**EXPERIMENT 1.1**

**Name**

**-**

**SANSKAR AGRAWAL**

**Branch**

**-**

**CSE**

**Semester**

**-**

**5**

**th**

**Subject**

**Name**

**–**

**Machine**

**Learning**

**Lab**

**Subject**

**Code**

**-**

**20**

**CSP**

**-**

**317**

**UID**

**-**

**20**

**BCS**

**5914**

**Section/Group**

**-**

**806**

**B**

**Date**

**of**

**Performance**

**-**

**28/08/2022**

**AIM -EXPLORATORY DATA ANALYSIS (EDA).**

**OBJECTIVE –**To Understand the data i.e., Data is clean , it doesn’t have any null values , missing values , remove noise , identify variables in dataset and relationship between variables to conclude the values.

**S/W Requirement: -** VS Code or Jupyter Notebook

**INPUT AND OUTPUT –**

**Importing Libraries: -**

import pandas as pd

import numpy as np

%matplotlib inline

import matplotlib.pyplot as plt

df = pd.read\_csv('train.csv')

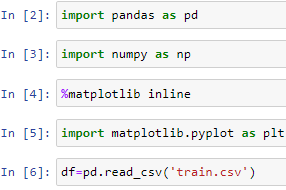
df.head()

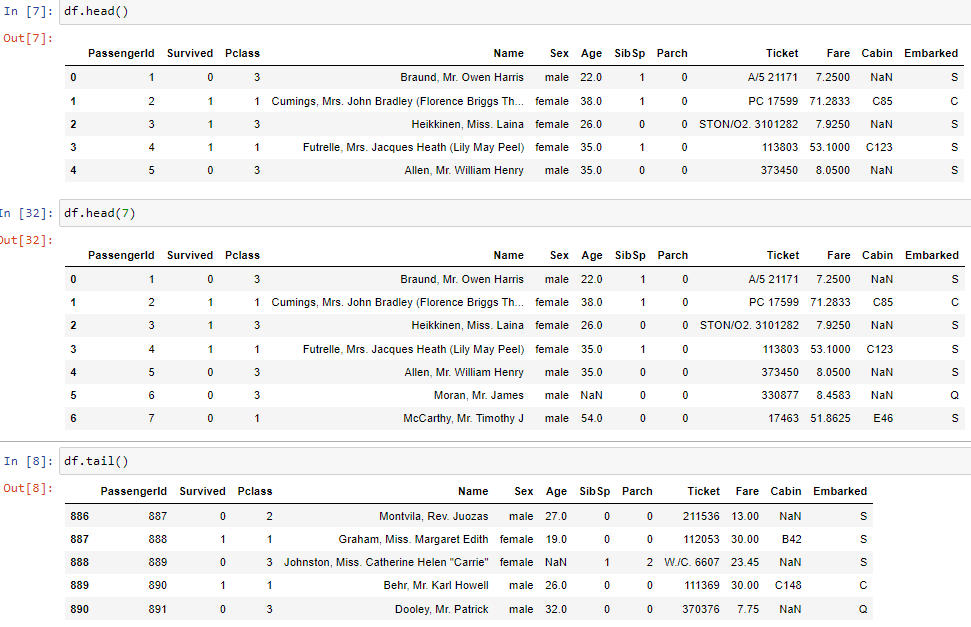
df.head(7)

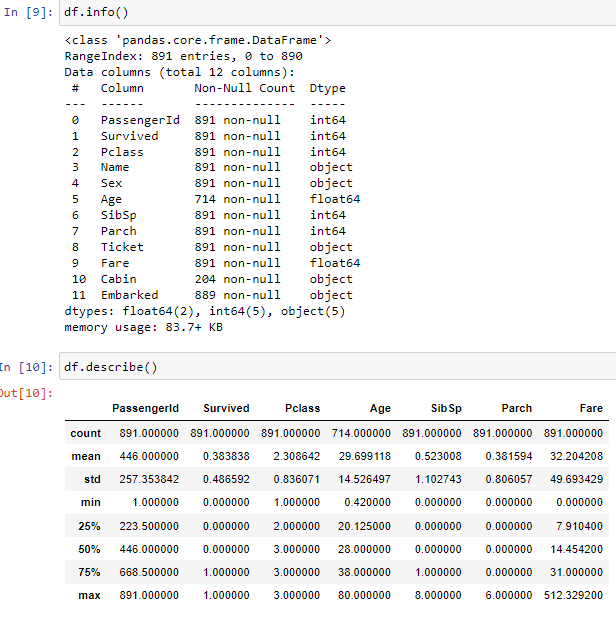
df.tail()

