Lending Club Case Study

Group Members:
Mayank Bhardwaj
Ritu Sharma

Index

Objective

- Lending Club is America’s largest online credit marketplace, and the first marketplace bank connecting borrowers and investors.

- When the company receives the loan application, it has to decide for loan approval on the basis of loan applicant

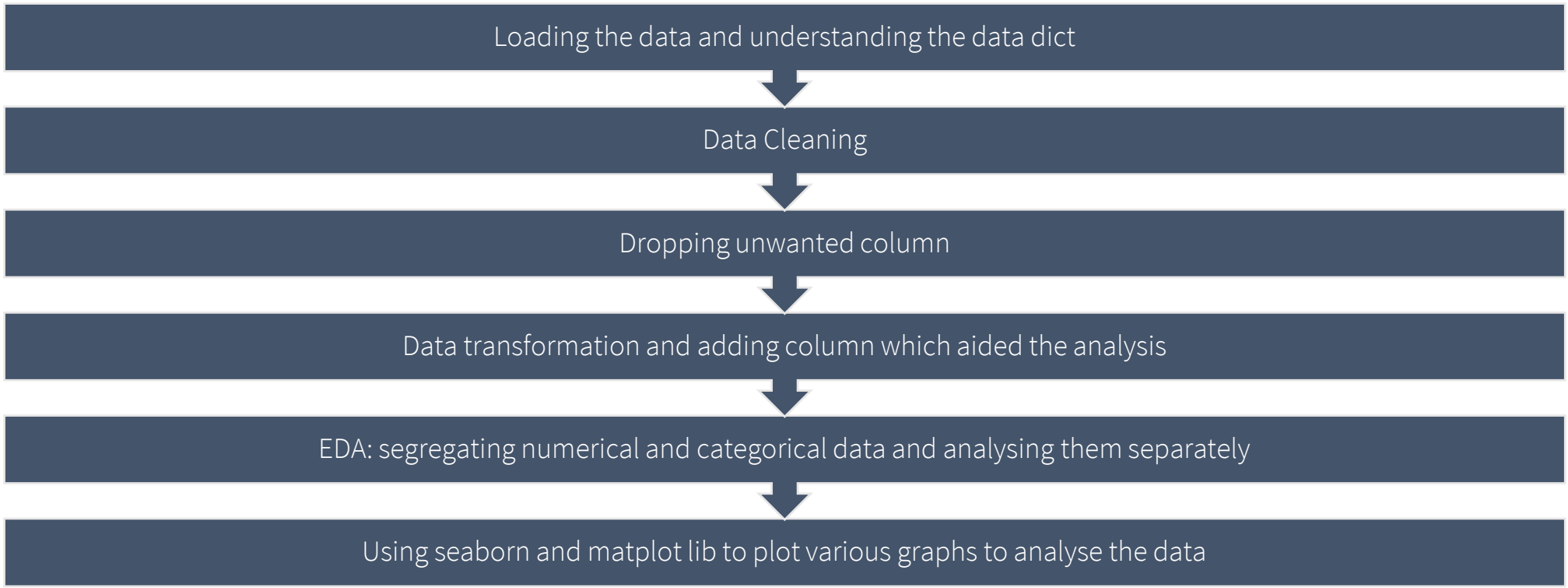
- The objective is to use EDA to understand how consumer attributes and loan attributes influence the tendency to default

Data Understanding

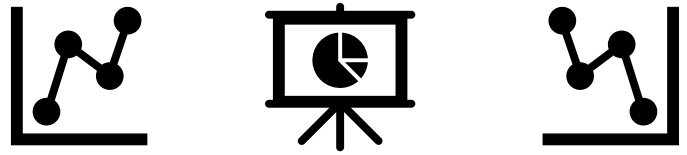
Excluded Columns:

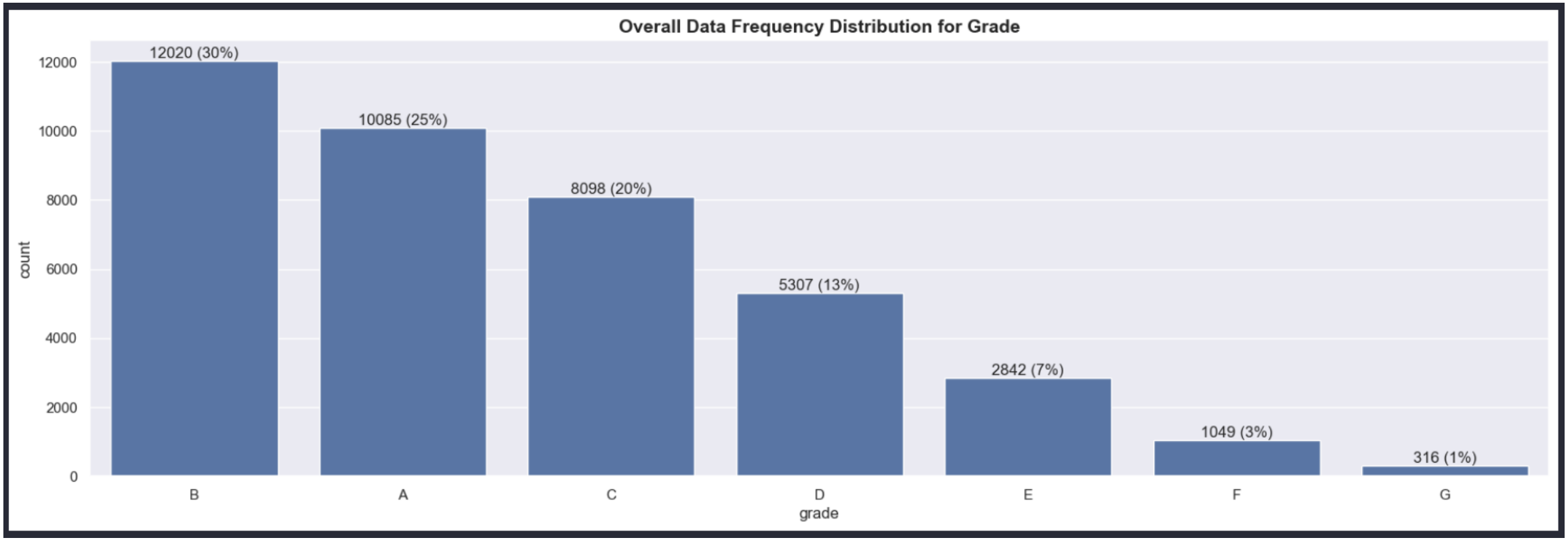
- In our analysis, we will not consider certain types of columns. It's important to note that this is a general categorization of the columns we will exclude from our approach, and it does not represent an exhaustive list.
- **Customer Behavior Columns**-Columns that describe customer behavior will not be factored into our analysis. The current analysis focuses on the loan application stage, while customer behavior variables pertain to post-approval actions. Consequently, these attributes will not influence the loan approval/rejection process.
- **Granular Data** -Columns providing highly detailed information that may not be necessary for our analysis will be omitted. For example, while the "grade" column may have relevance in creating business outcomes and visualizations, the "sub grade" column is excessively granular and will not be utilized in our analysis.
- **54** columns contain **NA** values only, and these columns will be removed namely **acc_open_past_24mths, all_util, annual_inc_joint, avg_cur_bal, bc_open_to_buy, bc_util, dti_joint, il_util, inq_fi, inq_last_12m, max_bal_bc, mo_sin_old_il_acct, mo_sin_old_rev_tl_op, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl, mort_acc, mths_since_last_major_derog, mths_since_rcnt_il, mths_since_recent_bc, mths_since_recent_bc_dlq, mths_since_recent_inq, mths_since_recent_revol_delinq, num_accts_ever_120_pd, num_actv_bc_tl, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl, num_rev_accts, num_rev_tl_bal_gt_0, num_sats, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, open_acc_6m, open_il_12m, open_il_24m, open_il_6m, open_rv_12m, open_rv_24m, pct_tl_nvr_dlq, percent_bc_gt_75, tot_coll_amt, tot_cur_bal, tot_hi_cred_lim, total_bal_ex_mort, total_bal_il, total_bc_limit, total_cu_tl, total_il_high_credit_limit, total_rev_hi_lim, verification_status_joint**
- Certain columns contain only **0** values, and these columns will also be dropped.
- **9** Columns with **single value** that do not contribute to the analysis will be removed.
- Columns with values that are single value but have other values as NA will be treated as constant and dropped.
- Columns with more than **65%** of data being empty (**mths_since_last_delinq, mths_since_last_record**) will be dropped.
- Columns (**id, member_id**) will be dropped as they are index variables with unique values and do not contribute to the analysis.
- Columns (**emp_title, desc, title**) will be dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- The redundant column (**url**) will be dropped. Further analysis reveals that the URL is a static path with the loan ID appended as a query, making it redundant compared to the (**id**) column.
- These columns capture customer behavior recorded after loan approval and are not available at the time of loan approval. Thus, these variables will not be included in the analysis.
- Columns to be dropped: (**delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, out_prncp, out_prncp_inv, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type**)

Steps Taken to do analysis in python file:

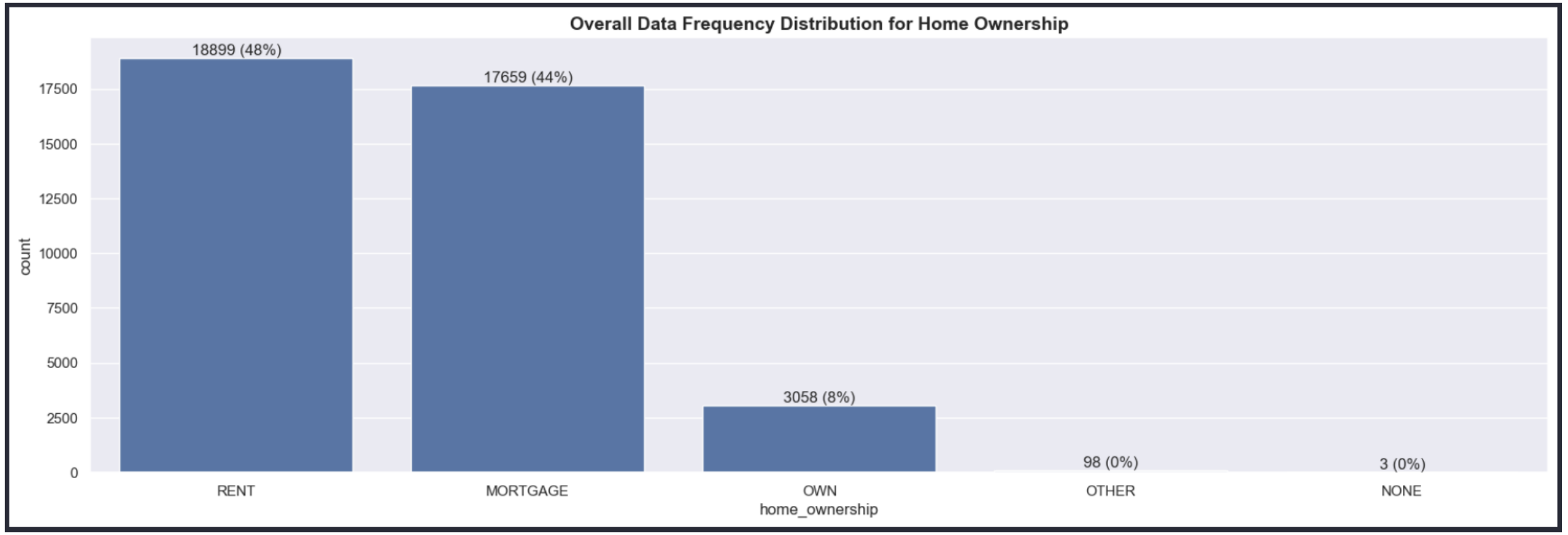


Overall Analysis: Frequency Distribution of Grade, Home Ownership and Purpose

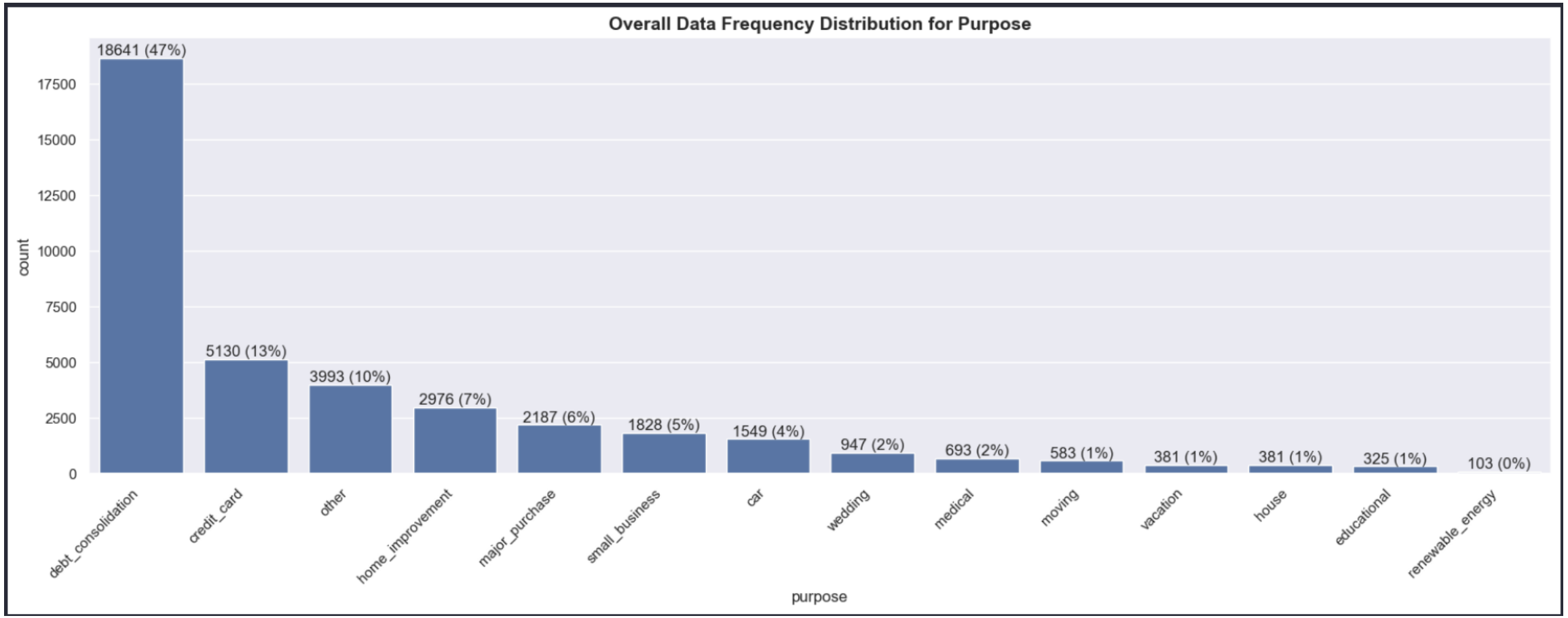




Overall Analysis based on grade show grade B customers are took highest Loan with a total count of 12020 which 30 % of overall customer count

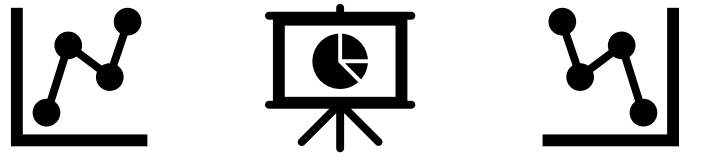


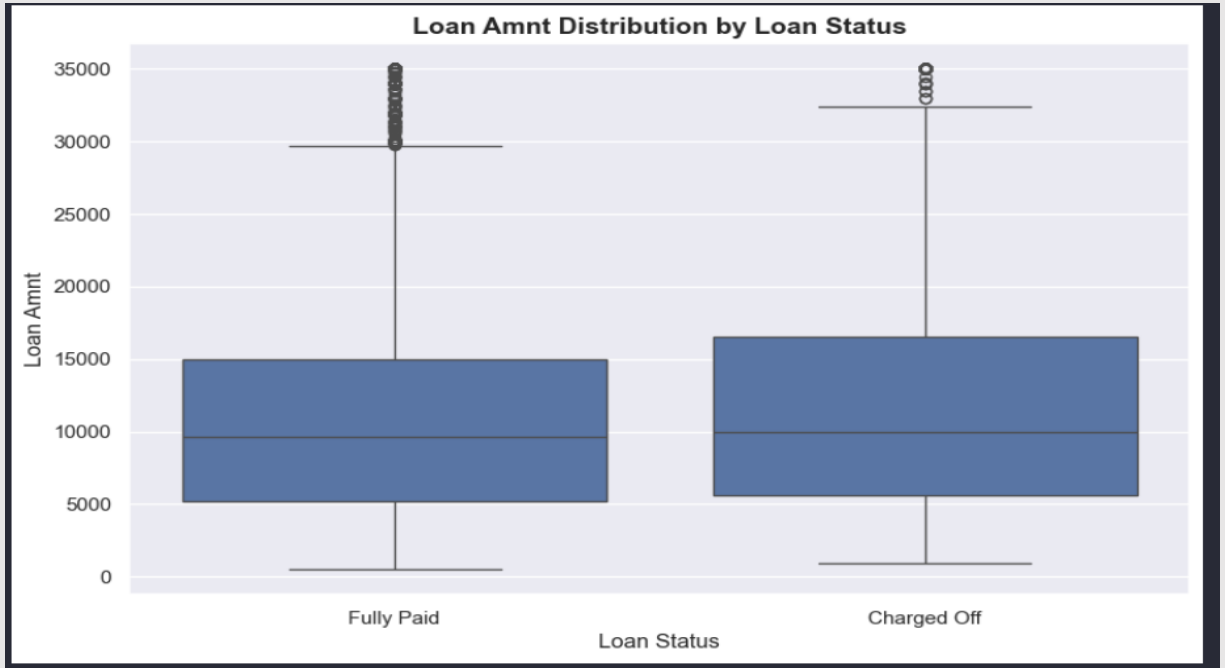
Overall Analysis based on Home Ownership show count of Rent and Mortgage are highest among all other ownership with a total of 18899 (48%) and 17659 (44%) respectively.

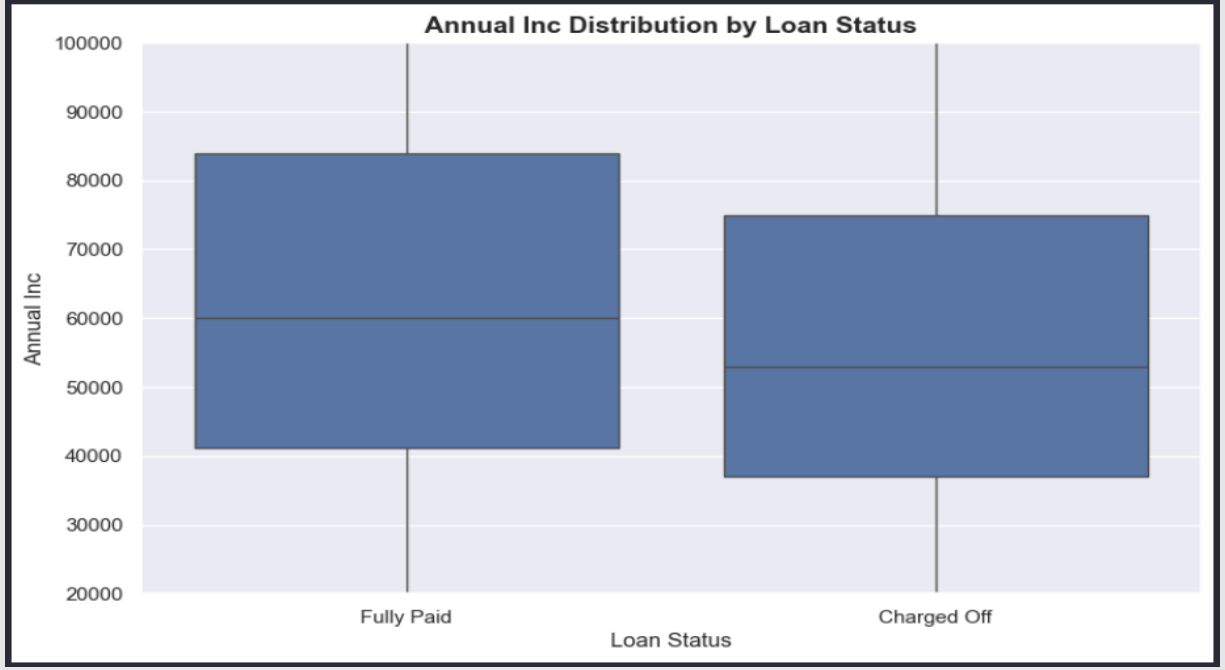


Overall Analysis based on Purpose show Debt Consolidation purpose grabs the highest count with a count of 18641 and with 47%

Bivariate Analysis (between Fully Paid and Charged Off Customers) of Loan Amount, Interest Rate, Instalment with respect to Loan status

