Predicting Housing Prices in London

Using the Linear Regression Model

Group 3

Meet Anna!

- A young professional
- Just relocated to London
- Needs a spacious, one-bedroom home that fits her budget and commute

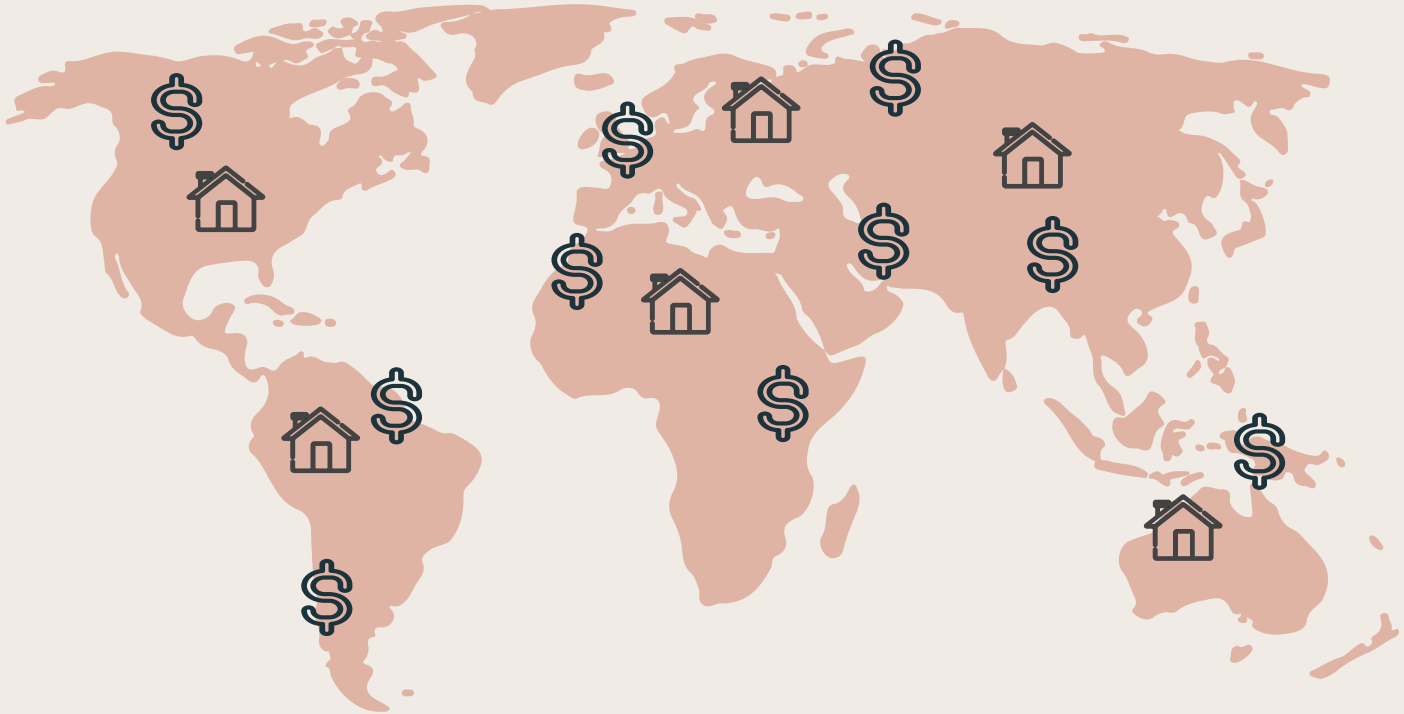
Her criteria

700 sq ft	01 Bedroom	Type: flat	Kensington	Budget: £1M

Research Question

To what extent can we predict the price of an apartment in London based on its characteristics?

How can Anna choose the best apartment for her needs based on these predictors?



“Linear regression is a method that explains how one or more independent variables relate to a dependent variable by estimating the best-fitting straight line through the data.”

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

Dependent Variable


Sales


Costumer behavior


Prices


Brand Metrics

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

Constant term

Value of Y when all independent variables equal zero

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

The "effect"

Tells you how much Y will change when X rises by 1

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

The Independent variable

Predicts the effect on Y, for example:

Advertising spending

Brand exposure

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

Extra control variables

Control for confounding factors

Improve explanatory power R^2

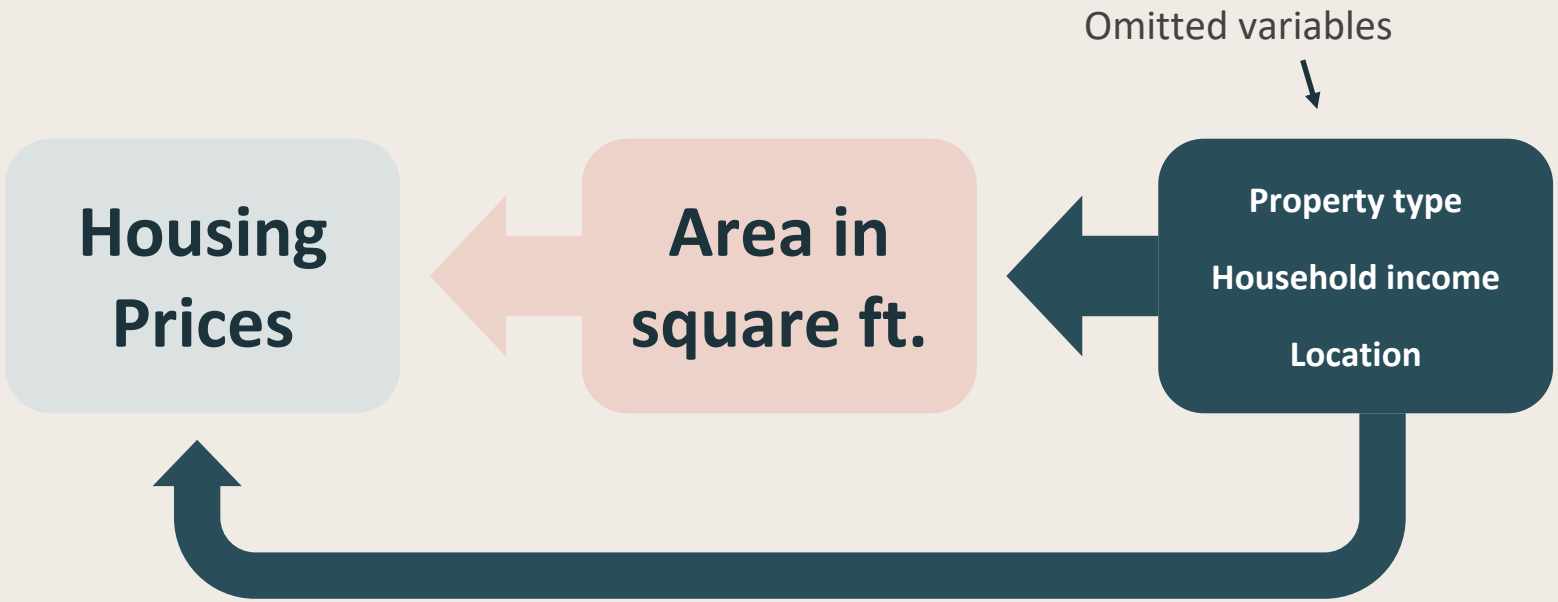
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_i X_i + \varepsilon_i$$

Error term

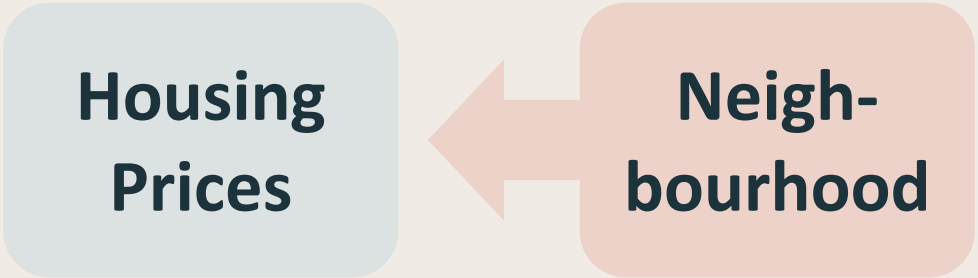
The difference between the actual outcome and predicted outcome

