**LAPORAN PRA-PEMROSESAN DATA**

Nama : Faliqul Ishbah  
Nim : 202210370311148

**1. Pendahuluan**

**1.1 Latar Belakang**

Dalam proses Knowledge Discovery in Databases (KDD), kualitas data sangat menentukan hasil analisis dan model prediksi. Data yang kotor dapat menyebabkan kesalahan dalam analisis dan prediksi yang tidak akurat. Oleh karena itu, tahapan pra-pemrosesan data menjadi sangat penting untuk meningkatkan kualitas data sebelum digunakan dalam proses analisis lebih lanjut.

Laporan ini menjelaskan langkah-langkah pra-pemrosesan data yang telah dilakukan menggunakan bahasa pemrograman **Python** dengan pustaka **Pandas, NumPy, dan Scikit-learn**. Fokus utama dari pra-pemrosesan ini meliputi:

* Menangani nilai yang hilang
* Menghapus duplikasi
* Normalisasi dan standarisasi data
* Memeriksa konsistensi
* Deteksi outlier

Untuk melihat Keseluruhan Code bisa Langsung akses [repository Github](https://github.com/fall1sh/Data-Informasi-dan-Pengetahuan)

**2. Dataset yang Digunakan**

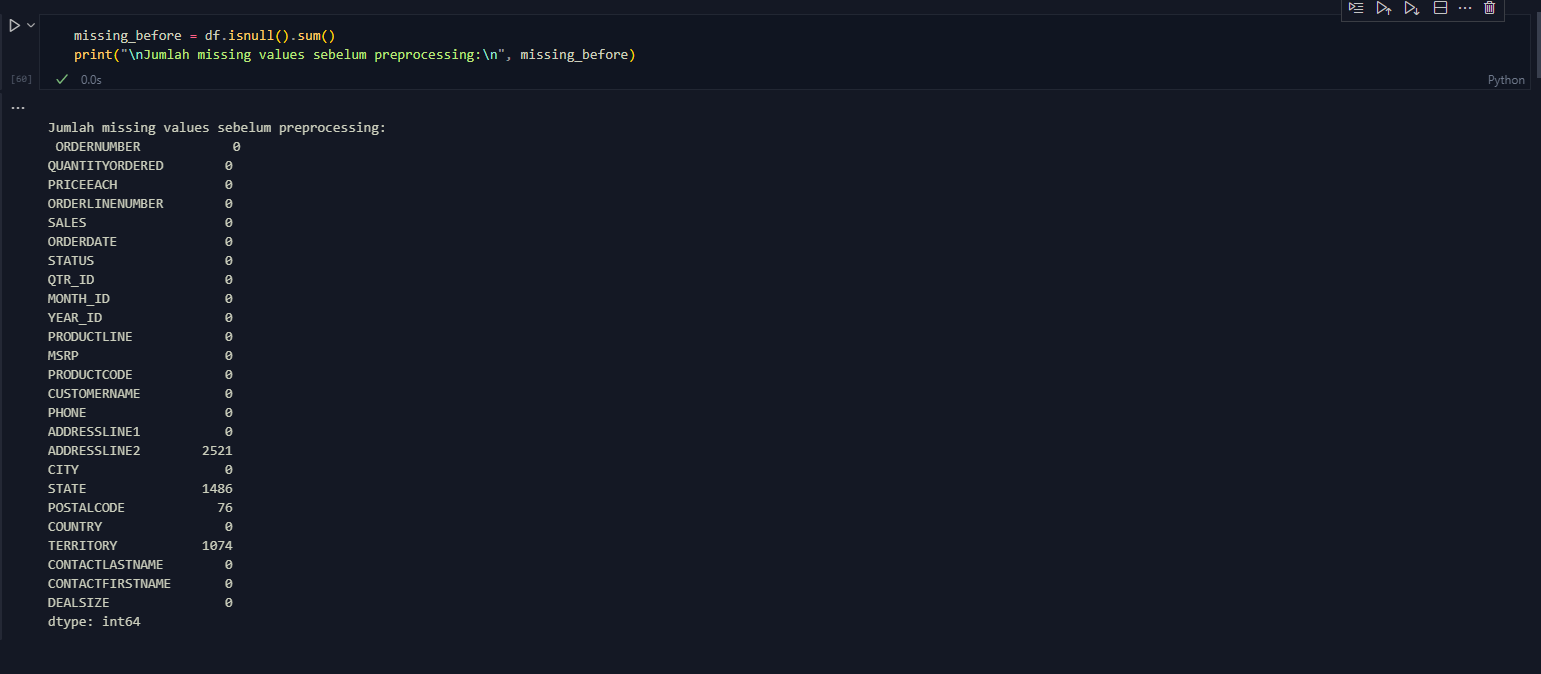
Dataset yang digunakan dalam proses ini adalah dataset [penjualan dari sebuah perusahaan](https://www.kaggle.com/datasets/kyanyoga/sample-sales-data/data). Dataset ini berisi informasi transaksi penjualan, termasuk **jumlah pesanan, harga satuan, total penjualan, status pesanan, tanggal pesanan, dan lain-lain**.

