UNIVERSITÀ DI PISA

Report in

Programmazione e analisi di dati

Modulo di analisi di dati

Analisi statistiche sul dataset Titanic

Ilenia Bari

Matr. 590445

Anno accademico 2020/21

Indice

Indice delle figure

Comprensione del dataset

Per una corretta analisi statistica è opportuno comprendere l'ambito e la natura dei dati da analizzare. Il dataset fa riferimento ai passeggeri del Titanic, il transoceanico britannico che affondò durante il suo viaggio inaugurale, a causa della collisione con un iceberg avvenuta nella notte. La prima tratta prevedeva il viaggio da **Southampton** a New York, passando per **Cherbourg** in Francia e **Queenstown**, l'attuale Cobh in Irlanda.

Il dataset è stato scaricato dal link (Titanic - Machine Learning from Disaster | Kaggle), considerando unicamente il file contenente il training set. Le variabili contenute in questo sono:

- Survived, la variabile target, i cui valori 0 e 1 indicano se il passeggero è sopravvissuto (1) o meno (0);
- Pclass indica la classe per la quale avevano pagato, può essere 1, 2 o 3;
- Name, i nomi dei passeggeri;
- Sex il genere di questi;
- Age l'età;
- SibSp indica il numero per ogni persona di parenti stretti come fratelli o sorelle e coniugi;
- Parch riguarda invece il numero di genitori o figli a bordo;
- Ticket rappresenta il codice del biglietto;
- Fare il prezzo pagato per il biglietto;
- Cabin il numero di cabina;
- Embarked indica il porto di imbarcazione: Southampton (**S**), Cherbourg (**C**), Queenstown (**Q**).

Le variabili *Fare* e *Age* sono variabili quantitative continue, *SIbSp* e *Parch* sono discrete. *Name*, *Sex*, *Ticket*, *Cabin* sono variabili qualitative nominale, *Sex* è qualitativa dicotomica, *Survived* e *Embarked* nominale codificate.

Pulizia del dataset

La prima parte di questa analisi si prefigge l'obiettivo di verificare che non ci siano valori mancanti e outlier e sulla base di questi effettuare delle operazioni risolutive.

Il dataset presenta 2 valori mancanti per la variabile Embarked, 687 per Cabin e 177 per la variabile Age. I primi sono stati sostituiti con la moda e quindi il valore "S", e i terzi con la media (30 anni), mentre la colonna Cabin è stata completamente eliminata dal dataset in quanto l'alto numero di valori nulli rendeva impossibile qualsiasi analisi statistica. In assenza di significatività statistica, è stata eliminata anche la variabile PassengerID, mentre la variabile Ticket è stata eliminata successivamente, in quanto utilizzata come supporto ad alcune

ipotesi. Con questa prima operazione il dataset si è ridotto a 10 variabili, con 891 valori ciascuno.

Successivamente sono stati costruiti dei boxplot per le variabili continue Fare e Age.

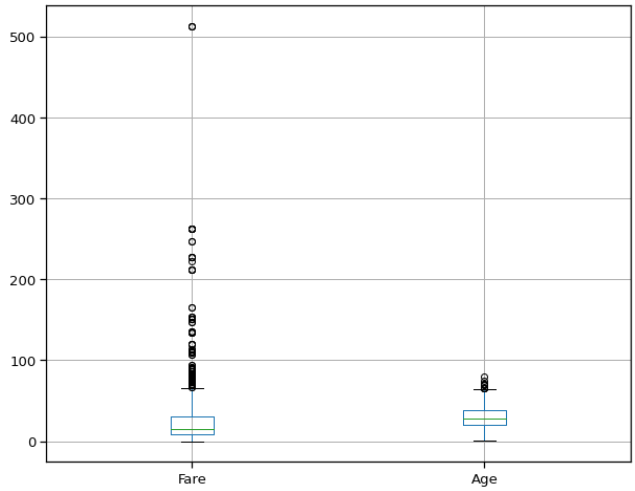


Figura 1 BoxPlot distribuzioni continue

Il boxplot mostra degli outlier per entrambe le variabili. L'eccessivo numero di valori anomali per Fare porta a chiedersi se siano o meno parte importante della distribuzione, e se eliminandoli si perderebbe una importante unità statistica. Anche per quanto riguarda la variabile Age, la continuità dei valori e la poca numerosità di questi porta a una riflessione aggiuntiva sulla decisione di eliminarli o meno

Occorre anche considerare che per entrambi i boxplot, il baffo inferiore è in corrispondenza del numero 0, ciò significa che nella distribuzione sono presenti dei valori mancanti o inesatti.

Per Age sono presenti 7 valori minori di 1, che vanno da 0,42 a 0,97; in questo caso un arrotondamento consente di avere numeri interi ed eliminare i valori al di sotto di 1. Per Fare invece sono 15 i valori inesatti, vengono perciò rimossi dal dataset.

