**Loan Approval Prediction**

#### **DATA CLEANING:**

#### 1. Handling Missing Values

* Numerical Columns:
  + Missing values in numerical columns were filled using the mean of the respective columns.
* Categorical Columns:
  + Missing values in categorical columns were filled using the forward fill method. This method propagates the last valid observation to the next missing value.

#### 2. Correcting Data Types

* Certain columns were converted to the categorical data type to accurately represent their nature.

#### 3. Addressing Inconsistent or Erroneous Data Entries

* Self\_employed and Loan\_status columns were standardized to have consistent 'Yes' or 'No' values.

#### 4. Standardizing Formats

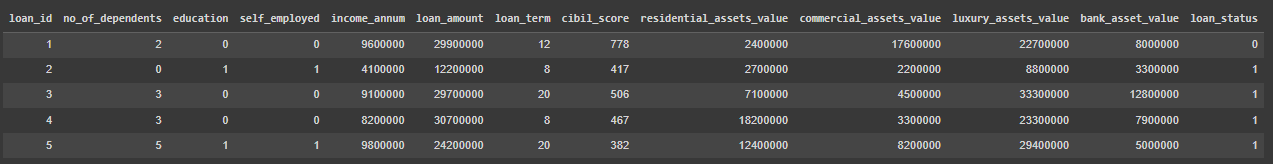
* Monetary values were standardized to integers, ensuring consistent data types for financial attributes.
* Example: If 'income\_annum' had decimal values, they were converted to integers.
* Other columns such as 'cibil\_score' and 'loan\_term' were also ensured to be of integer type for consistency.

#### 5. Label Encoding

* Label encoding was applied to categorical columns to convert them into numerical format for model training..

### **RESULTS:**

* Numerical Columns:
  + Mean imputation helped in retaining data without introducing bias.
  + All numerical columns were free of missing values.
* Categorical Columns:
  + Forward fill method ensured no missing values while maintaining the sequence of data.
  + Categorical columns were converted to the appropriate data types, ensuring correct interpretation by models.
* Binary Columns:
  + Consistent 'Yes'/'No' values improved data quality and reduced errors during analysis.
* Monetary Values:
  + Standardized to integers, ensuring uniformity in financial attributes.
* Label Encoding:
  + Categorical columns were transformed into numerical format, ready for model training.



**OUTLIERS DETECTION AND HANDLING**

#### **1. Handling Outliers**

Outliers were handled based on the provided cap parameter:

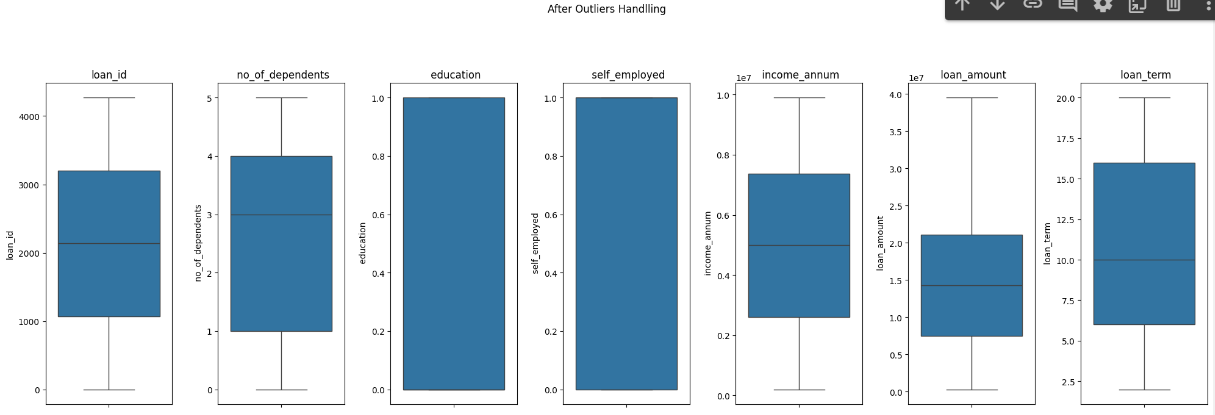
* **Capping Outliers:**
  + If cap is set to True, outliers were capped to the nearest lower or upper bound.
  + This approach retains all data points while limiting the influence of extreme values.
* **Removing Outliers:**
  + If cap is set to False, rows containing outliers were removed from the dataset.
  + This method ensures that extreme values do not skew the data analysis.

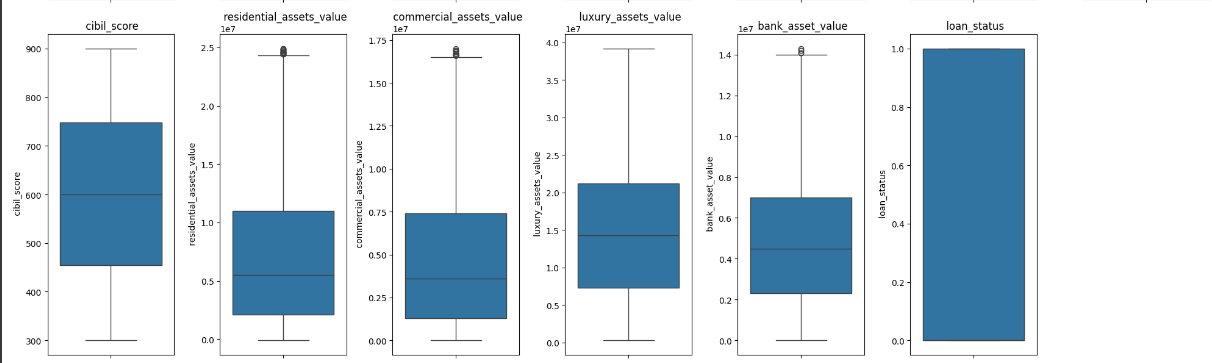
#### **2. Outlier Handling Procedure**

* **Identify Numerical Columns:**
  + The function first identifies columns with numerical data types.
* **Apply Capping or Removal:**
  + For capping: Outliers in each numerical column were capped to the lower or upper bounds.
  + For removal: Rows containing outliers were eliminated based on the IQR method.

