# Loan Data Analysis and Recommendations

Analyzing Key Factors for Lending Decisions

BY: Harsha Punati , Dinesh Reddy

3rd July 2024

# Introduction

- Objective: To analyze loan data and identify key factors influencing lending decisions.
- Data Source: Loan dataset given in website
- Scope: Data understanding, cleaning, analysis, and recommendations.

# Data Understanding

- Steps Performed:
  - Data Information
  - Info on Missinng Values
  - Summary statistics for numerical columns
  - Check data types and unique values for categorical columns

- **Observations:**
  - 1 : 39717 rows and 111 columns provided,
  - 2 : dtypes: float64(74), int64(13), object(24)
  - 3 : Missing values - 68 columns have null values and 54 of them have all null values
  - 4 : From the plot, almost a third of the columsn seems to have outliers, especially column annual_income seems to have significant number of outliers

- 5: Summary statistics for numerical columns. Few observations

  - Loan Amounts: The loan amounts vary widely with a mean of $11,219 and a standard deviation of $7,456. The minimum loan amount is $500, while the maximum is $35,000. This indicates a wide range of loan sizes provided to the borrowers.

  - Annual Income: The annual income of the borrowers also shows significant variation, with a mean of $68,969 and a high standard deviation of $63,794. The minimum annual income is $4,000, and the maximum reaches $6,000,000. This indicates a diverse borrower base in terms of income levels.

  - Debt-to-Income Ratio (DTI): The average DTI ratio is 13.31%, with a minimum of 0% and a maximum of 29.99%. This suggests that most borrowers have a moderate level of debt relative to their income, though there are some outliers with very high DTI ratios.

  - Inquiries in Last 6 Months: On average, borrowers have approximately 0.87 credit inquiries in the last 6 months, with a standard deviation of 1.07. The minimum number of inquiries is 0, and the maximum is 8, indicating varying levels of recent credit activity among the borrowers.

  - Delinquency in Last 2 Years: The data shows a mean delinquency rate of 0.15 over the last 2 years, with a standard deviation of 0.49. Most borrowers have had no delinquencies, but there are some with up to 11 delinquencies, indicating that while most borrowers manage their credit well, there are some with significant credit issues.

- Installment Payments: The installment payments (monthly payments on the loan) have a mean of $324.56 with a standard deviation of $208.87. The minimum installment is $15.69, and the maximum is $1,305.19. This indicates a wide range of monthly payments among the borrowers.

- Funded Amount by Investors: The amount funded by investors is slightly lower on average than the loan amount, with a mean of $10,397.45 and a standard deviation of $7,128.45. This suggests that there are some loans which were not fully funded by investors.

- Public Record Bankruptcies: There is a very low incidence of public record bankruptcies among the borrowers, with a mean of 0.043 and a standard deviation of 0.204. Most borrowers have no bankruptcies on their record, but a small number have up to 2 bankruptcies.

- Credit Inquiries and Delinquency Rates: The average number of credit inquiries in the last 6 months is 0.87, and the delinquency rate in the last 2 years is low at 0.15. This suggests that most borrowers are not frequently seeking new credit and have low rates of delinquencies, indicating relatively good credit health.

- Total Balances Excluding Mortgages: The column for total_bal_ex_mort is not populated (NaN values), indicating that there may be missing or incomplete data for the total balance excluding mortgages. This could be an area that requires further investigation or data cleaning.

- 6: unique values :

  - High Number of Unique Values in Employee Title: The emp_title column has 28,820 unique values, indicating a very diverse set of job titles among the borrowers. This high variability can make it challenging to analyze without some form of categorization or grouping.

  - Unique Identifiers and Descriptions: The url and desc columns have 39,717 and 26,526 unique values, respectively. The url likely serves as a unique identifier for each loan, while the desc provides unique descriptions, which might be valuable for text analysis but could also introduce significant variability.

  - Loan-related Categories: Columns like term, grade, and sub_grade have relatively few unique values (2, 7, and 35, respectively), making them suitable for categorical analysis. These columns can provide insights into loan terms and borrower creditworthiness and can be easily used for grouping and summarization in analysis.

- 7: on loan_status column :

  - Majority of Loans Fully Paid: The majority of the loans, 32,950 out of 39,717, are fully paid. This suggests that a significant portion of the borrowers were able to meet their loan obligations successfully.

  - Charged Off Loans: There are 5,627 loans that have been charged off. Charged off loans represent cases where the lender has given up on collecting the owed amount, indicating a failure in loan repayment. This accounts for about 14.2% of the total loans, which is a notable proportion and highlights the risk associated with lending.

  - Current Loans: There are 1,140 loans that are currently active and being repaid. This shows that there is a small portion of the dataset where the loans are still in the repayment phase. Monitoring these loans can provide insights into ongoing borrower behavior and potential future defaults or full repayments.