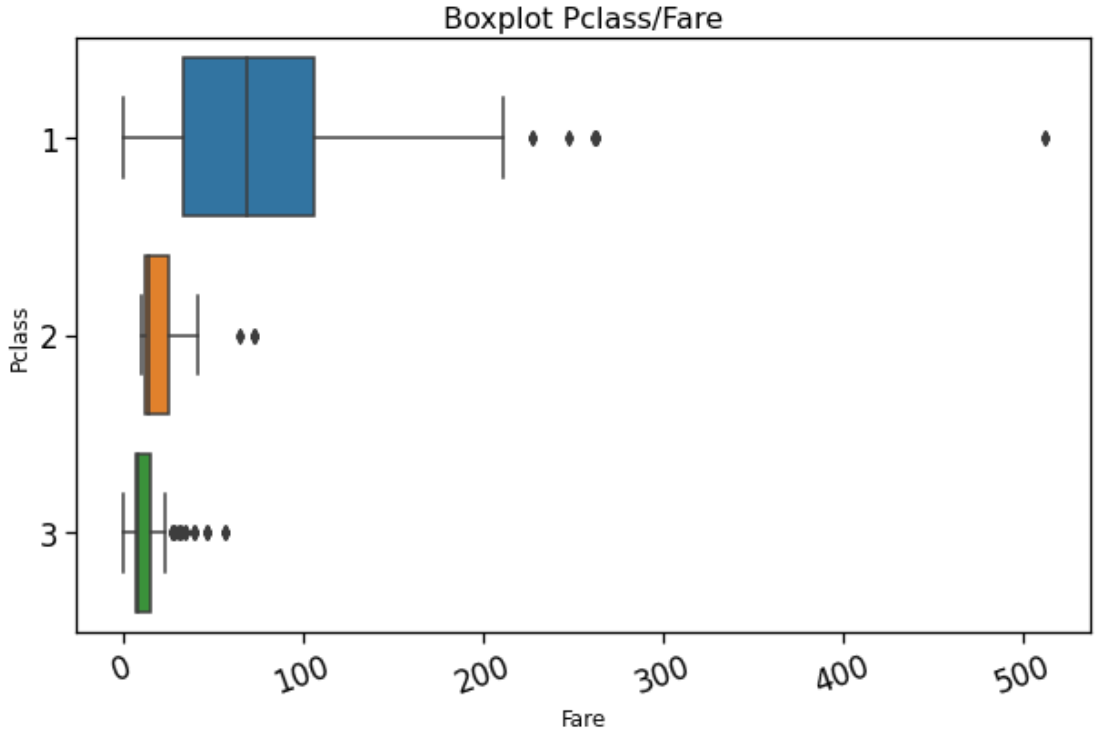


Figura 2 Boxplot per classe

Continuando con l'analisi degli outlier, sono stati costruiti dei nuovi boxplot per Fare in relazione però alle classi dei passeggeri (Pclass). Ogni classe ha mostrato la presenza di outlier. Data la sovrapposizione, per alcune fasce di prezzo, delle diverse classi si ipotizza che il prezzo dipenda da qualche servizio aggiuntivo o dal tipo di cabina, i cui dati sono impossibili da reperire.

Per la terza classe sono stati considerati outlier i valori al di sopra di 35, per la seconda classe quelli superiori a 60, per la prima i valori superiori a 160 ed è stato eliminato l'unico valore isolato pari a 5.

Ulteriore analisi per l'identificazione degli outlier è stata l'applicazione dello z score per determinare i valori inusuali, impostando un threshold pari 2.

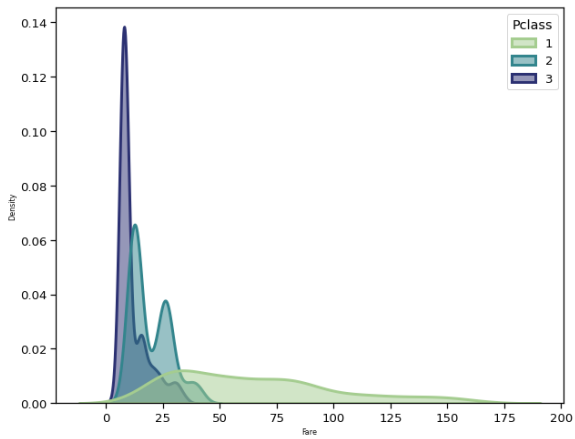


Figura 3 Plot Pclass-Fare

Una volta eliminati totalmente i valori considerati anomali, la distribuzione delle diverse classi per la fascia di prezzo si presenta come nella figura 3. Grazie a questo tipo di visualizzazione è possibile vedere quali siano i prezzi sui quali le varie classi si sovrappongano. Un biglietto di terza classe, quindi poteva essere acquistato a un prezzo anche superiore di quelli di seconda o prima, mentre la fascia di prezzo di prima classe si estende per l'intero asse orizzontale.

Da queste operazioni il dataset presenta 10 classi con 820 valori.

Statistica Descrittiva

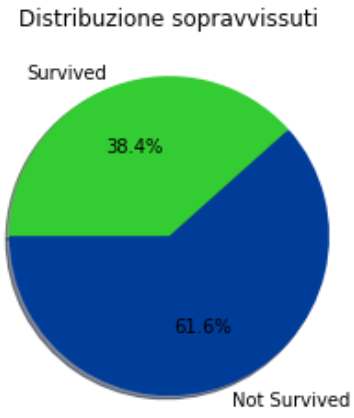
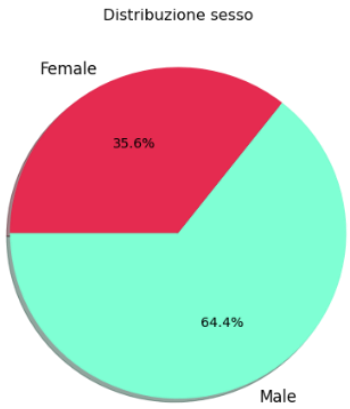


Figura 4 Pie chart Survived

Survived

Il 61,6% dei passeggeri non è sopravvissuto all'incidente del Titanic, mentre il restante 38,4% si. La moda di tale distribuzione è rappresentata dal valore 0 che indica i non sopravvissuti e ha una frequenza di 502.



