# Data Cleaning dan Data Duplikat Pada Python

Data cleaning adalah suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Caranya adalah dengan mendeteksi adanya error atau corrupt pada data, kemudian memperbaiki atau menghapus data jika memang diperlukan.

Proses data cleaning yang sering dilakukan adalah sebagai berikut

## Handling Missing Value :

A screenshot of a computer

AI-generated content may be incorrect.

Bisa berupa Lupa input , sistem error.

**Mengetahui Keberadaan Missing Value**

A screenshot of a computer

AI-generated content may be incorrect. 

A screenshot of a computer

AI-generated content may be incorrect.

1. Menangani missing value
   1. Delegasi (hapus)
      1. Metode ini tidak disarankan terutama untuk tipe MNAR
      2. Kekurangan dari metode ini adalah adanya kemungkinan data yang penting ikut terhapus

A close up of a word

AI-generated content may be incorrect.

* 1. Imputasi (diisi dengan suatu nilai)
     1. Imputasi (imputation) adalah proses mengisi missing values dengan suatu nilai
     2. Pada pandas, imputasi dapat menggunakan fillna()



* + 1. dapat diganti dengan nilai skalar atau hasil perhitungan seperti mean, min, max, atau modus
    2. Selain itu, missing values dapat diisi dengan nilai valid sebelumnya atau sesudahnya (backfill vs frontfill)

## Mengatasi Missing Value Versi lengkap

Untuk mengatasi hilangnya beberapa data yang telah diperoleh hal ini ada beberapa cara yang bisa dilakukan, diantaranya adalah :

* Leave as it is (dibiarkan)
* Filling the missing value (imputasi) with statistics value or other valid value
* Drop them (hapus row yang mengandung missing value)

A diagram of a patient treatment

AI-generated content may be incorrect.

* 1. Leave as it is

Mengatasi missing data dengan cara **meninggalkannya begitu saja** (leave as it is) atau tidak melakukan tindakan apa pun terhadap nilai yang hilang disebut dengan ignoring missing data atau leave as it is (biarkan saja). Ini **berarti Anda membiarkan missing data tetap ada dalam dataset** tanpa melakukan imputasi atau penghapusan.

Biasanya kolom yang tidak penting

**Ketentuan**

Beberapa alasan umum mengapa memilih untuk membiarkan missing data:

* Missing data termuat pada kolom yang **tidak digunakan dalam analisa data**, contoh : Missing data pada Alamat rumah, hal ini dikarenakan alamat sering tidak digunakan dalam proses analisa sehingga jika kolom tersebut tidak terisi maka lebih baik untuk membiarkan saja

## Imputation

Mengisi data yang hilang atau null tadi menjadi suatu data ada

Proses imputasi merujuk pada metode atau teknik untuk mengisi nilai yang hilang dalam data. Ketika kita bekerja dengan data yang tidak lengkap, di mana beberapa nilai tidak tersedia atau hilang, imputasi digunakan untuk menggantikan nilai yang hilang dengan estimasi yang masuk akal berdasarkan informasi yang ada dalam data yang lengkap.

**Ketentuan**

Ada beberapa alasan mengapa proses imputasi penting:

* Mencegah kehilangan informasi: Ketika kita memiliki data yang hilang, menghapus baris atau kolom yang mengandung nilai yang hilang dapat mengakibatkan kehilangan informasi yang berharga. Dengan mengimputasi nilai yang hilang, kita dapat mempertahankan sebagian besar data yang tersedia dan memaksimalkan penggunaannya.
* Mempertahankan ukuran sampel yang cukup besar: Dalam analisis statistik, ukuran sampel yang besar sering diinginkan untuk mendapatkan hasil yang lebih dapat diandalkan dan signifikan. Dengan mengisi nilai yang hilang, kita dapat mempertahankan ukuran sampel yang cukup besar dan mengurangi bias dalam analisis.
* Meminimalkan distorsi hasil analisis: Jika kita memiliki data yang hilang secara acak, menghapus baris atau kolom dengan nilai yang hilang dapat menghasilkan bias dalam analisis. Dengan mengimputasi nilai yang hilang, kita dapat meminimalkan distorsi hasil analisis dan menghasilkan estimasi yang lebih akurat.
* Mempertahankan hubungan antar variabel: Jika kita memiliki data yang hilang dalam beberapa variabel yang saling terkait, menghapus baris atau kolom dengan nilai yang hilang dapat mengganggu hubungan antar variabel tersebut. Dengan mengimputasi nilai yang hilang, kita dapat mempertahankan hubungan antar variabel yang penting dalam analisis.
* Proses imputasi dapat dilakukan menggunakan berbagai metode, seperti imputasi sederhana (seperti mengisi nilai yang hilang dengan rata-rata atau median), metode regresi, metode MICE, atau metode lainnya yang sesuai dengan karakteristik data dan tujuan analisis.
* Memungkinkan penggunaan algoritma yang mengharuskan data lengkap: Beberapa algoritma atau metode analisis statistik memerlukan data yang lengkap untuk memberikan hasil yang valid. Dengan mengimputasi nilai yang hilang, kita dapat memungkinkan penggunaan algoritma ini dan memperoleh hasil yang akurat.

