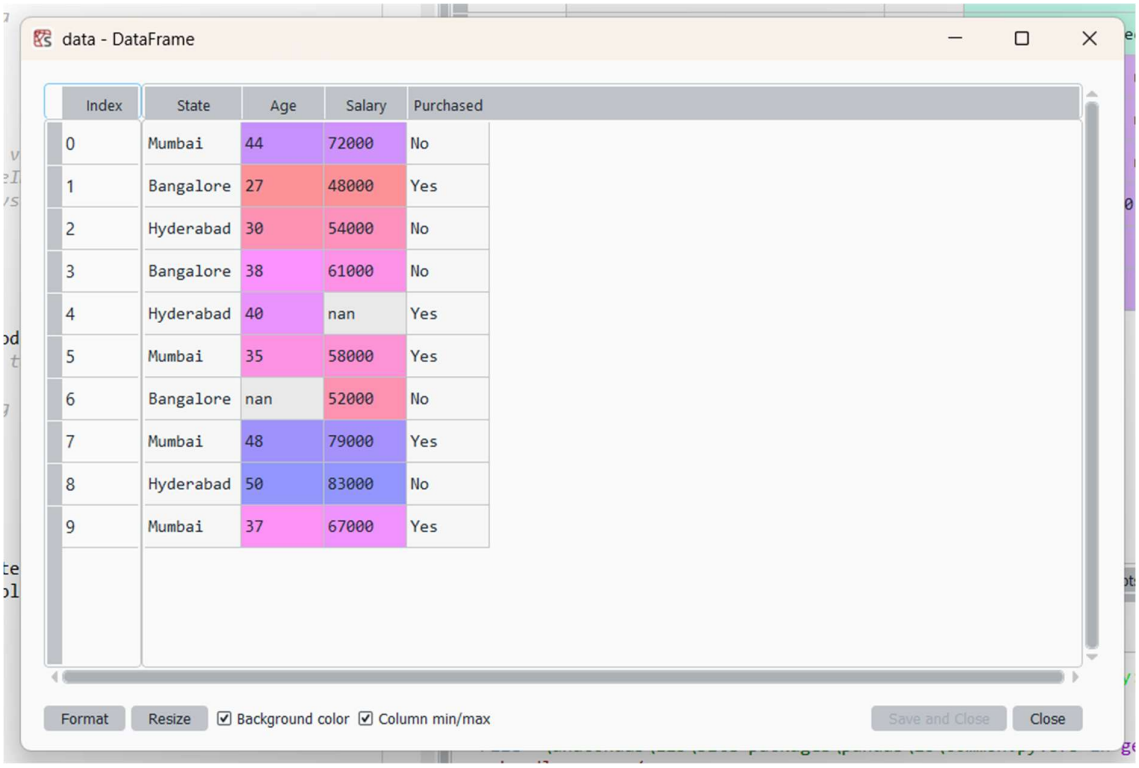
import pandas as pd
data=pd.read_csv(r"C:\Users\TharunMahendra\NIT\5.MachineLearning\Data.csv")



Index	State	Age	Salary	Purchased
0	Mumbai	44	72000	No
1	Bangalore	27	48000	Yes
2	Hyderabad	30	54000	No
3	Bangalore	38	61000	No
4	Hyderabad	40	nan	Yes
5	Mumbai	35	58000	Yes
6	Bangalore	nan	52000	No
7	Mumbai	48	79000	Yes
8	Hyderabad	50	83000	No
9	Mumbai	37	67000	Yes

#splitting the data to x&y

x=data.iloc[:,:-1].values #independent data

y=data.iloc[:,3].values #dependent data

#filling missing values

from sklearn.impute import SimpleImputer #Univariate imputer for completing missing values with simple strategies.

imputer=SimpleImputer() #Definition : SimpleImputer(*, missing_values=np.nan, strategy="mean", fill_value=None, copy=True, add_indicator=False, keep_empty_features=False)

#hyper parameter tuning if we change the system startegy to our startegy if not parameter tuning default startegy

imputer=imputer.fit(x[:,1:3])

x[:,1:3]=imputer.transform(x[:,1:3])

#transforming the categorical to integer

from sklearn.preprocessing import LabelEncoder #Encode target labels with value between 0 and n_classes-1.

#This transformer should be used to encode target values, i.e. y, and not the input X.

