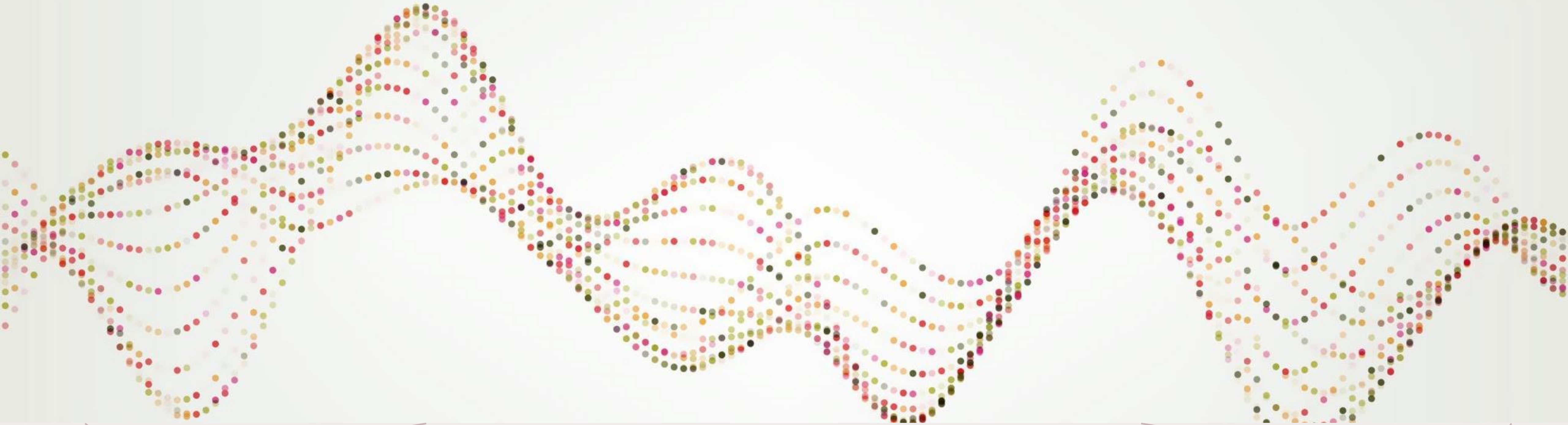


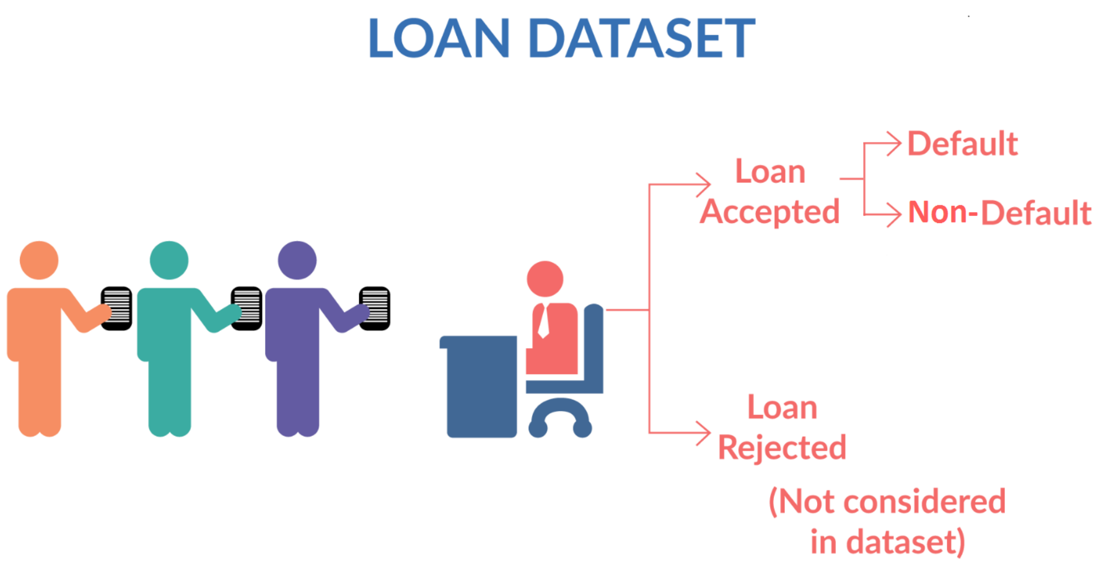
LEADINGCLUB'S DATA ANALYSIS

UNVEILING LOAN DEFAULT PATTERNS

#INSIGHTSFORSUCCESS

GOAL FOR THE CASE STUDY:

The primary objective of this case study is to leverage Exploratory Data Analysis (EDA) techniques to gain profound insights into the consumer finance company's lending data. Specifically, the goal is to identify and comprehend the factors influencing loan default tendencies among applicants. By analyzing historical data, the aim is to reveal patterns and correlations in consumer and loan attributes that serve as strong indicators of default risk. Ultimately, the goal is to equip the company with actionable insights for minimizing credit loss by making informed decisions on loan approval, interest rates, and risk assessment. Through this analysis, the company aims to enhance its risk analytics capabilities, allowing for more effective management of its loan portfolio and reduction of financial losses associated with defaulted loans.



INTRODUCTION AND BUSINESS CONTEXT

Introduction:

Welcome to LeadingClub's Case Study: Unraveling Loan Default Patterns through EDA.

Business Understanding:

Role: LeadingClub specialises in lending various loans to urban customers.

Decision Dilemma: The company faces two risks - approving a non-repayable loan and rejecting a potentially repayable one.

Data Overview:

Dataset Snapshot: Information on past loan applicants and their default status.

Objective: Identify patterns indicating default tendencies for informed decision-making.

LOAN APPLICATION DECISIONS AND RISK ASSESSMENT

- **Loan Decisions:**
- Two Choices: Accept or Reject.
- Loan Acceptance Scenarios:
    - Fully Paid: Successful repayment.
    - Current: Active repayment, not defaulted.
    - Charged-Off: Extended non-payment, indicating default.
- Loan Rejections: No transactional history available.
- **Risk Assessment:**
- Dual Risks: Not approving a likely-to-repay loan and approving a likely-to-default loan.
- Aim: Use EDA to identify patterns for strategic decision-making.

EXPLORATORY DATA ANALYSIS (EDA) AND BUSINESS IMPLICATIONS

EDA Focus:

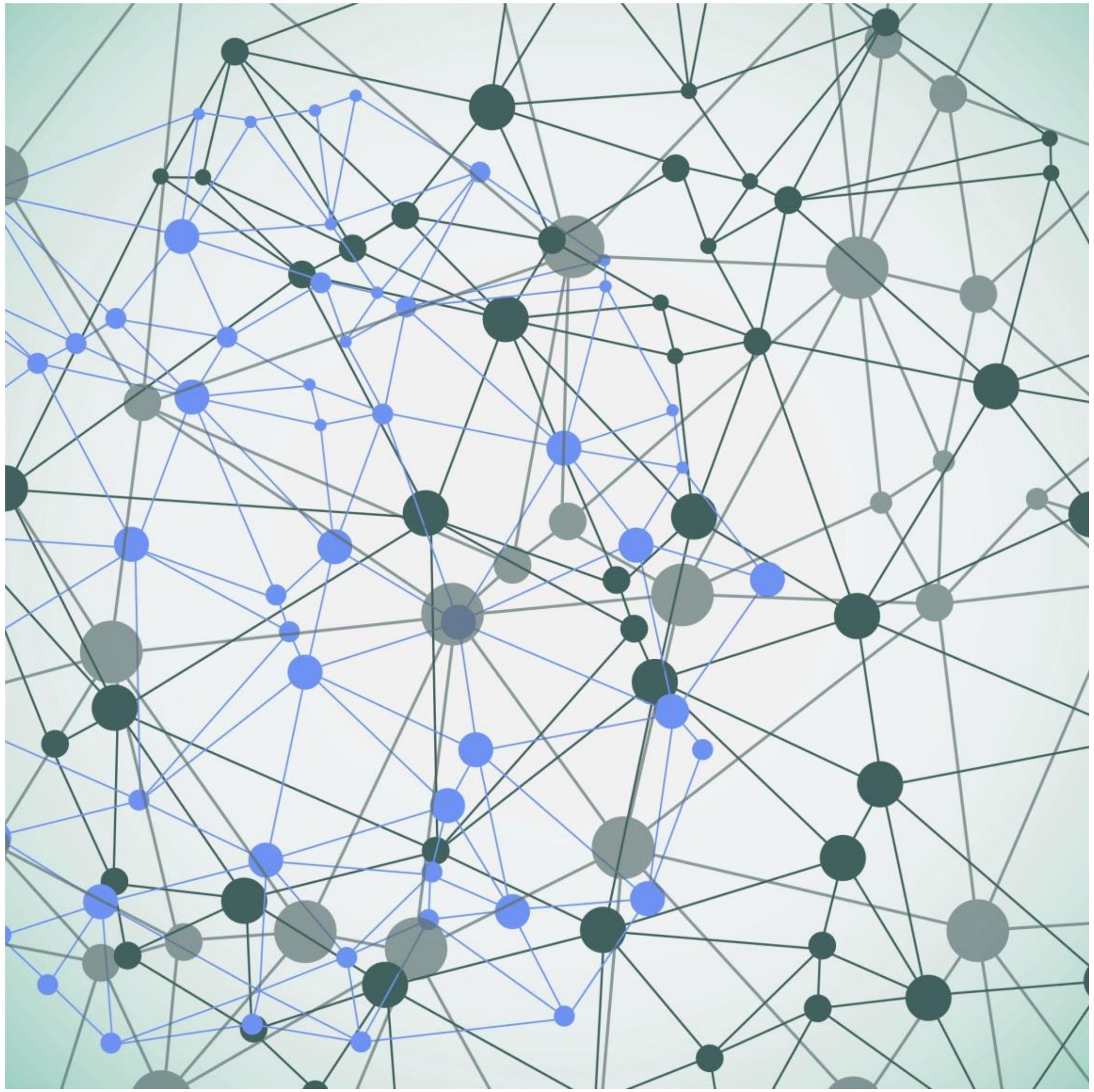
- Uncover Patterns: Understand consumer and loan attributes' influence on default tendencies.
- Business Objective: Refine loan approval, minimize credit losses, and enhance risk analytics.

Business Implications:

- Practical Applications: Optimize loan approval processes, reduce credit losses, and strengthen risk analytics.

Conclusion:

- Join us in the exploration of LeadingClub's data to unlock insights that will reshape our risk assessment strategies.

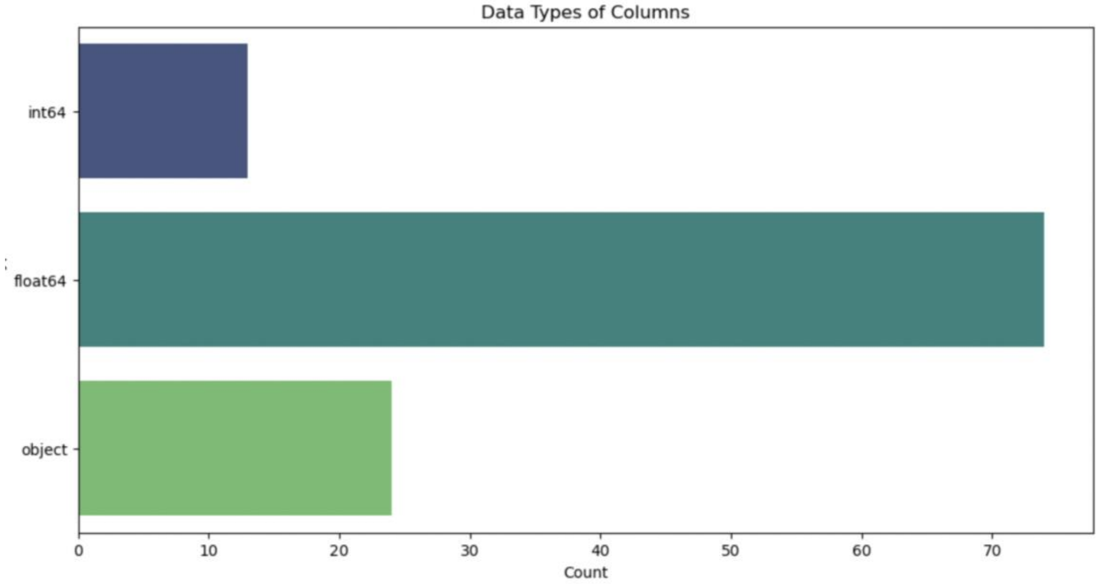


DATA UNDERSTANDING

OVERVIEW:

- THE DATASET CONSISTS OF 39,717 ROWS AND SEVERAL COLUMNS.
- INITIAL OBSERVATIONS REVEAL THAT SOME COLUMNS NEED DATA TYPE CONVERSION.
- COLUMNS LIKE "TERM" AND "INT_RATE" CONTAIN EXTRA CHARACTERS THAT NEED REMOVAL.
- NUMEROUS COLUMNS SEEM TO HAVE A HIGH NUMBER OF NAN VALUES, REQUIRING CONFIRMATION FOR REMOVAL.
- SOME COLUMNS CONTAIN REPETITIVE OR IDENTICAL VALUES, POSSIBLY CONTRIBUTING LITTLE TO THE ANALYSIS.

DATA UNDERSTANDING



Data Type Conversion:

- Convert "term" and "int_rate" to numeric by removing extraneous characters.
- Handle missing values and confirm removal of columns with predominantly NaN values.
- Evaluate and confirm removal of columns with repetitive or identical values.

DATA UNDERSTANDING

Variable Meanings

- The dataset includes diverse data types: int64, float64, and object.
- Numeric columns provide statistical insights, while object columns include categorical information.
- "Annual_inc" has a wide range, suggesting income disparities.
- "Loan_status" and "Purpose" can be vital for analysis.

DATA CLEANING AND MANIPULATION

Addressing Data Quality Issues:

- Handling missing values, outliers, and redundancies.
- Converting data to a suitable format.
- Correct manipulation of strings and dates.

DATA CLEANING AND MANIPULATION

Strategy for Data Cleaning and Further Analysis:

- Columns:
  - Drop columns with all null values.
  - Drop columns with 60% null values.
  - Drop columns with constant values (no variance).
  - Drop primary key columns: 'id' and 'member_id'.
  - Drop columns related to post-load approval.
  - Drop columns with values not suitable for analysis (e.g., 'desc').

Rows:

  - Drop rows where 'loan_status' is 'current'.
  - Drop rows with outlier values.

