

The Titanic sink

Analisi esplorativa sul naufragio del Titanic

Studenti:

Luca Federico

Puglisi Leone Emanuel

Relatori:

Elena Reas

Davide Di Grande

Introduzione

L'RMS Titanic è stato un transatlantico britannico della classe Olympic, divenuto celebre per essere naufragato nelle prime ore del 15 aprile 1912, durante il suo viaggio inaugurale, a causa della collisione con un iceberg, costando la vita a 1502 persone tra un totale di 1317 passeggeri e 892 membri dell'equipaggio.*

L'obiettivo è sviluppare un'analisi esplorativa in grado di determinare se, tra i passeggeri, ci fossero possibilità più o meno alte di sopravvivere rispetto ad altri, in base alle loro caratteristiche.

Tecnologie usate

Per la nostra analisi ci serviremo delle seguenti librerie, facenti parte dell'ecosistema di programmazione Python:

- Pandas per la manipolazione e l'analisi dei dati;
- Matplotlib per la visualizzazione dei dati in forma grafica.



Variabili

Piccola classificazione delle variabili all'interno del Dataset in modo da fare mente locale e capire cosa ci potrà essere utile (e in che modo).

Categoricals:

- Sex (male/female);
- Pclass (3rd/2nd/1st);
- Embarked (port);
- Survived (true/false).

Ordinals:

- Pclass (3rd/2nd/1st);
- Sibsp (siblings+spouse);
- Parch (parents+children).

Continual:

- Age;
- Fare (price of ticket).

Everything else:

- PassengerId;
- Name;
- Cabin.



Fasi di Analisi

- Controllo sui valori nulli nel set di dati per evitare che compromettano la bontà dell'analisi.
- Regressione logistica per determinare l'impatto di variabili demografiche, socioeconomiche e comportamentali sulla probabilità di sopravvivenza, utilizzando il parametro “Survived” come variabile dipendente.
- Analisi delle relazioni tra “Survived” e:
 - Genere;
 - Classe di appartenenza (prima, seconda o terza classe);
 - Età.
- Rappresentazione grafica dei dati ottenuti e conclusioni.



Got some null?

