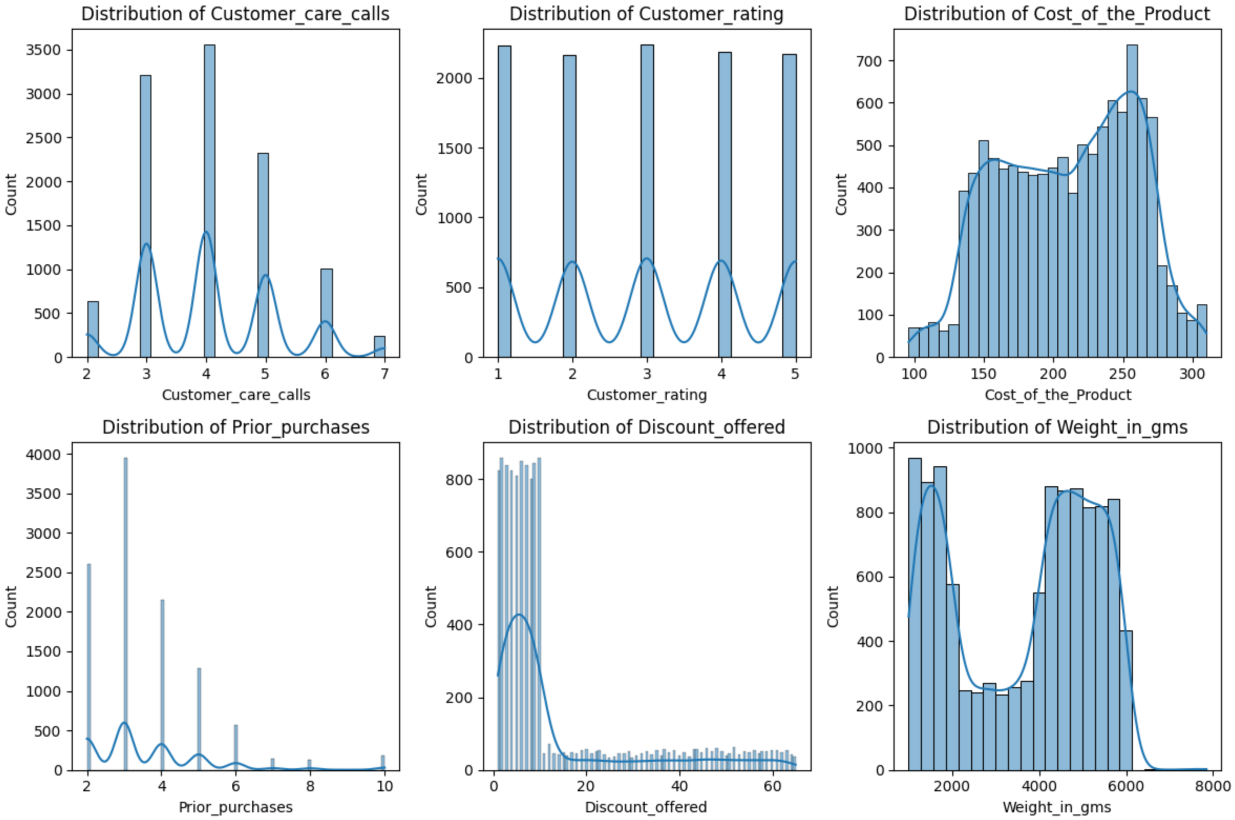
Data Understanding

The dataset used in this study is retrieved from online business data gathered from an online dataset repository, Kaggle, source mentioned in the appendices. The dataset contains 10,999 rows of data having the warehouse block where the product was stored prior to delivery, the number of calls that the customer care received, the customer rating, cost of the product, and whether it has reached in time or not.

- ID - id of the individual sales.
- Warehouse block - place where the Product was located.
- Mode of shipment - either Flight, Ship, or road based shipment.
- Customer care calls - The number of times the customer called about the status of the product.
- Customer rating - The satisfaction rating of the customer.
- Cost of the product - Cost in dollars.
- Prior purchase - the number of times the customer purchased the product.
- Product Importance - Shipment priority of the product.
- Gender - The Gender of the customer.
- Discount offered - How much was the customer offered in discounts (dollars).
- Weight in Grams - Weight of the product
- Reached on time - Whether the product arrived on schedule or not

The total number of sales in the dataset is 10,999 completed sales, we can visualize how these different features relate with each other by performing exploratory data analysis.



- Customer Care Calls - The distribution shows that the majority of customers made between 3 to 6 calls, with noticeable peaks at 3, 4, and 5 calls. The distribution has multiple modes, indicating specific numbers of calls are more common, possibly due to the customer service process.

