

EDA Assignment – E-Commerce Dataset

QUESTION 1 – Bagian 1

1. Jelaskan konsep Exploratory Data Analysis (EDA) dan mengapa penting dalam Data Science Workflow.

Exploratory Data Analysis (EDA) adalah proses memahami struktur, karakteristik, dan pola dari dataset melalui analisis statistik dan visualisasi. Tujuannya adalah untuk menemukan anomali, memverifikasi asumsi, dan mendapatkan wawasan awal sebelum melakukan pemodelan. EDA penting karena membantu data scientist memahami konteks data, mengidentifikasi outlier, serta memastikan kualitas data agar hasil model lebih akurat.

2. Jelaskan manfaat dari beberapa teknik visualisasi berikut:

- Histogram: Menunjukkan distribusi frekuensi dari variabel numerik, membantu melihat pola penyebaran data.
- Boxplot: Berguna untuk mendeteksi outliers dan memahami rentang interkuartil data.
- Scatter Plot: Memvisualisasikan hubungan antara dua variabel numerik dan pola korelasinya.
- Heatmap: Menunjukkan kekuatan hubungan (korelasi) antar variabel dalam bentuk visual matriks warna.

3. Studi Kasus Nyata:

Dalam industri e-commerce, EDA digunakan untuk memahami perilaku pelanggan. Misalnya, analisis hubungan antara jumlah pembelian (Quantity), diskon (Discount), dan profit. Dengan EDA, tim bisnis dapat mengidentifikasi bahwa diskon besar tidak selalu meningkatkan profit, sehingga mereka dapat menyesuaikan strategi promosi.

QUESTION 2 – Bagian 2: Praktik Eksplorasi Data

Dataset yang digunakan diambil dari E-commerce Dataset (Kaggle) yang berisi informasi penjualan pelanggan, seperti Gender, Sales, Quantity, Discount, Profit, dan Shipping Cost.

Langkah-langkah Eksplorasi Data

1. Menampilkan 5 baris pertama dataset:

```
df.head()
```

2. Menampilkan jumlah missing values per kolom:

```
df.isnull().sum()
```

3. Menampilkan statistik deskriptif dasar:

```
df.describe(include='all')
```

4. Visualisasi Data

Kode di bawah ini digunakan untuk menghasilkan visualisasi menggunakan Seaborn dan Matplotlib.

- Histogram:

```
if "Sales" in df.columns:  
    plt.figure(figsize=(7,5))  
    sns.histplot(df["Sales"], bins=30, kde=True)  
    plt.title("Histogram: Distribution of Sales")  
    plt.xlabel("Sales")  
    plt.ylabel("Frequency")  
    plt.show()
```

- Boxplot:

```
if "Profit" in df.columns:  
    plt.figure(figsize=(7,5))  
    sns.boxplot(y=df["Profit"])  
    plt.title("Boxplot: Profit Distribution")  
    plt.ylabel("Profit")  
    plt.show()
```

- Scatter Plot:

```
if "Quantity" in df.columns and "Profit" in df.columns and "Gender" in df.columns:  
    plt.figure(figsize=(7,5))  
    sns.scatterplot(data=df, x="Quantity", y="Profit", hue="Gender", alpha=0.7)  
    plt.title("Scatter Plot: Quantity vs Profit by Gender")  
    plt.xlabel("Quantity")  
    plt.ylabel("Profit")  
    plt.show()
```

- Heatmap:

```
numeric_df = df.select_dtypes(include=["float64", "int64"])  
if not numeric_df.empty:  
    plt.figure(figsize=(8,6))  
    sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm", fmt=".2f")  
    plt.title("Heatmap: Korelasi Antar Variabel Numerik")  
    plt.show()  
else:  
    print("Tidak ada kolom numerik yang tersedia untuk korelasi.")
```

5. Interpretasi Hasil Eksplorasi

- Histogram menunjukkan sebagian besar nilai Sales berada pada kisaran rendah hingga menengah, menandakan sebagian besar transaksi bernilai kecil.
- Boxplot memperlihatkan adanya beberapa outlier pada Profit, menunjukkan ada transaksi dengan laba sangat tinggi atau sangat rendah.
- Scatter Plot memperlihatkan bahwa hubungan Quantity dan Profit bervariasi antar Gender. Beberapa pelanggan melakukan pembelian besar dengan profit tinggi.
- Heatmap menunjukkan korelasi positif antara Sales dan Profit, serta korelasi negatif kecil antara Discount dan Profit.

6. Tantangan dan Solusi

- Tantangan: Dataset memiliki missing values dan outlier yang bisa memengaruhi analisis.
- Solusi: Missing values diatasi dengan imputasi atau penghapusan baris yang tidak lengkap. Outlier diperiksa lebih lanjut apakah wajar (misalnya dari promo besar) atau perlu dihapus.

EDA membantu memahami struktur data sebelum membuat model prediksi atau laporan bisnis yang akurat.