# train data

August 19, 2022

[1]:
```
import numpy as np # linear algebra
import pandas as pd
from matplotlib import pyplot as plt
```

[2]:
```
!pip install pyforest
```

```
Requirement already satisfied: pyforest in c:\users\bbala\anaconda3\lib\site-
packages (1.1.0)
```

[3]:
```
import pyforest
```

[4]:
```
data=pd.read_csv("Titanic-Train-Data.csv")
data
```

[4]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th… | female | 38.0 | 1 |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 |
| .. | … | … | … | … | … | … | … |
| 886 | 887 | 0 | 2 | Montvila, Rev. Juozas | male | 27.0 | 0 |
| 887 | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 |
| 888 | 889 | 0 | 3 | Johnston, Miss. Catherine Helen "Carrie" | female | NaN | 1 |
| 889 | 890 | 1 | 1 | 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]
```

[5]:
```
data.shape
```

[5]:
```
(891, 12)
```

[6]:
```
data.isna().sum()
```

[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
```

[7]:
```
data.describe()
```

[7]:
```
        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
```

[8]:
```
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
```

[9]:
```
data.dtypes
```

[9]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
```

```
dtype: object
```

[10]:
```
from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()

data['Sex']=label_encoder.fit_transform(data['Sex'])

data['Sex'].value_counts()
```

[10]:
```
1    577
0    314
Name: Sex, dtype: int64
```

[11]:
```
data
```

[11]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | 1 | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th… | 0 | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | 0 | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | 0 | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | 1 | 35.0 | 0 | 0 | | | | |
| .. | … | … | … | … | … | … | … | … | … | … | … | … |
| 886 | 887 | 0 | 2 | Montvila, Rev. Juozas | 1 | 27.0 | 0 | | | | | |
| 887 | 888 | 1 | 1 | Graham, Miss. Margaret Edith | 0 | 19.0 | 0 | | | | | |
| 888 | 889 | 0 | 3 | Johnston, Miss. Catherine Helen "Carrie" | 0 | NaN | 1 | | | | | |
| 889 | 890 | 1 | 1 | Behr, Mr. Karl Howell | 1 | 26.0 | 0 | | | | | |
| 890 | 891 | 0 | 3 | Dooley, Mr. Patrick | 1 | 32.0 | 0 | | | | | |

```
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]
```

[12]:
```
data=data.drop(['Ticket','Cabin','Name'],axis=1)
data
```

[12]:

| | PassengerId | Survived | Pclass | Sex | Age | SibSp | Parch | Fare | Embarked |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | 1 | 22.0 | 1 | 0 | 7.2500 | S |
| 1 | 2 | 1 | 1 | 0 | 38.0 | 1 | 0 | 71.2833 | C |
| 2 | 3 | 1 | 3 | 0 | 26.0 | 0 | 0 | 7.9250 | S |
| 3 | 4 | 1 | 1 | 0 | 35.0 | 1 | 0 | 53.1000 | S |
| 4 | 5 | 0 | 3 | 1 | 35.0 | 0 | 0 | 8.0500 | S |
| .. | … | … | … | … | … | … | … | … | … |
| 886 | 887 | 0 | 2 | 1 | 27.0 | 0 | 0 | 13.0000 | S |
| 887 | 888 | 1 | 1 | 0 | 19.0 | 0 | 0 | 30.0000 | S |
| 888 | 889 | 0 | 3 | 0 | NaN | 1 | 2 | 23.4500 | S |
| 889 | 890 | 1 | 1 | 1 | 26.0 | 0 | 0 | 30.0000 | C |
| 890 | 891 | 0 | 3 | 1 | 32.0 | 0 | 0 | 7.7500 | Q |

[891 rows x 9 columns]

[13]:
```
data['Age'].median()
```

[13]: 28.0

[14]:
```
data['Age']=data['Age'].fillna(value=28)
data
```

[14]:

| | PassengerId | Survived | Pclass | Sex | Age | SibSp | Parch | Fare | Embarked |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 3 | 1 | 22.0 | 1 | 0 | 7.2500 | S |
| 1 | 2 | 1 | 1 | 0 | 38.0 | 1 | 0 | 71.2833 | C |
| 2 | 3 | 1 | 3 | 0 | 26.0 | 0 | 0 | 7.9250 | S |
| 3 | 4 | 1 | 1 | 0 | 35.0 | 1 | 0 | 53.1000 | S |
| 4 | 5 | 0 | 3 | 1 | 35.0 | 0 | 0 | 8.0500 | S |
| .. | … | … | … | … | … | … | … | … | … |
| 886 | 887 | 0 | 2 | 1 | 27.0 | 0 | 0 | 13.0000 | S |
| 887 | 888 | 1 | 1 | 0 | 19.0 | 0 | 0 | 30.0000 | S |
| 888 | 889 | 0 | 3 | 0 | 28.0 | 1 | 2 | 23.4500 | S |
| 889 | 890 | 1 | 1 | 1 | 26.0 | 0 | 0 | 30.0000 | C |
| 890 | 891 | 0 | 3 | 1 | 32.0 | 0 | 0 | 7.7500 | Q |

