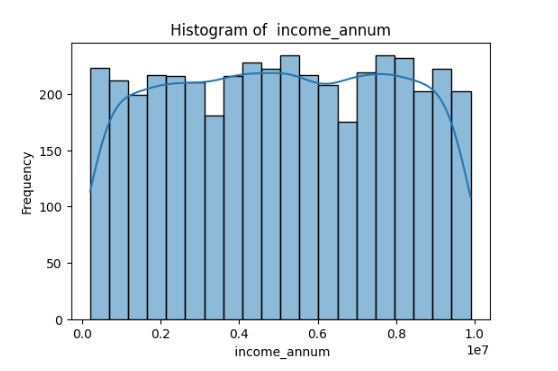
Project \_1

**Introduction of Dataset**: This Dataset contain the records of financial information which determine the eligibility of individuals or organizations to get the loan from institute. It includes features such as cibil score, income, employment status, loan term, loan amount, assets value, and loan status. This dataset is commonly used in machine learning and data analysis to develop models and algorithms that predict the likelihood of loan approval based on the given features.

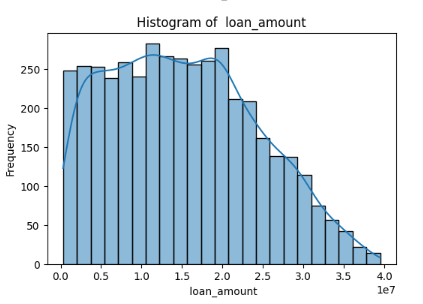
**Data Cleaning:**  To train model firstly need to clean the data. Steps of cleaning dataset such as.

1. load dataset: using pandas library .
2. Handle Missing Value: using mean strategy missing values are handled in data set if it have missing value , but in this data set missin value not found.
3. Outlier detect: outlier detect through IQR method and handle using mean strategy.
4. EDA : this approach highlight the presence of outlier by using box plot visualization and then it make us easy to handle, histogram shows the frequency of distribution of numeric values and box plot clear view of categorical distribution using this approach in this project it make easy to analyse key features or distribution.

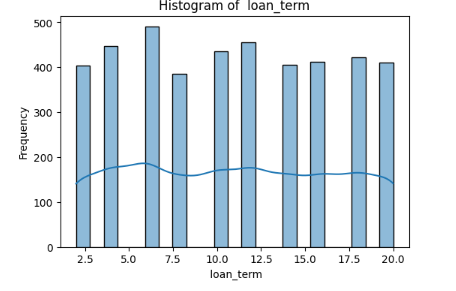


In income\_annum shows the count lies in which range where y axis shows the count and x-axis shows the interval like 0.0-0.1 , 0.1- 0.2 and so on

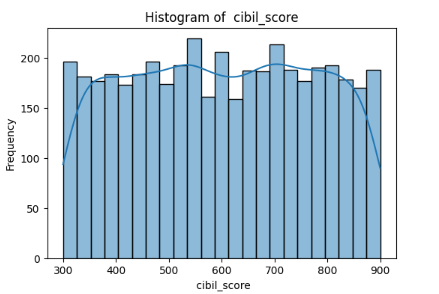
Peaks shows This is **bimodal and symmetrical graph**



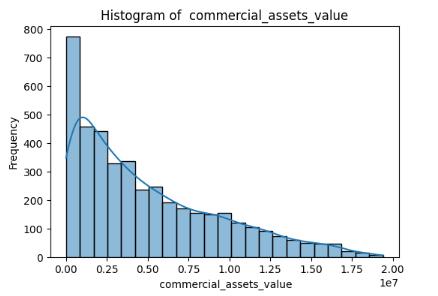
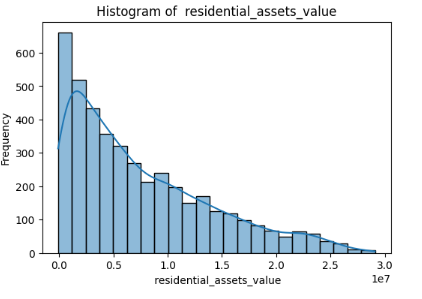
Loan\_amount graph is **skewed right** it show small amount of values are lies in interval [3.5-4] and almost data is lies between [0-3]



