

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN HỆ THỐNG THÔNG TIN**

**NHẬP MÔN LẬP TRÌNH PYTHON**

🙠🙟🕮🙝🙢



Đề Tài:

**PHÂN TÍCH DỮ LIỆU**

**DỰ ĐOÁN SỐNG SÓT TITANIC**

**GVHD**: Th.s Trần Quang Khải

**SVTH:**

1. Võ Văn Tú 24810046
2. Đỗ Bảo Ân 24810022
3. Mai Da Duy 24810028

Thành phố Hồ Chí Minh, Tháng 5 năm 2025

***Nhận xét của giảng viên***

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*...........................................................................................................................................*

*........................................................................................................................................... ...........................................................................................................................................*

*TP. Hồ Chí Minh, ngày … tháng… năm 2024*

*Giảng viên ký tên*

**MỤC LỤC**

[PHẦN 1: MỞ ĐẦU 1](#_Toc199435887)

[1. Lý do chọn đề tài 1](#_Toc199435888)

[2. Mục tiêu đề tài 1](#_Toc199435889)

[3. Đối tượng và phạm vi nghiên cứu 2](#_Toc199435890)

[PHẦN 2. NỘI DUNG 2](#_Toc199435891)

[CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT 2](#_Toc199435892)

[1.1. Giới thiệu về ngôn ngữ Python 2](#_Toc199435893)

[1.2. Tại sao Python thường được dùng để phân tích dữ liệu 3](#_Toc199435894)

[1.3. Giới thiệu về các thư viện thường được sử dụng trong Phân tích dữ liệu 3](#_Toc199435895)

[1.3.1. NumPy (Numerical Python) 3](#_Toc199435896)

[1.3.2. Pandas 4](#_Toc199435897)

[1.3.3. Matplotlib 4](#_Toc199435898)

[1.3.4. Missingno 5](#_Toc199435899)

[1.3.5. TTKBootstrap 5](#_Toc199435900)

[CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU DỰ ĐOÁN SỐNG SÓT TITANIC 7](#_Toc199435901)

[2.1. Quá trình thu thập dữ liệu 7](#_Toc199435902)

[2.2. Tổng quan về dữ liệu 7](#_Toc199435903)

[2.3. Các hạn chế và đặc điểm của dữ liệu 8](#_Toc199435904)

[CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN SỐNG SÓT TITANIC 9](#_Toc199435905)

[3.1. Xử lý dữ liệu 9](#_Toc199435906)

[3.1.1. Kiểm tra dữ liệu ban đầu 9](#_Toc199435907)

[3.1.2. Xử lý các giá trị bị thiếu 9](#_Toc199435908)

[3.1.3. Loại bỏ các giá trị trùng lặp 10](#_Toc199435909)

[3.1.4. Kiểm tra và xác thực dữ liệu hợp lệ 10](#_Toc199435910)

[3.1.5. Lưu dữ liệu đã làm sạch 10](#_Toc199435911)

[3.1.6. Tổng hợp 10](#_Toc199435912)

[3.2. Trực quan hóa dữ liệu 11](#_Toc199435913)

[3.3. Thiết kế giao diện 16](#_Toc199435914)

[3.4. Các chức năng chính của chương trình 18](#_Toc199435915)

[3.4.1. Thêm trong Quản lý hành khách 18](#_Toc199435916)

[3.4.2. Đọc bản ghi trong Quản lý hành khách 19](#_Toc199435917)

[3.4.4. Cập nhật dữ liệu hành khách trong Quản lý hành khách 19](#_Toc199435918)

[3.4.5. Xóa bản ghi trong Quản lý hành khách 20](#_Toc199435919)

[3.4.6. Chức năng Xóa form 20](#_Toc199435920)

[3.4.6. Chức năng lọc dữ liệu 21](#_Toc199435921)

[3.4.7. Chức năng tìm kiếm dữ liệu 22](#_Toc199435922)

[3.4.8. Chức năng phân trang 23](#_Toc199435925)

[PHẦN KẾT LUẬN 25](#_Toc199435926)

[1. Kết quả đạt được 25](#_Toc199435927)

[2. Hướng phát triển 26](#_Toc199435928)

[PHẦN TÀI LIỆU THAM KHẢO 26](#_Toc199435929)

**DANH MỤC HÌNH ẢNH**

[Hình 3. 1 Biểu đồ tỉ lệ sống sót và gặp nạn 11](#_Toc199438952)

[Hình 3. 2 Biểu đồ sự tương quan giữa sống sót với các đặc trưng 12](#_Toc199438953)

[Hình 3. 3 Biểu đồ tương quan giữa sống sót với đặc trưng gia đình 13](#_Toc199438954)

[Hình 3. 4 Biểu đồ biểu diễn sống sót theo tuổi 14](#_Toc199438955)

[Hình 3. 5 Biểu đồ phân bố tuổi theo hạng vé và giới tính 15](#_Toc199438956)

[Hình 3. 6 Biểu đồ số lượng hành khách qua giới tính, cảng biển và hạng vé 16](#_Toc199438957)

[Hình 3. 7 Giao diện người dùng 17](#_Toc199438958)

[Hình 3. 8 Giao diện Biểu đồ 18](#_Toc199438959)

[Hình 3. 9 Chức năng Thêm hành khách mới 19](#_Toc199438960)

[Hình 3. 10 Chức năng Đọc bản ghi 19](#_Toc199438961)

[Hình 3. 11 Chức năng Cập nhật dữ liệu cho hành khách 20](#_Toc199438962)

[Hình 3. 12 Chức năng Xóa một bản ghi của hành khách 20](#_Toc199438963)

[Hình 3. 13 Chức năng Xóa form 21](#_Toc199438964)

[Hình 3. 14 Lọc dữ liệu theo Giới tính Nam 21](#_Toc199438965)

[Hình 3. 15 Lọc dữ liệu theo Hạng vé (Pclass) 22](#_Toc199438966)

[Hình 3. 16 Lọc dữ liệu theo giới tính (male ) và Hạng vé ( hạng 3) 22](#_Toc199438967)

[Hình 3. 17 Tìm kiếm dữ liệu dựa trên PassengerId 23](#_Toc199438968)

[Hình 3. 18 Tìm kiếm dữ liệu dưa trên Name 23](#_Toc199438969)

[Hình 3. 19 Tìm kiếm dữ liệu dưa trên Name kết hợp bộ lọc 23](#_Toc199438970)

[Hình 3. 20 Chức năng phân trang 24](#_Toc199438971)

**KẾ HOẠCH PHÂN CÔNG NHIỆM VỤ THỰC HIỆN ĐỀ TÀI CUỐI KỲ MÔN LẬP TRÌNH PYTHON**

**HỌC KỲ I NĂM HỌC 2024-2025**

1. **Mã lớp môn học:** 2427IPPA233277\_04
2. **Giảng viên hướng dẫn:** ThS. Trần Quang Khải
3. **Tên đề tài: Phân tích dữ liệu tàu Titanic**
4. **Bảng phân công nhiệm vụ:**

|  |  |
| --- | --- |
| **Sinh viên thực hiện** | **Nội dung thực hiện** |
| Võ Văn Tú | ❖ Phân tích, trực quan hóa dữ liệu, tìm kiếm |
| Đỗ Bảo Ân | ❖ Thiết kế giao diện, crud dữ liệu, phân trang, lọc dữ liệu |
| Mai Gia Duy | ❖ Làm sạch dữ liệu, thiết kế sidebar |

**LỜI CẢM ƠN**

Để thực hiện được đề tài này, em xin gửi lời cảm ơn chân thành đến thầy Trần

Quang Khải - giảng viên bộ môn Lập trình Python của lớp 2427IPPA233277\_04 vì đã trang bị cho chúng em đầy đủ kiến thức và chỉ dẫn, hỗ trợ chúng em hết mình trong quá trình học cũng như thực hiện đồ án cuối kỳ.

Với đề tài nghiên cứu lần này nhóm chúng em đã có rất nhiều cố gắng, dành nhiều thời gian nhất có thể để đồng hành cùng nhau vận dụng các kiến thức đã học để giải quyết những vấn đề được đặt ra với dữ liệu. Tuy nhiên với thời gian có hạn kèm theo sự phức tạp về mặt dữ liệu cũng như sự non nớt trong kinh nghiệm thu thập, phân tích , giải quyết vấn đề, nên không thể tránh khỏi sự thiếu xót hay những điểm sai không mong muốn, kính mong nhận được sự góp ý, động viên và khích lệ từ Thầy cũng như các bạn.

Với lòng biết ơn sâu sắc, một lần nữa chúng em xin trân trọng cảm ơn sự quan tâm giúp đỡ của Thầy trong suốt hành trình vừa qua.

# PHẦN 1: MỞ ĐẦU

## 1. Lý do chọn đề tài

Trong triết học có một câu rất hay là: "Mọi sự vật, hiện tượng trong thế giới đều tồn tại trong mối liên hệ, ràng buộc và tác động lẫn nhau. Không có gì tồn tại một cách biệt lập hay không phụ thuộc vào các yếu tố khác." Để tìm ra những mối liên hệ đó chúng ta cần đặt chúng trong các mối quan hệ với nhau để nghiền ngẫm, phân tích, xử lý dữ liệu một cách kỹ lưỡng và chính xác. Quá trình này ngày càng trở nên quan trọng cho bất kỳ một cá nhân, tổ chức nào khi muốn đưa ra một quyết định mà quyết định ấy đem về hiệu quả tối ưu nhất. Để thực hành khả năng phân tích xử lý dữ liệu nhóm chúng em chọn tập dữ liệu hành khác trên tàu Titanic. Có thể nói đây là tập dữ liệu kinh điển trong khoa học dữ liệu sánh ngang với “hello world” trong các ngôn ngữ lập trình, nó dễ dàng tiếp cận với tất cả các trình độ từ sơ cấp đến kỳ cựu, nó cung cấp đa dạng các loại dữ liệu (dữ lệu phân loại, số, hỗn hợp lẫn dữ liệu khuyết,…) phù hù hợp để thực hiện đầy đủ các yêu cầu làm sạch, phân loại, thống kê và cả trực quan hóa dữ liệu… Hơn thế tập dữ liệu Titanic chính là bài toán thực tế về tối ưu hóa việc đưa ra quyết định, lựa chọn, nhằm tăng khả năng sống xót khi xảy ra thảm họa, đồng thời còn có thể mở rộng để phát triển mô hình học máy dự đoán khả năng sống xót của một người thông qua các yếu tố đã biết….

## 2. Mục tiêu đề tài

Mục tiêu của đề tài là khai thác và phân tích dữ liệu tàu Titanic để đưa ra những thông tin hữu ích về các đặc trưng quan trọng ảnh hưởng đến sự sống xót của một khách. Cụ thể, đề tài hướng đến việc xử lý và tổ chức dữ liệu bằng Python, từ đọc, làm sạch, chuẩn hóa dữ liệu từ các file Excel đến xây dựng một bảng dữ liệu hợp lý, dễ dàng phân tích và trực quan hóa. Tiếp đó, thực hiện phân tích để tìm hiểu mối quan hệ giữa các yếu tố như độ tuổi, giới tính, hạng vé, số lượng thành viên gia đình,... Dữ liệu sẽ được trực quan hóa qua các biểu đồ và báo cáo bằng các công cụ như matplotlib và seaborn để phát hiện xu hướng và mẫu hình quan trọng. Từ đó, đề tài không chỉ đưa ra những thông tin ý nghĩa, mà còn những nhận định có giá trị về đưa ra quyết định nhằm tăng tỉ lệ sống xót trong một số thảm họa đáng tiếc.

## 3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài:

*Dữ liệu hành khách trên tàu Titanic:* Bao gồm các thông tin liên quan đến giới tính, độ tuổi, hạng vé, cabin đã ngồi, giá vé phải trả, cảng biển nơi họ xuất phát, số lượng thành viên gia đình.

Phạm vi nghiên cứu, chúng em sẽ tập trung vào việc:

*Phân tích dữ liệu:* Tiến hành xử lý, tổ chức và phân tích bộ dữ liệu thông tin hành khách trên tàu Titanic.

*Trực quan hóa dữ liệu:* Xây dựng các biểu đồ trực quan để làm nổi bật các đặc trung liên quan đến tỉ lệ sống xót của hành khách.

*Ứng dụng và đánh giá:* Đưa ra các nhận định, kết luận dựa trên phân tích dữ liệu, đồng thời đề xuất các quyết định đúng đắn khi lựa chọn một phương tiện di chuyển cho chuyến đi.

# PHẦN 2. NỘI DUNG

# CHƯƠNG 1: TÓM TẮT CƠ SỞ LÝ THUYẾT

### 1.1. Giới thiệu về ngôn ngữ Python

Được sáng tạo bởi Guido van Rossum và ra mắt vào năm 1991, Python đã trở thành một trong những ngôn ngữ lập trình phổ biến nhất thế giới nhờ sự đơn giản, mạnh mẽ và khả năng ứng dụng rộng rãi. Với cú pháp gần gũi, dễ đọc, Python là lựa chọn lý tưởng cho cả những người mới nhập môn lập trình lẫn các kỹ sư phần mềm dày dạn kinh nghiệm.

