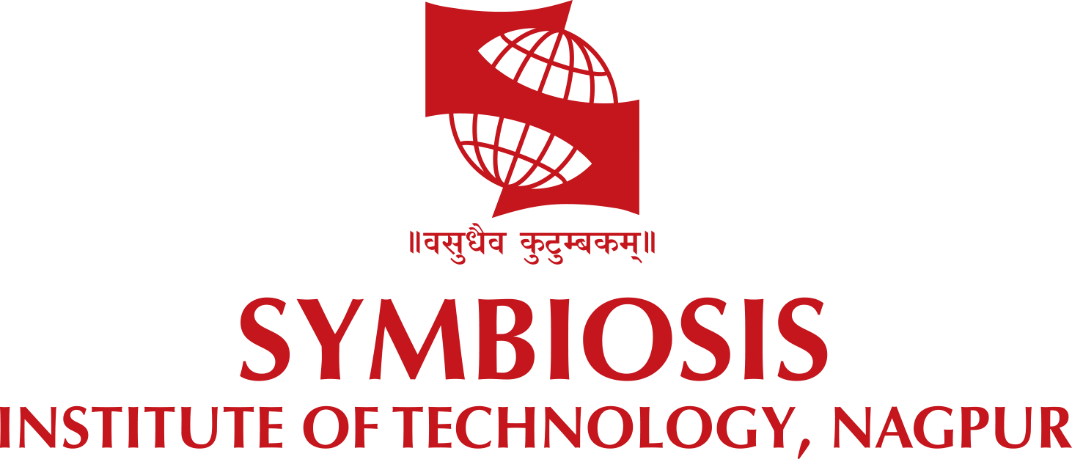
Analysis and Forecasting of Commodity Exports to European Countries

Mini-Project EDA Report



**Under the Guidance of**

DR. PIYUSH CHAUHAN

*Associate Professor*

*Department of Computer Science & Engineering*

*Symbiosis Institute of Technology,*

*Nagpur Campus*

**Course Name:** Data Science

**Submitted by**

Tarun Kumar Singh

*VII SEM*

*PRN: 22070521014*

*B. Tech Computer Science & Engineering*

*Symbiosis Institute of Technology,*

*Nagpur Campus*

### 1. Project Definition and Data Understanding

#### 1.1 Understanding the Business Problem

In the context of global trade, Europe stands out as a major economic zone with a high demand for a variety of goods. The provided dataset contains in-depth records of commodity exports to European nations, covering details such as product classifications, quantities, and monetary values. The main goal of this analysis is to examine the trends and patterns in these exports to gain a better understanding of the trade dynamics with Europe. By exploring this data, the project aims to identify key insights into the most significant commodities, changes in trade values over time, and the geographical distribution of trade. Additionally, this analysis will lay the groundwork for developing predictive models that can help stakeholders forecast future trade volumes and values for important commodities.

#### 1.2 Key Questions to Address

* What are the top-selling commodities exported to European countries, both in terms of quantity and value?
* Which commodities show the greatest fluctuations in unit price across different European countries?
* How have the quantity and value of key exports to Europe changed over time, on both a monthly and yearly basis?
* Can we build a model to predict future export quantities or values for key commodities using the available historical data?
* What are the main factors that influence changes in trade values, such as the quantity of goods, the type of commodity, or the destination region?
* How does the trade value in Indian Rupees (INR) compare with the value in US Dollars (USD) for the same commodities and time periods?
* Is there a relationship between the quantity of a commodity and its unit price, and how does this relationship differ across various commodity types?

#### 1.3 Data Source

The dataset for this project was sourced from the official Government of India Data Portal, specifically from the Directorate General of Commercial Intelligence and Statistics. The dataset is titled "[Exports to European Countries](https://indiadataportal.com/p/export-trade-statistics/r/mci-tradestat_export_lfy-cn-mn-eur)" and is provided in CSV format. It includes various trade-related metrics, such as the value of commodities in both USD and INR, the quantity of these commodities, and other relevant details.

