Lending Club Case Study

- Krantisinh Deshmukh
- Yamini Dadhirao

Contents

Overview

- A consumer finance company specializes in lending various types of loans to urban customers
- The company has provided dataset (covers loan data issued between 2007 and 2011) that has the information about past loan applicants and whether they ‘defaulted’ or not
- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.
- We will use EDA to understand how consumer attributes and loan attributes that influence the tendency of default.

Problem Statement

1. Understanding Loan Default Patterns and Risk Factors
   - Use EDA techniques to assess the influence of consumer behavior and loan characteristics on repayment tendencies.
2. Optimizing Loan Approval Decisions
   - Develop insights to assist the company in making data-driven decisions regarding loan approvals, rejections, or adjusted terms
3. Reducing Financial Risk and Enhancing Profitability
   - Identify actionable recommendations to reduce financial risk while maintaining profitability in loan approvals.

Approach

- Data Exploration and Understanding
- Data Cleaning and Preparation
- Analyzing Data with Exploratory Data Analysis (EDA) Techniques
  - Univariate Analysis
  - Segmented Univariate Analysis
  - Bivariate Analysis
- Deriving Insights from Analysis
- Formulating Data-Driven Recommendations

Data Exploration and Understanding

- The data set has 39717 records and 111 columns
- The Data Dictionary is helpful to understand the meanings of the columns
- The columns can be classified as –
  - Consumer Attributes
  - Loan Attributes
  - Delinquency and Risk Indicators
  - Credit Utilization and Account Details
  - Metadata (these columns can be excluded from the analysis)
    - Unique Identifiers
    - Administrative
    - Less Relevant Financial Details

Data Exploration and Understanding

Category	Variables
Consumer Attributes	annual_inc, annual_inc_joint, emp_length, home_ownership, zip_code, addr_state, verification_status, verification_status_joint
Loan Attributes	loan_amnt, issue_d, term, loan_status, installment, out_prncp, out_prncp_inv, int_rate, total_pymnt, total_rec_prncp, total_rec_int, grade, sub_grade, purpose, policy_code
Delinquency Indicators	delinq_2yrs, mths_since_last_delinq, collections_12_mths_ex_med, chargeoff_within_12_mths, pub_rec_bankruptcies, delinq_amnt, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, inq_last_6mths, inq_last_12mths, acc_open_past_24mths, tot_coll_amt
Credit Utilization	revol_bal, revol_util, open_acc, total_acc, bc_util, avg_cur_bal, total_bal_ex_mort, total_bc_limit, all_util

Data Cleaning and Preparation

Sr No.	Anomalies	Correction
1	Few columns have NA Values	Remove columns that have all rows as NA
2	Columns int_rate and revol_util have a trailing percentage	Remove the trailing % from these columns Convert the datatype of these from object to numerical
3	last_pymnt_d & issue_d columns have object datatype	Convert the last_pymnt_d column to datetime format for proper date handling.
4	zip_code has a trailing xx	Remove the trailing 'xx' from the zip_code column to retain only the relevant numeric part.
5	emp_length is an object	Convert the emp_length to numerical data type by getting rid of trailing string year or years

Analyzing Data

- We've used the following techniques of the Exploratory Data Analysis to analyze the data –

Technique	Description
Univariate Analysis	Examining a single variable to summarize its characteristics and identify patterns
Segmented Univariate Analysis	Analyzing a single variable across different segments or groups to uncover variations within those segments
Bivariate Analysis	Studying the relationship between two variables to understand how one may affect the other.

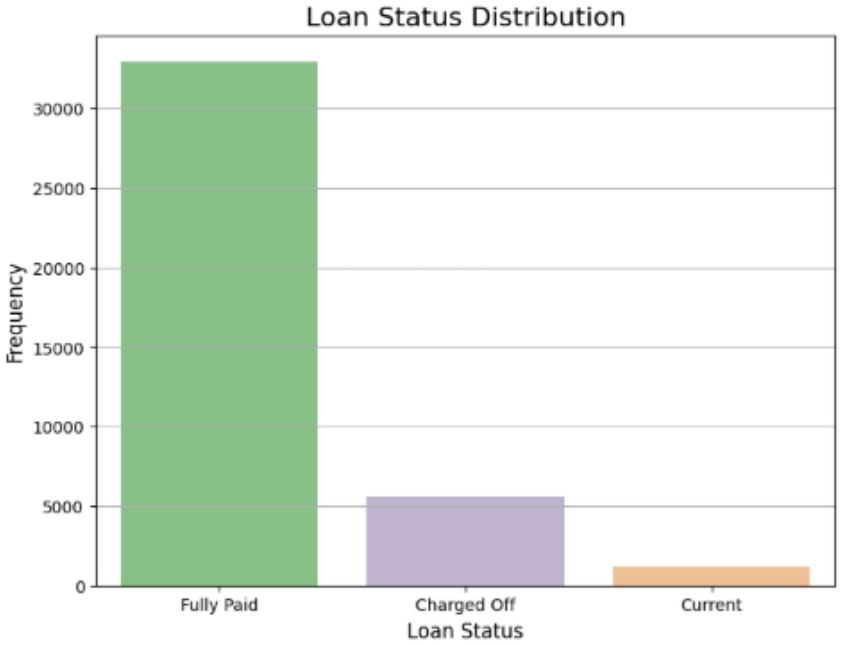
Univariate Analysis

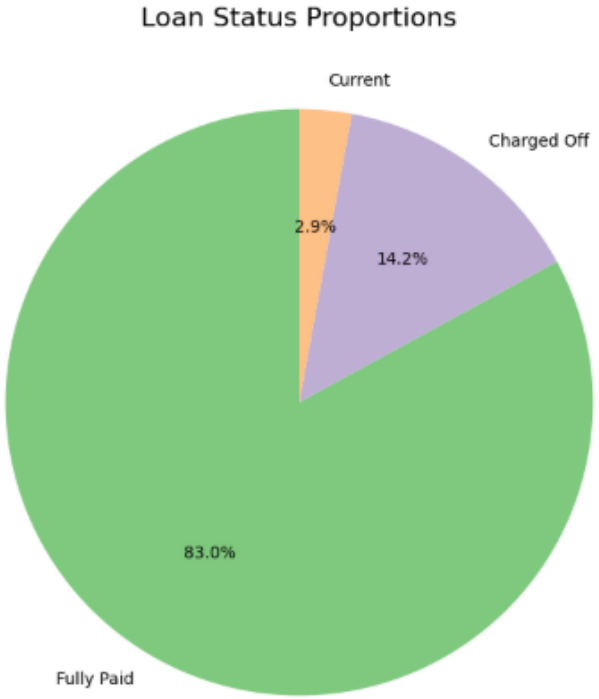
Summary of Insights

- Borrowers with high DTI (>20), recent delinquencies (<12 months), and high utilization rates (>80%) are key risk groups.
- Lower-grade loans (D, E, F, G) and high-interest loans (>17%) correlate strongly with defaults.
- Geographic trends (specific zip codes and states like NE and NV) highlight regions requiring risk-adjusted lending policies.
- Debt consolidation and credit card refinancing loans dominate but show higher risk, necessitating stricter thresholds.
- Verification status and revolving balances are critical indicators of borrower reliability and repayment capacity.

loan_status distribution

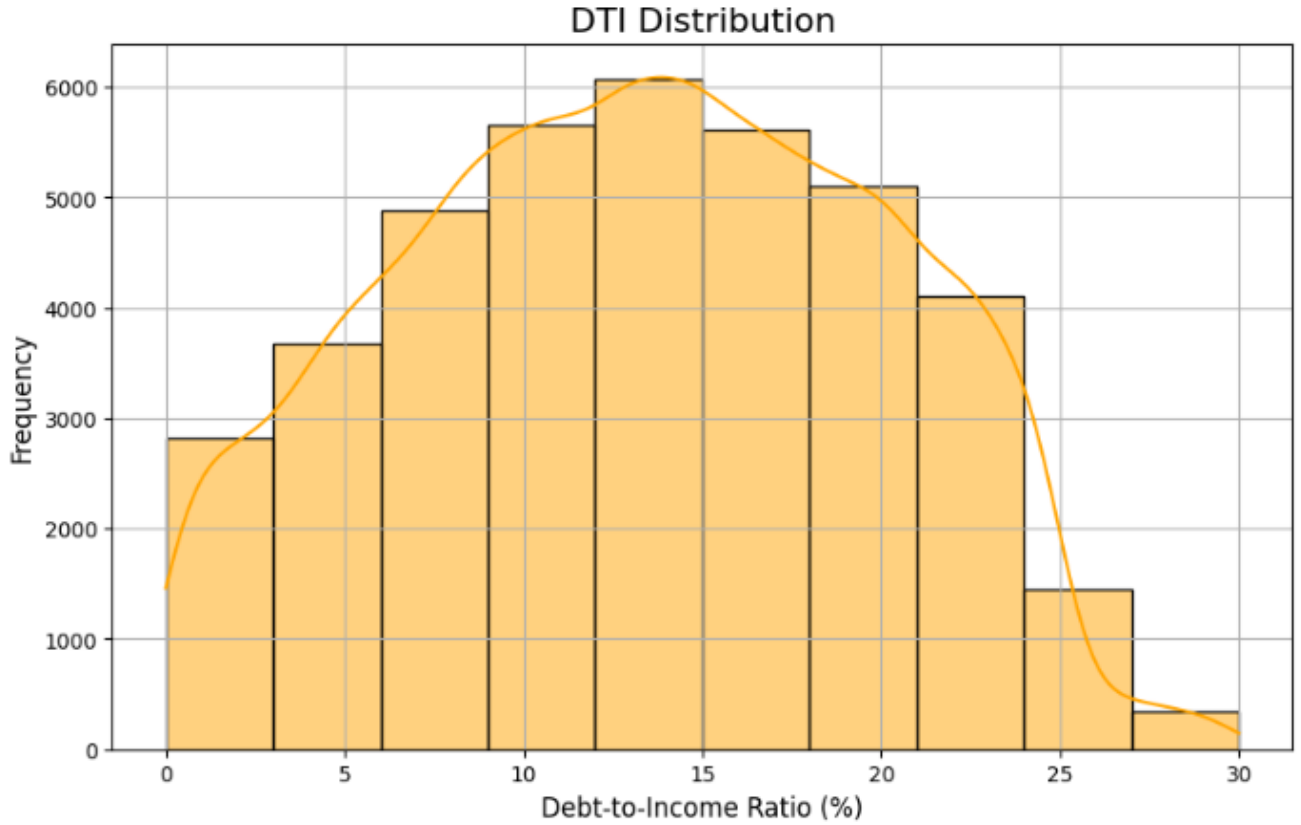
- 83% loans are fully paid, 14.2% loans are charged off while 2.9% loans are active





