Assignment No- 01

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Data Wrangling I

In [1]:
```
import pandas as pd
import urllib.request
```

In [4]:
```
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.d
```

In [6]:
```
column_names = ["Sepal_Length", "Sepal_Width", "Petal_Length", "Petal_Width"
```

In [24]:
```
iris = pd.read_csv(url, names=column_names)
iris.head()
```

Out[24]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [10]:
```
iris.head()
iris.tail()
iris.index
iris.columns
iris.shape
iris.dtypes
iris["Sepal_Length"]
iris.iloc[5]
iris[0:3]
iris.loc[:,["Sepal_Length","Sepal_Width"]]
iris.iloc[:5,:]
iris.iloc[:,:5]
iris.iloc[:5,:5]
```

Out[10]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [12]:
```
iris.isnull().any()
```

Out[12]:
```
Sepal_Length    False
Sepal_Width     False
Petal_Length    False
Petal_Width     False
Class           False
dtype: bool
```

In [14]:
```
iris.isnull().sum()
iris.isnull().sum().sum()
```

Out[14]:
```
0
```

In [16]:
```
iris.isnull().sum(axis=1)
```

Out[16]:
```
0      0
1      0
2      0
3      0
4      0
      ..
145    0
146    0
147    0
148    0
149    0
Length: 150, dtype: int64
```

In [18]:
```
iris.dtypes
```

Out[18]:
```
Sepal_Length    float64
Sepal_Width     float64
Petal_Length    float64
Petal_Width     float64
Class            object
dtype: object
```

In [22]:
```
iris["Petal_Length"] = iris["Petal_Length"].astype('int')
iris.head()
```

Out[22]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1	0.2	Iris-setosa
1	4.9	3.0	1	0.2	Iris-setosa
2	4.7	3.2	1	0.2	Iris-setosa
3	4.6	3.1	1	0.2	Iris-setosa
4	5.0	3.6	1	0.2	Iris-setosa

In [26]:
```
from sklearn import datasets
df = pd.DataFrame(iris)
df.head()
```

Out[26]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [32]:
```
features = df.iloc[:, :-1].astype(float)
labels = df.iloc[:, -1]
features.head()
```

Out[32]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

In [38]:
```
from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
features_scaled = scaler.fit_transform(features)
iris_normalized = pd.DataFrame(features_scaled, columns=column_names[:-1])
iris_normalized["class"] = labels
print(iris_normalized.head())
```

```
   Sepal_Length  Sepal_Width  Petal_Length  Petal_Width        class
0      0.222222     0.625000      0.067797     0.041667  Iris-setosa
1      0.166667     0.416667      0.067797     0.041667  Iris-setosa
2      0.111111     0.500000      0.050847     0.041667  Iris-setosa
3      0.083333     0.458333      0.084746     0.041667  Iris-setosa
4      0.194444     0.666667      0.067797     0.041667  Iris-setosa
```

In [40]:
```
df["Class"].unique()
```

Out[40]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

In [44]:
```
label_encoder=preprocessing.LabelEncoder()
df["Class"]=label_encoder.fit_transform(df["Class"])
df["Class"].unique()
df.head()
```

Out[44]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

In [ ]:
```

```

Assignment N0- 02

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Data Wrangling II

In [2]:
```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings("ignore")
```

In [6]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
```

In [8]:
```
df.shape
```

Out[8]: (30, 7)

In [14]:
```
df.head()
```

Out[14]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [16]:
```
df.describe()
```

Out[16]:

	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club_Join
count	27.000000	27.000000	27.000000	30.000000	30.0
mean	72.000000	83.666667	70.629630	86.400000	2019.7
std	6.101702	6.342773	5.917043	6.926212	1.3
min	60.000000	75.000000	61.000000	76.000000	2018.0
25%	68.000000	78.000000	66.000000	80.250000	2018.0
50%	74.000000	83.000000	70.000000	85.500000	2020.0
75%	77.000000	88.000000	76.000000	92.500000	2021.0
max	80.000000	95.000000	79.000000	100.000000	2021.0

In [18]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Gender                30 non-null     object
 1   Math_Score            27 non-null     float64
 2   Reading_Score         27 non-null     float64
 3   Writing_Score         27 non-null     float64
 4   Placement_Score       30 non-null     int64
 5   Club_Join_Date        30 non-null     int64
 6   Placement_Offer_Count  30 non-null     int64
dtypes: float64(3), int64(3), object(1)
memory usage: 1.8+ KB
```

In [22]:
```
df.isnull().head()
```

Out[22]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	False	False	False	False	False	
1	False	False	False	False	False	
2	False	False	False	False	False	
3	False	False	False	False	False	
4	False	False	False	False	False	

In [24]:
```
df.isnull().sum()
```

Out[24]:
```
Gender                 0
Math_Score             3
Reading_Score          3
Writing_Score          3
Placement_Score        0
Club_Join_Date         0
Placement_Offer_Count  0
dtype: int64
```

In [26]:
```
series = pd.isnull(df["Math_Score"])
df[series]
```

Out[26]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Clu
8	Female	NaN	85.0	68.0	81	
14	Female	NaN	76.0	66.0	86	
21	Male	NaN	86.0	NaN	85	

In [30]:
```
df.notnull().head()
```

Out[30]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	True	True	True	True	True	
1	True	True	True	True	True	
2	True	True	True	True	True	
3	True	True	True	True	True	
4	True	True	True	True	True	

In [32]:
```
df.notnull().sum()
```

Out[32]:
```
Gender                  30
Math_Score              27
Reading_Score           27
Writing_Score           27
Placement_Score         30
Club_Join_Date          30
Placement_Offer_Count   30
dtype: int64
```

In [36]:
```
series1 = pd.notnull(df["Math_Score"])
df[series1].head()
```

Out[36]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [40]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['Gender'] = le.fit_transform(df['Gender'])
newdf = df
newdf.head()
```

Out[40]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	1	69.0	95.0	71.0	93	
1	1	77.0	82.0	78.0	86	
2	1	76.0	90.0	70.0	83	
3	1	63.0	88.0	78.0	77	
4	0	70.0	76.0	63.0	98	

In [44]:
```
missing_values = ["Na","na"]
ead_csv("Downloads/Students_Performance.csv", na_values =
```

```
missing_values)
df.head()
```

Out[44]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [46]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
ndf = df
ndf.fillna(0).head()
```

Out[46]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [48]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df['Math_Score'] = df['Math_Score'].fillna(df['Math_Score'].mean())
df.head()
```

Out[48]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [50]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df['Math_Score'] = df['Math_Score'].fillna(df['Math_Score'].median())
df.head()
```

Out[50]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [52]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df['Math_Score'] = df['Math_Score'].fillna(df['Math_Score'].std())
df.head()
```

Out[52]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [54]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df['Math_Score'] = df['Math_Score'].fillna(df['Math_Score'].min())
df.head()
```

Out[54]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [56]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df['Math_Score'] = df['Math_Score'].fillna(df['Math_Score'].max())
df.head()
```

Out[56]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [58]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
mean_value=df['Math_Score'].mean()
df['Math_Score'].fillna(value=mean_value, inplace=True)
df.head()
```

Out[58]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [62]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df.replace(to_replace = np.nan, value = -99).head()
```

Out[62]:

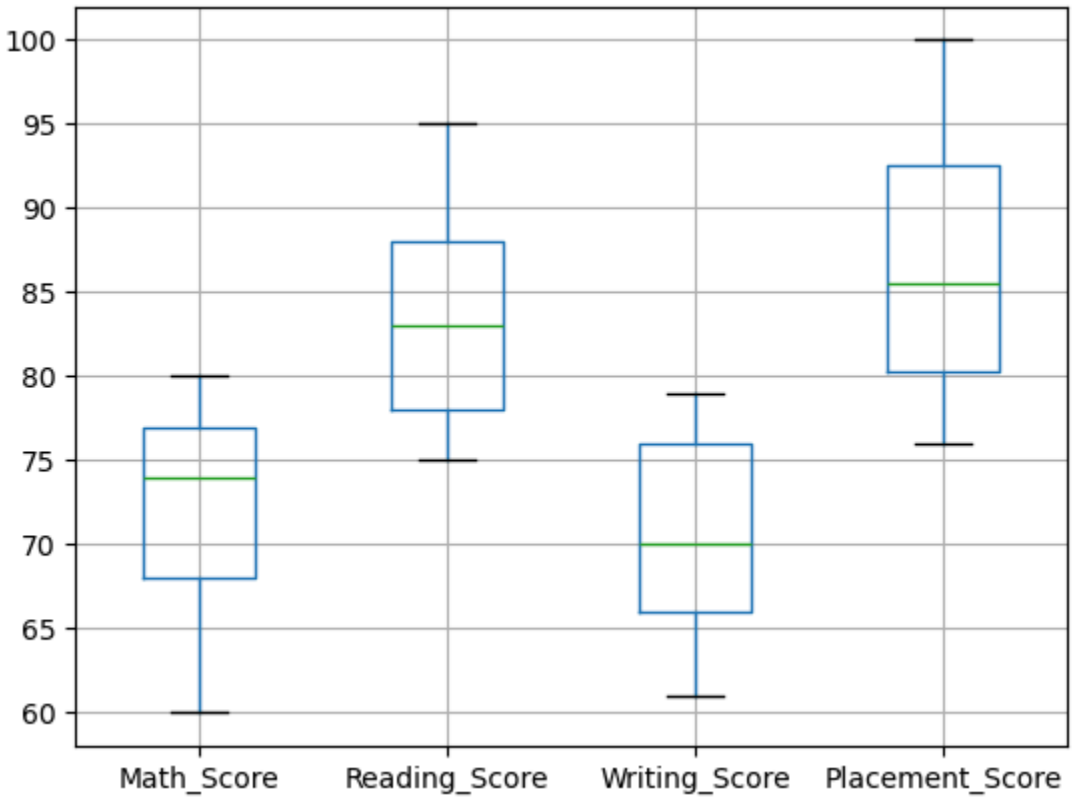
	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [64]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df.dropna().head()
```

Out[64]:

	Gender	Math_Score	Reading_Score	Writing_Score	Placement_Score	Club
0	Male	69.0	95.0	71.0	93	
1	Male	77.0	82.0	78.0	86	
2	Male	76.0	90.0	70.0	83	
3	Male	63.0	88.0	78.0	77	
4	Female	70.0	76.0	63.0	98	

In [68]:
```
import matplotlib.pyplot as plt
df = pd.read_csv("Downloads/Students_Performance.csv")
df.columns = df.columns.str.strip()
col = ['Math_Score', 'Reading_Score', 'Writing_Score', 'Placement_Score']
df.boxplot(column=col)
plt.show()
```



In [70]:
```
print(np.where(df['Math_Score']>80))
print(np.where(df['Reading_Score']<75))
print(np.where(df['Writing_Score']>80))
```

```
(array([], dtype=int64),)
(array([], dtype=int64),)
(array([], dtype=int64),)
```

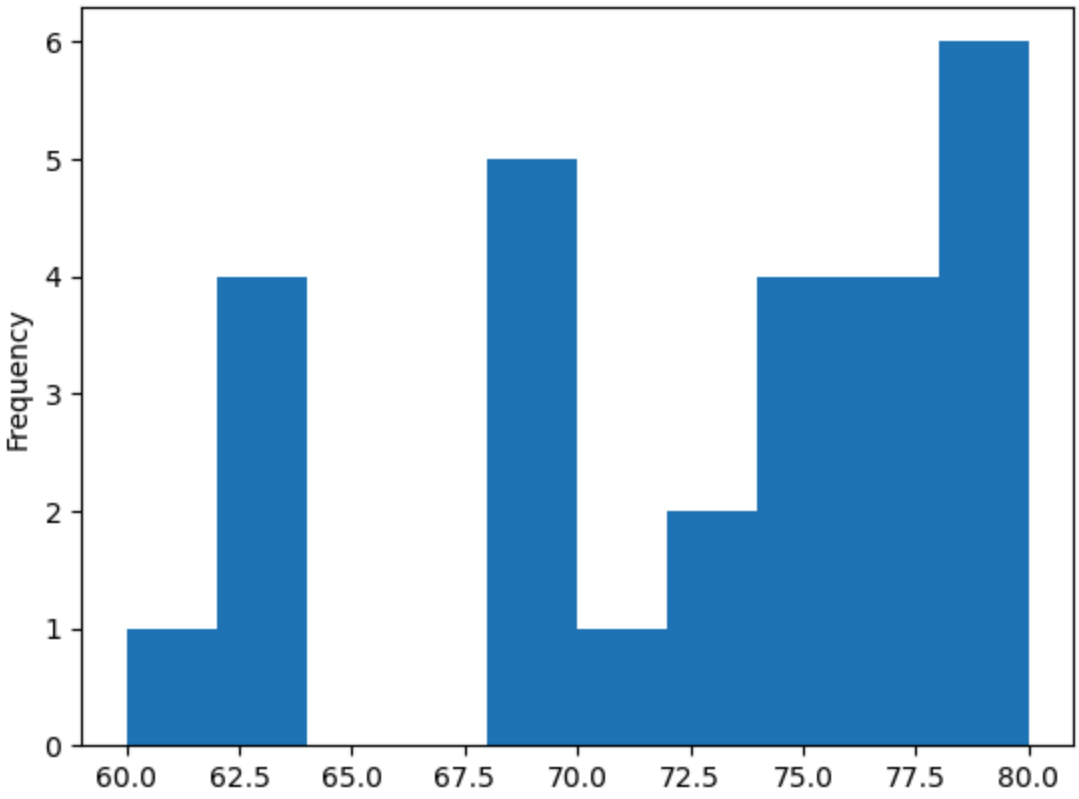
In [72]:
```
df = pd.read_csv("Downloads/Students_Performance.csv")
df.columns = df.columns.str.strip()
mean_math = df["Math_Score"].mean()
std_math = df["Math_Score"].std()
df["Zscore"] = (df["Math_Score"] - mean_math) / std_math
print(df[["Math_Score", "Zscore"]].head())
```

```
   Math_Score    Zscore
0        69.0 -0.491666
1        77.0  0.819443
2        76.0  0.655555
3        63.0 -1.474998
4        70.0 -0.327777
```

In [74]:
```
import matplotlib.pyplot as plt
df['Math_Score'].plot(kind = 'hist')
```

Out[74]: <Axes: ylabel='Frequency'>

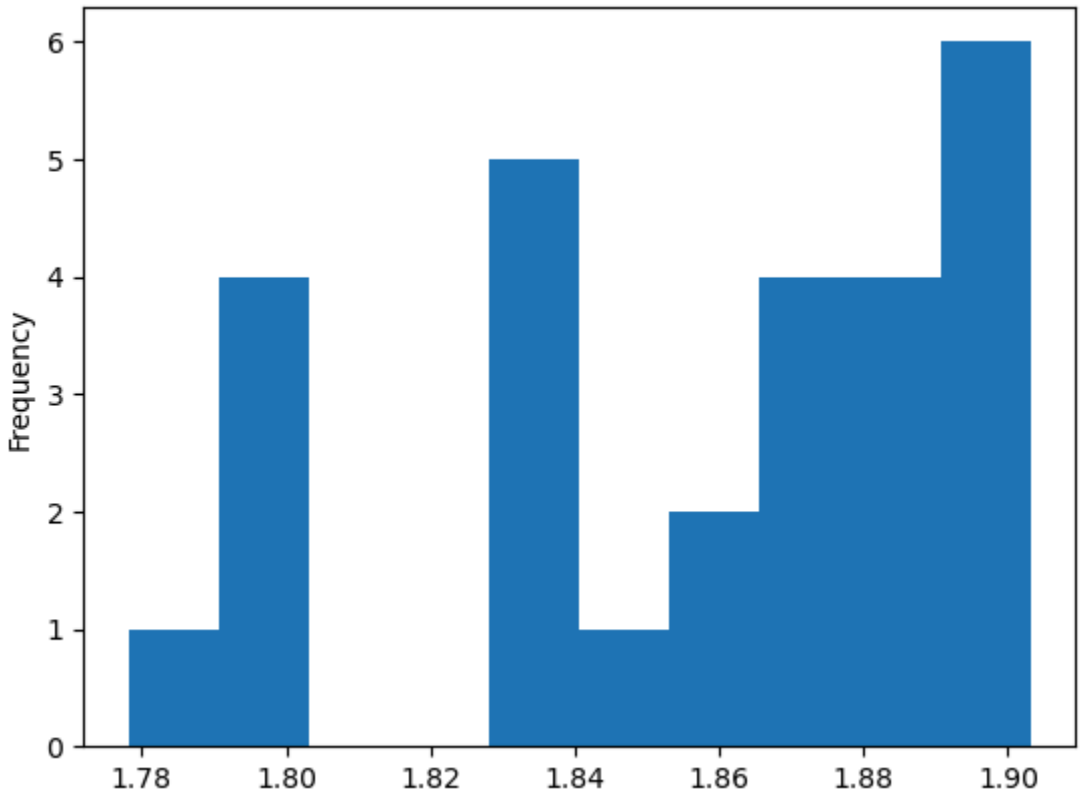
In [76]:
```
plt.show()
```



In [78]:
```
df['log_math'] = np.log10(df['Math_Score'])
df['log_math'].plot(kind = 'hist')
```

Out[78]: <Axes: ylabel='Frequency'>

In [80]:
```
plt.show()
```



In [ ]:

Assignment No- 03

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Descriptive Statistic

In [1]:
```
import numpy as np
import pandas as pd
df = pd.read_csv("Downloads/HR-Employee-Attrition.csv")
df.head()
```

Out[1]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome
0	41	Yes	Travel_Rarely	1102	Sales	1
1	49	No	Travel_Frequently	279	Research & Development	8
2	37	Yes	Travel_Rarely	1373	Research & Development	2
3	33	No	Travel_Frequently	1392	Research & Development	3
4	27	No	Travel_Rarely	591	Research & Development	2

5 rows × 35 columns

In [3]:
```
print("The mean of monthly income is :",df.loc[:,"MonthlyIncome"].mean())
print("The mean of age is :",df.loc[:,"Age"].mean())
```

```
The mean of monthly income is : 6502.931292517007
The mean of age is : 36.923809523809524
```

In [5]:
```
print("The median of monthly income is :",df.loc[:,"MonthlyIncome"].median()
print("The median of age is :",df.loc[:,"Age"].median())
```

```
The median of monthly income is : 4919.0
The median of age is : 36.0
```

In [7]:
```
print("The mode of monthly income is :",df.loc[:,"MonthlyIncome"].mode())
print("The mode of age is :",df.loc[:,"Age"].mode())
```

```
The mode of monthly income is : 0    2342
Name: MonthlyIncome, dtype: int64
The mode of age is : 0    35
Name: Age, dtype: int64
```

In [11]:
```
print("The standard deviation of monthly income is:",df.loc[:,"MonthlyIncome
print("The standard deviation of age is :",df.loc[:,"Age"].std())
```

```
The standard deviation of monthly income is: 4707.956783097995
The standard deviation of age is : 9.135373489136734
```

In [13]:
```
array1 = np.array(df['MonthlyIncome'])
array2=np.array(df["Age"])
print("Income",array1)
print("Age array",array2)
print("Maximum income among the employees is :",max(array1))
print("Minimum income among the employees is :",min(array1))
print("Maximum age among the employees is :",max(array2))
print("Minimum age among the employees is :",min(array2))
```

```
Income [5993 5130 2090 ... 6142 5390 4404]
Age array [41 49 37 ... 27 49 34]
Maximum income among the employees is : 19999
Minimum income among the employees is : 1009
Maximum age among the employees is : 60
Minimum age among the employees is : 18
```

In [17]:
```
df.head()
df["BusinessTravel"].replace({"Travel_Rarely":1, "Travel_Frequently":0},inpl
df["Attrition"].replace({ "Yes":1, "No":0}, inplace=True)
df.head()
```

Out[17]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome
0	41	1	1	1102	Sales	1
1	49	0	0	279	Research & Development	8
2	37	1	1	1373	Research & Development	2
3	33	0	0	1392	Research & Development	3
4	27	0	1	591	Research & Development	2

5 rows × 35 columns

In [19]:
```
df.describe()
```

Out[19]:

	Age	Attrition	DailyRate	DistanceFromHome	Educatior
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	0.161224	802.485714	9.192517	2.912925
std	9.135373	0.367863	403.509100	8.106864	1.024165
min	18.000000	0.000000	102.000000	1.000000	1.000000
25%	30.000000	0.000000	465.000000	2.000000	2.000000
50%	36.000000	0.000000	802.000000	7.000000	3.000000
75%	43.000000	0.000000	1157.000000	14.000000	4.000000
max	60.000000	1.000000	1499.000000	29.000000	5.000000

8 rows × 27 columns

In [ ]:

Assignment No- 04

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset.

