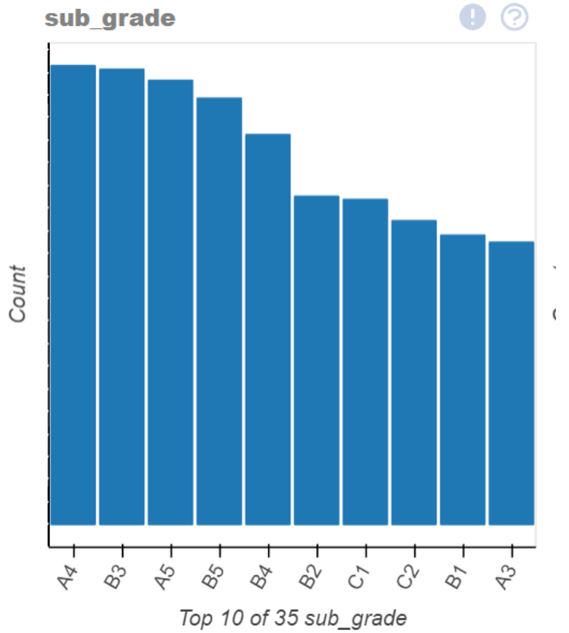
DATA MANIPULATION : OUTLIER IDENTIFICATION

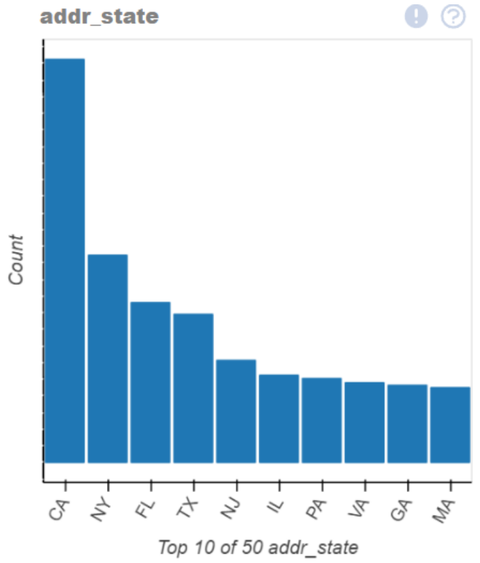
- Observation/s and insight/s. I have gone through the above plots to extract meaningful insights. The observations are captured below.
  - Highest number of loans are distributed loans have the ticket size in between 9900-10500.
  - 77.9% loans distributed have term of 36 months
  - The highest number of LC assigned loan grade are in B category
  - The highest Job titles belong to US Army and Bank of America
  - More than 80% of borrowers have either rent or mortgaged their home.
  - While 85% have fully paid their loans but 15% have singed off loan status.
  - 46% of borrowers stated debt consolidation as the purpose of taking loan the second category represents 13% stated credit card as purpose.
  - 945xx - Zip code represents the highest number of borrowers
  - 17.9% borrowers stated CA as their state

