PROJECT: MODULE 1

Data Loading & Basic Preprocessing

Harsha Priya Putta

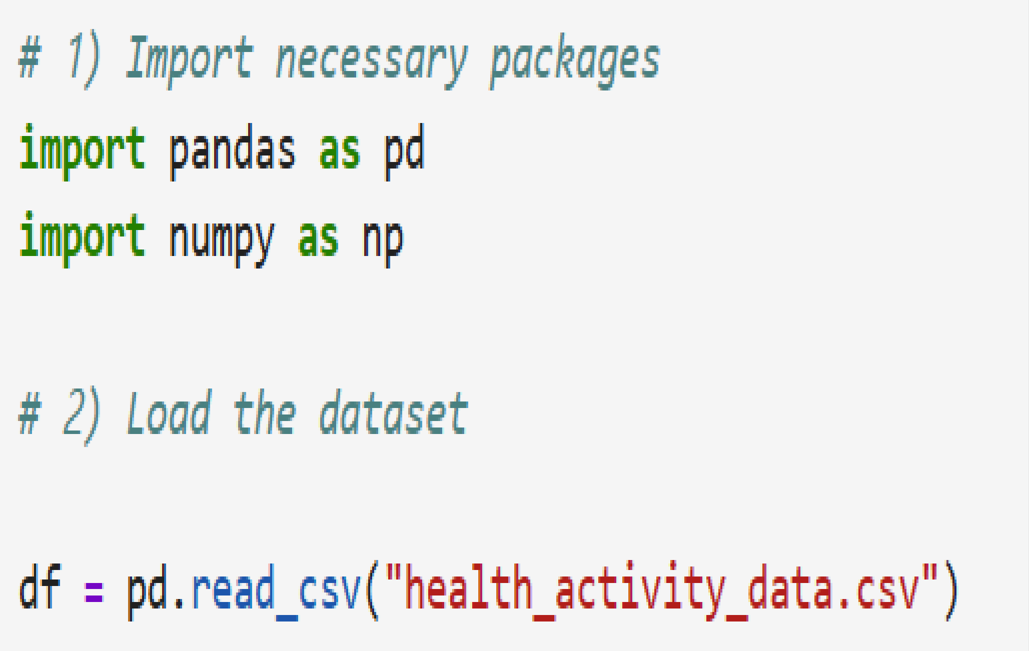
WELLBOT GLOBAL WELLNESS ASSISTANCE CHATBOT

Objective: Load raw health and lifestyle dataset and perform initial cleaning and preprocessing to prepare for analysis and model building.

Outcome: By the end of Module 1, the dataset is **clean, consistent, and ready** for deeper analysis, setting a solid foundation for the WellBot Global Wellness Assistance Chatbot project.

STEP1: IMPORT & LOAD DATASET

- **Libraries Used:**
  `pandas` → Data manipulation
  `numpy` → Numerical computations
- **Dataset:** `health_and_lifestyle_dataset.csv`
  Loaded using `pd.read_csv()`
- **Initial inspection:**
  Shape: Number of rows and columns
  Columns: List of features

