Strategic Insights for Minimizing Credit Risk

Lending Club Case Study: Deep Dive into Data-Driven Recommendations

ML C59 EPGP ML&AI Batch

- Presented By
  - Anirban Gangopadhyay
  - KK Rahul

Overview of Business Objectives and Credit Risk

Lending Club Overview: Largest online marketplace for personal, business, and medical loans.

Accessible Financing: Fast online interface providing lower interest rate loans.

Credit Loss Challenge: High-risk applicants are the primary source of financial loss.

Defaulters Impact: Notably, 'charged-off' customers represent a significant default risk.

Risk Identification Goal: Aims to reduce credit loss by identifying risky loan applicants.

EDA Utilization: Perform Exploratory Data Analysis to determine key default indicators.

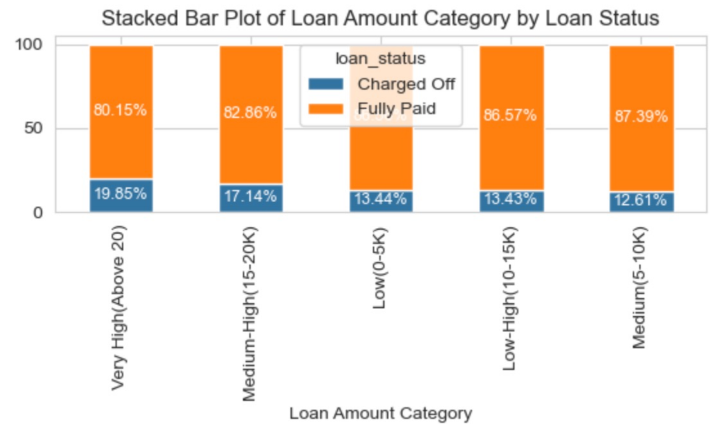
Strategic Importance: Knowledge of risk factors essential for portfolio and risk assessment.

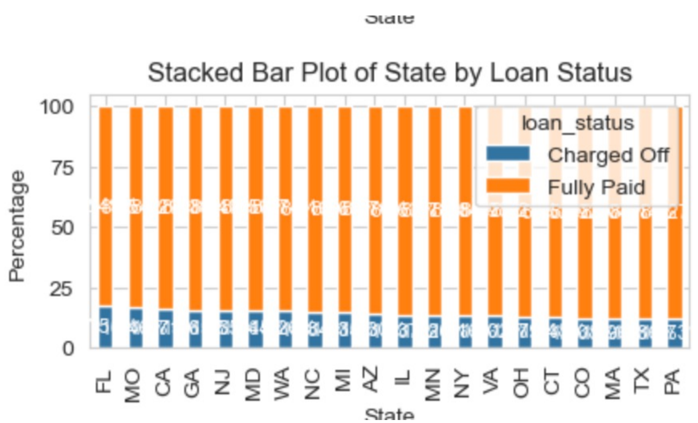
Data Cleaning and Preprocessing

- **Initial Dataset Composition:**
  - Started with 39,717 rows and 111 columns.
- **Feature Reduction:**
  - Excluded features not available at the loan application stage.
  - Removed unique-only features and high-unique-value features
  - Dropped those over 60% null.
  - Eliminated irrelevant features
  - Discarded the ones with singular values or NaNs.
- **Row Exclusion Criteria:**
  - Dropped rows with 1%-3% null values in certain columns.
  - Excluded 'current' loan status entries.
- **Data Refinement:**
  - Purged unnecessary text from feature values.
  - Adjusted data types for better analysis compatibility.
- **Date Correction:**
  - Corrected data entry errors in “Earliest Credit Line(earliest_cr_line)” feature from '20' to '19’.
- **Final Dataset:**
  - Ended up with 36,789 rows and 28 columns.

