

# Descriptive Statistics — *Understanding Your Data*

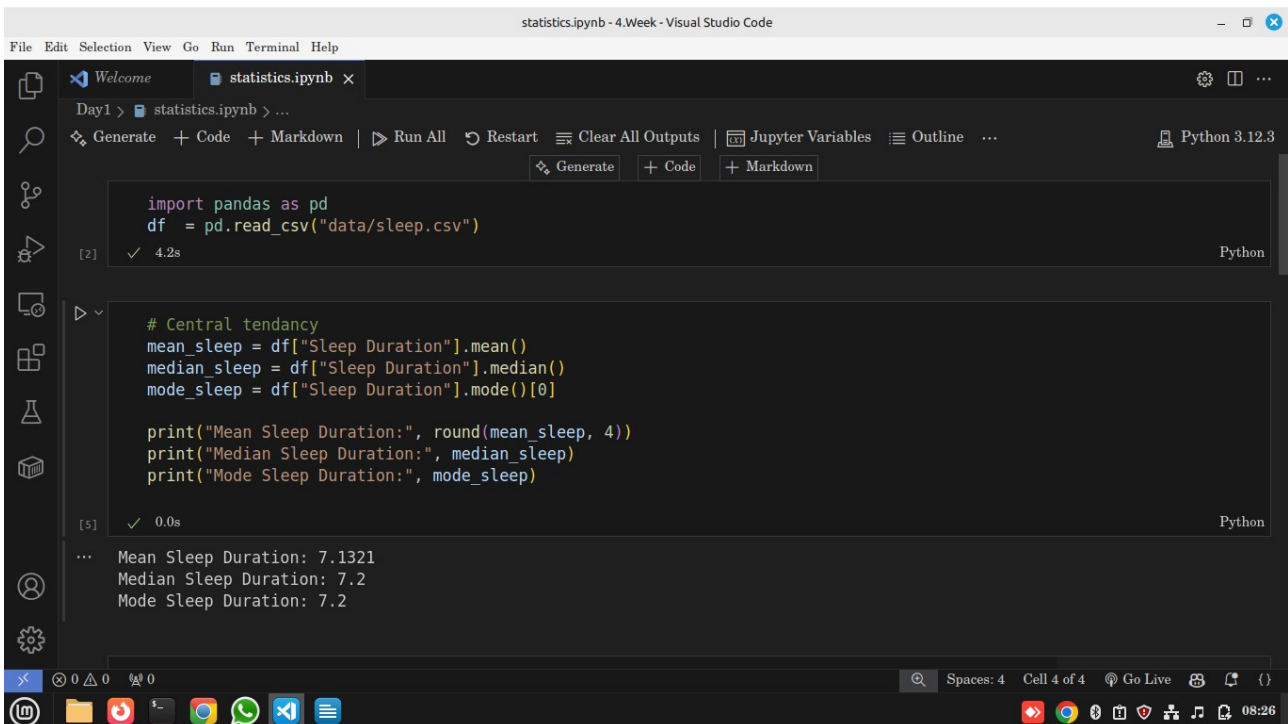
## Objective

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.

## 1 Central Tendency

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



```
statistics.ipynb - 4.Week - Visual Studio Code
File Edit Selection View Go Run Terminal Help

Welcome | statistics.ipynb x
Day1 > statistics.ipynb > ...
Generate + Code + Markdown | Run All Restart Clear All Outputs Jupyter Variables Outline ... Python 3.12.3

import pandas as pd
df = pd.read_csv("data/sleep.csv")

[2] ✓ 4.2s Python

# Central tendency
mean_sleep = df["Sleep Duration"].mean()
median_sleep = df["Sleep Duration"].median()
mode_sleep = df["Sleep Duration"].mode()[0]

print("Mean Sleep Duration:", round(mean_sleep, 4))
print("Median Sleep Duration:", median_sleep)
print("Mode Sleep Duration:", mode_sleep)

[5] ✓ 0.0s Python

...
Mean Sleep Duration: 7.1321
Median Sleep Duration: 7.2
Mode Sleep Duration: 7.2
```