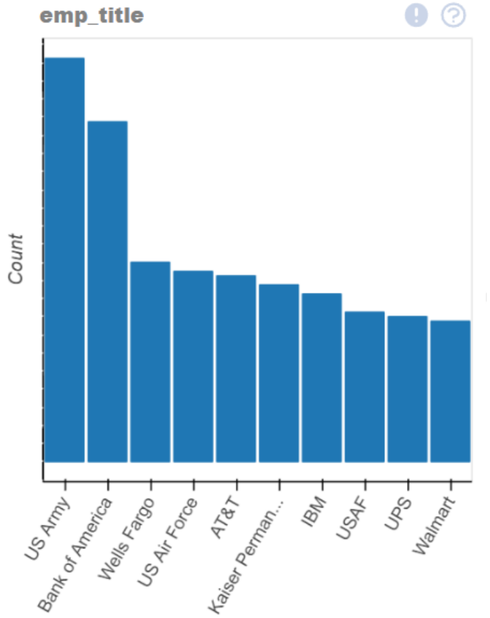
UNIVARIATE ANALYSIS

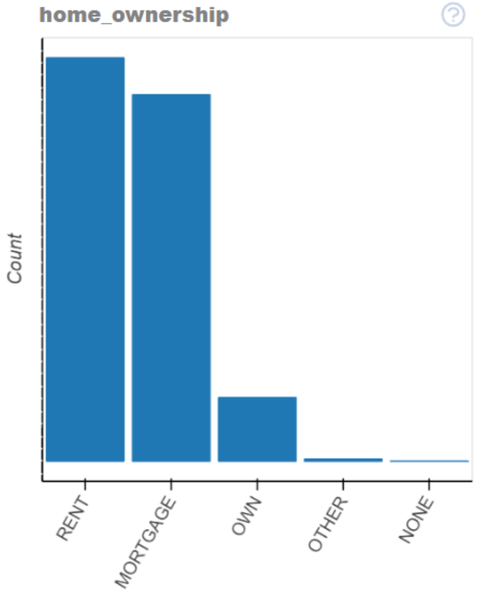


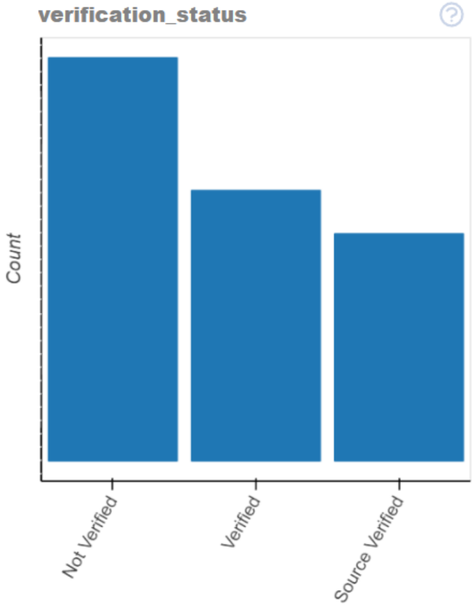


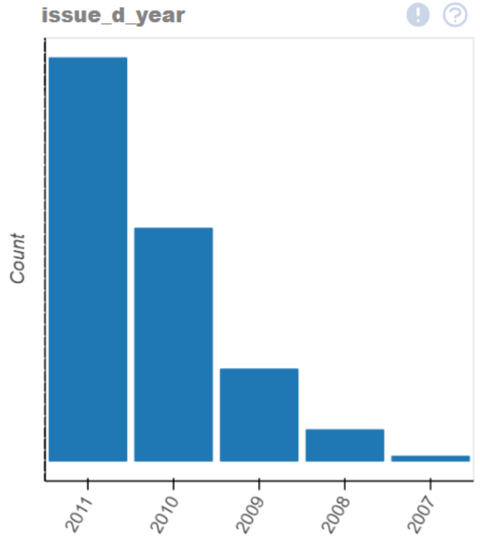




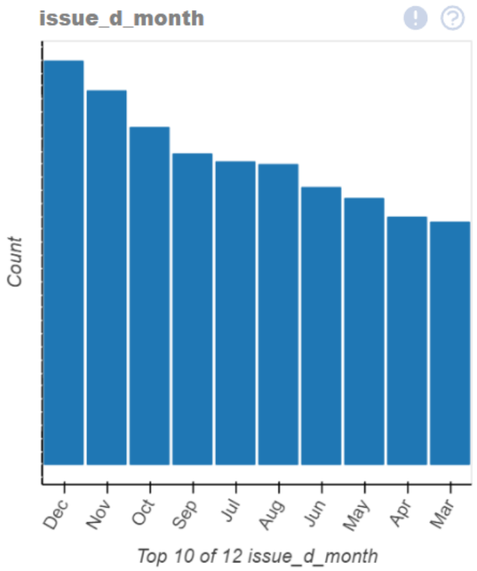


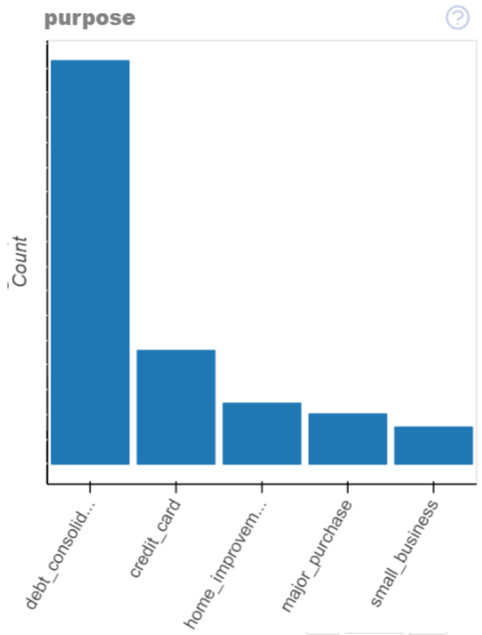


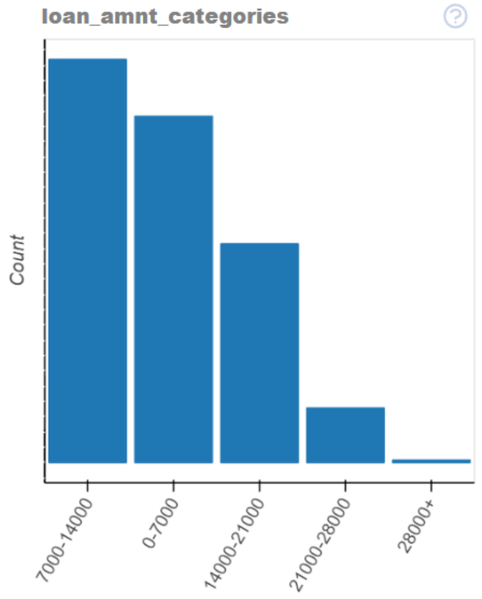


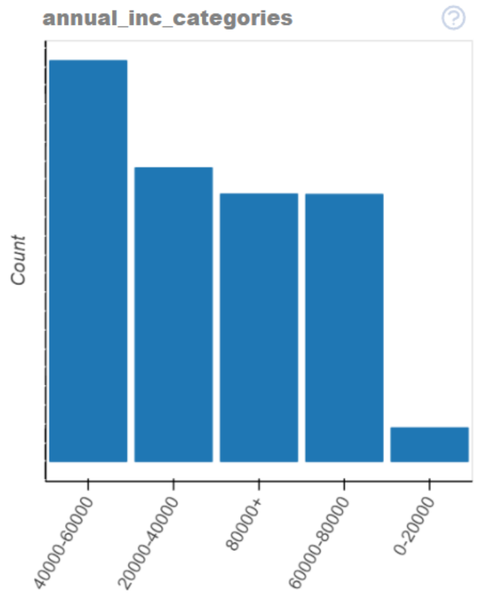


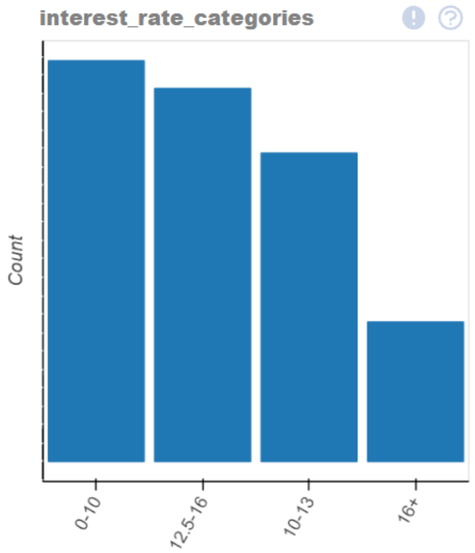
UNIVARIATE ANALYSIS

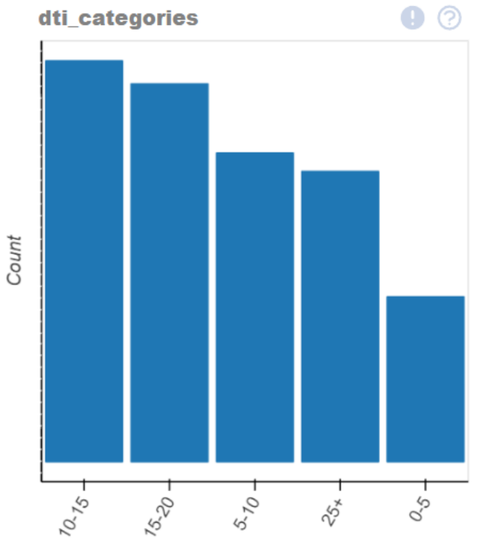






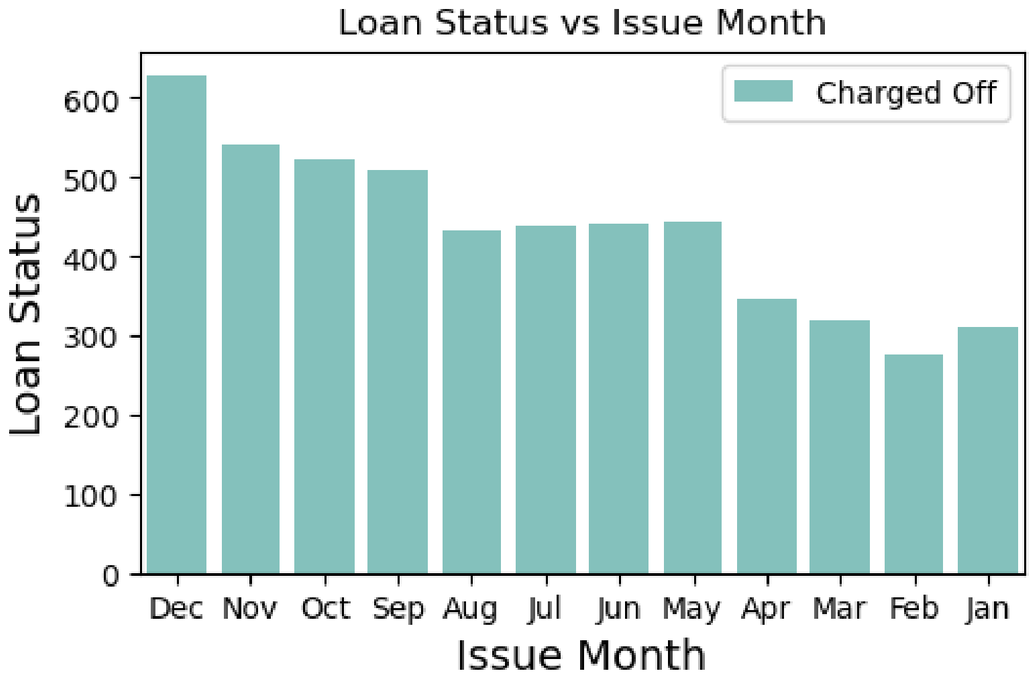


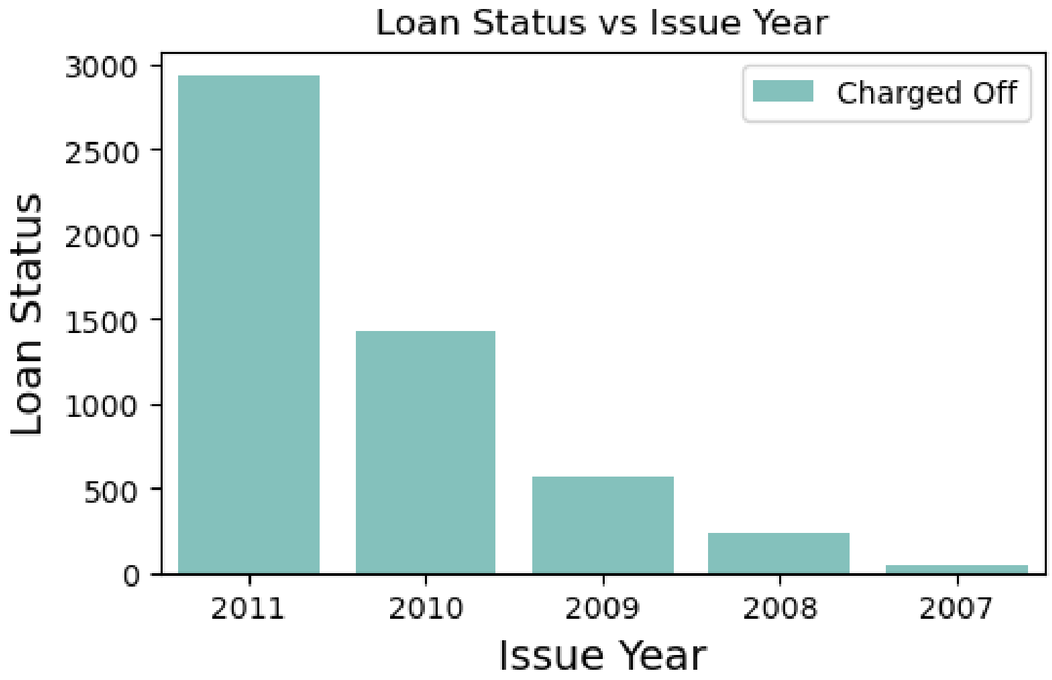




UNIVARIATE ANALYSIS : CHARGED OFFS VS M/Y

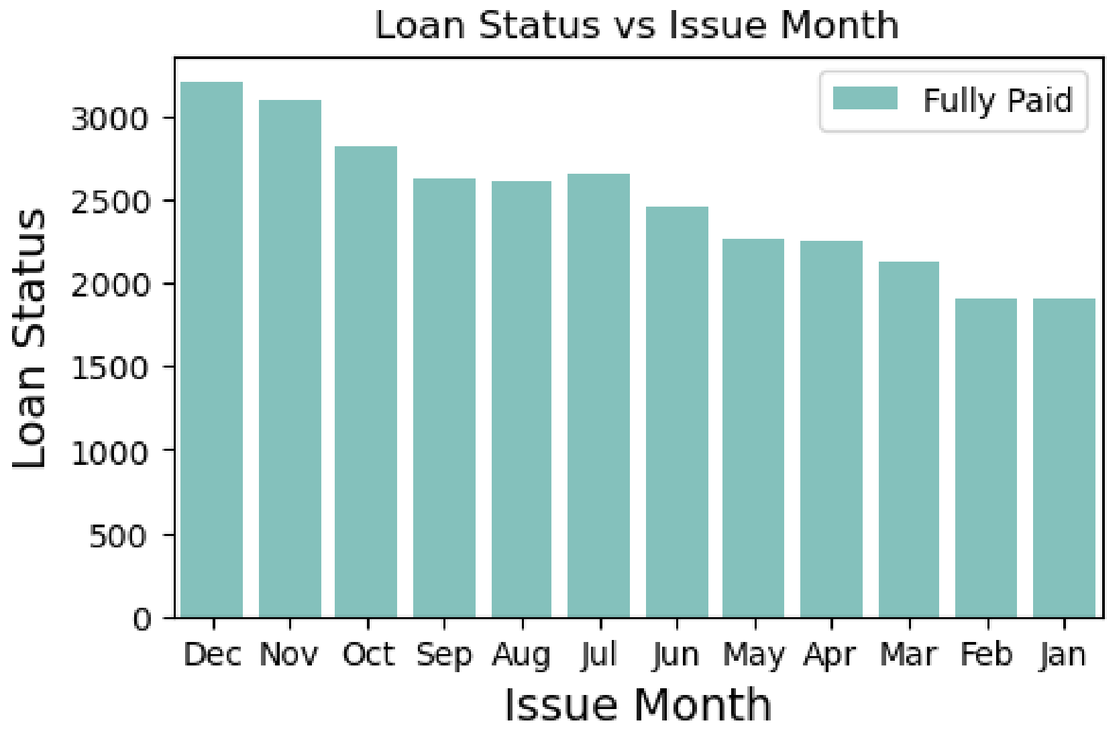
- Charged off loans are mostly issued in the last quarter of the year.
- highest number of charged off loans are issued in 2011. There is a stead increase in the charged off loans year after year.

