Data Science Analysis of Housing Affordability Dynamics in London Boroughs

Myo Shwe Sin Ei

I. Introduction

The housing market in UK plays an important part in the country's economy with real estate sector contributing £295 billion to gross value in 2024 [1]. Housing market dynamics affect the economy through three main channels: consumer confidence and spending, financial system stability through mortgage debt and GDP growth primarily through new construction activity rather than existing home transactions, affecting the economy through interest rate policy [2]. However, there still are challenges since housing completions in UK have consistently fallen short of the government's annual target (around 140,000 completions annually to 300,000 target), while housing prices are rising faster than wages [3].

This housing affordability issue is worse in London as the median price-to-earning ratio is at 11.1, meaning prices are 11 times of median-annual-earnings. Only 3% of London properties sold in 2024 met the traditional affordability threshold of five times earnings, down dramatically from 57% in 1997 [4]. It is also estimated that 4.3 million homes are 'missing' from UK housing market due to planning system restrictions on supply, with shortages highly concentrated in high-wage urban areas including London [5].

Fingleton, Fuerst and Szumilo demonstrated challenges with simple supply-side solutions where they used dynamic spatial panel models that increase housing supply in London and found out it has surprisingly weak effects on affordability [6]. Their simulations show that 15% increase in London's housing stock would reduce the mean affordability ration from 11.7 to 10.85 under supply-only effects. However, it reduces only to 11.07 when the demand was induced from increased employment. This forms a cycle where supply increase attracts workers, raises local income and thus, demand offsets price reductions. This paper builds on these insights by examining whether social housing provision, which may operate differently from market supply, shows similar or different patterns across London boroughs.

II. Analytical Questions and Data Description

A. Research Questions

In this paper, three research questions aimed to identify predictive relationships and casual mechanisms will be examined.

I. What is the relationship between changes in social housing stock and changes in housing affordability ratios across London boroughs?
II. Can London boroughs be clustered into housing market clusters and do they exhibit different affordability trajectories over time?
III. Can a composite gentrification index predict affordability outcomes, and which factors contribute the most to accuracy?

These questions are relevant towards London's current housing crisis where affordability ratios have doubled in some boroughs since 2004. Policy debates frequently link to declining social housing to worsening affordability yet empirical borough-level analysis remains limited. Understanding these dynamics can inform evidence-based housing policy making.

B. Data Sources and Sample

A panel dataset is created from nine datasets which were collected from authoritative government sources and merged on ONS borough codes (E09XXXXXX) and year. The dataset contains 693 observations across 31 variables and spans from 2004-2024, capturing both of pre-crisis period, the 2008 financial crisis, post-crisis recovery and recent market cooling.

Variables	Description
Affordability_ratio	House price to median earnings ratio (target variable)
House_price	Average transaction price
Social_housing_stock	Total social landlord properties
Social_rent_weekly	Registered social landlord average weekly rent
La_rent_weekly	Local authority average weekly rent
Vacant_dwellings	Empty housing count
Waiting_list	Households on local authority waiting list
Affordable_completions	New completed affordable housing units
Population	Mid-year population estimates
Borough_code	Geography code (E09XXXXXX)
Borough_name	Borough name
Year	Calendar year

C. Data Limitations

There are data limitations such as missing data from local authority rent observations where boroughs transferred council housing to housing associations, which created discontinuities resulting in 104 missing data. The affordability ratio's imputation from residence-based-earnings may not reflect local earnings as well. Small residential population (City of London) produces outliers, and some datasets contain suppressed values marked as '[x]'. These limitations were addressed through missing data handling with sensitivity to outliers in interpretation with functions such as clean_numeric().

III. Data and Analysis

A. Data Preparation

To get the main panel dataset, merging and balancing of nine datasets with different structures was necessary. Government statistics used different formats such as financial years ('2004-2005'), years as integers and mixed column naming. Four functions were used to standardise these mishaps where clean_numeric() was used for values like '[x]', clean_borough_code() for filtering borough codes, wide_to_long() to reshape clear year as column into panel format and parse_financial_year() to extract calendar years from financial year strings. A borough lookup table created from the affordability dataset was used as reference for merging on borough code to ensure consistent identification across the datasets with varied code systems.

