

EDA CASE STUDY

BY – VAISHALI SINGH & MANISHARAN



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:
- **1. 'loan.csv' contains It contains the complete loan data for all loans issued through the time period 2007 to 2011.**
- **2. 'Data_Dictionary.csv' is data dictionary which describes the meaning of the variables.**

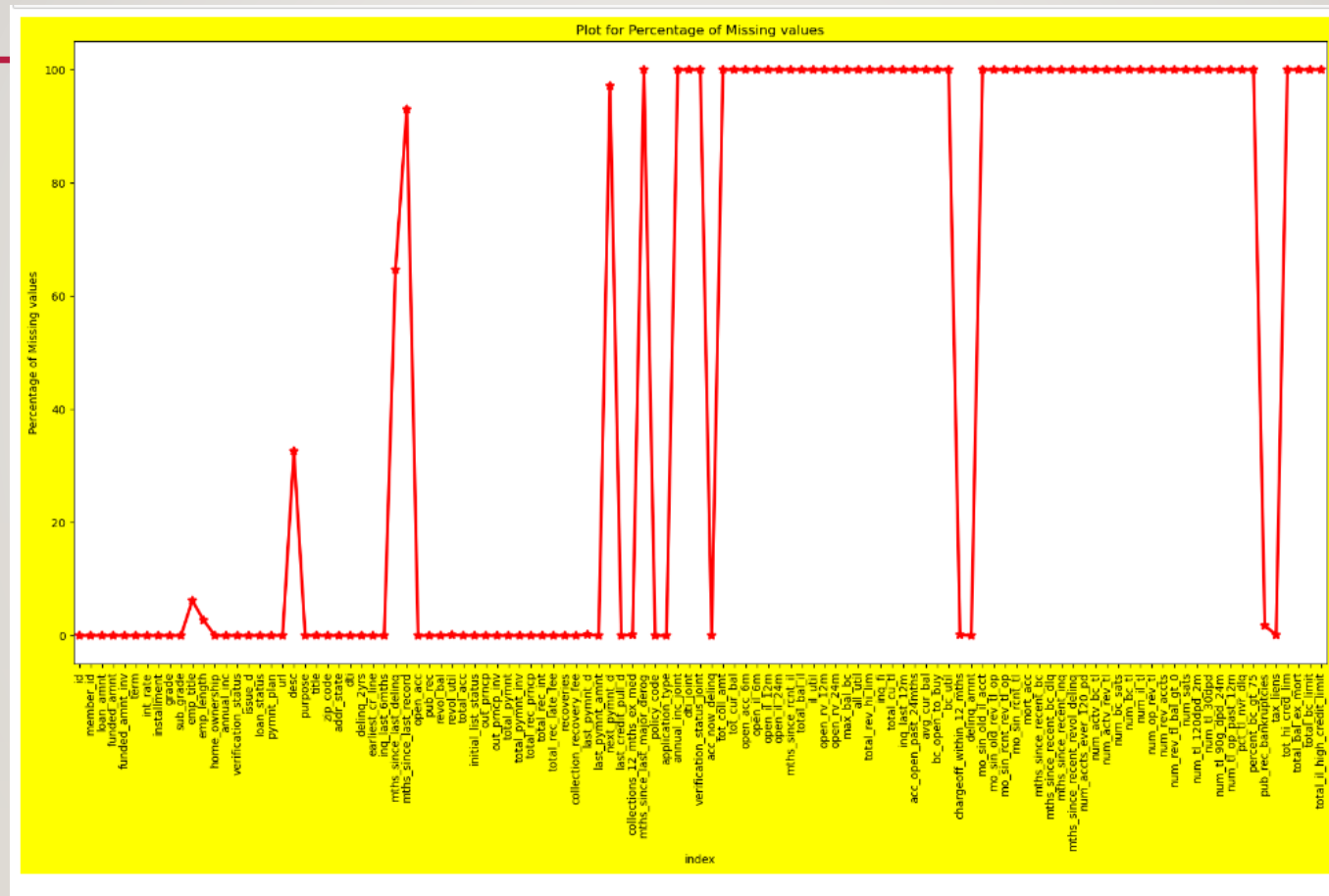
MAJOR STEPS IN ANALYSIS

- ❖ 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 Dataset

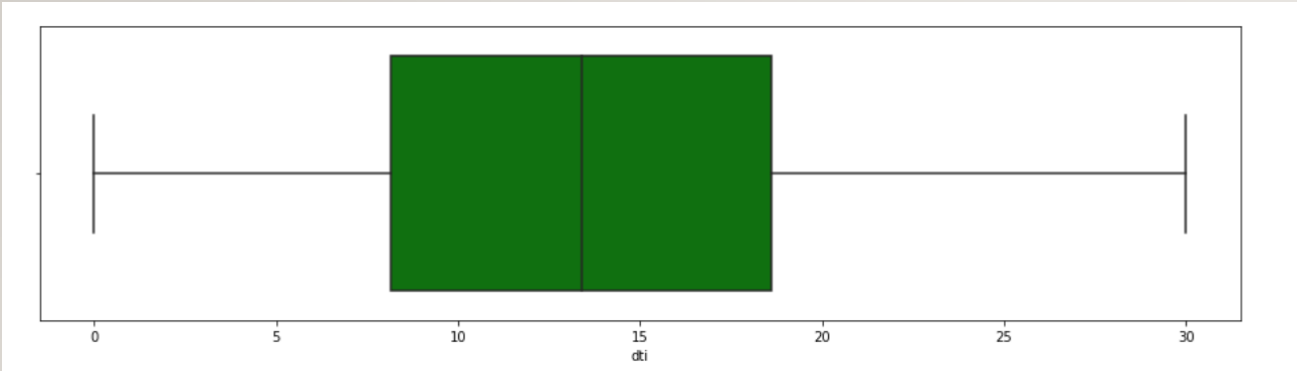
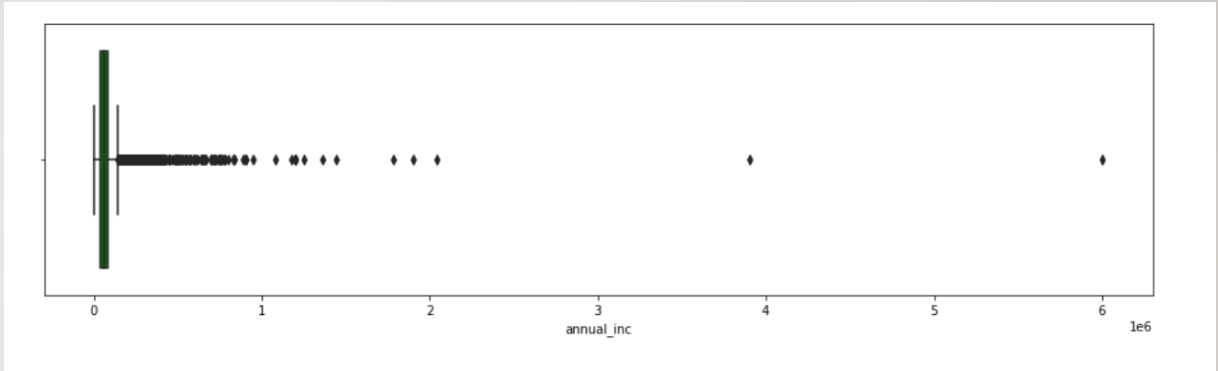
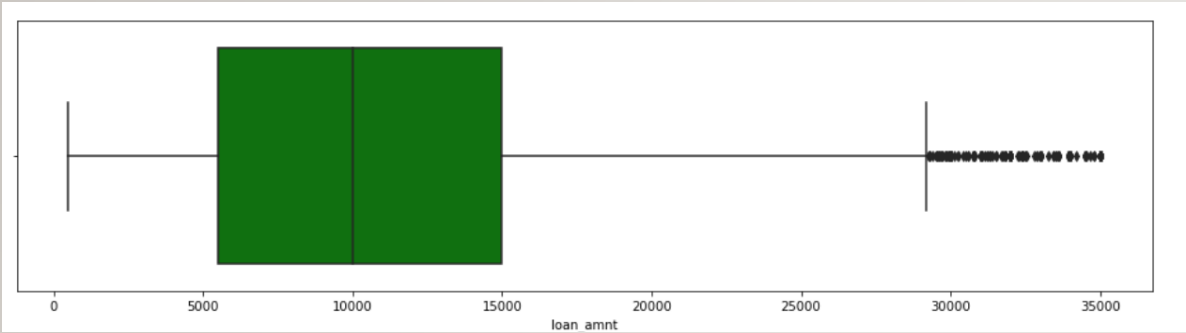