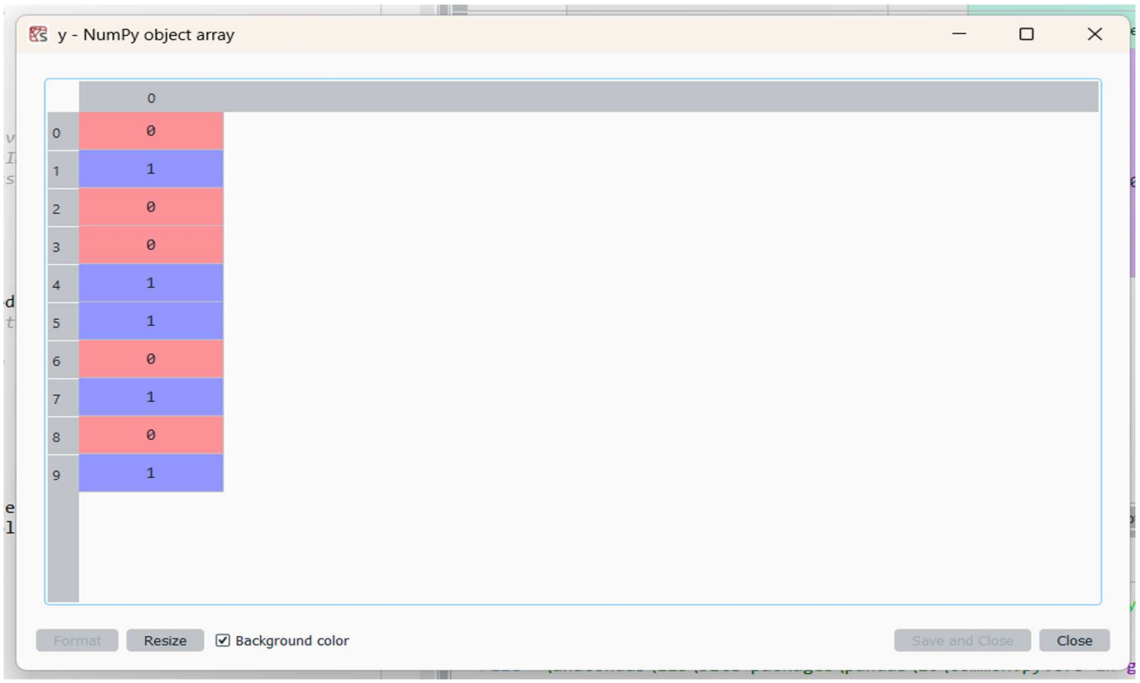
```
encoder_x=LabelEncoder()

encoder_x.fit_transform(x[:,0]) #converting states to 0,1,2..

x[:,0]=encoder_x.fit_transform(x[:,0])

encoder_y=LabelEncoder()

y=encoder_y.fit_transform(y)
```

