**Project Report**

**Project Title**

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

**Team Leader:**

A.Harshith sai 21R11A0504

**Team Members:**

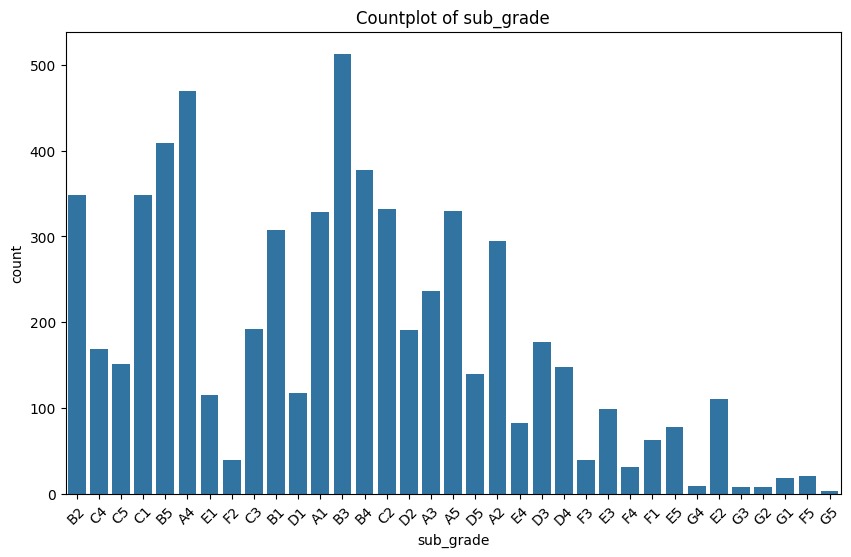
G.Pranava Sai 21R11A0515

G.Harshit 21R11A0519

P.Rohith 21R11A0538

**Exploratory Data Analysis**

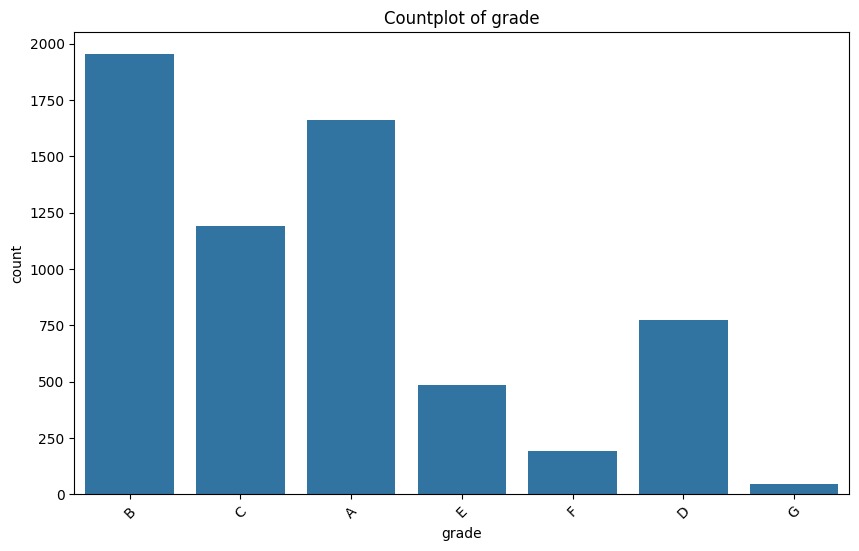
**1. Distribution of Loan Sub-Grades**



**Observations:**

* + The x-axis represents different **sub-grades**.
  + These sub-grades could be related to various aspects, such as academic performance, credit scores, or any other categorical variable.
  + In your specific dataset, the sub-grades range from **B1** to **G5**.
  + The y-axis represents the **count** or **frequency** of each sub-grade.
  + Each bar in the graph corresponds to a specific sub-grade, and its height indicates how many times that particular sub-grade appears in the dataset.
  + Some sub-grades have higher counts, meaning they occur more frequently in the dataset. For example:**B3**, **C1**, and **C2** have relatively higher counts, as indicated by their taller bars.
  + Conversely, some sub-grades occur less frequently. For instance:
  + Grades from **F1** to **G5** have significantly lower counts, as shown by their shorter bars.

