**UTS DATA MINING DAN BUSINES INTELLIGENCE**

Nama : Uday Akhmal Firmansyah

Nim : 32602200128

1. Jawab pertanyaan berikut:
2. Pada kolom apa saja ada missing data?

Jawaban :

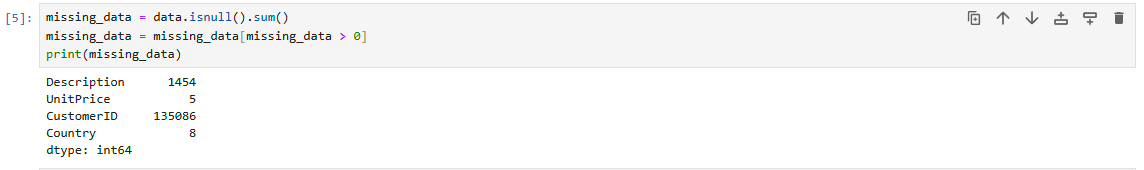
Kolom yang memiliki missing data yaitu  
Description 1454

UnitPrice 5

CustomerID 135086

Country 8

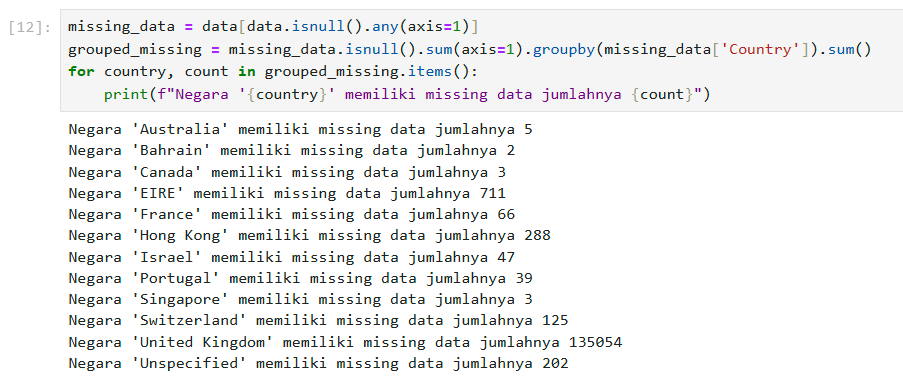
Berikut buktinya :



1. Pada negara apa saja ada missing data?

Jawaban :

Australia , Bahrain, Canada, EIRE, France, Hong Kong, Israel, Portugal, Singapore, Switzerland, United Kingdom, Unspecified

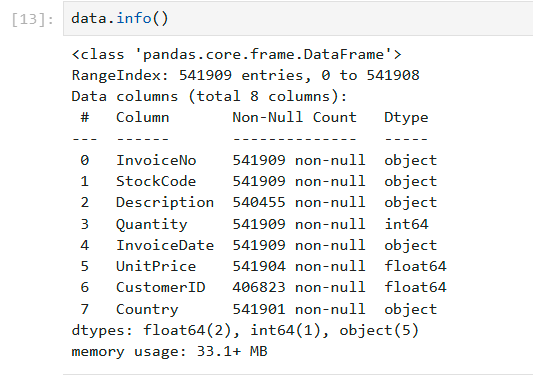


1. Lakukan dan jawab:
2. Jelaskan bagaimana Anda melakukan data cleaning?

**Langkah 1: Memahami Struktur Data**

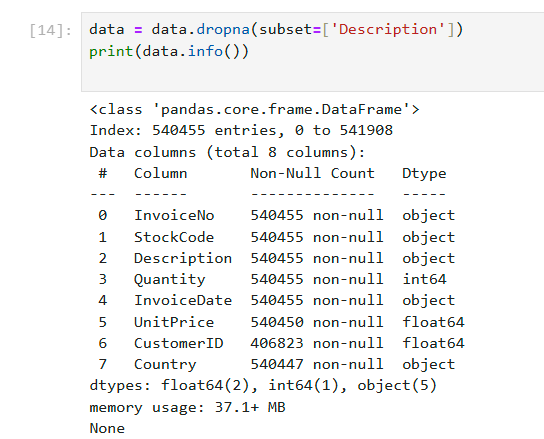
Sebelum membersihkan data, kita harus memahami strukturnya, seperti:

* Kolom yang tersedia.
* Jenis data di setiap kolom (numerik, string, tanggal, dll.).
* Apakah ada nilai yang hilang atau inkonsisten.

