

ASWINI- 21BEC7252

Assignment-3

1 Importing Libraries

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler,
LabelEncoder
from sklearn.preprocessing import
MinMaxScaler
```

2 Importing the dataset

```
[2]: dataset=pd.read_csv("Titanic-Dataset.csv")
```

```
[3]: dataset
```

```
[3]: PassengerId Survived Pclass \
0      1      0      3
1      2      1      1
2      3      1      3
3      4      1      1
4      5      0      3
..      ...      ...      ...
886     887      0      2
887     888      1      1
888     889      0      3
889     890      1      1
890     891      0      3
```

```
Name Sex Age SibSp \0 Braund, Mr. Owen Harris male 22.0
```

defined.

```
1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0
```

Error! Bookmark not

2

```
3 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 2
```

```
2           Heikkinen, Miss. Laina female 26.0 0
4           Allen, Mr. William Henry male 35.0 0
..           ... ..
886          Montvila, Rev. Juozas male 27.0 0
887          Graham, Miss. Margaret Edith female
19.0 0
888          Johnston, Miss. Catherine Helen
"Carrie" female NaN 1
889          Behr, Mr. Karl Howell male 26.0 0
890          Dooley, Mr. Patrick male 32.0 0
```

```
      Parch      Ticket    Fare Cabin Embarked
0         0   A/5 21171  7.2500    NaN    S
1         0   PC 17599 71.2833 C85    C
2         0 STON/O2. 3101282  7.9250    NaN    S
3         0  113803 53.1000 C123    S
4         0  373450   8.0500    NaN    S
..      ...      ...      ...      ...
886        0  211536 13.0000    NaN    S
887        0  112053 30.0000   B42    S
888        2   W./C. 6607 23.4500    NaN    S
889        0  111369 30.0000 C148    C
890        0  370376   7.7500    NaN    Q
```

```
[891 rows x 12 columns]
```

```
[4]: dataset.head()
```

```
[4]: PassengerId Survived Pclass \
```

```
0         1         0         3
1         2         1         1
2         3         1         3
3         4         1         1
4         5         0         3
```

```
      Name      Sex  Age SibSp \
0      Braund, Mr. Owen Harris male 22.0 1
1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1
2      Heikkinen, Miss. Laina female 26.0 0
3 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1
4      Allen, Mr. William Henry male 35.0 0
```

	Parch	Ticket	Fare	Cabin	Embarked
0	0	A/5 21171	7.2500	NaN	S
1	0	PC 17599	71.2833	C85	C
2	0	STON/O2. 3101282	7.9250	NaN	S
3	0	113803	53.1000	C123	S
4	0	373450	8.0500	NaN	S

```
[5]: dataset.tail()
```

```
[5]:      PassengerId  Survived  Pclass                    Name \
886           887     0         2    Montvila, Rev. Juozas
887           888     1         1    Graham, Miss. Margaret Edith
888           889     0         3  Johnston, Miss. Catherine Helen "Carrie"
889           890     1         1    Behr, Mr. Karl Howell
890           891     0         3    Dooley, Mr. Patrick
```

	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
886	male	27.0	0	0	211536	13.00	NaN	S
887	female	19.0	0	0	112053	30.00	B42	S
888	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	male	26.0	0	0	111369	30.00	C148	C
890	male	32.0	0	0	370376	7.75	NaN	Q

```
[6]: dataset.shape
```

```
[6]: (891, 12)
```

```
[7]: dataset.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
```

```
[8]: dataset.describe()
```

```
[8]:      PassengerId  Survived  Pclass    Age  SibSp \
count    891.000000    891.000000    891.000000   714.000000
      891.000000
mean     446.000000     0.383838     2.308642   29.699118   0.523008
std      257.353842     0.486592     0.836071   14.526497   1.102743
min        1.000000     0.000000     1.000000    0.420000   0.000000
25%      223.500000     0.000000     2.000000   20.125000   0.000000
50%      446.000000     0.000000     3.000000   28.000000   0.000000
75%      668.500000     1.000000     3.000000   38.000000   1.000000
max      891.000000     1.000000     3.000000   80.000000   8.000000

      Parch    Fare
count    891.000000    891.000000
mean       0.381594   32.204208
std        0.806057   49.693429
min         0.000000    0.000000
25%         0.000000    7.910400
50%         0.000000   14.454200
75%         0.000000   31.000000
max         6.000000
      512.329200
```

3 Checking for Null Values

```
[9]: dataset.isnull().any()
```

```
[9]: PassengerId  False
     Survived    False
     Pclass      False
     Name        False
     Sex         False
     Age         True
     SibSp       False
     Parch       False
     Ticket      False
     Fare        False
     Cabin       True
     Embarked    True
     dtype: bool
```

```
[10]: dataset.isnull().sum()
```

```
[10]: 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
```

4 Handling the null value

```
[11]: dataset['Age'].fillna(dataset['Age'].median(), inplace=True)
```

```
[12]: dataset.drop(['Cabin'], axis=1, inplace=True)
```

```
[13]: mode_embarked = dataset['Embarked'].mode()[0]
dataset['Embarked'].fillna(mode_embarked, inplace=True)
```

```
[14]: dataset.isnull().sum()
```

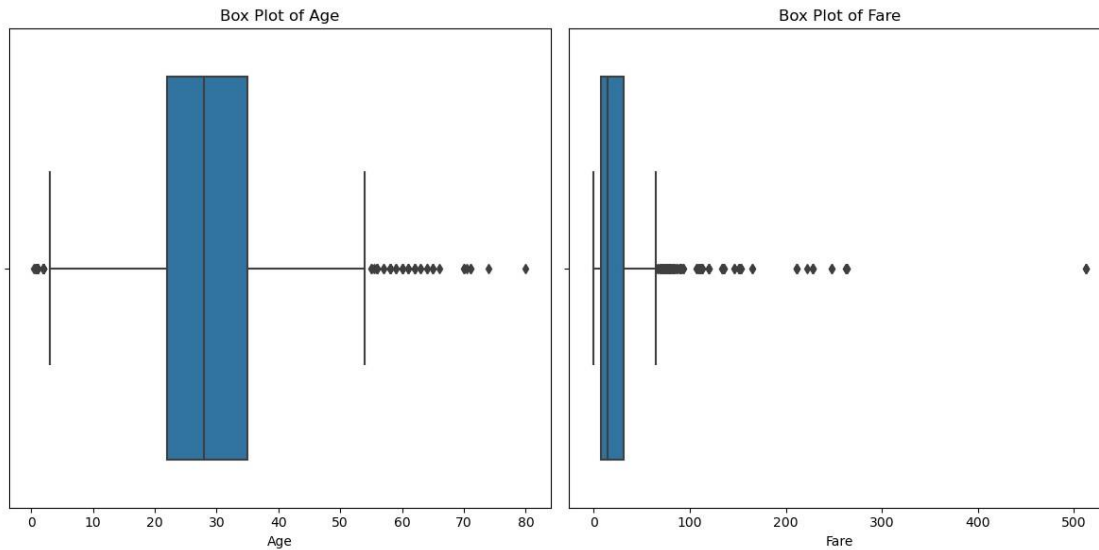
```
[14]: PassengerId    0
Survived          0
Pclass            0
Name              0
Sex               0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         0
dtype: int64
```

5 Data Visualization

```
[15]: plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
sns.boxplot(data=dataset, x="Age")
plt.title('Box Plot of Age ')

plt.subplot(1, 2, 2)
sns.boxplot(data=dataset, x="Fare")
plt.title('Box Plot of Fare ')

plt.tight_layout()
plt.show()
```



6 Outlier Detection:-

```
[16]: Q1_age = dataset['Age'].quantile(0.25)
      Q3_age = dataset['Age'].quantile(0.75)
      IQR_age = Q3_age - Q1_age
      lower_bound_age = Q1_age - 1.5 * IQR_age
      upper_bound_age = Q3_age + 1.5 * IQR_age

      outliers_age = dataset[(dataset['Age'] < lower_bound_age) | (dataset['Age'] >
      ↪upper_bound_age)]
```