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

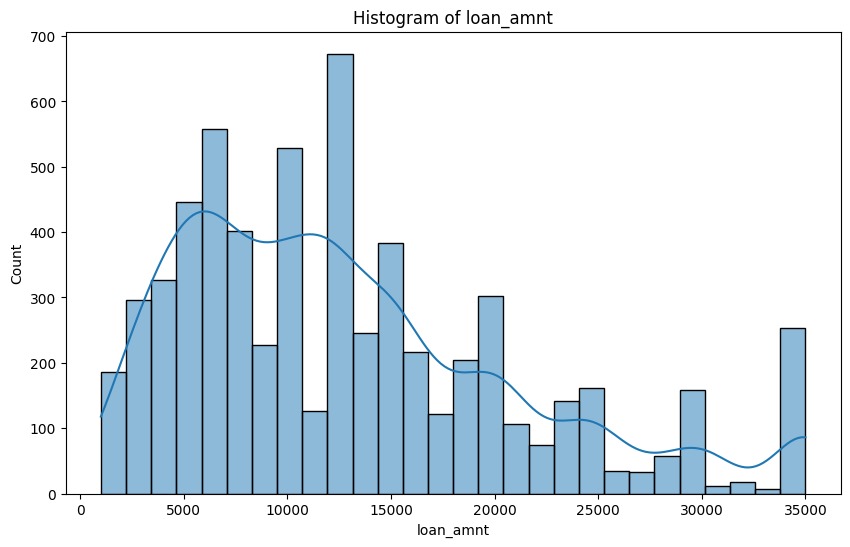


**Observations:**

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

* + The x-axis displays different **grade categories**.
  + However, we can infer that there are **six distinct categories** represented.
  + The y-axis represents the **count** or **frequency** of each grade.
  + Each bar corresponds to a specific grade category, and its height indicates how many times that particular grade appears in the dataset.
  + The first bar (leftmost) is significantly taller than the others, suggesting that this grade category was received most frequently.
  + As we move from left to right, the bars decrease in height.
  + The fourth category has the **lowest count** (shortest bar).
  + There’s a slight increase in counts for the fifth and sixth categories, but they are still lower than the initial ones.

