Business Value through Integrative Analytics
Update 1

Loan Data Analysis and Insights

Kojo Dokyi
Srushti Kadam
Nivedha Gautham Raj

Business Objective

Jasmin, a data-savvy investor, aims to leverage Lending Club data to invest in peer-to-peer loans.

- **Objective:** Maximize returns and manage risk using data-driven strategies
- **Business Questions:**
  - Which loans are profitable and low risk?
  - How can data improve investment decisions?
- **Key Metrics (KPIs):**
  - Net Return on Investment (ROI)
  - Loan Default Rate
  - Portfolio Diversification (by grade)

Data Overview & Preprocessing

- **Source:** Lending Club data (2014 & 2015 via Wayback Machine)
- **Raw Dataset:** 145 features per loan (loan amount, interest rate, employment info, etc.)
- **Preprocessing Steps:**
  - Selected 25 key features based on relevance and availability at investment time
  - Removed leaked/derived variables (total payment, recoveries)
  - Cleaned missing values, standardized formats
  - Saved as a clean dataset for modeling

Key Insights

Loan Status Breakdown: Significant portion of loans defaulted

Grade Correlation: Loans with grades E-G had higher default likelihood

Trend Observed:

- Higher income generally leads to larger loans
- Longer employment tenure correlated with lower default risk

Hypotheses Formed:

- High debt-to-income ratio increases default probability
- Higher interest rates are associated with higher risk and return

Derived Loan Features

Introduced 3 return metrics per loan:

1. **Optimistic Return:** Assumes full repayment
2. **Pessimistic Return:** Assumes no recovery after default
3. **Intermediate Return:** Partial recovery + interest before default

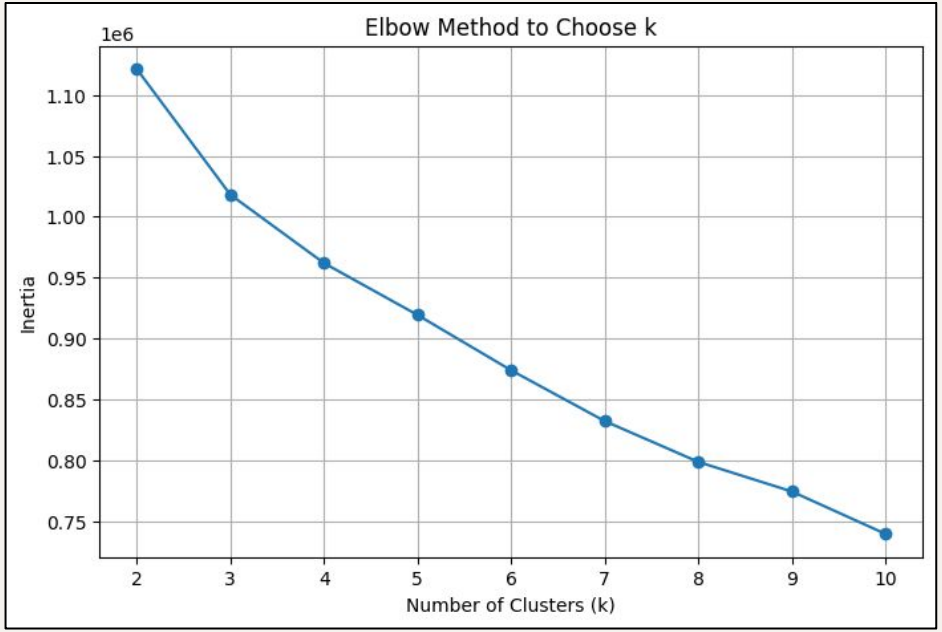
- **Why Intermediate:** Balances realism and usability for modeling
- **Outcome:** New return features added to dataset for analysis and model training

Cluster Analysis

3

Optimal Number of Clusters

As obtained from elbow point



11

Numeric Features / Labels

Loan amount, installments, interest rate, annual income, debt-to-income ratio, revolving utilization ratio, delinquencies in the last 2 years, number of open accounts, public records, revolving balance, months since last delinquency

