Lending Club Case Study

By

Debasish Mondal and Sudha G. Lakshmaiah

- This project work is done as a part of the **Statistics Essentials** module of **upGrad**'s Executive PG Programme in Machine Learning.

- Team members of this project are –
  - Debasish Mondal
  - Sudha G. Lakshmaiah

Problem Statement

- To identify the bank customers who could turn into possible loan defaulters in the future.

Introduction

- Dataset Name: loan.csv
- Total No. of Rows: 39717
- Total No. of Columns: 111

Data Analysis Chronology

- Data Reading and Import Requisite Libraries
- Data Cleansing
  - Remove Null Valued Columns
  - Remove Single-valued Columns
  - Remove Irrelevant Columns
  - Standardize Data formats
  - Remove Outliers
- Data Extraction
- Exploratory Data Analysis (EDA)
  - Univariate Analysis
  - Segmented Univariate Analysis
  - Bivariate Analysis
- Observations

Imported Libraries

- Pandas: For dataframe manipulation and analysis.
- Matplotlib: For data visualization and graphical plotting.
- Seaborn: For data visualization and graphical plotting in more sophisticated manner.
- Warnings: To control warning statements.

Data Cleansing

- Removing null valued columns : After removing them, row numbers are unchanged, but only 57 columns are still left.
- Removing single valued columns: After removing them, row numbers are unchanged, but only 48 columns are still left.

Data Cleansing

- Removing irrelevant columns: The following columns are not considered in our analysis, and the reasons are stated below:
  - 'id', 'member_id', 'emp_title', 'url', 'title', 'zip_code', 'addr_state', 'desc': These variables only provide non-technical information, which is irrelevant for exploratory data analysis (EDA).
  - 'issue_d', 'delinq_2yrs', 'inq_last_6mths', 'earliest_cr_line', 'mths_since_last_delinq', 'mths_since_last_record', 'revol_bal', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt', 'total_pymnt_inv', 'last_pymnt_d', 'last_pymnt_amnt', 'next_pymnt_d', 'last_credit_pull_d': These variables are indicate only some invester oriented or post loan approval features or just simply do not put any signifiacnce on the loan defaulter analysis.
  - 'funded_amnt': As' funded_amnt_inv 'prefered over it as a more reliable measure of loan amount.
- After removing them, row numbers are unchanged, but only 19 columns are still left.

Data Cleansing

- Data Standardization: The data formats of the following columns have been standardized: ‘term’, ‘int_rate’, ‘revol_util’, ‘pub_rec_bankruptcies’, ‘emp_length’ , ‘home_ownership’.

Data Cleansing

- Outliner Removal: Outliers have been removed from the following columns: ‘loan_amnt’, ‘funded_amnt_inv’, ‘ ‘annual_inc’, ‘installment’
- For outlier treatment, we took values within the 95% confidence level range.

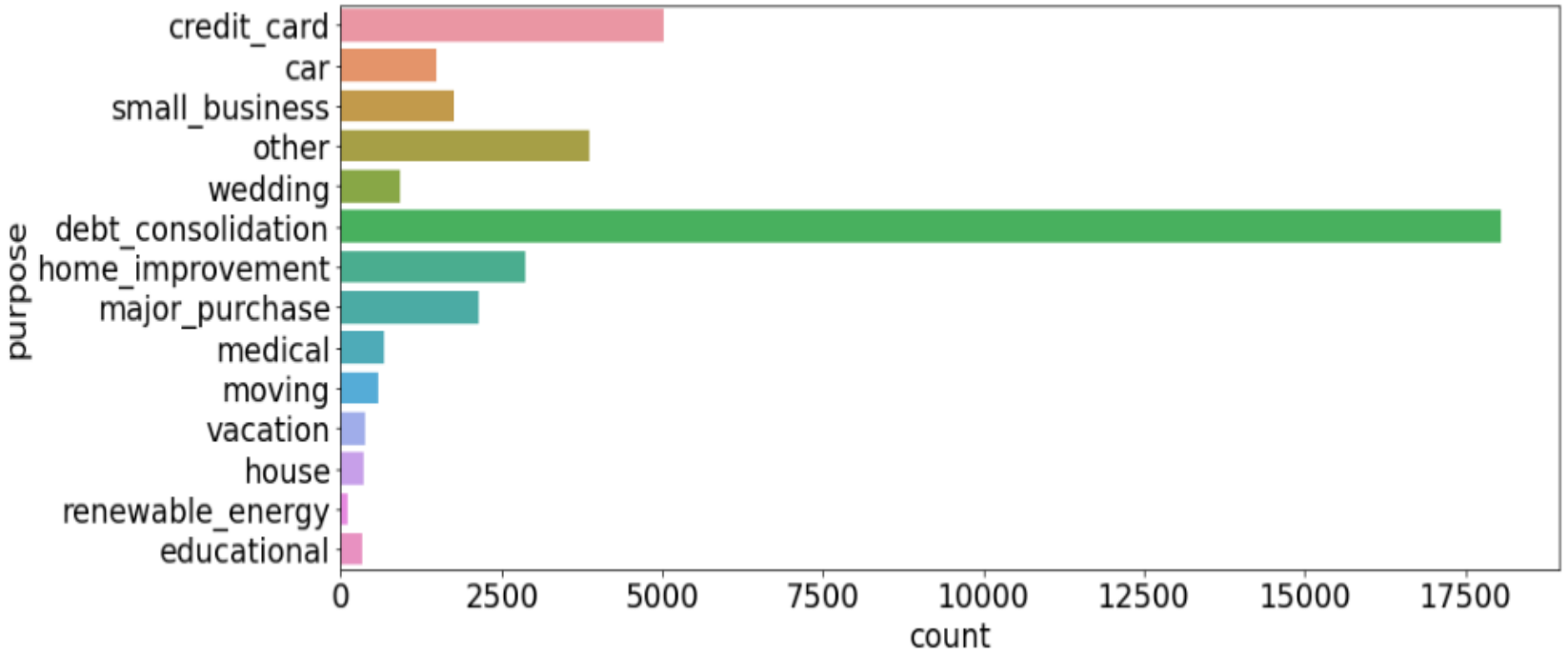
Data Extraction

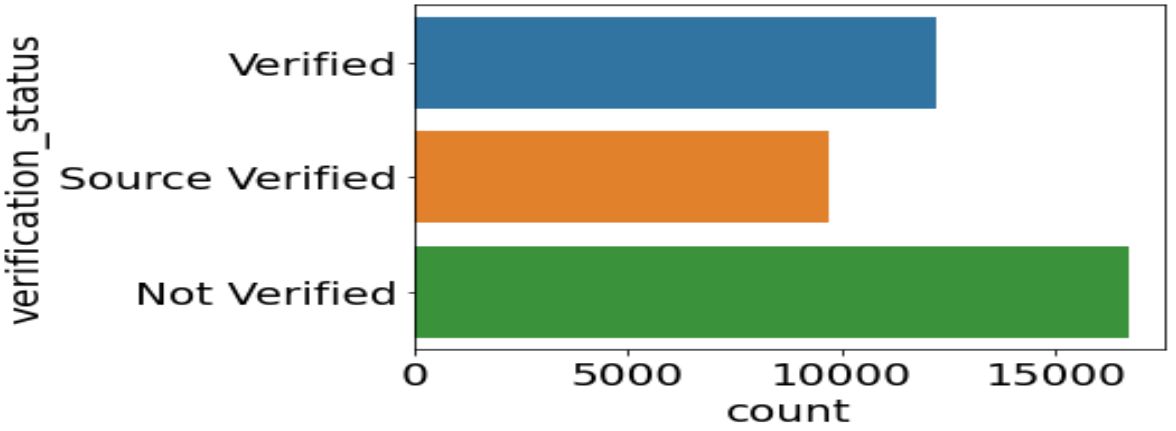
- Loan status marked as 'Charged Off' due to on-time payments of their loan installments is excluded from the provided dataset as there is not much weightage on those customers in the loan defaulter analysis.
- After removing them, 38577 rows and 19 columns are still left.

Exploratory Data Analysis (EDA)

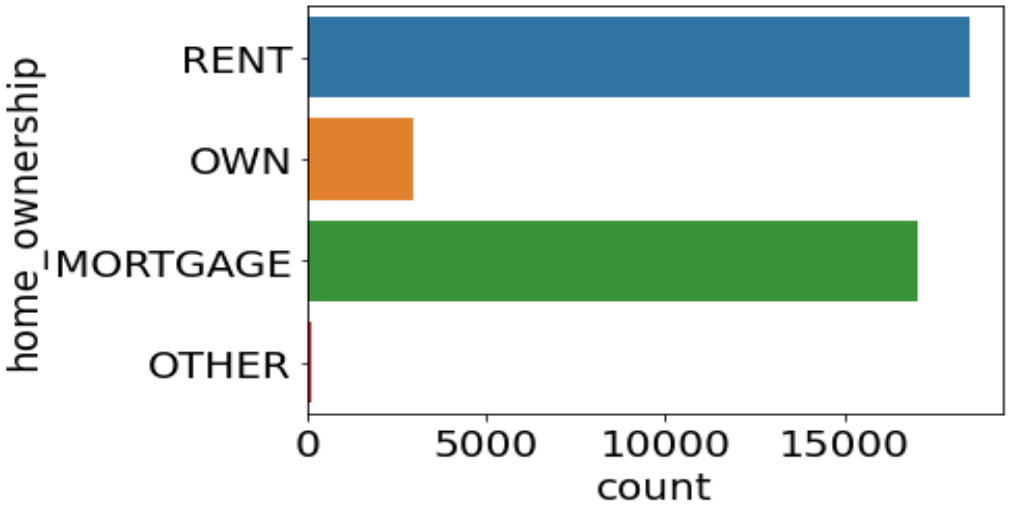
Univariate and Segmented Univariate Analysis

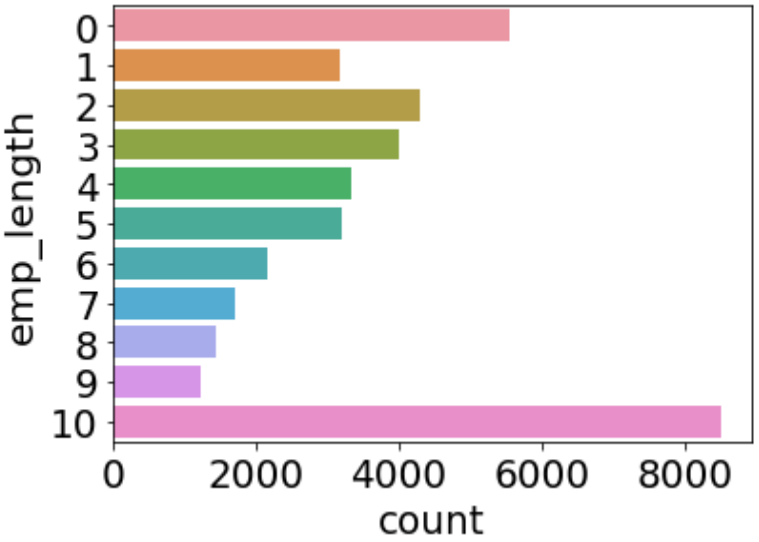
Exploratory Data Analysis (EDA)

