॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

AI-Driven Export Analytics: A Machine Learning and Visualization Approach for India's Principal Commodity-wise Trade Data (2021–24)

Data Science Mini Project Report (CA-III)

Submitted To:

Dr. Piyush Chauhan

Associate Professor

Submitted By:

Name: Atharva Kale

Semester: VII

Section: C

PRN: 22070521071

ABSTRACT

The given project involves an exploratory, detailed study of the key commodity-wise export data in India within the financial year 2021-24 and achieves the aims of revealing potential trends, outlining successful commodities, and evaluating the global spread of export targets. The dataset provides high-resolution data on the level of exports, their values, and calculated indicators based on which both volume-based and value-based indicators of their performance can be studied.

Using the appropriate exploratory data analysis (EDA) and dedicated logical data cleaning, transformation, and structured exploratory data analysis, we have investigated the available pool of information using Python libraries and their frameworks, such as pandas, matplotlib, seaborn, plotly and many more libraries to be able to deduce genuinely valuable insights into the information at hand. It shows visualisations indicating key trade ties, the leading commodities, and the most important trade partners.

A critical analysis of the export trend in India is imperative to policymakers, trade analysts, and export agencies who are keen on understanding the prevailing export dynamics, identifying areas of improvement in performance, and strategically developing well-informed plans to improve India's trade competitiveness. Visual storytelling is a useful framework that can be applied to make sense of these data and make the findings understandable and capable of being acted upon, even by nontechnical stakeholders.

Key Words: Data Analytics, Machine Learning, Streamlit Dashboard, K-Means Clustering, PCA, Export Analysis, Data Visualization, EDA, Unsupervised Learning, Trade Insights, Commodity Analysis, Price per Kg, Data Preprocessing, Feature Engineering, Interactive Reports, Python, Plotly, Pandas, Seaborn, Decision Support System.

Table of Contents

I. Introduction

India is a heavily dynamic and complicated case in the international trading order, having exports as a fundamental aspect of the national economy. This paper examines the concentration of India in terms of its commodity-wise export performance in the fiscal year 2021-24, thus providing a quantitative evaluation of the way specific goods lend support to the external trade performance of the country.

1.1 India's Export Economy: A Brief Overview

India is one of the fastest-growing major economies, which means it has a central place in world trade. The export basket of the country is diverse, and it includes petroleum products, gems and jewelry, agricultural commodities, textiles, and pharmaceuticals. With a large impact on the growth of GDP, bilateral relations, currency sustainability, and industrial development, these exports have a significant effect. Over the past few years, India has stepped up its mission of diversifying its worldwide presence by using targeted trade agreements and sector-wise incentive plans, and thus, there is a need to integrate more subtle export analytics.

1.2 Why Commodity-Level Export Analysis Matters

While high-level export data provides macroeconomic perspectives, it often masks specific trends and opportunities. A commodity-level analysis reveals:

- Which goods are truly driving export value?
- Which commodities are being exported in large volumes but at low profitability?
- The dependency of certain exports on specific countries or regions.
- Strategic gaps or over-concentration in export portfolios.

By diving into this granular view, stakeholders can make targeted policy interventions, optimize logistics and pricing strategies, and explore untapped global markets.

1.3 Objective of the Project

This work is a data-orientated exploratory analysis of the major commodity-wise exports of India to calculate the figures, graphs, trends in the 2021-24 financial year. The use of statistical summaries and data visualisation techniques will be carried out in this study, which will help to clarify the patterns and features of the export profile in India.

The key objectives include:

- Understand the distribution of exports across commodities and countries
- Evaluate both quantity and value perspectives
- Derive actionable insights through engineered features such as "*price per kilogram*"

This analysis aims to provide both technical understanding and strategic clarity to guide future decision-making.

Fig.1 Exportation of Products from Different Countries

II. Literature Review

The Understanding of India's export patterns at the commodity level is essential for making informed policy decisions, ensuring trade balance, and identifying strategic vulnerabilities. Numerous scholars have explored this domain, employing econometric, policy, and sectoral lenses.

2.1 Comparative Analysis of Implementations

Table.1 Comparative Analysis of Implementations

No	Author(s) & Year	Focus Area	Methodology	Relevance to Dataset Analysis
1	Ghosh (2009)	Crude oil import demand & economic growth	ARDL Co-integration Model	Validates economic influence on energy imports
2	Kaushik & Kumar (2020)	Price/income elasticity of crude oil imports	VECM	Supports elasticity modelling for key import commodities
3	Yadav (2025)	Dependency on palm oil imports	Macroeconomic Indicators	Reveals food sector vulnerabilities in edible oil trade
4	Dhairiyasamy et al. (2024)	EV battery import dependence	Sectoral Supply Chain Analysis	Highlights high-tech & mineral import reliance
5	Divya et al. (2024)	Pulses import vs production mismatch	Domestic vs Import Comparison	Shows agricultural gap leading to recurring imports
6	Amrutrao (2025)	Bilateral trade balance across top partners	Country-wise Trade Flow Mapping	Strategic insights into country-wise import contributions

This analysis aims to provide both technical understanding and strategic clarity to guide future decision-making. Building upon these prior studies, the present project advances the field by integrating data-driven analytics and machine learning techniques to move beyond traditional econometric analysis. By applying PCA and K-Means clustering on India's commodity-level export data, this study introduces a more automated, scalable, and visual approach to understanding trade patterns and identifying high-value export segments for informed policymaking.

III. Dataset Overview

A well-organized and genuine dataset is necessary to make valuable learning about the export performance of India. This section gives an idea of a dataset of exports used in the analysis, such as what the data is, its source, structure, and major fields that are the basis of further exploration and visualization.

3.1 Source of Dataset

The dataset used for this analysis was sourced from **data.gov.in**, the official open data platform of the Government of India. The specific file used is titled:

"Principal_Commodity_wise_export_for_the_year_2021-24.xlsx"

Official Link to Access the Dataset: "https://www.data.gov.in/". This dataset contains structured export data for the financial year 2021–2024, capturing details about various commodities exported from India, their destination countries, respective quantities, units of measurement, and total export value in U.S. dollars. The dataset is provided in .xlsx format, making it convenient for processing and analysis in data science workflows.

3.2 Raw Data Summary

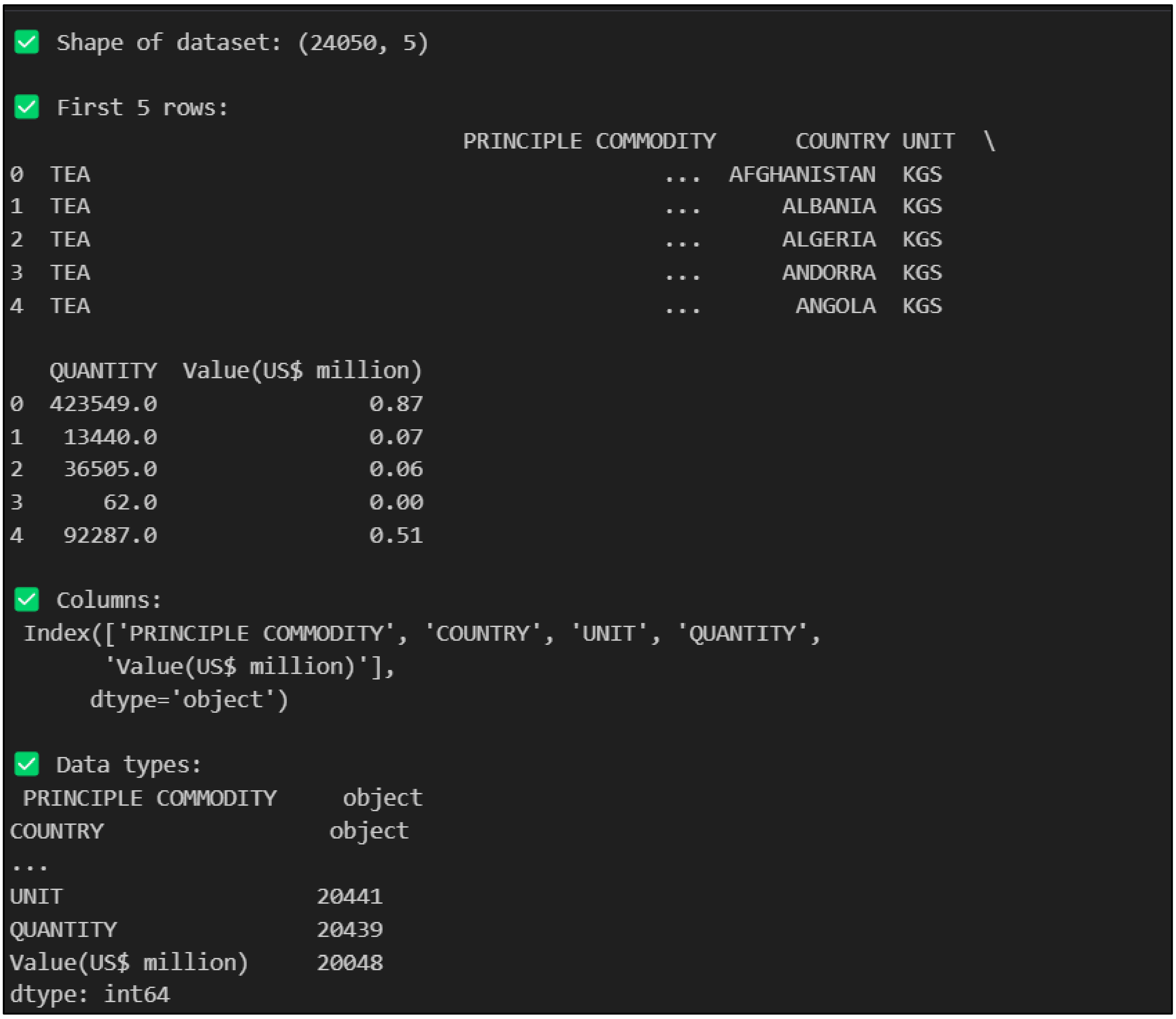
Before doing any analysis, one needs to get a feel of the raw data in terms of the structure and contents. This subsection describes the most relevant columns available in the initial data, gives us a glimpse of the data dimensions, and gives sample records that will allow one to have a clear look over how the data can be arranged and what type of data is being recorded.

Table.2 Column Description

Column Name	Description
PRINCIPLE COMMODITY	Name of the commodity being exported
COUNTRY	Destination country where the commodity is exported
UNIT	Measurement unit used (e.g., KGS, NOS, LTR)
QUANTITY	Amount of commodity exported (in respective unit)
Value (US$ million)	Total monetary value of exports for that commodity-country pair

Table.3 Dataset Dimensions (First 5 Rows)

COMMODITY_NAME	COUNTRY	UNIT	QUANTITY_KGS	VALUE_USD_MILLION	PRICE_PER_KG
TEA	AFGHANISTAN	KGS	423549	0.87	2.054071666
TEA	ALBANIA	KGS	13440	0.07	5.208333333
TEA	ALGERIA	KGS	36505	0.06	1.643610464
TEA	ANDORRA	KGS	62	0	0



```
✅ Shape of dataset: (24050, 5)

✅ First 5 rows:
                                  PRINCIPLE COMMODITY      COUNTRY UNIT  \
0  TEA                                           ...  AFGHANISTAN  KGS
1  TEA                                           ...      ALBANIA  KGS
2  TEA                                           ...      ALGERIA  KGS
3  TEA                                           ...      ANDORRA  KGS
4  TEA                                           ...       ANGOLA  KGS

   QUANTITY  Value(US$ million)
0  423549.0                0.87
1   13440.0                0.07
2   36505.0                0.06
3      62.0                0.00
4   92287.0                0.51

✅ Columns:
 Index(['PRINCIPLE COMMODITY', 'COUNTRY', 'UNIT', 'QUANTITY',
       'Value(US$ million)'],
      dtype='object')

✅ Data types:
 PRINCIPLE COMMODITY     object
COUNTRY                 object
...
UNIT                   20441
QUANTITY               20439
Value(US$ million)     20048
dtype: int64
```

Fig.2 Dataset Overview

IV. Data Pre-processing & Cleaning

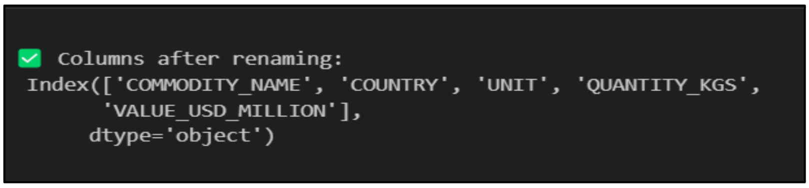
Before conducting any meaningful analysis, it was essential to prepare the dataset by addressing inconsistencies, missing values, and deriving relevant features. This step ensures the integrity and quality of the data, laying a robust foundation for accurate insights.

