ANALISI ESPLORATIVA CON PYTHON

DATASET

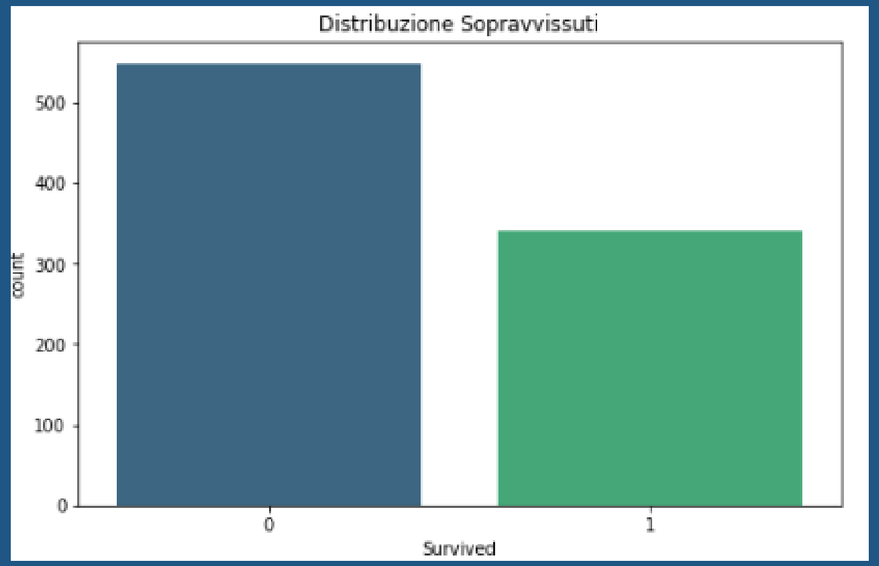
- *Fonte:* https://www.kaggle.com/competitions/titanic
- *Contenuti principali*: Età, genere, classe, biglietto, prezzo, cabina, porto di imbarco, Informazione su sopravvivenza (Survived).
- *Obiettivo*: Analizzare i dati dei passeggeri del Titanic per capire quali fattori hanno influenzato la sopravvivenza.

DISTRIBUZIONE SOPRAVVISSUTI

Obiettivo: comprendere la proporzione generale di sopravvissuti e non sopravvissuti nel dataset. Questo fornisce un primo sguardo al bilanciamento delle classi e al problema.

Risultato: il grafico mostra quante persone sono sopravvissute (valore 1) e quante no (valore 0).

```
plt.figure(figsize=(8,5))
sns.countplot(data=titanic_df, x='Survived',
                palette='viridis')
plt.title('Distribuzione Sopravvissuti')
plt.show()
```

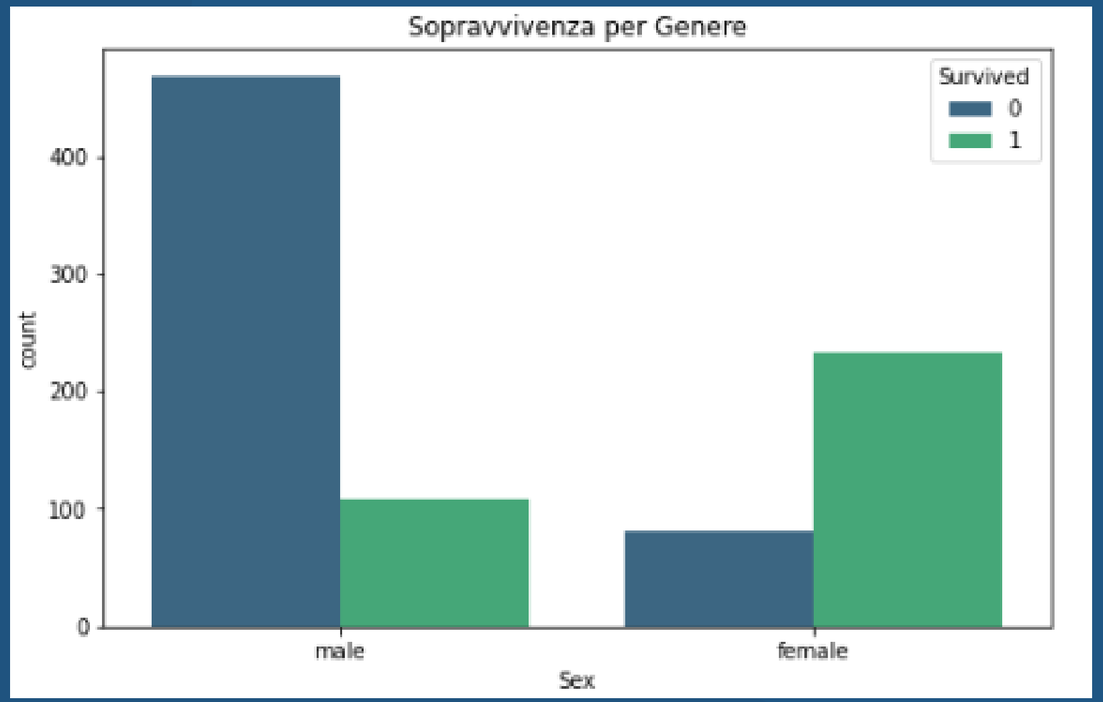


SOPRAVVIVENZA PER SESSO

Obiettivo: valutare l'impatto del genere sulla sopravvivenza.