```
[17]: outliers_age
```

```
[17]:
```

	PassengerId	Survived	Pclass	\
7	8	0	3	
11	12	1	1	
15	16	1	2	
16	17	0	3	
33	34	0	2	
..	
827	828	1	2	
829	830	1	1	
831	832	1	2	

```

851      852      0      3
879      880      1      1

```

```

                                Name    Sex   Age SibSp \
7                                Palsson, Master. Gosta   male  2.00    3
                                Leonard
11 Bonnell, Miss. Elizabeth female 58.00 0 15 Hewlett, Mrs.
(Mary D Kingcome) female 55.00 0
16 Rice, Master. Eugene male 2.00 4 33 Wheadon, Mr. Edward H
male 66.00 0
..
827                                Mallet, Master. Andre   male  1.00    0
829  Stone, Mrs. George Nelson (Martha Evelyn) female 62.00    0
831                                Richards, Master. George Sibley male  0.83    1
851                                Svensson, Mr. Johan   male 74.00    0
879 Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) female 56.00    0

```

```

      Parch      Ticket    Fare Embarked
7         1      349909 21.0750        S
11        0      113783 26.5500        S
15        0  248706 16.0000    S
16        1  382652 29.1250    Q
33        0      C.A. 24579 10.5000        S
..      ...      ...      ...      ...
827      2 S.C./PARIS 2079 37.0042        C
829      0      113572 80.0000        S
831      1      29106 18.7500        S
851      0      347060 7.7750        S
879      1      11767 83.1583        C

```

[66 rows x 11 columns]

```

[18]: Q1_fare = dataset['Fare'].quantile(0.25)
      Q3_fare = dataset['Fare'].quantile(0.75)
      IQR_fare = Q3_fare - Q1_fare
      lower_bound_fare = Q1_fare - 1.5 * IQR_fare
      upper_bound_fare = Q3_fare + 1.5 * IQR_fare

      outliers_fare = dataset[(dataset['Fare'] < lower_bound_fare) | (dataset['Fare']_
      ↪ > upper_bound_fare)]

```

```

[19]: outliers_fare

```

```

[19]: PassengerId  Survived  Pclass  \
1             2         1         1
27            28         0         1

```

31	32	1	1
34	35	0	1
52	53	1	1
..
846	847	0	3
849	850	1	1
856	857	1	1
863	864	0	3
879	880	1	1

	Name	Sex	Age	SibSp	\
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	
27	Fortune, Mr. Charles Alexander	male	19.0	3	
31	Spencer, Mrs. William Augustus (Marie Eugenie)	female	28.0	1	
34	Meyer, Mr. Edgar Joseph	male	28.0	1	
52	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	
..	
846	Sage, Mr. Douglas Bullen	male	28.0	8	
849	Goldenberg, Mrs. Samuel L (Edwiga Grabowska)	female	28.0	1	856
Wick,	Mrs. George Dennick (Mary Hitchcock)	female	45.0	1	863 Sage,
Miss.	Dorothy Edith "Dolly"	female	28.0	8	
879	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	

	Parch	Ticket	Fare	Embarked
1	0	PC 17599	71.2833	C
27	2	19950	263.0000	S
31	0	PC 17569	146.5208	C
34	0	PC 17604	82.1708	C
52	0	PC 17572	76.7292	C
..
846	2	CA. 2343	69.5500	S
849	0	17453	89.1042	C
856	1	36928	164.8667	S
863	2	CA. 2343	69.5500	S
879	1	11767	83.1583	C

[116 rows x 11 columns]

7 Splitting Dependent and Independent variables


```
[20]: x = dataset.drop('Survived', axis=1)
      y = dataset['Survived']
```

```
[21]: x
```

```
[21]:
```

	PassengerId	Pclass	Name \					
0	1	3	Braund, Mr. Owen Harris					
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...					
2	3	3	Heikkinen, Miss. Laina					
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)					
4	5	3	Allen, Mr. William Henry					
..					
886	887	2	Montvila, Rev. Juozas					
887	888	1	Graham, Miss. Margaret Edith					
888	889	3	Johnston, Miss. Catherine Helen "Carrie"					
889	890	1	Behr, Mr. Karl Howell					
890	891	3	Dooley, Mr. Patrick					
	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	
0	male	22.0	1	0	A/5 21171	7.2500	S	
1	female	38.0	1	0	PC 17599	71.2833	C	
2	female	26.0	0	0	STON/O2. 3101282	7.9250	S	
3	female	35.0	1	0	113803	53.1000	S	
4	male	35.0	0	0	373450	8.0500	S	
..	
	886	male	27.0	0	0	211536	13.0000	S
	887	female	19.00	0	0	112053	30.0000	S
	888	female	28.01	2	W./C. 6607	23.4500	S	
	889	male	26.0	0	0	111369	30.0000	C
	890	male	32.0	0	0	370376	7.7500	Q
	891	rows x 10 columns]						

```
[22]:
```

```
[22]: 0 0
```

```

1      1
2      1
3      1
4      0
..
886    0
887    1
888    0
889    1
890    0
Name: Survived, Length: 891, dtype: int64

```

8 Encoding

8.0.1 • Label encoding on Pclass Column :-

```
[23]: le=LabelEncoder()
```

```
[24]: x["Pclass"]=le.fit_transform(x["Pclass"])
```

```
[25]: x["Pclass"]
```

```

[25]: 0      2
      1      0
      2      2
      3      0
      4      2
      ..
      886    1
      887    0
      888    2
      889    0
      890    2
Name: Pclass, Length: 891, dtype: int64

```

```
[26]: x["Pclass"].value_counts()
```

```

[26]: 2
      491 0
      216
      1   184
Name: Pclass, dtype: int64

```

```
[27]: x["Pclass"].nunique()
```

```
[27]: 3
```

```
[28]: x.head()
```

```
[28]: PassengerId Pclass                               Name \
0          1      2      Braund, Mr. Owen Harris
1          2      0 Cumings, Mrs. John Bradley (Florence Briggs Th...
2          3      2      Heikkinen, Miss. Laina
3          4      0 Futrelle, Mrs. Jacques Heath (Lily May Peel)
4          5      2      Allen, Mr. William Henry
```

```
      Sex Age SibSp Parch      Ticket     Fare Embarked
0  male 22.0   1     0   A/5 21171   7.2500      S
1  female 38.0  1     0   PC 17599 71.2833 C
2  female 26.0  0     0 STON/O2. 3101282   7.9250      S
3  female 35.0  1     0   113803 53.1000      S
4  male 35.0   0     0   373450   8.0500      S
```

8.0.2 • One hot encoding on Sex and Embarked Column :-

```
[29]: x.shape
```

```
[29]: (891, 10)
```

```
[31]:
```

```
[30]: Sex = pd.get_dummies(x['Sex'], drop_first=False)
```

```
Sex [31]: female male
0      0 1
1      1 0
2      1 0
3      1 0
4      0 1
..      ... ..
886     0 1
887     1 0
888     1 0
889     0 1
```

