titanic_9

May 11, 2022

0.0.1 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.

[2]:
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns

df = sns.load_dataset('titanic')

df.head(10)
```

[2]:

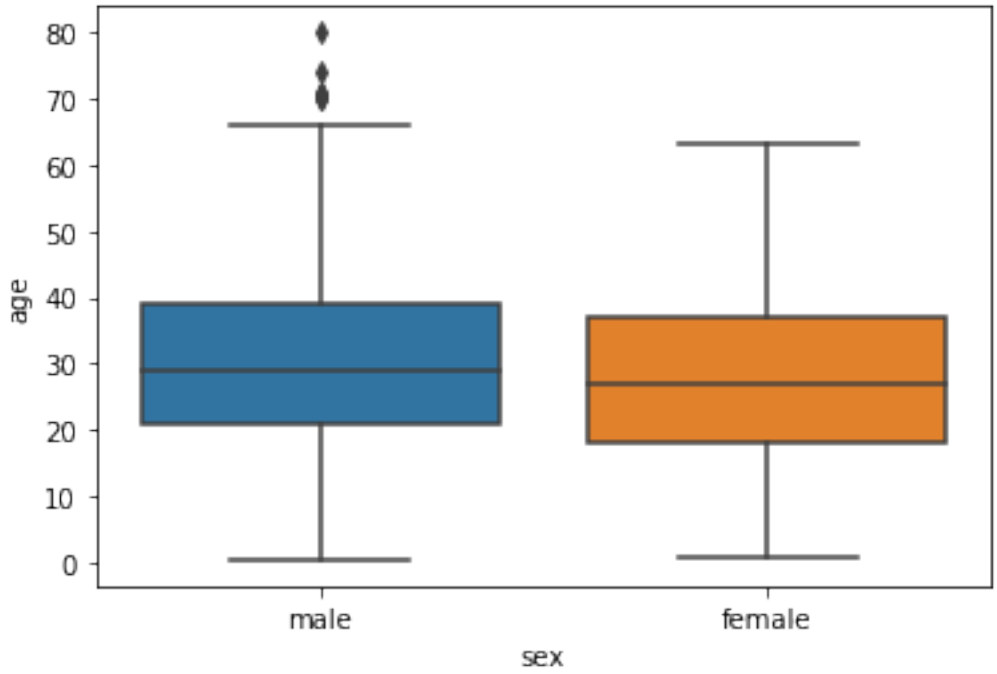
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	
5	0	3	male	NaN	0	0	8.4583	Q	Third	
6	0	1	male	54.0	0	0	51.8625	S	First	
7	0	3	male	2.0	3	1	21.0750	S	Third	
8	1	3	female	27.0	0	2	11.1333	S	Third	
9	1	2	female	14.0	1	0	30.0708	C	Second	

	who	adult_male	deck	embark_town	alive	alone
0	man	True	NaN	Southampton	no	False
1	woman	False	C	Cherbourg	yes	False
2	woman	False	NaN	Southampton	yes	True
3	woman	False	C	Southampton	yes	False
4	man	True	NaN	Southampton	no	True
5	man	True	NaN	Queenstown	no	True
6	man	True	E	Southampton	no	True
7	child	False	NaN	Southampton	no	False
8	woman	False	NaN	Southampton	yes	False

```
    9  child       False  NaN    Cherbourg   yes  False
```

[3]:
```
sns.boxplot(x='sex', y='age', data=df)
```

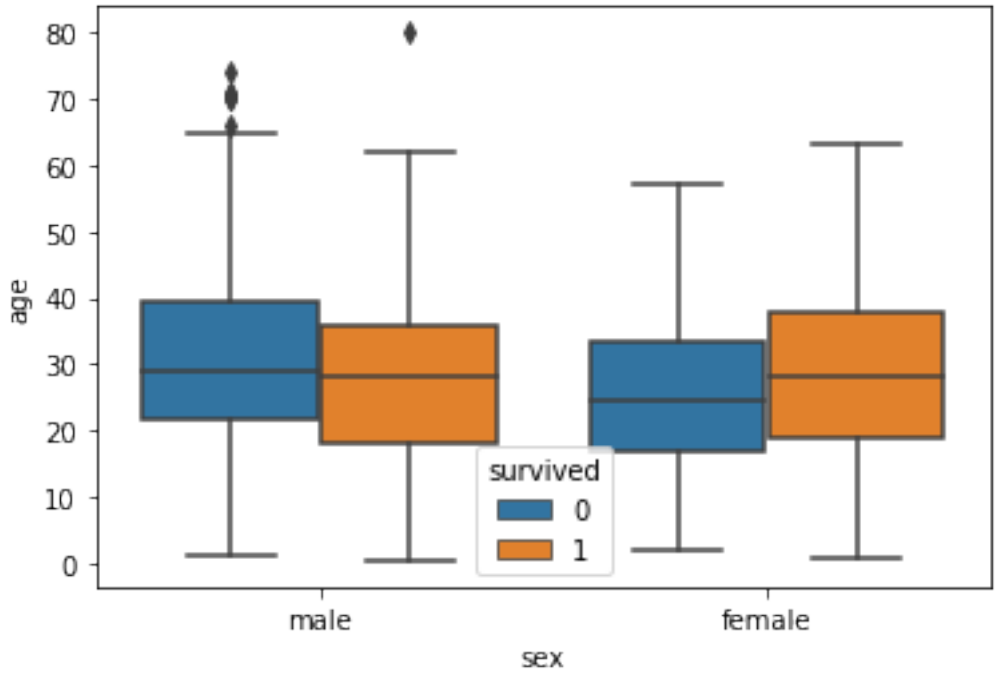
[3]:
```
<AxesSubplot:xlabel='sex', ylabel='age'>
```



Let's try to understand the box plot for female. The first quartile starts at around 5 and ends at 22 which means that 25% of the passengers are aged between 5 and 25.The second quartile starts at around 23 and ends at around 32 which means that 25% of the passengers are aged between 23 and 32. Similarly, the third quartile starts and ends between 34 and 42, hence 25% passengers are aged within this range and finall the fourth or last quartile starts at 43 and ends around 65. If there are any outliers or the passengers that do not belong to any of the quartiles, they are called outliers and are represented by dots on the box plot.

[6]:
```
sns.boxplot(x='sex', y='age', data=df, hue="survived")
```

[6]:
```
<AxesSubplot:xlabel='sex', ylabel='age'>
```


80
70
60
50
40
30
20
10
0
age
survived
0
1
male
female
sex