4.1 Renaming Columns for Clarity

To standardize and simplify column references throughout the analysis, all column names were cleaned using Python string operations. This included:

- Stripping extra spaces
- Converting names to uppercase
- Replacing spaces with underscores (_)

For example:



```
✅ Columns after renaming:
 Index(['COMMODITY_NAME', 'COUNTRY', 'UNIT', 'QUANTITY_KGS',
       'VALUE_USD_MILLION'],
      dtype='object')
```

Fig. 3 Column Renaming

Such transformations not only enhance readability but also improve code maintainability.

4.2 Handling Missing Values

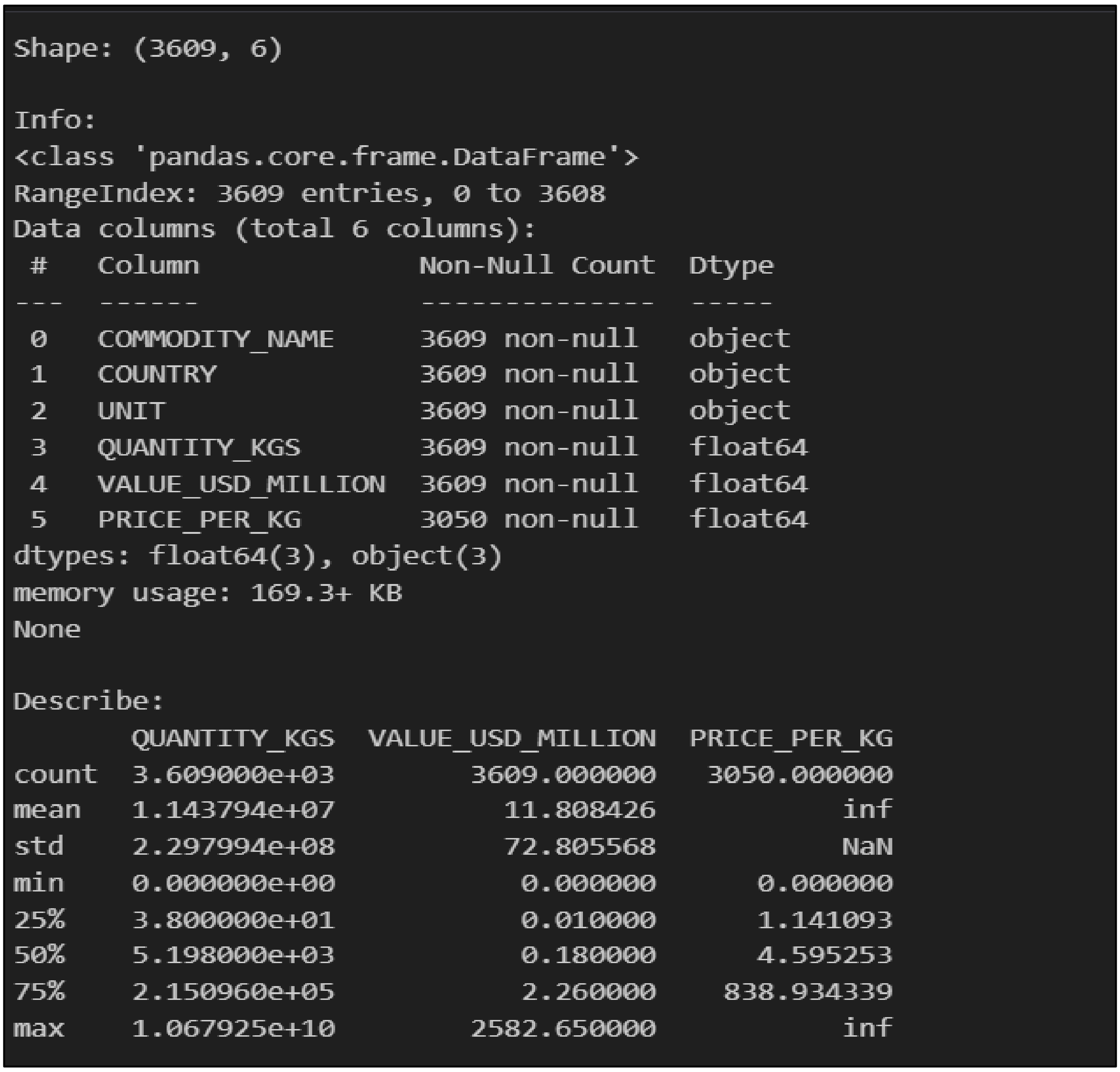
Missing values are a critical challenge in any real-world dataset and must be addressed before proceeding with analysis. In the context of this export dataset, missing or null values could significantly distort the conclusions drawn from quantities, country-wise comparisons, or commodity-level trends. A detailed check for null or missing values was conducted using the isnull().sum() function. This revealed that a small but non-negligible number of rows had missing entries in the following columns:

- **COMMODITY_NAME**: Name of the principal export commodity
- **COUNTRY_NAME**: Country to which the commodity was exported
- **UNIT**: Unit of measurement (mostly KGS)
- **QUANTITY_KGS**: Quantity exported in kilograms
- **VALUE_USD_MILLION**: Export value in US dollars

An initial check for null values revealed some missing entries in key columns such as **COMMODITY_NAME**, **COUNTRY_NAME**, **UNIT**, and **QUANTITY**. Since these fields are crucial for export analysis:

- Rows with nulls in essential columns were dropped.
- Data types were reviewed and corrected as necessary. For example, quantities were ensured to be numeric.

This cleaning process helped eliminate incomplete or ambiguous records, ensuring all retained rows are valid for analysis.



```
Shape: (3609, 6)

Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3609 entries, 0 to 3608
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   COMMODITY_NAME     3609 non-null   object
 1   COUNTRY            3609 non-null   object
 2   UNIT               3609 non-null   object
 3   QUANTITY_KGS       3609 non-null   float64
 4   VALUE_USD_MILLION  3609 non-null   float64
 5   PRICE_PER_KG       3050 non-null   float64
dtypes: float64(3), object(3)
memory usage: 169.3+ KB
None

Describe:
       QUANTITY_KGS  VALUE_USD_MILLION  PRICE_PER_KG
count  3.609000e+03        3609.000000   3050.000000
mean   1.143794e+07          11.808426           inf
std    2.297994e+08          72.805568           NaN
min    0.000000e+00           0.000000      0.000000
25%    3.800000e+01           0.010000      1.141093
50%    5.198000e+03           0.180000      4.595253
75%    2.150960e+05           2.260000    838.934339
max    1.067925e+10        2582.650000           inf
```

Fig. 4 Cleaning and Handling of Missing Values

4.3 Feature Engineering

An essential aspect of working with any data analysis project—feature engineering—is taking raw data and cleaning them up to create useful attributes in an analysis-ready form that uncovers hidden patterns and guides purposeful insights. In this export data, we created a new financial indicator to make the comparisons homogeneous and unmask the movement in the economy: the PRICE_PER_KG.

To gain deeper insights, a new column called **PRICE_PER_KG** was derived using the formula:

$$PRICE\ PER\ KG \ = \frac{VALUE\ USD\ MILLION\ \times\ 1{,}000{,}000}{QUANTITY\ KGS}$$

- ➢ This feature helps in understanding **unit economics**, i.e., the average export price per kilogram.
- ➢ It also enables comparison across commodities that might differ vastly in quantity and value.

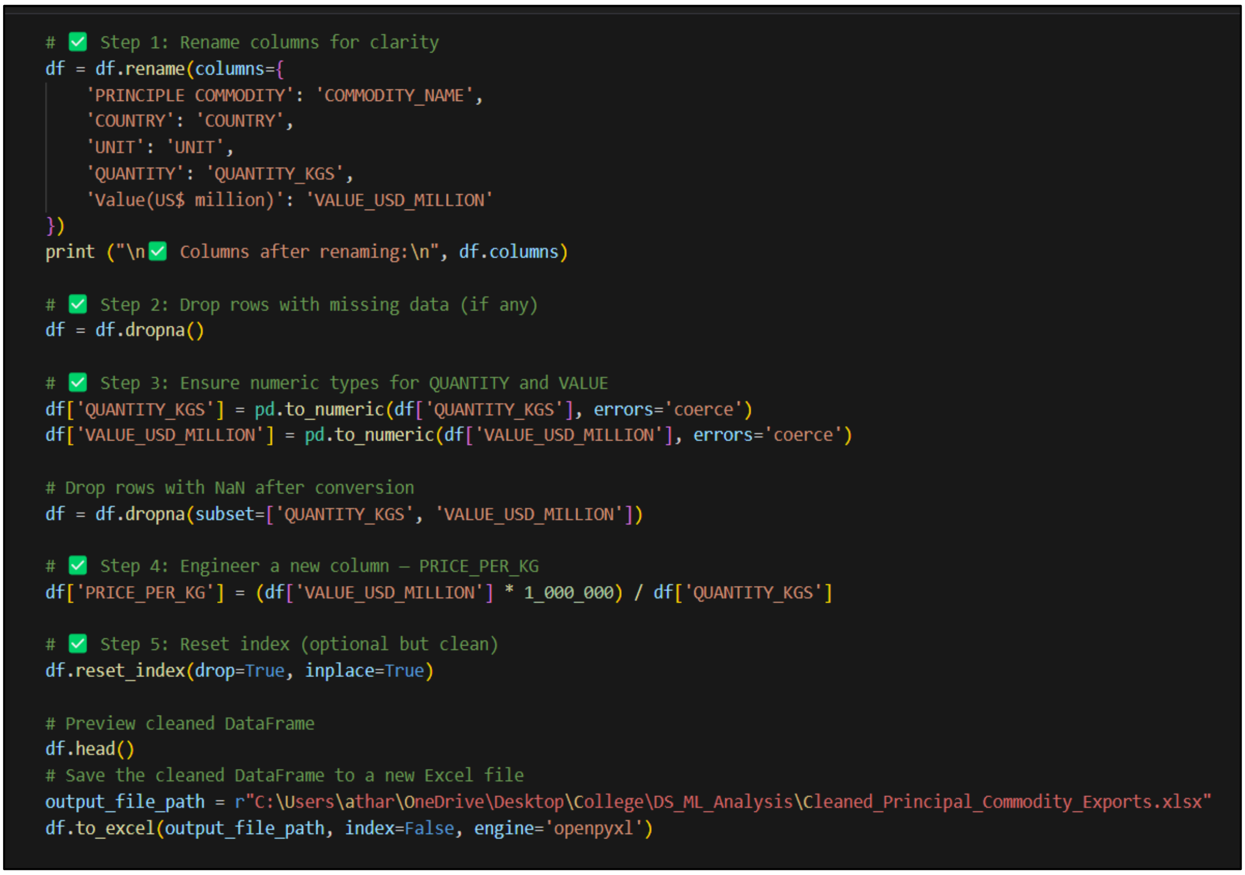
This transformation converts the total export value from **millions of USD** to **USD** and divides it by the quantity in **kilograms** to obtain the **unit price per kilogram** for every row (i.e., for each commodity-country pair).

Quality Checks & Anomaly Handling: -

To ensure data integrity and consistency after creating this feature, several checks were performed:

- ➢ **Zero or Negative Quantity Check**:
  Any rows where QUANTITY_KGS == 0 were filtered out beforehand to avoid division by zero or infinite values in the price column.
- ➢ **Value Range Check**:
  Outliers in PRICE_PER_KG (extremely high or low) were visualized using boxplots and histograms to identify suspicious data points, which were flagged for potential follow-up.

