LENDING CLUB CASE STUDY

Table of Contents

- ✓Problem Statement
- ✓Why this is required
- ✓Solution to the problem

Presented by –

Abhishek Pandey & Phani Sharma

PROBLEM STATEMENT

The data given contains information about past loan applicants and whether they 'defaulted' or not. 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

Approach

Analyzing past loan applicant data to identify default patterns, enabling informed decisions on loan approvals, amounts, and interest rates for risk management.

STEPS TO SOLVE THE PROBLEM

- Read the data from the .csv file
- Identify the null columns and rows
- Treat the null values with relevant information
- Outlier treatment
- Understanding of categorical variables
- Univariate Analysis
- Bivariate Analysis
- Conclusion

READING THE DATA FROM THE .CSV FILE

```python
## import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


#Importing the dataset
ld = pd.read_csv(r"C:\Users\abhis\Documents\IIITB\Class1\lending loan case study\loan.csv",low_memory=False)
ld.head()
```

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dp
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	
3	1076863	1277178	10000	10000	10000.0	36	13.49%	339.31	C	C1		

IDENTIFY THE NULL COLUMNS AND ROWS

```
# Getting the number of rows and columns
ld.shape
```

```
(39717, 111)
```

```
# Missing Value Check

100*ld.isnull().sum()/ld.shape[0]
```

```
id                           0.000000
member_id                    0.000000
loan_amnt                    0.000000
funded_amnt                  0.000000
funded_amnt_inv              0.000000
                               ...
tax_liens                    0.098195
tot_hi_cred_lim            100.000000
total_bal_ex_mort          100.000000
total_bc_limit             100.000000
total_il_high_credit_limit 100.000000
Length: 111, dtype: float64
```

- **In the above we can see lot of variables with missing values which we can't keep in our analysis hence we are going to remove them**

TREAT THE NULL VALUES WITH RELEVANT INFORMATION

- Discard columns with >40% or 50% missing values.

```
# First we need to identify the number of columns which are having the missig values

ld_clean = ld.dropna(axis=1, how='all')
```

```
# Again checking the missing values to see how many more columns have the missing values

100*ld_clean.isnull().mean()
```

- Impute missing values in columns within an acceptable range.

```
# Since now we have now relevant variable lets fill the missing values with the relevant information

loan_data2.emp_length.mode()[0]
```

```
'10+ years'
```

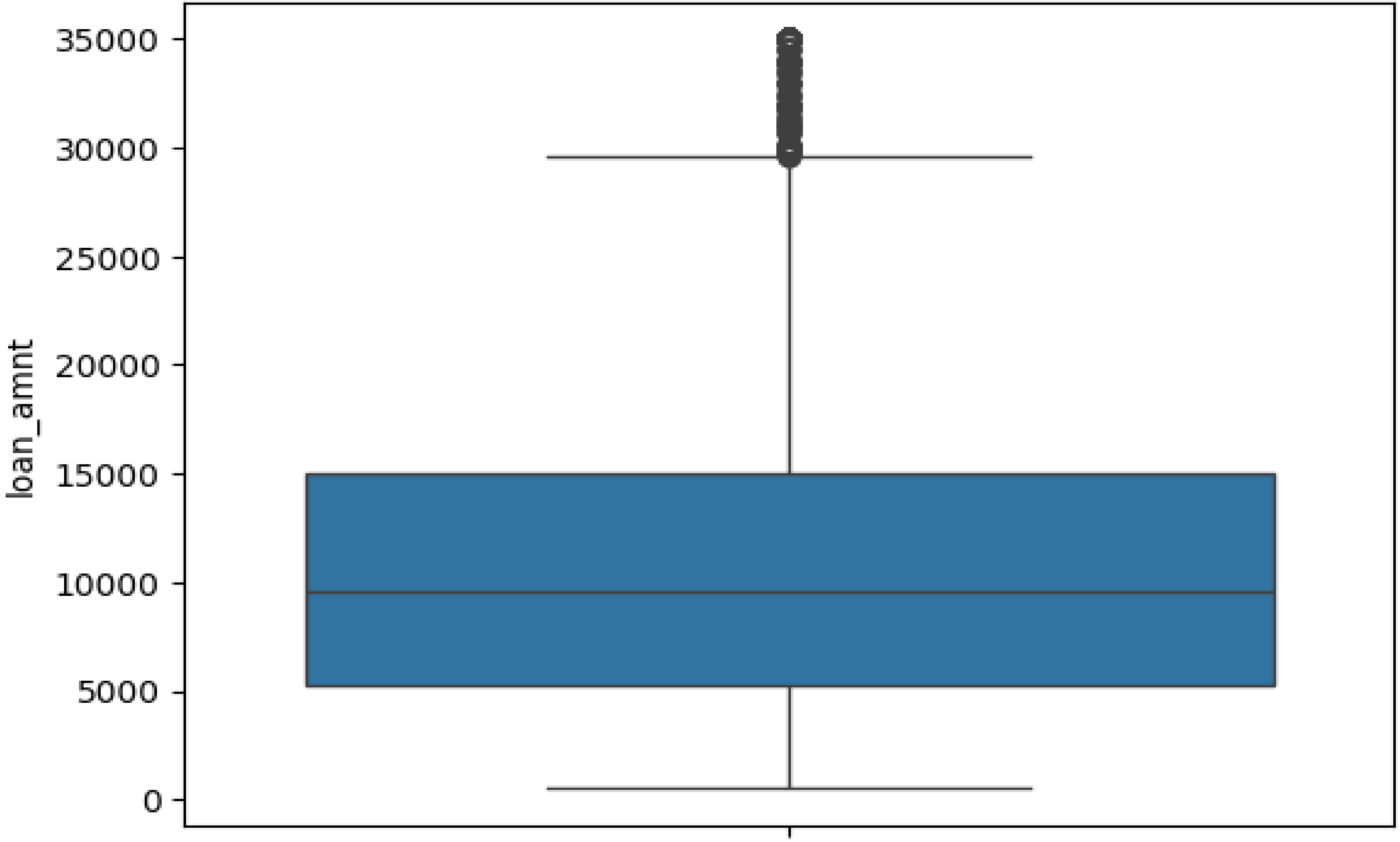
```
mod = loan_data2.emp_length.mode()[0]

loan_data2["emp_length"]=loan_data2["emp_length"].fillna(mod)
```

OUTLIER TREATMENT

Loan Amount Analysis

- After cleaning missing values and current loan status
- Observed outliers in the loan amount range of 30,000 - 35,000