**3. Langkah-Langkah Pra-Pemrosesan Data**

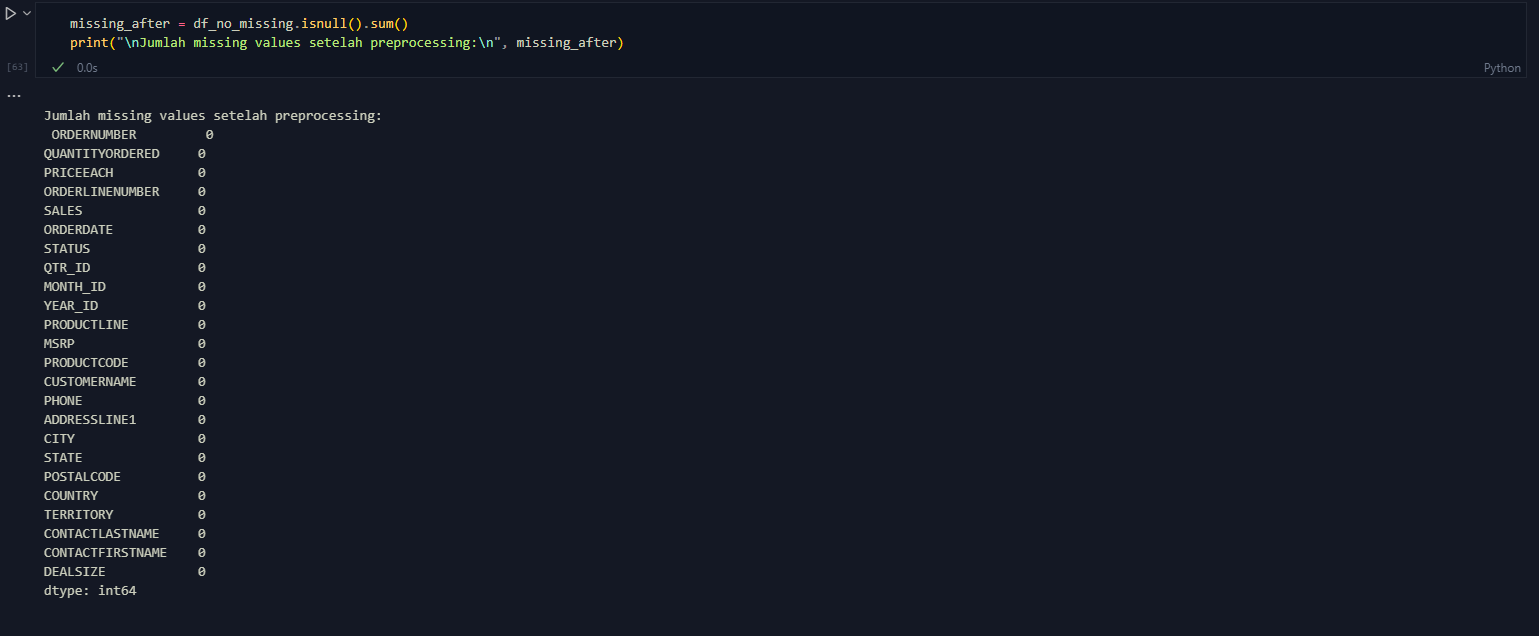
**3.1 Menangani Nilai yang Hilang**

Nilai yang hilang dalam dataset dapat mengganggu analisis, sehingga perlu ditangani dengan beberapa teknik seperti imputasi, penghapusan, atau pengisian dengan nilai default. Berikut adalah langkah-langkah yang dilakukan:

* Mengecek jumlah nilai yang hilang dalam dataset:



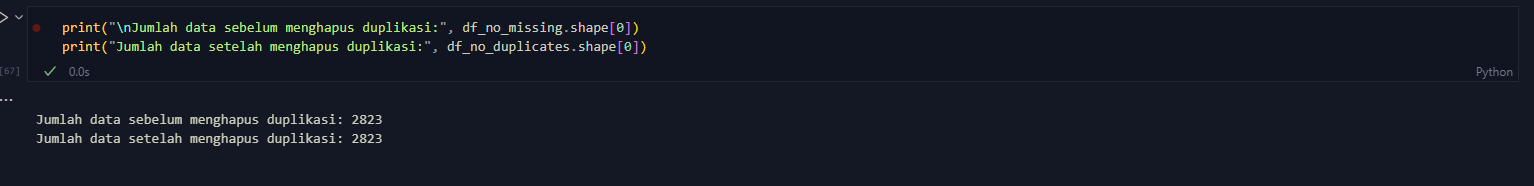
* Menghapus kolom dengan terlalu banyak nilai hilang:
* Mengisi nilai yang hilang pada kolom kategori dengan modus:



**3.2 Menghapus Duplikasi Data**

Duplikasi data dapat menyebabkan hasil analisis yang bias. Oleh karena itu, dilakukan pengecekan dan penghapusan duplikasi dengan cara berikut:

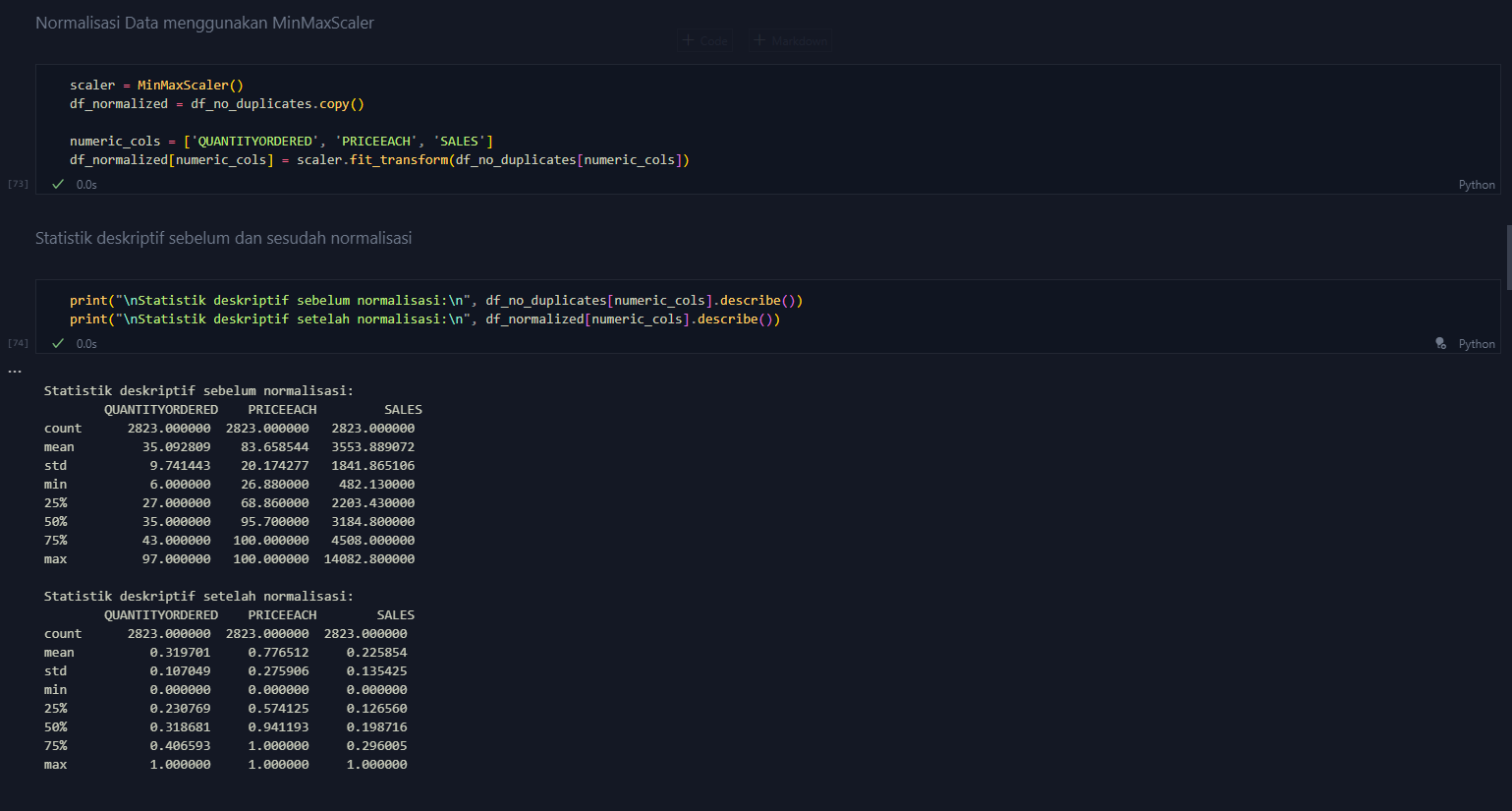
* Mengecek jumlah duplikasi:
* Menghapus data duplikat:



**3.3 Normalisasi dan Standarisasi Data**

