

**Tech Saksham**

**Capstone Project Report**

**“Agricultural Raw Material Analysis”**

**“Madras Institute of Technology, Anna University”**

|  |  |
| --- | --- |
| **NM ID** | **NAME** |
| au2021507040 | SELVALAKSHMI M |

|  |  |
| --- | --- |
|  |  |
|  | Ramar Bose |
|  | Sr. AI Master Trainer |

**ABSTRACT**

In the agriculture sector, where countless decisions must be made daily amidst complex factors, accurate yield estimation is crucial for effective planning. Data mining techniques offer a practical and efficient means to address this challenge. Agriculture, being a prime candidate for big data applications, benefits from the analysis of various factors such as environmental conditions, soil variability, input levels, and commodity prices.

This paper aims to leverage data mining techniques, including PAM, CLARA, DBSCAN, and Multiple Linear Regression, to analyze agricultural data and identify optimal parameters for maximizing crop production. By mining extensive datasets encompassing crop, soil, and climatic information, as well as analyzing new, non-experimental data, the agricultural sector can enhance production and bolster resilience against climate change.

PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications), and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are clustering algorithms that can identify patterns and groupings within agricultural data, aiding in understanding factors influencing crop yields. Multiple Linear Regression, on the other hand, provides a statistical method to model the relationship between various input variables and crop production, enabling predictive analysis and optimization.

By harnessing these data mining techniques, farmers and agribusinesses can make informed decisions, optimize resource allocation, and adapt to changing environmental conditions, ultimately enhancing productivity and sustainability in agriculture.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Table of Contents** | **Page No.** |
| 1 | Chapter 1: Introduction | 4 |
| 2 | Chapter 2: Services and Tools Required | 8 |
| 3 | Chapter 3: Project Architecture | 11 |
| 4 | Chapter 4: Modeling and Project Outcome | 15 |
| 5 | Conclusion | 22 |
| 6 | Future Scope | 23 |
| 7 | References | 24 |
| 8 | Links | 25 |

**25 pages**

**CHAPTER 1**

**INTRODUCTION**

* 1. **Problem Statement**

In the field of agricultural economics, the volatility of raw material prices presents a significant obstacle for stakeholders, complicating decision-making and resource management. Historical data alone often fails to provide reliable forecasts, intensifying market uncertainty and risk. To address this challenge, there's a critical demand for robust predictive models that harness machine learning algorithms to scrutinize historical price trends and forecast future values accurately. Developing such models would equip stakeholders with invaluable insights to navigate market fluctuations, fine-tune pricing strategies, and promote sustainability within the agricultural sector.

* 1. **Proposed Solution**

The proposed solution revolves around the creation of predictive models tailored to forecast agricultural raw material prices, employing machine learning algorithms. These models will capitalize on historical data and sophisticated analytical techniques like Linear Regression and Random Forest Regression to achieve precise predictions of future prices. By undergoing thorough data preprocessing, meticulous model selection, and rigorous evaluation processes, stakeholders within the agricultural sector stand to benefit from invaluable insights into market dynamics. Armed with this knowledge, they can optimize resource allocation strategies and make informed decisions to mitigate risks linked with price fluctuations.

* 1. **Feature**
* **Preprocessing Analysis**:

The code preprocesses the dataset by addressing missing values, encoding categorical variables, and dividing the data into training and testing sets. This ensures that the dataset is properly prepared for training machine learning models.

* **Customer Segmentation**:

To address the volatility of raw material prices in agricultural economics, stakeholders require advanced predictive models leveraging machine learning algorithms. These models scrutinize historical price trends to forecast future values accurately, offering invaluable insights to navigate market fluctuations and fine-tune pricing strategies. Additionally, such models facilitate customer segmentation, enabling stakeholders to tailor their approaches based on distinct customer behaviors and preferences. By understanding their customer base more comprehensively, stakeholders can optimize resource allocation, enhance decision-making, and promote sustainability within the agricultural sector.

**Trend Analysis**:

The ups and downs of agricultural raw material prices are a real headache for everyone involved in farming and food production. It's like playing a guessing game, making it tough to figure out how much stuff to plant or how much to charge.

The problem is, the current methods for predicting prices aren't exactly the best. They're like trying to win a prize at a carnival with a rusty dartboard - not very reliable.

Here's where some cool tech comes in: machine learning! Imagine a super-powered calculator that crunches tons of past price data to figure out what might happen next. This could be a game-changer for farmers, food companies, and anyone who relies on these raw materials.

With better predictions, everyone can make smarter decisions. Farmers can plan their crops with more confidence, businesses can set better prices, and maybe, just maybe, we can create a more stable and sustainable food system for everyone.

* 1. **Advantages**

**1.Comprehensive Data Analysis:** The code facilitates a comprehensive analysis of agricultural raw material prices by conducting exploratory data analysis (EDA), identifying high and low-range materials, analyzing percentage changes, and exploring correlations between raw materials. This holistic approach provides stakeholders with valuable insights into market dynamics and pricing trends.

**2. Model Flexibility:** The code allows for the selection and training of multiple machines learning algorithms, including Linear Regression and Random Forest Regression.

**3. Performance Evaluation:** The code includes thorough model evaluation using root mean squared error (RMSE) as the evaluation metric. By comparing the performance of different models, stakeholders can assess their accuracy in predicting raw material prices and make informed decisions about model deployment and refinement.

**4. Scalability and Adaptability:** The generated code is scalable and adaptable to accommodate future research and development efforts. Stakeholders can expand the analysis to include additional features, integrate external factors, or deploy the predictive models in real world applications. This scalability ensures that the code remains relevant and valuable in addressing evolving challenges and opportunities in agricultural economics.

* 1. **Scope**

The generated code for predictive modeling of agricultural raw material prices offers a powerful toolset for enhancing decision-making and resource allocation within the agricultural sector. Its extensive data analysis capabilities empower stakeholders to glean valuable insights into market dynamics, pricing trends, and interrelationships among various raw materials. Moreover, the flexibility to employ and fine-tune multiple machine learning algorithms enables experimentation and adaptation to diverse data characteristics and modeling requirements. Through rigorous evaluation metrics, stakeholders can confidently gauge the accuracy of predictive models, facilitating informed decisions regarding their deployment and refinement. Additionally, the code's scalability paves the way for future expansion, including the integration of additional features, external factors, and real-world applications. In sum, the code presents a robust framework for predictive modeling, poised to catalyze innovation and optimization in agricultural economics.

**CHAPTER 2**

**SERVICES AND TOOLS REQUIRED**

**2.1 Services Used**

**1.** **Pandas**: Pandas is a versatile data manipulation library in Python, essential for preprocessing, exploring, and manipulating data. Its key data structure, DataFrame, efficiently handles tabular data, making it indispensable for various tasks in data analysis.

**2. Scikit-learn (sklearn):** Scikit-learn is a comprehensive machine learning library in Python, offering a wide range of algorithms for regression, classification, clustering, and more. In the provided code, it facilitates model selection, training, and evaluation, with implementations for Linear Regression and Random Forest Regression algorithms.

**3. NumPy:** NumPy serves as a fundamental package for scientific computing in Python, empowering users with support for large, multi-dimensional arrays and matrices. Alongside, it provides a rich collection of mathematical functions tailored for operating on these arrays. Frequently paired with Pandas, NumPy enhances numerical computations within the code.

**4. Seaborn and Matplotlib:** Seaborn and Matplotlib represent indispensable visualization libraries in Python, employed for crafting static, animated, and interactive visualizations. Within the code, Seaborn aids in generating a heatmap to visualize the correlation matrix, while Matplotlib fulfills general plotting requirements.

**5. Scikit-learn (sklearn.metrics):** The `sklearn.metrics` module within Scikit-learn facilitates the computation of evaluation metrics, such as mean squared error (MSE) and root mean squared error (RMSE). These metrics serve as crucial indicators for assessing the predictive model's performance and guiding decision-making processes.

**2.2 Tools and Software used**

**Tools**:

**1. Python:** Python is a widely-used programming language for data analysis, machine learning, and scientific computing. Ensure you have Python installed on your system.

**2. Integrated Development Environment (IDE**): You can use any Python IDE or text editor of your choice for writing and executing the code. Popular options include PyCharm, Jupyter Notebook, Spyder, Visual Studio Code, and Sublime Text.

**3. Python Libraries:**

* Pandas: Install Pandas using `pip install pandas`. This library is essential for data manipulation and analysis.
* NumPy: Install NumPy using `pip install numpy`. It is a fundamental library for numerical computations.
* Matplotlib: Install Matplotlib using `pip install matplotlib`. It is a plotting library for creating static, animated, and interactive visualizations.

**4. Dataset:** You need a dataset containing historical records of agricultural raw material prices. Ensure the dataset is in a compatible format such as Excel.

