Đại học Quốc gia Thành phố Hồ Chí Minh

Đại học Khoa học tự nhiên

Khoa Công nghệ thông tin

Nhóm 3

MIDTERM: Dashboard trực quan hóa dữ liệu

Môn học	Trực quan hóa dữ liệu
Lớp	22KHDL
Students	**22127064 – Phạm Thành Đạt** **22127225 – Trần Thị Thiên Kim** **22127357 – Phạm Trần Yến Quyên** **22127374 – Lê Thanh Tâm** **22127449 – Mai Đức Vân**
Github:	Thu nhập dữ liệu và Trực quan hóa dữ liệu

Thành phố Hồ Chí Minh, 2025

1 Giới thiệu đồ án

Trong những năm gần đây, bệnh phổi đã trở thành một trong những nguyên nhân hàng đầu gây tử vong trên toàn thế giới. Việc phân tích dữ liệu liên quan đến bệnh phổi không chỉ giúp hiểu rõ hơn về các yếu tố nguy cơ mà còn hỗ trợ trong việc đưa ra các quyết định y tế hiệu quả. Đồ án này tập trung vào việc xây dựng một dashboard phân tích dữ liệu bệnh phổi bằng cách sử dụng thư viện Streamlit trong Python. Dashboard này cho phép người dùng tải lên dữ liệu về bệnh phổi dưới dạng tệp CSV, sau đó thực hiện các phân tích thống kê, trực quan hóa dữ liệu và đưa ra các nhận xét về các yếu tố ảnh hưởng đến bệnh phổi như tuổi, giới tính, tình trạng hút thuốc, loại bệnh, và kết quả hồi phục.

2 Mục tiêu của đồ án

- Phân tích dữ liệu bệnh phổi: Tìm hiểu các yếu tố ảnh hưởng đến bệnh phổi như tuổi, giới tính, tình trạng hút thuốc, loại bệnh, và kết quả hồi phục.
- Trực quan hóa dữ liệu: Sử dụng các biểu đồ như biểu đồ cột, biểu đồ tròn, biểu đồ hộp, biểu đồ phân tán, và heatmap để hiển thị các mối quan hệ giữa các biến số.
- Giúp người dùng (bác sĩ, nhà nghiên cứu, bệnh nhân) dễ dàng khám phá dữ liệu, hiểu xu hướng bệnh, và có thể dự đoán nguy cơ mắc bệnh phổi.
- Đưa ra nhận xét và kết luận: Dựa trên các phân tích và trực quan hóa, đưa ra các nhận xét về các yếu tố ảnh hưởng đến bệnh phổi và khả năng hồi phục của bệnh nhân.

3 Các công nghệ sử dụng trong đồ án

- `Streamlit`: Thư viện Python để xây dựng giao diện web tương tác.
- `Pandas`: Thư viện để xử lý và phân tích dữ liệu.
- `Matplotlib và Seaborn`: Thư viện để vẽ các biểu đồ trực quan hóa dữ liệu.
- `Plotly`: Thư viện để tạo các biểu đồ tương tác dùng trực tiếp với `streamlit`.
- `Scikit-learn`: Thư viện để mã hóa dữ liệu dạng phân loại (`categorical`).

4 Tiền xử lý dữ liệu

Dữ liệu tải về từ **Lungs Diseases Dataset** và được tiền xử lý ở file `jupyter notebook` trước khi đưa vào dashboard, bao gồm xử lý giá trị thiếu và dịch sang tiếng Việt.

4.1 Các bước tiền xử lý chính

- **Nhập dữ liệu:** Dữ liệu được nhập vào từ tệp `lung_disease_data.csv` bằng `pd.read_csv()`.

- **Kiểm tra kích thước và thông tin dữ liệu:** Sử dụng `df.shape` để kiểm tra số dòng và cột, `df.info()` để xem thông tin kiểu dữ liệu của từng cột.
- **Kiểm tra giá trị thiếu:** Hàm `df.isnull().sum()` được dùng để đếm số giá trị thiếu trong mỗi cột.
- **Xử lý giá trị thiếu:**
  - Các cột số (`Tuổi`, `Dung Tích Phổi`, `Số Lượt Khám Bệnh`) được điền giá trị thiếu bằng trung vị (`median`) của cột đó.
  - Các cột phân loại (`Giới Tính`, `Tình Trạng Hút Thuốc`, `Loại Bệnh`, `Loại Điều Trị`, `Hồi Phục`) được điền giá trị thiếu bằng mode (giá trị xuất hiện nhiều nhất).
- **Dịch sang tiếng Việt:**
  - Đổi tên cột từ tiếng Anh sang tiếng Việt bằng từ điển `column_map` và phương thức `df.rename()`.
  - Áp dụng ánh xạ giá trị từ tiếng Anh sang tiếng Việt (ví dụ: "Yes"thành "Có", "No"thành "Không", "COPD"thành "Bệnh Phổi Tắc Nghẽn Mãn Tính") bằng từ điển `value_map` và phương thức `df.replace()`.
- **Lưu dữ liệu đã xử lý:** Dữ liệu sau khi tiền xử lý được lưu vào tệp `lung_disease_data_cleaned.csv` để sử dụng trong dashboard.

4.2 Kết quả sau tiền xử lý

Sau khi thực hiện các bước trên, dữ liệu không còn giá trị thiếu và được chuyển hoàn toàn sang tiếng Việt, đảm bảo tính nhất quán và dễ hiểu cho người dùng tại Việt Nam. Tệp dữ liệu đã được chuẩn hóa sẵn sàng để đưa vào dashboard Streamlit cho các phân tích tiếp theo.

