**Project Report**

**Project Title**

**Dimensionality Reduction for Credit Risk Management of P2P Lending Loans**

**Team Leader:**

Shivanoor Vignesh 21R11A0597

**Team Members:**

G . Sandhya Rani 21R11A0569

Dhathri 21R11A0545

Hasini 21R11A0518

Pooja 21R11A0501

PSBN Sriya 21R11A0595

MD.Akhbar Ali 21R11A0586

N sai Charan 21R11A0588

**Exploratory Data Analysis**

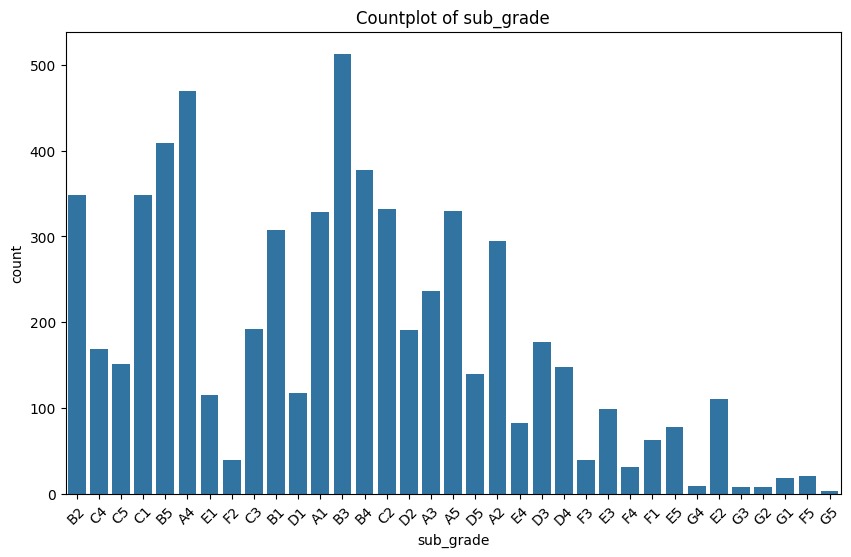
**Import data set**

For importing the dataset and to perform Exploratory Data Analysis we have to import some packages or library which are essential.

* import pandas
* import NumPy
* import seaborn

import matplotlib

1. **Distribution of Loan Sub-Grades**



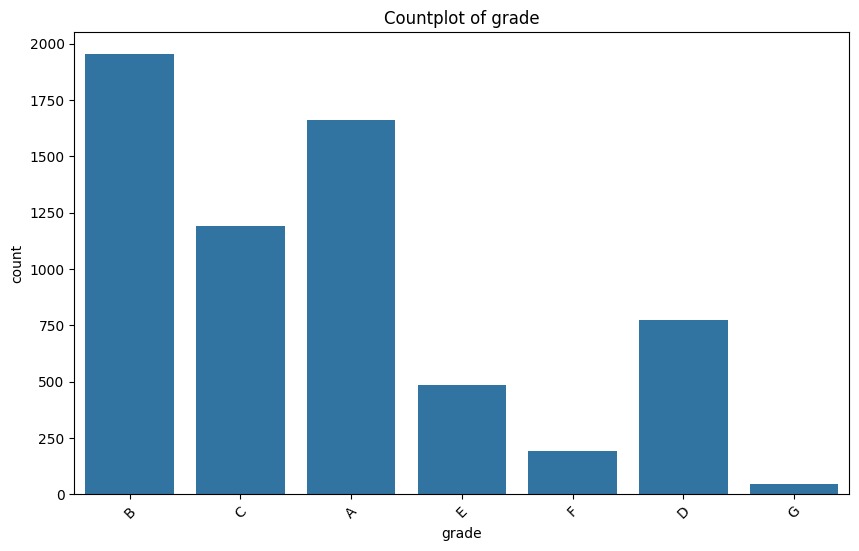
**OBSERVATION**

The above observations highlight several key trends and patterns in the distribution of sub-grades. Firstly, there is a notable frequency distribution, with certain sub-grades such as B4, C1, and B3 being more common. Among these, C1 stands out as the most prevalent sub-grade. This distribution suggests that certain sub-grades are more frequently assigned, possibly due to specific criteria or characteristics.