B. Feature Engineering

Derived features were created for further analytical objectives. Features such as year-over-year changes are calculated using grouped diff() and pct_change() to enable analysis of dynamics rather than levels. Per-capita normalisation (waiting_list_per_100, social_stock_per_100, vacancies_per_1000) addressed borough population differences (e.g. a waiting list of 10,000 has different meanings for a borough with high population compared to low population). Weekly social rent was annualised (social_rent_annual) by multiplying with 52 weeks. The rent_gap_ratio was constructed by house price divided by annualised social rent, as a gentrification pressure indicator, following the theoretical framework of measuring divergence between market and regulated housing costs. A composite gentrification index was created as a weighted combination of house_price_pct_change, social_housing_stock_pct_change, vacant_dwellings_pct_change, and rent_gap_ratio with an alternative PCA-based gentrification index (gentrification_index_pca) also computed. An inner/outer London binary indicator was created using ONS borough classifications. Lagged affordability variables (1-3 years) were also created for predictive modelling to prevent data leakage while ensuring historical values predict current outcomes.

C. Exploratory Data Analysis

In Figure 1, the affordability ratios are seen to be persistent inner/outer London divergence with inner London peaking at 17.3 in 2017 before declining to 13.9 while outer London peaked at 13.2 in 2021 before falling to 11.3. house prices diverged highly with inner London reaching £754k compared to outer London's £510k at their respective peaks. Social rents followed parallel trajectories regardless of geography, more than doubling from £68-72/weeks to £139-143/week with inner and outer London tracking within £3-5/week throughout the timeline. Social housing stock showed steady growth from 316k to 429k units, contradicting the narratives of social housing decline. Waiting lists exhibited and inverse U-shape, peaking at 380k households in 2012 before declining in 2016, then recovering to 336k by 2024. Vacancies followed a U-shaped pattern as well, declining from early 2000s before recovering to 95k by 2024.

Figure 1. Time Series Analysis

Observing the correlation heatmap produced by Pearson correlation analysis, the relationship between affordability ratio and house price can be seen to have near-perfect correlation (r=0.95), which is a given. Rent_gap_ratio is also highly correlated with both house_price (r=0.92) and affordability_ratio (r=0.85). This raised concerns of multicollinearity for regression modelling and informed the decision to use percent changes rather than levels for the gentrification index.

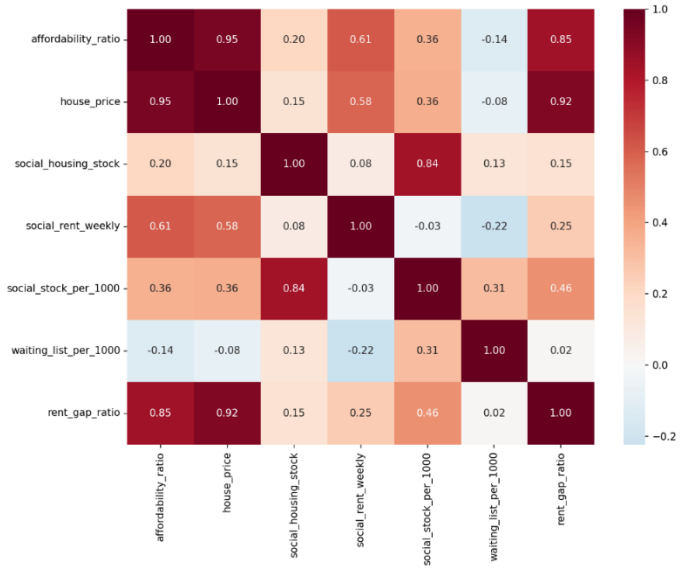

Figure 2. Correlation Heatmap

Pearson correlation and OLS regression with clustered standard errors were used for analysis because both stock change percent and affordability change are continuous and the hypothesis requires linear association. Clustered standard errors are necessary because panel data contains repeated observations within boroughs and standard OLS assumes independence between observations. Clustering shows changes as in one group rather than several unrelated events and widens the margin of errors, which results in avoiding false precision.