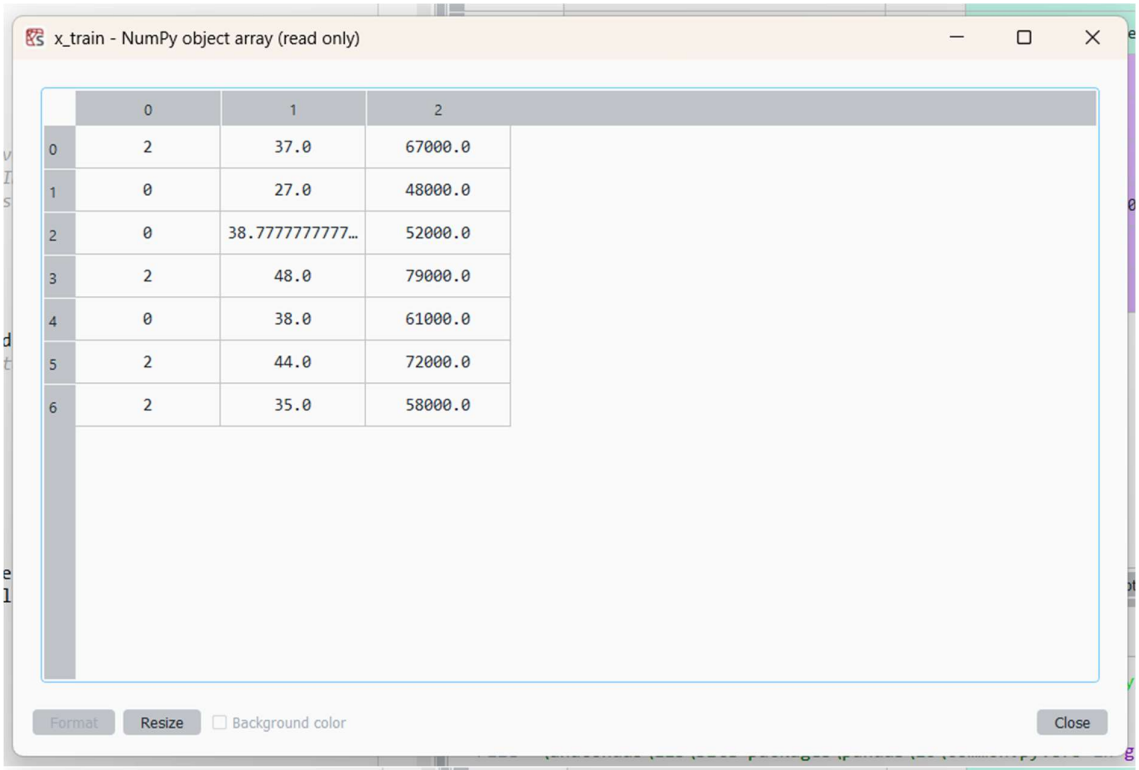


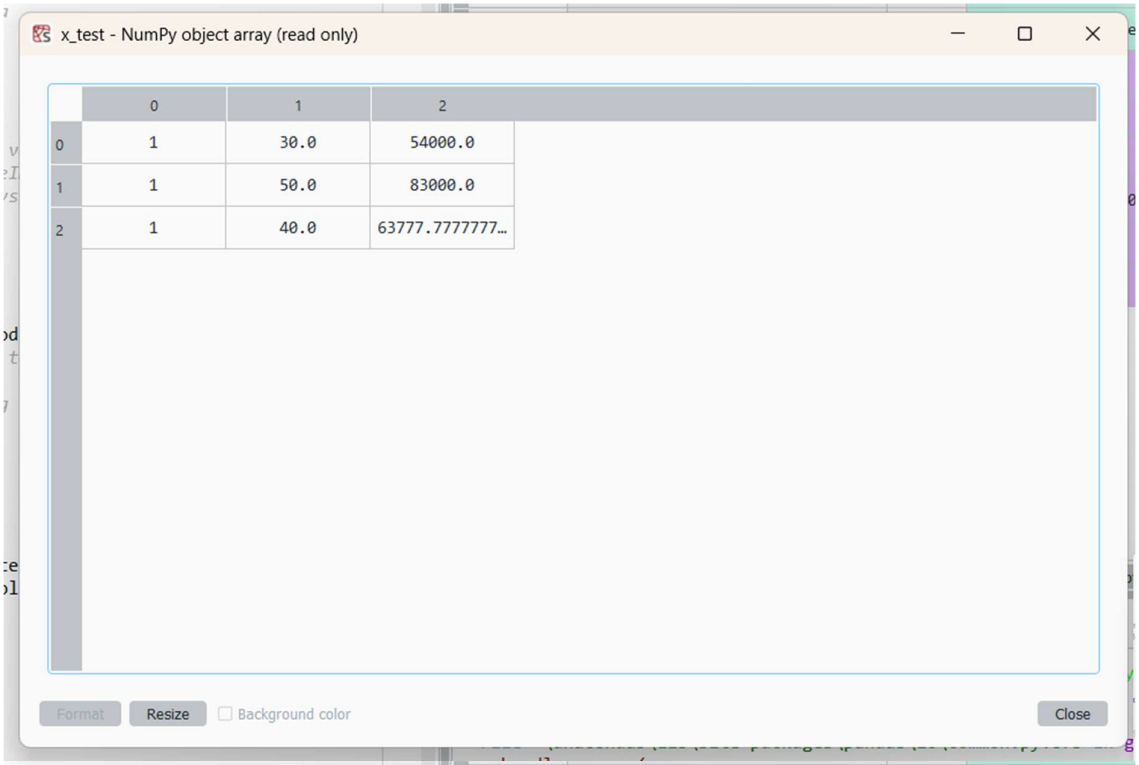
	0
0	0
1	1
2	0
3	0
4	1
5	1
6	0
7	1
8	0
9	1

