Global Supermarket Analysis

Q1. Exploratory Data Analysis (EDA)

Basic Dataset Information

Data shape: 26 features, 51,290 unique rows

Data summary

- Missing values: 0
- Duplicates: 0

	data type	#missing	%missing	#unique	min	max
customer_id	object	0	0	4873	NaN	NaN
customer_name	object	0	0	795	NaN	NaN
customer_segment	object	0	0	3	NaN	NaN
order_id	object	0	0	25035	NaN	NaN
order_city	object	0	0	3636	NaN	NaN
order_region	object	0	0	13	NaN	NaN
order_date	datetime64[ns]	0	0	1430	2019-01-01	2022-12-31
order_year	int64	0	0	4	2019.0	2022.0
order_weeknum	int64	0	0	53	1.0	53.0
quantity	int64	0	0	14	1.0	14.0
sales	int64	0	0	2246	0.0	22638.0
product_id	object	0	0	10292	NaN	NaN
product_name	object	0	0	3788	NaN	NaN
profit	float64	0	0	24575	-6599.978	8399.976
discount	float64	0	0	27	0.0	0.85
category	object	0	0	3	NaN	NaN
sub_category	object	0	0	17	NaN	NaN
market_country	object	0	0	147	NaN	NaN
market_area	object	0	0	7	NaN	NaN
market_city	object	0	0	1094	NaN	NaN
ship_date	datetime64[ns]	0	0	1464	2019-01-03	2023-01-07
ship_mode	object	0	0	4	NaN	NaN
shipping_cost	float64	0	0	16877	0.002	933.57
row_id	int64	0	0	51290	1.0	51290.0
pre_sales	float64	0	0	6601	0.0	45276.0
uni_cost	float64	0	0	9285	0.0	7546.0

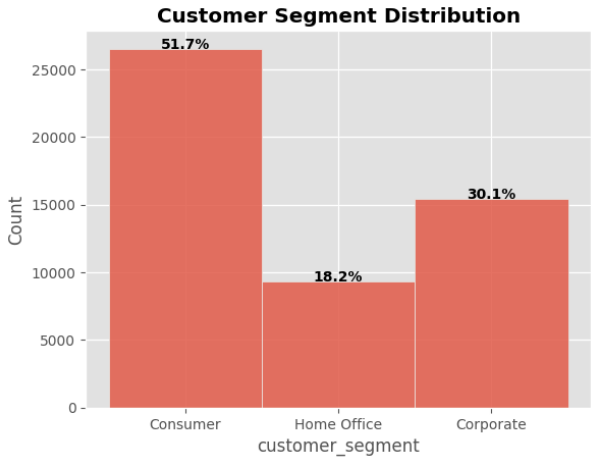
The EDA was conducted by dividing the analysis into 6 major and minor topics. Additional features (`pre_sales`, `uni_cost`) were created to perform a more in-depth EDA.

```
data['pre_sales'] = data['sales'] / (1-data['discount']) # Sales before discount is applie
d
data['uni_cost'] = data['pre_sales'] / data['quantity'] # unit cost per product
```

1. Customer Information (ID, Name, Segment)

```
print('Unique Customer ID:', data['customer_id'].nunique())
print('Unique Customer Names:', data['customer_name'].nunique())
--------------------------------------
Unique Customer ID: 4873
Unique Customer Names: 795
```

The supermarket has a total of 4,873 unique customer IDs and 795 unique customer names.



- Consumer: Personal
- Corporate: Companies
- Home Office: Remote Office

- This graph shows the proportion of each customer segment. Individual customers account for more than half of the customer base. Corporate customers follow at approximately 30%.

2. Order Information

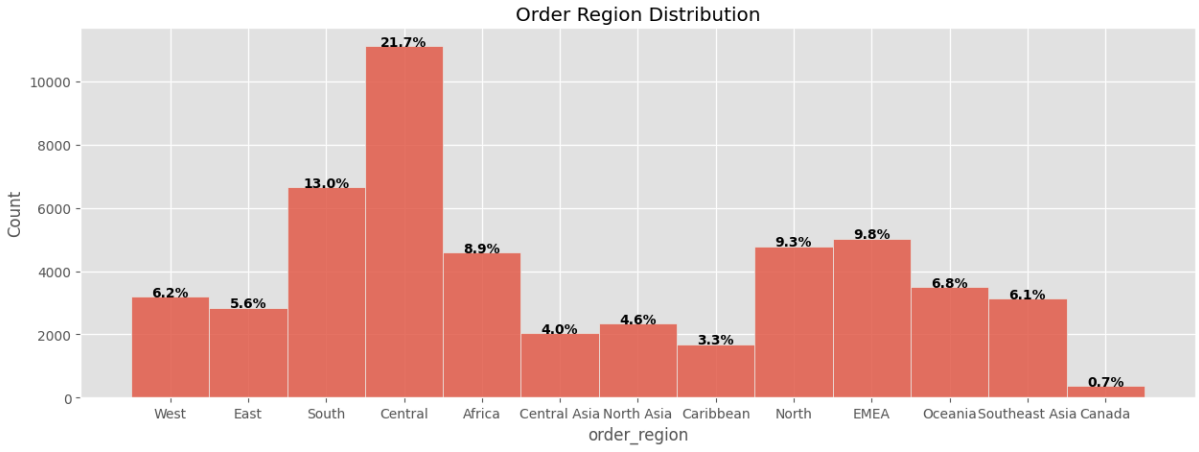
- Three perspectives were used to analyze the “Order” features.

```
print('Unique order_id:', data['order_id'].nunique())
print('Number of duplicated rows:', data.duplicated().sum())
-----------------------------------------------------
Unique order_id: 25035
Number of duplicated rows: 0
```

There are 25,035 unique order IDs recorded and no duplicate rows. This indicates that multiple line items are recorded under the same `order_id`.

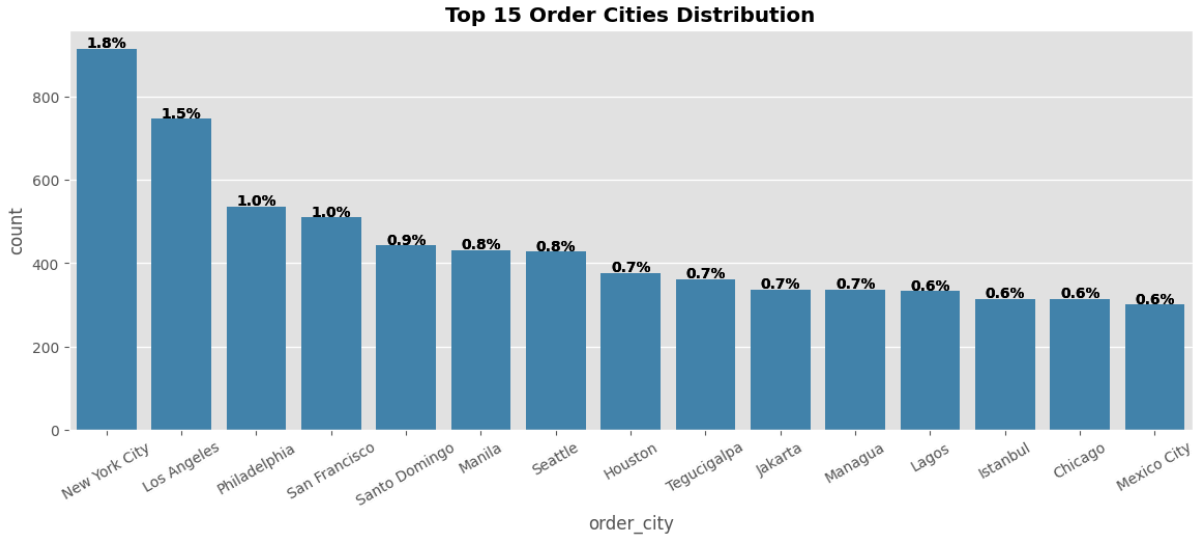
2-1. Regional Characteristics (Regions, Cities)

Order Region



The region with the highest share of orders is **Central**, while Canada has the lowest order count. Although “Central / West / East” might originally be interpreted as US regions, some non-US locations (for example, Berlin) are labeled as Central. This suggests that `order_region` is being used in a broader, more inclusive sense than US-only regions.