In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
x=np.array([95,85,80,70,60])
y=np.array([85,95,70,65,70])
model= np.polyfit(x, y, 1)
model
```

Out[4]:
```
array([ 0.64383562, 26.78082192])
```

In [6]:
```
predict = np.poly1d(model)
predict(65)
```

Out[6]:
```
68.63013698630137
```

In [8]:
```
y_pred = predict(x)
y_pred
```

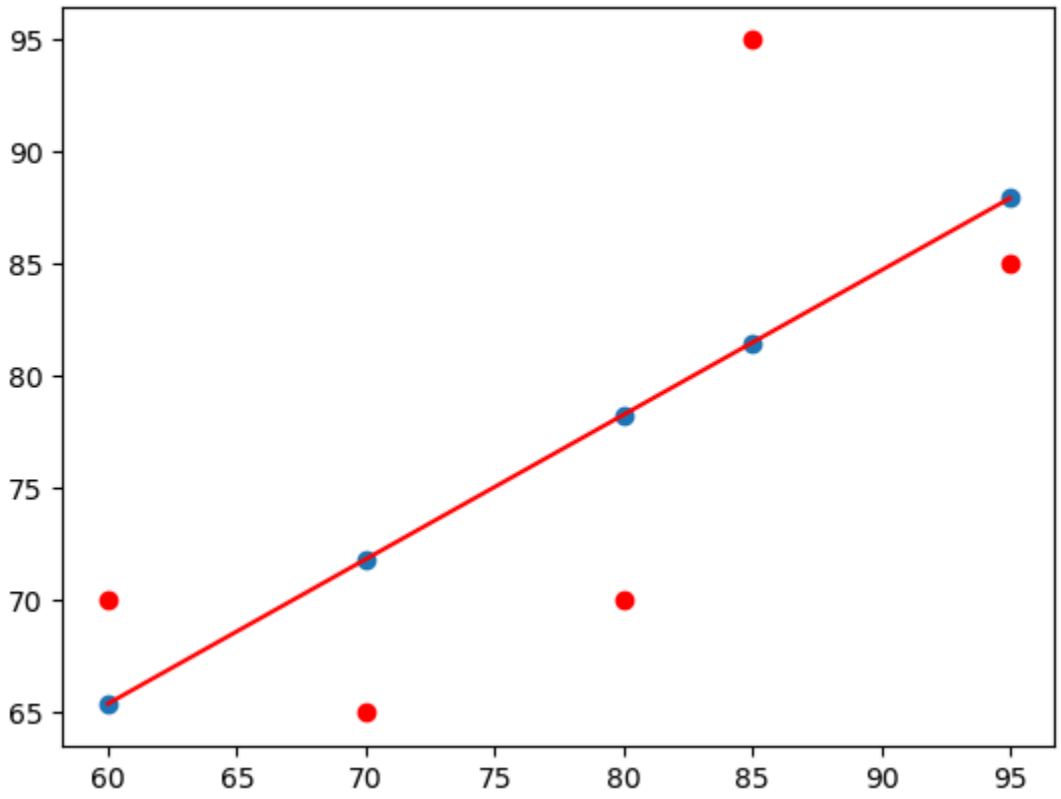
Out[8]:
```
array([87.94520548, 81.50684932, 78.28767123, 71.84931507, 65.4109589 ])
```

In [10]:
```
from sklearn.metrics import r2_score
r2_score(y, y_pred)
```

Out[10]:
```
0.4803218090889326
```

In [12]:
```
y_line = model[1] + model[0]* x
plt.plot(x, y_line, c = 'r')
plt.scatter(x,y_pred)
plt.scatter(x,y,c='r')
```

Out[12]:
```
<matplotlib.collections.PathCollection at 0x1a42cf30ce0>
```



In [14]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [16]:
```
from sklearn.datasets import fetch_california_housing
housing = fetch_california_housing()

housing
```

Out[16]:

```
{'data': array([[   8.3252    ,   41.        ,    6.98412698, ...,    2.555
55556,
           37.88      , -122.23      ],
        [   8.3014    ,   21.        ,    6.23813708, ...,    2.10984183,
           37.86      , -122.22      ],
        [   7.2574    ,   52.        ,    8.28813559, ...,    2.80225989,
           37.85      , -122.24      ],
        ...,
        [   1.7       ,   17.        ,    5.20554273, ...,    2.3256351 ,
           39.43      , -121.22      ],
        [   1.8672    ,   18.        ,    5.32951289, ...,    2.12320917,
           39.43      , -121.32      ],
        [   2.3886    ,   16.        ,    5.25471698, ...,    2.61698113,
           39.37      , -121.24      ]]),
 'target': array([4.526, 3.585, 3.521, ..., 0.923, 0.847, 0.894]),
 'frame': None,
 'target_names': ['MedHouseVal'],
 'feature_names': ['MedInc',
  'HouseAge',
  'AveRooms',
  'AveBedrms',
  'Population',
  'AveOccup',
  'Latitude',
  'Longitude'],
 'DESCR': '.. _california_housing_dataset:\n\nCalifornia Housing dataset\n-
-------------------------\n\n**Data Set Characteristics:**\n\n:Number of In
stances: 20640\n\n:Number of Attributes: 8 numeric, predictive attributes a
nd the target\n\n:Attribute Information:\n    - MedInc        median income
in block group\n    - HouseAge      median house age in block group\n    -
AveRooms      average number of rooms per household\n    - AveBedrms     av
erage number of bedrooms per household\n    - Population    block group pop
ulation\n    - AveOccup      average number of household members\n    - Lat
itude      block group latitude\n    - Longitude     block group longitude
\n\n:Missing Attribute Values: None\n\nThis dataset was obtained from the S
tatLib repository.\nhttps://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housin
g.html\n\nThe target variable is the median house value for California dist
ricts,\nexpressed in hundreds of thousands of dollars ($100,000).\n\nThis d
ataset was derived from the 1990 U.S. census, using one row per census\nblo
ck group. A block group is the smallest geographical unit for which the U.
S.\nCensus Bureau publishes sample data (a block group typically has a popu
lation\nof 600 to 3,000 people).\n\nA household is a group of people residi
ng within a home. Since the average\nnumber of rooms and bedrooms in this d
ataset are provided per household, these\ncolumns may take surprisingly lar
ge values for block groups with few households\nand many empty houses, such
as vacation resorts.\n\nIt can be downloaded/loaded using the\n:func:`sklea
rn.datasets.fetch_california_housing` function.\n\n.. rubric:: References\n
\n- Pace, R. Kelley and Ronald Barry, Sparse Spatial Autoregressions,\n  St
atistics and Probability Letters, 33 (1997) 291-297\n'}
```


In [18]:
```
data = pd.DataFrame(housing.data)
data.columns = housing.feature_names
data.head()
```

Out[18]:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitud
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.8
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.8
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.8
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.8
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.8

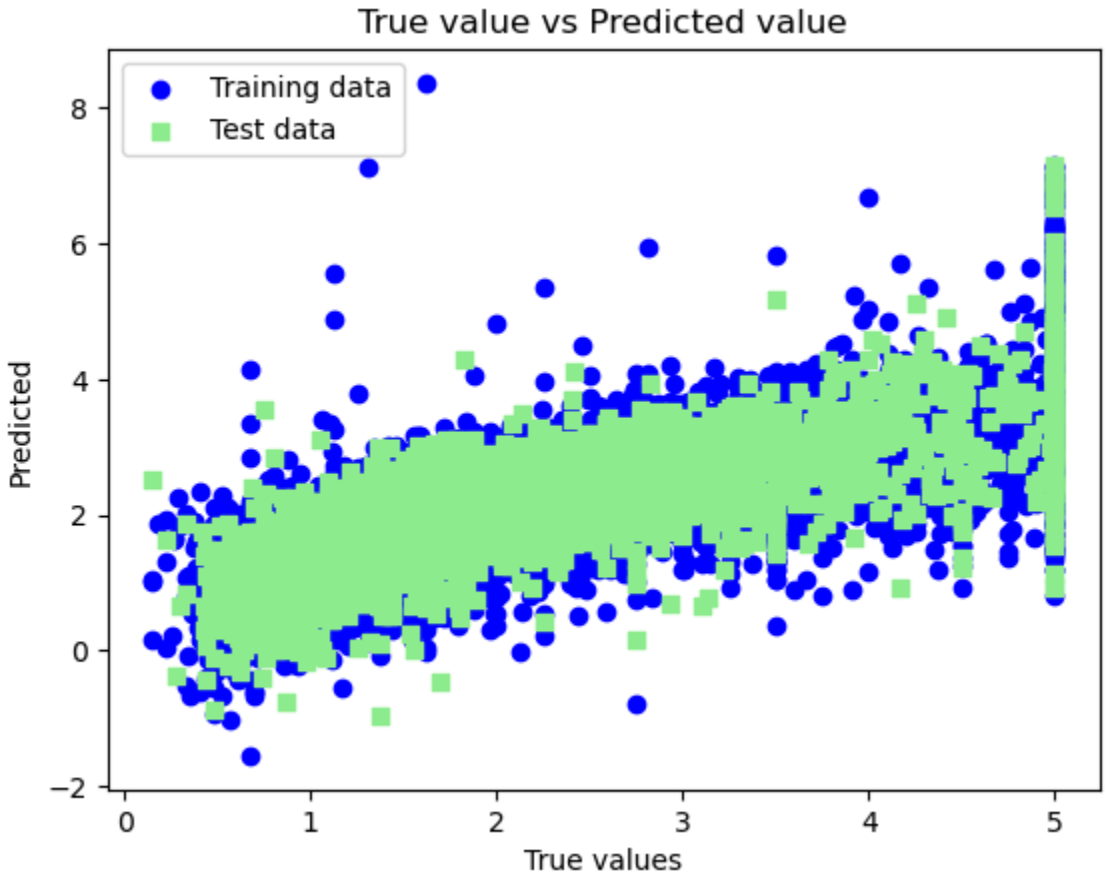
In [20]:
```
data['PRICE'] = housing.target
data.isnull().sum()
```

Out[20]:
```
MedInc        0
HouseAge      0
AveRooms      0
AveBedrms     0
Population    0
AveOccup      0
Latitude      0
Longitude     0
PRICE         0
dtype: int64
```

In [22]:
```
x = data.drop(['PRICE'], axis = 1)
y = data['PRICE']
from sklearn.model_selection import train_test_split
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size
=0.2,random_state = 0)
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
model = lm.fit(xtrain, ytrain)
ytrain_pred = lm.predict(xtrain)
ytest_pred = lm.predict(xtest)
df = pd.DataFrame(ytrain_pred,ytrain)
df = pd.DataFrame(ytest_pred,ytest)
from sklearn.metrics import mean_squared_error, r2_score
mse = mean_squared_error(ytest, ytest_pred)
print(mse)
mse = mean_squared_error(ytrain_pred,ytrain)
print(mse)
```

```
0.5289841670367221
0.5234413607125449
```

In [24]:
```
plt.scatter(ytrain ,ytrain_pred,c='blue',marker='o',label='Training data')
plt.scatter(ytest,ytest_pred,c='lightgreen',marker='s',label='Test data')
plt.xlabel('True values')
plt.ylabel('Predicted')
plt.title("True value vs Predicted value")
plt.legend(loc='upper left')
plt.plot()
plt.show()
```



In [ ]:

Assignment No- 05

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Implement logistic regression using Python/R to perform classification on Social_Network_Ads.csv dataset.

In [1]:
```
import pandas as pd # Data handling
import numpy as np # Numerical operations
import matplotlib.pyplot as plt # Data visualization
from sklearn.model_selection import train_test_split # Train-test split
from sklearn.preprocessing import StandardScaler # Feature scaling
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,precision_score
from sklearn.preprocessing import LabelEncoder
```

In [3]:
```
df=pd.read_csv("Downloads/Social_Network_Ads.csv")
df.head()
```

Out[3]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

In [5]:
```
label_encoder = LabelEncoder()
df["Gender"] = label_encoder.fit_transform(df["Gender"])
df.head()
```

Out[5]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	1	19	19000	0
1	15810944	1	35	20000	0
2	15668575	0	26	43000	0
3	15603246	0	27	57000	0
4	15804002	1	19	76000	0

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
User ID            0
Gender             0
Age                0
EstimatedSalary    0
Purchased          0
dtype: int64
```

In [9]:
```
df.cov()
```

Out[9]:

	User ID	Gender	Age	EstimatedSalary
User ID	5.134915e+09	-905.617719	-541.682870	1.737143e+08
Gender	-9.056177e+02	0.250526	-0.386917	-1.031404e+03
Age	-5.416829e+02	-0.386917	109.890702	5.548738e+04
EstimatedSalary	1.737143e+08	-1031.403509	55487.380952	1.162603e+09
Purchased	2.448363e+02	-0.010201	3.131165	5.924367e+03

In [11]:
```
X = df.drop(columns=["Purchased"]) # Assuming "Purchased" is the targetvaria
Y = df["Purchased"]
xtrain, xtest, ytrain, ytest = train_test_split(X, Y, test_size=0.2,
random_state=42)
from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
features_scaled = scaler.fit_transform(df)
df_normalized = pd.DataFrame(features_scaled, columns = df.columns)
df_normalized
```

Out[11]:

	User ID	Gender	Age	EstimatedSalary	Purchased
0	0.232636	1.0	0.023810	0.029630	0.0
1	0.982732	1.0	0.404762	0.037037	0.0
2	0.409926	0.0	0.190476	0.207407	0.0
3	0.147083	0.0	0.214286	0.311111	0.0
4	0.954801	1.0	0.023810	0.451852	0.0
...	...	...	...	...	...
395	0.503623	0.0	0.666667	0.192593	1.0
396	0.560787	1.0	0.785714	0.059259	1.0
397	0.352477	0.0	0.761905	0.037037	1.0
398	0.757720	1.0	0.428571	0.133333	0.0
399	0.110048	0.0	0.738095	0.155556	1.0

400 rows × 5 columns

In [13]:
```
logreg = LogisticRegression()
logreg.fit(xtrain, ytrain)
```

Out[13]:

```
▼ LogisticRegression
LogisticRegression()
```

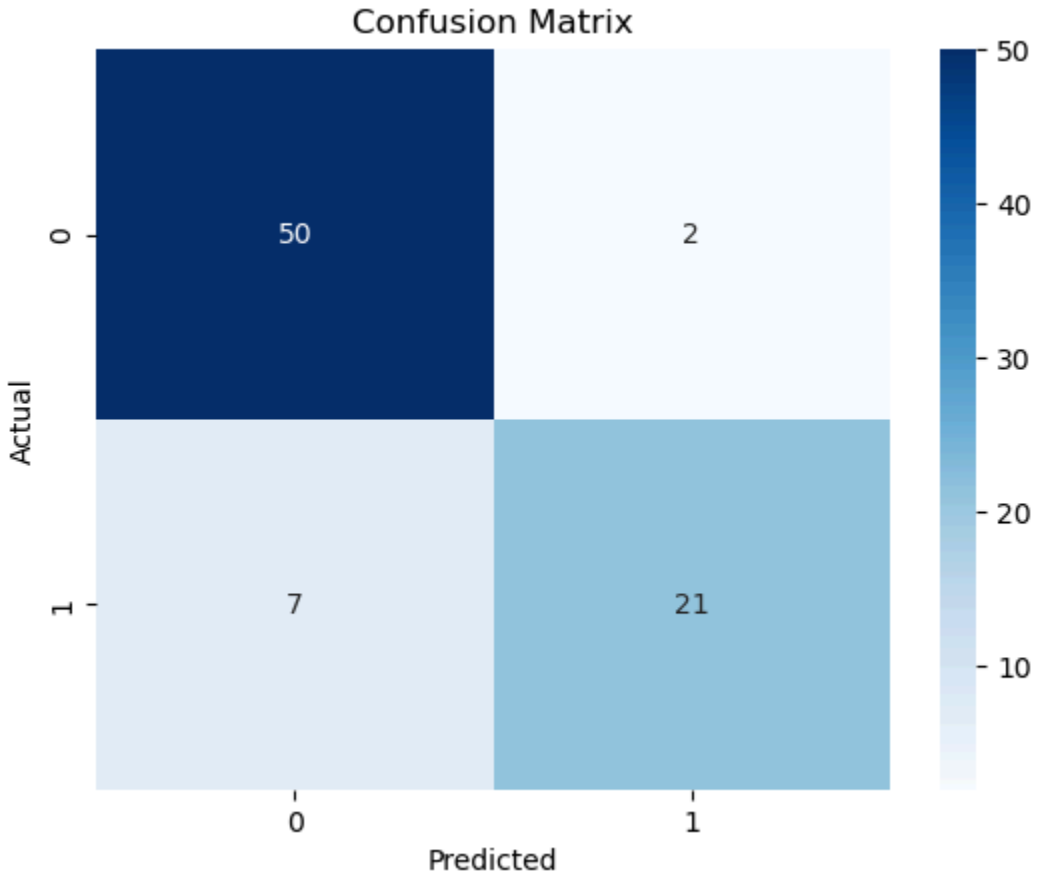
In [15]:

```
y_pred_train = logreg.predict(xtrain)
y_pred_test = logreg.predict(xtest)
train_acc = accuracy_score(ytrain, y_pred_train)
test_acc = accuracy_score(ytest, y_pred_test)
cm = confusion_matrix(ytest, y_pred_test)
precision = precision_score(ytest, y_pred_test)
recall = recall_score(ytest, y_pred_test)
print("Training Accuracy:", train_acc)
print("Testing Accuracy:", test_acc)
print("Confusion Matrix:\n", cm)
print("Precision:", precision)
print("Recall:", recall)
```

```
Training Accuracy: 0.840625
Testing Accuracy: 0.8875
Confusion Matrix:
 [[50  2]
 [ 7 21]]
Precision: 0.9130434782608695
Recall: 0.75
```

In [17]:

```
import seaborn as sns
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```



In [ ]:

Assignment No- 06

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title- Naive Bayes Classificaation Algorithm

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import urllib.request
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, confusion_matrix,precision_score
import seaborn as sns
```

In [3]:
```
url ="https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.da

column_names = ["Sepal_Length", "Sepal_Width", "Petal_Length", "Petal_Width"
"Class"]
iris = pd.read_csv(url, names=column_names)
```

In [5]:
```
iris.head()
```

Out[5]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

In [7]:
```
label_encoder = LabelEncoder()
iris["Class"] = label_encoder.fit_transform(iris["Class"])
```

In [9]:
```
iris.head()
```

Out[9]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

In [11]:
```
iris.isnull().sum()
```

Out[11]:
```
Sepal_Length    0
Sepal_Width     0
Petal_Length    0
Petal_Width     0
Class           0
dtype: int64
```

In [13]:
```
X = iris.drop(columns=["Class"])
Y = iris["Class"]
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,
random_state=42)
```

In [15]:
```
from sklearn import preprocessing
scaler = preprocessing.MinMaxScaler()
features_scaled = scaler.fit_transform(X)
iris_normalized = pd.DataFrame(features_scaled, columns=X.columns)
iris_normalized["Class"] = Y
iris_normalized.head()
```

Out[15]:

	Sepal_Length	Sepal_Width	Petal_Length	Petal_Width	Class
0	0.222222	0.625000	0.067797	0.041667	0
1	0.166667	0.416667	0.067797	0.041667	0
2	0.111111	0.500000	0.050847	0.041667	0
3	0.083333	0.458333	0.084746	0.041667	0
4	0.194444	0.666667	0.067797	0.041667	0

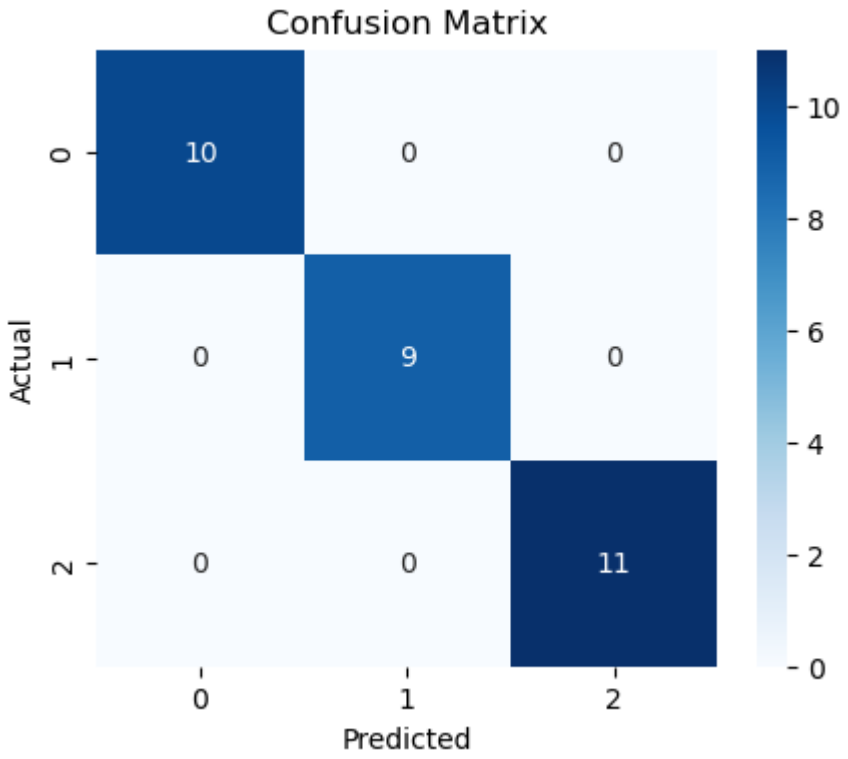
In [17]:
```
naive_bayes = GaussianNB()
naive_bayes.fit(X_train, y_train)
```

