In [1]:
```
import numpy as np
import pandas as pd
```

Data Set

In [2]:
```
data = pd.read_csv('Titanic-Dataset.csv')
```

In [4]:
```
data.head()
```

Out[4]:

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

In [5]:
```
data.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
```

Data Cleaning

In [7]:
```
data.isnull().sum()
```

Out[7]:
```
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 [8]:
```
data.drop(columns=['Cabin'],inplace=True)
```

In [16]:
```
data['Age'].fillna(data['Age'].median(),inplace=True)
data['Embarked'].fillna(data['Embarked'].mode()[0],inplace=True)
```

In [17]:
```
data.isnull().sum()
```

Out[17]:
```
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 [20]:
```
data.drop_duplicates(inplace=True)
```

In [25]:
```
data.head()
```

Out[25]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	1	0	3	male	22.0	1	0	7.2500	S
1	2	1	1	female	38.0	1	0	71.2833	C
2	3	1	3	female	26.0	0	0	7.9250	S
3	4	1	1	female	35.0	1	0	53.1000	S
4	5	0	3	male	35.0	0	0	8.0500	S

In [26]:
```
data.drop(columns=['PassengerId'],inplace=True)
```

In [30]:
```
data
```

Out[30]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S
887	1	1	female	19.0	0	0	30.0000	S
888	0	3	female	28.0	1	2	23.4500	S
889	1	1	male	26.0	0	0	30.0000	C
890	0	3	male	32.0	0	0	7.7500	Q

891 rows × 8 columns

Preprocessing

In [33]:
```
data = pd.get_dummies(data,columns=['Sex','Embarked'],dtype=int)
```

In [35]:
```
data
```

Out[35]:

	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Em
0	0	3	22.0	1	0	7.2500	0	1	0	
1	1	1	38.0	1	0	71.2833	1	0	1	
2	1	3	26.0	0	0	7.9250	1	0	0	
3	1	1	35.0	1	0	53.1000	1	0	0	
4	0	3	35.0	0	0	8.0500	0	1	0	
...	...	...	...	...	...	...	...	...	...	
886	0	2	27.0	0	0	13.0000	0	1	0	
887	1	1	19.0	0	0	30.0000	1	0	0	
888	0	3	28.0	1	2	23.4500	1	0	0	
889	1	1	26.0	0	0	30.0000	0	1	1	
890	0	3	32.0	0	0	7.7500	0	1	0	

891 rows × 11 columns

In [37]:
```
from sklearn.preprocessing import StandardScaler
stscaler = StandardScaler()
```

In [39]:
```
data[['Age','Fare']] =  stscaler.fit_transform(data[['Age','Fare']])
```

In [46]:
```
data[['Age','Fare']]
```

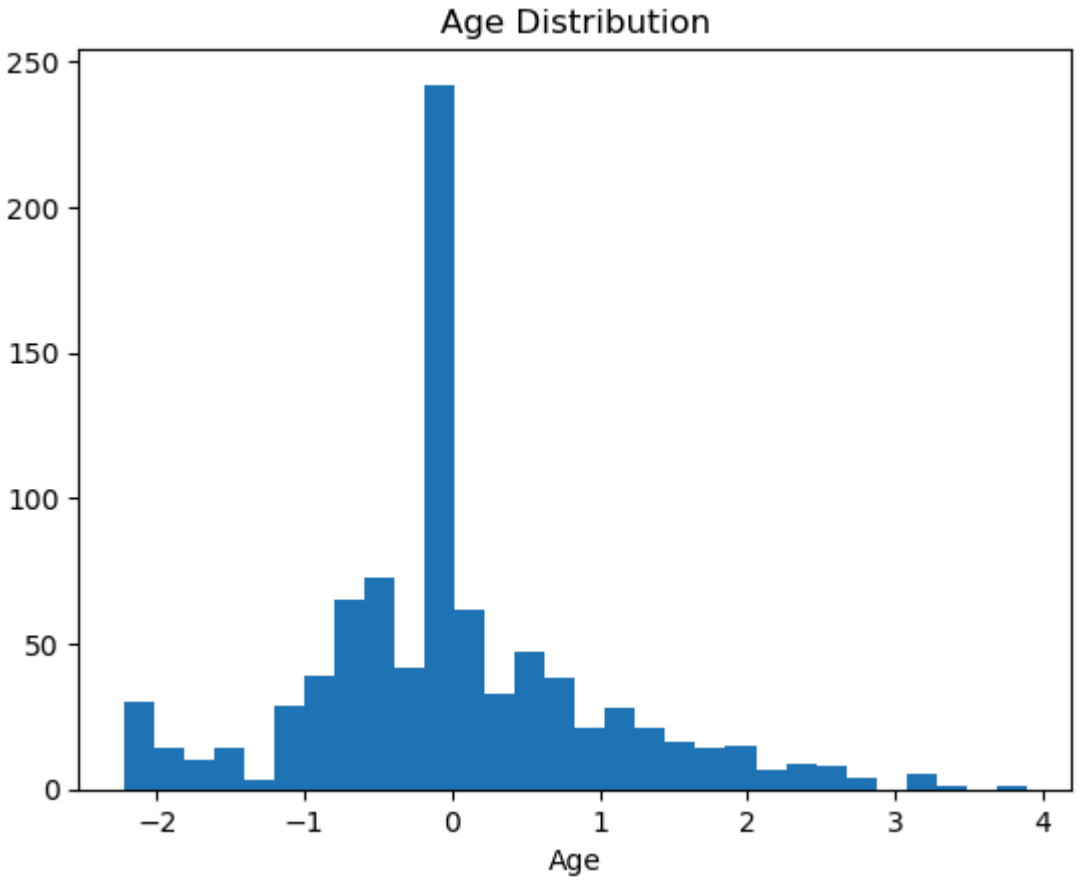
Out[46]:

	Age	Fare
0	-0.565736	-0.502445
1	0.663861	0.786845
2	-0.258337	-0.488854
3	0.433312	0.420730
4	0.433312	-0.486337
...	...	...
886	-0.181487	-0.386671
887	-0.796286	-0.044381
888	-0.104637	-0.176263
889	-0.258337	-0.044381
890	0.202762	-0.492378

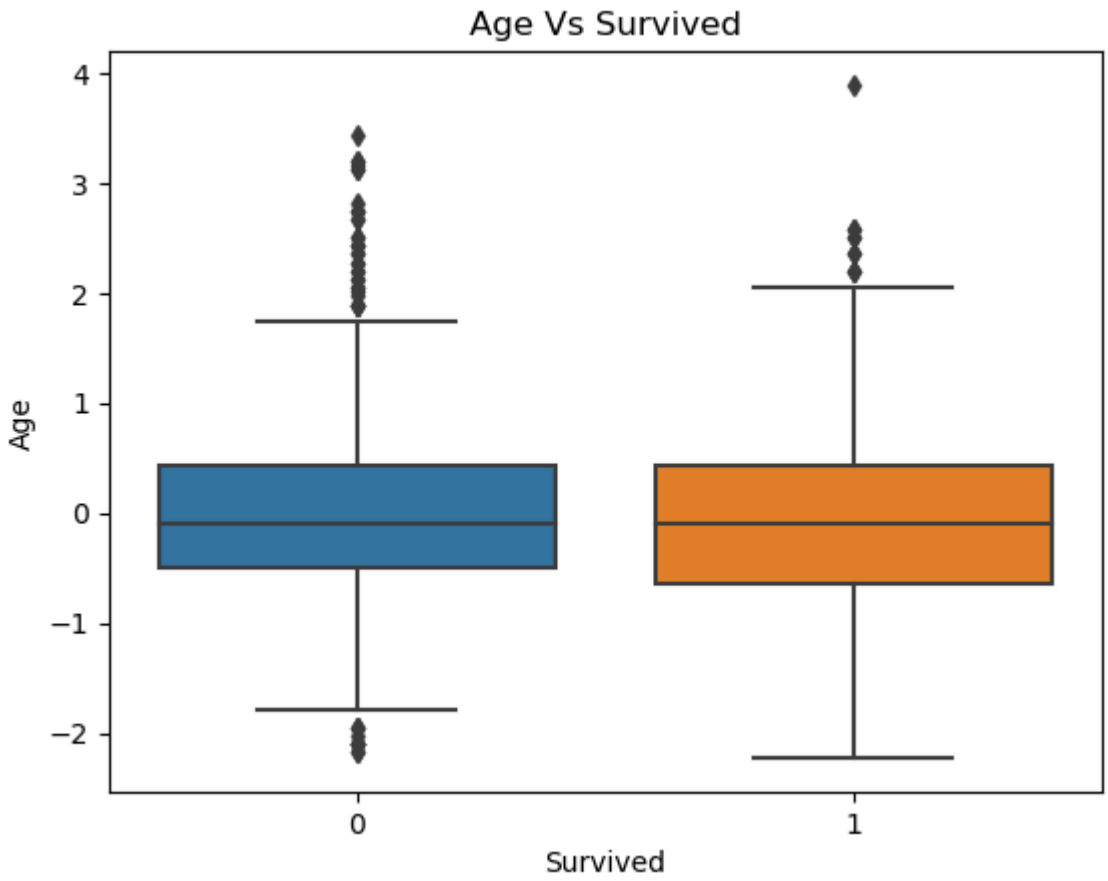
891 rows × 2 columns

In [47]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [50]:
```
plt.hist(data['Age'],bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.show()
```

