Lending Club Case Study

Prepared By:

Saurabh Tayde

Jyoti Kumari

Date: 25th Dec 2024

Contents

Problem Statement

- **Overview:**
  - The company needs to improve its loan approval process to identify applicants likely to default. This involves balancing two competing risks
  - Rejecting a loan application from a creditworthy individual, leading to lost business
  - Approving a loan application from an individual likely to default, leading to financial loss

- **Dataset and Objective:**
  - The dataset contains information on past loan applicants, including their loan characteristics and whether they defaulted ("charged-off").
  - The objective is to use EDA to identify patterns and "driver variables" that strongly indicate default risk.

- **Business Implications:**

  Identifying risky applicants allows the company to take mitigating actions such as:
  - Denying the loan application.
  - Reducing the approved loan amount.
  - Charging a higher interest rate to compensate for increased risk.

- **Data Scope:**

  The dataset only includes applicants who were approved for a loan. It does not include rejected applications because there is no subsequent loan performance data for these individuals. The included loan statuses are:
  - Fully Paid: Loan repaid successfully.
  - Current: Loan repayment in progress.
  - Charged-off: Loan defaulted.

- **Company Background:**

  The company is the largest online loan marketplace, offering various loan types (personal, business, medical). Its online platform provides borrowers with access to lower interest rates.

- **Key Goals:**
  - The primary goal is to reduce credit loss – the financial loss incurred when borrowers default – by identifying risky applicants before loan approval.
  - Understanding the factors that contribute to default and using this knowledge to improve the company's risk assessment strategies.
  - Perform EDA to address the challenge of loan default prediction for a consumer finance company.

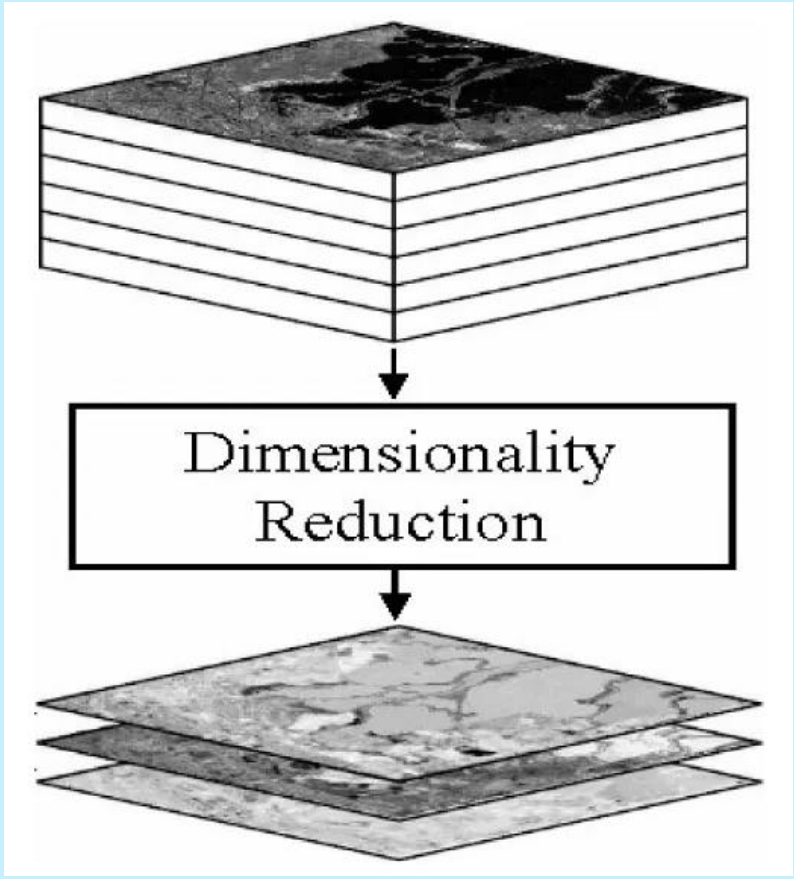
Dimensionality Reduction

- **Dimensionality Reduction** (Dropping Uninformative Features)
  - Goal: Streamline data, improve model efficiency, and mitigate overfitting.
  - Steps: Handling missing data, addressing high-cardinality features, and leveraging domain knowledge.

- **Excessive Missing Values** (Removing Features with Excessive Missing Values)
  - Criteria: Removed features with >60% missing values.
  - Rationale: High missing values lead to unreliable analysis and biased models.
  - Outcome: Ensures sufficient data for meaningful insights.

- **High Cardinality** (Removing Features with Redundant Values)
  - Criteria 1: No Variability - Features with only one unique value.
  - Criteria 2: High Repetition - Features with >98% identical values.
  - Rationale: Simplifies dataset and prevents potential overfitting.

- **Domain Expertise based Features Exclusion**
  - Removing **Unique Identifiers**
    - Rationale: Do not contribute to predicting loan default likelihood.
    - Outcome: Streamlines dataset without loss of predictive information.
  - Removing **Post-Approval** Features
    - Rationale: Irrelevant to pre-approval default prediction, prevents data leakage.
    - Outcome: Focuses analysis on pre-approval characteristics.
  - Removing Other **Irrelevant** Features
    - Rationale: Based on domain expertise, these features do not contribute to loan default prediction.



(Similar to features, the records that are not useful are also dropped. For example, records where 'loan_status' = 'Current' will not contribute to our analysis).

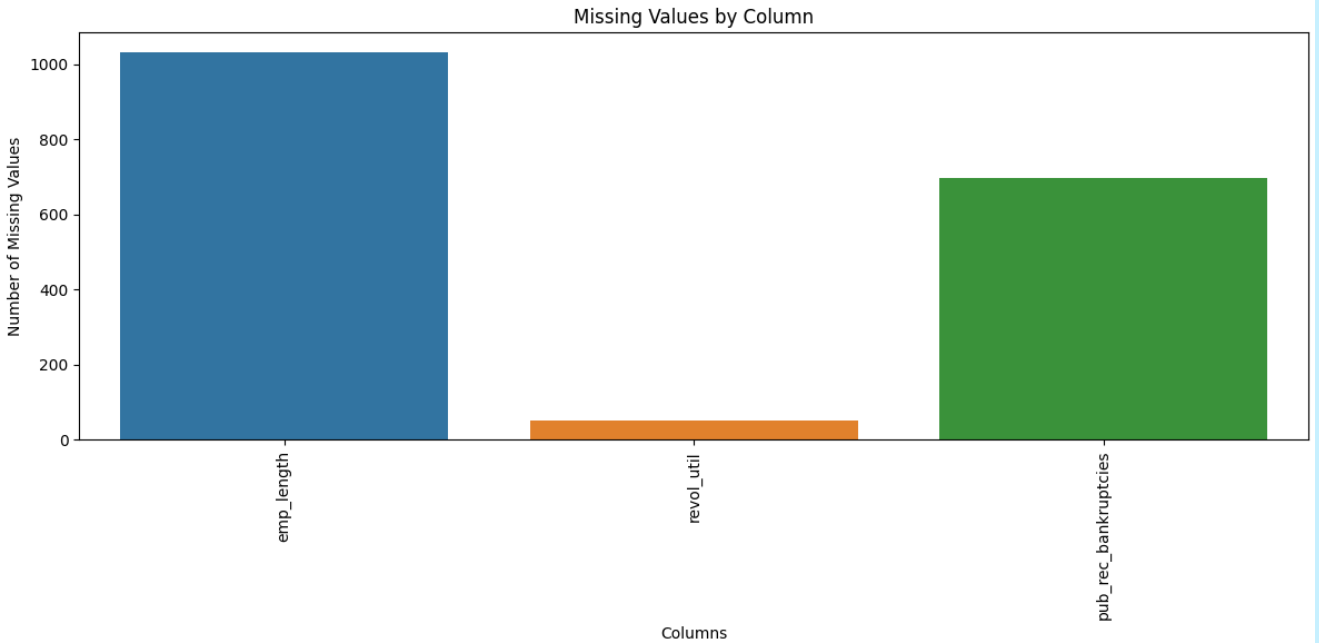
- **Dataset Size**

  Dataset Size after Dimensionality Reduction and removing irrelevant records = (38577, 22)

Data Cleaning and Manipulation

- **Converting Column Formats:**
  - Removed '%' sign from 'int_rate' and 'revol_util' and converted to float.
  - Converted 'issue_d' to 'yyyy-mm-dd' format.
  - Removed 'months' text from 'term' values.
  - Converted 'emp_length' to integer.

- **Missing Data Imputation:**
  - 'emp_length':
    - Number of unique values for emp_length (excluding NaN): 11
    - Even through it is numeric data type, Low cardinality (11 unique values) justifies **mode** imputation.
  - 'revol_util':
    - Number of unique values for revol_util (excluding NaN): 1088
    - Missing values imputed using the **median**.
  - 'pub_rec_bankruptcies':
    - Number of unique values for pub_rec_bankruptcies (excluding NaN): 3
    - Even through it is numeric data type, Low cardinality (3 unique values) justifies **mode** imputation.



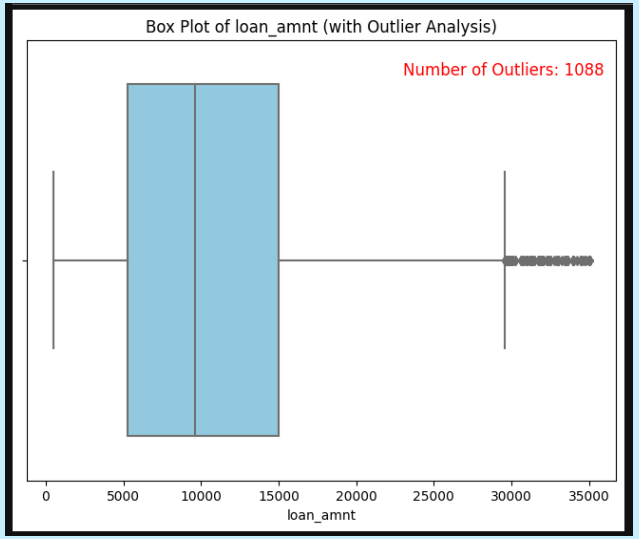
Outlier Detection and Mitigation

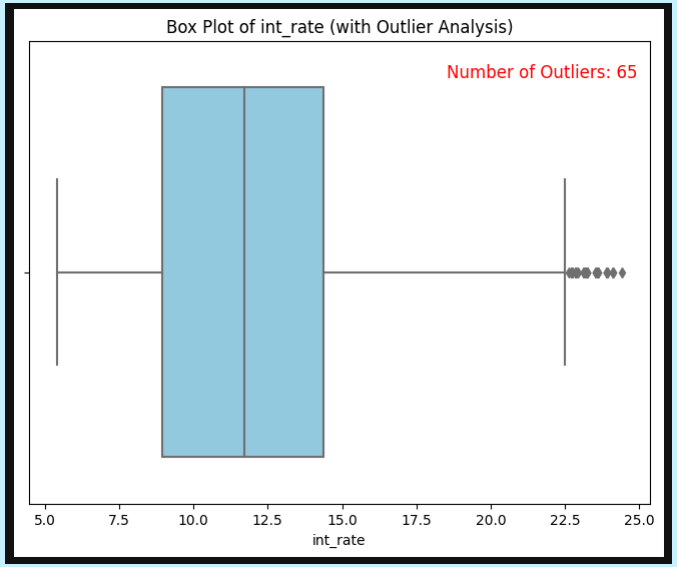
Justification for Outlier Removal in Loan Data Analysis

- Initial State:
  - 38,577 total loan records.
  - Outlier removal using IQR method resulted in 32,967 records (≈14% reduction).
- Why Removal is Preferred over Capping / Winsorizing:
  - Data Integrity:
    - Outliers can represent unusual cases, errors, or fraudulent activity.
    - Capping/Winsorizing introduces artificial data points, distorting the true distribution.
    - Removal ensures remaining data is more representative of typical loans.
  - Model Robustness:
    - Outliers can unduly influence model parameters and reduce predictive power.
    - Removal improves model robustness and reliability of findings.
  - Interpretability:
    - Outliers skew descriptive statistics, making interpretation difficult.
    - Removal normalizes data distribution, facilitating meaningful conclusions.
  - Sample Size Justification:
    - 32,967 records is still substantial for robust statistical analysis and understanding loan process.
    - Ample statistical power remains for detecting significant relationships.