4.3 Xử lý trên Dashboard

Sau khi người dùng tải tệp CSV đã tiền xử lý lên, dashboard sẽ tiến hành xử lý dữ liệu để đảm bảo tính nhất quán và chính xác. Các bước tiền xử lý bao gồm:

- **Xử lý dữ liệu trống:** Các giá trị thiếu được xác định bằng các ký hiệu như `"None"`, `""`, `"UNKNOWN"`, `-1`, `999`, `"NA"`, `"N/A"`, `"NULL"`, `""` và được chuyển thành `NaN`. Sau đó, các giá trị này được xử lý phù hợp (loại bỏ hoặc thay thế) trong từng phân tích cụ thể.
- **Chuẩn hóa định dạng dữ liệu:**
  - Các cột số như `Tuổi`, `Dung Tích Phổi`, `Số Lượt Khám Bệnh` được chuyển thành kiểu `numeric` bằng `pd.to_numeric`.
  - Cột `Hồi Phục` được chuẩn hóa thành `0` (Không) hoặc `1` (Có) từ các giá trị như `"Có - Yes"`, `"Không - No"`.
  - Các cột phân loại như `Giới Tính`, `Tình Trạng Hút Thuốc`, `Loại Bệnh`, `Loại Điều Trị` được chuyển thành kiểu `category`.
- **Đổi tên cột:** Các cột được đổi tên sang tiếng Việt (ví dụ: `Age` thành `Tuổi`, `Gender` thành `Giới Tính`) để phù hợp với giao diện người dùng.

5 Cấu trúc của dashboard

Dashboard được chia thành 4 trang chính:

5.1 Trang 1: Giới Thiệu Dữ Liệu

Hình 1: Trang Giới thiệu dữ liệu.

- **Nguồn gốc dữ liệu:** Giới thiệu về nguồn dữ liệu từ nền tảng **Kaggle**, bao gồm các thông tin về bệnh nhân mắc các bệnh phổi như hen suyễn, viêm phế quản, COPD, ung thư phổi, và viêm phổi.
- **Mô tả dữ liệu:** Mô tả các cột dữ liệu như `Tuổi`, `Giới Tính`, `Tình Trạng Hút Thuốc`, `Dung Tích Phổi`, `Loại Bệnh`, `Loại Điều Trị`, `Số Lượt Khám Bệnh`, và `Hồi Phục`.
- **Xem dữ liệu thô (raw):** Hiển thị bảng dữ liệu đã được dịch sang tiếng Việt dưới dạng tùy chọn mở rộng.

Ví dụ dữ liệu: Dưới đây là một số dòng dữ liệu mẫu từ tập dữ liệu:

Tuổi	Giới Tính	Tình Trạng Hút Thuốc	Dung Tích Phổi	Loại Bệnh...
71	Nữ	Không	4.49	Tắc Nghẽn Mãn Tính...
34	Nữ	Có	3.48	Viêm Phế Quản...
80	Nam	Có	1.95	Tắc Nghẽn Mãn Tính...
22	Nữ	Không	3.65	Viêm Phế Quản...
72	Nam	Có	2.61	Ung Thư Phổi...

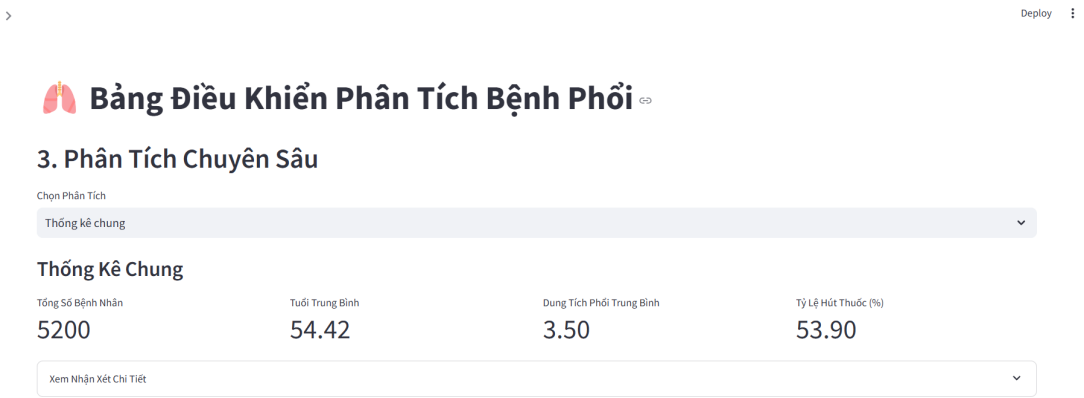
5.2 Trang 2: Thống Kê Mô Tả

	Tuổi	Dung Tích Phổi	Số Lượt Khám Bệnh	Hồi Phục
count	5200	5200	5200	5200
mean	54.4238	3.5006	7.5558	0.5369
std	19.5378	1.4184	3.8809	0.4987
min	20	1	1	0
25%	38	2.29	4	0
50%	54	3.48	8	1
75%	71	4.73	11	1
max	89	6	14	1

Hình 2: Trang Thống kê mô tả.