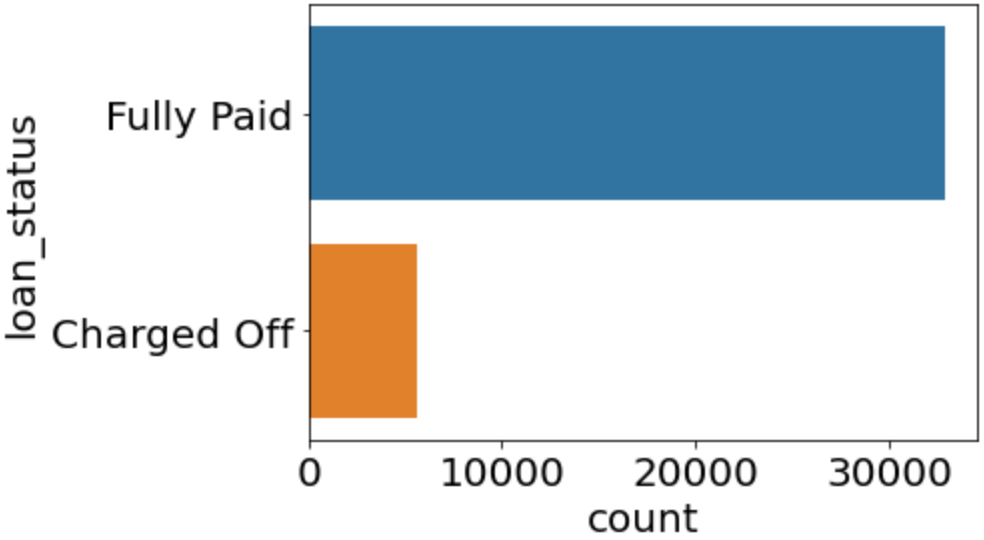


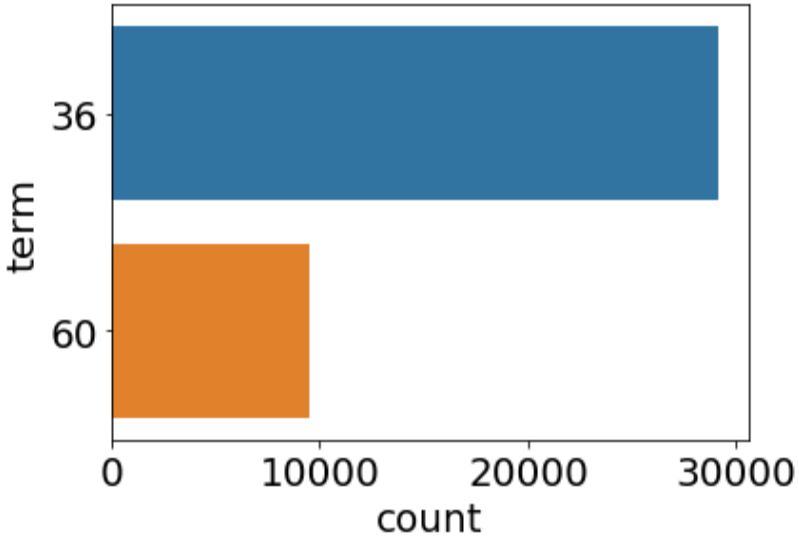


Exploratory Data Analysis (EDA)





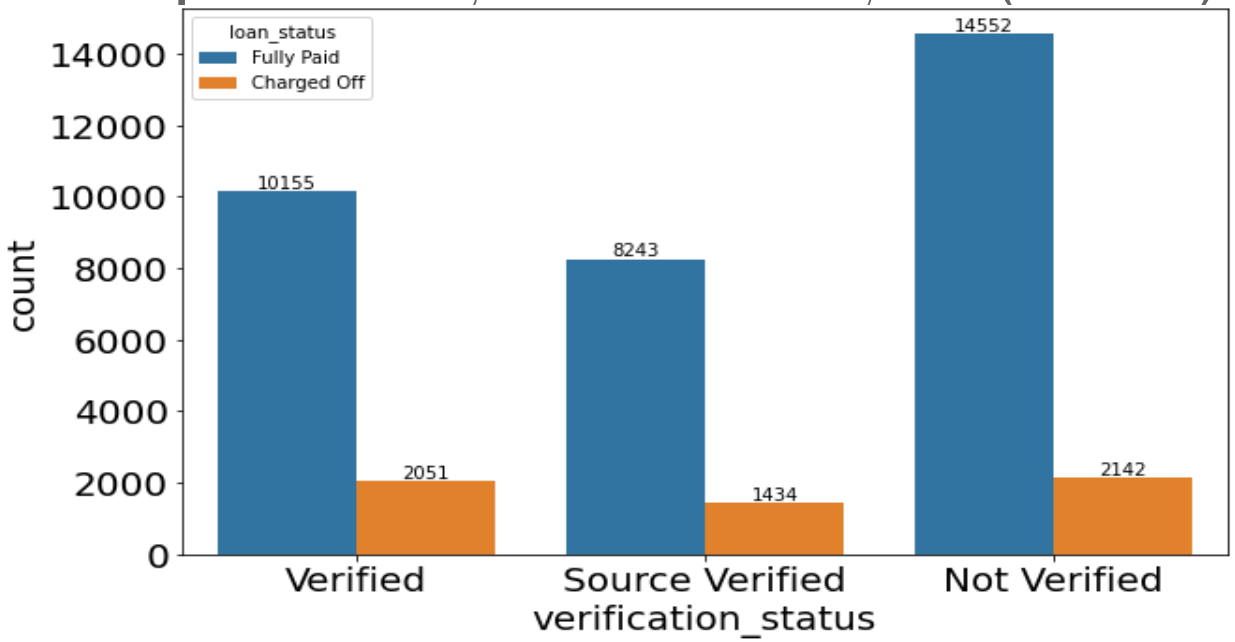


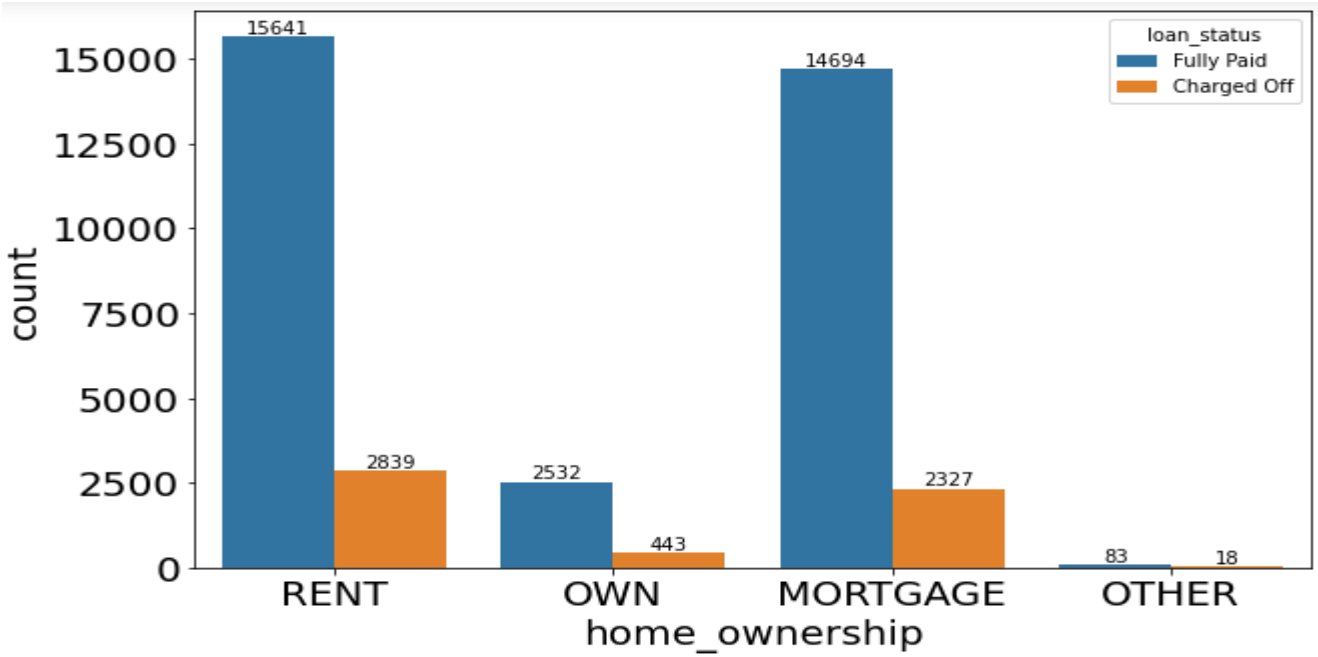


Exploratory Data Analysis (EDA)