```
[891 rows x 9 columns]
```

[15]:
```
data['Age'].isna().sum()
```

[15]:
```
0
```

[16]:
```
data.isna().sum()
```

[16]:
```
PassengerId    0
Survived       0
Pclass         0
Sex            0
Age            0
SibSp          0
Parch          0
Fare           0
Embarked       2
dtype: int64
```

[17]:
```
data['Embarked'].value_counts()
```

[17]:
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

[18]:
```
g=data.groupby('Survived')
g['Embarked'].value_counts()
```

[18]:
```
Survived  Embarked
0         S           427
          C            75
          Q            47
1         S           217
          C            93
          Q            30
Name: Embarked, dtype: int64
```

[19]:
```
data['Embarked']=data['Embarked'].fillna(value='S')
data
```

[19]:

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

```
..           …       …     …  …   …    …    …       …       …
886          887         0       2    1  27.0       0       0  13.0000        S
887          888         1       1    0  19.0       0       0  30.0000        S
888          889         0       3    0  28.0       1       2  23.4500        S
889          890         1       1    1  26.0       0       0  30.0000        C
890          891         0       3    1  32.0       0       0   7.7500        Q

[891 rows x 9 columns]
```

[20]:
```
data.head(830)
```

[20]:
```
     PassengerId  Survived  Pclass  Sex   Age  SibSp  Parch     Fare Embarked
0              1         0       3    1  22.0      1      0   7.2500        S
1              2         1       1    0  38.0      1      0  71.2833        C
2              3         1       3    0  26.0      0      0   7.9250        S
3              4         1       1    0  35.0      1      0  53.1000        S
4              5         0       3    1  35.0      0      0   8.0500        S
..           …       …     …  …   …    …    …       …       …
825          826         0       3    1  28.0      0      0   6.9500        Q
826          827         0       3    1  28.0      0      0  56.4958        S
827          828         1       2    1   1.0      0      2  37.0042        C
828          829         1       3    1  28.0      0      0   7.7500        Q
829          830         1       1    0  62.0      0      0  80.0000        S

[830 rows x 9 columns]
```

[21]:
```
from sklearn import preprocessing

label_encoder = preprocessing.LabelEncoder()

data['Embarked']=label_encoder.fit_transform(data['Embarked'])