Out[17]:
```
▼ GaussianNB
GaussianNB()
```

In [19]:
```
y_pred_train = naive_bayes.predict(X_train)
y_pred_test = naive_bayes.predict(X_test)
train_accuracy = accuracy_score(y_train, y_pred_train)
test_accuracy = accuracy_score(y_test, y_pred_test)
train_accuracy = accuracy_score(y_train, y_pred_train)
test_accuracy = accuracy_score(y_test, y_pred_test)
precision = precision_score(y_test, y_pred_test, average="micro")
recall = recall_score(y_test, y_pred_test, average="micro")
cm = confusion_matrix(y_test, y_pred_test)
print("\nTraining Accuracy:", train_accuracy)
print("Testing Accuracy:", test_accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("\nConfusion Matrix:\n", cm)
```

```
Training Accuracy: 0.95
Testing Accuracy: 1.0
Precision: 1.0
Recall: 1.0

Confusion Matrix:
 [[10  0  0]
 [ 0  9  0]
 [ 0  0 11]]
```

In [21]:
```
plt.figure(figsize=(5, 4))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues")
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```



In [ ]:
```
```

Assignment No- 07

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title-Tokenization,POS Tagging,stop words removal,Stemming and Lemmatization.

In [1]:
```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
```

Out[1]: True

In [3]:
```
text= "Tokenization is the first step in text analytics. The process ofbreak
```

In [7]:
```
import nltk
nltk.download('punkt_tab')

from nltk.tokenize import sent_tokenize
tokenized_text= sent_tokenize(text)
print(tokenized_text)
from nltk.tokenize import word_tokenize
tokenized_word=word_tokenize(text)
print(tokenized_word)
```

```
[nltk_data] Downloading package punkt_tab to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt_tab.zip.
```

```
['Tokenization is the first step in text analytics.', 'The process ofbreakin
g down a text paragraph into smaller chunksuch as words or sentencesis calle
d Tokenization.']
['Tokenization', 'is', 'the', 'first', 'step', 'in', 'text', 'analytics',
'.', 'The', 'process', 'ofbreaking', 'down', 'a', 'text', 'paragraph', 'int
o', 'smaller', 'chunksuch', 'as', 'words', 'or', 'sentencesis', 'called', 'T
okenization', '.']
```

In [9]:

```
from nltk.corpus import stopwords
import re
stop_words=set(stopwords.words("english"))
print(stop_words)
text= "How to remove stop words with NLTK library in Python?"
text= re.sub('[^a-zA-Z]', ' ',text)
tokens = word_tokenize(text.lower())
filtered_text=[]
for w in tokens:
 if w not in stop_words:
  filtered_text.append(w)
print("Tokenized Sentence:",tokens)
print("Filterd Sentence:",filtered_text)
```

```
{"we'll", 'if', "they'll", 'into', "weren't", 'here', 'only', 't', 'too', "w
ouldn't", 'weren', 'but', 'shan', "they've", 'did', 'after', 'this', 'unde
r', 'hers', 'itself', "they're", 'can', 'wouldn', 'any', "haven't", 'off',
"won't", 'about', 'an', 'nor', 'hadn', 'couldn', 'she', 'what', 'each', 'o',
'were', 'i', "they'd", 'where', 'a', 'they', 'was', "that'll", 'or', 'are',
"wasn't", 'further', 'myself', 'isn', 'shouldn', 'of', 'with', "we're", "h
e'll", 'very', 'won', 'don', 'which', 'yourself', 'whom', 'll', "you'd", 'ai
n', 'been', 'it', 'hasn', 'as', 're', 'again', 'y', 'mustn', 'yours', 'befor
e', 'not', 'because', 'when', 'we', 'while', 'that', 's', 'same', 'being',
'wasn', 'aren', 'doing', "hadn't", 'some', 'm', "it'll", "shan't", 'down',
"it'd", "you're", 'has', 'themselves', 'above', 'have', "mightn't", 'than',
'those', 'and', 'at', 'needn', 'until', 'below', 'both', 'his', "he'd", 'sho
uld', "needn't", 'who', 'in', 'having', 'by', 'doesn', 'your', 'ma', 'throug
h', 'will', 'no', 'herself', "mustn't", 'why', 'am', "she's", 'against', "is
n't", 'once', 'himself', 'its', "couldn't", 'then', 'he', 'yourselves', 'd',
'ours', "you've", 'up', 'own', "he's", 'them', 'from', 'just', 'now', "you'l
l", "i'll", 'is', 'during', 'so', "shouldn't", 'theirs', 'these', 'had', "sh
e'd", 'there', 've', "aren't", 'be', 'our', 'for', "we've", 'out', 'to', 'ou
rselves', "it's", 'their', "she'll", 'more', 'the', "i've", 'between', 'me',
'my', "i'm", "didn't", "we'd", 'do', 'does', 'on', 'him', 'few', 'such', 'ho
w', "hasn't", 'other', 'over', "should've", 'you', 'her', 'all', 'haven', 'd
idn', "doesn't", "don't", "i'd", 'mightn', 'most'}
Tokenized Sentence: ['how', 'to', 'remove', 'stop', 'words', 'with', 'nltk',
'library', 'in', 'python']
Filterd Sentence: ['remove', 'stop', 'words', 'nltk', 'library', 'python']
```

In [11]:

```
from nltk.stem import PorterStemmer
e_words= ["wait", "waiting", "waited", "waits"]
ps=PorterStemmer()
for w in e_words:
    rootWord=ps.stem(w)
    print(rootWord)
```

```
wait
wait
wait
wait
```

In [13]:
```
nltk.download('omw-1.4')
from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()
text = "studies studying cries cry"
tokenization = nltk.word_tokenize(text)
for w in tokenization:
    print(f"Lemma for '{w}' is '{wordnet_lemmatizer.lemmatize(w)}'")
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
Lemma for 'studies' is 'study'
Lemma for 'studying' is 'studying'
Lemma for 'cries' is 'cry'
Lemma for 'cry' is 'cry'
```

In [17]:
```
import nltk
nltk.download('averaged_perceptron_tagger_eng')
import nltk
from nltk.tokenize import word_tokenize
data="The pink sweater fit her perfectly"
words=word_tokenize(data)
for word in words:
    print(nltk.pos_tag([word]))
```

```
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     C:\Users\HP.DESKTOP-
[nltk_data]     853SQC2\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger_eng.zip.
[('The', 'DT')]
[('pink', 'NN')]
[('sweater', 'NN')]
[('fit', 'NN')]
[('her', 'PRP$')]
[('perfectly', 'RB')]
```

In [19]:
```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
documentA = 'Jupiter is the largest Planet'
documentB = 'Mars is the fourth planet from the Sun'
bagOfWordsA = documentA.split(' ')
bagOfWordsB = documentB.split(' ')
uniqueWords = set(bagOfWordsA).union(set(bagOfWordsB))
numOfWordsA = dict.fromkeys(uniqueWords, 0)
for word in bagOfWordsA:
    numOfWordsA[word] += 1
    numOfWordsB = dict.fromkeys(uniqueWords, 0)
    for word in bagOfWordsB:
        numOfWordsB[word] += 1
def computeTF(wordDict, bagOfWords):
    tfDict = {}
```

```
    bagOfWordsCount = len(bagOfWords)
    for word, count in wordDict.items():
        tfDict[word] = count / float(bagOfWordsCount)
        return tfDict
tfA = computeTF(numOfWordsA, bagOfWordsA)
tfB = computeTF(numOfWordsB, bagOfWordsB)
import math
def computeIDF(documents):
    N = len(documents)
    idfDict = dict.fromkeys(documents[0].keys(), 0.0)
    for document in documents:
        for word, val in document.items():
            if val > 0:
                idfDict[word] += 1
    for word, val in idfDict.items():
        idfDict[word] = math.log(N / float(val))
    return idfDict
idfs = computeIDF([numOfWordsA, numOfWordsB])
idfs
```

Out[19]:

```
{'largest': 0.6931471805599453,
 'from': 0.6931471805599453,
 'the': 0.0,
 'fourth': 0.6931471805599453,
 'Sun': 0.6931471805599453,
 'Mars': 0.6931471805599453,
 'is': 0.0,
 'planet': 0.6931471805599453,
 'Planet': 0.6931471805599453,
 'Jupiter': 0.6931471805599453}
```

In [21]:

```
def computeTFIDF(tfBagOfWords, idfs):
    tfidf = {}
    for word, val in tfBagOfWords.items():
        tfidf[word] = val * idfs[word]
    return tfidf
tfidfA = computeTFIDF(tfA, idfs)
tfidfB = computeTFIDF(tfB, idfs)
df = pd.DataFrame([tfidfA, tfidfB])
df
```

Out[21]:

	largest
0	0.138629
1	0.000000

In [ ]:

Assignment No- 08

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title- Data Visualization I

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
dataset = sns.load_dataset('titanic')
dataset.head()
```

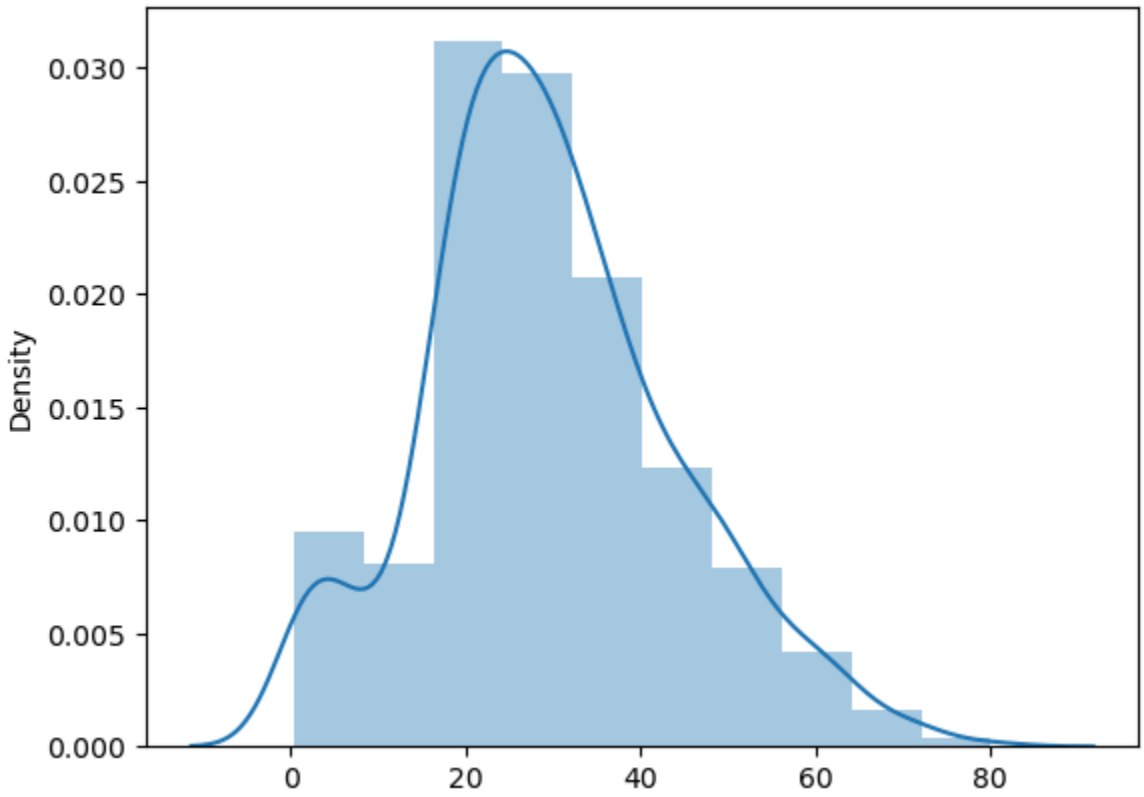
Out[4]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	w
0	0	3	male	22.0	1	0	7.2500	S	Third	m
1	1	1	female	38.0	1	0	71.2833	C	First	wom
2	1	3	female	26.0	0	0	7.9250	S	Third	wom
3	1	1	female	35.0	1	0	53.1000	S	First	wom
4	0	3	male	35.0	0	0	8.0500	S	Third	m

In [6]:
```
sns.distplot(x = dataset['age'], bins = 10)
```

```
C:\Users\HP.DESKTOP-853SQC2\AppData\Local\Temp\ipykernel_3212\3209197554.py:
1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(x = dataset['age'], bins = 10)
```

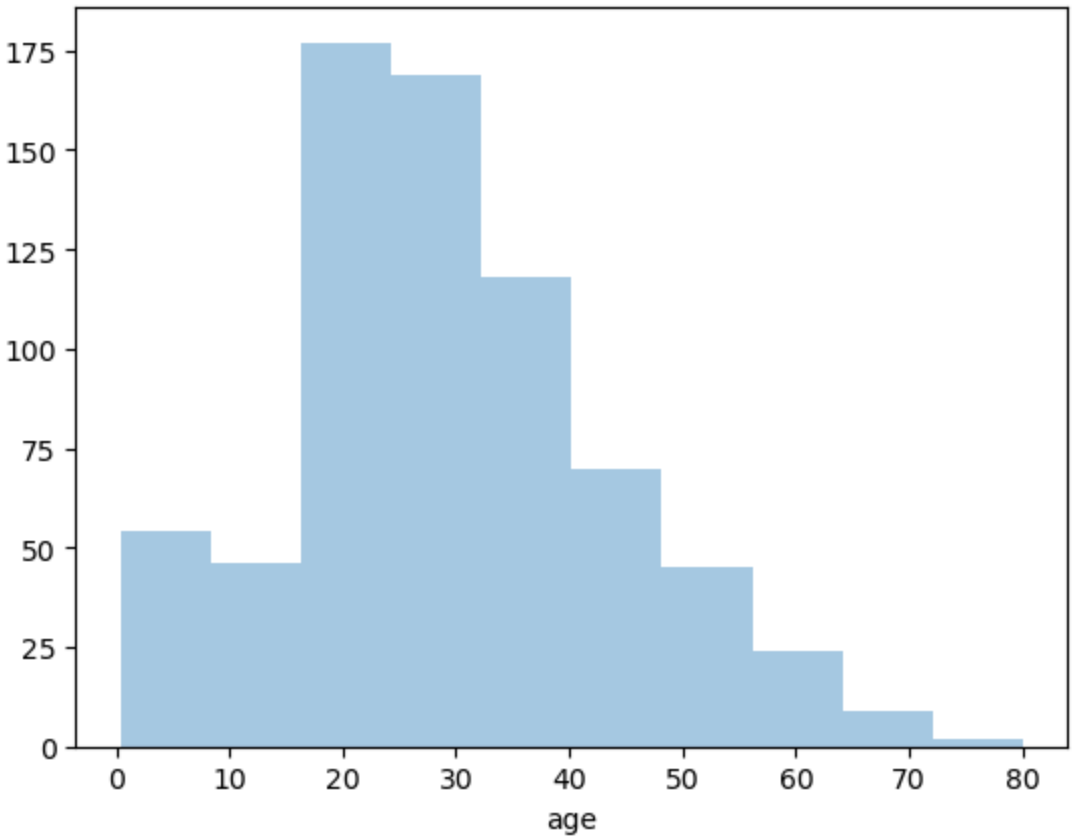
Out[6]: <Axes: ylabel='Density'>



In [8]:
```
sns.distplot(dataset['age'], bins = 10,kde=False)
```

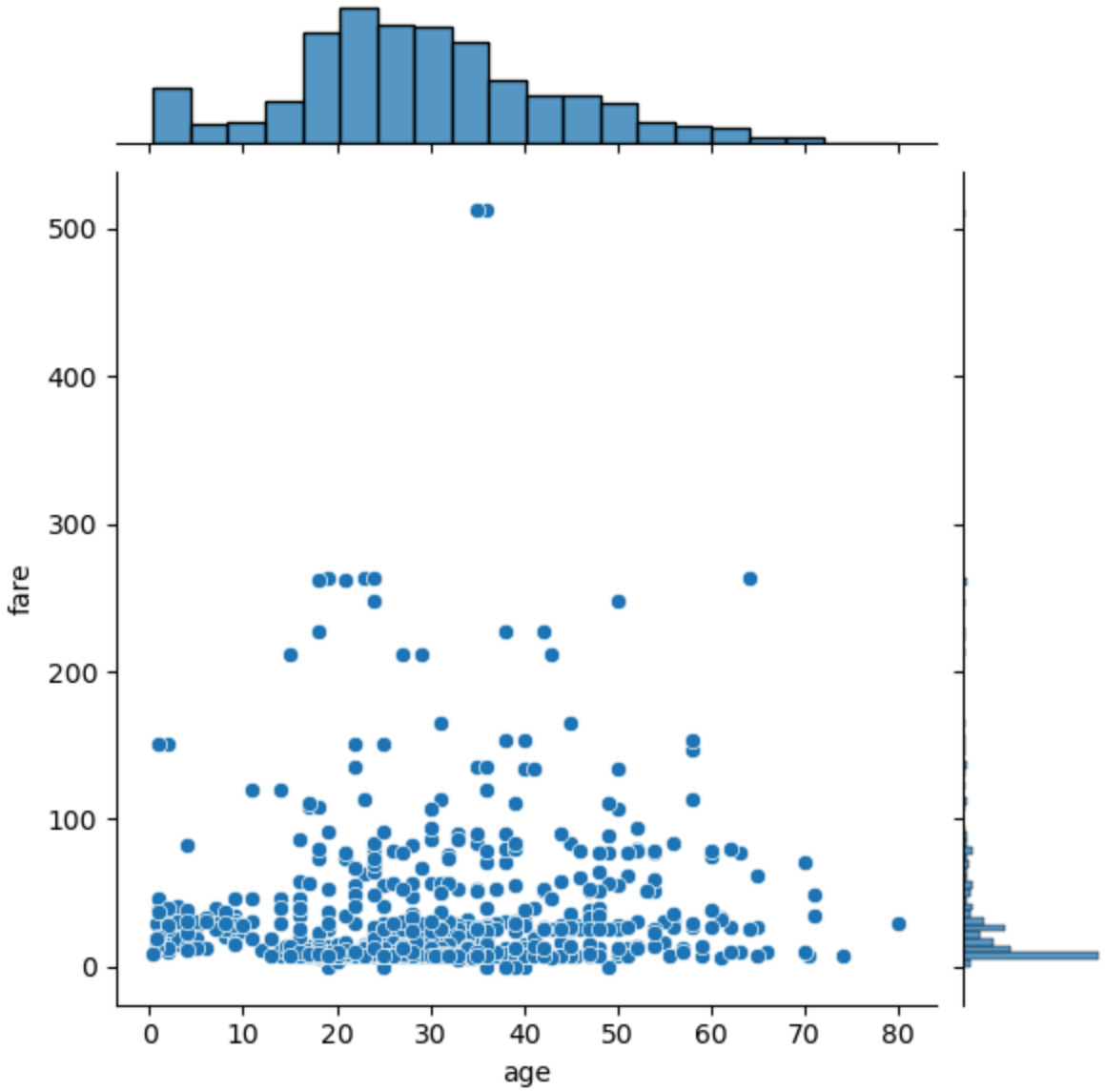
```
C:\Users\HP.DESKTOP-853SQC2\AppData\Local\Temp\ipykernel_3212\3517108427.py:
1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dataset['age'], bins = 10,kde=False)
```

