YZV475E - Data Visualization Report

Anıl Dervişoğlu
150220344

Ömer Faruk Zeybek
150220743

1 Introduction

The project analyzes world trade trends using the "Global Commodity Trade Statistics" dataset on Kaggle. The aim here is to analyze the key patterns and relationships in these data to understand the dynamics of international trade in terms of volume and commodities. This was done in the Jupyter Notebook, where cleaning, exploration, and visualization of data were performed. In addition, an interactive Tableau dashboard was also created to visualize the results in a dynamic and accessible format. In this way, the tools used in the project will highlight the importance of data analysis and visualization to understand trends in global trade.

2 Dataset

This data set covers import and export volumes for 5,000 commodities in the majority of countries on Earth over the last 30 years. It also covers the value of the trade and the weight of the commodity in kilograms.
Link to data set : Global Commodity Trade Statistics Dataset

2.1 Dataset's Content

The dataset we chose is pretty large and contains 8225871 rows and 10 columns. Each row corresponds to a country. You can see the columns of the data set and their definition in Table [1]

Column Name	Description
Country or Area	The country or area involved in the trade
Year	The year the trade occurred
Commodity	The type of commodity being traded
Comm_code	Specific index of the commodity
Trade USD	The value of the trade in USD
Flow	Flow of trade
Quantity	The amount of the commodity traded
Weight (kg)	Weight of the commodity in kilograms
Quantity Name	Description of the quantity measurement type given the type of item
Category	Category to identify commodity

Table 1: Columns of the Global Commodity Trade Dataset

We selected the "Global Commodity Trade Statistics" dataset because of its comprehensive and dynamic characteristics, making it ideal for our analysis. This data set provides a lot of

information on global trade flows, including time-based, geographical, and commodity-specific data. This variety allows us to examine international trade from multiple angles, such as trade volume and the economic significance of different commodities.

3 Exploratory Data Analysis

Before we start with the analysis, we need to understand and check for the missing values or any other anomalies in the data. At the very first step, we can see that quantity and 'weight_kg' column has missing values. Normally, we could have used other methods as taking mean, mode or any global value instead, but since the data has far too much rows compared to missing ones and the missing values does not have a certain pattern, we decided on deleting them from the data.

Then starting with each attribute, we have first analyzed the distribution of the countries in the dataset. Having more of a linear distribution, the maximum number of entries for any country was about $170K$ which was ideal for our visualizations. But an important here is that we needed to change some of the country names to match the country names in the Tableau map visualization.

Looking into details, we can observe the start of the linear fall in the graph as well as the maximum number of entries for the top countries. The other details are also shown in the notebook as data frame analysis.

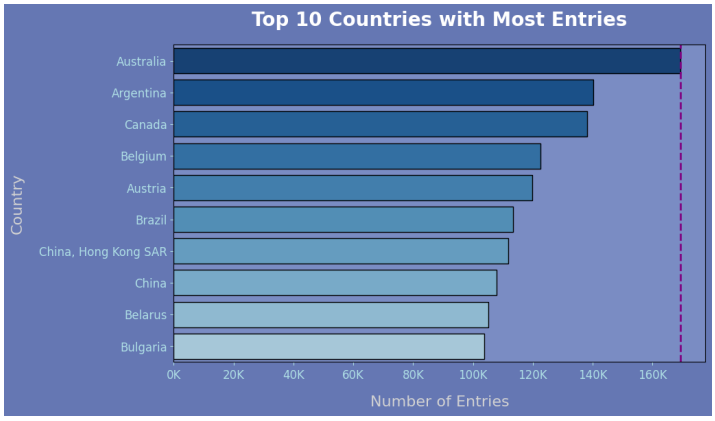

Figure 1: Country Counts

Continuing on, we have years data from 2000 to 2016 in which these transactions were made. It is important to note that the transactions from the second quarter of the 2016 are gone as it is as much data we were provided with.

We also had to make sure that our data was balanced in years as well so that we could get the same attributial data from each for countries. To make sure about that, we used a distribution graph where in Table 2 we calculated each import or export transaction made and found out the main statistical information about them. Being quite the distributed data, we can safely do our process upon it without further processing.

As one other important part to look for, we have commodity data. The data is actually given in 3 separate parts of comm_code, commodity and the category. Here comm_code specifies the commodity taken and the category is the general title for categories. Given an example, comm_code is an index indicating specific commodity (let's say the commodity is 'Sheep, live') and category would be generally explaining what the commodity is ('live animals' in this example).

