3 Prac

1. Mean (Average)

- **Definition**: The sum of all values divided by the total number of values.
- **Formula**:

Mean=∑xin\text{Mean} = \frac{\sum x_i}{n}Mean=n∑xi

Where xix_ixi are the data points and nnn is the number of data points.

- **Use**: Gives a central value of the dataset.
- **Example**: If the salaries are ₹10000, ₹20000, and ₹30000,

Mean=10000+20000+300003=₹20000\text{Mean} = \frac{10000 + 20000 + 30000}{3} = ₹20000Mean=310000+20000+30000=₹20000

---

2. Median

- **Definition**: The middle value when the data is arranged in order.
- **Use**: It's a better measure than mean when data has **outliers** (extremely high or low values).
- **Example**: Salaries = ₹10000, ₹20000, ₹30000 → **Median** = ₹20000
  Salaries = ₹10000, ₹20000, ₹1000000 → **Median** = ₹20000 (mean would be misleading here)

---

3. Minimum

- **Definition**: The **smallest** value in the dataset.
- **Use**: Shows the lower boundary of the data.
- **Example**: If ages are 21, 23, 25, **min = 21**

---

4. Maximum

- **Definition**: The **largest** value in the dataset.
- **Use**: Shows the upper boundary of the data.
- **Example**: If incomes are ₹25000, ₹50000, ₹90000 → **max = ₹90000**

---

5. Standard Deviation (std)

- **Definition**: A measure of **how spread out** the values are from the mean.
- **Formula**:

std=1n∑i=1n(xi−mean)2\text{std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \text{mean})^2}std=n1i=1∑n(xi−mean)2

- **Use**:
  - Low std = data points are **close to the mean**
  - High std = data points are **spread out**
- **Example**:
  - Salaries = ₹10000, ₹10500, ₹11000 → std is **low**
  - Salaries = ₹10000, ₹50000, ₹1000000 → std is **high**

Let me know if you'd like a visual plot to understand these better!

4o

You said:

what is data points

ChatGPT said:

Great question, Omkar!

Data Points – simply put:

A **data point** is a single value or observation in your dataset.

📊 **Example:**

If you have student scores:

csharp

CopyEdit

[85, 90, 78, 92, 88]

Each number (85, 90, 78, etc.) is a **data point**. So this dataset has **5 data points**.

🧠 **In real-world terms:**

- If you're collecting heights of 10 people → each person's height is a **data point**.
- If you're analyzing rainfall each day for a month → each day's rainfall is a **data point**.

📌 **Why it's important:**

We use **data points** to:

- Calculate mean, median, standard deviation, etc.
- Visualize data (plots, graphs)
- Detect patterns, trends, or outliers

Would you like a simple chart or visual to show how data points form a dataset?

Objective

You're performing **descriptive statistics** (measures of central tendency and variability) on two datasets:

1. Employee_Salary_Dataset.csv — Grouping numerical variables (like salary) by categorical (like age).
2. Iris.csv — Grouping flower features (sepal, petal dimensions) by species and computing statistics.

🧩 Part 1: Employee Salary Dataset Analysis

🔹 Read and Display Dataset

python

CopyEdit

```
df = pd.read_csv('datasets/Employee_Salary_Dataset.csv')

df.head()
```

- Loads employee salary data into a DataFrame.
- Columns: ID, Experience_Years, Age, Gender, Salary.

🔹 Basic Descriptive Statistics

python

CopyEdit

```
print(df.mean(numeric_only=True))

print(df.median(numeric_only=True))

print(df.min(numeric_only=True))

print(df.max(numeric_only=True))

print(df.std(numeric_only=True))

print(df.describe(include='all'))

print(df.corr(numeric_only=True))
```

These lines compute:

- **Mean, Median, Min, Max, Std. Deviation** — for numeric columns.
- describe() — gives a full summary of stats including count, percentiles, and unique value counts.
- corr() — shows correlation between numerical variables. For example:
    - Experience_Years and Age have **strong positive correlation** (0.97), as expected.

◆ Grouped Summary Statistics

python

CopyEdit

```
grouped = df.groupby('Age')['Salary'].agg(['mean', 'median', 'min', 'max', 'std'])

grouped_list = df.groupby('Age')['Salary'].apply(list)
```

- Groups Salary by Age, and computes:
    - Mean salary per age, median, min, max, and standard deviation.
- Also builds a **list of all salaries** for each age (useful for plotting or further analysis).