**Software Requirements**:

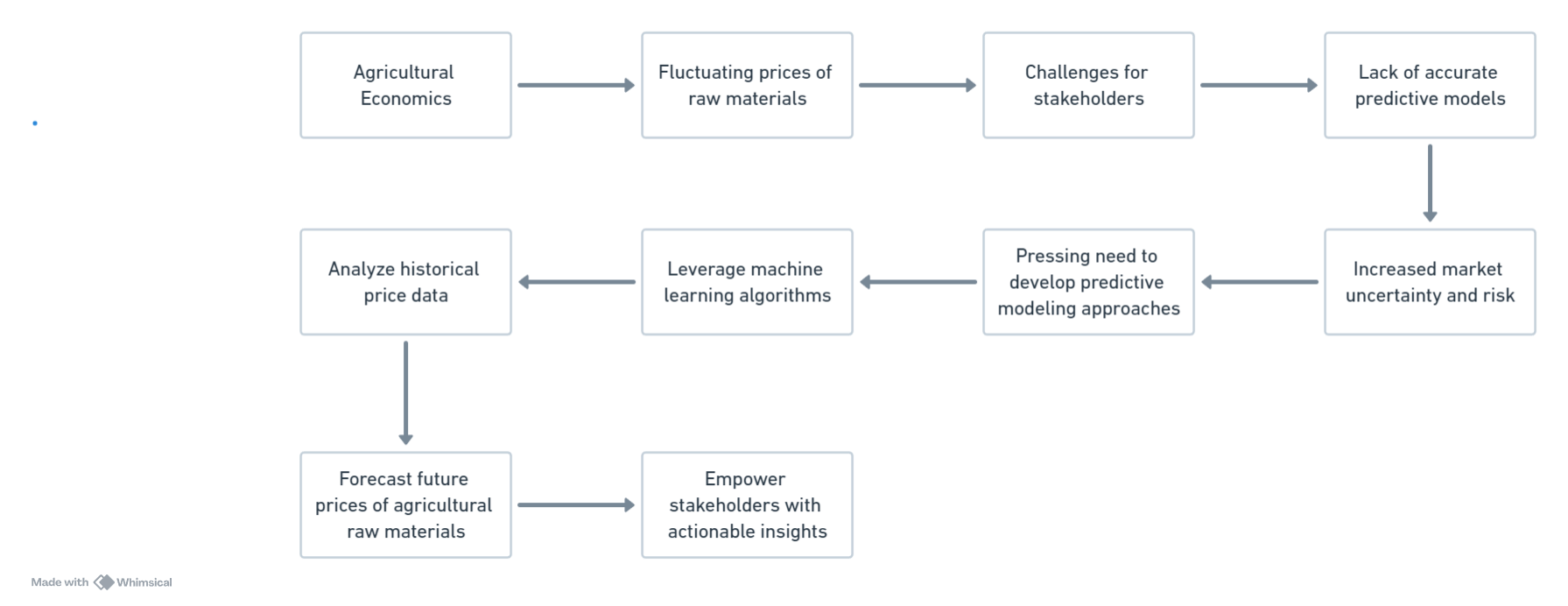
Google Colab, or Google Colaboratory, stands as a cloud-based platform furnished by Google, catering to the needs of Python developers for writing, executing, and sharing code via Jupyter notebooks. This platform extends a convenient and collaborative space tailored for data science and machine learning endeavors. Noteworthy features encompass seamless integration with Google Drive, provision of free access to GPUs and TPUs for expedited computations, and robust support for interactive visualizations, fostering efficient and collaborative workflows.