```
890      0 1
```

```
[891 rows x 2 columns]
```

```
[32]: Embarked = pd.get_dummies(x["Embarked"],drop_first=False)
```

```
[33]: Embarked
```

```
[33]:      C  Q  S
```

```
0      0  0  1
```

```
1      1  0  0
```

```
2      0  0  1
```

```
3      0  0  1
```

```
4      0  0  1
```

```
... ..
```

```
886    0  0  1
```

```
887    0  0  1
```

```
888    0  0  1
```

```
889    1  0  0
```

```
890    0  1  0
```

```
[891 rows x 3 columns]
```

```
[34]: x=pd.concat([x,Sex,Embarked],axis=1)
```

```
[35]: x.head()
```

```
[35]: PassengerId  Pclass                                     Name \
```

```
0              1      2      Braund, Mr. Owen Harris
```

```
1              2      0  Cumings, Mrs. John Bradley (Florence Briggs Th...
```

```
2              3      2      Heikkinen, Miss. Laina
```

```
3              4      0  Futrelle, Mrs. Jacques Heath (Lily May Peel)
```

```
4              5      2      Allen, Mr. William Henry
```

```
      Sex Age SibSp Parch      Ticket     Fare Embarked female \
0  male 22.0   1     0    A/5 21171  7.2500     S      0
```

```
1  female 38.0   1     0    PC 17599 71.2833 C      1
```

```
2  female 26.0   0     0  STON/O2. 3101282    7.9250     S      1
```

```
3  female 35.0   1     0    113803 53.1000     S      1
```

```
4  male 35.0   0     0    373450    8.0500     S      0
```

```
      male C  Q  S
```

```
0      1  0  0  1
```

```
1      0  1  0  0
```

```
2      0  0  0  1
```

```
3      0  0  0  1
```

```
4      1  0  0  1
```

```
[36]: x.drop(["Sex", "Embarked"], axis=1, inplace=True)
```

```
[37]: x.head(10)
```

```
[37]: PassengerId  Pclass                                Name \
0             1      2      Braund, Mr. Owen Harris
1             2      0  Cumings, Mrs. John Bradley (Florence Briggs Th...
2             3      2      Heikkinen, Miss. Laina
3             4      0  Futrelle, Mrs. Jacques Heath (Lily May Peel)
4             5      2      Allen, Mr. William Henry
5             6      2      Moran, Mr. James
6             7      0  McCarthy, Mr. Timothy J
7             8      2      Palsson, Master. Gosta Leonard
8             9      2  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
9            10      1      Nasser, Mrs. Nicholas (Adele Achem)
```

	Age	SibSp	Parch	Ticket	Fare	female	male	C	Q	S
0	22.0	1	0	A/5 21171 7.2500	0	1	0	0	1	
1	38.0	1	0	PC 17599 71.2833	1	0	1	0	0	
2	26.0	0	0	STON/O2. 3101282	7.9250	1	0	0	0	1
3	35.0	1	0	113803 53.1000	1	0	0	0	1	
4	35.0	0	0	373450 8.0500	0	1	0	0	1	
5	28.0	0	0	330877 8.4583	0	1	0	1	0	
6	54.0	0	0	17463 51.8625	0	1	0	0	1	
7	2.0	3	1	349909 21.0750	0	1	0	0	1	
8	27.0	0	2	347742 11.1333	1	0	0	0	1	
9	14.0	1	0	237736 30.0708	1	0	1	0	0	

```
[38]: x.shape
```

```
[38]: (891, 13)
```

9 Feature Scaling

9.0.1 • Normalization:-

```
[39]: scaler = MinMaxScaler()
x[['Age', 'Fare']] = scaler.fit_transform(x[['Age', 'Fare']])
```

```
[40]: x[['Age', 'Fare']]
```

```
[40]:      Age      Fare
0      0.271174
0.014151 1 0.472229
0.139136 2 0.321438
0.015469 3 0.434531
0.103644
4      0.434531 0.015713
..      ...      ...
```

```

886 0.334004 0.025374
887 0.233476 0.058556
888 0.346569 0.045771
889 0.321438 0.058556
890 0.396833 0.015127

```

```
[891 rows x 2 columns]
```

9.0.2 • Standardization :-

```
[41]: scaler = StandardScaler()
      x[['Age', 'Fare']] = scaler.fit_transform(x[['Age', 'Fare']])
```