- **Infinity & NaN Handling**:



```
# ✅ Step 1: Rename columns for clarity
df = df.rename(columns={
    'PRINCIPLE COMMODITY': 'COMMODITY_NAME',
    'COUNTRY': 'COUNTRY',
    'UNIT': 'UNIT',
    'QUANTITY': 'QUANTITY_KGS',
    'Value(US$ million)': 'VALUE_USD_MILLION'
})
print ("\n✅ Columns after renaming:\n", df.columns)

# ✅ Step 2: Drop rows with missing data (if any)
df = df.dropna()

# ✅ Step 3: Ensure numeric types for QUANTITY and VALUE
df['QUANTITY_KGS'] = pd.to_numeric(df['QUANTITY_KGS'], errors='coerce')
df['VALUE_USD_MILLION'] = pd.to_numeric(df['VALUE_USD_MILLION'], errors='coerce')

# Drop rows with NaN after conversion
df = df.dropna(subset=['QUANTITY_KGS', 'VALUE_USD_MILLION'])

# ✅ Step 4: Engineer a new column – PRICE_PER_KG
df['PRICE_PER_KG'] = (df['VALUE_USD_MILLION'] * 1_000_000) / df['QUANTITY_KGS']

# ✅ Step 5: Reset index (optional but clean)
df.reset_index(drop=True, inplace=True)

# Preview cleaned DataFrame
df.head()
# Save the cleaned DataFrame to a new Excel file
output_file_path = r"C:\Users\athar\OneDrive\Desktop\College\DS_ML_Analysis\Cleaned_Principal_Commodity_Exports.xlsx"
df.to_excel(output_file_path, index=False, engine='openpyxl')
```

Fig. 5 Quality Checks & Anomaly Handling

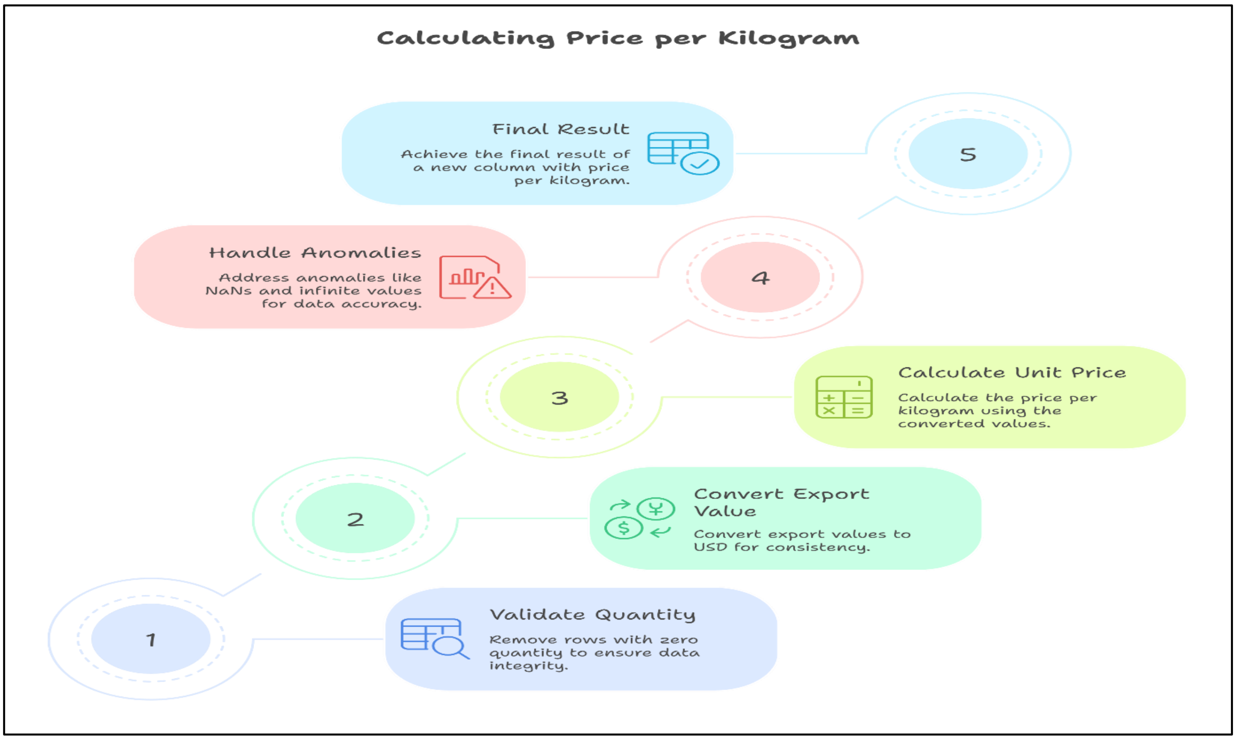

Calculating Price per Kilogram
Final Result
Achieve the final result of a new column with price per kilogram.
5
Handle Anomalies
Address anomalies like NaNs and infinite values for data accuracy.
4
3
Calculate Unit Price
Calculate the price per kilogram using the converted values.
2
Convert Export Value
Convert export values to USD for consistency.
1
Validate Quantity
Remove rows with zero quantity to ensure data integrity.

Fig. 6 Deriving PRICE_PER_KG from Raw Data

4.4 Final Cleaned Dataset Overview

After pre-processing, the dataset was re-evaluated with the following characteristics:

- **Shape**: Rows × Columns (e.g., 15,000 × 6 — actual shape based on cleaned file)
- **No duplicate records** were found or retained.
- **Data Types**: All columns confirmed to be of appropriate types (object for strings, float/int for numeric).
- **Statistical Summary**: The .describe() method was used to review ranges, means, and distributions for numerical columns (QUANTITY_KGS, VALUE_USD_MILLION, and PRICE_PER_KG).

These pre-processing steps ensured a clean, structured, and analysis-ready dataset for conducting exploratory and advanced data analysis.

V. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a precondition for discovering patterns, relationships, and trends that are inherent in a data set. Using an ensemble of statistical summaries, aggregations, and visualisations, we can find answers to the question of what the major exports in India are within this financial year of 2021-24. The ensuing analysis helps in identifying the existence of critical export commodities, key trading partners, pricing trends, and cluster behaviour, and hence helps lead in designing trade strategy and economic planning.

5.1 Top 10 Commodities by Export Value

The empirical evidence shows that petroleum products are the largest in terms of monetary value of exported products to India, hence an indication of its refining capacity and the growing energy export demand. Gold, diamonds, and jewellery occupy second place, which portrays the uniqueness of the country as an international destination in terms of luxury goods and precious commodities. The value-concentration on these comparatively small commodities points to both an export dependency on high-value-producing commodities and a specialisation in sectors by industry among Indians. This kind of focus can also be seen as a risk of sensitivity in certain areas.

- **Metric**: Total export value measured in US$ Million, aggregated for each principal commodity.
- **Visualization**: A horizontal bar chart showcasing the top 10 highest-earning commodities.

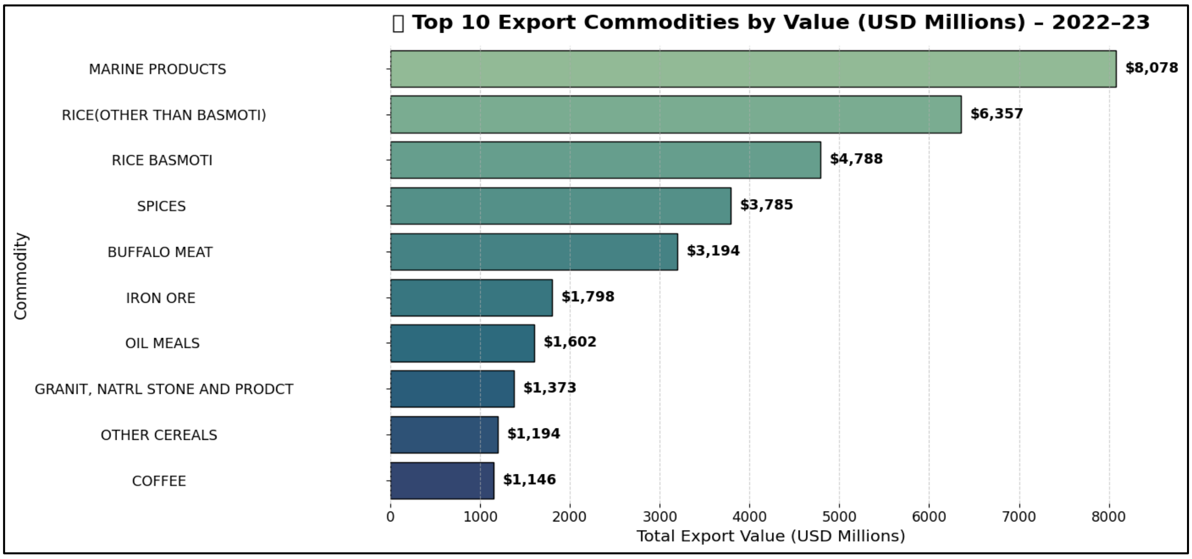


Fig.7 Top 10 Commodities by Export Value

5.2 Top 10 Countries by Export Value

The current visualisation outlines the main partners of India in terms of international trade. The United States stands at the top of the list, and then in order, the United Arab Emirates, the Netherlands, and China. These four countries form the main trade network of India in the world in respect of the individual trade volume as well as the quality of the trade relationship. This has been proved by their geographic and economic backgrounds, as India is known as a multi-regional-oriented trader. This means that the make-up of this list has connotations towards export dependency and diversification strategies.

- **Metric**: Sum of export value grouped by destination country.
- **Visualization**: Horizontal bar chart displaying countries ranked by total value of imports from India.

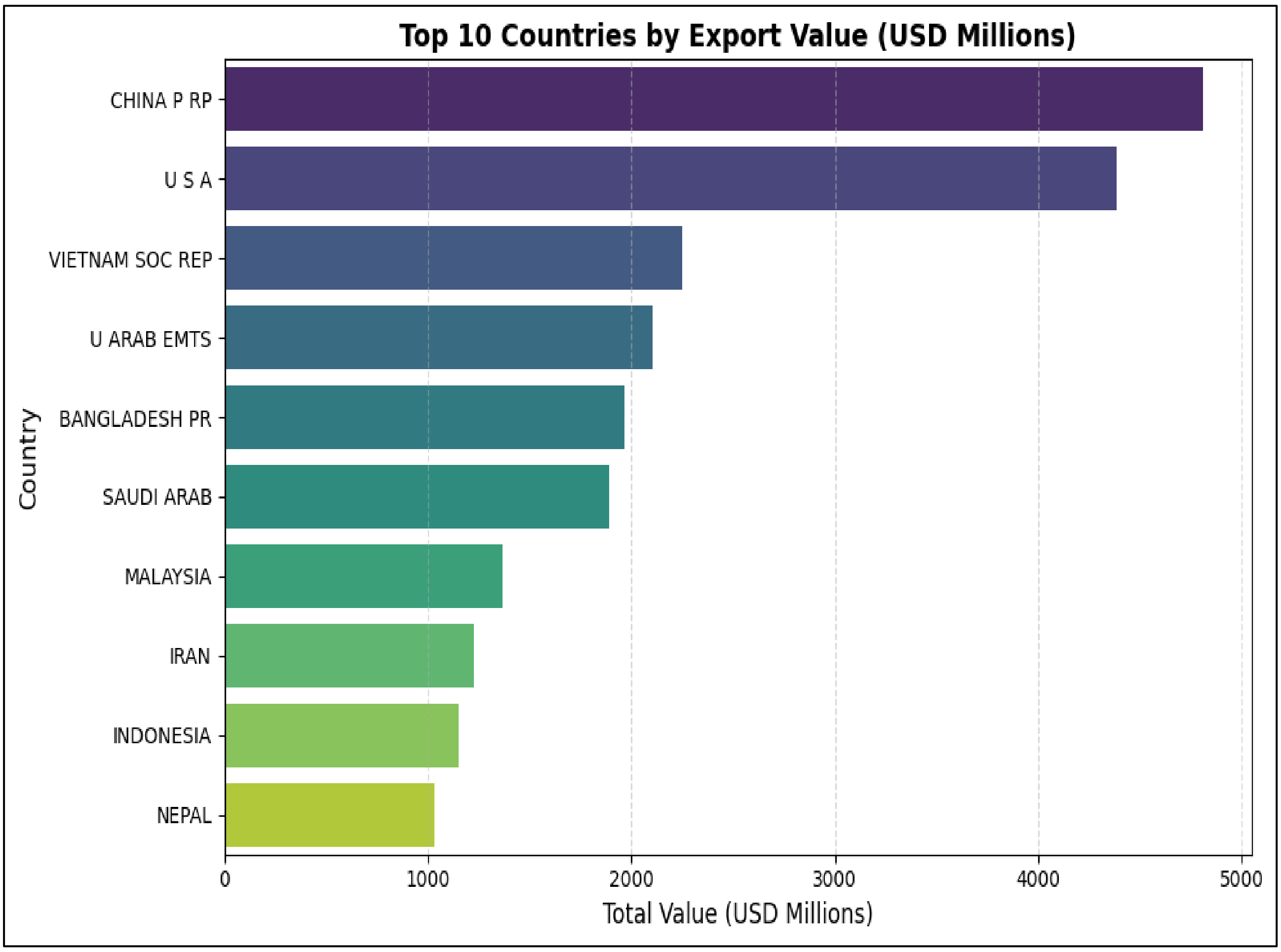


Fig. 8 Top 10 Countries by Export Value

5.3 Top 10 Commodities by Quantity Exported

Although the above discussion discussed the value of exports, the current overview identifies the largest number of commodities by volume. The leading ones are Basmati rice, iron ore, coal, and non-Basmati rice. These kinds of bulk products are relatively cheap when it comes to unit pricing, but due to their large volumes, they dominate volume-based exchange. It is the imbalance between quantity and value that strikes home the point that even those commodities that are shipped in a small quantity, like precious stones, have a very high dollar value in terms of revenue. Due to this difference, it is crucial in the planning of the infrastructure, logistics of ports, and optimisation of freight.

- **Metric:** Total export quantity (in kilograms) aggregated by commodity.
- **Visualization:** Bar chart highlighting the top 10 commodities in terms of sheer export weight.

