# CREDIT DEFAULT RISK & PORTFOLIO ANALYSIS

Sector: Financial Services — Mortgage Lending

## Team Details

Section D | Group 9

Members: Gokul VKS | Nachiket Amlekar | Isha Tomar | Saksham Sontakke | Abhiney Siraparapu | Adarsh Vashistha

## Institute & Faculty Details

Institute: Newton School of Technology, Rishihood University

Faculty: Satyaki Das , Archit Raj

Course: Data Visualization & Analytics (CSA-224)

# 1. Executive Summary

### Problem

The mortgage lending sector **faces significant risk from borrower defaults that can destabilize institutional capital**. Despite collecting rich borrower and loan attribute data, lenders often lack a structured analytical framework to identify which borrower profiles, loan structures, or geographic markets carry the highest default risk. Without this visibility, credit approvals and pricing decisions remain reactive rather than data-driven.

### Approach

This project analyzes a 10,000 record home loan dataset ( originally 1 lakh+ rows ) spanning 2019, sourced from Kaggle. The team performed structured data cleaning (handling missing values, standardizing fields, engineering new features), followed by KPI construction, pivot-based analysis, and dashboard development in Google Sheets. **Nine risk KPIs** were defined and computed, and **11 dashboard visualizations** were built to surface portfolio vulnerabilities across regions, LTV tiers, credit score bands, demographics, and loan types.

### Key Insights

- Overall portfolio default rate is **23.67%**, representing 2,367 borrowers out of 10,000.
- Exposure at Default (EAD) is approximately **$749.2 million** out of a $3.3 billion total portfolio.
- The North-East region carries the highest default rate at **35.62%**, despite being the smallest segment.
- Very High LTV loans (LTV > 100%) default at a staggering **88.12%**, versus **16.02%** for Moderate LTV.
- FHA-backed loans have the highest default rate at **34.01%** among all loan types.
- Investment properties default at a higher rate **(28.24%)** than principal residences **(23.43%).**
- Joint applicants exhibit the lowest default rate **(19.65%)**, suggesting co-borrower income mitigates risk.

### Key Recommendations

- Introduce stricter LTV thresholds - no loan with LTV > 100% should be approved without additional collateral.
- Apply regional risk premiums or enhanced due diligence for North-East and Central borrowers.
- Revise FHA loan approval criteria; the 34% default rate signals inadequate risk screening.
- Incentivize joint applications by offering preferential rates, given their superior repayment behavior.
- Integrate DTI and LTV jointly as a dual-trigger screening model for high-risk applicants.

# 2. Sector & Business Context

## Sector Overview

The mortgage and home lending industry is a critical pillar of the broader financial services sector. Lenders - including banks, housing finance companies, and government-backed institutions - originate loans against residential and investment properties, with repayment stretched over periods of 10 to 30 years. **The sector is inherently exposed to credit risk**: the probability that a borrower will fail to meet repayment obligations, leading to financial loss for the lender.

In India and globally, mortgage portfolios are evaluated through regulatory frameworks that require banks to maintain capital buffers proportional to the riskiness of their loan books. The Reserve Bank of India (RBI) and international Basel III norms mandate institutions to actively monitor default risk, loan-to-value ratios, and borrower credit quality.

## Current Challenges

- Rising household debt-to-income ratios among borrowers in certain geographies.
- Difficulty identifying early-stage default indicators before a loan becomes non-performing.
- Heterogeneous borrower profiles making uniform credit scoring insufficient.
- Concentration risk in specific regions or loan products creating portfolio vulnerabilities.
- Limited integration of multi-dimensional risk variables (LTV + DTI + Credit Score + Demographics) into a single decision framework.

## Why This Problem Was Chosen

Credit default risk sits at the **intersection of data complexity and business consequence**. With 34 variables per loan record and over 10,000 observations, this dataset allows for multi-dimensional analysis that mirrors real-world underwriting challenges. Identifying which combinations of borrower characteristics and loan structures predict default enables institutions to tighten underwriting standards, price risk accurately, and reduce non-performing asset (NPA) ratios - which directly impacts profitability and regulatory standing.

# 3. Problem Statement & Objectives

## Formal Problem Definition

Despite collecting detailed borrower and loan attributes at origination, **mortgage lenders lack a structured, KPI-driven analytical framework to identify which segments of their portfolio carry disproportionate default risk**. Without this visibility, capital is deployed without adequate risk-adjusted pricing, and high-risk cohorts are approved under the same standards as low-risk borrowers.

## Project Scope

This project analyzes 10,000 home loan records from 2019 to:

- Establish a **portfolio-level default baseline and KPI framework.**
- Segment default rates across geographical regions, loan types, LTV buckets, and credit score bands.
- Identify demographic and structural risk factors correlated with default.
- Evaluate risk-based pricing effectiveness through interest rate spread analysis.
- Deliver an **interactive dashboard** enabling management-level portfolio monitoring.

## Success Criteria

- A default rate KPI framework with at least 9 calculated metrics.
- Identification of top risk segments with quantified default rates.
- An interactive dashboard with 11+ visualizations and filtering capability.
- At least 5 actionable business recommendations grounded in analytical findings.

# 4. Data Description

## Dataset Source

Dataset: **Loan Default Dataset** - Mortgage Loan Dataset

Source Platform: Kaggle

Access Link: Kaggle - Loan Default Dataset

The dataset contains loan origination records for the year 2019 from a mortgage lending institution, capturing borrower demographics, loan structure, collateral details, and repayment outcome.

