Lending Club Case Study

By:
Lakshmi Prasanna
Deepak Lamba

Problem Statement -

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Steps followed -

- Data Sourcing
- Data Cleaning
- Univariate / Bivariate Analysis / Derived Metrics
- Observations

Data Sourcing -

- Source File – loan.csv
- Python code -
  - df_loan = pd.read_csv('loan.csv', low_memory=False)

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_pas
0	1077501	1296599	5000	5000	4975.00	36 months	10.65%	162.87	B	B2	...	NaN	
1	1077430	1314167	2500	2500	2500.00	60 months	15.27%	59.83	C	C4	...	NaN	
2	1077175	1313524	2400	2400	2400.00	36 months	15.96%	84.33	C	C5	...	NaN	
3	1076863	1277178	10000	10000	10000.00	36 months	13.49%	339.31	C	C1	...	NaN	
4	1075358	1311748	3000	3000	3000.00	60 months	12.69%	67.79	B	B5	...	NaN	

5 rows × 111 columns

Data Cleaning -

- It has been observed that some of columns have blank or NA as values. We cannot use those columns to do any analysis, hence we will delete them from our dataset.

    - df_loan = df_loan.dropna(axis=1, how='all’)

- Following columns have same values for all the rows and therefore they are not very useful for analysis

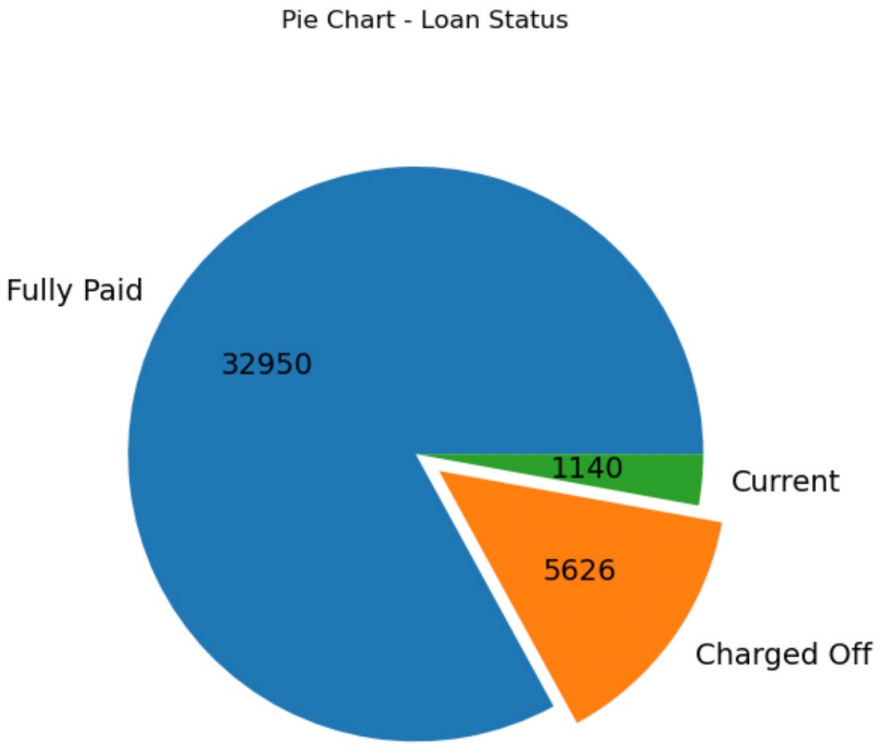
    1. pymnt_plan - value for this column is 'n' for all rows.
    2. policy_code - value for this column is '1' for all rows.
    3. application_type - value for this column is 'INDIVIDUAL' for all rows.
    4. initial_list_status - value for this column is 'f' for all rows.
    5. collections_12_mths_ex_med - value for this column is either '0' or NA for all rows.
    6. acc_now_delinq - value for this column is '0' for all rows.
    7. chargeoff_within_12_mths - value for this column is '0' for all rows.
    8. tax_liens - value for this column is '0' for all rows.

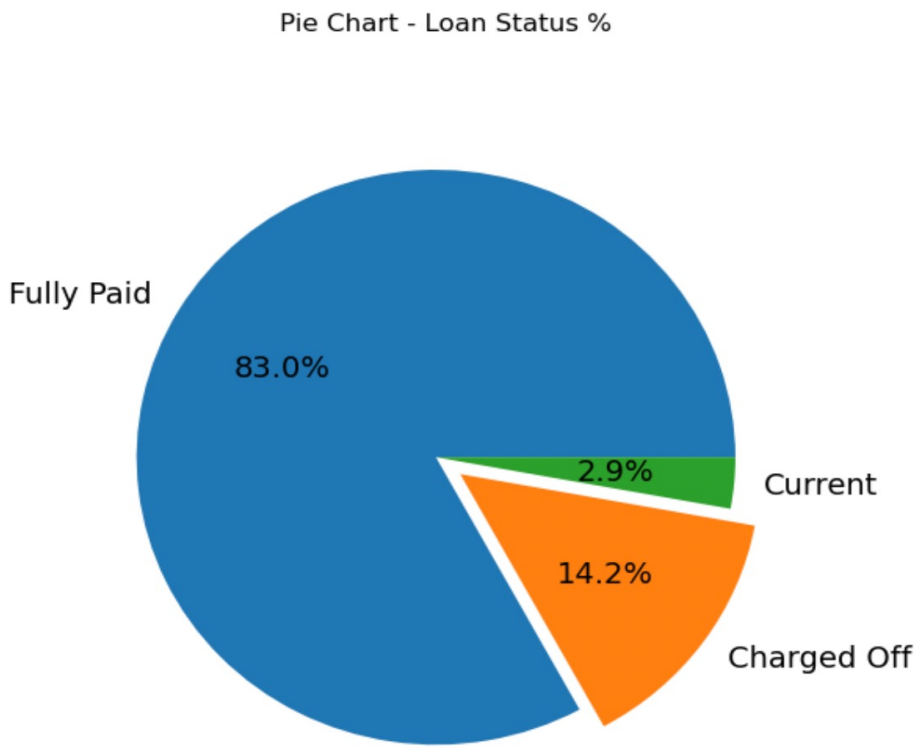
- Remove '%' from int_rate (Interest Rate) column and convert its datatype to float.
    - df_loan['int_rate'] = df_loan['int_rate'].apply(lambda x: float(x[0:-1]))

- Remove blank values, '%' from revol_util (Revolving line utilization rate) column and convert its datatype to float.
    - df_loan['revol_util'].fillna(df_loan['revol_util'].mode()[0], inplace=True)
    - df_loan['revol_util'] = df_loan['revol_util'].apply(lambda x: float(x[0:-1]))