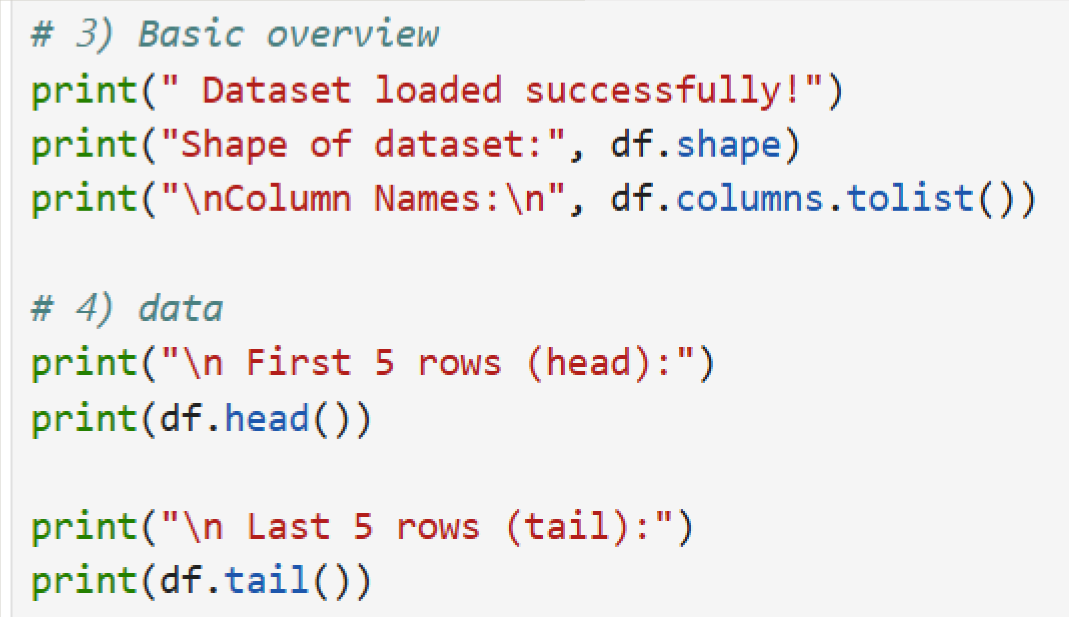


```
# 1) Import necessary packages
import pandas as pd
import numpy as np


# 2) Load the dataset


df = pd.read_csv("health_activity_data.csv")
```

STEP2: BASIC OVERVIEW



```
# 3) Basic overview
print(" Dataset loaded successfully!")
print("Shape of dataset:", df.shape)
print("\nColumn Names:\n", df.columns.tolist())

# 4) data
print("\n First 5 rows (head):")
print(df.head())

print("\n Last 5 rows (tail):")
print(df.tail())
```

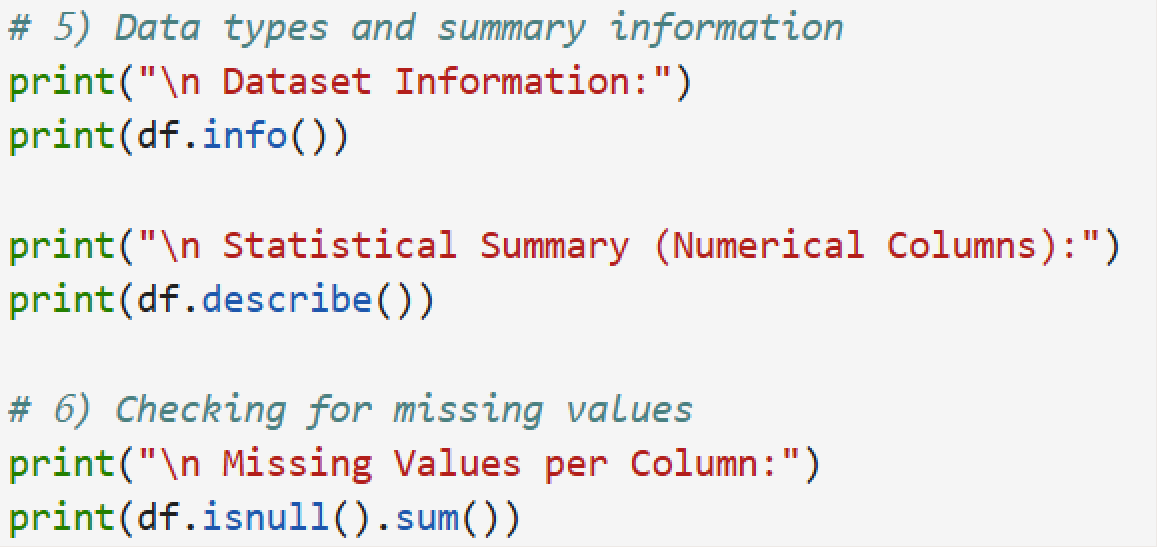
Peek at dataset:

- First 5 rows (head)
- Last 5 rows (tail)

Shape & columns:

- Shape of dataset: (1000, 16)
- Column Names:
- ['ID', 'Age', 'Gender', 'Height_cm', 'Weight_kg', 'BMI', 'Daily_Steps', 'Calories_Intake', 'Hours_of_Sleep', 'Heart_Rate', 'Blood_Pressure', 'Exercise_Hours_per_Week', 'Smoker', 'Alcohol_Consumption_per_Week', 'Diabetic', 'Heart_Disease']

STEP3: SUMMARY INFORMATION & MISSING VALUES



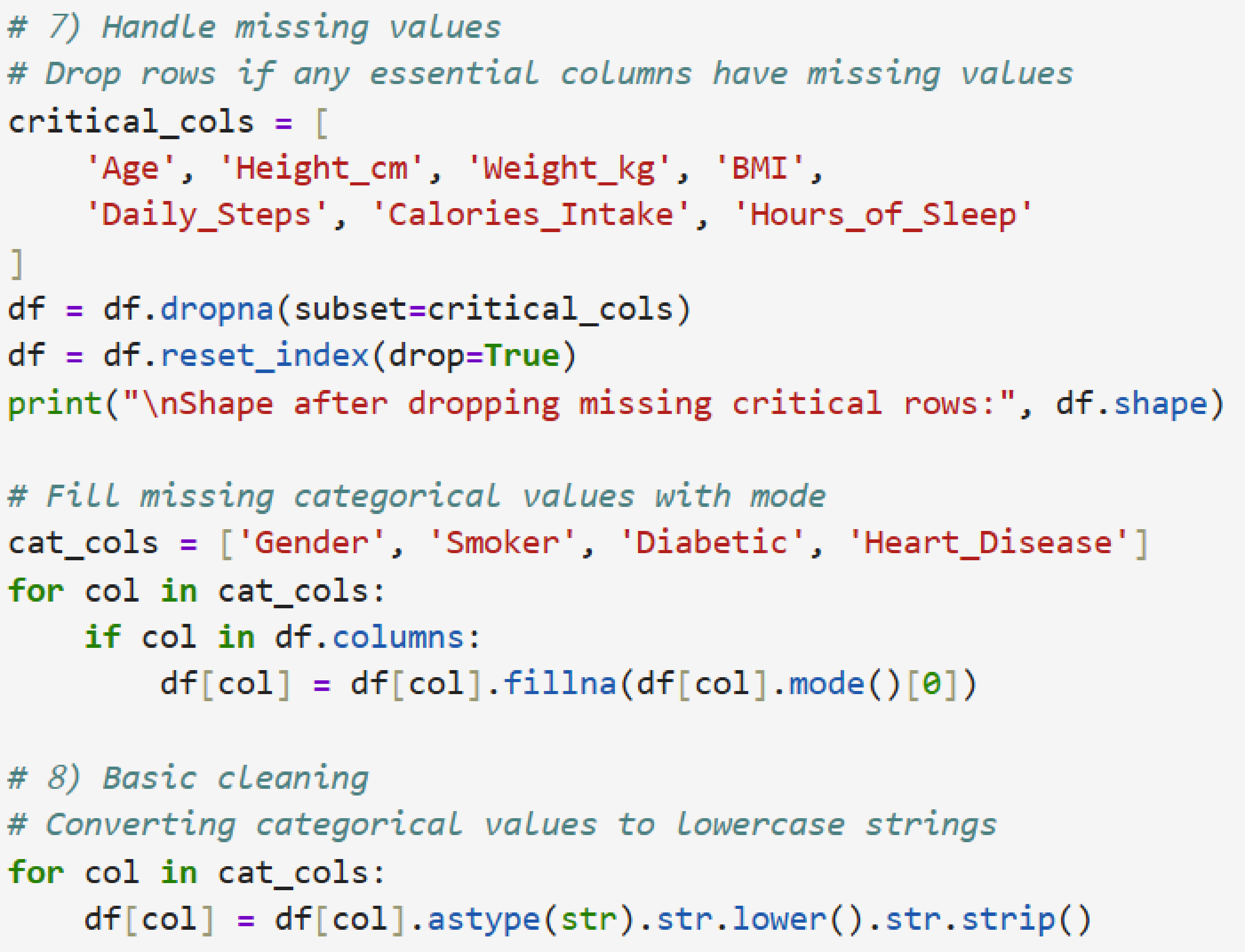
```
# 5) Data types and summary information
print("\n Dataset Information:")
print(df.info())

print("\n Statistical Summary (Numerical Columns):")
print(df.describe())

# 6) Checking for missing values
print("\n Missing Values per Column:")
print(df.isnull().sum())
```

