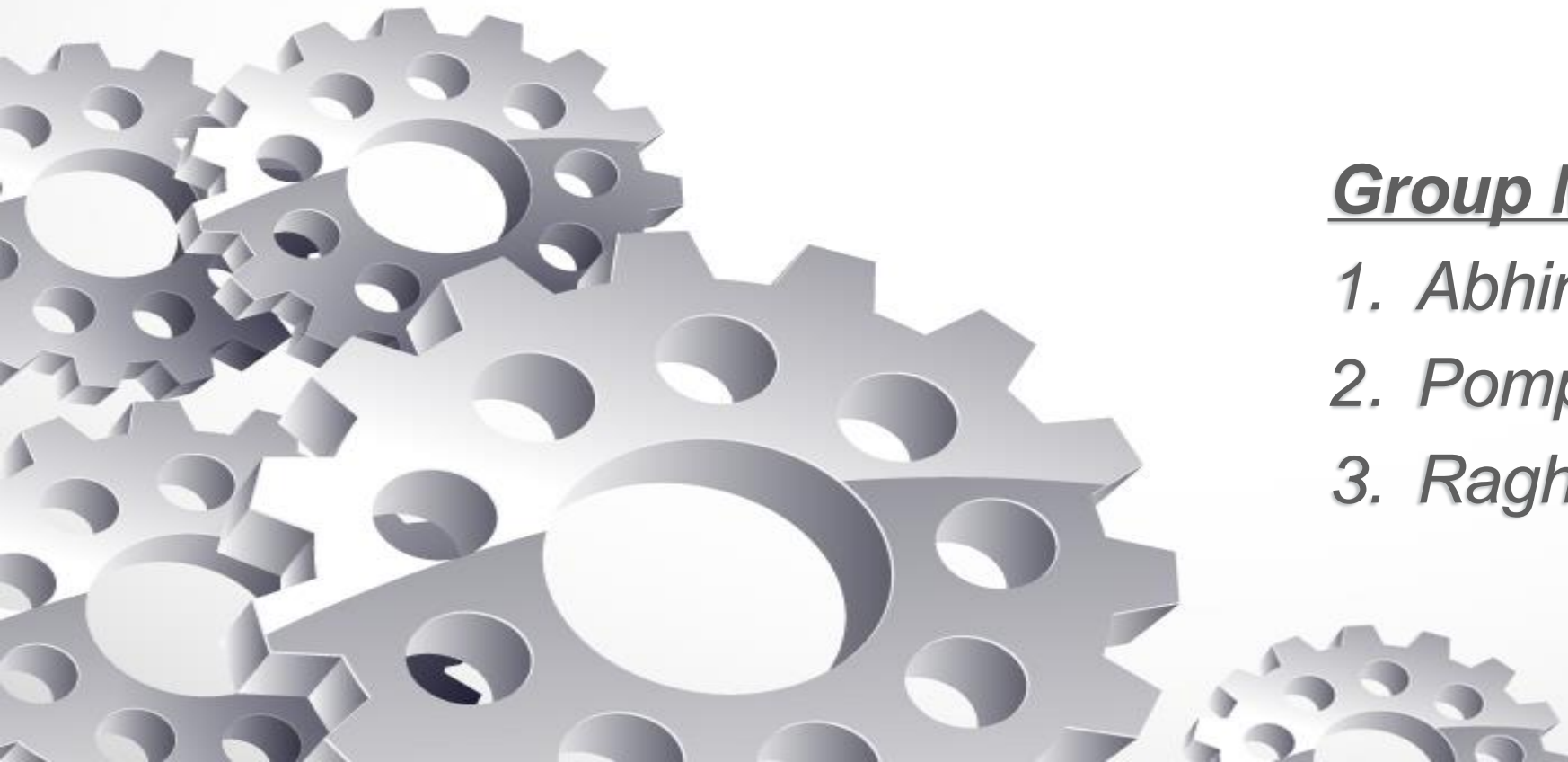


BFSI CREDIT RISK ASSESSMENT

Group Members:

- 1. Abhinav*
- 2. Pompy Mukhopadhyay*
- 3. Raghuveer Vempaty*



PROBLEM STATEMENT



Expected credit loss (ECL) computation is a method used in credit risk management to determine the amount of loss a bank is expected to incur in the event a borrower defaults on their loan. Banks use different methodologies for calculating the expected credit loss (ECL) and provisioning.

Expected credit loss (ECL) = Exposure at default (EAD) x Probability of Default (PD) x
Loss given default (LGD)

ECLs are calculated based on the exposure at default (EAD), probability of default (PD) and the loss given default (LGD) for each borrower. Banks can calculate the ECL for different points in time based on their risk management strategy and regulatory requirements.

For this assignment, we will consider the latest date from which the data is available as the point in time. This means we will estimate the expected credit loss (ECL) for the borrower assuming that the borrower has defaulted at the present point in time.

BUSINESS OBJECTIVE



- In this case study, our business objective is to calculate the LGD(Loss Given Default) and build a model that can predict the loss given default (LGD) for defaulted accounts and evaluate it based on the performance metric that will be described in the subsequent segments.

$$\text{LGD} = \frac{\text{Loan Amount} - (\text{Collateral value} + \text{Sum of Repayments})}{\text{Loan Amount}}$$

- To build a statistical model to estimate LGD of borrower for the defaulted accounts. For this assignment, we are focusing only on LGD component of ECL computation.

Data Cleaning and Pre processing



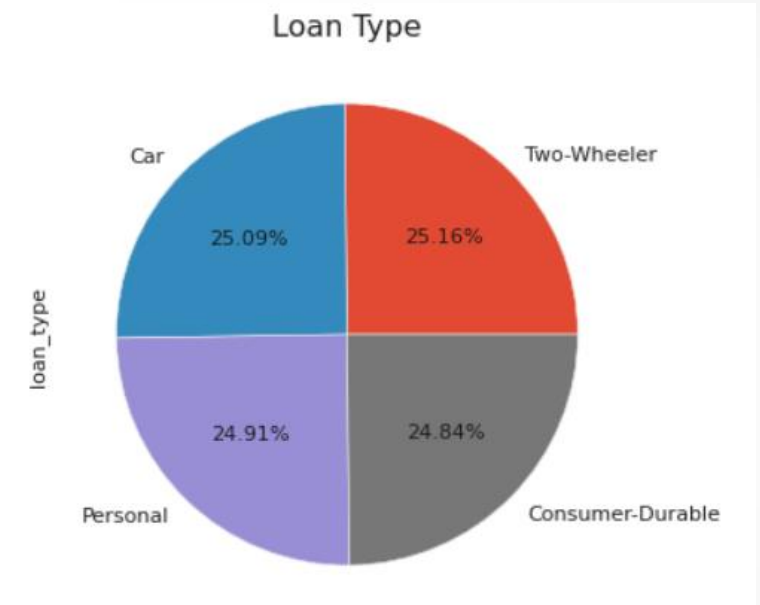
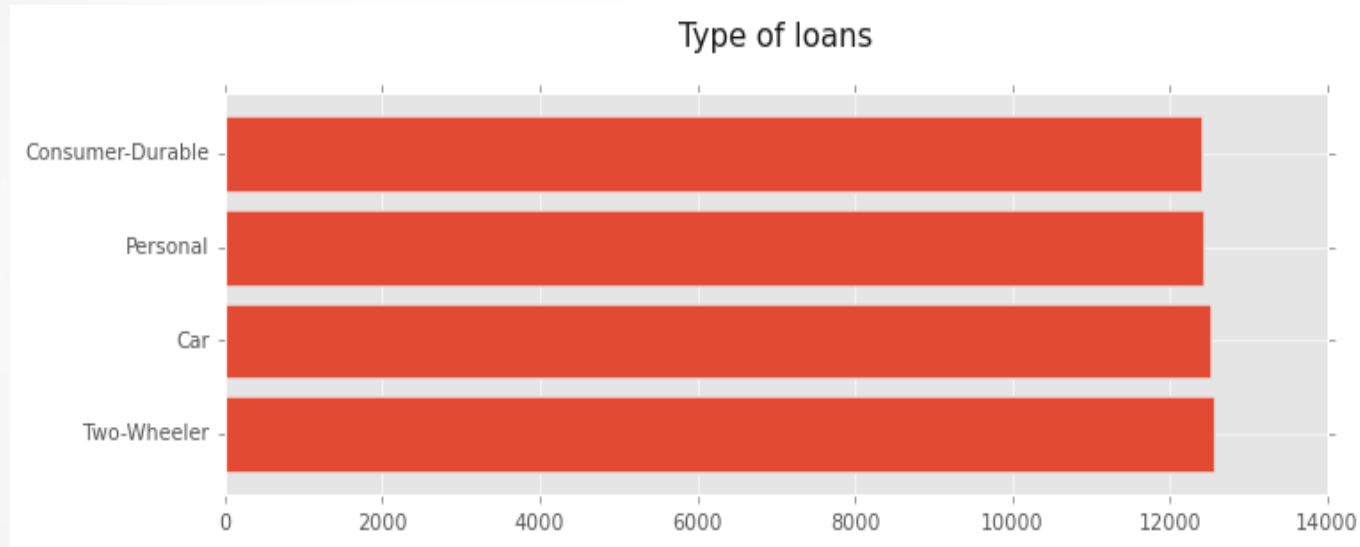
- ☐ **Data Understanding, cleaning and data manipulation.**
 - ☐ **Check and handle duplicate data.**
 - ☐ **Check and handle NAN values and missing values.**
 - ☐ **Imputation of the values.**
 - ☐ **Check and handle outliers in data.**
- ☐ **EDA**
 - ☐ **Univariate data analysis: value count, distribution of variable etc.**
 - ☐ **Bivariate & Multivariate data analysis: correlation coefficients and pattern between the variables etc.**
- ☐ **Feature scaling & dummy variables and encoding of the data.**
- ☐ **Model Selection**
- ☐ **Validation of the model.**

