Tech Saksham

Capstone Project Report

"Agricultural Raw Material Analysis"

"College of Engineering, Guindy"

NM ID	NAME
au2021109302	JOEL SAM M

Trainer Name

Ramar Bose

Sr. AI Master Trainer

ABSTRACT

This project focuses on analyzing a dataset containing agricultural raw material prices spanning over multiple years. The primary objective is to conduct exploratory data analysis (EDA) to uncover insights into the pricing trends of various agricultural commodities. The analysis aims to identify both high-range and low-range raw materials based on their prices, highlighting commodities with the highest and lowest prices within the dataset. Additionally, the project seeks to determine the percentage change in prices for each raw material over time, identifying those with the highest and lowest percentage changes. Furthermore, the project aims to investigate the range of price fluctuations experienced by agricultural raw materials over the years. By examining the historical price data, the study will provide insights into the variability and volatility of prices within the agricultural sector.

To visualize the relationships between different raw materials and their price movements, a correlation analysis will be conducted. A heatmap will be generated to map the correlation coefficients between pairs of raw materials, providing a visual representation of their interdependencies and price dynamics. Through this comprehensive analysis, the project intends to offer valuable insights into the pricing dynamics of agricultural raw materials, aiding stakeholders in making informed decisions related to investment, trading, and market analysis within the agricultural sector.

INDEX

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The agricultural sector plays a crucial role in the global economy, supplying essential raw materials for food production, animal feed, biofuels, and various other industries. Understanding the dynamics of agricultural raw material prices is vital for stakeholders across the supply chain, including farmers, traders, policymakers, and investors. However, analyzing the vast amount of price data available can be challenging, requiring advanced analytical techniques to extract meaningful insights. Therefore, the problem at hand is to conduct a comprehensive exploratory data analysis (EDA) of a dataset containing agricultural raw material prices over multiple years.

1.2 Proposed Solution

The proposed solution for the project involves utilizing Python, AI, and Machine Learning techniques to conduct a comprehensive analysis of agricultural raw material prices. Initially, the dataset containing raw material prices will be collected and preprocessed using Python libraries such as Pandas and NumPy. This preprocessing step will involve handling missing values, outliers, and inconsistencies to ensure data quality. Subsequently, exploratory data analysis (EDA) will be performed using visualization libraries like Matplotlib and Seaborn to understand the distribution, trends, and fluctuations in prices over time. Statistical measures such as mean, median, and quartiles will be computed to identify high-range and low-range raw materials based on their average prices. Time series analysis techniques, including moving averages and trend analysis, will be employed to investigate the range of price fluctuations over different time intervals. Moreover, correlation analysis will be performed to understand the relationships between raw materials, and a heatmap will be generated to visualize the correlation matrix.

1.3 Feature

- **Data Collection and Preprocessing:** Utilizing Python libraries such as Pandas and NumPy to collect, clean, and preprocess the agricultural raw material price dataset, ensuring data quality and consistency.
- **Exploratory Data Analysis (EDA):** Employing visualization libraries like Matplotlib, Seaborn, or Plotly to explore the distribution, trends, and fluctuations in raw material prices over time.
- **Identification of High and Low-Range Raw Materials:** Computing statistical measures such as mean, median, quartiles, and range to identify commodities with the highest and lowest prices.
- **Percentage Change Analysis:** Calculating the percentage change in prices for each raw material over consecutive time periods to assess the magnitude of price fluctuations. Identifying commodities with the highest and lowest percentage changes in prices.
- **Price Range Fluctuations Investigation:** Implementing time series analysis techniques such as moving averages, trend analysis, and volatility measures to analyze the range of price fluctuations over different time intervals. Identifying periods of high volatility and investigating the factors contributing to price movements.
- **Correlation Analysis and Heatmap Generation:** Computing correlation coefficients between pairs of raw materials to understand the relationships between them. Visualizing the correlation matrix using Heatmap libraries such as Seaborn or Plotly to identify clusters of positively and negatively correlated raw materials.