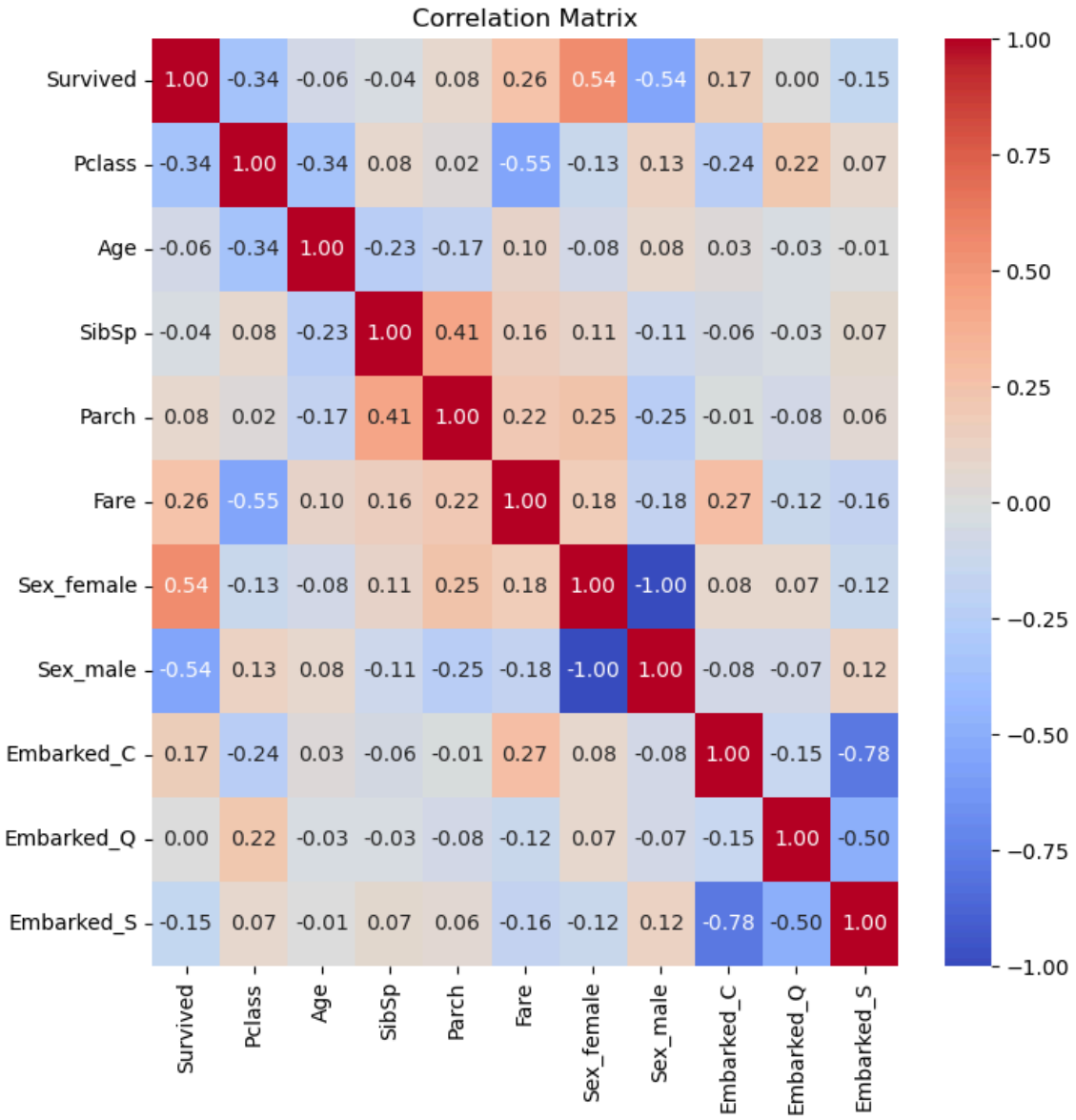


In [54]:
```
sns.boxplot(x='Survived',y='Age',data=data)
plt.title('Age Vs Survived')
plt.show()
```