Exploratory Data Analysis(EDA)

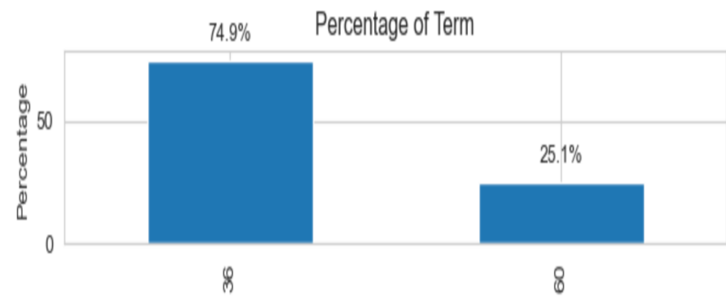
- **Comprehensive Analysis Approach:** Conducted univariate, bivariate, and multivariate analysis to uncover patterns and correlations.
- **Derived Features:** Developed Derived Features for enhanced insights and accuracy.
- **Predictive Analysis:** Executed predictive analysis to identify key driver variables influencing loan defaults.
- **Outlier Management:** Addressed and treated outliers for more robust data interpretation.
- **Focused Exploration:** The following slides will focus on key aspects of our analysis, providing a deep dive into selected areas of our EDA, demonstrating its thoroughness and analytical scope.

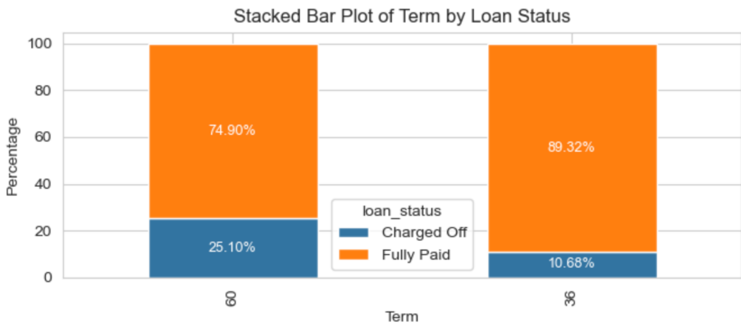




Loan Amount, Geographic Trends, and Default Risk

- **Data Range and Outliers:** Observed significant outliers in the loan amount data, necessitating binning into categories: 'Low (0-5K)', 'Medium (5-10K)', 'Low-High (10-15K)', 'Medium-High (15-20K)', 'Very High (Above 20K)'.
- **Popularity Trends:** Medium-sized loans are more popular, while loans above 15K, especially those over 20K, are less common.
- **Risk of Default:** Notable increase in 'Charged-off' loans in 'Medium-High' and 'Very High' categories.
- **Zip Code Analysis:** Variations in loan application frequencies across zip codes, with certain areas showing higher loan defaults.
- **State-Based Loan Distribution:** States like CA, NY, FL, and TX lead in loan numbers, with CA having the highest. Loans from FL, MO, and CA show higher default probabilities.
- **Predictive Insights:** Loan amount grouped into bins, zip code, and state analysis offer strong predictions of default risk, with combined state and loan amount features being particularly potent indicators.