Là một nền tảng mã nguồn mở, Python có thể hoạt động trơn tru trên mọi hệ điều hành phổ biến như Windows, macOS hay Linux. Sức mạnh thực sự của Python nằm ở tính đa năng – nó được ứng dụng trong vô số lĩnh vực: từ xây dựng website, phát triển AI, phân tích dữ liệu lớn (Big Data), đến tự động hóa tác vụ và nghiên cứu khoa học.

Phiên bản Python 3.x hiện đang thống trị thị trường nhờ những cải tiến vượt trội về hiệu năng và bảo mật. Đặc biệt, Python sở hữu kho thư viện khổng lồ như NumPy (tính toán khoa học), Pandas (xử lý dữ liệu), TensorFlow (học sâu) và Django (phát triển web), biến nó thành công cụ không thể thiếu trong các dự án công nghệ hiện đại. Không ngạc nhiên khi Python liên tục giữ vị trí "ông vua" trong làng lập trình, đặc biệt trong kỷ nguyên bùng nổ của trí tuệ nhân tạo và khoa học dữ liệu.

### 1.2. Tại sao Python thường được dùng để phân tích dữ liệu

Python không phải ngẫu nhiên mà trở thành ngôn ngữ hàng đầu trong phân tích dữ liệu. Sức hút của nó nằm ở sự dễ học, dễ dùng nhưng vẫn cực kỳ mạnh mẽ nhờ hàng loạt thư viện chuyên biệt:

* Pandas – Xử lý dữ liệu dạng bảng (DataFrame) nhanh như Excel trên steroid.
* NumPy – Tính toán số học siêu tốc, xử lý ma trận và đại lượng lớn dữ liệu.
* Missingno – là một thư viện Python cực kỳ hữu ích để trực quan hóa và phân tích dữ liệu bị thiếu (missing data)
* Matplotlib – Vẽ biểu đồ đẹp, trực quan chỉ với vài dòng code.
* Ttkbootstrap – Thư viện mở rộng của tkinter, cung cấp giao diện đẹp hơn với nhiều theme hiện đại

Không chỉ thế, Python còn "nuốt trôi" mọi định dạng dữ liệu – từ file CSV, Excel đến cơ sở dữ liệu SQL hay JSON. Nó cũng tích hợp tốt với các công cụ Big Data như Spark, Hadoop, giúp xử lý lượng dữ liệu khổng lồ một cách hiệu quả.

Python cực kỳ linh hoạt – bạn có thể dùng nó để phân tích dữ liệu, rồi ngay lập tức xây dựng web (Django/Flask), triển khai AI (TensorFlow/PyTorch), hay tự động hóa báo cáo. Hơn nữa, cộng đồng lớn mạnh và tài liệu phong phú giúp Python trở thành công cụ lý tưởng cho cả người mới bắt đầu và các chuyên gia trong lĩnh vực phân tích dữ liệu. Những ưu điểm này đã biến Python thành một công cụ không thể thiếu, giúp các nhà phân tích nhanh chóng xử lý, trực quan hóa và đưa ra những nhận định chính xác dựa trên dữ liệu.

### 1.3. Giới thiệu về các thư viện thường được sử dụng trong Phân tích dữ liệu

Python cung cấp một hệ sinh thái phong phú với nhiều thư viện mạnh mẽ, hỗ trợ toàn diện cho việc xử lý, phân tích và trực quan hóa dữ liệu. Dưới đây là những thư viện thường được sử dụng:

#### 1.3.1. NumPy (Numerical Python)

NumPy là một thư viện cốt lõi trong Python, cung cấp các công cụ hiệu suất cao để thao tác với mảng đa chiều (ndarray) và thực hiện các phép toán số học.

*Xử lý mảng hiệu suất cao:* NumPy hỗ trợ các phép toán vector hóa, cho phép áp dụng cùng một phép tính cho toàn bộ mảng, tăng tốc đáng kể so với việc sử dụng vòng lặp.

*Phép toán tuyến tính:* Hỗ trợ các phép toán phức tạp như nhân ma trận, nghịch đảo ma trận, và tính định thức, giúp giải quyết các bài toán toán học và khoa học.

*Khả năng tương thích:* NumPy tích hợp tốt với các thư viện khác như pandas, matplotlib, và scikit-learn, tạo điều kiện thuận lợi cho toàn bộ chuỗi phân tích dữ liệu.

Nhờ vào hiệu suất cao và tính linh hoạt, NumPy là nền tảng cho nhiều thư viện khác và là công cụ không thể thiếu trong xử lý dữ liệu lớn.

#### 1.3.2. Pandas

Pandas là một thư viện mạnh mẽ, được xây dựng trên NumPy, chuyên để thao tác và phân tích dữ liệu dạng bảng. Thư viện này được sử dụng rộng rãi nhờ các đặc điểm sau:

*Cấu trúc dữ liệu linh hoạt:* Pandas cung cấp hai cấu trúc chính: Series (dữ liệu một chiều) và DataFrame (dữ liệu hai chiều, tương tự bảng tính Excel).

*Đọc và ghi dữ liệu:* Pandas hỗ trợ đọc và ghi dữ liệu từ nhiều định dạng phổ biến như CSV, Excel, SQL, JSON, giúp dễ dàng làm việc với dữ liệu từ các nguồn khác nhau.

*Làm sạch và chuẩn hóa dữ liệu:* Pandas có các công cụ xử lý dữ liệu mạnh mẽ, giúp loại bỏ các giá trị không hợp lệ, xử lý dữ liệu bị thiếu, và chuẩn hóa dữ liệu một cách nhanh chóng.

*Thao tác dữ liệu hiệu quả:* Pandas hỗ trợ lọc, nhóm, sắp xếp, và tính toán trên dữ liệu, giúp quá trình phân tích dữ liệu trở nên đơn giản và trực quan hơn.

Pandas là một công cụ không thể thiếu khi xử lý và phân tích dữ liệu dạng bảng, đặc biệt trong các dự án khoa học dữ liệu.

#### 1.3.3. Matplotlib

Matplotlib là thư viện cơ bản và phổ biến nhất để trực quan hóa dữ liệu trong Python.

*Đa dạng biểu đồ:* Matplotlib hỗ trợ tạo nhiều loại biểu đồ như biểu đồ cột, biểu đồ đường, biểu đồ phân tán, và biểu đồ tròn.

*Khả năng tùy chỉnh cao:* Người dùng có thể tùy chỉnh màu sắc, kiểu đường, nhãn trục, và các yếu tố khác để tạo ra các biểu đồ chuyên nghiệp.

*Tích hợp với Pandas và NumPy:* Matplotlib có thể trực tiếp làm việc với các DataFrame hoặc mảng, giúp việc trực quan hóa trở nên nhanh chóng.

Matplotlib là công cụ nền tảng cho việc trình bày dữ liệu trực quan, hỗ trợ tốt cho các bài toán phân tích dữ liệu.

#### 1.3.4. Missingno

Thư viện Missingno cung cấp một cách rất hay để trực quan hóa sự phân phối của các giá trị NaN. Missingno là một thư viện Python và tương thích với Pandas.

*Biểu đồ ma trận (Matrix Plot):* Hiển thị sự phân bố của dữ liệu khuyết thiếu dưới dạng ma trận, giúp nhận biết các mẫu dữ liệu bị thiếu theo từng cột.

Màu sắc rõ ràng: Giá trị NaN được đánh dấu màu trắng, trong khi dữ liệu hợp lệ hiển thị màu đen.

*Biểu đồ thanh (Bar Chart):* Thống kê số lượng giá trị khuyết thiếu trong từng cột, giúp đánh giá mức độ nghiêm trọng của việc thiếu dữ liệu.

*Biểu đồ nhiệt (Heatmap):* Phân tích mối tương quan giữa các cột có dữ liệu khuyết thiếu, giúp phát hiện các nhóm cột có xu hướng thiếu dữ liệu cùng nhau*.*

*Biểu đồ cây (Dendrogram):* Nhóm các cột có mô hình khuyết thiếu tương tự nhau, hỗ trợ phân cụm dữ liệu để xử lý hiệu quả hơn*.*

#### 1.3.5. TTKBootstrap

TTKBootstrap - Thư viện tạo giao diện hiện đại cho Tkinter

TTKBootstrap (hay tkinter-bootstrap) là một thư viện Python mở rộng dựa trên Tkinter, giúp xây dựng ứng dụng desktop với giao diện hiện đại, đẹp mắt và dễ sử dụng. Thư viện này cung cấp các theme (chủ đề) được thiết kế sẵn, các widget nâng cao và hỗ trợ tùy chỉnh CSS-like, giúp cải thiện đáng kể trải nghiệm người dùng so với Tkinter thông thường.

*Giao diện đẹp với nhiều theme có sẵn:* Hỗ trợ nhiều theme như "darkly", "solar", "superhero", "minty", "flatly"….

*Widgets nâng cao:* Cung cấp nhiều widget hiện đại hơn Tkinter thông thường

* Meter (thanh tiến trình tròn)
* DateEntry (chọn ngày)
* FlipSwitch (nút bật/tắt kiểu hiện đại)
* Toast (thông báo popup)
* Tableview (hiển thị dữ liệu dạng bảng)

*Tương thích với Tkinter & TTK*

* Có thể sử dụng song song với các widget Tkinter/ttk cũ.
* Không phá vỡ code hiện có, chỉ cần thêm theme là giao diện tự động đẹp hơn.

# CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU DỰ ĐOÁN SỐNG SÓT TITANIC

### 2.1. Quá trình thu thập dữ liệu

Trong quá trình triển khai đề tài, nhóm chúng em đã lựa chọn sử dụng tập dữ liệu [*Titanic*](https://www.kaggle.com/c/titanic/data) từ nguồn dữ liệu uy tín của cộng đồng học máy và trí tuệ nhân tạo *Kaggle*. Bộ dữ liệu này được cộng đồng rộng rãi sử dụng nhờ tính toàn diện và độ tin cậy cao, đồng thời phản ánh bức tranh toàn cảnh về hành khách trên con tàu Titanic huyền thoại.

Dữ liệu Titanic chứa đựng những thông tin quan trọng như giới tính, độ tuổi, hạng vé (Pclass), số lượng người thân đi cùng (SibSp, Parch), giá vé (Fare), cảng lên tàu (Embarked) và đặc biệt là trạng thái sống sót (Survived). Tập dữ liệu này đã trở thành một chuẩn mực trong các nghiên cứu phân tích dữ liệu, cung cấp cho nhóm em nền tảng vững chắc để khai thác và phân tích mối quan hệ giữa các yếu tố xã hội và cơ hội sống sót của hành khách.

Việc lựa chọn bộ dữ liệu Titanic là hoàn toàn hợp lý bởi tính sẵn có, sự phong phú về thông tin cũng như khả năng ứng dụng cao trong việc xây dựng các mô hình phân tích và dự đoán. Những đặc điểm đó đã giúp nhóm chúng em thuận lợi trong việc xử lý, khám phá, trực quan hóa và rút ra các kết luận ý nghĩa, đóng góp vào việc hoàn thiện và phát triển đề tài.

### 2.2. Tổng quan về dữ liệu

1. Lịch sử hình thành: Tập dữ liệu Titanic được Kaggle giới thiệu từ năm 2012, và nó vẫn duy trì tính “kinh điển” qua thời gian, thường xuyên xuất hiện trong các khóa học và dự án phân tích dữ liệu.
2. Nguồn gốc: Bộ dữ liệu được xây dựng dựa trên thông tin của các hành khách thực tế trên con tàu Titanic, được tổng hợp từ nhiều nguồn tài liệu lịch sử và lưu trữ bởi các nhà nghiên cứu.
3. Công dụng: Dữ liệu này thường được sử dụng để phân tích và dự đoán tỷ lệ sống sót của hành khách dựa trên các yếu tố như độ tuổi, giới tính, hạng vé, số người đi cùng và các đặc điểm liên quan khác. Đây cũng là bài toán kinh điển trong việc học tập và thực hành các kỹ thuật khai thác dữ liệu, mô hình hóa và trực quan hóa.
4. Dữ liệu gồm 1309 dòng được đánh số từ 0 đến 1308 và có 12 cột dữ liệu.
5. Danh sách các đặc trưng:

* PassengerId: Mã số định danh của hành khách.
* Pclass: Hạng vé (1 – cao nhất,2 – Trung bình, 3 – thấp nhất).
* Name: Tên đầy đủ của hành khách.
* Sex: Giới tính (male/female).
* Age: Tuổi của hành khách.
* SibSp: Số anh chị em/vợ chồng đi cùng.
* Parch: Số cha mẹ/con cái đi cùng.
* Ticket: Số hiệu vé.
* Fare: Giá vé.
* Cabin: Mã cabin.
* Embarked: Cảng lên tàu (C, Q, S).
* Survived: Trạng thái sống sót (0 – không sống sót, 1 – sống sót).