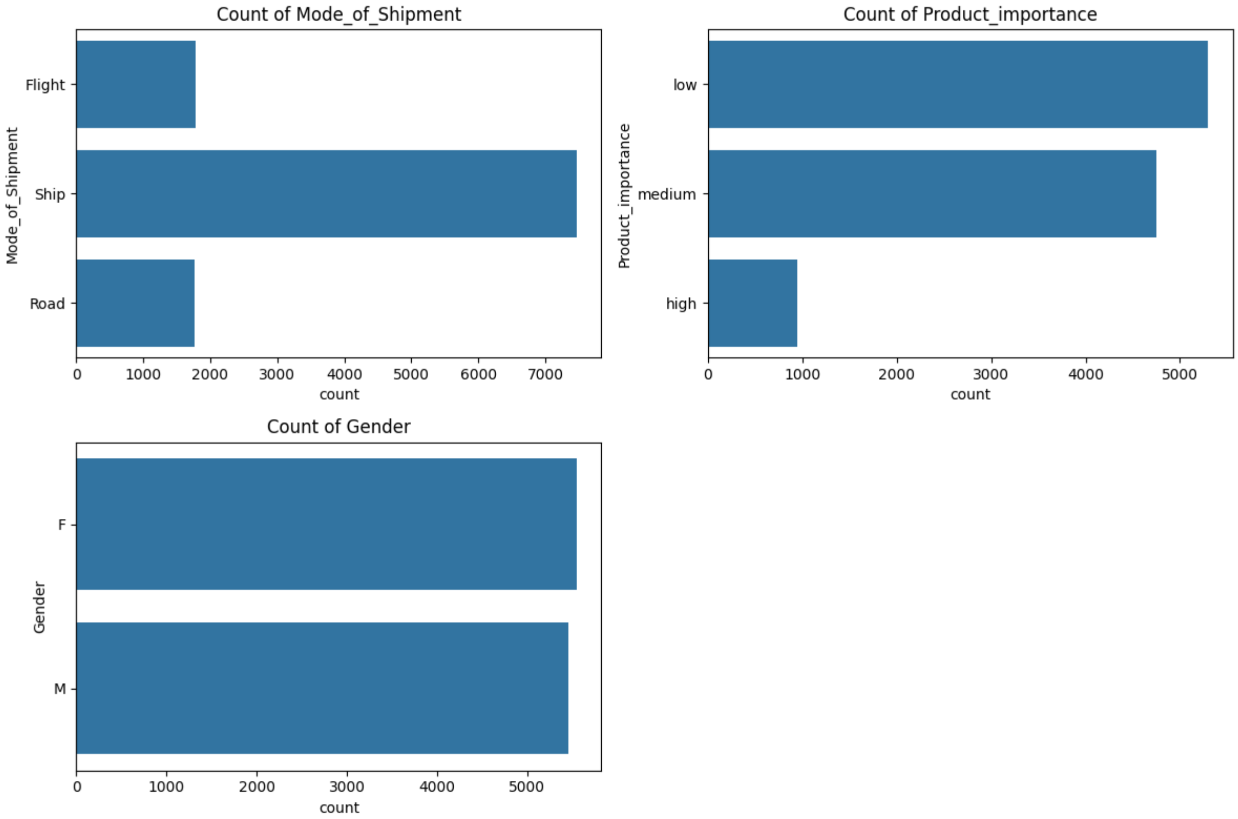
- Customer Rating - The customer rating distribution is almost uniform across all ratings from 1 to 5, suggesting a balanced feedback from customers, with no single rating overwhelmingly predominant.

- Cost of the Product - This distribution is roughly normal but skewed slightly to the right. The majority of products cost between 150 and 250 units, with a peak around 250. There are fewer products at the higher and lower ends of the cost spectrum.

- Prior Purchases - The distribution shows a steep decline as the number of prior purchases increases, with a large number of customers having made only 2 prior purchases. This indicates a trend where repeat purchases drop off significantly after a couple of transactions.

- Discount Offered - The distribution of discounts is heavily skewed towards the lower end, with most discounts offered being below 20 units. The frequency diminishes quickly as the discount value increases, indicating higher discounts are less commonly offered.