Sex

Sulla totalità dei passeggeri, i maschi rappresentano la maggioranza, il 64,4% con una frequenza assoluta di 528, le femmine invece rappresentano il 35,6% con frequenza di 292. Sono però quest'ultime a rappresentare la maggioranza dei sopravvissuti.

Figura 5 Pie Chart Sex

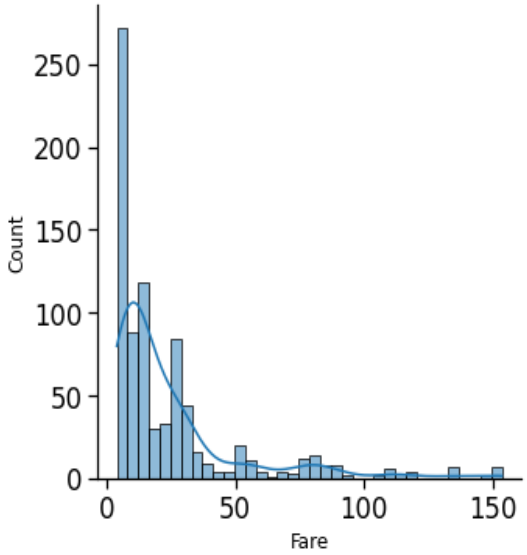


Figura 6 Istogramma Fare

Fare

Il prezzo medio che i passeggeri hanno pagato è di 25.4557, la distribuzione va da un minimo di 4,0125 a un massimo di 153.4625, con mediana pari a 13 e deviazione standard pari a 28.4413. La distribuzione è asimmetrica positiva con un picco che si riscontra nella prima barra del piano per poi scendere rapidamente. Per quanto riguarda gli indici di dispersione, sono stati calcolati anche la varianza pari a 808.9092 e lo scarto assoluto medio (MAD) di 19.1111.

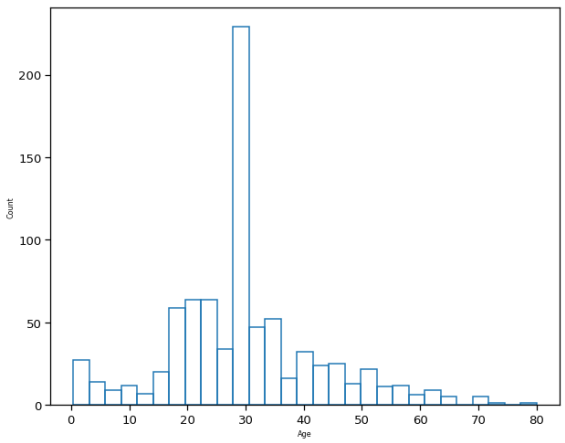


Figura 7 Istogramma Age

Age

L'età media dei passeggeri è di 29,87 anni, che con approssimazione diventa 30. La distribuzione va da un minimo di 1 anno a un massimo di 80, dove la maggior parte della popolazione ha meno di 35 anni. La distribuzione è simmetrica in quanto la media corrisponde con la moda, la cui frequenza è 157, e approssimato a 30 diventa 184.

Questo però deriva anche dalle azioni di rimpiazzo dei valori mancanti con la media. La mediana coincide con la media, mentre presenta una varianza di 172.04, una deviazione standard di 13.12 e MAD di 9.17.

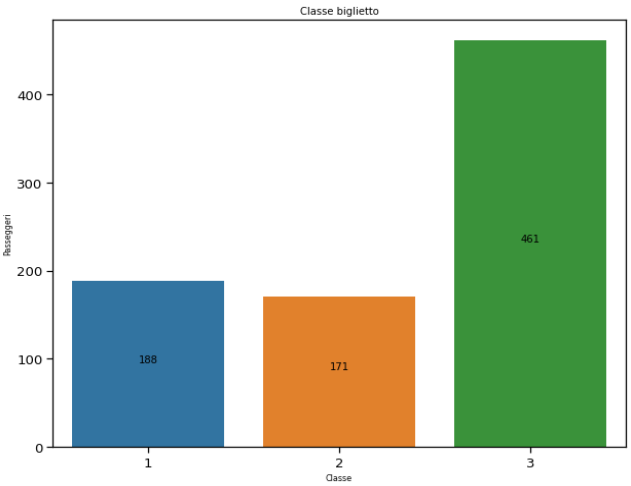


Figura 8 Bar Chart Pclass

Pclass

La maggioranza dei passeggeri (491) viaggiava in 3° classe, mentre una percentuale minore occupava la 1° (188 persone) e la 2° (171). La moda, come si nota anche graficamente è rappresentato dalla 3° classe.

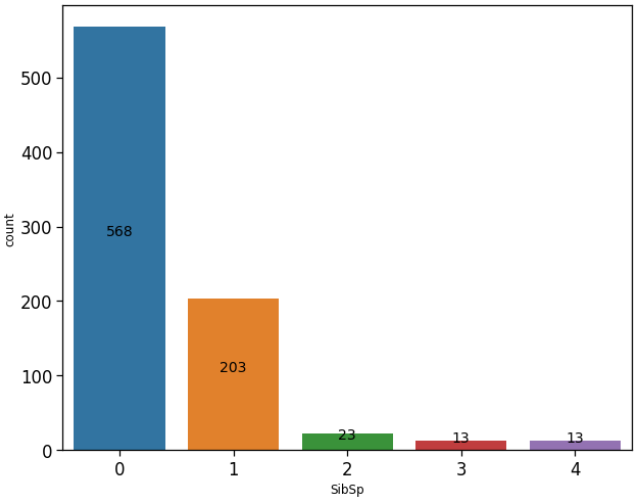


Figura 9 BarChart SibSp

SibSP

La distribuzione assume valori discreti che vanno da 0 a 4; come rappresentato dal grafico, la netta maggioranza dei passeggeri del Titanic (69,27%), viaggia senza fratelli o sorelle, e coniugi. Secondi per numerosità sono coloro che hanno una sola persona. La media della distribuzione è pari 0,4146, dovuto all'altissima presenza di valori pari a 0, che giustifica il valore uguale della mediana. La varianza è di 0.5849, deviazione standard di 0.7648 e MAD di 0,5744.

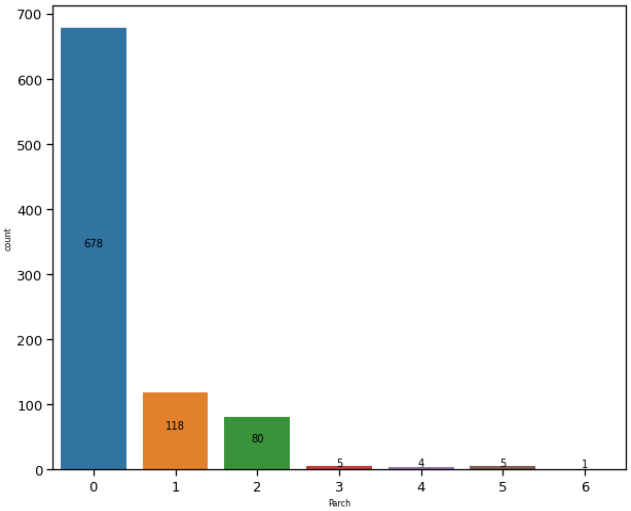


Figura 10 BarChart Parch

Parch

Per tale variabile i valori vanno da un minimo di 0, la percentuale più alta della popolazione, a un massimo di 6, la più bassa con un solo caso. Similmente alla variabile SipSp, la maggioranza viaggia senza figli o genitori al seguito, dove il restante si accompagna di più con una sola persona. La pulizia del dataset ha comportato la rimozione di alcuni valori pari a 8.

Era questo il caso della famiglia Sage, dove i coniugi viaggiavano con i loro 9 figli. In maniera analoga il dataset conteneva per la variabile SibSp 9 record con valore 8. Il calcolo degli indici

di posizione identifica la media come 0,3317, la mediana 0, quello degli indici di dispersione invece restituisce una varianza di 0,5442, deviazione standard pari a 0,7378 e lo scarto assoluto medio 0,5194.

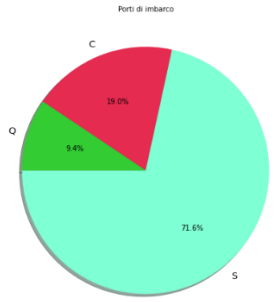


Figura 11 Pie Chart Embarked

Embarked

Su 820 passeggeri, il 71,6% si sono imbarcati a Southampton, nettamente minori sono le percentuali di passeggeri saliti sul Titanic nel porto francese a Chebourg (19%) e irlandese a Queenstown (9,45%). La moda in questo caso è rappresentata dal valore S, che indica il porto inglese.

Correlazione tra le variabili

Le variabili Age, Fare, SibSp, Parch e Pclass sono state correlate tra di loro, utilizzando il coefficiente di correlazione di Parsons. La seguente heatmap evidenzia i risultati ottenuti.

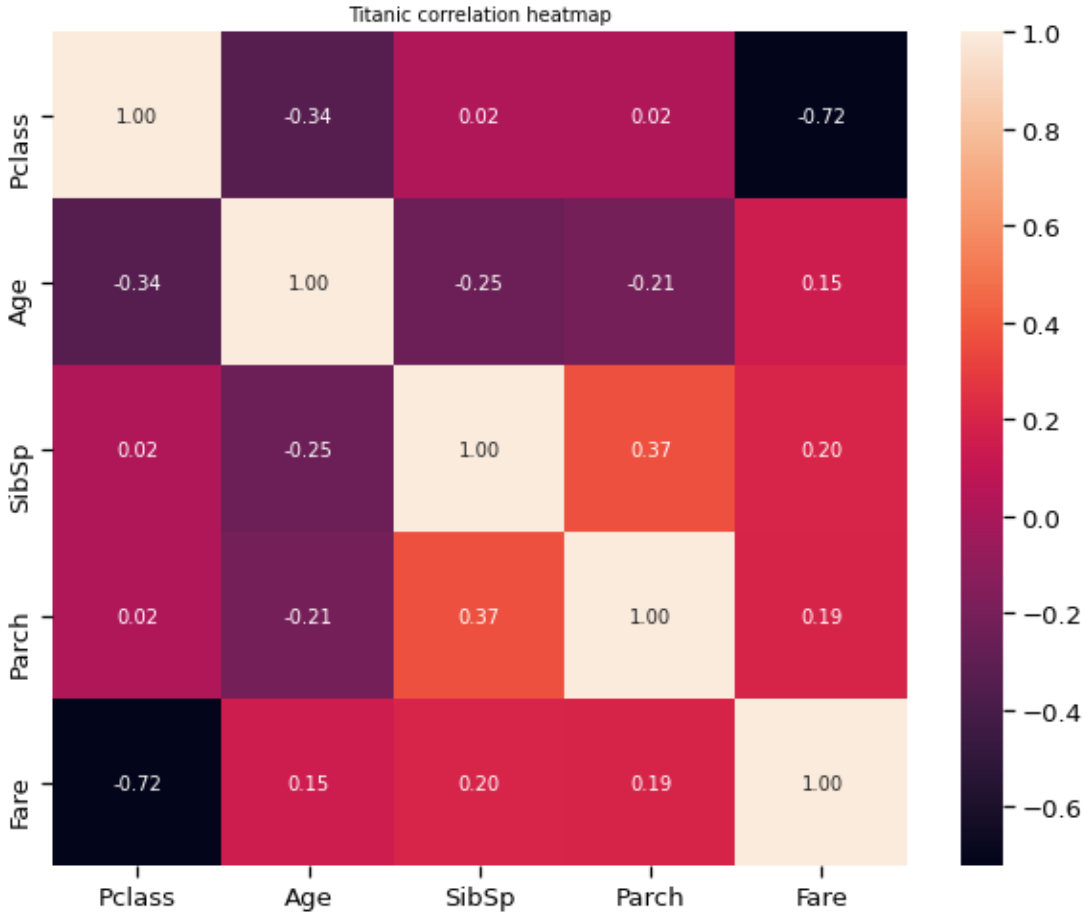


Figura 12 HeatMap correlation

La prima correlazione che salta all'occhio è sicuramente quella tra Fare e Pclass, quindi tra il prezzo del biglietto e della classe dei passeggeri. C'è una correlazione negativa abbastanza alta di -0,72, il che significa che all'aumentare del prezzo diminuisce la classe. Occorre ricordare per non trarre in inganno, che la prima classe, il valore 1, indica la classe per cui è previsto un biglietto più costoso. Tuttavia, il coefficiente non è talmente alto da poter considerare le variabili ridondanti, ma solo correlate.

Seconda per importanza è la correlazione tra SibSp e Parch con un coefficiente di 0,37. La correlazione non è molto alta, ma entrambe le variabili riguardano la sfera dei familiari, vien da sé che sia presente una correlazione grazie a coloro che viaggiano con i propri parenti.

Pare esserci una correlazione negativa anche tra l'età e la classe, questo porta a pensare che all'aumentare dell'età si riduce il valore numerico della classe, quindi aumenta l'esclusività della classe.

Correlazioni minori si hanno per SibSp e Parch con Age, con coefficienti di -0,25 e -0,21, con Fare con valore 0,20 e 0,19 e tra Age e Fare con 0,19.

Per andare a fondo nell'analisi della relazione tra variabili, quelle più interessanti sono state rappresentate graficamente. In questo modo possono essere comprese le relazioni anche tra le variabili categoriche per le quali non può essere calcolato il coefficiente.