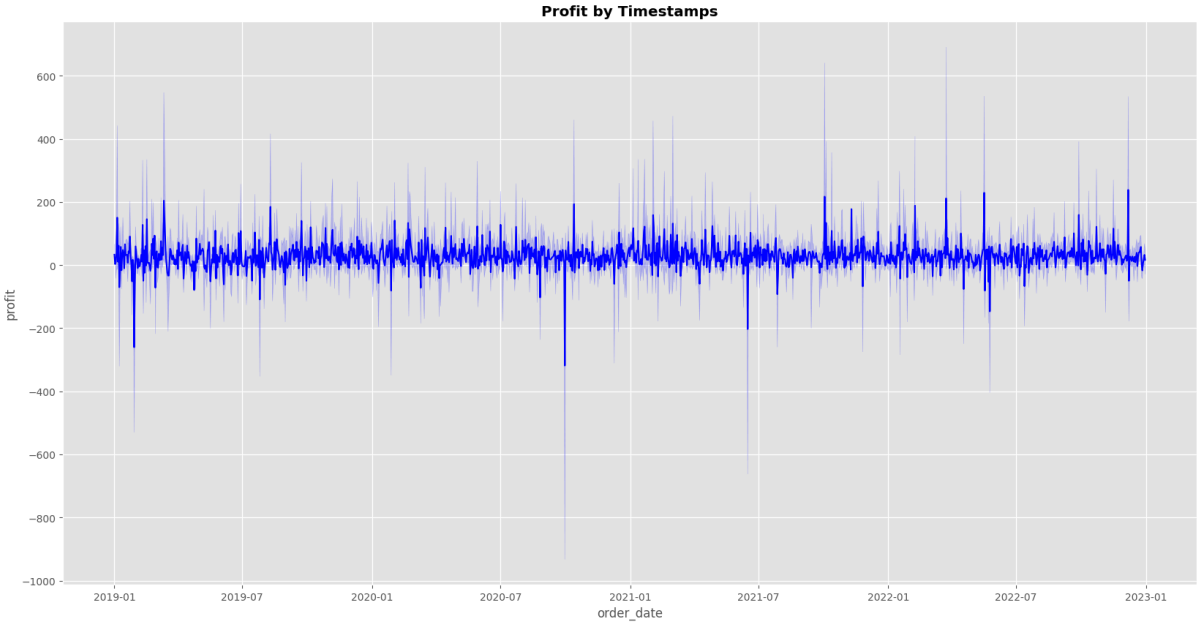
Order City



This chart shows the top 15 cities by order volume. Since the highest city accounts for only about 1.8% of orders, the distribution suggests that many cities contribute a similar share of total orders.

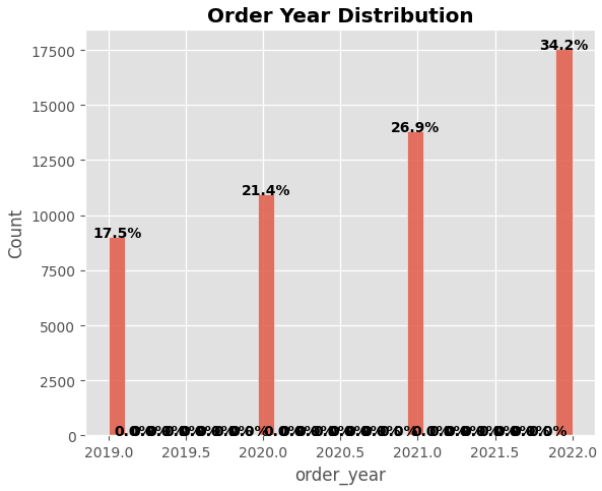
2-2. Time Characteristics (Year, Week Number, etc.)

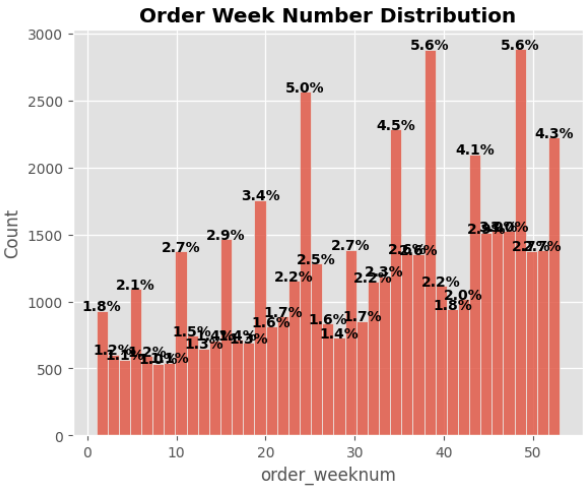
Order Date



This plot shows changes in `profit` over time from 2019-01-01 to 2022-12-31. Although there are occasional spikes, most values tend to stay relatively close to 0.

Order Year / Week Number



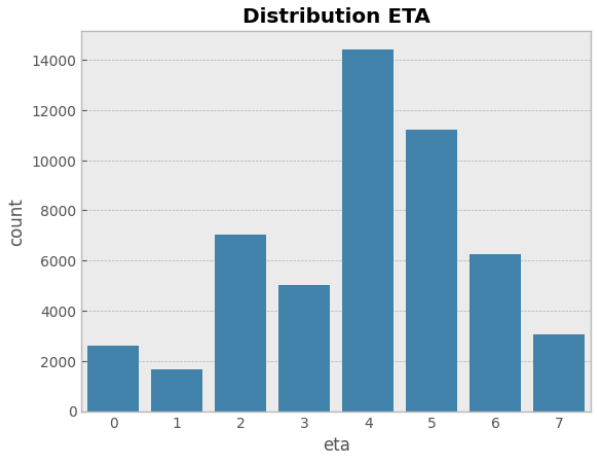


The largest number of orders occurred in 2022 (34.2%). Orders also tend to increase during the middle and end of the year rather than at the beginning.

✳️ ETA (Estimated Time of Arrival)

A derived feature was created from `order_date` and `ship_date`.

```
data['eta'] = (data['ship_date'] - data['order_date']).dt.days
```



- The most common delivery time is 4 days. This is likely influenced by the shipping class.

3. Product (ID, Categories)

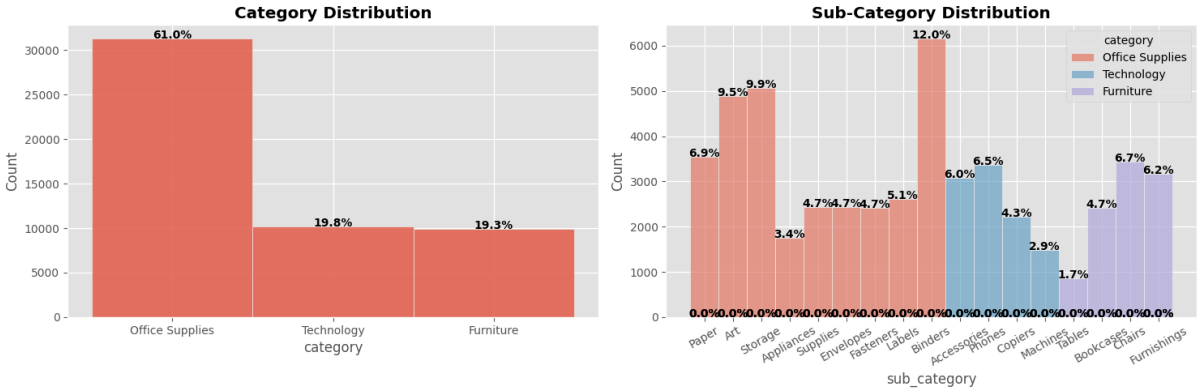
Product information is critical for a business. Different products have different profitability. Some products act as “stars” that drive profits, while others can harm profitability. Businesses should improve or replace unprofitable products to maintain a financially stable operation.

