Project: Dataset Description & Exploration – SuperStore Sales

Course code: BDM200

Instructor: Hemant Sangwan

Group 2

Group Member: Brian, Tai, Nhi, Tin, Victor

Date: February 13th,2026

Dataset Description & Exploration: SuperStore Sales

1. Dataset Structure Identification

The dataset **Group Project Data–SuperStore-Sales.xlsx** contains two worksheets: **Index** and **Data**. The analysis uses the **Data** sheet.

The dataset contains **9,994 rows and 21 columns** in the raw file. Two trailing empty columns (Unnamed: 19, Unnamed: 20) appear in the raw data; these can be removed for analysis, resulting in **19 usable variables**.

Each row represents a **single product-level sales transaction within a customer order**, meaning one order may appear multiple times when multiple products are purchased (line-item level).

Variable names (19 variables):

order_id, order_date, ship_date, customer, manufactory, product_name, segment, category, subcategory, region, zip, city, state, country, discount, profit, quantity, sales, profit_margin.

Variable types (data structure):

- **Date (2 variables):** order_date, ship_date (interpreted in YYYY-MM-DD format)
- **Numerical (5 variables):** continuous—sales, profit, discount, profit_margin; discrete—quantity
- **Categorical (12 variables):** order_id, customer, manufactory, product_name, segment, category, subcategory, region, zip, city, state, country

Categorical levels (examples):

- **segment**: Consumer, Corporate, Home Office (3 levels)
- **category**: Office Supplies, Furniture, Technology (3 levels)
- **region**: West, East, Central, South (4 levels)
- **subcategory**: 17 unique values

2. Variable Description and Data Quality Assessment

This section describes key variables used to understand sales performance and customer behavior, and checks for missing values, inconsistencies, and outliers.

2.1 Key variable descriptions (examples)

- order_id (**Categorical**): Unique order identifier; can repeat because an order may contain multiple products (line-item dataset). Repeated order_id values are expected and not an error.
- order_date (**Date**): Date the order was placed; range 2019-01-03 to 2022-12-30; no invalid or missing dates detected.

- ship_date (**Date**): Date the order shipped; range 2019-01-07 to 2023-01-05; 0 cases where ship_date is earlier than order_date.
- segment (**Categorical**): Customer classification with three levels: Consumer, Corporate, Home Office; no missing values and consistent labels.
- category (**Categorical**): Product grouping including Office Supplies, Furniture, and Technology; clean and consistent.
- sales (**Continuous**, **currency**): Revenue per line item; range 0.444 to 22,638.480; mean 229.858, median 54.490; distribution appears right-skewed.
- profit (**Continuous**, **currency**): Profit per line item can be negative; range −6,599.978 to 8,399.976; 1,871 records have negative profit; outliers present by IQR rule.
- discount (**Continuous**): Discount rate 0.00 to 0.80; values fall in a reasonable range (no negatives, none above 1).
- quantity (**Discrete**): Number of units sold; range 1 to 14; positive integers with no inconsistencies.
- profit_margin (**Continuous**): Ratio of profit to sales; range −2.75 to 0.50; negative values occur when transactions generate losses.

2.2 Data quality summary

- **Missing values**: None detected across the dataset's 19 columns (after removing the empty Unnamed columns).
- **Inconsistencies**: 0 cases where ship_date is earlier than order_date (date logic is consistent).
- **Outliers**: Outliers exist in sales (high-end values), profit (extreme negative and positive), and profit_margin (extreme ratios). These are flagged as notable values for later analysis rather than automatically removed.
- **Type caution**: zip should be treated as categorical/identifier rather than a numeric measure.

3. Descriptive Statistics and Exploratory Findings

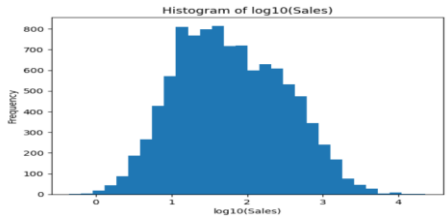
This section summarizes descriptive statistics and explores patterns and relationships using tables and plots produced in the group Python notebook. （Due to limited amount of space the graph would be extra small, can zone in or view the original graph from our Google Colab.）

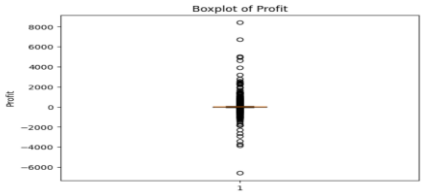
3.1 Descriptive statistics (5 key variables)

We summarize key numerical variables using **central tendency** (mean and median) and spread (quartiles/IQR) to describe typical values and variability.

	count	mean	std	min	25%	50%	75%	max
sales	9994.0	229.858001	623.245101	0.444	17.28000	54.4900	209.940	22638.480
profit	9994.0	28.656896	234.260108	-6599.978	1.72875	8.6665	29.364	8399.976
discount	9994.0	0.156203	0.206452	0.000	0.00000	0.2000	0.200	0.800
quantity	9994.0	3.789574	2.225110	1.000	2.00000	3.0000	5.000	14.000
ship_days	9994.0	3.958475	1.747603	0.000	3.00000	4.0000	5.000	7.000

3.2 Distributions and Anomalies(Sales histogram and Profit boxplot)





3.3 Relationships across variables

(Correlation matrix for 5 key variables and Discount vs Profit scatter plot)

	sales	profit	discount	quantity	ship_days
sales	1.000000	0.479064	-0.028190	0.200795	-0.007285
profit	0.479064	1.000000	-0.219487	0.066253	-0.004672
discount	-0.028190	-0.219487	1.000000	0.008623	0.000306
quantity	0.200795	0.066253	0.008623	1.000000	0.018494
ship_days	-0.007285	-0.004672	0.000306	0.018494	1.000000

3.4 Grouped comparison(Category pivot/summary table and Total Profit by Category chat)

category	orders	total_sales	total_profit	avg_discount	avg_ship_days
Technology	1847	836154.0330	145454.9481	0.132323	3.924201
Office Supplies	6026	719047.0320	122490.8008	0.157285	3.983239
Furniture	2121	741999.7953	18451.2728	0.173923	3.917963

3.5 Insights (evidence-based)

- **Sales is right-skewed:** Mean (229.86) is much higher than median (54.49), indicating a long right tail where a small number of high-value orders pull the average upward. This suggests median/quartiles are more representative of a “typical” sale than the mean. (**supported by 3.2 Sales histogram**).
- **Discount impacts profit:** Discount is negatively correlated with profit (corr = −0.219), and the scatter shows higher discounts are more often associated with low or negative profit. This implies discounting increases loss risk and is an important factor in profitability comparisons. (**supported by 3.3 scatter + correlation**).
- **Category performance differs:** Category summaries show Technology has the highest total profit while Furniture is the lowest, and the bar chart makes this gap clear. This suggests profitability varies meaningfully by category and should be explored further at the subcategory level. (**supported by 3.4 category pivot + profit bar chart**).

Reference

Doshi, V. (2020). *How to Describe a Dataset*. (Course reading)