Boxplots

(Only showing two plots here. Please refer code for all the plots and other details)

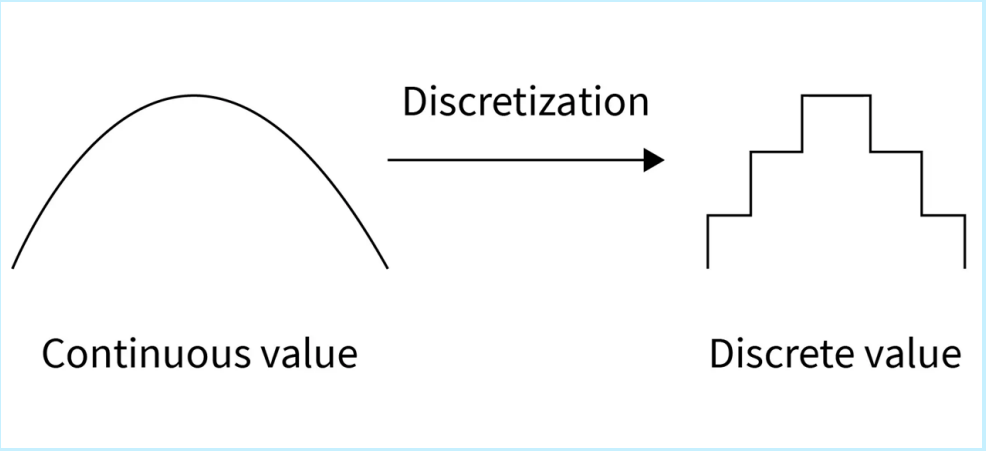




Derived Metrics / Features

- **Date-Based Features** (from `issue_d`)
  - `issue_year`: Allows analysis of yearly trends and macroeconomic influences.
  - `issue_month`: Enables examination of seasonal effects on loan defaults.

- **Binned Features:**
  - **Rationale:**
    - Improves univariate, segmented univariate, and bivariate analysis.
    - Facilitates clearer visualization and understanding of variable distributions and relationships, especially for loan default prediction.
    - Enhanced Univariate Analysis: Easier visualization of individual variable distributions (e.g., loan amounts).
    - More Powerful Segmented Univariate Analysis: Targeted analysis of loan status within bins of other variables (e.g., default rates by income bracket).
    - Improved Bivariate Analysis: Clearer examination of relationships between binned variables (e.g., loan amount vs. income).
  - **Binning Strategies (with justifications):**
    - **`loan_amnt`:** Ranges (0-4K, 4K-8K, etc.) based on distribution statistics (min, max, median).
    - **`funded_amnt`:** Same as `loan_amnt` for consistency and comparison.
    - **`int_rate`:** Ranges (0%-5%, 5%-9%, etc.) based on observed range. Reflects typical interest rate tiers and risk variations.
    - **`installment`:** Bins (0-150, 150-300, etc.) based on data range. Analyzes impact of monthly payments.
    - **`annual_inc`:** Income brackets (0-25K, 25K-50K, etc.) based on distribution. Represents meaningful income levels for default risk analysis.
    - **`dti`:** Ranges (< 5, 5-10, etc.) based on typical range. Reflects borrower debt burden.
    - **`open_acc`:** Groups (2-5, 5-8, etc.) based on number of accounts. Captures credit behavior patterns.
    - **`revol_util`:** Usage levels (0-17, 17-34, etc.) based on percentage utilization. Reflects revolving credit usage and risk.
    - **`total_acc`:** Groups (2-8, 8-16, etc.) based on total credit lines. Provides insight into credit history.



- **Final Dataset Details**
  - Dataset Size: 32,967 records and 33 variables
  - Composition:
    - 11 derived features (engineered)
    - 22 original features
  - This processed dataset has been used for all subsequent analysis and loan default prediction modeling

(Please refer code for list of all the variables and other details)

Univariate Analysis

Univariate Analysis (Numerical & Binned Features)

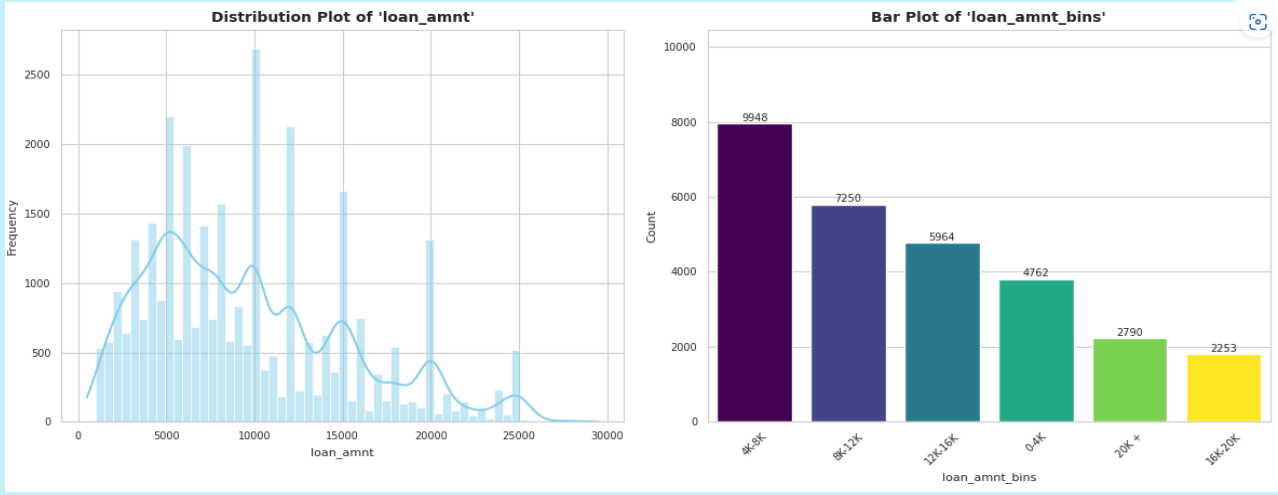
Interpretations / Findings of Univariate Analysis (Numerical & Binned Features):

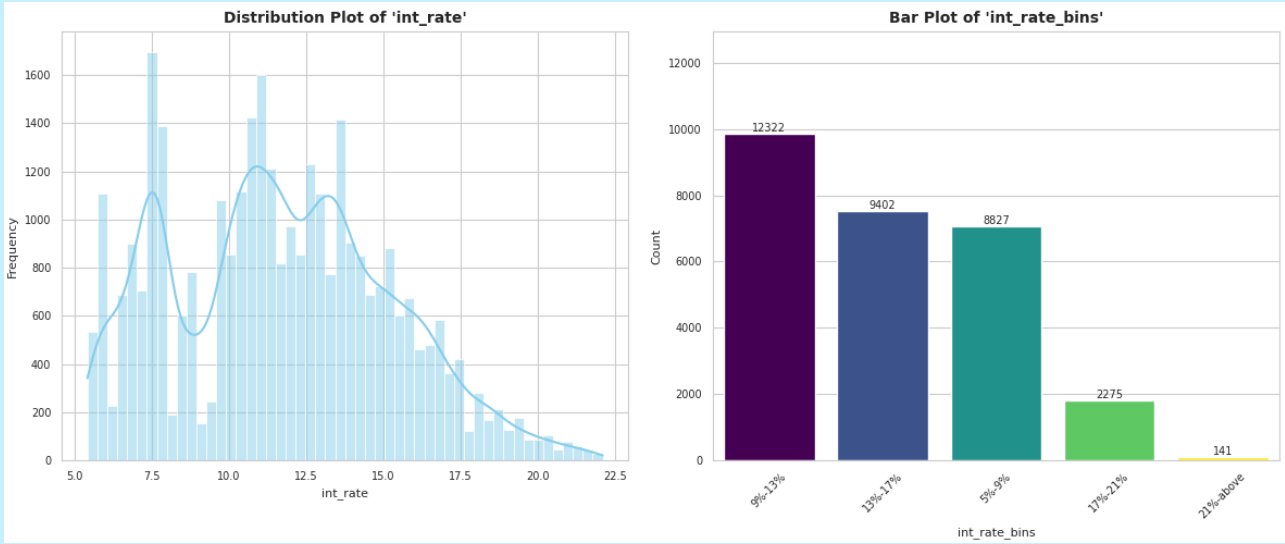
(Total Records Considered: **32,967** (Records with loan status as 'Fully Paid' as well as 'Charged Off'))

- **Loan Amounts and Funding:**
  - Concentration: Most loans and funded amounts are in the $8K-$12K and $4K-$8K ranges.
  - Implication: Suggests a typical mid-range loan amount with loan requests generally fulfilled.
- **Interest Rates and Installments:**
  - Interest Rate Concentration: Majority of loans have interest rates between 9% and 17%.
  - Installment Concentration: Most installments fall between $150 and $450.
  - Implication: Indicates a potential sweet spot for lenders balancing risk and return, and provides insight into typical borrower monthly payments.
- **Borrower Income and Debt:**
  - Income Distribution: Most borrowers have annual incomes between $25K and $75K.
  - DTI Distribution: Many borrowers have DTIs between 10 and 20.
  - Implication: Suggests a concentration of middle-income borrowers with moderate to moderately high debt burdens.
- **Open Credit Lines and Utilization:**
  - Open Credit Lines: Most borrowers have 5 to 11 open credit lines.
  - Revolving Utilization: Fairly even distribution across lower utilization bins, with slight peaks between 34% and 68%.
  - Implication: Indicates a moderate number of active accounts and diverse credit management practices.
- **Total Credit Accounts and History:**
  - Account Distribution: Most borrowers have 8 to 24 total accounts.
  - Implication: Suggests established credit histories among applicants, which could be correlated with default risk.

Univariate Plots

(Only showing two plots here. Please refer code for all the plots)





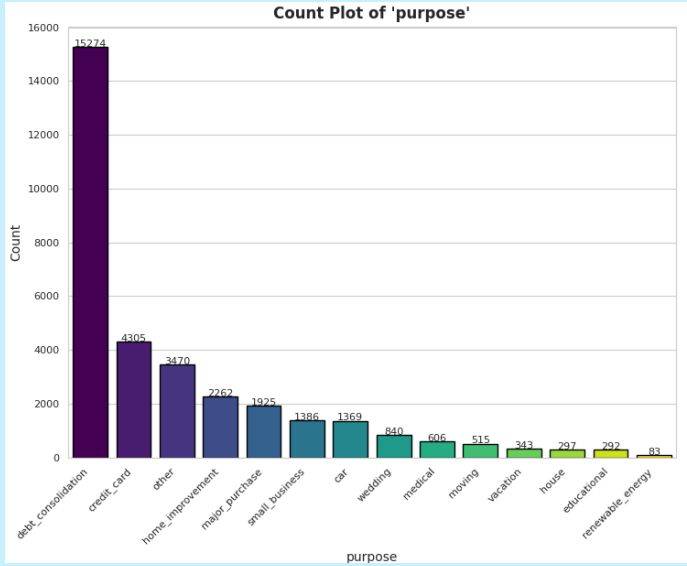
Univariate Analysis (Categorical Features)

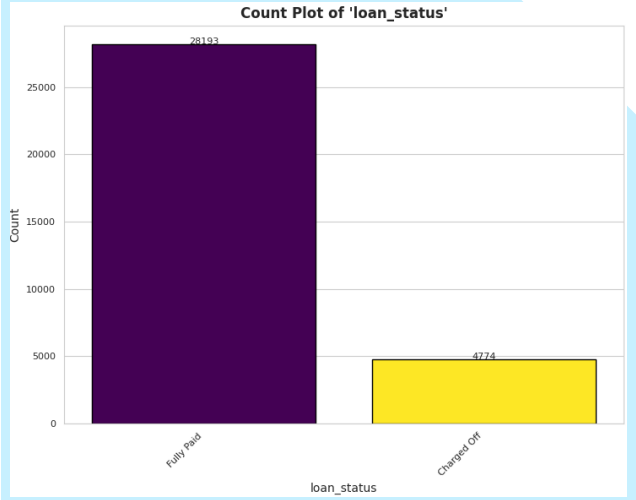
Interpretations / Findings of Univariate Analysis (Categorical Features) are as follows:

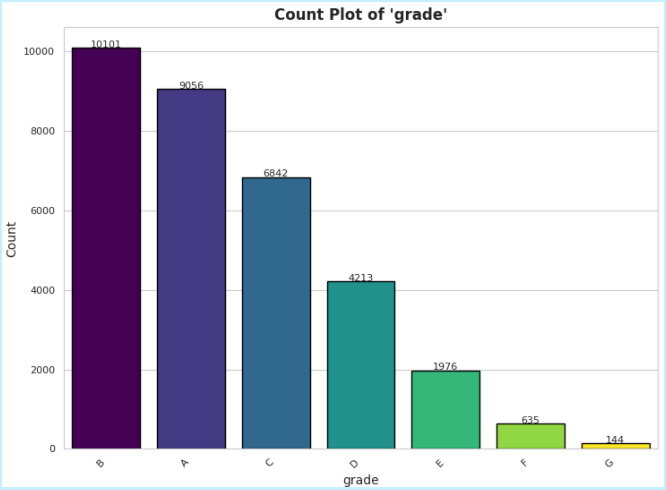
(Total Records Considered: **32,967** (Records with loan status as 'Fully Paid' as well as 'Charged Off'))

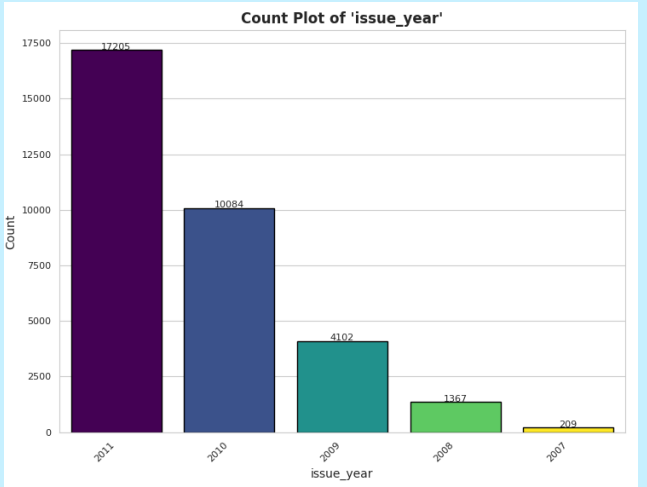
- **Loan Term and Grade Distribution:**
  - Loan Term: Majority of loans have a 36-month term.
  - Loan Grade: Grades A and B are the most frequent.
  - Implication: Indicates a preference for shorter-term loans and a concentration of higher-grade borrowers.
- **Loan Purpose:**
  - Primary Purpose: Debt consolidation is the most common loan purpose, followed by credit card refinancing.
  - Implication: Suggests many borrowers are using loans to manage existing debt, a key factor in risk assessment.
- **Sub-Grade and Employment Length:**
  - Sub-Grade: A4 and B3 are the most common sub-grades.
  - Employment Length: Most borrowers report 10+ years or <1 year of employment.
  - Implication: Suggests clustering within specific sub-grades and a potential relationship between employment stability and loan applications.
- **Loan Issuance Patterns:**
  - Issue Year: A large number of loans were issued in 2011.
  - Issue Month: Relatively consistent distribution across months, with slight peaks in December and November.
  - Implication: Points to potential influence of external factors and seasonality on loan origination.
- **Loan Status and Home Ownership:**
  - Home Ownership: Most borrowers are renters, followed by mortgage holders.
  - Loan Status: Majority of loans are fully paid.
  - Implication: Provides a baseline understanding of loan performance and borrower housing stability.