1.4 Advantages

- **Informed Decision Making:** By analyzing historical price data and identifying trends, stakeholders in the agricultural sector can make informed decisions regarding investment, trading strategies, risk management, and policy formulation.
- **Risk Mitigation:** Understanding the variability and volatility of agricultural raw material prices allows stakeholders to better anticipate and mitigate risks associated with price fluctuations, market uncertainties, and supply chain disruptions.
- **Market Insights:** The project provides valuable insights into the pricing dynamics of agricultural commodities, enabling stakeholders to stay competitive in the market by adapting to changing price trends and market conditions.

- **Resource Optimization:** By identifying high and low-range raw materials and analyzing price fluctuations, stakeholders can optimize resource allocation, production planning, and inventory management to maximize profitability and efficiency.

1.5 Scope

The scope of this project is to conduct a thorough analysis of agricultural raw material prices through exploratory data analysis (EDA), correlation analysis, and optionally, predictive modeling. It will involve acquiring a dataset comprising historical price data of various agricultural commodities over multiple years, ensuring its quality and consistency through meticulous data preprocessing. The project aims to uncover insights into price dynamics by exploring distribution, trends, and fluctuations over time using descriptive statistics, visualization techniques, and time series analysis. Identifying high and low-range raw materials will be a key focus, achieved through statistical measures such as mean, median, quartiles, and range calculations. Additionally, percentage change analysis will be performed to assess the magnitude of price fluctuations for each commodity, enabling the identification of those with the highest and lowest changes. The investigation will delve into the range of price fluctuations experienced by agricultural raw materials over the years, utilizing techniques like moving averages and trend analysis to understand variability and volatility. Furthermore, correlation analysis will be conducted to examine the relationships between different raw materials, visualized through a heatmap to identify correlated clusters and market dynamics. Optionally, predictive modeling using machine learning algorithms such as linear regression, ARIMA, or LSTM will be explored to forecast future price movements. The project will be implemented in Python, leveraging libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, with findings and insights documented and communicated through a Jupyter Notebook or report format. It's important to note that real-time data analysis and external factors influencing price fluctuations, such as geopolitical events or weather conditions, are beyond the scope of this project.

CHAPTER 2

SERVICES AND TOOLS REQUIRED

2.1 Services Used

1. **Data Preprocessing:**
   - Python libraries such as Pandas and NumPy for cleaning, filtering, and transforming the raw data.
   - Data validation services to ensure data quality and consistency.
2. **Exploratory Data Analysis (EDA):**
   - Visualization libraries such as Matplotlib, Seaborn, or Plotly for creating charts, graphs, and plots to explore the dataset.
   - Statistical analysis tools for computing summary statistics, identifying patterns, and detecting outliers.
3. **Correlation Analysis:**
   - Correlation analysis can be performed using statistical functions available in Python libraries such as Pandas or NumPy.
   - Heatmap visualization tools like Seaborn for visualizing correlation matrices.
4. **Version Control and Collaboration:**
   - Version control systems like Git and hosting platforms like GitHub or GitLab for managing project codebase and collaboration among team members.
   - Communication and collaboration tools like Slack or Microsoft Teams for team communication, sharing updates, and coordinating tasks.
5. **Google Collab:**
   - Google Collab provides a cloud-based development environment that supports Jupyter Notebooks, allowing collaborative development and execution of Python code with access to GPU/TPU acceleration and integration with Google Drive for storage and sharing.