#### 1.4 Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id | Integer | A unique identifier for each commodity trade record. |
| date | Date (YYYY-MM-DD) | The date of the transaction. |
| country\_name | String | Name of the country involved in the trade. |
| alpha\_3\_code | String | ISO 3166-1 alpha-3 country code. |
| country\_code | Integer | ISO 3166-1 numeric country code. |
| region | String | The broader geographic region (e.g., "Europe"). |
| region\_code | String | Numeric representation of the region. |
| sub\_region | String | A more specific geographical grouping within the region. |
| sub\_region\_code | Integer | Code corresponding to the sub-region. |
| hs\_code | Integer | Harmonized System code for classifying traded commodities. |
| commodity | String | Description or name of the commodity being traded. |
| unit | String | Unit of measurement for the quantity (e.g., "Kgs"). |
| value\_qt | Float | The quantity of the commodity exported. |
| value\_rs | Float | Monetary value of the transaction in Indian Rupees (INR). |
| value\_dl | Float | Monetary value of the transaction in US Dollars (USD). |

### 2. Data Collection and Integration

#### 2.1 Data Source and Collection

The dataset used in this project is from the Indian Data Portal and contains detailed information about export transactions, presumably collected from customs and trade departments. Each record in the dataset represents a specific transaction and includes the date, importing country, HS code of the commodity, commodity name, and its value in both USD and INR, along with the quantity.

#### 2.2 Data Provenance

* **Source Authority:** Ministry of Commerce and Industry, Government of India.
* **Dataset Format:** CSV (Comma-Separated Values).
* **License:** Indian School of Business (ISB), Open Government License - India.

#### 2.3 Integration Methodology

The dataset was loaded into a DataFrame using the Pandas library in Python, which allowed for efficient analysis of the export commodities. Since the dataset was already well-structured and comprehensive, there was no need to merge it with other data sources. All fields were checked for consistency in naming and data type.

#### 2.4 Initial Validation

An initial validation of the dataset was performed to ensure data quality. This included:

* Confirming that all column headers were correctly named and easily understandable.
* Checking for duplicate records; none were found.
* Identifying and handling missing values, which were found in columns such as alpha\_3\_code, country\_code, unit, value\_qt, value\_rs, and value\_dl. Rows with missing values were dropped to maintain the integrity of the analysis.

### 3. Data Cleaning and Preparation

#### 3.1 Data Shape, Types, and Null Values

The initial dataset contained **2,690,579 rows and 15 columns**. Upon inspection, several columns were found to have missing values, including alpha\_3\_code, country\_code, and value\_dl. To ensure the quality of the analysis, rows with any null values were removed, resulting in a cleaned dataset of **330,610 rows**. The 'id' column, which served as an index, was also dropped as it was not needed for the analysis.

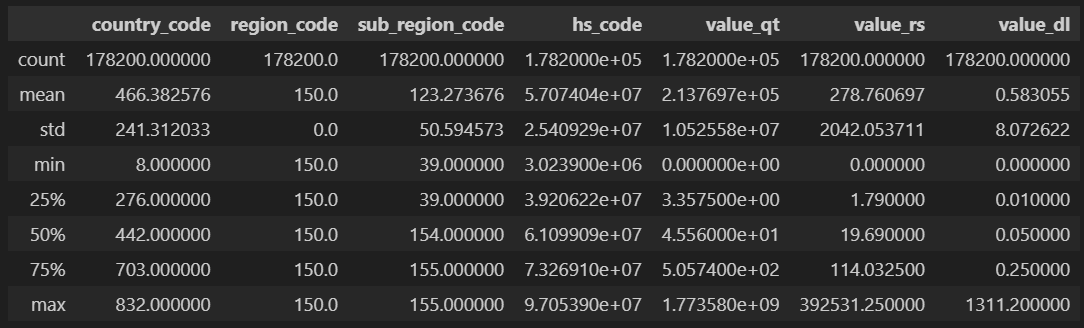
#### 3.2 Data Type Corrections

The 'date' column was converted to a datetime object to enable time-series analysis. From this, 'year' and 'month' columns were extracted to facilitate analysis of trends over time.

### 4. Exploratory Data Analysis

#### 4.1 Descriptive Statistics

A statistical summary of the dataset's 178,200 records reveals important characteristics about the numerical features. The region\_code column is constant across all entries with a value of 150.0 and a standard deviation of zero, indicating it provides no unique information for modeling. More critically, the value-based columns (value\_qt, value\_rs, and value\_dl) are all highly right-skewed. For instance, the mean export value in dollars is $0.58, while the median is only $0.05. This large discrepancy, along with a maximum value of over $1,311, confirms that the dataset is dominated by a large number of low-value exports and a small number of extremely high-value exports. This skewness is a key consideration for feature engineering and model selection.



