Loan Data Analysis

Exploration, Cleaning, and Insights

Name : Bryan Japheth Charles
Subject: Statistics and Exploratory Data Analysis
College: Bharath Institute of Higher Education And Research

Table Of Content

1. INTRODUCTION

Purpose of the Project:

•Explore, clean, and analyze a loan dataset.
•Understand borrower characteristics, loan funding details, repayment performance, and risk factors.

Analysis Approach:

- **Data Cleaning:** Handle missing values, duplicates, and inconsistencies.
- **Feature Engineering:** Create new features like debt ratio, principal/interest paid ratios, and credit history length.
- **Exploratory Data Analysis (EDA):** Visualize loan trends, borrower profiles, and repayment behavior.
- **Insights & Patterns:** Identify trends in loan issuance, borrower demographics, and financial risk.

2. Libraries and Tools Used

Library / Tool	Purpose
pandas	Data manipulation
numpy	Numerical computation
matplotlib	Data visualization (plots)
seaborn	Statistical & aesthetic plots
Tableau	Interactive dashboards

3. Dataset Overview

3.1 Dataset Dimensions:

Stage	Rows	Columns	Data Size (bytes)
Initial	39,717	111	4,408,587

3.2 Data Types:

Data Type	Description
Numerical	Continuous / integer values
Categorical	Labels or categorical values
Date	Timestamps / loan issue dates

4. Data Cleaning

4.1 Missing Values

- Checked for null values for all the 111 columns.
- Got 68 columns containing missing values.
- Dropped 54 columns with 100% missing values.
- Removed columns with >30% missing data.
- The columns which had < 30% was filled using fillna().
- Categorical missing values filled with "Unknown" or mode().
- Numerical missing values filled using mean imputation.

4.2 Duplicates & Dates

- Checked for duplicate rows, none found.
- Converted date columns to datetime format: ‘issue_d’, ‘earliest_cr_line’, ‘last_pymnt_d’.
- Converted string-based numeric columns to numeric: ‘int_rate’, ‘emp_length’.

5. Feature Reduction

- Dropped irrelevant and redundant columns to finalize cleaned data

Categorized columns after cleaning:

- **Discrete columns:**
  'delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal',
  'total_acc', 'total_rec_prncp', 'total_rec_int', 'pub_rec_bankruptcies'.

- **Continuous columns:**
  'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'annual_inc', 'installment',
  'dti', 'last_pymnt_amnt', 'total_pymnt', 'total_pymnt_inv', 'revol_util'.

- **Categorical columns:**
  'grade', 'home_ownership', 'verification_status', 'loan_status', 'purpose', 'term'.

- **Categorical to numeric:**
  'int_rate', 'emp_length'.

- **Date columns:**
  'issue_d', 'earliest_cr_line'.

6. Feature Engineering

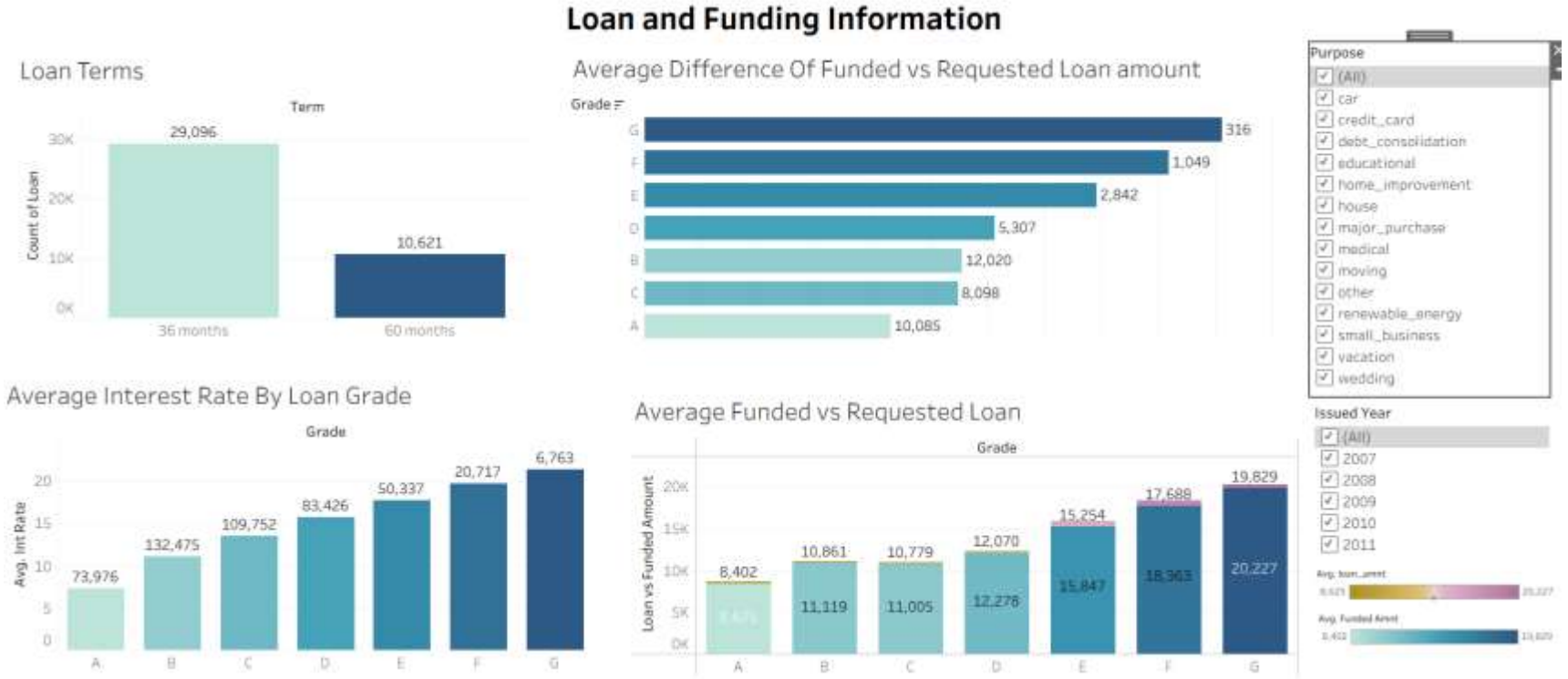
- Took the cleaned data, and create some new columns for more insights

Derived the following features:

- principal_paid_ratio = total_rec_prncp / loan_amnt
- interest_paid_ratio = total_rec_int / loan_amnt
- debt_ratio = installment / annual_inc
- credit_history_years = issue_d.year - earliest_cr_line.year
- delinq_flag = Categorized delinquency
- Converted the emp_length column to the following values-

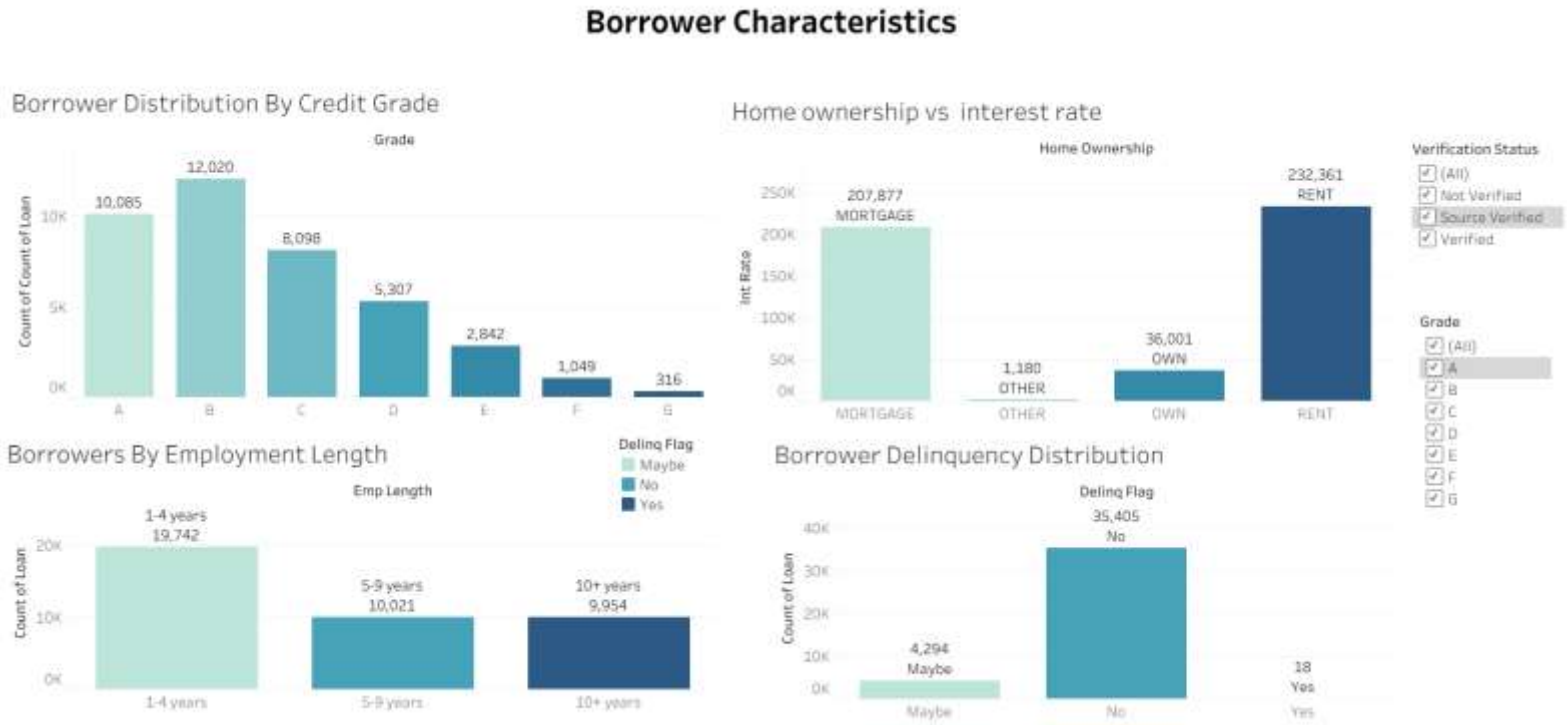
  "<1 year" → 0 years
  "1-4 years" → 1 to 4 years
  "5-9 years" → 5 to 9 years
  "10+ years" → 10 or more years

7. Loan & Funding Analysis



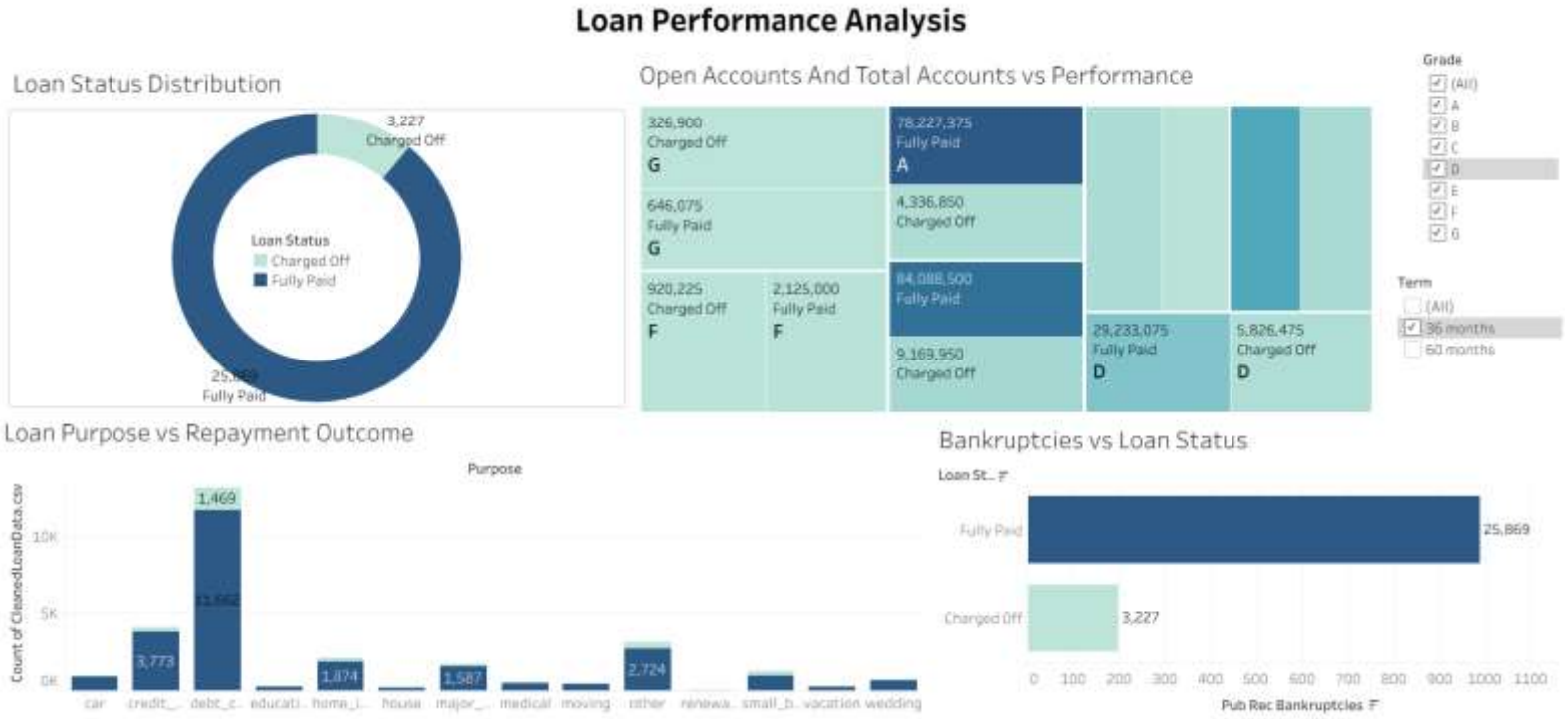
- Distribution of loan amounts requested by borrowers
- Most common loan terms
- Average funded vs requested loan amount
- Difference between funded amount (lenders vs investors)
- Percentage of loans issued per term
- Average interest rate by loan grade
- Average installment amount across loan categories

8. Borrower Characteristics Analysis



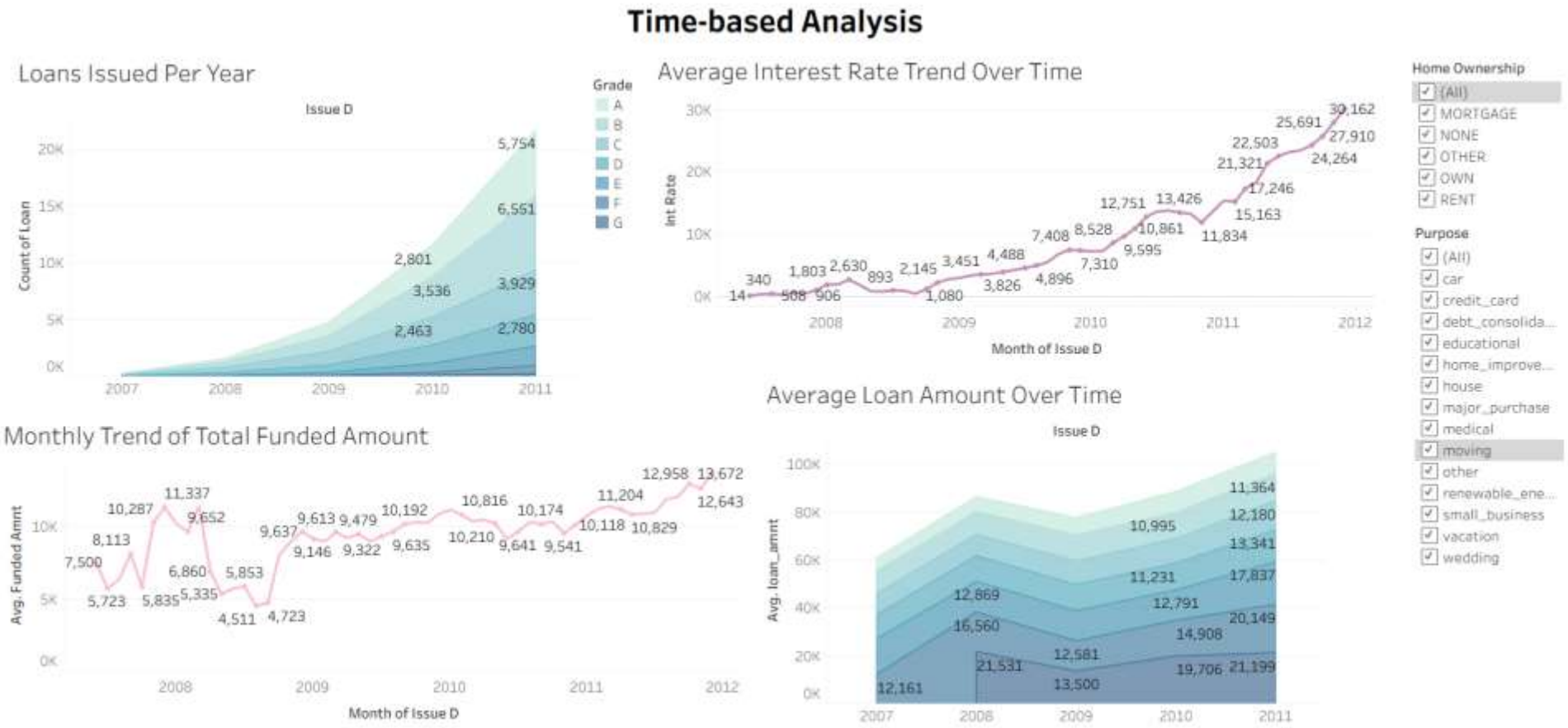
- Borrower distribution by credit grade
- Home ownership vs loan approval / interest rate
- Income verification vs funding & interest
- Borrower delinquency distribution
- Borrowers by employment length

9. Loan Performance & Status



- Loan status distribution
- Loan purpose vs repayment outcome
- Debt-to-Income Ratio (DTI) vs default risk
- Delinquencies in last 2 years vs repayment
- Credit inquiries (last 6 months) vs default
- Open accounts & total accounts vs performance
- Average repayment by loan grade
- Bankruptcies vs loan status

10. Time-Based Analysis



- Loan issue date vs default trends
- Loans issued per year
- Average loan amount over time
- Loan purpose distribution over time
- Monthly trend of total funded amount

11. Conclusion

Dataset cleaned: missing values handled, duplicates removed, irrelevant columns dropped, dates/numerics standardized.
Feature engineering: created debt ratio, principal/interest paid ratios, credit history years, delinquency flag for deeper insights.
Key borrower insights: mid-level credit grades dominate; verified income & home ownership improve funding; higher DTI & recent delinquencies lead to higher default risk.
Loan trends: standard term loans most common; funded amounts slightly lower than requested; average loan amounts increasing over time.
Correlation & ratios: strong positive correlation between loan, funded amount, and installment; principal/interest ratios and debt ratio effectively indicate repayment patterns.

Overall, analysis reveals patterns in loan issuance, borrower demographics, repayment behavior, and supports credit risk assessment and predictive modeling.