**3. Analysis of Loan Amount Distribution**

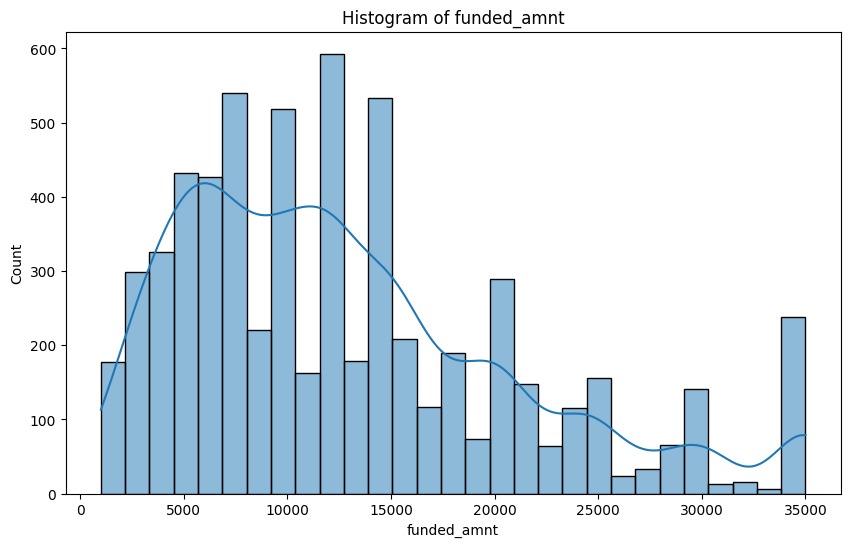


**Observations:**

**The histogram displays the distribution of loan amounts.**

* + The x-axis represents different **loan amounts**.
  + The range of loan amounts spans from **0** to **35,000** units.
  + The y-axis represents the **count** or **frequency** of loans issued at each amount.
  + Each bar corresponds to a specific loan amount, and its height indicates how many times that particular loan amount appears in the dataset
* The histogram bars are uneven in height.
* The first bar (leftmost) is significantly taller than the others, suggesting that loans around **10,000 units** were issued most frequently.
* There’s another peak in loan counts around **15,000 units**.
* Overall, the distribution shows that loans are commonly issued at these specific amounts.

**4. Analysis on the Distribution of Funded Amounts**

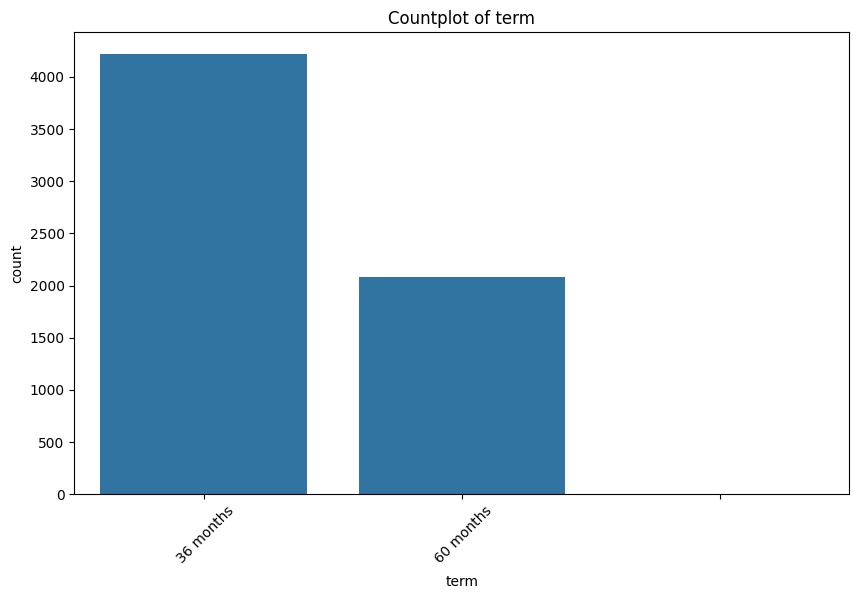


**Observations:**

**The histogram depicts the distribution of funded amounts.**

* + The x-axis represents different **funded amounts**.
  + The range of funded amounts spans from **0** to **35,000** units.
  + The y-axis represents the **count** or **frequency** of loans issued at each funded amount.
  + Each bar corresponds to a specific funded amount, and its height indicates how many times that particular funded amount appears in the dataset.
  + The histogram bars are uneven in height.
  + The first bar (leftmost) is significantly taller than the others, suggesting that loans around **10,000 units** were issued most frequently.
  + There’s another peak in loan counts around **15,000 units**.
  + Overall, the distribution shows that loans are commonly issued at these specific amounts