Assumption 1: Exogeneity

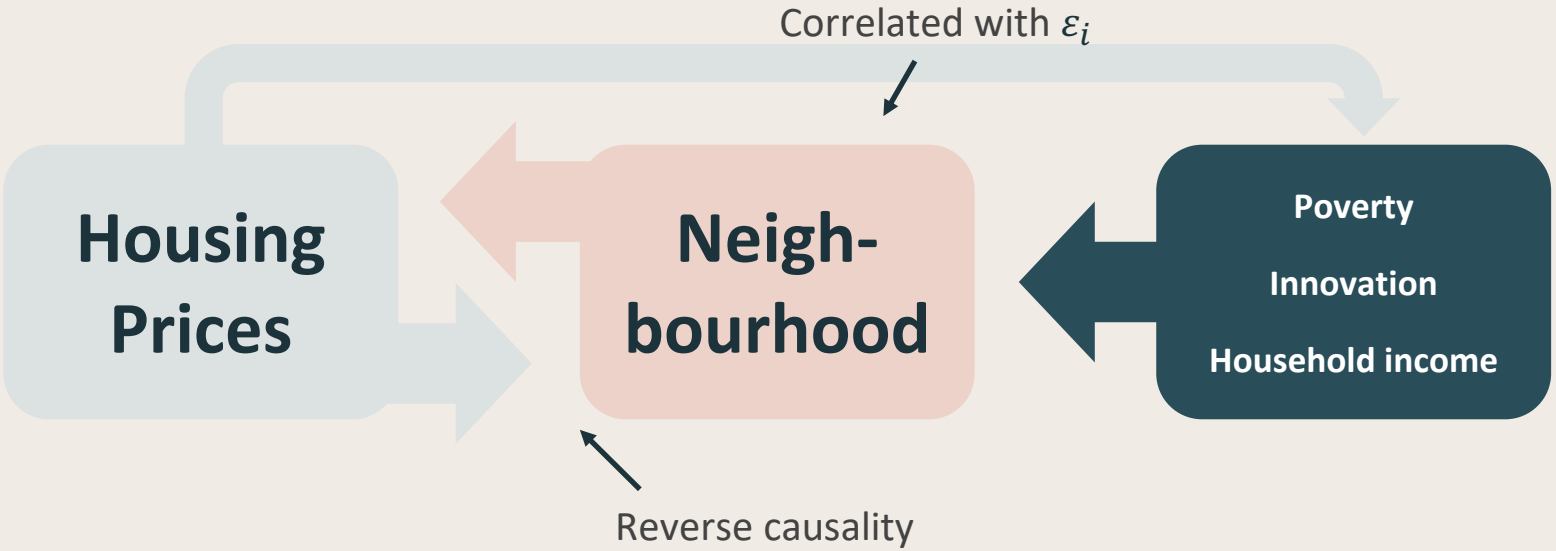
Assumption 1: Exogeneity



Assumption 1: Exogeneity



Assumption 1: Exogeneity



Other Assumptions

2. Linearity

Relationship between Y and X is linear

E.g.
Exponential price increases in property types -> nonlinearity

3. No perfect multicollinearity

Independent variables should not be perfectly linear to each other

E.g.
If you add all neighbourhood dummy's to the model, their sum is always 1

4. Homoscedasticity

ε_i has the same variance across all values of X

5. Normality

Errors are normally distributed

Literature Review

Linear Regression in Housing Price Analysis

Linear regression models the relationships between variables

Widely used in housing economics

Helps quantify effects of size, type, and location on price

Literature Review

What other researchers found?

Ng (2015)

Finding: Square footage and location are strong predictors.

Relevance to our study: Validates our focus on **Area** and **Location** as primary variables.

Saraf et al. (2021)

Finding: Controlling structural variables (bedrooms, bathrooms) improves accuracy.

Relevance to our study: Motivates inclusion of **No. of Bedrooms**, **No. of Bathrooms**, and **House Type** as controls.

Yin (2023)

Finding: Location is the dominant factor in London housing prices.

Relevance to our study: Motivates inclusion of **Location** as control.

Literature Review

What makes our study unique

Two-step approach

Demonstrates how adding relevant variables strengthens accuracy

Applies regression in a beginner-friendly, interpretable way

Data Description

Numeric variables

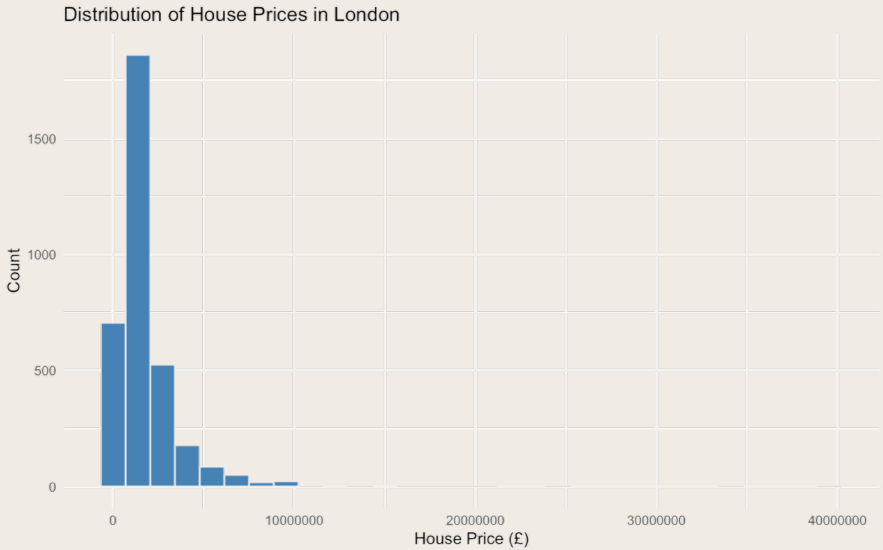
Dataset: House Price in London
Souce: Kaggle
No. of observations: 3840
Our key variables:

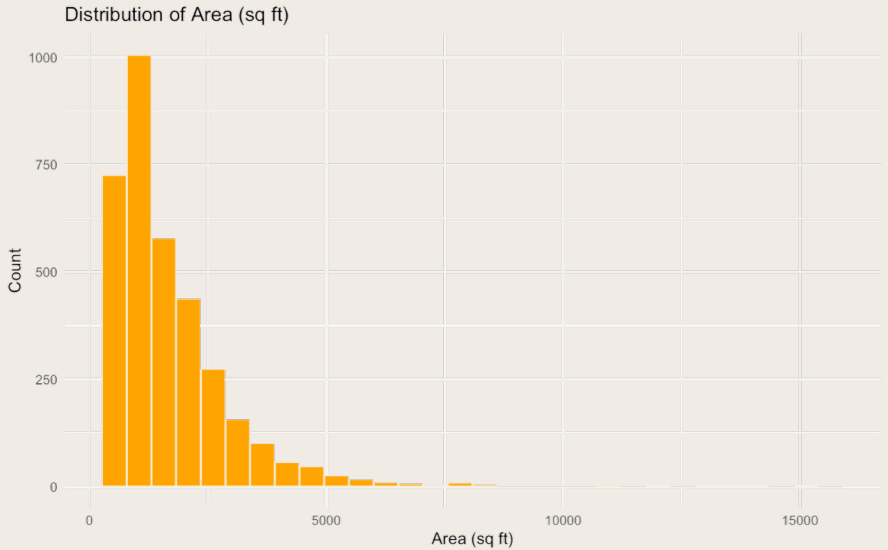
- Price (GBP)
- Area (sq ft)
- No. of Bedrooms
- No. of Bathrooms
- Location

	Variables	Mean	Min	Max	SD
1	Price	1,864,173	180,000	39,750,000	2,267,282.958
2	Area in sq ft	1,713	274	1,5405	1,364.259
3	No. of bedrooms	3.104	0.000	10.000	1.518
4	No. of bathrooms	3.104	0.000	10.000	1.518

Data Description

Numeric variables

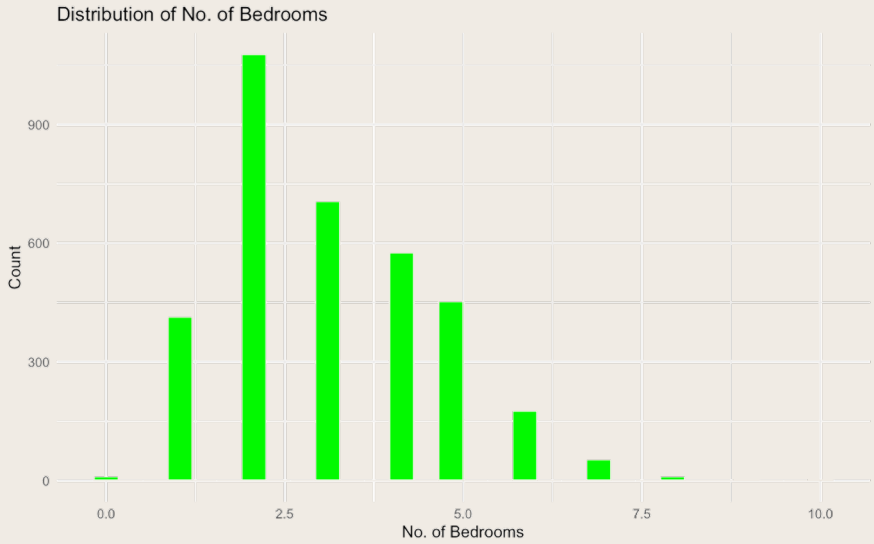


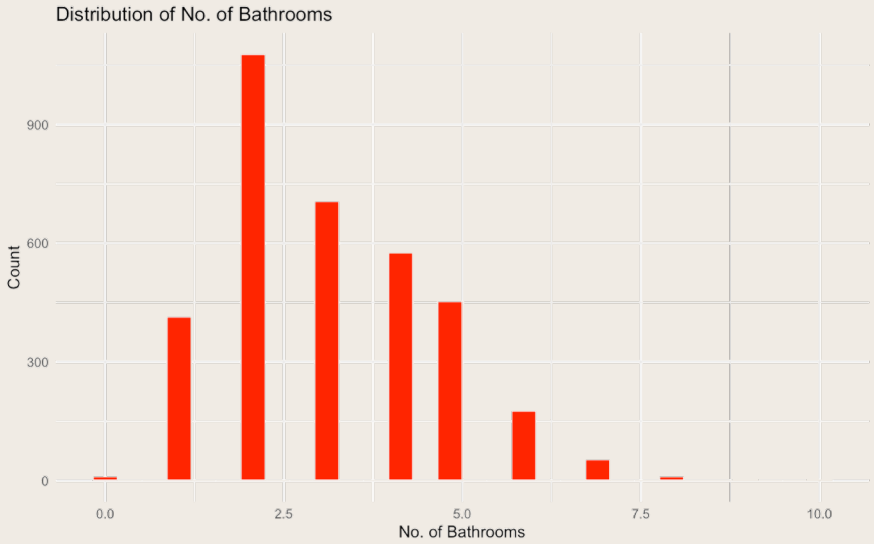


Distribution of numeric variables

Data Description

Numeric variables

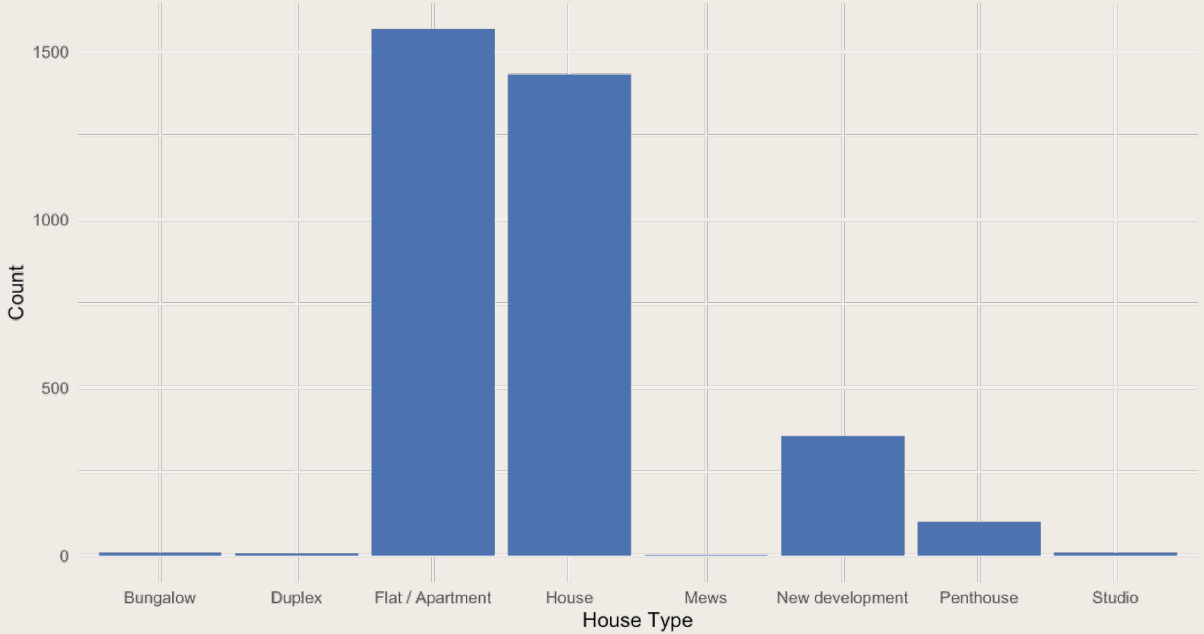




Distribution of numeric variables

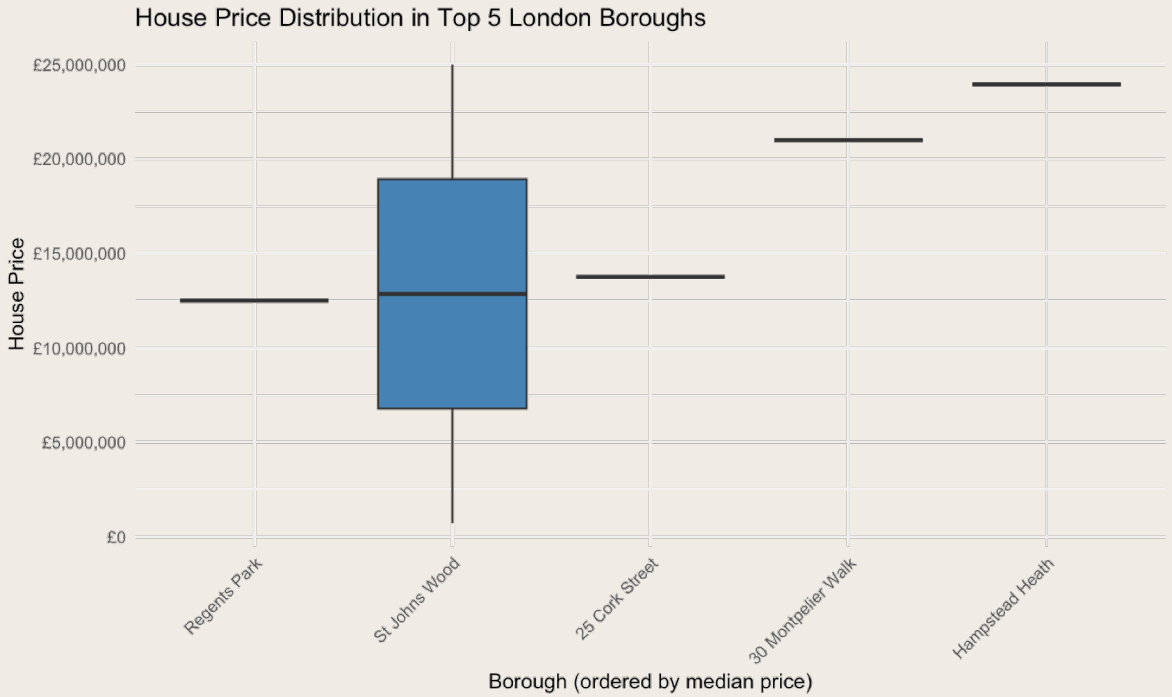
Data Description

Properties by House Type



Data Description

Price distribution in Top 5 Boroughs



Single Regression

```
m_simple <- lm(Price ~ Area.in.sq.ft, data = df)
```

	House price	P-value	R^2
Area in sq ft	1109.68 (20.98)	<2e-16***	0.446
Constant term	-36674.05 (45936.19)	0.425	
Observations	3480		

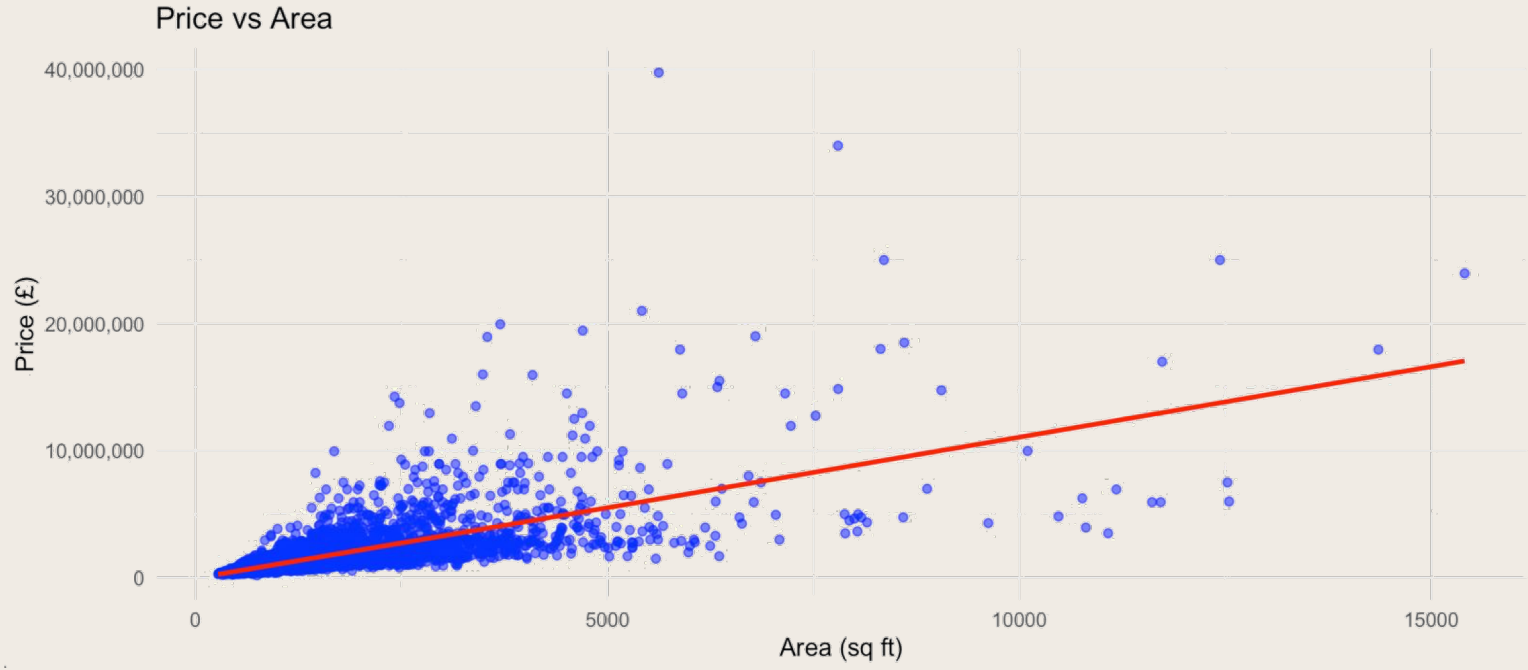
*p- value<0.10, **p-value<0.05, ***p-value<0.01. Standard errors are in brackets.*

Single Regression

$$House\ Price_i = \beta_0 + \beta_1 \text{Area in sq ft} + \varepsilon_i$$

$$House\ Price_i = -36674.05 + 1109.68 * Area\ in\ sq\ ft + \varepsilon_i$$

Single Regression



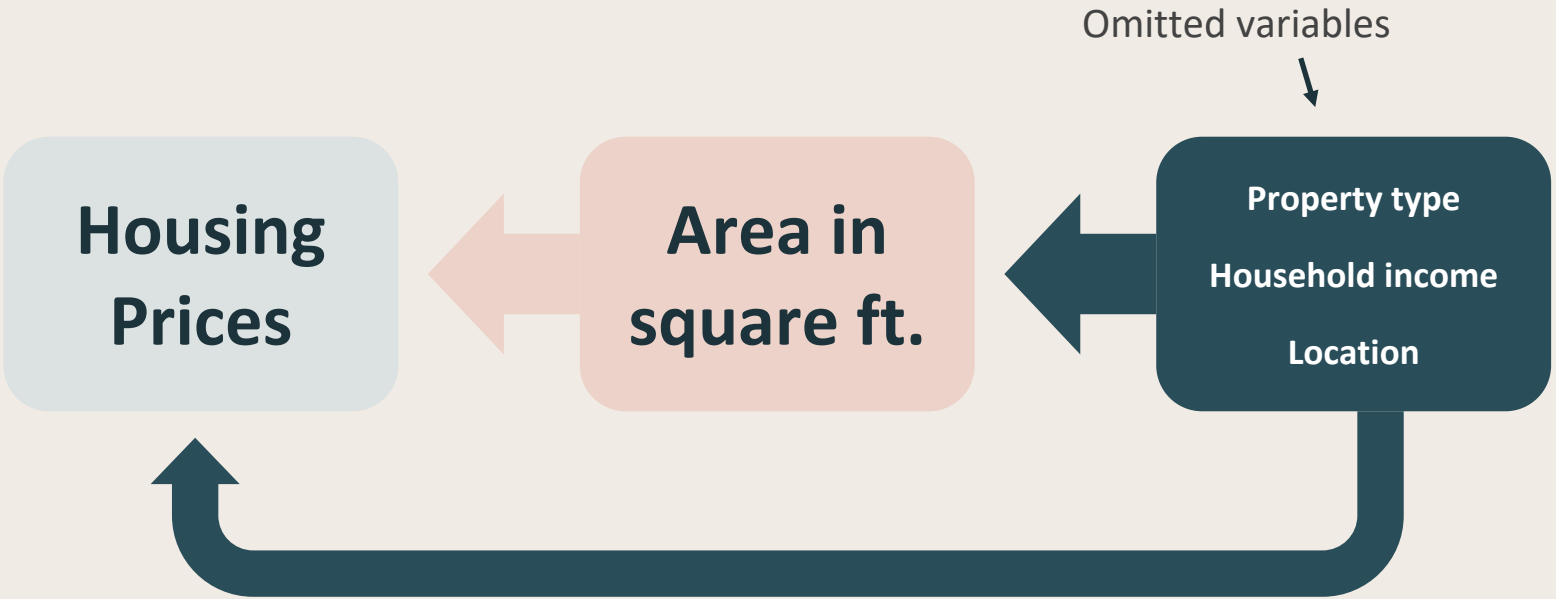
What it means

1 extra square foot → **£1110**

1000 difference in sq ft → **£1.1m difference in price**

! $R^2 = 0.446$

However, does Assumption 1 hold?



Multiple Regression

```
m_multi  <- lm(Price ~ Area.in.sq.ft + House.Type + No..of.Bedrooms +
                 No..of.Bathrooms + Location, data = df)
summary(m_simple); summary(m_multi)
```

	House price	P-value	R2
Area in sq ft	1573 (35.09)	<2e-16***	0.732
Housetype			
Duplex	-894800 (787800)	0.256	
Flat/Apartment	-610300 (517900)	0.239	
House	-743600 (515800)	0.149	
Mews	-1241000 (1135000)	0.274	
New development	-148200 (525700)	0.778	
Penthouse	706700 (542600)	0.896	
Studio	-1011000 (683800)	0.108	
No. of bedrooms	-269300 (3462)	1.02e-14***	

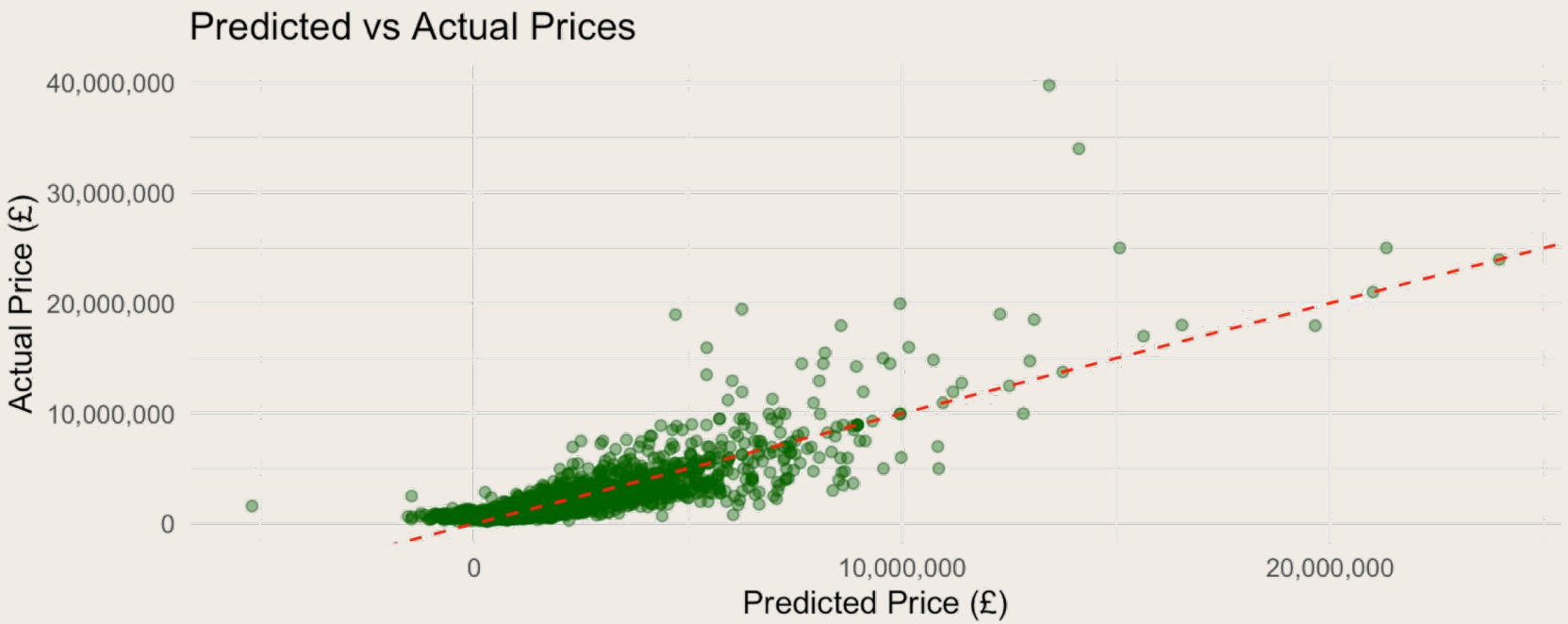
	Housing Prices	P-value
No. Of bathrooms	NA	NA
Location		
95 dummy's	High variance	<2e-16***
Constant term	817100 (522100)	0.425
Observations	3472	

