Linear Regression on House Prices in London

Business problem: - Predict the price of a house with the House price Data
Data set: - [/kaggle/input/housing-prices-in-london]

First Import all Libraries and Read CSV

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns


import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

Read CSV File

```
data =pd.read_csv("/kaggle/input/housing-prices-in-london/London.csv")
```

Data Exploration and Cleaning

Check for null Values

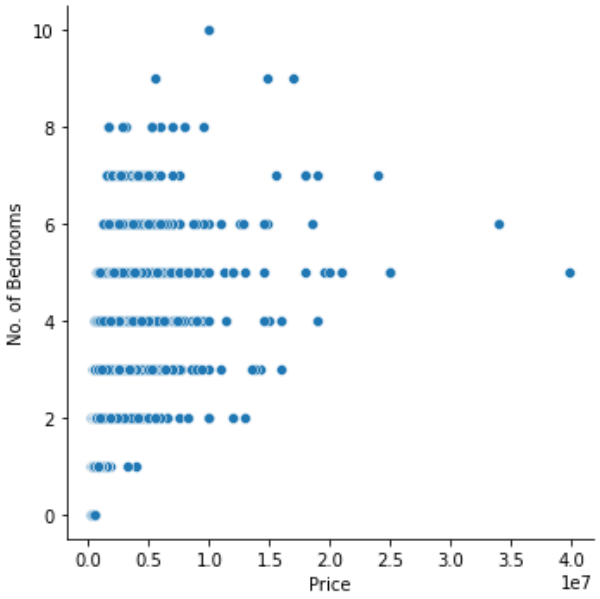
```
data.isnull().sum()
```

```
Unnamed: 0              0
Property Name           0
Price                   0
House Type              0
Area in sq ft           0
No. of Bedrooms         0
No. of Bathrooms        0
No. of Receptions       0
Location              962
City/County             0
Postal Code             0
dtype: int64
```

Exploration to understand the data.

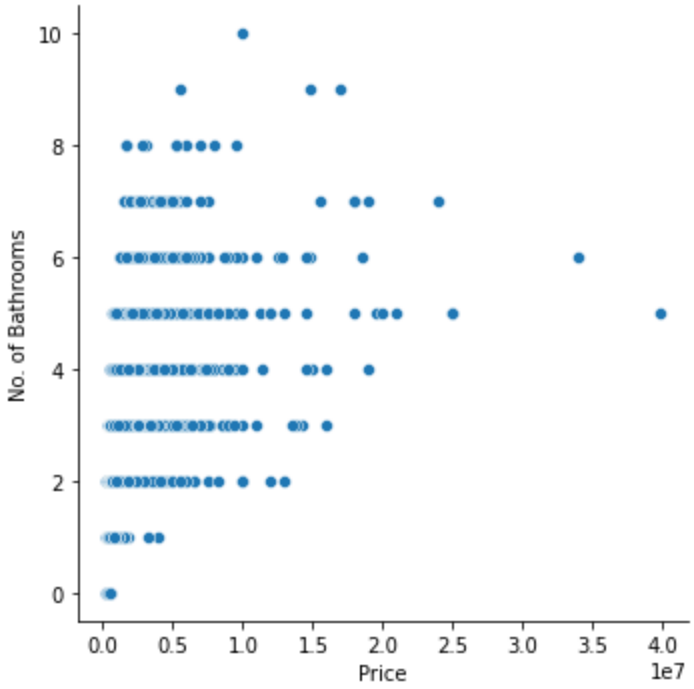
```
sns.relplot(x='Price', y = 'No. of Bedrooms', data =data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e746e2d0>
```



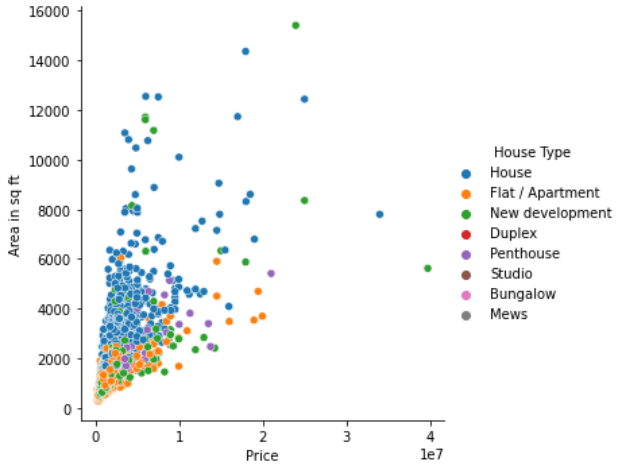
```
sns.relplot(x='Price', y = 'No. of Bathrooms', data =data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e4fee250>
```




```
sns.relplot(x='Price', y = 'Area in sq ft', hue ='House Type', data =data)
```

```
<seaborn.axisgrid.FacetGrid at 0x7f48e4f85210>
```



Through my data Exploration i can see that there is a trend and continuous Variable in the house Dataset. Price is the Independent variable

Data Modeling

```
data
```

	Unnamed: 0	Property Name	Price	House Type	Area in sq ft	No. of Bedrooms	No. of Bathrooms	No. of Receptions	Location	City/County	Postal Code
0	0	Queens Road	1675000	House	2716	5	5	5	Wimbledon	London	SW19 8NY
1	1	Seward Street	650000	Flat / Apartment	814	2	2	2	Clerkenwell	London	EC1V 3PA
2	2	Hotham Road	735000	Flat / Apartment	761	2	2	2	Putney	London	SW15 1QL
3	3	Festing Road	1765000	House	1986	4	4	4	Putney	London	SW15 1LP
4	4	Spencer Walk	675000	Flat / Apartment	700	2	2	2	Putney	London	SW15 1PL
...	...	...	...	...	...	...	...	...	...	...	...
3475	3475	One Lillie Square	3350000	New development	1410	3	3	3	NaN	Lillie Square	SW6 1UE
3476	3476	St. James's Street	5275000	Flat / Apartment	1749	3	3	3	St James's	London	SW1A 1JT
3477	3477	Ingram Avenue	5995000	House	4435	6	6	6	Hampstead Garden Suburb	London	NW11 6TG
3478	3478	Cork Street	6300000	New development	1506	3	3	3	Mayfair	London	W1S 3AR
3479	3479	Courtenay Avenue	8650000	House	5395	6	6	6	Highgate	London	N6 4LP

3480 rows × 11 columns

Import Libraries for Regression

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

Dropped the Unwanted data for Regression modeling

```
train = data.drop(['Price','Unnamed: 0','Property Name','House Type', 'Location','No. of Receptions','City/County', 'Postal Code'], axis=1)
test = data['Price']
```

```
X_train, X_test, y_train, y_test = train_test_split(train, test, test_size=0.3,random_state = 2)
```

```
regr = LinearRegression()
```

```
regr.fit(X_train, y_train)
```

```
LinearRegression()
```

```
pred = regr.predict(X_test)
```

```
regr.fit(X_train, y_train)
```

```
LinearRegression()
```

```
pred = regr.predict(X_test)
```

```
pred
```

```
array([1381943.88971662, 3265110.46262222, 1576568.53509975, ...,
       2480934.04257056, 2250281.95455301, 1278453.00685415])
```

```
regr.score(X_test, y_test)
```

```
0.37385544784332614
```

The score is 0.37 . I would assume weak. A good score with demographic data, we generally consider correlations above 0.75 to be relatively strong; correlations between 0.45 and 0.75 are moderate, and those below 0.45 are considered weak.