ASSIGNMENT 13/TASK13

SHWETA JHA

REGISTRATION ID- GO_STP_12574

Exploratory Data Analysis (EDA) of Titanic Survival Problem.

To do the same we will use the Pandas, Seaborn and Matplotlib library.

Dataset contains the details of the passengers who had boarded the ship.

Dataset can be downloaded from Kaggle.

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

```
df=pd.read_csv("/content/titanic_original.csv")
df.head()
```

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
0	1.0	1.0	Allen, Miss. Elisabeth Walton	female	29.0000	0.0	0.0	24160	211.3375	B5
1	1.0	1.0	Allison, Master. Hudson Trevor	male	0.9167	1.0	2.0	113781	151.5500	C22 C26
2	1.0	0.0	Allison, Miss. Helen Loraine	female	2.0000	1.0	2.0	113781	151.5500	C22 C26

Explore All Data

```
df.describe()
```

	pclass	survived	age	sibsp	parch	fare	
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000	12
mean	2.294882	0.381971	29.881135	0.498854	0.385027	33.295479	16
std	0.837836	0.486055	14.413500	1.041658	0.865560	51.758668	9
min	1.000000	0.000000	0.166700	0.000000	0.000000	0.000000	
25%	2.000000	0.000000	21.000000	0.000000	0.000000	7.895800	7
50%	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	15

```
df.tail()
```

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabi
1305	3.0	0.0	Zabour, Miss. Thamine	female	NaN	1.0	0.0	2665	14.4542	NaN
1306	3.0	0.0	Zakarian, Mr. Mapriededer	male	26.5	0.0	0.0	2656	7.2250	NaN
1307	3.0	0.0	Zakarian, Mr. Ortin	male	27.0	0.0	0.0	2670	7.2250	NaN

```
df.shape
```

```
(1310, 14)
```

```
df.size
```

```
18340
```

```
df.info
```

```
<bound method DataFrame.info of       pclass  survived  ...   body
0        1.0       1.0  ...    NaN                        St Louis, MO
1        1.0       1.0  ...    NaN  Montreal, PQ / Chesterville, ON
2        1.0       0.0  ...    NaN  Montreal, PQ / Chesterville, ON
3        1.0       0.0  ...  135.0  Montreal, PQ / Chesterville, ON
4        1.0       0.0  ...    NaN  Montreal, PQ / Chesterville, ON
...      ...       ...  ...    ...                              ...
1305     3.0       0.0  ...    NaN                              NaN
1306     3.0       0.0  ...  304.0                              NaN
1307     3.0       0.0  ...    NaN                              NaN
1308     3.0       0.0  ...    NaN                              NaN
1309     NaN       NaN  ...    NaN                              NaN

[1310 rows x 14 columns]>
```

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

```
pclass          1
survived        1
```

```
name            1
sex             1
age           264
sibsp           1
parch           1
ticket          1
fare            2
cabin        1015
embarked        3
boat          824
body         1189
home.dest     565
dtype: int64
```

```
df.isna().count()
```

```
pclass       1310
survived     1310
name         1310
sex          1310
age          1310
sibsp        1310
parch        1310
ticket       1310
fare         1310
cabin        1310
embarked     1310
boat         1310
body         1310
home.dest    1310
dtype: int64
```

```
df.columns
```

```
Index(['pclass', 'survived', 'name', 'sex', 'age', 'sibsp', 'parch', 'ticket',
       'fare', 'cabin', 'embarked', 'boat', 'body', 'home.dest'],
      dtype='object')
```

```
df['sex'].value_counts()
```

```
male      843
female    466
Name: sex, dtype: int64
```

```
df['survived'].value_counts()
```

```
0.0    809
1.0    500
Name: survived, dtype: int64
```

```
df['fare'].value_counts()
```

```
8.0500     60
13.0000    59
7.7500     55
26.0000    50
```

```
7.8958     49
           ..
13.7917     1
10.7083     1
7.7417      1
7.8208      1
34.6542     1
Name: fare, Length: 281, dtype: int64
```

```
df['ticket'].value_counts()
```

```
CA. 2343        11
1601             8
CA 2144          8
347082           7
S.O.C. 14879     7
                ..
11752            1
365235           1
349212           1
347468           1
A/5. 2151        1
Name: ticket, Length: 929, dtype: int64
```

Missing Values

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

```
df['cabin']=df['cabin'].fillna(np.random.choice(['A','B','C','D','E','F']))
```

