Analisi di dati

Chiara Bettini Eleonora Rossi

Introduzione

Il dataset era costituito dai dati di passeggeri del Titanic. Era composto da 891 tuple, per 12 colonne. Di queste dodici colonne, 'PassengerId', 'Age', 'SibSp', 'Parch', 'Fare' erano quantitative, discrete; 'Class' era qualitativa ordinabile; 'Name', 'Cabin' e 'Ticket' erano qualitative e non ordinabili, e 'Survived', 'Sex' e 'Embarked' erano delle variabili qualitative non ordinabili e categoriali.

Abbiamo analizzato le variabili, esaminandone anche la distribuzione rispetto ad altre variabili. Abbiamo però escluso dalle analisi 'PassengerId' e 'Name', e analizzato le variabili 'Cabin' e 'Ticket' solo in modo limitato.

PassengerId	Id numerico univoco
Survived	0 = non sopravvissuto, 1 = sopravvissuto
Pclass	1 = prima classe, 2 = seconda classe, 3 = terza classe
Name	Nome del passeggero
Sex	{male, female}
Age	Età (float)
SibSp	Numero (intero) di fratelli/sposi
Parch	Numero (intero) di genitori/figli
Ticket	Codice del biglietto (non univoco)
Fare	Prezzo
Cabin	Nome della cabina, se presente/conosciuta (non univoco)
Embarked	Luogo di imbarco; S = Southampton, Q = Queenstown, C = Cherbourg

Statistica descrittiva

Descrizione dei passeggeri

I passeggeri presenti nel dataset erano il 64,8% maschi e il 35,2% femmine, con una schiacciante maggioranza di maschi. I passeggeri erano a loro volta suddivisi in classi, con il 24,2% in prima classe, il 20,7% in seconda, e la maggior parte, il 55,1%, in terza. La composizione delle classi prima e seconda rispetto al sesso era abbastanza uniforme (fig. 1), con una prevalenza leggera di maschi, mentre per la terza classe la prevalenza dei maschi sulle femmine era schiacciante.

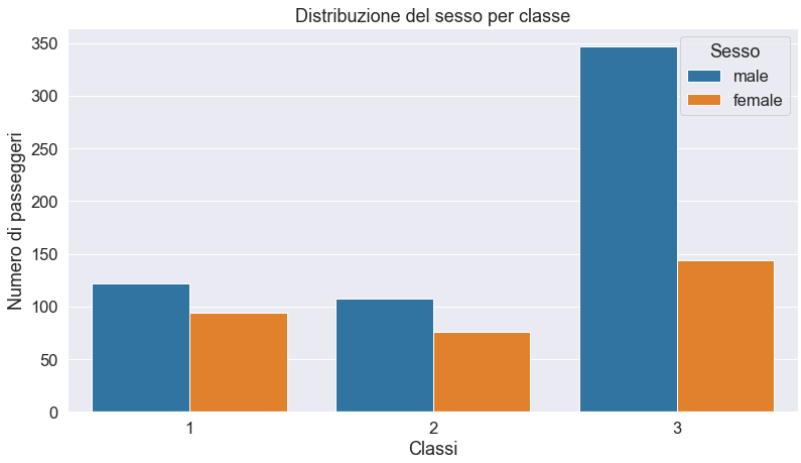

Figura 1

Abbiamo analizzato i dati riguardanti le cabine, calcolando che solo il 22,9% risultasse avere una cabina assegnata nel dataset. Tuttavia, non conoscendo informazioni specifiche sull'organizzazione della nave e avendo constatato che nel dataset sono presenti dati parzialmente mancanti per alcuni passeggeri, bisogna considerare la possibilità che la grossa quantità di dati nulli per l'attributo 'Cabin' possa essere dovuta anche ad una

mancanza di informazioni. In questa fase abbiamo constatato inoltre che alcune cabine erano condivise da più passeggeri.

Abbiamo quindi osservato le età dei passeggeri. La media per il dataset era di 29,7 anni, con una deviazione standard di 14,5. Tuttavia, osservando il grafico (fig. 2) si nota immediatamente che la distribuzione delle età non rispetta una distribuzione normale[1]. Per quanto molti dati si attestino attorno alla media, e media e mediana siano simili, c'era anche una significativa porzione di bambini nel dataset. Erano invece presenti solo pochi passeggeri sopra i sessant'anni.

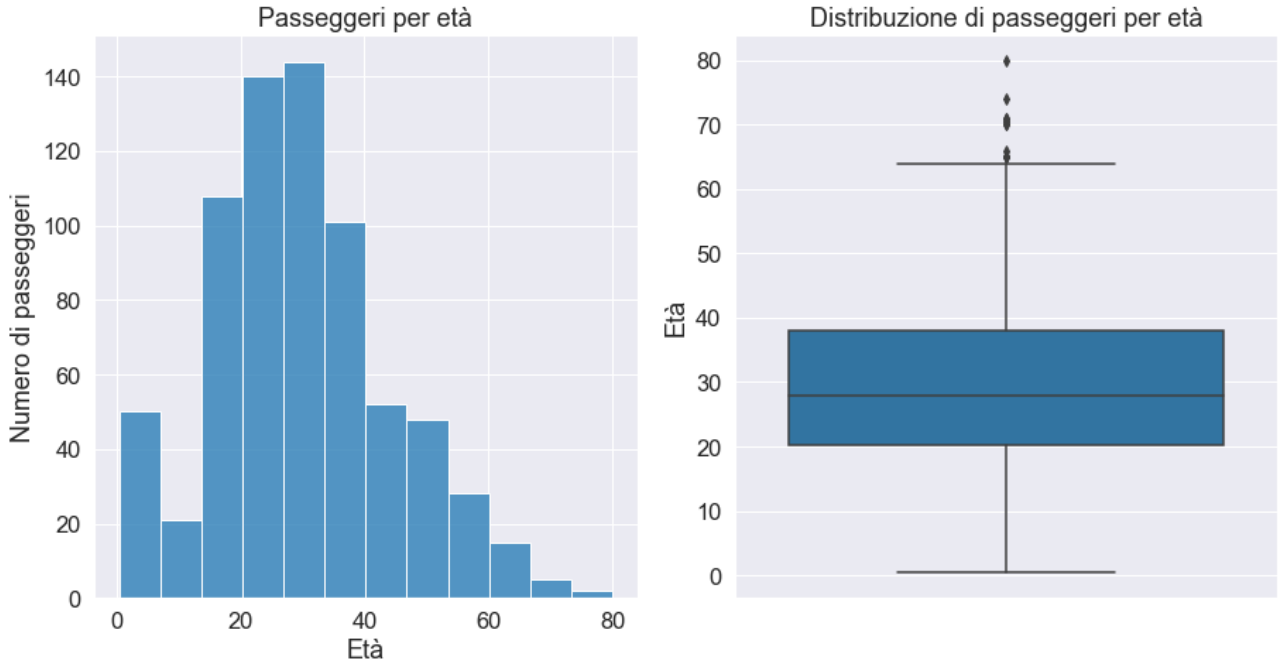

Figura 2. Descrizione età dei passeggeri

Quasi i tre quarti dei passeggeri si erano imbarcati a Southampton, il 72,4%. I passeggeri imbarcati a Queenstown erano il 18,9%, e quelli imbarcati a Cherbourgh erano solo l'8,7%.

Descrizione dei sopravvissuti

Durante l'analisi della distribuzione dei sopravvissuti (e non) siamo partiti dall'ipotesi secondo cui l'indice di sopravvivenza potesse essere correlato alla classe di appartenenza e al sesso dei passeggeri.

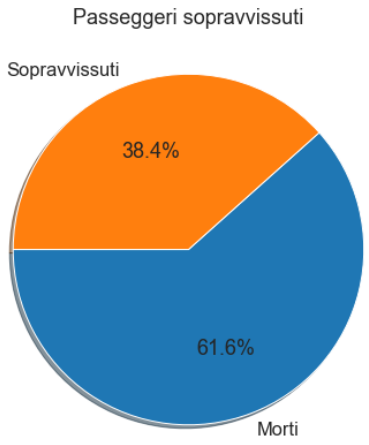

Figura 3

Dal grafico (fig. 3) possiamo osservare che in generale più della metà dei passeggeri non è sopravvissuta all'incidente. Analizzando poi il rapporto della variabile 'Survived' rispetto alla variabile 'Pclass', possiamo inferire che la maggior parte delle persone morte provenivano dalla terza classe (67,8%); a seguire abbiamo i viaggiatori della seconda (17,7%) ed in maniera sensibilmente più bassa quelli di prima (14,6%). Questo sembrerebbe spiegabile dal fatto che c'erano molti più passeggeri di terza classe.

Analizzando i sopravvissuti in base alla classe, possiamo vedere che nonostante ci sia una maggiore percentuale di sopravvissuti nella prima classe (34,8%), lo scarto con le altre due diminuisce sensibilmente (25,4% per la seconda e 39,8% per la terza). I grafici (figg. 4 e 5) mostrano però quanto nella terza classe il rapporto tra passeggeri sopravvissuti e morti sia molto sbilanciato con una percentuale nettamente maggiore di morti, mentre nella prima

[1] Confermato anche effettuando un test di Shapiro-Wilk su R.

abbiamo una tendenza quasi speculare, con una percentuale di morti nettamente inferiore rispetto a quella dei sopravvissuti. Questo probabilmente è dovuto al fatto che in fase di salvataggio veniva data maggiore precedenza alle classi più abbienti, oppure al fatto che spesso gli alloggi della terza classe erano collocati in luoghi della nave con accesso meno diretto al ponte dove si trovavano le scialuppe.

