Day 34

DIY

Q1. Problem Statement: Splitting Data for Testing and Training

You are given a dataset – "*housing.csv*." Load the dataset into a DataFrame. Considering the *"median_house_value"* column as output/target and the rest of the columns as input, perform the following tasks:

1. Find the number of null values in each column and replace them with the mean values
2. Split the data into test and train fragments using `train_test_split()` function in 80:20 ratio (80% train, 20% test)
3. Print the size of test and train data (For both input and output)

Dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Sample Output:

1. Find the number of null values in each column and replace them with the mean values

```
longitude                 0
latitude                  0
housing_median_age        0
total_rooms               0
total_bedrooms          207
population                0
households                0
median_income             0
median_house_value        0
ocean_proximity           0
dtype: int64
```

```
Replacing null values with mean:
longitude                 0
latitude                  0
housing_median_age        0
total_rooms               0
total_bedrooms          207
population                0
households                0
median_income             0
median_house_value        0
ocean_proximity           0
dtype: int64
```

2. Split the data into test and train fragments using the train_test_split() function in an 80:20 ratio (80% train, 20% test) and printing the size of test and train data (For both input and output)

```
shape of original dataset : (20640, 10)
shape of input - training set (16512, 9)
shape of output - training set (16512,)
shape of input - testing set (4128, 9)
shape of output - testing set (4128,)
```