- **Thống kê cơ bản:** Hiển thị bảng thống kê mô tả (mean, min, max, v.v.) bằng `df.describe()`.
- **Phân phối dữ liệu:**
  - Người dùng chọn biến số (numeric) để xem histogram với đường KDE.
  - Người dùng chọn biến phân loại (categorical) để xem biểu đồ cột.
- **Nhận xét:** Cung cấp nhận xét về tuổi trung bình, dung tích phổi, số lượt khám bệnh, và tỷ lệ hồi phục.

5.3 Trang 3: Phân Tích Chuyên Sâu

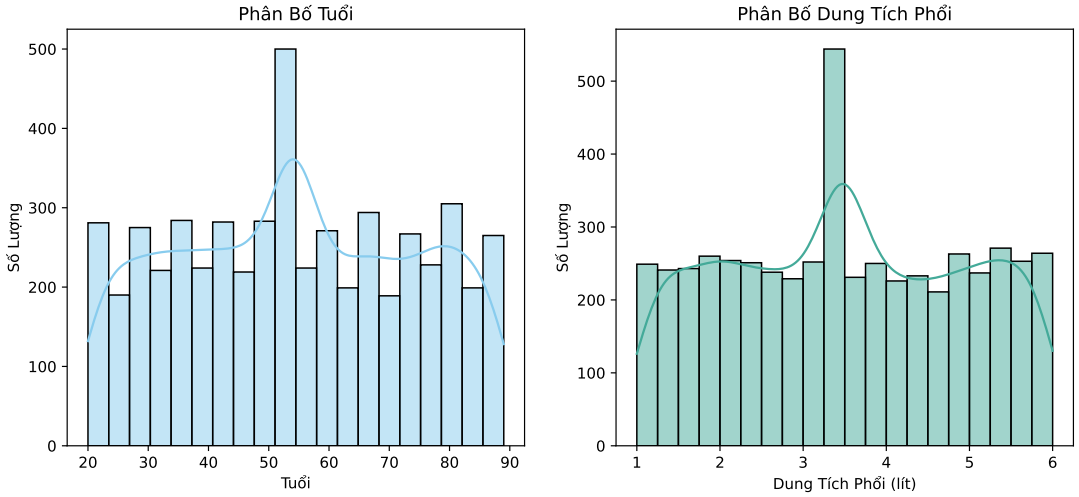


Hình 3: Trang Phân tích chuyên sâu.

- **Thống kê chung:** Hiển thị các chỉ số như tổng số bệnh nhân, tuổi trung bình, dung tích phổi trung bình, và tỷ lệ hút thuốc.

- Tổng số bệnh nhân: 5200
- Tuổi trung bình: 54.42
- Dung tích phổi trung bình: 3.50
- Tỷ lệ hút thuốc: 53.9
- **Nhận xét:** Các chỉ số này cung cấp cái nhìn tổng quan về đặc điểm của tập dữ liệu. Tuổi trung bình và dung tích phổi trung bình giúp hiểu rõ hơn về nhóm bệnh nhân được nghiên cứu. Tỷ lệ hút thuốc là một yếu tố quan trọng cần được theo dõi.

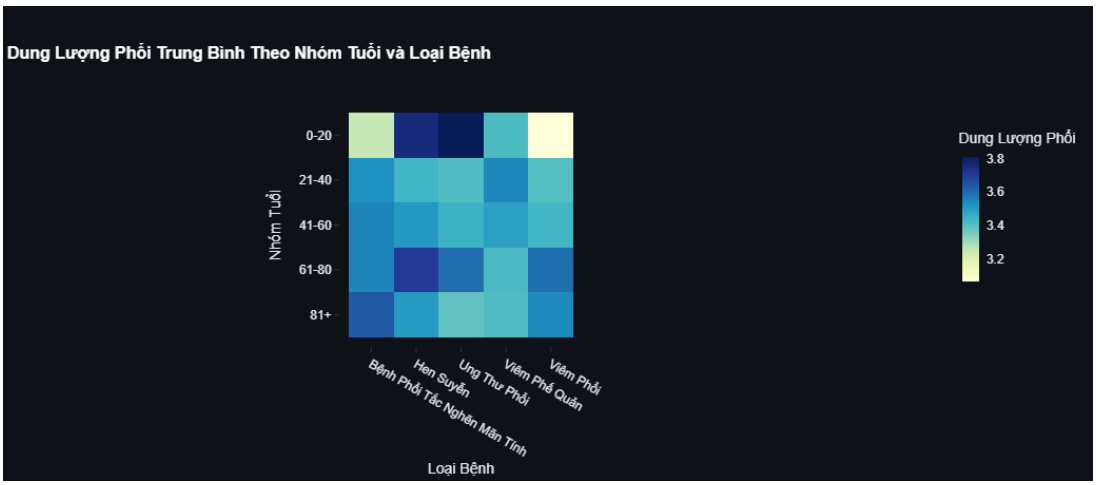
- **Tuổi & Dung Tích Phổi:** Histogram phân bố tuổi và dung tích phổi với thanh trượt chọn khoảng tuổi.



Hình 4: Biểu đồ phân bố tuổi và dung tích phổi.

- **Nhận xét:** Biểu đồ histogram cho thấy sự phân bố của tuổi và dung tích phổi. Tuổi có xu hướng phân bố đều hoặc lệch phải, trong khi dung tích phổi có thể có phân bố chuẩn hoặc lệch trái. Điều này cho thấy mối quan hệ giữa tuổi và dung tích phổi, với xu hướng dung tích phổi giảm khi tuổi tăng.