## 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

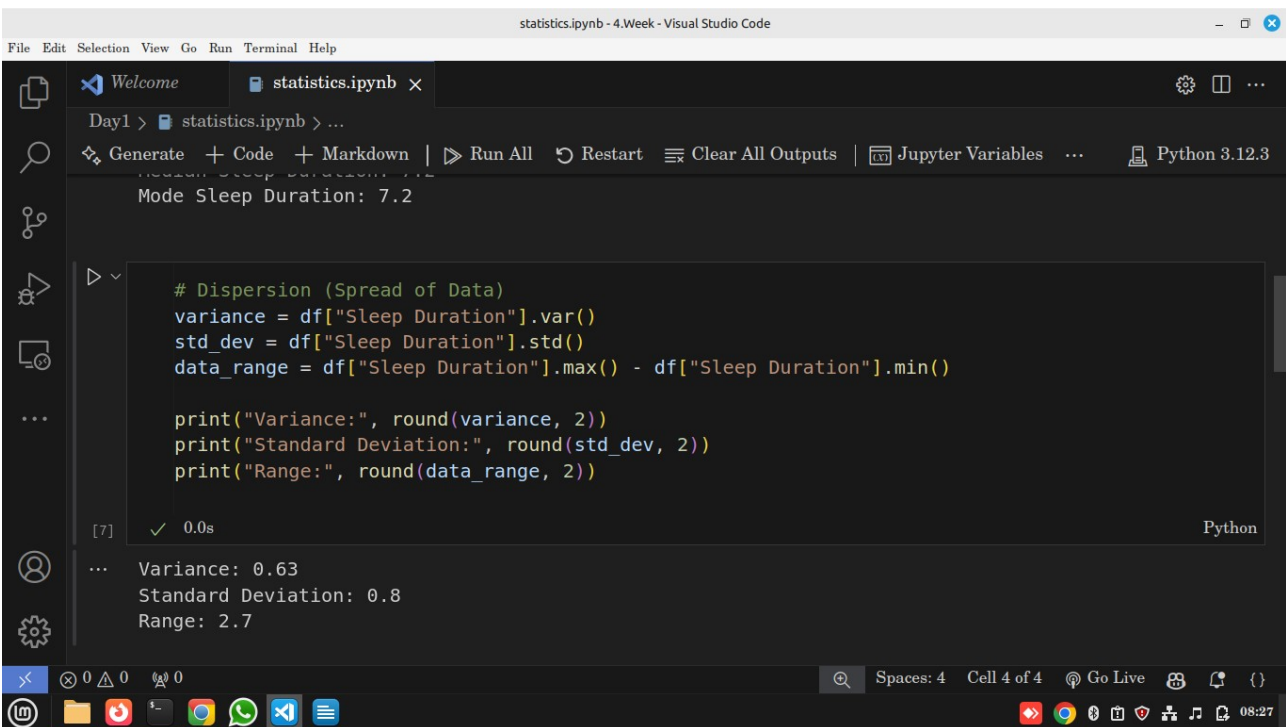
- If your dataset has outliers (e.g., someone sleeping only 2 hours), the **mean** will drop.

- 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.

## 2 Dispersion (Spread of Data)



The screenshot shows a Jupyter Notebook titled 'statistics.ipynb' in Visual Studio Code. The notebook is running a Python script that calculates dispersion statistics for a dataset. The code is as follows:

```
# Dispersion (Spread of Data)
variance = df["Sleep Duration"].var()
std_dev = df["Sleep Duration"].std()
data_range = df["Sleep Duration"].max() - df["Sleep Duration"].min()

print("Variance:", round(variance, 2))
print("Standard Deviation:", round(std_dev, 2))
print("Range:", round(data_range, 2))
```

The output of the script is displayed in the console:

```
[7] ✓ 0.0s
...
Variance: 0.63
Standard Deviation: 0.8
Range: 2.7
```

## Explanation

- **Variance** → Average of squared distances from the mean.
- **Standard Deviation ( $\sigma$ )** → 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.
- 

