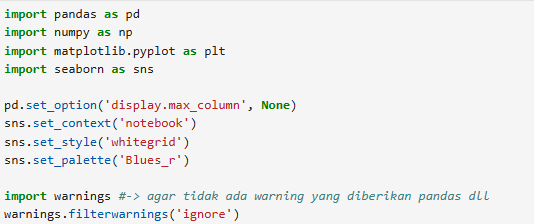
# Data Quality

1. Apakah ada features yang data type-nya tidak sesuai?
2. Adakah duplicate, missing values atau outliers?
3. Apakah terdapat distribusi yang tidak masuk akal?

Bagaimana mengatasi hal-hal tersebut?



Mengubah Tipe data. Hapus Spasi, dan Hapus Tanda

1. Menghapus tanda $ pada kolom

A screenshot of a computer program

AI-generated content may be incorrect.

1. Ubah tipe data

A white rectangular sign with red and blue text

AI-generated content may be incorrect.

1. Menghilangkan spasi ganda



Missing Value

## Mencari

# missing value per kolom

**df.isnull().sum().sort\_values(ascending=False)**

**Jika ingin mengisi missing values ini kita dapat mengisinya dengan mean atau median**

### **Melihan distribusi dari suatu kolom dengan visualisasi**

****

**Digunakan untuk menampilkan distribusi data (dalam hal ini kolom Income) menggunakan Seaborn — library visualisasi berbasis matplotlib.**

1. **sns.distplot(...)**

 Fungsi ini digunakan untuk **membuat visualisasi distribusi frekuensi** dari data.

 Sudah **deprecated (tidak disarankan lagi)** di Seaborn versi 0.11 ke atas.

 Fungsinya digantikan oleh sns.histplot() dan sns.kdeplot().

1. **df['Income']**

* Ini adalah kolom yang akan divisualisasikan.
* Berisi data numerik (contohnya: pendapatan pelanggan).

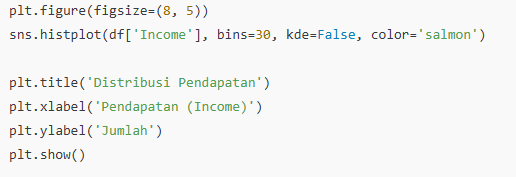
1. **kde=False**

* Artinya: **tidak menampilkan kurva KDE (Kernel Density Estimation)**.
* Jika kde=True, grafik akan menampilkan **kurva halus** di atas histogram untuk menunjukkan **perkiraan distribusi** data.

1. **hist=True**

* Artinya: tampilkan **histogram**, yaitu grafik batang frekuensi data.
* Ini akan menampilkan **berapa banyak data** yang berada dalam interval tertentu (bin).

1. **Melihan distribusi dari suatu kolom dengan visualisasi**



* 1. **plt.figure(figsize=(8, 5))**
* plt.figure() digunakan untuk membuat figure (kanvas gambar) baru.
* figsize=(8, 5) artinya ukuran kanvas adalah 8 inci (lebar) × 5 inci (tinggi).
  + Ukuran ini akan memengaruhi skala tampilan grafik.\
  1. **sns.histplot(...)**

Fungsi Seaborn untuk membuat histogram, yaitu grafik batang distribusi frekuensi.

* 1. df['Income']
* Kolom Income dari DataFrame df yang ingin divisualisasikan.
* Harus bertipe numerik (int atau float).
  1. **bins=30**
     + Jumlah **interval (kelompok)** nilai pada sumbu X dibagi jadi 30.
     + Misalnya jika income berkisar 0–300, maka tiap bin mencakup 10 unit.
  2. **kde=False**
* Tidak menampilkan **kurva KDE (Kernel Density Estimation)**.
* Fokus hanya pada histogram (batang frekuensi).

Outlier



* Ini memilih kolom yang bertipe float dan int
* Dan filter kolom yang diperlukan saja dengan menghapaus(drop) kolom yang tidak diperlukan

1. Visulaisasi oulier dengan boxplot



Penjelasasan:

* subplots=True → setiap kolom diplot dalam subplot terpisah
* layout=(4, 4) → tampilkan subplot dalam 4 baris × 4 kolom
* kind='box' → jenis plotnya adalah boxplot (untuk melihat **outlier**)
* figsize=(12, 14) → ukuran keseluruhan figure

A graph with numbers and a line

AI-generated content may be incorrect.

**Intepretasinya :**

Boxplot akan memperlihatkan:

* Garis tengah (median)
* Kotak (kuartil 1 ke kuartil 3)
* Garis batas (whiskers)
* Titik-titik di luar whisker → itu adalah outlier

Titik-titik hitam yang berada **di luar garis whisker** pada boxplot—dalam hal ini sekitar **tahun 1900**—itulah yang disebut sebagai **outlier**.

Lalu bisa dihapus data yang tidak masuk akalnya

Apakah terdapat distribusi yang tidak masuk akal?

## Penjelasan

## Menggunakan histogram dimana hal ini berfungsi agar bisa melihat distribusi nilai seluruh kolom yang datanya numeric

## A screenshot of a computer program AI-generated content may be incorrect. Tentu! Gambar yang kamu kirim berisi kode Python (menggunakan matplotlib dan pandas) untuk membuat histogram banyak kolom sekaligus.

1. fig = plt.figure(figsize=(15, 10))

* Membuat objek figure (kanvas) untuk semua histogram.
* figsize=(15, 10) artinya ukuran lebar 15 inci, tinggi 10 inci.

1. ax = plt.gca()

* gca() = **get current axes**, mengambil axis aktif saat ini.
* Namun ini **tidak diperlukan** untuk df.hist() (jadi ini bisa dihapus jika tidak dipakai eksplisit).

1. df.hist(bins=50, ax=ax, layout=(4, 4), column=numeric)

Ini bagian utama yang:

* Membuat **histogram** untuk banyak kolom numerik.
* **Parameter:**
  + bins=50 → Membagi sumbu X menjadi 50 interval untuk setiap histogram.
  + ax=ax → Melewatkan axis (tapi kurang tepat di sini, penjelasan di bawah).
  + layout=(4, 4) → Membuat grid 4 baris × 4 kolom = total 16 plot.
  + column=numeric → Menentukan kolom mana yang akan diplot (seharusnya numeric adalah list kolom numerik)

1. plt.tight\_layout()

* Memastikan **jarak antar subplot tidak saling tumpang tindih**.
* Otomatis menyesuaikan padding.

Fungsi kode ini adalah untuk **menampilkan histogram dari semua kolom numerik** dalam bentuk grid (4x4) untuk melihat **distribusi data tiap kolom**.

.