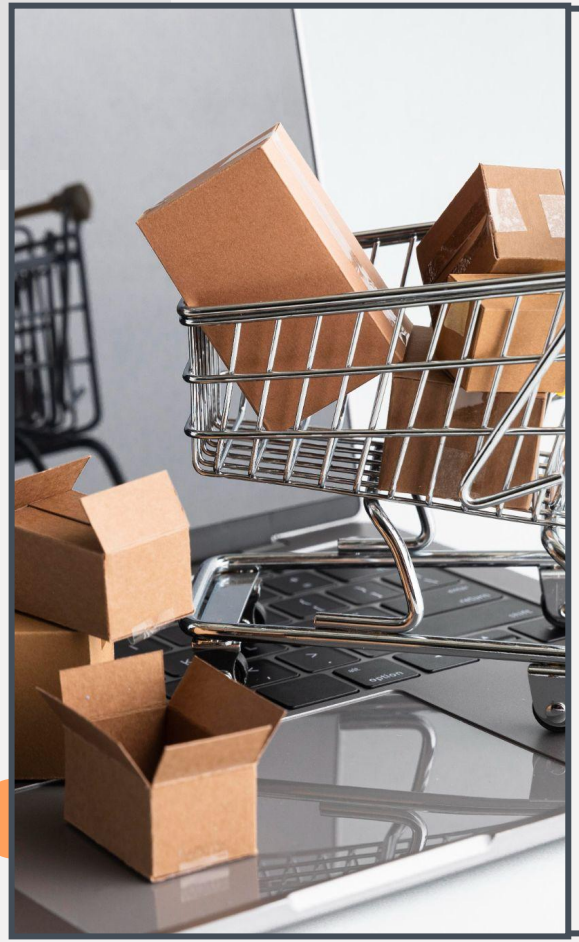


Unlocking Revenue: Which Factors Impact Retail Sales

Executor: Trần Thị Thuỳ Dung
Class: 12312024 - Data Analysis

Table of contents

01 Introduction

Purpose, outcomes, and dataset

Purpose statement

This project aims to **analyze available dataset, develop a robust predictive model for retail sales** and **provide actionable insights for optimizing sales revenue**. By leveraging historical sales, marketing, and discount data, I seek to understand the key drivers of sales performance.

Expected key outcomes:

- Sales evaluation: Analyze how features like marketing, discounts, holidays affect, etc. impact sales
- Sales prediction: Build model to predict retail sales
- Actionable suggestions: Propose reasonable actions to drive growth

Dataset description

1 Dataset name

Retail Sales Data with Seasonal Trends & Marketing

View on Kaggle

2 Description

This dataset provides detailed insights into retail sales, featuring a range of factors that may influence sales performance: units sold, discount, marketing, and holiday effect.

- Year of dataset: 2022 and 2023
- Data volume: 30,000 rows
- Numbers of column: 11

Dataset description

3

Data structure

01. **Store_ID:** Identifier for the retail store. (Categorical)
02. **Product_ID:** Identifier for the product. (Numerical)
03. **Date:** The date when the sale occurred. (Temporal - Key column)
04. **Units_Sold:** Quantity of items sold. (Numerical - Secondary Target Variable)
05. **Sales_Revenue_USD:** Total revenue generated from sales. (Numerical - Primary Target Variable)
06. **Discount_Percentage:** The percentage discount applied to products. (Numerical)
07. **Marketing_Spend_USD:** Budget allocated to marketing efforts. (Numerical)
08. **Store_Location:** Geographic location of the store. (Categorical) => Continent
09. **Product_Category:** The category to which the product belongs. (Categorical)
10. **Day_Week:** Day when the sale took place. (Categorical) => Weekday/Weekend
11. **Holiday_Effect:** Indicator of whether the sale happened during a holiday period. (Categorical/Binary)

Dataset description

3 Data structure

Store_ID	Product_ID	Date	Units_Sold	Sales_Revenue_USD	Discount_Percentage	Marketing_Spend_USD	Store_Location	Product_Category	Day_Week	Holiday_Effect
Spearsland	52372247	1/1/2022	9	2741.69	20	81	Tanzania	Furniture	Saturday	False
Spearsland	52372247	1/2/2022	7	2665.53	0	0	Mauritania	Furniture	Sunday	False
Spearsland	52372247	1/3/2022	1	380.79	0	0	Saint Pierre and Miquelon	Furniture	Monday	False
Spearsland	52372247	1/4/2022	4	1523.16	0	0	Australia	Furniture	Tuesday	False
Spearsland	52372247	1/5/2022	2	761.58	0	0	Swaziland	Furniture	Wednesday	False
Spearsland	52372247	1/6/2022	8	3046.32	0	41	Bhutan	Furniture	Thursday	False
Spearsland	52372247	1/7/2022	6	2284.74	0	0	Suriname	Furniture	Friday	False
Spearsland	52372247	1/8/2022	9	3427.11	0	83	Taiwan	Furniture	Saturday	False
Spearsland	52372247	1/9/2022	7	2665.53	0	0	Papua New Guinea	Furniture	Sunday	False
Spearsland	52372247	1/10/2022	1	380.79	0	164	Canada	Furniture	Monday	False

4 Tools

Google Colab: Handle data, test prediction models

PowerBI: EDA, visualize data

Github: Store project

02
Data handling

Data cleaning and re-structure
with Google Colab

Original dataset (11 columns)

Store_ID
object,
1 value only

Product_ID
int64,
42 unique values

Date
object,
731 unique values

Units_Sold
Int64
value: 0 - 56

Sales_Revenue_USD
float64,
Value: 0 - 27,165.88

Discount_Percentage
int64,
5 unique values

Marketing_Spend_USD
Int64
value: 0 - 199

Store_Location
object,
243 unique values

Product_Category
object,
4 unique values

Day_Week
object,
7 unique values

Holiday_Effect
bool,
value: True/False

Re-structure dataset (12 columns)

Product_ID
category,
42 unique values

Date
datetime64[ns],
731 unique values

Units_Sold
Int64
value: 0 - 56

Sales_Revenue_USD
float64,
Value: 0 - 27,165.88

Discount_Percentage
int64,
5 unique values

Marketing_Spend_USD
Int64
value: 0 - 199

Store_Location → **Continent**
category,
243 unique values

category,
5 unique values

Product_Category
category,
4 unique values

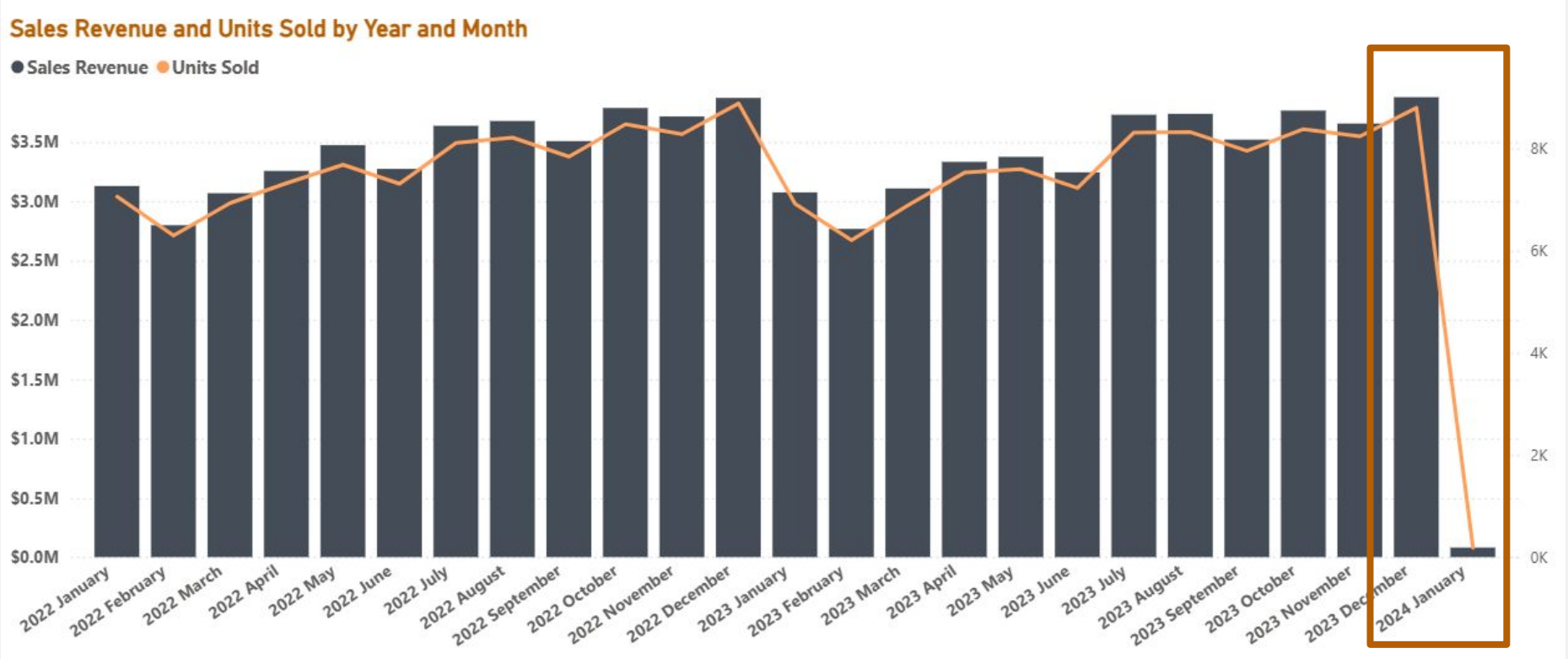
Day_Week → **Is_Weekend**
category,
7 unique values