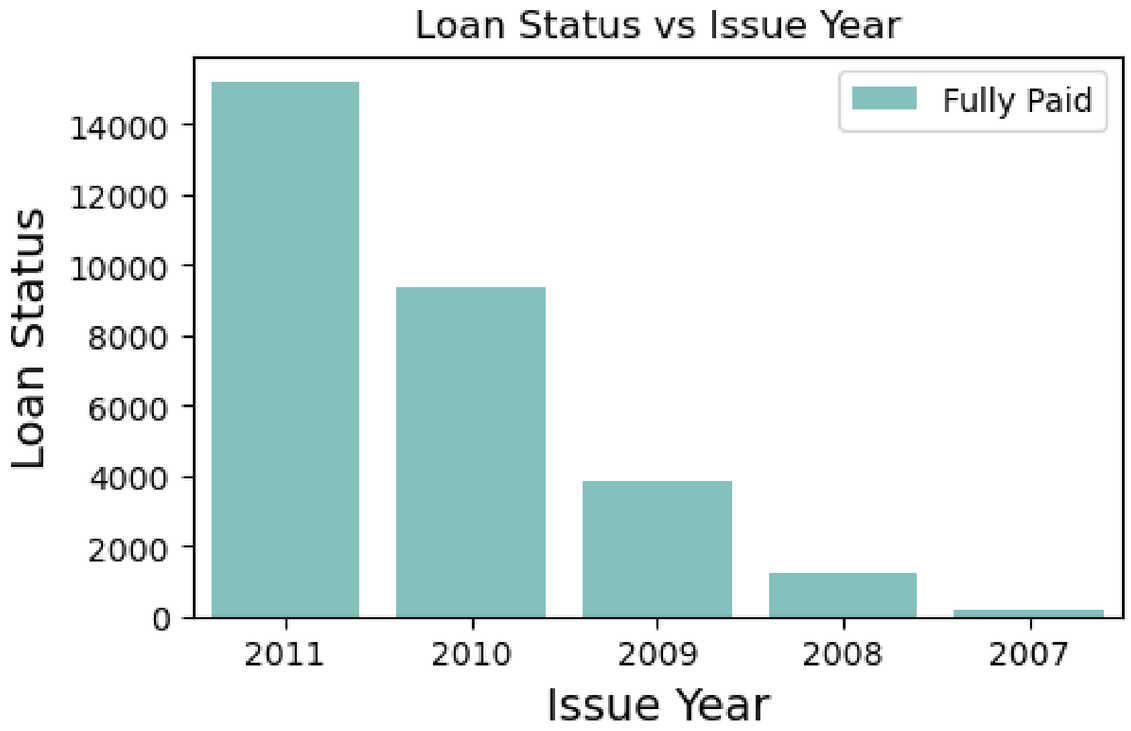




UNIVARIATE ANALYSIS : FULLY PAID VS M/Y

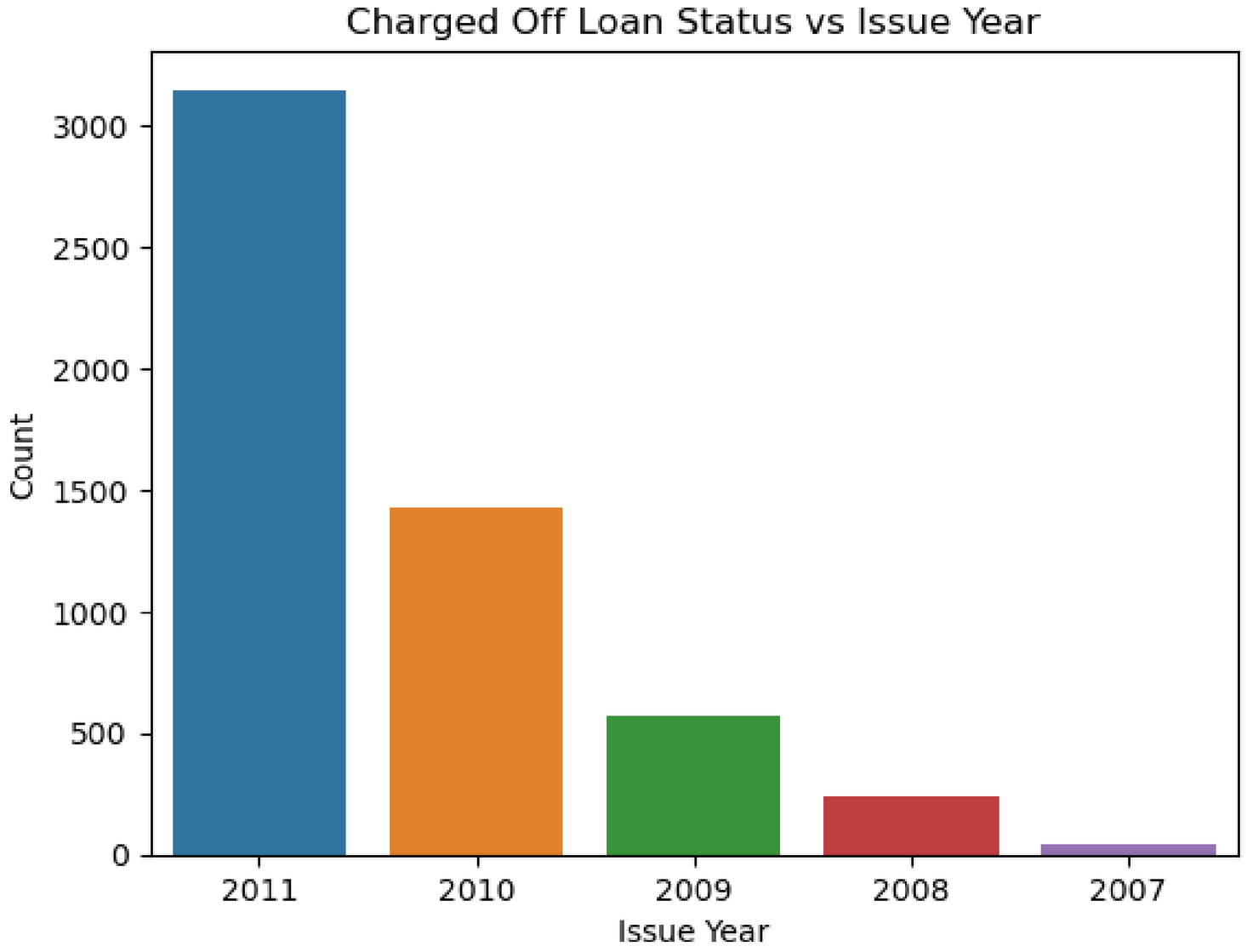
- Fully paid loans are mostly issued in the last quarter of the year.
- highest number of charged off loans are issued in 2011. There is a stead increase in the charged off loans year after year.
- No solid conclusion/s can be made since both fully paid and charged off loans have similar characteristics.