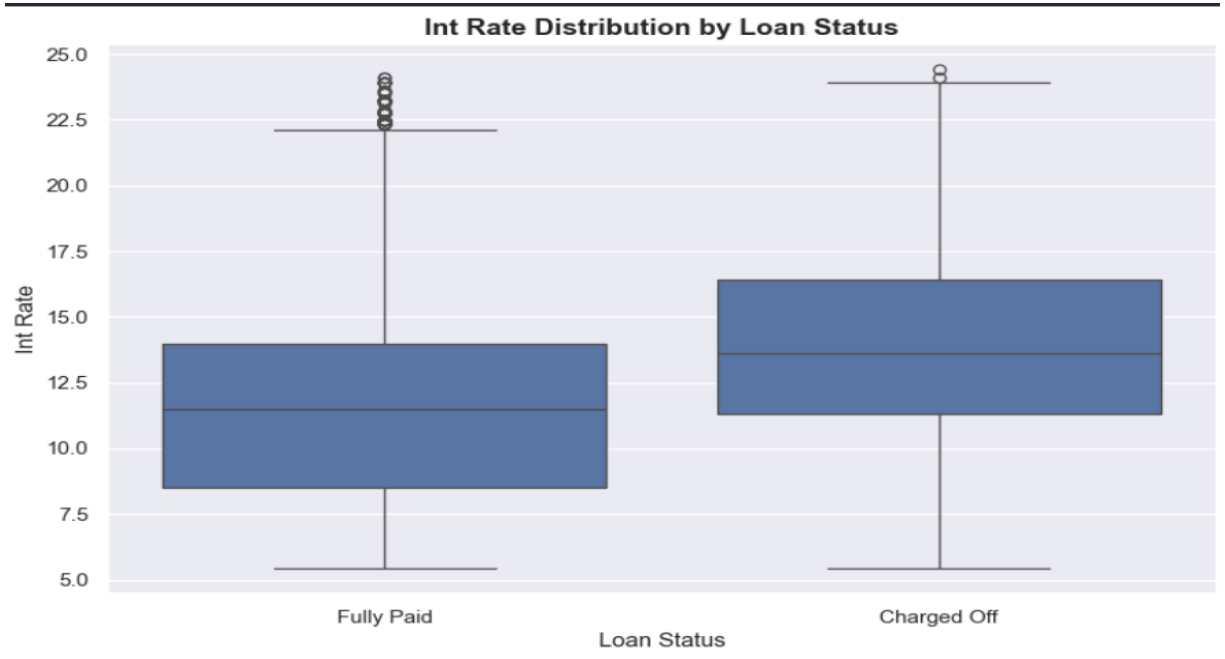


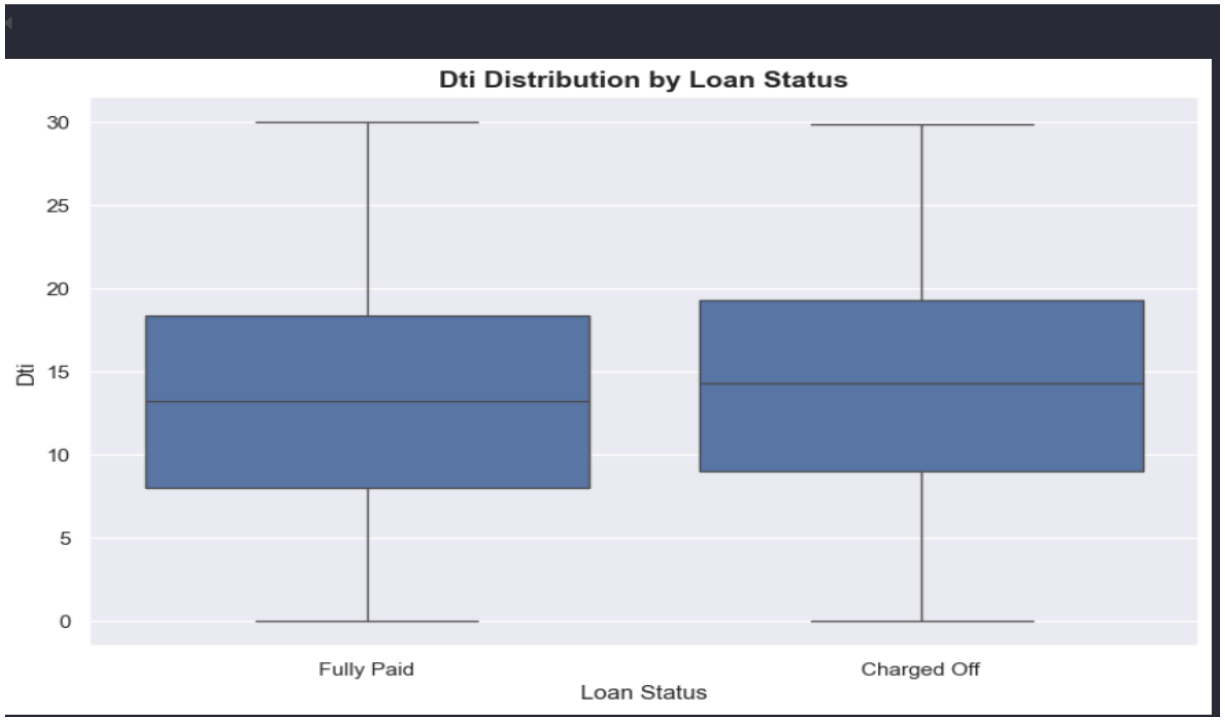




- There is a tendency to default when loan amount is higher as it is evident from the box plot.

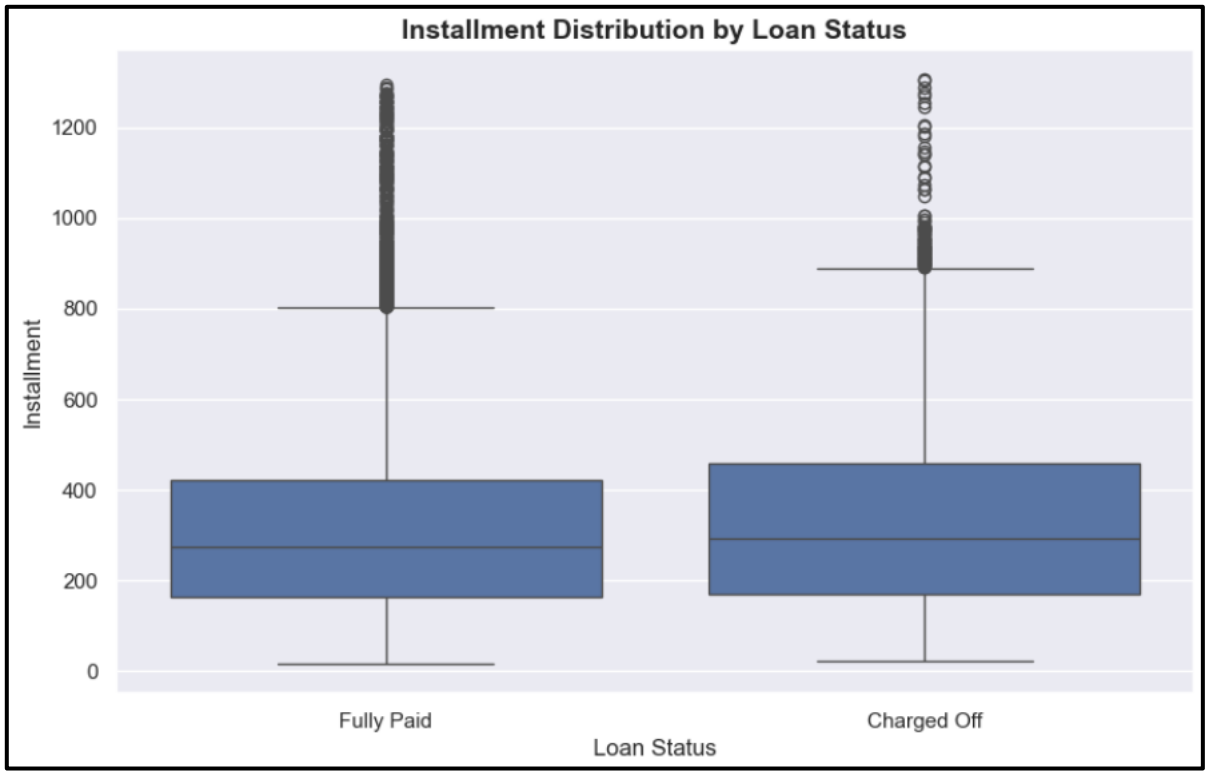
- Lower income levels have higher tendency to default.