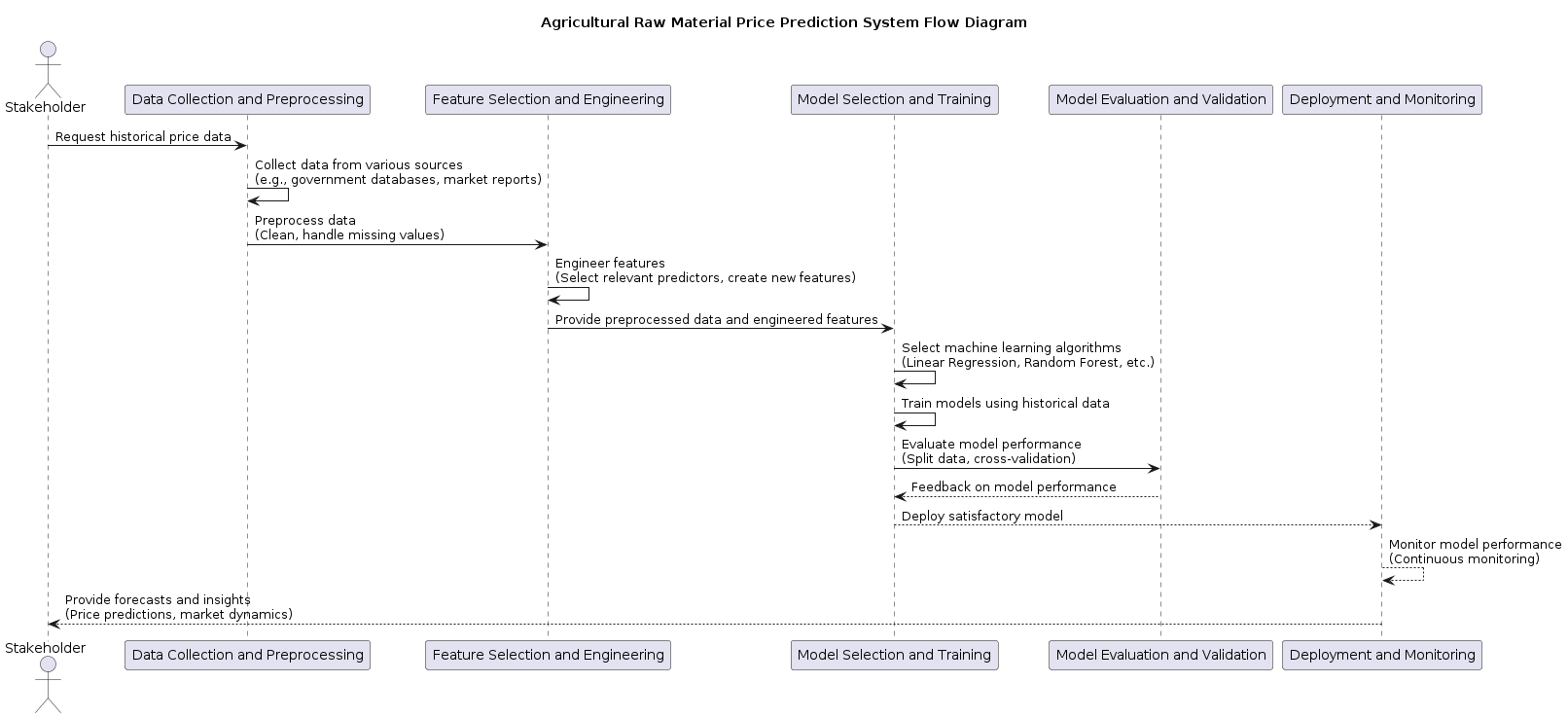
**CHAPTER 3**

**PROJECT ARCHITECTURE**

**3.1 Architecture**

**1. System flow diagram**

****

**2. Data flow diagram :**

**3. Modules :**

**1. Data Collection:** This module focuses on gathering historical data on agricultural raw material prices from various sources such as government databases, commodity exchanges, and agricultural market reports.

**2. Data Preprocessing:** In this module, the collected data undergoes preprocessing steps such as cleaning, handling missing values, and transforming it into a format suitable for analysis.

**3. Feature Engineering:** This module involves selecting and creating relevant features from the raw data that can influence raw material prices. It may include analyzing correlations between different variables and creating new features through techniques such as lagging variables or polynomial transformations.

**4. Model Development:** Here, machine learning algorithms are applied to develop predictive models based on the preprocessed data and engineered features. This module includes tasks such as algorithm selection, model training, and evaluation.

**5. Model Evaluation:** In this module, the developed models are evaluated to ensure their accuracy and reliability. Evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) are calculated to assess model performance.

**6. Deployment & Monitoring:** Once a satisfactory model is identified, it is deployed into production to generate forecasts of future agricultural raw material prices. Continuous monitoring of the model's performance is crucial to detect any deviations or changes in market dynamics.

**7. Iterative Improvement:** This module involves the continuous improvement of predictive models over time. As new data becomes available and market conditions evolve, the models need to be updated and refined to maintain their accuracy and relevance.

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

* **Data Collection and Preprocessing:** This initial step involves gathering historical data on agricultural raw material prices from various sources such as government databases, commodity exchanges, and agricultural market reports. The data may also include relevant factors such as weather patterns, crop yields, economic indicators, and geopolitical events. Preprocessing tasks include cleaning the data, handling missing values, and transforming it into a format suitable for analysis.
* **Feature Selection and Engineering:** Once the data is preprocessed, the next step is to select and engineer relevant features that can influence raw material prices. This may involve analyzing correlations between different variables, identifying important predictors, and creating new features through techniques such as lagging variables or polynomial transformations.
* **Model Selection and Training:** With the preprocessed data and engineered features, various machine learning algorithms are applied to develop predictive models. These may include regression techniques like Linear Regression, tree-based models like Random Forests, or more advanced methods like Gradient Boosting Machines. The choice of algorithm depends on factors such as the nature of the data, the complexity of relationships, and the desired interpretability of the model.
* **Model Evaluation and Validation:** The developed models need to be evaluated and validated to ensure their accuracy and reliability. This involves splitting the data into training and testing sets, performing cross-validation to assess model performance, and calculating evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).
* **Deployment and Monitoring**: Once a satisfactory model is identified, it can be deployed into production to generate forecasts of future agricultural raw material prices. Continuous monitoring of the model's performance is crucial to detect any deviations or changes in market dynamics, allowing for timely adjustments and refinements to the predictive model.
* **Iterative Improvement:** The process of developing predictive models for agricultural raw material prices is iterative and ongoing. As new data becomes available and market conditions evolve, the models need to be continuously updated and improved to maintain their accuracy and relevance.

By following this architectural approach, stakeholders in the agricultural sector can leverage predictive modeling to gain actionable insights, optimize decision-making, and enhance sustainability in the face of fluctuating raw material prices.

