Relazione del progetto di analisi dati

Mattia Poggioli Gianmario Ercoli

Luglio 2021

1 Introduzione e preparazione dei dati

*I*l dataset utilizzato per la realizzazione di questo progetto è Train.csv. Si tratta di un formato basato su file di testo costituito da righe e colonne, separati da caratteri appositi (comma-separated value). Il dataset è composto da 12 variabili e 891 osservazioni di passeggeri della nave del Titanic. Per l'analisi e la visualizzazione dei dati abbiamo utilizzato Python come linguaggio di programmazione e Jupyter Notebook come ambiente di lavoro.

Il primo passo è stato quello importare le varie librerie utili per l'elaborazione dei dati stessi (Pandas e NumPy) e per la visualizzazione dei dati (Seaborn). Trattandosi di un dataset piuttosto numeroso, con diversi tipi di variabili e molti valori nulli, prima di poter compiere qualsiasi operazione sullo stesso, abbiamo analizzato gli elementi che lo compongono. Abbiamo riscontrato alcune variabili superflue o problematiche da utilizzare. Inoltre, abbiamo analizzato le variabili con valori nulli e anomali. La preparazione dei dati si è svolta nelle seguenti fasi:

- Rimozione delle variabili non ritenute interessanti per la nostra analisi (*Name*, *Ticket*, *Id*).
- Trattamento dei valori nulli (*Embarked*, *Age*, *Cabin*).
- Gestione degli outlier nelle variabili numeriche continue *Age* e *Fare*.

Per quanto riguarda il trattamento dei valori nulli, sono state adottate strategie differenti a seconda della variabile in questione. I 2 valori nulli nella variabile *Embarked* sono stati sostituiti con la moda (*S*). I 117 di *Age* sono stati sostituiti con la mediana in base al genere e alla classe. La variabile *Cabin*, dato il numero elevato di valori mancanti (687) è stata eliminata, nonostante il confronto tra le osservazioni con e senza valori mancanti mostrasse risultati interessanti [1].

[1] In particolare, abbiamo riscontrato un tasso di sopravvivenza sensibilmente maggiore tra i passeggeri la cui cabina era registrata e quelli con valore mancante (rispettivamente 67% e 30%). È probabile che tra quest'ultimi ci sia una componente di passeggeri che effettivamente non viaggiasse in cabina. Abbiamo inoltre estratto dalle stringhe della cabina la lettera identificativa; anche in questo caso abbiamo riscontrato statistiche molto diverse tra loro per quanto riguarda il tasso di sopravvivenza. Tuttavia, abbiamo ritenuto di non poter fare troppo affidamento su una variabile con così tanti valori mancanti.

Il trattamento degli outlier ha interessato le variabili continue *Age* e *Fare*. Come da noi riscontrato e come spesso succede nell'analisi dei dati, abbiamo dovuto rimuovere gli outlier, poiché causavano problemi con il calcolo delle statistiche e nella visualizzazione.