Come possiamo vedere abbiamo valori
mancanti solo sulle colonne

- Età (Age);
- Porto d'imbarco (Embarked).

```
DataFrame(df.isna().sum(), columns=['']).T
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	0	0	0	0	177	0	0	0	0	687	2

L'età è un parametro importante per la nostra analisi e per questo cercheremo di recuperarla.

Vi sarebbero diversi approcci matematici e statistici: la scelta è infine ricaduta su un approccio informato. Ma ci torneremo...



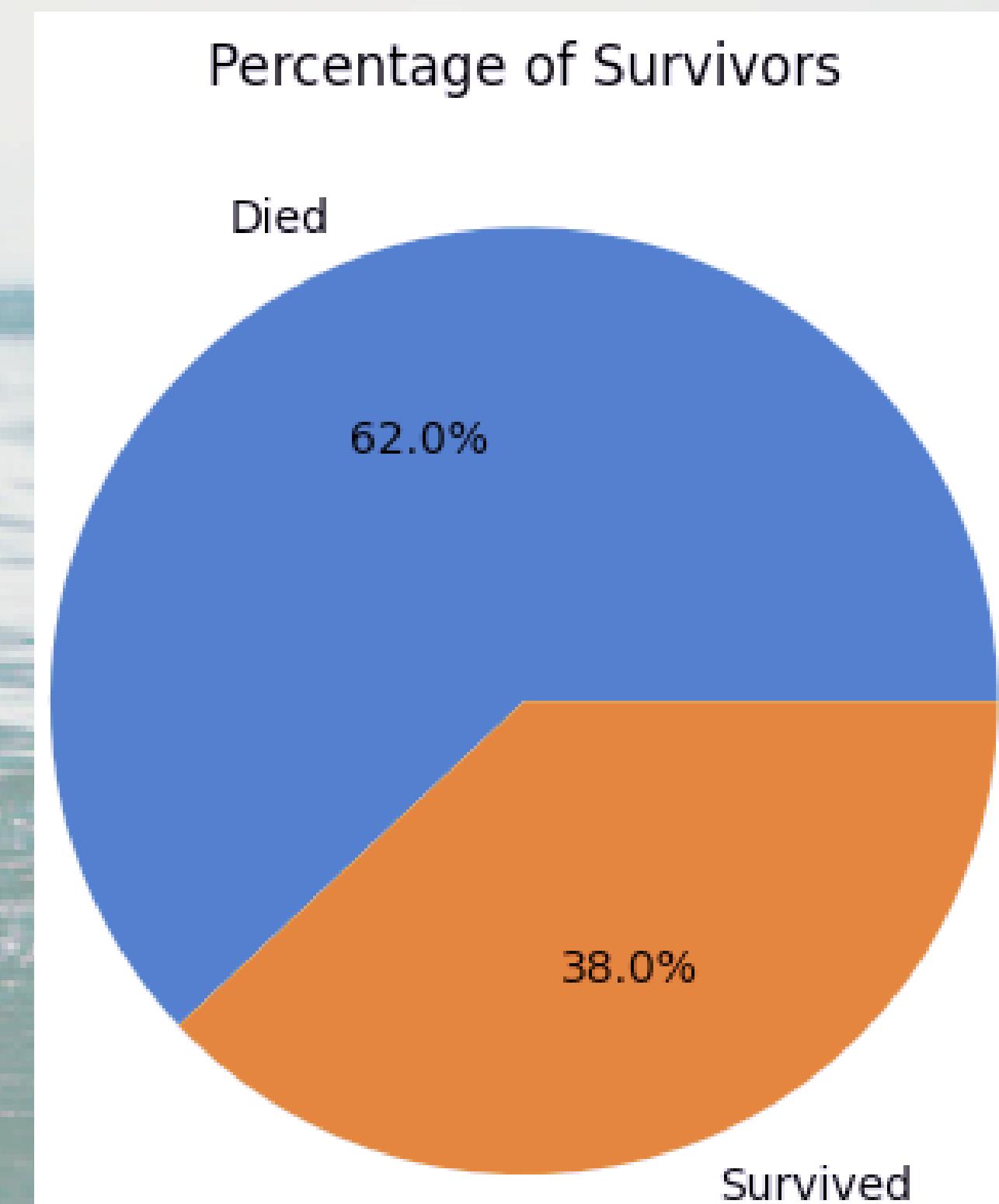
Name	Sex	Age
Braund, Mr. Owen Harris	male	22.0
Cumings, Mrs. John Bradley (Florence Briggs Th...)	female	38.0
