Exploratory Data Analysis Assignment

In this assignment, you will perform **Exploratory Data Analysis (EDA)** on the attached CSV file containing Superstore order data. The purpose of this exercise is to help you build the habit of **understanding data before modeling or dashboarding** by checking data quality, interpreting distributions, and extracting meaningful business insights. Your submission should not only include code and plots, but also clear interpretations in plain language

Data Description:

Feature	Description
category	The category of products sold in the superstore.
city	The city where the order was placed.
country	The country in which the superstore is located.
customer_id	A unique identifier for each customer.
customer_name	The name of the customer who placed the order.
discount	The discount applied to the order.
market	The market or region where the superstore operates.
ji_lu_shu	An unknown or unspecified column.
order_date	The date when the order was placed.
order_id	A unique identifier for each order.
order_priority	The priority level of the order.
product_id	A unique identifier for each product.
product_name	The name of the product.
profit	The profit generated from the order.
quantity	The quantity of products ordered.
region	The region where the order was placed.
row_id	A unique identifier for each row in the dataset.

sales	The total sales amount for the order.
segment	The customer segment (e.g., consumer, corporate, or home office).
ship_date	The date when the order was shipped.
ship_mode	The shipping mode used for the order.
shipping_cost	The cost of shipping for the order.
state	The state or region within the country.
sub_category	The sub-category of products within the main category.
year	The year in which the order was placed.
market2	Another column related to market information.
weeknum	The week number when the order was placed.

Tasks:

Q1: Data Cleaning:

i. Count and handle Missing values.

ii. Deal with Duplicate values

iii. Delete unknown columns if present.

iv. Check shape, size, and datatypes of the dataset features.

Q2: Perform Univariate Analysis across all numerical features,

i. Which features seem useless in the analysis? Explain why?

ii. Which features are uniformly distributed or normally distributed?

iii. Which features are right-skewed/left-skewed? What does this signify?

iv. Which features have a high number of outliers, and discuss the impact.

Q 3: Perform Univariate Analysis across all categorical features.

i. Which features seem inaccurate and are not useful as categorical “insights” directly?

ii. What’s the issue with treating **Customer? Name** as a categorical feature for modeling?

iii. Is the **Category** distribution balanced or skewed? Give a 1-line reason.

iv. In **Country**, does one country dominate strongly? What does that imply about geographic bias?

v. Is the **City** dataset concentrated in a few cities or spread out?

Q 4: Perform Bivariant Analysis for numerical-to-numerical features:

i. Which two features are most strongly correlated?

ii. Also name features that are negatively correlated.

iii. If your goal is to understand **profit**, which are the most useful next bivariate checks? Perform them and give clear insights.

iv. Look for Time Effects Clues. Mention any information you find about the time relationship with any feature.

Q 5: Perform Bi-variant Analysis for categorical to numerical features.:

i. **Profit by Category:** Which category has the **highest median profit**? Which has the **lowest**? Which category shows the **widest spread (largest IQR)** in Profit? What does that suggest about profit consistency?

ii. **Sales by Category:** Which category has the **highest median sales**? Does it also have the highest median profit?

iii. **Profit by Segment:** Which segment has the **highest median profit**? Which segment has the **most negative/low profit outliers**?

iv. **Sales by Segment:** Which segment has the **highest median sales**? Is the profit pattern consistent with sales?

Q 6: Perform Bivariate Analysis for **Market** features against **Region, Category, and Country**:

i. Is the **Market** is not randomly spread across all **Regions**?

ii. Which country has negligible office supply orders?

iii. What are the most useful insights?

Submission Instructions:

To submit your assignment, please follow these guidelines:

- Ensure that your assignment is fully completed.
- Write answers to questions or insights right after the visualization.
- Make Sure to mention the question index before each solution. (E.g. : Q1 i, Q3 iii)
- Only the ". ipynb" file needs to be submitted. The ".py" format is not acceptable.
- Name the file as "firstname_lastname_Batch.ipynb" (E.g.: Kartik_Mudaliar_DEDEAI1.ipynb)
- Upload this file directly to Vlearn or upload it to GitHub, then share the repository link by including it in a text, Word, or PDF file. Make sure to keep the GitHub repo public. Submit the file/text containing the repository link via Vlearn.