Thông tin nhóm **0**

1. Thông tin bộ dữ liệu **0**

2. Phân tích và Đặt câu hỏi **2**

3. Cài đặt và Trả lời câu hỏi **2**

3.1. Thống kê số lượng và tỉ lệ phần trăm những người sống sót và tử vong sau tai nạn đắm tàu? 2

3.2. Điều gì giúp gia tăng khả năng sống sót? 3

4. Tổng kết và Nêu phát hiện nổi bật **10**

Thông tin nhóm

Họ tên	**MSSV**	**Công việc**	**Hoàn thành**	**Đóng góp**
Nguyễn Minh Đạt	21127592	Phân tích & Đặt câu hỏi	100%	50%
		Thống kê số lượng và tỉ lệ sống sót	100%	
		Phân tích mối quan hệ giữa giá vé, hạng vé và tỉ lệ sống sót	100%	
Nguyễn Nhật Truyền	21127191	Phân tích & Đặt câu hỏi	100%	50%
		Phân tích mối quan hệ giữa giới tính, độ tuổi đối với tỉ lệ sống sót	100%	
		Phân tích mối quan hệ giữa số lượng người thân đi cùng đối với tỉ lệ sống sót	100%	

1. Thông tin bộ dữ liệu

Tên bộ dữ liệu: Titanic - Machine Learning from Disaster

Nguồn: Kaggle

Cấu trúc:

- Bao gồm các file train.csv và test.csv.
- Trong bài lab này, chúng em chỉ sử dụng file train.csv.

Danh sách thuộc tính:

Tên thuộc tính	**Ý nghĩa**
Survival	Tình trạng sống/tử vong
Pclass	Hạng vé
Sex	Giới tính
Age	Độ tuổi
SibSp	Số lượng anh chị em và vợ/chồng cùng đi trên tàu
Parch	Số lượng bố mẹ và con cái cùng đi trên tàu.
Ticket	Mã số vé
Fare	Giá vé
Cabin	Mã số cabin
Embarked	Cảng khởi hành

Cách thức xử lý dữ liệu:

- Đầu tiên, chúng em lưu bộ dữ liệu từ trên Kaggle về máy tính cá nhân.

- Tiếp theo, chúng em đẩy dữ liệu lên Github Repo của một thành viên trong nhóm (Nguyễn Minh Đạt).
- Cuối cùng, chúng em thực hiện lệnh “git clone" trong file notebook để lưu và sử dụng dữ liệu. Trong đó:
    - Môi trường mà nhóm chọn để viết code là Google Colab.
    - Ngôn ngữ Python phiên bản 3.0.

2. Phân tích và Đặt câu hỏi

Chúng em đặt ra những câu hỏi từ đơn giản đến phức tạp để khám phá và rút trích những thông tin quan trọng từ dữ liệu.

- Thống kê số lượng và tỉ lệ phần trăm những người sống sót, và tử vong sau tai nạn đắm tàu?
- Điều gì giúp gia tăng khả năng sống sót?
    - Liệu giá vé và hạng vé có ảnh hưởng đến khả năng sống sót?
    - Liệu giới tính và độ tuổi có liên quan đến khả năng sống sót?
    - Liệu đi cùng gia đình, người thân (bố, mẹ, vợ, chồng, anh, chị, em) có giúp gia tăng khả năng sống sót?

3. Cài đặt và Trả lời câu hỏi

3.1. Thống kê số lượng và tỉ lệ phần trăm những người sống sót và tử vong sau tai nạn đắm tàu?

Thông tin cài đặt:

- Chúng em đọc dữ liệu từ file train.csv và lưu dữ liệu dưới dạng Dataframe (train_df).
- Chúng em tạo hai Dataframe mới từ Dataframe ban đầu, bao gồm:
    - Một Dataframe lưu thông tin những người sống sót.

    - Một Dataframe lưu thông tin những người tử vong.
- Chúng em thực hiện đếm số dòng. Trong đó:
    - Số dòng có thuộc tính Survived = 0 thể hiện số người tử vong.
    - Số dòng có thuộc tính Survived = 1 thể hiện số người sống sót.

