Group6_Task_5_Output

Task-1-Standard scaler to replace null values with the mean

- Showing Dataframe after renaming the columns and also the pandas dataframe ‘df_pd’ used for computing the correlation.

```
+----------+----+-------+--------+-------+-------+--------+----------+--------+--------------------+--------------+
|    Suburb|Type|  Price|Distance|Zipcode|Bedroom|Bathroom|Car_Garage|Lot_size|         Region_name|Property_count|
+----------+----+-------+--------+-------+-------+--------+----------+--------+--------------------+--------------+
|Abbotsford|   h|1165000|     2.5|   3067|      3|       2|         0|      92|Northern Metropol...|          4019|
|Abbotsford|   h|1050000|     2.5|   3067|      2|       1|         0|     129|Northern Metropol...|          4019|
|Abbotsford|   h|1465000|     2.5|   3067|      3|       2|         0|     134|Northern Metropol...|          4019|
|Abbotsford|   h| 911000|     3.0|   3067|      2|       1|         0|     141|Northern Metropol...|          4019|
|Abbotsford|   h|1635000|     3.0|   3067|      3|       1|         0|     142|Northern Metropol...|          4019|
|Abbotsford|   h|1315000|     2.5|   3067|      2|       1|         0|     147|Northern Metropol...|          4019|
|Abbotsford|   h|1035000|     2.5|   3067|      2|       1|         0|     156|Northern Metropol...|          4019|
|Abbotsford|   h|1180000|     2.5|   3067|      2|       1|         0|     162|Northern Metropol...|          4019|
|Abbotsford|   h| 941000|     2.5|   3067|      2|       1|         0|     181|Northern Metropol...|          4019|
|Abbotsford|   h|1012500|     2.5|   3067|      2|       1|         0|     182|Northern Metropol...|          4019|
|Abbotsford|   h| 955000|     2.5|   3067|      3|       1|         0|     183|Northern Metropol...|          4019|
|Abbotsford|   h|1000000|     2.5|   3067|      3|       1|         0|     198|Northern Metropol...|          4019|
|Abbotsford|   h|1876000|     2.5|   3067|      4|       2|         0|     245|Northern Metropol...|          4019|
|Abbotsford|   h|1375000|     3.0|   3067|      3|       1|         0|     299|Northern Metropol...|          4019|
|Abbotsford|   h| 940000|     3.0|   3067|      2|       1|         0|     424|Northern Metropol...|          4019|
|Abbotsford|   h| 855000|     3.0|   3067|      3|       2|         1|      92|Northern Metropol...|          4019|
|Abbotsford|   h| 850000|     2.5|   3067|      3|       2|         1|      94|Northern Metropol...|          4019|
|Abbotsford|   h|1200000|     2.5|   3067|      3|       2|         1|     113|Northern Metropol...|          4019|
|Abbotsford|   h|1195000|     2.5|   3067|      3|       2|         1|     120|Northern Metropol...|          4019|
|Abbotsford|   h|1100000|     2.5|   3067|      2|       2|         1|     124|Northern Metropol...|          4019|
+----------+----+-------+--------+-------+-------+--------+----------+--------+--------------------+--------------+
only showing top 20 rows

>>> df_pd
         Price  Distance  Zipcode  #Bedroom  #Bathroom  ...  Lot_size  Property_count  Suburb_indexed  Type_indexed  Region_name_indexed
0       950000      11.0     3018         3          1  ...       399            5301              52             0                    2
1       931000      11.1     3025         3          2  ...       590            5132              69             0                    2
2       565000       5.9     3032         2          2  ...       103            6567              24             2                    2
3      2215000      11.0     3147         4          2  ...       681            3052              62             0                    0
4      1650000      11.0     3147         3          1  ...       697            3052              62             0                    0
...        ...       ...      ...       ...        ...  ...       ...             ...             ...           ...                  ...
11574   980000       1.9     3205         2          1  ...        64            5943              75             0                    0
11575  2640000       8.0     3016         4          2  ...       650            6380              37             0                    2
11576   600000      20.6     3064         3          2  ...       630           15510              84             0                    1
11577  1150000       7.5     3040         3          1  ...       658            9264               5             0                    2
11578  2500000      13.8     3188         5          4  ...       661            5454              23             0                    0

[11579 rows x 11 columns]
```

- Output obtained from the standard scaler and storing the output in the new columns.

```
+----------+--------+-----------+------------+
|Car_Garage|Bathroom|out_garrage|out_Bathroom|
+----------+--------+-----------+------------+
|         0|       2|          0|           2|
|         0|       1|          0|           1|
|         0|       2|          0|           2|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         0|       2|          0|           2|
|         0|       1|          0|           1|
|         0|       1|          0|           1|
|         1|       2|          1|           2|
|         1|       2|          1|           2|
|         1|       2|          1|           2|
|         1|       2|          1|           2|
|         1|       2|          1|           2|
+----------+--------+-----------+------------+
only showing top 20 rows
```