# Data Cleaning and Manipulation

- Steps Performed:
  - Identify Columns with Significant Missing Data
  - Drop columns with significant missing data
  - chanding vlaues of columns into required format
  - Converting columns into corresponding correct DTypes
  - Handling rows with missing values
  - OutLier Treatment

- # Observations:
- 1: For handling missing values:

  - 50% missing columns are dropped which resulted in 54 columns remaining
  - Handle emp_title with 'Unknown': Fills missing values in the emp_title column with 'Unknown' to handle incomplete data and ensure consistency.
  - Handle emp_length with median imputation: Fills missing values in the emp_length column with the median value. Median imputation is used here because the median is less sensitive to outliers compared to the mean, providing a robust estimate for missing values.
  - Drop the desc column: Drops the desc column because it contains a large number of missing values. Removing this column simplifies the dataset and avoids potential biases that could arise from imputing or including incomplete data.
  - Handle title with 'Unknown': Fills missing values in the title column with 'Unknown'. Similar to emp_title, this step ensures that all records have a value for the title column.
  - Impute revol_util with mean: Fills missing values in the revol_util column with the mean value. Mean imputation is used here to estimate missing values based on the average utilization rate, providing a reasonable estimate for incomplete data.
  - Drop rows with missing last_pymnt_d and last_credit_pull_d: Drops rows where either last_pymnt_d or last_credit_pull_d has missing values. Since these columns are date-related and crucial for analysis, dropping rows with missing dates ensures data integrity and consistency.
  - Impute collections_12_mths_ex_med, chargeoff_within_12_mths, and tax_liens with mode: Fills missing values in these columns with the mode (most frequent value). Mode imputation is suitable here to handle categorical or count data where the most common occurrence provides a reasonable estimate for missing values.

- 2: changing values of columns into required format:
    - Term Column: Extracts the numeric part of the term column and converts it to an integer.
    - Interest Rate Column: Removes the '%' sign from the int_rate column values and converts them to floats.
    - Revolving Utilization Column: Removes the '%' sign from the revol_util column values and converts them to floats.
    - Date Columns:Converts the issue_d, earliest_cr_line, last_pymnt_d, and last_credit_pull_d columns to datetime format using the specified date format.
    - Employment Length Column:Replaces specific values in the emp_length column and then extracts the numeric part, converting it to a float.
    - Public Record Bankruptcies Column: Fills any missing values in the pub_rec_bankruptcies column with 0 and converts the values to integers.
    - Tax Liens Column: Fills any missing values in the tax_liens column with 0.0.

# Univariate Analysis

- Steps Performed:
  - Univariate analysis of Numerical columns
  - Univariate analysis of Categorical columns

- **Observations**:
  - Loan Amounts and Interest Rates:
    - Borrowers typically request loans around $10,000 to $11,000, with some taking larger amounts, leading to a right-skewed distribution.
    - Interest rates average around 12%, with relatively stable rates across loans.

  - Borrower income and Financial Capability:
    - Average borrower income is approximately $65,286, but with a wide range indicating diverse financial capacities.
    - Positive skewness suggests some borrowers have significantly higher incomes.

- Repayment Behaviour:
  - Borrowers make substantial repayments, reflected in moderate to high amounts like total_pymnt and total_rec_prncp.
  - Skewed distributions indicate occasional larger payments or interest amounts.

- Credit Metrics:
  - Credit profiles, measured by open_acc and total_acc, are moderately varied
  - revol_bal shows positive skewness, indicating some borrowers have higher revolving balances.

- Employment and Debt Ratio:
  - Borrowers typically have around 5 years of employment (emp_length).
  - Debt-to-income ratio (dti) skewness suggests most borrowers manage debt relative to their income effectively.

- Borrower Preferences:
  - There is a clear preference for shorter loan terms, indicating sensitivity to interest costs and a desire for quicker repayment.

- Risk Distribution:
  - Higher-grade loans (A and B) dominate, suggesting a generally low-risk borrower profile in the dataset.

- Home OwnerShip:
  - Most borrowers rent or have a mortgage, with fewer owning outright or falling into other categories.

- Verification Status:
  - Majority of loans are not verified, followed by verified and source verified.

- Loan Status:
  - A large majority of loans are fully paid, with a significant portion charged off and a few still current.

- Payment Plan:
  - Almost all loans do not have a payment plan.

- Loan Purpose
  - Debt consolidation and credit card usage are the primary reasons for loans.

- Geographic Distribution:
  - Loans are concentrated in California, New York, and Florida, with varying counts across other states.

