Spark Mini Project Report

Prepared By: CH.Aditya

Course: Big Data Analytics

Date: October 2025

1. Dataset Description

Dataset Description

- **File name:** Market commodity prices.csv
- **Rows:** 16,039
- **Columns:** 10
- **Columns Detected:** State, District, Market, Commodity, Variety, Grade, Arrival_Date, Min_x0020_Price, Max_x0020_Price, Modal_x0020_Price
- **Data types:** mixture of categorical (State, District, Market, Commodity, Variety, Grade) and numeric (Min/Max/Modal price). Arrival_Date parsed as date (most entries on 06/10/2025).
- **Date range:** 06-Oct-2025 (dataset contains records for the same arrival date in this sample)

2. Observations from Executed Cells

- The CSV was successfully loaded into a Pandas DataFrame (16,039 rows × 10 columns).
- There are **no missing values** reported in the main columns after parsing (all columns have 0 missing counts in the provided sample).
- Arrival_Date values parse to a single date (2025-10-06) — indicates data is a daily snapshot.
- Price columns (Min_x0020_Price, Max_x0020_Price, Modal_x0020_Price) contain large ranges and clear outliers (Modal price min = 2, max = 102,500).
- The most frequent commodity in the dataset is **Onion** (top commodity by record count).

3. Plots Observed

The following plots were generated and saved to /mnt/data/report_outputs:

1. **Distribution of Modal Prices** — modal_price_hist.png
    - Interpretation: Modal prices are right-skewed with heavy tails; outliers exist at very high values.
2. **Top 10 Commodities by Record Count** — top_commodities_bar.png
    - Interpretation: Shows which commodities have the most observations (Onion is top). Useful to prioritize analysis.
3. **Top 10 States by Average Modal Price** — state_modal_top10_bar.png
    - Interpretation: Karnataka, Nagaland, Meghalaya, Tripura, Kerala and others show the highest average modal prices in the sample.
4. **Time Series — Average Modal Price for Top Commodity (Onion)** — top_commodity_timeseries.png
    - Interpretation: Because the snapshot is for a single arrival date, the time series is degenerate in this dataset; however the code groups by date and would show trends in multi-date datasets.

4. Key Insights

- **Dataset type:** This looks like a daily market snapshot dataset (single-date sample: 06-Oct-2025) for many markets across India.
- **Top commodity: Onion** is the most frequently reported commodity in the file.
- **Modal price central tendency:** mean Modal Price ≈ **₹4,551.68** (rounded). The Modal_x0020_Price values span a wide range (min 2, max 102,500) indicating extreme outliers or data-entry errors for some rows.
- **State-level variation:** Highest average modal prices (top 10 states) — **Karnataka, Nagaland, Meghalaya, Tripura, Kerala, Tamil Nadu, Himachal Pradesh, Jammu and Kashmir, Gujarat, West Bengal**. These states show higher average modal prices than others in this snapshot.
- **No missing data:** All columns present values — simplifies analysis, but still requires quality checks.

5. Recommendations

1. **Outlier handling:** Review rows with extremely high or low prices (e.g., Modal ≤ 2 or Modal ≥ 100,000). Investigate whether these are currency-unit mismatches,

reporting errors, or legitimate extreme values. Apply winsorization or remove obvious errors when computing aggregated metrics.

2. **Normalize price units:** Ensure all price values are in the same unit (per quintal, per kg, per 100 kg, etc.). The dataset does not include unit columns; confirm and add unit metadata if available.
3. **Add time coverage:** For trend analysis, collect or merge additional days. The current file is a snapshot — multi-day data will enable seasonality and volatility analysis.
4. **Add derived fields:** Compute Spread = Max - Min, Relative_Spread = (Max - Min)/Modal, and Price_Change across days (if available) to quantify market volatility.
5. **State/market-level dashboards:** Build alerts for unusual jumps in modal price at state or market level using z-score or rolling-window methods.
6. **Data validation rules:** Add automated checks in your ingestion pipeline, e.g., price >= 0, reasonable upper bounds, valid commodity names (use a master list to avoid typos).

6. Conclusion:-

This dataset is a rich daily snapshot of market commodity prices across Indian markets. The sample contains 16,039 records and is ready for exploratory analysis. Key actionable outcomes from the current snapshot are the identification of Onion as the most-observed commodity, existence of large price outliers, and a set of states with noticeably higher average modal prices.

To progress from snapshot analysis to operational insights, I recommend cleaning outliers, confirming price units, and ingesting additional dates to enable time-series and volatility analyses.