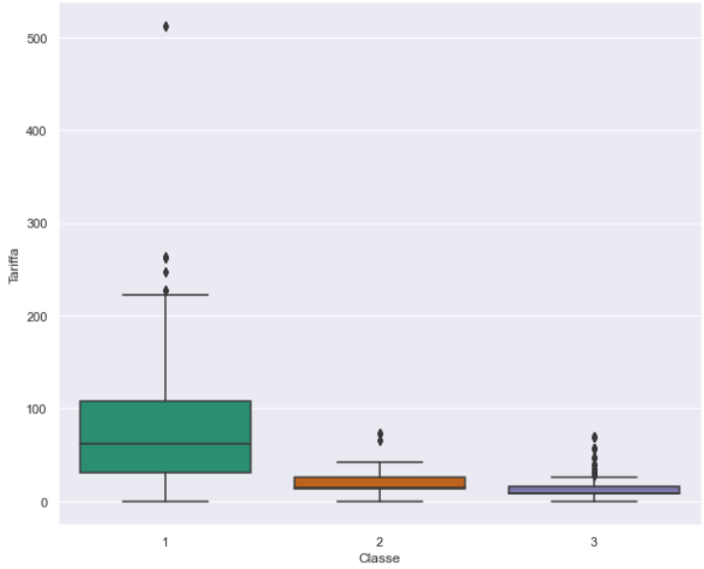

Figure 1: Boxplot delle tariffe per ogni classe

Per il rilevamento degli outlier abbiamo utilizzato il metodo del range interquartile. Nel caso di *Age*, abbiamo individuato 28 osservazioni di passeggeri relativamente anziani (a partire da 60 anni ca.). Trattandosi di un numero non particolarmente elevato, le osservazioni sono state eliminate. Nel caso di *Fare*, abbiamo calcolato il range interquartile per ogni classe, data la correlazione della variabile numerica continua con *Pclass*. Per la classe 1, 2 e 3 abbiamo individuato rispettivamente 14, 7 e 52 outlier. I valori sono stati sostituiti con la mediana per classe, potendo fare affidamento su un valore presumibilmente più veritiero (essendo questi calcolato localmente e non globalmente).

Infine, abbiamo eseguito una conversione di tipo intero sui valori della variabile *Age*, dopo aver arrotondato per eccesso i minori di 1 anno.

Il trattamento dei dati ci ha garantito maggiore praticità e un utilizzo più corretto dei dati, nel calcolo delle statistiche e nella visualizzazione dei dati. Dopo questo passaggio abbiamo proceduto a un'analisi descrittiva più approfondita delle variabili.

2 Descrizione delle variabili

La prima valutazione che siamo andati a fare è stata quella dei sopravvissuti. Come vediamo dal grafico qui riportato, il numero dei decessi supera nettamente quello dei superstiti, attestandosi a 528, il 61% dei decessi e a 335, cioè il 39% per i sopravvissuti.

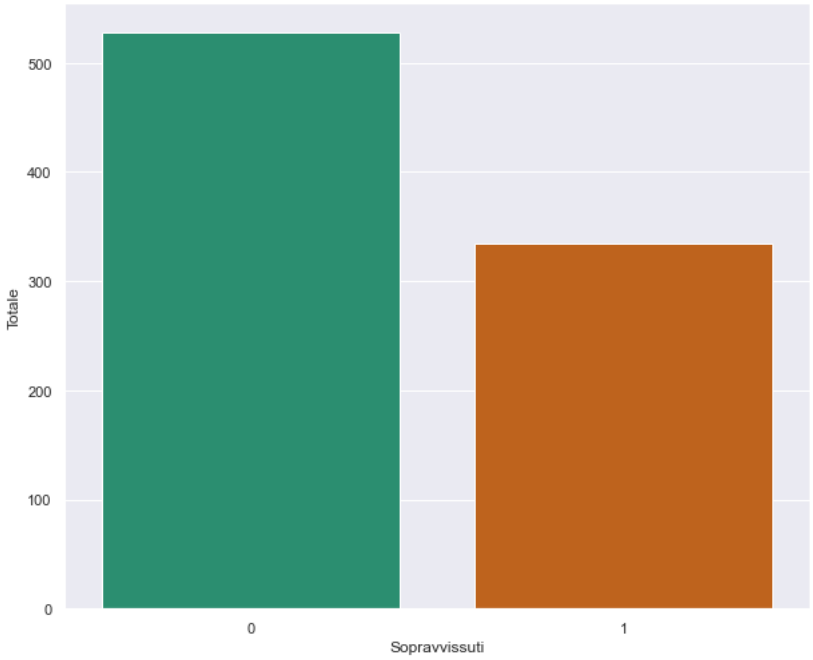

Figure 2: Distribuzione sopravvissuti

La descrizione dei dati ci permette di capire nel dettaglio alcune informazioni riguardo ai passeggeri, in particolar modo l'età, la classe e la tariffa che i passeggeri del Titanic hanno pagato per salire a bordo. Tutte queste caratteristiche sono state osservate singolarmente e in rapporto a ad altre variabili. In particolare, abbiamo considerato la variabile *Survived* come nostro oggetto d'indagine.

Il barplot qui sotto mostra come la maggior parte delle persone fosse collocata in terza classe, con un valore di 485 individui, mentre le altre due classi avevano un numero di individui nettamente inferiore, rispettivamente 199 per la prima e 179 per la seconda. Ci siamo chiesti quale fosse il numero di deceduti e di sopravvissuti rispetto alla classe in cui questi viaggiavano. Il barplot successivo ci permette di visualizzare la distribuzione in rapporto alla variabile *Survived*. Il numero maggiore di decessi si è verificato in terza. Al calcolo delle percentuali risulta che ben il 75% dei viaggiatori nella classe più bassa è deceduto, scendendo al 51% per la seconda e al 34% nella prima.

