Loan Risk Analytics
EDA case Study

Understanding Data to Mitigate Risks

- Vishal Verma
- Vinti Singh

Introduction

- **What is EDA?**

  A preliminary analysis process to summarize the main characteristics of the data.

- **Why is EDA important for loan risk analytics?**

  Identifies patterns and anomalies.

  Provides insights for feature engineering.

  Improving Risk Stratification.

  Identifying Key Risk Factors.

EDA workflow

Data Understanding:

Inspect the dataset structure (dimensions, types, and summary).

Data Cleaning:

Handle missing values.

Remove duplicates and correct data types.

Outlier Detection and Treatment.

Univariate Analysis:

Examine each variable’s distribution.

Bivariate Analysis:

Explore relationships between independent variables and the target.

Multivariate Analysis:

Feature Engineering and Insights.

emp_length

Data Cleaning

- **Issues Identified:**
  - Removing irrelevant columns as per Data Dictionary sheet.
  - Missing values in employee year of service and public record bankruptcies columns.
- **Resolution:**
  - Imputation techniques (mean, median, mode).
  - Converting dates into numerical formats (e.g., year,month since application).

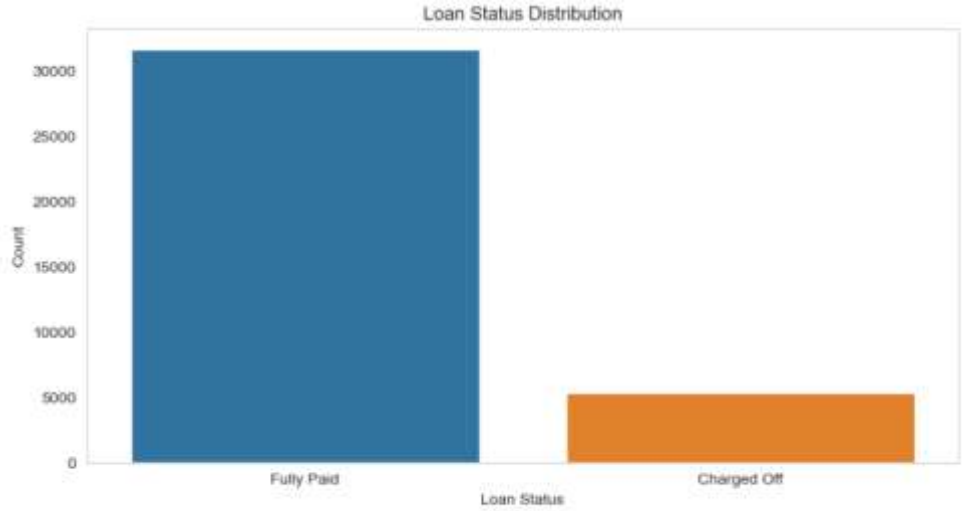
Outlier Treatment

- **Techniques Used:**
  - Boxplots and Histogram to detect anomalies in loan amount, income, etc.
  - IQR methods to quantify outliers.
- **Resolution:**
  - Cap extreme values for variables like income or loan amount.
  - Verify outliers for potential fraud indicators.
  - Standardization techniques .

