Tech Saksham

Capstone Project Report

“Agricultural Raw Material Analysis”

“College of Engineering, Guindy”

NM ID	NAME
au2021109037	SIVA PRABHA P

Trainer Name

Ramar Bose

Sr. AI Master Trainer

ABSTRACT

The primary objective of this study is to analyze a dataset that includes pricing for agricultural raw materials across several years. The main goal is to perform exploratory data analysis (EDA) to find out more about the patterns in the prices of different agricultural commodities. The goal of the analysis is to determine which raw materials are high- and low-range in terms of price, emphasizing the commodities in the dataset with the highest and lowest prices. To help discover which raw materials have the biggest and lowest percentage changes in price over time, the project also aims to calculate the % change in price for each raw material.

Additionally, the research endeavor seeks to examine the spectrum of price variations that agricultural raw materials have encountered throughout time. The study will provide light on the fluctuations and volatility of by looking at the historical pricing data.

INDEX

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The production of food, animal feed, biofuels, and other key raw resources are all made possible by the agricultural sector, which is vital to the world economy. Stakeholders in the supply chain, such as farmers, traders, legislators, and investors, must comprehend the dynamics of agricultural raw material prices. However, it might be difficult to analyze the large amount of price data that is available; in order to derive useful insights, sophisticated analytical approaches are needed. As a result, the current challenge is to carry out a thorough exploratory data analysis (EDA) of a dataset that includes pricing for agricultural raw materials across several years.

1.2 Proposed Solution

The project's suggested approach is using machine learning, artificial intelligence, and Python to perform a thorough examination of the costs of agricultural raw materials. First, preprocessing and dataset collection for raw material pricing will be done with Python tools like NumPy and Pandas. In order to guarantee data quality, this preprocessing stage will handle missing values, outliers, and inconsistencies. To comprehend the distribution, trends, and variations in pricing over time, exploratory data analysis (EDA) will subsequently be carried out utilizing visualization packages such as Matplotlib and Seaborn. The average prices of raw materials will be used to determine whether raw materials are high- and low-range by computing statistical measures like mean, median, and quartiles. The range of price swings across various time intervals will be examined using time series analysis techniques, such as trend analysis and moving averages. In addition, a heatmap will be created to show the correlation matrix and correlation analysis will be carried out to comprehend the connections between the raw materials.

1.3Feature

• **Percentage Change Analysis:** To determine the extent of price variations, compute the percentage change in prices for each raw material across a series of time periods. determining which commodities have the biggest and lowest percentage price fluctuations.

• **Price Range variations Investigation:** To examine the range of price variations across various time periods, time series analysis techniques such moving averages, trend analysis, and volatility measures are used. recognizing times of extreme volatility and looking into the causes of price changes.

• **Correlation Analysis and Heatmap Generation:** To determine the relationships between raw material pairs, compute the correlation coefficients between them. Using heatmap libraries like Plotly or Seaborn, visualize the correlation matrix to find raw material groupings that are positively and negatively connected.

• **Data Collection and Preprocessing:** To ensure data quality and consistency, the agricultural raw material price dataset is collected, cleaned, and preprocessed using Python libraries like Pandas and NumPy.

• **Exploratory Data Analysis (EDA):** This method involves examining the distribution, trends, and variations in raw material pricing over time using visualization tools such as Matplotlib, Seaborn, or Plotly.

• **Finding High and Low-Range Raw Materials:** To find the commodities with the highest and lowest prices, statistical measures including mean, median, quartiles, and range are computed.

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 analysing price fluctuations, stakeholders can optimize resource allocation, production planning, and inventory management to maximize profitability and efficiency.

1.5 Scope

This research aims to provide a comprehensive examination of agricultural raw material pricing by means of correlation analysis, exploratory data analysis (EDA), and maybe predictive modeling. It will need obtaining a dataset that includes historical price data for different agricultural commodities over a number of years, carefully preparing the data to ensure its consistency and quality. Through the use of time series analysis, visualization techniques, and descriptive statistics, the project seeks to shed light on pricing dynamics by examining distribution, trends, and variations across time. One of the main goals will be to determine which raw materials are high and low range. This will be done by using statistical tools like mean, median, quartiles, and range calculations.

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:

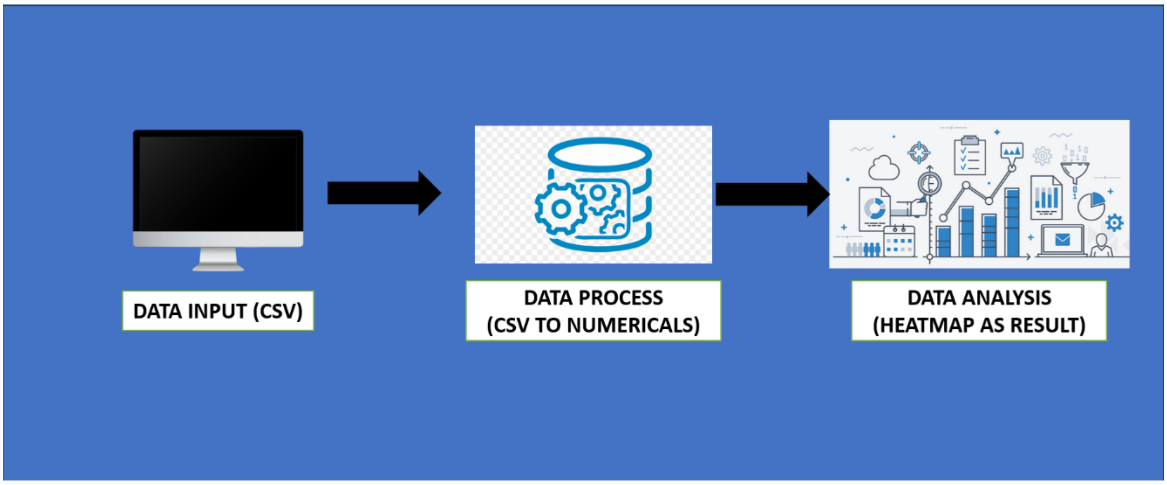
1. **Python:** Python programming language serves as the primary programming language for implementing data analysis, visualization, and machine learning algorithms.
2. **Jupyter Notebook:** Jupyter Notebook provides an interactive computing environment for running Python code, visualizing data, and 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

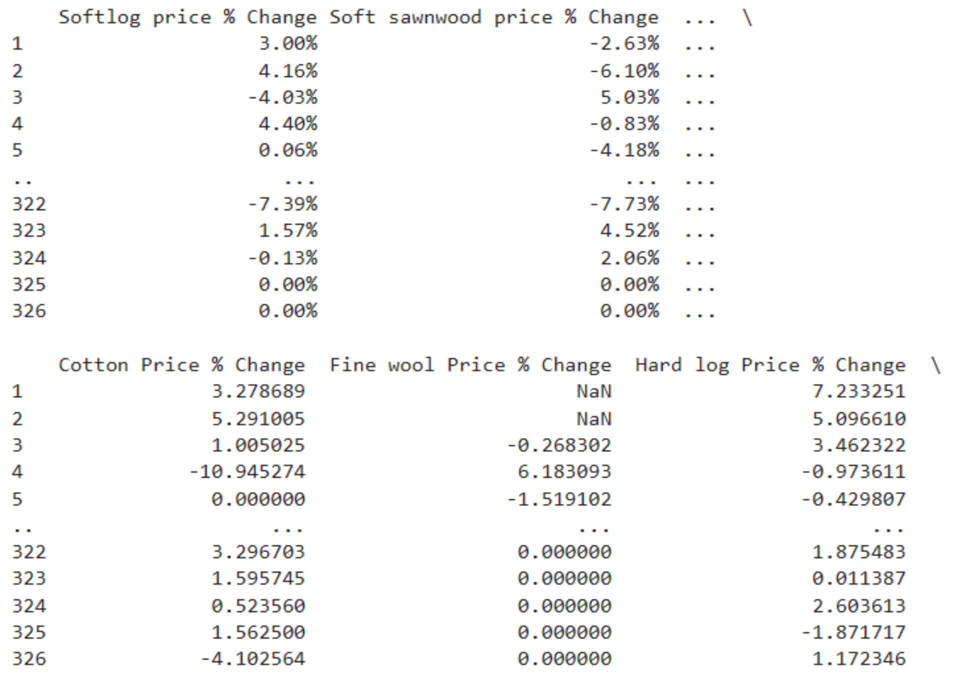
The missing data in the project is handled by either dropping the line or replacing missing values with median values of the data.

Code:

```
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:



```
    Softlog price % Change Soft sawnwood price % Change  ...  \
1                    3.00%                       -2.63%  ...
2                    4.16%                       -6.10%  ...
3                   -4.03%                        5.03%  ...
4                    4.40%                       -0.83%  ...
5                    0.06%                       -4.18%  ...
..                     ...                          ...  ...
322                 -7.39%                       -7.73%  ...
323                  1.57%                        4.52%  ...
324                 -0.13%                        2.06%  ...
325                  0.00%                        0.00%  ...
326                  0.00%                        0.00%  ...

