Global sales data for electronics company

Sales data for a fictitious global electronics retailer, including tables containing information about transactions, products, customers, stores and currency exchange rates.

Project Objective.

The goal of this project was to analyze sales data from multiple sources to generate actionable insights that can drive business decisions. The analysis focused on revealing revenue trends, customer behavior, delivery efficiency, and product performance across regions and sales channels. Furthermore a profitability analysis was also conducted to get a clearer picture of the business performance.

The data was processed, cleaned, and analyzed using SQL, and insights were gathered for potential visualization in Tableau Public.

The business seeks to get insights for the following questions:

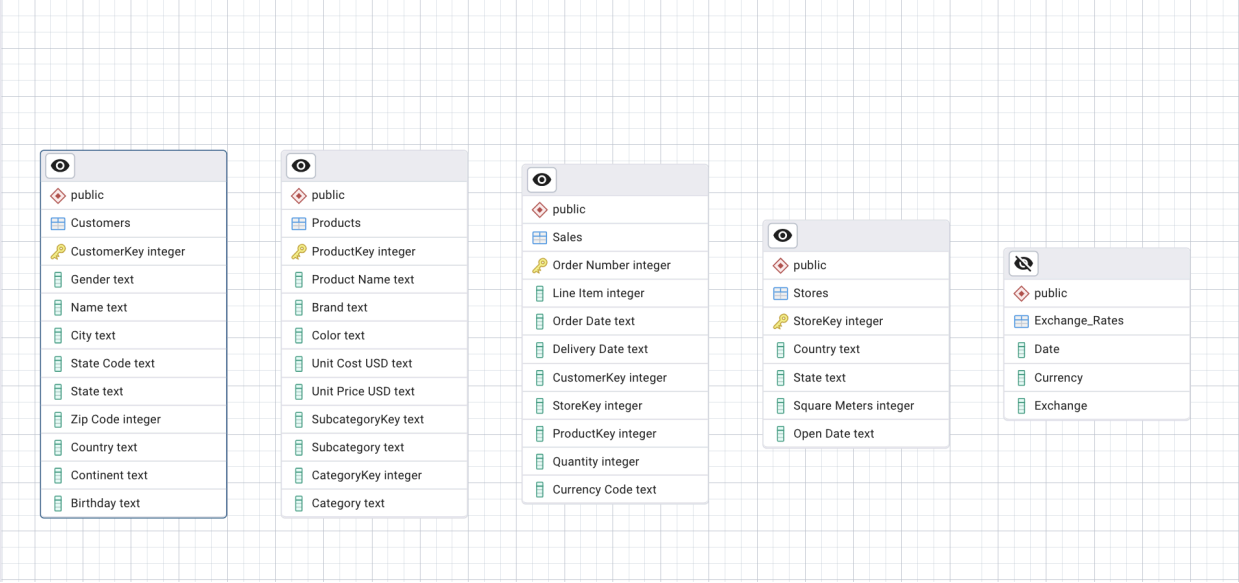
1. What types of products does the company sell, and where are customers located?

2. Are there any seasonal patterns or trends for order volume or revenue?

3. How long is the average delivery time in days? Has that changed over time?

4. Is there a difference in average order value (AOV) for online vs. in-store sales?

Initial ER diagram based on the dataset provided.



It can be seen that some of the tables need to be normalized.

Table ‘Customers’: Already in BCNF. No changes necessary for now.

`CustomerKey` is a candidate key (uniquely identifies each customer).
No transitive or partial dependencies exist.

Table ‘Products’:

ProductKey is a candidate key
SubcategoryKey -> Subcategory, CategoryKey -> Category unrelated dependency to Products (decomposition required)

Tables ‘Sales’:

Order Number, Line Item is the composite candidate key for this table.
No partial or transitive dependencies exist within the current structure.

