For eda

amount, assuming a higher DTI and higher loan amount imply higher credit risk.

UNI variate analysis

The univariate analysis of key variables in your loan dataset provides the following insights:

Loan Amount ('loan\_amnt'): The distribution appears to be multi-modal, indicating common specific amounts at which loans are frequently disbursed.

Funded Amount ('funded\_amnt'): Similar to the loan amount, the distribution of funded amounts shows peaks at certain loan values, suggesting that these amounts are popular among borrowers.

Interest Rate ('int\_rate'): The distribution of interest rates is somewhat skewed, with a concentration in a specific range. This indicates common interest rates at which most loans are offered.

Loan Term ('term'): The count plot shows the frequency of different loan terms. It appears that certain terms are more popular or standard in the loan industry.

Grade ('grade'): This count plot of loan grades (which usually reflect the risk or quality of the loans) can show which grades are most common in your dataset.

Annual Income ('annual\_inc'): The distribution of annual income is likely right-skewed, showing that most borrowers fall into lower income brackets, with fewer borrowers having higher incomes.

Debt-to-Income Ratio ('dti'): The DTI ratio's distribution can provide insights into the financial health of borrowers. A high DTI ratio suggests that a significant portion of borrowers' income is going towards debt repayment.

Segmented univariate analysis by grade

1. **Loan Amount ('loan\_amnt')**: The distribution of loan amounts across different grades shows how the average loan amount, as well as the range and outliers, vary by grade. There may be trends indicating that certain grades tend to have higher or lower loan amounts.
2. **Annual Income ('annual\_inc')**: By observing the distribution of borrowers' annual incomes across loan grades, we can infer if there's a relationship between income levels and the risk category of the loan. For instance, higher income levels might be more prevalent in lower-risk grades.
3. **Debt-to-Income Ratio ('dti')**: The DTI ratio across different grades can highlight if borrowers with certain risk profiles tend to have higher or lower debt burdens relative to their income.

**The segmented analysis across different loan statuses for key variables provides valuable insights:**

1. **Loan Amount ('loan\_amnt')**: This shows how the loan amount varies across different loan statuses, which can indicate if larger loans are more likely to be in certain statuses like 'Charged Off' or 'Fully Paid'.
2. **Annual Income ('annual\_inc')**: Observing annual income across loan statuses can reveal if there's a trend where borrowers with different income levels are more likely to be in specific loan statuses.
3. **Debt-to-Income Ratio ('dti')**: The distribution of DTI across loan statuses can highlight if borrowers in certain statuses (like 'Default' or 'Charged Off') tend to have higher debt burdens relative to their income.
4. **Interest Rate ('int\_rate')**: Analyzing interest rates across loan statuses can provide insights into whether loans with higher interest rates are more likely to end up in adverse statuses.
5. **Loan Utilization Rate ('loan\_utilization\_rate')**: This can show if borrowers who utilize a higher proportion of their approved loan amount are more likely to fall into certain loan statuses.
6. **Payment-to-Income Ratio ('pti')**: Observing how the PTI ratio varies with loan statuses can indicate if borrowers with higher loan payment burdens relative to their income are often in statuses like 'Late' or 'Charged Off'.
7. **Loan to Income Ratio ('loan\_to\_income')**: This analysis can reveal how the ratio of loan amount to annual income differs among various loan statuses.

**Credit risk**

1. Annual Income ('annual\_inc'): The boxplot shows the distribution of annual incomes across credit score categories, revealing trends in income relative to creditworthiness.
2. Interest Rate ('int\_rate'): This visualization illustrates the variation in interest rates offered to different credit score categories, a key aspect of risk-based pricing in lending.
3. Home Ownership ('home\_ownership'): The count plot for home ownership status across credit score categories indicates how different types of homeownership (like renting, owning, or having a mortgage) are distributed among borrowers with different credit scores.
4. Employment Length ('emp\_length'): Similarly, this count plot shows the distribution of employment lengths across credit score categories, providing insights into employment stability among different credit risk groups.

Bi variate

1. **Interest Rate vs. Credit Score**: The scatter plot shows the relationship between the credit score and the interest rate. Ideally, higher credit scores should correlate with lower interest rates, reflecting lower credit risk.
2. **Loan Amount vs. Annual Income**: This scatter plot explores the relationship between a borrower's annual income and the loan amount they are granted. A positive correlation is often expected, indicating that borrowers with higher incomes are approved for larger loans.
3. **DTI Ratio vs. Loan Status**: The box plot demonstrates how the debt-to-income ratio varies across different loan statuses. Higher DTI ratios in certain loan statuses (like 'Charged Off' or 'Default') can indicate financial strain.
4. **Loan Amount vs. Term**: This box plot shows the distribution of loan amounts across different loan terms. Longer terms might be associated with larger loans, as they are typically spread out over a longer period.
5. **PTI (Payment-to-Income Ratio) vs. Credit Score Category**: The box plot for this derived variable against the credit score category illustrates how the burden of loan payments relative to income varies across different levels of creditworthiness.