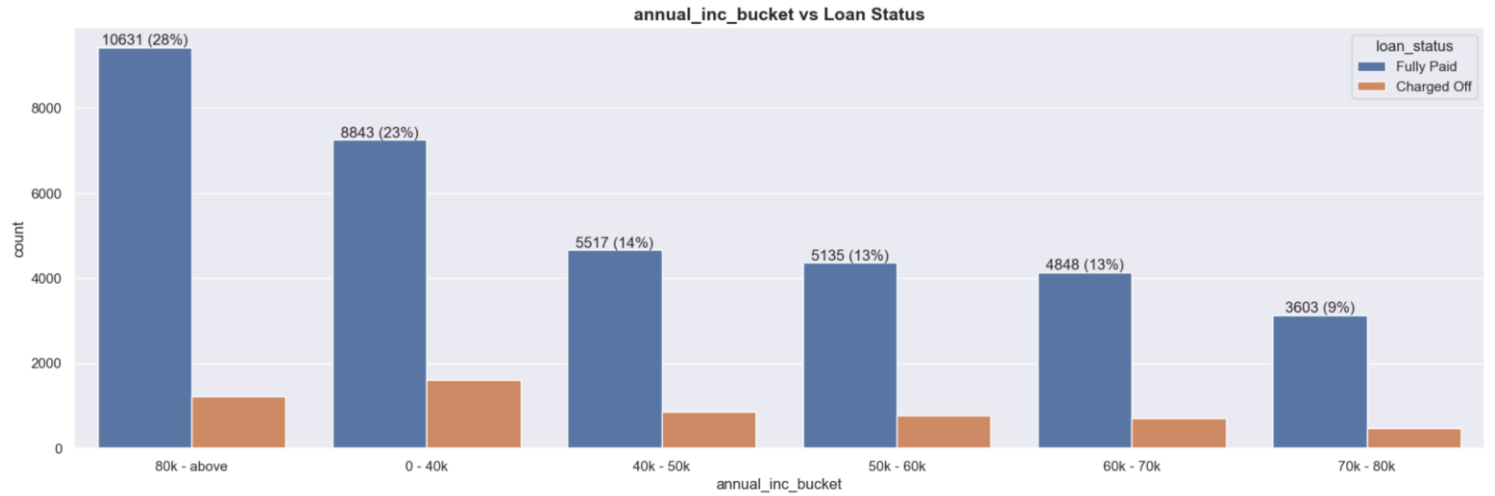


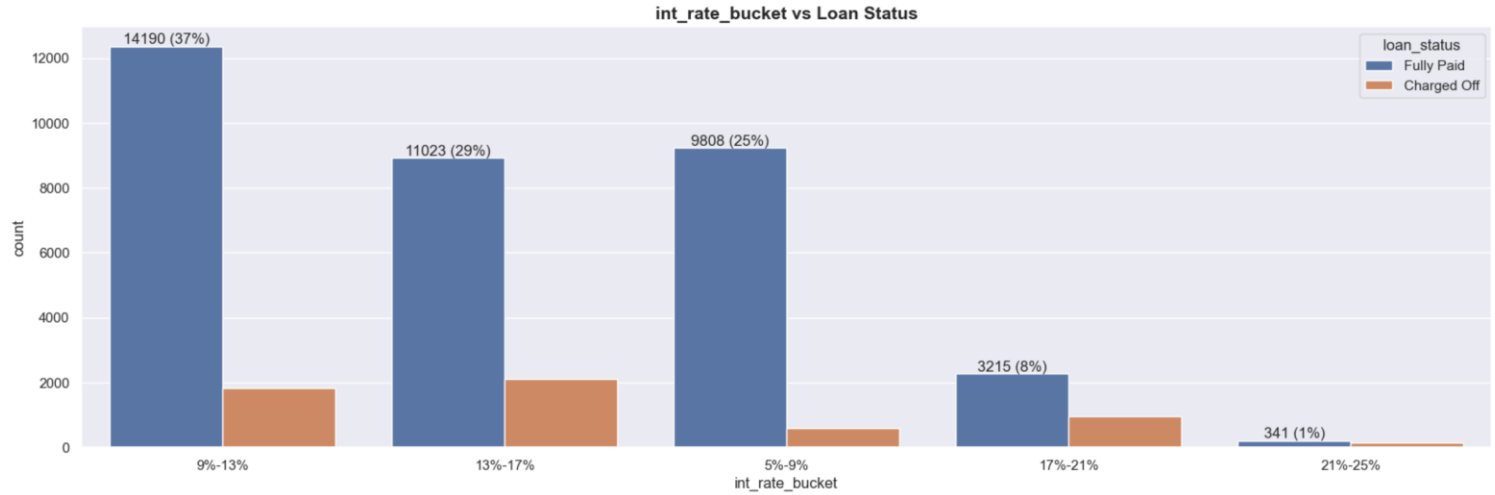
- There is a higher tendency to default when Interest rates are higher as box plot suggest Lower income levels have higher tendency to default.

- Debt-to-Income Ratio does not provide any indication of default.



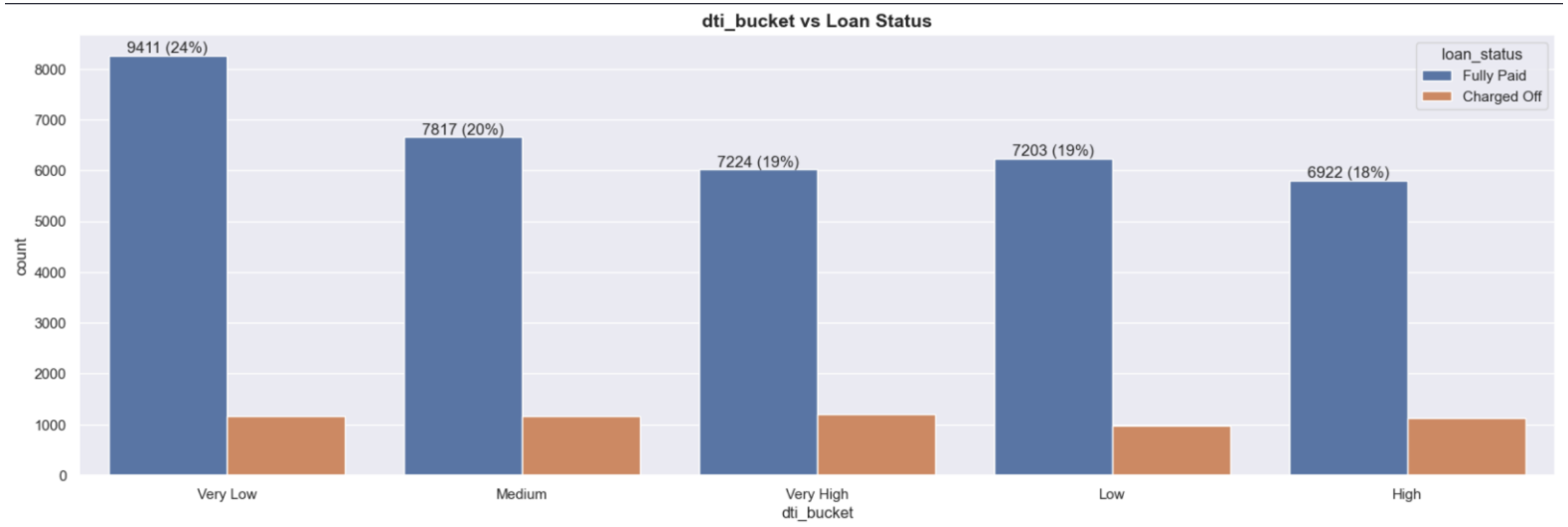
- Installment does not provide any indication of default.

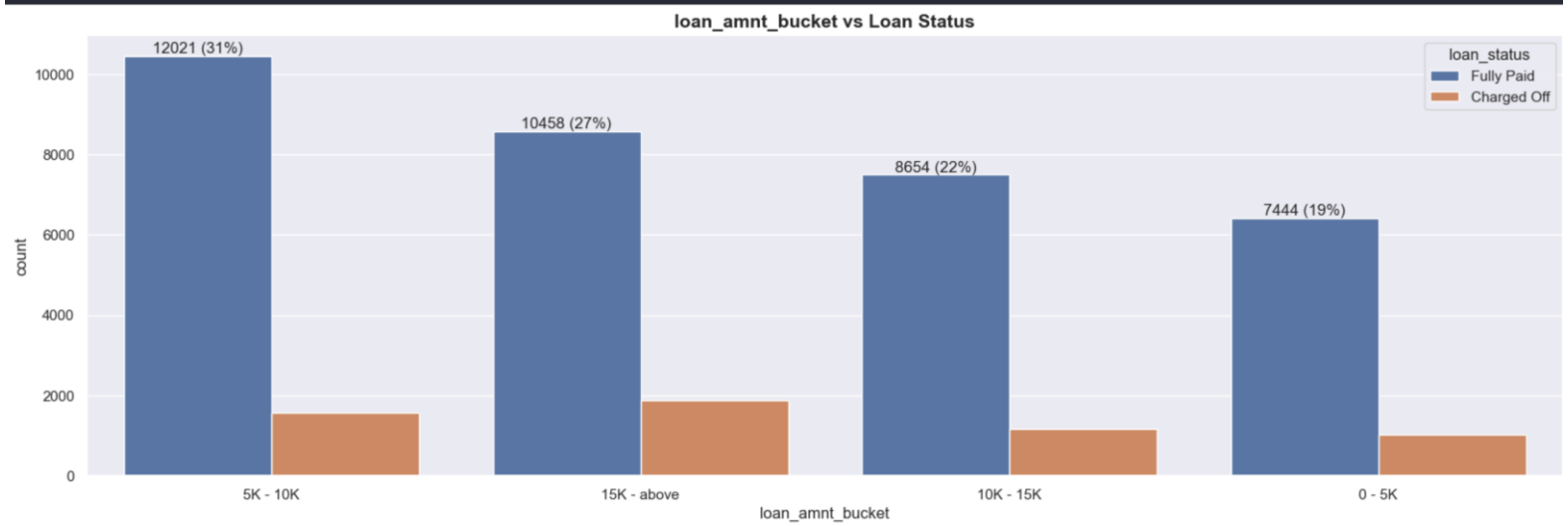




- A significant portion of loan applicants who defaulted received loans with interest rates falling within the range of 13% to 17%.

- A majority of the loan applicants who charged off reported an annual income of less than $40,000.