Additionally, a trend analysis reveals that grades B and C are assigned more frequently than lower grades F and G. This indicates a grading trend that favors mid-range grades, which could reflect the overall risk profile of the dataset or the grading criteria used.

Moreover, the distribution of sub-grades appears non-uniform, with noticeable variations in counts across different sub-grades. This variability suggests that there may be differing criteria or considerations for assigning sub-grades, leading to the observed distribution.

**2.** **Analysis of Grade Distribution**



**Observation:**

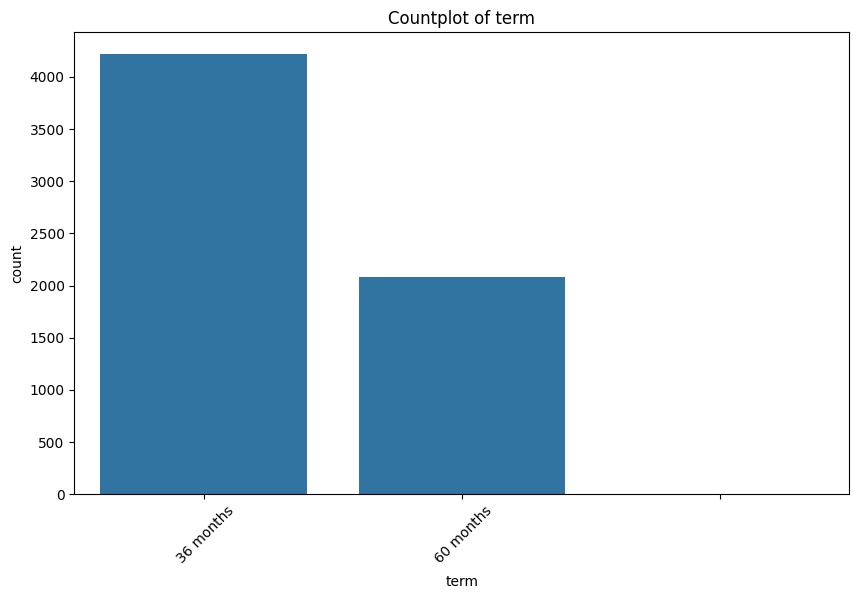
The graph is a bar chart titled “count plot of grade” displaying the frequency of the different grades.

The grade b has the highest count nearing 2000 instances

The grade c follows with just above 1000

The grade E exhibits the lowest count falling below 500, while grades g and f record counts below 750 each suggesting a hierarchical risk distribution within a dataset

**3. Comparative Analysis of Term Frequencies in Loan Data**



**Observations**

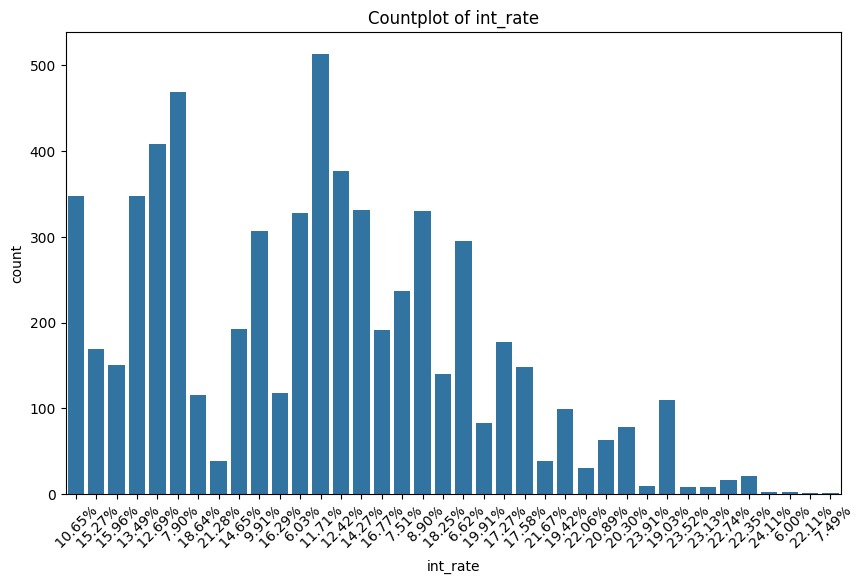
36 months term is significantly more common than 60 months the count for 36 months exceeds 4000

The count for 36 months exceeds 4000 indicating its prevalence

In contrast the count “60 months “ is just above 1500, suggesting its less commonly chosen