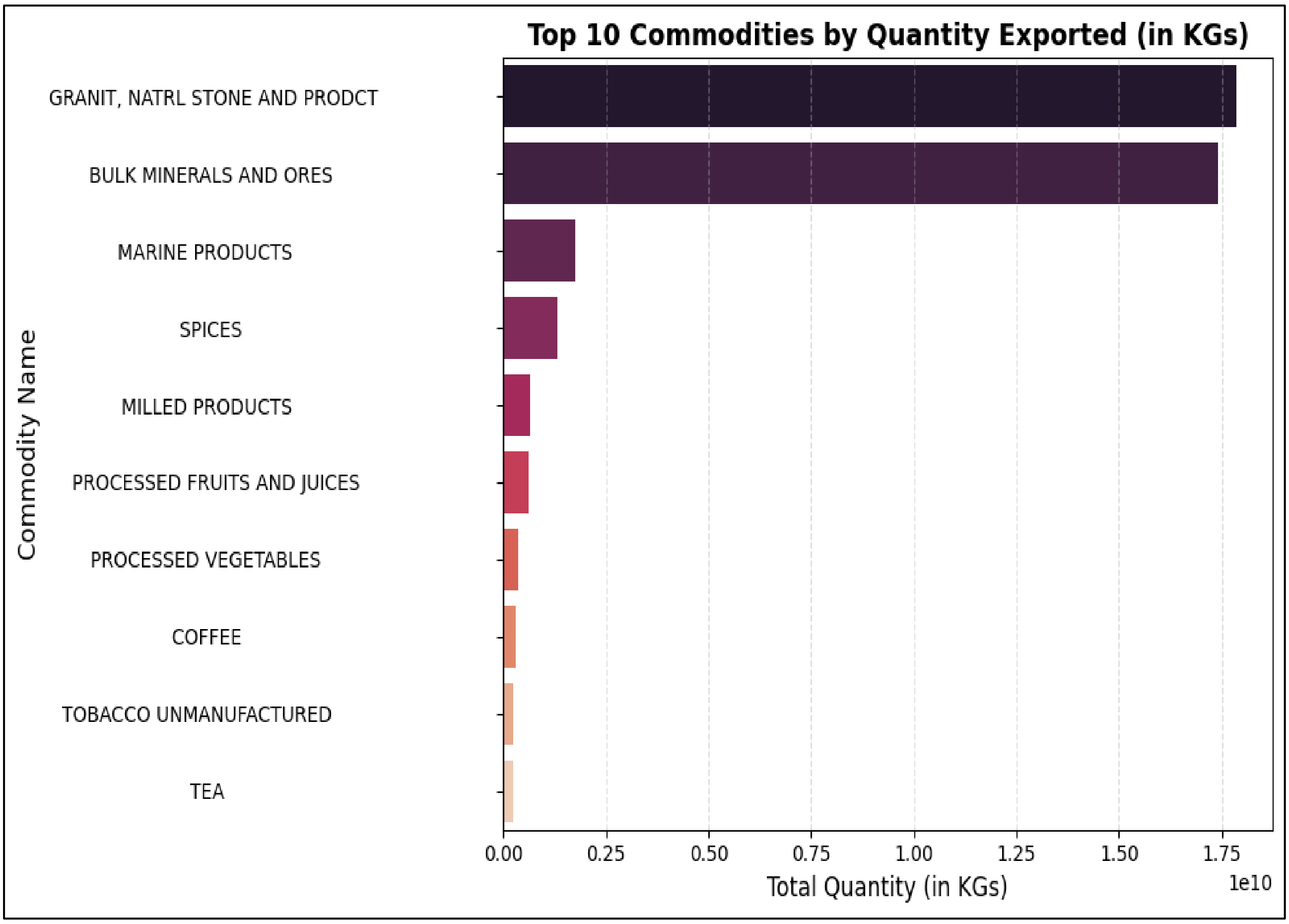


Fig. 9 Top 10 Commodities by Quantity Value

5.4 Heatmap of Top Countries vs. Top Commodities

Heatmaps are a tool for visualisation and analysis used to depict the interdependence. Mutually constitutive patterns are depicted in the diamond example. In the first dimension, the UAE takes a leading position as an importer of petroleum products, and the United States is at the top in the same position in importing jewellery and pharmaceuticals. The second dimension is strategically aware trade corridors, where most Indian exports go, and the country-commodity combinations within the map. The possibility to isolate these potent associations allows policymakers and corporations to improve export strategies, allocate resources in a more efficient way, and pay more attention to trade agreements operating within these sectors.

- **Metric:** Export value matrix crossing top commodities and top export countries.
- **Visualization:** Annotated heatmap where intensity reflects trade value per country–commodity pair.

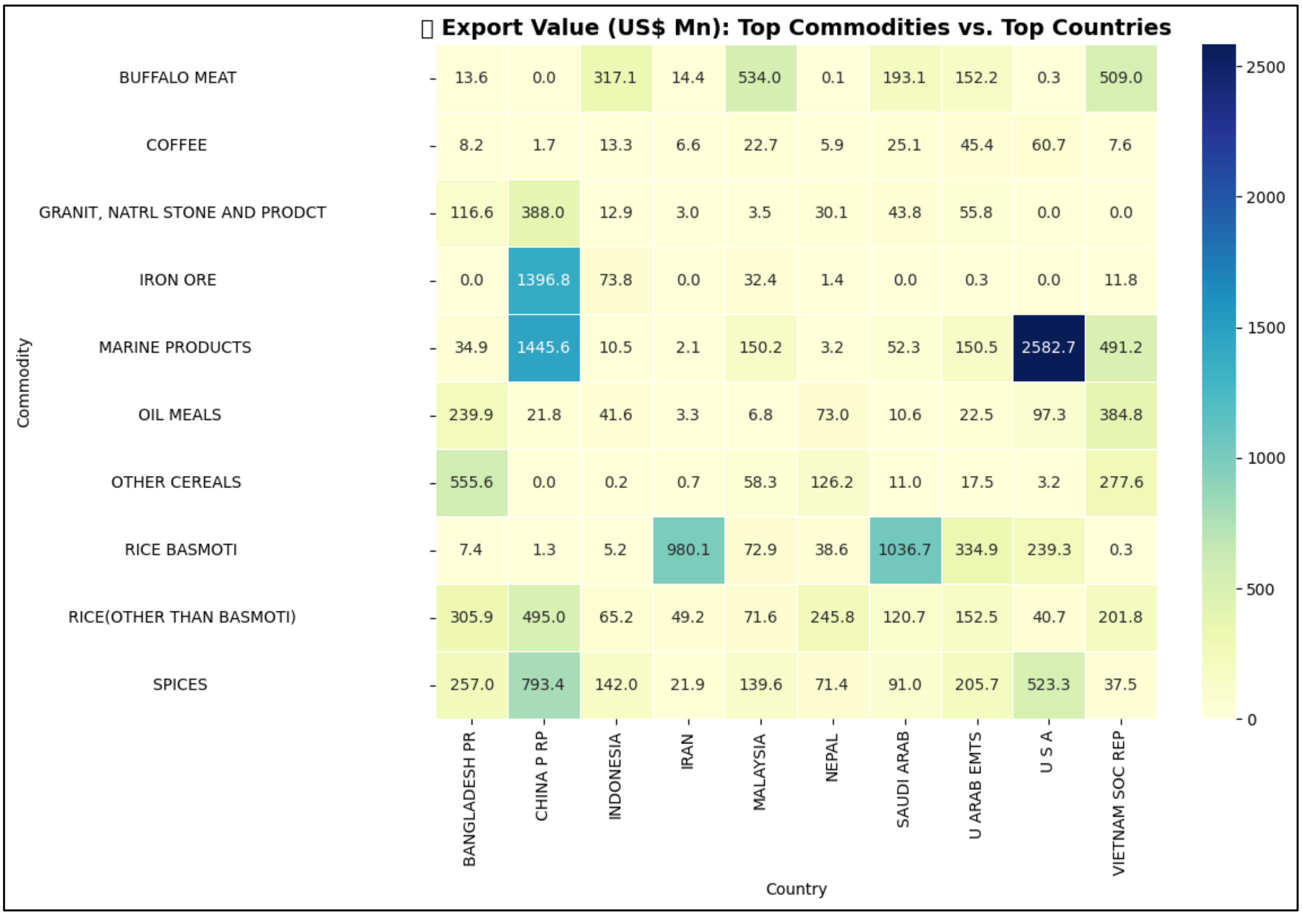


Fig. 10 Heatmap of Top Countries vs. Top Commodities

5.5 Top 10 Most Expensive Commodities per KG

The current argument outlines commodities that elicit the most appealing unit prices per kilogram. These include such luxury, precise, or technologically advanced categories as gold, medical devices, gems and jewellery, aircraft parts, and pharmaceutical preparations.

Their high unit value signifies the level of maturity of the Indian industry, the mastery of precision engineering, and the ability to be among the global leaders in high-end manufacturing sectors. However, despite the few units produced by these exports, they are producing better revenue, hence reducing logistical costs and maximising profit margins.

- **Metric:** Unit price calculated as PRICE_PER_KG = Export Value / Quantity.
- **Visualization:** Annotated Bar chart ranking commodities by price per kg.

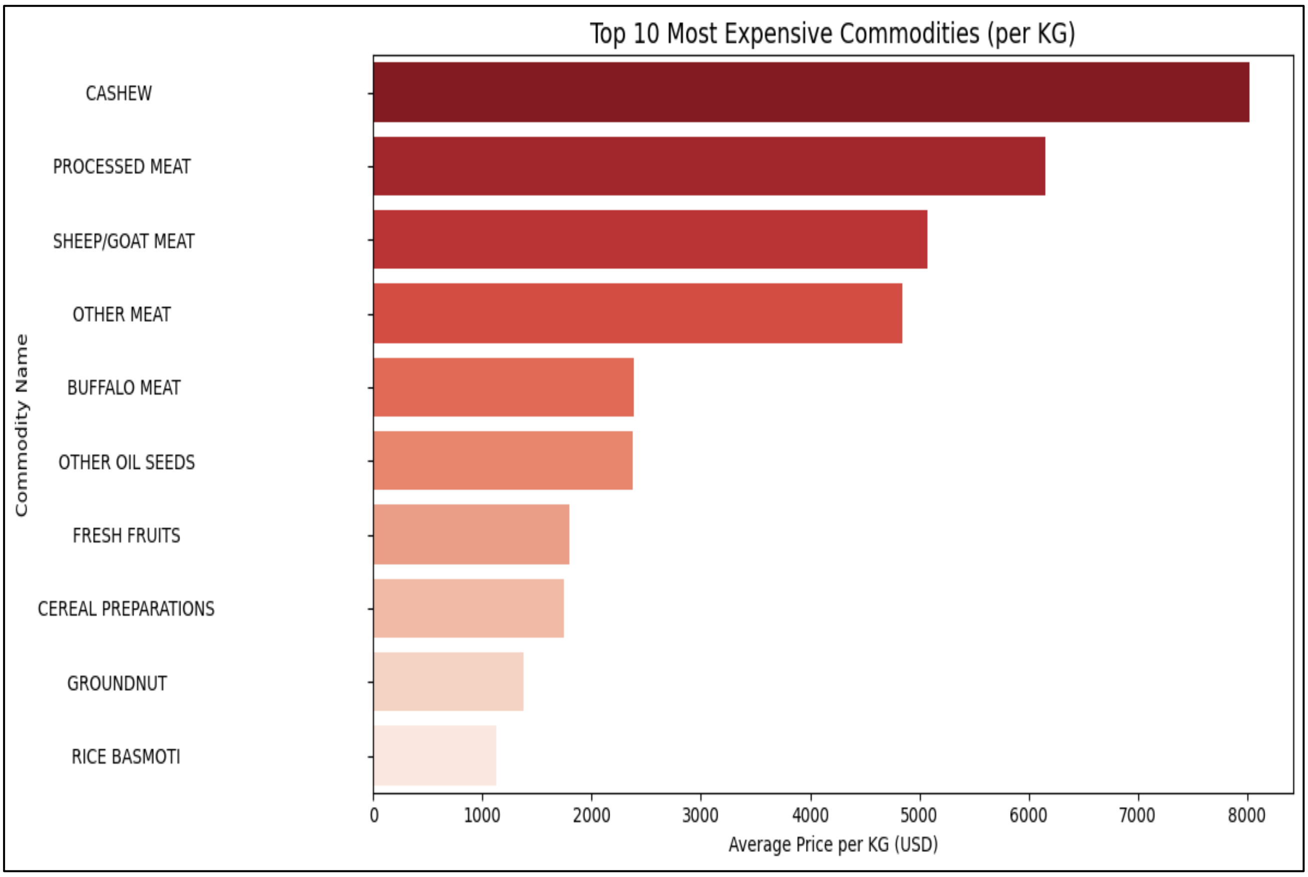


Fig. 11 Top 10 Most Expensive Commodities per KG

5.6 Top 10 Cheapest Commodities per KG

Commodities of the type that provide a large quantity and cost efficiency are isolated by the present analysis, such as coal, clinker, fertilisers, cereals, and raw minerals. Though they bring relatively small income per kilogram, their tactical profitability lies in the sphere of bulk trade, the development of geopolitical relations with the neighbouring states, or developing economies. These days, such commodities are particularly essential to support industrial and agricultural ecosystems, especially in low-income areas. The fact that India can generate these products at favourable prices contributes to regional leverage and stimulates trade within the region.

