London Calling: Exploring the influence of short-term rentals on housing affordability in pre-pandemic inner London

jsrobson
42 min read · 19 hours ago

Photo © Benjamin Davies / Unsplash

*This post is written in satisfaction of the second deliverable of the term project for **CIS-5450: Big Data Analytics** for the Fall 2023 semester of the University of Pennsylvania's Master of Computer and Information Technology program.*

Study prepared by C. Peinado, J. Robson, and C.F. Uba, December 2023.

Introduction

Short-term rental (STR) services have been a major driver of urban transformation through their impacts on business travel and tourism, neighbourhood composition, and their challenge to the historical dominance of the hospitality sector. Since its founding in 2008, the STR marketplace provider Airbnb has expanded to over 4.5 million listings in countless cities across the world and has resulted in over 300 million unique stays and USD $41 billion in revenue to its legion of hosts (Deboosere et al., 2019).

The company and its adherents cite a number of benefits made possible by the platform. In the ideal, Airbnb–and STR providers, more generally–offers customers cheaper and more diverse housing options with greater geographic spread relative to traditional hoteliers that tend to dominate central business districts within an urban area. For property owners, Airbnb supports the monetization of otherwise underused housing assets. For the broader community, Airbnb provides new economic development opportunities by attracting tourists to previously unvisited parts of the city (Hati et al., 2021).

Conversely, critics note the distorting effect of STR platforms on housing markets. Under an STR model, property owners are incentivized to convert long-term rental dwellings into short-term rentals that yield an increase in monthly rental incomes. This had led to a number of impacts at urban and regional scales, including a decrease in housing affordability, rental unit consolidation, illegal operation outside of municipal licensing requirements, noise, and (perceived) reductions in safety for residents (Koster et al., 2021; Wachsmuth, 2021).

Problem statement: As the capital of the United Kingdom (UK), Greater London is a bellwether for the impacts of the STR sector on the UK housing market. A 2020 report by the Greater London Authority (GLA) reflected on a decade of broad-scale STR sector operation in the region's housing market; it revealed that there were over 80,000 properties listed on the service with 23 per cent thought to be in violation of the authority's 90-day rental limit for STRs (GLA, 2020). In some of the region's inner boroughs, upwards of seven

per cent of the housing stock is thought to have been converted by owners for short-term rental use (Temperton, 2020).

> *"As Mayor, I know that our housing crisis is the biggest threat to London's future. It is the main reason why all Londoners cannot share in our city's success."* — **Sadiq Khan, Mayor of London, 2018**

Greater London is a prime contributor to the UK's gross domestic product, and so with it, its labour market. The region's predominant role in the national economy, coupled with the growth of the STR sector and the continued speculation in the region's housing market have strained the ability of Londoners to afford safe and stable housing within their means (Mayor of London, 2018).

Study approach: In this study, we will use regional housing purchase price data, STR listing data, and supplemental indicators for measures of deprivation and taxation to generate insights about the relationship between the STR sector and property pricing in Greater London.

We will apply standard data cleaning, exploratory data analysis (EDA) and regression modelling techniques using Python data science libraries to build a predictive linear model for median housing prices by borough given local STR sector activity and the confounding influences of urban dynamics and taxation. We will also consider, as a corollary, the impact of amenities and general property attributes on the pricing of STR properties.

Given the body of research on the relationship between housing affordability and the STR sector, we reason that a density of STRs within a borough is positively correlated with increased housing costs (see for example: Shabrina et al., 2021).

Data Foundations

To facilitate our study of the STR sector and the housing market in London, we draw on two primary and two supplementary data sources:

Attribute-linked residential property price dataset for England and Wales, 2017–2019. This consolidated dataset was produced by University College London (UCL) researchers to synthesize distinct official property price and domestic energy performance certification datasets prepared by the UK Government for the period 2011–2019; it is made available through an Open Government Licence for use.

For the purposes of this study, the dataset was reduced in temporal scope to the years 2017–2019 inclusive to ensure feasibility while meeting project data row count requirements. The resulting `properties` dataset contains feature columns for sale price, transfer date, postcode, property type, sale duration, street, locality (borough), floor area, and a host of (interesting but unused) environmental performance measures.

The dataset is notable in that it allows for a calculation of GBP-per-metre-squared (cost per floor area) given the consolidation of the two distinct datasets with one holding purchase information and the other holding building information. This will allow a common reference for pricing regardless of the building type or floor area.

Airbnb listings data, to Q2 2020. This dataset was generated and maintained by the data-driven advocacy group Inside Airbnb and is available under a Creative Commons A4: Attribution licence. The advocacy group explores the impact of the STR sector on local communities through a critical lens. The archival dataset exceeds 50,000 rows for the period 2019–2020 and contains feature columns for the borough, geolocation, property type, room type, number of beds and baths, amenities, price per night, minimum and maximum nights, long-term availabilities, and unique identifiers. The 2020 dataset contains information about all Airbnb hosts dating to 2015. Two flavours of dataset were used in this project:

1. `listings_cond` : a condensed set of features for the principal property price/listing count investigation, and
2. `listings_detail` : a detailed set of features for the supplemental STR feature/pricing investigation.

Supplemental data sources, 2019. The English Indices of Deprivation (UK Government, 2019) will be consulted to provide potential confounding variables within the study.

The English Indices of Deprivation are the central government's measure of relative deprivation for borough and sub-borough scales in England (UK MHCLG, 2019a). The 2019 iteration is informed by 39 separate indicators, which are themselves organized across seven distinct domains of deprivation at both borough and sub-borough scales. The latter scale is officially defined as the Lower-layer Super Output Area (LSOA): a "homogenous small area of relatively even size containing approximately 1,500 people" (UK MHCLG, 2019b, 13).

Our study explores the domains of deprivation for income, education, employment, health deprivation and disability, crime, barriers to housing and services, and living environment at the sub-borough level.

We also attempted integration of an additional confounding variable in the form of local authority taxation bands (Greater London Authority, 2019). However, this data is collected at the borough-level such that we saw significant repetition and noise from its inclusion in a dataset composed at the LSOA level. With this in mind, taxation bands were excluded from consideration.

Study Design Considerations

To maintain project feasibility, we narrowed the scope of analysis from Greater London in total (32 boroughs) to the City of London and the Inner London boroughs (13) of Camden, Greenwich, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Lewisham, Newham, Southwark, Tower Hamlets, Wandsworth, and Westminster. The majority of STR properties in Greater London are found in these 14 areas (GLA, 2020).

The study is limited to a pre-COVID-19 context given the availability and alignment of data, and the distorting influence of the pandemic on housing, tourism, and travel since the second quarter of 2020.

For modelling purposes, we use the LSOA sub-borough scale rather than the more conventional borough scale as it provides an opportunity to introduce a sufficient amount of complexity by focusing on a more fine-grained geographic scale. For example, Greater London comprises 32 boroughs and the City of London. In comparison, there are 4,835 LSOA within the same boundary area. Conversely, when talking about general trends within the Indices of Deprivation — as in the following **Exploratory Data Analysis** section — we will apply a borough-level scale to ensure a sufficient geographic area on which to draw and discuss conclusions.

Study Organization

The study is organized into three conceptual sections that each contribute to our understanding of the dynamics of the housing market in London. These are:

1. Exploration of the relationship between STR density and the impact of purchase price within selected boroughs
2. Integration of potential confounding variables focused on English Indices of Deprivation for 2019, and
3. Supplemental investigation of the importance of STR characteristics as a determinant of cumulative pricing (defined as the sum of weekly rental cost plus security deposit and cleaning fees).

These sections are arrayed across multiple Google Colab (Jupyter) Notebooks. Each notebook used in the project is described below for reference:

1. **CIS5450_TermProject_01_acquire_data.ipynb:** This notebook — generally archival in nature — reflects the workflow to identify, collect, and parse resources, discretize large datasets, clean and parse data, and conduct preliminary data investigations. It captures the whole of the foundational data acquisition work, including ideas that were tried and later discarded in favour of more optimal approaches.
2. **CIS5450_TermProject_02_load_data.ipynb:** This notebook reflects a foundational step of the overall project process. Here, we read project datasets into a Dask dataframe format for modification and analysis in later project steps. In this notebook, we mount a shared drive of project data in comma-separated value (.csv) format, and define and run a series of functions to read the datasets.
3. **CIS5450_TermProject_03_compose_data.ipynb:** This notebook includes the necessary step to collate, clean, and compile a project dataset from the distinct data sources introduced in **Data Foundations**. The resulting product from this notebook is a project dataset comprising a series of summary data joined with the English Indices of Deprivation to illustrate

data trends at the sub-borough level. This project dataset is used as the basis for modelling tasks.

4. **CIS5450_TermProject_04_map_data.ipynb:** This notebook includes code to visualize several cartographic products that help a selection of influences on the housing and STR markets in inner London.
5. **CIS5450_TermProject_05_Bivariate_Model.ipynb:** This notebook explores a bivariate relationship between a dependent variable `cost_fl_area` (the average property cost per square metre in an LSOA geographic area) and an independent variable `count_list` (the number of short-term rental [STR] listings in a LSOA geographic area) using standard Linear Regression, Random Forest, and Decision Tree (both regression) modelling approaches. This is the first of three notebooks that apply a modelling approach to the data we have sourced, cleaned, and collated.
6. **CIS5450_TermProject_06_Expanded_Variable_Model.ipynb:** This notebook continues exploration of the linear relationship between the dependent and independent variable above. In this notebook, we introduce several **confounding variables** in the form of the scores for the domains established in the English Indicies of Deprivation 2019, collected at the LSOA level for Greater London. This is the second of three notebooks that apply a modelling approach to the data we have sourced, cleaned, and collated. We apply several linear regression techniques–including the introduction of neural network architecture–to explore this relationship.
7. **CIS5450_TermProject_STR_Feature_Review.ipynb:** This notebook reflects a corollary data investigation premised on a detailed variant of the STR listings dataset that explores the relationship between STR property features and their cumulative pricing. In this notebook, we build a condensed data pipeline by loading, cleaning, and analyzing detailed STR data before applying a series of modelling approaches to explore the relationship.