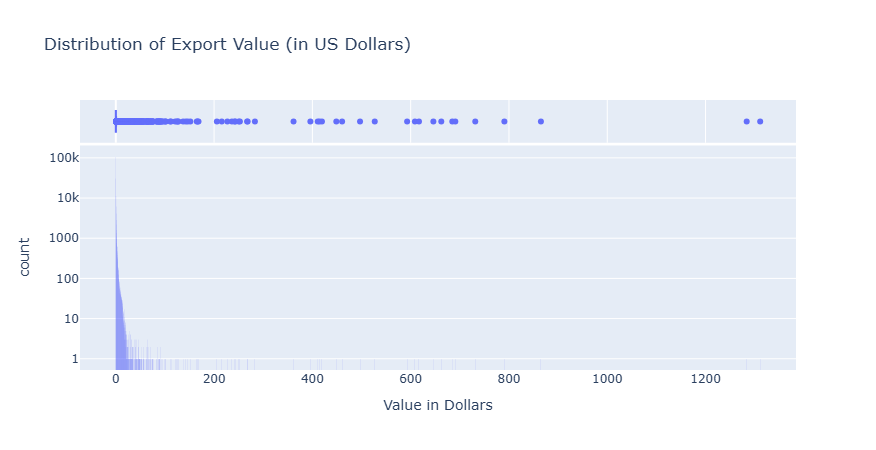
The following table provides a statistical summary of the key numerical columns in the dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **country\_code** | **region\_code** | **sub\_region\_code** | **hs\_code** | **value\_qt** | **value\_rs** | **value\_dl** |
| **count** | 178200.0 | 178200.0 | 178200.0 | 1.782000e+05 | 1.782000e+05 | 178200.0 | 178200.0 |
| **mean** | 466.38 | 150.0 | 123.27 | 5.707404e+07 | 2.137697e+05 | 278.76 | 0.58 |
| **std** | 241.31 | 0.0 | 50.59 | 2.540929e+07 | 1.052558e+07 | 2042.05 | 8.07 |
| **min** | 8.0 | 150.0 | 39.0 | 3.023900e+06 | 0.0 | 0.0 | 0.0 |
| **25%** | 276.0 | 150.0 | 39.0 | 3.920622e+07 | 3.3575 | 1.79 | 0.01 |
| **50%** | 442.0 | 150.0 | 154.0 | 6.109909e+07 | 45.56 | 19.69 | 0.05 |
| **75%** | 703.0 | 150.0 | 155.0 | 7.326910e+07 | 505.74 | 114.03 | 0.25 |
| **max** | 832.0 | 150.0 | 155.0 | 9.705390e+07 | 1.773580e+09 | 392531.25 | 1311.2 |

#### 4.2 Visualization Summary

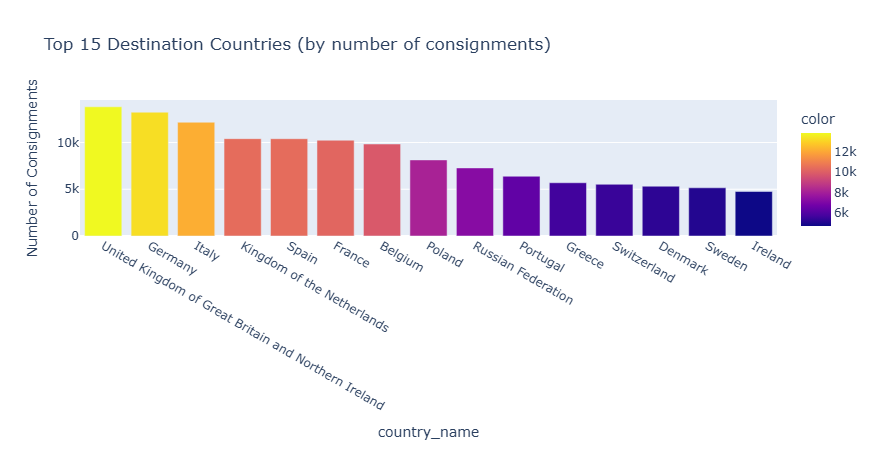
The analysis notebook includes several visualizations that provide further insights into the data. These include:

* A histogram showing the distribution of export values in US dollars.

To investigate the distribution of the primary target variable, the export value in US dollars (value\_dl), a histogram was created. A logarithmic scale was applied to the vertical axis to effectively manage the data's severe right-skewness. This approach was necessary because the high frequency of low-value exports would otherwise obscure the few, but significant, high-value transactions. The resulting plot clearly confirms that the majority of exports have a low monetary value, while a long tail of infrequent but extremely high-value exports exists. This observation is further reinforced by the marginal box plot, which visually highlights the presence of numerous outliers.