category,
value: Weekday/Weekend

Holiday_Effect
bool,
value: True/False

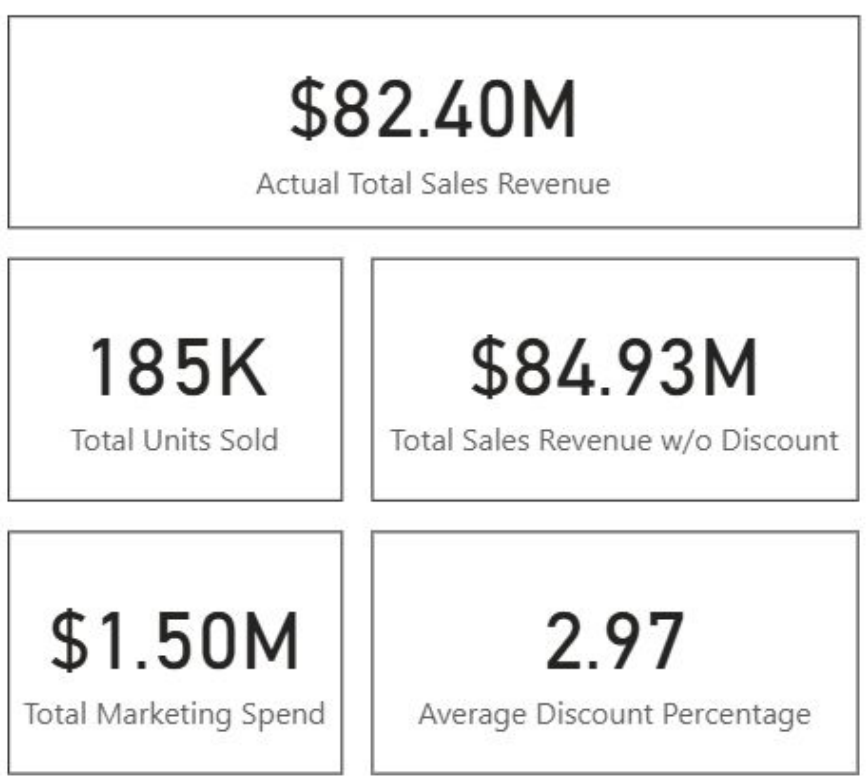
03
EDA

Exploratory Data Analysis
with Power BI



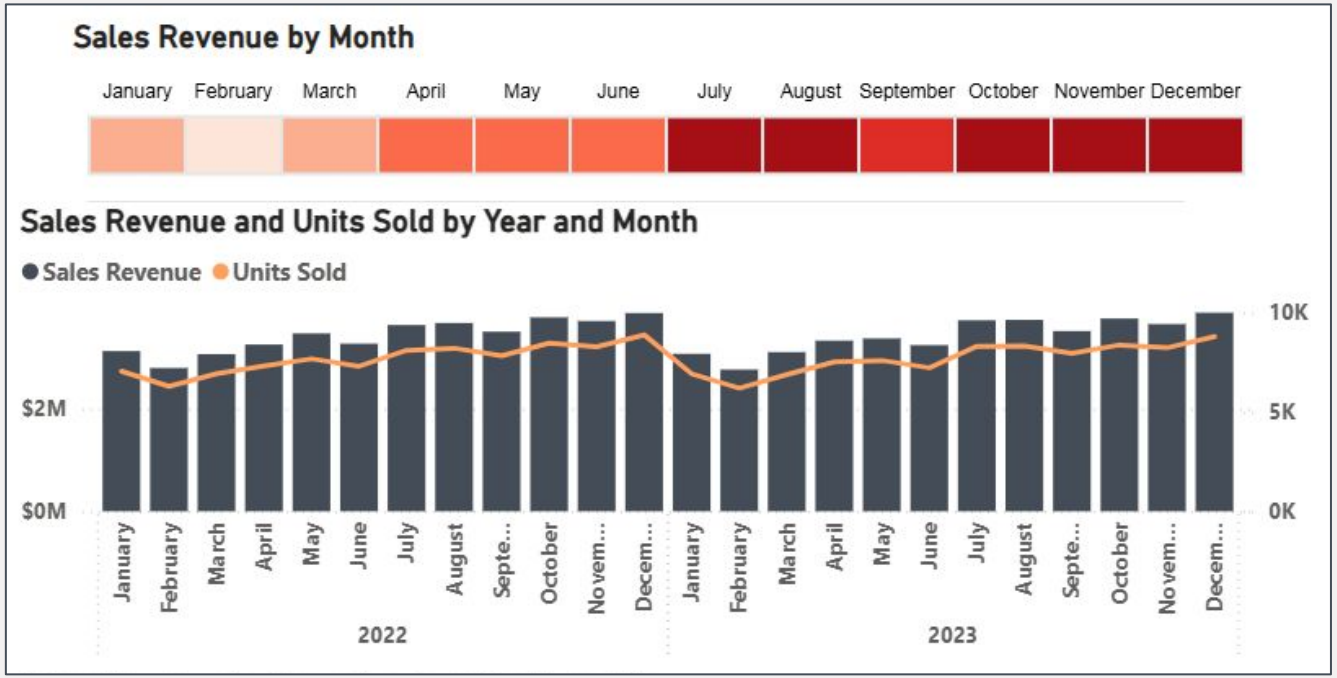
Sales in January 2024 appear unusually low because data was only collected up to January 1, 2024. For a clearer analysis, we **excluded this date from the EDA**.

Overview



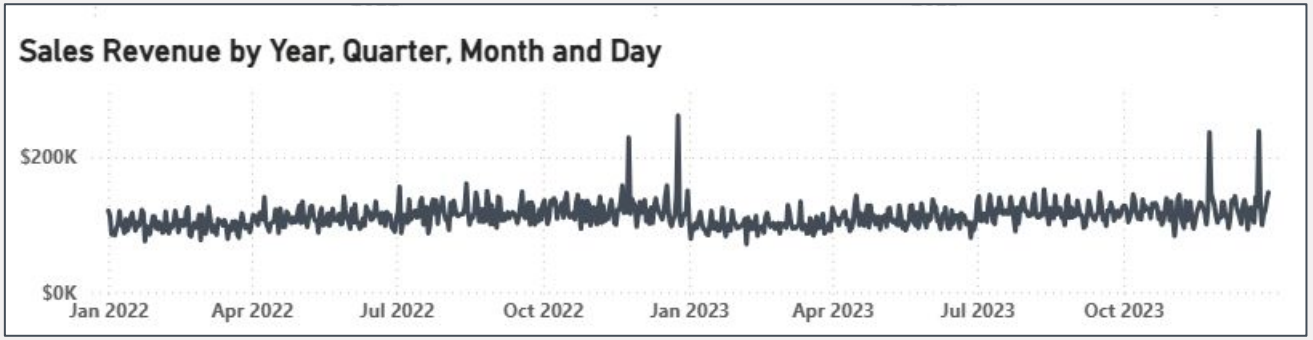
- $84.9M in sales without discount, which means about **$2.5M was given away through discounts**.
- $1.5M spent on **marketing**, which is **relatively small compared to total sales revenue**.
- And the average discount percentage is only 2.97%, suggesting **discounts are used sparingly**.