Loan term this feature shows the length of time loans can be either long-term or short term loans and this graph shows how many numbers of features lies in short term and long term . And this visualization is  **bimodal and symmetric.**

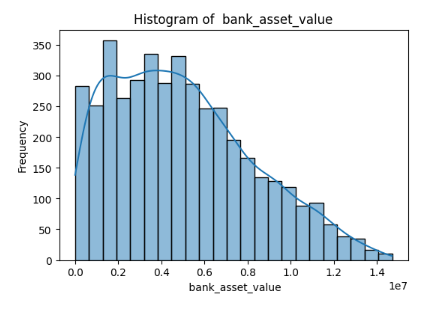
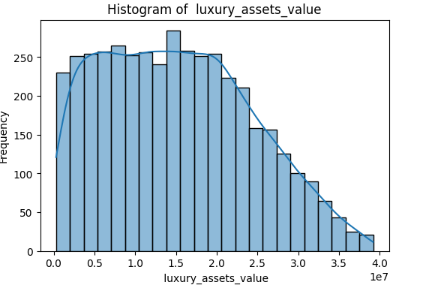


This is **bimodal and symmetric histogram** this graph shows the cibil score which start from 300 to 900 and this graph shows how many features have same cibil score



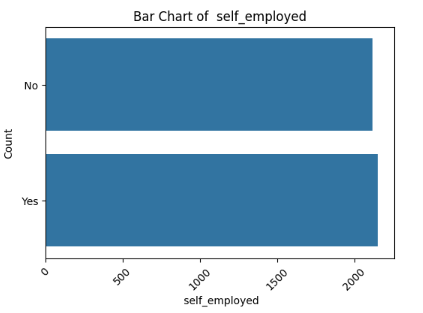
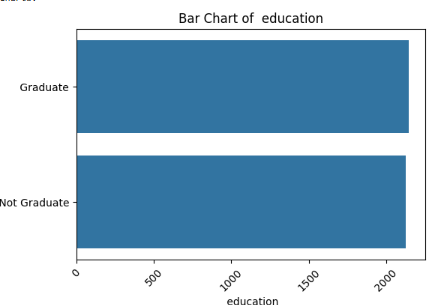
Both These are  **skewed right graphs** which shows the most assets values lies in interval of [0.0-0.5].

And small amount of assets values lies in interval of [1.5-3.0]

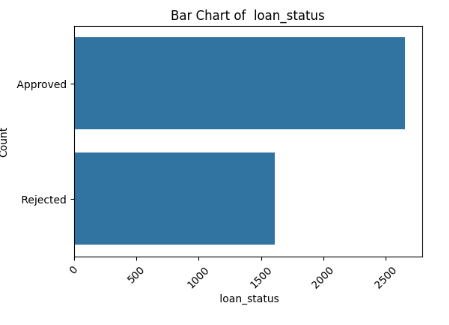


Both These are  **skewed right graphs** which shows the most assets values lies in interval of [0.0-2.5] and [0.0-0.8].

And small amount of assets values lies in interval of [3.0-4] and [1-1.4]

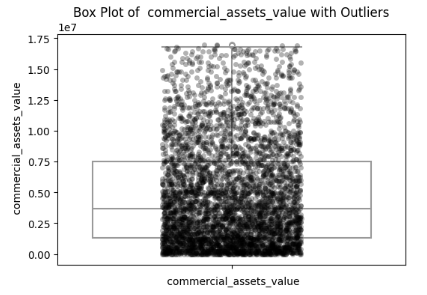
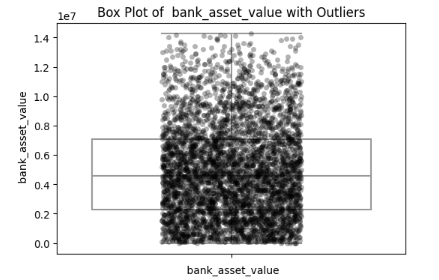
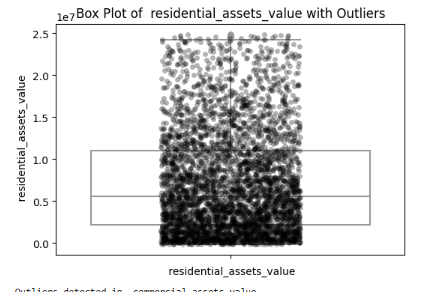


