omkargaikwad23 updates History

1 contributor

705 lines (705 sloc) | 312 KB

Assignment 09

Omkar Gaikwad

31126

Data Visualization II

1. Use the inbuilt dataset 'titanic' as used in the above problem. 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')
2. Write observations on the inference from the above statistics.

Importing Libraries

In [1]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
sns.get_dataset_names()
```

Out[2]:

```
['anagrams',
 'anscombe',
 'attention',
 'brain_networks',
 'car_crashes',
 'diamonds',
 'dots',
 'exercise',
 'flights',
 'fmri',
 'gammas',
 'geyser',
 'iris',
 'mpg',
 'penguins',
 'planets',
 'taxis',
 'tips',
 'titanic']
```

Importing Dataset

In [3]:

```
df = sns.load_dataset('titanic')
```

In [4]:

```
df.head()
```

Out[4]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	d
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	N
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	N
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	N

Cleaning Null Values

In [5]:

```
df.isnull().sum()
```

Out[5]:

```
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
```

In [6]:

```
df['age'] = df['age'].fillna(df['age'].mean())
```

In [7]:

```
deck_mode = df['deck'].mode()[0]
print("Mode of deck is:", deck_mode)
df['deck'] = df['deck'].fillna(deck_mode)
```

```
Mode of deck is: C
```

In [8]:

```
embarked_mode=df['embarked'].mode()[0]
print("Embarked mode: ", embarked_mode)
df['embarked'] = df['embarked'].fillna(embarked_mode)
```

```
Embarked mode:  S
```

In [9]:

```
df = df.dropna()
df = df.reset_index()
df = df.drop('index',axis=1)
```

In [10]:

```
df.isnull().sum()
```

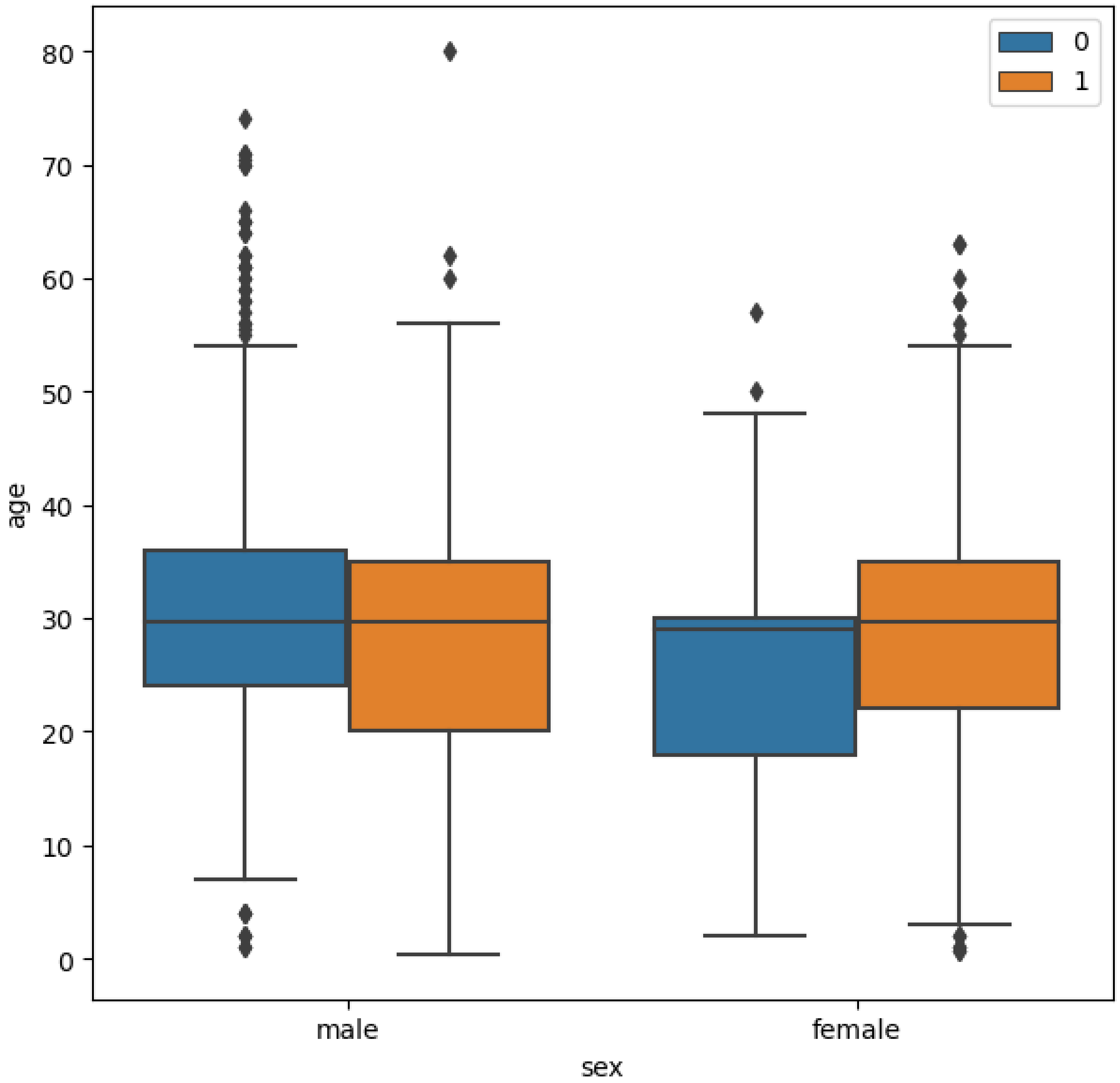
Out[10]:

```
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
```

Plotting Graphs

In [11]:

```
plt.figure(figsize=(7,7),dpi=100)
sns.boxplot(x="sex", y="age", data=df,hue='survived')
plt.legend()
plt.show()
```



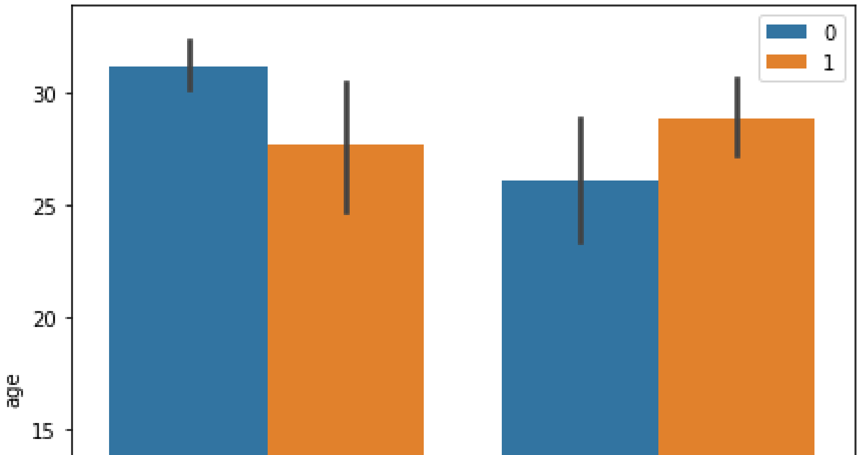
Observations: 1) There are more number of male having age above 65 than female

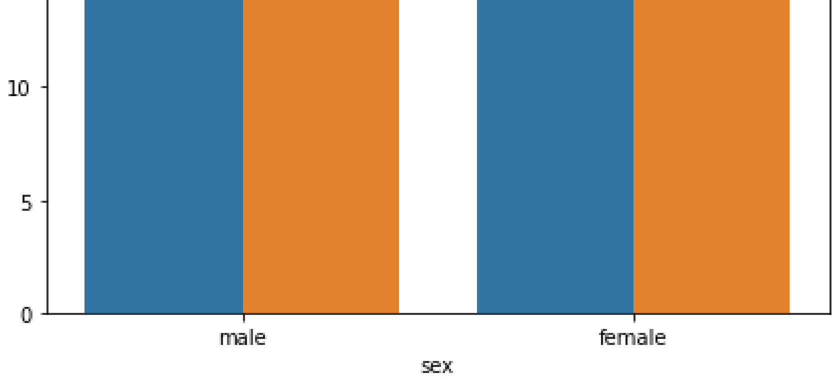
2) Survival rate of female is more than male

3) More older males could survive than older female

In [13]:

```
plt.figure(figsize=(7,7))
ax = sns.barplot(data=df, x="sex",y="age",hue='survived')
plt.legend()
plt.show()
```

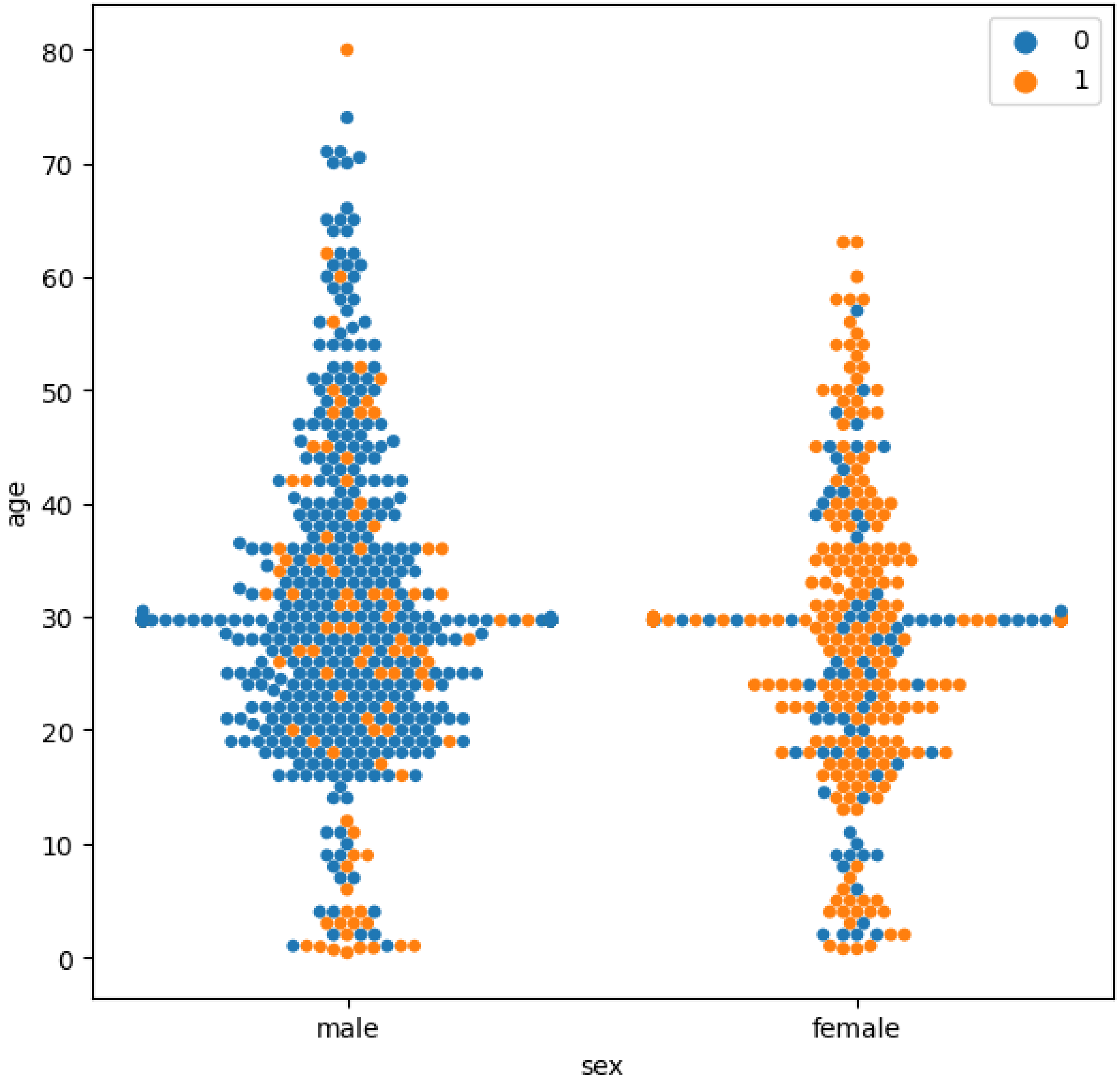




In [14]:

```
# Draw a categorical scatterplot with non-overlapping points.
plt.figure(figsize=(7,7),dpi=100)
ax = sns.swarmplot(x="sex",y="age",data=df,hue='survived')
plt.legend()
plt.show()
```

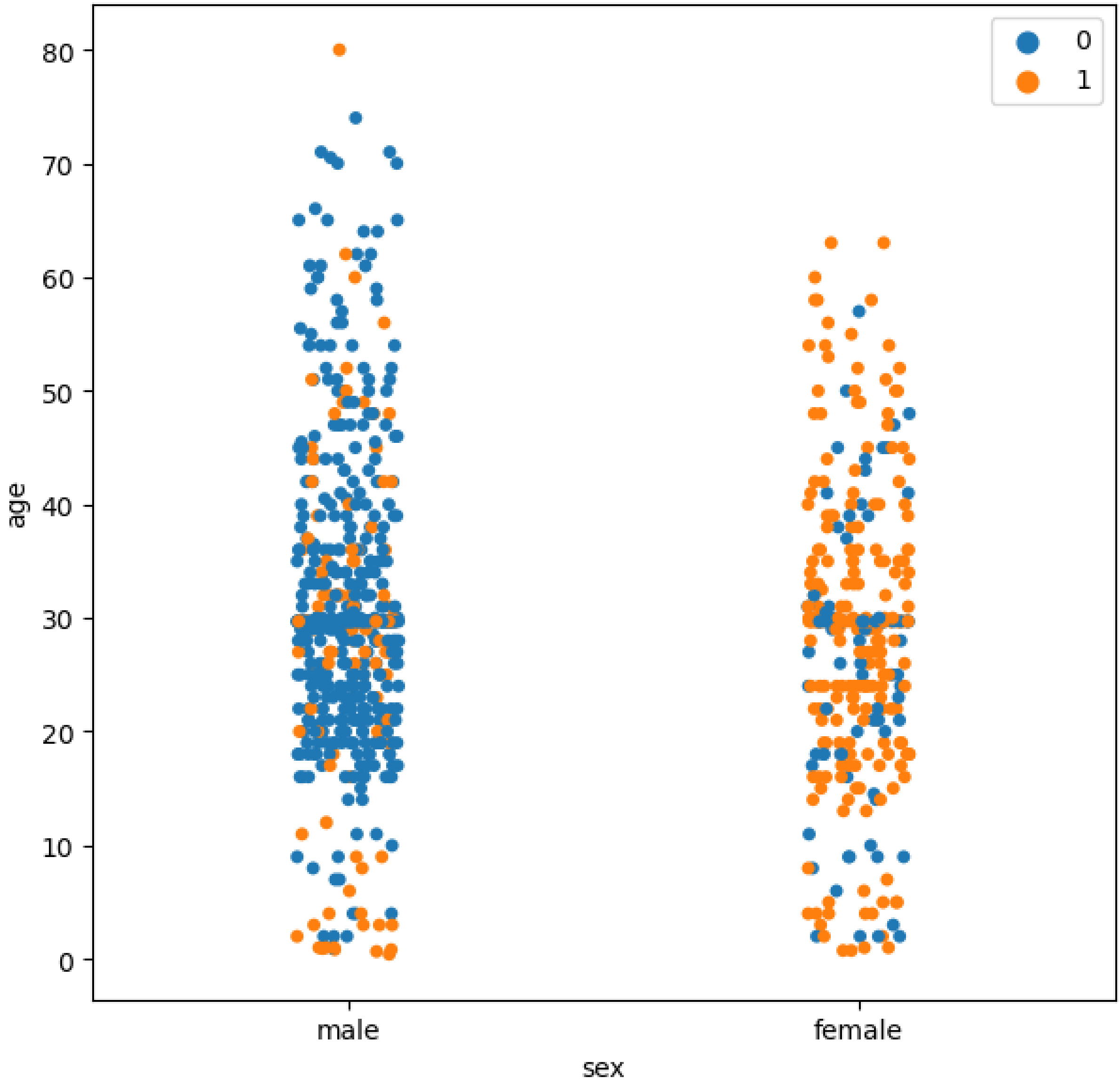
```
C:\Users\omkar madhav gaikwad\anaconda3\lib\site-packages\seaborn\categorica
l.py:1296: UserWarning: 19.1% of the points cannot be placed; you may want to
decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
C:\Users\omkar madhav gaikwad\anaconda3\lib\site-packages\seaborn\categorica