```
[42]: x[['Age', 'Fare']]
```

```
[42]:      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 x 2 columns]
```

10 Splitting Data into Train and Test

```
[43]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,
      random_state=42)
```

```
[44]: x_train, x_test, y_train, y_test
```

```
[44]: (      PassengerId  Pclass                                Name      Age  SibSp
\
331          332         0                Partner, Mr. Austen 1.240235      0
733          734         1      Berriman, Mr. William John -0.488887      0
382          383         2                Tikkanen, Mr. Juho 0.202762      0
704          705         2      Hansen, Mr. Henrik Juul -0.258337      1
813          814         2  Andersson, Miss. Ebba Iris Alfrida -1.795334      4
..          ...      ...
106          107         2  Salkjelsvik, Miss. Anna Kristine -0.642586      0
270          271         0      Cairns, Mr. Alexander -0.104637      0
860          861         2      Hansen, Mr. Claus Peter 0.894411      2
435          436         0      Carter, Miss. Lucile Polk -1.180535      1

```

102 103 0 White, Mr. Richard Frasar -0.642586 0

Parch Ticket Fare female male C Q S 331 0 113043 -
0.074583 0 1 0 0 1
733 0 28425 -0.386671 0 1 0 0 1
382 0 STON/O 2. 3101293 -0.488854 0 1 0 0 1
704 0 350025 -0.490280 0 1 0 0 1
813 2 347082 -0.018709 1 0 0 0 1
..
106 0 343120 -0.494391 1 0 0 0 1
270 0 113798 -0.024246 0 1 0 0 1
860 0 350026 -0.364355 0 1 0 0 1
435 2 113760 1.767741 1 0 0 0 1
102 1 35281 0.907738 0 1 0 0 1

[712 rows x 13 columns],

	PassengerId	Pclass	Name \
709	710	2	Moubarek, Master. Halim Gonios ("William George")
439	440	1	Kvillner, Mr. Johan Henrik Johannesson
840	841	2	Alhomaki, Mr. Ilmari Rudolf
	720	721	1 Harper, Miss. Annie
			Jessie "Nina"
39	40	2	Nicola-Yarred, Miss. Jamila
..
433	434	2	Kallio, Mr. Nikolai Erland
773	774	2	Elias, Mr. Dibo
25	26	2	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...
84	85	1	Ilett, Miss. Bertha
10	11	2	Sandstrom, Miss. Marguerite Rut

	Age	SibSp	Parch	Ticket	Fare	female	male	C	Q	\	709	-
0.104637	1	1	2661	-0.341452	0	1	1	0				
439	0.125912	0	0	C.A. 18723	-0.437007	0	1	0	0			
840	-0.719436	0	0	SOTON/O2 3101287	-0.488854	0	1	0	0			
720	-1.795334	0	1	248727 0.016023	1	0	0	0	39	-1.180535		
	1	0	2651	-0.422074	1	0	1	0				
..		
433	-0.949986	0	0	STON/O 2. 3101274	-0.504962	0	1	0	0			
773	-0.104637	0	0	2674	-0.502949	0	1	1	0			
25	0.663861	1	5	347077	-0.016444	1	0	0	0			
84	-0.949986	0	0	SO/C 14885	-0.437007	1	0	0	0			
10	-1.949034	1	1	PP 9549	-0.312172	1	0	0	0			

S
709 0
439 1
840 1

```

720 1
39 0
.. ..
433 1
773 0
25 1
84 1
10 1

[179 rows x 13 columns],
331 0
733 0
382 0
704 0
813 0
..
106 1
270 0
860 0
435 1
102 0
Name: Survived, Length: 712, dtype: int64,
709 1
439 0
840 0
720 1
39 1
..
433 0
773 0
25 1
84 1
10 1
Name: Survived, Length: 179, dtype: int64)

```

```
[45]: x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

```
[45]: ((712, 13), (179, 13), (712,), (179,))
```