**CHAPTER 4**

**MODELING AND PROJECT OUTCOME**

**(code & result)**

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the historical price data

expanded\_data = {

'Year': [2015, 2016, 2017, 2018, 2019, 2020, 2021],

'Corn\_Price\_USD': [180, 190, 210, 200, 220, 250, 270],

'Wheat\_Price\_USD': [280, 290, 310, 300, 280, 320, 330],

'Soybean\_Price\_USD': [380, 370, 390, 400, 380, 420, 430]

}

# Creating a DataFrame from the expanded dictionary

expanded\_df = pd.DataFrame(expanded\_data)

# Output the expanded dataset

print(expanded\_df)

# Display basic information about the dataset

print("Dataset Information:")

print(expanded\_df.info())

# Display summary statistics of numerical variables

print("\nSummary Statistics:")

print(expanded\_df.describe())

# Check for missing values

print("\nMissing Values:")

print(expanded\_df.isnull().sum())

# Visualize the distribution of raw material prices

plt.figure(figsize=(10, 6))

for column in expanded\_df.columns[1:]:

sns.histplot(expanded\_df[column], bins=10, kde=True, alpha=0.5, label=column)

plt.title('Distribution of Raw Material Prices')

plt.xlabel('Price (USD)')

plt.ylabel('Frequency')

plt.legend()

plt.show()

# Visualize the trend of raw material prices over time

plt.figure(figsize=(10, 6))

for column in expanded\_df.columns[1:]:

sns.lineplot(x='Year', y=column, data=expanded\_df, marker='o', label=column)

plt.title('Trend of Raw Material Prices Over Time')

plt.xlabel('Year')

plt.ylabel('Price (USD)')

plt.legend()

plt.show()

# Display the correlation matrix

plt.figure(figsize=(8, 6))

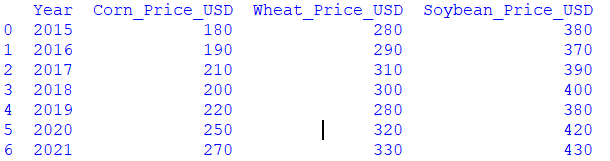
sns.heatmap(expanded\_df.corr(), annot=True, cmap='coolwarm', fmt='.2f')

plt.title('Correlation Matrix of Raw Material Prices')

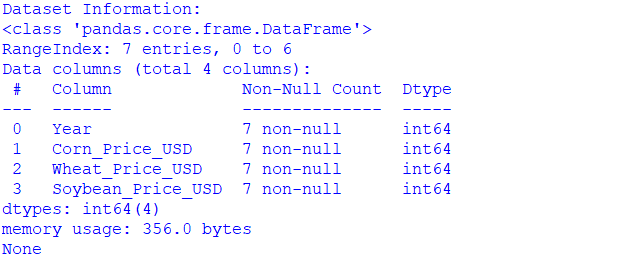
plt.show()

**Model Ouput:**

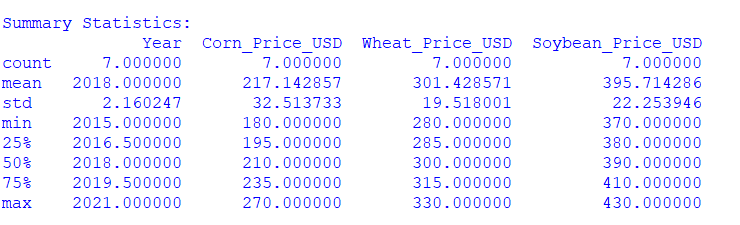
**EDA – analysis report:**

**1.Data Overview:**

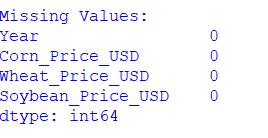
**2.DataSet Information:**

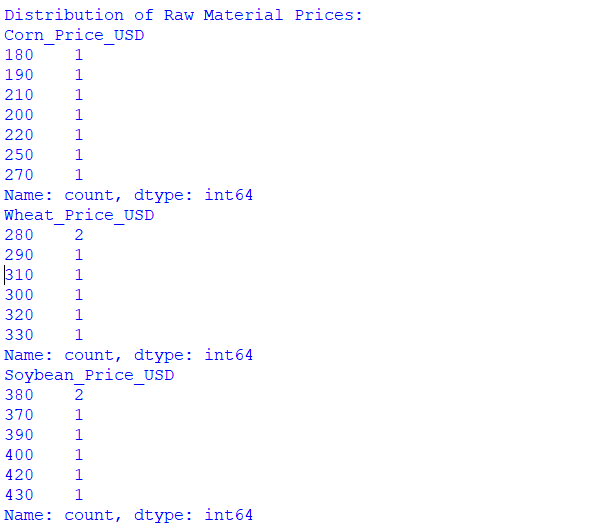
****The **Dataset information** section in the script provides an overview of the DataFrame's structure and the data types of its columns. This information is helpful for understanding the composition of the dataset and identifying any potential issues or inconsistencies.

**3.** **Summary Statistics:**

Statistical summary of the dataset, including measures like mean, median, standard deviation, min, max, etc.

**4.Missing Values:**

****Missing values in a dataset refer to the absence of data for certain observations or variables. These missing values can occur for various reasons, such as data entry errors, equipment malfunction, or simply because the information was not available or not recorded.