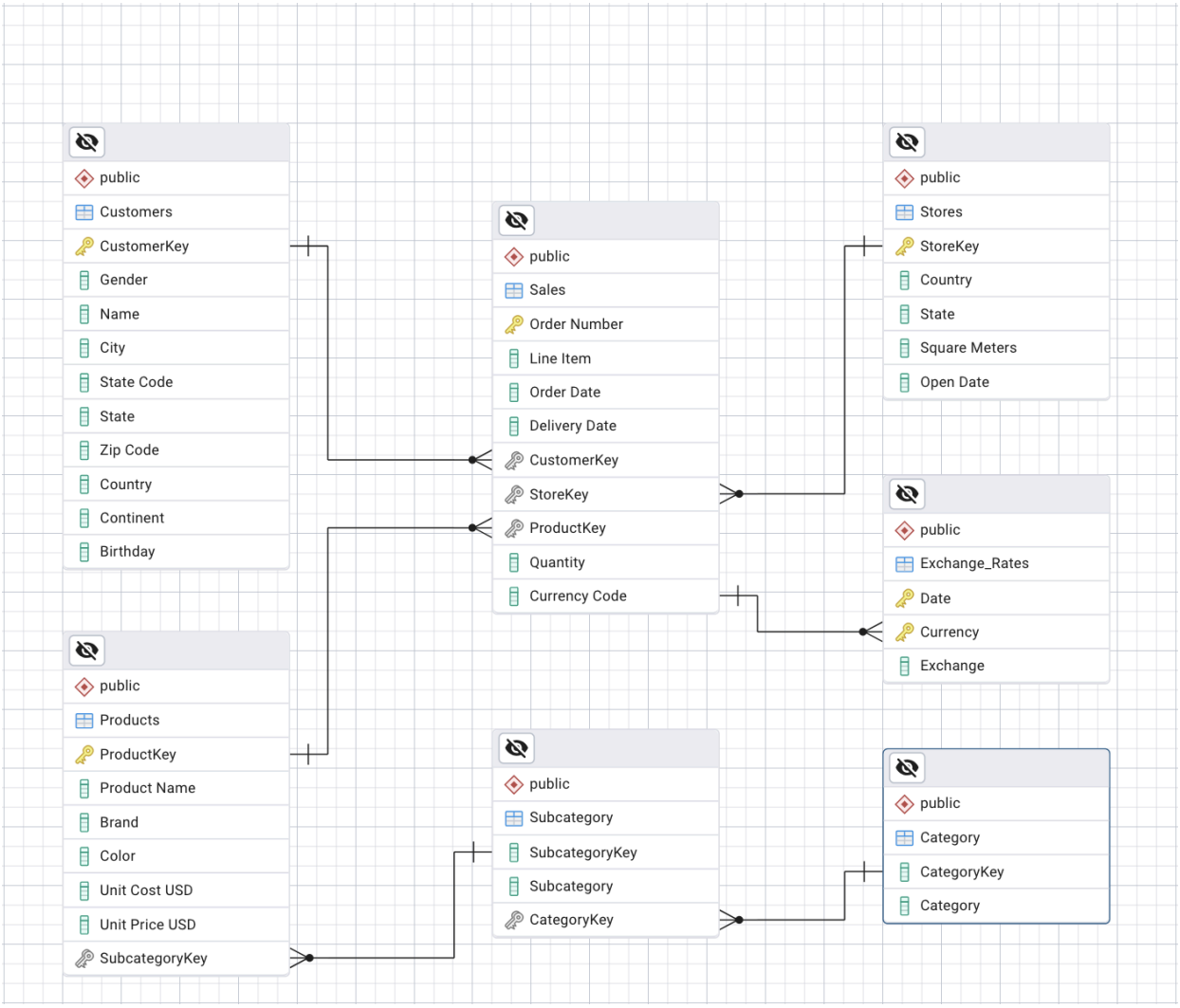
Already in BCNF.

Table ‘Stores’: Already in BCNF

Table ‘Exchange Rates’: Already in BCNF.

By doing these transformations, an updated Entity Relationship Diagram of Global Electronics Sales data that is normalized and with relationships is established.

New ER - Diagram with updated relationships and cardinalities.



Implementation.

Writing SQL scripts to create tables.

