Unlocking Equity & Efficiency: A Data-Driven Review of India’s Crop Insurance Schemes (2018–2022)

Project Overview

This report presents a unified, in-depth review of India’s two flagship crop-insurance programs—the **Pradhan Mantri Fasal Bima Yojana (PMFBY)** and the **Weather Based Crop Insurance Scheme (WBCIS)**. The analysis is based on a dataset of **6,161 district-level entries** from 625 districts across 26 states, covering the five-year period from 2018 to 2022. The objective is to identify critical patterns in the schemes' operational scale, demographic inclusivity, geographic distribution, and financial dynamics to inform data-driven policy decisions.

Report Structure:

This document follows a systematic analytical workflow:

1. **Data Exploration & Diagnostics:** Initial assessment of the raw dataset's quality and integrity.
2. **Data Cleaning Process:** Detailed steps taken to validate, clean, and prepare the data for analysis.
3. **Descriptive & Statistical Analysis:** A summary of the dataset's statistical properties and feature classifications.
4. **Univariate Analysis:** Exploration of individual variables to understand their distributions.
5. **Bivariate Analysis:** Analysis of relationships between pairs of variables.
6. **Multivariate Analysis:** Investigation of complex, multi-variable interactions, including Principal Component Analysis (PCA).
7. **Key Findings & Summary:** A consolidated summary of the most critical insights derived from the analysis.
8. **Conclusion & Strategic Recommendations:** Actionable policy implications based on the evidence.

1. Data Exploration & Diagnostics

An initial diagnostic of the raw dataset (6,161 rows × 28 columns) revealed several key quality issues that required attention before analysis.

Data Quality Issue	**Diagnostic Finding**
Farmer Count Discrepancy	The farmer_count column was inconsistent with the sum of loanee and non_loanee fields (mean difference of ~35,169).
Missing & Invalid Data	~12% of rows were missing district_code, ~1.1% were

	missing demographic data, and 11 rows had invalid negative percentages.
Zero-Value Records	A significant number of rows contained zero values for sum_insured, gross_premium, and area_insured.
Data Integrity	Financial data was highly consistent (premium_shares summed to gross_premium), but demographic percentages required validation.

2. Data Cleaning Process

A systematic, 7-step cleaning strategy was applied to a copy of the dataset, resulting in a final, reliable dataset of **5,694 rows** (~7.6% reduction).

Step	**Action**	**Justification**
1	**Drop Invalid Rows**	Removed 10 rows with negative demographic percentages, as these are clear data entry errors.
2	**Reconcile Farmer Count**	Created a new, reliable total_policies column from the sum of loanee and non_loanee; dropped the original, misleading columns.
3	**Handle Invalid Zeros**	Dropped 457 rows representing logically impossible scenarios (e.g., positive policies but zero insured sum).
4	**Preserve Valid Zeros**	Intentionally retained rows where both policy count and insured values were zero, as these represent valid "zero-uptake" findings.
5	**Clip Demographic Bounds**	Corrected any minor data entry errors by clipping demographic percentages to the valid [0, 100] range.
6	**Correct Data Types**	Converted identifier columns like district_code to appropriate integer formats (Int64) for proper analysis.
7	**Reset Index**	Finalized the cleaned dataset with a clean, sequential index.

3. Descriptive & Statistical Analysis

3.1. Visual Analysis Highlights

- **Distributions:** Histograms revealed that all key financial and scale metrics (sum_insured, total_policies, etc.) are heavily **right-skewed**, indicating that a majority of records have low values with a long tail of high-value outliers.
- **Categorical Frequencies:** Bar charts confirmed that **PMFBY** has far more records than WBCIS, and the **Kharif** season has slightly more entries than the Rabi season.
- **Outliers:** Box plots visually confirmed the presence of extreme outliers in financial metrics, necessitating the use of log scales in many visualizations.

3.2. Feature Classification

Classification	**Columns**
Categorical (Nominal)	season, scheme, state_name, district_name
Numerical (Discrete)	year, state_code, district_code, iu_count, total_policies
Numerical (Continuous)	area_insured, sum_insured, gross_premium, all share and demographic % columns

4. Univariate Analysis

Objective:

The univariate analysis examined each feature in isolation to assess distribution, central tendency, spread, and data quality. This foundational step supports accurate modeling and interpretation in subsequent multivariate analyses.

Variable Classification:

Features were systematically categorized into five types:

Category	**Variables**
A. **Categorical (Nominal)**	season, scheme, state_name, district_name
B. **Numerical (Discrete)**	year, state_code, district_code, iu_count, total_policies
C. **Numerical (Continuous)**	area_insured, sum_insured, gross_premium, farmer_share, goi_share, state_share
D. **Demographics**	male, female, transgender

(Continuous %)	
E. **Caste & Farmer Type (%)**	sc, st, obc, gen, marginal, small, other

A. Categorical Variables

- **Scheme**:
  - **PMFBY** is the predominant scheme.
  - **WBCIS** appears regionally limited with lower record counts.
- **Season**:
  - **Kharif** dominates participation, aligning with India's primary growing season.
- **Yearly Trend**:
  - Policies increased steadily from 2018 to 2022, peaking in 2021–22, likely due to expanded adoption or policy mandates.