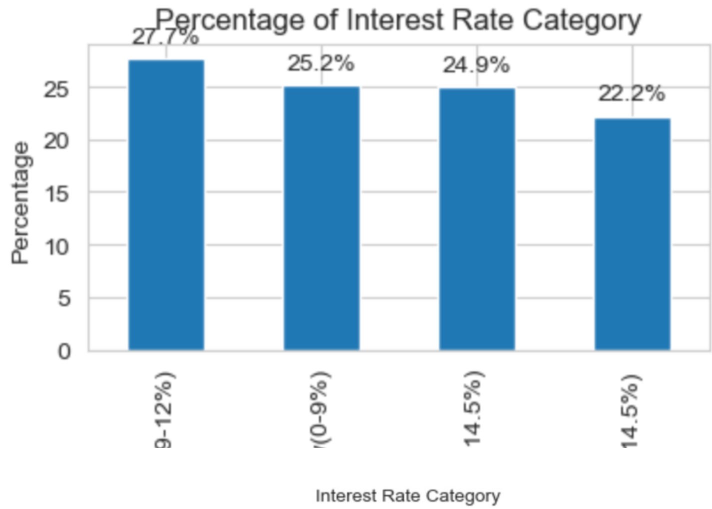


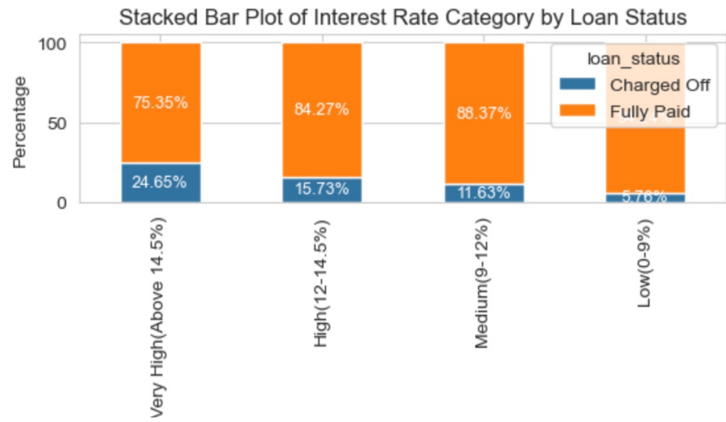
Analysis of Loan Term Preferences and Risks

- **Borrower Preferences:** A clear preference for 36-month loan terms over the 60-month options among borrowers.
- **Risk Assessment:** There is an elevated risk associated with the 60-month term loans in terms of repayment reliability.

Interest Rate Analysis and Default Risk

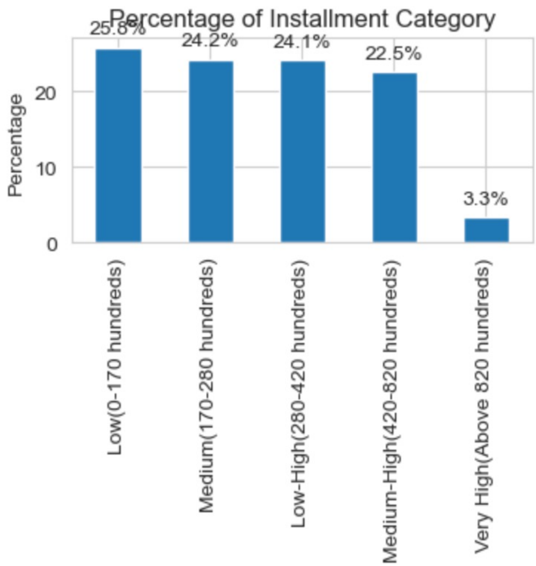
- **Loan Demand Across Interest Categories:** Univariate analysis indicates evenly distributed demand for loans across different interest rate categories.
- **Risk Escalation with Higher Rates:** Bivariate analysis reveals a significant increase in 'Charged Off' loans as interest rates rise, particularly in high and very high categories.
- **Default Probability:** Loans with interest rates above 12% show a higher default likelihood, escalating markedly for rates over 14.4%.
- **Predictive Strength:** The interest rate is a strong predictor of loan default risk.