Plots (Only a few two plots here. Please refer code for all the plots)









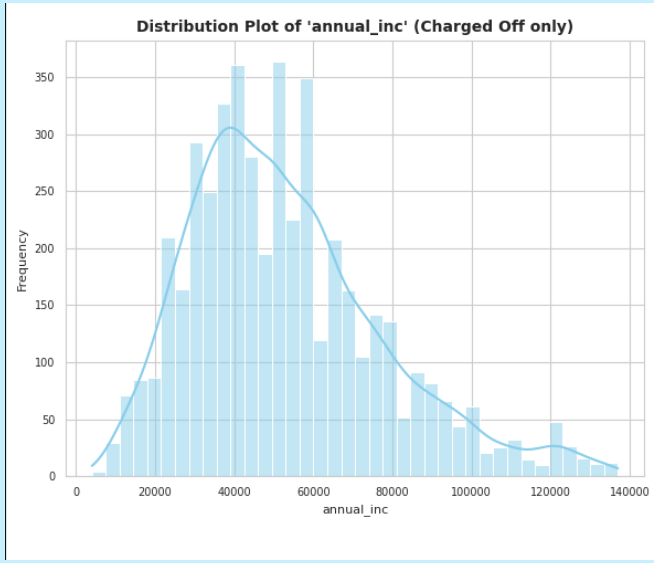
Univariate Analysis (Numerical & Binned Features): Defaults Only

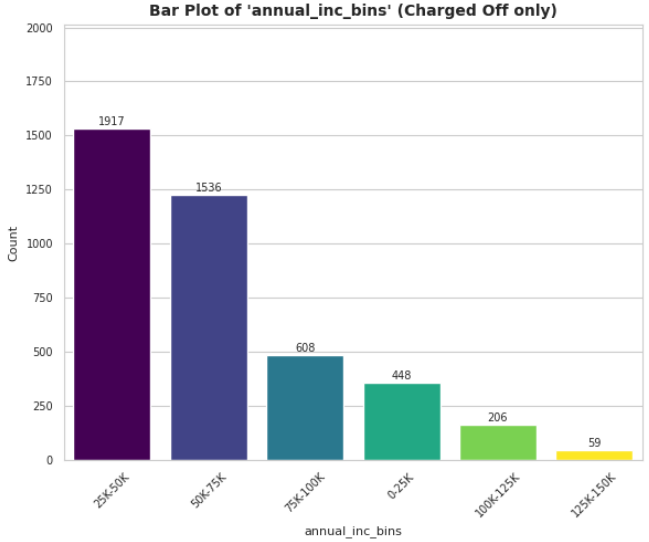
Interpretations / Findings of Univariate Analysis (Numerical & Binned Features):

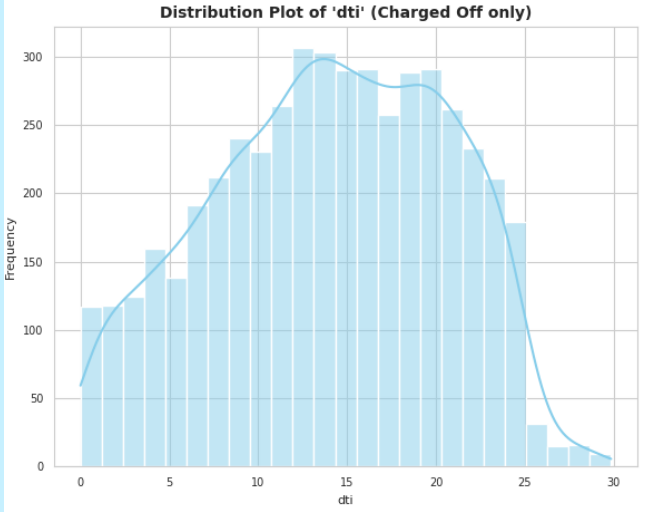
(Total Records Considered: **4774** (Records with loan status as 'Charged Off' only))

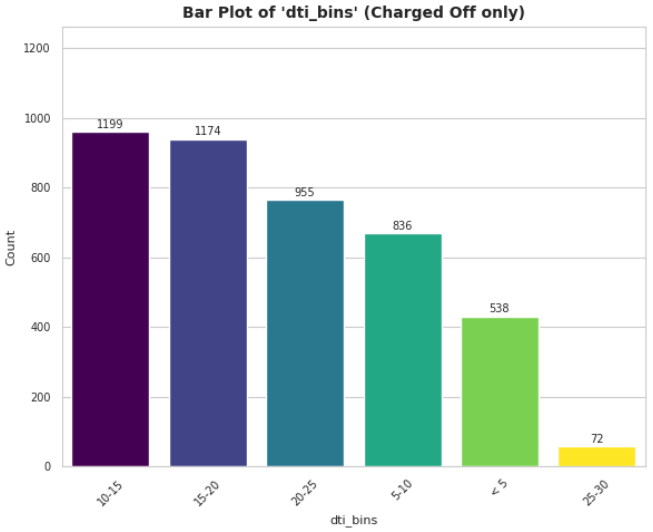
- **Loan and Funded Amounts (Defaults):**
  - Concentration: Most defaults occur with loan/funded amounts between $4K-$8K and $8K-$12K.
  - Implication: Smaller loan sizes may correlate with higher default risk.
- **Interest Rates and Installments (Defaults):**
  - Interest Rates: Most defaults have interest rates between 13%-17% and 9%-13%.
  - Installments: Most defaults have installments between $150-$300 and $300-$450.
  - Implication: Mid-range interest rates and typical installment amounts are prevalent among defaults, suggesting affordability isn't the sole driver.
- **Annual Income and DTI (Defaults):**
  - Annual Income: Most defaulters have incomes between $25K-$50K and $50K-$75K.
  - DTI: Most defaults occur with DTIs between 10-15 and 15-20.
  - Implication: Mid-range incomes and moderate to moderately high DTIs are common among defaulters.
- **Open Credit Lines and Utilization (Defaults):**
  - Open Credit Lines: Most defaults have 5-8 and 8-11 open lines.
  - Revolving Utilization: Defaults are spread across utilization bins, with higher frequencies between 51%-85%.
  - Implication: Moderate open credit lines and varying utilization rates are observed among defaulters, indicating other factors at play.
- **Total Accounts (Defaults):**
  - Distribution: Most defaults have 8-16 and 16-24 total accounts.
  - Implication: A moderate to high number of total accounts is common among defaulters, suggesting further analysis of account types is warranted.

Univariate Plots (Only showing two plots here. Please refer code for all the plots)









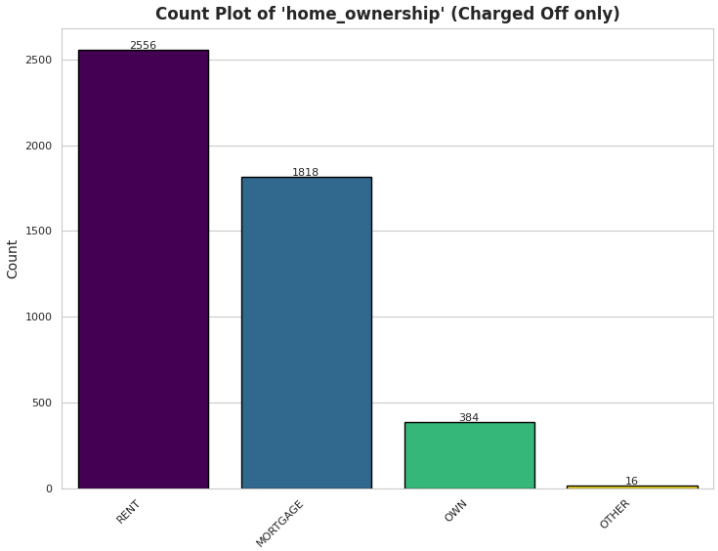
Univariate Analysis (Categorical Features): Defaults Only

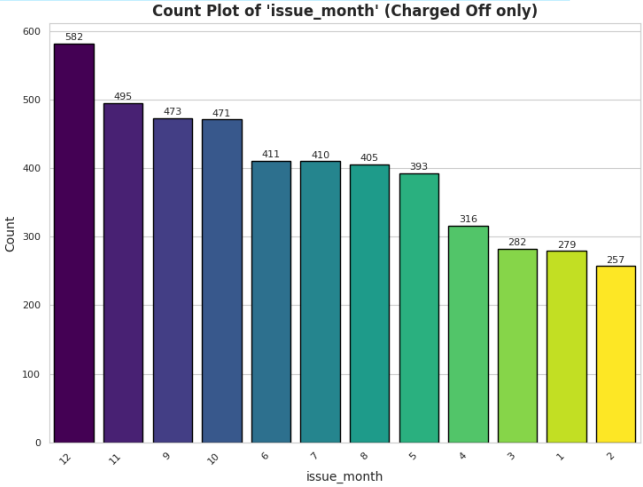
Interpretations / Findings of Univariate Analysis (Categorical Features) are as follows:

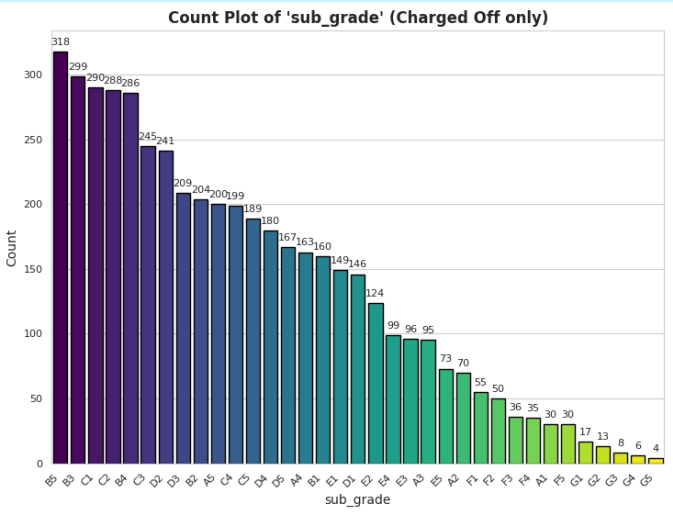
(Total Records Considered: **4774** (Records with loan status as 'Charged Off' only))

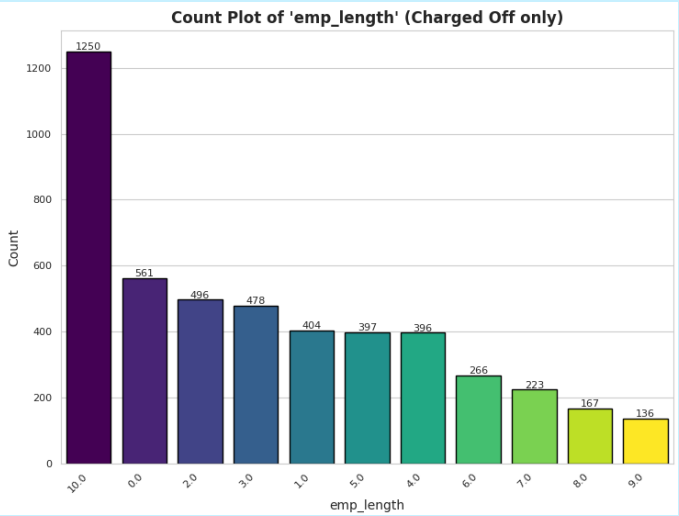
- **Loan Term and Grade (Defaults):**
  - Term: Significant defaults for both 36-month and 60-month terms.
  - Grade: Grades B and C have the highest default counts.
  - Implication: Indicates higher risk associated with these loan grades, regardless of term.
- **Sub-Grade and Employment (Defaults):**
  - Sub-grade: B5 and C1 are the most frequent sub-grades among defaults.
  - Employment Length: Surprisingly, borrowers with 10+ years and <1 year of employment show high default rates.
  - Implication: Requires further investigation into factors beyond employment length.
- **Loan Issuance (Defaults):**
  - Year: Most defaults originated in 2011.
  - Month: December and November have the highest default counts.
  - Implication: Suggests potential influence of economic conditions or seasonal factors.
- **Home Ownership (Defaults):**
  - Status: Renters and mortgage holders have the highest default counts.
  - Implication: Aligns with the general borrower population distribution.
- **Loan Purpose (Defaults):**
  - Primary Purpose: Debt consolidation is the most common purpose for defaulted loans.
  - Secondary Purposes: 'Other,' 'credit_card,' and 'small_business' follow.
  - Implication: Suggests debt consolidation may be used as a last resort, and other purposes also contribute significantly to defaults.