Cluster Analysis

Cluster 1

Grades B and C

A middle-risk segment

Cluster 0

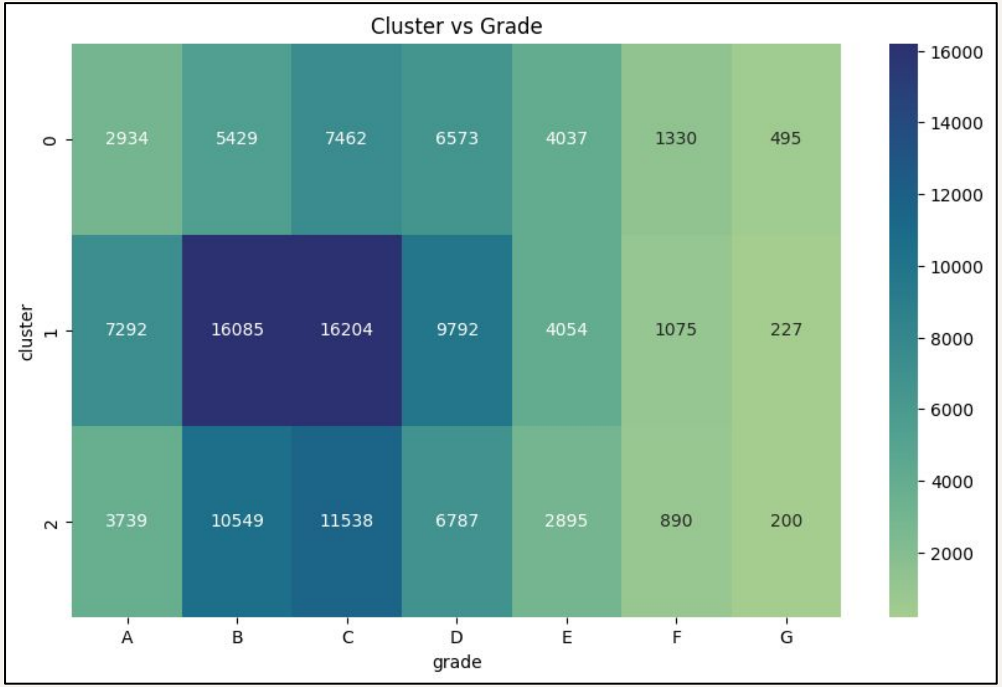
Fewer Loans

Riskier borrowers with lower creditworthiness.