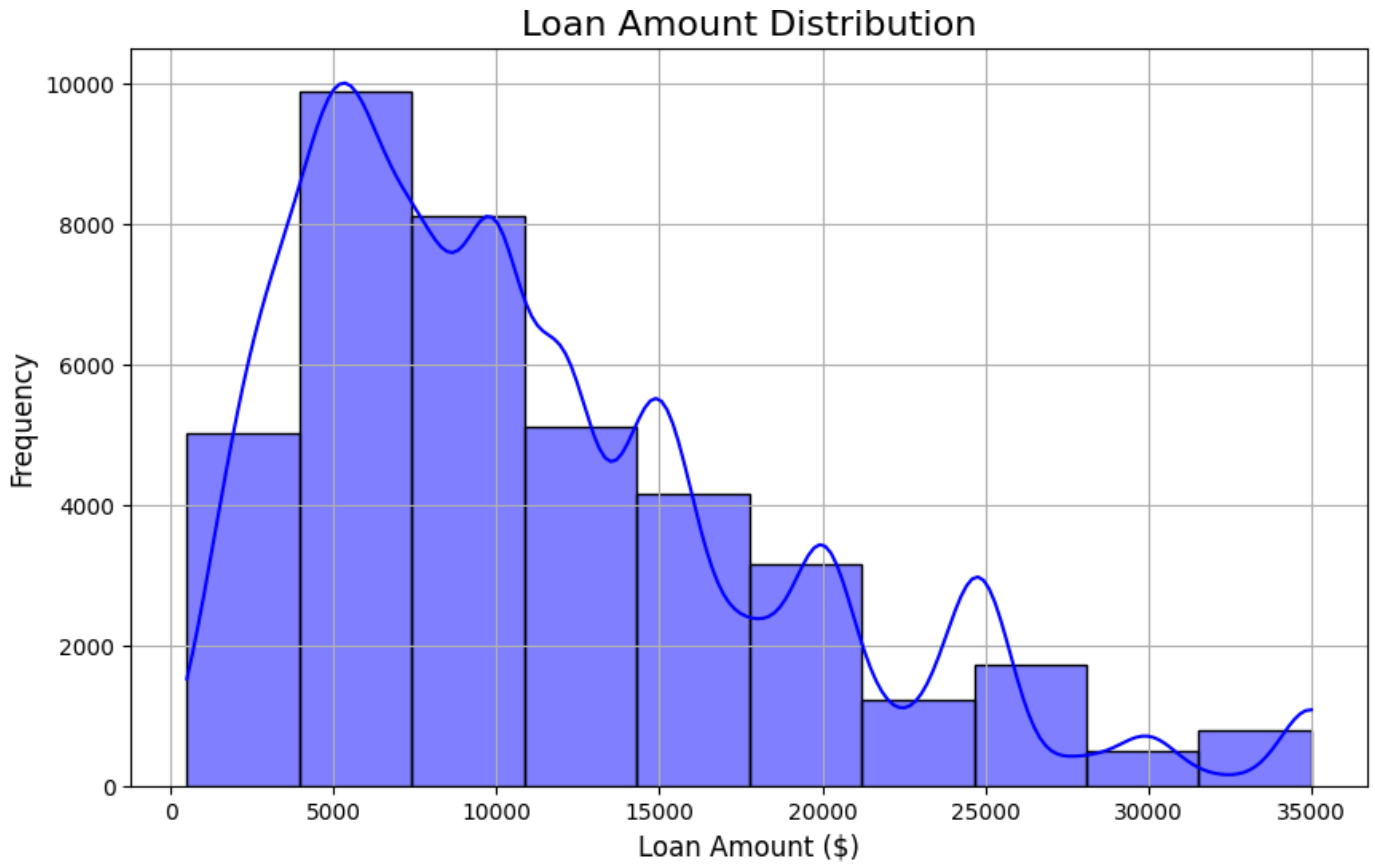
dti (debt to income ratio) distribution

- Few borrowers have dti > 20 and they can be categorized as medium risk.
- Most borrowers have acceptable values of dti and are a low risk.



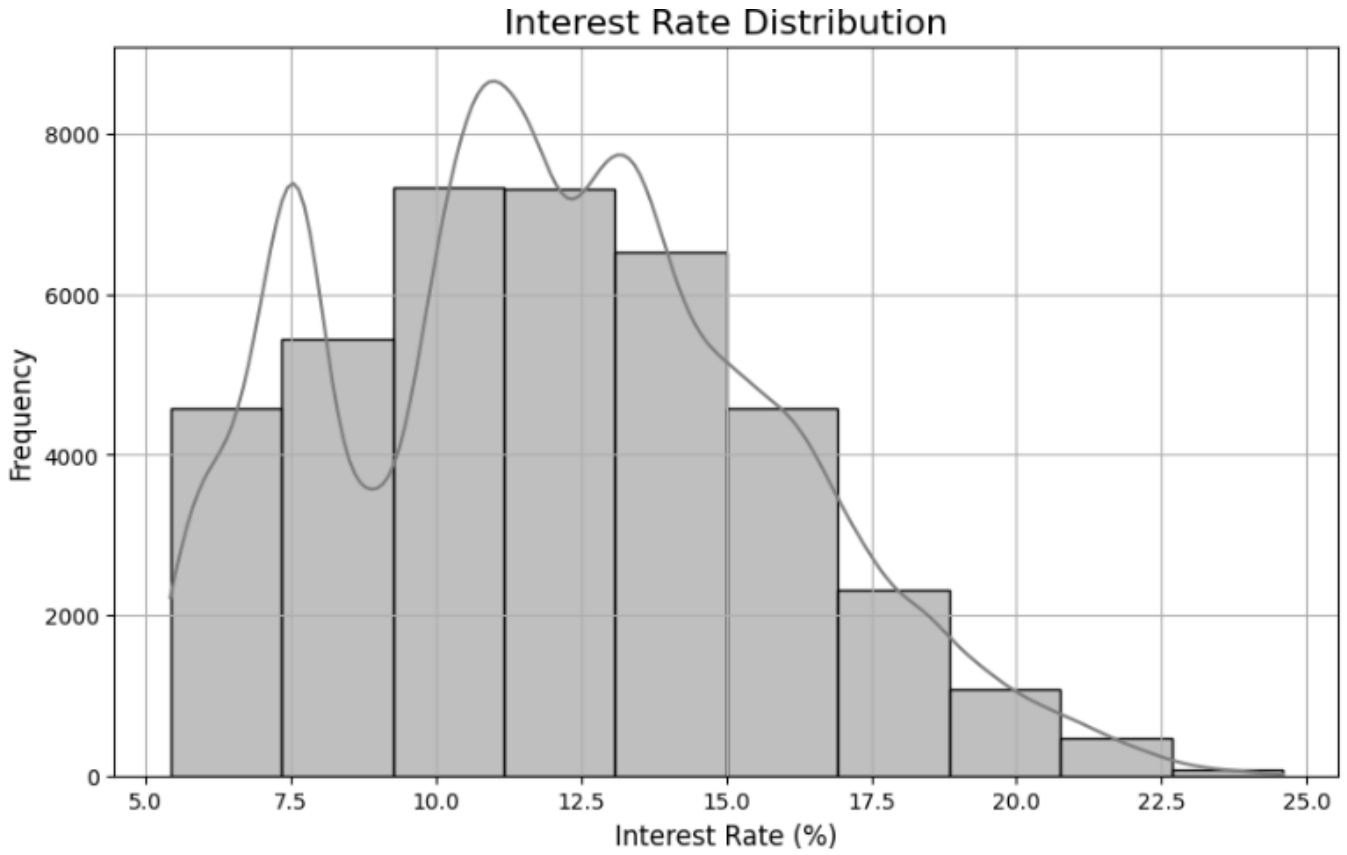
loan_amnt distribution

- Most users prefer to take loan of $10,000, possibly influenced by lender policies or borrower needs
- The distribution plot of the loan amount is right-skewed which indicates that most borrowers request smaller loans, with a few requesting much larger amounts.



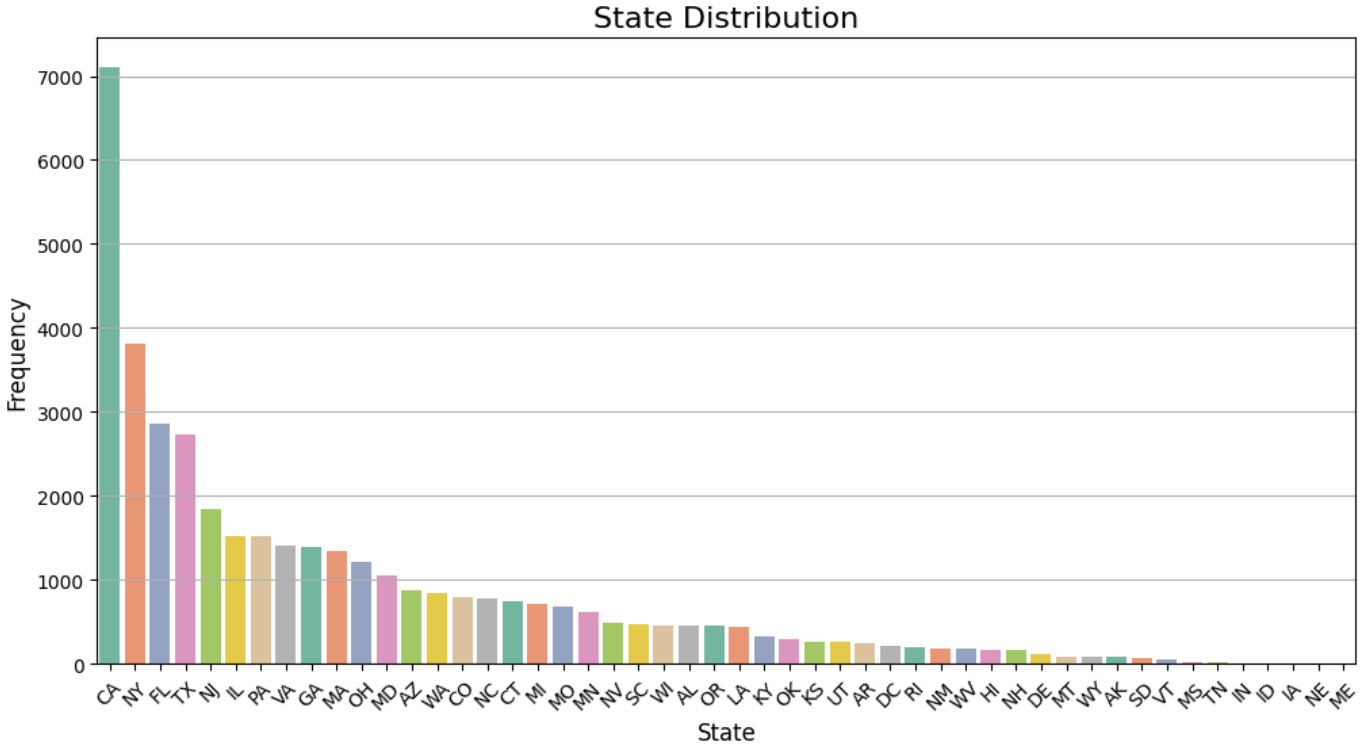
int_rate distribution

- The median interest rate is 11.86%
- Extremely high interest rates may be associated with subprime borrowers or risk-based pricing
- It is a right-skewed distribution indicating that most loans have moderate interest rates, with a few high-interest loans.