Before creating tables to load data to PostgreSQL, staging tables can be created in order to adjust potential data inconsistencies, so that later, data can be added with constraints that match with the updated semantic schema.

Data Import Workflow for Staging Tables

This document outlines the steps taken to import and process data from CSV files into PostgreSQL staging tables.

1. Preparation of Data Files

Initial Issue: The CSV files had encoding issues, with some files containing special characters (e.g., ü for ürstenberg).

Resolution:

- Detected file encodings using the chardet Python library.
- Converted all files to UTF-8 encoding for compatibility with PostgreSQL.
- Cleaned special characters where necessary while preserving data quality.

Tools Used: Python, pandas, chardet

2. Creating Staging Tables

Creating a staging table allows us to import all of the data first, regardless of cleanliness, and then perform data cleaning and validation before moving it to the final destination table with constraints.

Example: Creating staging_sales table

```
`CREATE TABLE Staging_Sales (
        "Order Number" VARCHAR(255),
        "Line Item" VARCHAR(255),
        "Order Date" VARCHAR(255),
        "Delivery Date" VARCHAR(255),
        CustomerKey VARCHAR(255),
        StoreKey VARCHAR(255),
        ProductKey VARCHAR(255),
        Quantity VARCHAR(255),
        "Currency Code" VARCHAR(255)
);`
```

3. Data Import into Staging Tables

Bulk Data Import:

Used PostgreSQL's COPY command to load data directly from the CSV files.

Ensured file paths and permissions were correct.

Following command is used in PSQL

/COPY staging_customers
FROM '/path/to/your/converted/Customers.csv'
WITH (FORMAT csv, HEADER true, DELIMITER ',');

Handling Permissions:

Resolved server-side permission issues by moving files to accessible directories or using the \COPY command for client-side imports.

Validation After Import:

Verified successful imports using queries like:

SELECT COUNT(*) FROM staging_customers;
SELECT * FROM staging_customers LIMIT 10;

Checked for null values and duplicates:

SELECT * FROM staging_customers WHERE customer_id IS NULL;
SELECT customer_id, COUNT(*) FROM staging_customers GROUP BY customer_id HAVING COUNT(*) > 1;

4. Data Validation and Cleaning

Identified and resolved encoding-related issues, ensuring that special characters (e.g., ü) were correctly imported.

Performed basic data validation to ensure completeness and consistency.

Verified critical columns for null values and duplicates.

5. Outcome

Successfully loaded all data into staging tables for further processing.

Ensured data integrity and compatibility with PostgreSQL.

Cleaning the dataset.

To ensure data is in the correct format for analysis, certain cleaning methods were used. By cleaning data, inconsistent raw data can be handled appropriately and then standardize it.

1. Column Renaming

- Certain columns were renamed renamed for consistency (e.g., Line Item → line_item, Currency Code → currency_code).

  Example: ` `

2. Data Type Changes

- Converted date columns (order_date, delivery_date) into proper DATE format for time-based calculations.

  Example: ` `

- Changed numeric columns (order_number, quantity, productkey, etc.) to INTEGER for better data consistency.

  Example: ``

3. Data Integrity Checks

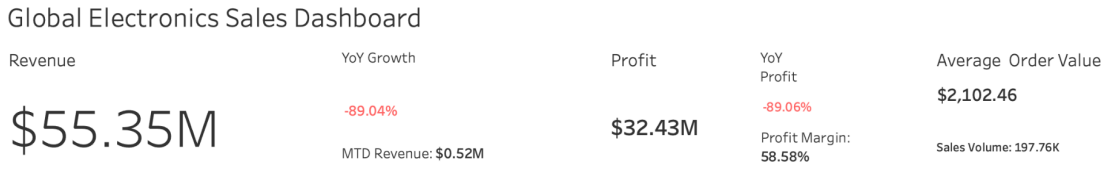
- Missing Product Keys: Checked for product keys in staging_sales that don't exist in staging_products.

  Example: ``

- Null Values:Checked for NULL values in critical fields like productkey and quantity.
  Example: ``

Performing the Analysis

Develop key performance metrics



Analysis 1:

What types of products does the company sell, and where are customers located?