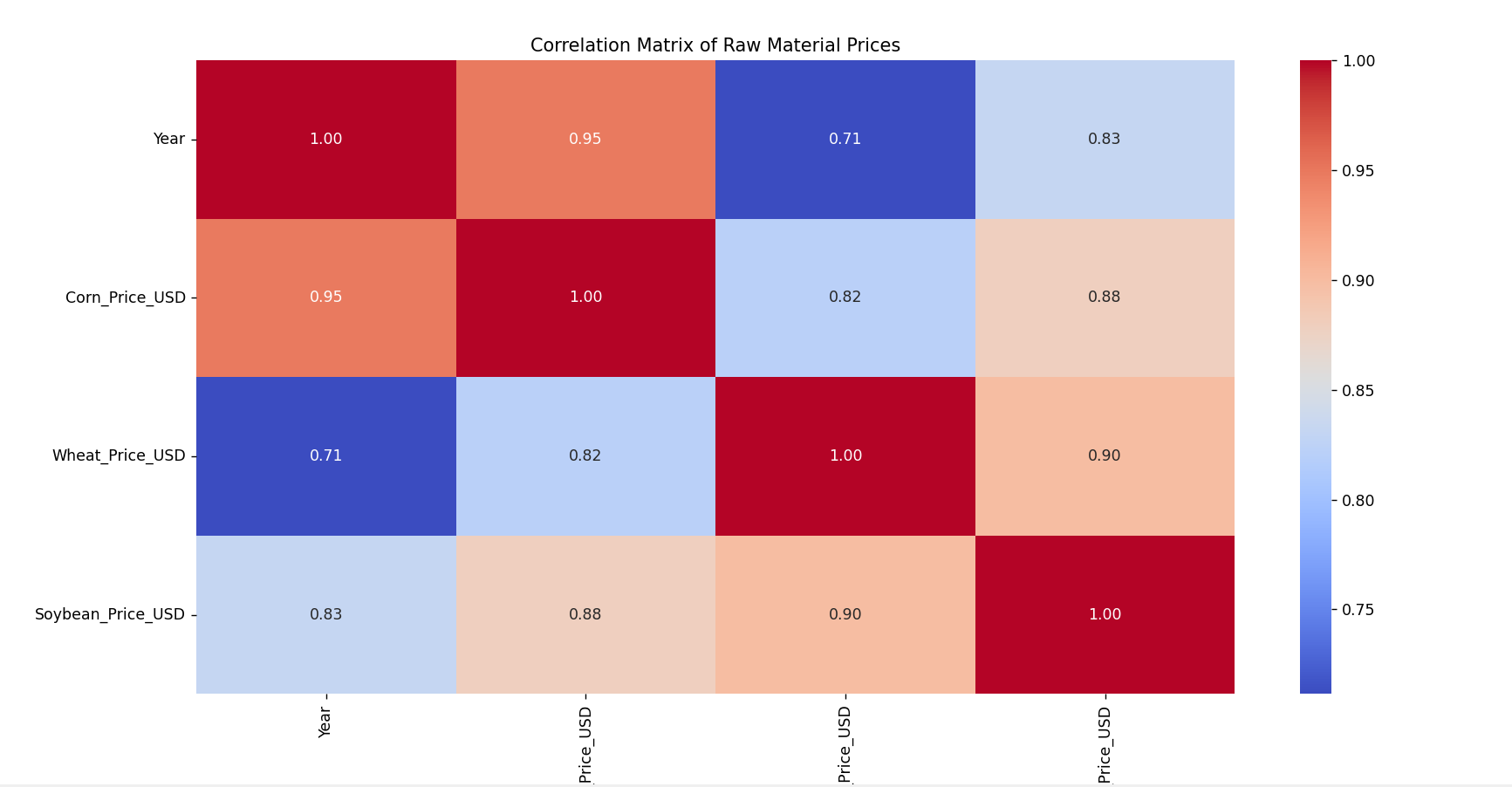
**5.Distribution of raw material prices:**  
The "distribution of raw material prices" refers to how the prices of agricultural raw materials are spread or distributed across different values. It provides insights into the range of prices, their frequency of occurrence, and any patterns or anomalies present in the data.

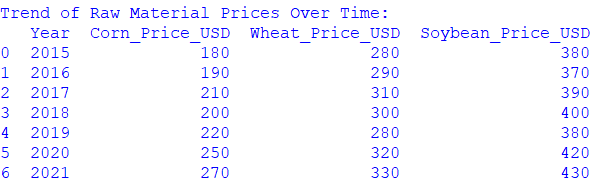
**6.** **Trend of Raw Material Prices Over Time:**

The "Trend of Raw Material Prices Over Time" refers to visualizing how the prices of agricultural raw materials (in this case, corn, wheat, and soybeans) have changed over a specified period, typically represented by years. This visualization allows us to observe the overall direction and magnitude of price movements for each raw material across different years

