1. PDF Keywords – Meaning, How They Work, and Where Used

These keywords are **Pandas DataFrame operations**, mainly used in **Employee / HR data analysis**.

---

1. `head()`

Meaning:
Shows the **first 5 rows** of the dataset (by default).

How it works:
Helps quickly preview the data structure.

Where used:
Initial data inspection.

---

2. `tail()`

Meaning:
Shows the **last 5 rows** of the dataset.

How it works:
Useful to verify the ending records.

Where used:
Checking recent or final entries.

---

3. `describe()`

Meaning:
Generates **statistical summary** of numeric columns.

How it works:
Displays count, mean, std, min, max, and quartiles.

Where used:
Understanding employee salary, age, experience, etc.

---

4. `info()`

Meaning:
Provides **dataset structure** information.

How it works:
Shows column names, data types, non-null counts.

Where used:
Data quality checks before analysis.

---

5. `display()`

Meaning:
Visually displays the dataset in tabular form.

How it works:
Better formatting than print.

Where used:
Exploratory Data Analysis (EDA).

---

6. `shape`

Meaning:
Returns **rows and columns count**.

How it works:
Output format → `(rows, columns)`

Where used:
Understanding dataset size.

---

7. `dropna()`

Meaning:
Removes rows with **missing values**.

How it works:
Deletes NaN-containing records.

Where used:
Data cleaning stage.

---

8. `fillna()`

Meaning:
Fills missing values with a given value.

How it works:
Replaces NaN with mean, median, 0, etc.

Where used:
Preventing data loss during analysis.

---

9. `drop_duplicates()`

Meaning:
Removes duplicate rows.

How it works:
Keeps unique employee records.

Where used:
Avoiding double-counting employees.

---

10. `columns`

Meaning:
Returns column names.

How it works:
Helps identify available attributes.

Where used:
Feature selection and understanding schema.

---

11. `groupby()`

Meaning:
Groups data based on a column.

How it works:
Performs aggregation (mean, sum, count).

Where used:
Department-wise or role-wise analysis.

---

2. Statistical Numeric Keywords (From PDF)

Mean

Average value (e.g., average salary)

Max / Min

Highest / lowest value in a column

`sort_values()`

Sorts data based on a column

Ascending / Descending

Controls sort order

`value_counts()`

Counts unique values (e.g., attrition Yes/No)

`size`

Total number of rows in a group

`std`

Standard deviation (data spread)

Quartiles (25%, 50%, 75%)

Data distribution points:

- 25% → Lower range
- 50% → Median
- 75% → Upper range

def

---

3. What Is Attrition? (From PDF)

Attrition means:
👉 **Employees leaving the organization** (resignation, retirement, termination).

Why it matters:
Used to analyze:

- Employee turnover
- Job satisfaction
- HR retention strategies

def

---

4. Main Columns Seen in Employee Data (Typical)

Based on the operations in the PDF, Employee datasets usually contain:

Core Employee Columns

- EmployeeID
- Age
- Gender
- Department
- JobRole
- Education
- Experience / YearsAtCompany

Compensation Columns

- MonthlyIncome / Salary
- Bonus
- PercentSalaryHike

Performance & Work

- JobSatisfaction
- PerformanceRating
- WorkLifeBalance
- Overtime

Attrition Column (Very Important)

- Attrition (Yes / No)