G. H. Raisoni College Of Engineering And Management, Wagholi Pune

2021- 2022

Group C :-Assignment no :- 1

Department	CE [SUMMER 2022 (Online)]		
Term / Section	III/B	Date Of **submission**	08-10-2021
Subject Name /Code	Python for Data Science / UCSP204		
Roll No.	SCOB77	Name	Pratham Rajkumar pitty
Registration Number	2020AC0E1100107		

Group C → Assignment NO 1 (16)

Aim: using the sample dataset
(i) Handle the null values if any by removing them or perform imputation.

(ii) Import the ~~ness~~ necessary package and perform the train and test split on the dataset.

Theory:→

pandas is a python library for data analysis.

- isnull() function:→

The isnull() function is used to detect missing values for an array-like object.

This Function takes a Scalar or array-like object and indicates whether values are missing (NaN in numeric arrays None or NaN in object array, NaT in datetimelike).

Syntax:

```
pandas.isnull(obj)
```

- Parameters:

obj – Type→ Scalar or array-like. – To check for null values

Returns: bool or array-like of bool.

- Fillna()

Datafram.Fillna()

- The fillna() method replaces the NULL values with a specified value.

The Fillna() method returns a new DataFrame object unless the inplace parameter is set to True, in that case the Fillna() method does the replacing in the orignal DataFrame instead.

Syntax:

```
dataframe.fillna(value, method, axis, inplace, limit,
                                          downcast)
```

- Pandas DataFrame : dropna() function

The dropna() Function is used to remove missing values

syntax:

```
Dataframe.dropa(Self, axis=0, how='any', thresh=None,
                          Subset=None, inplace=False)
```

Sometimes CSV file has null values, which are later displayed as NaN in Data Frame.
Pandas dropa() method allows the user to analyze and drop Rows/columns with Null values in different ways.

- iloc in python

iloc is a Build in python Function. It is used to retrive rows from the data set. This Function is used when the indexing in the Dataset is not a number (0,1,2,3 .. n) or when the user does not know the exact name given to the index

Syntax:
pandas.dataframe.iloc[]

Parameters:

(1) index position
(2) Return type

- Matpiotib :-

It is open Sorce project of Num Focus matplotlib is a comprehensive library for Creating Static, animated, and interactive visualizations in python

- Train Test split ()

The Sklearn library method train test split is used to Split data in train and Test sets.

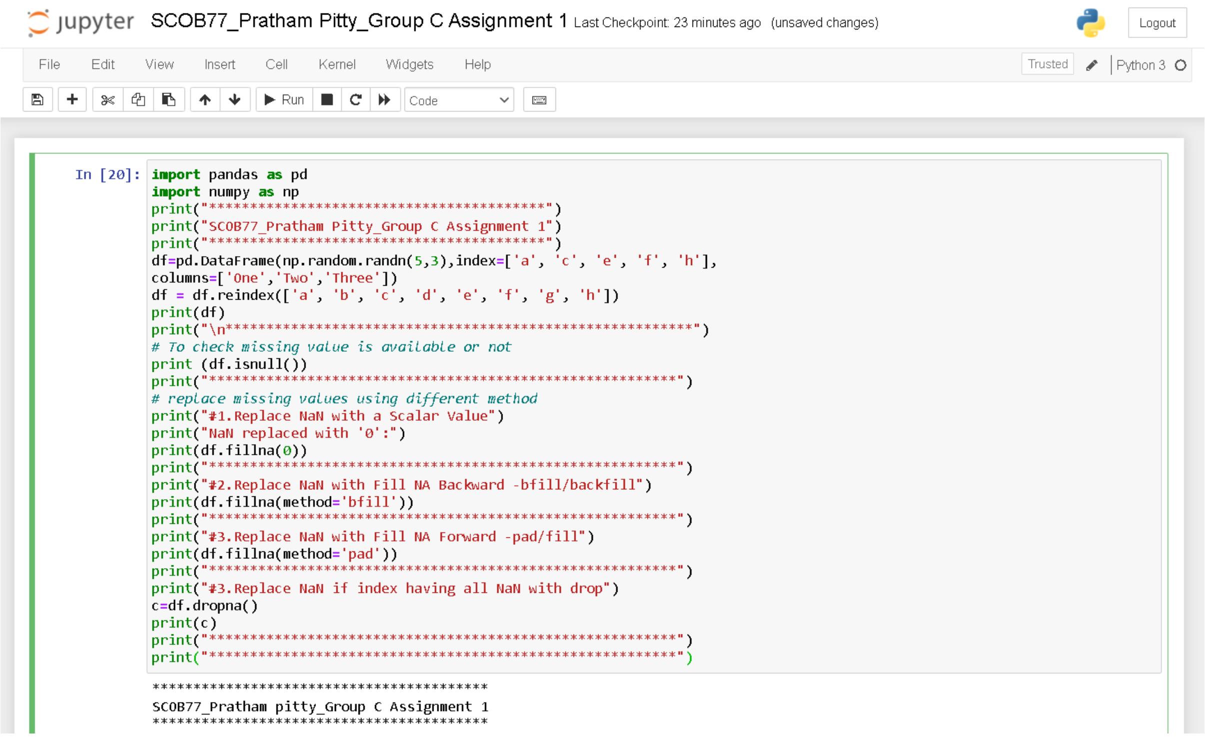
The process of Train and Test split splitting the dataset into two different sets called train and test sets.