DATA MANIPULATION

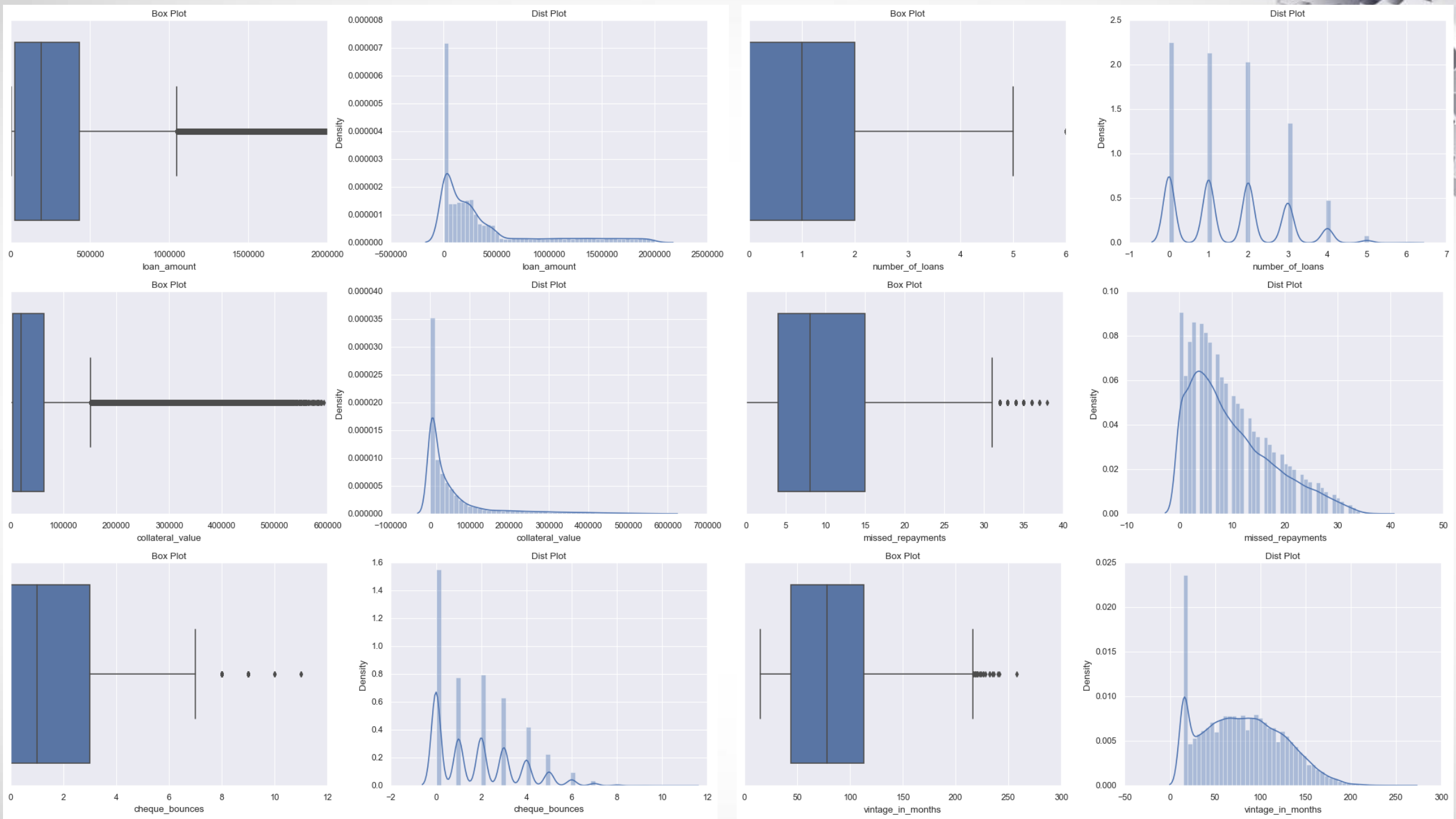


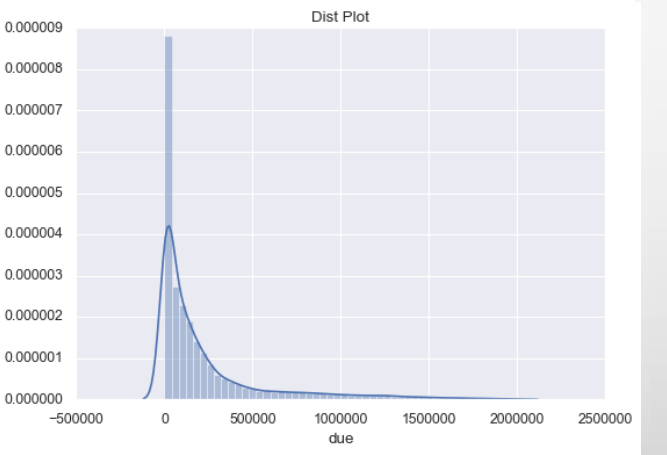
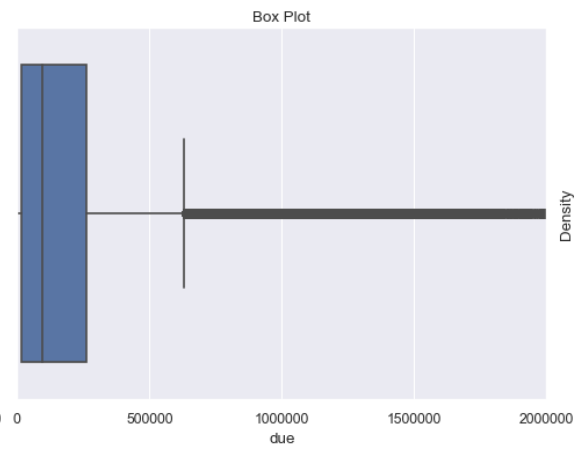
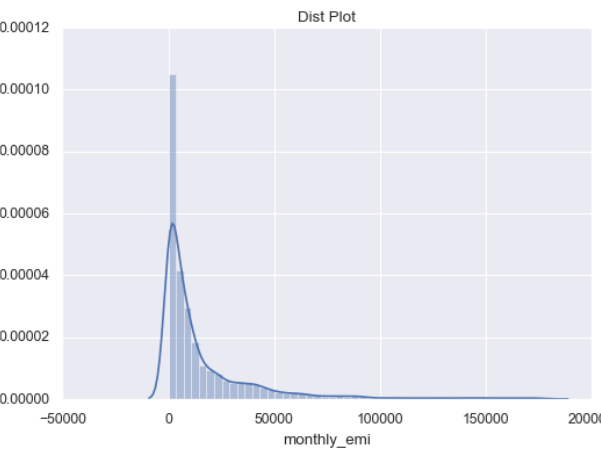
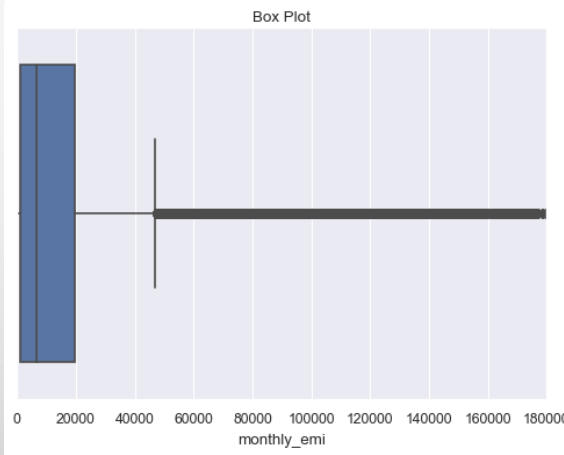
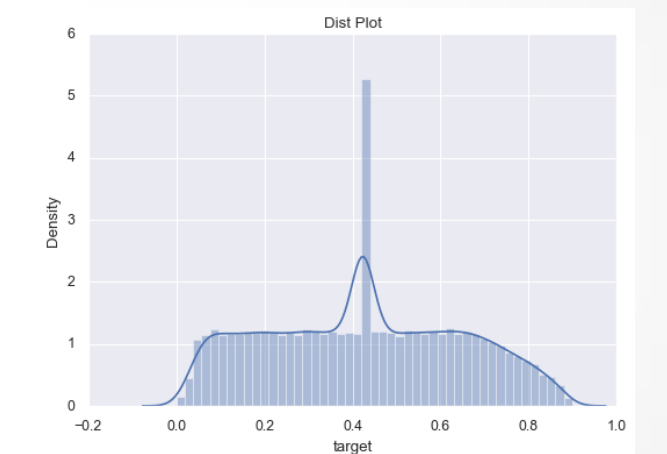
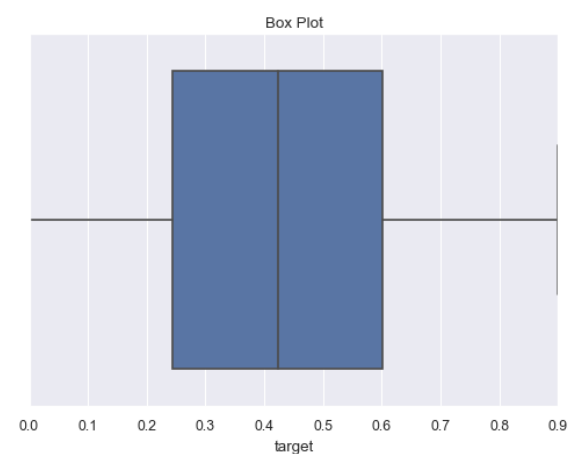
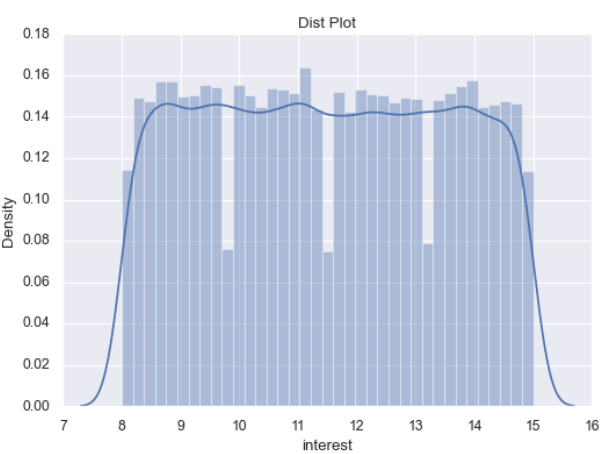
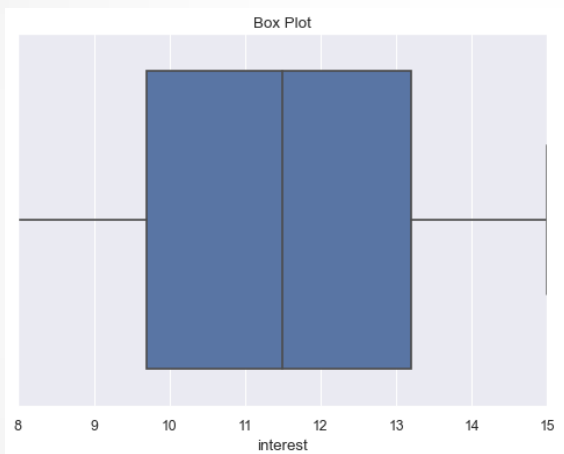
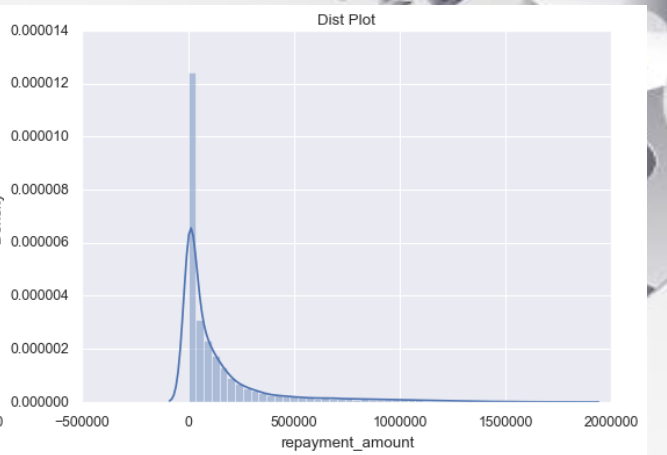
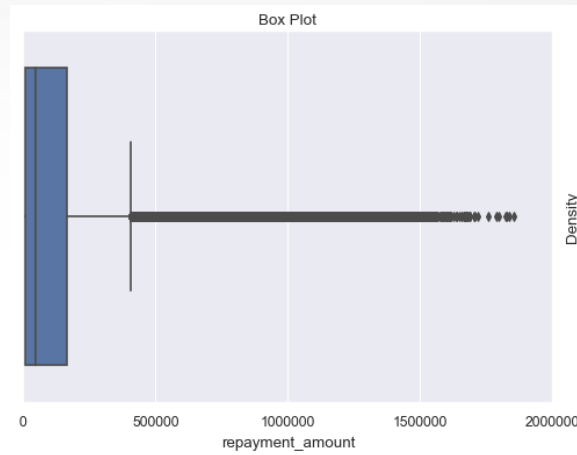
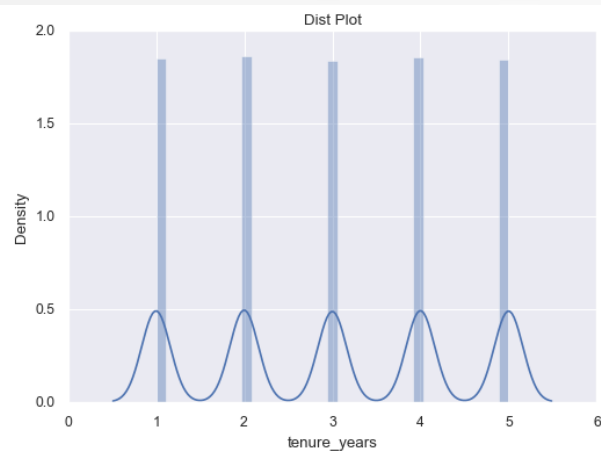
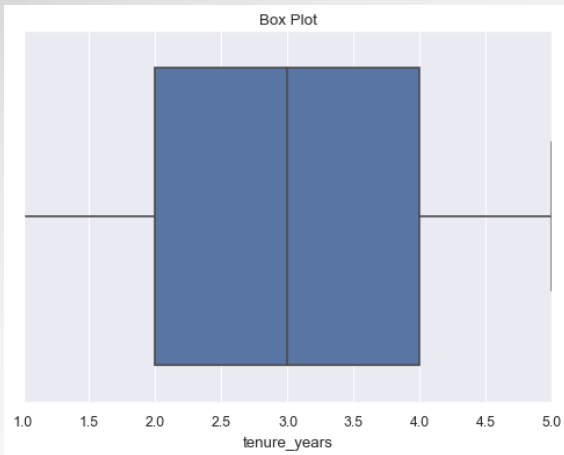
- There are three data sets i.e. main_loan_base.csv, repayment_base.csv, and monthly_balance_base.csv which needs to be aggregated and merged to derive the relevant data set
- Total no of row: 49,985
- Total no of columns: 19
- Test data sets also have three similar data set with different loan account numbers, which needs to be merged similarly
- Target variable is calculated using the formula of LGD
- Missing values in the repayment columns & avg. monthly balance are imputed with 0 and target variable is imputed with mean value
- New column of due amount is added for feature engineering
- Outliers in the Target variable column is dropped