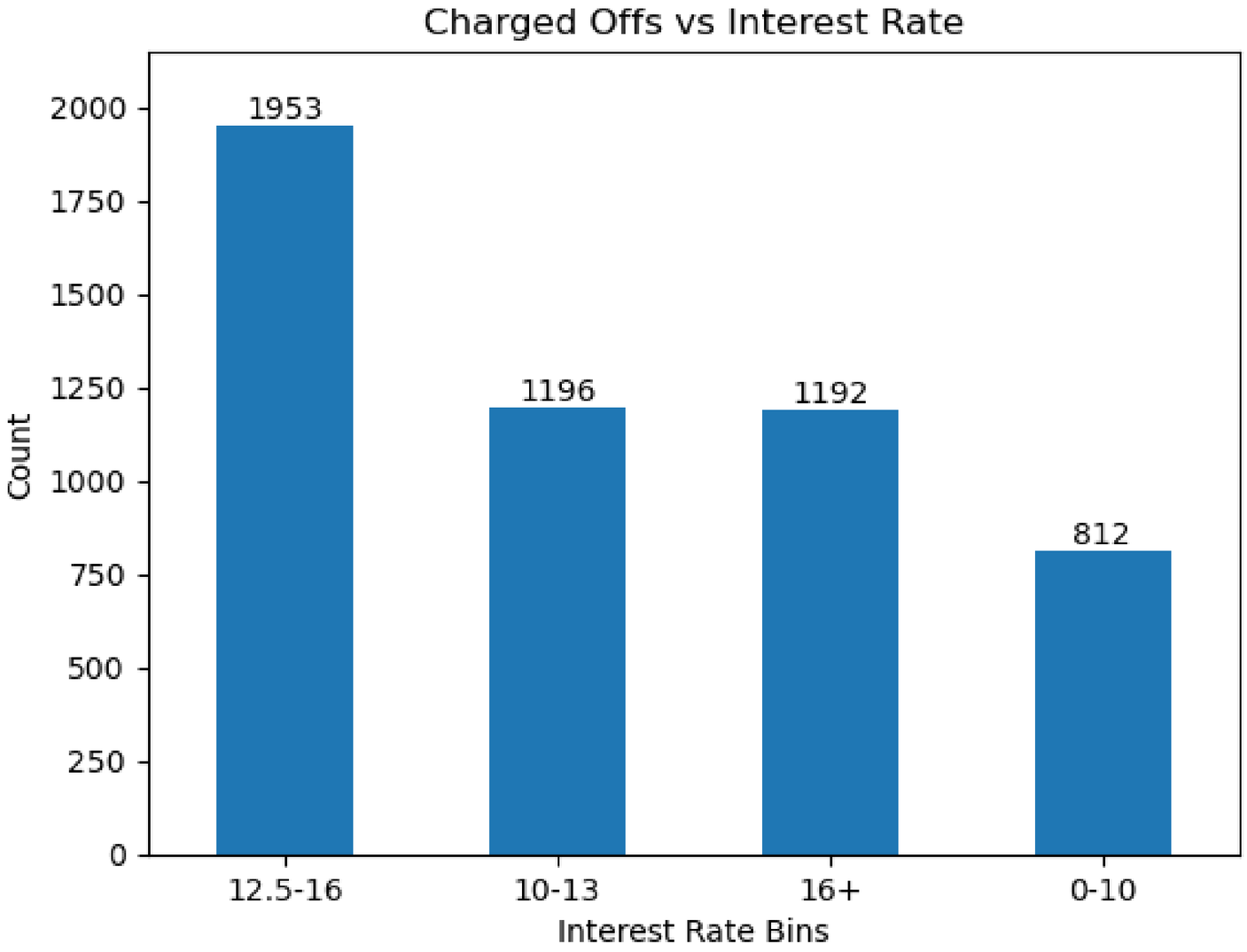
UNIVARIATE ANALYSIS : CHARGED OFFS VS ISSUE YEAR

- Observation: 2011 year observed major charged offs applicants



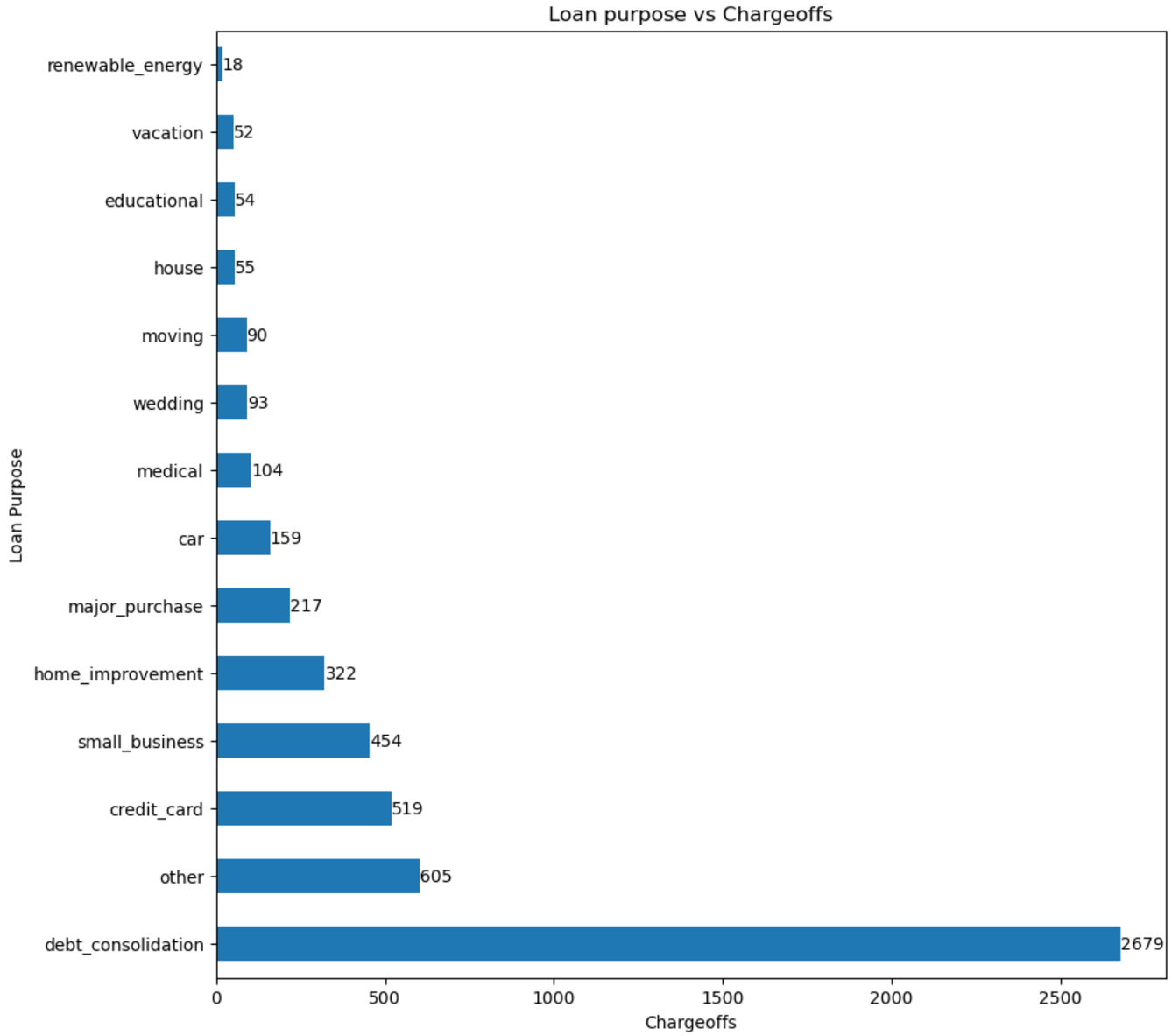
UNIVARIATE ANALYSIS : CHARGED OFFS VS INTEREST RATE

- Observation: Major Charge Offs are seen with Interest rate 12.5 to 16 percentage.



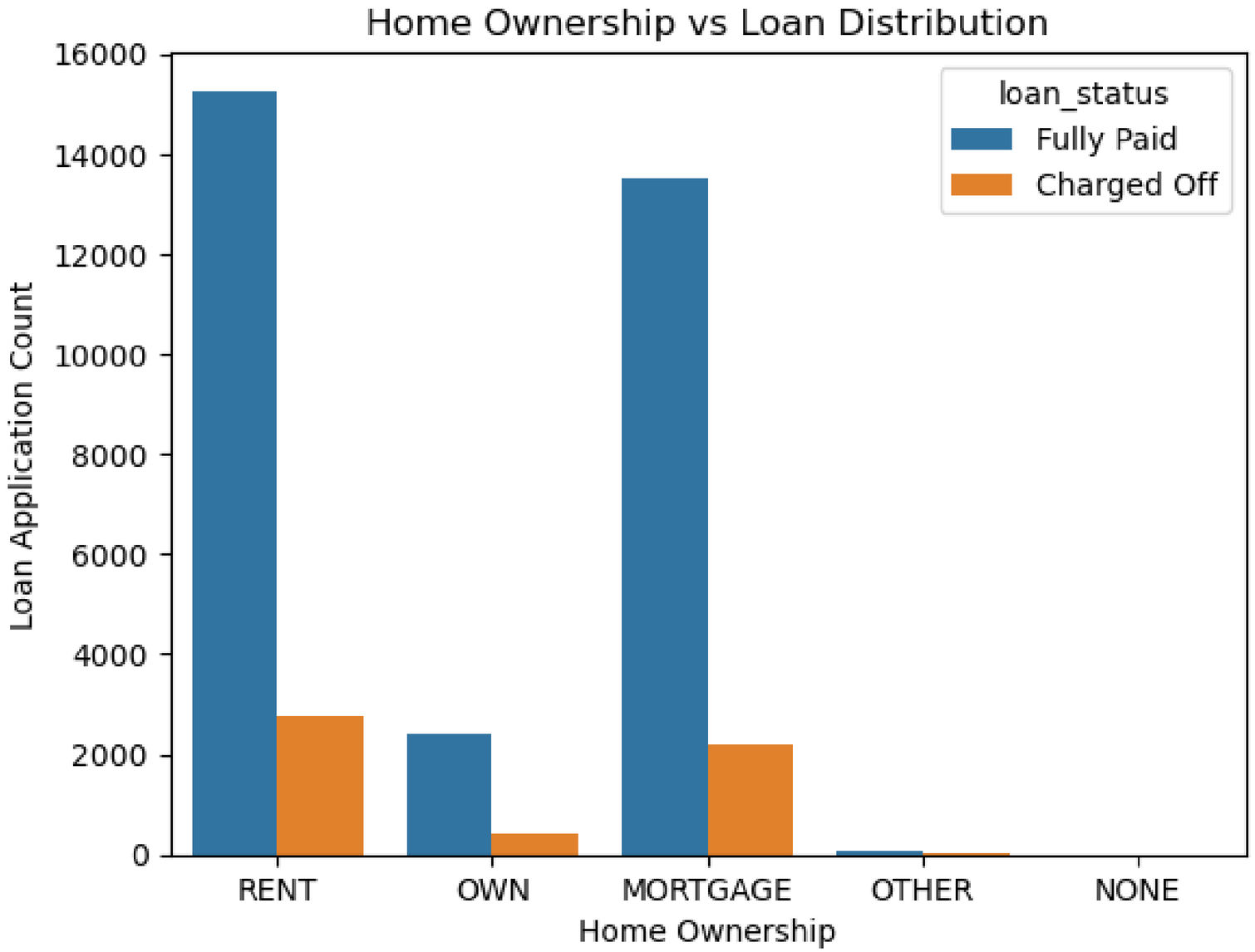
UNIVARIATE ANALYSIS: CHARGED OFFS VS PURPOSE

- Observation: Major Charge Offs are observed with Applicants with Purpose Debt Consolidation



