Lending Club case study

EXPLORATORY DATA ANALYSIS

Problem Objective

Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

identify these risky loan applicants using EDA is the aim of this case study.

Loan Dataset

Contains the complete loan data for all loans issued through the time period 2007 t0 2011.

Total records 39717 and 111 feature columns.

EDA approach

1. Data Cleaning

2. Data understanding

3. Univariate Analysis

4. Segmented univariate analysis

5. Recommendations

Data Cleaning

- *Excluding the "loan_status" category 'current' for analysis.*

- *Imputing the feature "emp_length" and dropping the record for "revol_util" feature.*

- *Data type conversion and standardizing the data.*

- *After dropping the null values and dropping the features which would not contribute for the loan defaulter prediction analysis, the total columns considered for analysis is 21.*

# Data Understanding

Table with the selected features for analysis

Feature dictionary with the feature description and feature type

0	annual_inc	The self-reported annual income provided by th...	N
1	dti	A ratio calculated using the borrower's total ...	N
2	earliest_cr_line	The month the borrower's earliest reported cre...	N
3	emp_length	Employment length in years. Possible values ar...	N
4	funded_amnt_inv	The total amount committed by investors for th...	N
5	grade	LC assigned loan grade	C
6	home_ownership	The home ownership status provided by the borr...	C
7	inq_last_6mths	The number of inquiries in past 6 months (excl...	N
8	installment	The monthly payment owed by the borrower if th...	N
9	int_rate	Interest Rate on the loan	N
10	issue_d	The month which the loan was funded	N
11	loan_amnt	The listed amount of the loan applied for by t...	N
12	loan_status	Current status of the loan	C
13	open_acc	The number of open credit lines in the borrowe...	N
14	pub_rec	Number of derogatory public records	N
15	purpose	A category provided by the borrower for the lo...	C
16	revol_util	Revolving line utilization rate, or the amount...	N
17	sub_grade	LC assigned loan subgrade	C
18	term	The number of payments on the loan. Values are...	N
19	total_acc	The total number of credit lines currently in ...	N
20	verification_status	Indicates if income was verified by LC, not ve...	C

Univariate Analysis

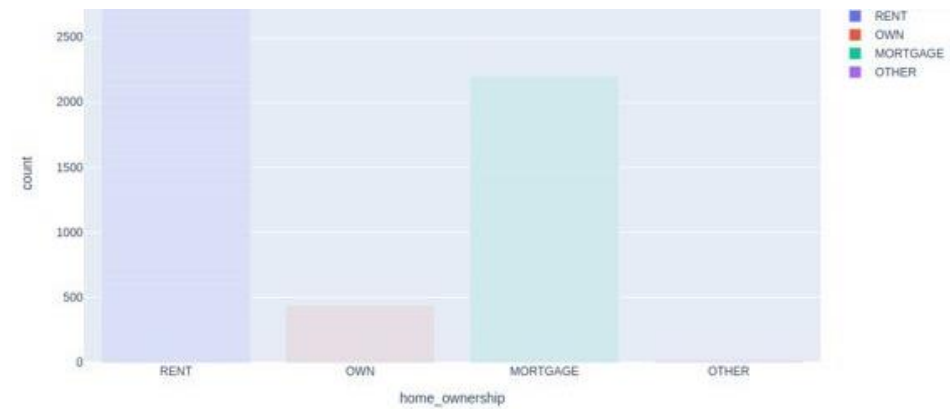**On analysing the Bar-plot for the following features,**

1. **'grade' and 'sub-grade'** , majority of customer loans charged-off falls on the assigned grade category 'B and 'C' ,Majority of the defaulters are in the assigned grade 'B' and Sub-grade '5'.

2. **'home_ownership'** , majority of defaulters are in the 'home_ownership' category 'Rent' or 'Mortgage'.

3. **'purpose'** , majority of defaulters are in the 'purpose' category 'debt_consolidation'.

4. **'int_rate_grp'** , majority of defaulters are in the 'int_rate_grp' category '11%-15%'.

5. **'open_acc_grp'** , majority of defaulters are in the 'open_acc_grp' category '2-10'.

6. **'revol_util_grp'** , majority of defaulters are in the 'revol_util_grp' category '80-100'.