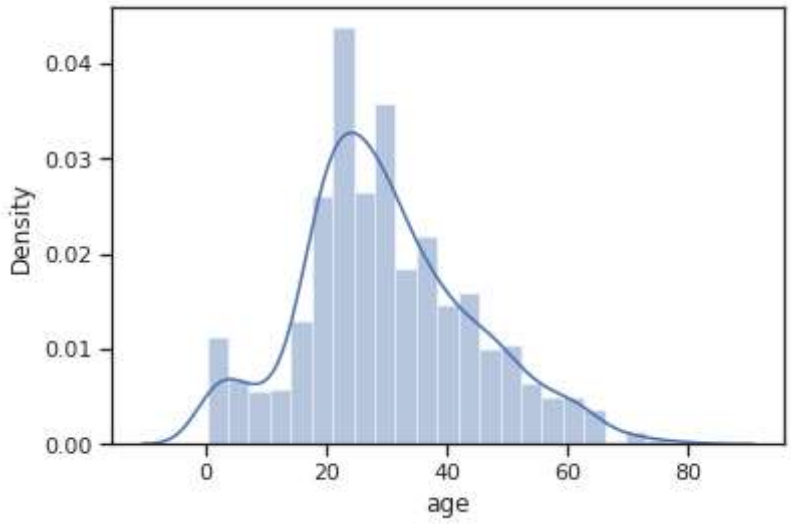
```
df.dropna(inplace=True)
df.isna().sum()
```

```
pclass       0
survived     0
name         0
sex          0
age          0
sibsp        0
parch        0
ticket       0
fare         0
cabin        0
embarked     0
boat         0
body         0
home.dest    0
dtype: int64
```

Data Visualizations

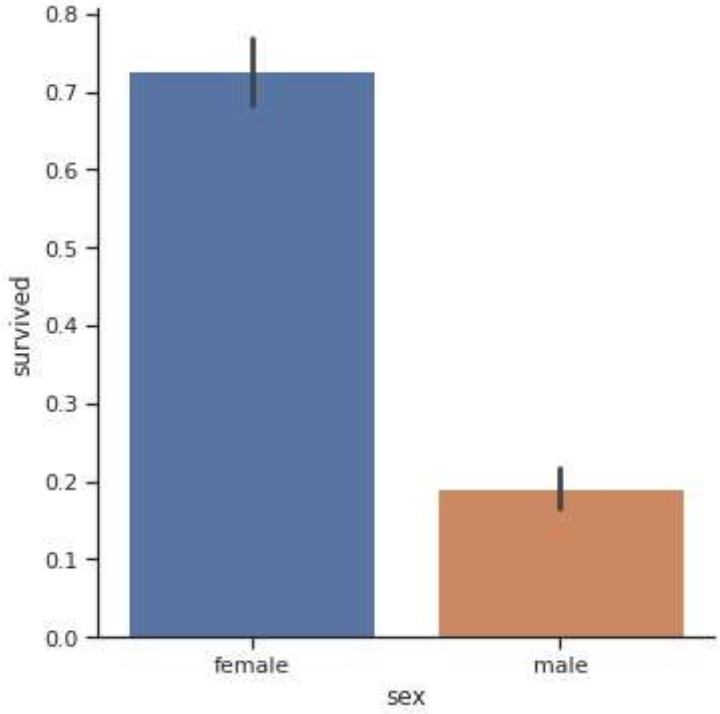
```
sns.distplot(df['age'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7eff8df69950>
```



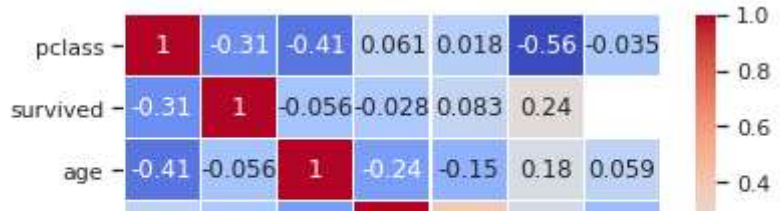

```
sns.catplot(x="sex", y="survived", kind="bar", data=df)
```

```
<seaborn.axisgrid.FacetGrid at 0x7eff8da0cf50>
```