- Train sets - Used to fit data into your machine Learning model.

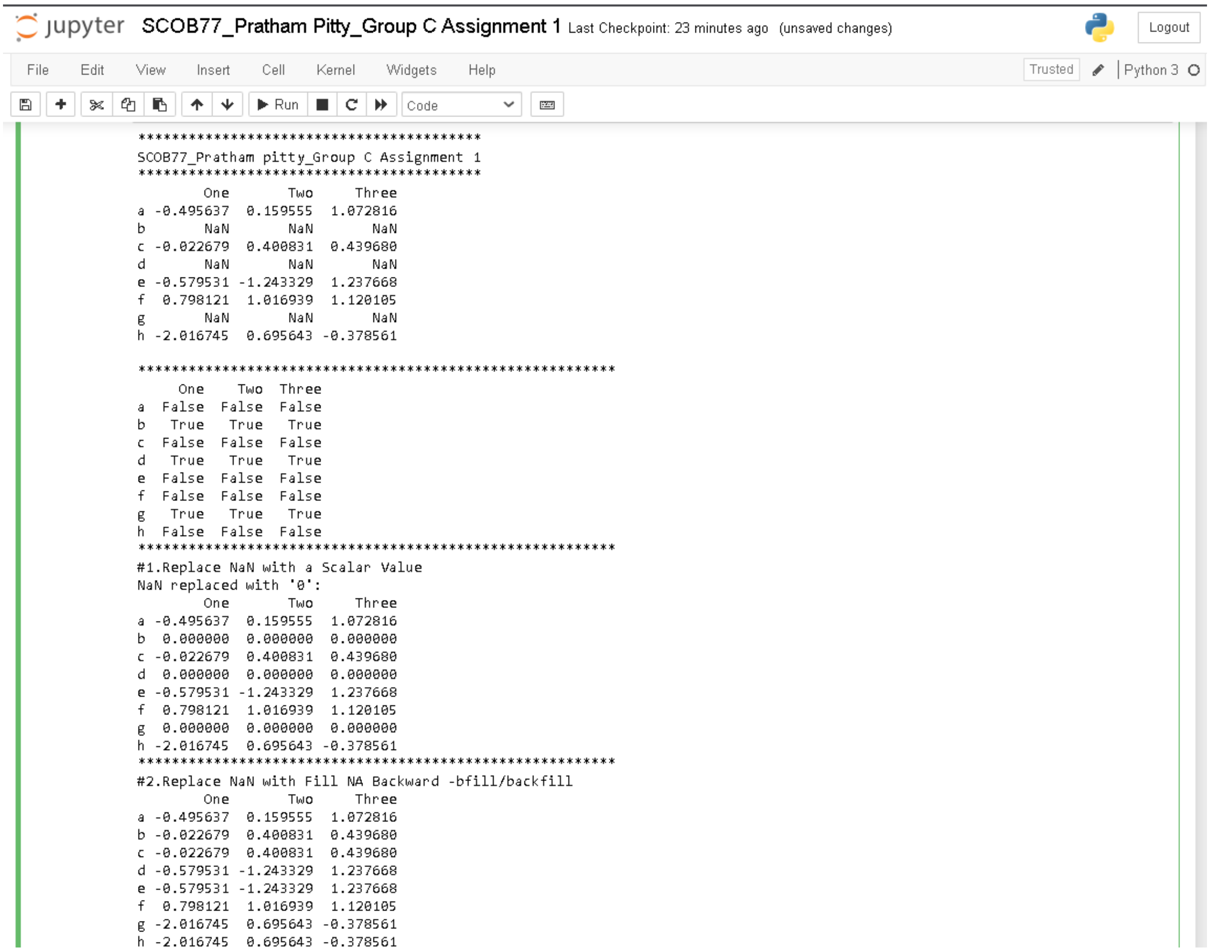
- Test sets -> Used to evaluate the Fit in your machine Learning model.