Indicates a preference among browsers for the shorter loan term option

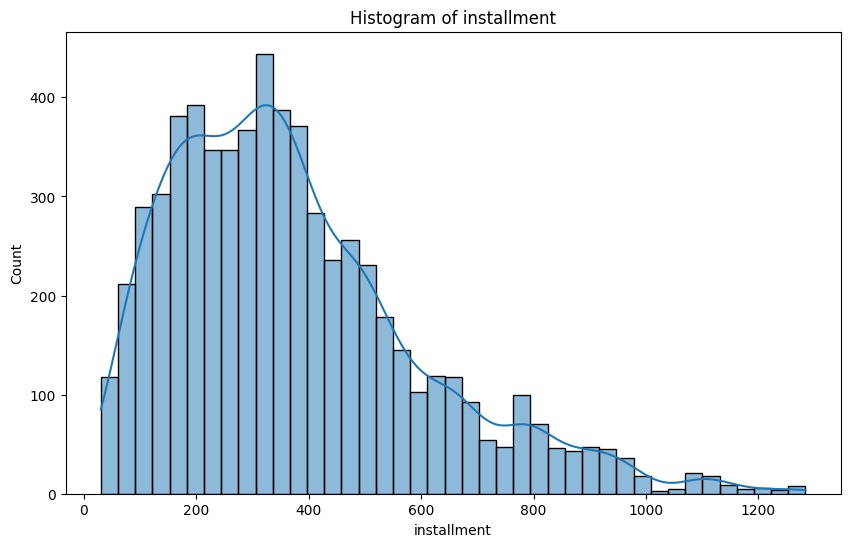
**4. Comparative Analysis of Term Frequencies in Loan Data**



**Observation:**

The graph is a count plot illustrating the frequency of two different loan terms: "36 months" and "60 months." It is evident from the plot that there are significantly more counts for "36 months" compared to "60 months." The count for "36 months" exceeds 4000, indicating that this term is much more common in the dataset. In contrast, the count for "60 months" is just above 1500, suggesting that loans with a term of 60 months are less prevalent.

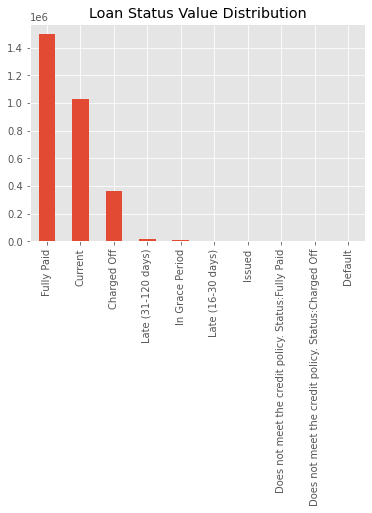
**5. Analyzing Installment Amounts**



**Observations**

The histogram provides a visual representation of the distribution of installment amounts. It is evident that the majority of the data clusters between the 200 to 400 range, indicating that most individuals have installment amounts within this interval. As installment amounts surpass 400, the frequency notably decreases. Installments above 800 are rare, with very few individuals falling into this category. This pattern suggests a concentration of installment amounts in the lower to mid-range values, with a steep decline in frequency as amounts increase.

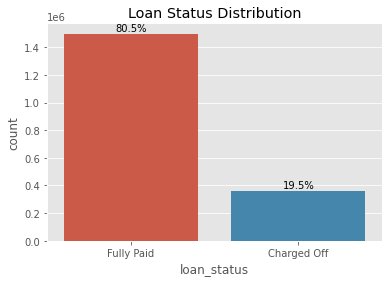
**6.Loan Status Value Distribution**



**Observation**

The bar graph provides an overview of the distribution of loan statuses, which is crucial for assessing the performance and risk associated with a loan portfolio. A significant proportion of loans fall into the "fully paid" category, indicating successful loan repayment and financial stability. Another substantial segment comprises "current" loans, which are active loans with timely payments, reflecting a healthy loan portfolio. The graph may also depict other loan statuses, such as "late payments," "default," or "charged off." Analyzing these categories helps assess risk exposure and potential losses. It's important to note that a well-balanced loan portfolio aims to minimize risk while maximizing returns. Therefore, monitoring these loan statuses is essential for effective financial management.