Proses imputasi dapat dilakukan menggunakan berbagai metode, seperti imputasi sederhana (seperti mengisi nilai yang hilang dengan rata-rata atau median), metode regresi, metode MICE, atau metode lainnya yang sesuai dengan karakteristik data dan tujuan analisis.

**Backfill vs Frontfill**

A screenshot of a computer

AI-generated content may be incorrect.

* + 1. Backfill

Backfill mengisi missing value dengan nilai valid yang ada setelah baris itu.

* + 1. Frontfill

Frontfill mengisi missing value dengan nilai valid yang ada sebelum baris itu.

teknik ini tepat untuk kolom numerik atau kategorikal yang tidak wajib unik. Jangan gunakan teknik ini untuk kolom seperti ID unik (sale\_id, user\_id, dll).

* 1. **Drop Missing Value**

Menghapus data dengan nilai yang hilang adalah salah satu cara untuk mengatasi missing data, tetapi keputusan ini harus dibuat dengan hati-hati dan dipertimbangkan. Namun, penting untuk mempertimbangkan konsekuensi penghapusan data, terutama jika missing data tersebut tidak bersifat acak atau jika variabel yang dihapus memiliki signifikansi dalam analisis. Seperti kehilangan informasi, mengubah total sampel, mengubah nilai statistik secara signifikan, dll.

Beberapa alasan umum mengapa memilih untuk menghapus missing data:

**Ketentuan**

* **Pentingnya Data yang Hilang**: Jika data yang hilang tidak kritis untuk tujuan analisis atau tidak memiliki dampak signifikan pada kesimpulan, beberapa orang mungkin memilih untuk menghapusnya untuk menyederhanakan analisis.
* **Ketidakberdayaan untuk Diperbaiki**: Terkadang, missing data mungkin terjadi pada variabel atau kolom yang sulit atau tidak mungkin diimputasi dengan benar. Dalam situasi seperti itu, menghapus data yang hilang mungkin menjadi alternatif yang lebih baik daripada imputasi yang tidak dapat diandalkan.
* **Ukuran Sampel yang Cukup Besar**: Jika jumlah total data sangat besar dan jumlah missing data relatif kecil, menghapusnya mungkin tidak signifikan secara statistik.

**Hapus baris yang missing value**

**df\_cleaned = df.dropna(subset=['data\_id'])**

**Hapus Kolom yang missing value**

## Data Duplikat

Redudansi data adalah duplikasi atau penyimpanan data yang sama secara berulang dalam satu atau lebih tabel, sehingga data yang sama di simpan di dalam lebih dari 1 lokasi.

Data bisa terjadi duplikasi karena kesalahan manusia (Human Error) atau bisa jadi karena kesalahan sistem. Untuk mengatasinya kadang perlu diperiksa terlebih dahulu baru akan dilakukan tindakan seperti membiarkan saja atau perlu di hapus

* **Untuk memeriksa banyak nilai duplikat pada kolom tertentu gunakan sintaks (Subset ini nama kolom yang ingin dicari)**



* **Hapus row yang duplikat** pada bagian sale\_id dan keep bagian terakhir

Gunakan sintaks berikut untuk menghapus data duplikat



## Mengubah tipe data di setiap kolom

1. **Mengubah data Rp**

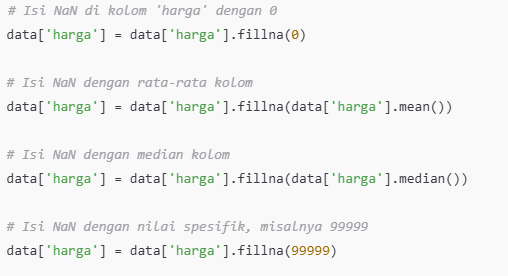
**A computer code with text

AI-generated content may be incorrect.**

Maksudnya df[‘harga’] -> akan mengubah tipe data kolom harga nah

df[’harga’] yang keduanya -> ini untuk menjalanksn fungsinya

## Memasukkan data yang missing dengan suatu nilai



## Rename kolom

Berikut strukturnya

