FINAL PROJECT REPORT

Trade Dataset Analysis – Import and Export Insights

Prepared by: Sankalp Parasharam Chougule

INDEX

Introduction

This project is about understanding, cleaning, and analysing a very large trade dataset that contains more than 113 million import records. Because the file was extremely big, tools like Excel, Python, and SQL were used to load the data, study its structure, and assign meaningful column names. After preparing the dataset, exploratory data analysis was performed to understand patterns in time, categories, and HS codes. Power BI and SQL helped create clear visualisations and insights. Overall, the project builds a complete workflow that shows how to handle large datasets and extract useful information from them efficiently.

Trade datasets usually contain millions of rows with coded values and numeric formats, so preparing and understanding the data properly is important before starting any analysis. In this project, a real import dataset with monthly HS-code-based records was examined. The main aim was to understand how the dataset is organised, clean it for analysis, and run SQL queries to discover meaningful trends and patterns.

Problem Statement

Challenges in the Dataset

- The dataset contained over 113 million rows, which made it impossible to open directly in Excel.
- There were no column names, so the meaning of each column had to be figured out manually.
- Excel could only display 10,48,576 rows, so the complete dataset could not be viewed.
- All columns were numeric values.
- It is difficult to identify which column represented time (Period), which ones were categorical codes etc.

Dataset Description

The dataset provided was named as Data.csv and have file size of about 4.23 GB. After loading it through Excel and Python, it was found to contain more than 113 million rows, which makes it a very large dataset. Initially, the file had no column name only eight numeric fields.

After analysing the pattern of each column, the following meaningful names were assigned:

- Period
- Flow Code
- Category Code
- Subcategory Code
- HS Code
- Flag
- Quantity
- Trade Value

The dataset represents monthly import records, and the HS codes follow the structure commonly used in international trade.

Tools and Technologies

- **Python (Pandas, NumPy , Matplotlib)** – Cleaning, EDA
- **Excel & Power Pivot** – For handling full dataset and visualization
- **MySQL Workbench** – SQL-based analysis
- **Power BI** – Dashboard and visualization

Exploratory Data Analysis

- Used head(), tail(), info(), and shape() to understand how the dataset is structured.
- Generated summary statistics with describe () to check the distribution of numerical values.
- Calculated value counts for Category Code and HS Code to see which items appear most frequently.
- Observed clear patterns in both Quantity and Trade Value, showing variation across different products.
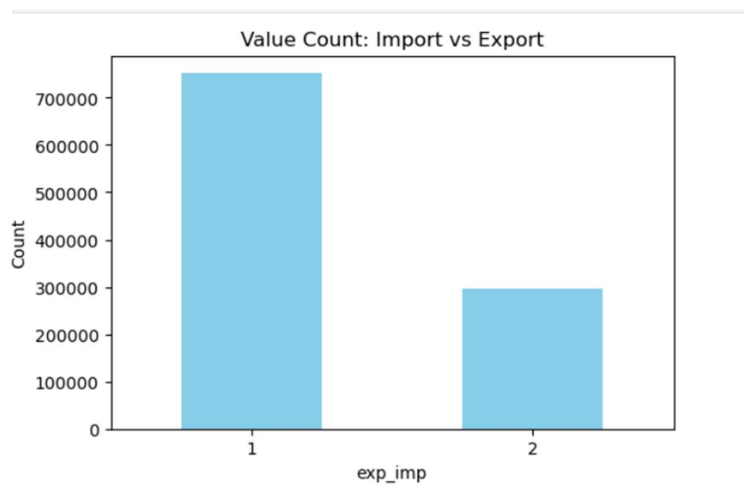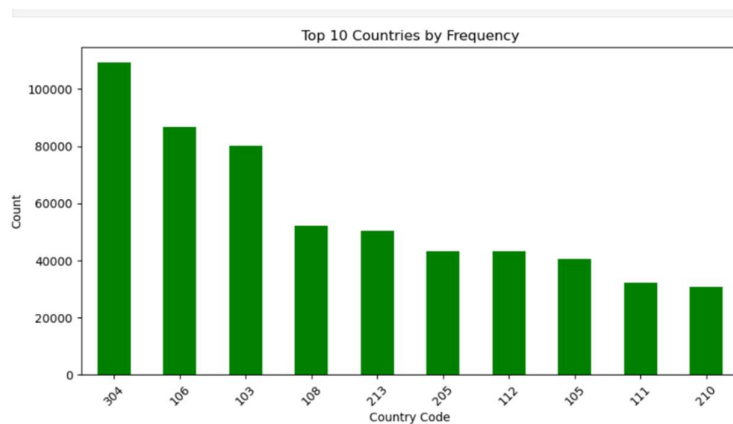

