Problem Statement

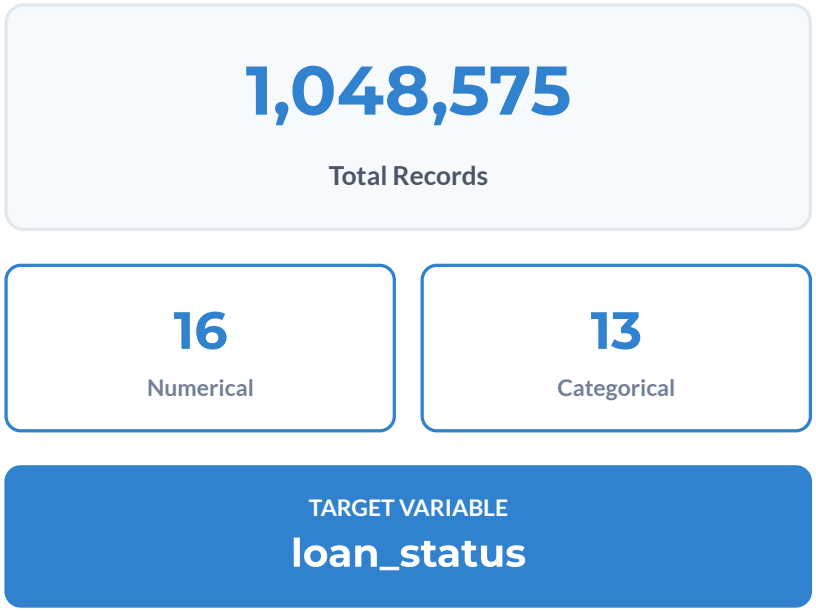
- What borrower and account attributes align with higher repayment risk?
- How are outcomes distributed across BNPL use cases (purposes)?
- Which signals (pricing, limits, DTI) co-occur with delinquency/charge-off?
- Where are data quality issues that can bias risk insights?

Objectives

- Summarize distributions of key financial and behavioral variables
- Examine bivariate and multivariate relationships with loan_status
- Quantify segment differences via pivots (purpose, home_ownership, application_type)
- Establish clean, analysis-ready data with documented assumptions
- Surface actionable insights for BNPL risk policy and operations

Dataset Overview

- **Data source:** Kaggle - BNPL Dataset (v1)
- **Records:** 1,048,575; **Variables:** 29 (numeric: 16, categorical: 13)
- **Core features:** loan_amount, interest_rate, loan_term, monthly_payment
- **Credit profile:** annual_income, total_dti, credit_limit, total_bal_ex_mort
- **Behavior/history:** delinq_2yrs, num_accts_120_pd, mort_acc
- **Target:** loan_status (Current, Fully Paid, Charged Off, etc.)



Data Cleaning & Preparation

- **Parsed text fields:** loan_term → int (months); emp_length → numeric (years)
- **Missing values:** emp_length, total_dti imputed with median; emp_title dropped (8.6% missing)
- **Duplicates:** None found; types validated for analysis
- **Outliers:** IQR capping applied to stabilize extreme financial values
- **Result:** Clean, analysis-ready dataset with 1,048,575 records

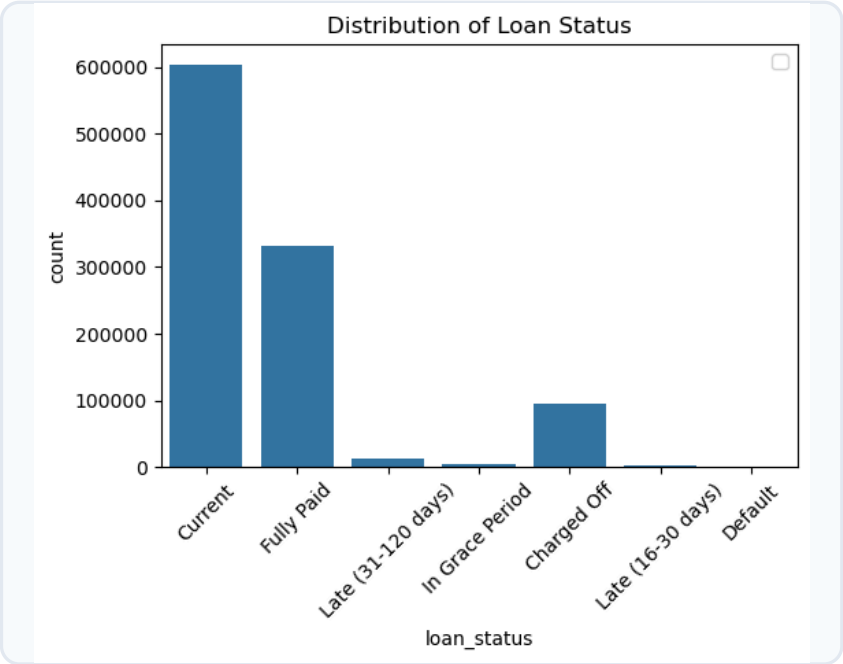
MINIMAL CODE EXAMPLE

```
# loan_term
df['loan_term'] =
df['loan_term'].str.replace('months',
'').astype(int)

# emp_length
df['emp_length'] = (df['emp_length']
    .str.replace('years', '')
    .str.replace('+', '')
    .astype(float)
    .fillna(df['emp_length'].median()))

# Handle missing values
df.drop(columns=['emp_title'], inplace=True)
df['total_dti'].fillna(df['total_dti'].median(),
inplace=True)
```