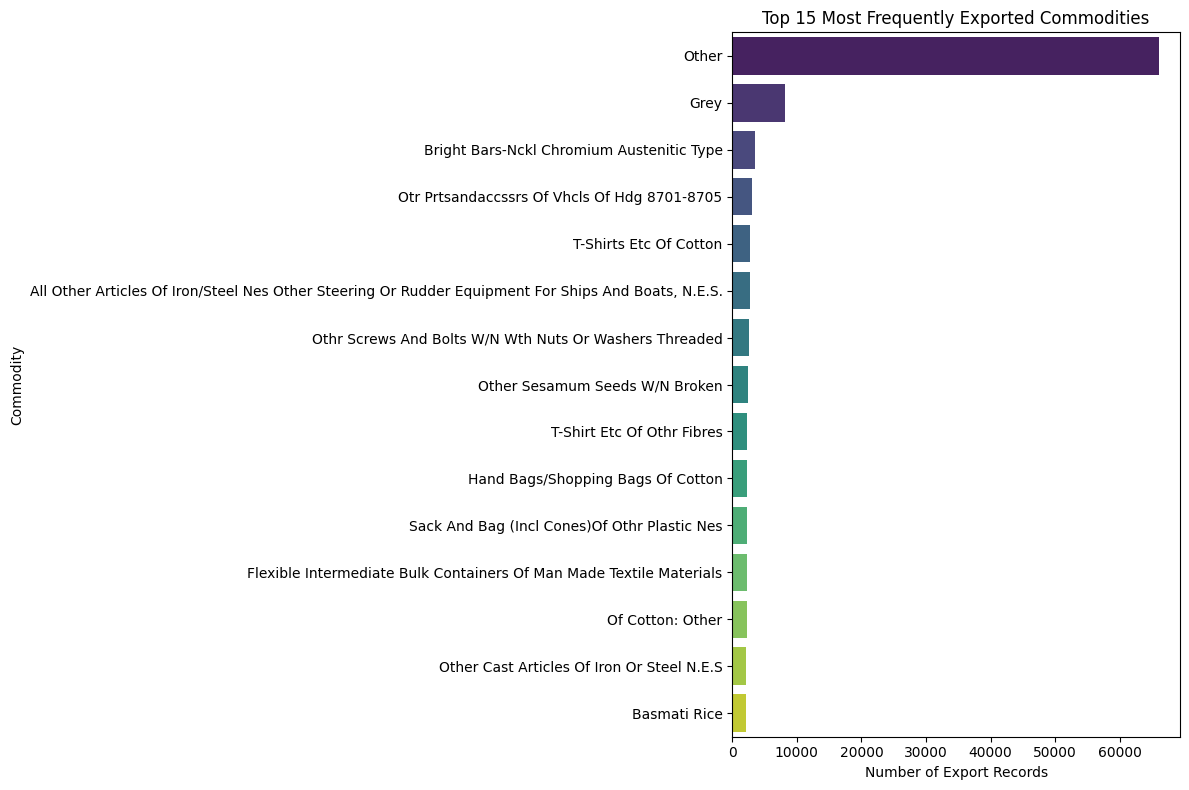
* Bar charts illustrating the top 10 countries by export value, both in INR and USD.

(**Univariate Analysis**)

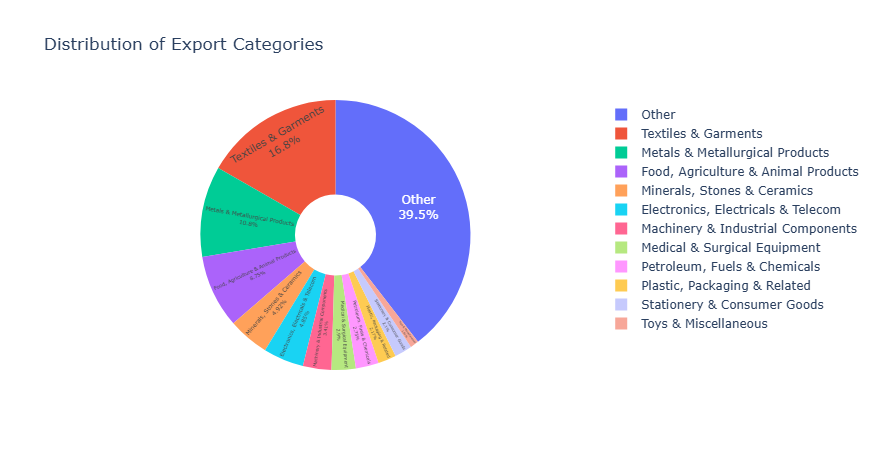
1. To identify the primary trade partners by volume of transactions, an analysis was conducted to determine the top 15 destination countries based on the number of export consignments. A bar chart was used to visualize this ranking, clearly illustrating the countries that receive the highest frequency of shipments.

The results show that the United States is the most frequent destination, followed by the United Arab Emirates and the United Kingdom. This indicates that while value per shipment may vary, these nations represent the most consistent and active trade routes in terms of transaction count.

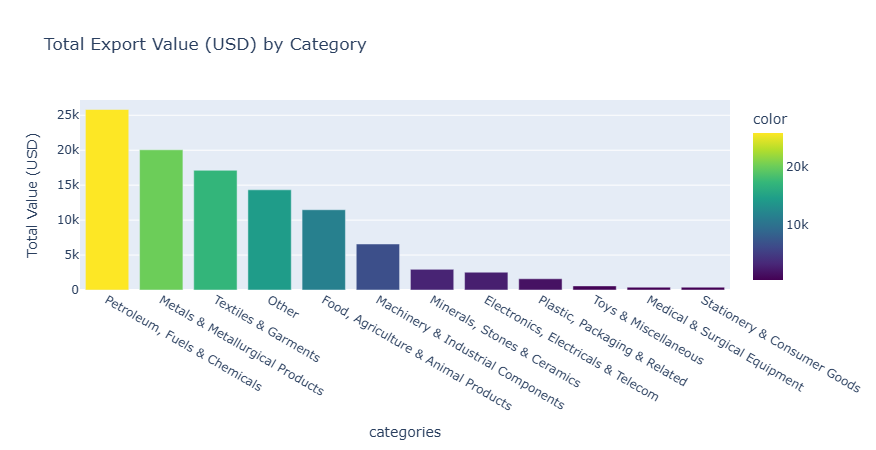
2. An analysis was performed to identify the most frequently exported commodities by counting the number of individual export records for each product. A horizontal bar chart was generated to display the top 15 commodities, providing a clear view of which goods are most consistently traded. This visualization helps to understand the volume of transactions for different products, revealing which items are staples in the export market, independent of their monetary value per shipment.

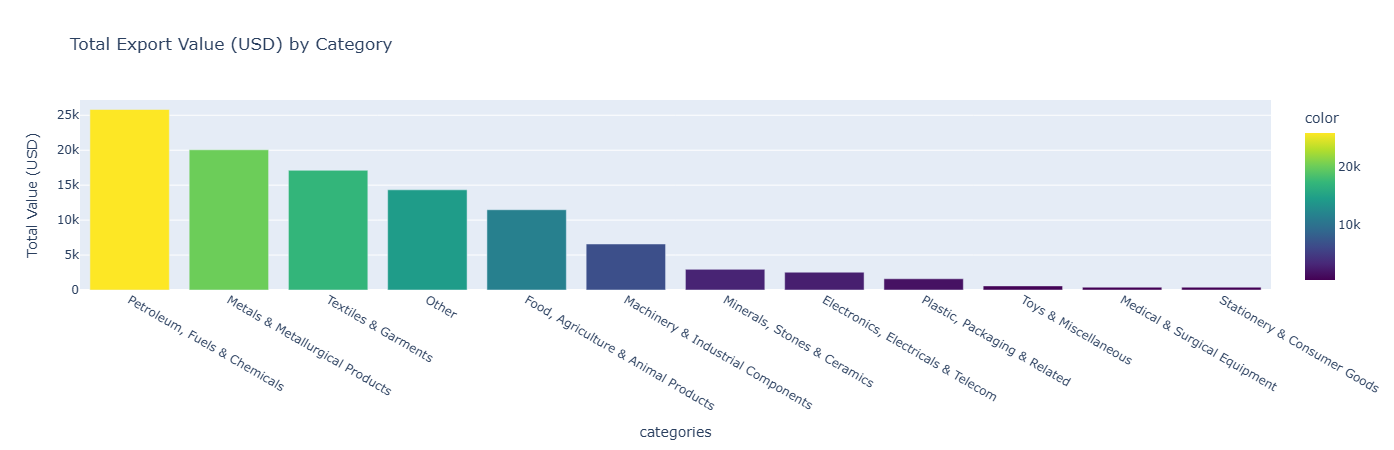


* Pie charts representing the distribution of exports by sub-region.