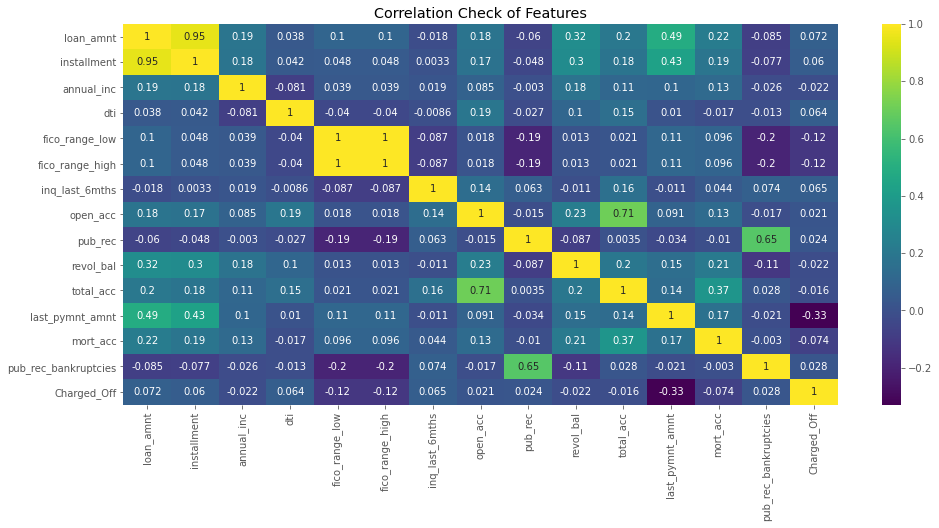
**7.Loan Status Distribution**



**Observation:**

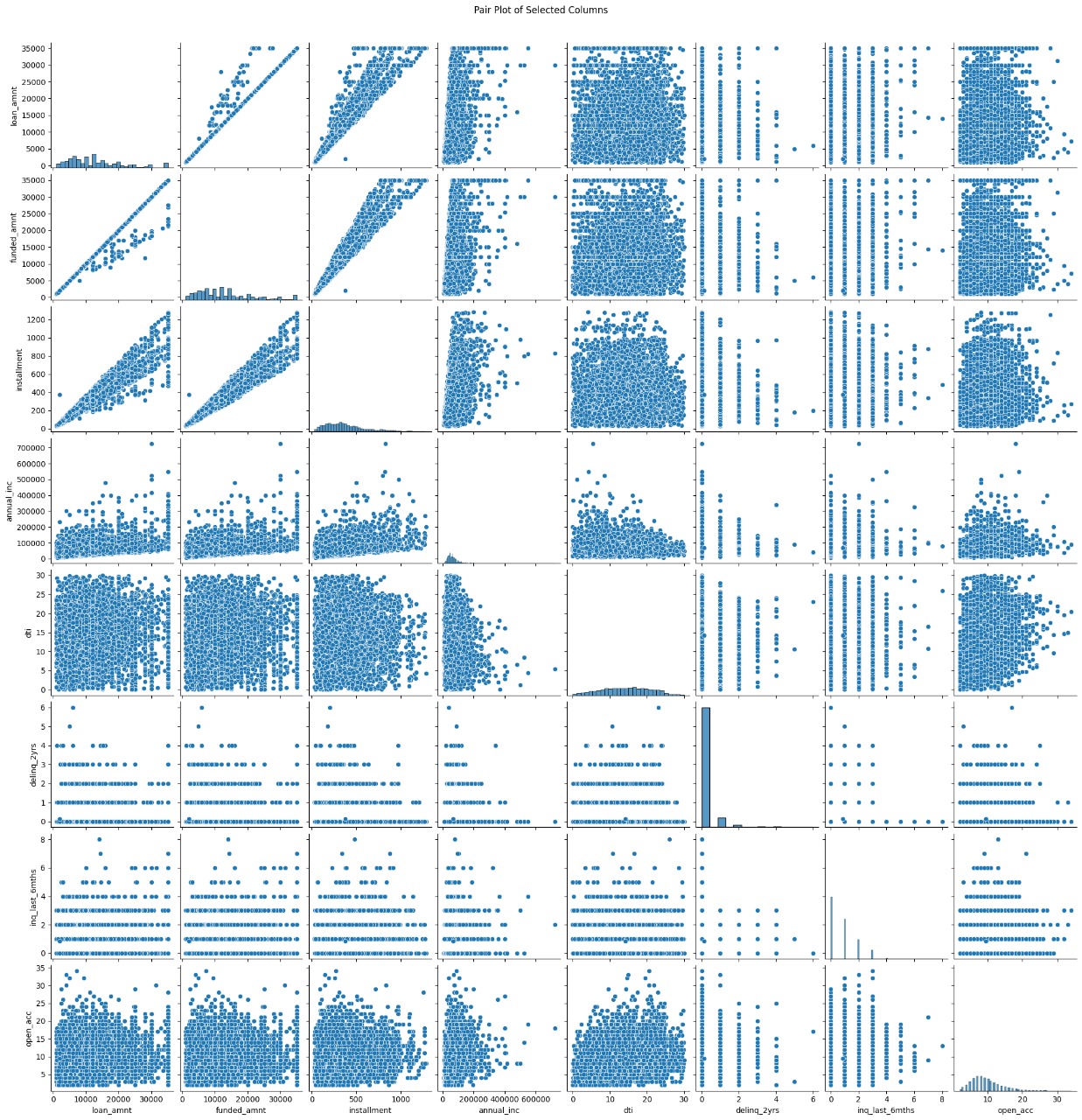
**From the above graph we can say that** that a significant majority of loans, **80.5%**, are fully paid, while a smaller portion, **19.5%**, are charged off. This distribution highlights the difference in loan outcomes and is relevant for assessing loan performance and risk.

