Name :Yehia Ahmed Hassan EL-Boudy (20191311656).
Ahmed Hany Mohamed Abdulawahab (20191310663).
Mamoun Mohamed Hassan Abdelbaey (20191312895).
ID Number : 20191311656.
Subject : Data science methodolgy.
To Professor :Dr/ Magda Matbouly

Project

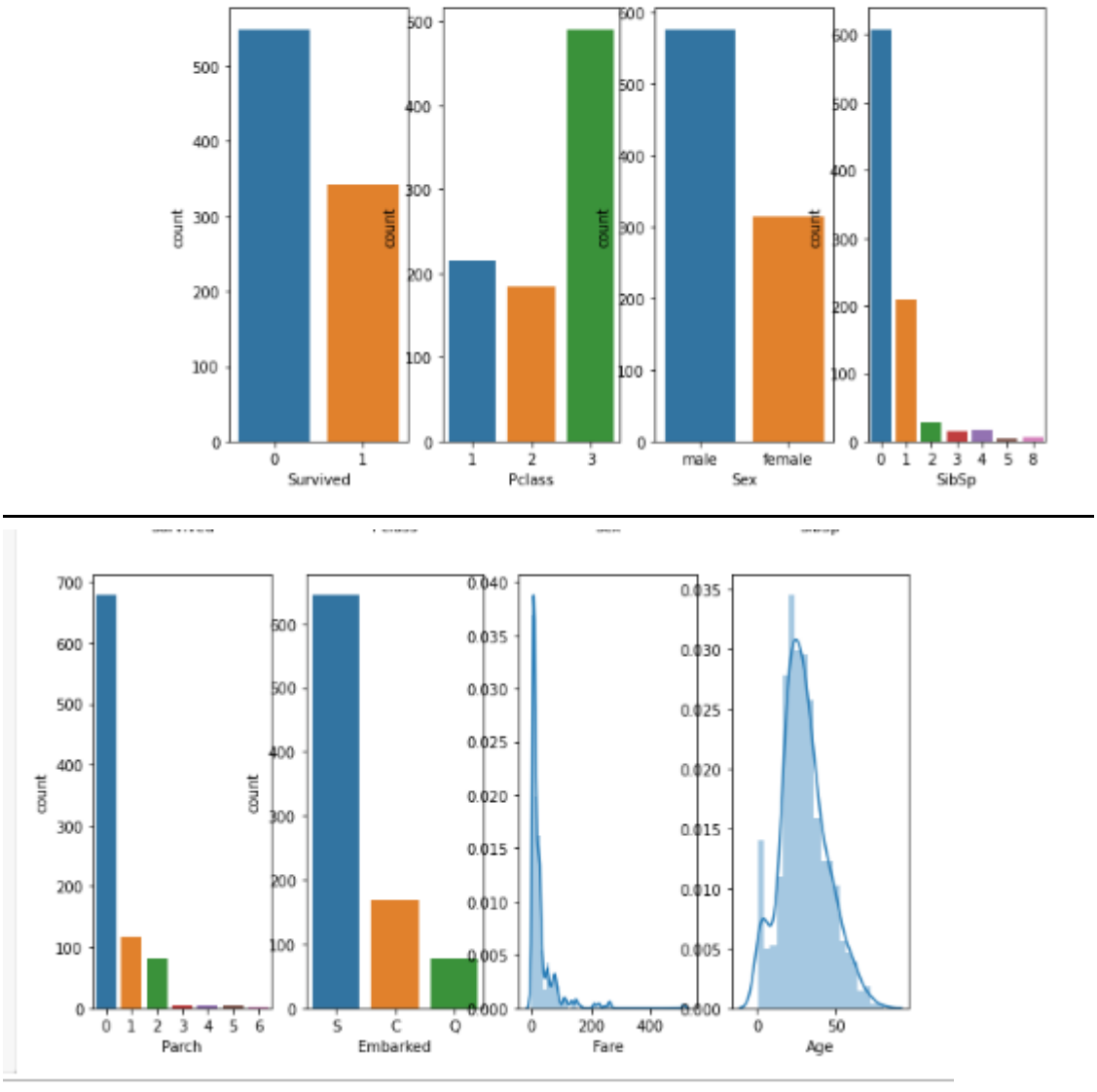
- In this report we are going to visualize our titanic data set. This visualization helps us to understand our data and the relationship between every feature and the goal feature (Survived featured).
- We used some packages that contains and provide us with the right tools that helps in the visualization process :

```
import matplotlib.pyplot as plt
import seaborn as sns
from seaborn import distplot
from IPython.display import Image,display
%matplotlib inline
```