df.info()

df.describe()





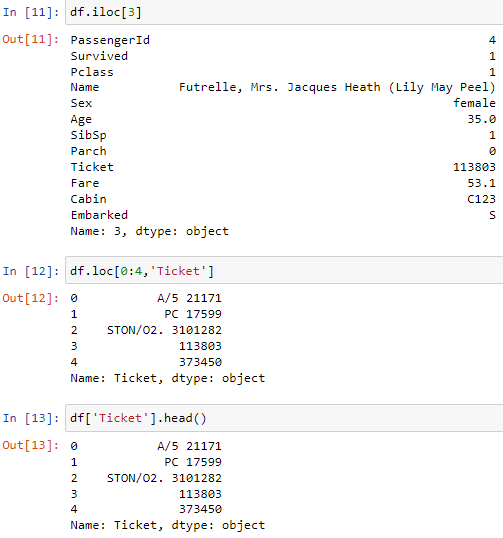


**Indexing: -**

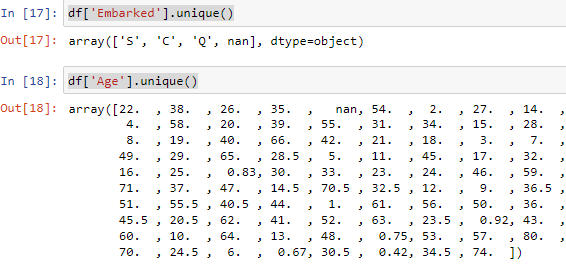
df.iloc[3]

df.loc[0:4,'Ticket']

df['Ticket'].head()



**Distinct Elements: -**



**Selections: -**

df[df.Age>65]

df[(df.Age==11)&(df.SibSp==5)]

df[(df.Age==11)|(df.SibSp==5)]

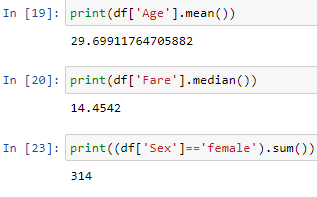


**Missing values find and treatment: -**

print(df['Age'].mean())

print(df['Fare'].median())

print((df['Sex']=='female').sum())

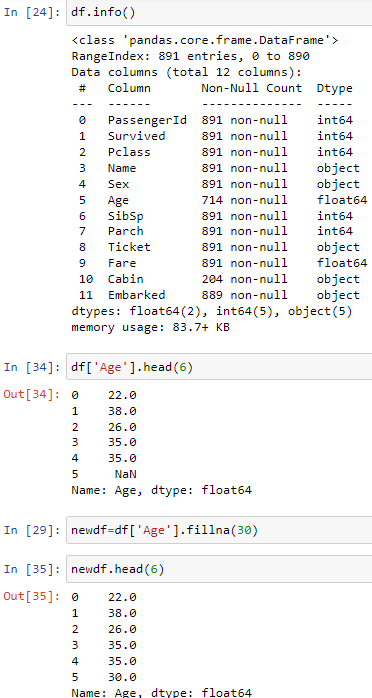


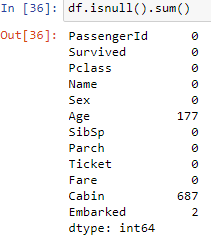
**Missing Data: -**

df[df.Age>65]

df[(df.Age==11)&(df.SibSp==5)]

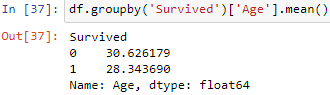
df[(df.Age==11)|(df.SibSp==5)]





