Report- Commodity Price Analysis Using PySpark

Dataset Description

The dataset used in this project contains agricultural commodity market information collected from multiple states and markets across India. It includes pricing details such as minimum, maximum, and modal prices along with categorical fields like State, District, Market, Commodity, Variety, and Grade. The dataset serves as a crucial representation of market behavior, supply patterns, and region-wise pricing differences.

Dataset Size and Structure

The dataset comprises several thousand rows and multiple columns representing both categorical and numerical attributes. It is structured in CSV format and loaded into PySpark for distributed processing.

The structure includes fields such as State, District, Market, Commodity, Variety, Grade, Arrival_Date, Min_Price, Max_Price, and Modal_Price.

Attribute Overview

State, District, Market: Categorical fields describing the geographic origin of the commodity data.

Commodity, Variety, Grade: Attributes describing the type and quality of the agricultural product.

Arrival_Date: The date when the commodity was recorded in the market.

Min_Price, Max_Price, Modal_Price: Numerical pricing attributes indicating market fluctuations.

Domain Context

The dataset belongs to the agricultural market analytics domain. It is useful for identifying market trends, analyzing price fluctuations, and studying region-wise commodity performance.

It can support studies related to crop economics, inflation analysis, supply-chain behavior, and predictive modeling for agricultural prices.

Data Characteristics

The dataset exhibits variation across states, markets, and commodity types.

Missing values were present in some pricing fields and categorical columns, requiring cleaning before analysis.

The dataset contains real-world market data that may include inconsistencies such as spelling variations, grade differences, and null entries.

Purpose of Use

The dataset was used to perform data cleaning, preprocessing, exploratory analysis, visualization, and correlation exploration using PySpark and Python.

Its primary goal is to understand price behavior, region-wise distribution, and top commodities across India.

Issues Observed

Missing values were found in commodity pricing and some categorical attributes.

Inconsistent attribute naming required column renaming for clarity.

Some rows contained invalid or blank entries requiring filtering.

Variation in data distribution across states created imbalance in analysis.

Duplicate records were detected and removed to preserve data integrity.

Overall Impact

The issues identified highlight the need for thorough data cleaning to ensure accuracy in price comparison and statistical analysis.

After cleaning, the dataset became suitable for further EDA, correlation analysis, and visualization.

Analysis and Observations

State-wise analysis revealed that certain states contribute more market entries than others.

Commodity frequency distribution showed which products dominate Indian markets.

Pricing analysis indicated strong correlations among Min_Price, Max_Price, and Modal_Price.

Scatter plots demonstrated positive relationships, confirming consistent pricing trends.

Heatmap visualization validated the strong correlation among numeric price attributes.

Visualizations

Heatmaps showed strong positive correlation between all price variables.

Bar charts highlighted top commodities and highest average pricing by state.

Pie charts illustrated grade distributions across markets.

Scatter plots depicted relationships between minimum, maximum, and modal prices.

Conclusion

The project successfully cleaned, analyzed, and visualized a large agricultural commodity dataset using PySpark.

Findings revealed price variation patterns, regional differences, and commodity dominance.

The study emphasizes the value of big data tools in agricultural analytics and market intelligence.

The results can support farmers, policymakers, and traders in making informed decisions.