Task-2 Correlating the label “Price” with all features

```
>>> print("Distance correlation is   :",Distance_corr)
Distance correlation is   : -0.2621722992278138
>>> print("Zipcode correlation is    :",Zipcode_corr)
Zipcode correlation is    : 0.1139082643230997
>>> print("Bedroom correlation is    :",Bedroom_corr)
Bedroom correlation is    : 0.4034004631219416
>>> print("Bathroom correlation is   :",Bathroom_corr)
Bathroom correlation is   : 0.43529085038326526
>>> print("Car_Garage correlation is :",Car_Garage_corr)
Car_Garage correlation is : 0.18018501693287553
>>> print("Lot_size correlation is   :",Lot_size_corr)
Lot_size correlation is   : 0.023238431407413256
>>> print("Property correlation is            :",Property_count_corr)
Property correlation is            : -0.02492886748660884
>>> print("Suburb_indexed correlation is      :",Suburb_indexed_corr)
Suburb_indexed correlation is      : -0.16142479313158709
>>> print("Type_indexed correlation is        :",Type_indexed_corr)
Type_indexed correlation is        : -0.24144346825030533
>>> print("Region_name_indexed correlation is :",Region_name_indexed_corr)
Region_name_indexed correlation is : -0.3119131290970111
>>> print("Lot_size correlation is            :",Lot_size_corr)
Lot_size correlation is            : 0.023238431407413256
```

The below figure shows the Pandas dataframe ‘df_pd’

	Suburb	Type	Price	Distance	...	Property_count	Suburb_indexed	Type_indexed	Region_name_indexed
0	Abbotsford	h	1165000	2.5	...	4019	89.0	0.0	1.0
1	Abbotsford	h	1050000	2.5	...	4019	89.0	0.0	1.0
2	Abbotsford	h	1465000	2.5	...	4019	89.0	0.0	1.0
3	Abbotsford	h	911000	3.0	...	4019	89.0	0.0	1.0
4	Abbotsford	h	1635000	3.0	...	4019	89.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...
11574	Yarraville	h	985500	7.0	...	6543	13.0	0.0	2.0
11575	Yarraville	h	725000	7.0	...	6543	13.0	0.0	2.0
11576	Yarraville	h	1240000	6.3	...	6543	13.0	0.0	2.0
11577	Yarraville	h	1400000	6.3	...	6543	13.0	0.0	2.0
11578	Yarraville	h	1100000	6.3	...	6543	13.0	0.0	2.0

Task-3(a)Predicting label ‘price’ with highest correlation features

```
>>> model1 = pipeline1.fit(trainingData1)
>>> predictions1 = model1.transform(testData1)
>>> predictions1.select("prediction", "Price","normFeatures").show(7)
+------------------+-------+--------------------+
|        prediction|  Price|        normFeatures|
+------------------+-------+--------------------+
|1238826.4260338494| 911000|[5.93173360458614...|
|1258446.2475492414|1030000|[4.94217421685400...|
|1386780.7772708095|1100000|[4.94354657113562...|
|1362432.8832976627|1190000|[5.93153278861791...|
|1410138.7364272473|1290000|[4.94294348328334...|
|1609240.8243209526|1350000|[4.94060829941353...|
|1351083.3519281223|1480000|[4.94109105289031...|
+------------------+-------+--------------------+
only showing top 7 rows
```

Task-4(a) RMSE metric

```
>>> print(rmse)
498723.06416245725
```

Task-3(b)Predicting label ‘price’ with all features

```
>>> model1 = pipeline1.fit(trainingData1)
>>> predictions1 = model1.transform(testData1)
>>> predictions1.select("prediction", "Price","normFeatures").show(7)
+------------------+-------+--------------------+
|        prediction|  Price|        normFeatures|
+------------------+-------+--------------------+
|1238826.4260338494| 911000|[5.93173360458614...|
|1258446.2475492414|1030000|[4.94217421685400...|
|1386780.7772708095|1100000|[4.94354657113562...|
|1362432.8832976627|1190000|[5.93153278861791...|
|1410138.7364272473|1290000|[4.94294348328334...|
|1609240.8243209526|1350000|[4.94060829941353...|
|1351083.3519281223|1480000|[4.94109105289031...|
+------------------+-------+--------------------+
only showing top 7 rows
```

Task-4(b)RMSE Metric

```
>>> print(rmse1)
 454976.4144312548
```