**Groupby: -**

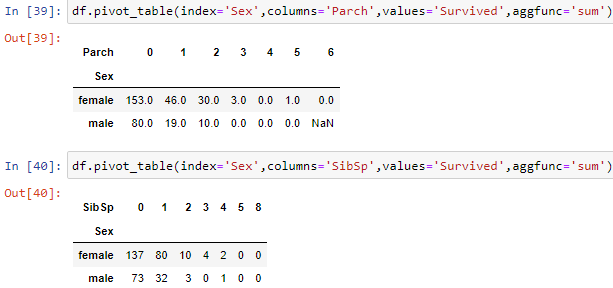
df[df.Age>65]



**Missing Data: -**

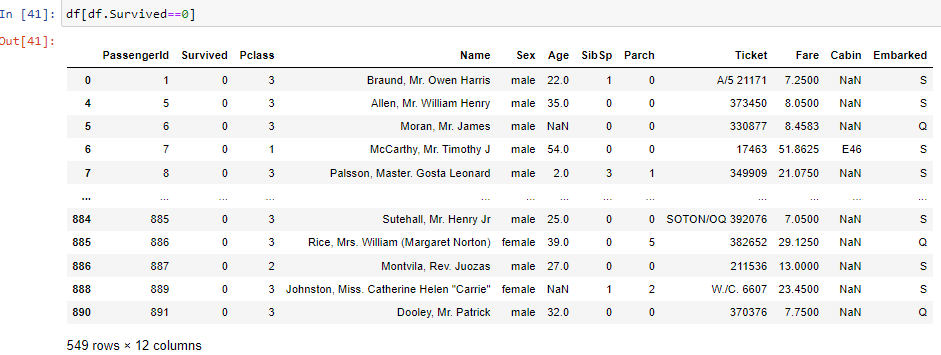
df. pivot\_table(index='Sex’, columns='Parch',values='Survived',aggfunc='sum')

df. pivot\_table(index='Sex’, columns='SibSp',values='Survived',aggfunc='sum')

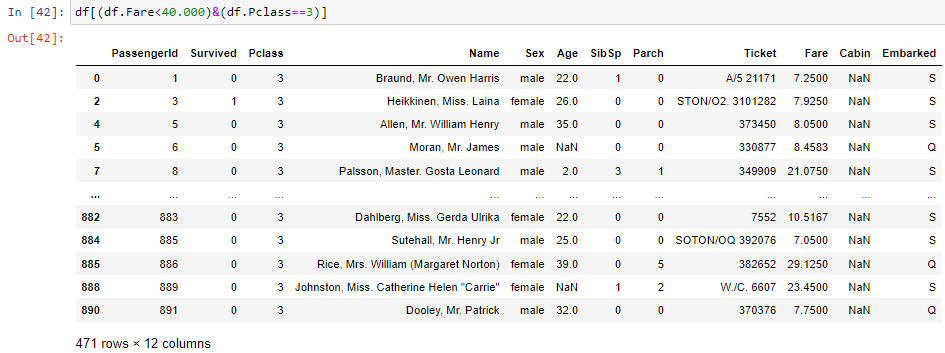


**Exercises:**

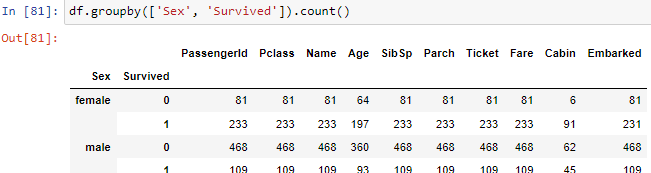
select passengers that died



select passengers who paid less than 40.000 and were in third class



count the number of survived and dead per each gender



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

1. Identify the faulty points so that we can clean the data.
2. How to deal with missing values of variables (Columns) in dataset.
3. To Deal with Outliers.
4. To find Relationship between different variables and map different type of Graphs.

**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. |  |  |  |
| 4. |  |  |  |