

BÁO CÁO PHÂN TÍCH HIỆU SUẤT BỆNH VIỆN / PHÒNG KHÁM N4_CASE 9

I. GIỚI THIỆU VÀ ĐẶT VẤN ĐỀ

Mục tiêu của Case Study 09 là thực hành đầy đủ các kỹ năng Pandas cốt lõi, bao gồm làm sạch dữ liệu, truy vấn, phân nhóm, kết hợp (merge), và xây dựng bảng xoay (pivot) trong bối cảnh Phân tích Y tế (Health Analytics).

1. Dữ liệu sử dụng Dữ liệu gốc bao gồm ba tệp CSV/Excel chứa thông tin về bệnh nhân, hồ sơ khám bệnh, và thông tin thanh toán:

- + patient_info.csv: Chứa mã bệnh nhân (patient_id), tên (full_name), tuổi (age), và giới tính (gender).
- + visit_records.csv: Chứa mã lượt khám (visit_id), mã bệnh nhân (patient_id), ngày khám (visit_date), khoa khám (department), và chẩn đoán (diagnosis).
- + billing_info.csv: Chứa mã lượt khám (visit_id), tổng phí (total_fee), và phần trăm BHYT chi trả (insurance_coverage).

II. PHƯƠNG PHÁP XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU

Phương pháp phân tích chủ yếu dựa trên các kỹ thuật lập trình hướng mảng và xử lý dữ liệu của thư viện **Pandas**, được xây dựng trên nền tảng NumPy

Giai đoạn phân tích	Kỹ thuật/Hàm Pandas chủ đạo	Mục đích sử dụng
Làm sạch dữ liệu	pd.to_datetime(), str.strip(), str.title(), dropna(), fillna(), astype()	Chuẩn hóa kiểu dữ liệu cột ngày (datetime64[ns]) và xử lý các giá trị không đồng nhất (ví dụ: tuổi, giới tính, phí) và dữ liệu thiếu (NaN/None)
Truy vấn & Thống kê	.loc, .iloc, Boolean Masking, .describe(), .mean()	Truy xuất dữ liệu theo nhãn hoặc vị trí, tạo mặt nạ logic để lọc dữ liệu phức tạp, và cung cấp tóm tắt thống kê mô tả
Group nhóm (GroupBy)	.groupby(), .agg(), .apply(), .mean(), .sum()	Group nhóm dữ liệu theo các tiêu chí (ví dụ: Khoa) để tính toán các đặc trưng tổng hợp như chi phí trung bình hoặc số lượt khám
Kết hợp (Merge)	pd.merge(), how='inner'/'left', on	Kết hợp các bảng dữ liệu (Bệnh nhân, Lượt khám, Thanh toán) dựa trên các khóa chung (patient_id, visit_id)
Bảng xoay (Pivot)	pd.pivot_table(), pd.cut(), .stack(), .unstack()	Tạo các báo cáo tổng hợp đa chiều trực quan (ví dụ: Chi phí theo Khoa x Nhóm tuổi).

III. KẾT QUẢ VÀ PHÂN TÍCH CHÍNH (THEO NHIỆM VỤ)

1. Nhiệm vụ 1 – Đọc & làm sạch dữ liệu

+ **Mục tiêu:** Xử lý các vấn đề về định dạng, khoảng trắng thừa, giá trị không chuẩn (ví dụ: tuổi âm, total_fee có ký tự tiền tệ, gender dạng text/viết thường).

+ **Kết quả chính:** Ba tập dữ liệu sạch, không còn các lỗi cấu trúc cơ bản, sẵn sàng cho các nhiệm vụ tiếp theo.

2. Nhiệm vụ 2 – Truy vấn & thống kê mô tả

+ **Phương pháp:** Sử dụng Boolean Masking (.loc) và các hàm tổng hợp cơ bản (.count(), .mean()).

+ **Các bảng kết quả chính:**

- **Thống kê số bệnh nhân theo khoa:** Cho thấy phân bố lượt khám giữa các khoa (Nội, Ngoại, Sản, Nhi).

- **Lượt khám không có chẩn đoán:** Xác định số lượng visit_id có trường diagnosis là NaN/None (sử dụng isnull() hoặc notnull()).

- **Thống kê tuổi trung bình theo khoa:** Cung cấp bức tranh tổng quan về nhóm tuổi mà mỗi khoa phục vụ.

3. Nhiệm vụ 3 – GroupBy & Tổng hợp

+ **Phương pháp:** Áp dụng nguyên tắc Split-Apply-Combine (groupby()) để tính toán các chỉ số kinh doanh cốt lõi.

+ **Các bảng kết quả chính:**

- **Chi phí trung bình theo khoa:** Sử dụng df.groupby('department')['total_fee'].mean() để xác định mức chi phí trung bình mà bệnh nhân phải trả ở mỗi khoa.

