SuperStore Dataset

Data Source

Data sourcing:
SuperStore Dataset
This is an external data source. The data is from Kaggle. This dataset is valuable for analyzing sales trends, customer behavior, and order history.

Data Collection:
Data is collected by Saad Haroon on Kaggle.

Data contents:
This dataset provides information related to sales orders. It includes details about orders, customers, products, and sales transactions. The data set contains 5901 rows and 23 columns.

Data ethics:
The Dataset does not contain any sensitive information, except customer name and city.

Limitations:
Data set has 2 empty columns which is not explained why they are left empty. There is no information on the data collection methods and it is unknown if the data set is trustworthy.

Data Profile

Variable	Description	Time variant/ invariant	Structured/ Unstructured	Qualitative/ Quantitative	Sub-types of qual/quant
Row ID	An identifier for each row of data.	Invariant	Structured	Quantitative	Discrete
Order ID	A unique identifier for each order made.	Invariant	Structured	Quantitative	Discrete
Order Date	The date when the order was placed.	Invariant	Structured	Quantitative	Discrete
Ship Date	The date when the order was shipped.	Invariant	Structured	Quantitative	Discrete
Ship Mode	The mode of shipping for the order.	Invariant	Unstructured	Qualitative	Binary
Customer ID	A unique identifier for each customer.	Invariant	Unstructured	Qualitative	Nominal
Customer Name	The name of the customer who placed the order.	Invariant	Structured	Qualitative	Nominal
Segment	The customer segment to which the customer belongs (e.g., retail, corporate).	Invariant	Structured	Qualitative	Nominal
Country	The country where the order was placed.	Invariant	Structured	Qualitative	Nominal
City	The city where the order was placed.	Invariant	Structured	Qualitative	Nominal

State	The state where the order was placed.	Invariant	Structured	Qualitative	Nominal
Region	The region where the order was placed.	Invariant	Structured	Qualitative	Nominal
Product ID	A unique identifier for each product.	Invariant	Structured	Qualitative	Ordinal
Category	The category to which the product belongs.	Invariant	Structured	Qualitative	Nominal
Sub-Category	A more specific sub-category within the product category.	Invariant	Structured	Qualitative	Nominal
Product Name	The name of the product.	Invariant	Structured	Qualitative	Nominal
Sales	The total sales amount for the order.	Invariant	Structured	Quantitative	Continuous
Quantity	The quantity of the product ordered.	Invariant	Structured	Quantitative	Discrete
Profit	The profit earned from the order.	Invariant	Structured	Quantitative	Continuous
Returns	Information about returns (if any).	Invariant	Structured	Quantitative	Discrete
Payment Mode	The mode of payment used for the order.	Invariant	Structured	Qualitative	Nominal
ind1	(Empty column)				
ind2	(Empty column)				

Descriptive analysis

	Row ID+O6G3A1:R6	Sales	Quantity	Profit	Returns	ind1	ind2
count	5901	5901	5901	5901	287	0	0
mean	5022.422471	265.345589	3.781901	29.700408	1	NaN	NaN
std	2877.977184	474.260645	2.212917	259.589138	0	NaN	NaN
min	1	0.836	1	-6599.978	1	NaN	NaN
25%	2486	71.976	2	1.7955	1	NaN	NaN
50%	5091	128.648	3	8.5025	1	NaN	NaN
75%	7456	265.17	5	28.615	1	NaN	NaN
max	9994	9099.93	14	8399.976	1	NaN	NaN

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5901 entries, 0 to 5900
Data columns (total 23 columns):

#	Column	Non-Null Count	Dtype
0	Row ID+O6G3A1:R6	5901 non-null	int64
1	Order ID	5901 non-null	object
2	Order Date	5901 non-null	object
3	Ship Date	5901 non-null	object
4	Ship Mode	5901 non-null	object
5	Customer ID	5901 non-null	object
6	Customer Name	5901 non-null	object
7	Segment	5901 non-null	object
8	Country	5901 non-null	object
9	City	5901 non-null	object
10	State	5901 non-null	object
11	Region	5901 non-null	object
12	Product ID	5901 non-null	object
13	Category	5901 non-null	object
14	Sub-Category	5901 non-null	object
15	Product Name	5901 non-null	object
16	Sales	5901 non-null	float64
17	Quantity	5901 non-null	int64
18	Profit	5901 non-null	float64
19	Returns	287 non-null	float64
20	Payment Mode	5901 non-null	object
21	ind1	0 non-null	float64
22	ind2	0 non-null	float64

dtypes: float64(5), int64(2), object(16)
memory usage: 1.0+ MB

Data Cleaning

Columns	Changes	Explanation
'ind1', 'ind2', 'Row ID+O6G3A1:R6', 'Order ID', 'Customer ID', 'Product ID','Ship Date','Order Date','Region','Product Name'	dropped	Not needed for analysis

Consistency checks

Consistency	Column name	Changes
Missing values	Returns column has 5614 missing values, which is more than 5%, ind1, ind2 columns are empty from the beginning	No changes were made for the returns column, so I could get some insights. ind1, ind 2 were dropped
Duplicates	none	

Questions to explore with analysis

1. What is the most profitable product?
2. Who is the client? (Corporates or consumers?)
3. What state needs more ad pushing in order to sell more?
4. What is the percentage of returns?
5. What is the most preferred shipping method? Do clients pay for first class shipping often?
6. Are there loyal customers to get promotions?