Plots (Only a few two plots here. Please refer code for all the plots)









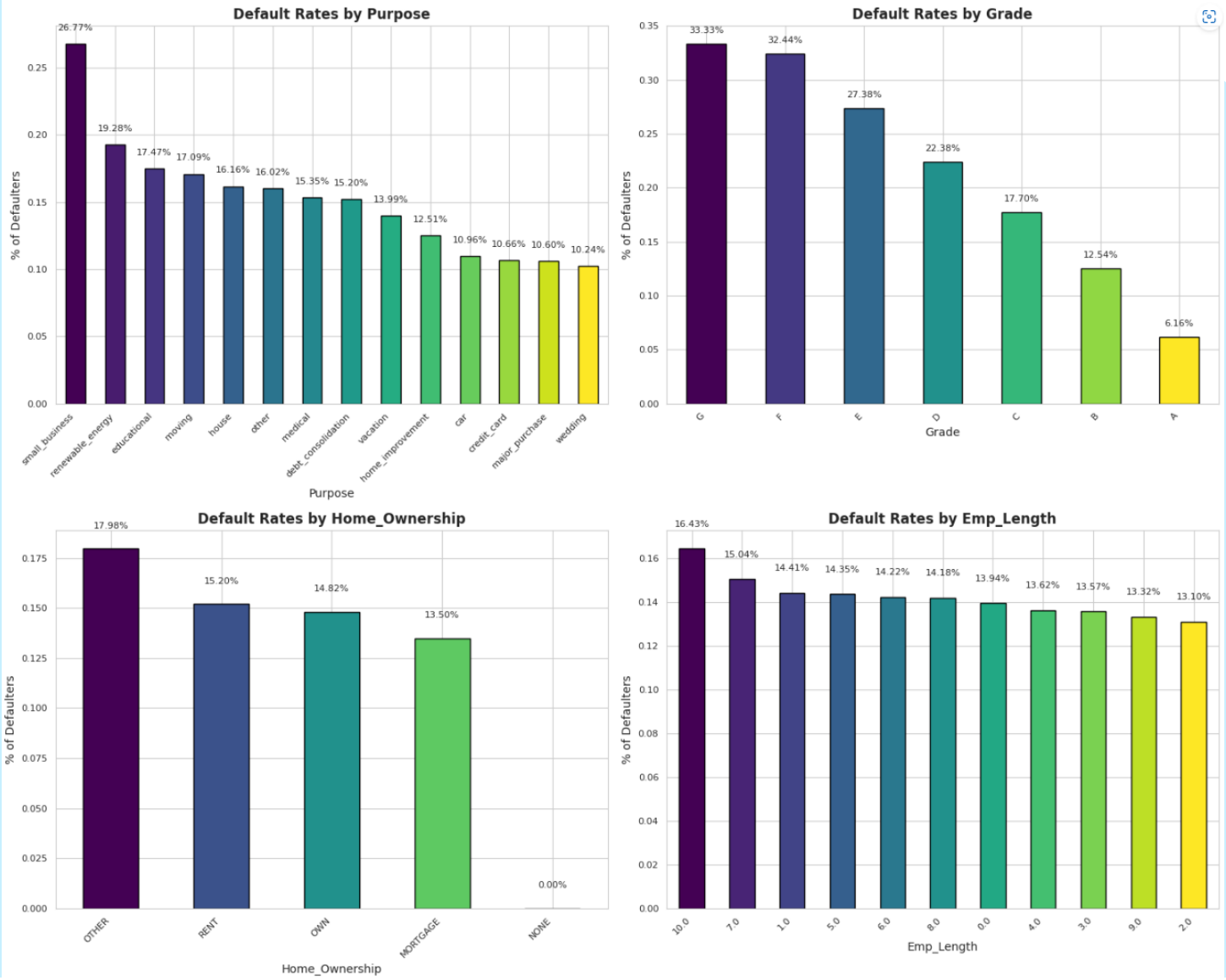
Segmented Univariate Analysis

Segmented Univariate Analysis (Loan Status Segmentation)

Interpretations / Findings of Segmented Univariate Analysis (Loan Status Segmentation):

- **Default Rates by Loan Purpose:**
  - Highest Risk: Small business and renewable energy loans exhibit the highest default rates.
  - Implication: These loan purposes inherently carry greater risk.
- **Default Rates by Loan Grade:**
  - Grade-Risk Relationship: Default rates increase as loan grade decreases (G and F being highest).
  - Implication: Confirms the lender's risk assessment reflected in the grading system.
- **Default Rates by Home Ownership:**
  - Housing Stability: "Other" and renter categories have higher default rates.
  - Implication: Non-traditional or less stable housing situations may increase default risk.
- **Default Rates by Employment Length:**
  - Unexpected Trend: Relatively consistent default rates across employment lengths, with a slight peak at 10+ years.
  - Implication: Employment length alone is not a strong predictor of default.
- **Default Rates by Loan Amount:**
  - Loan Size and Risk: Larger loan amounts (over $20K) show higher default rates.
  - Implication: Loan size is a contributing factor to default risk.

Plots

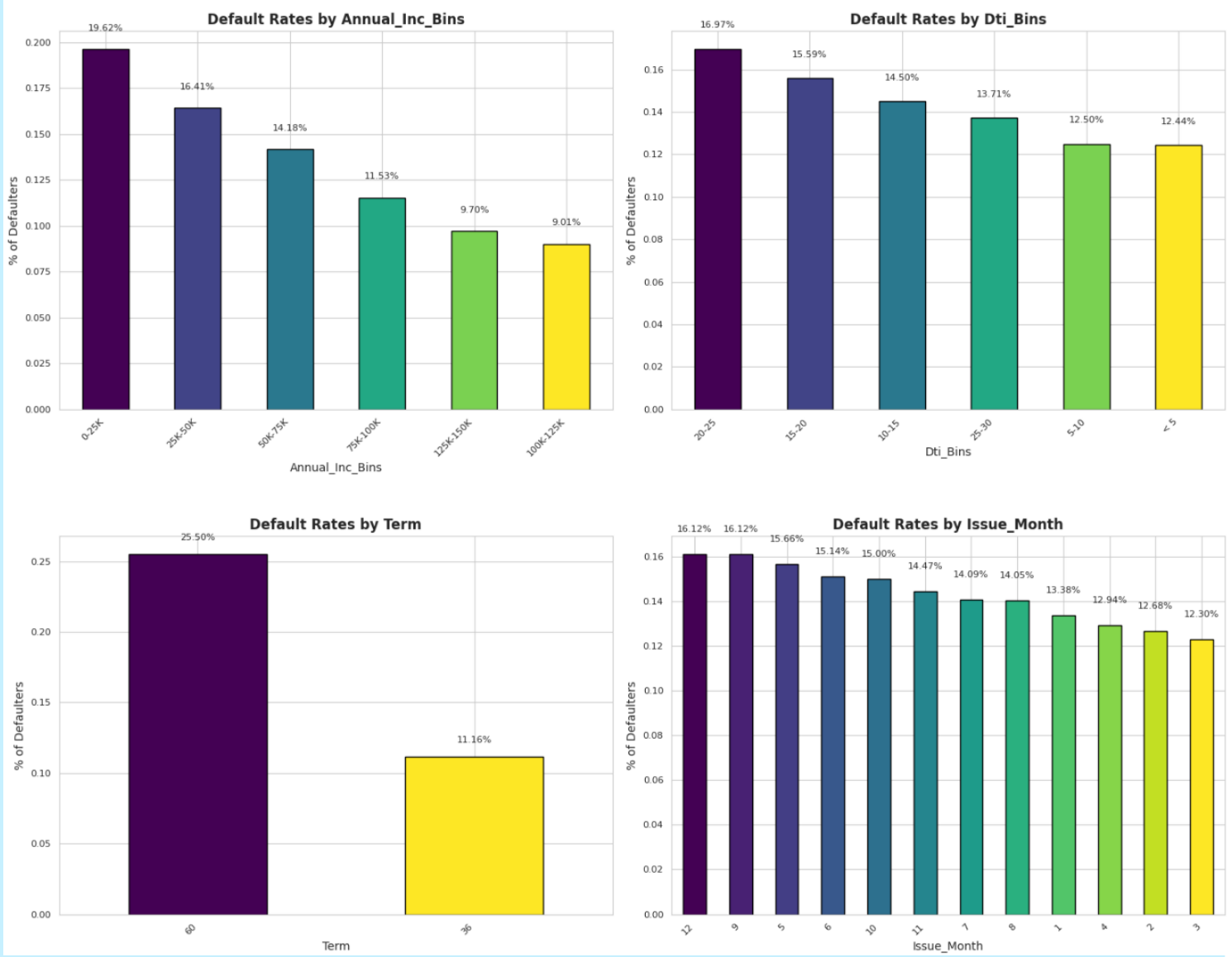


Segmented Univariate Analysis (Loan Status Segmentation)

Interpretations / Findings of Segmented Univariate Analysis (Loan Status Segmentation):

- **Default Rates by Interest Rate:**
  - Interest Rate and Risk: Strong positive correlation between interest rate and default rate.
  - Implication: Higher interest rates indicate higher perceived risk and increased borrower burden.
- **Default Rates by Annual Income:**
  - Income and Stability: Default rates generally decrease with increasing annual income.
  - Implication: Higher income provides greater financial stability and reduces default risk.
- **Default Rates by DTI:**
  - DTI and Risk: Higher DTIs (20-25 and 15-20) correlate with higher default rates.
  - Implication: High DTI increases vulnerability to default, regardless of income level.
- **Default Rates by Loan Term:**
  - Term Length and Risk: 60-month loans have significantly higher default rates than 36-month loans.
  - Implication: Longer loan terms increase the likelihood of default.
- **Default Rates by Issue Month:**
  - Seasonality: Minor variations in default rates across months, with slight increases in Nov, Dec, and Sept.
  - Implication: Month of issuance has a relatively small impact on default compared to other factors.