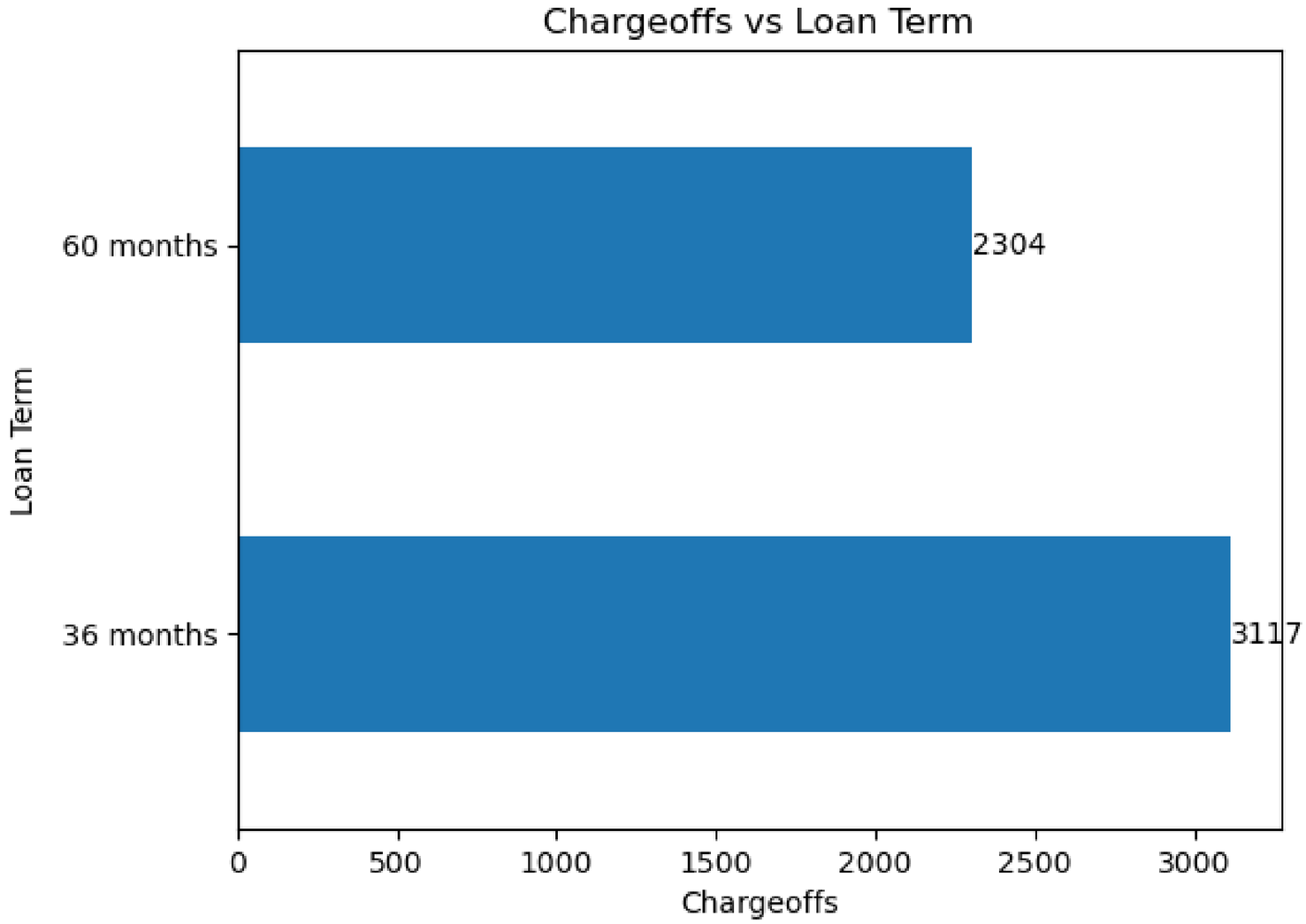
UNIVARIATE ANALYSIS : Charged Offs vs Home Ownership

- Observation: Major Charge Offs are seen with Rented and Mortgaged Ownership Applicants



UNIVARIATE ANALYSIS : CHARGED OFFS VS TERM

- Observation: Major Charge Offs are seen with Applicants taking 36 Months as the loan term

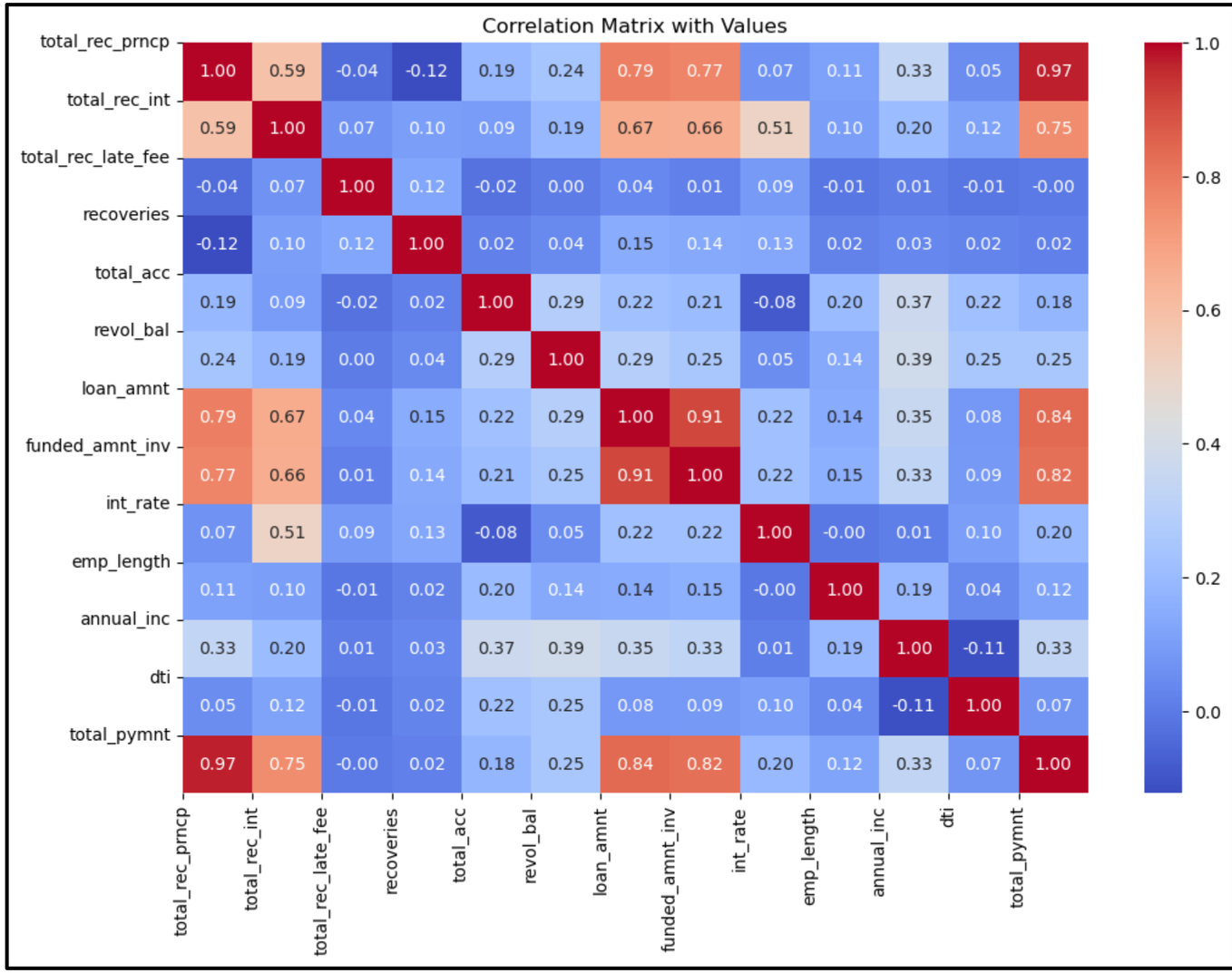


DATA ANALYSIS : BIVARIATE ANALYSIS

The objective of this section is to conduct bivariate analysis and identify combinations of driver variables relevant to the business objective