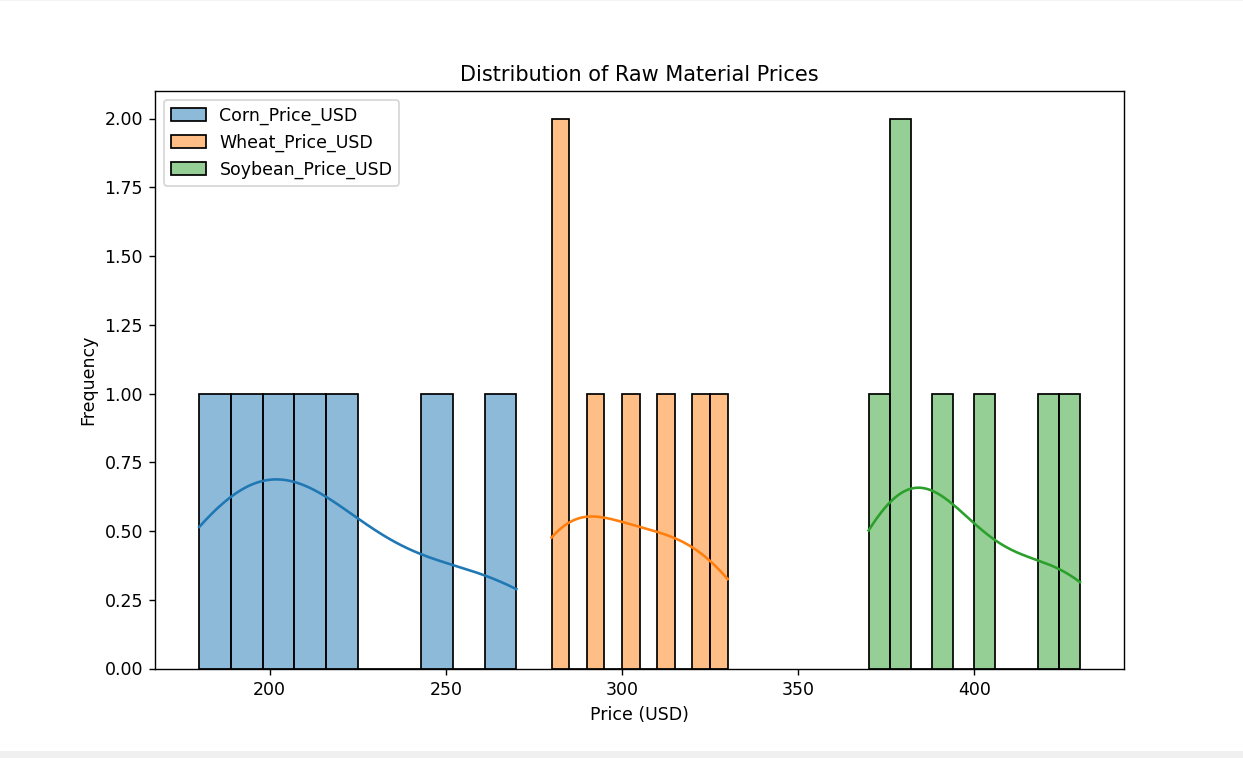
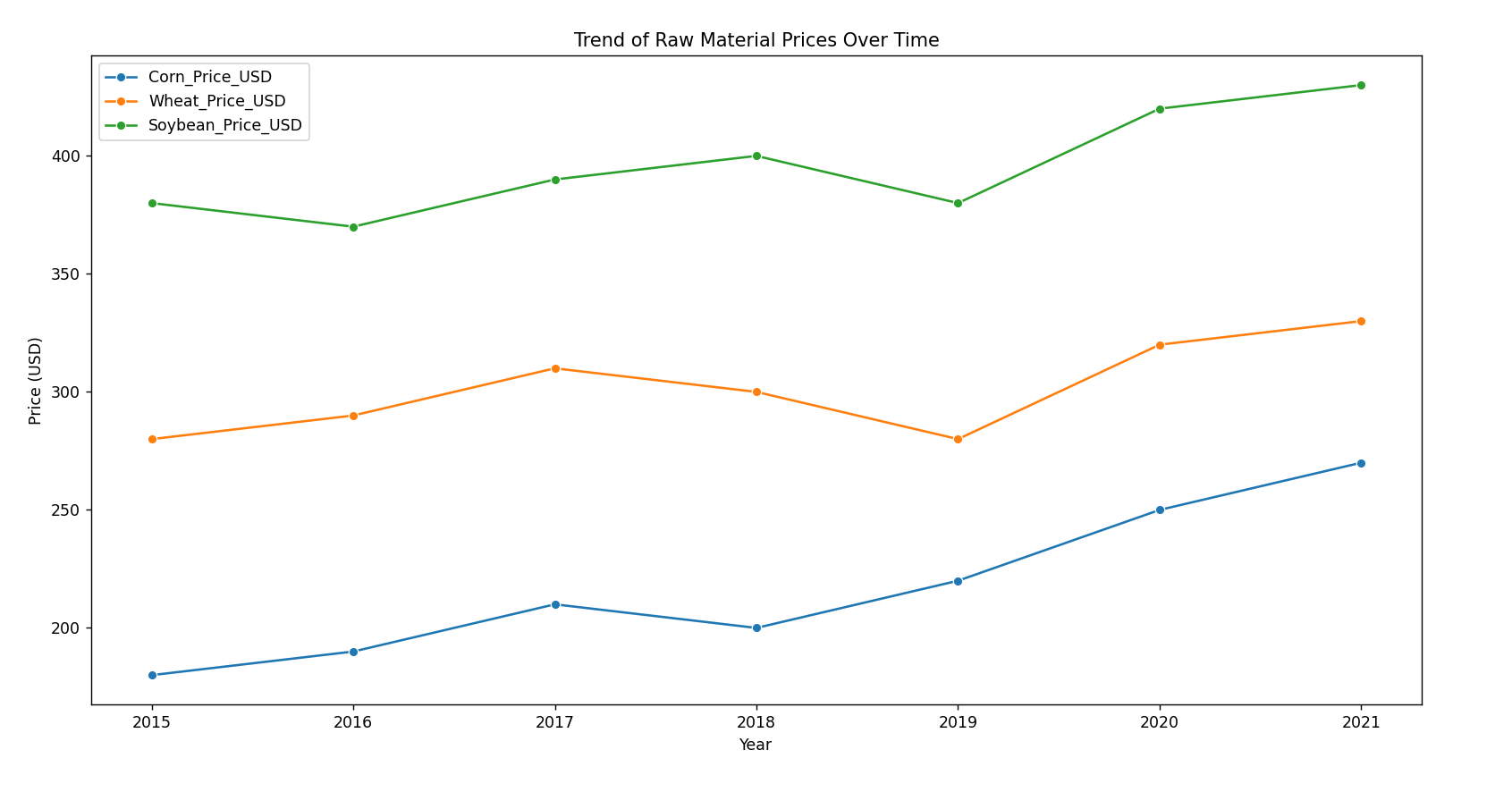
**7.** **Correlation Matrix:**