B. Numerical (Discrete)

- **IU Count & Total Policies**:
  - Extremely **right-skewed**; a few districts account for most policies.
- **Year, State Code, District Code**:
  - Uniformly distributed; behave like categorical identifiers.

C. Financial Metrics (Continuous)

Metric	**Skewness**	**Insight**
Area Insured	+11.36	Highly skewed; dominated by small plots, few extremely large outliers.
Sum Insured	+3.96	Long-tailed; mean far exceeds median.
Gross Premium	+5.25	Rare, high-value premiums distort the average.

Premium Shares (Farmer, GoI, State)	All right-skewed	Heavy regional/subsidy-driven variability.

D. Demographics (%)

Group	**Avg Proportion**	**Skewness**	**Interpretation**
Male	~86%	−2.11	Majority male participation; many districts ~100%.
Female	~14%	+2.01	Underrepresented; some districts near 0%.
Transgender	~0.02%	+42.34	Extremely rare entries; outlier-driven.

E. Caste & Farmer Type (%)

Category	**Avg Proportion**	**Skewness**	**Key Insight**
Gen > OBC > SC > ST	—	SC: +3.61 ST: +2.52	SC/ST concentrated in fewer districts.
Small Farmers	~60.7%	−0.46	Most represented; relatively balanced.
Marginal	~21.5%	+1.43	High presence in some areas; skewed.
Other Farmers	~17.8%	+1.94	Least represented and more varied.

Summary of Skewness (Highlights)

Column	**Skewness**	**Interpretation**
area_insured	+11.36	Few districts cover very large land areas.
gross_premium	+5.25	Some policies are exceptionally expensive.
farmer_share	+4.16	Uneven farmer contributions across districts.
transgender	+42.34	Almost all values are near 0 with rare spikes.

male	−2.11	Strong male-dominated participation.
small	−0.46	Slightly left-skewed; consistent participation.

Key Insights & Recommendations

- **High Skewness**: Most financial and participation metrics are **highly right-skewed**, suggesting the need for **log or quantile transformations** before modeling.
- **Scheme & Season Dominance**: **PMFBY** and **Kharif** season dominate, highlighting a **core segment** of insurance activity.
- **Gender & Caste Gaps**: Participation is heavily **male and General caste-skewed**. This signals a need for **inclusive policy interventions**.
- **Farmer Profile**: The dataset primarily serves **small and marginal farmers**, aligning with subsidy targets but requiring consideration for their limited land size.
- **Data Preparation**: Visual and statistical diagnostics point to essential preprocessing steps: **handling outliers**, **scaling**, and **transforming skewed features** for fair modeling.

5. Bivariate Analysis Report

Objective:

Examine relationships between pairs of variables to uncover key trends in insurance scale, cost structures, demographics, and regional patterns.

Key Variable Relationships:

#	Variable Pair	Pearson r	Strength & Insight	Use / Interpretation
1	Sum Insured vs. Area Insured	≈ 0.60	Moderate positive correlation; larger land = higher coverage	Identify anomalies (e.g., over-insured small plots)
2	Gross Premium vs. Total Policies	≈ 0.72	Strong positive correlation	Analyze cost variability and policy efficiency
3	Sum Insured vs. Gross Premium	≈ 0.95	Near-perfect linearity; standardized premium rates	Simplifies pricing and forecasting models
4	Total Policies vs. Area Insured	≈ 0.35–0.45	Weak-to-moderate; more policies ≠ more land	Reflects high participation by small/marginal farmers

5	Gross Premium vs. Area Insured	≈ 0.58	Moderate trend; influenced by crop type/value	Use in premium load adjustment based on crop risk
6	Farmer Share vs. Gross Premium	≈ 0.85	Strong linearity; farmer contributions scale with premiums	Useful proxy for affordability analysis
7	% Male vs. Total Policies	≈ 0.05	Negligible relationship	Gender mix does not drive policy volume
8	% Small Farmers vs. Total Policies	≈ –0.02	No correlation	Farmer size distribution has no direct impact on policy uptake

Scheme & Season Comparison

- **PMFBY vs. WBCIS**:
  - PMFBY has higher values in sum insured, premiums, and policy count.
  - WBCIS is niche and lower-scale.
- **Kharif vs. Rabi**:
  - Kharif dominates in land and sum insured; Rabi remains secondary.

State-Level Highlights

- **Maharashtra**: High premium, moderate scale → risk-based pricing.
- **Rajasthan**: Highest policies → scale-driven adoption.
- **Himachal Pradesh**: Large area, low premium → low-value crops.
- **Andaman & Nicobar**: Low volume, high premium → specialized market.

Conclusion

- **Strong financial coupling**: Premiums tightly linked to sum insured and farmer share.
- **Land–policy decoupling**: High policy volume ≠ large land coverage.
- **Demographics neutral**: Gender and smallholder share don't predict uptake.
- **PMFBY & Kharif dominate**: Largest scale and value across metrics.
- **Regional variation**: States differ in premium structures and coverage—calls for localized policy tuning.