EDA Interpretation



- Two wheeler loan are highest in number followed by car loan as per the above representation





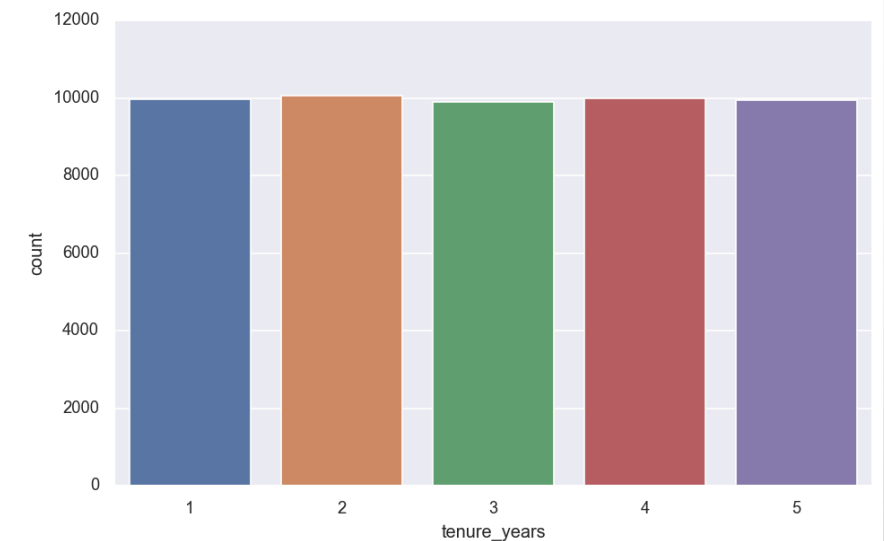
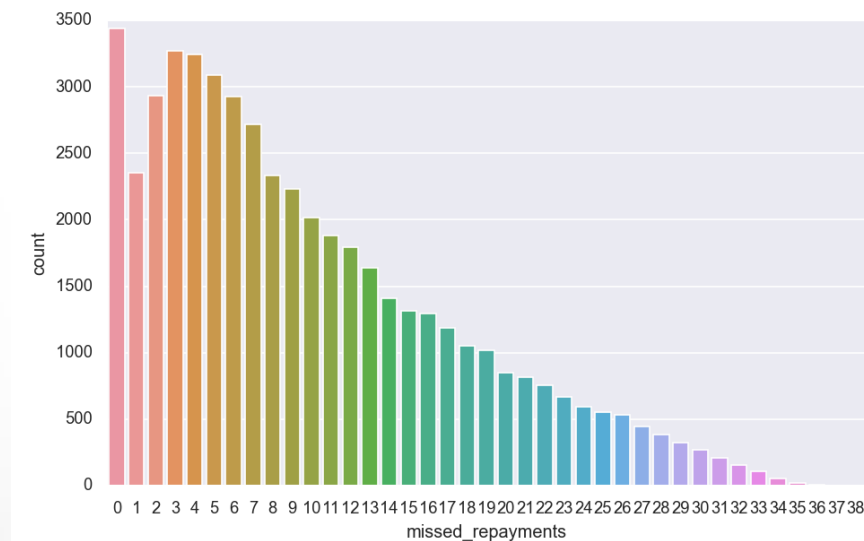
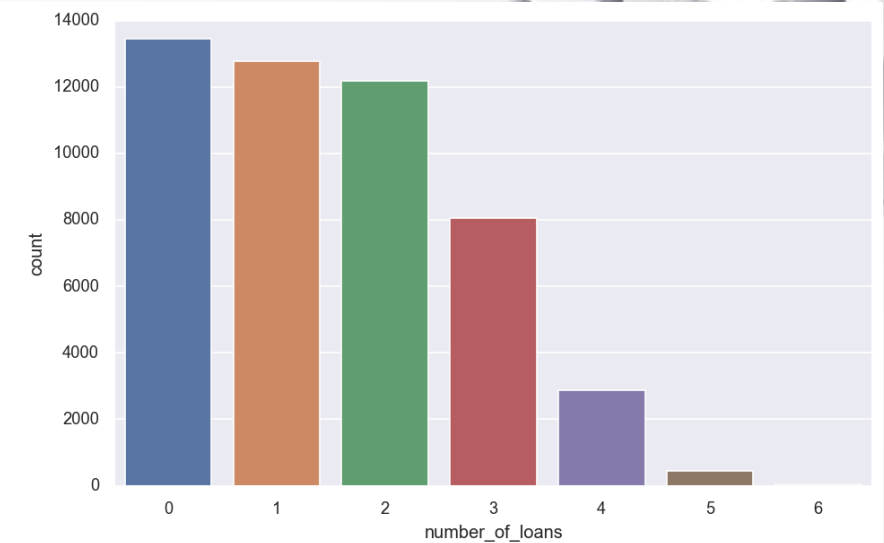
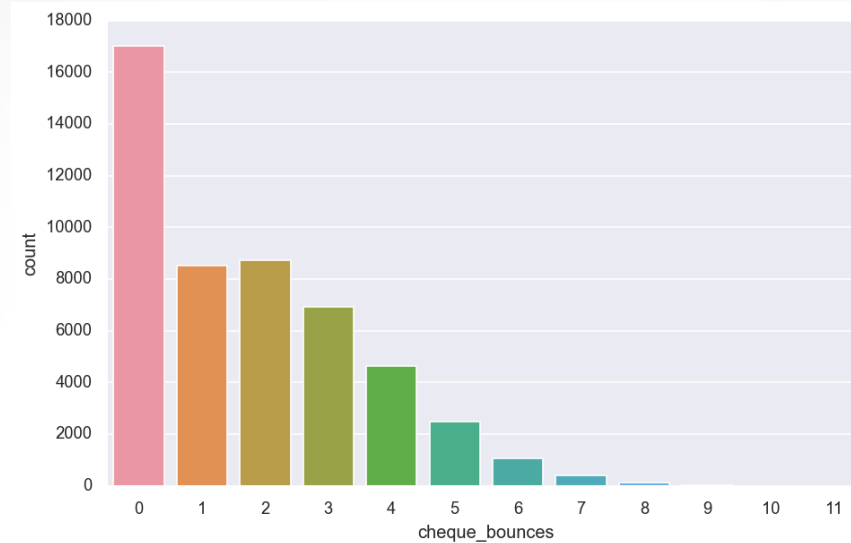


