Lending Club Case Study EDA Presentation

Cohort 68

By

Group Facilitator: Saurabh Kumar

Problem Statement

You work for a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company must decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

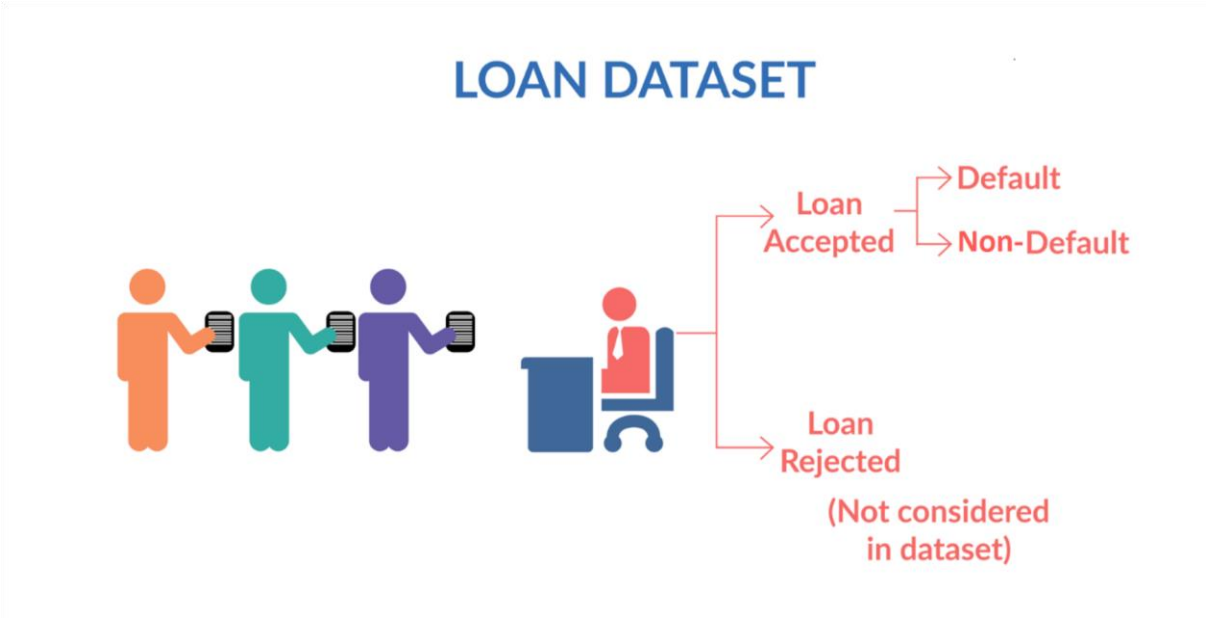
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below 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, etc.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Through Exploratory Data Analysis we need to identify how various factors impact the loan repayment of a customer, also how various customer and loan attributes influence tendency of load default. In this case study we will apply our knowledge of Univariate analysis, bivariate analysis, derived fields/ deriving factors to find insights in data set.

To start with different data analysis, we would need to pre-process, clean, and make data ready for analysis.



Case Study Objectives

- **Identification of the defaulter loan applicants**
- **Understanding the reason behind the load defaults**
- **Current standing of the lending club**
- **Areas of improvement**

Technology Used

- Anaconda 3
- Python - version 3.11.4
- Jupyter Notebook - version 6.5.4
- NumPy - version 1.24.3
- Pandas - version 1.5.3
- Matplotlib - version 3.7.1
- Seaborn - version 0.12.2

Packages to Import

Before we start data cleaning, we need to import all libraries that we will need in our EDA.

```
import pandas as pd
pd.set_option('display.max_rows', 130, 'display.max_columns', 130)
pd.options.display.float_format = '{:,.2f}'.format

import matplotlib.pyplot as plt
import seaborn as sns

import numpy as np
```

Data Cleaning & Pre – processing

1. First thing would be to import the data in a data frame.
2. Then we checked for data rows and columns , there are 111 columns with 39717 columns
3. Then we identified columns having all nulls or NaN (missing values)
4. We identified about 54 columns that were having missing values, we removed. Dropped then from dataset.

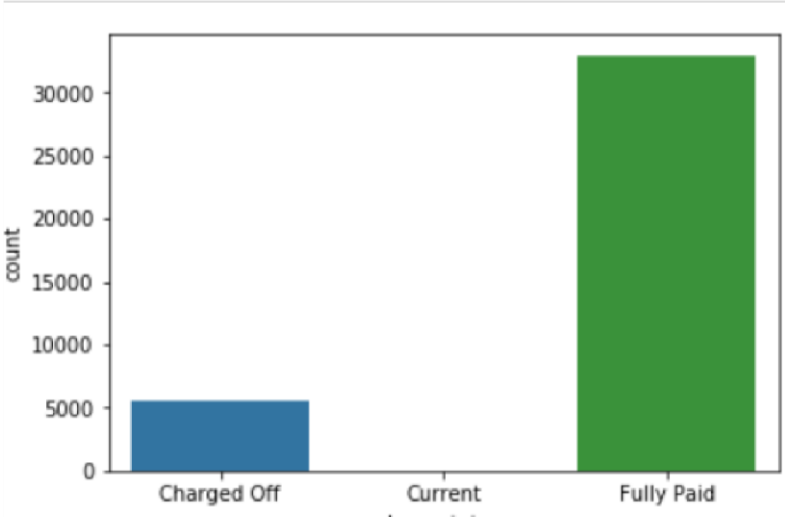
5. We further dropped columns which have too many missing in them.
6. After that we further identified columns which were having same values at most of places we dropped them as well.
7. After removing all columns we are left with 40 from 111 columns.

