EDA CASE STUDY

By - Prabhat Kumar & Preetha Buddhan

Problem Statement

- The aim is to identify patterns which indicate if a person is likely to default,which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate,etc.

ABOUT THE DATASET

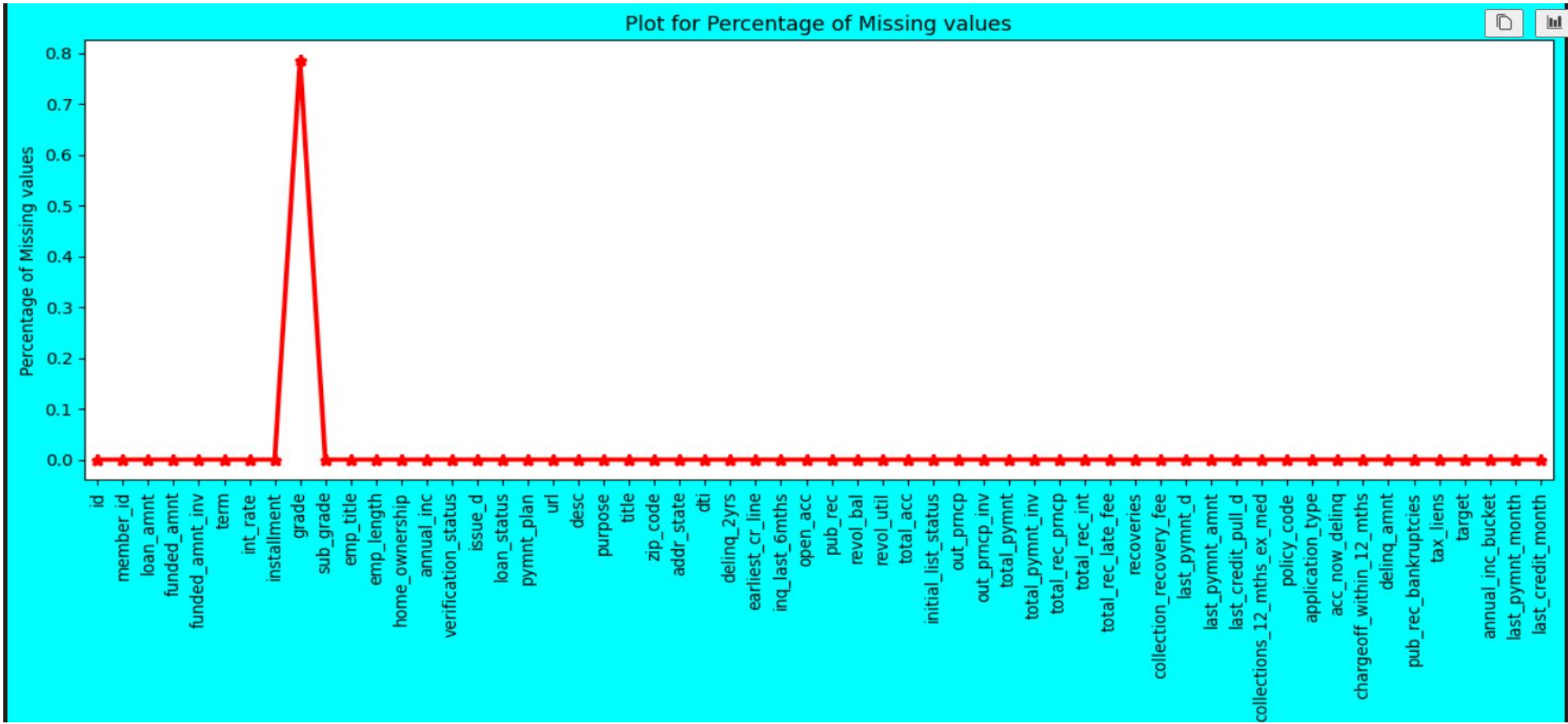
- This dataset has 2 files as explained below:
- ‘loan.csv’ contains It contains the complete loan data for all loans issued through
- the time period 2007 t0 2011.
- ‘Data_Dictionary.csv’ is data dictionary which describes the meaning of the
- variables.