Out[8]: <Axes: xlabel='age'>



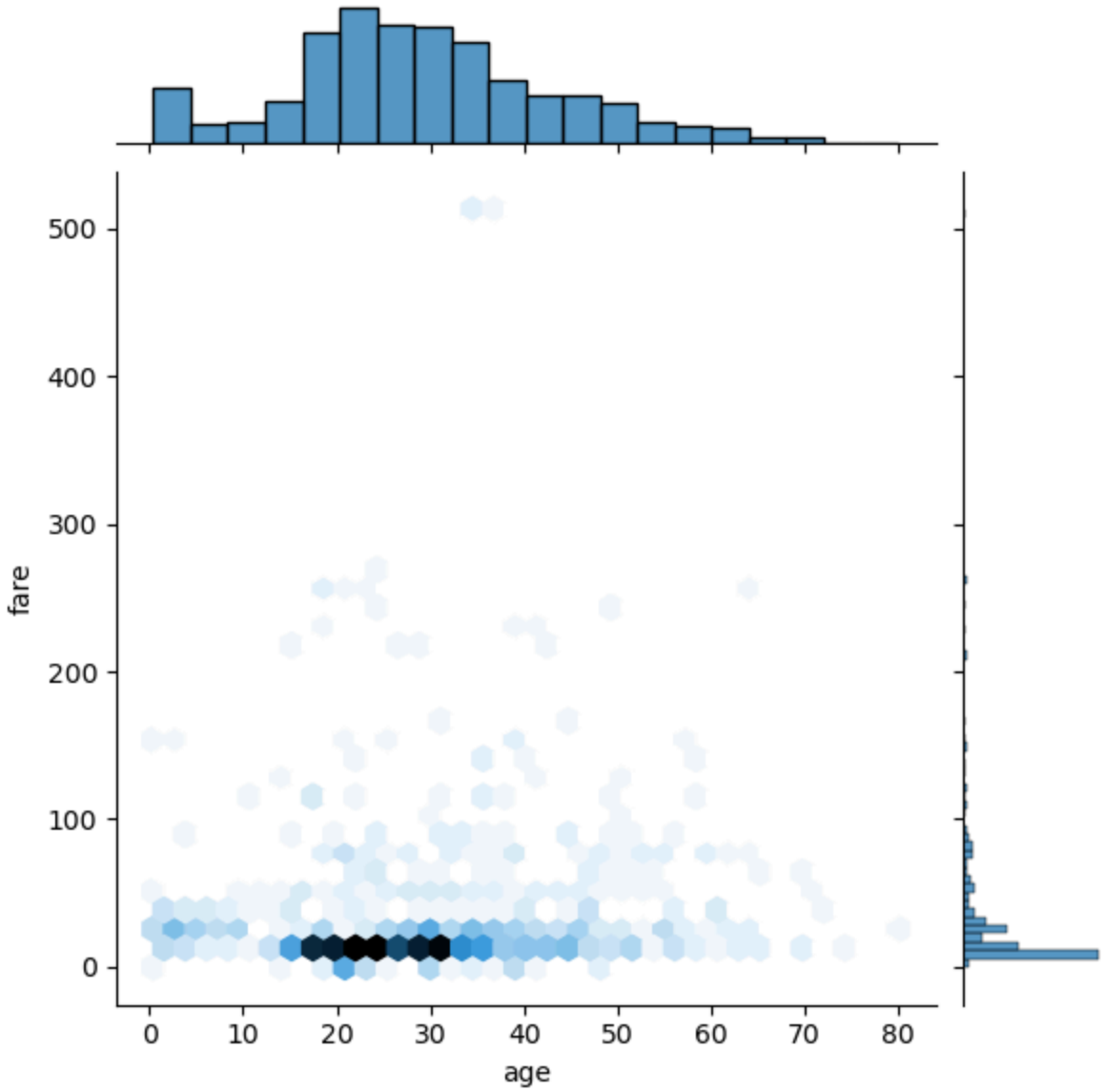
In [10]:
```
sns.jointplot(x = dataset['age'], y = dataset['fare'], kind ='scatter')
```

Out[10]:
```
<seaborn.axisgrid.JointGrid at 0x269a3502ae0>
```



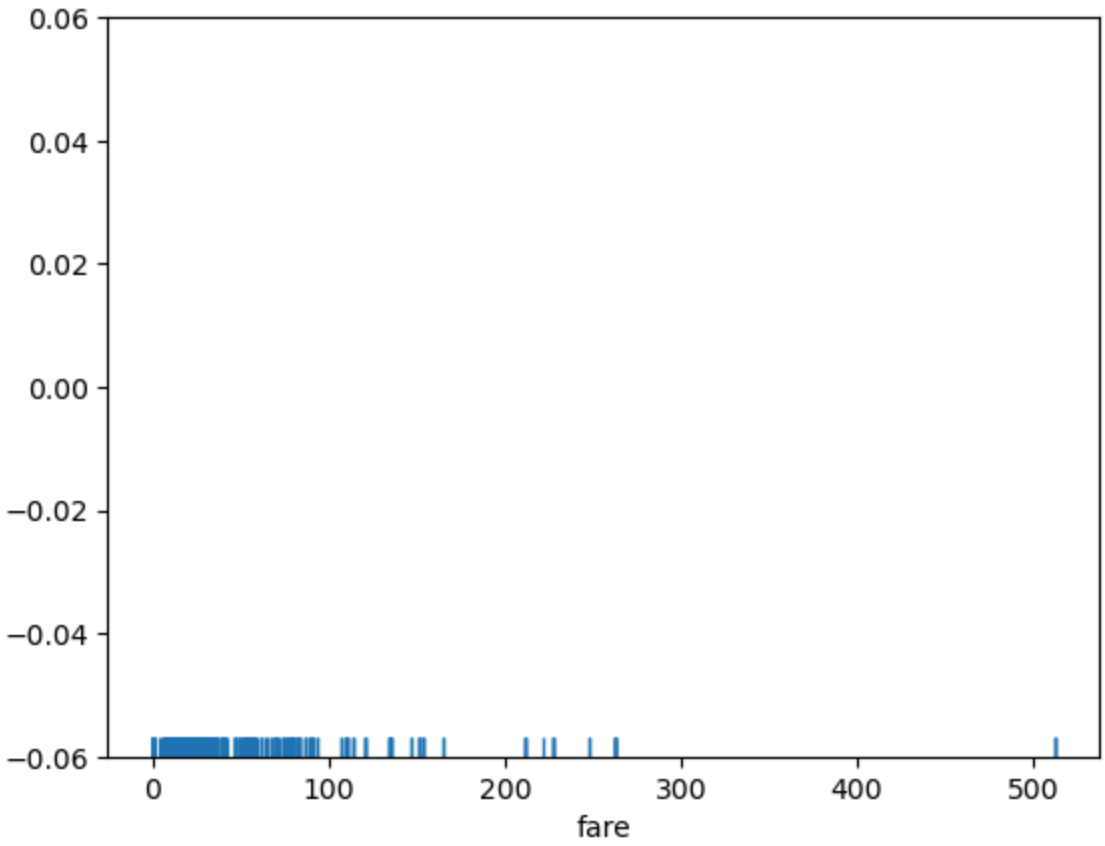
In [12]:
```
sns.jointplot(x = dataset['age'], y = dataset['fare'], kind = 'hex')
```

Out[12]: <seaborn.axisgrid.JointGrid at 0x269a3543530>



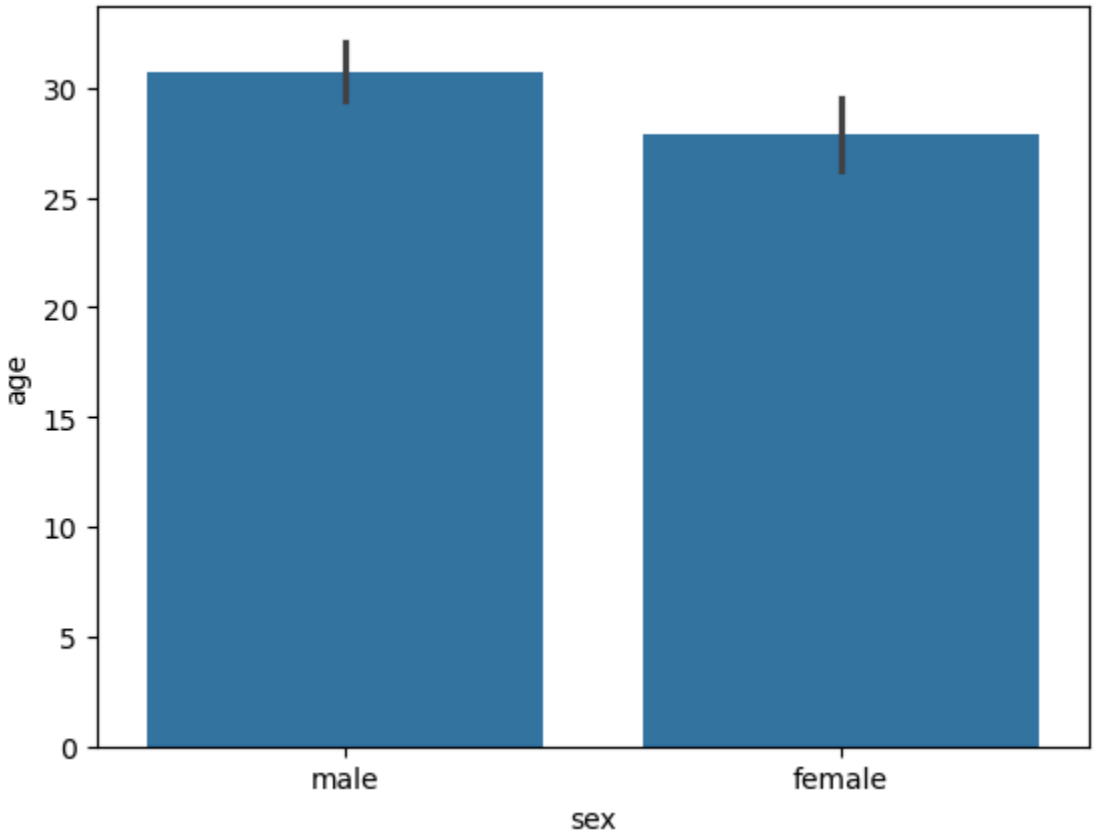
In [15]:
```
sns.rugplot(dataset['fare'])
```

Out[15]: <Axes: xlabel='fare'>



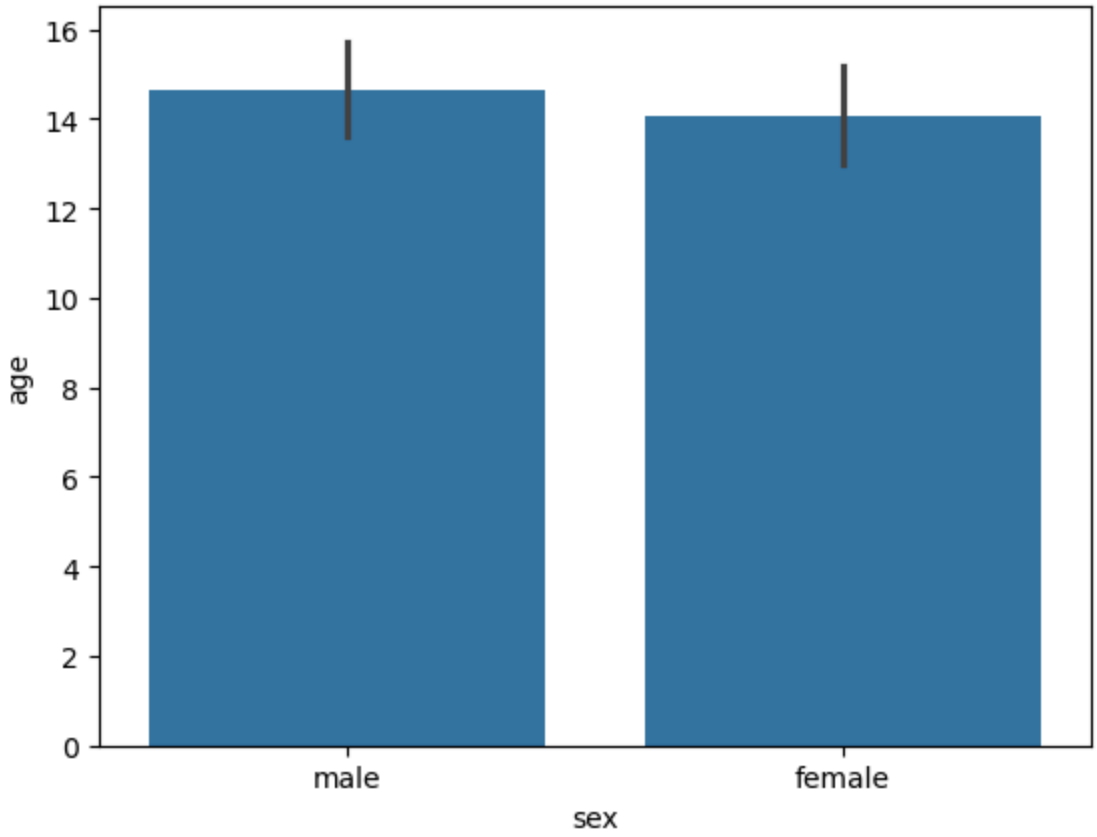
In [17]:
```
sns.barplot(x='sex', y='age', data=dataset)
```

Out[17]: <Axes: xlabel='sex', ylabel='age'>



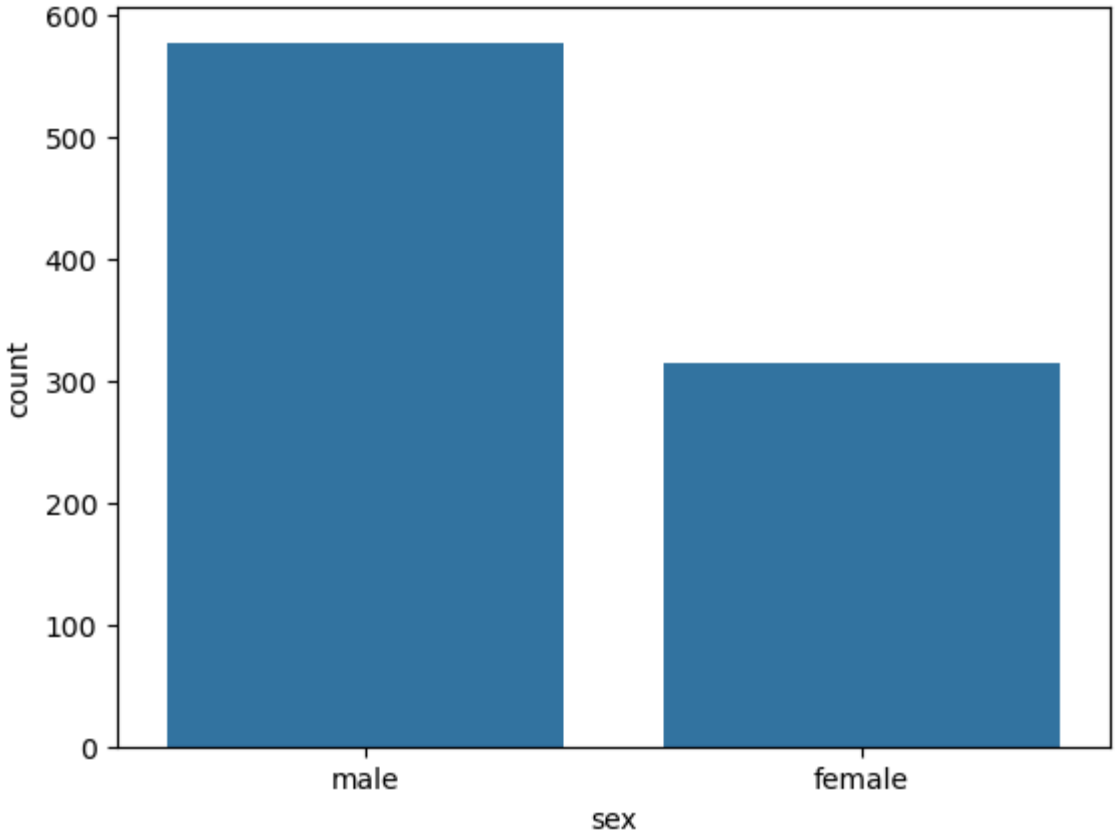
In [19]:
```
sns.barplot(x='sex', y='age', data=dataset, estimator=np.std)
```

Out[19]: <Axes: xlabel='sex', ylabel='age'>



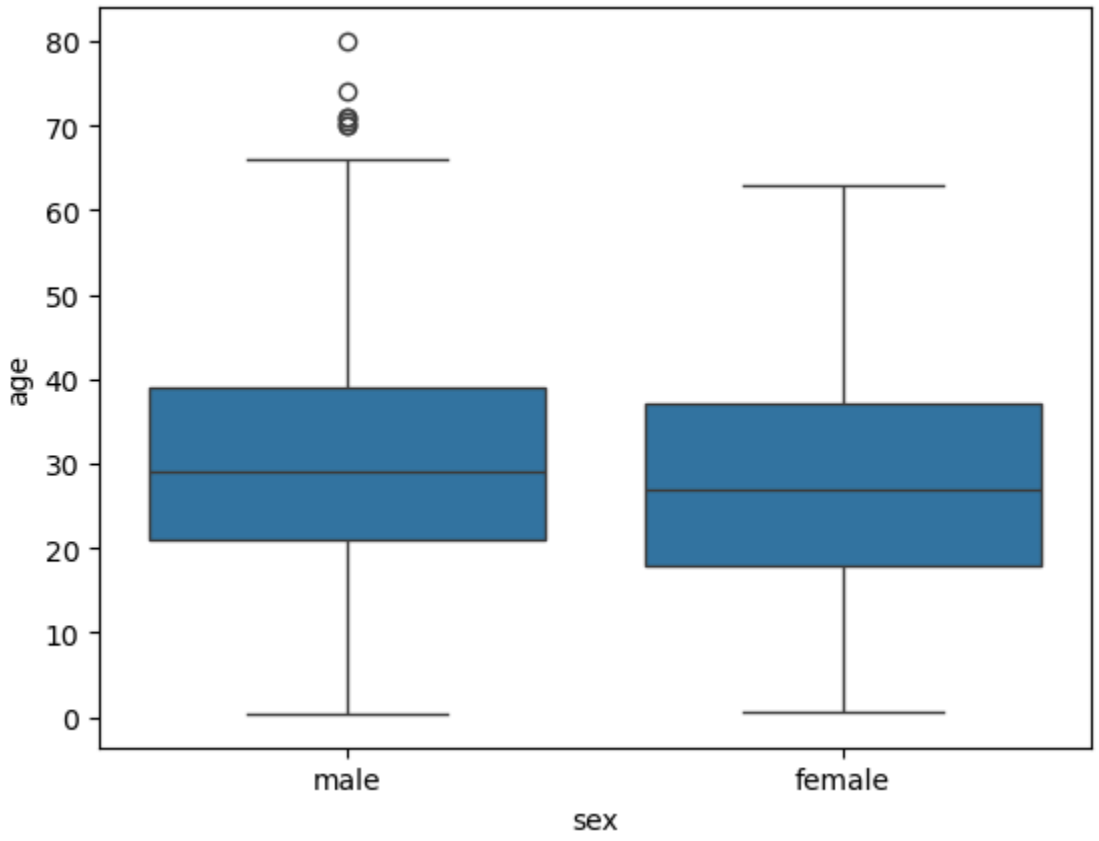
In [21]:
```
sns.countplot(x='sex', data=dataset)
```

Out[21]: <Axes: xlabel='sex', ylabel='count'>



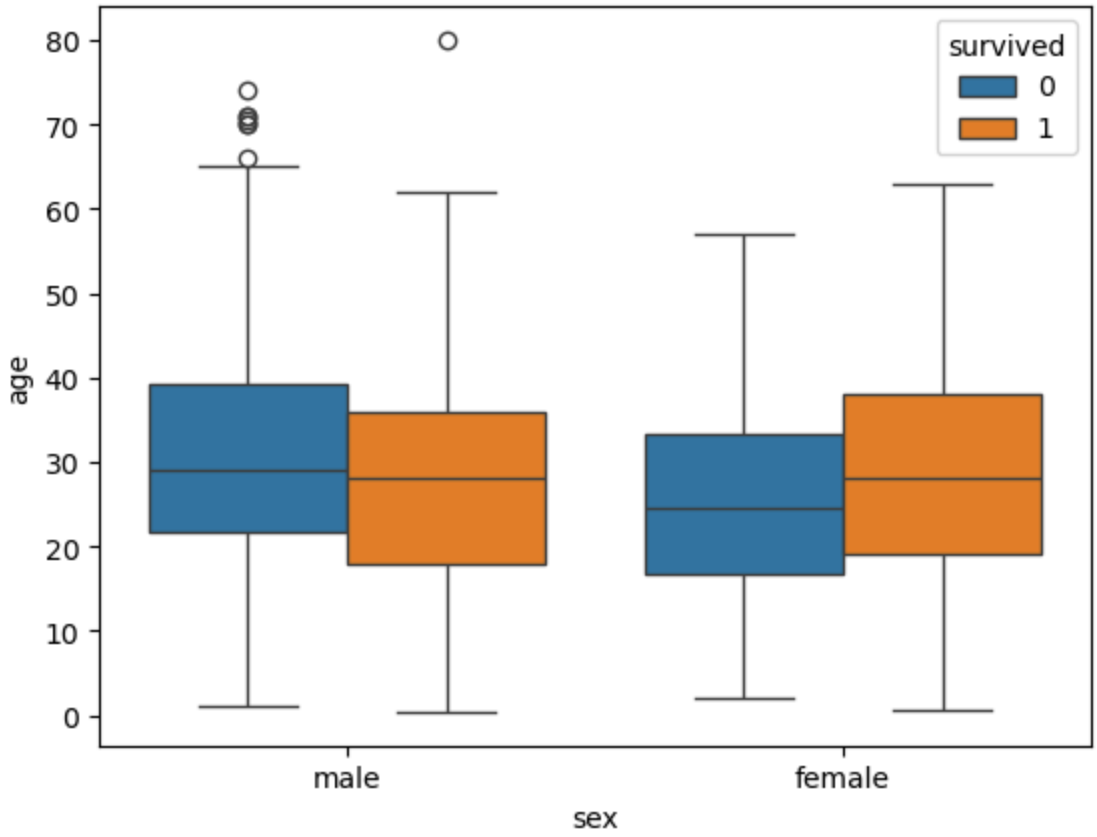
In [23]:
```
sns.boxplot(x='sex', y='age', data=dataset)
```

