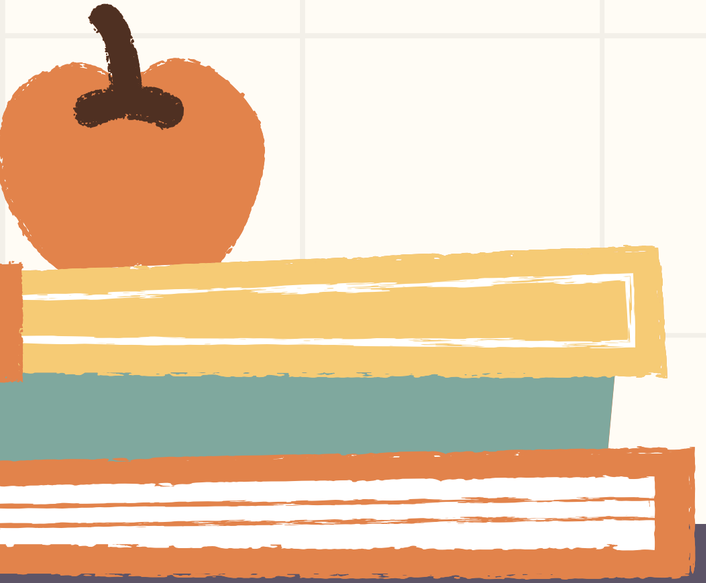
Sales Forecasting analysis for Urban Markets



ASLAM MUSTHAFA
AVIJEET
BASITH
BLESSY EVANGELINE
DEVENDRAN
DHANUSRI

INTRODUCTION

- In today’s data-driven world, organizations generate massive volumes of transactional data on a daily basis. However, raw data in its original form holds limited value unless it is effectively processed and analyzed to reveal meaningful patterns. Our project, titled "Sales Data Analysis and Visualization with Streamlit", is centered around transforming such raw sales data into strategic business insights using modern data analytics techniques.

- By leveraging Python’s data manipulation and visualization capabilities, this project aims to explore, clean, and dissect a large dataset of over half a million sales transactions. Beyond just analyzing static numbers, the project introduces an interactive dashboard built with Streamlit, allowing stakeholders to interact with the data in real time, drill down into patterns, and make informed decisions with ease.

- This end-to-end workflow—from raw data to dashboard—bridges the gap between technical data science processes and practical business use, making analytics more accessible to non-technical users like sales managers, marketers, and decision-makers.

PURPOSE

The primary purpose of this project is to unlock hidden business intelligence from sales transaction data and deliver it through intuitive visualizations and interactive tools. It seeks to answer critical business questions such as:

- Which products are generating the most revenue?
- Which countries contribute the most to sales?
- What are the seasonal trends or buying patterns?
- How do customer behaviors vary across regions?

Through meticulous data processing and visualization, this project allows us to quantify performance, identify opportunities, and highlight anomalies.

In addition, by deploying the analysis on a Streamlit dashboard, we provide an accessible platform where business users can explore KPIs, visualize trends, and filter information based on their needs—without writing a single line of code.

Ultimately, the project stands as a blueprint for how organizations can democratize data insights, ensure data transparency, and foster a culture of evidence-based decision-making.

OBJECTIVES

- Clean and preprocess raw transactional data.
- Identify high-performing products, customers, and regions.
- Perform Exploratory Data Analysis (EDA) to spot patterns.
- Visualize findings using charts and maps.
- Build an interactive dashboard with Streamlit for user-friendly exploration.

RELEVANCE

Sales data is a rich source of intelligence. With over 530,000+ transactions, analyzing this dataset reveals operational bottlenecks, revenue-driving factors, and customer behavior—valuable for retail, marketing, and strategic planning.

DATASET DESCRIPTION

Dataset Overview

The dataset used in this project is titled "Sales Transaction v.4a.csv" and consists of over 536,000 individual transaction records. It reflects real-world e-commerce or retail business operations, capturing granular details about customer purchases over time.

This rich dataset forms the foundation for all downstream analysis, visualization, and dashboarding activities in the project.

Data Highlights

- Total Rows: 536,350
- Number of Unique Products: ~3,768
- Number of Unique Customers: ~5,800 (after filtering nulls)
- Number of Countries Represented: 38
- Top Country by Volume: United Kingdom (dominates with ~90% of transactions)
- Most Frequent Product: Cream Hanging Heart T-Light Holder

TECHNICAL FLOW

```
[Data Cleaning] → [Analysis] → [Visualization] → [Deployment]
   (Pandas)    (Stats/ML)   (Plotly/Seaborn)  (Streamlit + Ngrok)]
```

Workflow:

Data Cleaning → 2. Analysis → 3. Visualization → 4. Deployment

Tools:

Python Libraries: Pandas (cleaning), matplotlib,Seaborn/Plotly (visualization).

UI: Streamlit for interactive dashboards.

Deployment: Ngrok for secure tunneling.

METHODOLOGY

Data Cleaning & Preprocessing

- Converted Date column to datetime.
- Dropped null values in CustomerNo.
- Removed transactions with non-positive Quantity or Price.
- Added TotalSales column.

Exploratory Data Analysis (EDA)

- Computed descriptive stats: mean, median, frequency.
- Aggregated data by country, product, date, and customer.
- Identified top products, customers, and countries.

Tools Used

- Data Handling: pandas, numpy
- Visualization: matplotlib, seaborn, plotly
- Dashboard UI: streamlit
- Deployment Tools: ngrok for tunnel-based access

Statistical Techniques

- GroupBy and aggregation
- Pivot tables and cross-tabulation
- Time series decomposition (trend analysis)
- Pairwise correlation via heatmaps

Exploratory Data Analysis (EDA)

```
# Remove negative or zero quantities/prices
df = df[(df['Quantity'] > 0) & (df['Price'] > 0)]
```

```
# Preview cleaned data
df.head()
```

	TransactionNo	Date	ProductNo	ProductName	Price	Quantity	CustomerNo	Country	TotalSales
0	581482	2019-12-09	22485	Set Of 2 Wooden Market Crates	21.47	12	17490.0	United Kingdom	257.64
1	581475	2019-12-09	22596	Christmas Star Wish List Chalkboard	10.65	36	13069.0	United Kingdom	383.40
2	581475	2019-12-09	23235	Storage Tin Vintage Leaf	11.53	12	13069.0	United Kingdom	138.36
3	581475	2019-12-09	23272	Tree T-Light Holder Willie Winkie	10.65	12	13069.0	United Kingdom	127.80
		2019-		Set Of 4 Knick				United	

Remove negative or zero quantities/prices

```
# Summary stats for numerical columns
df.describe()
```

	Date	Price	Quantity	CustomerNo	TotalAmount
count	527764	527764.000000	527764.000000	527764.000000	5.277640e+05
mean	2019-07-04 05:58:58.445213952	12.629640	10.594679	15231.626733	1.193069e+02
min	2018-12-01 00:00:00	5.130000	1.000000	12004.000000	5.130000e+00
25%	2019-03-28 00:00:00	10.990000	1.000000	13813.000000	1.717000e+01
50%	2019-07-20 00:00:00	11.940000	3.000000	15159.000000	4.383000e+01
75%	2019-10-19 00:00:00	14.090000	11.000000	16729.000000	1.194000e+02
max	2019-12-09 00:00:00	660.620000	80995.000000	18287.000000	1.002718e+06
std	NaN	7.933224	156.786795	1716.522182	1.851192e+03

Summary stats for numerical columns

```
# Unique values
df.nunique()
```

	0
TransactionNo	19789
Date	305
ProductNo	3753
ProductName	3753
Price	514
Quantity	375
CustomerNo	4718
Country	38
TotalAmount	5626