Sales revenue

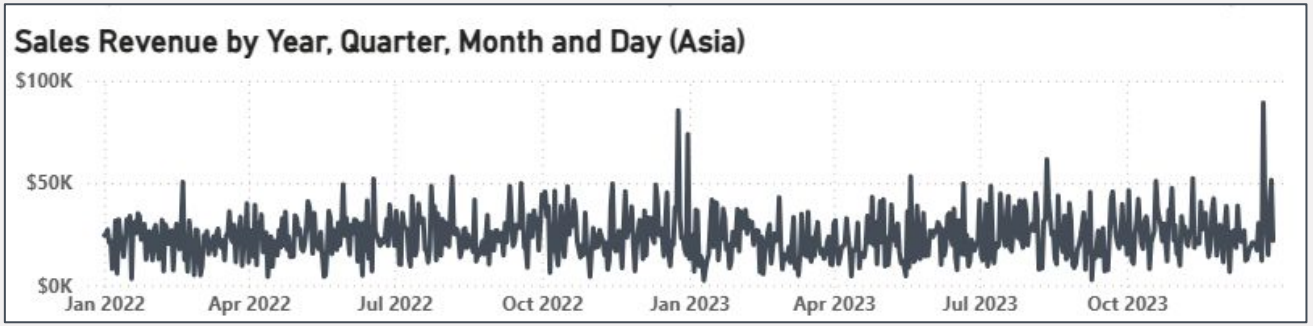


The second half of year had better sales performance. There was a noticeable dip in February both years. Units sold and revenue tend to move together, showing a healthy correlation.

Sales revenue



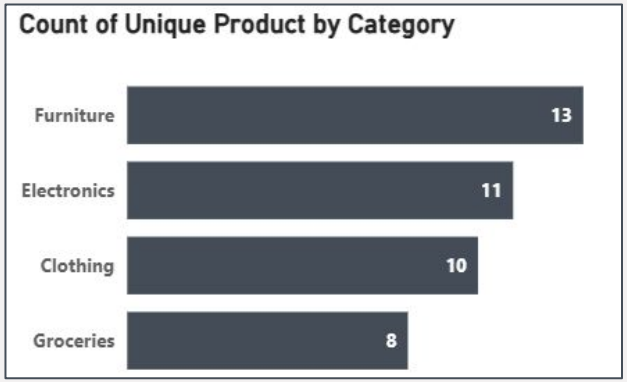
General: Peak at Thanksgiving and Christmas



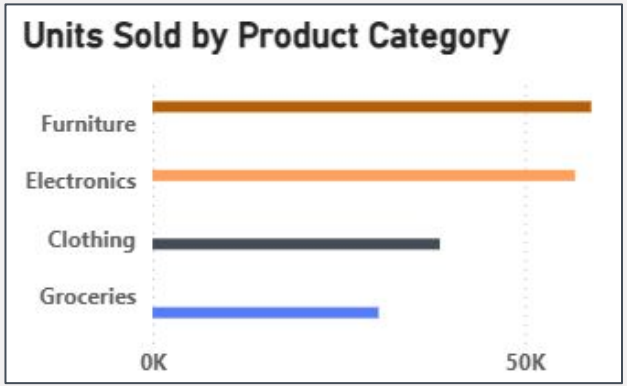
Asia: Peak before Lunar New Year

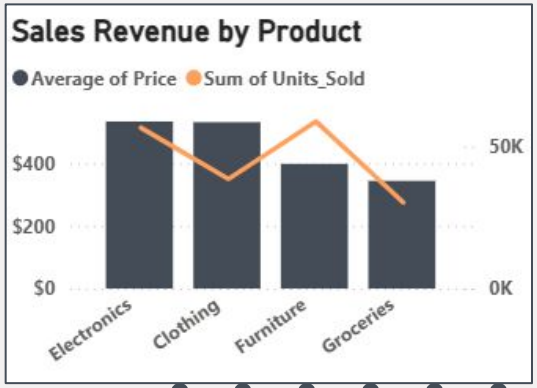
The fluctuate daily sales amount highlights occasional revenue spikes, driven by big holidays of each areas.

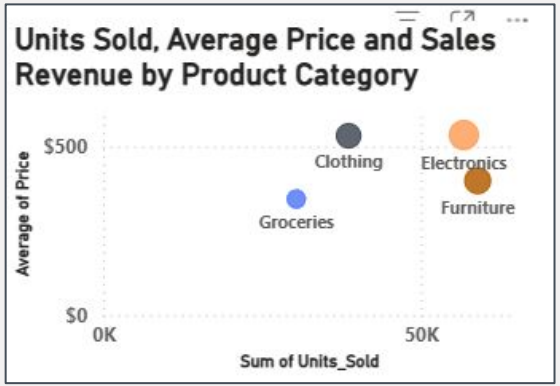
Product Category



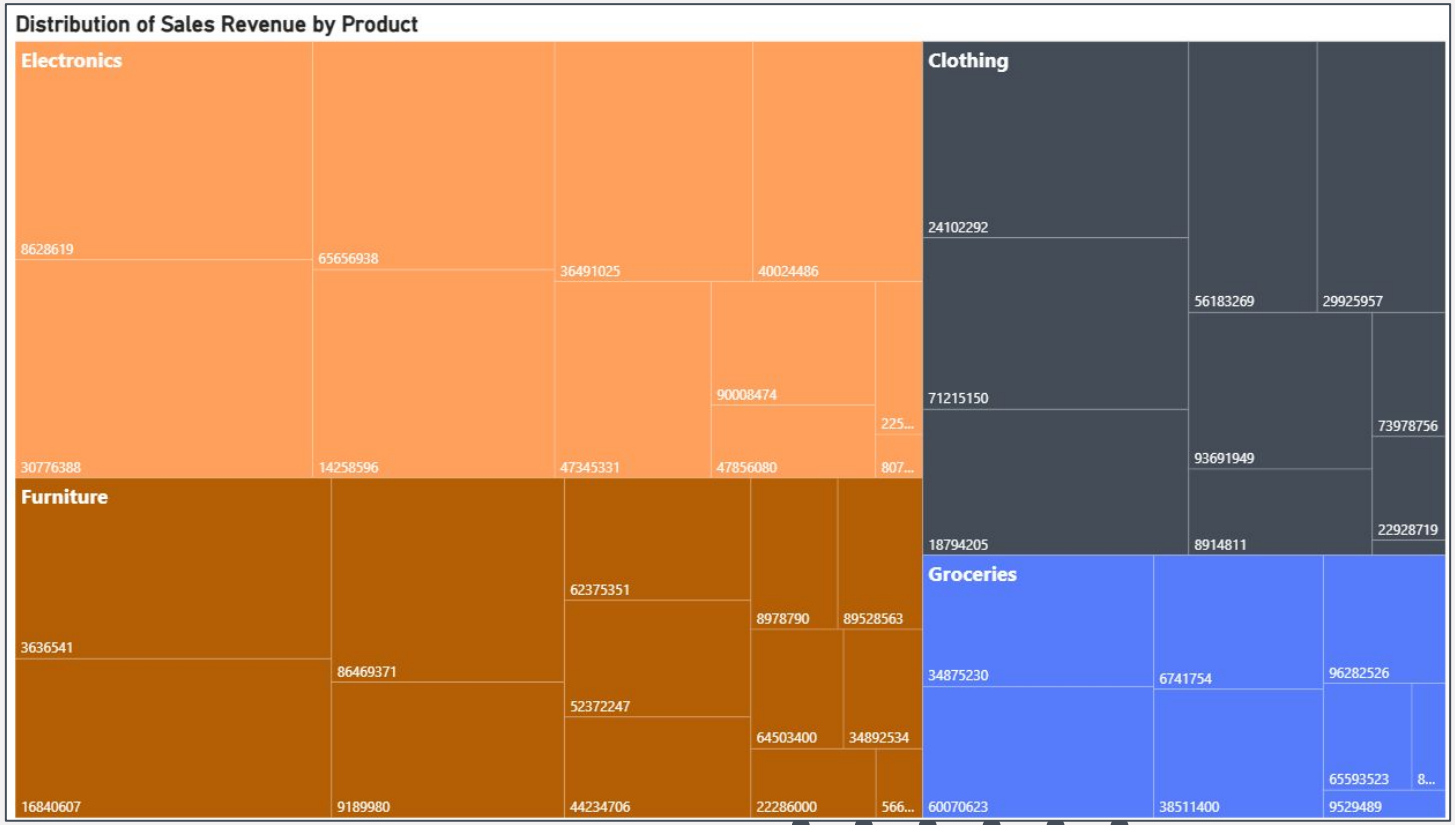
- **Furniture** led with 13 unique products, following by **Electronics** and **Clothing**.
- **Furniture and Electronics sold the most units**, while Clothing had fewer units but a higher average price.
- **Groceries sold in large volumes but at lower prices**, which explains their smaller revenue share.







