Bank Loan Data Analysis — Case Study

BIG DATA ANALYTICS PROJECT

Prepared by

Shreyash Pandey

1240259048

Table of Contents

1. Project Overview

This project analyzes bank loan application records for Indian branches to identify approval patterns, default risk, demographic influences, and branch-level performance. The analysis is implemented using Cloudera/HiveQL style SQL queries (demonstrative) and supplemented with visualizations to aid stakeholder decisions.

2. Dataset Description

Dataset file: Bank_Loan_Data_India.csv (synthetic, for academic use). Rows: 500 (approx).

Columns and descriptions:

Column Name	Description
customer_id	Unique customer identifier (STRING)
gender	Customer gender (Male/Female)
age	Age in years (INT)
marital_status	Marital status (STRING)
city	Branch / customer city (STRING)
monthly_income	Monthly income in INR (INT)
employment_type	Employment type (Salaried/Self-Employed/etc.)
loan_type	Type of loan applied (Home/Personal/Education/Business/Auto)
loan_amount	Applied loan amount in INR (INT)
loan_term_months	Loan tenure in months (INT)
interest_rate_percent	Interest rate offered (DOUBLE)
credit_score	Credit bureau score (INT, nullable)
approval_status	Approved/Rejected (STRING)
default_status	Yes/No (STRING) - indicates if borrower later defaulted
application_date	Date of application (DATE)

branch_city	Bank branch city (STRING)
existing_loans	Number of existing loans (INT)

3. Project Objectives

- Analyze approval rates across loan types and branches.
- Investigate relationship between credit score and approval/default.
- Segment customers by income and risk profile.
- Estimate default probabilities by cohort and credit band.
- Provide SQL queries and visual dashboards.

4. Technologies Used

Tool	Purpose
Cloudera / HiveQL	SQL aggregation and querying
HDFS / Hadoop	Storage (optional)
Python / Pandas	Prototyping
Power BI / Tableau	Dashboarding
Matplotlib	Charts for report

5. Steps Performed (Data Loading & Cleaning)

Sample SQL to create table and load CSV into Hive shown below:

CREATE DATABASE bank_loans_db; USE bank_loans_db;
CREATE TABLE bank_loans (...);

6. Exploratory Analysis & SQL Queries

total_apps	approval_pct
500	60.4

Insight: Overall approval rate and volume.

Approval rate by loan type:

loan_type	approval_pct
Auto	47.5
Business	69.23
Education	69.86
Home	16.8
Personal	85.33

Insight: Product-wise approval differences.

Average loan amount and approval by city:

city	avg_loan	approval_pct
Ahmedabad	828222.08	58.49
Bengaluru	697790.17	66.67
Chennai	835997.37	56.52
Delhi	726563.92	61.22
Hyderabad	608599.4	68.89

Jaipur	803809.37	57.14
Kolkata	679129.9	62.75
Lucknow	714984.14	51.02
Mumbai	722272.66	62.26
Pune	555480.58	59.65

Insight: Branch ticket sizes and approval behavior.

Default rate by loan type (approved loans only):

loan_type	default_pct
Auto	42.11
Business	50.0
Education	58.82
Home	76.19
Personal	32.48

Insight: Product default risk.

