Data Source

1. Dataset: Superstore Dataset
    a. Source: https://www.kaggle.com/datasets/vivek468/superstore-dataset-final/data
    b. Data Type: Internal Data (as data such as this would be obtained through sales information within the business).
    c. Owner: Kaggle
    d. Trustworthiness: Dataset appears to be for educational purpose, so I am treating it as trustworthy.
2. Data Collection Method
    a. Data Type: Administrative Data
    b. Collection Method: The data collection is mainly automatic as the information would be collected through checkout software, but sales order and invoicing could be altered by employee manually to adjust for refunds and thus sometimes manual entries.
    c. Time Lag: I do imagine some time lag between payment received. In my personal ecommerce selling role, we do not receive payment right away and will sometimes get order cancelations and therefore the amount we could have received is no longer applicable to the profit amount. Therefore, I see there to be some time lag at times, but not much.
3. Reason for Choosing Dataset
    a. I chose this dataset based on my current work in ecommerce sells and marketing. Also, I hope to find a future position continuing my work in sales and marketing as a future sales analyst or marketing analyst. I felt that this dataset of product sales and profit with customer behavior patterns was a perfect fit towards this goal to showcase future employers.
4. Data Contents Overview
    a. Row ID: Identifying unique ID of row.
    b. Order ID: Identifying unique ID of orders.
    c. Order Date: Date product was ordered by customer.
    d. Ship Date: Date product was shipped out to customer.
    e. Customer ID: Identifying unique ID that is specific to each customer.
    f. Customer Name: Name of customer.
    g. Segment (Changed to Customer Segment): identifies segment in which customer belongs.
    h. Country: Country in which customer is located.
    i. City: City in which customer is located.
    j. State: State in which customer is located.
    k. Postal Code: Identifies customer's postal code.
    l. Region: Region in which customer is located.
    m. Product ID: Identifying unique ID of product.
    n. Category: States the category a product belongs.
    o. Sub-Category: States the more detailed category in which a product belongs.
    p. Product Name: States the specific name of product.
    q. Sales: Amount the product was sold for.
    r. Quantity: States the number of product(s) sold.
    s. Discount: States the percentage discount of product.
    t. Profit: States the revenue amount gained or lost from sale of product.

Data Profile

1. Data Cleaning Process
    a. Visualization checks of dataset.
        i. Inspected dataset column names.
        ii. Inspected data types.
    b. Converted data types of columns that were numerical to string types.
        i. This was done because the columns did not warrant numerical data type classification.
    c. Renamed columns.
        i. This was done to all columns to uniform to lower case as this is best practice for python column names.
        ii. Renamed original column “segment” to “customer_segment” as I felt this better portrayed column data.
    d. Dropped columns.
        i. Dropped column Row ID as this was irrelevant to the dataset.
        ii. Dropped customer name column for PII protections as the data contains a customer ID column.
    e. Checked data for mixed-type data
        i. Data did not result in any mixed-data types
    f. Checked data for duplicates.
        i. One duplicate was found and removed from dataset.
    g. Ran descriptive analysis check of processed dataset.
        i.

	sales	quantity	discount	profit
count	9993.000000	9993.000000	9993.000000	9993.000000
mean	229.852867	3.789753	0.156188	28.661048
std	623.276104	2.225149	0.206457	234.271571
min	0.440000	1.000000	0.000000	-6599.980000
25%	17.280000	2.000000	0.000000	1.730000
50%	54.480000	3.000000	0.200000	8.670000
75%	209.940000	5.000000	0.200000	29.360000
max	22638.480000	14.000000	0.800000	8399.980000

        ii. Checked for outliers using box-whisker plots for above pictured columns.
        iii. Checked frequency counts in segmented related columns.
    h. Derived New variable.
        i. Created a new column though grouping customer_id and order_id to create a total_orders column.
        ii. Created a new column monitoring customer order behavioral frequency classifying customers into three classes: Occasional Buyer, Consistent Buyer, and Loyal Buyers
    i. Final double checks of data

i. Ran one more check for missing values as I had computer technical difficulties with computer suddenly restarting for update and found a random missing value in all columns that was not there originally.
    1. Removed all missing values.
ii. Exported final dataframe.

Limitations & Ethic Considerations

Limitations:

- Dataset may not be fully representative of the company's entire customer base and could be missing data on diversity, such as customer demographics.
- Seasonal or economic events that impact certain items or timeframes may not be accurately represented that could have influenced customer purchasing behaviors.
- Data could focus too much on profit and ignore customer satisfaction and opinions.

Potential Biases:

- Sampling bias could have occurred favoring certain customers who frequently purchased or who had orders of noticeable losses in profit that could have affected dataset.
- Geographic bias could occur if the dataset favors data collection in one region over others and will not yield a generalized national market base.
- Data may be bias towards certain demographic customers and underrepresent other demographics as this information is not provided.

Potential Ethical Dilemmas:

- Without removing customer names from the dataset, the data could face customer privacy ethical issues with customer information being leaked or hacked.
- Possible transparency of data usage if customer have not been properly informed how their information may be used by the company.
- Data could focus too much on profit and ignore customer satisfaction and trust if company is too aggressive in pricing increase strategies.

Questions:

1) What are the purchasing patterns of buyer behavior segments and what pricing strategies can be implemented to maximize profit while maintaining customer satisfaction?
2) How have discounts impacted profits based on customer segments and buyer behavior segments?
3) What are the purchasing patterns of customers based on product category to so marketing and sales can tailor strategies to specific customers?
4) Which regions have the highest and lowest performing sales?
5) Which regions is the company's primary market base located?
6) Which customer segment or buyer segment contributes to the most profit and what strategies can help encourage better customer engagement from poorly performing segments?

7) Which months show to be peak order times and which months show to be the slowest for customer order placements?