Leading Club Case Study

Contributor : Vijayita Azad Vereesh

Problem Statement :

The consumer finance company is focused on lending various types of loans to urban customers and must navigate two main risks in loan approval decisions:

1. **Loss of Business Risk**: Not approving loans for applicants who are likely to repay leads to lost business opportunities.
2. **Financial Loss Risk**: Approving loans for applicants who are likely to default results in significant financial losses for the company.

Historical data on past loan applicants, including whether they defaulted, is available for analysis. The objective is to identify patterns that indicate an applicant's likelihood to default. This analysis will help the company make informed decisions regarding loan approvals, including potential actions like loan denial, adjusting loan amounts, or applying higher interest rates for risky applicants.

The company has historical data on past loan applicants, including whether they defaulted or not. The objective is to analyze this data using EDA to identify patterns that indicate the likelihood of default. Insights gained will inform decisions such as loan denial, adjusting loan amounts, or setting higher interest rates for higher-risk applicants.

Objective:

Ultimately, the goal is to develop a comprehensive understanding of the driving factors behind loan defaults, allowing for better risk assessment and portfolio management.

Data Source and detailing :

- Source is laon.CSV file. (added zip file in git repo)
- Data dictionary details every column of data. Link

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts

* As a first step deep dive into each column to understand data and its relevance. This would help in data analysis and data cleaning as well .

Data Cleaning :

Data cleaning is a crucial step in exploratory data analysis (EDA) that ensures the dataset is accurate, complete, and ready for analysis. Here are the key steps used in data cleaning during EDA of loan.csv:

1. **Handling Missing Values**
- **Identify Missing Values**: Check for missing values in the dataset using methods like .isnull() or .isna().
- **Imputation**: Fill in missing values using appropriate techniques:
    - Dropping rows or columns with excessive missing values.

2. Removing Duplicates

- **Identify Duplicates**: Use methods like .duplicated() to find duplicate entries.
- **Remove Duplicates**: Drop duplicate rows to ensure each record is unique using .drop_duplicates().

3. Data Type Conversion

- **Check Data Types**: Use .dtypes or df.info() to inspect the data types of each column.
- **Convert Data Types**: Change data types as necessary (e.g., *converting strings to datetime, or floats to integer*s) to facilitate accurate analysis.
- Removing “Months” from term to make it float field
- Removing “+years” from amp length .

4. Outlier Detection and Treatment

- **Identify Outliers**: Use statistical methods and box plot to detect outliers in numerical features.
- **Handle Outliers**: Decide on a strategy to treat outliers (e.g., *capping via using quantiles, transformation, or remova*l).

6. Feature Engineering

- **Binning/ Bucketing** : Group continuous variables into discrete bins for better analysis.(eg : *annual income , loan amount , funded amount*)

7. Count Values :

- Identifying unique values in column . If column have all values same in entire column its of no use in analysis.

Univariate Analysis :

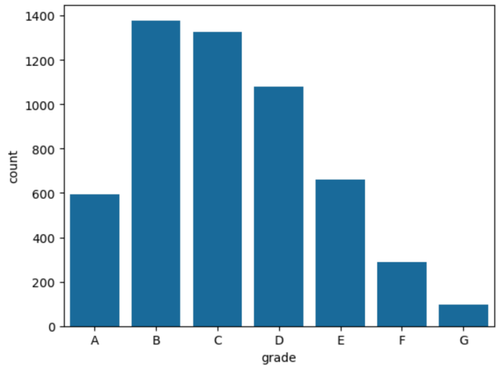
Data analysis, focusing on a single variable at a time. It provides insights into the distribution, central tendency, and variability of that variable.

