In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv('titanic_survivors.csv')
```

In [3]:
```
df.head()
```

Out[3]:

	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 [4]:
```
df.tail()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

In [5]:
```
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 [6]:
```
df.isnull().sum()
```

Out[6]:
```
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 [7]:
```
df.shape
```

Out[7]:
```
(891, 12)
```

In [8]:
```
df['Age'].fillna(df['Age'].median(), inplace = True)
```

In [9]:
```
df.drop('Cabin', axis=1, inplace=True)
```

In [10]:
```
df = df.dropna(subset=['Embarked'])
```

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

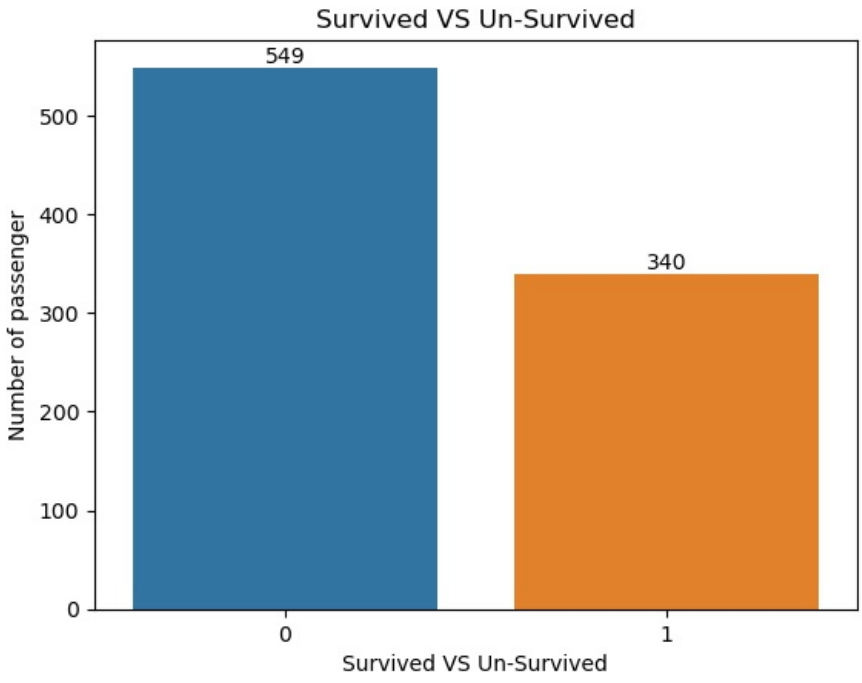
In [12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  889 non-null    int64  
 1   Survived     889 non-null    int64  
 2   Pclass       889 non-null    int64  
 3   Name         889 non-null    object 
 4   Sex          889 non-null    object 
 5   Age          889 non-null    float64
 6   SibSp        889 non-null    int64  
 7   Parch        889 non-null    int64  
 8   Ticket       889 non-null    object 
 9   Fare         889 non-null    float64
 10  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(4)
memory usage: 83.3+ KB
```

In [13]:
```
df['Survived'].value_counts()
```

Out[13]:
```
0    549
1    340
Name: Survived, dtype: int64
```

In [14]:
```
ax = sns.countplot(x='Survived', data= df)
for p in ax.containers:
    ax.bar_label(p)
plt.xlabel('Survived VS Un-Survived')
plt.ylabel('Number of passenger')
plt.title('Survived VS Un-Survived')
plt.show()
```



In [15]:
```
df.corr()
```

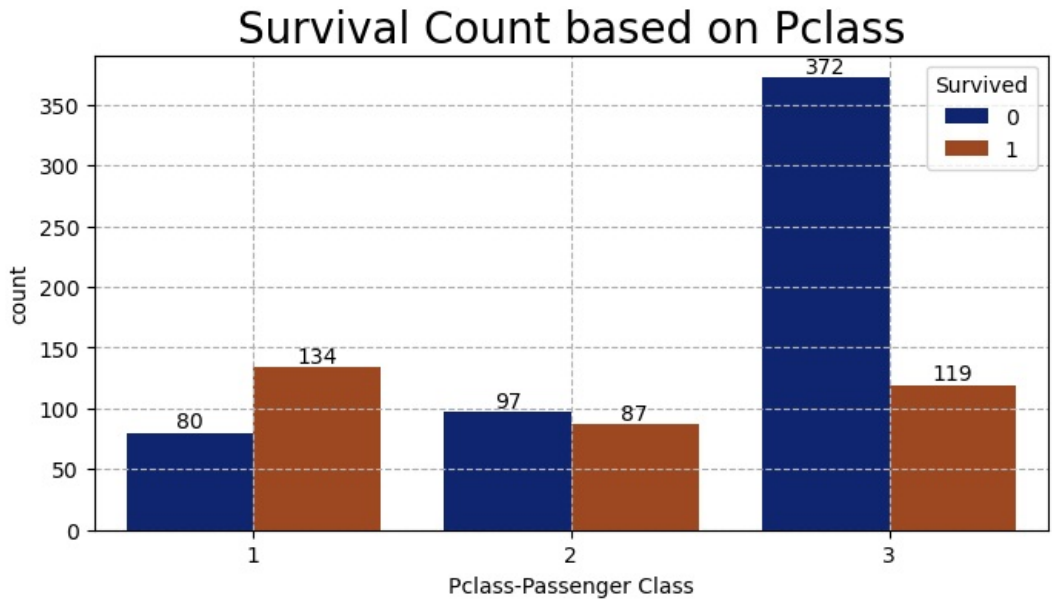
```
C:\Users\Admin\AppData\Local\Temp\ipykernel_8828\1134722465.py:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  df.corr()
```

Out[15]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005028	-0.035330	0.031319	-0.057686	-0.001657	0.012703
Survived	-0.005028	1.000000	-0.335549	-0.069822	-0.034040	0.083151	0.255290
Pclass	-0.035330	-0.335549	1.000000	-0.336512	0.081656	0.016824	-0.548193
Age	0.031319	-0.069822	-0.336512	1.000000	-0.232543	-0.171485	0.093707
SibSp	-0.057686	-0.034040	0.081656	-0.232543	1.000000	0.414542	0.160887
Parch	-0.001657	0.083151	0.016824	-0.171485	0.414542	1.000000	0.217532
Fare	0.012703	0.255290	-0.548193	0.093707	0.160887	0.217532	1.000000

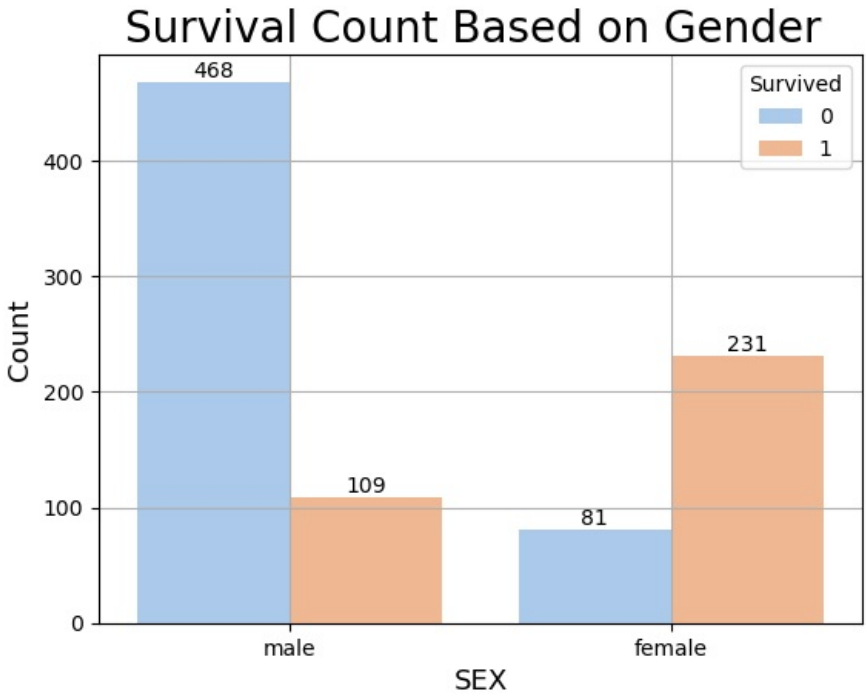
In [16]:
```
plt.figure(figsize=(8,4))
ax = sns.countplot(x='Pclass',hue='Survived',data=df, palette='dark')
for z in ax.containers:
    ax.bar_label(z)

