Lending Club Case Study assignment

Prashant Kumar Misra
Pranav Kavishwar

Introduction

Lending club case study deals with a consumer finance company specialized in providing loans to urban customers based on requirement. The company has to make a decision of whether to sanction the loan or not based on the profile on the applicant. We performed EDA (Exploratory Data Analysis) on the dataset provided.

Analysis:-

1. Analyzing the dataset and performed data cleaning by removing unwanted columns and null rows.
2. Perform univariate and bivariate analysis on the data columns and draw insights from it.
3. Use data visualization to display few boxplots, graphs etc. for pictorial view and better understanding of the data.

Data Cleaning

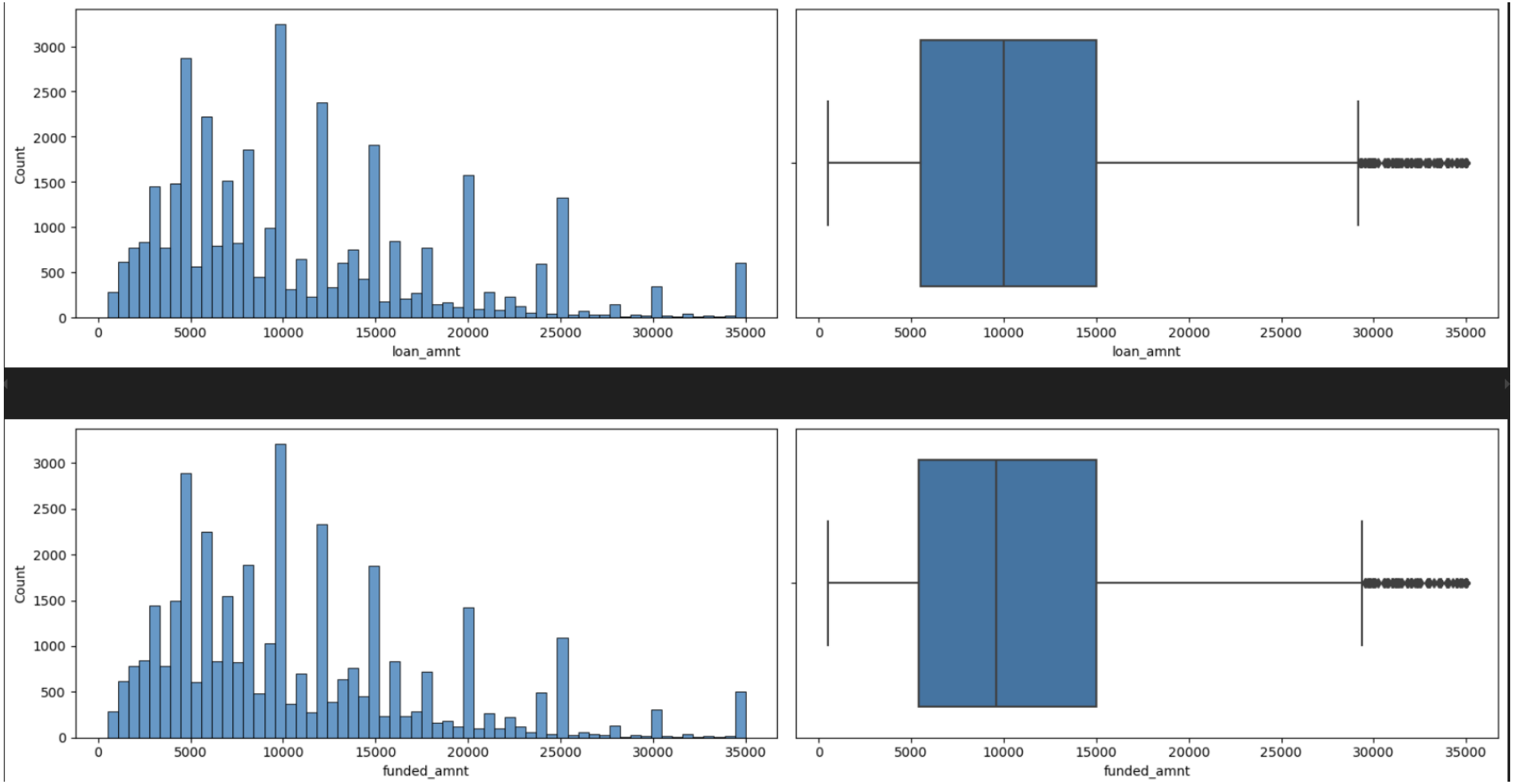
The initial process was to remove unwanted data from the dataset and make it in an interpretable format. As a part of data cleaning activity, we removed columns with a significant number of null values. Later, we removed those columns with only one single value.

Also, we classified the data columns based on either numeric or categorical variables.

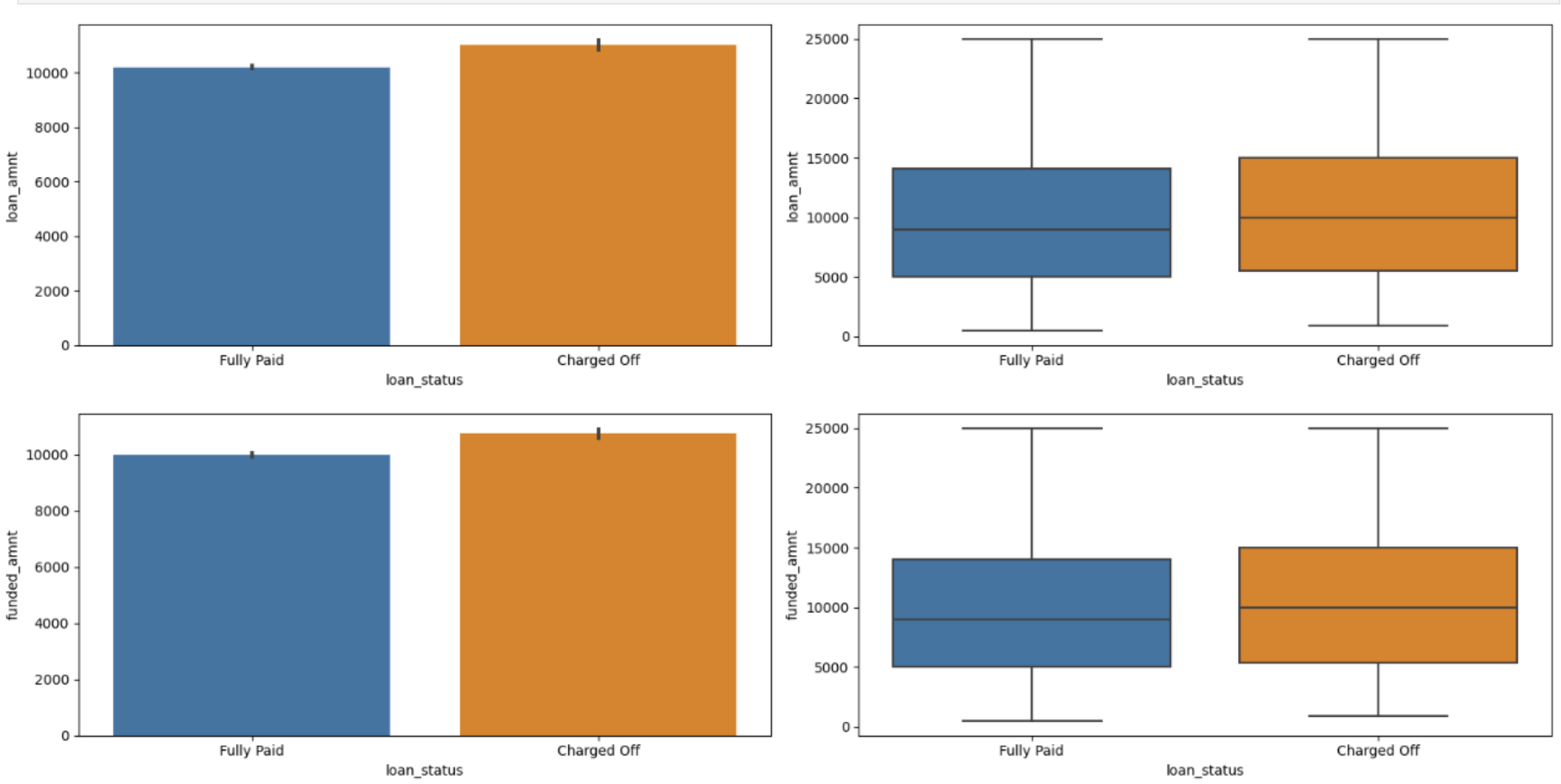
Univariate Analysis

Univariate analysis as the name suggest deals with a single column. Here, we have used Python functions to identify various percentile values for any particular data column. For ex. Loan Amount is a data column of numeric type. We displayed it's stats like mean, standard deviation, 50th percentile, 75th percentile and so on. We similarly used a box plot for pictorial view and it displays where major chunk lies and are there any outlier values. In this way, we performed for similar variables too.

Univariate analysis



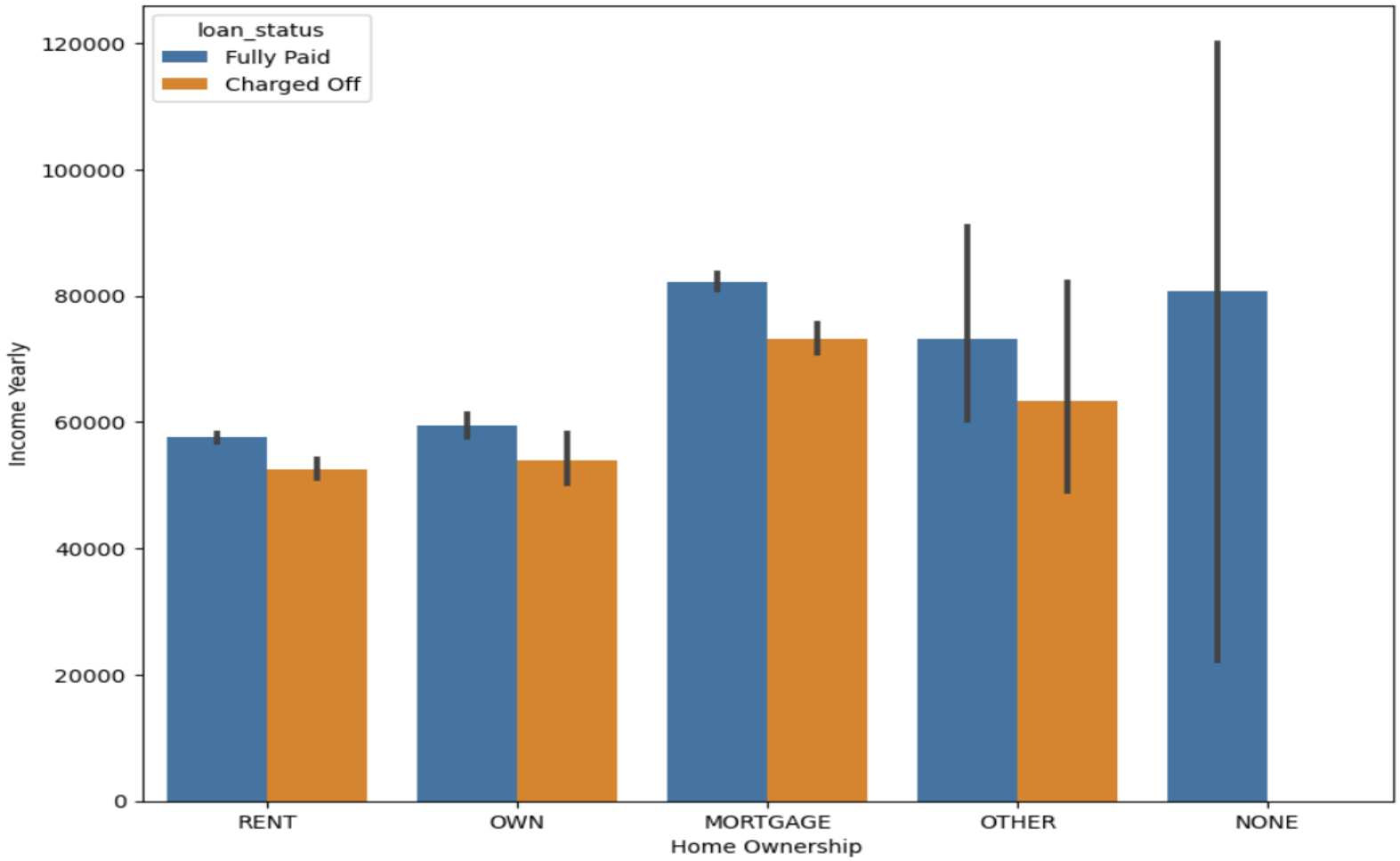
Segmented Univariate Analysis

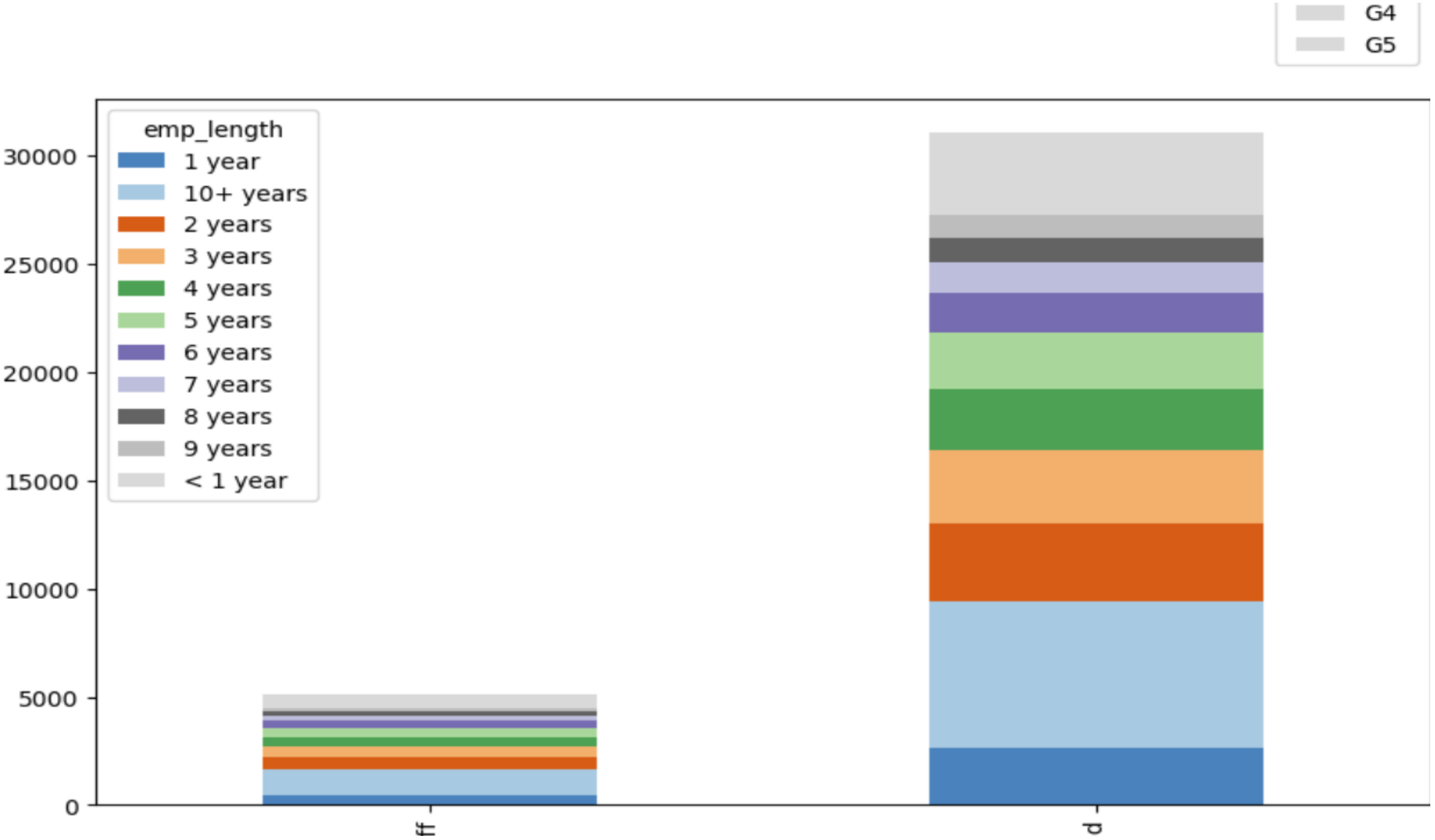


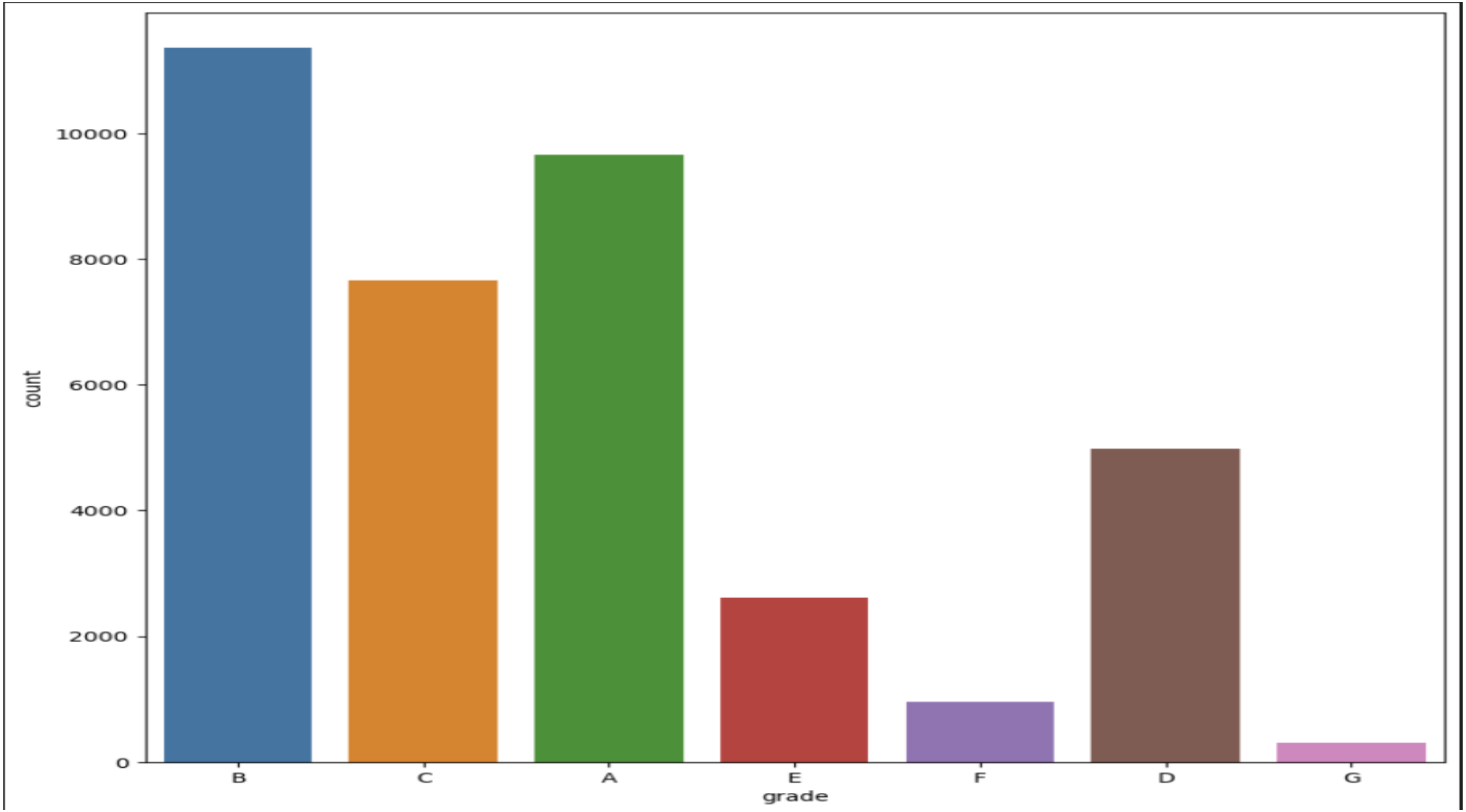
Bivariate analysis

Bivariate analysis comes from the term “correlation”. Two variables are positively correlated, two are negatively correlated to each other and in some cases two variables can be neutral to each other i.e. 0. Bivariate analysis is an important aspect of EDA as it helps to understand the dependency of one column on another. Below is an example of a bivariate analysis of yearly income vs Home ownership. Similarly, we can find the same for other variables too.

Bivariate Analysis




G4
G5
emp_length
1 year
10+ years
2 years
3 years
4 years
5 years
6 years
7 years
8 years
9 years
< 1 year
30000
25000
20000
15000
10000
5000
0
ff
d


10000
8000
6000
4000
2000
0
count
B
C
A
E
F
D
G
grade

Conclusion

Considering all the above analysis multiple factors can be taken into account while accepting or rejecting loan application

- Term → Duration of the loan (lesser the duration, lower the chances of defaulters)
- Interest rate → (lesser the interest rate higher the chances of loan repayment)
- Purpose for which the loan has been availed - debt_consolidation, credit card, home_improvement small businesses are major categories where loss is more.
- Grades → Loan Grades A,B,C have higher chances of repayment and C, D, E, F, G grades have most loan defaulters
- annual_inc —> higher the annual income lesser the loan defaulters
- addr_state —> Few states has high loan defaulters as compared to others eg: CA and FL

Summary

Hence, we performed Exploratory Data Analysis to understand and draw some decisive insights from the Loan dataset. We used some visualization through Python libraries like matplotlib and seaborn so as to get some better understanding.

Thank you

Prashant Kumar Misra

Pranav Kavishwar