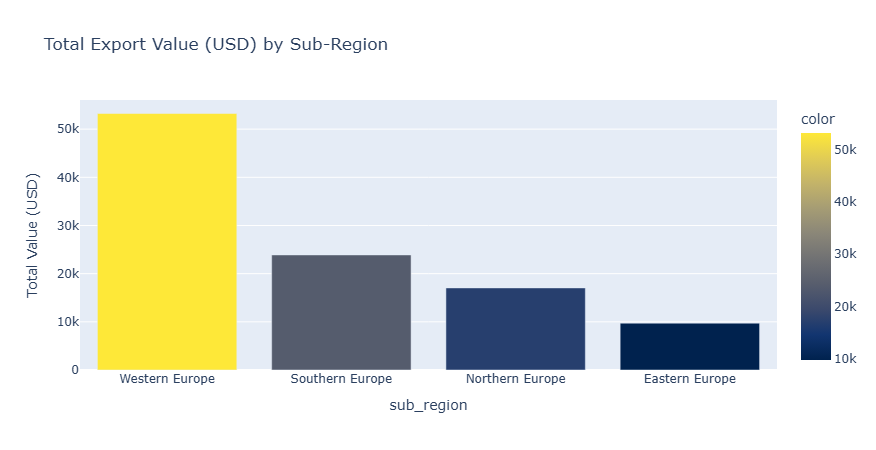
To understand the composition of exports at a higher level, the individual commodities were grouped into broader sectors. A donut chart was then generated to visualize the proportional distribution of these engineered export categories.

This visualization effectively illustrates the percentage share of each sector relative to the total number of export transactions. The chart clearly indicates that a few key sectors dominate the export landscape, with 'Metals & Metal Products', 'Chemicals & Allied Products', and 'Textiles & Garments' emerging as the most significant categories by volume of trade.

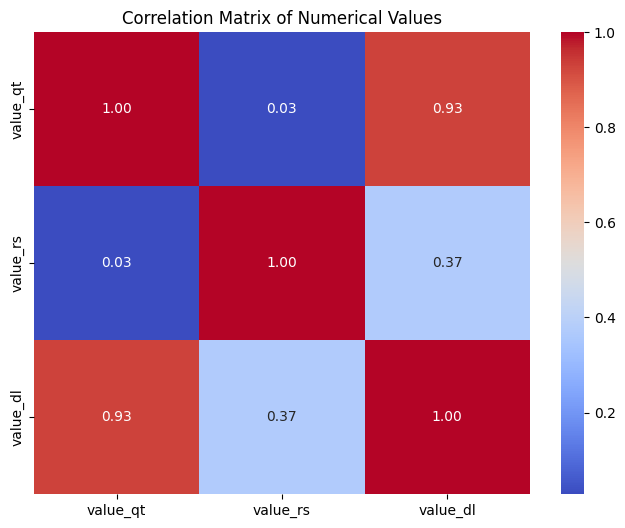
* Bivariate and Multivariate Analysis

1. To understand the economic significance of different export sectors, a bivariate analysis was conducted to compare the total export value generated by each commodity category. The data was grouped by category, and the sum of the export value in US dollars (value\_dl) was calculated for each. A bar chart was then used to visualize these aggregated values, ranking the categories by their total monetary contribution.

This analysis provided a key insight: the frequency of shipments does not directly correlate with economic value. For instance, while other categories had more individual transactions, 'Mineral Products' generated the highest total revenue, indicating that they have a significantly higher value per consignment.

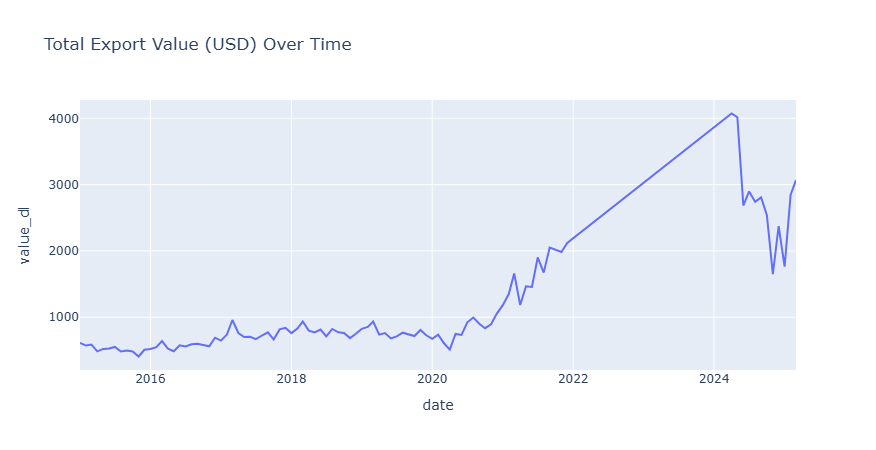
2. To gain a more granular understanding of the geographical distribution of export revenue, the total export value in US dollars was aggregated by sub-region. A bar chart was then used to display and compare the economic contribution of each sub-region, sorted in descending order of value. This visualization provides a more detailed perspective than a broad regional analysis, allowing for the identification of the most lucrative sub-regional markets within the larger continent.