The notebook programming environment draws upon several pre-defined Python modules including, among others: Contextily, Dask, Geopandas, Matplotlib, Numpy, Pandas, SciKit-Learn, Seaborn, Statsmodels, and Requests.

We will include major inflection points, methodologies, and important takeaways in the remainder of this post in satisfaction of the **Deliverable 2** requirement. Each notebook used in the project has been packaged and submitted in satisfaction of **Deliverable 1** requirements. We have included a `deprecated_archive` folder, which includes earlier iterations of data cleaning, data composition, and linear regression modelling premised on borough-level data.

Exploratory Data Analysis

The project team conducted several exploratory data analyses across the three conceptual sections. Within each section, the analysis conducted falls into one of two categories:

- **Assessment of dataset quality and shape:** In this category, we use both textual and visualization tools drawn from various Python libraries to investigate the distribution, quality, and characteristics of the dataset itself. The respective datasets are considered for null instances, frequency of statistical outliers, value counts, and feature composition.
- **Summary of data outcomes and trends:** In this category, we use tools drawn from geopandas, matplotlib, and seaborn libraries to output data visualizations summarizing outcomes and trends from the data that are relevant to our study topic. For instance, we used the contextily and geopandas libraries to output a combination of STR data and projected cartographic data into choropleth maps that illustrate the geography and data trends that characterize inner London.

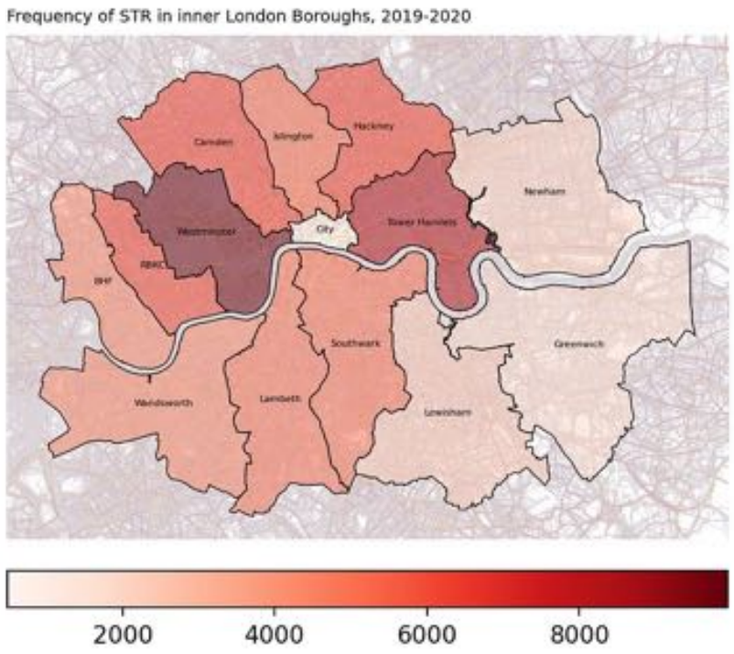
In this section, we will predominantly focus on the second type of analysis. For review of both types, a comprehensive exploratory data analysis process is included in detail within the provided notebooks under **Deliverable 1**. Cost is expressed in Pound Sterling (£GBP) and time range is for the period 2019–2020 unless otherwise specified.

In this section, we will explore:

- Frequency of STR properties by borough
- Average value by floor area (£GBP/m^2) of properties by borough

- Average local council taxation rate (all bands) by borough
- Prevalence of STR property amenities
- Types of housing forms as STR
- Review of levels of deprivation for domain areas of crime, employment, health, housing, income, and living environment
- Notable correlations observed in respective datasets

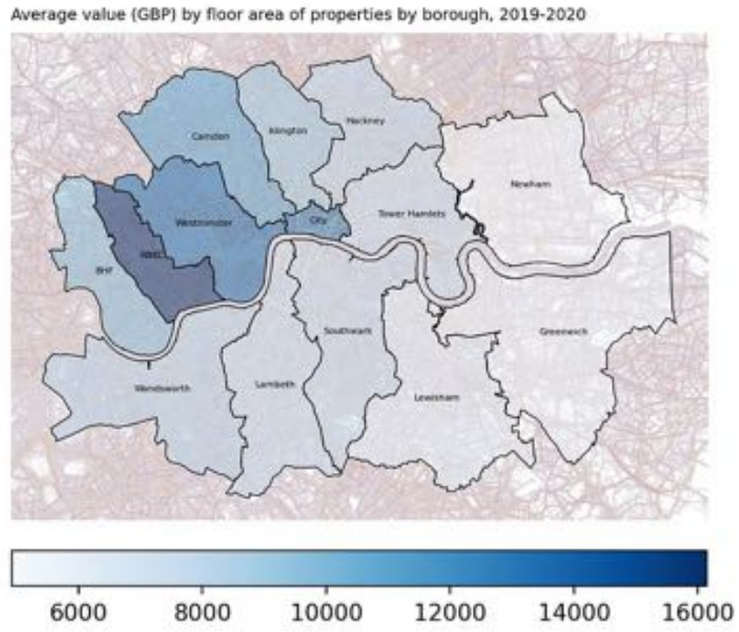
Frequency of STR properties by borough: Within the survey area of inner London and after preliminary data cleaning, we have identified a total of 63,671 STR properties. The boroughs with the most STR listings are Westminster (9,433) and Tower Hamlets (7,627). Conversely, the boroughs with the least STR listings are the City of London proper (461) and Greenwich (1,683) at a distant second.



Data courtesy of Inside Airbnb / GLA, 2019–2020

The median count of STR properties is 4,764, with the boroughs of Southwark (4,803) and Lambeth (4,725) closest. We observe that those boroughs north of the River Thames generally exhibit a greater count of STR listings within their bounds than do those south of the river.

Average value by floor area of properties by borough: Within inner London, the boroughs with the highest average value by floor area are Kensington and Chelsea (£16,136/m^2), Westminster (£13,431/m^2), and the City of London (£12,990/m^2). These outcomes are not surprising, as the leading two boroughs feature some of the most expensive postcodes in which to reside in the UK, and the City is the seat of the UK's financial services industry with a small residential footprint.

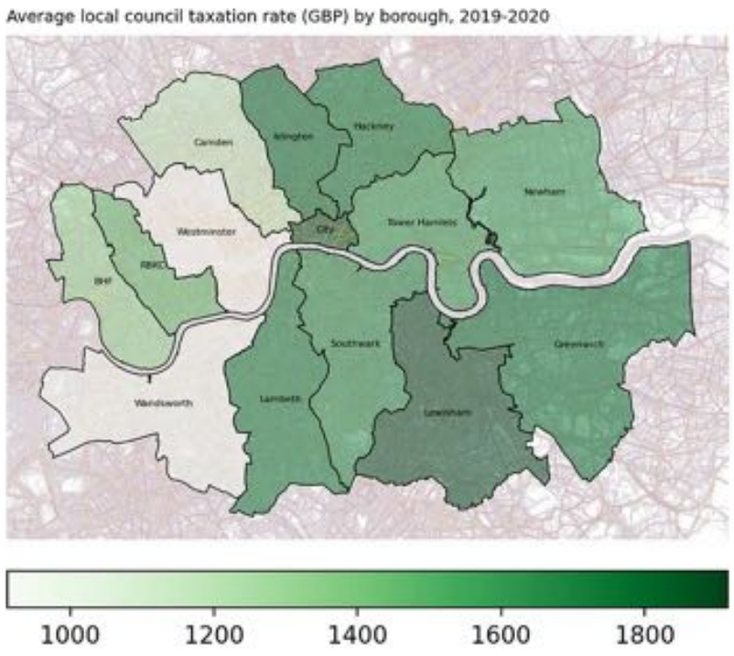


Data courtesy of UK Government / UCL / GLA, 2019–2020

We generally observe that the western boroughs north of the River Thames exhibit higher property prices than those to south of the river, and particularly, those to the east.

The surveyed boroughs with the lowest average value by floor area are Newham (£4,932/m^2), Greenwich (£5,277/m^2), and Lewisham (£5,814/m^2). Similarly, these outcomes follow our expectations, as the boroughs are located on the eastern fringes of inner London in a capital area that has historically prioritized its western boroughs.

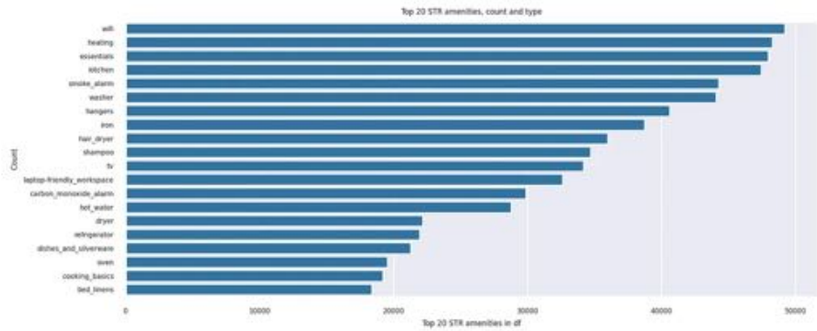
Average local council taxation rate (all bands) by borough: In England, each property is assigned to one of eight council tax bands on the basis of its value. The assignment of a property to an identified band determines how much council tax is paid by its owner. For convenience, we have collapsed the distinct band rates into a single average band tax rate.



Data courtesy of GLA, 2019

The surveyed boroughs with the highest average band tax rate are Lewisham (£1,914), the City of London (£1,888), Greenwich (£1,800), and Islington (£1,800). Conversely, the boroughs with the lowest rate are Westminster (£913), Wandsworth (£931), and Camden (£1,175).