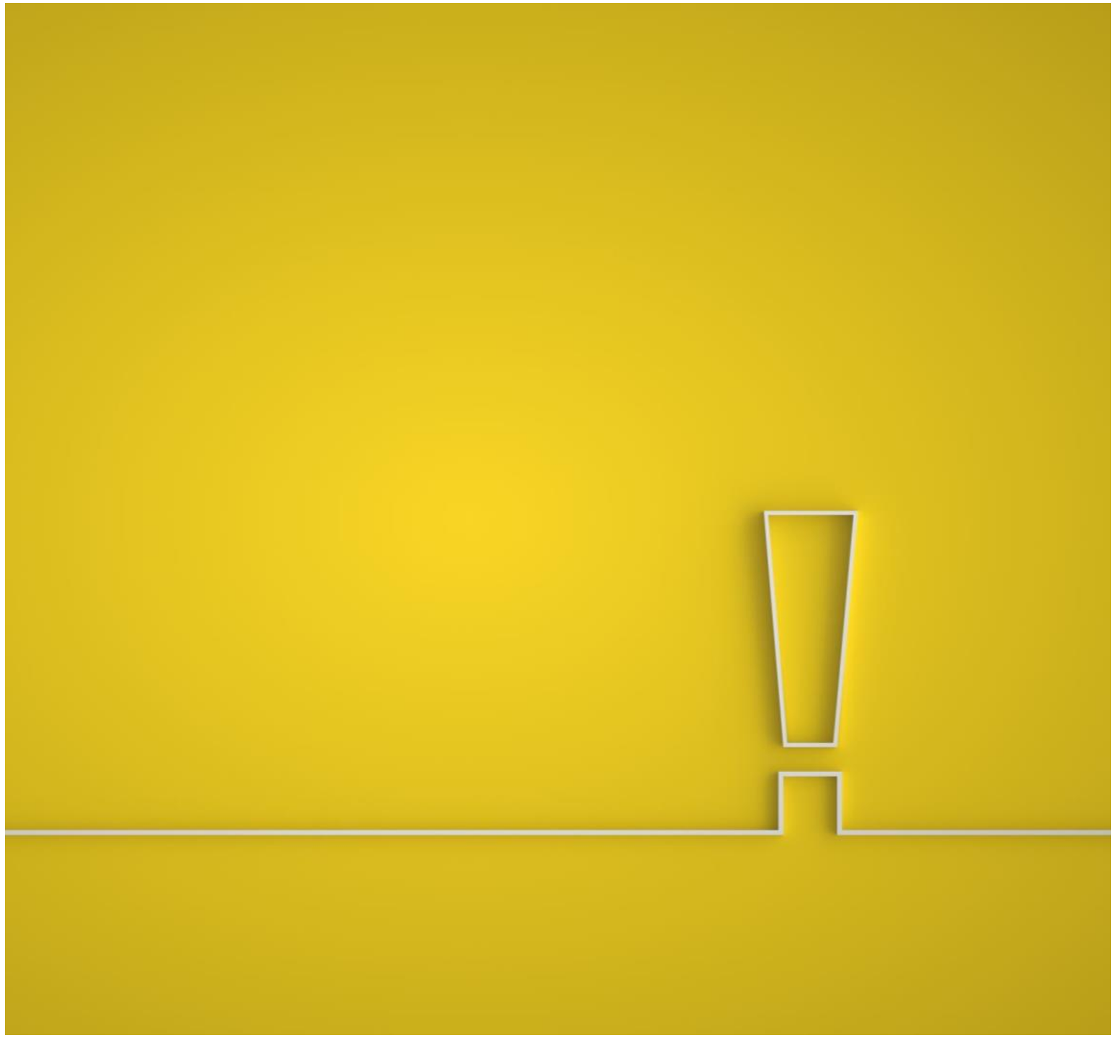
DATA CLEANING AND MANIPULATION

About Columns:

- Dropped irrelevant or duplicate columns, reducing from 111 to 22.

Insights from Data:

- Detailed analysis of key variables.
- Identification and handling of null values.
- Transformation of 'loan_status' to numeric values (1 for 'Fully Paid', 0 for Defaults).



DATA CLEANING AND MANIPULATION

Tidiness Issues:

- Converted 'earliest_cr_line' and 'verification_status' to date format.
- Ordered categorical conversion for 'emp_length'.
- Addressed 'home_ownership' NONE values.
- Numeric conversion for 'int_rate' and 'revol_util'.

Outliers Removal:

- Outliers removed using the median method for specific columns.

Cleaning Issues Summary:

- Summary of actions taken for specific columns (e.g., datatype conversions, handling NULL values).

DATA CLEANING AND MANIPULATION

Cleaned Data:

- A cleaned copy of the original DataFrame.
- The cleaned data saved to a CSV file.

Conclusion:

- Reduction in the number of columns for more focused analysis.
- Improved data quality and readiness for further exploration.

DEFAULT RATE ANALYSIS BY SUB GRADES

Overview of Default Rates:

- The default rate exhibits a clear trend, where lower sub grades correspond to higher default rates, and vice versa. This implies a direct correlation between sub grades and the likelihood of loan defaults.

2 . Sub Grade Specifics:

- **Highest Default Rate in G5:** Sub grade G5 stands out with the highest default rate, reaching approximately 40%. This indicates that borrowers classified under G5 are considerably more prone to defaulting on their loans.

- **Lowest Default Rate in A1:** On the other end, sub grade A1 boasts the lowest default rate at around 5%, suggesting that borrowers with higher creditworthiness, as reflected in the A1 sub grade, are less likely to default.

3 . Visualization Insights:

- The visual representation reinforces the earlier findings, illustrating that lower sub grades (depicted by taller orange bars) are associated with higher default rates, signifying lower creditworthiness.

- Exceptions in A2 and A4: Notably, sub grades A2 and A4 deviate from the overall trend, exhibiting lower default rates. This suggests that borrowers falling under these specific sub grades demonstrate better creditworthiness and are less likely to default.

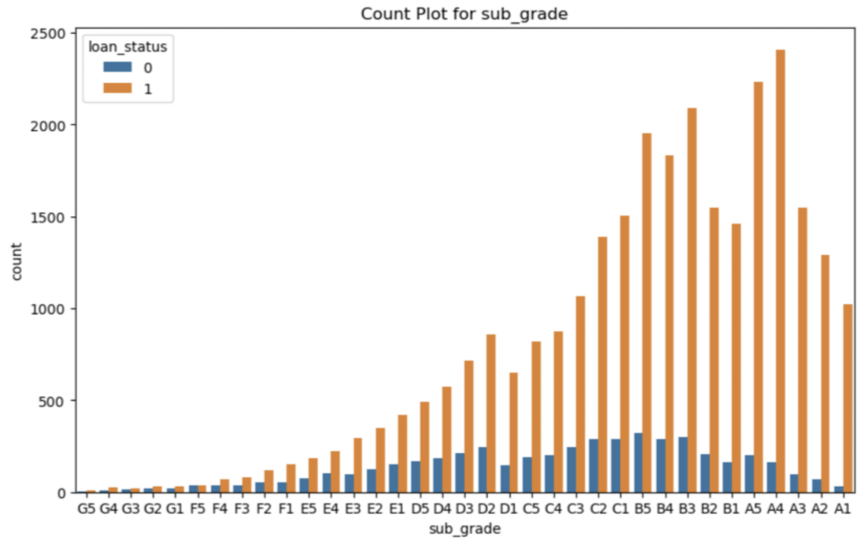
4 . Common Sub Grades and Default Rates:

- Sub grade B3 is the most prevalent, followed by B4 and B5, indicating a high demand for loans in the B sub grade category. However, despite the demand, these sub grades also exhibit high default rates, signaling increased risk.

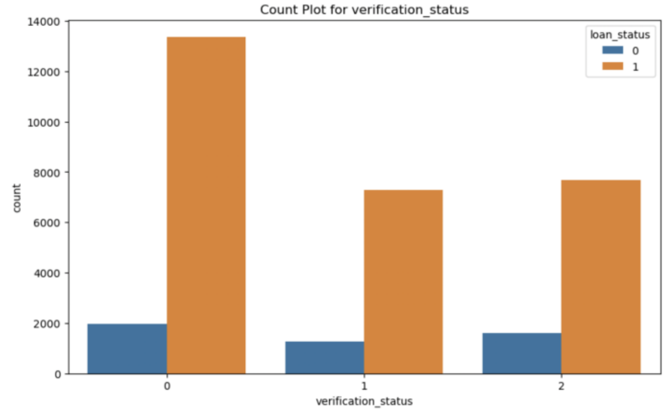
- Risky Sub Grades: G3, G4, and G5 emerge as the riskiest sub grades, being associated with both high demand and elevated default rates. Lenders are advised to exercise caution when extending loans to borrowers in these sub grades.