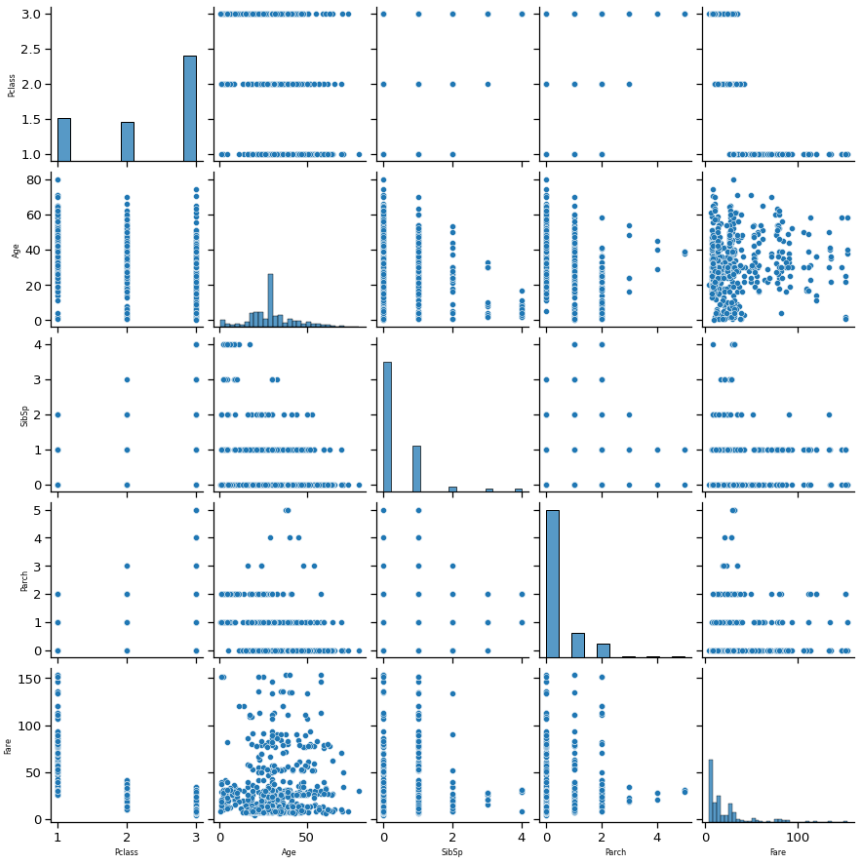


Figura 13 PairPlot variabili

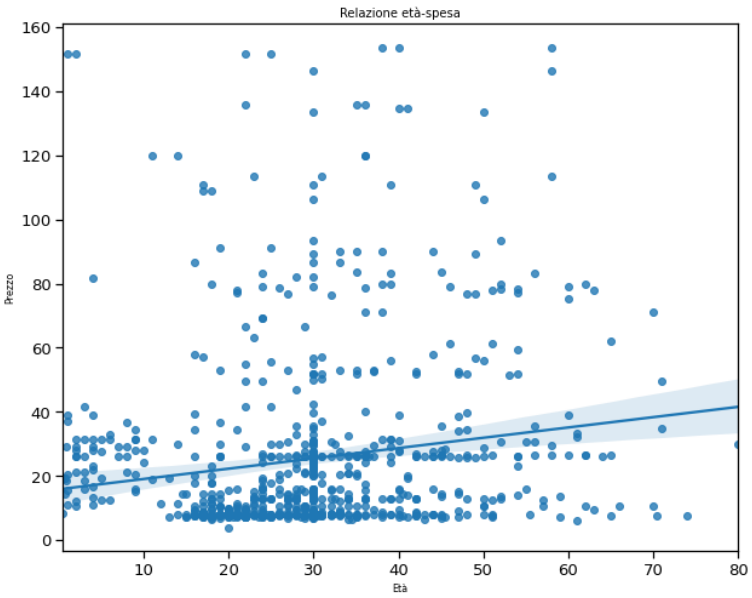


Figura 14 Scatter Plot con interpolazione lineare

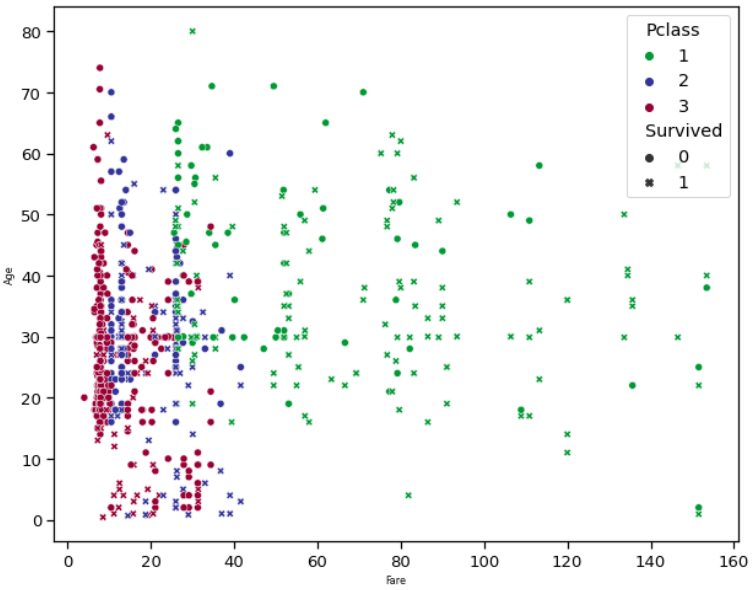


Figura 15 Plot multiclasse

Nella figura 14 viene presentato lo scatter plot con regressione lineare per le variabili Age e Fare. Anche graficamente vediamo come siano molti i punti che si discostano dalla retta di interpolazione. Il dataset inoltre non sembra seguire l'andamento della retta. Viene così confermata la scarsa correlazione tra le due variabili.

La figura 15 mostra la relazione tra le variabili Pclass, Age, Fare e Survived. La terza classe presenta il più alto numero di non sopravvissuti (Survived=0 75.488%), ma in relazione all'età, solo al di sotto dei 10 anni la mortalità si riduce. Per la seconda classe il numero di superstiti è quasi uguale a quello degli scomparsi, anche in questo caso sotto i 10 anni la variabile Survived ha valore 1. Il grafico mostra inoltre come i passeggeri della prima classe siano quasi interamente sopravvissuta.

Nel complesso le varie classi ricoprono quasi tutto il range di età, solo per la prima classe sono pochissimi i passeggeri che viaggiano al di sotto dei 10 anni, in totale 3 di cui 2 sopravvissuti e 1 no. La distribuzione per questa classe è sparsa, ha una bassa densità di 0.20, mentre le altre classi sono più (3 classe) e meno dense (2 classe).

