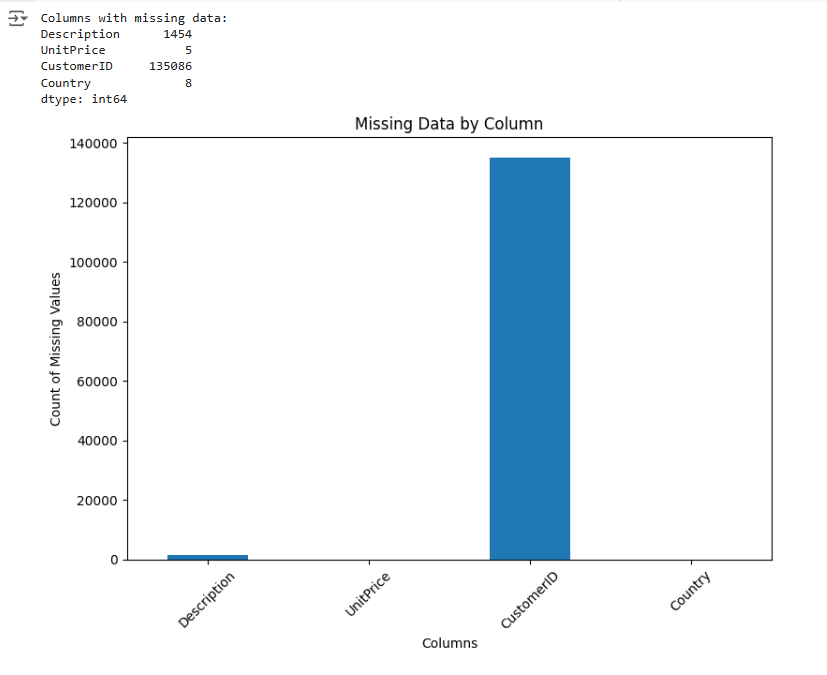
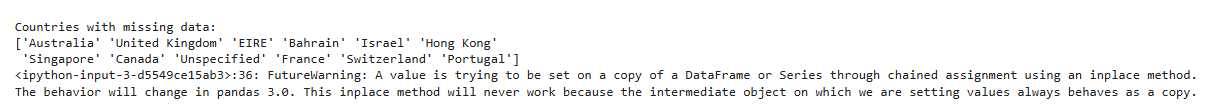
Jawaban UTS DATA MINING

1. Jawab Pertanyaan :
2. Colum Kosong



1. Negara missing data



1. Lakukan dan jawab:

b.

Langkah-langkah yang diambil dalam melakukan data cleaning pada dataset:

1. Identifikasi Missing Data:

- Missing data diidentifikasi menggunakan isnull() yang menghitung jumlah nilai NaN di setiap kolom.

- Kolom dengan missing data adalah:

- Description

- UnitPrice

- CustomerID

- Country

2. Metode Pengisian Missing Data:

- Kolom Numerik (UnitPrice, CustomerID):

- Missing data pada kolom numerik diisi dengan median. Median dipilih karena tidak dipengaruhi oleh outlier, sehingga hasil imputasi lebih stabil.

- Kolom Kategorikal (Descriptio, Country):

- Missing data pada kolom kategorikal diisi dengan mode. Mode dipilih karena merepresentasikan nilai yang paling sering muncul dan mempertahankan konsistensi data.

3. Implementasi dalam Kode:

- Dilakukan iterasi pada semua kolom dataset.

- Jika kolom bertipe numerik (float64 atau int64), missing data diisi menggunakan fillna(median).

- Jika kolom bertipe kategorikal (object), missing data diisi menggunakan `fillna(mode)`.

- Semua perubahan diterapkan langsung pada dataset dengan inplace=True.

4. Validasi Cleaning:

- Setelah cleaning selesai, dataset diperiksa kembali menggunakan isnull().sum() untuk memastikan tidak ada lagi nilai NaN.

1. Bagaimana nilai untuk data baru ditentukan:

1. Kolom Numerik:

- UnitPrice:

- Nilai missing diisi dengan median `UnitPrice` dari dataset karena median memberikan representasi tengah dari distribusi data tanpa terpengaruh oleh outlier.

- CustomerID:

- Nilai missing diisi dengan median `CustomerID` karena diasumsikan setiap ID bersifat unik tetapi tetap mencerminkan urutan dalam dataset.

2. Kolom Kategorikal:

- Description:

- Nilai missing diisi dengan mode (nilai yang paling sering muncul). Deskripsi barang yang paling sering muncul diasumsikan relevan untuk menggantikan data yang hilang.

- Country:

- Nilai missing diisi dengan mode, yaitu negara dengan frekuensi terbanyak dalam dataset. Hal ini diasumsikan untuk menjaga konsistensi dengan pola data.

3. Justifikasi Penggunaan Median dan Mode:

- Median digunakan untuk kolom numerik karena:

- Tidak terpengaruh oleh outlier yang dapat mendistorsi nilai rata-rata.

- Representasi nilai tengah memberikan hasil imputasi yang lebih konsisten.

- Mode digunakan untuk kolom kategorikal karena:

- Menggantikan nilai yang hilang dengan data yang paling sering muncul, sehingga mempertahankan pola distribusi dataset.