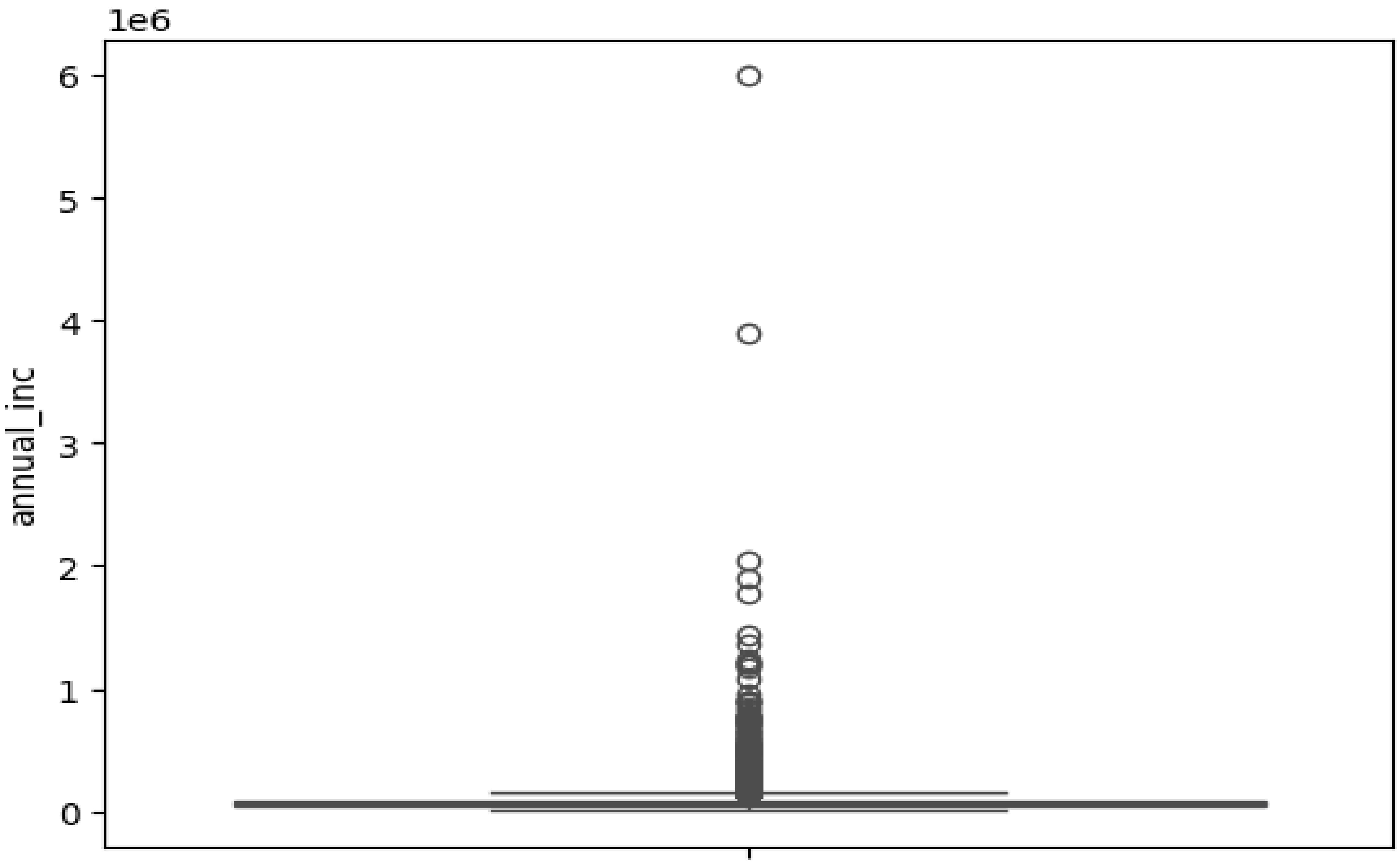


OUTLIER TREATMENT

Annual Income Analysis

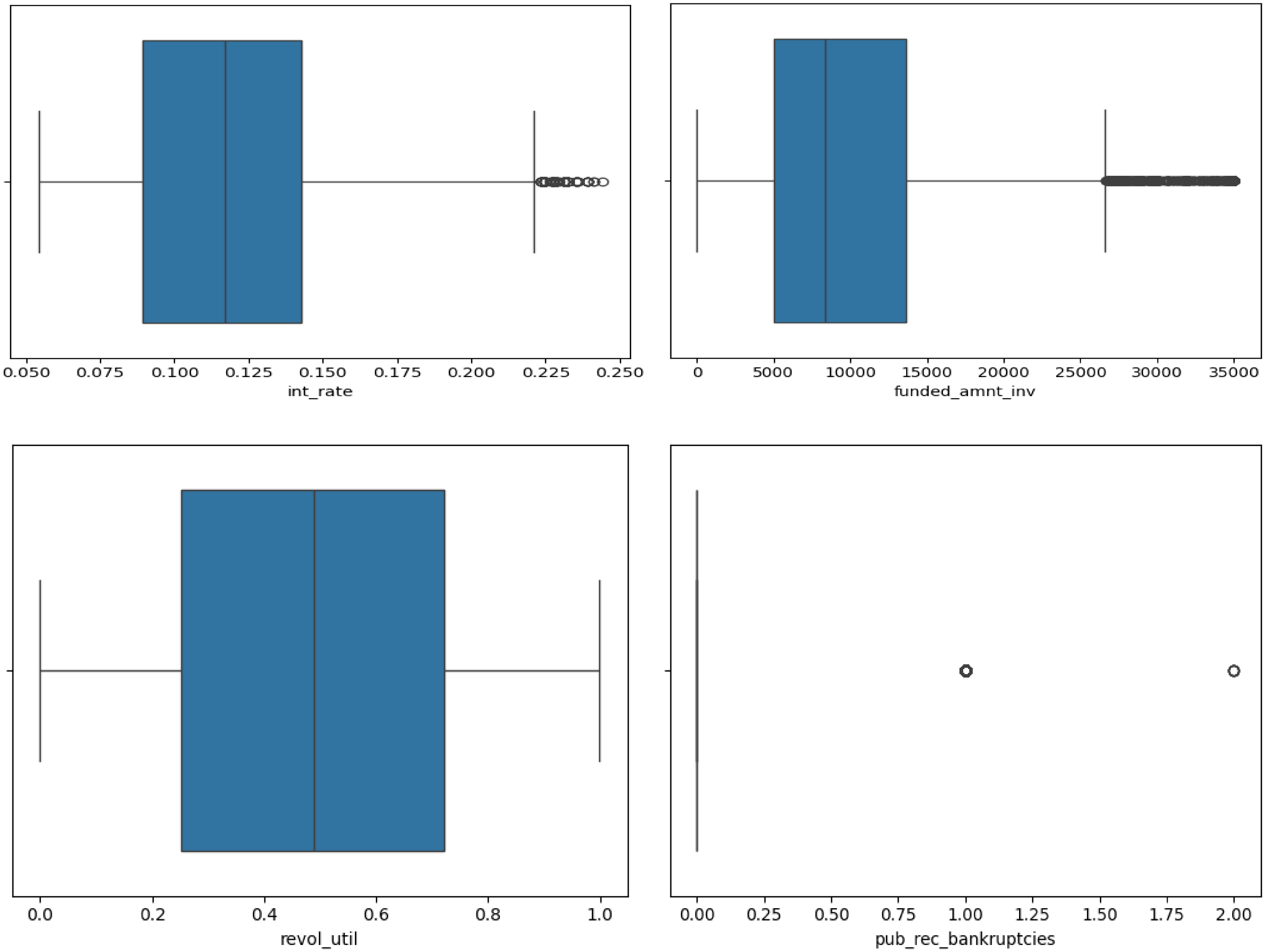
- Upon examining the plot, it becomes evident that there are outliers present in the data.

0.50	58868.0
0.75	82000.0
0.90	115000.0
0.95	140004.0
0.97	165000.0
0.98	187000.0
0.99	234144.0



The values beyond the 95th percentile appear to deviate significantly from the overall distribution. Hence, we will remove them from analysis since they are outliers in data

OUTLIER TREATMENT



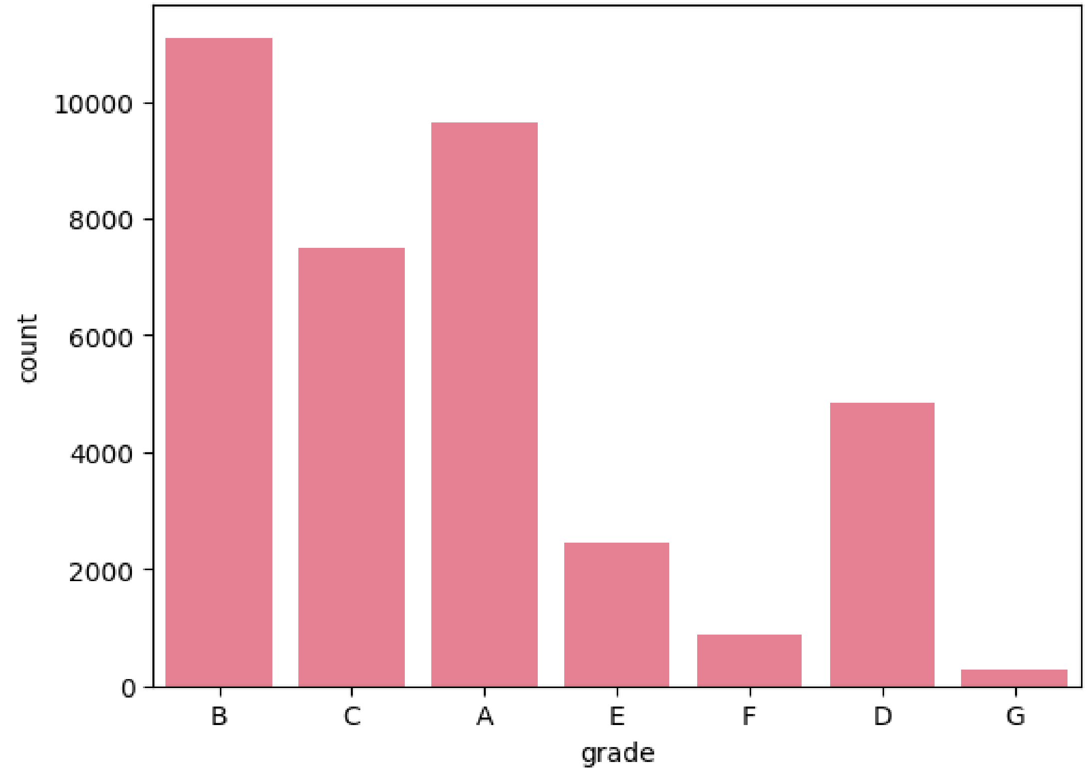
Here we have all numeric variables plotted to understand their distribution and outliers

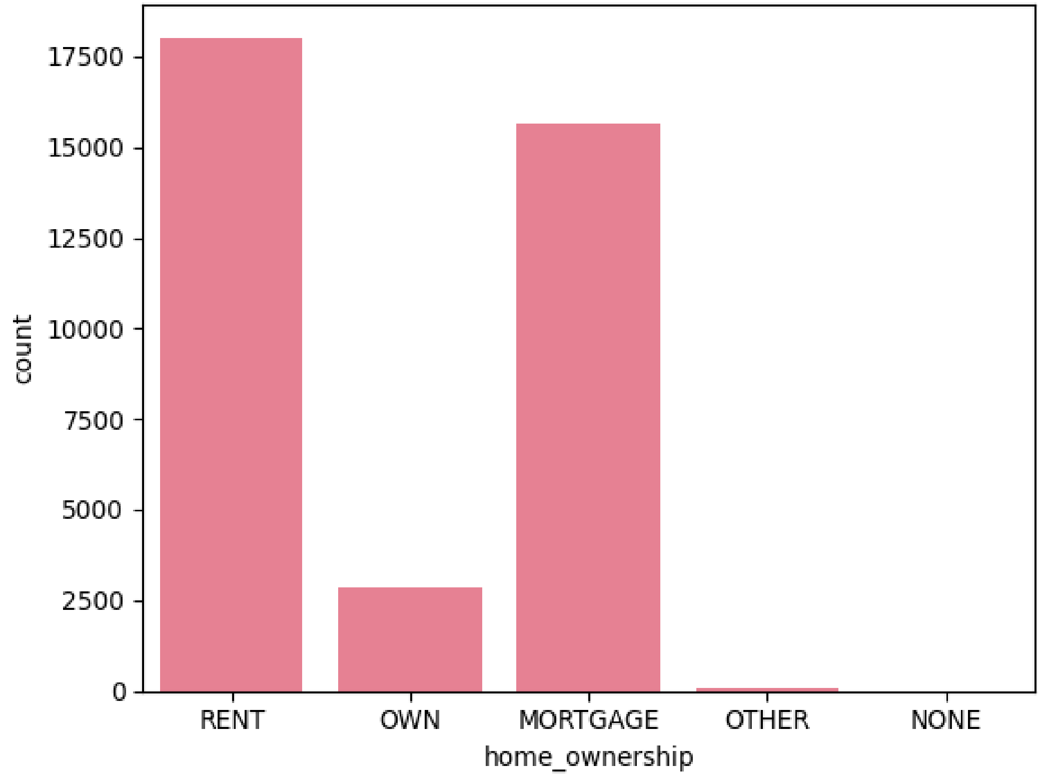
- ✓ It is evident that all numerical variables appear to be within expected ranges, with the exception of 'pub_rec_bankruptcies,' where the majority of values cluster around 0. Hence, we will remove