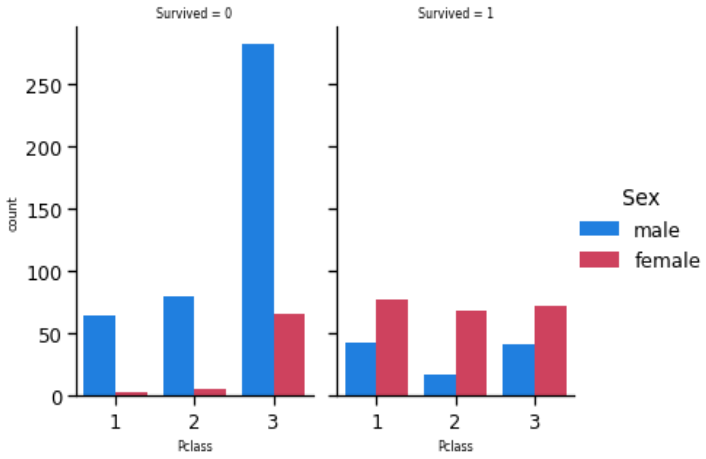


Figura 16 Istogramma Pclass-Sex-Survived

Gli istogrammi in figura 16 consentono una rappresentazione distinta per i sopravvissuti o non, distinguendo in base alla classe e al sesso. Considerando il numero totale di passeggeri, come anche si era visto precedentemente nella fase di analisi descrittiva, il numero di maschi è nettamente superiore a quello delle donne. Tuttavia, ciò che è interessante è vedere come nella prima e nella seconda classe le donne siano quasi totalmente sopravvissute, soprattutto in relazione al numero di uomini. Si presuppone, quindi, che vi sia stata una preferenza nel permettere a queste di salvarsi. Il tasso di donne decedute più alto riguarda la terza classe, dove anche il numero di uomini con cui condividono la sorte è altissimo. Questo può riguardare la loro posizione all'interno della nave che non ha permesso a questi di salvarsi.

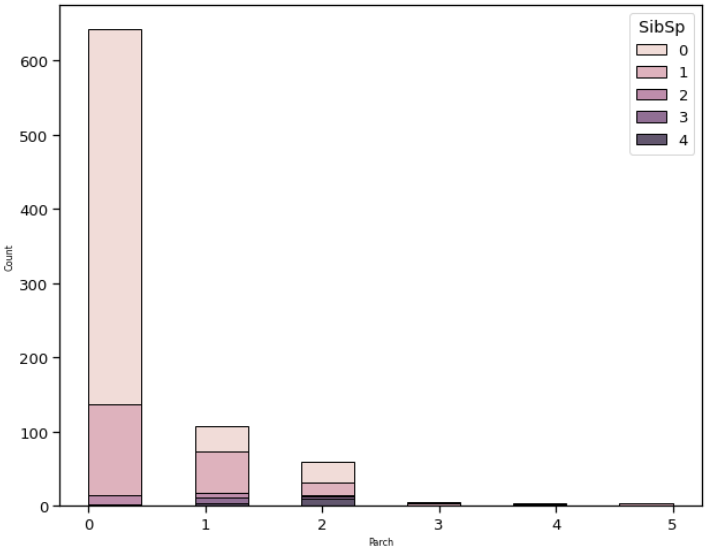


Figura 17 Istogramma Parch-SibSp

In questo grafico vengono relazionate le due variabili relative alla sfera familiare, per le quali il calcolo della correlazione ha fornito un coefficiente tale da poter ipotizzare una correlazione. Avendo una visualizzazione grafica, è possibile analizzare in che modo queste siano collegate tra loro. Sono circa 500 i passeggeri che viaggiano senza familiari, quindi Parch e SibSp uguale a 0.

Tra gli accompagnati, la maggior parte ha un solo fratello, sorella o coniuge, sia per coloro che non hanno genitori o figli, sia per chi ne ha 1 o 2. Valori minimi si hanno man mano che il numero di accompagnatori sale, di entrambe le tipologie. Coloro che hanno al loro seguito 5 tra figli e genitori, solo 3 su 4 hanno un appartenente all'altra classe. In sintesi, sono poco frequenti le famiglie con molti componenti (genitori e figli, fratelli e coniugi). Di seguito, con la figura 18, si pone attenzione anche sulla variabile Survived, per capire se il fatto di avere una famiglia a bordo possa in qualche modo influenzare il valore della variabile target.

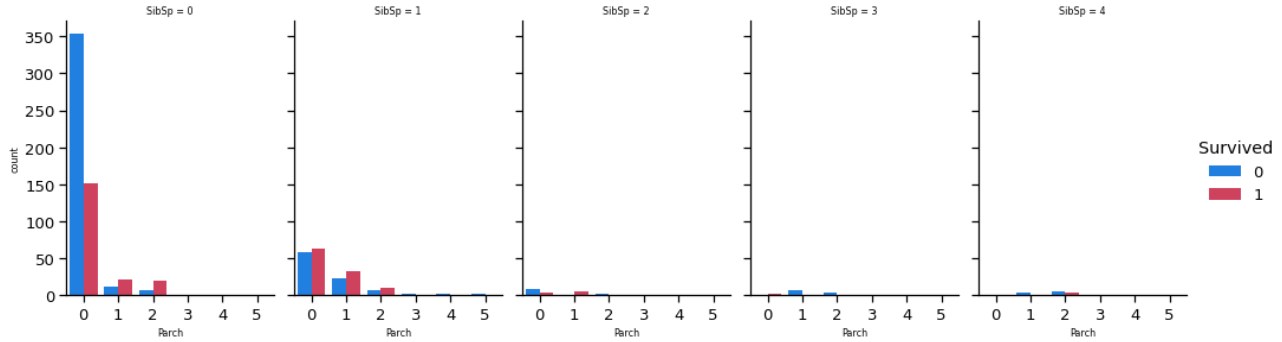


Figura 18 CatPlot Parch-SibSp-Survived

Diversamente da quanto si ipotizzava il grafico ha confermato che aumentando il numero di figli o genitori non c'è un aumento considerevole. Se solo prendessimo in considerazione le osservazioni con Parch pari a 1,2 e 3 il numero di sopravvissuti è maggiore, talvolta anche di poco, tuttavia le frequenze non sembrano tali da poter costituire una regola.