Top Categories

Category	
Computers	$19.14M
Home Appliances	$10.75M
Cameras and camcorders	$6.50M
Cell phones	$6.12M
TV and Video	$5.86M
Audio	$3.15M
Music, Movies and Audio ..	$3.11M
Games and Toys	$0.72M

Top 10 Subcategories

Subcategory	
Desktops	$9.78M
Televisions	$4.26M
Projectors & Screens	$3.77M
Water Heaters	$3.53M
Camcorders	$3.37M
Laptops	$3.15M
Movie DVD	$3.11M
Touch Screen Phones	$3.05M
Smart phones & PDAs	$2.77M
Refrigerators	$2.14M

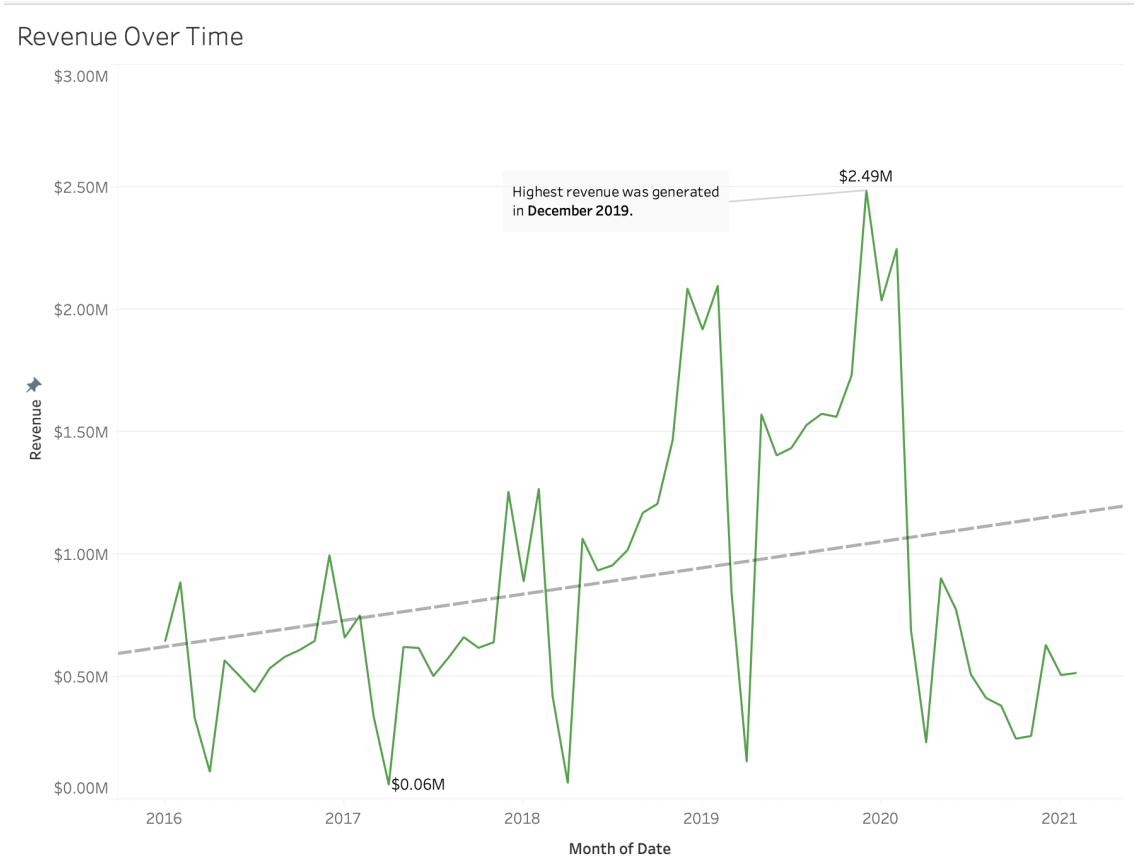
Insights:

The company's sales are dominated by the **Computers** category, which generates a total revenue of **$19.14M**, followed by **Home Appliances** at **$10.75M** and **Cameras and Camcorders** at **$6.50M**. These categories are the key drivers of the company's overall revenue. Within subcategories, **Desktops** contribute the highest revenue at **$9.78M**, reinforcing the significance of the Computers category. Other strong subcategories include **Televisions** ($4.26M) and **Projectors & Screens** ($3.77M).