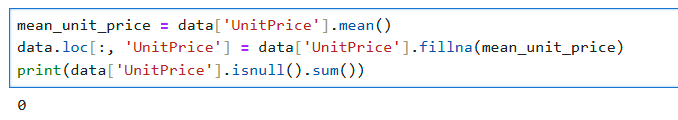


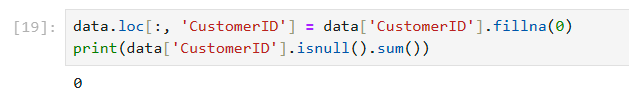
**Langkah 2: Menangani Nilai yang Hilang**

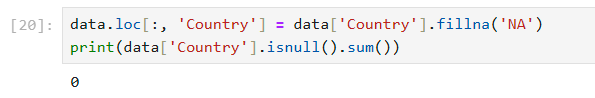
* Menghapus Baris dengan Missing Data di Kolom Description

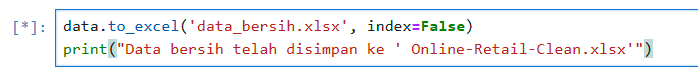


* Mengisi Nilai Hilang di Kolom UnitPrice dengan Rata-Rata



* Mengisi Nilai Hilang di Kolom CustomerID dengan 0  
  
* Mengisi Nilai Hilang di Kolom Country dengan "NA"



* Langkah terakhir adalah menyimpan data yang sudah bersih  
  

1. Jika Anda mengisi missing data dengan data baru, jelaskan bagaimana anda mendapatkan setiap value untuk data baru tersebut masing-masing!

Jawaban :

**Kolom UnitPrice: Diisi dengan nilai rata-rata**

* **Mengapa rata-rata dipilih?**
  1. Kolom ini merepresentasikan harga satuan suatu produk. Nilai rata-rata memberikan estimasi yang netral, tidak terlalu rendah atau terlalu tinggi.
  2. Rata-rata sering digunakan saat tidak ada informasi lain untuk mengisi data hilang, karena nilai ini merepresentasikan titik tengah data yang ada.
* **Bagaimana nilai rata-rata diperoleh?**
  1. Nilai rata-rata dihitung menggunakan:

mean\_unit\_price = data['UnitPrice'].mean()

* 1. Metode ini menjumlahkan semua nilai yang tersedia lalu membaginya dengan jumlah nilai non-hilang.

**Kolom CustomerID: Diisi dengan 0**

* **Mengapa 0 dipilih?**
  1. Nilai ini dipilih untuk menunjukkan bahwa data tentang pelanggan tidak tersedia (tidak dikenal).
  2. 0 adalah pilihan logis karena kolom ini bertipe numerik, dan ini menghindari konflik tipe data.
* **Bagaimana nilai 0 diperoleh?**

data.loc[:, 'CustomerID'] = data['CustomerID'].fillna(0)

**Kolom Country: Diisi dengan "NA"**

* **Mengapa "NA" dipilih?**
  1. "NA" adalah label tekstual yang menunjukkan "Tidak Tersedia" tanpa membuat asumsi tentang negara.
  2. Pilihan ini mempertahankan data lainnya di baris yang sama tanpa perlu menghapus baris tersebut.
* **Bagaimana nilai NA diperoleh?**

data.loc[:, 'Country'] = data['Country'].fillna('NA')