LENDING CLUB CASE STUDY

BY:

HEMANTH KUMAR SURENDRA BABU

CONTENTS

PROBLEM STATEMENT

Lending Club, a consumer finance marketplace serving urban customers, faces a critical challenge in optimizing its loan approval process. The goal is to minimize **credit losses**, which occur when borrowers default on their loans—especially those labeled as "charged-off," causing the most significant financial impact.

The challenge lies in balancing two key risks:

1. **Missed Opportunities**: Rejecting applicants likely to repay results in lost business and potential revenue.
2. **Financial Losses**: Approving high-risk applicants leads to substantial defaults and monetary loss.

- **Objective:** Leverage data insights to identify risky applicants, enabling smarter decisions that minimize credit losses while maximizing profit potential.

DATA DESCRIPTION

The dataset contains detailed information about past loan applicants, their attributes, and repayment statuses. These attributes are essential for understanding applicant profiles and predicting the likelihood of loan default.

Key Components of the Dataset

- **Loan Information:**
  - loan_amnt: Loan amount requested by the applicant.
  - funded_amnt: Loan amount approved by the lender.
  - funded_amnt_inv: Final loan amount funded by investors.
  - term: Loan tenure (e.g., 36 or 60 months).
  - int_rate: Interest rate charged on the loan.
  - grade & sub_grade: Credit grades assigned to the applicant.

DATA DESCRIPTION

- **Applicant Information:**
  - emp_length: Length of the applicant's employment.
  - home_ownership: Housing status (e.g., rent, mortgage, own).
  - annual_inc: Applicant's annual income.
  - purpose: Purpose of the loan (e.g., debt consolidation, home improvement).
- **Repayment Status:**
  - loan_status: Current status of the loan (e.g., Fully Paid, Current, Charged-Off).
  - charged_off: Indicates whether the loan was defaulted.
- **Financial Metrics:**
  - dti: Debt-to-Income ratio, representing the borrower's debt burden.
  - delinq_2yrs: Number of delinquencies in the last two years.
  - revol_bal: Total credit card balance.

DATA UNDERSTANDING

- **Primary Attribute:**
  - **Loan Status**: The principal attribute of interest, consisting of three values:
    - **Fully Paid**: Customers who have successfully repaid their loans.
    - **Charged Off**: Customers who have defaulted on their loans.
    - **Current**: Loans that are still in progress and not included in the analysis.
- **Decision Matrix:**
  - **Loan Acceptance Outcomes:**
    - **Fully Paid**: Applicants who have repaid both principal and interest.
    - **Current**: Applicants currently making loan installments.
    - **Charged Off**: Applicants who have defaulted on their loans.
  - **Loan Rejection**: No transactional history available for rejected applications.

DATA UNDERSTANDING

- **Customer Demographics:**
  - **Annual Income (annual_inc)**: Higher income increases loan approval likelihood.
  - **Home Ownership (home_ownership)**: Indicates if the customer owns or rents a home.
  - **Employment Length (emp_length):** Longer employment tenure signifies financial stability.
  - **Debt to Income (dti)**: Lower DTI indicates a higher chance of loan approval.
  - **State (addr_state)**: Used for demographic analysis.
- **Loan Characteristics:**
  - **Loan Amount (loan_amt)**: Amount requested by the borrower.
  - **Grade (grade)**: Credit rating indicating loan risk.
  - **Term (term)**: Loan duration in months.
  - **Loan Date (issue_d)**: Date the loan was issued.
  - **Purpose of Loan (purpose)**: Reason for the loan (e.g., debt consolidation).
  - **Verification Status (verification_status)**: Whether the borrower's information is verified.
  - **Interest Rate (int_rate)**: Annual interest rate charged.
  - **Installment (installment)**: Monthly payment amount.

DATA CLEANING

STEPS:

1. Loading data from loan CSV
2. Checking for null values in the dataset
3. Checking for unique values
4. Checking for duplicated rows in data
5. Dropping records
6. Handling Missing Values
7. Outlier Treatment

DATA CLEANING & PRE-PROCESSING

After loading the data, followed the below steps:

1. Removed the null values, we reduced the data set from 111 columns to 57 columns.
2. Checked for unique values.
3. No duplicated values were found in the dataset
4. Removed these columns from my dataset as they don't add value to my approach 'pymnt_plan', "initial_list_status",'collections_12_mths_ex_med','policy_code','acc_now_delinq', 'application_type', 'pub_rec_bankruptcies', 'tax_liens', 'delinq_amnt'
5. After bring the data down to 48 columns, I inspected and further removed "id", "member_id", "url", "title", "emp_title", "zip_code", "last_credit_pull_d", "addr_state","desc","out_prncp_inv","total_pymnt_inv","funded_amnt", "delinq_2yrs", "revol_bal", "out_prncp", "total_pymnt", "total_rec_prncp", "total_rec_int", "total_rec_late_fee", "recoveries", "collection_recovery_fee", "last_pymnt_d", "last_pymnt_amnt", "next_pymnt_d" , "chargeoff_within_12_mths", "mths_since_last_delinq", "mths_since_last_record"
6. The columns were finally at 21 and now started to address the missing values.

DATA CLEANING & PRE-PROCESSING

Handling Missing values and outliers:

1. Columns with missing values are emp_length and revol_util, used mode and added values to them.
2. Dropped all the NA values from all the columns.
3. Standardized the data.
4. Removed '%', '+' and ‘<‘ from revol_util, int_rate and emp_length.
5. Removed outliers above 95% using quantile approach.
6. Referred to the column named sub_grade in the loan DataFrame. This column contains strings representing sub-grades such as "A1", "B2", "C3", etc. I have converted the extracted string values (e.g., "1", "2", "3") into numeric data types (int or float).

UNIVARIATE ANALYSIS

Univariate analysis is the simplest form of data analysis, focusing on a single variable. The goal is to understand the distribution and characteristics of that variable.

A **countplot** is a type of bar plot used to visualize the count of observations in each categorical bin using bars. It's particularly useful for displaying the distribution of a categorical variable.