7. **'total_acc_grp'** , majority of defaulters are in the 'total_acc_grp' category '2-20'.

8. **'term'** , majority of defaulters are in the 'term' category '36 months'.
   'verification_status' , majority of defaulters are in the 'verification_status' category 'Not Verified'.

9. **'inq_last_6mths'** , majority of defaulters are in the 'inq_last_6mths' category '1 and 2'.

10. **'pub_rec'** , majority of defaulters are in the 'pub_rec' category '0'.

11. **'funded_amnt_inv_grp',** majority of defaulters are in the >'funded_amnt_inv_grp' category '0K-10K'.

12. **'loan_amnt_grp'** , majority of defaulters are in the 'loan_amnt_grp' category '5K-10K'.

13. **'dti_grp'** , majority of defaulters are in the 'dti_grp' category '12-18'.

14. **'installment_grp'** , majority of defaulters are in the 'installment_grp' category '145-274'.

15. **'issue_year'** , majority of defaulters are in the 'issue_year' category '2011'.
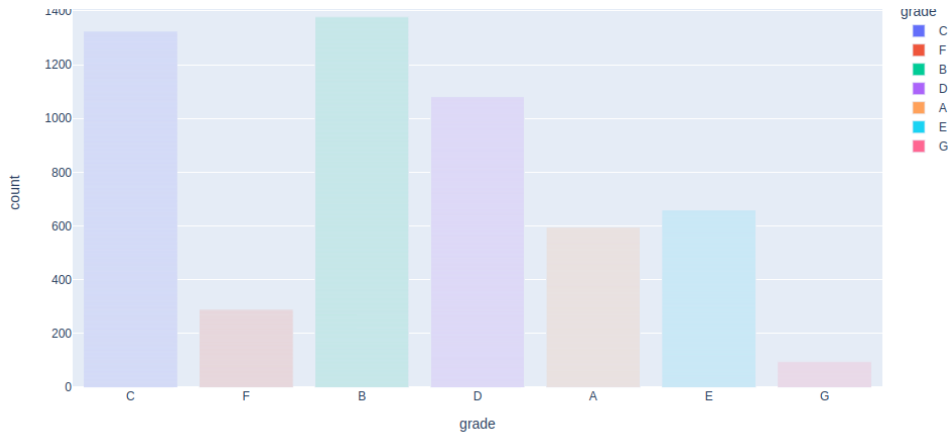
Univariate Analysis Visualization

Feature 'grade' and 'sub-grade' analysis

- On analyzing the Bar-plot for the feature 'grade' and 'sub-grade', we could visualize that the majority of customer loans charged-off loans falls on the assigned grade category 'B and 'C'
- Majority of the defaulters are in the assigned grade 'B' and Sub-grade '5'.

