LENDING CLUB CASE STUDY

Prepared By
Sai Krishna Rali & Manjiri Gajmal

PROBLEM STATEMENT

Problem Statement

Business Introduction

A consumer finance company that caters to urban customers must evaluate loan applications to decide whether to approve them or not. This decision involves assessing two types of risks:

- When an applicant is deemed likely to repay the loan, rejecting the application means the company loses potential business.
- Conversely, if the applicant is likely to default and not repay the loan, approving it could result in financial losses for the company.

Business Objective

- The primary objective of this case study is to pinpoint high-risk loan applicants to minimize the number of such risky loans and consequently reduce credit losses.
- To achieve this, we must identify the key factors or variables that strongly correlate with loan defaults, serving as reliable indicators of potential default.

Data Provided

- We have been given a loan dataset that includes comprehensive information about all loans disbursed between 2007 and 2011.
- Additionally, a data dictionary has been provided, offering descriptions and explanations for the attributes or variables present in the dataset.

Note:

The company declined the loan applications of certain candidates due to their failure to meet the companies criteria or other reasons. As a result, these rejected applicants lack any transactional history with the company, and therefore, such data is unavailable in this dataset.

APPROACH

Our Approach

- We have taken the loan dataset and imported it into a pandas DataFrame, where we performed data cleansing tasks such as handling missing values, outliers, standardization, and removing redundant information.
- Subsequently, we proceeded to systematically analyze the data to determine the key factors contributing to loan defaults.
  - Our analytical process involved various stages, including univariate analysis, segmented univariate analysis, bivariate analysis, and ultimately, an examination of derived metrics.
- Throughout each of these stages, we pinpointed the variables that exhibited correlations with loan defaults, and we thoroughly documented all our findings.
- Finally. We summarized the key driving factors and the key combinations

DATA LOADING

Libraries & Display options

Imported the required libraries and configured some display options

In [1]:
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

Display Settings

In [2]:
```
pd.set_option('display.max_rows', 150)
pd.set_option('display.max_columns', 150)

pd.options.display.float_format = '{:,.2f}'.format
```

Data Loading

Loaded the given loan data into pandas data frame for analysis

```
In [3]: 1 df = pd.read_csv('loan.csv')

In [4]: 1 df.head(2)
```

Out[4]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	a
0	1077501	1296599	5000	5000	4,975.00	36 months	10.65%	162.87	B	B2	NaN	10+ years	RENT	
1	1077430	1314167	2500	2500	2,500.00	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	

Dimensions & Duplicate data

```
In [5]:  1 df.shape
Out[5]: (39717, 111)
```

```
In [8]:  1 # Checking if there are any duplicate values
         2
         3 df[df.duplicated()].shape
Out[8]: (0, 111)
```

- The loan dataset contains 39,717 rows and 111 variables, and there are no duplicated rows in the dataset

Description of Data

In [7]:
```
# Understanding the statistical metrics of numerical data columns

df.describe()
```

Out[7]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	mths_since_last_d
count	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	39,717.00	14,0
mean	683,131.91	850,463.56	11,219.44	10,947.71	10,397.45	324.56	68,968.93	13.32	0.15	0.87	
std	210,694.13	265,678.31	7,456.67	7,187.24	7,128.45	208.87	63,793.77	6.68	0.49	1.07	
min	54,734.00	70,699.00	500.00	500.00	0.00	15.69	4,000.00	0.00	0.00	0.00	
25%	516,221.00	666,780.00	5,500.00	5,400.00	5,000.00	167.02	40,404.00	8.17	0.00	0.00	
50%	665,665.00	850,812.00	10,000.00	9,600.00	8,975.00	280.22	59,000.00	13.40	0.00	1.00	
75%	837,755.00	1,047,339.00	15,000.00	15,000.00	14,400.00	430.78	82,300.00	18.60	0.00	1.00	
max	1,077,501.00	1,314,167.00	35,000.00	35,000.00	35,000.00	1,305.19	6,000,000.00	29.99	11.00	8.00	1

- After reviewing the data description, we observed that certain columns are entirely empty
- In the next step, we examined the presence of missing values in each column and, based on our findings, decide whether to remove the column or fill in the missing values.

DATA CLEANING

Missing Value Columns Deletion

- During this step, we have identified and removed columns with a missing value percentage exceeding 40%
- There are a total of 57 columns in the data frame with a missing value percentage exceeding 40%.
- After eliminating these columns, we now have a new data frame containing 54 remaining columns.

In [11]:
```
# First, we will remove any columns that contain over 40% missing values.
# This is because handling such a significant amount of missing data is not feasible for imputation, and eliminating
# too many rows is also impractical, as it would result in a substantial loss of valuable data and potentially biased insigh

df_null_m40 = list(df_null_per[df_null_per > 40].index)
df_null_m40
```

In [12]:
```
# Creating a new dataframe and loading those columns that have less than 40% of missing values

df1 = df.drop(df_null_m40, axis=1)
```

In [13]:
```
# Shape of new dataframe is

print('The shape of new dataframe is', df1.shape)
```

```
The shape of new dataframe is (39717, 54)
```

Single Unique Value Columns Identification

- Upon reviewing the metrics once more, we have observed that certain columns have identical minimum and maximum values, essentially containing only a single unique value.
- In the data frame, there are a total of nine columns that consist of only a single unique value

In [18]:
```
# identifying those numerical columns that have a zero standard deviation

df1_std = df1[num_col].std(axis=0)
single_val_num_col = list(df1_std[(df1_std==0)].index)
```

In [19]:
```
single_val_num_col
```

Out[19]:
```
['collections_12_mths_ex_med',
 'policy_code',
 'acc_now_delinq',
 'chargeoff_within_12_mths',
 'delinq_amnt',
 'tax_liens']
```

In [22]:
```
# identifying those categorical columns that have a single value

df1_single = df1[cat_col].nunique()
single_val_cat_col = list(df1_single[(df1_single==1)].index)
```

In [23]:
```
single_val_cat_col
```

Out[23]:
```
['pymnt_plan', 'initial_list_status', 'application_type']
```

Single Unique Value Columns Deletion

- We have removed all columns with a single unique value as they do not provide any meaningful information.
- Consequently, we now have 45 columns remaining

In [20]:
```
# As these numerical columns contain only one value, these columns will not give us any insights.
#Hence, dropping these columns

df1.drop(single_val_num_col, axis=1, inplace=True)
```

In [24]:
```
# As these categorical columns contain only one value, these columns will not give us any insights.
# Hence, dropping these columns

df1.drop(single_val_cat_col, axis=1, inplace=True)
```

In [25]:
```
df1.shape
```

Out[25]: (39717, 45)

Identification of Unnecessary Columns

Through research and analysis, we have determined that the following columns are not essential for our analysis, as they do not significantly influence loan default

In [28]:
```
xtra_columns = ['id', 'member_id','funded_amnt', 'funded_amnt_inv','emp_title','url', 'desc','title', 'zip_code',
       'total_acc','out_prncp', 'out_prncp_inv', 'total_pymnt',
       'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int','last_pymnt_d','revol_bal','recoveries',
       'total_rec_late_fee','collection_recovery_fee','last_pymnt_amnt','last_credit_pull_d']
```

- id and member_id serve solely for identification purposes and are not influential in determining Loan Default.
- loan_amnt and funded_amnt exhibit high correlation, leading us to drop funded_amnt.
- Investor-related details are not relevant for this analysis, so well remove columns associated with investors.
- url and desc provide extra but unessential information for the analysis.
- emp_title lacks strong predictive power for Loan Default since it contains 28,820 unique values.
- Although title and purpose contain similar information, purpose is more structured, prompting us to drop the title column.
- open_acc and total_acc convey similar information, but we will keep open_acc for its relevance to the current situation.
- The remaining columns on the list do not have a significant impact on Loan Default and contain post-loan approval information.

Unnecessary Columns Deletion

- We have removed the unnecessary columns from the data frame, leaving us with a total of 22 columns for our analysis

```
In [29]:
# Dropping extra columns from the dataframe as these will not any value add

df1.drop(xtra_columns, axis=1, inplace=True)
```

```
In [30]:
print('The shape of the dataframe after dropping extra columns', df1.shape)
```

```
The shape of the dataframe after dropping extra columns (39717, 22)
```

Now that weve removed the columns with excessive missing values, single unique value, and those that are unnecessary, we will proceed with the next step, which is imputing missing values

Identification of Columns with Missing Values

2.2.2 Missing Values Treatment

In [31]:
```
# Missing values percentage in each column

df1_null_per = df1.isnull().mean() * 100

```

In [32]:
```
# columns with Null value percentage more than zero (mtz) in the new dataframe

df1_null_mtz = list(df1_null_per[df1_null_per > 0].index)
df1_null_mtz
```

Out[32]:
```
['emp_length', 'revol_util', 'pub_rec_bankruptcies']
```

- There are just three columns with some percentage of missing values.
- We will analyze each of them individually and make the required data adjustments

Imputing Missing values

- We have filled the missing values in emp_length and revol_util with the mode, considering them as categorical columns.
- For pub_rec_bankruptcies, which is a numerical variable, we have imputed the missing values with the median

```
In [35]:
# Imputing the missing values with the mode

df1.emp_length = df1.emp_length.fillna(df1.emp_length.mode()[0])
```

```
In [39]:
# Imputing the missing values with the mode

df1.revol_util = df1.revol_util.fillna(df1.revol_util.mode()[0])
```

```
In [43]:
# Imputing the missing values with the mode

df1.pub_rec_bankruptcies = df1.pub_rec_bankruptcies.fillna(df1.pub_rec_bankruptcies.median())
```

- All missing values in all columns have been addressed, and we no longer have any columns with missing values

Standardising the data

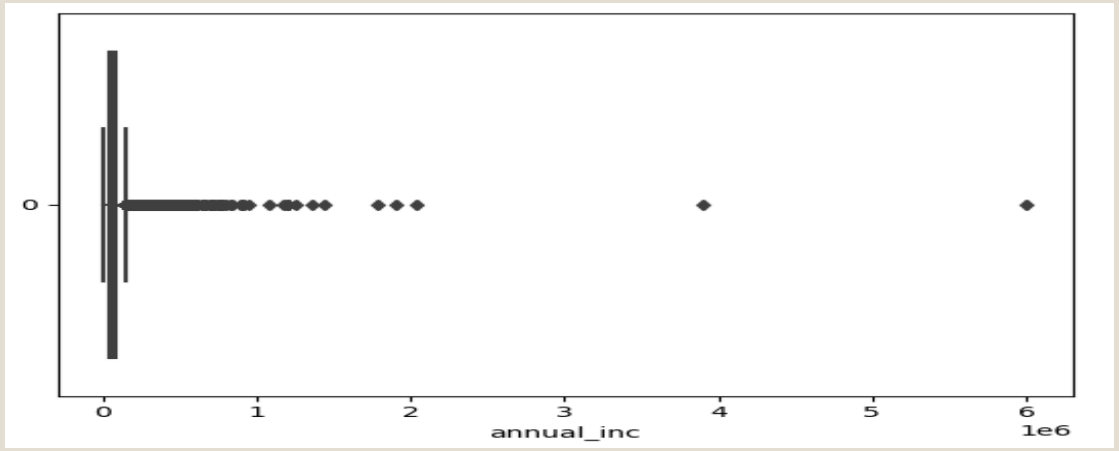
- We subsequently moved on to the next step, which involves standardizing the values in certain columns by removing the percentage symbol. This was done for the int_rate and revol_util columns
- We employed lambda and replace functions to remove the % symbol

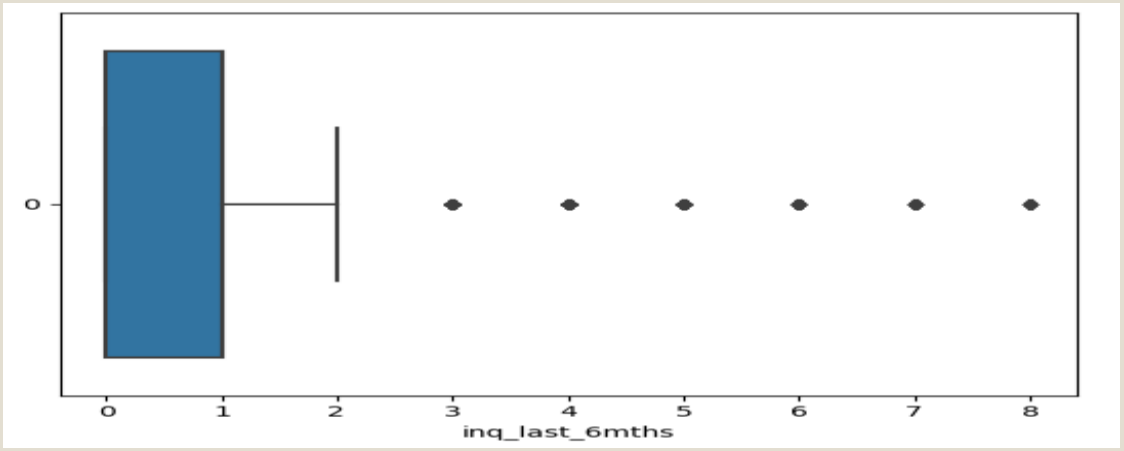
```
In [47]:
1 rem_per = lambda x : float(x.replace('%',''))

In [48]:
1 df1.int_rate = df1.int_rate.apply(rem_per)
2 df1.revol_util = df1.revol_util.apply(rem_per)
```

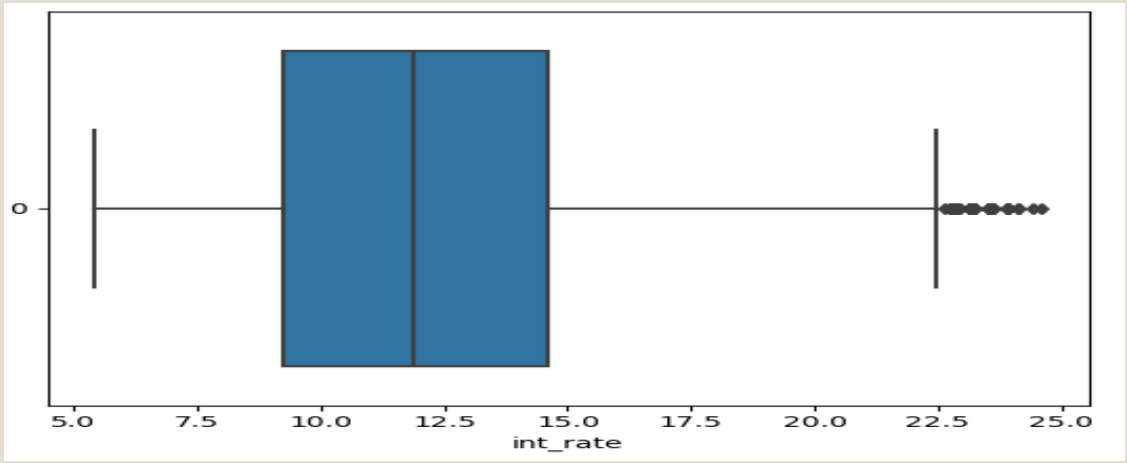
- We also observed that the data types of all the columns match the type of values they contain

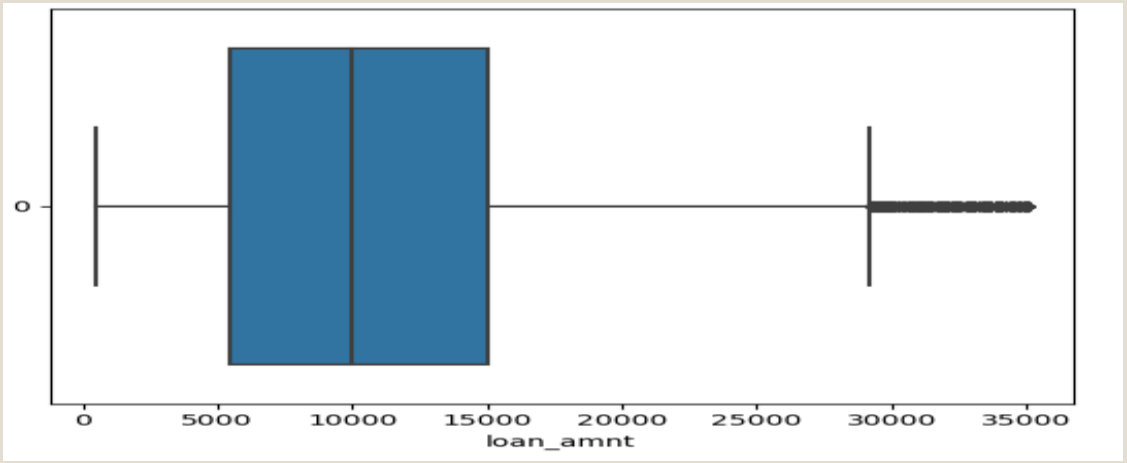
Outliers' treatment

- We have identified outliers in the Annual income column and set a threshold (99th percentile of annual_inc) to remove them.
- However, for the remaining columns, while there are values outside the plot, they exhibit more of a continuous nature, so we have opted not to delete those outliers









DATA ANALYSIS

Data Analysis – Preparatory Step

- We excluded the data where the loan status is 'Current' because our goal is to identify the key variables associated with loan default.
- Subsequently, we divided the dataset into two separate data frames based on the loan status, which will facilitate our analysis.