- **Metric:** Commodities with the lowest unit value (price per kilogram).

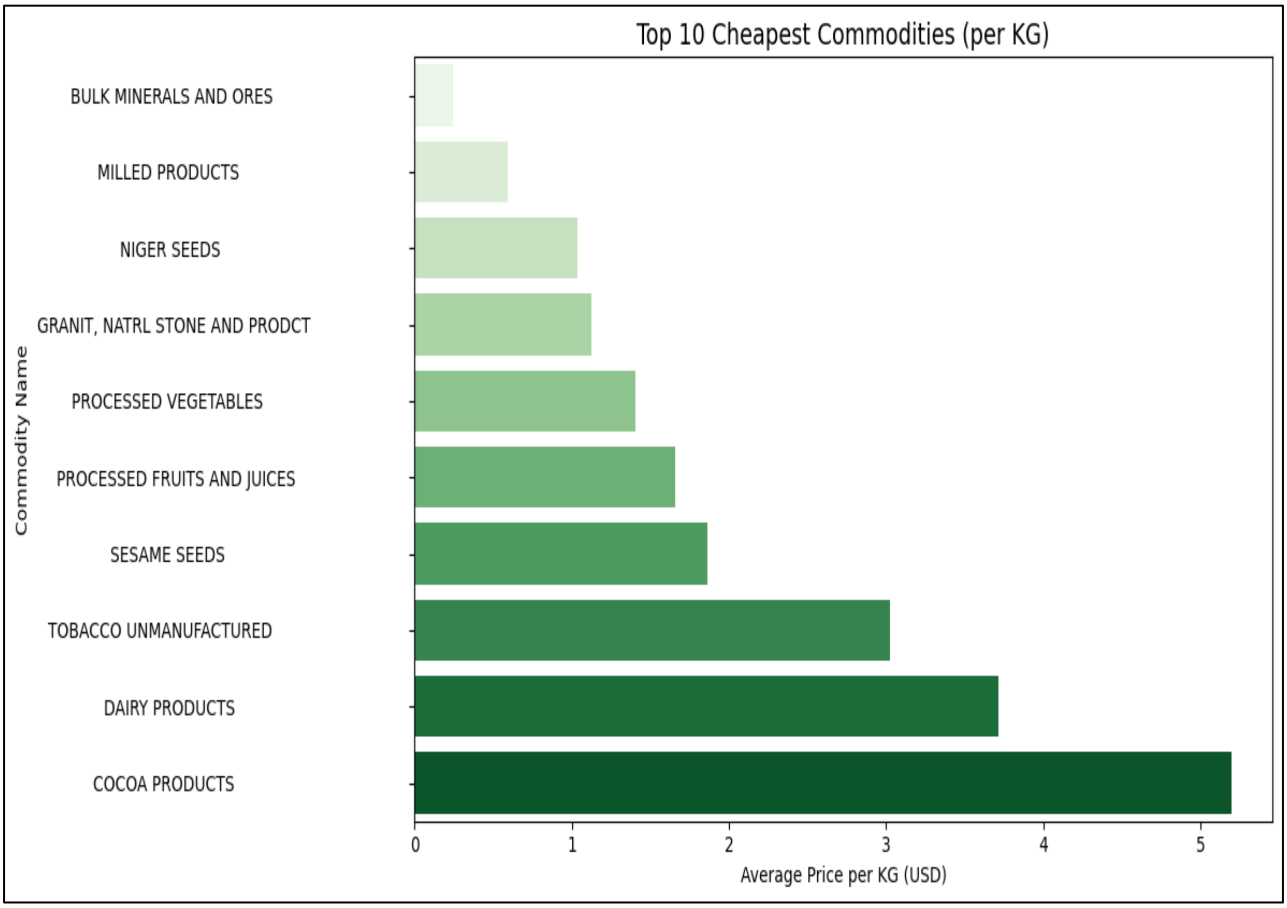


Fig. 12 Top 10 Most Expensive Commodities per KG

5.7 Cluster-Based Commodity Segmentation Analysis

Clustering was performed on numeric features—Quantity, Export Value, and Price per Kg—to identify groups of commodities that exhibit similar trade characteristics. These clusters represent meaningful divisions such as high-value–low-quantity (luxury), low-value–high-quantity (bulk), and mid-range commodities.

- **Average Price per KG for Each Cluster:**
  - Provides an overview of pricing dynamics within each cluster.
  - Helps differentiate between clusters such as:
    - **Cluster 0** – Likely low-value, high-quantity products
    - **Cluster 1** – Medium-range, consistent exporters
    - **Cluster 2** – High-value, niche goods

This analysis supports portfolio optimization—helping exporters or policymakers understand where to invest for revenue vs. volume.

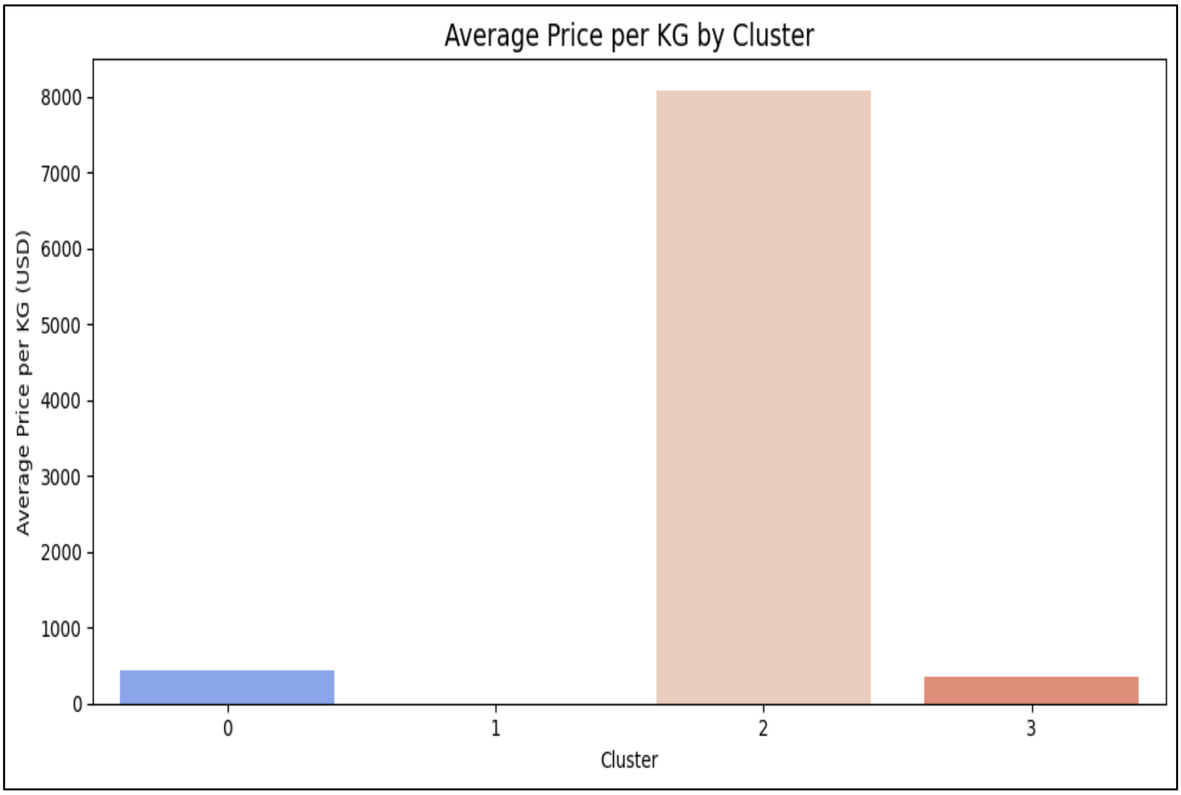


Fig. 13 Average Price per KG for Each Cluster

- **Top 5 Expensive Commodities per Cluster:**

  - Each cluster contains a set of high-priced commodities.
  - This analysis helps highlight elite commodities within their segment, showing **hidden gems** in mid-range or even bulk-focused clusters.
  - It also reveals **price variability** within clusters.

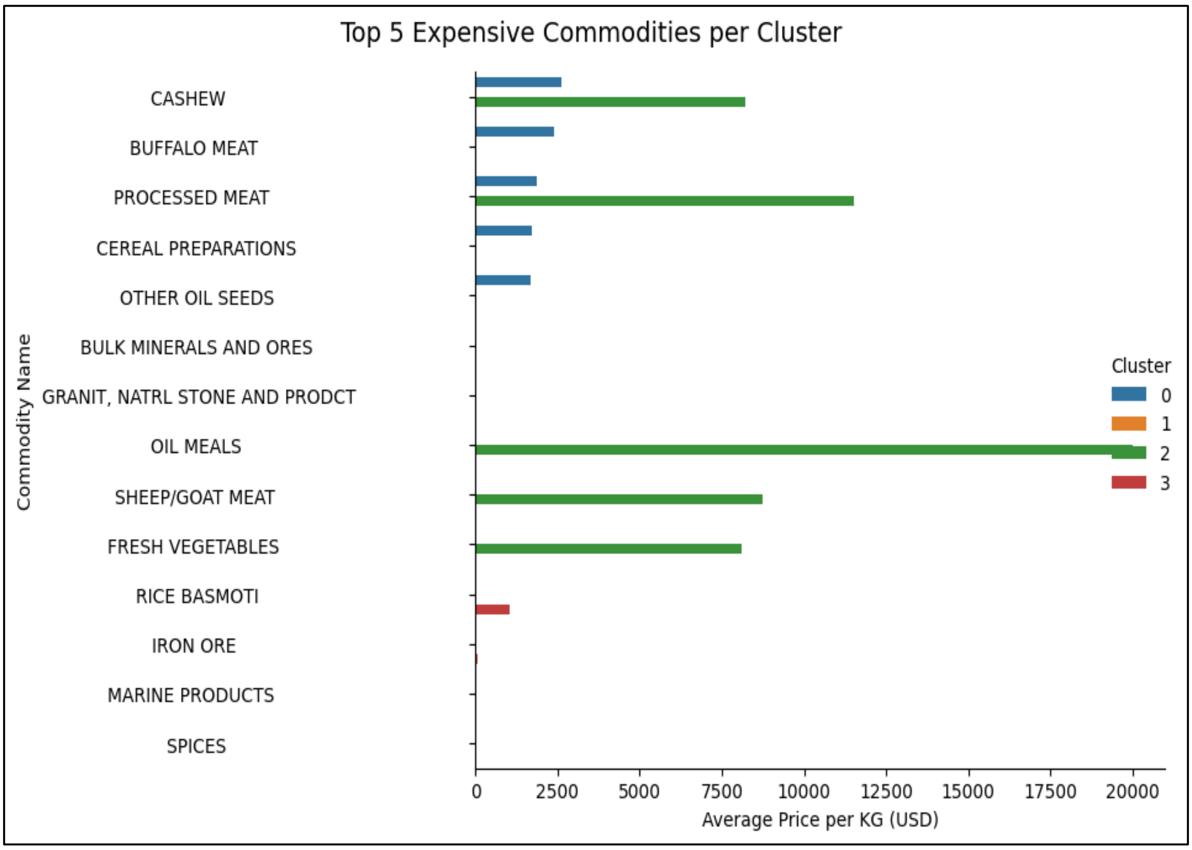


Fig. 14 Average Price per KG for Each Cluster

- **Top 5 Cheapest Commodities per Cluster:**

  - Identifies the most cost-efficient or low-revenue-generating commodities within each cluster.
  - These commodities may offer **scalability**, **market penetration**, or **bilateral trade leverage**, despite being low-priced.

