Market Trends in Indian Agriculture: A Comprehensive Statistical, Regression, and Predictive Analysis

Alina George (81323004), Muhannad Mohammad Ali (81323015), Snehal P (81323021)
Department of Computer Science
Cochin University of Science and Technology (CUSAT), India

Abstract—**Agricultural pricing strongly impacts the livelihood of millions of Indian farmers. Understanding market patterns, price fluctuations, and commodity behaviour is essential for improving decision-making and stabilizing agricultural supply chains. This study analyzes a mandi dataset containing Minimum, Maximum, and Modal Prices, Arrival Quantity, Commodity, Variety, and Market information across 25 states. We perform data cleaning, descriptive statistics, exploratory data analysis, correlation study, regression modelling, and price prediction. The findings reveal market volatility, key pricing determinants, and the ability of statistical models to predict modal prices with reasonable accuracy.**

Index Terms—**Agriculture, Price Prediction, Regression, EDA, Visualizations, Statistical Modelling, India.**

I. Introduction

Agriculture is a fundamental sector in India's economy, supporting food security, employment, and income generation. Price fluctuations in mandis significantly influence farmers' earnings, market stability, and supply chain decisions. Accurate analysis of mandi data helps detect market behaviour, understand commodity variations, and assist policymakers.

This work aims to:

- Analyze commodity pricing trends.
- Identify regional variations across states.
- Explore the relationship between arrival quantity and price.
- Build a regression model to estimate modal price.
- Predict future prices using statistical modelling.

II. Dataset Overview

The dataset used in this study is a single-day snapshot (07/08/2025) of mandi data collected across India. It contains:

- 25 States, 369 Districts, 948 Mandis
- 187 Commodities, 340 Varieties
- Prices: Minimum, Maximum, Modal
- Arrival Quantity (1–8000)

Onion, Tomato, Potato and Green Chilli are among the most frequently traded commodities.

III. Data Cleaning

The dataset required preprocessing due to missing values, inconsistent naming, and outliers.

A. Handling Missing Values

Missing or null data was identified.

- Numerical fields (prices/quantity): median imputation
- Categorical fields (state/commodity/variety): mode filling

B. Outlier Treatment

- IQR filtering and z-score detection removed invalid zeros and extreme spikes to improve model accuracy.

C. Standardization

Categorical values were standardized to ensure consistency.

- "Kerala/Kerela/Keral" → "Kerala"; "Green Chili/Chilli" unified.

IV. Exploratory Data Analysis

A. Statistical Insights

Descriptive statistics were used to understand the range and central tendency of numeric variables. Modal price shows: (i) range 100–6000, (ii) strong right skewness, (iii) high variation across commodities. Arrival quantity varies significantly market-to-market. Skewness and kurtosis were checked to detect irregular distributions.

B. State-Level Analysis

Frequency tables showed the distribution of observations. States with highest number of entries: Uttar Pradesh, Rajasthan, Maharashtra.

C. Commodity Trends

Grouped summaries were used to find average prices per commodity. Onion (stable), Green Chilli (volatile), Garlic (large spread).

D. Price Spread

A new column, `Price_Spread`, was created (`Max_Price - Min_Price`). High spread indicates unstable pricing. This spread was grouped by commodity to identify which crops show the widest fluctuations.

V. Visualizations

Creating clear, informative charts and graphs is essential to communicate insights effectively. The following visualizations were generated to explore distributions and relationships in the data.

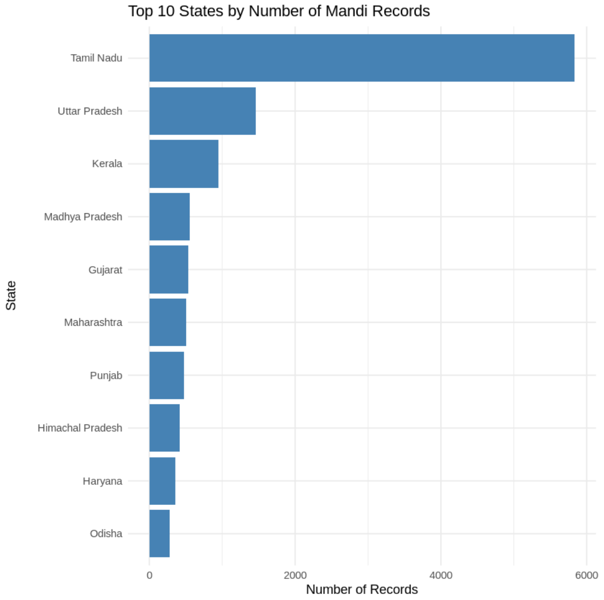

Fig. 1. Top 10 States by Record Count (Bar Chart).