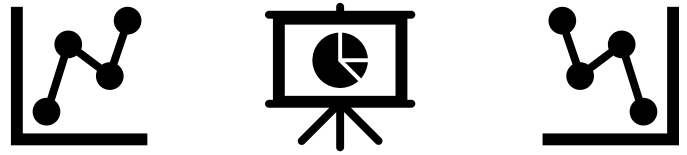


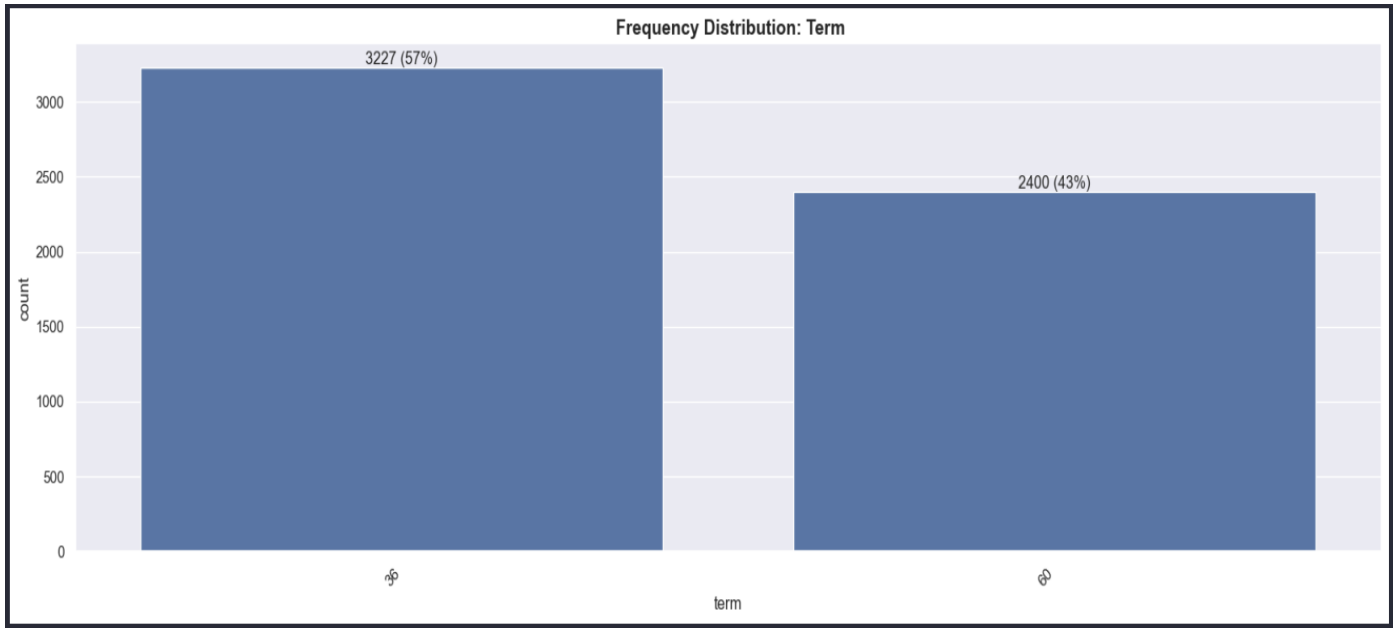


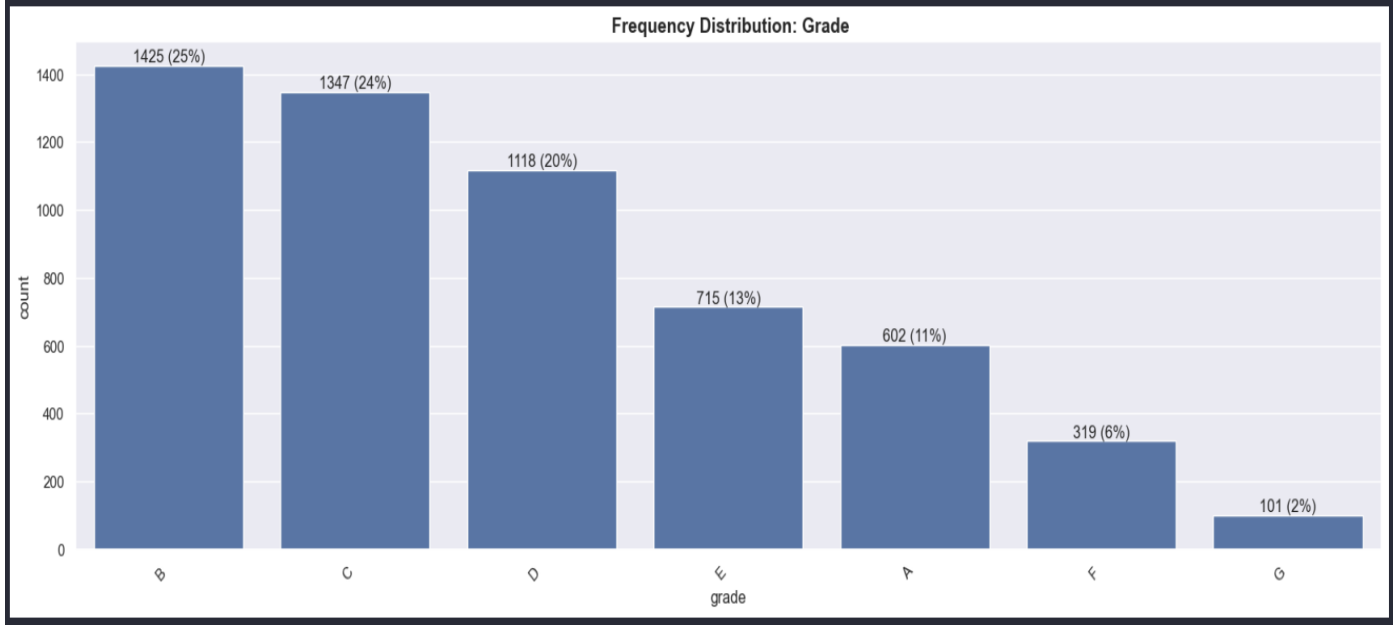
- A majority of the loan applicants who defaulted received loan amounts of $15,000 or higher.

- The majority of loan applicants who charged off had significantly high Debt-to-Income (DTI) ratios.

Univariate Analysis (Only for Charged Off Customers) of Term, Grade, Emp Length, Home Ownership and Verification Status