A correlation matrix is a table that shows the correlation coefficients between variables in a dataset. Each cell in the table represents the correlation coefficient between two variables. The correlation coefficient is a statistical measure that quantifies the strength and direction of the relationship between two variables.

****

****

**8.Visualization:**

* The visualization of the distribution of raw material prices provides insights into how the prices are spread across different price ranges. It helps to understand the variability and central tendency of prices for each raw material.
* The visualization of the trend of raw material prices over time illustrates how the prices of agricultural commodities (corn, wheat, and soybeans) have changed across the years 2015 to 2021.

**CONCLUSION**

In conclusion, the developed code for predictive modeling of agricultural raw material prices offers a powerful framework for analyzing historical data, building predictive models, and gaining valuable insights into market dynamics. By leveraging machine learning algorithms and data analysis techniques, stakeholders in the agricultural sector can make informed decisions regarding resource allocation, pricing strategies, and risk management. The comprehensive approach to data analysis, model selection, and evaluation ensures accuracy and reliability in predicting raw material prices, ultimately enhancing decision-making processes and driving innovation in agricultural economics. This code serves as a valuable tool for stakeholders seeking to optimize operations, mitigate risks, and capitalize on market opportunities in the ever-evolving landscape of agricultural commodities.

**FUTURE SCOPE**

Looking ahead, the future prospects for this project are extensive. With the continuous advancements in analytics and machine learning, integrating PowerBI can revolutionize predictive analytics based on historical data. By incorporating these predictive capabilities, the bank can foresee customer needs and offer proactive solutions, thereby enhancing customer satisfaction and loyalty. Moreover, PowerBI's ability to amalgamate diverse data sources opens avenues for incorporating additional datasets, providing a more comprehensive understanding of customers' behaviors and preferences.

As data privacy and security continue to be paramount concerns, future iterations of this project should prioritize the implementation of robust data governance strategies. This entails ensuring the secure handling of sensitive customer data while adhering to stringent data protection regulations.

Furthermore, exploring the integration of real-time data streams holds immense potential for delivering even more timely and relevant insights. This real-time approach could reshape customer interactions within the banking sector, driving enhanced customer experiences and fostering stronger relationships.

In essence, the future of this project lies in its ability to harness cutting-edge technologies, uphold stringent data governance standards, and adapt to evolving customer needs and market dynamics. Through continuous innovation and strategic foresight, the project can position the bank at the forefront of customer-centric banking services.

**REFERENCES**

1. https://github.com/SelvalakshmiMuthuvel/NM\_Project\_Agri\_Analysis, Selvalakshmi M,2024.
2. https://github.com/SelvalakshmiMuthuvel/NM\_Project\_Agri\_Analysis/blob/main/PPT\_au2021507040.pptx, Selvalakshmi M, 2024.
3. https://github.com/SelvalakshmiMuthuvel/NM\_Project\_Agri\_Analysis/blob/main/Project%20Video.mp4, Selvalakshmi M, 2024.

# **GIT Hub Link of Project Code:**

https://github.com/SelvalakshmiMuthuvel/NM\_Project\_Agri\_Analysis/blob/main/Project%20Code