In Figure 3, borough level analysis examined the relationship between 20-years of percentage change in social housing stock and affordability ratio. The scatter plot reveals a negative relationship (r=-0.4275), which means boroughs with larger stock increases experienced smaller affordability deterioration. Inner London boroughs cluster in the upper left with smaller stock gains and worse affordability trajectories. Contrary to the initial narrative about 'loss' of social housing, most boroughs gained stock with an exception for Kensington and Chelsea.

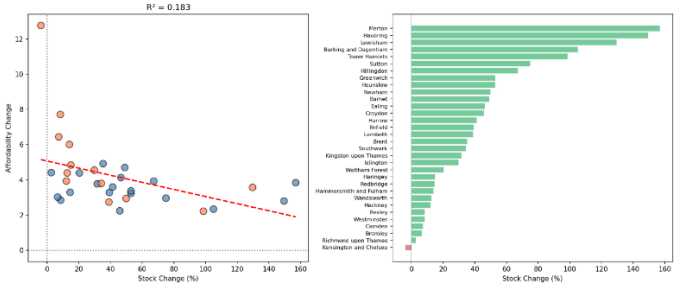

Figure 3. Housing stock change vs Affordability

K-means clustering was selected for identifying borough typologies since the variables are continuous and clusters are expected to be roughly spherical in feature area, which will produce borough clusters for interpretation in policy discussion about housing market ranges. Z-score was mandatory before clustering since K-means uses Euclidean distance. Otherwise, house prices would dominate affordability ratios and social stock per 100. These clusters were specified into three groups (affordable, mid-range and unaffordable). ANOVA validated that clusters represent distinct groups rather than arbitrary divisions.

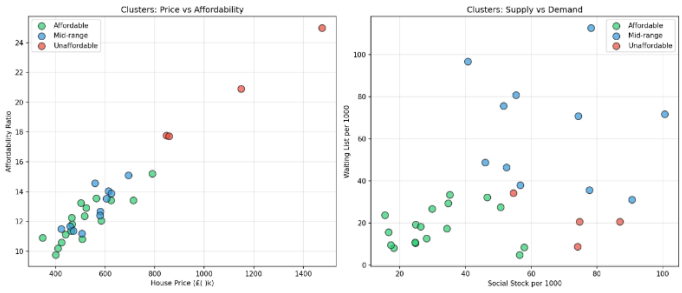

Figure 4. Housing affordability clusters by borough

The left plot shows clear separation on price affordability, and the right plot shows that 'Unaffordable' boroughs show high social stock per capita (60-100 per 1000) but relatively low waiting lists (10-35 per 1000) in Figure 4. This suggests inherited large social housing estates in expensive central boroughs, but allocation policies or affordability barriers affect who joins waiting lists. The statistical test

proved that there is significant difference across the clusters ($F=37.3$, $p<0.001$). This indicates variation between different groups are greater than variation within them.

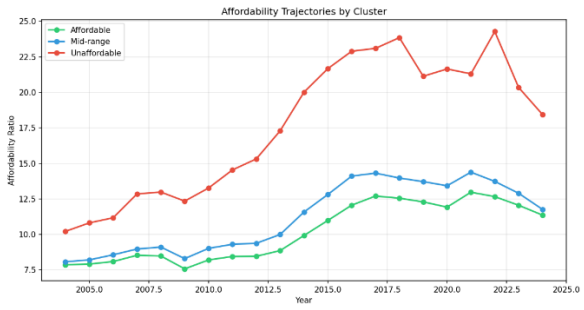

Figure 5. Cluster Trajectories

Tracking affordability over time by clusters reveals widening inequality between them. In 2004, they were relatively close (8-10). By 2017-2022, the 'Unaffordable' cluster had diverged sharply to ratios of 21-24 while 'Affordable' and 'Mid-range' remained at 11-14.