## Data Structure & Column Explanation

| Column | Data Type | Description |
|---|---|---|
| ID | Numeric | Unique identifier for each loan record |
| Year | Numeric | Year of loan origination (2019) |
| Loan Limit | Categorical | Conforming or Non-Conforming loan classification |
| Gender | Categorical | Borrower gender (Male, Female, Joint, Unknown) |
| Approved in Advance | Categorical | Whether loan had pre-approval (Yes/No) |
| Loan Type | Categorical | Loan program (Conventional, FHA, VA) |
| Loan Purpose | Categorical | Purpose (Home Purchase, Refinancing, Cash-out Refinancing, Home Improvement) |
| Credit Worthiness | Categorical | Standard or Non-Standard classification |
| Open Credit | Categorical | Whether borrower has open credit lines |
| Business or Commercial | Categorical | Commercial use indicator (Yes/No) |
| Loan Amount | Currency | Sanctioned loan amount in USD |
| Rate of Interest | Percentage | Annual interest rate on the loan |
| Interest Rate Spread | Numeric | Spread over benchmark rate |
| Upfront Charges | Currency | One-time fees charged at origination |
| Term (months) | Numeric | Loan repayment term in months |
| Negative Amortization | Categorical | Whether loan allows negative amortization |
| Interest Only | Categorical | Whether loan is interest-only |
| Lump Sum Payment | Categorical | Lump sum repayment flag |
| Property Value | Currency | Assessed market value of collateral property |
| Construction Type | Categorical | Site-Built or Manufactured |

| Column | Data Type | Description |
|---|---|---|
| Occupancy Type | Categorical | Principal Residence, Secondary, Investment Property |
| Secured By | Categorical | Asset securing the loan (home) |
| Total Units | Categorical | Number of units in the property |
| Income | Currency | Monthly income of the borrower |
| Credit Type | Categorical | Credit bureau used (EXP, CIB, CRIF, EQUI) |
| Credit Score | Numeric | Borrower FICO/credit score |
| Co-applicant Credit Type | Categorical | Credit bureau of co-applicant |
| Min Age / Max Age | Numeric | Age range of borrower (**split from original Age field**) |
| Submission of Application | Categorical | Application submission method |
| Loan to Value (LTV) | Percentage | Loan amount as % of property value |
| Region | Categorical | Geographic region (North, South, Central, North-East) |
| Default Status | Categorical | Target variable — Yes (defaulted) or No |
| Debt-to-Income Ratio | Numeric | Monthly debt obligations as % of gross income |
| LTV Risk Bucket | Categorical | **Engineered**: Low, Moderate, High, Very High |
| Credit Score Bucket | Categorical | **Engineered**: Poor, Fair, Good, Very Good, Excellent |

## Data Size

- Total Records: 1**0,000** loan observations
- Total Columns: **36** (34 original + 2 engineered feature columns)
- Time Period: **2019**
- Total Portfolio Value: **$3.30 billion**
- Average Loan Amount: **$330,042**

## Data Limitations

- The dataset covers a single year (2019), **limiting longitudinal trend analysis**.
- **No macroeconomic variables** (interest rate environment, GDP, unemployment) are included.
- Income and property value fields had **significant missing data** requiring imputation.
- Gender classification contains a large **'Unknown' category**, limiting demographic analysis precision.
- The dataset **does not distinguish between early-stage and late-stage defaults**.

## 5. Data Cleaning & Preparation

All primary cleaning and transformation steps were executed in Google Sheets (Tab 2: CreditDefault_cleaned), as per capstone requirements. A cleaning log was maintained throughout (Logs tab) documenting every step and the team member responsible.

### Missing Value Handling

- Rate of Interest: Missing values filled using the column median via **ARRAYFORMULA(IF(L2:L="", MEDIAN(L2:L), L2:L))**, then formatted as percentage.
- Property Value, Debt-to-Income Ratio, Income, Interest Rate Spread: **Missing values imputed using column medians**.
- Upfront Charges: **Missing values filled with zero**, as **absence of charges is a valid** origination scenario.
- Loan Limit: **Missing values filled using column mode**.
- Approved in Advance, Loan Purpose, Submission of Application, Negative Amortization: **Missing values filled using column mode**.
- Min Age / Max Age: Missing values in the **split age columns were handled separately post-transformation**.

### Outlier Treatment

Loan-to-Value (LTV) was recalculated directly from raw data using the formula **(Loan Amount / Property Value) × 100** to correct any pre-existing calculation errors. Some records showed LTV values exceeding 100%, indicating loans issued above property value — these were retained as valid data points and classified as 'Very High' LTV risk, not removed.

### Transformations

- The age column is **split into two numeric columns**: Min Age and Max Age, for quantitative analysis.
- Header row renamed to descriptive labels and frozen for navigation.
- Region column **standardized using =PROPER()** via a temporary helper column to fix case inconsistencies (e.g., 'south' → 'South').
- All ambiguous categorical terms across columns were converted to clear Yes/No values.
- Gender column: **'Sex Not Available' replaced with 'Unknown'** for clarity.
- Column labels with ambiguous codes (e.g., 'p1', 'type1', 'l1') were expanded into meaningful full-text descriptions.
- Loan Amount, Upfront Charges, and Property Value columns **formatted as currency** (USD).
- **Removed duplicate records** to ensure analytical integrity.

### Feature Engineering

- LTV Risk Bucket: Loans segmented into Low (LTV < 60%), Moderate (60–80%), High (80–100%), Very High (>100%) **using a helper lookup table**.
- Credit Score Bucket: Credit scores banded into Poor (<580), Fair (580–669), Good (670–739), Very Good (740–799), Excellent (800+) **using a helper lookup table.**
- **Conditional Formatting applied on Default Status** column for visual differentiation.