OBSERVATIONS:

1. Monthly EMIs have enormous outliers. Though it will not effect our analysis
2. There are few loan accounts who have missed the repayments more than 30 times
3. Generally people are taking one loan at a time
4. The loan tenure is 2-4 years
5. There are few loan account whose cheque has bounced more than 8 times

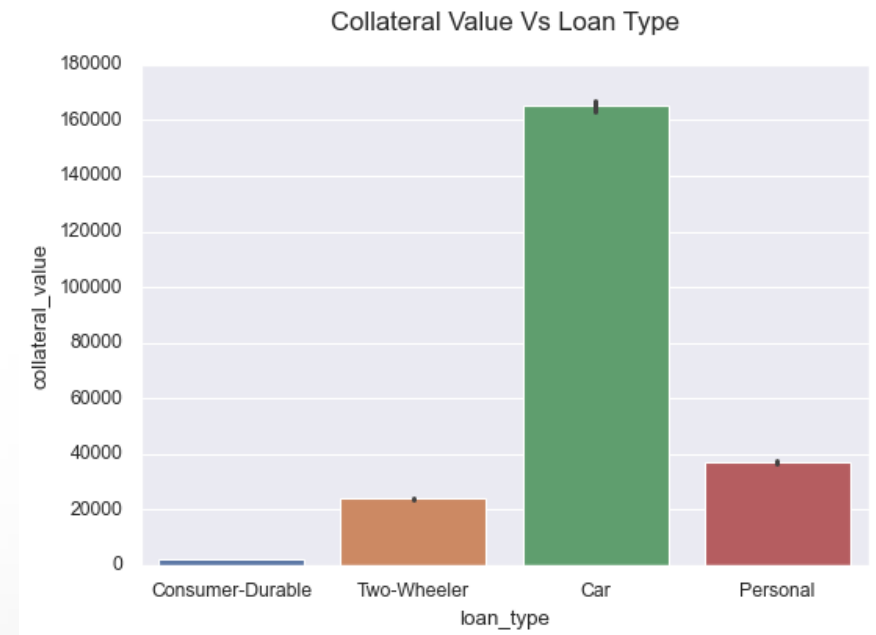
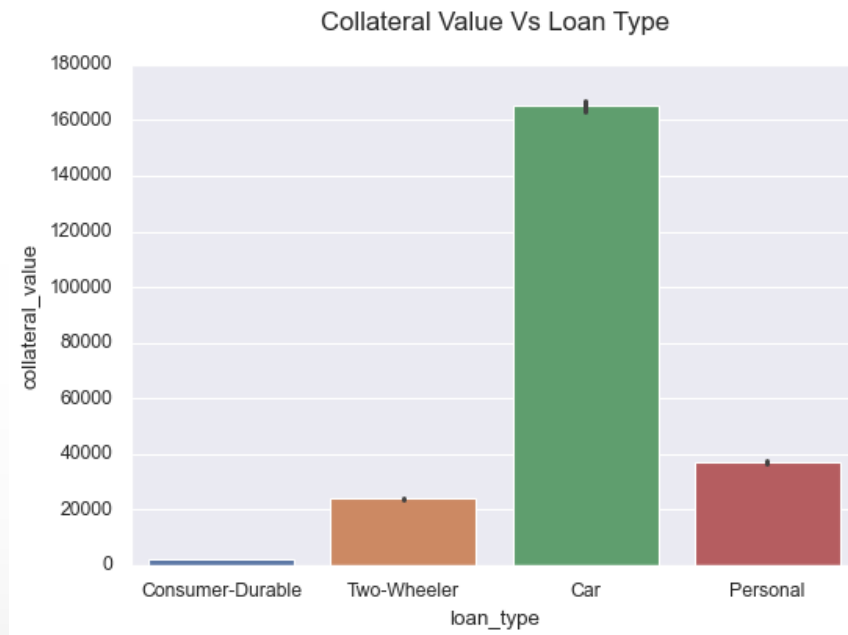
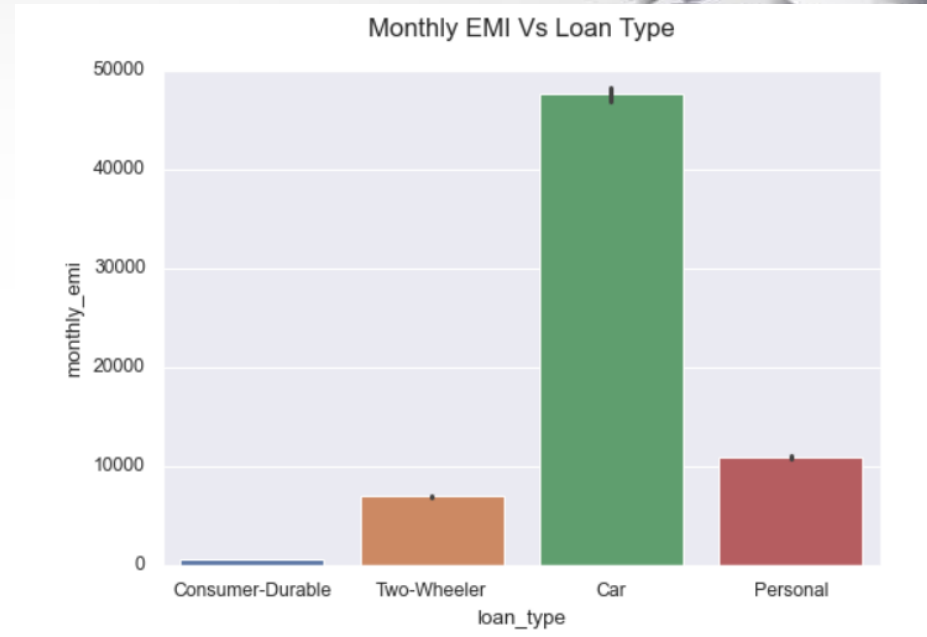
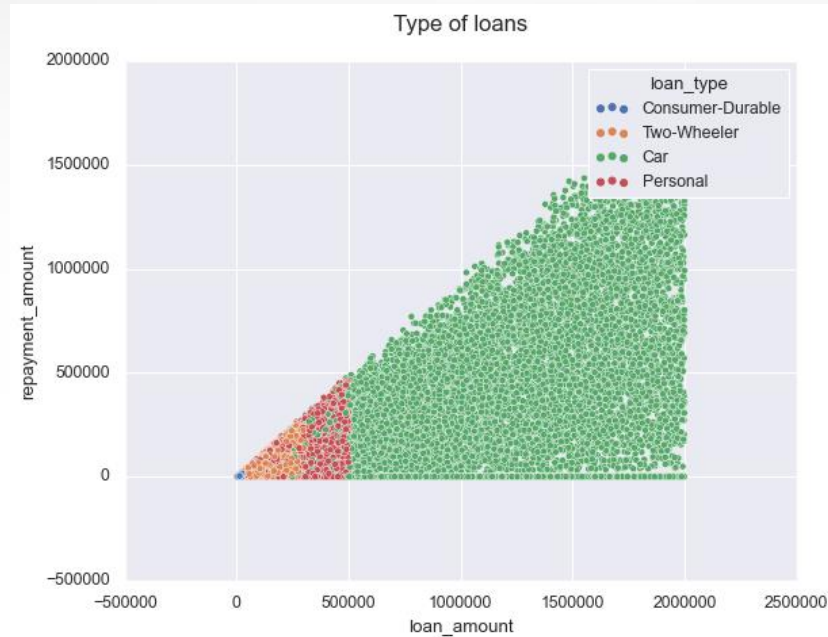
Observation:

- The cheques of over 16,000 loan accounts have bounced.
- There is decreasing trend found on missed repayment of loan amount
- Equal number of people are taking one and two loan at a time

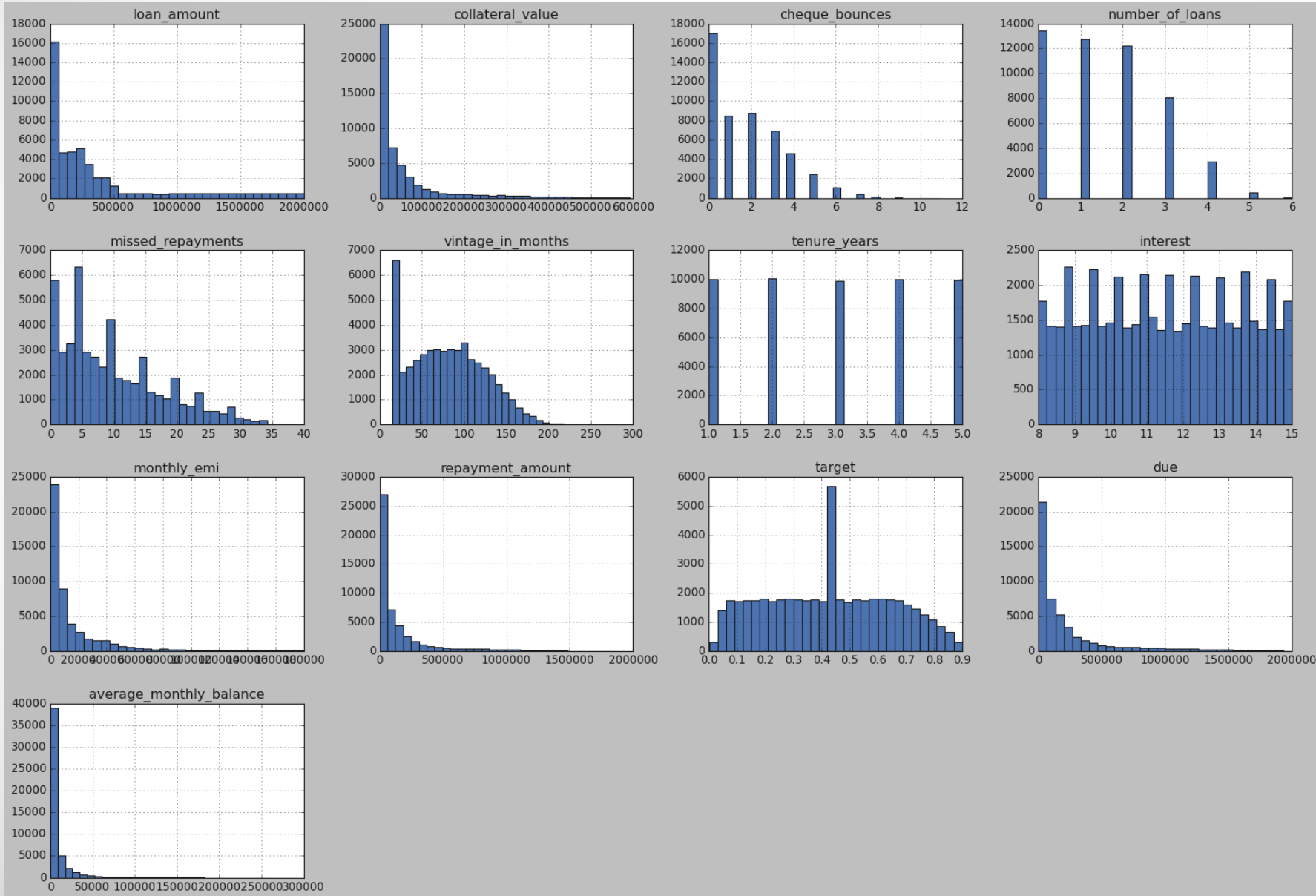


Observation:

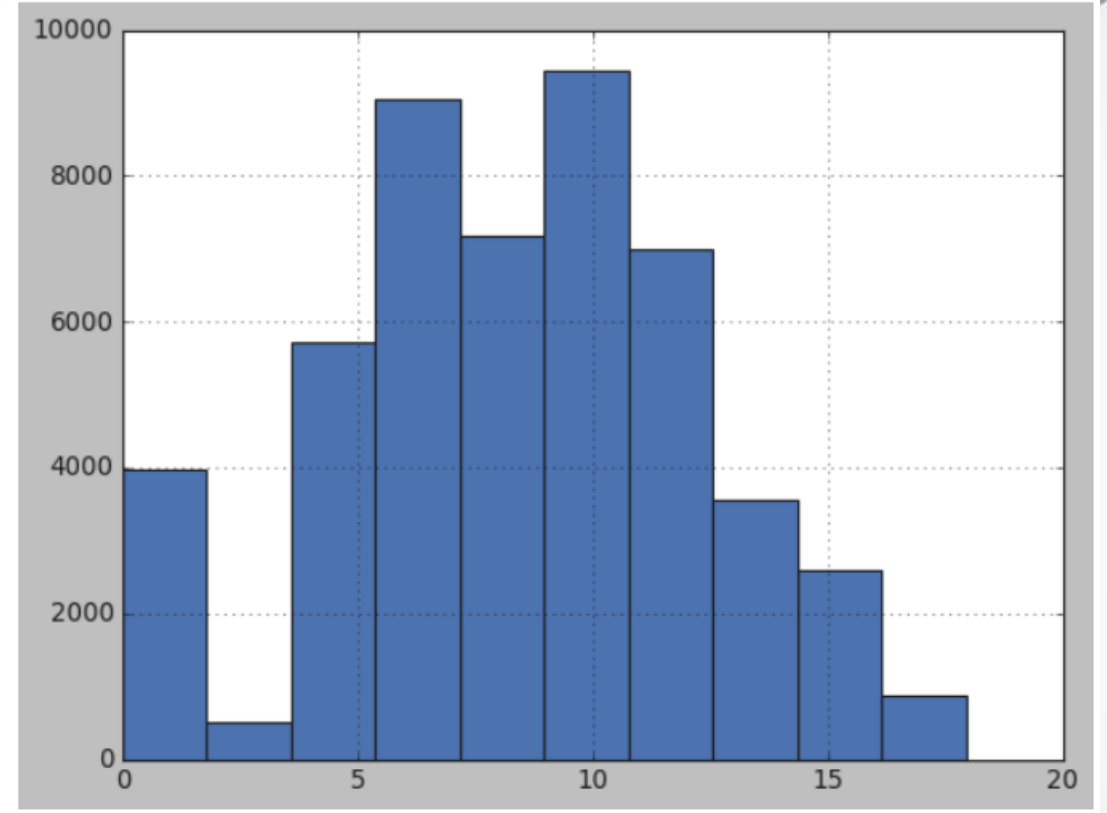
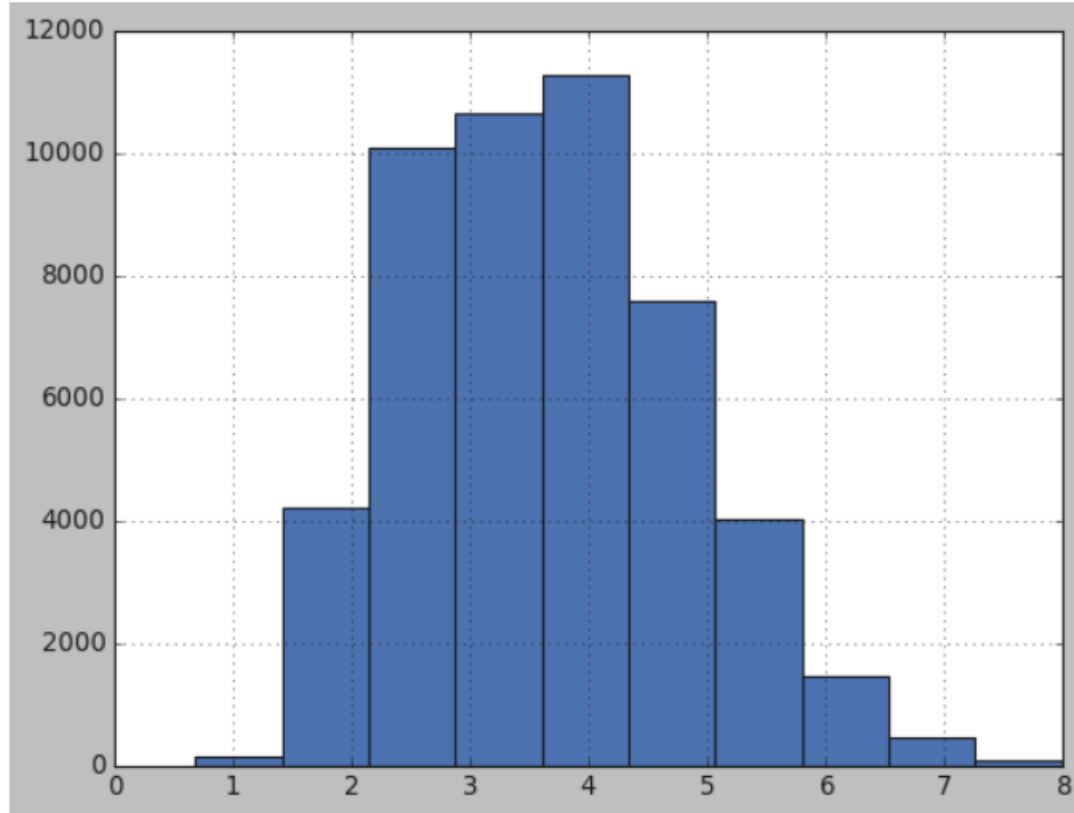
- Repayment amount & EMI for Car Loans are higher than other loan types



Variable transformation, Feature engineering

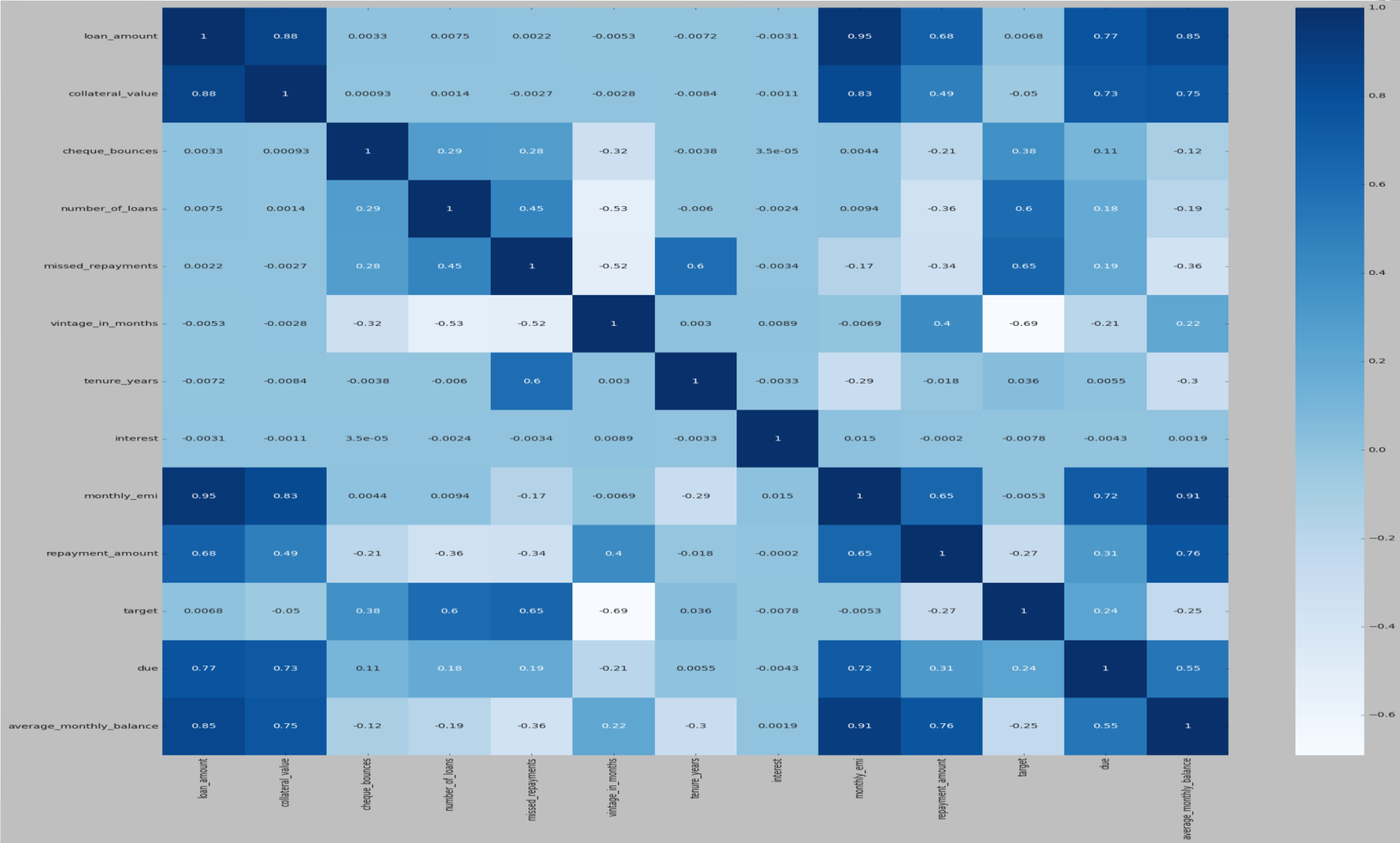


Linear Regression Interpretation



- **Power transformation is done here for linear regression considering all the independent variables are normally distributed**