2.2 Tools and Software used

Tools:

1. **Python:** Python programming language serves as the primary programming language for implementing data analysis, visualization, and machine learning algorithms.
2. **Jupyter Notebook:** Jupyter Notebook provides an interactive computing environment for running Python code, visualizing data, and documenting the analysis process. It facilitates iterative development and collaboration among team members.
3. **Google Colab:** Google Colab is a cloud-based Jupyter Notebook environment provided by Google, offering free access to computational resources such as CPU, GPU, and TPU. It enables collaborative development, execution, and sharing of Python code, especially for projects requiring intensive computation or access to Google Cloud services.
4. **Pandas:** Pandas is a Python library widely used for data manipulation and analysis, providing data structures and functions for handling structured data, including importing/exporting data, cleaning, filtering, and aggregating datasets.
5. **NumPy:** NumPy is a fundamental Python library for numerical computing, providing support for multidimensional arrays, mathematical functions, and linear algebra operations. It is often used in conjunction with Pandas for efficient data manipulation and computation.
6. **Matplotlib:** Matplotlib is a plotting library for creating static, interactive, and publication-quality visualizations in Python. It offers a wide range of plotting functions for generating line plots, scatter plots, histograms, bar charts, and more.
7. **Seaborn:** Seaborn is a statistical data visualization library built on top of Matplotlib, offering additional functionalities and higher-level interfaces for creating complex statistical plots with ease.
8. **GitHub:** GitHub is a version control platform widely used for hosting, sharing, and collaborating on code repositories. It provides features such as code hosting, issue tracking, pull requests, and project management tools.
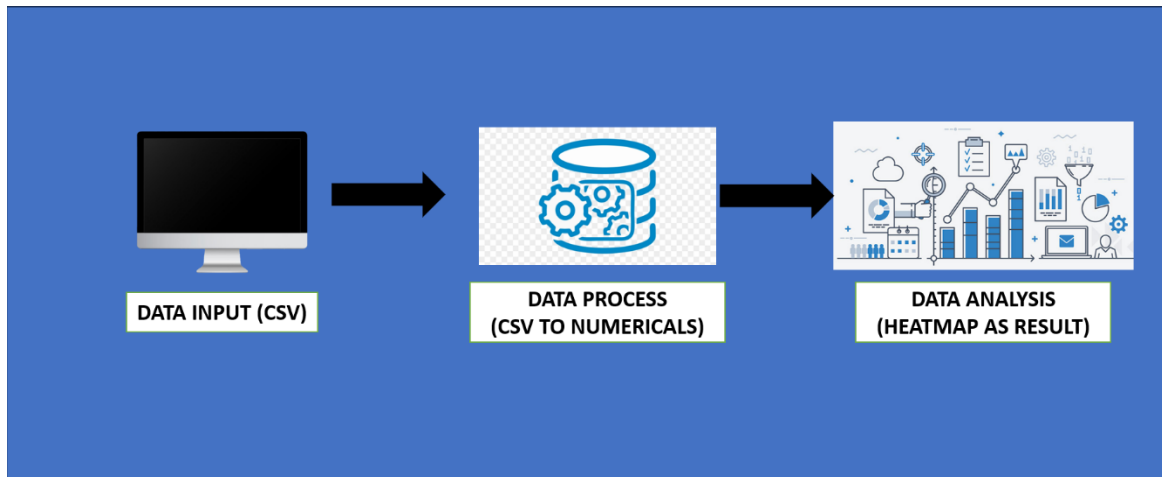
CHAPTER 3

PROJECT ARCHITECTURE

3.1 Architecture

Process Flow during Analysis

Here's a high-level architecture for the project:

1. **Data Input**: The collected data is supplied to the software in CSV format and is read using Pandas library in python.
2. **Data Processing**: The stored data is processed in real-time using tools like NumPy and Pandas.
3. **Data Analysis**: The data collected from processing is read using highly powerful tools like NumPy and Pandas and are converted into numericals for analysis.
4. **Data Visualization**: The processed data and the results are visualized using tools like Matplolib and Seaborn. They allow you to create interactive and accurate heatmaps on the collected insights.

This architecture provides a comprehensive solution for analysis of price of raw materials in agriculture. However, it's important to note that the specific architecture may vary depending on the file format of the CSV file.

