Data Visualization I

1. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. 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 [3]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
%matplotlib inline
```

Load data and basic stats

In [5]:

```
df = pd.read_csv("train.csv")
```

In [8]:

```
df.shape
```

Out[8]:

```
(891, 12)
```

In [10]:

```
df.head()
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [12]:

```
df.info()
```

```
<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
```

In [16]:

```
df.columns
```

Out[16]:

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [18]:

```
df.describe()
```

Out[18]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [20]:

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

Out[20]:

```
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 [22]:

```
df["Age"] = df["Age"].fillna(df["Age"].mean())
```

In [24]:

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

Out[24]:

```
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
```

Visualization

In [27]:

```
df["Name"]
```

Out[27]:

```
0                                Braund, Mr. Owen Harris
1      Cumings, Mrs. John Bradley (Florence Briggs Th...
2                                 Heikkinen, Miss. Laina
3           Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                               Allen, Mr. William Henry
                             ...                        
886                                Montvila, Rev. Juozas
887                         Graham, Miss. Margaret Edith
888             Johnston, Miss. Catherine Helen "Carrie"
889                                Behr, Mr. Karl Howell
890                                  Dooley, Mr. Patrick
Name: Name, Length: 891, dtype: object
```

In [29]:

```
df["Sex"].value_counts()
```

Out[29]:

```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [31]:

```
df["Ticket"].value_counts()
```

Out[31]:

```
Ticket
347082      7
CA. 2343    7
1601        7
3101295     6
CA 2144     6
           ..
9234        1
19988       1
2693        1
PC 17612    1
370376      1
Name: count, Length: 681, dtype: int64
```

In [33]:

```
df["Cabin"].value_counts()
```

Out[33]:

```
Cabin
B96 B98        4
G6             4
C23 C25 C27    4
C22 C26        3
F33            3
              ..
E34            1
C7             1
C54            1
E36            1
C148           1
Name: count, Length: 147, dtype: int64
```

In [35]:

```
df["Embarked"].value_counts()
```

Out[35]:

```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [37]:

```
def fun1(value):
    if (value == "male"):
        return 1
    else:
        return 0
```

In [39]:

```
def fun2(value):
    if (value == 'S'):
        return 0
    elif (value == 'C'):
        return 1
    elif (value == 'Q'):
        return 2
    else:
        return 0
```

In [41]:

```
df["Sex"] = df["Sex"].apply(fun1)
```

In [43]:

```
df["Embarked"] = df["Embarked"].apply(fun2)
```

In [45]:

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

Out[45]:

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

In [47]:

```
df = df.drop("Cabin", axis=1)
```

In [49]:

```
df.shape
```

Out[49]:

```
(891, 11)
```

In [51]:

```
df.shape
```

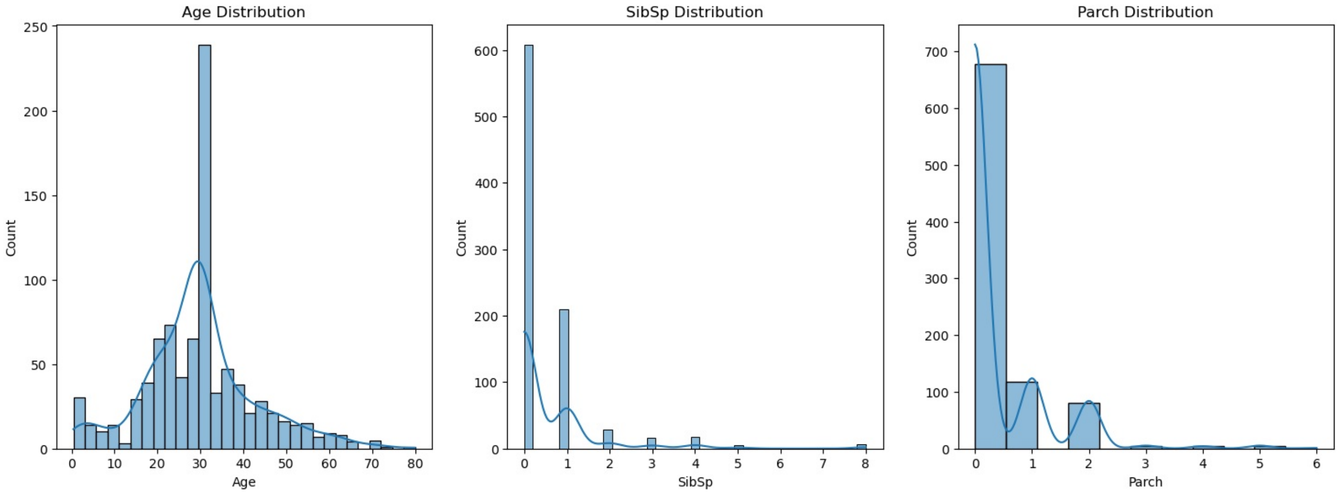
Out[51]:

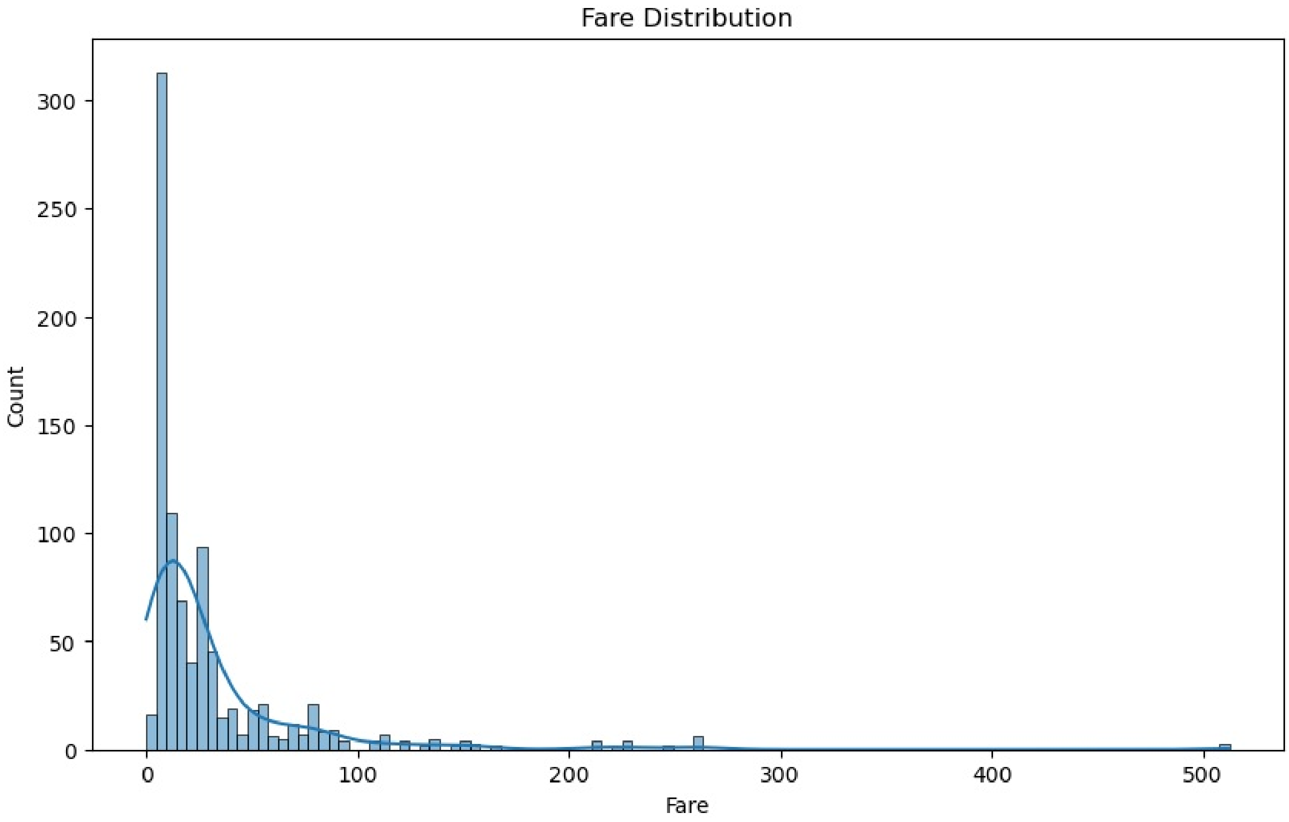
```
(891, 11)
```

In [59]:

```
 #Set up the figure and axes
fig, axes = plt.subplots(1, 3, figsize=(18, 6))
```

```python
# Age Distribution
sns.histplot(data=df, x='Age', kde=True, ax=axes[0])
axes[0].set_title('Age Distribution')

# SibSp Distribution
sns.histplot(data=df, x='SibSp', kde=True, ax=axes[1])
axes[1].set_title('SibSp Distribution')

# Parch Distribution
sns.histplot(data=df, x='Parch', kde=True, ax=axes[2])
axes[2].set_title('Parch Distribution')

#plt.tight_layout()
#plt.show()

# Fare Distribution
plt.figure(figsize=(10, 6))
sns.histplot(data=df, x='Fare', kde=True)
plt.title('Fare Distribution')
plt.show()
```





In [61]:

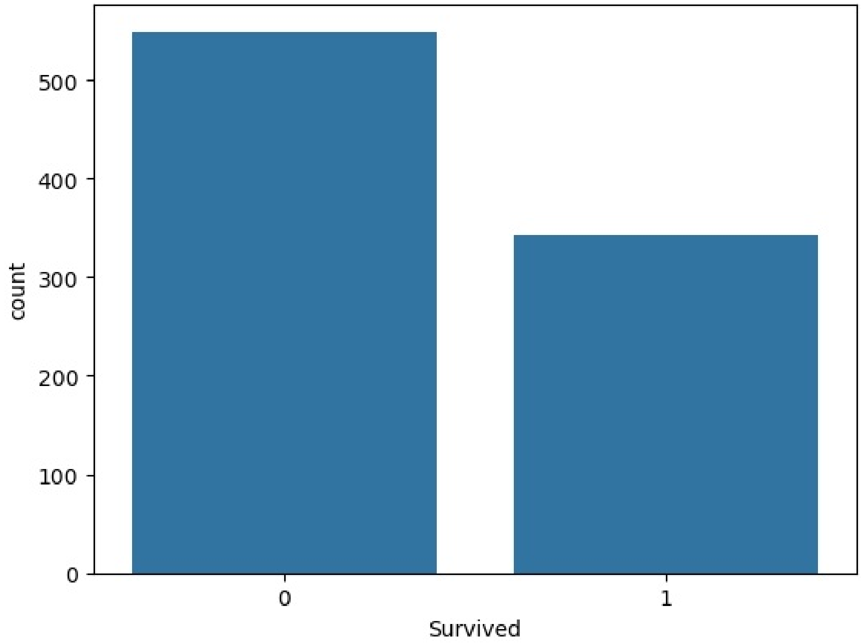
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 11 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    int64  
 5   Age          891 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  Embarked     891 non-null    int64  
dtypes: float64(2), int64(7), object(2)
memory usage: 76.7+ KB
```

"Survived" is the label

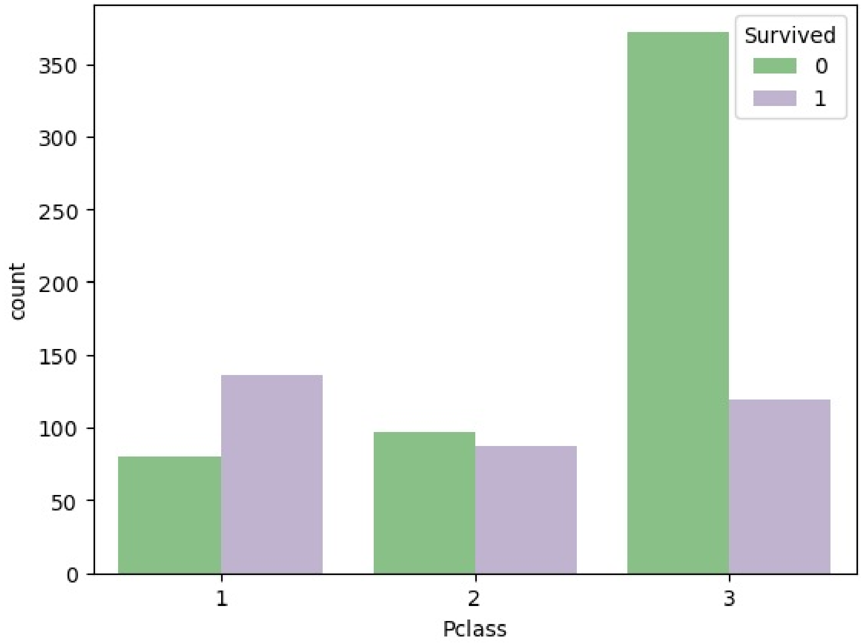
In [64]:

```
sns.countplot(df, x="Survived")
plt.show()
```