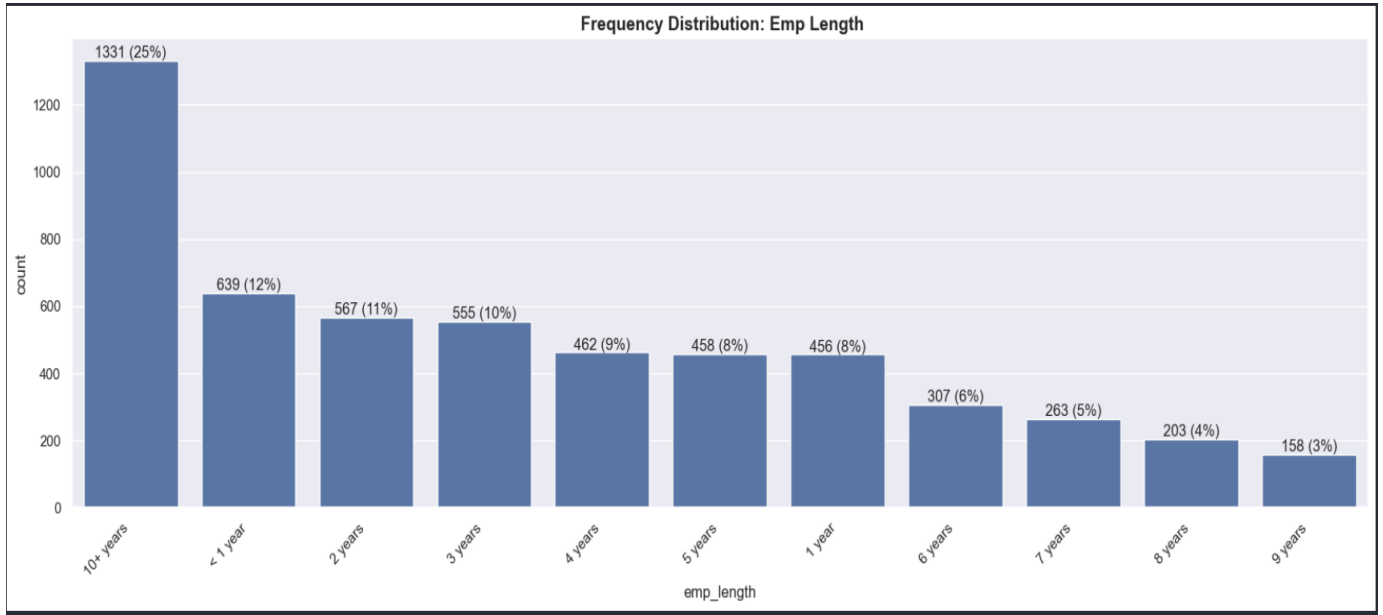


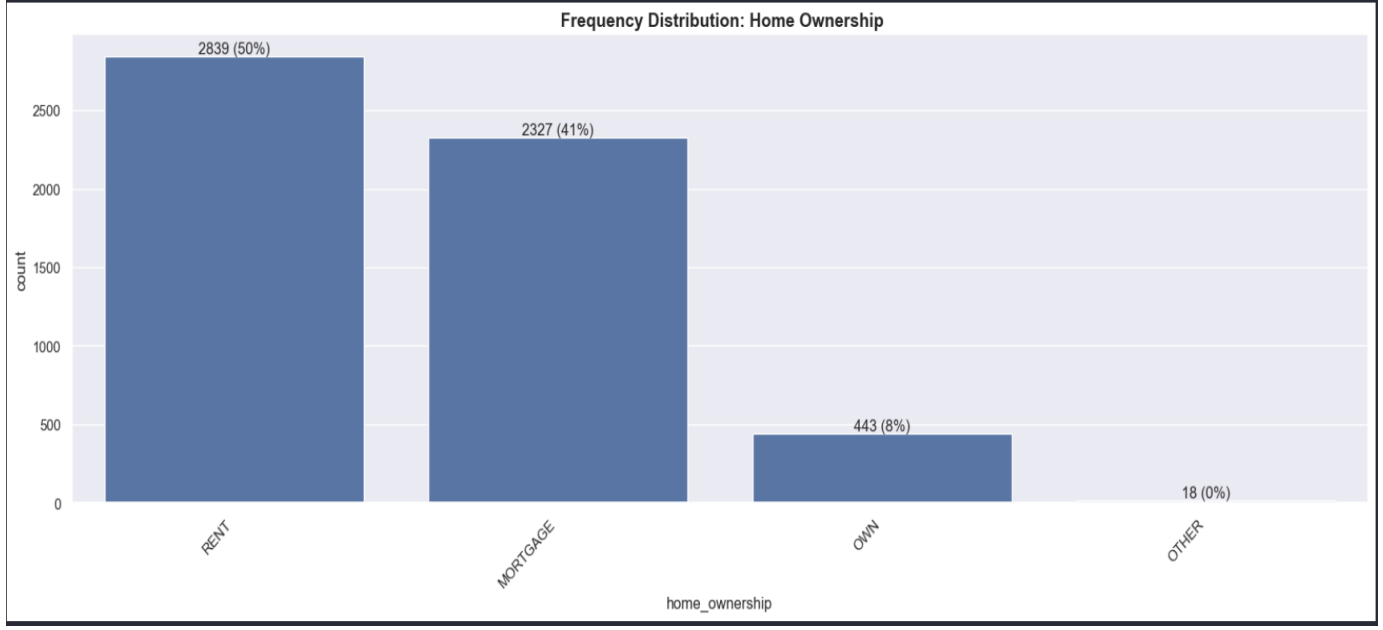




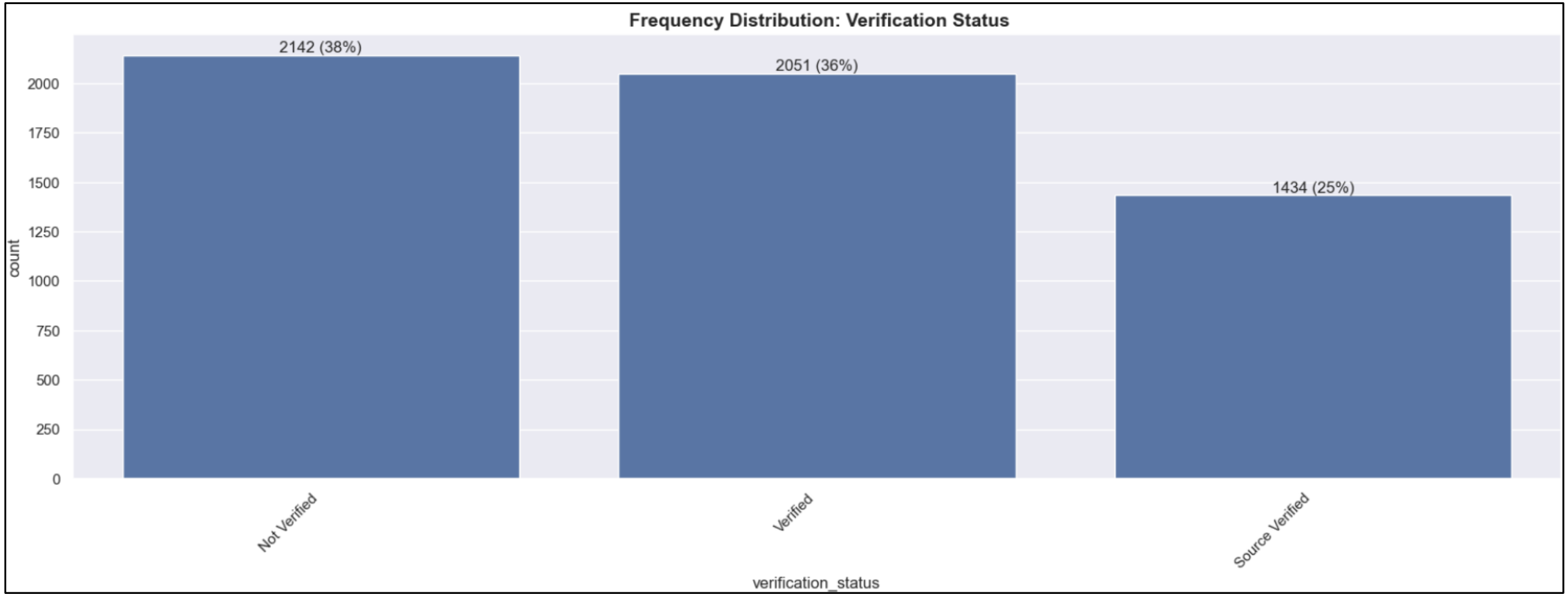
- There is a higher tendency to default when term of loan is for 36 months.

- There is a higher tendency to default when grade of customer is either B or C

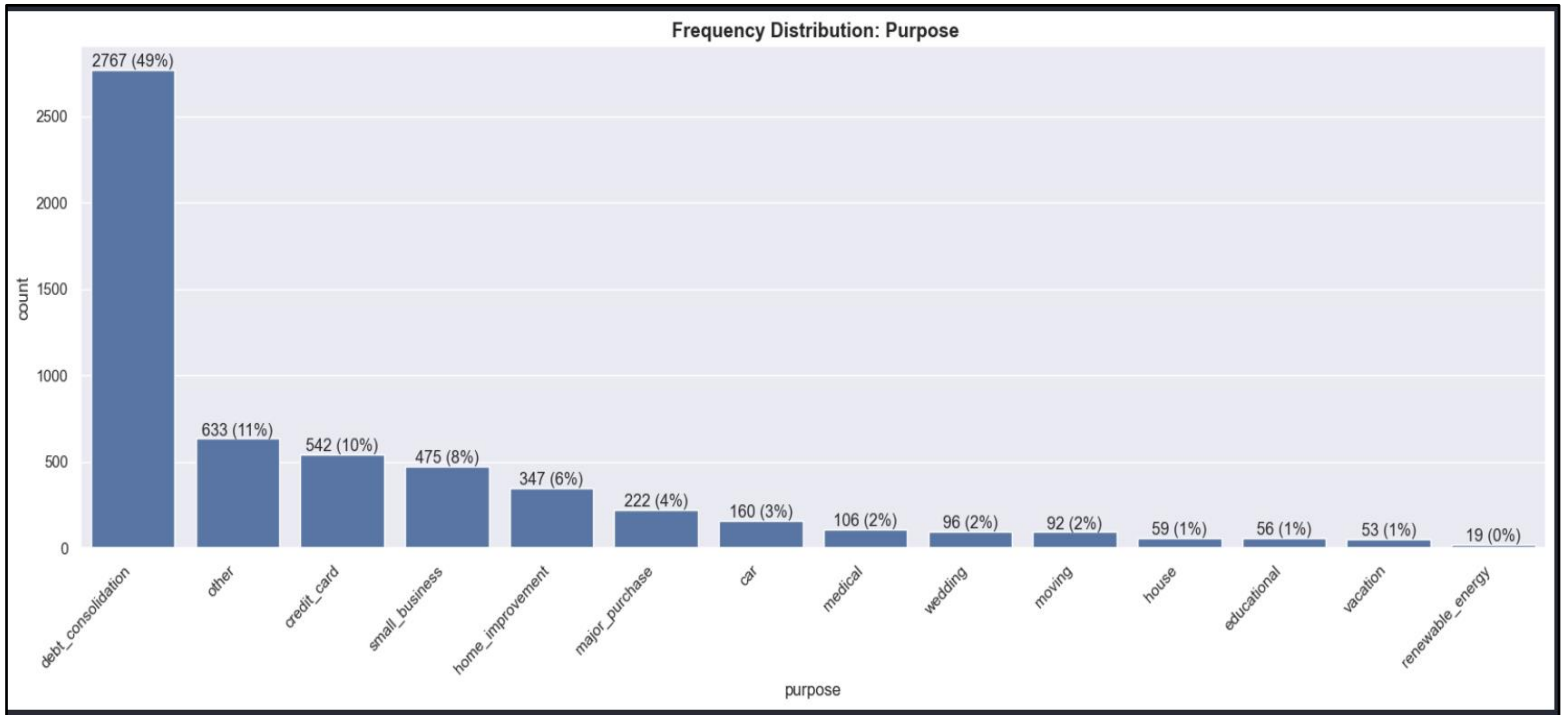


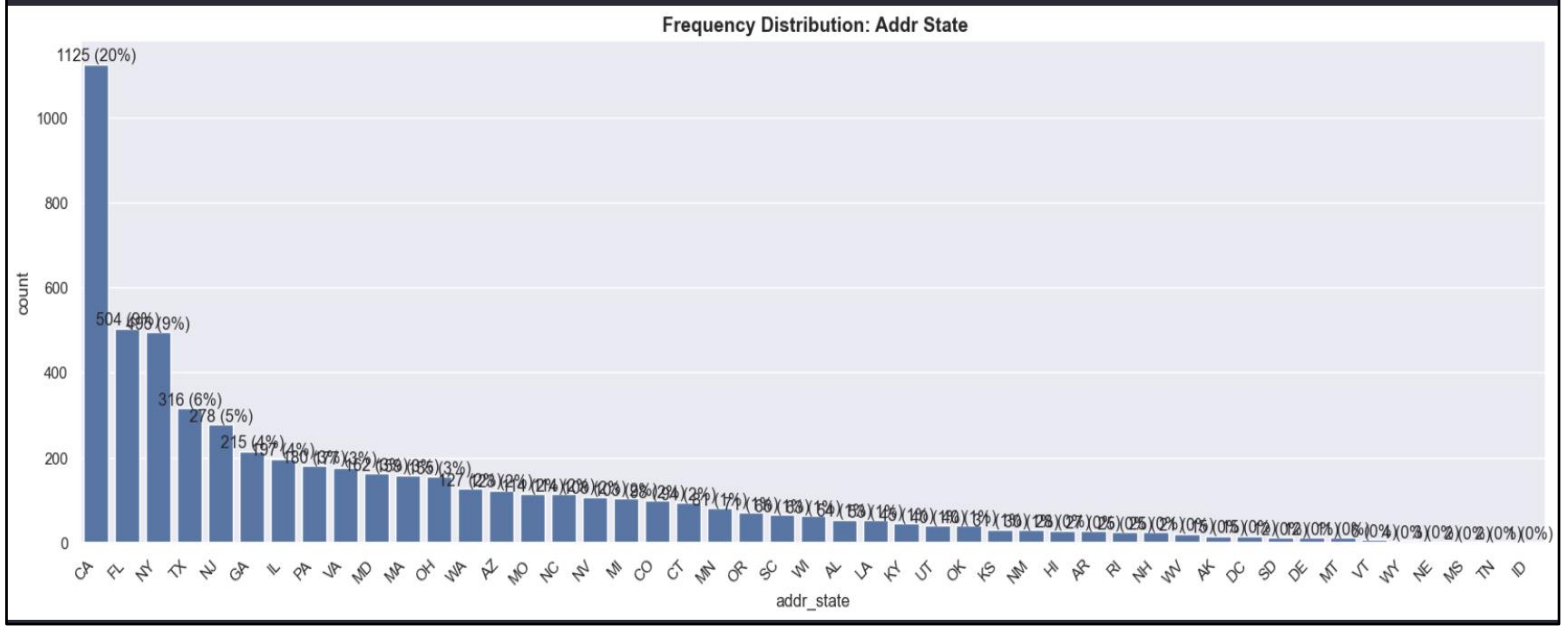


- There is a higher tendency to default when employment length is 10+ years.
- There is a higher tendency to default when home ownership is either on Rent or Mortgage



- There is a higher tendency to default when verification status is not ‘Source Verified’

