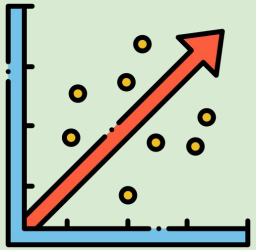
Housing Price Regression

Hokeun Go
BANA320

Data Source: Kaggle.com

Problem statement : Which Independent variable impacts the target variable, Price, the most.

Step 1: Data/Business Understanding

Type	Variables	
	Independent	**Dependent**
Regression Model (Linear) 	- Bathroom - Bedroom - Guestroom - Hot Water - Mainroad - AC - Preferred Area - Basement - Furnitured Status - Parking	- Price

Step 2. Data Preparation

Data Source: Kaggle.com

1. Select variables: (11) / dropped “area”, “stories”

```
housing_select = housing[['price', 'bedrooms', 'mainroad', 'bathrooms','basement', 'prefarea','furnishingstatus', 'parking', 'airconditioning', 'guestroom', 'hotwat
housing_select.head
```

2. No Missing Data Shown

```
housing_select.isna().sum()

price               0
area                0
bedrooms            0
mainroad            0
bathrooms           0
basement            0
prefarea            0
furnishingstatus    0
parking             0
airconditioning     0
guestroom           0
hotwaterheating     0
dtype: int64
```

3. Changing Data Types

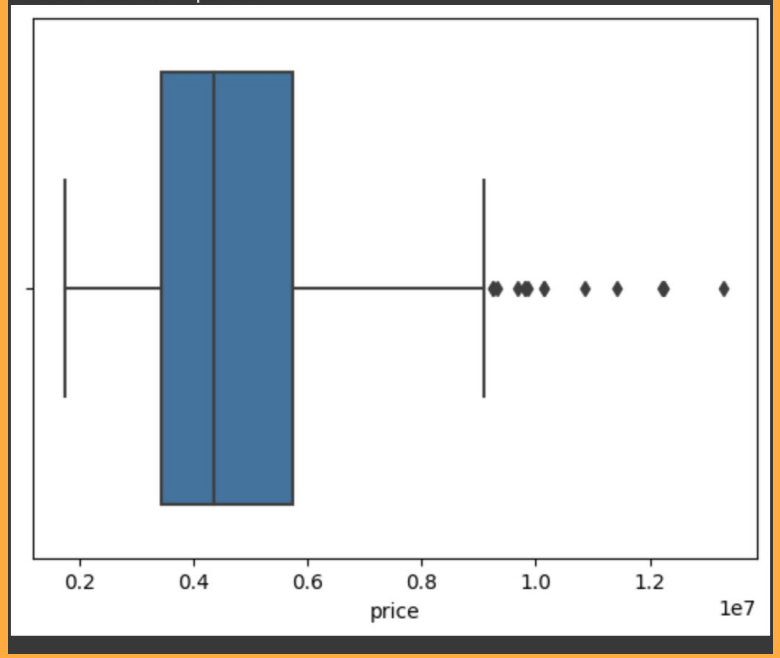
```
housing_select.dtypes

price               int64
area                int64
bedrooms            int64
mainroad           object
bathrooms           int64
basement           object
prefarea           object
furnishingstatus   object
parking             int64
airconditioning    object
guestroom          object
hotwaterheating    object
dtype: object
```

```
housing_select['mainroad'] = housing_select['mainroad'].astype('category')
housing_select['airconditioning'] = housing_select['airconditioning'].astype('category')
housing_select['guestroom'] = housing_select['guestroom'].astype('category')
housing_select['hotwaterheating'] = housing_select['hotwaterheating'].astype('category')
housing_select['prefarea'] = housing_select['prefarea'].astype('category')
housing_select['furnishingstatus'] = housing_select['furnishingstatus'].astype('category')
```

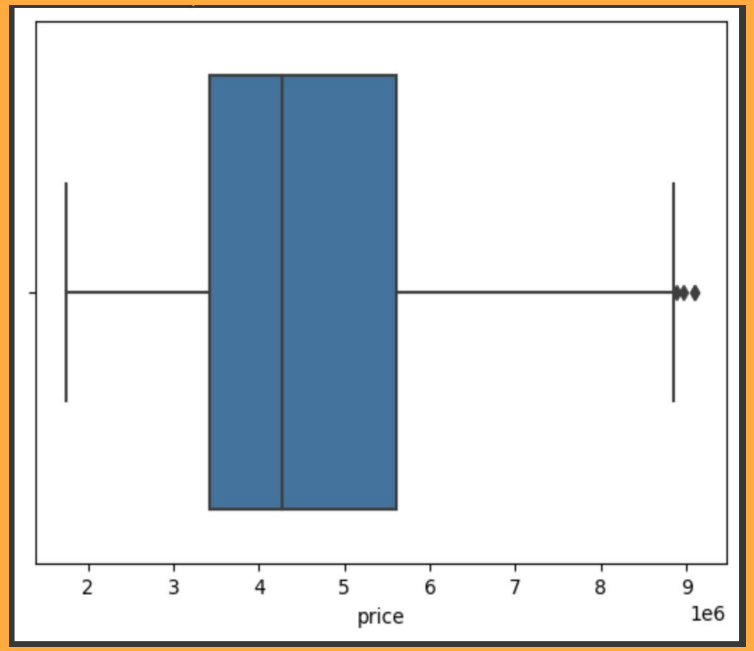
Step 2. Data Preparation: Elimination of Outliers

```
sns.boxplot(x=housing_select['price'])
```

```
<Axes: xlabel='price'>
```




```
q1 = housing_select['price'].quantile(.25)
q3 = housing_select['price'].quantile(.75)
iqr = q3 - q1
threshold = 1.5
lower_fence = q1 - threshold * iqr
upper_fence = q3 + threshold * iqr
housing_normal = housing_select[(housing_select['price'] > lower_fence)&(housing_select['price']<upper_fence)]
```



Step 2. Data Preparation: Make Dummies

	price	bedrooms	mainroad	bathrooms	basement	prefarea	furnishingstatus	parking	airconditioning	guestroom	hotwaterheating
0	13300000	4	yes	2	no	yes	furnished	2	yes	no	no
1	12250000	4	yes	4	no	no	furnished	3	yes	no	no
2	12250000	3	yes	2	yes	yes	semi-furnished	2	no	no	no
3	12215000	4	yes	2	yes	yes	furnished	3	yes	no	no
4	11410000	4	yes	1	yes	no	furnished	2	yes	yes	no

```
housing_select = pd.get_dummies(housing_select, columns=mainrd, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=ac, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=gst_room, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=ht_wtr, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=prefarea, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=frniture, drop_first=True)
housing_select = pd.get_dummies(housing_select, columns=base, drop_first=True)

housing_select.head()
```

	price	bedrooms	bathrooms	parking	mainroad_yes	airconditioning_yes	guestroom_yes	hotwaterheating_yes	prefarea_yes	furnishingstatus_semi-furnished	f
0	13300000	4	2	2	1	1	0	0	1	0	
1	12250000	4	4	3	1	1	0	0	0	0	
2	12250000	3	2	2	1	0	0	0	1	1	
3	12215000	4	2	3	1	1	0	0	1	0	
4	11410000	4	1	2	1	1	1	0	0	0	

Step 2. Data Preparation: Standardizing Data

```
numeric_var = ['price']
scaler = StandardScaler()
housing_select[numeric_var] = scaler.fit_transform(housing_select[numeric_var])

housing_select.head()
```

	price	bedrooms	bathrooms
0	13300000	4	2
1	12250000	4	4
2	12250000	3	2
3	12215000	4	2
4	11410000	4	1

	price	bedrooms	bathrooms
0	4.566365	4	2
1	4.004484	4	4
2	4.004484	3	2
3	3.985755	4	2
4	3.554979	4	1

Train_Test_Split to get the R squared

```
X4 = housing_select[['bedrooms', 'bathrooms', 'parking', 'mainroad_yes','airconditioning_yes', 'guestroo
Y4 = housing_select['price']

X_train, X_test, Y_train, Y_test = train_test_split(X4, Y4, test_size=0.2, random_state=42)

model = LinearRegression()

model.fit(X_train, Y_train)

y_pred = model.predict(X_test)
```

```
r_squared = r2_score(Y_test, y_pred)
print(f"R-squared: {r_squared}")

n = X_test.shape[0]
p = X_test.shape[1]
adjusted_r_squared = 1 - (1 - r_squared) * (n - 1) / (n - p - 1)
print(f"Adjusted R-squared: {adjusted_r_squared}")

mse = mean_squared_error(Y_test, y_pred)
print(f"Mean squared error: {mse}")
```

Step 3. Model Comparison (R^2 , MSE)

Model 1

```
X1 = housing_select[['bedrooms', 'bathrooms']]
Y1 = housing_select['price']
```

```
R-squared: 0.2605701076698278
Adjusted R-squared: 0.2466186002673717
Mean squared error: 1.070265332447618
```

Model 2

```
X2 = housing_select[['bedrooms', 'bathrooms', 'parking']]
Y2 = housing_select['price']
```

```
R-squared: 0.35927577259623034
Adjusted R-squared: 0.3409693660989799
Mean squared error: 0.9273968166049447
```

Model 3

```
X3 = housing_select[['bedrooms', 'bathrooms', 'parking', 'mainroad_yes']]
Y3 = housing_select['price']
```

```
R-squared: 0.4048614103660153
Adjusted R-squared: 0.38197146461086195
Mean squared error: 0.8614152701884653
```

Model 4

```
X4 = housing_select[['bedrooms', 'bathrooms', 'parking', 'mainroad_yes','airconditioning_yes', 'guestroom_yes', 'hotwaterheating_yes','prefarea_yes', 'furnishingstatus_semi-furnished','furnishing
Y4 = housing_select['price']
```

```
R-squared: 0.5708854770213114
Adjusted R-squared: 0.5222230053433158
Mean squared error: 0.6211087790170299
```

Step 4.Modeling

```
intercept = model.intercept_
print(intercept)
```

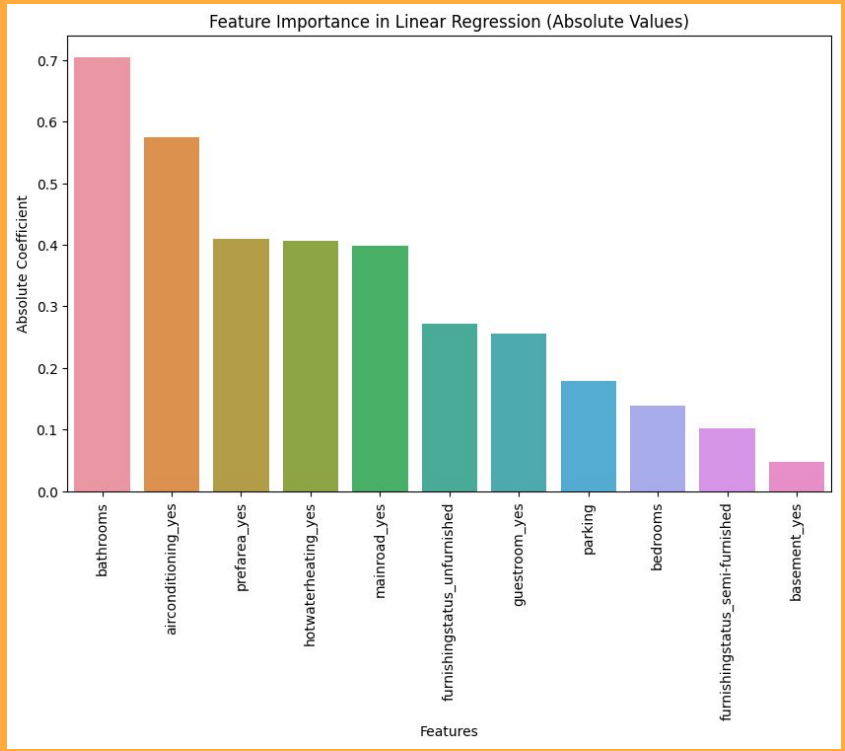
```
-2.0281951076890494
```

	Feature	Coefficient
0	bedrooms	0.139692
1	bathrooms	0.704481
2	parking	0.178460
3	mainroad_yes	0.399087
4	airconditioning_yes	0.574134
5	guestroom_yes	0.255771
6	hotwaterheating_yes	0.406683
7	prefarea_yes	0.409315
8	furnishingstatus_semi-furnished	-0.102068
9	furnishingstatus_unfurnished	-0.272431
10	basement_yes	0.047975

	price	bedrooms	bathrooms	parking	mainroad_yes	airconditioning_yes	guestroom_yes	hotwaterheating_yes	prefarea_yes	furnishingstatus_semi-furnished	furnishingstatus_unfurnished
0	13300000	4	2	2	1	1	0	0	1	0	0

7,900,000 = -2.0281951076890494 + 0.14*Bed(4) + 0.70 *Bath(2) + 0.18*Parking(2) + 0.4*mainroad(1) + 0.58*AC(1) + 0.26*Guestroom(0) + 0.41*Hotwater(0) + 0.41*preferArea(1) -0.10*Semi-furnitured(0) -0.28*Unfurnitured(0) + 0.05*basement(0)

Step 5. Insights



1. Bathrooms impacts the most on housing price

Options:

- Even though model 4 has the best R^2 values compared to other models, the value is still lower than 0.7, this indicates linear regression might not be the best fit for housing price data.
- From the data and model, basement has the least impact.

Share the Bathroom!

Thank You