*p- value<0.10, **p-value<0.05, ***p-value<0.01. Standard errors are in brackets.*

Multiple Regression

$$\begin{aligned} &\boldsymbol{House\ price_i} \\ &= \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\textbf{Area in sq ft} + \boldsymbol{\beta_2}\textbf{House type} + \boldsymbol{\beta_3}\textbf{No of bedrooms} \\ &+ \boldsymbol{\beta_4}\textbf{No of bathrooms} + \boldsymbol{\beta_5}\textbf{Location} + \boldsymbol{\varepsilon_i} \end{aligned}$$

Multiple Regression



What it means

1 extra square foot → **£1573**

1 extra bedroom → **£270000 decline in price**

Location dominates inequality

Values are in milion £.

Area in sq ft explains within-neighborhood differences

Location explains between-neighborhood inequality

Multicollinearity test

```
> vif(m_full)
                     GVIF  Df GVIF^(1/(2*Df))
Area.in.sq.ft    4.189784   1        2.046896
No..of.Bedrooms  5.179911   1        2.275942
House.Type      12.819211   7        1.199867
Location        13.880276 571        1.002306
```

All VIFs are below 10

Area and bedrooms are moderately correlated ~ 2

House type and Location look big at first due to dummies

Diagnostics

1. Exogeneity

Cannot be tested, there might be omitted variable bias

2. Linearity

Area shows a strong positive trend with price

3. No perfect multicollinearity

Bedrooms and area overlap, making bedrooms less reliable as a predictor, but the model is still considered reliable.

4. Homoscedasticity

Higher-priced properties show more variation (luxury outliers)

5. Normality

Mostly normal, but extreme values (e.g., Knightsbridge luxury flats) distort it

Implications for Anna

1

Most reliable predictors: Area and location

2

Bedrooms alone are misleading (correlated with area)

3

Predictions are best for typical mid-range apartments, less so for luxury extremes

4

Location effects: (Chelsea, Knightsbridge, Marylebone, etc.) are very strong

Conclusion for Anna

700 sq ft	01 Bedroom	Type: flat	Kensington	Budget: £1M

$$House\ price_i = 817100 + 1573 * 700sq\ ft - 269500 * 1\ bedroom - 610600 * flat + 2405000 * Kensington$$

Predicted price:

£3,443,100

Conclusions & Recommendations

Area and **location** are the strongest drivers of price.

Multiple regression is **essential for accurate predictions**.

Anna can use this model to focus on apartments with the best combination of size and location, avoiding overpaying for features that don't significantly increase value.

- Prioritize Apartment Size (Area in sq ft)
- Focus on High-Value Locations
- Balance Bedrooms and Bathrooms with Area
- Consider Overall Value, Not Just Price

References

- Ng, A. (2015). *Machine Learning for a London Housing Price Prediction Mobile Application*. Imperial College London.
- Saraf, S., et al. (2021). *House Price Prediction Using Linear Regression*. International Journal for Research in Applied Science & Engineering Technology (IJRASET).
- Yin, M. (2023). *Model for Predicting London House Prices*. ResearchGate.

Appendix

Literature review

Study	Key predictors & Findings	Implications for our study
Ng (2015)	Square footage and location were the most influential predictors of London housing prices; linear regression proved effective for real-world appraisal.	Validates our focus on **Area in sq ft** and **Location** as primary variables.
Saraf et al. (2021)	Controlling structural features—bedrooms and bathrooms—significantly improved predictive accuracy in multiple regression models.	Motivates inclusion of **No. of Bedrooms**, **No. of Bathrooms**, and **House Type** as controls.
Yin (2023)	Using the same London dataset, location emerged as the dominant factor in price variation across boroughs.	Motivates inclusion of **Location** as control.

Reference

- *Ng, A. (2015). Machine Learning for a London Housing Price Prediction Mobile Application. Imperial College London.*
- *Saraf, S., et al. (2021). House Price Prediction Using Linear Regression. International Journal for Research in Applied Science & Engineering Technology (IJRASET).*
- *Yin, M. (2023). Model for Predicting London House Prices. ResearchGate.*