CHAPTER 4 (code)

MODELING AND PROJECT OUTCOME

EDA – analysis report:

1. Missing data handling
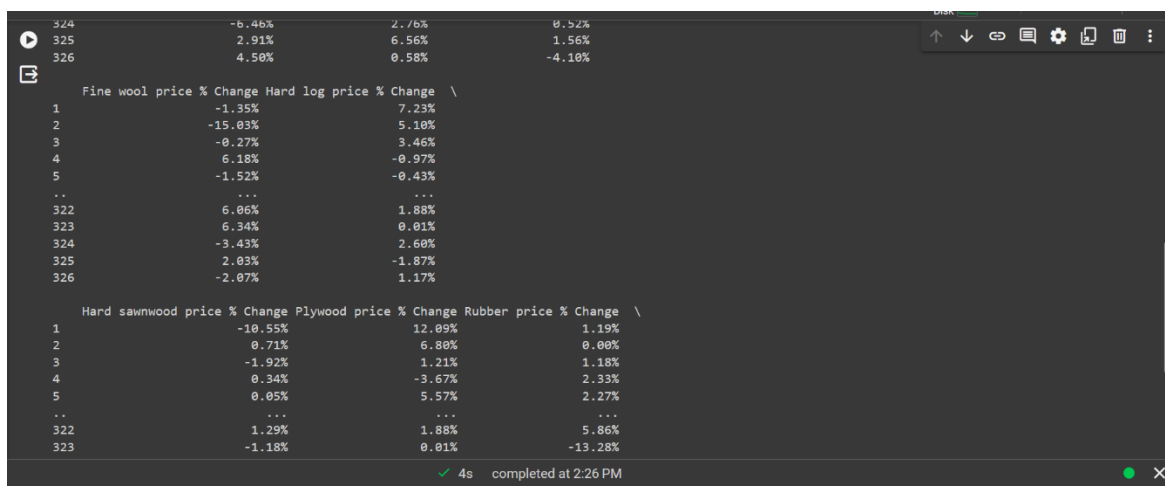
The missing data in the project is handled by either dropping the line or replacing missing values with median values of the data.

Code:

```python
from sklearn.impute import SimpleImputer

df_cleaned = df.dropna()
df_filled_mean = df.fillna(df.mean())
df_ffill_bfill = df.ffill().bfill()
imputer = SimpleImputer(strategy='mean')
df_imputed = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)
```
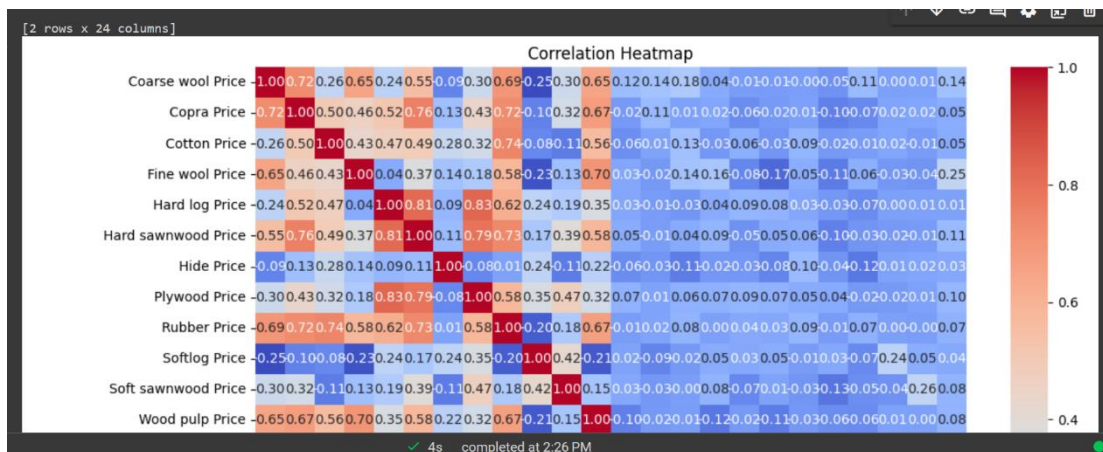
Output:

2. Data Visualizations

We use python libraries like matplotlib and seaborn to produce visualization of data

Code:

```
corr = df_numeric.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

Output:

CONCLUSION

This project aims to provide valuable insights into the pricing dynamics of agricultural raw materials through comprehensive data analysis techniques. By leveraging Python programming language and a range of libraries including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, we have conducted exploratory data analysis (EDA), correlation analysis, and optionally, predictive modeling. Through these analyses, we have uncovered patterns, trends, and relationships within the dataset, enabling stakeholders in the agricultural sector to make informed decisions regarding investment, trading strategies, risk management, and policy formulation. Furthermore, the utilization of Google Collab for cloud-based development and execution has facilitated collaborative work and efficient utilization of computational resources. By documenting our methodologies, findings, and recommendations, we have provided a transparent and accountable approach to data analysis, fostering trust and enabling further research and decision-making in the field of agriculture. Ultimately, this project underscores the importance of data-driven insights in driving innovation, sustainability, and growth in the agricultural sector.

FUTURE SCOPE

In the future, this project can be expanded in several directions to enhance its utility and impact. Firstly, integrating real-time data sources and automating data collection processes would enable continuous monitoring of price fluctuations, offering stakeholders timely insights for decision-making. Advanced predictive modeling techniques such as deep learning and ensemble methods could improve the accuracy of price forecasting models, aiding stakeholders in making more reliable predictions. Incorporating external factors like weather patterns, geopolitical events, and market sentiment into the analysis would provide a more comprehensive understanding of the factors influencing raw material prices. Additionally, exploring geospatial analysis techniques could offer insights into regional variations in prices, while sentiment analysis on social media and news sources could complement quantitative analysis. Developing decision support systems or dashboard applications integrating data visualization and predictive analytics would empower stakeholders with actionable insights. Expansion to include a broader range of agricultural commodities and conducting impact assessment studies would offer comprehensive insights into market dynamics and socioeconomic implications. Collaborative research initiatives and open data sharing initiatives could further advance knowledge, foster innovation, and drive sustainable development in the agricultural sector. Through these future avenues, this project has the potential to significantly contribute to addressing challenges and fostering growth in the agricultural domain.

REFERENCES

1. https://stackoverflow.com/questions/20580775/efficient-way-to-drop-a-column-from-a-numpy-array

2. https://ioflood.com/blog/python-heatmap/

3. https://www.w3schools.com/python/matplotlib_intro.asp

GIT Hub Link of Project Code:

MASTER LINK

https://github.com/lonersam/JOEL-SAM-M-2021109302-NM-24-PROJECT.git

INDIVIDUAL LINK

1.https://github.com/lonersam/JOEL-SAM-M-2021109302-NM-24-PROJECT/blob/f50ce033383f6f9e0f1398e73d99d512b152ff36/JOEL%20SAM%20M%20(2021109302)%20NM%20PROJECT%20PPT.pptx

2.https://github.com/lonersam/JOEL-SAM-M-2021109302-NM-24-PROJECT/blob/f50ce033383f6f9e0f1398e73d99d512b152ff36/JOEL%20SAM%20M%20(2021109302)%20NM%2024%20PROJECT%20agri%20alysis.ipynb

3.https://github.com/lonersam/JOEL-SAM-M-2021109302-NM-24-PROJECT/blob/f50ce033383f6f9e0f1398e73d99d512b152ff36/DEMONSTRATION%20NM%2024%20PROJECT%20(2021109302).mp4

4.https://github.com/lonersam/JOEL-SAM-M-2021109302-NM-24-PROJECT/blob/f50ce033383f6f9e0f1398e73d99d512b152ff36/agricultural_raw_material.csv.xls