Product

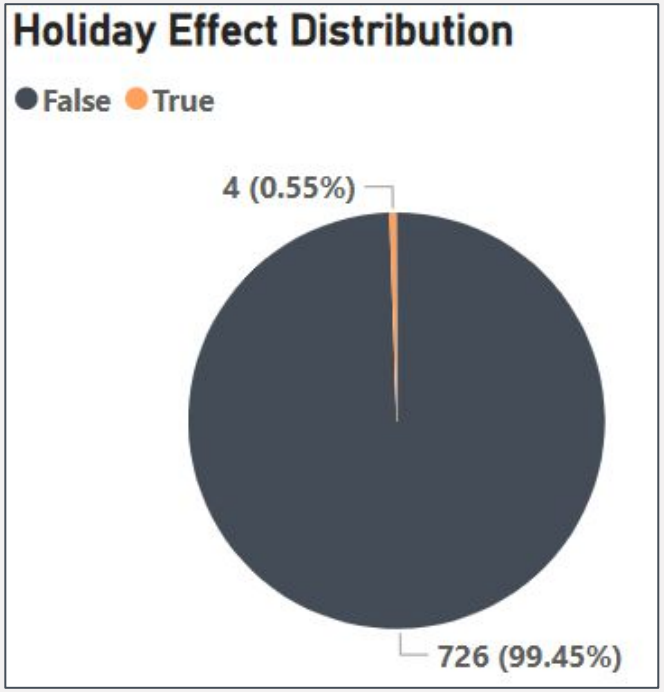


Dataset did **not provide the exact products**, just IDs.

Revenue was concentrated in a few high performing products, especially in **Electronics** & **Furniture**.

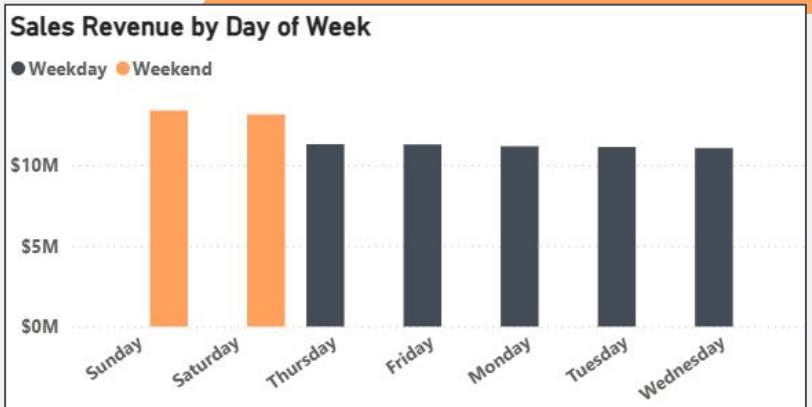
Holiday

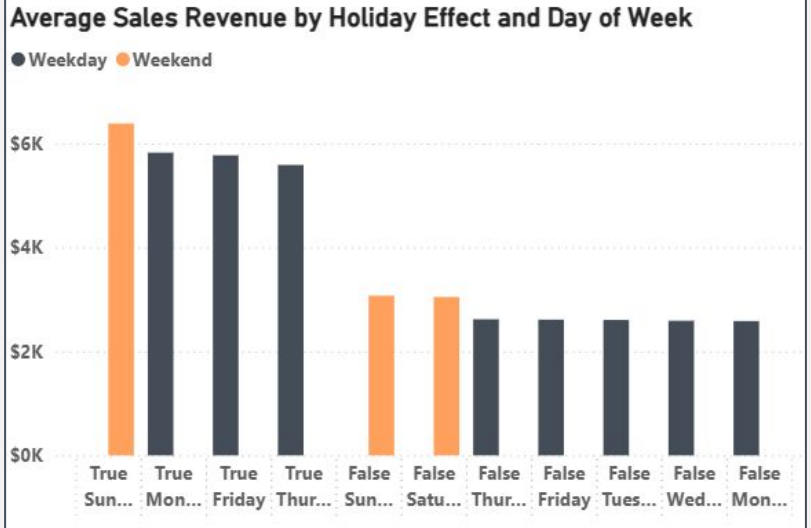
- The holiday effect is mostly 'False' in 99.45% of the cases, which is expected, as holidays are not daily occurrences.
- The 4 holidays counted were Thanksgiving and Christmas.



Day of the Week

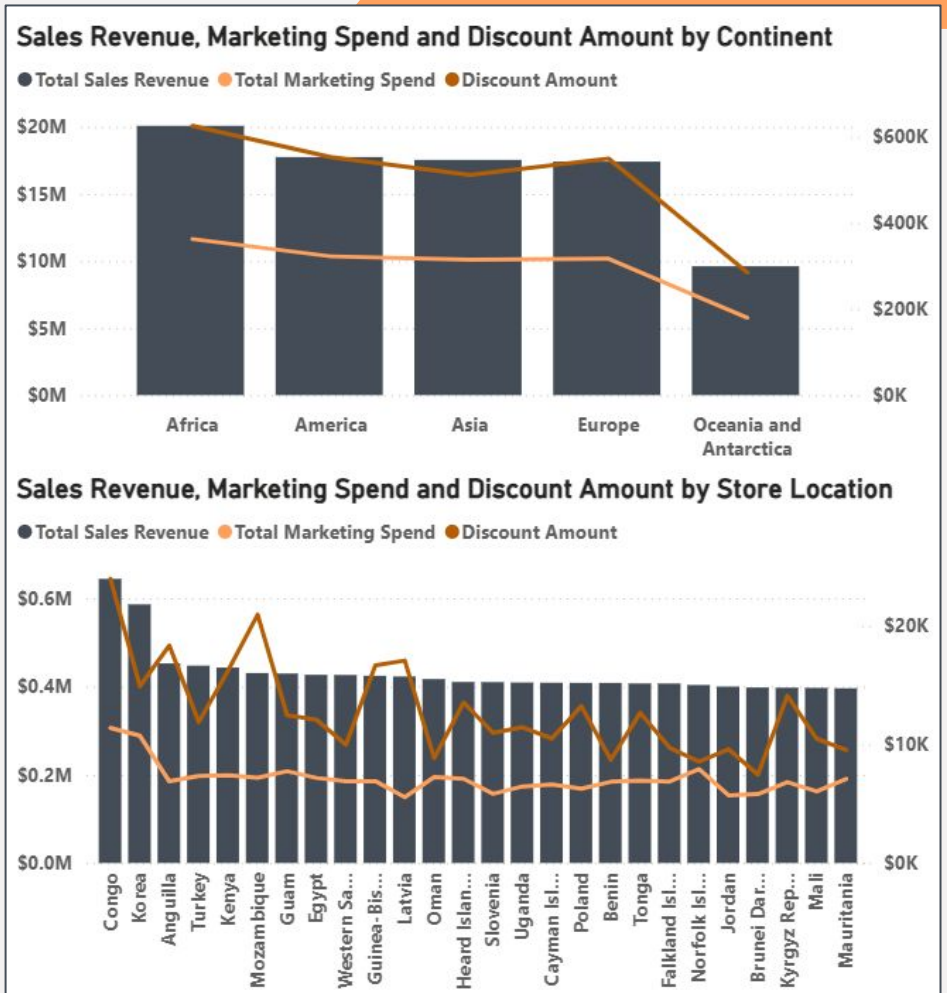
- **Sales peak on weekends.** Weekdays have similar and just slightly lower revenue levels.
  ⇒ **Weekend did not have much impact on revenue.**
- Sunday holidays drive the highest average revenue and all average sales amount on holidays was higher than the other days.
  ⇒ **Holidays significantly boosted sales** although there were 4 holidays counted.