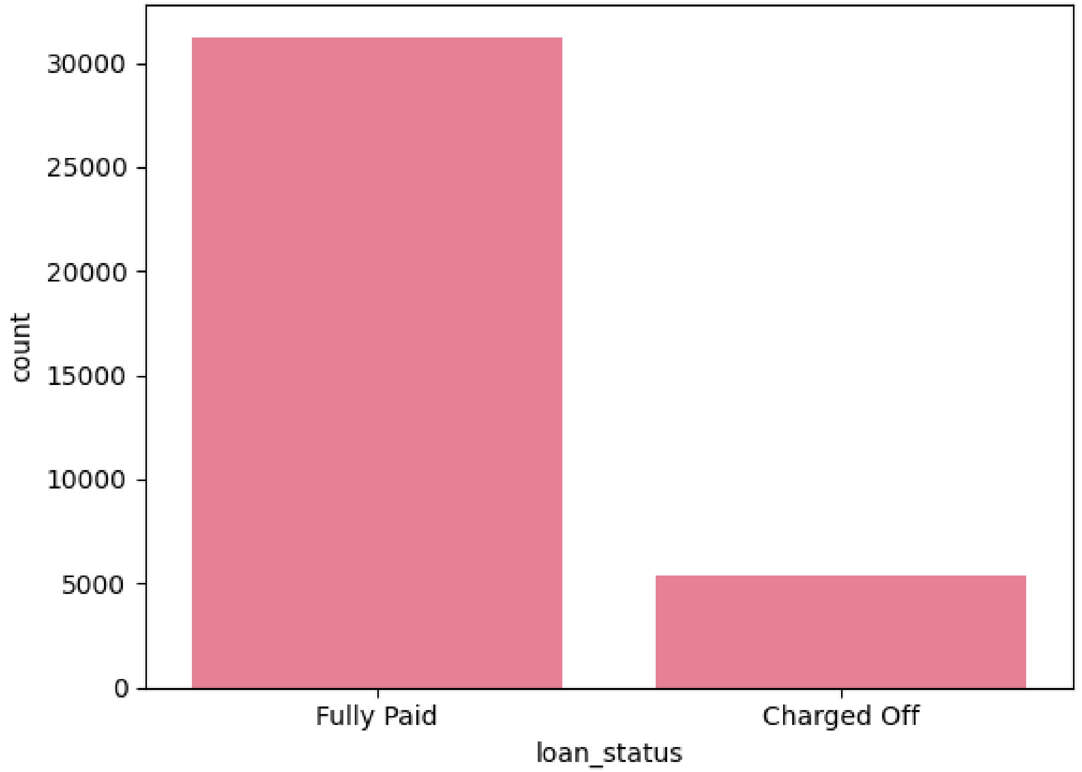
```
0.500    0.0
0.750    0.0
0.900    0.0
0.950    0.0
0.970    1.0
0.975    1.0
0.980    1.0
0.985    1.0
0.990    1.0
1.000    2.0
Name: pub_rec_bankruptcies, dtype: float64
```

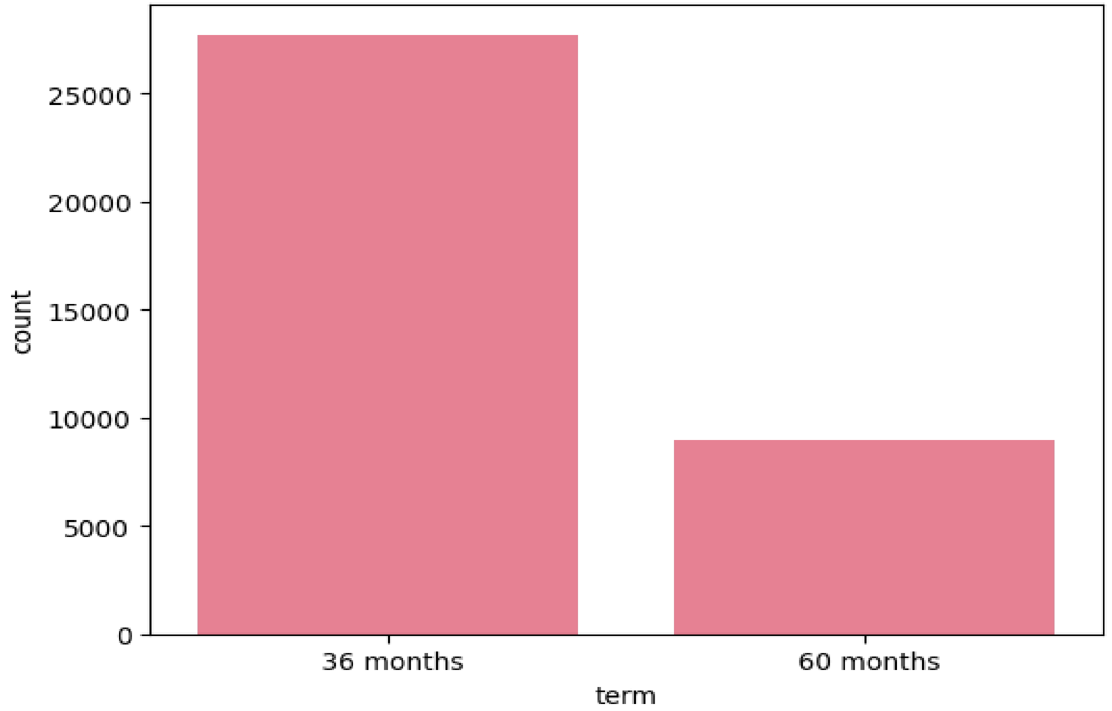
UNDERSTANDING THE CATEGORICAL VARIABLES

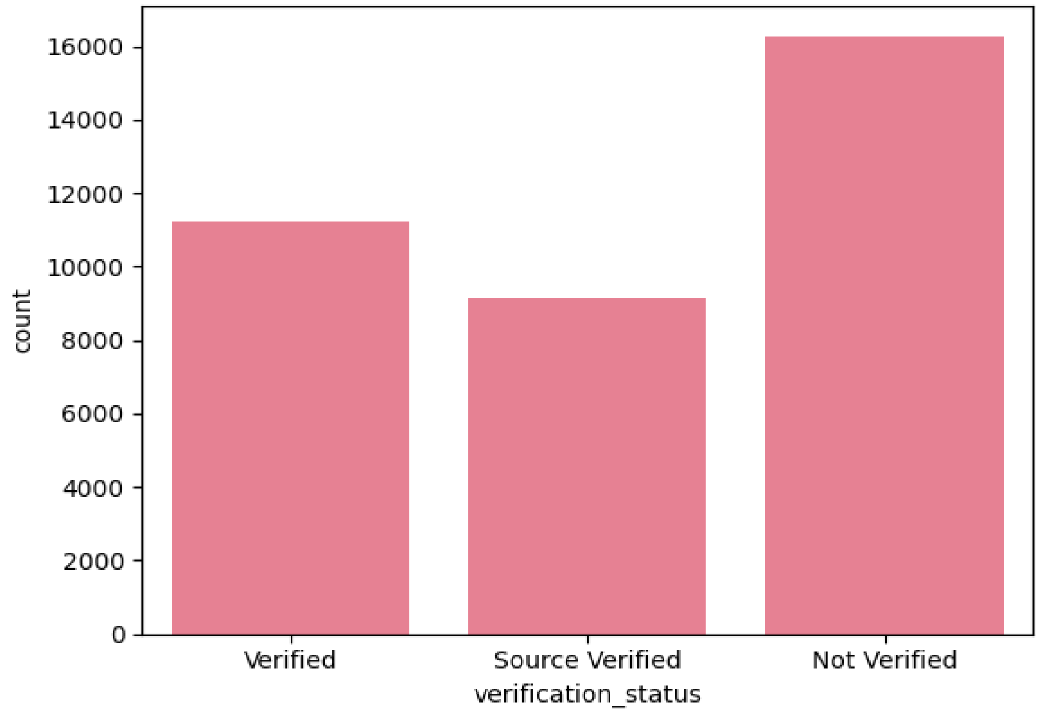
We have plotted all categorical variables and found subgrade is values needs to treated so that we can present it plot