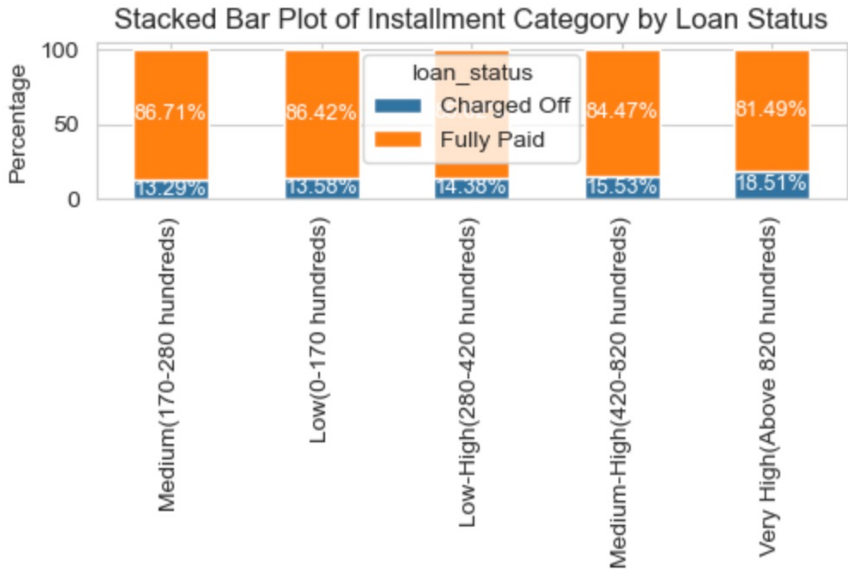




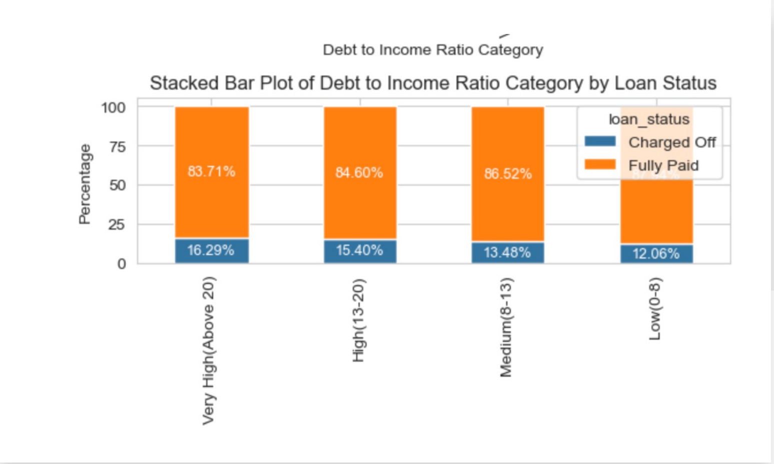
Installment Payments and Loan Default Risk

- **Loan Distribution by Installment:** Univariate analysis shows a consistent distribution of loans with installment payments up to $820, followed by a decline in loans with higher installments.
- **Increased Default Beyond Certain Thresholds:** Bivariate analysis indicates a rise in 'Charged Off' loans for installments beyond $420, with a significant spike in defaults for installments over $820.