1. Kiểu dữ liệu: Đa biến

|  |  |
| --- | --- |
| PassengerId | int64 |
| Pclass | int64 |
| Name | object |
| Sex | object |
| Age | float64 |
| SibSp | int64 |
| Parch | int64 |
| Ticket | object |
| Fare | float64 |
| Cabin | object |
| Embarked | object |
| Survived | int64 |

### 2.3. Các hạn chế và đặc điểm của dữ liệu

* Trong bộ dữ liệu Titanic, trường PassengerId đã tồn tại và được sử dụng như một khóa chính, đảm bảo tính duy nhất cho từng hành khách.
* Một số bản ghi thiếu dữ liệu
* Một số cột dữ liệu có giá trị bị thiếu (NaN), điển hình là:
* Age: Có nhiều hành khách không có thông tin tuổi.
* Cabin: Đa số hành khách không có thông tin cabin.
* Embarked: Một vài bản ghi không có cảng lên tàu.
* Một số bản ghi dữ liệu không hợp lệ
* Ticket: Có thể có định dạng không thống nhất (chữ-số, ký hiệu đặc biệt), không ảnh hưởng lớn nhưng cần chú ý nếu dùng để phân tích sâu

# CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU DỰ ĐOÁN SỐNG SÓT TITANIC

### 3.1. Xử lý dữ liệu

Trong bước xử lý dữ liệu, chúng em tiến hành kiểm tra và làm sạch toàn bộ dữ liệu bằng các thư viện Pandas và NumPy để đảm bảo rằng dữ liệu đầu vào có chất lượng tốt nhất cho quá trình phân tích. Các bước xử lý bao gồm:

#### 3.1.1. Kiểm tra dữ liệu ban đầu

Đầu tiên, chúng em đọc 3 file dữ liệu train.csv, test.csv, và gender\_submission.csv bằng thư viện pandas để nạp dữ liệu vào DataFrame. File test được merge với gender\_submission để thêm cột Survived, sau đó kết hợp với train data thành một dataset thống nhất. Chúng em kiểm tra missing values, shape của dataset và data types của từng cột để có cái nhìn tổng quan về dữ liệu ban đầu.

#### 3.1.2. Xử lý các giá trị bị thiếu

Sau khi chuẩn hóa missing values, chúng em xử lý các giá trị NaN theo từng cột:

* Age: Sử dụng median để fill các giá trị NaN
* Fare: Sử dụng median để fill các giá trị NaN
* Cabin: Sử dụng mode (giá trị xuất hiện nhiều nhất) để fill
* Embarked: Sử dụng mode để fill các giá trị NaN

#### 3.1.3. Loại bỏ các giá trị trùng lặp

Ngay sau khi load dữ liệu, chúng em thực hiện loại bỏ các dòng trùng lặp bằng hàm drop\_duplicates() của pandas để đảm bảo mỗi hành khách chỉ xuất hiện một lần trong tập dữ liệu.

#### 3.1.4. Kiểm tra và xác thực dữ liệu hợp lệ

Để đảm bảo tính chính xác, chúng em thực hiện validation dữ liệu:

* Chuyển đổi string 'nan': Các giá trị string 'nan' trong các cột object được chuyển thành NaN thực
* Chuẩn hóa missing values: Các dạng khác nhau của missing values ('', ' ', 'null', 'NULL', 'None', 'none') được chuyển thành NaN
* Kiểm tra invalid entries: Xác minh các cột numeric không chứa giá trị không thể chuyển đổi thành số
* Validation cột Sex: Kiểm tra phân phối giá trị trong cột giới tính

#### 3.1.5. Lưu dữ liệu đã làm sạch

Sau khi dữ liệu được làm sạch hoàn tất, chúng em lưu lại dữ liệu vào file CSV mới có tên cleaned.csv trong thư mục dataset/titanic/. Việc lưu dữ liệu sạch vào file mới giúp thuận tiện hơn cho các bước phân tích tiếp theo.

#### 3.1.6. Tổng hợp

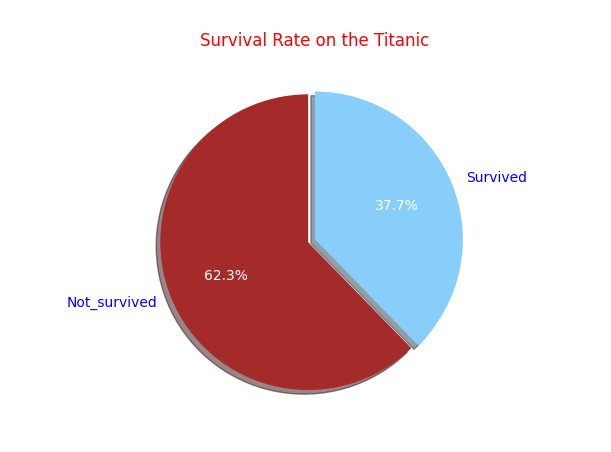
Các bước trên đảm bảo rằng dữ liệu đầu vào:

* Được merge và kết hợp từ nhiều file thành dataset thống nhất
* Không chứa giá trị trùng lặp
* Các dạng missing values khác nhau được chuẩn hóa thành NaN
* Các giá trị numeric đều hợp lệ và có thể chuyển đổi
* Không còn giá trị bị thiếu (đã được fill phù hợp)
* Được tổ chức và lưu trữ gọn gàng trong file CSV mới
* Có function load\_cleaned\_data() để dễ dàng load lại dữ liệu đã clean

Quá trình xử lý dữ liệu tuần tự này giúp giảm thiểu lỗi, đảm bảo chất lượng và tính nhất quán của dữ liệu, từ đó hỗ trợ cho các phân tích và trực quan hóa về sau.

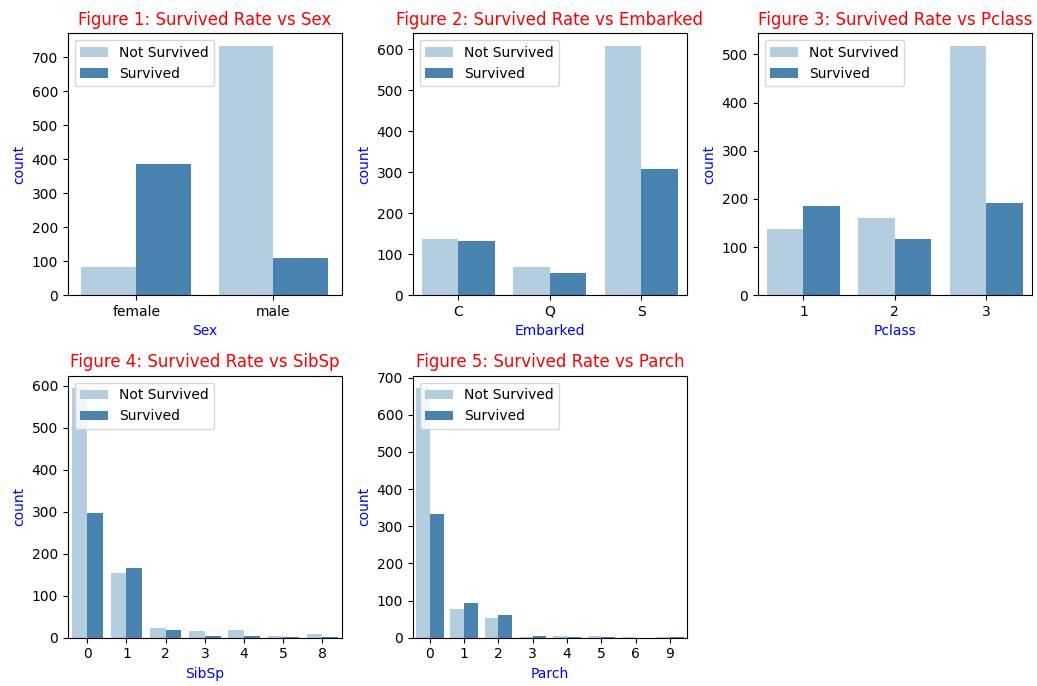
### 3.2. Trực quan hóa dữ liệu

Chúng em tiến hành tạo các biểu đồ và đồ thị để trực quan hóa dữ liệu hành khách từ tập dữ liệu đã được làm sạch, nhằm minh họa rõ ràng các mối quan hệ giữa các yếu tố đặc trưng với tỉ lệ sống sót. Quá trình trực quan hóa này không chỉ giúp phân tích đặc điểm của từng nhóm đối tượng mà còn hỗ trợ nhận diện xu hướng, sự chênh lệch và mối tương quan giữa các yếu tố khác nhau, từ đó đưa ra những nhận định và kết luận chính xác hơn. Chúng em sử dụng các thư viện trực quan hóa mạnh mẽ như Seaborn và Matplotlib để tạo ra các loại biểu đồ như biểu đồ cột, biểu đồ tròn, biểu đồ phân tán và biểu đồ hộp, mang lại cái nhìn trực quan và dễ hiểu về dữ liệu. Cụ thể các biểu đồ như sau:



Hình 3. 1 Biểu đồ tỉ lệ sống sót và gặp nạn

Biểu đồ trên thể hiện tỉ lệ phần trăm sống sót và không sống sót trên con tàu Titanic. Vào ngày 15 tháng 4 năm 1912, trong chuyến đi đầu tiên của mình, RMS Titanic được nhiều người coi là "không thể chìm" đã chìm sau khi va chạm với một tảng băng trôi. Thật không may, không có đủ xuồng cứu sinh cho tất cả mọi người trên tàu, dẫn đến cái chết của 1502 trong số 2224 hành khách và thủy thủ đoàn. Nhìn vào biểu đồ có thể thấy khoảng 2/3 số lượng hành khách đã không may mắn gặp nạn, đây là một con số kinh hoàng cho một thảm họa đường thủy thời bấy giờ, gây ra sự đau buồn tột độ và phẫn nộ sâu sắc. Một câu hỏi được đặt ra là tại sao có quá ít người sống sót và những người sống sót có đặc trưng gì, tại sao họ lại may mắn đến vậy?



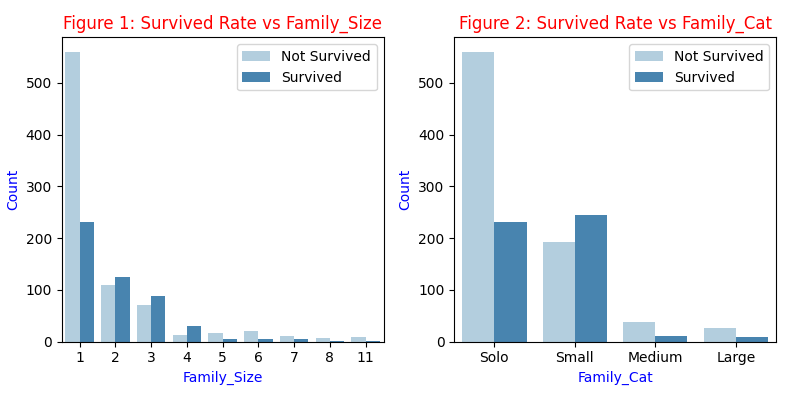
Hình 3. 2 Biểu đồ sự tương quan giữa sống sót với các đặc trưng

Xét mối liên hệ giữa Sống sót (Survived) – Giới tính (Sex): phân bố khá chênh lệch, khả năng sống sót của nữ giới lên tới 80%, trong khi đó nam giới chỉ khoảng 12,5%. Ở đây có thể lí giải do dành cho phụ nữ “lady first” mà Phương Tây vẫn đang áp dụng.

Sống sót – Cảng biển (Embarked): chỉ có cảng Cherbourg thì tỉ lệ sống và gặp nạn gần tương đương nhau, cảng Southampton chiếm số lượng hành khách lớn nhất nhưng tỉ lệ sống sót lại khá thấp, chỉ khoảng 30% mà thôi. Ở đây ta sẽ khó giải thích được lý do nhưng ta sẽ đề cập lại sau.

Sống sót – Hạng vé (Pclass): Tỉ lệ sống sót ở hạng 1 khoảng 57%, cao hơn hẳn 2 hạng khác, hành khách hạng vé thứ 3 chiếm số lượng cao nhất trên tàu nhưng khả năng sống sót chỉ khoảng 28% . Một điểm khá tương đồng với Embarked

Sống sót – Số lượng anh/chị/em (SibSp) và sống sót – số lượng ba mẹ (Parch): Hai dữ liệu này có khá nhiều mục và chênh lệch khá lớn. Điều này khá phức tạp khi phân tích, vì thế em xin tổng hợp lại thành đặc trưng gia đình và biểu thị bằng biểu đồ bên dưới.