MAJOR STEPS IN ANALYSIS

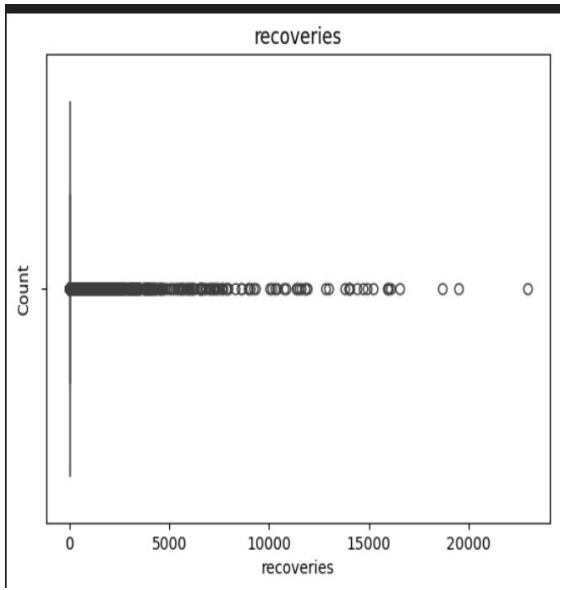
- ❖ Data Sourcing
- ❖ Data Understanding
- ❖ Checking and Handling Missing values in the data
- ❖ Handling Data Errors
- ❖ Outlier Identification and Analysis
- ❖ Univariate Analysis
- ❖ Bivariate and Multivariate Analysis
- ❖ Finding Top Correlated Features those Support Target Column.