In [146]:
```
# As we need to find out driving factor for default, the individuals who are currently paying installments are of no use.
# Hence dropping those rows from the datafram

df1 = df1[df1.loan_status != 'Current']
```

In [147]:
```
df1.shape
```

Out[147]: (38193, 22)

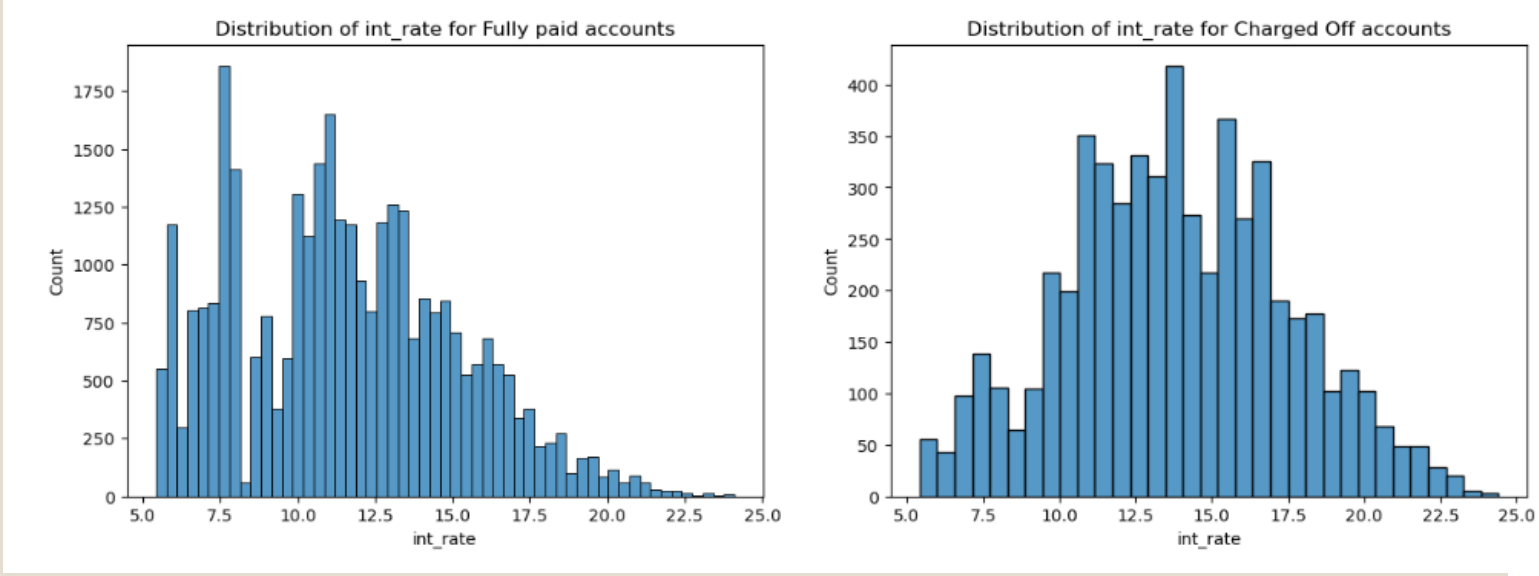
In [148]:
```
# we are dividing the dataframe into two separate dataframes based on loan status

fully_paid = df1[df1.loan_status == 'Fully Paid']
charged_off = df1[df1.loan_status == 'Charged Off']
```

UNIVARIATE ANALYSIS

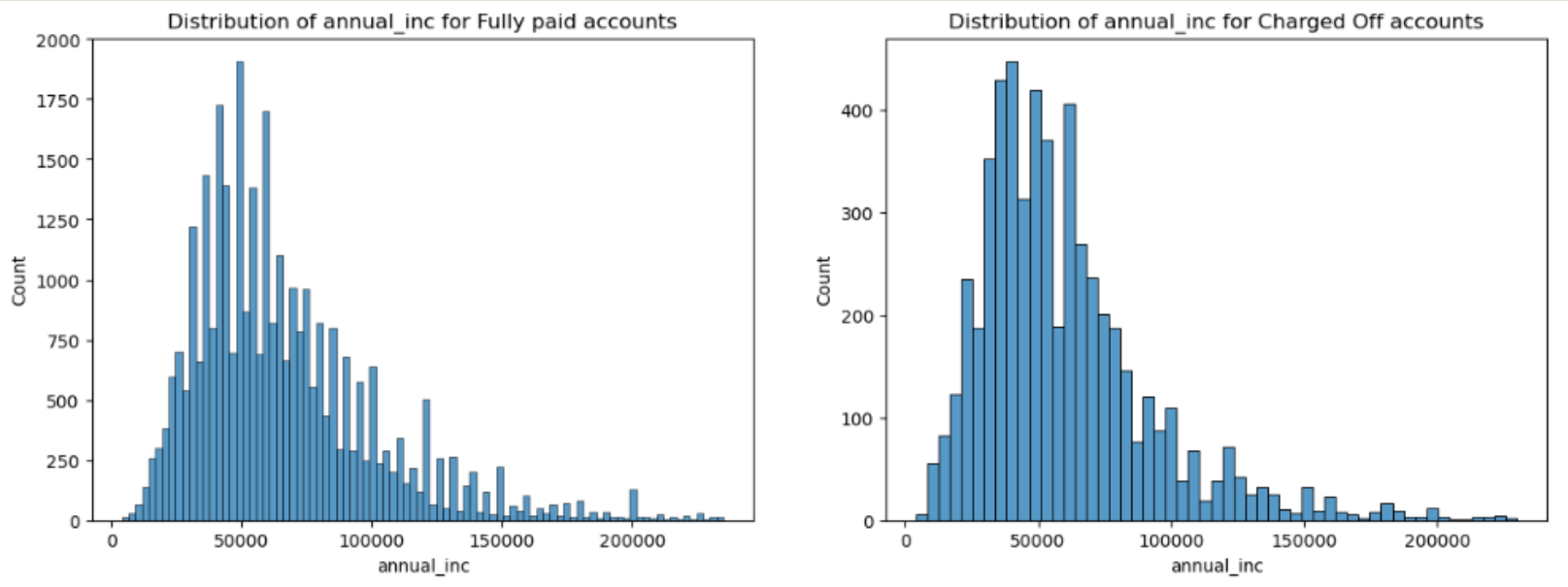
Distribution of Numerical Variables

- Borrowers with loan amounts ranging from 5,000 to 10,000 are at a higher risk of defaulting.

- Additionally, the interest rate distribution is relatively uniform among borrowers who defaulted, whereas borrowers who fully paid their loans tended to have lower interest rates

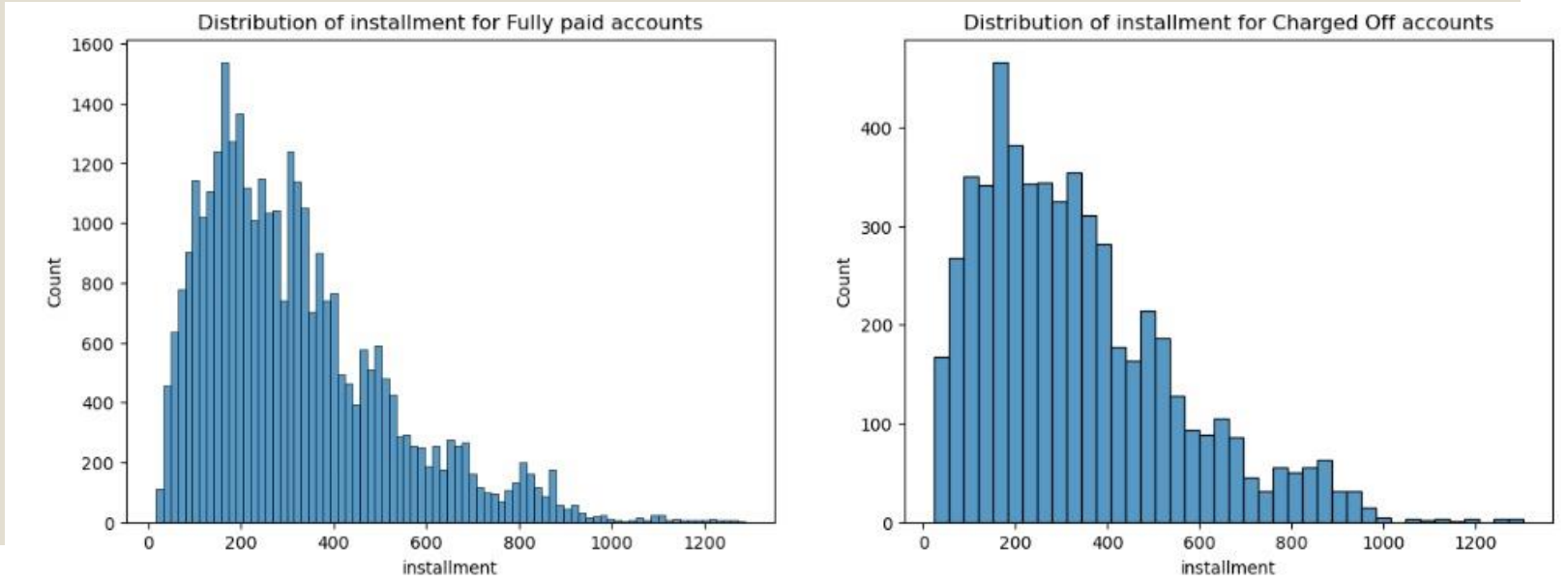


Distribution of Numerical Variables

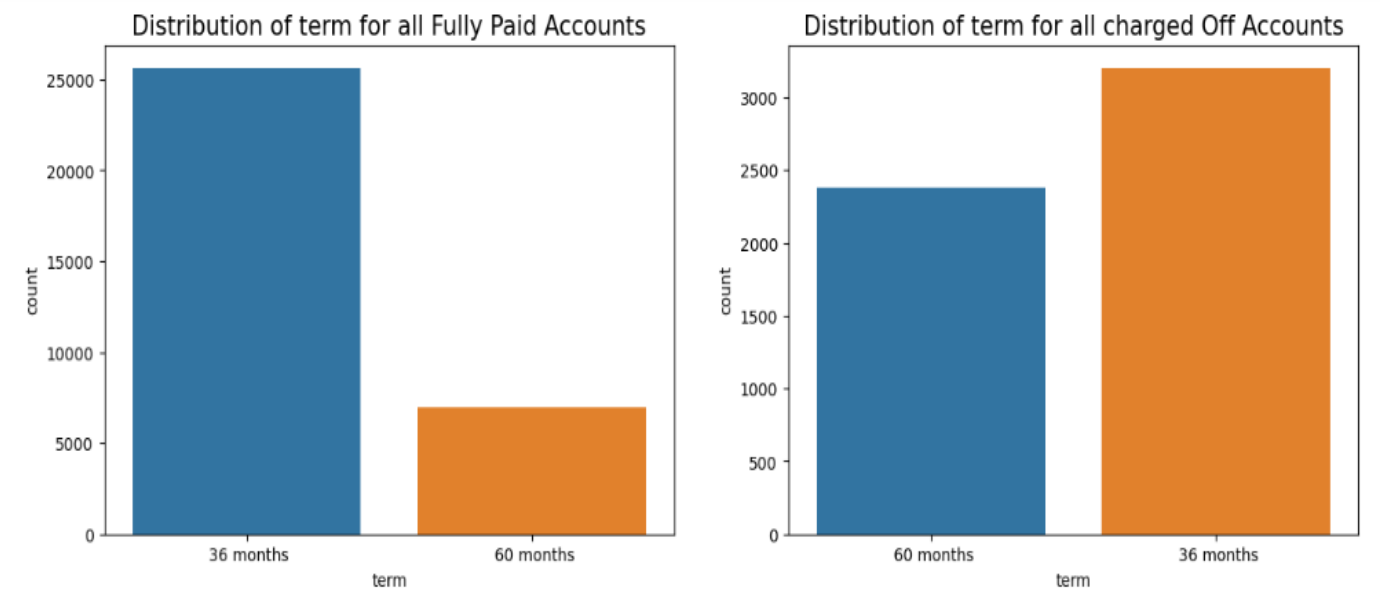


- A significant portion of defaulters has lower income levels

- Most defaulters have monthly installments falling within the 100 to 400 range

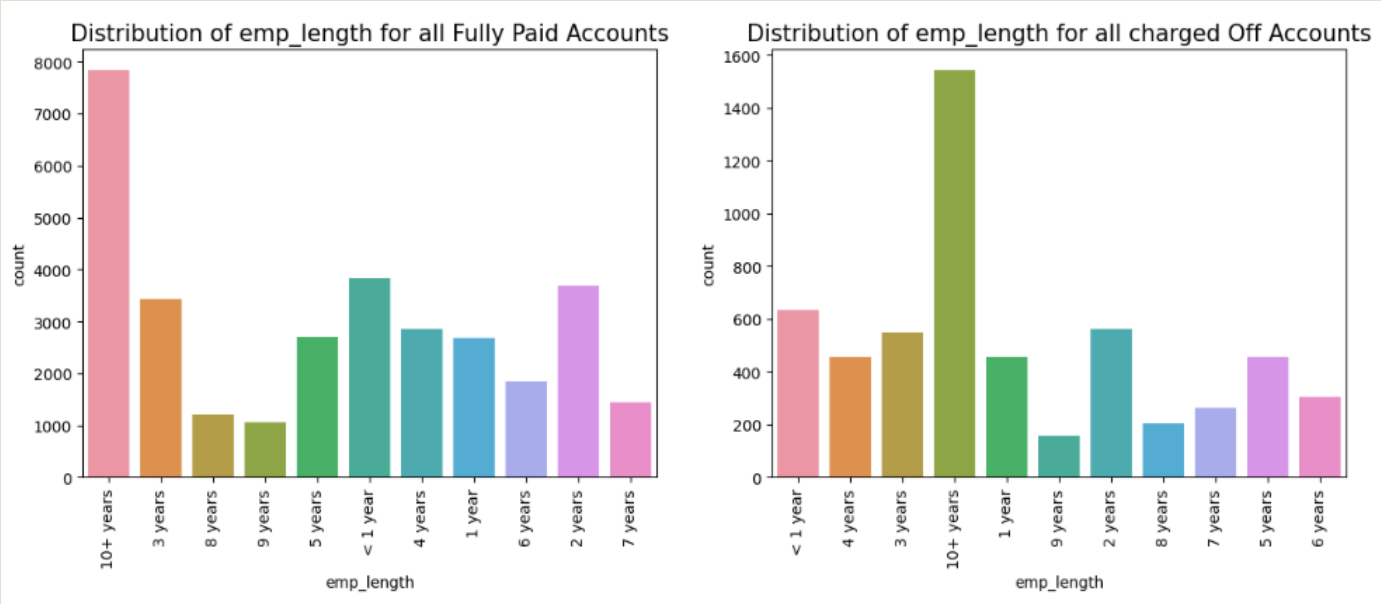


Distribution of Categorical Variables

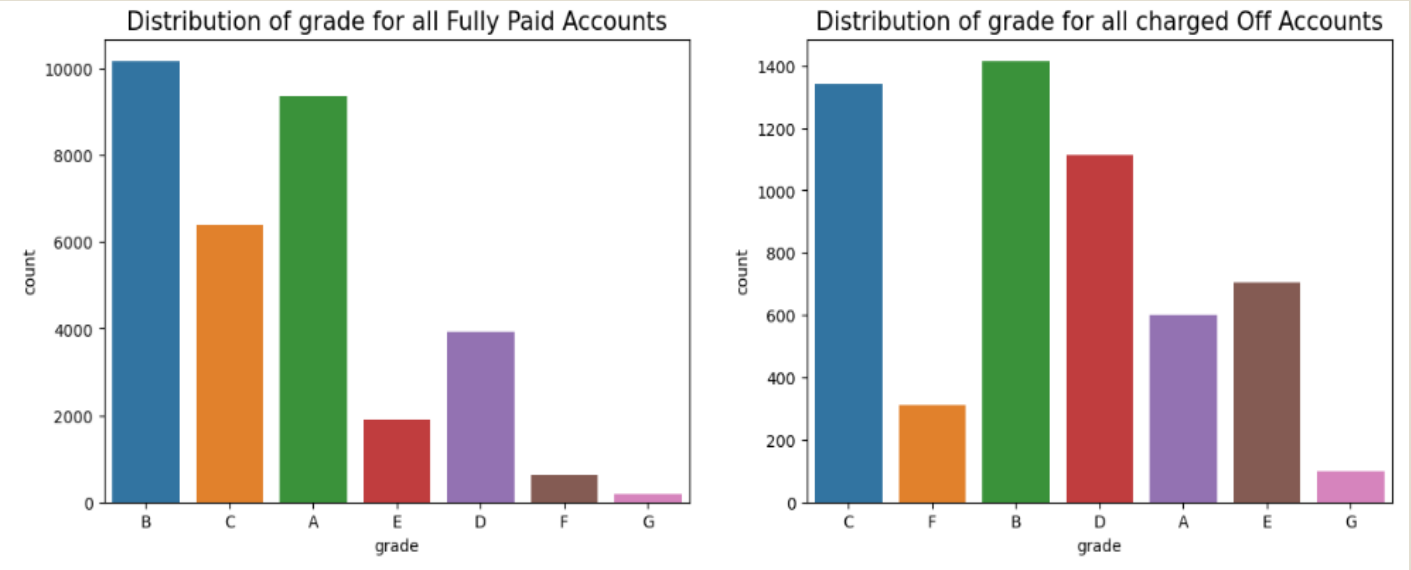


- Many borrowers favor a 36-month loan tenure, and this choice is associated with a lower likelihood of loan default

- Borrowers with more experience (10+ years) are at a higher risk of loan defaults.