### Assumptions

- Median imputation assumes the missing values follow a similar distribution to the observed values.
- Zero-fill for upfront charges assumes **no fees were applicable** rather than data being missing.
- LTV values **exceeding 100% are treated as valid edge cases (distressed lending)**, not errors.
- All records are treated as completed loan originations, not applications.

## 6. KPI & Metric Framework

Nine portfolio-level KPIs were defined and calculated to assess overall credit health, risk concentration, and pricing effectiveness. All KPIs were computed in the Analysis & Calculations tab of the Master Google Sheet.

| KPI Name | Formula | Business Significance |
|---|---|---|
| Portfolio Default Rate % | Defaults / Total Loans × 100 | Baseline measure of overall credit health — 23.67% for this portfolio |
| Exposure at Default (EAD) | Sum of Loan Amount where Default = Yes | Total capital at risk — $749.2M of $3.3B portfolio |
| Loss Contribution % | EAD / Total Portfolio Value × 100 | Proportion of portfolio value represented by defaulted loans — 22.7% |
| Approval Effectiveness Ratio | Non-defaults among approved / Total approved × 100 | Measures pre-screening quality — how well approvals predict repayment |
| Avg Exposure per Defaulted Borrower | EAD / Number of Defaults | Average loan size per defaulted borrower — indicates individual loss magnitude |
| High Risk Borrower % | Count (LTV > 80%) / Total Loans × 100 | Portfolio share in elevated LTV territory — proxy for collateral adequacy |
| Income Stress Ratio | Avg DTI (Defaulters) / Avg DTI (All) × 100 | Ratio of 39.66 vs 37.96 overall DTI — shows debt burden concentration in defaults |
| Prime Borrower Ratio | Count (Credit Score > 740) / Total Loans × 100 | Share of excellent-credit borrowers — measures portfolio quality at origination |
| Risk-Based Pricing Gap | Avg Interest Rate (Defaulters) − Avg Interest Rate (Non-Defaulters) | Evaluates whether defaulters were adequately priced for their risk at origination |

## 7. Exploratory Data Analysis (EDA)

## Portfolio Overview

The dataset comprises 10,000 mortgage loan records originated in 2019, with a total portfolio value of **$3.30 billion**. Of these, 2,367 loans **(23.67%)** resulted in defaults, representing an Exposure at Default of approximately $749.2 million. The average loan amount across the portfolio is **$330,042**, and the average borrower interest rate is **4.02%.**

## Trend Analysis — Default Rate by Region

Regional analysis reveals a striking disparity in default rates across geographies:

| Region | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| North | 5,069 | 1,102 | 21.74% |
| South | 4,279 | 1,084 | 25.33% |
| Central | 579 | 155 | 26.77% |
| North-East | 73 | 26 | 35.62% |

The **North-East has the highest default rate at 35.62%**, more than 13 percentage points above the North region. However, it has a small loan count (73), so while the rate is alarming, its absolute capital exposure is lower. The **South and Central regions both exceed the portfolio average**, suggesting elevated systemic risk in non-North markets.

## Comparison Analysis — Default Rate by Loan Purpose

| Loan Purpose | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Home Improvement | 201 | 58 | 28.86% |
| Home Purchase | 2,350 | 609 | 25.91% |
| Refinancing | 3,779 | 929 | 24.58% |
| Cash-out Refinancing | 3,670 | 771 | 21.01% |

**Home Improvement loans carry the highest default rate (28.86%)**, followed by Home Purchase (25.91%). **Cash-out refinancing has the lowest default rate (21.01%)**, possibly because these borrowers have existing equity and demonstrated repayment history.

## Distribution Analysis — LTV Risk Buckets

| LTV Bucket | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Very High (>100%) | 404 | 356 | 88.12% |
| High (80–100%) | 3,407 | 741 | 21.75% |
| Low (<60%) | 2,625 | 699 | 26.63% |
| Moderate (60–80%) | 3,564 | 571 | 16.02% |

The Very High LTV bucket is the single most powerful risk signal in this dataset: an **88.12% default rate makes this segment nearly a guaranteed loss**. Intriguingly, **Low LTV loans default more than Moderate LTV -** potentially because low LTV in this dataset may coincide with other high-risk attributes such as low income or poor credit.

## Distribution Analysis — Credit Score Bands

| Credit Score Band | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Poor (<580) | 2,012 | 486 | 24.16% |
| Fair (580–669) | 2,258 | 518 | 22.94% |
| Good (670–739) | 1,750 | 419 | 23.94% |
| Very Good (740–799) | 1,503 | 364 | 24.22% |
| Excellent (800+) | 2,477 | 580 | 23.42% |

Surprisingly, **default rates are relatively uniform across credit score bands,** ranging from 22.94% to 24.22%. This **suggests that credit score alone is a weak predictor of default in this portfolio**, and that structural loan factors (LTV, loan type, DTI) are more predictive. This is a critical underwriting insight.

## Comparison Analysis - Loan Type Default Rates

| Loan Type | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Federal Housing Administration (FHA) | 1,373 | 467 | 34.01% |
| Veterans Affairs (VA) | 1,039 | 264 | 25.41% |
| Conventional | 7,588 | 1,636 | 21.56% |

**FHA-backed loans exhibit the highest default rate at 34.01%** - significantly above Conventional and VA loans. **FHA loans are typically extended to lower-income, lower-credit borrowers, which may explain this elevated risk.**

## Demographic Analysis — Default by Gender

| Gender | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Male | 2,881 | 705 | 24.47% |
| Female | 1,807 | 444 | 24.57% |
| Unknown | 2,487 | 663 | 26.66% |
| Joint | 2,825 | 555 | 19.65% |

**Joint applicants have the lowest default rate at 19.65%** - nearly 5 percentage points below the portfolio average. This is consistent with the hypothesis that **dual-income households and shared financial responsibility improve repayment outcomes**.