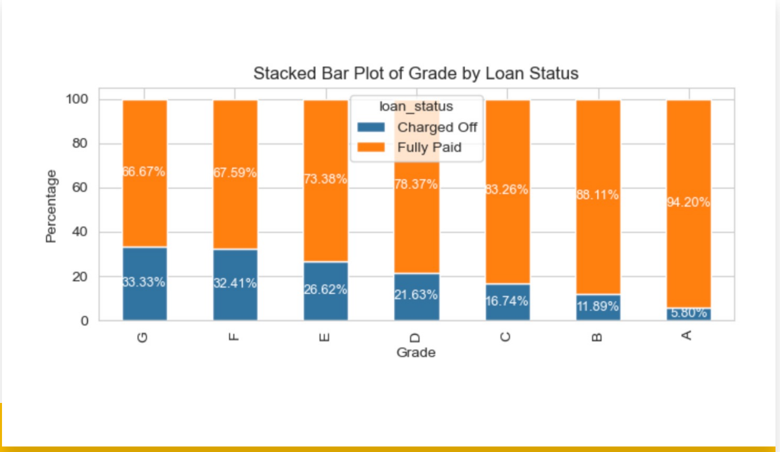




Grade, DTI, and Default Risk Insights

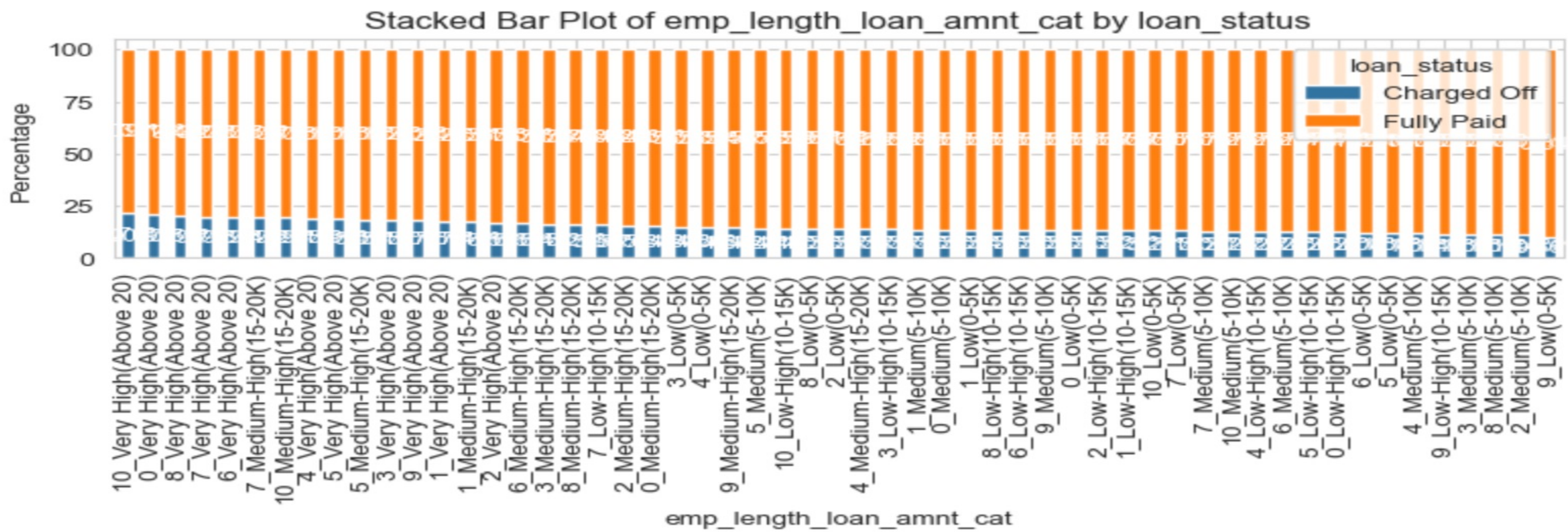


- **Grade Distribution:** Majority loans in 'A' and 'B' grades. Higher grades indicate lower default risk, while 'F' and 'G' grades are most risky.
- **DTI Binning and Analysis:** New feature 'dti_category' with bins 'Low', 'Medium', 'High', and 'Very High'. Higher DTI categories show increased default likelihood.
- **Grade and DTI Correlation:** Bivariate analysis reveals that loans with 'F' and 'G' grades tend to default even at low (0-5) or medium (8-13) DTI ratios.
- **Resilience of High Grades:** In contrast, 'A' and 'B' grade loans show a higher likelihood of full repayment, even with very high DTI ratios (above 20).
- **Predictive Power:** These insights underline the strong predictive ability of grades and DTI ratios in assessing loan default risks.



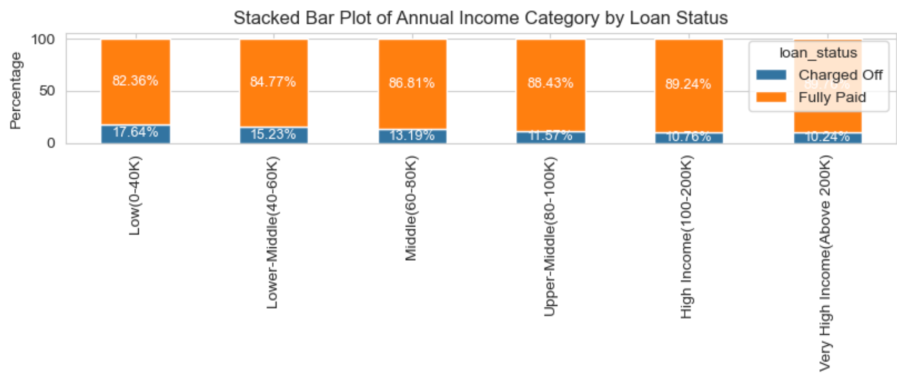
Employment Length, Home Ownership, and Loan Risk