**8. Correlation**

 **Observations**

Firstly, there might be a strong negative correlation between interest rates and loan grades, indicating that higher-grade loans (lower risk) could have lower interest rates. This relationship is expected as lenders typically offer lower interest rates to borrowers with lower risk profiles. Secondly, there may be a positive correlation between loan amount and annual income, suggesting that borrowers with higher incomes might qualify for larger loans. This relationship aligns with standard lending practices, where borrowers with higher incomes are often considered more creditworthy and eligible for larger loan amounts. Lastly, investigating correlations between loan status (fully paid or charged off) and other features such as credit history, employment, and debt-to-income ratio could provide valuable insights

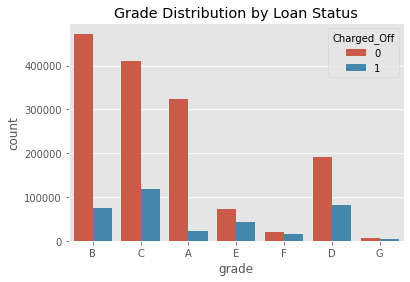
**9. Multivariate Data Analysis and Insights Derived from Scatter Plot Matrices**



**Observations**

The scatter plot matrix offers a comprehensive view of the multivariate data, with each cell displaying a scatter plot of two variables. The diagonal cells provide insights into the univariate distributions, highlighting individual variable characteristics. Some variable pairs exhibit strong linear relationships, as evidenced by clustered points forming clear lines, indicating potential correlations between the variables. In contrast, other pairs show dispersed point patterns, suggesting less obvious or no linear relationships. Additionally, distinct groupings or clusters in some plots suggest potential underlying categories or classes within the data, which could be further explored for segmentation or classification purposes.

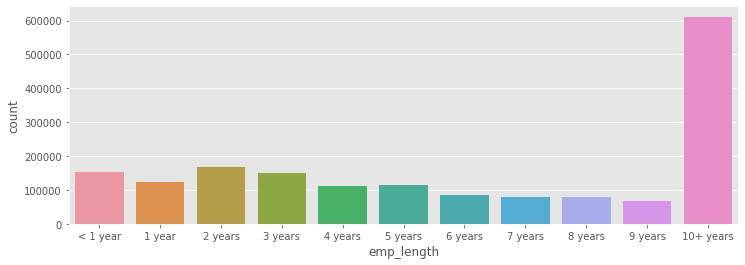
**10. Grade Distribution By Loan status**



**Observation:**

The bar graph illustrates the distribution of loan grades by loan status, specifically focusing on the occurrence of charge-offs. The x-axis likely represents different loan grades (such as A, B, C, etc.), while the y-axis represents the number of loans or loan count for each grade. A clear trend is observed in the graph: as the loan grade increases, the number of loan charge-offs also increases. This suggests that higher-grade loans are more likely to result in charge-offs, contrary to the expectation that higher-grade loans are safer. The occurrence of charge-offs indicates that borrowers have failed to repay their loans, leading to financial losses for lenders. Despite being considered safer, the graph suggests that higher-grade loans are still associated with some level of risk, as evidenced by the presence of charge-offs in these categories.