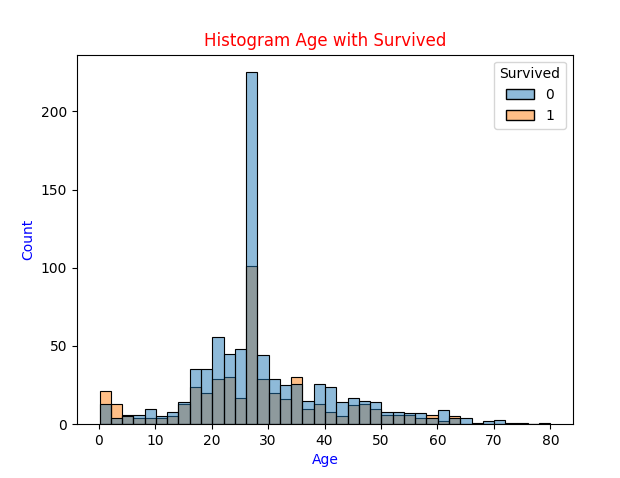
Hình 3. 3 Biểu đồ tương quan giữa sống sót với đặc trưng gia đình

Ở đây số lượng thành viên trong gia đình sẽ được tính bằng công thức:

Family\_Size = SibSp + Parch + 1

Như vậy số thành viên trong gia đình của tập dữ liệu trên sẽ nằm trong khoảng từ 1 đến 11 người, cũng còn khá nhiều mục và chênh lệch lớn, vì thế tiếp tục phân loại theo số lượng thành viên thành 4 mức, đi một mình (Solo) chỉ 1 người, gia đình nhỏ (Small) từ 2 – 3 người, gia đình trung bình (Medium) từ 4 – 5 người, gia đình lớn (Large) từ 6-11 người.

Với biểu đồ “Survived vs Family\_Cat” ta có thể thấy rằng số lượng các hành khách đi một mình chiếm khá cao nhưng tỉ lệ sống sót lại thấp chỉ khoảng 28%. Ở đây tỉ lệ sống sót lượng hành khách đi theo gia đình nhỏ 2-4 người là cao nhất.

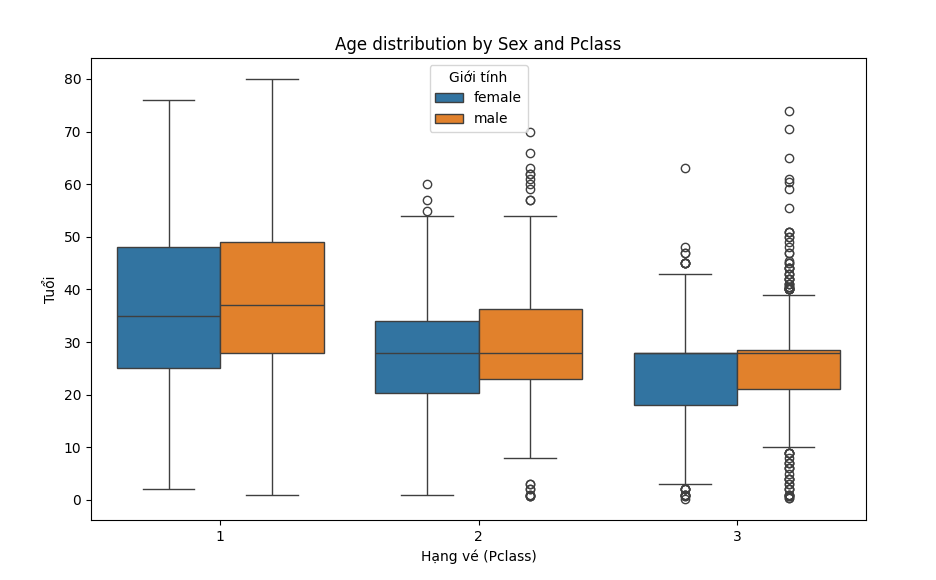


Hình 3. 4 Biểu đồ biểu diễn sống sót theo tuổi

Biểu đồ trên cho thấy sự phân bố tuổi đối với tỉ lệ sống sót của hành khách. Phần lớn khách hàng nằm trong độ tuổi từ 18 đến 42 tuổi, nổi bật nhất là tuổi 26, tuy nhiên tỉ lệ sống sót hầu hết đều thấp chỉ cá biệt một số trường hợp là trẻ em dưới 6 tuổi đa phần đều được an toàn. Một số cá biệt khác như tuổi 36, tuổi 58, tuổi 62 cũng khá cao, đặc biệt nhất là 1 cụ già 80 tuổi vẫn bình an.

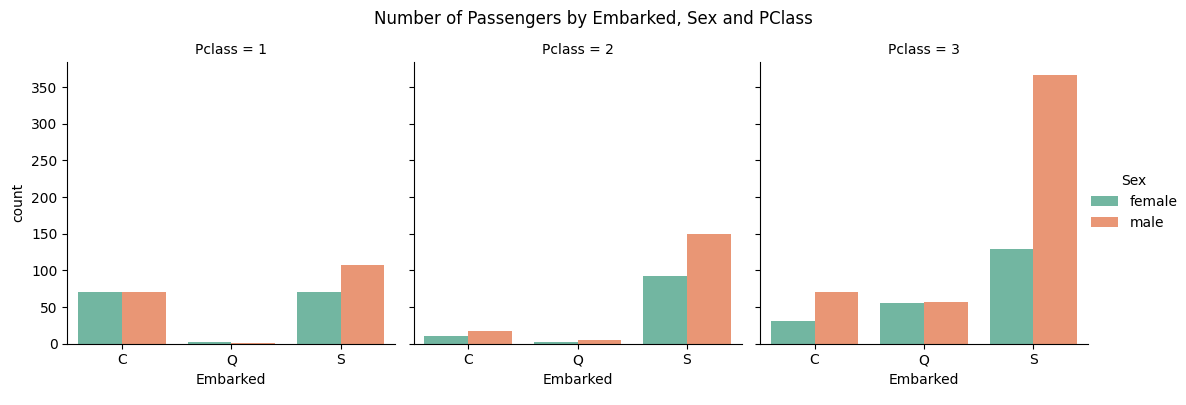
Sau khi phân tích khả năng sống sót thông qua các đơn yếu tố đặc trưng ta có thể rút ra nhận định sơ bộ như: những hành khách mang đặc trưng là nữ hoặc trẻ em, đi vé hạng 1, xuất phát từ cảng Cherbourg, đi cùng gia đình nhỏ (2-3 người) sẽ có tỉ lệ sống sót cao hơn. Đối với những hành khách là nam giới, vé hạng 3 hoặc xuất phát từ Southampton sẽ có tỉ lệ gặp nạn cao hơn.

Để lý giải những nhận định ban đầu, chúng em sẽ phân tích các đặc trưng trong mối liên hệ với nhau.

**

Hình 3. 5 Biểu đồ phân bố tuổi theo hạng vé và giới tính

Biểu đồ cho thấy sự phân bố tuổi theo hạng giới tính và hạng vé. Ở vé hạng 1, ta thấy có sự phân bố tuổi khá rộng từ 1 đến 80, số lượng nữ giới trong hạng này nhiều hơn nam giới, phân bố chủ yếu từ 26 đến 48 tuổi, cụ bà 80 tuổi may mắn sống sót cũng nằm trong hạng vé này. Ở vé hạng 2 phân bố hẹp hơn từ 20 đến 36 tuổi, xuất hiện khá nhiều giá trị outliers ngoại lai, số lượng nam và nữ gần tương đương nhau. Ở vé hạng 3, số lượng nam giới nhiều hơn nữ, phân bố hẹp hơn 2 hạng còn lại, xuất hiện nhiều giá trị ngoại lai, đây có thể là những giá trị đặc biệt hoặc cũng có thể là do sự sai xót trong tập dữ liệu.



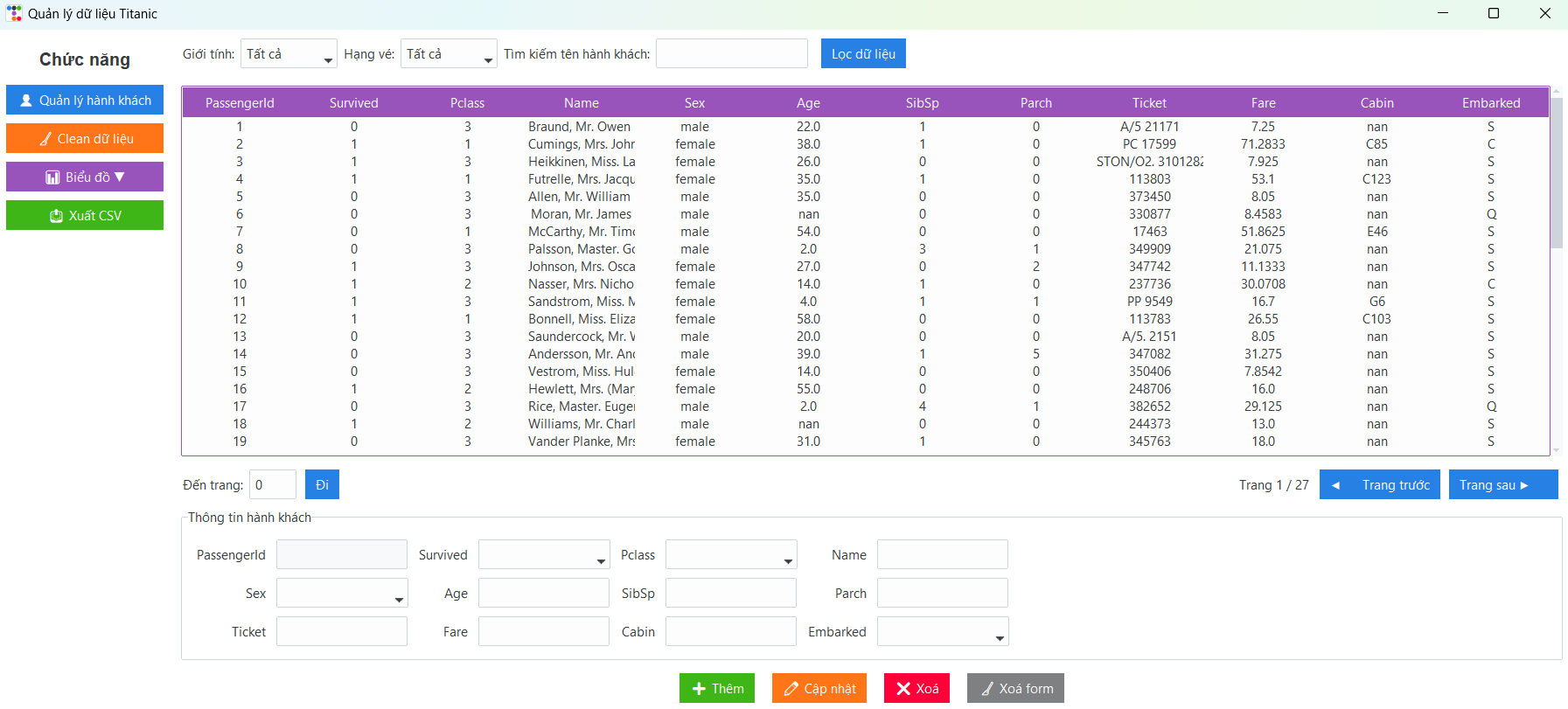
Hình 3. 6 Biểu đồ số lượng hành khách qua giới tính, cảng biển và hạng vé

Biểu đồ trên cho thấy ở cảng Cherbourg, số lượng hành khách nữ nhiều hơn số lượng hành khách nam, hơn nữa lượng hành khách tập trung ở hạng 1 là khá đông. Đây là điểm tương đồng lớn lý giải tại sao vé hạng 1 và cảng Cherbourg lại có tỉ lệ sống sót cao hơn. Ngược lại số lượng hành khách ở cảng Southampton chiếm khá lớn trong tổng hành khách trên tàu, tuy nhiên đa phần đều là nam giới và đi hạng 3 vì thế tỉ lệ gặp nạn ở đây cao hơn 2 cảng còn lại.

.

### 3.3. Thiết kế giao diện

Để cho người dùng dễ dàng tương tác với dữ liệu, có thể tương tác trực tiếp nhờ có các nút chức năng cơ bản và trực quan thì nhóm chúng em đã dùng thư viện TTKBootstrap trong python để xây dựng một giao diện người dùng (GUI) đơn giản.



Hình 3. 7 Giao diện người dùng

Giao diện này được thiết kế bao gồm các chức năng cơ bản như tải dữ liệu, quản lí dữ liệu, trực quan dữ liệu, thoát chương trình. Điều này giúp người dùng dễ dàng quản lí và tổ chức dữ liệu mà không cần phải có nhiều kiến thức chuyên sâu về lập trình. Cụ thể:

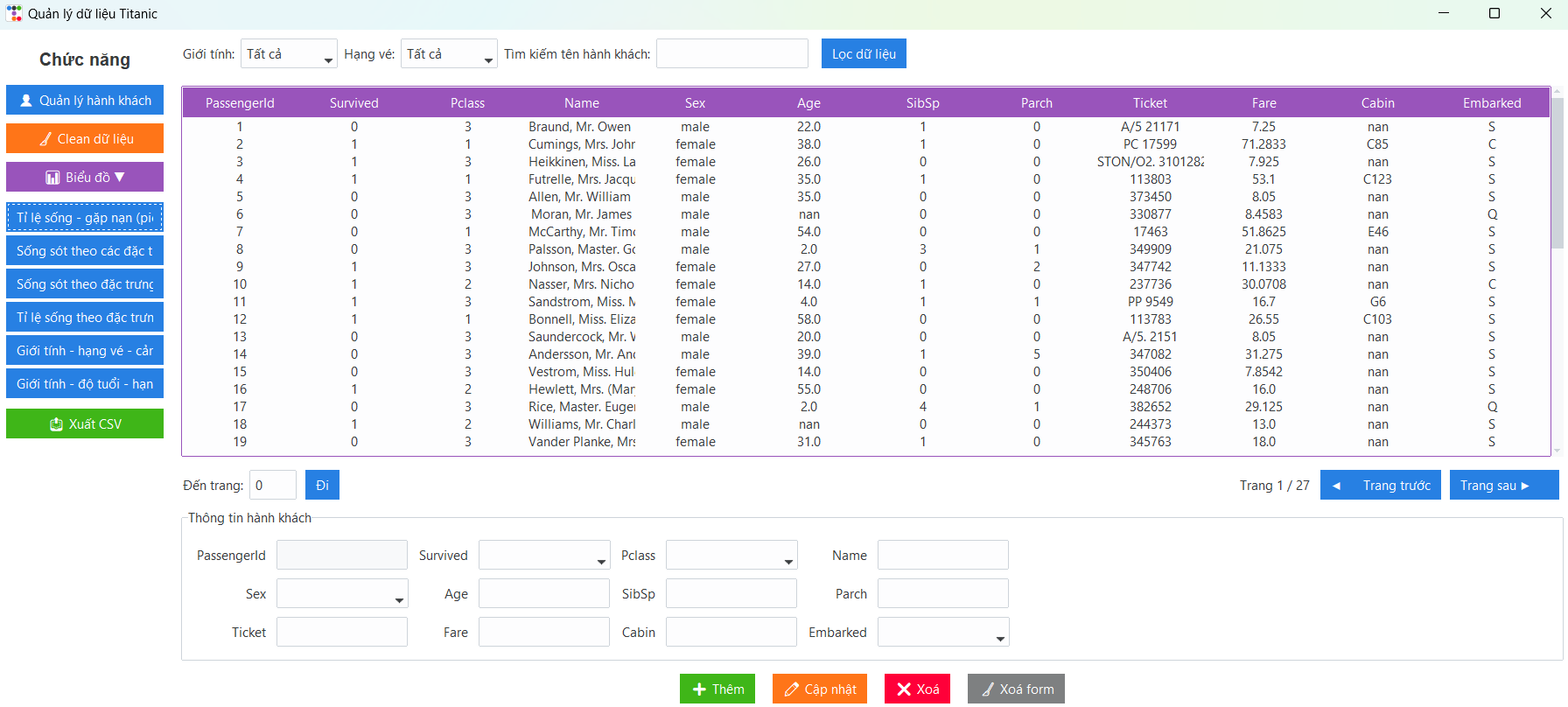
Bảng dữ liệu: Giao diện chính là một bảng dữ liệu cho phép người dùng dễ dàng xem và quan sát các bản ghi dữ liệu.

Các nút điều khiển: Có bốn nút cơ bản là Quản lí hành khách, Làm sạch hóa dữ liệu Biểu đồ, Xuất file dữ liệu và Thoát chương trình. Mỗi nút được thiết kế dễ nhìn và dễ hiểu, phù hợp cho người dùng dễ dàng thao tác và sử dụng, dễ dàng cho người mới làm quen thao tác.

Giao diện quản lý dữ liệu bao gồm các chức năng chính như:

* Thêm: Dùng để tạo và nhập dữ liệu mới cho hành khách.
* Thông tin hành khách: Xem thông tin của hành khách.
* Cập nhật: Cập nhật thông tin của hành khách.
* Xóa: Xóa dữ liệu của hành khách.

Làm sạch dữ liệu: Một trong những bước quan trọng trong quá trình phân tích dữ liệu là xử lý dữ liệu, bao gồm việc loại bỏ các giá trị thiếu, xử lý các giá trị không hợp lệ, và chuẩn hóa dữ liệu. Thì ứng dụng này đã rút gọn nó quá trình này và tự động xử lí và làm sạch dữ liệu khi dữ liệu được tải lên nhờ nút “Làm sạch dữ liệu”.



Hình 3. 8 Giao diện Biểu đồ

Biểu đồ giao diện trực quan hóa dữ liệu dùng để phân tích dữ liệu học sinh thông qua biểu đồ, bao gồm các biểu đồ như sau:

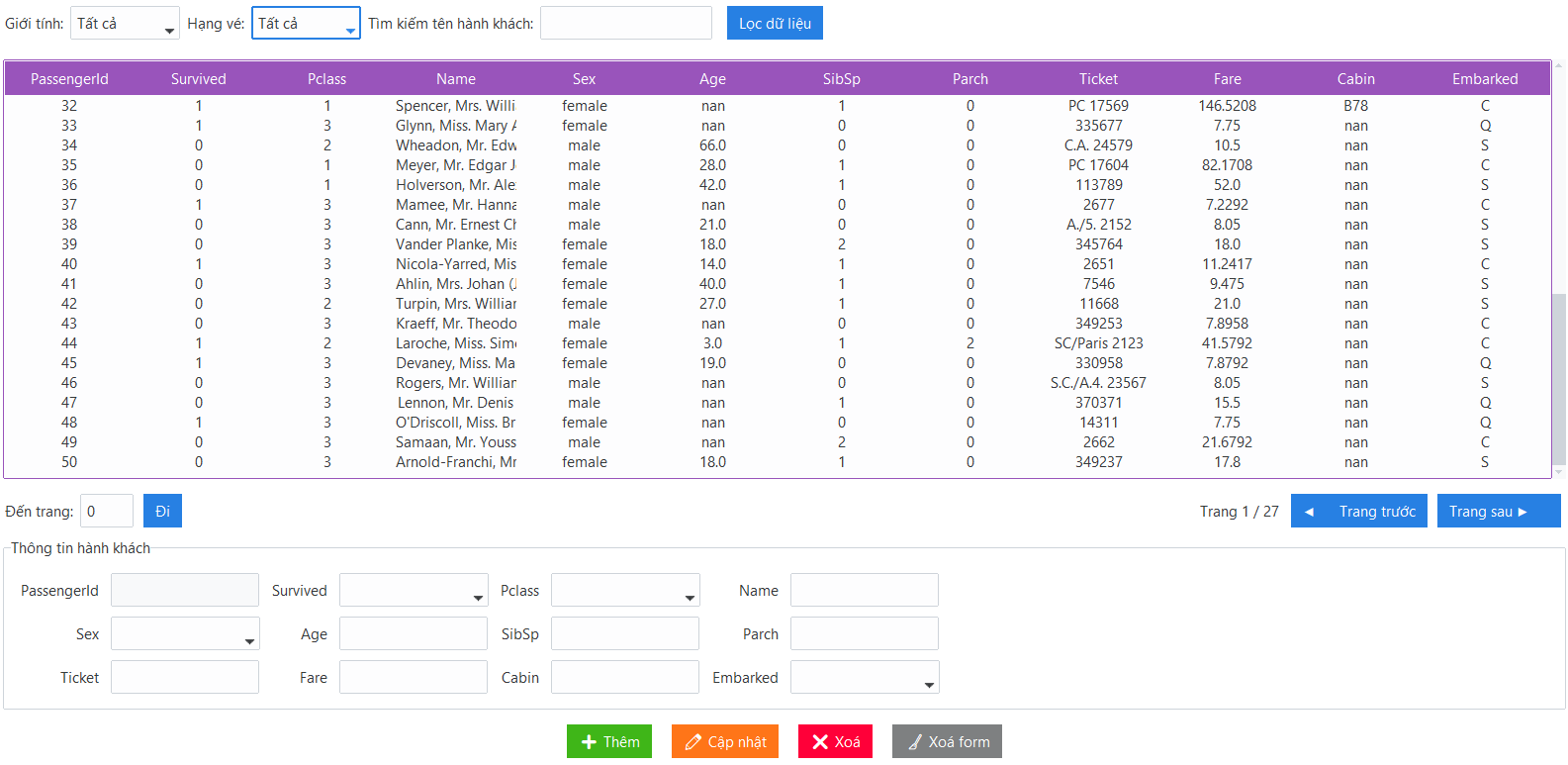
* Biểu đồ tỉ lệ sống – gặp nạn (pie)
* Biểu đồ sống sót theo các đặc trưng (subplot)
* Biểu đồ sống sót theo Sex
* Biểu đồ sống sót theo Embarked
* Biểu đồ sống sót theo Pclass
* Biểu đồ sống sót theo SibSp
* Biểu đồ sống sót theo Parch
* Biểu đồ sống sót theo đặc trưng gia đình (countplot)
* Biểu đồ tỉ lệ sống theo đặc trưng tuổi (hist)
* Biểu đồ tương quan Giới tính - hạng vé - cảng biển (countplot)
* Biểu đồ tương quan Giới tính - độ tuổi - hạng vé (boxplot)

Xuất file dữ liệu: Sau khi chúng làm sạch dữ liệu bằng chức năng làm sạch dữ liệu, có thể xuất file dữ liệu đó ra một file dữ liệu mới không làm ảnh hưởng đến dữ liệu cũ

### 3.4. Các chức năng chính của chương trình

#### 3.4.1. Thêm trong Quản lý hành khách

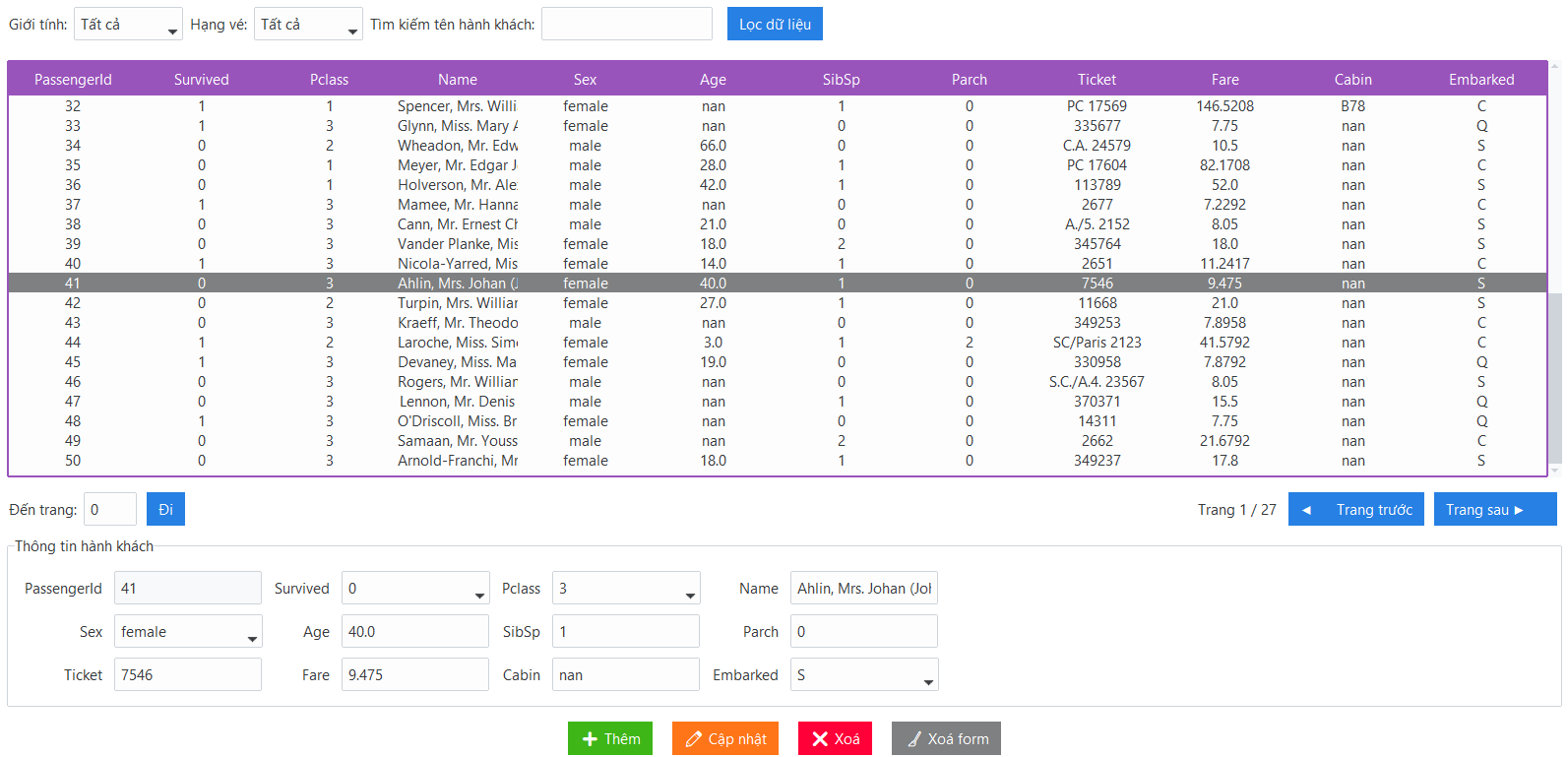
Chức năng tạo dữ liệu cho phép người dùng (người quản lý) nhập dữ liệu trực tiếp cho một hành khách.



Hình 3. 9 Chức năng Thêm hành khách mới

#### 3.4.2. Đọc bản ghi trong Quản lý hành khách

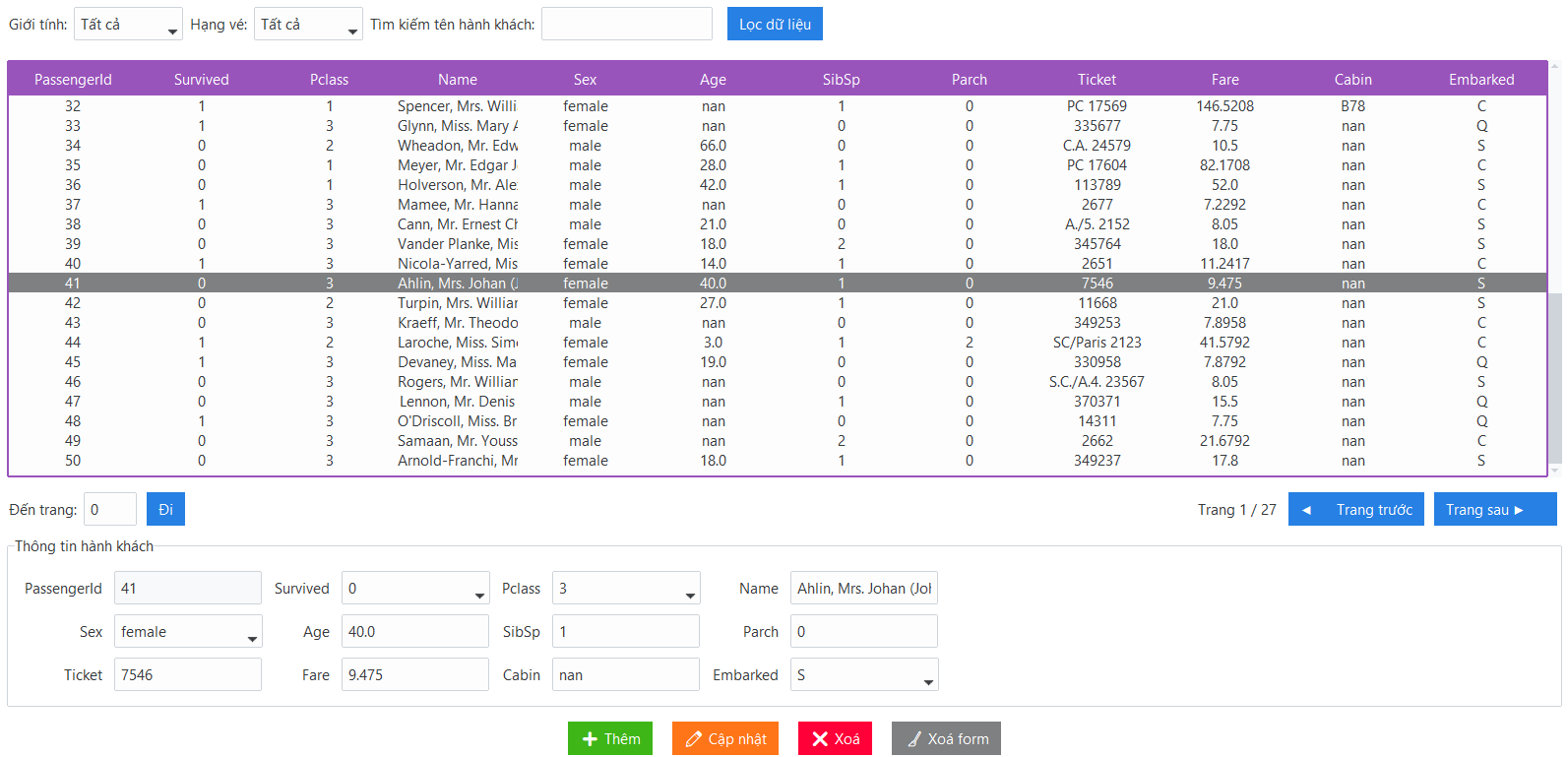
Chức năng đọc bản ghi cho phép người dùng (người quản lý) đọc dữ liệu cụ thể của một hành khách.



Hình 3. 10 Chức năng Đọc bản ghi

#### 3.4.4. Cập nhật dữ liệu hành khách trong Quản lý hành khách

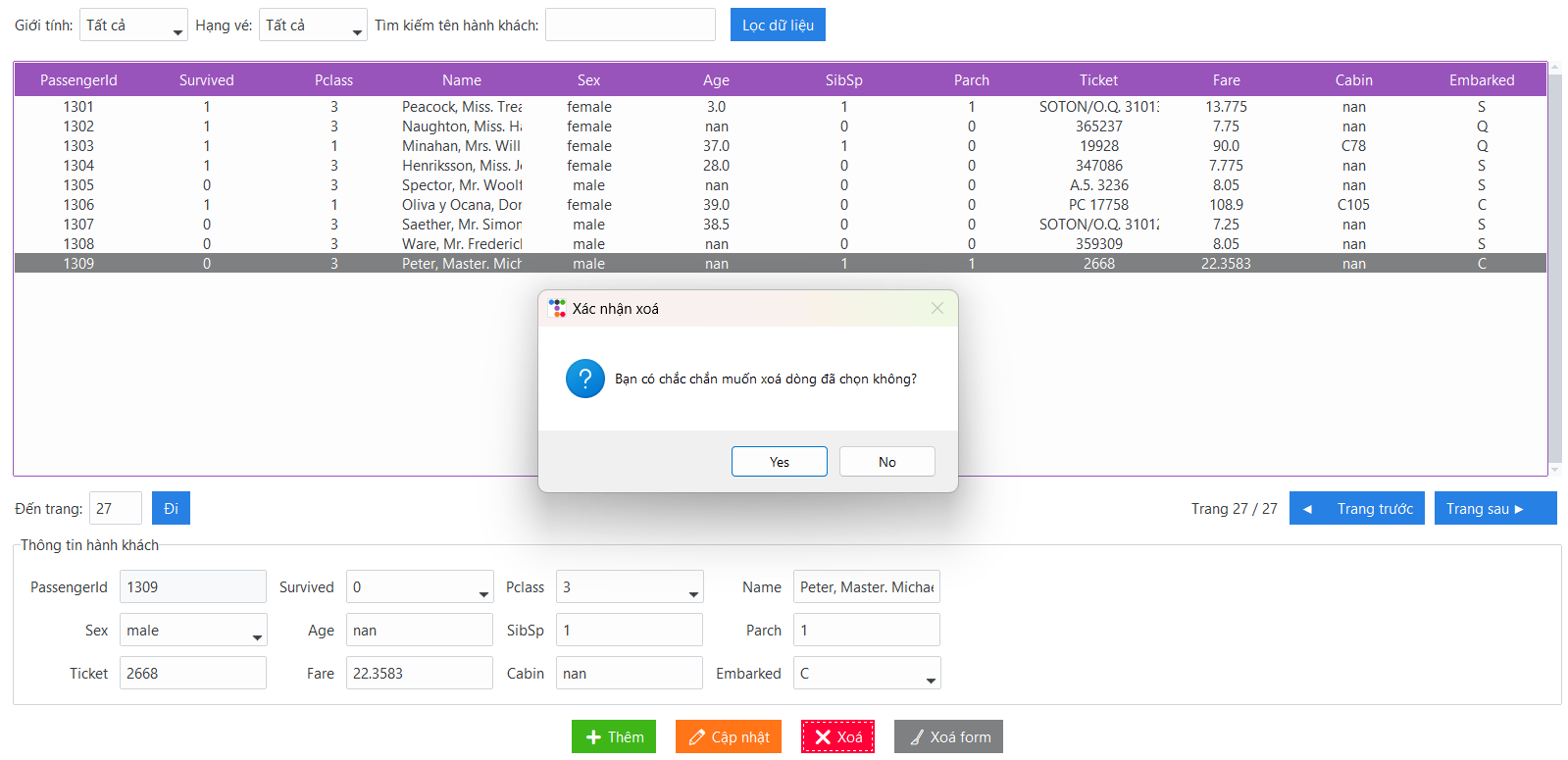
Chức năng cập nhật dữ liệu cho phép người dùng (người quản lý) được truy cập, sửa đổi dữ liệu của hành khách khi trường hợp dữ liệu của hành khách đó là không đúng.



Hình 3. 11 Chức năng Cập nhật dữ liệu cho hành khách

#### 3.4.5. Xóa bản ghi trong Quản lý hành khách

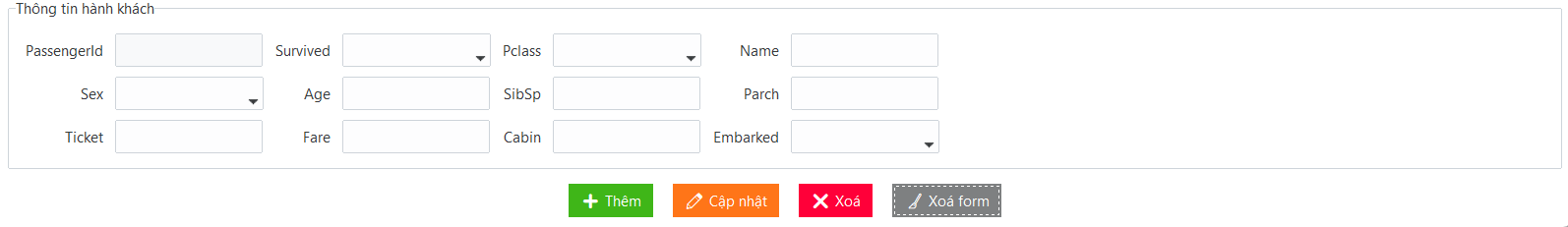
Chức năng Xóa cho phép người dùng (người quản lý) được truy cập, xóa dữ liệu của hành khách khi trường hợp dữ liệu của hành khách đó là không đúng



Hình 3. 12 Chức năng Xóa một bản ghi của hành khách

#### 3.4.6. Chức năng Xóa form

Chức năng Xóa form cho phép người dùng (người quản lý) dọn dẹp các ô nhập liệu

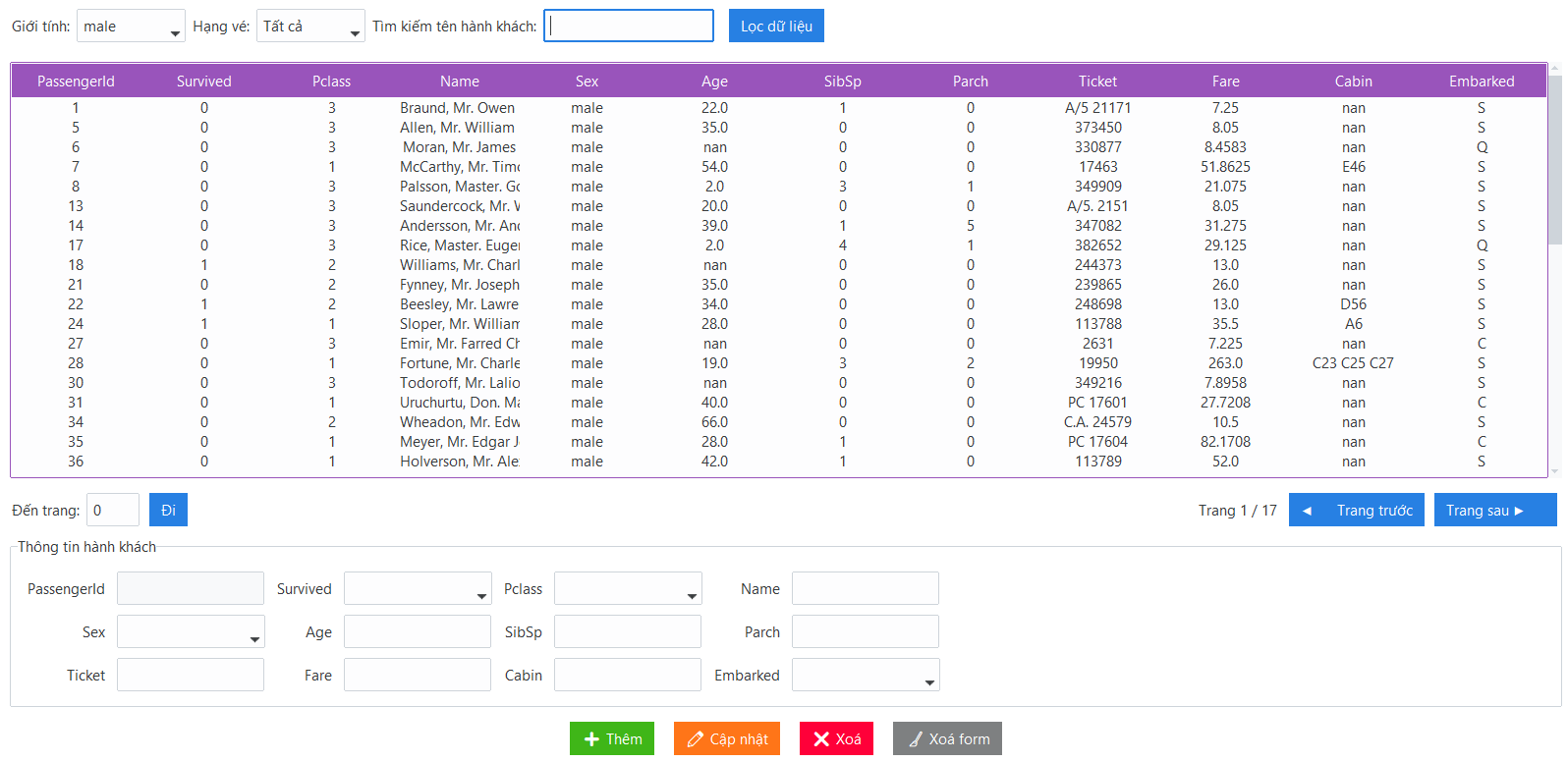


Hình 3. 13 Chức năng Xóa form

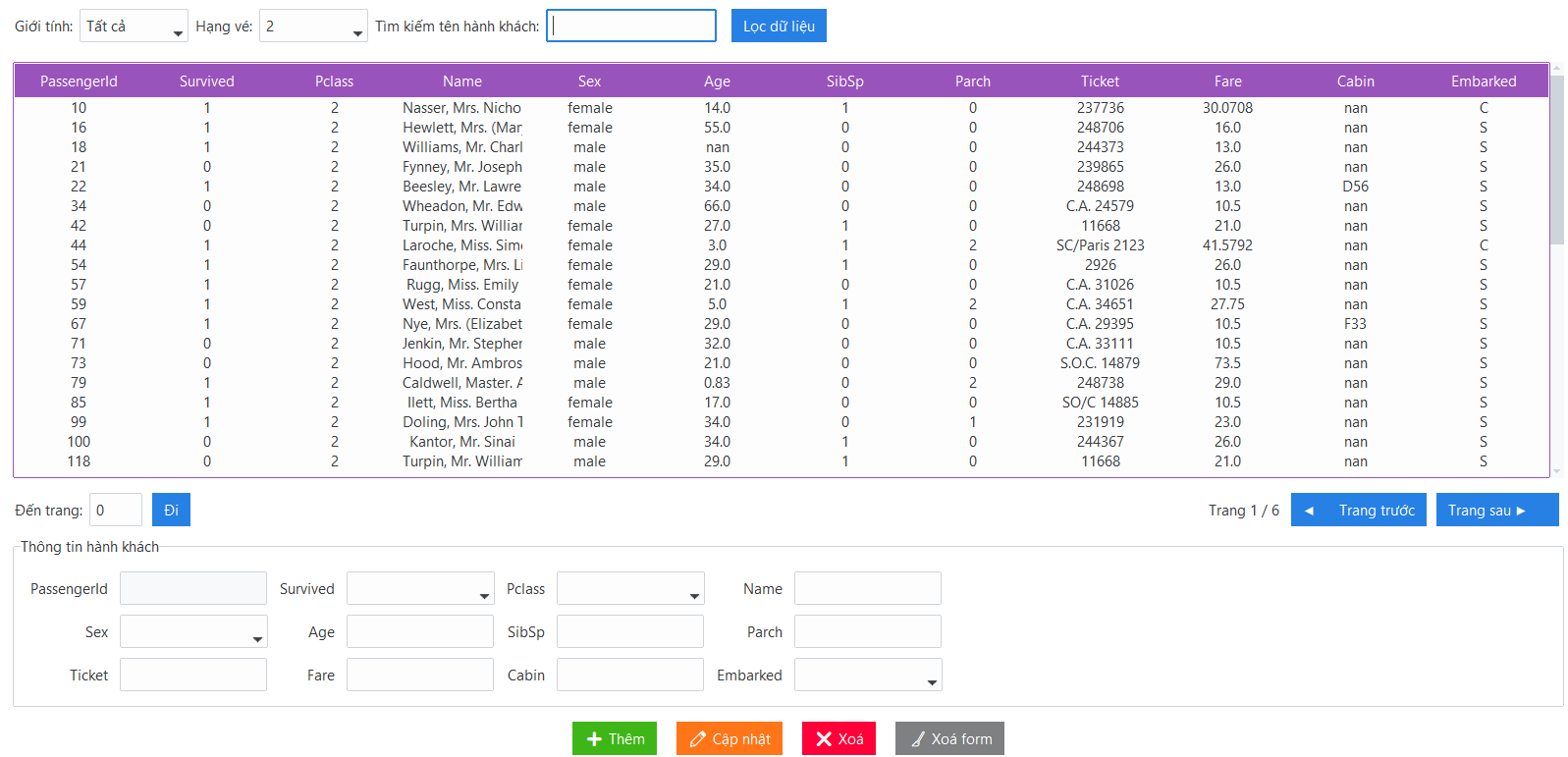
#### 3.4.6. Chức năng lọc dữ liệu

Chức năng lọc dữ liệu của chương trình cho phép người dùng (người quản lý) có thể lọc dựa trên các tiêu chí để người dùng có thể đánh giá một cách trực quan nhất trên tập dữ liệu.

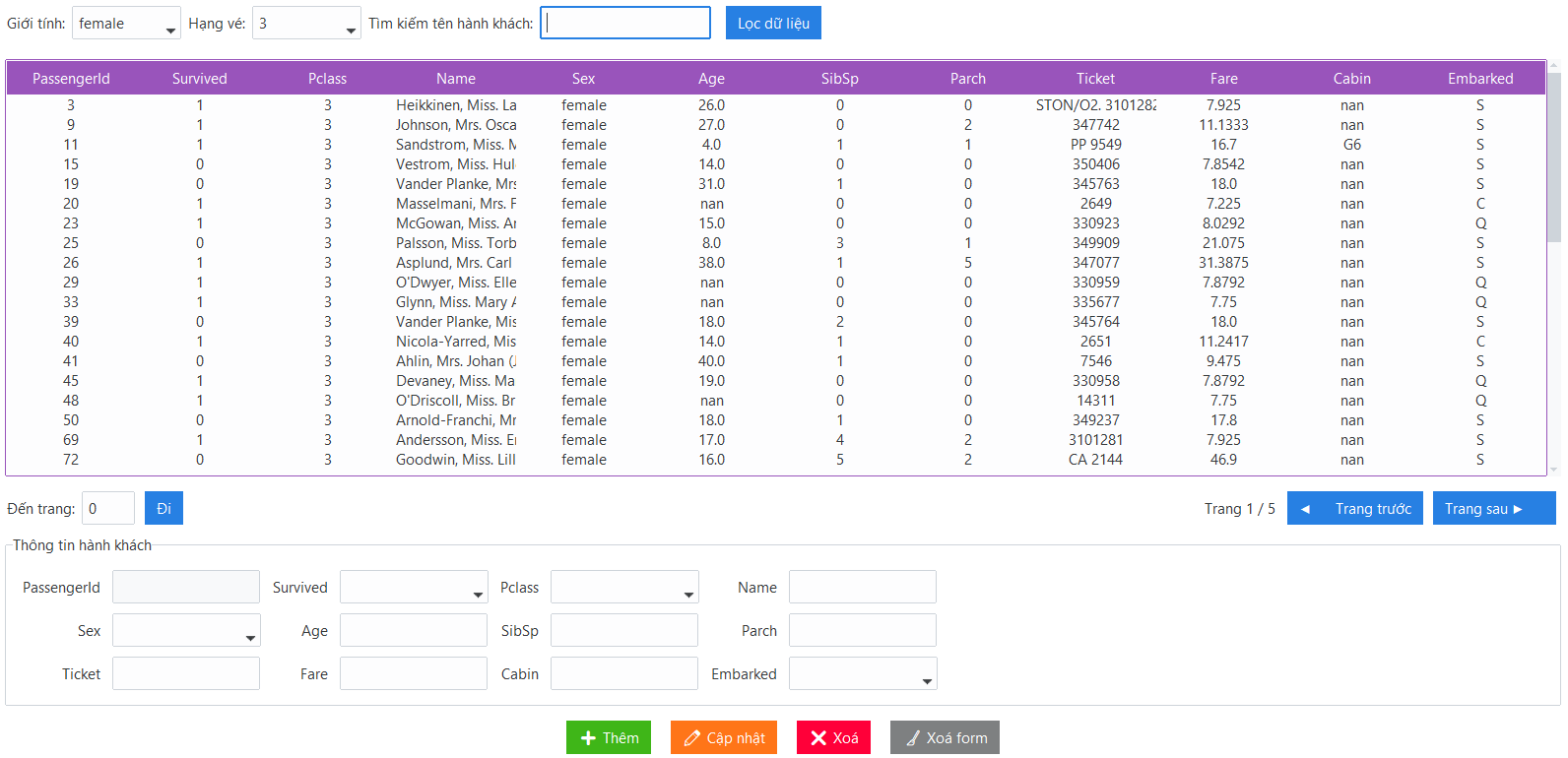
* Lọc theo Giới tính
* Hành khách Nam (male)
* Hành khách Nữ (female)
* Lọc theo Hạng vé
  + - Hạng 1
    - Hạng 2
    - Hạng 3



Hình 3. 14 Lọc dữ liệu theo Giới tính Nam



Hình 3. 15 Lọc dữ liệu theo Hạng vé (Pclass)

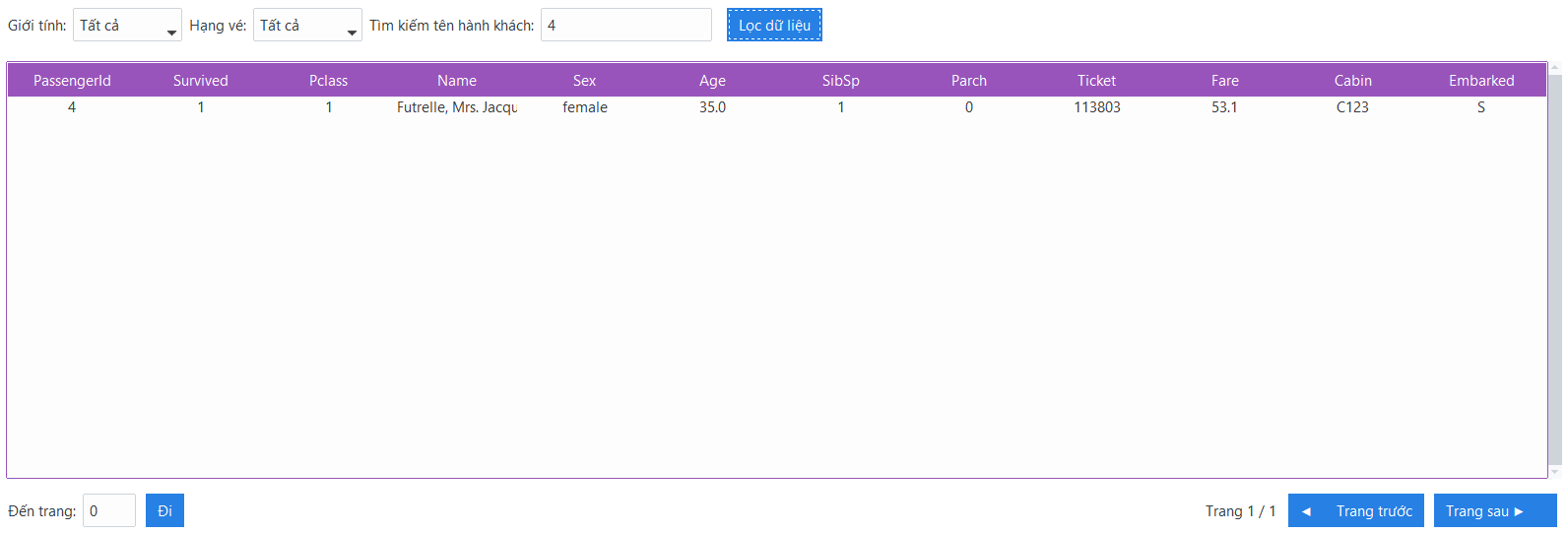


Hình 3. 16 Lọc dữ liệu theo giới tính (male ) và Hạng vé ( hạng 3)

