**Introduction**

Objective: To identify factors influencing loan defaults within the largest online loan marketplace data. This analysis focuses on key variables related to loan defaults, aiming to enhance risk assessment and improve decision-making in loan approvals.

**Data Cleaning Steps**

1. **Normalization**: Key variables such as Debt-to-Income Ratio (DTI) and Loan Amount were normalized to align scales and improve comparability.
2. **Percentage Conversion**: Converted percentage strings (e.g., Interest Rate) to numerical values for accurate calculations and comparisons.
3. **Imputation of Missing Values**: Missing values, particularly in numerical columns, were imputed using median values to maintain data integrity.
4. **One-Hot Encoding**: Categorical variables like 'home\_ownership', 'term', and 'grade' were one-hot encoded to transform them into a format suitable for modeling.
5. **Derived Metrics Creation**: Developed new metrics such as 'credit\_score', 'loan\_utilization\_rate', and 'payment-to-income ratio (pti)' for deeper analysis.

Note :**for credit score**

loan\_data['credit\_score'] = 100 - (loan\_data['normalized\_dti'] \* 0.7 + loan\_data['normalized\_loan\_amnt'] \* 0.3) \* 100

The formula used to calculate the credit score in your dataset was based on a combination of the normalized Debt-to-Income Ratio (DTI) and normalized Loan Amount, with respective weights of 70% and 30%. This specific formula was likely chosen based on certain assumptions and objectives:

1. **Debt-to-Income Ratio (DTI) Weight**: The DTI ratio is a crucial indicator in credit scoring. It measures a borrower's monthly debt payments relative to their income. A higher DTI ratio suggests a higher burden of debt relative to income, which could indicate a higher risk of default. Giving it a 70% weight emphasizes its importance in assessing financial stress and repayment ability.
2. **Loan Amount Weight**: The size of the loan can also be an important factor in credit risk. Larger loans might represent a higher risk, especially if they constitute a significant portion of the borrower's financial commitments. However, this factor is usually less predictive of default risk than DTI, hence the lower weighting of 30%.
3. **Normalization**: Both DTI and Loan Amount are normalized to ensure they are on a similar scale. This standardization is crucial when combining variables that may originally have different ranges and distributions.
4. **100-Scaling**: The formula scales the weighted sum to a 100-based scale, which is a common practice in credit scoring. This makes the score more interpretable, with higher scores generally indicating better creditworthiness.

It's important to note that this formula and the weights used are somewhat arbitrary and based on certain assumptions about the relative importance of DTI and Loan Amount

**Handling Outliers**

Techniques Used:

1. **Capping Outliers**: Outliers in key financial metrics like annual income or loan amount were capped. This involves setting a threshold (at. the 95th percentile), and values above this threshold are replaced with the threshold value itself. This method retains the data point while reducing the impact of extreme values.
2. **Normalization**: Normalizing the data, as done with DTI and Loan Amount, also helps in mitigating the impact of outliers. Normalization transforms the data into a standard scale, making the dataset less sensitive to extreme values.
3. **Log Transformation**: For highly skewed data, log transformation is a common technique. It helps in reducing the impact of extreme values, making the data more normally distributed, which is a desirable property for many statistical models.

Rationale:

* **Preserve Data Integrity**: Completely removing outliers can lead to loss of valuable information. Capping and transformations allow for retaining these data points while minimizing their potential to skew the analysis.
* **Improve Model Performance**: Outliers can disproportionately influence the model's training process, leading to poor generalization on unseen data. By handling outliers appropriately, models can achieve better performance and more reliable predictions.
* **Statistical Requirements**: Many statistical methods and machine learning algorithms assume data is normally distributed. Transformations help in meeting these assumptions, leading to more valid inferences and predictions.