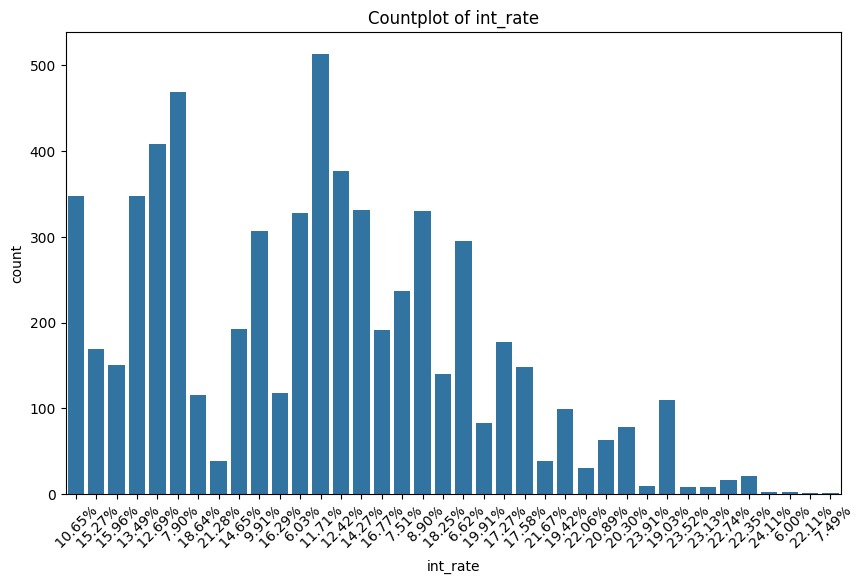
**5. Comparative Analysis of Term Frequencies in Loan Data**



**Observations:**

**The countplot illustrates the frequency of two loan term options: "36 months" and "60 months".**

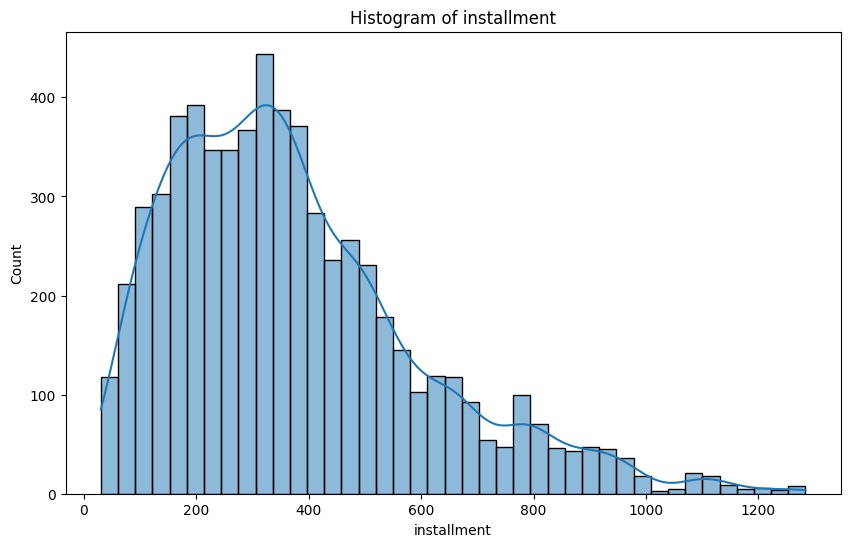
1. 36 months" term is significantly more common than "60 months".
2. The count for "36 months" exceeds 4000, indicating its prevalence.
3. In contrast, the count for "60 months" is just above 1500, suggesting it's less commonly chosen.
4. This distribution indicates a preference among borrowers for the shorter loan term option.
   * 6. Comparative Analysis of Term Frequencies in Loan Data



**Observations:**

* + This is a **bar graph** that visualizes the distribution of interest rates (**int\_rate**) in a dataset.
  + The x-axis represents various interest rate values, while the y-axis shows the count or frequency of each rate.
  + Each bar corresponds to a specific interest rate, and its height indicates how frequently that rate occurs in the dataset.
  + **Peaks**: There are peaks at certain interest rates, suggesting that these rates are more common.
  + **Variability**: Some interest rates occur more frequently than others, leading to varying bar heights.
  + **Prominent Rate**: One specific interest rate stands out with a high count.

