Descriptive Statistics — *Understanding Your Data*

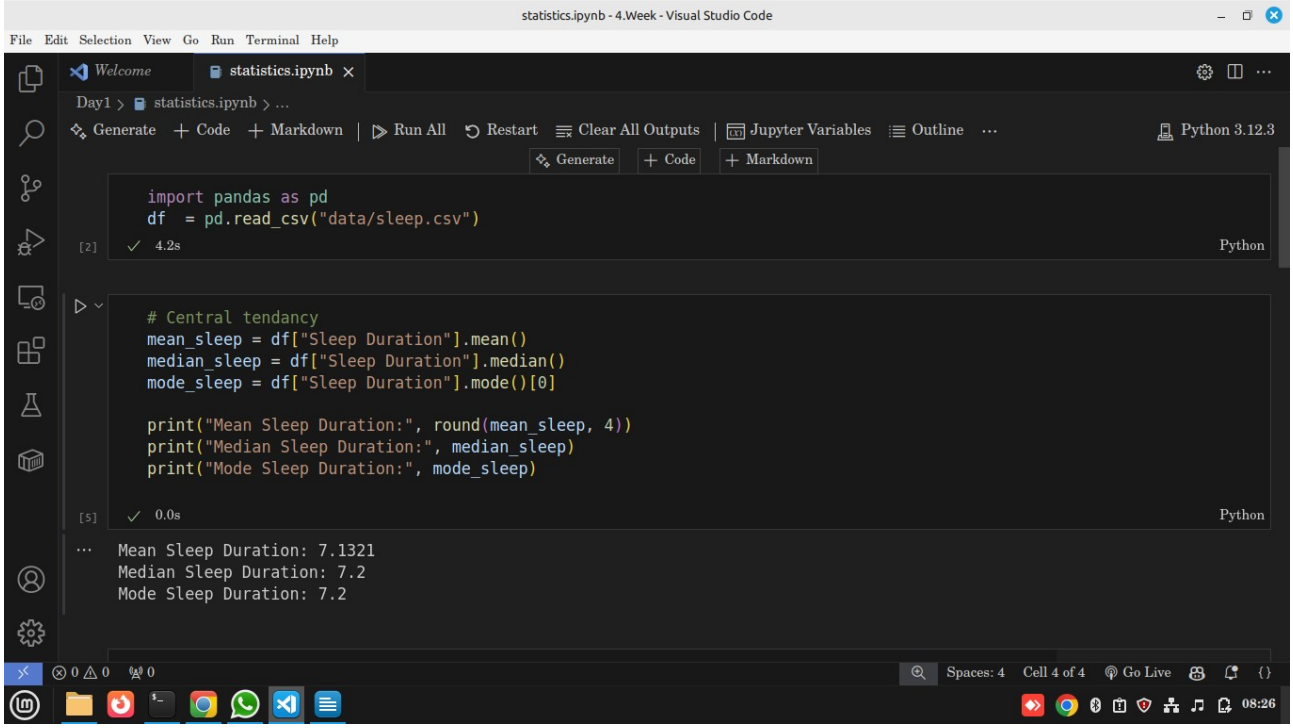
🎯 Objective

To summarize and describe the main features of your dataset using **central tendency** (mean, median, mode) and **dispersion** (variance, standard deviation, range).

We'll use your dataset **`sleep.csv`** to understand how people sleep differently based on lifestyle and other factors.

1️⃣ Central Tendency

These values help us understand *where most of the data lies*.



🔍 Explanation

- **Mean** → The arithmetic average; gives the *typical value*.
- **Median** → The *middle value* when sorted; less affected by outliers.
- **Mode** → The *most frequent value*; shows the most common sleep duration.