Plots

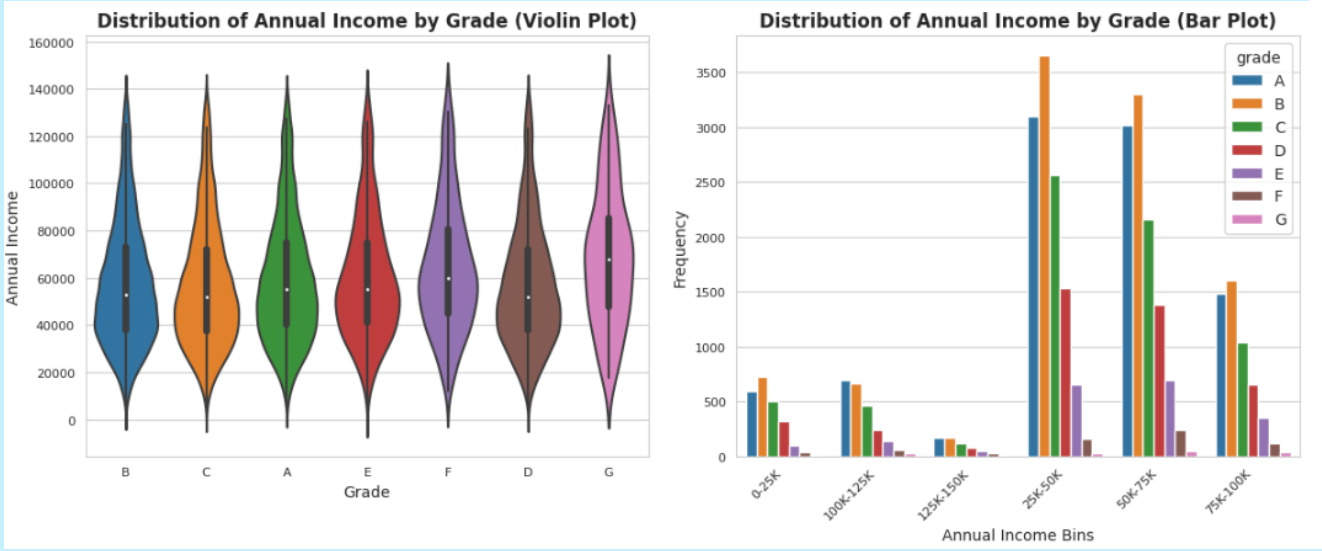


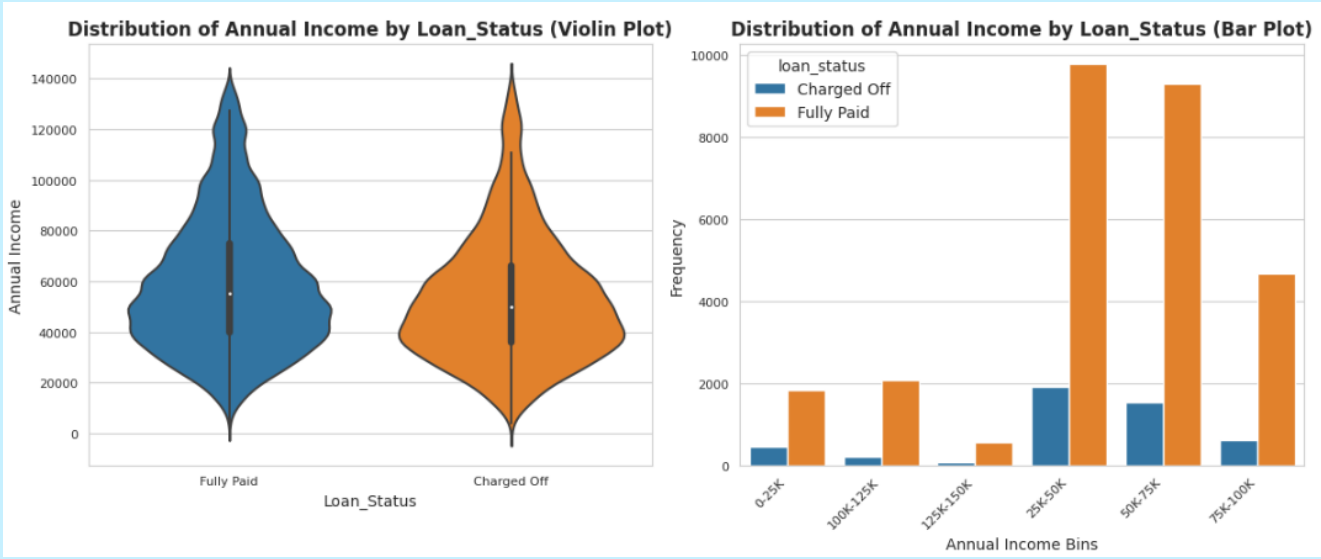
Segmented Univariate Analysis (Annual Income Segmentation)

Interpretations / Findings of Segmented Univariate Analysis (Annual Income Segmentation):

- **Income Distribution by Loan Grade:**
  - Positive Correlation: Higher loan grades (A, B, C) tend to have higher median annual incomes.
  - Implication: Higher earners are more likely to qualify for better loan terms.
- **Income Distribution by Homeownership:**
  - Renters vs. Owners: Renters tend to have lower incomes compared to mortgage holders.
  - "NONE" Category: Wide income distribution within the "NONE" category suggests diverse borrower profiles.
  - Implication: Homeownership status reflects differences in financial stability.
- **Income Distribution by Loan Status:**
  - Fully Paid vs. Charged Off: Fully paid loans show a slight skew towards higher incomes compared to charged-off loans.
  - Implication: Higher income correlates with successful loan repayment.

Plots (Please refer code for all the plots)



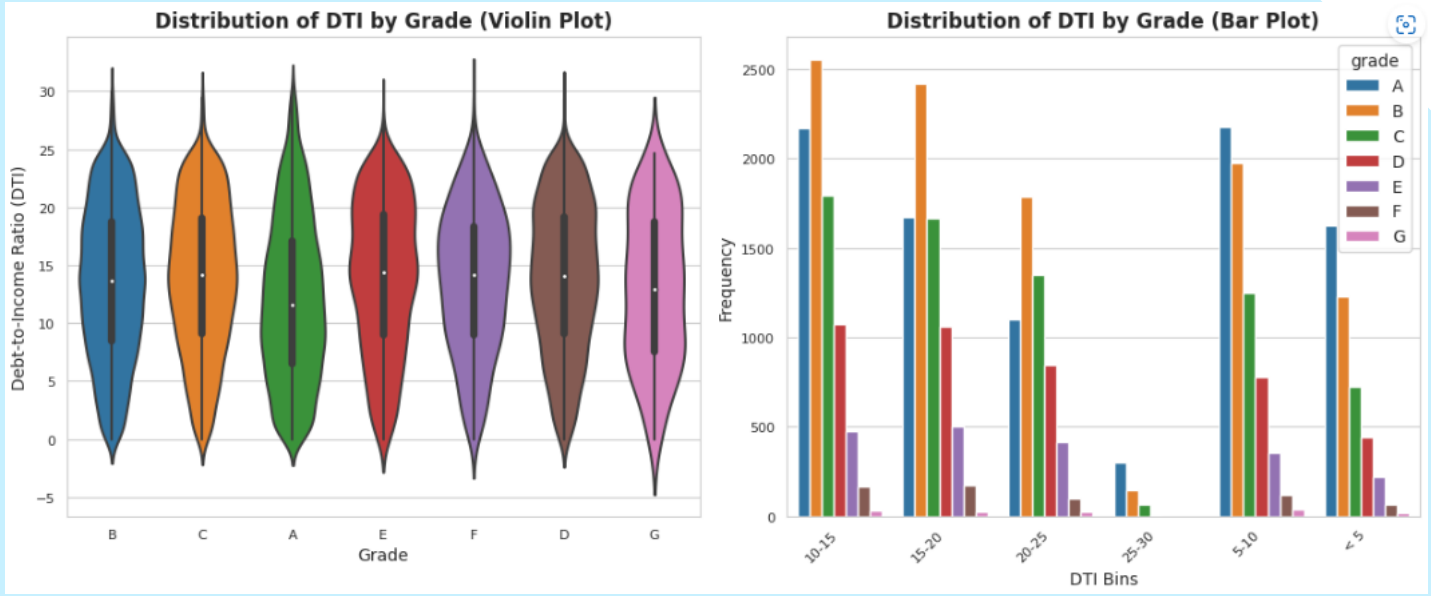


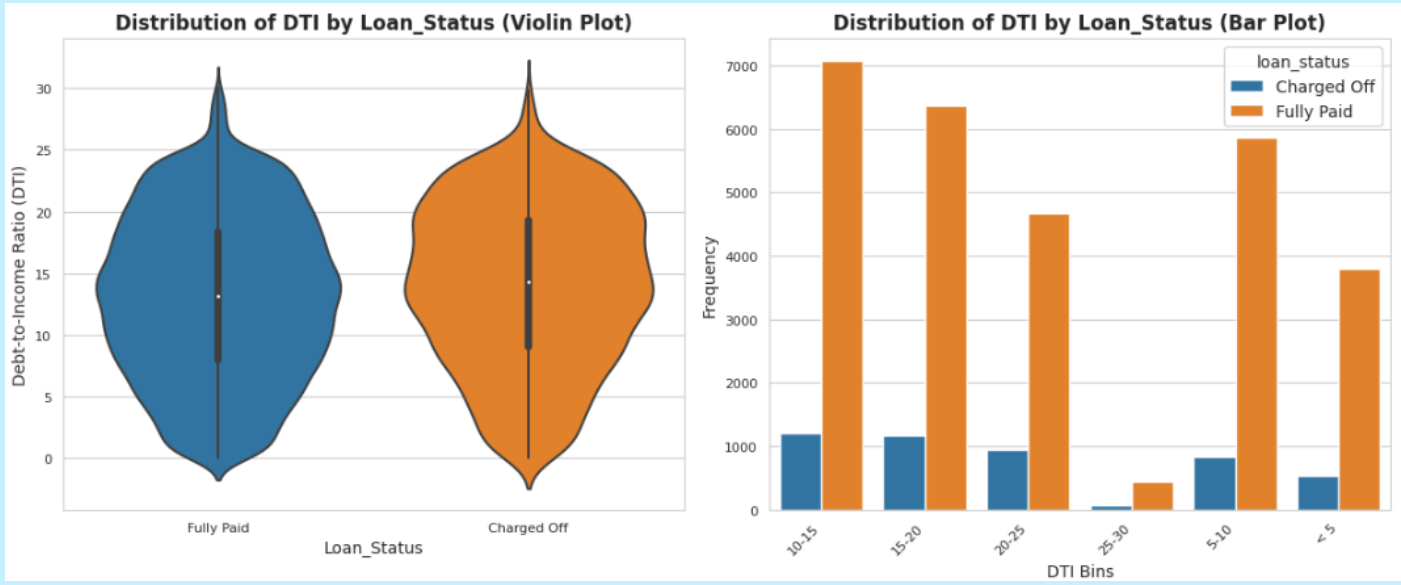
Segmented Univariate Analysis (DTI Segmentation)

Interpretations / Findings of Segmented Univariate Analysis (DTI Segmentation):

- **DTI Distribution by Loan Grade:**
  - General Consistency: DTI remains relatively consistent across loan grades, with similar median values.
  - Nuance: Higher loan grades show a slightly higher frequency of lower DTI values.
  - Implication: While not a strong differentiator, lower DTI might be slightly favored for higher grades.
- **DTI Distribution by Loan Status:**
  - Similar Medians: Both 'Fully Paid' and 'Charged Off' loans exhibit similar median DTI values, around 13-14, according to the violin plot.
  - Distribution Differences: The bar plot reveals that 'Charged Off' loans have a slightly higher representation in the 15-20 and 20-25 DTI bins. 'Fully Paid' loans are more frequent in the 10-15 bin.
  - Implication: While the overall DTI distributions are similar, a slightly higher DTI appears to be somewhat more prevalent among 'Charged Off' loans. This suggests DTI might be a factor, but not a dominant one, in predicting loan status.
- **DTI Distribution by Employment Length**:
  - Overall Consistency: DTI remains largely consistent across different employment lengths.
  - Variation: Shorter employment lengths show a higher concentration in the 15-20 DTI bin.
  - Implication: Individuals with shorter employment histories might have a slightly higher tendency towards higher DTIs.

Plots (Please refer code for all the plots)



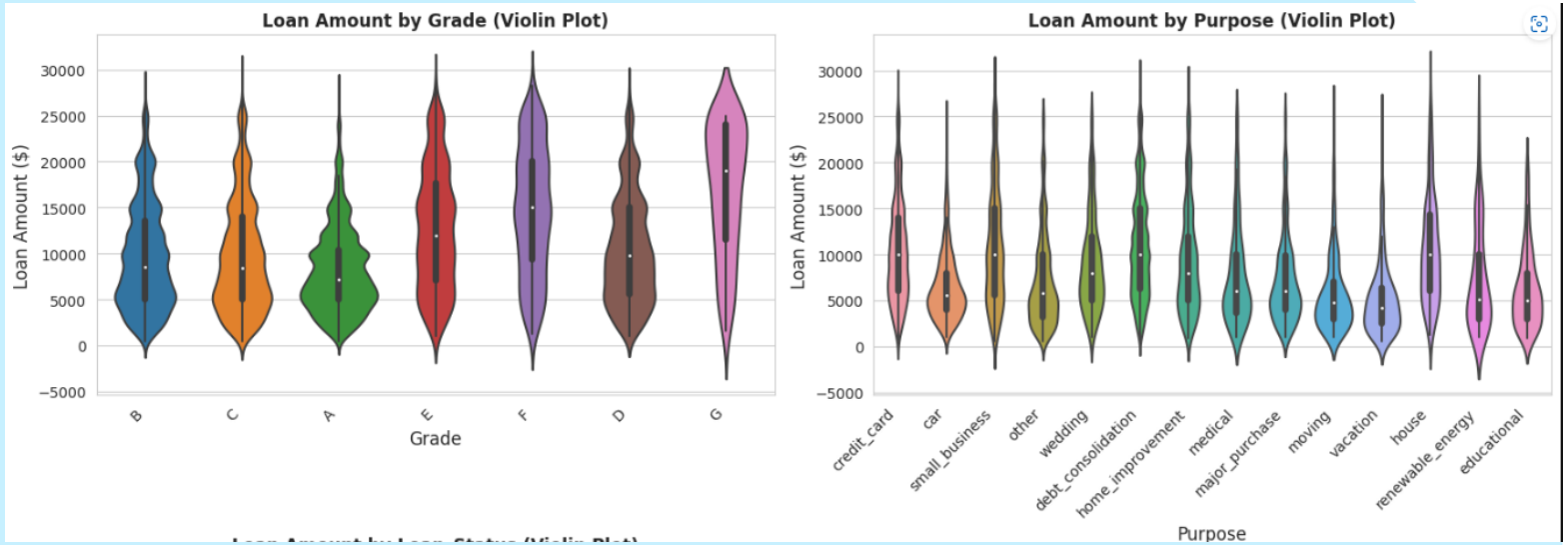


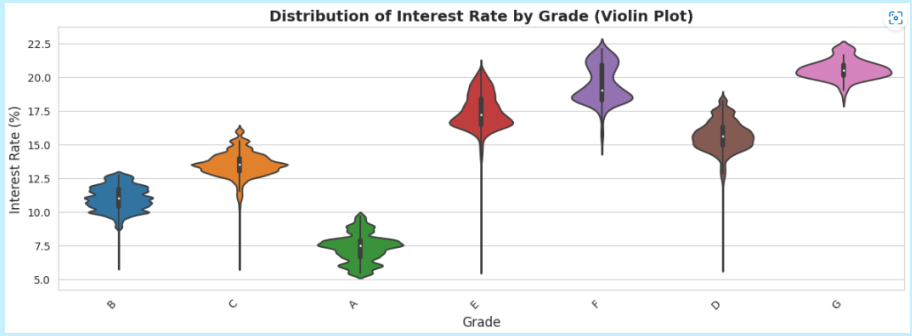
Segmented Univariate Analysis (Interest Rate, Loan Amount, Emp Length, Home Ownership)

