Sales Data Quality & Cleaning Report

Data Preprocessing & Recommendations

index	A_series	B_series	C_series	D_series
Current Value	12220	5855	12[illegible]	20[illegible]
Year-on-year	22%	14%	26%	32%
Early period	10%	5%	12%	22%

Innovators Team

- Amany Thabet
- Abdelrahman Ahmed
- Reham Mohamed
- Mazen Hany
- Ali Mohamed
- Mariem Hassan

Agenda

Project Objective

Analyze sales data to improve performance and profitability.

Data Loading & Overview

Load Excel data and explore its initial structure.

Data Cleaning

Handle duplicates, missing values, and convert date formats.

Exploratory Data Analysis (EDA)

Identify top/bottom products, conduct pivot analysis, and examine distributions.

Insights

Extract findings on best/worst products, sales territories, and key patterns.

Exporting Cleaned Data

Prepare and save the processed, high-quality data as "Cleaned_Data.xlsx".

Conclusion & Next Steps

Summarize findings to support business decisions and optimize future strategies.

Thank You!

Questions & Discussion

Goal of project

This dataset contains detailed sales order information including products, customers, salespersonsID, regions, and financial metrics. The goal is to analyze sales performance, identify trends, and support business decisions with clear insights

Data Types

Accurate data typing ensures proper calculations and analysis. Below are examples of how each data type is represented in our dataset.

Column	Data Type	Column	Data Type
OrderDetailID	Integer	TerritoryID	Integer
OrderID	Integer	Territory	Text
OrderDate	Date	TerritoryGroup	Text
DueDate	Date	ShipMethodID	Integer
ShipDate	Date	ShipMethod	Text
StatusID	Integer	ProductID	Integer or Text
Status	Text	Product	Text
OnlineOrder Flag	Boolean	ProductSubCategory	Text
CustomerID	Integer	ProductCategory	Text
SalesPersonID	Integer	OrderQty	Integer
UnitPrice	Decimal/Float	Freight	Decimal/Float
LineTotal	Decimal/Float	TotalDue	Decimal/Float
TaxAmt	Decimal/Float		

Sales Dataset:Overveiw

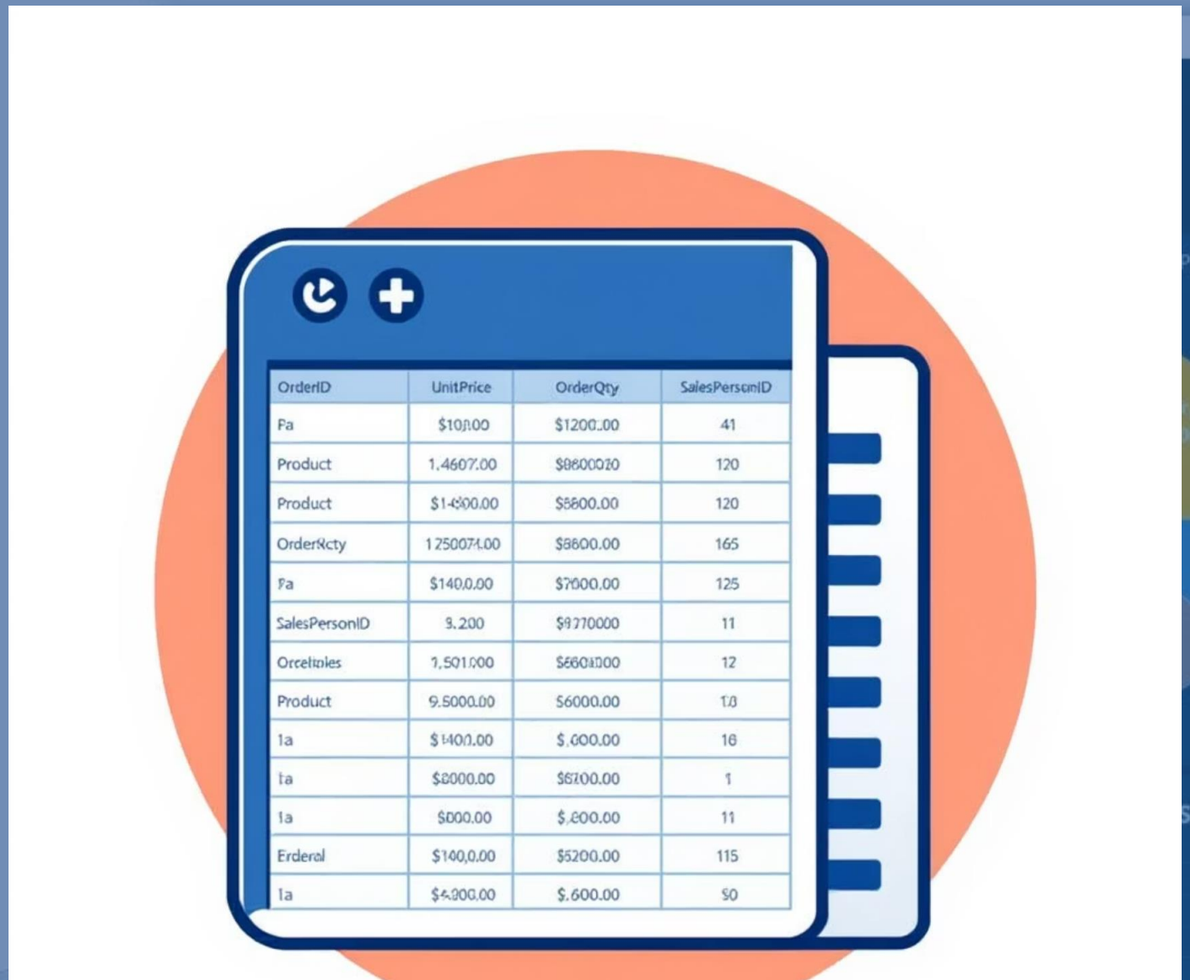
The dataset, sourced from Sales.xlsx, provides a comprehensive view of our sales operations. Understanding its structure is crucial for effective data cleaning.