On the other hand, **Games and Toys** perform the worst, with only **$0.72M** in revenue, indicating a need for focused marketing or product improvements in this segment. These insights suggest prioritizing investment in top-performing categories and subcategories while exploring strategies to boost sales in underperforming areas.

Analysis 2:

Are there any seasonal patterns or trends for order volume or revenue?



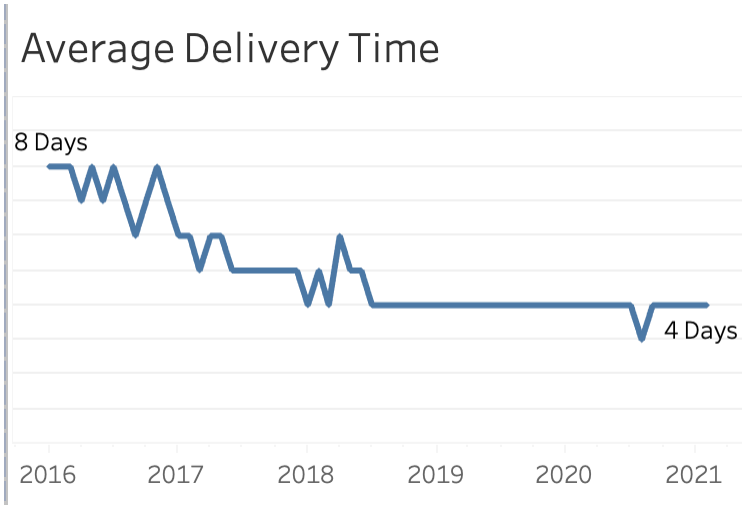
Insights:

The revenue trends reveal a clear seasonality, with noticeable peaks during the holiday season in December, as reflected by the highest revenue of **$2.49M** in December 2019. This indicates that the company benefits significantly from seasonal shopping periods. Additionally, there is an overall upward trend in revenue over the years, despite short-term fluctuations. However, the trend started to decline post-2020, possibly due to external factors such as market conditions or operational disruptions.

These insights highlight the importance of aligning marketing and inventory strategies with seasonal patterns to maximize revenue during peak months. It is also critical to investigate the factors contributing to the decline after 2020 to address potential challenges proactively.

Analysis 3:

How long is the average delivery time in days? Has that changed over time?



Tables: Sales
Columns: order_date, delivery_date, order_number

In this case we are assuming the null records in the delivery_date column are orders which are not completed. As such, only records with non null records are considered for this delivery.

```
`SELECT
	AVG(delivery_date - order_date) AS avg_delivery_time
FROM staging_sales
WHERE delivery_date IS NOT NULL;`
```