## 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

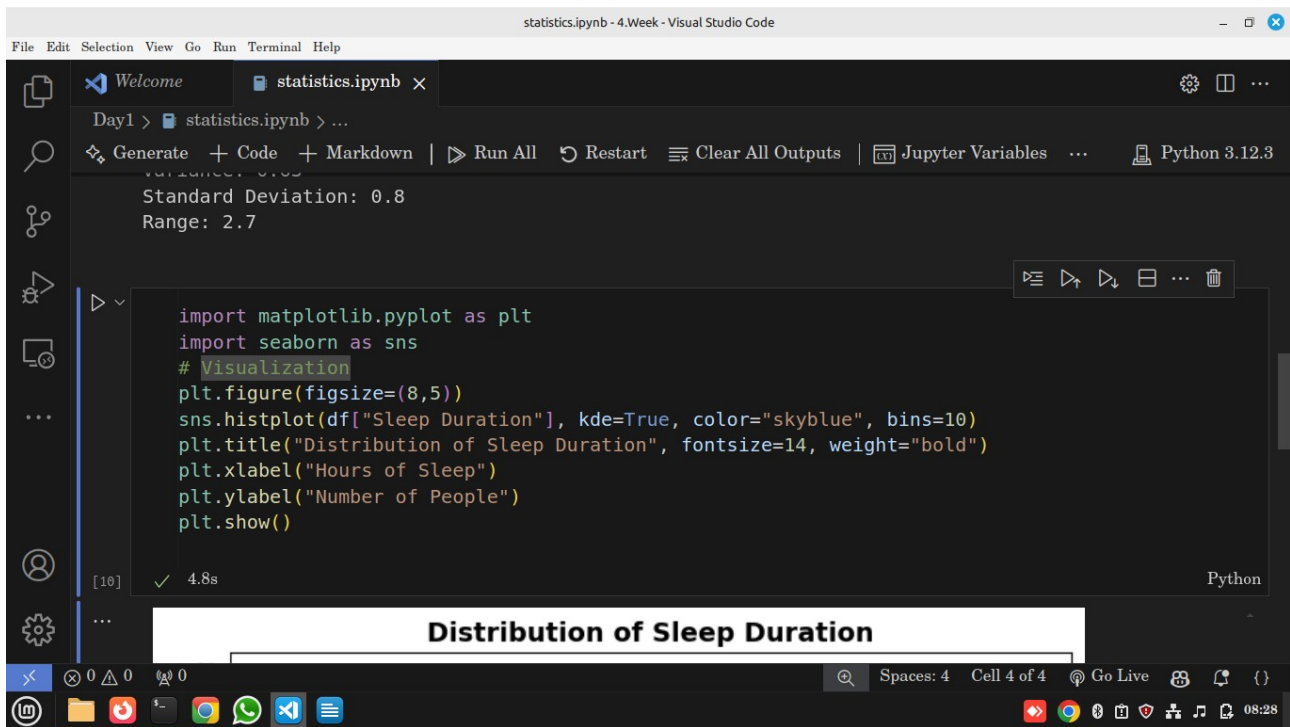
### Insights

- 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**.

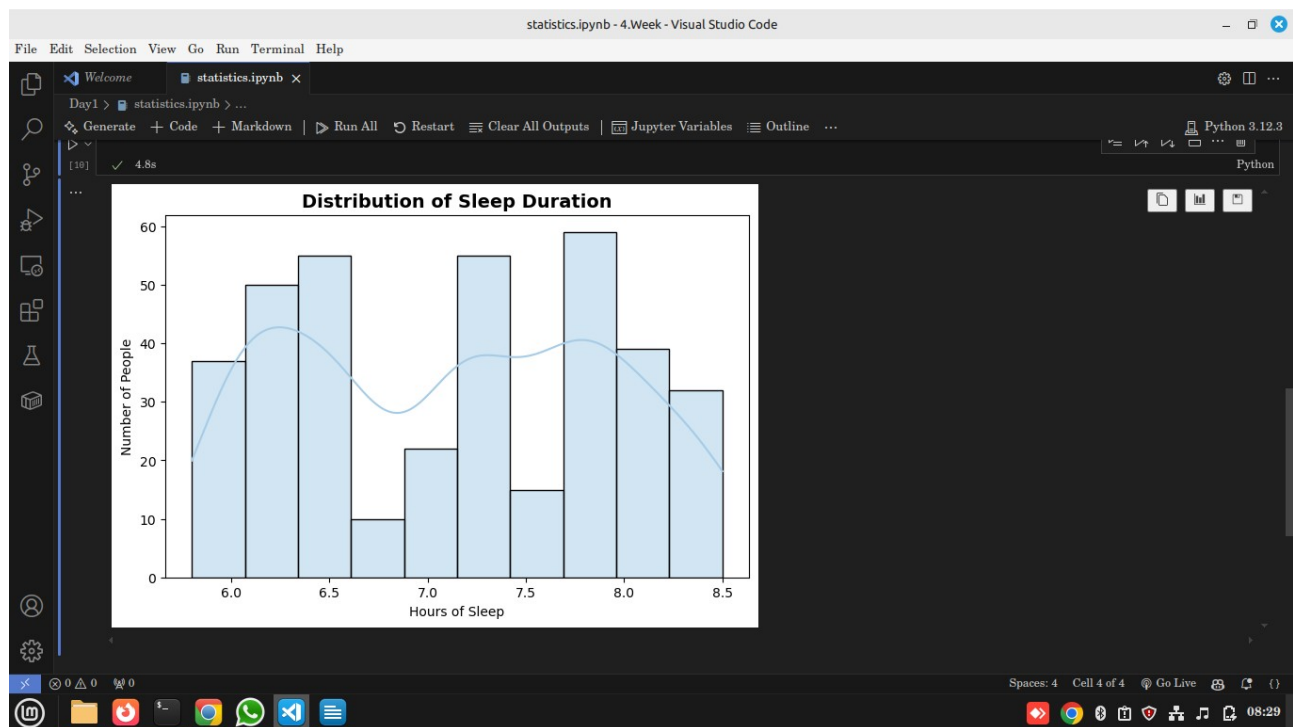
## Real-world Use

- 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.
-

## 4 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	<code>.mean()</code>	Average customer spend, average sleep hours
Median	Middle value	<code>.median()</code>	House price or income where 50% lie below
Mode	Most frequent	<code>.mode()</code>	Common product purchased
Variance	Spread	<code>.var()</code>	Financial volatility

Concept	Meaning	Python	Real-world Use
Std Dev	Typical deviation	<code>.std()</code>	Performance consistency
Range	Max–Min	<code>.max() - .min()</code>	Temperature differences
Describe	Summary stats	<code>.describe()</code>	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?