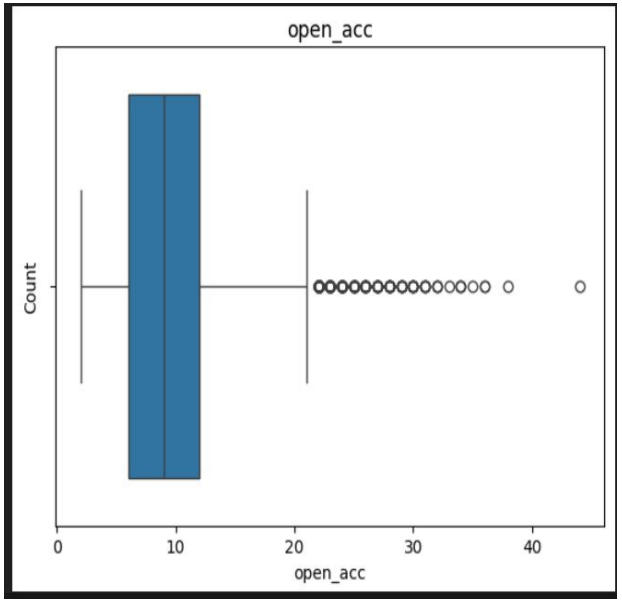
Results on Loan.csv

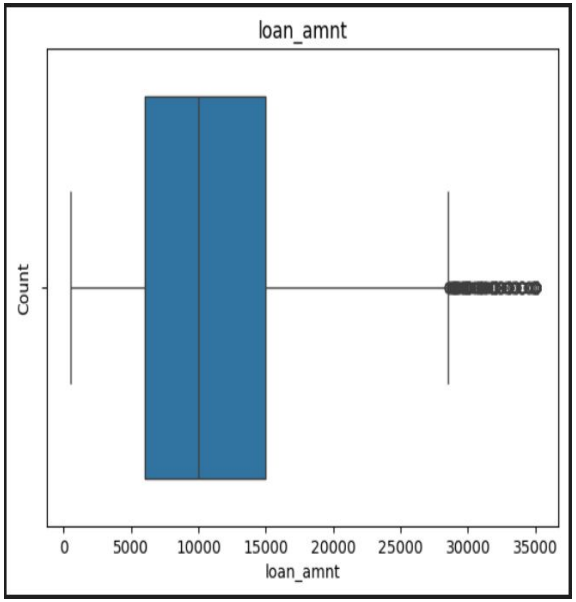
HANDLING MISSING DATA



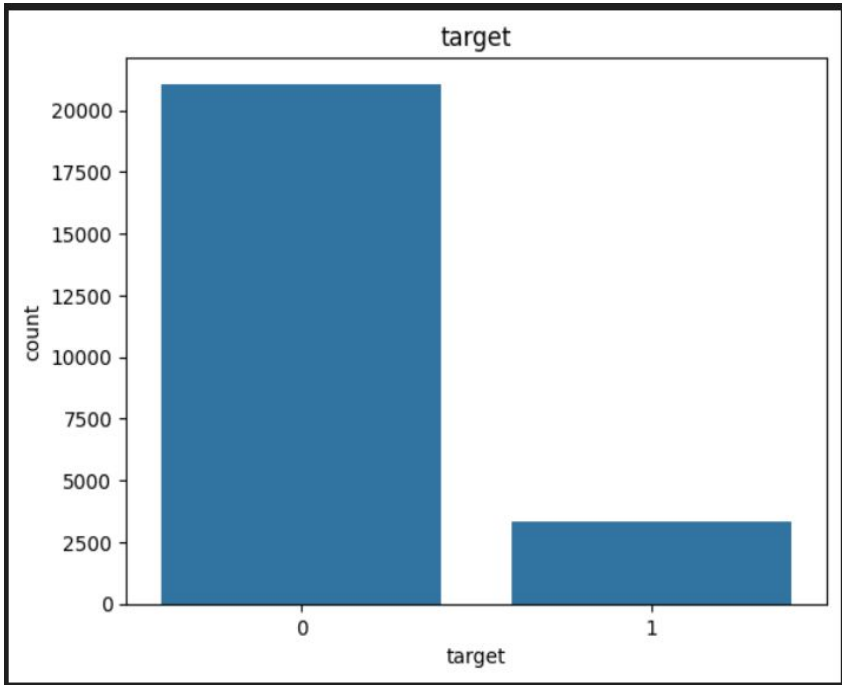
Outlier Analysis

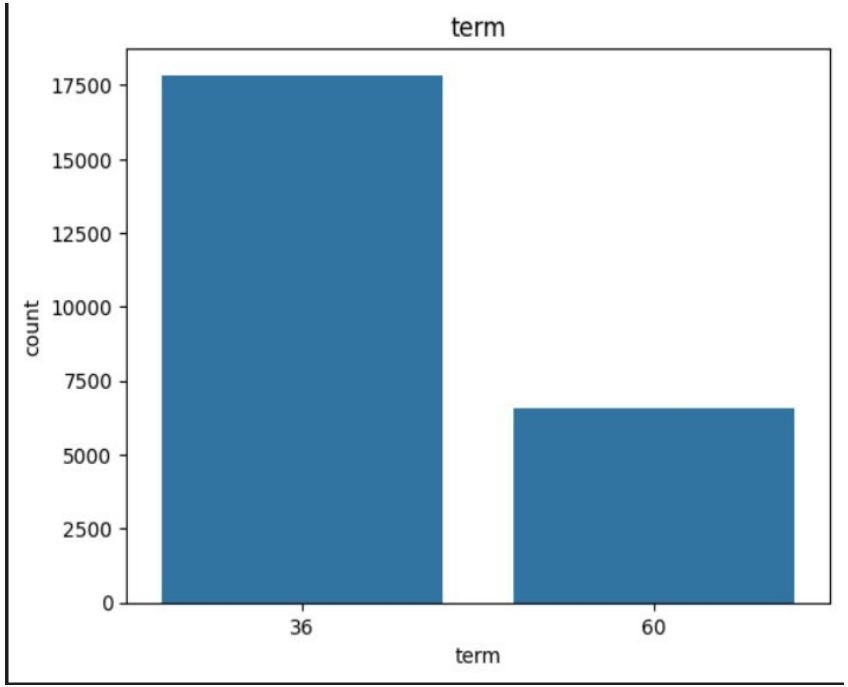




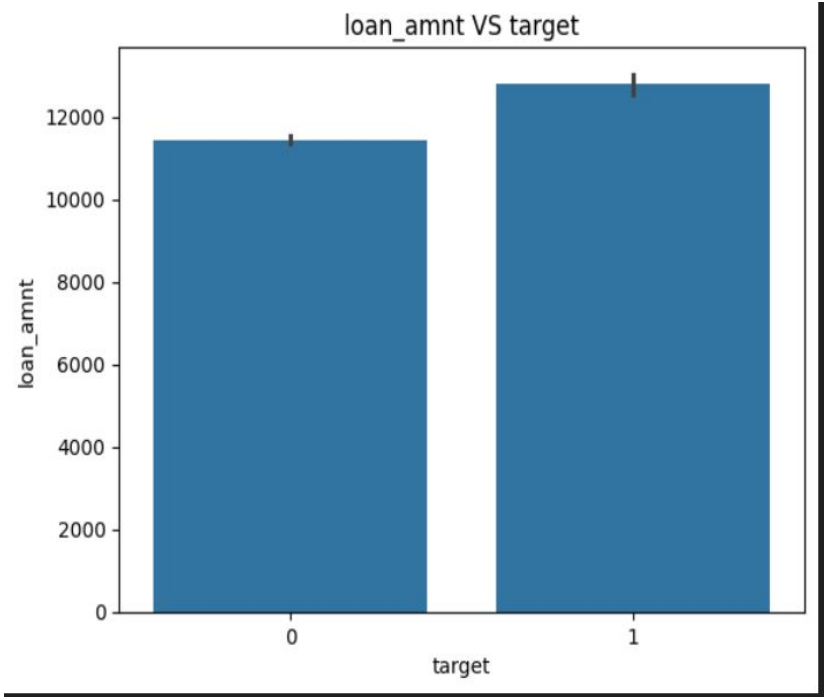


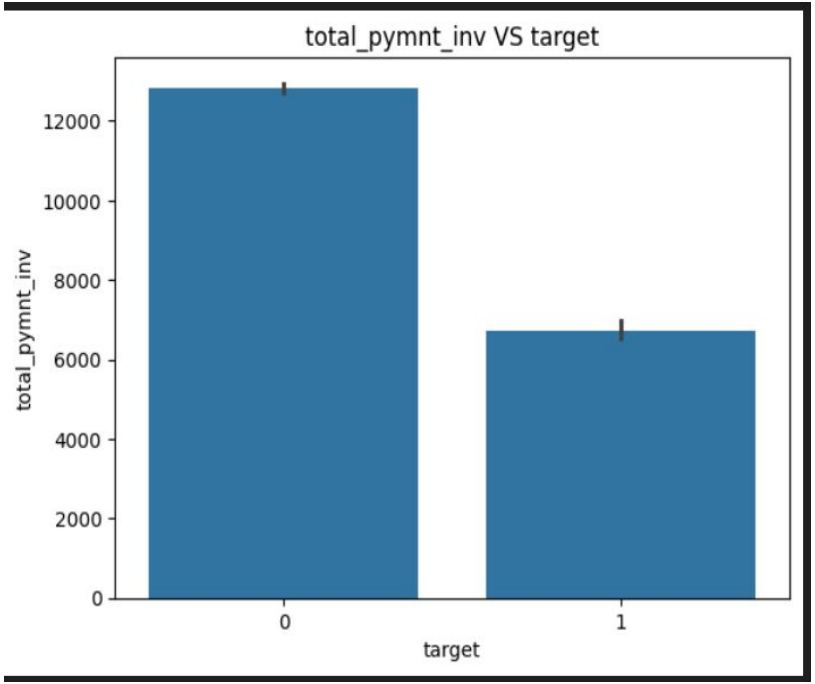
Univariate Analysis

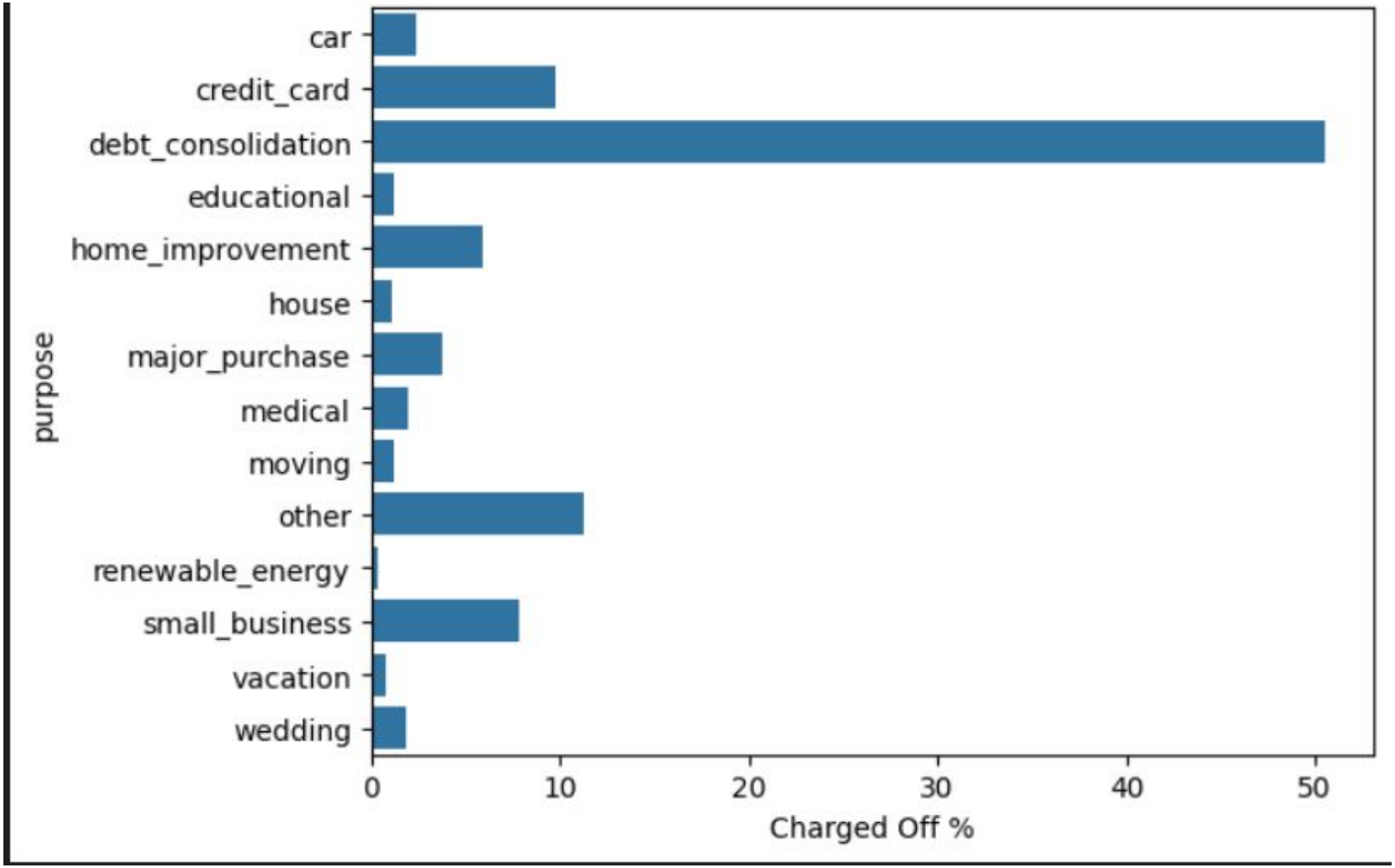




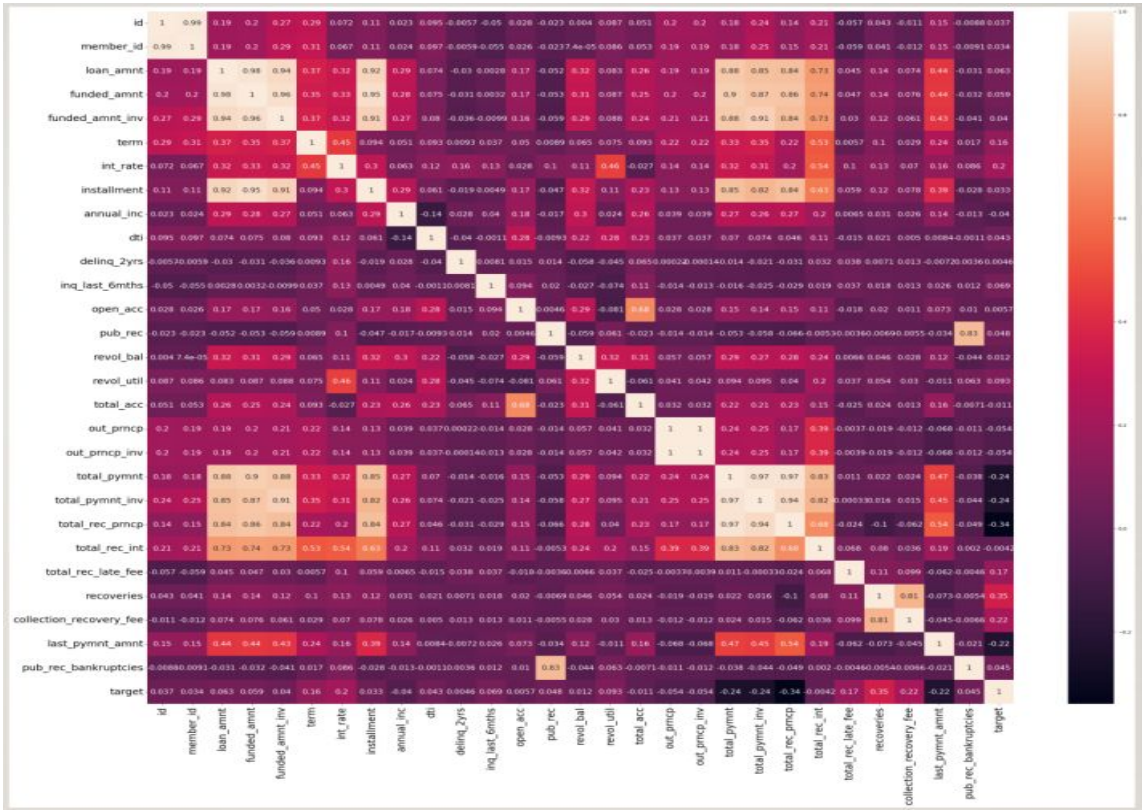
Bivariate Analysis






car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
purpose
0
10
20
30
40
50
Charged Off %

Cluster Map


id
member_id
loan_amnt
funded_amnt
funded_amnt_inv
term
