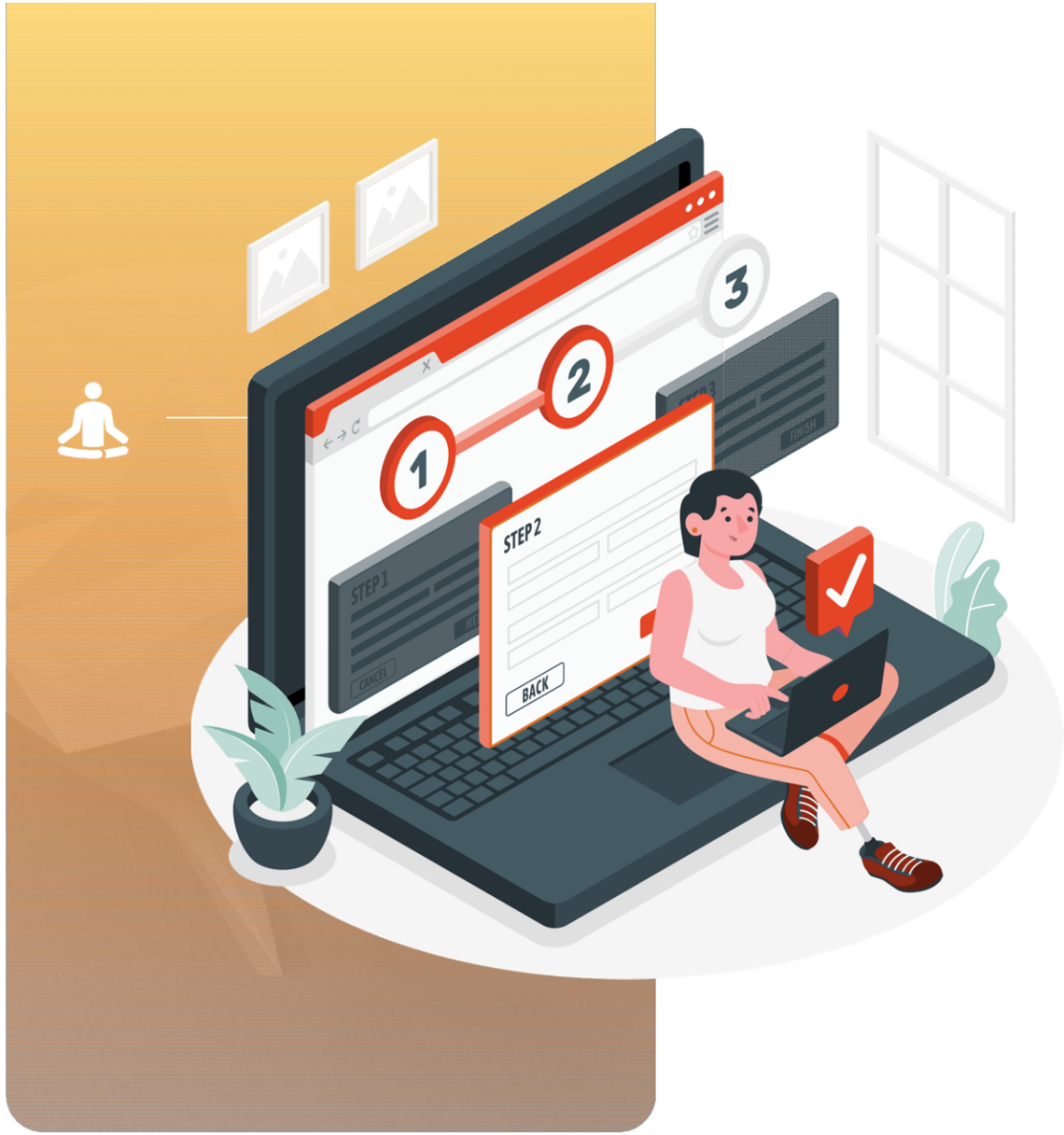
DATA AND WEBSITE ANALYTICS

ARI PURNOMO, S.KOM. M.MT

PERTEMUAN 5

Seputar Data and Website Analytics



DATA CLEANSING ATAU PEMBERSIHAN DATA

Next Step Slide Presentation

PENGANTAR

Di era digital, data menjadi aset yang sangat berharga. Namun, data yang diperoleh dari berbagai sumber sering kali tidak bersih atau tidak terstruktur dengan baik. Data seperti ini sulit digunakan untuk analisis maupun pengambilan keputusan.

Oleh karena itu, proses *data cleansing* menjadi sangat penting dalam siklus pengolahan data.

APA ITU DATA CELANSING

Data cleansing adalah proses membersihkan data dari kesalahan atau ketidaksesuaian yang dapat memengaruhi analisis.

Kesalahan ini bisa berupa data duplikat, nilai kosong, data yang tidak sesuai format, atau nilai ekstrem yang tidak logis. Proses ini bertujuan meningkatkan kualitas dan keakuratan data sebelum digunakan lebih lanjut.

Next Step Slide Presentation

JENIS MASALAH PADA DATA

- **Missing Values**: Data yang seharusnya diisi tetapi kosong.
- **Inconsistent Entries**: Penulisan yang tidak seragam, seperti "Laki-laki" dan "laki2".
- **Duplicate Data**: Baris data yang sama muncul lebih dari sekali.
- **Wrong Format**: Misalnya tanggal ditulis sebagai "12-30-2024" padahal seharusnya "30-12-2024".
- **Outlier (Data tidak Logis)** : Nilai ekstrem yang tidak sesuai, seperti usia 250 tahun

TAHAP DATA CLEANSING

- **Deteksi Error**: Mengidentifikasi data yang bermasalah.
- **Analisis Sumber Masalah**: Menelusuri asal mula data salah, misalnya dari input manual.
- **Perbaikan Data**: Mengisi, menghapus, atau mengganti data dengan benar.
- **Validasi Hasil**: Memastikan bahwa data yang dibersihkan sudah sesuai standar.
- **Dokumentasi**: Mencatat proses yang dilakukan untuk transparansi dan replikasi

TEKNIK DETEKSI MASALAH DATA

- **Visualisasi**: Membantu melihat outlier dan distribusi data.
- **Statistik Deskriptif**: Menggunakan mean, median, dan range untuk mengenali data aneh.
- **Tools**: Gunakan Excel (conditional formatting), Python (pandas profiling), atau OpenRefine.

PENANGANAN MISSING DATA

- **Hapus** jika jumlahnya sedikit dan tidak signifikan.
- **Isi dengan rata-rata atau median,** jika data numerik.
- **Gunakan prediksi atau interpolasi** untuk nilai penting.
- **Biarkan kosong** jika tidak berdampak besar, tergantung konteks analisis

Mengapa Menggunakan Rata-rata atau Median?

1. **Agar Data Lengkap dan Bisa Diproses**
   - Kalau nilai hilang dibiarkan kosong, banyak tools analisis (seperti Excel, Python, SPSS) **tidak bisa menghitung** dengan benar.
   - Maka, nilai kosong diisi agar dataset bisa dipakai
2. **Rata-rata:** Cocok jika **data tidak banyak outlier**
   - Jadi kita isi data kosong dengan **12.5**.
   - Ini menjaga distribusi data tetap alami

```
Data: 10, 12, 13, (kosong), 15
Rata-rata = (10+12+13+15)/4 = 12.5
```

3. **Median**: Lebih aman jika ada **outlier**
   - Rata-rata = 33.75 (karena 100 terlalu besar)
   - Tapi **median tetap 12.5**, jadi lebih **tidak terpengaruh outlier.**
   - Maka, **median lebih cocok jika data miring atau banyak outlier.**

```
Data: 10, 12, 13, (kosong), 100
Median dari 10, 12, 13, 100 = (12+13)/2 = 12.5
```

Mengapa Menggunakan prediksi atau interpolasi untuk nilai penting

1. Prediksi (Prediction / Estimasi)

- Kita **gunakan model statistik atau machine learning** untuk **memprediksi nilai yang hilang** berdasarkan pola dari data lainnya

  Misal : Jika kolom yang hilang adalah “penghasilan”, bisa diprediksi dari kolom umur, pendidikan, pekerjaan, dll.

2. Interpolasi

- Digunakan untuk **data berurutan** (misal: data waktu/suhu, penjualan per hari).
- **Isi nilai kosong dengan memperkirakan berdasarkan titik data sebelum dan sesudahnya**

```
Hari ke-1: 80
Hari ke-2: (kosong)
Hari ke-3: 100
→ Interpolasi: isi Hari ke-2 = 90
```

MENANGANI DUPLIKASI DATA

- **Gunakan Excel** (remove duplicates) atau Python (drop_duplicates()).
- **Cek manual** untuk data penting.
- Tetapkan kolom identifikasi unik **(ID)** untuk mencegah duplikasi di masa depan.

NORMALISASI DATA

- Huruf besar/kecil dijadikan konsisten.
- Format tanggal distandarisasi.
- Nama kategori diseragamkan ("Perempuan", bukan "wanita", "cewek", dll).

Normalisasi penting untuk menghindari duplikasi yang tidak perlu dan meningkatkan kualitas klasifikasi.

VALIDASI DATA

Data valid artinya data harus sesuai dengan aturan atau standar yang ditentukan.

- Contoh aturan: NIM harus 10 digit, nilai tidak boleh >100.
- Gunakan Regex, data validation di Excel, atau pengkodean Python untuk memeriksa validitas. Validasi membantu menjaga integritas dan mencegah error sebelum data dianalisis.

ARI PURNOMO, S.KOM. M.MT

SEKIAN TERIMAKASIH