Lending Club Case Study

By:
Deepak TM
Shaily

Problem Statement

The primary goal of this exploratory data analysis (EDA) case study is to identify the key factors and patterns that drive loan defaults, with a specific focus on distinguishing between loans that are "Fully Paid" and those that are "Charged Off."

The dataset contains the complete loan data for all loans issued through the time period 2007 to 2011.

Understanding Loan Status and Data Limitations

The loan dataset includes a loan_status column, which categorizes loans into three states:

- **Fully Paid:** The borrower has successfully repaid the loan in full.
- **Charged Off:** The loan has been deemed uncollectible and is considered a loss for the lender (default).
- **Current:** The borrower is currently making payments, and the loan is in good standing.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

By focusing on loans with definitive outcomes ("Fully Paid" or "Charged Off"), we aim to uncover patterns and relationships that can help business:

- **Identify High-Risk Borrowers:** Determine the borrower characteristics, loan attributes, and credit history factors that are strong predictors of loan defaults.
- **Inform Lending Decisions:** Use the insights from EDA to refine lending policies, adjust interest rates based on risk, and develop more effective underwriting criteria.

Approach

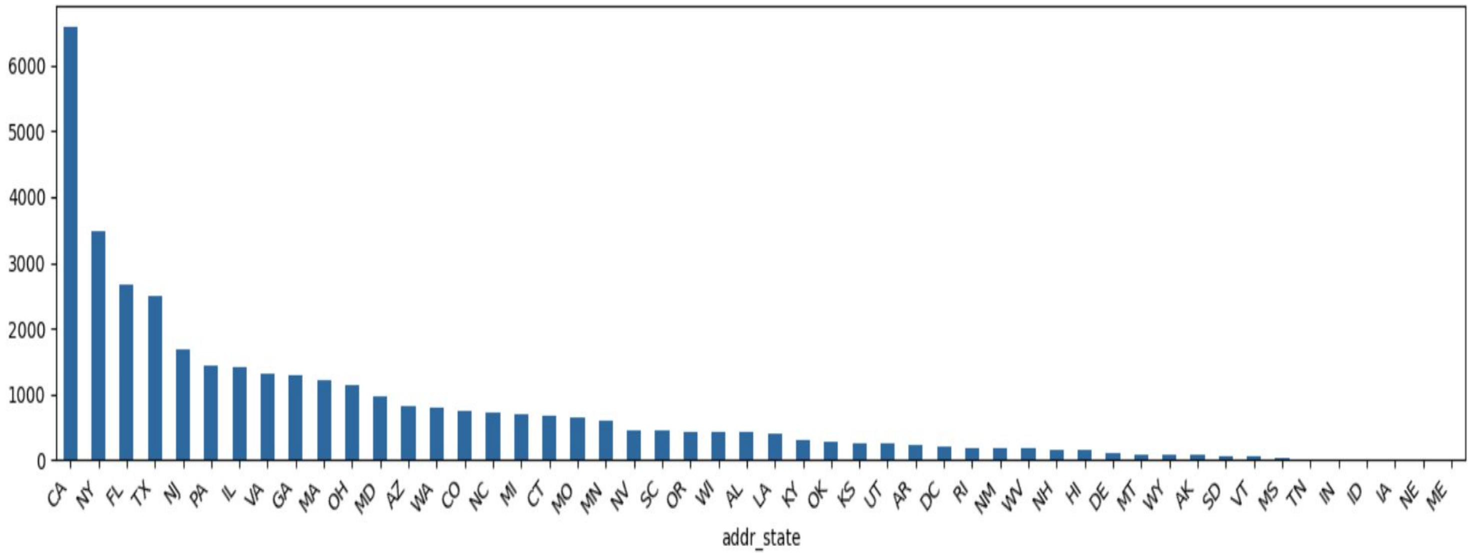
1. Data Discovery
    a. Import Libraries
    b. Load Dataset
    c. Explore Dataset
2. Data Handling and Cleaning
    a. Drop Columns with more than 60% null values.
    b. Handle Missing Values under acceptable range and Incorrect Data Types
        i. Imputation
            1. For categorical columns, we prefer to use mode
            2. For numerical values, we prefer to use mean or median (Median is preferred)
    c. Drop Columns with Duplicates
    d. Drop Columns with unique rows
3. Outlier detection and removal
4. Univariate Analysis
5. Bivariate Analysis

Data Understanding

Numerical Columns	**Categorical Columns**
Continuous Loan Amount Funded Amount Funded Amount Inv Installment Interest Rate Annual Income Debt To Income Ratio Revolving Balance Revolving Utilization	Ordered Term Grade Sub Grade Employee Length Earliest Credit Year Issue Quarter Year Issue Year
Discrete Number of Delinquencies in last 2 years Open Account Public Account Total Account Public record bankruptcies	Unordered Home ownership Verification Status Loan Status Purpose Address State

