EXPLORATORY DATA ANALYSIS

MYSQL WORKBENCH

DENIS KOMBE
BS.ICT | DATA ANALYST

Table of Contents

INTRODUCTION TO EXPLORATORY DATA ANALYSIS (EDA) 2
OBJECTIVES 2
THE ANALYTICAL APPROACH 3
THE SQL QUERIES 4
Overview of the dataset 4
Analyzing Sales Trends 5
Analyzing Sales Trends 5
Analyzing Sales Trends 6
Channel performance 6
Salespersons 7
Data Validation 8
Product & Sales Analysis 8
Channel, Leads and Conversion Rates. 11
City 12
CONCLUSION 13

INTRODUCTION TO EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis is a crucial step in the data analysis process, providing initial insights into the structure and characteristics of the data. Majorly, it involves summarizing the main features of the dataset to uncover patterns and trends. It helps in understanding the data better and lays the foundation for further analysis.

In this analysis, we focus on a sales dataset that includes various attributes such as product details, salesperson performance, advertising channel/medium effectiveness and also geographical (Country, City) sales distribution.

The primary objective is to explore these attributes to derive meaningful insights that can inform strategic decisions. Also to identify key trends, patterns and insights that can improve overall business performance.

OBJECTIVES

The key objectives of this EDA are:

i **Analyze Product Performance:**
   - Understand the distribution and revenue contribution of different products and product categories.

ii **Evaluate Sales Channels:**
   - Assess the effectiveness of different sales channels in terms of transaction volume, lead generation and conversion rates.

iii **Asses Salesperson Performance:**
   - Determine the performance of salespeople based on the number of transactions and total sales generated.

iv **Geographical Analysis:**
   - Examine the distribution of sales across various cities and countries, identifying the top performing regions in terms of sales volume and revenue.

THE ANALYTICAL APPROACH

The analysis is conducted through the following steps using MySQL Queries in MySQL Workbench and it includes the following.

a. **Product Analysis:**
   - Distribution of product names.
   - Top products. (measured by different conditions).

b. **Product category analysis:**
   - Distribution of product categories.
   - Top product categories. (measured by different conditions).

c. **Channel analysis:**
   - Distribution of sales by channel.
   - Total leads generated by channel.
   - Conversion rate by channel.

d. **Salesperson analysis:**
   - Performance of salesperson.
   - Top salesperson. (measured by different conditions).

e. **Geographical analysis:**
   - Distribution of sales by city.
   - Top cities by total sales revenue.
   - Distribution of sales by country.
   - Top countries by total sales revenue.

THE SQL QUERIES

Overview of the dataset

1.1. Summary Statistics

The summary statistics provide an overview of the dataset like the total number of records, average, minimum and maximum values for quantities, unit prices and total prices. This helps understand the overall scale and variability of the data.

```
-- 1. Overview of the Dataset
-- 1.1. Summary Statistics
SELECT
    COUNT(*) AS Total_Records,
    AVG(Quantity) AS Avg_Quantity,
    MIN(Quantity) AS Min_Quantity,
    MAX(Quantity) AS Max_Quantity,
    AVG(UnitPrice_In_Dollars) AS Avg_UnitPrice,
    MIN(UnitPrice_In_Dollars) AS Min_UnitPrice,
    MAX(UnitPrice_In_Dollars) AS Max_UnitPrice,
    AVG(TotalPrice_In_Dollars)  AS  Avg_TotalPrice,
    MIN(TotalPrice_In_Dollars)  AS  Min_TotalPrice,
    MAX(TotalPrice_In_Dollars)  AS  Max_TotalPrice
FROM sales_datac1;
```

1.2. Data distribution

Quantity:

This shows how frequent different quantities appear in the dataset.
Some salespeople had the similar number of products sold in terms of their quantities.

```
-- 1.2. Data Distribution
-- Quantity
SELECT Quantity, COUNT(*) AS Frequency
FROM sales_datac1
GROUP BY Quantity
ORDER BY Frequency DESC;
```

Unit Price:

This shows how frequent different unit prices appear in the dataset.
Some products had similar unit prices.

```
-- UnitPrice
SELECT UnitPrice_In_Dollars, COUNT(*) AS Frequency
FROM sales_datac1
GROUP  BY  UnitPrice_In_Dollars
ORDER BY Frequency DESC;
```

Analyzing Sales Trends

2.1. Sales Over Time

This helps in identifying trends.

```
-- 2. Analyzing Sales Trends
-- 2.1. Sales Over Time
SELECT
    OrderDate,
    SUM(TotalPrice_In_Dollars) AS Total_Sales
FROM sales_datac1
GROUP BY OrderDate
ORDER BY OrderDate;
```

2.2. Top products

This helps focus on high performing products.

```
-- 2.2. Top Products
SELECT
    ProductName,
    SUM(Quantity) AS Total_Quantity,
    SUM(TotalPrice_In_Dollars) AS Total_Revenue
FROM sales_datac1
GROUP BY ProductName
ORDER BY Total_Revenue DESC
LIMIT 10;
```

2.3. Sales by category

This helps understand which product categories are driving revenue.

```
-- 2.3. Sales by Category
SELECT
    ProductCategory,
    SUM(TotalPrice_In_Dollars) AS Total_Revenue
FROM sales_datac1
GROUP BY ProductCategory
ORDER BY Total_Revenue DESC;
```

Analyzing Sales Trends

3.1. Sales by Country

This helps identify the most profitable markets.

```
-- 3. Geographic Analysis
-- 3.1. Sales by Country
SELECT
    Country,
    SUM(TotalPrice_In_Dollars) AS Total_Revenue
FROM sales_datac1
GROUP BY Country
ORDER BY Total_Revenue DESC;
```

3.2. Sales by City

This provides insights into regional performance.

```
-- 3.2. Sales by City
SELECT
    City,
    SUM(TotalPrice_In_Dollars) AS Total_Revenue
FROM sales_datac1
GROUP BY City
ORDER BY Total_Revenue DESC;
```

Analyzing Sales Trends

4.1. Customer Segmentation

Segmenting customers based on their total spend helps identify high-value customers.

```
-- 4. Customer Analysis
-- 4.1. Customer Segmentation
SELECT
    CustomerID,
    SUM(TotalPrice_In_Dollars) AS Total_Spend
FROM sales_datac1
GROUP BY CustomerID
ORDER BY Total_Spend DESC;
```

Channel performance

5.1. Sales by channel

Helps in identifying the most effective sales channels.

```
-- 5. Channel Performance
-- 5.1. Sales by Channel
SELECT
    Channel,
    SUM(TotalPrice_In_Dollars) AS Total_Revenue
FROM sales_datac1
GROUP BY Channel
ORDER BY Total_Revenue DESC;
```

5.2. Conversion rates by channel

This is crucial to under the efficiency of each channel in turning leads to sales.

```
-- 5.2. Conversion Rates by Channel
SELECT
    Channel,
    AVG(ConversionRate) AS Avg_ConversionRate
FROM sales_datac1
GROUP BY Channel
ORDER BY Avg_ConversionRate DESC;
```

Salespersons

6.1. Performance of salespersons

Analyzing the number of sales per salesperson helps measure the performance of each salesperson.

```
-- 6. SalesPersons
-- 6.1. Performance of SalesPerson
SELECT
    Salesperson,
    COUNT(*) AS Number_Of_Sales
FROM sales_datac1
GROUP BY Salesperson
ORDER BY Number_Of_Sales DESC;

--  Finding the Number of SalesPersons
SELECT
    COUNT(DISTINCT  SalesPerson)  AS  Total_SalesPersons
FROM sales_datac1;
```

6.2. Total sales by each salesperson

```
-- 6.2. Total Sales by each SalesPerson
SELECT
    SalesPerson,
    SUM(TotalPrice_In_Dollars)  AS  Total_Sales_In_Dollars
FROM sales_datac1
GROUP BY Salesperson
ORDER BY Total_Sales_In_Dollars DESC;
```

6.3. Top products sold by each salesperson

```
-- 6.3. Top Products Sold by each Salesperson
SELECT
    SalesPerson,
    ProductName,
    COUNT(*) AS ProductCount
FROM sales_datac1 AS s1
GROUP BY Salesperson, ProductName
HAVING ProductCount = (
    SELECT COUNT(*) AS ProductCount
    FROM sales_datac1 AS s2
    WHERE s2.Salesperson = s1.Salesperson
    GROUP BY s2.Salesperson, s2.ProductName
    LIMIT 1
)
ORDER BY Salesperson, ProductCount DESC;
```

Data Validation

Validating the total price.