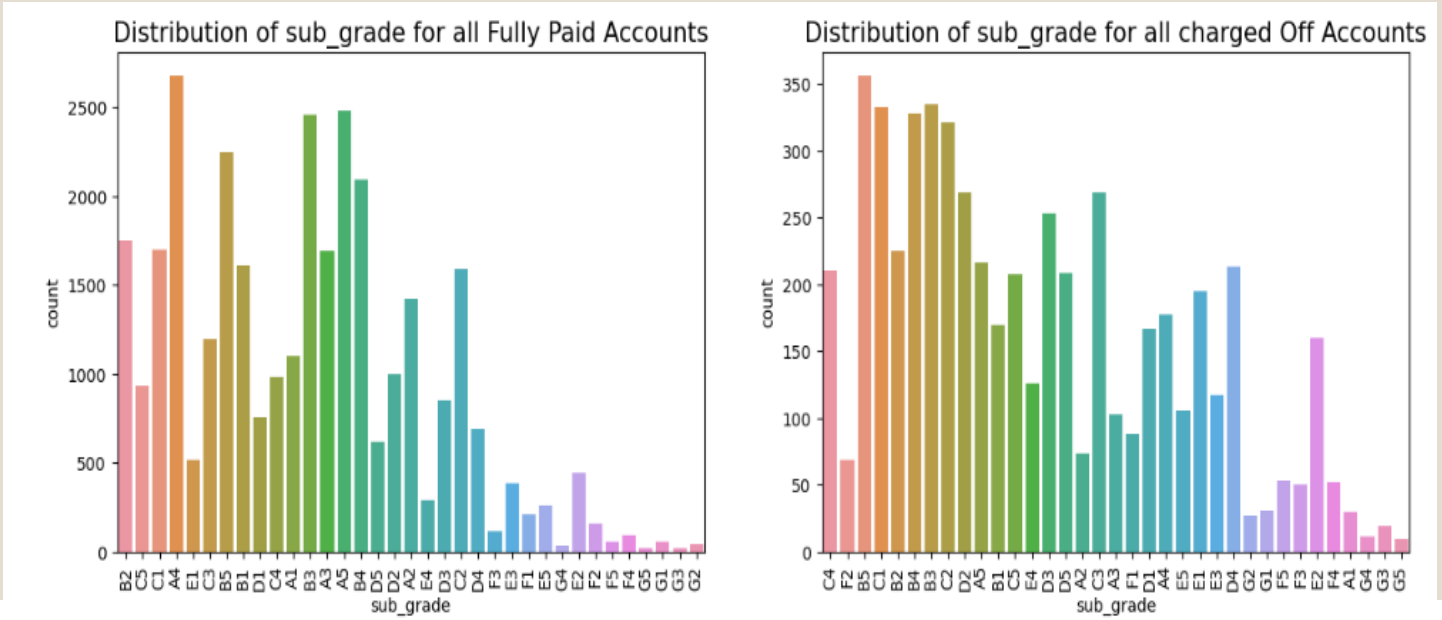


Distribution of Categorical Variables

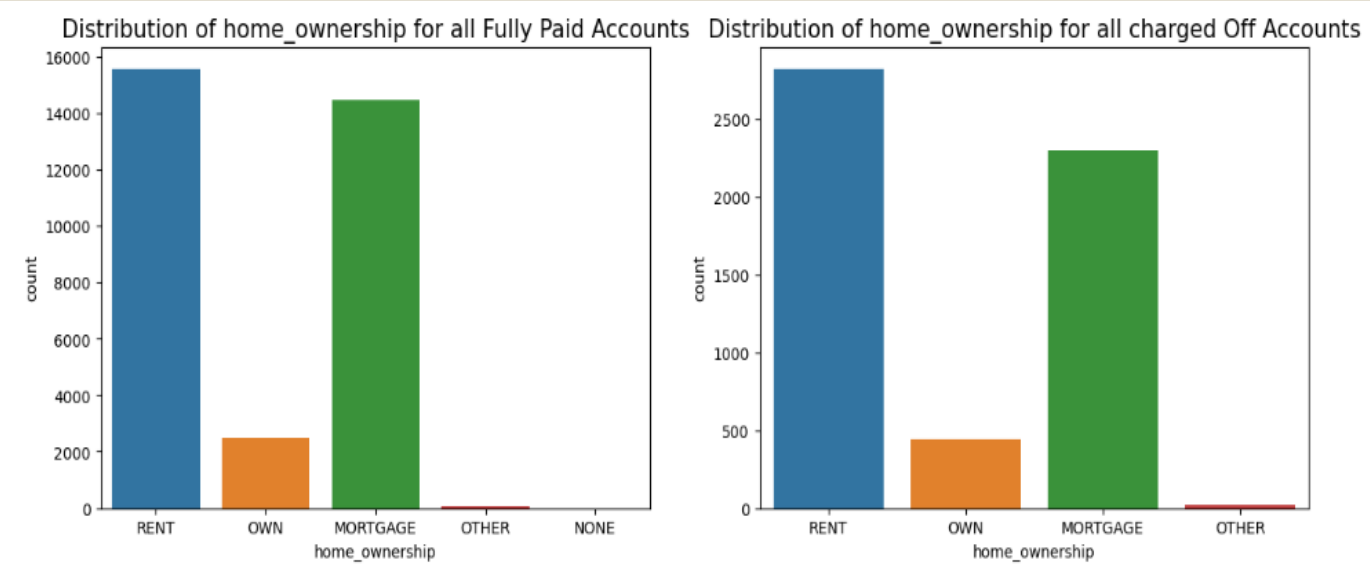


- Borrowers from Grade B have a higher default rate, and they are followed by those in Grade C in terms of default rates.

- Many defaulters fall under the B5 subgrade.

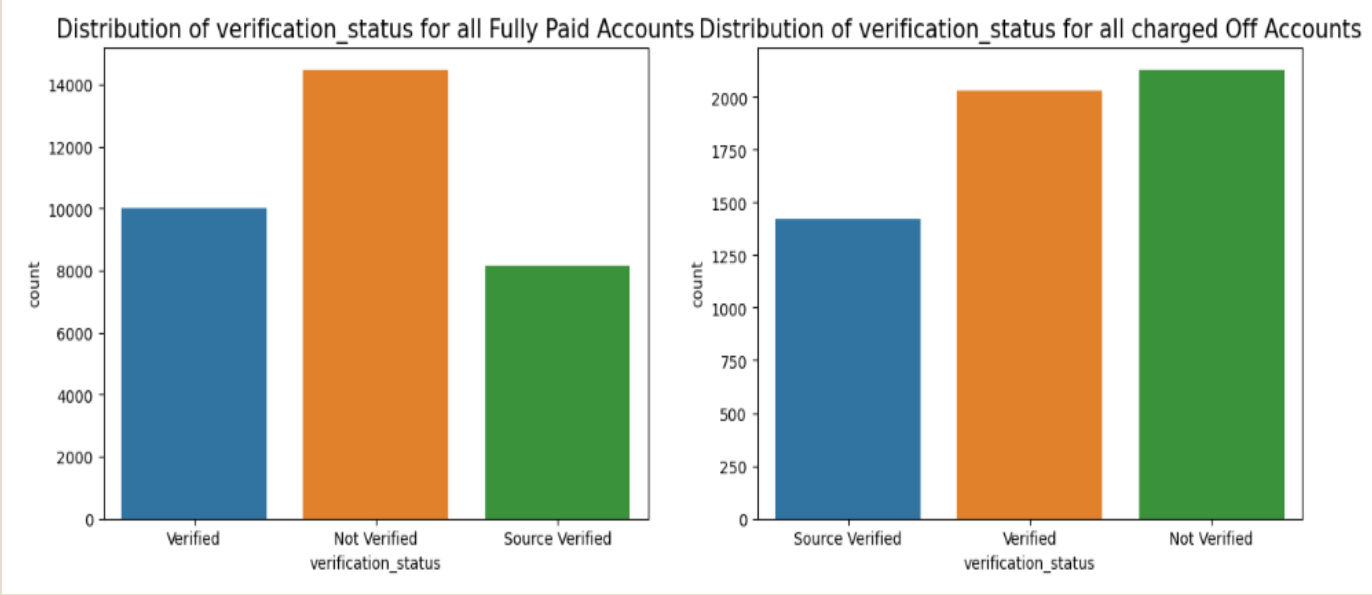


Distribution of Categorical Variables

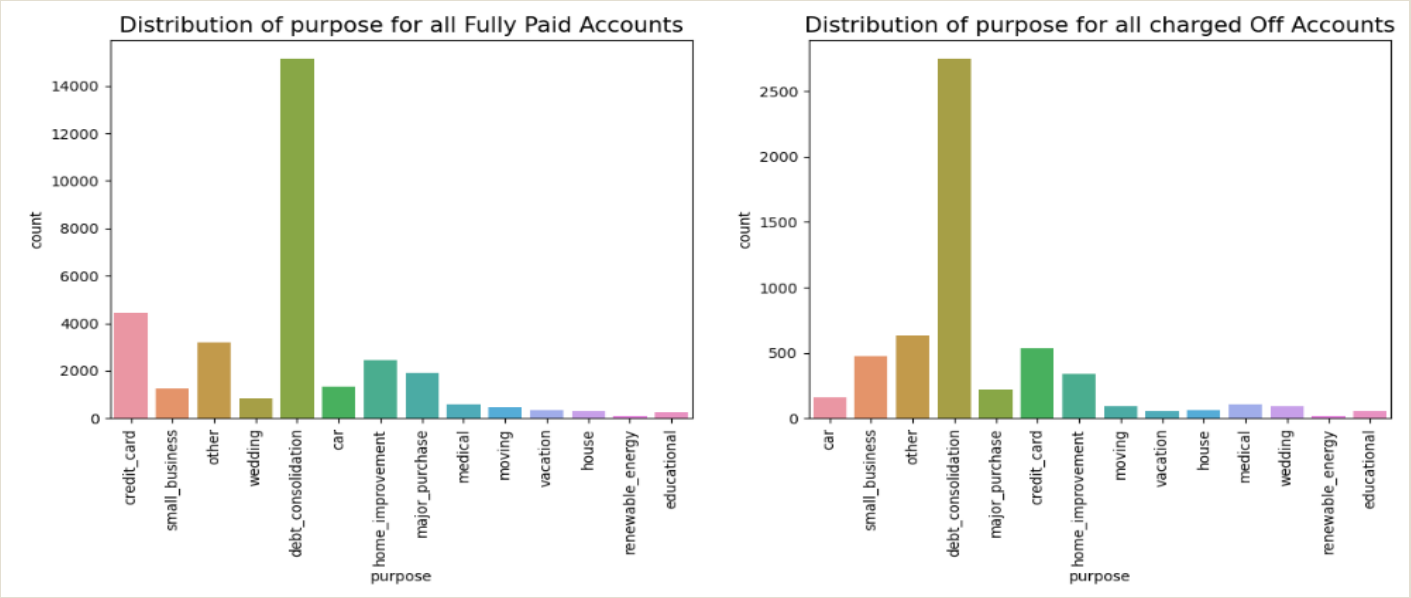


- Borrowers who rent a house have a higher likelihood of defaulting on their loans.

- Most borrowers did not have their income verified, and among them, the not-verified category has the highest number of defaulters, followed by the verified category.

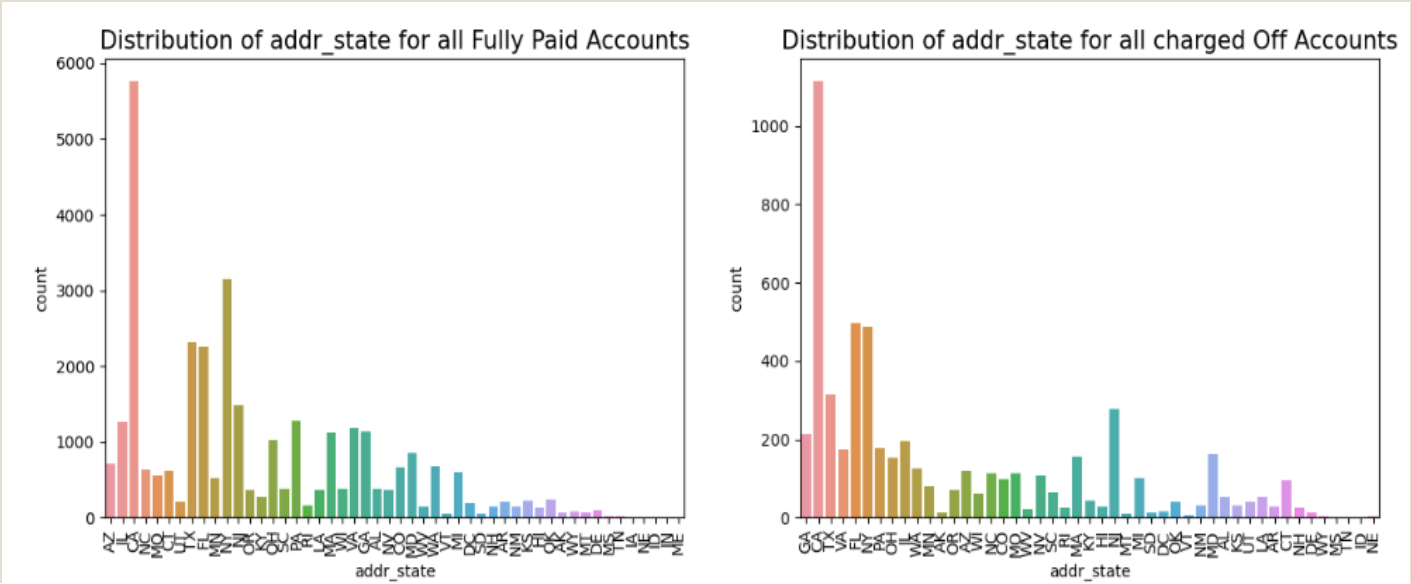


Distribution of Categorical Variables



- Borrowers who used the loan for debt consolidation are at a greater risk of defaulting.

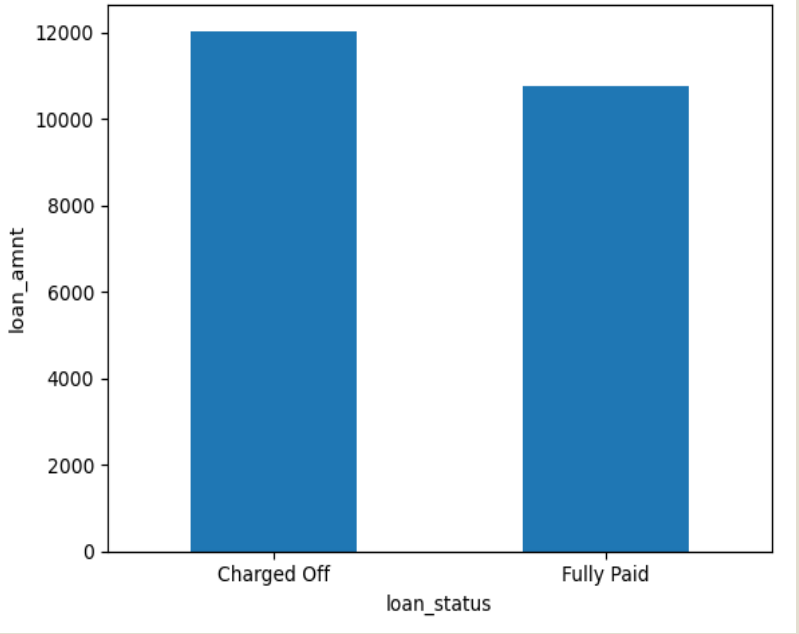
- Borrowers from California, Florida, and New York are more likely to default on their loans.

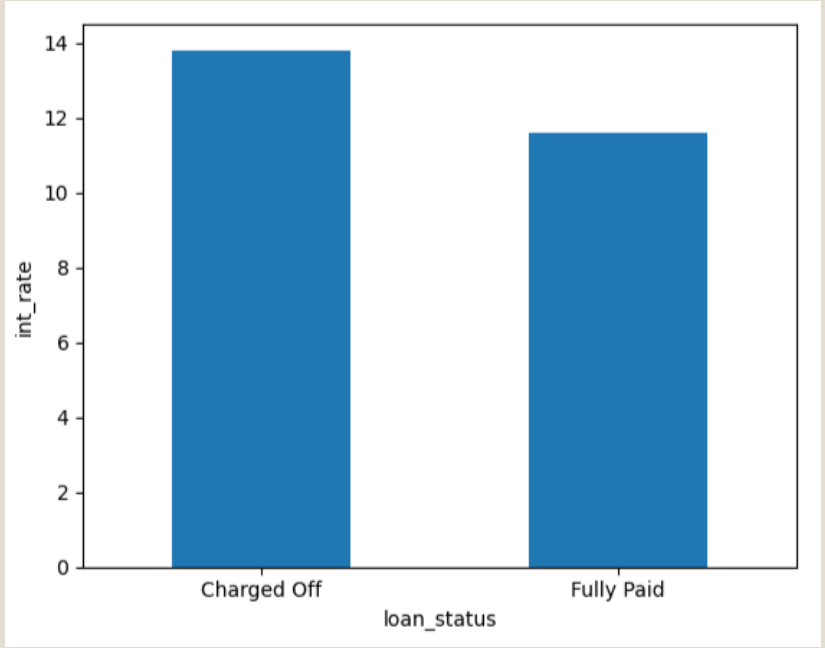


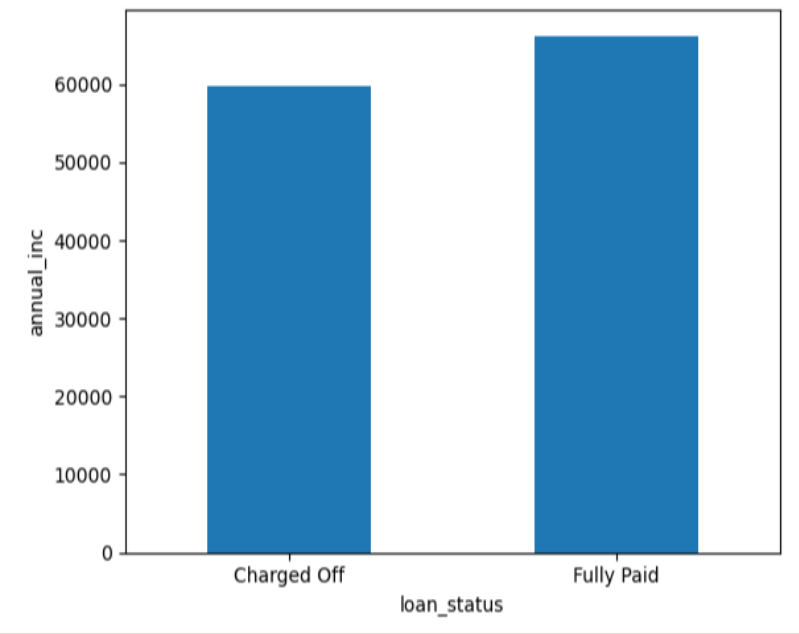
SEGMENTED UNIVARIATE ANALYSIS

Segmented Univariate Analysis

During this step, we have segmented the data based on the "oan_status variable, given that it is the target variable of interest.