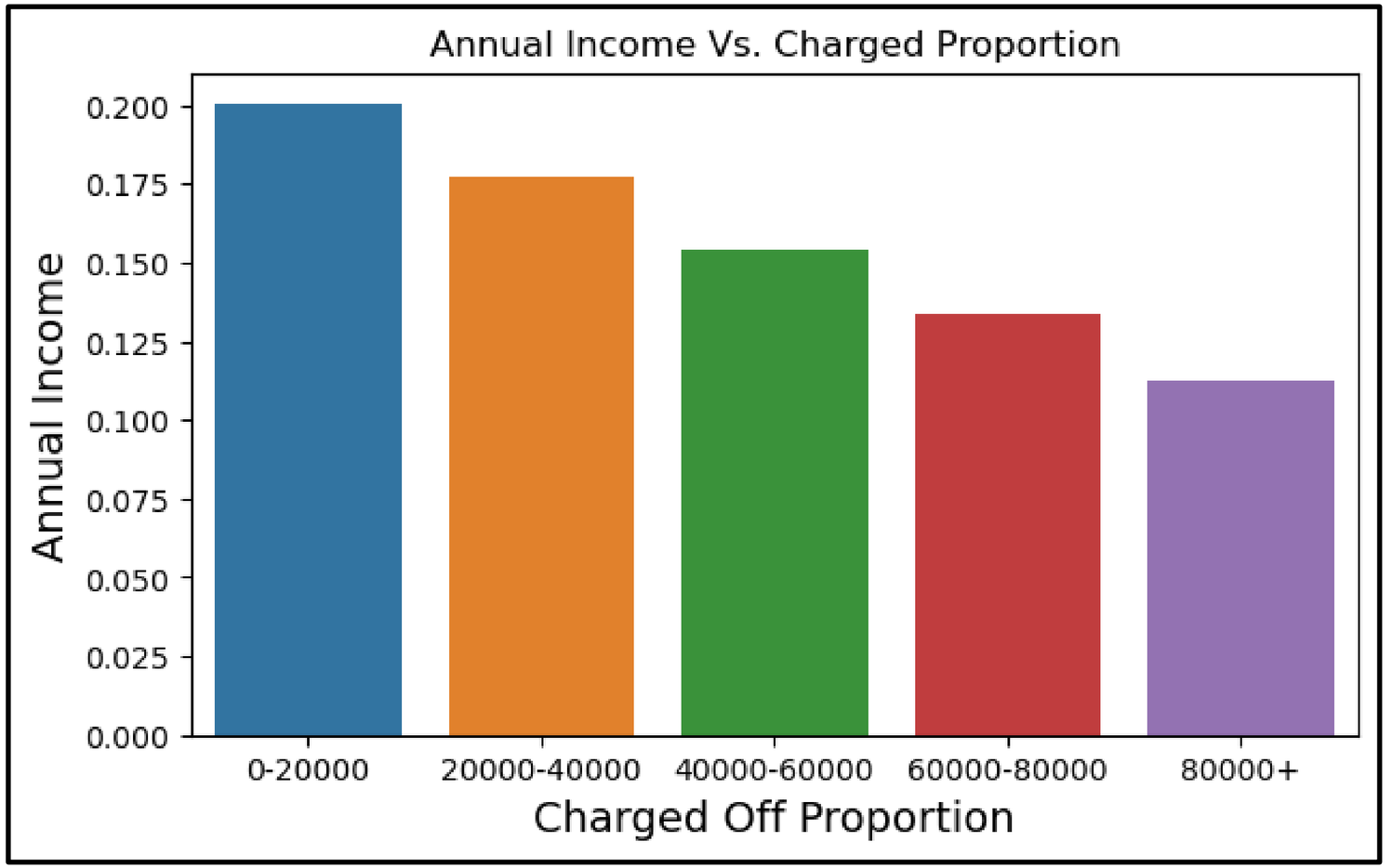
BIVARIATE ANALYSIS : CORRELATION MATRIX

- Obsevation/s and insight/s
- Positive correlation is observed between principal received to date an interest received to date, loan amount, funded amount by investros and total payment. This is expected
- There is negetive correlation between principal received to date with late fees received to date & recoveries
- This means when late fees and recoveries decreases principal received will increase
- When loan amount is high toal interest received is also high. High ticket loans are briningin in more interest to the business.
- Interestingly the number of accounts/credit lines by borrower is negetively correlated to late fees received to date and interest rate.
- This means borrowers with higher number of multiple accounts are able to manage their credit better as compared to borrowers with less number of accounts.
- When funded amount by investors goes up loan amount also goes up.
- Interest rate is lesser for employees having longer employement.
- Total late fees received to date decreases as employement length increases
- Annual income is negetively correlated to debt to income ratio.



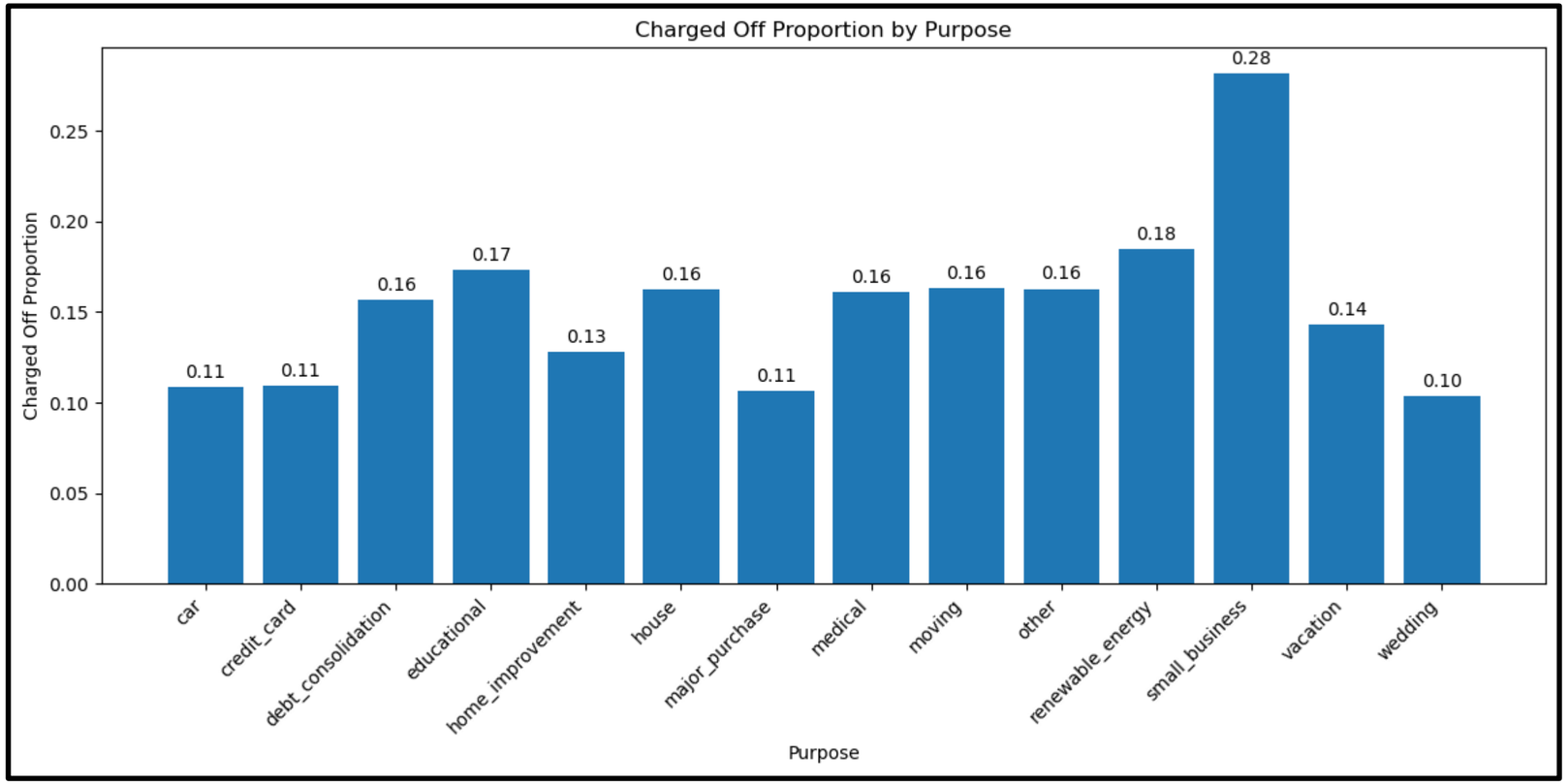
BIVARIATE ANALYSIS : CHARGED OFF LOAN CATEGORY

- Income range 0-20000 has the highest chances of being charged off.



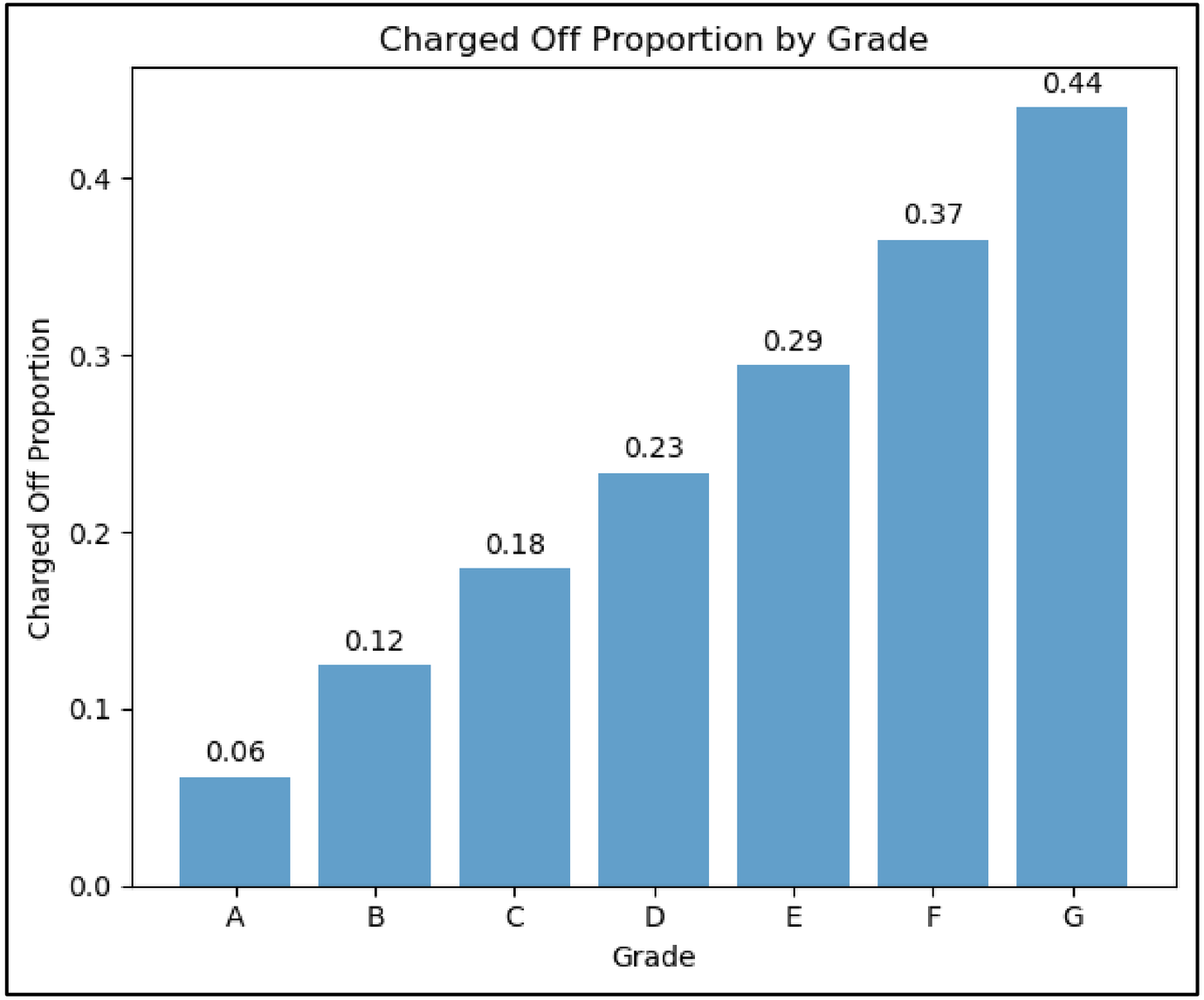
BIVARIATE ANALYSIS : CHARGED OFF LOAN vs. PURPOSE

