NAME: Gayatri Gosavi

ROLL NO.: TCOB13

ASSIGNMENT-9

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 required library*

import pandas as pd import numpy as np import seaborn as sns

import matplotlib.pyplot as plt

*#loading dataset*

data =

pd.read\_csv('https://raw.githubusercontent.com/dphi-official/Datasets/ master/titanic\_data.csv')

data.head()

PassengerId Survived Pclass \

1

1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1

0

3 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1

0

4 Allen, Mr. William Henry male 35.0

2 Heikkinen, Miss. Laina female 26.0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 0 | 3 |  | | |
| 1 | 2 | 1 | 1 |
| 2 | 3 | 1 | 3 |
| 3 | 4 | 1 | 1 |
| 4 | 5 | 0 | 3 |
|  |  |  |  |
|  |  |  |  | Name | Sex | Age |
| SibSp \ |  |  |  |  |  |  |
| 0 |  |  | Brau | nd, Mr. Owen Harris | male | 22.0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0 | Parch  0 | Ticket A/5 21171 | Fare Cabin Embarked 7.2500 NaN S | | |
| 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 0 | 373450 | 8.0500 | NaN | S |

data.describe()

PassengerId Survived Pclass

Name

0

0

0

0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PassengerId | Survived | Pclass | Age | SibSp | \ |
| count | 891.000000 | 891.000000 | 891.000000 | 714.000000 | 891.000000 |  |
| mean | 446.000000 | 0.383838 | 2.308642 | 29.699118 | 0.523008 |  |
| std | 257.353842 | 0.486592 | 0.836071 | 14.526497 | 1.102743 |  |
| min | 1.000000 | 0.000000 | 1.000000 | 0.420000 | 0.000000 |  |
| 25% | 223.500000 | 0.000000 | 2.000000 | 20.125000 | 0.000000 |  |
| 50% | 446.000000 | 0.000000 | 3.000000 | 28.000000 | 0.000000 |  |
| 75% | 668.500000 | 1.000000 | 3.000000 | 38.000000 | 1.000000 |  |
| max | 891.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parch | Fare |  |
| count | 891.000000 | 891.000000 |  |
| mean | 0.381594 | 32.204208 |  |
| std | 0.806057 | 49.693429 |  |
| min | 0.000000 | 0.000000 |  |
| 25% | 0.000000 | 7.910400 |  |
| 50% | 0.000000 | 14.454200 |  |
| 75% | 0.000000 | 31.000000 |  |
| max | 6.000000 | 512.329200 |  |

data.info()

<class 'pandas.core.frame.DataFrame'> RangeIndex: 891 entries, 0 to 890 Data columns (total 12 columns):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Column  PassengerId Survived Pclass | Non-Null Count  891 non-null  891 non-null  891 non-null | Dtype |  |
| 0  1  2 | int64 int64 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

data.isnull().sum()

|  |  |  |
| --- | --- | --- |
|  |  | |
| Sex | 0 |
| Age | 177 |
| SibSp | 0 |
| Parch | 0 |
| Ticket | 0 |
| Fare | 0 |
| Cabin | 687 |
| Embarked | 2 |
|  | dtype: int64 |  |

Here, we can see there are Null values in the dataset. Hence, we need to replace these values by mean (in case of numerical variables) or mode (in case of categorical variables)

data['Age'] = data['Age'].fillna(np.mean(data['Age']))

data['Cabin'] = data['Cabin'].fillna(data['Cabin'].mode()[0]) data['Embarked'] = data['Embarked'].fillna(data['Embarked'].mode()[0])

data.isnull().sum()

sns.boxplot(data['Sex'], data["Age"], data["Survived"], palette = 'Set2').set\_title('Plot for distribution of age with respect to each gender along with the information about whether they survived or not') plt.show()

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 0

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 0

Embarked 0

dtype: int64