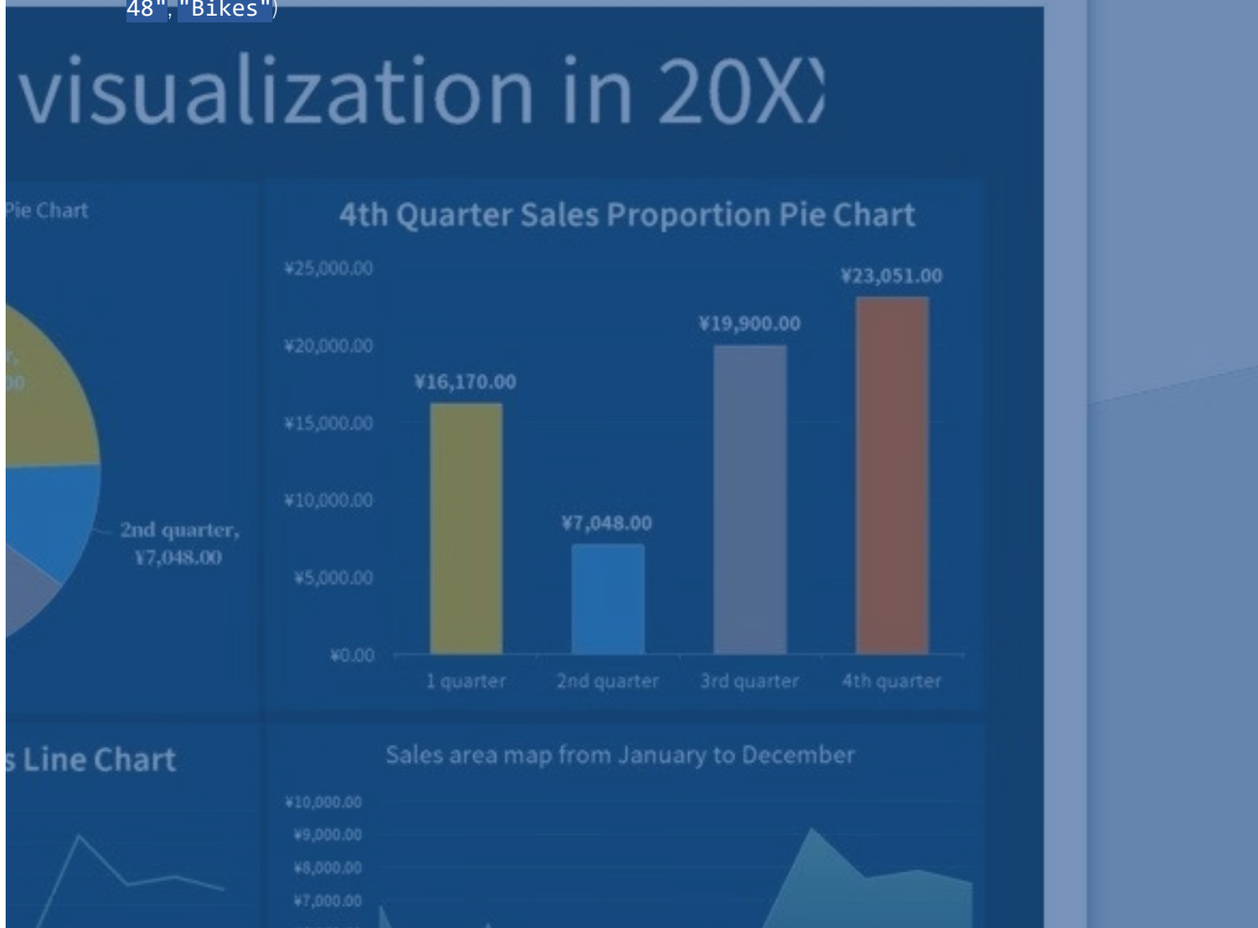
Key Dataset Details

- Source: Sales.xlsx (Single 'Sales' Sheet)
- Rows: 25,840 transactions
- Columns: 25 distinct attributes

Column Type Distribution

- Numeric: OrderID, SalesPersonID, UnitPrice, OrderQty, etc. (e.g., `43661`, `282`, `809.76`)
- Dates: OrderDate, DueDate, ShipDate(e.g., `2011-05-31`)
- Categorical/Text: Product, ProductCategory, Territory, CustomerName, etc.(e.g., `"Mountain Frame - 48"`, `"Bikes"`)