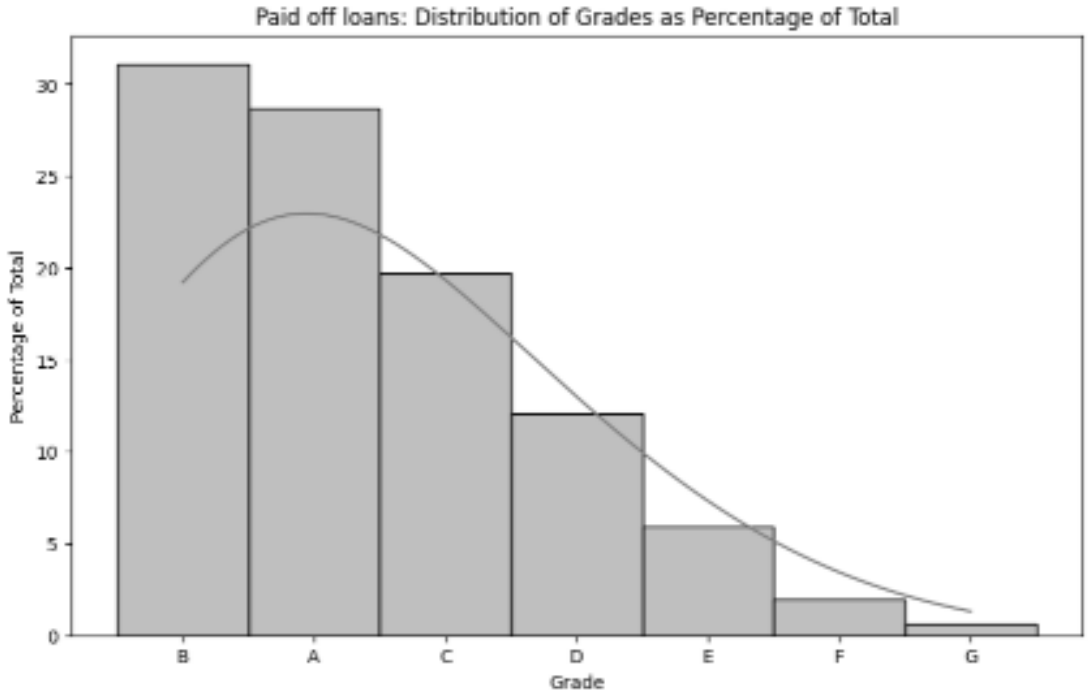
addr_state distribution

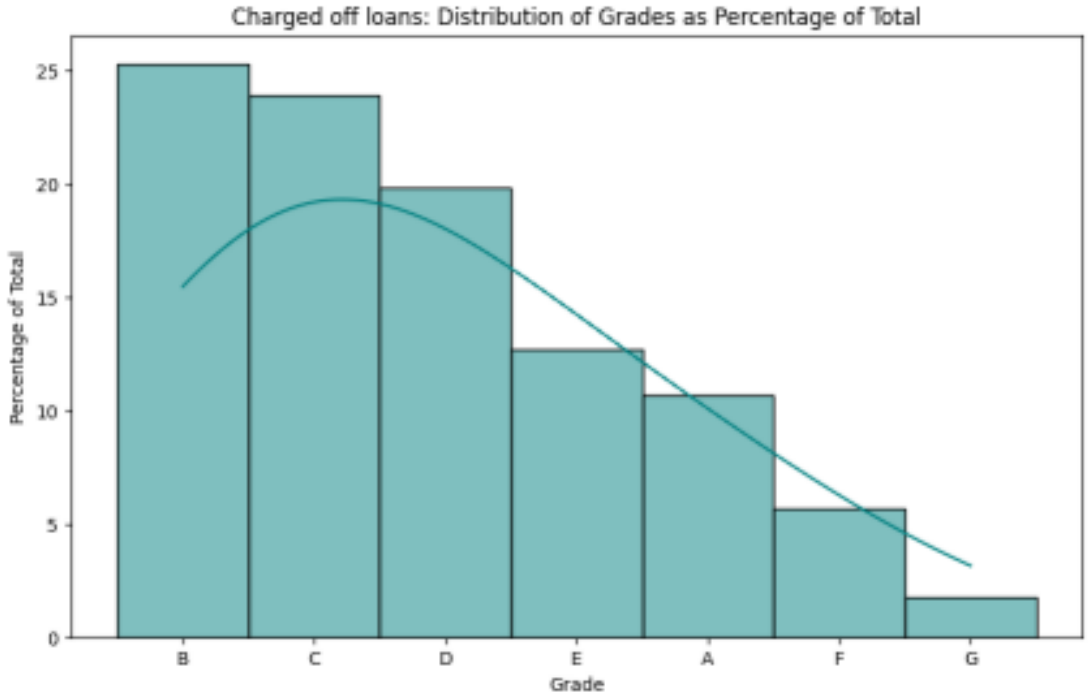
- States like CA, NY, FL and TX dominate the dataset, reflecting population density or lending activity.



Grade distribution

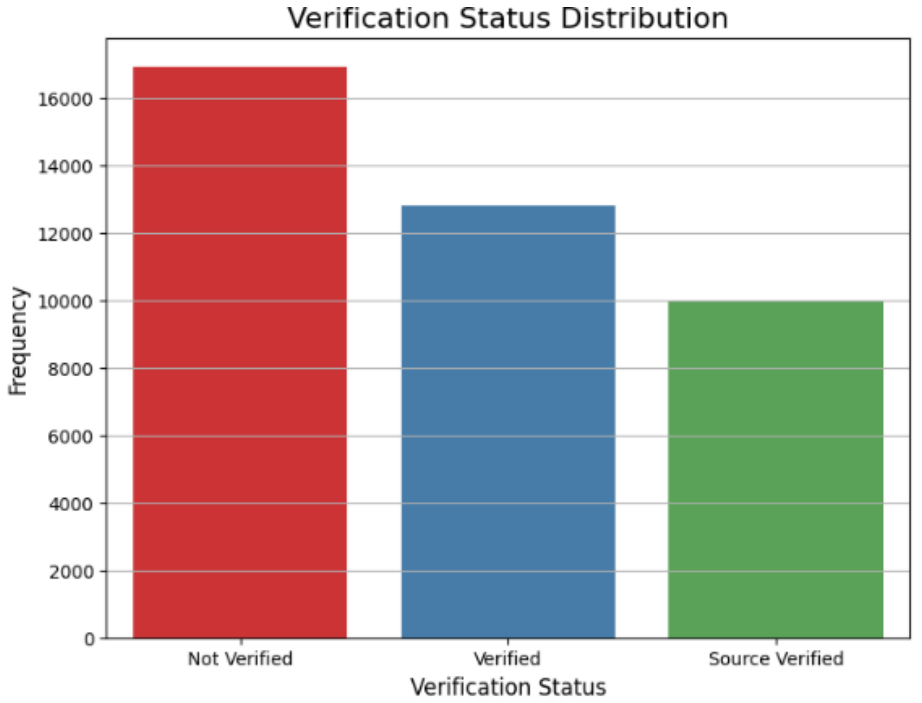
- Those loans that were charged off have higher percentage of loans of grade D, E, F & G as compared to the ones that were paid off

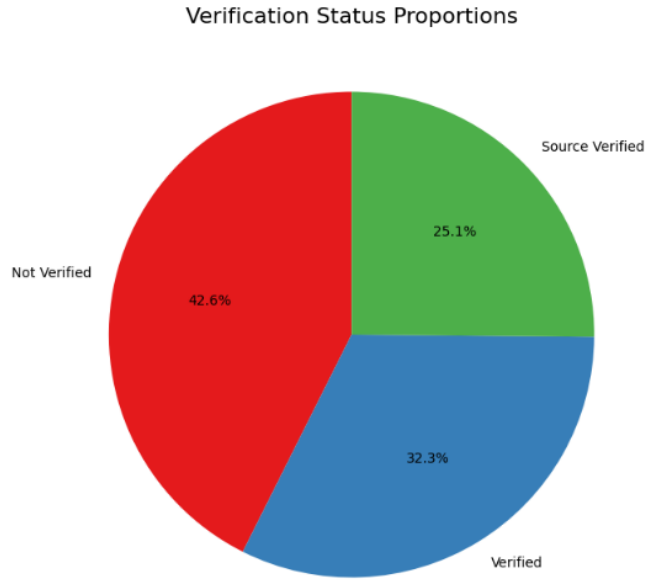




Loan Verification Status Distribution

- 42.6% which is a significant proportion of loans are Not Verified. It may pose a higher risk.