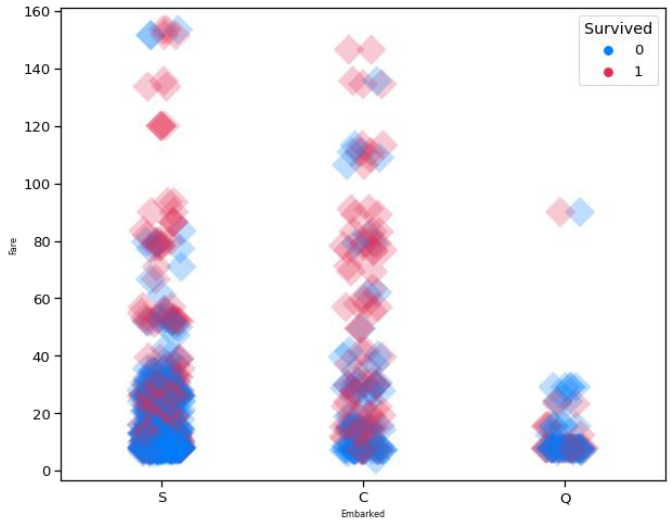


Figura 19 StripPlot Embarked-Fare-Survived

Un ulteriore quesito che può essere risolto è quello relativo al porto di imbarco, il prezzo e se questo possa cambiare in relazione al primo.

Non sembrerebbe essere applicato un sovrapprezzo relativo alla lunghezza del viaggio, solo per quanto riguarda Queenstown non ci sono quasi passeggeri che hanno pagato al di sopra delle 40 sterline, ma questo non sembra in linea con l'analisi in corso.

La tendenza alla sopravvivenza è simile per tutti i porti, c'è un maggior tasso di sopravvissuti quanto più si alza il prezzo, ma è costante la superiorità di non sopravvissuti in ogni porto.

Conclusioni

In seguito a tutte le analisi statistiche qui presentate possiamo determinare che, stando ai dati inseriti nel dataset, ci sono stati più morti che sopravvissuti. Per quanto riguarda questi ultimi possiamo vedere come appartengano per di più a classi alte, e che ci sia stata una tendenza a salvare di più le donne e i bambini.

Il sesso non sembra influenzare le altre variabili, mentre il porto di imbarco non sembra avere alcuna influenza sulle altre variabili.

Il grafico di seguito tende a confermare tali conclusioni.

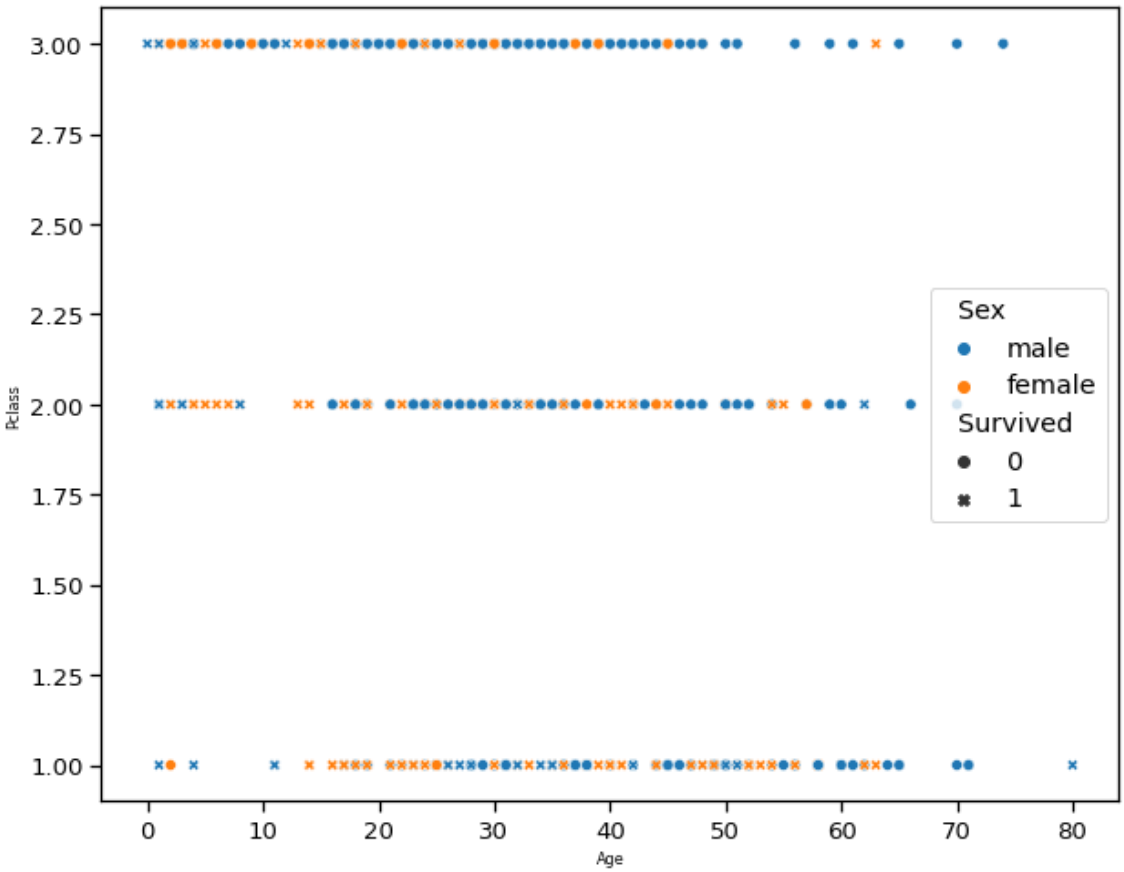


Figura 20 Grafico finale