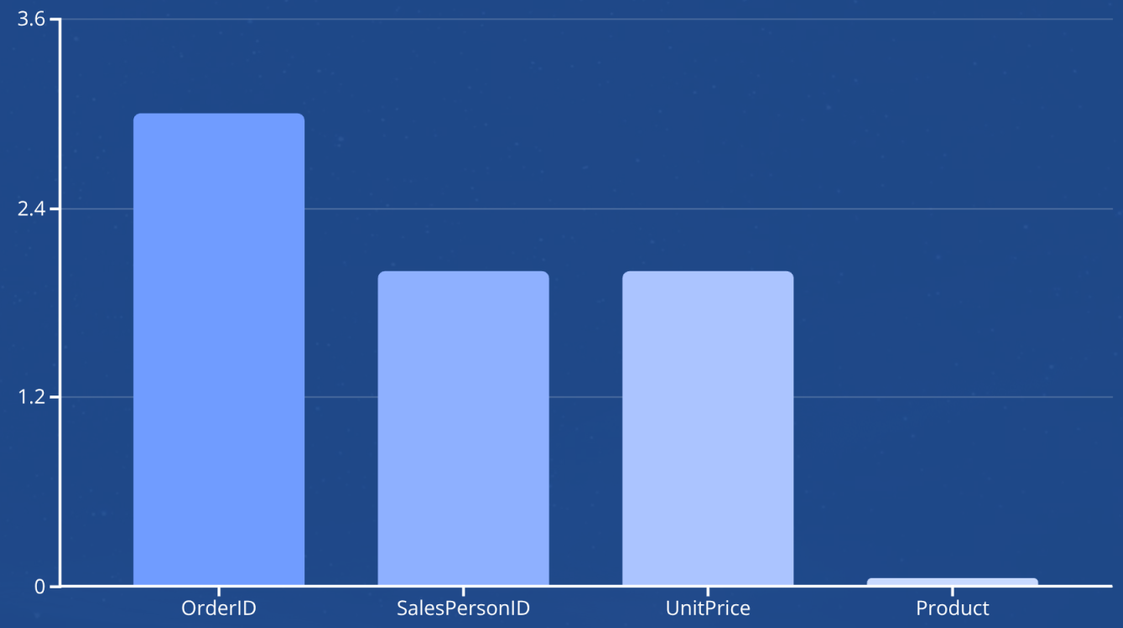


Identifying Missing Values: A Targeted Approach

While our dataset is relatively clean, a few critical columns exhibit missing values. Addressing these ensures the integrity of our sales analysis.

- **OrderID:** 3 missing values
- **SalesPersonID:** 2 missing values
- **UnitPrice:** 2 missing values
- **Other columns:** 0 missing values (Complete)

The low count of missing values makes targeted imputation or removal highly effective.