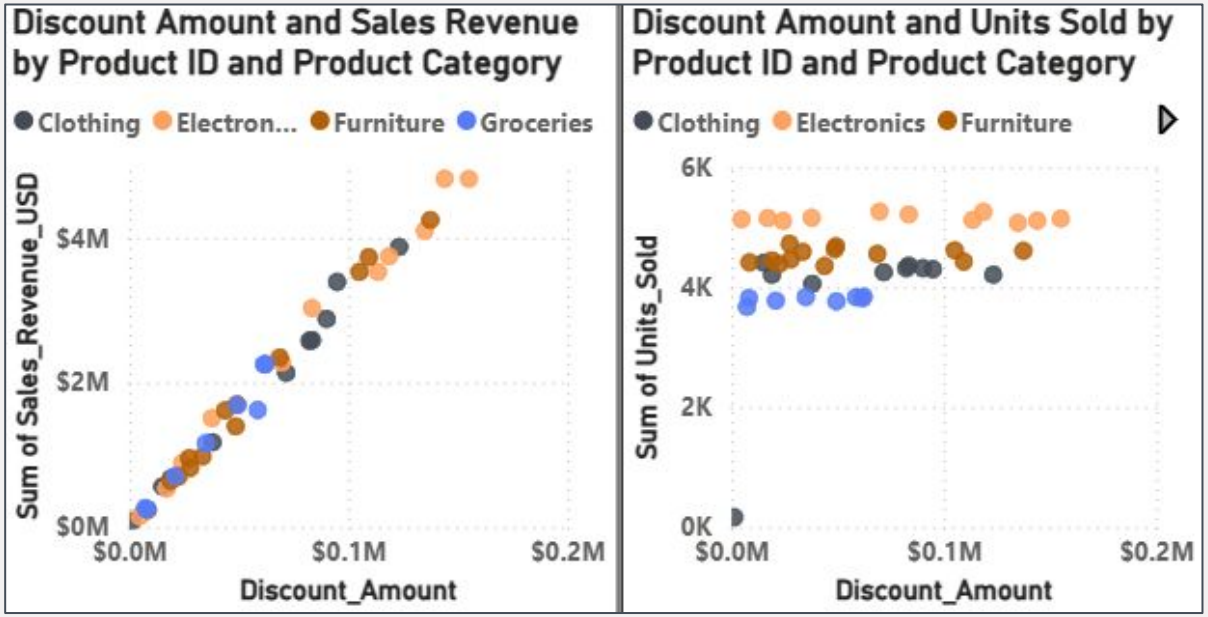


Location

- Africa led in total sales revenue (nearly 20M USD), despite moderate marketing spend. Oceania and Antarctica generated the least revenue (nearly 10M USD).
- Congo and Korea have the highest store-level sales (around 0,6M USD). Discount amounts varied but didn't strongly correlate with revenue.



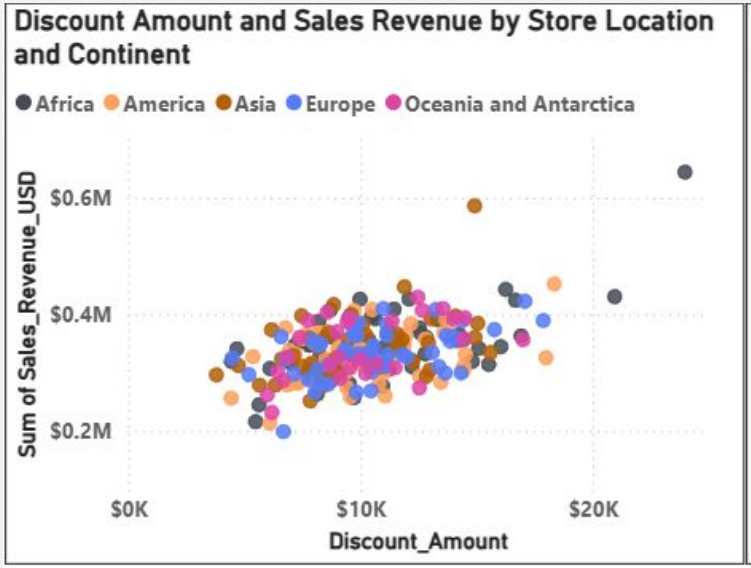
Sales & Discount (by Product)

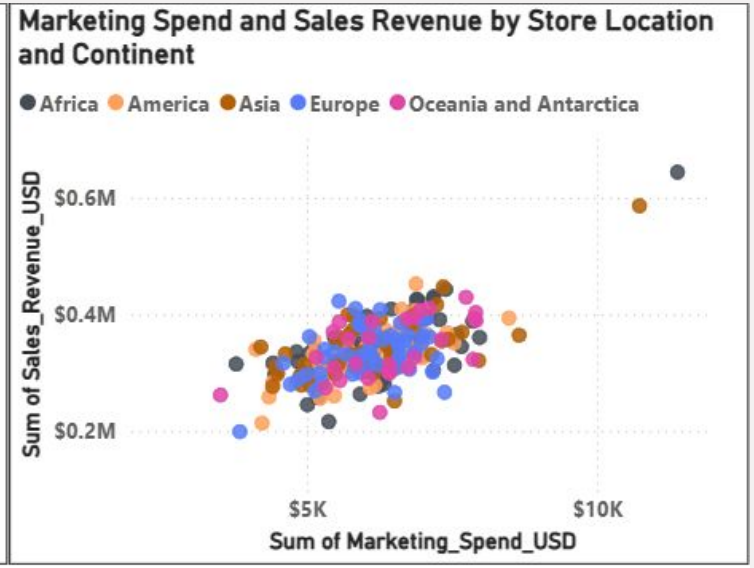


There's a positive relationship between discount amount and both sales revenue & units sold, especially with Electronics and Furniture.

Groceries show more stable sales, even at lower discount levels.

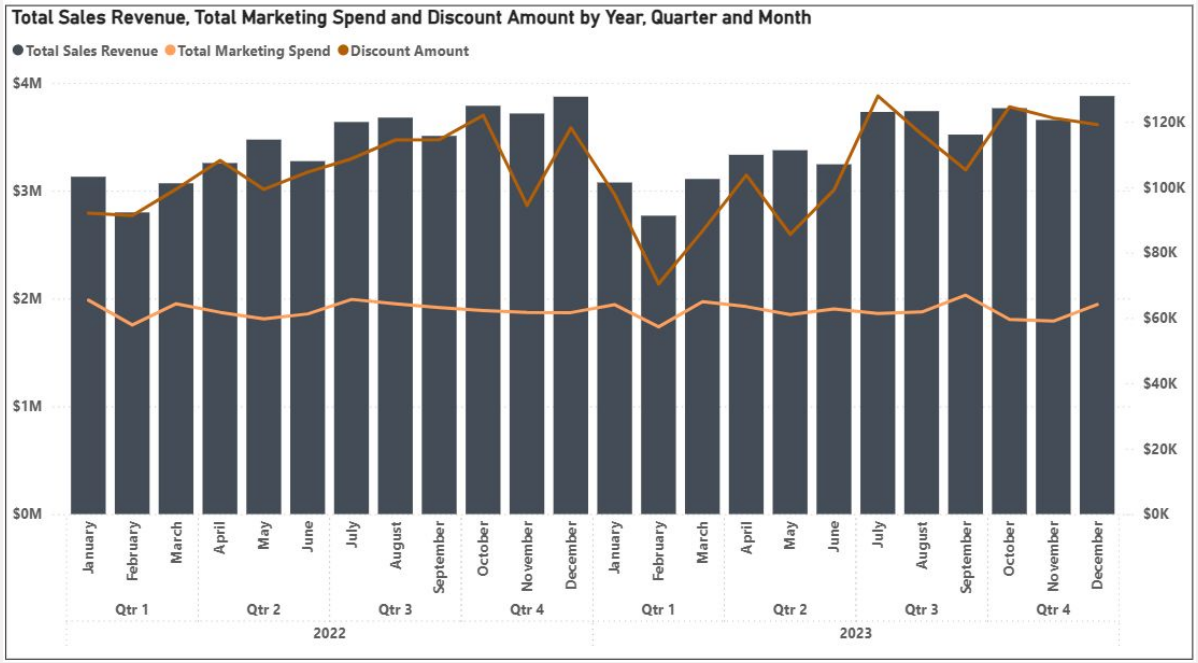
Sales, Marketing & Discount (Location)





In general, higher discounts and marketing spend lead to higher revenue across store locations and continents, suggesting the impact of promotions is geographically broad.

