Tech Saksham

Capstone Project Report

“Agricultural Raw Material Analysis”

“College of Engineering, Guindy”

NM ID	NAME
au2021111044	MOHAMED WAFEEQ

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.

Through correlation analysis, an investigation into the relationships between different raw materials and their price changes will begin. To graphically represent the relationships and pricing behaviors between pairs of raw materials, a heatmap displaying the correlation coefficients between them will be created. Through this in-depth analysis, the project hopes to provide insightful knowledge on the pricing dynamics of agricultural raw materials, enabling stakeholders to make educated decisions about trade, investment, and market research in 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 approach for the project entails the utilization of Python, AI, and Machine Learning methodologies to conduct an exhaustive examination of agricultural raw material prices. Initially, data comprising raw material prices will be gathered and processed using Python libraries such as Pandas and NumPy. This preprocessing stage will involve addressing missing values, outliers, and discrepancies to uphold data integrity. Following this, exploratory data analysis (EDA) will be undertaken utilizing visualization tools like Matplotlib and Seaborn to grasp the distribution, patterns, and fluctuations in prices over time. Statistical metrics such as mean, median, and quartiles will be computed to discern raw materials with high and low average prices. Techniques in time series analysis, including moving averages and trend analysis, will be applied to explore the extent of price fluctuations across various time spans. Additionally, correlation analysis will be conducted to elucidate the connections between raw materials, with a resultant heatmap serving to visually represent the correlation matrix.

1.3Feature

- **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

ľools:

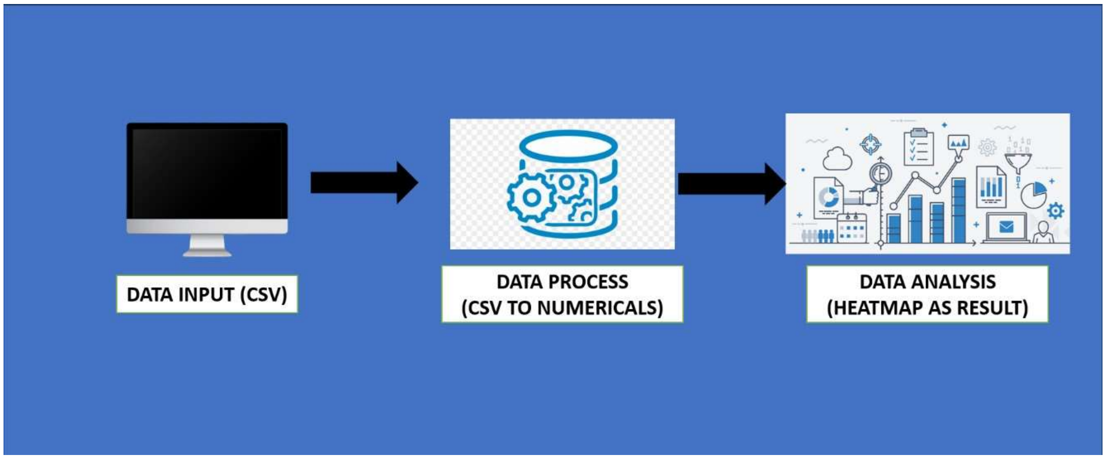
1. **Python:** Serving as the principal programming language, Python is utilized for executing data analysis, visualization, and machine learning algorithms.
2. **Jupyter Notebook:** Offering an interactive computing environment, Jupyter Notebook supports the execution of Python code, data visualization, and documentation of the analytical process. It fosters iterative development and facilitates collaboration among team members.
3. **Google Colab:** Provided by Google, Google Colab is a cloud-based Jupyter Notebook environment that grants free access to computational resources like CPU, GPU, and TPU. It enables collaborative coding, execution, and sharing of Python scripts, particularly beneficial for projects necessitating extensive computation or access to Google Cloud services.
4. **Pandas:** Widely employed for data manipulation and analysis, Pandas is a Python library that furnishes data structures and functions for managing structured data. Its capabilities encompass importing/exporting data, cleaning, filtering, and aggregating datasets.
5. **NumPy**: Fundamental to numerical computing in Python, NumPy supports multidimensional arrays, mathematical functions, and linear algebra operations. Often used alongside Pandas, it enhances efficient data manipulation and computation.
6. **Matplotlib**: Functioning as a plotting library, Matplotlib facilitates the creation of static, interactive, and publication-quality visualizations in Python. Offering a diverse array of plotting functions, it caters to generating line plots, scatter plots, histograms, bar charts, and more.
7. **Seaborn:** Built on top of Matplotlib, Seaborn is a statistical data visualization library that provides additional functionalities and higher-level interfaces for effortlessly crafting complex statistical plots.
8. **GitHub:** A widely utilized version control platform, GitHub serves as a hub for hosting, sharing, and collaborating on code repositories. It offers features such as code hosting, issue tracking, pull requests, and project management tools.