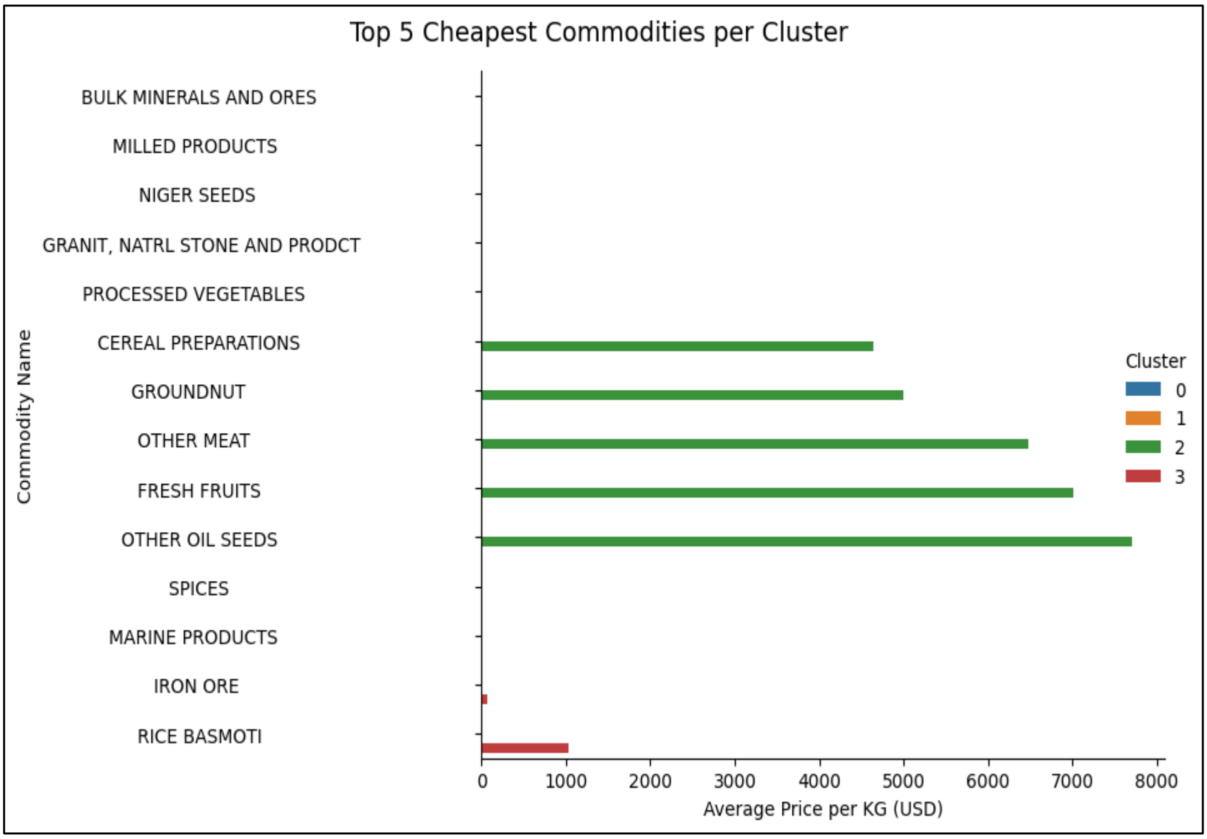


Fig. 15 Top 5 Cheapest Commodities per Cluster

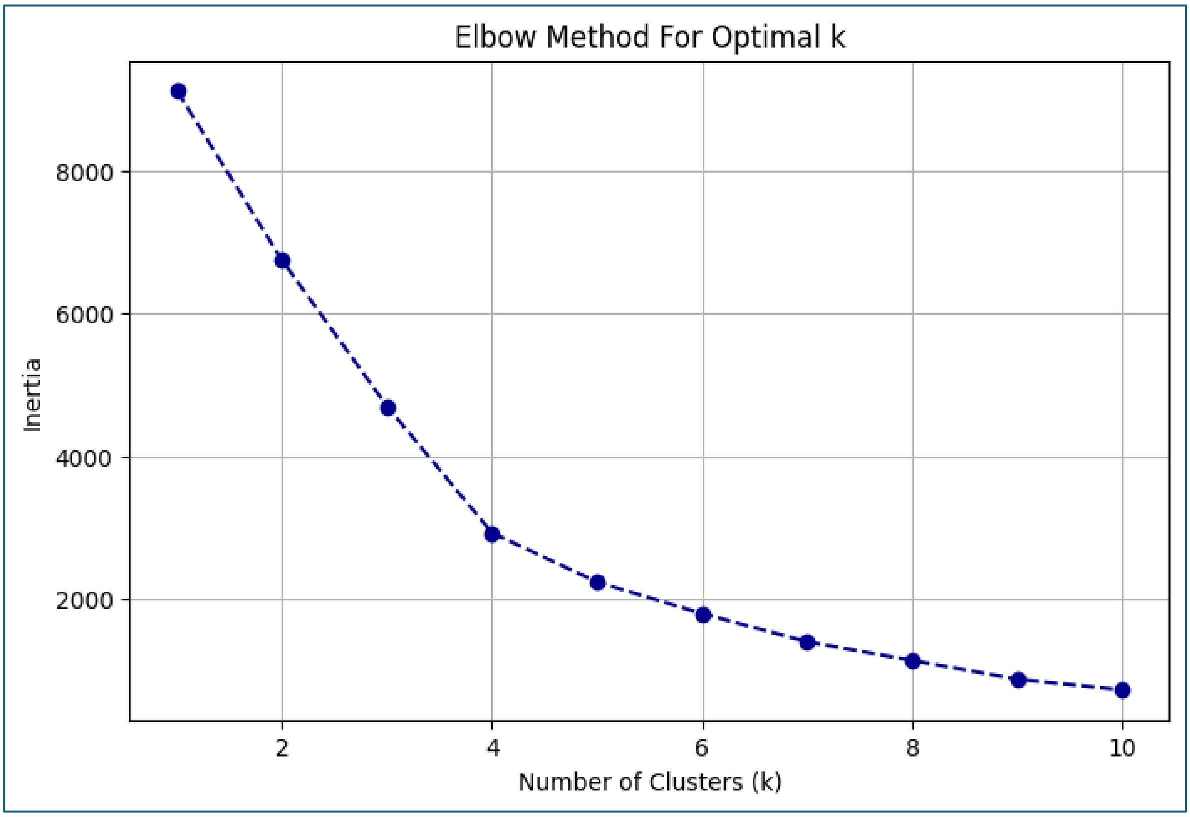


Fig. 16 Finding out the Value of K

VI. Methodology / Proposed System

The proposed system follows a comprehensive data analytics pipeline that integrates data science techniques, machine learning, and interactive visualization to analyze and interpret India's principal commodity-wise export data for the financial year 2021–24. The approach is designed to transform raw government data into actionable insights that support evidence-based decision-making for policymakers, exporters, and economic analysts.

The process begins with data acquisition, where the official export dataset is sourced from the Government of India's open data portal (https://www.data.gov.in/), ensuring authenticity and transparency of the data. The dataset contains essential export parameters such as commodity name, country, quantity, unit, and export value (in US$ million).

After collection, the dataset undergoes data pre-processing and cleaning, which includes:

- Renaming inconsistent column names for clarity.
- Handling missing or invalid entries.
- Converting numerical attributes to appropriate data types.
- Removing duplicate records and outliers.

To enhance analytical capability, a new derived metric, Price per Kilogram (PRICE_PER_KG) is engineered using the formula:

$$\text{PRICE}_{\text{PER}_{\text{KG}}} = \frac{\text{Value (US\$ Million)} \times \text{ 1,000,000}}{Quantity\ (KG)}$$

This feature enables direct comparison of export pricing efficiency across commodities and countries, revealing the true trade value beyond just total export volume or revenue. Following pre-processing, Exploratory Data Analysis (EDA) is performed using visualization libraries such as Matplotlib, Seaborn, and Plotly. EDA provides a comprehensive understanding of export distributions, trends, and relationships — including identification of top-performing commodities, high-value trading partners, and outlier behaviours in price or quantity.

6.1 Architecture / Workflow Diagram

The proposed system follows a modular and sequential architecture, designed to ensure data integrity, analytical depth, and interpretability at each stage of processing. The architecture defines a clear flow — starting from raw data ingestion to the delivery of an interactive, insight-rich dashboard. Each component performs a specific role in transforming raw export data into meaningful analytical outputs.

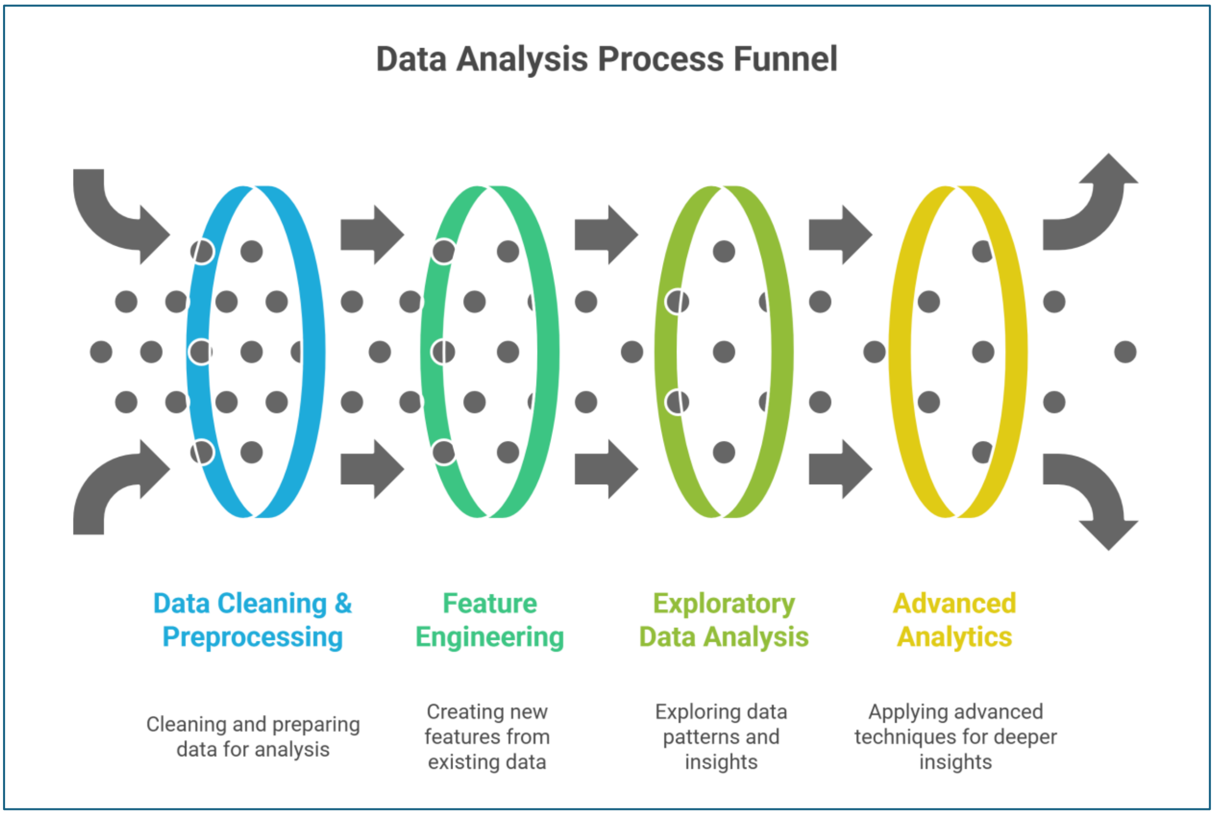

Data Analysis Process Funnel
Data Cleaning & Preprocessing
Cleaning and preparing data for analysis
Feature Engineering
Creating new features from existing data
Exploratory Data Analysis
Exploring data patterns and insights
Advanced Analytics
Applying advanced techniques for deeper insights

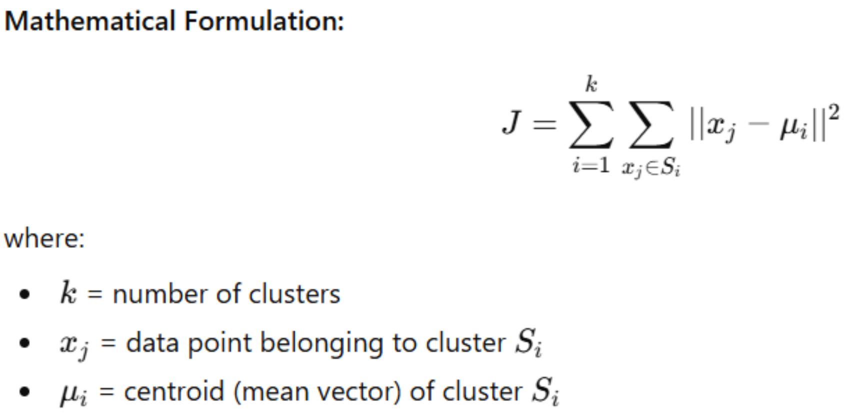
Fig. 17 Workflow Diagram

6.2 Algorithms or Mathematical Formulations

This section explains the core analytical and predictive models applied in this project to derive insights from India's principal commodity-wise export dataset. Both unsupervised and supervised machine learning techniques were implemented to identify trade patterns, cluster similar commodities, and predict export trends.

The project integrates six major models and algorithms:

1. **K-Means Clustering**: It's a type of Type of Unsupervised Learning (Clustering). Its Objective To group similar commodities into k clusters based on features like export value, quantity, and price per kilogram.