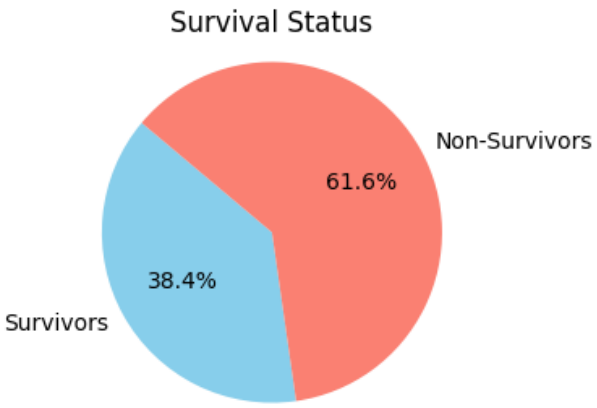
```
non_survivors = train_df[train_df['Survived'] == 0]
survivors = train_df[train_df['Survived'] == 1]
print("Number of people who did not survive:", len(non_survivors))
print("Number of people who survived:", len(survivors))
```

Kết quả:

- Số người tử vong: 549 (chiếm 61.6%)
- Số người sống sót: 342 (chiếm 38.4%)

Chúng em lựa chọn biểu đồ tròn để trực quan kết quả thu được.

- Lý do: Biểu đồ tròn thích hợp để biểu diễn tỉ lệ, đặc biệt trong trường hợp số lượng lớp ít. Từ đó giúp cho người dùng dễ dàng quan sát và so sánh.



3.2. Điều gì giúp gia tăng khả năng sống sót?

Thuộc tính: Hạng vé

- Thông tin cài đặt:
    - Dựa vào hai Dataframe lưu thông tin những người tử vong và sống sót, chúng em đếm số lượng từng loại hạng vé từ mỗi Dataframe.

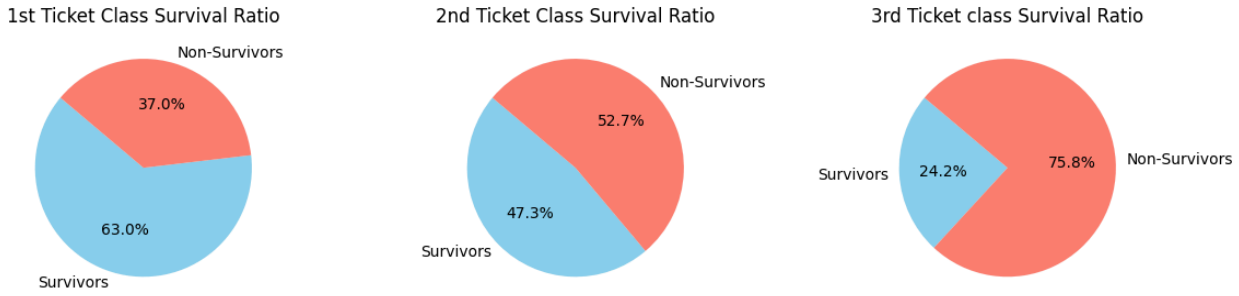
```
survivor_ticket_class1 = survivors[survivors['Pclass'] == 1]
```

```
survivor_ticket_class2 = survivors[survivors['Pclass'] == 2]
survivor_ticket_class3 = survivors[survivors['Pclass'] == 3]
non_survivor_ticket_class1 = non_survivors[non_survivors['Pclass'] == 1]
non_survivor_ticket_class2 = non_survivors[non_survivors['Pclass'] == 2]
non_survivor_ticket_class3 = non_survivors[non_survivors['Pclass'] == 3]
```