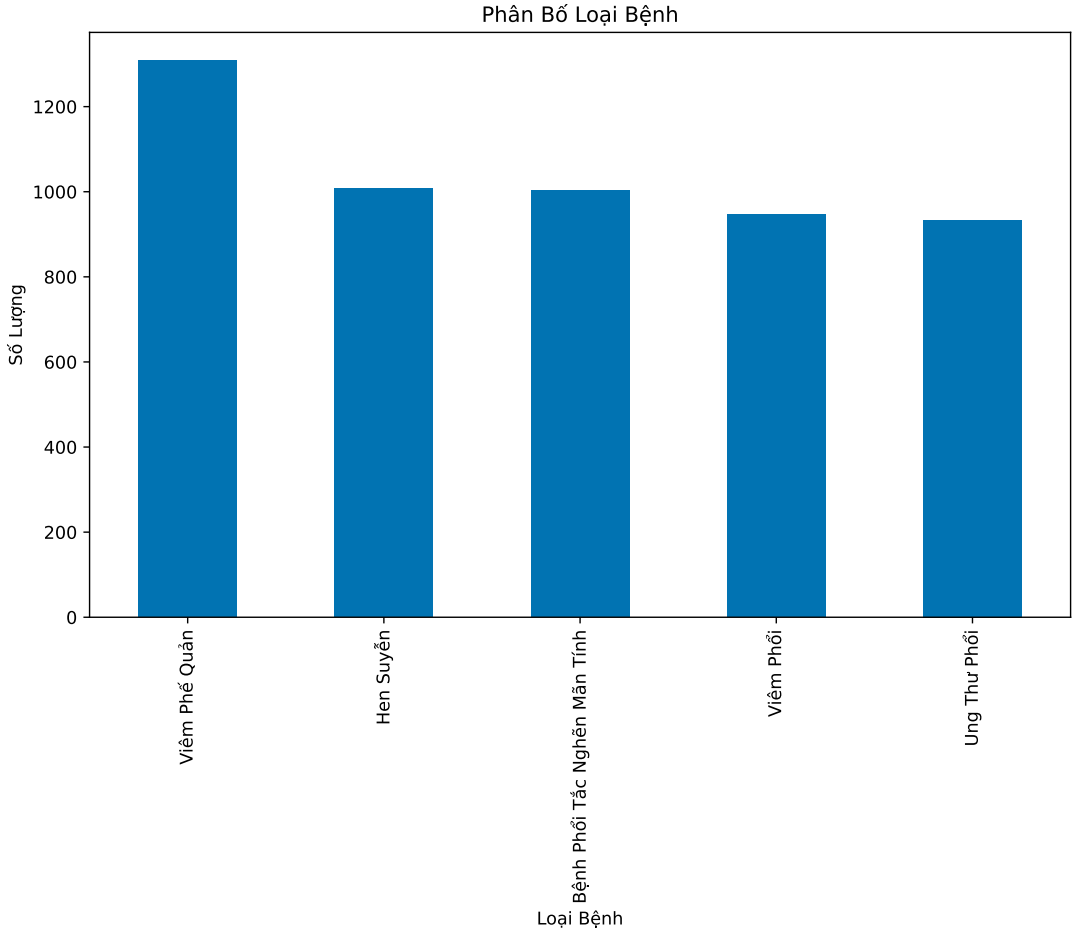
- **Dung Lượng Phổi Trung Bình Theo Nhóm Tuổi và Loại Bệnh:** Heatmap tương tác (Plotly) hiển thị dung lượng phổi trung bình theo nhóm tuổi và loại bệnh.



Hình 5: Heatmap Dung Lượng Phổi Trung Bình Theo Nhóm Tuổi và Loại Bệnh.

  - **Nhận xét:** Nhóm 0-20 có dung lượng cao nhất 3.8 ở Ung Thư Phổi và Hen Suyễn, thấp nhất 3.1 ở Viêm Phổi; nhóm 21-40 dao động 3.2-3.6, với Viêm Phế Quản cao nhất; nhóm 81+ thấp nhất 3.38 ở Ung Thư Phổi, cao nhất 3.63 ở Bệnh Phổi tắc nghẽn mãn tính. Kết luận: Dung lượng phổi giảm theo tuổi, với Ung Thư Phổi và Hen Suyễn bất thường cao ở trẻ, Viêm Phế Quản ổn định ở người cao tuổi.