Prevalence of STR property amenities: In the exploration of the feature importance of STR amenities as a determinant of cumulative price, we explored the amenities most and least present in STRs within inner London. We reasoned that in both cases, they were largely immaterial; people would not pay a premium for a class of amenities available in most STRs nor those in so few that it is clear no demand exists. Note that these values were collected following a first pass drop of unneeded rows.

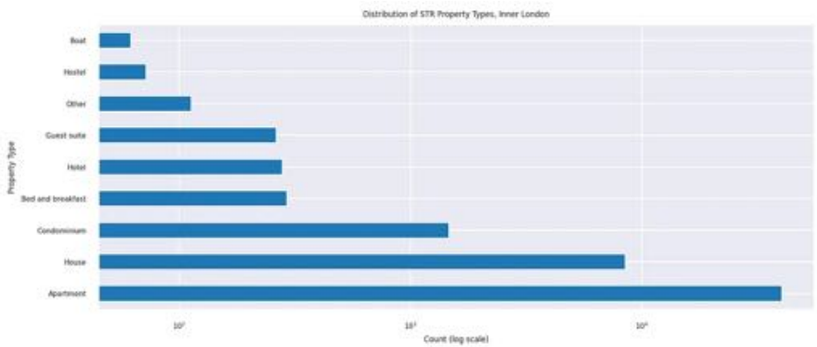


Data courtesy of Inside Airbnb, 2019–2020

The most prevalent STR amenities were wireless internet (49,206), heating (48,291), an all-encompassing and vague set of “essentials” (47,976), kitchen (47,555), and smoke alarm (44,299). Conversely, the least prevalent amenities were a private pool, a mountain view, a beach view, a mobile hoist, and an electric profiling bed (each with one instance).

One would argue that as London possesses neither a mountain view nor a beach view, some of these amenities may be misplaced (or hopeful).

Types of housing forms as STR: During the same exploration of feature importance, we sought to uncover the types of housing forms commonly found as STR properties. Prior to developing the count, we truncated the least occurring types as outliers or, in some cases, misallocations.



Displayed using logarithmic scale. Data courtesy of Inside Airbnb, 2019–2020

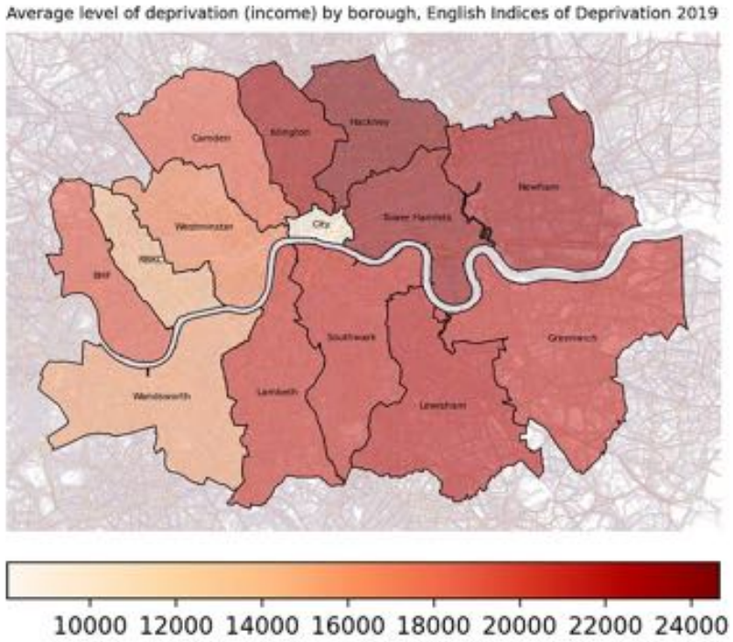
The most frequent — unsurprising, given London's population density and historical development trends — were apartments (40,206). In a distant second and third were houses (8,503) and condominiums (1,468). The least frequent were houseboats (62), hostels (72), and a nebulous "other" category (113).

Review of levels of deprivation for domain areas:

In this section, we explore observed trends in the English Indices of Deprivation (2019) for the select domains of interest in inner London. We use the average rank calculations used by the UK Government; these aggregate the LSOA geographic scale into higher-level and more approachable borough units. The ranking is interpreted such that geographic areas with higher ranks are considered more uniformly deprived. Those with lower ranks are more unevenly deprived (that is, there is more polarization among the constituent LSOA units (UK MHCLG, 2019b, Sections 3.3.7–3.3.9). The cartographic visualizations are performed at the borough level as the scale is more readily approachable for discussion than that of the LSOA (14 boroughs to 1,800+ LSOA).

Here, deprivation is used as a general term. For example, a borough with a high ranking for health deprivation will exhibit uniformly poor access to health services and concomitant health outcomes. Similarly, a borough with a high ranking for housing deprivation will uniformly experience barriers to housing access, quality, and safety.

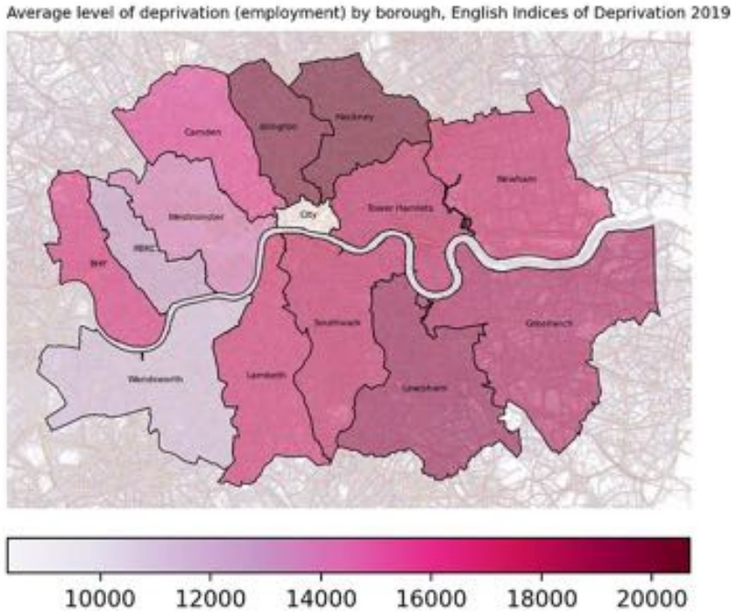
We will very briefly explore findings from EDA for the English Indices of Deprivation across the domains of income, employment, health deprivation and disability, crime, barriers to housing and services, and living environment at the borough level. This will provide quick insight into the socioeconomic composition of each study borough.



Data courtesy of GLA, UK Government, 2019

Income: This domain measures "the proportion of the population in an area experiencing deprivation relating to low income" (UK MHCLG, 2019a).

Review of income deprivation reveals a distinct east/west income disparity in inner London. The western boroughs — such as Kensington and Chelsea, Westminster, and Camden — exhibit far more polarization in their ranking. This suggests a wealthier strata of residents at the top of the income hierarchy. By contrast, the eastern boroughs — such as Hackney, Tower Hamlets, and Newham — display far less polarization but comparatively more income deprivation according to the index.

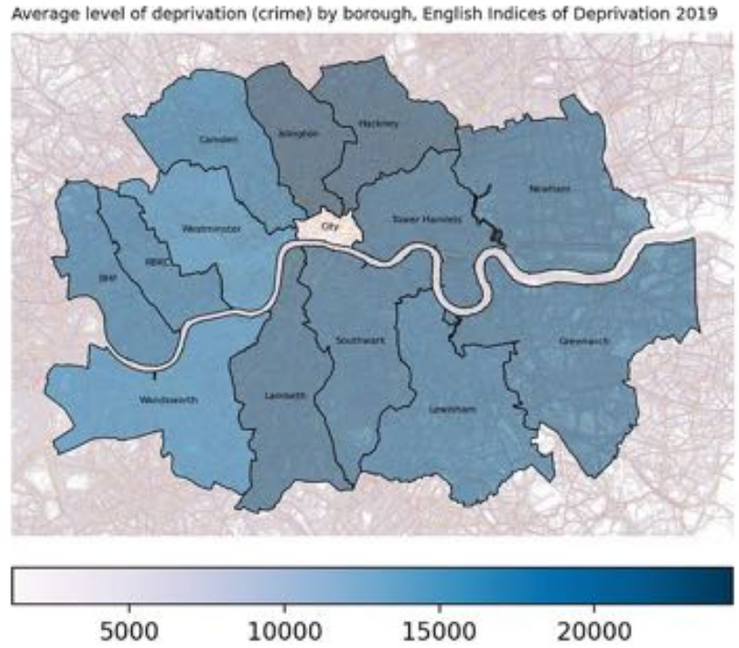


Data courtesy of GLA, UK Government, 2019

Employment: This domain measures "the proportion of the working age population in an area involuntarily excluded from the labour market" (UK MHCLG, 2019a).

The deprivation index for employment presents a similar picture to that of income. The western boroughs again show far more polarization and therefore less uniform deprivation in access to employment, which we might suspect from historically wealthy parts of Greater London.

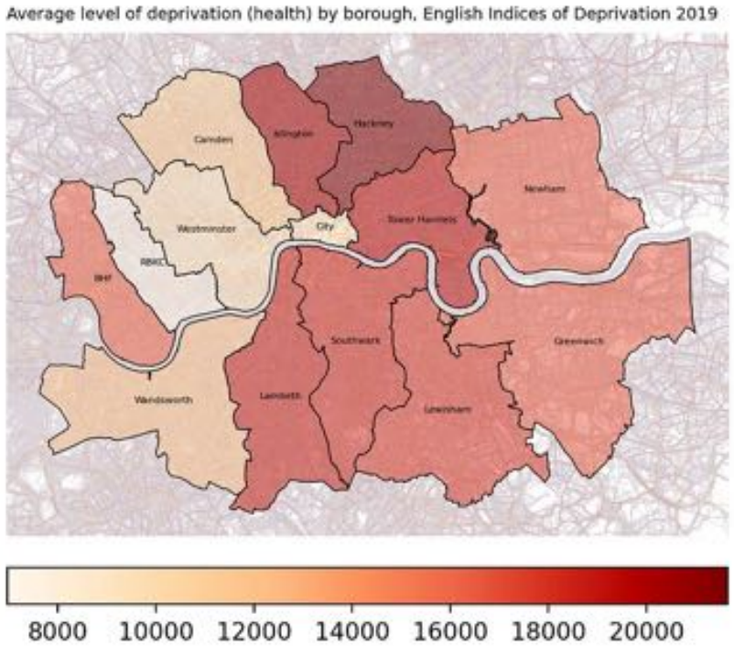
Conversely, the boroughs to the north and east of the City of London exhibit more uniform employment deprivation, which affirms the prior conclusions on income.