A composite gentrification index was constructed using heuristic weights reflecting gentrification theory with house_price_pct_change (0.4: highest as primary displacement mechanism), social_housing_stock_pct_change (-0.3: negative as social housing expansion oppose gentrification), vacant_dwellings_pct_change (-0.1: negative as vacancy reduction might signal demand), and rent_gap_ratio (0.2: due to Smith's rent gap theory) [7]. They were chosen to reflect the relative theoretical importance with directions based on expected relationships with gentrification pressure. While exact weight values involve heuristic judgement, the findings that this index underperforms simpler approaches suggests results are robust to weight specification. Standardisation before weighting ensures comparability across scales. PCA also provided a data-driven alternative (29% explained variance). Train-test split (70/30) was used for training and testing. Upon comparing Linear Regression, Random Forest and Gradient Boosting, similar performance across models was found ($R^2=0.94$), validating linear regression framework. Feature importance analysis revealed affordability_lag1 dominates (importance of 0.953) with gentrification_index contributing minimally (0.046) and inner_london_location negligible (0.001).

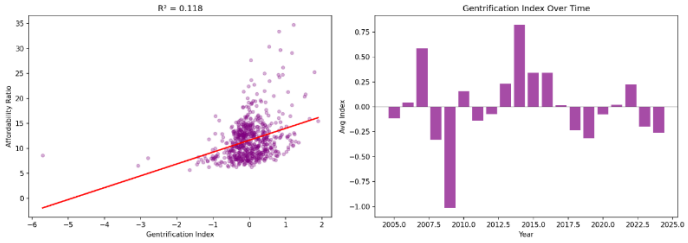

Figure 6. Gentrification Index

The left plot shows the index-affordability relationship ($R^2=0.118$), which is weaker explanation than the stock change analysis ($R^2=0.183$). This shows that more complex composite features do not necessarily improve data analysis. The right plot shows gentrification pressure over two decades, capturing economic cycles such as positive pre-2008 boom, sharp negative during crisis, recovery during 2014 and recent decline in 2018.

IV. Findings, reflections and further work

A. Key Findings

Upon analysis, it is found that boroughs building more social housing experienced smaller affordability deterioration although the relationship is small. Most boroughs gained stock over twenty years except Kensington and Chelsea which declined. These findings support the simulation research showing that even 15% supply increase reduces affordability ratio by less than one point [6]. K-means clustering identified three distinct typologies with diverging trajectories, and that the widening gap between Unaffordable cluster and others rose from 10.2 to 23.1 from 2004-2017. Compared to it, the Affordable cluster increased from 7.9 to 12.7, widening the between them from 2.4 to 10.4. This pattern shows the feedback mechanism where expensive areas attract high-wage workers who can afford high prices, creating spatial inequality [6]. Despite theoretical point of view, the gentrification index explained only 11.8% of variation, which failed to predict affordability outcomes effectively. Predictive models were found to be more persistent. Lagged affordability contributed the most to predictive power with 95.3% compared to gentrification index with 4.6%.

B. Reflections and Limitations

This paper addressed all three research questions with findings that challenge simple supply-focused housing narratives. Social housing stock growth shows modest negative association with affordability deterioration, although most boroughs built stock and experienced worsening affordability. Spatial clustering reveals a huge gap between three affordability clusters

and most strikingly, predictive modelling is found to be persistent which implies limited scope for short-term policy interventions. However, the high predictive R^2 shows more of explanation on past affordability rather than forecasting insights. The observational design prevents casual claims, which means other factors like local economic conditions and policies may have driven outcomes. Most significantly, treating boroughs as independent units ignored spatial dependencies and controlling for unchangeable borough traits like historical significance or geographical limitations would've separated temporal changes from fixed traits between boroughs.

C. *Further Work*

Incorporating borough-level income data would breakdown affordability changes into price versus earning effects and clarify whether crises primarily reflect housing costs outpacing wages or stagnant earnings. Adding transport accessibility would allow for a better analysis of if commuting patterns shape housing demand distribution. Extending coverage pre-2004 would also capture earlier housing cycles and Right to Buy impacts. For further work, explicitly modelling demand feedback mechanisms would test whether social housing operates differently from market supply in attracting population and address the puzzle of why supply increases fail to improve affordability.