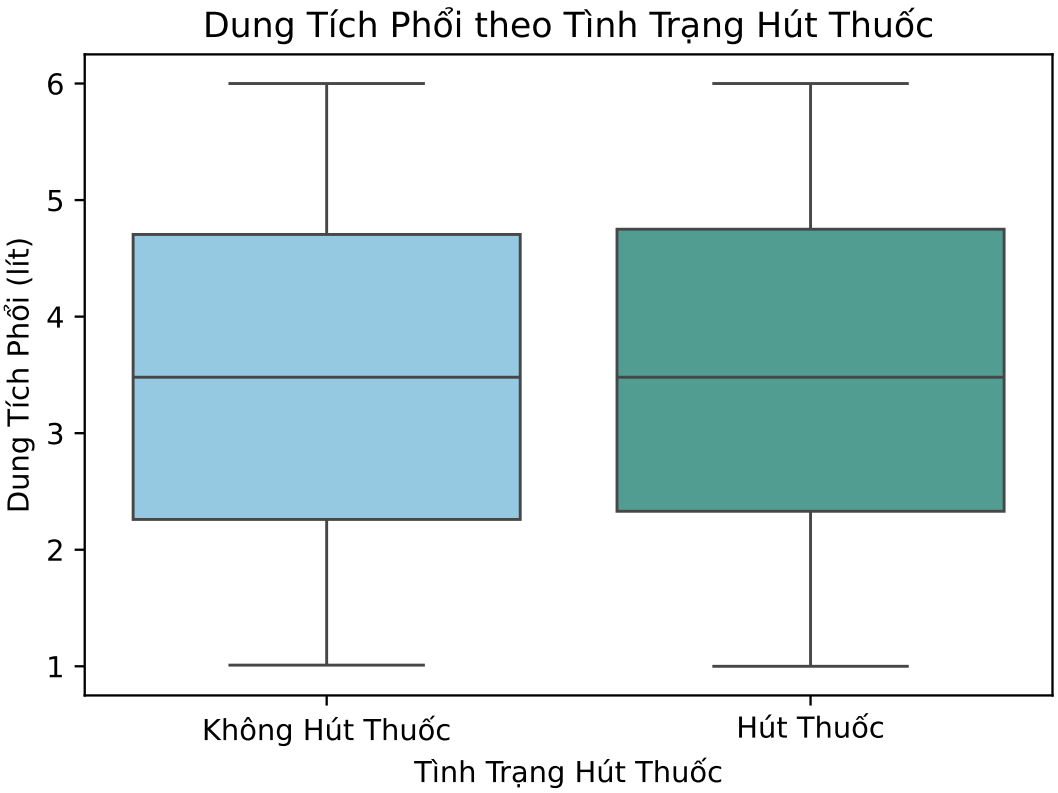
- **Loại Bệnh:** Biểu đồ cột thể hiện số lượng bệnh nhân theo từng loại bệnh.



Hình 6: Phân bố các loại bệnh phổi.

- **Nhận xét:** Phân tích dữ liệu từ biểu đồ cho thấy Viêm Phế Quản là bệnh lý hô hấp nổi trội với khoảng 1,300 ca (chiếm khoảng 25 % tổng số), vượt trội hơn 30% so với Ung Thư Phổi và Viêm Phổi (mỗi loại khoảng 950 ca). Bệnh Phổi Tắc Nghẽn và Hen Suyễn đứng ở mức trung bình với khoảng 1,000 ca cho mỗi loại. Sự phân bố này phản ánh gánh nặng đáng kể của các bệnh đường hô hấp, đặc biệt là Viêm Phế Quản, đòi hỏi các chiến lược y tế công cộng phù hợp với tỷ lệ mắc bệnh. Đáng chú ý là mặc dù Ung Thư Phổi có số ca thấp hơn, nhưng vẫn chiếm tỷ lệ đáng kể (khoảng 19%), phản ánh nhu cầu cấp thiết về các biện pháp sàng lọc và phát hiện sớm.

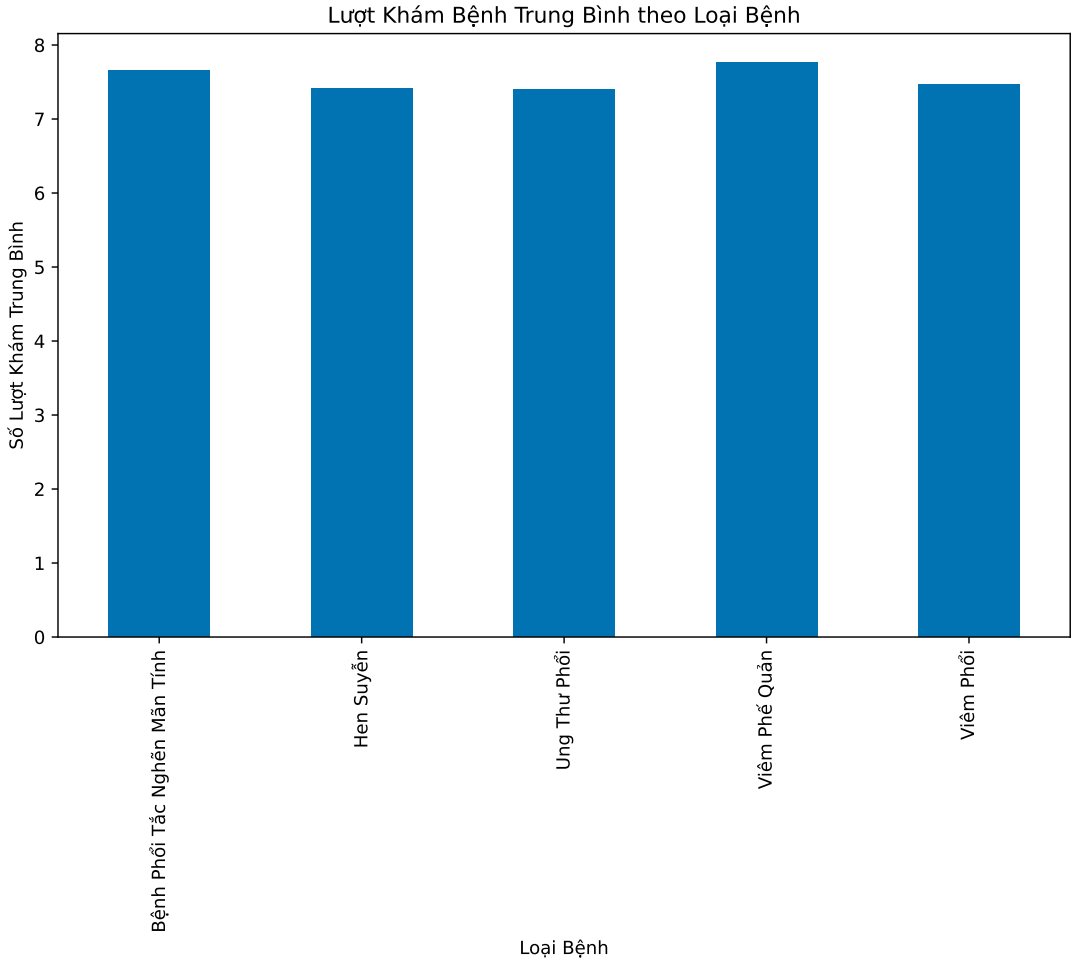
- **Hút Thuốc & Dung Tích Phổi:** Boxplot và scatter plot so sánh dung tích phổi giữa người hút thuốc và không hút thuốc, kèm kiểm định t-test.