Here are some key points about countplots:

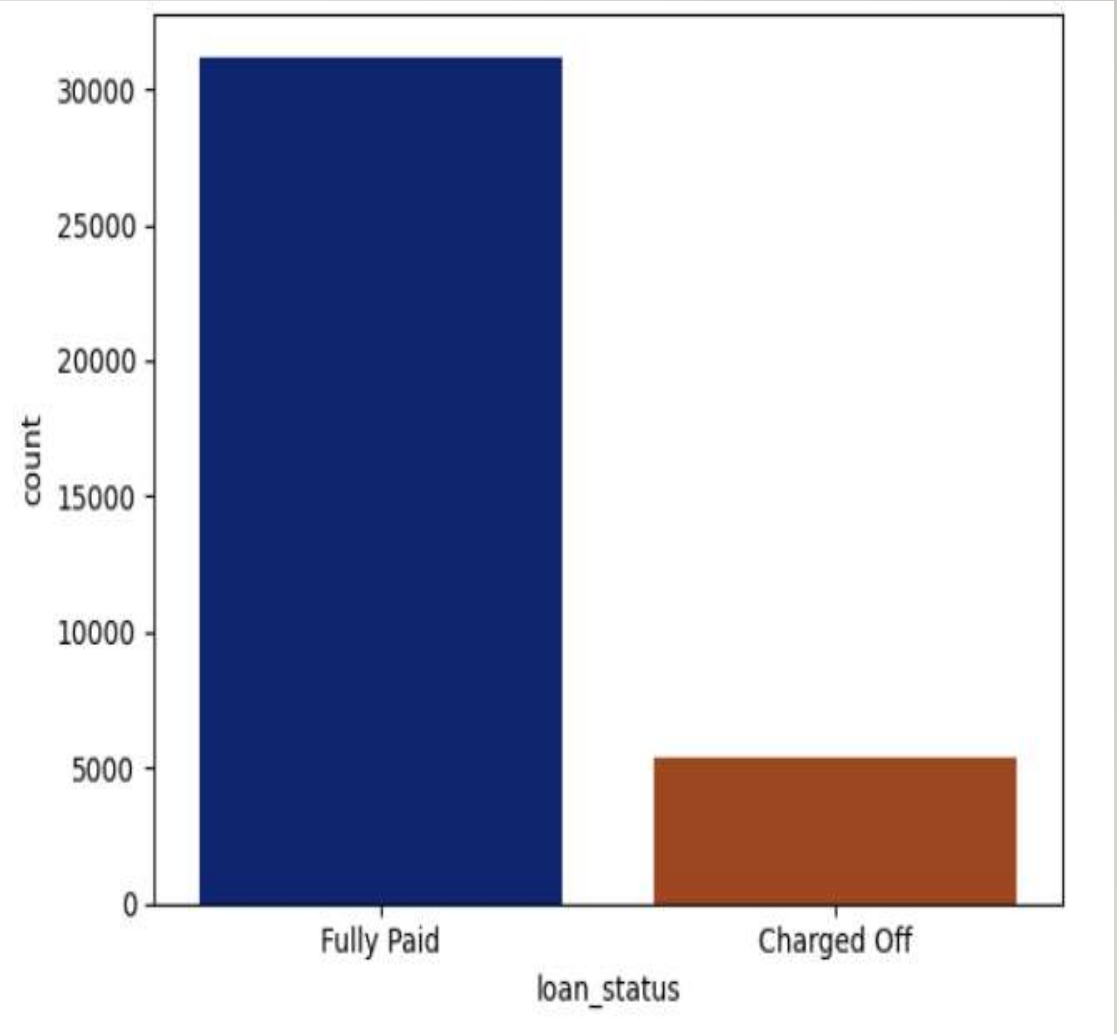
- **Purpose**: Countplots show the frequency of each category in a dataset. This helps in understanding the distribution and identifying any imbalances in the data.
- **Usage**: They are commonly used in exploratory data analysis (EDA) to get a quick overview of the data. For example, you can use a countplot to see how many passengers were in each class on the Titanic.
- **Customization**: Countplots can be customized in various ways, such as changing the color palette, adding a hue to show additional categorical variables, and adjusting the orientation (vertical or horizontal).
- **Implementation**: In Python, countplots are typically created using the Seaborn library

UNIVARIATE ANALYSIS

- **Categorical Variables**:
  - Ordered:
    - Grade (grade)
    - Sub grade (sub_grade)
    - Term (36 / 60 months) (term)
    - Employment length (emp_length)
    - Issue year (issue_year)
    - Issue month (issue_month)
  - Unordered:
    - Loan purpose (purpose)
    - Home Ownership (home_ownership)
    - Loan status (loan_status)
- **Quantitative Variables:**
    - Interest rate group (int_rate_groups)
    - Open account bucket (open_acc_groups)
    - Revolving line utilization group(revol_util_groups)
    - Total account group (total_acc_groups)
    - Annual Income group (annual_inc_groups)