In [55]:
```
corr_matrix = data.corr()
```

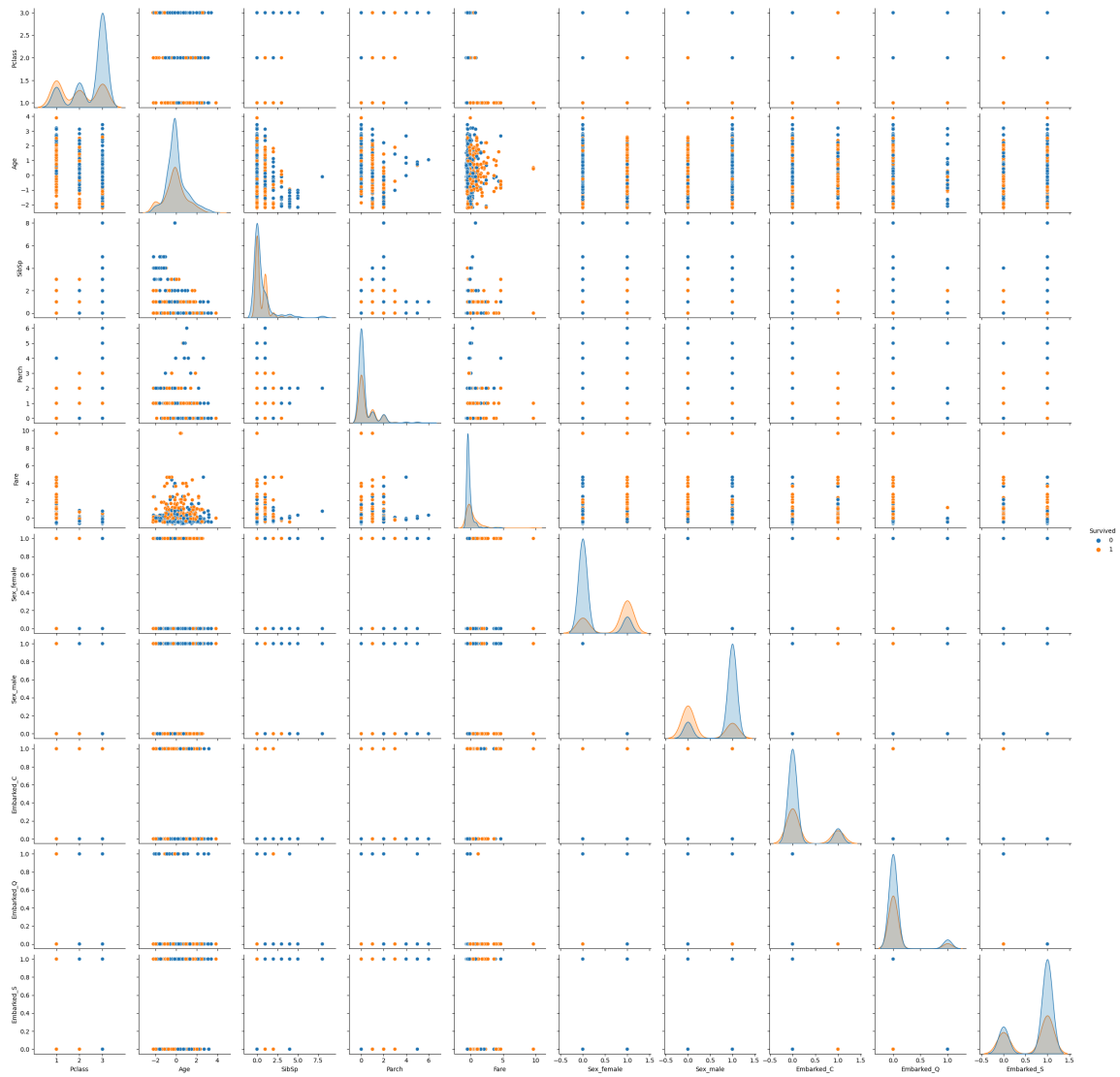
In [66]:
```
plt.figure(figsize=(8,8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm',fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```



Correlation Analysis

In [69]:
```
sns.pairplot(data,hue='Survived',diag_kind='kde')
plt.show()
```

```
C:\Users\lahir\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWa
rning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```



Conclusion

EDA is a crucial step in the data analysis process.

It helps in understanding the underlying patterns and relationships in the data.

Provides a solid foundation for further modeling and analysis.

By Lahiru Sadakelum