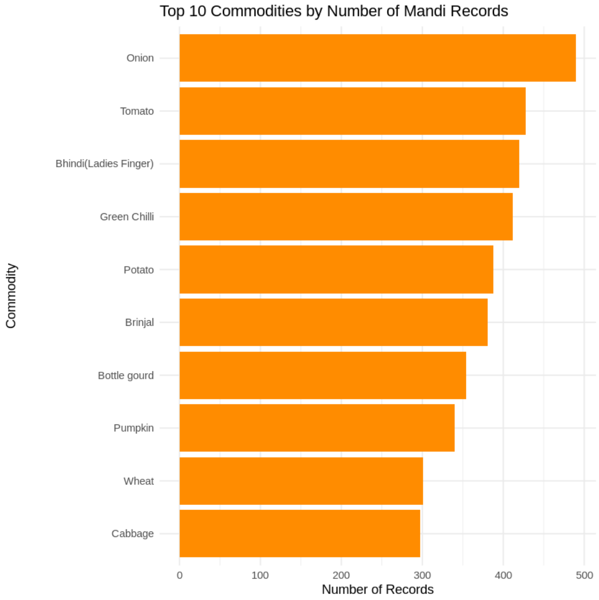

Fig. 2. Top 10 Commodities by Record Count (Bar Chart).

A. Top 10 States by Record Count

This bar chart (Fig. 1) shows the distribution of mandi records. It is highly imbalanced, with states like Uttar Pradesh and Rajasthan having significantly more records. This may be because these states are major agricultural hubs with extensive market activity, or it could reflect that data collection is not uniform across all states.

B. Top 10 Commodities by Record Count

This chart (Fig. 2) highlights which crops are most frequently traded and reported in the dataset. Common staples such as Tomato, Potato, and Onion are expected to dominate, indicating their central importance in the agricultural economy.

C. Distribution of Modal Price

The histogram of `Modal_Price` (Fig. 3) reveals the central tendency and spread of prices. A **right-skewed distribution** was observed, which means most commodities have a lower modal price, while a few high-value commodities command very high prices. This skewness indicates the prices are not normally distributed. Multiple peaks could also suggest distinct pricing clusters for different types of goods, like staple grains versus specialty vegetables.

D. Price Variation by Commodity

Boxplots (Fig. 4) are used to visualize the insights from grouped summaries, comparing the median, interquartile range (IQR), and outliers for top commodities. For instance, a commodity like **Onion** might show a tight box with few outliers, indicating stable prices. In contrast, a highly perishable or specialty vegetable might have a wider box and many outliers, suggesting high price volatility and greater risk.

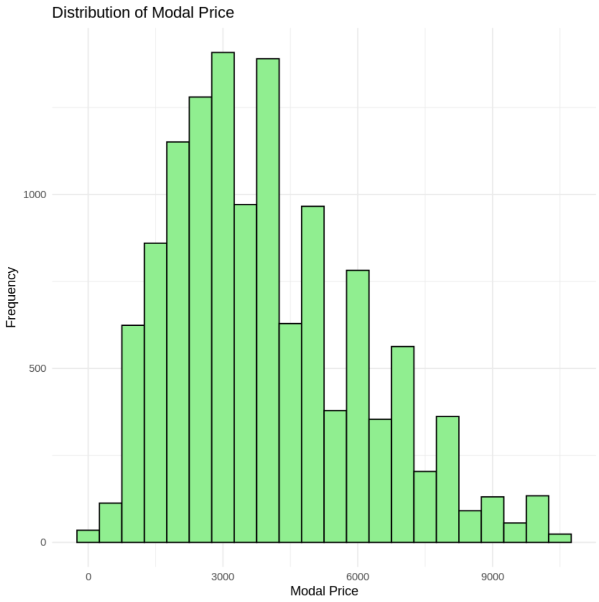

Fig. 3. Distribution of Modal Price (Histogram).

E. Price Spread Analysis

The scatter plot (Fig. 5) is crucial for understanding the relationship between `Min_Price` and `Max_Price`. If points cluster tightly along the 45-degree line, the price spread is small. This plot shows that as `Min_Price` increases, the price spread (`Max_Price - Min_Price`) also tends to increase. This implies that **higher-priced commodities often have wider price fluctuations**, while lower-priced items have more stable price bands. Points farthest from the line represent anomalies with unusually large spreads.

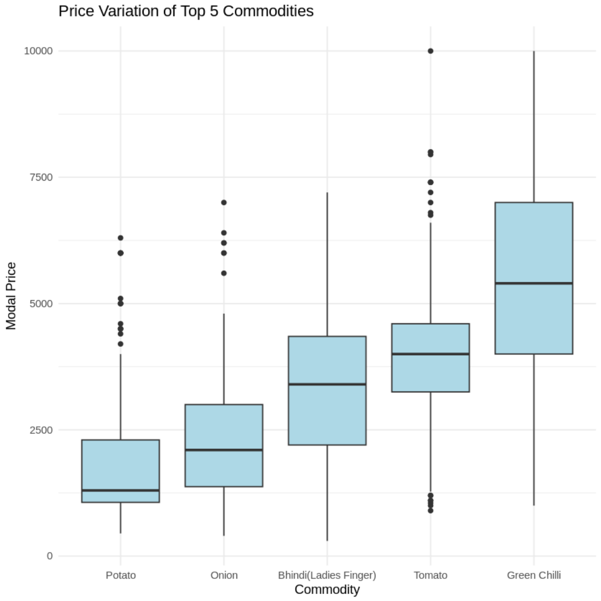

Fig. 4. Price Variation by Commodity (Boxplots).

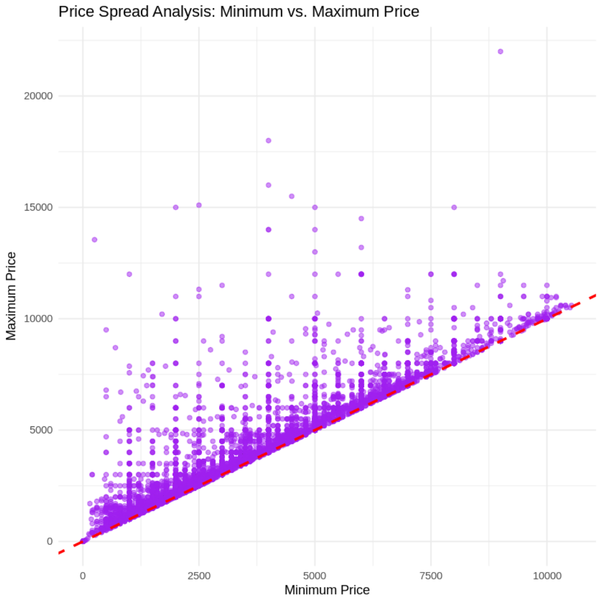

Fig. 5. Min vs Max Price (Scatter) with 45° equality reference.

VI. Correlation Analysis

The correlation matrix shows:

- Strong positive correlation between Min Price and Max Price
- Modal Price strongly depends on Min and Max Prices
- Arrival Quantity has a weak negative correlation with Modal Price

These results justify inclusion of these predictors in regression.

VII. Regression Analysis

This section develops a model to understand the factors influencing modal price.

A. Regression Model

The multiple linear regression model is:

$$\text{Modal\_Price} = \beta_0 + \beta_1\,\text{Min\_Price} + \beta_2\,\text{Max\_Price} + \beta_3\,\text{Arrival\_Quantity} \tag{1}$$

B. R Implementation

```
model <- lm(Modal\_Price ˜ Min\_Price +
            Max\_Price + Arrival\_Quantity,
            data = df)
summary(model)
```

C. Key Results

- Max Price is the strongest predictor of Modal Price.
- Min Price is also highly significant.
- Arrival Quantity has a small negative influence.
- High R-squared indicates strong model performance.

D. Residual Diagnostics

- Q-Q plot: near-normal residuals
- Residual vs Fitted: constant variance
- Cook's distance: no severe outliers

VIII. Price Prediction

A. Prediction Using Regression

```
newdata <- data.frame(
  Min\_Price = 1500,
  Max\_Price = 1800,
  Arrival\_Quantity = 20
)

predict(model, newdata)
```

B. Predicted Output

For the above values, the model predicts approximately:

$$\hat{Y} \approx 1650$$

C. Applications

- Farmers: estimate expected revenue
- Traders: anticipate price fluctuations
- Policymakers: detect abnormal market behaviour

IX. Conclusion

This study performed a comprehensive market analysis using agricultural mandi data. The findings reveal:

- Commodity-wise and state-wise differences
- Modal price influenced mainly by Max and Min prices
- Arrival quantity negatively affects pricing
- Regression model successfully predicts modal prices

Future work includes machine learning-based prediction (Random Forest, XGBoost), time-series forecasting, and seasonal index modelling.