Result: ~ 5 days

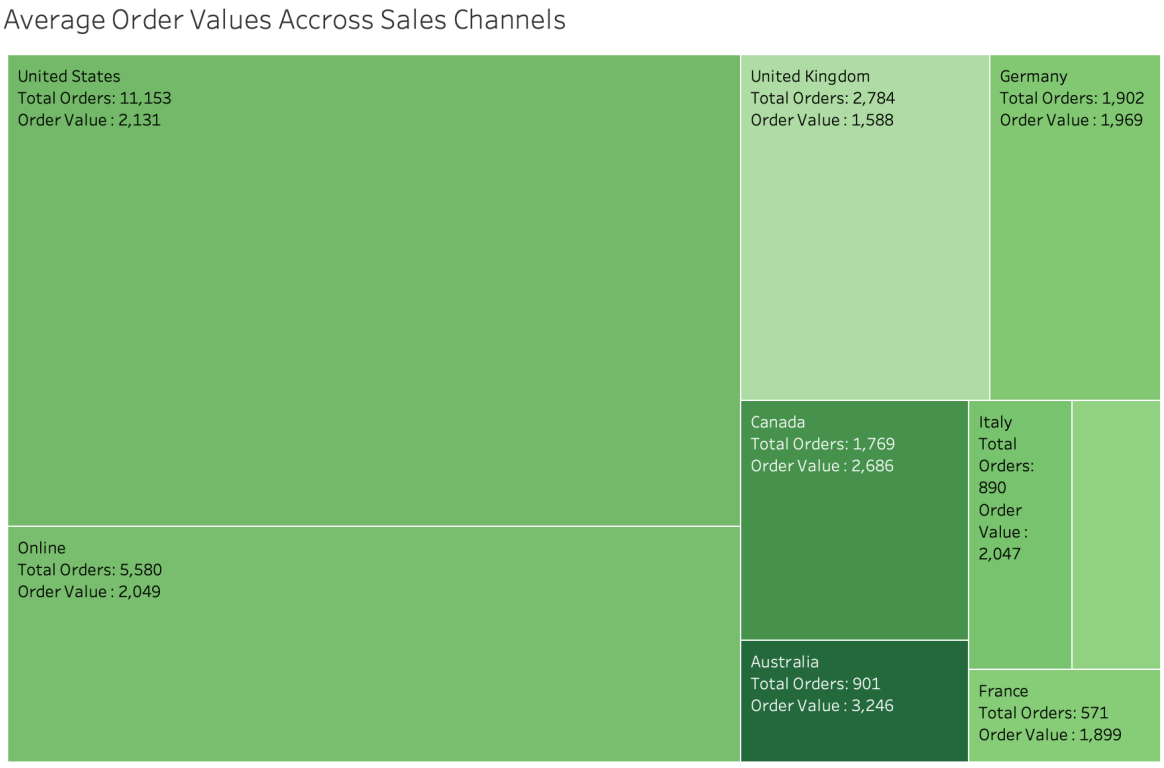
How has it changed over time?

Has reduced over time from 7 to 4 days. That is a ~43% decrease in delivery time.

Over the course of months there has been a reduction in average delivery days. This could be a result of overall reduction in the average order volume. Further analysis whether delivery time correlates with factors like **order size**, **order volume**, or **specific regions** can be conducted to identify problematic areas.

Analysis 4:

Is there a difference in average order value (AOV) for online vs. in-store sales?



Understanding the differences in AOV for online vs in-store sales can help the business to understand the **customer behaviour** across different channels, **help allocate resources** based on revenue contribution from each channel and provide **insights to profitability** and efficiency of online vs in-store operations across different regions.

Tables: Sales, Store, Customer
Columns: order_number, store_number, customer_id, product or line_items
Calculated Colums: total_value, AOV

