Step 1: Load Train dataset and preview

In [ ]:
```
import pandas as pd

df = pd.read_csv("train.csv")

print("Shape:", df.shape)

df.head()
```

```
Shape: (891, 12)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25(
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28:
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92!
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10(
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05(

Observation:

- The dataset has 891 rows and 12 columns.
- The first few rows show columns such as PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked.

Step 2: Dataset info and missing values

In [ ]:
```
df.info()
df.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Observation: The dataset contains the following data types:

- Integer columns: 2 (PassengerId, Survived, Pclass, SibSp, Parch)
- Float columns: 2 (Age,Fare)
- Object (string) columns: 5 (Name, Sex, Ticket, Cabin, Embarked)

Missing values are present in columns such as:

- Age: 177
- Cabin: 687
- Embarked: 2

Step 3: Statistical summary of numerical columns

In [ ]:
```
df.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fa
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.0000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.2042
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.6934
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.0000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.9104
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.4542
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.0000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.3292

Observation:
From the statistical summary:

- The average age of passengers is around 29.7 years, with a minimum of 0.42 and a maximum of 80.
- Fare ranges from 0 to 512.33, with a mean of 32.20.
- The most common passenger class (median Pclass = 3) indicates that a majority of passengers traveled in 3rd class.

Step 4: Frequency counts for categorical columns

In [ ]:
```
print("Sex:\n", df['Sex'].value_counts(), "\n")
print("Embarked:\n", df['Embarked'].value_counts(), "\n")
print("Pclass:\n", df['Pclass'].value_counts(), "\n")
```

```
Sex:
 Sex
male      577
female    314
Name: count, dtype: int64

Embarked:
 Embarked
S    644
C    168
Q     77
Name: count, dtype: int64

Pclass:
 Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

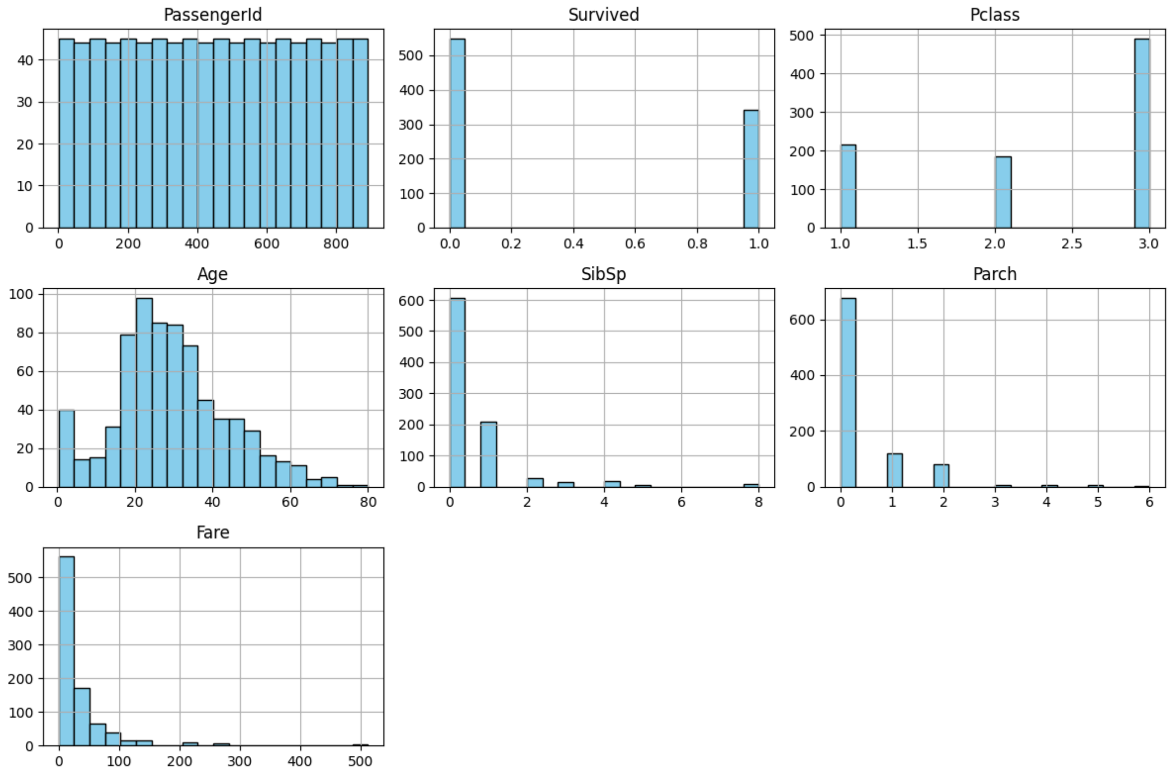
Observation:

- Majority of passengers are male (577) compared to female (314).
- Most passengers embarked from port 'S' (Southampton), followed by 'C' (Cherbourg) and 'Q' (Queenstown).
- Passenger class distribution shows that class 3 has the highest count.

Step 5.1: Histograms for numerical columns

In [ ]:
```
import matplotlib.pyplot as plt
import seaborn as sns

df.hist(figsize=(12, 8), bins=20, color='skyblue', edgecolor='black')
plt.tight_layout()
plt.show()
```



Observation:

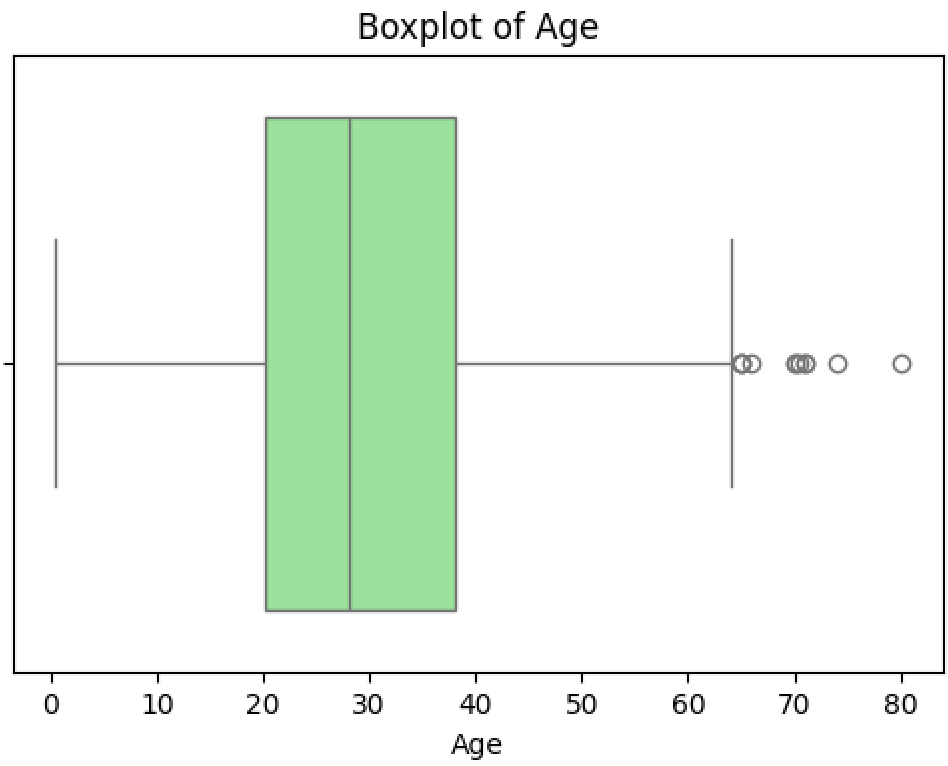
- Age distribution is roughly right-skewed, with most passengers aged between 20–40 years.
- Fare distribution is highly right-skewed, with most fares below 100.
- SibSp and Parch have many passengers with 0 relatives onboard.