Data courtesy of GLA, UK Government, 2019

Crime: This domain measures "the risk of personal and material victimization at the local level" with reference to violence, burglary, theft, and criminal damage (UK MHCLG, 2019a).

The deprivation index for crime does not show the same extremes as the previous indices. While the western boroughs show greater polarization — i.e. LLSOA units in Westminster will cumulatively exhibit a greater gulf between minimal and maximal crime rates — relative to their eastern counterparts, this is less severe than the disparity among boroughs in other domains.

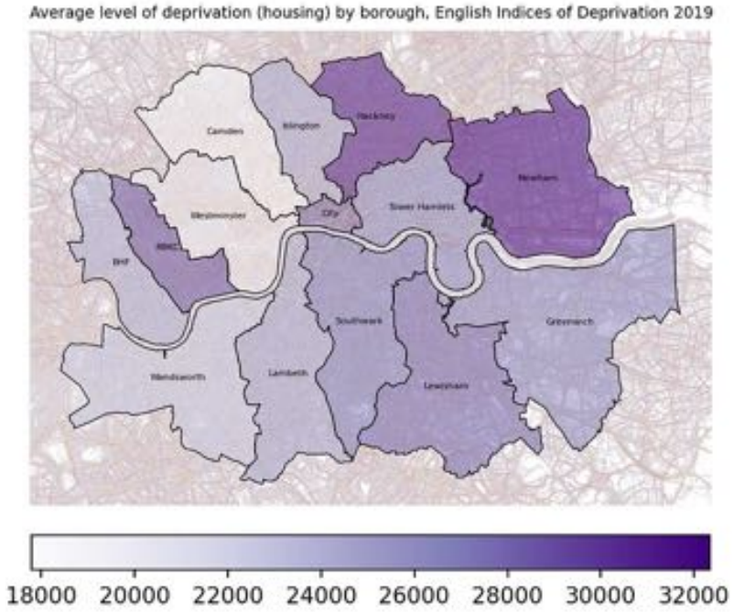


Data courtesy of GLA, UK Government, 2019

Health: This domain measures "the risk of premature death and the impairment of quality of life through poor physical or mental health" (UK MHCLG, 2019a).

Unsurprisingly, the deprivation index for health follows a similar geographic distribution as with income and employment. The western boroughs, with the exception of Hammersmith and Fulham, exhibit far more polarization and less uniform deprivation in the domain of health than do the eastern boroughs.

The relationship between health outcomes, income, and employment are well-reported in the literature (see for example: Akanni et al., 2022; Local Government Association, 2020).

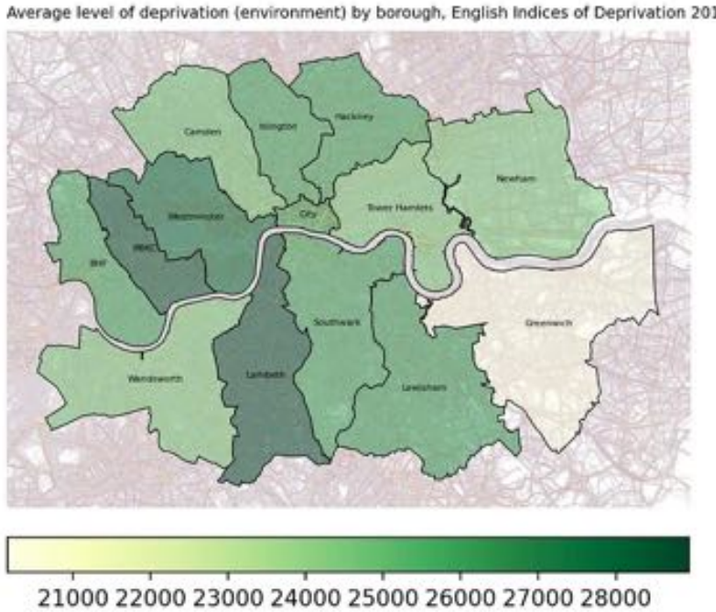


Data courtesy of GLA, UK Government, 2019

Housing: This domain measures "the physical and financial accessibility of housing and local services" with focus on geographic, social, and economic barriers to housing (UK MHCLG, 2019a).

The deprivation index for housing barriers reveals some interesting insights relative to the other domains. At first glance, the state of housing in inner London looks less severe than previously anticipate.

However, when we consider the minimum and maximum values of the indices, we observe a far higher minimum value relative to the minimums of the other indices we have explored. That is, the wealthier western boroughs exhibit far less uniform deprivation in accessing housing, which we might expect given an overlay of the income and employment indices. Conversely, the other boroughs exhibit far greater uniform deprivation, with particular emphasis on the boroughs of Hackney and Newham.



Data courtesy of GLA, UK Government, 2019

Living Environment: This domain measures "the quality of the local environment" including quality of housing, air quality, and road traffic accidents (UK MHCLG, 2019a).

The deprivation index for environment quality reveals an interesting converse to what we have observed previously. First, we again note that the minimum index values are far greater than other explored indices. Second, we observe that the western boroughs — particularly Kensington and Chelsea, Lambeth, and Westminster — exhibit higher levels of uniform deprivation than their eastern counterparts. On first glance, this is contrary to what we might expect given the historical development of west London and the Indices of Deprivation for income and employment.

The broad scope of the domain category makes it difficult to explain this observed behaviour. Are these uniformly deprived boroughs suffering from uniform deprivation because of poor housing quality? Or, equally likely, are the wealthier boroughs suffering from traffic congestion and road accidents due to their popularity and disposable income yielding more automotive traffic on local roads?

Modelling

In this section, we discuss the modelling approaches taken across the three part of the study process specified in **Study Organization**. Conclusions from each modelling approach are discussed within their respective sections. Challenges and broader directions for future study are discussed in subsequent sections.

Data composition

The basis of our modelling workflow is a LSOA-level summary dataset constructed from the foundational project datasets for property price paid / environmental performance `properties` and STR listings `listings_cond`, each containing ≥ 50,000 rows of data. The work performed to clean, collate, and compose the summary dataset was performed in the notebook titled in part **compose_data.ipynb.**

Design considerations for the data composition were two-fold. First, we observe there is no one-to-one match between listings and property data; that is, we could not know which properties in inner London were listed on the STR platform Airbnb. As a consequence, we needed to introduce some generalization where geographic scale was concerned.

Second, as discussed in **Study Design Considerations**, the borough-level was *too* general a scale (14 rows) at which to conduct our study. This necessitated using LSOA-level scale to introduce a sufficient geographic granularity at which to conduct the analysis. As is discussed in the modelling notebooks themselves and in the **Challenges Encountered** section, this was sufficient for modelling but, as always, more data would assist in the development of a more accurate model.

The LSOA-level summary dataset includes the following features:

LSOA11CD: Given LSOA code, 2011 boundaries
avg_str_price: The average price (by night) of a STR property within the LSOA
cost_fl_area: The average cost by floor area (m^2) of a property within the LSOA
count_prop: The count of properties within the LSOA boundary
count_list: The count of STR listings within the LSOA boundary
borough: The borough (local government authority) to which the LSOA is a part
imd_score: The index of multiple deprivation (cumulative) score, UK Index of Deprivation (ID)

income_score: The income deprivation score, UK ID
employ_score: The employment deprivation score, UK ID
educ_score: The education, skills and training score, UK ID
health_score: The health deprivation and disability score, UK ID
crime_score: The crime deprivation score, UK ID
house_score: The barriers to housing and services score, UK ID
env_score: The living environment score, UK ID

In addition, for the corollary analysis of STR features (third of three modelling notebooks), we draw upon the detailed variant of STR listings for the same time period `listings_detail`.