6. Multivariate Analysis Report

Objective

To analyze complex interactions among financial metrics, demographics, geography, schemes, and seasons using multivariate techniques like PCA.

Key Insights:

1. Scheme Dynamics

- **PMFBY** covers wide financial range; **WBCIS** operates on a smaller scale, with **AP as an**

outlier (avg. premium ₹5,589).

- Both follow a **uniform premium-to-sum insured ratio** (r ≈ 0.95).
- **Farmer Premium Share**: PMFBY (14%), WBCIS (19%).

→ PMFBY is dominant in scale; WBCIS is niche with higher farmer cost.

2. Demographics

- High policy volumes seen in **male-majority districts** (80–85%).
- No clear link between **farmer type** (small/marginal) and area insured.

→ Inclusion doesn't equate to coverage; design must target fragmented land and female outreach.

3. Financial Relationships

- **Strong links**: Sum ↔ Premium (r ≈ 0.95).
- **Moderate links**: Area ↔ Sum/Premium (r ≈ 0.58–0.60).

→ Focus on financial indicators for modeling—area alone is less predictive.

4. Seasonal Patterns

- **Kharif and Rabi** seasons have similar average scale and coverage.

→ Seasonality has minimal impact compared to regional or demographic factors.

5. Subsidy Composition

- Avg. Premium: PMFBY ₹2,440; WBCIS ₹1,380.
- Share Split:
  - PMFBY: Farmer 14%, GoI 44%, State 42%
  - WBCIS: Farmer 19%, GoI 40%, State 41%

→ Both rely on subsidies; WBCIS shifts more burden to farmers.

6. Regional Variation

- **High premiums** in Rajasthan, Maharashtra, AP (PMFBY).
- **WBCIS premiums** in AP far exceed other states.

→ Region-specific pricing and subsidy strategies needed.

7. Correlation Matrix

- Strong financial correlations:
  - sum_insured ↔ gross_premium (0.88),
  - total_policies ↔ sum_insured (0.71).

- Strong demographic patterns:
  - male vs female: –1.00,
  - small vs marginal: –0.51.
- Scheme/season have **weak linear relationships** with other variables.

→ Financial metrics cluster tightly; scheme/season are secondary influencers.

8. Principal Component Analysis (PCA)

Component	Captures	Top Loadings
PC1	Financial Scale	Premium, sum_insured, policies, area
PC2	Demographic Inclusivity	Female (+), Male (–), Small Farmer (+)

- PMFBY = High PC1 (scale); WBCIS = Low PC1
- Kharif/Rabi = No clear separation

→ Two key dimensions: Scale vs Inclusivity.

Summary Table

Dimension	Conclusion
Financial Scale	Driven by PMFBY; large, consistent premiums and subsidies
Demographic Equity	Needs improvement—female and smallholder inclusion lags
Seasonality	Minimal influence on scale or uptake
Regional Differences	Significant—requires custom subsidy & risk models per state

7. Key Findings

7.1. National Scale & The “Vulnerable Farmer Paradox”

The national trend shows explosive growth in reach but not in the physical or financial depth of coverage per policy. The number of farmer policies surged by **80%** (from ~58M to ~104M), but the total insured area and average sum insured declined concurrently. This is explained by the scheme's deep penetration among **small and marginal farmers**.

7.2. Demographic Inclusivity & Equity

The schemes have succeeded in reaching their core target of smallholders, but significant equity gaps remain. **Small and Marginal Farmers** make up **81.1%** of beneficiaries. However, a significant **gender gap** persists, with **female farmers** comprising only **14–20%** of the total.

7.3. Geographic Patterns & Regional Disparities

Insurance activity is not uniform. A high-volume **"Insurance Heartland"** exists in **Maharashtra, Rajasthan, Madhya Pradesh, and Karnataka**. States exhibit different profiles, with **Rajasthan** being a **volume leader**, **Telangana & Haryana** being **value leaders**, and **Maharashtra** being the **costliest** in terms of average premium.

7.4. District-Level Insights

- **Policy Hotspots (>5M policies):** Hanumangarh, Beed, Sri Ganganagar.
- **High-Value Districts (>₹1 Lakh/policy):** Nuh, Nellore, Warangal.
- **Seasonal Volatility:** Districts like Bilaspur show extreme cyclical swings, reflecting the Kharif-Rabi seasons.

8. Summary & Conclusion

This analysis reveals that India's crop insurance landscape is defined by two dominant, largely independent dimensions: **Financial Scale** and **Demographic Inclusivity**. The PMFBY scheme is the clear engine of scale, successfully reaching an unprecedented number of small and marginal farmers. However, this success in volume has not yet translated into equitable social outcomes, with a significant gender gap and regional disparities remaining key challenges. The data confirms that uptake is driven more by geography and crop patterns than by the demographic mix of a region.

Therefore, a "one-size-fits-all" policy is insufficient. The path forward requires a more nuanced, data-driven approach that uses advanced segmentation (like the PCA-derived axes) to design state-sensitive models that can simultaneously optimize for both financial efficiency and social equity.