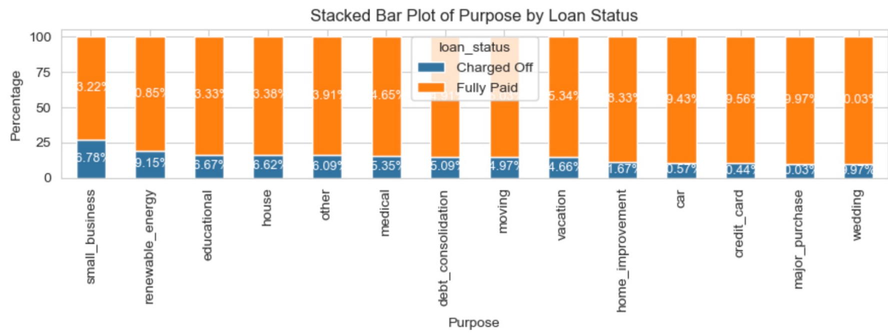
- **Employment Length Insight:** Longer employment suggests stability but is not a conclusive predictor of loan repayment, with similar default rates across varying employment lengths.
- **Home Ownership Patterns:** A mix of renters, mortgage holders, and fewer outright homeowners, with 'Others' showing a higher default tendency. Moderately predictive of loan status.
- **Derived Feature:** Home Ownership and Loan Amount: 'Other' home status with high loans (> $10K) indicates higher default risks. Renters default more at very high loan amounts (> $20K), while mortgages with smaller loans (up to $15K) tend to be safer. Combining home ownership status with loan amount reveals stronger predictive potential.

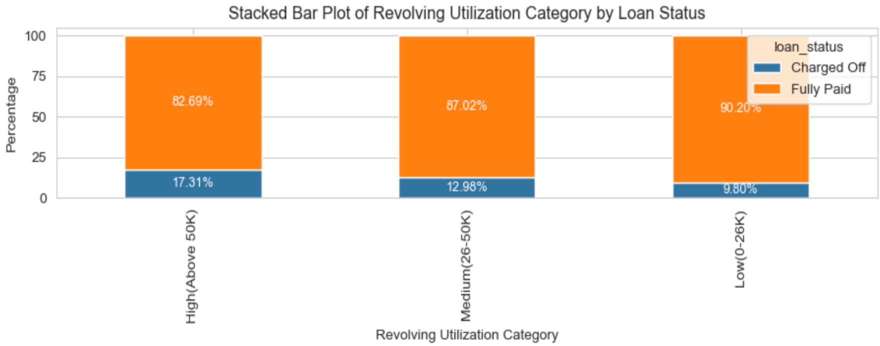


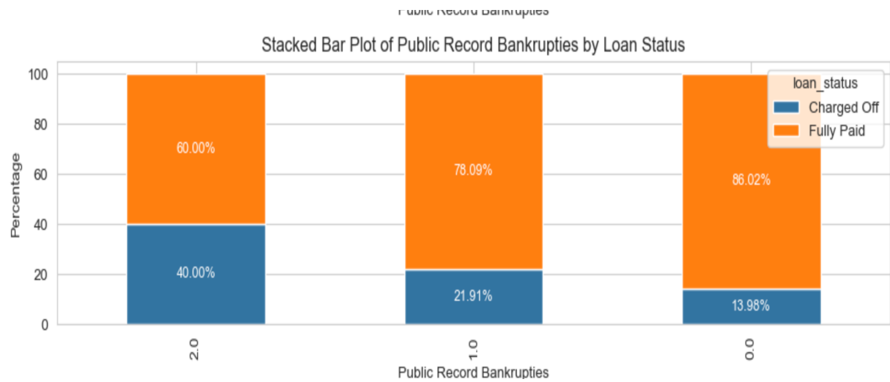
Annual Income and Loan Purpose: Insights

- **Annual Income Insights**: Created 'annual_inc_category' to analyze income patterns. Higher default risk noted in those earning below $40,000, with a slight risk up to $80,000.
- **Loan Purpose Trends:** 'Debt_consolidation' and 'credit_card' are the most common purposes. 'Small_business' and 'renewable_energy' loans exhibit higher charge-offs.
- **Predictive Analysis:** Income and loan purpose, especially when combined, offer strong predictive insights into loan performance, albeit with some variability due to different overlapping terminologies in the purpose category.