HANDLING MISSING DATA

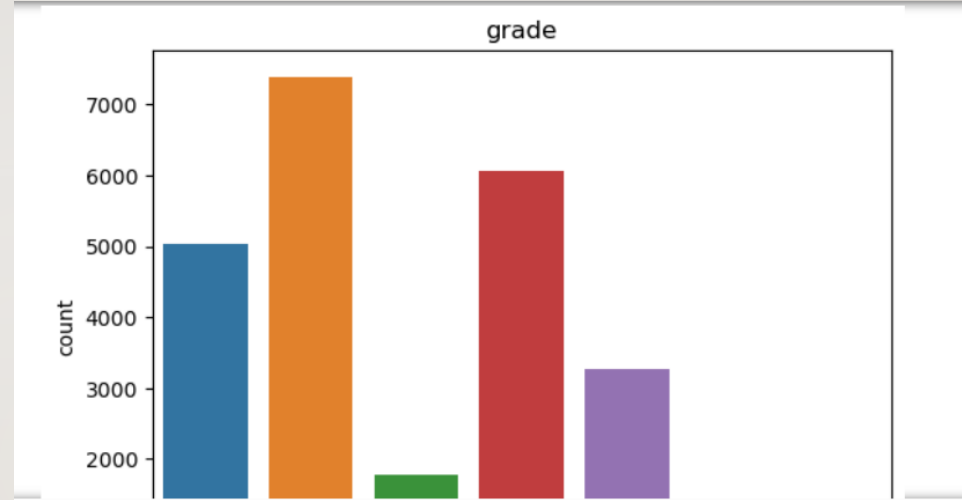
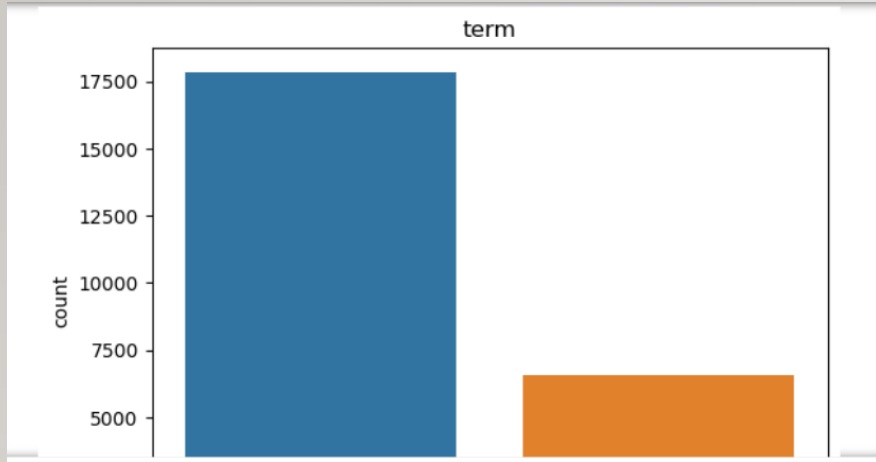