Sales, Marketing & Discount

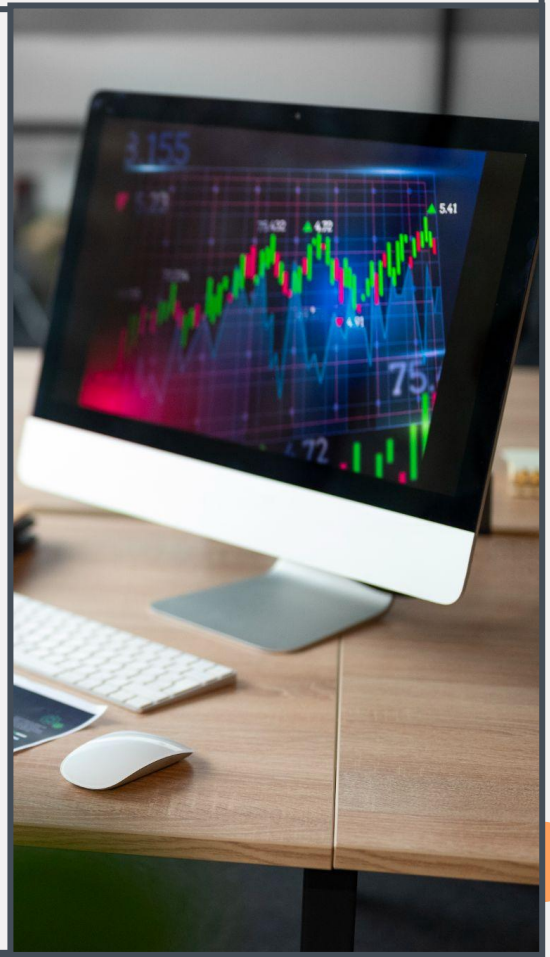


Both discount and marketing spend showed no significant upward or downward trend. ⇒ Consistent promotional efforts over time.

Revenue increased with marketing spend, though the relationship was not as strong as with discounts. ⇒ Discount had a more direct impact.

Preliminary conclusion 1

- Higher sales performance on weekends and holidays.
- Potential positive impact from both discount and marketing spend on sales revenue.
- The seasonal peak in late fall & winter and high revenue within Africa and America stores suggest some key insights for sales growth.



04
Sales prediction

Data encoding, correlation, prediction models with Google Colab

04a. Data encoding

Remove redundant columns

The **“Continent”** and **“Is_Weekend”** columns were created to simplify the granular data in the **“Store_Location”** and **“Day_Week”** columns.

When building a predictive model, it’s best to remove these original columns to **avoid data redundancy and potential multicollinearity**.

Remove redundant columns (10 columns)

Product_ID
category,
42 unique values

Date
datetime64[ns],
731 unique values

Units_Sold
Int64
value: 0 - 56

Sales_Revenue_USD
float64,
Value: 0 - 27,165.88

Discount_Percentage
int64,
5 unique values

Marketing_Spend_USD
Int64
value: 0 - 199

~~**Store_Location**~~
~~category,~~
243 unique values

Continent
category,
5 unique values

Product_Category
category,
4 unique values

~~**Day_Week**~~
~~category,~~
7 unique values

Is_Weekend
category,
value: Weekday/Weekend

Holiday_Effect
bool,
value: True/False

Encode data

Encode columns without numeric data type
(Product_Category, Continent, Is_Weekend, Holiday_Effect)
and convert new columns into numeric data type.

Encoded dataset (19 columns)

Product_ID	Date	Units_Sold	Sales_Revenue_USD	Discount_Percentage	Marketing_Spend_USD
category	datetime64[ns]	Int64	float64,	int64,	Int64

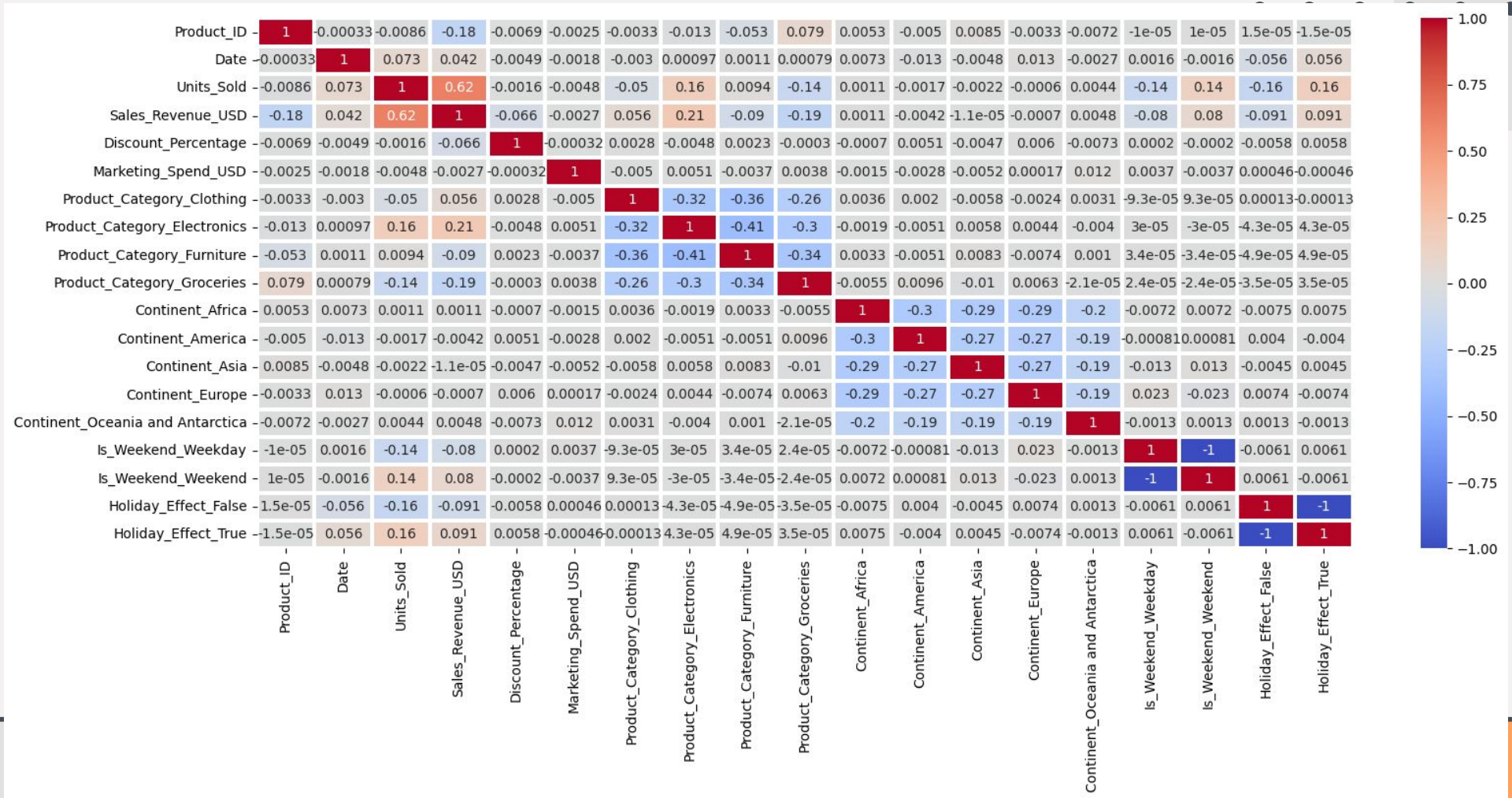
Product_Category_Clothing	Product_Category_Electronies	Product_Category_Furniture	Product_Category_Groceries
int64	int64	int64	int64

Continent_Africa	Continent_America	Continent_Asia	Continent_Europe	Continent_Oceania and Antarctica
int64	int64	int64	int64	int64