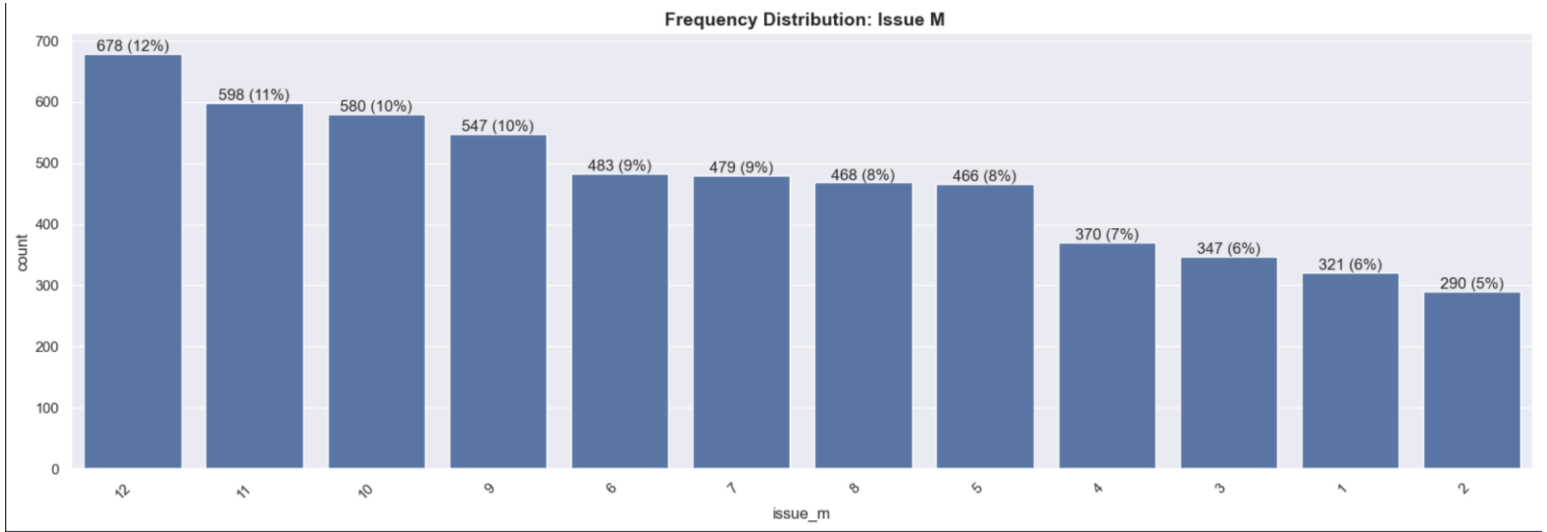


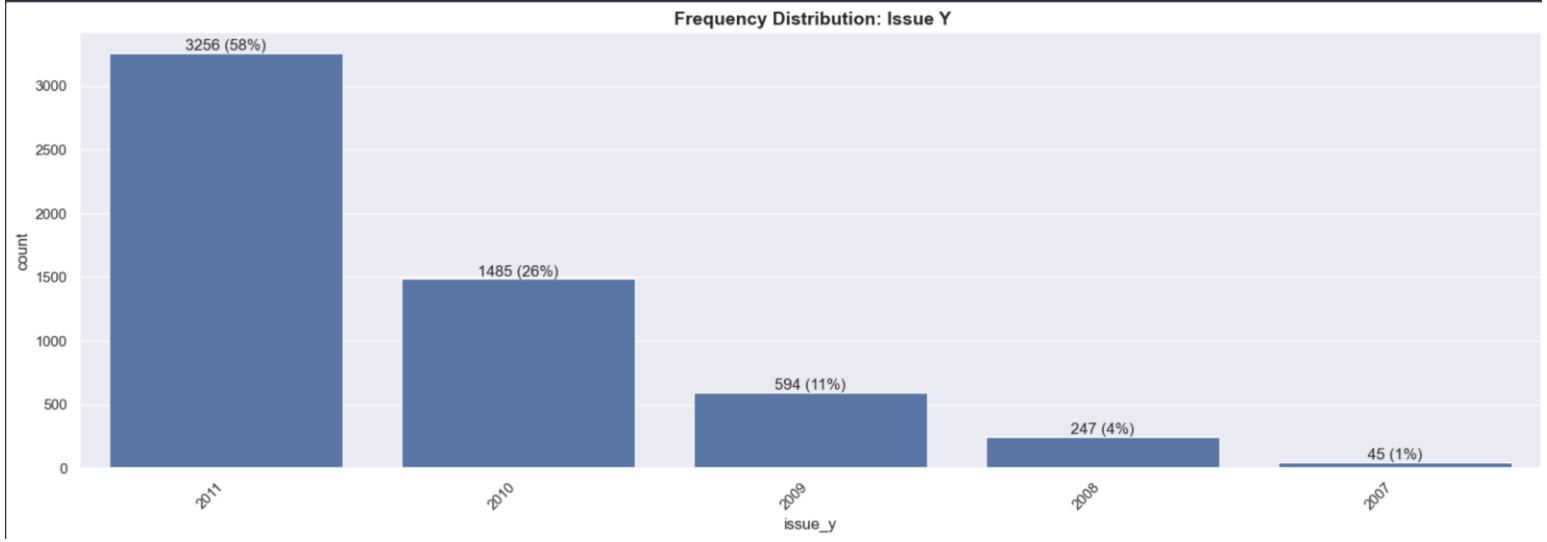


- There is a higher tendency to default when purpose is of loan is Debt Consolidation

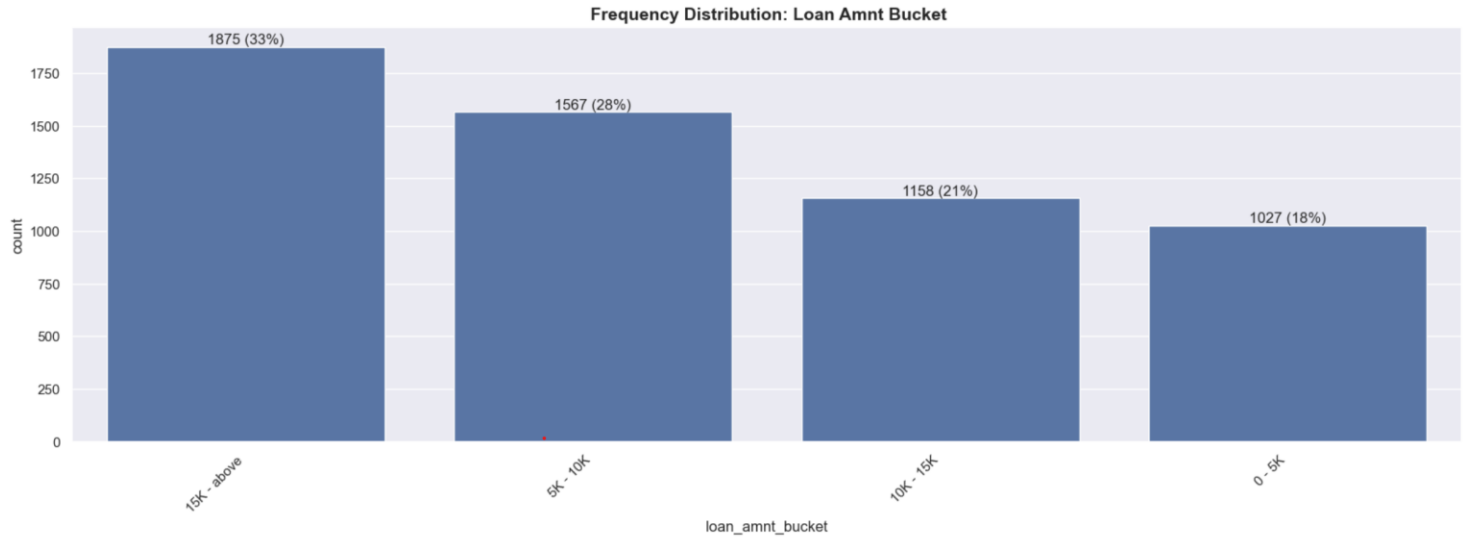
- There is a higher tendency to default when Customer is from CA, FL or NY

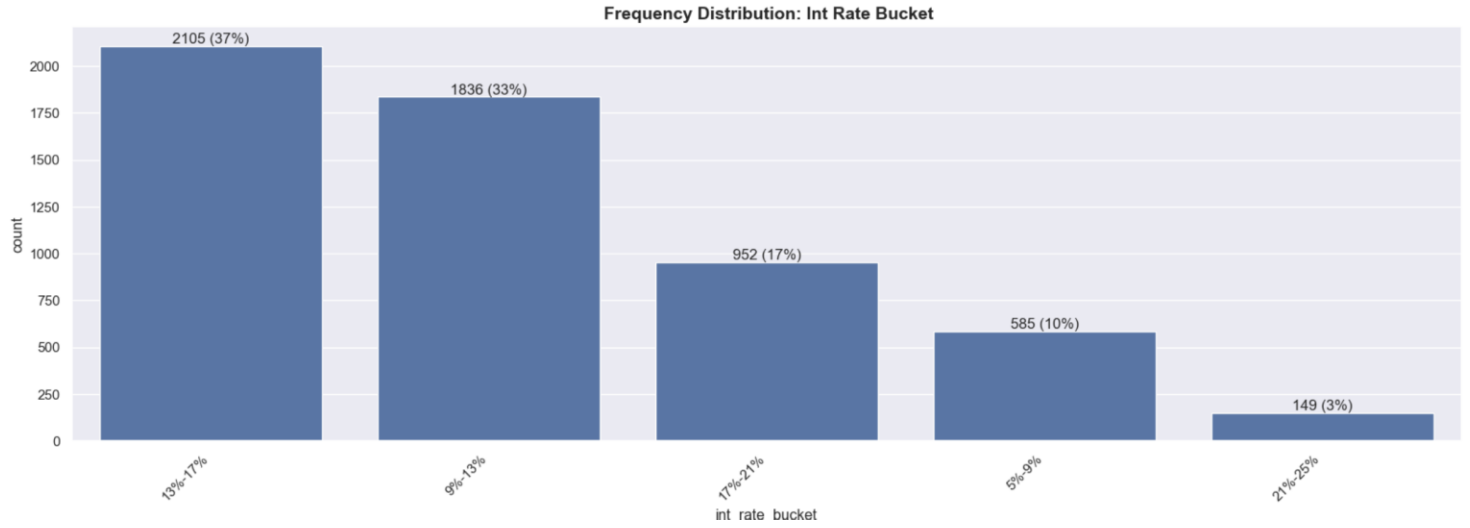
Note: CA: California , FL: Florid, NY: New York





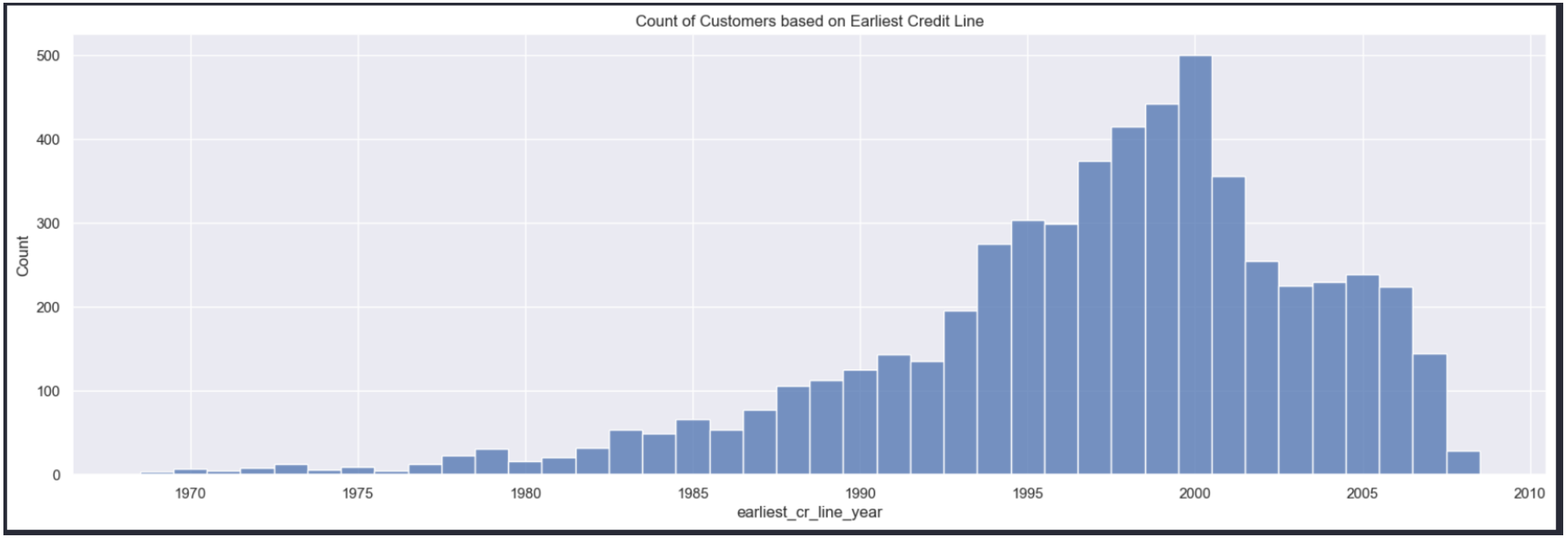
- Higher tendency to default was observed for year 2011 and for later months of the year