Missing Value percentage in the given Dataset

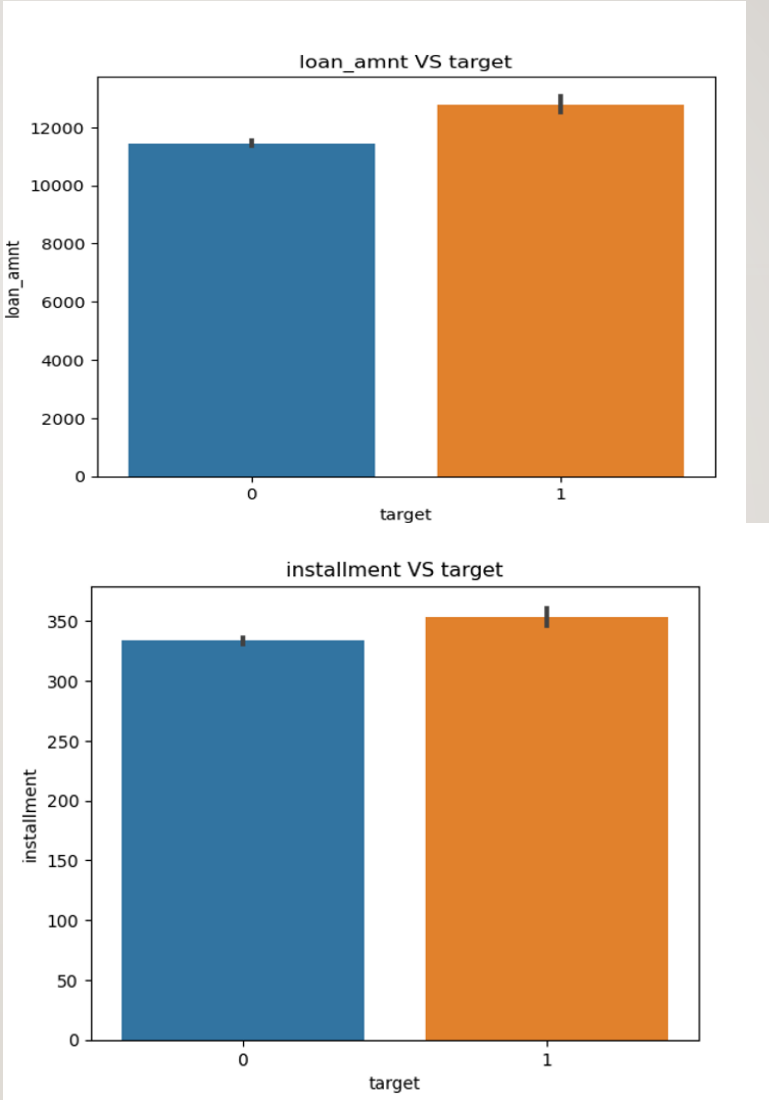
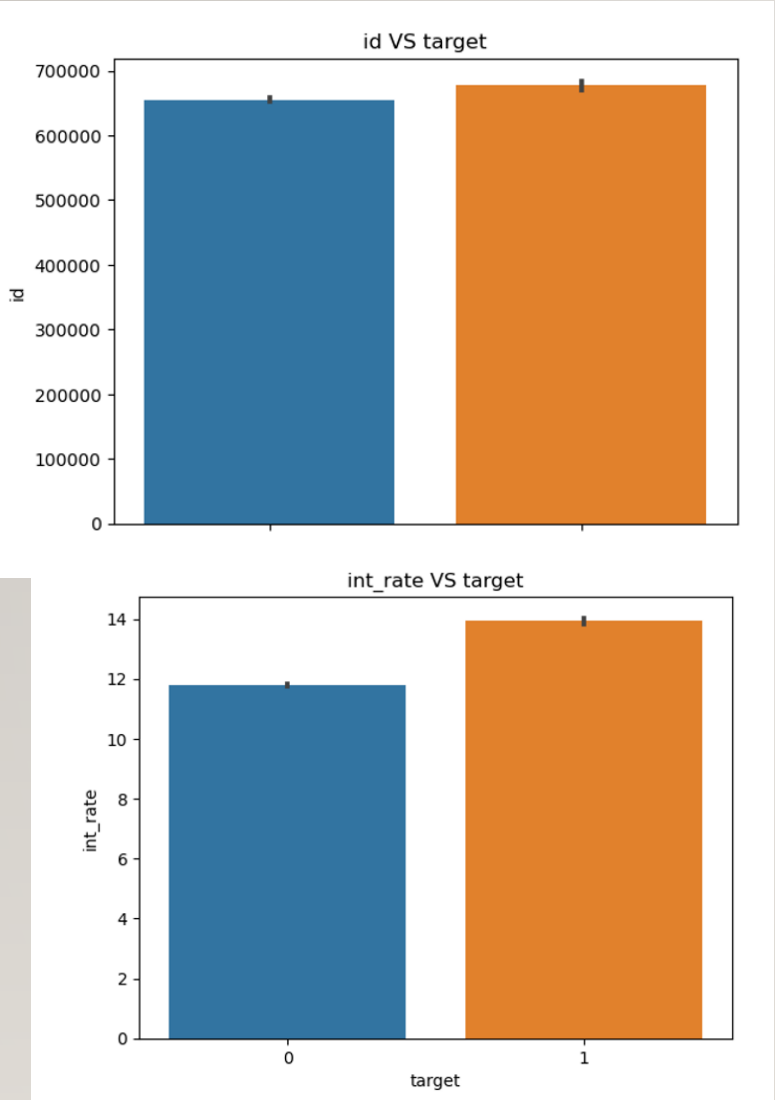
Outlier Analysis