Out[23]: <Axes: xlabel='sex', ylabel='age'>



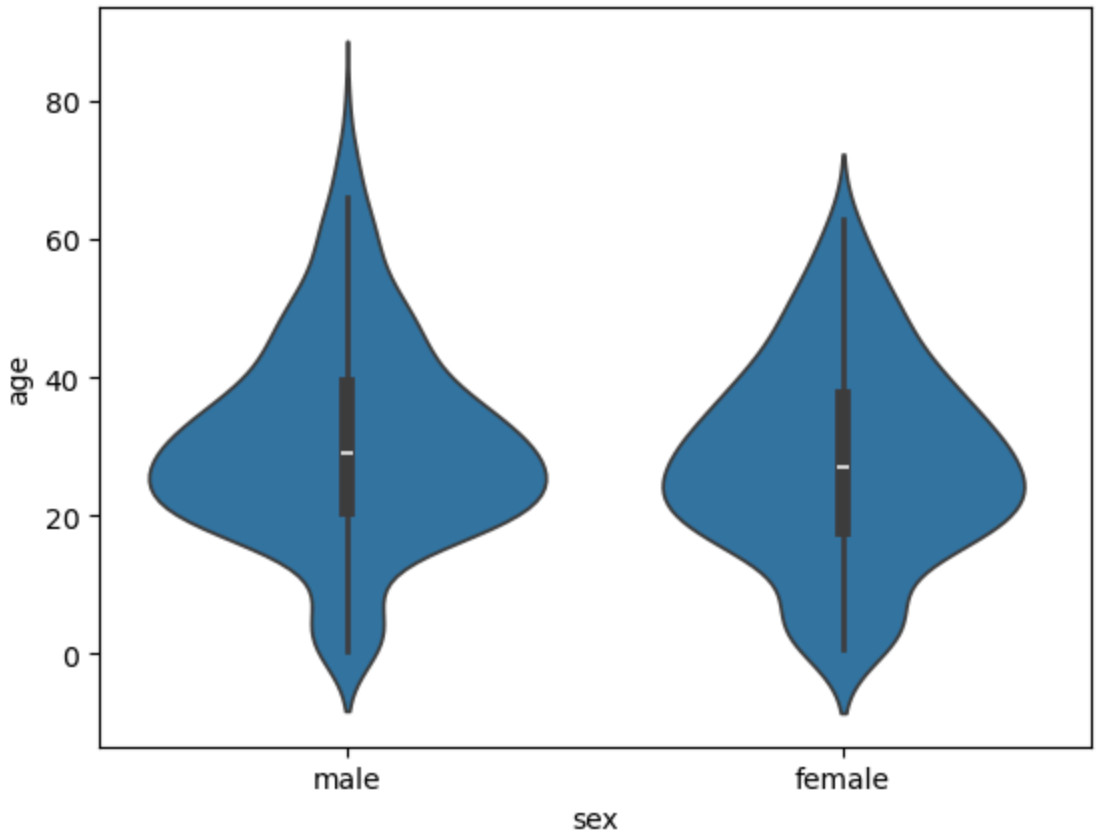
In [25]:
```
sns.boxplot(x='sex', y='age', data=dataset, hue="survived")
```

Out[25]: <Axes: xlabel='sex', ylabel='age'>



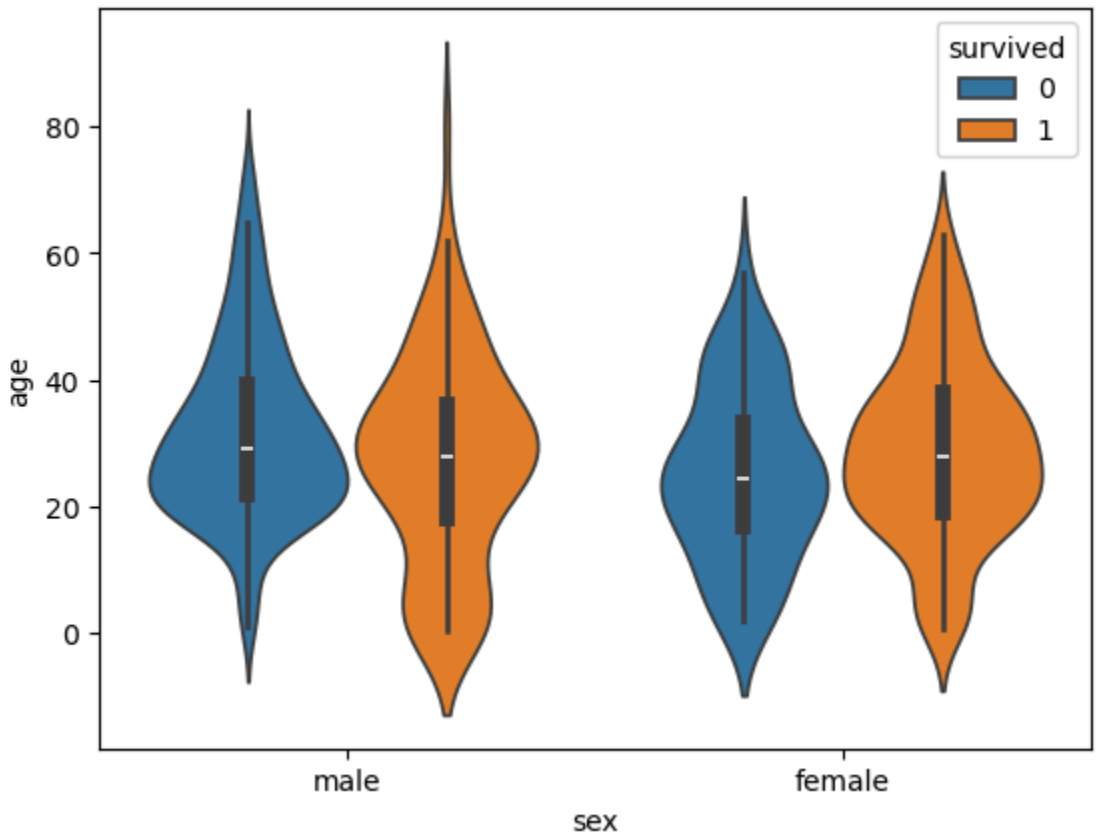
In [27]:
```
sns.violinplot(x='sex', y='age', data=dataset)
```

Out[27]: <Axes: xlabel='sex', ylabel='age'>



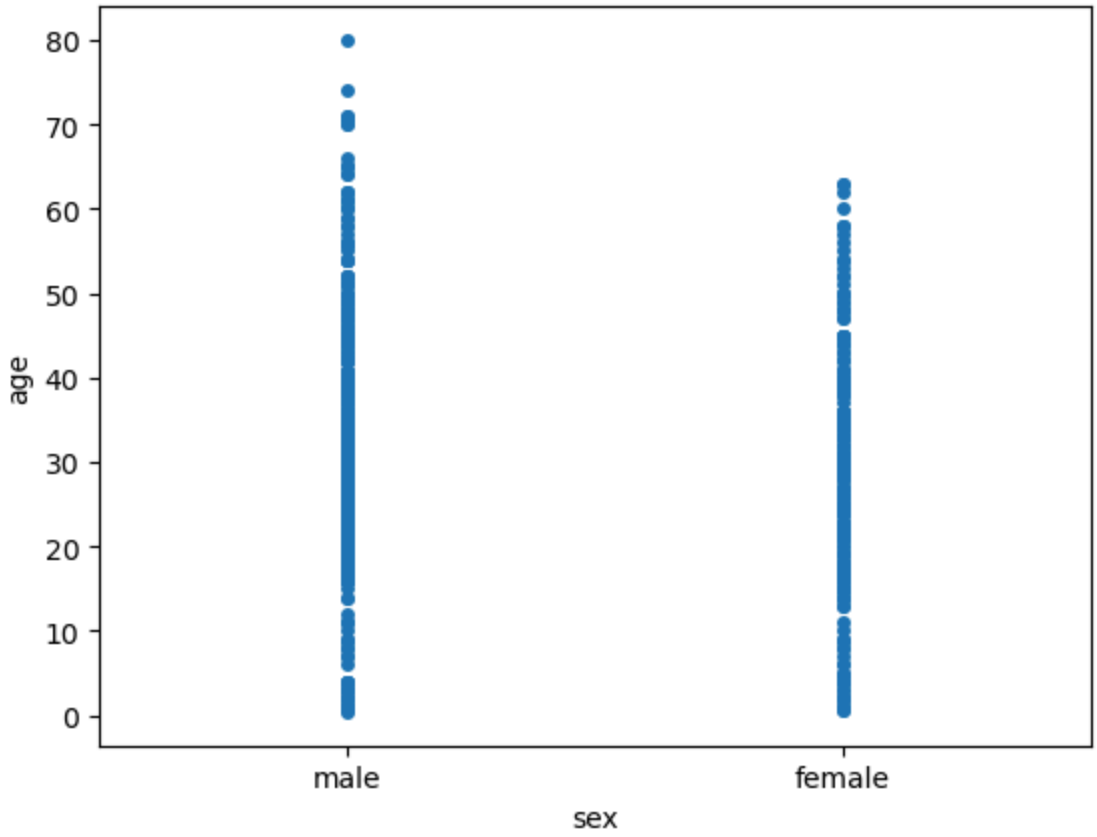
In [29]:
```
sns.violinplot(x='sex', y='age', data=dataset, hue='survived')
```

Out[29]: <Axes: xlabel='sex', ylabel='age'>



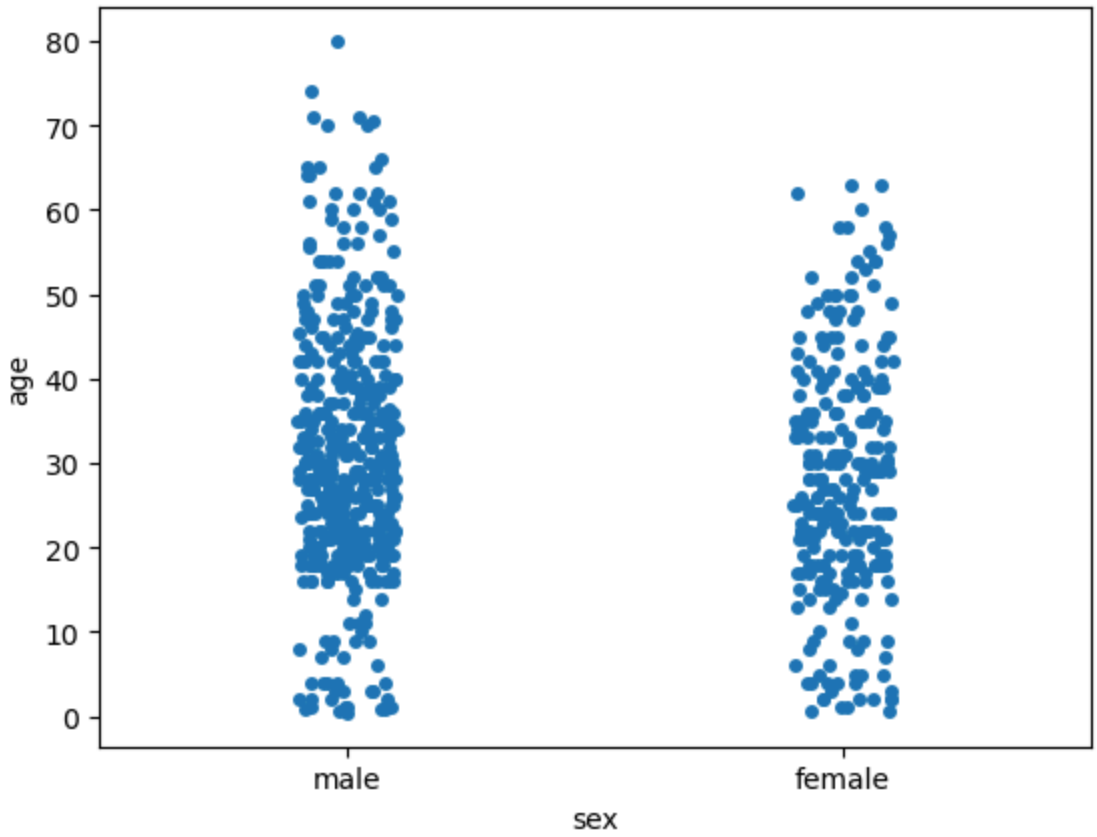
In [31]:
```
sns.stripplot(x='sex', y='age', data=dataset, jitter=False)
```

Out[31]: <Axes: xlabel='sex', ylabel='age'>



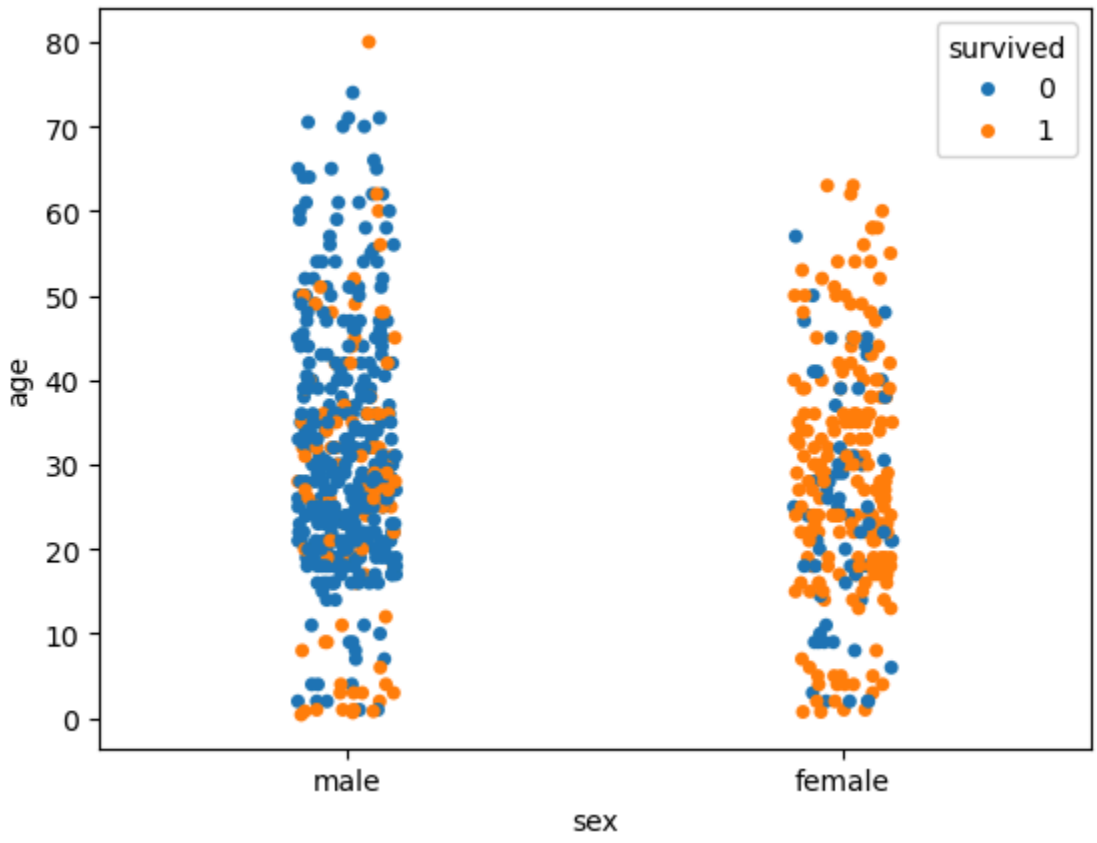
In [33]:
```
sns.stripplot(x='sex', y='age', data=dataset, jitter=True)
```

Out[33]: <Axes: xlabel='sex', ylabel='age'>



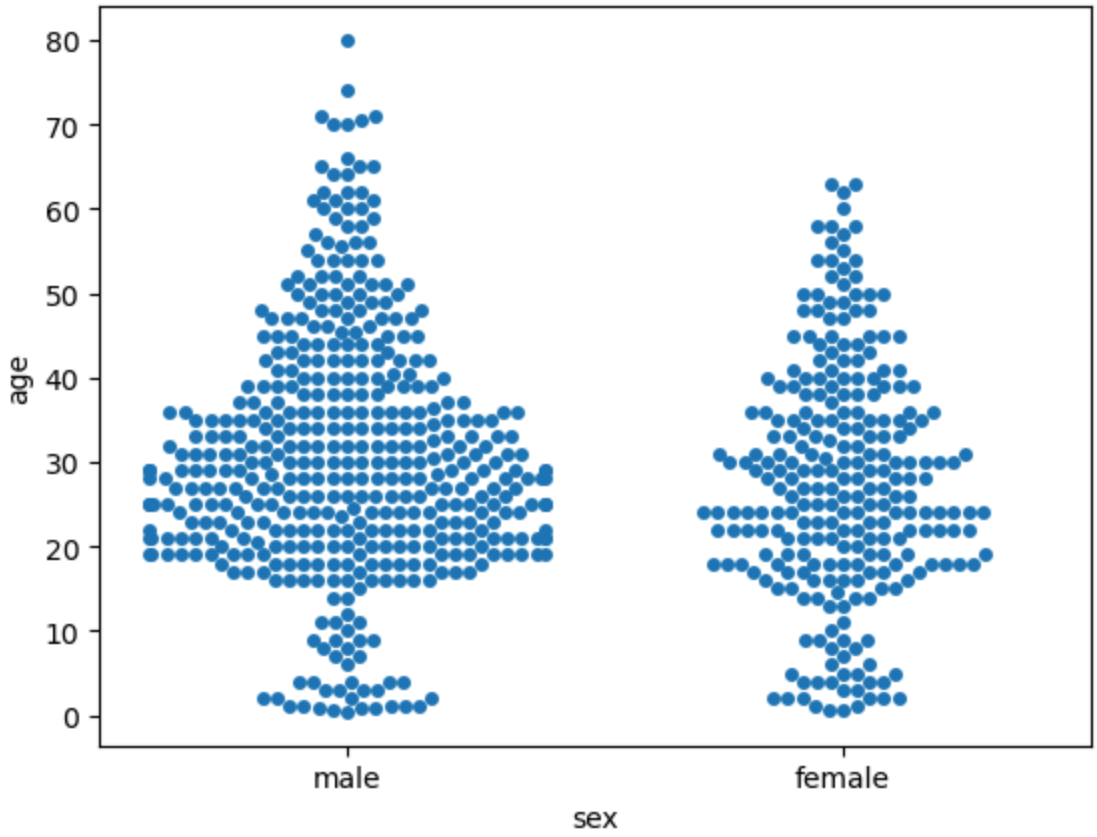
In [35]:
```
sns.stripplot(x='sex', y='age', data=dataset, jitter=True, hue='survived')
```

