

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score


file_path = "house price data.csv"

df = pd.read_csv(file_path)

print("Initial Dataset Info:")

print(df.info())

print("\nMissing Values:\n", df.isnull().sum())

df.fillna(df.median(numeric_only=True), inplace=True)

df.fillna(df.mode().iloc[0], inplace=True)

label_encoders = {}

for col in df.select_dtypes(include=['object']).columns:

    le = LabelEncoder()

    df[col] = le.fit_transform(df[col])

    label_encoders[col] = le

target_column = "price"

features = df.drop(columns=[target_column])

target = df[target_column]

X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2,
random_state=42)

scaler = StandardScaler()

X_train = scaler.fit_transform(X_train)
```

```

X_test = scaler.transform(X_test)

model = LinearRegression()

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)

mse = mean_squared_error(y_test, y_pred)

rmse = np.sqrt(mse)

r2 = r2_score(y_test, y_pred)


print(f"\nModel Performance:")

print(f"Mean Absolute Error (MAE): {mae}")

print(f"Mean Squared Error (MSE): {mse}")

print(f"Root Mean Squared Error (RMSE): {rmse}")

print(f"R2 Score: {r2}")

plt.figure(figsize=(10, 6))

plt.scatter(y_test, y_pred, alpha=0.5, color="blue")

plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red',
linestyle='dashed')

plt.xlabel("Actual Prices")

plt.ylabel("Predicted Prices")

plt.title("Actual vs Predicted House Prices")

plt.show()

```

output:

Initial Dataset Info:

```
<class 'pandas.core.frame.DataFrame'>
```

RangeIndex: 4600 entries, 0 to 4599

Data columns (total 18 columns):

#	Column	Non-Null Count	Dtype
---	--------	----------------	-------

```

--- -----
0  date      4600 non-null  object
1  price     4600 non-null  float64
2  bedrooms  4600 non-null  float64
3  bathrooms 4600 non-null  float64
4  sqft_living 4600 non-null  int64
5  sqft_lot   4600 non-null  int64
6  floors     4600 non-null  float64
7  waterfront 4600 non-null  int64
8  view       4600 non-null  int64
9  condition  4600 non-null  int64
10 sqft_above 4600 non-null  int64
11 sqft_basement 4600 non-null  int64
12 yr_built   4600 non-null  int64
13 yr_renovated 4600 non-null  int64
14 street     4600 non-null  object
15 city       4600 non-null  object
16 statezip   4600 non-null  object
17 country    4600 non-null  object
dtypes: float64(4), int64(9), object(5)
memory usage: 647.0+ KB

None

```

Missing Values:

```

date      0
price     0
bedrooms  0
bathrooms 0

```

sqft\_living 0  
sqft\_lot 0  
floors 0  
waterfront 0  
view 0  
condition 0  
sqft\_above 0  
sqft\_basement 0  
yr\_built 0  
yr\_renovated 0  
street 0  
city 0  
statezip 0  
country 0  
dtype: int64

#### Model Performance:

Mean Absolute Error (MAE): 207816.8294166553

Mean Squared Error (MSE): 986145473005.3936

Root Mean Squared Error (RMSE): 993048.575350367

R<sup>2</sup> Score: 0.03304504390209173