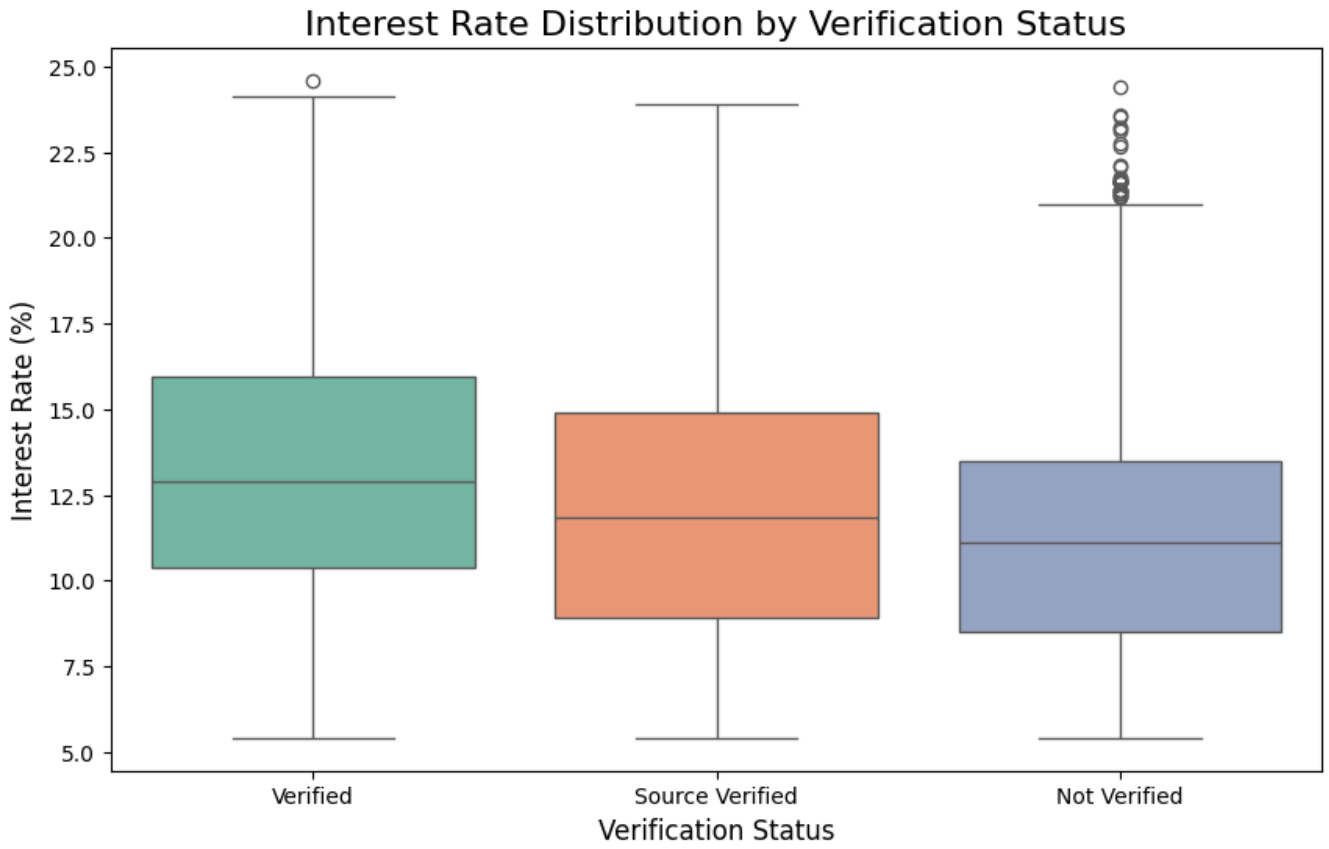
Segmented Univariate Analysis

Summary of Insights

- Income Verification: Verified borrowers should receive favorable terms, while unverified borrowers require additional scrutiny.
- Shorter Loan Terms (36 months): Lower risk and more appealing to borrowers with competitive rates.
- DTI and Loan Grades: Strong correlation between higher DTIs, lower grades, and default risk. Refine approval and pricing policies accordingly.
- Homeownership Stability: Borrowers with mortgages or ownership demonstrate better financial stability and repayment potential.
- Geographic Risk Factors: High-default states (e.g., NE, NV) demand stricter lending policies; low-risk states (e.g., IA, IN) present growth opportunities.
- Loan Purpose Risk: High-risk purposes like debt consolidation require stricter thresholds, while lower-risk purposes like vacation can have more competitive rates.

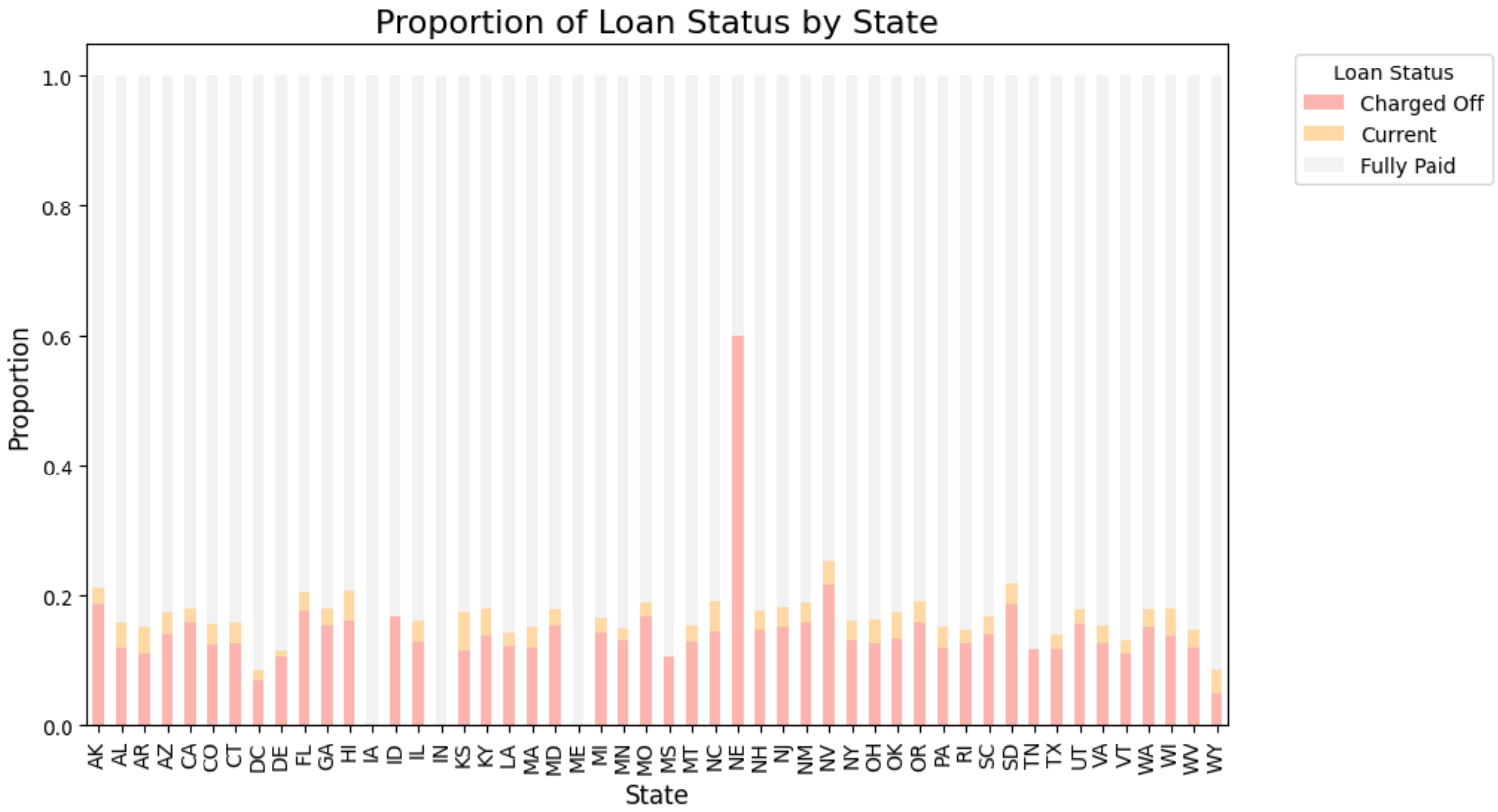
int_rate vs verification_status

- Borrowers with income **not verified** have the least interest rate and those with their loans verified have the highest interest rates.



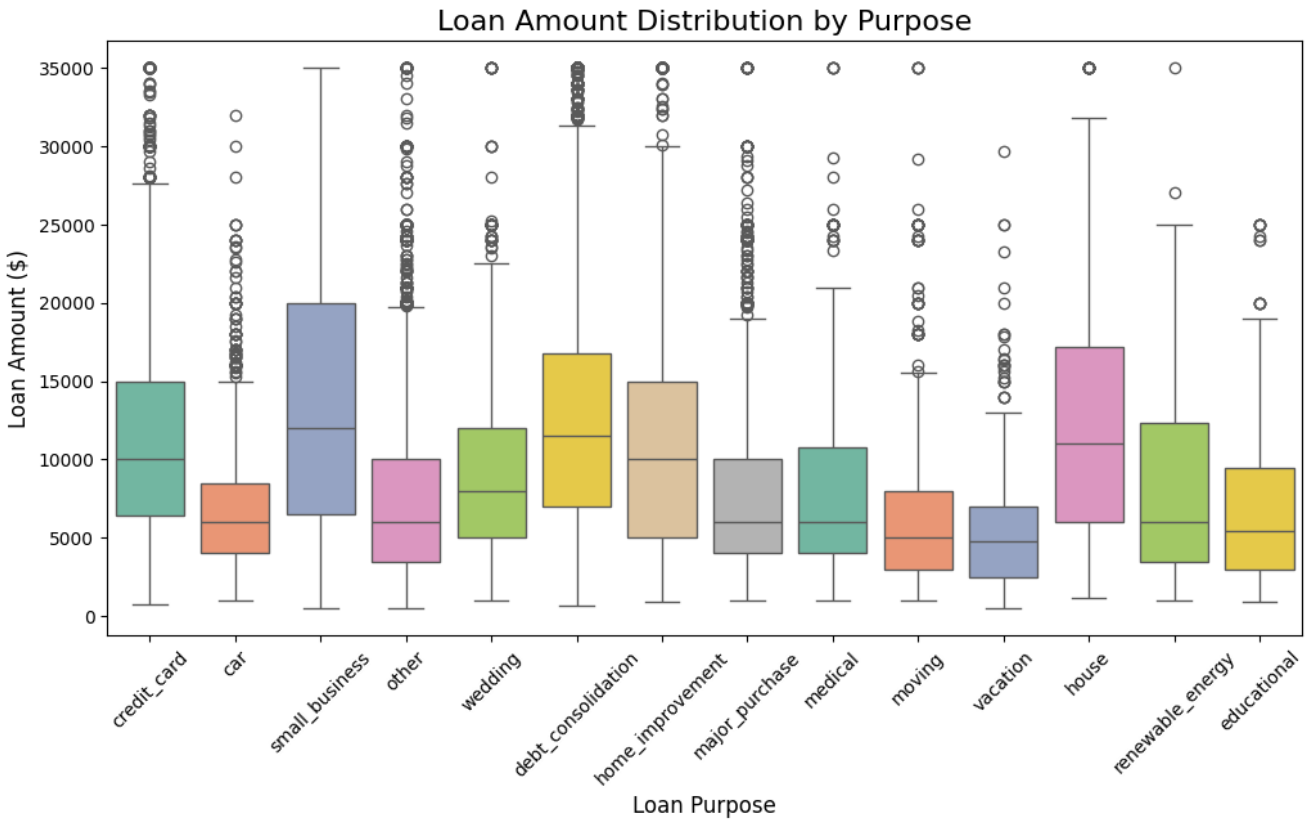
addr_state vs loan_status

- The state NE has the highest number of defaults, followed by NV, indicating higher default risk in this area.
- States like IA, IN, ME have highest percentage of fully paid loans and can be considered as low risk regions.