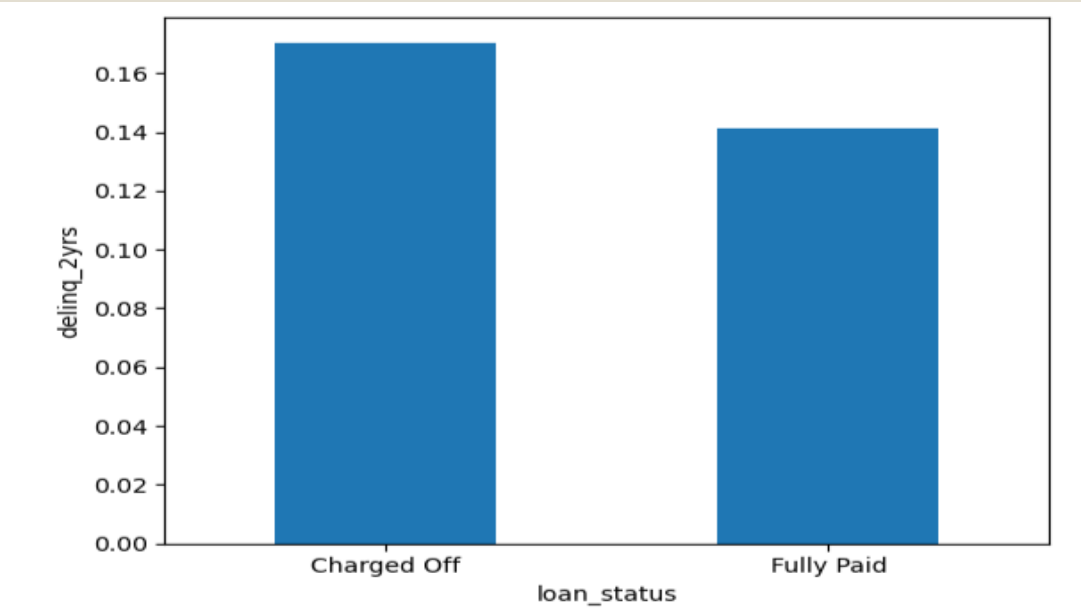


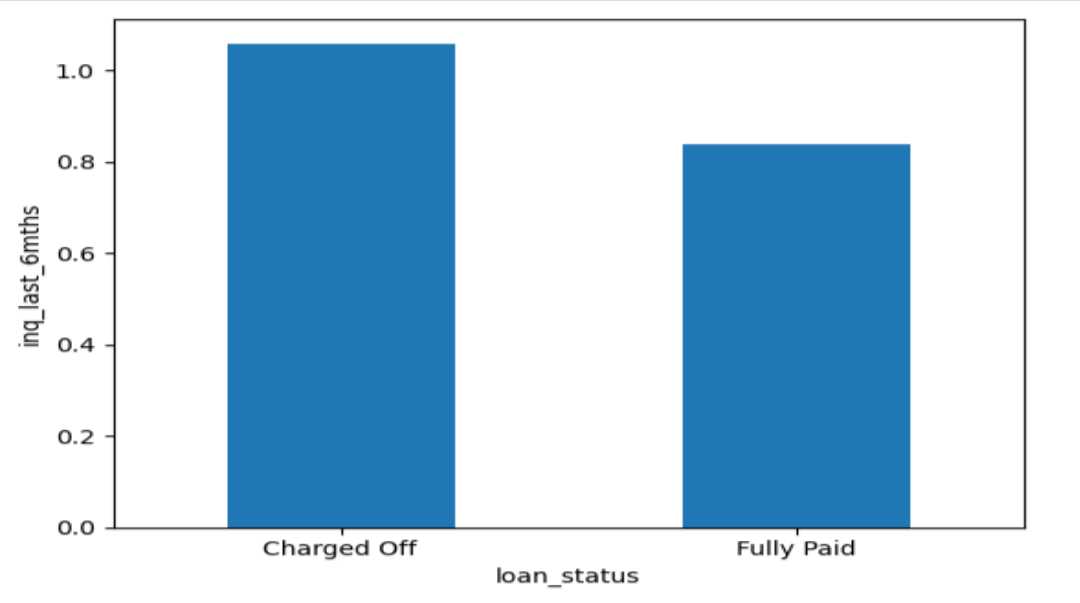




- Borrowers who choose larger loan amounts are at a higher risk of defaulting.
- Loans with higher interest rates and larger installment payments are more likely to result in defaults.
- Borrowers with lower annual incomes have an increased likelihood of defaulting.

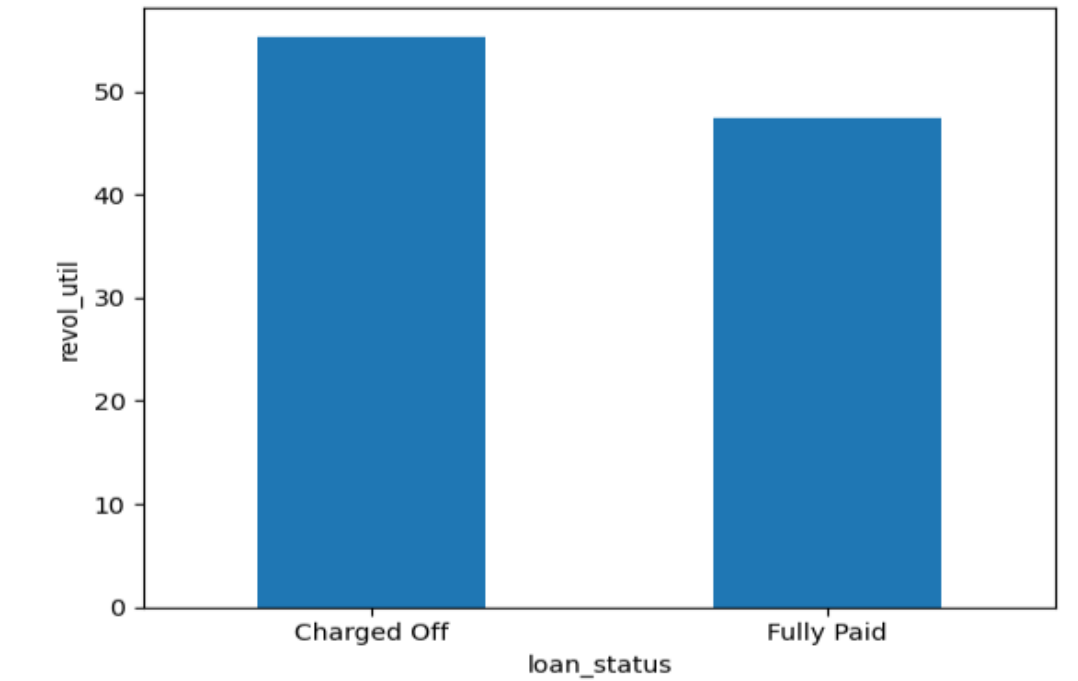
Segmented Univariate Analysis

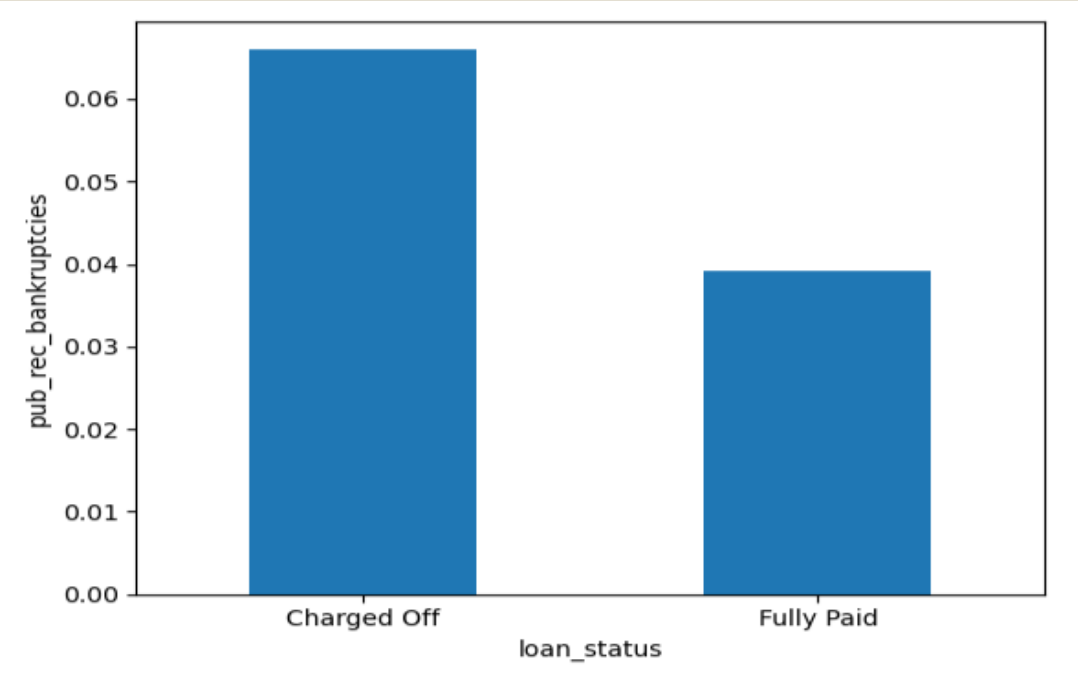




- Borrowers with high DTI (Debt-to-Income) ratios are more likely to default on their loans.
- A history of delinquencies in a borrower's credit history increases the likelihood of loan defaults.
- A higher number of inquiries in the last 6 months on a borrower's credit report also raises the chances of defaulting on a loan.

Segmented Univariate Analysis

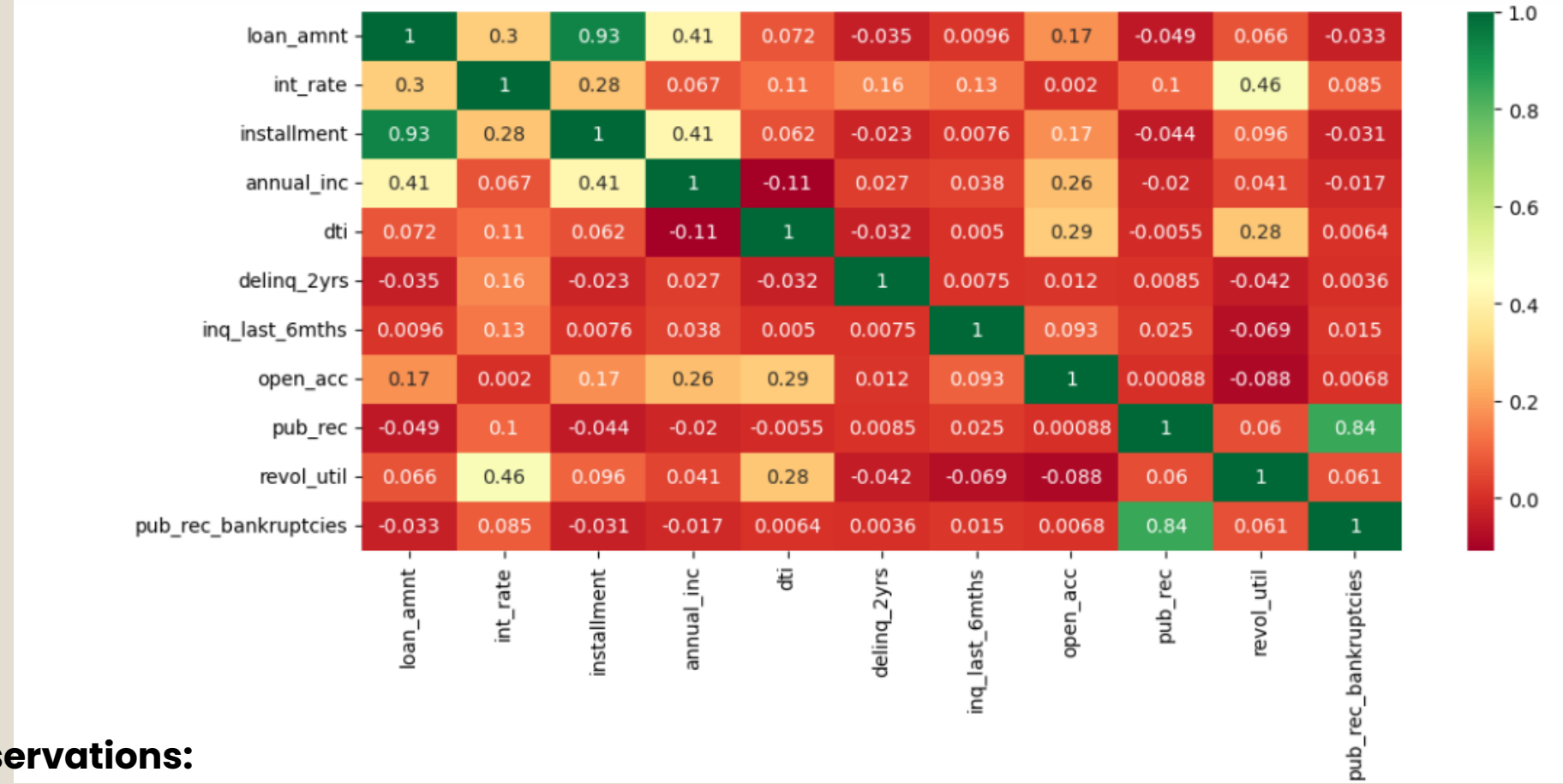




- Borrowers with greater revolving credit utilization are at a higher risk of defaulting on their loans.
- A higher count of public records of bankruptcies in a borrower's credit history increases the likelihood of loan defaults.