IDs and Names

```
print('Unique product ID:', data['product_id'].nunique())
print('Unique product names:', data['product_name'].nunique())
---------------------------------------------------------------
Unique product ID: 10292
Unique product names: 3788
```

There are 10,292 unique product IDs and 3,788 unique product names.

Categories

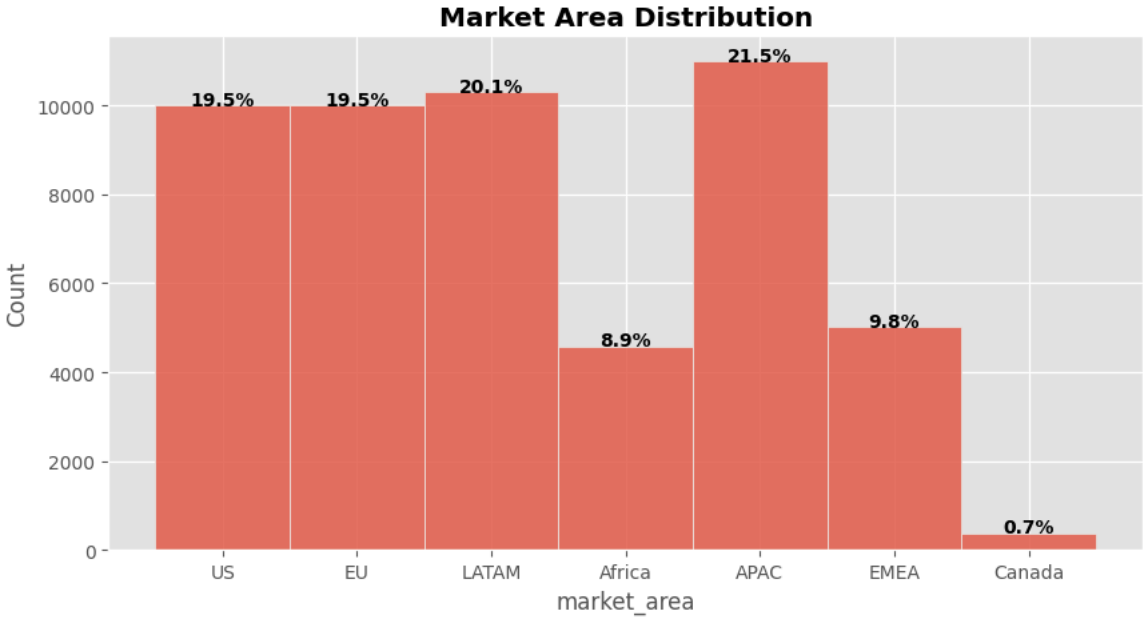


Within categories, Office supplies make up the largest share. Within that category, Binder, Art, and Storage take up a substantial portion.

4. Supermarket (Area, Country, City)

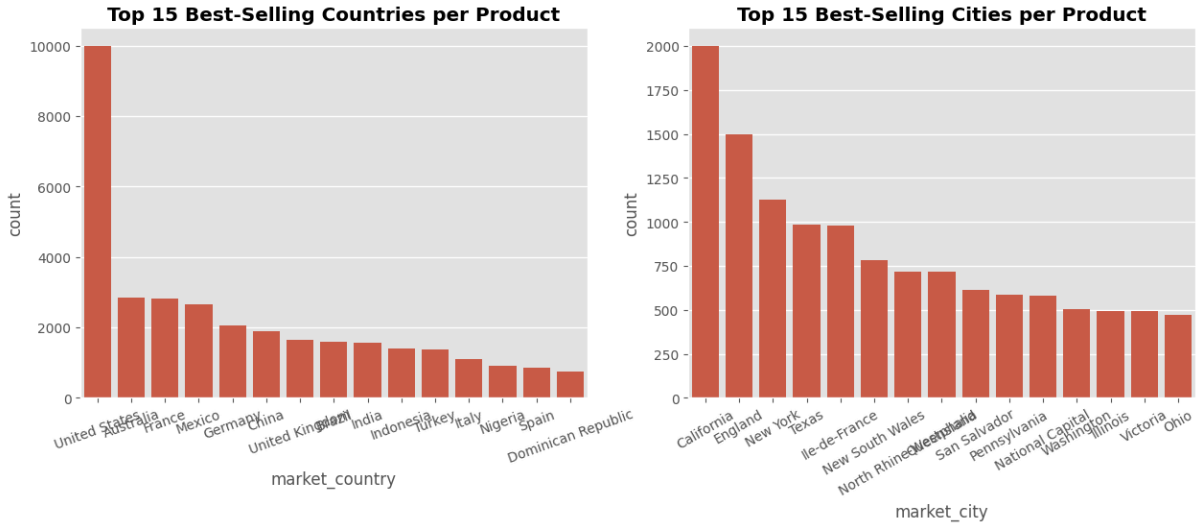
Knowing where the supermarket operates is an important factor for effective sales strategy, since product quantity and assortment can be tailored by country and city.

Area



Market areas are distributed across major regions (US, EU, LATAM, APAC). The US is notable because it represents a single country (excluding Canada), yet still accounts for a large portion.

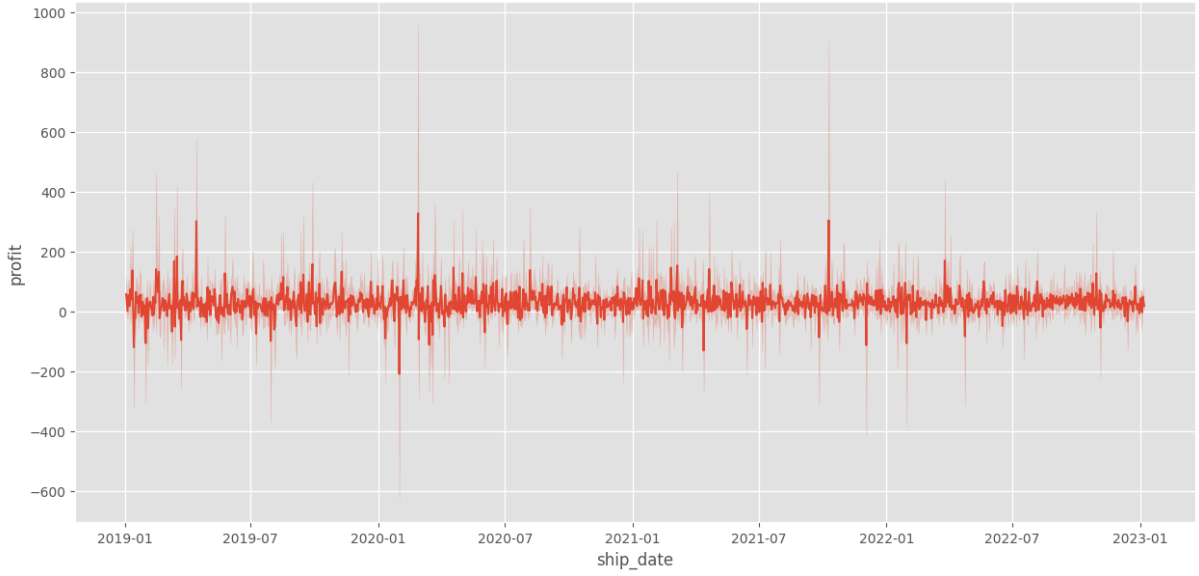
Country & City (Top 15)



- As mentioned earlier, the US has the largest transaction volume (about 20%). Among US locations, **CA (West), NY (East), and TX (Central)** rank 1st, 3rd, and 4th.
- England also deserves attention as the 2nd highest in transactions per store. The transaction volume associated with England is comparable to that of many other countries.
  - As a side note, cities recorded under the United Kingdom appear to be **"England", "Scotland", and "Wales"**.