V. References

[1] Office for National Statistics (UK) (2025). *UK real estate GVA 2021*. [online] Statista. Available at: https://www.statista.com/statistics/760134/real-estate-sector-gross-value-added-in-the-uk/.

[2] Bank of England (2020). *How does the housing market affect the economy?* [online] www.bankofengland.co.uk. Available at: https://www.bankofengland.co.uk/explainers/how-does-the-housing-market-affect-the-economy.

[3] Keep, M. (2023). Housing Market: Key Economic Indicators. *commonslibrary.parliament.uk*. [online] Available at: https://commonslibrary.parliament.uk/research-briefings/sn02820/.

[4] Office for National Statistics (2025). *Housing affordability in England and Wales*. [online] Ons.gov.uk. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/bulletins/housingaffordabilityinenglandandwales/2024.

[5] Rollison, C. (2024). *Do falling house prices mean the housing crisis will soon be over?* [online] Centre for Cities. Available at: https://www.centreforcities.org/blog/do-falling-house-prices-mean-the-housing-crisis-will-soon-be-over/.

[6] Fingleton, B., Fuerst, F. and Szumilo, N. (2018). Housing affordability: Is new local supply the key? *Environment and Planning A: Economy and Space*, [online] 51(1), pp.25–50. doi:https://doi.org/10.1177/0308518x18798372.

[7] Smith, N. (1979). Toward a Theory of Gentrification: A Back to the City Movement by Capital, not People. *Journal of the American Planning Association*, 45(4), pp.538–548.

[8] GeeksforGeeks (2017). *Regression in machine learning*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/regression-in-machine-learning/.

[9] GeeksforGeeks (2021). *Feature Selection Techniques in Machine Learning*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/feature-selection-techniques-in-machine-learning/.

Datasets

[1] Department of Communities and Local Government (2024). *Affordable Housing Supply, by Borough* [Dataset]. Available at: https://data.london.gov.uk/dataset/dclg-affordable-housing-supply-borough-e64g0/.

[2] Department of Communities and Local Government (2025) *Median and Lower Quartile Ratio of House Prices to Residence-Based Earnings* [Dataset]. Available at: https://data.london.gov.uk/dataset/ratio-of-house-prices-to-earnings-borough-2z04n/.

[3] Department of Communities and Local Government (2025) *Households on Local Authority Waiting List* [Dataset]. Available at: https://data.london.gov.uk/dataset/households-on-local-authority-waiting-list-borough-vd6oo/.

[4] Department of Communities and Local Government (2025) *Vacant Dwellings* [Dataset]. Available at: https://data.london.gov.uk/dataset/vacant-dwellings-e79gj/.

[5] Department of Communities and Local Government (2025) *Local Authority Average Weekly Rents* [Dataset]. Available

at: https://data.london.gov.uk/dataset/local-authority-average-rents-2g1k1/.
[6] Office of National Statistics (2025) *Population estimates – local authority based by single year of age* [Dataset]. Available at: https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?theme=32.
[7] Department of Communities and Local Government (2025) *Registered Social Landlord Housing Stock* [Dataset]. Available at: https://data.london.gov.uk/dataset/registered-social-landlords-average-rents-24rn6/.
[8] Department of Communities and Local Government (2025) *Registered Social Landlord Average Weekly Rents* [Dataset]. Available at: https://data.london.gov.uk/dataset/registered-social-landlord-housing-stock-2n83y/.
[9] HM Land Registry (2025) *UK Housing Price Index* [Dataset]. Available at: https://data.london.gov.uk/dataset/uk-house-price-index-2zwx6/.

Word Count

Introduction	297
Analytical question and data	247
Analysis	954
Findings, reflections and further work	387
Total	1882

AI Declaration

I hereby declare that AI was used to assist in the completion of this coursework. Please find below a list of the prompts used, along with details on where and how they were applied.

Prompt used	How it was used
"My housing dataset has name columns like 2004-2005 as financial years but I need just the first year to merge with other datasets based on year alone. How can I deal with it"	This is for dataset merging based on years.
"my housing dataset has strange values like [x], LSVT in numeric values. Explain it briefly to me and how can I covert these to NaN for calculations?"	This is for dataset cleaning for unusual values.