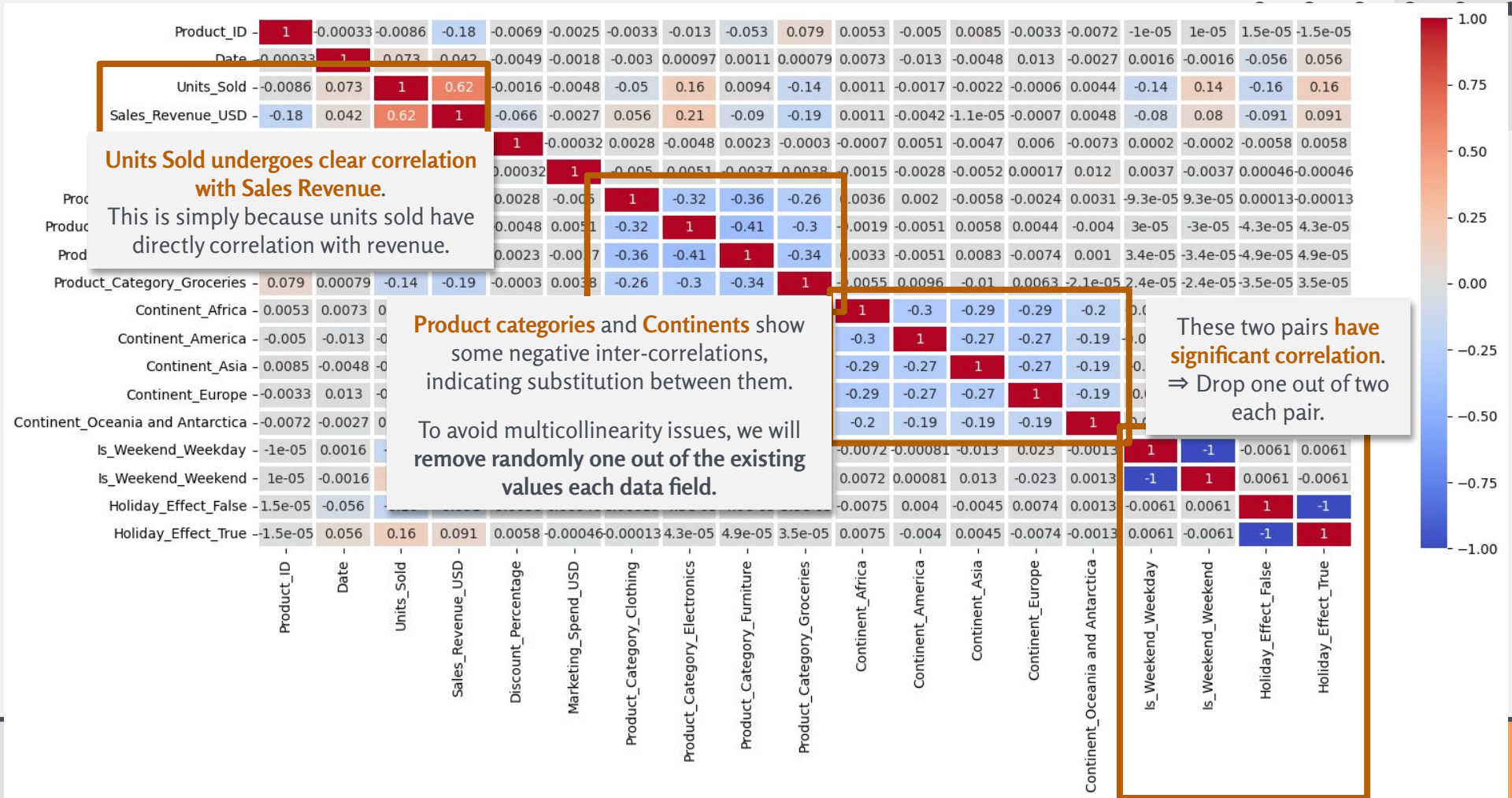
Is_Weekend_Weekday	Is_Weekend_Weekend	Holiday_Effect_False	Holiday_Effect_True
int64	int64	int64	int64

04b. Correlation

Part 04b. Correlation



Part 04b. Correlation



Units Sold undergoes clear correlation with Sales Revenue.
This is simply because units sold have directly correlation with revenue.

Product categories and **Continents** show some negative inter-correlations, indicating substitution between them.

To avoid multicollinearity issues, we will **remove randomly one out of the existing values each data field.**

These two pairs **have significant correlation.**
⇒ Drop one out of two each pair.

New dataset (15 columns)

Product_ID	Date	Units_Sold	Sales_Revenue_USD	Discount_Percentage	Marketing_Spend_USD
category	datetime64[ns]				

Product_Category_Clothing	Product_Category_Electronies	Product_Category_Furniture	Product_Cate ory_Groceries

Continent_Africa	Continent_America	Continent_Asia	Continent_Europe	Continent_Ocea a and Antarctica

Is_Weeken Weekday	Is_Weekend_Weekend	Holiday_E ect_False	Holiday_Effect_True

Check multicollinearity by VIF

	feature	VIF
0	const	19.590899
1	Product_ID	1.007408
2	Units_Sold	1.091546
3	Holiday_Effect_True	1.028226
4	Discount_Percentage	1.000227
5	Marketing_Spend_USD	1.000259
6	Product_Category_Clothing	1.673054
7	Product_Category_Electronics	1.812752
8	Product_Category_Furniture	1.828297
9	Continent_Africa	2.354640
10	Continent_America	2.258110
11	Continent_Asia	2.241632
12	Continent_Europe	2.235936
13	Is_Weekend_Weekend	1.022886

All factors have low VIFs (around from 1 to 2.4).

⇒ **Multicollinearity is not a problem.**

04c. Sales forecasting

Prediction models

1. OLS Linear Regression,
2. Ridge,
3. Lasso,
4. Elastic Net,
5. SGDRegressor,
6. Random Forest Regressor,
7. Gradient Boosting Regressor

Calculate MSE and R^2

```
OLS LinearRegression
  Mean Squared Error: 3928865.64
  R²: 0.41
Ridge
  Mean Squared Error: 3928872.15
  R²: 0.41
Lasso
  Mean Squared Error: 3928835.66
  R²: 0.41
ElasticNet
  Mean Squared Error: 3956505.75
  R²: 0.41
SGDRegressor
  Mean Squared Error: 16646754423837336618926080
  R²: -248070496981847968.00
RandomForest
  Mean Squared Error: 4487654.83
  R²: 0.33
GradientBoosting
  Mean Squared Error: 3894043.92
  R²: 0.42
```

All R^2 values are quite low (< 0.50).

Linear Models (4 first models): Perform similarly and moderately. Reasonable fit but room for improvement.

SGDRegressor: Failed model, extremely high error, negative R^2.

Tree-Based Models: While Random Forest has low performance, **Gradient Boosting provides best overall and slight improvement over linear models.**

Predict sales revenue

	Actual Sales	Predicted (OLS LinearRegression)	Predicted (Ridge)	Predicted (Lasso)	Predicted (ElasticNet)	Predicted (RandomForest)	Predicted (GradientBoosting)
2308	5424.03	1707.497057	1708.448150	1709.708216	2243.239509	1735.729762	1819.230727
22404	5503.23	2932.206595	2932.119365	2932.425539	2835.421478	2951.824569	2902.601828
23397	334.80	1368.215910	1369.233895	1370.546772	1920.600898	784.880715	1534.866064
25058	11591.76	4052.618835	4052.423065	4052.410423	3794.453038	4308.165371	4174.508701
2664	572.54	1570.048882	1571.007946	1572.271718	2107.092405	2040.788879	1693.478320
8511	4851.05	4035.027883	4034.758878	4034.792928	3766.644061	3843.244330	4108.909475
5148	4337.55	2293.365390	2293.466919	2293.805047	2325.177515	5950.527887	2356.764873
7790	91.14	3584.116670	3584.041413	3584.418893	3419.890973	3609.990601	3594.896400
11311	957.85	1674.031184	1674.991419	1676.233351	2215.167514	1715.628498	1791.188595
19043	1038.40	2315.625140	2315.656928	2316.003852	2329.580237	1034.972398	2173.281407

All remaining models provide **similar predictive values** which have large differences with atual number, with the exception of **Random Forest.**

OLS Linear Regression

OLS Regression Results

Dep. Variable:	Sales_Revenue_USD	R-squared:	0.414
Model:	OLS	Adj. R-squared:	0.414
Method:	Least Squares	F-statistic:	1412.
Date:	Thu, 21 Aug 2025	Prob (F-statistic):	0.00
Time:	15:35:18	Log-Likelihood:	-2.1602e+05
No. Observations:	24000	AIC:	4.321e+05
Df Residuals:	23987	BIC:	4.322e+05
Df Model:	12		
Covariance Type:	nonrobust		

41.4% of the variance in revenue can be explained by the variables in the model.