Deeper bivariate analysis

The deeper bivariate analysis provides several key insights:

1. **Statistical Correlation Analysis**:
   * Pearson Correlation between Annual Income and Loan Amount: 0.44. This indicates a moderate positive linear relationship.
   * pearson Correlation between Annual Income and Loan Amount: 0.43. This suggests a moderate positive monotonic relationship. Since Spearman's correlation is higher, it implies that the relationship may not be strictly linear.
2. **Segmented Bivariate Analysis - Loan Amount vs. Annual Income for Different Credit Score Categories**:
   * These scatter plots for each credit score category show how the relationship between annual income and loan amount varies with credit score. You can observe whether higher credit scores are associated with larger loans or higher incomes.
3. **Advanced Visualization - Joint Plot for Annual Income and Loan Amount**:
   * The joint plot combines scatter and regression line plots with histograms, providing a comprehensive view of the relationship between these two variables. It shows both the distribution of each variable and their bivariate relationship.

Corelation insights

1. **Loan Amount and Funded Amount**: There's likely a very high positive correlation between 'loan\_amnt' and 'funded\_amnt'. This suggests that the amount requested by borrowers is typically close to the amount funded.
2. **Interest Rate and Credit Score**: If there's a significant negative correlation between 'int\_rate' and the derived 'credit\_score', it indicates that borrowers with higher credit scores tend to receive loans with lower interest rates.
3. **Annual Income and Loan Amount**: A moderate positive correlation between 'annual\_inc' and 'loan\_amnt' suggests that borrowers with higher incomes tend to take larger loans.
4. **Debt-to-Income Ratio and Loan Amount**: If there's a noticeable correlation between 'dti' and 'loan\_amnt', it could indicate that borrowers with larger loans also tend to have higher debt-to-income ratios.
5. **Interest Rate and Loan Term**: A potential positive correlation between 'int\_rate' and 'term' might imply that longer-term loans often come with higher interest rates.
6. **Credit Score and DTI Ratio**: If the derived 'credit\_score' and 'dti' show significant negative correlation, it implies that higher credit scores are associated with lower debt-to-income ratios.
7. **Loan Utilization and Credit Score**: A negative correlation between 'loan\_utilization\_rate' and 'credit\_score' could suggest that borrowers with higher credit scores use a lower proportion of their approved loan amount.
8. **Payment-to-Income Ratio and Credit Score**: A significant negative correlation here would indicate that borrowers with higher credit scores tend to have lower payment-to-income ratios, suggesting better affordability.
9. **Annual Income and Interest Rate**: A negative correlation would suggest that borrowers with higher annual incomes generally receive loans with lower interest rates, possibly reflecting a lower perceived risk.
10. **Loan Term and Payment-to-Income Ratio (PTI)**: If a positive correlation exists, it might indicate that loans with longer terms generally have higher payment-to-income ratios, potentially due to larger loan amounts being spread over longer periods.

Based on the analyses conducted, including univariate, segmented univariate, bivariate, and correlation analyses, along with the derived metrics, here are key factors that might indicate a higher likelihood of loan default:

1. **Credit Score (Derived Metric)**: Lower credit scores typically indicate higher risk. A negative correlation with interest rates suggests that lower credit scores are associated with higher rates, often reflecting a higher risk of default.
2. **Debt-to-Income Ratio (DTI)**: Higher DTI ratios, as observed in the univariate analysis, may indicate a borrower's financial stress, increasing the risk of default.
3. **Loan Amount**: The bivariate analysis between loan amount and loan status might reveal that higher loan amounts are associated with a greater risk of default.
4. **Interest Rate**: Higher interest rates, possibly due to risk-based pricing, could correlate with a higher likelihood of default. This relationship can be further explored through bivariate analysis.
5. **Loan Term**: Longer loan terms might be associated with an increased risk of default, as suggested by certain bivariate analyses.
6. **Annual Income**: While higher incomes generally correlate with larger loan amounts, it's essential to consider this in conjunction with other factors like DTI and loan amount for default risk assessment.
7. **Employment Length**: Shorter employment lengths might indicate less stability, potentially correlating with a higher risk of default.
8. **Home Ownership**: The segmented univariate analysis can reveal if certain types of homeownership (like renting or owning) are more associated with defaults.
9. **Loan Utilization Rate (Derived Metric)**: A higher loan utilization rate might indicate financial stress, possibly leading to a higher likelihood of default.
10. **Payment-to-Income Ratio (PTI) (Derived Metric)**: Higher PTIs might suggest that loan payments take up a significant portion of a borrower's income, which could increase the risk of default.
11. **History of Delinquency (Derived from Data Dictionary)**: Past delinquencies can be a strong indicator of future risks of default.
12. **Number of Charge-offs (Derived from Data Dictionary)**: A history of charge-offs within the last 12 months can signal an increased risk of default.

