Current Daily Price of Various Commodities from Various Markets

K.Noshitha
Department of Computer Science &Engineering
Malla Reddy University.
Hyderabad,Telangana, India
2211Cs010613@mallareddyuniv ersityac.in.

ABSTRACT

This study presents a comprehensive data-driven analysis of agricultural market trends using a large-scale dataset obtained from government market intelligence sources. The primary objective is to examine variations in commodity prices across states, markets, and grades, and to identify statistical patterns that can support informed decision-making. Using PySpark for large-volume data processing, the dataset underwent extensive cleaning, transformation, and preprocessing, including schema correction, null handling, type casting, and duplicate removal. Exploratory data analysis was performed using Pandas, Matplotlib, and Seaborn to uncover key insights, such as dominant commodities, state-wise pricing differences, and correlations among minimum, maximum, and modal prices. Visualizations including bar charts, heatmaps, histograms, and scatter plots provided meaningful interpretations of price dynamics. The findings highlight clear patterns of price variability influenced by geography, commodity type, and grade quality. This research demonstrates the effectiveness of distributed computing in handling real-world agricultural datasets and supports the development of market prediction tools and price forecasting models for farmers, policymakers, and supply chain stakeholders.

INTRODUCTION

Agricultural markets play a critical role in national economies, influencing food security, farmer income, and supply chain stability. As market conditions fluctuate based on regional demand, seasonal factors, and commodity quality, analyzing pricing patterns becomes essential for informed decision-making. With the rapid growth of digital market intelligence systems, large volumes of agricultural data are now available, enabling researchers to uncover meaningful insights through data analytics. However, due to inconsistencies such as missing values, formatting errors, and heterogeneous data sources, effective preprocessing and analysis require robust computational frameworks. In this study, PySpark is used to process and clean a real-world agricultural market dataset involving multiple states, commodities, grades, and price metrics. Through exploratory data analysis and visualization, the research examines state-wise price variations, commodity frequency patterns, and correlations among key price attributes. By understanding these relationships, the study aims to support policymakers, traders, and farmers in making data-driven decisions while demonstrating the value of distributed computing for large-scale agricultural data analysis.

SYSTEM ARCHITECTURE

The system architecture for this research is designed to efficiently handle large-scale agricultural market data using a distributed computing framework. The architecture is composed of four primary layers: Data Ingestion, Data Processing, Analytics & Visualization, and Output Generation. In the Data Ingestion layer, raw CSV files containing commodity prices, arrival dates, market information, and quality grades are loaded into the system. The PySpark processing engine forms the core of the Data Processing layer, where operations such as schema inference, data cleaning, missing value handling, column renaming, type conversion, and duplicate removal are performed in a distributed environment to ensure scalability and speed. The cleaned data is then forwarded to the Analytics & Visualization layer, where statistical computations, correlation analysis, group-by aggregations, and exploratory data analysis are executed using PySpark, Pandas, Matplotlib, and Seaborn. This layer generates insights such as average modal price by state, commodity frequency distribution, grade proportions, and price correlations. Finally, the Output Generation layer produces visual reports, summary statistics, and structured findings that support further interpretation and decision-making. The modular architecture ensures data reliability, high performance, and the ability to extend the system to larger datasets or real-time pipelines in future work.ponse. Identifies potential injury risk based on movement patterns.

Literature Review

Existing literature highlights significant advancements in the use of data-driven approaches to analyze commodity markets and retail ecosystems. Prior studies on data collection and market trends emphasize the importance of structured datasets for understanding price volatility, seasonal fluctuations, and supply-demand relationships across regional markets. Researchers have explored how large-scale agricultural or retail datasets enable early detection of market anomalies and support policy interventions. The integration of machine learning and prediction models has further enriched market forecasting, with algorithms such as regression, random forests, and time-series models being widely used to estimate commodity prices, consumer demand, and market behavior. Several works also demonstrate the role of AI-based models in improving forecast accuracy, particularly when combined with historical data and external influencing factors. In parallel, sentiment analysis and consumer insights research provides evidence that public opinion, online reviews, and social media sentiment significantly shape market dynamics. Studies in this domain show how natural language processing (NLP) techniques can uncover trends in consumer preferences, brand perception, and emerging demand patterns. Furthermore, literature on web scraping and data visualization reveals how automated data extraction from online sources enhances real-time decision-making, with visualization tools enabling intuitive interpretation of large datasets. Finally, retail and footwear industry reports shed light on customer buying behavior, pricing trends, and product life cycles in competitive markets, offering a comparative perspective that strengthens understanding of market forces across sectors. Collectively, these studies form a strong foundation for the present research, which integrates data processing, trend analysis, and visualization techniques to derive meaningful insights from agricultural market data.