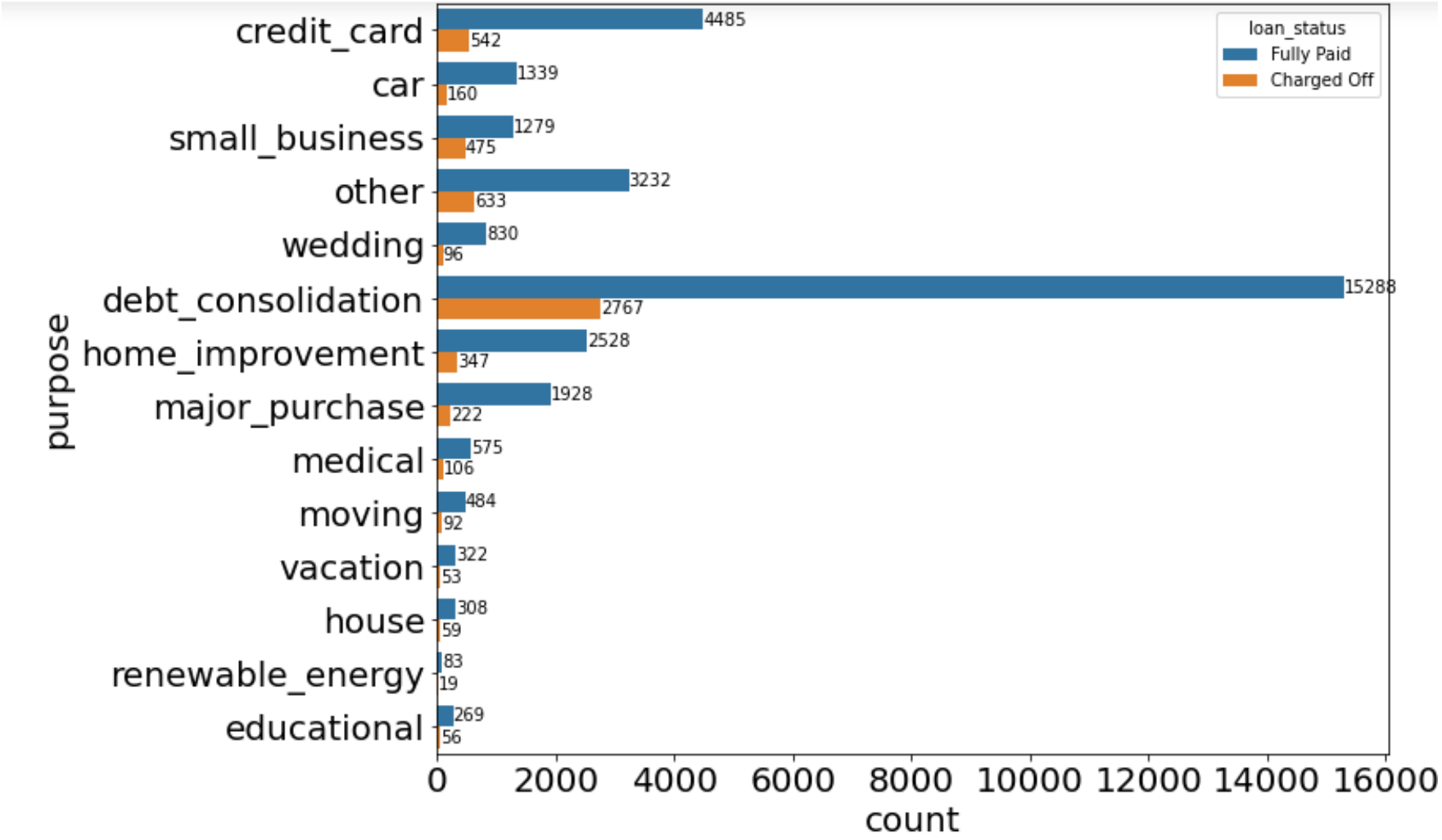
- Important Observations:
  - The vast majority of people use the loan for debt consolidation purposes.
  - The majority of people's incomes have not been verified by LC.
  - Most people have either a rented or mortgaged home.
  - The proportion of people who fully repay their loans is quite high.
  - The loan has mostly been taken out by people who have worked for at least ten years.
  - The majority of people prefer to repay their loans in 36 month installments.

Exploratory Data Analysis (EDA)

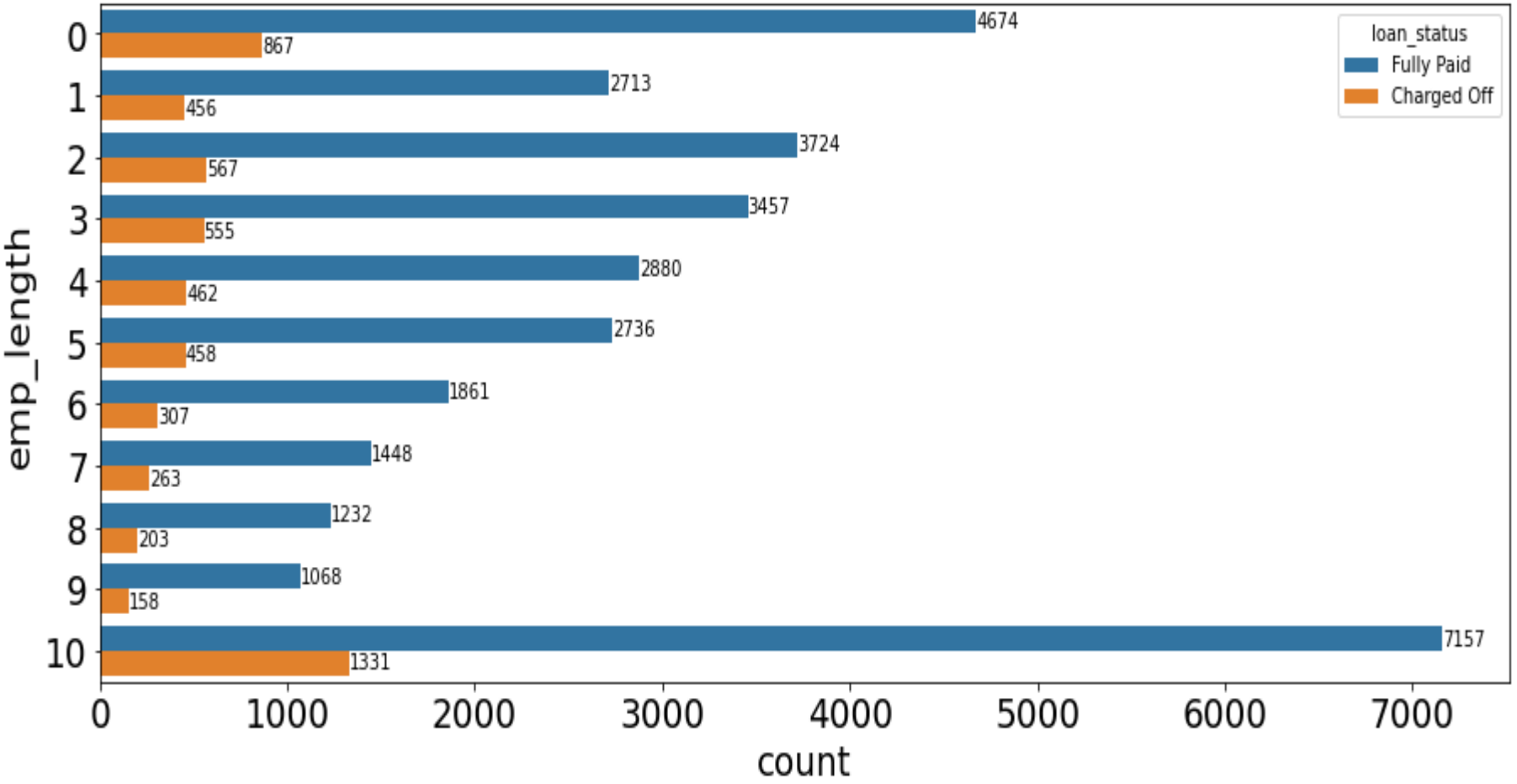




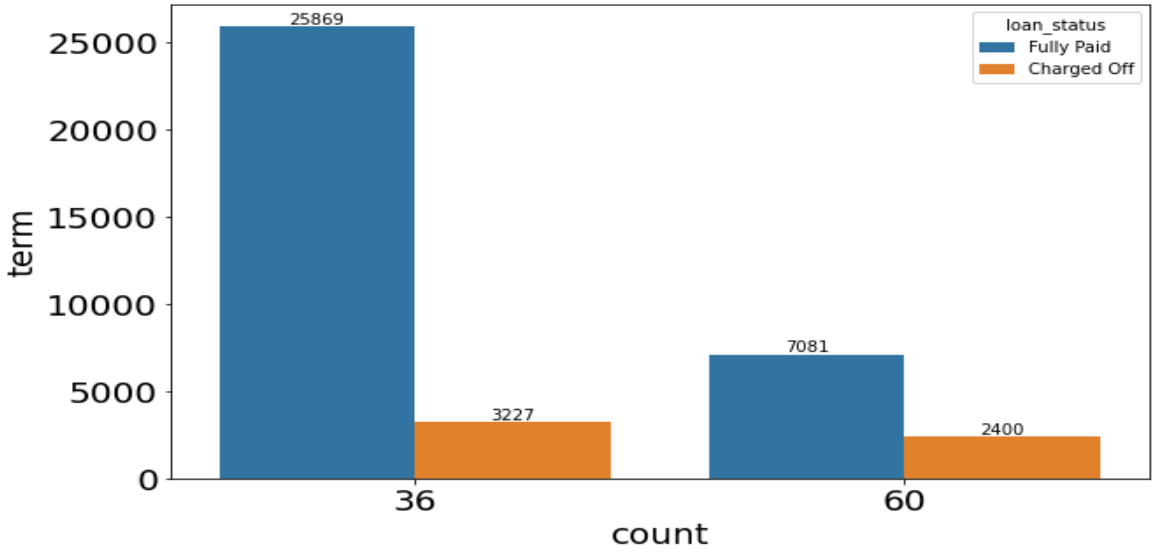
Exploratory Data Analysis (EDA)

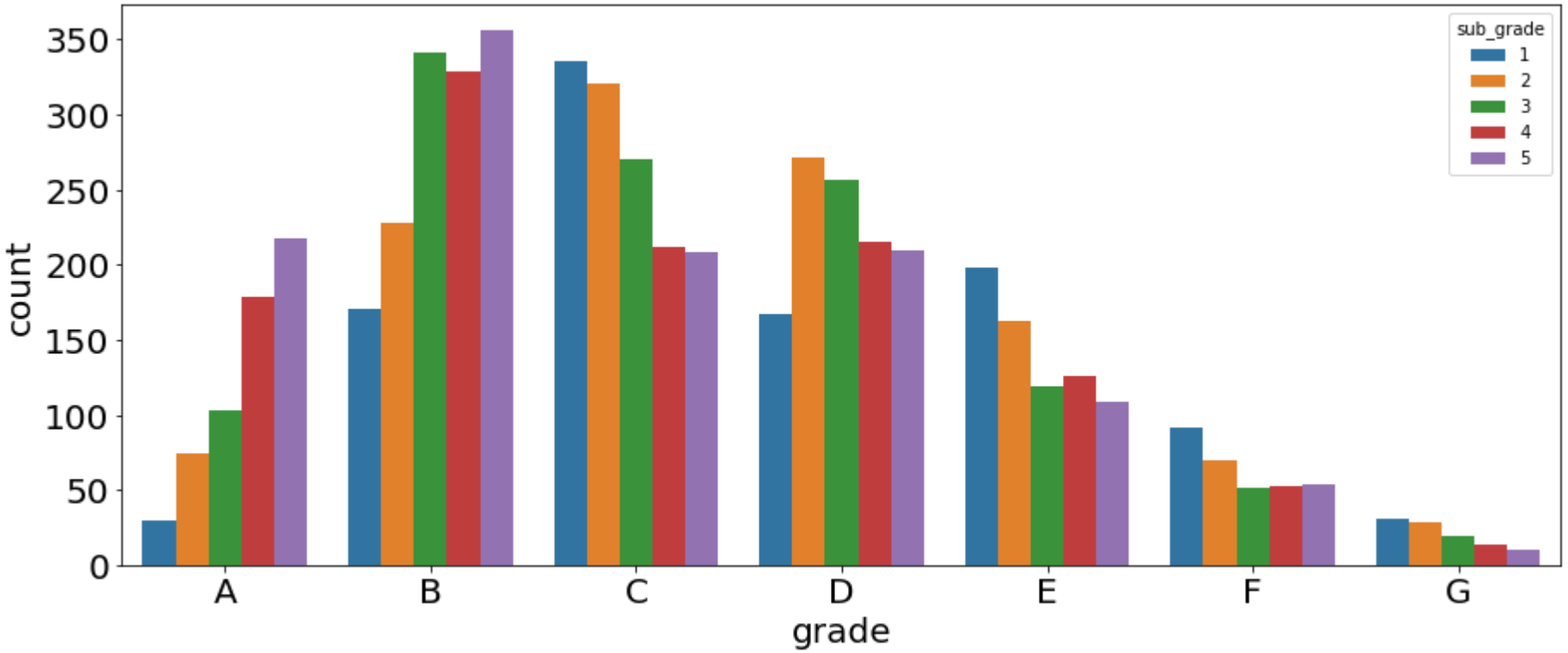


Exploratory Data Analysis (EDA)

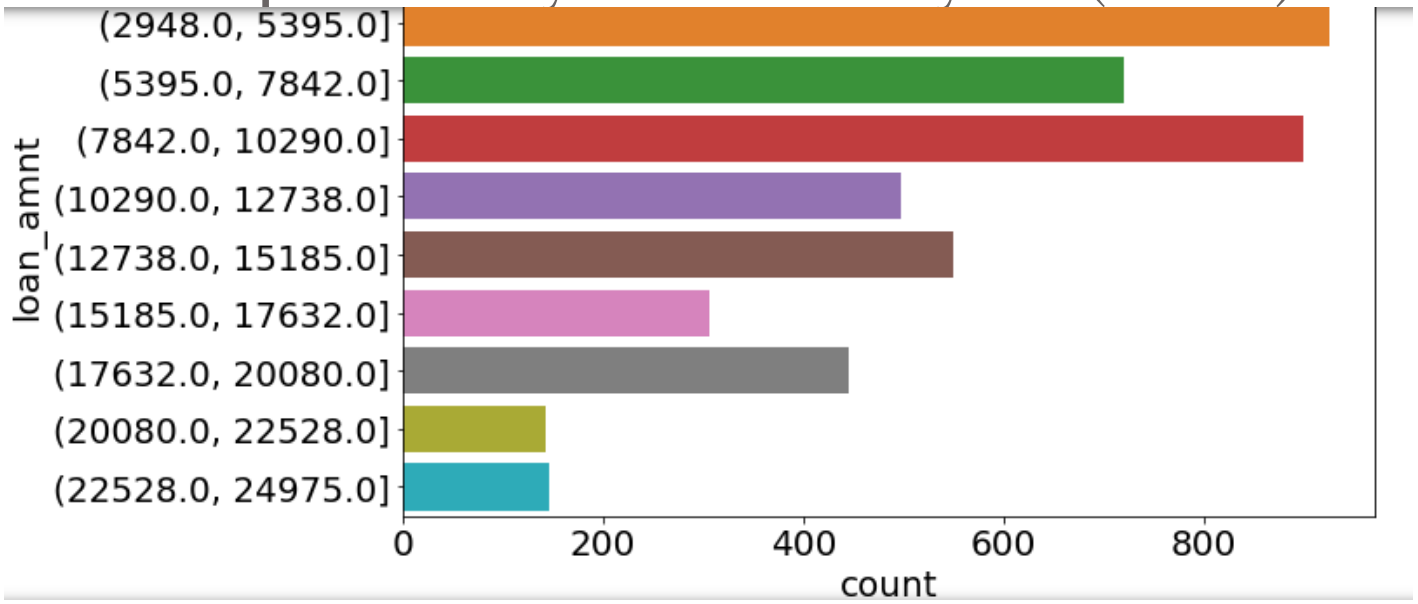


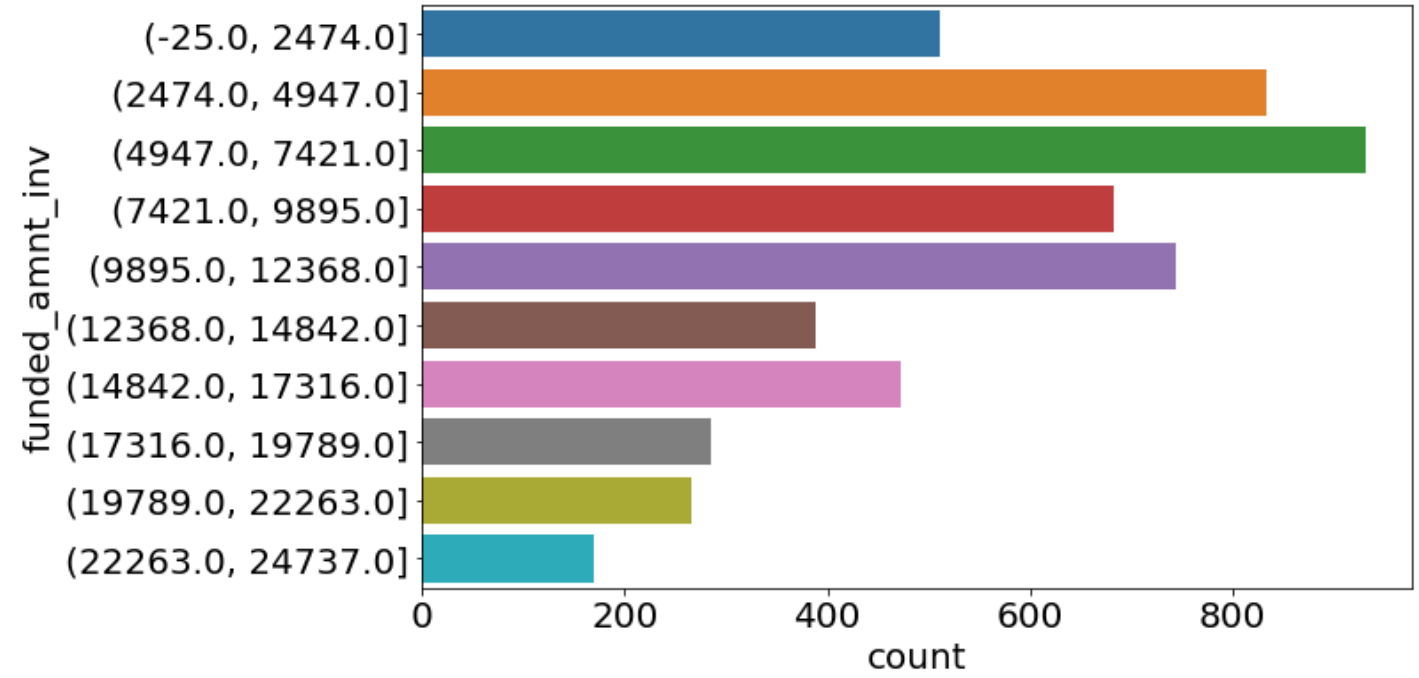
Exploratory Data Analysis (EDA)



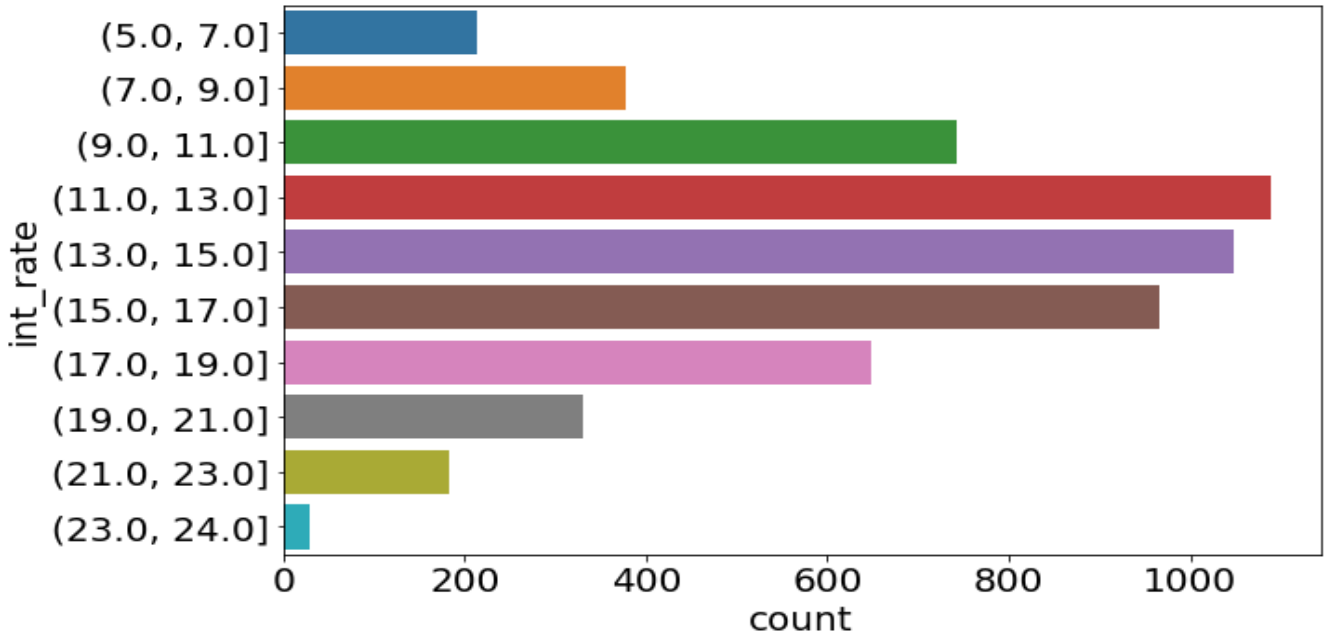


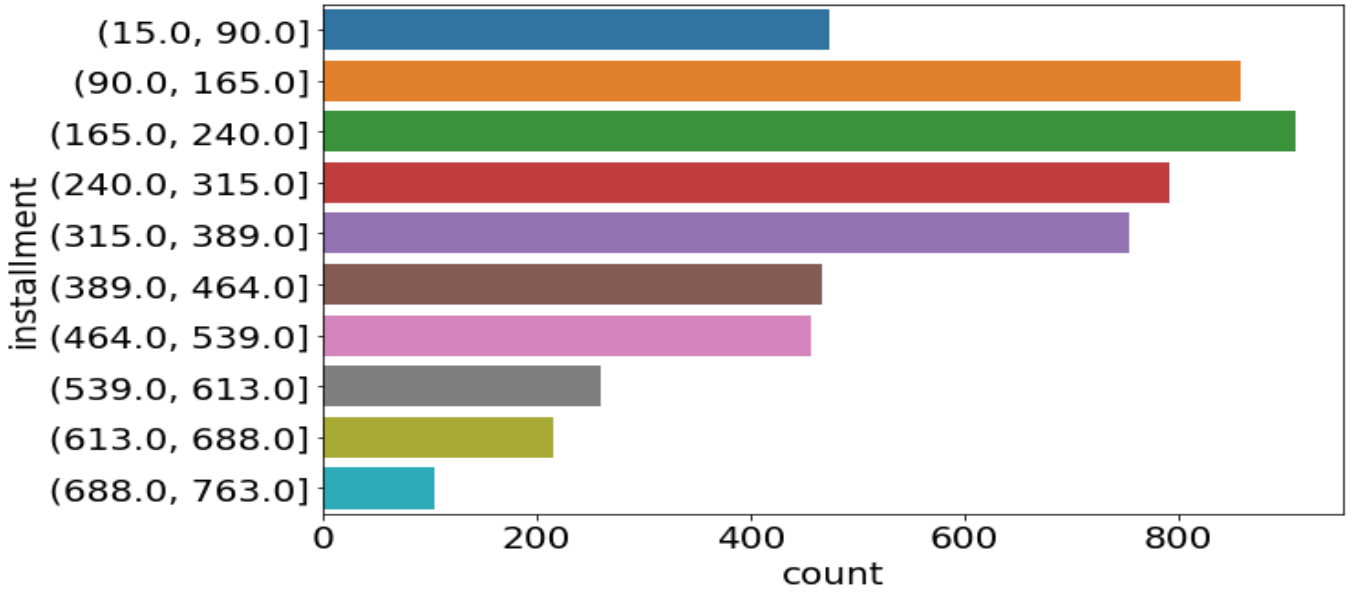
Exploratory Data Analysis (EDA)



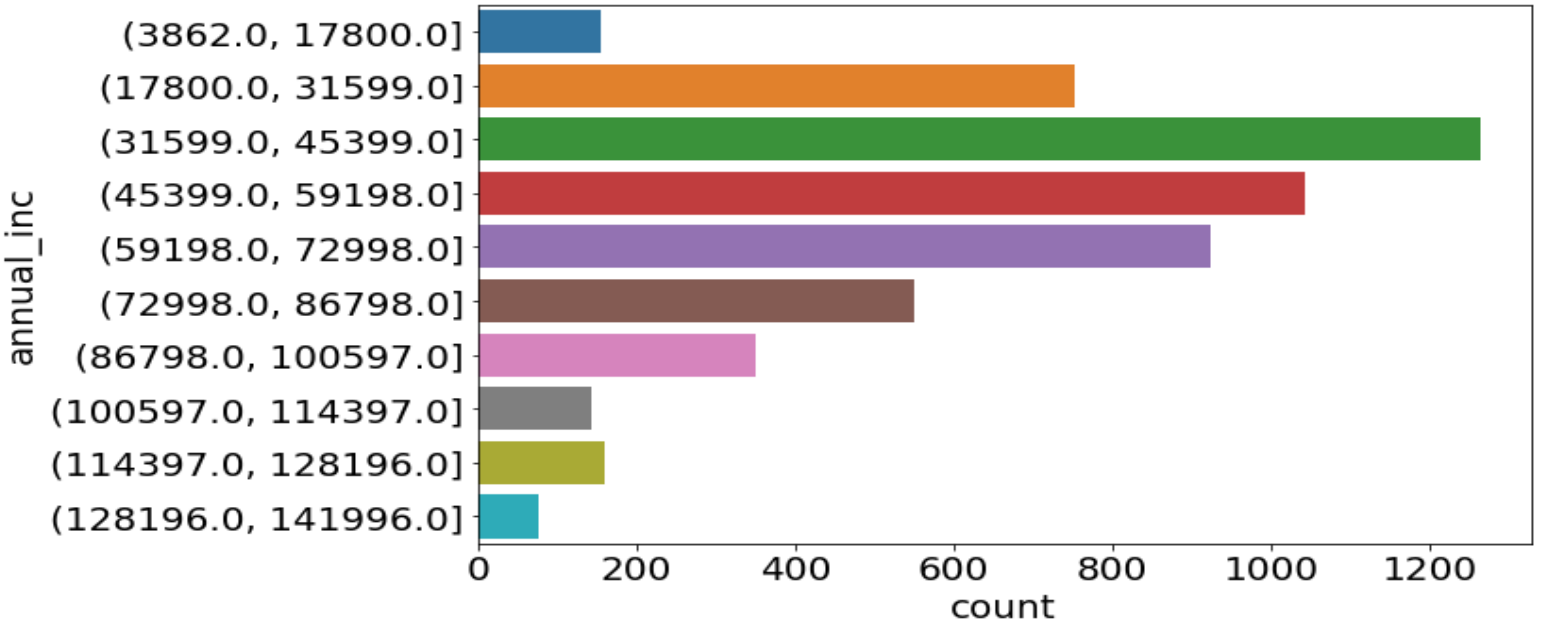


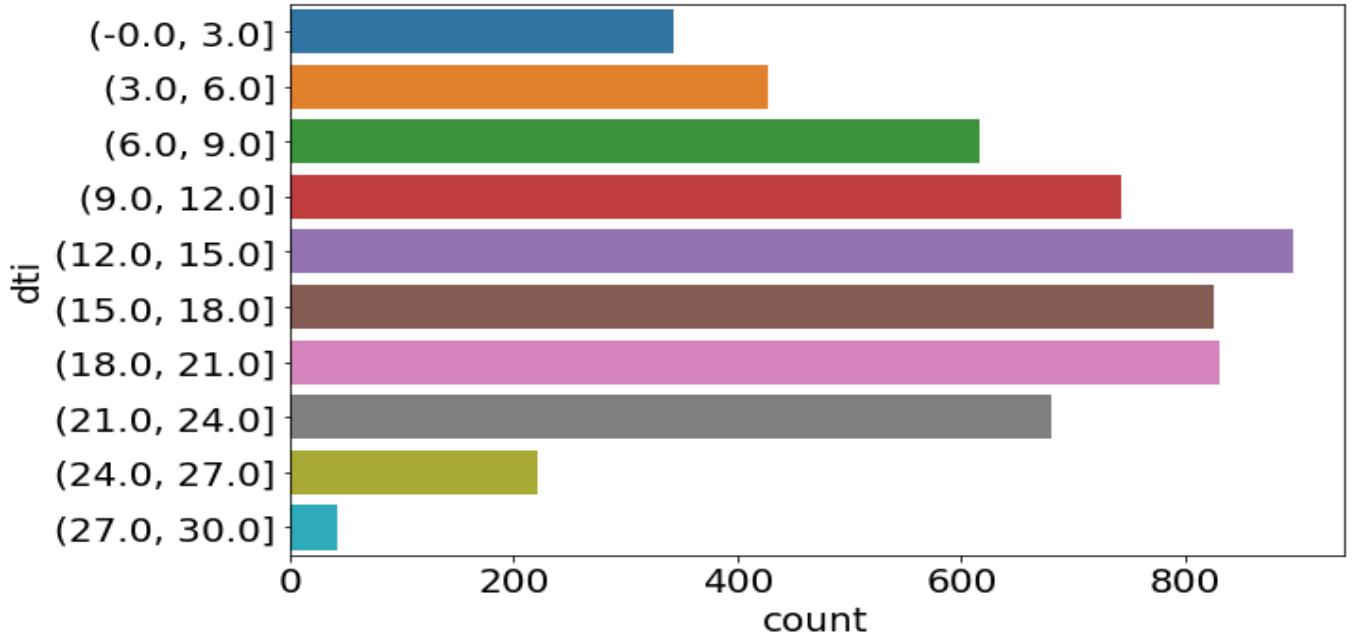
Exploratory Data Analysis (EDA)



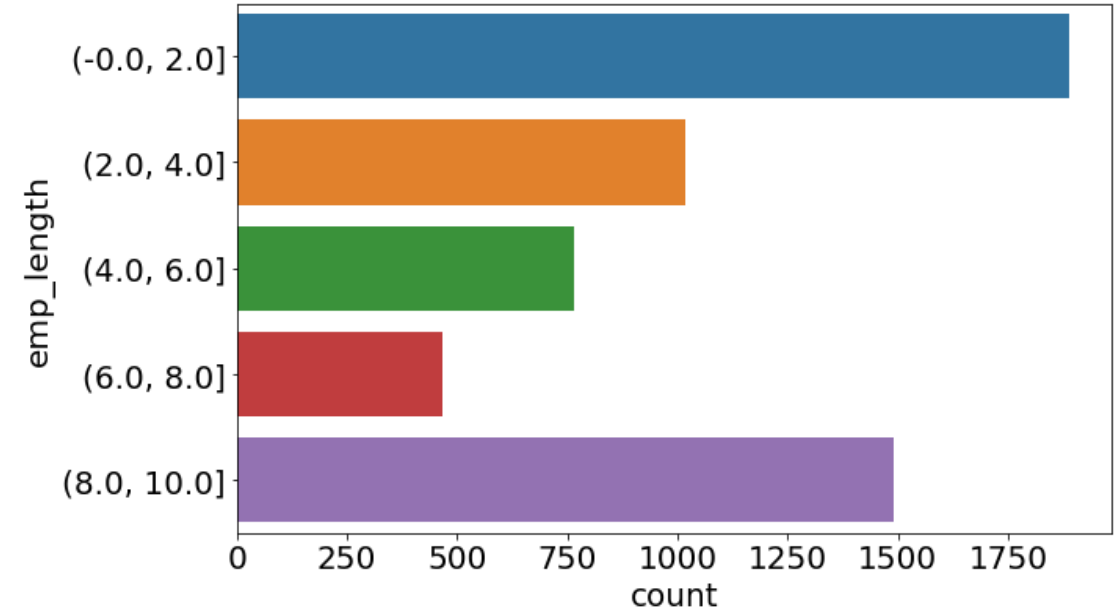


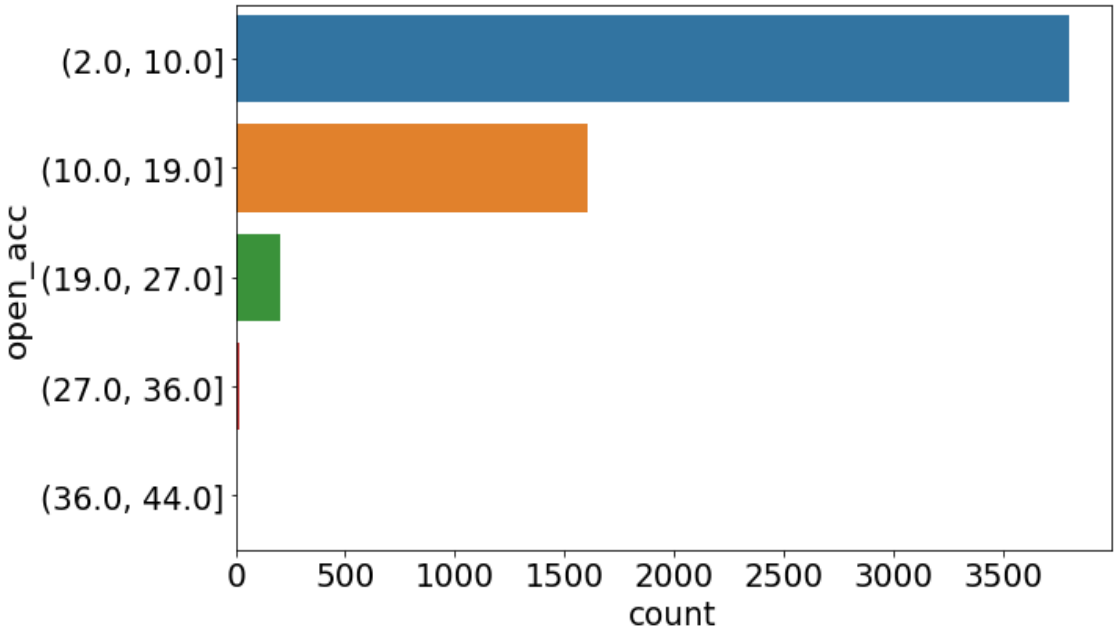
Exploratory Data Analysis (EDA)



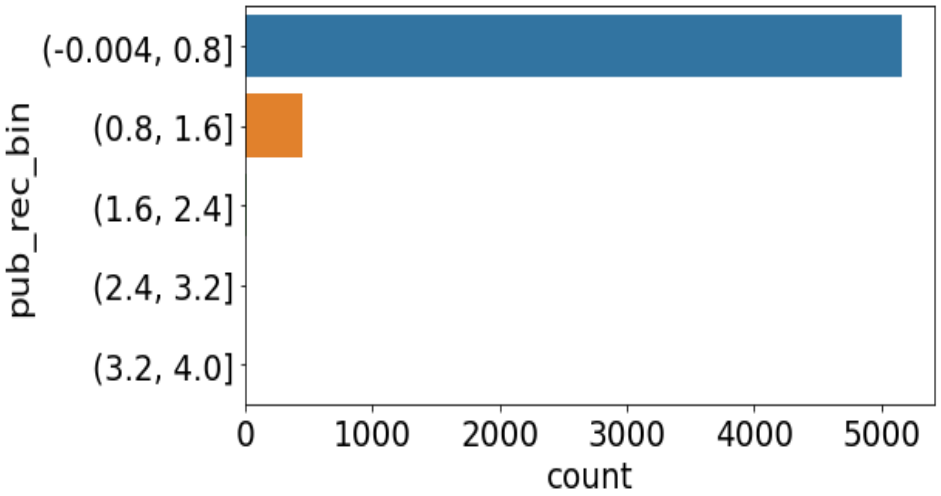


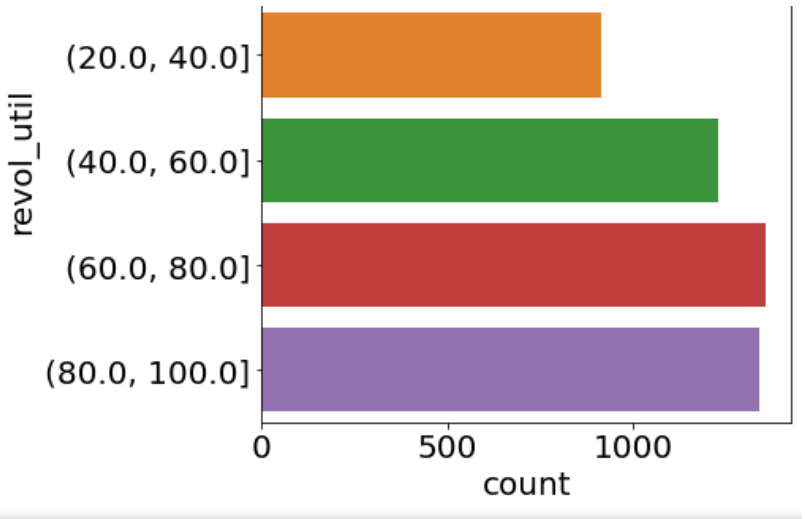
Exploratory Data Analysis (EDA)

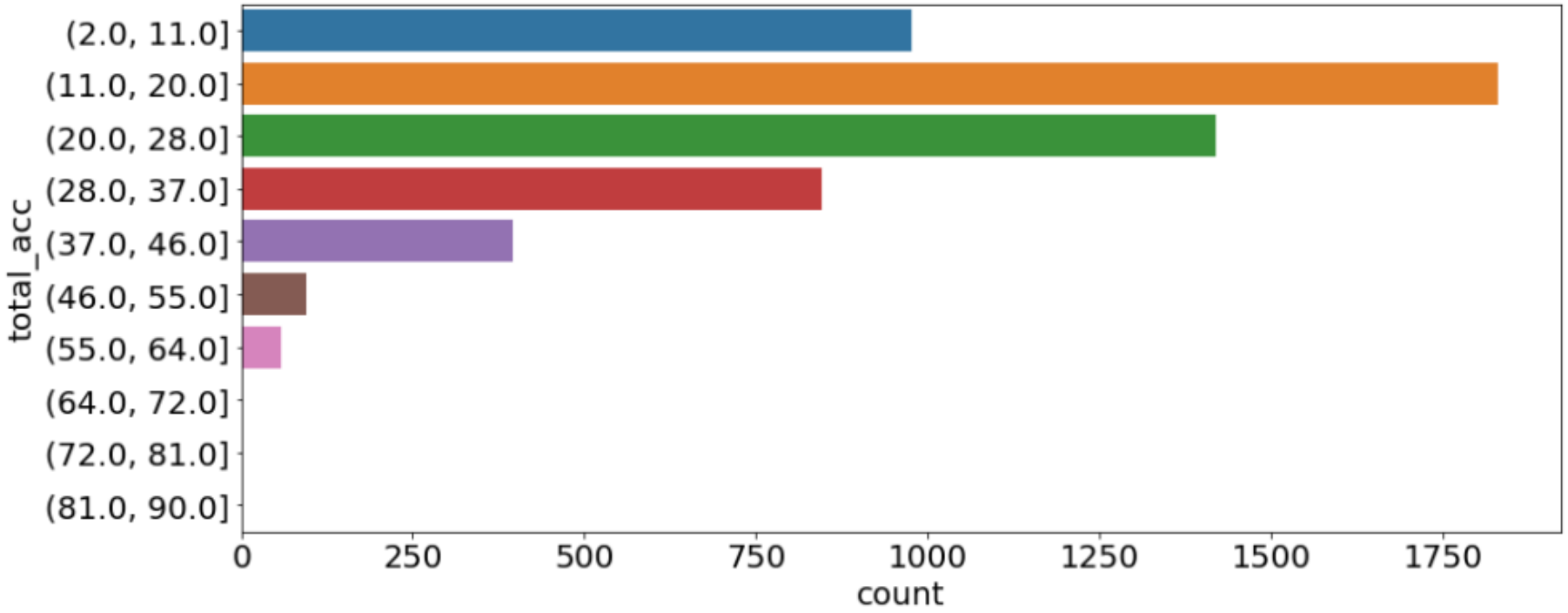




Exploratory Data Analysis (EDA)







Exploratory Data Analysis (EDA)

- **The most possible loan defaulters are -**
  - The customers whose income source has not been verified by LC.
  - The customers who have rented home.
  - The customers who have taken loans for debt consolidation.
  - The customers who have been labelled as B5 in terms of loan grade.
  - Customers whose loan amount is between 2948 and 5395 INR.
  - Customers whose invested fund amount is between 4947 and 7421 INR.

Exploratory Data Analysis (EDA)

- Customers with an interest rate of between 11 and 13.
- Customers whose installment amount is between 165 and 240 INR.
- Customers with an annual income between 31599 and 45399 INR.
- Customers with DTIs ranging from 12 to 15.
- Customers with an employment duration of 0 to 2 years.
- The customers whose number of open credit lines in the borrower's credit file is within 2 to 10.

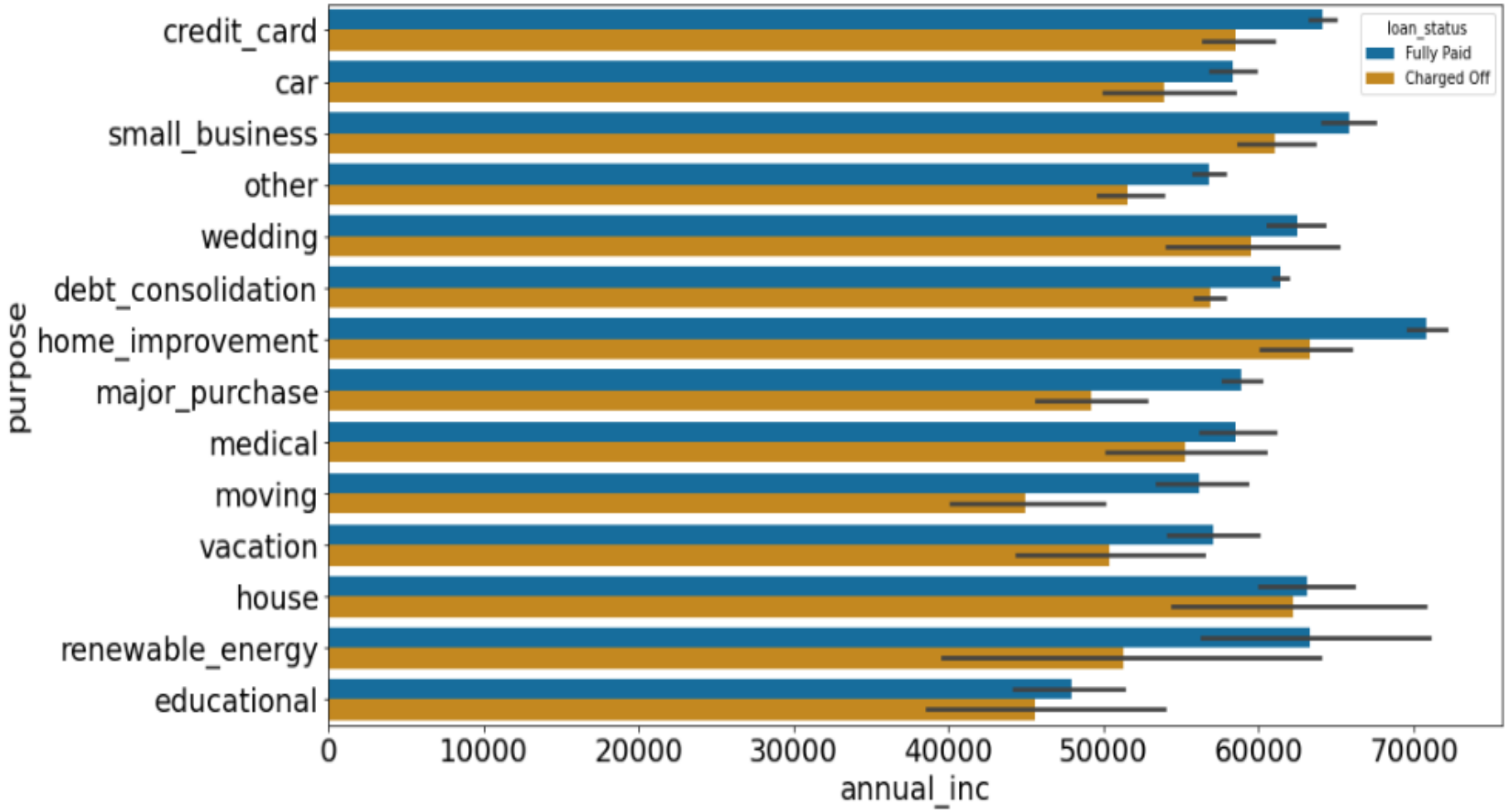
Exploratory Data Analysis (EDA)

- Customers who have one derogatory public record.
- Customers with a revolving line utilization rate of 60 to 80 percent.
- The customers whose total number of credit lines currently in the borrower's credit file is between 11 and 20.

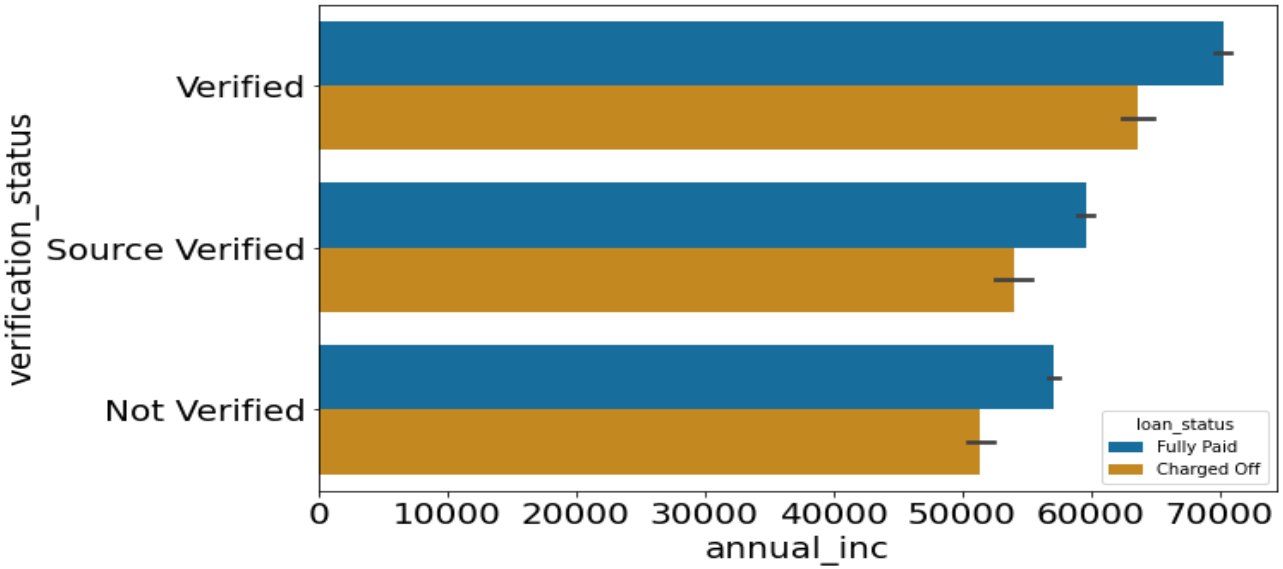
Exploratory Data Analysis (EDA)

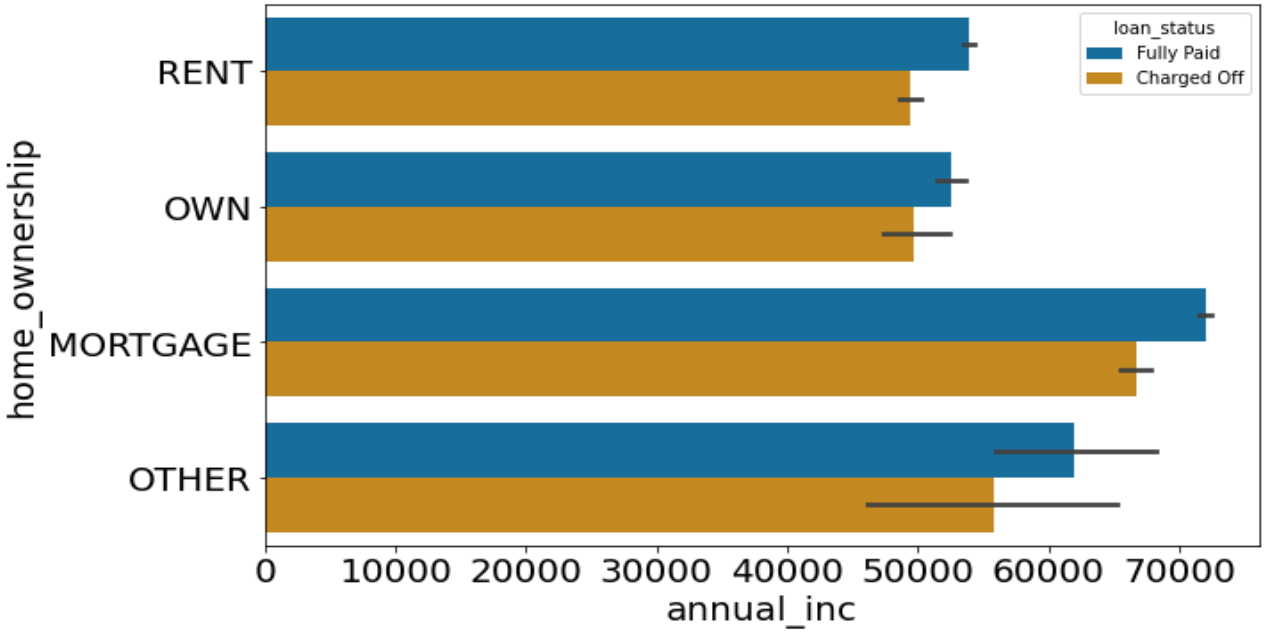
Bivariate Analysis

Exploratory Data Analysis (EDA)

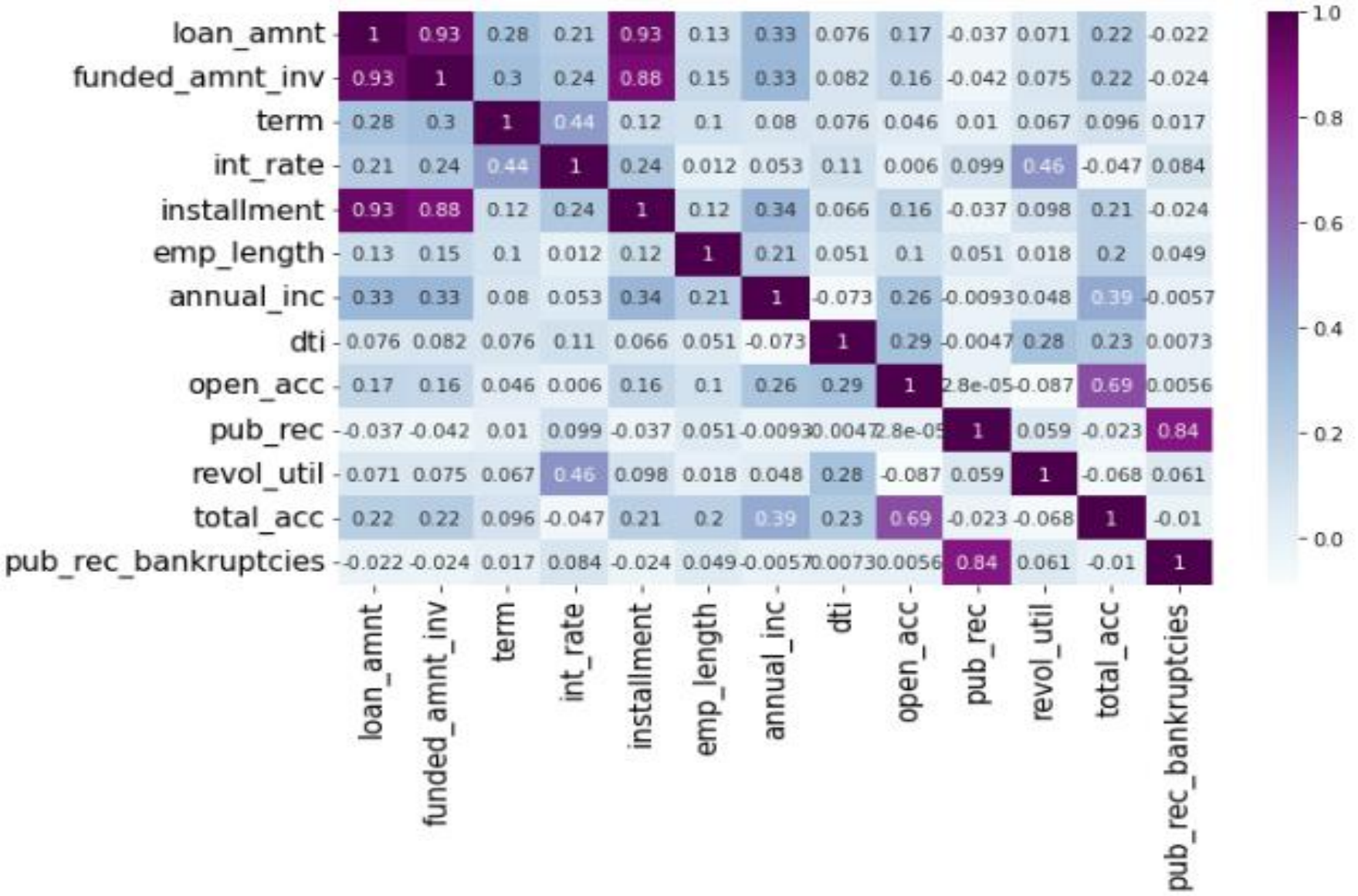


Exploratory Data Analysis (EDA)