**Univariate Analysis Insights**

* **Loan Amount ('loan\_amnt')**: Exhibits a multi-modal distribution, indicating common specific amounts at which loans are frequently disbursed.
* **Funded Amount ('funded\_amnt')**: Similar to loan amount, indicating popularity among specific loan values.
* **Interest Rate ('int\_rate')**: The distribution shows a concentration in a specific range, indicating common rates for most loans.
* **Loan Term ('term')**: Certain loan terms are more prevalent, likely reflecting industry standards.
* **Grade ('grade')**: Distribution shows the frequency of different risk categories, indicating common risk assessments in the dataset.
* **Annual Income ('annual\_inc')**: Likely right-skewed, with most borrowers in lower income brackets.
* **DTI (Debt-to-Income Ratio)**: Distribution provides insights into borrowers' financial health, with higher DTI suggesting significant income going towards debt repayment.

**Segmented Univariate Analysis by Interest Rate**

Insights:

1. **Loan Amount ('loan\_amnt')**:
   * Visualization: Box plots for loan amounts across different interest rate categories.
   * Insight: Observe if higher loan amounts correspond to higher interest rates, possibly indicating risk-based pricing strategies.
2. **Annual Income ('annual\_inc')**:
   * Visualization: Box plots showing the distribution of borrowers' annual incomes across interest rate categories.
   * Insight: Determine if higher incomes are associated with lower interest rates, reflecting creditworthiness.
3. **Debt-to-Income Ratio ('dti')**:
   * Visualization: Box plots for DTI ratios across different interest rate categories.
   * Insight: Explore whether borrowers with higher DTI ratios tend to have higher interest rates, indicating increased credit risk.
4. **Credit Score (Derived Metric)**:
   * Visualization: Box plots or violin plots showing the distribution of credit scores across different interest rate categories.
   * Insight: Assess how credit scores vary with interest rates, with an expectation of lower scores corresponding to higher rates.
5. **Loan Utilization Rate (Derived Metric)**:
   * Visualization: Box plots for loan utilization rates across interest rate categories.
   * Insight: Investigate if higher utilization rates are linked to higher interest rates, which might indicate borrowers maximizing available credit.

**Relationship with Credit Score**

* **Annual Income**: Distribution across credit score categories shows income trends relative to creditworthiness.
* **Interest Rate**: Variation in rates offered to different credit scores underlines risk-based pricing.
* **Home Ownership**: Distribution among credit score categories reveals correlations between homeownership types and credit scores.
* **Employment Length**: Indicates employment stability across different credit risk groups.

**Deeper Bivariate Analysis**

1. **Statistical Correlation Analysis**:
   * Pearson Correlation between Annual Income and Loan Amount shows a moderate positive relationship, indicating that borrowers with higher incomes tend to take larger loans.
   * Spearman Correlation suggests a moderate positive monotonic relationship, indicating the possibility of a non-linear correlation.
2. **Segmented Analysis by Credit Score Categories**:
   * The relationship between annual income and loan amount across different credit score categories. This analysis shows whether higher credit scores correlate with larger loans.

**Correlation Insights with Default Status**

The analysis of correlation coefficients between 'is\_default' and various columns in the dataset reveals significant insights:

1. **Loan Amount (loan\_amnt)**: Positive correlation (0.048217) suggests a slight increase in default risk with higher loan amounts.
2. **Interest Rate (int\_rate)**: A more substantial positive correlation (0.196253) indicates a significant increase in default probability with higher interest rates, reflecting risk-based pricing.
3. **Annual Income (annual\_inc)**: Negative correlation (-0.068794) implies that higher income borrowers have a lower likelihood of default.
4. **DTI (Debt-to-Income Ratio)**: Shows a positive correlation (0.041701), suggesting that borrowers with higher DTI are slightly more likely to default.
5. **PTI (Payment-to-Income Ratio)**: Positive correlation (0.080584) indicates that higher PTI ratios increase the risk of default.
6. **Loan Utilization Rate**: The negative correlation (-0.017406) suggests that higher loan utilization might slightly reduce the likelihood of default.
7. **Recoveries**: Strong positive correlation (0.340297) suggests that higher recovery amounts are significantly associated with defaults, likely due to post-default recoveries.
8. **Last Payment Amount (last\_pymnt\_amnt)**: Negative correlation (-0.214949) indicates that higher last payment amounts reduce the likelihood of default.
9. **Total Payment (total\_pymnt)**: Strong negative correlation (-0.238844) signifies that higher total payments are strongly associated with lower default risks.