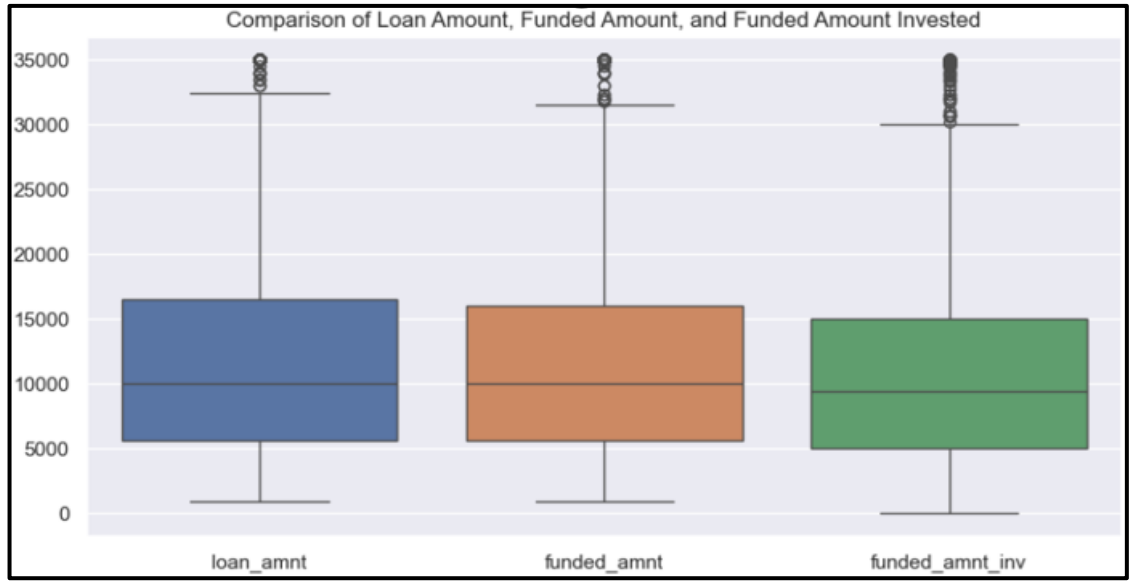
- Higher tendency to default is observed for customers with higher loan amount i.e. 15k – above.

- Higher tendency to default is observed for int bucket 13% to 17%



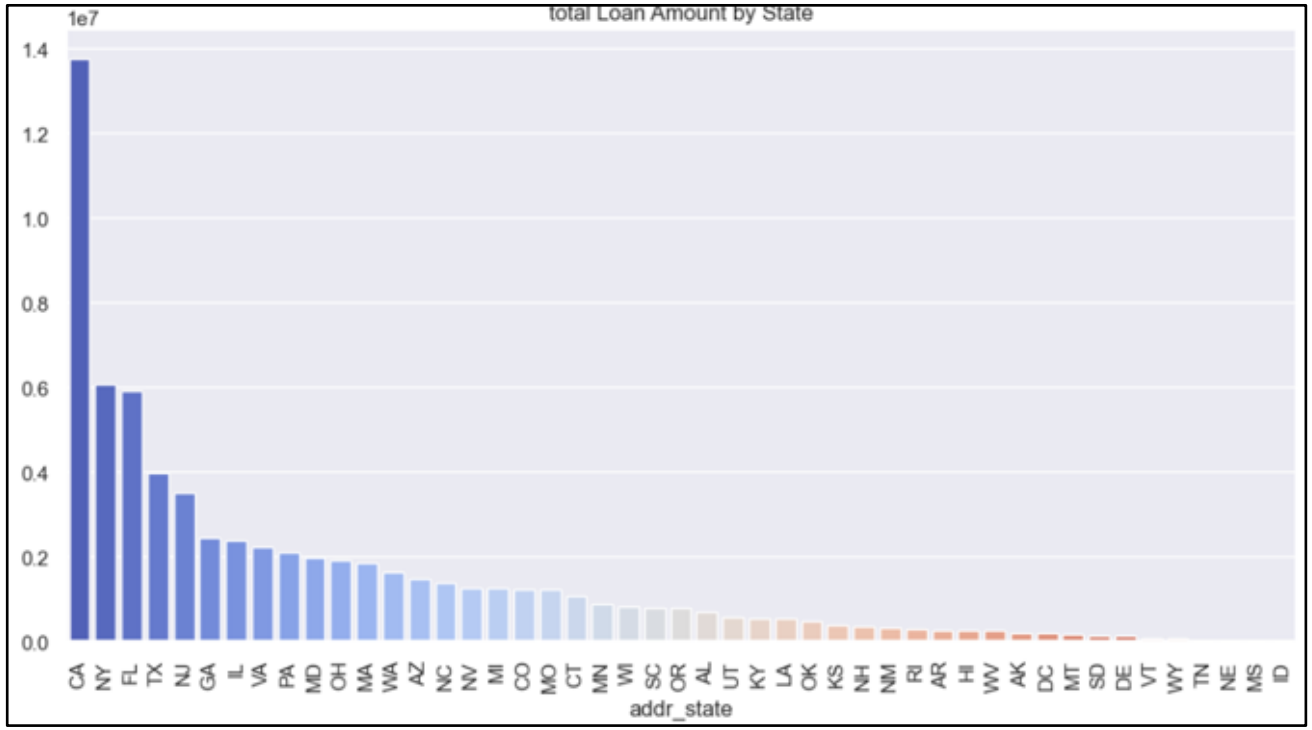
The above frequency distribution chart clearly shows most of the default customers were the ones which have there first credit line year between 1997 - 2000

Bivariate Analysis (for Charged Off Customers) of Employment Length, State with respect to Loan Amount



- There is no much difference in loan amount, funded amount and funded amount investor this behavior is in sync with the overall analysis as well so we can infer for defaulters the behavior is same there is no much difference between these amounts

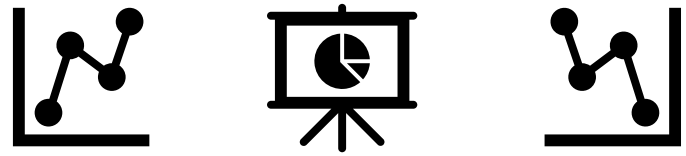
- The Loan amount Average value is higher for customers with employment length 10+, 9 and 8 years. So, this infer the customer that have higher employment length intend to take more loan

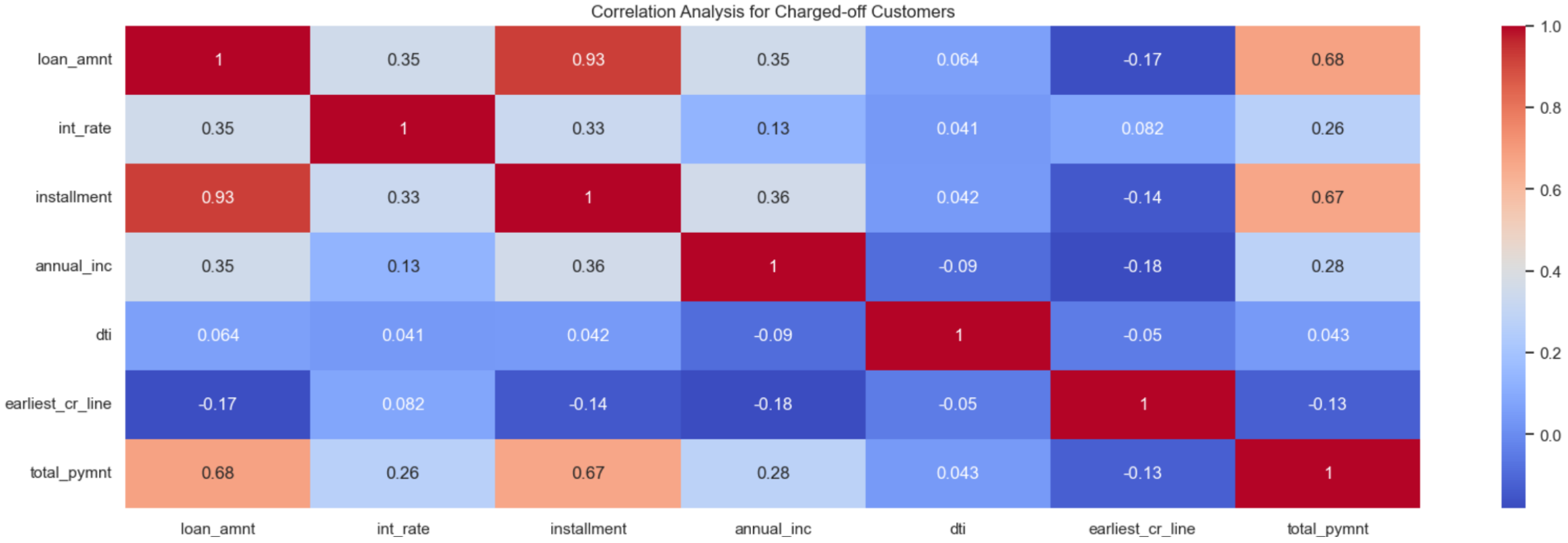


- Customers from State : CA, FL or NY have tendency to take more loan amount and this is in sync with total count of customers as well.

Note: CA: California , FL: Florid, NY: New York

Multivariate Analysis: Analysing multiple variables and checking correlation between them on a heat map





The above heat map shows there is not much correlation between the numerical variables except a high correlation between loan amount and instalment i.e. 93% but the correlation between loan amount and total payment is only 68% which clearly infer that charged off customers failed to pay the full payment on time.