Steps for Univariate Analysis:

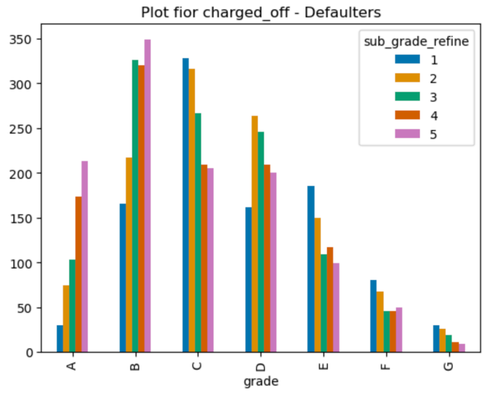
1. **Load Data** : Ensure dataset is well cleaned and is ready for analysis.
2. **Analyze Numerical Variables**
    - Calculate descriptive statistics.
    - Visualize using histograms, box plots, and density plots.
3. **Analyze Categorical Variables**
    - Count and visualize using bar charts or pie charts
    - Columns Considered : Grade ,Sub Grade, Address state , Term , Issue_d , Home Ownership , Loan Status, Purpose
4. **Analyze Analyze Numerical Variables**
    - Count and visualize using box plot
    - Column Considered : Annual Income , Funded Amount , Loan Amount , DTI , Interest Rate .

Categorical and Numerical bar plots :

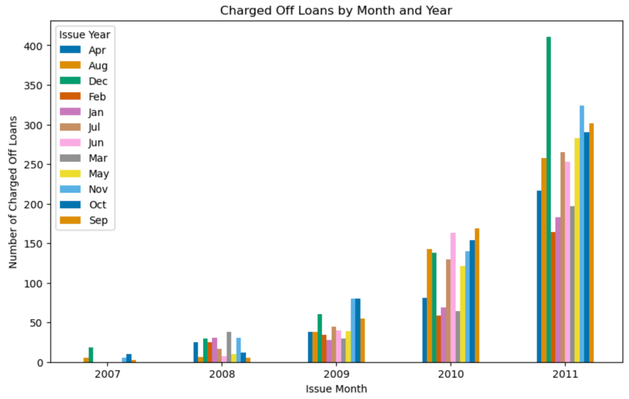
** Categorical is covering ordered and unordered both here



Analysis: Applicants belonging to "grade B" accounted for highest number of "charged off"



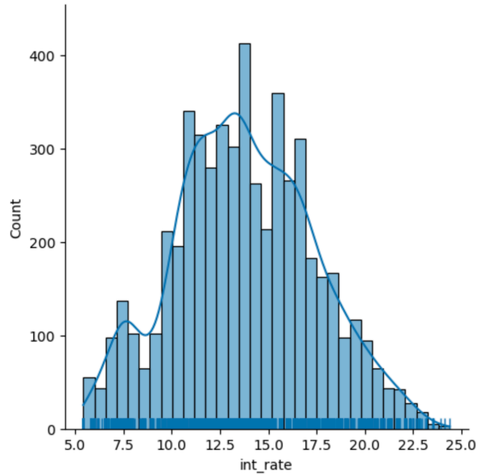
Analysis: Applicants belonging to "Sub grade B5" accounted for highest number of "charged off"



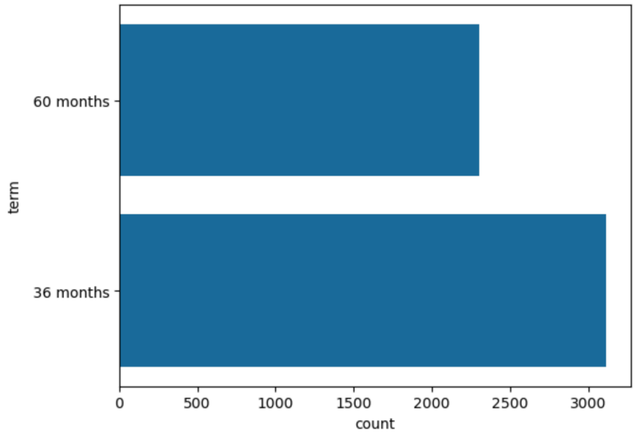
Analysis : Dec month of 2011 have max "Charged Off " applicants

Categorical and Numerical Count plots :

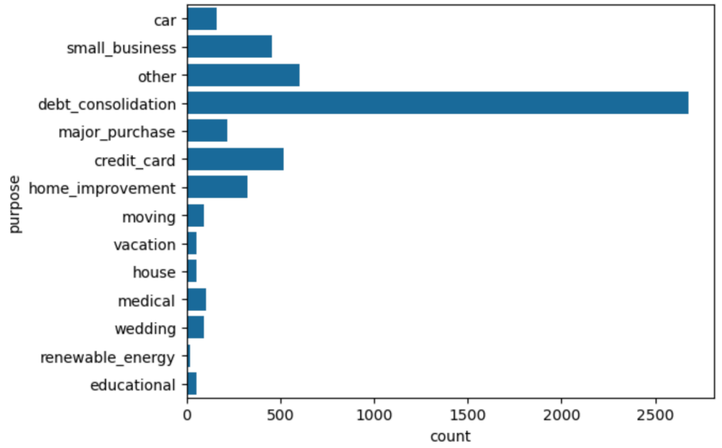
** Categorical is covering ordered and unordered both here



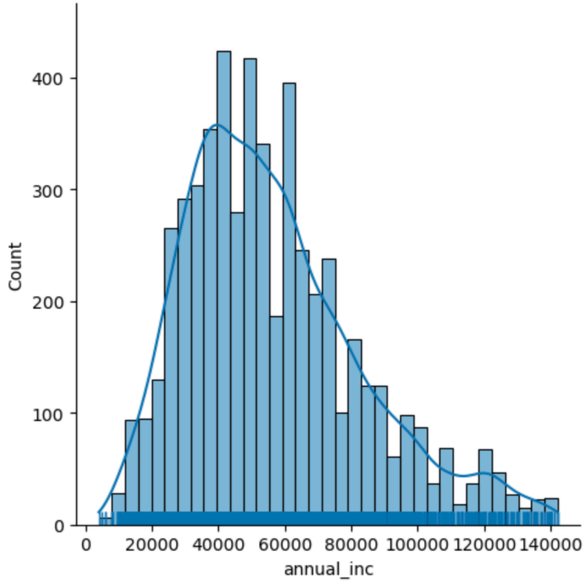
Analysis : This shows avg or most of interest rates are in range of 13-17% for charged off aplicants

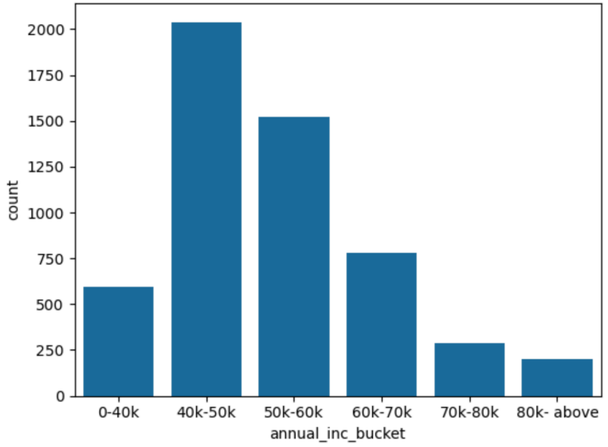


Analysis : Applicants with term of 36 months likely to be "charged off"

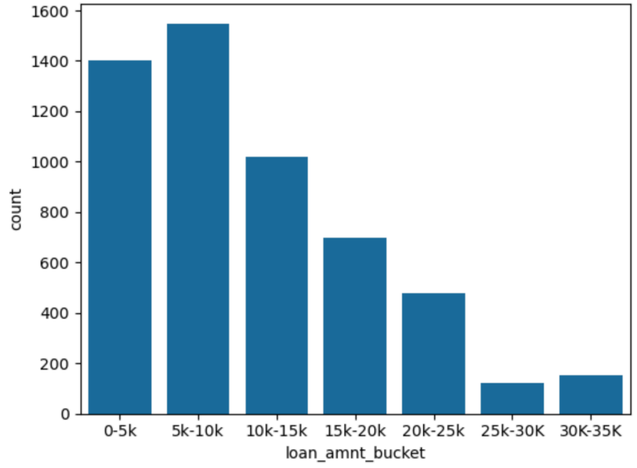


Analysis : Applicants with purpose of "dept consideration" accounting for charged off





Analysis : Applicants with annual income of 40-50K accounting for charged off



Analysis : Applicants with loan ammount of 5-10K are mostly being defaulters .