BIVARIATE ANALYSIS

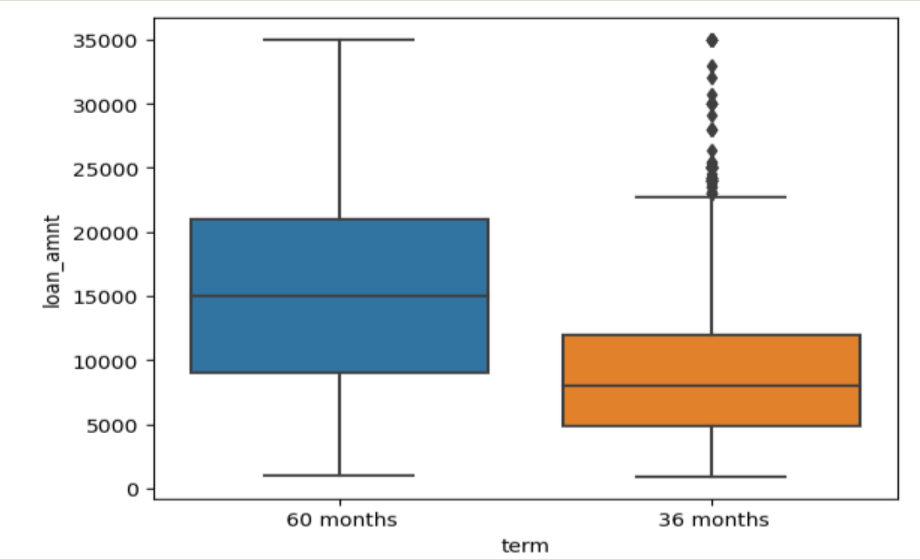
Bivariate Analysis - Numerical Vs Numerical

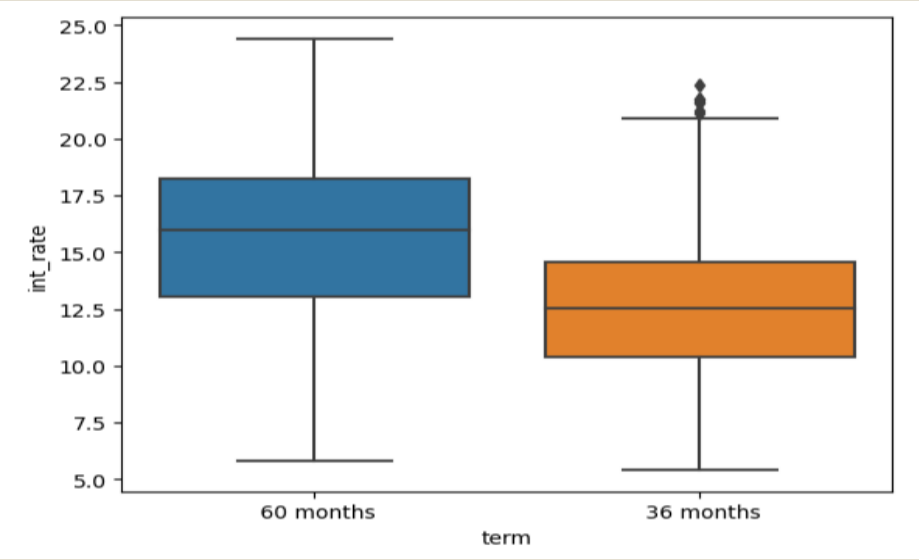


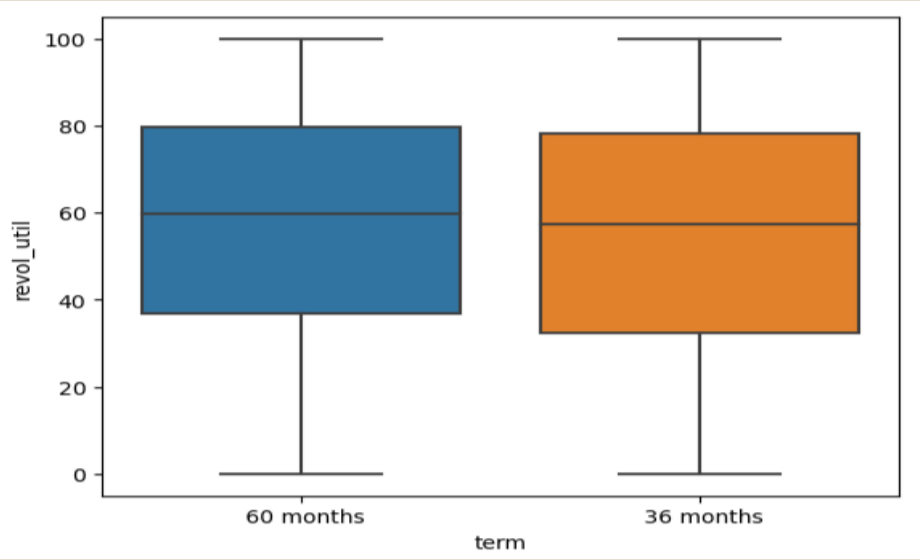
Observations:

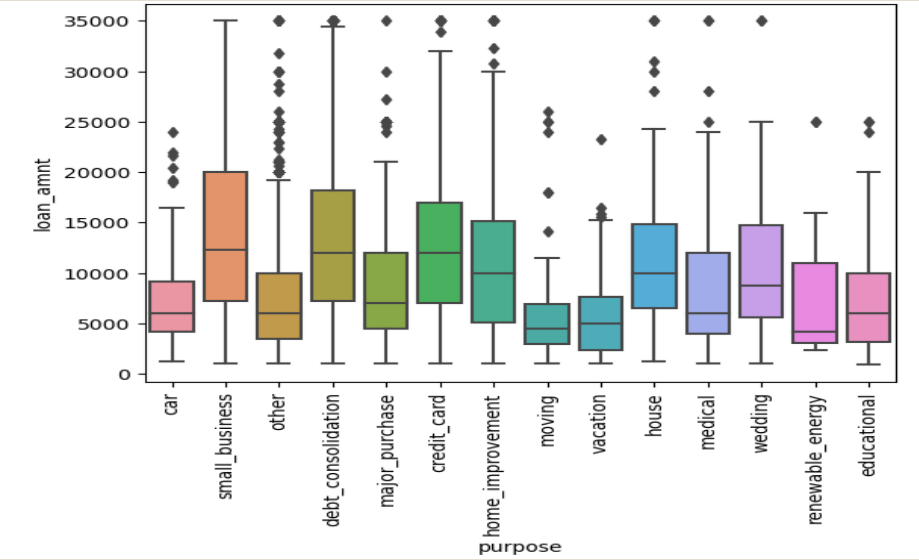
- Loan amount and installments exhibit a strong correlation.
- Derogatory public records and public records of bankruptcies are significantly correlated.
- Apart from the mentioned four columns, there is a limited linear relationship among the remaining variables.