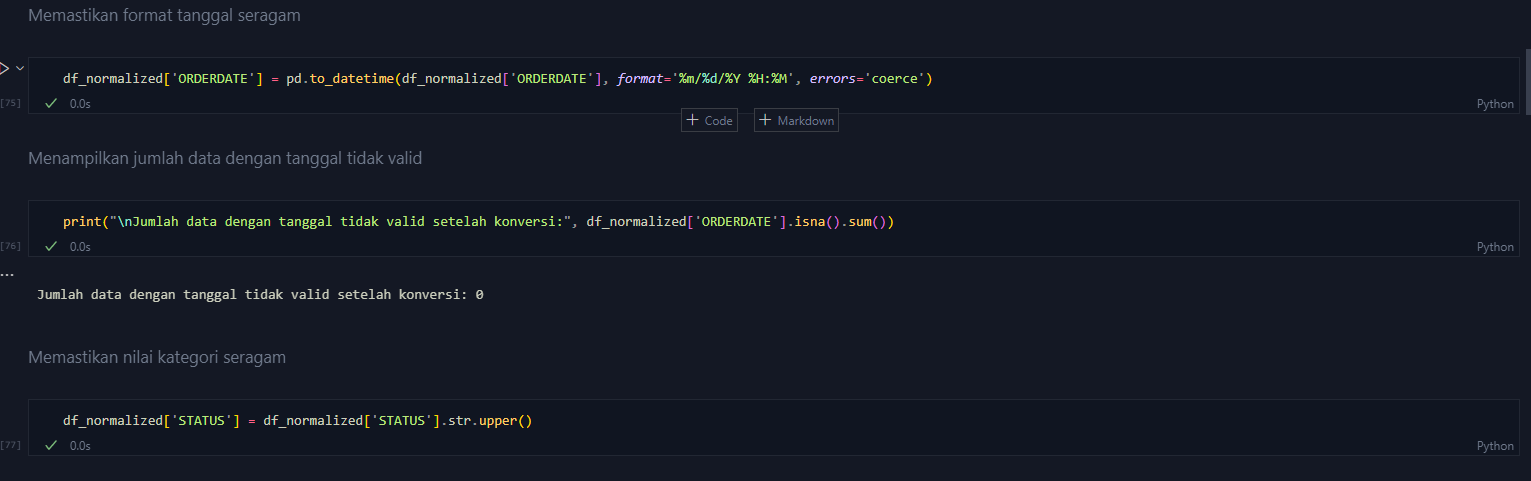
Agar data memiliki skala yang seragam, dilakukan normalisasi menggunakan **Min-Max Scaling** pada beberapa kolom numerik:

* Normalisasi dengan Min-Max Scaling:



**3.4 Memeriksa Konsistensi Data**

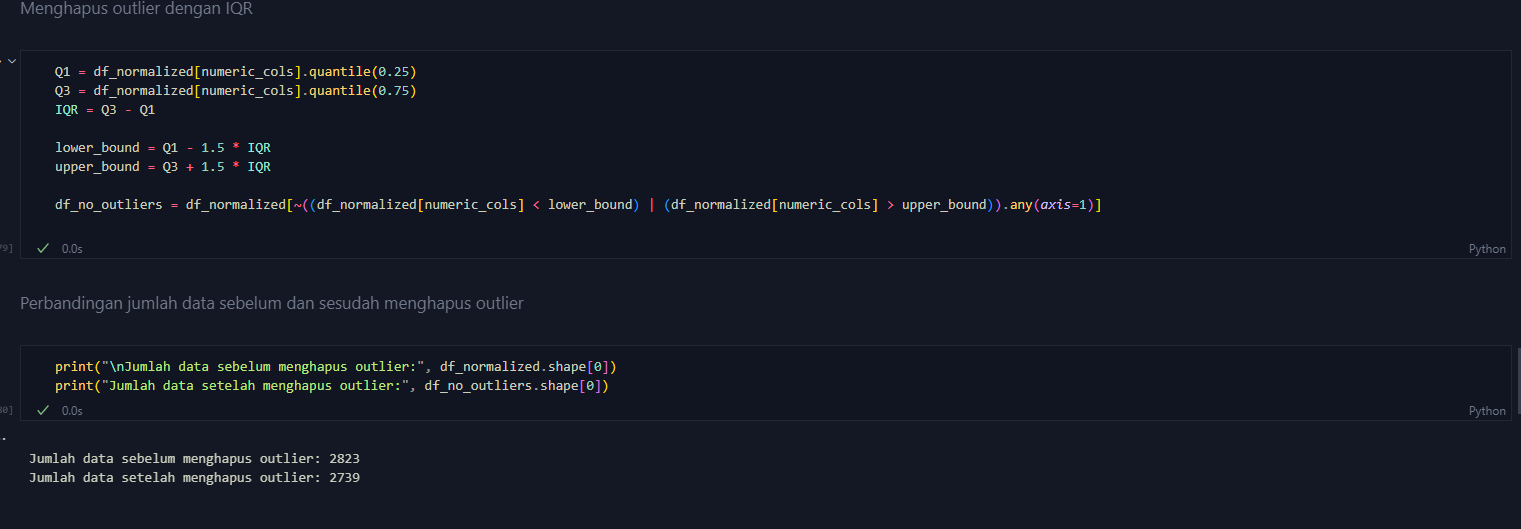
* Memastikan format tanggal seragam:
* Menyeragamkan format teks pada kolom kategori:



**3.5 Deteksi dan Penanganan Outlier**

Boxplot digunakan untuk mendeteksi outlier dalam data numerik, kemudian outlier dihapus berdasarkan **Interquartile Range (IQR)**.

* Menghitung batas IQR untuk deteksi outlier:

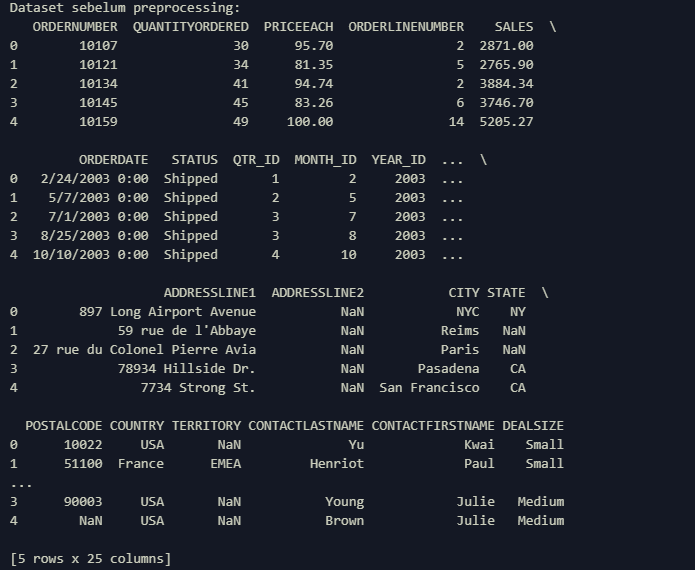


**4. Hasil dan Perbandingan Sebelum vs Sesudah Preprocessing**

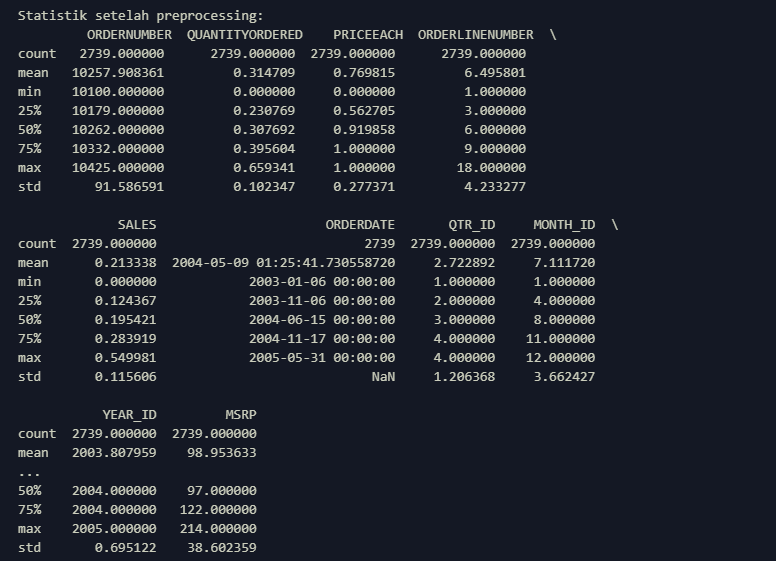
**4.1 Statistik Deskriptif Sebelum dan Sesudah**

Setelah melakukan pra-pemrosesan, dilakukan perbandingan statistik deskriptif sebelum dan sesudah pembersihan data:

* Sebelum preprocessing:



* Setelah preprocessing:



Dari hasil di atas, terlihat bahwa:

* Data lebih bersih dengan **tidak ada nilai yang hilang**
* **Distribusi data lebih seragam** setelah normalisasi
* **Outlier sudah dihilangkan**, sehingga hasil analisis akan lebih akurat

**5. Kesimpulan**

Pra-pemrosesan data sangat penting untuk memastikan kualitas data sebelum digunakan dalam analisis lebih lanjut. Dalam laporan ini, berbagai teknik telah diterapkan untuk membersihkan dan menstandarkan data. Dengan melakukan preprocessing, data menjadi lebih akurat dan siap untuk dianalisis lebih lanjut.

Dengan adanya proses ini, kualitas data meningkat secara signifikan dan dapat digunakan untuk analisis mendalam dalam proses Knowledge Discovery in Databases (KDD).