Group C Assignment 1 program code

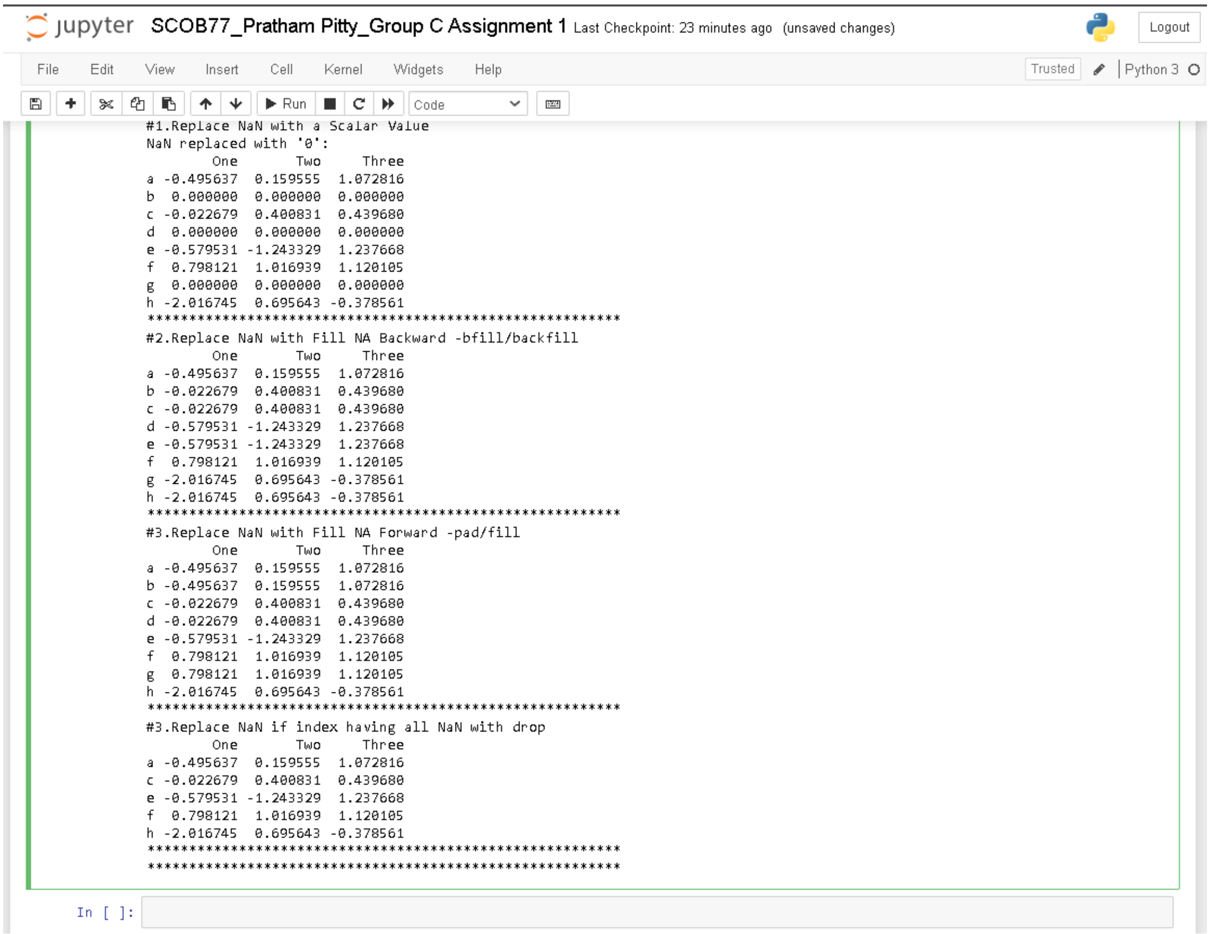
```
#part1. Handle the null values if any by removing them or perform imputation

import pandas as pd

import numpy as np

print("*****************************************")

print("SCOB77_Pratham pitty_Group C Assignment 1")

print("*****************************************")

df=pd.DataFrame(np.random.randn(5,3),index=['a', 'c', 'e', 'f', 'h'],

columns=['One','Two','Three'])

df = df.reindex(['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h'])

print(df)

print("\n*********************************************************")

# To check missing value is available or not

print (df.isnull())

print("**********************************************************")

# replace missing values using different method

print("#1.Replace NaN with a Scalar Value")

print("NaN replaced with '0':")

print(df.fillna(0))

print("**********************************************************")

print("#2.Replace NaN with Fill NA Backward -bfill/backfill")

print(df.fillna(method='bfill'))

print("**********************************************************")

print("#3.Replace NaN with Fill NA Forward -pad/fill")

print(df.fillna(method='pad'))

print("**********************************************************")

print("#3.Replace NaN if index having all NaN with drop")

c=df.dropna()

print(c)

print("**********************************************************")
print("*********************************************************")
```


jupyter
SCOB77_Pratham Pitty_Group C Assignment 1 Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [20]: import pandas as pd
         import numpy as np
         print("*****************************************")
         print("SCOB77_Pratham Pitty_Group C Assignment 1")
         print("*****************************************")
         df=pd.DataFrame(np.random.randn(5,3),index=['a', 'c', 'e', 'f', 'h'],
         columns=['One','Two','Three'])
         df = df.reindex(['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h'])
         print(df)
         print("\n*********************************************************")
         # To check missing value is available or not
         print (df.isnull())
         print("*********************************************************")
         # replace missing values using different method
         print("#1.Replace NaN with a Scalar Value")
         print("NaN replaced with '0':")
         print(df.fillna(0))
         print("*********************************************************")
         print("#2.Replace NaN with Fill NA Backward -bfill/backfill")
         print(df.fillna(method='bfill'))
         print("*********************************************************")
         print("#3.Replace NaN with Fill NA Forward -pad/fill")
         print(df.fillna(method='pad'))
         print("*********************************************************")
         print("#3.Replace NaN if index having all NaN with drop")
         c=df.dropna()
         print(c)
         print("*********************************************************")
         print("*********************************************************")

         *****************************************
         SCOB77_Pratham pitty_Group C Assignment 1
         *****************************************
```

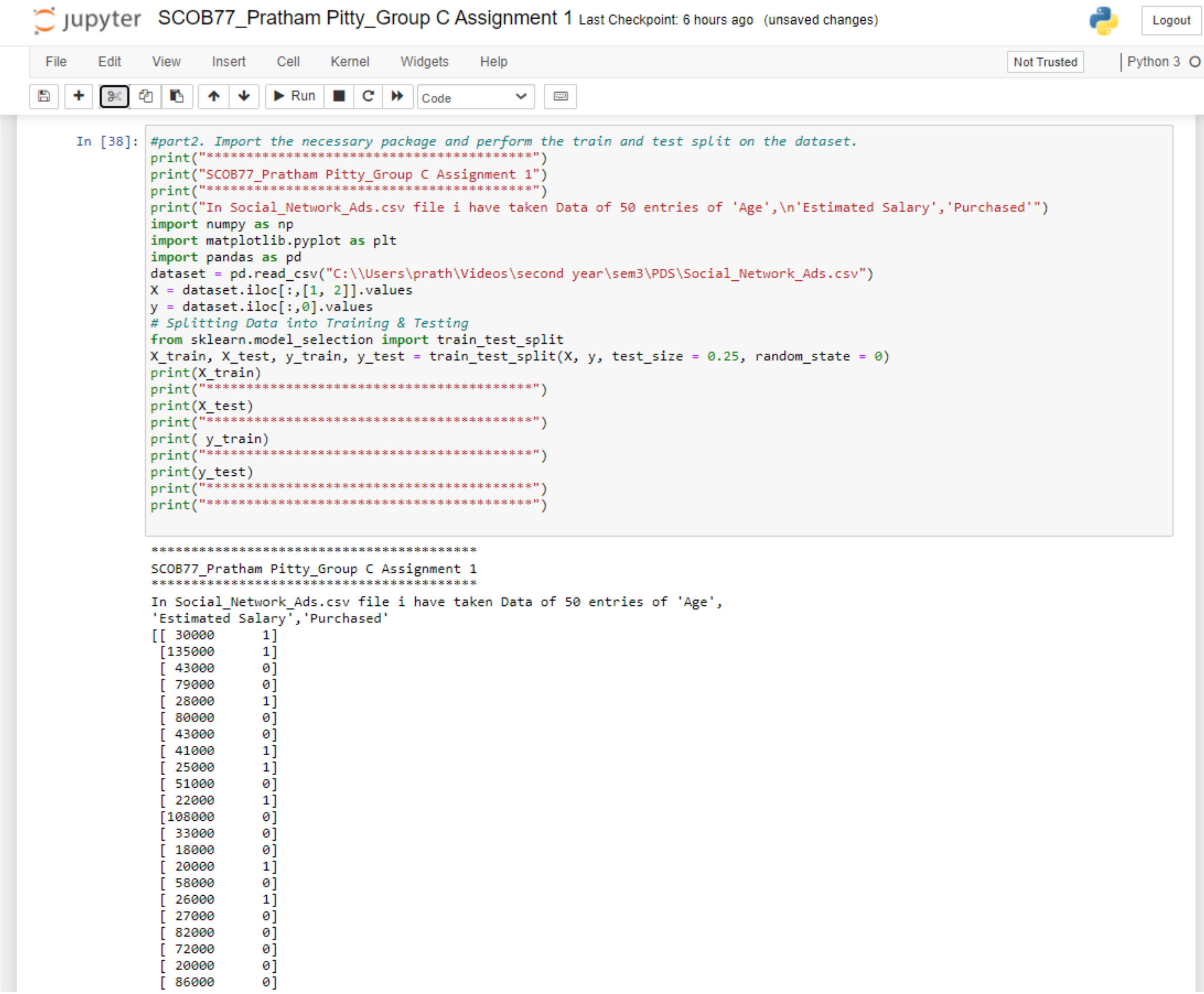

jupyter
SCOB77_Pratham Pitty_Group C Assignment 1 Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
*****************************************
SCOB77_Pratham pitty_Group C Assignment 1
*****************************************
        One       Two     Three
a -0.495637  0.159555  1.072816
b       NaN       NaN       NaN
c -0.022679  0.400831  0.439680
d       NaN       NaN       NaN
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g       NaN       NaN       NaN
h -2.016745  0.695643 -0.378561

*********************************************************
     One    Two  Three
a  False  False  False
b   True   True   True
c  False  False  False
d   True   True   True
e  False  False  False
f  False  False  False
g   True   True   True
h  False  False  False
*********************************************************
#1.Replace NaN with a Scalar Value
NaN replaced with '0':
        One       Two     Three
a -0.495637  0.159555  1.072816
b  0.000000  0.000000  0.000000
c -0.022679  0.400831  0.439680
d  0.000000  0.000000  0.000000
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g  0.000000  0.000000  0.000000
h -2.016745  0.695643 -0.378561
*********************************************************
#2.Replace NaN with Fill NA Backward -bfill/backfill
        One       Two     Three
a -0.495637  0.159555  1.072816
b -0.022679  0.400831  0.439680
c -0.022679  0.400831  0.439680
d -0.579531 -1.243329  1.237668
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g -2.016745  0.695643 -0.378561
h -2.016745  0.695643 -0.378561
```


jupyter
SCOB77_Pratham Pitty_Group C Assignment 1 Last Checkpoint: 23 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
In [ ]:

```
#1.Replace NaN with a Scalar Value
NaN replaced with '0':
        One       Two     Three
a -0.495637  0.159555  1.072816
b  0.000000  0.000000  0.000000
c -0.022679  0.400831  0.439680
d  0.000000  0.000000  0.000000
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g  0.000000  0.000000  0.000000
h -2.016745  0.695643 -0.378561
*********************************************************
#2.Replace NaN with Fill NA Backward -bfill/backfill
        One       Two     Three
a -0.495637  0.159555  1.072816
b -0.022679  0.400831  0.439680
c -0.022679  0.400831  0.439680
d -0.579531 -1.243329  1.237668
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g -2.016745  0.695643 -0.378561
h -2.016745  0.695643 -0.378561
*********************************************************
#3.Replace NaN with Fill NA Forward -pad/fill
        One       Two     Three
a -0.495637  0.159555  1.072816
b -0.495637  0.159555  1.072816
c -0.022679  0.400831  0.439680
d -0.022679  0.400831  0.439680
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
g  0.798121  1.016939  1.120105
h -2.016745  0.695643 -0.378561
*********************************************************
#3.Replace NaN if index having all NaN with drop
        One       Two     Three
a -0.495637  0.159555  1.072816
c -0.022679  0.400831  0.439680
e -0.579531 -1.243329  1.237668
f  0.798121  1.016939  1.120105
h -2.016745  0.695643 -0.378561
*********************************************************
*********************************************************
```