- **Dataset info** (`df.info()`) → data types & non-null counts
- **Statistical summary** (`df.describe()`) → numerical columns overview
- **Purpose:** Understand structure, data types, and spot anomalies
- Checked missing values per column (`df.isnull().sum()`)

STEP4: HANDLING MISSING VALUES



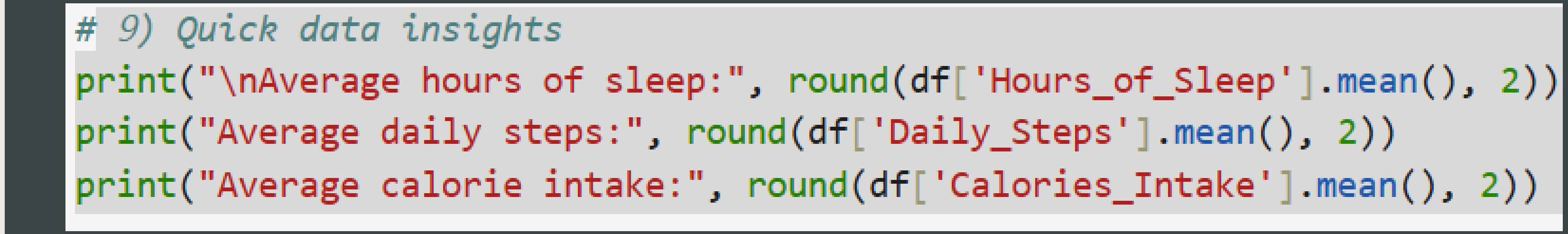
```
# 7) Handle missing values
# Drop rows if any essential columns have missing values
critical_cols = [
    'Age', 'Height_cm', 'Weight_kg', 'BMI',
    'Daily_Steps', 'Calories_Intake', 'Hours_of_Sleep'
]
df = df.dropna(subset=critical_cols)
df = df.reset_index(drop=True)
print("\nShape after dropping missing critical rows:", df.shape)

# Fill missing categorical values with mode
cat_cols = ['Gender', 'Smoker', 'Diabetic', 'Heart_Disease']
for col in cat_cols:
    if col in df.columns:
        df[col] = df[col].fillna(df[col].mode()[0])

# 8) Basic cleaning
# Converting categorical values to lowercase strings
for col in cat_cols:
    df[col] = df[col].astype(str).str.lower().str.strip()
```

- Drop rows with missing values in critical numeric columns:
  Age, Height_cm, Weight_kg, BMI, Daily_Steps, Calories_Intake, Hours_of_Sleep
- Fill missing categorical columns with mode:
  Gender, Smoker, Diabetic, Heart_Disease
  Standardized categorical columns:
- Convert to lowercase
- Remove extra spaces