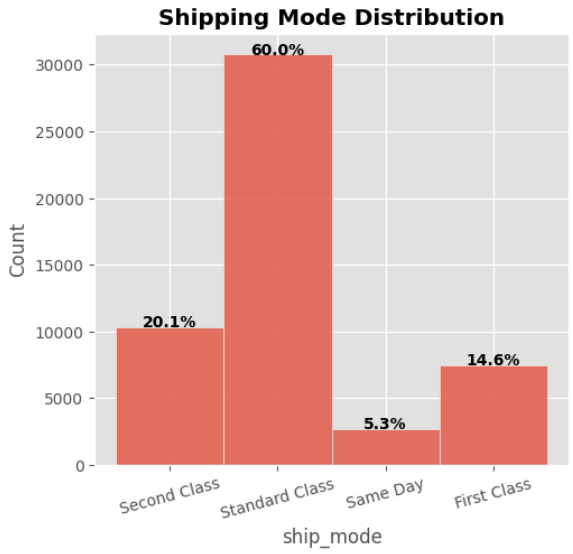
5. Shipping

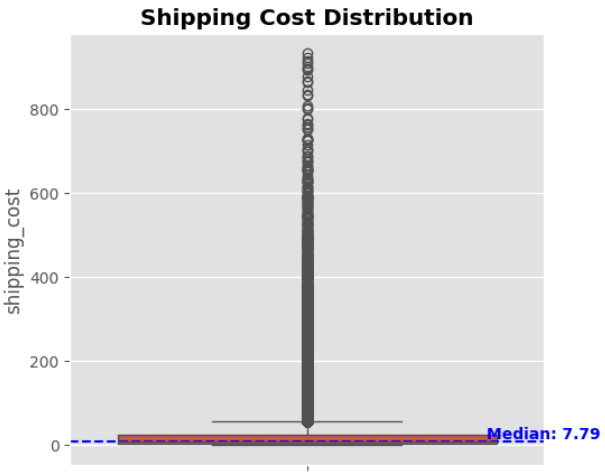
Shipping Date



Delivery dates generally follow a similar trend to shipping dates.

Shipping Mode / Cost



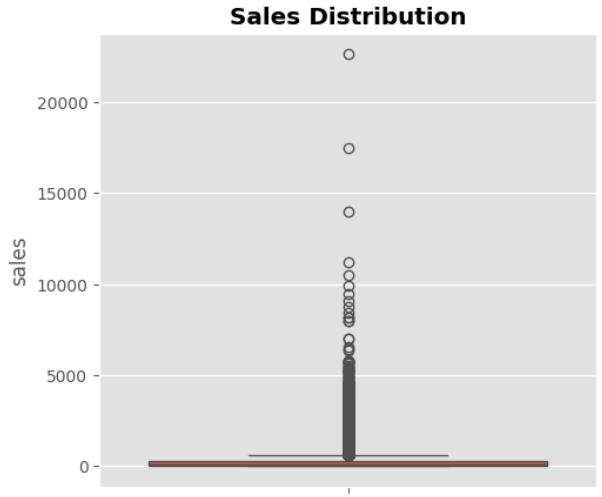


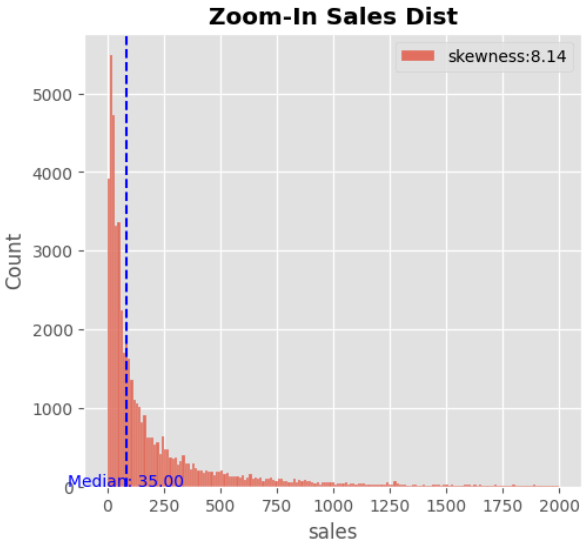
- “Standard Class” accounts for the largest share of shipping modes.
- Most shipping costs are between 0 and 50, but the boxplot indicates a small number of very high shipping cost orders.

6. Prices (Sales, Profit, etc.)

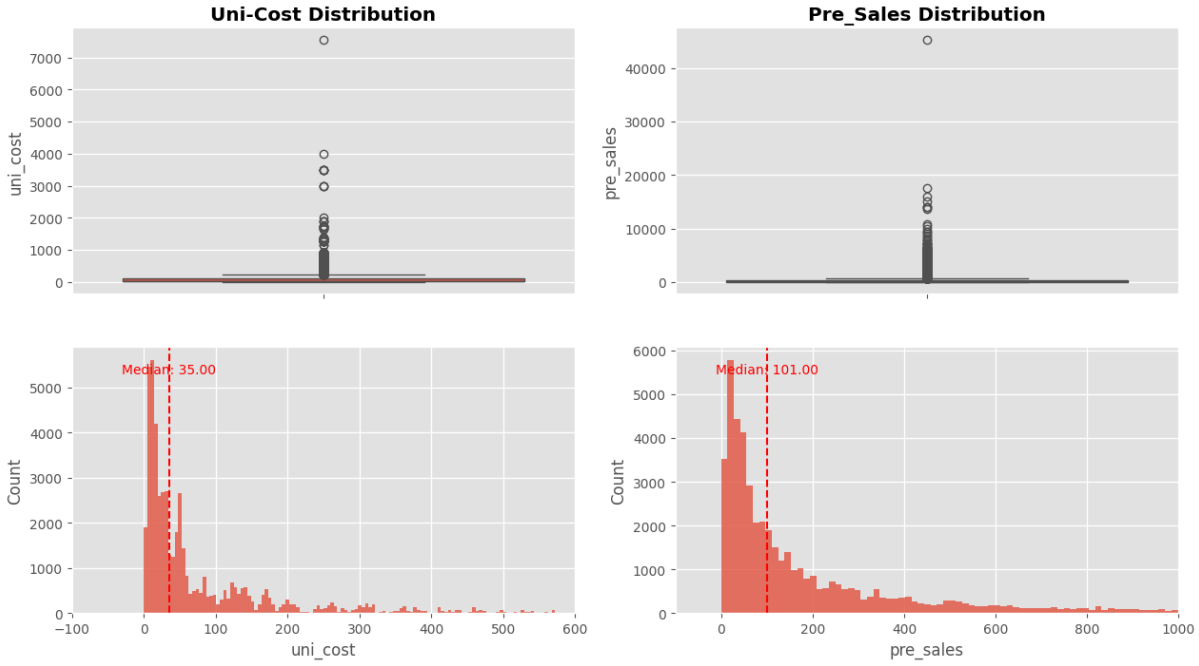
Financial metrics are crucial for a business. A company aims to generate profit, and profitability can strongly affect long-term success.

Sales (incl. Pre-Sales and Unit Cost)



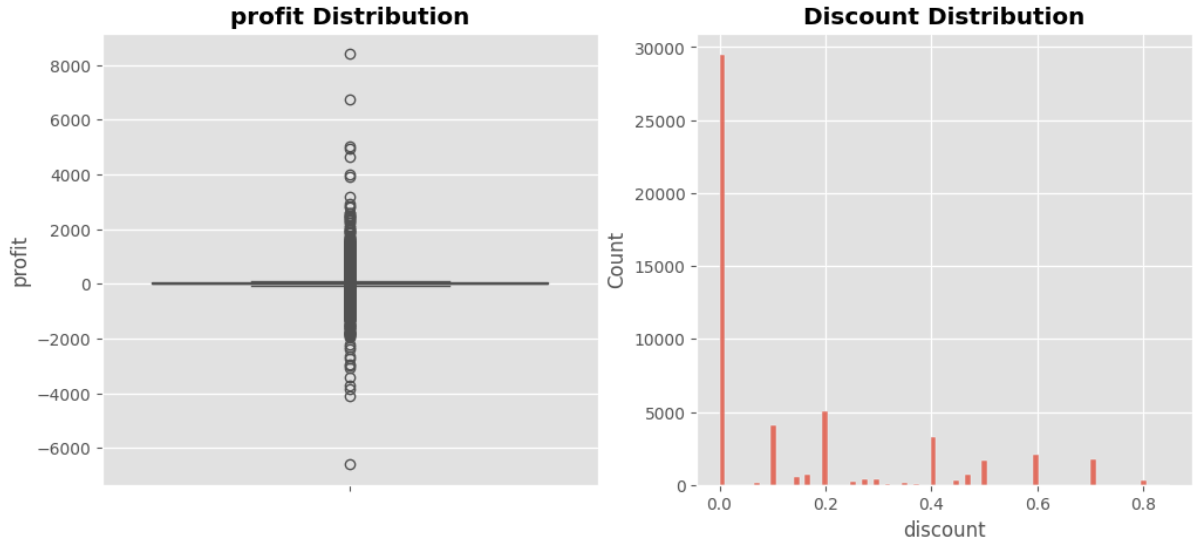


Sales are widely distributed. Most sales fall between 0 and 250, although there are also high-priced items.