Out[35]: <Axes: xlabel='sex', ylabel='age'>



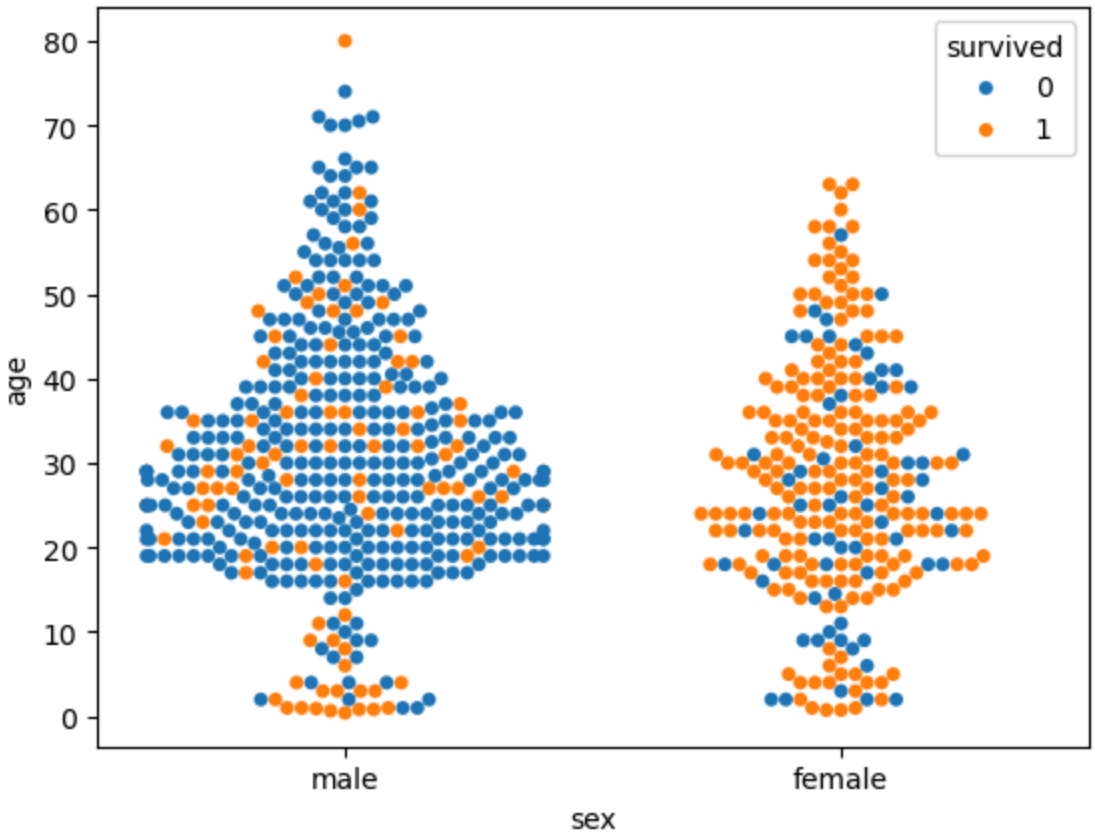
In [37]:
```
sns.swarmplot(x='sex', y='age', data=dataset)
```

Out[37]: <Axes: xlabel='sex', ylabel='age'>



In [39]:
```
sns.swarmplot(x='sex', y='age', data=dataset, hue='survived')
```

Out[39]: <Axes: xlabel='sex', ylabel='age'>



In [41]:
```
dataset
```

Out[41]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class
0	0	3	male	22.0	1	0	7.2500	S	Third
1	1	1	female	38.0	1	0	71.2833	C	First
2	1	3	female	26.0	0	0	7.9250	S	Third
3	1	1	female	35.0	1	0	53.1000	S	First
4	0	3	male	35.0	0	0	8.0500	S	Third
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Second
887	1	1	female	19.0	0	0	30.0000	S	First
888	0	3	female	NaN	1	2	23.4500	S	Third
889	1	1	male	26.0	0	0	30.0000	C	First
890	0	3	male	32.0	0	0	7.7500	Q	Third

891 rows × 15 columns

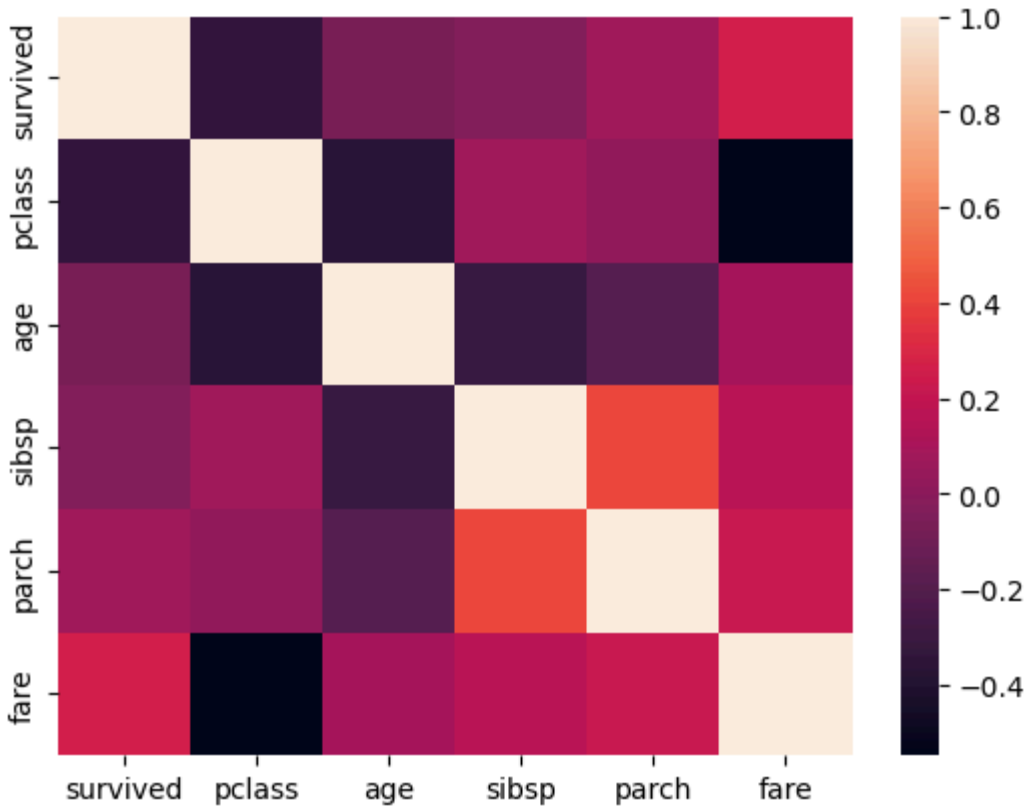
In [43]:
```
df = dataset.drop(columns=['sex','embarked','class','who','adult_male','deck
df.corr()
```

Out[43]:

	survived	pclass	age	sibsp	parch	fare
survived	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
pclass	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
age	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
sibsp	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
parch	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
fare	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

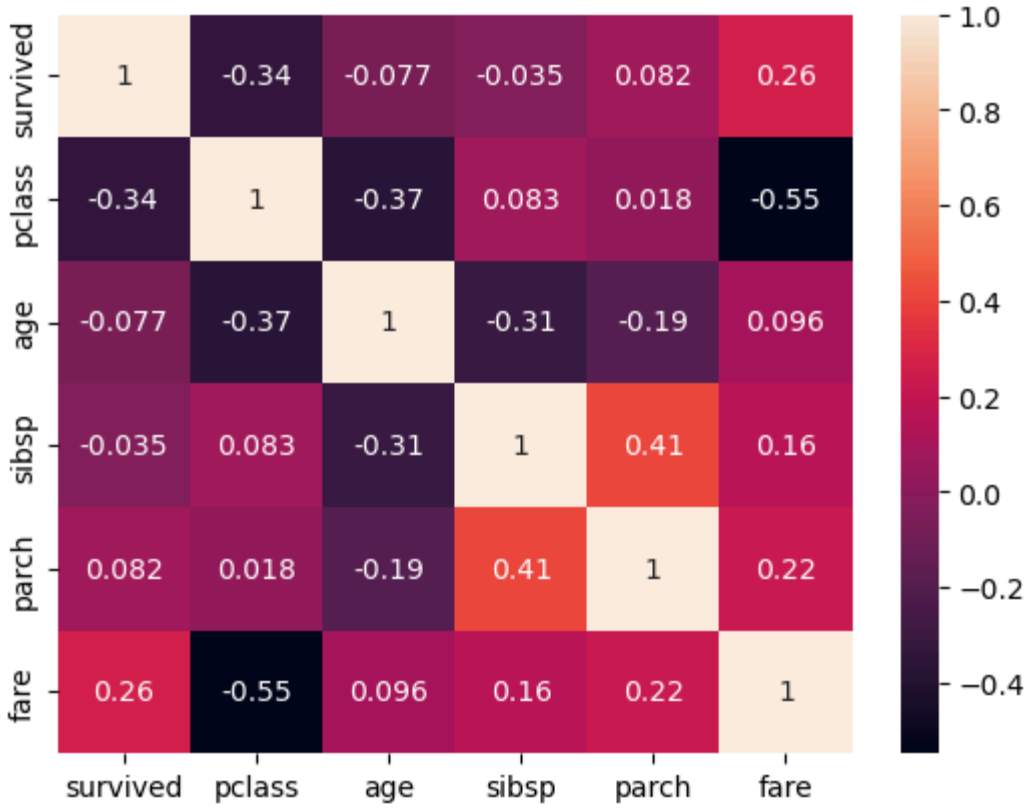
In [45]:
```
corr = df.corr()
sns.heatmap(corr)
```

Out[45]: <Axes: >



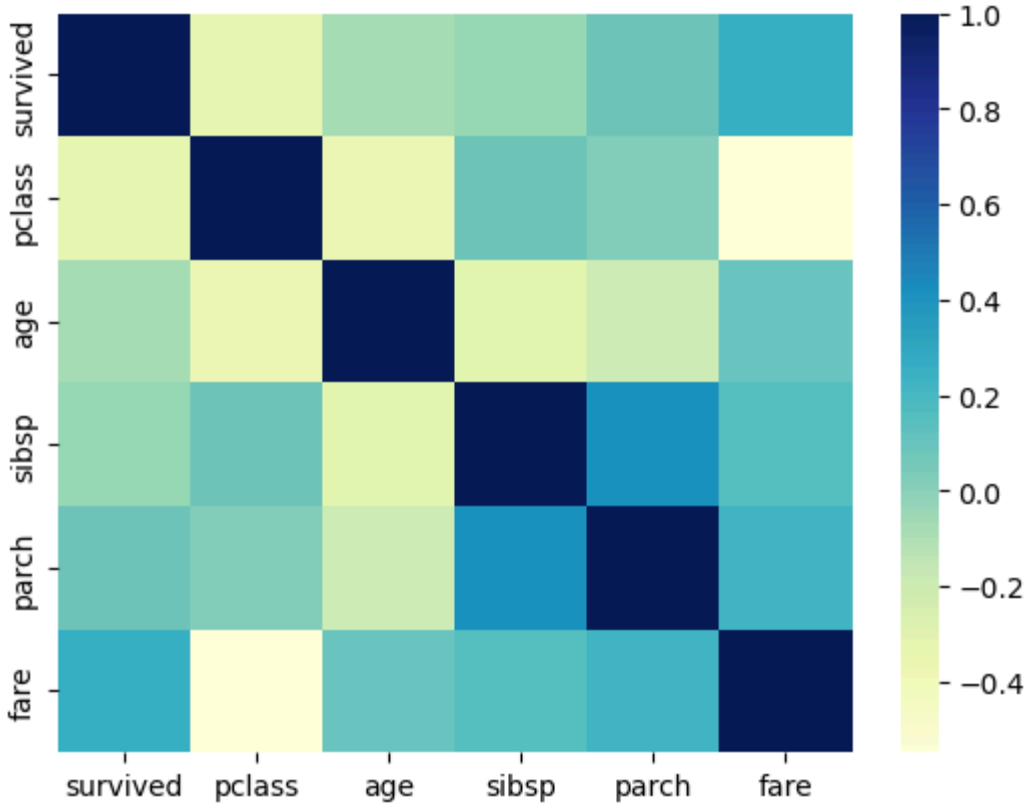
In [47]:
```
corr = df.corr()
sns.heatmap(corr,annot=True)
```

Out[47]: <Axes: >



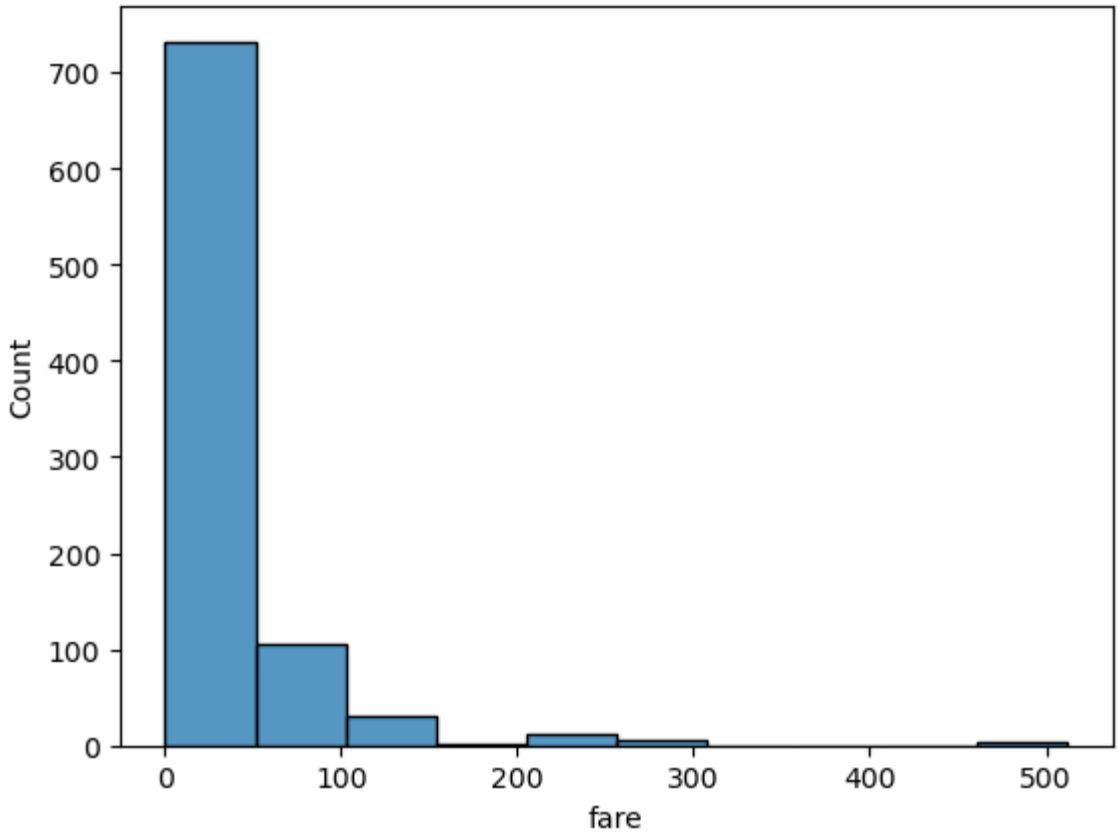
In [49]:
```
corr = df.corr()
sns.heatmap(corr,cmap="YlGnBu")
```

Out[49]: <Axes: >



In [51]:
```
sns.histplot(dataset['fare'],kde=False, bins=10)
```

Out[51]: <Axes: xlabel='fare', ylabel='Count'>



In [ ]:
```

```

Assignment No- 09

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title- Data Visualization II

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

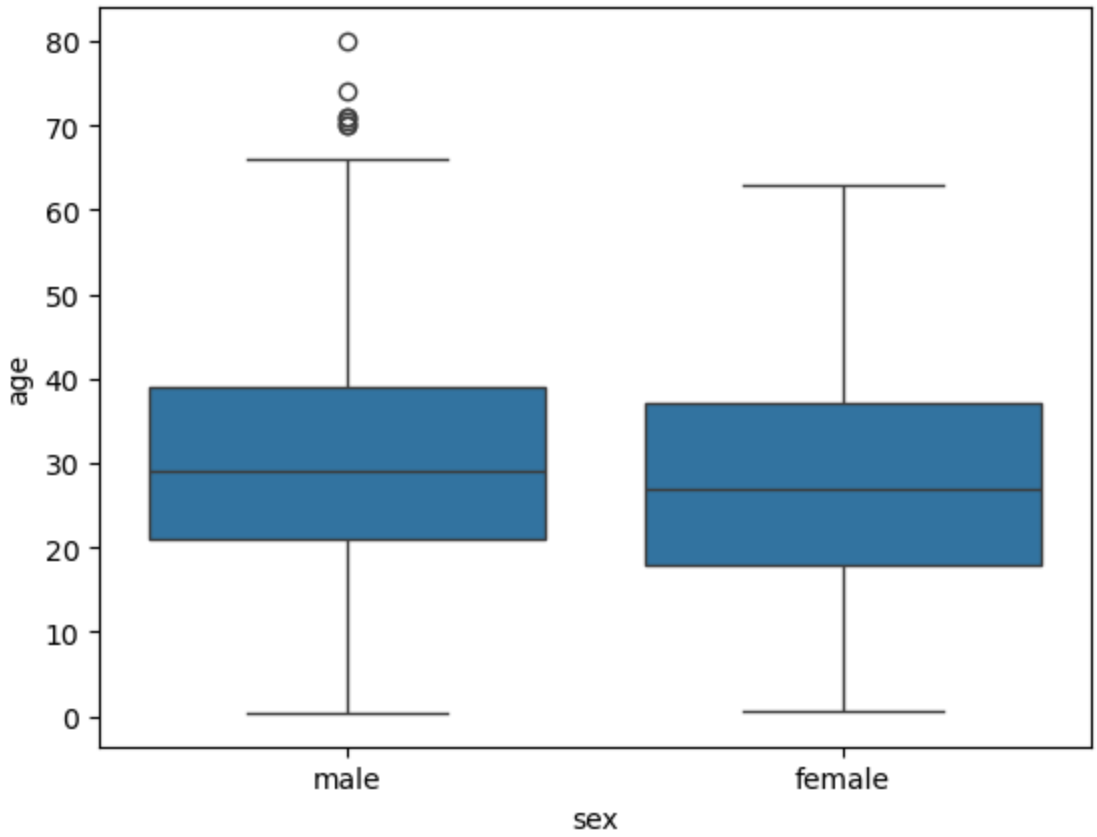
In [3]:
```
ds = sns.load_dataset('titanic')
```

In [5]:
```
ds.head()
```

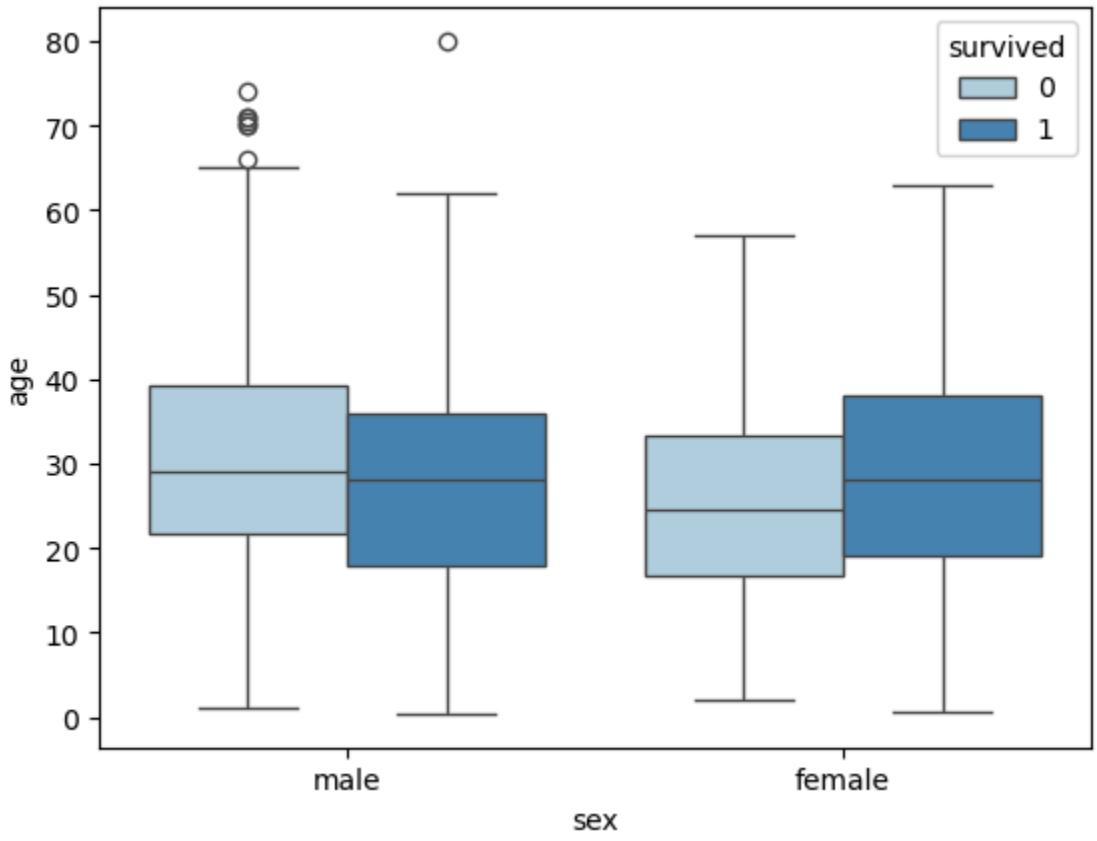
Out[5]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	w
0	0	3	male	22.0	1	0	7.2500	S	Third	m
1	1	1	female	38.0	1	0	71.2833	C	First	wom
2	1	3	female	26.0	0	0	7.9250	S	Third	wom
3	1	1	female	35.0	1	0	53.1000	S	First	wom
4	0	3	male	35.0	0	0	8.0500	S	Third	m

In [7]:
```
sns.boxplot(x='sex', y='age', data=ds)
plt.show()
```



In [9]:
```
sns.boxplot(x='sex', y='age', data=ds, hue='survived', palette="Blues")
plt.show()
```