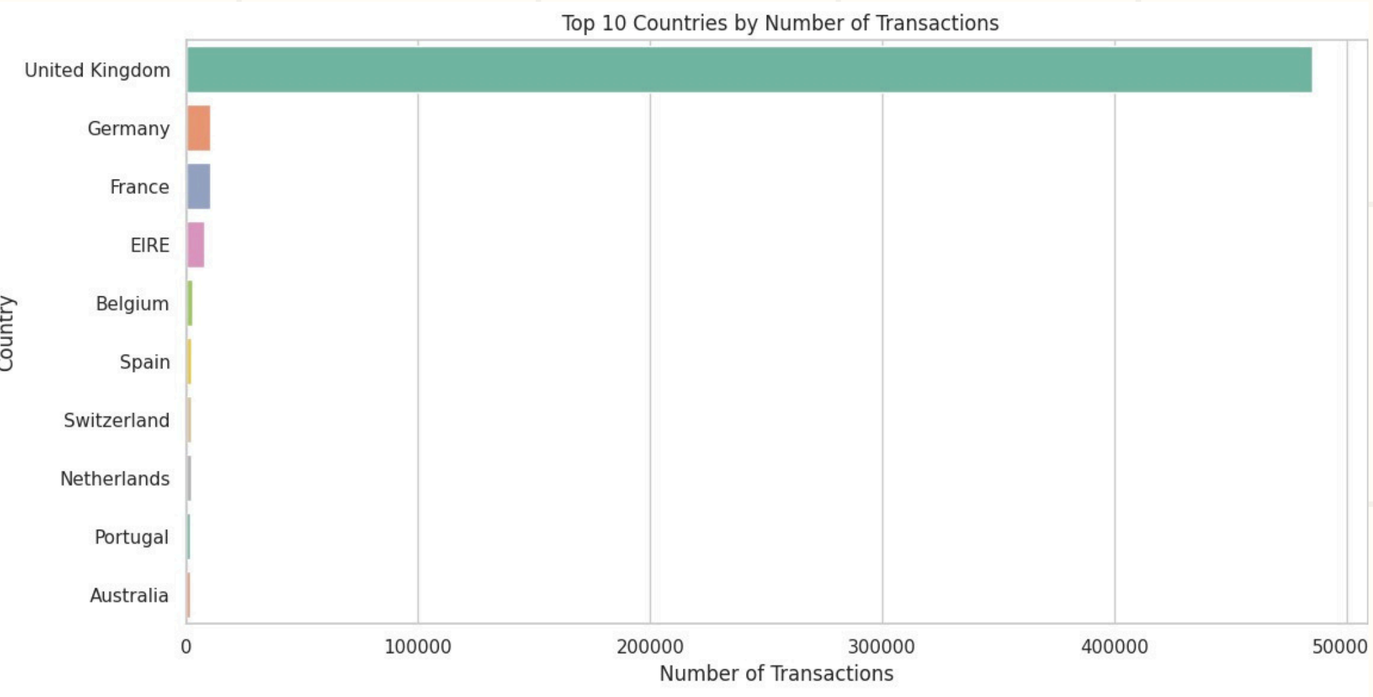
Unique values

```
(variable) df: DataFrame
df.isnull().sum()
```