l.py:1296: UserWarning: 11.2% of the points cannot be placed; you may want to
decrease the size of the markers or use stripplot.
  warnings.warn(msg, UserWarning)
```



1) Survival rate of female is more than male

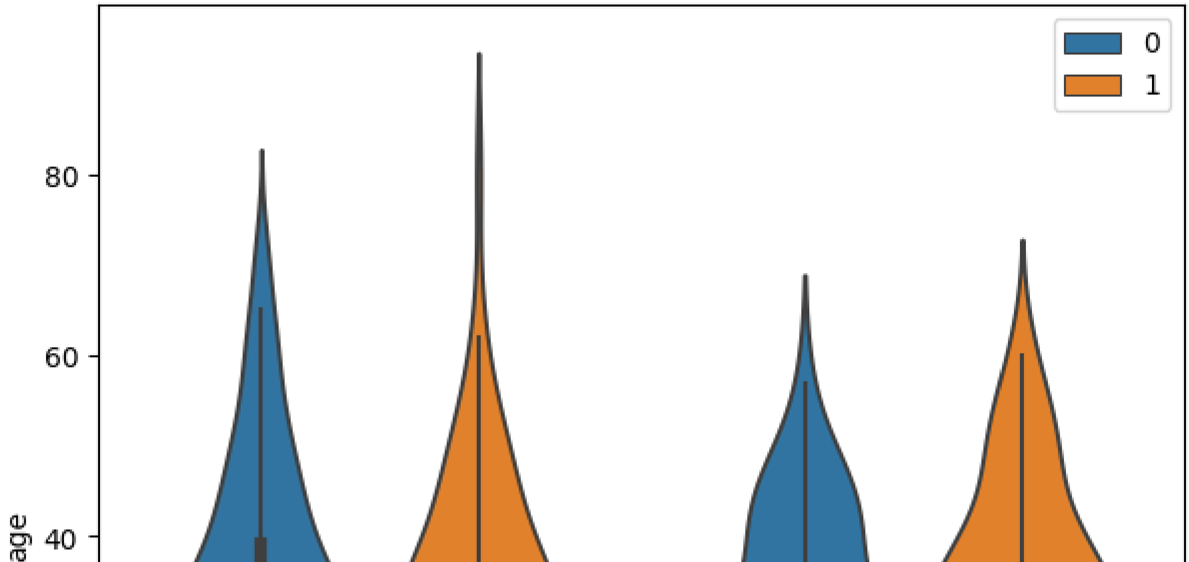
In [15]:

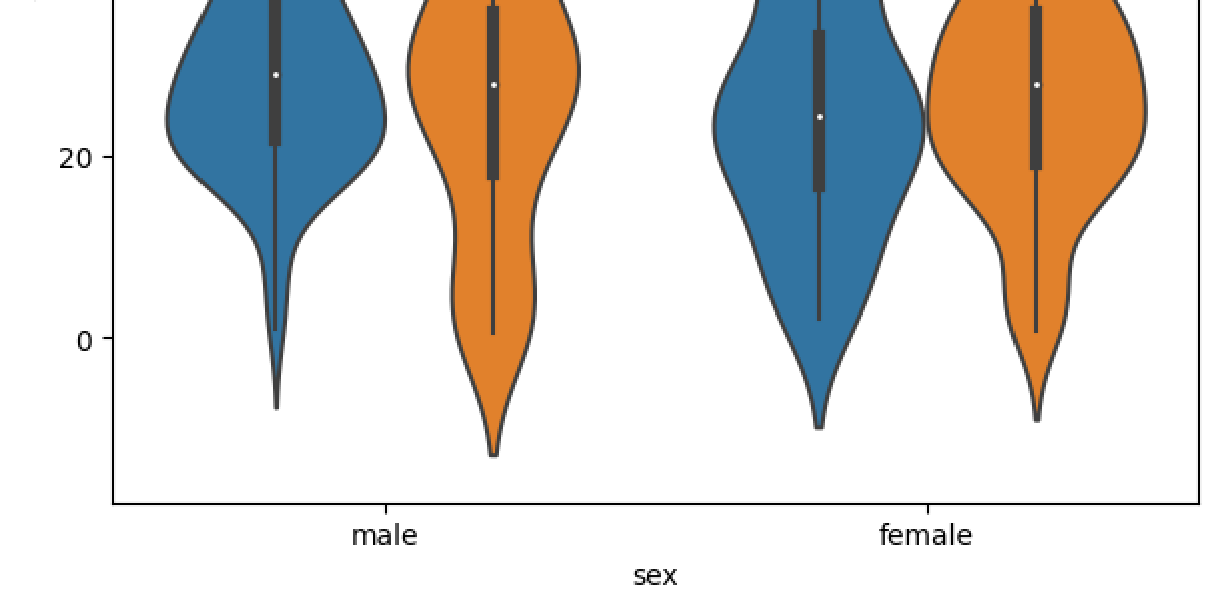
```
# similar to scatter plot differentiate category
plt.figure(figsize=(7,7),dpi=100)
sns.stripplot(x="sex",y="age",data=df,hue='survived')
plt.legend()
plt.show()
```



In [30]:

```
# violin plot features a kernel density estimation of the underlying distribu
plt.figure(figsize=(7,7),dpi=100)
ax = sns.violinplot(x="sex",y="age",data=df,hue='survived')
plt.legend()
plt.show()
```





1) Wider section of violine plot represent higher probability that members of the population will take on the given value

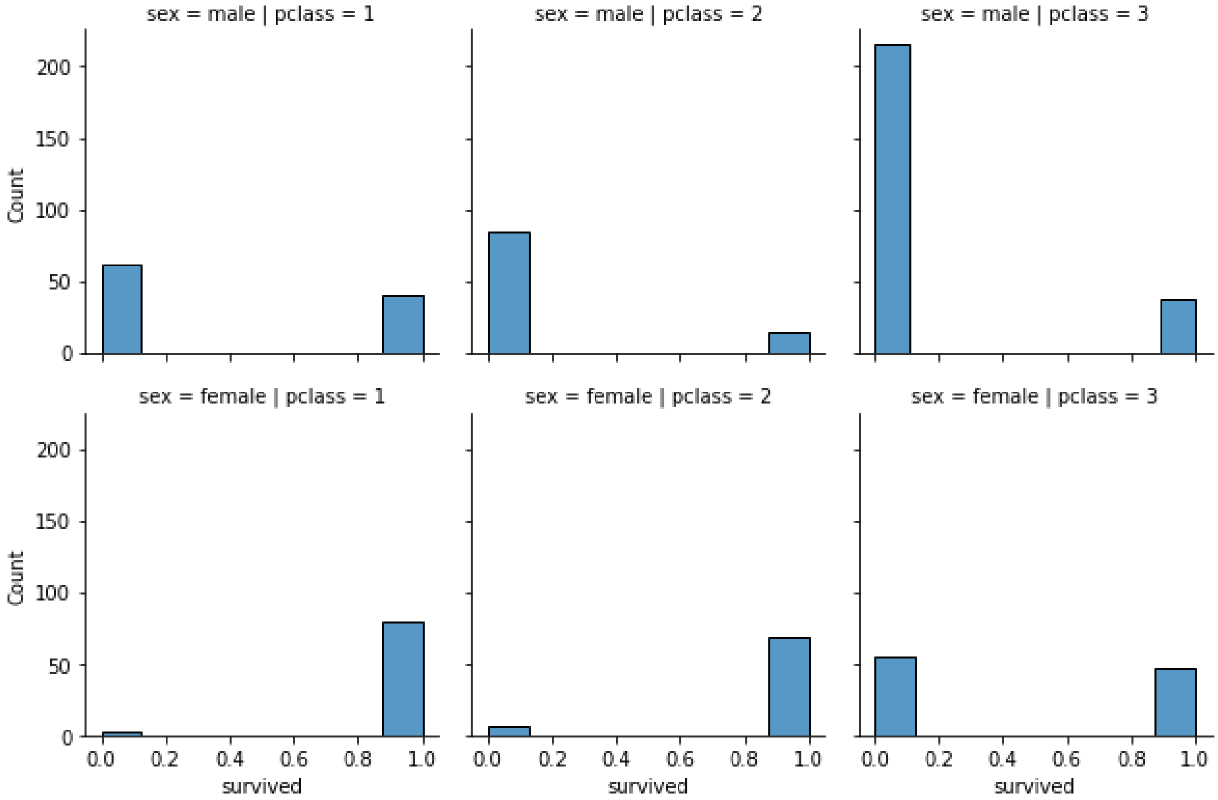
- the white dot represents the median
- the thick gray bar in the center represents the interquartile range
- the thin gray line represents the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the interquartile range.

In [29]:

```
# FacetGrid class helps in visualizing distribution of one variable as well a
# variables separately within subsets of your dataset using multiple panels.

plt.figure(figsize=(7,7),dpi=100)
ax = sns.FacetGrid(df,col="pclass",  row="sex")
ax.map(sns.histplot,"survived")
plt.show()
```

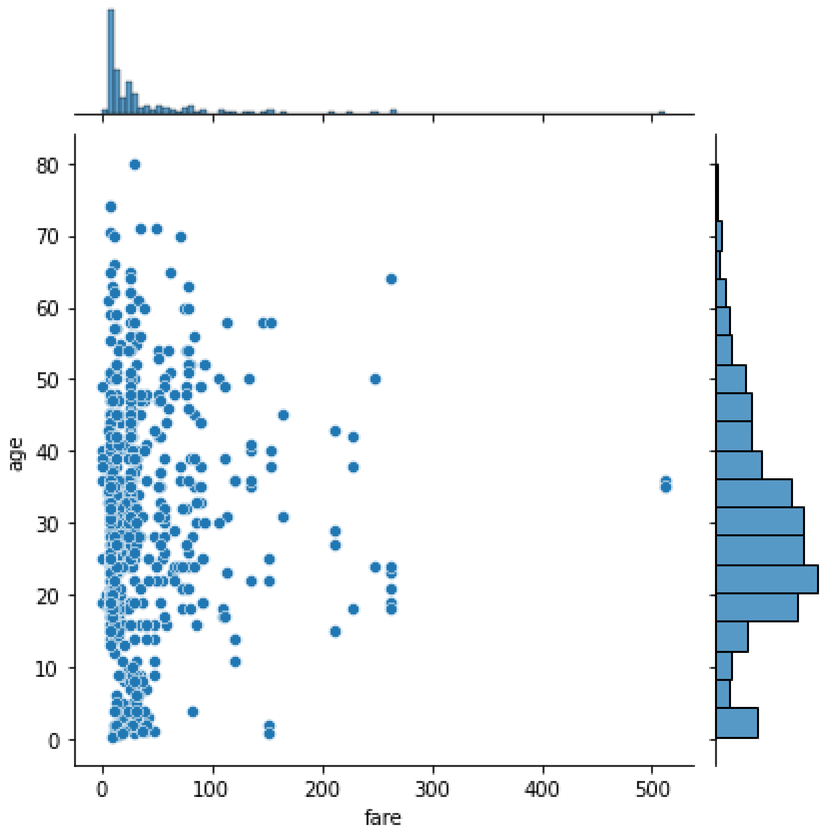
```
<Figure size 700x700 with 0 Axes>
```



In [33]:

```
sns.jointplot(x="fare", y="age", data=df)
```

```
Out[33]: <seaborn.axisgrid.JointGrid at 0x7f53efd179d0>
```



from above representation: 1) More number of people having fare below 100

2) children generally have lower fare rate