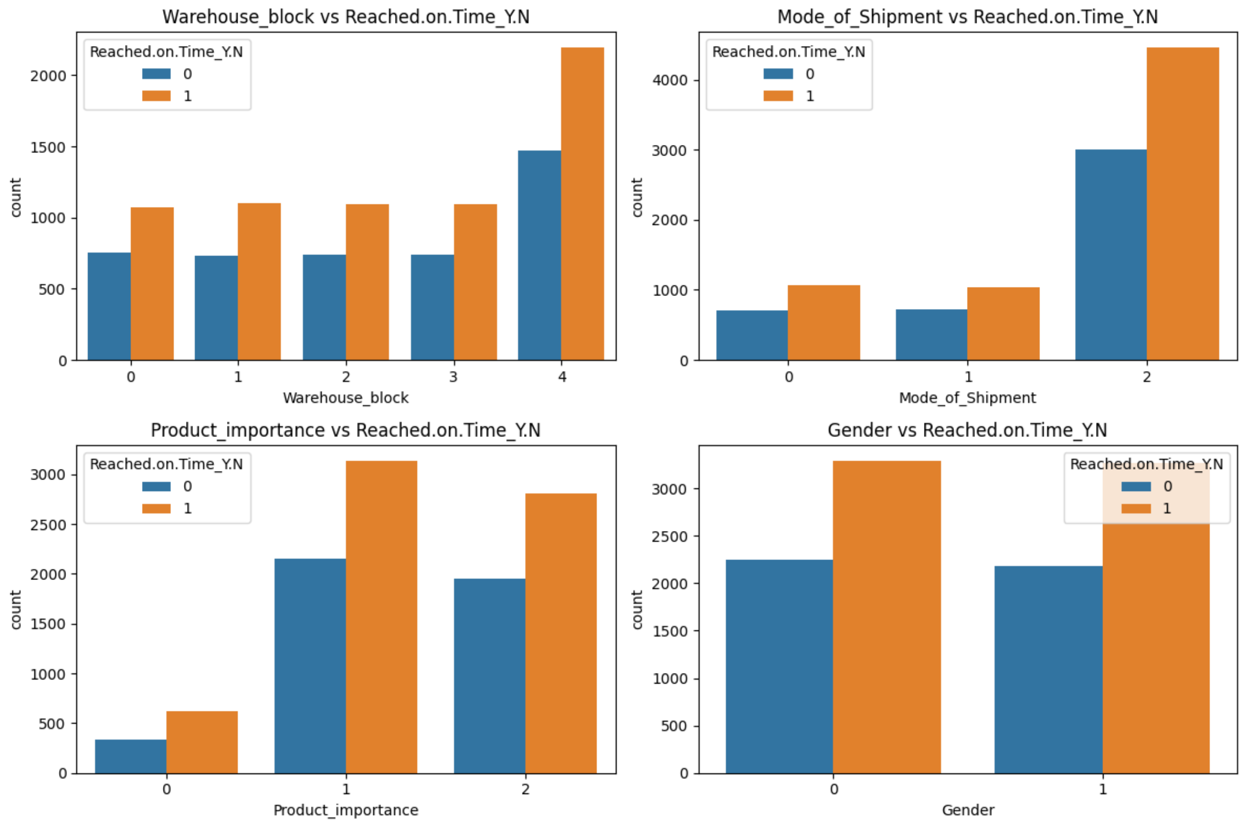
- Weight in Grams - The distribution of product weights shows a bimodal distribution with peaks around 2000-3000 grams and 5000-6000 grams. This suggests that the products fall into two main weight categories.



- Mode of Shipment - The bar plot shows that shipping by sea is the most common mode, followed by road and flight. This indicates a preference or practicality for shipping large quantities or heavy items by sea.

- Product Importance - The majority of products are categorized as either of low or medium importance, with high-importance products being the least common. This may reflect the nature of the inventory or customer purchasing patterns.

- Gender - The count of male and female customers appears to be nearly equal, indicating a balanced gender representation among the customer base.



Warehouse Block vs. Reached on Time:

- Warehouse Block 4 has the highest count of deliveries reaching on time, significantly more than those not reaching on time.
- Other blocks (0 to 3) show a more balanced distribution between on-time and delayed deliveries, with Block 2 and 3 having slightly more on-time deliveries than delayed ones.

Mode of Shipment vs. Reached on Time:

- Mode of Shipment 2 (Ship) has the highest count of on-time deliveries, followed by Mode 0 (Flight).
- Mode of Shipment 1 (Road) shows a more balanced distribution, but on-time deliveries still outnumber delayed ones.
- Across all shipment modes, on-time deliveries generally outnumber delayed ones.

Product Importance vs. Reached on Time:

- Products with low importance (0) have fewer counts overall, but the number of on-time deliveries is slightly higher than delayed ones.
- Medium importance products (1) have a higher count of on-time deliveries compared to delayed ones.

- High importance products (2) also show more on-time deliveries than delayed ones, but the difference is not as pronounced as in medium importance products.

Gender vs. Reached on Time:

- Female (0) and male (1) customers both show more on-time deliveries than delayed ones.
- The distribution is similar for both genders, with on-time deliveries consistently outnumbering delayed ones.

Based on this exploratory data analysis the key features are customer care calls, customer rating, product cost, delivery timeliness, Mode of Shipment, Product Importance based on their relationship with each other.