	0
TransactionNo	0
Date	0
ProductNo	0
ProductName	0
Price	0
Quantity	0
CustomerNo	0
Country	0
TotalAmount	0

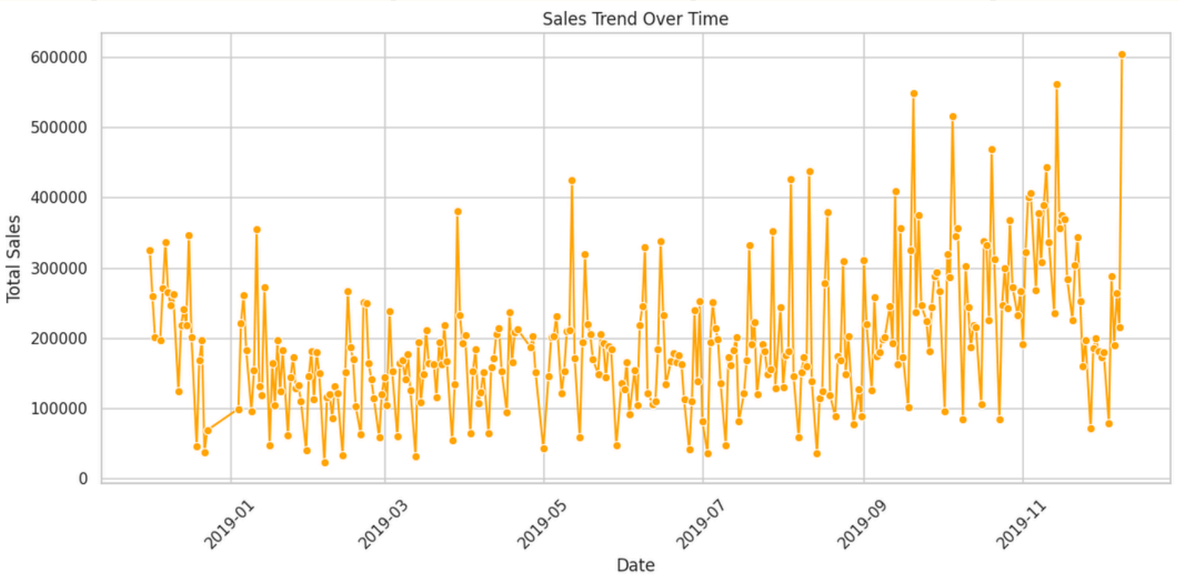
dtype: int64

Null value check



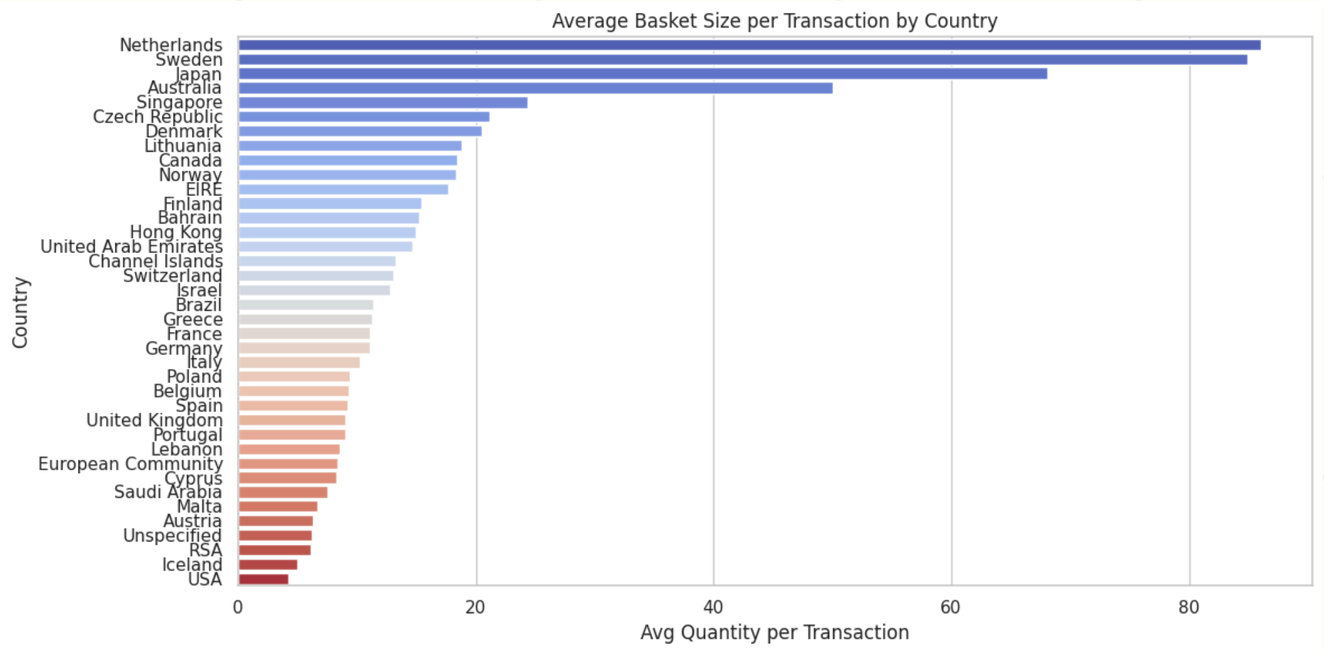
***Graph Inference (Top 10 Countries by Transactions)*:**