**11.Frequency of emp\_length:**

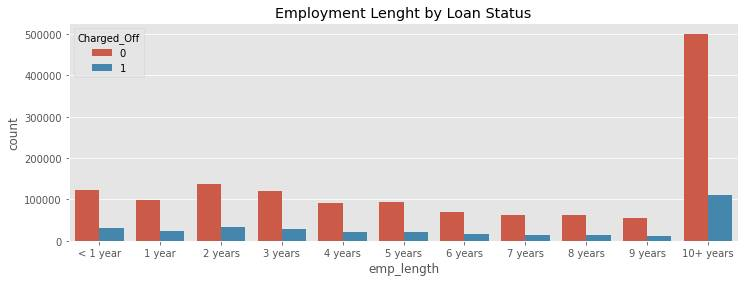


**Observation:**

The majority of people fall into the **0-5 years** employment category.

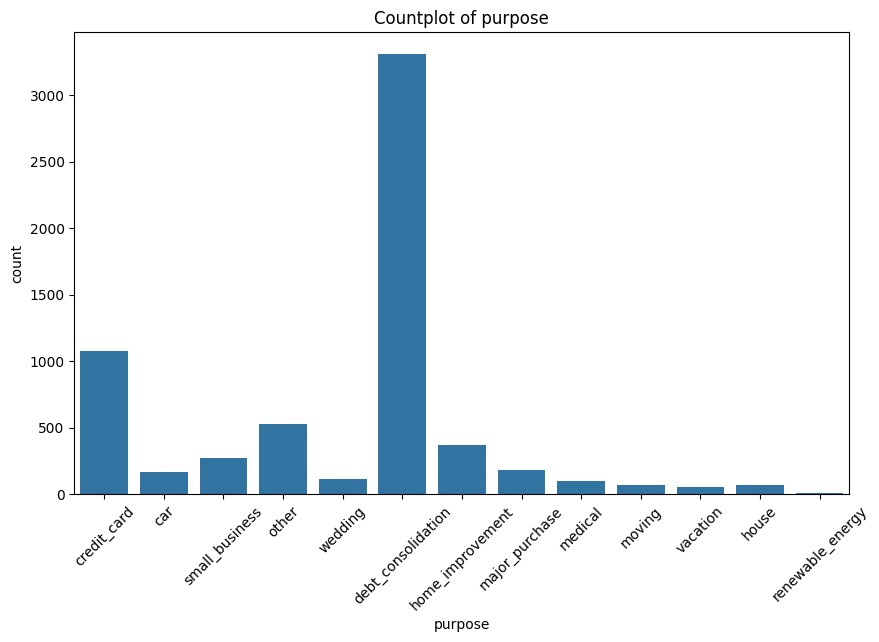
There is a noticeable increase in the number of individuals with **10+ years** of employment experience.