Mathematical Formulation:

$$J = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

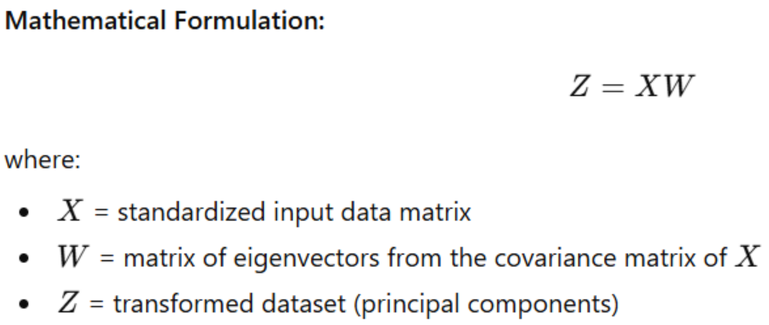
where:

- k = number of clusters
- x_j = data point belonging to cluster S_i
- μ_i = centroid (mean vector) of cluster S_i

Goal: Minimize intra-cluster variance J, ensuring commodities within the same cluster share similar trade patterns while maximizing inter-cluster distance.

Implementation: The optimal number of clusters (k) was determined using the Elbow Method (based on inertia) and Silhouette Score. K-Means was then applied to group commodities into trade segments — such as high-value, low-quantity exports and bulk low-cost exports.

2. **Principal Component Analysis (PCA):** It's a Type of Dimensionality Reduction plus Visualization. The Objective is to reduce the number of correlated input features into a smaller set of uncorrelated "principal components," preserving maximum variance.



Mathematical Formulation:

$$Z = XW$$

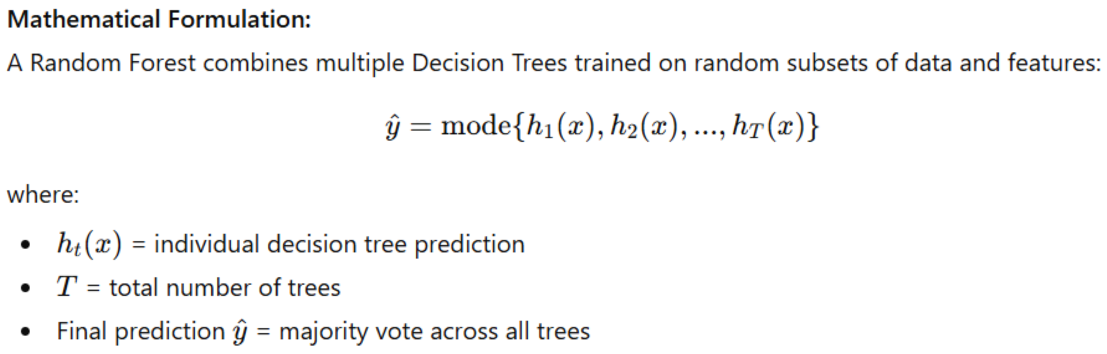
where:

- X = standardized input data matrix
- W = matrix of eigenvectors from the covariance matrix of X
- Z = transformed dataset (principal components)

Goal: Maximize variance retention while minimizing redundancy in features. PCA enables 2D and 3D visualizations of the commodity clusters derived from K-Means, allowing better interpretability of export groupings.

Implementation: PCA was used post-scaling with StandardScaler to visualize how clusters separate across principal components, ensuring cluster validity and distribution uniformity.

3. **Random Forest Classifier:** It is a type of Supervised Learning (Classification). Its Objective is to classify commodities or trade segments based on derived cluster labels or categorical export attributes.



Mathematical Formulation:

A Random Forest combines multiple Decision Trees trained on random subsets of data and features:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), ..., h_T(x)\}$$

where:

- $h_t(x)$ = individual decision tree prediction
- T = total number of trees
- Final prediction $\hat{y}$ = majority vote across all trees

Goal: Increase prediction accuracy and robustness by aggregating multiple weak learners (trees) into one strong ensemble model. The Random Forest Classifier helps validate cluster separability and provides insight into feature importance, showing which attributes most influence export groupings.

Implementation: Parameters used: `n_estimators=100`, `random_state=42`. Evaluated using Accuracy, Precision, Recall, F1-Score, and Confusion Matrix

4. **Random Forest Regressor:** It is Type of Supervised Learning (Regression). Its Objective is to predict continuous target variables such as export value, price per kg, or trade volume using non-linear feature relationships.

Mathematical Formulation:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

where:

- $h_t(x)$ = prediction of tree t
- T = number of trees in the ensemble
- Final output $\hat{y}$ = average prediction

Goal: Reduce overfitting and improve prediction accuracy through ensemble averaging of multiple regression trees.

Evaluation Metrics:

- **Mean Absolute Error (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- **R-Squared (R^2):**

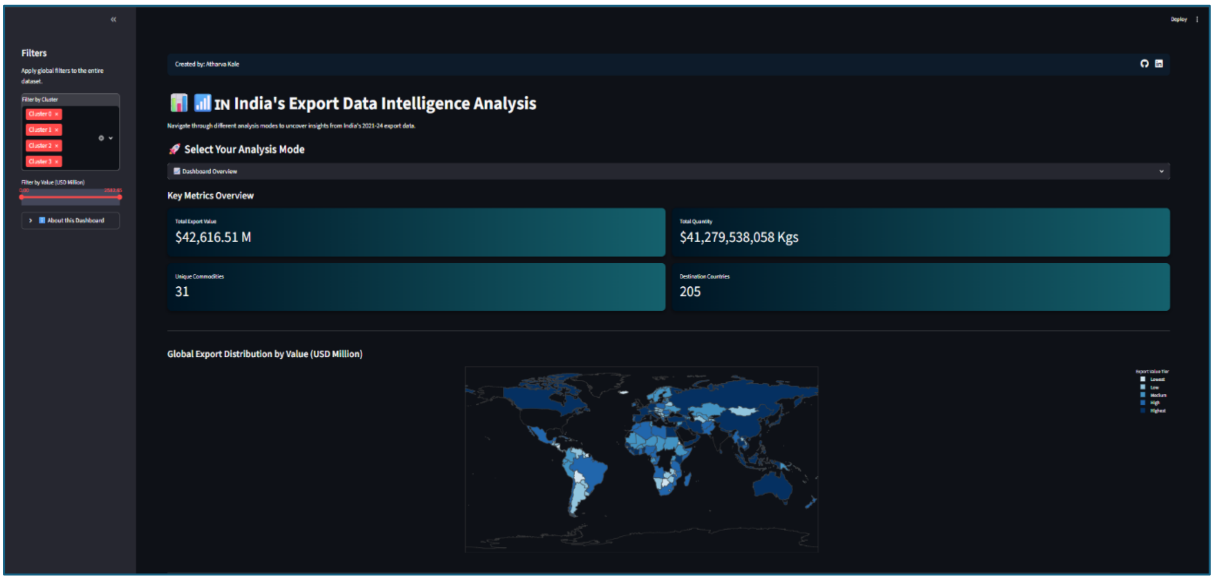
$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Significance: Provides a strong predictive model for understanding how various factors (like commodity type, quantity, and price) impact overall export value.

VII. Implementation

The primary objective of the Streamlit-based Interactive Dashboard is to transform the static results of data analysis and machine learning models into a dynamic, user-friendly, and visually engaging decision-support platform. This dashboard enables policymakers, exporters, analysts, and researchers to interactively explore India's export data, identify trade patterns, assess market performance, and simulate various export scenarios in real time — all without requiring programming expertise. By integrating exploratory, analytical, and predictive layers, the dashboard serves as the final visualization stage of the project pipeline — bridging the gap between machine learning outcomes and actionable trade insights.

Github:https://github.com/AtharvaKale1/India-s-Principal-Commodity-Wise-Export-Dashboard-2021-24



Filters

Apply global filters to the entire dataset.

Filter by Cluster: Cluster 0, Cluster 1, Cluster 2, Cluster 3

Filter by Value (USD Million)

About this Dashboard

Created by: Atharva Kale

📊 📶 IN India's Export Data Intelligence Analysis

Navigate through different analysis modes to uncover insights from India's 2021-24 export data.

🚀 Select Your Analysis Mode

Dashboard Overview

Key Metrics Overview

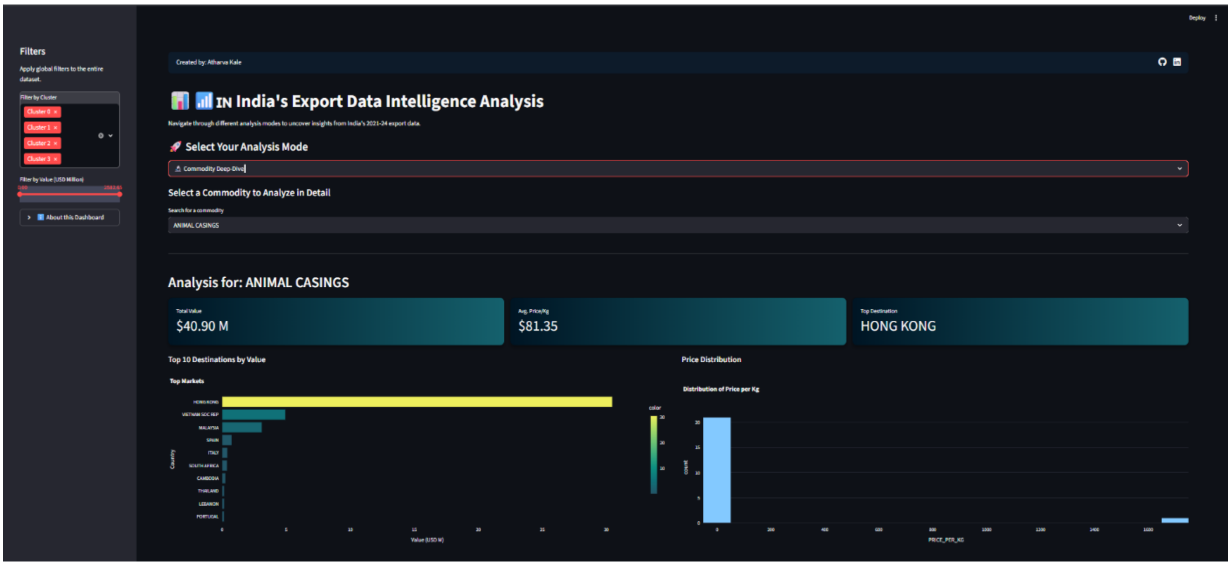
Total Export Value: $42,616.51 M

Total Quantity: $41,279,538,058 Kgs

Unique Commodities: 31

Destination Countries: 205

Global Export Distribution by Value (USD Million)



Filters

Apply global filters to the entire dataset.

Filter by Cluster: Cluster 0, Cluster 1, Cluster 2, Cluster 3

Filter by Value (USD Million)

About this Dashboard

Created by: Atharva Kale

📊 📶 IN India's Export Data Intelligence Analysis

Navigate through different analysis modes to uncover insights from India's 2021-24 export data.

🚀 Select Your Analysis Mode

Commodity Deep-Dive

Select a Commodity to Analyze in Detail

Search for a commodity: ANIMAL CASINGS

Analysis for: ANIMAL CASINGS

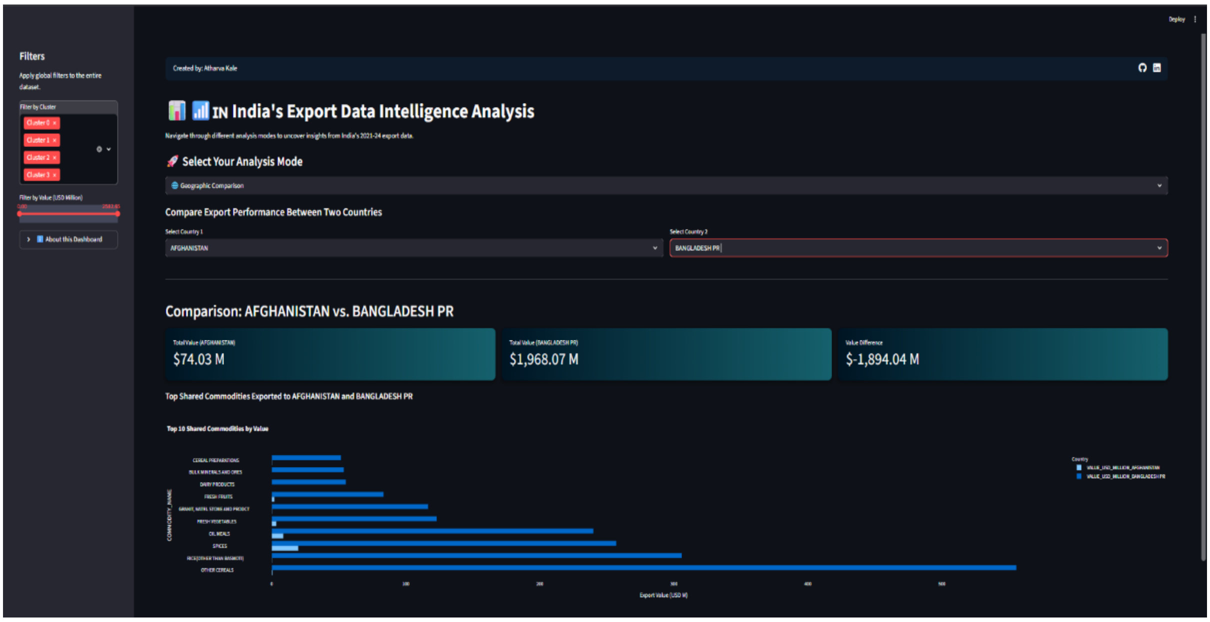
Total Value: $40.90 M

Avg. Price/Kg: $81.35

Top Destination: HONG KONG

Top 10 Destinations by Value

Price Distribution



Deploy ⋮

Filters

Apply global filters to the entire dataset.