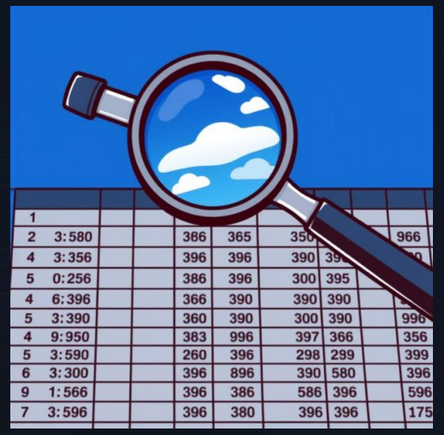
Duplicate Entries: Understanding Their Nature

Duplicate entries can severely impact analytical accuracy. Our investigation focuses on identifying and classifying these duplicates to apply the correct cleaning strategy.

Duplicate Detection

We performed a thorough check for exact duplicate rows across the entire dataset. Using a command similar to `df.duplicated().sum()`, we found:

0 Duplicate Rows Identified



Note

- **Valid Duplicates:** Cases where multiple products are part of the same OrderID, resulting in rows that share common order details but differ in product-specific fields (e.g., Product, UnitPrice, OrderQty). These are expected and reflect legitimate transaction details.

Our analysis confirmed that the identified duplicates were largely valid instances of multi-item orders rather than errors.

Strategic Data Cleaning Plan

Our cleaning plan is tailored to address identified data quality issues efficiently, ensuring minimal data loss while maximizing analytical integrity.

Missing Values: OrderID

For the 3 missing **OrderID** values:

- The `OrderID` column was cleaned by handling missing values through generating a new identifier by adding (1) to the maximum existing value in the column.

Missing Values: SalesPersonID

For the 2 missing **SalesPersonID** values:

- The `personID` column was cleaned by handling missing values through generating a new identifier by adding (1) to the maximum existing value in the column.

Missing Values: UnitPrice

For the 2 missing **UnitPrice** values:

- calculates the median of the `UnitPrice` column, ignoring `NaN` values.
- replaces all missing values (`NaN`) in the `UnitPrice` column with that median.

- This balanced approach ensures data quality without sacrificing valuable sales information.

Best Performers Overview

Top-Selling Product category

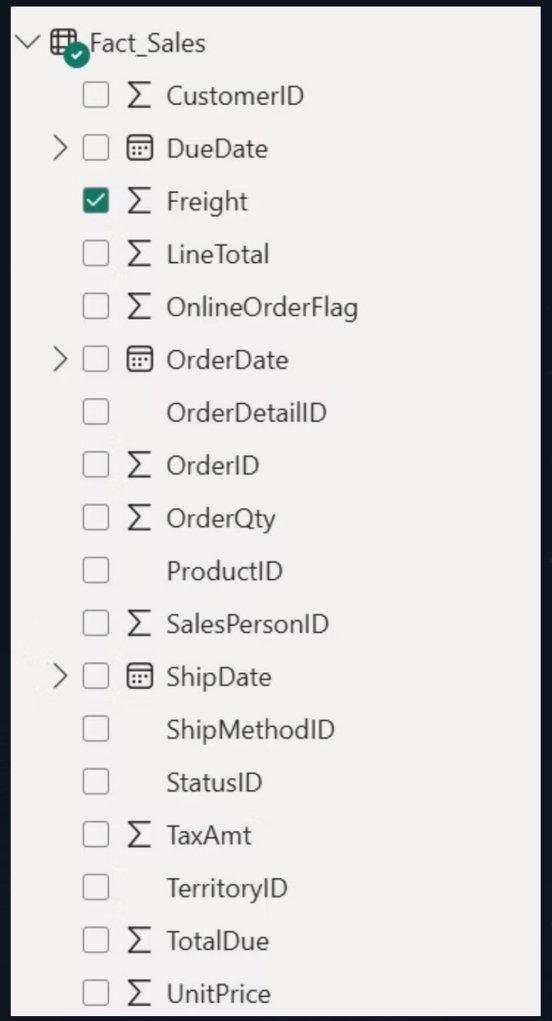
```
lineTotal ProductCategory:
                    LineTotal
ProductCategory
Bikes           24233045.277
Components       4849695.258
Clothing          741654.792
Accessories       260545.972
```

