ULTIMATE DATA CLEANING GUIDE + UNIVERSAL TEMPLATE

----

1. HOW TO KNOW WHAT TO CLEAN IN ANY DATASET

STEP 1 — Inspect the dataset

- df.head()
- df.info()
- df.describe(include="all")
- df.isnull().sum()

STEP 2 — Check datatype problems

- Convert dates
- Convert numeric columns
- Fix boolean values

STEP 3 — Handle missing values

- Fill numeric with median
- Fill categorical with mode or 'Unknown'
- Fill booleans with False
- Drop rows with missing keys (like IDs)

STEP 4 — Clean categorical columns

- Strip spaces
- Lowercase
- Check unique values
- Fix typos

STEP 5 — Clean numeric columns

- Remove negative values
- Fix impossible values
- Recalculate totals (price * quantity)

STEP 6 — Remove duplicates

- df.drop_duplicates()

STEP 7 — Understand each column's meaning

STEP 8 — Validate business logic

- Quantity > 0
- Price > 0
- Total == price * quantity

STEP 9 — Rename columns

- Lowercase
- Replace spaces with underscores

----

2. UNIVERSAL CLEANING TEMPLATE (COPY–PASTE READY)

```
import pandas as pd
import numpy as np

df = pd.read_csv("your_dataset.csv")

# 1. Clean column names
df.columns = (
df.columns.str.strip()
.str.lower()
.str.replace(" ", "_")
.str.replace(r"[^a-z0-9_]", "", regex=True)
)

# 2. Inspect
print(df.info())
print(df.isnull().sum())

# 3. Fix dtypes
date_cols = ["date", "transaction_date"]
for col in date_cols:
if col in df.columns:
```

```
df[col] = pd.to_datetime(df[col], errors="coerce")

# Numeric conversion
num_cols = ["price", "quantity", "total"]
for col in num_cols:
if col in df.columns:
df[col] = pd.to_numeric(df[col], errors="coerce")

# 4. Clean categorical columns
cat_cols = df.select_dtypes(include="object").columns
for col in cat_cols:
df[col] = df[col].astype("string").str.strip().str.lower()

# 5. Fill missing values
for col in num_cols:
if col in df.columns:
df[col] = df[col].fillna(df[col].median())

for col in cat_cols:
df[col] = df[col].fillna("unknown")

# Boolean fix
bool_cols = ["discount_applied", "is_returned"]
for col in bool_cols:
if col in df.columns:
df[col] = df[col].astype(str).str.lower().replace({
"true": True, "false": False, "nan": np.nan
}).fillna(False)

# 6. Recalculate totals if needed
if all(c in df.columns for c in ["price", "quantity", "total"]):
df["total"] = df["price"] * df["quantity"]

# 7. Remove duplicates
df.drop_duplicates(inplace=True)
```

```
# 8. Save cleaned data
df.to_csv("cleaned_dataset.csv", index=False)

print("Cleaning complete!")
```