lab-09

April 13, 2024

Data Visualization

Problem Statement

Use the inbuilt dataset 'titanic' as used in the above problem. 1. 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.

[1]:
```
#imports
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

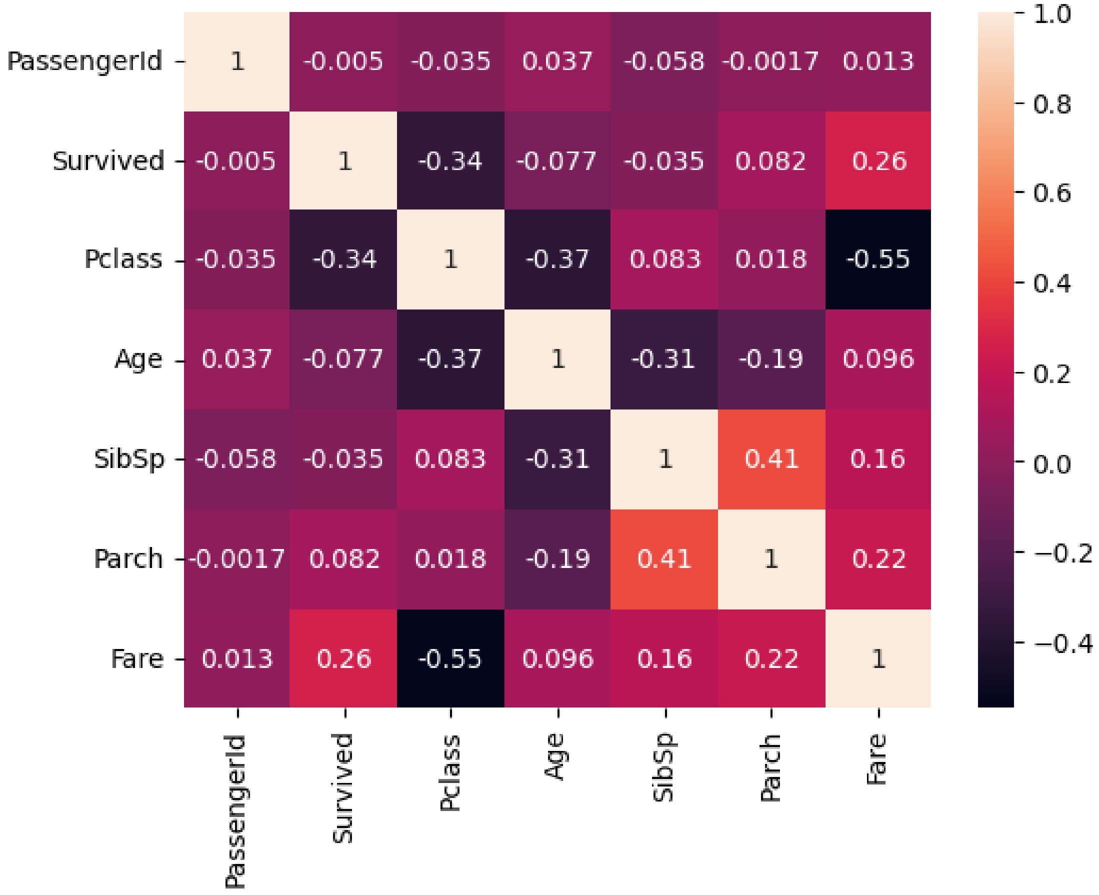
[3]:
```
data = pd.read_csv('train.csv')
data.sample(5)
```

[3]:
```
     PassengerId  Survived  Pclass  \
194          195         1       1
299          300         1       1
96            97         0       1
554          555         1       3
864          865         0       2

                                                Name     Sex   Age  SibSp  \
194        Brown, Mrs. James Joseph (Margaret Tobin)  female  44.0      0
299  Baxter, Mrs. James (Helene DeLaudeniere Chaput)  female  50.0      0
96                          Goldschmidt, Mr. George B    male  71.0      0
554                                Ohman, Miss. Velin  female  22.0      0
864                            Gill, Mr. John William    male  24.0      0

     Parch    Ticket      Fare    Cabin Embarked
194      0  PC 17610   27.7208       B4        C
299      1  PC 17558  247.5208  B58 B60        C
96       0  PC 17754   34.6542       A5        C
554      0    347085    7.7750      NaN        S
864      0    233866   13.0000      NaN        S
```

[4]:
```
data.isna().sum()
```

```
[4]: 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
```

```
[6]: #Age has a lot of null values and is one of the attributes we need to use.
     sns.heatmap(data.corr(), annot = True);
```



From the above corealtion matrix we can see that the attribute ‘Age’ is not highly dependant on any other attribute This means we can randomly fill in the missing data for ‘Age’ within the valid distribution.

[7]:

```
age_null_mask = data['Age'].isnull()

age_mean = data['Age'].mean()
age_std = data['Age'].std()

# generate random ages based on the age distribution of the dataset
age_random = np.random.normal(loc=age_mean, scale=age_std, size=age_null_mask.
 ↪sum())

# fill in missing age values with random ages
data.loc[age_null_mask, 'Age'] = age_random
```

[17]:

```
# 177 normal random values generated for 177 missing data points
age_random.size
```

[17]:

```
177
```

[8]:

```
data.isna().sum()
```

[8]:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

[15]:

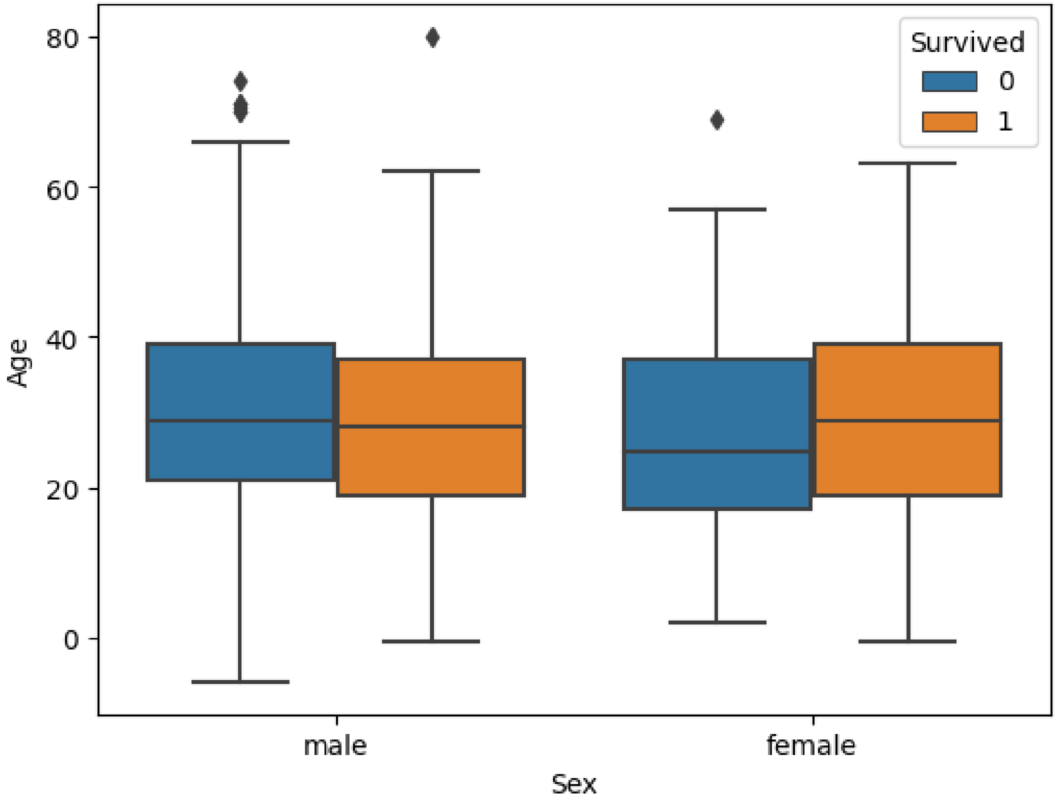
```
data.sample(7)
```

[15]:

```
     PassengerId  Survived  Pclass                            Name     Sex  \
205          206         0       3       Strom, Miss. Telma Matilda  female
794          795         0       3            Dantcheff, Mr. Ristiu    male
598          599         0       3                Boulos, Mr. Hanna    male
743          744         0       3                McNamee, Mr. Neal    male
810          811         0       3           Alexander, Mr. William    male
47            48         1       3         O'Driscoll, Miss. Bridget  female
604          605         1       1  Homer, Mr. Harry ("Mr E Haven")    male

           Age  SibSp  Parch  Ticket     Fare Cabin Embarked
205   2.000000      0      1  347054  10.4625    G6        S
```

```
794  25.000000      0      0  349203    7.8958   NaN        S
598   4.419244      0      0    2664    7.2250   NaN        C
743  24.000000      1      0  376566   16.1000   NaN        S
810  26.000000      0      0    3474    7.8875   NaN        S
47   35.287735      0      0   14311    7.7500   NaN        Q
604  35.000000      0      0  111426   26.5500   NaN        C
```

[14]:
```python
sns.boxplot(x='Sex', y='Age', hue='Survived', data=data);
```


Survived
Age
male
female
Sex