Hình 7: Dung tích phổi theo tình trạng hút thuốc.

- **Nhận xét:** Trung vị dung tích phổi của cả hai nhóm đều nằm ở khoảng 3.5 lí. Khoảng tứ phân vị (IQR) của nhóm hút thuốc có vẻ tương đương với nhóm không hút thuốc. t-statistic = -1.26 cho thấy sự khác biệt nhỏ giữa hai nhóm (giá trị âm gợi ý dung tích phổi nhóm hút thuốc thấp hơn nhóm không hút) p-value = 1.793 lớn hơn ngưỡng ý nghĩa thống kê thông thường (0.05), điều này chỉ ra rằng không có bằng chứng đủ mạnh để kết luận có sự khác biệt về dung tích phổi giữa người hút và không hút thuốc. Các biểu đồ này chỉ ra rằng các yếu tố khác ngoài tình trạng hút thuốc có thể đóng vai trò quan trọng hơn trong việc quyết định dung tích phổi, hoặc có thể cần mẫu lớn hơn để có thể phân tích được.

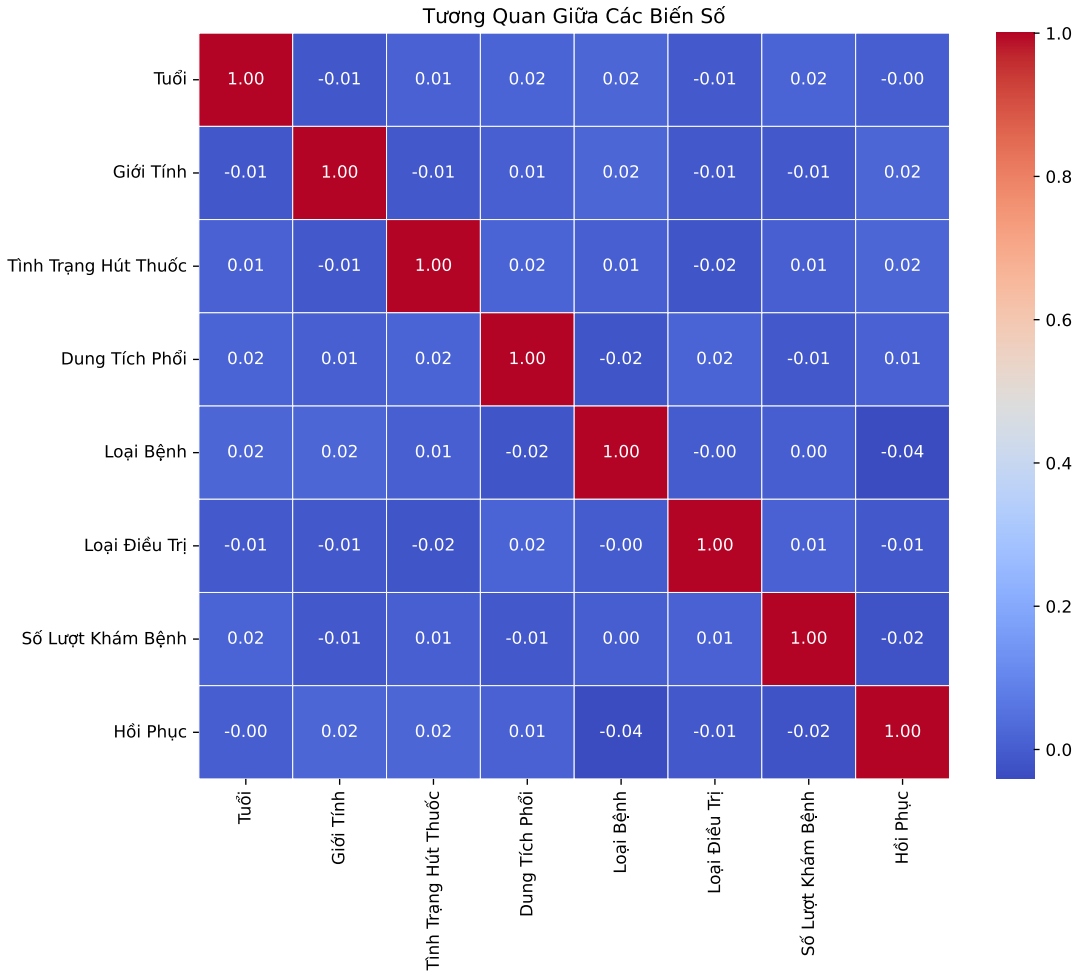
- **Lượt Khám Bệnh:** Biểu đồ cột hiển thị số lượt khám trung bình theo loại bệnh (lựa chọn nhiều loại bệnh).