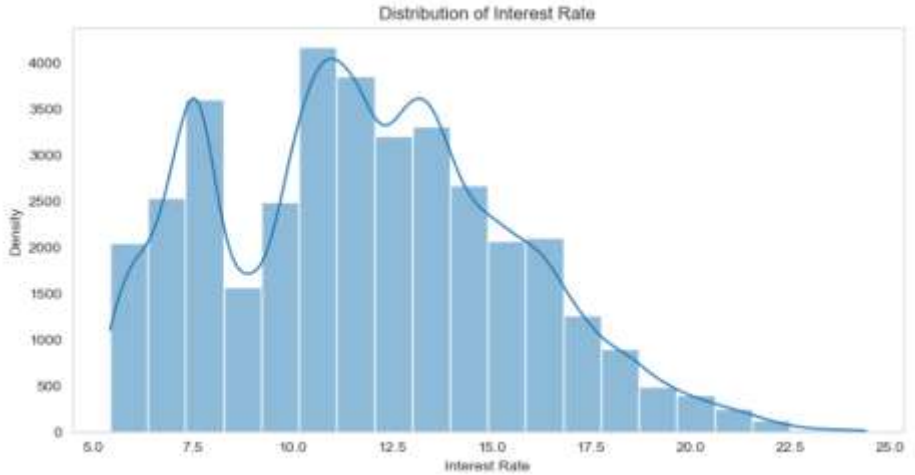
Univariate Analysis

- **Numerical Features:**
  - Distribution of DTI, loan amounts, interest rates, term, verification status, Installment , and income using histograms ,count plot and Bar chart etc.
  - Key statistics (mean, median, standard deviation,info).
- **Categorical Features:**
  - Count plots for loan status, Grade, and ownership of the loan applicant.
- **Insights:**
  - Most loans applied from rented peoples .
  - Majority of applicants fall into the mid-income bracket.

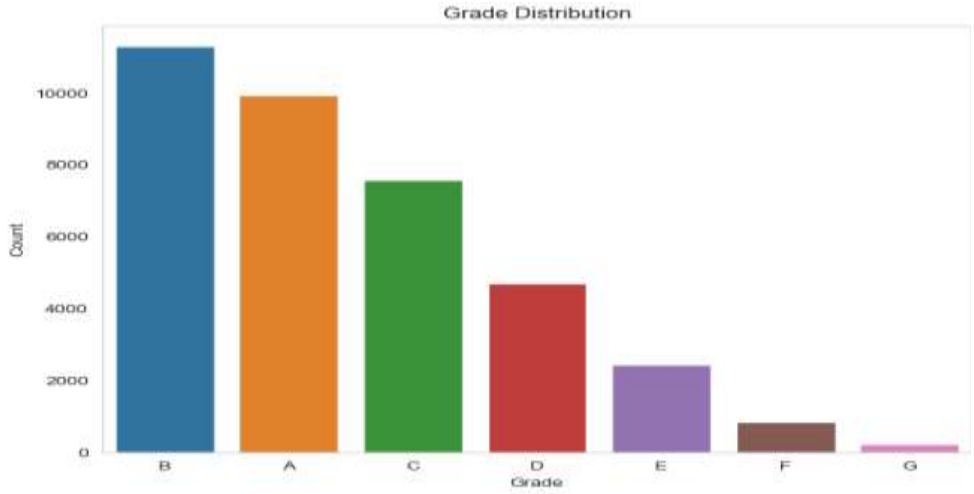
Univariate Analysis



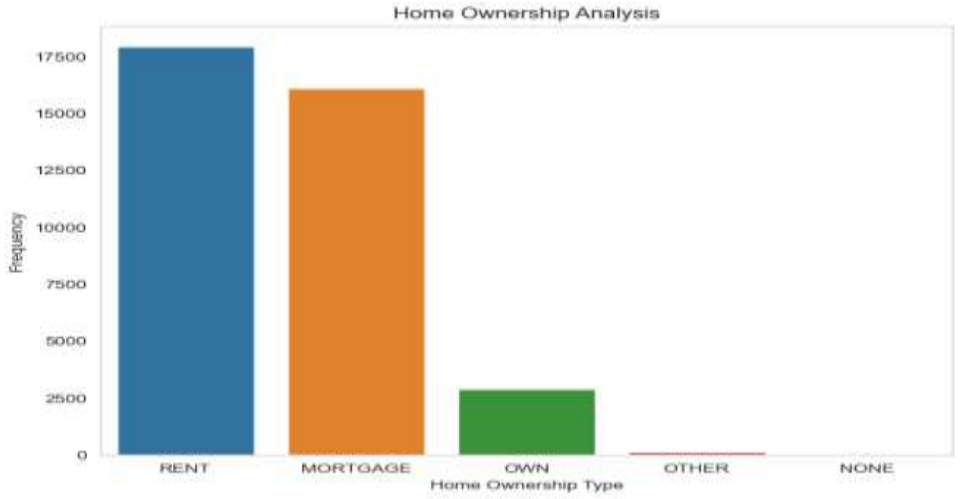
Observation - Less number of loans are defaulted



Obserbation - We found between 5-8% interest and 10-13% being opted more.



Observation - We Found frequency of grade B is higher in comparision and followed by A second and C on thir

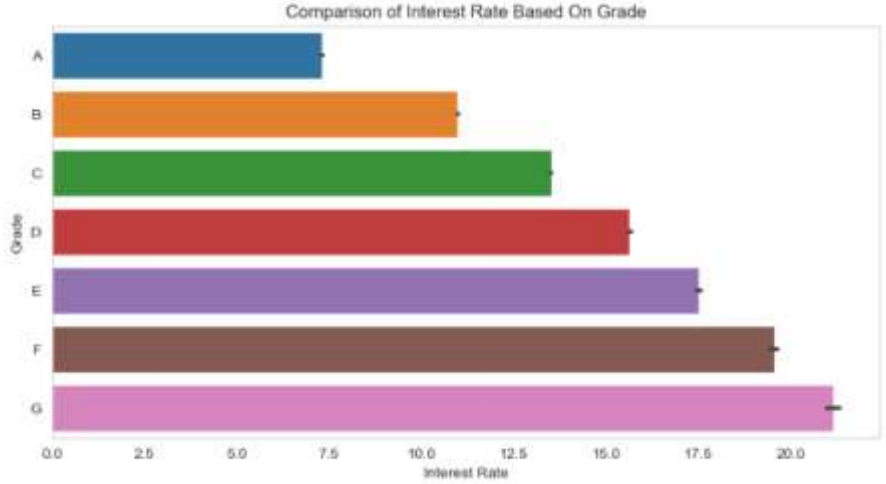


Observation - The highest applicants have either rented or on Mortgage.

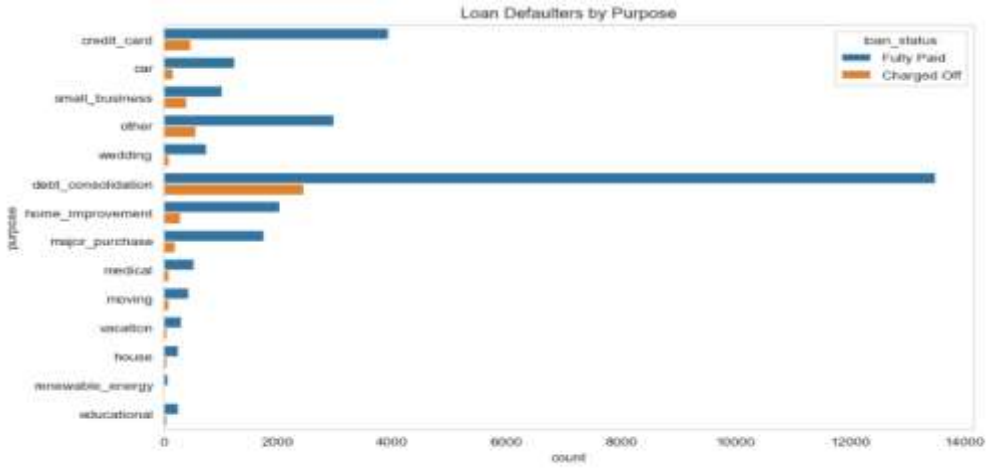
Bivariate Analysis

- **Bivariate analysis b/w Features:**
  - Interest rate vs grade (Bar Plot)
  - Purpose vs loan status (Count Plot)
  - term vs Interest rate (Bar Plot)
  - Loan status vs State(Count Plot)
  - Loan status vs home ownership(Count Plot)
  - Loan status vs loan amount (Dist Plot)
  - Emp Length vs Interest rate (Hist Plot)
- **Key Findings:**
  - Default rates are higher for lower credit scores.
  - Default rates are higher for high rate of interest