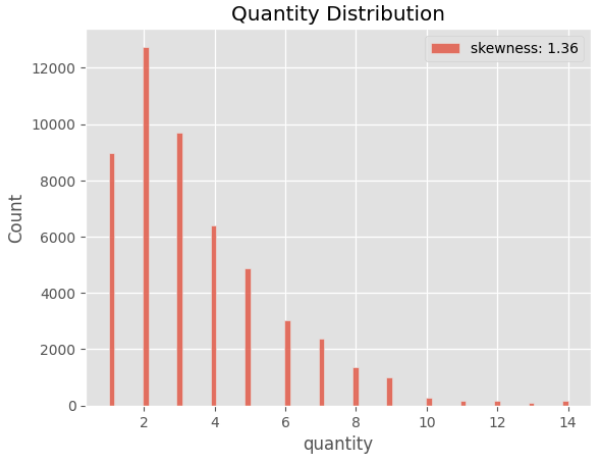


`pre_sales` and `uni_cost` show distributions similar to sales.

Profit, Discount, Quantity



- `profit` values are centered around 0.
- `discount` has a large mass at 0.
- Roughly 80% of transactions are profitable (`profit > 0`), while about 20% are loss-making.



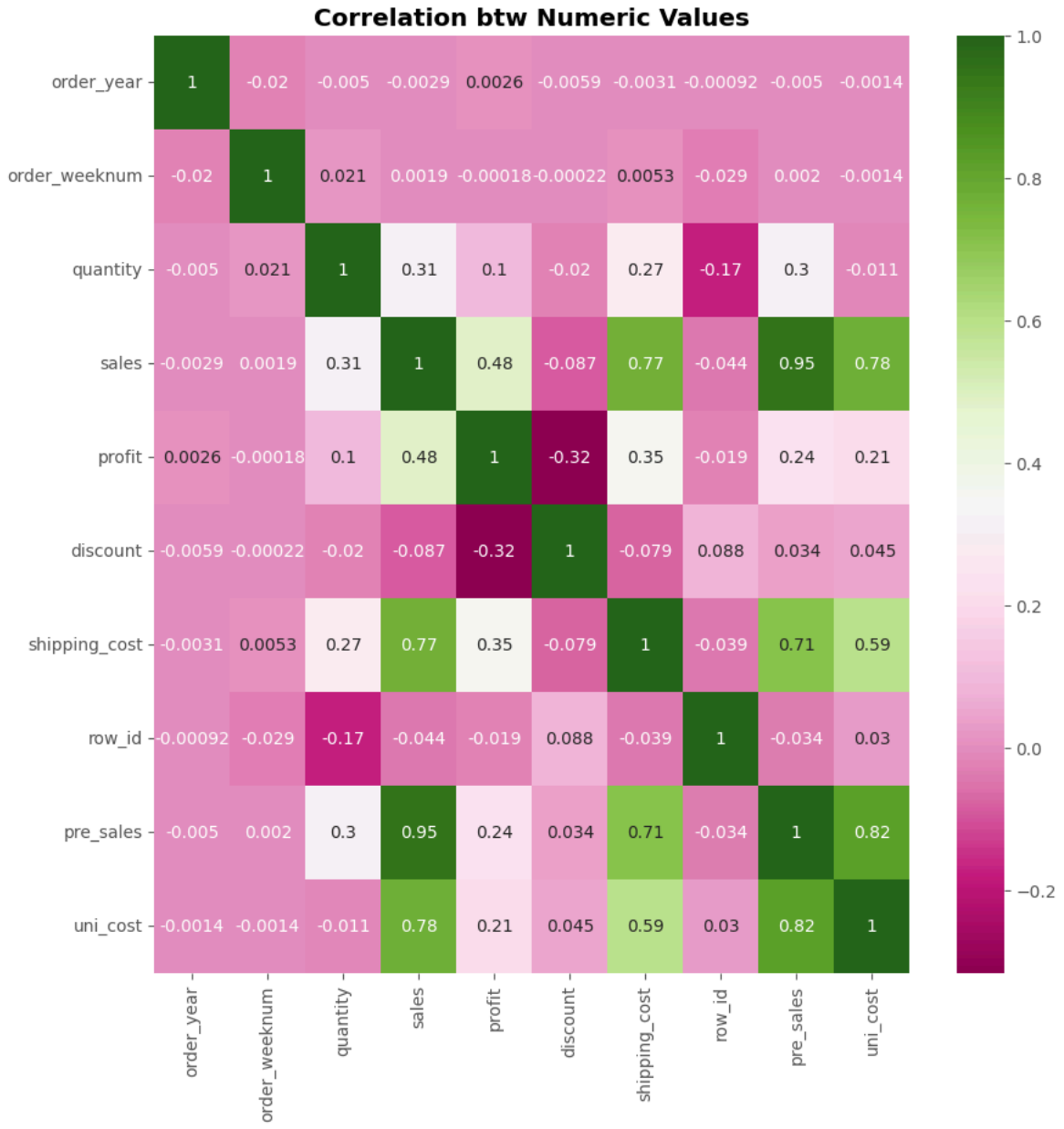
- Quantity is most commonly 2 units per order. Compared to other continuous variables, it appears more normalized.

7. Correlation Analysis

All features were categorized as follows.

Column Types	Column Names
Individual identifiers (ind_col)	customer_id, customer_name, order_id
Categorical data (cat_col)	customer_id, customer_name, customer_segment, order_id, order_city, order_region, order_date, order_year, order_weeknum, product_id, product_name, market_area, market_city, ship_date, ship_mode, category, sub_category, market_country
Continuous data (con_col)	quantity, sales, shipping_cost, profit, discount, row_id, uni_cost, pre_sales

To understand relationships between variables, a heatmap was used to visualize correlation coefficients.



- **Discount vs. Profit (0.32)**: Higher discounts tend to be associated with lower profitability, showing a negative tendency overall. More analysis is needed, but understanding and improving this relationship can increase profitability.
    - For example, `pre_sales` (sales before discount) shows a positive relationship relative to discount, suggesting that reducing discount rates could help improve profitability.
- **Sales**: Sales has noticeable correlations with several continuous variables. In particular, higher shipping cost orders are associated with higher sales (0.77). Selling larger quantities also correlates positively with sales (0.31).