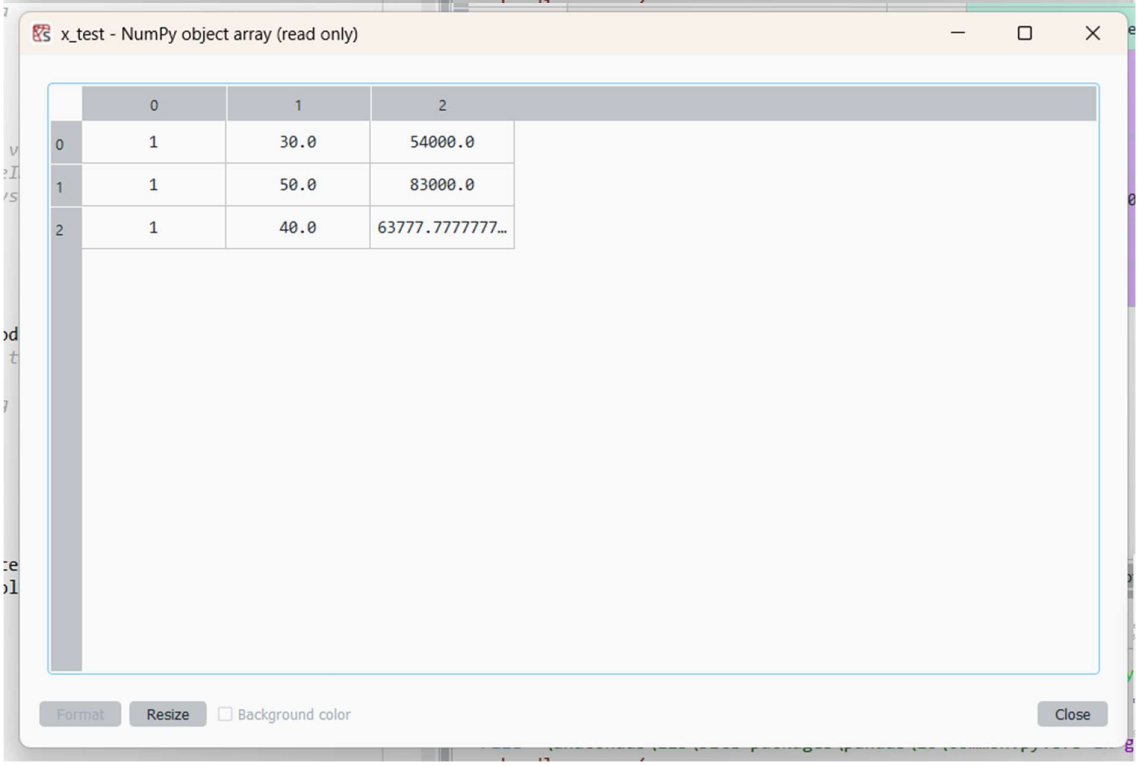
	0	1	2
0	2	44.0	72000.0
1	0	27.0	48000.0
2	1	30.0	54000.0
3	0	38.0	61000.0
4	1	40.0	63777.7777777…
5	2	35.0	58000.0
6	0	38.7777777777…	52000.0
7	2	48.0	79000.0
8	1	50.0	83000.0
9	2	37.0	67000.0

#splitting to tarin&test data

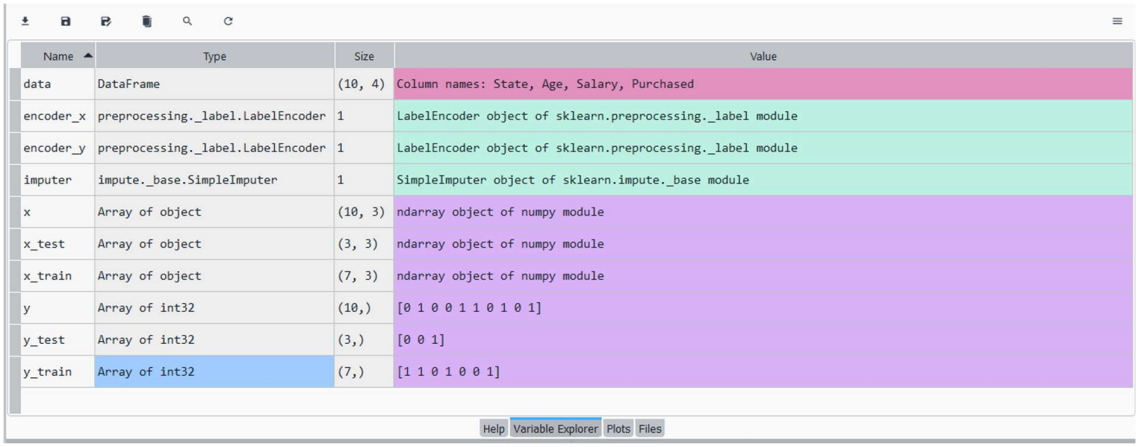
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
```




x_test - NumPy object array (read only)
0
1
2
0
1
30.0
54000.0
1
1
50.0
83000.0
2
1
40.0
63777.7777777…
Format
Resize
Background color
Close


x_test - NumPy object array (read only)
0
1
2
0
1
30.0
54000.0
1
1
50.0
83000.0
2
1
40.0
63777.7777777…
Format
Resize
Background color
Close

Variables Created Till Now



Name	Type	Size	Value
data	DataFrame	(10, 4)	Column names: State, Age, Salary, Purchased
encoder_x	preprocessing._label.LabelEncoder	1	LabelEncoder object of sklearn.preprocessing._label module
encoder_y	preprocessing._label.LabelEncoder	1	LabelEncoder object of sklearn.preprocessing._label module
imputer	impute._base.SimpleImputer	1	SimpleImputer object of sklearn.impute._base module
x	Array of object	(10, 3)	ndarray object of numpy module
x_test	Array of object	(3, 3)	ndarray object of numpy module
x_train	Array of object	(7, 3)	ndarray object of numpy module
y	Array of int32	(10,)	[0 1 0 0 1 1 0 1 0 1]
y_test	Array of int32	(3,)	[0 0 1]
y_train	Array of int32	(7,)	[1 1 0 1 0 0 1]