- ***Visualizing the train data set:***
  - We used the count plot to count the discrete values and the distplot for the continuous variables.

```
In [126]: #visualizing our data
          fig,axes=plt.subplots(2,4,figsize=(10,12))
          sns.countplot('Survived',data=train,ax=axes[0,0])
          sns.countplot('Pclass',data=train,ax=axes[0,1])
          sns.countplot('Sex',data=train,ax=axes[0,2])
          sns.countplot('SibSp',data=train,ax=axes[0,3])
          sns.countplot('Parch',data=train,ax=axes[1,0])
          sns.countplot('Embarked',data=train,ax=axes[1,1])
          sns.distplot(train['Fare'].dropna(),kde=True,ax=axes[1,2])
          sns.distplot(train['Age'].dropna(),kde=True,ax=axes[1,3])
```

```
Out[126]: <matplotlib.axes._subplots.AxesSubplot at 0x161097831c0>
```

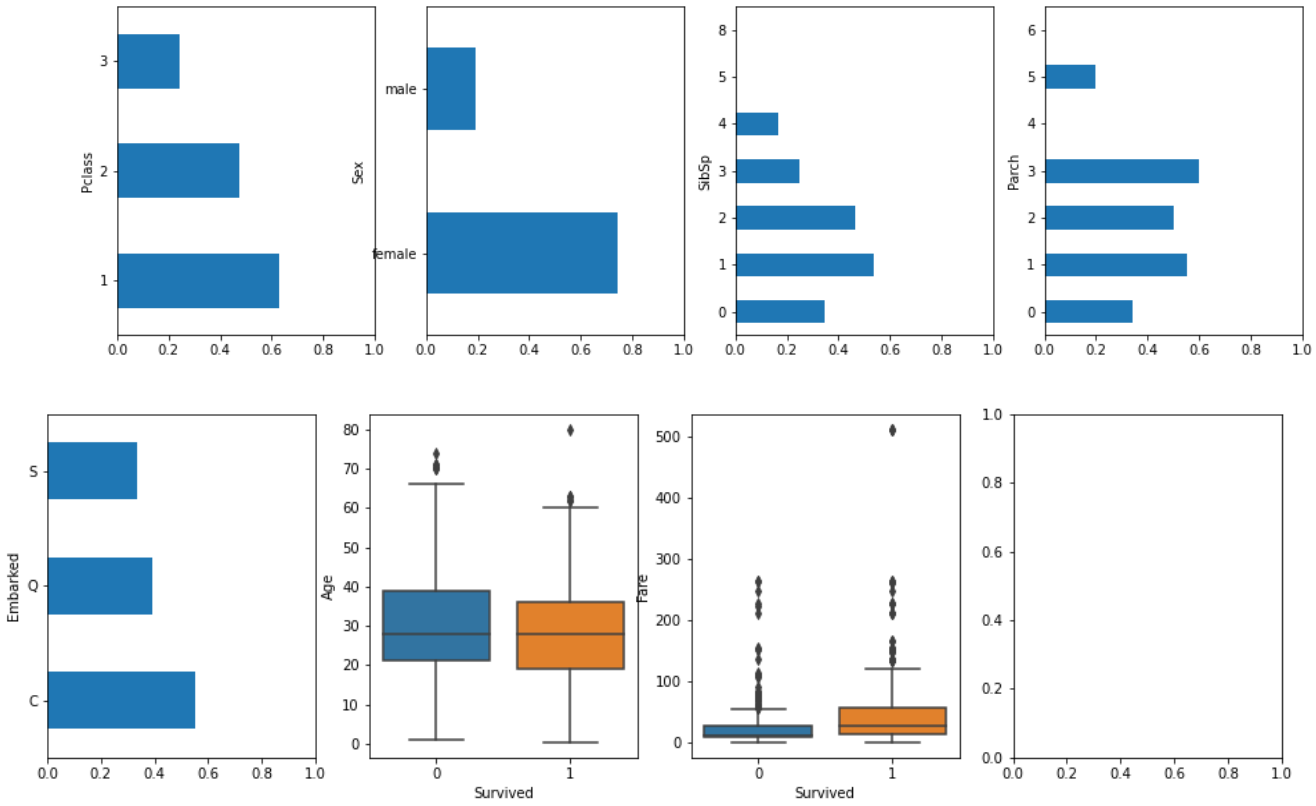


- From the previous picture, we have noticed that the fare data is rightly skewed, so we can't fill the nulls with the mean or median.
- However, the age is normal distributed so we can fill the nulls with the mean of all the values of the column.

- Visualizing every feature with the survive feature (the goal) by using groupby function.

In [127]:
```
#visualizing the relationship between every feature with the servived feature which is the goal