plt.title('Survival Count based on Pclass', fontsize=20)
plt.xlabel('Pclass-Passenger Class')
plt.ylabel('count')
plt.grid(True,linestyle='--')
plt.show()
```



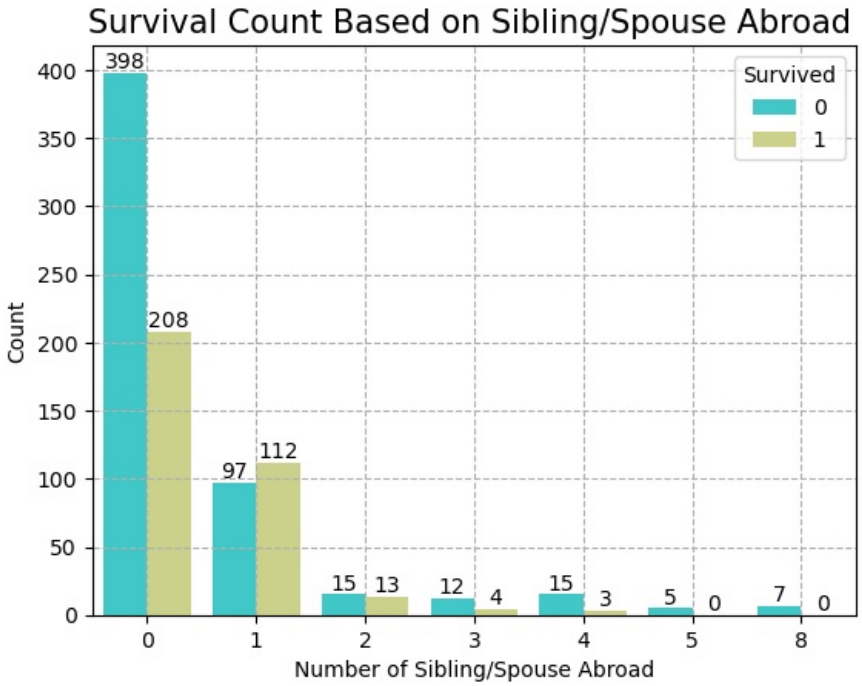
In [17]:
```
ax = sns.countplot(x='Sex',hue='Survived', data=df, palette='pastel')
for z in ax.containers:
    ax.bar_label(z)