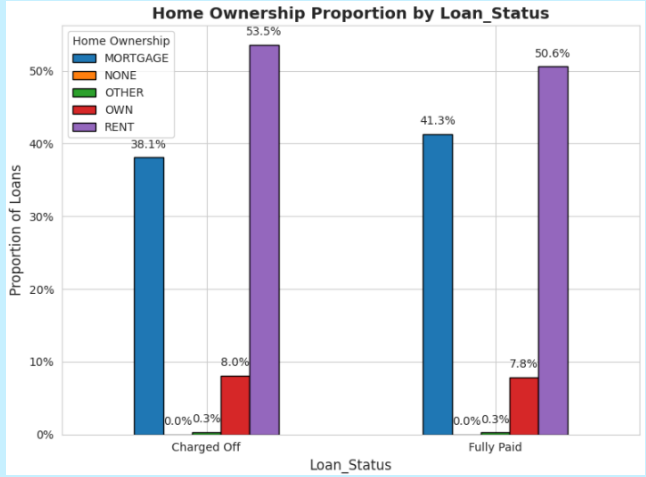
Interpretations / Findings of Segmented Univariate Analysis (Interest Rate, Loan Amount, Emp Length, Home Ownership):

- **Interest Rate Segmentation:**
  - Grade-Interest Relationship: Interest rates increase with decreasing loan grade (higher risk).
  - Purpose-Interest Relationship: Small business loans tend to have higher interest rates.
  - Term-Interest Relationship: 60-month loans generally have higher interest rates than 36-month loans.
- **Loan Amount Segmentation:**
  - Grade-Amount Relationship: Loan amounts tend to increase with lower loan grades.
  - Purpose-Amount Relationship: Debt consolidation and small business loans typically have larger amounts.
  - Status-Amount Relationship: Fully paid and charged-off loans have similar distributions, with charged-off loans showing slightly wider distribution at higher amounts.
- **Employment Length Segmentation:**
  - Grade-Employment Relationship: Higher grades tend to have more borrowers with longer employment histories.
  - Status-Employment Relationship: Charged-off loans have a higher representation of borrowers with shorter employment lengths.
- **Home Ownership Segmentation:**
  - Status-Home Ownership Relationship: Homeownership proportions are similar for both charged-off and fully paid loans.

Plots (Please refer code for all the plots)







Bivariate Analysis

Bivariate Analysis (Core Financial Relationships)

➢ **Interest Rate vs. Loan Amount:**

- Positive Correlation: Larger loans tend to have higher interest rates, reflecting increased lender risk.
- Default Concentration: Charged-off loans cluster in high interest rate/high loan amount regions, highlighting these as key risk factors.

➢ **DTI vs. Annual Income:**

- Complex Relationship: While higher income generally correlates with lower DTI, exceptions exist.
- DTI as Risk Factor: Charged-off loans are present even at higher income levels with moderate to high DTIs, indicating DTI's importance in risk assessment.

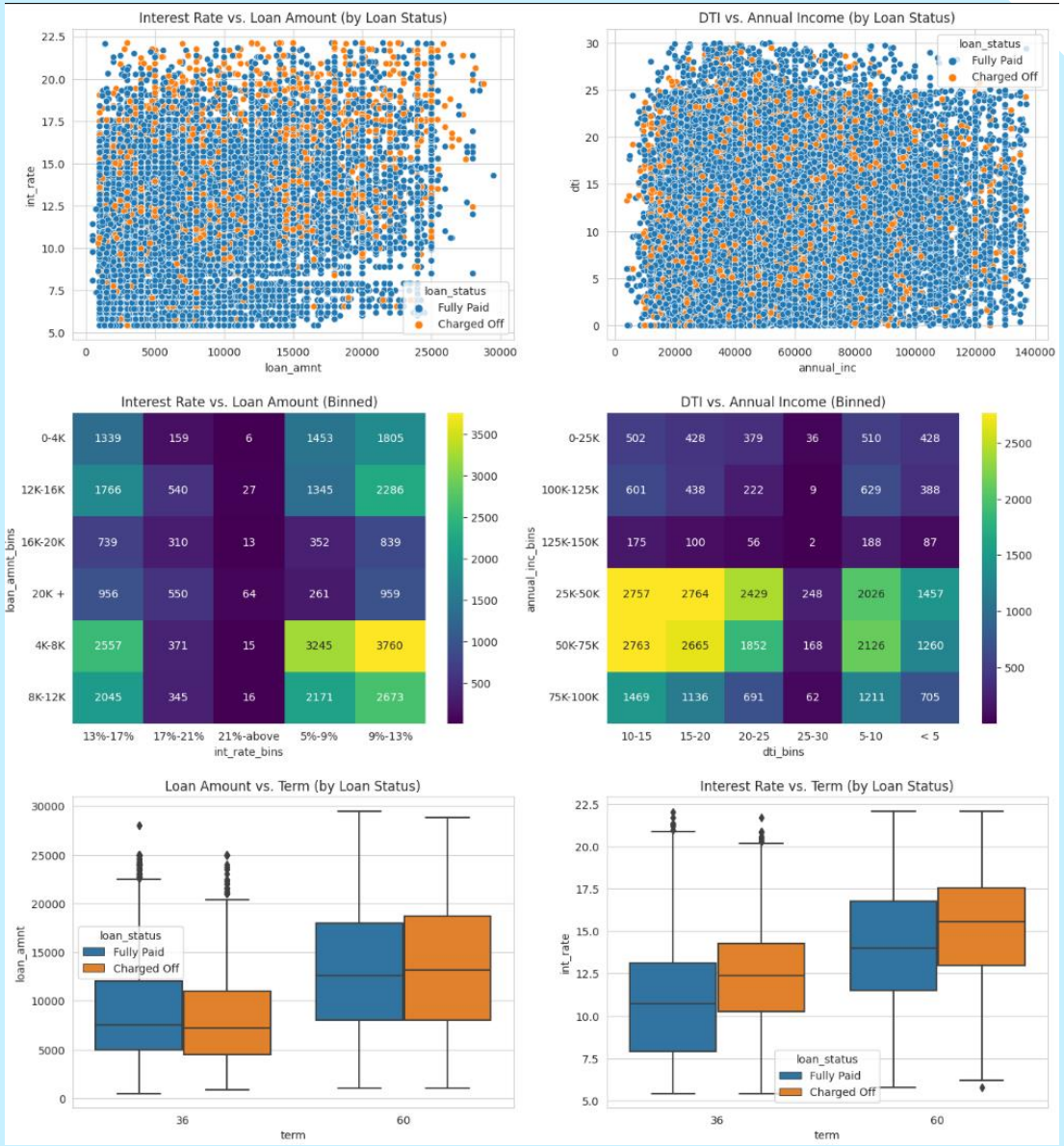
➢ **Loan Amount vs. Term:**

- Term and Loan Size: 60-month loans generally have higher loan amounts.
- Risk of Longer Terms: Charged-off loans within the 60-month term show a wider range of loan amounts, suggesting increased risk with larger, longer-term loans.

➢ **Interest Rate vs. Term:**

- Term and Interest Rate: 60-month loans consistently have higher interest rates than 36-month loans.
- Consistent Influence: The impact of term on interest rate is similar for both fully paid and charged-off loans, although higher rates on longer terms could contribute to financial stress and default.

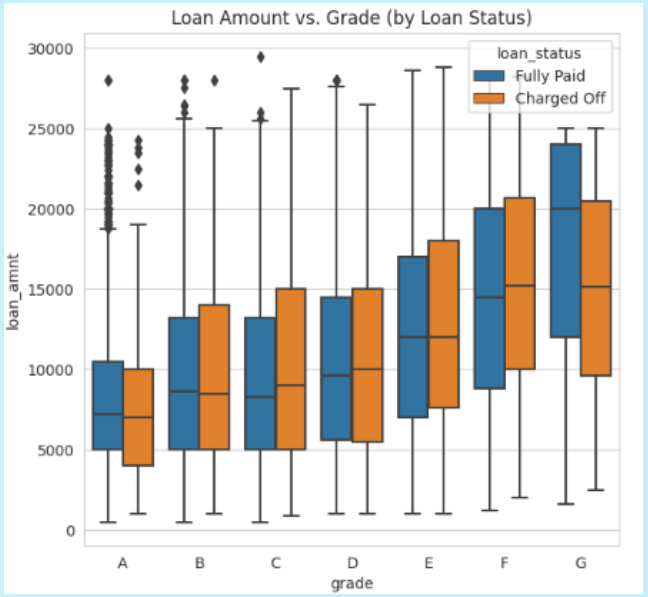
Plots

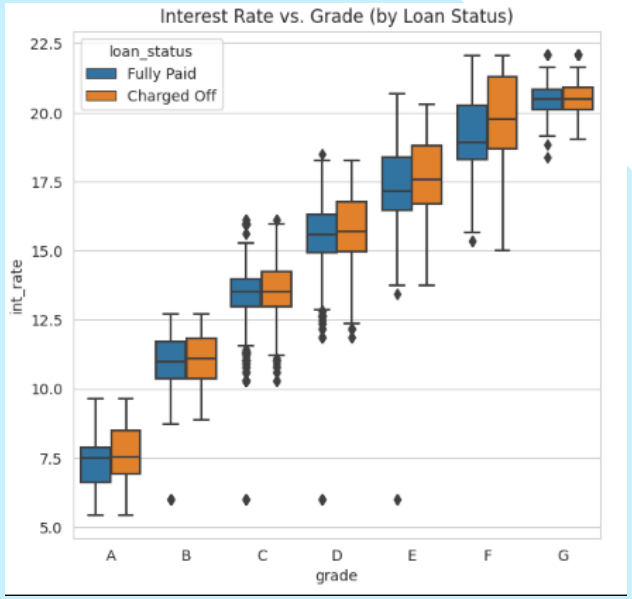


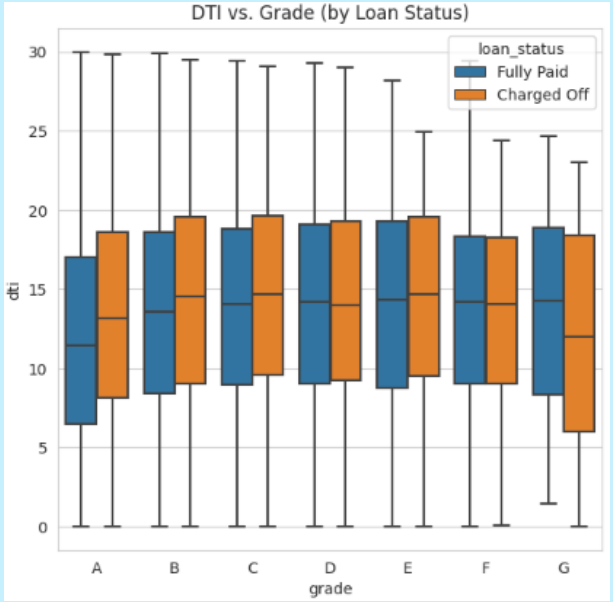
Bivariate Analysis (Loan Characteristics and Grade)

- **Loan Amount vs. Grade:**
  - Amount-Grade Relationship: Loan amounts generally increase with decreasing grade (higher risk).
  - Risk in Lower Grades: Charged-off loans within each grade often have wider and higher loan amount distributions than fully paid loans, particularly in lower grades.
- **Interest Rate vs. Grade:**
  - Rate-Grade Relationship: Interest rates increase with decreasing grade, as expected.
  - Grade as Primary Driver: Distributions of interest rates are similar for both charged-off and fully paid loans within each grade, indicating grade is the primary determinant of interest rate.
- **DTI vs. Grade:**
  - DTI-Grade Relationship: DTI shows a slight upward trend with decreasing grade.
  - DTI as a Factor: Charged-off loans tend to have slightly higher median DTIs within each grade compared to fully paid loans, suggesting DTI is a contributing risk factor.

