3_praktikum

October 17, 2025

[1]:
```
import pandas as pd
df = pd.read_csv("data_penjualan.csv")
print(df.head())
```

```
   nama  umur gender  qty  price
0  Sari  24.0      M  NaN   9000
1  Hadi  32.0      M  NaN   8000
2  Sari  24.0      F  NaN  15000
3  Xena  21.0      F  5.0  15000
4  Xena  20.0      M  5.0   8000
```

[2]:
```
# Hapus kolom yang tidak diperlukan
df = df.drop(columns=["Unnamed: 0"], errors="ignore")

# Hapus data duplikat
df = df.drop_duplicates()

# Rename kolom
df = df.rename(columns={"qty": "jumlah", "price": "harga"})

print("--- Data setelah rename & drop duplicates ---")
print(df.head())
```

```
--- Data setelah rename & drop duplicates ---
   nama  umur gender  jumlah  harga
0  Sari  24.0      M     NaN   9000
1  Hadi  32.0      M     NaN   8000
2  Sari  24.0      F     NaN  15000
3  Xena  21.0      F     5.0  15000
4  Xena  20.0      M     5.0   8000
```

[3]:
```
#Catat berapa banyak nilai kosong SEBELUM
print("--- Nilai kosong SEBELUM perbaikan ---")
print(df.isnull().sum())

#Hitung rata-rata kolom umur
rata_rata_umur = df["umur"].mean()
print(f"\nRata-rata umur: {rata_rata_umur}")
```

```
#Isi nilai kosong di kolom umur dengan rata-rata
df["umur"] = df["umur"].fillna(rata_rata_umur)

#Isi nilai kosong di kolom jumlah dengan 0
df["jumlah"] = df["jumlah"].fillna(0)

#Isi nilai kosong di kolom harga dengan median
print("\n--- Nilai kosong SESUDAH perbaikan ---")
print(df.isnull().sum())
```

```
--- Nilai kosong SEBELUM perbaikan ---
nama       0
umur      70
gender     0
jumlah    64
harga      0
dtype: int64

Rata-rata umur: 26.974025974025974

--- Nilai kosong SESUDAH perbaikan ---
nama      0
umur      0
gender    0
jumlah    0
harga     0
dtype: int64
```

[4]:
```
# Replace nilai kategorikal
df["gender"] = df["gender"].replace({"M": "Male", "F": "Female"})

# Normalisasi kolom harga
df["harga_norm"] = (df["harga"] - df["harga"].min()) / \
                   (df["harga"].max() - df["harga"].min())

# Feature engineering (total_price)
df["total_price"] = df["jumlah"] * df["harga"]

print("--- Data setelah Transformasi ---")
print(df.head())
```

```
--- Data setelah Transformasi ---
   nama  umur  gender  jumlah  harga  harga_norm  total_price
0  Sari  24.0    Male     0.0   9000         0.4          0.0
1  Hadi  32.0    Male     0.0   8000         0.3          0.0
2  Sari  24.0  Female     0.0  15000         1.0          0.0
3  Xena  21.0  Female     5.0  15000         1.0      75000.0
4  Xena  20.0    Male     5.0   8000         0.3      40000.0
```

[5]:
```
print("\n--- Latihan 1: interpolate() ---")
# Kita baca ulang data mentah untuk demo
df_latihan_1 = pd.read_csv("data_penjualan.csv")

print("Data 'umur' SEBELUM interpolate (baris 0-5):")
print(df_latihan_1[['nama', 'umur']].head())

# Gunakan interpolate()
df_latihan_1['umur'] = df_latihan_1['umur'].interpolate()

print("\nData 'umur' SESUDAH interpolate (baris 0-5):")
print(df_latihan_1[['nama', 'umur']].head())
```

```
--- Latihan 1: interpolate() ---
Data 'umur' SEBELUM interpolate (baris 0-5):
   nama  umur
0  Sari  24.0
1  Hadi  32.0
2  Sari  24.0
3  Xena  21.0
4  Xena  20.0

Data 'umur' SESUDAH interpolate (baris 0-5):
   nama  umur
0  Sari  24.0
1  Hadi  32.0
2  Sari  24.0
3  Xena  21.0
4  Xena  20.0
```

[6]:
```
print("\n--- Latihan 2: describe() Sebelum vs Sesudah ---")

# Sebelum cleaning
df_awal = pd.read_csv("data_penjualan.csv")

print("\n--- Ringkasan Statistik SEBELUM Cleaning (df_awal) ---")
print(df_awal.describe())

# Setelah cleaning
print("\n--- Ringkasan Statistik SESUDAH Cleaning (df) ---")
print(df.describe())

print("\nPerbedaan utama:")
print("- 'count' (jumlah data) di df (sesudah) lebih tinggi untuk 'umur' dan␣
 ↪'jumlah' karena NaN sudah diisi.")
```

```
print("- 'count' total baris di df (159) lebih rendah dari df_awal (163) karena␣
  ↪4 baris duplikat dihapus.")
print("- 'mean', 'std', 'min' 'umur' dan 'jumlah' (qty) berubah karena nilai␣
  ↪NaN sudah diisi (diisi mean & 0).")
print("- 'df' (sesudah) memiliki kolom baru: 'harga_norm' dan 'total_price'.")
```

```
--- Latihan 2: describe() Sebelum vs Sesudah ---

--- Ringkasan Statistik SEBELUM Cleaning (df_awal) ---
            umur        qty         price
count  77.000000  85.000000    150.000000
mean   26.974026   2.988235   9353.333333
std     4.463224   1.409949   2912.911933
min    20.000000   1.000000   5000.000000
25%    24.000000   2.000000   7000.000000
50%    27.000000   3.000000   9000.000000
75%    30.000000   4.000000  12000.000000
max    35.000000   5.000000  15000.000000

--- Ringkasan Statistik SESUDAH Cleaning (df) ---
             umur      jumlah         harga  harga_norm   total_price
count  147.000000  147.000000    147.000000  147.000000    147.000000
mean    26.974026    1.673469   9367.346939    0.436735  15877.551020
std      3.220172    1.817583   2935.567511    0.293557  18944.004712
min     20.000000    0.000000   5000.000000    0.000000      0.000000
25%     26.974026    0.000000   7000.000000    0.200000      0.000000
50%     26.974026    1.000000   9000.000000    0.400000   9000.000000
75%     27.000000    3.000000  12000.000000    0.700000  29000.000000
max     35.000000    5.000000  15000.000000    1.000000  75000.000000

Perbedaan utama:
- 'count' (jumlah data) di df (sesudah) lebih tinggi untuk 'umur' dan 'jumlah'
karena NaN sudah diisi.
- 'count' total baris di df (159) lebih rendah dari df_awal (163) karena 4 baris
duplikat dihapus.
- 'mean', 'std', 'min' 'umur' dan 'jumlah' (qty) berubah karena nilai NaN sudah
diisi (diisi mean & 0).
- 'df' (sesudah) memiliki kolom baru: 'harga_norm' dan 'total_price'.
```