**7. Analyzing Installment Amounts**

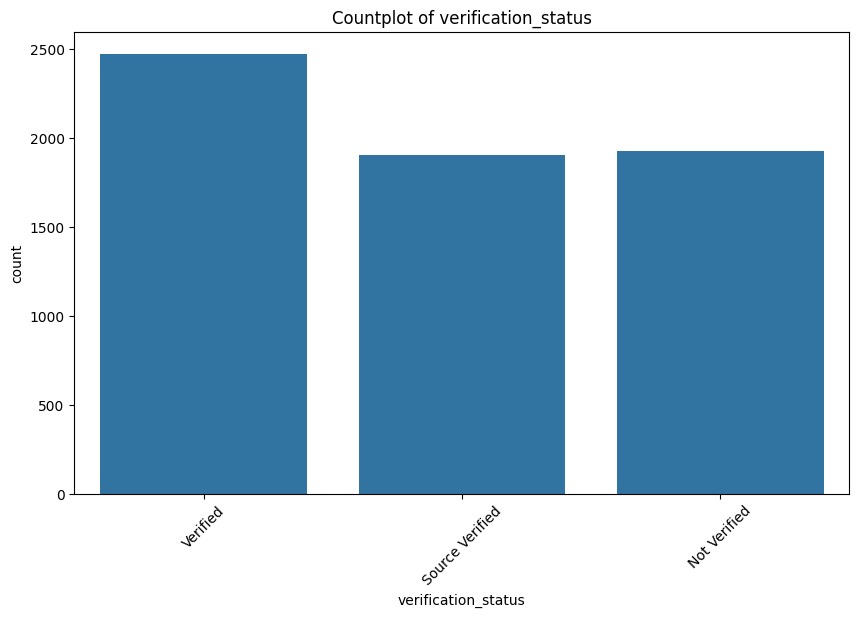


**Observations:**

**The histogram displays the distribution of installment amounts.**

* + This is a **bar graph** that visualizes the distribution of home ownership types among a population.
  + The x-axis represents different types of home ownership: **RENT**, **OWN**, and **MORTGAGE**.
  + The y-axis shows the count or frequency of each type of home ownership.
  + Each blue bar corresponds to one of the three categories.
  + **Renters**: The highest count is for the **RENT** category, indicating that a significant number of people in the dataset are renters.
  + **Mortgage Holders**: The **MORTGAGE** category also has a substantial count, suggesting that many individuals have mortgages.
  + **Homeowners**: The **OWN** category has the lowest count, implying that fewer people outright own their homes.

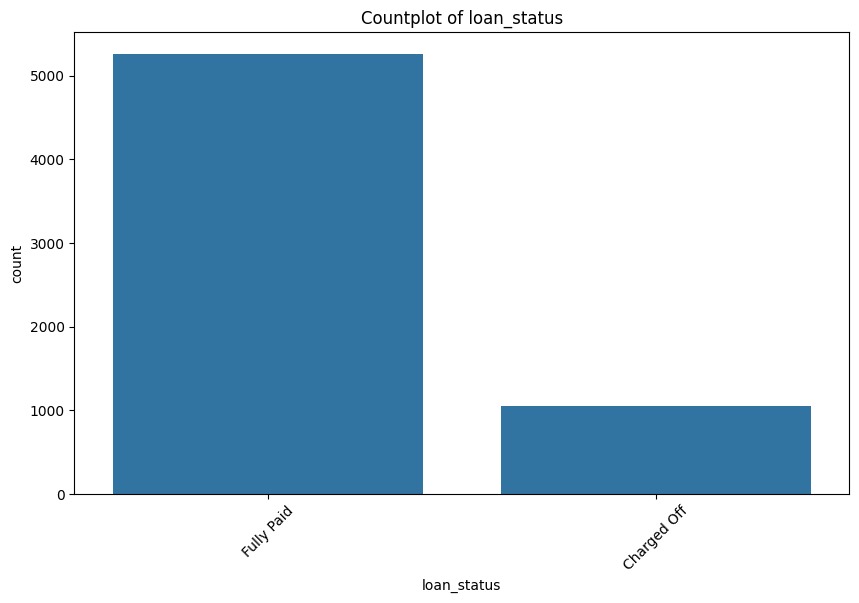
**8. Analyzing Verification Status**



**Observations:**

* The y-axis represents the **count** of items falling into each verification status category.
* **Verified**: Over 2500 items
* **Source Verified**: Approximately 2000 items
* **Not Verified**: Around 2000 items
* The majority of items are **verified**, followed by **source verified**, and a similar count of items falls under the **not verified** category.

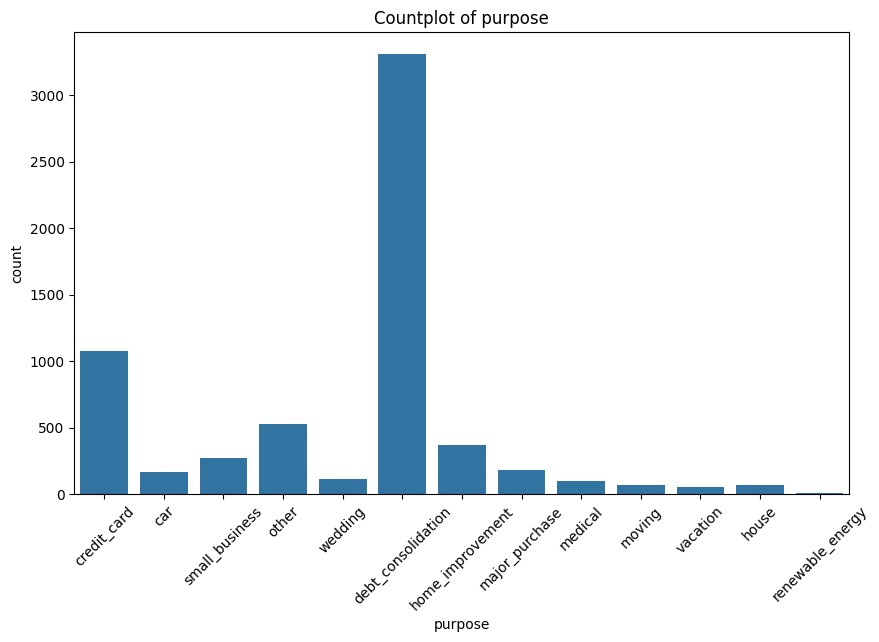
**9 Analyzing Loan Repayment Efficiency**



**Observations:**

* The graph represents a countplot of loan statuses, specifically comparing the number of loans that have been “Fully Paid” versus those that have been “Charged Off.”
* A significant majority of the loans, over 5000 in count, have been fully paid.
* A smaller number, less than 1000, have been charged off.
* This indicates a high rate of loan repayment and a lower incidence of default or charge-offs.