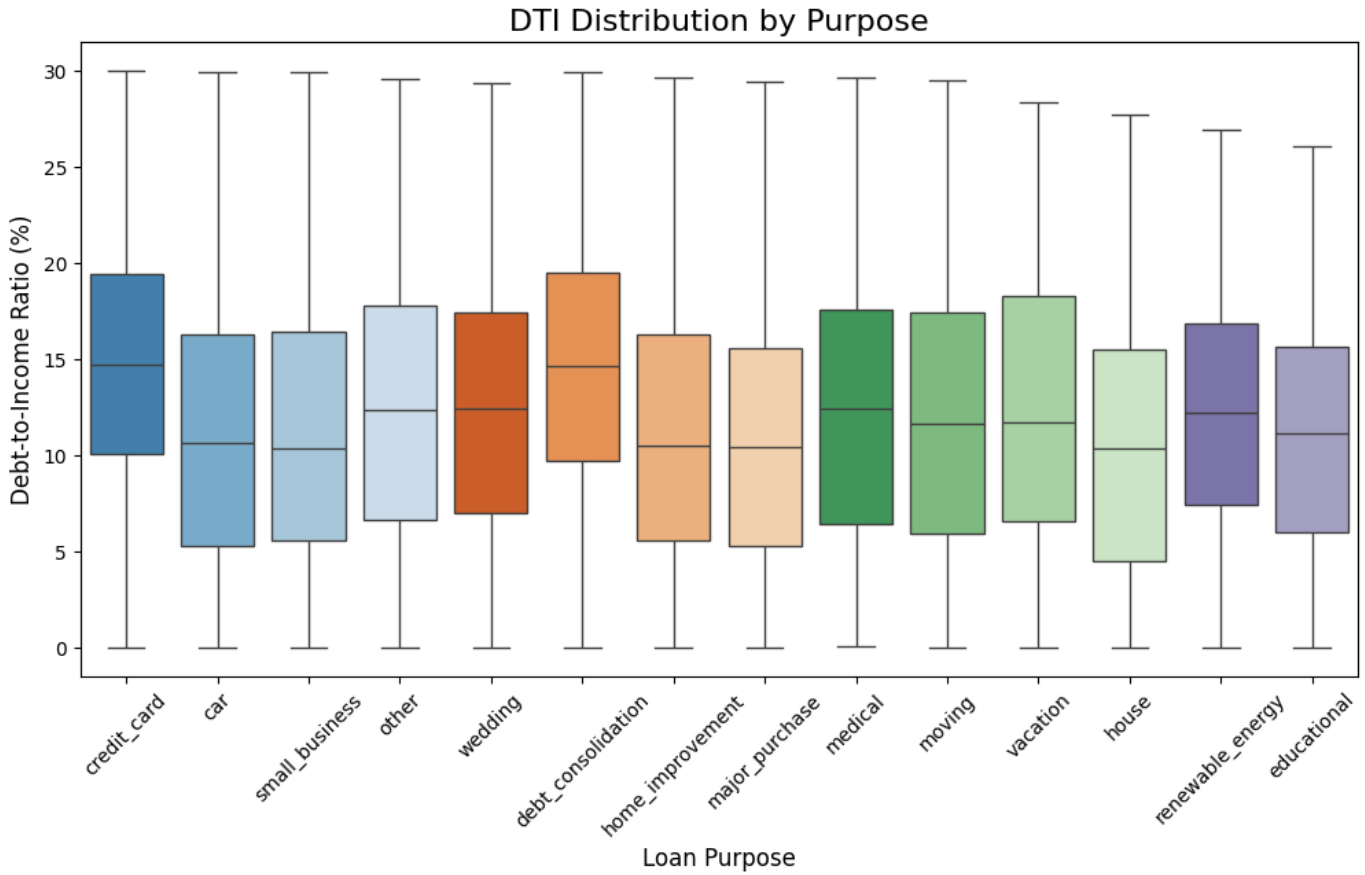
purpose vs loan_amnt

- Debt Consolidation have higher median loan amounts as borrowers consolidate multiple debts.
- Vacation purpose have lower loan amounts due to a comparatively smaller expense need.



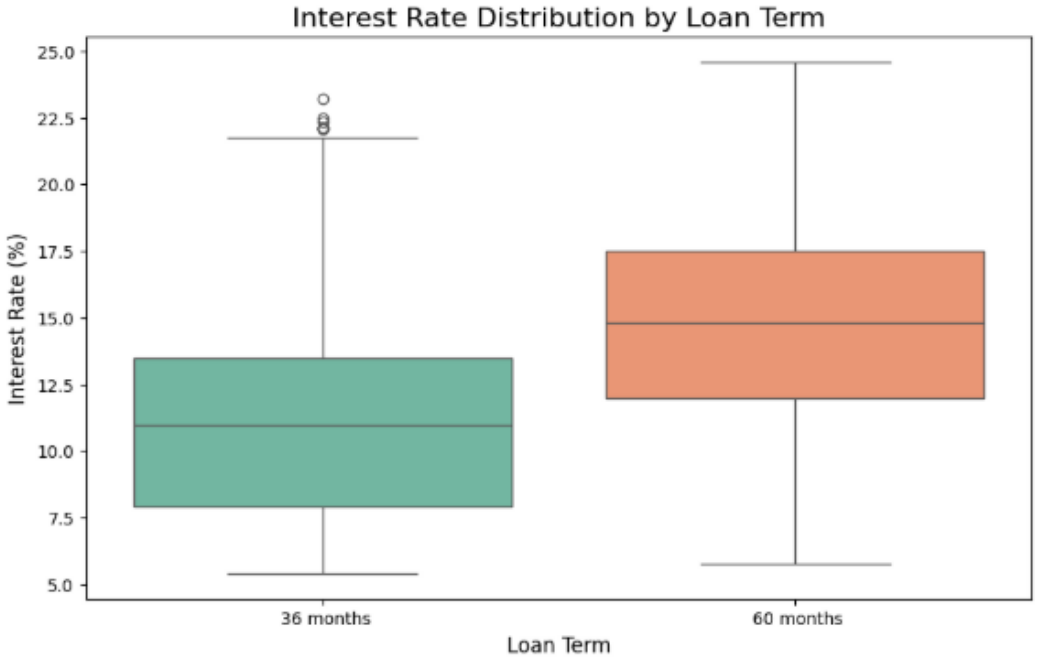
purpose vs dti

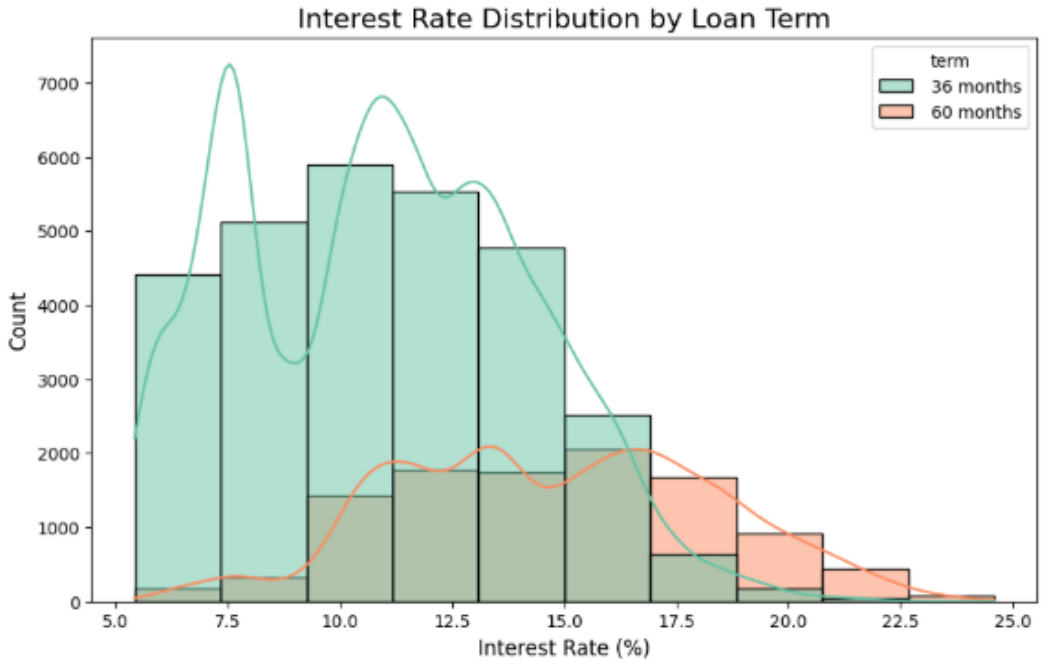
- Borrowers with Credit Card and Debt Consolidation purposes often have higher dti, reflecting existing financial strain.



int_rate vs loan_term

- Longer loan terms (60 months) are associated with higher interest rates due to the increased risk over a longer repayment period.
- Shorter loan terms (36 months) tend to have lower interest rates and less variability.





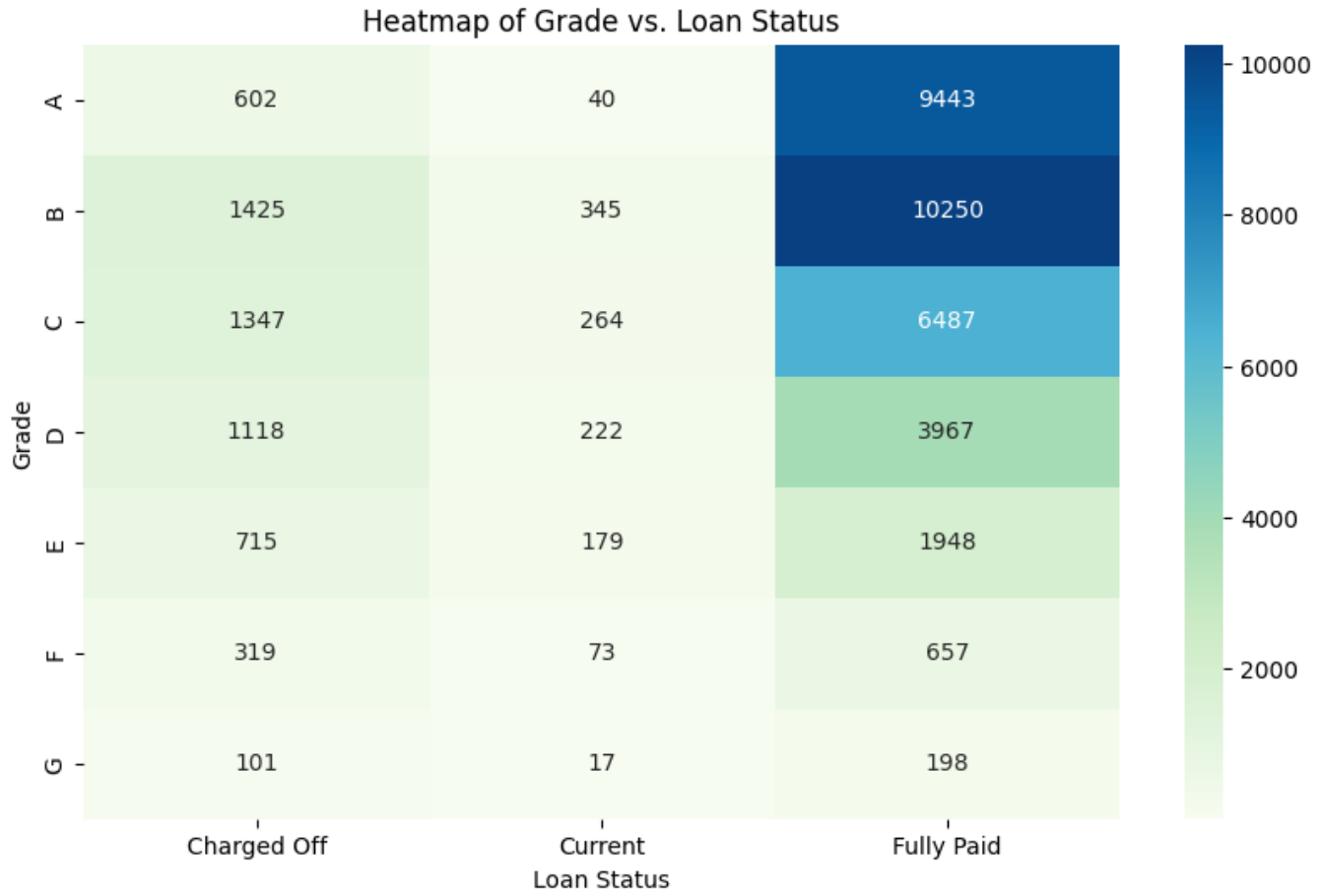
Bivariate Analysis

Summary of Insights

- Interest Rate and Default Risk: Higher interest rates correlate with repayment difficulties, making them a key variable for identifying high-risk borrowers.
- Loan Grades and Risk: Lower-grade loans exhibit higher default risks, requiring stricter criteria and dynamic pricing strategies.
- Loan Purpose and Risk: Debt consolidation loans are riskier, warranting higher interest rates and closer monitoring.
- Macroeconomic Influences: Interest rate trends reflect broader economic conditions, underscoring the need for dynamic adjustment based on external factors.

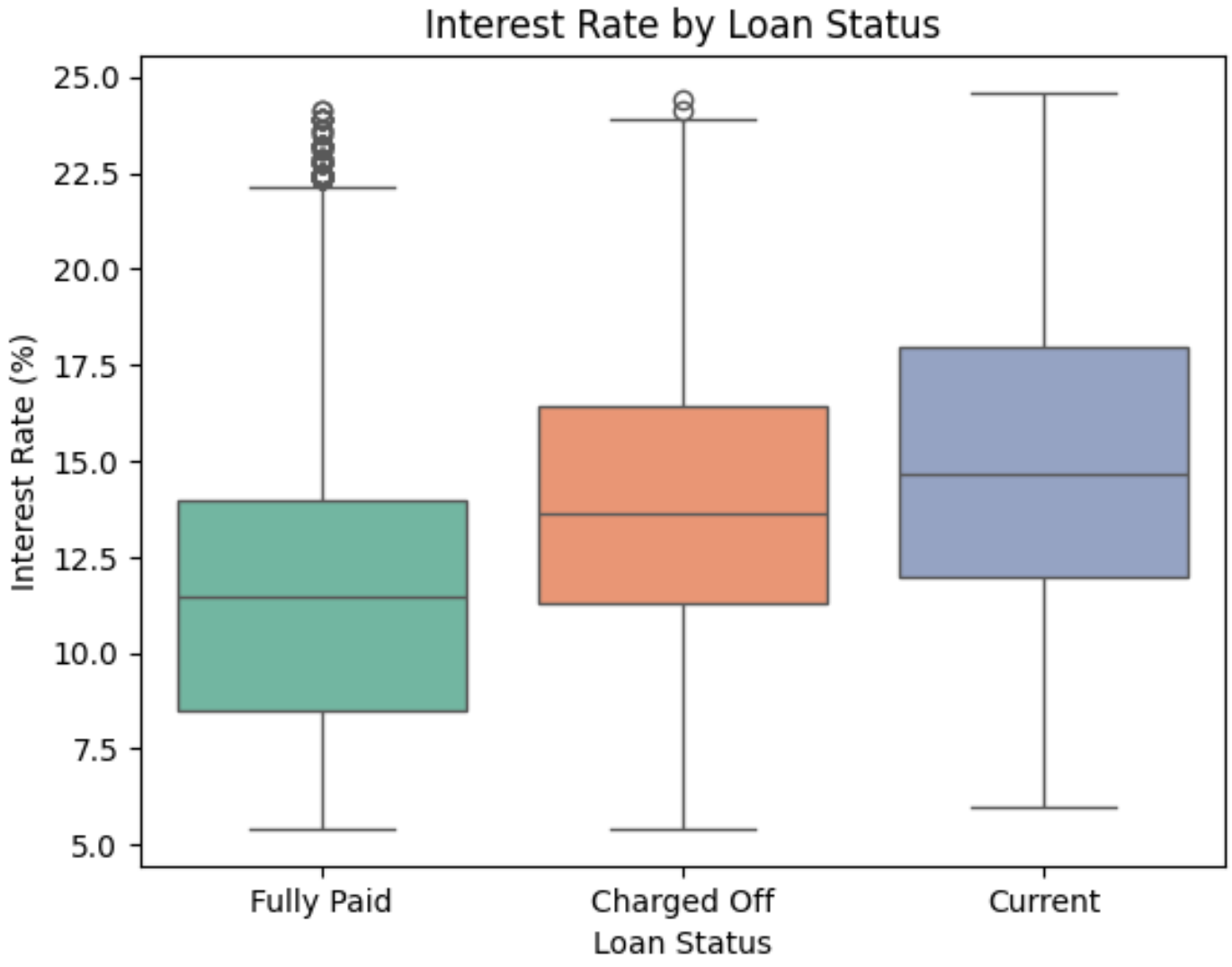
grade vs loan_status

- Fully paid loans have a greater number of higher grades (A, B, C) whereas Charged off loans have a larger number of lower grades (D, E, F, G).
- This indicates that there's a strong correlation between grades and loan status, lower the grade, higher are the chances of default.



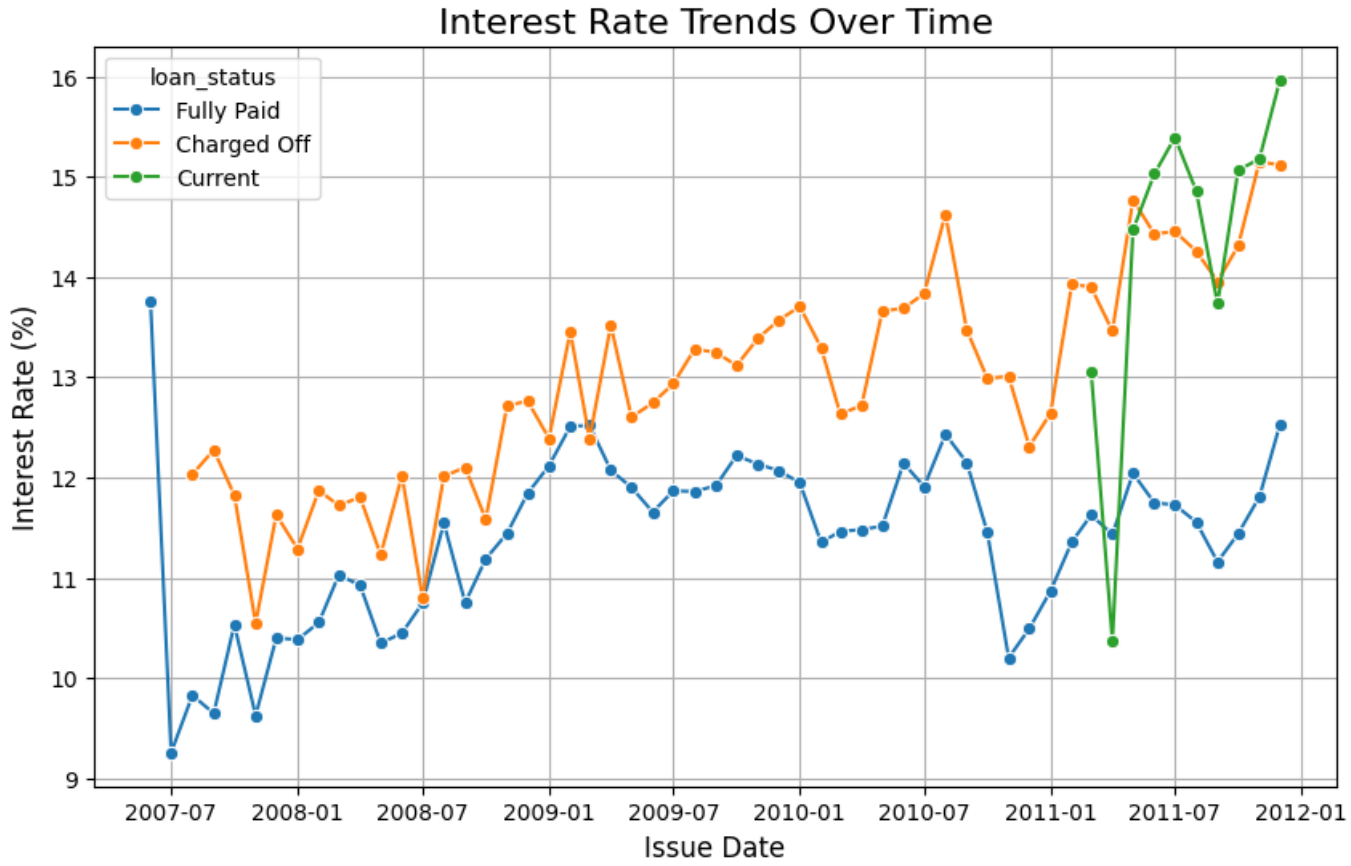
int_rate vs loan_status

- Higher median interest rate for Charged Off loans might indicate that high interest rates contribute to repayment difficulties.



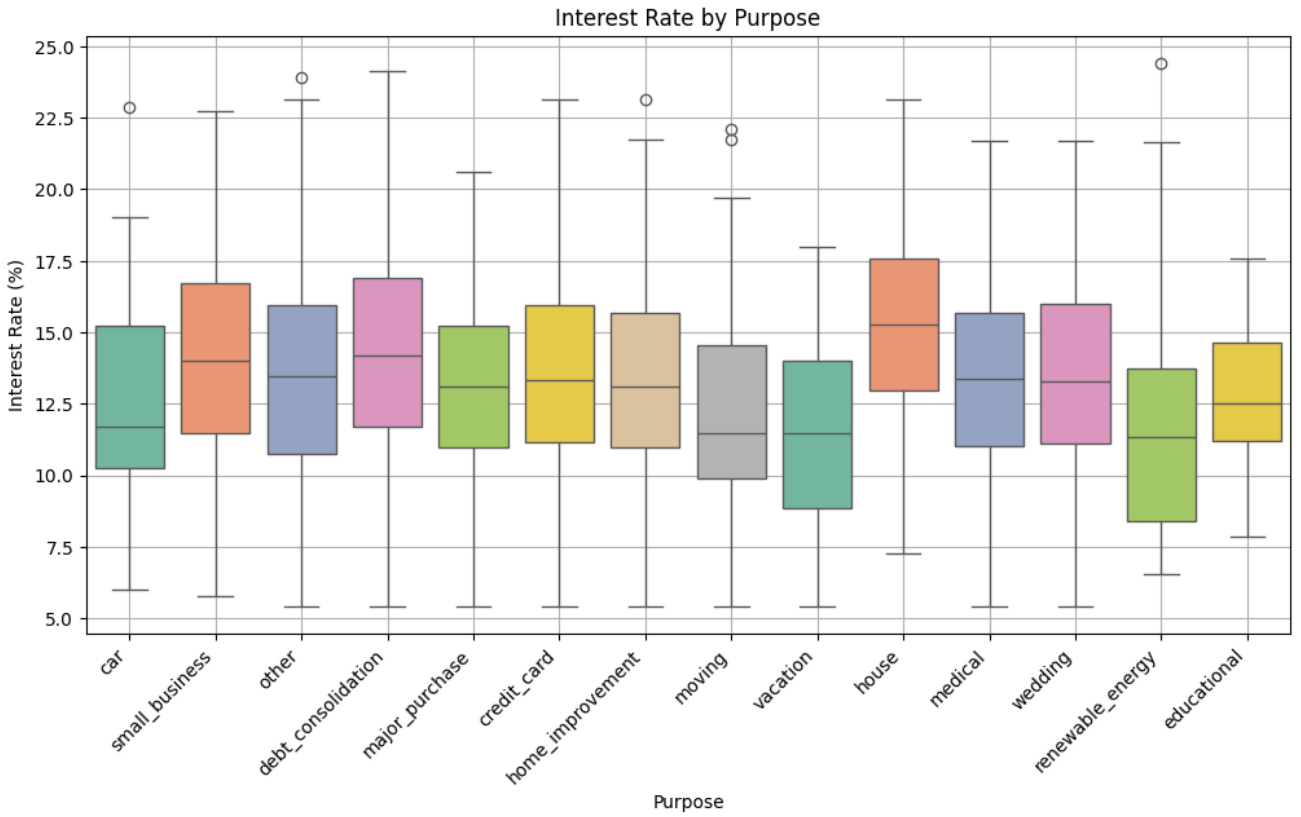
int_rate vs issue_d

- The interest rate trends closely mirror the economic turbulence during and after the 2008 financial crisis.
- The lender responded dynamically to changes in borrower risk profiles, economic conditions, and Federal Reserve policies.



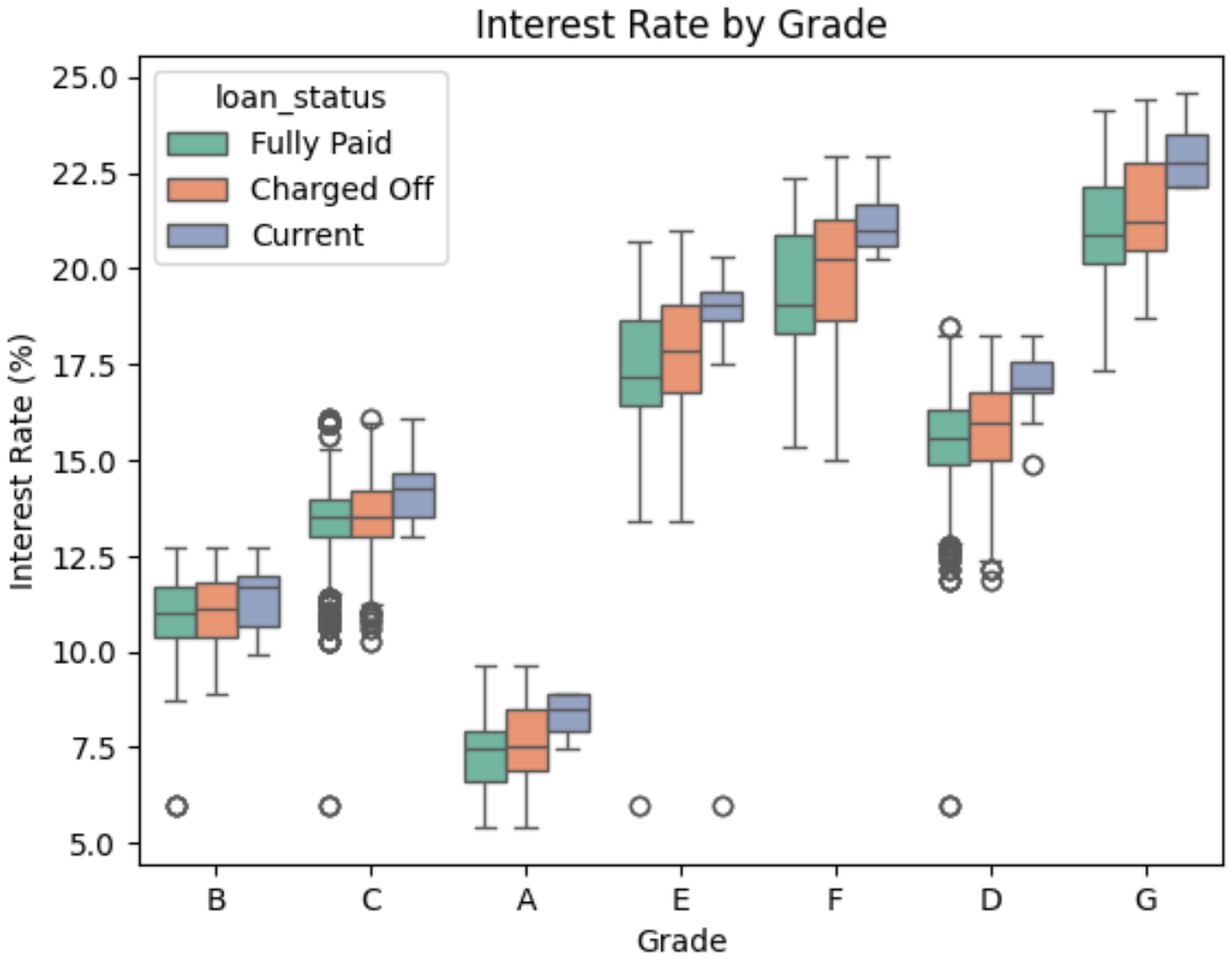
purpose vs int_rate

- Loans taken for house purpose have the highest interest rate. However, earlier analysis shows that loans taken for debt_consolidation are more prone to default
- The interest rate for loans taken for debt_consolidation should be increased and those for house can be reduced.



int_rate vs grade

- Lower grade (D, E, F, G) loans have the highest interest rate due to increased risk of lending



Recommendations

Understanding Loan Default Patterns and Risk Factors

Objective:

Identify patterns in consumer and loan attributes that indicate a higher likelihood of loan default and evaluate variables associated with loan default.

Recommendations:

1. **Debt-to-Income Ratio (DTI)**: Flag borrowers with DTI > 20 as medium risk and apply stricter approval criteria for such applicants to reduce defaults.
2. **Loan Grades**: Focus on lower-grade loans (D, E, F, G) as they have a higher likelihood of default. Implement stricter lending criteria and monitor these loans closely.
3. **Loan Purpose**: Monitor loans for Debt Consolidation and Credit Card purposes, which show higher default rates. Apply stricter DTI thresholds and track performance closely.
4. **Interest Rates**: High-interest loans (e.g., >17%) are associated with defaults. Monitor these accounts more rigorously and evaluate if high rates are contributing to repayment difficulties.
5. **Employment and Income**: Scrutinize borrowers with less than 1 year of employment and annual incomes below $40,000, as they may have limited repayment capacity.
6. **Public Records**: Borrowers with public records or bankruptcies should be flagged for additional risk assessment and closer monitoring.
7. **Credit Utilization (revol_util)**: Flag borrowers with utilization rates >80% as high risk and consider adjusting interest rates upward to mitigate credit loss.

Optimizing Loan Approval Decisions

Objective:

Address risks of rejecting reliable borrowers and approving loans for risky applicants while making data-driven loan decisions.

Recommendations:

- **Term and Interest Rates**:
  - Offer lower interest rates for shorter-term loans (36 months) to attract low-risk borrowers.
  - Educate borrowers about the increased cost of longer-term loans (60 months) during the approval process to encourage shorter-term borrowing.
- **Grade and DTI**:
  - Approve loans for lower grades (D, E, F, G) only if DTI is within acceptable limits (e.g., <9).
- **Verification Status**:
  - Apply stricter thresholds for Not Verified borrowers to mitigate default risks. Increase interest rates for non verified borrowers to reduce the risk.
- **Homeownership**:
  - Offer favorable terms (e.g., higher loan amounts or lower interest rates) to borrowers in MORTGAGE or OWN categories due to better financial stability.
- **Loan Purpose**:
  - Tailor loan products for high-demand purposes like Debt Consolidation and Credit Card refinancing, while applying stricter criteria for high-risk purposes.

Reducing Financial Risk and Enhancing Profitability

Objective:

Provide actionable recommendations to minimize financial risk while maintaining profitability.

Recommendations:

- **Risk-Based Pricing**:
  - Use credit utilization, grades, and DTI to set interest rates dynamically. Charge higher rates for borrowers with high utilization (>80%), lower grades, or DTI >20.
- **Geographic Targeting**:
  - Apply stricter approval criteria and higher interest rates in high-risk states like NE and NV while focusing marketing efforts on low-risk states like IA, IN, and ME.
- **Loan Amount Caps**:
  - Cap loan amounts for high-risk borrowers (e.g., those with high DTI or low grades) to limit exposure.
- **Instalment Design**:
  - Align repayment plans with popular brackets ($200 to $400) for borrower convenience while maintaining profitability.
- **Historical Trends**:
  - Use historical patterns (e.g., trends from int_rate vs issue_d) to anticipate economic impacts on borrower behavior and adjust policies accordingly.
- **Public Records**:
  - Borrowers with public records or bankruptcies pose significant risks. Set stricter approval criteria for these applicants to avoid losses.

Appendix

Detailed Insights from Univariate Analysis

Variable	Insights	Applications
dti Debt to income ratio	Most borrowers have acceptable values of dti and are a low risk.	Few borrowers have dti > 20 and they can be categorized as medium risk. Stricter policies need to be in place for borrowers having > 20 dti to reduce the default risk.
mths_since_last_delinq The number of months since borrower's last delinquency	The median for the months since last delinquencies is 34 months which is a low risk	Few borrowers have mths_since_last_delinq < 12 which can be at a high risk of default
annual_inc	Common income brackets are 40000 to 70000	Develop loan products tailored to common income brackets (e.g., 40,000–70,000).
revol_util The revolving credit utilization rate	The average typical credit utilization is 49.3% with variability (std) in the credit utilization as 28%.	There are a few borrowers with high utilization rates (>80%) and they should be flagged for additional monitoring. Use revol_util as a factor in setting interest rates. Higher utilization rates may warrant higher interest rates.
grade	Those loans that were charged off have higher percentage of loans of grade D, E, F & G as compared to the ones that were paid off	Focus monitoring efforts on lower-grade loans (D, E, F, G) to mitigate default risks.
revol_bal The revol_bal variable represents the total amount of revolving credit balance the borrower owes.	The median revolving balance is 8850 whereas mean is 13382. This suggests that it's a right skewed distribution with presence of outliers	There are borrowers with high revolving balances, and they can be flagged for additional evaluation or stricter approval criteria. Use revol_bal to adjust interest rates, with higher balances indicating increased risk and justifying higher rates.
open_acc Number of open credit accounts of the borrower	On an average, there are 9 credit accounts per user with a min of 2. Borrowers with fewer accounts (e.g., <5) may indicate limited credit history, possibly higher risk.	Borrowers with too many open accounts (>20) or very few (<3) may be flagged for further scrutiny.
home_ownership	Borrowers from RENT and MORTGAGE category have formed the majority	Borrowers in the RENT category may be flagged for additional risk evaluation, given the potential lack of asset backing. Adjust approval criteria or interest rates based on homeownership status.

Variable	Insights	Applications
zip_code	Customers from certain localities whose zip codes start with 663, 94, 999, 669, 385, 373, 833, 689 have never paid loans	1. Adjust the lending policies for the regions with zip codes that have never paid loans or have high default rate. 2. Assess geographic concentration in the portfolio to diversify risk across regions.
addr_state	States like CA, NY, FL and TX dominate the dataset, reflecting population density or lending activity.	1.Focus marketing efforts in states with high borrower density to capture opportunities. 2.Assess the geographic concentration of the portfolio to reduce regional risk exposure. 3.Compare state-level borrower profiles or economic factors to tailor loan offerings and terms.
loan_amnt	Most users prefer to take loan of 10000, possibly influenced by lender policies or borrower needs. The distribution plot of the loan amount is right-skewed which indicates that most borrowers request smaller loans, with a few requesting much larger amounts.	Tailor loan products that focus on popular loan amounts
int_rate	The median interest rate is 11.86%	Few loan accounts have high interest rates (> 17) and these need closer monitoring, as they often indicate higher borrower risk.
term	A higher frequency of 36 months (73.3%) loans suggests borrowers prefer shorter repayment periods to minimize interest costs.	Design marketing campaigns targeting borrowers with short term preferences.
verification_status	42.6% which is a significant proportion of loans are Not Verified. It may pose a higher risk.	Focus monitoring efforts on Not Verified loans, as they may carry higher default risk.
purpose	The categories like debt_consolidation 46.9% and credit card 12.9% dominate. Suprizingly, house category is less frequent.	Develop tailored products for high-demand purposes (e.g., Debt Consolidation and Credit Card refinancing).
installment	Mborrowers have moderate monthly payments (centering around $280), with a few having significantly higher installments.	Design repayment plans that align with popular installment brackets ($200 to $400) for borrower convenience.

Detailed Insights from Segmented Univariate Analysis

Variables	Insights	Applications
term vs int_rate	1. Longer loan terms (60 months) are associated with higher interest rates due to the increased risk over a longer repayment period. 2. Shorter loan terms (36 months) tend to have lower interest rates and less variability.	1. Offer lower interest rates for shorter terms to attract low-risk borrowers. 2. Highlight the increased cost of longer-term loans to customers during the approval process to encourage shorter-term borrowing when appropriate.
grade vs dti	1. Highet DTIs (> 9) correlate to lower credit grades (e.g. C, D, etc) 2. Lower DTIs are associated with higher grades (A, B) indicating better repayment capacity	1. Use dti thresholds to refine approval guidelines for each grade. For example: Approve loans for grades C or below only if dti is within acceptable limits. 2. Adjust interest rates based on the combined impact of grade and dti. Borrowers with lower grades and high dti may warrant higher interest rates. 3. Flag applications with high dti and lower grades for closer review before approval.
home_ownership vs annual_inc	1. Borrowers with MORTGAGE have the highest median incomes, likely reflecting better financial stability 2. Borrowers with MORTGAGE have the highest median incomes.	1. Use home_ownership as an additional criterion for assessing borrower risk, especially for lower-income renters. 2. Offer more favorable terms (e.g., lower interest rates or higher loan amounts) to borrowers in the OWN and MORTGAGE categories with higher incomes.
verification_status vs int_rate	1. Borrowers with income not verified have the least interest rate and those with their loans verified have the highest interest rates. It should have been the other way around.	1. Offer lower interest rates to borrowers with verified income, encouraging them to apply and boosting approval rates for lower-risk individuals. 2. Set stricter thresholds or additional requirements for Not Verified applicants to mitigate risk while maintaining profitability. 3. Encourage income verification to move borrowers into lower-risk categories and justify lower interest rates.

Variables	Insights	Applications
addr_state vs loan_status	1. The state NE has the highest number of defaults, followed by NV, indicating higher default risk in this area. 2. States like IA, IN, ME have highest percentage of fully paid loans and can be considered as low risk regions.	1. Risk-Based Loan Approvals: Apply stricter approval criteria or adjusted loan terms (e.g., higher interest rates) in high-risk states. 2. Geographic Targeting: Focus marketing and loan offerings in states with higher proportions of Fully Paid loans to capture reliable borrowers. 3. Regional Monitoring: Monitor economic conditions and policies in high-default states to adjust risk models dynamically.
purpose vs loan_amnt and dti	1. Debt Consolidation have higher median loan amounts as borrowers consolidate multiple debts. 2. Vacation purpose have lower loan amounts due to a comparatively smaller expense need. 3. Borrowers with Credit Card and Debt Consolidation purposes often have higher dti, reflecting existing financial strain.	1. Apply stricter DTI thresholds for purposes like Credit Card and Debt Consolidation to minimize risk. 2. Offer more competitive interest rates for loans taken for Vacation and moving purposes, as they may have lower default risk. 3. Monitor loan performance closely for Debt Consolidation loans due to higher loan amounts and potential risk factors.

Detailed Insights from Bivariate Analysis

Variables	Insights	Applications
int_rate vs loan_status	1. Higher median interest rate for Charged Off loans might indicate that high interest rates contribute to repayment difficulties.	1. Use interest rate thresholds to identify high-risk borrowers. 2. Offer competitive interest rates to low-risk borrowers to encourage timely repayments. 3. Monitor loans with high interest rates more closely.
grade vs loan_status	1. Fully paid loans have a greater number of higher grades whereas Charged off loans have a larger number of lower grades. 2. This indicates that there's a strong correlation between grades and loan status, lower the grade, higher are the chances of default.	1. Implement stricter lending criteria for lower grades (D, E, F, G). 2. Focus monitoring efforts on lower grades to reduce default risk. 3. Develop risk-adjusted pricing strategies based on grades.
grade vs int_rate	1. Lower grade loans have the highest interest rate due to increased risk of lending	1. Adjust interest rates dynamically based on grades to balance risk and profitability. 2. Use grade-specific policies to attract reliable borrowers and discourage defaults among lower grades.
purpose vs int_rate	1. Loans taken for house purpose have the highest interest rate. However, earlier analysis shows that loans taken for debt_consolidation are more prone to default	1. Increase interest rates for high-risk purposes like debt consolidation. 2. Offer competitive rates for low-risk purposes like house-related loans. 3. Align loan pricing with default risk levels.
int_rate vs issue_d	1. The interest rate trends closely mirror the economic turbulence during and after the 2008 financial crisis. 2. The lender responded dynamically to changes in borrower risk profiles, economic conditions, and Federal Reserve policies.	1. Monitor economic conditions to adjust interest rates proactively. 2. Use historical patterns to forecast rate adjustments during economic volatility. 3. Align lending policies with macroeconomic trends.

Thank you