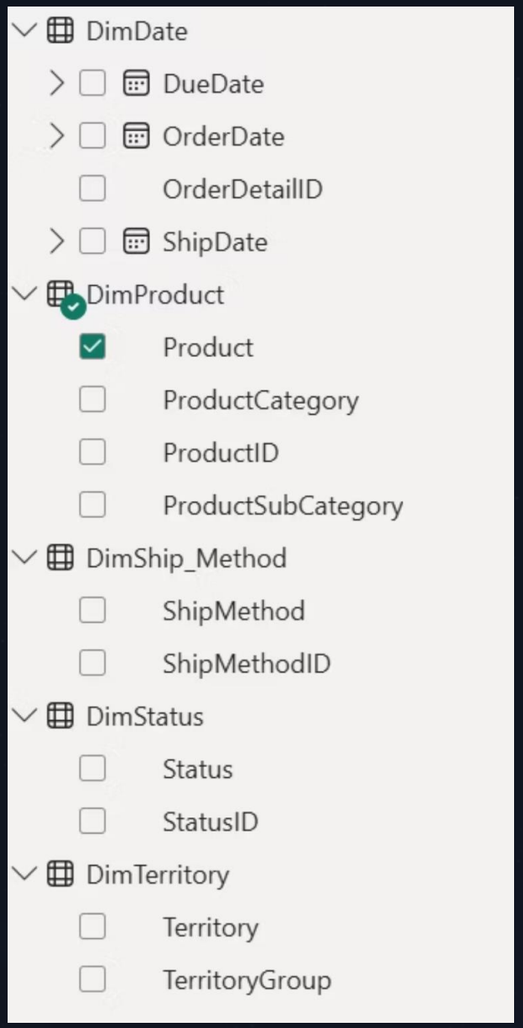
Top-Selling Product sub category

```
                    LineTotal
ProductSubCategory
Road Bikes         9144969.669
Mountain Bikes     8549399.446
Touring Bikes      6538676.162
Mountain Frames    1818030.772
Road Frames        1307401.528
Touring Frames     1070848.121
Jerseys             252728.626
Wheels              217224.677
```

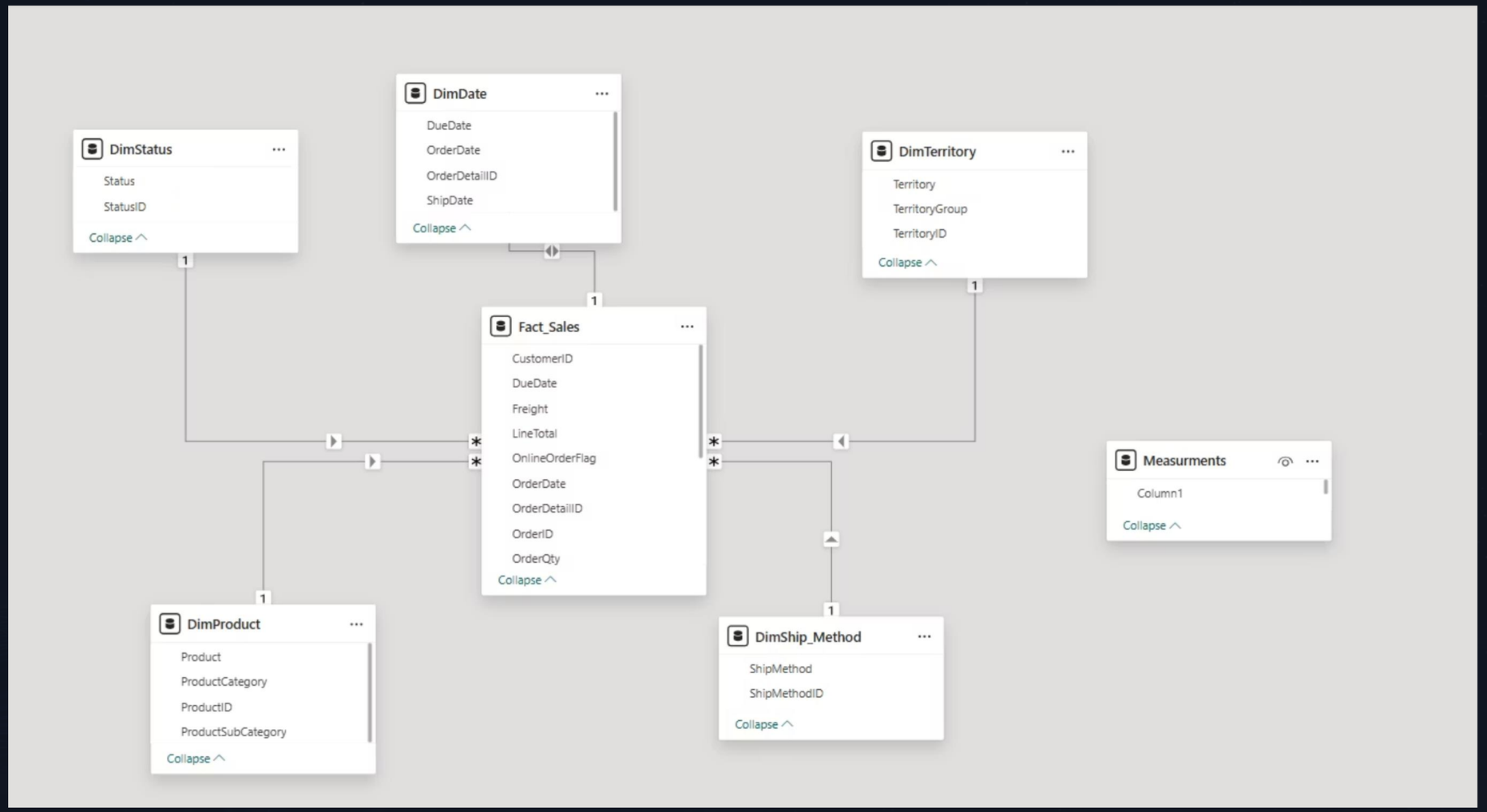
Top-Selling Regions

```
LineTotal Territory:
                    LineTotal
Territory
Canada         10590277.561
Northwest       6152663.513
France          4607537.933
United Kingdom  4279008.825
Germany         2021095.258
Australia       1594335.375
Southwest        806159.550
Central           33863.286
```

