AI VIET NAM
@aivietnam.edu.vn

AI VIET NAM – AI COURSE 2026

Pandas và SQL

Nguyễn Thế Hào Đinh Quang Vinh

I. Tư duy SQL trong Pandas

Pandas - "SQL"dành cho Python

Nếu bạn đã quen thuộc với ngôn ngữ truy vấn SQL, việc tiếp cận Pandas sẽ trở nên rất tự nhiên. Hãy hình dung DataFrame trong Pandas chính là một Table trong cơ sở dữ liệu.

- **SELECT (Chọn cột)**: Tương tự việc `SELECT column` trong SQL, Pandas cho phép chọn đặc trưng bằng cú pháp `df['column']`.
- **WHERE (Lọc hàng)**: Thay vì dùng mệnh đề `WHERE`, Pandas sử dụng phương thức `.loc` kết hợp với điều kiện logic để lọc bản ghi
- **AND/OR (Kết hợp điều kiện)**: Các toán tử logic trong SQL được chuyển đổi thành toán tử bitwise (`&`, `|`) trong Pandas.

Thay vì viết câu lệnh dạng chuỗi văn bản, Pandas cho phép ta thực hiện các thao tác "truy vấn"này trực tiếp thông qua cú pháp lập trình Python.

II. Ví dụ minh họa

Bộ dữ liệu Advertising Simple

Chúng ta sẽ sử dụng bộ dữ liệu mẫu `advertising_simple.csv`. Đây là dữ liệu mô phỏng chi phí quảng cáo trên các kênh khác nhau và doanh số bán hàng tương ứng.

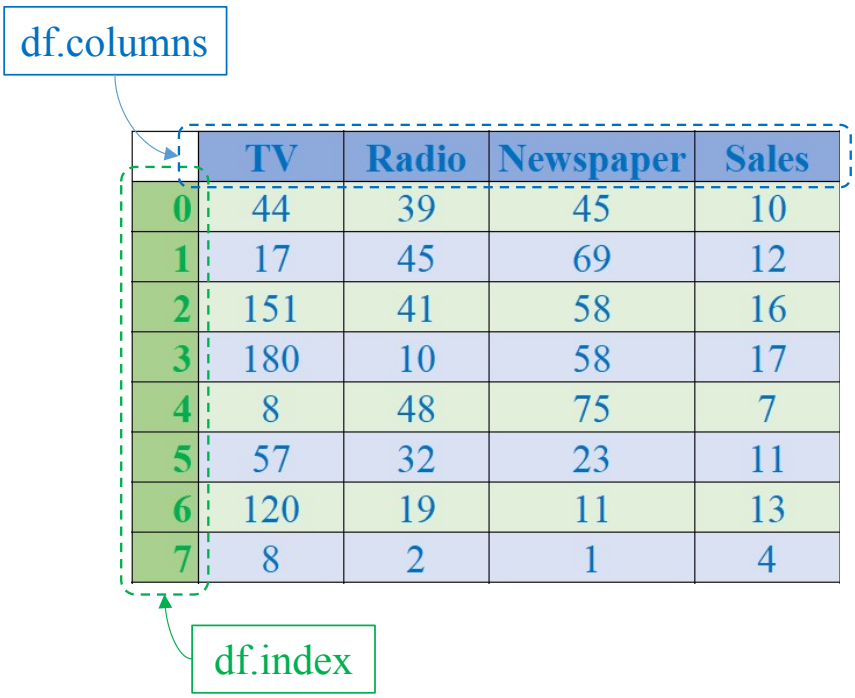
- **Các cột (Features)**:
  - `TV`: Chi phí quảng cáo trên TV.
  - `Radio`: Chi phí quảng cáo trên Radio.
  - `Newspaper`: Chi phí quảng cáo trên Báo chí.
- **Mục tiêu (Target)**:
  - `Sales`: Doanh số bán hàng thu được.

II.1. Thực hành thao tác dữ liệu

Bước 1: Đọc dữ liệu và kiểm tra cấu trúc

```python
import pandas as pd

# 1. Đọc dữ liệu từ file CSV
df = pd.read_csv('advertising_simple.csv')
```



	TV	Radio	Newspaper	Sales
0	44	39	45	10
1	17	45	69	12
2	151	41	58	16
3	180	10	58	17
4	8	48	75	7
5	57	32	23	11
6	120	19	11	13
7	8	2	1	4

Hình 1: Dữ liệu mẫu Advertising Simple

II.2. Thao tác 1: Chọn cột dữ liệu (SQL SELECT)

Trong SQL, để lấy một cột cụ thể, ta dùng `SELECT column_name`. Trong Pandas, ta dùng cú pháp `df['column_name']`.

Lấy riêng cột Radio

```python
df['Radio']
```

SQL Tương ứng

```sql
SELECT Radio FROM df
```

	Radio
0	39
1	45
2	41
3	10
4	48
5	32
6	19
7	2

Hình 2: Kết quả khi chọn riêng cột Radio

II.3. Thao tác 2: Lọc theo điều kiện đơn (SQL WHERE)

Đây là thao tác phổ biến nhất để loại bỏ nhiễu hoặc tìm các mẫu dữ liệu quan tâm. Ta sử dụng toán tử so sánh ($>, <, ==$) bên trong dấu ngoặc vuông hoặc phương thức `.loc`.

Lọc dữ liệu có Newspaper < 30

```
low_newspaper = df.loc[df.Newspaper < 30]

print("\nCác bản ghi có Newspaper < 30:")
print(low_newspaper)
```

SQL Tương ứng

```
SELECT * FROM df WHERE Newspaper < 30
```

	TV	Radio	Newspaper	Sales
5	57	32	23	11
6	120	19	11	13
7	8	2	1	4

Hình 3: Kết quả lọc dữ liệu có Newspaper < 30

Lưu ý

Lưu ý về cú pháp ⚠ Khi lọc dữ liệu, biểu thức `df.Newspaper < 30` thực chất sẽ trả về một `Series` kiểu Boolean (True/False). Pandas sử dụng Series này để quyết định giữ lại hay loại bỏ hàng nào trong DataFrame.

II.4. Thao tác 3: Lọc đa điều kiện (SQL AND/OR)

Khi cần kết hợp nhiều điều kiện, trong SQL ta dùng `AND`, `OR`. Trong Pandas, ta phải sử dụng các toán tử bitwise:

- & (thay cho AND)
- | (thay cho OR)
- ∼ (thay cho NOT)

Kết hợp điều kiện Sales và Newspaper

```
df.loc[(df.Sales > 10) & (df.Newspaper < 30)]
```

SQL Tương ứng

```
SELECT * FROM df WHERE Sales > 10 AND Newspaper < 30
```

	TV	Radio	Newspaper	Sales
5	57	32	23	11
6	120	19	11	13

Hình 4: Dữ liệu sau khi áp dụng nhiều điều kiện lọc

Phụ lục

1. **Code:** Các file code được đề cập trong bài có thể được tải tại **đây**.
2. **Q&A:** Bạn có thể đặt thêm câu hỏi về nội dung bài đọc trong group Facebook hỏi đáp tại đây. Tất cả câu hỏi sẽ được trả lời trong vòng tối đa 4 tiếng.

Hình 1: Hình ảnh group facebook AIO Q&A.