Bivariate Analysis - Numerical Vs Categorical

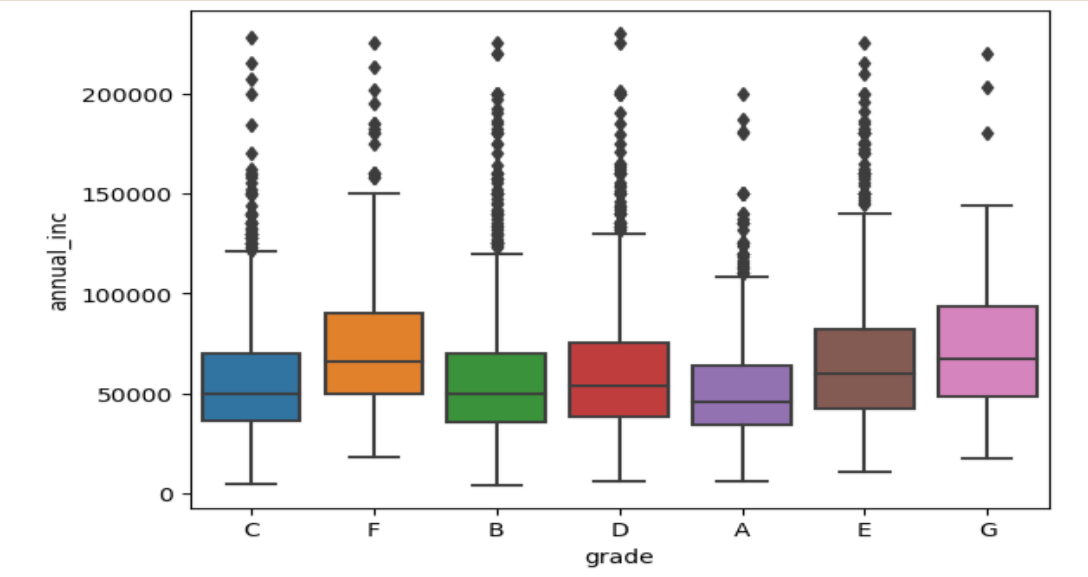


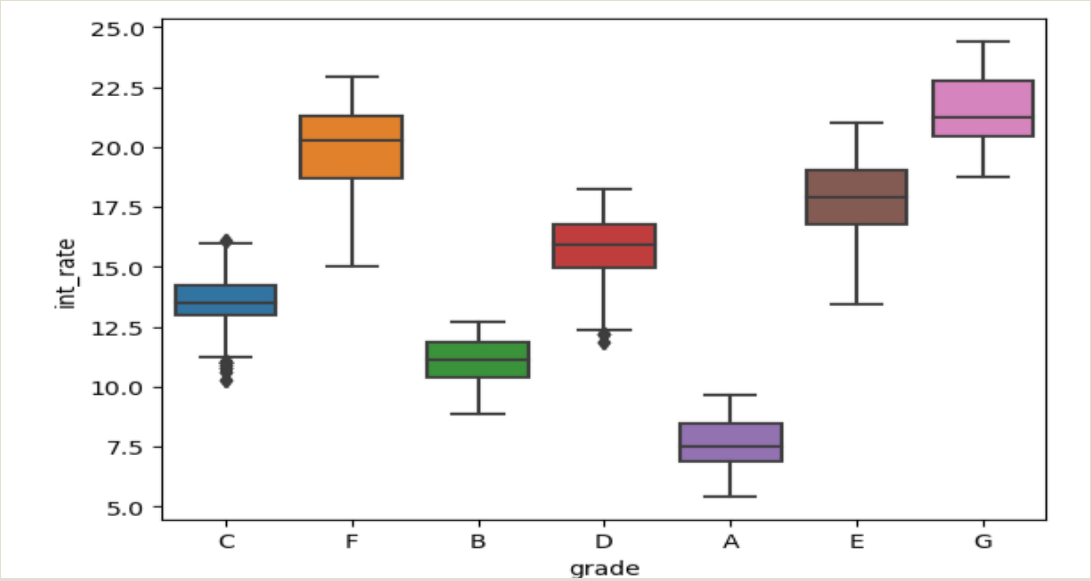


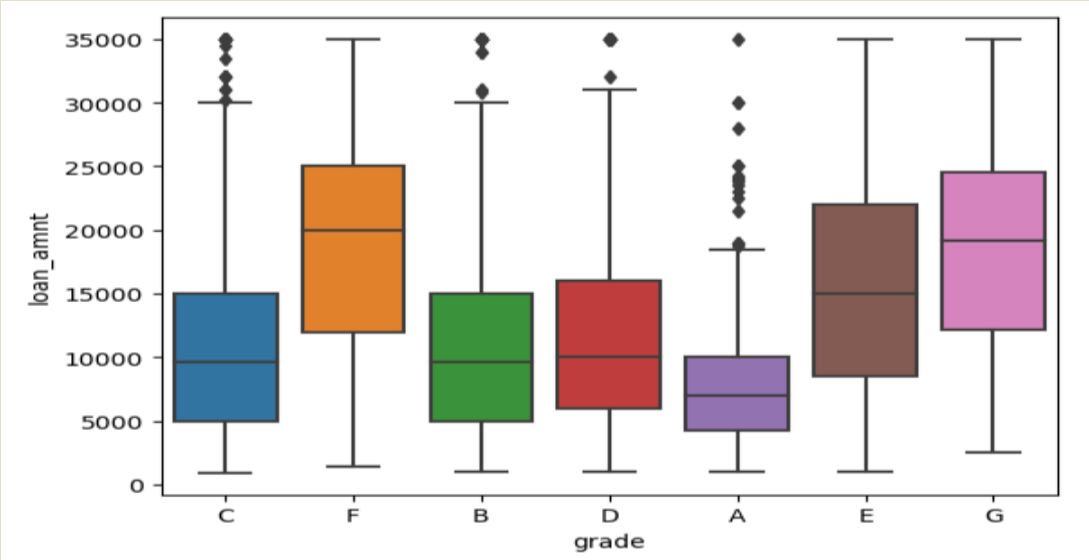


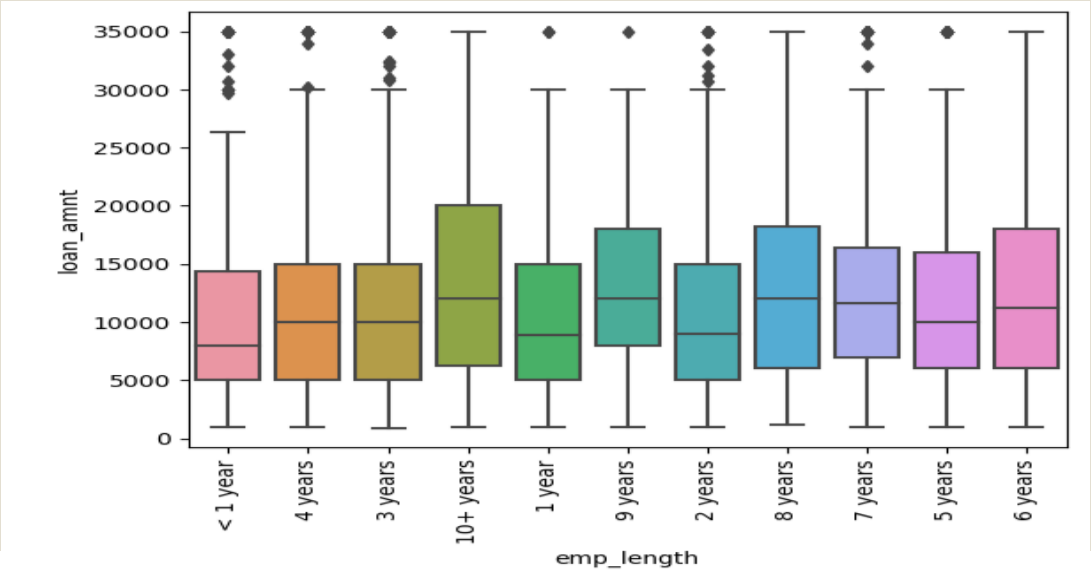


Bivariate Analysis - Numerical Vs Categorical

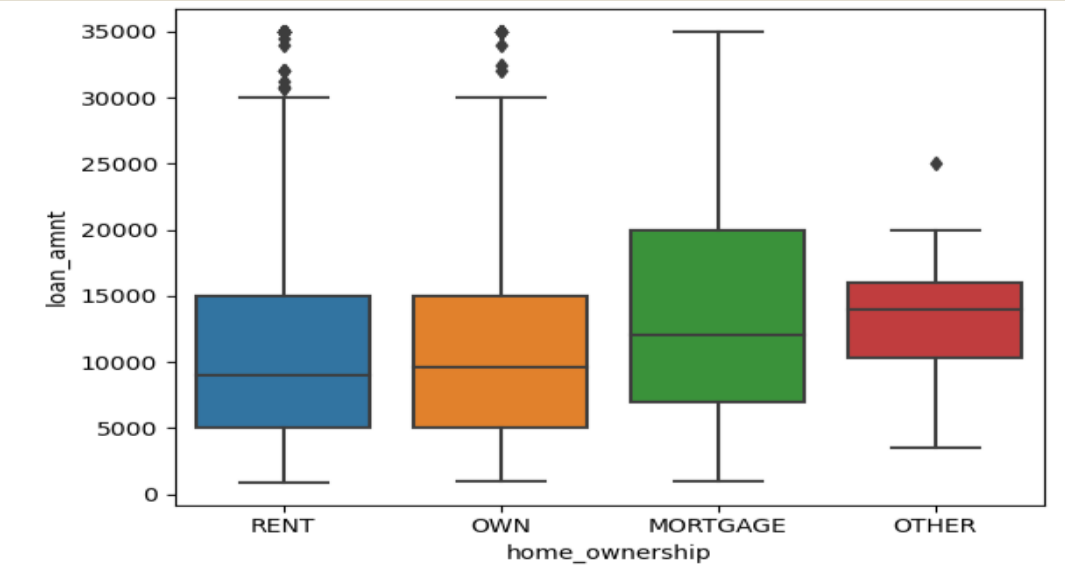


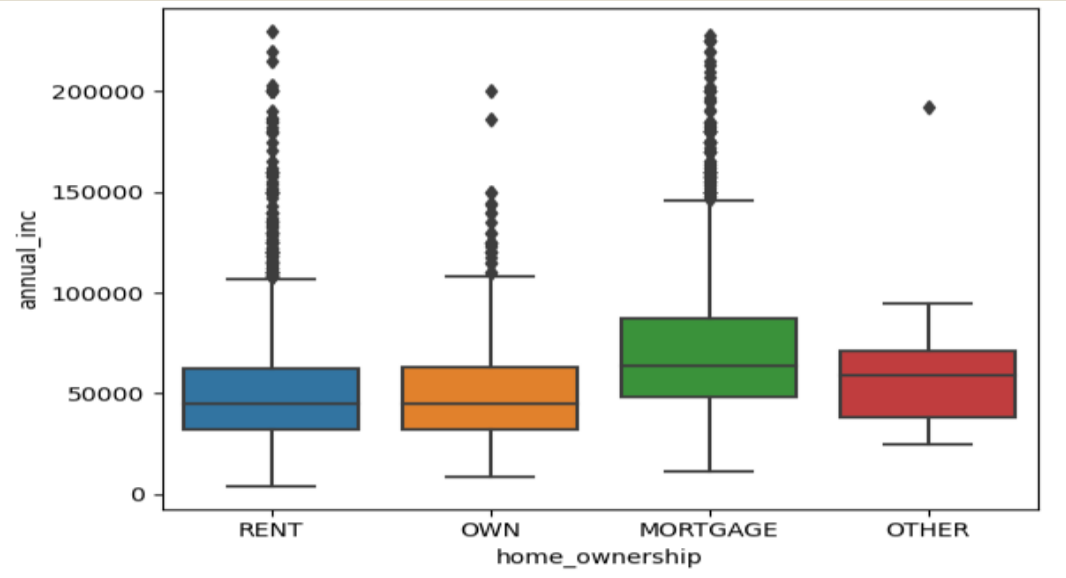


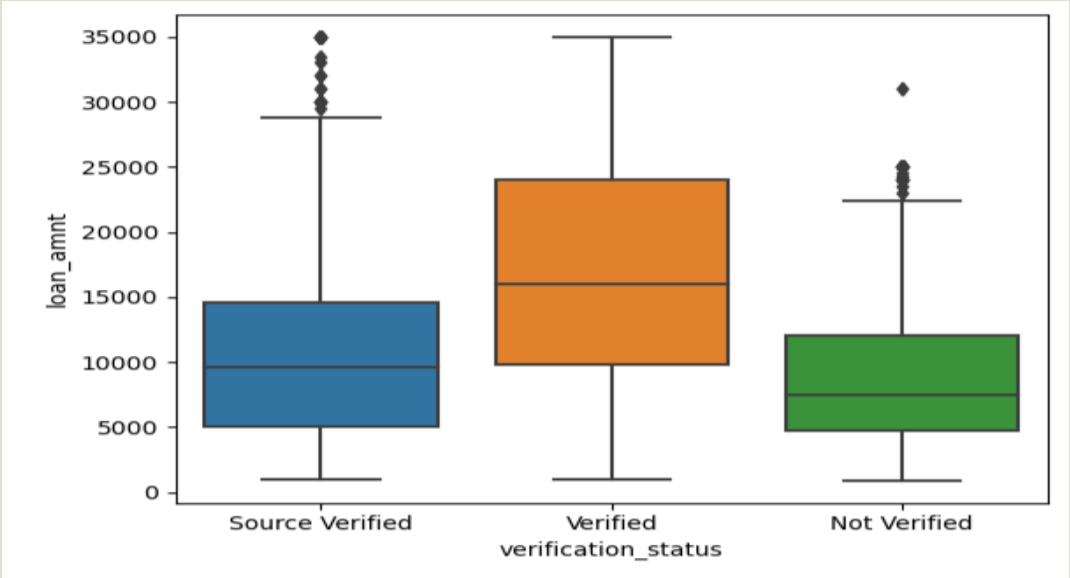


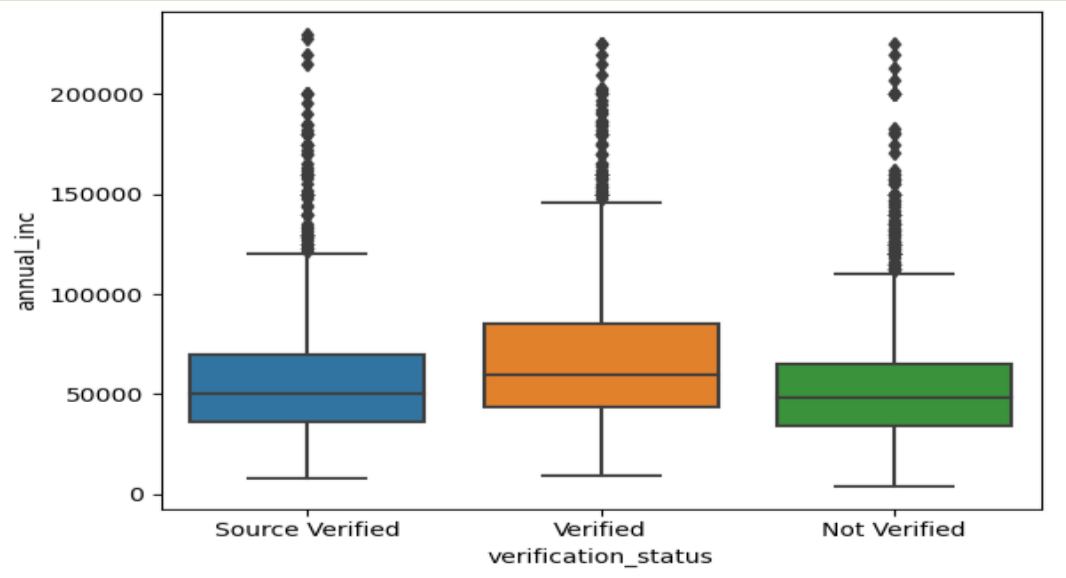


Bivariate Analysis - Numerical Vs Categorical





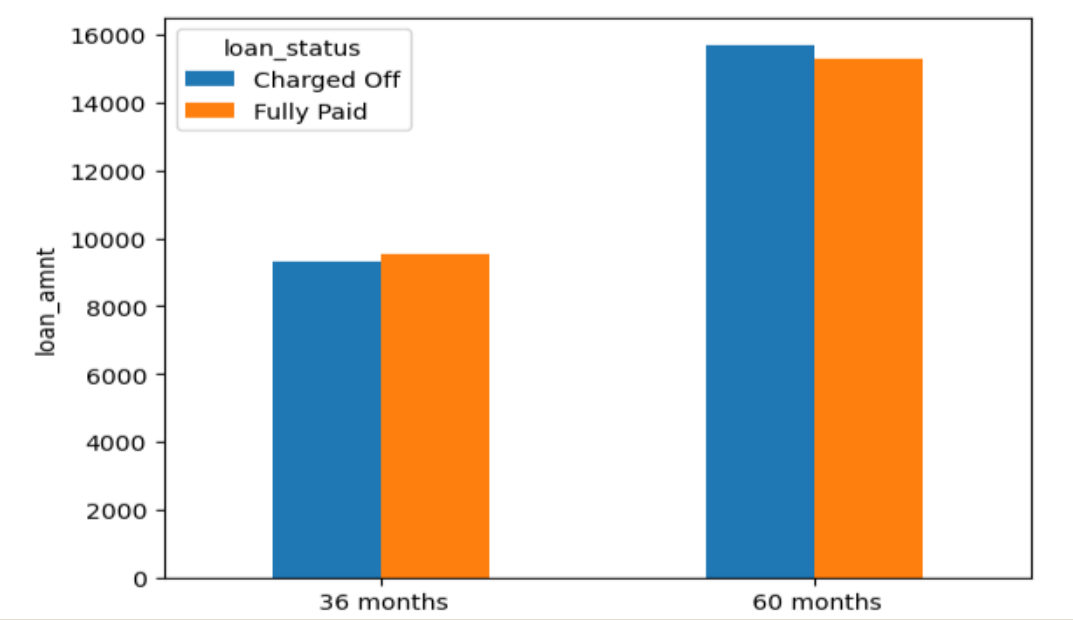


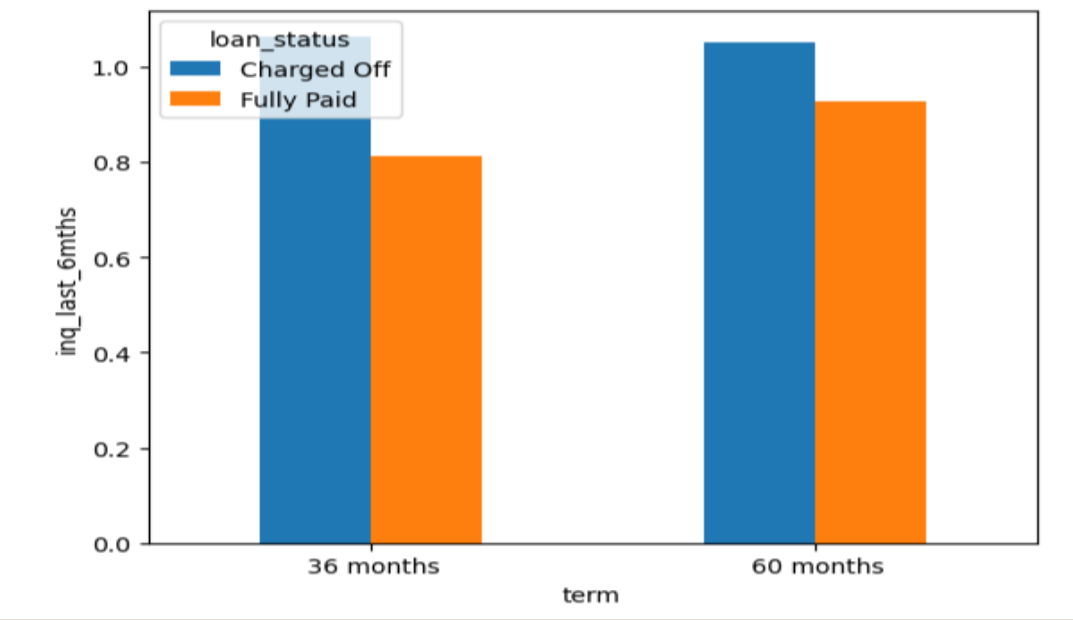


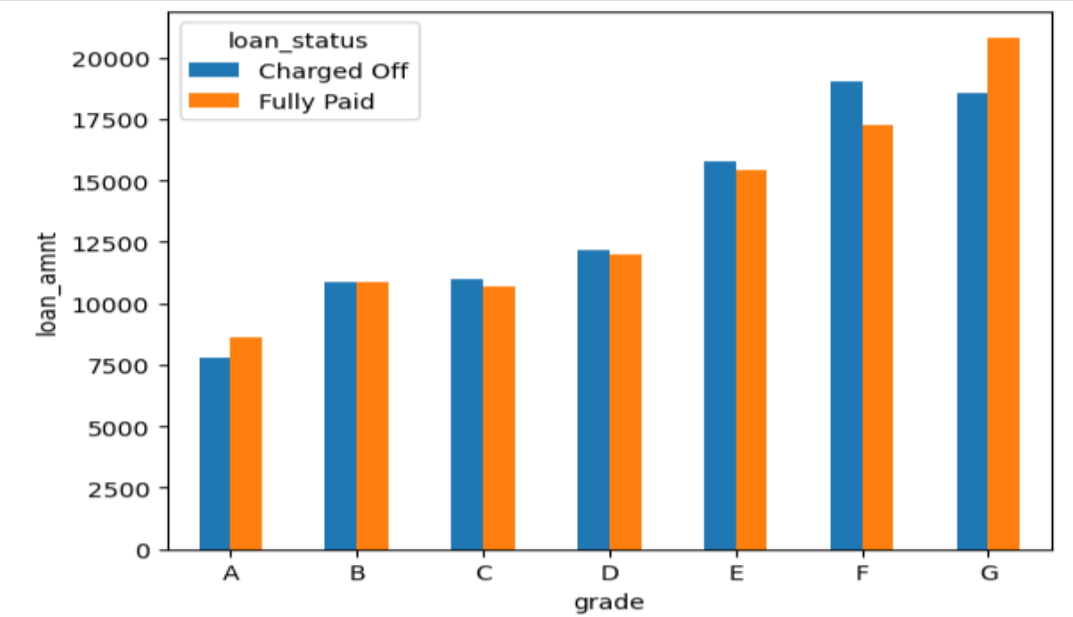
Observations - Numerical Vs Categorical

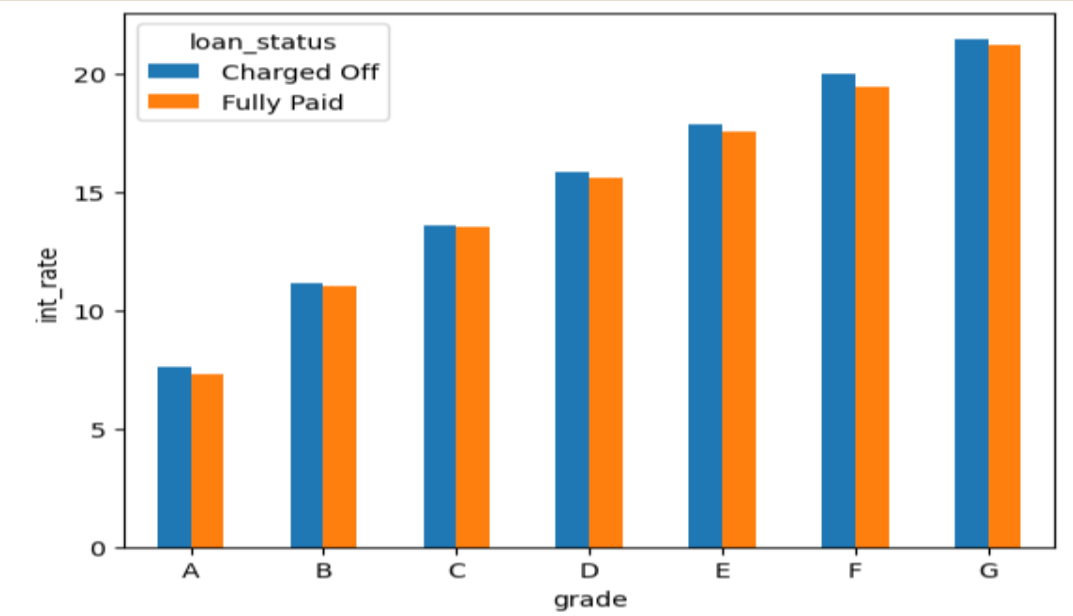
- Borrowers who choose a 60-month loan tenure tend to request larger loan amounts, and longer tenures generally result in higher interest payments. Additionally, loans with longer tenures have higher monthly installments.
- Borrowers with longer tenures also tend to have slightly higher revolving credit utilization.
- Borrowers in Grades F and G typically apply for larger loan amounts and face higher interest rates. Moreover, these borrowers tend to have higher annual incomes.
- Grade G borrowers tend to have a wider range of open credit accounts.
- Borrowers with mortgages tend to request higher loan amounts and have higher annual incomes.
- Borrowers in the "other" category of home ownership have higher monthly installment payments.
- Borrowers with more experience tend to request larger loan amounts.
- Borrowers with verified incomes are approved for larger loan amounts and have higher income levels.
- The majority of loans are approved for small businesses, with debt consolidation being the second most common loan purpose.