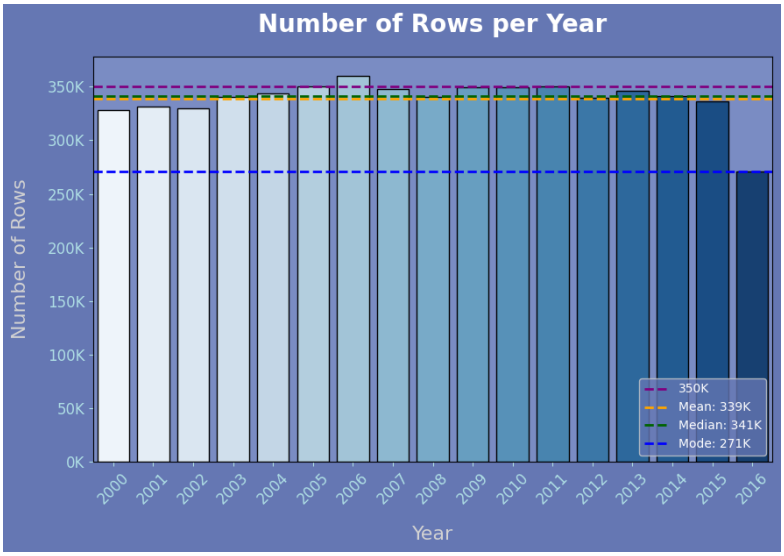

Figure 2: Year Counts

Knowing this information, we thought about keeping our presentation and project precise and clean without too many information to mess up. So we used the general titles in rather than other specific ones. Meaning that we dropped the 'comm_code' and 'commodity' columns for further clarence.

In the end, being left with category data for commodities. We again checked the statistical information about it from Table 3 and we found out that approximately 70% of the categories do have $60K$ to $80K$ examples in the data which is more than enough for our processing as the previous ones.

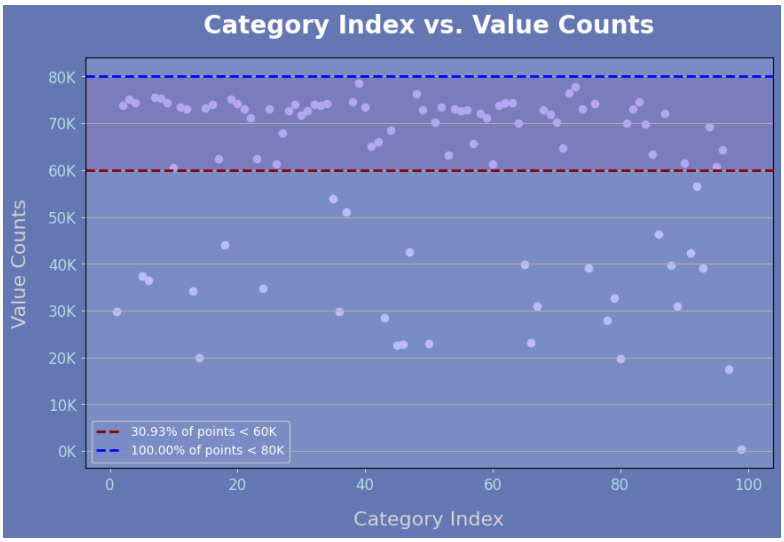

Figure 3: Category Counts

One other important attribute was the flow, showing whether the transaction was done as import, export, re-import or re-export. Here, re-import means importing the same elements which was exported and vice versa for re-export. To simply and better present our visualizations, we also believed removing taking re-import and re-export as import and export would have come in profit.

