ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

TRỰC QUAN HÓA DỮ LIỆU

< Lab 01 - TITANIC >

Sinh viên thực hiện: 21127115 - Trần Thanh Ngân
21127229 - Dương Trường Bình
21127616 - Lê Phước Quang Huy

Giảng viên hướng dẫn: TS. Bùi Tiến Lên

Lớp: 21KHDL

Mục lục

Thông tin nhóm và phân công công việc

MSSV	Họ và tên	Công việc được phân công	Mức độ hoàn thành
21127115	Trần Thanh Ngân	• A. Mô tả dữ liệu • Trả lời câu hỏi 1, 2	100%
21127229	Dương Trường Bình	• EDA dữ liệu Numerical • Trả lời câu hỏi 3, 4	100%
21127616	Lê Phước Quang Huy	• EDA dữ liệu Categorical • Trả lời câu hỏi 5, 6	100%

Tiến độ công việc

Phần	Nội dung	Mức độ hoàn thành
A. Mô tả dữ liệu	1. Đếm số dòng và số cột.	100%
	2. Viết bảng mô tả về các cột.	100%
	3. Phân tích tỷ lệ missing rate.	100%
	4. Fill missing rate.	100%
B. EDA	1. Dữ liệu numerical	100%
	2. Dữ liệu categorical	100%
C. Phân tích dữ liệu	4. Đặt câu hỏi về bộ dữ liệu và trả lời	100%

1 Mô tả dữ liệu

Tổng quan

Bộ dữ liệu Titanic chứa thông tin về dân số và hành khách từ 891 trong số 2224 hành khách và phi hành đoàn trên tàu Titanic. Tổng cộng có **1,309** dòng và **11** cột *(sau khi bỏ đi cột Survived)*. Ý nghĩa của từng cột được thể hiện ở bảng sau:

STT	Tên thuộc tính	Mô tả	Giá trị	Kiểu dữ liệu
1	PassengerId	Mã định danh của hành khách		Integer
2	Pclass	Hạng vé	1 = hạng 1, 2 = hạng 2, 3 = hạng 3	Integer
3	Name	Tên hành khách		String
4	Sex	Giới tính		String
5	Age	Tuổi		Decimal
6	SibSp	Số anh chị em / vợ chồng trên tàu		Integer
7	Parch	Số cha mẹ / con cái trên tàu		Integer
8	Ticket	Số vé		String
9	Fare	Giá vé hành khách		Decimal
10	Cabin	Số cabin		String
11	Embarked	Cảng lên tàu	C = Cherbourg, Q = Queenstown, S = Southampton	String

Tỷ lệ missing rate

Thuộc tính	Tổng dữ liệu bị thiếu	Tỉ lệ thiếu dữ liệu (%)
Cabin	1014	77.5
Age	263	20.1
Embarked	2	0.2
Fare	1	0.1
PassengerId	0	0.0
Pclass	0	0.0
Name	0	0.0
Sex	0	0.0
SibSp	0	0.0
Parch	0	0.0
Ticket	0	0.0