Heikkinen, Miss. Laina	female	26.0
Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0
Allen, Mr. William Henry	male	35.0

I will survive!

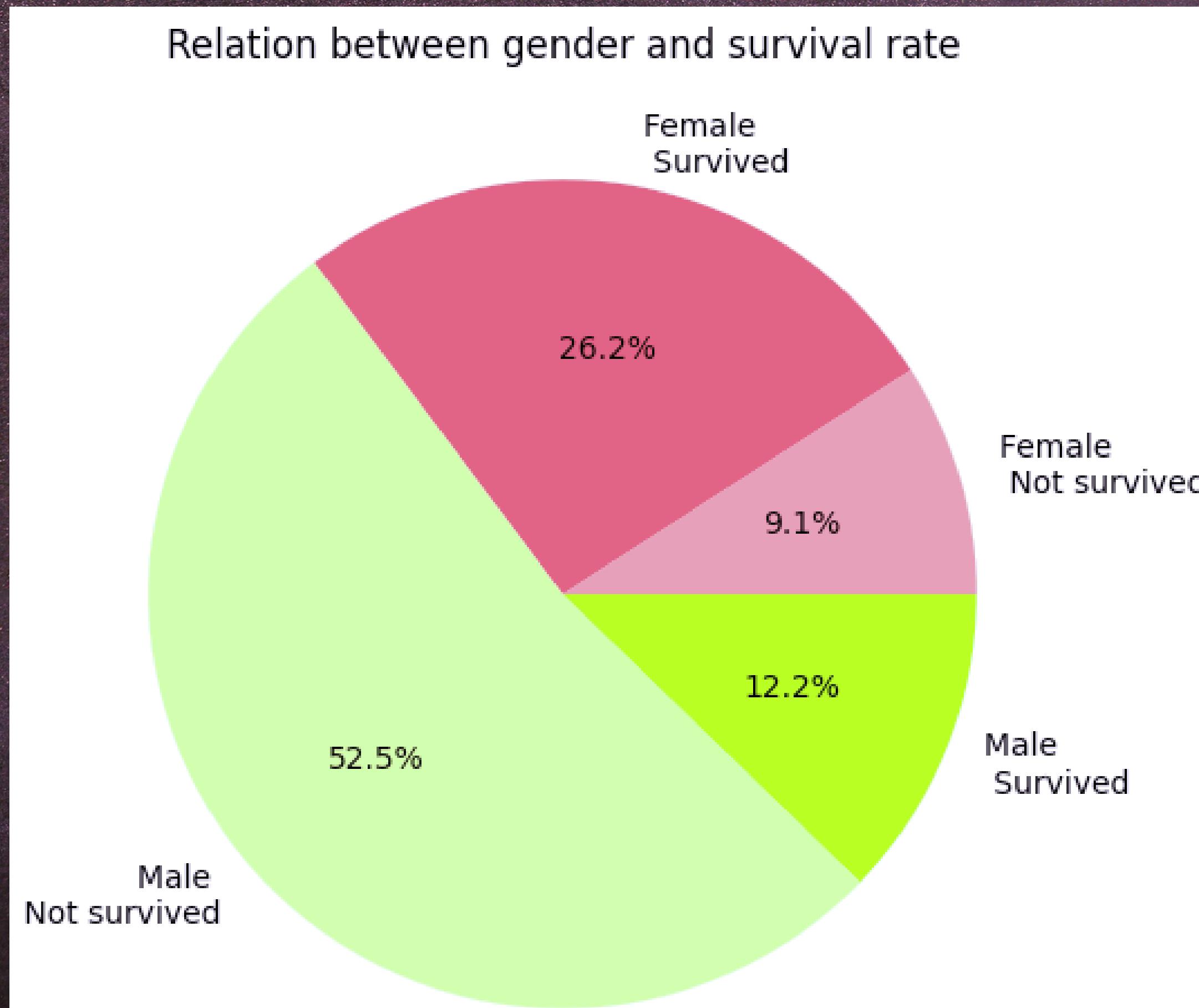
La seconda parte della nostra analisi prevede l'esplorazione della variabile relativa ai sopravvissuti "Survived" e le correlazioni con le altre caratteristiche dei passeggeri, per capire se e come una di queste abbia contribuito ad aumentare le percentuali di salvezza.



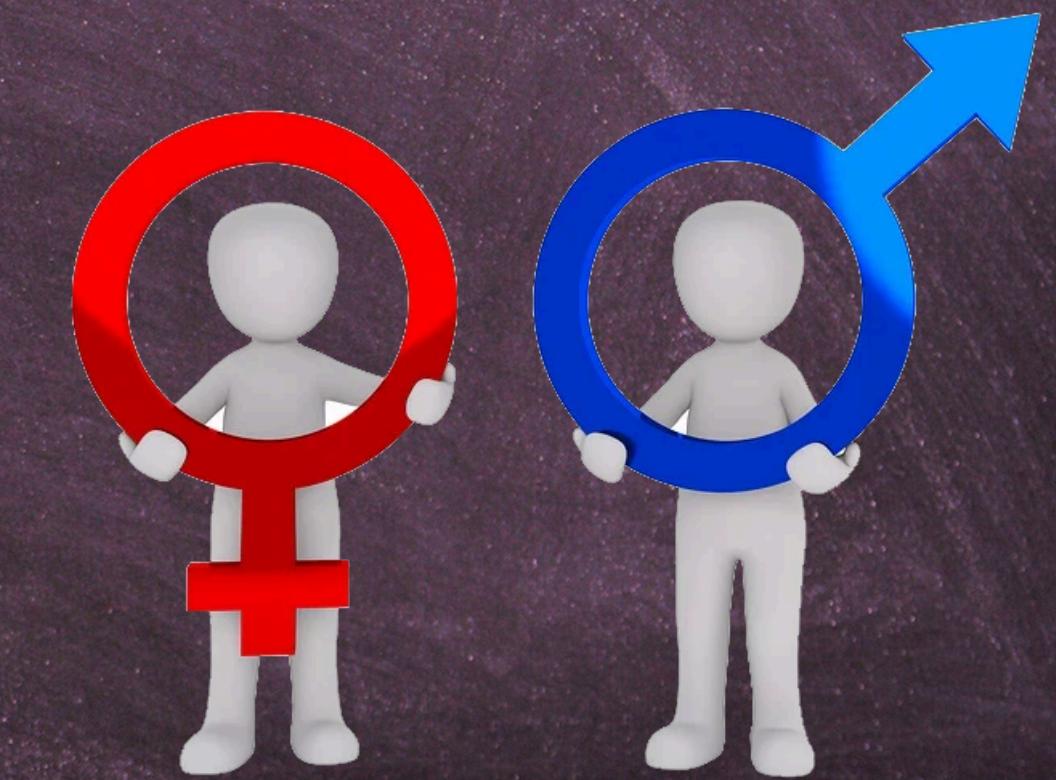
Purtroppo dai dati ricavati sembra che una vasta maggioranza delle persone sia morta nel naufragio.



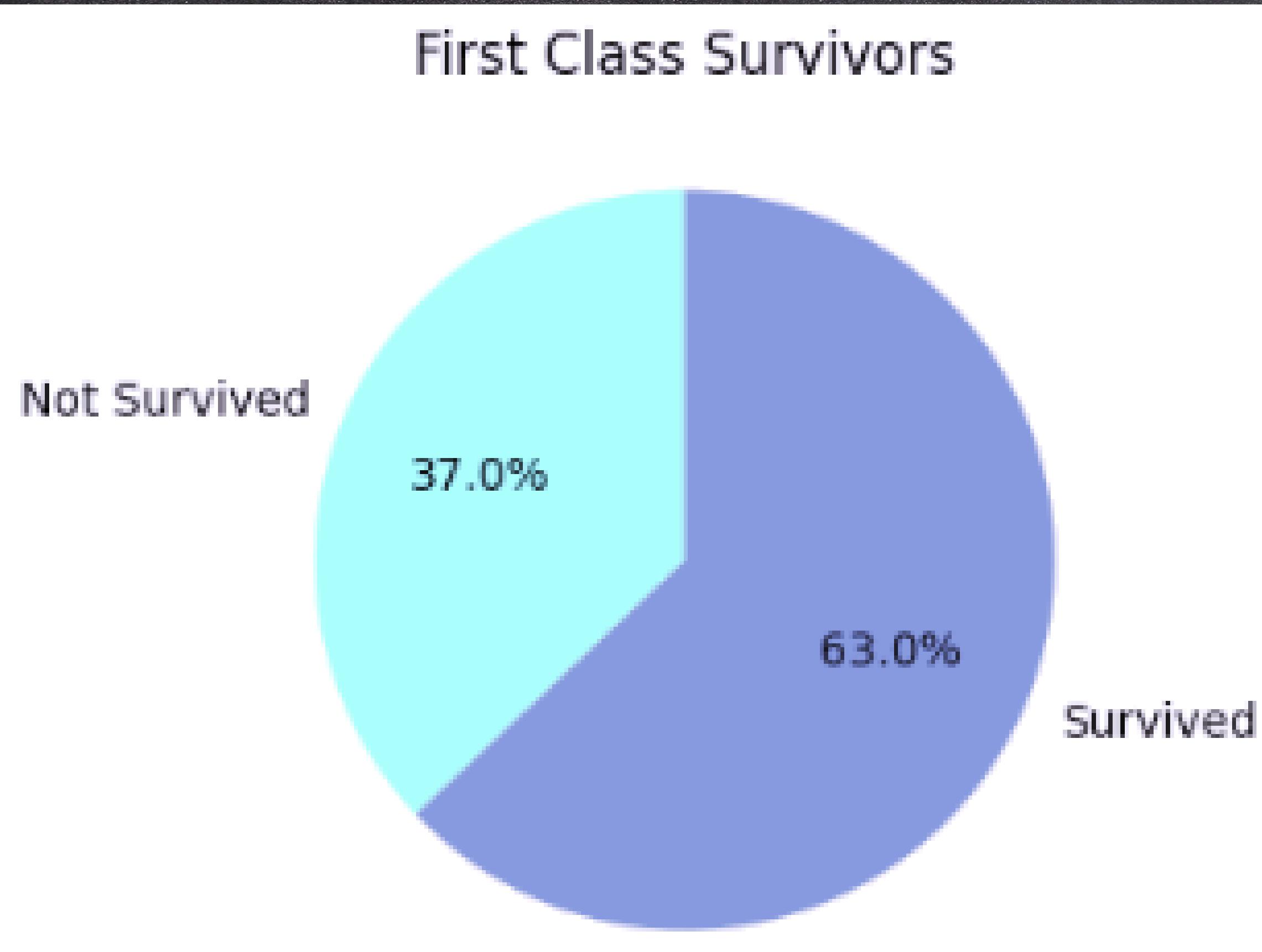
Variabile 1: Genere



Come si vede dal grafico è evidente la differenza tra maschi e femmine in termini di sopravvivenza.



Variabile 2: Classe

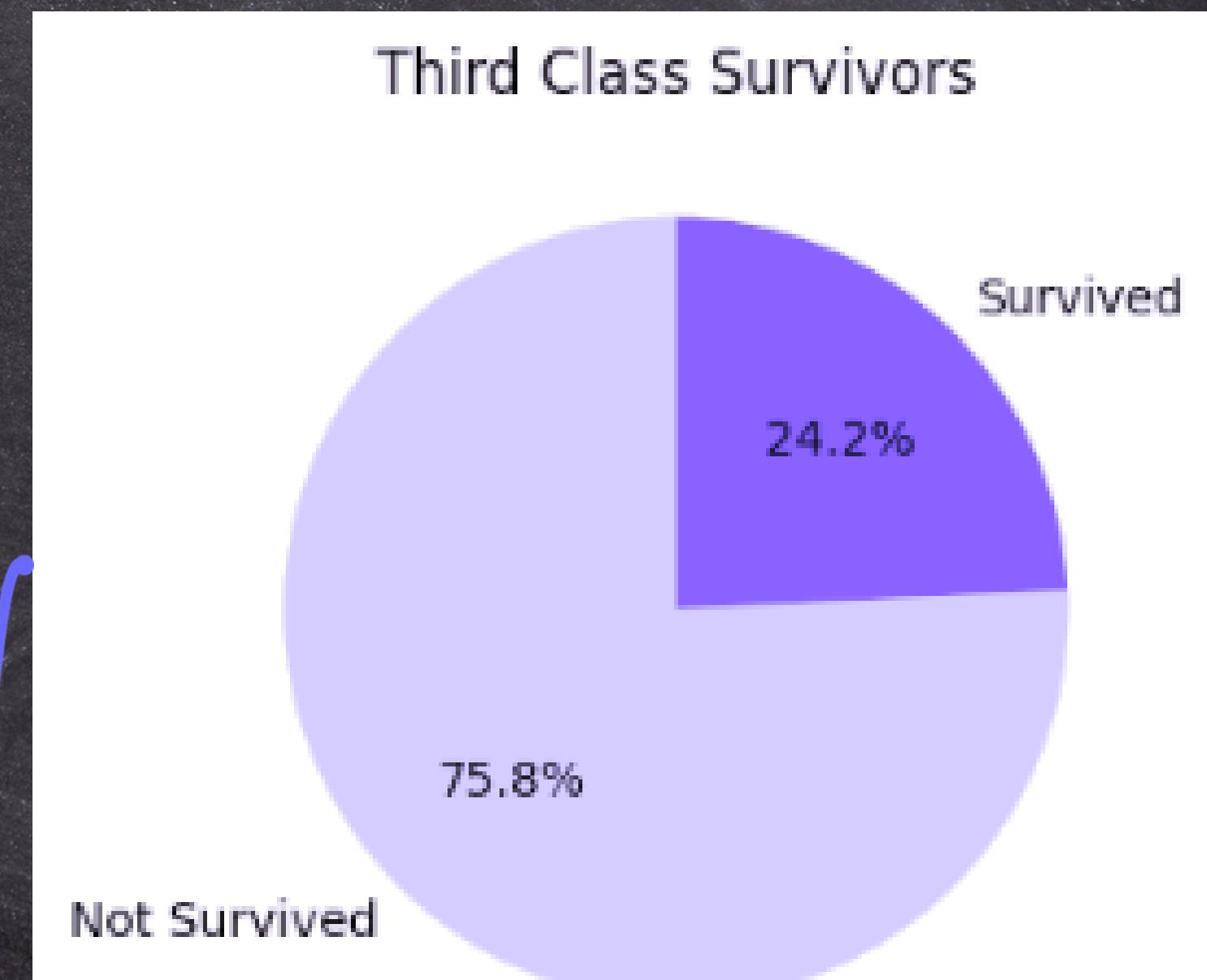
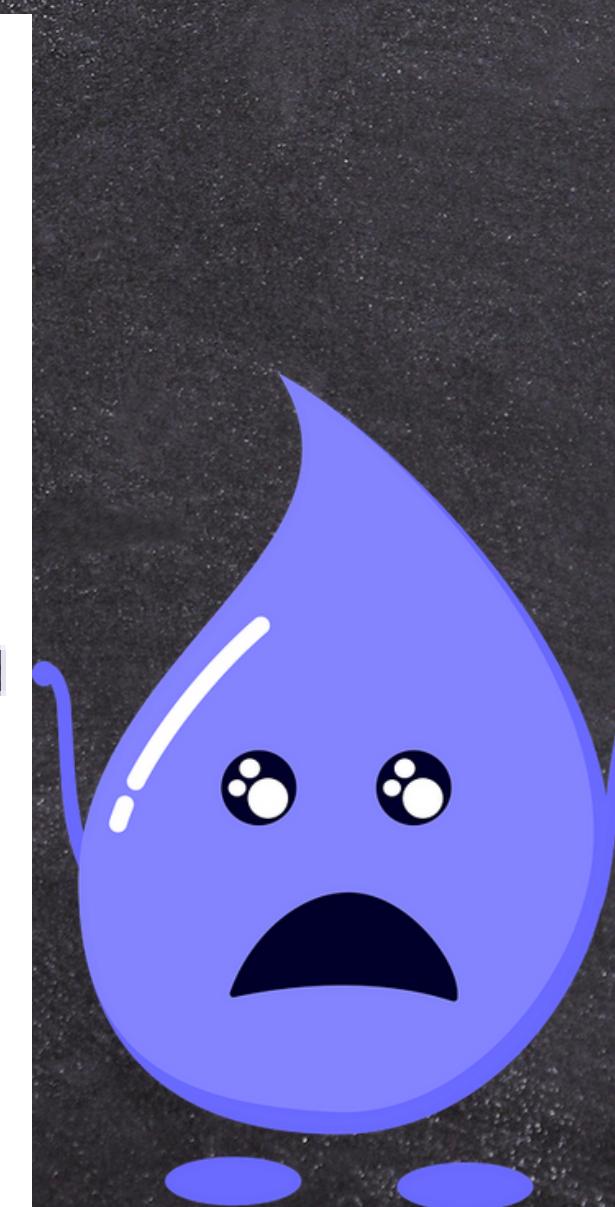
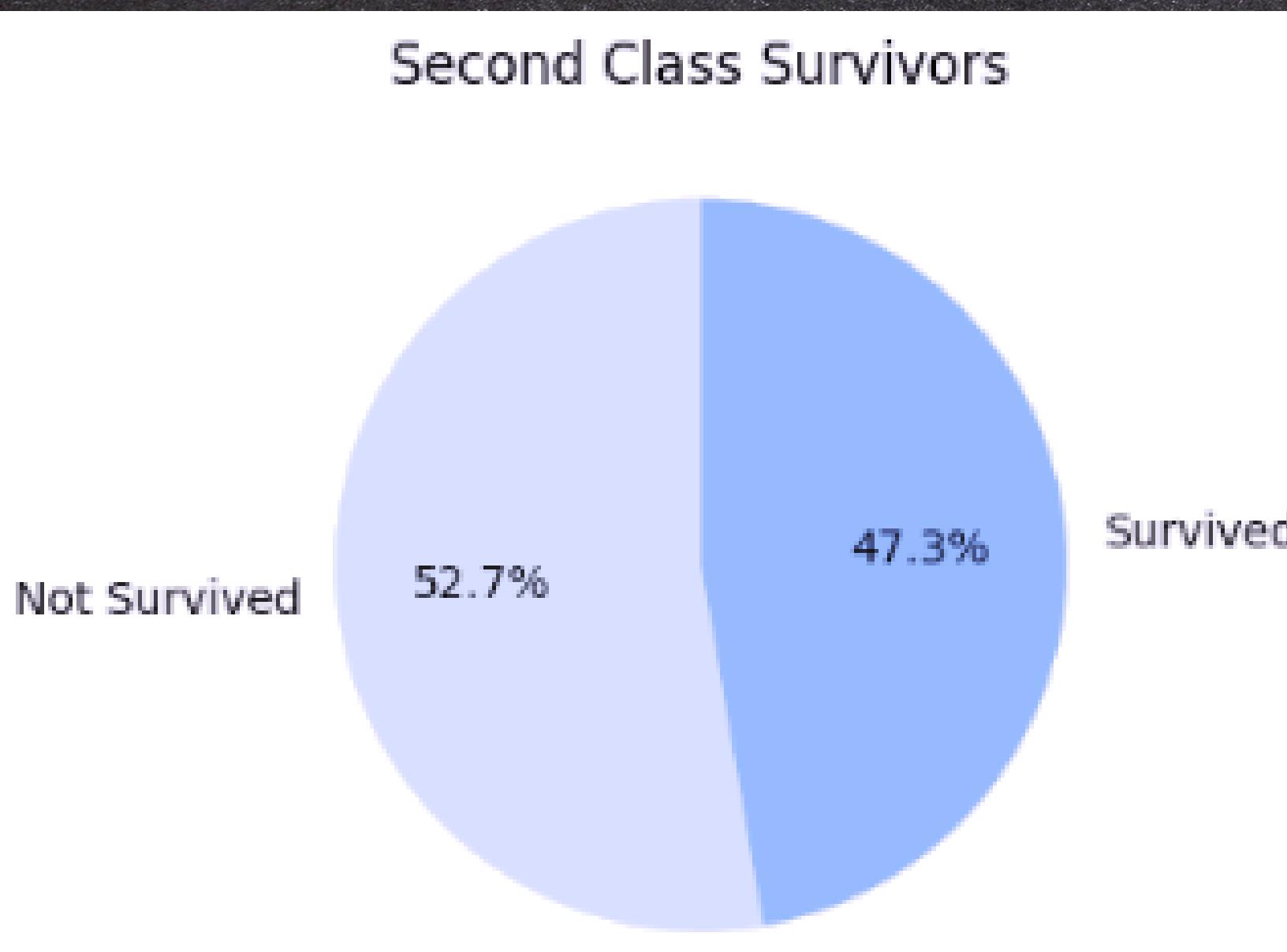


I biglietti di prima classe possiamo assumere appartenessero a persone di rango sociale elevato che sono in gran parte sopravvissuti. Intervento tempestivo dello staff? Posizione favorevole?



Variabile 2: Classe... parte 3

Scendendo di classe la situazione peggiora, fino al punto in cui, le persone detentrici del biglietto più economico, si sono ritrovate ad essere le più sfortunate di tutte, con una percentuale di sopravvivenza appena sopra il 24 %.

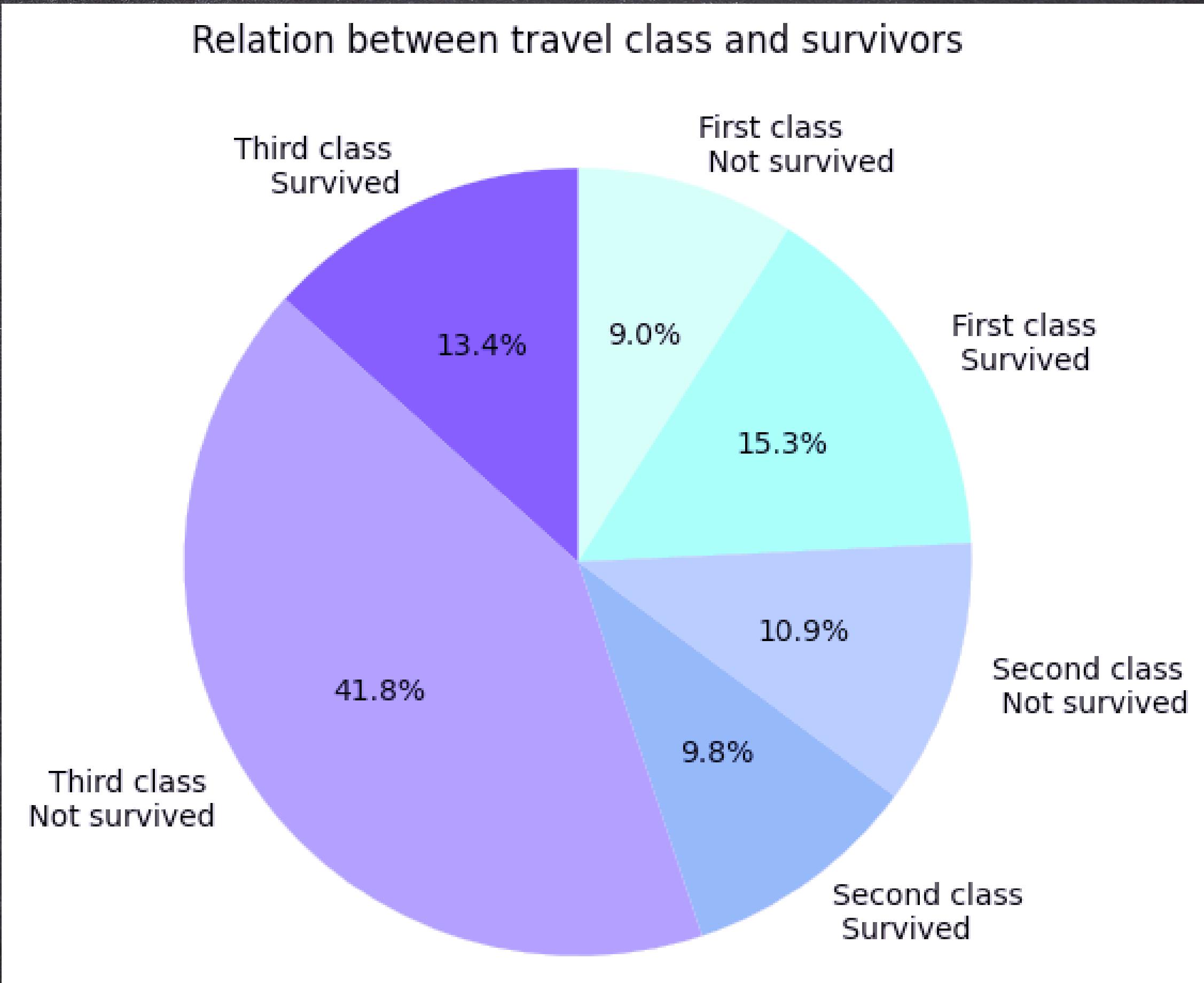


Variabile 2: Classe + Genere

Cosa succede se mettiamo assieme i dati appena ottenuti con quelli relativi alla sopravvivenza per genere?

Apparentemente lo status sociale ha influito in maniera tutt'altro che superficiale nella sopravvivenza.

Ma sarebbe valido anche l'inverso?



Variabile 2: Classe Femminile

Female First Class Survivors

Not Survived

3.2%

96.8%

Survived

Female Second Class Survivors

Not Survived

7.9%

92.1%

Survived

Female Third Class Survivors

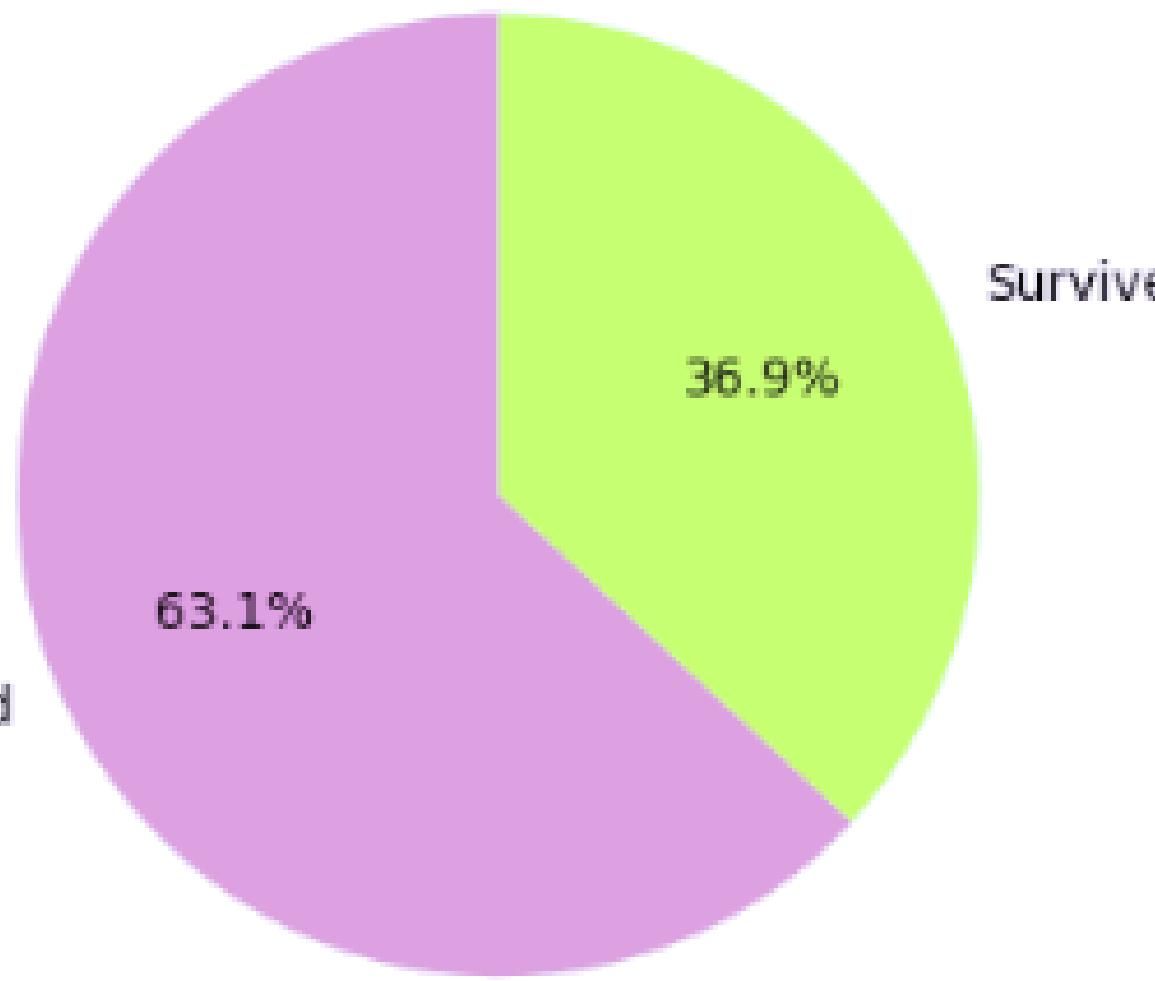
50.0%

50.0%

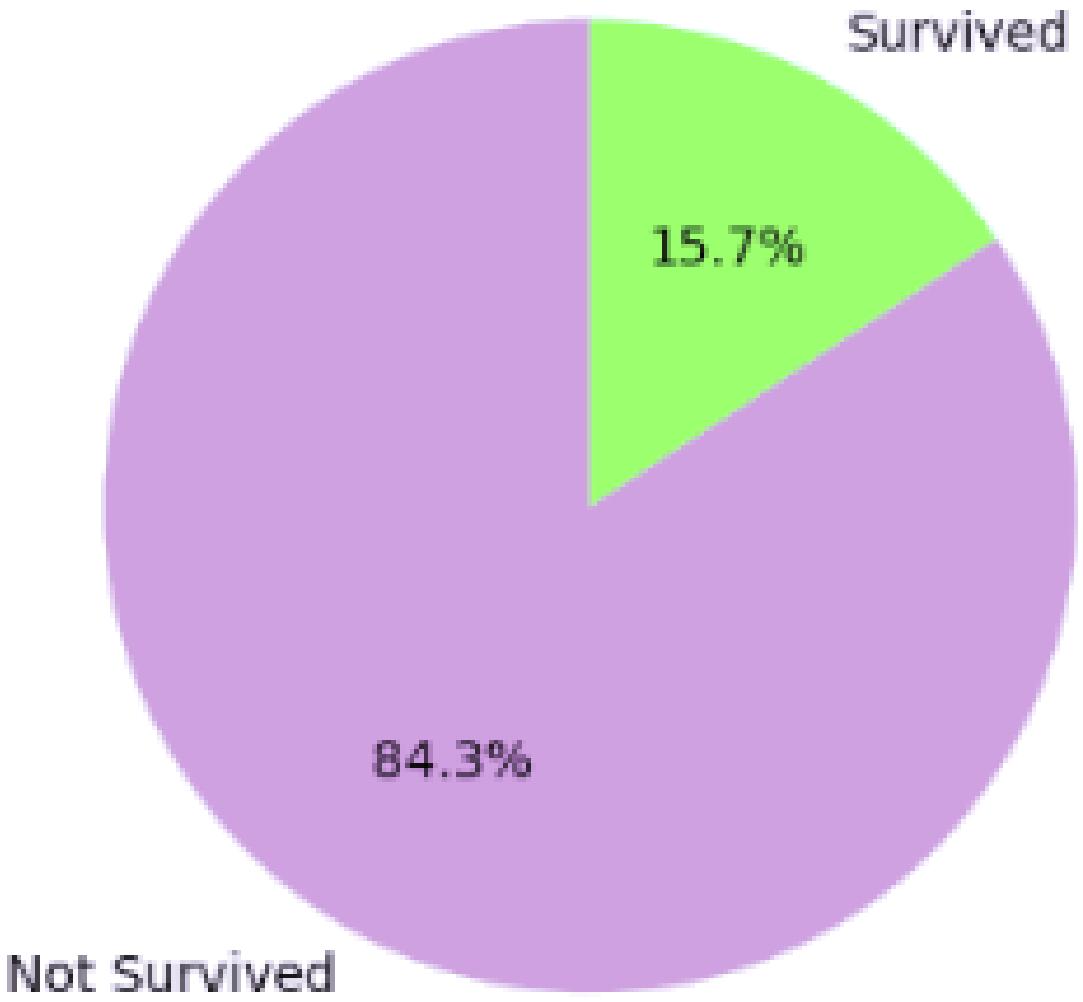
Poche sorprese qui. Le donne di prima e seconda classe sono quasi tutte tornate a casa. Peggior sorte per le passeggerie della terza.

Variabile 2: Classe Maschile

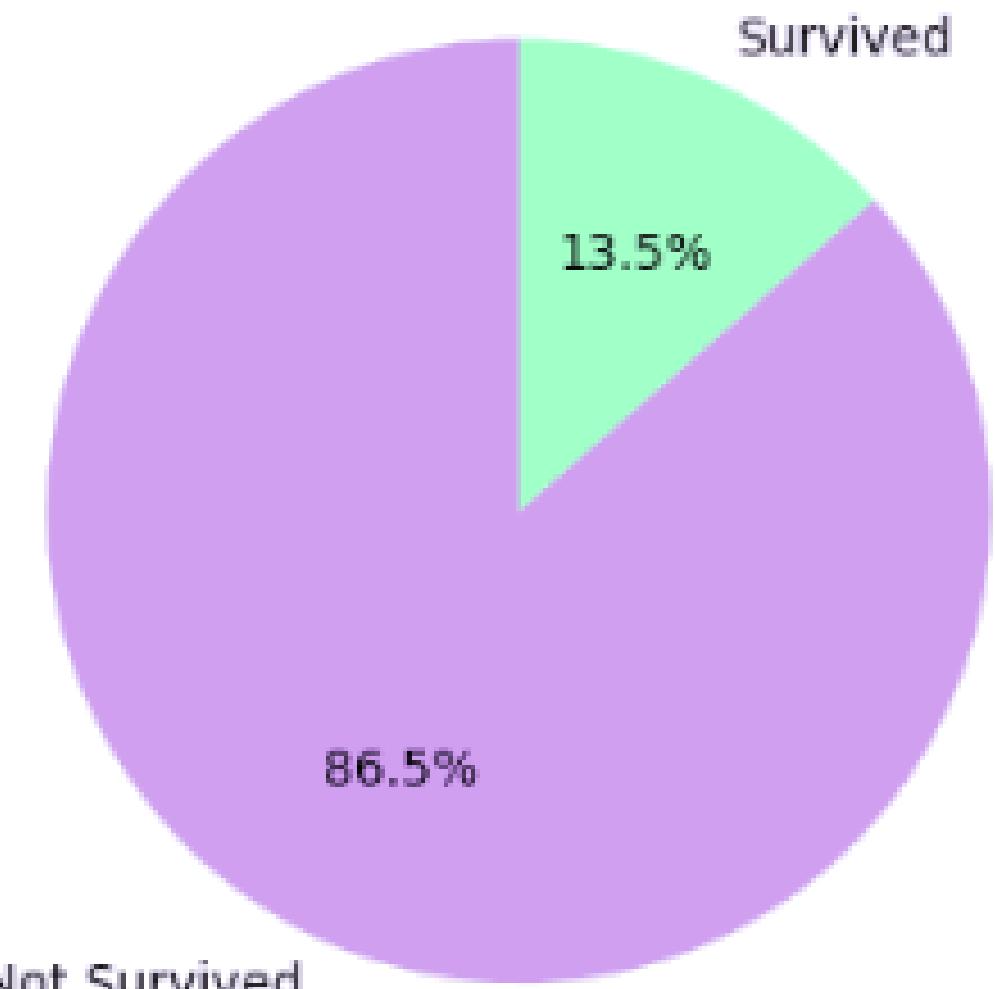
Male First Class Survivors



Male Second Class Survivors



Male Third Class Survivors



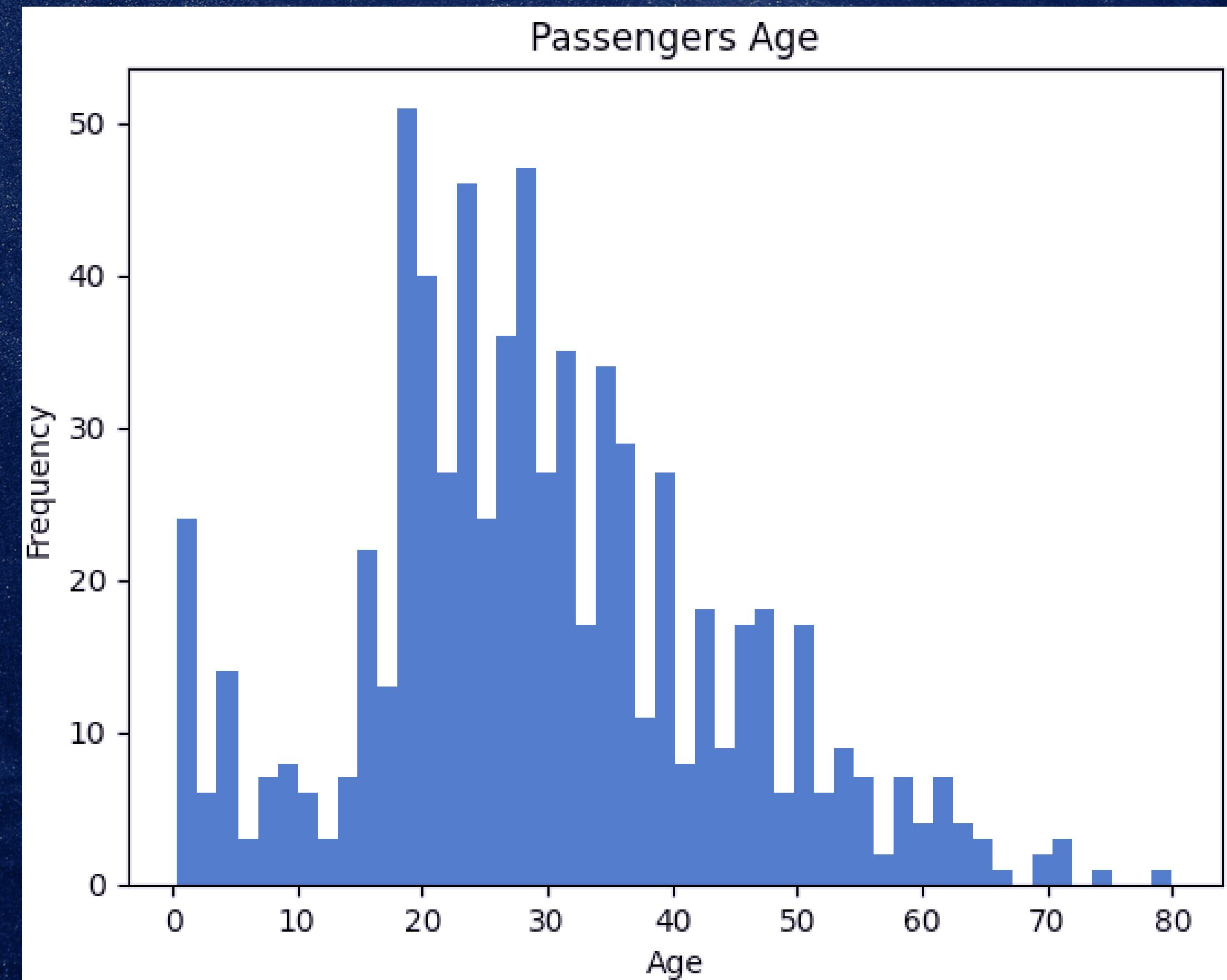
In questi grafici la differenza di sopravvivenza tra classi è meno marcata rispetto alla controparte femminile, che quindi sembra aver beneficiato in misura maggiore del proprio status sociale più elevato, quando presente.



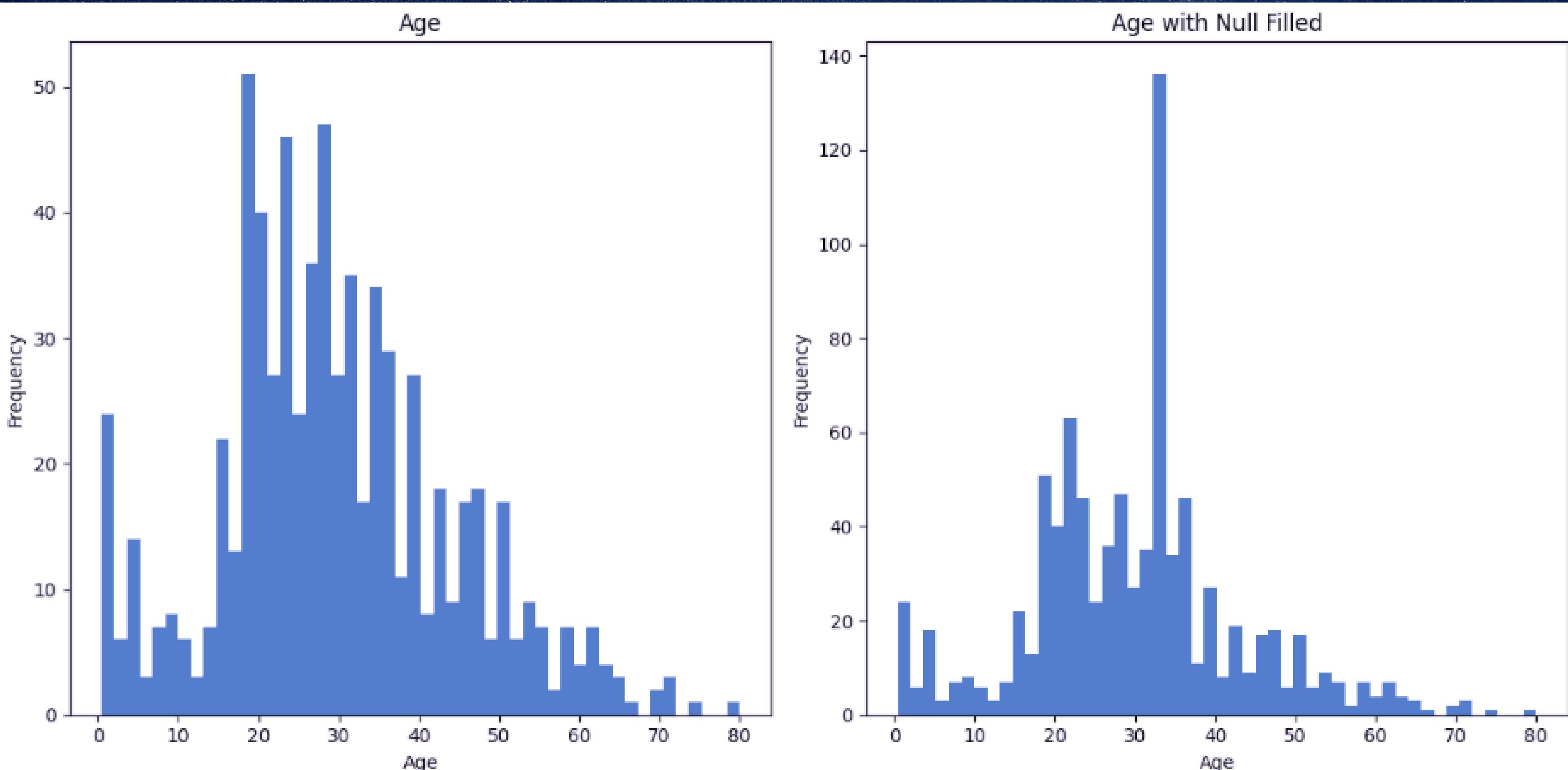
Variabile 3: Età

Nel rilevare l'età abbiamo notato un'insolita quantità di zeri. Come visto qualche diapositiva fa, il dataset presentava molti valori nulli nel campo età.

La cosa importante quindi diventa gestire questa carenza.



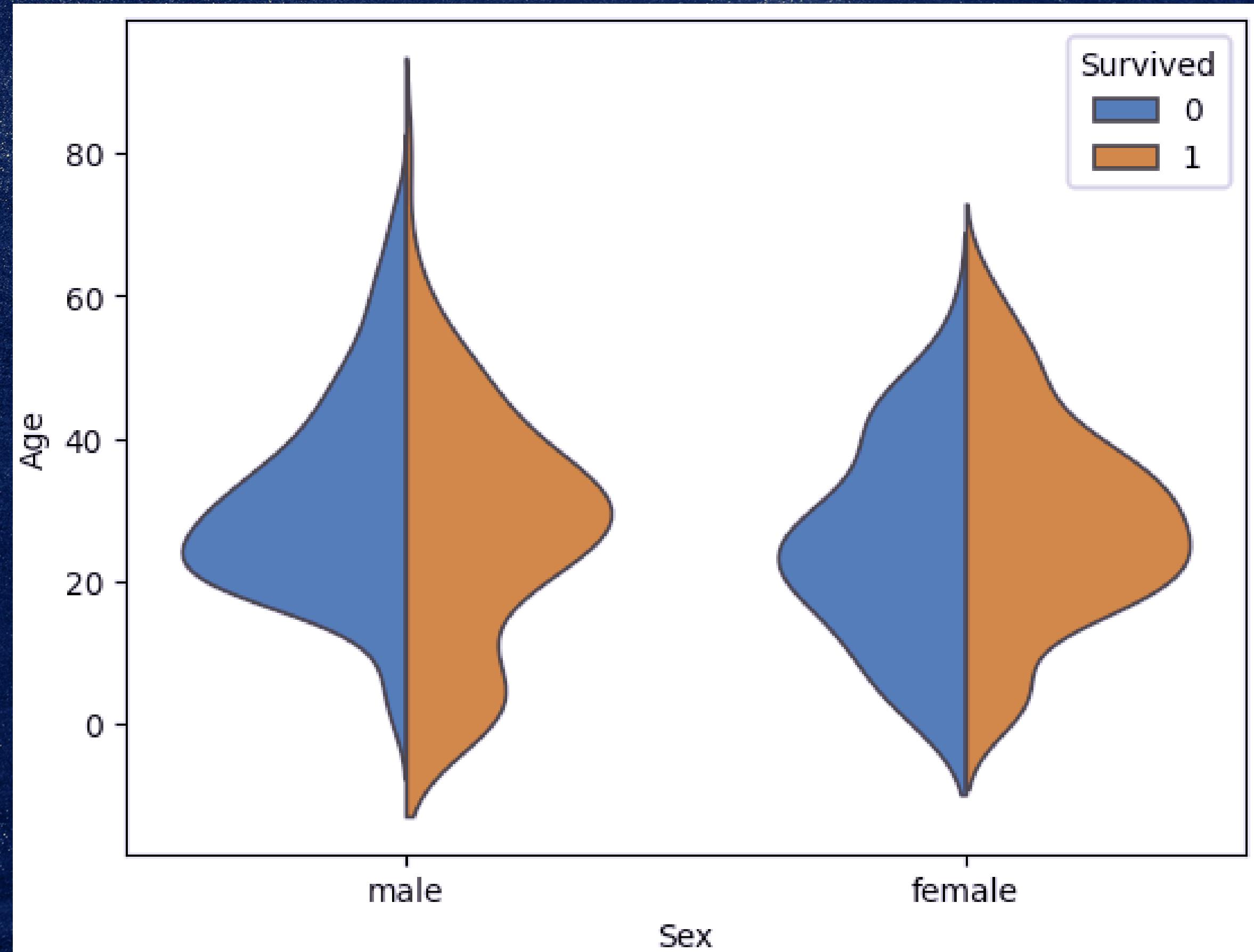
Variabile 3: Età... senza i nulli!



Variabile 3: Età + Genere

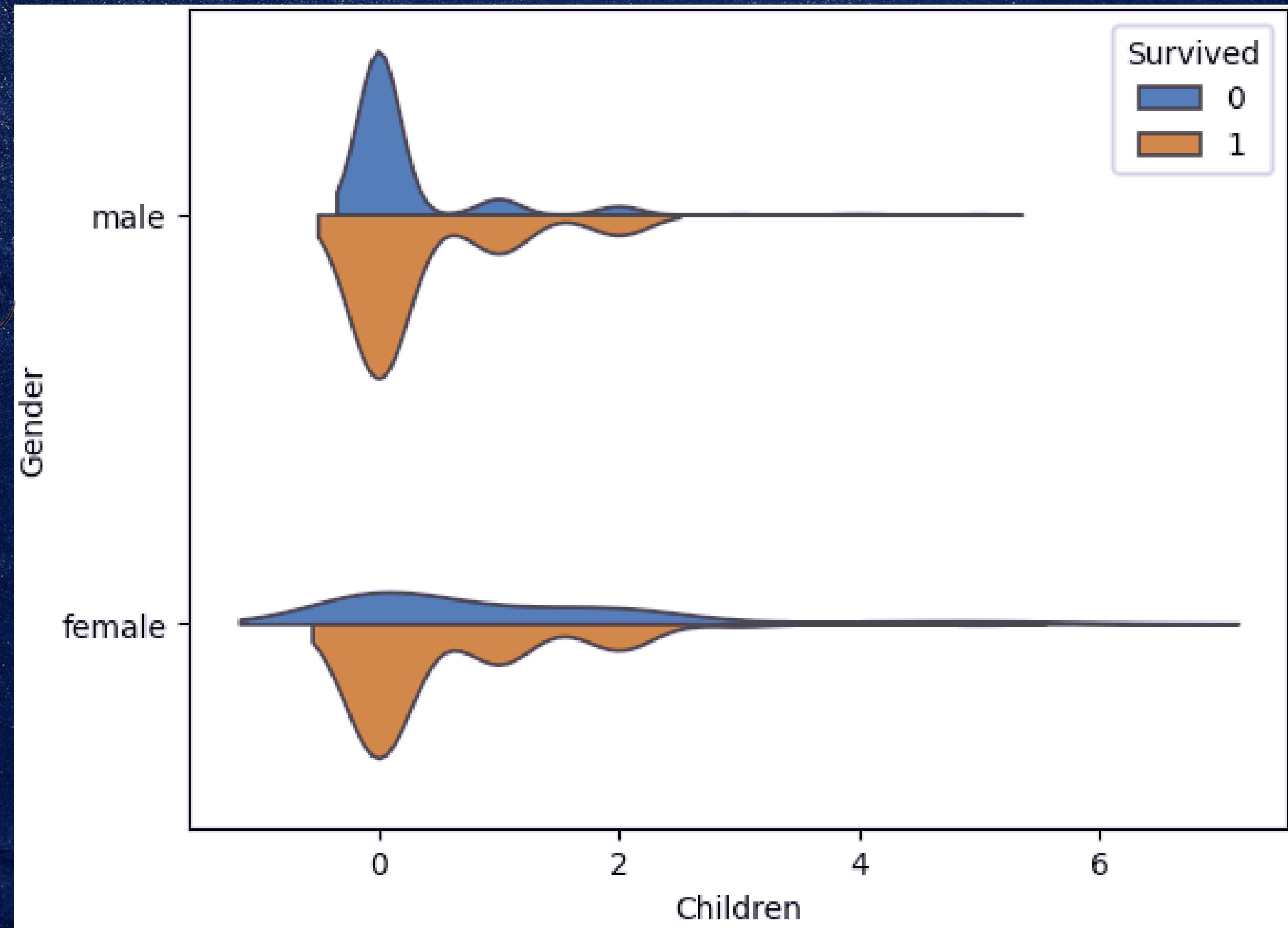
Con dati più consistenti sull'età dei passeggeri, possiamo integrare l'informazione con quelle già ottenute.

Questo grafico ci suggerisce che i sopravvissuti si trovano intorno ai 20/25 anni d'età su entrambi i generi. Più giovani quindi più prestanti?



Variabile 3: Età + Figli

L'ultima domanda che ci siamo posti è: la presenza di bambini ha in qualche modo incentivato la sopravvivenza? Apparentemente più per gli uomini che per le donne, che in questo grafico vedono il tasso di sopravvivenza crollare a picco.





**GRAZIE
PER LA VISIONE**