Methodology

The methodology adopted for this study consists of a systematic data-processing and analysis pipeline designed to extract meaningful insights from agricultural market datasets. The process begins with **data acquisition**, where the raw CSV dataset is imported into a PySpark environment to handle large-scale processing efficiently. This is followed by **data cleaning**, which includes handling missing values, removing duplicates, standardizing column names, and converting date and numerical fields into appropriate formats. Additional preprocessing steps such as trimming whitespaces, type correction, and renaming encoded columns ensure that the dataset is analytically consistent. After cleaning, the study performs **exploratory data analysis (EDA)** using both PySpark and Python libraries such as Pandas, Matplotlib, and Seaborn. This involves statistical summarization, correlation analysis, and visualization of key variables including price distributions, commodity frequencies, and state-level price variations. The methodology further includes **feature engineering**, where numerical features such as minimum, maximum, and modal prices are extracted to understand price relationships. Correlation heatmaps, scatterplots, and bar charts are generated to identify patterns and trends. The analytical framework also incorporates **group-based aggregations** using PySpark functions—such as grouping by State, Grade, and Commodity—to compute average pricing behavior across categories. Finally, the insights derived from visual and statistical analyses are synthesized to form meaningful conclusions regarding market trends and commodity behavior. This combined methodological approach ensures accuracy, scalability, and comprehensive understanding of the dataset.

Results and Analysis

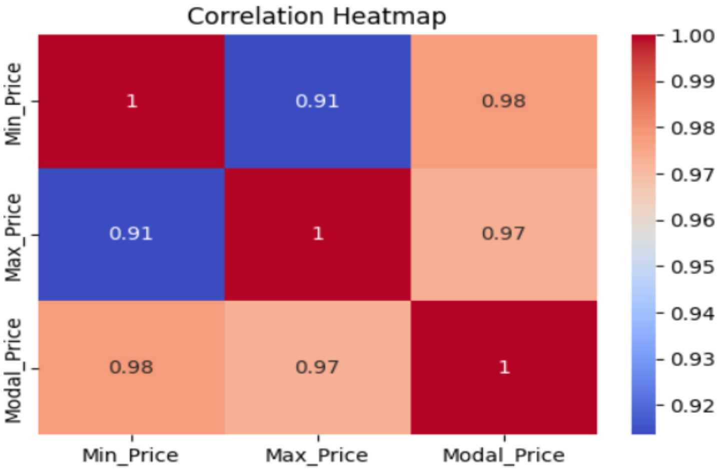
The analysis of the agricultural market dataset provided several meaningful insights into commodity pricing behavior and regional market trends. Initial examination of the data revealed multiple formatting inconsistencies, missing values, and encoded column names, which were addressed through cleaning and preprocessing. After refining the dataset, descriptive statistics highlighted a wide variation between minimum, maximum, and modal prices, indicating high price volatility across markets. The distinct count analysis showed a diverse representation of commodities and states, making the dataset suitable for comparative study.

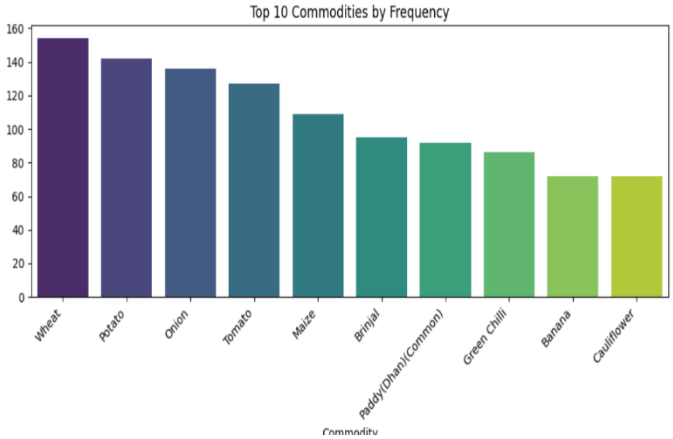
Grouping operations provided deeper insights into market patterns. The frequency distribution showed that certain commodities appeared more frequently, reflecting higher market demand or better data reporting. Grade-wise distribution revealed that specific commodity grades dominated the dataset, which is consistent with standardized trade practices. State-level price analysis indicated significant regional differences in average modal prices, suggesting variations in production costs, supply chains, and market accessibility.