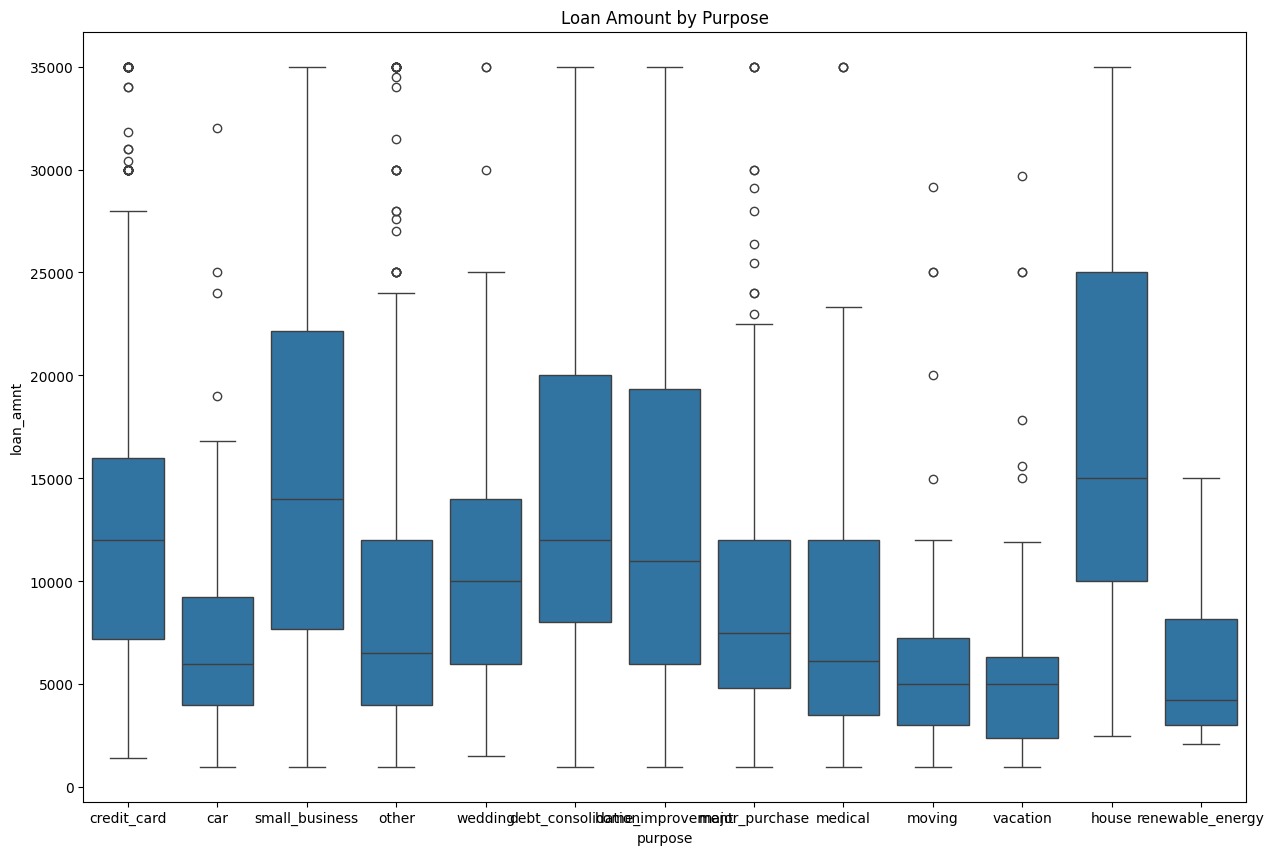
**10. Analyzing Loan Purposes**



**Observations:**

* The countplot depicts loan purposes with debt consolidation being the most common 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 others trailing behind.