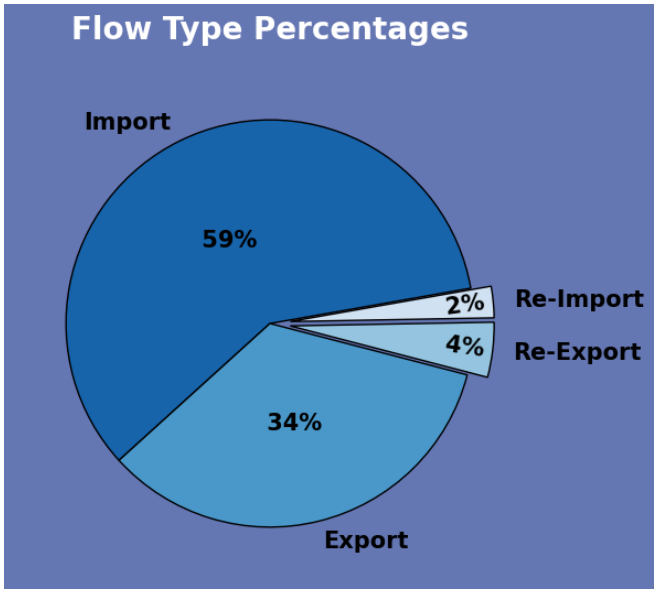

Figure 4: Flow Counts in Pie Chart

'trade_usd', 'weight_kg' and 'quantity' are the numerical features in our data set that we can use for understanding the transactions in more detail. But one of the problems is that there are too many outliers in the graph as the data set is not distributed evenly in these features among all the countries. Some of the countries are able to buy more while some other are able to buy less. To solve this issue, we have applied logarithmic scale to the data set to get a better understanding on the data.

Having over 1 million outliers in each of the stated features, we have our observations improved after logarithmic scaling as most of the outliers came also in this new outlier boundaries. We also tried to analyze the relations between these features if there was any as can be seen below.

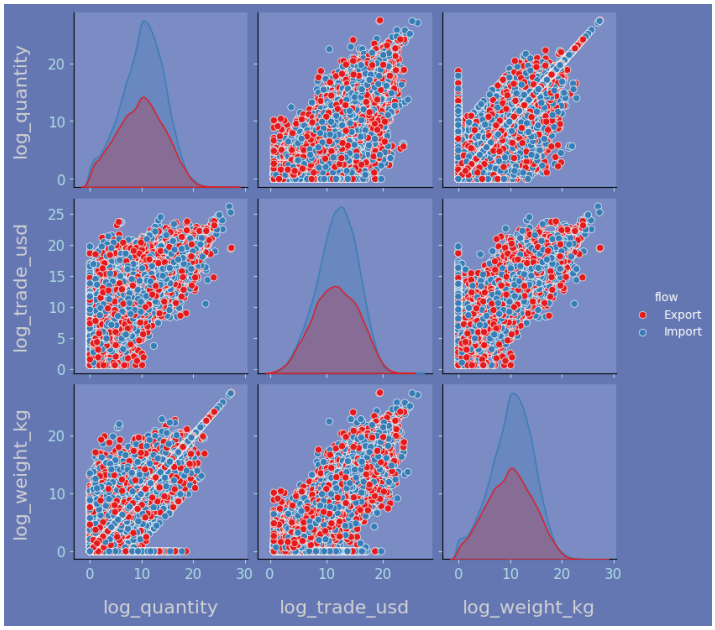

Figure 5: Pair Plot of 'trade_usd', 'weight_kg' and 'quantity' in log scales

Almost being randomly distributed, we tried another approach in which we have used different quantity names from the column to group these numerical features of 'trade_usd', 'weight_kg' and 'quantity'. And to our surprise, we came to some observations showing that if the quantities are number of packages or number of pairs, all of these features are also correlated with each other. In other scenearios if quantity is electrical energy, the 'trade_usd' is correlated with 'quantity' itself. Which shows us the idea similar to that countries can use means of electrical energy decrease the transportation cost as weight is not related with them and can increase the total profit by increasing the quantity only.

Figure 6: Correlation Matrices of relations between 'trade_usd', 'weight_kg' and 'quantity'

Eventually, the question comes down to, what is the best profitable and profited category both in import and export flows. And to understand this even further, we have used multiple subgraphs simultaneously in Table 7 and found out the top import and export categories as in the figure. And sharing the same y axises, we can say that using minerals both in import and export seems to be the most profitable one so far in all the years between 2000 and 2016.

Now, knowing which categories were the most profitable and wanted in the global, we can also check what are the countries open for these trades. And similarly to the previous graph, we have found out the most imported and exported countries in total and plotted their import and export rates simultaneously in sub graphs.

Having shared y axises in Table 8, we can see that china leads the overall trade capacity by far and especially japan have more of its capacity on importing rather then exporting etc.

In our final approach before Tableau dashboard visualization with Table 9, we tried to figure out the most profited import and export materials in the top 10 country which has made the most amount of transactions in usd. And similar to the prior graph of category profits, we can see that especially minerals are being used both as an import and export material in transactions made.