- Market Leaders: Identifies which countries have the highest purchase frequency (longest bars = most transactions).
- Growth Opportunities: Short bars reveal untapped markets needing
- Strategic Focus: Compare with revenue data—high transactions ≠ high sales (e.g., many small orders vs. few large ones).
- Actionable Insight: Prioritize marketing/operations in top countries while investigating low-volume regions.



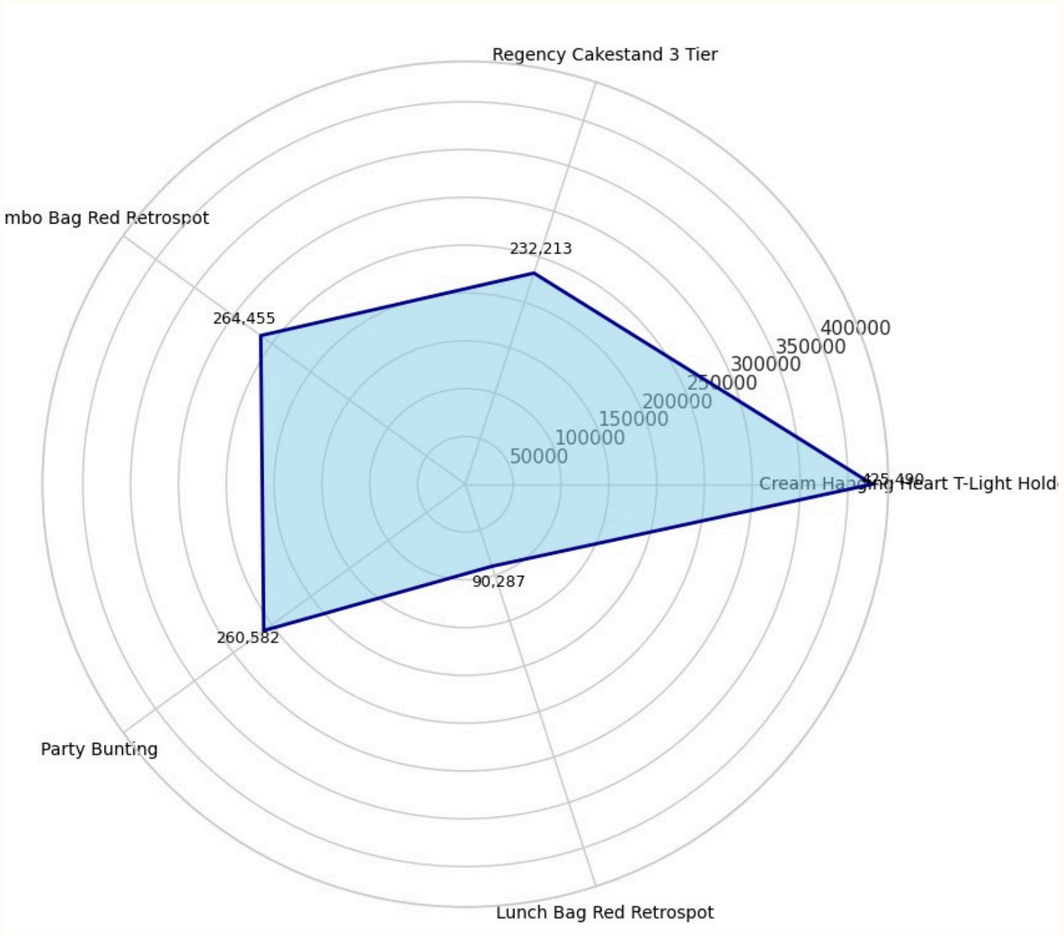
Sales Trends

1. Growth Direction - Up (good), Down (problem), Flat (stagnant)
2. Seasonal Peaks - Regular spikes show best sales periods
3. Special Events - Sudden jumps/drops reveal campaign impacts
4. Action Items - Double down on what works, fix what doesn't



Basket Size Graph Shows:

- Top Countries: Where customers buy the most items per order (best for upselling)
- Low Countries: Need strategies to increase order size (e.g., bundles, discounts)
- Action: Target high-basket countries with bulk deals, improve incentives in low-basket markets