plt.title('Survival Count Based on Gender ', fontsize=20)
plt.xlabel('SEX', fontsize=13)
plt.ylabel('Count', fontsize=13)
plt.grid()
plt.show()
```



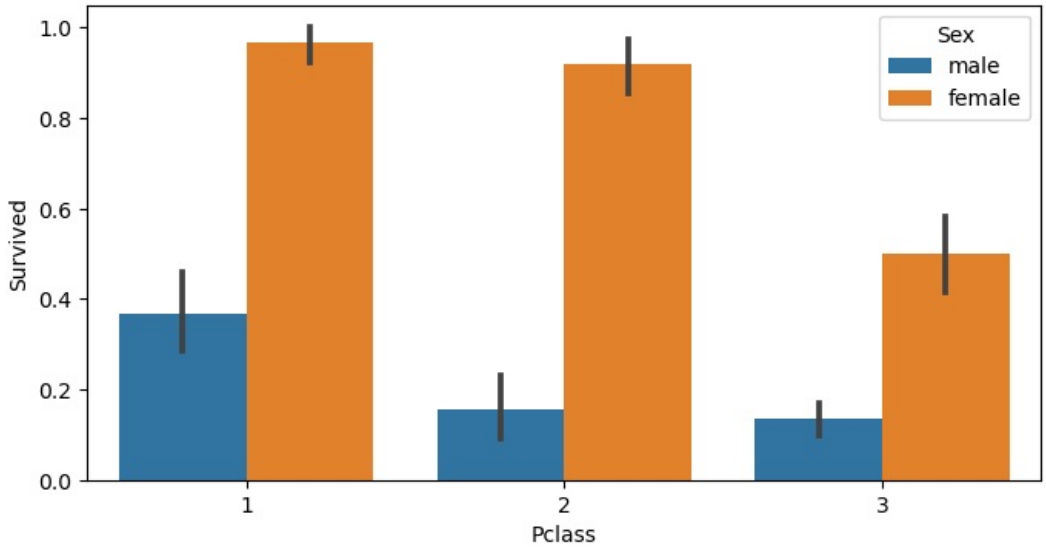
In [18]:
```
ax = sns.countplot(x='SibSp',hue='Survived', data=df, palette='rainbow')
for z in ax.containers:
    ax.bar_label(z)

plt.title('Survival Count Based on Sibling/Spouse Abroad ', fontsize=15)
plt.xlabel('Number of Sibling/Spouse Abroad')
plt.ylabel('Count')
plt.grid(True, linestyle='--')
plt.show()
```



In [19]:
```
plt.figure(figsize=(8,4))
sns.barplot(x='Pclass', y='Survived', hue='Sex', data=df)
```

Out[19]:
```
<Axes: xlabel='Pclass', ylabel='Survived'>
```



In [20]:
```
df = pd.get_dummies(df)
```

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Columns: 1581 entries, PassengerId to Embarked_S
dtypes: float64(2), int64(5), uint8(1574)
memory usage: 1.4 MB
```

In [22]:
```
df.head()
```

Out[22]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Name_Abbing, Mr. Anthony	Name_Abbott, Mr. Rossmore Edward	Name_Abbott, Mrs. Stanton (Rosa Hunt)	...	Ticket_W./C. 14263	Ticket_W./C. 6607
0	1	0	3	22.0	1	0	7.2500	0	0	0	...	0	0
1	2	1	1	38.0	1	0	71.2833	0	0	0	...	0	0
2	3	1	3	26.0	0	0	7.9250	0	0	0	...	0	0
3	4	1	1	35.0	1	0	53.1000	0	0	0	...	0	0
4	5	0	3	35.0	0	0	8.0500	0	0	0	...	0	0

5 rows × 1581 columns

In [23]:
```
x = df.drop('Survived', axis=1)
y = df['Survived']
```

In [24]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.2, random_state=42)
```

In [25]:
```
x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

Out[25]:
```
((711, 1580), (178, 1580), (711,), (178,))
```

In [26]:
```
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier()
model
```

Out[26]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [27]:
```
model.fit(x_train,y_train)
```

Out[27]:
```
▾ RandomForestClassifier
RandomForestClassifier()
```

In [28]:
```
y_pred = model.predict(x_test)
y_pred
```

Out[28]:
```
array([0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0,
       0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0,
       0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0,
       0, 1], dtype=int64)
```

In [29]:
```
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test,y_pred)
accuracy
```

Out[29]:
```
0.8202247191011236
```

In [ ]: