Panduan Praktis Data Cleaning & Missing Value Handling

1. Cek Tipe Data Tiap Kolom

```
df.dtypes
df.info()
```

- Pastikan tahu mana numeric, categorical, date, object. - Tipe data saja tidak menjamin formatnya benar.

2. Lihat Isi Unik / Sample Tiap Kolom

```
df['Property'].unique()
df['Locality'].value_counts()
df[['carpet_area','Estimated Value']].head(10)
```

- Tujuan: menemukan nilai aneh atau placeholder missing ('?', 'NA', 0, '').

3. Standardisasi Missing Value

```
import numpy as np
df.replace({'?': np.nan, 'NA': np.nan, '': np.nan, 0: np.nan}, inplace=True)
# Bisa juga kolom spesifik:
df['Property'] = df['Property'].replace('?', np.nan)
```

4. Pastikan Tipe Data Sesuai

```
df['carpet_area'] = pd.to_numeric(df['carpet_area'], errors='coerce')
df['Estimated Value'] = pd.to_numeric(df['Estimated Value'], errors='coerce')
df['Sale Date'] = pd.to_datetime(df['Sale Date'], errors='coerce')
```

5. Cek Missing Value & Persentase

```
missing_count = df.isnull().sum()
missing_pct = df.isnull().mean() * 100
```

```
print(missing_count)
print(missing_pct)
```

6. Tentukan Strategi Imputasi

Tipe Kolom	Strategi Imputasi
Numeric	mean / median / prediksi
Categorical	mode / 'Unknown' label
Date	forward/backward fill

7. Imputasi / Isi Missing Value

• Numeric (median lebih aman untuk outlier):

```
df['carpet_area'] = df['carpet_area'].fillna(df['carpet_area'].median())
```

• Categorical (mode atau Unknown):

```
df['Property'] = df['Property'].fillna(df['Property'].mode().iloc[0])
df['Locality'] = df['Locality'].fillna('Unknown')
```

8. Cek Hasil & Validasi

```
print(df.isnull().sum())
print(df.head())
```

- Pastikan semua kolom sudah bersih.

9. (Opsional) Simpan Hasil Bersih

```
df.to_csv("real_estate_clean.csv", index=False)
```

- Supaya bisa load lagi tanpa harus ulangi proses cleaning.