Risultato: il confronto tra generi mostra che una percentuale significativamente maggiore di donne è sopravvissuta rispetto agli uomini.

```
plt.figure(figsize=(8,5))
sns.countplot(x='Sex', hue='Survived', data=titanic_df,
              palette='viridis')
plt.title('Sopravvivenza per Genere')
plt.show()
```

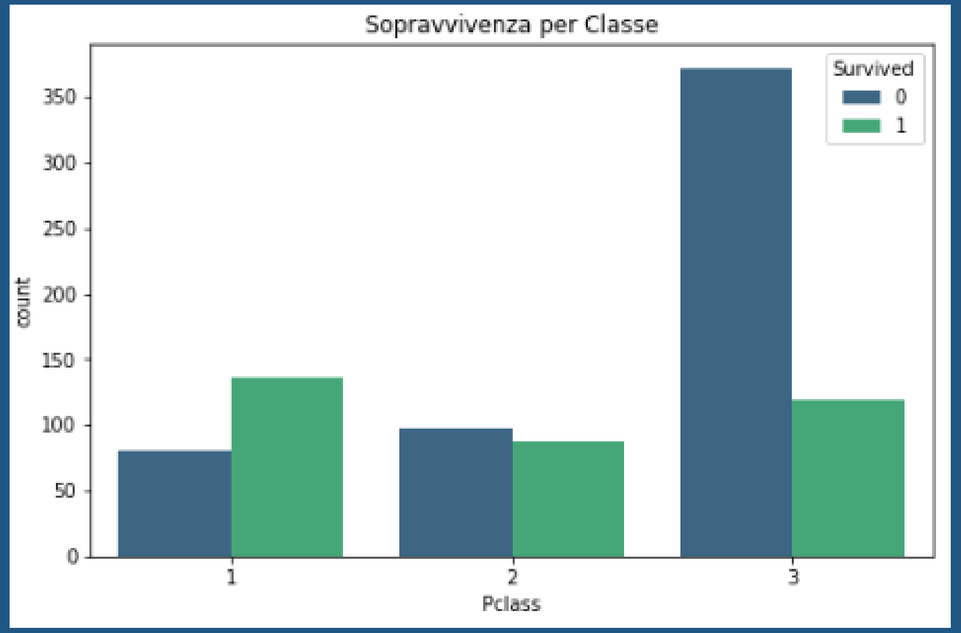


SOPRAVVIVENZA PER CLASSE

Obiettivo: esaminare la correlazione tra classe di viaggio e probabilità di sopravvivenza.

Risultato: analizzando le classi di viaggio, si nota che i passeggeri della prima classe avevano un tasso di sopravvivenza più alto.

```
plt.figure(figsize=(8,5))
sns.countplot(data=titanic_df, x='Pclass',
              hue='Survived', palette='viridis')
plt.title('Sopravvivenza per Classe')
plt.show()
```