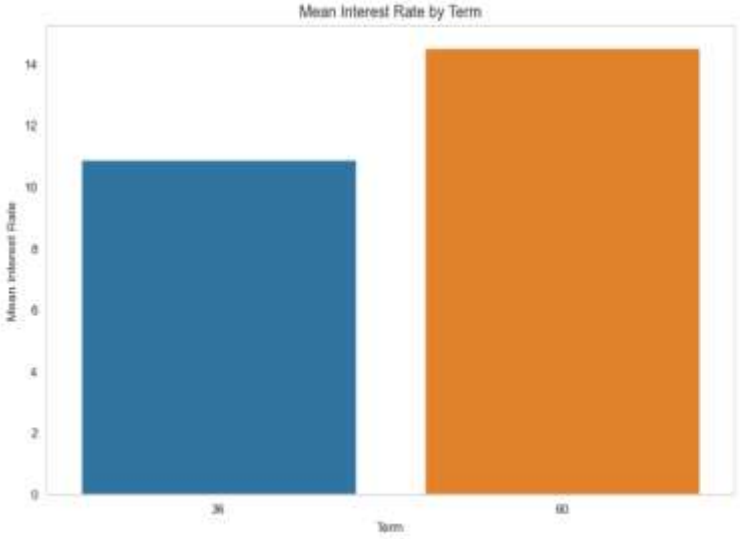
Bivariate Analysis



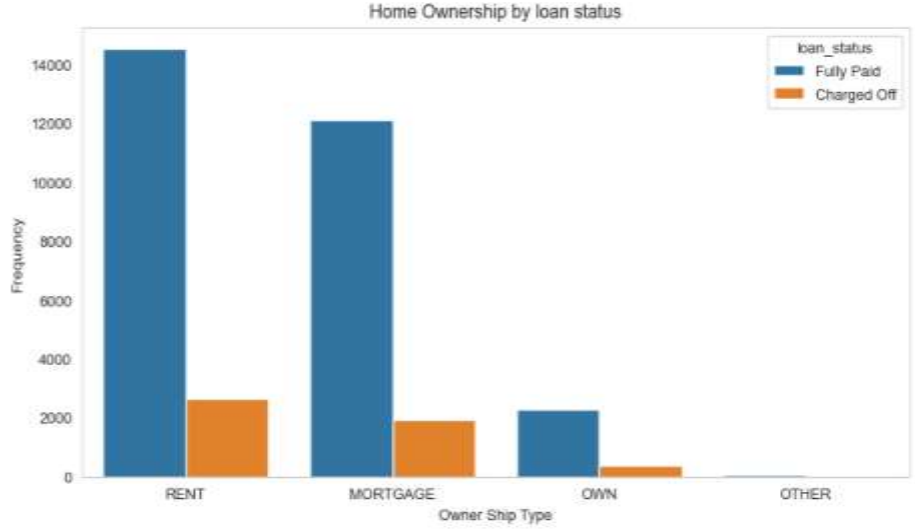
Observation: We found A approx 7.5 has the lowest interest rate and interest rate keep rising until our last Grade G i.e.approx 22%.



Observation - We found debt consolidation has the most of defaulters and fully paid borrower count.



Observation - We have observed that the average interest rate for 36 month term is 10.9676115% and for 60 month it is 14.667568%. So we can conclude the higher the term the interest rate rises.

