In [1]:

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

In [2]:
```
df = pd.read_csv("titanic_train.csv")
df
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	C
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	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	
...	...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	

891 rows × 12 columns

In [3]:
```
df["Age"].fillna(df["Age"].mean(),inplace=True)
```

In [4]:
```
1 df
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.050
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.450
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.000
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.750

891 rows × 12 columns

In [5]:
```
1 df.drop(["Cabin"],axis=1,inplace=True)
```

In [6]:
```
1 df
```

Out[6]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.050
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.450
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.000
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.750

891 rows × 11 columns

In [7]:
```
1 df.drop(["PassengerId"],axis=1,inplace=True)
```

In [8]:
```
1 df
```

Out[8]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	S
1	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C
2	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	S
3	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	S
4	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...
886	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	S
887	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	S
888	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4500	S
889	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C
890	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	Q

891 rows × 10 columns

In [9]:
```
1 df["Embarked"].value_counts()
```

Out[9]:
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [10]:
```
1 df.isnull().sum()
```

Out[10]:
```
Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
Parch       0
Ticket      0
Fare        0
Embarked    2
dtype: int64
```

In [11]:
```
1 df.dropna(subset=["Embarked"], inplace=True)
```

In [12]:
```
1 df
```

Out[12]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	S
1	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	C
2	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	S
3	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	S
4	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...
886	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	S
887	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	S
888	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4500	S
889	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C
890	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	Q

889 rows × 10 columns

In [13]:
```
1 df.isnull().sum()
```

Out[13]:
```
Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
Parch       0
Ticket      0
Fare        0
Embarked    0
dtype: int64
```

In [14]:
```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Survived  889 non-null    int64  
 1   Pclass    889 non-null    int64  
 2   Name      889 non-null    object 
 3   Sex       889 non-null    object 
 4   Age       889 non-null    float64
 5   SibSp     889 non-null    int64  
 6   Parch     889 non-null    int64  
 7   Ticket    889 non-null    object 
 8   Fare      889 non-null    float64
 9   Embarked  889 non-null    object 
dtypes: float64(2), int64(4), object(4)
memory usage: 76.4+ KB
```

In [15]:
```
1 from sklearn.impute import SimpleImputer
```

In [16]:
```
1 si = SimpleImputer(strategy="mean")
```

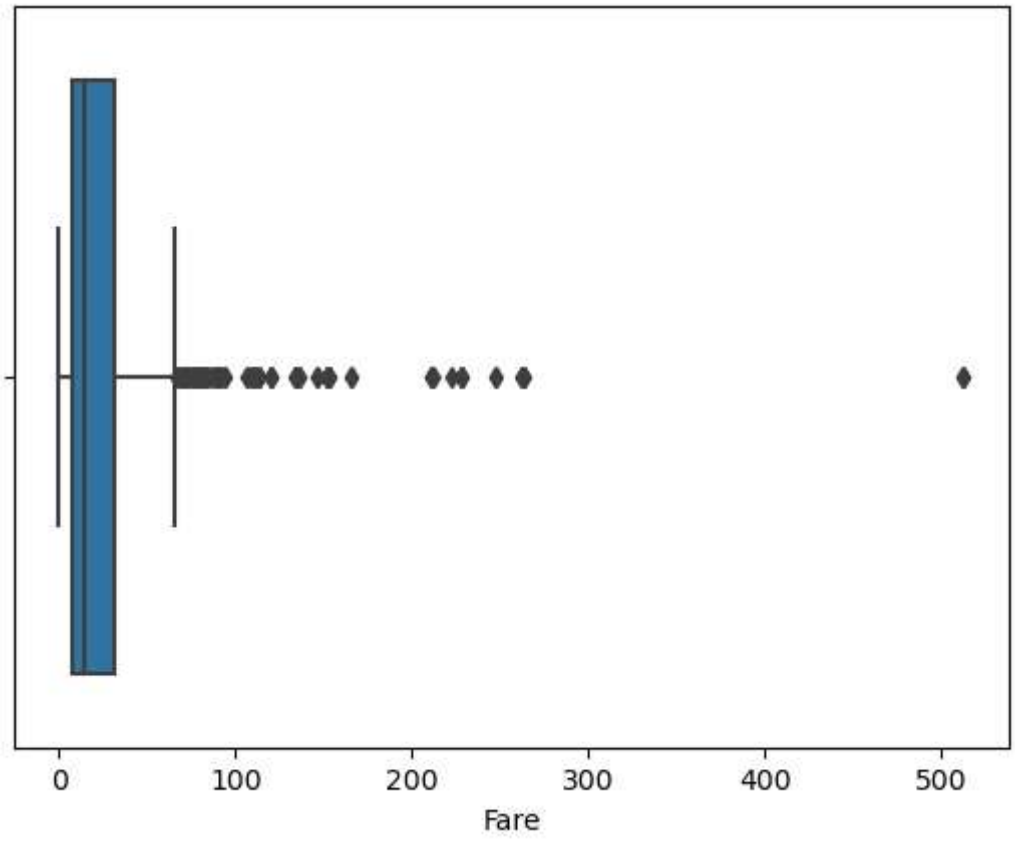
In [17]:
```
1 df[["Survived","Fare"]]=si.fit_transform(df[["Survived","Fare"]])
```

In [18]:
```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 889 entries, 0 to 890
Data columns (total 10 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   Survived  889 non-null    float64
 1   Pclass    889 non-null    int64  
 2   Name      889 non-null    object 
 3   Sex       889 non-null    object 
 4   Age       889 non-null    float64
 5   SibSp     889 non-null    int64  
 6   Parch     889 non-null    int64  
 7   Ticket    889 non-null    object 
 8   Fare      889 non-null    float64
 9   Embarked  889 non-null    object 
dtypes: float64(3), int64(3), object(4)
memory usage: 76.4+ KB
```

In [19]:
```
1 sns.boxplot(data=df,x="Fare")
```

Out[19]: <AxesSubplot:xlabel='Fare'>



In [20]:
```
1 colname=df.select_dtypes(["int64","float64"]).columns
```

In [21]:
```
1 colname
```

Out[21]:
```
Index(['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare'], dtype='object')
```

In [22]:
```
1 df[colname]
```

Out[22]:

	Survived	Pclass	Age	SibSp	Parch	Fare
0	0.0	3	22.000000	1	0	7.2500
1	1.0	1	38.000000	1	0	71.2833
2	1.0	3	26.000000	0	0	7.9250
3	1.0	1	35.000000	1	0	53.1000
4	0.0	3	35.000000	0	0	8.0500
...	...	...	...	...	...	...
886	0.0	2	27.000000	0	0	13.0000
887	1.0	1	19.000000	0	0	30.0000
888	0.0	3	29.699118	1	2	23.4500
889	1.0	1	26.000000	0	0	30.0000
890	0.0	3	32.000000	0	0	7.7500

889 rows × 6 columns

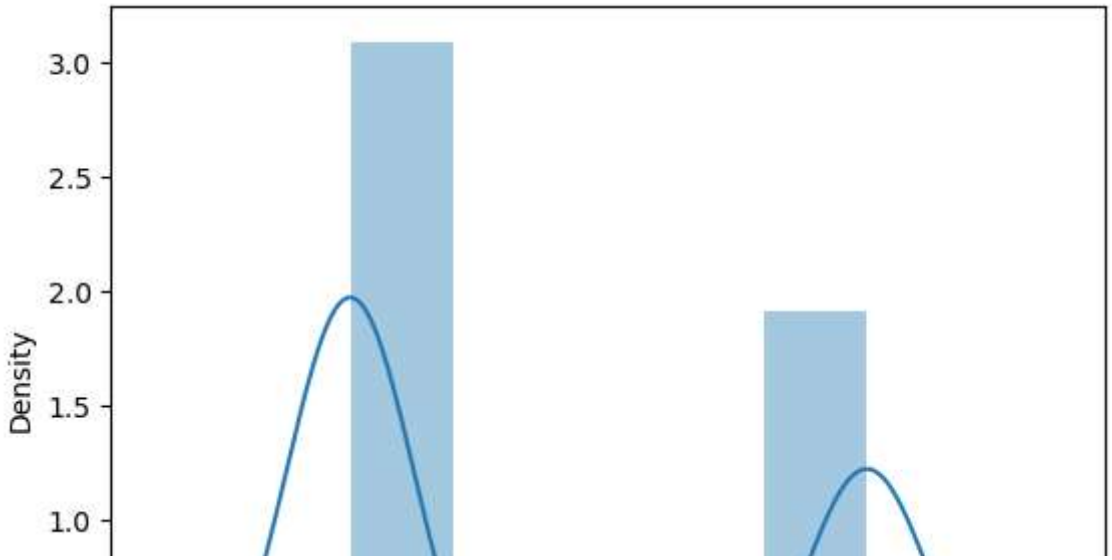
In [23]:
```
1 from scipy.stats import skew
```

In [24]:
```
for col in df[colname]:
    print(col)
    print(skew(df[col]))

    sns.distplot(df[col])
    plt.show()

```

```
Survived
0.4837496405947267
```



In [25]:
```
df.corr()
```

Out[25]:

	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.000000	-0.335549	-0.074673	-0.034040	0.083151	0.255290
Pclass	-0.335549	1.000000	-0.327954	0.081656	0.016824	-0.548193
Age	-0.074673	-0.327954	1.000000	-0.231875	-0.178232	0.088604
SibSp	-0.034040	0.081656	-0.231875	1.000000	0.414542	0.160887
Parch	0.083151	0.016824	-0.178232	0.414542	1.000000	0.217532
Fare	0.255290	-0.548193	0.088604	0.160887	0.217532	1.000000

In [26]:
```
np.log(-5)
```

Out[26]: nan

In [27]:
```
np.sqrt(-3)
```

Out[27]: nan

In [28]:
```
df1=df.corr().style.background_gradient()
```

In [29]:
```
df.corr().style.background_gradient()
```

Out[29]:

	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.000000	-0.335549	-0.074673	-0.034040	0.083151	0.255290
Pclass	-0.335549	1.000000	-0.327954	0.081656	0.016824	-0.548193
Age	-0.074673	-0.327954	1.000000	-0.231875	-0.178232	0.088604
SibSp	-0.034040	0.081656	-0.231875	1.000000	0.414542	0.160887
Parch	0.083151	0.016824	-0.178232	0.414542	1.000000	0.217532
Fare	0.255290	-0.548193	0.088604	0.160887	0.217532	1.000000

In [30]:
```
skew(df["Age"])
```

Out[30]: 0.4309914863386608

In [31]:
```
df["Age"]=np.log(df["Age"])
```

In [32]:
```
skew(df["Age"])
```

Out[32]: -2.6798307790180864

In [33]:
```
df.corr().style.background_gradient()
```

Out[33]:

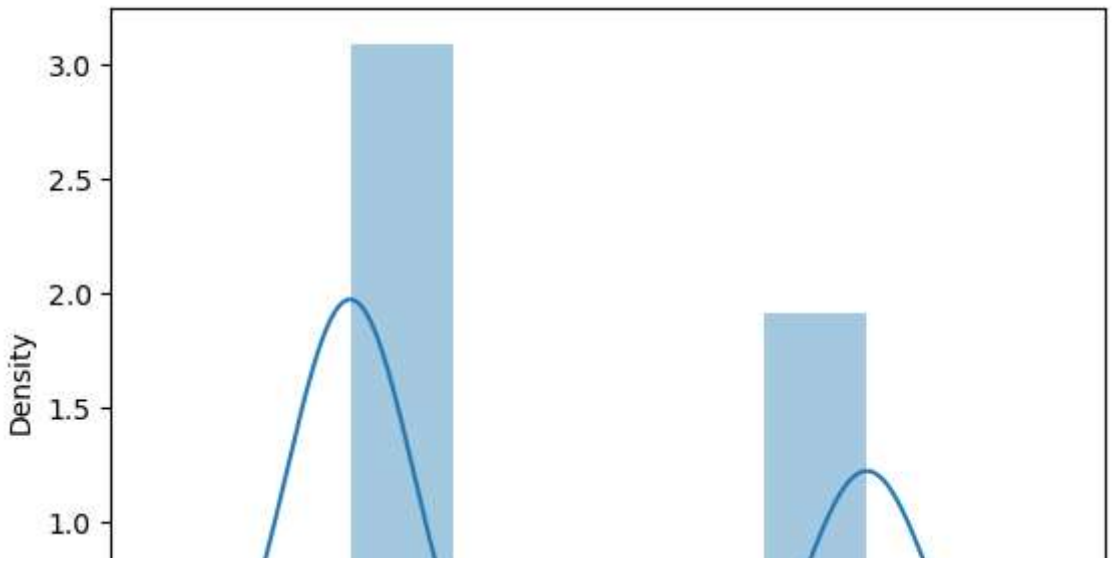
	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.000000	-0.335549	-0.128040	-0.034040	0.083151	0.255290
Pclass	-0.335549	1.000000	-0.218503	0.081656	0.016824	-0.548193
Age	-0.128040	-0.218503	1.000000	-0.290044	-0.287102	0.045438
SibSp	-0.034040	0.081656	-0.290044	1.000000	0.414542	0.160887
Parch	0.083151	0.016824	-0.287102	0.414542	1.000000	0.217532
Fare	0.255290	-0.548193	0.045438	0.160887	0.217532	1.000000

In [34]:
```
for col in df[colname]:
    print(col)
    print(skew(df[col]))

    sns.distplot(df[col])
    plt.show()

```

```
Survived
0.4837496405947267
```



In [35]:
```
df.head()
```

Out[35]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0.0	3	Braund, Mr. Owen Harris	male	3.091042	1	0	A/5 21171	7.2500	S
1	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	3.637586	1	0	PC 17599	71.2833	C
2	1.0	3	Heikkinen, Miss. Laina	female	3.258097	0	0	STON/O2. 3101282	7.9250	S
3	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	3.555348	1	0	113803	53.1000	S
4	0.0	3	Allen, Mr. William Henry	male	3.555348	0	0	373450	8.0500	S

In [36]:
```
from sklearn.preprocessing import OrdinalEncoder
```

In [37]:
```
oe = OrdinalEncoder()
```

In [38]:
```
1 oe.fit_transform(df[["Sex","Embarked"]])
```

Out[38]:
```
array([[1., 2.],
       [0., 0.],
       [0., 2.],
       ...,
       [0., 2.],
       [1., 0.],
       [1., 1.]])
```

In [39]:
```
df
```

Out[39]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0.0	3	Braund, Mr. Owen Harris	male	3.091042	1	0	A/5 21171	7.2500	S
1	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	3.637586	1	0	PC 17599	71.2833	C
2	1.0	3	Heikkinen, Miss. Laina	female	3.258097	0	0	STON/O2. 3101282	7.9250	S
3	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	3.555348	1	0	113803	53.1000	S
4	0.0	3	Allen, Mr. William Henry	male	3.555348	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...
886	0.0	2	Montvila, Rev. Juozas	male	3.295837	0	0	211536	13.0000	S
887	1.0	1	Graham, Miss. Margaret Edith	female	2.944439	0	0	112053	30.0000	S
888	0.0	3	Johnston, Miss. Catherine Helen "Carrie"	female	3.391117	1	2	W./C. 6607	23.4500	S
889	1.0	1	Behr, Mr. Karl Howell	male	3.258097	0	0	111369	30.0000	C
890	0.0	3	Dooley, Mr. Patrick	male	3.465736	0	0	370376	7.7500	Q

889 rows × 10 columns

In [40]:
```
catcol = df.select_dtypes(object).columns
```

In [41]:
```
from sklearn.preprocessing import OrdinalEncoder
oe = OrdinalEncoder()
df[catcol]=oe.fit_transform(df[catcol])
```

In [42]:
```
1 df
```

Out[42]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	0.0	3	108.0	1.0	3.091042	1	0	522.0	7.2500	2.0
1	1.0	1	190.0	0.0	3.637586	1	0	595.0	71.2833	0.0
2	1.0	3	353.0	0.0	3.258097	0	0	668.0	7.9250	2.0
3	1.0	1	272.0	0.0	3.555348	1	0	48.0	53.1000	2.0
4	0.0	3	15.0	1.0	3.555348	0	0	471.0	8.0500	2.0
...	...	...	...	...	...	...	...	...	...	...
886	0.0	2	547.0	1.0	3.295837	0	0	100.0	13.0000	2.0
887	1.0	1	303.0	0.0	2.944439	0	0	14.0	30.0000	2.0
888	0.0	3	412.0	0.0	3.391117	1	2	674.0	23.4500	2.0
889	1.0	1	81.0	1.0	3.258097	0	0	8.0	30.0000	0.0
890	0.0	3	220.0	1.0	3.465736	0	0	465.0	7.7500	1.0

889 rows × 10 columns

In [43]:
```
1 from sklearn.preprocessing import StandardScaler
```

In [44]:
```
1 ss = StandardScaler()
```

In [45]:
```
1 df.iloc[:,:-1]=ss.fit_transform(df.iloc[:,:-1])
```

In [46]:
```
df
```

Out[46]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	-0.786961	0.825209	-1.309268	0.735342	-0.201216	0.431350	-0.474326	0.917018	-0.50024
1	1.270711	-1.572211	-0.989744	-1.359911	0.572664	0.431350	-0.474326	1.281353	0.78894
2	1.270711	0.825209	-0.354593	-1.359911	0.035325	-0.475199	-0.474326	1.645689	-0.48665
3	1.270711	-1.572211	-0.670220	-1.359911	0.456219	0.431350	-0.474326	-1.448669	0.42286
4	-0.786961	0.825209	-1.671654	0.735342	0.456219	-0.475199	-0.474326	0.662482	-0.48413
...	...	...	...	...	...	...	...	...	
886	-0.786961	-0.373501	0.401353	0.735342	0.088763	-0.475199	-0.474326	-1.189142	-0.38447
887	1.270711	-1.572211	-0.549425	-1.359911	-0.408799	-0.475199	-0.474326	-1.618360	-0.04221
888	-0.786961	0.825209	-0.124692	-1.359911	0.223676	0.431350	2.006119	1.675635	-0.17408
889	1.270711	-1.572211	-1.414477	0.735342	0.035325	-0.475199	-0.474326	-1.648305	-0.04221
890	-0.786961	0.825209	-0.872845	0.735342	0.329332	-0.475199	-0.474326	0.632536	-0.49017

889 rows × 10 columns

In [47]:
```
x = df.iloc[:,:-1]
y = df.iloc[:,-1]
```

In [48]:
```
x
```

Out[48]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	-0.786961	0.825209	-1.309268	0.735342	-0.201216	0.431350	-0.474326	0.917018	-0.50024
1	1.270711	-1.572211	-0.989744	-1.359911	0.572664	0.431350	-0.474326	1.281353	0.78894
2	1.270711	0.825209	-0.354593	-1.359911	0.035325	-0.475199	-0.474326	1.645689	-0.48665
3	1.270711	-1.572211	-0.670220	-1.359911	0.456219	0.431350	-0.474326	-1.448669	0.42286
4	-0.786961	0.825209	-1.671654	0.735342	0.456219	-0.475199	-0.474326	0.662482	-0.48413
...	...	...	...	...	...	...	...	...	
886	-0.786961	-0.373501	0.401353	0.735342	0.088763	-0.475199	-0.474326	-1.189142	-0.38447
887	1.270711	-1.572211	-0.549425	-1.359911	-0.408799	-0.475199	-0.474326	-1.618360	-0.04221
888	-0.786961	0.825209	-0.124692	-1.359911	0.223676	0.431350	2.006119	1.675635	-0.17408
889	1.270711	-1.572211	-1.414477	0.735342	0.035325	-0.475199	-0.474326	-1.648305	-0.04221
890	-0.786961	0.825209	-0.872845	0.735342	0.329332	-0.475199	-0.474326	0.632536	-0.49017

889 rows × 9 columns

In [49]:
```
1 y
```

Out[49]:
```
0      2.0
1      0.0
2      2.0
3      2.0
4      2.0
      ...
886    2.0
887    2.0
888    2.0
889    0.0
890    1.0
Name: Embarked, Length: 889, dtype: float64
```

In [50]:
```
1 x.shape
```

Out[50]:
```
(889, 9)
```

In [51]:
```
1 y.shape
```

Out[51]:
```
(889,)
```

In [52]:
```
1 from sklearn.model_selection import train_test_split
2 xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.3,random_state=1)
```

In [53]:
```
1 xtrain
```

Out[53]:

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
115	-0.786961	0.825209	0.779326	0.735342	-0.267086	-0.475199	-0.474326	1.630716	-0.48665
874	1.270711	-0.373501	-1.714517	-1.359911	0.140258	0.431350	-0.474326	1.176544	-0.1630
77	-0.786961	0.825209	0.459802	0.735342	0.223676	-0.475199	-0.474326	0.667473	-0.48413
876	-0.786961	0.825209	-0.518252	0.735342	-0.336171	-0.475199	-0.474326	0.822191	-0.44797
682	-0.786961	0.825209	0.662427	0.735342	-0.336171	-0.475199	-0.474326	0.802227	-0.46047
...	...	...	...	...	...	...	...	...	
716	1.270711	-1.572211	-0.791016	-1.359911	0.572664	-0.475199	-0.474326	1.351226	3.93457
768	-0.786961	0.825209	0.420836	0.735342	0.223676	0.431350	-0.474326	0.647509	-0.15999
73	-0.786961	0.825209	-1.102746	0.735342	0.035325	0.431350	-0.474326	-0.675080	-0.35519
236	-0.786961	-0.373501	-0.288351	0.735342	0.780248	0.431350	-0.474326	-0.705025	-0.12274
37	-0.786961	0.825209	-1.207955	0.735342	-0.267086	-0.475199	-0.474326	0.872100	-0.48413

622 rows × 9 columns

In [54]:
```
1 ytrain
```

Out[54]:
```
115    2.0
874    0.0
77     2.0
876    2.0
682    2.0
      ...
716    0.0
768    1.0
73     0.0
236    2.0
37     2.0
Name: Embarked, Length: 622, dtype: float64
```

In [ ]:
```
1
```