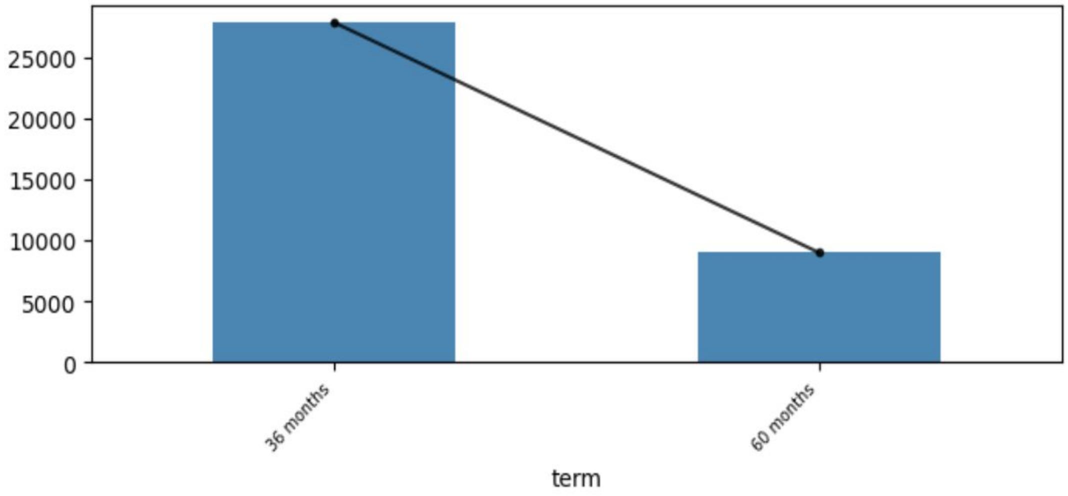
Univariate Analysis

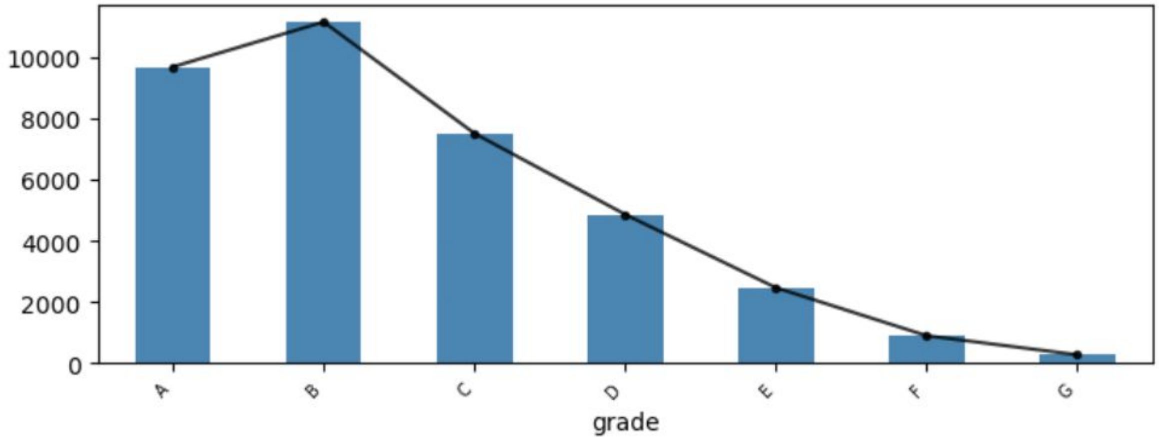
Categorical Columns



Address State - Loan volume varies significantly by state, with California (CA) having the highest and many states having very few loans.

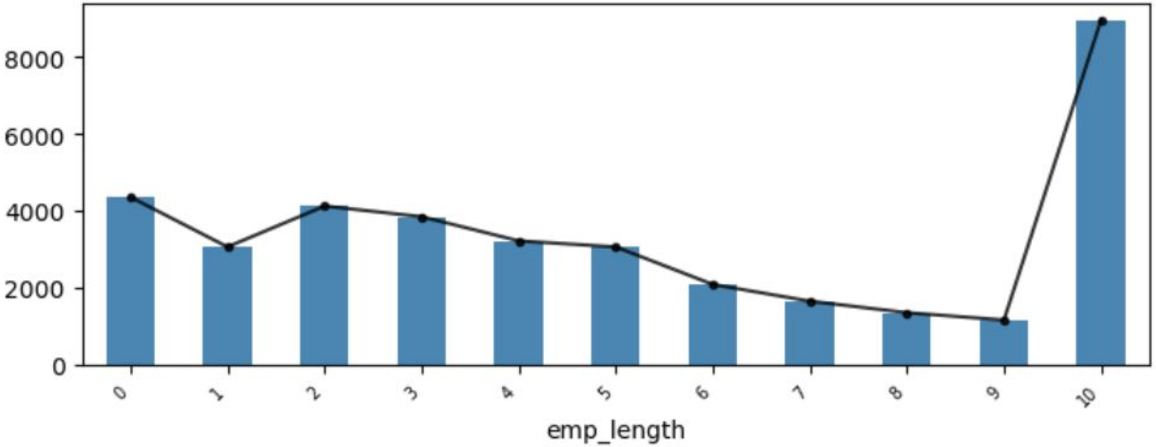
Term - The majority of loans have a 36-month term.



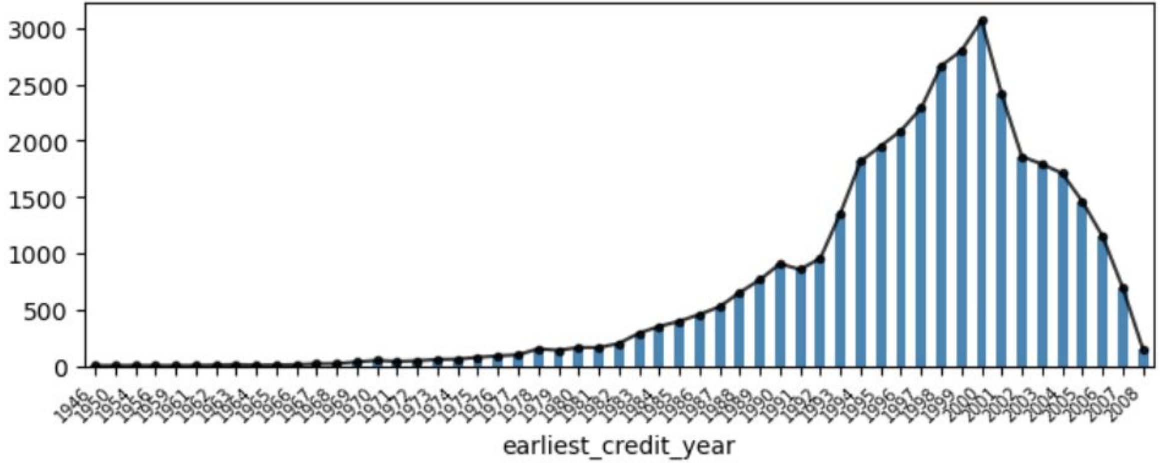


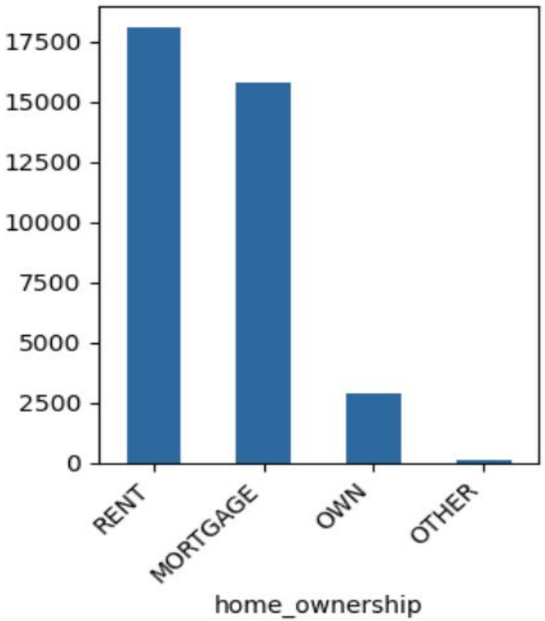
Grade - Loan volume decreases as the loan grade moves from A to G (riskier). Higher-grade loans likely have lower default rates, while lower grades require closer monitoring.

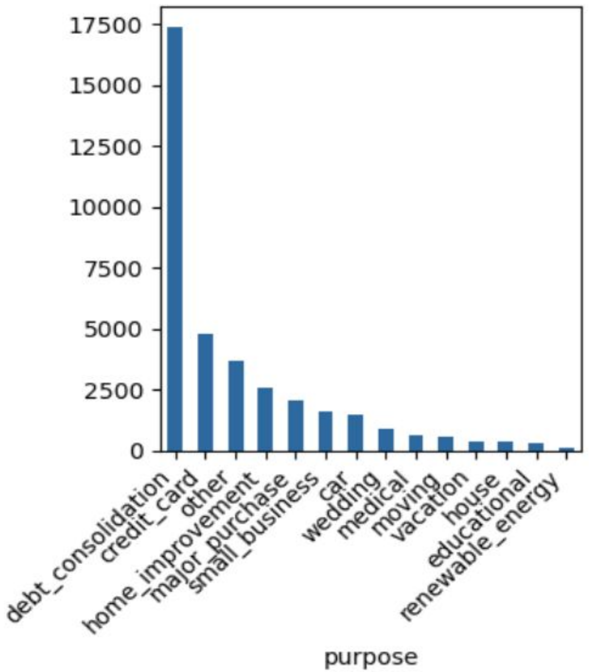
Employee Length - The relationship between employment length and loan volume is **not strictly linear**. Borrowers with **10+ years** of employment have the highest loan volume, followed by those with **< 1 year**.