Assignment No- 10

Name- Thorve Avishkar Shrikrushna

Roll No- 63

Title- Data Visualization III

In [5]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.d
column_names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width'
# Read the dataset into a DataFrame
iris_df = pd.read_csv(url, names=column_names)
```

In [7]:
```
iris_df
```

Out[7]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

In [10]:
```
iris_df.tail()
```

Out[10]:

	sepal_length	sepal_width	petal_length	petal_width	species
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

In [12]:
```
iris_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal_length  150 non-null    float64
 1   sepal_width   150 non-null    float64
 2   petal_length  150 non-null    float64
 3   petal_width   150 non-null    float64
 4   species       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

In [14]:
```
from sklearn.preprocessing import LabelEncoder , MinMaxScaler
le = LabelEncoder()
iris_df['species'] = le.fit_transform(iris_df['species'])
iris_df
```

Out[14]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2
146	6.3	2.5	5.0	1.9	2
147	6.5	3.0	5.2	2.0	2
148	6.2	3.4	5.4	2.3	2
149	5.9	3.0	5.1	1.8	2

150 rows × 5 columns

In [16]:
```
iris_df.describe()
```

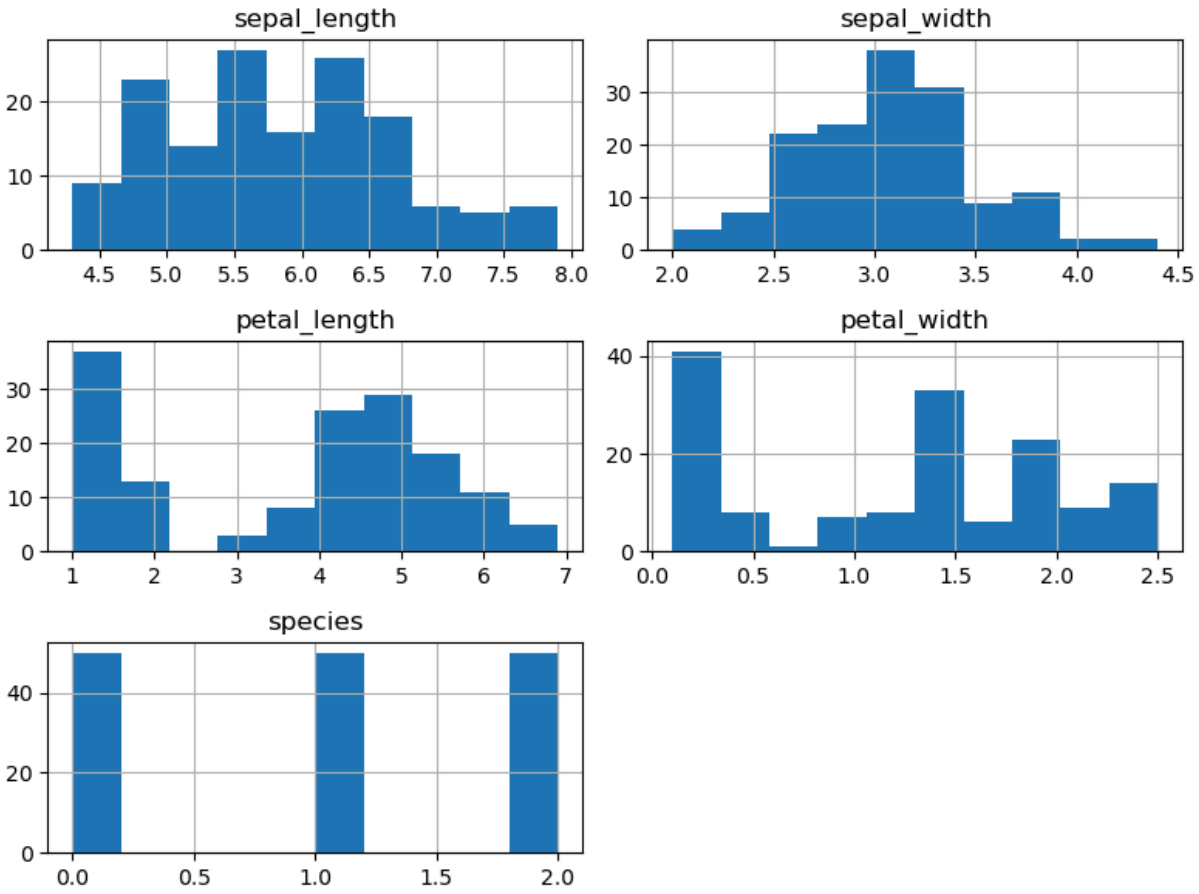
Out[16]:

	sepal_length	sepal_width	petal_length	petal_width	species
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667	1.000000
std	0.828066	0.433594	1.764420	0.763161	0.819232
min	4.300000	2.000000	1.000000	0.100000	0.000000
25%	5.100000	2.800000	1.600000	0.300000	0.000000
50%	5.800000	3.000000	4.350000	1.300000	1.000000
75%	6.400000	3.300000	5.100000	1.800000	2.000000
max	7.900000	4.400000	6.900000	2.500000	2.000000

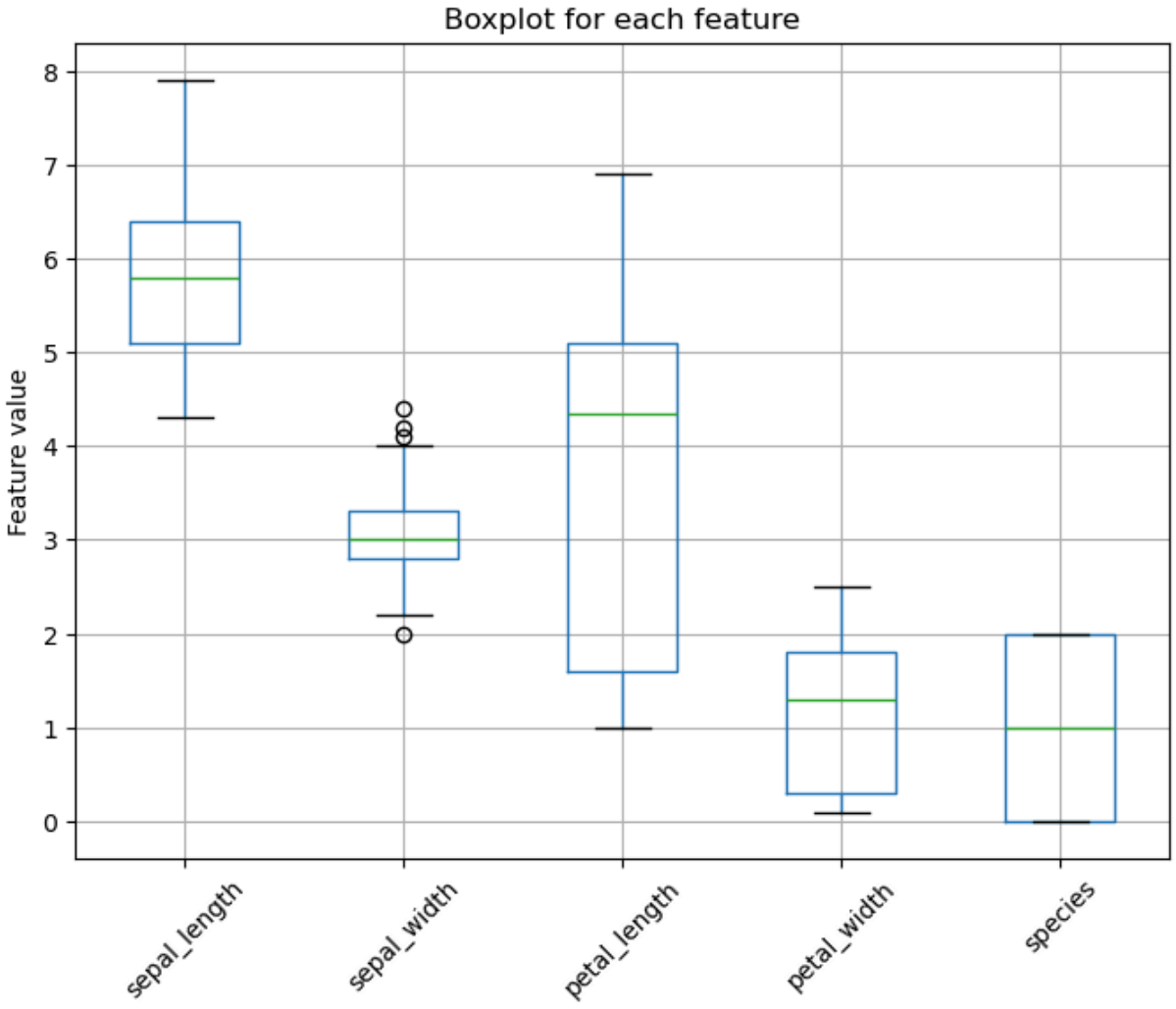
In [18]:
```
iris_df.isnull().sum()
```

Out[18]:
```
sepal_length    0
sepal_width     0
petal_length    0
petal_width     0
species         0
dtype: int64
```

In [20]:
```
iris_df.hist(figsize=(8, 6))
plt.tight_layout()
plt.show()
```



In [22]:
```
plt.figure(figsize=(8, 6))
iris_df.boxplot()
plt.title('Boxplot for each feature')
plt.ylabel('Feature value')
plt.xticks(rotation=45)
plt.show()
```



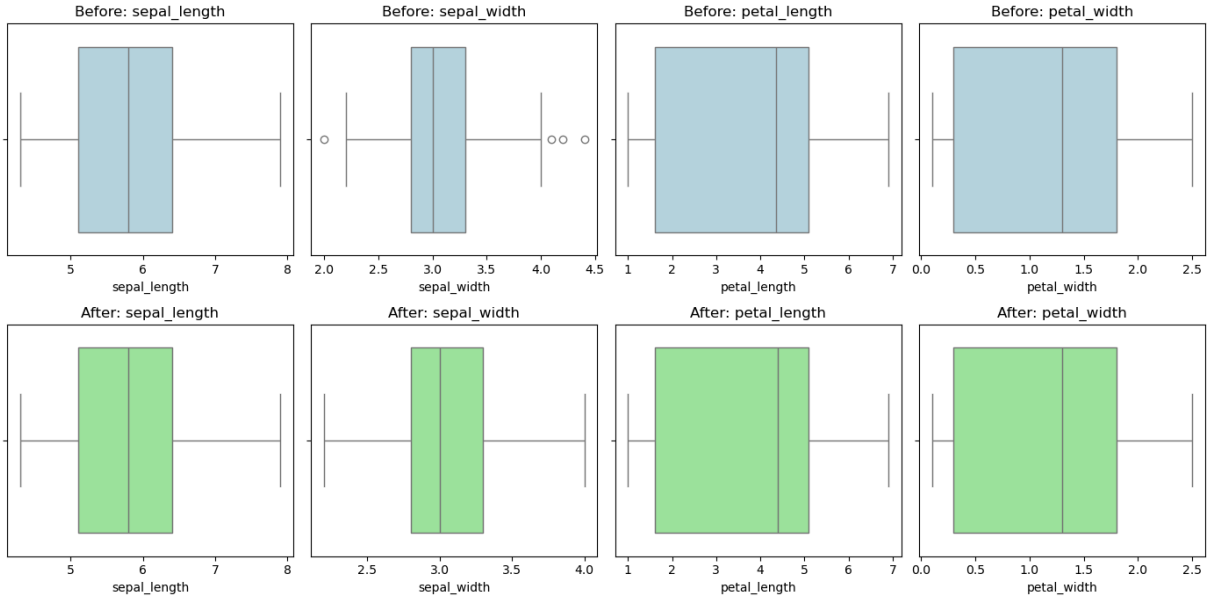
In [24]:
```
def remove_outliers(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_limit = Q1 - 1.5 * IQR
    upper_limit = Q3 + 1.5 * IQR
    return df[(df[column] >= lower_limit) & (df[column] <= upper_limit)]
```

In [26]:
```
columns_to_check = ['sepal_length', 'sepal_width', 'petal_length', 'petal_wi
cleaned_df = iris_df.copy()
for col in columns_to_check:
    cleaned_df = remove_outliers(cleaned_df, col)
print('Before removing outliers:', len(iris_df))
print('After removing outliers:', len(cleaned_df))
print('Outliers removed:', len(iris_df) - len(cleaned_df))
```

```
Before removing outliers: 150
After removing outliers: 146
Outliers removed: 4
```

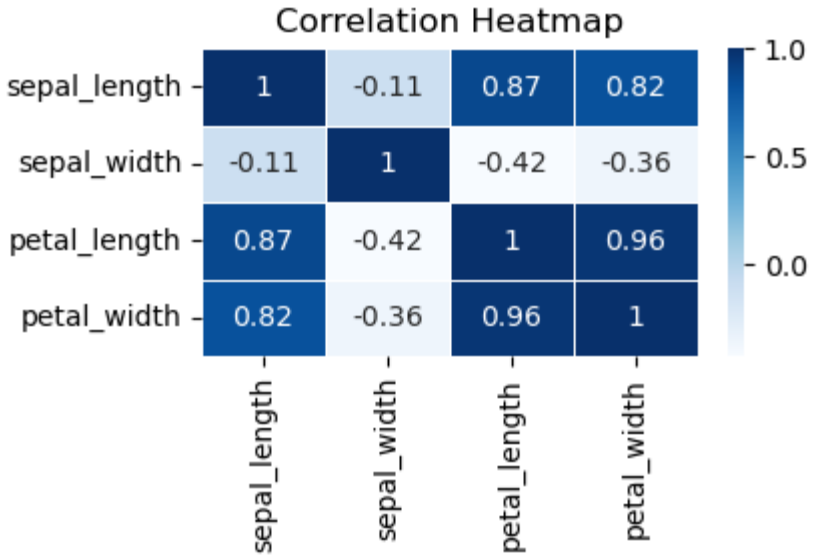
In [28]:
```
fig, axes = plt.subplots(nrows=2, ncols=4, figsize=(14, 7))
for i, col in enumerate(columns_to_check):
    sns.boxplot(x=iris_df[col], ax=axes[0, i], color='lightblue')
    axes[0, i].set_title(f'Before: {col}')
```

```
    sns.boxplot(x=cleaned_df[col], ax=axes[1, i], color='lightgreen')
    axes[1, i].set_title(f'After: {col}')
plt.tight_layout()
plt.show()
```



In [30]:
```
print("Heatmap for the Correlation")
subset_df = iris_df.iloc[:, :4]
plt.figure(figsize=(4, 2))
sns.heatmap(subset_df.corr(), annot=True, cmap="Blues", linewidths=0.5)
plt.title("Correlation Heatmap")
plt.show()
```

```
Heatmap for the Correlation
```



In [ ]:
```

```

Practical No 11

NAME: Thorve Avishkar Shrikrushna

Roll No: 63

Title: Create databases and tables, insert small amounts of data, and run simple queries using Impala