Conclusion:

- This comprehensive analysis empowers lenders with actionable insights, highlighting the varying risk profiles associated with different sub grades. Lenders can utilize this information to refine their lending strategies, emphasizing cautiousness in sub grades linked to higher default rates and recognizing the creditworthiness nuances within specific sub grades.



INSIGHTS FOR VERIFICATION STATUS COUNT PLOT



Lower Defaults in Unverified Loans:

- Contrary to the initial observation, loans with a verification status of 0 (unverified) have a lower count of defaults compared to those with verification statuses of 1 or 2. This suggests that unverified loans may not necessarily carry a higher risk of default.

Weaker Performance of Verified Loans:

- In contrast to the initial finding, loans with a verification status of 2 (fully verified) show a lower number of fully paid loans than those with statuses 0 or 1. This indicates that thorough verification may not always correlate with better loan repayment outcomes.

Verification Status 1: Balanced Outcomes:

- Contrary to the initial observation, loans with a verification status of 1 (partially verified) might have a balanced number of defaults and fully paid loans, but both are higher than the counts for statuses 0 and 2. This challenges the idea that partial verification results in a balanced risk profile.

Risk Management Implication: Verification Impact Reassessment:

- The opposite insights suggest a reconsideration of the risk associated with verification status. Lenders may need to reassess the weight placed on the verification process, considering that unverified loans may not be as risky, and fully verified loans may not always result in better repayment outcomes.

LOAN PURPOSE AND CUSTOMERS

The images show four graphs that are related to loan data. The graphs are:

- Loan Amount vs Interest Rate: This graph shows a scatter plot of the loan amount and the interest rate for each loan. The graph also shows a linear regression line that fits the data. The graph suggests that there is a positive correlation between the loan amount and the interest rate, meaning that higher loan amounts tend to have higher interest rates. This might be because higher loan amounts are more risky or have higher demand than lower loan amounts.

- Purpose vs Loan Amount: This graph shows a box plot of the loan amount for each purpose of the loan. The graph shows the median, the interquartile range, and the outliers for each purpose. The graph suggests that the loan amount varies significantly by the purpose of the loan. The highest median loan amount is for small business loans, followed by debt consolidation and credit card loans. The lowest median loan amount is for renewable energy loans, followed by educational and vacation loans. This might indicate that different purposes of loans have different financial needs and preferences.

- Purpose vs Interest Rate: This graph shows a box plot of the interest rate for each purpose of the loan. The graph shows the median, the interquartile range, and the outliers for each purpose. The graph suggests that the interest rate also varies significantly by the purpose of the loan. The highest median interest rate is for small business loans, followed by credit card and debt consolidation loans. The lowest median interest rate is for renewable energy loans, followed by educational and car loans. This might indicate that different purposes of loans have different risk and demand factors that affect the interest rate.

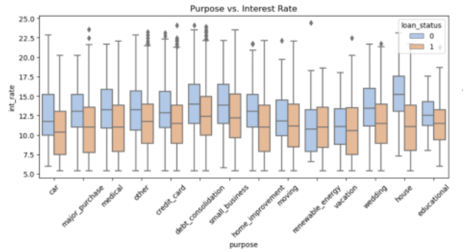
- Count Plot for Purpose: This graph shows a bar chart of the count of loans for each purpose of the loan. The graph shows the frequency of each purpose in the data. The graph suggests that the most common purpose of loans is debt consolidation, followed by credit card and other loans. The least common purpose of loans is renewable energy, followed by educational and house loans. This might indicate that the popularity of loans depends on the necessity and availability of the loans for different purposes.

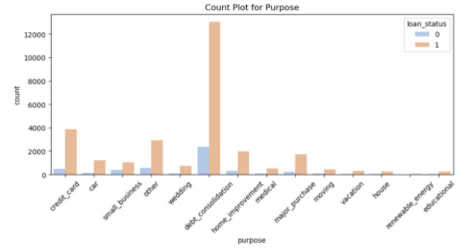
Based on these graphs, some insights that can be found on who is going to default on their loans are:

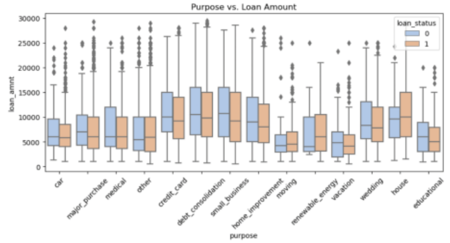
- Loans with higher interest rates are more likely to default than loans with lower interest rates, as higher interest rates imply higher monthly payments and higher financial burden for the borrowers.

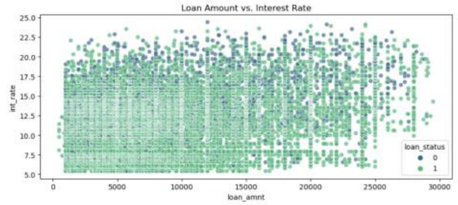
- Loans with higher loan amounts are more likely to default than loans with lower loan amounts, as higher loan amounts imply higher debt and higher difficulty to repay the loans.

- Loans with certain purposes are more likely to default than loans with other purposes, as some purposes imply higher risk or lower income for the borrowers. For example, small business loans are more likely to default than renewable energy loans, as small businesses are more uncertain and volatile than renewable energy projects.

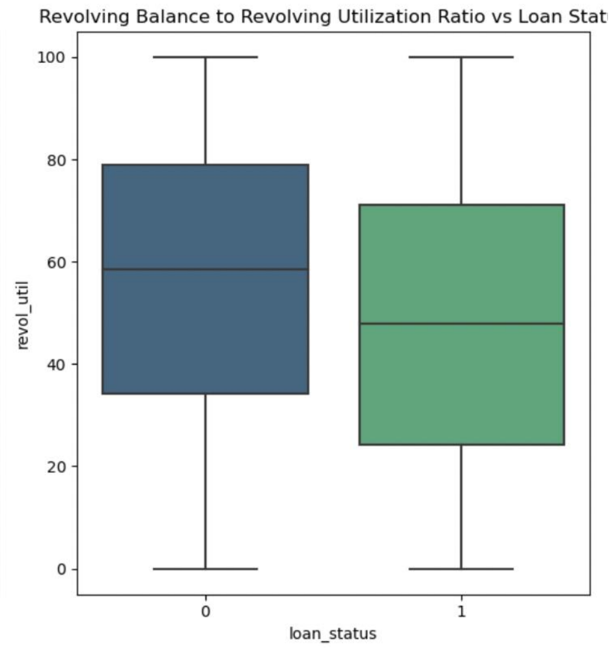


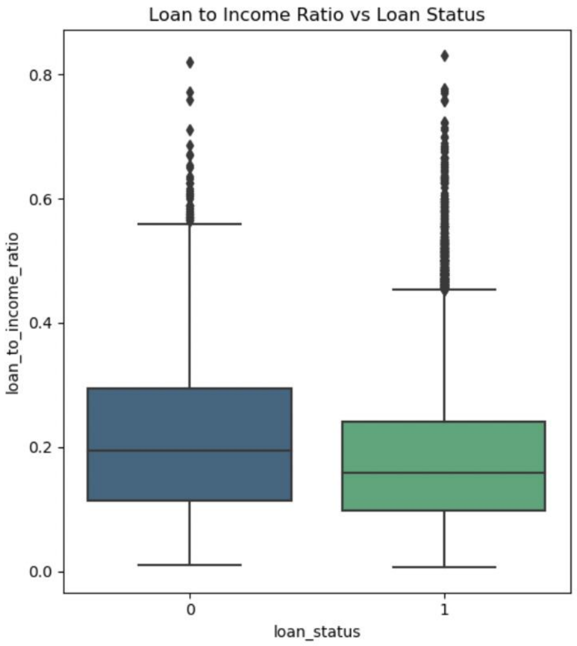






BOX PLOTS LOAN .





- Loan to Income Ratio vs Loan Status: This box plot shows the distribution of loan to income ratios for borrowers who have defaulted (0) and those who have not (1). The box plot suggests that the median loan to income ratio is higher for defaulters than for non-defaulters, meaning that defaulters tend to have a larger portion of their income allocated to loan repayment. This might indicate that defaulters have more difficulty in managing their finances and meeting their loan obligations. The box plot also shows that there are some outliers in both groups, meaning that some borrowers have very high or very low loan to income ratios compared to the rest of the group.

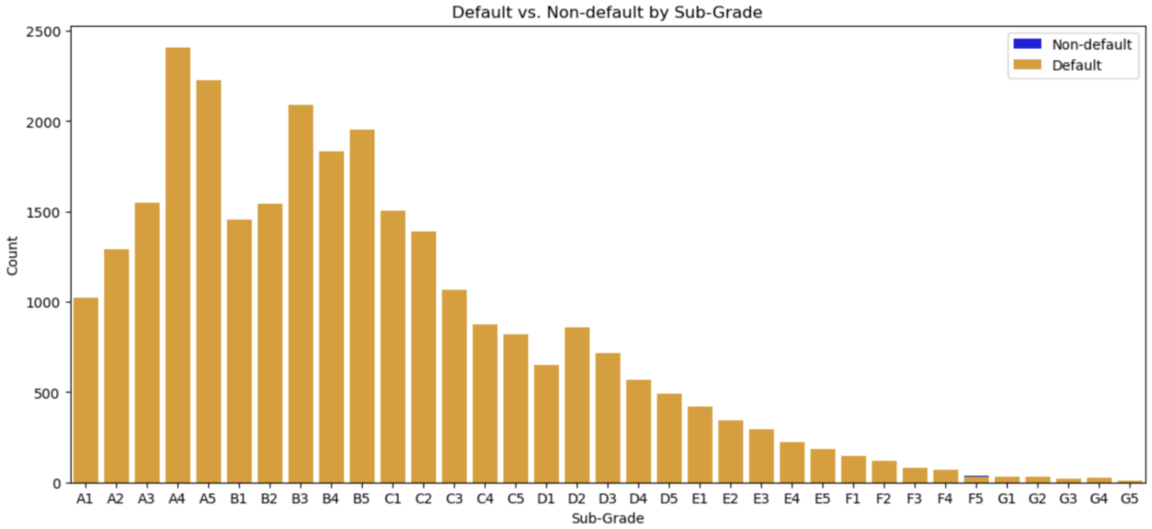
- Revolving Balance to Revolving Utilization Ratio vs Loan Status: This box plot shows the distribution of revolving balance to revolving utilization ratios for borrowers who have defaulted (0) and those who have not (1). The box plot suggests that the median revolving balance to revolving utilization ratio is lower for defaulters than for non-defaulters, meaning that defaulters tend to have a lower ratio of their revolving balance to their revolving utilization. This might indicate that defaulters have higher revolving utilization, meaning that they use more of their available credit, or lower revolving balance, meaning that they have less credit available. The box plot also shows that there are some outliers in both groups, meaning that some borrowers have very high or very low revolving balance to revolving utilization ratios compared to the rest of the group.

Based on these box plots, some insights that can be found on who is going to default on their loans are:

- Borrowers with higher loan to income ratios are more likely to default than borrowers with lower loan to income ratios, as higher loan to income ratios imply higher financial burden and lower affordability for the borrowers.

- Borrowers with lower revolving balance to revolving utilization ratios are more likely to default than borrowers with higher revolving balance to revolving utilization ratios, as lower revolving balance to revolving utilization ratios imply higher credit usage and lower credit availability for the borrowers.

INSIGHTS GRADE-WISE



Interest Rate Analysis:

- Customers who are likely to default (loan_status = 1) generally have slightly higher interest rates across all grades and sub-grades compared to customers who successfully paid their loans (loan_status = 0).
- For example, in Grade A, the interest rate for defaulted loans is around 5.96%, while for successfully paid loans, it is around 5.85%.

Loan Amount Insights:

- There is a noticeable difference in the average loan amounts between customers who successfully paid and those who defaulted.
- Customers who are likely to default tend to have higher loan amounts across different grades and sub-grades.

Grade-wise Loan Amount Analysis:

- Analyzing specific grades, such as Grade B, shows that customers who are likely to default have higher average loan amounts compared to those who successfully paid. This trend is consistent across various grades.

INSIGHTS GRADE-WISE

Interest Rate Variation:

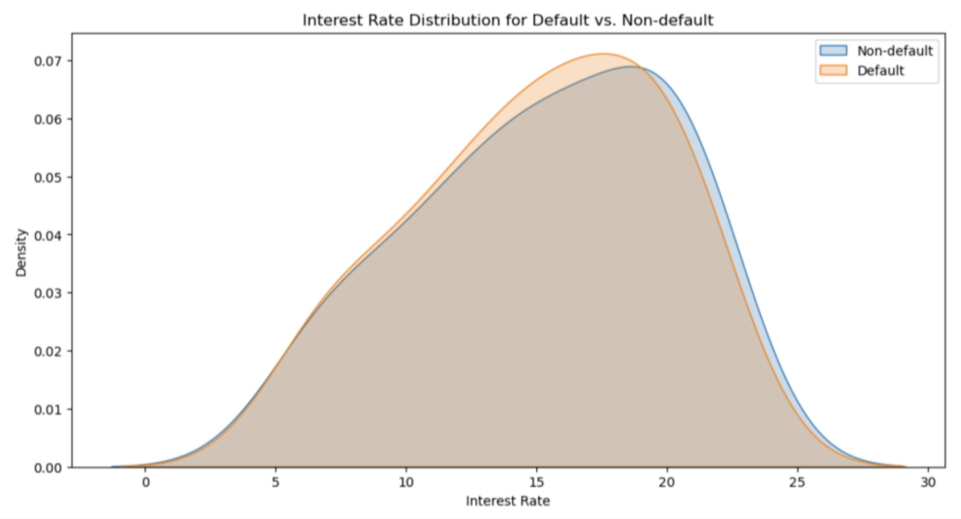
- While interest rates generally increase with higher grades, the difference in interest rates between successfully paid and defaulted loans remains relatively consistent within each grade.

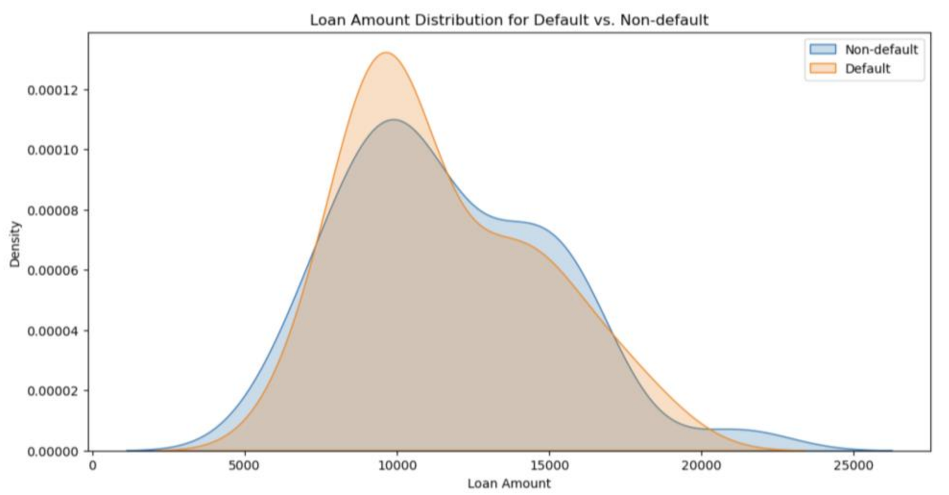
Risk by Grade:

- The analysis reveals a clear trend that higher-grade loans have lower default counts, indicating that customers in lower-grade categories are more likely to default.

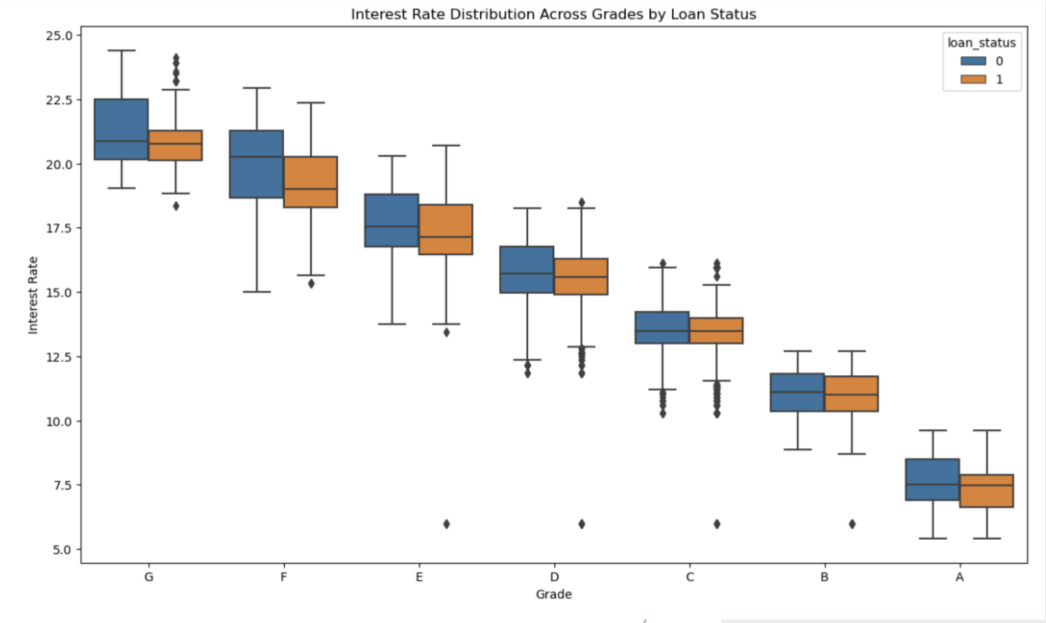
Loan Amount Variation:

- The difference in loan amounts between successful and defaulted loans tends to increase in higher-grade categories. This suggests that customers in higher-grade categories may be more cautious with their borrowing.





INSIGHTS GRADE-WISE



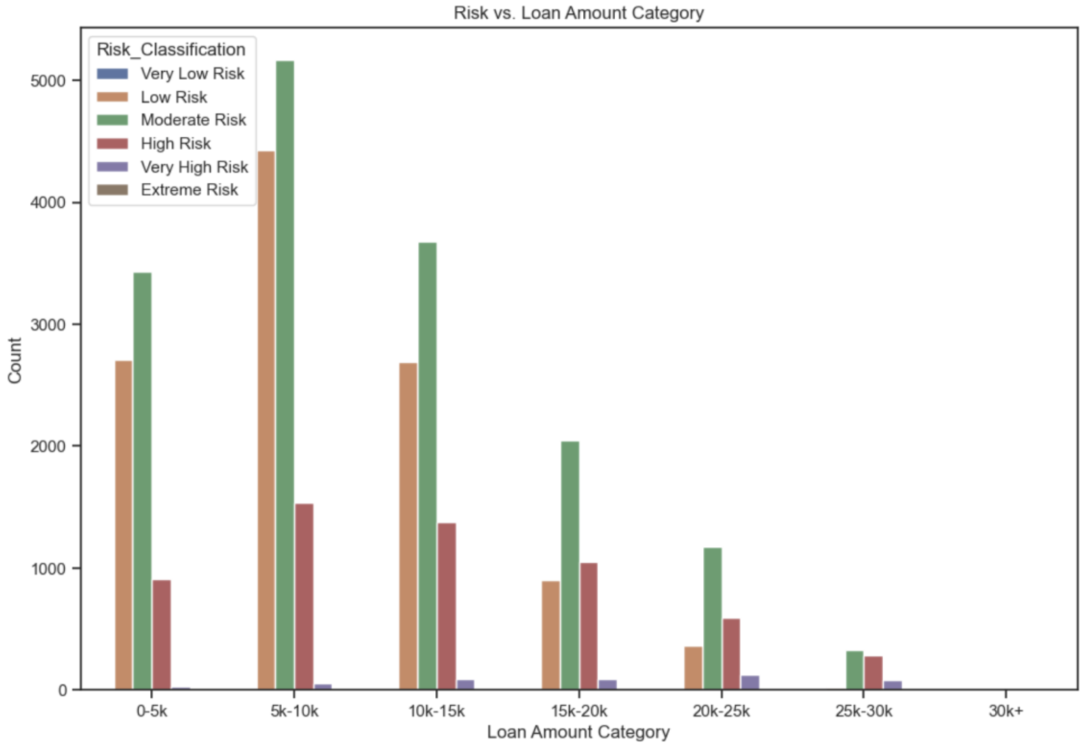
Insights for Business Strategy:

- The business should consider implementing strategies to manage risk, especially in lower-grade categories where default rates are higher.
- Monitoring and adjusting interest rates based on grade-specific default patterns can help optimize profitability.
- Personalized credit limit adjustments for customers in higher-grade categories may enhance risk management and customer satisfaction.

Opportunities for Improvement:

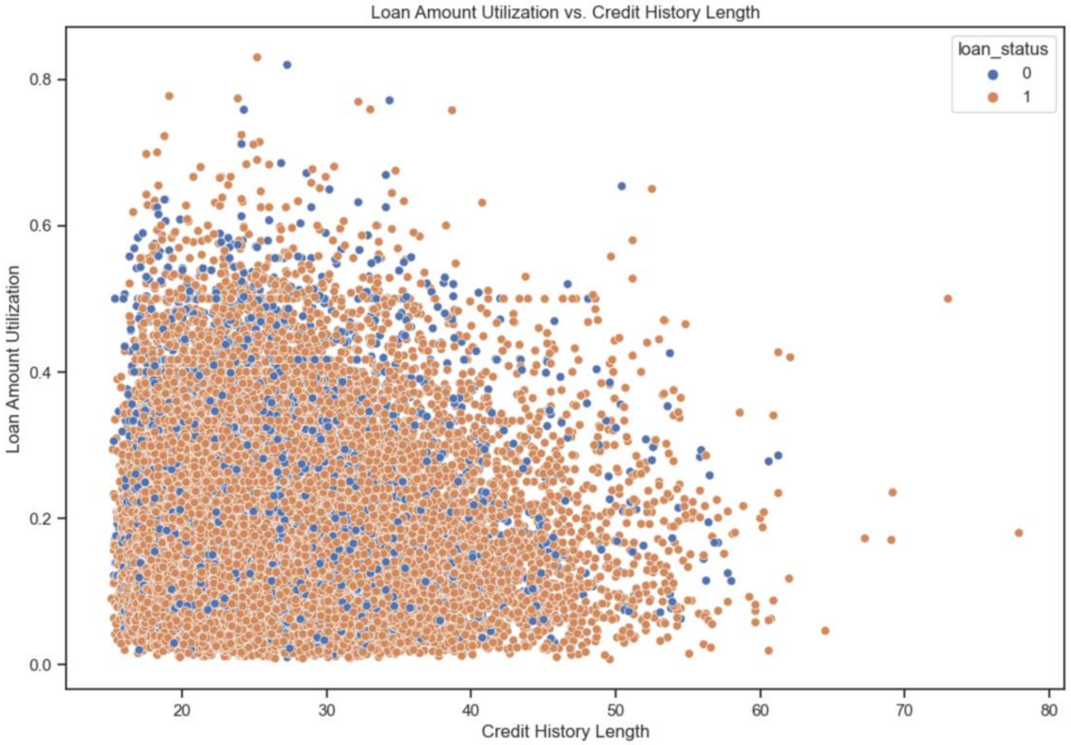
- The insights provide an opportunity for the business to refine its lending criteria, considering the observed patterns in interest rates, loan amounts, and default rates across different grades.

BIVARIATE ANALYSIS OF NEWLY DERIVED COLUMNS



Loan Amount Categories:

- 0-5k: This category has the second most overall count of loans, with a significant portion being classified as low risk. It indicates that smaller loans are generally considered less risky.
- 5-10k: This category has the highest overall count of loans, with a significant portion being classified as low risk. There is a noticeable increase in moderate risk loans compared to the 0-5k category, suggesting a slight uptick in risk as the loan amount increases.
- 10-15k: The trend of increasing moderate risk continues, and we start to see a rise in high-risk loans as well.
- 15-20k: This category marks a shift with a higher count of high-risk loans, indicating that loans of this size are more likely to be associated with higher risk.
- 20-25k: While the total count of loans decreases, the proportion of high and very high-risk loans remains significant.
- 25-30k: The count of loans further decreases, but the proportion of very high and extreme risk loans increases, highlighting the increased risk associated with larger loan amounts.
- 30k+: Although the count is the lowest in this category, the risk is predominantly high or very high, with a small but notable presence of extreme risk loans.



LOAN AMOUNT UTILIZATION VS. CREDIT HISTORY LENGTH

Distribution:

- There is a **higher concentration of orange circles** in the **lower left corner** of the plot, suggesting that loans with shorter credit history lengths and lower loan amount utilization are more likely to be not approved.
- The **blue circles** are more evenly distributed across the plot, indicating that loans with a variety of credit history lengths and loan amount utilizations have been approved.

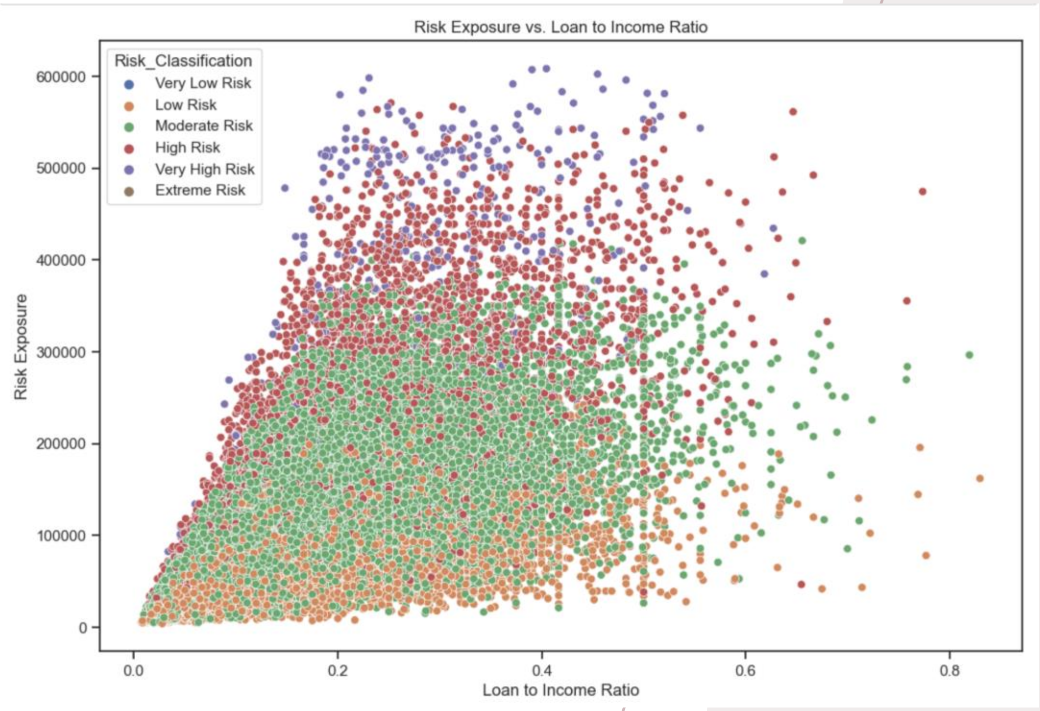
Trends:

- There appears to be no clear linear relationship between credit history length and loan amount utilization for approved loans, as indicated by the scattered distribution of blue circles.
- However, for not approved loans, there seems to be a trend where a shorter credit history length correlates with a lower loan amount utilization.

Insights:

- Lenders may be considering credit history length as a factor in the loan approval process, with a preference for borrowers who have a longer credit history.
- Loan amount utilization alone does not seem to be a decisive factor for loan approval, as approved loans show a wide range of utilization ratios.
- The combination of a short credit history and low loan amount utilization might be perceived as a higher risk, leading to a higher likelihood of loan disapproval.

RISK EXPOSURE VS. LOAN TO INCOME RATIO

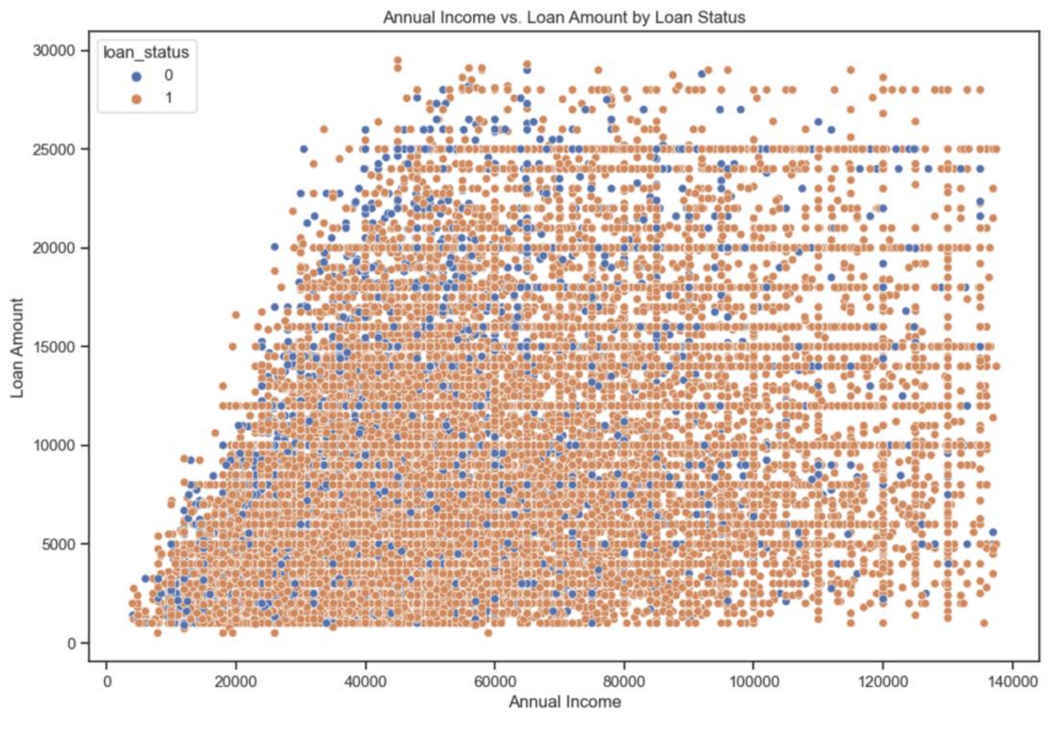


Data Point Distribution:

- There is a higher concentration of data points in the lower left corner of the graph, indicating that many loans with a low loan to income ratio also have low risk exposure.
- As we move towards the right side of the graph (higher loan to income ratio), the risk exposure tends to increase, suggesting that a higher debt burden relative to income is associated with greater risk.
- The top right corner has fewer data points, but these are likely to be classified as high or extreme risk, indicating that borrowers with a high loan to income ratio and high risk exposure are less common but potentially more concerning for lenders.

Insights:

- Lenders may use the loan to income ratio as a key factor in assessing the risk exposure of a loan.
- Borrowers with a lower loan to income ratio are generally considered safer bets for lenders.
- The graph can help lenders identify thresholds for risk exposure based on the loan to income ratio, aiding in decision-making for loan approvals and interest rates.



ANNUAL INCOME VS. LOAN AMOUNT BY LOAN STATUS

Data Point Distribution:

- The data points form a **diagonal line** from the bottom left to the top right of the graph.
- This pattern indicates a **positive correlation** between loan amount and risk exposure, suggesting that as the loan amount increases, so does the risk exposure.

Trends and Patterns:

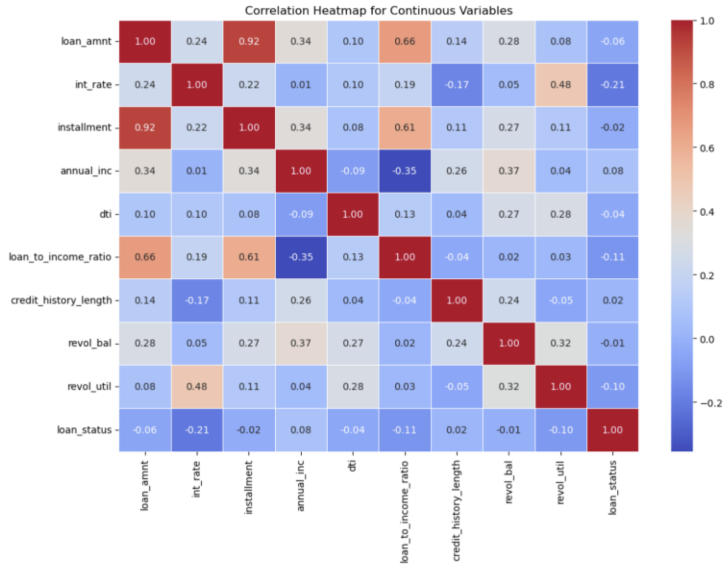
- The concentration of data points is denser at the lower end of the loan amount axis, which could imply that there are more loans with smaller amounts and correspondingly lower risk exposure.
- As we move towards higher loan amounts, the data points spread out, and the risk exposure values vary more widely.

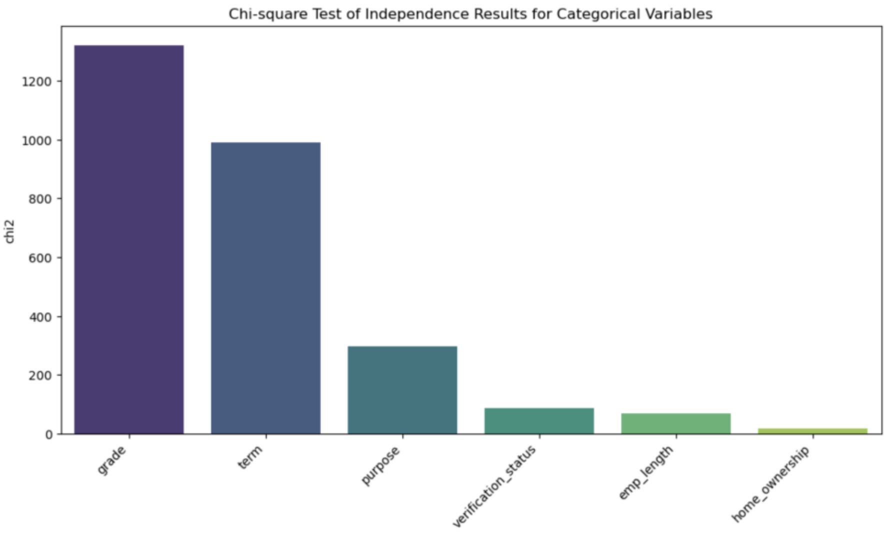
Insights for Lenders:

- Lenders might use this graph to assess the risk associated with different loan amounts and to set interest rates or lending terms accordingly.
- The positive correlation between loan amount and risk exposure could lead lenders to be more cautious with larger loans, possibly requiring additional security or higher interest rates to mitigate the risk.

Considerations for Borrowers:

- Borrowers seeking larger loans should be aware that lenders may view these as higher risk and may impose stricter lending criteria or higher costs.





CORRELATION HEATMAP FOR CONTINUOUS VARIABLES

- As it is clearly visible that, some of the variables have a high positive correlation with each other, such as 'loan_amnt' and 'installment', or 'revol_bal' and 'revol_util'.
- This means that as one variable increases, so does the other.
- Some of the variables have a high negative correlation with each other, such as 'int_rate' and 'annual_inc', or 'dti' and 'verification_status'.
- This means that as one variable increases, the other decreases.
- Some of the variables have a low or no correlation with each other, such as 'loan_amnt' and 'pub_rec', or 'annual_inc' and 'delinq_2yrs'.
- This means that there is no clear relationship between them.

BASED ON THE DATA ANALYSIS, HERE ARE SOME FINAL BUSINESS TIPS AND SUGGESTIONS FOR LEADINGCLUB'S LEADER:

Loan Approval Strategy:

- The company could refine its loan approval strategy by considering the purpose of the loan. For instance, customers applying for educational loans or small business loans have a higher likelihood of defaulting.

Interest Rates and Loan Amounts:

- Consider adjusting interest rates based on the purpose of the loan. Customers with higher-risk purposes tend to default more; hence, adjusting interest rates accordingly might help mitigate risks.
- There is a positive correlation between loan amounts and the likelihood of default. The company may want to reassess its loan amount approval criteria to reduce risk exposure.

Grade and Subgrade Analysis:

- Analyzing loans based on grade and subgrade provides insights into the risk associated with different categories. It's important to monitor and adjust strategies for higher-risk grades.

Risk Exposure:

- Evaluate and manage risk exposure carefully. The company might want to focus on strategies to reduce risk in loan portfolios.

BASED ON THE DATA ANALYSIS, HERE ARE SOME FINAL BUSINESS TIPS AND SUGGESTIONS FOR LEADINGCLUB'S LEADER:

Employee Stability and Promotion Potential:

- The Employee Stability Index and Promotion Potential are consistent across loan statuses. However, it's essential to monitor these indices over time and ensure a stable and motivated workforce.

Income Range and Income Deviation:

- Analyzing income ranges and deviations can help in tailoring loan products to different income groups. It's crucial to ensure that loan products align with the financial capacity of the borrowers.

Affordability Score:

- The Affordability Score seems to be NaN in the provided data. It's essential to investigate and rectify this issue to utilize it as a valuable metric for loan assessment.

Interest Spread:

- The negative Interest Spread for loans with status 1 indicates potential issues with profit margins. The company should revisit interest rate policies to maintain profitability.

BASED ON THE DATA ANALYSIS, HERE ARE SOME FINAL BUSINESS TIPS AND SUGGESTIONS FOR LEADINGCLUB'S LEADER:

Loan Amount Utilization:

- Monitoring the loan amount utilization can provide insights into how customers are using the funds. This information can be useful for risk assessment and product development.

Risk Classification:

- The risk classification provides a breakdown of risk levels. The company should focus on mitigating risks associated with high and very high-risk classifications.

New Variables Insights:

- Evaluate the new variables like Risk_Exposure, Loan_Amount_Category, Profitability_Index, and others for ongoing monitoring and potential refinement of business strategies.