Earliest Credit Year - A downward trend in loan volume for borrowers who established credit more recently (especially from the mid-2000s onwards). Most loan borrowers got their earliest credit line opened in the year 2001.

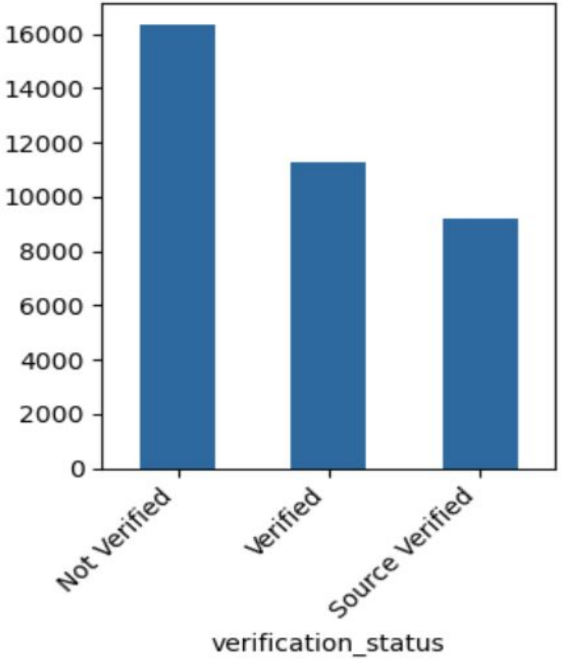




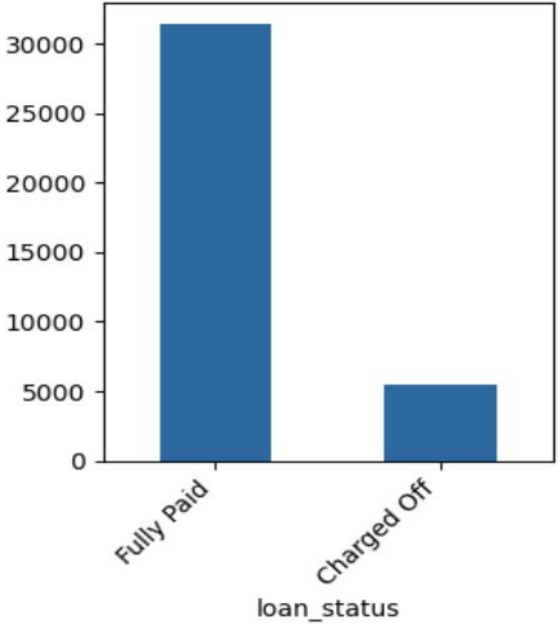


Home Ownership - A large majority of borrowers are either renting or have a mortgage. Very few own their homes outright or fall into the "other" category.

Purpose - "Debt consolidation" is the most common loan purpose, followed by "credit card", "other" and "home improvement".



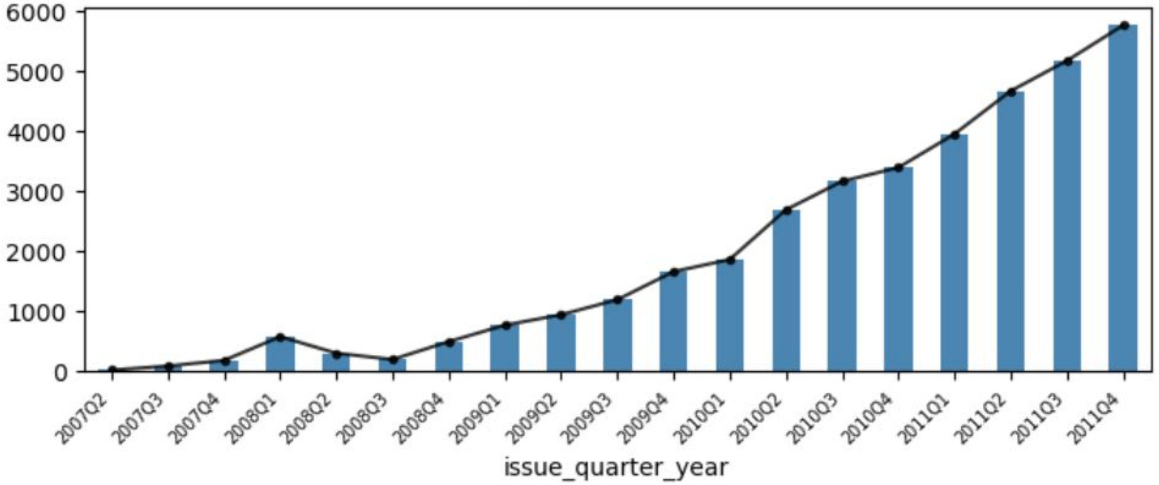
Income Verification Status - A significant portion of borrowers have not had their income verified. Majority of the dataset have borrowers who have their income or income source verified.

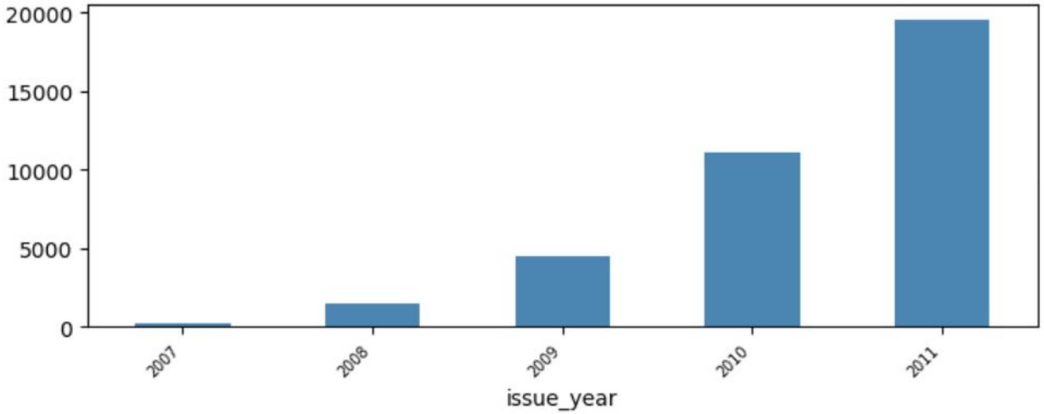


Loan Status - A large majority of loans are marked as "Fully Paid" which is a positive sign. However, there's still a notable proportion of "Charged Off" loans, indicating defaults.

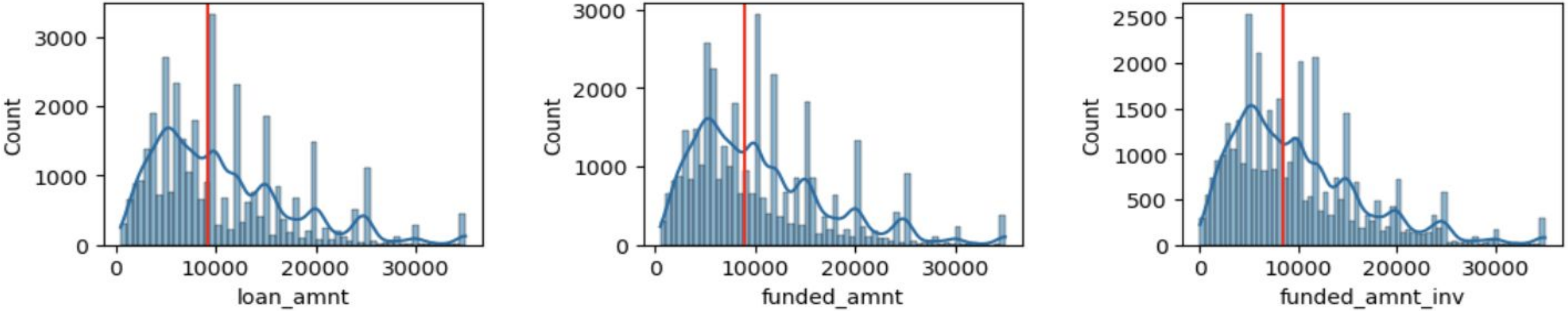
Issue Year & Issue Quarter by Year

- There is a general **upward trend** in loan originations over time, with some seasonality (potentially higher volume in later quarters).
- **2011** has the most number of accepted loans.





Numerical Columns

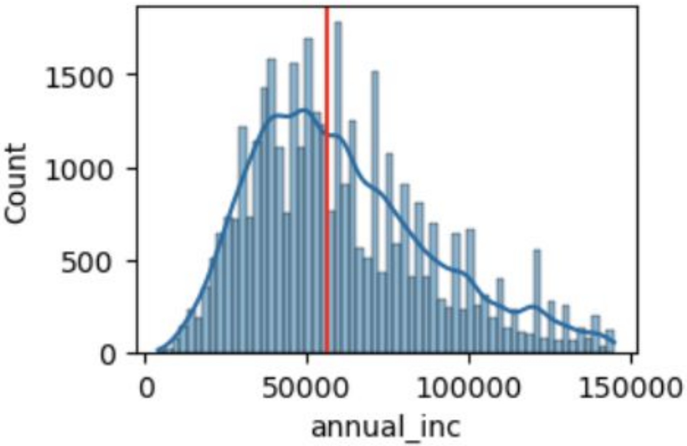


Loan Amount (Funded Amount and Funded Amount Invested)

- All three loan amount variables show a strong right skew, meaning most loans are concentrated at lower amounts with a tail extending towards larger loans.
- There are peaks around common loan amounts (e.g., 5,000, 10,000, 15,000 USD), likely reflecting standard loan offerings.

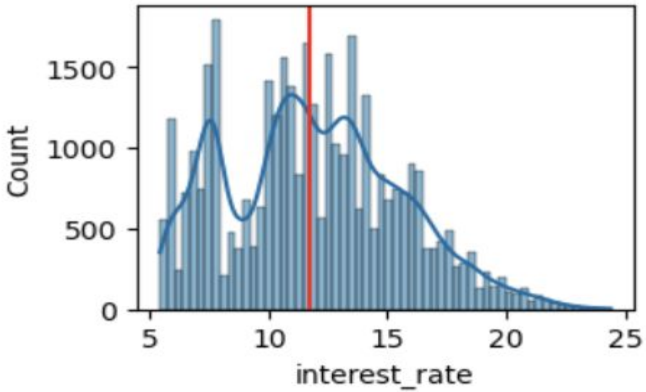
Annual Income

Annual income is heavily right-skewed, indicating that a large portion of borrowers have lower incomes, with a long tail of higher earners



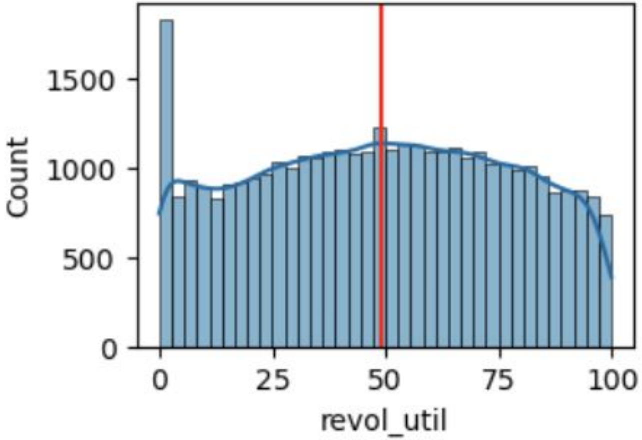
Interest Rate

The interest rate distribution is multimodal (multiple peaks), where we could see multiple peaks around 7.5% and in the range of 10%-15%. There are fewer loans with larger interest rate , mostly after 17.5%.



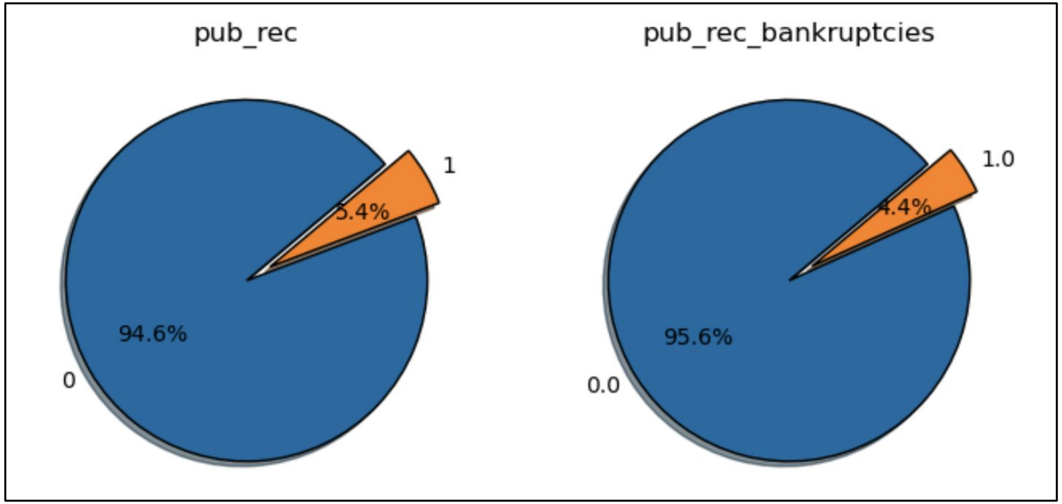
Revolving utilization

Revolving line utilization (the percentage of available credit being used) had a uniform distribution with more borrowers having revol_util from 0-2.



Public Records & Public Records Bankruptcies

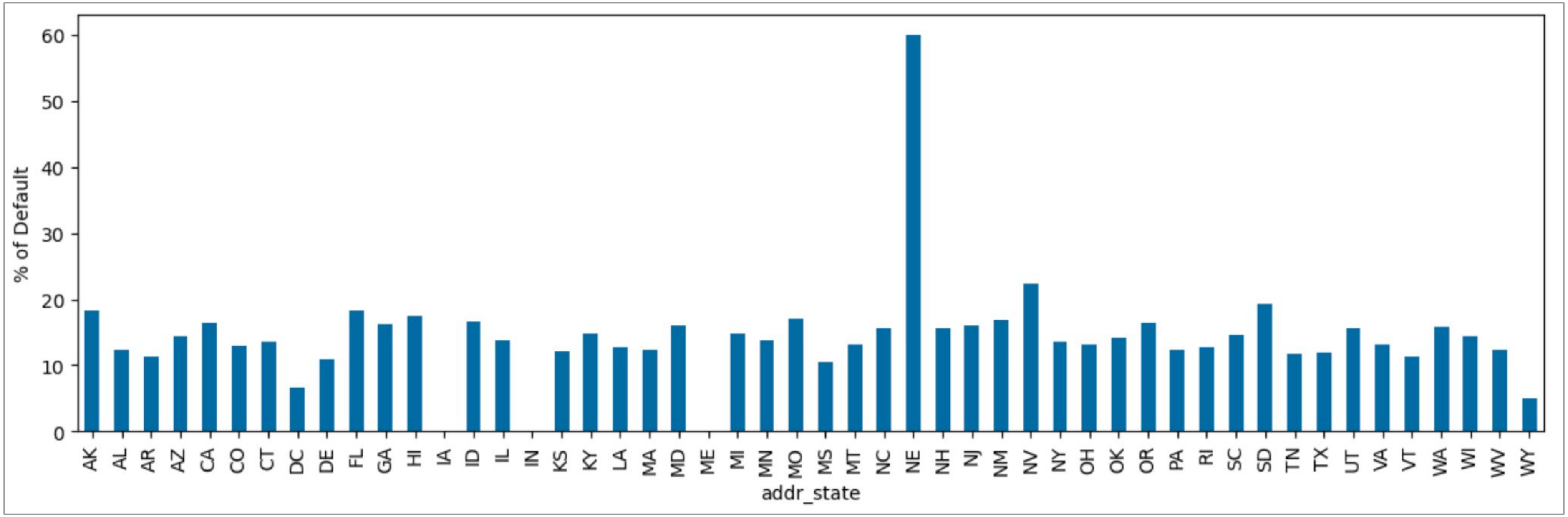
- Overwhelming Majority with No Records: Both pie charts show a dominant majority of borrowers (over 94%) have no derogatory public records or bankruptcies. This suggests that most loan applicants in your dataset have relatively clean financial histories.
- Small Percentage with Negative Events: Only a small percentage of borrowers have any negative public records (5.4%) or bankruptcies (4.4%).



Bivariate Analysis

Categorical Vs Categorical (vs Default Rate)

Address State - Default rates vary significantly by state. One state (NE - Nebraska) stands out with a much higher default rate.



Income Verification Status

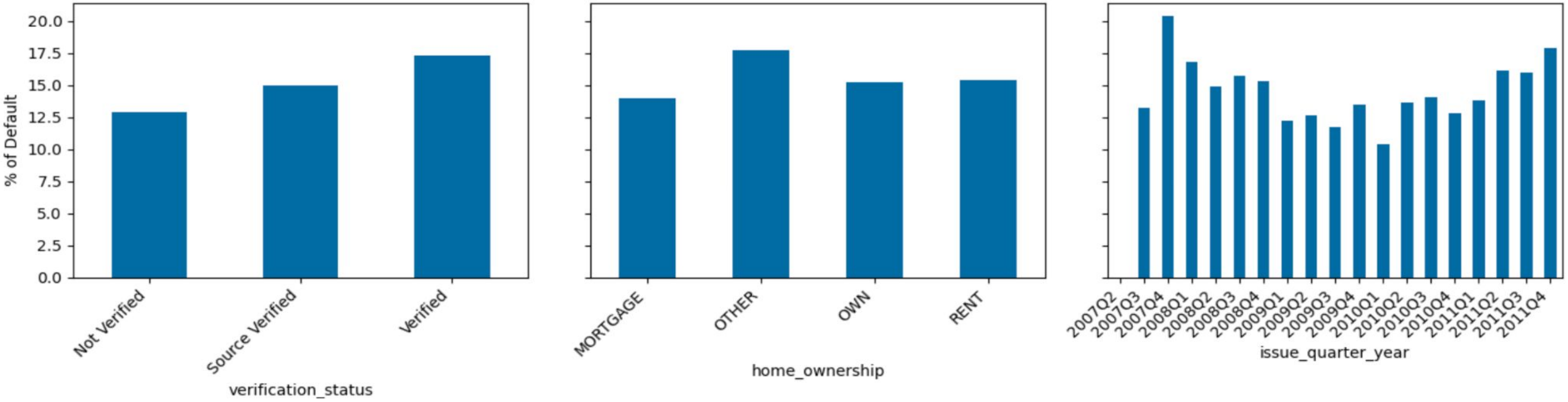
Loans with "Verified" income have a noticeably higher default rate than "Not Verified" or "Source Verified".

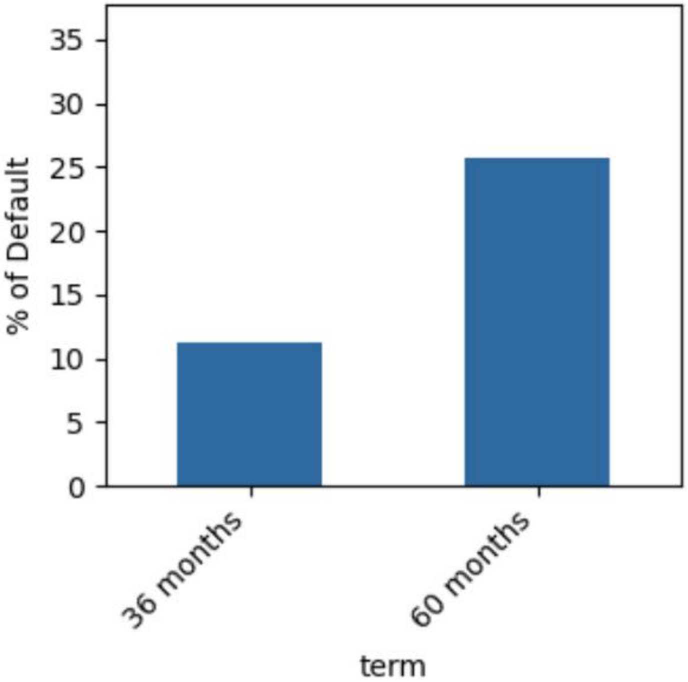
Home Ownership

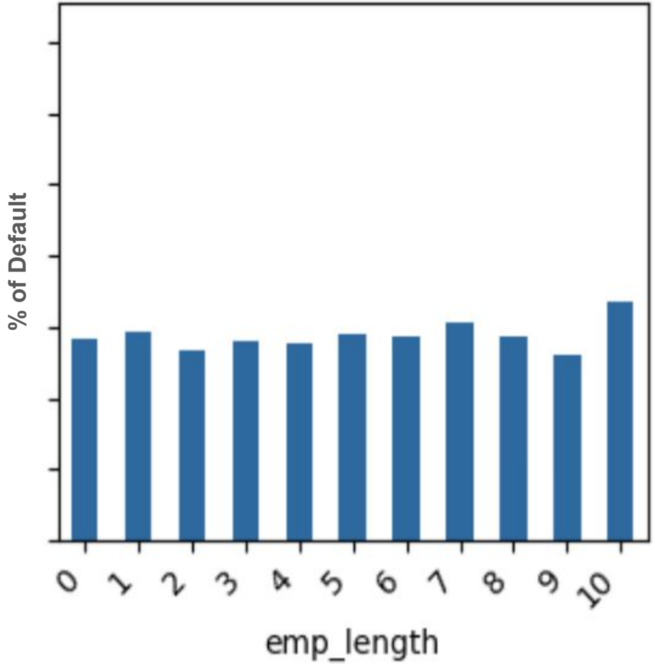
"OTHER" categories have higher default rates compared to others.

Issue Quarter Year

There's some fluctuation in default rates across different quarters and years, but no strong consistent trend. Also since, issue year is in past, it cannot help us analyse future default applicants.





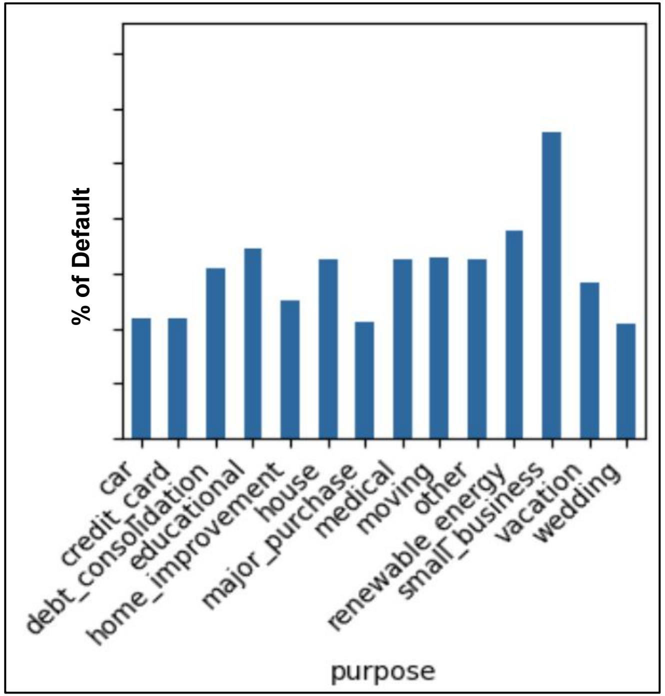


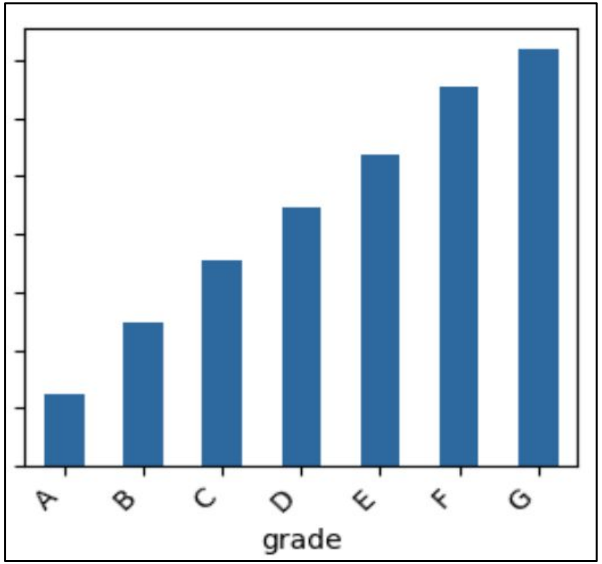
Term Vs Employee Length

Loans with a 60-month term have a significantly higher default rate compared to 36-month loans.

Default Percentage Vs Employee Length

There's no clear linear relationship between employment length and default rate. Borrowers with very short (0-1 year) or longer (9-10 years) employment histories have slightly higher default rates.



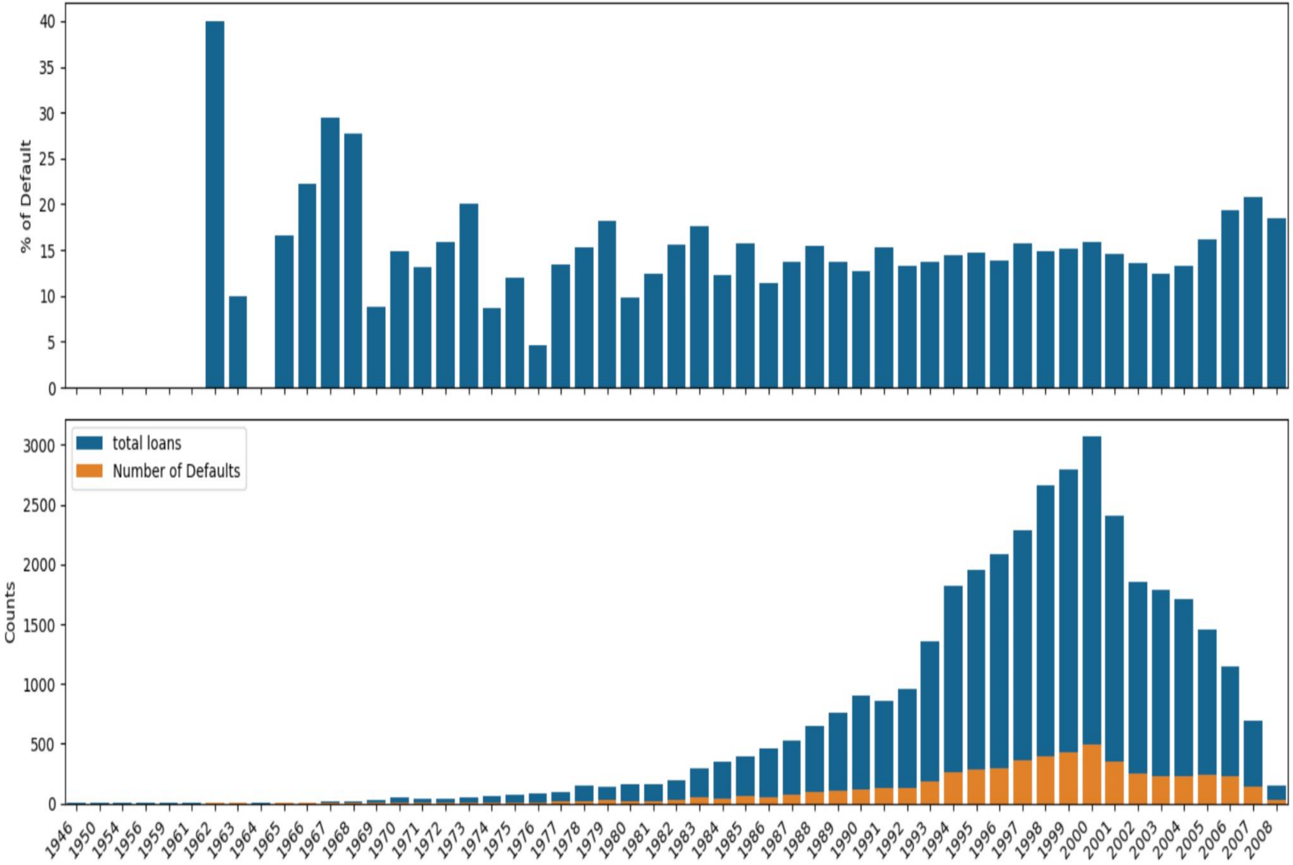


Purpose

"Small business" loans exhibit the highest default rate, followed by "renewable_energy" and "Educational". Car, Credit card, Major purchase and Wedding have lower default rates.

Grade

Default rates increase as loan grade deteriorates (from A to G). This pattern is a core principle of credit risk: Lower grades reflect higher risk.



Earliest Credit Opening Year

1946 to early 1980s: Borrowers with the earliest credit opened in this range had a very small number of loans issued in our dataset for time period 2007 t0 2011, with defaults being negligible or non-existent.

Mid-1980s to early 2000s: Borrowers with the earliest credit opened in this range have a significant increase in number of loans issued, peaking around 2000. However, defaults also increased but at a smaller rate compared to total loans.

Post-2000: For borrowers with the earliest credit opened in this range, number of loans issued in 2007-2011 decreased after 2000, with the number of defaults remaining relatively stable but declining as well after 2005.

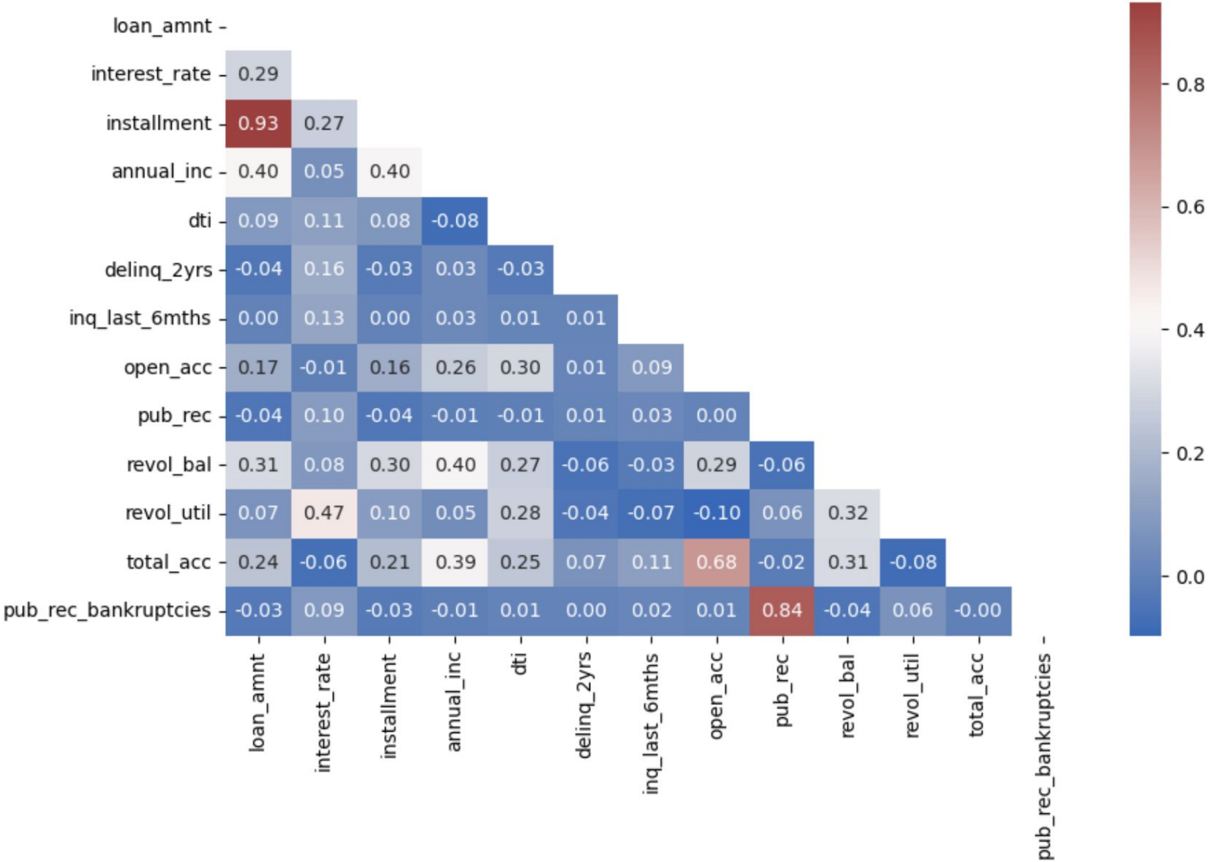
Numerical Vs Numerical

Correlation Matrix

1. Positive Relationship: One variable increase leads to increase in another variable (Positive Correlation)

2. Negative Relationship: One variable increase leads to decrease in another variable (Negative Correlation)

3. No Relationship: One variables seems unaffected from other variable (No Correlation)



	loan_amnt	interest_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	pub_rec_bankruptcies
loan_amnt													
interest_rate	0.29												
installment	0.93	0.27											
annual_inc	0.40	0.05	0.40										
dti	0.09	0.11	0.08	-0.08									
delinq_2yrs	-0.04	0.16	-0.03	0.03	-0.03								
inq_last_6mths	0.00	0.13	0.00	0.03	0.01	0.01							
open_acc	0.17	-0.01	0.16	0.26	0.30	0.01	0.09						
pub_rec	-0.04	0.10	-0.04	-0.01	-0.01	0.01	0.03	0.00					
revol_bal	0.31	0.08	0.30	0.40	0.27	-0.06	-0.03	0.29	-0.06				
revol_util	0.07	0.47	0.10	0.05	0.28	-0.04	-0.07	-0.10	0.06	0.32			
total_acc	0.24	-0.06	0.21	0.39	0.25	0.07	0.11	0.68	-0.02	0.31	-0.08		
pub_rec_bankruptcies	-0.03	0.09	-0.03	-0.01	0.01	0.00	0.02	0.01	0.84	-0.04	0.06	-0.00	

Observations

1. **Strong Positive Correlations (Darker Red):**

   Loan Attributes: Loan Amount and installment show a very strong positive correlation (0.93), which is expected. Larger loan amounts typically come with larger monthly installments.