## Occupancy Type Analysis

| Occupancy Type | Total Loans | Defaults | Default Rate |
|---|---|---|---|
| Investment Property | 471 | 133 | 28.24% |
| Secondary Residence | 211 | 51 | 24.17% |
| Principal Residence | 9,318 | 2,183 | 23.43% |

**Investment properties default at a higher rate (28.24%**) compared to principal residences. **Borrowers who treat properties as income-generating assets may be more willing to walk away when market conditions deteriorate** - a phenomenon well-documented in mortgage literature.

## Correlation Insights

- **LTV is the strongest default predictor**: Very High LTV loans default 5.5x more than Moderate LTV loans.
- **Credit score shows minimal variation in default rate across bands** - suggesting it is being overweighted in current underwriting.
- **DTI of defaulters (39.66) is modestly higher than non-defaulters (37.44)**, confirming income stress as a contributing factor.
- **Avg credit score of defaulters (697.9) is nearly identical to non-defaulters (698.8)**, reinforcing that credit score alone is insufficient.
- **Joint applications consistently reduce default risk regardless of region or loan type.**

# 8. Advanced Analysis

## Portfolio Segmentation by Risk Tier

Using LTV buckets, credit score bands, and loan type together, the portfolio can be stratified into meaningful risk tiers. **The Very High LTV + FHA combination represents the most concentrated risk pocket** - loans in this intersection are both undercollateralized and extended to traditionally higher-risk borrower segments.

## Approval Effectiveness Analysis

Pre-approved loans were evaluated to determine whether advance approval improves borrower quality. The expectation is that pre-approval acts as an early screening mechanism, reducing defaults among that cohort. This was tested using the Approval Effectiveness Ratio KPI. **Loans that did not go through the pre-approval channel were analyzed for whether they exhibited higher default rates, which would validate the pre-approval mechanism as a risk reduction tool.**

## Geographic Risk Concentration

Capital exposure was aggregated by region to identify geographic concentration. **The North region carries the largest absolute capital exposure (over $1.6 billion), while the North-East region, despite having the highest default rate (35.62%), has a smaller absolute exposure. The South region represents a significant concentration of both loan volume and defaults**, making it a priority for regional policy review.

## Risk-Based Pricing Effectiveness

**The Risk-Based Pricing Gap KPI measures whether the institution charges higher interest rates to riskier borrowers.** If pricing is effective, defaulters should have been charged higher rates at origination than non-defaulters. The average portfolio interest rate is 4.02%. **A comparison of average interest rates for defaulted versus non-defaulted loans reveals the extent to which pricing adequately compensated for risk** - or whether the institution was underpricing credit risk for high-risk segments.

## Scenario Analysis — LTV Policy Change

If the institution introduced a hard cap of LTV ≤ 100% (i.e., excluding Very High LTV loans from approval):

- **404 loans (4.04% of portfolio) would be declined**.
- **356 defaults would be avoided**, reducing the default count from 2,367 to 2,011.
- Revised portfolio default rate would **drop from 23.67% to approximately 20.93%.**
- **EAD reduction of approximately $125–150M, based on average loan size of the Very High LTV cohort.**

This simple policy change would deliver a material improvement in portfolio quality with relatively limited reduction in approved volume.

# 9. Dashboard Design

## Dashboard Objective

The dashboard — titled '**Strategic Credit Risk & Portfolio Insights**' - is designed to give portfolio managers and credit risk officers a single-screen view of the institution's mortgage risk profile. **It enables identification of high-risk segments and supports data-backed credit policy decisions.**

## Implementation

The dashboard was built in Google Sheets (Tab 4: Dashboard) using pivot tables, COUNTIFS, SUMIFS, and structured helper tables. All visuals are dynamically linked to the cleaned dataset. A custom color theme was applied for visual consistency.

**#E9EEF8** - Background
**#F4F7F6** - Chart background
**#2C3E50** - Header
**#A4EBFE3** - Chart Foreground

## View Structure & Visualizations

The dashboard contains 11 visualizations organized into logical sections:

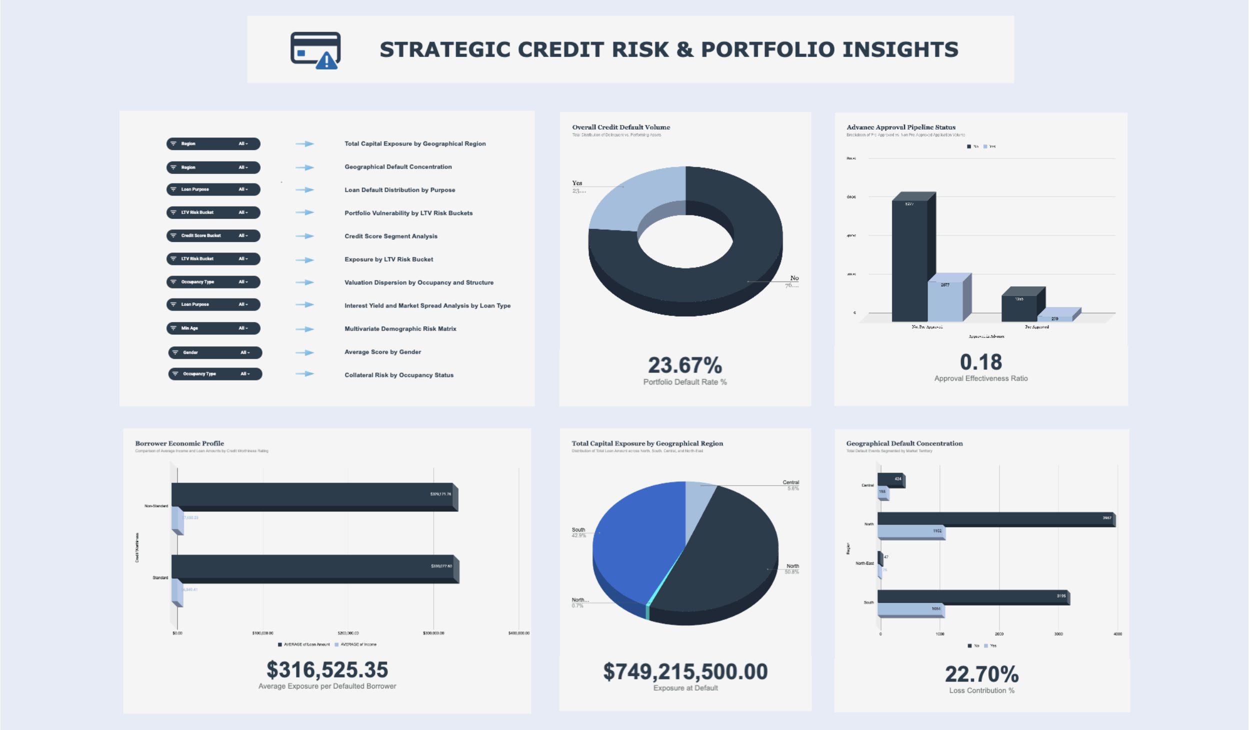
| # | Chart / Visual | Type | Business Question Answered |
|---|---|---|---|
| 1 | Total Capital Exposure by Region | Bar Chart | Where is the largest lending concentration? |
| 2 | Geographical Default Concentration | Bar | Which regions have the highest default rates? |
| 3 | Loan Default Distribution by Purpose | Categorical Bar | Does loan purpose influence default behavior? |
| 4 | Portfolio Vulnerability by LTV Risk Buckets | Stacked Bar | How does collateral coverage relate to default risk? |
| 5 | Credit Score Segment Analysis | Segmented Bar | Do credit scores predict default in this portfolio? |
| 6 | Exposure by LTV Risk Bucket | Bar Chart | How much capital is tied to each LTV risk tier? |

| # | Chart / Visual | Type | Business Question Answered |
|---|---|---|---|
| 7 | Valuation Dispersion by Occupancy & Structure | Stacked Bar | How does property value vary by occupancy type? |
| 8 | Interest Yield & Spread Analysis by Loan Type | Dual-Axis Chart | Are riskier loan types priced appropriately? |
| 9 | Multivariate Demographic Risk Matrix | Radar Chart | Which age + income + gender combinations carry most risk? |
| 10 | Average Credit Score by Gender | Bar Chart | Are there credit quality differences across demographic groups? |
| 11 | Collateral Risk by Occupancy Status | Clustered Bar | Do investment properties carry higher collateral risk? |

## Filters & Drilldowns

The dashboard supports interactive filtering by Region, Loan Type, LTV Risk Bucket, and Credit Score Band. Users can isolate a specific market segment to evaluate its risk profile independently, **enabling targeted policy recommendations**.

## Snapshots of the Dashboard



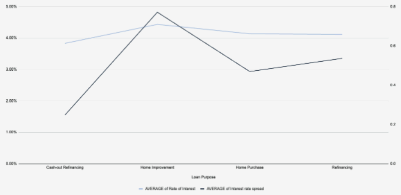
STRATEGIC CREDIT RISK & PORTFOLIO INSIGHTS

| | |
|---|---|
| Region All | Total Capital Exposure by Geographical Region |
| Region All | Geographical Default Concentration |
| Loan Purpose All | Loan Default Distribution by Purpose |
| LTV Risk Bucket All | Portfolio Vulnerability by LTV Risk Buckets |
| Credit Score Bucket All | Credit Score Segment Analysis |
| LTV Risk Bucket All | Exposure by LTV Risk Bucket |
| Occupancy Type All | Valuation Dispersion by Occupancy and Structure |
| Loan Purpose All | Interest Yield and Market Spread Analysis by Loan Type |
| Min Age All | Multivariate Demographic Risk Matrix |
| Gender All | Average Score by Gender |
| Occupancy Type All | Collateral Risk by Occupancy Status |

Overall Credit Default Volume
Total Distribution of Delinquent vs. Performing Assets
Yes
23....
No
76....
23.67%
Portfolio Default Rate %
Advance Approval Pipeline Status
Breakdown of Pre Approved vs. Non Pre Approved Application Volume
No
Yes
0
2000
4000
6000
8000
5277
2077
1355
250
Not Pre Approved
Pre Approved
Approval in Advance
0.18
Approval Effectiveness Ratio
Borrower Economic Profile
Comparison of Average Income and Loan Amounts by Credit Worthiness Rating
Credit Worthiness
Non-Standard
Standard
$329,171.76
$330,077.60
$0.00
$100,000.00
$200,000.00
$300,000.00
$400,000.00
AVERAGE of Loan Amount
AVERAGE of Income
$316,525.35
Average Exposure per Defaulted Borrower
Total Capital Exposure by Geographical Region
Distribution of Total Loan Amount across North, South, Central, and North-East
Central
5.6%
South
42.9%
North
50.8%
North...
0.7%
$749,215,500.00
Exposure at Default
Geographical Default Concentration
Total Default Events Segmented by Market Territory
Region
Central
North
North-East
South
424
155
3967
1102
47
3195
1054
0
1000
2000
3000
4000
No
Yes
22.70%
Loss Contribution %