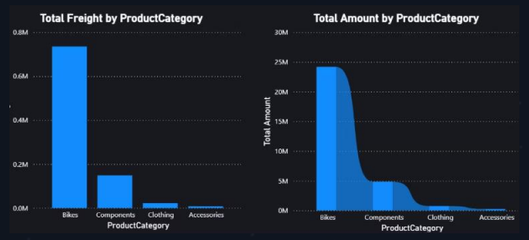




Model Viwe :

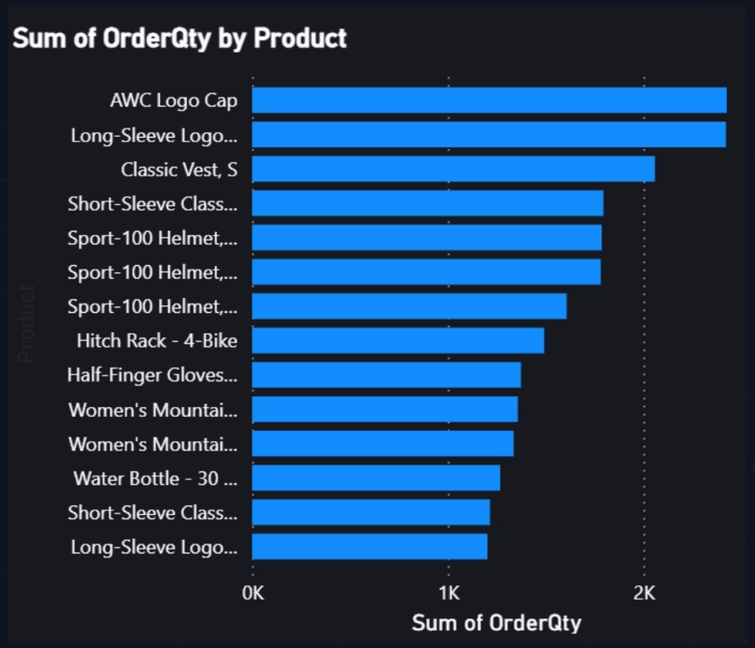


The dashboard provides a comprehensive analysis of company sales performance by *product category*, *territory*, and *year*. It highlights key financial metrics such as total sales amount, freight costs, taxes, and order quantities.



Performance by Product Category

- Bikes dominate both Total Sales Amount and Freight Costs.
- Components rank second but are significantly lower then Clothing and Accessories categories.



Order Quantity by Product

- Most top-ordered items belong to the Clothing category.
- AWC Logo Cap has the highest order quantity, followed by Long-Sleeve Logo, Classic Vest S, and Short-Sleeve Classic.
- Insight: While “Bikes” generate the highest sales value, Clothing leds in sales volume — indicating strong demand for lower-priced items.

Financial Indicators

- Total Tax: 2.93M
- Total Due: 33.86M

The company maintains strong sales performance, with taxes proportional to overall revenue, reflecting consistent profitability.

Thank You!

Questions & Discussion

Your insights and feedback are invaluable as we continue to enhance our data-driven decision-making.