Plots



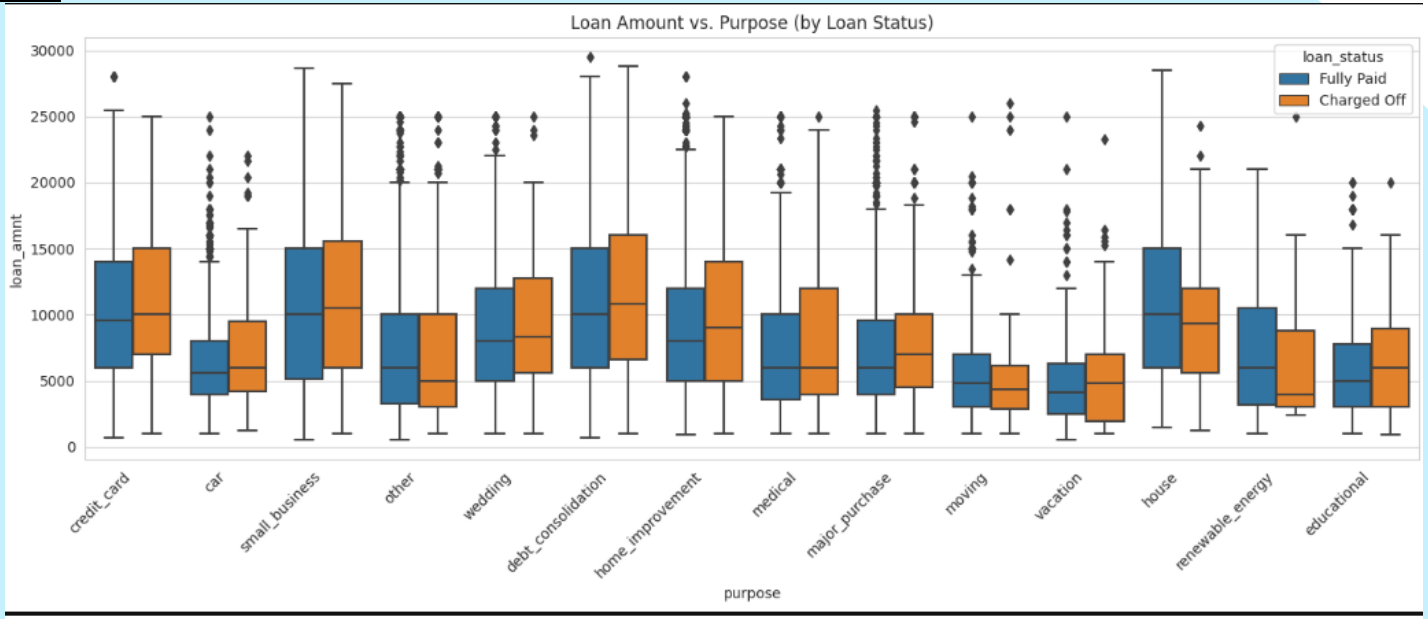


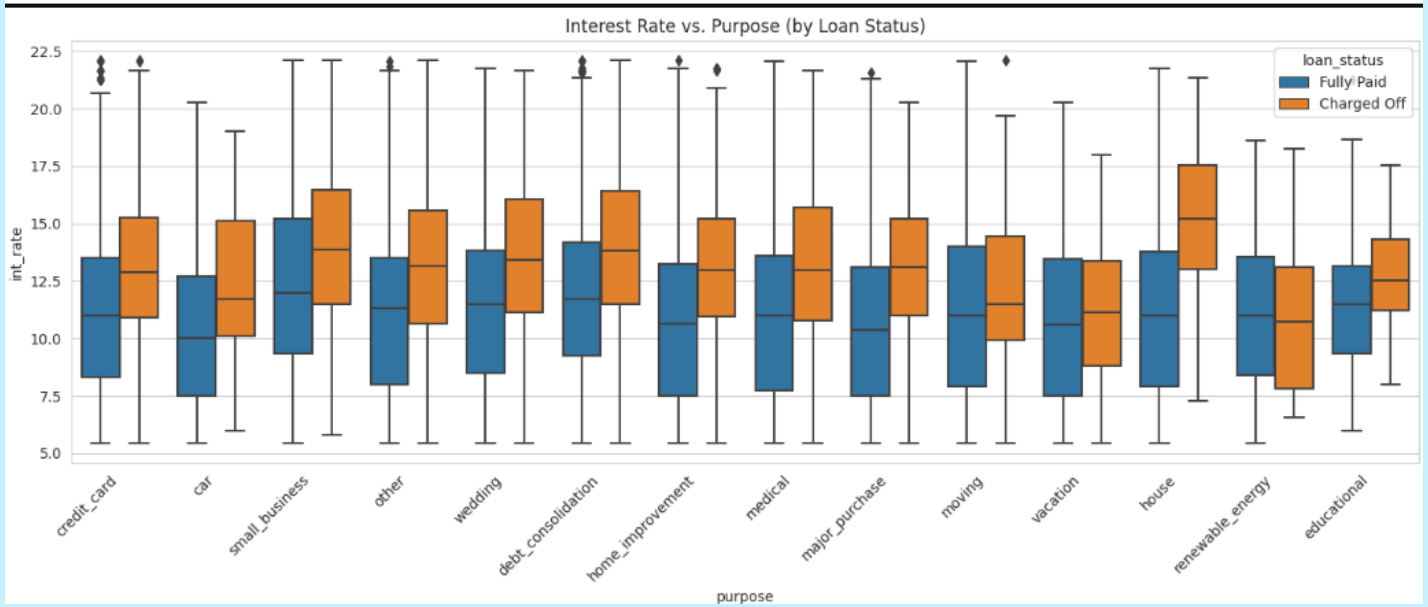


Bivariate Analysis (Loan Purpose and Risk)

- **Loan Amount vs. Purpose:**
  - Purpose and Loan Size: Small business and debt consolidation loans typically have the largest loan amounts.
  - Risk and Loan Size: Charged-off loans generally exhibit wider and higher loan amount distributions, especially for small business loans, indicating increased risk with larger loans within specific purposes.
- **Interest Rate vs. Purpose:**
  - Purpose and Interest Rate: Small business loans carry the highest interest rates, aligning with their higher risk profile.
  - Purpose as Primary Driver: Interest rate distributions are similar for charged-off and fully paid loans within each purpose, indicating that purpose is the main determinant of interest rate. However, the high rates on small business loans likely contribute to their elevated default rates.

Plots

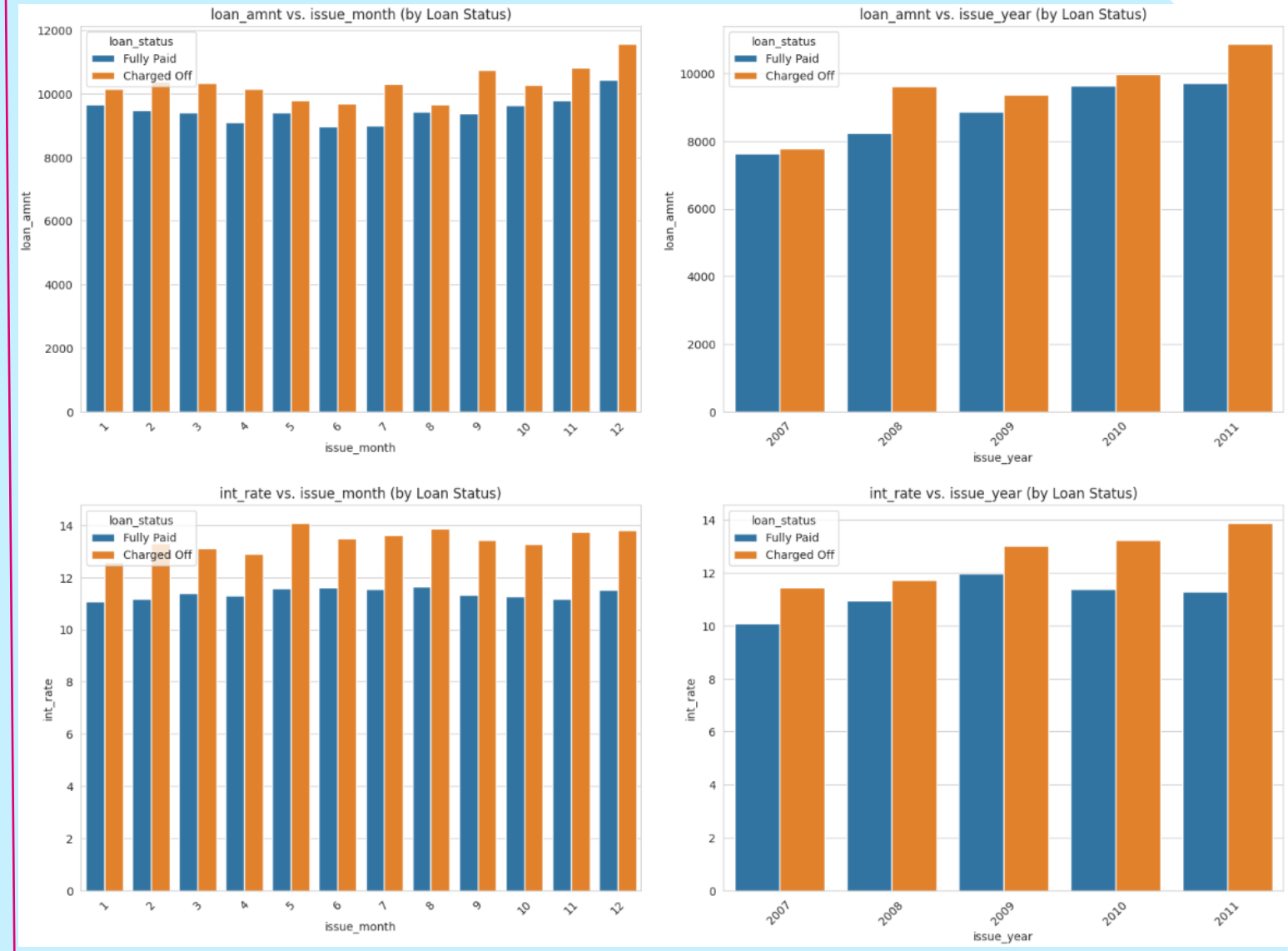




Bivariate Analysis (Time-Based Analysis)

Plots

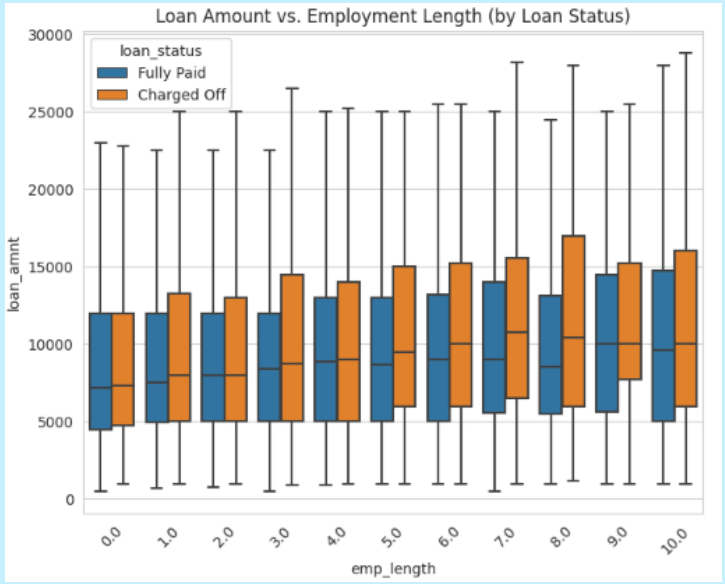
- **Loan Amount vs. Issue Month/Year:**
  - No Seasonal Pattern:
    Loan amounts show some monthly variation but no clear seasonal trend.
  - Increasing Trend Over Time:
    Loan amounts increased significantly from 2007 to 2011, potentially due to changing economic conditions or lending practices. This increase might have contributed to the higher default rates observed in 2011.
- **Interest Rate vs. Issue Month/Year:**
  - No Seasonal Pattern:
    Interest rates show monthly fluctuations but no consistent seasonal trend.
  - Increasing Trend Over Time:
    Interest rates also increased over the years, especially for charged-off loans, likely reflecting higher perceived risk by lenders. This increase in rates could be a contributing factor to the higher default rates observed in 2011.

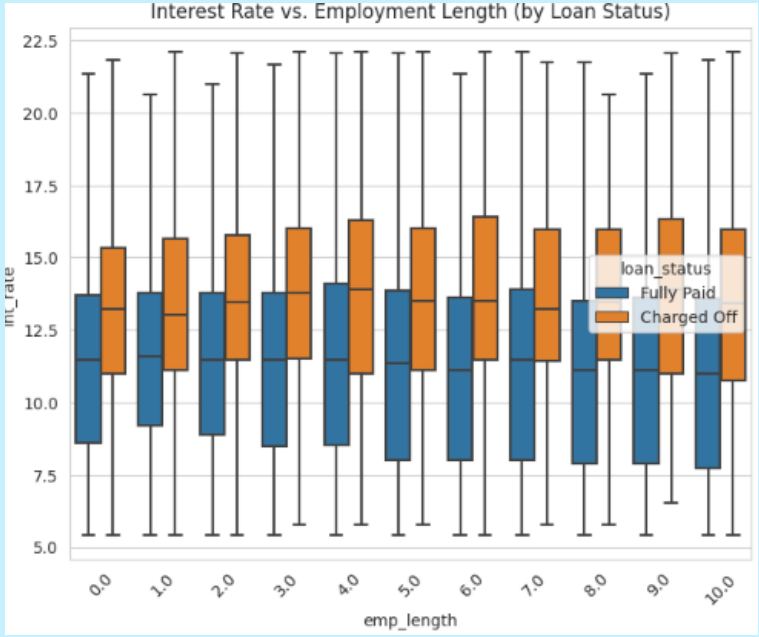


Bivariate Analysis (Employment Length and Loan Characteristics)

Plots

- **Loan Amount vs. Employment Length:**
  - Slight Upward Trend: Borrowers with longer employment histories tend to take out slightly larger loans, but the difference is not substantial.
  - Limited Predictive Power: Charged-off loans have slightly higher loan amounts across all employment lengths, but significant overlap in distributions suggests employment length alone is not a strong predictor of loan amount or default risk.
- **Interest Rate vs. Employment Length**:
  - No Clear Trend: Interest rates show no discernible relationship with employment length.
  - Employment Length Not a Key Driver: Similar interest rate distributions for charged-off and fully paid loans across all employment lengths reinforce that employment length is not a primary driver of interest rates or default likelihood.