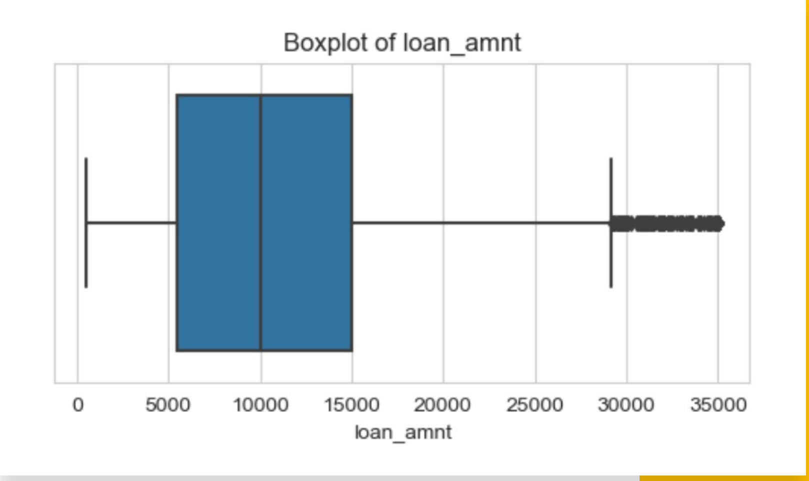


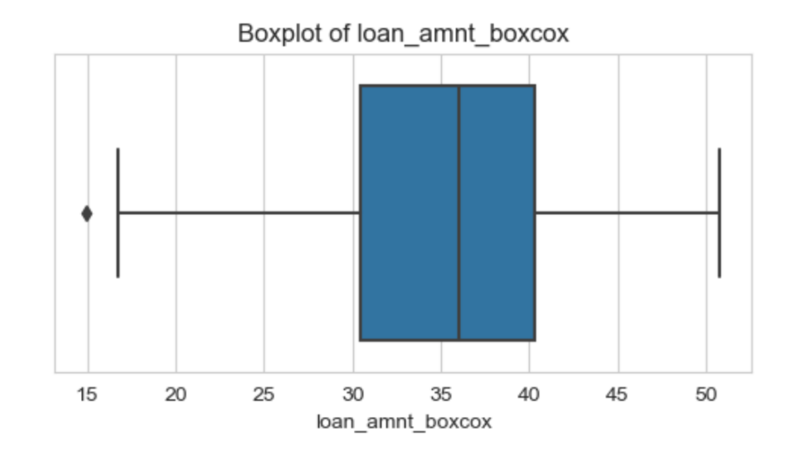
Credit History and Utilization: Predictors of Loan Default

- **Derogatory Public Records & Bankruptcies**: Majority have no derogatory records or bankruptcies, yet their presence significantly increases charge-off risk.
- **Credit Utilization Trends:** High utilization rates correlate with increased default probability.
- **Predictive Analysis:** Both public records and revolving utilization are strong predictors, indicating financial stress and risk of default among borrowers.

Outlier Transformation Techniques in Key Features

- **Transformations Applied:** Employed various transformations to normalize data and reduce the impact of outliers in key columns.
    - **Loan Amount:** Box-Cox transformation for a more normal distribution.
    - **Interest Rate:** Log transformation to address skewness.
    - **Installment and Annual Income:** Box-Cox transformation for normalization.
    - **Revolving Balance:** Yeo-Johnson transformation to manage extreme values.
    - **Revolving Utilization:** Z-score transformation for standardization.
- **Objective:** Enhance data quality for more accurate predictive modeling.





Conclusive Insights from EDA

- **Key Predictive Variables:** Several features emerged as strong predictors after binning and transformation, enhancing predictive accuracy.
    - Loan Term, Grade, Zip Code, State, and Public Record Bankruptcies independently show strong predictive power.
    - Loan Amount, Interest Rate, Installment, Annual Income, DTI, Inquiries in Last 6 Months, Public Records, and Revolving Utilization are particularly strong after categorization.
- **Binning Effectiveness:** Binned categories in various features like loan amount, interest rate, installment, annual income, and revolving utilization significantly improve their predictive strength.
- **Conclusion:** These findings offer a comprehensive understanding of risk factors in loan approval, crucial for effective risk management and decision-making.