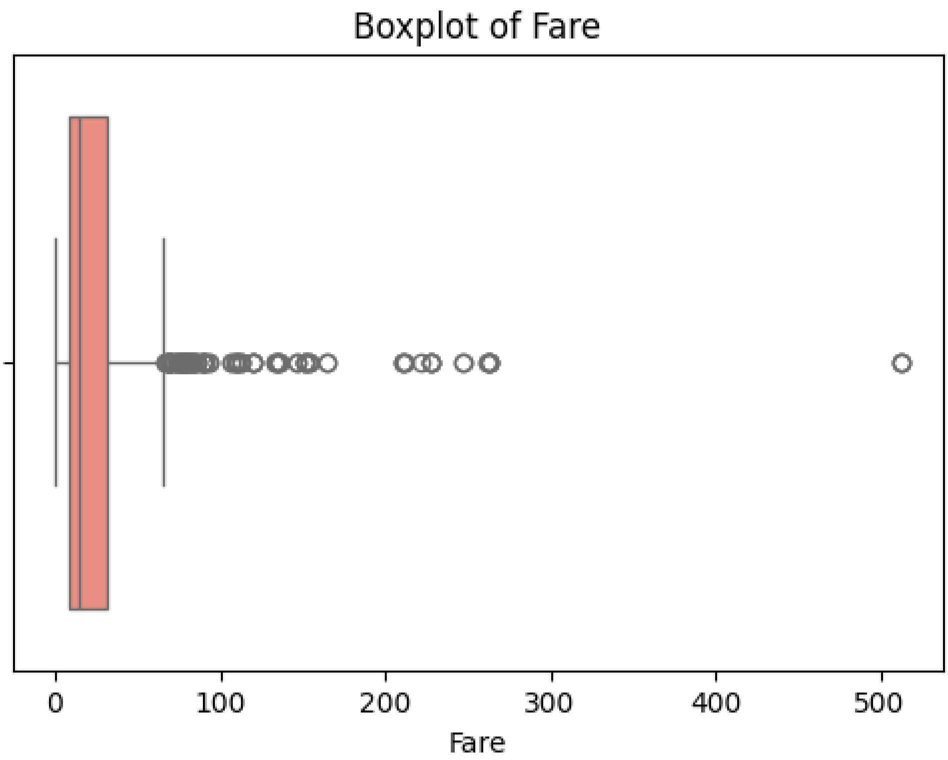
Step 5.2: Boxplots for detecting outliers

In [ ]:
```
plt.figure(figsize=(6,4))
sns.boxplot(x=df['Age'], color='lightgreen')
plt.title('Boxplot of Age')
plt.show()

plt.figure(figsize=(6,4))
sns.boxplot(x=df['Fare'], color='salmon')
```

```
plt.title('Boxplot of Fare')
plt.show()
```





