Day 35

DIY

Q1. Problem Statement: Linear Regression

Load the housing_price.csv dataset to a DataFrame and perform the following tasks:

The dataset contains only numeric data, and the median house value column is our target variable, so with the help of linear regression, build a model that can predict accurate house prices. Perform the below tasks and build a model:

1. Load the *housing_price* dataset into DataFrame
2. Find the null value and drop then, If any
3. Split data into two DataFrames x and y based on dependent and independent variables
4. Split x and y into 80% training set and 20% testing set. Set the random state to 10. Call the LinearRegression model, then fit the model using train data
5. Print the R2 value, coefficient, and intercept
6. Compare actual and predicted values.
7. Print the final summary

Dataset:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-114.31	34.19	15	5612	1283	1015	472	1.4936	66900
1	-114.47	34.40	19	7650	1901	1129	463	1.8200	80100
2	-114.56	33.69	17	720	174	333	117	1.6509	85700
3	-114.57	33.64	14	1501	337	515	226	3.1917	73400
4	-114.57	33.57	20	1454	326	624	262	1.9250	65500

Sample Output:

3. Split data into two DataFrame x and y based on dependent and independent variables

```
independent data
 [[-114.31     34.19     15.      ... 1015.      472.        1.4936]
 [-114.47     34.4      19.      ... 1129.      463.        1.82  ]
 [-114.56     33.69     17.      ...  333.      117.        1.6509]
 ...
 [-124.3      41.84     17.      ... 1244.      456.        3.0313]
 [-124.3      41.8      19.      ... 1298.      478.        1.9797]
 [-124.35     40.54     52.      ...  806.      270.        3.0147]]

dependent data
 [ 66900  80100  85700 ... 103600  85800  94600]
```

4. Split x and y into train and test data set based on test size as 0.2 and random_state as 10

```
x_train and x_test dataset shape (13600, 8) (3400, 8)
y_train and y_test dataset shape (13600,) (3400,)
```

6. Print R2 vallue,coefficient and intercept

```
R2 value: 0.6484403017760402

coefficient:
  [-4.34225673e+04 -4.34584915e+04  1.15417922e+03 -8.34683693e+00
  1.14234465e+02 -3.87425498e+01  5.04252279e+01  4.02554220e+04]

intercept: -3635200.010897698
```

7. Compare actual and predicted values.

	Actual	Predicted
0	96100	-8475.675202
1	500001	490876.394233
2	177200	112662.107990
3	55000	218093.753334
4	220800	207600.925885
5	158300	121540.170888
6	37900	180602.126583
7	115600	104694.108104
8	359700	310765.123759
9	203300	265864.990208

8. Print the final summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.641
Model:                            OLS   Adj. R-squared:                  0.641
Method:                 Least Squares   F-statistic:                     3798.
Date:                Wed, 30 Mar 2022   Prob (F-statistic):               0.00
Time:                        15:27:55   Log-Likelihood:            -2.1365e+05
No. Observations:               17000   AIC:                         4.273e+05
Df Residuals:                   16991   BIC:                         4.274e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -3.621e+06   6.92e+04    -52.312      0.000   -3.76e+06   -3.48e+06
x1         -4.314e+04    789.568    -54.637      0.000   -4.47e+04   -4.16e+04
x2         -4.293e+04    745.804    -57.556      0.000   -4.44e+04   -4.15e+04
x3          1150.6949     47.577     24.186      0.000    1057.438    1243.951
x4            -8.3783      0.863     -9.711      0.000     -10.069      -6.687
x5           117.6485      7.687     15.305      0.000     102.582     132.715
x6           -38.4888      1.186    -32.456      0.000     -40.813     -36.164
x7            45.4360      8.445      5.380      0.000      28.883      61.989
x8          4.051e+04    368.172    110.022      0.000    3.98e+04    4.12e+04
==============================================================================
Omnibus:                     4032.682   Durbin-Watson:                   1.162
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            15559.395
Skew:                           1.141   Prob(JB):                         0.00
Kurtosis:                       7.094   Cond. No.                     5.14e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.14e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```