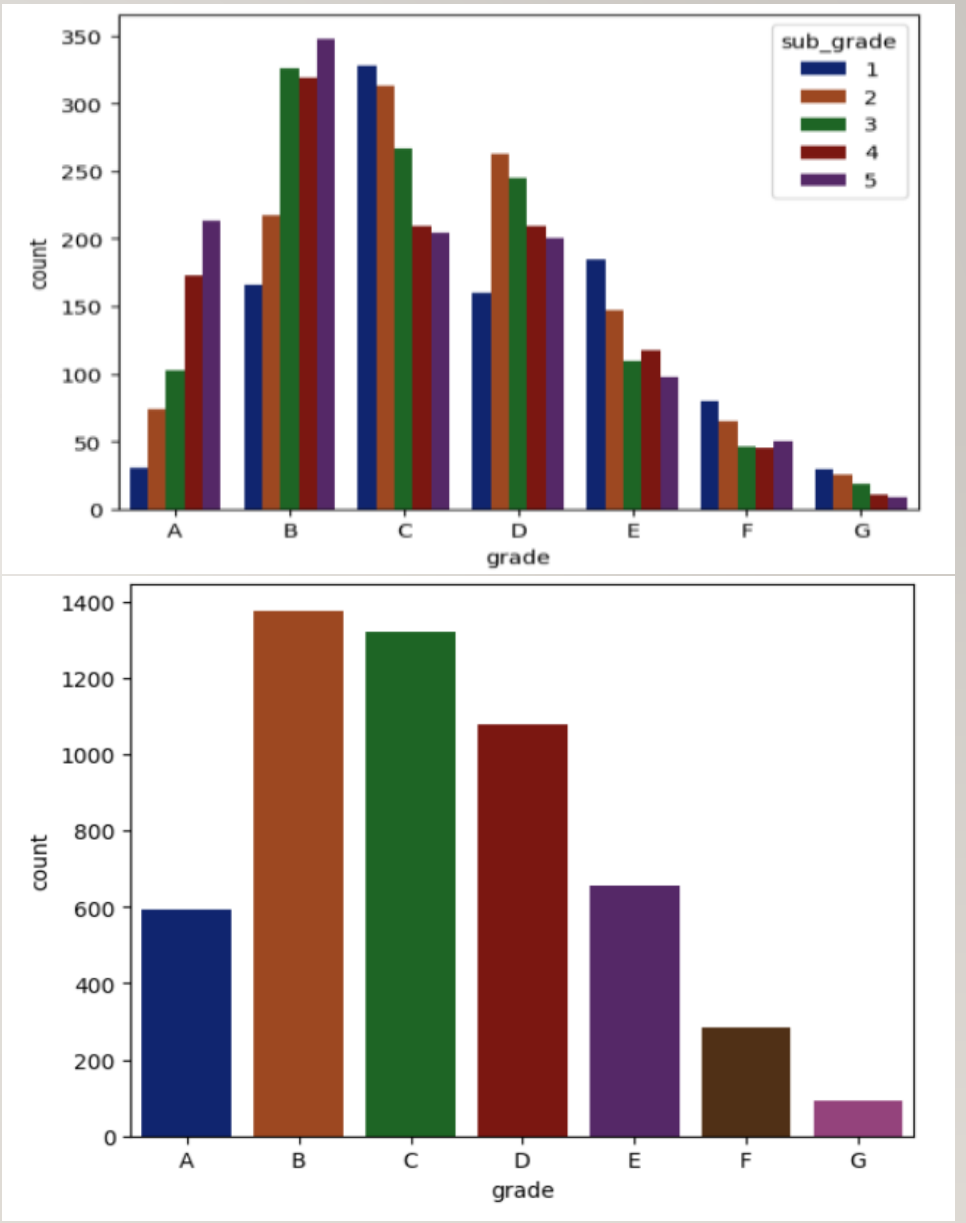
UNIVARIATE ANALYSIS

- We are only looking at defaulter data and sub-setting the data while plotting only for 'Charged Off’



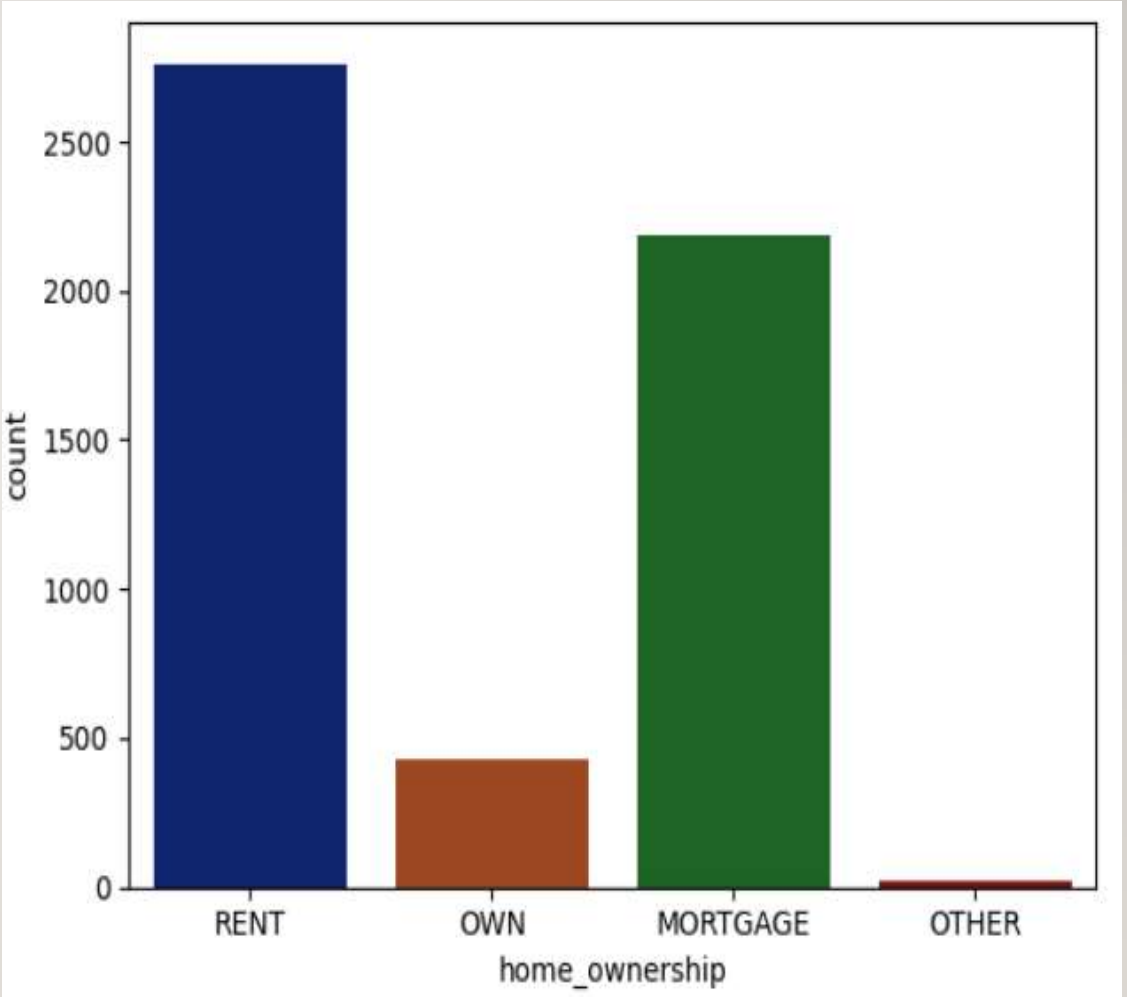
UNIVARIATE ANALYSIS

- Loan applicants in Grades B, C, and D account for the majority of "Charged Off" loans.
- Subgrades B3, B4, and B5 are associated with a higher likelihood of charge-offs.



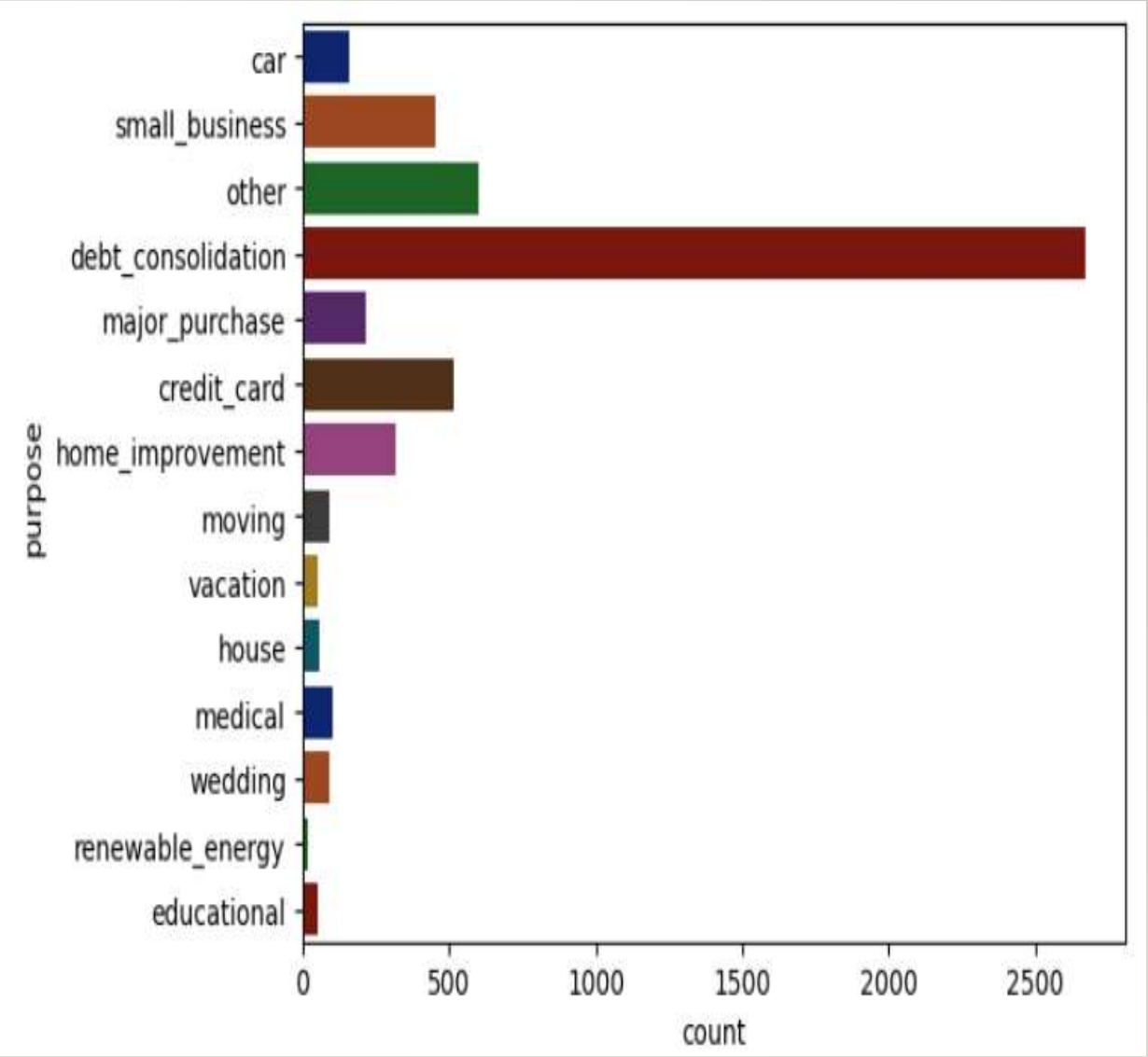
UNIVARIATE ANALYSIS

- Applicants living in rented or mortgaged houses are more likely to default. This information can be considered in the underwriting process to assess housing stability and its impact on repayment ability.



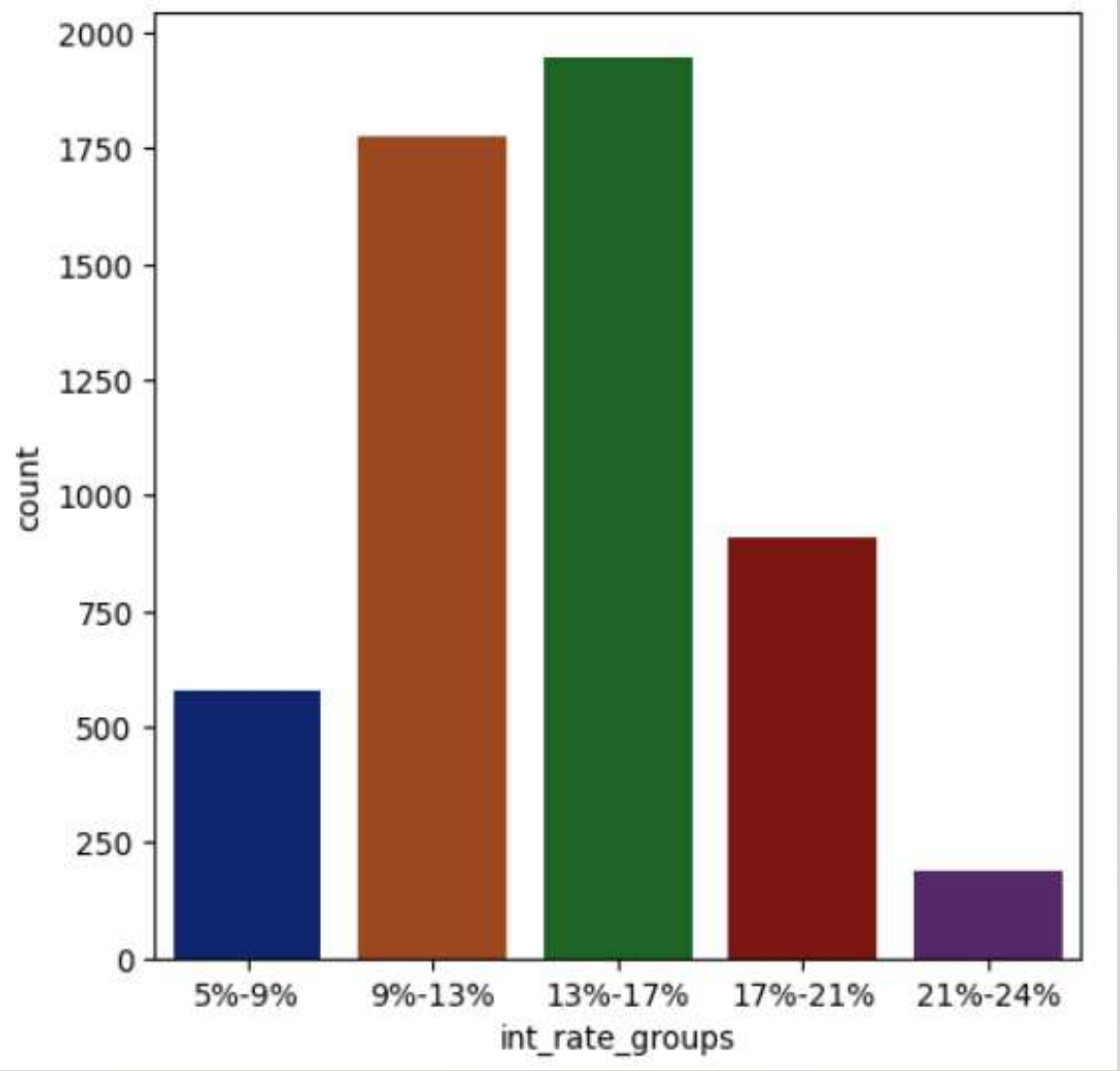
UNIVARIATE ANALYSIS

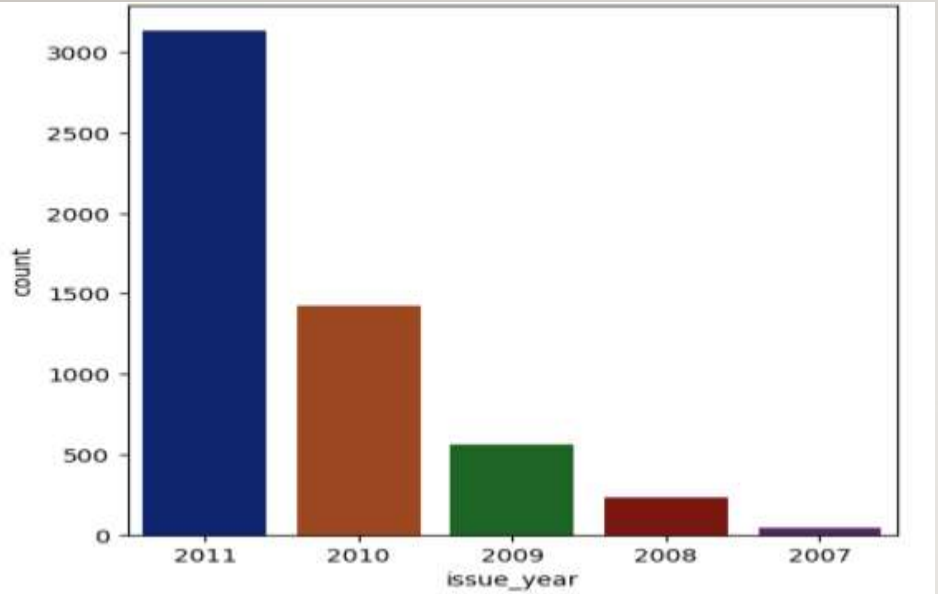
- Debt consolidation is the most common loan purpose, and it also has the highest number of defaults.

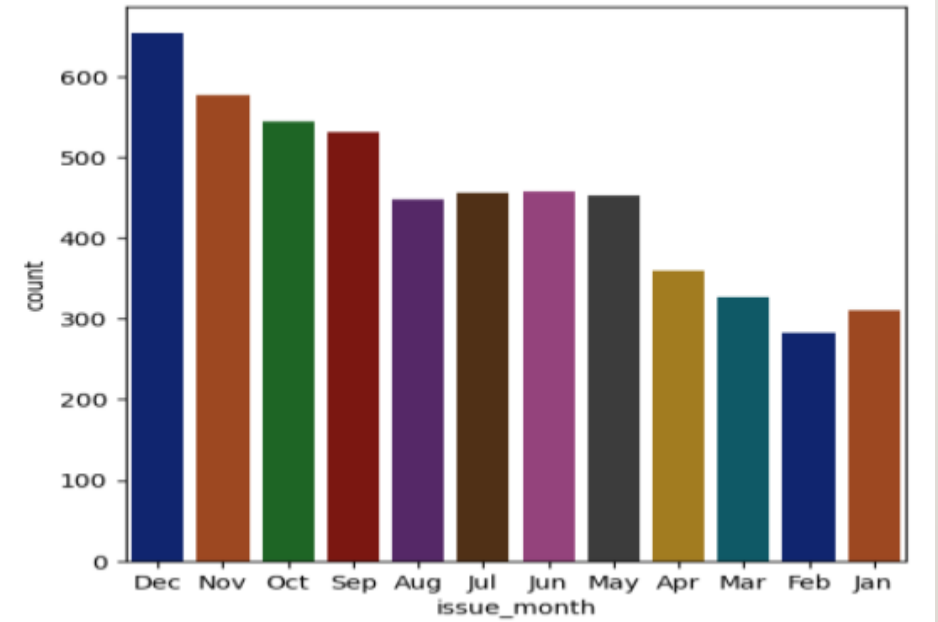


UNIVARIATE ANALYSIS

- A notable proportion of applicants who defaulted had loans with interest rates between 13% and 17%.







UNIVARIATE ANALYSIS

- The number of loan applicants steadily increased from 2007 to 2011, suggesting a growth trend in subsequent years.
- The visualizations indicate the defaulted applicants were sanctioned the maximum number of loans during December 2011 compared to other years.
- December is the most popular month for taking loans, likely due to holiday-related expenses.

BIVARIATE ANALYSIS

Bivariate analysis is a statistical method used to explore the relationship between two variables. It helps determine if there is an association between the two variables and the strength of that relationship

A **Bar Plot** is a useful tool in bivariate analysis, especially when dealing with categorical data. It helps visualize the relationship between two categorical variables by displaying the frequency or proportion of combinations of the categories.

Here are some common types of barplots used in bivariate analysis:

Purpose: Bar plots in bivariate analysis are used to visualize the relationship between two categorical variables. They help in understanding the distribution and frequency of categories and how they interact with each other. This can reveal patterns, trends, and potential correlations between the variables.

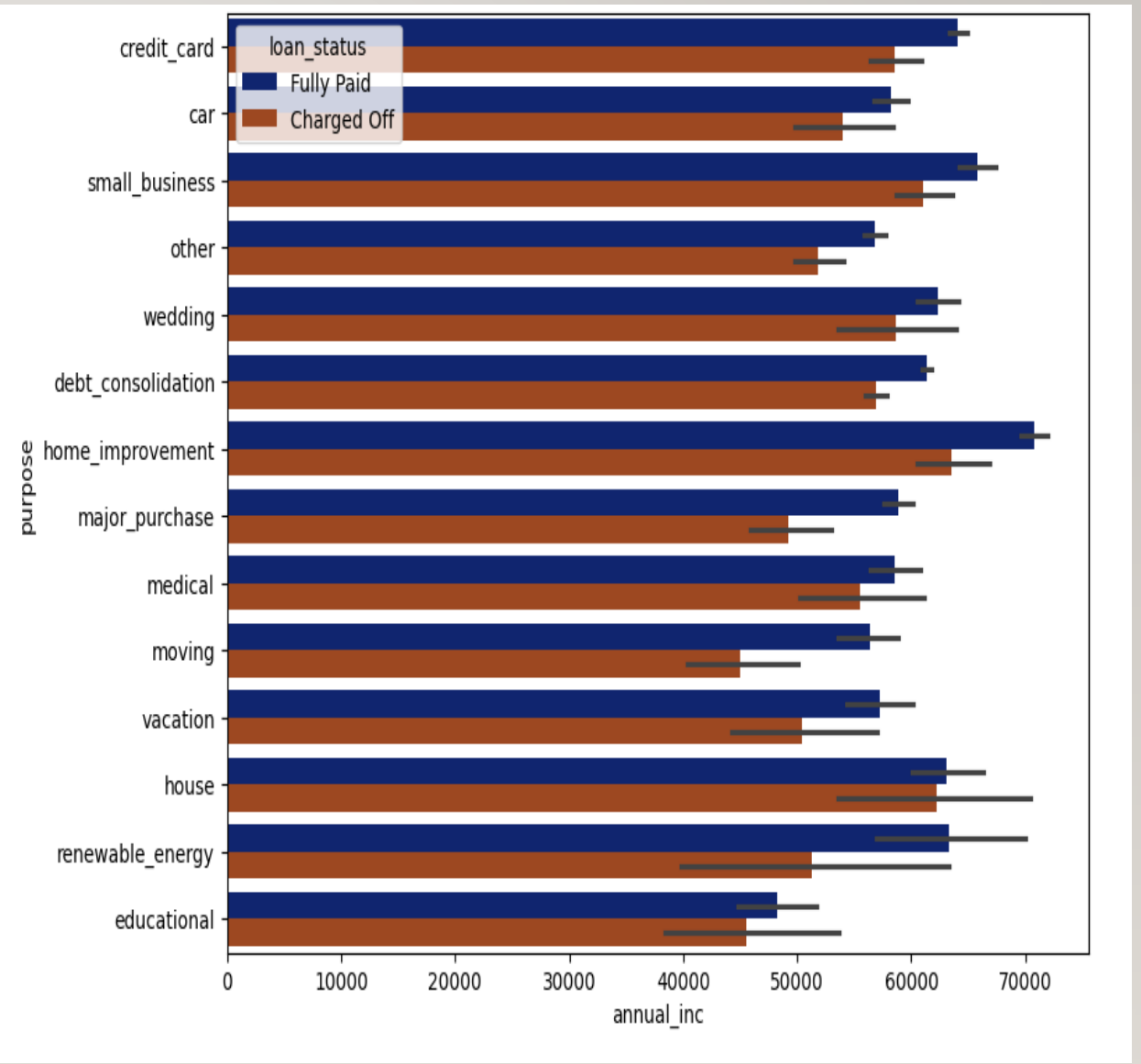
Usage: Bar plots are particularly useful when you want to compare the frequency of categories across different groups, Identify patterns or anomalies in the data, Present categorical data in a clear and concise manner

Customization: Bar plots can be customized in various ways to enhance their readability and visual appeal like Colors, Labels, Orientation, Bar Width and Spacing

Implementation: In Python, Bar plots are typically created using the

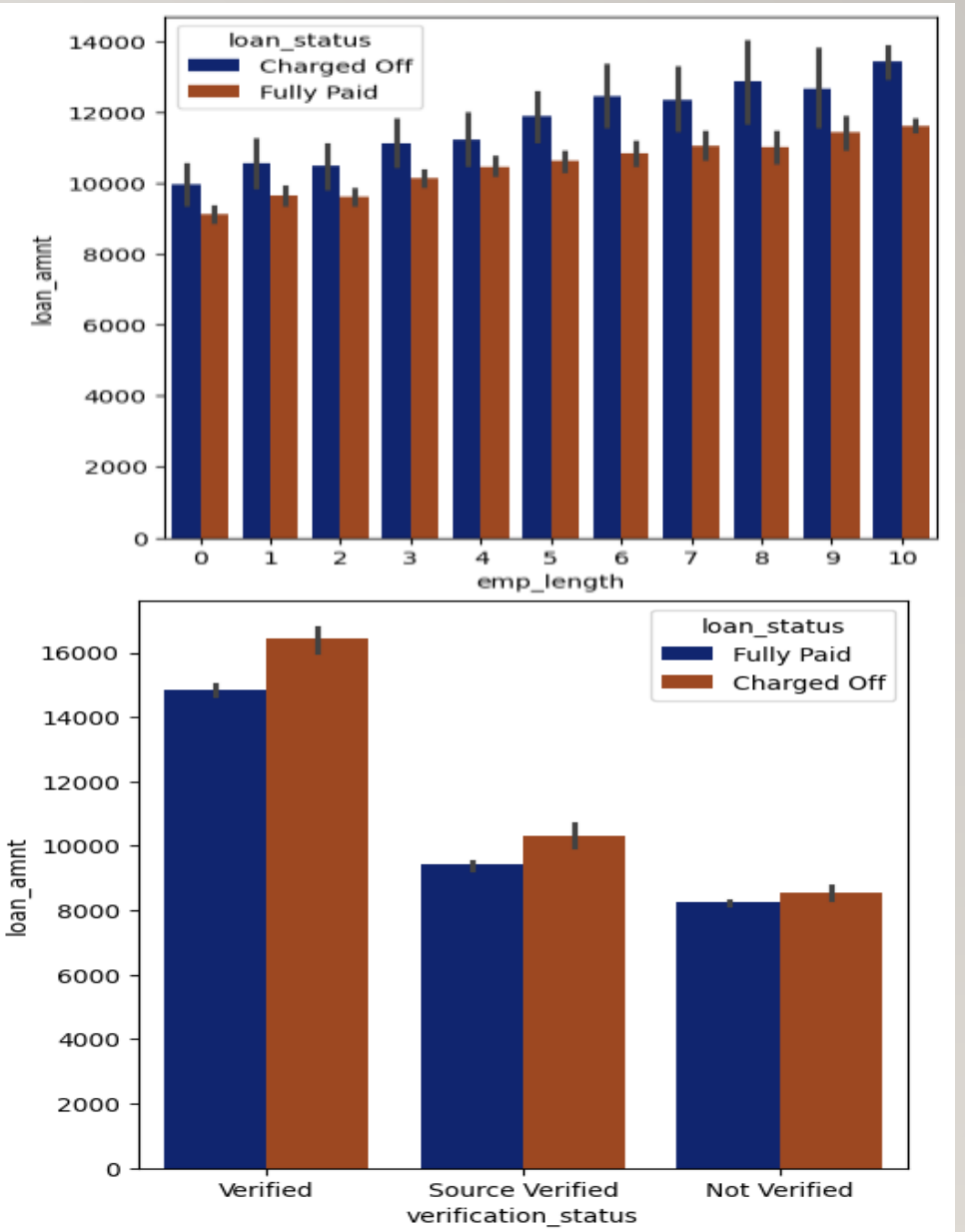
BIVARIATE ANALYSIS

- The previous observations indicated that "debt_ consolidation" defaulted the highest among the others. However, the annual income is comparatively less for those applicants.
- The Applicants with higher salaries have mostly taken loans for "home improvement", "house", "renewable_enery" and "small_business"



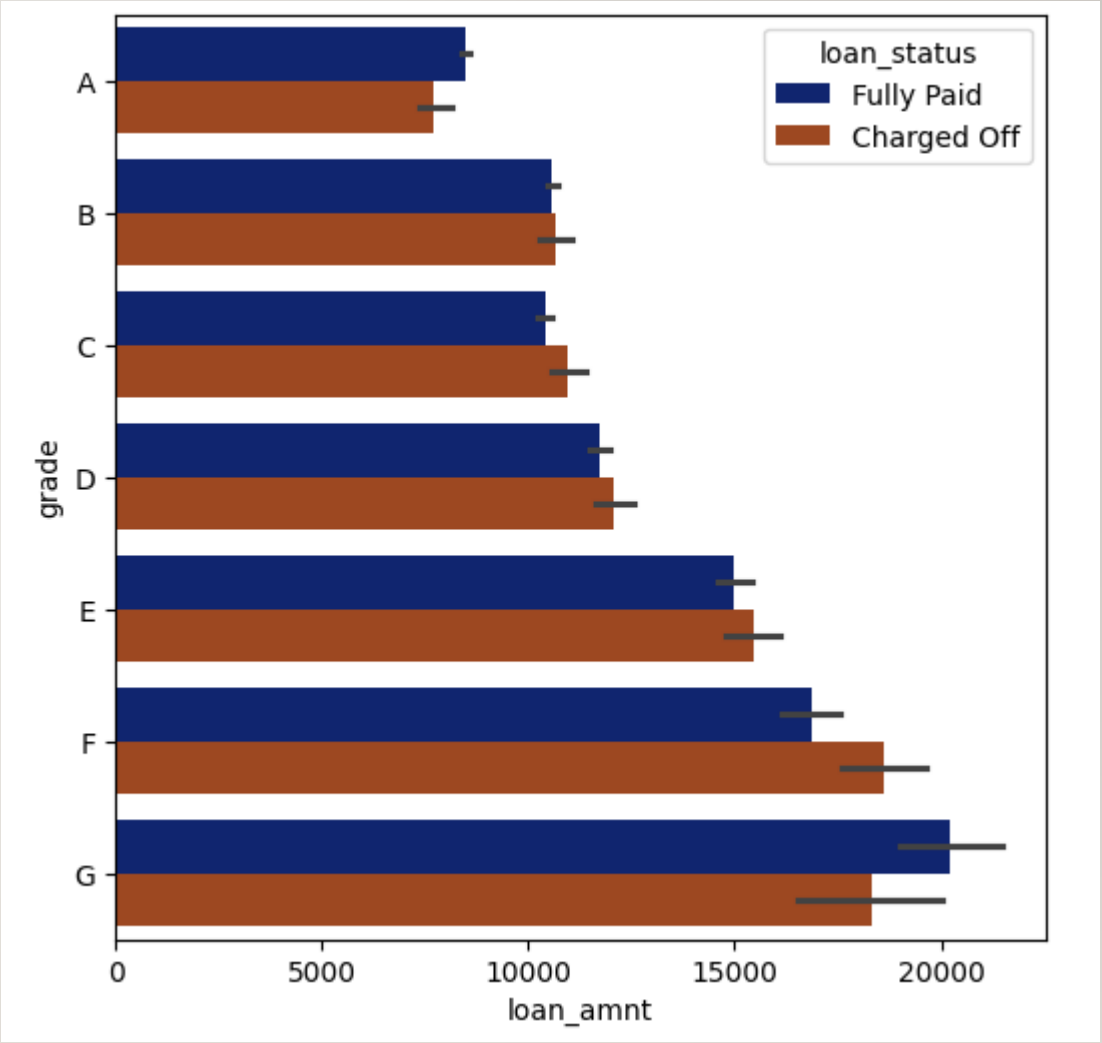
BIVARIATE ANALYSIS

- Loan applicants with ten or more years of experience are more likely to default. This suggests that experience alone may not be a reliable indicator of creditworthiness. The company should use a more comprehensive credit scoring system that factors in other risk-related attributes.
- Applicants Verified loan applicants default more than those who are not verified. The company should review its verification process to ensure it effectively assesses applicant creditworthiness and consider improvements or adjustments.



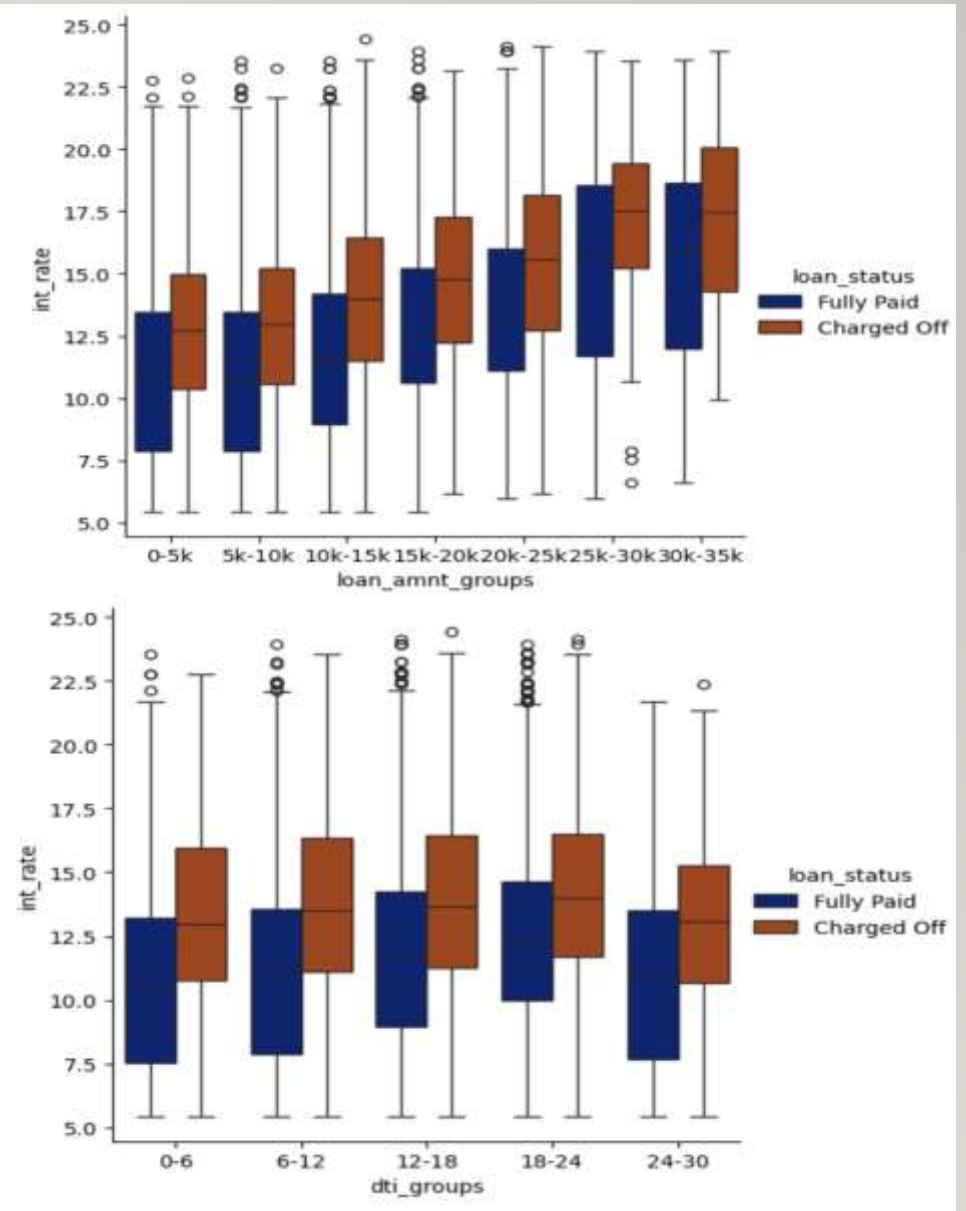
BIVARIATE ANALYSIS

- Loan applicants from Grades B, C, and D account for the majority of "Charged Off" loans. To address this, the company should consider applying stricter risk assessment and underwriting criteria for applicants in these grades.
- Focus on applicants with Subgrades B3, B4, and B5, as they have a higher likelihood of charging off. To mitigate this risk, consider implementing additional risk management strategies or offering these applicants lower loan amounts.



BIVARIATE ANALYSIS

- Applicants who receive loan amounts of $15,000 or more are at a higher risk of defaulting. To reduce this risk, the company should consider implementing more rigorous evaluations for larger loan requests and possibly setting a cap on loan amounts for applicants deemed higher risk.
- High Debt-to-Income (DTI) ratios and interest rates between 13% and 17% are linked to higher default rates. To mitigate this, the company should reassess its interest rate determination process and consider adjusting rates based on DTI ratios to better match borrowers' repayment capacities.



SUMMARIZATION

- Loan applicants in Grades B, C, and D account for most "Charged Off" loans.
- Subgrades B3, B4, and B5 are associated with a higher likelihood of charge-offs.
- Applicants seeking loans with a 60-month term are more prone to default than those opting for 36-month terms.
- Most loan applicants have ten or more years of experience, and this group also exhibits the highest default rates.
- The number of loan applicants steadily increased from 2007 to 2011, suggesting a positive growth trend in subsequent years.
- December is the most popular month for taking loans, likely due to holiday-related expenses.
- Debt consolidation is the most common loan purpose, and it also has the highest number of defaults.
- Applicants living in rented or mortgaged homes are more likely to default on their loans.
- Verified loan applicants have higher default rates compared to those who are not verified.
- Loan applicants who charged off typically had significantly high Debt-to-Income (DTI) ratios.
- A notable proportion of applicants who defaulted had loans with interest rates between 13% and 17%.

FINAL CONCLUSION

- **Grades B, C, and D:** Tighten risk assessment and underwriting criteria for these grades.
- **Subgrades B3, B4, and B5:** Introduce additional risk controls, lower loan amounts, or higher rates.
- **Term Length:** Limit 60-month loans or adjust interest rates for longer terms.
- **Experience and Default Probability:** Develop a holistic credit scoring model beyond experience alone.
- **Positive Growth Trend:** Leverage market growth while maintaining robust risk management.
- **Debt Consolidation Risk:** Enhance screening and offer financial counseling for debt consolidation loans.
- **Housing Status and Default Risk:** Incorporate housing stability into risk assessments and loan terms.
- **Verification Process:** Review and refine the verification process to improve accuracy.
- **High Loan Amounts:** Cap loan amounts and implement stricter evaluations for large requests.
- **DTI and Interest Rates:** Align interest rates with DTI ratios and impose stricter DTI thresholds.