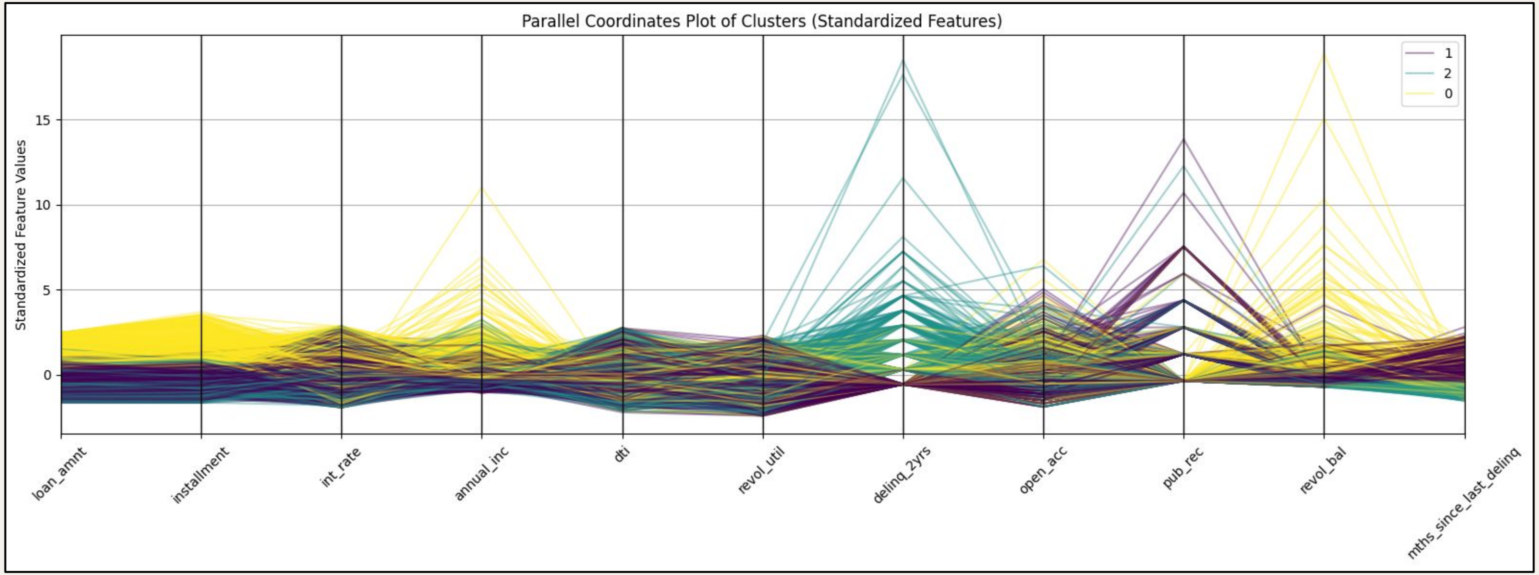
Cluster 2

Higher in B and C

Mix of lower-risk loans along with middle-risk loans



Cluster Analysis



Cluster 0: High-Income, High Revolving Balance

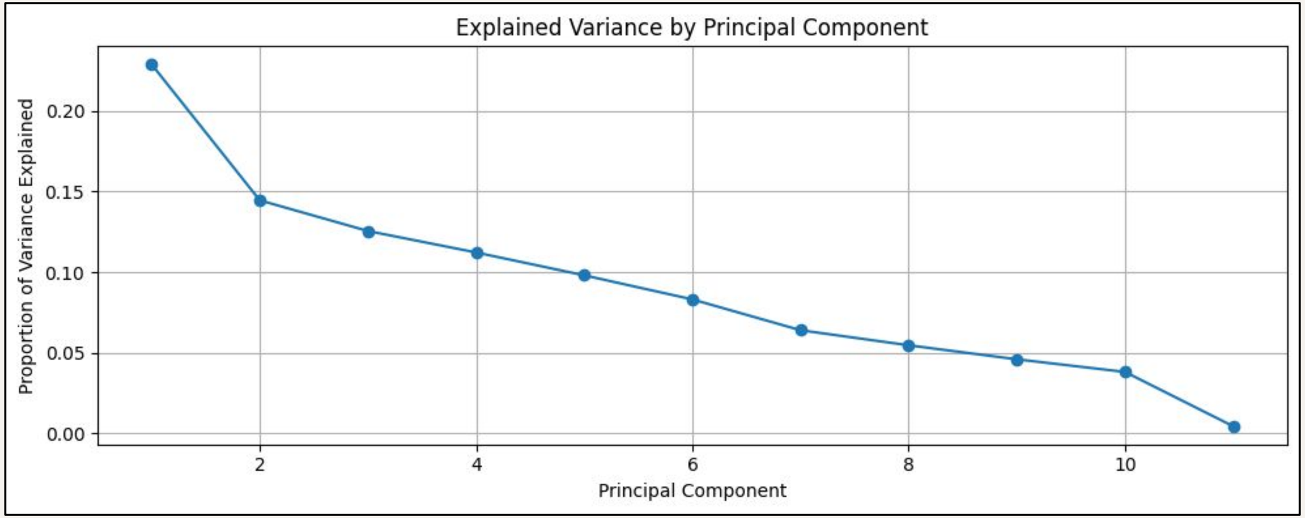
Cluster 1: Higher Public Records & Moderate Loan Amounts

Cluster 2: High Delinquencies, Higher Loan Amounts

PCA Insights

To better understand structure of the dataset and reduce dimensionality, PCA was applied to the same features as clustering analysis

PCA Insights - Explained Variance



PC1

Explains ~22% of the variance

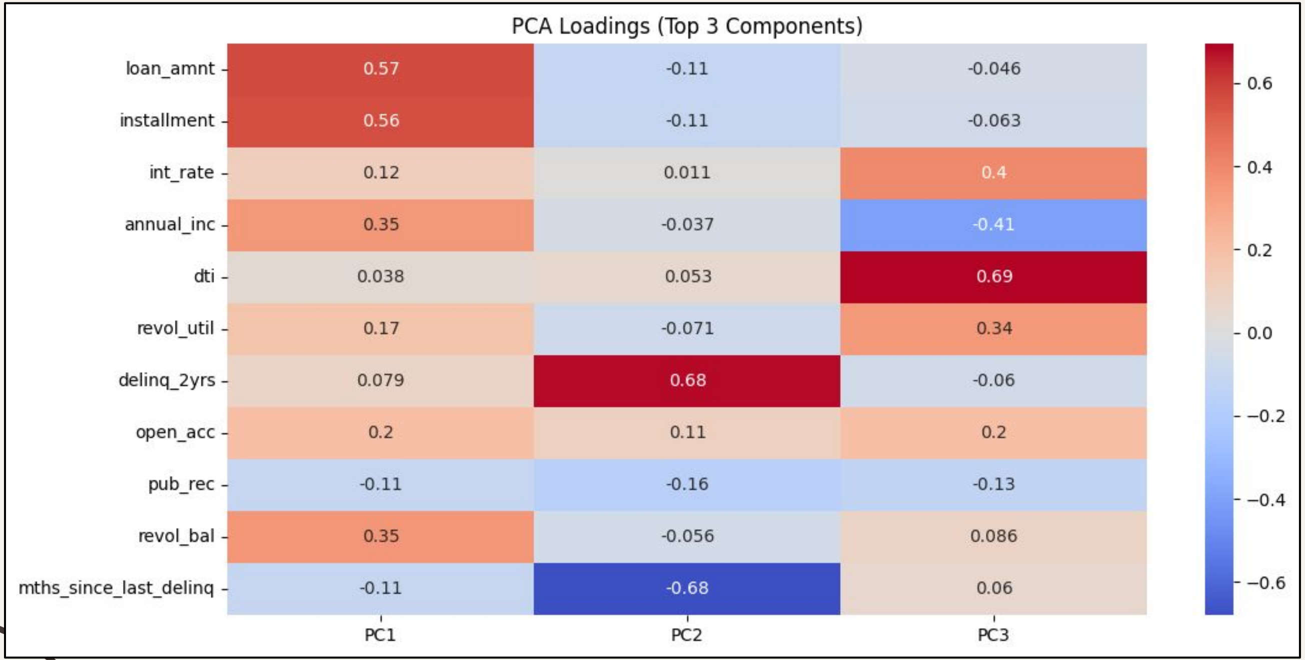
PC2

Additional ~14.5%

PC3

Rounds total to ~48% for the top 3

PCA Insights - Loadings Interpretation



PC1

Strongly influenced by Loan Size & Borrower Capacity

PC2

Strongly influenced by Delinquency History / Riskiness

PC3

Strongly influenced by Delinquency History / Riskiness

Key Takeaways and Business Implications

Loan Profitability and Risk Identification

Higher loan grades (E-G) showed a strong correlation with default likelihood

Feature Engineering Enhanced Decision-Making

PCA helped reduce dimensionality while preserving key patterns.

Investor Strategy Recommendations

- Focus on clusters with moderate risk and high return potential.
- Avoid loans with high delinquencies and low creditworthiness.
- Diversify investments across Grade B and C loans for balanced risk.

Thanks