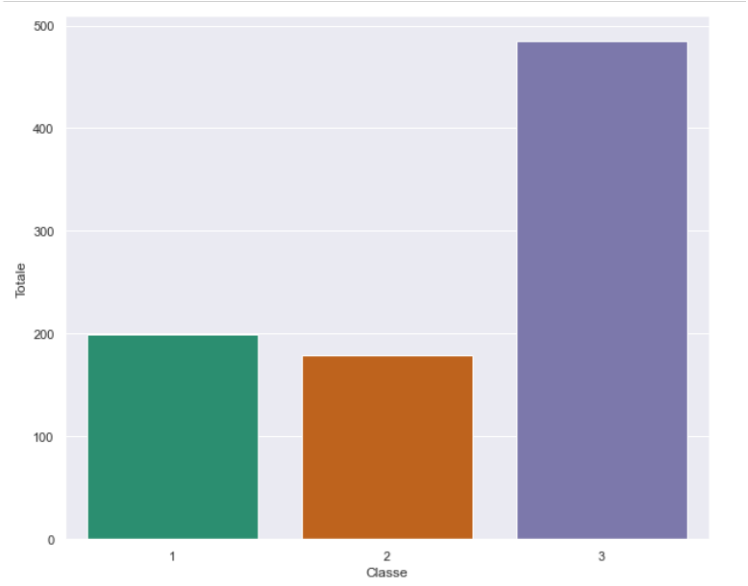

Figure 3: Distribuzione dei passeggeri per classe

Attraverso i due piechart possiamo vedere anche le percentuali dei sopravvissuti e dei deceduti totali divisi per classe.

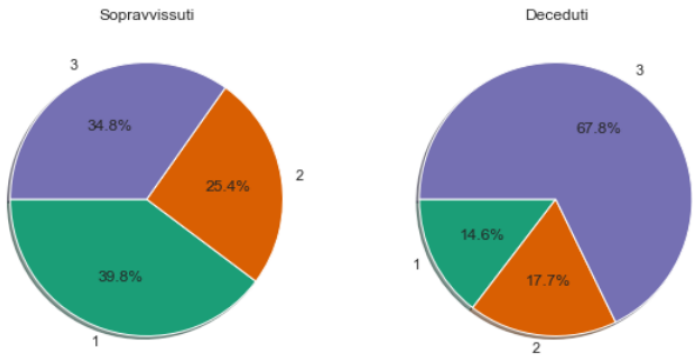

Figure 4: Percentuali di sopravvissuti e deceduti per classe

In seguito, abbiamo analizzato la distribuzione per genere per classe, che ha mostrato come risultato un'evidente preponderanza degli uomini rispetto alle donne in tutte le classi, differenza che tende ad assottigliarsi nelle prime, mentre vi è una netta separazione nella terza. Complessivamente, il nostro dataset è composto da 310 donne (36%) e 553 uomini (64%). Nonostante la divisone in classe, le donne sono comunque decedute in numero e in percentuale

nettamente inferiore rispetto agli uomini. Il tasso di sopravvivenza di uomini e donne è rispettivamente 19% e 74%. Tra i sopravvissuti troviamo il 68% di donne e il 32% di uomini.

