**Data Visualization - I**

**Problem statement :**

**Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship.  
1. Use the Seaborn library to see if we can find any patterns in the data.  
2. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.**

In [2]:

*#imports*

**import** pandas **as** pd

**import** numpy **as** np

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

In [4]:

data **=** pd**.**read\_csv('train.csv')

In [5]:

data**.**sample(5)

Out[5]:

|  | **PassengerId** | **Survived** | **Pclass** | **Name** | **Sex** | **Age** | **SibSp** | **Parch** | **Ticket** | **Fare** | **Cabin** | **Embarked** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **787** | 788 | 0 | 3 | Rice, Master. George Hugh | male | 8.0 | 4 | 1 | 382652 | 29.1250 | NaN | Q |
| **671** | 672 | 0 | 1 | Davidson, Mr. Thornton | male | 31.0 | 1 | 0 | F.C. 12750 | 52.0000 | B71 | S |
| **14** | 15 | 0 | 3 | Vestrom, Miss. Hulda Amanda Adolfina | female | 14.0 | 0 | 0 | 350406 | 7.8542 | NaN | S |
| **702** | 703 | 0 | 3 | Barbara, Miss. Saiide | female | 18.0 | 0 | 1 | 2691 | 14.4542 | NaN | C |
| **413** | 414 | 0 | 2 | Cunningham, Mr. Alfred Fleming | male | NaN | 0 | 0 | 239853 | 0.0000 | NaN | S |

In [7]:

data**.**info(), data**.**isna()**.**sum()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 891 entries, 0 to 890

Data columns (total 12 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 PassengerId 891 non-null int64

1 Survived 891 non-null int64

2 Pclass 891 non-null 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

Out[7]:

(None,

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

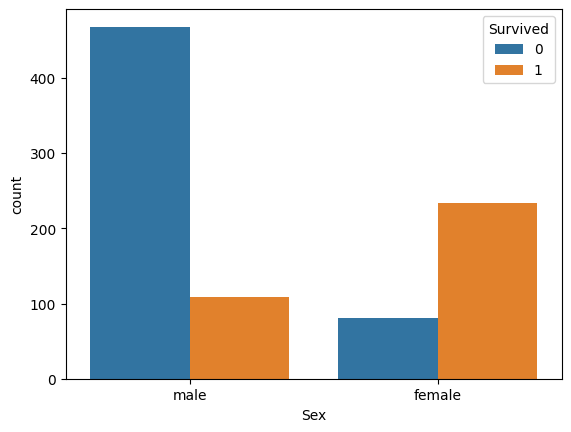
dtype: int64)

In [10]:

sns**.**countplot(data**=**data, x**=**'Sex', hue**=** 'Survived')

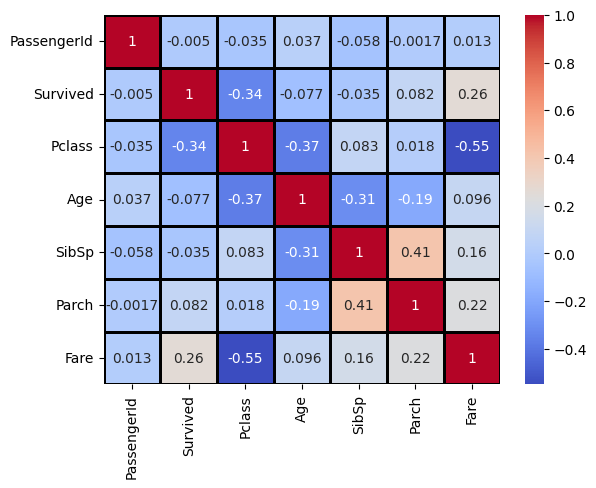
Out[10]:

<AxesSubplot:xlabel='Sex', ylabel='count'>



In [19]:

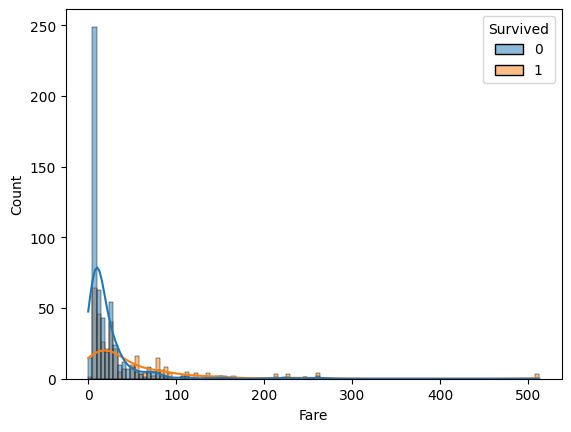
sns**.**heatmap(data**.**corr(), annot**=** **True**, cmap**=** 'coolwarm', linewidths **=** 1, linecolor **=** 'black');



From the above corelation matrix, it is clear that 'Fare' and 'Survived' have a positive corelation.  
Meaning higher the cost of the ticket, higher is the chance of survival.

In [23]:

sns**.**histplot(data **=** data, x **=** 'Fare', hue **=** 'Survived',kde **=** **True**);



The corelation can be confirmed from the above histplot.  
The number of people having cheap tickets was extremely high. However only a few of them survived.  
On the other hand, most of the handful people with expensive tickets survived.