- **Số lượt khám theo bệnh nhân:** Cho biết tần suất khám bệnh của từng bệnh nhân.

- **Phân tích mối liên hệ giữa tuổi và chi phí khám:** Đòi hỏi phải nhóm tuổi (sử dụng pd.cut()) và tính toán tương quan chi phí.

- **Xác định khoa có chi phí cao nhất:** Dễ dàng suy ra từ kết quả chi phí trung bình.

4. Nhiệm vụ 4 – Merge dữ liệu

+ **Mục tiêu:** Tạo ra một DataFrame hoàn chỉnh chứa tất cả thông tin (Bệnh nhân, Lượt khám, Thanh toán).

+ **Phương pháp:**

1. Merge patient_info và visit_records (key: patient_id).

2. Merge kết quả với billing_info (key: visit_id).

+ **Kết quả chính:** DataFrame tổng hợp (final_data.csv) phục vụ cho các phân tích đa chiều. Việc phát hiện các lượt khám thiếu thông tin thanh toán được thực hiện bằng cách kiểm tra giá trị NaN ở các cột billing_info sau khi thực hiện merge (ví dụ: how='left').

5. Nhiệm vụ 5 – Pivot Table + Stack/Unstack

+ **Phương pháp:** Sử dụng pd.pivot_table() để tái cấu trúc dữ liệu tổng hợp thành định dạng báo cáo hai chiều (index x columns).

+ **Các bảng kết quả chính:**

- **Pivot chi phí trung bình theo Khoa x Nhóm tuổi:** Cần sử dụng pd.cut() để chuyển cột tuổi (biến liên tục) thành các nhóm rời rạc trước khi tạo Pivot Table. Đây là ví dụ điển hình về việc tổng hợp dữ liệu đa chiều (multidimensional aggregation).

- **Pivot số lượt khám theo Tháng x Khoa:** Cho thấy mô hình sử dụng dịch vụ theo thời gian.

- **Nhận xét:** Dựa trên kết quả Pivot (Số lượt khám/Tháng/Khoa), nhận xét khoa nào có **mật độ bệnh nhân cao nhất** (tức là số lượt khám lớn nhất).

IV. NHẬN XÉT VÀ KẾT LUẬN

1. Nhận xét từ kết quả phân tích (Phần này sẽ được điền chi tiết sau khi thực hiện các nhiệm vụ 1-5. Dưới đây là khung nhận xét cần có, dựa trên các yêu cầu phân tích):

+ *Tính nhất quán dữ liệu*: Đánh giá mức độ phức tạp của việc làm sạch, đặc biệt ở các trường age, total_fee, và visit_date.

Phân bổ tải dịch vụ: Phân tích khoa có tần suất lượt khám cao nhất (Nhiệm vụ 2, 5).

+ *Phân bổ tài chính*: Khoa có chi phí trung bình cao nhất và tỷ lệ BHYT trung bình.

+ *Hiệu suất báo cáo*: Đánh giá tính trực quan và hiệu quả của Pivot Table so với việc sử dụng chuỗi lệnh groupby().mean().unstack().

2. Kết luận Tóm tắt các phát hiện quan trọng nhất về hiệu suất của bệnh viện/phòng khám (ví dụ: mô hình chi phí theo khoa, xu hướng khám bệnh theo thời gian), khẳng định việc áp dụng Pandas đã giúp chuyển đổi dữ liệu thô sang thông tin chi tiết hỗ trợ ra quyết định

V. NHẬT KÍ NHÓM

Số thứ tự	Họ và Tên	Mã sv	Nhiệm vụ	Nội dung	Đánh Giá
1	Đỗ Mạnh Huân	24174600112	NV 1	Nhóm trưởng , Làm sạch dữ liệu: chuẩn hóa ID, tên, tuổi, giới tính, ngày khám, khoa, chẩn đoán, chi phí	Tích cực
2	Đặng Văn Hùng	24174600113	NV 2	Truy vấn & thống kê: số BN theo khoa, tuổi TB, lượt khám, tỷ lệ BHYT	Chủ động
3	Nguyễn Đức Dinh	24174600061	NV 3	GroupBy & tổng hợp: chi phí TB, số lượt khám, tỷ lệ BHYT, phân tích tuổi-chi phí	Đúng tiến độ
4	Phùng Quốc Khánh	24174600114	NV 4	Merge dữ liệu theo patient_id, visit_id; kiểm tra thiếu thông tin thanh toán	Phối hợp tốt
5	Bùi Thị Thanh Hòa	24174600064	NV 5	Pivot, stack/unstack; phân tích chi phí, lượt khám theo khoa, tháng, viết báo cáo chi tiết	Hoàn thành tốt