Hình 8: Biểu đồ lượt khám bệnh trung bình theo loại bệnh.

  - **Nhận xét:** Biểu đồ cho thấy Viêm Phế Quản có dung lượng phổi trung bình cao nhất (7.76), trong khi Ung Thư Phổi thấp nhất (7.39), phản ánh mức độ tác động khác nhau của các loại bệnh lên chức năng hô hấp. Hen Suyễn và Ung Thư Phổi là hai bệnh gây ảnh hưởng nghiêm trọng nhất, cần được quản lý chặt chẽ. Các giá trị tương đối gần nhau (7.39-7.76) cho thấy sự đồng đều trong mẫu dữ liệu, nhưng vẫn nhấn mạnh tầm quan trọng của việc can thiệp sớm để bảo vệ chức năng phổi, đặc biệt ở bệnh nhân ung thư phổi và hen suyễn.

- **Tương Quan:** Heatmap thể hiện tương quan giữa các biến số (sau khi mã hóa dữ liệu phân loại).



Hình 9: Heatmap tương quan giữa các biến.

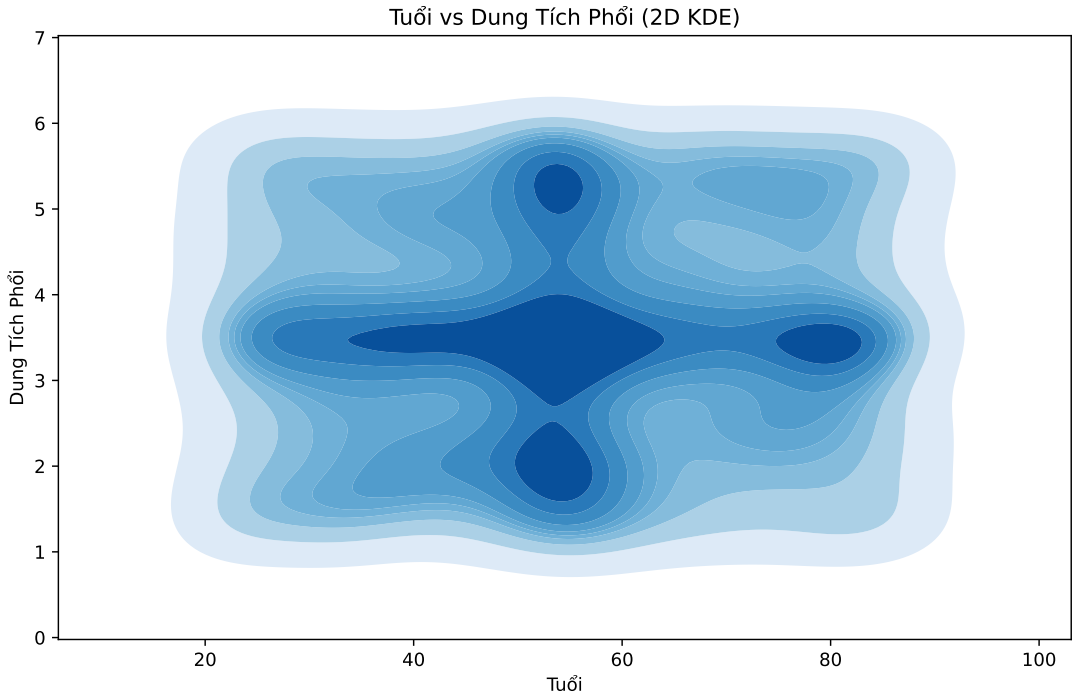
– **Nhận xét:**

∗ Tương quan yếu (gần 0): Hầu hết các cặp biến số khác (ví dụ: Tuổi-Giới Tính, Tuổi-Dung Tích Phổi, Loại Bệnh-Số Lượt Khám) có tương quan rất thấp (0.00 đến 0.02 hoặc -0.04), cho thấy không có mối liên hệ đáng kể giữa các yếu tố này.

∗ Tương quan âm yếu: Một số cặp như Dung Tích Phổi-Loại Bệnh (-0.04) và Hồi Phục-Loại Điều Trị (-0.02) có tương quan âm nhẹ, nhưng không đủ mạnh để kết luận ảnh hưởng rõ rệt.

∗ **Kết luận:** Dữ liệu cho thấy các biến số độc lập với nhau, ngoại trừ sự tự tương quan trong cùng một biến. Cần phân tích sâu hơn để xác định tác động tiềm ẩn, đặc biệt với các yếu tố như Dung Tích Phổi và Loại Bệnh.