### **RESULT:**

#### **Numerical Columns:**

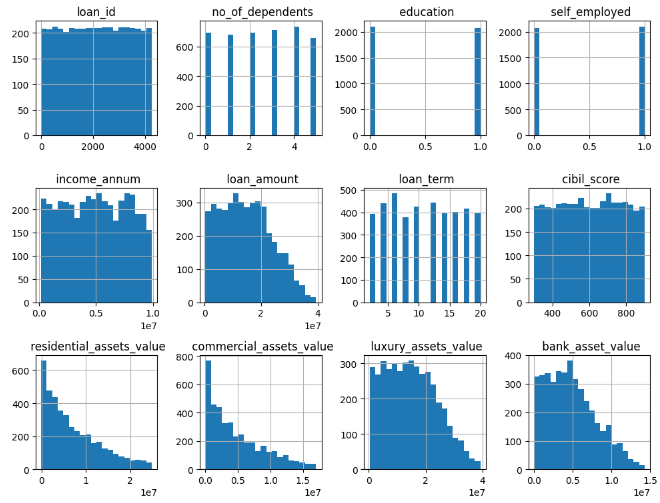
* **Capping:**
  + Outliers were capped to the nearest bound, ensuring that no data points were lost while minimizing the impact of extreme values.
  + This approach maintained the integrity of the dataset, especially in columns where extreme values were expected but needed to be controlled.
* **Removal:**
  + Outliers were removed, resulting in a cleaner dataset with reduced variability.
  + The removal ensured that extreme values did not adversely affect the analysis, providing more reliable results.



**EXPLORATORY DATA ANALYSIS**

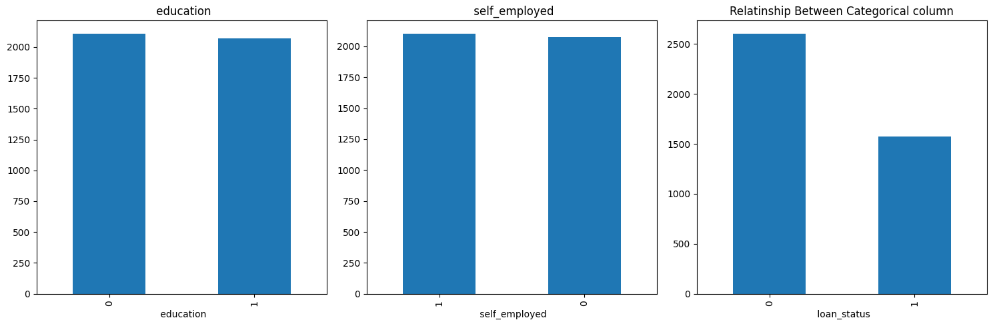
#### **1. Histogram Analysis**

The histograms reveal that several features exhibit right-skewed distributions, such as income\_annum, loan\_amount, and various asset values. This indicates a concentration of lower values with a long tail of higher values. Features like loan\_id and loan\_term are uniformly distributed, and binary features such as education and self\_employed show distinct categories.



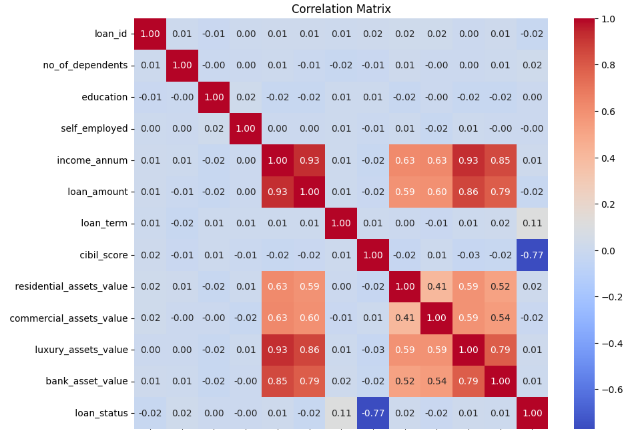
#### **2. Bar Plot Analysis**

The bar plots for categorical variables education, self\_employed, and loan\_status indicate balanced distributions for education and self\_employed. However, the loan\_status variable is imbalanced, with a higher count for one class, suggesting unequal distribution in loan approvals or denials.



#### **3. Correlation Matrix Analysis**

The correlation matrix highlights significant relationships between numerical features. High positive correlations suggest that certain features move together, while negative correlations imply inverse relationships. Key features such as income\_annum and loan\_amount show notable correlations with various asset values, pointing to important drivers in the dataset.



**MODEL EVALUATION:**

* The K-Nearest Neighbors (KNN) model was evaluated using several performance metrics. The test accuracy achieved was 0.5538, indicating that approximately 55% of the predictions were correct.
* The precision score was 0.3858, reflecting that about 39% of the positive predictions were true positives.
* The recall was 0.3111, showing that the model identified 31% of the actual positive cases.
* The F1 score, which balances precision and recall, was 0.3445, highlighting the overall effectiveness of the model in managing the trade-off between precision and recall.

Overall, the model shows moderate performance with room for improvement.