data['Embarked'].value_counts()
```

[21]:
```
2    646
0    168
1     77
Name: Embarked, dtype: int64
```

[22]:
```
data
```

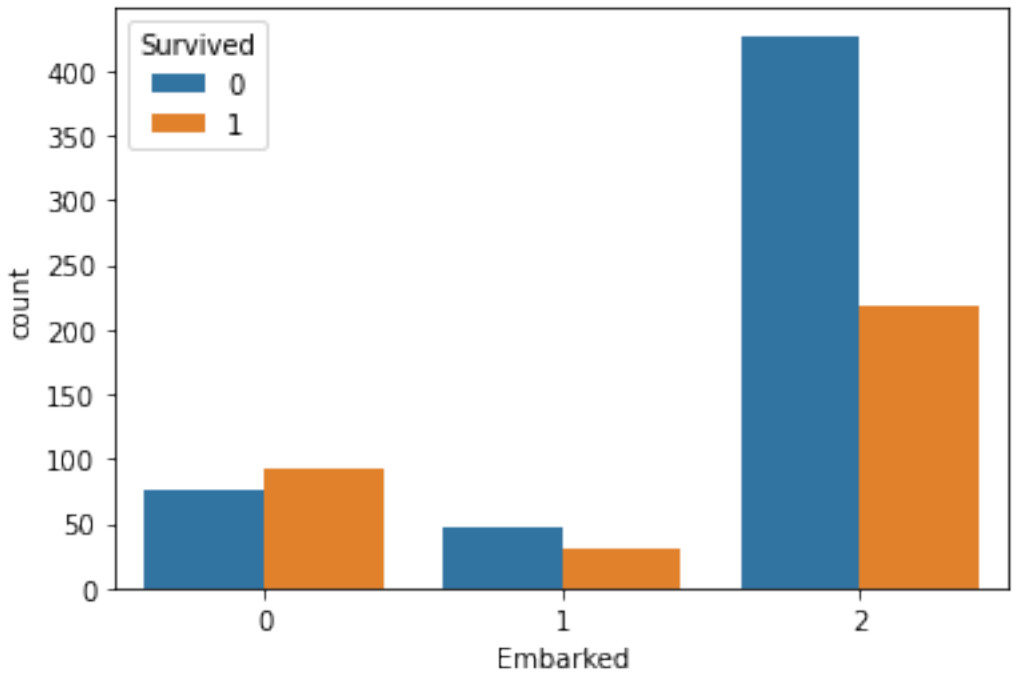
[22]:
```
     PassengerId  Survived  Pclass  Sex   Age  SibSp  Parch     Fare  Embarked
0              1         0       3    1  22.0      1      0   7.2500         2
1              2         1       1    0  38.0      1      0  71.2833         0
2              3         1       3    0  26.0      0      0   7.9250         2
3              4         1       1    0  35.0      1      0  53.1000         2
4              5         0       3    1  35.0      0      0   8.0500         2
```

```
..          ...       ...      ...  ...  ...   ...    ...      ...       ...
886         887         0        2    1   27.0    0      0   13.0000         2
887         888         1        1    0   19.0    0      0   30.0000         2
888         889         0        3    0   28.0    1      2   23.4500         2
889         890         1        1    1   26.0    0      0   30.0000         0
890         891         0        3    1   32.0    0      0    7.7500         1

[891 rows x 9 columns]
```

[23]:
```
import seaborn as sns
```

[24]:
```
sns.countplot(data['Embarked'],hue=data['Survived'])
plt.show()
```

```
C:\Users\bbala\anaconda3\lib\site-packages\seaborn\_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```