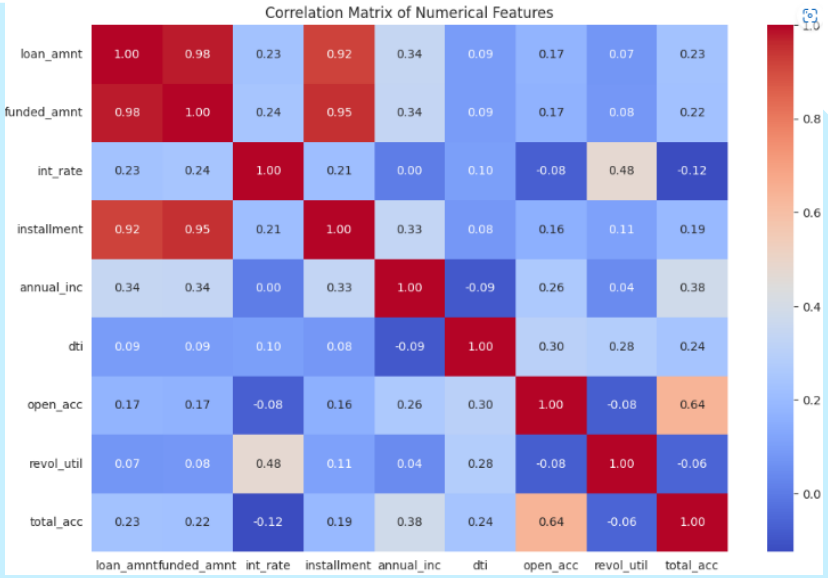


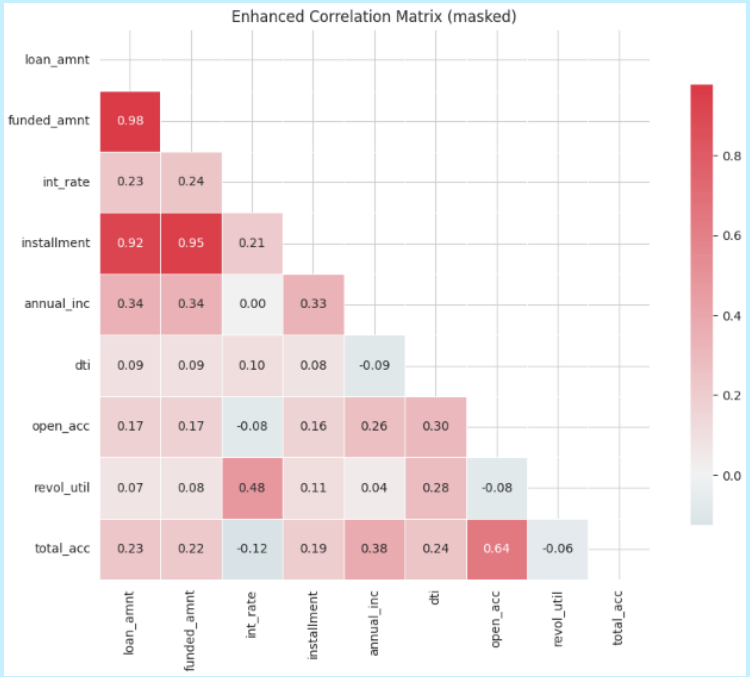


Correlation Analysis

- **Strong Positive Correlations:**
  - `loan_amnt` and `funded_amnt` (0.98): Indicates almost perfect agreement between requested and funded loan amounts. Suggests potential multicollinearity.
  - `loan_amnt` and `installment` (0.92): Larger loans lead to higher installment payments.
  - `funded_amnt` and `installment` (0.95): Similar to above; reinforces the relationship between funded amount and installment payments.
- **Moderate Positive Correlations:**
  - Several variables exhibit moderate positive correlations (0.2-0.4), indicating some tendency to move together, but not as strongly as the highly correlated variables.
- **Weak Correlations:**
  - Most remaining variable pairs show weak correlations (near zero), suggesting limited linear relationships.
- **Negative Correlations:**
  - A few weak negative correlations are present, but none are strong enough to indicate substantial inverse relationships.
- **Multicollinearity Concerns:**
  - The extremely high correlation between `loan_amnt` and `funded_amnt` raises strong multicollinearity concerns. In modeling, only one of these variables should be used.
  - The high correlations involving `installment` also suggest potential multicollinearity issues, though less severe. Careful variable selection is recommended for modeling.

Plots:





Driver Variables for Identifying Potential Loan Defaulters

1. **Grade (grade (and sub_grade)):** Loan grade is a strong predictor of default, with lower grades (B-G) showing significantly higher default rates. This is because lenders assign grades based on their assessment of borrower risk, considering factors like credit history and existing debt levels. Sub-grades offer even finer granularity for risk differentiation.

2. **Interest Rate (int_rate):** High interest rates are a key indicator of risk, and the analysis confirms a strong positive relationship between interest rate and default rate. Higher rates often reflect the lender's perception of increased borrower risk and contribute to a heavier repayment burden, making default more likely.

3. **Term (term):** Longer loan terms (60 months) have a substantially higher default rate compared to shorter terms (36 months). While 36-month loans may have more defaults in absolute numbers due to higher volume, the proportional risk is greater with longer terms, possibly due to increased exposure to financial hardship over time.

4. **Debt-to-Income Ratio (dti):** A high debt-to-income ratio (DTI), particularly in the 15-25 range, is a strong predictor of default. This indicates that a large portion of the borrower's income is already allocated to debt repayment, leaving less room to handle unexpected expenses or financial shocks, thus increasing their vulnerability to default.

5. **Loan Purpose (purpose):** Certain loan purposes are associated with higher default risk. Small business loans, for instance, have the highest default rate due to the inherent risks of business ventures, coupled with often larger loan amounts and higher interest rates. Debt consolidation loans also signal elevated risk, as they often indicate borrowers are already struggling to manage existing debt.

6. **Loan Amount (loan_amnt):** Larger loan amounts, especially those above 16k, correlate with higher default rates. This is likely because larger loans result in higher monthly payments and a greater overall repayment burden, increasing the financial strain on borrowers and making them more susceptible to default.

7. **Home Ownership (home_ownership):** Renters and those with "OTHER" homeownership status exhibit higher default rates compared to homeowners. This suggests that homeownership may serve as a proxy for greater financial stability and access to alternative financial resources in times of hardship.

8. **Annual Income (annual_inc):** While not a strong standalone predictor, annual income plays a role in default risk. Lower annual incomes (<25k) are associated with higher default rates, indicating that borrowers with limited financial capacity may be more vulnerable to unforeseen circumstances that lead to default. It's important to consider income in conjunction with other factors like DTI and loan amount.

Summary: Lending Club Case Study

1. **Loan Amount and Interest Rate**: Smaller loans (4k-12k) with mid-range interest rates (9-17%) are common among defaults, suggesting borrowers may be overextending their finances. Larger loans, particularly those with higher interest rates, also present elevated risk due to the increased repayment burden. This highlights the combined effect of loan amount and interest rate in assessing default risk.

2. **Financial Health (Income and DTI):** Defaults are concentrated among borrowers with annual incomes of 25k-75k, indicating that income alone is not a reliable predictor of default. DTI plays a crucial role, with moderate to moderately high DTIs (10-20) being prevalent among defaulters. High DTI, even with a reasonable income, is a significant risk factor, suggesting potential difficulty managing debt and financial overextension.

3. **Credit History (Open Lines, Utilization, Total Accounts**): Defaulters typically have a moderate number of open credit lines (5-11), indicating established credit but possible challenges in managing multiple accounts. High revolving credit utilization (68-85%) is a risk factor, although defaults also occur at lower utilization levels, suggesting other contributing factors. A moderate to high total number of accounts (8-24) is also common, potentially reflecting a history of seeking credit.

4. **Loan Characteristics (Grade, Term, Purpose):** Several loan characteristics are associated with increased default risk. These include lower loan grades (B, C, F, G), sub-grades B5 and C1, 36-month loan terms, and debt consolidation as the loan purpose. Longer terms (60 months), while less frequent among defaults, carry higher risk due to potentially larger loan amounts and accumulated interest. Small business loans are particularly risky due to their larger amounts, higher rates, and inherent business uncertainties.

5. **Employment Length**: Both very short (<1 year) and very long (10+ years) employment histories show elevated default rates. Short employment may indicate instability, while long employment might lead to overconfidence in borrowing capacity. This counterintuitive finding warrants further investigation.

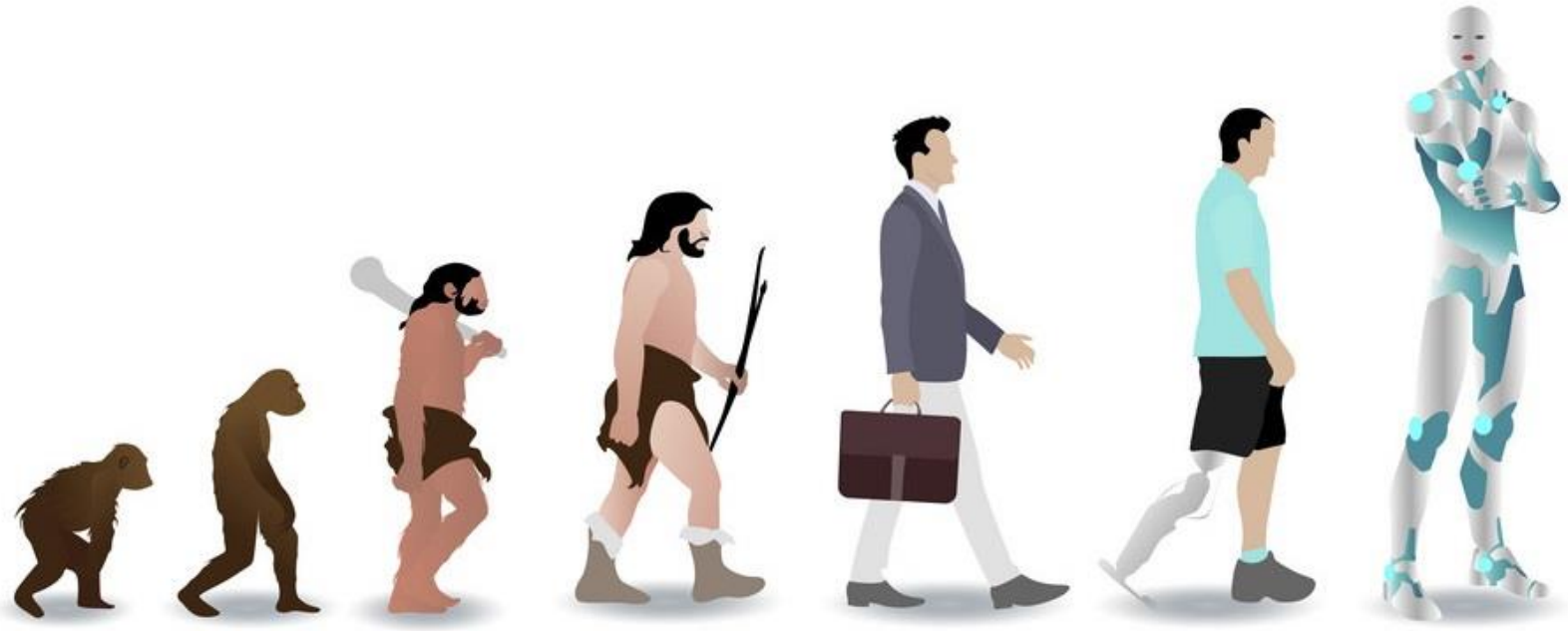
6. **Time of Loan (Year, Month):** Most defaults occurred in 2011, possibly due to economic conditions or lending practices specific to that period. Increased defaults in November and December suggest seasonal influences, such as financial pressures related to holiday spending.

7. **Home Ownership and Loan Purpose:** Renters and individuals with non-traditional homeownership ("OTHER") experience higher default rates, potentially due to lower financial stability. Debt consolidation is a frequent loan purpose among defaulters, often indicating a last-resort attempt to manage existing debt. Small business loans remain a high-risk category.

8. **Interaction of Loan Grade and Loan Amount:** Lower loan grades are associated with both higher default rates and larger loan amounts. This combination significantly increases the risk of default, as borrowers with weaker credit profiles take on more substantial debt.

9. **Risk Factors for Small Business Loans:** Small business loans are inherently risky due to factors such as larger loan amounts, higher interest rates, and the general uncertainties of business ventures. These factors contribute to the elevated default rates observed in this category.

10. **DTI and Loan Grade Interaction:** While DTI is a risk factor across all loan grades, its influence is amplified for lower grades. Borrowers with lower loan grades and higher DTIs are particularly vulnerable to default. Lenders should carefully consider the combined effect of these two factors



THANK YOU !