# LOANTAP PROJECT

**Index:**

1. Introduction to the project
2. Objectives of the Project
3. Methodology
4. Answering the Questions based on Analysis.
5. Observations
6. Conclusion

**Introduction to the project:**

The Loan Tap dataset provides a comprehensive view of loan-related information collected from borrowers applying through the Loan Tap online platform. This dataset encompasses diverse aspects of loan applications, borrower profiles, and financial attributes crucial for risk assessment and loan management.

Key Features Include:

loan\_amnt: The amount initially requested by borrowers.

term: The duration of loan repayment in months (36 or 60 months).

int\_rate: The interest rate applied to loans.

installment: Monthly payment obligations for borrowers.

grade and sub\_grade: Assigned loan grades and subgrades by Loan Tap.

emp\_title and emp\_length: Borrowers' job titles and employment lengths.

home\_ownership: Status of home ownership as reported by borrowers.

annual\_inc: Self-reported annual income of borrowers.

verification\_status: Verification status of borrowers' income by Loan Tap.

issue\_d: The month in which loans were funded.

loan\_status: Status of loans, serving as the target variable.

purpose: Borrowers' stated purposes for loan requests.

dti: Debt-to-income ratio, a financial metric for borrowers.

earliest\_cr\_line: Date when borrowers' earliest credit lines were opened.

open\_acc: Number of open credit lines in borrowers' credit files.

pub\_rec: Count of derogatory public records associated with borrowers.

revol\_bal and revol\_util: Revolving credit balances and utilization rates.

total\_acc: Total number of credit lines in borrowers' credit files.

initial\_list\_status: Initial listing status of loans (W for whole loans, F for fractional loans).

application\_type: Type of loan application (individual or joint).

mort\_acc: Number of mortgage accounts reported by borrowers.

pub\_rec\_bankruptcies: Number of public record bankruptcies.

This dataset serves as a valuable resource for data analysis and mining, offering insights into borrower demographics, loan characteristics, and risk factors essential for decision-making processes in the lending domain.

**Objectives of the project:**

Given the complexity of the dataset with numerous columns, the analysis can reveal insights into various attributes through multivariate analysis, particularly when visualized. My primary objectives are as follows:

1. To understand the distribution of loan amounts across different purposes, grade, subgrade, and employment length, providing insights into funding needs for each aspect. This includes comparing average loan amounts, total loan amounts, and percentage contributions across loan purposes to identify trends.
2. Analysing the percentage contribution of each aspect to the total loan portfolio, highlighting the relative significance of variables in loan distribution.
3. Identifying borrowing patterns, such as loan amount utilization, interest rates, and their relation to borrowers' annual income, to understand market demand and financial priorities.
4. Generating insights into borrower preferences, financial needs, and market demand for different loan purposes, aiding lenders in decision-making and targeted product design.

**Methodology:**

Data Cleaning and Transformation:

EDA is the first step to understand the dataset before performing analysis and visualization. After doing a thorough exploratory data analysis on the structure of the dataset, I encounter multiple correction viewpoint in multiple columns.

Data Cleaning and Data Transformation plays a crucial and very prominent role in giving a shape to the dataset that the dataset will become more reliable and consistent. As the quote says “garbage in, garbage out”, emphasizing that if you input low-quality or inaccurate data, the output or analysis based on that data will also be flawed or unreliable. In other words, the quality of the output is directly dependent on the quality of the input data.

1. Handling missing data:
   1. For the **"emp\_title"** column, which had over 22k missing values, instead of replacing them with the mode of employee titles, I opted to replace NaN values with 'Unknown' to prevent data loss.
   2. Similarly, for the **"emp\_length"** column with over 18k missing values and being a categorical column, I used the same treatment as emp\_title and replaced NaNs with 'Unknown.'
   3. In the **"mort\_acc"** column, with 37k missing values and being a numerical column, I initially considered replacing them with the median. However, due to concerns about the significant impact this might have on other aspects of the dataset, I chose to drop the NaN values instead.
   4. Regarding the **"pub\_rec\_bankruptcies"** column, which denotes the number of public record bankruptcies, replacing NaNs could artificially enhance the perceived completeness of the dataset, potentially masking underlying data collection or recording issues.
2. Feature Engineering:
3. extraction of year and month from the column that has the dates for the commencement as well as the termination of the loan after converting those columns’ datatype as datetime for further analysis.
4. extracting state from the address column.
5. from term column, extracting the digits only.

C. Rectifying the columns datatypes of numerical columns and dropping irrelevant columns.

D. Outlier Treatment: identifying the outliers in every numerical column and then applying quantile method (0.25 to 0.75 quantile) to remove outliers from the necessary columns.

1. Data Visualization: Data Visualization is the best way to replicate figures into visual form to display the structure of the variables and to get an initial idea of the dataset. After cleaning the data and dealt with the necessary column aspects, I have presented the data in different ways:
2. Count plot of category variables with Pie plot of them to see the division of variables in each column.
3. And histogram and distribution plot of numerical columns to show the counts and dispersion of them.

**Question and Answer based on the Analysis:**

**Ques. How many categorical and numerical columns are there in the dataset?**

A. Initially, there were 12 numerical columns and 15 object datatype columns in the dataset. After transformation and cleaning, the figures changed to 19 numerical columns and 12 object datatype columns.

**Ques. What correlations have been observed among different columns and what do they indicate?**

Several correlations have been observed:

Loan Amount (loan\_amnt) and Installment (installment): strong positive correlation (0.955), indicating that as the loan amount increases, the monthly installment amount also increases significantly.

Loan Amount (loan\_amnt) and Annual Income (annual\_inc): moderate positive correlation (0.502), suggesting that borrowers with higher annual incomes tend to apply for larger loan amounts.

Term and loan\_amnt: moderate positive correlation (0.40), indicating a tendency for longer-term loans to have larger loan amounts.

Term and int\_rate: moderate positive correlation (0.43), suggesting that longer-term loans tend to have slightly higher interest rates.

Loan Amount (loan\_amnt) and Revolving Balance (revol\_bal): moderate positive correlation (0.327), implying that borrowers with higher revolving balances may seek larger loan amounts.

Loan Amount (loan\_amnt) and Total Accounts (total\_acc): moderate positive correlation (0.217), suggesting that borrowers with more total accounts may apply for larger loan amounts.

Installment (installment) and Annual Income (annual\_inc): moderate positive correlation (0.485), indicating that borrowers with higher incomes may have higher monthly installment amounts.

Open Accounts (open\_acc) and Total Accounts (total\_acc): very strong positive correlation (0.994), showing that the number of open accounts is highly correlated with the total number of accounts.

Credit Line Delta (credit\_line\_delta) and Earliest Credit Line Quarter (earliest\_cr\_quarter): strong negative correlation (-0.988), suggesting an inverse relationship between the difference in credit lines and the quarter when the earliest credit line was opened.

**Ques. What observations can be made with respect to grades and subgrades?**

The dominant grades in the dataset are B (29%), C (27%), D (16%), and A (15%). The most dominant subgrades are from grades B and C, encompassing about 57% of the grade distribution. As grades decrease, subgrades also decrease.

**Ques. Is there a significant difference in borrower behaviors or creditworthiness based on the difference between issue year and earliest credit line year?**

Yes, there is a credit\_line\_delta column showing the difference between earliest\_cr\_line and issue\_year, ranging from 3 years to over 70 years with a mean of 15 years, indicating varied loan durations.

**Ques. How do loan amounts, interest rates, and installment amounts vary for different aspects of the borrowers?**

Loan amounts, interest rates, and installment amounts vary based on employment length, home ownership, loan grades, subgrades, verification status, loan purposes, and loan status.

**Ques. What observations can be made on borrowers’ attributes in the context of the loan status attribute?**

Borrowers’ attributes such as employment length, home ownership, loan purposes, loan grades, and subgrades show variations in loan status (fully paid or charged off), indicating trends in repayment behavior and creditworthiness.

**Ques. What observations can be made on borrowers’ attributes in the context of the term of the loan?**

Borrowers’ attributes such as employment length, home ownership, loan purposes, loan grades, and subgrades influence the choice of loan term (36 months or 60 months), reflecting financial stability, risk perception, and repayment preferences.

**Ques. What type of trends can be noticed with pub bankruptcies and different aspects of borrower’s attributes?**

A. Bankruptcy trends show variations over time and across borrower attributes, with employment length and loan purposes contributing significantly to bankruptcy cases.

**Ques. What trends can be observed with debt-to-income ratio?**

A. Debt-to-income ratios vary based on loan grades, subgrades, employment length, loan purposes, and loan status, indicating varying levels of debt management and financial stability among borrowers.

**Ques. Explain the terms: open account and total account. Also, explain their correlation.**

A. Open\_acc refers to the count of open credit lines, while total\_acc represents the total number of credit lines, including both open and closed accounts. They have a very strong positive correlation (0.99), indicating a close relationship between the number of open accounts and the total number of accounts.

**Ques. How much data imbalance is there in the loan status?**

There is an imbalance in loan status, with about 80% of borrowers fully paying the loan and the remaining 20% falling into the charged-off category.

**Observations:**

**Correlations:**

* 1. Strong positive correlations were observed between loan amounts and installment amounts, indicating higher loan amounts correspond to higher monthly installments.
  2. Moderate positive correlations were noted between loan amounts and annual incomes, as well as between loan terms and loan amounts.
  3. Home ownership, employment length, loan grades, subgrades, and verification status correlated with financial metrics such as annual incomes, installment amounts, and interest rates.

**Loan Grades and Subgrades:**

* 1. Grades B and C were dominant, encompassing about 57% of the grade distribution.
  2. Lower subgrades within each grade category exhibited higher default risks and lower repayment rates.

**Borrower Behaviors and Creditworthiness:**

1. Employment length influenced annual incomes and loan installment amounts but showed standardized interest rates across categories.
2. Homeownership statuses correlated with income levels, financial obligations, and loan stability.
3. Higher loan grades indicated better creditworthiness, lower installment amounts, and lower interest rates.
4. Verified loans exhibited higher installment amounts, interest rates, and income variability.

**Loan Purposes and Financial Characteristics:**

1. Different loan purposes showed varying repayment rates and default risks, with educational loans exhibiting the highest repayment rates.
2. Loan purposes impacted interest rates, with renewable energy loans carrying higher rates.

**Loan Status and Borrower Attributes:**

1. Fully paid loans were more common across different employment lengths, indicating higher repayment reliability with longer employment.
2. Homeownership status influenced charged-off loan rates, with property owners showing lower rates.
3. Loan grades and subgrades correlated with loan status, with lower grades indicating higher charged-off rates.

**Term Preferences:**

1. Borrowers generally preferred 36-month loan terms, with variations based on employment length, home ownership, loan purposes, loan grades, and subgrades.
2. Higher-grade loans tended to opt for shorter terms, while lower-grade loans leaned towards longer terms.

**Bankruptcy Trends**:

1. Bankruptcy cases varied across employment lengths and loan purposes, with debt consolidation and credit card issues contributing significantly.

**Debt-to-Income Ratios:**

1. DTIs varied based on loan grades, subgrades, employment length, loan purposes, and loan status, reflecting differing levels of debt management and financial stability.

**Conclusion: Leveraging Insights for Informed Decisions**

The comprehensive analysis of LoanTap's dataset has unearthed valuable insights into borrower behaviours, creditworthiness indicators, and loan performance metrics. These insights serve as a cornerstone for making informed decisions, optimizing lending strategies, and enhancing risk management practices. By delving into various aspects such as loan amounts, interest rates, instalment amounts, loan grades, subgrades, borrower attributes, loan purposes, loan status, term preferences, bankruptcy trends, debt-to-income ratios, and correlations among different variables, we have painted a detailed picture of the lending landscape.

One of the key takeaways from this analysis is the strong correlation between loan amounts and instalment amounts, indicating a predictable pattern where higher loan amounts lead to higher monthly instalments. This understanding can guide loan structuring and repayment planning for both borrowers and lenders. Additionally, the insights into loan grades and subgrades reveal distinct risk profiles associated with different borrower segments, highlighting the importance of tailored risk assessment strategies.

The analysis also sheds light on the impact of borrower attributes such as employment length, home ownership status, and verification status on financial metrics and loan performance. This knowledge empowers LoanTap to design targeted loan products, offer personalized terms, and enhance customer experiences. Moreover, the observations regarding loan purposes and their financial characteristics provide actionable intelligence for pricing loans, managing default risks, and aligning product offerings with market demands.

Furthermore, the findings related to loan status, term preferences, bankruptcy trends, and debt-to-income ratios offer strategic insights for optimizing portfolio performance, mitigating credit risks, and fostering sustainable growth. By leveraging these insights, LoanTap can refine its underwriting criteria, fine-tune risk pricing models, and proactively address potential challenges in the lending ecosystem.

In conclusion, the analysis serves as a compass for navigating the dynamic landscape of lending, enabling LoanTap to stay ahead of market trends, mitigate risks effectively, and deliver value-driven financial solutions to its customers. Continued data analysis, monitoring of key performance indicators, and iterative refinement of lending strategies will be instrumental in sustaining competitive advantage and fostering long-term success in the lending industry.