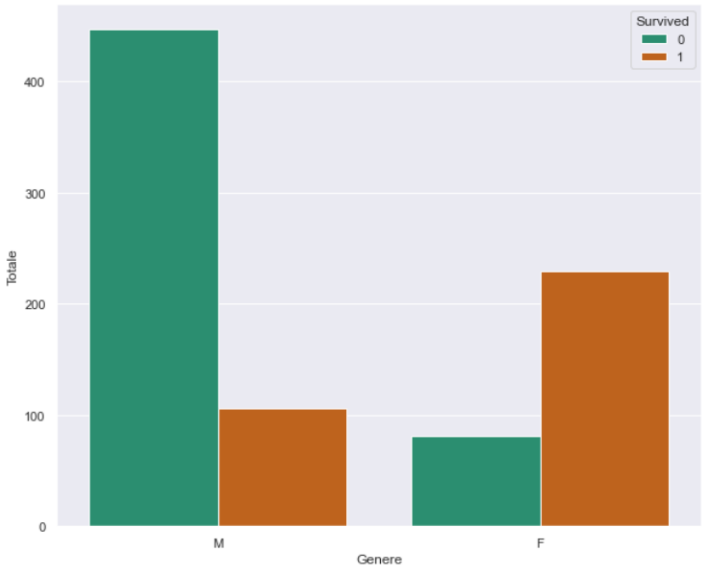

Figure 5: Distribuzione sopravvissuti e deceduti per genere

Andando a visualizzare le statistiche di età per genere e prezzo per genere abbiamo trovato risultati interessanti e con esiti inaspettati. In particolar modo la variabile *Age* ha una media più alta per quanto riguarda il genere maschile, attestandosi intorno ai 28,6 mentre per le donne è 26,7. Le tariffe invece si distribuiscono in modo inverso rispetto all'età. I viaggiatori di genere femminile hanno pagato una tariffa media molto più alta rispetto agli uomini, quasi il doppio, raggiungendo il valore di 35,69 sterline inglesi mentre quello degli uomini raggiunge 19,72. La distribuzione effettiva dell'età è dimostrata nel seguente istogramma unimodale, dove il picco è raggiunto nella fascia tra 20 e 30. La maggior parte delle persone a bordo della nave era quindi costituito da ragazzi e ragazze giovani.

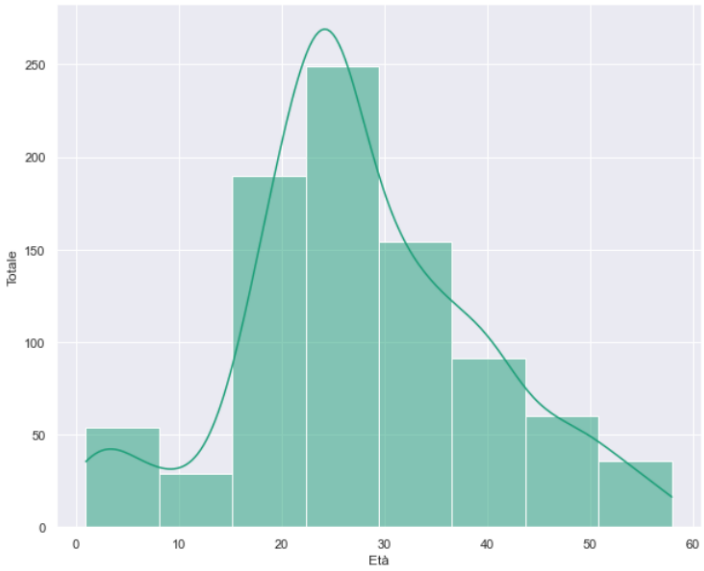

Figure 6: Distribuzione dell'età dei passeggeri

Per quanto riguarda la distribuzione dei passegeri, nel density plot qui sotto possiamo vedere come la fascia d'età tra i 20 e i 40 anni è quella più densamente popolata, e che la maggior parte dei passeggeri di questa fascia erano alloggiati nelle cabine della terza. Questo però è ancora una realizzazione parziale, che abbiamo colmato con l'incrocio di età, densità e genere per capire meglio quanti uomini e quante donne, di quanti anni e con quale concentrazione fossero presenti. Ci siamo quindi chiesti se l'età e il genere abbiano contribuito in qualche modo alla sopravvivenza di un determinato gruppo di passeggeri. La risposta a questo quesito sembra proprio essere positiva. Andando a considerare il grafico qui riportato:

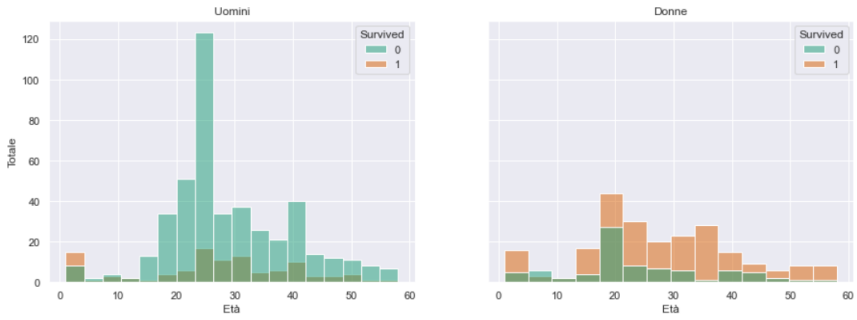

Figure 7: Distribuzione dell'età tra deceduti e deceduti in base al genere

Come già sottolineato, è possibile notare chiaramente come il numero di donne sopravvissute alla catastrofe sia stato nettamente superiore a quello degli uomini. Si potrebbe ipotizzare quindi che nel salvataggio dei passeggeri sulle scialuppe sia stata data la precedenza ai viaggiatori di sesso femminile rispetto a quelli di sesso maschile, tanto che alcune fasce di età il numero di decessi si avvicina allo 0, cosa che non si verifica nel grafico degli uomini.