💡 Insights

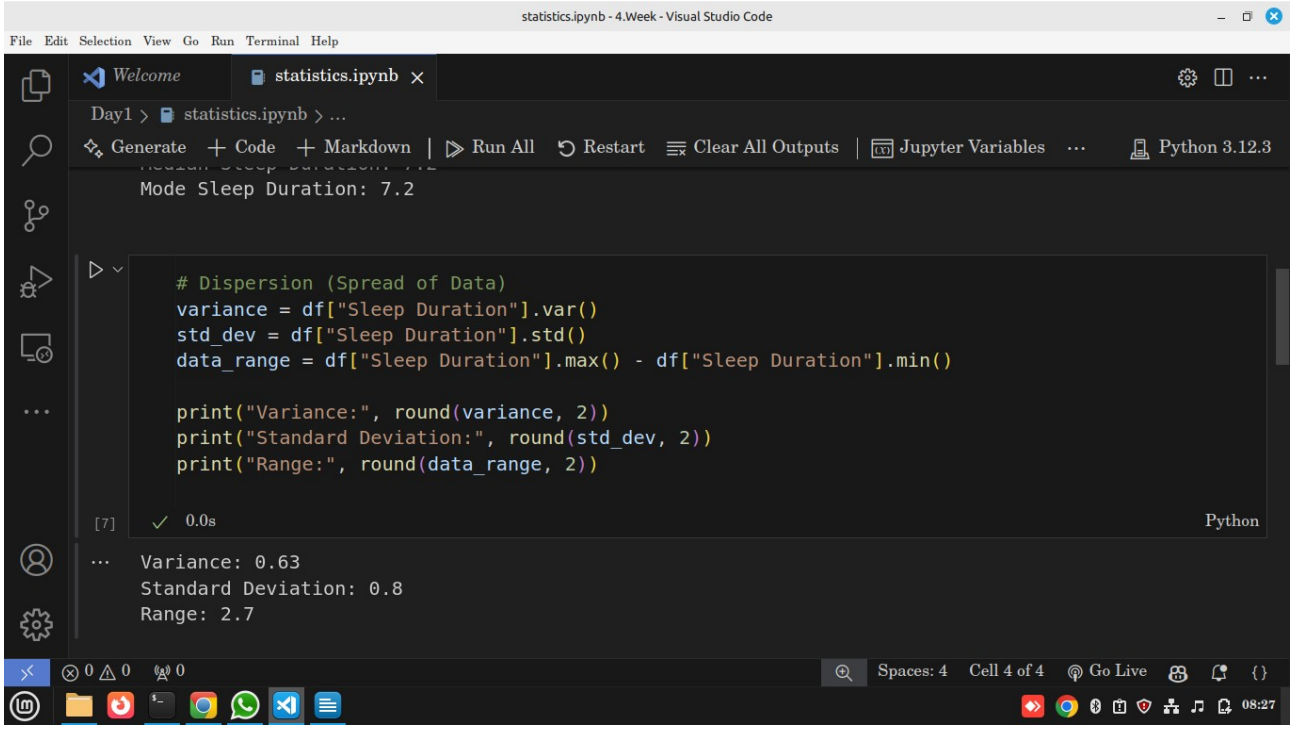
- If your dataset has outliers (e.g., someone sleeping only 2 hours), the **mean** will drop.

- The **median** still stays stable — it's more reliable when there are extreme values.
- The **mode** shows the most common habit — e.g., if mode = 7, most people sleep 7 hours.

🌍 Real-world Use

- In **health research**, mean and median sleep help define "normal" sleep hours.
- In **finance**, they summarize typical daily returns.
- In **education**, mean marks show class performance levels.
- **Example:** A hospital studying patient recovery times can use mean & median to set treatment expectations.

2️⃣ Dispersion (Spread of Data)



🔍 Explanation

- **Variance** → Average of squared distances from the mean.
- **Standard Deviation (σ)** → Typical distance from the mean; tells how *consistent* or *variable* the data is.
- **Range** → Simple difference between highest and lowest value.

💡 Insights

- A **small standard deviation** means most people sleep similarly (consistent patterns).
- A **large std** means huge variation — some sleep too much, others too little.

🌍 Real-world Use

- **Health:** Detect irregular sleep patterns or outliers (e.g., insomnia cases).
- **Business:** Measure volatility in sales or prices.
- **Quality control:** Factories use std dev to monitor product consistency.