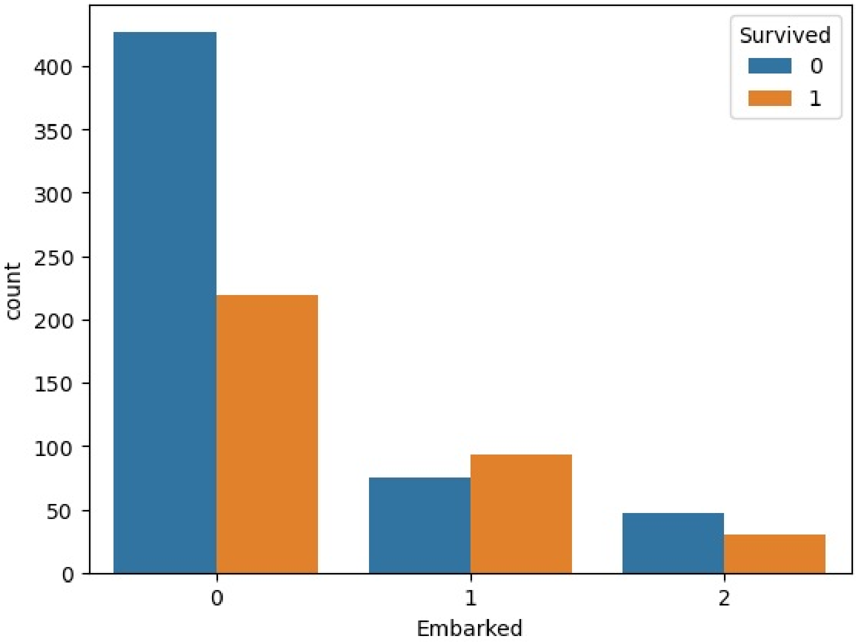
In [325…

```
sns.countplot(df,x="Pclass", hue="Survived",palette="Accent")
plt.show()
```



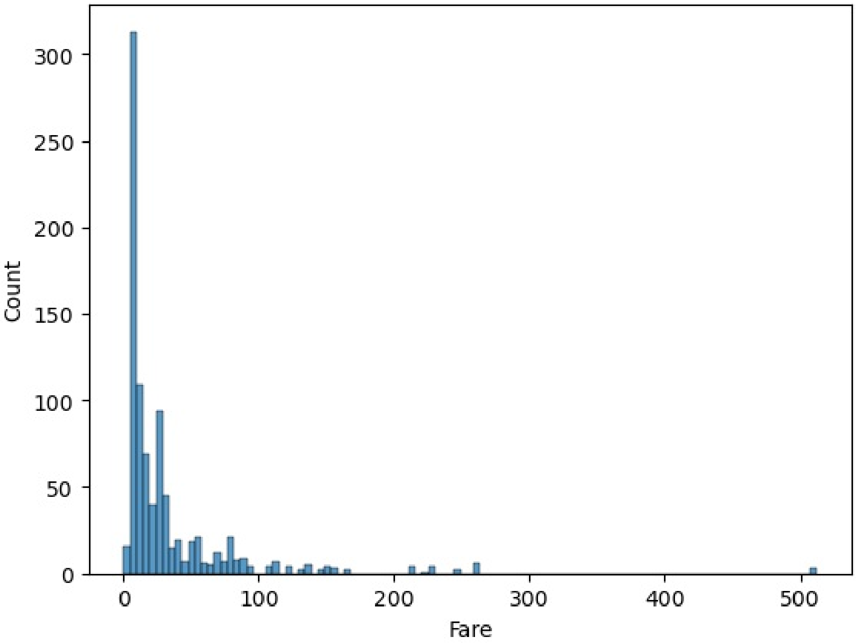
In [68]:

```
sns.countplot(df,x="Embarked",hue="Survived")
plt.show()
```



In [70]:

```
sns.histplot(df["Fare"])
plt.show()
```



In [ ]:

In [ ]:

In [ ]: