Identifying drivers of house prices in England and Wales

Interview Task

Bruno Chereque

Prepared for:

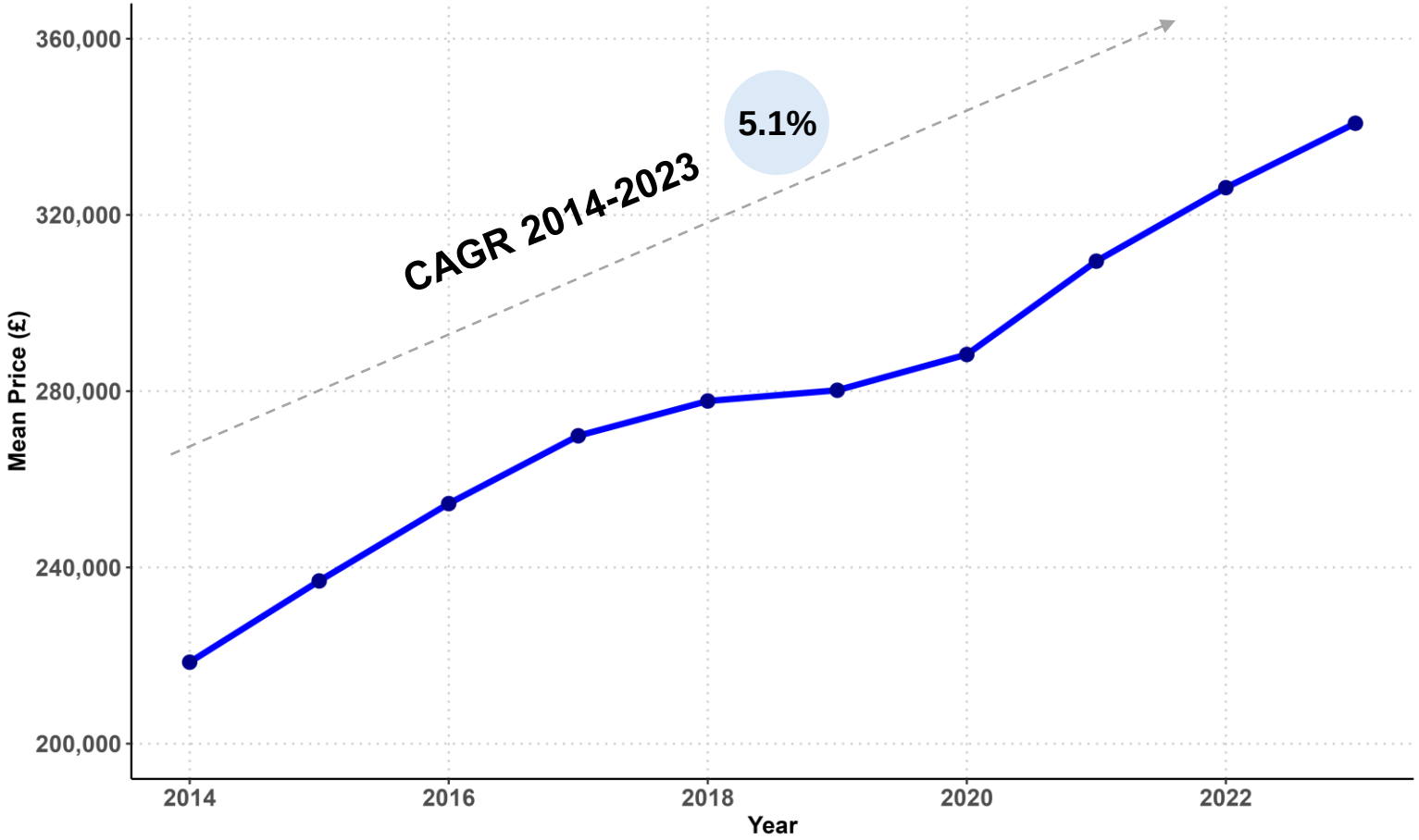
frontier economics

19 December 2024

Contents

Context: Residential property prices in England and Wales have consistently grown in the last decade.

Annual Mean Price for Residential Property in England and Wales (2014-2023):
(Increasing trend over the years)



Source: ONS (2024). Own elaboration.

+56%

percentage change in mean property prices between 2014 and 2023

Objective of the analysis

- Explore the drivers of residential property prices in England and Wales.

Explanatory Variables: Dataset with 180 variables that will help us identify the drivers of property prices.

Sources and Variables Selected (Data Consistent Across Sources: 2014–2017)

Valuation Office Agency

Council tax data

- Property counts across council tax bands (A – H)

Department for Transport

Journey times data

- Travel time in minutes to the nearest:
  - Employment Centre
  - Secondary School
  - Hospital

Office for National Statistics

School characteristics

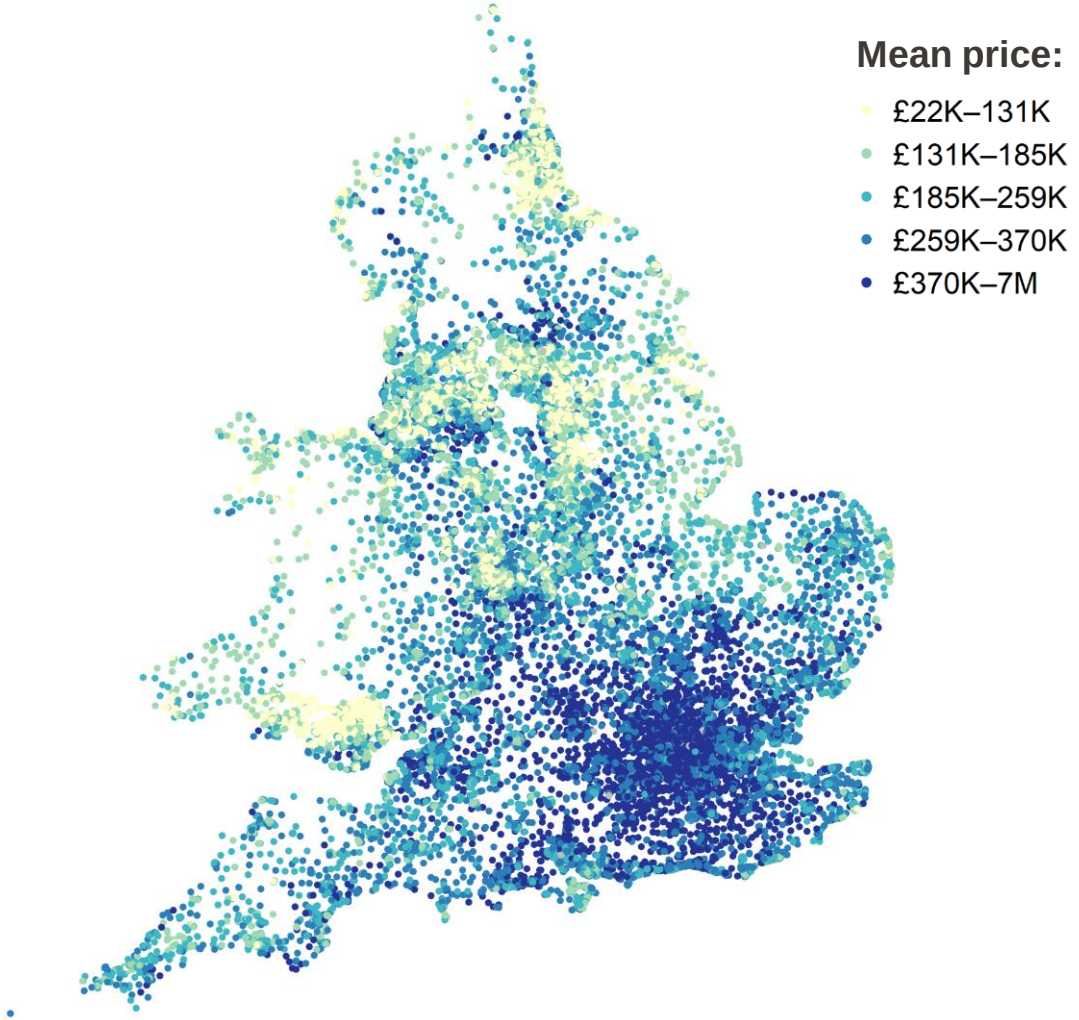
- Schools by type:
  - Primary, secondary
  - Mixed gender
  - Religious
  - Among other aggregated metrics

Additional for refined analysis (not prioritised due to time constraints):

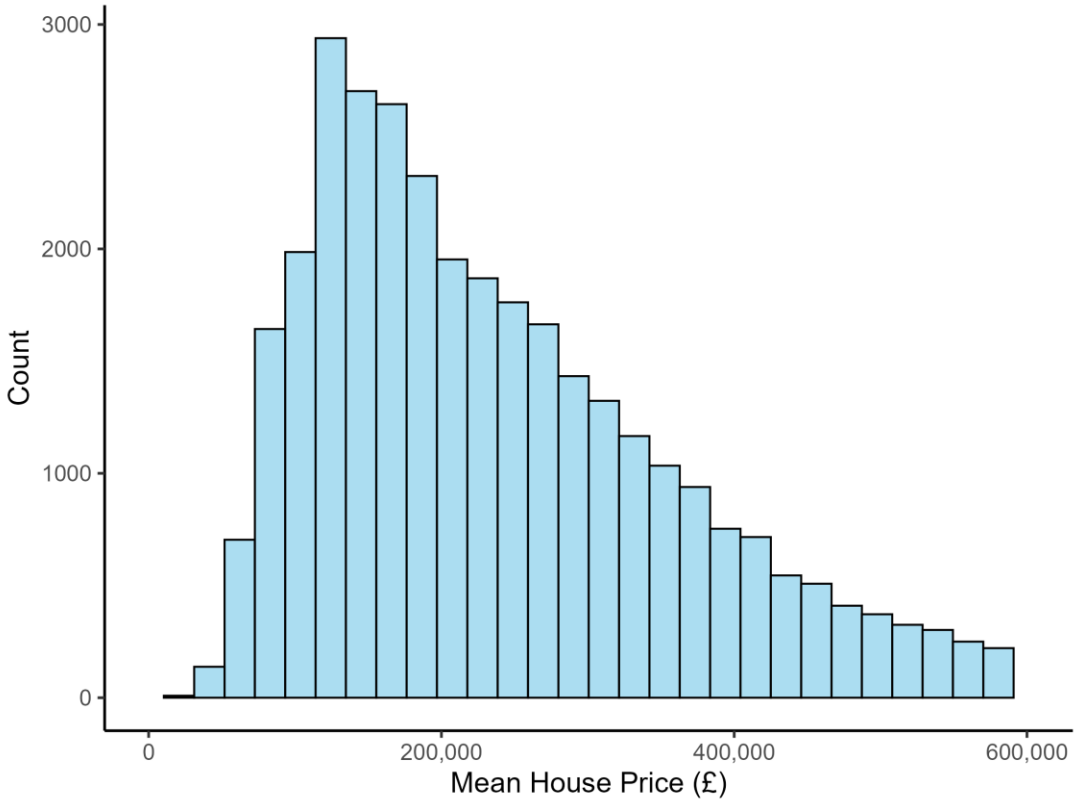
- Property type
- Build period
- Travel times to GPs
- Travel times to food stores
- Primary or secondary school performance
- Additionally, household income proxy variables like the index of multiple deprivation at the LSOA level.

Regional Disparities: Affordability challenges concentrate in London and the South East. The distribution skews toward lower-priced housing but reveals high-value properties driving disparities.

Mapping Housing Affordability in England and Wales (2017)



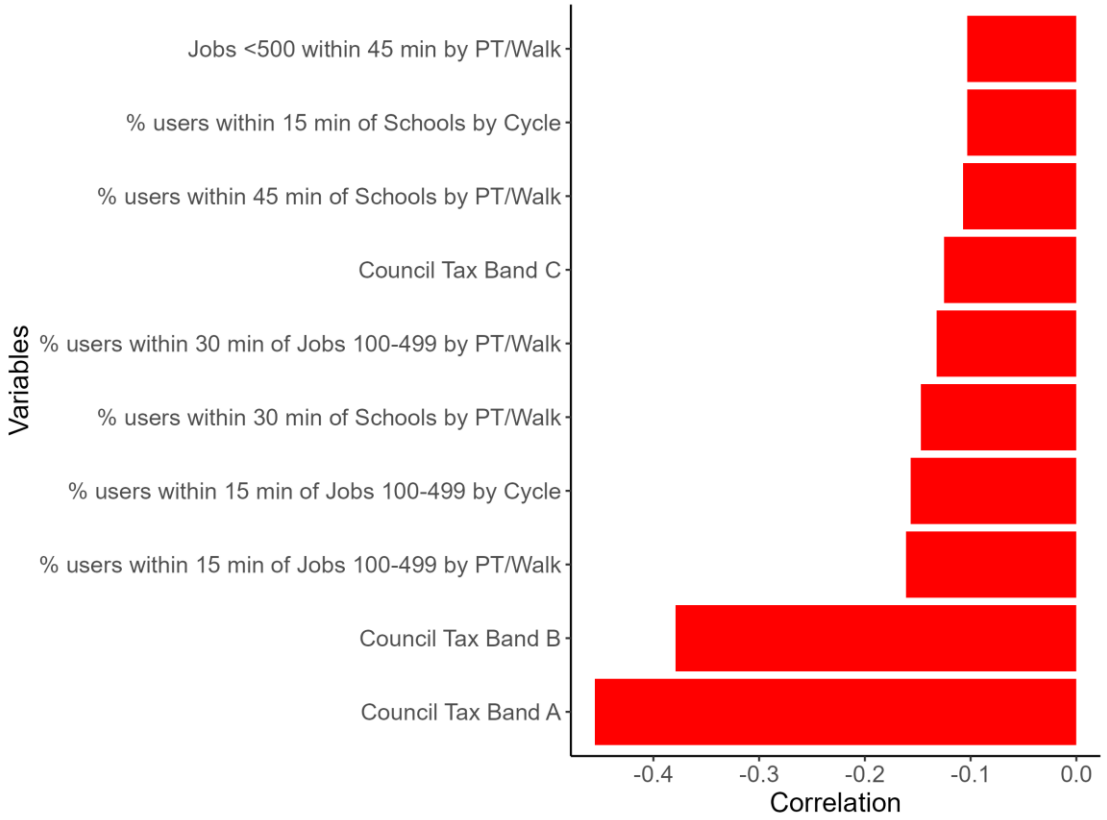
Distribution of Mean House Prices LSOAs (2017)



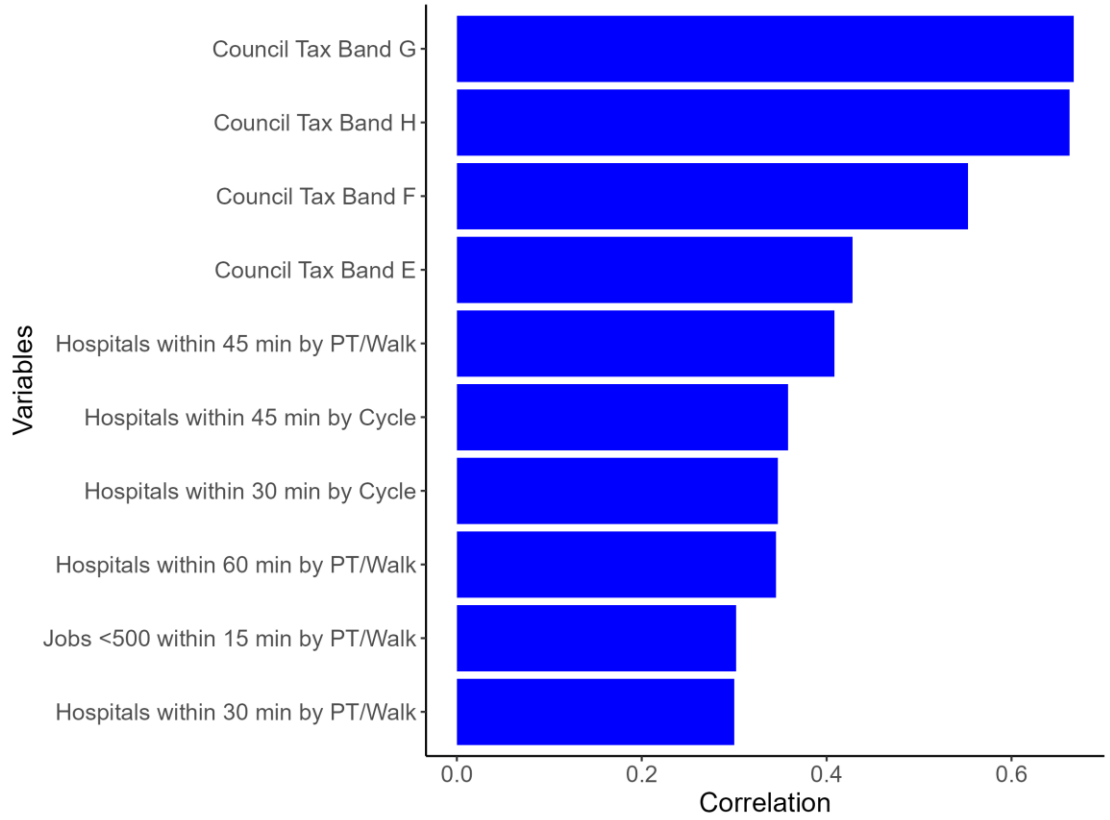
Source: ONS (2024). Own elaboration.

Visualising Correlations: This analysis shows variables most positively and negatively correlated with property prices. These insights highlight potential drivers but should not be interpreted as causal.

Top 10 most negatively correlated variables



Top 10 most positively correlated variables

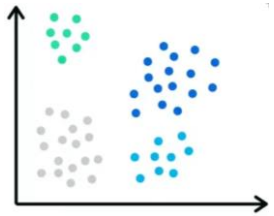


- The observed correlations reflect statistical associations but do not imply causal relationships.
- External factors or omitted variables could drive both the explanatory variables and house prices.

Source: Valuation Office Agency (2024), Department of Transport (2024), ONS (2024). Own elaboration.

Unsupervised Learning: Utilising data science techniques like PCA and clustering to group and characterise observations based on their publicly available features at the LSOA level.

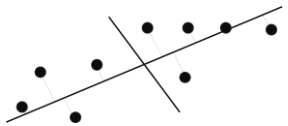
Key methodologies to characterise properties at the LSOA level[1/]



1

Clustering analysis (K means)

Creates k groups based on similarity, where observations within the same cluster are more alike than those in other clusters.



2

Principal Component Analysis (PCA)

PCA is a tool that reduces the number of features (dimensions) in the data while preserving the most important information.

Raw Data

- Large number of correlated variables
- Make the analysis noisy and challenging

PCA

- Simplify data reducing dimensionality
- Kept with components that explain most of the variation

K means

- Implement clustering algorithm on PCAs
- Assigns LSOAs to groups based on similarities

Results

- Presentation of final segmentation highlighting key group differences

1/ These methodologies require the standardisation of variables.

Methodology Considerations: Handling missing data and key model parameters.

Feature-observation ratio:

PCA and clustering require complete observations (no NAs)

139k
observations including rows with NAs

22k
observations after removing rows with NAs

161
numeric variables ready for the analysis

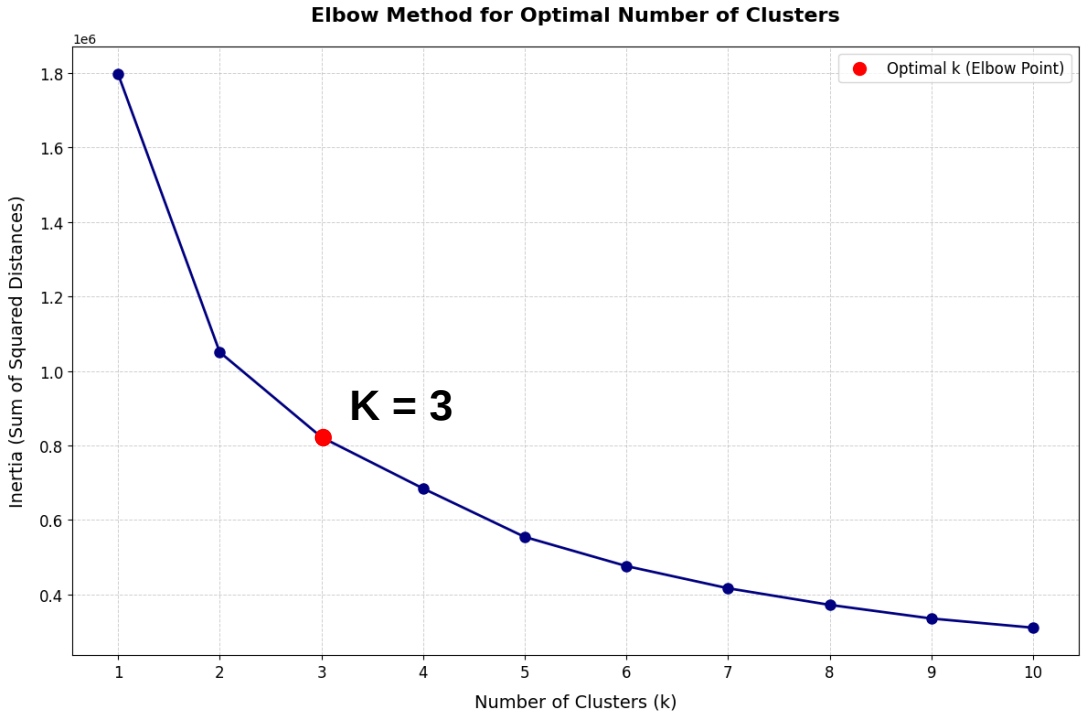
Model parameters

Principal Component Analysis:

5
principal components chosen for the analysis

50%
of the cumulative variance in the data explained

K-means[1/]:



3
clusters were chosen for the analysis

1/ A Silhouette Score of 0.4 indicates moderate clustering performance: clusters are somewhat distinct but not highly separated.

Clustering analysis: Identifying three property segments in England and Wales via cluster analysis at the LSOA level.

		Remote and affordable areas	Accessible mid-tier areas	Premium areas
	Cluster size	(39% of the total obs.)	(46% of the total obs.)	(15% of the total obs.)
	Mean price	**£282k**	**£285k**	**£330k**
Council Tax	Band G % properties (LSOA)	8%	4%	11%
	Band F % properties (LSOA)	9%	5%	13%
	Band B % properties (LSOA)	17%	19%	13%
	Band A % properties (LSOA)	14%	23%	9%
Travel Times	**Minutes by public transport to the nearest:**			
	Employment Centre (+5k jobs)	43	23	79
	Secondary School	22	14	46
	Hospital	50	29	80
School	Avg primary schools (LSOA)	**1.3**	**1.2**	**1.6**
	Avg maintained schools (LSOA)	**0.9**	**0.8**	**1.2**
	Avg religious schools (LSOA)	**1.3**	**1.2**	**1.5**

Source: Valuation Office Agency, Department of Transport, ONS. Own elaboration.

Causal Inference: We could implement a panel regression model, but a more localised and specific case study would be necessary to develop a robust identification strategy for assessing causality.

Proposed methodology: Fixed Effects Panel Regression

$$PropertyPrices_{it} = X_{it}'\beta + \alpha_i + \theta_t + \varepsilon_{it}$$

- X_{it}: Vector including key explanatory variables (e.g., council tax bands, school quality, accessibility metrics).
- α_i: LSOA fixed effects to control for time-invariant, unobservable characteristics.
- θ_t : Time fixed effects to capture year-specific shocks.

Challenges:

- Time-varying unobserved factors (e.g., changes in local policies and economic shocks**)** could bias the results.
- Regression coefficients reflect correlation, not causation, without addressing confounding or endogeneity.
- We would need to exploit localised interventions, shocks or exogenous variation to apply robust identification strategies such as Difference-in-Differences or Instrumental Variables.

Predictive Modelling: Identifying key drivers for property prices and quantifying model accuracy.

Outline for predictive modelling

Data split

- Split the data into training and testing sets.
- Ensure consistent normalization for numerical features.

Model training

- Train OLS model for comparison as well as:
- Lasso Regression
- Decision Trees
- Random Forest

Validation

- Evaluate models on the testing set using unseen data.
- Perform repeated cross-validation to ensure results are not dataset-specific

Metric evaluation

- Compare models using:
  - Root Mean Square Error (RMSE)
  - Number of features used
  - Feature importance score

Conclusions:

- **Exploratory Data Analysis (EDA):** Identified key drivers such as council tax bands, school quality, and travel times, showing strong correlations with house prices.
- **Clustering Results:** Segments Highlighted regional disparities in affordability and accessibility.
  - Remote and Affordable Areas
  - Accessible Mid-Tier Areas
  - Premium Areas
- **Challenges:** Feature vs observation trade-off. Handling missing values for data science methodologies.
- **Future Work:**
  - Expand clustering with more variables to enhance segmentation.
  - Develop predictive models to forecast property prices and prioritise drivers.