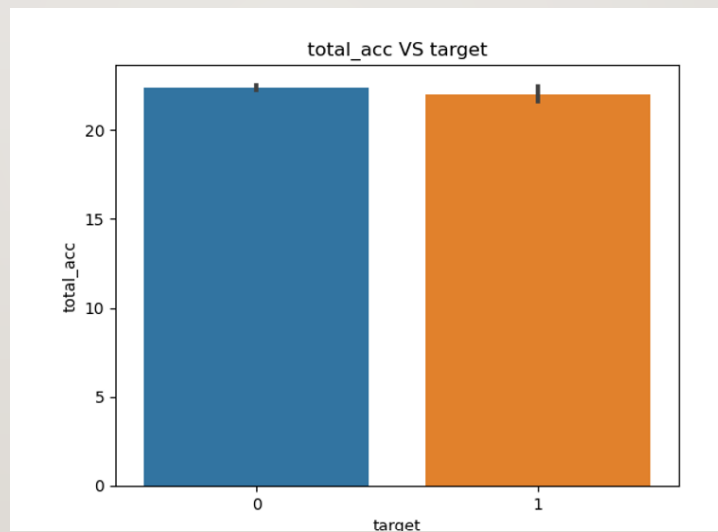
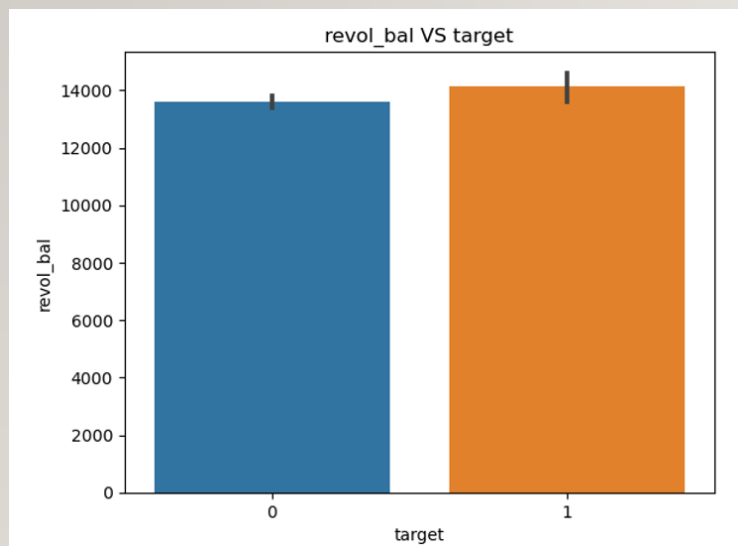
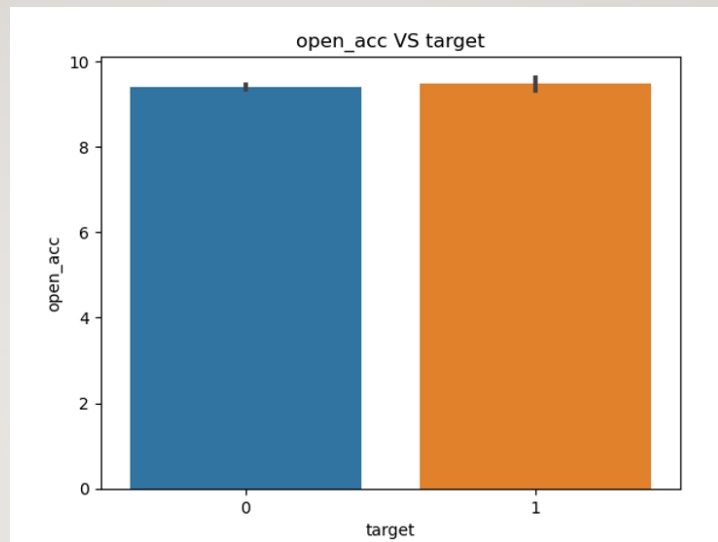
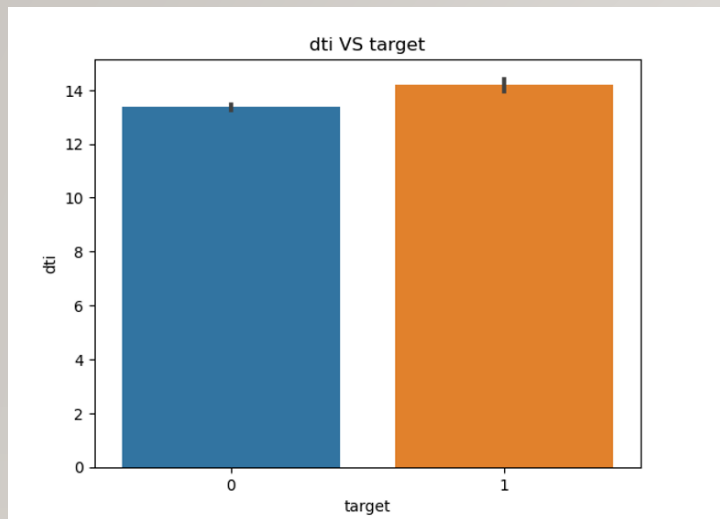


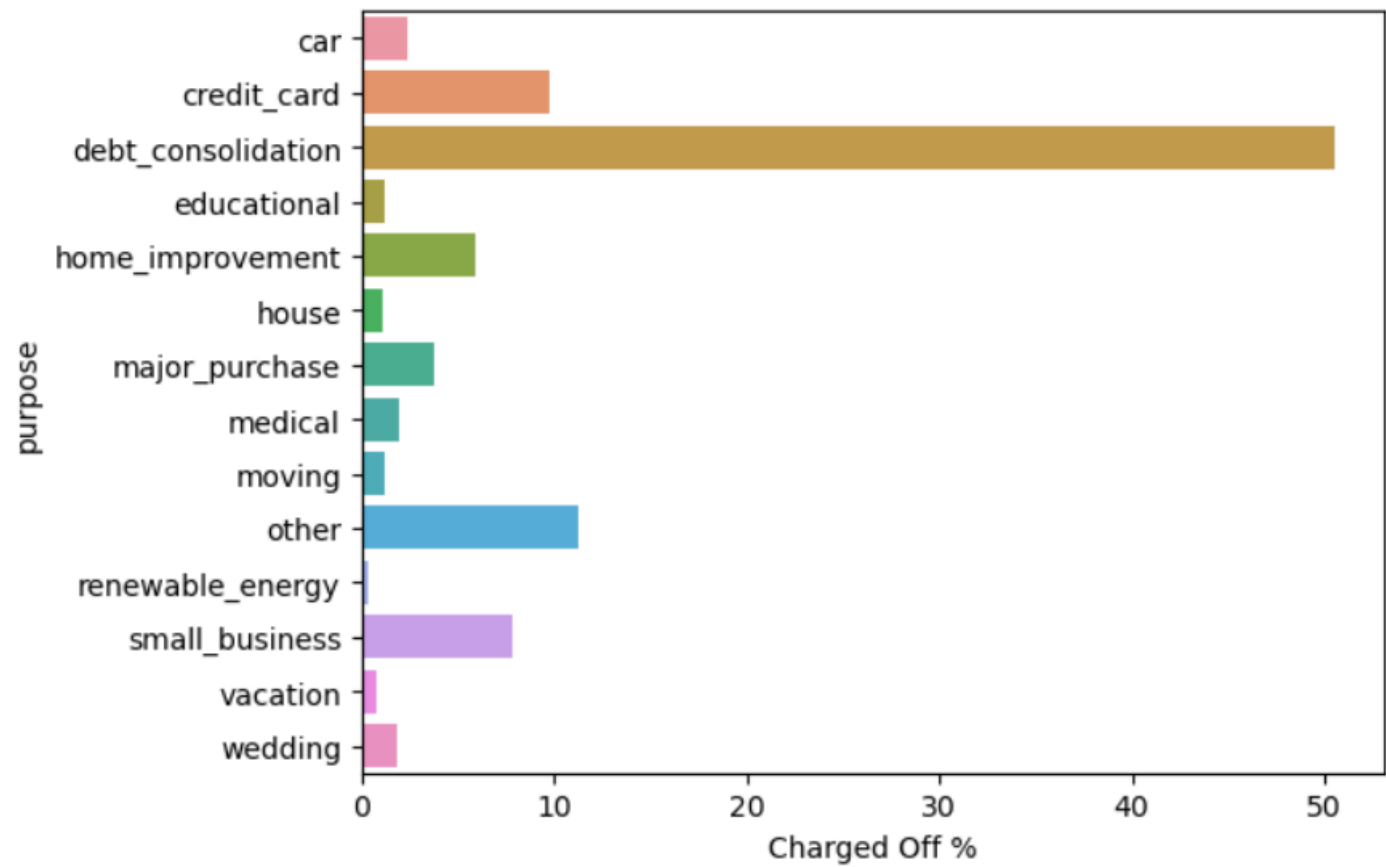
Univariate Analysis



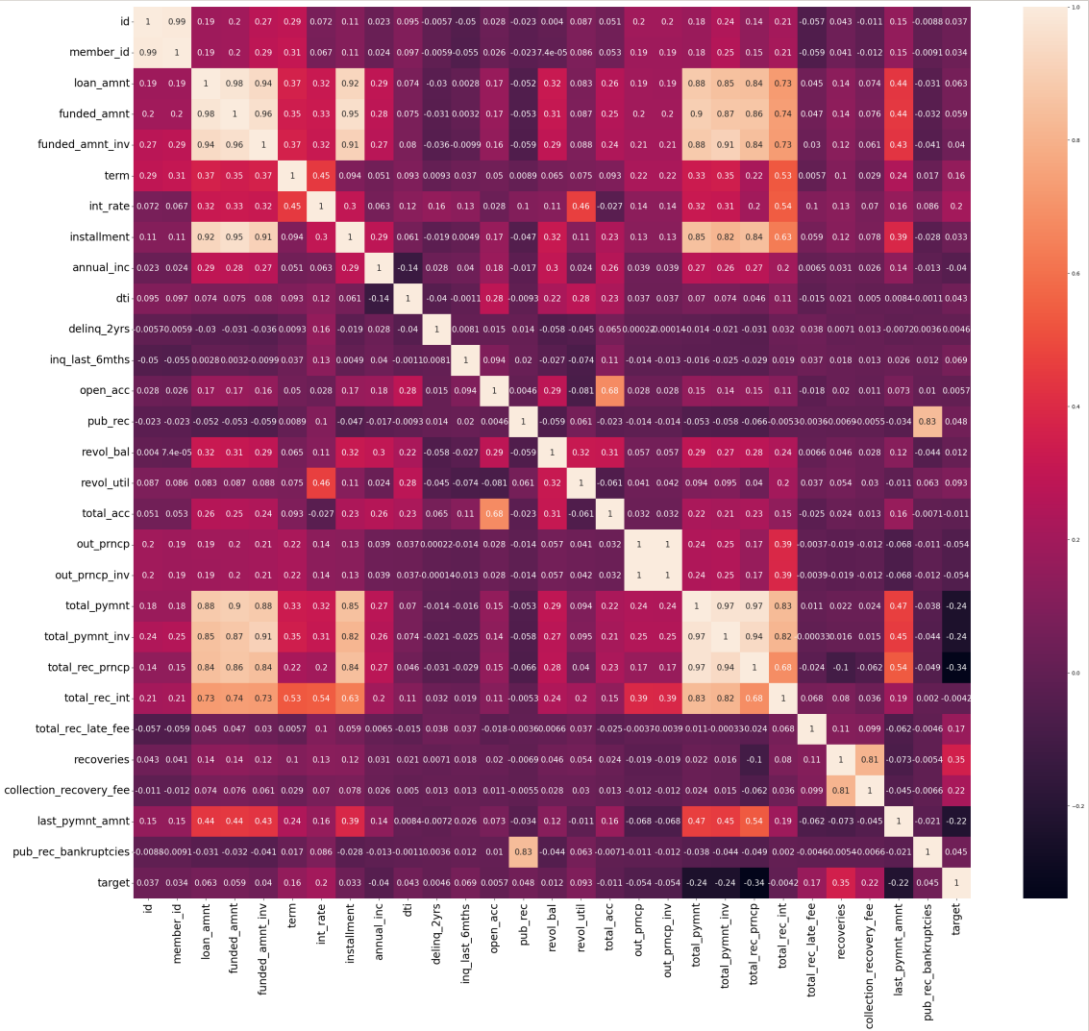
Bivariate Analysis







Cluster Map



IMPORTANT OBSERVATIONS FROM THE EDA

- NO major differentiation seen in interest rates with income segment however, as income 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

- 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



Key Takeaway:

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 with
code, corresponding
results and important
Deductions can be found
in the .ipynb file.***