Multivariate analysis for coefficeint of correlation



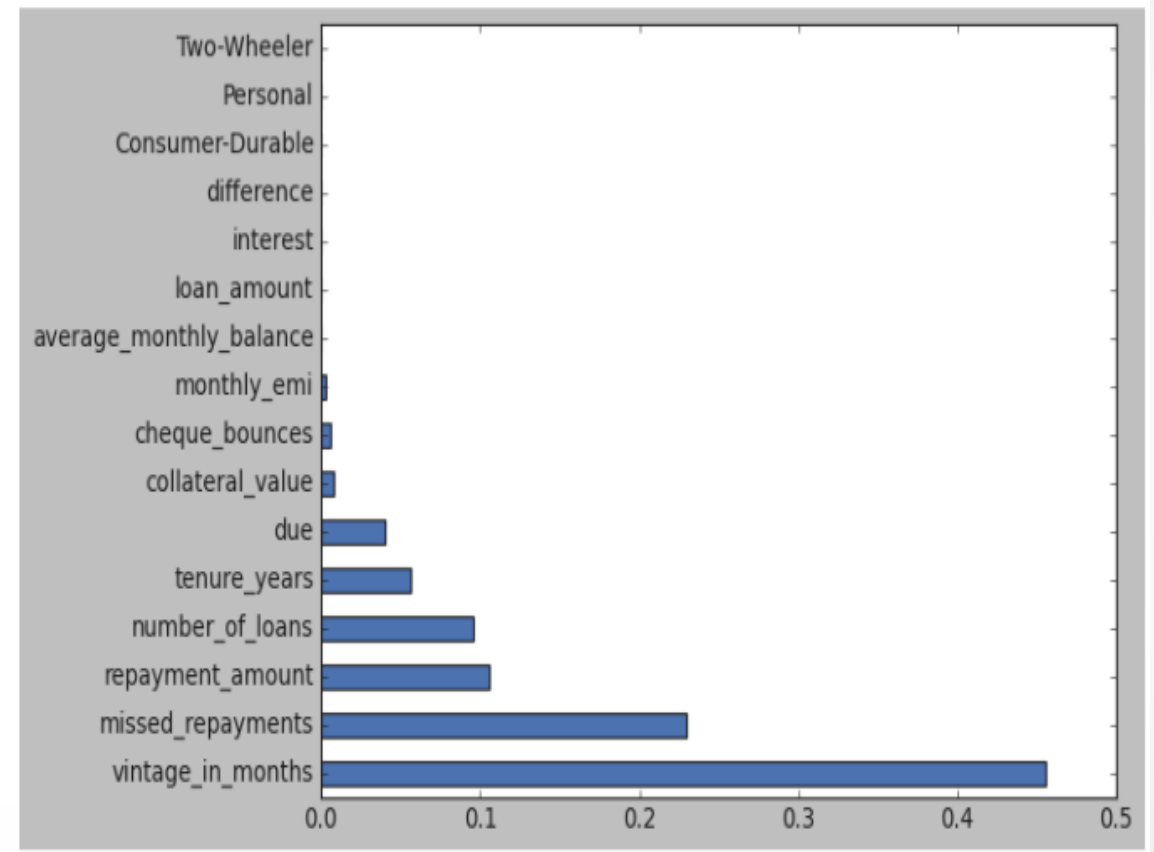
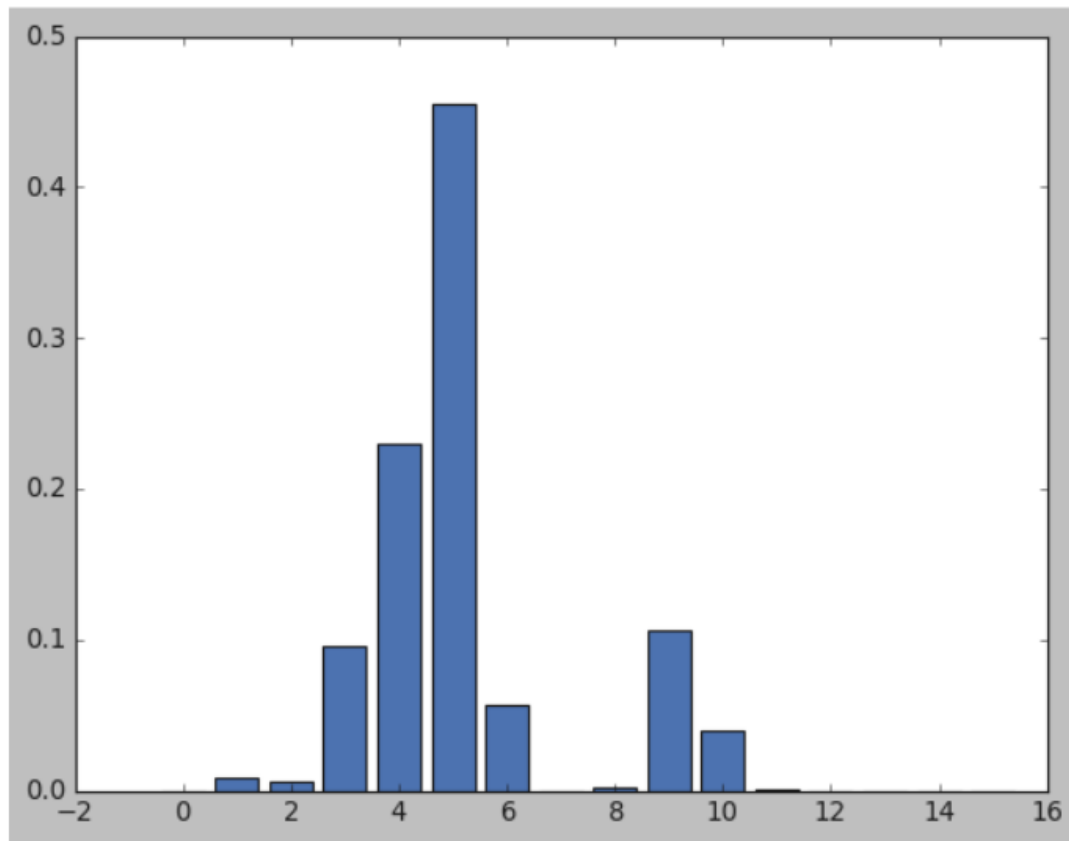
Model Building



- ❑ The given target variable is a continuous variable, hence this fall under the category of Supervised Machine Learning so we used the following models:
 - Recursive Feature Elimination (RFE)
 - Multiple Linear Regression
 - Random Forest Regression
 - Gradient Boost Regression
 - Extreme Gradient (XG) Boost Regression
 - Adaboost Regressor
 - ElasticNet : Hybrid Regularized Model
 - LightGBM

- ❑ Used R Squared as a performance metrics.

Using Feature Importance



Performance result after evaluating the Test data




<u>Model Name</u>	<u>R²value</u>	<u>Model</u>	<u>R²Train</u>	<u>R²Test</u>
RFE	0.77	RandomForestRegressor	0.75	0.75
Multiple LinearRegression	0.78	GradientBoostRegression	0.91	0.89
		XG BoostRegression	0.995	0.997

- Finalize the XG Boost Regression model as it is giving best R2 Metric for both train and test data sets

Result

- These are the Predicted Values of LGD upon the successful execution of the unseen test data



	id	LGD
9792	LN45331474	0.603697
7080	LN20930370	0.114627
3592	LN95300406	0.758101
3131	LN96891062	0.140937
95	LN46736792	0.453451
1858	LN50412896	0.600577
6619	LN14810124	0.049314
6377	LN33945384	0.065070
5085	LN56005390	0.586086
3311	LN17605602	0.337783

Recommendations



- The organization can focus on borrowers with high predicted LGD as they are likely to cause more Credit Loss. LGD is directly proportional to ECL, so the higher the LGD, higher the ECL
- In case, when LGD is equal to 0, loan amount can be recovered with Collateral value and chances of BFSI losing the loan amount is less.
- When LGD is equal to 1, it indicates that the entire loan amount is expected to be lost.
- Customer's due factors and tenure are another subset of influencers to predict the Loss Given Default of the customers.
- The BFSI organization will not expect any loss and will be able to recover loan amount for borrowers who are below threshold, in case of default.
- However, BFSI should focus on the borrowers whose LGD is above threshold, for which organization can expect loss in case of default.
- In order to avoid the Credit Loss, Collateral value should be given more importance.