7. Risk & Default Analysis (Visuals)

Figure: Approval Rate by Loan Type

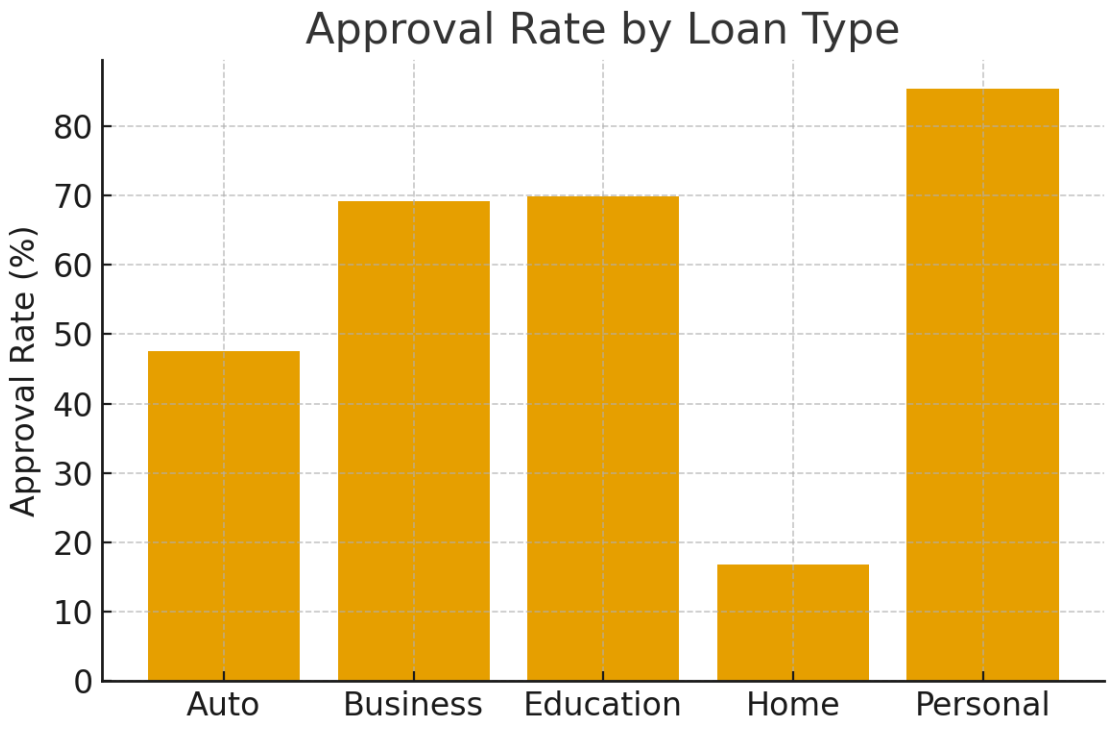


Figure: Average Loan Amount by City (Lakh INR)