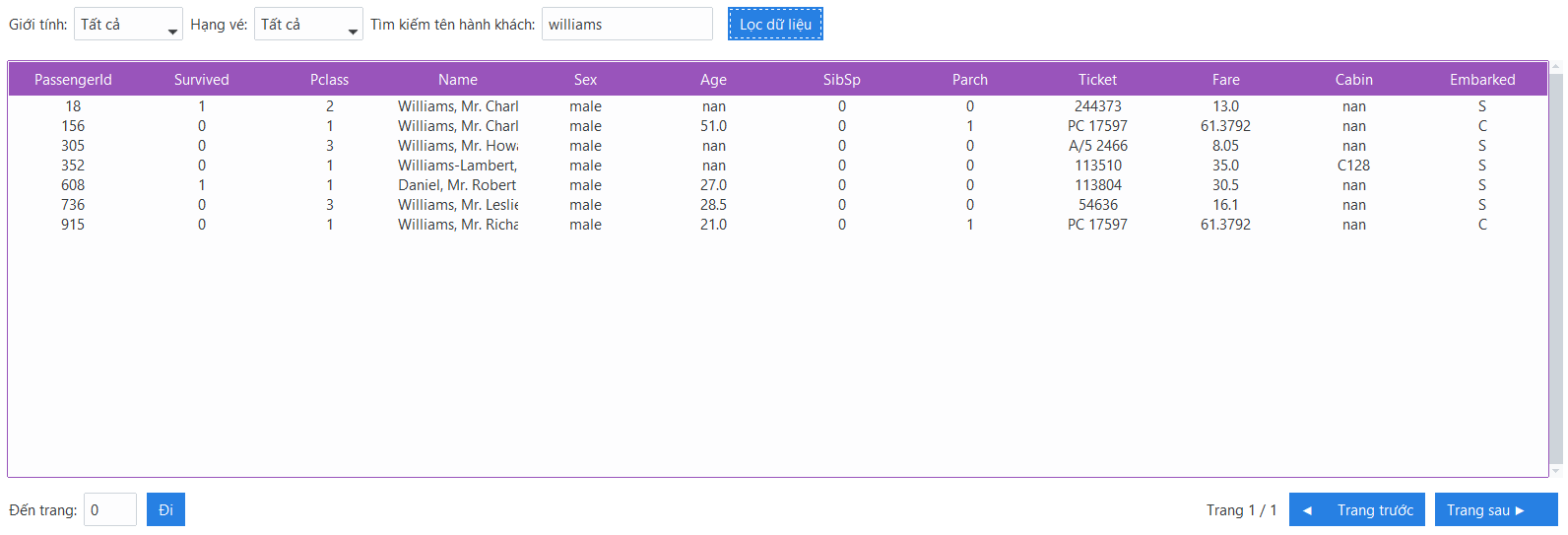
**3.4.7. Chức năng tìm kiếm dữ liệu**

Tìm kiếm theo PassengerId hoặc Name

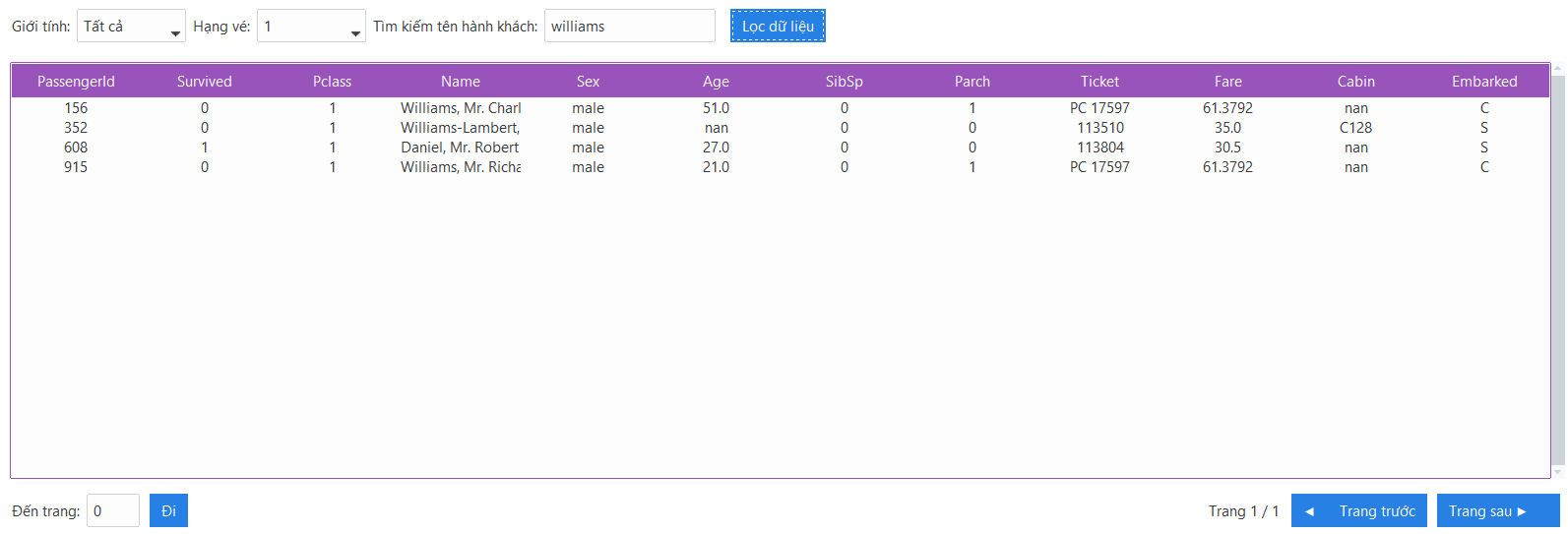
* Tìm kiếm PassengerId hoặc Name theo các giá trị đã lọc



Hình 3. 17 Tìm kiếm dữ liệu dựa trên PassengerId



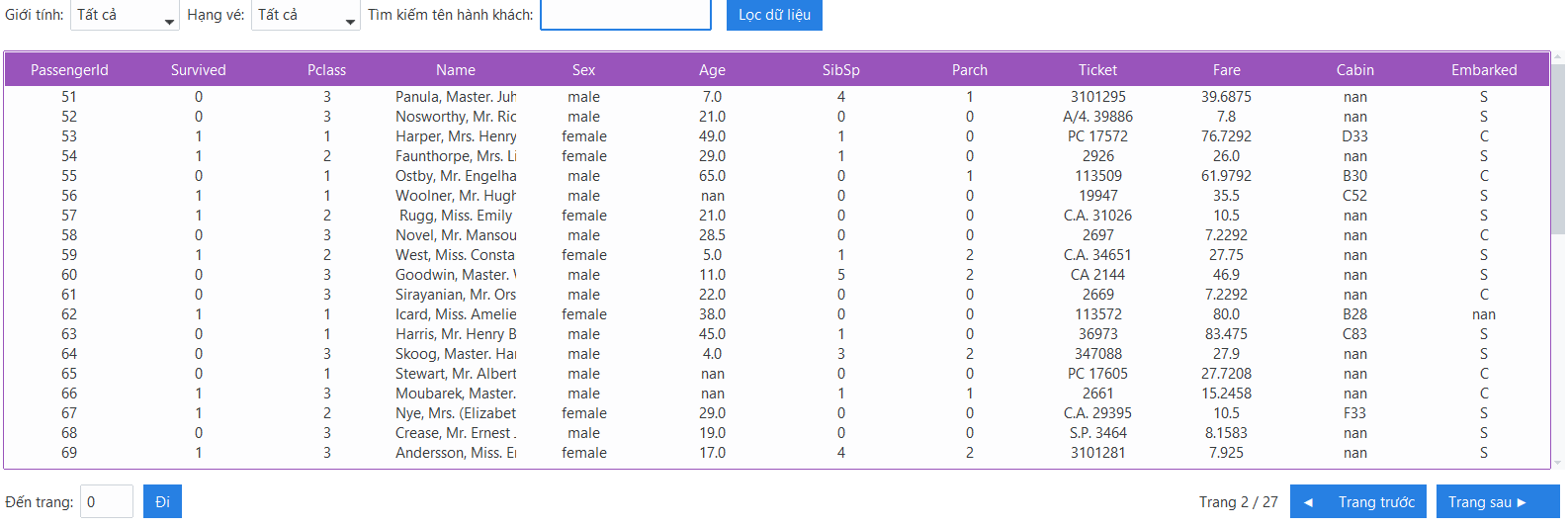
Hình 3. 18 Tìm kiếm dữ liệu dưa trên Name



Hình 3. 19 Tìm kiếm dữ liệu dưa trên Name kết hợp bộ lọc

#### 3.4.8. Chức năng phân trang

Bộ dữ liệu người dùng được phân thành từng trang và được hiển thị trên màn hình chính, người quản lý có thể chọn Trang sau để đi tới trang kế tiếp, Trang trước để quay lại trang trước, nhập số trang vào ô nhập dữ liệu Đến trang chọn Đi để đến trang muốn đến, mỗi trang có 50 trường dữ liệu của hành khách



Hình 3. 20 Chức năng phân trang

# PHẦN KẾT LUẬN

## 1. Kết quả đạt được

**Về kiến thức:**

Qua quá trình thực hiện đề tài, chúng em đã có cơ hội làm việc với tập dữ liệu thực tế, giúp hiểu rõ hơn về cách phân tích và xử lý dữ liệu lớn.Chúng em tập trung nhận diện các yếu tố đặc trưng ảnh hưởng đến khả năng sống sót của hành khách trên tàu Titanic, cũng như mối tương quan giữa chúng với nhau.

Các bước làm sạch và chuẩn hóa dữ liệu giúp chúng em nhận ra tầm quan trọng của việc đảm bảo chất lượng dữ liệu trước khi phân tích. Những lỗi như dữ liệu bị thiếu, trùng lặp, hay không hợp lệ cần được xử lý kỹ lưỡng để đảm bảo độ chính xác và tính toàn vẹn của dữ liệu đầu vào.

Quá trình phân tích và trực quan hóa đã giúp nhóm chúng em nắm rõ hơn về việc sử dụng các công cụ như pandas, matplotlib, seaborn để khai thác thông tin từ dữ liệu. Chúng em nhận thấy rằng trực quan hóa là một bước quan trọng để nhận diện các xu hướng và mối quan hệ phức tạp mà dữ liệu thô không thể hiện rõ.

Bên cạnh đó, qua việc phân tích các biểu đồ, chúng em đã học cách trình bày và diễn giải kết quả phân tích sao cho trực quan, dễ hiểu, từ đó hỗ trợ quá trình đưa ra nhận định chính xác.

**Về sản phẩm:**

Sản phẩm của đề tài là một tập dữ liệu đã được làm sạch và chuẩn hóa, sẵn sàng cho các bước phân tích và áp dụng trong các dự án tiếp theo. Việc xây dựng quy trình làm sạch và lưu trữ dữ liệu đã giúp cải thiện đáng kể chất lượng dữ liệu, đảm bảo tính chính xác và hiệu quả.

Hệ thống giao diện quản lý dữ liệu, với các chức năng như tải dữ liệu, làm sạch, chuẩn hóa, thêm, xóa, cập nhật và trực quan hóa dữ liệu, giúp đơn giản hóa các tác vụ phân tích cho người sử dụng. Giao diện này không chỉ hỗ trợ tốt cho nhóm mà còn có thể áp dụng rộng rãi trong các nghiên cứu khác.

Từ kết quả phân tích, chúng em đã rút ra được các xu hướng và mối quan hệ quan trọng trong dữ liệu, hỗ trợ cho các nhận định và quyết định trong giáo dục.

Những biểu đồ và báo cáo được tạo ra có thể được sử dụng làm cơ sở cho các nghiên cứu hoặc mô hình dự đoán trong tương lai. Ưu và nhược điểm:

Ưu điểm: Mô hình tự động hóa hiệu quả việc làm sạch và xử lý dữ liệu, giao diện thân thiện, dễ sử dụng, trực quan hóa dữ liệu mạnh mẽ với các biểu đồ đa dạng, minh họa rõ ràng các mối quan hệ trong dữ liệu.

Nhược điểm: Tốc độ xử lý còn hạn chế khi làm việc với các tập dữ liệu lớn, chưa tích hợp các thuật toán học máy để dự đoán và phân loại,

Giao diện còn đơn giản, chưa hỗ trợ các tính năng như tải file động hoặc lưu biểu đồ dưới dạng file hình ảnh.

## 2. Hướng phát triển

* Mở rộng chức năng phân tích.
* Phát triển giao diện người dùng.
* Ứng dụng thực tiễn.

# PHẦN TÀI LIỆU THAM KHẢO

**[1].** Kaggle (2024). *Student Performance.* Truy cập tháng 10/2024, đường dẫn: <https://www.kaggle.com/code/zabihullah18/student-performance/notebook>

**[2].** GeeksforGeeks. *Python Numpy*. Đường dẫn:<https://www.geeksforgeeks.org/python-numpy/>

**[3].** GeeksforGeeks. *Pandas Tutorial*. Đường dẫn: <https://www.geeksforgeeks.org/pandas-tutorial/>

**[4].** GeeksforGeeks. *Matplotlib Tutorial*. Đường dẫn: <https://www.geeksforgeeks.org/matplotlib-tutorial/>

**[5].** GeeksforGeeks. *Python Seaborn Tutorial*. Đường dẫn: <https://www.geeksforgeeks.org/python-seaborn-tutorial/>

**[6].** GeeksforGeeks. *Learning Model Building in Scikit-learn*. Đường dẫn:

[https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machinelearning-library/](https://www.geeksforgeeks.org/learning-model-building-scikit-learn-python-machine-learning-library/)

**[7].** GeeksforGeeks. *PythonTkinter*. Đường dẫn: <https://www.geeksforgeeks.org/python-gui-tkinter/>

**[8].** GeeksforGeeks. *Python Tkinter Tutorial*. Đường dẫn: <https://www.geeksforgeeks.org/python-tkinter-tutorial/>