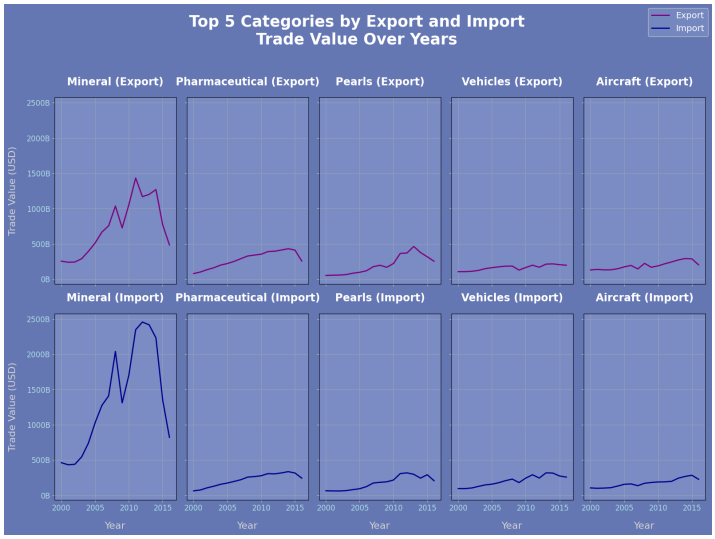

Figure 7: Categories with the most valuable transactions

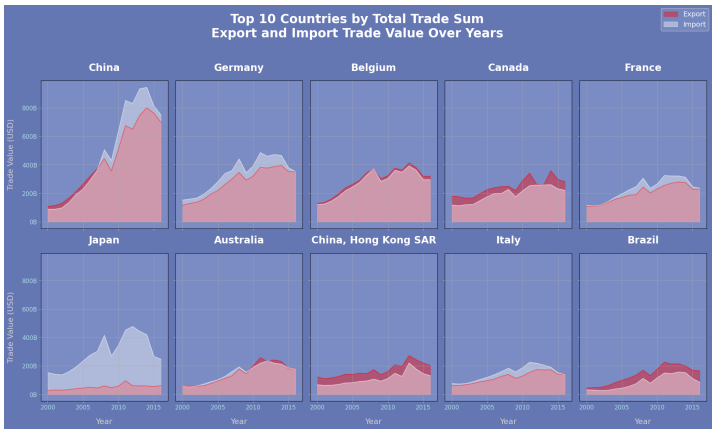

Figure 8: Countries with the most capable transactions

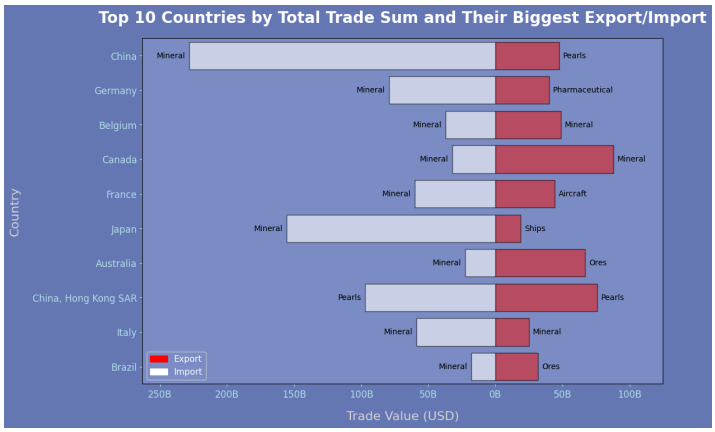

Figure 9: Best protifited import and export category in Top 10 countries

4 Tableau Analysis

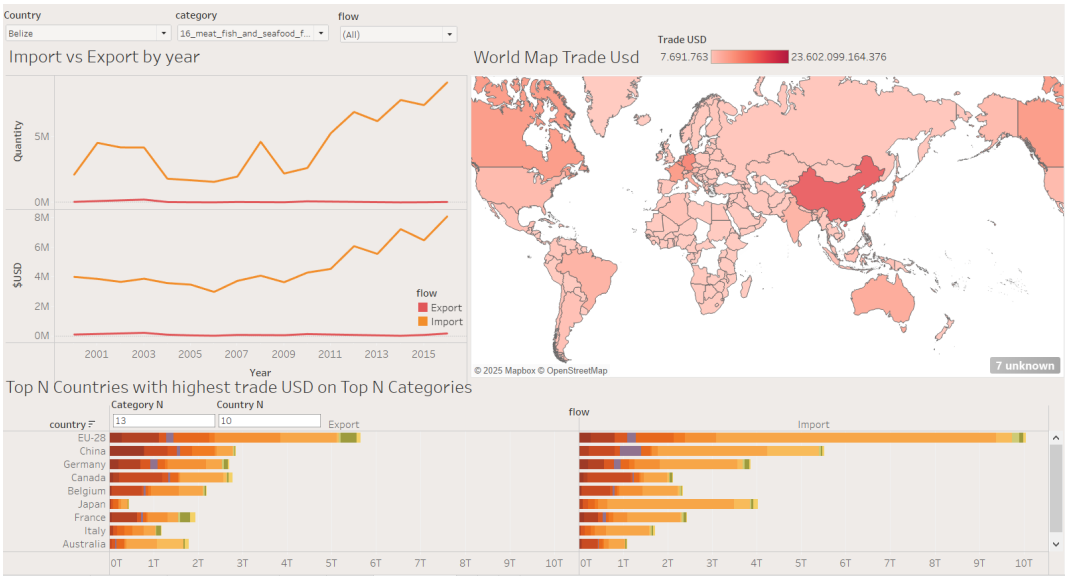

Figure 10: Tableau Dashboard

For the Tableau dashboard in Figure 10, we made 3 different visualizations:

1) **Line chart:** You can select the country, flow type, and commodity category and observe changes in quantity and trade values over the years.

2) **Map:** In the map, you can observe the trade values for each country. In addition, when you click on any country on the map, the other charts also update accordingly.

