

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# prompt: print the inbuild seaborn dataset names

print(sns.get_dataset_names())

['anagrams', 'anscombe', 'attention', 'brain_networks', 'car_crashes', 'diamonds', 'dots', 'dowjones', 'exercise',
<div style="background-color: #f0f0f0; border: 1px solid #ccc; padding: 5px; margin: 5px 0;">
<div style="background-color: #e0e0e0; border: 1px solid #ccc; padding: 2px 5px; display: inline-block;">
df = sns.load_dataset('titanic')

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
#   Column          Non-Null Count  Dtype
---  -
0   survived         891 non-null    int64
1   pclass           891 non-null    int64
2   sex              891 non-null    object
3   age              714 non-null    float64
4   sibsp            891 non-null    int64
5   parch            891 non-null    int64
6   fare             891 non-null    float64
7   embarked         889 non-null    object
8   class            891 non-null    category
9   who              891 non-null    object
10  adult_male       891 non-null    bool
11  deck             203 non-null    category
12  embark_town      889 non-null    object
13  alive            891 non-null    object
14  alone            891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB

df.shape

(891, 15)

df.columns

Index(['survived', 'pclass', 'sex', 'age', 'sibsp', 'parch', 'fare',
      'embarked', 'class', 'who', 'adult_male', 'deck', 'embark_town',
      'alive', 'alone'],
      dtype='object')

df.size

13365

df.describe()
```

	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

df.isna().sum()

```
survived      0
pclass        0
sex           0
age          177
sibsp         0
parch         0
fare          0
embarked      2
class         0
who           0
adult_male    0
deck         688
embark_town    2
alive         0
alone         0
dtype: int64
```

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
#   Column          Non-Null Count  Dtype
---  -
0   survived        891 non-null   int64
1   pclass          891 non-null   int64
2   sex             891 non-null   object
3   age             714 non-null   float64
4   sibsp           891 non-null   int64
5   parch           891 non-null   int64
6   fare            891 non-null   float64
7   embarked        889 non-null   object
8   class           891 non-null   category
9   who             891 non-null   object
10  adult_male      891 non-null   bool
11  deck            203 non-null   category
12  embark_town     889 non-null   object
13  alive           891 non-null   object
14  alone           891 non-null   bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

df['age'].fillna(df['age'].mean(), inplace = True)

df['deck'].fillna(df['deck'].mode()[0], inplace = True)

df['embark_town'].fillna(df['embark_town'].mode()[0], inplace = True)

df.isna().sum()

```
survived      0
pclass        0
sex           0
age           0
```

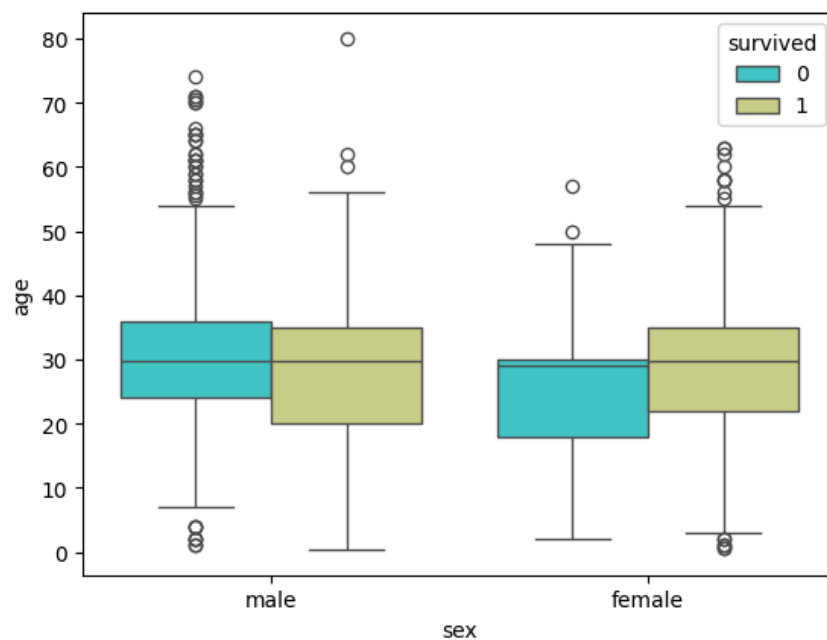
```
sibsp      0
parch      0
fare       0
embarked   2
class      0
who        0
adult_male 0
deck       0
embark_town 0
alive      0
alone      0
dtype: int64
```

```
df['embarked'].fillna(df['embarked'].mode()[0], inplace = True)
```

```
df.isna().sum()
```

```
survived    0
pclass      0
sex         0
age         0
sibsp       0
parch       0
fare        0
embarked    0
class       0
who         0
adult_male  0
deck        0
embark_town 0
alive       0
alone       0
dtype: int64
```