This helps answer questions like: *“What is the average salary of all 29-year-olds?”*

🧩 Part 2: Iris Dataset Analysis

◆ Load Iris Dataset

python

CopyEdit

```
dff = pd.read_csv('datasets/Iris.csv')

iris = dff.groupby('Species')
```

- Loads the classic **Iris dataset**, which contains:
    - Features like SepalLengthCm, PetalWidthCm, etc.
    - Categorical class: Species — Iris-setosa, Iris-versicolor, Iris-virginica.

◆ Descriptive Statistics by Species

python

CopyEdit

```
iris.describe(percentiles=[.25, .5, .75])

iris.mean(numeric_only=True)
iris.median(numeric_only=True)

iris.quantile(numeric_only=True)

iris.std(numeric_only=True)

iris.min(numeric_only=True)

iris.max(numeric_only=True)
```

This part answers:

- What are the **average sepal/petal sizes** per species?
- What are the **minimum/maximum/median values**?
- What's the **spread** (standard deviation)?
- How different are Iris-setosa vs. Iris-virginica in terms of petal length?

The describe() command gives a concise statistical summary with percentiles.

🎯 Summary: What You've Achieved

Dataset	Action	Result
Employee Salary	Summarized age-wise salary stats	Insight into salary trends across ages
Employee Salary	Correlation between age/experience/salary	Age and experience are highly correlated with salary
Iris	Computed species-wise statistics	Understand how flower features vary across species

---

Would you like help plotting any of these statistics as graphs (e.g., boxplots or histograms)?

4o

You said:

explain code

ChatGPT said:

Sure Omkar! Let me walk you through the code you've written step by step. Your code is divided into **two major parts**:

🔹 Part 1: Descriptive Statistics on Employee Dataset

✅ 1. Loading Libraries and Dataset

python

CopyEdit

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('datasets/Employee_Salary_Dataset.csv')
df.head()
```

- **pandas**, **numpy**, **matplotlib**, and **seaborn** are imported for data handling and visualization.
- df.head() shows the first 5 rows of the dataset.

✅ 2. Basic Summary Statistics

python

CopyEdit

```
print("Mean")
print(df.mean(numeric_only=True))
```

- **mean()**, **median()**, **min()**, **max()**, **std()** are used to compute basic statistics **only for numeric columns**.
- describe() gives a full statistical summary.
- corr() calculates correlation between numeric columns.

✅ 3. Grouped Summary Statistics by Age

python

CopyEdit

```
grouped = df.groupby('Age')['Salary'].agg(['mean', 'median', 'min', 'max', 'std'])
```

- Groups the dataset by **Age** and computes **summary statistics** for **Salary**.

python

CopyEdit

```
grouped_list = df.groupby('Age')['Salary'].apply(list)
```

- Converts each group into a **list of Salary values per age**.

◆ Part 2: Descriptive Statistics on Iris Dataset

✅ 1. Load Iris Dataset

python

CopyEdit

```
dff = pd.read_csv('datasets/Iris.csv')
dff.head()
```

- Loads the classic **Iris flower dataset**.

✅ 2. Group by Species

python

CopyEdit

```
iris = dff.groupby('Species')
```

- Grouping the dataset by **Species**: Iris-setosa, Iris-versicolor, Iris-virginica.

✅ 3. Statistics by Species

Each of these provides different descriptive stats **per species**:

python

CopyEdit

```
print(iris.describe(percentiles=[.25, .5, .75]))
```

- **Percentiles**, **mean**, **std**, **min**, **max** for each numeric column.

python

CopyEdit

```
print(iris.mean(numeric_only=True))
print(iris.median(numeric_only=True))
print(iris.quantile(numeric_only=True))
```

- These give **mean**, **median**, and **quantiles** per species.

python

CopyEdit

```
print(iris.std(numeric_only=True))
print(iris.min(numeric_only=True))
print(iris.max(numeric_only=True))
```

- Shows standard deviation, min, and max per species.

python

CopyEdit

```
print(iris.describe(include='all'))
```

- Gives a complete summary for **each group** (species).

🔚 Summary

You successfully:

1. Computed overall and grouped statistics on the **Employee dataset**.
2. Computed detailed per-species descriptive stats from the **Iris dataset**.

This covers **measures of central tendency (mean, median)** and **variability (std, min, max, percentiles)** — exactly what the practical asked for.

Would you like me to help visualize any of these results using plots or graphs?

4o