Un altro dato che ha suscitato particolare interesse durante lo studio del dataset è stato il prezzo del biglietto che i passeggeri hanno pagato per questo viaggio. In generale abbiamo appurato che il maggior numero di passeggeri abbia pagato una cifra compresa tra 0 e 50 sterline, come è possibile vedere nell'istogramma qui riportato.

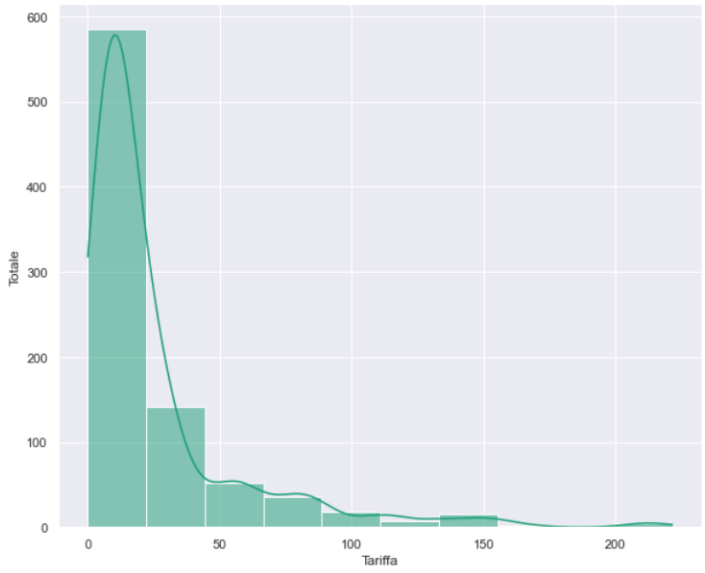

Figure 8: Distribuzione tariffe

A questo punto abbiamo deciso di studiare la distribuzione del costo del biglietto in base alla classe di viaggio, scoprendo che in terza classe la maggior parte dei viaggiatori ha pagato una cifra compresa tra le 5 e le 10 sterline, in seconda tra le 10 e le 30 e infine in prima tra le 25 e le 100 sterline.

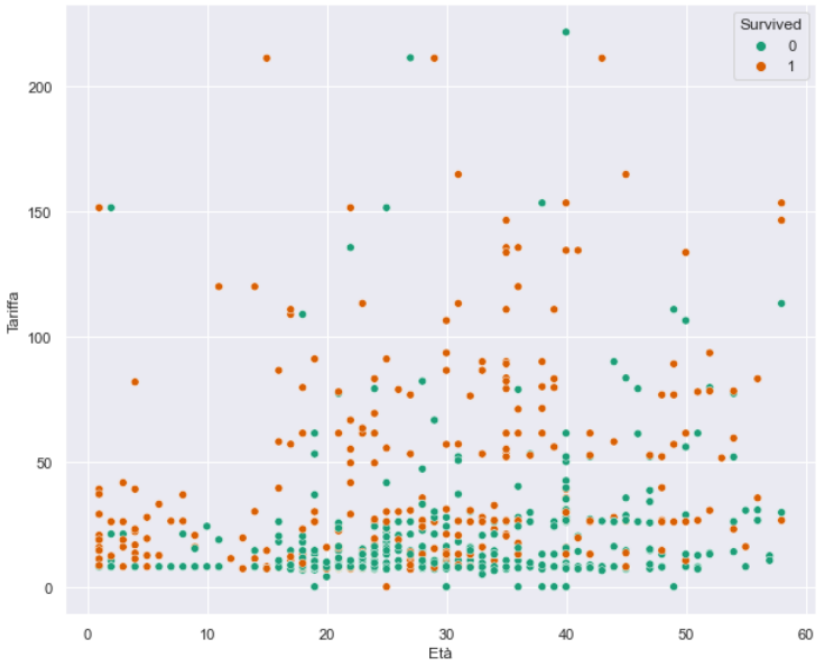

Figure 9: Età e tariffa dei passeggeri di sopravvissuti e deceduti

Nel grafico è possibile notare anche come la stragrande maggioranza dei viaggiatori che hanno pagato una cifra superiore alle 50 sterline sia sopravvissuta e questo dimostra anche che nel momento di salvare i passeggeri, coloro che alloggiavano nelle cabine di prima siano stati tra i primi ad essere salvati.

Il confronto tra porto di imbarco, tariffa e età non ha mostrato tendenze particolari, con una distribuzione molto variegata, anche se dal grafico è possibile osservare come la maggioranza dei passeggeri sia stata imbarcata nel porto di Southampton. Più di 600 persone salirono a bordo del Titanic dal porto inglese, che corrispondono al 72%, mentre nel porto francese di Cherbourg si imbarcarono 163 passeggeri, il 19% e infine a Queenstown soltanto 75 persone, cioè il 9%.

Calcolando la percentuale dei passeggeri per classe di viaggio e in base al porto di imbarco abbiamo ottenuto le seguenti informazioni: nei porti di Southampton e Queenstown la percentuale maggiore di passeggeri si è imbarcata per la terza, mentre nel porto di Cherbourg la percentuale maggiore riguarda i passeggeri di prima. Questa tesi è sostenuta anche dal calcolo statistico delle tariffe per porto di imbarco, tanto che nel porto francese la media del costo del biglietto è di 40,83 sterline, il doppio rispetto al porto inglese e più del triplo rispetto a quello irlandese.

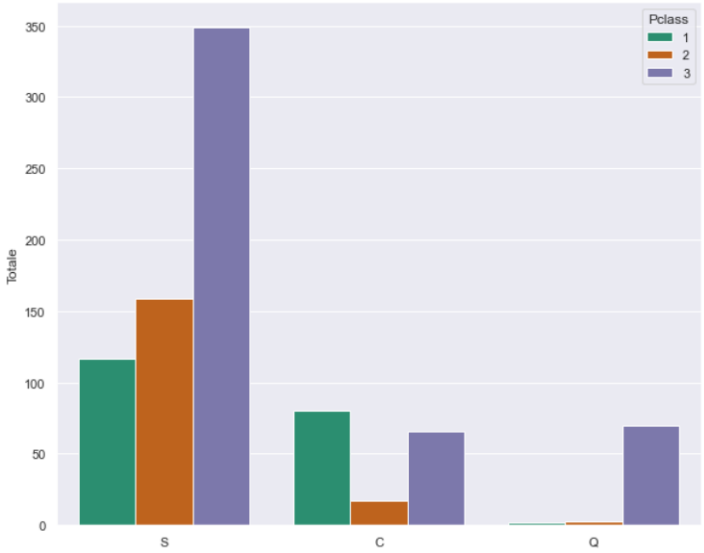

Figure 10: Distribuzione delle classi in base al porto di imbarco

Tra i dati disponibili nel dataset è presente anche il numero di fratelli/coniugi e genitori/figli a bordo del Titanic. Per quanto riguarda i primi, indicati con la dicitura *SibSp*, le percentuali mostrano come la maggior parte dei passeggeri viaggiasse senza famiglia, con una percentuale del 68%, il 24% era accompagnato da 1 familiare, il 3% da 2, e 2% da 4. Grazie a questo calcolo possiamo capire che pochissime famiglie erano a bordo del Titanic. Un discorso molto simile vale anche per i genitori/figli (*Parch*), con percentuali leggermente superiori per i viaggiatori solitari che ci permettono di capire come effettivamente la maggior parte delle persone non viaggiasse in famiglia.

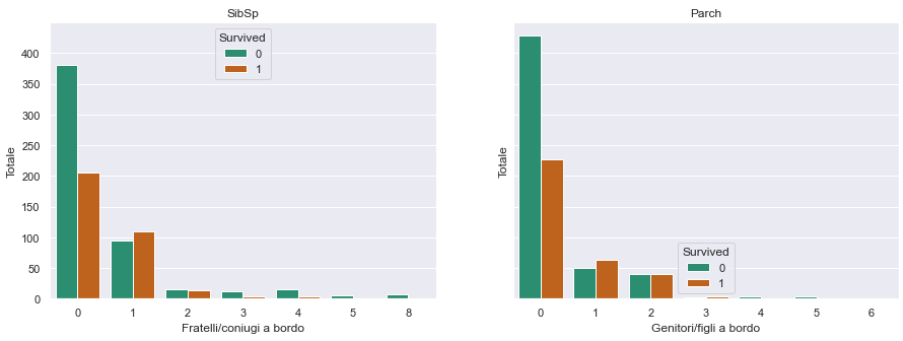

Figure 11: Distribuzione deceduti e sopravvissuti per le variabili *SibSp* e *Parch*