CHAPTER 3

PROJECT ARCHITECTURE

3.1 Architecture

Process Flow during Analysis



DATA INPUT (CSV)

DATA PROCESS
(CSV TO NUMERICALS)

DATA ANALYSIS
(HEATMAP AS RESULT)

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

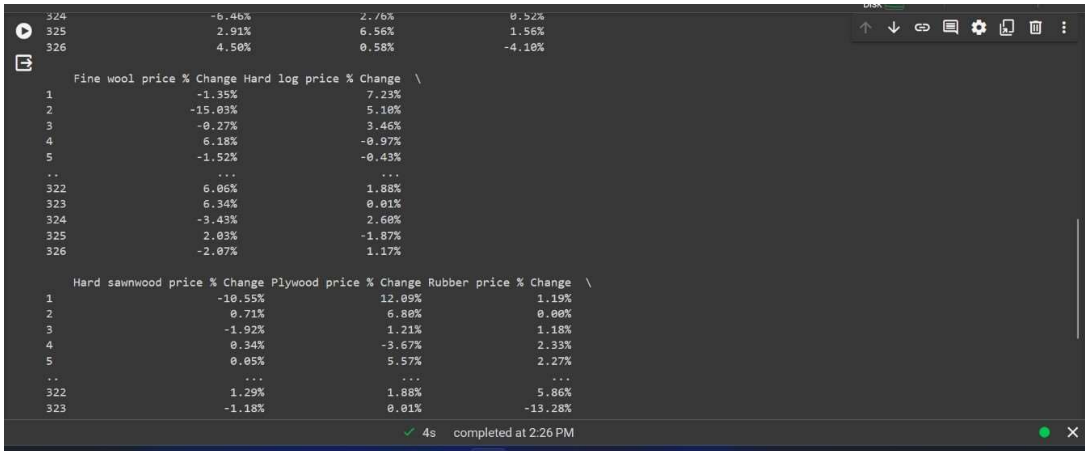
The missing data in the project is handled by either dropping the line or replacing missing values with median values of the data.

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:



```
324                  -6.46%              2.76%               0.52%
325                   2.91%              6.56%               1.56%
326                   4.50%              0.58%              -4.10%

    Fine wool price % Change Hard log price % Change  \
1                     -1.35%                   7.23%
2                    -15.03%                   5.10%
3                     -0.27%                   3.46%
4                      6.18%                  -0.97%
5                     -1.52%                  -0.43%
..                       ...                     ...
322                    6.06%                   1.88%
323                    6.34%                   0.01%
324                   -3.43%                   2.60%
325                    2.03%                  -1.87%
326                   -2.07%                   1.17%

    Hard sawnwood price % Change Plywood price % Change Rubber price % Change  \
1                        -10.55%                 12.09%                 1.19%
2                          0.71%                  6.80%                 0.00%
3                         -1.92%                  1.21%                 1.18%
4                          0.34%                 -3.67%                 2.33%
5                          0.05%                  5.57%                 2.27%
..                           ...                    ...                   ...
322                        1.29%                  1.88%                 5.86%
323                       -1.18%                  0.01%               -13.28%
```

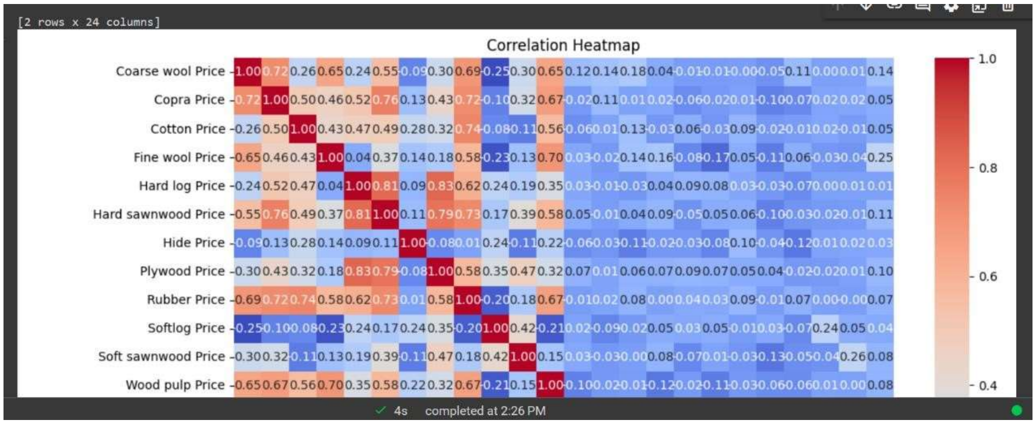
2. Data Visualizations

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:



```
[2 rows x 24 columns]
```