Summary of Univariate Analysis :

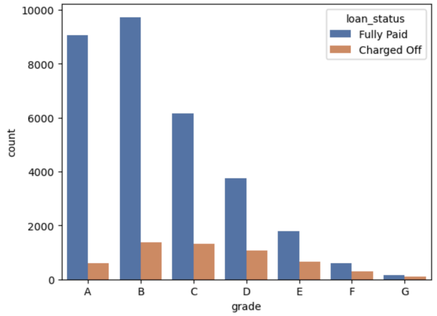
Applicants Likely to Default Are Those Who:

1. **Credit Grade**:
    - Have a grade of **'B'**.
2. **Sub-Grade**:
    - Fall into the sub-grade **'B5'**.
3. **Interest Rate**:
    - Have an interest rate between **10% and 17%**.
4. **Annual Income**:
    - Have an annual income ranging from **$40,000 to $60,000**.
5. **Employment Length**:
    - Have an employment length of **10 years or more**.
6. **Loan Amount**:
    - Request a loan amount between **$5,000 and $10,000**.
7. **Loan Term**:
    - Opt for a loan term of **36 months**.
8. **Loan Purpose**:
    - Specify the purpose as **'debt consolidation'**.
9. **Loan Issuance Date**:
    - Were issued loans in **December 2011**, potentially due to festive spending.

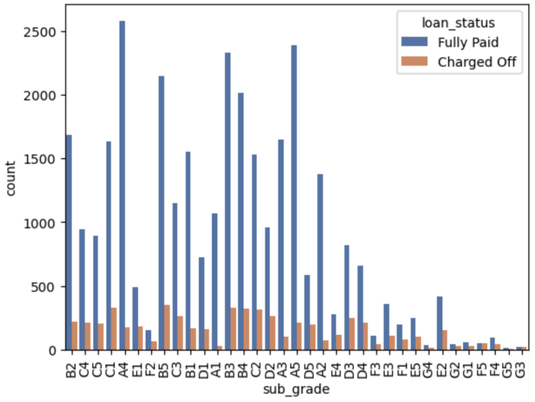
Bivariate Analysis :

Bivariate analysis examines the relationship between two variables to identify patterns and correlations that may indicate risk factors for loan defaults. In the context of loan applicants, several key insights emerge:

1. **Credit Grade vs Default Status**: A significant relationship exists between credit grades and default rates. Applicants in lower grades, particularly grades **B, C, and D**, are more likely to default compared to those in higher grades. This trend highlights the importance of credit grading in assessing risk.

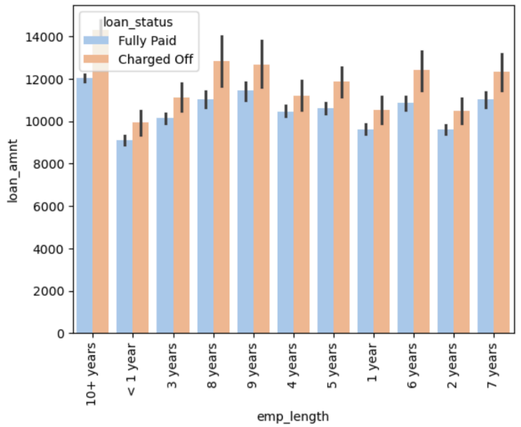


2. **Sub-Grade Analysis**: A closer examination of sub-grades reveals that specific sub-grades, such as **B3, B4, B5, and C1**, show elevated default rates. This indicates that even within a single grade, certain sub-grades carry more risk.

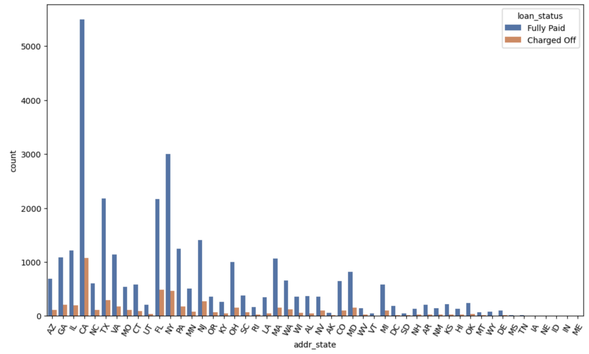


Bivariate Analysis :

3. **Employment Length**: The duration of employment also correlates with default likelihood. Applicants with **10 or more years** of employment demonstrate a higher propensity to default, suggesting that long-term employment may not always equate to financial stability.