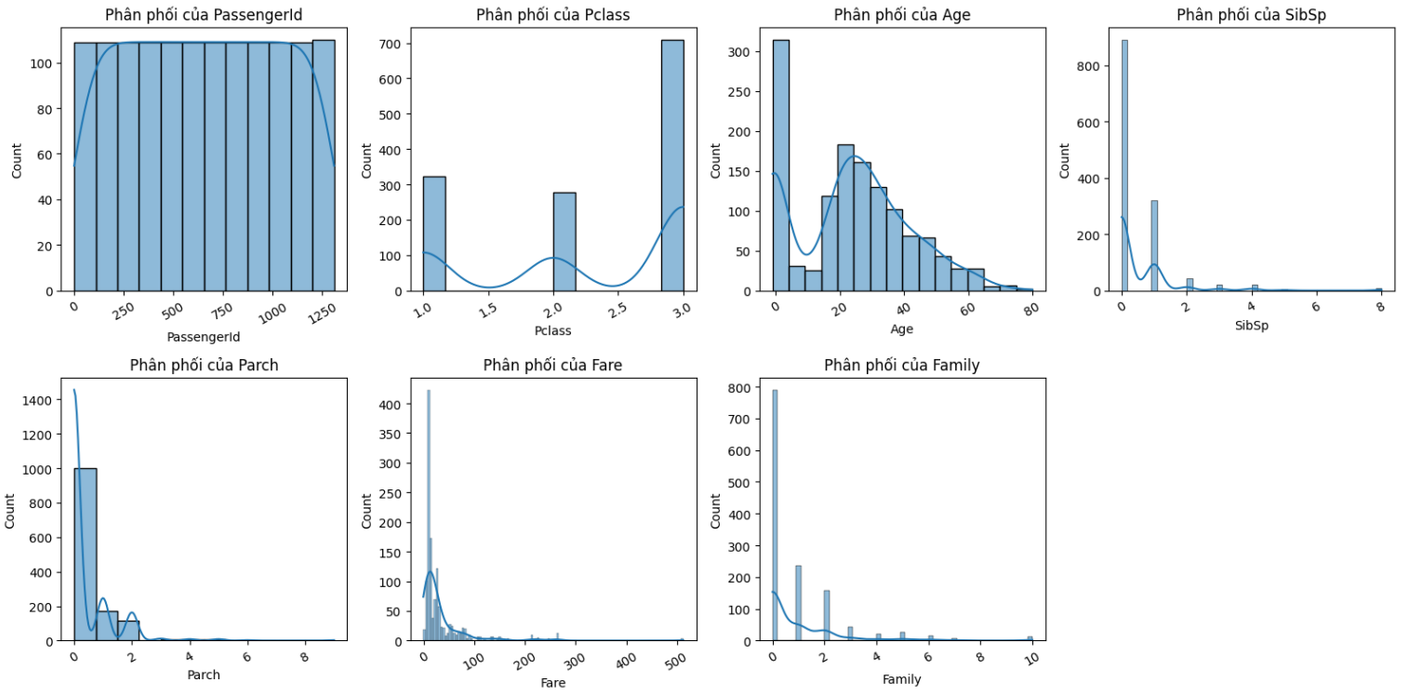
Sau khi tính tỉ lệ missing ở các cột dữ liệu, nhóm đã điền vào giá trị `-1` để có thể tiếp tục phân tích

2 EDA cho các biến

Sau khi tách các biến ra thành 2 loại numerical và categorical. Nhóm tiến hành phân tích cho từng loại.

Phân tích phân phối đối với biến numerical

- Nhóm dùng hàm `describe` để tính toán các giá trị thống kê mô tả cho các biến: `min, max, mean, std, 25%,50%,75%,`
- Nhóm sử dụng histogram để trực quan phân phối cho tất cả các biến numerical được thể hiện ở Hình **??**



Nhận xét:

- **PassengerId**: Bộ dữ liệu hiện tại chứa thông tin về 1043 hành khách.
- **Pclass**: Hạng vé trung bình là khoảng 2.2, cho thấy hầu hết hành khách ở hạng vé thứ hai hoặc thứ ba.
- **Age**: Tuổi trung bình của hành khách là khoảng 29.8 tuổi, với độ lệch chuẩn là 14.37. Độ tuổi dao động từ 0.17 (khoảng 2 tháng) đến 80 tuổi.
- **SibSp**: Trung bình, hành khách có khoảng 0.5 anh chị em hoặc vợ chồng trên tàu, với số lớn nhất là 8.
- **Parch**: Trung bình, hành khách có khoảng 0.42 bố mẹ hoặc con cái trên tàu, với số lớn nhất là 6.
- **Fare**: Giá trung bình mà hành khách trả là $36.60, với độ lệch chuẩn là $55.75. Giá vé thấp nhất là $0.00, và giá vé cao nhất là $512.33.

- **Family**: Trung bình, hành khách có khoảng 0.9 người thân trên tàu, với số lớn nhất là 10.

Phân tích phân phối đối với biến categorical

Nhóm dùng hàm `describe` để tính toán các giá trị thống kê mô tả cho các biến: `unique`, `top`, `freq`.

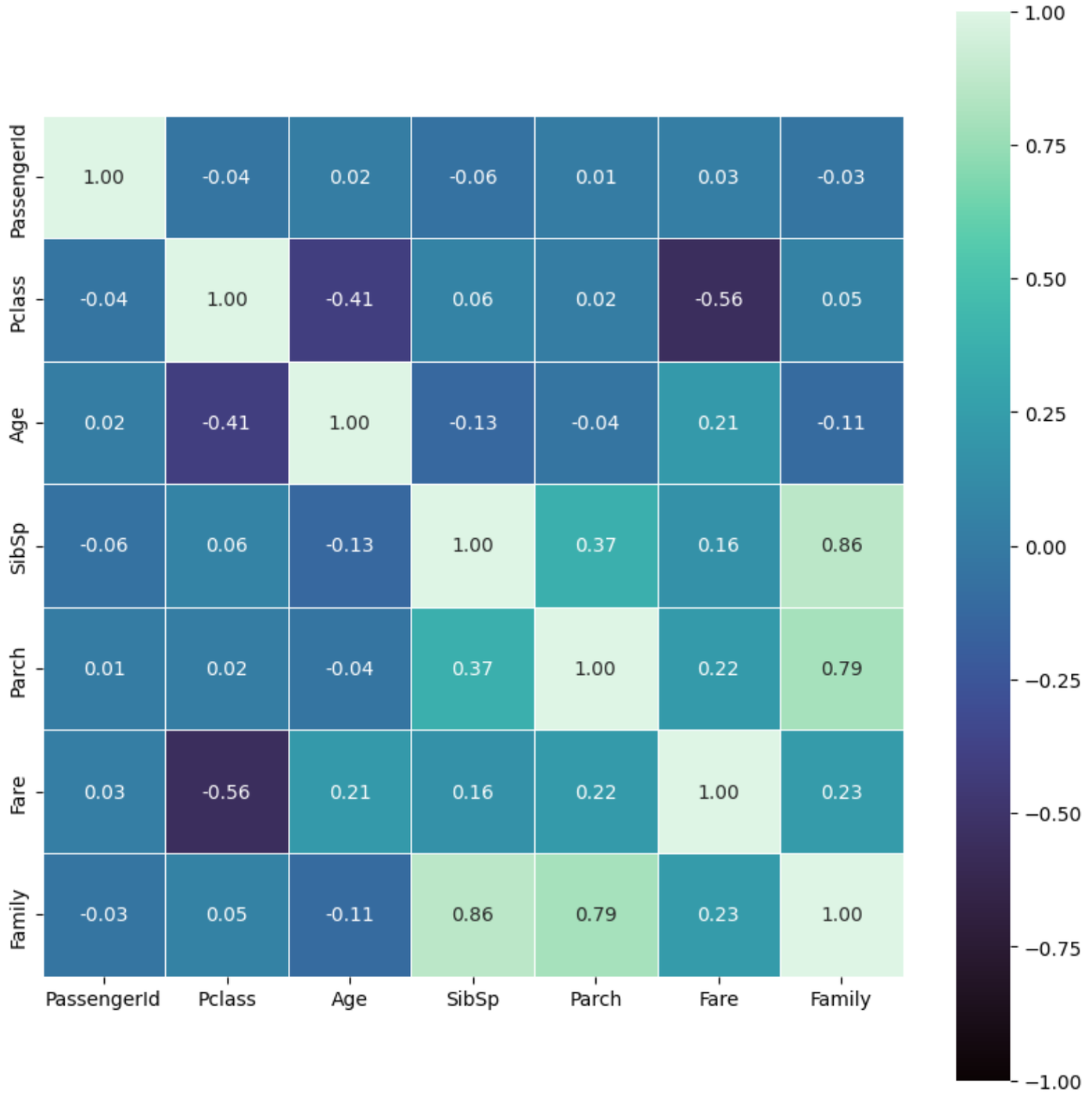
	Name	Sex	Ticket	Cabin	Embarked
count	1309	1309	1309	1309	1309
unique	1307	2	929	187	4
top	Connolly, Miss. Kate	male	CA. 2343	-1	S
freq	2	843	11	1014	914

Nhận xét:

- **Name**: Có 1307 giá trị duy nhất trong số 1309 mẫu dữ liệu, chỉ có 2 mẫu dữ liệu trùng lặp. Mỗi hành khách có một tên duy nhất, tuy nhiên có thể có các trường hợp trùng tên.
- **Sex**: Chỉ có 2 giá trị duy nhất là ’male’ và ’female’, được chia thành 843 nam và 466 nữ.
- **Ticket**: Có 929 giá trị duy nhất trong số 1309 mẫu dữ liệu. Có một số vé trùng lặp, đặc biệt là vé ’CA. 2343’ với tần suất xuất hiện cao nhất là 11 lần.
- **Cabin**: Có 187 giá trị duy nhất trong số 1309 mẫu dữ liệu. Có 1014 mẫu dữ liệu không có thông tin về phòng (-1).
- **Embarked**: Có 3 giá trị về cảng là ’S’, ’C’, ’Q’, và 1 giá trị không xác định. Cảng S có tần suất xuất hiện cao nhất với 914 mẫu dữ liệu.

Phân tích tương quan giữa các thuộc tính numerical

Nhóm sử dụng biểu đồ heatmap để trực quan ma trận tương quan cho các biến numerical



Nhận xét: Nhìn vào biểu đồ ta nhận thấy có các biến tương quan với nhau khá cao bao gồm:

- `SibSp` và `Family`
- `Parch` và `Family`

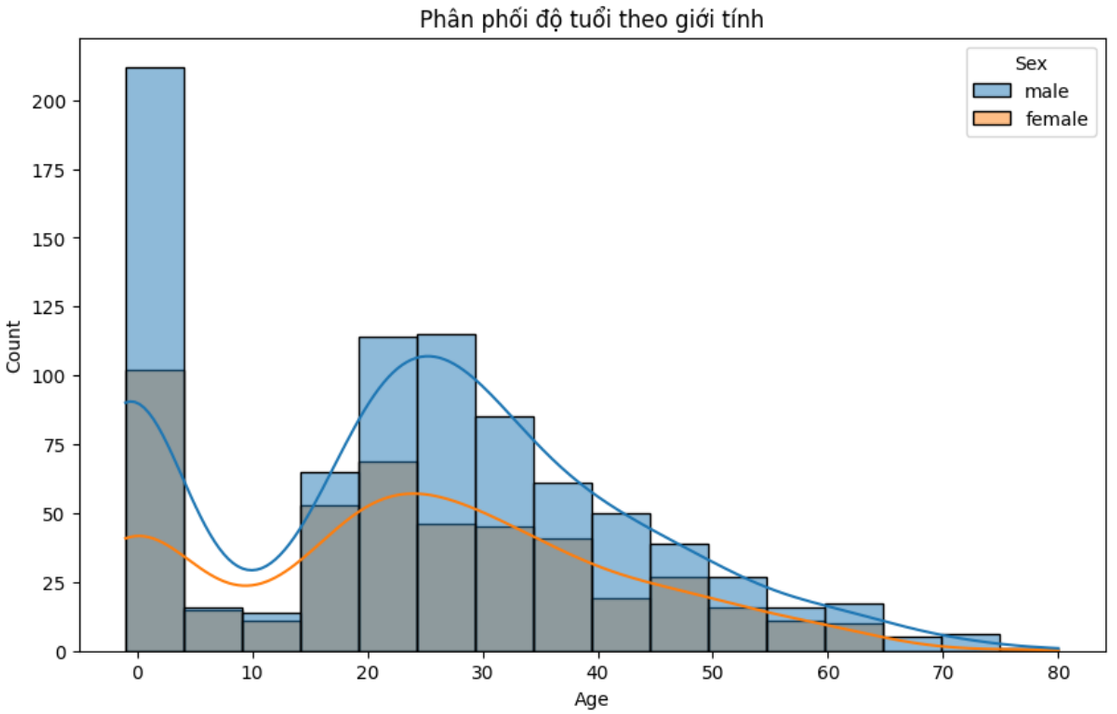
3 Phân tích dữ liệu

3.1 Câu hỏi 0: Thông tin về thuyền trưởng

Thuyền trưởng là một người đàn ông có tên là `Edward Gifford` tuổi 70.

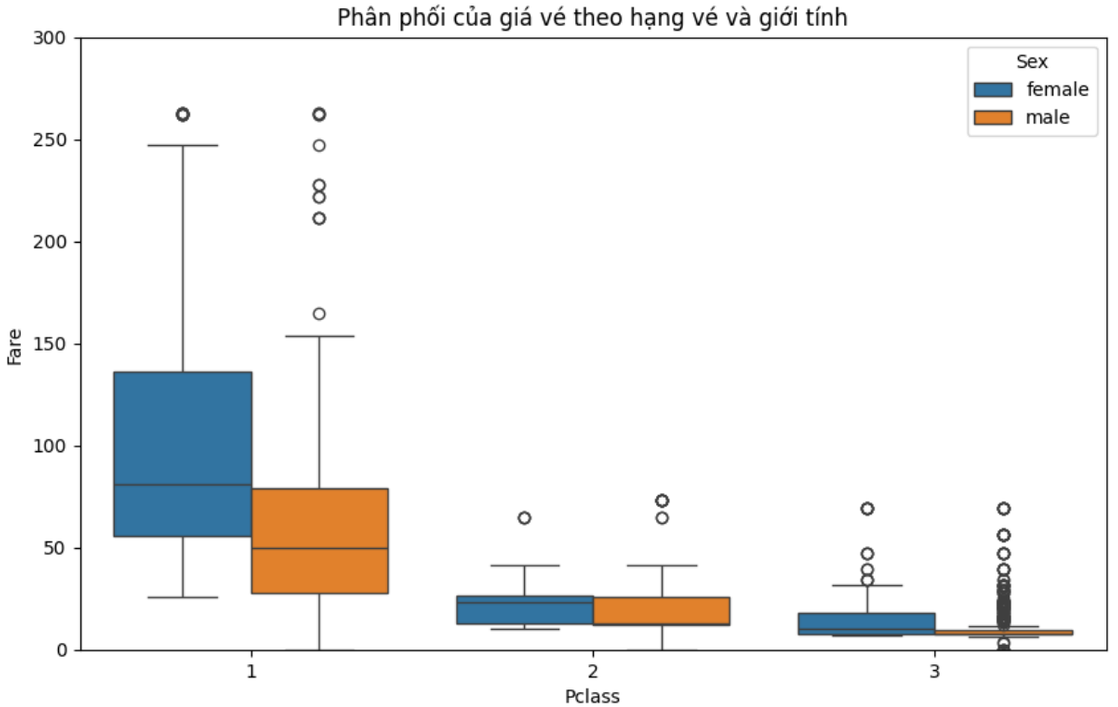
	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Family
745	746	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0	B22	S	2

3.2 Câu hỏi 1: Phân phối độ tuổi của hành khách theo giới tính là như thế nào và liệu có sự khác biệt đáng kể giữa nam và nữ không?



- Dựa vào biểu đồ ta thấy được, Số lượng hành khách nam nhiều hơn số lượng hành khách nữ nhưng tỷ lệ độ tuổi ở 2 giới của hành khách khá cân bằng

3.3 Câu hỏi 2: Liệu có sự ảnh hưởng nào giữa giới tính đối với giá vé và hạng vé không?



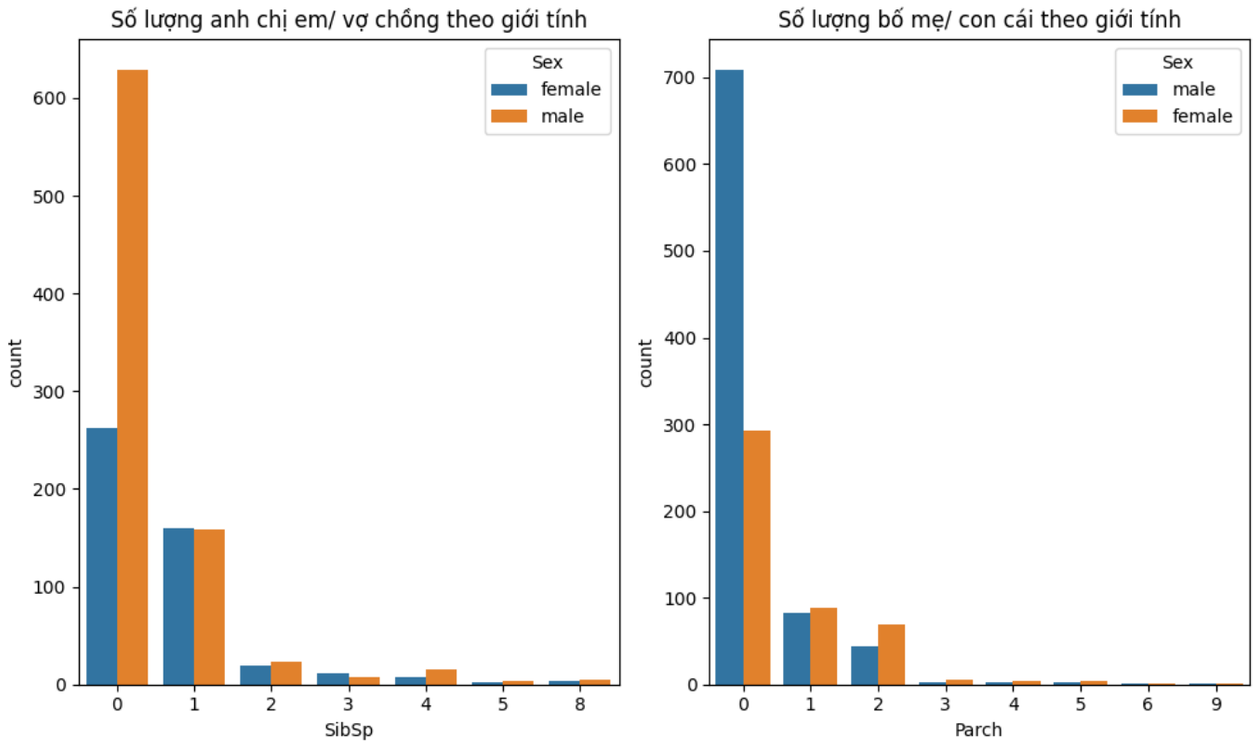
Dựa vào biểu đồ này, có một số nhận xét có thể thấy về mối quan hệ giữa giới tính, hạng vé, và giá vé:

- 'Hạng 1': Phân phối giá vé cho cả nam và nữ có xu hướng cao hơn so với các hạng khác, với giá trung vị của nữ giới cao hơn nam giới. Giá vé của nữ giới cũng có sự biến động rộng lớn hơn so với nam giới, như thể hiện qua khoảng từ Q1 đến Q3 , cùng các điểm nằm ngoại lai ở phía trên.
- 'Hạng 2': Giá vé giữa nam và nữ tỏ ra ít biến động hơn so với hạng 1. Giá trung vị giữa hai giới tính khá gần nhau, và không có sự khác biệt lớn về phạm vi giá giữa hai giới.
- 'Hạng 3': Phân phối giá vé đã rất chặt chẽ và ít biến động hơn so với hai hạng trước. Tuy nhiên, cũng như hạng 1 và 2, giá trung vị của nữ giới cao hơn nam giới một chút. Các điểm nằm ngoại lai cho thấy có một số vé giá rất cao hoặc rất thấp so với phần còn lại.

Từ những nhận xét trên, có thể kết luận rằng có một sự khác biệt nhất định giữa giá vé mà nam và nữ phải trả theo từng hạng vé, đặc biệt là trong hạng 1, nơi có sự biến động lớn về giá

vé giữa hai giới. Tuy nhiên, sự khác biệt này không rõ ràng ở hạng 2 và 3.

3.4 Câu hỏi 3: Phân tích mối tương quan giữa giới tính đối với số anh chị em/ vợ chồng và bố mẹ/ con cái mà một hành khách sẽ dẫn theo?



Từ biểu đồ "Số lượng anh chị em/vợ chồng theo giới tính"chúng ta có thể thấy:

- Phần lớn hành khách, cả nam và nữ, không có anh chị em hoặc vợ/chồng đi cùng (giá trị là 0).
- Số lượng nữ giới có một anh chị em hoặc vợ/chồng đi kèm là cao hơn so với số lượng nam giới.
- Số lượng giảm dần khi số lượng anh chị em/vợ chồng tăng lên, với nam giới có xu hướng giảm nhanh hơn so với nữ giới.

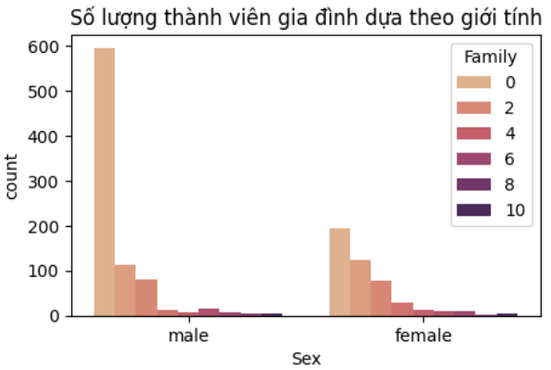
Từ biểu đồ "Số lượng bố mẹ/con cái theo giới tính"ta có thể nhận thấy:

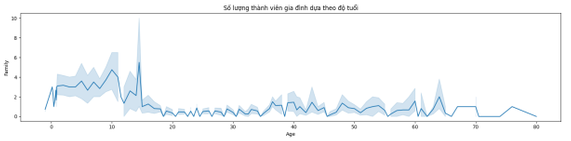
- Số lượng hành khách nam không có bố mẹ hoặc con cái (giá trị là 0) đi cùng là cao nhất, sau đó là nữ giới.

- Số lượng nữ giới có một người (có thể là bố/mẹ hoặc con cái) đi kèm là đáng kể và cao hơn so với nam giới.
- Số lượng tiếp tục giảm khi số người bố mẹ/con cái tăng lên, với việc hầu hết hành khách không có hơn 3 người bố mẹ/con cái đi cùng.
- Cũng giống như biểu đồ SibSp, xu hướng giảm số lượng cũng nhanh hơn ở nam giới so với nữ giới khi số lượng người bố mẹ/con cái tăng lên.

Những xu hướng trên có thể bị ảnh hưởng bởi nhiều yếu tố xã hội và văn hoá vào thời điểm dữ liệu được thu thập, chẳng hạn như vai trò giới trong gia đình, xu hướng kết hôn hoặc có con, cũng như các quy định về du lịch và di cư vào thời đó.

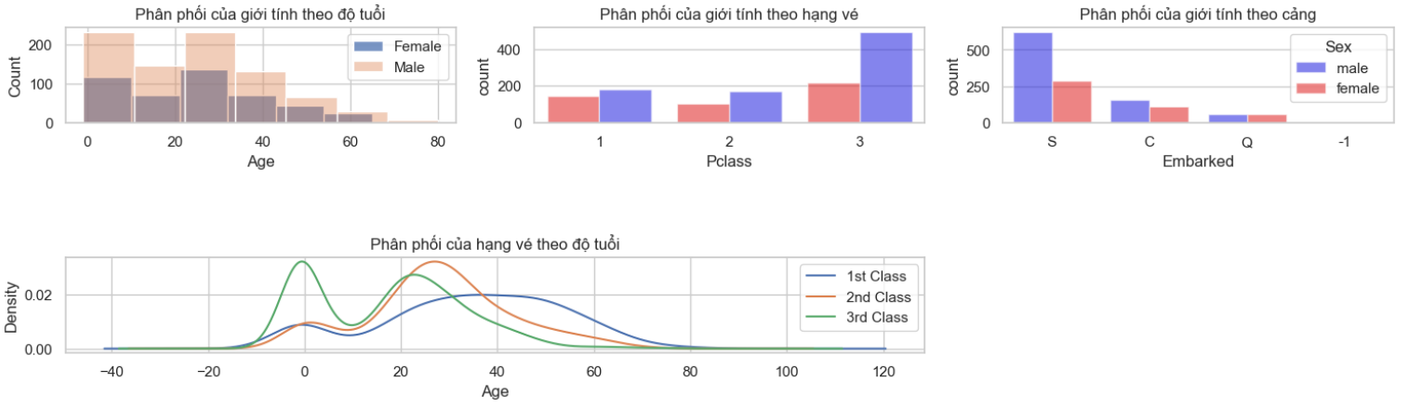
3.5 Câu hỏi 4: Mỗi hành khách nam và nữ có xu hướng dẫn theo bao nhiêu thành viên trong gia đình? Từ đó so sánh với mối quan hệ giữa độ tuổi và số lượng thành viên gia đình?





- Dựa vào biểu đồ phía dưới ta thấy ở độ tuổi dưới 20 có số lượng thành viên gia đình lớn hơn so với phần còn lại.
- Đặc biệt ở giới tính Nam số lượng 0 thành viên gia đình rất lớn có gần 600. Còn ở các mức số lượng thành viên khác thì cả nam và nữ đều ở mức ngang nhau

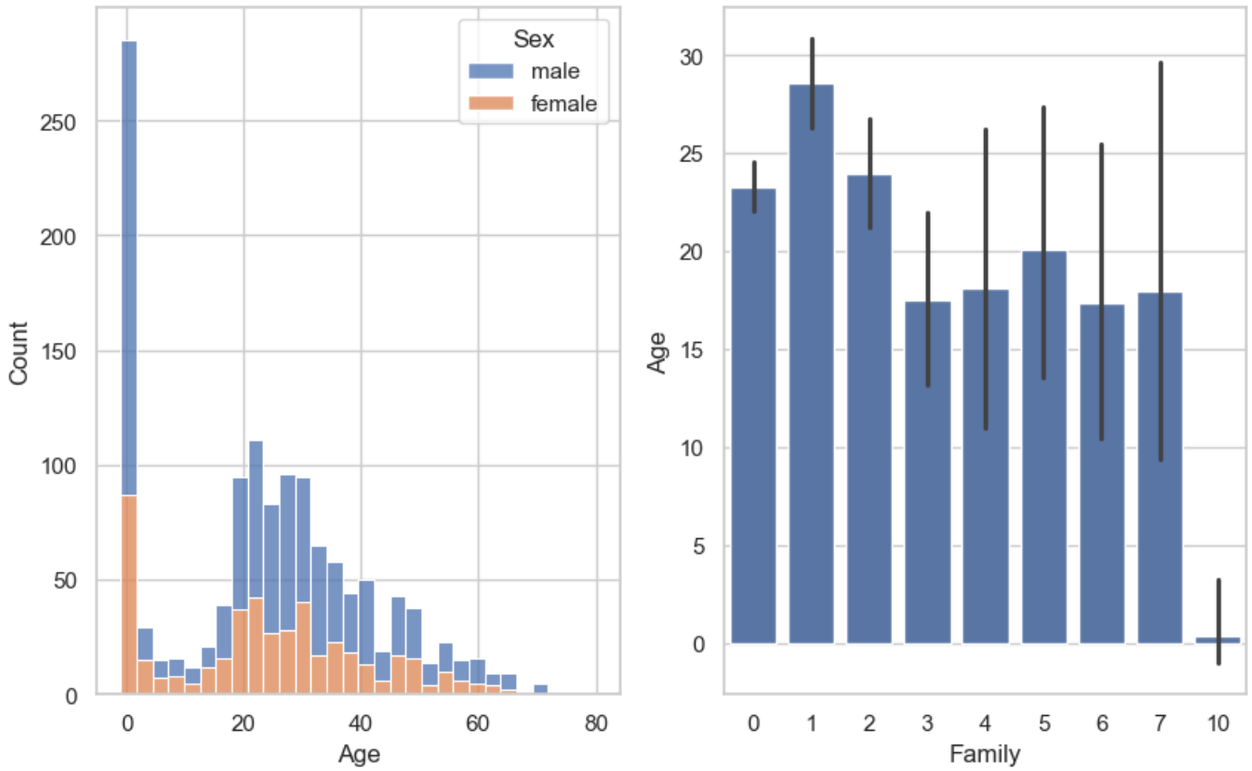
3.6 Câu hỏi 5: Phân tích mối liên hệ giữa Giới tính, Độ tuổi trong việc lựa chọn Hạng vé và Cảng?



Từ biểu đồ trên, chúng ta có thể thấy một số mối liên hệ giới tính và độ tuổi trong lựa chọn hạng vé và cảng khởi hành như sau:

- Hạng vé có thể phản ánh khả năng tài chính và tuổi tác. Người lớn tuổi có khả năng cao hơn chọn hạng vé cao cấp (1st Class), có thể do họ có nhiều tài chính hơn. Trong khi đó, hành khách trẻ tuổi hơn có xu hướng chọn hạng ba (3rd Class), có thể do mức giá phải chăng hơn hoặc do họ chấp nhận ít tiện nghi hơn.
- Phân phối độ tuổi giữa nam và nữ không có sự khác biệt đáng kể nào rõ ràng từ biểu đồ, nhưng nó cho thấy rằng phụ nữ có xu hướng tập trung vào một nhóm tuổi nhỏ hơn (20-40) so với nam giới.
- Sự lựa chọn cảng khởi hành có thể chịu ảnh hưởng bởi giới tính, với phụ nữ khởi hành từ cảng C chiếm tỉ lệ cao hơn so với các cảng khác. Điều này có thể liên quan đến các yếu tố văn hóa, kinh tế, hoặc thậm chí là lịch sử di cư tại cảng đó.

3.7 Câu hỏi 6: Phân tích mối quan hệ giữa Giới tính, Độ tuổi đối với số thành viên gia đình?



- Biểu đồ bên trái cho thấy sự phân bố độ tuổi theo giới tính. Cả nam và nữ đều có độ tuổi từ 0 đến khoảng 80. Cột màu xanh biểu thị cho nam và màu cam biểu thị cho nữ. Có vẻ như có một số lượng lớn trẻ em ở độ tuổi từ 0 đến 10 (đặc biệt là ở cột cao nhất của trẻ em sơ sinh và trẻ em dưới 5 tuổi). Số lượng người nam trong các nhóm độ tuổi này có vẻ cao hơn hơn so với người nữ. Tuy nhiên, ở độ tuổi từ 20 đến 40, số lượng nữ có vẻ cao hơn so với nam. Số lượng người giảm dần ở các nhóm độ tuổi cao hơn.
- Biểu đồ bên phải cho thấy mối quan hệ giữa độ tuổi trung bình và số thành viên gia đình. Trục ngang (Family) biểu thị cho số thành viên trong gia đình từ 0 đến 10, và trục dọc (Age) biểu thị cho độ tuổi trung bình. Đường kẻ trên mỗi cột biểu định khoảng tin cậy hoặc phân tán của độ tuổi (có thể là độ lệch chuẩn hoặc khoảng cách từ 25