**Correlation Insights**

1. **Loan Amount and Funded Amount**: High positive correlation, suggesting that requested amounts are typically close to funded amounts.
2. **Interest Rate and Credit Score**: Significant negative correlation might indicate that borrowers with higher credit scores tend to receive lower interest rates.
3. **Annual Income and Loan Amount**: Moderate positive correlation, indicating that borrowers with higher incomes usually take larger loans.
4. **DTI and Loan Amount**: A correlation here could indicate that borrowers with larger loans also have higher debt-to-income ratios.
5. **Interest Rate and Loan Term**: Potential positive correlation may suggest that longer-term loans often have higher interest rates.
6. **Credit Score and DTI**: Negative correlation implies higher credit scores are associated with lower debt-to-income ratios.
7. **Loan Utilization and Credit Score**: Negative correlation might suggest that borrowers with higher credit scores use a lower proportion of their loan amount.
8. **PTI and Credit Score**: Negative correlation here would indicate that borrowers with higher credit scores tend to have lower payment-to-income ratios, suggesting better loan affordability.

**Correlation Insights with Default Status**

The analysis of correlation coefficients between 'is\_default' and various columns in the dataset reveals significant insights:

1. **Loan Amount (loan\_amnt)**: Positive correlation (0.048217) suggests a slight increase in default risk with higher loan amounts.
2. **Interest Rate (int\_rate)**: A more substantial positive correlation (0.196253) indicates a significant increase in default probability with higher interest rates, reflecting risk-based pricing.
3. **Annual Income (annual\_inc)**: Negative correlation (-0.068794) implies that higher income borrowers have a lower likelihood of default.
4. **DTI (Debt-to-Income Ratio)**: Shows a positive correlation (0.041701), suggesting that borrowers with higher DTI are slightly more likely to default.
5. **PTI (Payment-to-Income Ratio)**: Positive correlation (0.080584) indicates that higher PTI ratios increase the risk of default.
6. **Loan Utilization Rate**: The negative correlation (-0.017406) suggests that higher loan utilization might slightly reduce the likelihood of default.
7. **Recoveries**: Strong positive correlation (0.340297) suggests that higher recovery amounts are significantly associated with defaults, likely due to post-default recoveries.
8. **Last Payment Amount (last\_pymnt\_amnt)**: Negative correlation (-0.214949) indicates that higher last payment amounts reduce the likelihood of default.
9. **Total Payment (total\_pymnt)**: Strong negative correlation (-0.238844) signifies that higher total payments are strongly associated with lower default risks.

**Bivariate Analysis Insights on Loan Default Status**

Interest Rate and Loan Default Status (Box Plot Analysis)

* **Interest Rate Distribution for Defaults vs. Non-Defaults**: The box plot reveals that loans which defaulted (charged off) tend to have higher interest rates compared to those that didn't default. This suggests risk-based pricing where riskier loans are assigned higher rates.
* **Median Interest Rate**: Differences in the median interest rates between defaulted and non-defaulted loans indicate that higher rates are a significant factor in loan defaults.
* **Spread and Outliers**: The range and outliers in the interest rate distribution for both defaulted and non-defaulted loans reveal the variance in risk pricing. Outliers, especially in the defaulted category, may indicate exceptionally high-risk loans.

Additional Insights from Bivariate Analysis

1. **Loan Amount ('loan\_amnt')**: Higher loan amounts are slightly more associated with defaults, possibly due to the increased financial burden on borrowers.
2. **Home Ownership ('home\_ownership')**: Default rates vary across different home ownership statuses, reflecting the financial stability and obligations associated with each type.
3. **Employment Length ('emp\_length')**: Shorter employment lengths might correlate with higher default rates, potentially due to less stable income or career uncertainty.
4. **Loan Term ('term')**: Longer loan terms are possibly linked to higher default rates, as extended repayment periods increase financial uncertainties over time.
5. **Grade ('grade')**: Higher default rates are associated with lower loan grades, validating the risk assessment in the grading system.
6. **Debt-to-Income Ratio ('dti')**: A higher DTI ratio is a significant predictor of default, indicating financial strain.
7. **Payment-to-Income Ratio ('PTI') (Derived Metric)**: Higher PTIs suggest that a larger portion of a borrower's income goes towards loan payments, increasing the risk of default.
8. **Loan Utilization Rate ('loan\_utilization\_rate') (Derived Metric)**: High loan utilization rates might indicate financial strain, leading to higher defaults.

**Key Insights for Loan Defaults**

1. **Higher Interest Rates**: Loans with higher rates show a greater likelihood of default, likely due to increased financial burden.
2. **Large Loan Amounts**: Larger loans are slightly more prone to defaults, possibly due to the greater financial commitments they represent.
3. **Low Credit Scores**: Lower credit scores are associated with higher defaults, indicating riskier borrower profiles.
4. **High Debt-to-Income Ratio (DTI)**: A higher DTI suggests borrowers are overleveraged, increasing the default risk.
5. **Payment-to-Income Ratio (PTI)**: Higher PTIs indicate a significant portion of income is dedicated to loan payments, leading to defaults under financial strain.
6. **Loan Utilization Rate**: Borrowers who utilize a larger portion of their loan amount tend to default more, reflecting potential financial stress.
7. **Short Employment Length**: Shorter employment history might correlate with financial instability, leading to higher default rates.
8. **Lower Loan Grades**: Loans with lower grades (indicating higher risk) have higher default rates.
9. **Longer Loan Terms**: Extended loan terms are associated with a higher risk of default, potentially due to prolonged financial exposure.
10. **Home Ownership Status**: Renters and those with mortgages show different default patterns than outright homeowners, possibly due to varying financial stability.
11. **Variations in Annual Income**: Fluctuations in borrowers' annual income over time can impact their ability to repay loans.
12. **Economic and Market Changes**: Macro-economic factors and market dynamics can significantly influence default rates.
13. **Policy and Regulation Changes**: Shifts in lending policies or financial regulations can affect default rates.
14. **Recovery Efforts**: The effectiveness of collection and recovery efforts plays a role in managing defaults.
15. **Borrower Demographics**: Changes in the borrower demographic over time can influence the overall risk profile and default rates.

**Suggestions to Better Identify Defaults**

1. **Enhanced Risk Modeling**: Incorporate a wider range of variables, including economic indicators and borrower behavior metrics, into risk assessment models.
2. **Dynamic Credit Scoring**: Update credit scoring models regularly to reflect current economic conditions and borrower trends.
3. **Segmented Strategies**: Develop differentiated strategies for various borrower segments based on risk profiles.
4. **Early Warning Indicators**: Implement systems to identify early warning signs of financial stress in borrowers.
5. **Flexible Repayment Options**: Offer flexible repayment plans for borrowers showing early signs of financial strain.
6. **Data-Driven Collection Practices**: Use data analytics to optimize collection and recovery efforts, focusing on high-risk accounts.
7. **Regular Policy Reviews**: Continuously review and update lending policies in response to changing market conditions and regulatory landscapes.
8. **Borrower Education and Support**: Provide financial education and advisory services to borrowers to help them manage their finances better.
9. **Technology Integration**: Leverage advanced technologies like AI and machine learning for predictive analytics in default identification.
10. **Diversification of Loan Portfolio**: Diversify the loan portfolio to spread risk across different loan types and borrower segments.

By focusing on these insights and implementing these suggestions, the company can enhance its ability to identify potential defaults early and manage them more effectively, thereby mitigating risk and ensuring sustainable lending practices.