# Segmented Univariate Analysis

- Steps Performed:
  - segmented analysis of numerical variables with loan_status
  - segmented analysis of categorical variables with loan_status

- **Observations**:
  - Loan Term:
    - Charged Off: Most loans with a 36-month term are charged off compared to those with a 60-month term.
    - Fully Paid: Majority of loans that are fully paid have a 36-month term rather than a 60-month term.
    - Current: Loans with a 60-month term are more common among current loans than those with a 36-month term.

- Loan Grade:
  - Charged Off: Grades B, C, and D have the highest counts for charged off loans.
  - Fully Paid: Grades B and A have the highest counts for fully paid loans.
  - Current: Grade B is the most common grade among current loans.

- Home Ownership:
  - Charged Off: Renters have the highest count of charged off loans, followed by those with mortgages.
  - Fully Paid: Renters also have the highest count of fully paid loans, followed closely by those with mortgages.
  - Current: Mortgages are most common among current loans.

- Analysis underscores the critical role of loan size, interest rates, installment amounts, income levels, DTI ratios, and credit utilization in predicting loan outcomes. Borrowers with higher loan amounts, interest rates, and installment obligations, coupled with lower incomes and higher debt burdens, appear to face increased risks of defaulting on their loans. Understanding these factors can help lenders assess and manage credit risk more effectively, potentially reducing default rates and improving overall loan portfolio performance.

# Key Inferences

Based on the analysis, these are the strong indicators of default  are

- 'int_rate'
- term'
- 'grade'
- 'sub_grade'
- 'verification_status'
- 'purpose'
- 'annual_inc'

# Bivariate Analysis

- Steps Performed:
  - Numerical vs Numerical
  - Numerical vs Categorical
  - Categorical vs Categorical

- Observations:
  - int_rate (interest rate) positively correlates with loan amounts and installments.

  - Higher annual_inc (annual income) correlates with larger loan requests.

- revol_bal (revolving balance) and revol_util (utilization rate) show moderate correlations, suggesting higher balances relate to increased credit utilization.

- open_acc (open credit lines) correlates positively with total_acc (total credit lines), indicating a relationship between active and total credit accounts.

- last_pymnt_amnt (last payment amount) correlates with total payment and interest received, reflecting its role in payment behaviors and dynamics.

- Loan Amounts: Highest loan amounts are generally associated with verified status borrowers, particularly in higher grade loans (A to E).

- Interest Rates: Lower grade loans (D and E) typically incur higher interest rates compared to higher grade loans (A to C).

- Annual Income: Variations in annual income are noticeable across all loan grades, home ownership types, and verification statuses, indicating diverse borrower demographics.

- on loan_status:
  - Charged Off Loans: Higher loan amounts and interest rates, with indications of financial strain based on higher DTI and revolving balances.
  - Current Loans: Higher current income and employment length, with moderate DTI.
  - Fully Paid Loans: Lower loan amounts and interest rates, with better DTI ratios and stable repayment behavior.

- All pairs (term vs. grade, purpose vs. verification_status, home_ownership vs. addr_state, grade vs. verification_status) show statistically significant associations based on the chi-square tests performed.

- These associations indicate that understanding one variable within each pair can provide insights into the distribution or status of the other variable.

# Derived Metrics

- Business-driven Metrics
  - Debt-to-Income Improvement (dti_improvement): Measures how much of the borrower's annual income (annual_inc) has been used to repay the loan (total_pymnt). Higher values indicate better debt management relative to income.
  - Annual Installment Percentage (annual_installment_pct): Calculates the percentage of annual income (annual_inc) dedicated to loan installments (installment). This metric gauges the affordability and financial burden of loan payments on borrowers.

- Type-driven Metrics
  - Log Transformation (log_annual_inc, log_revol_bal): Applies a natural logarithm transformation (np.log1p) to skewed numerical variables (annual_inc, revol_bal). This transformation helps normalize the distribution of these variables for better statistical modeling.
  - Ordinal Encoding (grade, sub_grade): Converts categorical variables (grade and sub_grade) into ordinal integers using OrdinalEncoder. This encoding preserves the order information inherent in these categories, facilitating their use in mathematical models.

- Data-driven Metrics

  - Interaction Terms (int_rate_dti, loan_amnt_annual_inc): Creates new variables based on interactions between existing variables:
  - int_rate_dti: Multiplies the interest rate (int_rate) by the debt-to-income ratio (dti). This captures the combined impact of interest rates and debt burden on loan performance.
  - loan_amnt_annual_inc: Computes the ratio of loan amount (loan_amnt) to annual income (annual_inc). This metric assesses the loan size relative to borrower income, influencing loan risk assessment.

- These metrics collectively enhance the understanding of borrower profiles, loan affordability, and risk factors within the dataset. They are instrumental in both business decision-making and analytical modeling, providing insights into loan repayment behaviors and financial health.

# Conclusion

- Summary of Analysis:
  - Key factors identified for lending decisions.
  - Effective data cleaning and analysis methods.