- Top three purpose leading to charge off
  - Small business
  - Renewable energy
  - Educational



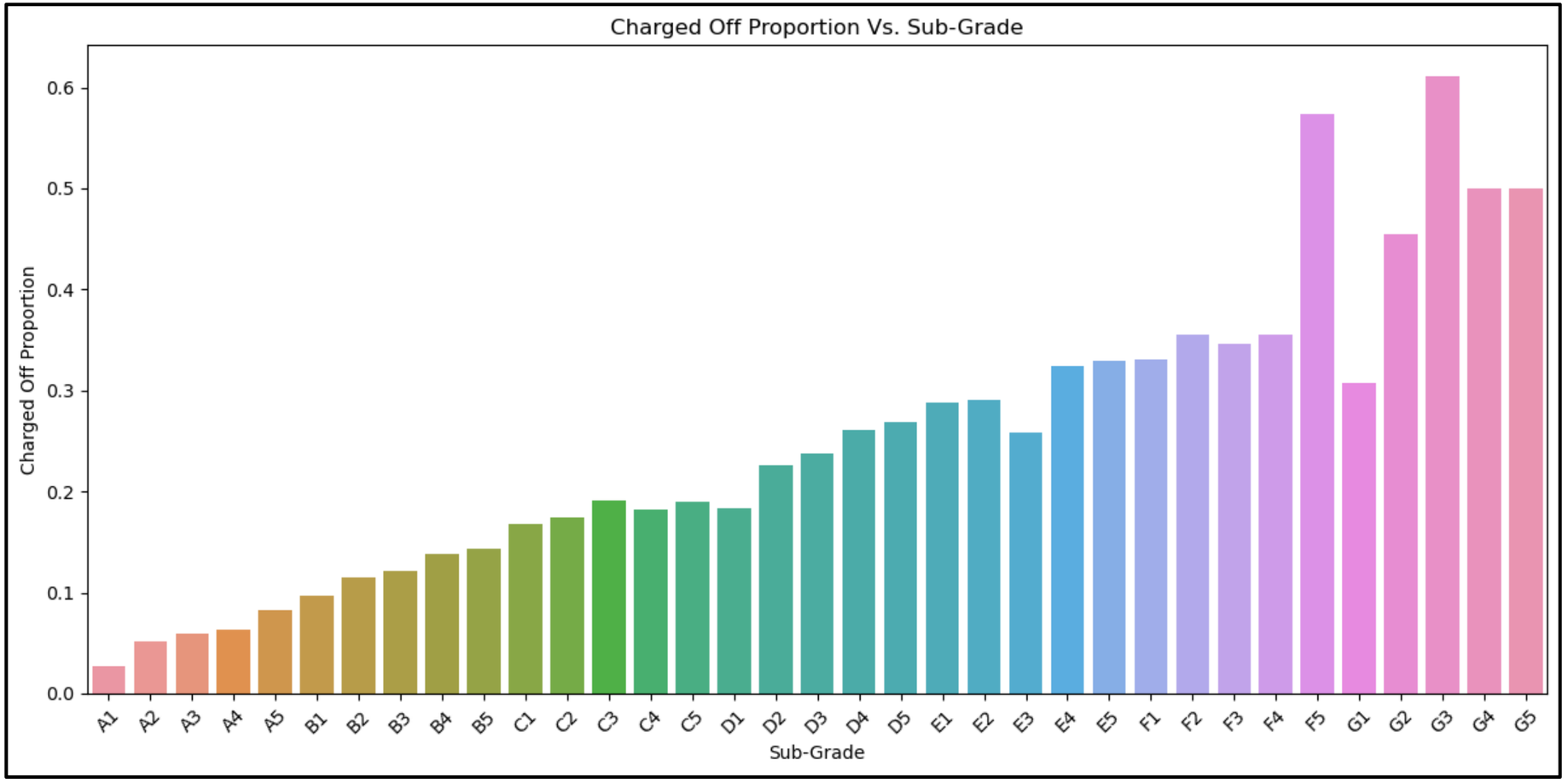
BIVARIATE ANALYSIS : CHARGED OFF LOAN VS. GRADE

- Grade A borrowers have the lowest likelihood of experiencing charge-offs.
- Conversely, Grade F and Grade G borrowers face a significantly higher risk of charge-offs.
- The probability of charge-offs appears to increase as you progress from Grade A to Grade G.



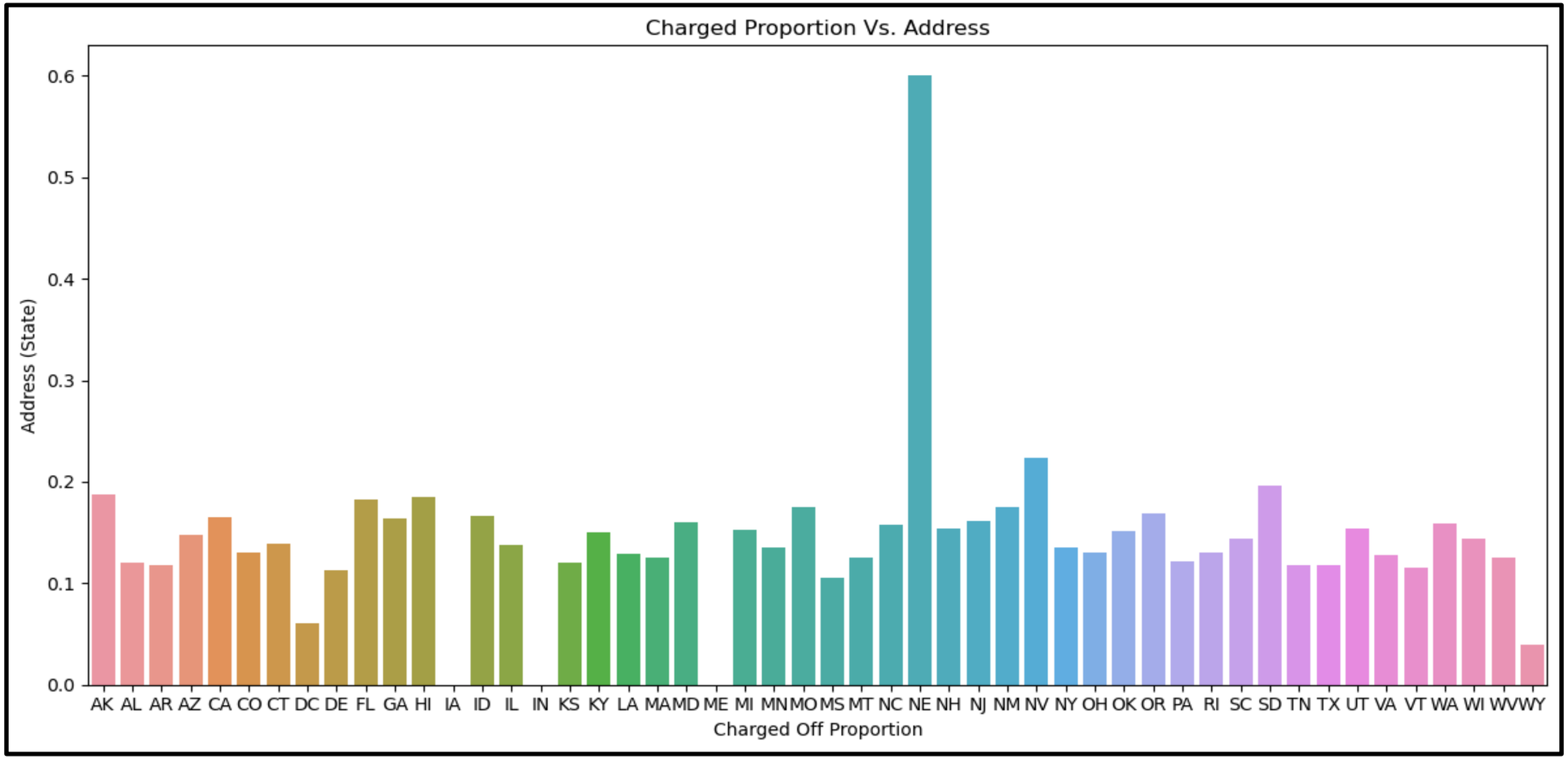
BIVARIATE ANALYSIS : CHARGED OFF LOAN VS. SUB-GRADE

- Sub-Grade A1 exhibits the lowest likelihood of experiencing charge-offs.
- Conversely, Sub-Grades F5 and G3 are associated with a significantly higher risk of charge-offs.



