

## JAMES ALEX EKORO

### PROJECT ON EXPLORATORY DATA ANALYSIS (EDA)

#### QUESTIONS.

- Ask meaningful questions about the dataset before analysis.
- Explore the data structure, including variables and data types.
- Identify trends, patterns and anomalies within the data.
- Test hypotheses and validate assumptions using statistics and visualization.
- Detect potential data issues or problems to address in further analysis.

#### ANSWERS.

- Ask meaningful questions about the dataset before analysis.

Before you start working with the data, it is important to clarify the analytical goals. The effectiveness of Exploratory Data Analysis (EDA) hinges on how well we can articulate the problem at hand.

##### A. Contextual Questions

- What's the purpose of the dataset?
- Who gathered the data and what was their method?
- What does each row signify (unit of observation)?
- What do the columns stand for (features/variables)?
- What period does the data cover?
- Are there any known issues with data collection?

##### B. Analytical Questions

- What are the main dependent and independent variables?
- Are there specific target variables (for prediction/classification)?
- What kind of relationships do we expect between the variables?
- Are there any possible confounding factors?
- What assumptions are made about the dataset (normality, independence, linearity)?

##### C. Data Quality Questions

- Are there any missing values?
- Are there duplicate entries?
- Are there any outliers?
- Are there any inconsistent formats or invalid data entries?

- Exploratory Data Analysis of Retail Sales Performance (Imagery Dataset)

Project Overview

This imaginary dataset was created to analyze sales performance for a retail company operating across 5 regions over 12 months.

Dataset Specifications

- Observations (Rows): 1,000 transactions
- Variables (Columns): 8
- Time Frame: January 2024 – December 2024

Variables Description

| Variable             | Type                 | Description                       |
|----------------------|----------------------|-----------------------------------|
| Transaction_ID       | Categorical          | Unique ID                         |
| Region               | Categorical          | North, South, East, West, Central |
| Month                | Categorical          | Jan–Dec                           |
| Units_Sold           | Discrete Numerical   | Quantity sold                     |
| Unit_Price (\$)      | Continuous Numerical | Price per unit                    |
| Revenue (\$)         | Continuous Numerical | Units_Sold × Unit_Price           |
| Marketing_Spend (\$) | Continuous Numerical | Advertising cost                  |
| Customer_Rating      | Continuous (1–5)     | Satisfaction score                |

- Meaningful Questions Before Analysis

1. Which region generates the highest revenue?
2. Does marketing spend significantly affect revenue?
3. Is customer satisfaction related to sales performance?
4. Are there seasonal sales trends?
5. Are there outliers affecting overall revenue?

- Data Structure Exploration

Summary Statistics (Imaginary Values)

- Units Sold: Mean=52, Median=50, Std Dev=18, Min=5, Max=120, Skewness=+0.85
- Unit Price: Mean=\$25, Std Dev=\$5, Range=\$15–\$40
- Revenue: Mean=\$1,300, Median=\$1,250, Std Dev=\$420, Max=\$4,200
- Marketing Spend: Mean=\$500, Std Dev=\$150, Range=\$200–\$900
- Customer Rating: Mean=3.8, Std Dev=0.6, Range=2.0–5.0

### Correlation Matrix (Key Relationships)

| Variables                 | Correlation (r) |
|---------------------------|-----------------|
| Marketing Spend & Revenue | 0.72            |
| Units Sold & Revenue      | 0.88            |
| Customer Rating & Revenue | 0.45            |

### Seasonal Trends - Monthly Average Revenue

| Month | Avg Revenue (\$) |
|-------|------------------|
| Jan   | 1,050            |
| Feb   | 1,100            |
| Mar   | 1,200            |
| Apr   | 1,250            |
| May   | 1,300            |
| Jun   | 1,350            |
| Jul   | 1,400            |
| Aug   | 1,380            |
| Sep   | 1,320            |
| Oct   | 1,450            |
| Nov   | 1,700            |
| Dec   | 2,100            |

### - Hypothesis Testing

Hypothesis 1: Marketing Spend significantly impacts revenue ( $r=0.72$ ,  $p<0.001$ ). Conclusion: Reject  $H_0$ .

Hypothesis 2: Regional revenue differences exist ( $F=6.45$ ,  $p=0.0003$ ). Conclusion: Significant regional differences.

### Multicollinearity (VIF)

| Variable   | VIF |
|------------|-----|
| Units Sold | 4.2 |

- **Key Findings**

1. You Strong relationship between marketing spend and revenue.
2. Units sold is the strongest driver of revenue.
3. Significant seasonal spikes in Q4.
4. Regional performance differences are statistically significant.
5. Customer satisfaction moderately impacts sales.
6. Few extreme revenue outliers driven by promotions.

- **Recommended Next Steps**

- Build multiple linear regression model:  $\text{Revenue} \sim \text{Units\_Sold} + \text{Marketing\_Spend} + \text{Region} + \text{Month}$
- Introduce interaction terms ( $\text{Marketing} \times \text{Region}$ ).
- Consider time-series forecasting for seasonal prediction.
- Optimize marketing budget allocation toward high-performing regions.

- **Final Conclusion**

The exploratory analysis reveals that revenue performance is primarily driven by units sold and marketing investment, with clear seasonal and regional variations. The dataset is suitable for predictive modeling after minor preprocessing.