The model as a whole is **statistically significant**.

OLS Linear Regression

	coef	std err	t	P>|t|	[0.02
const	-502.7767	51.300	-9.801	0.000	-603.32
Units_Sold	457.2574	3.975	115.036	0.000	449.46
Discount_Percentage	-27.5957	2.112	-13.065	0.000	-31.73
Marketing_Spend_USD	-0.0916	0.197	-0.465	0.642	-0.47
Product_Category_Clothing	972.1700	39.572	24.567	0.000	894.60
Product_Category_Electronics	1046.9112	38.418	27.250	0.000	971.60
Product_Category_Furniture	197.0707	36.701	5.370	0.000	125.13
Continent_Africa	-40.5926	45.106	-0.900	0.368	-129.00
Continent_America	-57.3383	46.021	-1.246	0.213	-147.54
Continent_Asia	-21.3292	46.095	-0.463	0.644	-1
Continent_Europe	-28.0698	46.247	-0.607	0.544	-1
Is_Weekend_Weekend	-18.7640	28.270	-0.664	0.507	-
Holiday_Effect_True	-100.6282	168.873	-0.596	0.551	-4

Omnibus:	1373.103	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4599.875
Skew:	0.227	Prob(JB):	0.00
Kurtosis:	5.096	Cond. No.	1.09e+03

Each additional **unit sold**, revenue is expected to increase by **~ 457.2 USD**.

A **higher percentage discount** might lead to **lower revenue** per transaction, even if it drives higher unit sales.

Electronics products generate about **~1,046.91 USD** more in sales than Groceries, **clothing products** generate **~972 USD** and **furniture products** generate **~197 USD** more.

The remaining variances do not have have a statistically significant relationship with sales revenue in this mode.

In real-world scenario

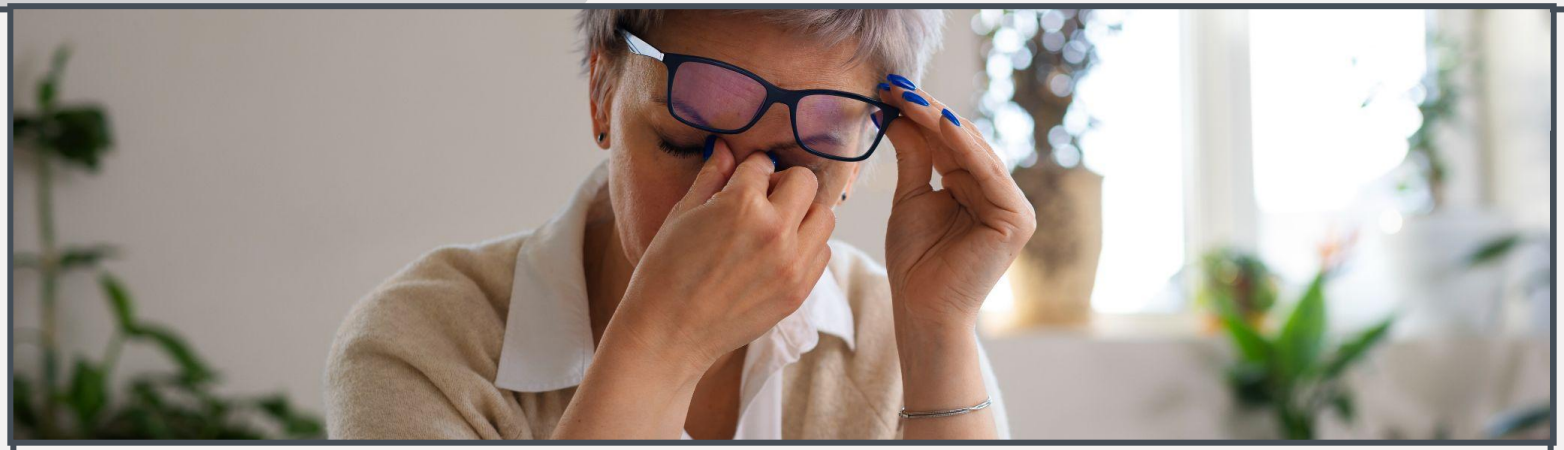
In reality, **we do not have Units Sold in advance** to predict sales amount. Thus, I will remove this variance to exam the prediction model change.

In real-world scenario

```
OLS LinearRegression
  Mean Squared Error: 6111662.81
  R²: 0.09
Ridge
  Mean Squared Error: 6111657.92
  R²: 0.09
Lasso
  Mean Squared Error: 6111612.09
  R²: 0.09
ElasticNet
  Mean Squared Error: 6223615.60
  R²: 0.07
SGDRegressor
  Mean Squared Error: 13243672826309822602280960.00
  R²: -197357660012283936.00
RandomForest
  Mean Squared Error: 7138826.96
  R²: -0.06
GradientBoosting
  Mean Squared Error: 6100406.65
  R²: 0.09
```

All R^2 values through all models drop significantly after removing Units Sold (from > 0.4 to < 0.1).

⇒ This result once again confirms that **Units Sold is a key driver of revenue prediction**.



04d.
Preliminary conclusion 2

Preliminary conclusion 2

Prediction models: **Gradient Boosting** provides the best predicted sales. However, it has not much difference from the other linear models and the predicted sales is not highly reliable.

Key predictors: **Units_Sold**, **Discount Percentage**, and **Categories**.

The model only accounts for a small portion of the variance in sales revenue. However, **the conclusion that Units Sold is a primary driver of revenue is robust and can be trusted**.

05 Suggestions

Actionable insights to drive growth in retail
Based on 2 above preliminary conclusions

Amending the Dataset

1. Invest in Data Collection Infrastructure
2. Add New Data Points to Existing Records

Let the dataset change your mindset.

- Hans Rosling -

The reasons why

Our initial conclusions after EDA were a good start, but the regression model's poor performance isn't a failure. It's a finding.

It tells us that **our current data is not enough to accurately predict day-to-day sales revenue**.

Invest in Data Collection Infrastructure

1

Start collecting transaction-level data

This is the most crucial missing piece. Instead of just a daily summary, we need a record for every single sale.

2

Link with current data structure

This data level allows us to link specific products to specific discounts and marketing efforts, see what customers buy together, and analyze the true impact of promotions on profitability, not just revenue.

Add New Data Points to Existing Records

1

Customer ID

This helps us analyze customer behavior, purchase history, and lifetime value.

2

Product Price & Cost

This is essential for calculating Discount Amount and for understanding profitability, not just revenue.

3

Channel

Was the sale online or in-store? This helps optimize the sales funnel.

Boosting Sales Revenue

1. Optimize Discounts Strategically
2. Marketing on Top Categories and Periods
3. Improve Cross-Selling Opportunities

Optimize Discounts Strategically

Continue using discounts to drive sales, as the EDA showed a strong positive correlation with revenue and **test new strategies** like time-limited offers.

Investigate the negative coefficient of Discount Percentage. It's likely that large % discounts on low-value items (e.g., groceries) are lowering overall revenue.

Focus on high-value items to apply higher dollar-amount discounts like furniture and electronics.

Marketing on Top Categories and Periods

Allocate Marketing budget based on category sales.

Top priority is Electronics, following by Furniture and Clothing.

Focus campaigns during high sales volume periods.

Promotion for weekends and peak holiday season (Nov-Jan) based on continents' culture.

Add more holidays to dataset.

Add Black Friday as a "global shopping holiday" and investigate major regional holidays (e.g. Independent Day) for localized efforts.

Improve Cross-Selling Opportunities

Design cross-selling campaigns to encourage customers to buy products from different categories together. E.g.: TV (Electronics) and TV stand (Furniture).

Take advantage of technology. Use website recommendations to suggest complementary products, offering a small discount for bundling items from different categories.

06
The ending

Documents

1. Original dataset: CSV file
2. Cleaned dataset: CSV file
3. Store Location and Continent: CSV file
4. Visualization dashboard: PBIX file and PDF file
5. Presentation: PPTX file or PDF file
6. Github repository: Website link

Thank you for your attention!

Should you need any further discussion, please feel free to contact me at:
thuydungtran.dtt@freepik.com
+84 903 949 561

Sincere gratitude is extended to

Mr. Cường from practice class,
Ms. Thảo from lecture class,
all my valued classmates,
and Swiss Coding Academy
for your enthusiasm and support.