3️⃣ Full Statistical Summary

```
df["Sleep Duration"].describe()
```

Metric	Meaning
count	How many people recorded sleep hours
mean	Average sleep hours
std	Variation in hours
min/max	Shortest and longest sleepers
25%, 50%, 75%	Quartiles showing data spread

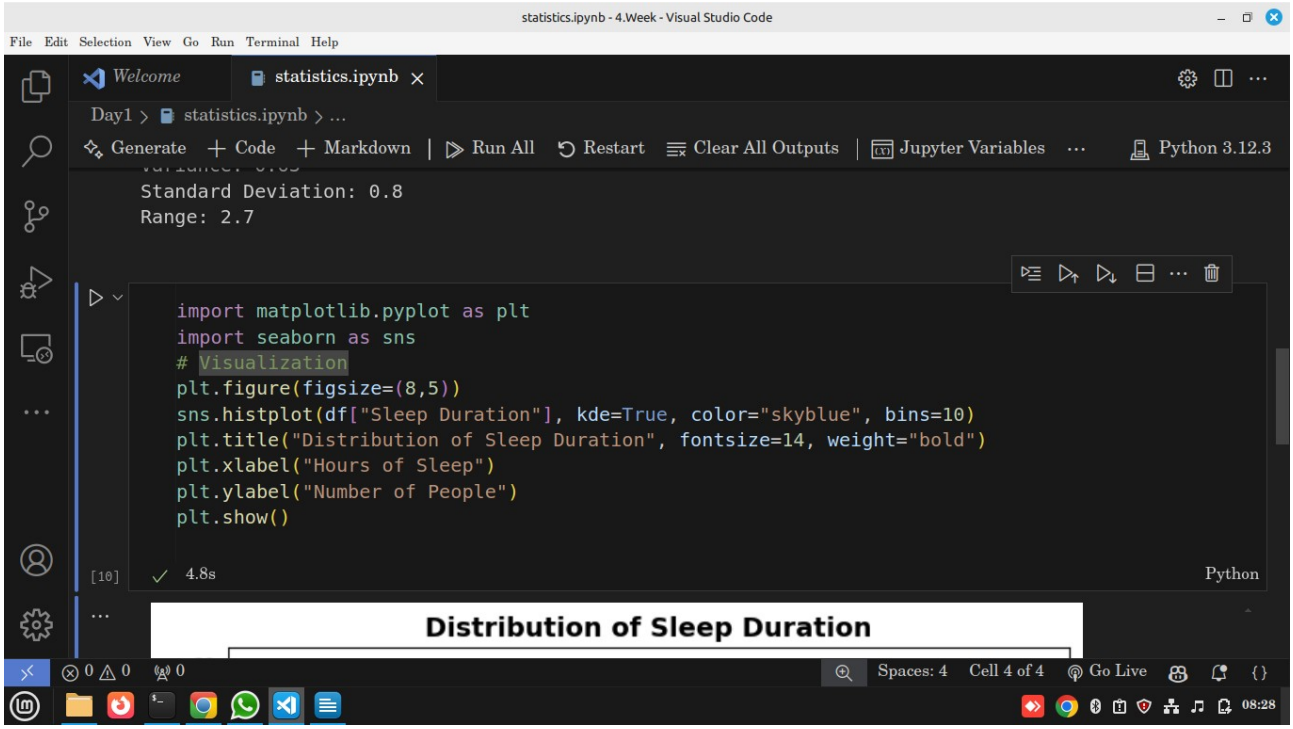
💡 Insights

- The **IQR (Q3–Q1)** shows where the middle 50% of the population lies.
- In your dataset, if Q1 = 6.1 and Q3 = 7.8, then most people sleep between **6–8 hours**.