Of a channel, AOV = Total Revenue / Number of Orders

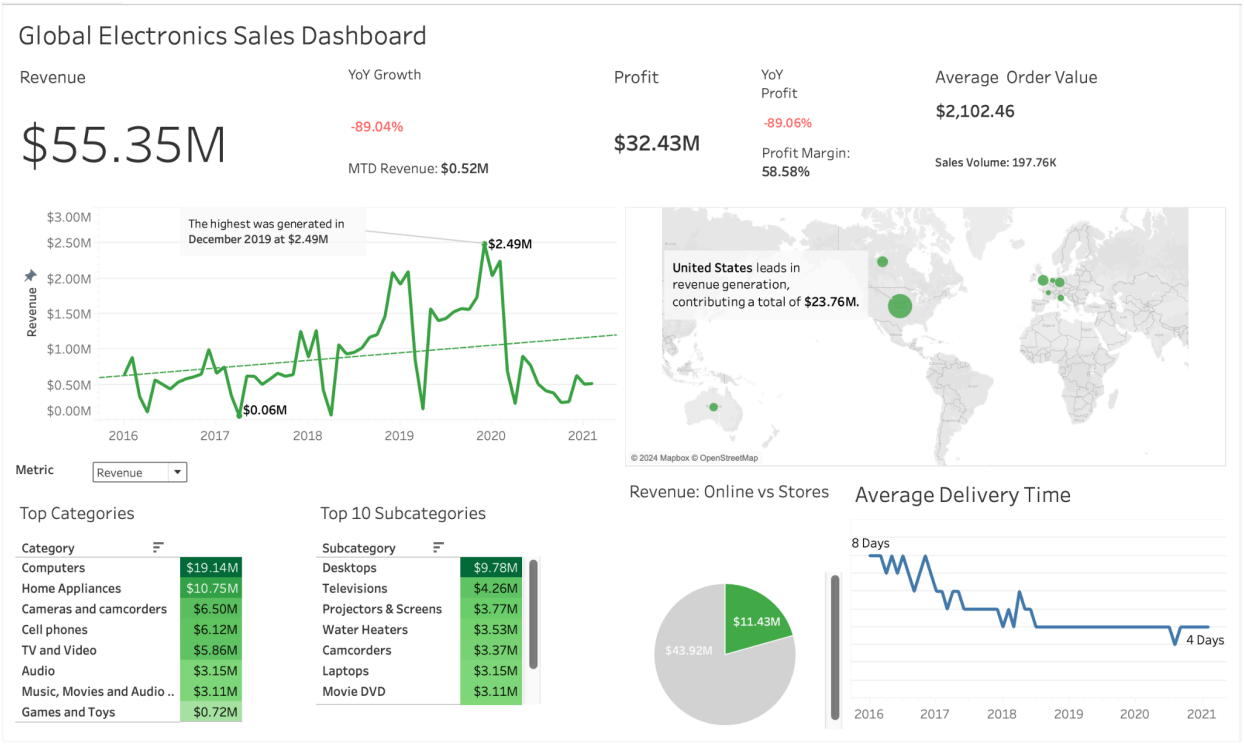
Insights:

The analysis shows a clear variation in **Average Order Value (AOV)** across sales channels and regions. **Online sales** have an AOV of **$2,049**, which is slightly lower than the **United States** in-store AOV of **$2,131**, highlighting strong performance for both channels.Similarly, **Canada** also has a high AOV of **$2,686**, demonstrating the potential profitability of specific regional markets.

Conversely, regions like the **United Kingdom** and **Germany** have lower AOVs at **$1,588** and **$1,969**, respectively, despite significant order volumes. This indicates opportunities to enhance customer spending in these markets through targeted promotions or product offerings. The overall insights suggest that while online sales are competitive, specific regions with high AOVs may offer more lucrative opportunities.

Australia stands out with the highest AOV of **$3,246** despite low total orders (901). This could be due to premium products, bulk purchases, or niche market strategies. Further analysis of product categories, pricing, and customer behavior is needed to uncover the reason and replicate success in other regions.

Dashboard


Global Electronics Sales Dashboard
Revenue
$55.35M
YoY Growth
-89.04%
MTD Revenue: $0.52M
Profit
$32.43M
YoY Profit
-89.06%
Profit Margin: 58.58%
Average Order Value
$2,102.46
Sales Volume: 197.76K
$3.00M
$2.50M
$2.00M
$1.50M
$1.00M
$0.50M
$0.00M
Revenue
The highest was generated in December 2019 at $2.49M
$2.49M
$0.06M
2016
2017
2018
2019
2020
2021
United States leads in revenue generation, contributing a total of $23.76M.
© 2024 Mapbox © OpenStreetMap
Metric
Revenue
Top Categories
Category
Computers $19.14M
Home Appliances $10.75M
Cameras and camcorders $6.50M
Cell phones $6.12M
TV and Video $5.86M
Audio $3.15M
Music, Movies and Audio .. $3.11M
Games and Toys $0.72M
Top 10 Subcategories
Subcategory
Desktops $9.78M
Televisions $4.26M
Projectors & Screens $3.77M
Water Heaters $3.53M
Camcorders $3.37M
Laptops $3.15M
Movie DVD $3.11M
Revenue: Online vs Stores
$11.43M
$43.92M
Average Delivery Time
8 Days
4 Days
2016
2017
2018
2019
2020
2021