sns.histplot

my graph shows the distribution of product prices.

Most products are low-priced (around ₹0–₹500).

Very few products are high-priced (₹5000, ₹10000, ₹20000).

The graph is right-skewed — meaning many cheap items and few expensive ones.

This tells us that low-cost products are much more common in your dataset.

sns. scatterplot

This graph shows the relationship between Discount and Price.

Most prices are spread out across different discount levels.

High-priced products are sometimes sold without any discount (near 0.0).

Even when discounts are high (like 0.5 or 0.7), prices are generally lower.

There is no strong clear trend — discounts are given on both expensive and cheap products.

Sns. heatmap correlation

Price and Cost of Goods are strongly related — when one increases, the other increases too.

Quantity and other features have very weak relationships — they behave almost independently.

Discount is almost uncorrelated with everything — giving discounts doesn't directly relate to price, cost, quantity, etc.

Most other features have very low or no meaningful correlations (close to 0).

Sns.boxplot

Most prices are very small, close to zero.

Few prices are extremely high (up to 20,000+), but these are very rare — they are outliers.

sns.pairplot

Price vs Cost of Goods:

A strong straight-line pattern (going up) — meaning they are strongly positively correlated.

Cost of Goods increases when Price increases — like two travelers climbing the same mountain!

Price vs Discount:

Mostly scattered.

No strong pattern — giving a feeling of randomness (like leaves flying in the wind).

Price vs Quantity:

Most quantities are small even if prices vary a lot.

A cluster near low quantities — most people buy few items no matter the price!

Discount vs Cost of Goods:

No strong relationship — wild scattering.

Quantity vs Cost of Goods:

A mild upward trend — as quantity increases, the cost of goods slightly increases too (but not strongly).

a.load data_data =pd.read_csv(‘ .csv’)

data.head()

data.shap

the dataset contains 9994 row and 22 columns there are no missing values

.describe()

1. Row ID:

Count of records: 4997 rows.

Row IDs start from 1 and go up to 9994.

The IDs are evenly spread, indicating they uniquely identify each row.

2. Postal Code:

Postal codes range from 1040 to 99301.

There is a lot of variation (standard deviation ~32,063), suggesting wide geographical coverage.

3. Cost of Goods:

Mean cost is around 160.9.

Costs range from very low (~0.03) to very high (15,846).

Wide spread indicates that products of all cost types are available (cheap to expensive).

4. Price:

Average selling price is 229.85.

Minimum price is just 0.44, maximum goes up to 22,638.

High standard deviation (~623) shows some products are much more expensive than others.

5. Quantity:

On average, 3 items are ordered.

Most orders are between 2 and 5 items (based on 25%, 50%, and 75% quartiles

6. Discount:

Average discount given is 15.6%.

Most discounts are 0% (no discount) to 20%.

Maximum discount observed is 80% (rare cases of very high discounting).

7. Order Date and Ship Date:

Orders span from 2014 to 2017.

There are no missing values in order and ship dates.

It seems orders are shipped promptly after placing.

Data.info()

dtypes: datetime64[ns](2), float64(3), int64(3), object(14)

memory usage: 1.7+ MB

data.value_counts()

ship mode; most of ship mode standard class [5968]

> Country: Most of the sales are from [united states-9994].

City: [915-new York city] has the highest number of orders, showing concentration of customers there.

State: [California-2001] has the highest contribution.

Category: ["office supplies"-6026 category is the most sold].

Sub-Category: ["binders" and "Paper" are the most popular sub-categories.]

Binders-[1523]

Paper-[1370]