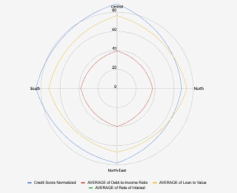
### Interest Yield and Market Spread Analysis by Loan Type

Analysis of Average Interest Rates and Market Spreads per Loan Purpose



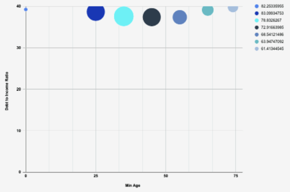
### Regional Risk Signature & Financial Profile

Normalized Performance Metrics across Geographical Markets



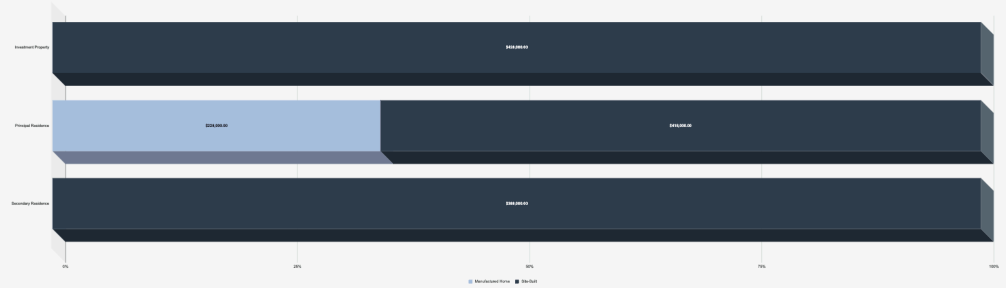
### Multivariate Demographic Risk Matrix

Exposure Analysis: Equity (LTV) vs. Debt Stress (DTI)



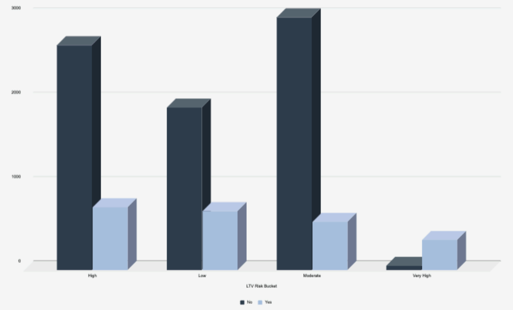
### Valuation Dispersion by Occupancy and Structure

Median Property Values categorized by Occupancy Type and Construction Method

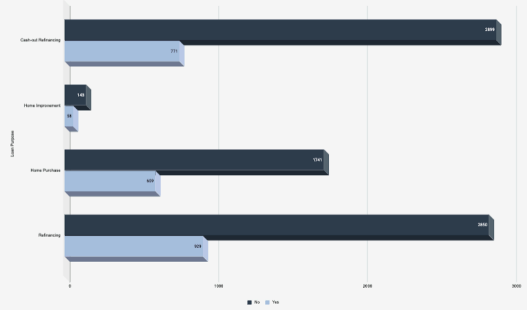


### Portfolio Vulnerability by LTV Risk Buckets

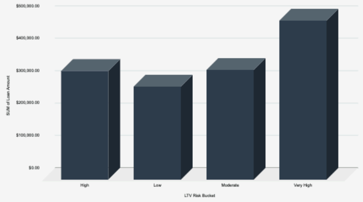
Volume of Defaults vs. Non-Defaults across Loan-to-Value Categories



### Loan Default Distribution by Purpose

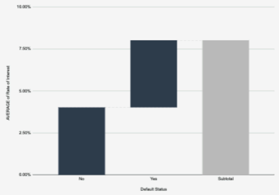


### Exposure by LTV Risk Bucket



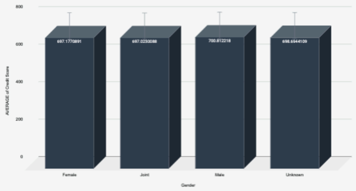
### Yield Variance by Loan Performance

Comparative Analysis of Weighted Average Interest Rates for Defaults vs. Non-Defaults



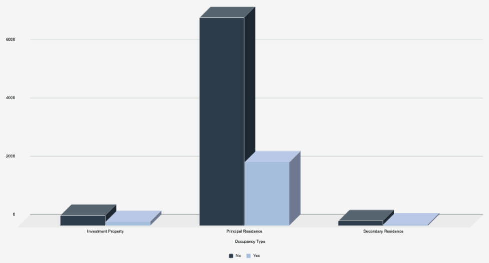
### Average Score by Gender

Benchmark of applicant creditworthiness across gender classifications



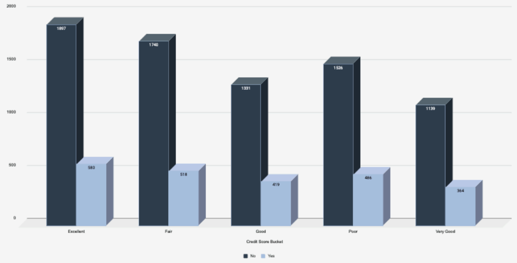
### Collateral Risk by Occupancy Status

ncidence of Defaults across Principal, Investment, and Secondary Residences



### Credit Score Segment Analysis

Default Concentration within Standardized Credit Score Bands



## 10. Insights Summary

- Overall portfolio default rate is 23.67% - **nearly 1 in 4 loans defaults**, indicating an elevated-risk book.