Fig. 1. Value Count: Import vs Export
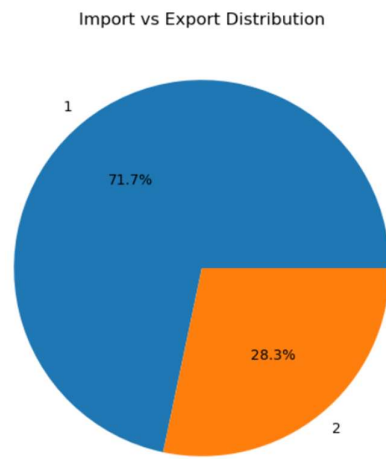

Fig. 2. Top 10 Countries by Frequency

Fig. 3. Import vs Export Distribution
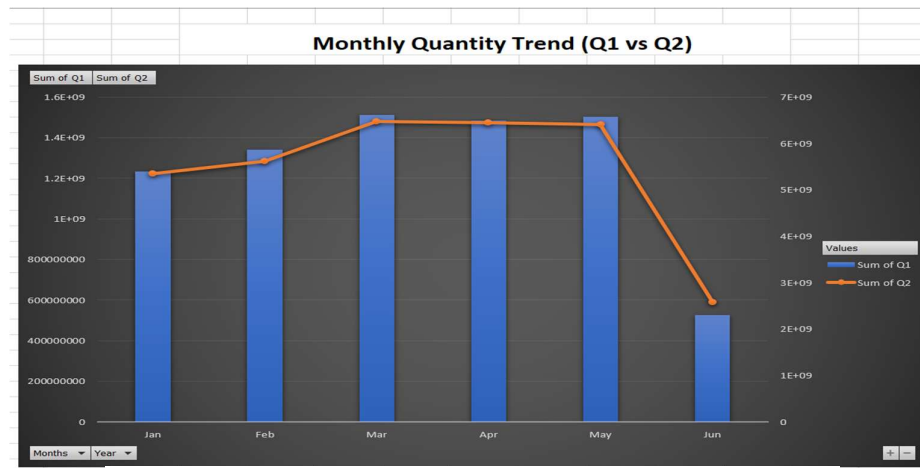
Excel Analysis

Fig. Monthly Quality Trend (Q1 vs Q2) (Bar Graph)