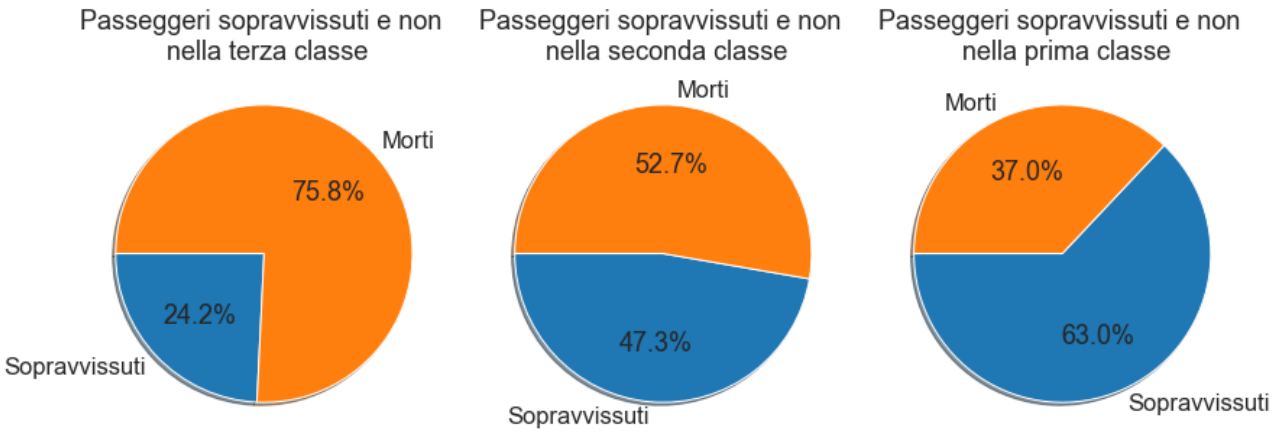

Figura 4

Figura 5

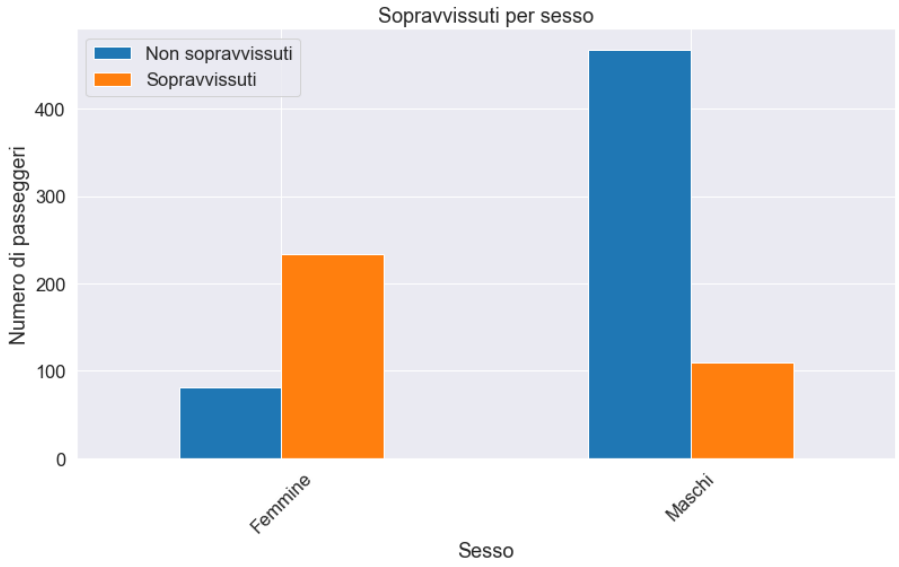

Figura 6

Per quanto riguarda il confronto tra passeggeri sopravvissuti (e non) e il sesso, i dati mostrano che, in effetti, è stata data la precedenza alle donne (fig. 6); tuttavia, il dislivello così grande dei non sopravvissuti tra maschi e femmine potrebbe essere dovuto anche al fatto che in generale la maggior parte dei passeggeri presenti fossero di sesso maschile.

Da figura 7 possiamo inferire che la maggior parte dei sopravvissuti come anche dei morti si era imbarcata nel porto di Southampton. Ma crediamo che questo sia semplicemente dovuto al fatto che in generale la maggior parte dei passeggeri proveniva da tale porto.

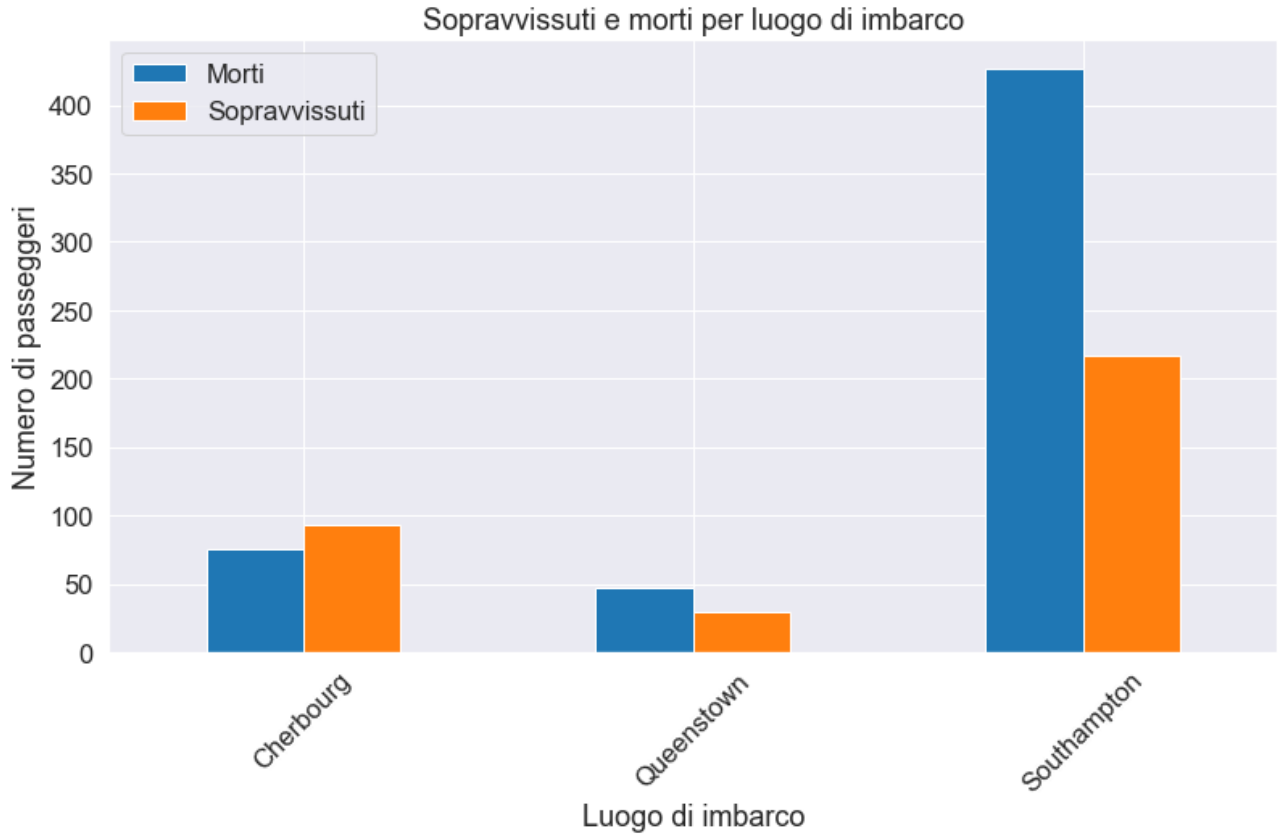

Figura 7

Descrizione dell'età

Sempre in base all'ipotesi di partenza relativa ai sopravvissuti, abbiamo pensato che anche l'età potesse essere una condizione determinante per la sopravvivenza o meno dei passeggeri. Tuttavia, il boxplot (fig. 8) mostra come la differenza non sia così rilevante: la mediana dell'età dei sopravvissuti e dei morti coincide,

come la metà centrale della distribuzione, nonostante quella dei morti sia leggermente più alta rispetto a quella dei sopravvissuti.