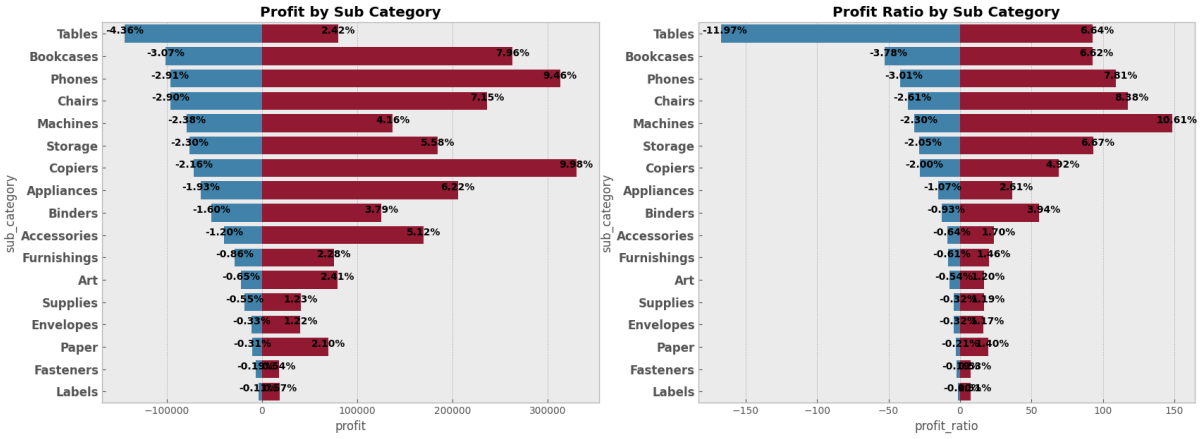
Loss Analysis

This section explores loss-making transactions to identify ways to reduce losses and improve operating profit. The analysis focuses on bivariate and multivariate perspectives related to profit loss.

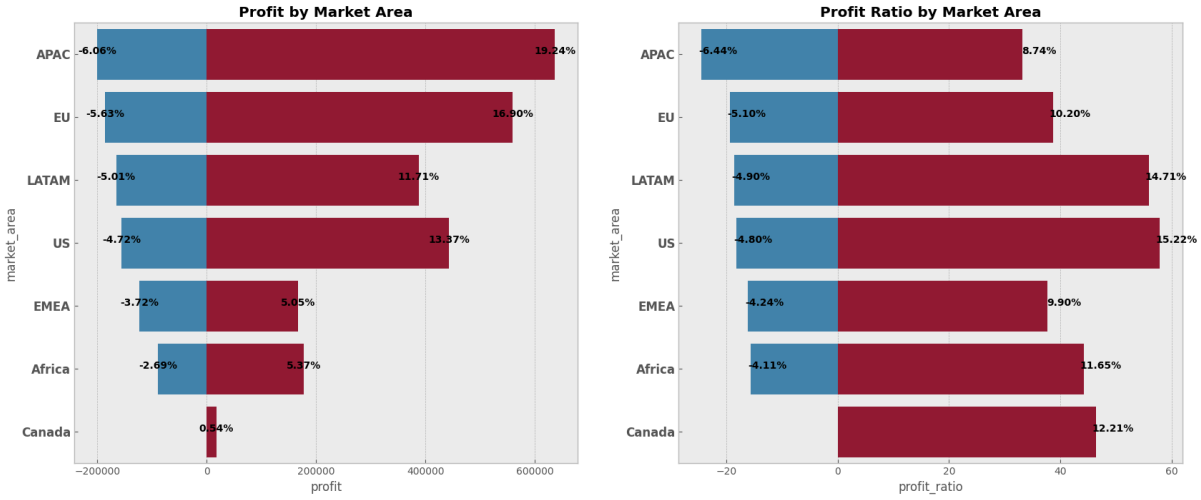
First, transactions with negative profit were extracted into a DataFrame named `minus`.

```
minus = data.loc[data['profit'] < 0]
print(minus.shape)
```

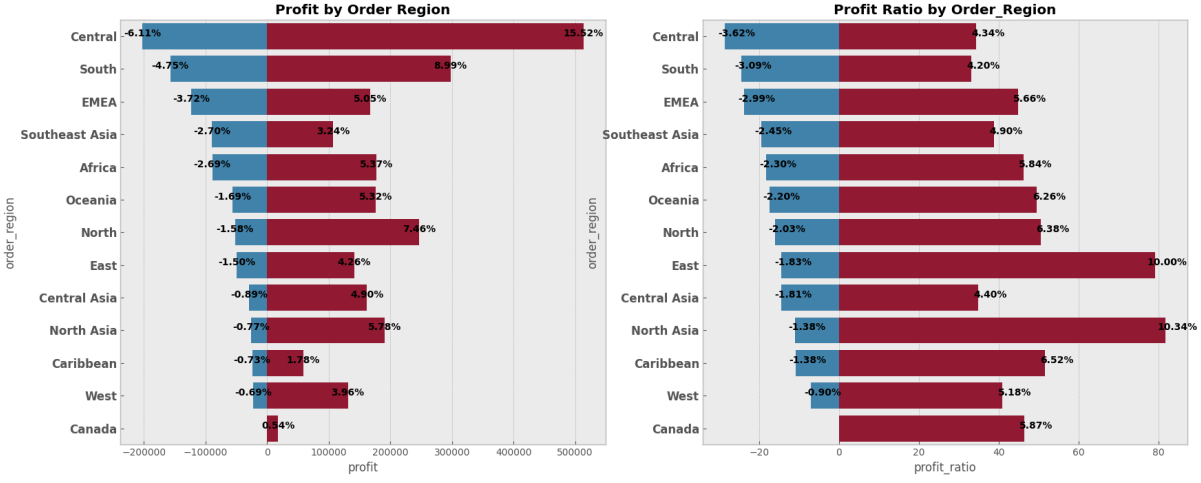
```
----------------
(12544, 26)
```



These graphs visualize loss/profit amounts and loss/profit ratios by sub-category. **Tables** show the largest losses in both amount and ratio. In contrast, **Copiers** appear to be one of the most profitable sub-categories.



By market area, all regions show overall positive profit. APAC has a large share of losses, but it also has high profitability overall.

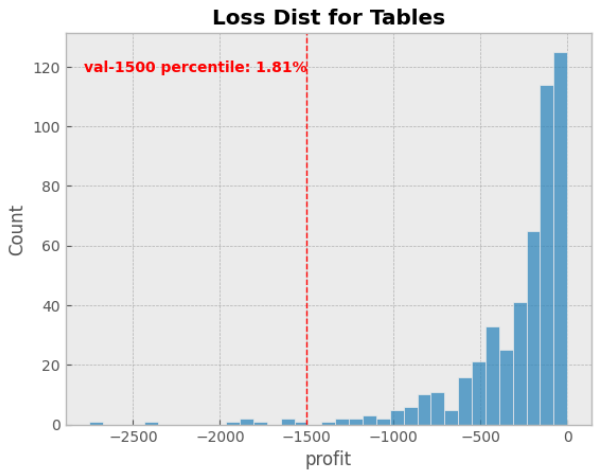


By order region, Central shows the largest losses but also the highest profits, making it one of the best-performing regions overall. Northeast Asia also shows high profit per transaction.

1. Product: “Tables”

```
mt = minus.loc[minus['sub_category'] == 'Tables']
mt['sub_category'].min()
------------------------------
-2750.28
```

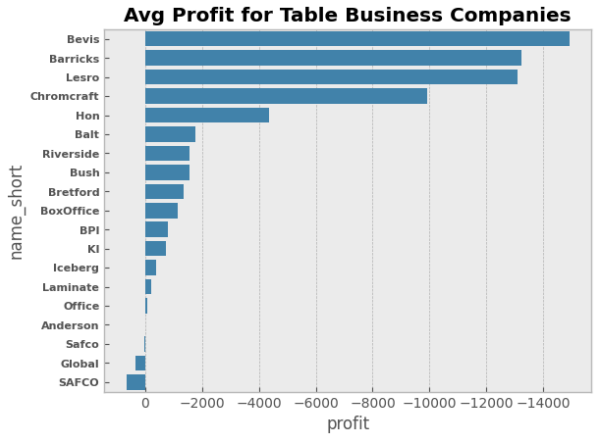
A DataFrame `mt` was created to analyze loss-making “Tables” transactions.