   Credit History: Open account (number of open credit lines) and Total account (total number of credit lines) have a strong positive correlation (0.68), indicating that individuals with more open accounts tend to have a greater total number of accounts.

   Derogatory Records: Public record (number of derogatory public records) and public record bankruptcies also exhibit a strong positive correlation (0.84). This implies that people with bankruptcies are more likely to have other derogatory records on their credit history.

2. **Moderate Positive Correlations (Lighter Red):**

   Interest Rate and Revolving Utilization: They show a moderate positive correlation of 0.47. This indicates that as the revolving utilization increases, interest rates at which loans are given also increases. It makes sense as interest rate is also based on the borrower's credit history.

   Loan Amount & Annual Income: These variables have a moderate positive correlation. It can be deduced that as annual income of a borrower increases, their borrowing capability also increases. Thus, the increase in least loan amount requested.

   Revolving Balance & Annual Income: They show a moderate positive correlation (0.40).

   Total Account & Annual Income: As described above, increase of annual income leads to slight increase in Revolving Balance which may also be a result of more accounts being opened.

   Revolving Balance & Utilization: revol_bal (revolving balance) and revol_util (revolving utilization rate) show a weak positive correlation (0.32). This makes sense as higher revolving balances often lead to higher utilization rates.

Categorical vs Numerical Bivariate Analysis (Segmented Univariate Analysis)

Annual Income:

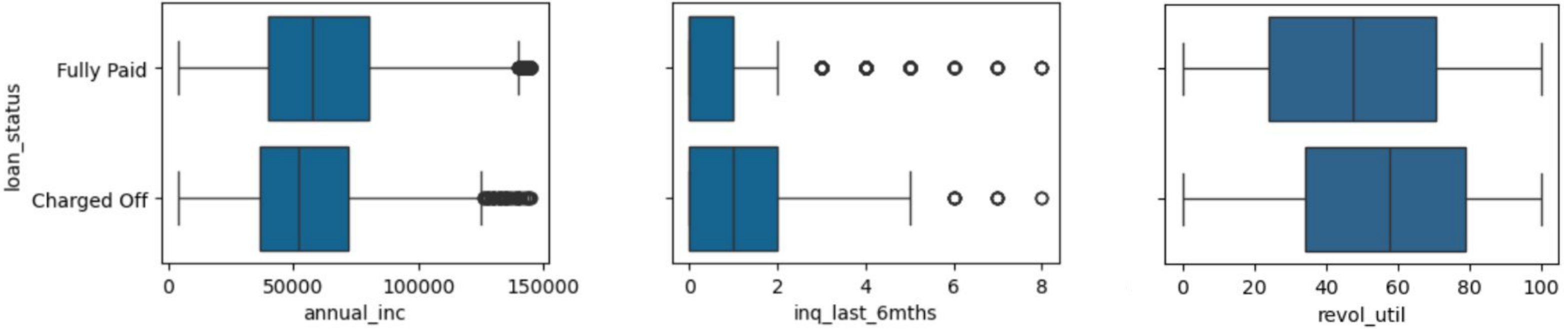
The median annual income for fully paid loans is slightly higher than for charged-off loans. However, there's considerable overlap in the distributions, and both groups have a wide range of incomes.

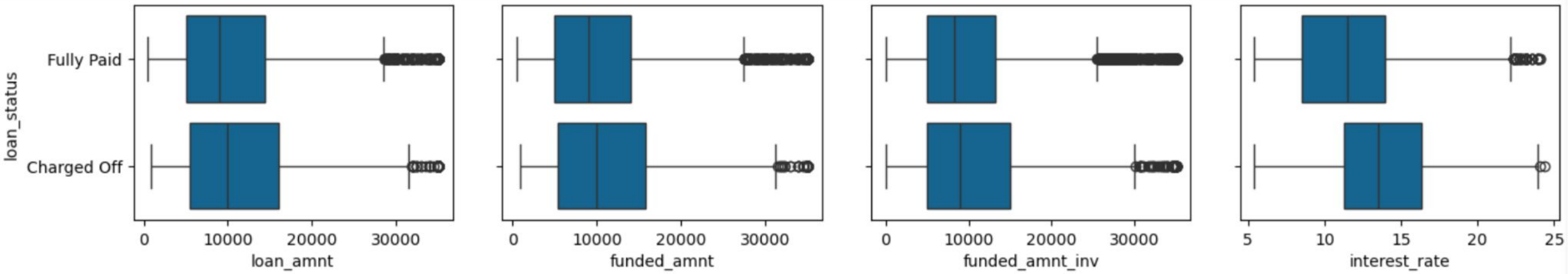
Inquiries in Last 6 Months:

Charged-off loans have median number of inquiries in the past 6 months equivalent to 75% of the paid off loans. This indicates that those borrowers are likely to default for whom more number of inquiries were done.

Revolving Line Utilization:

Charged-off loans have a significantly higher median of utilization of credits.





Loan Amount (Funded Amount and Funded Amount Invested):

The median loan amount for charged-off loans is slightly higher than for fully paid loans. There's more variability (a wider IQR) in loan amounts for charged-off loans. The maximum loan amount is also higher for charged-off as compared to fully paid.

Interest Rate:

25 percentile of charged off loans have approximately the same interest rate as the median of paid off loans. This indicates that loans with higher interest rate are very likely to being charged off. Charged-off loans have a significantly higher median interest rate compared to fully paid loans.

Driver Variables & Recommendations

1. **Inquiries in Last 6 Months:** Charged-off loans have median number of inquiries in the past 6 months equivalent to 75% of the paid off loans. This indicates that those borrowers are likely to default for whom more number of inquiries were done.
2. **Address State -** As we noticed, Nebraska state had the maximum number of defaulters. We can take the address state into consideration while assessing borrowers from next time.
3. **Term -** Borrowers with 60 month term tends to default more. It can be because more time to pay off a loan increases the chance of them charging off.
4. **Annual Income:** The median annual income for fully paid loans is slightly higher than for charged-off loans. However, there's considerable overlap in the distributions, and both groups have a wide range of incomes.
5. **Interest Rate:** 25 percentile of charged off loans have approximately the same interest rate as the median of paid off loans. This indicates that loans with higher interest rate are very likely to being charged off. Charged-off loans have a significantly higher median interest rate compared to fully paid loans.
6. **Purpose -** "Small business" loans exhibit the highest default rate, followed by "renewable_energy" and "Educational". Car, Credit card, Major purchase and Wedding have lower default rates.
7. **Home Ownership -** "Other" categories have higher default rates compared to others. The ways to verify home ownership of a borrower can be improved.
8. **Grade -** As the grade moves from A to G, credit worthiness decreases and risk factor increases. It also leads to increase in default rate. There are still default borrowers with grade A. This variable’s calculation can be improved.