figbi,axesbi=plt.subplots(2, 4,figsize=(16,10))
train.groupby('Pclass')['Survived'].mean().plot(kind='barh',ax=axesbi[0,0],xlim=[0,1])
train.groupby('Sex')['Survived'].mean().plot(kind='barh',ax=axesbi[0,1],xlim=[0,1])
train.groupby('SibSp')['Survived'].mean().plot(kind='barh',ax=axesbi[0,2],xlim=[0,1])
train.groupby('Parch')['Survived'].mean().plot(kind='barh',ax=axesbi[0,3],xlim=[0,1])
train.groupby('Embarked')['Survived'].mean().plot(kind='barh',ax=axesbi[1,0],xlim=[0,1])
sns.boxplot(x='Survived',y='Age',data=train,ax=axesbi[1,1])
sns.boxplot(x='Survived',y='Fare',data=train,ax=axesbi[1,2])
```

Out[127]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1610a81d940>
```

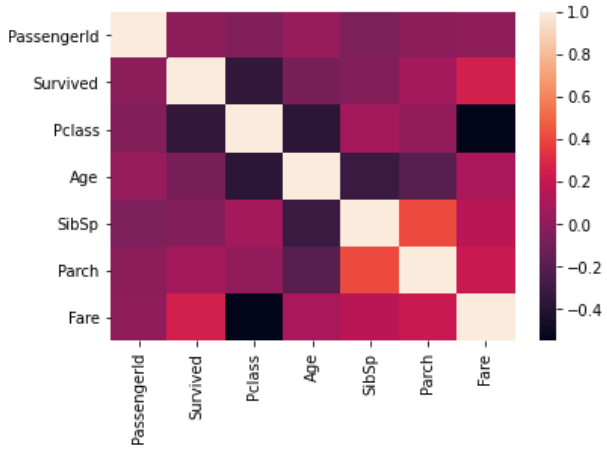


- From these visualizations, we familiarized ourselves that the higher you Pclass you are in the higher the survival rate; Females have more survivals rate than males; the embarked from C has more survival rate than the others.

- **Heat map:**

```
]: #correlation between features
   corr=train.corr()
   sns.heatmap(corr)
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x1610a98e730>
```



- Heat map helps us to know if the correlation between each feature is either a positive or negative correlation.
- For example, we can tell that the survival rate and the pclass have a negative correlation.