Univariate Analysis – Target Variable (Loan Status)



- Key Insight

  Imbalanced distribution: **Current** and **Fully Paid** dominate the portfolio

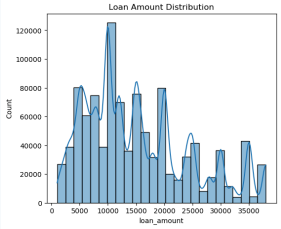
- Adverse outcomes (**Charged Off**, **Default**, **Late**) form minority classes

- **Analytical implication:** Evaluate risk using rates and ratios, not raw counts

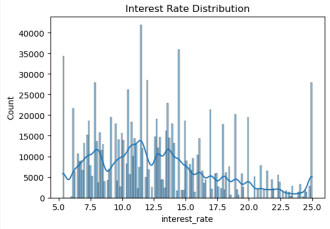
- Class imbalance reflects real-world lending but must be handled carefully downstream

Univariate Analysis – Numerical Variables

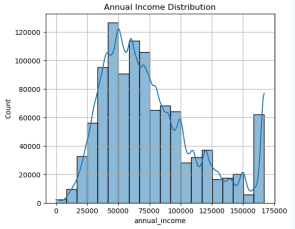
Loan Amount Distribution



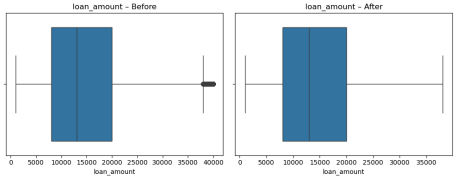
Interest Rate Distribution



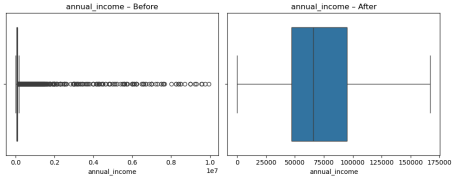
Annual Income Distribution



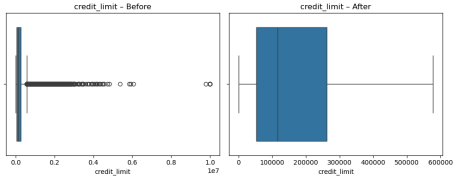
Loan Amount: Before/After Outlier Treatment