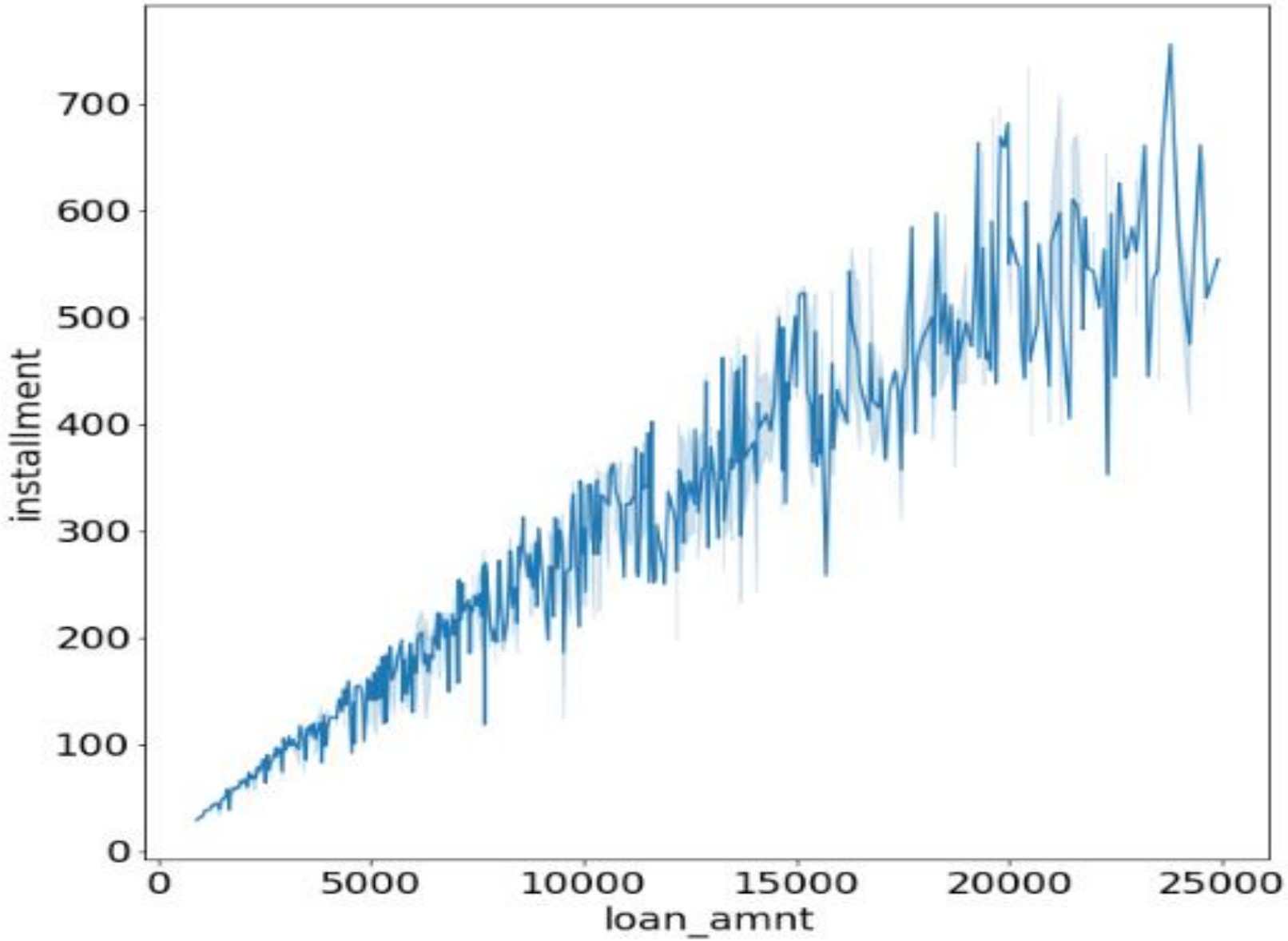




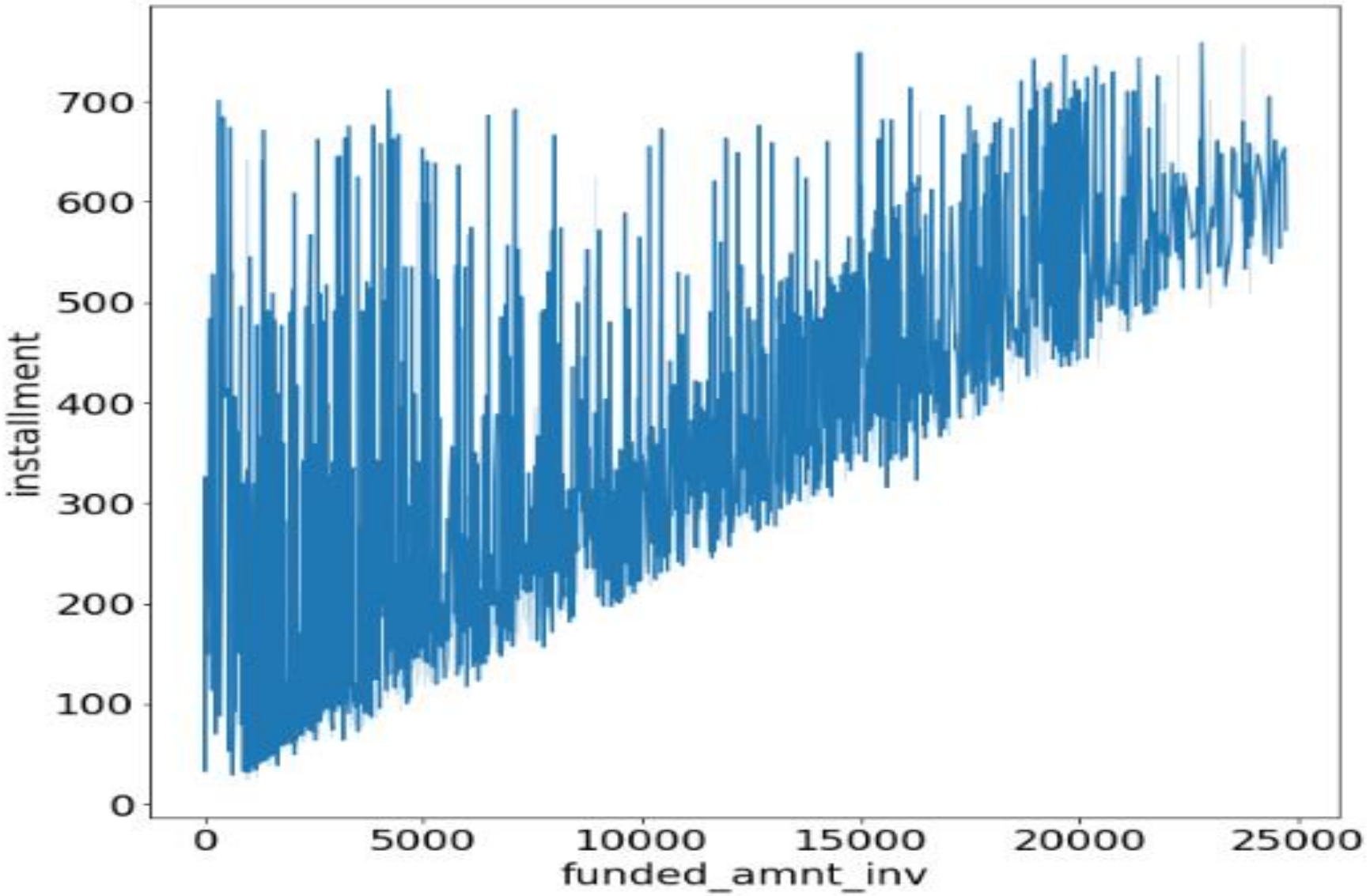
Exploratory Data Analysis (EDA)



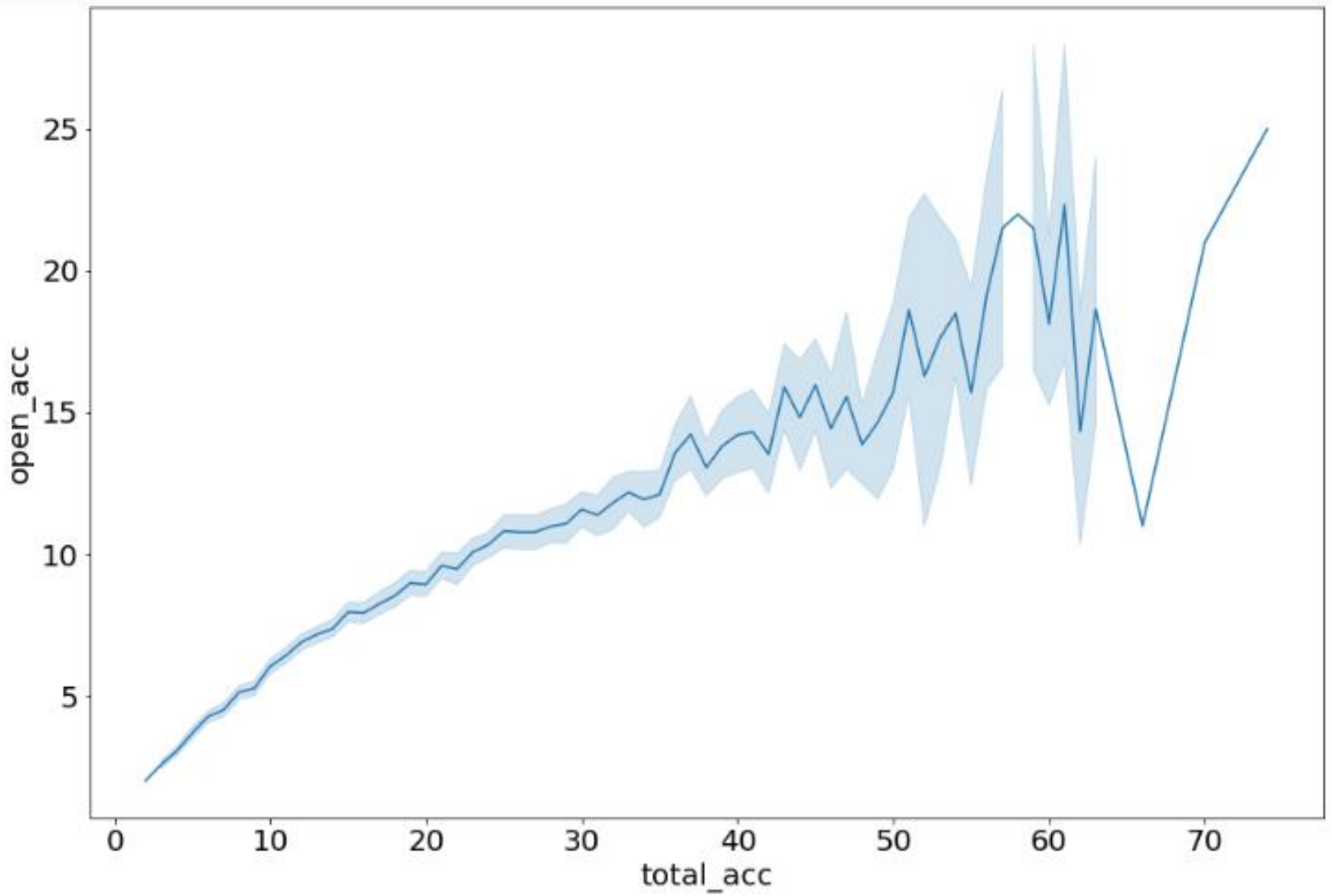
Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)



Exploratory Data Analysis (EDA)

- **The most possible loan defaulters are -**
  - Customers with an annual income of more than 60,000 INR who are taking out a loan for home improvement.
  - Customers who earn more than 60,000 INR per year and whose source of income has been verified by LC.
  - Customers who have an annual income of more than 60,000 INR and have a mortgaged home.

Exploratory Data Analysis (EDA)

- Other important observations are -
  - In the case of possible defaulters, the installment amount is strongly correlated with the loan amount.
  - In the case of possible defaulters, the invested fund amount is strongly correlated with the loan amount.
  - In the case of possible defaulters, the total number of credit lines currently in the borrower's credit file is strongly correlated with the number of open credit lines in the borrower's credit file.


The End