Feature 'home_ownership' analysis

- On analyzing the Bar-plot for the feature 'home_ownership', we could visualize that the majority of defaulters are in the 'home_ownership' category 'Rent' or 'Mortgage'.
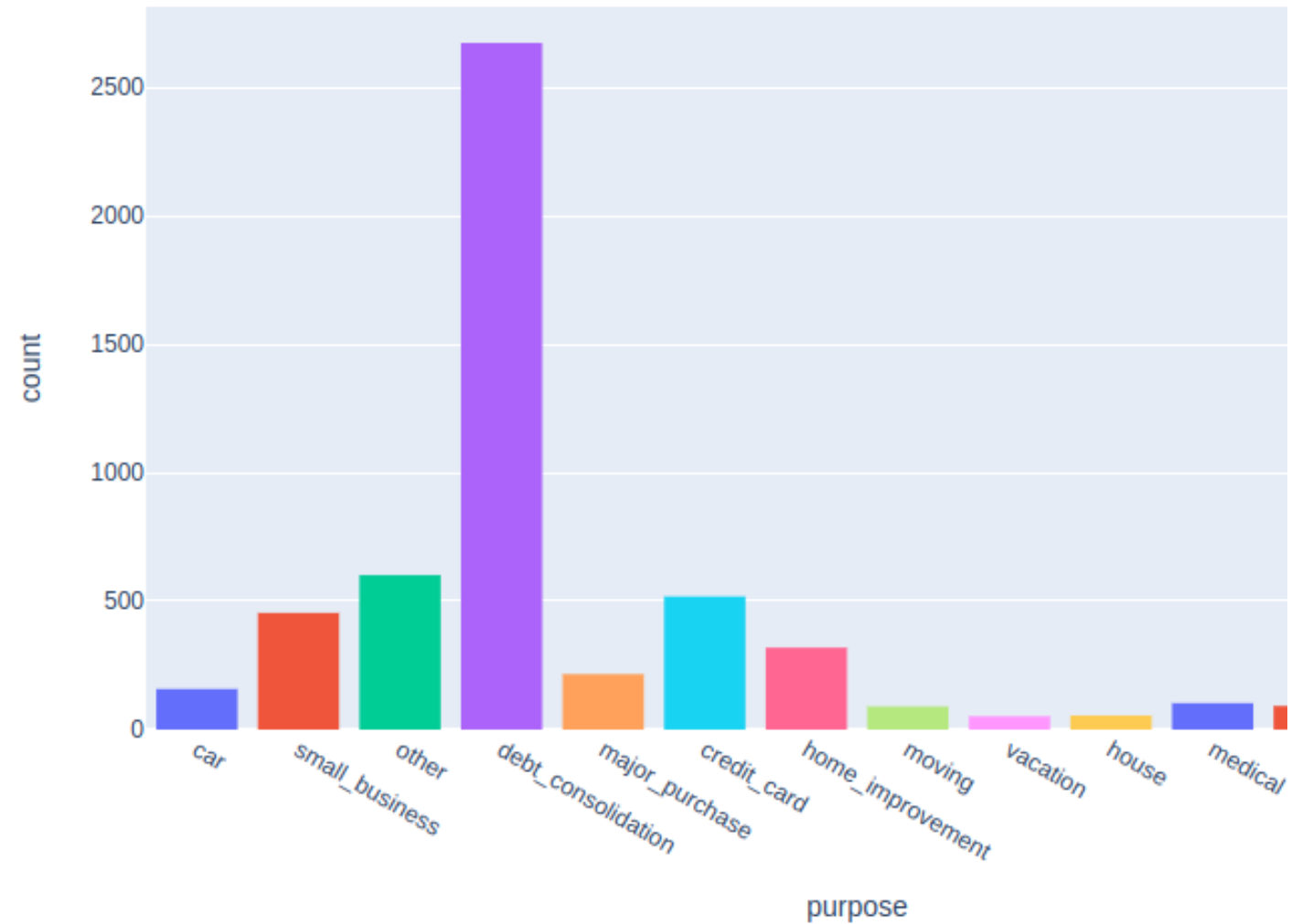
Univariate Analysis Visualization

Feature 'purpose' analysis

On analyzing the Bar-plot for the feature 'purpose', we could visualize that the majority of defaulters are in the 'purpose' category 'debt_consolidation'.

Segmented univariate analysis

Obtaining derived features by binning few numerical features ('int_rate', 'annual_inc', 'funded_amnt_inv') to convert them into categorical.