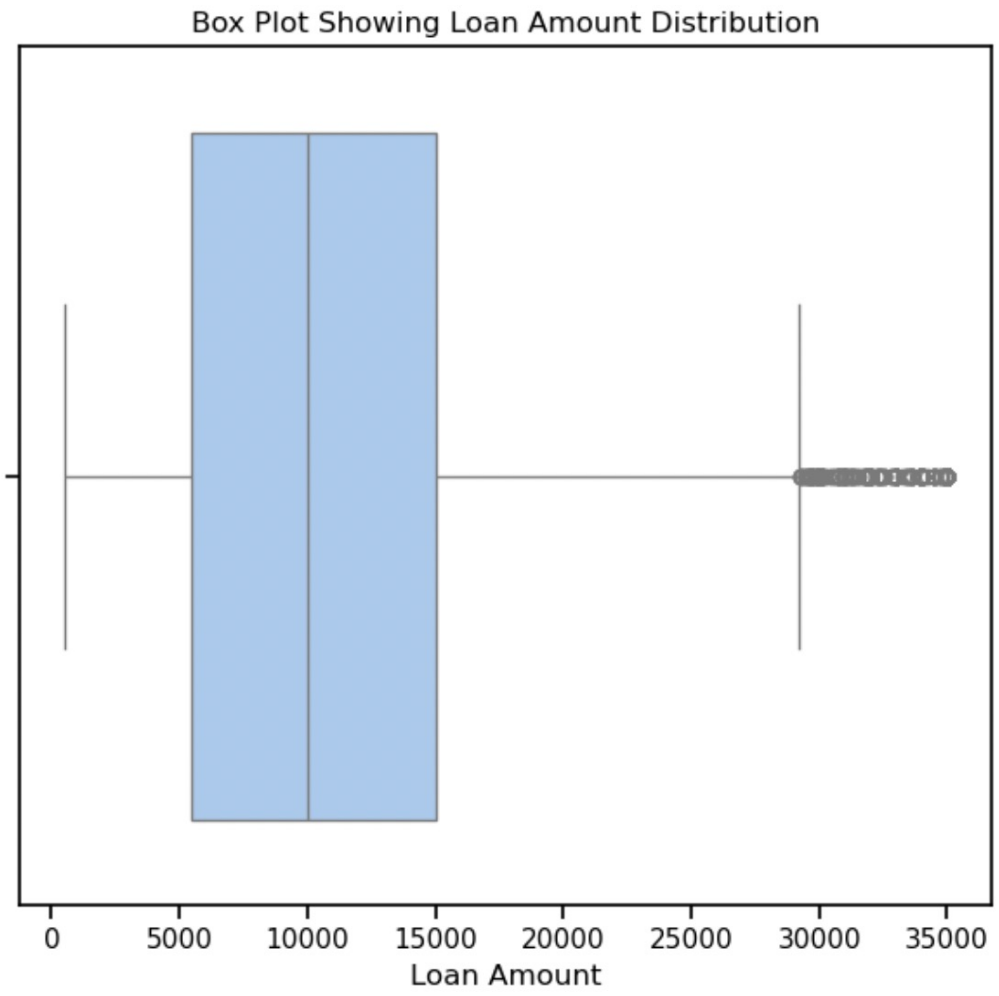
Univariate / Bivariate Analysis

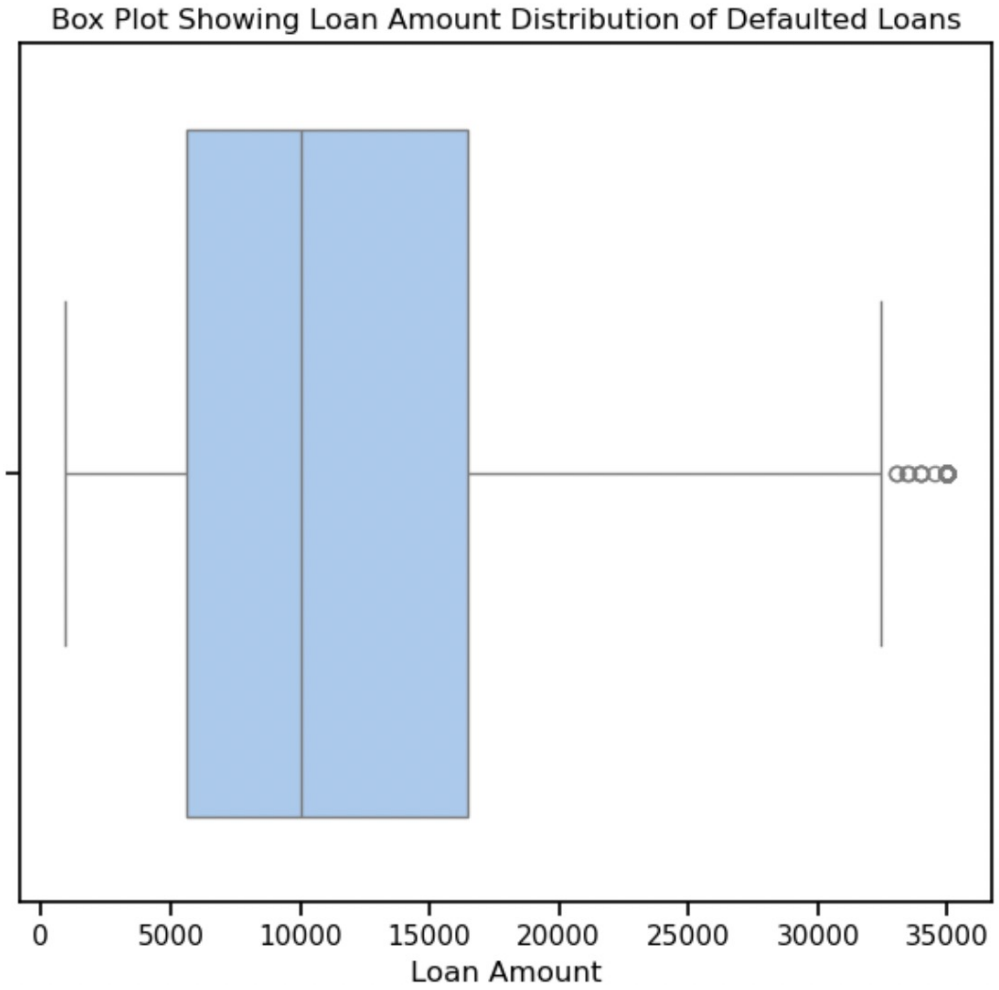
1. Variable - Loan status



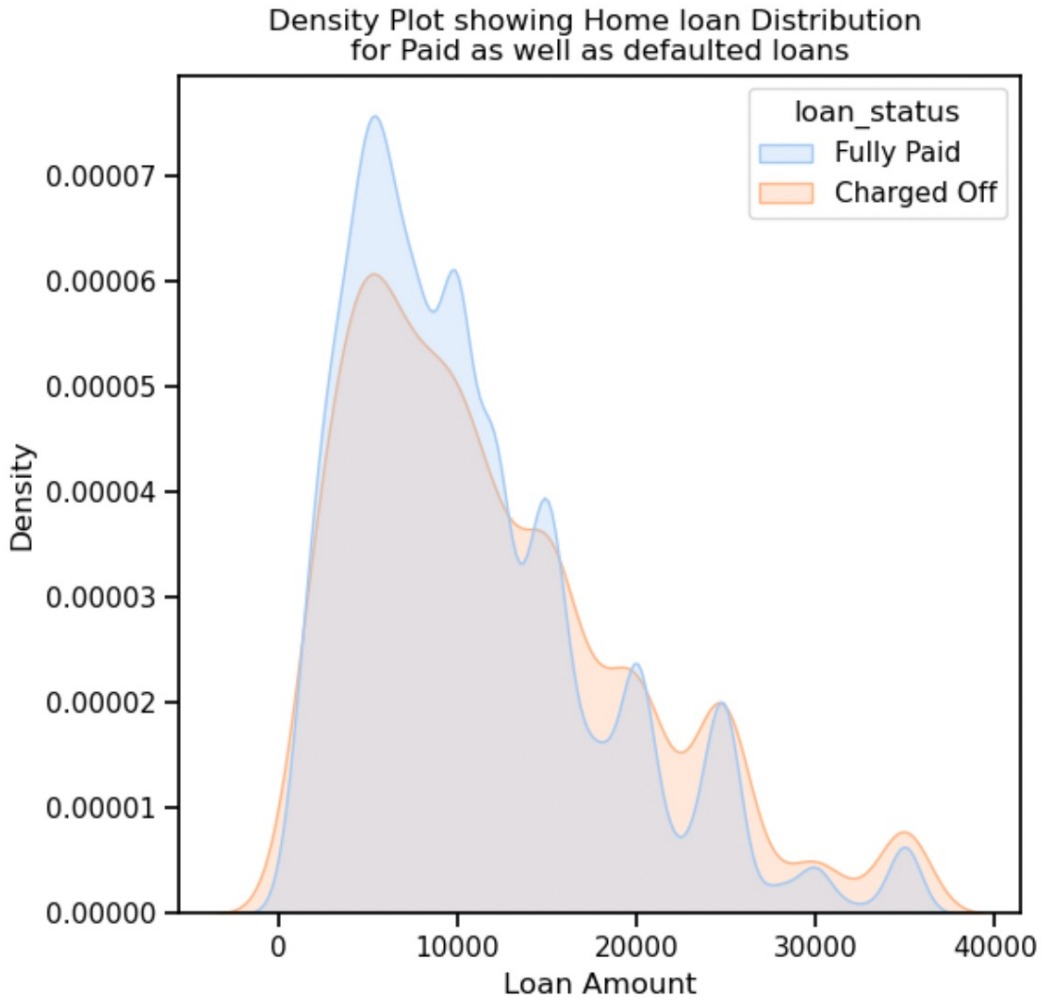


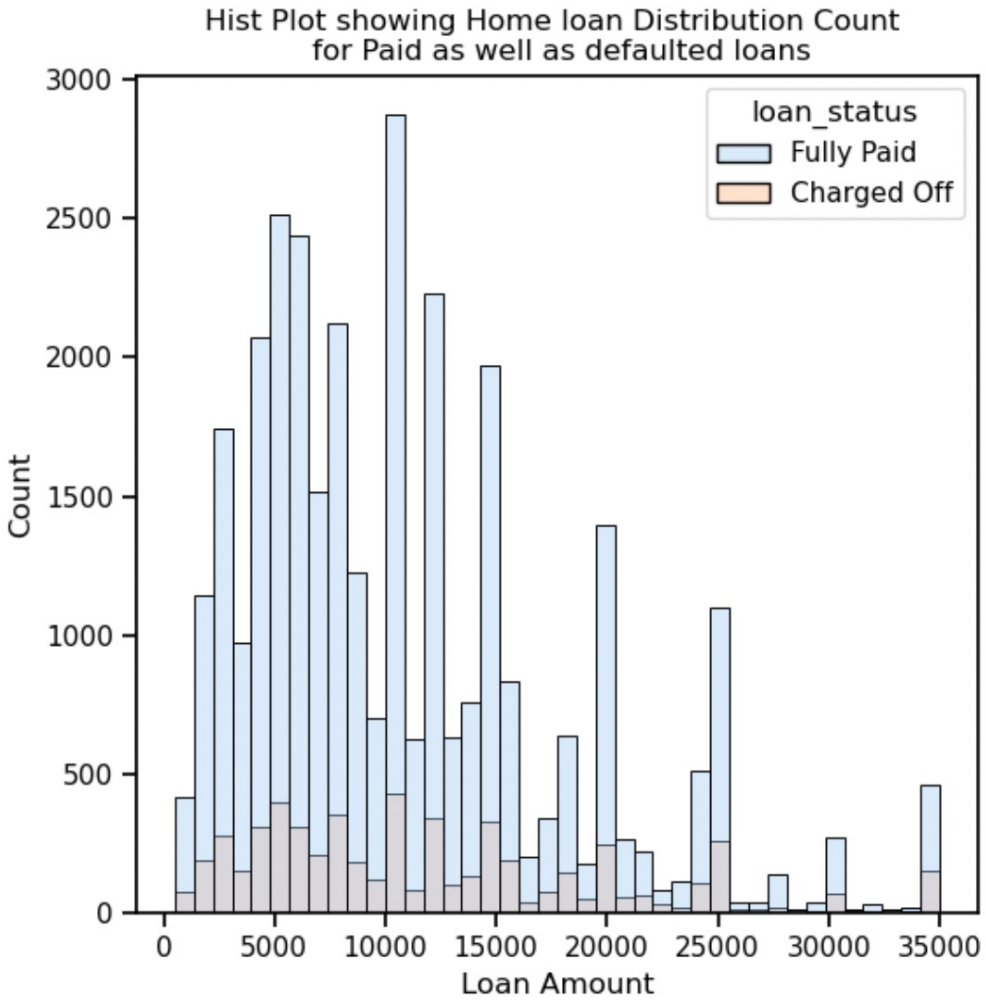
2. **Variable - Loan status**



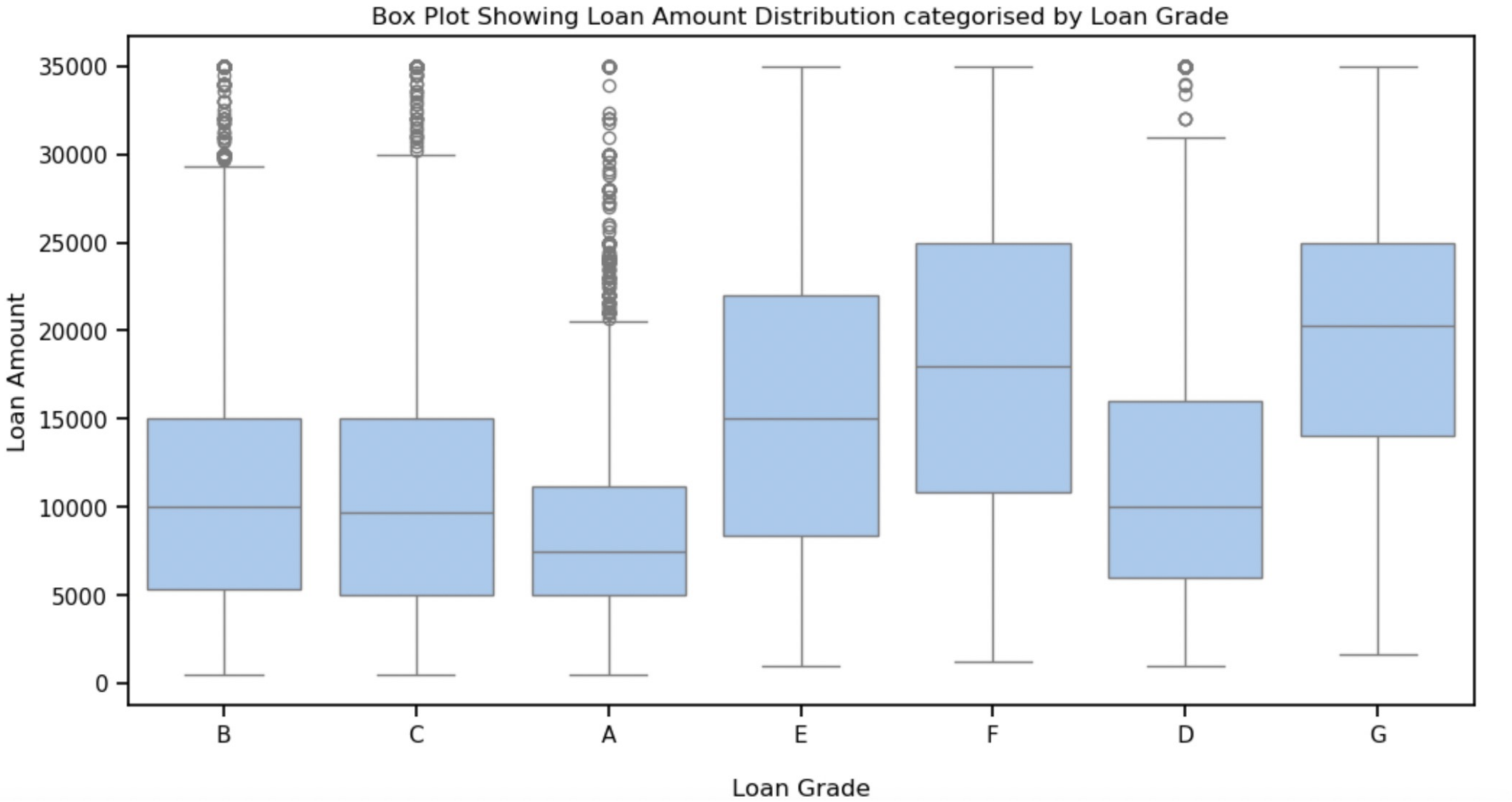


Variable - Loan status (cont..)

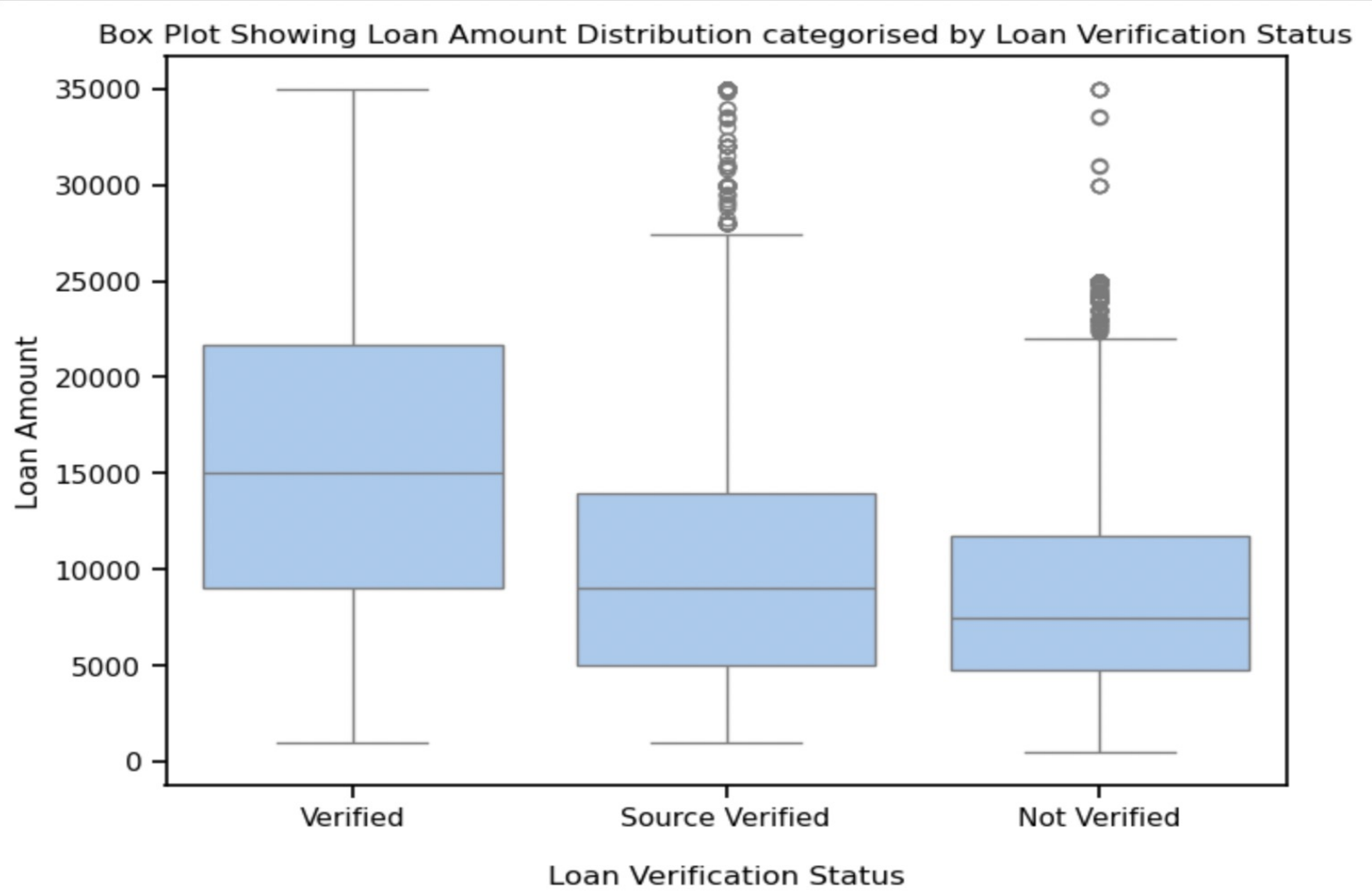




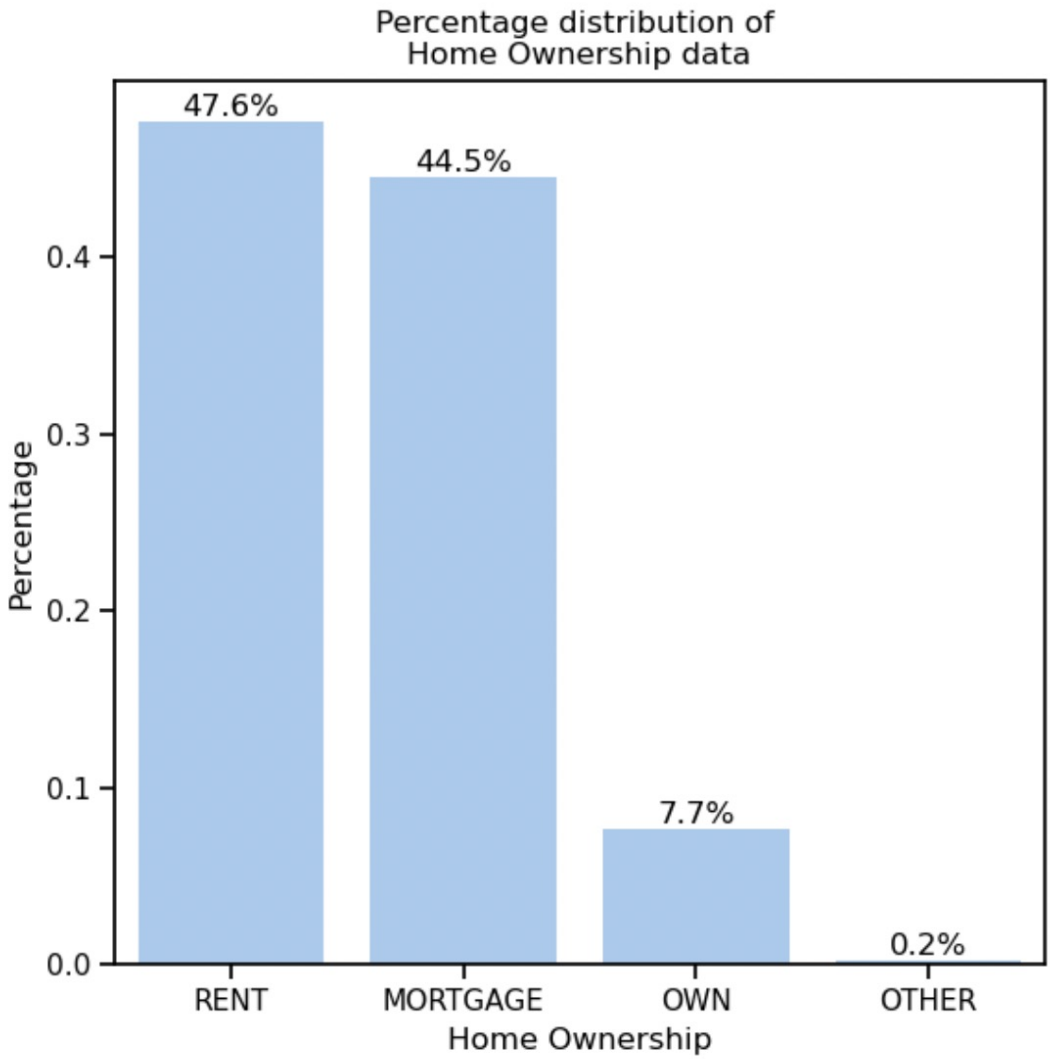
Variable - Loan status (cont..)

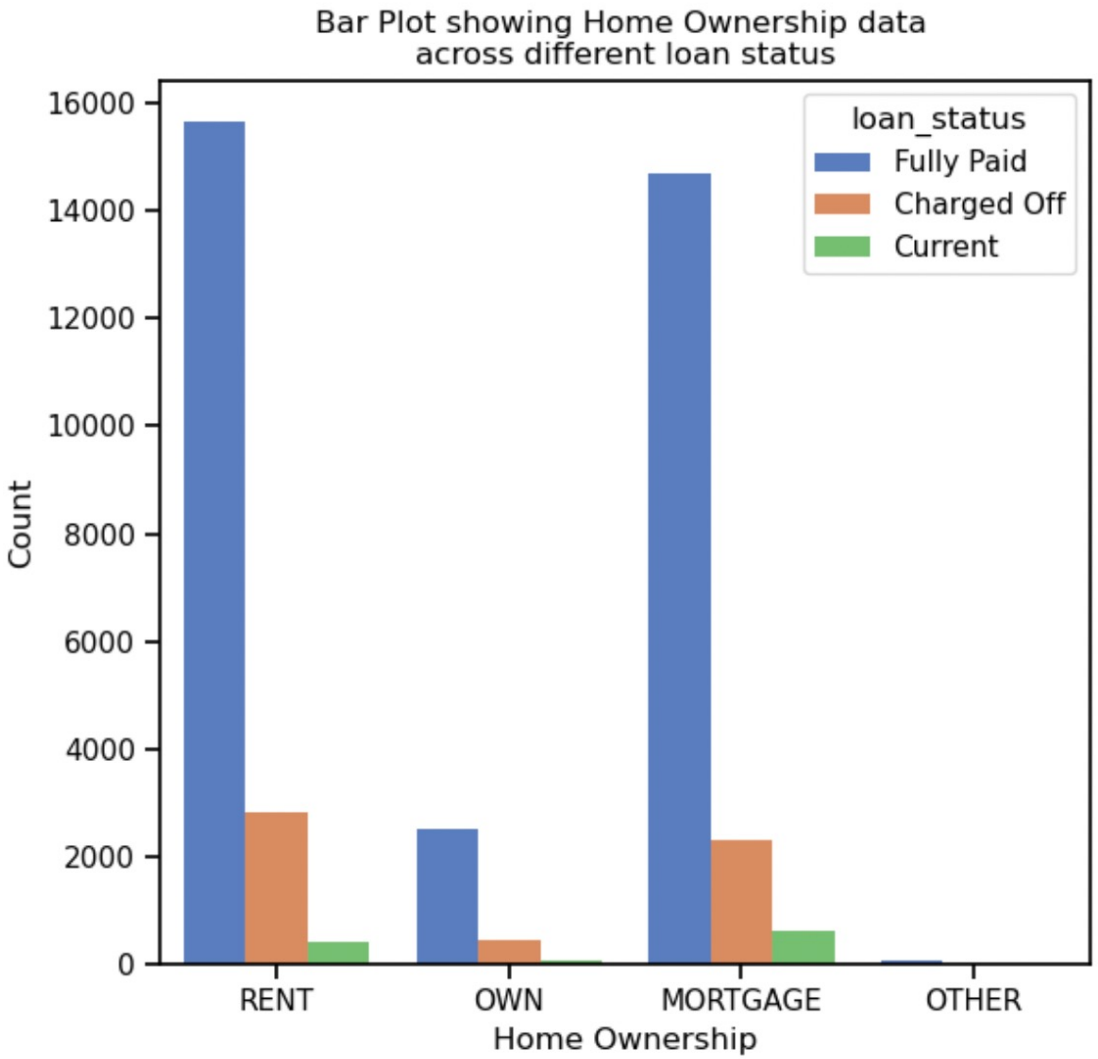


Variable - Loan status (cont..)

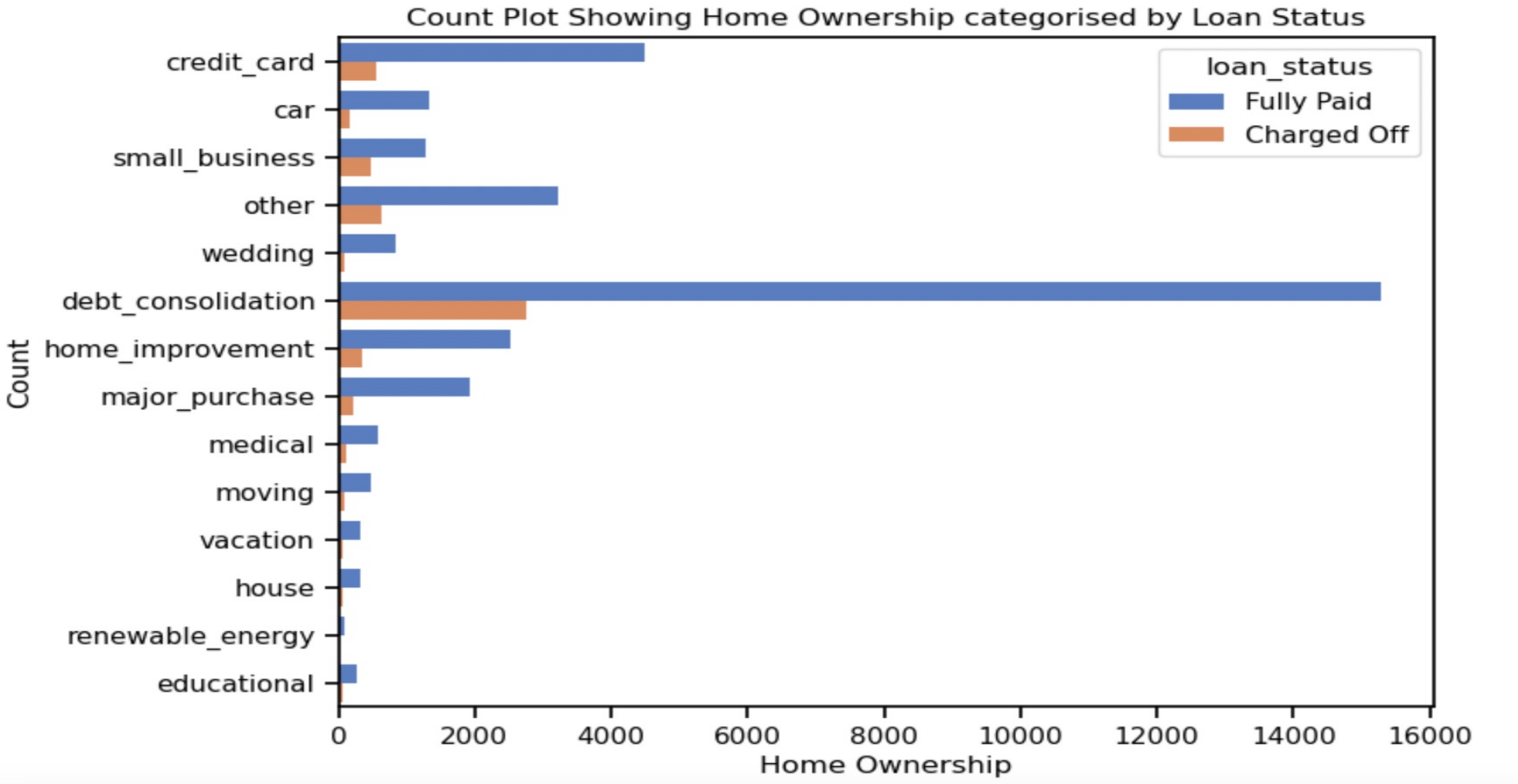


3. **Variable - Home Ownership status**

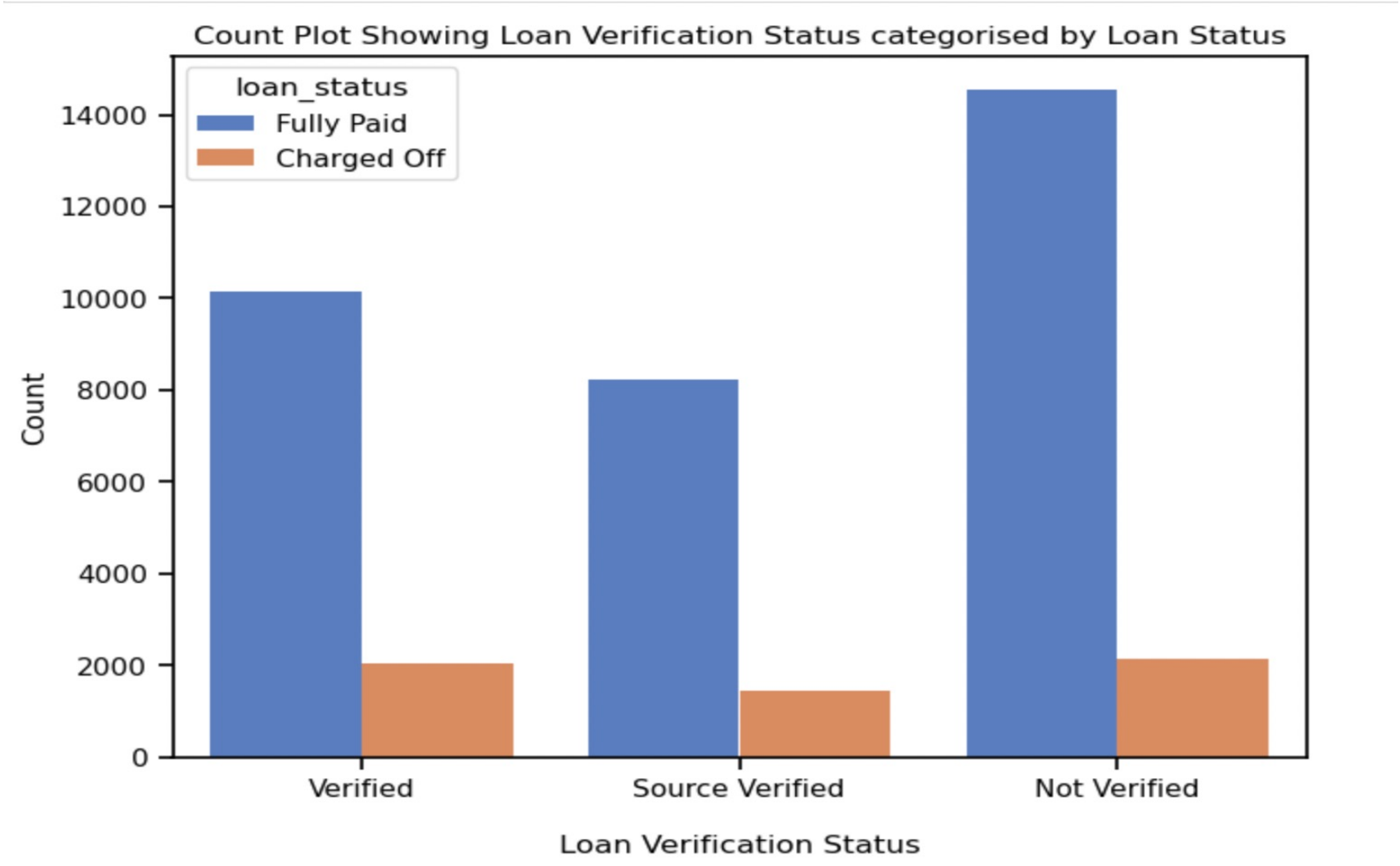




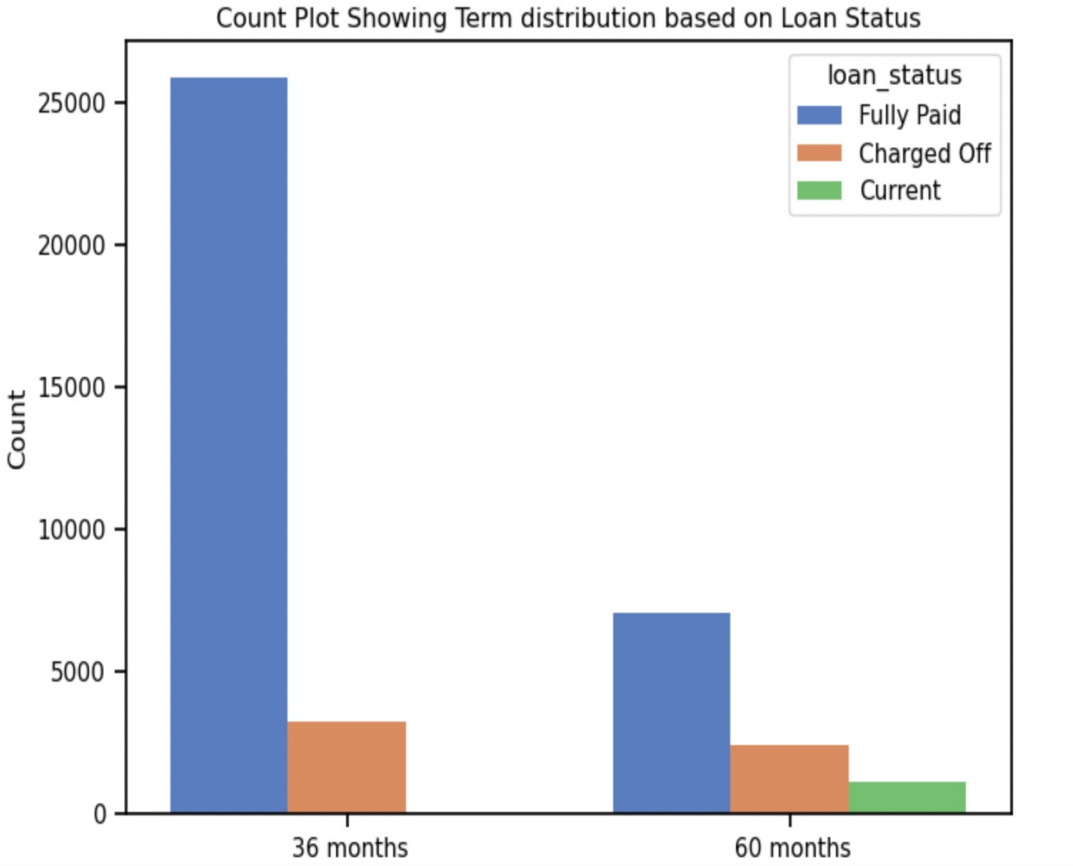
4. **Variable - Loan purpose**

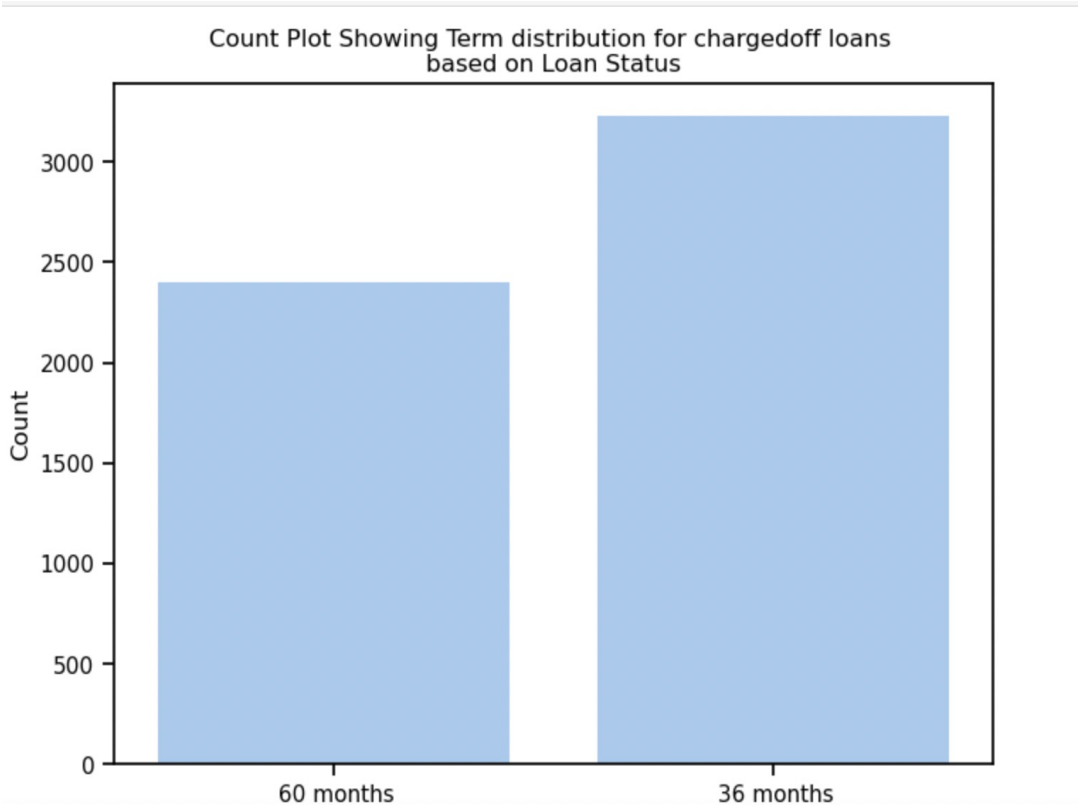


5. **Variable - Verification status**

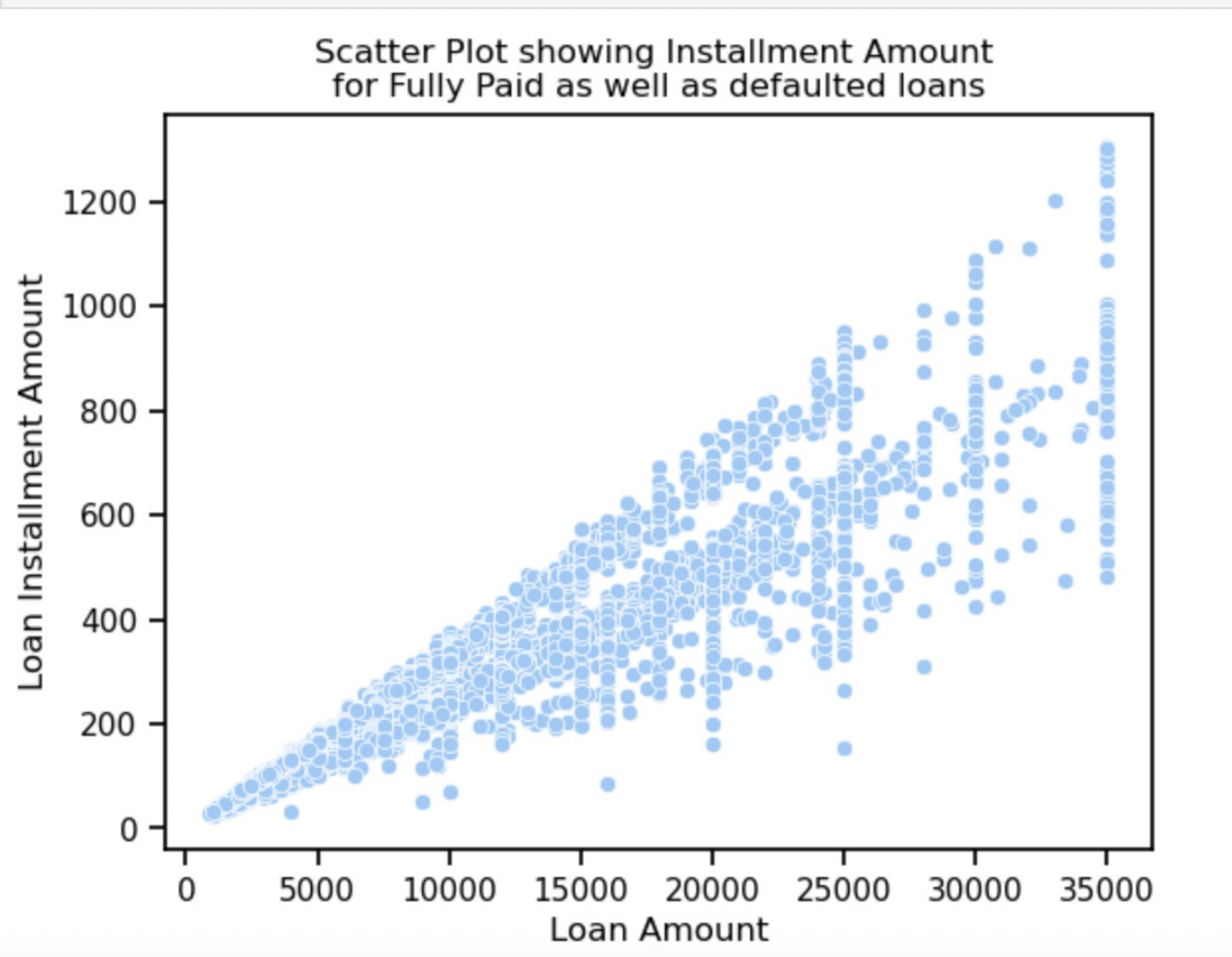


6. Variable - Term

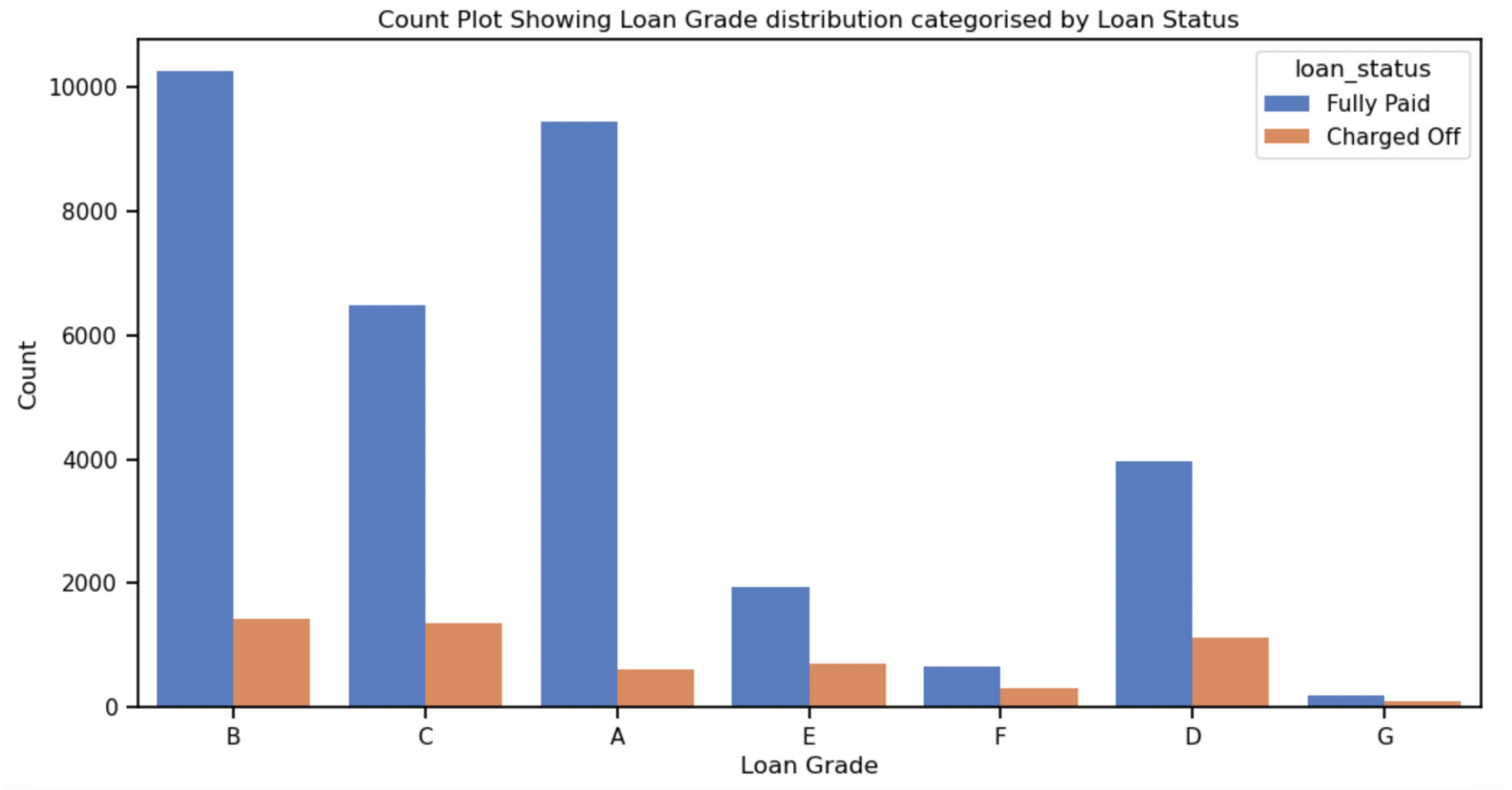




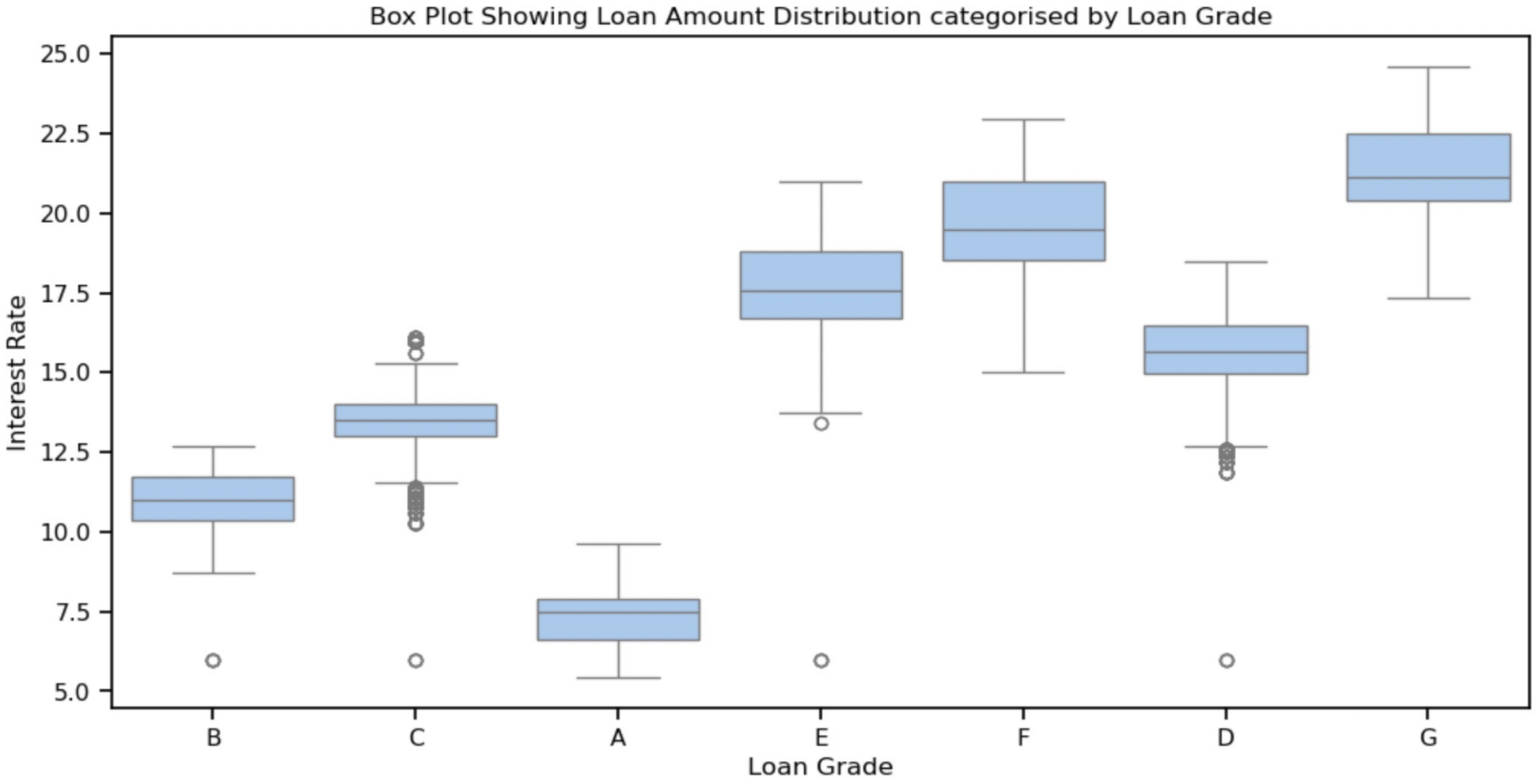
7. Variable - Installment