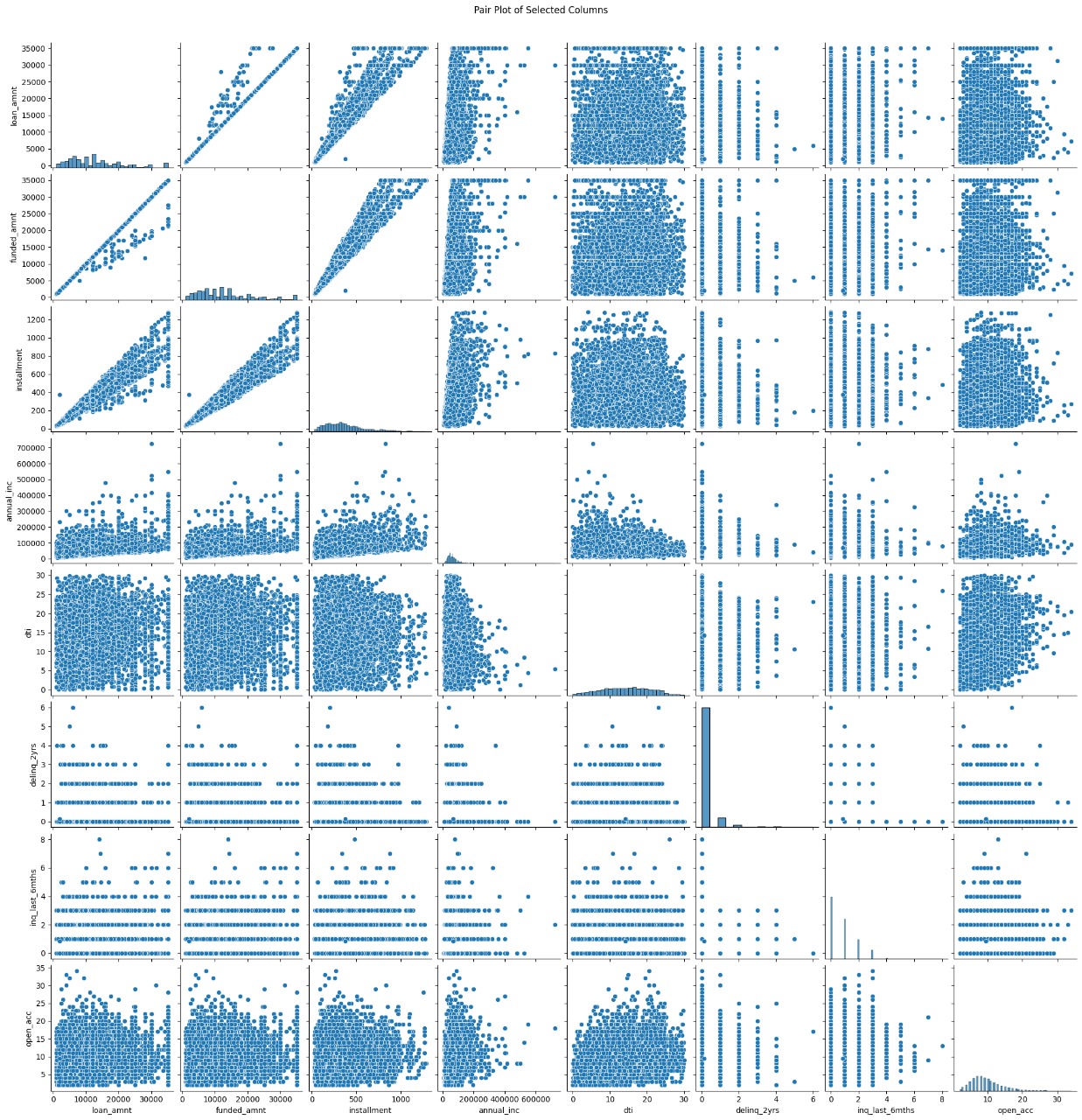
**11. Analyzing Loan Amount Distribution Across Various Purposes**



**Observations:**

* The graph shows loan amounts distributed by purpose.
  + **X-Axis Categories (Loan Purposes)**Credit card
  + Car,Small business,Wedding/debt consolidation/home improvement/purchase/medical purpose (grouped together)
  + Moving,Vacation,House renewable energy
* **Y-Axis**: Loan amount in dollars (ranging from 0 to 35,000)
  + Each category has a blue rectangular box representing the interquartile range (from the first quartile to the third quartile).
  + A line inside the box indicates the median loan amount.
  + Whiskers extend above and below each box, showing variability beyond the upper and lower quartiles.
  + Circles represent outliers.
  + The “house renewable energy” category stands out with a notably higher median loan amount compared to other purposes.
* Renewable energy projects have the highest median amount, followed by debt consolidation and home improvement.
* Debt consolidation displays significant variation with outliers and a wide range.
* Medical, vacation, and moving loans have relatively lower median values.
* Credit card loans show a compact data spread with fewer outliers compared to debt consolidation.

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



**Observations:**

* The scatter plot matrix represents multivariate data with each cell showing a scatter plot of two variables.
* Diagonal cells display univariate distributions, providing insights into individual variable characteristics.
* Some variable pairs exhibit strong linear relationships, evident from clustered points forming clear lines, suggesting potential correlations.
* Other pairs show dispersed point patterns, indicating less obvious or no linear relationships.
* Distinct groupings or clusters in some plots suggest potential underlying categories or classes within the data.