int_rate
installment
annual_inc
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
pub_rec_bankruptcies
target

IMPORTANT OBSERVATIONS FROM THE EDA

- NO major differentiation seen in interest rates with income segment however, asincome segment increases interest rates increases very slightly
- As grade changes from A to B and finally to F interest rates significantly increases.This means F are more risky customers as compares to A
- We can observe that the month May has more defaulted values compared to other in last_credit_pull_d months.
- Median incomes of all three categories of customers are nearly similar. However,many Fully Paid customers have higher income levels than charged off and current customers.
- People take Higher Loan amount for long term loans and vice versa i.e. Higher Loan amount in 60 months loan tenure

Major Observationas

- Median incomes of all three categories of customers are nearly similar with increasing trend from fully paid to charged off and current customers.
- However, Many Fully Paid customers have higher installments than charged off and current customers.
- Almost 49% loan are charged off when taken for the purpose of debt consolidation which is very high
- As income segment increases installment also increases

Take Keyaway:

From the above heatmap we can observe that the columns term, int_rate, revol_util has positive correlation with the target column and total_payment, total_payment_inv, total_rec_prncp, total_rec_late_fee, recoveries, collection_recovery_fee and last_payment_amount has negative correlation with the target column.

And we also observe that columns loan_amnt, funded_amnt, funded_amnt_inv, total_payment, total_payment_inv, total_rec_prncp, total_rec_int has high correlation among them self.

Detailed Analysis can be found in .ipynb file

THANK YOU