Data preprocessing

Code

```python
# Data preprocesing

# Importing liberies

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Import data and divide into dv & iv
dataset = pd.read_csv(r'D:\1. Professionall\Data Science\08-09-2023\Data.csv')

X = dataset.iloc[:, :-1].values
y = dataset.iloc[:,3].values

# Filling null values in X dataset

from sklearn.impute import SimpleImputer
# default startegy is mean
impute = SimpleImputer(missing_values=np.nan, strategy="median")

impute = impute.fit(X[:,1:3])
X[:, 1:3] = impute.transform(X[:,1:3])

# Convert categorical data into numerical data

from sklearn.preprocessing import LabelEncoder

labelencoder_X = LabelEncoder()
X[:,0] = labelencoder_X.fit_transform(X[:,0])


labelencoder_y = LabelEncoder()
y = labelencoder_y.fit_transform(y)

# Spliting data into training set and testing set

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)

```

Dataset

dataset - DataFrame

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

Format Resize Background color Column min/max Save and Close Close

X & y dataset

X - NumPy object array (read only)

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

Format Resize Background color Close

y - NumPy object array (read only)

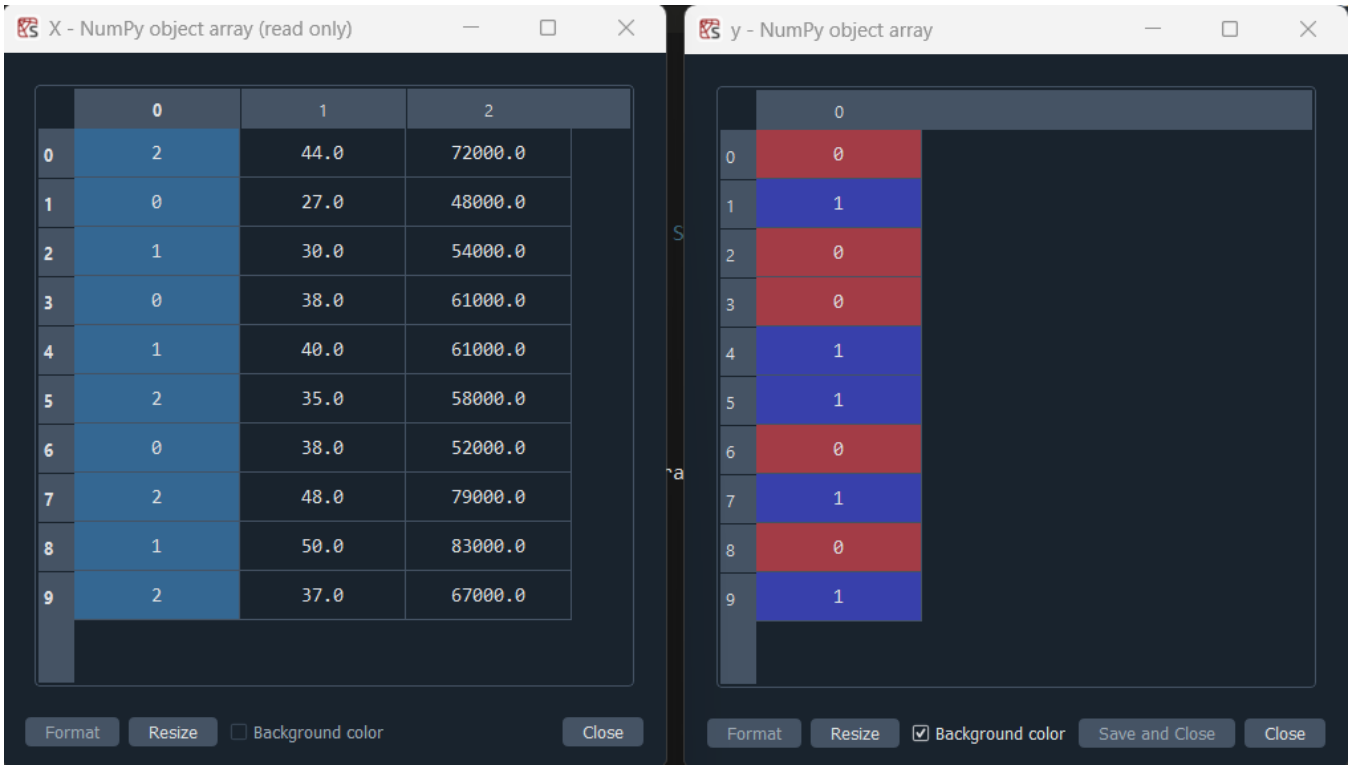
	0
0	No
1	Yes
2	No
3	No
4	Yes
5	Yes
6	No
7	Yes
8	No
9	Yes

Format Resize Background color Close

Simple imputed X dataset

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

Label encoded X & y



	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	61000.0
5	2	35.0	58000.0
6	0	38.0	52000.0
7	2	48.0	79000.0
8	1	50.0	83000.0
9	2	37.0	67000.0

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

Train test and split data