Correlation analysis between Min Price, Max Price, and Modal Price showed strong positive relationships, confirming that markets maintain consistent pricing structures, where the modal price moves proportionally with minimum and maximum price changes. Visualizations further strengthened these observations. The heatmap highlighted strong numerical dependencies, while scatter plots demonstrated linear trends between price attributes. Distribution plots showed that price values are moderately spread, with a few high-value outliers.

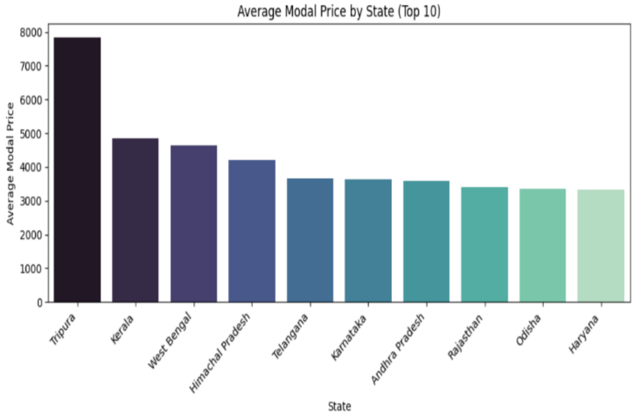
Commodity-level analysis revealed the top 10 most traded commodities and their corresponding price behavior, emphasizing the role of frequently traded items in overall market dynamics. State-wise bar charts showed which regions exhibit higher modal prices, offering insights into regional agricultural economics. Grade-wise pie charts provided an understanding of quality distribution across the dataset.

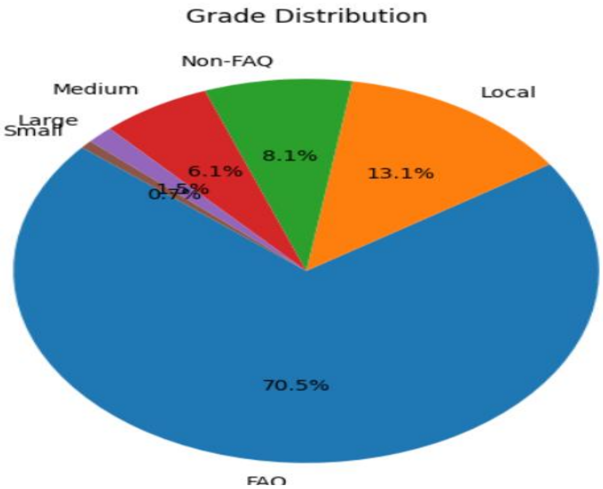
Overall, the results demonstrate that the dataset captures realistic market patterns, with pricing influenced by commodity type, regional factors, and grade classifications. The visual and statistical findings collectively confirm the reliability of the dataset and support meaningful interpretation for agricultural market trend analysis.