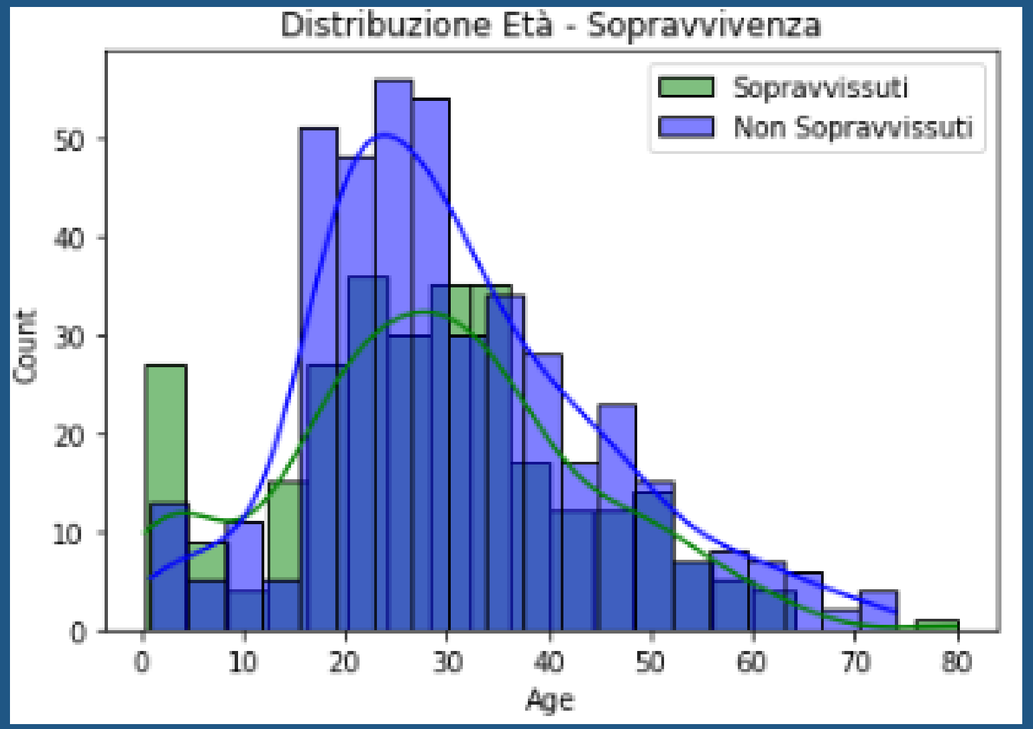


ETA' E SOPRAVVIVENZA

Obiettivo: individuare fasce d'età più vulnerabili o più protette.

Risultato: questo doppio istogramma mette in evidenza che i bambini piccoli avevano maggiori probabilità di sopravvivenza. Gli adulti, in particolare quelli tra i 20 e i 40 anni, mostrano una distribuzione più bilanciata.

```
sns.histplot(titanic_df[titanic_df['Survived']==1]['Age'], bins=20,
             label='Sopravvissuti', kde=True, color='green')
sns.histplot(titanic_df[titanic_df['Survived']==0]['Age'], bins=20,
             label='Non Sopravvissuti', kde=True, color='blue')
plt.legend()
plt.title('Distribuzione Età - Sopravvivenza')
plt.show()
```



HEATMAP CORRELAZIONI

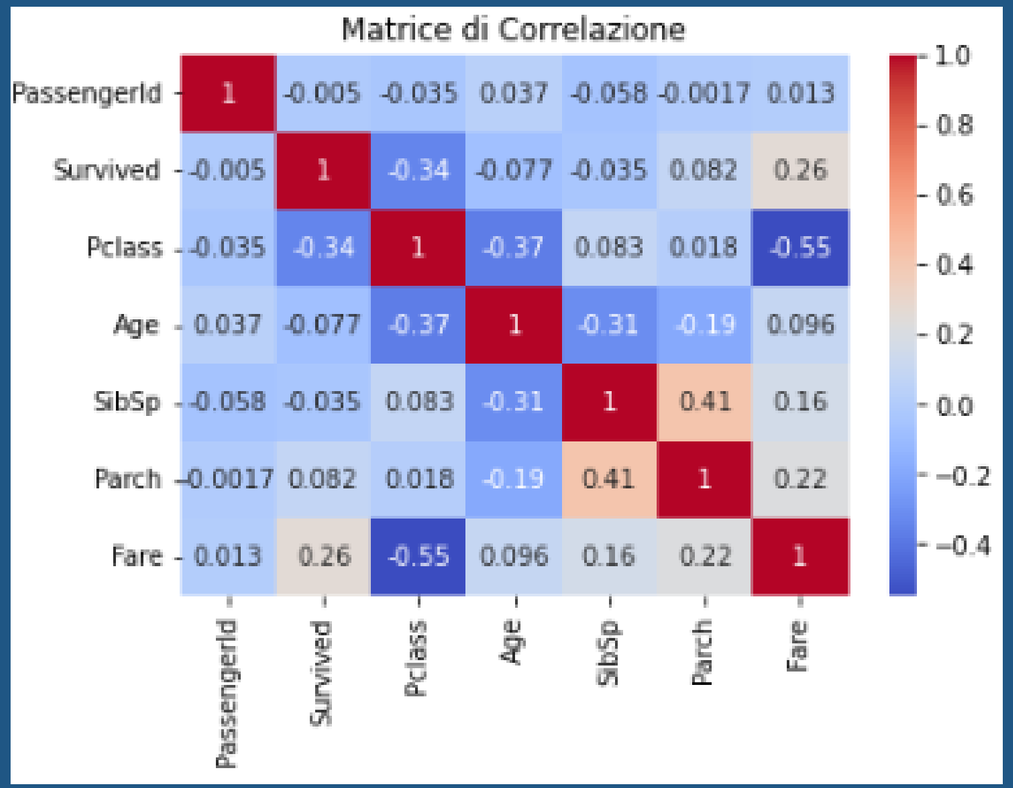
Obiettivo: studiare le correlazioni tra variabili numeriche per individuare possibili relazioni.

Risultato: la matrice di correlazione aiuta a identificare relazioni potenzialmente utili per modelli predittivi, anche se non implicano causalità.

```python
# Selezione colonne numeriche per evitare errori
numeric_data = titanic_df.select_dtypes(include='number')

# Calcolo della matrice di correlazione
corr_matrix = numeric_data.corr()

# Visualizzazione della heatmap
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Matrice di Correlazione')
plt.show()
```



CONCLUSIONE

- Le donne avevano una probabilità di sopravvivenza molto più alta.
- I passeggeri di prima classe avevano maggiori possibilità di sopravvivere.
- I bambini (età < 10) tendevano a sopravvivere più degli adulti.
- I maschi in terza classe erano i più vulnerabili.