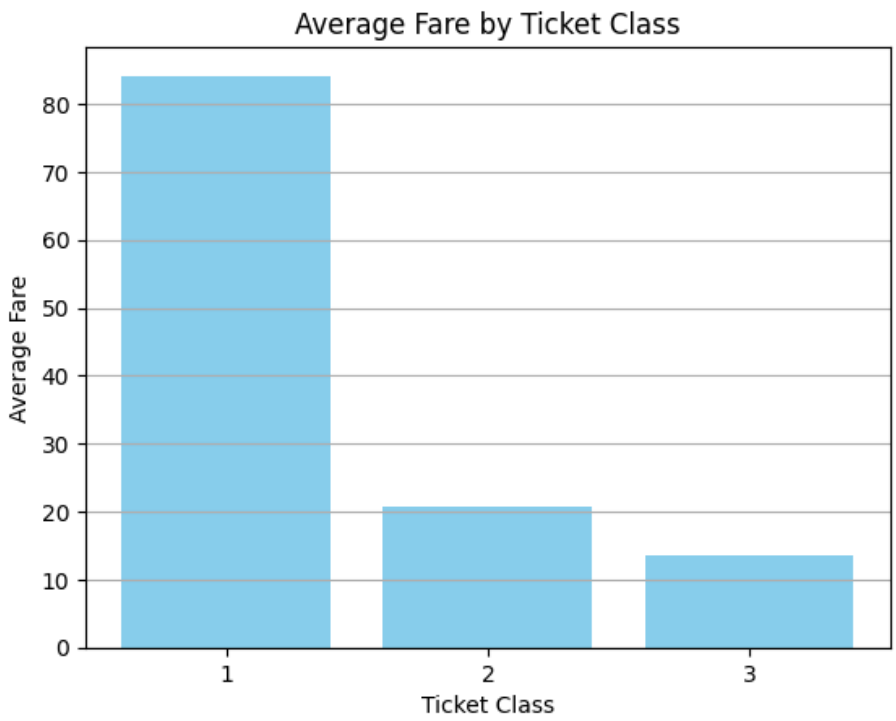
- Kết quả:

	Vé hạng 1	**Vé hạng 2**	**Vé hạng 3**
Sống sót	136 (63.0%)	87 (47.3%)	119 (24.2%)
Tử vong	80 (37.0%)	97 (52.7%)	372 (75.8%)

- Chúng em sử dụng biểu đồ tròn để trực quan kết quả với lý do tương tự như đã trình bày ở trên.



- Điều tra giá vé trung bình của từng hạng vé:



- Nhận xét xu hướng:
    - Hạng vé cao giúp gia tăng khả năng sống sót.
    - Tất nhiên, chúng ta phải trả nhiều tiền để có được hạng vé cao.

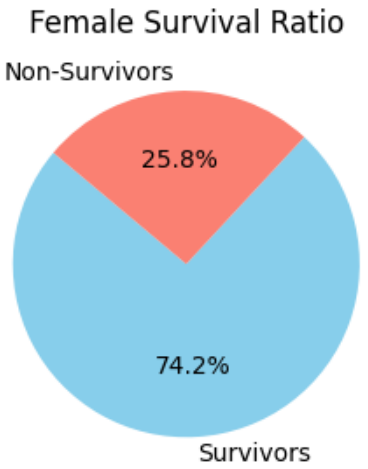
Thuộc tính: Giới tính

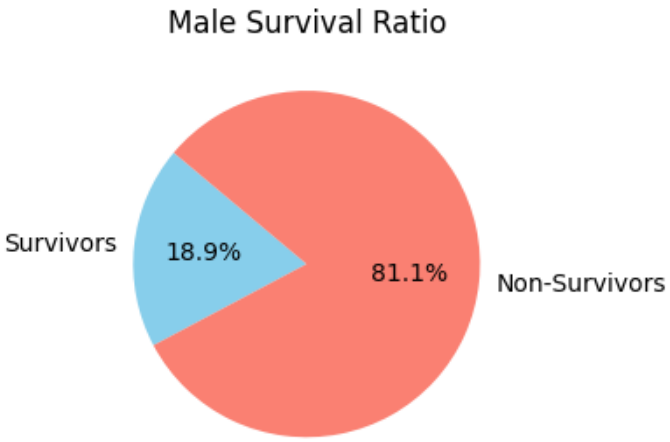
- Thông tin cài đặt:
    - Dựa vào hai Dataframe lưu thông tin những người tử vong và sống sót, chúng em đếm số lượng nữ giới và nam giới từ mỗi Dataframe.

```
female_survivors = survivors[survivors['Sex'] == 'female']
male_survivors = survivors[survivors['Sex'] == 'male']
female_non_survivors = non_survivors[non_survivors['Sex'] == 'female']
male_non_survivors = non_survivors[non_survivors['Sex'] == 'male']
```

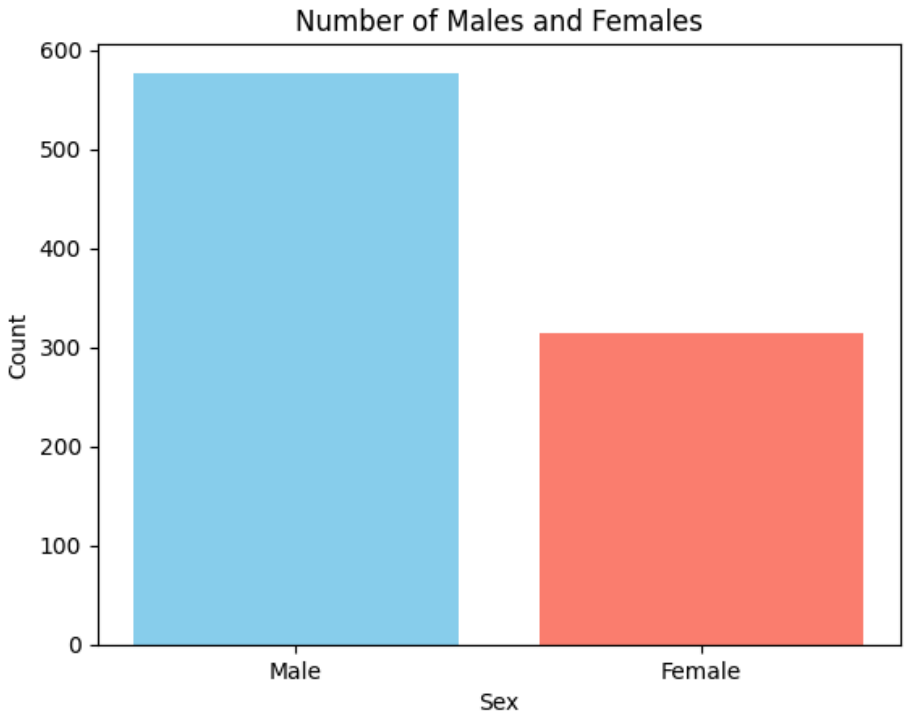
- Kết quả:
    - Nữ giới:
        - Số lượng tử vong: 81 (chiếm 25.8%)
        - Số lượng sống sót: 233 (chiếm 74.2%)

- Nam giới:
  - Số lượng tử vong: 468 (chiếm 81.1%)
  - Số lượng sống sót: 109 (chiếm 18.9%)





- Điều tra số lượng nam giới và nữ giới:
  - Tổng số lượng nữ giới: 314
  - Tổng số lượng nam giới: 577



- Nhận xét:

- Số lượng nam giới trên tàu nhiều gần gấp đôi so với số lượng nữ giới.
- Tuy nhiên, tỉ lệ tử vong ở nam giới là trên 80%
- Trong khi đó, tỉ lệ tử vong ở nữ giới là chưa đến 26%.
- Như vậy, giới tính có liên quan tới khả năng sống sót sau tai nạn.

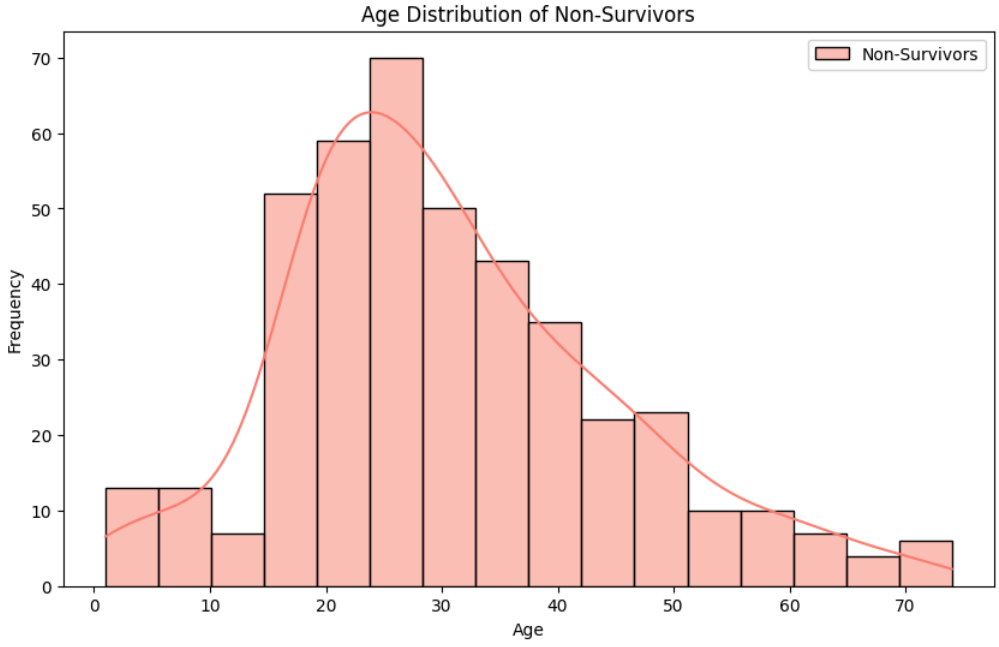
Thuộc tính: Độ tuổi

- Thông tin cài đặt:
  - Dựa vào hai Dataframe lưu thông tin những người tử vong và sống sót, chúng em tạo ra hai Dataframe mới để lưu thông tin tuổi tác của từng nhóm người.

```
age_survivors = survivors['Age'].dropna()  # Drop NaN values
age_non_survivors = non_survivors['Age'].dropna()
```

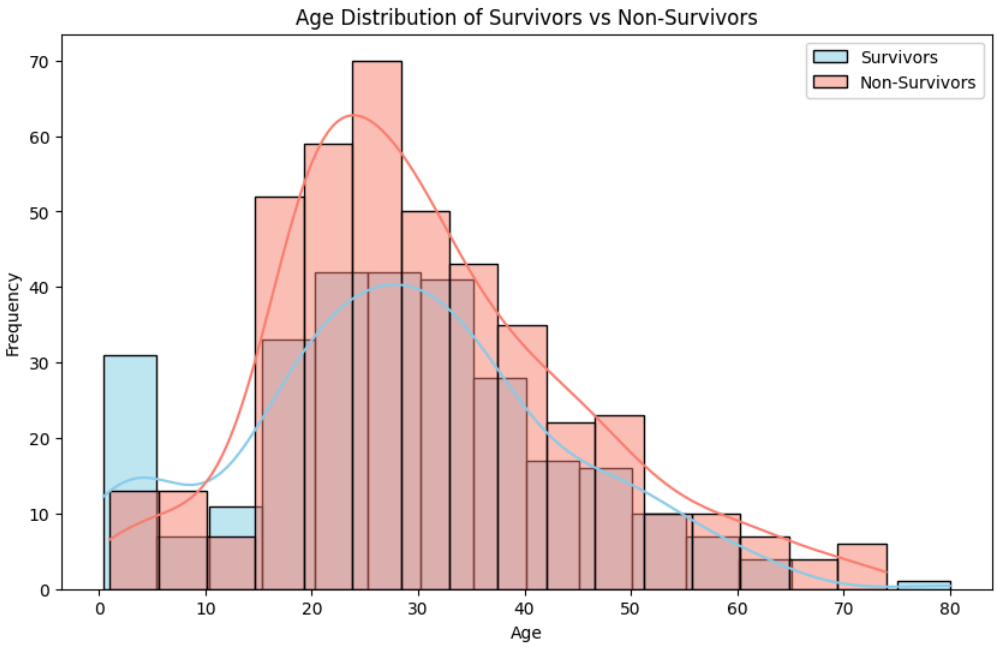
- Mục đích trực quan: Biểu diễn phân phối độ tuổi của hai nhóm người.
  - Nhóm những người tử vong.
  - Nhóm những người sống sót.
- Kết quả:
  - Phân phối độ tuổi của nhóm những người sống sót.