In base a questi dati abbiamo cercato di trovare una relazione tra la grandezza

della famiglia e la sopravvivenza della stessa. Abbiamo quindi creato una nuova variabile (*FamilySize*) che combinasse il numero dei fratelli/coniugi e quello di genitori/figli facendo la somma delle due variabile e aggiungendo 1 (considerando quindi chi viaggiava da solo come parte di una famiglia di grandezza 1). Il grafico sottostante mostra il tasso di sopravvivenza in base alla grandezza della famiglia. Dal grafico sembrerebbe che famiglie poco numerose (2-4 elementi) abbiano avuto maggiore possibilità di salvarsi rispetto a chi era da solo. Questa differenza potrebbe anche essere spiegata dalla distribuzione di uomini e donne in base alla grandezza della famiglia: quelli che viaggiavano da soli erano in larga maggioranza uomini (76%); la percentuale diminuisce sensibilmente nelle famiglie composte da 2 persone (46 %) e 3 persone (51%), scendendo al 34% nelle famiglie composte da 4 persone (dove per l'appunto si registra il tasso di sopravvivenza più alto).

Il tasso di sopravvivenza diminuisce drasticamente nelle famiglie più numerose, indicando che tra questi nessuno sia riuscito a salvarsi. Tuttavia, è bene sottolineare lo sbilanciamento tra le varie modalità della variabile, specialmente tra le famiglie molto numerose.

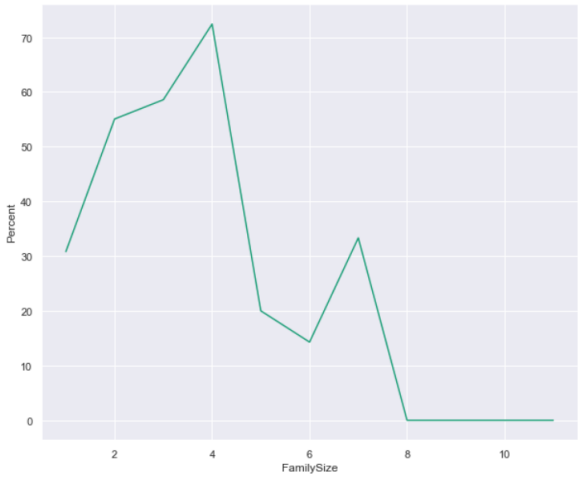

Figure 12: Variazione del tasso di sopravvivenza in base alla grandezza della famiglia

3 Correlazioni

Per poter calcolare la correlazione tra le variabili del dataset abbiamo avuto la necessità di trasformare *Sex* in variabile numerica dicotomica. Dopodiché, abbiamo calcolato le correlazioni. La heatmap permette di capire visivamente quali variabili sono correlate tra di loro.

Figure 13: Correlazioni tra le variabili

Come è visibile nel grafico, sono presenti diverse correlaizoni basse o trascurabili e alcune alte. Le correlazioni più forti sono quelle tra dimensione delle famiglie e *SibSp* e *Parch*, che raggiungono valori molto vicini a 1, mentre ha correlazione negativa l'incrocio tra *Fare* e *Pclass*. Si tratta di risutati perlopiù attesi, in particolare per il primo caso, essendo la variabile *FamilySize* la combinazione delle due variabili, anch'esse tra loro correlate.

Nel grafico tariffa-età vediamo che la linea rossa tende verso l'alto nella parte sinistra dell'immagine. Questo ci permette di capire che la tendenza delle tariffe aumenta all'aumentare dell'età del passeggero.