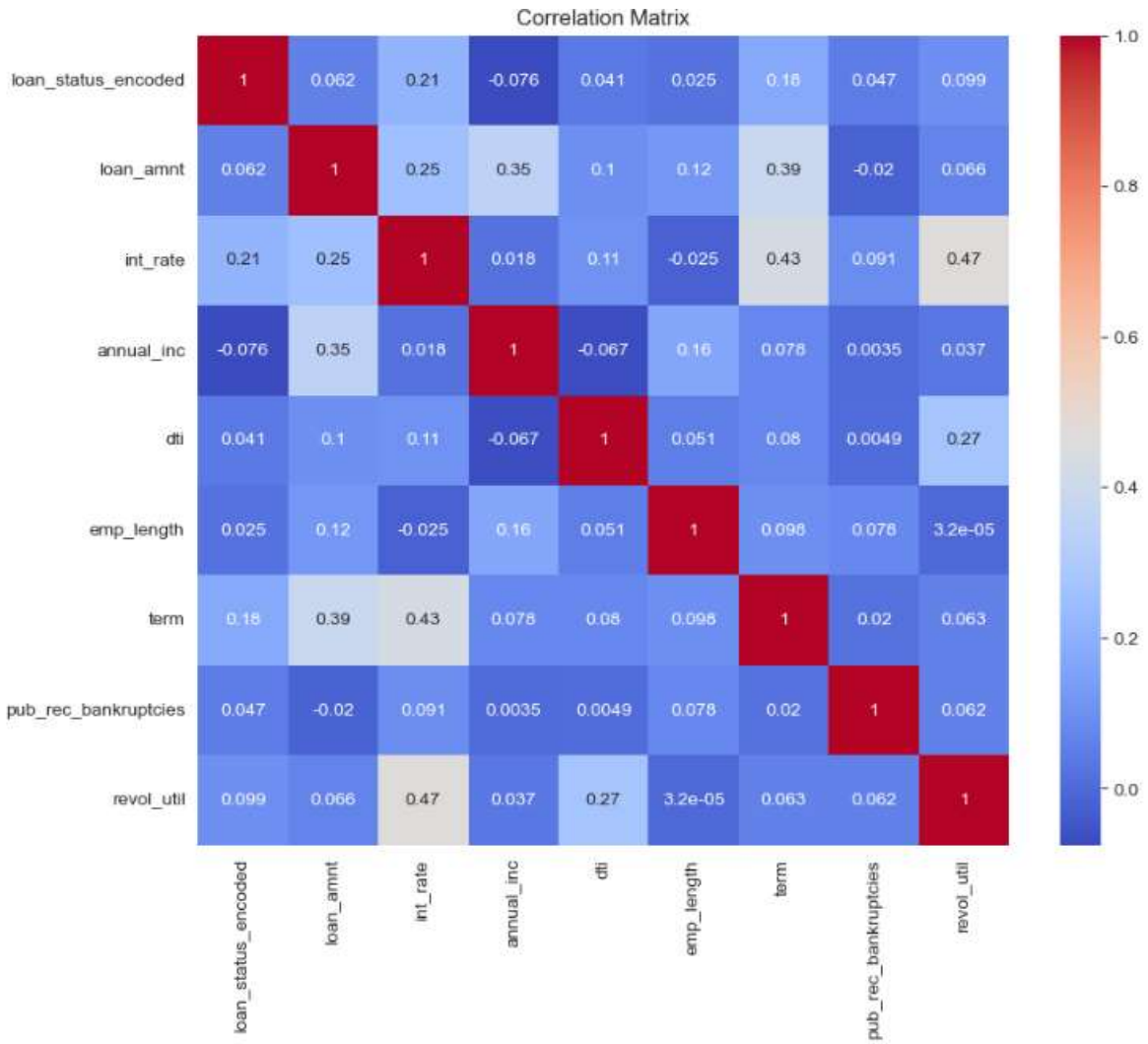


Observation - The highest borrower have either rented or on Mortgage and along with the loan status.

Multivariate Analysis

- **Correlation Heatmap:**
  - Relationship between numerical features (e.g., loan amount, loan status, income, Dti).
- **Key Findings:**
  - Default rates are higher for lower credit scores.
  - High loan amounts with low-income borrowers show increased risks.

Multivariate Analysis



Conclusion

- **Summary of EDA Findings:**
  - Data is ready for advanced modeling after cleaning and preprocessing.
  - Key risk factors have been identified for predictive modeling.
- **Next Steps:**
  - Develop risk prediction models using insights from EDA.
  - Continuously monitor data quality for future analyses.