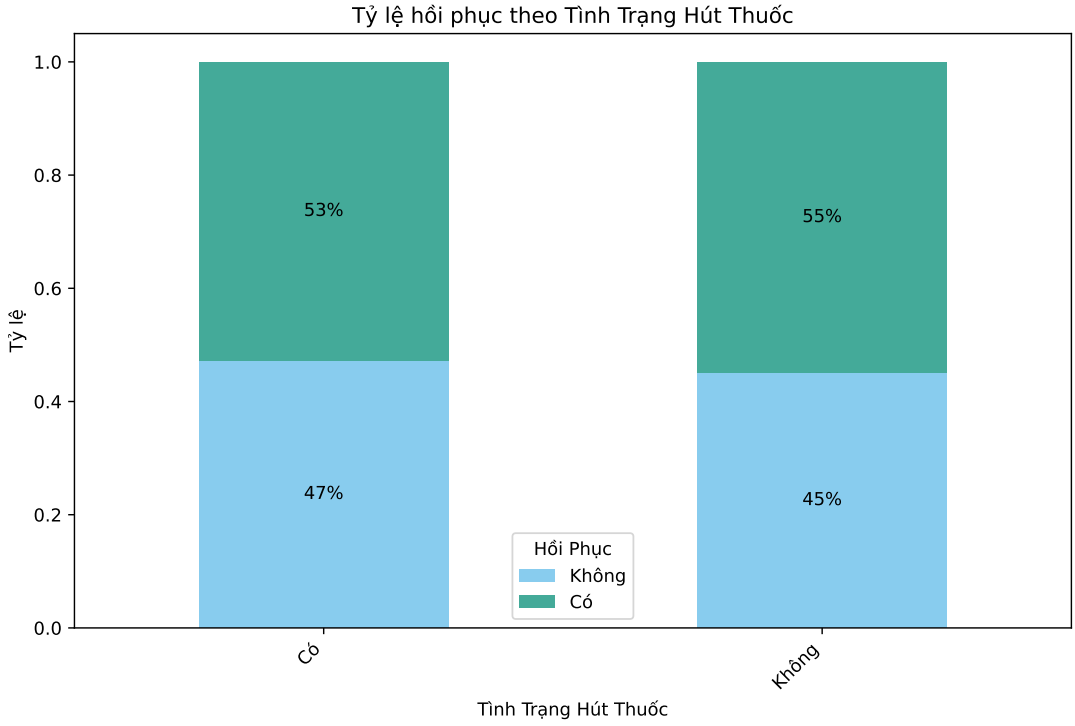
• **Phân Tích Song Biến:** Biểu đồ phân tán hoặc KDE 2D giữa hai biến số do người dùng chọn.



Hình 10: 2D KDE giữa Tuổi và Dung tích phổi.

  - Mối Quan Hệ giữa Dung Tích Phổi và Tuổi: Khi tuổi tăng, dung tích phổi có xu hướng giảm.
  - Tương Quan giữa Hút Thuốc và Dung Tích Phổi: Những người hút thuốc có xu hướng có dung tích phổi thấp hơn.
  - Tỷ Lệ Hồi Phục và Số Lượt Khám Bệnh: Theo dõi y tế tốt hơn giúp tăng cơ hội phục hồi.
  - Tác Động của Loại Bệnh lên Hồi Phục: Bệnh mãn tính có tỷ lệ hồi phục thấp hơn.
  - Ảnh Hưởng của Loại Điều Trị: Phẫu thuật hoặc liệu pháp có xu hướng có tương quan tích cực với hồi phục.

- **Tỷ lệ hồi phục:** Biểu đồ cột xếp chồng thể hiện tỷ lệ hồi phục theo các yếu tố như hút thuốc, loại bệnh, hoặc loại điều trị.



Hình 11: Tỉ lệ hồi phục theo tình trạng hút thuốc.

– **Nhận xét:**

  * Nhóm Không hút thuốc ghi nhận tỷ lệ hồi phục là 55%, trong khi tỷ lệ không hồi phục là 45%. Ngược lại, nhóm Có hút thuốc có tỷ lệ hồi phục là 53% và không hồi phục là 47%.
  * Khoảng cách giữa hai nhóm chỉ là 2%, cho thấy sự khác biệt nhỏ về tỷ lệ hồi phục liên quan đến tình trạng hút thuốc.
  * **Kết luận:** Dữ liệu cho thấy có sự khác biệt nhẹ về tỷ lệ hồi phục giữa hai nhóm, nhưng không đủ lớn để khẳng định hút thuốc là yếu tố quyết định. Cần nghiên cứu thêm để đánh giá chính xác tác động.

5.4 Trang 4: Nhận Xét Chung

Hình 12: Trang Nhận xét chung.

Bao gồm các nội dung:

- **Tổng quan:** Tóm tắt các phát hiện chính như mối quan hệ giữa hút thuốc và dung tích phổi, tỷ lệ hồi phục, và phân bố loại bệnh.
- **Hạn chế:** Đề cập đến kích thước mẫu, dữ liệu thiếu, và đơn vị đo chưa chuẩn hóa.

6 Nhận xét Đồ án

6.1 Kết luận

Đồ án đã cung cấp một dashboard trực quan hóa dữ liệu bệnh phổi hoàn thiện, giúp người dùng dễ dàng khám phá các xu hướng và mối quan hệ giữa các yếu tố như tuổi, giới tính, tình trạng hút thuốc, và kết quả hồi phục. Bằng cách kết hợp các biểu đồ đa dạng và phân tích chuyên sâu như t-test hay heatmap tương quan, nhóm đã mang đến cái nhìn sâu sắc về dữ liệu bệnh phổi.

7 Mức Độ Hoàn Thành Của Các Thành Viên

Dưới đây là bảng tổng quan về mức độ hoàn thành công việc của từng thành viên trong nhóm:

Công Việc	Mức Độ Hoàn Thành
Thiết kế giao diện Streamlit	100%
Tiền xử lý dữ liệu	100%
Phân tích thống kê mô tả	100%
Trực quan hóa phân tích chuyên sâu	100%
Viết báo cáo và nhận xét	100%

Kiểm tra và tối ưu hóa mã nguồn	100%