- From this plot, we can see that import quantities rise in February and March.
- April and May reach the highest levels, showing peak import activity.
- Q2 follows a similar upward trend as Q1 until the sudden drop in June.
- Overall, the trend starts strong early in the year but falls sharply at the end of Q2.
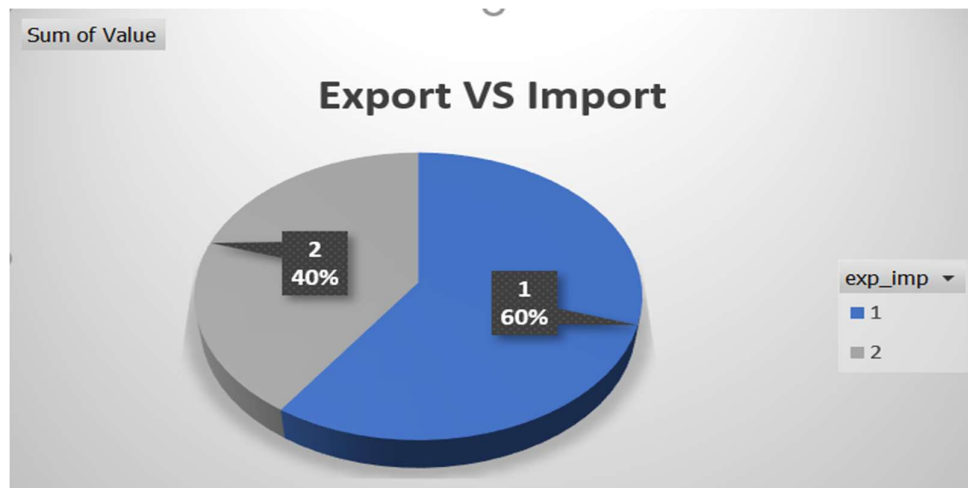

Fig. Export VS Import (Pie Chart)

- From this plot, we can see that imports (60%) make up the larger portion of the total trade value.
- Exports account for **40%**, showing they are lower compared to imports.
- This indicates that the country relies more on imports than exports in this dataset.
- The gap between imports and exports suggests a possible trade imbalance.
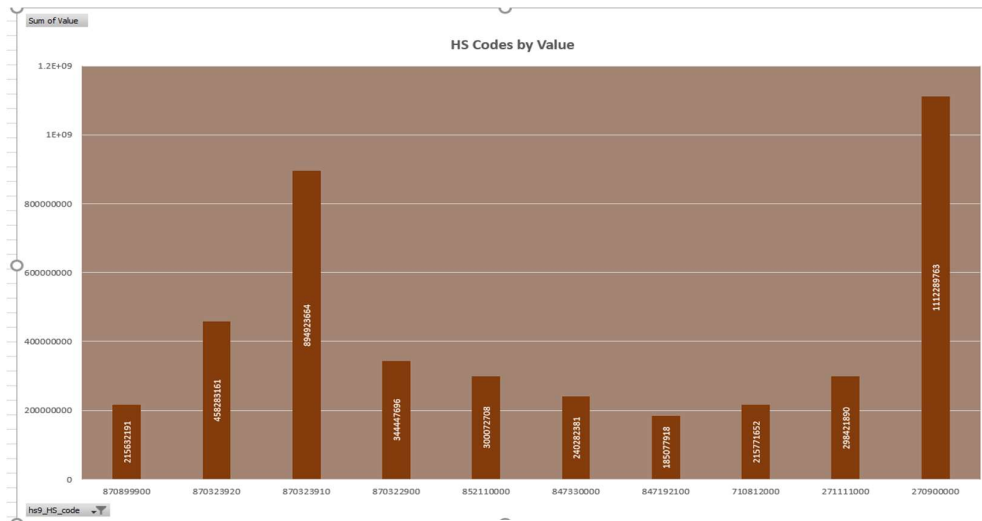- Overall, imports clearly dominate the trade activity in this comparison.

Fig. HS code by Value (Bar Chart)