This is bar graph for categorical values in this x-axis shows the frequency and this graph shows the almost equal values are self employed and not self employed. And are graduate and not graduate

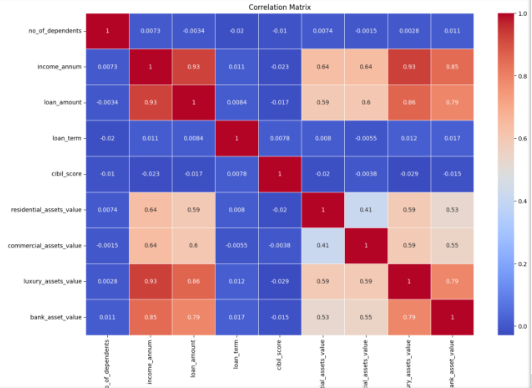


This is bar graph for categorical values in this x-axis shows the frequency and this graph shows the range of approval and rejected number of loans.

**Outlier detection:**  outlier detect in three features in bank, commercial and residential assets values



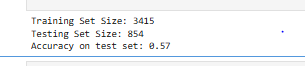
Correlation matrix Is used here to see the relation between features whether the relation is positive or negative or have no relation.



**Manual Data Splitting**

in manual splitting without using the library of train and test the data is split into 80/20

And size of rain and test is ..



**Model Training**

Using KNN model for data training and testing to train modal training set is used and to evaluate the performance of model test set is used and accuracy score is ..



Which shows the modal performance is poor

**Cross Validation:**

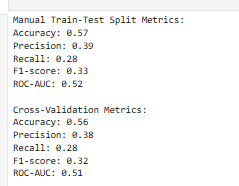
Cross validation technique is used to further train and check performance of model so data is divided into 5 folds where train and test data shuffled to see the performance by changing the parameter and the result of 5 folds and total mean of these 5 folds is this ..

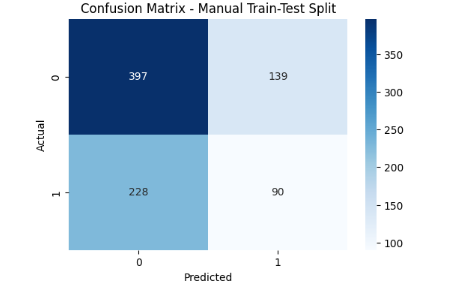
and using cross validation technique performance is not improve

**Model Evaluation**

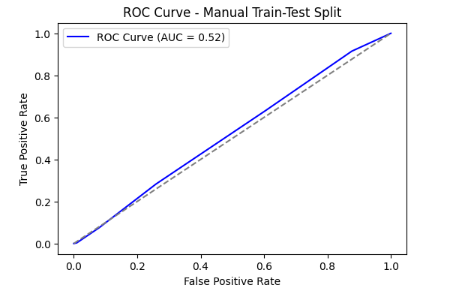
In model evaluation metrics such as accuracy, precision, recall,

F1-score, and ROC-AUC. Of manual split train test using knn and using cross validation technique the performance metric result are..





This confusion metrix shows the score of true positive , true negative false positive and false negative



The ROC curve (blue line) is very close to the diagonal line(doted line), indicating that the model's performance is almost equivalent to random guessing.

The AUC of 0.52 suggests that the model has a very low ability to distinguish between the positive and negative classes. It is only slightly better than a random classifier.