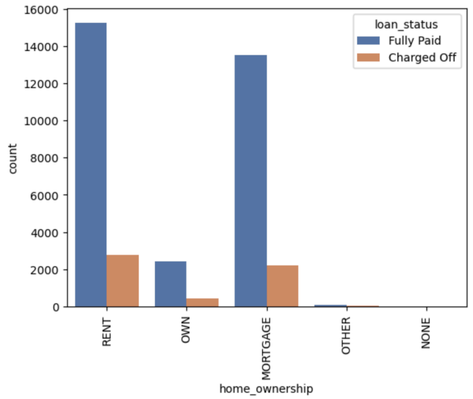


4. **Geographic Influence**: Geographic factors play a critical role in default rates. States such as **California (CA), Florida (FL), and New Jersey (NJ)** are associated with higher default rates, pointing to regional economic conditions that may affect borrowers' ability to repay.

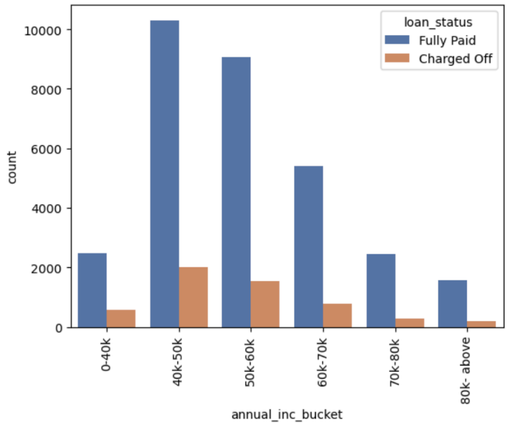


Bivariate Analysis :

5. **Housing Situation**: Analysis shows that applicants living in **rented or mortgaged homes** have a greater likelihood of default compared to homeowners. This may reflect the financial pressures faced by those who do not own their homes outright.

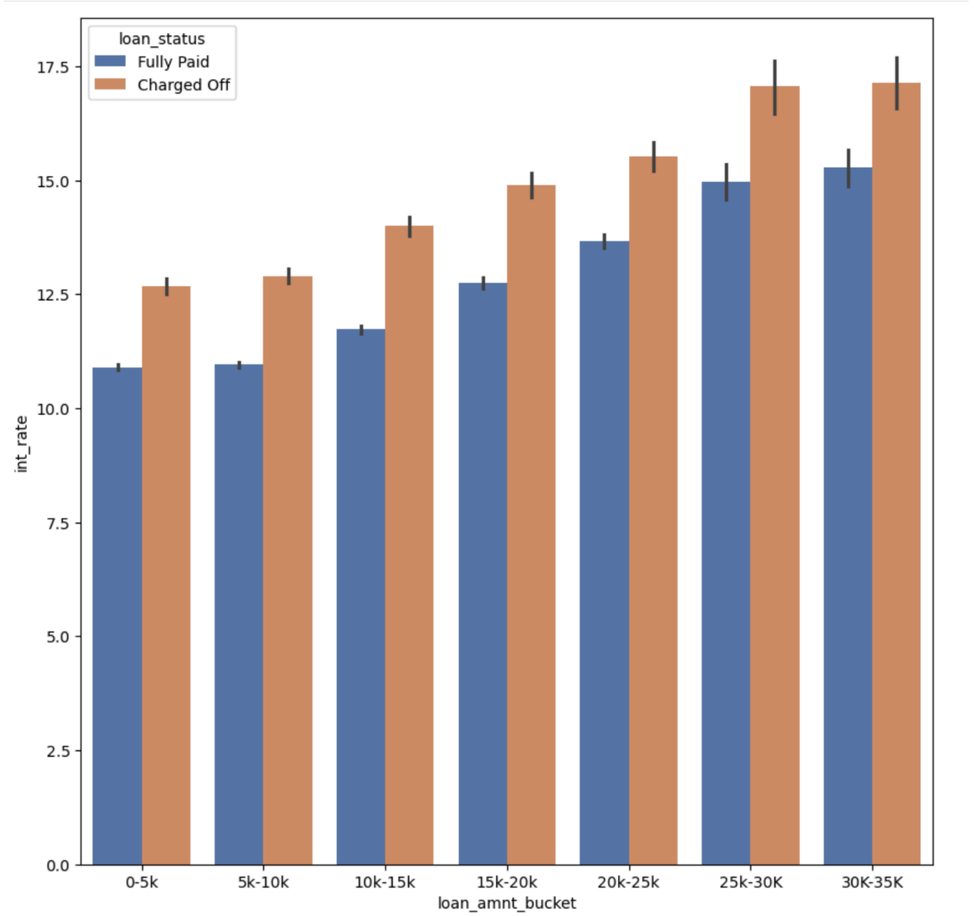


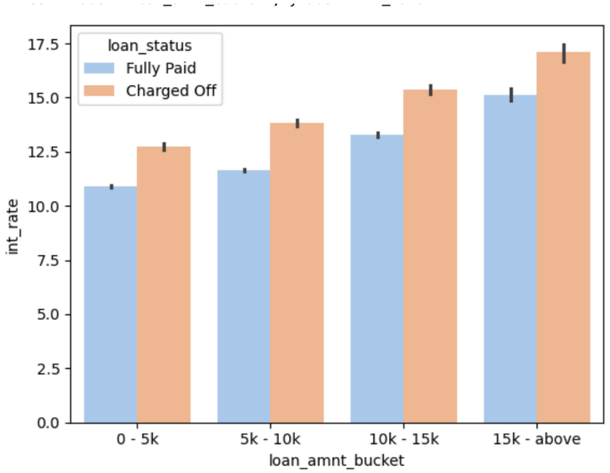
6. **Income Level**: A clear trend emerges as lower annual incomes correlate with higher default rates. This relationship emphasis the financial challenges faced by lower-income applicants.

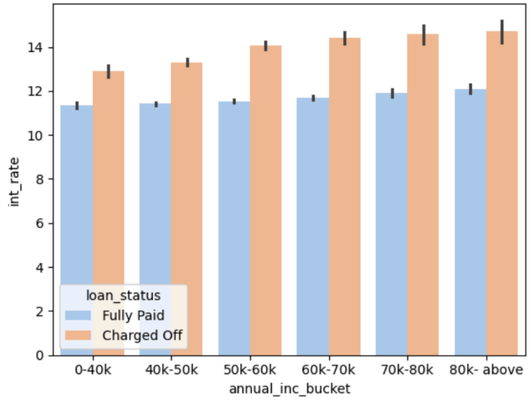


Bivariate Analysis :

7. **Interest Rate Impact vs loan amount** : The likelihood of default increases with rising interest rates, particularly affecting those with lower incomes or longer loan terms.







Summary of Bivariate Analysis :

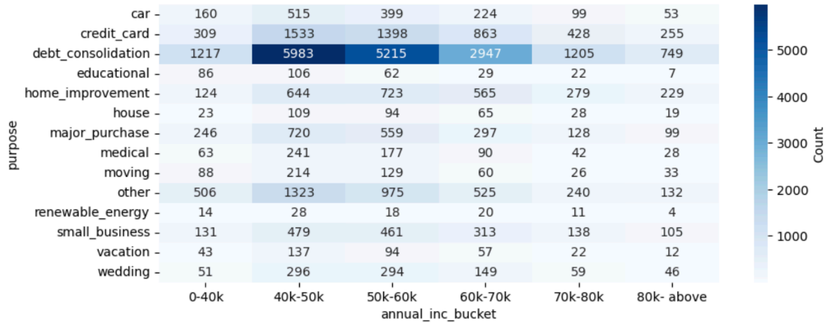
The bivariate analysis of loan applicants reveals several significant relationships that contribute to the understanding of default risk:

1. **Credit Grades and Charged Off Loans**: Applicants in grades **B, C, and D** account for the majority of "Charged Off" loans, indicating that lower credit grades are a strong predictor of default.
2. **Geographic Trends**: Loan applicants from **California (CA)**, **Florida (FL)**, and **New York (NY)** are found to be the most likely to default, highlighting regional economic factors that may influence repayment ability.
3. **Sub-Grade Impact**: Specific sub-grades, particularly **B3, B4, and B5**, show a higher propensity for charge-offs, suggesting that certain classifications within grades carry more risk.
4. **Housing Situation**: Borrowers who are renting or mortgaging their homes are more likely to charge off, indicating financial instability associated with housing costs.
5. **Interest Rates and Default Risk**: Applicants who defaulted typically received loans with interest rates exceeding **13%**, demonstrating that higher borrowing costs are linked to increased risk.
6. **Loan Amount and Purpose**: Applicants who took large loans for **Small Business** purposes show a higher likelihood of default, suggesting that the scale of the loan may impact repayment feasibility.
7. **Loan Volume Trends**: The number of loan applicants increased from **2007 to 2011**, with a notable spike in applications during **December**, potentially due to seasonal borrowing behaviors.
8. **Employment Length**: Applicants with **more than 10 years of employment** tend to default more often, particularly those who also take larger loans, indicating that longer employment does not guarantee financial stability.
9. **Loan Term Duration**: Applicants seeking loans with a **60-month term** are more likely to default compared to those taking loans for **36 months**, suggesting that longer-term loans may carry higher risk.
10. **Loan Purpose and Default Rates**: **Debt consolidation** is the most common loan purpose among those who default, indicating that this category is particularly vulnerable to non-repayment.
11. **Loan Amounts and Defaults**: Defaulting applicants often received loan amounts of **$15,000 or higher**, reflecting that larger loans are associated with higher default rates.
12. **Debt-to-Income Ratios**: Charged-off applicants generally reported medium to high **Debt-to-Income (DTI)** ratios, suggesting a strong correlation between financial strain and default risk.
13. **Annual Income**: Many applicants who defaulted reported an annual income between **$40,000 and $50,000**, indicating that lower income levels may contribute to repayment difficulties.

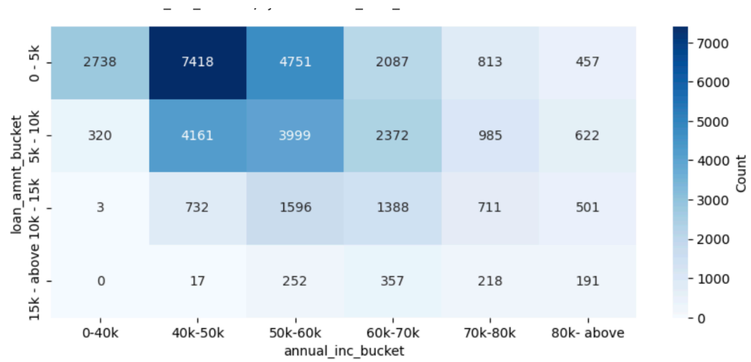
Multivariate Analysis :

Multivariate analysis examines the interactions between multiple variables to identify complex relationships that influence loan default risk. This approach provides a deeper understanding of how various factors collectively impact borrowers' likelihood of default

Heat map :



OBSERVATION : Debt consolidation is the category where the maximum number of loans are issued, and people have defaulted the most in the same category.



OBSERVATION : Low annual income applicant seems to be defaulter. (below 50k)

Summary :

1. **Credit Grade**: Applicants in grades B, C, and D are significantly more likely to default.
2. **Sub-Grade Risk**: Higher default rates are observed in sub-grades B3, B4, B5, and C1.
3. **Employment Length**: Applicants with 10+ years of employment are more prone to default.
4. **Geographic Factors**: Higher default rates are concentrated in states CA, FL, and NJ.
5. **Housing Situation**: Renters and mortgaged homeowners are at greater risk of defaulting.
6. **Income Level**: Lower annual income correlates with higher default likelihood. - Interest Rates: Default risk increases with rising interest rates

Multivariate Analysis :

Correlation Observation :

