PYTHONE & DATA ANALYTICS BOOTCAMP

PROJECT

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sb

 To read data frame

Data1=pd.DataFrame(pd.read\_csv('/content/train (1).csv'))

To see data set

Data1

| **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 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **886** | 887 | 0 | 2 | Montvila, Rev. Juozas | male | 27.0 | 0 | 0 | 211536 | 13.0000 | NaN | S |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 | 0 | 112053 | 30.0000 | B42 | S |
| **888** | 889 | 0 | 3 | Johnston, Miss. Catherine Helen "Carrie" | female | NaN | 1 | 2 | W./C. 6607 | 23.4500 | NaN | S |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.0000 | C148 | C |
| **890** | 891 | 0 | 3 | Dooley, Mr. Patrick | male | 32.0 | 0 | 0 | 370376 | 7.7500 | NaN | Q |

891 rows × 12 columns

# To get fist 5 data information

Data1.head()

| **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 |

# To get last 5 data information

Data1.tail()

| **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 |

# To get number of columns and rows

Data1.shape

(891, 12)

# To get column names

Data1.columns

Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'Ticket',

'Fare', 'Cabin', 'Embarked', 'FamilySize', 'Alone'],

dtype='object')

# To get rows information

Data1.index

RangeIndex(start=0, stop=891, step=1)

# To get all null values

x=Data1.isnull().sum()

x

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

drop\_col=x[x>(35/100\*Data1.shape[0])]

drop\_col

Cabin 687

dtype: int64

Data1.drop(drop\_col.index, axis=1, inplace=True)

Data1.fillna(Data1.mean(),inplace=True)

Data1.isnull().sum()

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 0

SibSp 0

Parch 0

Ticket 0

Fare 0

Embarked 2

dtype: int64

# To get information about Embaked columns

Data1['Embarked'].describe()

count 889

unique 3

top S

freq 644

Name: Embarked, dtype: object

Data1['Embarked'].fillna('S',inplace=True)

Data1.isnull().sum()

PassengerId 0

Survived 0

Pclass 0

Name 0

Sex 0

Age 0

SibSp 0

Parch 0

Ticket 0

Fare 0

Embarked 0

dtype: int64

# To get corrlation between other columns

Data1.corr()

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **assengerId** | 1.000000 | -0.005007 | -0.035144 | 0.033207 | -0.057527 | -0.001652 | 0.012658 |
| **Survived** | -0.005007 | 1.000000 | -0.338481 | -0.069809 | -0.035322 | 0.081629 | 0.257307 |
| **Pclass** | -0.035144 | -0.338481 | 1.000000 | -0.331339 | 0.083081 | 0.018443 | -0.549500 |
| **Age** | 0.033207 | -0.069809 | -0.331339 | 1.000000 | -0.232625 | -0.179191 | 0.091566 |
| **SibSp** | -0.057527 | -0.035322 | 0.083081 | -0.232625 | 1.000000 | 0.414838 | 0.159651 |
| **Parch** | -0.001652 | 0.081629 | 0.018443 | -0.179191 | 0.414838 | 1.000000 | 0.216225 |
| **Fare** | 0.012658 | 0.257307 | -0.549500 | 0.091566 | 0.159651 | 0.216225 | 1.000000 |

Data1['FamilySize']=Data1['SibSp']+Data1['Parch']

Data1.drop(['SibSp','Parch'],axis=1,inplace=True)

Data1.corr()

|  | **PassengerId** | **Survived** | **Pclass** | **Age** | **Fare** | **FamilySize** |
| --- | --- | --- | --- | --- | --- | --- |
| **PassengerId** | 1.000000 | -0.005007 | -0.035144 | 0.036847 | 0.012658 | -0.040143 |
| **Survived** | -0.005007 | 1.000000 | -0.338481 | -0.077221 | 0.257307 | 0.016639 |
| **Pclass** | -0.035144 | -0.338481 | 1.000000 | -0.369226 | -0.549500 | 0.065997 |
| **Age** | 0.036847 | -0.077221 | -0.369226 | 1.000000 | 0.096067 | -0.301914 |
| **Fare** | 0.012658 | 0.257307 | -0.549500 | 0.096067 | 1.000000 | 0.217138 |
| **FamilySize** | -0.040143 | 0.016639 | 0.065997 | -0.301914 | 0.217138 | 1.000000 |

Data1['Alone']=[0 if Data1['FamilySize'][i]>0 else 1 for i in Data1.index]

Data1.head()

| **PassengerId** | **Survived** | **Pclass** | **Name** | **Sex** | **Age** | **Ticket** | **Fare** | **Cabin** | **Embarked** | **FamilySize** | **Alone** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | A/5 21171 | 7.2500 | NaN | S | 1 | 0 |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | PC 17599 | 71.2833 | C85 | C | 1 | 0 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | STON/O2. 3101282 | 7.9250 | NaN | S | 0 | 1 |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 113803 | 53.1000 | C123 | S | 1 | 0 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 373450 | 8.0500 | NaN | S | 0 | 1 |

Data1.tail()

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

Data1.groupby(['Alone'])['Survived'].mean()

Alone

0 0.505650

1 0.303538

Name: Survived, dtype: float64

Data1.groupby(['Sex'])['Survived'].mean()

Sex

female 0.742038

male 0.188908

Name: Survived, dtype: float64

Data1.groupby(['Embarked'])['Survived'].mean()

Embarked

C 0.553571

Q 0.389610

S 0.336957

Name: Survived, dtype: float64

Data1.groupby(['Pclass'])['Survived'].mean()

Pclass

1 0.629630

2 0.472826

3 0.242363

Name: Survived, dtype: float64

**CONCLUSION**

* **There is the 891 rows and 12 columns.**
* **When we consider the correlation between other columns, then we can see there is a good correlation between Pclass and Fare value, that mean if Pclass is increase then Fare value is decrease.**
* **We can say that if person is alone then they have less chance of surviving.**
* **And also Females passengers have more chance of surviving.**
* **We see that the percentage of survived was more for Embarked is “C” and survive was minimum for Embarked is “S”.**
* **Finally, we can say that the Pclass 1 has high chance to survived and less chance to survive Pclass is 3.**