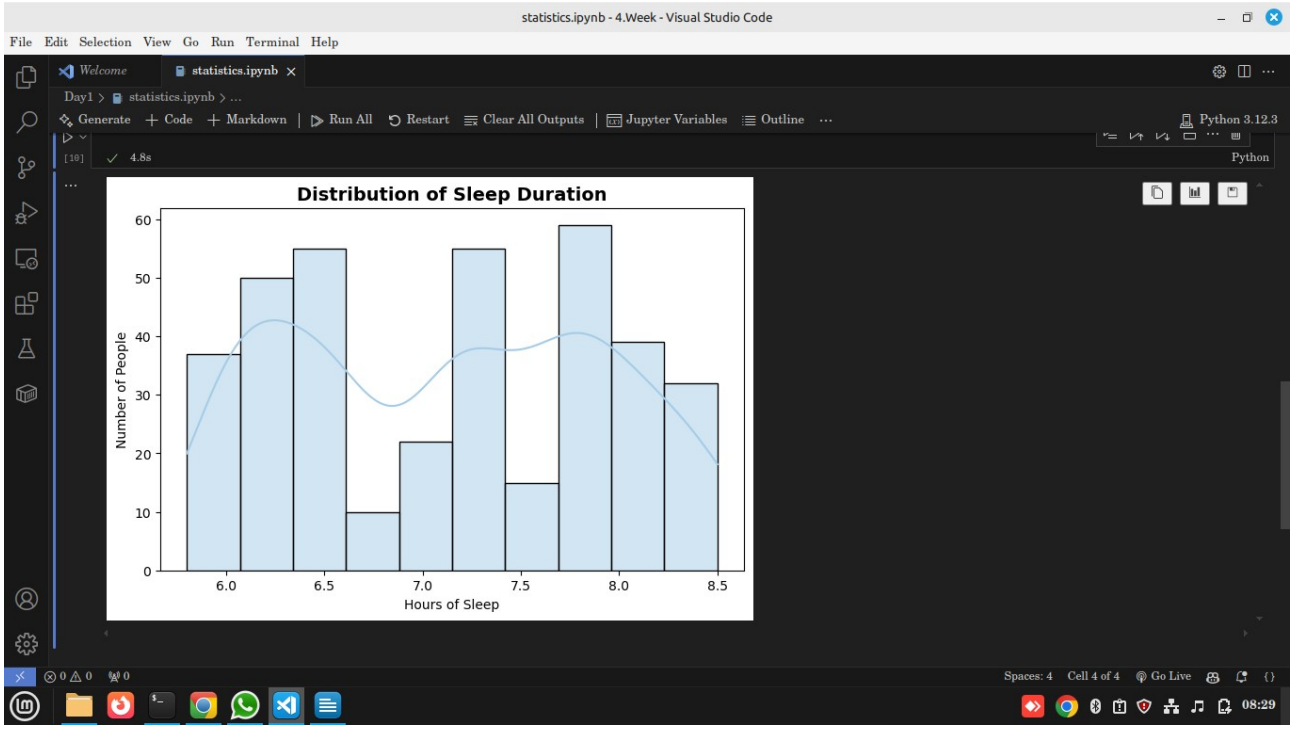
🌍 Real-world Use

- Hospitals or fitness apps (like Fitbit) use quartiles to create "normal sleep range" recommendations.
- Businesses can apply `.describe()` to customer spending data to understand common patterns.

4️⃣ Visualizing the Distribution



Output



🔍 Explanation

- **Histogram** → Shows how frequently each sleep duration occurs.
- **KDE (Kernel Density Estimate)** → Smooth curve showing data density.

💡 Insights

- The KDE peak tells the *most common sleep hours* — around **7–8 hours**.
- Long tails on either side mean a few people sleep unusually less or more.

🌍 Real-world Use

- In **data science**, histograms help check data normality before modeling.
- **Marketing analysts** use them to find common purchase frequencies or price points.
- **Healthcare** can detect sleep issues by spotting unusual sleep hour patterns.

✅ Summary

Concept	Meaning	Python	Real-world Use
Mean	Average	`.mean()`	Average customer spend, average sleep hours
Median	Middle value	`.median()`	House price or income where 50% lie below
Mode	Most frequent	`.mode()`	Common product purchased
Variance	Spread	`.var()`	Financial volatility

Concept	Meaning	Python	Real-world Use
Std Dev	Typical deviation	`.std()`	Performance consistency
Range	Max–Min	`.max() - .min()`	Temperature differences
Describe	Summary stats	`.describe()`	Quick dataset overview

🧩 Assignment

1. Compute mean, median, variance, and std for:
   - `"Stress Level"`
   - `"Quality of Sleep"`
   - `"Physical Activity Level"`
2. Plot a histogram for `"Stress Level"` with KDE.
3. Write **two insights per graph**:
   - What pattern do you see?
   - What could it mean in real life?