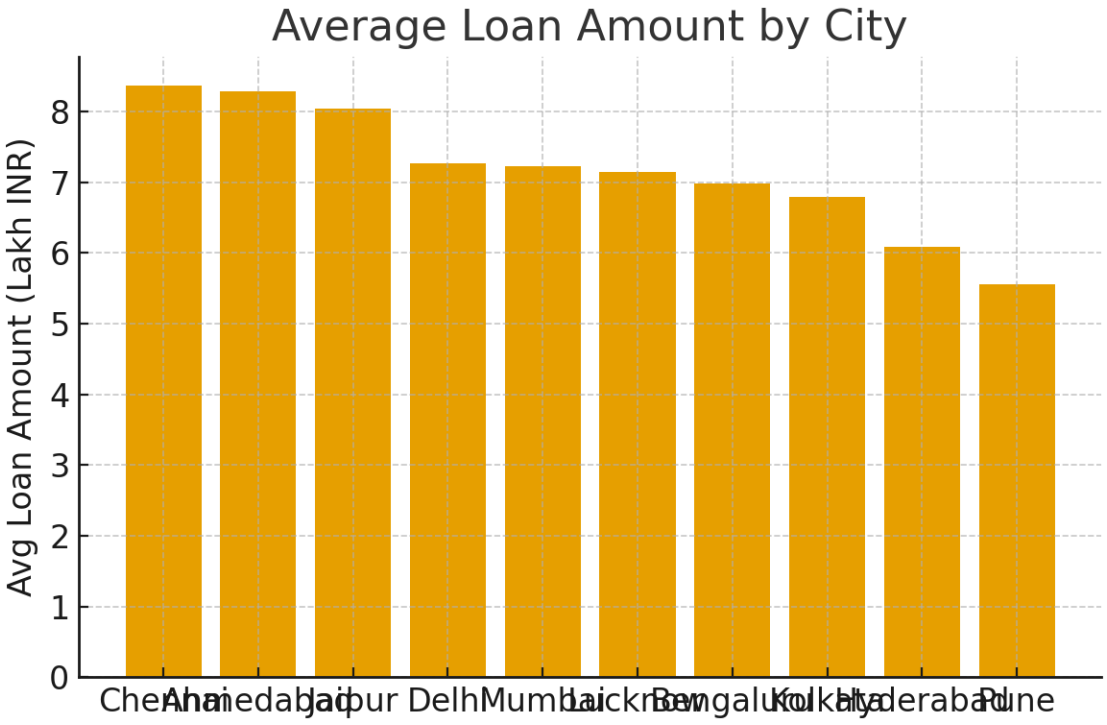


Figure: Credit Score Distribution

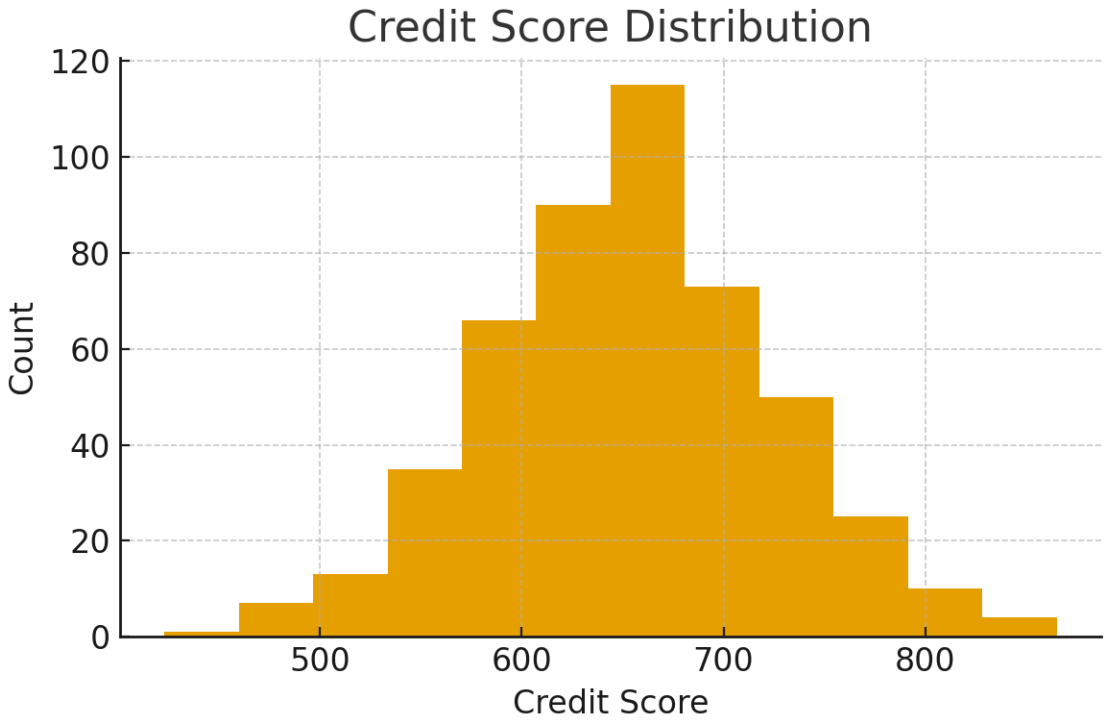
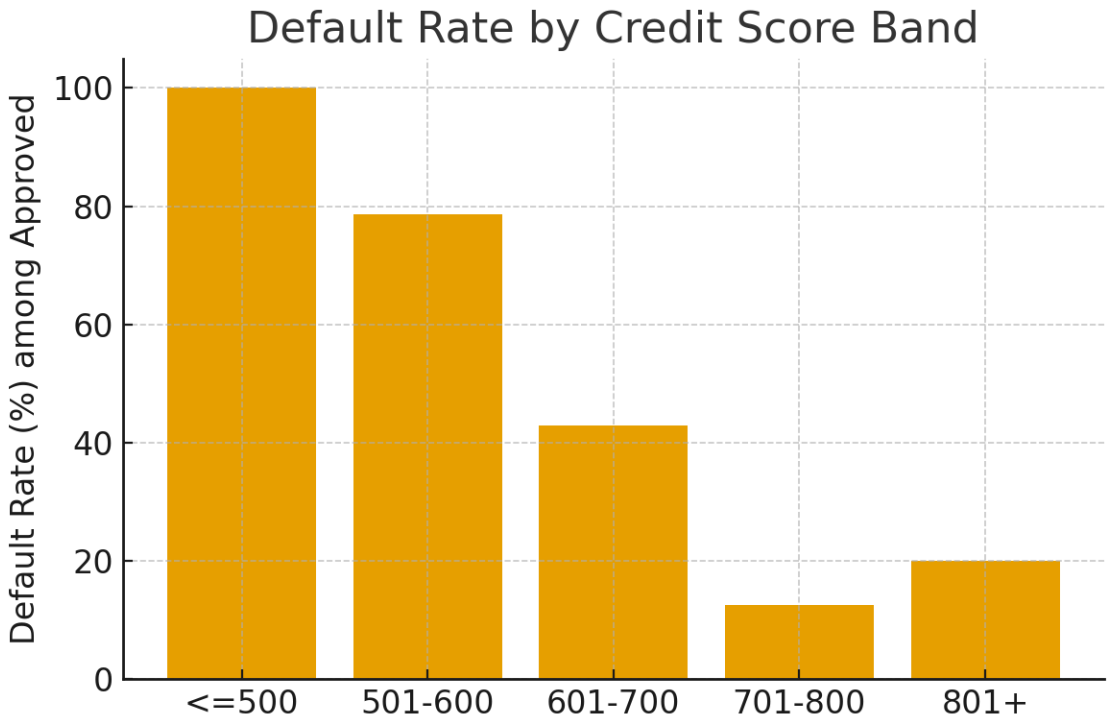


Figure: Default Rate by Credit Score Band


Default Rate by Credit Score Band
Default Rate (%) among Approved
0
20
40
60
80
100
<=500
501-600
601-700
701-800
801+

8. City & Demographic Insights

Sample SQL and insights on age, employment and product preference.

9. Recommendations & Conclusion

- Tighten underwriting for low credit-score bands.
- Promote secured products with risk-based pricing.
- Launch credit-building products for young customers.
- Implement branch KPIs and dashboards.

10. Appendix — SQL Snippets & Dataset Notes

CSV 'Bank_Loan_Data_India.csv' is synthetic for academic use. Load into Hive using LOAD DATA or external table.

```
-- Appendix Query 1
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 2
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 3
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 4
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 5
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 6
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 7
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 8
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 9
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 10
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 11
SELECT COUNT(*) FROM bank_loans WHERE ...;
```

```
-- Appendix Query 12
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 13
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 14
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 15
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 16
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 17
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 18
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 19
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 20
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 21
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 22
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 23
SELECT COUNT(*) FROM bank_loans WHERE ...;

-- Appendix Query 24
SELECT COUNT(*) FROM bank_loans WHERE ...;
```