These insights, derived from various analyses, can be instrumental in developing more effective credit scoring models, refining loan approval criteria, and enhancing overall risk management strategies. It's important to note that these factors should be considered in a comprehensive model that takes into account the interdependencies and nuances of borrower behavior and financial health. Additionally, compliance with lending regulations and ethical considerations in credit assessment should always be maintained.

Loan default tatus

The box plot visualizes the relationship between the interest rate and loan default status. From this plot, you can observe the following:

1. **Interest Rate Distribution for Defaults vs. Non-Defaults**: The plot shows the distribution of interest rates for loans that were charged off (labeled as default) compared to those that were not.
2. **Median Interest Rate**: Look for differences in the median interest rates (indicated by the line inside the boxes) between defaulted and non-defaulted loans.
3. **Spread and Outliers**: The range of the boxes and the whiskers indicate the spread of the interest rates. Outliers, if any, are shown as points outside the whiskers.

Based on the bivariate analysis conducted between various variables and loan default status, here are detailed insights:

1. **Loan Amount ('loan\_amnt')**:
   * A higher loan amount might correlate with an increased likelihood of default. This suggests that larger loans carry higher risk, potentially due to the greater financial burden they impose on borrowers.
2. **Home Ownership ('home\_ownership')**:
   * The default rates vary across different home ownership statuses. Renters or those with mortgages might show different default patterns compared to outright homeowners. This could be due to variations in financial stability and obligations associated with different types of home ownership.
3. **Interest Rate ('int\_rate')**:
   * Loans with higher interest rates might be more prone to default. This could reflect risk-based pricing, where higher rates are assigned to riskier loans, or it might indicate that higher rates increase the financial strain on borrowers, leading to defaults.
4. **Employment Length ('emp\_length')**:
   * Employment length could be an indicator of financial stability. Shorter employment lengths might be associated with higher default rates, possibly due to less stable income or career uncertainty.
5. **Loan Term ('term')**:
   * Longer loan terms could be linked to higher default rates. This may be because extended repayment periods increase the uncertainty and risks associated with borrowers’ financial situations over time.
6. **Grade ('grade')**:
   * The default rate likely increases with lower loan grades. This aligns with the expectation that lower grades (indicating higher risk) are more prone to defaults, validating the risk assessment criteria used to assign these grades.
7. **Debt-to-Income Ratio ('dti')**:
   * A higher DTI ratio could be a significant predictor of default. This implies that borrowers with a higher proportion of debt relative to their income are more likely to face financial difficulties, leading to defaults.
8. **Payment-to-Income Ratio ('PTI') (Derived Metric)**:
   * Higher PTI ratios might be associated with higher default rates. This indicates that when loan payments constitute a large portion of a borrower’s income, there's a greater risk of default due to limited financial flexibility.
9. **Loan Utilization Rate ('loan\_utilization\_rate') (Derived Metric)**:
   * If higher utilization rates correlate with higher defaults, it suggests that borrowers who use a large portion of their approved loan amount are at a greater risk of financial strain, leading to defaults.

The correlation coefficients between the loan default status and various variables are as follows:

1. **Credit Score**: -0.056
   * This negative correlation suggests that a higher credit score (indicating lower credit risk) is slightly associated with a lower likelihood of default.
2. **Loan Amount**: 0.048
   * A positive correlation, although weak, indicates that higher loan amounts are slightly more likely to be associated with defaults.
3. **Debt-to-Income Ratio (DTI)**: 0.042
   * A positive correlation, albeit weak, suggests that higher DTI ratios might be slightly associated with an increased likelihood of default.
4. **Payment-to-Income Ratio (PTI)**: 0.081
   * This shows a slightly stronger positive correlation, indicating that loans with higher PTIs (a larger portion of income going towards loan payments) have a somewhat higher likelihood of default.
5. **Loan Utilization Rate**: -0.017
   * This negative correlation, although very weak, suggests that higher loan utilization rates are slightly less likely to be associated with defaults.

These correlations provide insights into factors that might influence loan defaults. The negative correlation with credit score aligns with the expectation that higher creditworthiness reduces default risk. Similarly, the positive correlations with loan amount, DTI, and PTI suggest that higher financial burdens or larger loans are associated with a slightly increased risk of default. However, these correlations are relatively weak