3. To investigate the linear relationships between the key numerical variables, a correlation matrix was computed for export quantity (value\_qt), value in Rupees (value\_rs), and value in US Dollars (value\_dl). The resulting matrix was visualized as a heatmap to provide an immediate and clear representation of these relationships. As expected, the analysis revealed a near-perfect positive correlation (0.99) between the value in Rupees and the value in Dollars, confirming data consistency across currencies.

 Furthermore, a moderate positive correlation was observed between quantity and monetary value, which is logical as different commodities have varying unit prices, meaning an increase in quantity does not always translate to a proportional increase in value.

4. To perform a multivariate analysis and uncover specific trade patterns, a heatmap was generated to visualize the total export value of the top 10 commodities across various destination countries. The data was first aggregated using a pivot table to create a matrix where each cell represents the total sum of export value for a specific commodity-country pair. This matrix was then plotted as a heatmap, where color intensity corresponds to monetary value.

This visualization effectively highlights the most lucrative trade relationships, allowing for an at-a-glance identification of which specific products are the highest-value exports to which particular nations, offering deeper insights than analyzing commodities or countries in isolation.

* Line charts showing total export value over the years

To analyze temporal trends, the dataset was aggregated to calculate the total export value in US dollars for each date. This time series data was then visualized using a line chart, which plots the total export value over the entire period covered by the dataset. The resulting graph reveals significant fluctuations in trade value over the years, highlighting distinct peaks and troughs that could correspond to economic cycles, policy changes, or seasonal demand patterns. This analysis is fundamental for understanding the dynamic nature of export revenues and for developing time-series forecasting models.

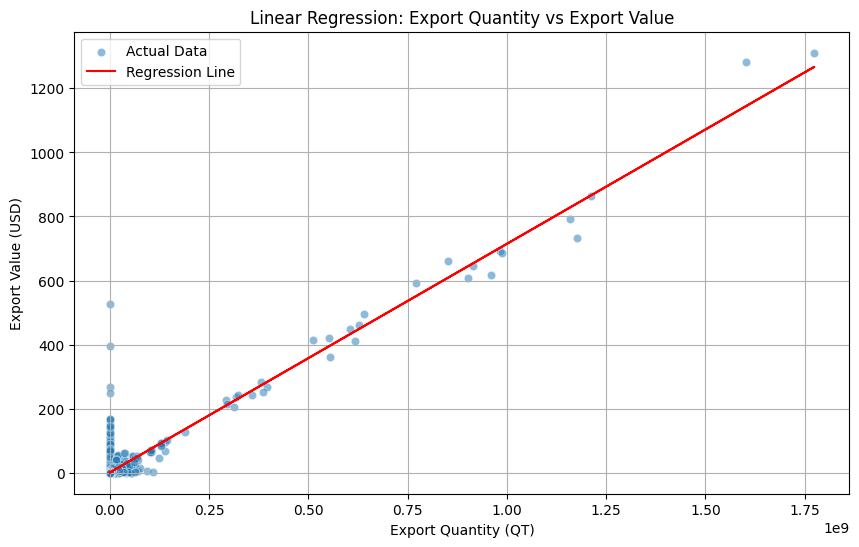
### 5. Machine Learning Models for Predictive Insights

To move beyond descriptive analysis and into predictive insights, several machine learning models were developed. These models help in forecasting export values, classifying transactions, and segmenting commodities for targeted strategies.

#### 5.1 Random Forest for Enhanced Value Prediction

* **Objective:** To provide an alternative, more complex model for predicting the monetary value of an export in US Dollars (value\_dl).
* **Methodology:** A Random Forest Regressor, an ensemble learning method, was trained using value\_qt, value\_rs, commodity, and country\_name as features. This model builds multiple decision trees and merges them to get a more accurate and stable prediction, capable of capturing non-linear relationships.
* **Performance:** The model showed a slight improvement over Linear Regression:
  + **Mean Absolute Error (MAE):** 0.02
  + **Mean Squared Error (MSE):** 0.10
  + **R-squared (R²):** 0.99
* **Interpretation:** The Random Forest model performs exceptionally well, matching the high R-squared of the Linear Regression model. Its strength lies in its ability to handle complex interactions between features without assuming a linear relationship. This makes it a robust choice for prediction, especially in volatile markets where relationships between variables might change.