Segmented Univariate Analysis Visualization

Feature 'int_rate_grp' analysis

On analyzing the barplot for the feature 'int_rate_grp', we could visualize that the majority of defaulters are in the 'int_rate_grp' category '11%-15%'.
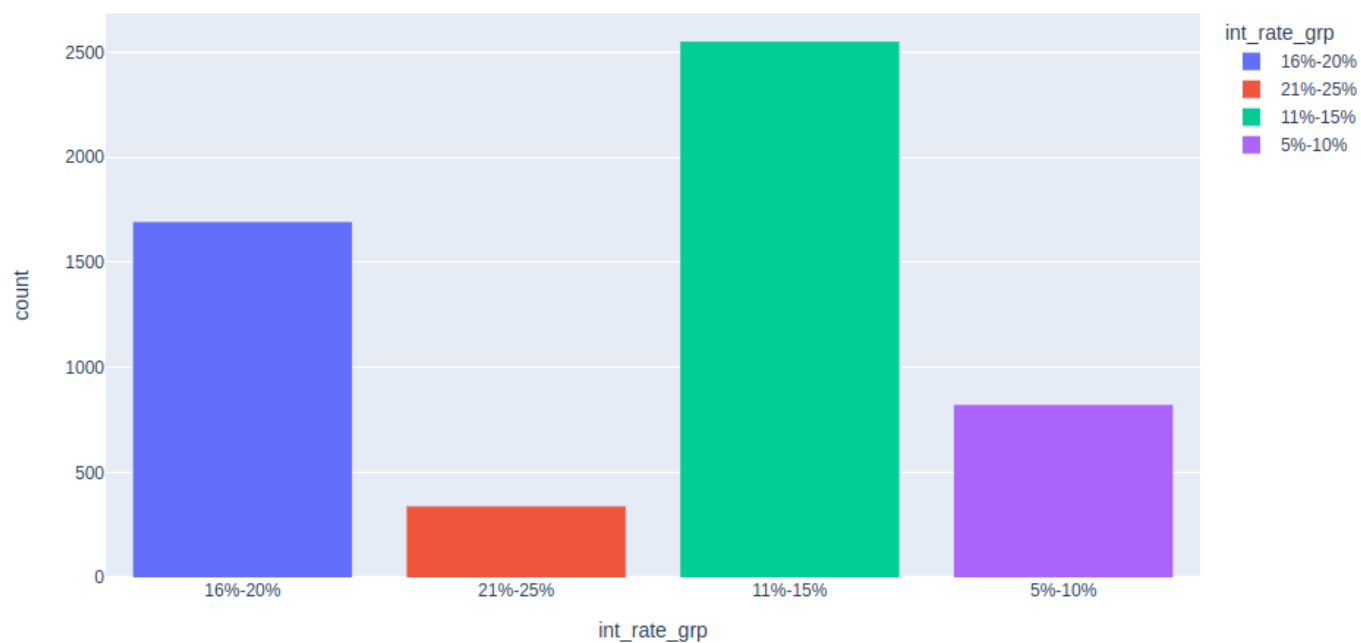
Bivariate Analysis
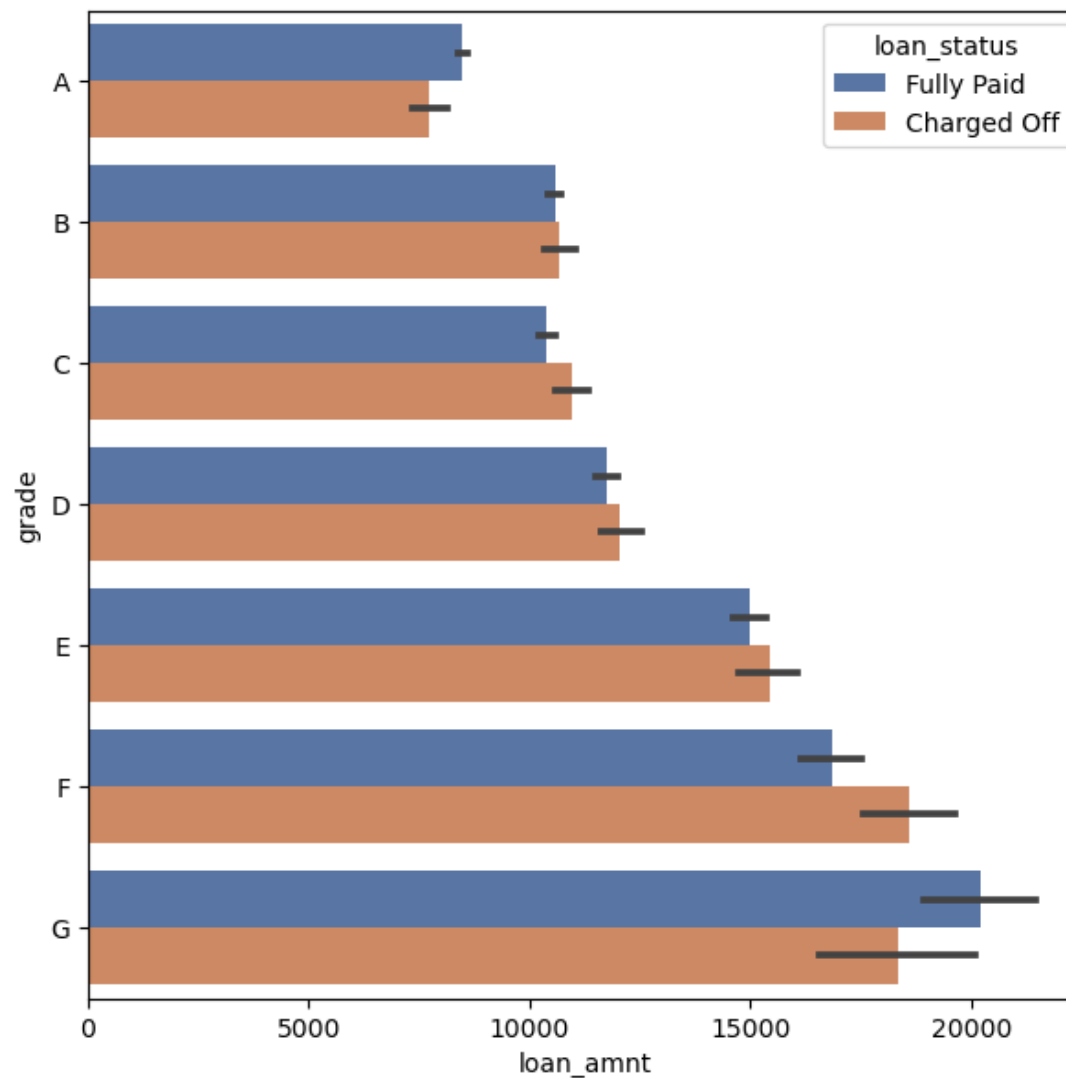
On analysing the Bar-plot for the following features,