****

- **WordCount.java**

```
import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration; import
org.apache.hadoop.fs.Path; import
org.apache.hadoop.io.IntWritable; import
org.apache.hadoop.io.Text; import
org.apache.hadoop.mapreduce.Job; import
org.apache.hadoop.mapreduce.Mapper; import
org.apache.hadoop.mapreduce.Reducer; import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutputForma
t;

public class WordCount {

        public static class TokenizerMapper
                extends Mapper<Object, Text, Text, IntWritable>{
        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context
              ) throws IOException,
        InterruptedException {       StringTokenizer itr =
        new StringTokenizer(value.toString());
        while (itr.hasMoreTokens()) {
        word.set(itr.nextToken());
        context.write(word, one);
    }

}
}
```

```
        public static class IntSumReducer
            extends
        Reducer<Text,IntWritable,Text,IntWritable> {
        private IntWritable result = new IntWritable();

         public void reduce(Text key, Iterable<IntWritable> values, Context context )
                 throws IOException, InterruptedException {

        int sum = 0;
        for (IntWritable val : values)
        {       sum += val.get();
        }
        result.set(sum);
        context.write(key,
        result);
        }
       }

        public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word
        count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class)
        ;
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));
        System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

Input.txt

Pune
Mumbai
Nashik
Pune

Nashik
Kolhapur
delhi Pune
Channai
Nashik
Pune

Program Running step on terminal(linux)

pansa@pansa-HP-Laptop-14s-dr1xxx:~$ su hduser
Password:
hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa$ cd
hduser@pansa-HP-Laptop-14s-dr1xxx:~$ hadoop version Hadoop
3.4.0
Source code repository git@github.com:apache/hadoop.git
bd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled by root on 2024-03-04T06:35Z
Compiled on platform linux-x86_64
Compiled with protoc 3.21.12
From source with checksum f7fe694a3613358b38812ae9c31114e
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-3.4.0.jar
hduser@pansa-HP-Laptop-14s-dr1xxx:~$ javac -version
javac 11.0.22

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ export HADOOP_CLASSPATH=$(hadoop classpath)

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ echo HADOOP_CLASSPATH
HADOOP_CLASSPATH
hduser@pansa-HP-Laptop-14s-dr1xxx:~$ echo $HADOOP_CLASSPATH
/usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/local/hadoop/share/hadoop/common/*:/usr/local/hadoop/share/hadoop/hdfs:/usr/local/hadoop/share/hadoop/hdfs/lib/*:/usr/local/hadoop/share/hadoop/hdfs/*:/usr/local/hadoop/share/hadoop/mapreduce/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ hadoop fs -mkdir /classes_files
2024-04-21 22:42:49,381 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable mkdir: Call From pansa-HP-Laptop-14s-dr1xxx/127.0.1.1 to localhost:54310 failed on connection exception: java.net.ConnectException: Connection refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ hadoop fs -mkdir /WordCountPractical

2024-04-21 22:44:04,835 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable mkdir: Call From pansa-HP-Laptop-14s-dr1xxx/127.0.1.1 to localhost:54310 failed on connection exception: java.net.ConnectException: Connection refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ hadoop fs -mkdir /WordCountPractical/Input 2024-04-21 22:44:50,883 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
mkdir: Call From pansa-HP-Laptop-14s-dr1xxx/127.0.1.1 to localhost:54310 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd Home/Desktop/WordCountPractical
bash: cd: Home/Desktop/WordCountPractical: No such file or directory

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /Home/Desktop/WordCountPractical
bash: cd: /Home/Desktop/WordCountPractical: No such file or directory

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd home
bash: cd: home: No such file or directory

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd'/home/pansa/Desktop/WordCountPractical'
bash: cd/home/pansa/Desktop/WordCountPractical: No such file or directory
hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /home/pansa/Desktop/WordCountPractical
bash: cd: /home/pansa/Desktop/WordCountPractical: Permission denied

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd '/home/pansa/Desktop/WordCountPractical'
bash: cd: /home/pansa/Desktop/WordCountPractical: Permission denied

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /home/pansa/Desktop/WordCountPractical bash: cd: /home/pansa/Desktop/WordCountPractical: Permission denied

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /home/pansa/Desktop/WordCountPractical
bash: cd: /home/pansa/Desktop/WordCountPractical: Permission denied

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /home/pansa/Desktop/WordCountPractical
bash: cd: /home/pansa/Desktop/WordCountPractical: Permission denied

hduser@pansa-HP-Laptop-14s-dr1xxx:~$ cd /home/pansa/Desktop/WordCountPractical

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$ javac classpath ${HADOOP_CLASSPATH} -d

'/home/pansa/Desktop/WordCountPractical/classes_files'
'/home/pansa/Desktop/WordCountPractical/WordCount.java'

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$
jar -cvf firstpractical.jar -c classes_files/ . -c : no such file or directory
added manifest
adding: classes_files/(in = 0) (out= 0)(stored 0%)
adding: classes_files/WordCount$IntSumReducer.class(in = 1755) (out= 749)(deflated 57%) adding: classes_files/WordCount.class(in = 1511) (out= 825)(deflated 45%)
adding: classes_files/WordCount$TokenizerMapper.class(in = 1752) (out= 764)(deflated 56%) adding: WordCount.java(in = 2148) (out= 711)(deflated 66%)
adding: input_data/(in = 0) (out= 0)(stored 0%)
adding: input_data/input.txt(in = 71) (out= 48)(deflated 32%)
java.util.zip.ZipException: duplicate entry:
classes_files/WordCount$IntSumReducer.class
atjava.base/java.util.zip.ZipOutputStream.putNextEntry(ZipOutputStream.java:233) at
java.base/java.util.jar.JarOutputStream.putNextEntry(JarOutputStream.java:109)
at jdk.jartool/sun.tools.jar.Main.addFile(Main.java:1208)
at jdk.jartool/sun.tools.jar.Main.create(Main.java:879)
at jdk.jartool/sun.tools.jar.Main.run(Main.java:319)
hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$ jar -cvf firstpractical.jar -c /home/pansa/Desktop/WordCountPractical/classes_files/.
-c : no such file or
directory
added manifest
adding: home/pansa/Desktop/WordCountPractical/classes_files/./(in = 0) (out= 0)(stored 0%)
adding:
home/pansa/Desktop/WordCountPractical/classes_files/./WordCount$IntSumReducer.class(i n = 1755) (out= 749(deflated 57%)

adding: home/pansa/Desktop/WordCountPractical/classes_files/./WordCount.class(in = 1511) (out= 825)(deflated 45%) adding:
home/pansa/Desktop/WordCountPractical/classes_files/./WordCount$TokenizerMapper.class (in = 1752) (out= 764)(deflated 56%)

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$
jar -cvf firstpractical.jar -c classes_files/ . -c : no such file or directory
added manifest
adding: classes_files/(in = 0) (out= 0)(stored 0%)
adding: classes_files/WordCount$IntSumReducer.class(in = 1755) (out= 749)(deflated 57%) adding: classes_files/WordCount.class(in = 1511) (out= 825)(deflated 45%)

adding: classes_files/WordCount$TokenizerMapper.class(in = 1752) (out= 764)(deflated 56%) adding: WordCount.java(in = 2148) (out= 711)(deflated 66%)
adding: input_data/(in = 0) (out= 0)(stored 0%)
adding: input_data/input.txt(in = 71) (out= 48)(deflated 32%)
java.util.zip.ZipException: duplicate entry:
classes_files/WordCount$IntSumReducer.class
atjava.base/java.util.zip.ZipOutputStream.putNextEntry(ZipOutputStream.j
ava:233) at
java.base/java.util.jar.JarOutputStream.putNextEntry(JarOutputStream.java:109)
at jdk.jartool/sun.tools.jar.Main.addFile(Main.java:1208)
at jdk.jartool/sun.tools.jar.Main.create(Main.java:879)
at jdk.jartool/sun.tools.jar.Main.run(Main.java:319)
at jdk.jartool/sun.tools.jar.Main.main(Main.java:1680)
adding: classes_files/WordCount$IntSumReducer.classhduser@pansa-HP-Laptop-14sdr1xxx:/home/pansa/Desktop/WordCountPractical$ jar -cvf firstpractical.jar
-c /home/pansa/Desktop/WordCountPractical/classes_files/.
-c : no such file or
directory
added manifest
adding: home/pansa/Desktop/WordCountPractical/classes_files/./(in = 0) (out= 0)(stored 0%)
adding:
home/pansa/Desktop/WordCountPractical/classes_files/./WordCount$IntSumReduce
r.class(i n = 1755) (out= 749)(deflated 57%) adding:
home/pansa/Desktop/WordCountPractical/classes_files/./WordCount.class(in = 1511)
(out= 825)(deflated
45%) adding:
home/pansa/Desktop/WordCountPractical/classes_files/./WordCount$TokenizerMapper
.class (in = 1752) (out= 764)(deflated 56%)

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$ jar -cvf firstpra.jar classes_files/.
added manifest
adding: classes_files/./(in = 0) (out= 0)(stored 0%)
adding: classes_files/./WordCount$IntSumReducer.class(in = 1755) (out= 749)(deflated 57%) adding: classes_files/./WordCount.class(in = 1511) (out= 825)(deflated 45%) adding: classes_files/./WordCount$TokenizerMapper.class(in = 1752) (out= 764)(deflated 56%)

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$

jar tf /home/pansa/Desktop/WordCountPractical/firstpra.jar
META-INF/
META-INF/MANIFEST.MF/

classes_files/./
classes_files/./WordCount$IntSumReducer.class classes_files/./WordCount.class
classes_files/./WordCount$TokenizerMapper.class

hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$
hadoop jar '/home/pansa/Desktop/WordCountPractical/firstpra.jar'
WordCount
/WordCountPractical/Input /WordCountPractical/Output
2024-04-22 01:04:33,489 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-04-22 01:04:34,648 INFO impl.MetricsConfig: Loaded properties from hadoopmetrics2.properties
2024-04-22 01:04:34,794 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-04-22 01:04:34,794 INFO impl.MetricsSystemImpl: JobTracker metrics system started 2024-04-22 01:04:35,047 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-22 01:04:35,332 INFO input.FileInputFormat: Total input files to process : 1
2024-04-22 01:04:35,401 INFO mapreduce.JobSubmitter: number of splits:1
2024-04-22 01:04:35,689 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local20263485_0001
2024-04-22 01:04:35,689 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-22 01:04:35,976 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-04-22 01:04:35,979 INFO mapreduce.Job: Running job: job_local20263485_0001
2024-04-22 01:04:35,980 INFO mapred.LocalJobRunner: OutputCommitter set in config null 2024-04-22 01:04:35,993 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-04-22 01:04:35,995 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-04-22 01:04:35,995 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-04-22 01:04:35,997 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-04-22 01:04:36,072 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-04-22 01:04:36,073 INFO mapred.LocalJobRunner: Starting task: attempt_local20263485_0001_m_000000_0
2024-04-22 01:04:36,111 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2024-04-22 01:04:36,112 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-04-22 01:04:36,112 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-04-22 01:04:36,139 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]

2024-04-22 01:04:36,149 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/WordCountPractical/Input/input.txt:0+71
2024-04-22 01:04:36,266 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-04-22 01:04:36,266 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-04-22 01:04:36,266 INFO mapred.MapTask: soft limit at 83886080

2024-04-22 01:04:36,266 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2024-04-22 01:04:36,266 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2024-04-22 01:04:36,277 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2024-04-22 01:04:36,520 INFO mapred.LocalJobRunner:
2024-04-22 01:04:36,523 INFO mapred.MapTask: Starting flush of map output
2024-04-22 01:04:36,523 INFO mapred.MapTask: Spilling map output
2024-04-22 01:04:36,523 INFO mapred.MapTask: bufstart = 0; bufend = 115; bufvoid = 104857600
2024-04-22 01:04:36,523 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214356(104857424); length = 41/6553600
2024-04-22 01:04:36,556 INFO mapred.MapTask: Finished spill 0
2024-04-22 01:04:36,576 INFO mapred.Task: Task:attempt_local20263485_0001_m_000000_0 is done. And is in the process of committing
2024-04-22 01:04:36,582 INFO mapred.LocalJobRunner: map
2024-04-22 01:04:36,582 INFO mapred.Task: Task 'attempt_local20263485_0001_m_000000_0' done.
2024-04-22 01:04:36,593 INFO mapred.Task: Final Counters for attempt_local20263485_0001_m_000000_0: Counters: 24

File System Counters

FILE: Number of bytes read=3469
FILE: Number of bytes written=712207
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=71
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework

Map input records=11
Map output records=11
Map output bytes=115
Map output materialized bytes=84
Input split bytes=122
Combine input records=11
Combine output records=6

Spilled Records=6

Failed Shuffles=0

Merged Map outputs=0 GC time elapsed (ms)=14

Total committed heap usage (bytes)=312475648

File Input Format Counters

Bytes Read=71

2024-04-22 01:04:36,594 INFO mapred.LocalJobRunner: Finishing task: attempt_local20263485_0001_m_000000_0

2024-04-22 01:04:36,595 INFO mapred.LocalJobRunner: map task executor complete.

2024-04-22 01:04:36,600 INFO mapred.LocalJobRunner: Waiting for reduce tasks 2024-04-22 01:04:36,601 INFO mapred.LocalJobRunner: Starting task: attempt_local20263485_0001_r_000000_0

2024-04-22 01:04:36,614 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory

2024-04-22 01:04:36,614 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2

2024-04-22 01:04:36,614 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false

2024-04-22 01:04:36,615 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ] 2024-04-22 01:04:36,619 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@15a917ae

2024-04-22 01:04:36,622 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!

2024-04-22 01:04:36,707 INFO reduce.MergeManagerImpl: closeInMemoryFile -> mapoutput of size: 80, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->80 2024-04-22 01:04:36,709 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning 2024-04-22 01:04:36,710 INFO mapred.LocalJobRunner: 1 / 1 copied.

2024-04-22 01:04:36,711 INFO reduce.MergeManagerImpl: finalMerge called with 1 inmemory map-outputs and 0 on-disk map-outputs

2024-04-22 01:04:36,725 INFO mapred.Merger: Merging 1 sorted segments

2024-04-22 01:04:36,725 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 70 bytes

2024-04-22 01:04:36,731 INFO reduce.MergeManagerImpl: Merged 1 segments, 80 bytes to disk to satisfy reduce memory limit

2024-04-22 01:04:36,732 INFO reduce.MergeManagerImpl: Merging 1 files, 84 bytes from disk

2024-04-22 01:04:36,733 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce

2024-04-22 01:04:36,733 INFO mapred.Merger: Merging 1 sorted segments

2024-04-22 01:04:36,733 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 70 bytes

2024-04-22 01:04:36,734 INFO mapred.LocalJobRunner: 1 / 1 copied.

2024-04-22 01:04:36,776 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords

2024-04-22 01:04:36,860 INFO mapred.Task:

Task:attempt_local20263485_0001_r_000000_0 is done. And is in the process of committing
2024-04-22 01:04:36,864 INFO mapred.LocalJobRunner: 1 / 1 copied.
2024-04-22 01:04:36,864 INFO mapred.Task: Task attempt_local20263485_0001_r_000000_0 is allowed to commit now
2024-04-22 01:04:36,887 INFO output.FileOutputCommitter: Saved output of task 'attempt_local20263485_0001_r_000000_0' to hdfs://localhost:54310/WordCountPractical/Output
2024-04-22 01:04:36,889 INFO mapred.LocalJobRunner: reduce > reduce
2024-04-22 01:04:36,889 INFO mapred.Task: Task 'attempt_local20263485_0001_r_000000_0' done.
2024-04-22 01:04:36,891 INFO mapred.Task: Final Counters for attempt_local20263485_0001_r_000000_0: Counters: 30

File System Counters

FILE: Number of bytes read=3669
FILE: Number of bytes written=712291
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=71
HDFS: Number of bytes written=54
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework

Combine input records=0
Combine output records=0
Reduce input groups=6
Reduce shuffle bytes=84
Reduce input records=6
Reduce output records=6
Spilled Records=6
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=312475648

Shuffle Errors

BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Output Format Counters
Bytes Written=54
2024-04-22 01:04:36,891 INFO mapred.LocalJobRunner: Finishing task: attempt_local20263485_0001_r_000000_0
2024-04-22 01:04:36,891 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-04-22 01:04:36,988 INFO mapreduce.Job: Job job_local20263485_0001 running in uber mode : false
2024-04-22 01:04:36,989 INFO mapreduce.Job: map 100% reduce 100%
2024-04-22 01:04:36,991 INFO mapreduce.Job: Job job_local20263485_0001 completed successfully
2024-04-22 01:04:37,005 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=7138
FILE: Number of bytes written=1424498
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=142
HDFS: Number of bytes written=54
HDFS: Number of read operations=15
HDFS: Number of large read operations=0
HDFS: Number of write operations=4
HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework
Map input records=11
Map output records=11
Map output bytes=115
Map output materialized bytes=84
Input split bytes=122
Combine input records=11
Combine output records=6
Reduce input groups=6
Reduce shuffle bytes=84
Reduce input records=6
Reduce output records=6
Spilled Records=12
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=14
Total committed heap usage (bytes)=624951296
Shuffle Errors
BAD_ID=0

```
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=71
    File Output Format Counters
        Bytes Written=54
```

//OUTPUT: hduser@pansa-HP-Laptop-14s-dr1xxx:/home/pansa/Desktop/WordCountPractical$ hadoop dfs -cat /WordCountPractical/Output/*

```
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

2024-04-22 01:07:10,889 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Channai  1
Kolhapur 1
Mumbai 1
Nashk 3
Pune     4
delhi     1
```

Practical 12

NAME: Thorve Avishkar Shrikrushna
Roll No: 63
Title: Write a simple program in SCALA using Apache Spark framework.

****

- **SalesMapper.java**

```
package SalesCountry;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable; import
org.apache.hadoop.io.LongWritable; import
org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class SalesMapper extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable> {
        private final static IntWritable one = new IntWritable(1);

        public void map(LongWritable key, Text value, OutputCollector<Text,
IntWritable> output, Reporter reporter) throws IOException {

                String valueString = value.toString();
                String[] SingleCountryData = valueString.split("-");
                output.collect(new Text(SingleCountryData[0]), one);
        }
}
```

- **SalesCountryReducer.java**

```
package SalesCountry;

import java.io.IOException;
import java.util.*;
```

```
import org.apache.hadoop.io.IntWritable; import
org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.*;

public class SalesCountryReducer extends MapReduceBase implements
Reducer<Text, IntWritable, Text, IntWritable> {

        public void reduce(Text t_key, Iterator<IntWritable> values,
OutputCollector<Text,IntWritable> output, Reporter reporter) throws IOException {
                Text key = t_key;
int frequencyForCountry = 0;
while (values.hasNext()) {
                        // replace type of value with the actual type of our value
        IntWritable value = (IntWritable) values.next();
                        frequencyForCountry += value.get();

                }
                output.collect(key, new IntWritable(frequencyForCountry));
        }
}
```

- **SalesCountryReducer.java**

```
package SalesCountry;

import org.apache.hadoop.fs.Path; import
org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;

public class SalesCountryDriver {
        public static void main(String[] args) {
                JobClient my_client = new JobClient();
                // Create a configuration object for the job
                JobConf job_conf = new JobConf(SalesCountryDriver.class);

                // Set a name of the Job
                job_conf.setJobName("SalePerCountry");

                // Specify data type of output key and value
job_conf.setOutputKeyClass(Text.class);
                job_conf.setOutputValueClass(IntWritable.class);
```

```
		// Specify names of Mapper and Reducer Class
	job_conf.setMapperClass(SalesCountry.SalesMapper.class);
	job_conf.setReducerClass(SalesCountry.SalesCountryReducer.class);

		// Specify formats of the data type of Input and output
	job_conf.setInputFormat(TextInputFormat.class);
	job_conf.setOutputFormat(TextOutputFormat.class);

		// Set input and output directories using command line arguments,
	//arg[0] = name of input directory on HDFS, and arg[1] =  name of output
	directory to be created to store the output file.

		FileInputFormat.setInputPaths(job_conf, new Path(args[0]));
		FileOutputFormat.setOutputPath(job_conf, new Path(args[1]));

		my_client.setConf(job_conf);
		try {
			// Run the job
			JobClient.runJob(job_conf);
		} catch (Exception e) {
			e.printStackTrace();
		}
	}
}
```

- **Access_log_short.csv (Some Samples From CSV input file)**

```
10.223.157.186 - - [15/Jul/2009:20:50:32 -0700] "GET /assets/js/the-associates.js HTTP/1.1" 304 - 10.223.157.186
- - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/home-logo.png HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/primary-news-2.jpg HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/primary-news-1.jpg HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/home-media-block-placeholder.jpg HTTP/1.1" 304
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/secondary-news-4.jpg HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/loading.gif HTTP/1.1" 304 - 10.223.157.186
- - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/search-button.gif HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/secondary-news-3.jpg HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/secondary-news-2.jpg HTTP/1.1" 304 -
10.223.157.186 - - [15/Jul/2009:20:50:33 -0700] "GET /assets/img/dummy/secondary-news-1.jpg HTTP/1.1" 304 -
```

10.216.113.172 - - [16/Jul/2009:02:51:31 -0700] "GET /assets/img/closelabel.gif HTTP/1.1" 200 979

10.216.113.172 - - [16/Jul/2009:02:51:31 -0700] "GET /favicon.ico HTTP/1.1" 404 209

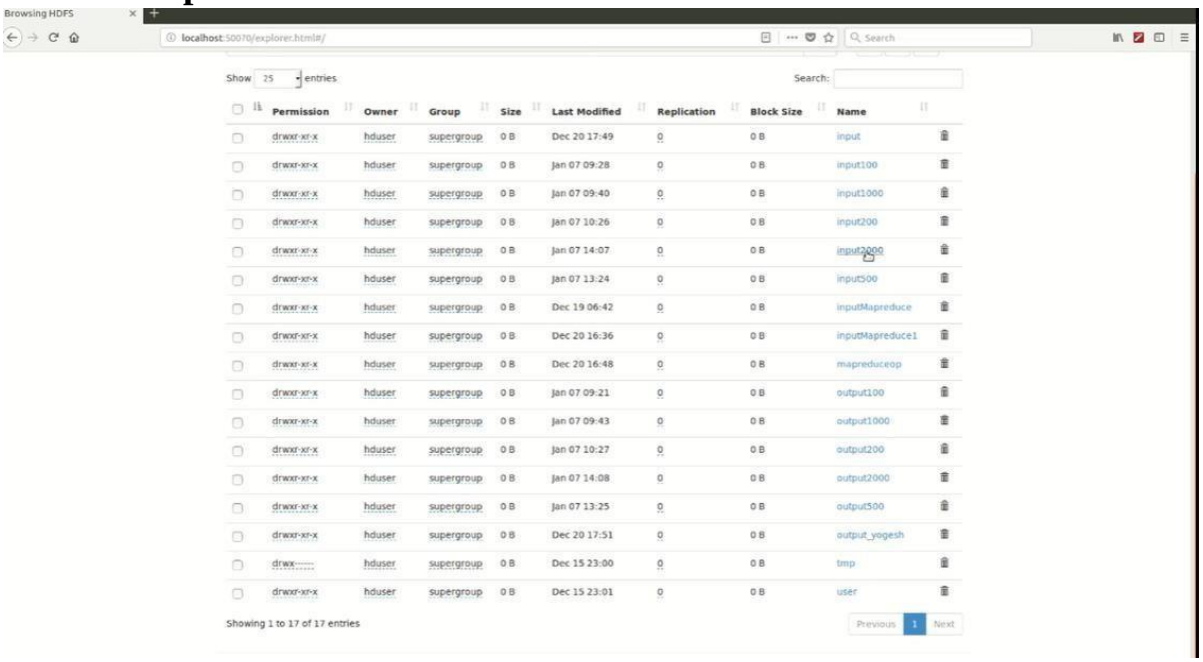
10.216.113.172 - - [16/Jul/2009:02:51:31 -0700] "GET /assets/swf/home-media-block.swf HTTP/1.1" 200 123884

10.216.113.172 - - [16/Jul/2009:02:51:41 -0700] "GET /films/district-13 HTTP/1.1" 301 268

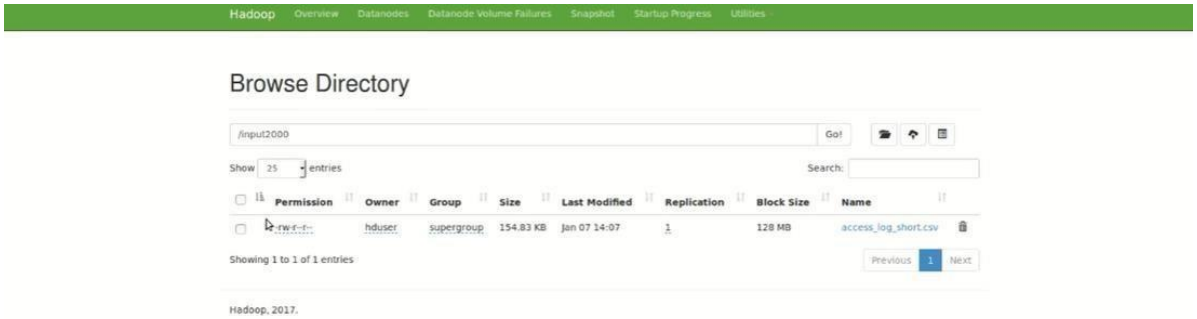
10.216.113.172 - - [16/Jul/2009:02:51:41 -0700] "GET /films/district-13/ HTTP/1.1" 200 12772

- **OUTPUT**

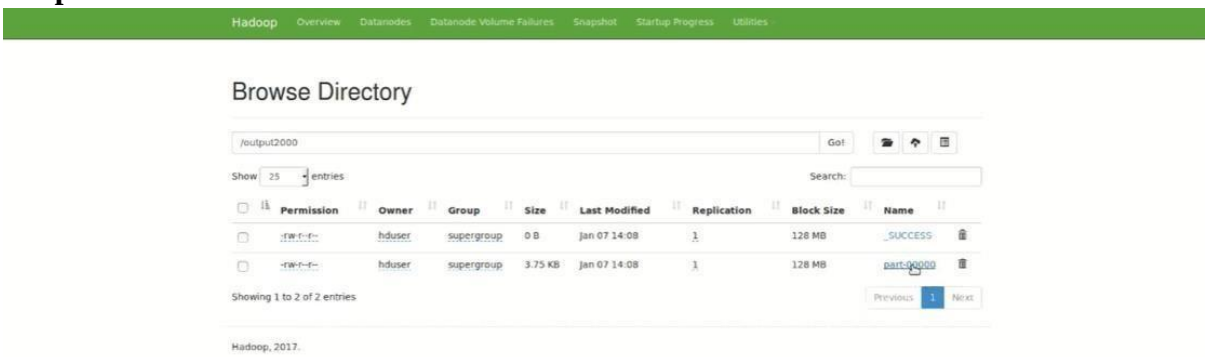
Hadoop Dashboard



Input Folder



Output Folder



Actual Output