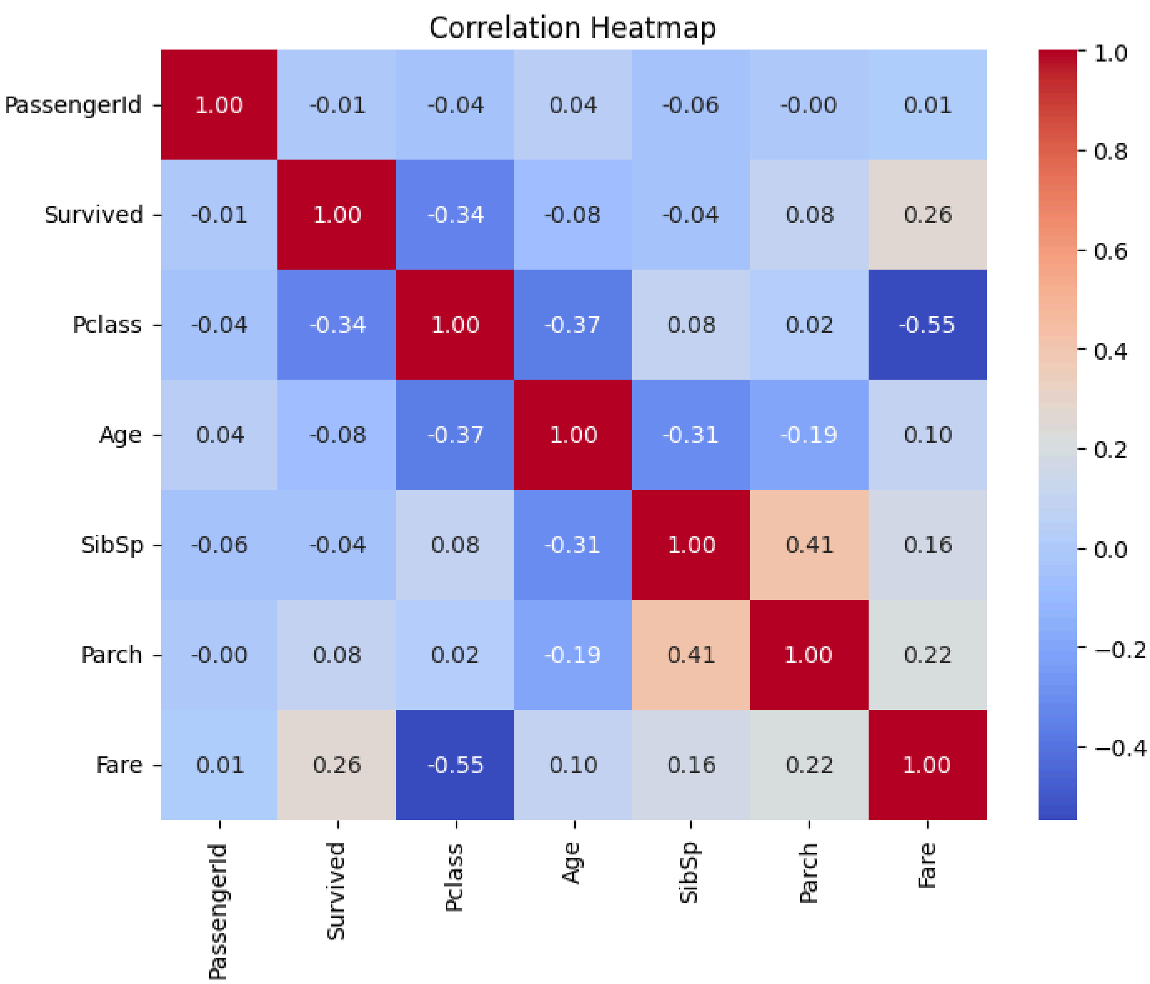
Observation:

- Age column has a few mild outliers above ~65 years.
- Fare column shows many extreme outliers, with some fares above 300–500.

Step 5.3: Correlation Heatmap

In [ ]:
```
corr_matrix = df.corr(numeric_only=True)

plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



Observation:

- Survived is positively correlated with Fare and negatively correlated with Pclass (higher class = higher survival chances).
- Age has very weak correlation with survival.
- SibSp and Parch show a small positive correlation with each other, indicating families often traveled together.

Step 5.4: Pairplot of selected features

In [ ]:
```
selected_cols = ['Survived', 'Pclass', 'Age', 'Fare', 'SibSp', 'Parch']
sns.pairplot(df[selected_cols], hue='Survived', palette='husl')
plt.show()
```

Observation:

- Passengers in Pclass 1 generally paid higher fares and had better survival rates.
- Younger passengers (children) had slightly higher chances of survival.
- Many passengers with 0 SibSp or Parch did not survive, but families with small group sizes had better outcomes.

Step 6: Summary of EDA Findings

Summary of Insights:

1. **Survival Rate:** Only ~38% of passengers survived.
2. **Gender Impact:** Females had a much higher survival rate than males.
3. **Passenger Class:** Pclass 1 passengers had the highest survival rate; Pclass 3 had the lowest.
4. **Age Distribution:** Most passengers were between 20–40 years; children had higher chances of survival.
5. **Fare:** Higher fares were associated with better survival chances (linked to higher class).
6. **Family Size:** Small family sizes (SibSp and Parch values of 1–2) had better survival rates than those traveling alone or with very large families.

7. **Embarkation Port:** Most passengers boarded at Southampton, but Cherbourg passengers had higher survival rates.
8. **Outliers:** Fare column contained extreme outliers; Age had a few mild outliers.
9. **Correlations:** Fare and Pclass had the strongest correlation with survival. Age had very weak correlation with survival.