# Experiment Title-1

## Implement Exploratory Data Analysis on Any Data Set.

**Student Name: ISHU RATTAN**
**UID: 20BCS7755**
**Branch: BE-CSE**
**Section/Group: 806/B**
**Semester: 5th**
**Date of Performance:08/08/2022**
**Subject Name: Machine Learning**

**Subject Code: 20CSP_806**

### 1. Aim/Overview of the practical:

To Implement Data Exploratory Analysis.

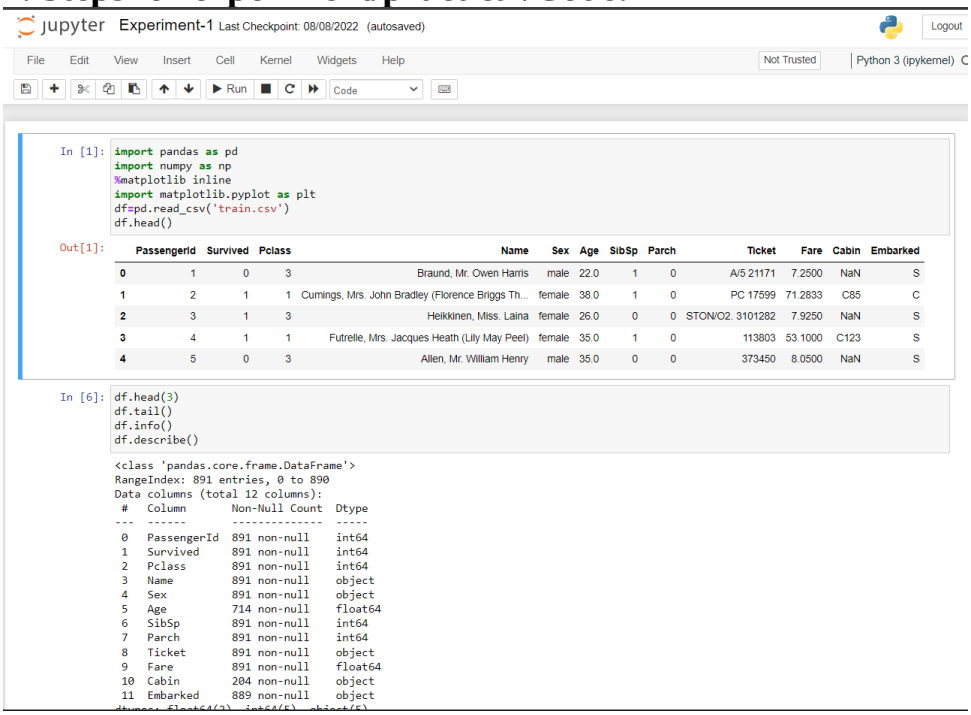
### 2. Task to be done/ Which logistics used:

- Goggle colab (Online Complier)
- Jupyter Notebook (Offline)

**Hardware Requirement:**

- Window 10
- Power Supply
- RAM 4GB

## 4. Steps for experiment/practical/Code:




In [1]:
```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
df=pd.read_csv('train.csv')
df.head()
```

Out[1]:

| | 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 |

In [6]:
```
df.head(3)
df.tail()
df.info()
df.describe()
```

```
<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)
```

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3 (ipykernel)

Run | Code

In [6]:
```
df.head(3)
df.tail()
df.info()
df.describe()
```

```
<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
```

Out[6]:

| | PassengerId | Survived | Pclass | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| count | 891.000000 | 891.000000 | 891.000000 | 714.000000 | 891.000000 | 891.000000 | 891.000000 |
| mean | 446.000000 | 0.383838 | 2.308642 | 29.699118 | 0.523008 | 0.381594 | 32.204208 |
| std | 257.353842 | 0.486592 | 0.836071 | 14.526497 | 1.102743 | 0.806057 | 49.693429 |
| min | 1.000000 | 0.000000 | 1.000000 | 0.420000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 223.500000 | 0.000000 | 2.000000 | 20.125000 | 0.000000 | 0.000000 | 7.910400 |
| 50% | 446.000000 | 0.000000 | 3.000000 | 28.000000 | 0.000000 | 0.000000 | 14.454200 |
| 75% | 668.500000 | 1.000000 | 3.000000 | 38.000000 | 1.000000 | 0.000000 | 31.000000 |
| max | 891.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 6.000000 | 512.329200 |

In [8]:
```
df.iloc[3]
df.loc[0:4,'Ticket']
```

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved)

File Edit View Insert Cell Kernel Widgets Help | Not Trusted | Python 3 (ipykernel)

Run | Code

```
 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
```

Out[6]:

| | PassengerId | Survived | Pclass | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| count | 891.000000 | 891.000000 | 891.000000 | 714.000000 | 891.000000 | 891.000000 | 891.000000 |
| mean | 446.000000 | 0.383838 | 2.308642 | 29.699118 | 0.523008 | 0.381594 | 32.204208 |
| std | 257.353842 | 0.486592 | 0.836071 | 14.526497 | 1.102743 | 0.806057 | 49.693429 |
| min | 1.000000 | 0.000000 | 1.000000 | 0.420000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 223.500000 | 0.000000 | 2.000000 | 20.125000 | 0.000000 | 0.000000 | 7.910400 |
| 50% | 446.000000 | 0.000000 | 3.000000 | 28.000000 | 0.000000 | 0.000000 | 14.454200 |
| 75% | 668.500000 | 1.000000 | 3.000000 | 38.000000 | 1.000000 | 0.000000 | 31.000000 |
| max | 891.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 6.000000 | 512.329200 |

In [8]:
```
df.iloc[3]
df.loc[0:4,'Ticket']
df['Ticket'].head()
```

Out[8]:
```
0           A/5 21171
1            PC 17599
2    STON/O2. 3101282
3              113803
4              373450
Name: Ticket, dtype: object
```

In [10]:
```
df[df.Age>65]
```

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| max | 891.000000 | 1.000000 | 3.000000 | 80.000000 | 8.000000 | 6.000000 | 512.329200 |

In [8]:
```
df.iloc[3]
df.loc[0:4,'Ticket']
df['Ticket'].head()
```

Out[8]:
```
0           A/5 21171
1            PC 17599
2    STON/O2. 3101282
3              113803
4              373450
Name: Ticket, dtype: object
```

In [10]:
```
df[df.Age>65]
```

Out[10]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **33** | 34 | 0 | 2 | Wheadon, Mr. Edward H | male | 66.0 | 0 | 0 | C.A. 24579 | 10.5000 | NaN | S |
| **96** | 97 | 0 | 1 | Goldschmidt, Mr. George B | male | 71.0 | 0 | 0 | PC 17754 | 34.6542 | A5 | C |
| **116** | 117 | 0 | 3 | Connors, Mr. Patrick | male | 70.5 | 0 | 0 | 370369 | 7.7500 | NaN | Q |
| **493** | 494 | 0 | 1 | Artagaveytia, Mr. Ramon | male | 71.0 | 0 | 0 | PC 17609 | 49.5042 | NaN | C |
| **630** | 631 | 1 | 1 | Barkworth, Mr. Algernon Henry Wilson | male | 80.0 | 0 | 0 | 27042 | 30.0000 | A23 | S |
| **672** | 673 | 0 | 2 | Mitchell, Mr. Henry Michael | male | 70.0 | 0 | 0 | C.A. 24580 | 10.5000 | NaN | S |
| **745** | 746 | 0 | 1 | Crosby, Capt. Edward Gifford | male | 70.0 | 1 | 1 | WE/P 5735 | 71.0000 | B22 | S |
| **851** | 852 | 0 | 3 | Svensson, Mr. Johan | male | 74.0 | 0 | 0 | 347060 | 7.7750 | NaN | S |

In [12]:
```
df[(df.Age==1)&(df.SibSp==5)]
```

Out[12]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **386** | 387 | 0 | 3 | Goodwin, Master. Sidney Leonard | male | 1.0 | 5 | 2 | CA 2144 | 46.9 | NaN | S |

In [13]:
```
df[(df.Age==1)|(df.SibSp==5)]
```

Out[13]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

In [10]:
```
df[df.Age>65]
```

Out[10]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **33** | 34 | 0 | 2 | Wheadon, Mr. Edward H | male | 66.0 | 0 | 0 | C.A. 24579 | 10.5000 | NaN | S |
| **96** | 97 | 0 | 1 | Goldschmidt, Mr. George B | male | 71.0 | 0 | 0 | PC 17754 | 34.6542 | A5 | C |
| **116** | 117 | 0 | 3 | Connors, Mr. Patrick | male | 70.5 | 0 | 0 | 370369 | 7.7500 | NaN | Q |
| **493** | 494 | 0 | 1 | Artagaveytia, Mr. Ramon | male | 71.0 | 0 | 0 | PC 17609 | 49.5042 | NaN | C |
| **630** | 631 | 1 | 1 | Barkworth, Mr. Algernon Henry Wilson | male | 80.0 | 0 | 0 | 27042 | 30.0000 | A23 | S |
| **672** | 673 | 0 | 2 | Mitchell, Mr. Henry Michael | male | 70.0 | 0 | 0 | C.A. 24580 | 10.5000 | NaN | S |
| **745** | 746 | 0 | 1 | Crosby, Capt. Edward Gifford | male | 70.0 | 1 | 1 | WE/P 5735 | 71.0000 | B22 | S |
| **851** | 852 | 0 | 3 | Svensson, Mr. Johan | male | 74.0 | 0 | 0 | 347060 | 7.7750 | NaN | S |

In [12]:
```
df[(df.Age==1)&(df.SibSp==5)]
```

Out[12]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **386** | 387 | 0 | 3 | Goodwin, Master. Sidney Leonard | male | 1.0 | 5 | 2 | CA 2144 | 46.9 | NaN | S |

In [13]:
```
df[(df.Age==1)|(df.SibSp==5)]
```

Out[13]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **59** | 60 | 0 | 3 | Goodwin, Master. William Frederick | male | 11.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **71** | 72 | 0 | 3 | Goodwin, Miss. Lillian Amy | female | 16.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **164** | 165 | 0 | 3 | Panula, Master. Eino Viljami | male | 1.0 | 4 | 1 | 3101295 | 39.6875 | NaN | S |
| **172** | 173 | 1 | 3 | Johnson, Miss. Eleanor Ileen | female | 1.0 | 1 | 1 | 347742 | 11.1333 | NaN | S |
| **183** | 184 | 1 | 2 | Becker, Master. Richard F | male | 1.0 | 2 | 1 | 230136 | 39.0000 | F4 | S |
| **381** | 382 | 1 | 3 | Nakid, Miss. Maria ("Mary") | female | 1.0 | 0 | 2 | 2653 | 15.7417 | NaN | C |
| **386** | 387 | 0 | 3 | Goodwin, Master. Sidney Leonard | male | 1.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **480** | 481 | 0 | 3 | Goodwin, Master. Harold Victor | male | 9.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **683** | 684 | 0 | 3 | Goodwin, Mr. Charles Edward | male | 14.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

Run Code

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **630** | 631 | 1 | 1 | Barkworth, Mr. Algernon Henry Wilson | male | 80.0 | 0 | 0 | 27042 | 30.0000 | A23 | S |
| **672** | 673 | 0 | 2 | Mitchell, Mr. Henry Michael | male | 70.0 | 0 | 0 | C.A. 24580 | 10.5000 | NaN | S |
| **745** | 746 | 0 | 1 | Crosby, Capt. Edward Gifford | male | 70.0 | 1 | 1 | WE/P 5735 | 71.0000 | B22 | S |
| **851** | 852 | 0 | 3 | Svensson, Mr. Johan | male | 74.0 | 0 | 0 | 347060 | 7.7750 | NaN | S |

In [12]:
```
df[(df.Age==1)&(df.SibSp==5)]
```

Out[12]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **386** | 387 | 0 | 3 | Goodwin, Master. Sidney Leonard | male | 1.0 | 5 | 2 | CA 2144 | 46.9 | NaN | S |

In [13]:
```
df[(df.Age==1)|(df.SibSp==5)]
```

Out[13]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **59** | 60 | 0 | 3 | Goodwin, Master. William Frederick | male | 11.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **71** | 72 | 0 | 3 | Goodwin, Miss. Lillian Amy | female | 16.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **164** | 165 | 0 | 3 | Panula, Master. Eino Viljami | male | 1.0 | 4 | 1 | 3101295 | 39.6875 | NaN | S |
| **172** | 173 | 1 | 3 | Johnson, Miss. Eleanor Ileen | female | 1.0 | 1 | 1 | 347742 | 11.1333 | NaN | S |
| **183** | 184 | 1 | 2 | Becker, Master. Richard F | male | 1.0 | 2 | 1 | 230136 | 39.0000 | F4 | S |
| **381** | 382 | 1 | 3 | Nakid, Miss. Maria ("Mary") | female | 1.0 | 0 | 2 | 2653 | 15.7417 | NaN | C |
| **386** | 387 | 0 | 3 | Goodwin, Master. Sidney Leonard | male | 1.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **480** | 481 | 0 | 3 | Goodwin, Master. Harold Victor | male | 9.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **683** | 684 | 0 | 3 | Goodwin, Mr. Charles Edward | male | 14.0 | 5 | 2 | CA 2144 | 46.9000 | NaN | S |
| **788** | 789 | 1 | 3 | Dean, Master. Bertram Vere | male | 1.0 | 1 | 2 | C.A. 2315 | 20.5750 | NaN | S |
| **827** | 828 | 1 | 2 | Mallet, Master. Andre | male | 1.0 | 0 | 2 | S.C./PARIS 2079 | 37.0042 | NaN | C |

In [14]:
```
df['Embarked'].unique()
```

Out[14]:
```
array(['S', 'C', 'Q', nan], dtype=object)
```

In [16]:
```
#missing value
print(df['Age'].mean())
print(df['Fare'].median())
```

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 827 | 828 | 1 | 2 | Mallet, Master. Andre | male | 1.0 | 0 | 2 | S.C./PARIS 2079 | 37.0042 | NaN | C |

In [14]:
```
df['Embarked'].unique()
```

Out[14]:
```
array(['S', 'C', 'Q', nan], dtype=object)
```

In [16]:
```
#missing value
print(df['Age'].mean())
print(df['Fare'].median())
```

```
29.69911764705882
14.4542
```

In [17]:
```
print((df['Sex']=='female').sum())
```

```
314
```

In [18]:
```
df.info()
df['Age'].fillna(30)
df.isnull().sum()
```

```
<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
```

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
 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
```

Out[18]:
```
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 [19]:
```
df.groupby('Survived')['Age'].mean()
```

Out[19]:
```
Survived
0    30.626179
1    28.343690
Name: Age, dtype: float64
```

In [22]:
```
df.pivot_table(index='Sex', columns='Parch',values='Survived',aggfunc='sum')
```

Out[22]:

| Parch | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Sex | | | | | | | |
| female | 153.0 | 46.0 | 30.0 | 3.0 | 0.0 | 1.0 | 0.0 |
| male | 80.0 | 19.0 | 10.0 | 0.0 | 0.0 | 0.0 | NaN |

jupyter Experiment-1 Last Checkpoint: 08/08/2022 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

In [19]:
```
df.groupby('Survived')['Age'].mean()
```

Out[19]:
```
Survived
0    30.626179
1    28.343690
Name: Age, dtype: float64
```

In [22]:
```
df.pivot_table(index='Sex', columns='Parch',values='Survived',aggfunc='sum')
```

Out[22]:

| Parch | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Sex | | | | | | | |
| female | 153.0 | 46.0 | 30.0 | 3.0 | 0.0 | 1.0 | 0.0 |
| male | 80.0 | 19.0 | 10.0 | 0.0 | 0.0 | 0.0 | NaN |

In [23]:
```
df.pivot_table(index='Sex', columns='SibSp',values='Survived',aggfunc='sum')
```

Out[23]:

| SibSp | 0 | 1 | 2 | 3 | 4 | 5 | 8 |
|---|---|---|---|---|---|---|---|
| Sex | | | | | | | |
| female | 137 | 80 | 10 | 4 | 2 | 0 | 0 |
| male | 73 | 32 | 3 | 0 | 1 | 0 | 0 |

In [ ]:

**Learning outcomes (What I have learnt):**

1. Learnt about Machine Learning

2.Learnt about how dataset is used.

3.Learnt about different types of commands

4.Learnt about how to use these commands on time.

**Evaluation Grid (To be created as per the SOP and Assessment guidelines by the faculty):**

| Sr. No. | Parameters | Marks Obtained | Maximum Marks |
|---|---|---|---|
| 1. | | | |
| 2. | | | |
| 3. | | | |
| | | | |