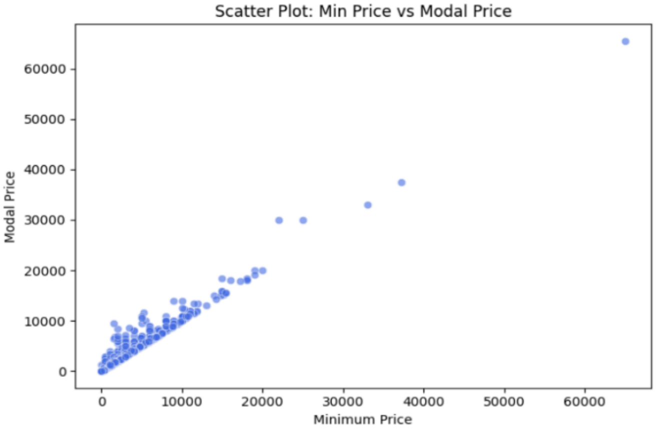


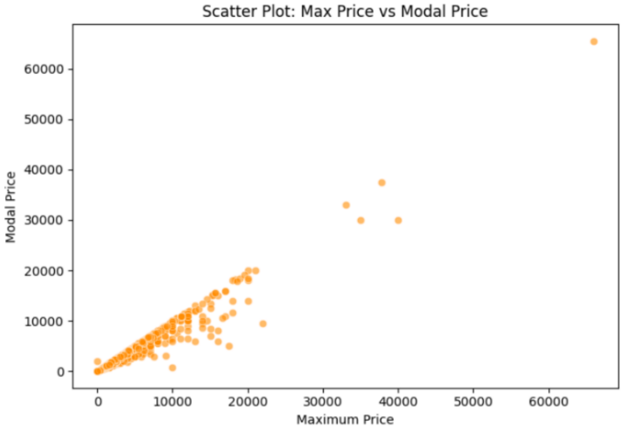


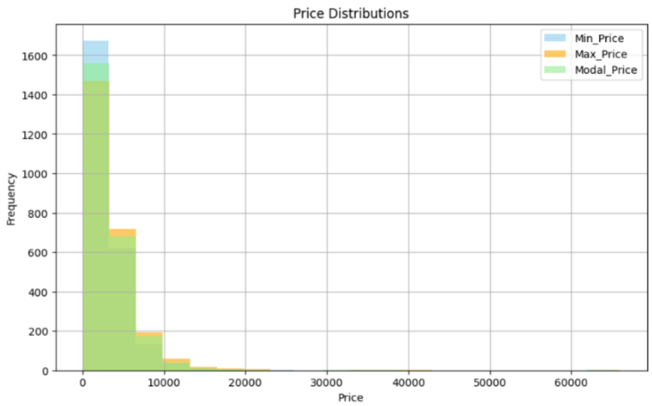
```
sns.barplot(data=avg_price_state, x='State', y='avg(Modal_Price)', palette='mako')
```











Conclusion

This study successfully analyzed agricultural market trends using PySpark-based data processing and a variety of visualization techniques. Through systematic data cleaning, transformation, and exploratory analysis, meaningful insights were derived regarding commodity pricing, regional market behavior, and grading patterns. The results highlighted strong correlations among price variables, significant regional disparities in modal prices, and clear dominance of certain commodities and grades in the dataset. Visualizations reinforced these findings, illustrating pricing distributions, market frequency, and inter-variable relationships. Overall, the project demonstrates that big-data tools like PySpark, combined with statistical and graphical analysis, provide an efficient and scalable approach for understanding agricultural markets. These insights can support policymakers, traders, and farmers by offering evidence-based perspectives on price trends, thereby enabling more informed decision-making and improved market transparency.

Future Enhancements:

□ Apache Spark Documentation. *PySpark SQL, DataFrame, and Dataset Guide*.
□ Dean, J., & Ghemawat, S. *MapReduce: Simplified Data Processing on Large Clusters*.
□ Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*.
□ Aggarwal, C. C. *Data Mining: The Textbook*.
□ Seabold, S., & Perktold, J. *Statsmodels: Econometric and Statistical Modeling with Python*.
□ Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*.
□ Bramer, M. *Principles of Data Mining*.
□ Sarker, I. H. *Machine Learning: Algorithms, Real-World Applications, and Research Directions*.
□ Rao, K. S., & Reddy, K. R. *Agricultural Market Price Forecasting and Trend Analysis: A Data-Driven Approach*.
□ Indian Ministry of Agriculture. *Market Price and Commodity Trend Reports*.
□ Tukey, J. W. *Exploratory Data Analysis*.
□ Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*.
□ Jain, A., & Singh, S. *Big Data Analytics in Agriculture: Applications and Challenges*.
□ National Agricultural Market (e-NAM). *Commodity Arrival and Price Statistics*.
□ Sharma, R. & Gupta, P. *Visualization Techniques for Large-Scale Market Datasets*.

References:

1. Apache Spark Documentation. *PySpark SQL, DataFrame, and Dataset Guide*.
2. Dean, J., & Ghemawat, S. *MapReduce: Simplified Data Processing on Large Clusters*.
3. Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*.
4. Aggarwal, C. C. *Data Mining: The Textbook*.
5. Seabold, S., & Perktold, J. *Statsmodels: Econometric and Statistical Modeling with Python*.
6. Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*.
7. Bramer, M. *Principles of Data Mining*.
8. Sarker, I. H. *Machine Learning: Algorithms, Real-World Applications, and Research Directions*.
9. Rao, K. S., & Reddy, K. R. *Agricultural Market Price Forecasting and Trend Analysis: A Data-Driven Approach*.
10. Indian Ministry of Agriculture. *Market Price and Commodity Trend Reports*.
11. Tukey, J. W. *Exploratory Data Analysis*.
12. Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*.
13. Jain, A., & Singh, S. *Big Data Analytics in Agriculture: Applications and Challenges*.
14. National Agricultural Market (e-NAM). *Commodity Arrival and Price Statistics*.
15. Sharma, R. & Gupta, P. *Visualization Techniques for Large-Scale Market Datasets*.