- Phân phối độ tuổi của nhóm những người tử vong



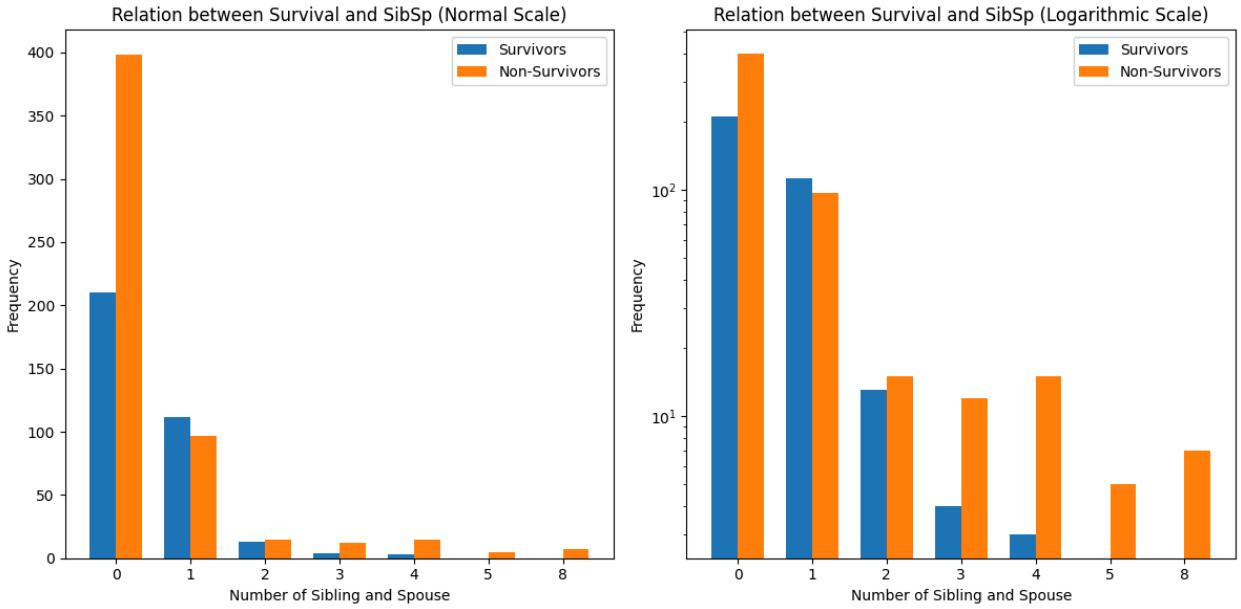
- Nhận xét:
  - Trẻ em dưới 10 tuổi có khả năng sống sót cao.
  - Tuy nhiên nhìn chung, phân phối độ tuổi không có nhiều đặc trưng khác biệt giữa hai nhóm người.
  - Độ tuổi có thể không ảnh hưởng nhiều đến khả năng sống sót.

- So sánh mối quan hệ giữa hai phân phối trên cùng một biểu đồ:



Thuộc tính: Số lượng anh/chị/em, vợ/chồng cùng đi trên tàu

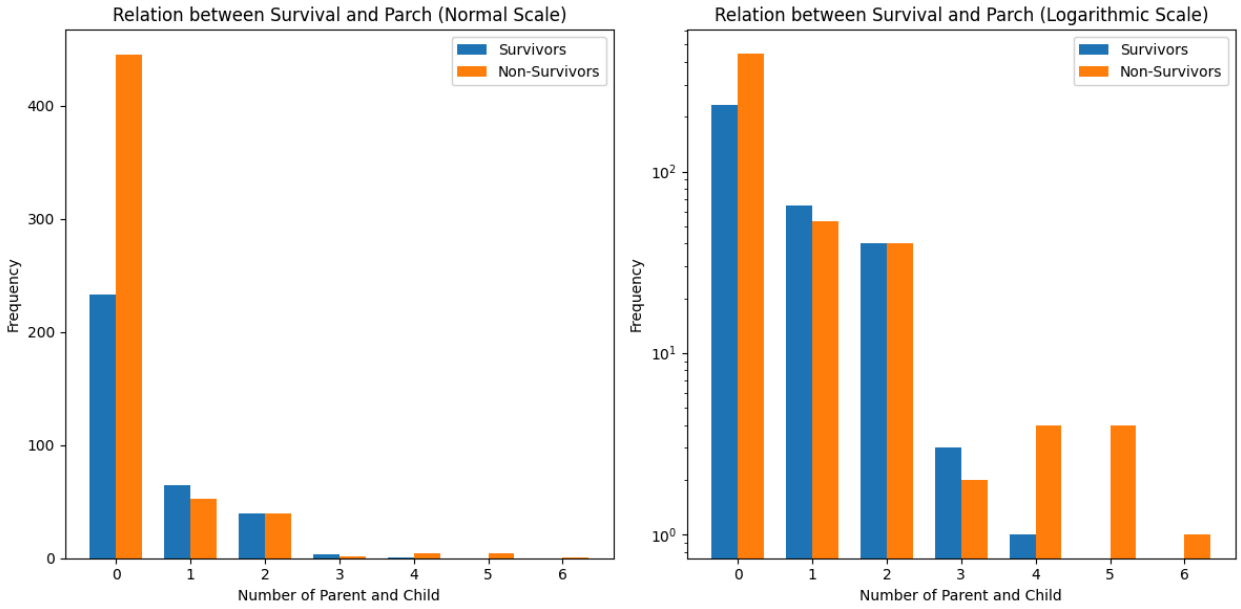
- Thông tin cài đặt: Sử dụng biểu đồ cột ghép với hai tỉ lệ biểu diễn.
    - Tỉ lệ bình thường: So sánh các tần suất có giá trị lớn, dễ quan sát.
    - Tỉ lệ logarit: So sánh các tần suất có giá trị nhỏ, khó quan sát ở tỉ lệ bình thường.
- Kết quả:



- Nhận xét:
    - Không đi với bất kì ai trong số anh, chị, em, vợ, chồng có tỉ lệ tử vong cao, gần như gấp đôi so với tỉ lệ sống sót.
    - Đi cùng 1 người trong số anh/chị/em, vợ/chồng có tỉ lệ sống sót nhỉnh hơn tỉ lệ tử vong.
    - Đi cùng từ 3 người trở lên trong số anh/chị/em, vợ/chồng có tỉ lệ tử vong cao nhất.
    - Như vậy, lựa chọn đi cùng từ 1 người trong số anh/chị/em, vợ/chồng là tốt nhất.

Thuộc tính: Số lượng bố mẹ và con cái cùng đi trên tàu.

- Thông tin cài đặt: Sử dụng biểu đồ cột ghép với hai tỉ lệ biểu diễn.
    - Tỉ lệ bình thường: So sánh các tần suất có giá trị lớn, dễ quan sát.
    - Tỉ lệ logarit: So sánh các tần suất có giá trị nhỏ, khó quan sát ở tỉ lệ bình thường.
- Kết quả:



- Nhận xét:
    - Không đi cùng bất kì ai trong số bố mẹ và con cái có tỉ lệ tử vong cao gần gấp đôi so với tỉ lệ sống sót.
    - Đi cùng 1 hoặc 3 người trong số bố/mẹ/con cái có tỉ lệ sống sót nhỉnh hơn tỉ lệ tử vong.
    - Đi cùng từ 4 người trở lên trong số bố/mẹ/con cái có tỉ lệ tử vong cao nhất.
    - Như vậy, đi cùng từ 1 hoặc 3 người trong số bố/mẹ/con cái là lựa chọn tốt nhất.

4. Tổng kết và Nêu phát hiện nổi bật

Công nghệ:

- pandas:
    - Lưu dữ liệu (Dataframe)
    - Tính toán trên Dataframe
- matplotlib, seaborn:
    - Vẽ biểu đồ

Phát hiện:

- Trên 60% số người đã tử vong sau tai nạn.
- Hạng vé càng cao, tương ứng với giá vé càng cao và khả năng sống sót càng cao.
- Nữ giới có khả năng sống sót cao hơn nam giới.
- Tuổi tác gần như không ảnh hưởng đến khả năng sống sót.
- Đi cùng với 1 đến 3 người thân (bố mẹ, con cái, anh chị em, vợ chồng) giúp gia tăng khả năng sống sót.