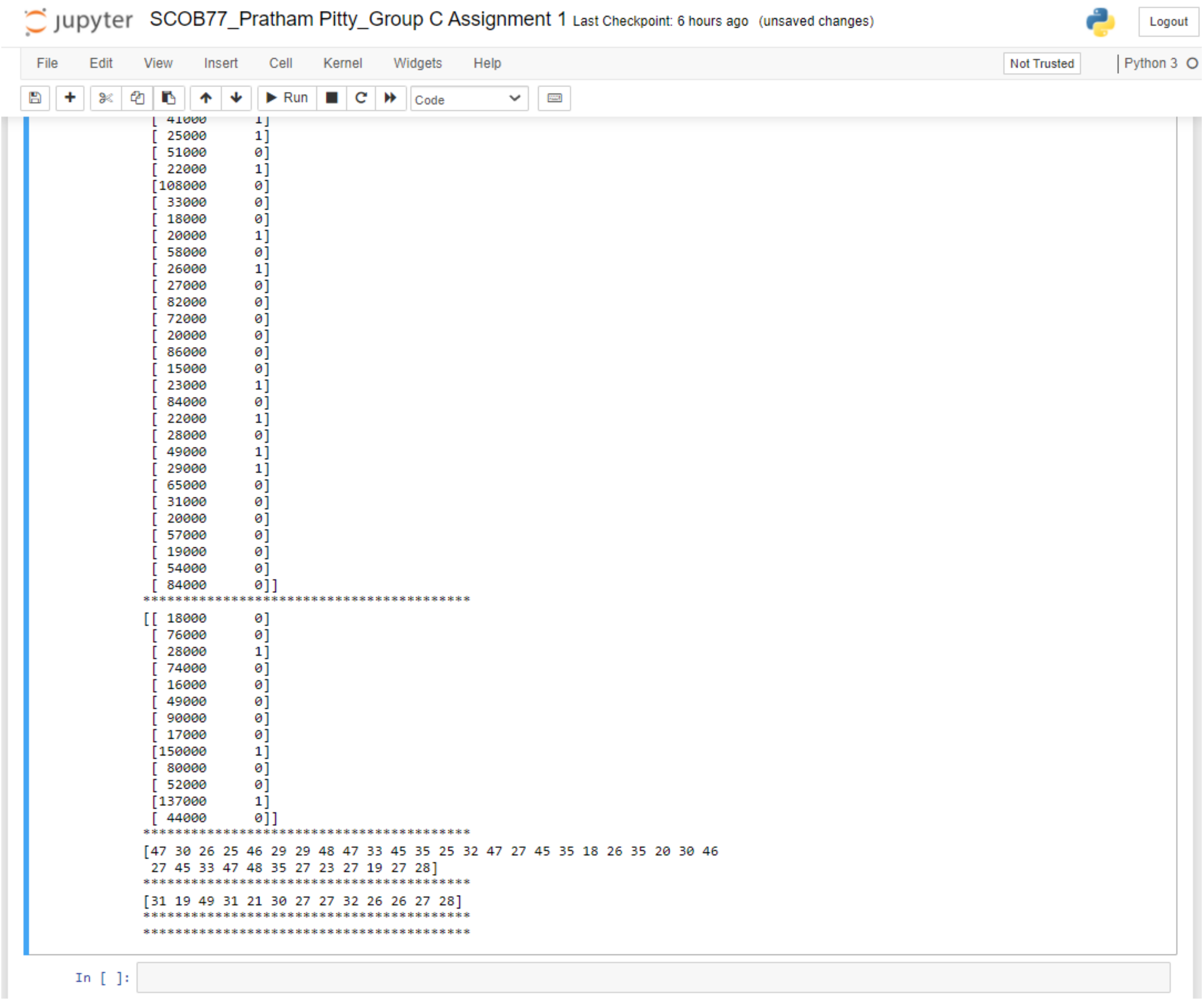
```
#part2. Import the necessary package and perform the train and test split on the dataset.

print("******************************************")

print("SCOB77_Pratham Pitty_Group C Assignment 1")

print("******************************************")

print("In Social_Network_Ads.csv file i have taken Data of 50 entries of 'Age',\n'Estimated Salary','Purchased'")

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

dataset = pd.read_csv("C:\\Users\prath\Videos\second year\sem3\PDS\Social_Network_Ads.csv")

X = dataset.iloc[:,[1, 2]].values

y = dataset.iloc[:,0].values

# Splitting Data into Training & Testing

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)

print(X_train)

print("******************************************")

print(X_test)

print("******************************************")

print( y_train)

print("******************************************")

print(y_test)

print("******************************************")

print("******************************************")
```