- From this plot, we can see that HS code 27090000 has the highest import value among all categories.
- HS code 87032390 also shows a very high value, indicating strong demand.
- Several other HS codes show moderate import values, mostly related to vehicles and machinery.
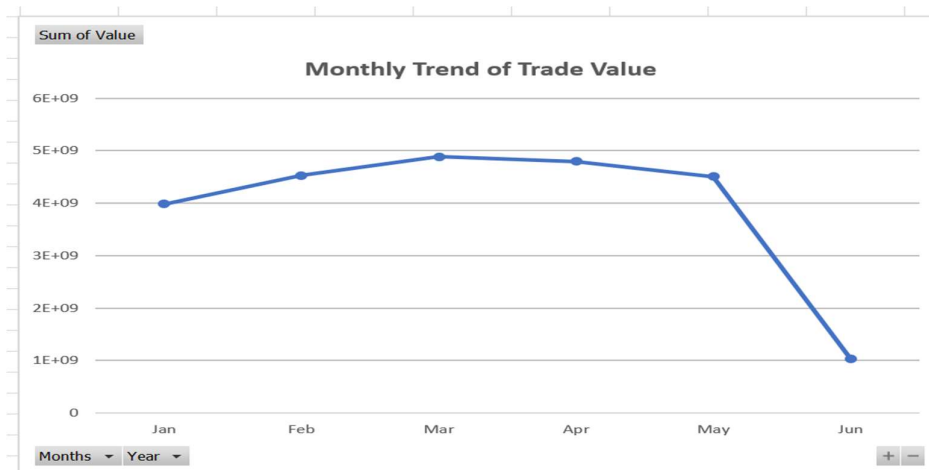- Categories like 85111000 and 84733000 appear with lower but steady import values.

Fig. Monthly Trend of Trade Value (Line Graph)

- From this plot, we can see that trade value rises from January to March.
- March reaches the highest trade value.
- April and May show a slight decline but still remain at high levels.
- June drops.
- Overall, the trade value increases early in the year and then decreases at the end of the period.
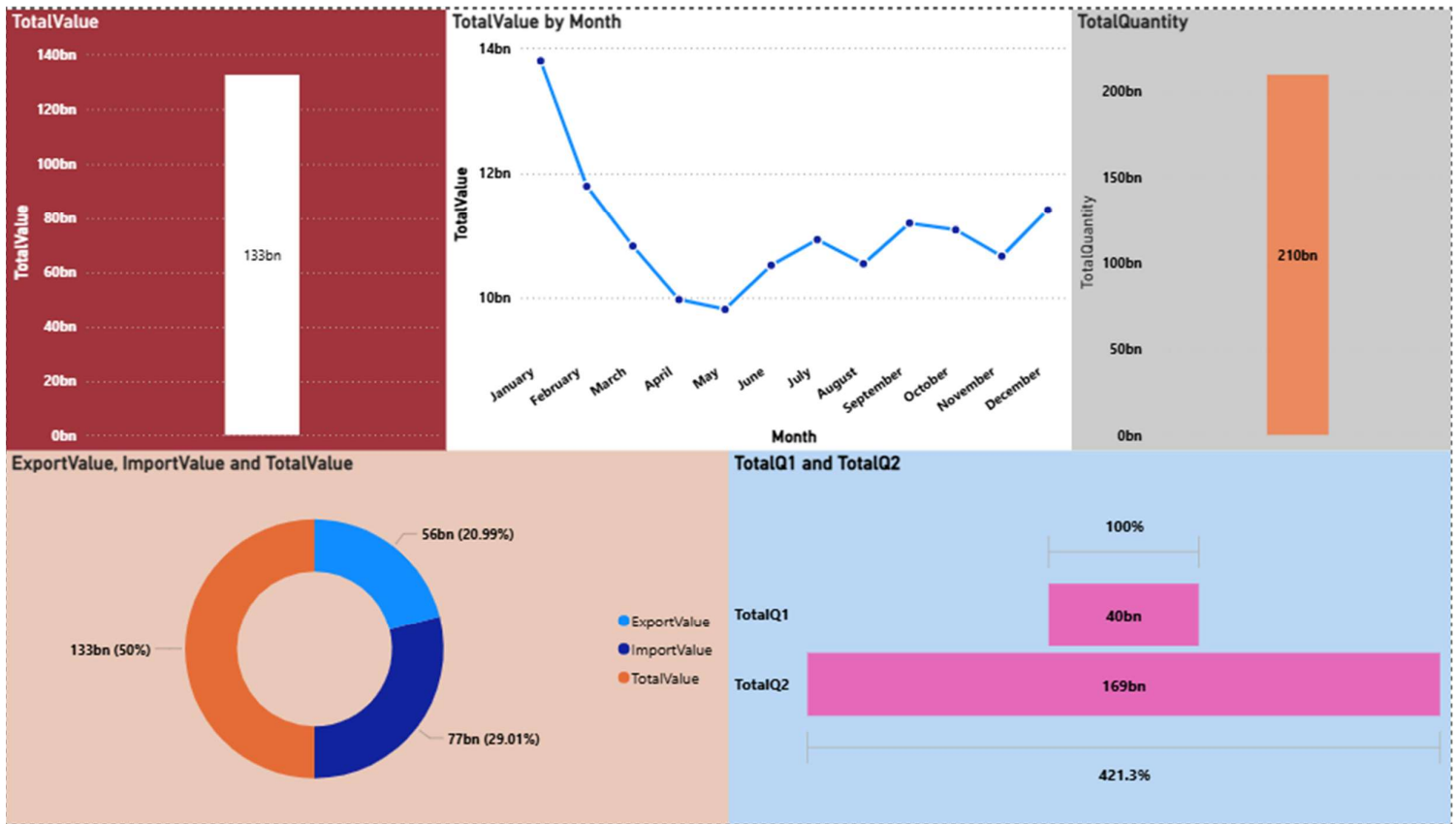
Power BI Dashboard

