7️⃣ Week 7: Data Cleaning & Preprocessing

This code demonstrates loading raw data, identifying/filling missing values, fixing data types, and renaming columns.

\# Week 7 Assignment – Cleaning and Preparing Basketball Data

```python
import pandas as pd

import numpy as np

from io import StringIO


# --- 1. Create Raw Dataset ---
# Data simulates missing values (None/np.nan) and improper types (floats)
data = {

    "Player": ["Player A", "Player B", "Player C", "Player D", "Player E"],

    "PTS": [25, None, 18, 30, 12],

    "REB": [7, 9, np.nan, 10, 5],

    "AST": [5, 6, 4, None, 3],

    "MIN": [33, 28, 30, 40, np.nan],

    "Position": ["G", "F", "G", "C", "G"]

}
df = pd.DataFrame(data)

print("Raw Dataset:")

display(df)


# --- 2. Check for Missing Values ---

print("\n--- Missing Values Summary ---")

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

```python
# --- 3. Fill Missing Data (Imputation) ---

# Note: Using .loc to avoid the SettingWithCopyWarning

# Fill PTS and REB with Mean

df.loc[:, "PTS"] = df["PTS"].fillna(df["PTS"].mean())

df.loc[:, "REB"] = df["REB"].fillna(df["REB"].mean())


# Fill AST with Median (a good robust choice for assist counts)

df.loc[:, "AST"] = df["AST"].fillna(df["AST"].median())


# Fill MIN with Mean

df.loc[:, "MIN"] = df["MIN"].fillna(df["MIN"].mean())


# --- 4. Fix Data Types ---

# Convert MIN column from float (due to imputation) to integer

df.loc[:, "MIN"] = df["MIN"].astype(int)


# --- 5. Rename Columns ---

df.rename(columns={"PTS": "Points", "REB": "Rebounds", "AST": "Assists"}, inplace=True)


# --- 6. Detect Outliers (Optional) ---

df["Outlier_MIN"] = df["MIN"] > 45


# --- 7. Export Clean Data ---

print("\n--- Cleaned Dataset ---")

display(df)
```

```
df.to_csv("cleaned_week7_assignment.csv", index=False)

print("\n✅ Cleaned dataset saved as 'cleaned_week7_assignment.csv'")
```