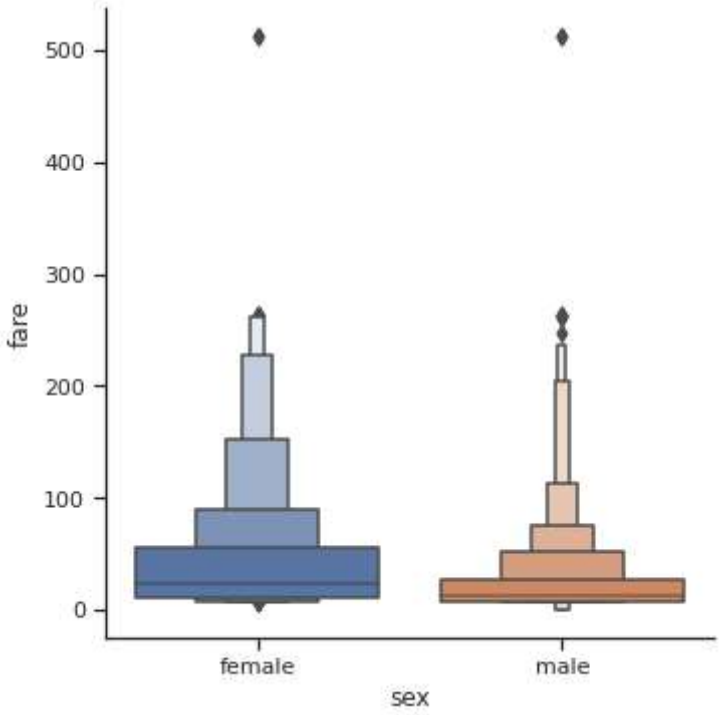

```
sns.heatmap(df.corr(), cmap='coolwarm', annot=True, linewidths=0.30)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7eff8dd285d0>
```



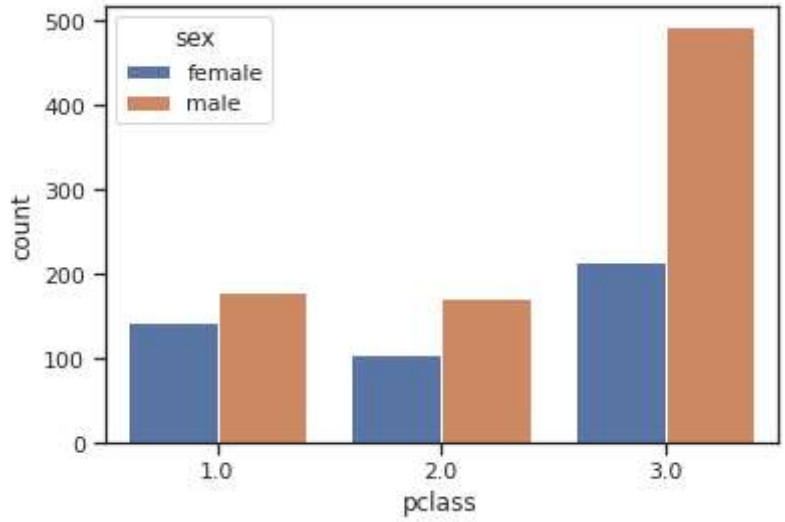

```
sns.catplot(x="sex",y="fare",data=df,kind="boxen")
```

```
<seaborn.axisgrid.FacetGrid at 0x7eff7ef29b50>
```

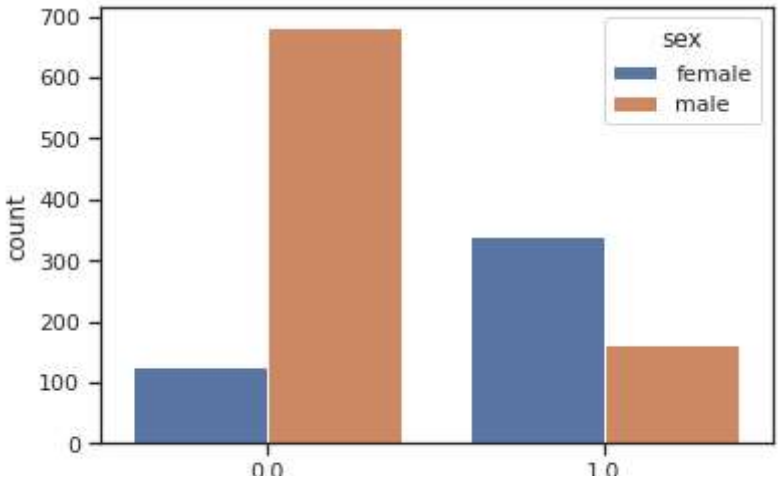



```
sns.countplot(x="pclass", hue="sex", data=df)
```

