task2-houseprice

July 8, 2024

1 House Price Prediction

[1]:
```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
house_df = pd.read_csv("HousePricePrediction.xlsx - Sheet1.csv")
```

[3]:
```
house_df.head()
```

[3]:
```
   Id  MSSubClass MSZoning  LotArea LotConfig BldgType  OverallCond  \
0   0          60       RL     8450    Inside     1Fam            5
1   1          20       RL     9600       FR2     1Fam            8
2   2          60       RL    11250    Inside     1Fam            5
3   3          70       RL     9550    Corner     1Fam            5
4   4          60       RL    14260       FR2     1Fam            5

   YearBuilt  YearRemodAdd Exterior1st  BsmtFinSF2  TotalBsmtSF  SalePrice
0       2003          2003     VinylSd         0.0        856.0   208500.0
1       1976          1976     MetalSd         0.0       1262.0   181500.0
2       2001          2002     VinylSd         0.0        920.0   223500.0
3       1915          1970     Wd Sdng         0.0        756.0   140000.0
4       2000          2000     VinylSd         0.0       1145.0   250000.0
```

[4]:
```
house_df.shape
```

[4]:
```
(2919, 13)
```

[5]:
```
house_df.drop_duplicates(inplace=True)
```

[6]:
```
house_df.shape
```

[6]:
```
(2919, 13)
```

[7]:
```
house_df.columns
```

[7]: Index(['Id', 'MSSubClass', 'MSZoning', 'LotArea', 'LotConfig', 'BldgType',
       'OverallCond', 'YearBuilt', 'YearRemodAdd', 'Exterior1st', 'BsmtFinSF2',
       'TotalBsmtSF', 'SalePrice'],
      dtype='object')

[8]:
```
house_df.drop(columns = ['Id'], inplace = True)
```

[9]:
```
house_df.shape
```

[9]: (2919, 12)

[10]:
```
house_df.isna().sum()
```

[10]:
```
MSSubClass         0
MSZoning           4
LotArea            0
LotConfig          0
BldgType           0
OverallCond        0
YearBuilt          0
YearRemodAdd       0
Exterior1st        1
BsmtFinSF2         1
TotalBsmtSF        1
SalePrice       1459
dtype: int64
```

[11]:
```
from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy='mean')

imputer.fit(house_df[['SalePrice']])
```

[11]: SimpleImputer()

[12]:
```
imputer.statistics_      # all null values of saleprice will be replaced by mean␣
 ↪of salePrice which is 180921.1958
```

[12]: array([180921.19589041])

[13]:
```
house_df['SalePrice'] = imputer.transform(house_df[['SalePrice']])
```

[14]:
```
house_df.isna().sum()
```

[14]:
```
MSSubClass      0
MSZoning        4
LotArea         0
```

```
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     1
BsmtFinSF2      1
TotalBsmtSF     1
SalePrice       0
dtype: int64
```

[15]:
```
house_df = house_df.fillna(0)
```

[16]:
```
house_df.isna().sum()  #we have replaced the rest null values with 0. now we
 ↪dont have have null values
```

[16]:
```
MSSubClass      0
MSZoning        0
LotArea         0
LotConfig       0
BldgType        0
OverallCond     0
YearBuilt       0
YearRemodAdd    0
Exterior1st     0
BsmtFinSF2      0
TotalBsmtSF     0
SalePrice       0
dtype: int64
```

[17]:
```
house_df.describe()
```

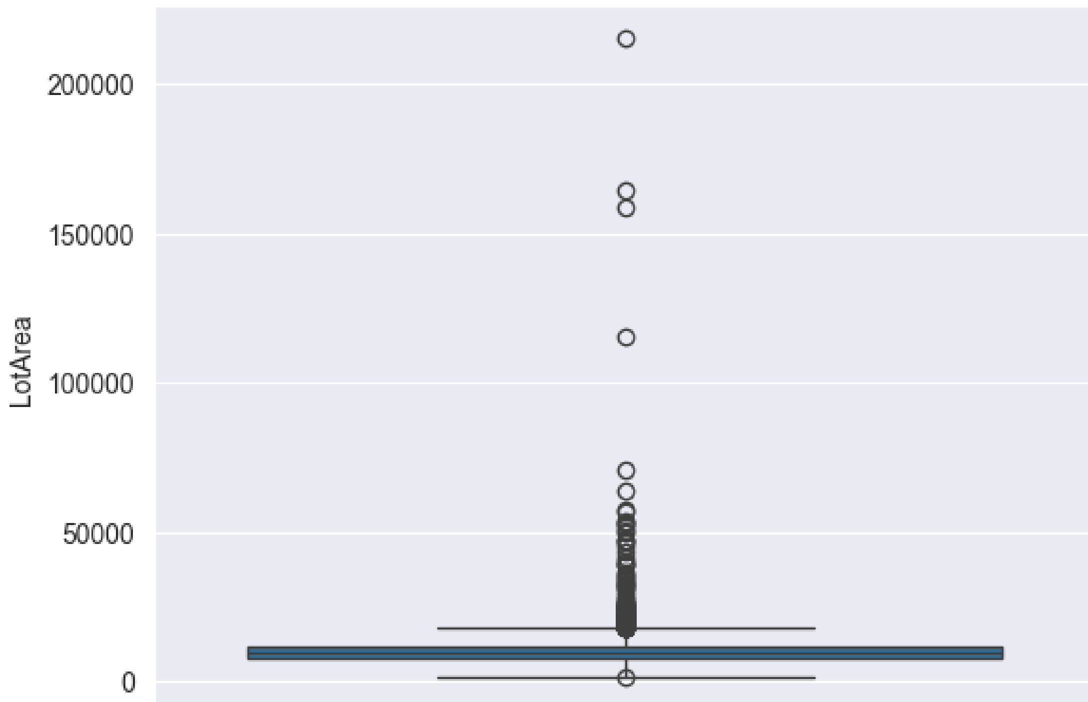
[17]:
```
        MSSubClass        LotArea  OverallCond    YearBuilt  YearRemodAdd  \
count  2919.000000    2919.000000  2919.000000  2919.000000   2919.000000
mean     57.137718   10168.114080     5.564577  1971.312778   1984.264474
std      42.517628    7886.996359     1.113131    30.291442     20.894344
min      20.000000    1300.000000     1.000000  1872.000000   1950.000000
25%      20.000000    7478.000000     5.000000  1953.500000   1965.000000
50%      50.000000    9453.000000     5.000000  1973.000000   1993.000000
75%      70.000000   11570.000000     6.000000  2001.000000   2004.000000
max     190.000000  215245.000000     9.000000  2010.000000   2010.000000

        BsmtFinSF2  TotalBsmtSF      SalePrice
count  2919.000000  2919.000000    2919.000000
mean     49.565262  1051.417266  180921.195890
std     169.179104   441.120498   56174.332503
min       0.000000     0.000000   34900.000000
```

```
25%        0.000000   793.000000  163000.000000
50%        0.000000   989.000000  180921.195890
75%        0.000000  1302.000000  180921.195890
max     1526.000000  6110.000000  755000.000000
```

[18]:
```
import matplotlib.pyplot as plt
import seaborn as sns
```

[19]:
```
sns.set_style('darkgrid')
sns.boxplot(house_df, y = 'LotArea');
```



[20]:
```
import numpy as np

Q1 = np.percentile(house_df['LotArea'], 25, interpolation = 'midpoint')
Q3 = np.percentile(house_df['LotArea'], 75, interpolation = 'midpoint')

IQR = Q3 - Q1
```

[21]:
```
lowerBound = Q1 - 1.5 * IQR
upperBound = Q1 + 1.5 * IQR
```

[22]:
```
df = house_df[(house_df.LotArea < upperBound) & (house_df.LotArea > lowerBound)]
df
```

[22]:
```
      MSSubClass MSZoning  LotArea LotConfig BldgType  OverallCond  YearBuilt  \
0             60       RL     8450    Inside     1Fam            5       2003
1             20       RL     9600       FR2     1Fam            8       1976
2             60       RL    11250    Inside     1Fam            5       2001
3             70       RL     9550    Corner     1Fam            5       1915
6             20       RL    10084    Inside     1Fam            5       2004
...          ...      ...      ...       ...      ...          ...        ...
2913         160       RM     1526    Inside    Twnhs            5       1970
2914         160       RM     1936    Inside    Twnhs            7       1970
2915         160       RM     1894    Inside   TwnhsE            5       1970
2917          85       RL    10441    Inside     1Fam            5       1992
2918          60       RL     9627    Inside     1Fam            5       1993

      YearRemodAdd Exterior1st  BsmtFinSF2  TotalBsmtSF     SalePrice
0             2003     VinylSd         0.0        856.0  208500.00000
1             1976     MetalSd         0.0       1262.0  181500.00000
2             2002     VinylSd         0.0        920.0  223500.00000
3             1970     Wd Sdng         0.0        756.0  140000.00000
6             2005     VinylSd         0.0       1686.0  307000.00000
...            ...         ...         ...          ...           ...
2913          1970     CemntBd         0.0        546.0  180921.19589
2914          1970     CemntBd         0.0        546.0  180921.19589
2915          1970     CemntBd         0.0        546.0  180921.19589
2917          1992     HdBoard         0.0        912.0  180921.19589
2918          1994     HdBoard         0.0        996.0  180921.19589

[2550 rows x 12 columns]
```

[23]:
```python
df.MSZoning.unique()
```

[23]:
```
array(['RL', 'RM', 'C (all)', 'FV', 'RH'], dtype=object)
```

[24]:
```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 2550 entries, 0 to 2918
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   MSSubClass    2550 non-null   int64
 1   MSZoning      2550 non-null   object
 2   LotArea       2550 non-null   int64
 3   LotConfig     2550 non-null   object
 4   BldgType      2550 non-null   object
 5   OverallCond   2550 non-null   int64
 6   YearBuilt     2550 non-null   int64
 7   YearRemodAdd  2550 non-null   int64
```

```
 8   Exterior1st   2550 non-null   object
 9   BsmtFinSF2    2550 non-null   float64
 10  TotalBsmtSF   2550 non-null   float64
 11  SalePrice     2550 non-null   float64
dtypes: float64(3), int64(5), object(4)
memory usage: 259.0+ KB
```

[25]:
```
cat_cols = df.select_dtypes('object').columns.tolist() # gives us all the␣
↪columns that are categorical
```

[26]:
```
cat_cols
```

[26]:
```
['MSZoning', 'LotConfig', 'BldgType', 'Exterior1st']
```

[28]:
```
from sklearn.preprocessing import OneHotEncoder

encoder = OneHotEncoder(sparse_output=False, handle_unknown='ignore')

encoder.fit(df[cat_cols])
```

[28]:
```
OneHotEncoder(handle_unknown='ignore', sparse_output=False)
```

[29]:
```
encoded_cols = encoder.get_feature_names_out(cat_cols)
encoded_cols
```

[29]:
```
array(['MSZoning_C (all)', 'MSZoning_FV', 'MSZoning_RH', 'MSZoning_RL',
       'MSZoning_RM', 'LotConfig_Corner', 'LotConfig_CulDSac',
       'LotConfig_FR2', 'LotConfig_FR3', 'LotConfig_Inside',
       'BldgType_1Fam', 'BldgType_2fmCon', 'BldgType_Duplex',
       'BldgType_Twnhs', 'BldgType_TwnhsE', 'Exterior1st_AsbShng',
       'Exterior1st_AsphShn', 'Exterior1st_BrkComm',
       'Exterior1st_BrkFace', 'Exterior1st_CBlock', 'Exterior1st_CemntBd',
       'Exterior1st_HdBoard', 'Exterior1st_ImStucc',
       'Exterior1st_MetalSd', 'Exterior1st_Plywood', 'Exterior1st_Stucco',
       'Exterior1st_VinylSd', 'Exterior1st_Wd Sdng',
       'Exterior1st_WdShing'], dtype=object)
```

[30]:
```
df[encoded_cols] = encoder.transform(df[cat_cols])
df
```

[30]:
```
      MSSubClass MSZoning  LotArea LotConfig BldgType  OverallCond  YearBuilt  \
0             60       RL     8450    Inside     1Fam            5       2003
1             20       RL     9600       FR2     1Fam            8       1976
2             60       RL    11250    Inside     1Fam            5       2001
3             70       RL     9550    Corner     1Fam            5       1915
6             20       RL    10084    Inside     1Fam            5       2004
…              …        …        …         …        …            …          …
```

```
2913          160        RM     1526    Inside    Twnhs             5       1970
2914          160        RM     1936    Inside    Twnhs             7       1970
2915          160        RM     1894    Inside   TwnhsE             5       1970
2917           85        RL    10441    Inside     1Fam             5       1992
2918           60        RL     9627    Inside     1Fam             5       1993

      YearRemodAdd Exterior1st  BsmtFinSF2  …  Exterior1st_CBlock  \
0             2003     VinylSd         0.0  …                 0.0
1             1976     MetalSd         0.0  …                 0.0
2             2002     VinylSd         0.0  …                 0.0
3             1970     Wd Sdng         0.0  …                 0.0
6             2005     VinylSd         0.0  …                 0.0
…              …          …          …  …                 …
2913          1970     CemntBd         0.0  …                 0.0
2914          1970     CemntBd         0.0  …                 0.0
2915          1970     CemntBd         0.0  …                 0.0
2917          1992     HdBoard         0.0  …                 0.0
2918          1994     HdBoard         0.0  …                 0.0

      Exterior1st_CemntBd  Exterior1st_HdBoard  Exterior1st_ImStucc  \
0                     0.0                  0.0                  0.0
1                     0.0                  0.0                  0.0
2                     0.0                  0.0                  0.0
3                     0.0                  0.0                  0.0
6                     0.0                  0.0                  0.0
…                     …                    …                    …
2913                  1.0                  0.0                  0.0
2914                  1.0                  0.0                  0.0
2915                  1.0                  0.0                  0.0
2917                  0.0                  1.0                  0.0
2918                  0.0                  1.0                  0.0

      Exterior1st_MetalSd  Exterior1st_Plywood  Exterior1st_Stucco  \
0                     0.0                  0.0                 0.0
1                     1.0                  0.0                 0.0
2                     0.0                  0.0                 0.0
3                     0.0                  0.0                 0.0
6                     0.0                  0.0                 0.0
…                     …                    …                   …
2913                  0.0                  0.0                 0.0
2914                  0.0                  0.0                 0.0
2915                  0.0                  0.0                 0.0
2917                  0.0                  0.0                 0.0
2918                  0.0                  0.0                 0.0

      Exterior1st_VinylSd  Exterior1st_Wd Sdng  Exterior1st_WdShing
0                     1.0                  0.0                  0.0
```

```
1                    0.0                  0.0                  0.0
2                    1.0                  0.0                  0.0
3                    0.0                  1.0                  0.0
6                    1.0                  0.0                  0.0
…                  …                  …                  …
2913                 0.0                  0.0                  0.0
2914                 0.0                  0.0                  0.0
2915                 0.0                  0.0                  0.0
2917                 0.0                  0.0                  0.0
2918                 0.0                  0.0                  0.0

[2550 rows x 41 columns]
```

[31]:
```python
df.drop(columns=cat_cols, inplace=True)
df
```

[31]:
```
      MSSubClass  LotArea  OverallCond  YearBuilt  YearRemodAdd  BsmtFinSF2  \
0             60     8450            5       2003          2003         0.0
1             20     9600            8       1976          1976         0.0
2             60    11250            5       2001          2002         0.0
3             70     9550            5       1915          1970         0.0
6             20    10084            5       2004          2005         0.0
…            …      …          …        …          …           …
2913         160     1526            5       1970          1970         0.0
2914         160     1936            7       1970          1970         0.0
2915         160     1894            5       1970          1970         0.0
2917          85    10441            5       1992          1992         0.0
2918          60     9627            5       1993          1994         0.0

      TotalBsmtSF     SalePrice  MSZoning_C (all)  MSZoning_FV  …  \
0           856.0  208500.00000               0.0          0.0  …
1          1262.0  181500.00000               0.0          0.0  …
2           920.0  223500.00000               0.0          0.0  …
3           756.0  140000.00000               0.0          0.0  …
6          1686.0  307000.00000               0.0          0.0  …
…            …           …               …          … …
2913        546.0  180921.19589               0.0          0.0  …
2914        546.0  180921.19589               0.0          0.0  …
2915        546.0  180921.19589               0.0          0.0  …
2917        912.0  180921.19589               0.0          0.0  …
2918        996.0  180921.19589               0.0          0.0  …

      Exterior1st_CBlock  Exterior1st_CemntBd  Exterior1st_HdBoard  \
0                    0.0                  0.0                  0.0
1                    0.0                  0.0                  0.0
2                    0.0                  0.0                  0.0
3                    0.0                  0.0                  0.0
```

```
6                     0.0                   0.0                   0.0
…                    …                    …                    …
2913                  0.0                   1.0                   0.0
2914                  0.0                   1.0                   0.0
2915                  0.0                   1.0                   0.0
2917                  0.0                   0.0                   1.0
2918                  0.0                   0.0                   1.0

      Exterior1st_ImStucc  Exterior1st_MetalSd  Exterior1st_Plywood  \
0                     0.0                  0.0                  0.0
1                     0.0                  1.0                  0.0
2                     0.0                  0.0                  0.0
3                     0.0                  0.0                  0.0
6                     0.0                  0.0                  0.0
…                    …                   …                   …
2913                  0.0                  0.0                  0.0
2914                  0.0                  0.0                  0.0
2915                  0.0                  0.0                  0.0
2917                  0.0                  0.0                  0.0
2918                  0.0                  0.0                  0.0

      Exterior1st_Stucco  Exterior1st_VinylSd  Exterior1st_Wd Sdng  \
0                    0.0                  1.0                  0.0
1                    0.0                  0.0                  0.0
2                    0.0                  1.0                  0.0
3                    0.0                  0.0                  1.0
6                    0.0                  1.0                  0.0
…                   …                   …                   …
2913                 0.0                  0.0                  0.0
2914                 0.0                  0.0                  0.0
2915                 0.0                  0.0                  0.0
2917                 0.0                  0.0                  0.0
2918                 0.0                  0.0                  0.0

      Exterior1st_WdShing
0                     0.0
1                     0.0
2                     0.0
3                     0.0
6                     0.0
…                    …
2913                  0.0
2914                  0.0
2915                  0.0
2917                  0.0
2918                  0.0
```

```
[2550 rows x 37 columns]
```

[32]:
```python
df.columns
```

[32]:
```
Index(['MSSubClass', 'LotArea', 'OverallCond', 'YearBuilt', 'YearRemodAdd',
       'BsmtFinSF2', 'TotalBsmtSF', 'SalePrice', 'MSZoning_C (all)',
       'MSZoning_FV', 'MSZoning_RH', 'MSZoning_RL', 'MSZoning_RM',
       'LotConfig_Corner', 'LotConfig_CulDSac', 'LotConfig_FR2',
       'LotConfig_FR3', 'LotConfig_Inside', 'BldgType_1Fam', 'BldgType_2fmCon',
       'BldgType_Duplex', 'BldgType_Twnhs', 'BldgType_TwnhsE',
       'Exterior1st_AsbShng', 'Exterior1st_AsphShn', 'Exterior1st_BrkComm',
       'Exterior1st_BrkFace', 'Exterior1st_CBlock', 'Exterior1st_CemntBd',
       'Exterior1st_HdBoard', 'Exterior1st_ImStucc', 'Exterior1st_MetalSd',
       'Exterior1st_Plywood', 'Exterior1st_Stucco', 'Exterior1st_VinylSd',
       'Exterior1st_Wd Sdng', 'Exterior1st_WdShing'],
      dtype='object')
```

[33]:
```python
X = df.drop(columns = 'SalePrice')
y = df['SalePrice']
```

[36]:
```python
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

scaler.fit(X)
```

[36]:
```
MinMaxScaler()
```

[37]:
```python
X[:] = scaler.transform(X)  # FOR data normalisation or standardisation i.e,␣
↪bringing the data into one scale with 1 as highest value and 0 as lowest
```

[38]:
```python
X
```

[38]:
```
      MSSubClass   LotArea  OverallCond  YearBuilt  YearRemodAdd  BsmtFinSF2  \
0       0.235294  0.574722        0.500   0.949275      0.883333         0.0
1       0.000000  0.669411        0.875   0.753623      0.433333         0.0
2       0.235294  0.805270        0.500   0.934783      0.866667         0.0
3       0.294118  0.665294        0.500   0.311594      0.333333         0.0
6       0.000000  0.709263        0.500   0.956522      0.916667         0.0
…            …         …           …          …            …           …
2913    0.823529  0.004611        0.500   0.710145      0.333333         0.0
2914    0.823529  0.038370        0.750   0.710145      0.333333         0.0
2915    0.823529  0.034911        0.500   0.710145      0.333333         0.0
2917    0.382353  0.738658        0.500   0.869565      0.700000         0.0
2918    0.235294  0.671634        0.500   0.876812      0.733333         0.0

      TotalBsmtSF  MSZoning_C (all)  MSZoning_FV  MSZoning_RH  …  \
```

```
0       0.266999                0.0         0.0         0.0  …
1       0.393637                0.0         0.0         0.0  …
2       0.286962                0.0         0.0         0.0  …
3       0.235808                0.0         0.0         0.0  …
6       0.525889                0.0         0.0         0.0  …
…          …               …         …          …  …
2913    0.170306                0.0         0.0         0.0  …
2914    0.170306                0.0         0.0         0.0  …
2915    0.170306                0.0         0.0         0.0  …
2917    0.284467                0.0         0.0         0.0  …
2918    0.310667                0.0         0.0         0.0  …

      Exterior1st_CBlock  Exterior1st_CemntBd  Exterior1st_HdBoard  \
0                    0.0                  0.0                  0.0
1                    0.0                  0.0                  0.0
2                    0.0                  0.0                  0.0
3                    0.0                  0.0                  0.0
6                    0.0                  0.0                  0.0
…                  …                  …                  …
2913                 0.0                  1.0                  0.0
2914                 0.0                  1.0                  0.0
2915                 0.0                  1.0                  0.0
2917                 0.0                  0.0                  1.0
2918                 0.0                  0.0                  1.0

      Exterior1st_ImStucc  Exterior1st_MetalSd  Exterior1st_Plywood  \
0                     0.0                  0.0                  0.0
1                     0.0                  1.0                  0.0
2                     0.0                  0.0                  0.0
3                     0.0                  0.0                  0.0
6                     0.0                  0.0                  0.0
…                   …                  …                  …
2913                  0.0                  0.0                  0.0
2914                  0.0                  0.0                  0.0
2915                  0.0                  0.0                  0.0
2917                  0.0                  0.0                  0.0
2918                  0.0                  0.0                  0.0

      Exterior1st_Stucco  Exterior1st_VinylSd  Exterior1st_Wd Sdng  \
0                    0.0                  1.0                  0.0
1                    0.0                  0.0                  0.0
2                    0.0                  1.0                  0.0
3                    0.0                  0.0                  1.0
6                    0.0                  1.0                  0.0
…                  …                  …                  …
2913                 0.0                  0.0                  0.0
2914                 0.0                  0.0                  0.0
```

```
2915                 0.0                 0.0                 0.0
2917                 0.0                 0.0                 0.0
2918                 0.0                 0.0                 0.0

      Exterior1st_WdShing
0                     0.0
1                     0.0
2                     0.0
3                     0.0
6                     0.0
…                    …
2913                  0.0
2914                  0.0
2915                  0.0
2917                  0.0
2918                  0.0

[2550 rows x 36 columns]
```

[39]:
```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
 ↪random_state=42)
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

[39]:
```
((2040, 36), (510, 36), (2040,), (510,))
```

[40]:
```python
from sklearn.linear_model import LinearRegression

model = LinearRegression()

model.fit(X_train, y_train)
```

[40]:
```
LinearRegression()
```

[41]:
```python
y_pred = model.predict(X_test)
y_test[:5]
```

[41]:
```
67      226000.00000
226     290000.00000
2546    180921.19589
268     120500.00000
2174    180921.19589
Name: SalePrice, dtype: float64
```

[42]:
```python
y_pred[:5]
```

[42]: array([205612., 202536., 174344., 138276., 212236.])

[43]:
```
from sklearn.metrics import mean_absolute_error
mean_absolute_error(y_test, y_pred)
```

[43]: 29950.854539349984

[44]:
```
from sklearn.linear_model import Lasso

lasso_reg = Lasso(alpha=50, max_iter=100, tol = 0.1)

lasso_reg.fit(X_train, y_train)
```

[44]: Lasso(alpha=50, max_iter=100, tol=0.1)

[45]:
```
lasso_pred = lasso_reg.predict(X_test)
mean_absolute_error(y_test, lasso_pred)
```

[45]: 29916.55331889878

[ ]:
```

```