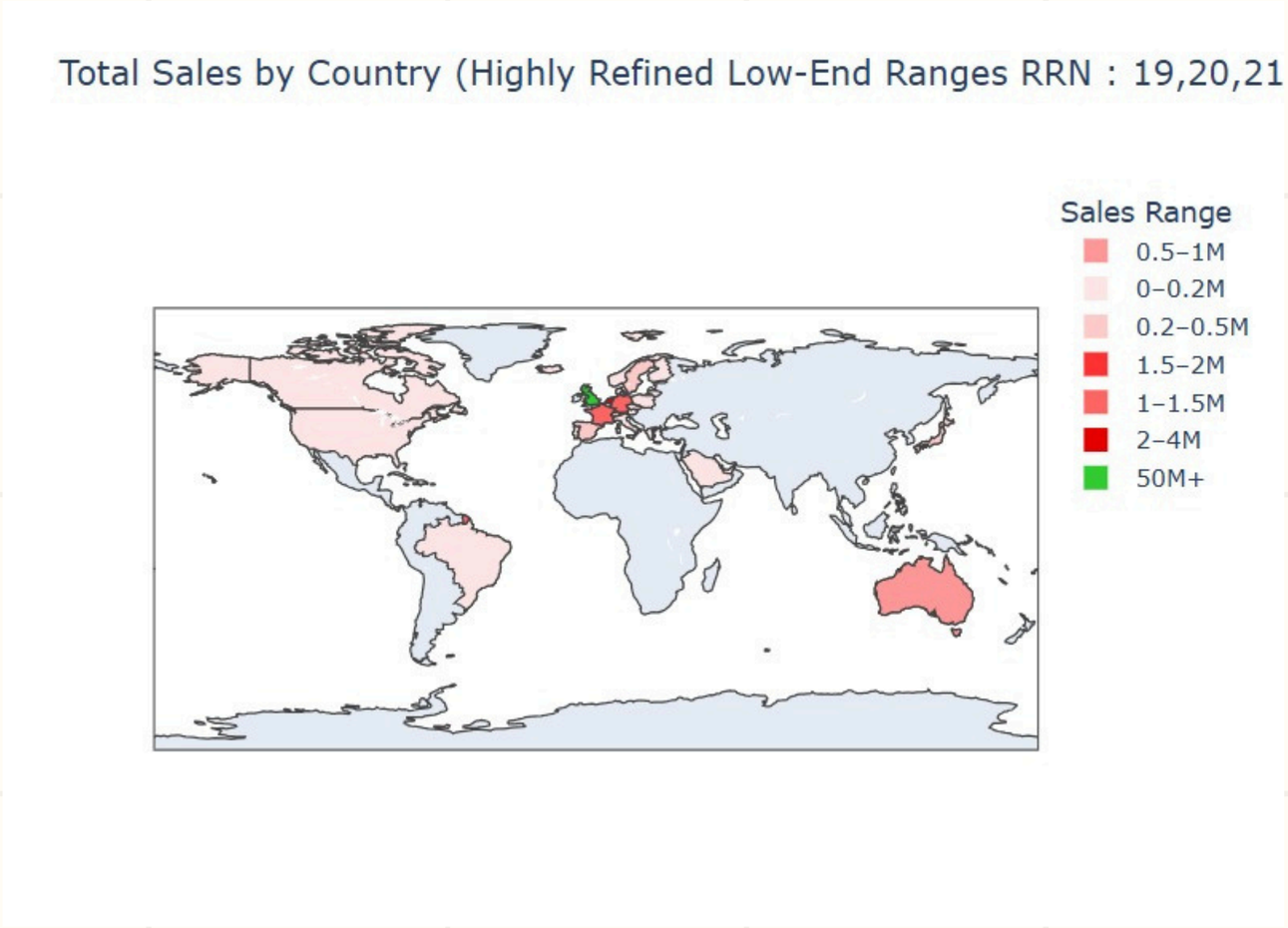
RADAR CHART

What it shows:

This radar chart highlights the performance of the top 5 best-selling products by visualizing their total revenue (or possibly