Correlation Heatmap

Coarse wool Price	1.00	0.72	0.26	0.65	0.24	0.55	-0.09	0.30	0.69	-0.25	0.30	0.65	0.12	0.14	0.18	0.04	-0.01	-0.01	-0.00	0.05	0.11	0.00	0.01	0.14
Copra Price	0.72	1.00	0.50	0.46	0.52	0.76	0.13	0.43	0.72	-0.10	0.32	0.67	-0.02	0.11	0.01	0.02	-0.06	0.02	0.01	-0.10	0.07	0.02	0.02	0.05
Cotton Price	0.26	0.50	1.00	0.43	0.47	0.49	0.28	0.32	0.74	-0.08	0.11	0.56	-0.06	0.01	0.13	-0.03	0.06	-0.03	0.09	-0.02	0.01	0.02	-0.01	0.05
Fine wool Price	0.65	0.46	0.43	1.00	0.04	0.37	0.14	0.18	0.58	-0.23	0.13	0.70	0.03	-0.02	0.14	0.16	-0.08	-0.17	0.05	-0.11	0.06	-0.03	-0.04	0.25
Hard log Price	0.24	0.52	0.47	0.04	1.00	0.81	0.09	0.83	0.62	0.24	0.19	0.35	0.03	-0.01	-0.03	0.04	0.09	0.08	0.03	-0.03	0.07	0.00	0.01	0.01
Hard sawnwood Price	0.55	0.76	0.49	0.37	0.81	1.00	0.11	0.79	0.73	0.17	0.39	0.58	0.05	-0.01	0.04	0.09	-0.05	0.05	0.06	-0.10	0.03	0.02	-0.01	0.11
Hide Price	-0.09	0.13	0.28	0.14	0.09	0.11	1.00	-0.08	0.01	0.24	-0.11	0.22	-0.06	0.03	-0.11	-0.02	-0.03	0.08	0.10	-0.04	-0.12	0.01	0.02	0.03
Plywood Price	0.30	0.43	0.32	0.18	0.83	0.79	-0.08	1.00	0.58	0.35	0.47	0.32	0.07	0.01	0.06	0.07	0.09	0.07	0.05	0.04	-0.02	-0.02	0.01	0.10
Rubber Price	0.69	0.72	0.74	0.58	0.62	0.73	0.01	0.58	1.00	-0.20	0.18	0.67	-0.01	0.02	0.08	0.00	0.04	0.03	0.09	-0.01	0.07	0.00	-0.00	0.07
Softlog Price	-0.25	-0.10	-0.08	-0.23	0.24	0.17	0.24	0.35	-0.20	1.00	0.42	-0.21	0.02	-0.09	0.02	0.05	0.03	0.05	-0.01	0.03	-0.07	0.24	0.05	0.04
Soft sawnwood Price	0.30	0.32	-0.11	0.13	0.19	0.39	-0.11	0.47	0.18	0.42	1.00	0.15	0.03	-0.03	0.00	0.08	-0.07	0.01	-0.03	0.13	0.05	0.04	0.26	0.08
Wood pulp Price	0.65	0.67	0.56	0.70	0.35	0.58	0.22	0.32	0.67	-0.21	0.15	1.00	-0.10	0.02	-0.01	-0.12	-0.02	-0.11	-0.03	0.06	0.06	0.01	0.00	0.08

✓ 4s completed at 2:26 PM

CONCLUSION

This project endeavors to offer insightful understandings into the dynamic pricing behaviors of agricultural raw materials via thorough data analysis methodologies. Employing the Python programming language and an array of libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, we have engaged in exploratory data analysis (EDA), correlation analysis, and potentially predictive modeling. Through these analytical endeavors, we have unveiled underlying patterns, trends, and interconnections within the dataset, empowering stakeholders within the agricultural domain to make well-grounded decisions concerning investment, trading strategies, risk mitigation, and policy development. Moreover, the integration of Google Colab for cloud-based development and execution has streamlined collaborative efforts and optimized the utilization of computational resources. By meticulously documenting our methodologies, discoveries, and recommendations, we have adopted a transparent and accountable approach to data analysis, nurturing trust and facilitating further exploration and decision-making within the agricultural sphere. Ultimately, this project underscores the pivotal role of data-driven insights in propelling innovation, sustainability, and advancement within 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:

https://github.com/mohamed-wafeeq/NM_AGRI_PROJ