    Cotton Price % Change  Fine wool Price % Change  Hard log Price % Change  \
1                3.278689                       NaN                 7.233251
2                5.291005                       NaN                 5.096610
3                1.005025                 -0.268302                 3.462322
4              -10.945274                  6.183093                -0.973611
5                0.000000                 -1.519102                -0.429807
..                    ...                       ...                      ...
322              3.296703                  0.000000                 1.875483
323              1.595745                  0.000000                 0.011387
324              0.523560                  0.000000                 2.603613
325              1.562500                  0.000000                -1.871717
326             -4.102564                  0.000000                 1.172346
```

2. Data Visualizations

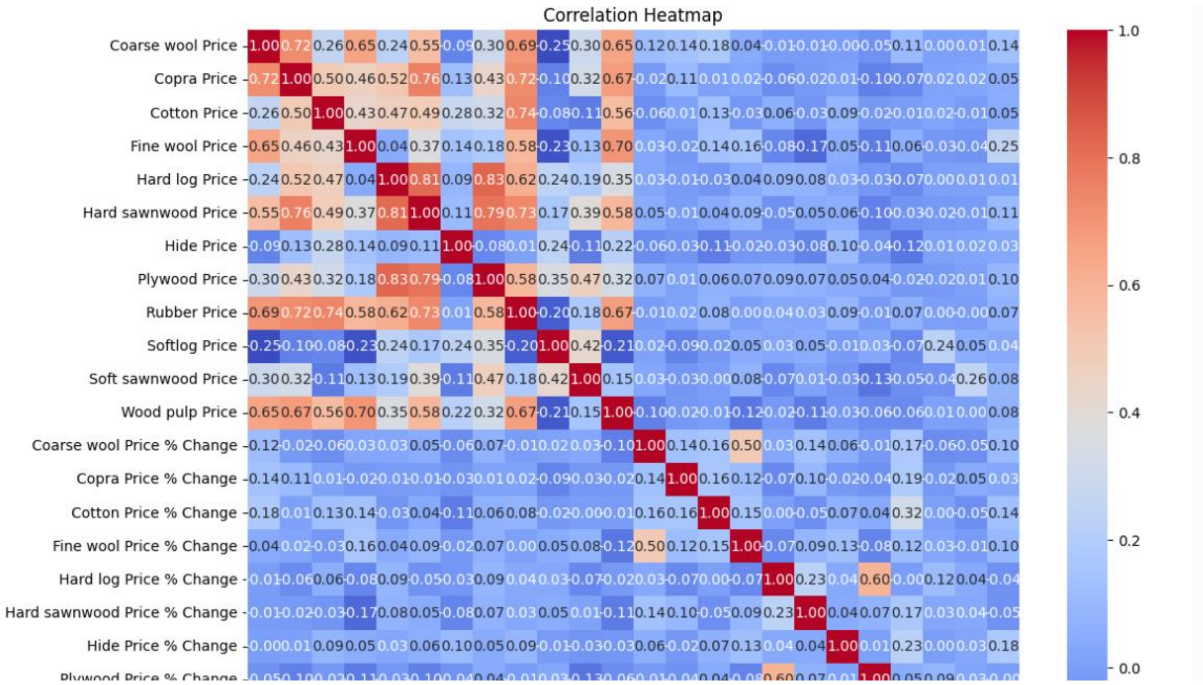
We use python libraries like matplotlib and seaborn to produce visualization of data

Code:

```
numeric_columns = df.select_dtypes(include=[np.number]).columns
df_numeric = df[numeric_columns]

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 uses extensive data analysis tools to try and offer insightful information about the dynamics of agricultural raw material pricing. Using the Python programming language and several libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, we have performed correlation analysis, exploratory data analysis (EDA), and predictive modeling as necessary. With the use of these studies, we have been able to identify patterns, trends, and linkages in the dataset that will help agricultural sector players make well-informed decisions about trading strategies, investment, risk management, and policy formation. In addition, the efficient use of computational resources and collaborative work have been made possible by the cloud-based creation and execution of Google Collab. 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

This project can be developed further in the future to increase its usefulness and impact. Firstly, automating data gathering procedures and integrating real-time data sources would allow for ongoing price fluctuation monitoring, providing stakeholders with fast decision-making insights. Deep learning and ensemble approaches are examples of advanced predictive modeling techniques that could increase the accuracy of price forecasting models and help stakeholders make more accurate predictions. For a more thorough knowledge of the factors impacting raw material prices, the analysis should take into account external factors such as weather patterns, geopolitical events, and market mood. Furthermore, investigating geospatial analytic methods may provide light on regional differences in pricing, and sentiment research on news sources and social media may enhance quantitative analysis. Creating dashboard apps or decision support systems that combine predictive analytics and data visualization will provide stakeholders with useful information. Including a wider variety of agricultural goods and carrying out effect assessment research would provide thorough understandings of market dynamics and social ramifications. In the agricultural industry, collaborative research projects and open data sharing programs have the potential to increase understanding, stimulate creativity, and promote sustainable development. This initiative has the potential to make a substantial contribution toward addressing issues and promoting development in the agriculture sector through these future paths.

REFERENCES

1. https://www.w3schools.com/python/matplotlib_intro.asp
2. https://stackoverflow.com/questions/10996140/how-to-remove-specific-elements-in-a-numpy-array?noredirect=1&lq=1
3. seaborn.heatmap — seaborn 0.13.2 documentation (pydata.org)
4. https://www.askpython.com/python/examples/heatmaps-in-python

GIT Hub Link of Project Code:

https://github.com/sivaprabha24/sivaprabha