Validating the total price ensures data accuracy by comparing calculated and recorded total prices.

```
-- 7. Data Validation
-- 7.1. Validating the Total Price
SELECT
    Quantity,
    UnitPrice_In_Dollars,
    TotalPrice_In_Dollars,
    (Quantity * UnitPrice_In_Dollars) AS Calculated_Total_Price,
CASE
    WHEN TotalPrice_In_Dollars != (Quantity * UnitPrice_In_Dollars)
        THEN 'Mismatch'
    ELSE
        'Match'
END AS Validation
FROM sales_datac1 LIMIT 10;

-- NOTE: Data Cleaning on ProductName Vs ProductCategory
UPDATE sales_datac1
    SET ProductCategory = CASE
        WHEN ProductName='Aloe Vera' THEN 'Personal Care'
        WHEN ProductName='Kikoy Throw' THEN 'Fashion and Accessories'
        WHEN ProductName='Maasai Shuka' THEN 'Fashion and Accessories'
        WHEN ProductName='Sukuma Wiki' THEN 'Food and Beverages'
        WHEN ProductName='Chai Tea' THEN 'Food and Beverages'
        WHEN ProductName='Kitenge Fabric' THEN 'Fashion and Accessories'
        WHEN ProductName='Wooden Bowls' THEN 'Handicrafts and Home Decor'
        WHEN ProductName='African Hair Care Products' THEN 'Fashion and
          Accessories'
    ELSE ConversionRate
    END;
```

Product & Sales Analysis

7.1. Total Number of Orders Received per Product Category by Country.

To understand country-based preferences.

- GHANA

```
-- 8. Products Analysis
-- Total Number of Orders Received per product category in Ghana.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
```

```
WHERE Country = 'Ghana'
GROUP BY ProductCategory;
```

- KENYA

```
-- Total Number of Orders Received per product category in Kenya.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'Kenya'
GROUP BY ProductCategory;
```

- TANZANIA

```
-- Total Number of Orders Received per product category in Tanzania.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'Tanzania'
GROUP BY ProductCategory;
```

- ANGOLA

```
-- Total Number of Orders Received per product category in Angola.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'Angola'
GROUP BY ProductCategory;
```

- ETHIOPIA

```
-- Total Number of Orders Received per product category in Ethiopia.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'Ethiopia'
GROUP BY ProductCategory;
```

- NIGERIA

```
-- Total Number of Orders Received per product category in Nigeria.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
```

```
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'Nigeria'
GROUP BY ProductCategory;
```

- SOUTH AFRICA

```
-- Total Number of Orders Received per product category in Soutn Africa.
SELECT
    ProductCategory,
    COUNT(*),
    Country,
    SUM(Quantity) AS Total_Number_of_Orders
FROM sales_datac1
WHERE Country = 'South Africa'
GROUP BY ProductCategory;
```

7.2. Distribution of products

```
-- 8.2. Distribution of productNames
SELECT
    ProductName,
    COUNT(*) AS Frequency
FROM sales_datac1
GROUP BY ProductName
ORDER BY Frequency DESC;
```

7.3. Top products by total sales

```
-- 8.3. Top Products by Total Sales
SELECT
    ProductName,
    SUM(TotalPrice_In_Dollars) AS TotalSales
FROM sales_datac1
GROUP BY ProductName
ORDER BY TotalSales DESC;
```

7.4. Top Product category by total sales

```
-- 8.4. Top ProductCategory by Total Sales
SELECT
    ProductCategory,
    SUM(TotalPrice_In_Dollars) AS TotalSales
FROM sales_datac1
GROUP BY ProductCategory
ORDER BY TotalSales DESC;
```

7.5. Number of sales Transactions per country

```
-- 8.5. Distribution of Sales by Country. Focuses on Number of Sales
Transactions
SELECT
    Country,
    COUNT(*) AS NumberOfSales
FROM sales_datac1
GROUP BY Country
ORDER BY NumberOfSales DESC;
```

7.6. Total sales revenue per country

```
-- 8.6. Top Countries by Total Sales. Focuses on the total sales
revenue
SELECT
    Country,
    SUM(TotalPrice_In_Dollars) AS  TotalSales
FROM sales_datac1
GROUP BY Country
ORDER BY TotalSales DESC;
```

Channel, Leads and Conversion Rates.