Derived Columns

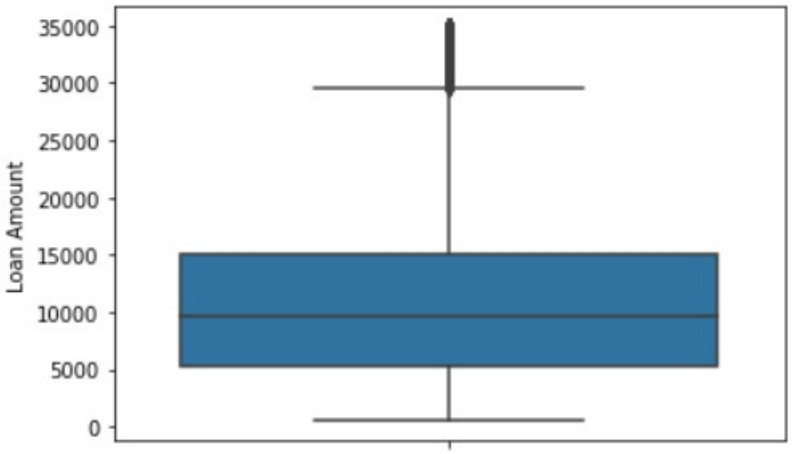
1. Let's create derived columns now: term_month to term , remove % from int_rate and create new column : int_rate_percentage
2. Derive emp_length_months from emp_length , remove years and + from values
3. Remove % from revol_util and create a column revol_util_percentage 4. Similarly we need to create few more and drop old ones
5. Create another derived column PnL and drop old ones.
6. Now Update columns to float
7. Now identify the types of these columns. There are 25 numerical, 12 categorical and 14 string features in the data set.

Univariate Analysis

1. For Loan Status, there are 32950 Fully paid records and 5627 charged off.



2. The loan amount varies from 0 to 35,000 having mean of 10,000.

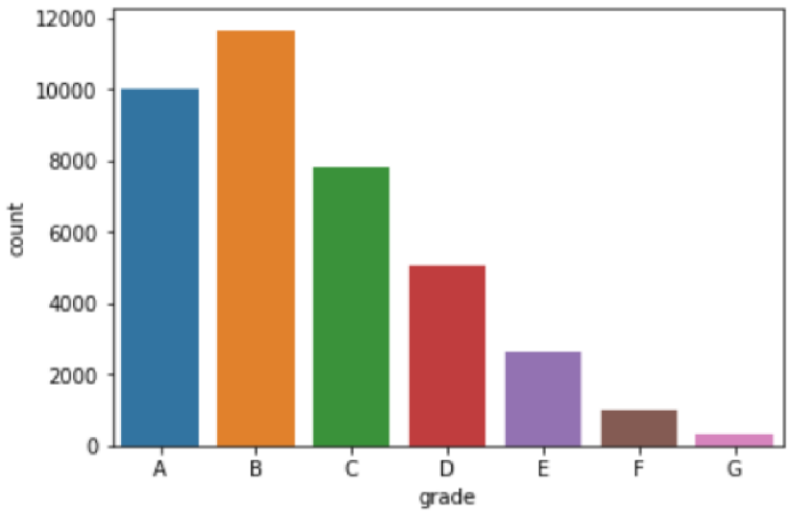


3. Most of the loans are Fully Paid.

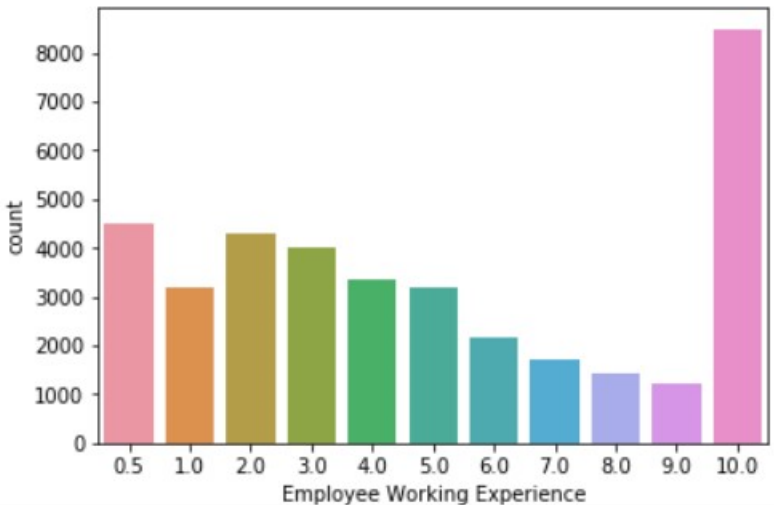
4. About 14% of loan are having status as defaulters.

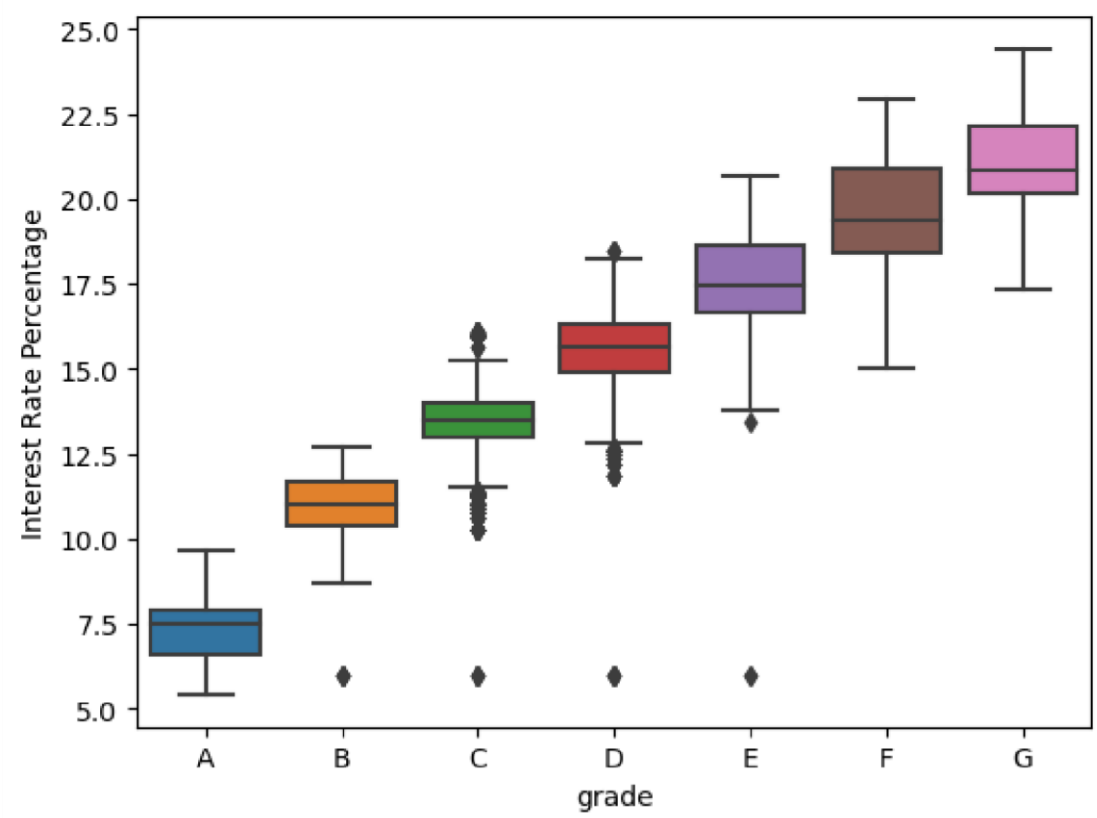
5. We have a class imbalance here.