1. **Positive Correlation**

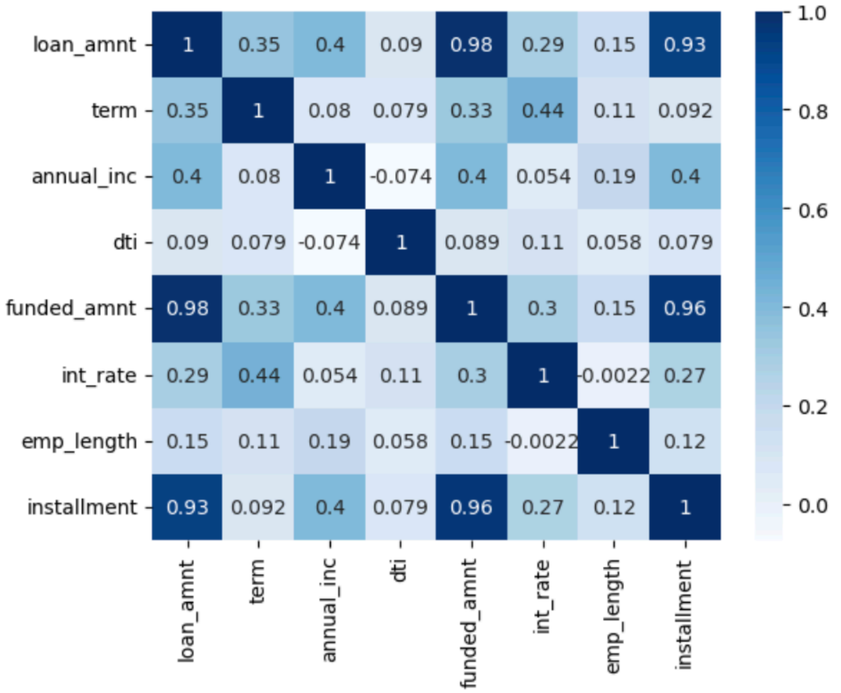
- **Loan Amount and Installment**: There is a strong positive correlation between loan amount and installment payments, indicating that as the loan amount increases, the installment payments also rise proportionately.
- **Annual Income and Loan Amount**: A strong positive correlation exists between annual income and loan amount, suggesting that higher-income applicants tend to take out larger loans.
- **Funded Amount and Loan Amount**: There is a strong correlation between the funded amount and the loan amount, indicating that loans are typically fully funded at or near the requested amounts.
- **Term and Interest Rate**: A strong relationship between the loan term and interest rate suggests that longer loan terms often come with higher interest rates, reflecting increased risk for lenders.

2. **Weak Correlation**

- **Debt-to-Income Ratio (DTI)**: The DTI shows no significant correlation with any other variable in the dataset. This implies that DTI may not be a strong predictor of loan-related behaviors or outcomes in this context.

3. **Negative Correlation**

- **DTI and Annual Income**: There is a negative correlation between DTI and annual income, indicating that as annual income increases, the DTI tends to decrease. This suggests that higher earners have a lower proportion of debt relative to their income.
- **Employment Length and Interest Rate**: A negative correlation exists between employment length and interest rate, suggesting that longer employment may be associated with lower interest rates, potentially reflecting the perceived stability and creditworthiness of long-term employees.



Recommendation :

Based on the findings from the univariate, bivariate, and multivariate analyses of loan default risk, following are comprehensive recommendations:

1. **Enhanced Risk Assessment Models**:
    - Implement advanced predictive algorithms that incorporates multiple variables, including credit grade, employment length, income levels, and geographic factors, to better identify high-risk applicants.
    - Develop more stringent criteria for applicants in lower credit grades (B, C, and D) and high-risk sub-grades (B3, B4, B5, C1) to minimise potential defaults.
3. **Targeted Financial Education Programs**:
    - Offer financial literacy programs focused on borrowers with lower annual incomes and those applying for debt consolidation loans. This could help them better understand their financial commitments and improve repayment strategies.
4. **Geographic Risk Strategies**:
    - Adjust lending policies based on geographic risk profiles, particularly in states with higher default rates (CA, FL, NJ). Tailor lending products and interest rates to handle regional economic conditions.
5. **Housing Stability Initiatives**:
    - Implement support programs for renters and mortgaged homeowners to enhance financial stability. Consider partnerships with housing organisations to offer resources that help borrowers manage their housing costs effectively.
6. **Interest Rate Transparency**:
    - Consider offering options for borrowers to lock in lower rates to encourage responsible borrowing.
7. **Custom Loan Products**:
    - Create loan products that meet the specific needs of different groups, like lower-income borrowers or those looking to consolidate debt. Offering flexible terms and payment options can make loans more affordable and lower the risk of default.
9. **Incentives for Responsible Borrowing**:
    - Create incentives for borrowers who demonstrate responsible financial behavior, such as making timely payments or improving their credit scores. This could include lower interest rates for future loans or reduced fees.