- The graph shows the loss distribution for “Tables”. Most losses fall between -1000 and 0.
- Reducing small losses matters, but preventing the largest-loss transactions is the highest priority.
- Below is a summary DataFrame for the most damaging transactions (with `uni_cost` rounded).

index	order_region	order_city	market_area	product_id	discount	uni_cost	profit
29652	EMEA	Vilnius	EMEA	FUR-BAR-10003532	0.7	904.5	-2750.28
30191	Central Asia	Lahore	APAC	FUR-TA-10002172	0.8	919.28	-2380.35
29974	Central	Hanover	EU	FUR-TA-10003963	0.85	925	-1924.54
29704	North	Stockholm	EU	FUR-TA-10003354	0.7	909	-1864.09
47284	South	Concord	US	FUR-TA-10000198	0.4	551.02	-1862.31
29390	EMEA	Ankara	EMEA	FUR-BEV-10002193	0.6	520.42	-1779.76
38582	South	Barcelona	EU	FUR-TA-10004054	0.6	857.5	-1629.54
40773	Africa	Zaria	Africa	FUR-CHR-10002278	0.7	469.2	-1576.82
29693	North	Stockholm	EU	FUR-TA-10004371	0.7	454.28	-1557.99

- `name_short` was created by extracting the substring before the first space in `product_name`. Some product names include a leading company/brand string.
- Overall, these high-loss table transactions tend to have high discount rates and are concentrated in EU stores.



- This chart shows the average profit/loss by table manufacturer.
- The top 5 companies with the largest average operating losses can also be seen in the table above.
- Preventing the transactions that generate the largest losses appears to be the most effective immediate solution.

In particular, **Bevis, Barricks, Lesro, and Chromcraft** show substantial losses. Downsizing or discontinuing table offerings from these companies could be an effective way to reduce losses.

- **Bevis**

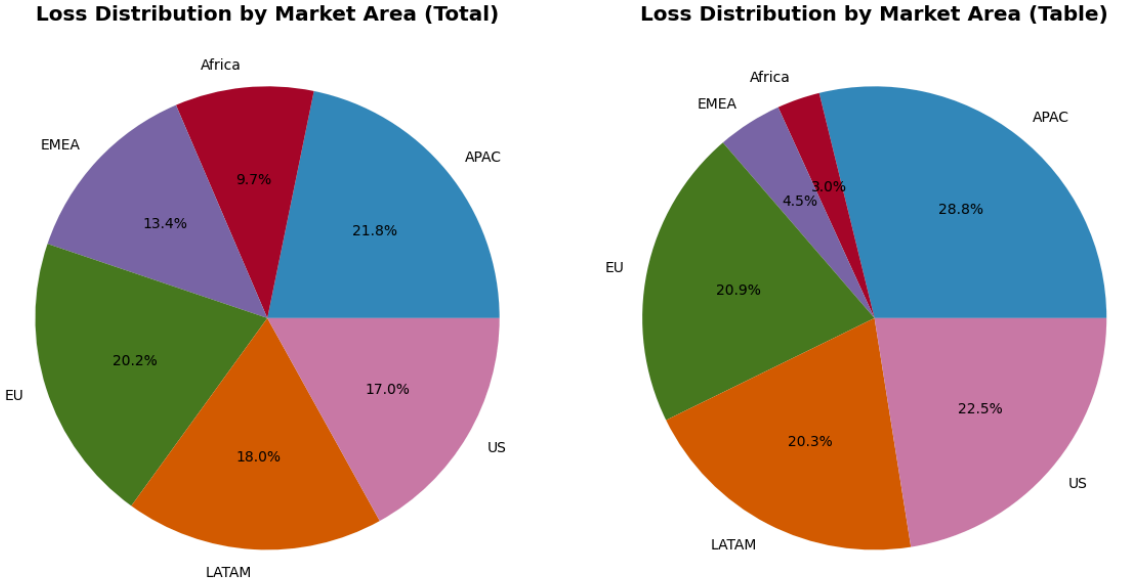
Bevis Training Table, Fully Assembled',
'Bevis Wood Table, Rectangular',
'Bevis Training Table, Rectangular',
'Bevis Computer Table, Fully Assembled',
'Bevis Computer Table, Adjustable Height', etc.

- **Barricks Furniture Solutions**

'Barricks Computer Table, Fully Assembled',
'Barricks Training Table, with Bottom Storage',
'Barricks Conference Table, Rectangular', etc.

- **Lesro Reception Furniture**

'Lesro Computer Table, Adjustable Height',
'Lesro Training Table, Fully Assembled',
'Lesro Wood Table, with Bottom Storage',
'Lesro Coffee Table, Adjustable Height', etc.



(Left) Share of total loss by market area (all products) vs. (Right) Share of loss by market area (tables only)

In **APAC, US, and LATAM**, the share of losses coming from table sales is higher than their share of overall losses. This suggests that tables generate relatively larger losses in those regions, and further regional follow-up analysis is needed.

Notes / Considerations

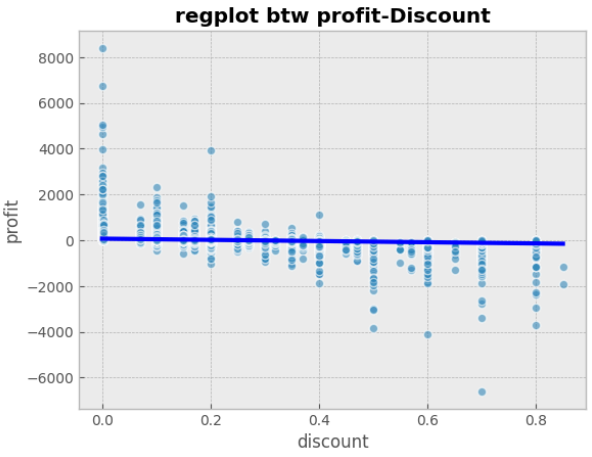
The number of transactions recorded for Tables is relatively small (about 861). While action seems urgent based on current data, collecting more data would be recommended to support stronger decision-making.

---

2. Discount

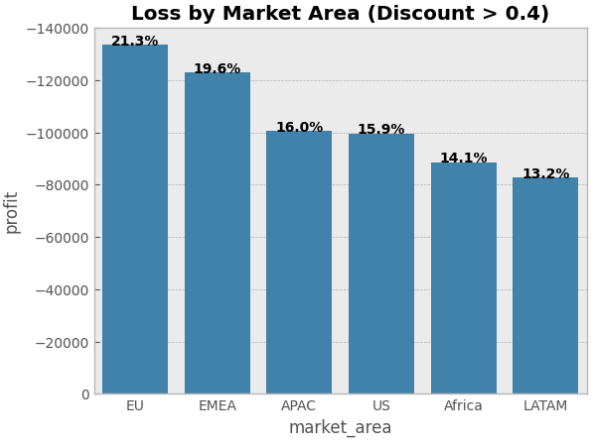
Discounts are an important driver of purchases, but aggressive discounting can cause losses. Discount may vary by **region**, **product**, and **customer segment**.

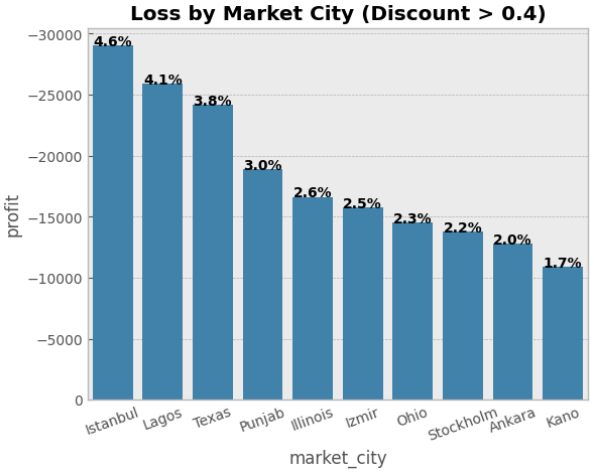
- Unique values: 27
- Most common value: 0 (29,009 rows, about 56%)



- This plot shows `profit` vs `discount` using a regression line.
- Above a discount rate of 0.4, there are very few profitable orders.
- Many orders above 0.4 generate losses, and the magnitude of losses also increases.

Transactions with discount rates greater than 0.4 account for about 13% of the data. This subset contributes only about 0.01% of total profit, but accounts for about **68% of total losses**.