Fig. Power BI Dashboard

- As we can see in above dashboard, this dashboard gives an easy to understand view of the total trade value and quantity, shows how the numbers change month by month, compares exports and imports clearly, and highlights how much Q2 increased compared to Q1.
- We will see insights of each plot one by one.

1. Total Value (First Bar Chart — 133bn)

This first chart gives a clear view of the overall trade value calculated from the entire dataset.

Let's discuss what we have observed from this plot.

Observation:

- The total trade value for the period analyzed is **133 billion**.
- This acts as the baseline metric for comparing import, export, and product performance

2. Total Value by Month (Line Chart)

This line chart displays the month-wise trend in total trade value across the year.

Observations:

- The highest trading month is **January.**
- We notice decline occurs between February → May, reaching the lowest point around May.
- A moderate recovery is observed from June to September.
- Slight fluctuations occur from September to December, ending the year stronger.

3. Total Quantity (Bar Chart — 210bn)

This chart shows the total combined traded quantity across Q1 and Q2.

Observations:

- Total traded volume is 210 billion units, significantly higher than the value scale.
- This suggests that the dataset contains large-quantity, low-value items in some HS code categories.

4. Export Value vs Import Value vs Total Value (Donut Chart)

This donut chart compares the share of exports, imports, and total value.

Observations:

- The total trade value is **133bn**.

- Exports are **56bn**, and imports are **77bn**.

- Imports are higher than exports, showing a **trade deficit**.

- This suggests the country's demand is higher

5. TotalQ1 vs TotalQ2 (Horizontal Bar Comparison)

This visual compares the two quantity metrics Q1 and Q2.

Observations:

- Total Q1 quantity is **40bn**, while Q2 reaches **169bn**.

- Q2 is about **421% higher** than Q1.

- This shows that most of the traded quantity comes from **Q2**, making it the main contributing period.

.

Observation and Conclusion

This project worked with a very large dataset that had more than **113 million trade records**. Even the file had no column names and everything was just numbers, the data was cleaned and understood by checking patterns and giving proper names to each column. Using SQL, Python, and Power BI made it easier to study how the trade values changed every month and across different HS codes and categories.

From the analysis, it was clear that the dataset is well-organized and follows a monthly pattern. The HS codes used in the data match the common codes used in global trade, like those from UN Comtrade. The project also showed clear trends such as month-to-month increases and decreases, differences between imports and exports, and strong changes between Q1 and Q2.

Overall, this project shows the complete process of taking raw data and turning it into useful insights. It proves that even very large datasets can be understood with the right tools, and the results can help us learn more about how trade activities change over time.