8.1. Sales by channel

```
-- 9. Analysis on Channel, Leads and ConversionRates.
-- 9.1. Distribution of Sales by Channel
SELECT
    Channel,
    COUNT(*) AS NumberOfSales
FROM sales_datac1
GROUP BY Channel
ORDER BY NumberOfSales DESC;
```

8.2. Leads generated by channel

```
-- 9.2. Total leads generated by channel
SELECT
    Channel,
    SUM(LeadsGenerated) AS TotalLeads
FROM sales_datac1
GROUP BY Channel
ORDER BY TotalLeads DESC;
```

8.3. Conversion Rates

```sql
-- 9.3. Average conversion rates by channel
SELECT
    Channel,
    AVG(ConversionRate) AS AverageConversionRate
FROM sales_datac1
GROUP BY Channel
ORDER BY AverageConversionRate DESC;
```

8.4. Top products by leads

```sql
-- 9.4. Top Products by leads generated
SELECT
    ProductName,
    SUM(LeadsGenerated) AS TotalLeads
FROM sales_datac1
GROUP BY ProductName
ORDER BY TotalLeads DESC;
```

8.5. Top products by conversion rates

```sql
-- 9.5. Top Products by Conversion Rate
SELECT
    ProductName,
    AVG(ConversionRate) AS AverageConversionRate
FROM sales_datac1
GROUP BY ProductName
ORDER BY AverageConversionRate DESC;
```

City

9.1. Sales by city

```sql
-- 10. City-Based Analysis
-- 10.1. Distribution of Sales by City
SELECT
    City,
    COUNT(*) AS Numberofsales
FROM sales_datac1
GROUP BY City
ORDER BY Numberofsales DESC;
```

9.2. Top cities by revenue

```sql
-- 10.2. Top Cities by sales revenue
SELECT
    City,
    SUM(TotalPrice_In_Dollars) AS Totalsales
FROM sales_datac1
GROUP BY City
ORDER BY Totalsales DESC;
```

CONCLUSION

This analysis provided valuable insights into various areas.
Key findings include:

I. **Sales trends:**
- This analysis identified significant trends in sales over time.
- The top performing products and product categories were identified, hence providing a clear focus for inventory and marketing strategies.
- Based on the dataset used, the top product was *"Maasai Shuka"* with *1684485 total sales*. While the top product category was *"African Hair Products"* with *4039956 total sales.*

II. **Geographical analysis:**
- This revealed the most profitable countries and cities.
- This helps to pinpoint regions with the highest sales potential.
- This information can guide targeted marketing campaigns and regional expansion efforts.
- Based on the dataset used, the top Country in terms of sales was *"Angola"* while the top City was *"Luanda".*

III. **Customer analysis:**
- This helps in realizing customer behavior.
- Also highlights the importance of customer loyalty programs and personalized marketing efforts by salespersons.

IV. **Channel performance:**
- The performance of different sales channels was evaluated with insights into the most effective channels and their conversion rates.
- This information could be used to optimize marketing spend and also improve channel strategies.
- Based on the dataset used, the top channel based on leads generated was *"TikTok".*
- Based on the dataset used, the top channel based on total sales was *"TikTok"* with *1562926 sales.*
- Based on the dataset used, the top channel based on conversion rate was *"Facebook"* with a conversion rate of ***0.9***

V. **Salesperson performance:**
- The analysis of salespersons performance identified top performers and their best-selling products providing performance incentives..
- Based on the dataset used, the best salesperson was *"Bob Johnson"* with a total sale of *2068550.*