1. **'annual_inc_grp'and 'loan_amnt' ,** we could visualize that across all the income category defaulters are higher.

2. **'annual_inc' and 'home_ownership' ,** we could visualize that for applicants whose 'home_ownership' category is 'Nortgage' has annual_inc range '60k-70k' .

3. **'annual_inc' and 'int_rate_grp'** , we could visualize that the majority of defaulters are in the 'annual_inc' range 60k-70k and opted for the interest rate category 21%-25%.

4. **'loan_amt_grp' and 'int_rate'** , we could visualize that the majority of defaulters are in the 'loan_amt_grp'category 25k-35k and opted for the interest rate range over 16%.

5. **'loan_amt' and 'purpose',** we could visualize that the majority of defaulters are in the 'loan_amt' range 12k-14k and opted for the purpose of 'debt_consolidation' ,'small business' and 'home_improvement'.

6. **'loan_amt' and 'home_ownership',** we could visualize that the majority of defaulters are in the 'loan_amt' range 12k-14k and under the 'home_ownership' category 'Mortgage'.

7. **"loan_amt' and 'grade',** we could visualize that the majority of defaulters are in the 'loan_amt' range 15k-20k and under the assigned 'grade' category 'E, F, G'.

8. **"loan_amt' with 'emp_length' and 'verification_status',** we could visualize that the loan_amount is higher for heigher emp_length category and 'Verfication_status' verified.

9. **"int_rate' and 'grade',** we could visualize that the majority of defaulters are in the 'int_rate' range 15%-20% and under the assigned 'grade' category 'E, F, G'.

10. **'term' and 'loan_amt' ,** we could visualize that the majority of defaulters are in the 'term' range 60 months and under the 'loan_amnt' range 14k-16k.

Bivariate Analysis Vizualization

Feature 'loan_amt' and 'grade' bivariate analysis

On analyzing the barplot for the features 'loan_amt' and 'grade', we could visualize that the majority of defaulters are in the 'loan_amt' range 15k-20k and under the assigned 'grade' category 'E, F, G'.

Bivariate Analysis Visualization

Feature 'loan_amt' and 'home_ownership' bivariate analysis

On analyzing the barplot for the features 'loan_amt' and 'home_ownership', we could visualize that the majority of defaulters are in the 'loan_amt' range 12k-14k and under the 'home_ownership' category 'Mortgage'.
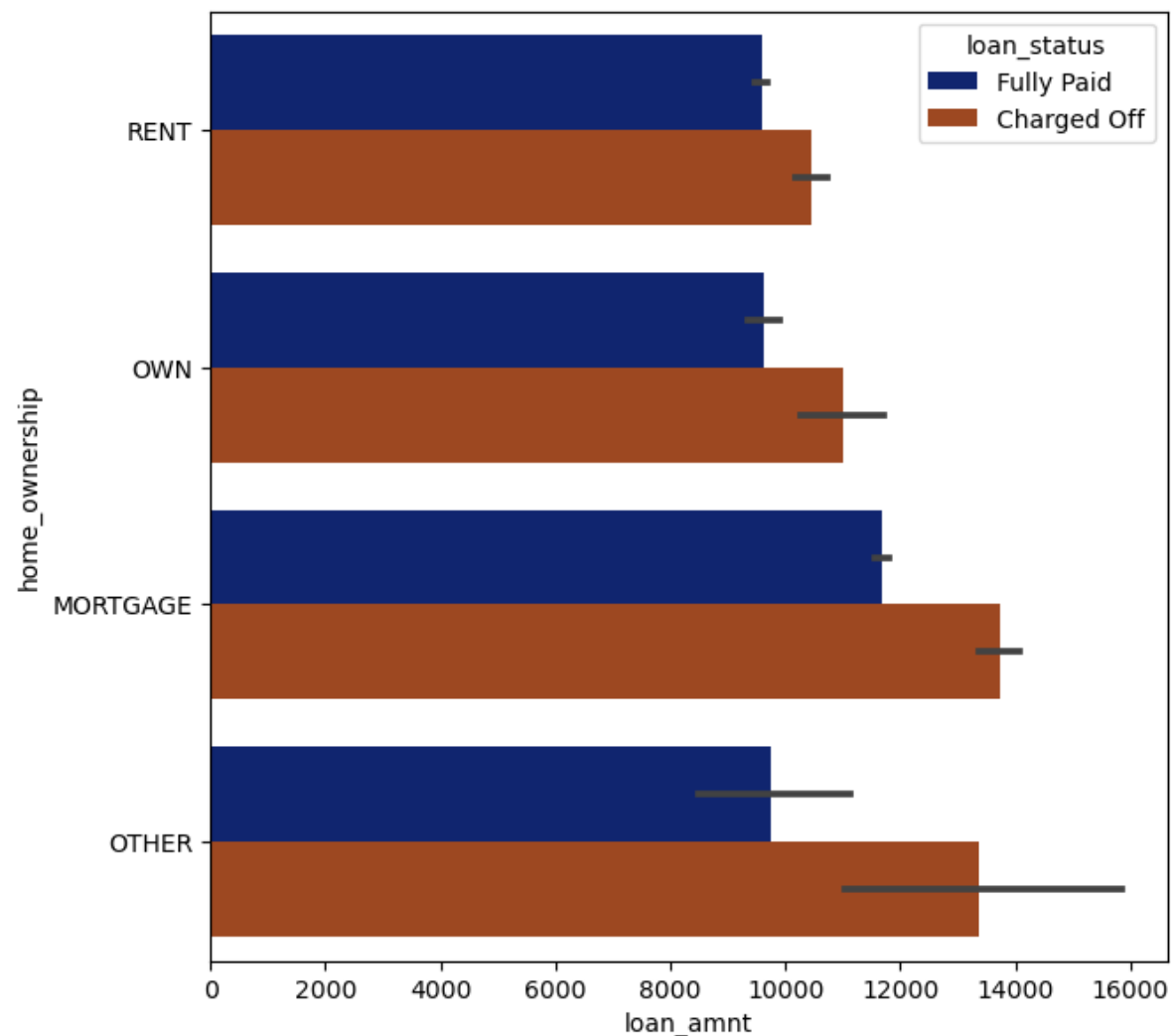
Recommendations

- Reduce the number of loans issued with high interest rate and term 60 months.

- For the customer grade category B and C, multiple levels of verification process to be performed before loan approval.

- For the Customers with home ownership category 'Mortgage' and 'Rent' , multiple levels of credit verification to be performed before loan approval and approval of loan amount higher than 12k to be restricted.

- Loan approvals for the purpose of 'debit consolidation' and 'small business' should be restricted.