Bivariate Analysis – Categorical Vs Categorical

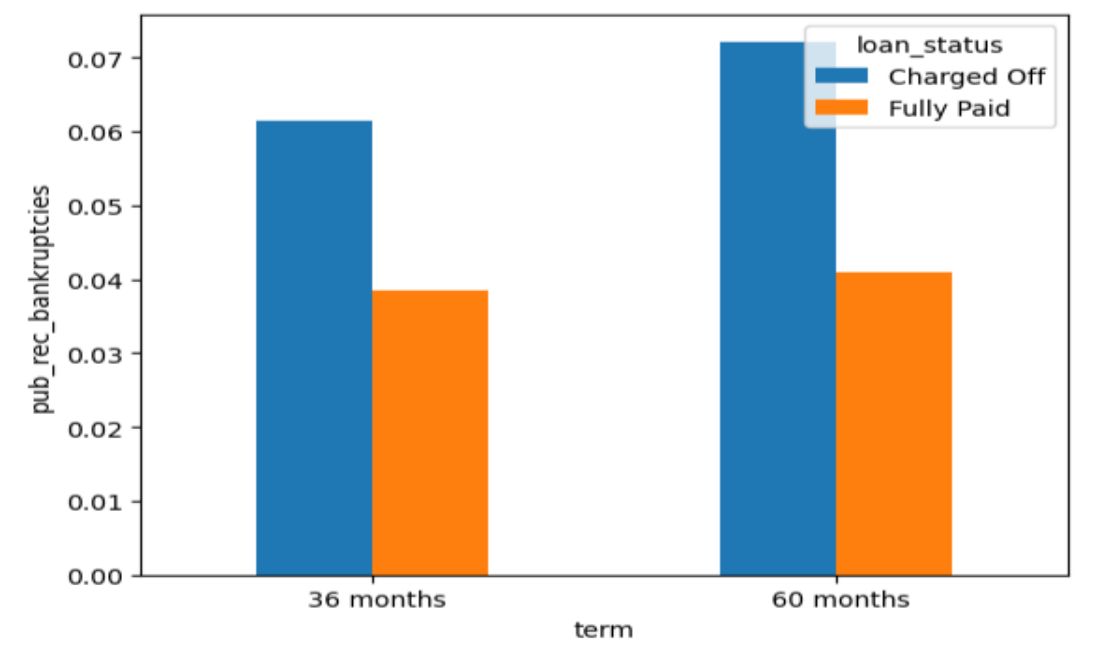


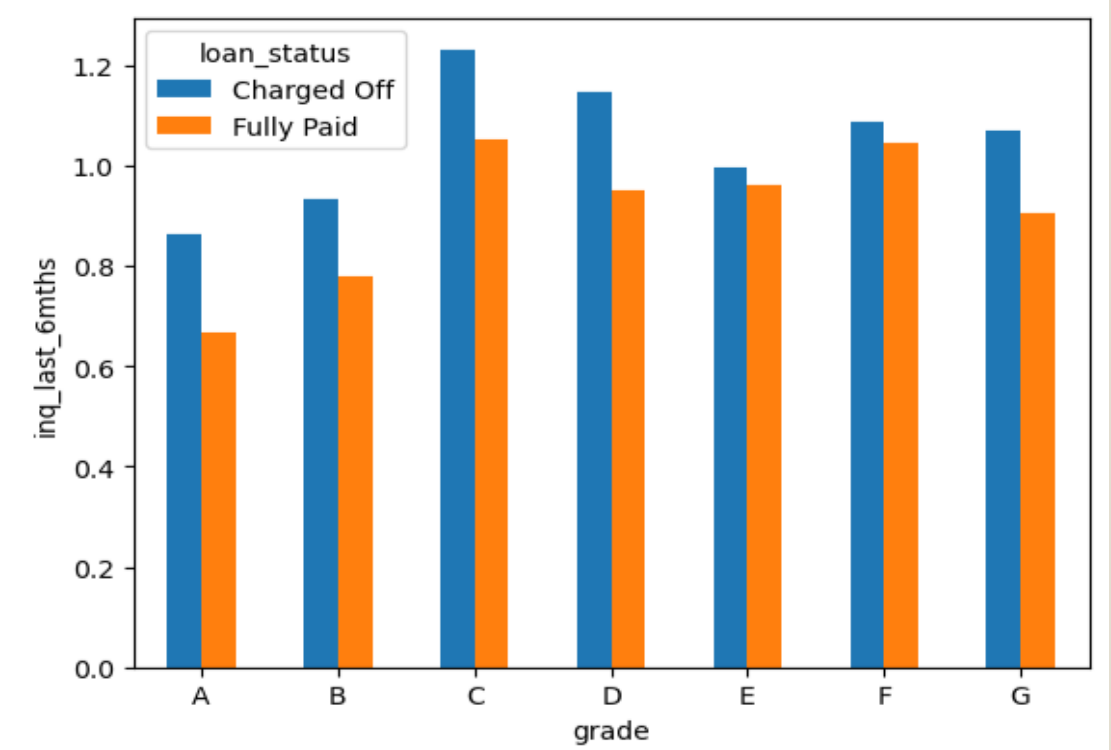




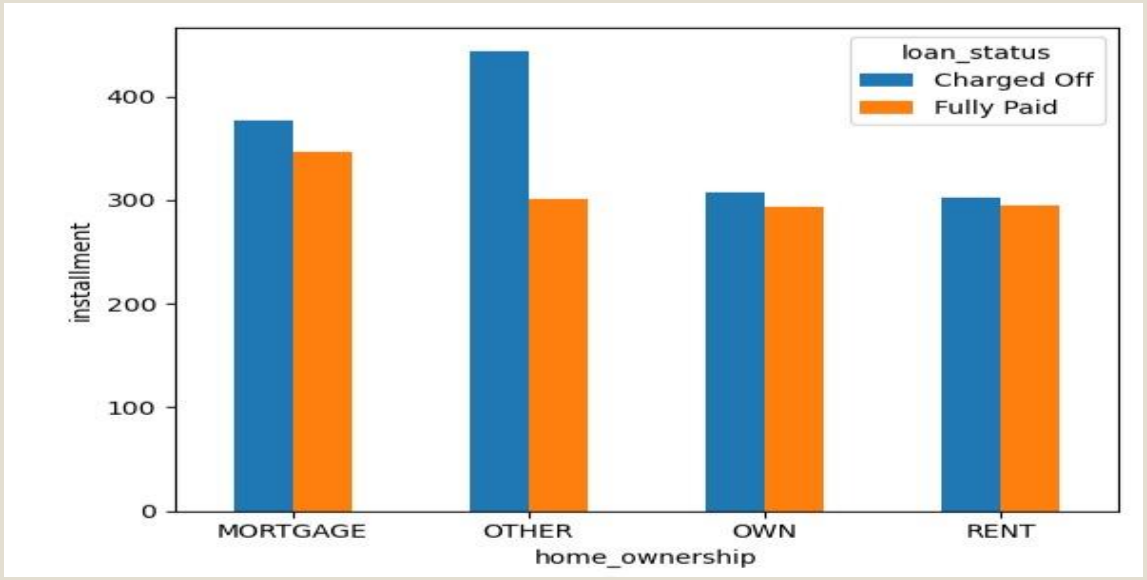


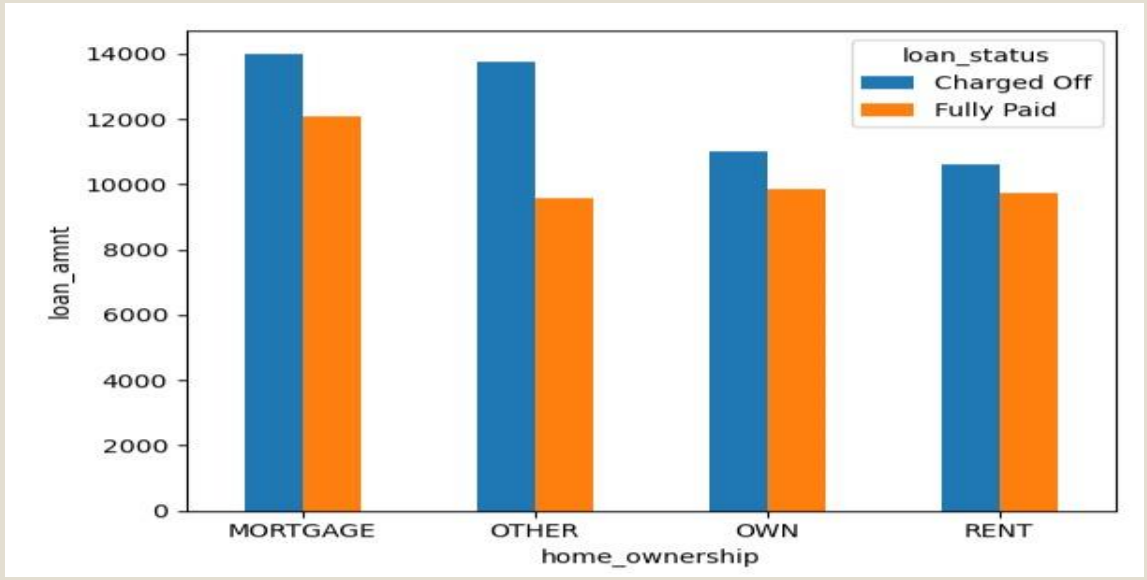
Bivariate Analysis – Categorical Vs Categorical

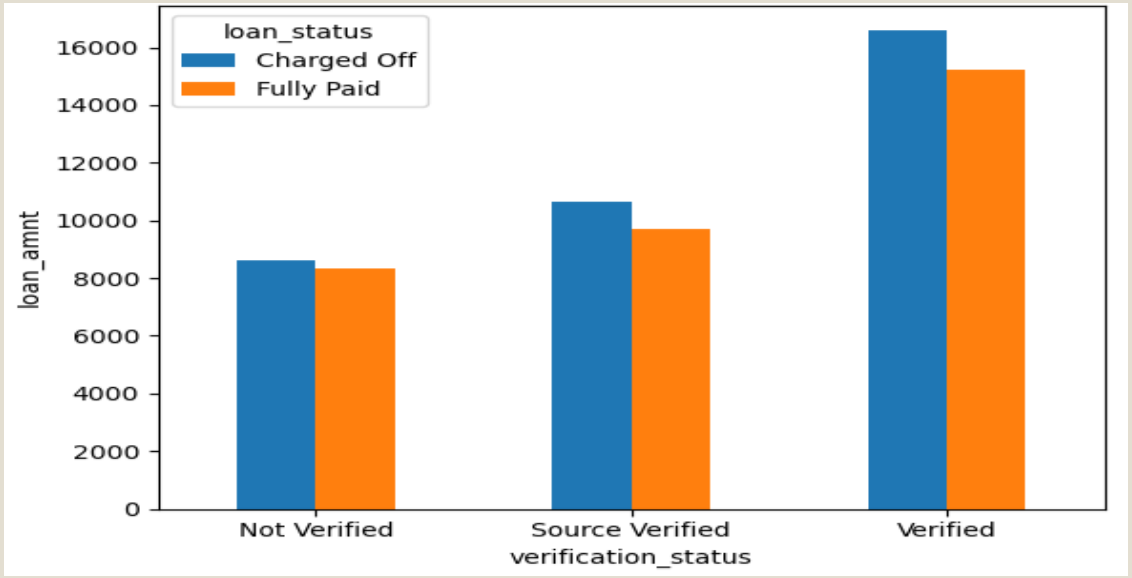


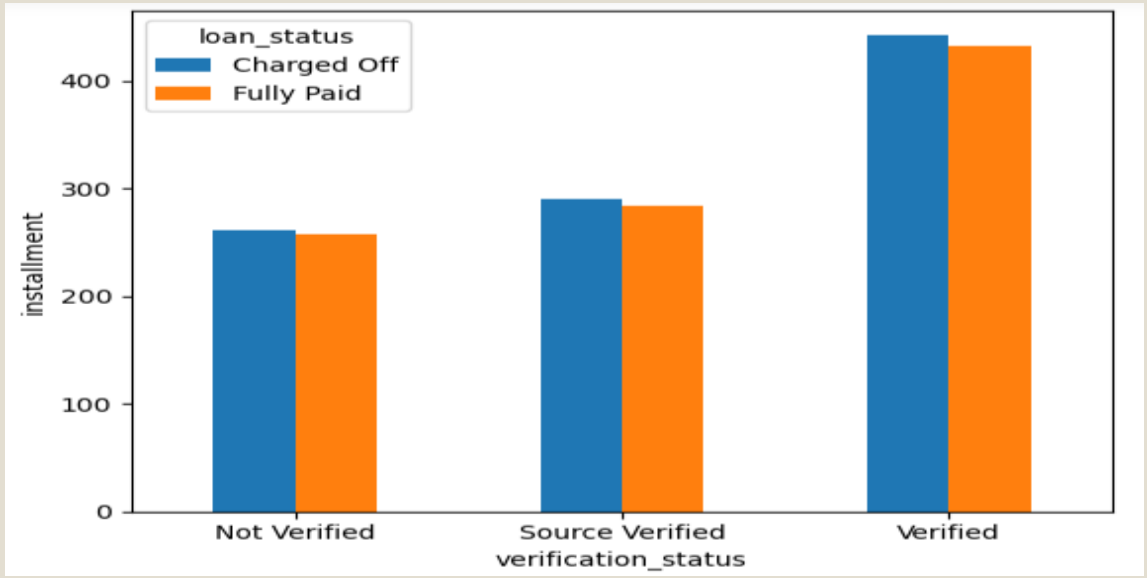


Bivariate Analysis – Categorical Vs Categorical

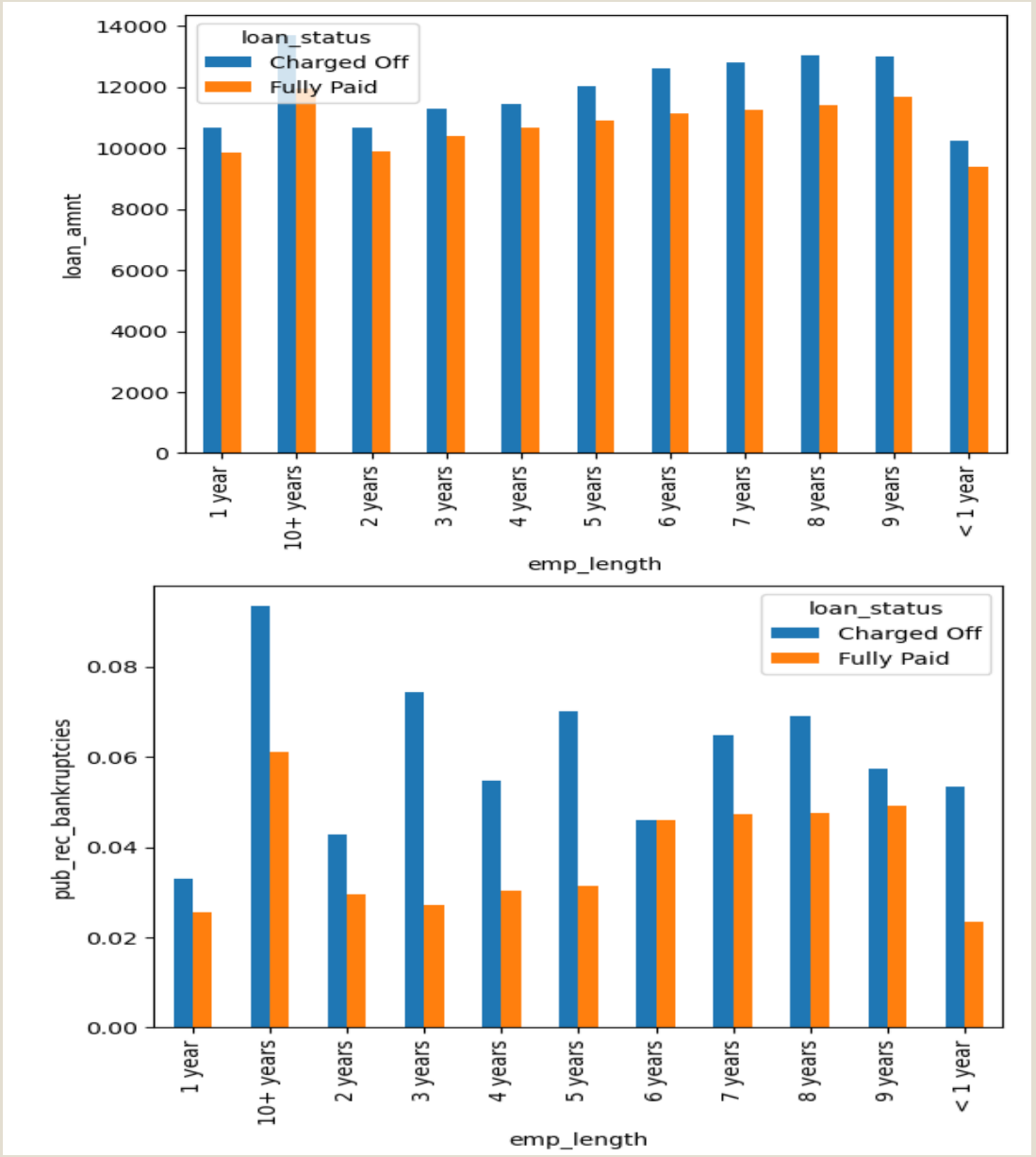


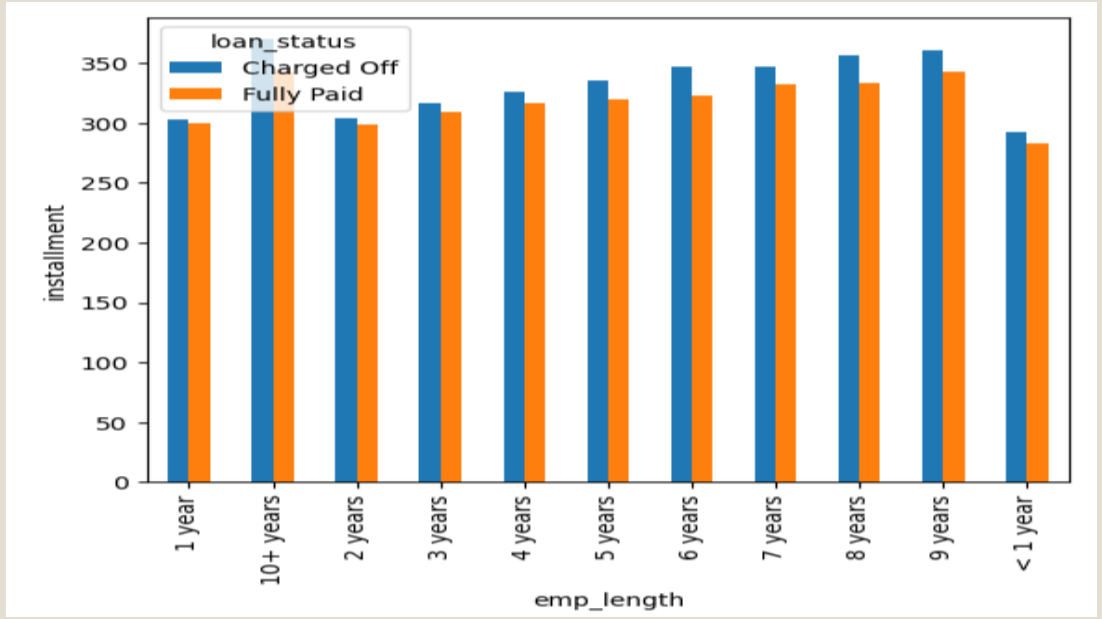


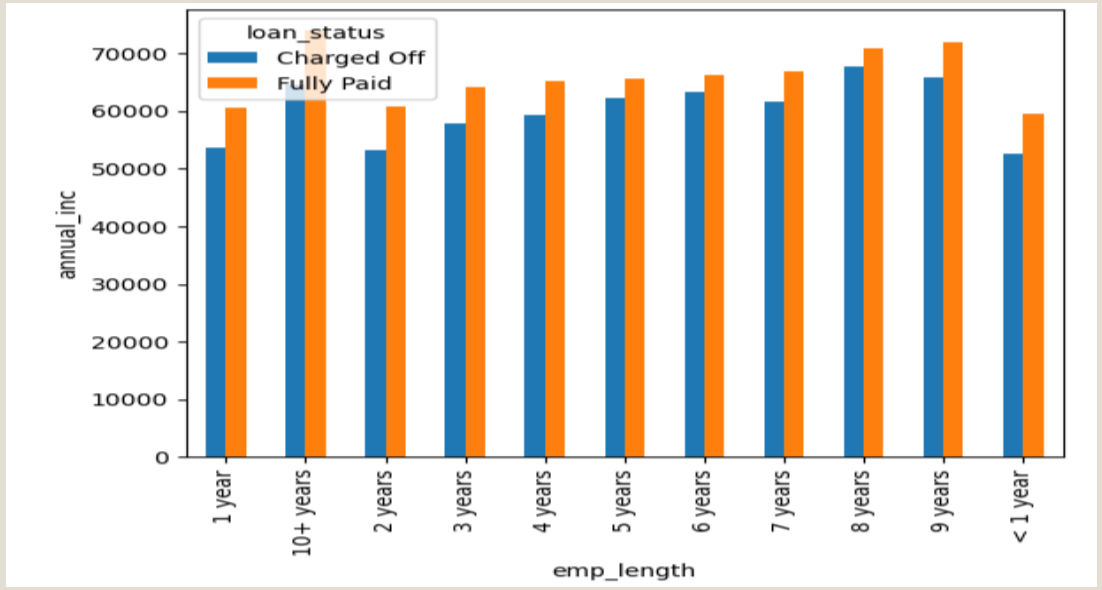




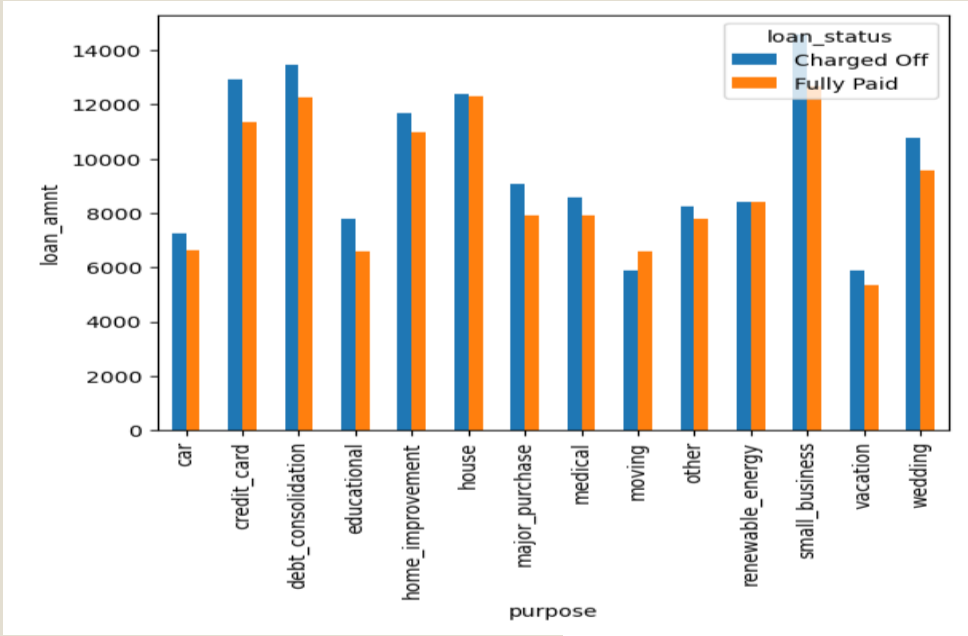
Bivariate Analysis – Categorical Vs Categorical