6. Most of the loans have grade of A and B. Therefore stating most of the loans are high graded loans

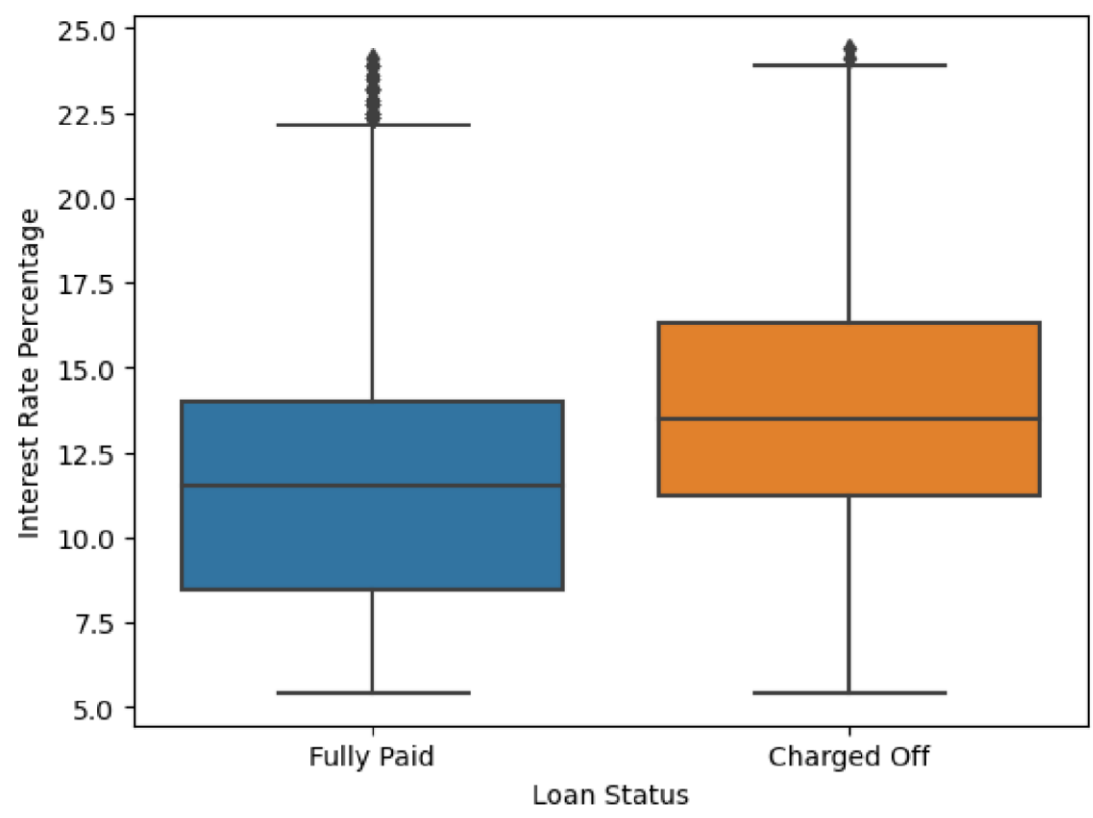


7. The loan default rate is high for high graded loans.
8. Majority of employees applying for the loan have more than 10 years of experience.

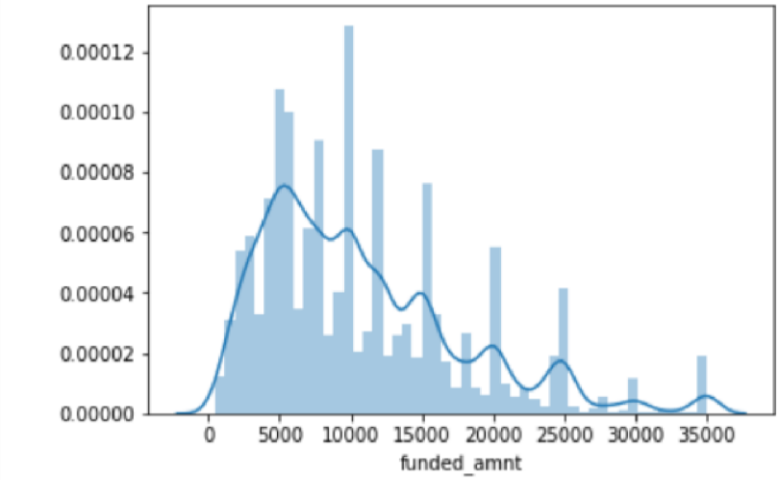




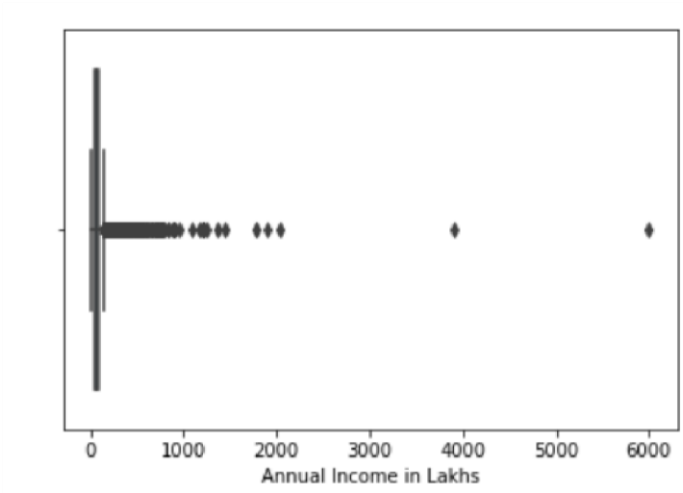
9. The highest interest rate applied is up to 25 & and high graded loans are bearing less i nterest.
10. People with higher interest rate tend to default more.



11. Funded amount is left skewed. Most of the loan amount given is 5 lakhs

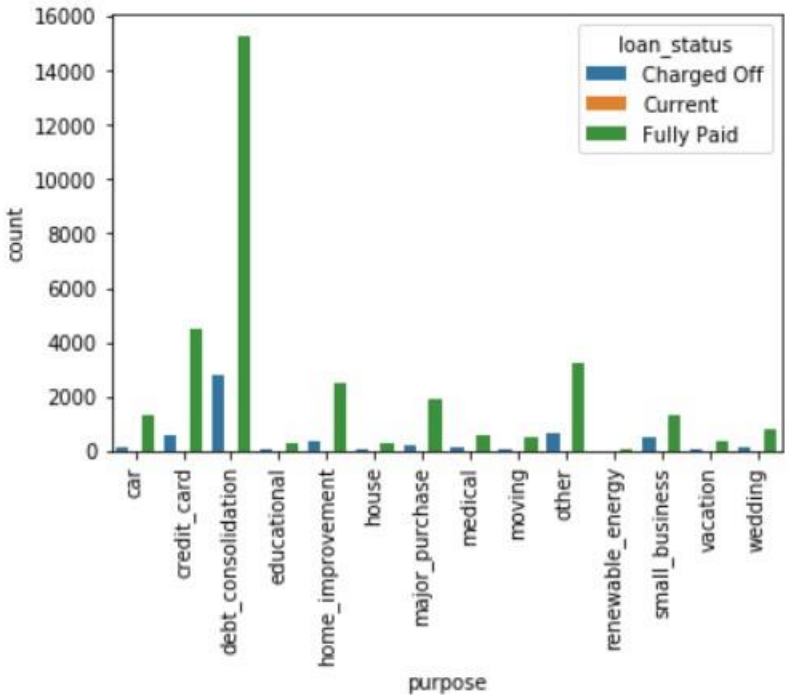


12. There are only two applicants having annual income of more than 30 lakhs

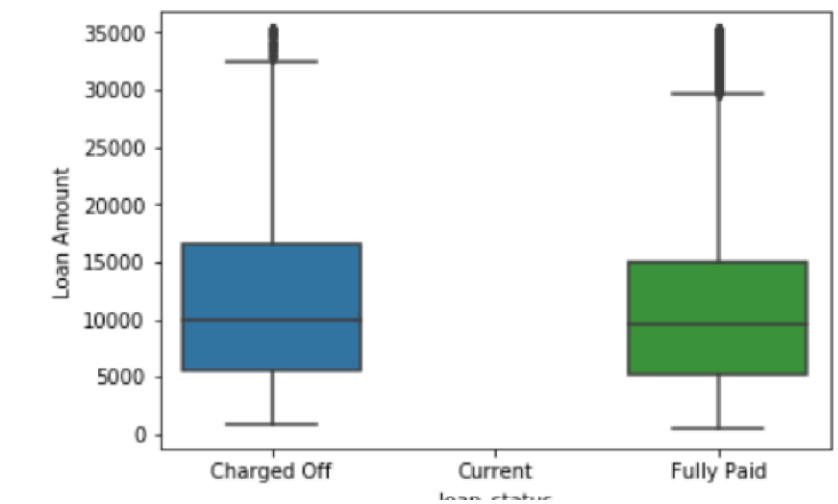


Segmented Univariate Analysis

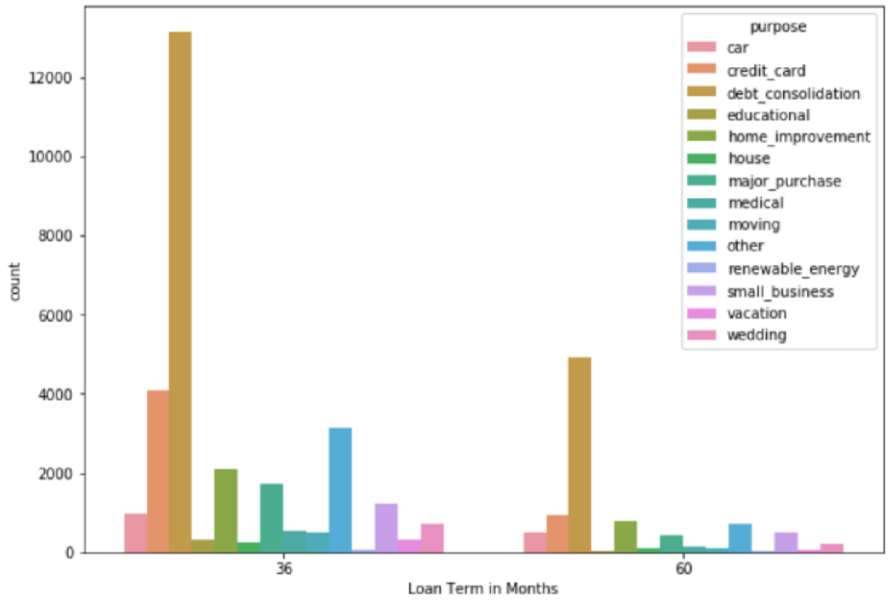
1. Majority of loan has been given for the debt consolidation purpose and has been f ully paid.



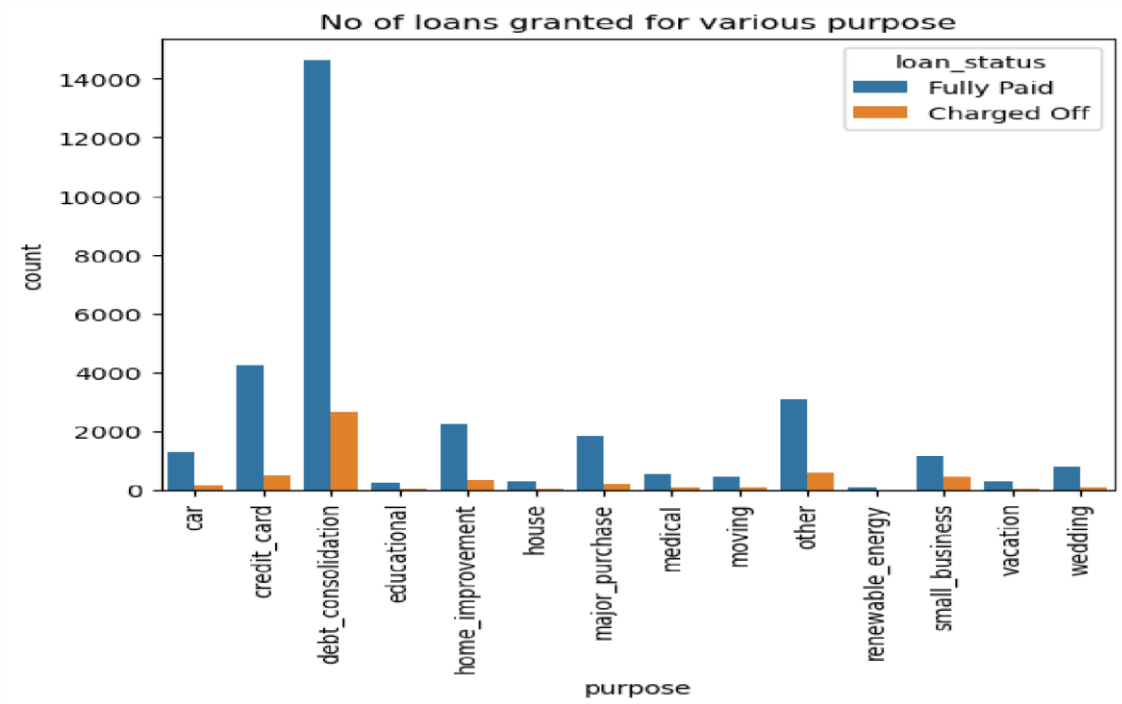
2. Mean, 25% and 75% Loan amount of Fully paid and charged off is exactly same



3. Tenure of 36 months have high chances to be defaulters

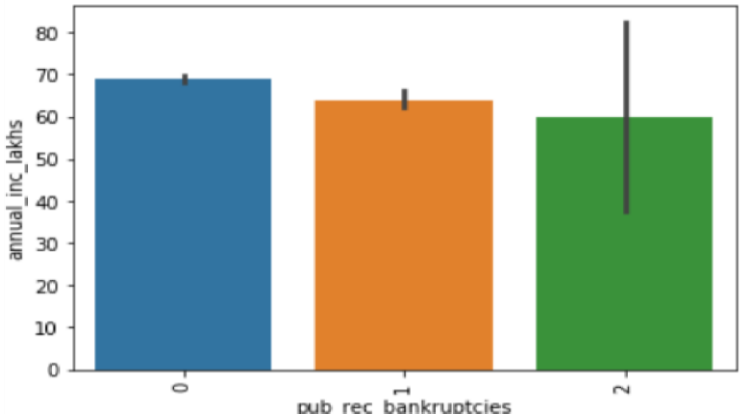


4. Wedding, Vacation, Moving, House and Education seems to be of Safe loan types.

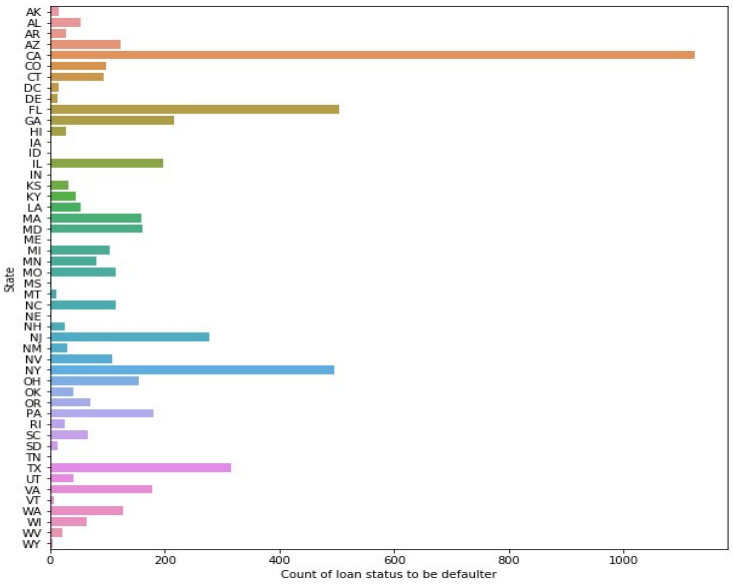


Bivariate Plots

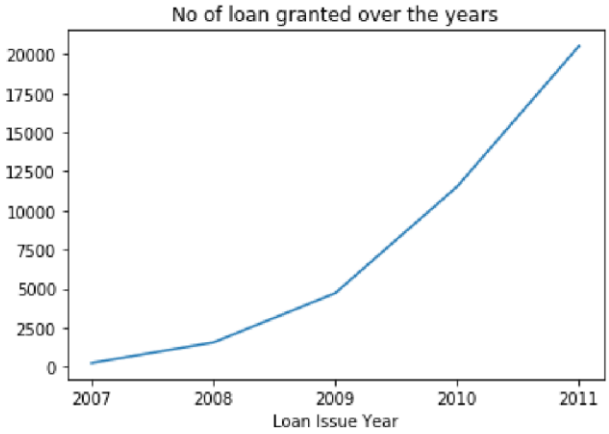
1. Off All Publicly Recorded Bankruptcies, Customers with slightly higher income has no bankruptcies.



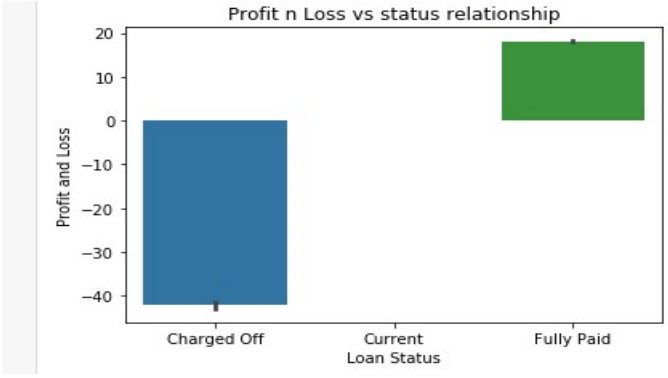
2. Applicants from the state CA are having high probability to be default.



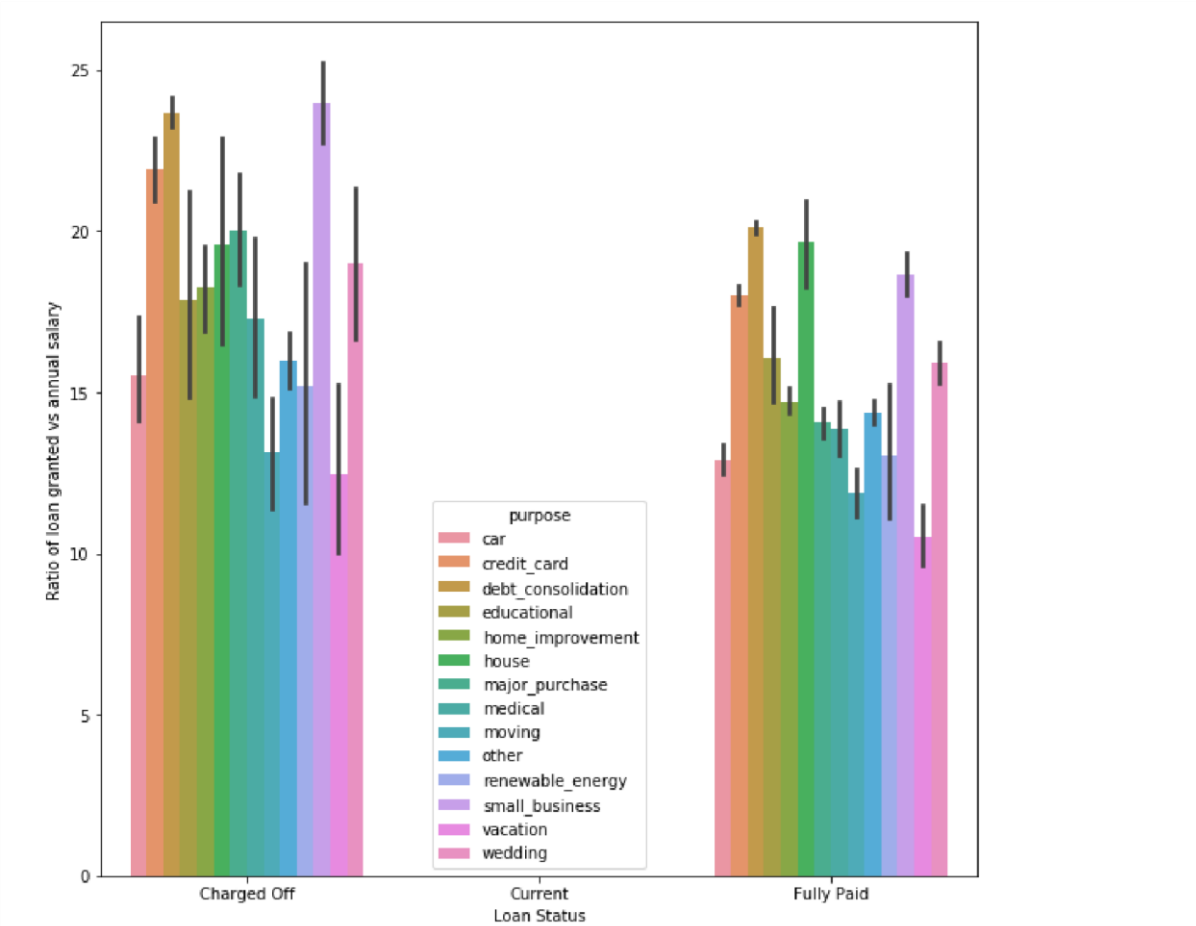
3. There is an increment in loan applicants from 2007 which increases with much highe r rate from 2010 to 2011.



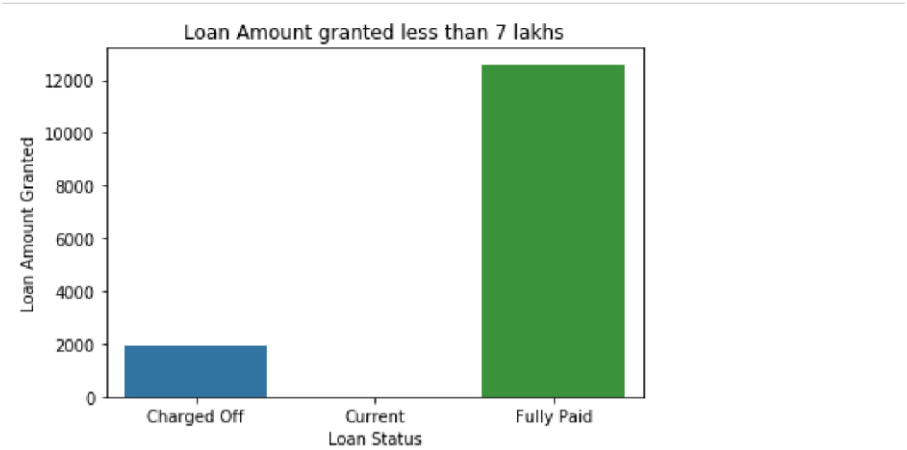
4. Most customers who have defaulted suffers losses.

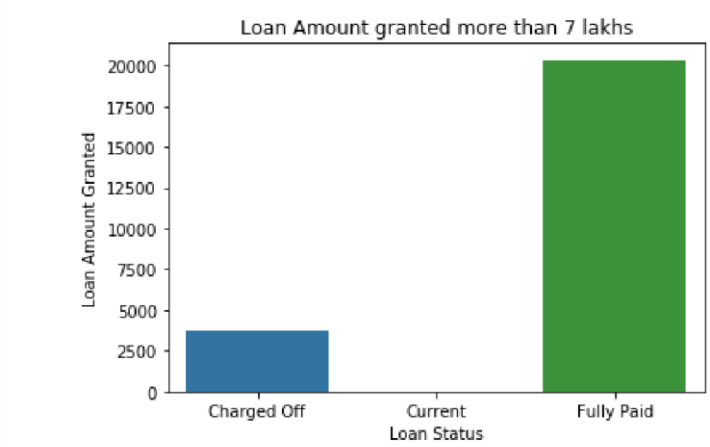


5. If we see from Ratio of loan granted to annual income, customers with higher ratio h ave defaulted more.

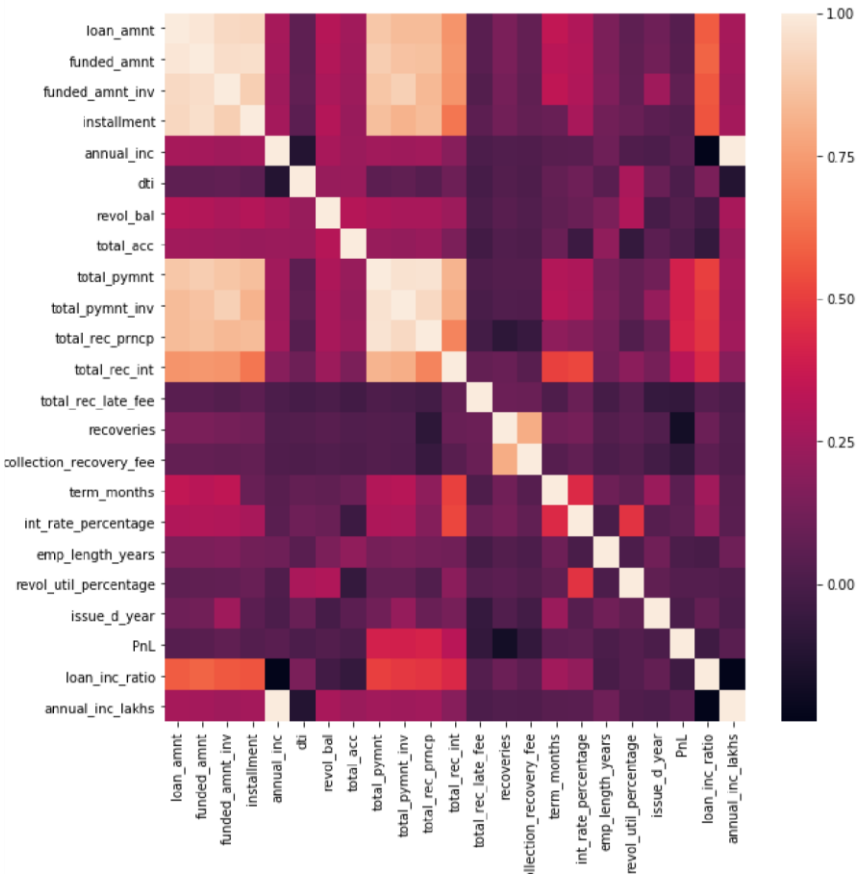


6. Another insight is customers with higher annual income have fully paid loan amount.

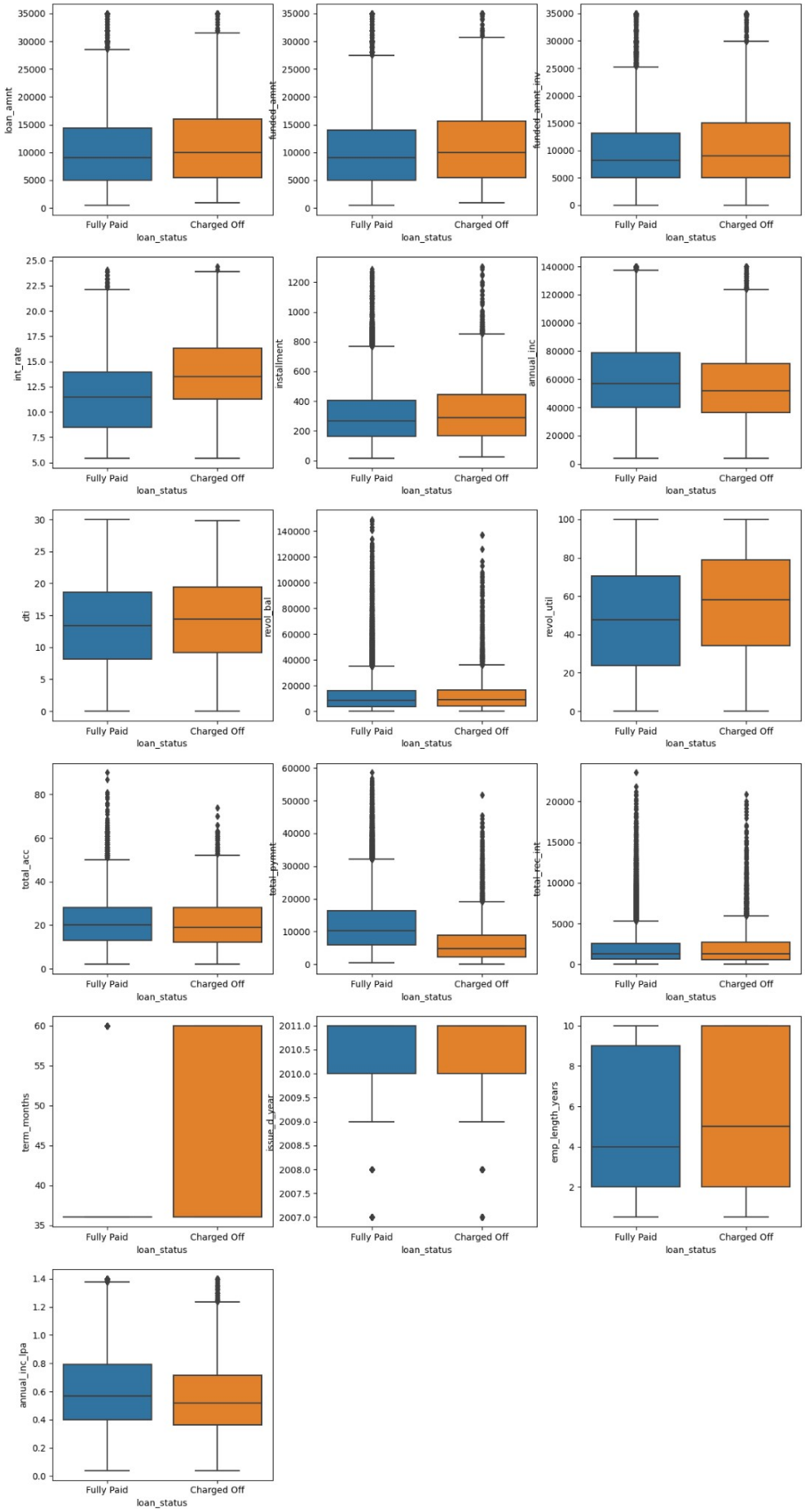




7. From Correlation Metric we can identify that Loan Amount, Funded Amount and Insta lment are highly correlated positively, Total Payment columns too are significantly relate d in a positive way to all mentioned in prior statement. Loan Amount is not too much rel ated to Emp_Length years, interest rate and term months.



8. The relationship between default rate and different parameters, it suggests a string relationship between dti, pub bankruptcies, term months and annual income.



Steps Performed

- As part of case study, we first

processed the data which takes most of the time in analysis, included fixing **rows,** columns, deriving new columns, managing missing values etc.

- Then we identified which are numerical variables and which are categorical, further identifying Integer, float, and strings.
- Then we performed Univariate analysis numeric variables to see how various variable are distributed to find out ranges and quartiles by graph.
- Similarly, we did Univariate analysis of segmented variables to find how they impact loan status in data set.
- Finally, we did bivariate analysis to identify how one variable is related / correlated with other and how change in values for one impact other.

What Lending Club is Doing Well?

- High Graded Loans: Lending Club is predominantly providing loans with high grades (A and B), which suggests that they are generally giving loans to borrowers with good creditworthiness. This is a positive sign as higher-grade loans tend to have lower default rates.
- Loan Purpose Analysis: Majority of loans have been given for debt consolidation purposes, and most of these have been fully paid. This indicates that loans for debt consolidation may be a relatively safer category.
- Income vs. Bankruptcies: Customers with slightly higher income have no recorded bankruptcies. This suggests that higher income levels may be a good indicator of creditworthiness.
- Income vs. Loan Status: Customers with higher annual income are more likely to have fully paid off their loans. This implies that higher income levels are associated with a lower risk of default.
- Correlation Analysis: Understanding the correlations between variables is crucial. For example, identifying that Loan Amount, Funded Amount, and Instalment are highly correlated is important for risk assessment and portfolio management.

Areas for Improvement