- $749.2 million is at risk from defaulted borrowers against a $3.3 billion portfolio - **a 22.7% capital exposure**.

- **North-East is a disproportionate risk pocket**, with a 35.62% default rate despite small loan volume.

- Very High LTV loans (LTV > 100%) default at 88.12% - **these loans should not exist in a well-managed portfolio.**

- **Credit score is a poor standalone predictor of default** - all five credit score bands default within a narrow 22–24% range.

- FHA loans default at 34.01%, **the highest among loan types** - nearly 12 points above Conventional loans.

- Joint applicants default at only 19.65% - **the lowest across all demographic groups** - validating dual-income risk mitigation.

- **Investment properties default 4.8 percentage points higher than principal residences**, reflecting strategic default risk.

- DTI for defaulters (39.66) is modestly elevated versus non-defaulters (37.44), **confirming income stress as a contributing factor**.

- Average credit scores of defaulters (697.9) and non-defaulters (698.8) are nearly identical, **challenging traditional credit-score-centric underwriting**.

- Moderate LTV (60–80%) has the lowest default rate (16.02%), **identifying the optimal risk zone for mortgage lending**.

## 11. Recommendations

### 1. Enforce a Hard LTV Cap at 100%

- **Mapped Insight:** Very High LTV loans default at 88.12%.
- **Action:** Decline or require additional collateral for any loan with LTV > 100%. Introduce a maximum LTV of 95% for FHA loans specifically.
- **Business Impact:** Eliminating Very High LTV approvals would reduce portfolio defaults by ~356 cases and cut EAD by an estimated $125–150M.
- **Feasibility:** High — a policy change requiring only updated underwriting guidelines.

### 2. Apply Regional Risk Premiums for North-East and Central

- **Mapped Insight:** North-East (35.62%) and Central (26.77%) exceed portfolio average by 12 and 3 points respectively.
- **Action:** Introduce 25–50 bps additional risk premium in interest rates for these regions, or require lower maximum LTV.
- **Business Impact:** Better risk-adjusted pricing protects margins; reduced origination in high-risk regions reduces NPA exposure.
- **Feasibility:** High — adjustable via pricing models without requiring regulatory approval.

### 3. Tighten FHA Loan Approval Standards

- **Mapped Insight:** FHA loans default at 34.01% — the highest of any loan type.
- **Action:** Require minimum credit scores of 620+ for FHA approvals; lower maximum LTV from current levels; add mandatory income verification.
- **Business Impact:** Even a 5% reduction in FHA default rate would preserve ~$25M in portfolio value.
- **Feasibility:** Medium — some constraints may require coordination with FHA program guidelines.

### 4. Incentivize Joint Applications

- **Mapped Insight:** Joint applicants default at 19.65% — nearly 4 points below portfolio average.
- **Action:** Offer preferential interest rates (10–15 bps discount) and slightly higher LTV allowances for joint applications.

- **Business Impact:** Shifting 10% of individual applications to joint could improve portfolio default rate by ~0.5%.
- **Feasibility:** High — a marketing and pricing adjustment.

### 5. Replace Single-Factor Credit Score Screening with Dual-Trigger Model

- **Mapped Insight:** Credit score alone shows near-zero predictive power — all bands default within 22–24%.
- **Action:** Implement a dual-trigger underwriting model requiring both LTV ≤ 80% AND DTI ≤ 36% for standard approval. Borrowers meeting only one criterion should face enhanced review.
- **Business Impact:** Multi-factor screening is significantly more predictive than credit score alone, based on the LTV and DTI patterns observed.
- **Feasibility:** Medium to High — requires credit model reconfiguration but uses existing data fields.

# 12. Impact Estimation

## LTV Cap Policy

- **Loans declined:** 404 (4.04% reduction in approvals)
- **Defaults avoided:** 356
- **Estimated EAD reduction:** $125–150 million
- **Portfolio default rate improvement:** 23.67% → ~20.93%

## FHA Tightening

- FHA default rate reduced from 34.01% to 29% (5-point improvement)
- **Defaults avoided:** ~69 loans
- **Capital preserved:** ~$23–28 million

## Joint Application Incentive

- Joint share increases from 28.25% to 35% of applications
- **Portfolio-level default rate reduction:** approximately 0.3–0.5 percentage points
- **Estimated annual savings:** $10–15M based on current portfolio size

## Dual-Trigger Underwriting

Adopting LTV + DTI dual-trigger screening could reduce approval of high-default-probability borrowers who currently pass credit score screening. Even a 2% improvement in post-origination default rate across the portfolio represents $66M in reduced EAD.

## 13. Limitations

- **Single-year dataset (2019)**: No multi-year trends available; cannot assess whether 23.67% is a structural or cyclical default rate.

- **No macroeconomic context**: Unemployment rates, interest rate environments, and housing market conditions at origination are not included.

- **Missing value imputation risk**: Median imputation for income, property value, and interest rate assumes the distribution of missing values mirrors observed values, which may not hold.

- **Credit score is FICO-approximated**: Different credit bureaus (EXP, CIB, CRIF, EQUI) may use different scoring scales, making cross-bureau comparison imprecise.

- **No delinquency timeline**: The dataset records only final default outcome, not when in the loan's term the default occurred. Early defaults carry different risk implications than late-stage defaults.

- **Unknown gender category**: 2,487 records with Unknown gender limit the precision of demographic risk analysis.

- **No cost or recovery data**: Loss Given Default (LGD) cannot be computed — the analysis estimates exposure but not net loss after recovery.