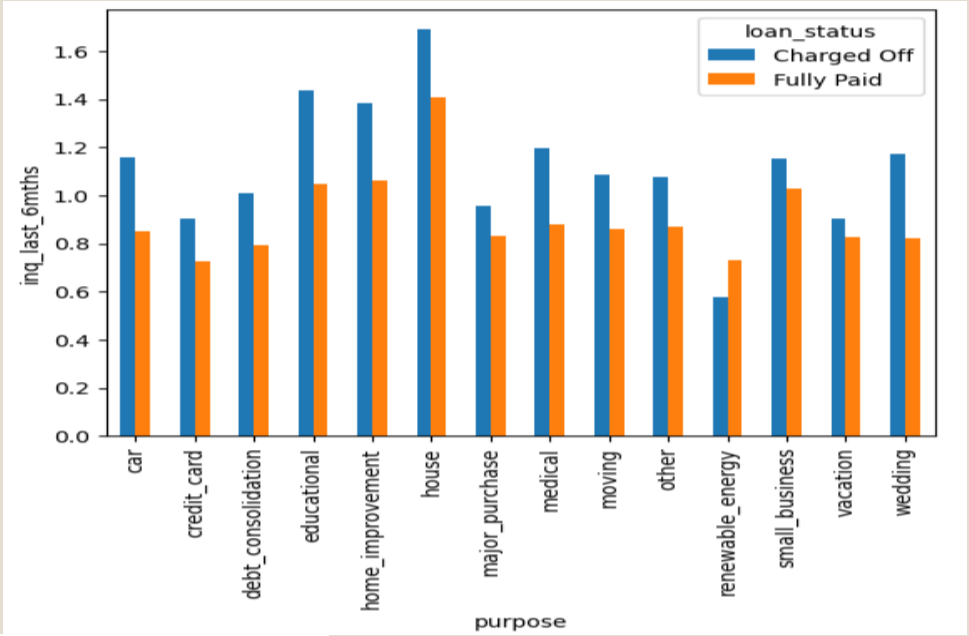


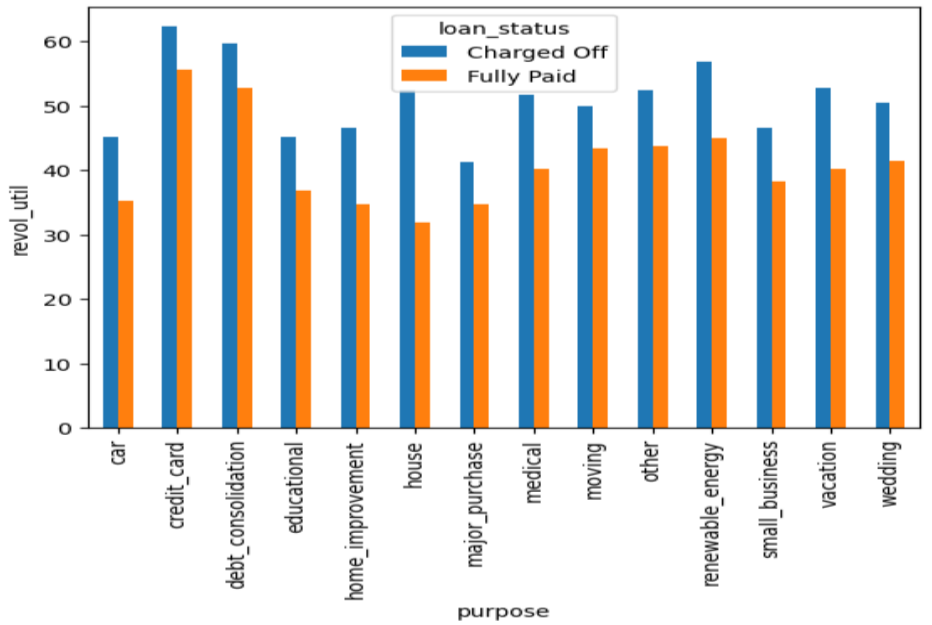




Bivariate Analysis – Categorical Vs Categorical







Observations - Categorical Vs Categorical

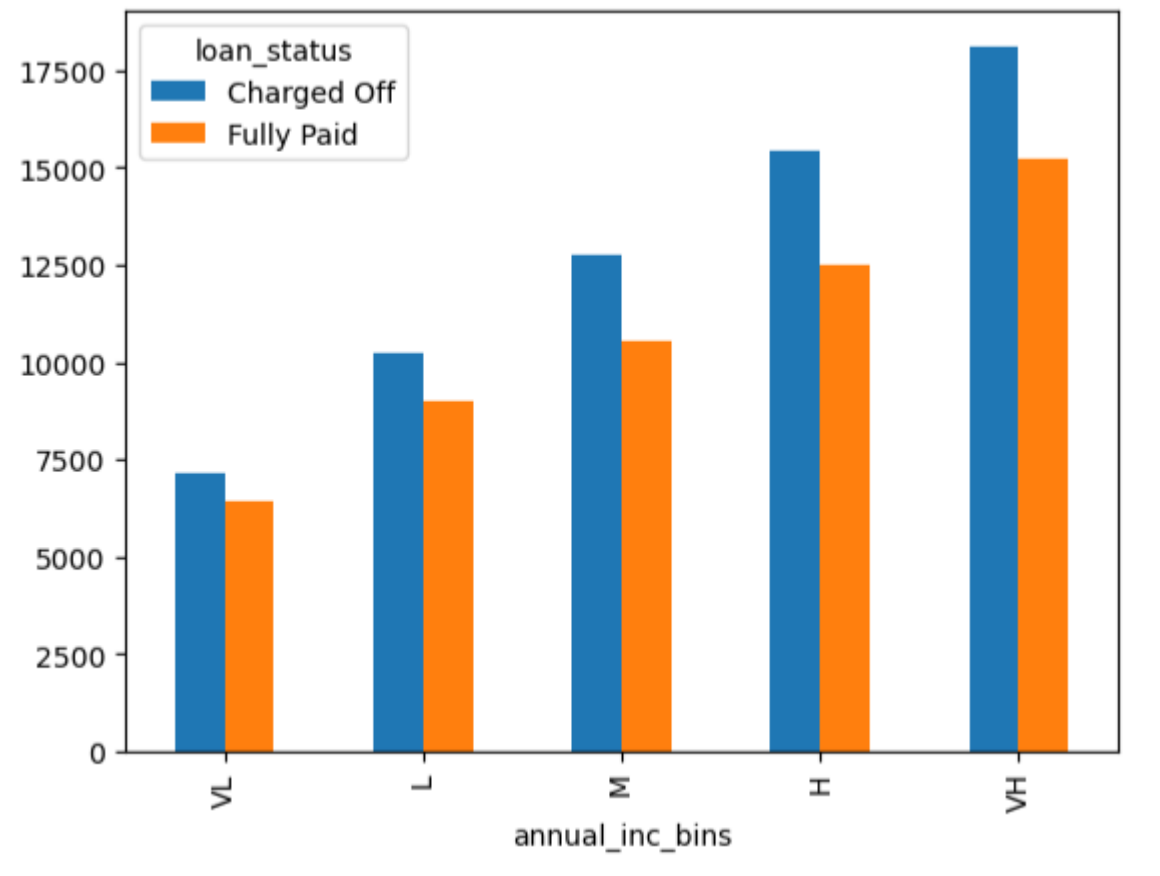
- Customers who opt for a 60-month loan term and later default tend to have borrowed larger amounts compared to those with a 36-month loan term.
- Borrowers with a credit grade (Grade F) who apply for larger loans are more likely to default.
- More experienced borrowers who request larger loan amounts are also at a higher risk of defaulting.
- Borrowers whose income has been verified and request larger loan amounts face a higher default risk.
- Grade G borrowers with higher interest rates are more likely to default.
- Borrowers with mortgages or those living in rental houses, with high interest rates and larger loan amounts, are more likely to default.
- Experienced individuals with higher monthly installments are more likely to default.
- Highly experienced borrowers with higher annual incomes are more likely to repay their loans.
- Borrowers who choose either 36 or 60 months of tenure and have more inquiries in the last 6 months are at a higher risk of defaulting.
- Grade C borrowers with more inquiries in the last 6 months are more likely to default.

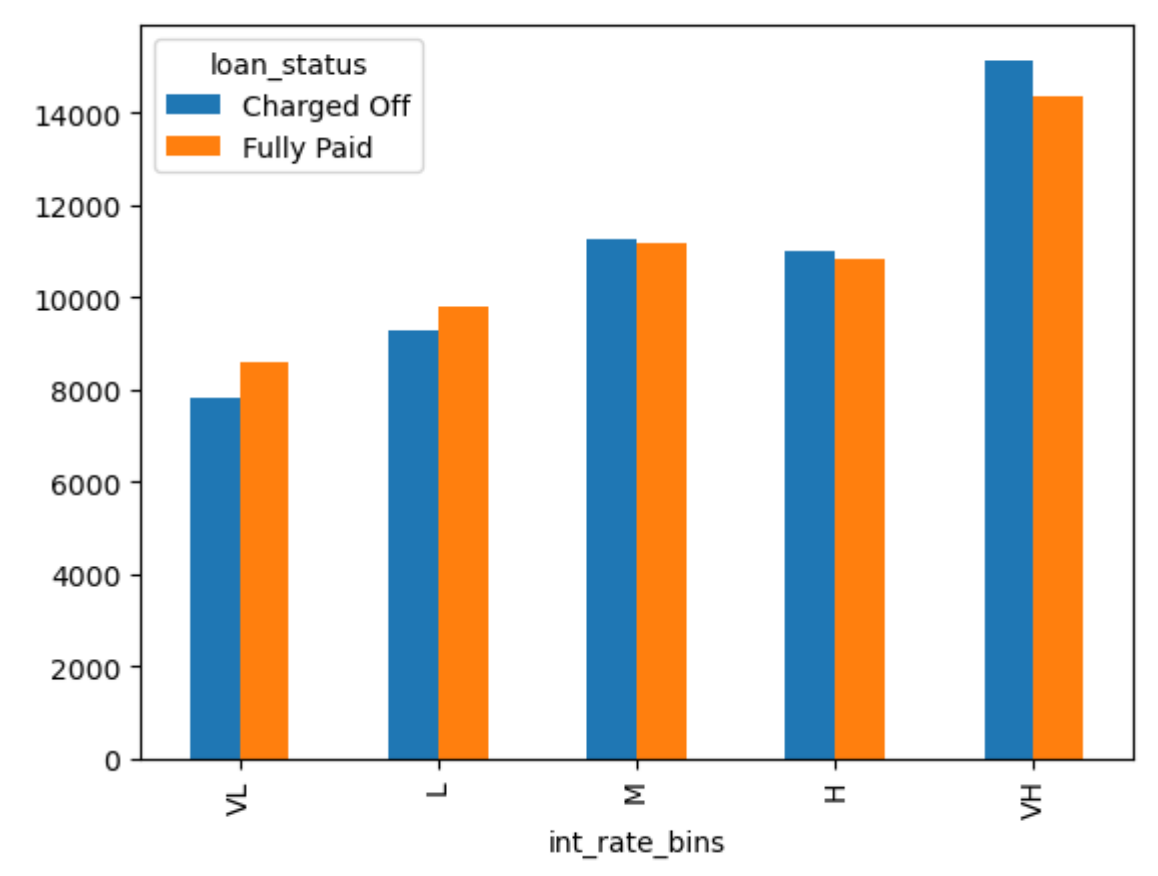
Observations - Categorical Vs Categorical

- Borrowers with 9 years of experience and more inquiries in the last 6 months are more likely to default.
- Those with "other" home ownership status and more inquiries in the last 6 months are more likely to default.
- Loans taken for housing purposes with more inquiries in the last 6 months are more likely to default.
- Grade G borrowers with more open credit accounts are more likely to default.
- Borrowers with mortgages and more open credit accounts are more likely to default.
- Loans taken for debt consolidation with more open credit accounts are more likely to default.
- Credit card users with higher revolving balance utilization are more likely to default.
- Longer tenure and a higher count of public record bankruptcies are indicators of potential defaults.
- Grade G borrowers with more public record bankruptcies are more likely to default.
- Highly experienced borrowers with more public record bankruptcies are more likely to default.
- Individuals with "other" home ownership status and more public record bankruptcies are more likely to default.
- Loans taken for housing purposes and more public record bankruptcies are more likely to default.

DERIVED METRICS

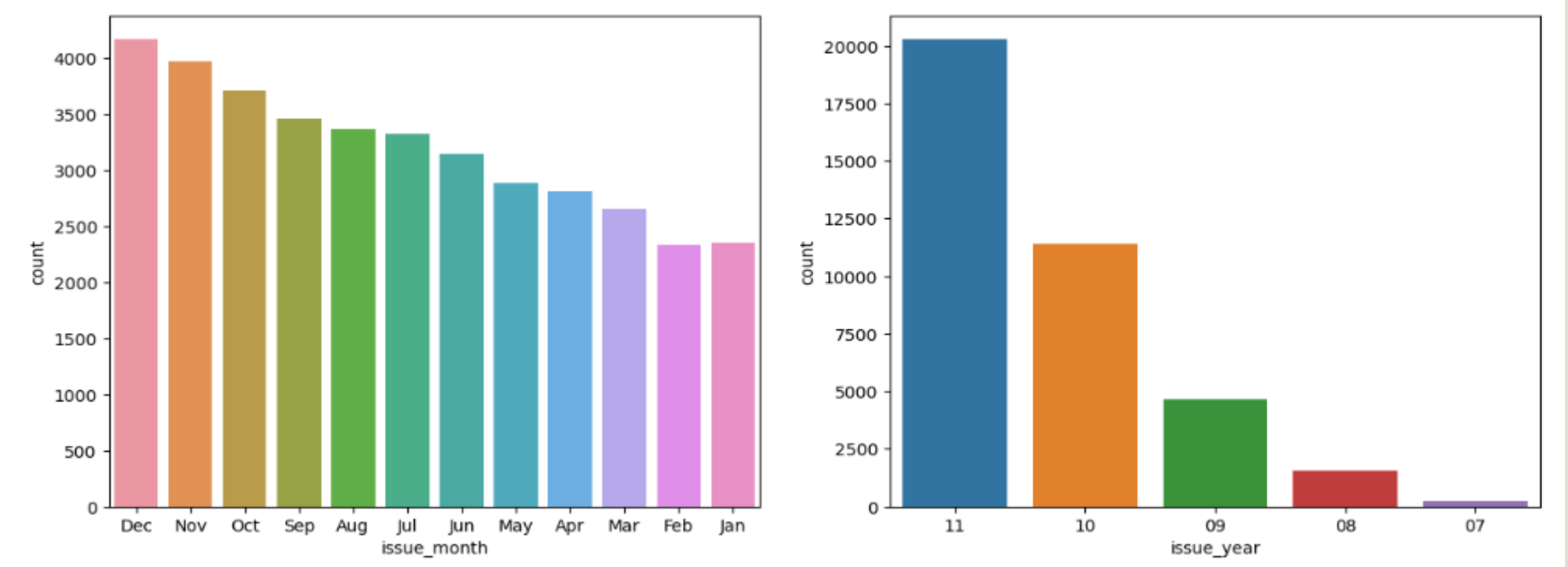
Derived Metrics Analysis





- Borrowers with higher incomes tend to request larger loan amounts, and they also have a higher likelihood of defaulting.
- Loans with very high amounts and very high interest rates are at a greater risk of defaulting.

Derived Metrics Analysis



- The majority of charged-off loans were issued in the month of December, with November being the second most common month for charged-off loans.
- Loans issued in the year 2011 experienced a high rate of charge-offs.

SUMMARY

Summary

After conducting a thorough data analysis, we have identified the crucial combinations to monitor closely

- Loan Amount in relation to Tenure, Credit Grades, Borrower Experience, Income Verification, and Interest Rates
- Annual Income concerning Home Ownership and Borrower Experience
- Credit Inquiries in connection with Credit Grades, Home Ownership, and Loan Purpose
- Number of Open Accounts associated with Loan Purpose
- Revolving Credit Utilization linked to Grades, Borrower Experience, Home Ownership, Loan Purpose, and Borrower's Address State
- Presence of Public Records of Bankruptcies with respect to Loan Tenure, Credit Grades, Borrower Experience, Home Ownership, and Loan Purpose

Our analysis has shown that the following attributes are the primary factors contributing to loan defaults

- Loan Amount, interest rates
- Annual Income
- Borrower Experience and Home ownership status
- Occurrences of Public Records and Bankruptcy History
- Recent Credit Inquiries made in the Previous Six Months
- Issue Month