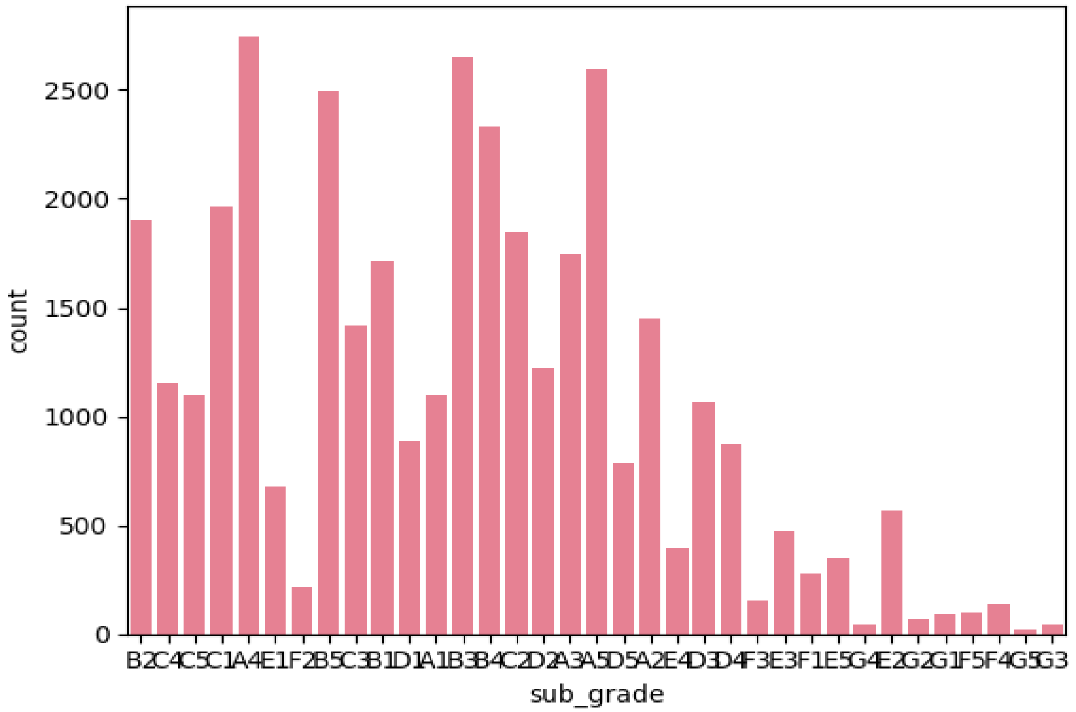








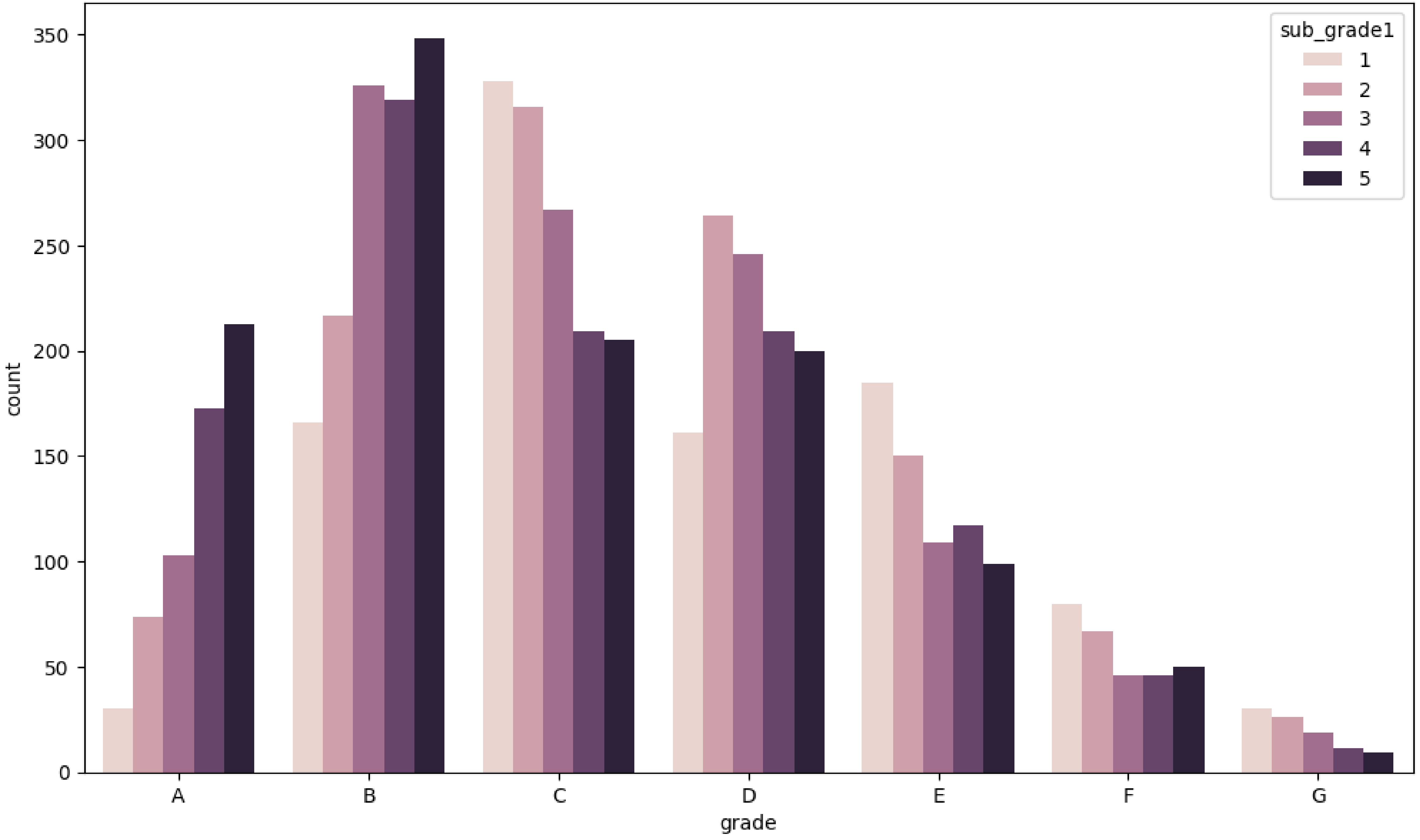




UNIVARIATE ANALYSIS

Subgrade Analysis

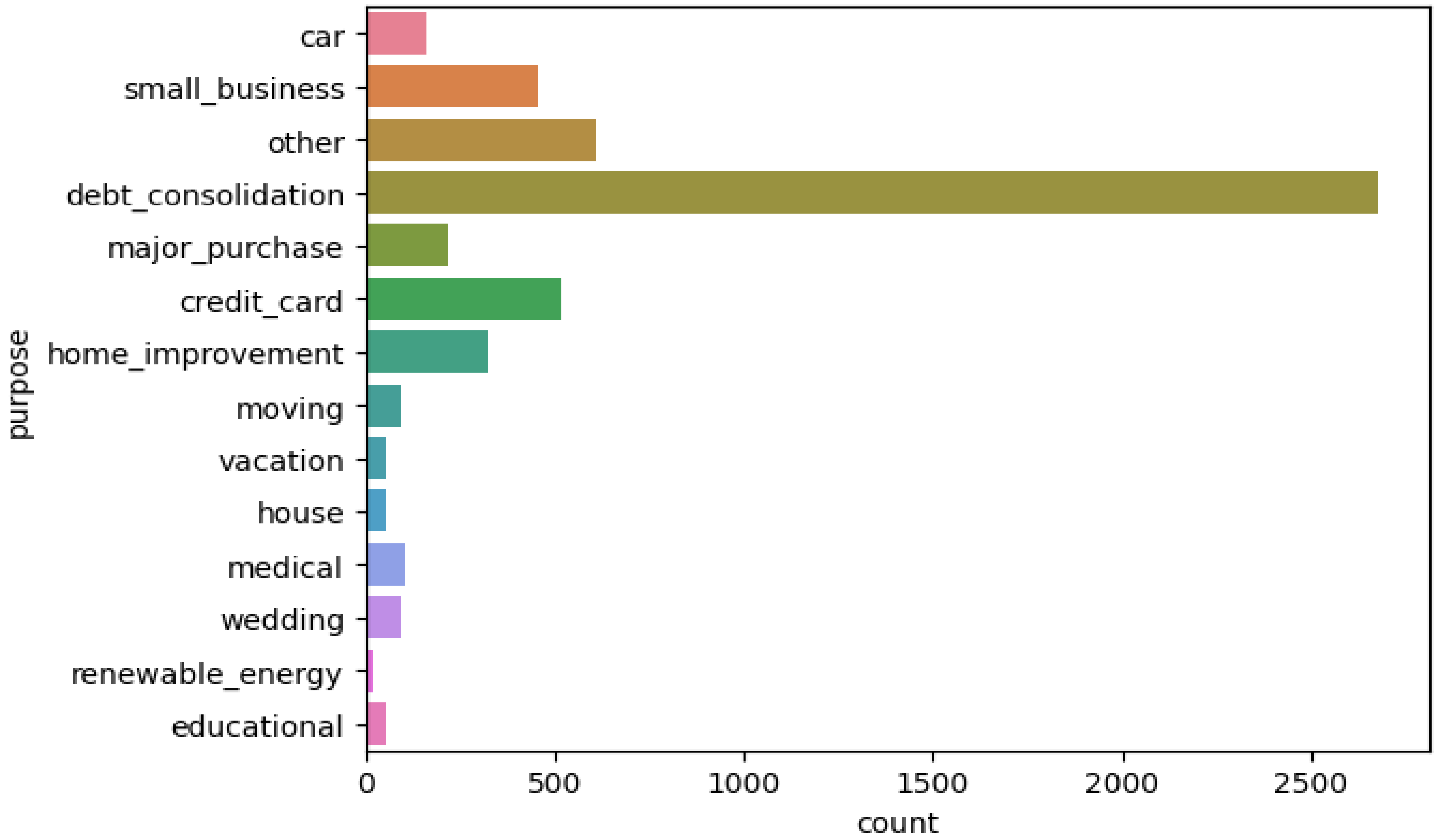
- Applicants with a grade of B and subgrade of 5 are more likely to default.



UNIVARIATE ANALYSIS

Loan Purposes Analysis

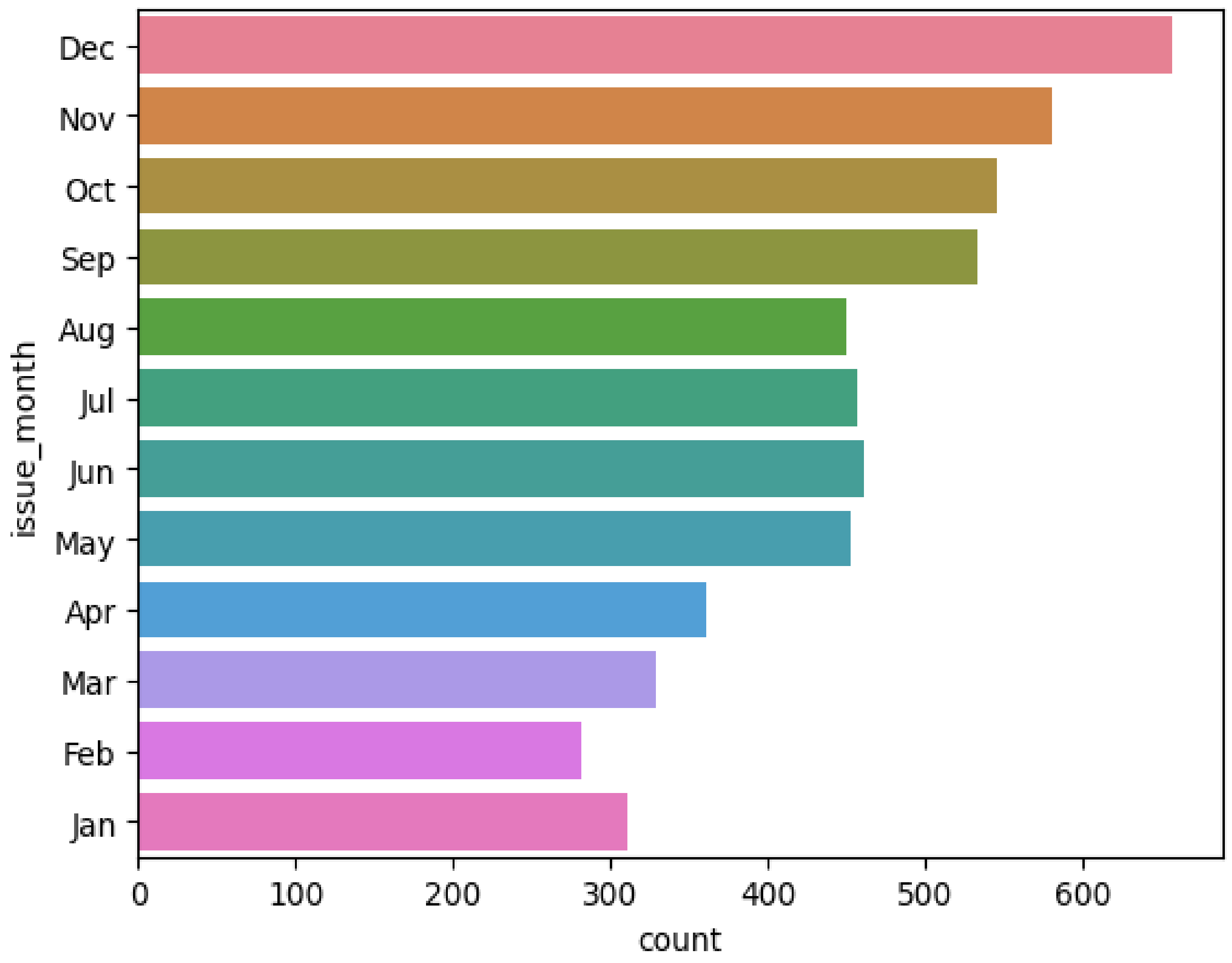
- Debt consolidation has the highest number of credit models.

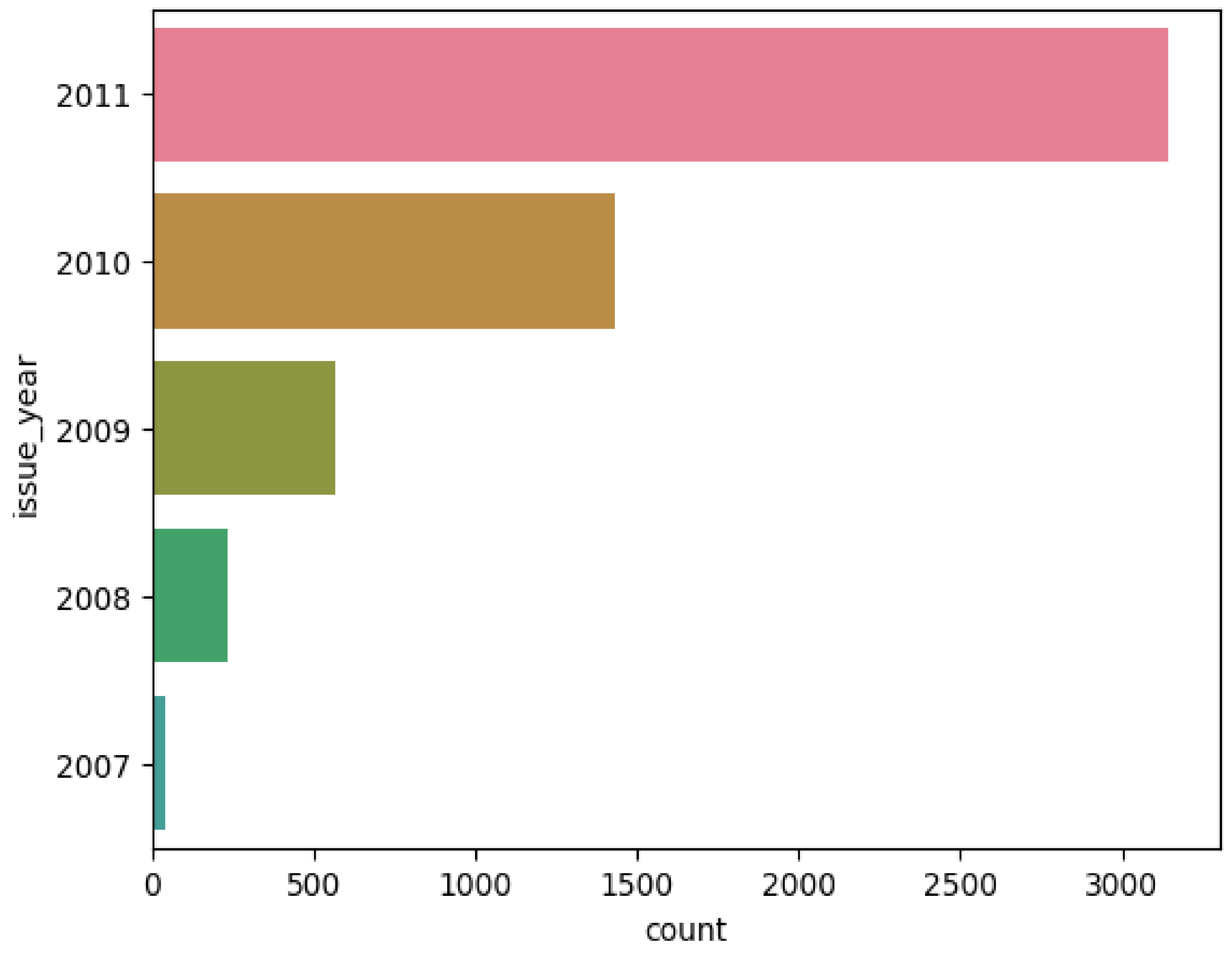


UNIVARIATE ANALYSIS

Analysis of Applicants Based On Year and Month

- Applicants from **December 2011** are likely to default

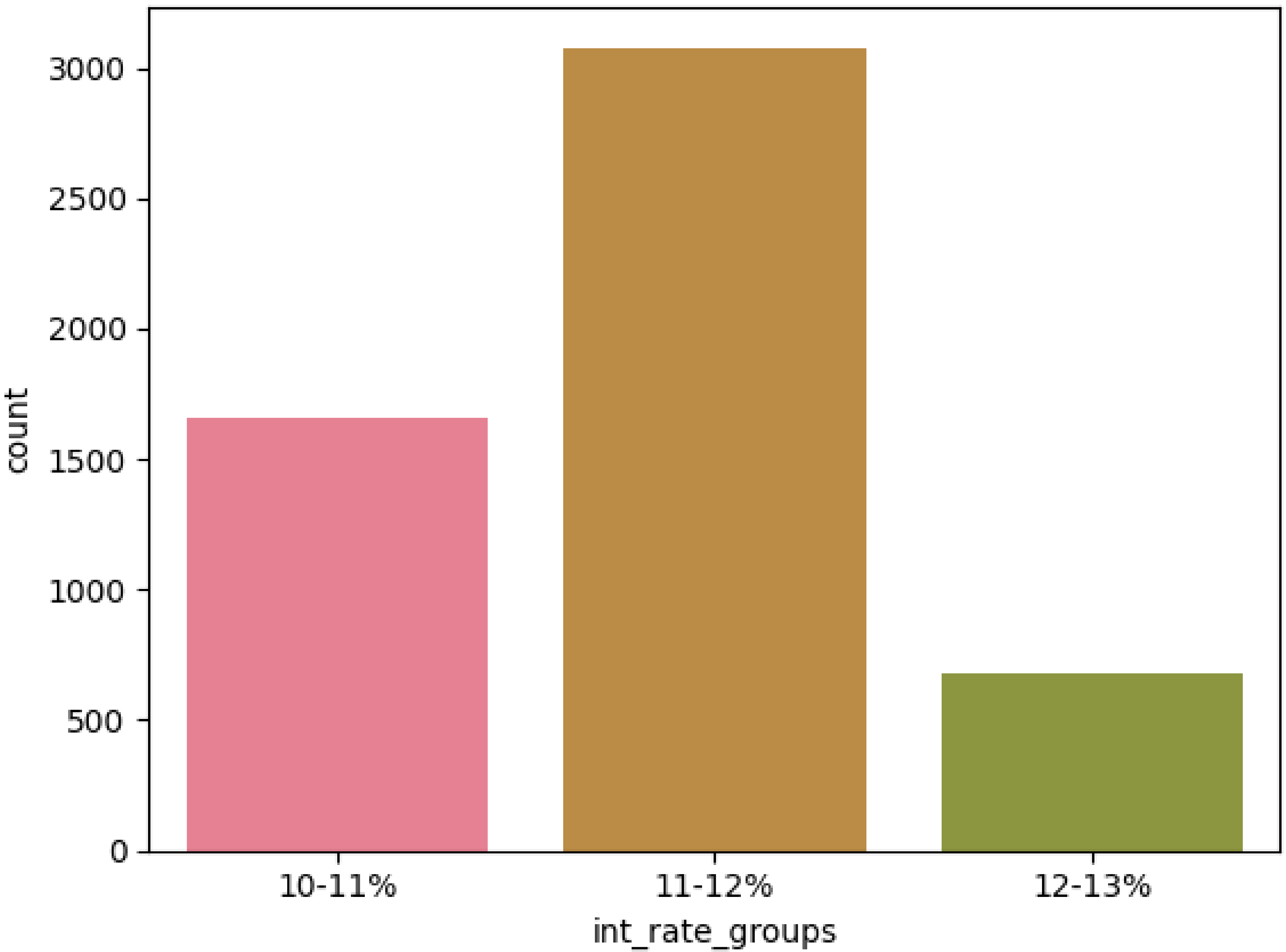




UNIVARIATE ANALYSIS

Applicant Default Probability Up On Interest Rates

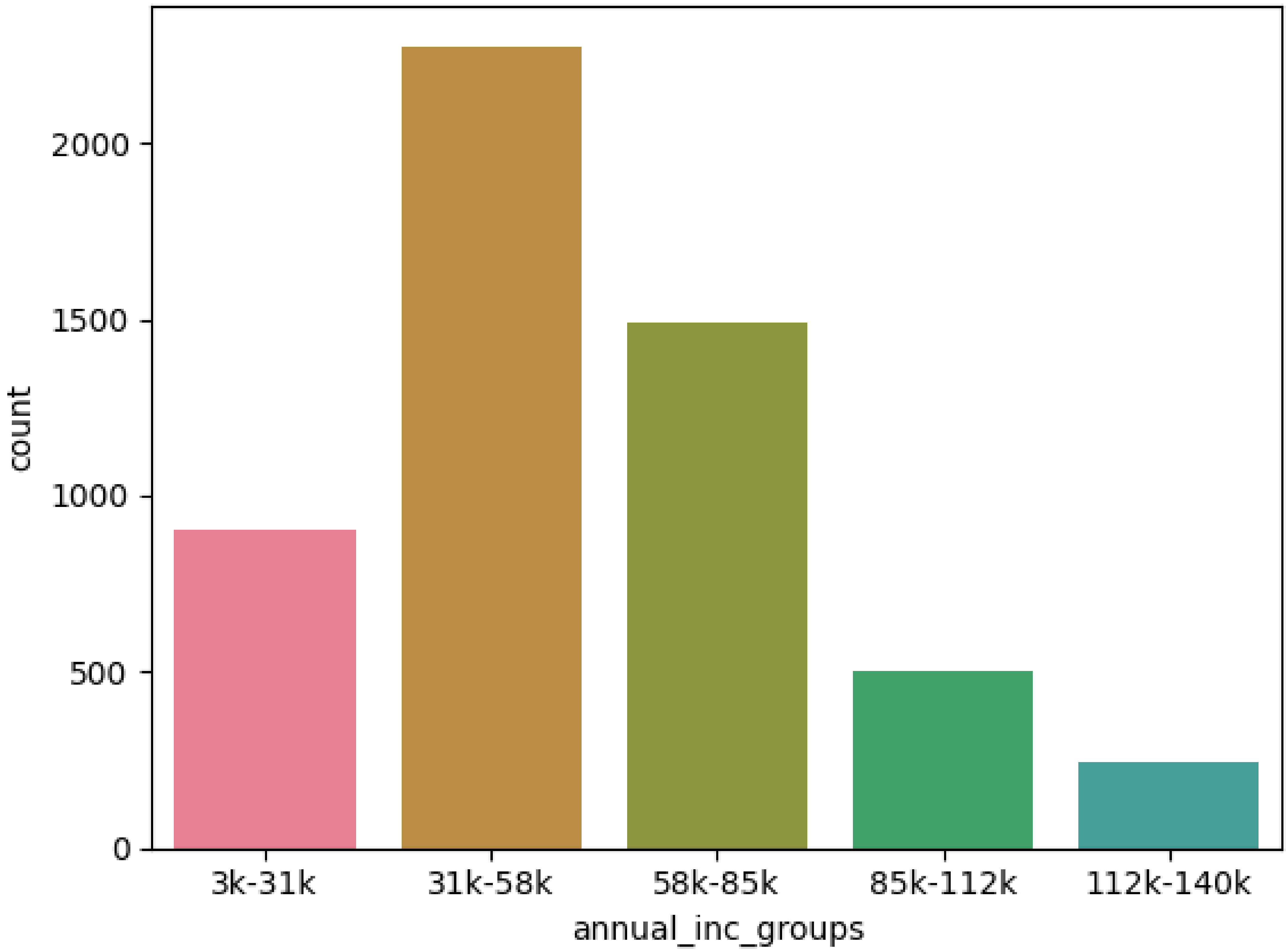
- Applicants with interest rates between 11% to 12% are more likely to default.



UNIVARIATE ANALYSIS

Annual Income Group Analysis

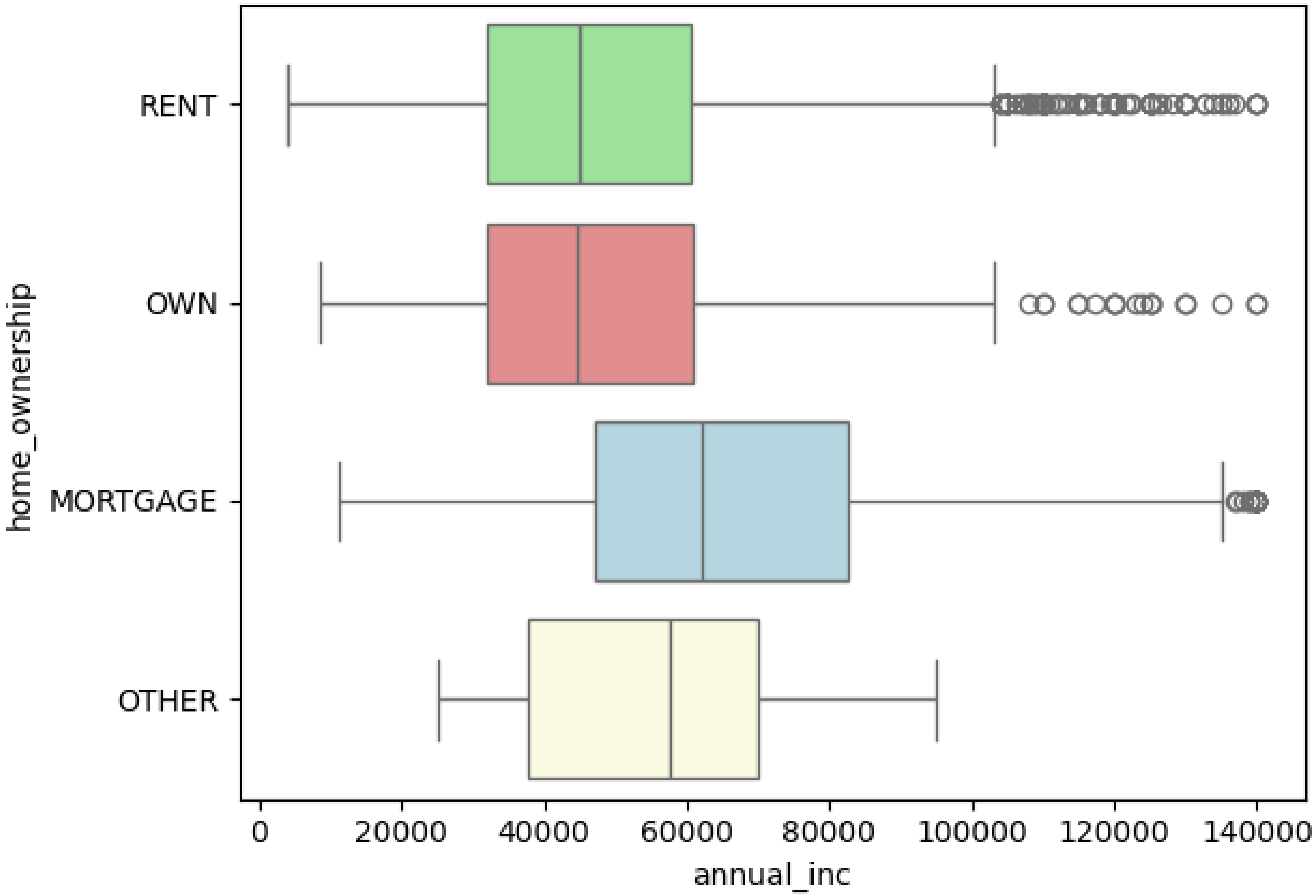
- Focus on applicants with income between 31 to 58



BIVARIATE ANALYSIS

Applicant Financial Overview

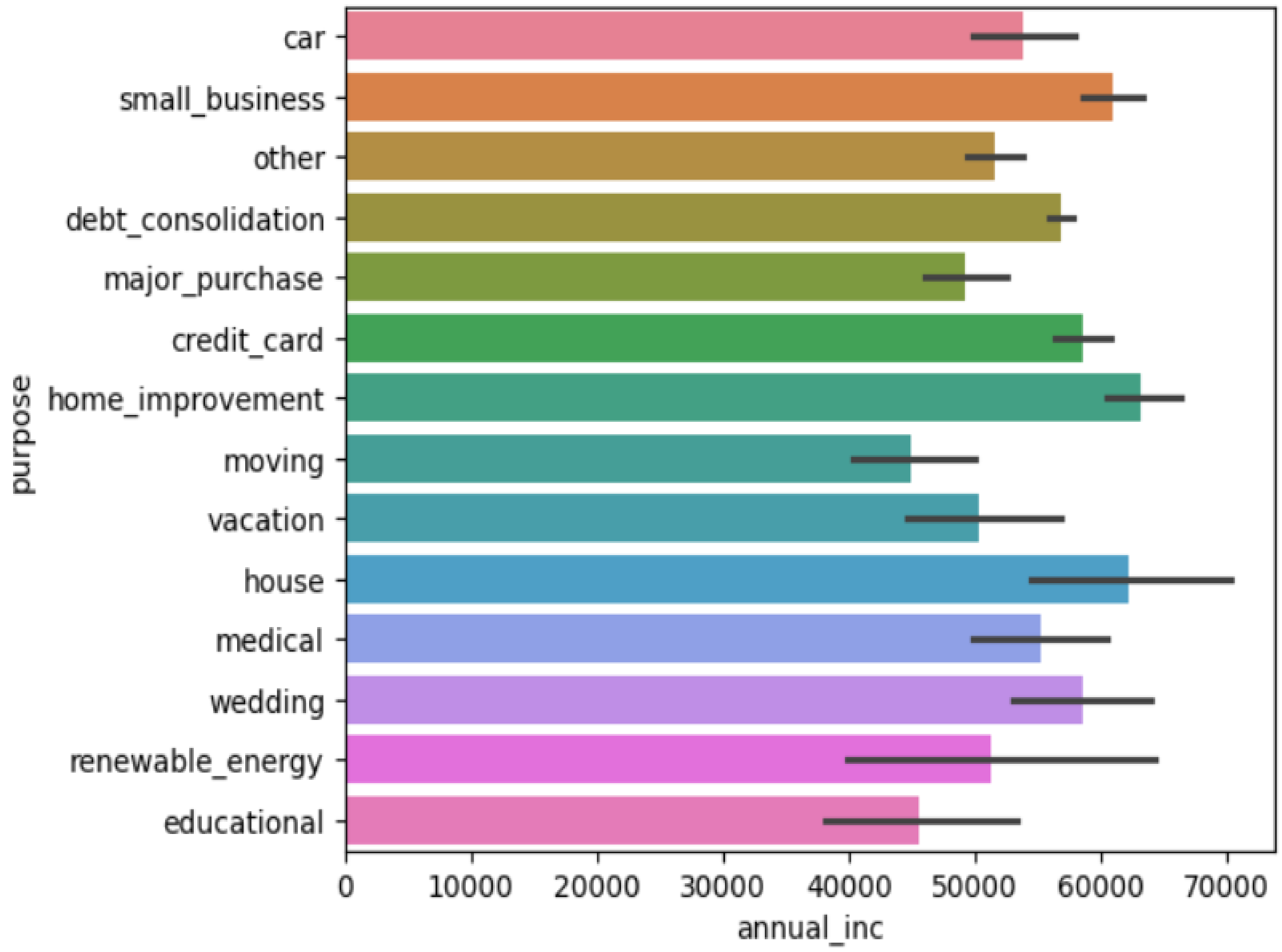
- High annual income, median around $60K
- Likely have a mortgage or other house ownership



BIVARIATE ANALYSIS

Applicant Preferences

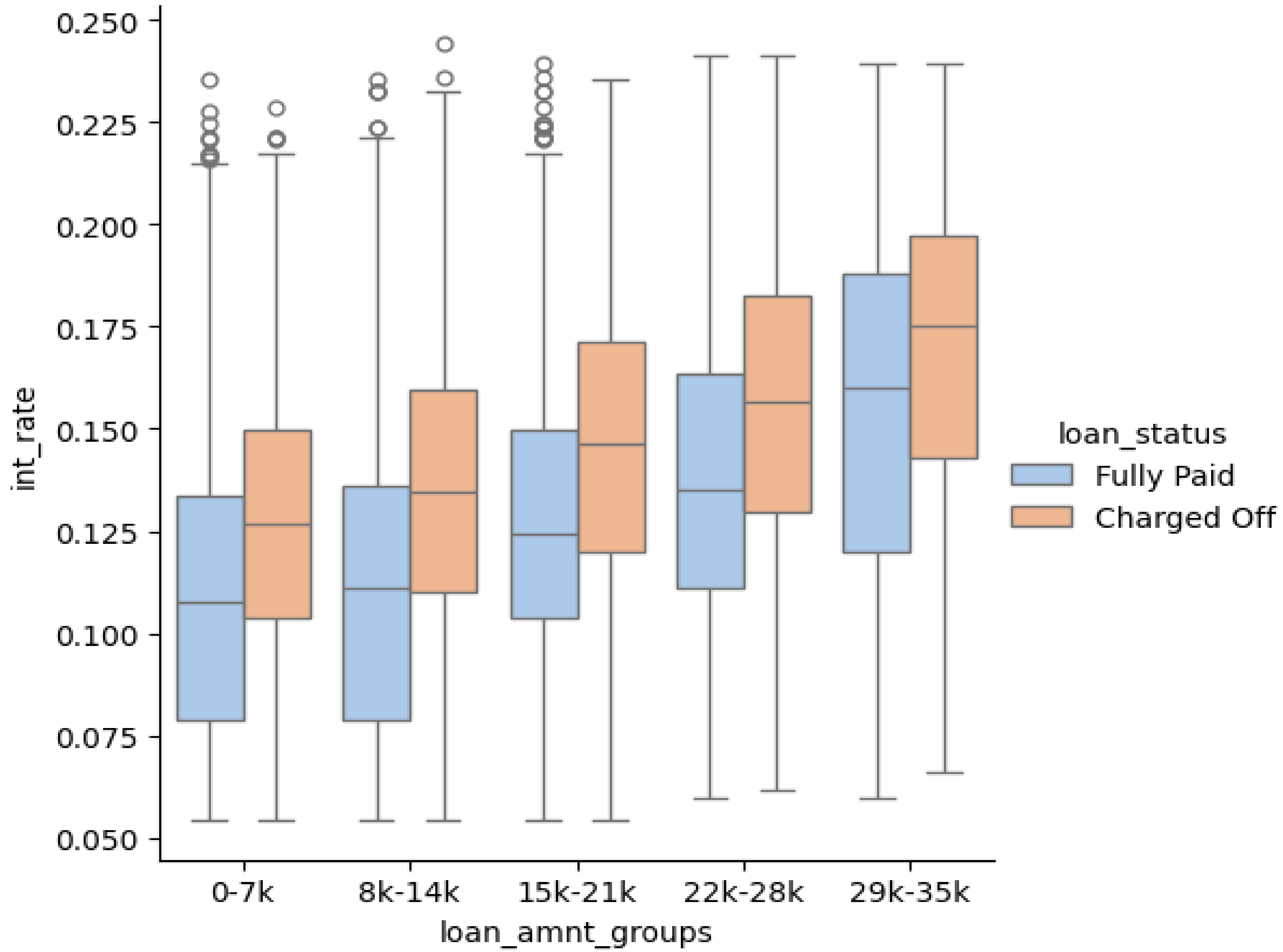
- **High-income** applicants typically apply for:
  - Home improvement loans
  - House loans
  - Small business loans are **likely to get defaulted**



BIVARIATE ANALYSIS

High Interest Rate Impact on Loan Defaults

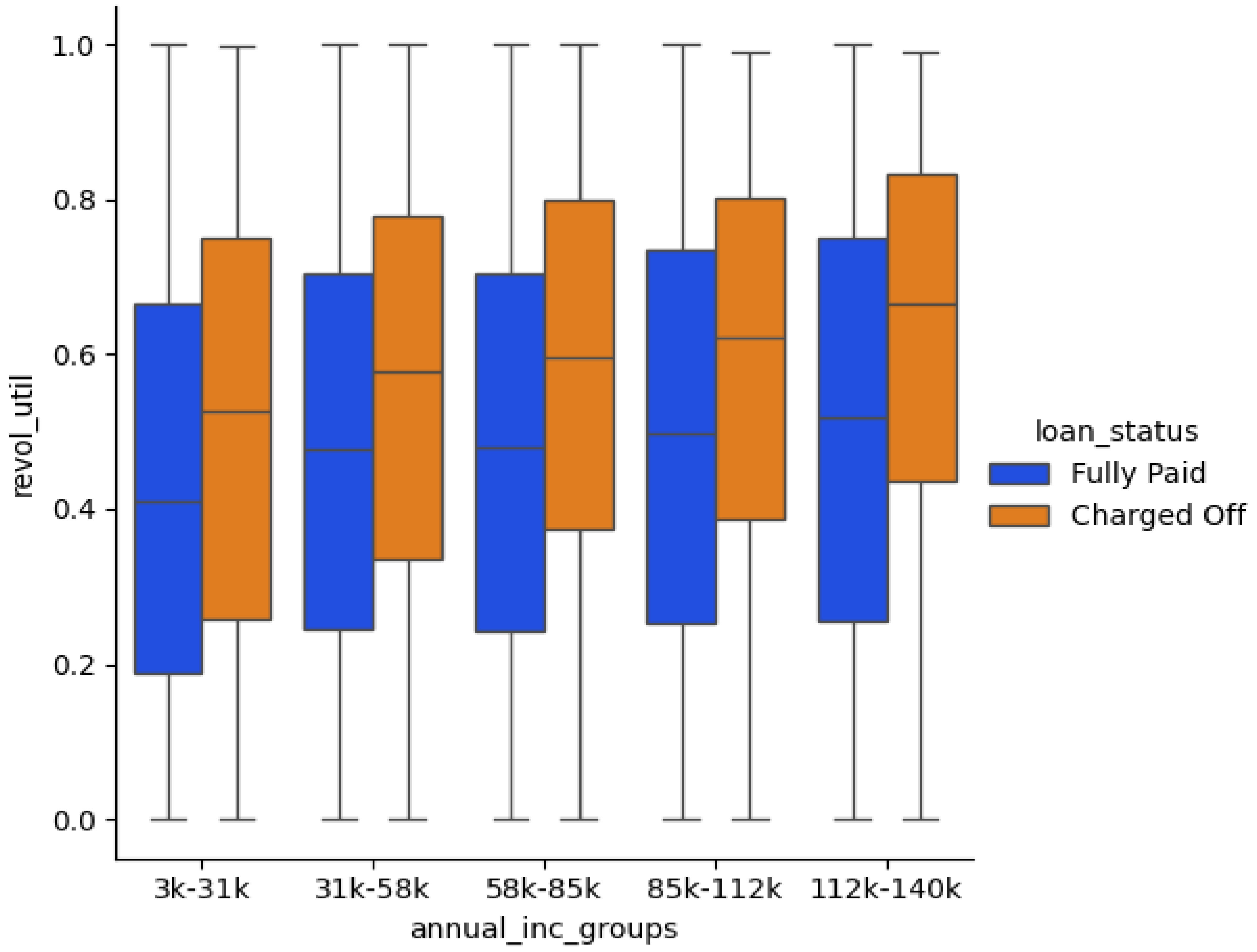
- Applicants with higher interest rates are more likely to default.
- Fully paid loan applicants experience lower default rates.



BIVARIATE ANALYSIS

Revolve Utilization and Annual Income

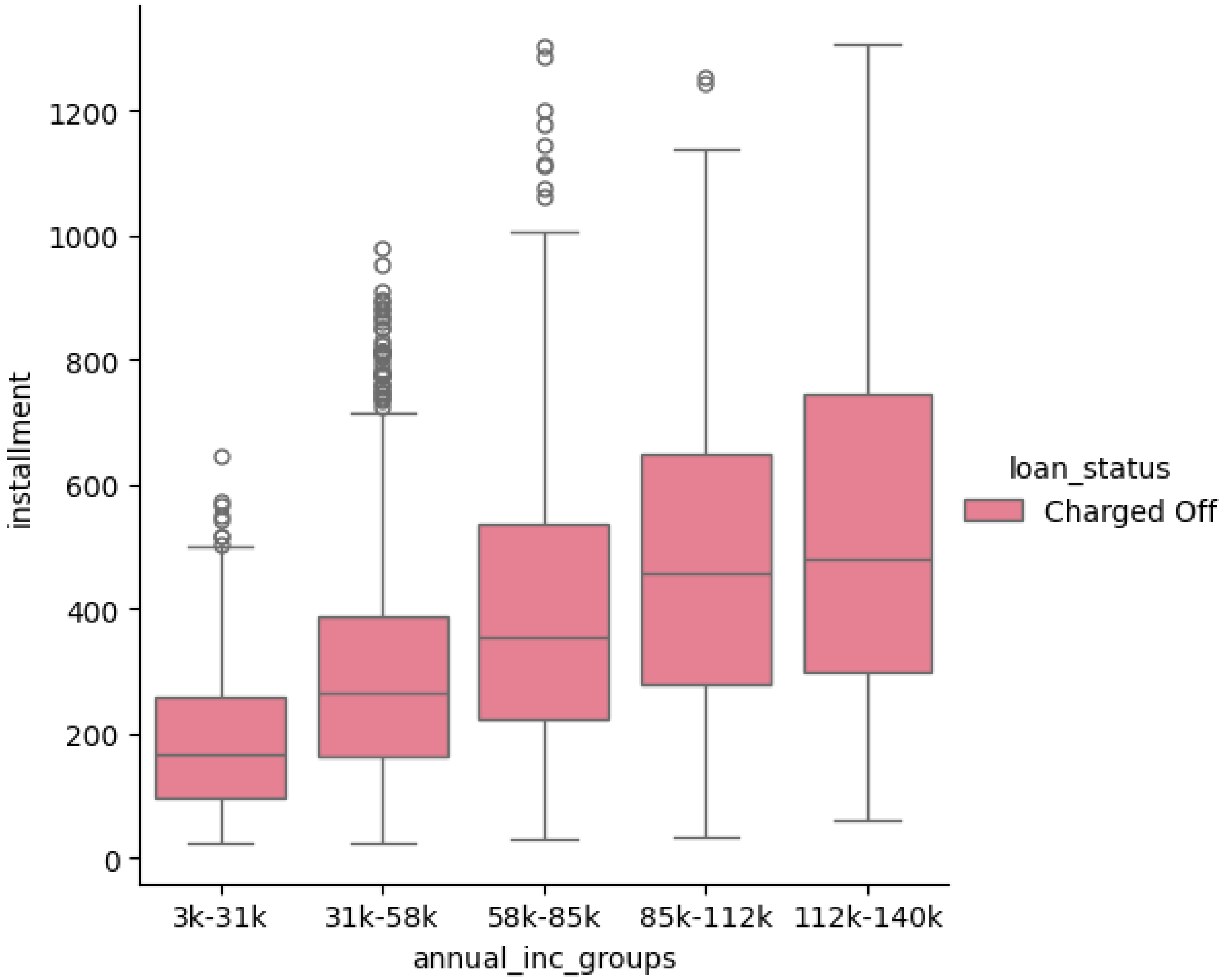
- Higher revolve utilization correlates with a higher tendency to default
- Applicants who revolve their amounts more frequently are more likely to default



BIVARIATE ANALYSIS

Annual Income vs Installment Graph

- Higher EMIs might lead to loan defaults.



CONCLUSION

Upon examining both univariate and bivariate data, specific trends emerge that suggest a higher likelihood of loan defaults.

Univariate Analysis:

- Applicants who use the loan to clear other debts
- Applicants who have taken short term loan that is around 36
- Applicants whose grade is B and subgrade is 5
- Applicants who have applied in 2011 and month Dec
- Applicants who have credit line between 1980 to 1999
- Applicants whose interest rate is between 11% to 12%
- Applicants who have just 2- 10 open credit lines
- Applicants who have income between 31 to 58
- Applicants whose income source is not verified

Bivariate Analysis

- Applicant have high annual income since median is around 60K either have Mortgage or other house ownership
- Applicant having the high income they usually apply for home improvement ,house or small business
- Applicant who is having the higher annual income and getting high loan amount
- Applicant whose rate of interest is very high as compared to the interest of the applicants who have fully paid the loan
- Applicant who revolve the spend amount more
- Applicant who is having the higher installment amount

These insights provide lenders with valuable information to identify high-risk applicants and develop strategies to reduce potential financial losses.