[25]:
```
data['Embarked'].value_counts()
```

[25]:
```
2    646
0    168
```

```
1     77
Name: Embarked, dtype: int64
```

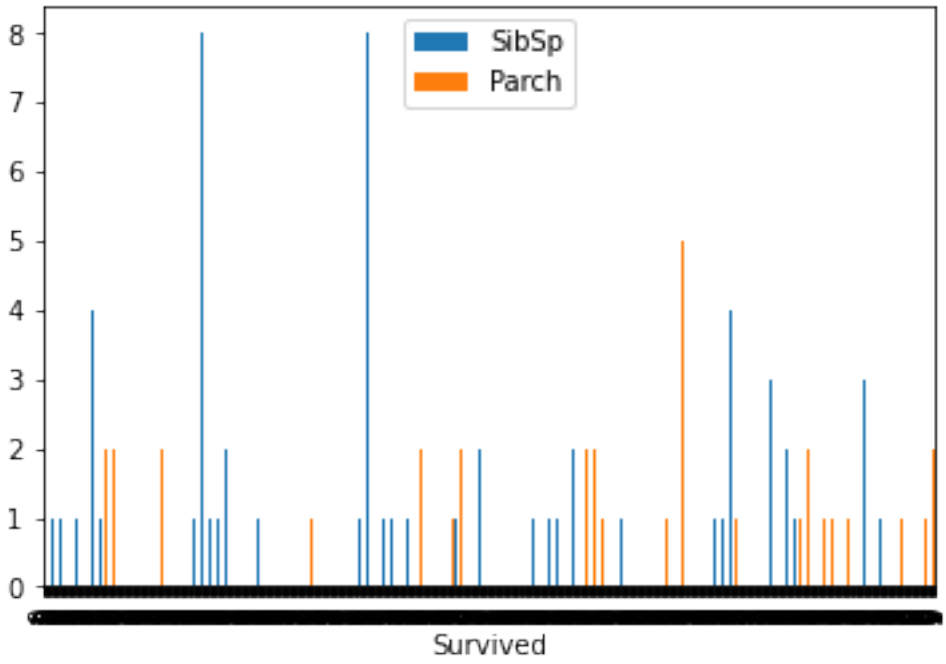
[26]:
```
data
```

[26]:
```
     PassengerId  Survived  Pclass  Sex   Age  SibSp  Parch     Fare  Embarked
0              1         0       3    1  22.0      1      0   7.2500         2
1              2         1       1    0  38.0      1      0  71.2833         0
2              3         1       3    0  26.0      0      0   7.9250         2
3              4         1       1    0  35.0      1      0  53.1000         2
4              5         0       3    1  35.0      0      0   8.0500         2
..           ...       ...     ...  ...   ...    ...    ...      ...       ...
886          887         0       2    1  27.0      0      0  13.0000         2
887          888         1       1    0  19.0      0      0  30.0000         2
888          889         0       3    0  28.0      1      2  23.4500         2
889          890         1       1    1  26.0      0      0  30.0000         0
890          891         0       3    1  32.0      0      0   7.7500         1

[891 rows x 9 columns]
```

[27]:
```
data.corr()
```

[27]:
```
             PassengerId  Survived    Pclass       Sex       Age     SibSp  \
PassengerId     1.000000 -0.005007 -0.035144  0.042939  0.034212 -0.057527
Survived       -0.005007  1.000000 -0.338481 -0.543351 -0.064910 -0.035322
Pclass         -0.035144 -0.338481  1.000000  0.131900 -0.339898  0.083081
Sex             0.042939 -0.543351  0.131900  1.000000  0.081163 -0.114631
Age             0.034212 -0.064910 -0.339898  0.081163  1.000000 -0.233296
SibSp          -0.057527 -0.035322  0.083081 -0.114631 -0.233296  1.000000
Parch          -0.001652  0.081629  0.018443 -0.245489 -0.172482  0.414838
Fare            0.012658  0.257307 -0.549500 -0.182333  0.096688  0.159651
Embarked        0.013128 -0.167675  0.162098  0.108262 -0.018754  0.068230

                Parch      Fare  Embarked
PassengerId -0.001652  0.012658  0.013128
Survived     0.081629  0.257307 -0.167675
Pclass       0.018443 -0.549500  0.162098
Sex         -0.245489 -0.182333  0.108262
Age         -0.172482  0.096688 -0.018754
SibSp        0.414838  0.159651  0.068230
Parch        1.000000  0.216225  0.039798
Fare         0.216225  1.000000 -0.224719
Embarked     0.039798 -0.224719  1.000000
```

[28]:
```
data.plot(x='Survived',y=['SibSp','Parch'],kind='bar')
plt.show()
```



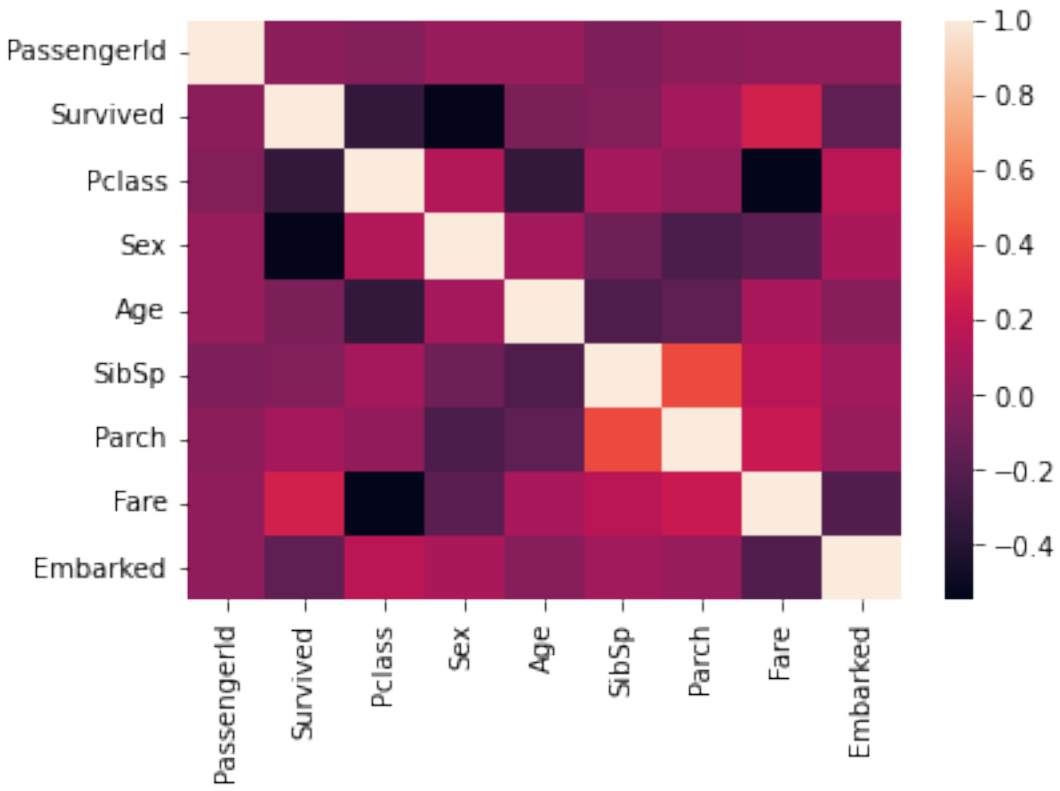

[29]:
```
correlation=data.corr()
correlation['Survived'].sort_values(ascending=False)
```

[29]:
```
Survived       1.000000
Fare           0.257307
Parch          0.081629
PassengerId   -0.005007
SibSp         -0.035322
Age           -0.064910
Embarked      -0.167675
Pclass        -0.338481
Sex           -0.543351
Name: Survived, dtype: float64
```

[30]:
```
import seaborn as sns
```

[31]:
```
sns.heatmap(data.corr())
```

[31]:
```
<AxesSubplot:>
```




[32]:
```
correlation['Fare'].sort_values(ascending=False)
correlation['Fare']
```

[32]:
```
PassengerId    0.012658
Survived       0.257307
Pclass        -0.549500
Sex           -0.182333
Age            0.096688
SibSp          0.159651
Parch          0.216225
Fare           1.000000
Embarked      -0.224719
Name: Fare, dtype: float64
```

[33]:
```
data
```

[33]:

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

```
..          …       …     …  …   …    …     …       …        …
886         887       0     2    1  27.0      0       0  13.0000         2
887         888       1     1    0  19.0      0       0  30.0000         2
888         889       0     3    0  28.0      1       2  23.4500         2
889         890       1     1    1  26.0      0       0  30.0000         0
890         891       0     3    1  32.0      0       0   7.7500         1

[891 rows x 9 columns]
```

[34]:
```
data['family']=data['SibSp']+data['Parch']+1
data=data.drop(['SibSp','Parch'],axis=1)
data=data.drop('PassengerId',axis=1)
data=data.drop('Embarked',axis=1)
data
```

[34]:
```
     Survived  Pclass  Sex   Age     Fare  family
0           0       3    1  22.0   7.2500       2
1           1       1    0  38.0  71.2833       2
2           1       3    0  26.0   7.9250       1
3           1       1    0  35.0  53.1000       2
4           0       3    1  35.0   8.0500       1
..        …     …  …   …       …     …
886         0       2    1  27.0  13.0000       1
887         1       1    0  19.0  30.0000       1
888         0       3    0  28.0  23.4500       4
889         1       1    1  26.0  30.0000       1
890         0       3    1  32.0   7.7500       1

[891 rows x 6 columns]
```

[35]:
```
x=data.drop('Survived',axis=1).values
y=data['Survived'].values
```

[ ]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_spilt
```

[ ]:
```
x_train,x_test,y_train,y_test=train_test_spilt(x,y,test_size=0.
↪3,random_state=100)
```

[ ]:
```
from sklearn.metrics import accuracy_score
```

[ ]:
```
lr=LogisticRegression()
lr.fit(x_train,y_train)
lrpred=lr.predict(x_test)
```

[ ]:
```
accuracy_score(y_test,lrpred)
```

[ ]:
```
from sklearn.model_selection import GridSearchCV
c_space = np.logspace(-5,8,15)
param_grid={'C':c_space}

logreg_cv= GridSearch(lr,param_grid,cv=5)

logreg_cv.fit(x_train,y_train)

print('Tuned Logistic Regression Parameters: {}'.format(logreg_cv.best_params_))
print('Best score is {}'.format(logreg_cv.best_score_))
```

[ ]:
```

```