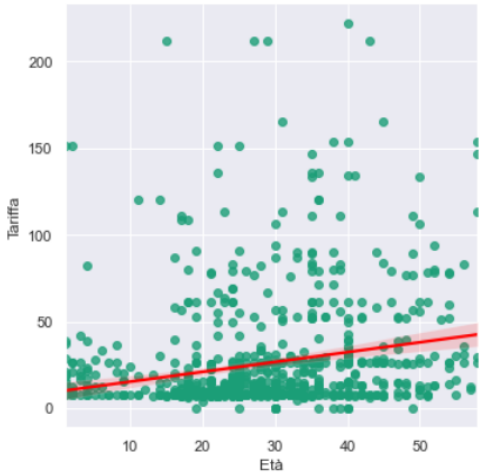

Figure 14: Correlazione tariffa-età

Inoltre, nel dividere i grafici per genere dei passeggeri è possibile notare come questo aumento sia particolarmente più evidente nel sesso femminile tanto che il calcolo del coefficiente di Pearson restituisce un risultato di 0.26 per le donne, mentre quello degli uomini raggiunge lo 0.20, quindi comunque positiva.

Figure 15: Correlazione tariffa-età in base al genere

La correlazione tra la tariffa e l'età rispetto alla classe restituisce invece una correlazione negativa, con valori per quanto riguarda la prima, seconda e terza rispettivamente di -0.20, -0.23 e -0.13.

La visualizzazione grafica dei del rapporto sopravvissuti-genere tramite heatmap

ci permette di vedere quanto sia stato importante l'impatto della catastrofe sul genere maschile, uccidendo 447 uomini su 553, mentre quella che riguarda il rapporto sopravvissuti-classe mostra chiaramente come la maggior parte dei decessi sia avvenuta nella terza.

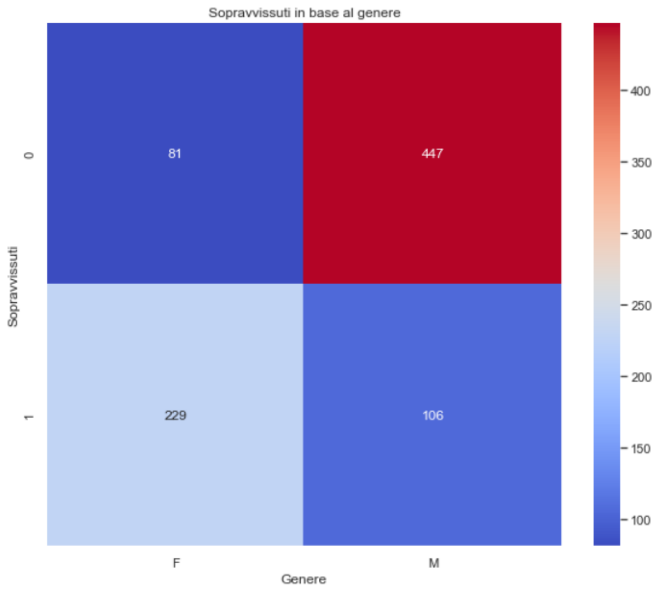

Figure 16: Heatmap sopravvissuti e genere

Per ulteriore chiarezza abbiamo realizzato due grafici boxplot che mostrano il rapporto fra classe e tariffa e quello tra età e classe, calcolando per entrambi il coefficiente di Pearson.

4 Conclusioni

L'incidente del Titanic rappresenta uno dei più gravi disastri dell'età moderna, in cui persero la loro vita centinaia di persone. Attraverso l'analisi del dataset dei passeggeri a bordo, abbiamo potuto comprendere alcune informazioni importanti. In primo luogo, la distribuzione del genere tra i passeggeri, con una netta superiorità maschile rispetto alla femminile. Questa prevalenza numerica ha ovviamente avuto effetti anche sul numero di decessi per genere, anche qui nettamente maggiore negli uomini. Inoltre, l'analisi ci ha permesso di capire che la maggior parte dei passeggeri a bordo era costituita da under 30, e che questa maggioranza era alloggiata nella terza. Per quanto riguarda i nuclei familiari, abbiamo scoperto che circa 500 persone viaggiavano in solitaria, ma erano presenti anche numerose famiglie con un numero di membri variabile da 2 a 10 membri. Grazie a questo dato abbiamo poi appurato che all'aumentare del numero di membri, aumentava la percentuale di deceduti, tanto che le famiglie

composte da più di 4 persone sono decedute con un valore superiore al 60%. Le tariffe sono state una variabile molto interessante dal punto di vista di correlazione, restituendo un valore positivo se messa in relazione con età e genere, mentre se relazionate con età e cabina queste avevano un coefficiente negativo, di poco ma comunque negativo.