STEP5: QUICK DATA INSIGHTS



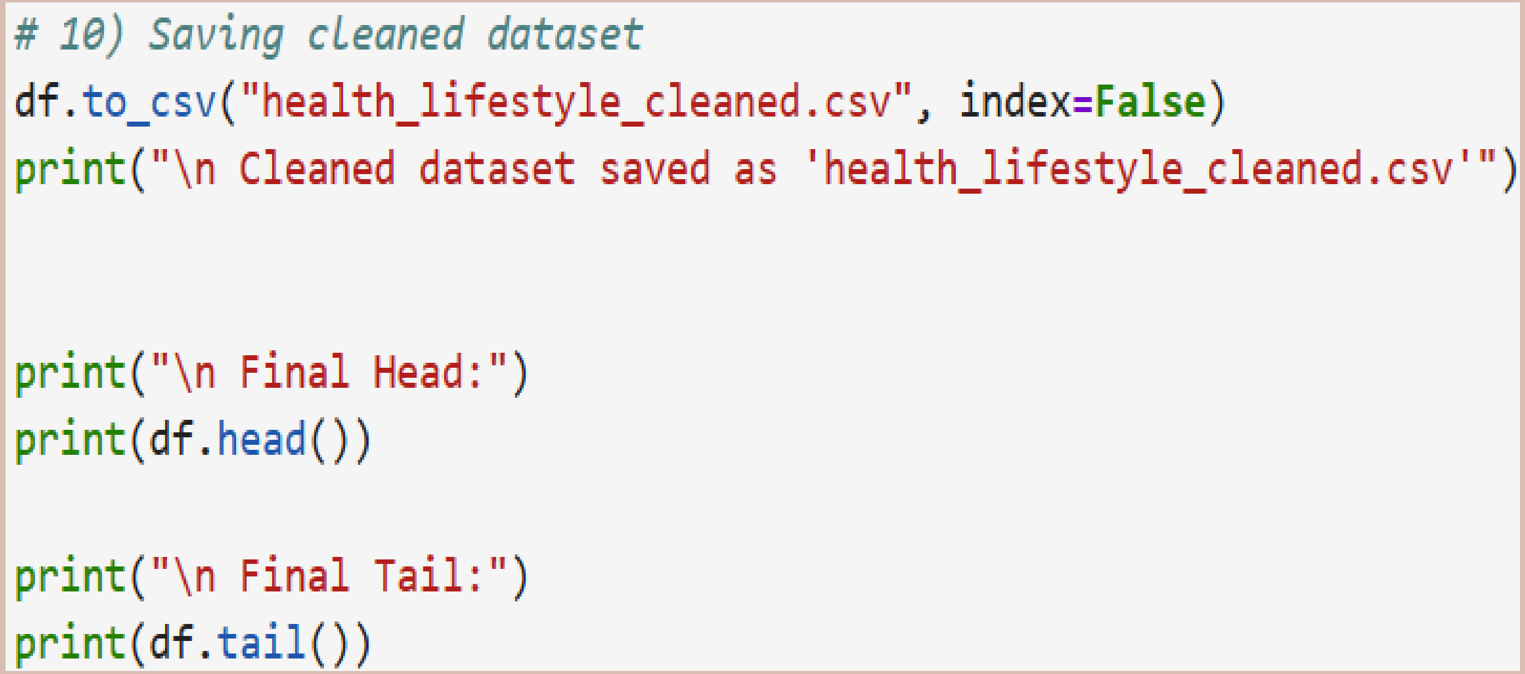
```
# 9) Quick data insights
print("\nAverage hours of sleep:", round(df['Hours_of_Sleep'].mean(), 2))
print("Average daily steps:", round(df['Daily_Steps'].mean(), 2))
print("Average calorie intake:", round(df['Calories_Intake'].mean(), 2))
```

Calculated averages for key columns:

- Hours of Sleep
- Daily Steps
- Calorie Intake

Provides initial understanding of health and lifestyle trends.

STEP7: SAVE THE DATASET



```
# 10) Saving cleaned dataset
df.to_csv("health_lifestyle_cleaned.csv", index=False)
print("\n Cleaned dataset saved as 'health_lifestyle_cleaned.csv'")


print("\n Final Head:")
print(df.head())

print("\n Final Tail:")
print(df.tail())
```

Save Cleaned Dataset :

- Saved cleaned dataset as `health_lifestyle_cleaned.csv`
- Verified by checking first (head) and last (`tail`) rows

SUMMARY

What was done:

What was done:

- Dataset loaded and inspected
- Missing values handled
- Categorical columns standardized
- Quick insights generated

How it helps the project:

- Provides a solid foundation for EDA and machine learning
- Ensures reproducibility and workflow efficiency in WellBot Global Wellness Assistance Chatbot

THANK YOU!!