3) **Bar chart:** In the bar chart, you can observe the top N countries based on their export and import quantities. Additionally, you can adjust N to view more or fewer categories and countries.

As seen in the line chart in Figure 11, there is a significant increase in import quantities. The global economic crisis in 2008 may have contributed to this surge.

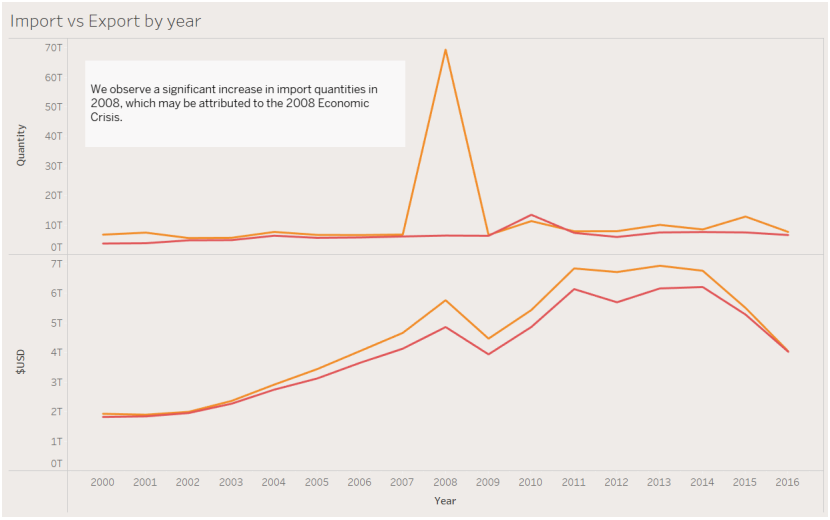

Figure 11: Line Chart

In the bar chart in Figure 12, we can observe that the EU-28 is the largest exporting and

importing area. This is because it is a combination of many countries.

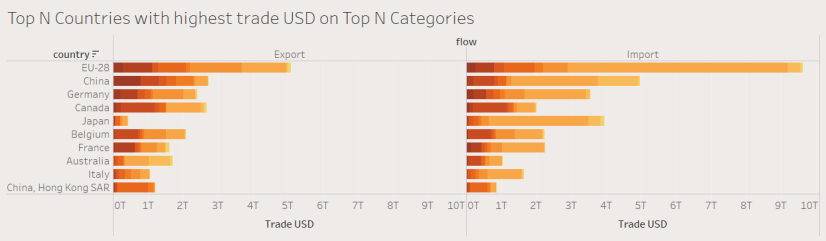

Figure 12: Bar Chart