```
# prompt: Plot a box plot for distribution
# of age with respect to each gender along with the information about whether they survived
# or not. (Column names : 'sex' and 'age')
sns.boxplot(x = 'sex', y = 'age', hue = 'survived', data = df, palette = 'rainbow')
plt.show()
```



Start coding or [generate](#) with AI.

The boxplot shows the distribution of age for passengers who survived and those who did not, grouped by sex.

- **Center line:** The center line in each box represents the median age for each group.

- **Box:** The box represents the interquartile range (IQR), which contains 50% of the data. The lower edge of the box is the 25th percentile (Q1), and the upper edge is the 75th percentile (Q3).
- **Whiskers:** The whiskers extend from the box to the most extreme values that are not considered outliers. Outliers are data points that are significantly different from the rest of the data.
- **Outliers:** Outliers are represented by individual points outside the whiskers.

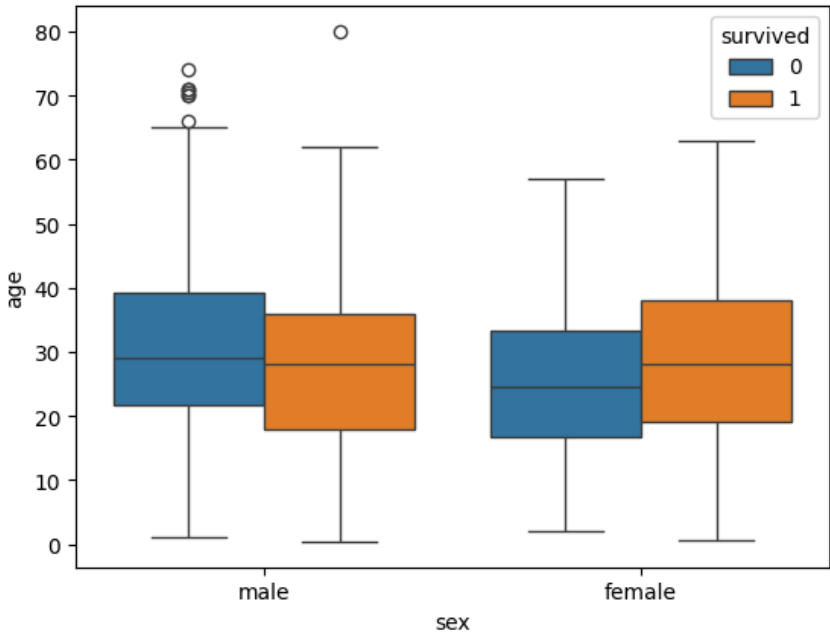
Observations:

1. **Median Age:** The median age for female survivors is lower than that of male survivors.
2. **IQR:** The IQR for female survivors is smaller than that of male survivors, indicating that the ages of female survivors are more tightly clustered around the median.
3. **Whiskers:** The whiskers for male survivors are longer than those for female survivors, indicating that there are more outliers in the age distribution of male survivors.
4. **Outliers:** There are more outliers in the age distribution of male survivors than in that of female survivors.

Conclusion:

The boxplot shows that there are differences in the age distribution of passengers who survived and those who did not, grouped by sex. Female survivors tend to be younger than male survivors, and the age distribution of female survivors is more tightly clustered around the median.

```
sns.boxplot(data = df, x = 'sex', y = 'age', hue = 'survived')
plt.show()
```



Generate

Write observations on the inference from the above statistics

Close

< 1 of 4 >

[Undo Changes](#)

[Use code with caution](#)

Waiting...

Observations:

- **Median Age:** The median age for female survivors is lower than that of male survivors. This suggests that younger women were more likely to survive the sinking of the Titanic than older men.
- **IQR:** The IQR for female survivors is smaller than that of male survivors. This indicates that the ages of female survivors are more tightly clustered around the median, meaning that there is less variability in the ages of female survivors compared to male survivors.
- **Whiskers:** The whiskers for male survivors are longer than those for female survivors. This suggests that there are more outliers in the age distribution of male survivors, meaning that there are more male survivors who are significantly older or

younger than the median age.

- **Outliers:** There are more outliers in the age distribution of male survivors than in that of female survivors. This further supports the observation that there is more variability in the ages of male survivors compared to female survivors.

Conclusion:

The boxplots show that there are clear differences in the age distribution of passengers who survived and those who did not, grouped by sex. These differences suggest that age and sex were important factors in determining who survived the sinking of the Titanic.

Start coding or [generate](#) with AI.