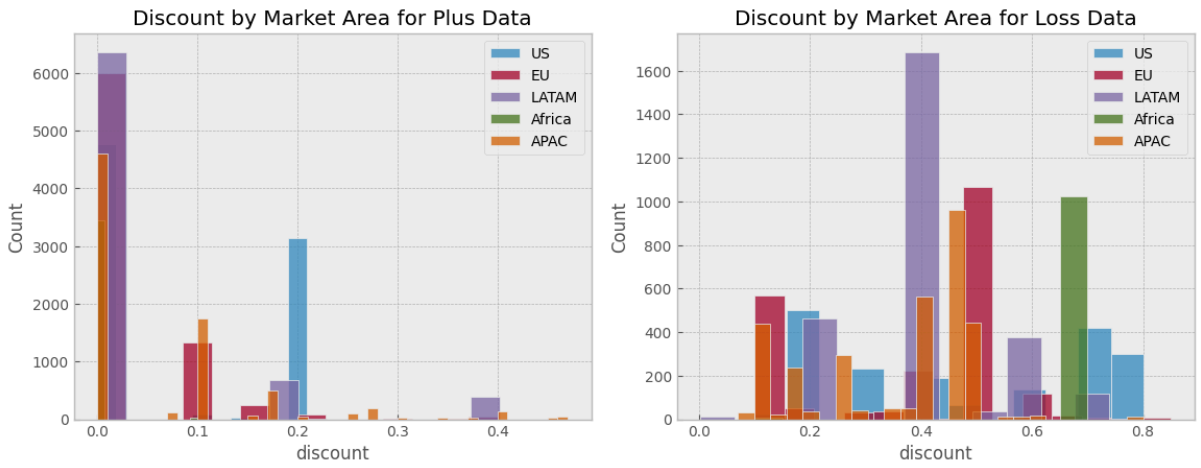




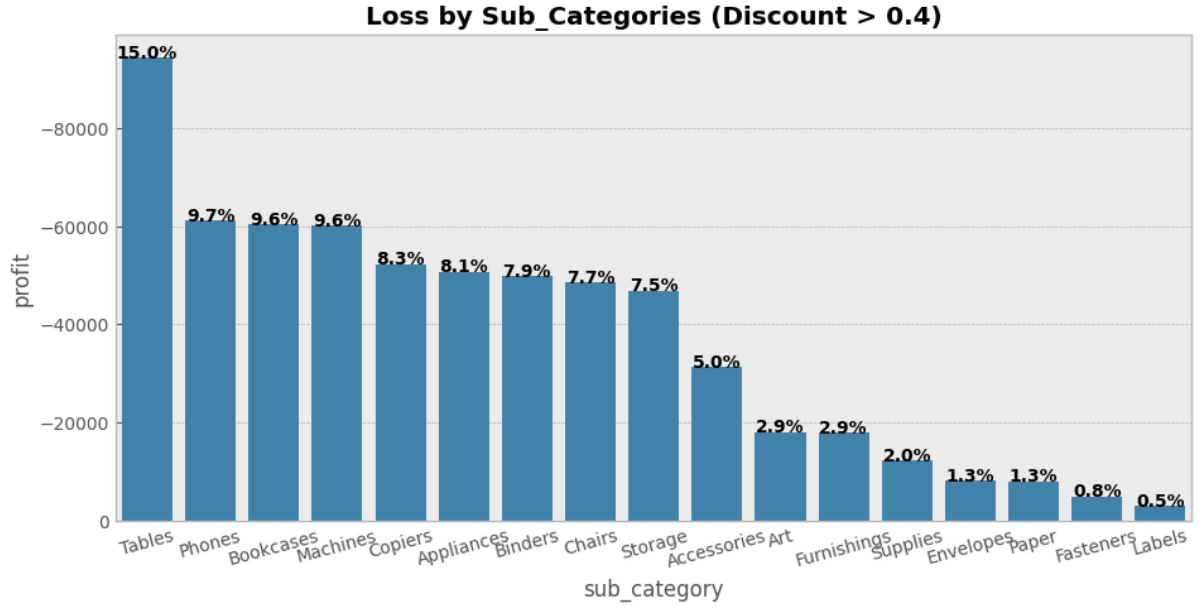
- By market area, the loss shares are broadly similar. Reducing the discount rates applied in the EU is expected to improve profitability.
- For the top 3 cities with high loss contribution, the observed discount values were as follows. Based on this, lowering discount rates below 0.4 in Istanbul and Lagos and reducing high-discount items in Texas are recommended.

City	Discount	Values
Istanbul	0.6	425 (total)
Lagos	0.7	333 (total)

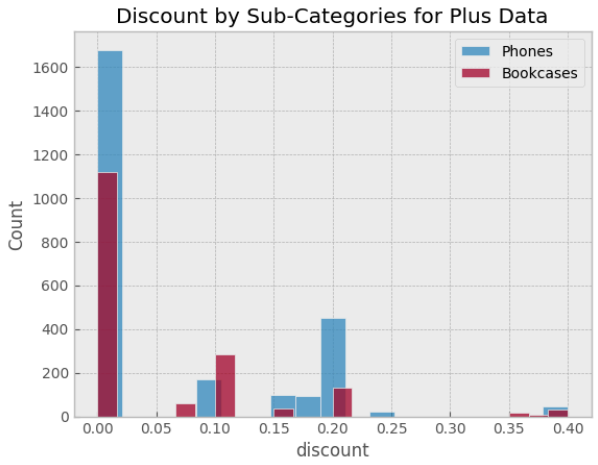
City	Discount	Values
Texas, US	{0.2, 0.8, 0.3, 0.6, etc.}	{570, 200, 94, 81, ...}

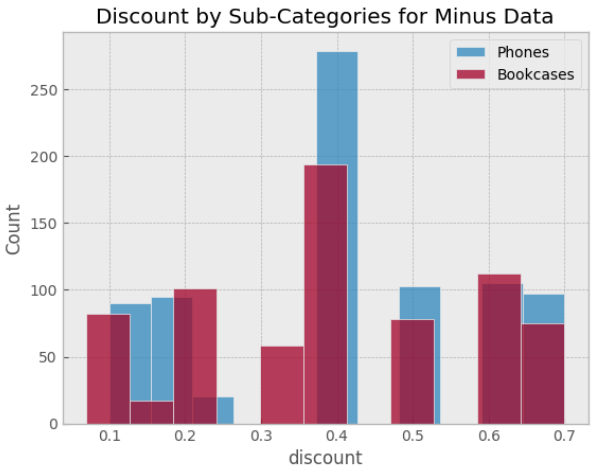


This plot shows discount distributions by market area, split into profitable vs loss-making transactions. In the EU (red) plot, most losses occur in the 0.4–0.6 discount range. Reducing discount rates for these products by about **10–30 percentage points** is expected to reduce losses.



Since improvements for Tables were discussed above, the analysis now focuses on **Phones** and **Bookcases**, which have both high profitability and high transaction volumes (about 6.5% and 6.7% of total transactions, respectively).





As above, these charts show discount distributions split by loss vs profit and by sub-category.

- For both categories, discount rates should be adjusted so that they do not exceed 40%.
- For Phones, the 20% discount band appears to generate meaningful profit. A strategy of selectively moving eligible products into the 20% discount band is recommended.

Q2. Final Business Decision-Making Report

💡 Based on the EDA and detailed loss analysis above, transactions involving products with high discount values were confirmed to be a major driver of corporate losses. In particular, products with discount rates of 40% or higher require discount rate reductions.
Key recommendations:

- Reduce discount rates for products sold in the EU by an average of 20 percentage points.
- Diversify and lower discount rates in Istanbul and Lagos, especially keeping discounts below 40%.
- For the Tables sub-category, consider large-scale business adjustments for products from companies such as "Bevis", "Barricks Furniture", and "Lesro" due to consistently large losses.
- APAC, US, and LATAM show disproportionately high losses compared to other regions, indicating a need for follow-up investigation and targeted improvements.