DS Report - 10

1.0 Data Gathering

- The dataset used in this analysis was obtained from Kaggle: **https://www.kaggle.com/datasets/ahmedmohamed2003/cafe-sales-dirty-data-for-cleaning-training**
- The dataset was obtained in CSV format.
- The file was checked to make sure it opened correctly
- The dataset was loaded into Python using the Pandas library.
- Loaded data contains transaction details including Transaction ID, Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, and Transaction Date.
- The data was stored in a DataFrame to allow easy handling and analysis.
- A quick preview of the data was done using:
- head() to view the first 5 rows.
- tail() to view the last 5 rows.

The shape of the dataset was checked to know the number of rows and columns which is **10000 rows × 8 columns.**

This step helped ensure that the data was successfully loaded and ready for assessment.

1.1 Data Assessment

- The structure of the dataset was explored using **info()** to:
    1. Check column names.

    2. Identify data types.
    3. Detect missing values.
- Missing values were checked using:
    1. **isnull()** and **sum()** to know how many missing values exist in each column.
- Duplicate records were identified using **duplicated()**.
- Basic statistics were reviewed using **describe()** to:
    1. Understand numerical data.
    2. Detect possible outliers.
- Unique values were checked for categorical columns to ensure consistency.
- The dataset was visually inspected to detect any unusual values or errors.

---

2. Project Objective

- The objective of this project is to analyze café sales data using Python and fundamental data science techniques.

  The project focuses on loading, cleaning, and organizing the dataset to make it suitable for analysis.
- Various visualizations were created, including:
    - Bar charts to represent total sales per item
    - Pie charts to show payment method distribution
    - Line charts to analyze sales trends over time
- The project aims to identify customer behavior patterns, popular products, and payment preferences.

  These insights can help the café improve business performance, plan future strategies, and make data-driven decisions.

3.0 Data Cleaning

To preserve the original dataset, a separate copy was created before performing any cleaning operations.

The following cleaning steps were applied:

- All text columns (Transaction ID, Item, Payment Method, Location) were converted to uppercase to ensure consistency.
- The Transaction Date column was converted to datetime format, and rows with invalid dates were removed.
- Quantity, Price Per Unit, and Total Spent columns were converted to numeric types, with invalid values coerced to NaN.
- Missing values in Quantity and Price Per Unit were filled using the median to reduce the effect of outliers.
- Missing Total Spent values were recalculated using the formula:
  Total Spent = Quantity × Price Per Unit
- Missing categorical values were replaced with **"UNKNOWN"**.

Logical validation was applied by removing records with non-positive values for Quantity, Price Per Unit, or Total Spent.

3.1 Outlier Detection and Handling

Outliers in the Total Spent column were detected using the Interquartile Range (IQR) method.

The first quartile (Q1) and third quartile (Q3) were calculated, and the IQR was derived as:
IQR = Q3 − Q1

Lower and upper bounds were defined as:

- Lower Bound = Q1 − 1.5 × IQR
- Upper Bound = Q3 + 1.5 × IQR

Instead of removing outliers, **capping** was applied to limit extreme values within acceptable ranges.

This approach preserves the dataset size while reducing the influence of extreme values.

3.2. Verification

After completing the cleaning process:

- The dataset was rechecked for missing values.
- No duplicate records were found.
- All numerical and categorical columns were confirmed to be consistent and standardized.

The dataset was verified to be clean and ready for analysis and modeling.

4.Data Visualization

(Bar Chart)

● The **Total Spent** column was converted to numeric format to ensure accurate calculations.

● Then the data was grouped by **Item** to calculate the total sales for each item.

● A **bar chart** was used to visualize and compare total sales across different item .

(Pie Chart)

● The data was analyzed to calculate the frequency of each **payment method** used.

● A **pie chart** was used to show the percentage distribution of different payment methods.

● This visualization helps in understanding the most and least commonly used payment Methods.

(line plot)

● The **Transaction Date** column was converted to datetime format to enable time-based analysis.

● Monthly sales were calculated by grouping the data by month and summing **Total Spent** values.

● A **line chart** was used to visualize the sales trend over time and identify changes in sales patterns

---

5. Data Science Techniques

- **Data Standardization:**
  Numerical features were standardized using StandardScaler to ensure equal contribution during clustering.
- **K-Means Clustering:**
  Transactions were segmented into three clusters representing low, medium, and high spenders.
  The Elbow Method was used to determine the optimal number of clusters.

---

6. Data insights

1. **Sales Performance**
   - Coffee is the top-selling item in terms of total revenue.
   - Cake and Cookie generate lower sales but act as important complementary products.
   - Some transactions have high Total Spent values, indicating bulk or multi-item purchases.

- Insight:
- Bundle offers such as Coffee + Dessert can increase the average transaction value.

2. **Payment Methods**
    - Credit Card and Digital Wallet are the most commonly used payment methods.
    - Customers show a clear preference for fast and convenient digital payments.
    - The presence of UNKNOWN values indicates incomplete payment data recording.
    - Recommendation:
    - Improving payment data accuracy will enhance customer behavior analysis.
3. **Location-Based Behavior**
    - In-store purchases are more frequent than Takeaway orders.
    - Most customers prefer consuming products inside the café.
    - UNKNOWN location values slightly affect the accuracy of location analysis.
    - Business Insight:
    - Enhancing the in-store customer experience may positively impact sales.
4. **Time-Based Sales Trends**
    - Monthly sales show a generally stable trend with moderate fluctuations.
    - Some months experience higher sales, possibly due to promotions or seasonal demand.
    - Insight:
    - Identifying high-performing months can help optimize future marketing strategies.
5. **Customer Segmentation**
    - Transactions were segmented into three clusters:
    - Cluster 0 – Low Spenders:
    - Low quantities and low total spending.
    - Cluster 1 – Medium Spenders:

- Moderate quantities and average spending levels. This is the largest segment.
- Cluster 2 – High Spenders:
- High quantities and the highest total spending. These represent high-value customers.
- Business Value:
- Targeting high spenders with loyalty programs can significantly increase revenue.

Final Conclusion

- Sales are driven mainly by a small number of core products.
- Digital payments dominate customer transactions.
- Customer segmentation provides actionable business insights.
- Data-driven analysis supports better strategic decision-making.

Team members

1. **Khloud sabry - 2406026: Data Gathering & Assessment.**
2. **Rana Nageh - 2306107: Project Objective**
3. **Basmala Sherif - 2406234: Data Cleaning.**
4. **Hajer Essam - 2406216: Data Visualization.**
5. **Eman saeed - 2406005: Data Science Technique.**
6. **Fatma Drar - 2406203 : Data insights**