jupyter
SCOB77_Pratham Pitty_Group C Assignment 1 Last Checkpoint: 6 hours ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code
In [38]:

```
#part2. Import the necessary package and perform the train and test split on the dataset.
print("*****************************************")
print("SCOB77_Pratham Pitty_Group C Assignment 1")
print("*****************************************")
print("In Social_Network_Ads.csv file i have taken Data of 50 entries of 'Age',\n'Estimated Salary','Purchased'")
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset = pd.read_csv("C:\\Users\prath\Videos\second year\sem3\PDS\Social_Network_Ads.csv")
X = dataset.iloc[:,[1, 2]].values
y = dataset.iloc[:,0].values
# Splitting Data into Training & Testing
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
print(X_train)
print("*****************************************")
print(X_test)
print("*****************************************")
print( y_train)
print("*****************************************")
print(y_test)
print("*****************************************")
print("*****************************************")
```

```
*****************************************
SCOB77_Pratham Pitty_Group C Assignment 1
*****************************************
In Social_Network_Ads.csv file i have taken Data of 50 entries of 'Age',
'Estimated Salary','Purchased'
[[ 30000      1]
 [135000      1]
 [ 43000      0]
 [ 79000      0]
 [ 28000      1]
 [ 80000      0]
 [ 43000      0]
 [ 41000      1]
 [ 25000      1]
 [ 51000      0]
 [ 22000      1]
 [108000      0]
 [ 33000      0]
 [ 18000      0]
 [ 20000      1]
 [ 58000      0]
 [ 26000      1]
 [ 27000      0]
 [ 82000      0]
 [ 72000      0]
 [ 20000      0]
 [ 86000      0]
```



jupyter
SCOB77_Pratham Pitty_Group C Assignment 1 Last Checkpoint: 6 hours ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Code

```
 [ 41000      1]
 [ 25000      1]
 [ 51000      0]
 [ 22000      1]
 [108000      0]
 [ 33000      0]
 [ 18000      0]
 [ 20000      1]
 [ 58000      0]
 [ 26000      1]
 [ 27000      0]
 [ 82000      0]
 [ 72000      0]
 [ 20000      0]
 [ 86000      0]
 [ 15000      0]
 [ 23000      1]
 [ 84000      0]
 [ 22000      1]
 [ 28000      0]
 [ 49000      1]
 [ 29000      1]
 [ 65000      0]
 [ 31000      0]
 [ 20000      0]
 [ 57000      0]
 [ 19000      0]
 [ 54000      0]
 [ 84000      0]]
******************************************
[[ 18000      0]
 [ 76000      0]
 [ 28000      1]
 [ 74000      0]
 [ 16000      0]
 [ 49000      0]
 [ 90000      0]
 [ 17000      0]
 [150000      1]
 [ 80000      0]
 [ 52000      0]
 [137000      1]
 [ 44000      0]]
******************************************
[47 30 26 25 46 29 29 48 47 33 45 35 25 32 47 27 45 35 18 26 35 20 30 46
 27 45 33 47 48 35 27 23 27 19 27 28]
******************************************
[31 19 49 31 21 30 27 27 32 26 26 27 28]
******************************************
******************************************
```

```
In [ ]:
```

Conclusion:

Hence we conclude that using imputation method handle the null values also perform the train and test split on the dataset.