## 14. Future Scope

- **Multi-Year Analysis:** Incorporating 2017–2022 data would enable detection of cyclical default patterns correlated with housing market cycles.
- **Predictive Default Modeling:** A logistic regression or decision tree model using LTV, DTI, Loan Type, and Region as features could provide borrower-level default probability scores.
- **Loss Given Default (LGD):** Adding property auction recovery data would enable full Expected Loss (EL = PD × LGD × EAD) calculation.
- **Macro-Integration:** Overlaying unemployment rates or Fed Funds Rate at origination could improve model accuracy by accounting for economic conditions.
- **Early Warning System:** Time-to-default analysis using additional servicing data could enable proactive intervention for at-risk borrowers.

Looker Studio Migration: The current Google Sheets dashboard could be migrated to Looker Studio for enhanced interactivity, automatic refresh, and management-level access.

## 15. Conclusion

This project developed a comprehensive, KPI-driven credit risk analysis of a 10,000-record mortgage loan portfolio using Google Sheets as the primary analytical platform. **The analysis revealed a portfolio operating at a 23.67% default rate, with $749.2 million in exposure at default - driven disproportionately by Very High LTV loans, FHA-backed mortgages, the North-East and Central regions, and investment property borrowers.**

Critically, **the analysis challenges the conventional over-reliance on credit scores in mortgage underwriting**: all five credit score bands exhibit nearly identical default rates, while LTV bucket demonstrates default rates ranging from **16% (Moderate) to 88% (Very High)**. This finding has significant implications for how the institution designs its credit approval framework.

By implementing targeted policy changes - LTV caps, regional risk premiums, FHA tightening, and a dual-trigger underwriting model - the institution could reduce defaults by an estimated 400–500 cases annually and preserve over $150M in portfolio value. **The interactive dashboard provides an ongoing tool for credit risk monitoring and supports data-backed decisions at the portfolio management level.**

**This project demonstrates how structured EDA, KPI frameworks, and dashboard analytics can transform raw loan origination data into actionable business intelligence for risk reduction and strategic capital deployment.**

# 16. Appendix

## A. Data Dictionary

| Column Name | Original Values | Cleaned Values | Transformation Applied |
|---|---|---|---|
| Loan Limit | cf / ncf | Conforming Loan / non Conforming Loan | Mode fill + text expansion |
| Gender | Male/Female/Joint/ Sex Not Available | Male/Female/Joint/Unknown | Renamed 'Sex Not Available' to 'Unknown' |
| Approved in Advance | pre / nopre | Pre-Approved / Not Pre-Approved | Mode fill + text expansion |
| Loan Type | type1/type2/type3 | Conventional / FHA / VA | Text expansion |
| Loan Purpose | p1/p2/p3/p4 | Home Purchase / Home Improvement / Refinancing / Cash-out Refinancing | Text expansion |
| Negative Amortization | neg_amm / not_neg | Yes / No | Binary standardization |
| Interest Only | int_only / not_int | Yes / No | Binary standardization |
| Lump Sum Payment | lpsm / not_lpsm | Yes / No | Binary standardization |
| Region | north/south/central | North/South/Central/North-East | =PROPER() standardization |
| Rate of Interest | Missing values present | Median imputed | ARRAYFORMULA + MEDIAN |
| Age | Single range field (e.g., 35-44) | Min Age / Max Age (two columns) | Split into two numeric columns and used if else conditions to fill the missing values |
| LTV | Calculated field | Recalculated: (Loan Amount/Property Value)×100 | Formula correction |
| LTV Risk Bucket | Not present | Low/Moderate/High/Very High | Engineered via helper table using Xlookup |
| Credit Score Bucket | Not present | Poor/Fair/Good/Very Good/Excellent | Engineered via helper table using Xlookup |

## B. KPI Formulas

| KPI | Formula Used in Google Sheets |
|---|---|
| Portfolio Default Rate % | =COUNTIF(Default_Status,"Yes")/COUNTA(ID)*100 |
| Exposure at Default | =SUMIF(Default_Status,"Yes",Loan_Amount) |
| Loss Contribution % | =EAD / SUM(Loan_Amount) * 100 |
| Avg Exposure per Defaulter | =EAD / COUNTIF(Default_Status,"Yes") |
| High Risk Borrower % | =COUNTIF(LTV_Risk_Bucket,"Very High")/COUNTA(ID)*100 |
| Income Stress Ratio | =AVERAGEIF(Default_Status,"Yes",DTI) / AVERAGE(DTI) * 100 |
| Prime Borrower Ratio | =COUNTIF(Credit_Score_Bucket,"Excellent")/COUNTA(ID)*100 |

## 17. Contribution Matrix

This matrix documents each team member's contribution across project phases. Contributions are verifiable through Google Sheets Version History and the Logs tab.

| Team Member | Dataset & Sourcing | Cleaning | KPI & Analysis | Dashboard | Report Writing | PPT | Overall Role |
|---|---|---|---|---|---|---|---|
| Gokul | ✓ | ✓ | ✓ | ✓ | ✓ | — | Project Lead & Strategy Lead |
| Nachiket | ✓ | ✓ | — | — | — | — | Data Cleaning Lead |
| Isha | ✓ | ✓ | — | ✓ | — | ✓ | Data Cleaning Lead & PPT Lead |
| Saksham | ✓ | ✓ | — | — | — | — | Data Cleaning & Strategic Lead |
| Adarsh | — | — | — | — | — | ✓ | PPT Lead |
| Abhiney | ✓ | — | — | — | — | ✓ | PPT Lead |

Declaration: We confirm that the above contribution details are accurate and verifiable through Google Sheets Version History and submitted artifacts.

Team Signature: Team Signature