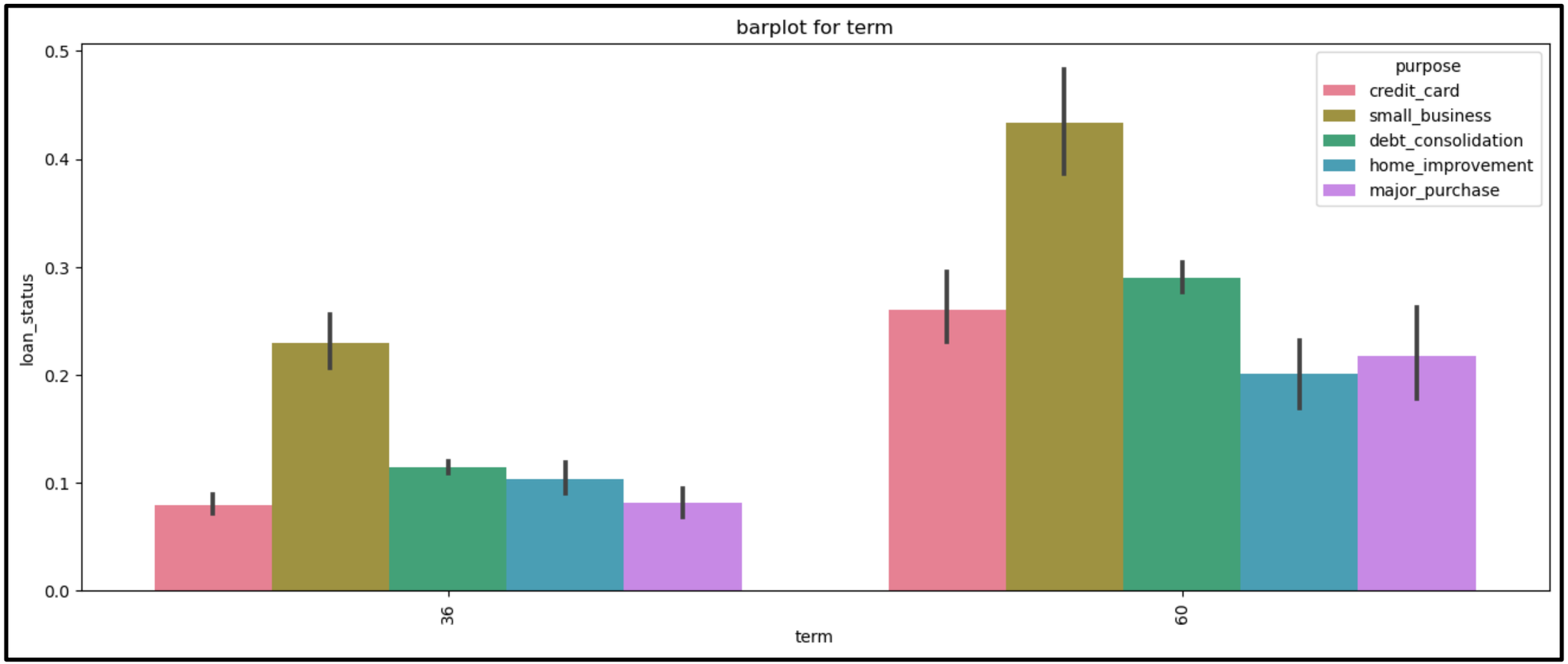
BIVARIATE ANALYSIS : CHARGED OFF LOAN VS. SUB-GRADE

- The state of NE exhibits a very high likelihood of charge-offs, but the number of loan applications from this state is insufficient to draw meaningful conclusions.
- On the other hand, states like NV, CA, and FL show a substantial number of charge-offs across a significant volume of loan applications, indicating a noteworthy trend of charge-offs in these regions.



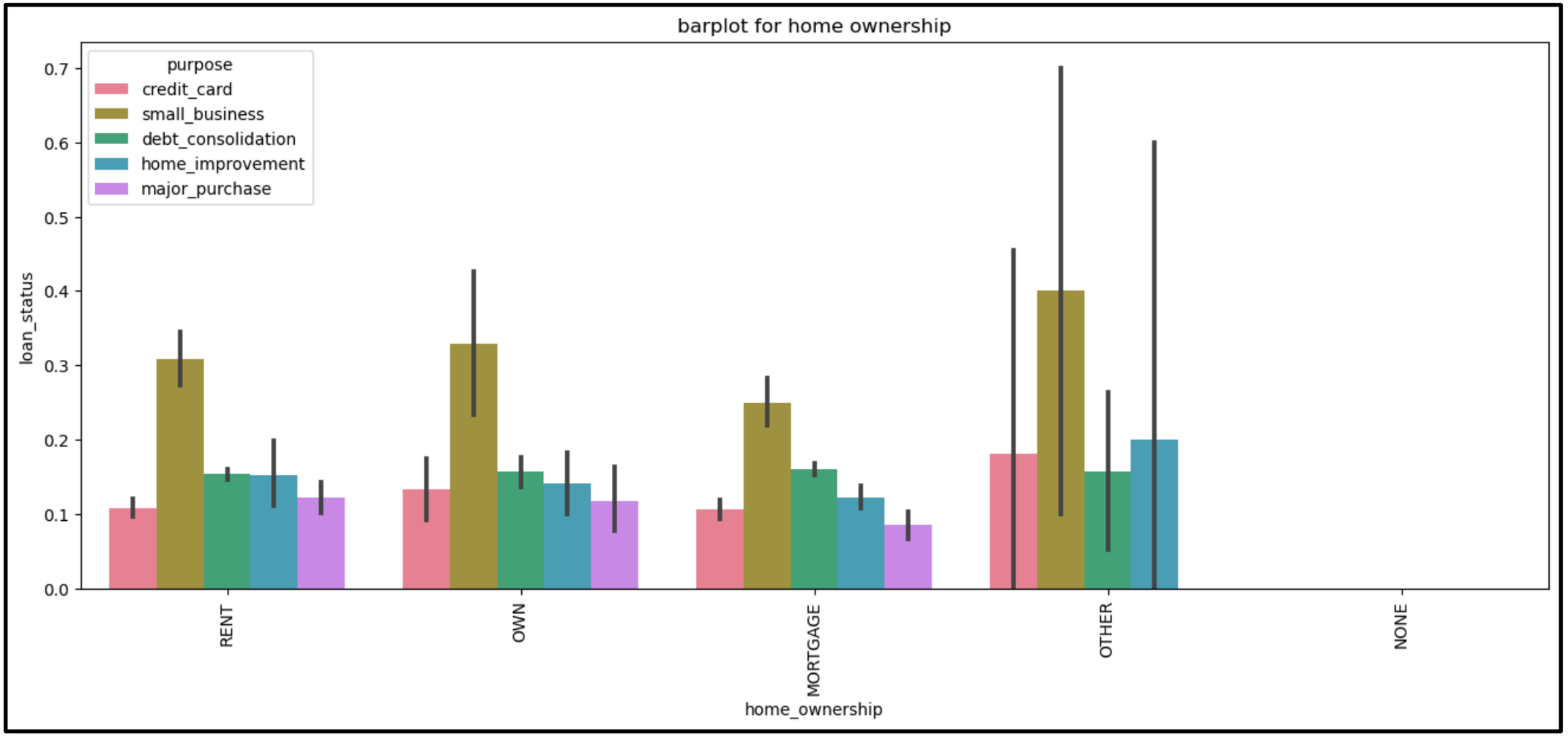
DATA ANALYSIS : SEGMENTED ANALYSIS

SEGMENTED ANALYSIS : LOAN STATUS, TERM, PURPOSE



- Small business has the most loans defualts, followed by debt_consolidation for both terms.

SEGMENTED ANALYSIS : LOAN STATUS, HOME OWNERSHIP, PURPOSE



- Small business has the most loans defaults across home ownership segment.

RECOMMENDATIONS

RECOMMENDATIONS

- Based on the observations you've made from your data analysis, here are some recommendations and insights:

  - Small Business Loans: Since most defaulters are observed to be taking loans for small business purposes, it may be advisable for lenders to implement stricter screening criteria for applicants seeking loans for small businesses. This could include a more thorough assessment of the business plan and financial stability of the business.
  - Higher Annual Income: The data shows that defaulters tend to have lower annual incomes. Lenders could consider setting minimum income thresholds for loan applicants to reduce the risk of default. Additionally, offering financial literacy resources to help borrowers manage their finances could be beneficial.
  - Grade G and F Loans: Lenders should be cautious when approving loans with Grade G and F, as these grades are associated with higher default rates. It may be wise to limit the issuance of such loans or charge higher interest rates to compensate for the increased risk.
  - Higher Interest Rates: Defaulters are more likely to have loans with interest rates greater than 16%. Lenders should be cautious about extending loans with very high-interest rates, as this may make it difficult for borrowers to meet their repayment obligations. Proper assessment of the borrower's ability to repay at such rates is crucial.
  - Economic Crisis Impact: The observation that most defaulters are from the year 2011, which was a recession year, highlights the impact of economic crises on loan defaults. Lenders should consider macroeconomic factors and market conditions when assessing loan applications during economic downturns.
  - Loan Term: Defaulters are more common among loans with a 36-month term. Lenders might consider offering shorter-term loans or closely monitoring loans with longer terms to mitigate the risk of default.
  - Experience Level: Borrowers with no years of experience are more likely to default. Lenders should assess the stability and income potential of borrowers with limited work experience more rigorously.
  - Bankruptcy Records: Borrowers with public bankruptcy records are at a higher risk of default. Lenders should exercise caution when dealing with applicants who have a history of bankruptcy and may implement stricter eligibility criteria or charge higher interest rates to mitigate the risk.

SUMMARY OF RECOMMENDATIONS

In summary, the data analysis suggests that lenders should focus on improving their risk assessment and monitoring processes, especially for loan applicants seeking funds for small businesses, those with lower income, poor credit grades, high-interest rates, limited experience, and a history of bankruptcy. Additionally, being aware of economic conditions and their potential impact on default rates is essential for prudent lending practices

THANK YOU

Have any question?