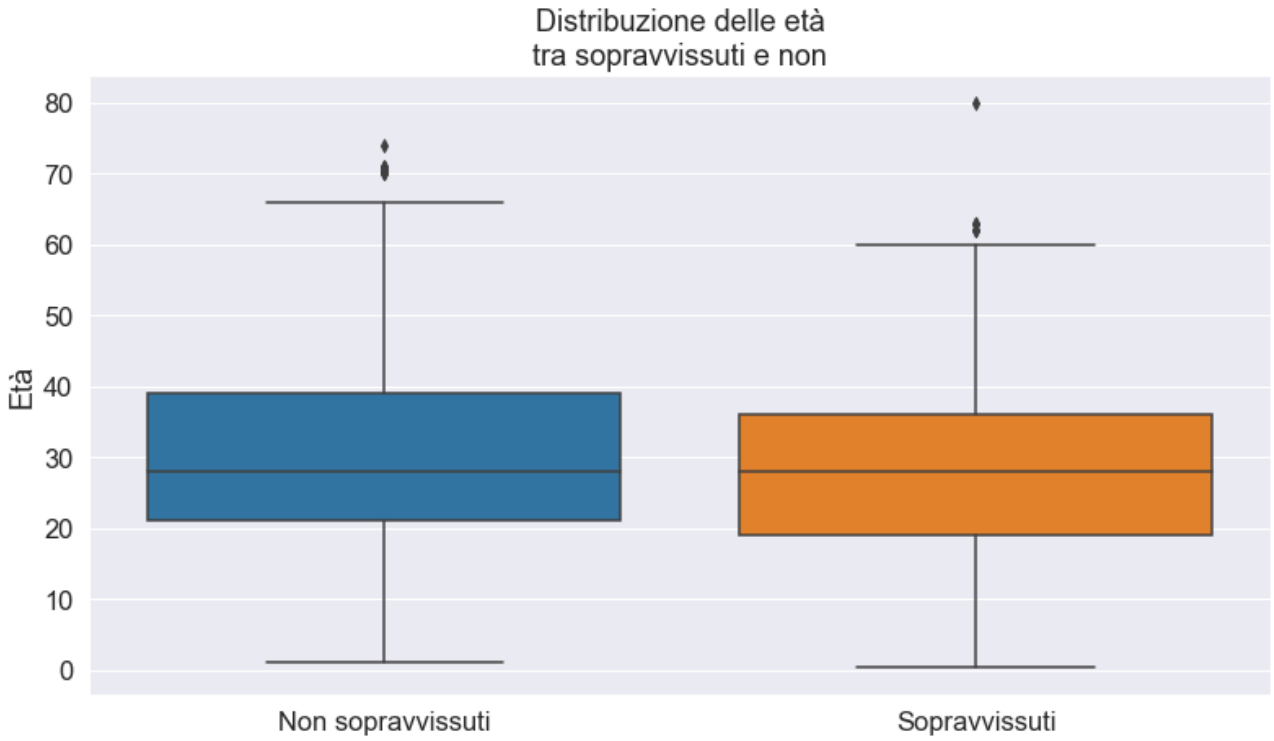

Figura 8

Dal grafico relativo all'età rispetto alla classe (fig. 9) è possibile vedere un andamento discendente: in prima classe la maggior parte dei passeggeri avevano un'età compresa tra i 27 anni e i 50, mentre nella terza classe erano presenti maggiormente passeggeri giovani, tra i 18 e i 32.

La distribuzione mostra che in generale i passeggeri di sesso femminile erano leggermente più giovani di quelli di sesso maschile, ma non crediamo che questo dato sia significativo.

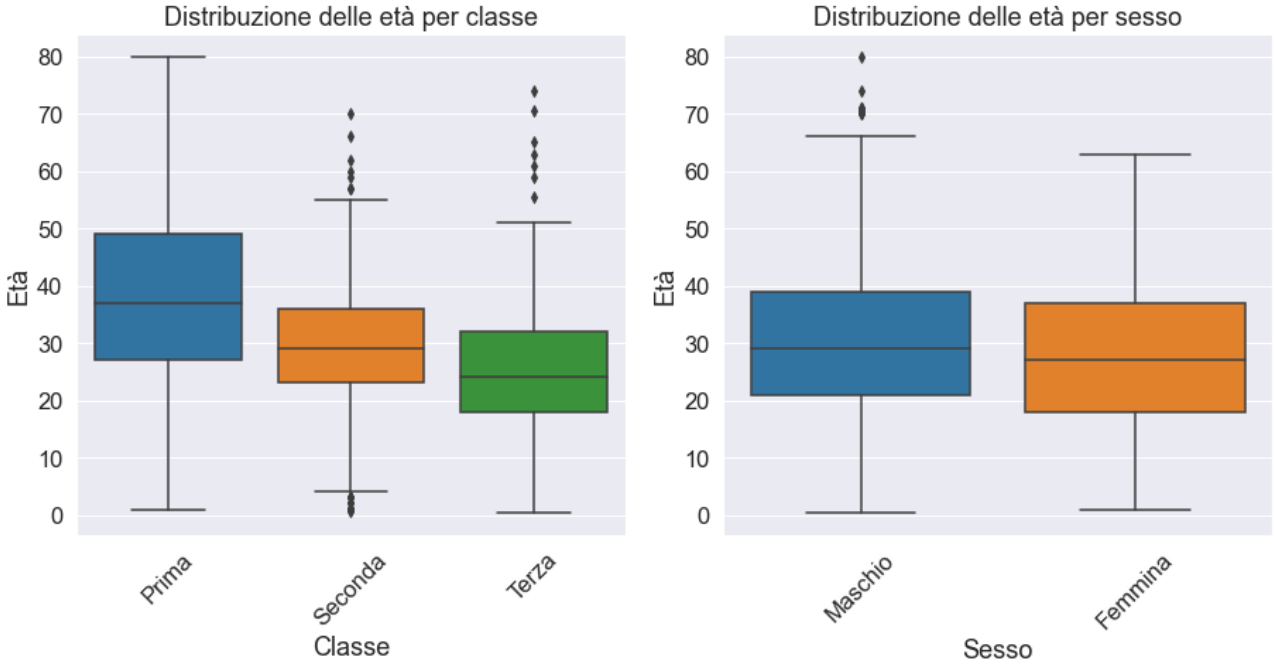

Figura 9

Infine, il grafico di figura 10 mostra quanto i viaggiatori imbarcati a Queenstown fossero tendenzialmente più giovani rispetto a quelli imbarcati negli altri porti, anche se la mediana dei tre boxplot è più o meno equivalente; questo andamento può essere spiegato dal fatto che la quasi totalità dei passeggeri provenienti dal porto di Queenstown erano appartenenti alla terza classe (cfr. paragrafo successivo), dove abbiamo visto esserci una maggiore distribuzione di passeggeri giovani.

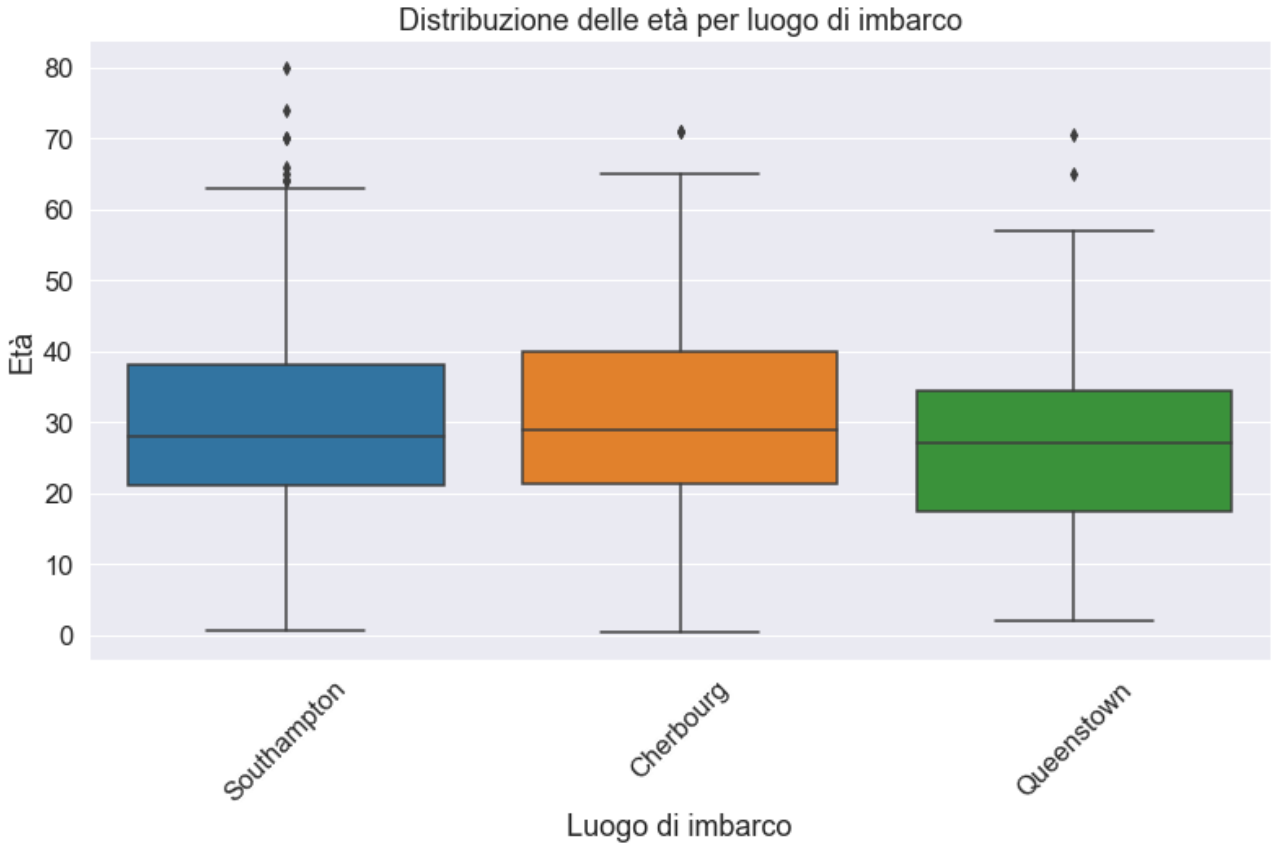

Figura 10

Descrizione delle classi

Da un'analisi superficiale dei dati (e tenuto conto di quanto discusso su 'Cabin' nel paragrafo Passeggeri) avevamo ipotizzato che i passeggeri con cabina fossero unicamente di prima classe. Questo però non ha avuto conferma dai dati: sebbene l'80% delle cabine fosse di passeggeri di prima classe, era presente una ridotta frazione di passeggeri con cabina di seconda o terza classe.

Ci siamo quindi chiesti se potesse esserci un numero diverso di passeggeri per cabina in base alla classe, immaginando che passando da una classe più "ricca" ad una più modesta aumentasse il numero di persone con cui i passeggeri condividevano la cabina. Questo non ha trovato totale conferma nei dati: la media di persone per cabina della prima classe era 1,32, il valore minore di tutti, ma per la seconda era 2,29 e per la terza 1,71.

È stato infine analizzato il dato della classe in base al luogo di imbarco. Ci chiedevamo se le classi fossero distribuite similmente nei vari luoghi di imbarco. Come si può vedere dal grafico (fig. 11), a Cherbourgh si nota un numero di imbarchi di passeggeri della prima classe maggiori di quelli di terza classe, in contrasto con la tendenza osservata a Queenstown e a Southampton. Si può inoltre notare come gli imbarchi a Queenstown fossero di passeggeri di terza classe in modo schiacciante rispetto a quelli delle altre classi.

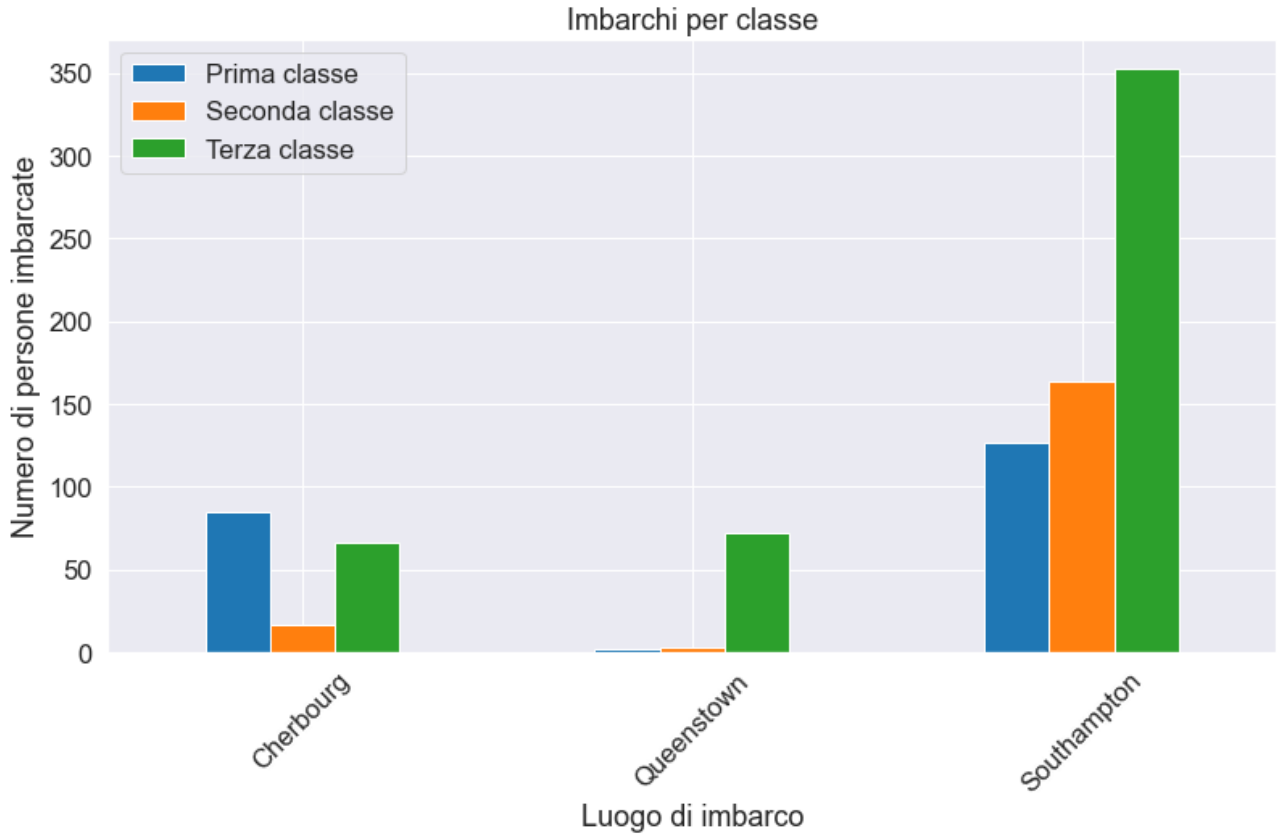

Figura 11

Descrizione dei parenti

La media di fratelli/sposi per passeggero era di 0,53, e di genitori/figli per passeggero di 0,38: solo il 31,8% dei passeggeri aveva fratelli o sposi indicati nel dataset, e solo il 23,9% genitori o figli.

I passeggeri che avevano fratelli o sposi tra i passeggeri ne avevano nella maggior parte dei casi solo uno. In misura molto minore erano indicate due, quattro o tre persone (fig. 12).

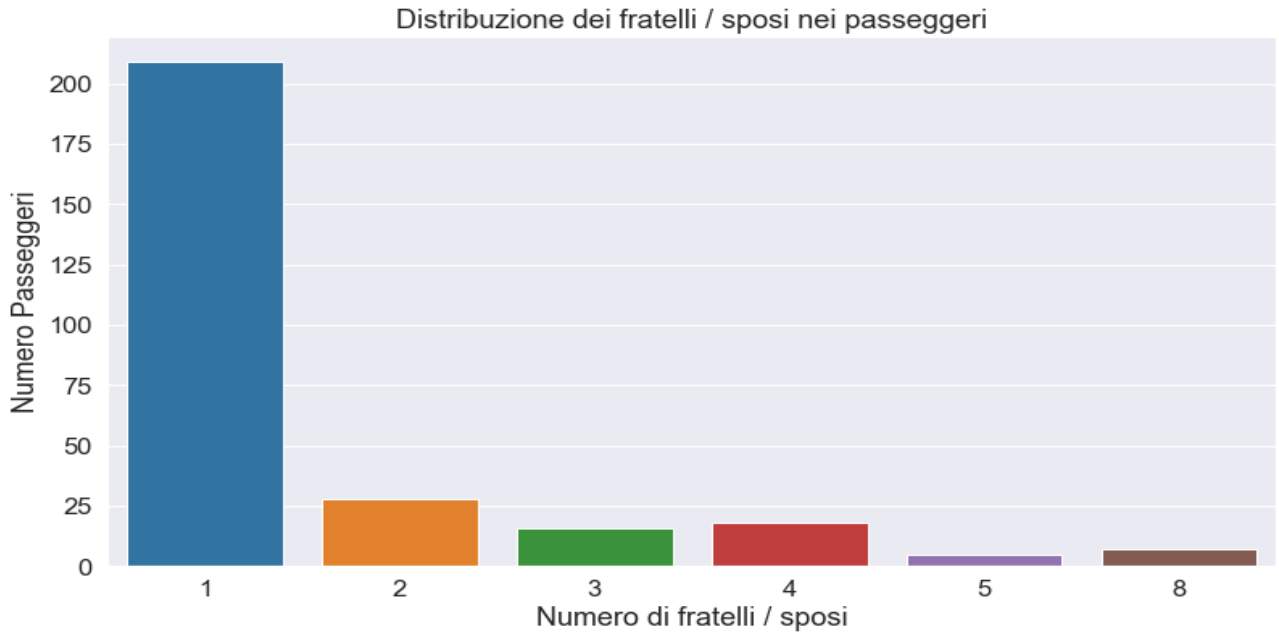

Figura 12

Diversa la situazione nel caso di genitori o figli: in questo caso, sebbene i passeggeri con un solo genitore o figlio a bordo fossero in numero maggiore, una frazione rilevante ne aveva due (fig. 13).

Analizzando le distribuzioni in base alla classe, si nota come quella per genitori/figli sia abbastanza uniforme in tutte le classi, attestandosi in tutti i casi tra 0,35 e 0,40 genitori/figli per passeggero (fig. 13). Nel caso di fratelli/sposi a bordo, invece, si è notato come più del 60% dei passeggeri di terza classe avesse fratelli o sposi a bordo, il 20% circa in più rispetto alle altre classi.

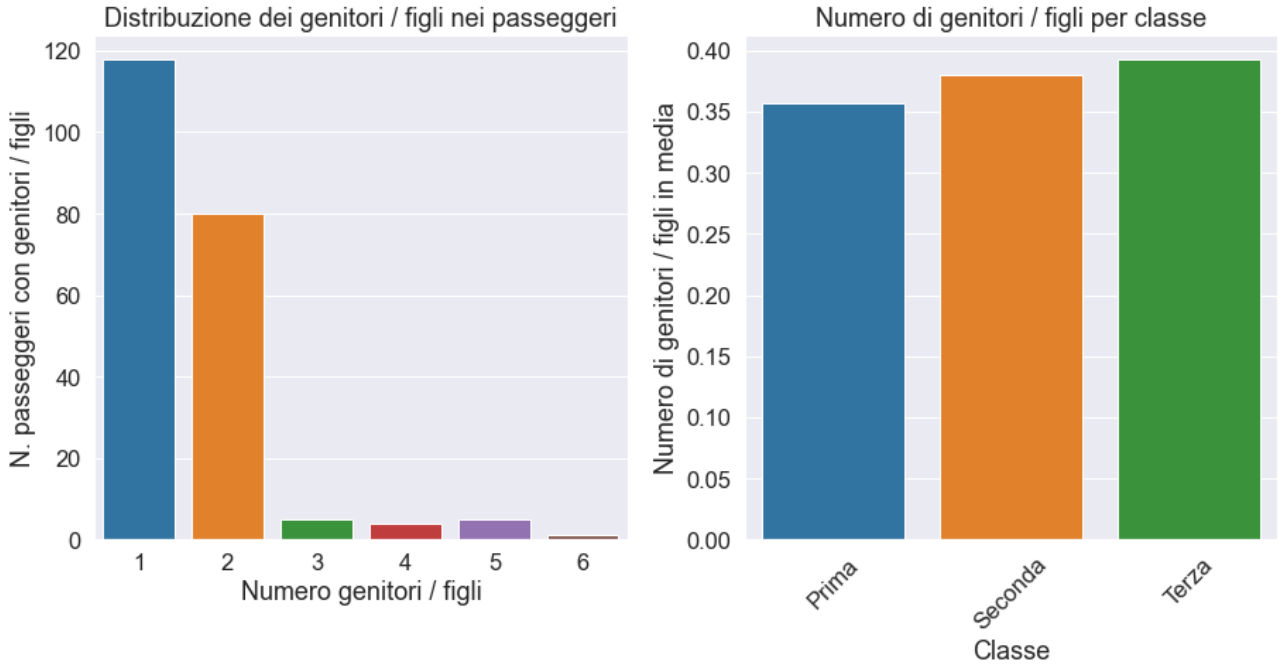

Figura 13

Sembrerebbe plausibile che i passeggeri di terza classe si spostassero con l'intera famiglia, probabilmente a scopo di emigrare, mentre quelli di classi più alte si spostassero più per divertimento, al massimo in coppia.

Descrizione del prezzo del biglietto

Analizzando i dati all'interno del dataset ci siamo resi conto che i biglietti non erano univoci, e una seconda analisi più approfondita ci ha infatti mostrato che, a parte un'unica eccezione, i prezzi dello stesso biglietto erano uguali. Abbiamo quindi ritenuto probabile che questo indicasse un prezzo complessivo per tutte le persone con lo stesso biglietto. Di conseguenza, abbiamo filtrato il dataset in modo che i prezzi dei biglietti fossero presi in considerazione una sola volta per biglietto.

A parte alcuni passeggeri che viaggiavano con un biglietto gratuito, i prezzi dei biglietti variavano da un minimo di 4 sterline a un massimo di 512, con una dispersione dei dati abbastanza alta; la parte centrale della distribuzione si concentrava infatti nella fascia più economica, con una presenza comunque massiccia di outliers a segnalare il fatto che erano presenti pochi biglietti molto costosi probabilmente dei passeggeri di prima classe. Dai seguenti boxplot (figg. 14-15) è possibile vedere infatti quanto sia sproporzionato il rapporto tra le distribuzioni delle tre classi. Andando ad analizzare la distribuzione relativa alle sole classi seconda e terza è possibile vedere quanto la terza sia caratterizzata da una presenza massiccia di outliers: ciò può essere spiegato dal fatto che la maggior parte dei passeggeri di terza viaggiavano in famiglie numerose, quindi con prezzi del biglietto piuttosto alti.

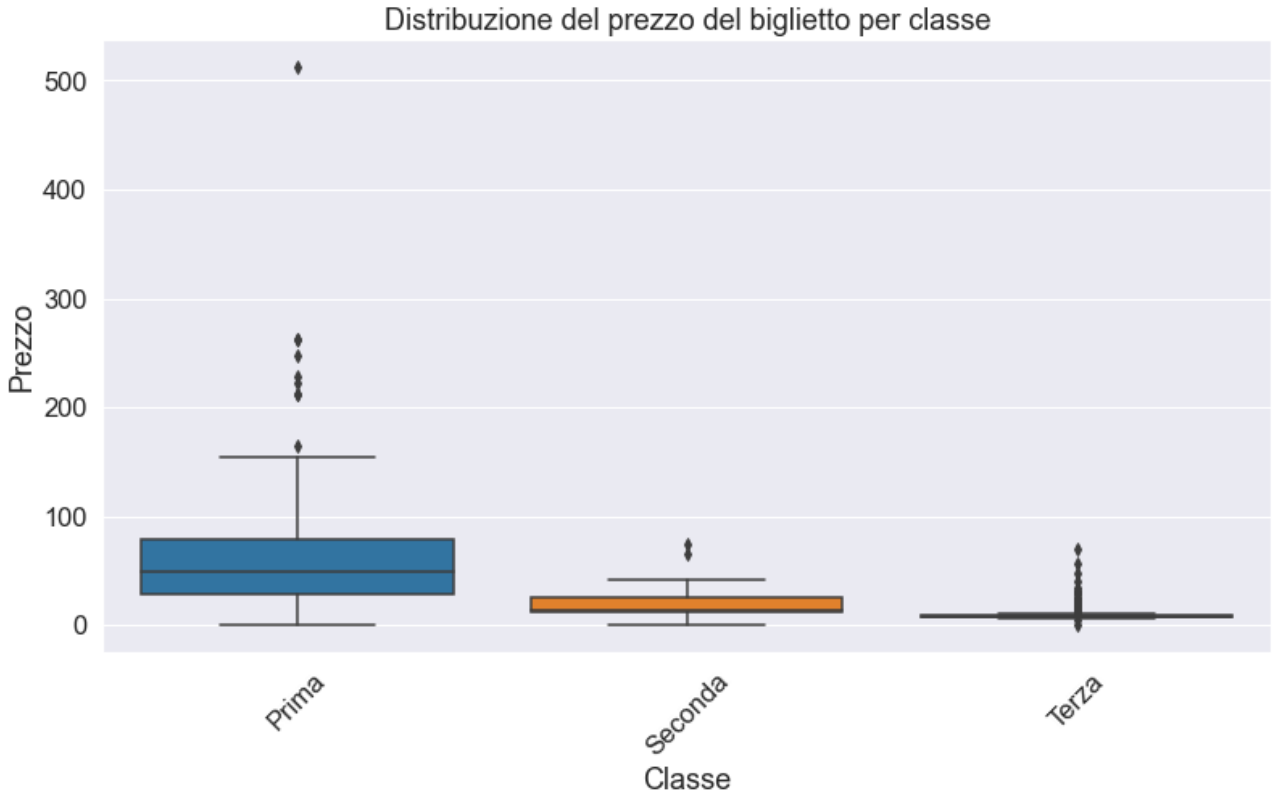

Figura 14

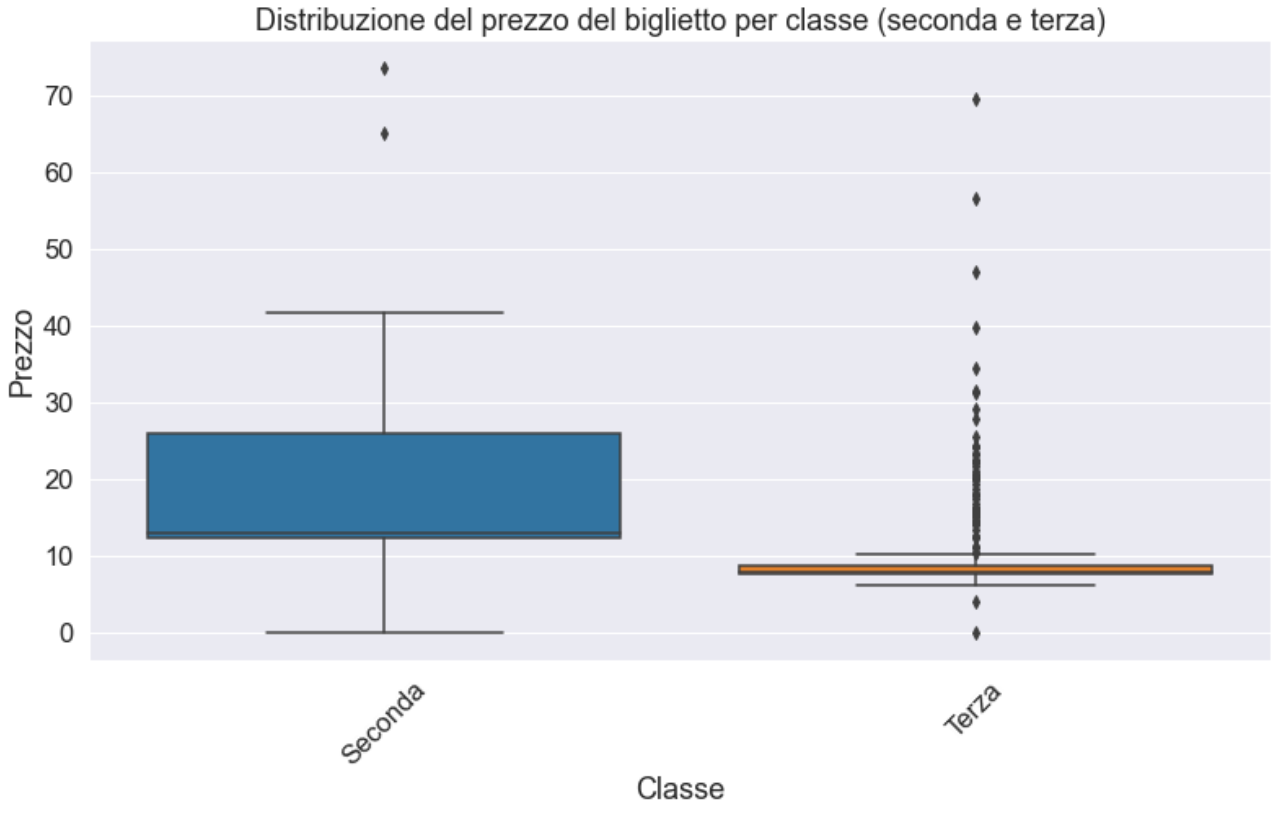

Figura 15

Naturalmente si può pensare che il prezzo aumenti con l'aumentare del numero di passeggeri con cui si viaggia, infatti andando ad analizzare i boxplot (fig. 16) è possibile vedere tale andamento crescente.

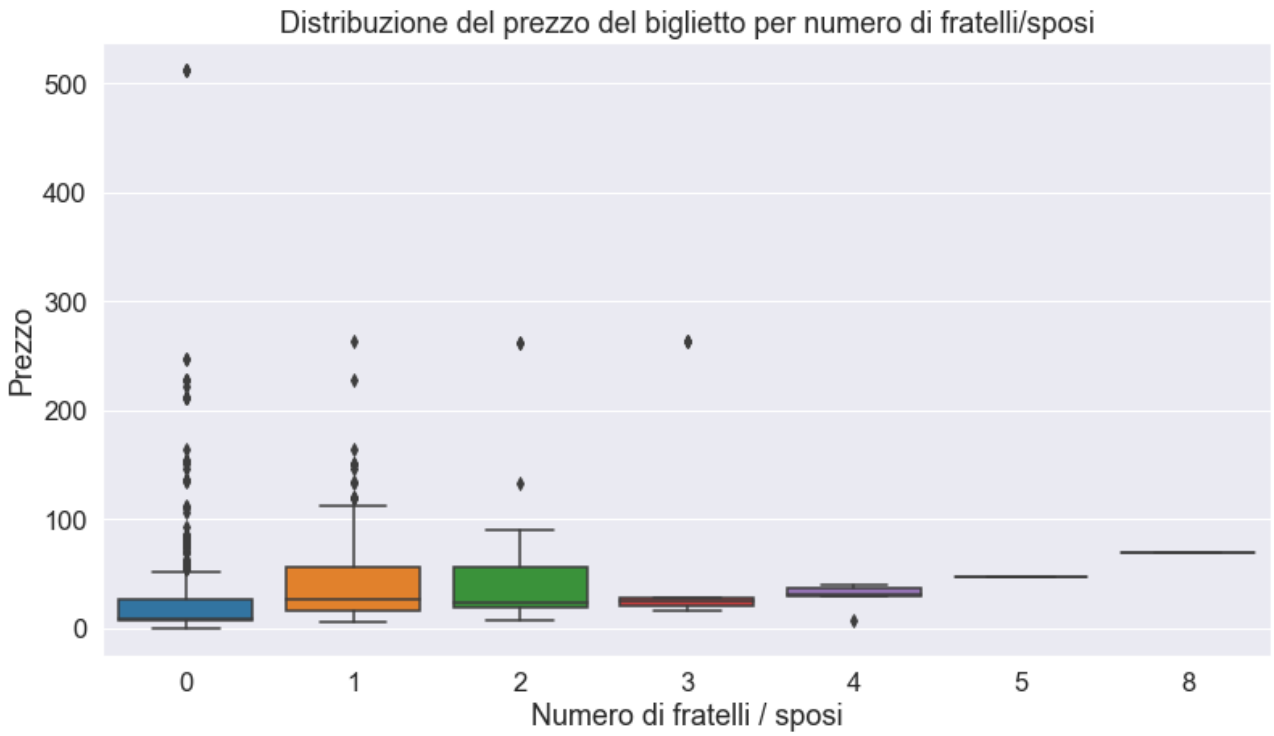

Figura 16

È possibile fare delle stesse considerazioni per quanto riguarda il boxplot relativo alla distribuzione del prezzo del biglietto per numero di genitori/figli (fig. 17), con la sola differenza che l'andamento crescente era meno pronunciato a partire da coloro che viaggiavano in compagnia di 3 altri viaggiatori.

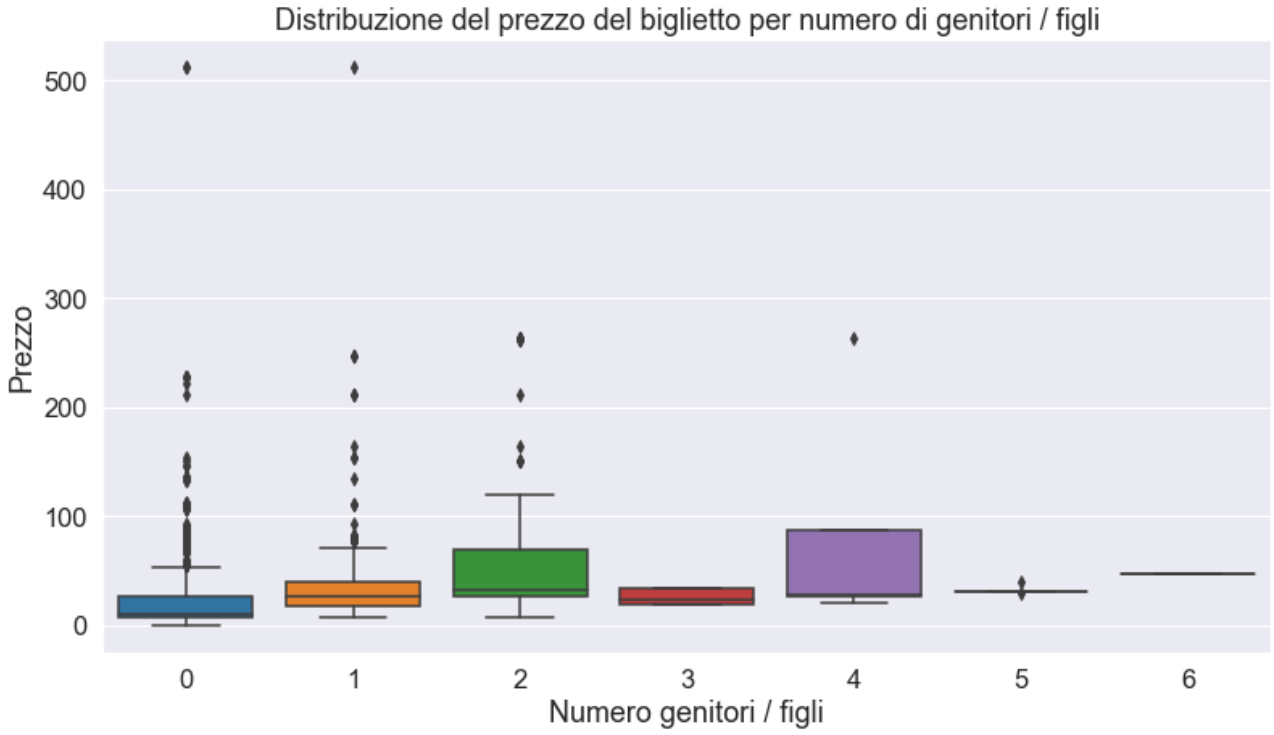

Figura 17

Correlazione

Pulizia del dataset

Abbiamo innanzitutto modificato i valori della variabile 'Sex': da valori {male, female} a {0, 1} per rendere possibile il calcolo della correlazione con le funzioni della libreria di python.

Successivamente, abbiamo controllato la presenza di valori nulli nel dataset per le variabili 'Survived', 'Pclass', 'Sex', 'SibSp', 'Parch', 'Fare' e 'Age'. Quest'ultimo era l'unico campo in cui erano presenti valori nulli (177, circa il 20% dei dati nel dataset). Di conseguenza, abbiamo creato un dataset contenente solo le tuple senza valori nulli in 'Age'.

Correlazione con outliers

Per prima cosa abbiamo realizzato il pairplot del dataset (fig. 18). Alcune correlazioni non erano rilevanti, ad esempio quelle con 'PassengerId'; ma anche gli scatterplot per variabili che potevano avere qualche correlazione, come 'Age' e 'Fare', si sono rivelati poco interessanti.

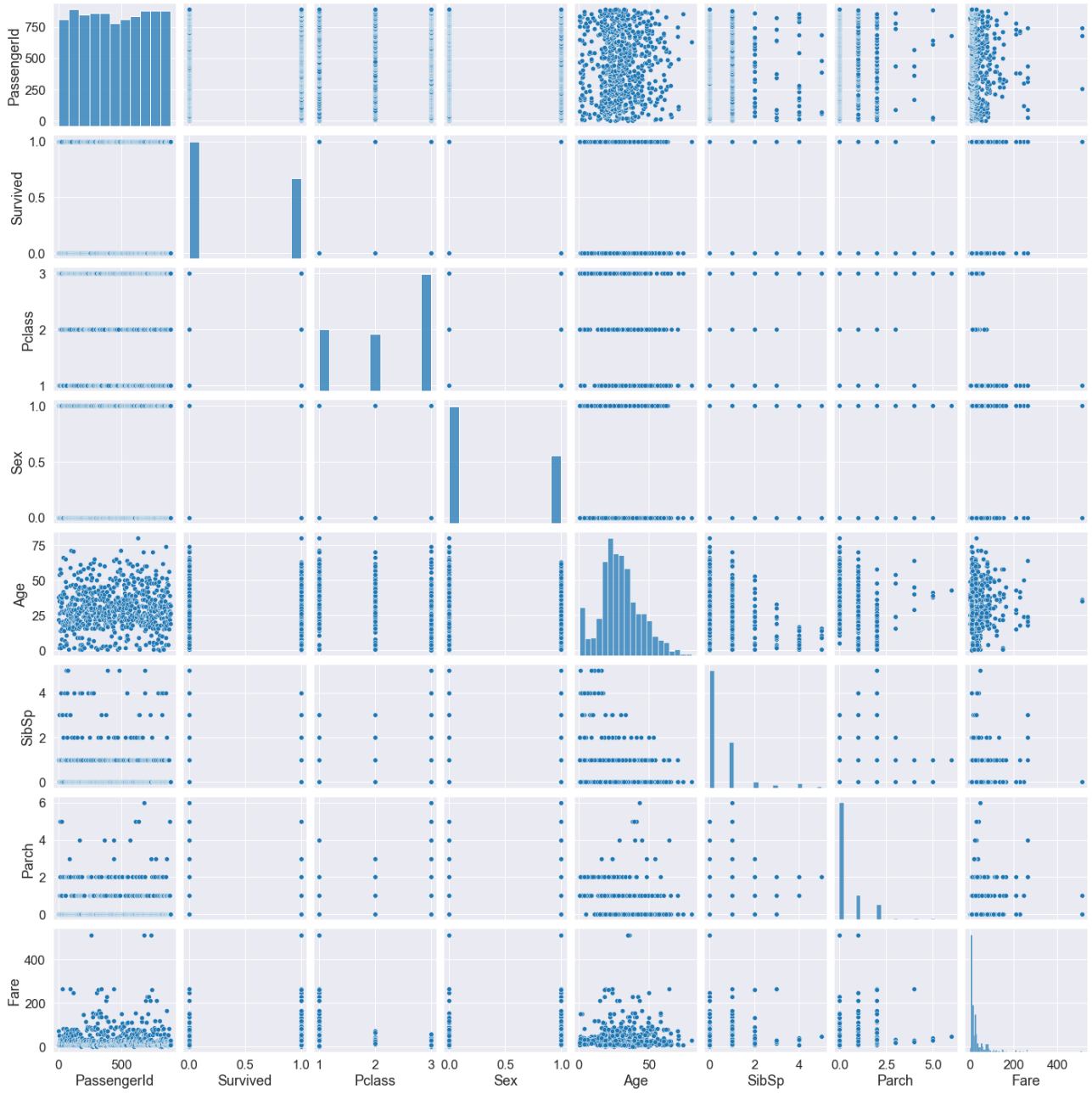

Figura 18

Abbiamo quindi realizzato una heatmap, rimuovendo prima dal dataset le colonne che non ci interessavano, ovvero 'PassengerId' e altre variabili di tipo stringa.

Per la realizzazione della heatmap (fig. 19) abbiamo utilizzato il coefficiente di Spearman. La scelta di questo coefficiente è stata motivata dal fatto che non avevamo eliminato gli outliers dal dataset, e che le distribuzioni dei dati non erano normali.

Figura 19

Dal grafico è possibile vedere che la correlazione più forte è tra 'Pclass' (la classe dei passeggeri) e 'Fare' (il prezzo del biglietto), pari a -0,73, indicando una correlazione negativa. Questo era prevedibile anche osservando semplicemente il boxplot della distribuzione del prezzo rispetto alla classe (fig. 14), evidentemente dovuto al fatto che la prima classe era la più lussuosa, e all'aumento di classe corrispondeva una condizione più modesta sulla nave.

Un'altra correlazione rilevante è quella tra il sesso dei passeggeri e l'essere sopravvissuti o meno, pari a 0,54. Ci sono infatti più sopravvissuti tra le femmine che tra i maschi, aspetto rilevante considerando che le femmine erano in numero decisamente minore rispetto ai maschi. In generale, è luogo comune dare la precedenza a donne e bambini in casi di emergenza; questo dato sembrerebbe riflettere questo, ma c'è da notare che non c'è invece correlazione tra età e l'essere sopravvissuti (-0,05).

Ci sono delle leggere correlazioni, inoltre, tra le seguenti variabili:

- il numero di fratelli/sposi e il numero di genitori/figli (0,43)
- il numero di fratelli/sposi e il prezzo del biglietto (0,42)
- il numero di genitori/figli e il prezzo del biglietto (0,41)
- la classe e l'essere sopravvissuti o meno, e classe ed età (entrambi -0,36)

La correlazione tra fratelli/sposi e genitori/figli sembrerebbe spiegabile con il fatto che chi viaggia con la famiglia viaggerebbe con la famiglia al completo. Se un passeggero ha tanti figli, inoltre, questo significa che i figli avranno tanti fratelli, e viceversa.

Per quanto riguarda la correlazione con il prezzo del biglietto, dato che quest'ultimo sembrerebbe essere complessivo, è ovvio che all'aumentare del numero di persone in un gruppo aumentasse il prezzo del biglietto.

La correlazione tra classe ed età è piuttosto ridotta, indicando all'aumentare della classe una leggera tendenza alla diminuzione dell'età, come si era osservato dalla figura 9.

La correlazione poco forte tra la classe e la sopravvivenza è stata invece sorprendente, soprattutto in confronto a quanto ci eravamo aspettati osservando le figure 4 e 5.

Analisi senza outliers

Abbiamo quindi deciso di effettuare le analisi su un dataset ripulito dagli outliers.

Abbiamo calcolato lo scarto interquartile per le variabili 'Age', 'Parch', 'SibSp' e 'Fare', e in base a questo calcolato upper e lower fence e realizzato due dataset, uno per gli outliers e uno con i dati senza outliers.

Una volta eliminate le colonne non rilevanti da quest'ultimo dataset, abbiamo realizzato nuovamente sia il pairplot (fig. 20) che l'heatmap.

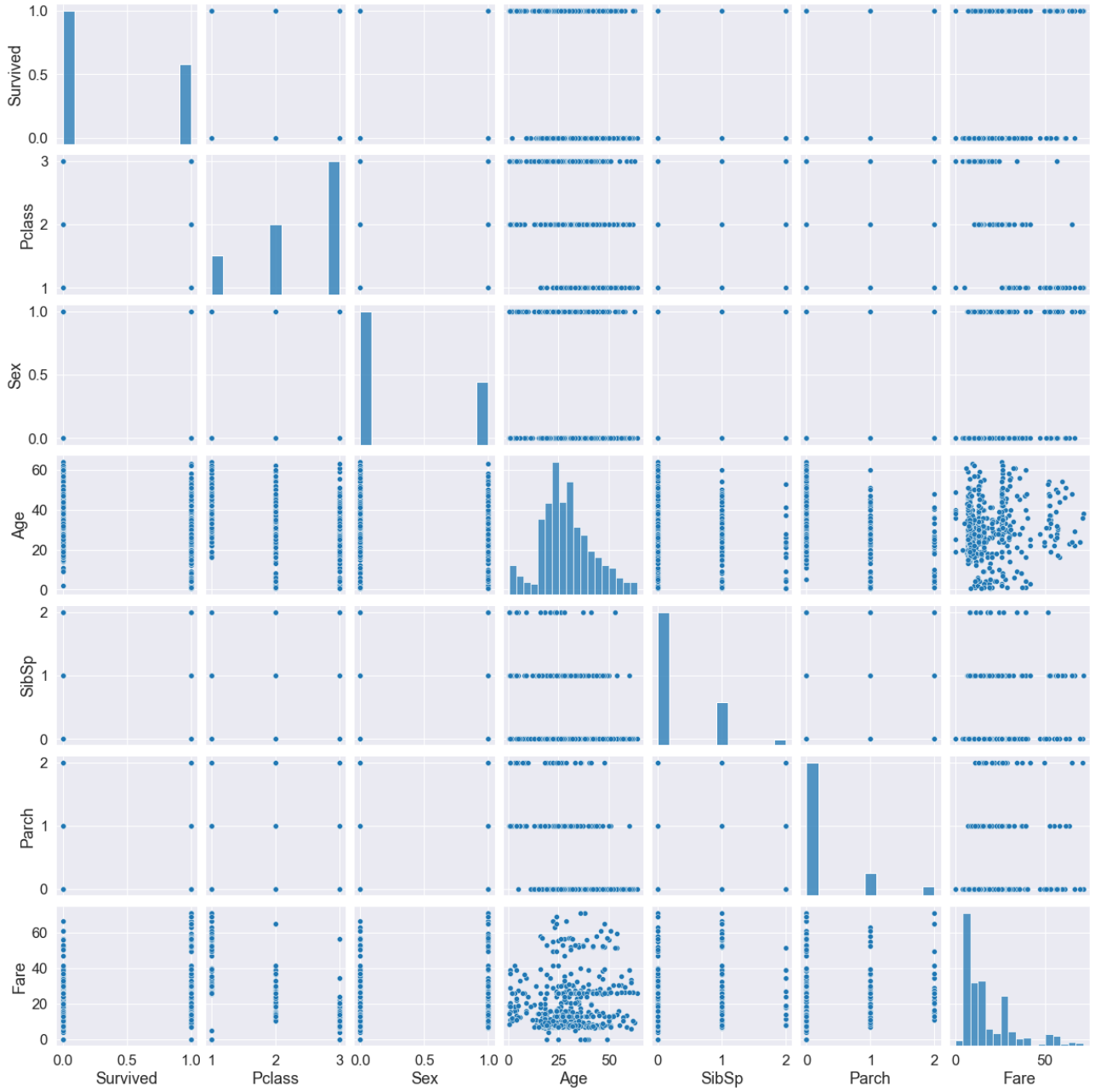

Figura 20

Questa modifica ha cambiato la distribuzione di alcuni dati, ovviamente, in particolare nello scatterplot relativo ad 'Age' e 'Fare' (fig. 21), pur rimanendo poco interessante ai fini della correlazione.

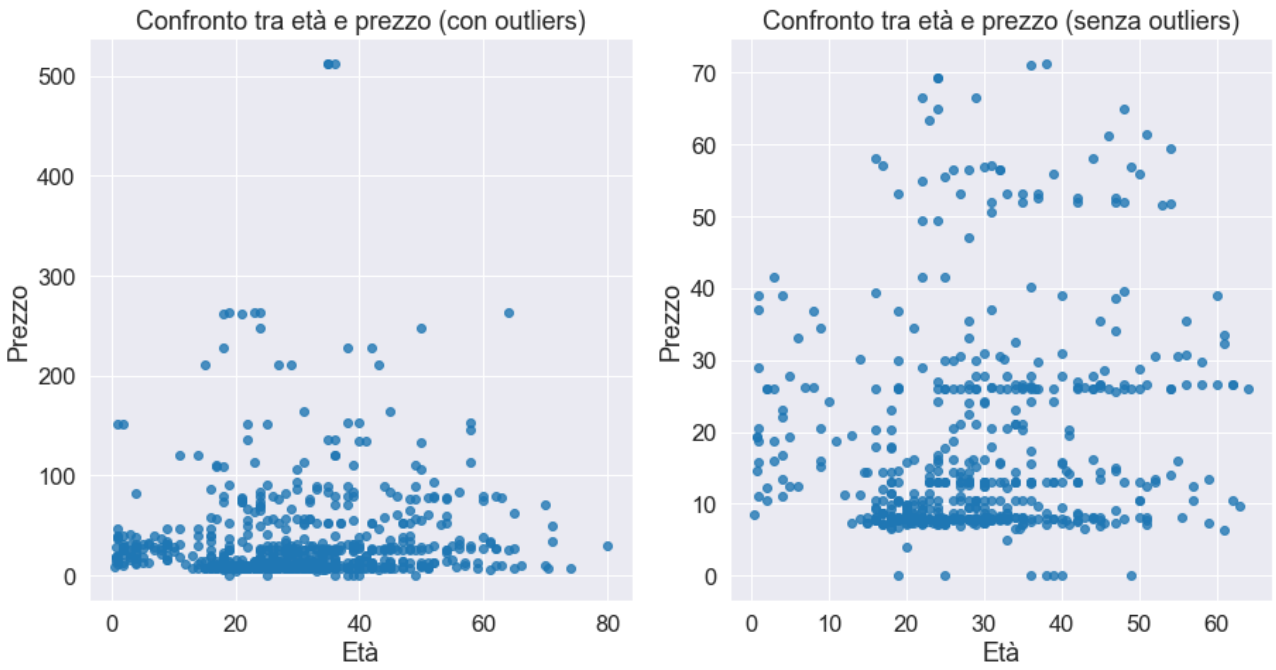

Figura 21

Abbiamo realizzato due heatmap, una utilizzando il tau di Kendall (in quanto più sensibile agli outliers, che qui non sono presenti, ma comunque adatto per distribuzioni non normali) e una con il coefficiente di Spearman (fig. 22), in modo da confrontare i dati del dataset senza outliers con quelli calcolati precedentemente.

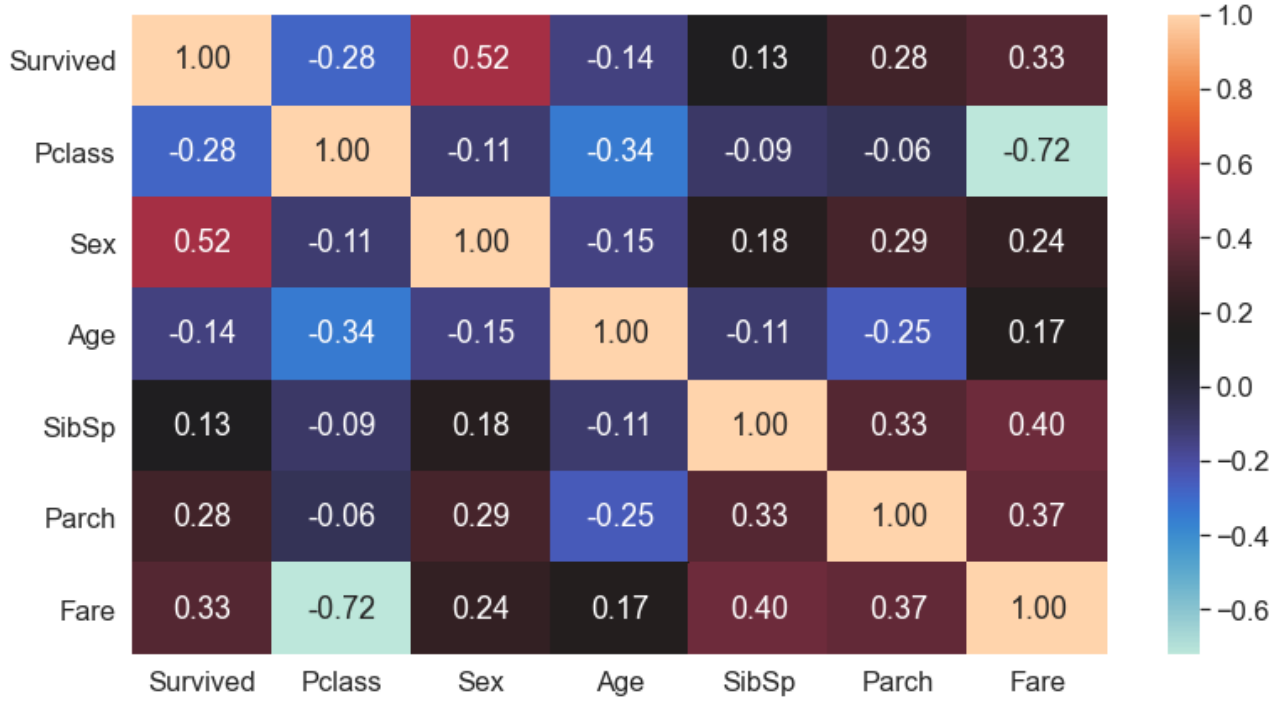

Figura 22

Le correlazioni nel dataset senza outliers erano generalmente più basse per le variabili che erano maggiormente correlate nel dataset originale, ma le tendenze erano comunque quelle rilevate nel dataset

con gli outliers. C'è stato però un leggero incremento per altre variabili, anche se comunque molto ridotto (ad esempio, 'Age' e 'Survived', passato da -0,05 a -0,12, e il numero di genitori/figli e 'Survived', passato da 0,16 a 0,28), e una riduzione per altre (ad esempio, 'Parch' e 'SibSp', da 0,43 a 0,33, e 'Pclass' e 'Survived', da -0,36 a -0,28).

In base a quanto osservato con la correlazione, abbiamo analizzato i dati delle colonne 'SibSp' e 'Parch' in forma di grafico. Il grafico ottenuto (fig. 23) era però poco informativo.

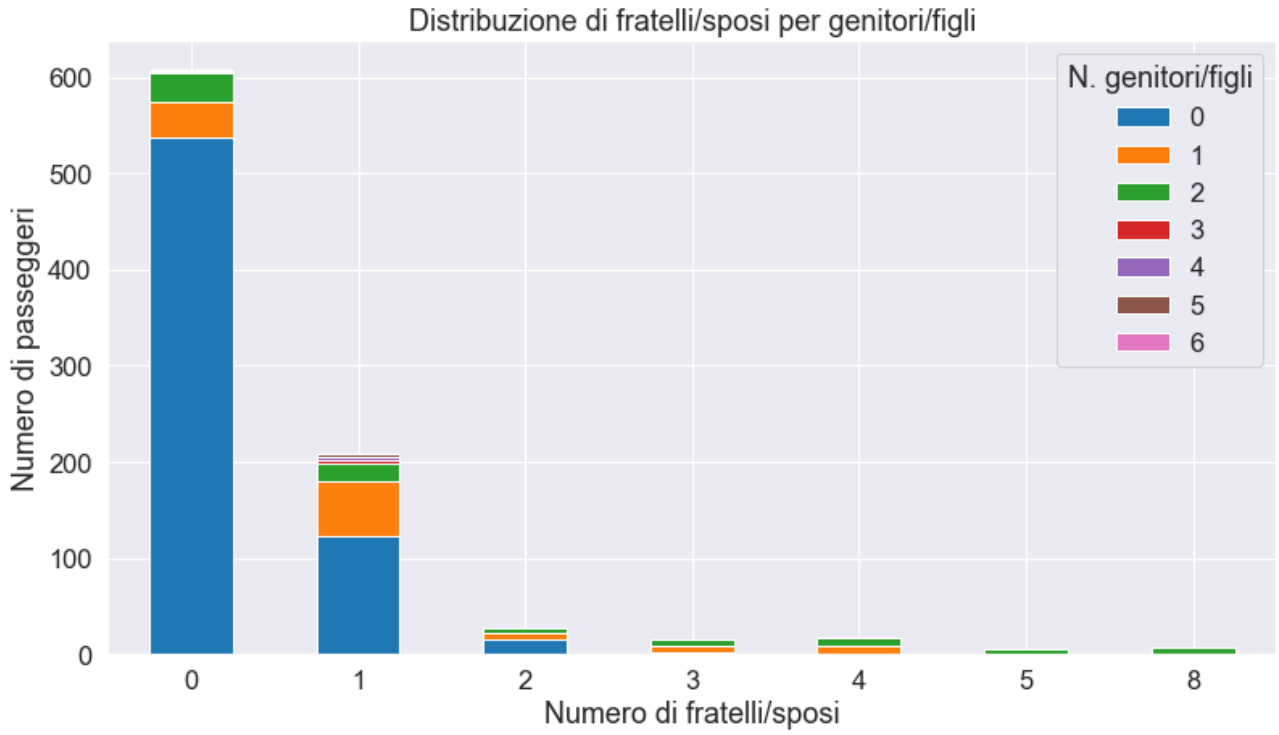

Figura 23

Conclusione

Dallo studio della correlazione abbiamo confermato quanto visto nella sezione della statistica descrittiva, a parte nel caso di 'Pclass' e 'Survived'.

In questo caso, avevamo ipotizzato che questo potesse essere derivato dal fatto che una quantità rilevante di dati era stata rimossa durante la pulizia del dataset dai valori nulli (per il parametro dell'età), ma anche includendo questi dati, l'indice di correlazione aumentava rispetto al valore del dataset senza outliers, e diminuiva rispetto a quello del dataset con gli outliers, ma sempre in quantità minima.

Abbiamo quindi concluso che probabilmente altri fattori influenzavano in maniera più determinante le possibilità di sopravvivere (ad esempio, come abbiamo visto, il sesso del passeggero).

Rimane aperto il dubbio riguardante i valori nulli di 'Cabin', che potrebbero essere dovuti anche a mancanza o perdita di dati e non al fatto che i passeggeri con un valore nullo per 'Cabin' non avessero una cabina assegnata. Per questo motivo non abbiamo approfondito ulteriormente lo studio di questo aspetto.