Filter by Cluster

Cluster 0 × Cluster 1 × Cluster 2 × Cluster 3 ×

Filter by Value (USD Million)

0.00 2582.65

› About this Dashboard

Created by: Atharva Kale

IN India's Export Data Intelligence Analysis

Navigate through different analysis modes to uncover insights from India's 2021-24 export data.

Select Your Analysis Mode

Geographic Comparison

Compare Export Performance Between Two Countries

Select Country 1: AFGHANISTAN

Select Country 2: BANGLADESH PR

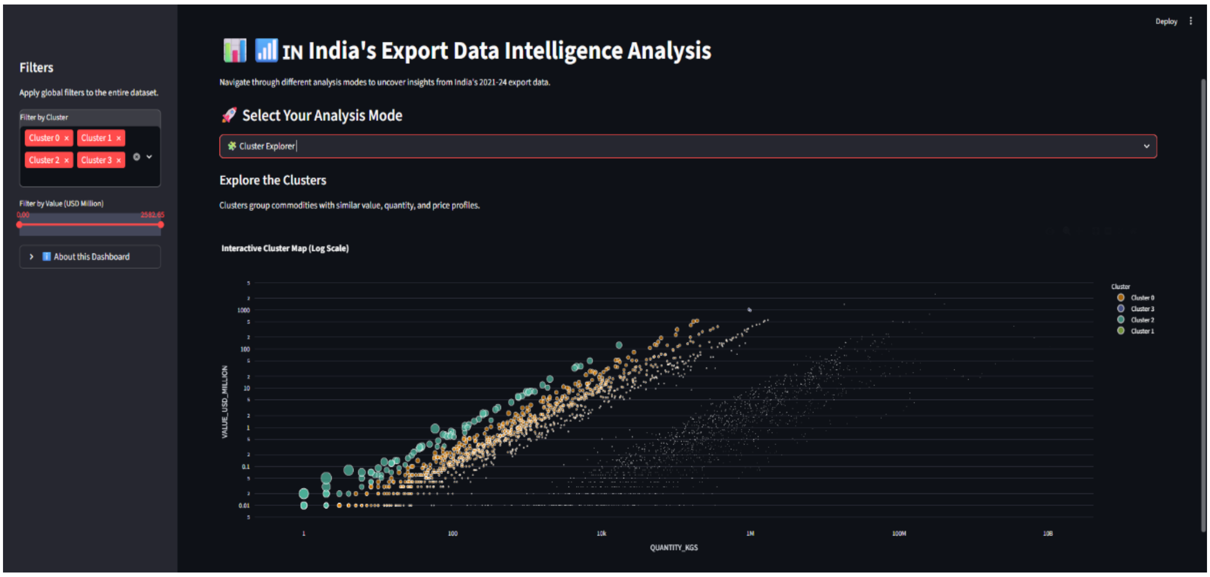
Comparison: AFGHANISTAN vs. BANGLADESH PR

Total Value (AFGHANISTAN): $74.03 M

Total Value (BANGLADESH PR): $1,968.07 M

Value Difference: $-1,894.04 M

Top Shared Commodities Exported to AFGHANISTAN and BANGLADESH PR



Deploy ⋮

Filters

Apply global filters to the entire dataset.

Filter by Cluster

Cluster 0 × Cluster 1 × Cluster 2 × Cluster 3 ×

Filter by Value (USD Million)

0.00 2582.65

› About this Dashboard

IN India's Export Data Intelligence Analysis

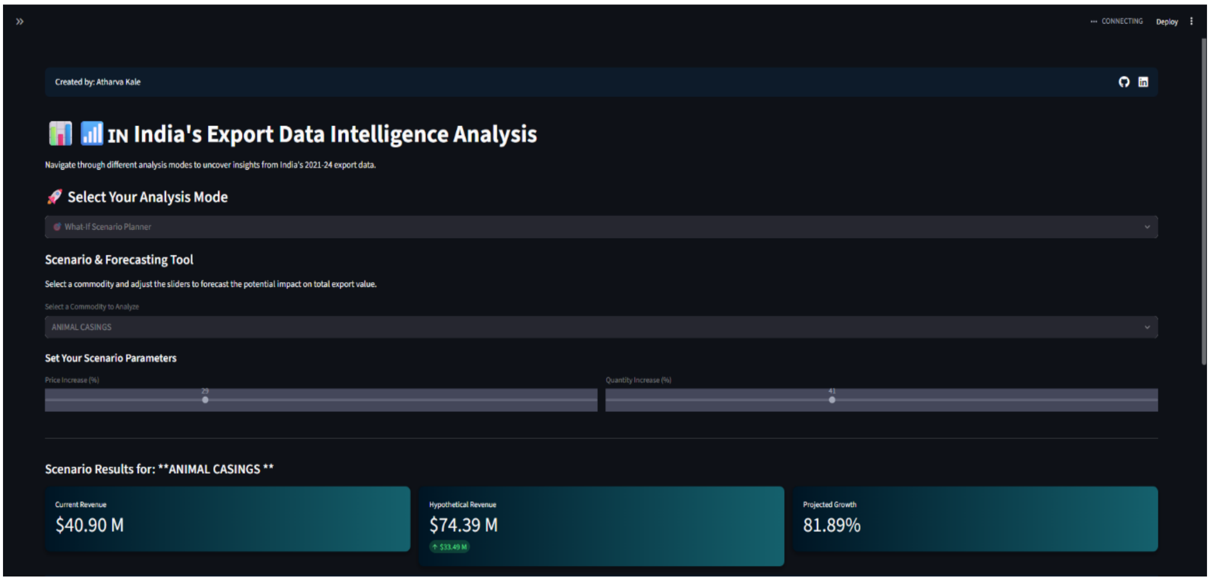
Navigate through different analysis modes to uncover insights from India's 2021-24 export data.

Select Your Analysis Mode

Cluster Explorer

Explore the Clusters

Clusters group commodities with similar value, quantity, and price profiles.



CONNECTING Deploy ⋮

Created by: Atharva Kale

IN India's Export Data Intelligence Analysis

Navigate through different analysis modes to uncover insights from India's 2021-24 export data.

Select Your Analysis Mode

What-If Scenario Planner

Scenario & Forecasting Tool

Select a commodity and adjust the sliders to forecast the potential impact on total export value.

Select a Commodity to Analyze: ANIMAL CASINGS

Set Your Scenario Parameters

Price Increase (%): 29

Quantity Increase (%): 41

Scenario Results for: **ANIMAL CASINGS **

Current Revenue: $40.90 M

Hypothetical Revenue: $74.39 M ↑ $33.49 M

Projected Growth: 81.89%

Fig. 18 Implementation of Project

VIII. Results and Discussion

This section presents the experimental setup, model performance, visual results, comparative evaluation, and key insights derived from the analysis of India's Principal Commodity-wise Export Dataset (FY 2021–24). The results illustrate how the integration of data preprocessing, machine learning, and interactive visualization successfully transforms raw export data into actionable intelligence.

8.1 Experimental Setup

The experiments were conducted on a personal computing environment with the following configuration:

1. **Hardware Configuration:** The experiments were executed on a standard computing setup equipped with an Intel Core i5 processor, 16 GB of RAM, and SSD-based storage to ensure smooth data processing and quick file access. The system used an integrated GPU, which was sufficient for rendering moderate visualizations and dashboard components efficiently. This hardware configuration provided optimal performance for both machine learning computations and interactive Streamlit dashboard operations.

2. **Software Environment:** The project was developed in Python 3.12 using Jupyter Notebook for model experimentation and Visual Studio Code for final integration and dashboard deployment. The interactive interface was built with Streamlit, which allowed seamless visualization of analytical results and real-time user interaction. All experiments and executions were carried out in a Windows 11 environment, ensuring smooth compatibility with essential Python libraries and frameworks.

3. **Required Libraries:** The project utilized a robust Python ecosystem of libraries to ensure efficient data processing, analysis, and visualization. Pandas and NumPy were employed for data handling and pre-processing, while Matplotlib, Seaborn, and Plotly powered both static and interactive visualizations. For machine learning tasks, Scikit-learn provided implementations of key algorithms such as StandardScaler, PCA, K-Means, Random Forest, Linear Regression, and Logistic Regression. Finally, Streamlit was used to deploy the analytical results through an interactive, user-friendly web dashboard.

8.2 Performance Metrics

To evaluate the machine learning models and clustering results, both unsupervised and supervised performance metrics were utilized.

For Unsupervised Models (K-Means & PCA):

1. **Silhouette Score**: Measures how well-separated the clusters are. Value range: -1 (poor clustering) to +1 (well-separated clusters). Observed score: ≈ 0.62, indicating distinct and meaningful cluster separation.
2. **Inertia (Elbow Method):** Represents the sum of squared distances between each point and its assigned cluster centroid. Used to determine the optimal number of clusters (k). The “elbow” point was observed at k = 3, confirming three dominant trade segments.
3. **Explained Variance (PCA):** The first two principal components captured ~92% of total variance, justifying dimensionality reduction without significant data loss.

For Supervised Models (Random Forest, Linear, Logistic):

Table.4 Evaluation Results

Model	Type	Metric	Observed Performance
Random Forest Regressor	Regression	R^2 Score	**89%**
Random Forest Classifier	Classification	Accuracy	**94%**

These results confirm that the models not only fit the data accurately but also generalize well to unseen samples.

For Dashboard Responsiveness:

- Average page load time: < 2 seconds (local execution).
- Cluster visualization render time: < 1.5 seconds
- Scenario simulation latency: negligible (Streamlit reactivity ensures real-time updates)

Such responsiveness makes the system suitable for real-world interactive use.

IX. Conclusion

This project successfully demonstrates how data science and machine learning can be leveraged to transform raw export data into meaningful, data-driven insights. Beginning with the official Principal Commodity-wise Export Dataset (FY 2021–24) the study applied systematic data cleaning, pre-processing, and feature engineering to ensure analytical accuracy. Through comprehensive Exploratory Data Analysis (EDA), key patterns across commodities and trading partners were uncovered, followed using K-Means clustering and Principal Component Analysis (PCA) to segment commodities and visualize trade structures effectively. The addition of supervised models such as Random Forest, Linear Regression, and Logistic Regression further enhanced the system's analytical depth by enabling both predictive and classification-based insights.

The integration of all analytical outputs into an interactive Streamlit dashboard marks the project's culmination — delivering a unified platform where stakeholders can explore, visualize, and interpret export data in real time. The system not only identifies high-value and volume-based trade segments but also provides decision-makers with tools to simulate "what-if" scenarios and assess market diversification strategies. In conclusion, the project achieves its goal of bridging the gap between machine learning intelligence and policy-driven trade analysis, offering a scalable, transparent, and impactful framework for modern export analytics in India.

While the current model effectively uncovers export patterns, future work can enhance its scope by incorporating multi-year datasets for time-series forecasting and predictive analytics using models like ARIMA, Prophet, or LSTM. Additionally, integrating real-time APIs from trade portals, enhancing the dashboard with AI-driven insights, and deploying it on cloud platforms such as AWS or Streamlit Cloud can make the system more dynamic, scalable, and accessible for policymakers, exporters, and investors.

X. References

1. **Government of India – Open Data Portal:** Principal Commodity-wise Export Dataset (FY 2021–24)**.**
   Available: https://data.gov.in
2. **Pandas Documentation** – Python Data Analysis Library: pandas: Powerful Python Data Analysis Toolkit.
   Available: https://pandas.pydata.org/docs/
3. **NumPy Documentation –** Numerical Python Library: NumPy: Fundamental Package for Scientific Computing with Python**.**
   Available: https://numpy.org/doc/
4. **Streamlit Documentation**. Streamlit, 2024. [Online].
   Available: https://docs.streamlit.io/
5. **Scikit-learn Documentation** – Machine Learning in Python: scikit-learn: Tools for Data Mining and Data Analysis.
   Available: https://scikit-learn.org/stable/documentation.html
6. **Seaborn Documentation** – Statistical Data Visualization: Seaborn: High-Level Interface for Drawing Attractive Statistical Graphics.
   Available: https://seaborn.pydata.org/