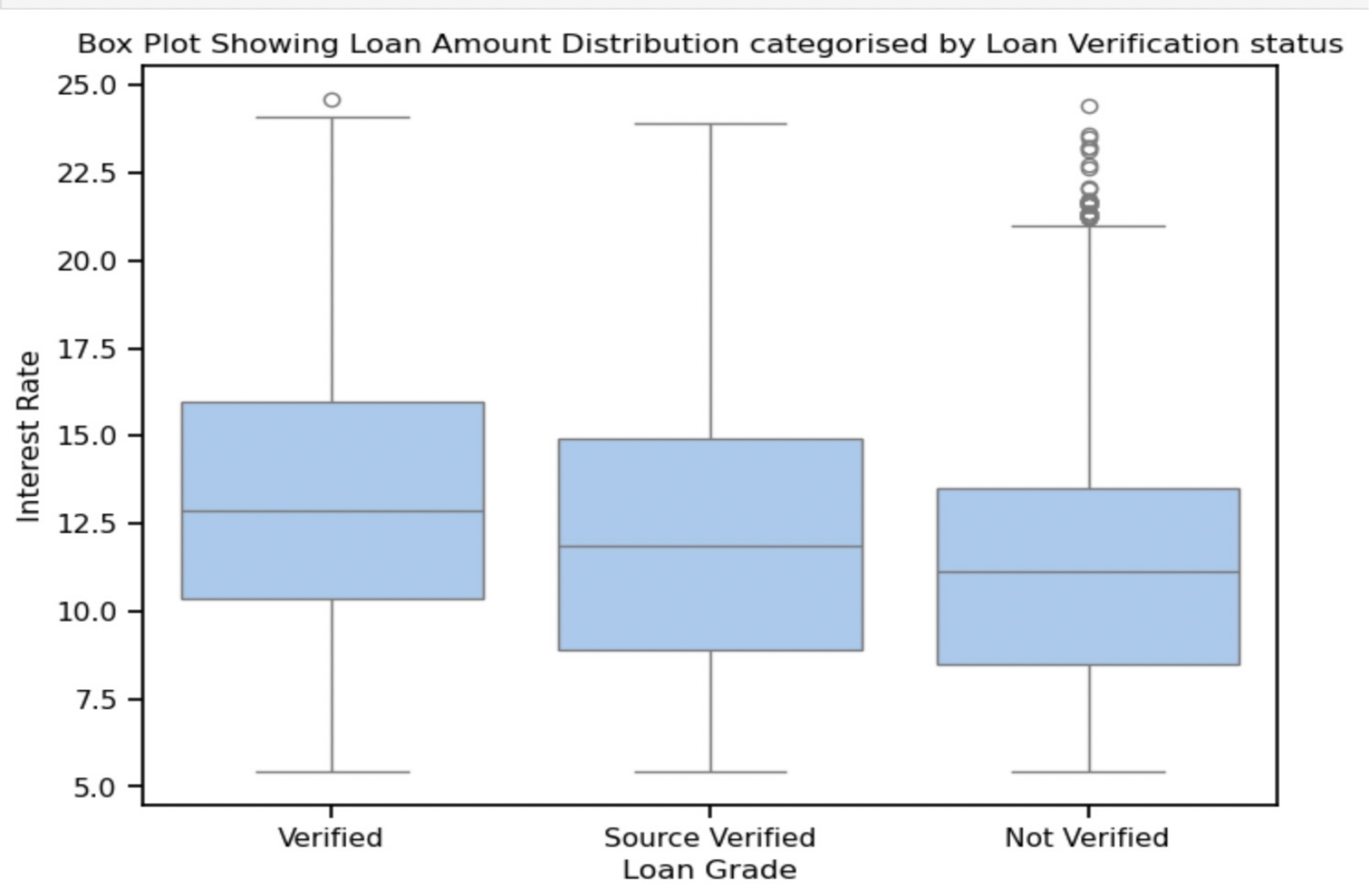
8. Variable - Installment



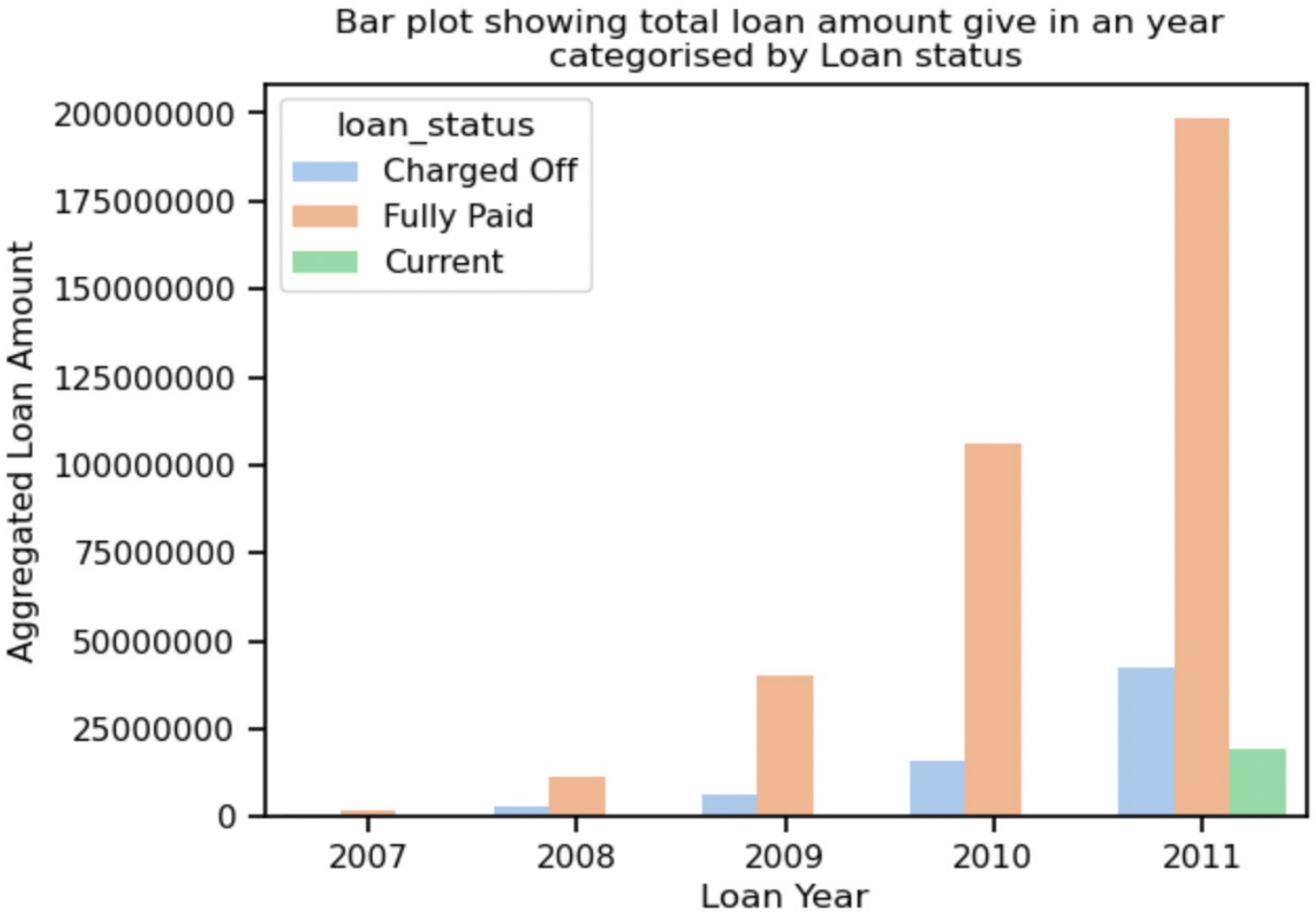
9. Variable - Interest rate



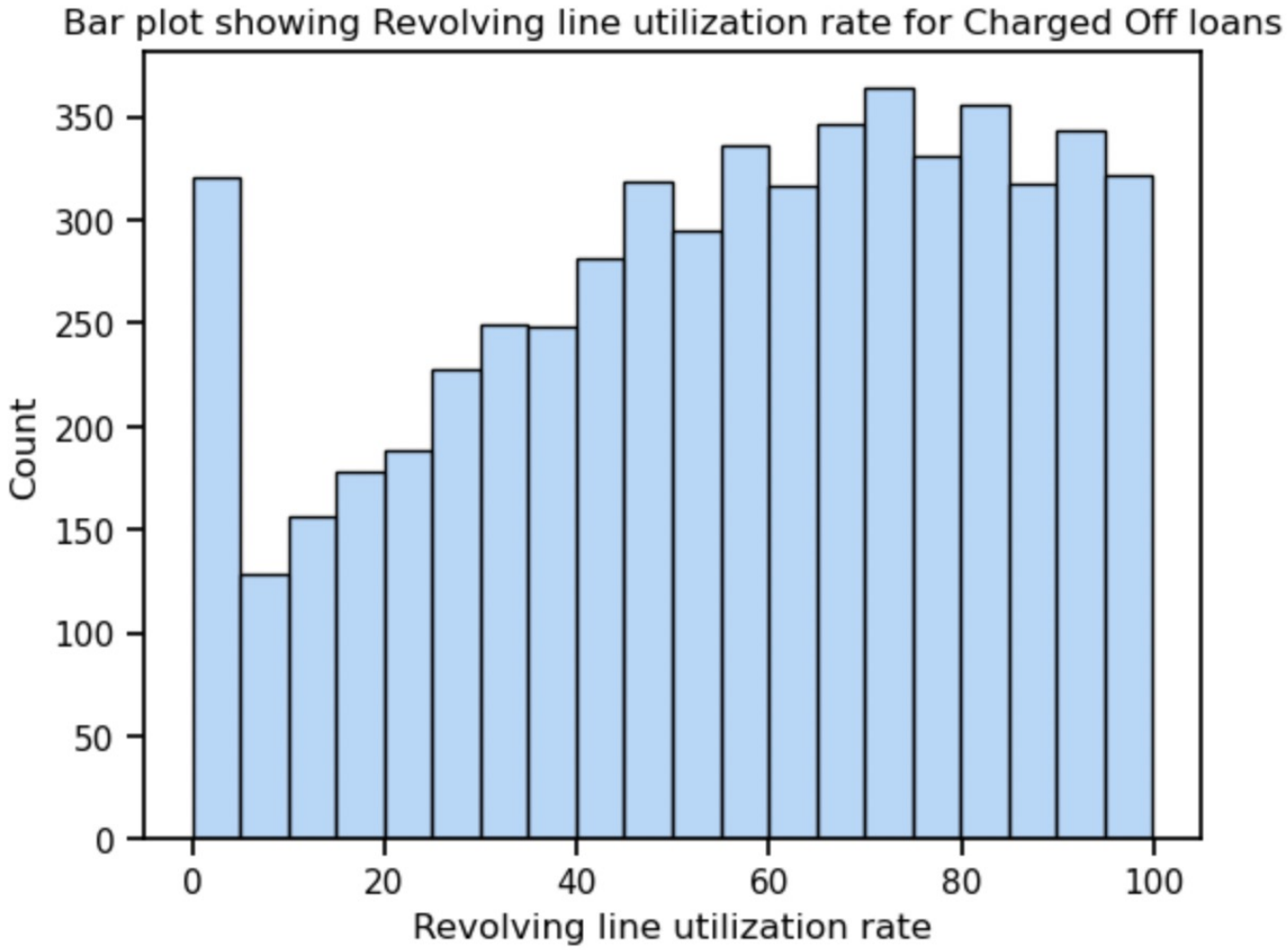
Variable - Interest rate (cont..)



10. Variable - Interest rate, Loan Year (Derived Metric)



10. Variable - Variable - Revolving line utilization rate



Observations -

1. It is observed that 14.2% of total loans get defaulted.
2. It is observed that for majority of loans the loan amount is between 5000 and 15000 both in case of fully paid and defaulted loans.
3. The loan amount for majority of defaulters are less then 15000.
4. Highest average loan amount are given against the loan graded as 'G’.
5. Borrower living in rented house or have mortgage their homes have defaulted more as compared to borrower living in own house.
6. The number of loans taken for the purpose of debt consolidation is highest and loan defaulters are also highest.
7. Loans with 'Verified' and 'Source Verified' status are most likely to default.
8. Higher the DTI, chances of default are more.
9. Higher the loan tenure the chances of default are high.
10. Loans graded as 'C' and 'D' have maximum chances of default.
11. Higher the loan grade, high is the interest rate.
12. Most number of defaults happened in year 2011.
13. Borrower having Revolving line utilization rate higher then 50 is most likely to default.