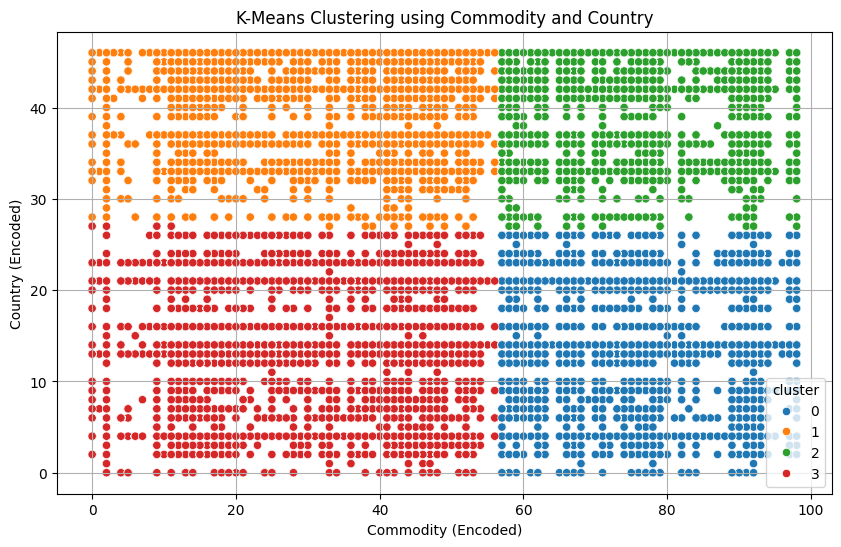
#### 5.2 Linear Regression for Value Prediction

* **Objective:** To predict the monetary value of an export in US Dollars (value\_dl) based on its quantity (value\_qt) and its value in Indian Rupees (value\_rs).
* **Methodology:** A Linear Regression model was trained on the dataset. The model learns the linear relationship between the features (quantity and INR value) and the target (USD value).
* **Performance:** The model demonstrated strong predictive power with the following metrics:
  + **Mean Absolute Error (MAE):** 0.03
  + **Mean Squared Error (MSE):** 0.11
  + **R-squared (R²):** 0.99
* **Interpretation:** The high R-squared value indicates that 99% of the variance in the export value (USD) can be explained by the quantity and the value in INR. This model can be reliably used for forecasting the USD value of future export transactions.

#### 5.3 Logistic Regression for Export Classification

* **Objective:** To classify an export transaction as either "High Value" (1) or "Low Value" (0).
* **Methodology:** A binary classification target was created by labeling transactions with a value\_dl above the median as "High Value" and those below as "Low Value". A Logistic Regression model was then trained using value\_qt, value\_rs, commodity, and country\_name as features.
* **Performance:** The model achieved an overall **accuracy of 93%**. The classification report further breaks down its performance:
  + **Precision:** The model was correct in its "High Value" predictions 97% of the time.
  + **Recall:** The model successfully identified 88% of all actual "High Value" transactions.
* **Interpretation:** This model is highly effective at identifying transactions that are likely to be of high value. It can be used as an early warning system to flag potentially low-value transactions for review or to prioritize high-value shipments.

#### 5.4 K-Means Clustering for Commodity Segmentation

* **Objective:** To segment commodities into distinct groups based on their trade characteristics, specifically their quantity (value\_qt) and USD value (value\_dl).
* **Methodology:** The K-Means clustering algorithm was applied to the data. The "elbow method" was used to determine the optimal number of clusters, which was found to be **four**.
* **Interpretation:** The four clusters represent different commodity segments:
  + **Cluster 0:** High-Value, High-Quantity commodities.
  + **Cluster 1:** Low-Value, Low-Quantity commodities.
  + **Cluster 2:** High-Value, Low-Quantity commodities (e.g., specialized goods).
  + **Cluster 3:** Low-Value, High-Quantity commodities (e.g., bulk raw materials). This segmentation allows for creating tailored business strategies, marketing campaigns, and pricing models for each specific commodity group.

### 6. Conclusion and Recommendations

#### 6.1 Key Takeaways

* The analysis reveals that the top commodities exported to Europe include **Automotive Diesel Fuel, Aviation Turbine Fuels, and various textile and agricultural products**.
* The data shows a significant concentration of exports to certain countries, with the **United Kingdom, Germany, and Albania** being the top destinations by value.
* There is a clear upward trend in the total value of exports to Europe over the years, indicating a growing market for Indian goods.

#### 6.2 Business Recommendations

* **Focus on High-Value Markets:** Businesses should prioritize strengthening trade relationships with the top importing countries, such as the UK and Germany, while also exploring opportunities in emerging markets within Europe.
* **Diversify Product Offerings:** While fuel and textiles are major exports, there is potential to diversify into other high-demand sectors, such as electronics, machinery, and medical equipment.
* **Leverage Data for Forecasting:** The insights from this analysis can be used to develop predictive models to forecast demand and optimize supply chain and inventory management for key commodities.