Annual Income: Before/After Outlier Treatment



Credit Limit: Before/After Outlier Treatment

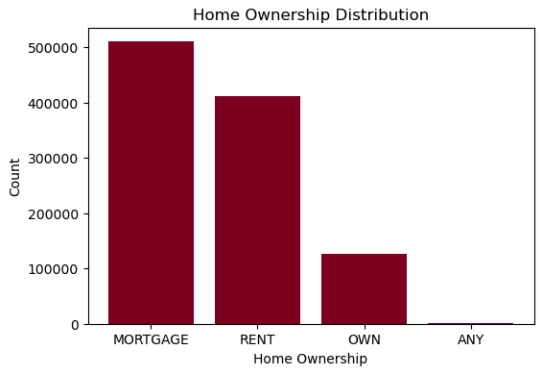


Key Observations

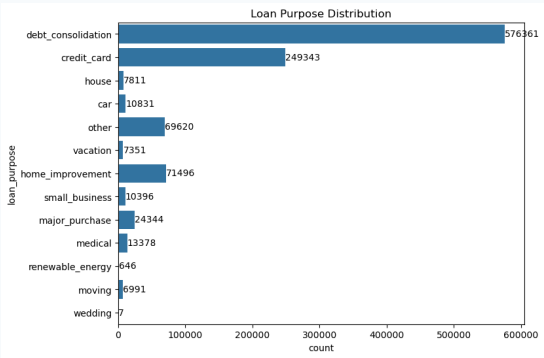
- **Right-skewed distributions:** Loan amounts, interest rates, and annual income show concentration in lower-to-mid ranges with long right tails
- **Outlier capping effect:** IQR-based capping preserved medians and quartiles while eliminating extreme values, improving distribution quality without shifting central tendency
- **Common loan amounts:** Peaks around $5k, $10k, and $20k indicate preferred loan sizes; interest rates cluster in 10-15% band

Univariate Analysis – Categorical Variables

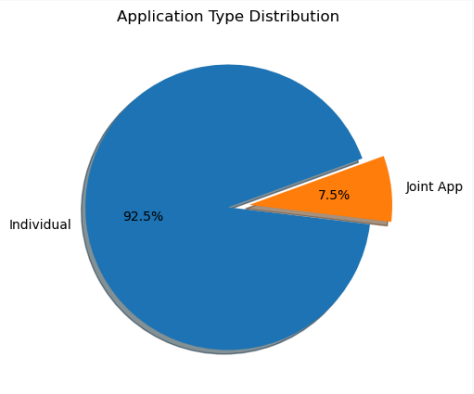
Home Ownership Distribution



Loan Purpose Distribution



Application Type Distribution

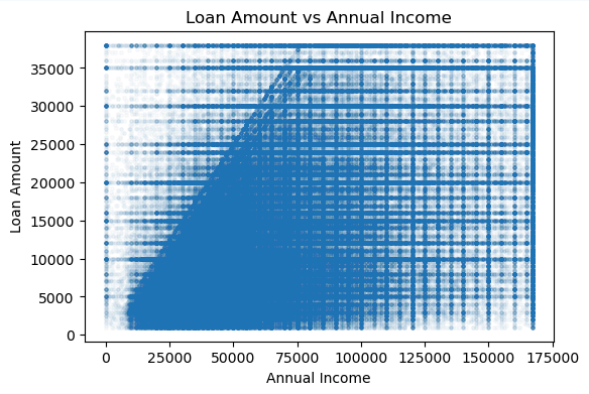


Key Observations

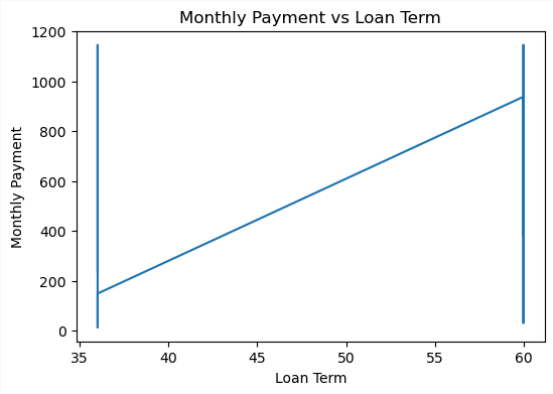
- **Home Ownership:** Mortgage (48.6%) and Rent (39.3%) dominate, with Own (12.1%) as a smaller segment; housing status indicates reliance on credit among non-homeowners
- **Loan Purpose:** Debt Consolidation (55%) and Credit Card (23.8%) account for nearly 80% of loans, highlighting borrowers' tendency to restructure existing debt
- **Application Type:** Individual applications dominate at 92.5%, with joint applications forming only 7.5%; portfolio mix is highly concentrated in specific categories

Bivariate Analysis – Numerical vs Numerical

Loan Amount vs Annual Income



Monthly Payment vs Loan Term

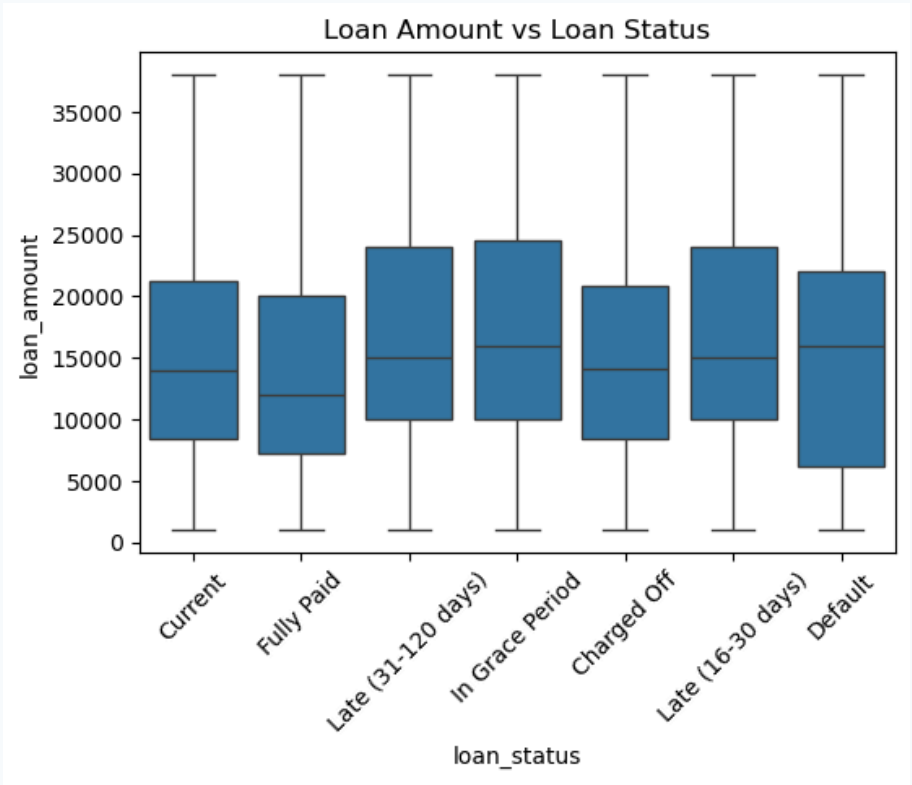


Key Observations

- **Moderate positive relationship:** Higher income borrowers tend to take larger loans, though considerable variability exists across all income levels
- **Heteroscedasticity observed:** Wide spread in loan amounts suggests additional controls are needed (purpose, credit limits) to explain loan size beyond income alone
- **Payment-term dynamics:** Monthly payments increase with loan term, suggesting longer-term loans in this dataset are for higher principals that outweigh the spreading effect

Bivariate Analysis – Categorical vs Numerical

Loan Amount by Loan Status (Boxplot)

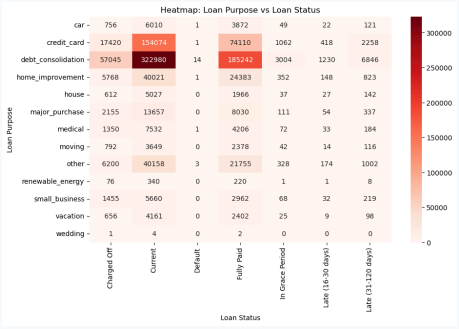


Key Insights

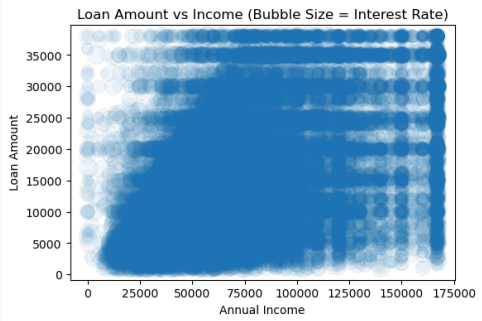
- **Variability**
  Loan amounts vary significantly across outcomes; **adverse statuses** show wider dispersion
- **Risk Indicator**
  Higher interest rates concentrate in **unfavorable outcomes**
- Interest rate overlaps indicate **non-price drivers** of repayment
- **Implication:** Segment-level pricing & limits are required

Bivariate Analysis – Categorical vs Categorical

Loan Purpose vs Loan Status (Heatmap)



Home Ownership vs Loan Status (Grouped Bars)



Segment-Level Risk Patterns

- **Purpose Variation**

  Loan outcomes vary by purpose; **debt consolidation** dominates volume with mixed performance

- **Housing Status**

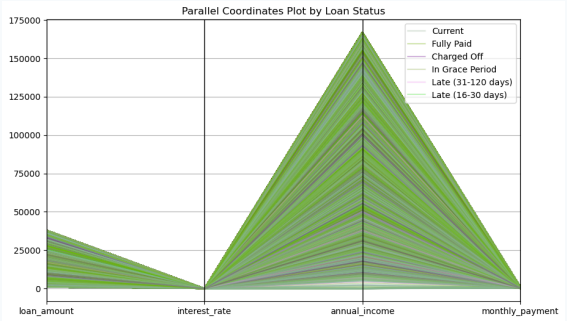
  Renters exhibit higher adverse outcomes than mortgage holders

- Heatmaps show **risk concentration** in specific purpose segments

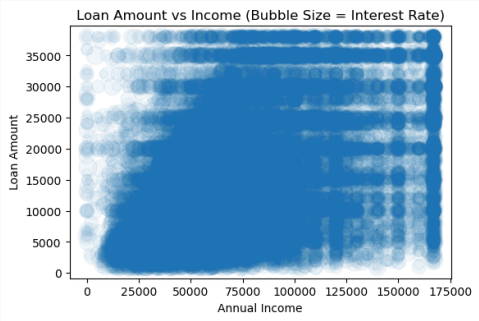
- **Implication:** Purpose & housing should drive underwriting and pricing

Multivariate Analysis – Overview

Parallel Coordinates Plot by Loan Status



Loan Amount vs Income (Bubble Size = Interest Rate)



Key Observations

- **Parallel coordinates reveal overlap:** Multivariate patterns show dense overlap across loan_status, indicating weak linear separation
- **Income–amount clustering:** Concentration at low–mid incomes; higher interest bubbles appear more frequently in lower segments
- **Implication:** Risk signals require categorical context (purpose, housing) beyond numeric features alone

Multivariate Analysis – Key Visual Insights

- Higher interest rates cluster with adverse outcomes, independent of amount bands
- At comparable incomes, larger amounts align with higher payment stress
- Purpose interacts with pricing: consolidation and credit card dominate adverse counts by volume
- Housing status moderates exposure: mortgage holders take larger loans with mixed outcomes

Pivot Table Analysis

Home Ownership × Loan Status Crosstab

Home Ownership	Current	Fully Paid	Charged Off
MORTGAGE	292,193	170,225	38,819
RENT	238,025	121,383	43,825
OWN	72,547	39,860	11,626

MINIMAL CODE

```
pd.crosstab(df['home_ownership'], df['loan_status'])
```

Segment Analysis Insights

- **Volume Distribution**

  MORTGAGE borrowers dominate portfolio (501,237 total), followed by RENT (403,233) and OWN (124,033)

- **Risk Metrics**

  Charge-off rates highest for **RENT** (10.9%), followed by OWN (9.4%) and MORTGAGE (7.7%)

- **Purpose Patterns**

  Crosstabs by **loan_purpose** show debt consolidation & credit cards dominate volume with mixed outcomes

- **Implication:** Differentiate underwriting by housing status; consider tighter limits or pricing for renters

Key Insights

Portfolio Distribution

Portfolio is skewed to **Current/Fully Paid** statuses; monitor minority adverse classes via **rates and proportions**

Class Imbalance

Pricing Alignment

Pricing aligns with risk, but **overlaps across statuses** require segment-level controls and multivariate underwriting

Risk Pricing

Segment Stratification

Purpose and housing strongly stratify repayment; consolidation/credit card and renter segments show higher risk

Key Drivers

Income–Loan Relationships

Income alone is insufficient; **interactions with purpose, limits, and housing** explain outcomes better

Multivariate Effects

Data Quality

Outlier capping improved distributions without shifting medians; **IQR preserved central tendency**

Preprocessing Success

Conclusion

- **EDA Deliverables:** This analysis delivered a **clean, analysis-ready dataset**, comprehensive univariate and bivariate distributions, multivariate relationship mapping, and segment-level pivots quantifying repayment patterns across borrower and loan attributes

- **Risk Signals:** Key indicators of elevated repayment risk include **higher interest rates, debt consolidation and credit card purposes, renter housing status**, and interactions between loan amount and income levels independent of univariate thresholds

- **Operational Implications:** Findings support **differentiated underwriting policies** by segment (purpose, housing), refinement of pricing tiers to reflect multivariate risk, tighter exposure limits for high-risk purposes, and targeted collection strategies for renter segments

- **Limitations:** Analysis is **observational and cross-sectional**; lacks temporal cohort tracking, causal inference frameworks, and predictive validation; class imbalance in target variable requires rate-based metrics and stratified sampling for future modeling