quantity sold) across a radial plot. It allows us to quickly compare which products lead in sales and how the others fare in comparison.

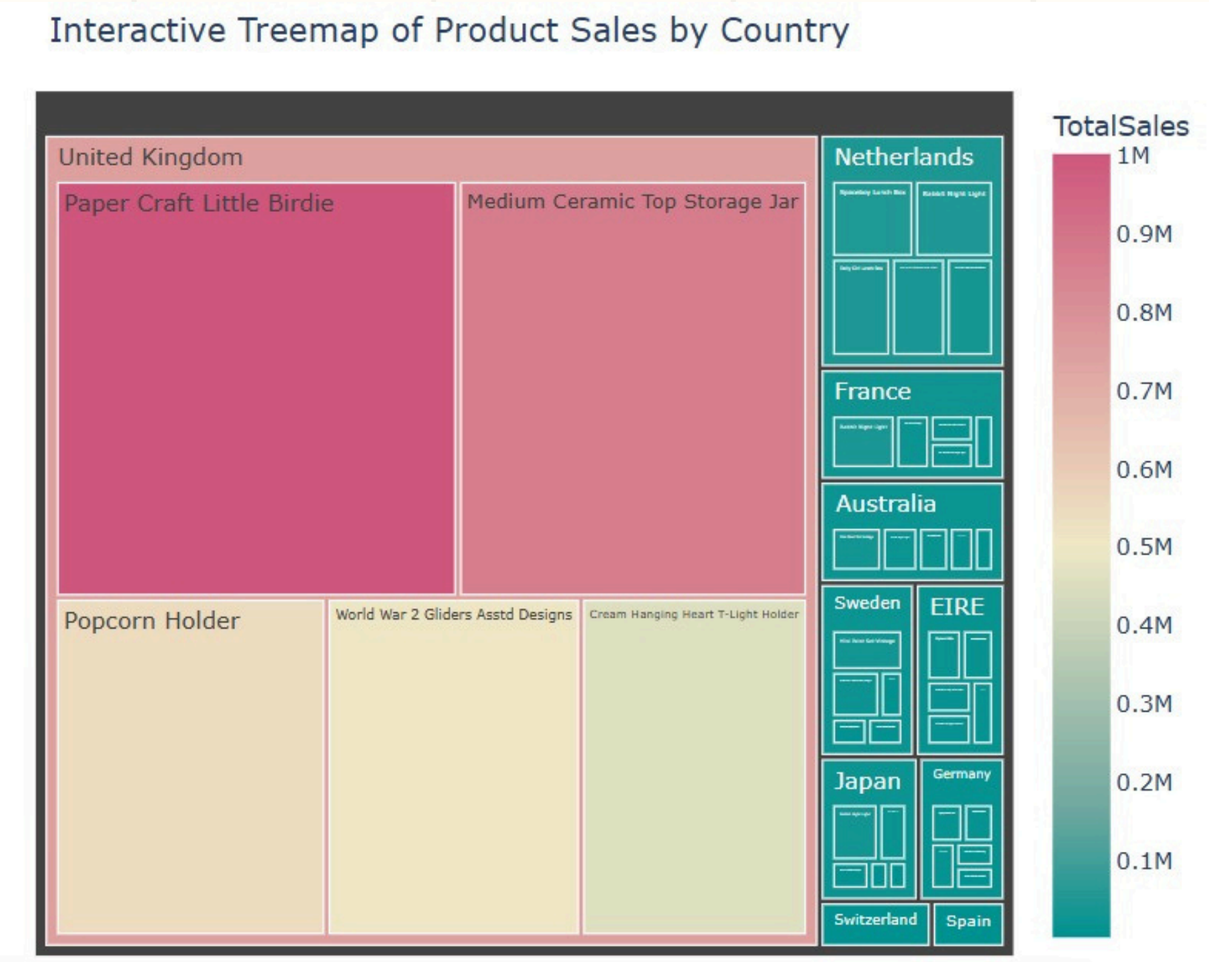
Products Featured:

- Cream Hanging Heart T-Light Holder - Highest sales (~439,494 units/revenue)
- Jumbo Bag Red Retrospot
- Party Bunting
- Regency Cakestand 3 Tier
- Lunch Bag Red Retrospot - Lowest among top 5 (~90,287)



Choropleth Map.

- The United Kingdom stands out with total sales exceeding 50 million, shown in green, far surpassing all other countries.
- Countries like the Netherlands and France fall into the 2-4M and 1.5-2M sales ranges, respectively.
- Australia, the United States, and a few European and South American countries show moderate sales between 0.5M to 1M.
- Most other regions, including Asia and Africa, reflect minimal sales activity, remaining in the 0-0.2M range.



TREE MAP

- The United Kingdom leads in product sales by a large margin compared to other countries.
- Top-selling products include "Paper Craft Little Birdie" and "Medium Ceramic Top Storage Jar" from the UK.
- Countries like the Netherlands, France, and Australia show significantly lower total sales
- The color intensity reflects sales volume, with the UK products nearing the 1M mark in total sales.

Key Insights

Seasonality & Trends

- Sales show spikes in late November and early December, possibly due to holiday season purchases.
- Weekday vs weekend analysis may show consistent low sales on Sundays.

Regional Performance

- Top Country: United Kingdom (over 90% of all transactions).
- Germany, France, and EIRE also showed notable revenue.
- Smaller countries like Norway and Austria contributed to niche markets.

Product Trends

- “Cream Hanging Heart T-Light Holder” was the most sold product.
- Products related to home decor and gifts performed best.

Additional Insight

- Basket size analysis revealed differences in average quantity bought per transaction across countries.
- Pairplot showed that while quantity and total sales correlate strongly, price has weaker correlation to volume.

CHALLENGES & SOLUTIONS

Challenge	Resolution
Missing CustomerNo values	Dropped such rows to maintain data quality
Negative values in Quantity/Price	Filtered out to retain valid transactional data
Date inconsistencies	Handled using errors='coerce' in pd.to_datetime()
Streamlit deployment from Colab	Resolved using cloudflared and ngrok tunnels for public access
Visual clutter	Limited plots to top 5 or 10 categories for clarity

CONCLUSION & FUTURE SCOPE

Summary of Takeaways

A robust data cleaning and visualization pipeline was established.

Interactive dashboards help business users explore insights without coding.

Foundational business questions around product demand and regional performance were answered.

Future Scope

Forecasting Models: Use Prophet or ARIMA for predicting future sales.

Customer Segmentation: Apply clustering (KMeans/DBSCAN) to segment shoppers.

Real-Time Dashboards: Integrate with live databases or APIs.

Alerting System: Auto-email stakeholders on threshold events like stockouts or sales spikes.

Enhanced UI: Add filters, download buttons, and drilldowns to Streamlit dashboard with open AI integration.


THANK YOU