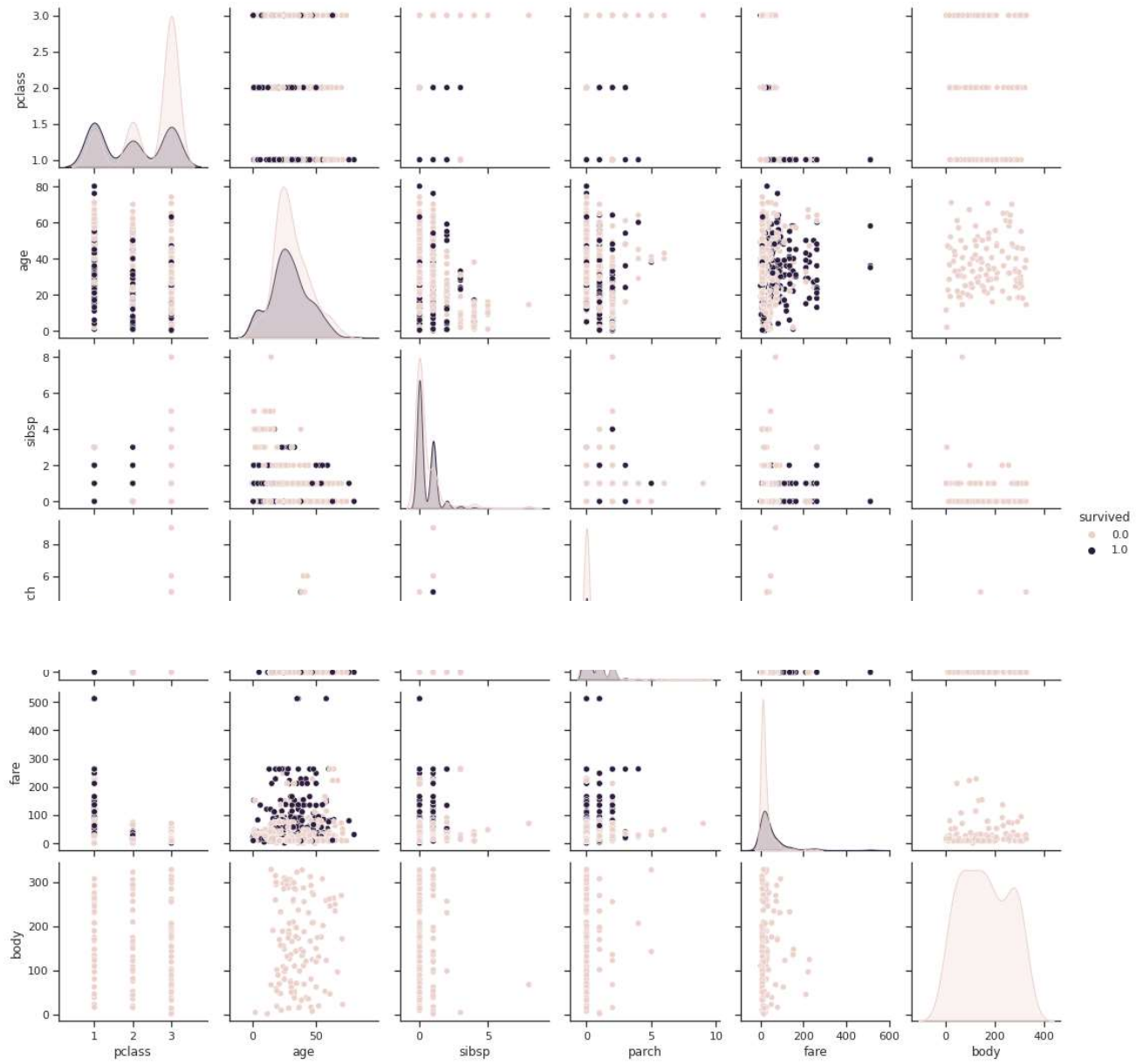
```
<matplotlib.axes._subplots.AxesSubplot at 0x7eff7f70a350>
```




```
sns.countplot(x="survived", hue="sex", data=df)
plt.show()
```




```
sns.pairplot(df,hue='survived')
plt.show()
```



27s completed at 1:56 AM