Modelling a bivariate relationship

In the first modelling notebook, we explore a bivariate relationship between the dependent (target) variable `cost_fl_area`, or the average property cost per square metre in an LSOA geographic area, and an independent (feature) variable `count_list`, or the number of STR listings in an LSOA area. We apply standard linear , random forest, and decision tree regression modelling approaches. As the central focus of our study, we pose the question:

> *Is an active (dense) STR sector within a LSOA geographic unit positively correlated with increased housing costs in the same unit?*

In this notebook, we explore only the bivariate relationship by removing all other features and assume that the cost per square-metre is directly informed by the count of STR listings.

```
                            OLS Regression Results
==============================================================================
===
Dep. Variable:           cost_fl_area   R-squared:
0.241
Model:                            OLS   Adj. R-squared:
0.241
Method:                 Least Squares   F-statistic:
598.5
Date:                Sat, 02 Dec 2023   Prob (F-statistic):          5.33e-
115
Time:                        10:08:55   Log-Likelihood:
-17623.
No. Observations:                1885   AIC:
3.525e+04
Df Residuals:                    1883   BIC:
3.526e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
===
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
---
Intercept   6546.0639     88.534     73.938      0.000    6372.428
6719.700
count_list    42.2477      1.727     24.464      0.000      38.861
45.635
==============================================================================
===
Omnibus:                      614.022   Durbin-Watson:
0.706
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
2698.257
Skew:                           1.505   Prob(JB):
0.00
Kurtosis:                       8.029   Cond. No.
70.9
==============================================================================
===
```

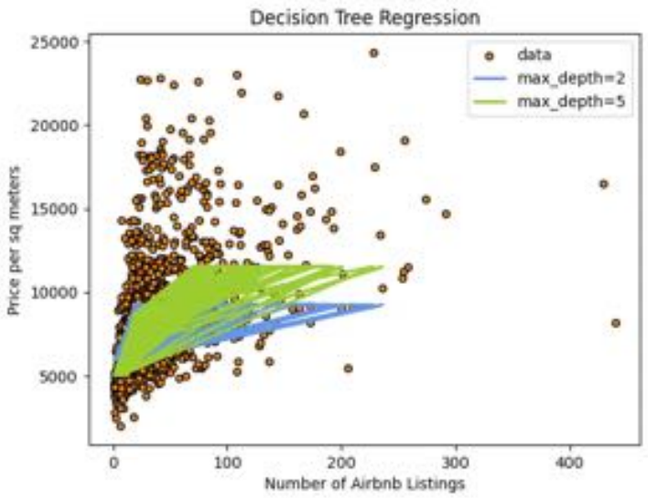
Standard linear regression: We run a standard linear regression model using the statsmodels package, and find that for every unit increase in the number of STR listings, the cost per square-metre increases by 42 per cent. We can easily develop a cost function in Python with this in mind:

```
# Given initial cost per square metre, a number of STR listings, and a percentag
# by which to increase cost, function calculates the cost per square metre incre

def calculate_cost_increase(initial_cost_per_sqm, num_STR_listings, increase_per
    # calculate the updated cost per square meter
    updated_cost_per_sqm = initial_cost_per_sqm * (1 + increase_percentage * num
    return updated_cost_per_sqm

# example usage:
initial_cost = 100    # initial cost per square metre
num_listings = 3      # number of STR listings
increase_pct = 0.42   # percentage by which to increase cost

# get cost increase from function
cost_increase = calculate_cost_increase(initial_cost, num_listings, increase_pct
print(f"The cost increase per square metre is: {cost_increase}")
```

Introducing other flavours of regression: A motivating factor to explore other regression models is to determine if any performance improvement can be found. Observe the R-squared value of the standard linear regression model only achieves 0.24; that is, the model explains only ~ 24% of

variability in the dependent variable. Can improvements be found? We explore **decision tree** and **random forest regression** variants.

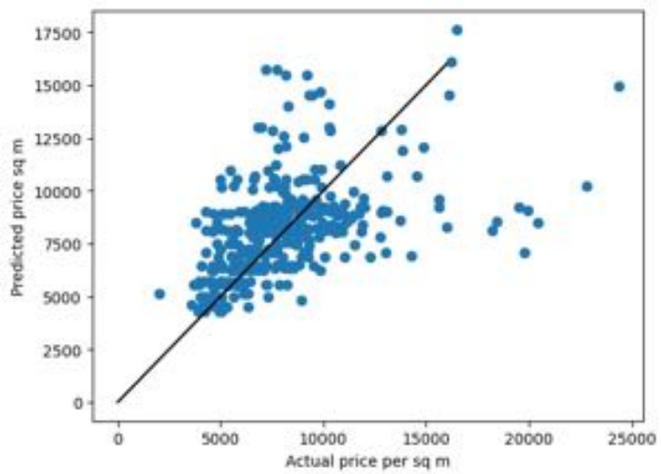
Decision tree may be suitable on the basis of a non-linear relationship between the two variables, while random forest may introduce resilience with respect to the presence of outliers and the potential for overfitting.



Decision tree model output

For the decision tree model, we find of trialled depths for `max_depth=[2,5]` that 5 is preferable for our data composition for its balance of bias and variance. We observe a root mean square error (RMSE) of 2,605.98 — suggesting an average predictive error — and a R-square score of 0.27, offering minor performance improvements over the standard linear regression model.

Although **random forest** is often used as a classification model, we select it for use in our regression trials. Instead of predicting a label or class as we might expect in a classification model, we use it to predict the cost per square-metre. We observe that our plot visually indicates a better model despite comparatively sub-optimal RMSE and R-square scores of 2,788.68 and 0.17 , respectively.

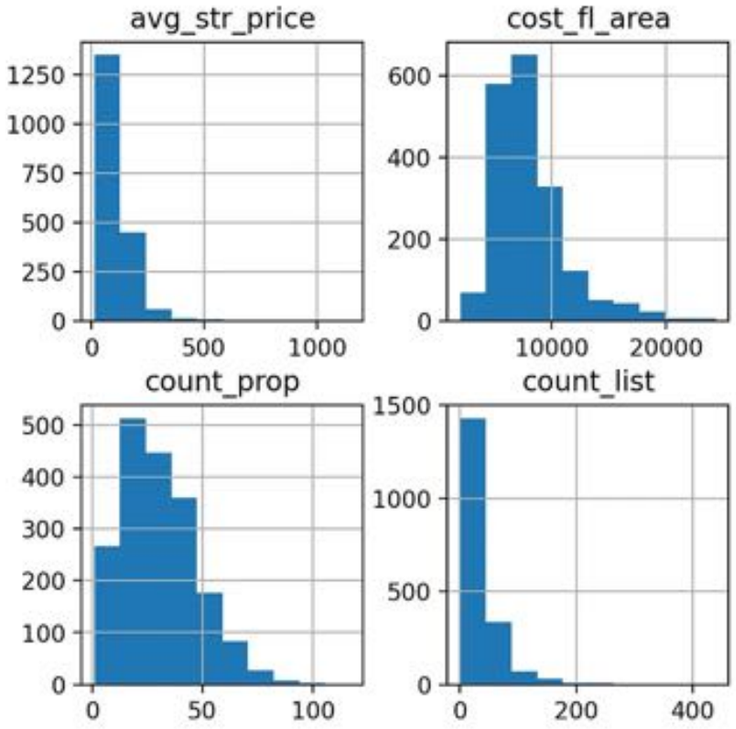


Random forest model output

We now conduct cross-validation using SciKit-Learn's GridSearchCV method to see if we might improve the random forest model performance. In GridSearchCV, we automate the tuning of hyper-parameters in an exhaustive search by passing predefined hyper-parameter options to the method, which will then try all combinations and return the best performing of the group. Our parameters of interest for GridSearchCV are:

- `cv` : number of cross-validation to perform. In this case due to the large dataset we will do 3 splits. That means 2 splits will be kept for training, and one for testing.
- `estimator` : the model instance.
- `params_grid:` the dictionary object that holds the hyper-parameters we want to adjust, as above.

The use of GridSearchCV to determine the optimal hyper-parameters for use in the random forest model improved performance with RMSE and R-square values of 2788.68 and 0.27, respectively. Contrary to our initial premise, the decision tree and random forest models converged in their overall predictive performance.



Histogram output of sentinel distributions, denoting potential outliers

Can we **iterate the random forest model** further? We revisit the shape of the data and discover potential outliers among each of our four sentinel features focused on average STR price, the cost per square-metre, and the count of properties and of listings. We remove rows exhibiting significant deviation from the mean to eliminate these outliers and, in doing so, potentially improve model performance.

The same workflow is performed by conducting an un-tuned fit using derived training data and tuning for optimal parameterization using GridSearchCV. Given the dataset with outliers removed and optimal hyper-parameters, the RMSE and R-square suggest improved predictive performance relative to prior trials, with values of 1,946.25 and 0.31, respectively.

In **conclusion**, by managing outlying data through iterative exploratory data analysis and cleaning, we have observed an improvement in RMSE scores, with improved performance from the tuned model with outliers. The error margin remains stubbornly greater than we'd like, with an average error per estimate of 1,946.25.

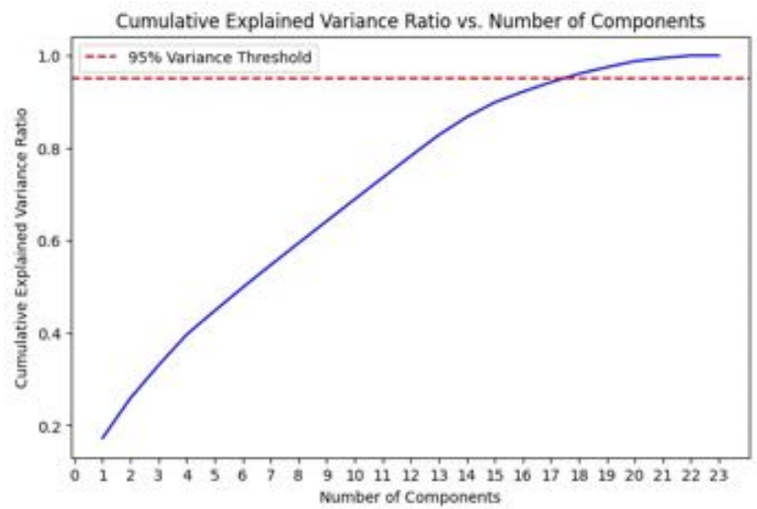
We posit that the introduction of confounding variables may help the predictive accuracy of subsequent models by better reflecting the complexity of the real world, albeit in a manageable way.

Introducing confounding variables

Despite our best attempts, the predictive performance of a model founded on the bivariate relationship remains stubbornly low. Can we improve performance with the introduction of confounding variables?

In the second modelling notebook, we use several **confounding variables** in the form of the scores for the domains established in the English Indices of Deprivation 2019, collected at the LSOA level for Greater London. We use the same LSOA summary dataset and this time included the indices for consideration.

We conduct standard data processing tasks and drop unneeded columns from consideration, including the LSOA code `LSOA11CD`, the employment index domain score `employ_score`, and the indices of multiple deprivation score `imd_score`. In particular, the latter feature exhibits a high correlation with other domain-based indices of deprivation as it is a cumulative composite of those scores.

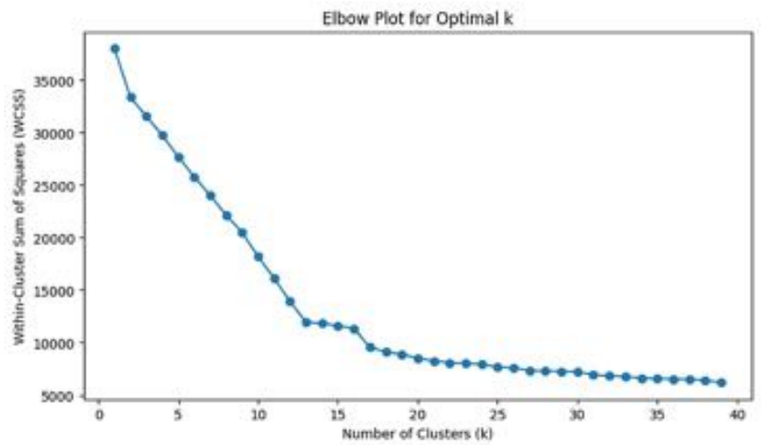


PCA analysis, visualized

We first conduct a series of **unsupervised learning** tasks in the form of principal component analysis (PCA) and K-means clustering. In **PCA**, we attend to the existence of multi-collinearity between features. We determine there are 19 components within our dataset that help explain 95% of the variance. With this in mind, we refit and transform the PCA object on the training set only. Using the PCA-inflected training data, we find improved performance using a standard linear regression model.

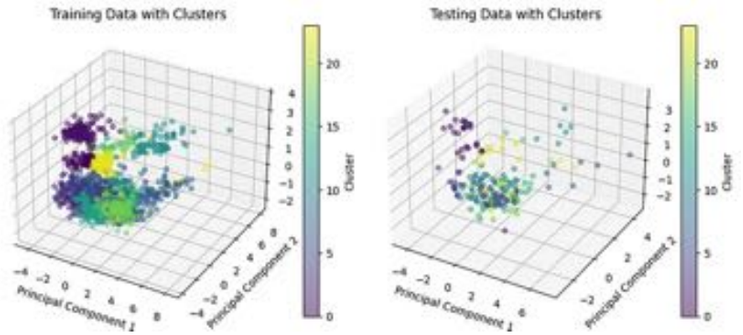
Here, the RMSE and R-square values are 1,278.53 and 0.85, respectively; the introduction of confounding variables represent a marked improvement for model performance relative to the simple bivariate relationship.

We now consider **K-means clustering**, or the possible groupings in our dataset. We visualize a graphical "elbow" plot to help us choose the number of clusters that minimize the Within-Cluster Sum of Square (WCSS) while avoiding over-fitting the model.



K-means clustering output

We observe that the elbow forms in the visualized plot around a *k* value of 23. We set this value and fit on the result of our reduced training set. Using two-dimensional visualization of the resulting clusters, we do not clearly observe the extant clusters in our dataset. We apply a three-dimensional visualization to see if we can better separate the clusters conceptually.



Three-dimensional visualization suggests the emergence of clustering

With the emergence of clustering, we represent the average values of features within each cluster and observe which emerge as the most significant. By a runaway margin, the feature that most influences the clustering result is the average pricing of STR properties `avg_str_price` within the LSOA area.

Given an emerging sense that the confounding variables might support improved predictive performance, we now move to a **supervised learning approach**. We split our dataset into test and train datasets as before, but this time allocate only 10 per cent of the data to the test set rather than the usual 20. Given the limited number of rows in the dataset, a 90 per cent training set allows the model to be exposed to a broader range of scenarios. While other methods-such as cross-validation-may be beneficial, our approach prioritizes maximizing training scenarios with the available data.

We conduct **four flavours of regression modelling**: standard linear regression, Lasso (L1 regularization), Ridge (L2 regularization), and ElasticNet (L1/L2 regularization) using the Scikit-Learn library. This follows the workflow established in the first modelling approach where a model is instantiated, fit on training data and used to predict on the test set. From this, we derive the RMSE and R-square values for comparative performance evaluation. The motivation for each flavour of linear regression is denoted below:

- Lasso Regression (L1 regularization) is used to reduce the possibility of overfitting and improve model generalization.
- Ridge Regression (L2 regularization) is used to attend to any extant multi-collinearity in the data in the event our hand-tuning of features was not as comprehensive as we would hope.
- ElasticNet (L1/L2 regularization) is used to benefit from a blend of both described L1 and L2 regularization approaches.

The RMSE and R-square values of the standard linear regression model deployed with our test and train data was 1,160.34 and 0.85, respectively. Indeed, these values were repeated across the different flavours of regression model with only minuscule difference, indicating similar fitting and predictive performances. From this, we infer that our **confounding variables** have generally supported significant gains in the predictive performance of a linear model with average cost per square-metre as the dependent variable.

As a final exploration of the confounding variables, we explore potential performance gains through deep learning techniques using the Pytorch package and neural network (NN) architecture underpinned by a Multi-Layer Perceptron (MLP) Regression model. This is largely conducted on an experimental basis; we post that the straightforward regression models explored to this point are likely sufficient for our purposes.

Here, we build two NN architectures: one with, and one without **dropout**. The concept of dropout helps to mitigate model overfitting by randomly "dropping out" (or, setting to zero) a proportion of architecture neurons during each training iteration. This prevents any single neuron from relying too much on specific features in the training data. This, in turn, makes the network more robust and generalizable.

Each NN is **characterized** by three layers traversing from 256 to 128 to 64 outputs before convergence on an output layer. A ReLU activation function is used to contend with any prospective vanishing gradient issue. The number of training epochs is set at 200.

For the NN **with dropout**, we observe RMSE and R-square values of 1,086.89 and .87. For the NN **without dropout**, we observe RMSE and R-square values of 1,093.76 and .87, suggesting very little overfitting exists within the dataset as interpreted by the NN architecture.

In **conclusion,** the deep learning NN models exhibit both a slightly lower RMSE and a slightly higher R-squared score relative to the prior approaches; this suggests enhanced performance. It appears that, for this dataset, the deep learning approach offers a modest improvement in predictive performance over the traditional supervised learning (linear regression) methods tested previously.

Across the models explored in this second notebook, we observe that the introduction of confounding variables has, perhaps unsurprisingly, improved model performance by sufficiently increasing the dimensionality of the surveyed dataset. This is ideal, both as it comports with our general assumptions on the importance of a richly contoured dataset and that real-world indicator data is an important influence on overall predictive accuracy.

Corollary: STR Feature Review

In the third of three notebooks, we depart from our study question and consider a corollary: what is the relationship between the **cumulative**

pricing of an STR property on a weekly basis and several geographic, programmatic, and quality-of-life based factors. That is, what are the factors that influence the dependent variable of cumulative weekly pricing?

For the purposes of our review, cumulative pricing is defined as the weekly price of an STR plus the levied cleaning and security fees.

Where academics in disciplines of economics, sociology, and urban planning consider our main study question, the corollary question we have posed is often the domain of entrepreneurs, real estate investors, and software developers interested in the housing investment sector. However, we will consider this corollary question from the lens of our central focus: given a comprehensive dataset of STR instances and their defining features, what might we learn about the factors that are correlated to STR pricing and what might that tell us about the influence of STRs on an urban housing market?

In this section, we draw on a more detailed variant of STR listings data. The listings_detail dataframe includes a row count exceeding 50,000 and several salient features for analysis. These are:

id: unique identifier for STR property
neighbourhood_cleansed: cleaned representation of borough name
property_type: the type of dwelling (i.e. apartment, townhouse)
room_type: the accommodation offered (i.e. entire home, private room)
accommodates: the number of people the property accommodates at one time
host_listings_count: the number of listings available by host
number_of_reviews: the number of reviews given to the STR property
bathrooms: the number of bathrooms within the STR property
beds: the number of beds within the STR property
bed_type: the type of beds available (i.e. regular bed, air mattress, futon)
amenities: the distinct amenities available within the STR property (as string list)
price: the nightly rate for accommodation in the STR property
security_deposit: a refundable amount of money held during the guest's stay
cleaning_fee: the fee associated with cleaning the STR property following the guest's stay
guests_included: the number of guests accommodated by the nightly rate
extra_people: the fee associated with hosting people additional to those accommodated by the nightly rate
minimum_nights: the minimum number of nights required to book the STR property
maximum_nights: the maximum number of nights possible to book the STR property
instant_bookable: t/f indication of whether property may be booked without communication to host
availability_90: the number of days within a 90-day window in which the STR property is available for rent
cancellation_policy: the in-place policy for cancellation (i.e. strict 14 days with grace period)

With a significant number of rows and features, there was a great deal of **pre-processing** required as a necessary precursor step to data modelling. The companion notebook includes a significant amount of cleaning activities, including standard data casting and pruning activities, outlier identification and management through delimiting of Z-score deviation, consolidation of similar feature values into overarching categories, log transformation of select features, and conversion of categorical types into numbered values where possible. For a detailed glimpse into these activities, please consult the companion notebook.

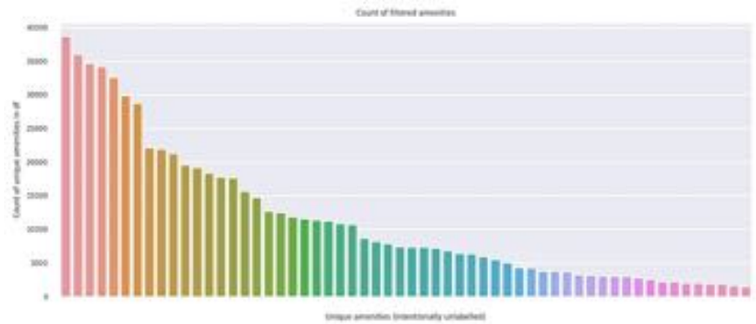
A central premise of the corollary question is the influence of **amenities** on pricing. The dataset includes an `amenities` feature that, while a string representation, effectively serves as an array of comma-separated values. Within the dataset of detailed listings, there are over 46,000 different variations on the array, with 186 different listed amenities overall.

We generate functions to parse amenity strings, build them as Python lists, and count the unique occurrences of each amenity across the whole dataframe as a means to assist our composition of the dataframe. This is performed using three functions:

- `parse_amens(string)` : Function parses a string representation of the listed amenities for each STR property, as defined above.
- `add_amens(l, d)` : Function adds parsed amenities from a list `l` to a pre-existing dictionary `d` ; counts the number of amenity instances at each addition.
- `build_dict_amenities(df)` : Given a dataframe `df` , function uses `parse_amens()` and `add_amens()` to build a sorted dictionary of cumulative amenity instances and their counts.

This workflow enables the ability to cast each amenity, in effect, as a one-hot encoding by answering the question “does this STR property have the identified amenity?”. However, given the overall number of amenities, we

first pruned away those that were nearly universal and those that were so few as to be non-factors.



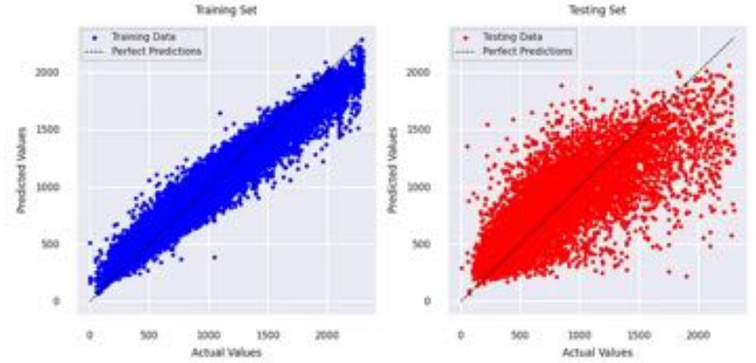
The count of amenities following pruning of both dominant and non-factor amenities.

A standard (immense) correlation matrix was used as the basis to drop any further features from the dataset exhibiting strong correlations as they are noted as redundant predictors. These activities resulted in the decrease of features in the dataset from 210 to 75.

With a cleaned and parsed dataset in place, we apply a series of **supervised learning** linear regression models as we have before: standard linear regression, Lasso regression (L1 regularization) Ridge regression (L2 regularization), ElasticNet regression (L1/L2 regularization) and random forest regression. We remain in the linear regression environment as we seek to model a continuous relationship between a set of independent variables (features) and the dependent cumulative price variable (target).

Using a standard linear regression, we observed an R-square score of 0.57 and a general tendency towards the importance of non-amenity features as a predictor for cumulative price. Indeed, in each regression instance apart from random forest, we observe the same R-square value with minor variation on the importance of features that later aggregate to gain comprehensive insight. The equivalent performance of each model suggests:

- general feature importance, such that Lasso did not vary significantly from the standard model, and
- little extant collinearity, such that Ridge did not differentiate itself in terms of performance.



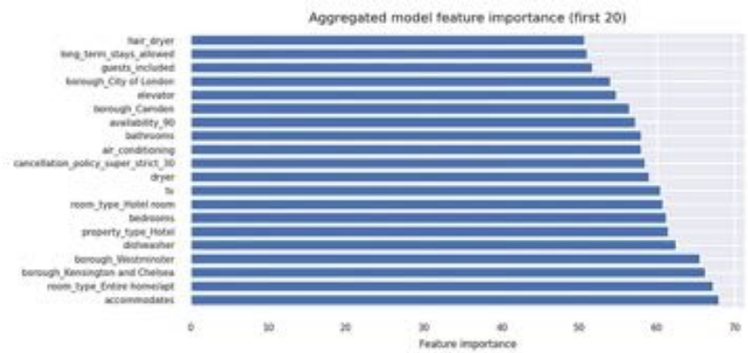
Visual reference of the predictive performance on the training and test data sets for our random forest model

We observed improved performance in the application of a random forest regression model, with a R-square score of 0.64. However, review of mean squared errors for the training and testing datasets and of a visualization of training and testing set predictions, we conclude that the model is overfitting and thus, for now, the shared performance of the other models remain as our benchmark.

As our final modelling approach, we look to **NN architecture** with a MLP Regression model. We arbitrarily set hyper-parameters; our hidden layer sizes are set to decreasing outputs of 150, 100, and 50, with a ReLU activation layer and a maximum iteration count of 500. Unlike prior models, there is no intrinsic feature importance that can be derived from the MLP model, so we are unable to comment on feature importance. Instead, we focus only on predictive performance: does the NN model exceed our benchmark? At least with arbitrary hyper-parameters, the answer is *no*. We observe a R-square value of 0.36, far below that of the standard regression model.

We use RandomizedSearchCV to help tune the hyper-parameters as for a dataset of this size, GridSearchCV proves too time-intensive. In contrast to the exhaustive approach, a fixed number of parameter settings are sampled from probability distributions specified by the randomized search. The resulting hyper-parameters are hidden layer sizes of (100, 50, 30), an identity activation function, a constant learning rate, and a maximum number of iterations set to 200. This considerably speeds model performance to a trivial degree, while causing convergence on the baseline R-square value, with a return of 0.56.

As the R-square value is equivalent to the standard supervised learning models despite the complexity, time, and performance requirements of the neural network, we eschew the use of a neural network for this dataset given optimal performance using a simpler approach.



Aggregated feature importance across the supervised learning models

Given convergence on the same R-square value, we now consider the aggregate feature importance using a function to gather common indicative features from the respective supervised learning models.

We observe that the features that contribute most to the variance in cumulative price `c_price` span categories of:

- **quality-of-life** (i.e. infrastructure and amenities such as elevator, hair dryer, television)
- **programmatic** (i.e. number of guests included in price, availability over 90 days, long-term stays allowed), and
- **geographic / spatial** (i.e. borough location of STR, type of room, number of beds and baths).

From a **quality-of-life** perspective, we observe that the provision of extra amenities can positively impact pricing, though in most instances, the amenities included in the graph are relatively trivial (i.e. a `hair_dryer` or a `tv`). The only listed amenity with a significant cost attached is the `elevator`, which is a sentinel indicator for the age of the building (newer than 1820), the height of the building, accessibility to those with mobility challenges, and maintenance costs (and their potential to be passed to the consumer). Across the amenity class of features, there is very little variation among them with regard to their influence on pricing (that is, the presence of an `elevator` is equally as likely to influence pricing as is a `hair_dryer`).

We consider the **programmatic** perspective; that is, the rules established by the STR proprietor or the hard constraints imposed by the STR property's design and orientation. Programmatic features include the number of `guests_included`, the number of people accommodated by the STR, the cancellation policy, and whether `long_term_stays_allowed`. Here, we observe the feature `accommodates`, as the carrying capacity of the STR property, as the most significant global predictor of STR pricing. This is intuitive such that we would expect a property able to accommodate multiple guests as being more generally expensive than a property accommodating one guest.

Finally, from a **geographic** perspective, we consider both borough membership and the defining characteristics of the STR property. Indeed, the largest influence on price according to the model is the feature `room_type_Entire home/apt`, which may suggest this room type is a differentiator between affordable and expensive STR stays. We see this similar logic applied in the number of `bedrooms` and `baths`. These are intuitive, as we might expect the number of bedrooms and baths to be indicative of the total floor area of the STR, and thus the cost required to rent it. Furthermore, it is perhaps no surprise that the features for location in the `borough_Kensington_and_Chelsea`, and `borough_Westminster` have an influence on pricing. The boroughs of Kensington and Chelsea, and Westminster are the first and second most expensive boroughs in London in which to live (and as such, the most expensive in the United Kingdom). We would correctly assume then that an STR's location in those boroughs would reflect both broader property pricing and more nebulous concepts such as exclusivity.

So what **may we infer from the results of this evaluation**? Contrary to our preliminary perspective as we began this work, amenities are not generally considered as significant influences on the pricing of STR properties in inner London. Instead, we find the most significant influences predominantly relate to the borough location and the built form (i.e. room type, number of beds and baths) of the STR.

In some respects, this is helpful to the broader discussion around STR frequency and purchase prices. While an STR can be augmented with new amenities to suit platform demands, geography and built form are stable. However, we know from data investigations that the geographic influence is not inflexible. The borough with the most STRs in the capital is Westminster; from this analysis, we might expect Kensington and Chelsea to contain the second-most. However, we find this borough to be the comparatively less well-off borough of Tower Hamlets. From this, we may infer that STRs with appealing built form exist throughout the inner boroughs of London and beyond and are not bound to the most desirable neighbourhoods.

Challenges Encountered

As we progressed through the study process, we encountered several challenges pertaining to practical data issues, project scale, and domain knowledge. Each challenge is described below in brief.

Structuring and organizing the work: The linear nature of the data analysis pipeline — acquisition, loading, cleaning, exploring and analyzing, and modelling — occasionally conflicted with the distributed nature of the team environment. Each team member was assigned a different portion of the study process to manage. While this generally functioned well, there were moments of project "bottlenecking" as one team member necessarily had to complete their allocation of study tasks before another was able to continue to the next process task.

This was exacerbated by the Colab Notebook environment as we could not rely on conventional version control systems such as Github to branch code and allow for collaborative coding and refactoring.

Working with large datasets: To ensure the project requirement of ≥ 50,000 data rows was met and to ensure the dataset was sufficiently robust to analyze and use as the basis for modelling, the team prioritized the integration of large datasets into our process. In terms of the study, we can conceptualize a large dataset as one with both a significant number of rows and many defining features.

This prioritization led to a property price dataset `ppd` that originally comprised every property sold in England and Wales over a 27-year period from 1995 to 2022 and a comparatively smaller STR dataset `listings` that included every STR listing in the Greater London area available on the service at time of scraping. The magnitude of these datasets required continued iteration of the study terms and assumptions to balance the size and complexity of the data such that it remained interesting and useful while not being overwhelming to both human comprehension and computational ability.

The team's attempts to scale the dataset required iteration. Initially, we conducted additional domain research to understand the scope of the study through the lens of STR coverage and the relationship between inner London and its suburban counterparts in outer London. We thus scaled to an appropriate geographic and temporal subset — from 32 boroughs to 13 and the City of London and from a decade-long span to a three-year period immediately prior to the pandemic. This simplification, it was believed, would ensure us able to conduct the study process efficiently.

Exploratory data analysis and preliminary attempts to build mathematical models from the resulting data revealed challenges that gave us pause. By condensing to 14 boroughs and working largely with summary data gleaned from operations on the datasets, our models were overfitting and yielding perfect predictions from the passed data. That is, our efforts to simplify the data proved too successful.

With this in mind, we introduced greater complexity into the dataset mid-process by moving to a more granular geographic scale by recalibrating our datasets at the LSOA scale. This immediately increased the extent of the datasets from 14 rows (one for each borough and the City of London) to nearly 1,900 (one for every LSOA with available listing and property price data) and allowed us to move forward with our data modelling tasks.

Composing an interesting dataset: Given the above considerations of scope, it became tempting to simply conclude our data selection activities given large datasets for property price data and STR listings with condensed features. However, we had to reflect on how those datasets would be used in concert in analytic and (especially) modelling activities. This would

ultimately constrain us to a single independent variable (STR count) when considering the relationship to the dependent variable of purchase price.

Additional effort was made to find supplemental data sources. The project team became acquainted with different public domain government datasets including environmental performance data, official indices for deprivation, and local council taxation band rates. While by no means comprehensive in context of a broad and complex urban system, these supplemental data and confounding variables help to introduce some additional nuance into the contours of data analysis and modelling.

Assessing urban systems data: The different scales of human settlement — whether regions, cities, or neighbourhoods — reflect a significant analytical challenge due to their complexity and dynamism. Throughout the study process, the team was aware of the narrowness of the perspective associated with the data. For instance, due to limitations in the data and for reasons of feasibility, we necessarily had to remove from consideration issues of local and regional policy formation, regional economies, global trends in tourism and real estate, and historical realities of London's growth trajectory in our analysis.

Consideration of any — to say nothing of *all* — of these factors would necessarily require more time not possible in a single semester course (and would be more reflective of an urban planning or sociology course, anyway). This meant that as it came time to evaluate results from the modelling process, there was a sense among the team that we may have built a compelling closed system but one that does not necessarily capture the complexity of its subject. In this respect, our final takeaways necessarily arrive with caveats given both these and other constraints.

Balancing emergent explorations and concrete process: With any project, one must balance the opportunity to explore new threads as they emerge with the need to follow a defined process to constrain the given scope and maintain feasibility.

The project team contended with this challenge by establishing a series of project milestones and building a regular meeting schedule to ensure the project remained on track while allowing flexibility to attend to ideas as they emerged. Team members were granted the opportunity to explore different ideas and approaches so long as they could provide rationale for their choices and integrate with the existing schedule.

The fragmented information environment held implications for how we proceeded in the study process; we simply did not know the extent of available information when we began. Had we started the project proposal process from a place where we had the finalized datasets for the project, we may have ultimately pursued an alternative approach to the same study or assumed a different study entirely. For example, a rich trove of residential environmental performance data was sourced in the data acquisition process, but only on the basis of requiring one feature — total floor area `tfarea` — based on prior identified needs. An alternative approach may have seen more or all of the performance data integrated into the study.

Scaling ambition given constraints: With any study design, there is a challenge to include the appropriate amount of study elements given time and resource constraints. Include too few elements and there is a missed opportunity to do more with the data we selected and the experience and knowledge we have gained through the course. Include too many elements and we are soon overwhelmed such that we are working only to complete the work and not, ultimately, to learn from it.

With this in mind, the project team would continually iterate and scale the project proposal and work as needed; if more time was available, we drew on the opportunity to trial new ideas and approaches such as use of XML syntax for web-scraping or geospatial data packages for cartographic visualization. Conversely, if we found sections in the study process were consuming too much time, we would evaluate alternative approaches to condense the work as much as possible and maintain scope.

Future Direction(s)

This study revealed in its composition and outcomes several future directions that support an expanded understanding of the subject area, with respect to regional housing markets and affordability, or could make use of the selected data differently. Potential future directions include:

Exploration of the potential policy implications of the findings: How do prospective regulatory and policy changes to STR impact property prices?

Are there optimal regulatory frameworks that balance the interests of the STR sector and the communities they impact? Is such an outcome even desirable? These policy implications could be tested through domain understanding of their quantitative impacts and therefore, their modelling using regression techniques.

>> **Alignment with Big Data Analytics:** Translation of policy outcomes into qualitative data that can be integrated into a similar workflow as was used in this study.

Refined spatial analysis of STR clusters: Building on the work performed in this study, how might a concurrent mapping of popular tourist destinations and neighbourhood-level amenities (among other features) support a more nuanced understanding of the relationship between property and STR pricing? Are these stronger determinants on pricing than the features explored in this study?

>> **Alignment with Big Data Analytics:** Integration of a new feature set into modelling centered on public amenities and attractions with respect to their economic impacts and visitation rates.

Temporal analysis of data pre- and post-STR arrival: How has the relationship between STR density and property pricing evolved over time? What longitudinal housing trends existed prior to the widespread arrival of STR following the 2008 founding of Airbnb, and how have they changed in response? Are these trends more significant predictors of housing price outcomes than a strong STR sector?

>> **Alignment with Big Data Analytics:** Expansion of pre-existing property price dataset to include pre-2008 housing price data and supplemental research to identify other datasets to assist in the identification of broader influencing trends. As with other recommendations, this would see the expansion of the project dataset to include both more instances and more features.

Comparative study of municipal experiences with STR and their respective policy responses: How has the STR sector shaped the housing market of different "global cities"? As with London, these are cities that are the cultural, economic, and/or political capitals of the respective high-income countries. Similarly, how has the STR sector shaped the housing market of different cities in the UK? Are there universal trends or do the relationships between housing market performance and the STR sector vary based on distinct localized economies, cultural preferences, and regulatory environments? In this sense, is London a singular case or reflective of the exertion of a broader spectrum of influences?

>> **Alignment with Big Data Analytics:** Parallel application of data-driven study process to other identified jurisdictions (i.e. Frankfurt, Paris, New York, Shanghai, Tokyo or Birmingham, Edinburgh, Glasgow, Manchester), subject to availability of comparable data and specific knowledge of local domains and languages.

Borough-level social and demographic analysis of the impacts of STR: How has the STR sector shaped the broader social and demographic contexts of the boroughs under review. How do characteristics of the local population, such as income, age distribution, or employment, interact with the STR sector and its influence on housing? Has community sentiment on the presence of STR in their neighbourhoods changed over time?

>> **Alignment with Big Data Analytics:** Introduction of new confounding variables drawn from publicly-available population census data to reflect the communities that exist within each subject borough. Inclusion of sentiment analysis from public-facing social media drawn from geocoded messages posted by residents of said boroughs.

We also consider a more general future direction that reflects the potential present in the selected datasets and study focus:

Augmenting data or focus: Given the experience of conducting this study, how might we augment the selected data or the study focus? As the review of STR amenities and their relationship to STR pricing demonstrated, there remains room to draw on existing-but-underused features in the dataset (such as total floor area) were we to be presented with a more robust variant of the dataset. Similarly, we note in the **Modelling** section that the development of predictive models could be helped by broadening the geographic or temporal scale of analysis to integrate more data into model development, and in doing so, improving predictive ability.

Alternatively given the exact same project data, what other research tasks could be considered and ultimately acted upon? There is a significant amount of environmental performance data present in the consolidated property price/residential environmental performance dataset that went unused as it was not needed to complete the study as it was defined. This data could be easily repurposed to model environmental performance of different housing forms and eras of construction in the UK capital completely separate from the STR focus we have taken in this study.

References

All references accessed 23 November 2023 unless otherwise specified.

Akanni, L., Lenhart, O., and Morton, A. "Income trajectories and self-rated health status in the UK." *SSM — Population Health* 17 (2022).

Deboosere, R., Kerrigan, D.J., Wachsmuth, D., and El-Geneidy, A. "Location, location, and professionalization: A multilevel hedonic analysis of Airbnb listing prices and revenue." *Regional Studies, Regional Science* 6.1 (2019): 143–156.

Greater London Authority (GLA). Housing and Land. "Housing Research Note 2020/04: Short-term and holiday letting in London," Policy report, authored by G. Cosh (2020).

Hati, S.R.H; Balgiah, T.E., Hananto, A., and Yuliati, E. "A decade of systematic literature review on Airbnb: The sharing economy from a multiple stakeholder perspective." *Heliyon* 7.10 (2021).

Koster, H.R.A., van Ommeren, J., and Volkhausen, N. "Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles." *Journal of Urban Economics* 124 (2021).

Local Government Association (UK). "Social determinants of health and the role of local government." Research report. July 2020.

Mayor of London — Greater London Assembly. "London Housing Strategy". May 2018.

Shabrina, Z., Arcaute, E., and Batty, M. "Airbnb and its potential impact on the London housing market." *Urban Studies* 1–25 (2021).

Temperton, J. "Airbnb has devoured London — and here is the data that proves it." *Wired* (13 February 2020).

Wachsmuth, D. "Short-term rentals in Los Angeles: Are the City's regulations working?" Report prepared for Better Neighbors LA (2021).

United Kingdom (UK). Ministry of Housing, Communities and Local Government. "The English Indices of Deprivation 2019: Statistical Release." (2019a).

UK. Ministry of Housing, Communities and Local Government. "The English Indices of Deprivation 2019: Research Report." (2019b).

Data Science | Housing | London | Urban Planning

Unlisted

Written by jsrobson

Edit profile

0 Followers