**12. Employment length by the Loan status**



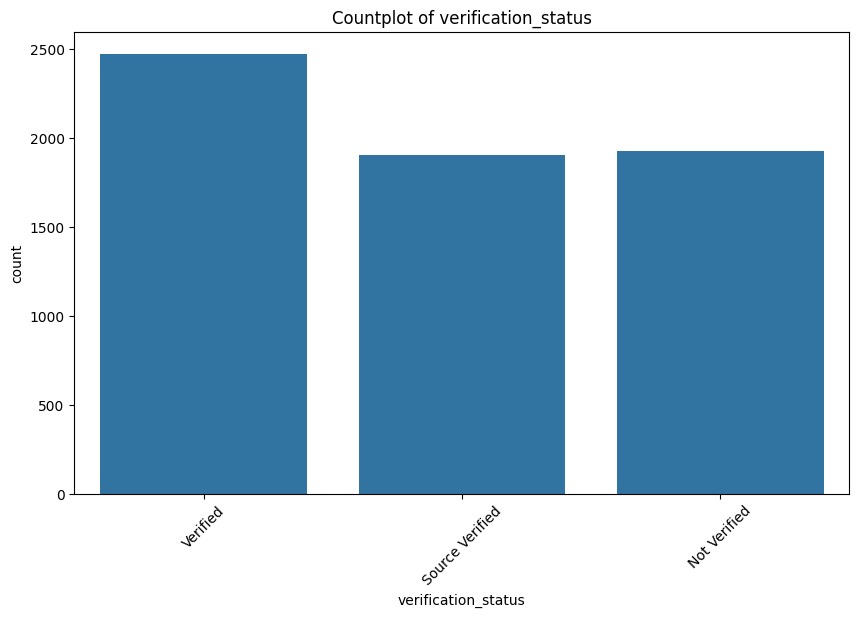
**Observation:**The graph displays the count of loans categorized by employment length and their status as charged off or not. The x-axis represents different employment lengths (e.g., 1 year, 2 years, 3 years, etc.), while the y-axis represents the number of loans. Charged-Off Loans (0) are represented by bars labeled "0," indicating loans where borrowers failed to repay. The number of charged-off loans remains relatively constant across different employment lengths. Non-Charged-Off Loans (1) are represented by bars labeled "1," indicating loans that are not charged off, meaning borrowers successfully repaid. Interestingly, as employment length increases, the number of non-charged-off loans generally increases. There is a notable spike for individuals employed for 10+ years, suggesting that borrowers with longer employment histories tend to have a higher likelihood of successfully repaying their loans. However, it's worth noting that even among longer-tenured employees, some loans still end up being charged off, indicating that employment length is not a definitive predictor of loan repayment success.

**13. Analyzing Loan Purposes**

**11**

**Observation:**The count plot illustrates the distribution of loan purposes, with debt consolidation being the most common purpose by a significant margin. Following debt consolidation, credit card and home improvement purposes are also notable. Less common purposes include wedding, car, small business, major purchase, medical, moving, vacation, house, and renewable energy. This distribution indicates a clear hierarchy in the frequency of loan purposes, with debt consolidation dominating and other purposes trailing behind.

**14. Analyzing Verification Status**

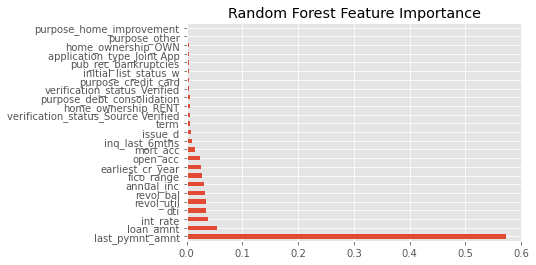


* The countplot shows verification status: Verified, Source Verified, and Not Verified.
* Verified has the highest count, over 2500.
* Source Verified and Not Verified have similar counts, around 1500 each.

This indicates a significant portion undergoes verification, with a slightly smaller proportion in Source Verified and Not Verified categories

**Model Building, Testing Different Models**

**1.Random Forest Classifier**



0 0.97 0.97 0.97 13156

1 0.88 0.88 0.88 2697

accuracy 0.96 15853

macro avg 0.93 0.93 0.93 15853

weighted avg 0.96 0.96 0.96 15853

**Observations:**

**Class Performance**: The precision and recall scores for both classes are high, indicating a balanced performance in predicting.

**Overall Model Accuracy**: The overall accuracy of the model is 96%, indicating its effectiveness in correctly classifying loan instances into their respective categories.

**F1-Score**: The F1-score, which combines precision and recall into a single metric, is also high for both classes, reflecting the model's robustness in achieving a balance between precision and recall.

**Macro and Weighted Averages**: The macro and weighted averages for precision, recall, and F1-score are consistent and high, indicating strong overall model performance across both classes.

**Feature Importance:**

* 'last\_pymnt\_amnt' emerges as the most significant feature, indicating its strong predictive power in determining the target variable.
* 'loan\_amnt' and 'revol\_bal' follow closely, also contributing significantly to the model's predictive ability.
* Conversely, features like loan purpose, home ownership status, and application type appear to have minimal impact on the model's predictive performance, as indicated by their lower importance scores.

**The Random Forest model accurately predicts loan default and non-default instances with high precision and recall. Key features like 'last\_pymnt\_amnt' and 'loan\_amnt' are crucial in assessing credit risk. While some factors have less impact, the model maintains balanced performance despite class imbalance. Its robustness makes it a reliable tool for credit analysis, though further monitoring and refinement are needed for ongoing accuracy.**