- Lending club can run some promotions during the end of the year to attract more customers during festive seasons.
- Lending club can focus on some safe debts for the purpose of wedding, education, vacation etc.
- Class Imbalance: There is a class imbalance in loan status, with about 14% of loans being charged off. Lending Club should focus on strategies to minimize defaults, such as refining their underwriting process or targeting higher creditworthy borrowers.
- Tenure Impact on Default: Loans with a tenure of 36 months have a higher likelihood of default. This could be a signal to reassess the risk associated with loans of this tenure.
- State-specific Risk: Applicants from the state of CA seem to have a higher probability of default. Lending Club should consider implementing strategies to mitigate this risk, such as stricter lending criteria or additional due diligence for applicants from this state.
- Income Verification: Given that there are only two applicants with an annual income of more than 30 lakhs, Lending Club should ensure that income verification processes are robust and reliable to prevent fraud or misrepresentation.
- Increase in Loan Applicants: There has been a notable increase in loan applicants from 2007, which accelerated from 2010 to 2011. This could indicate increased market penetration or increased risk exposure. It's important to assess whether the underwriting process has kept pace with this growth.
- Ratio of Loan Granted to Annual Income: Customers with higher ratios of loan amount to annual income have defaulted more. This suggests that lending amounts relative to income need to be carefully assessed and potentially capped to mitigate risk.
- Employee Experience: While not directly related to loan performance, it's worth noting that the majority of employees applying for loans have more than 10 years of experience. This could be indicative of a stable employment situation, which is a positive factor for creditworthiness.

Conclusion

Overall, Lending Club should continue to focus on improving there business processes, monitoring trends, and implementing risk mitigation strategies to minimize losses and nonperforming assets. Additionally, they should keep an eye on state-specific risks and consider adjusting lending criteria accordingly.

Acknowledgement/References:

- https://www.kaggle.com/code/lonewolf95/eda-101-univariate-analysis-forbeginners

- https://www.geeksforgeeks.org/get-unique-values-from-a-column-in-pandasdataframe/
- https://www.kaggle.com/code/kashnitsky/topic-1-exploratory-data-analysis-withpandas
- https://www.tutorialspoint.com/how-to-select-all-columns-except-one-in-a-pandasdataframe#:~:text=To%20select%20all%20columns%20except%20one%20column%2 0in%20Pandas%20DataFrame,%5D.