1. Introduction

The London housing market is one of the complex and dynamic system that is influenced by numerous factors. The city's popularity as a global hub for business and culture has driven demand for property, which is characterized by high levels of volatility, with prices fluctuating rapidly in response to changes in economic conditions, government policies, and local market conditions. As such, predicting housing prices accurately can be a challenging task.

Spatial analysis is an essential tool in investigating the relationships between spatial features and variables of interest. It is widely used in fields such as urban planning, environmental science, and epidemiology. Given London's status as one of the world's most sought-after locations, understanding the geographic distribution of house prices and the factors that affect them is critical. Spatial analysis can reveal patterns and relationships that may not be visible through traditional statistical methods, which is vital for our project.

In addition to spatial analysis, machine learning models can also provide valuable insights into the London housing market. By leveraging the power of algorithms and statistical models, these techniques can be used to predict future housing prices and identify key drivers of price fluctuations. This information can be used by investors and policymakers to make informed decisions about real estate investments and policy interventions. Overall, the combination of spatial analysis and machine learning can provide a powerful toolkit for analyzing the complex and dynamic London housing market. The machine learning models we have chosen in this project include Gradient Boosting Regressor (GB), Random Forest, K Nearest Neighbors (KNN), and Lasso.

Our project aims to provide valuable insights into the current state of the London housing market by utilizing machine learning models and spatial analysis techniques. Specifically, we will be implementing four machine learning models and evaluating their performance to choose the best model for predicting house prices across the London boroughs. We will also perform spatial analysis using Exploratory Spatial Data Analysis (ESDA), including spatial autocorrelation, to better understand the patterns and relationships within the data. Furthermore, we will also perform a time series visualization to gain deeper understanding of the changes in house prices over time. In general, we hope to contribute to the existing literature by providing a comprehensive analysis of house prices across different London boroughs.

The main analysis part is organized as follows. In 3.1-3.2 we perform data manipulation, preparation to conduct an exploratory analysis. This includes generating choropleth maps to visualize price trends and other important feature values. In section 3.3, we will conduct spatial autocorrelation analysis and visualization, including global and local measures. Spatial autocorrelation is a statistical technique that measures the degree to which data values are correlated based on their spatial weights. In 3.4, we have Point Pattern Analysis through Interactive Points clustering mapping. In 3.5, we focus on Machine learning, fits, evaluate and compare 4 machine learning models. Finally, we wrap up our project with conclusion.

2.Resarch Method

Data Processing:

The data we employed for this project included house price data, socio-economic data, and borough boundary data for London. The primary data source is London House Price Data, which can be downloaded from Price Paid Data (PPD).

We implement the project analysis in Python language for all data manipulation, geo-visualization, geospatial feature engineering and machine learning. Many libraries were involved including matplotlib, PySAL, Geopandas, Seaborn, ESDA, Sklearn, etc. We also utilized Mapbox for an interactive clustering map of house price points.

Spatial Analysis Methods:

With code in spatial join, merge, align crs system, aggregation etc., we created a new dataset 'lb2' ready for spatial mapping. Regarding spatial autocorrelation, we are going to calculate Moran's I for overall clustering and use Local Indicators of Spatial Association
(LISA) to analysis local autocorrelation and plot. Moran's I ranges from -1 to 1, corresponding to correlation. Global Moran's I is defined as below where N is the number of spatial units indexed by i and j, W is the sum of wij.

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Machine Learning and evaluation:

Machine learning is a method of data analysis that allows computers to learn pattern and predict without being explicitly programmed. We consider four crucial features for price prediction here: the floor area, number of rooms, property type, and old-new status. Our project analyzed the performance of four highly cited and popular supervised regression models, namely Gradient Boosting Regressor (GB), Random Forest, K Nearest Neighbors (KNN), and Lasso.

GB (Gradient Boosting Regressor) and RF (random forest regressor) are more advanced and complex models that belongs to Ensemble learning algorithms, which combine multiple models to improve their performance. Considering geolocation feature in our data, it's reasonable to have KNN regressor, which uses 'feature similarity' to predict the values of any new data points. Lasso linear model is a regularization technique. By selecting popular models from different categories and complexity levels, we can compare and evaluate their performance. We split the data into training (70%) and test (20%), and train model on training dataset with GridSearchCV.

We plan to use three evaluation metrics for regression models: R-squared or coefficient of determination, MAE or mean absolute error, and MSE or mean squared error. Each metric captures different aspects of model performance, and we will compare multiple metrics to provide a more comprehensive evaluation. R-squared will be the main evaluation metric, as it is an intuitive way to compare and explain the performance of the models in general. Therefore, we can obtain a more complete understanding of how well each model is performing and choose the best one for predicting house prices across London's boroughs.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total Variance}}$$

Part 3: Main Analysis with Code and Output

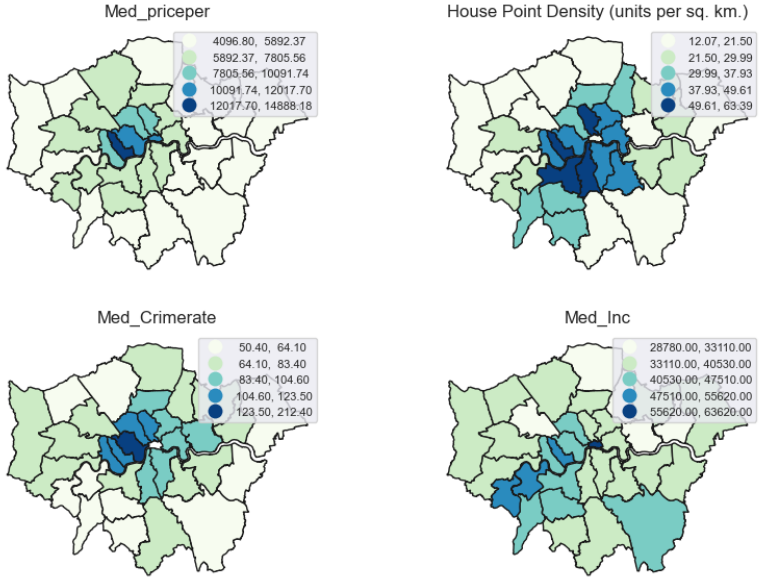
3.1 Data Preparation for Visualization

3.2 Choropleth mappings

We performed spatial join to aggregate the house points to London Polygon for geo-visualization.

1. We choose to use 'Med_priceper' because without the influence of floor area, price per square foot provides better representation of price-level.
2. Different classification schemes can yield varying results. For our choropleth mapping, we have chosen to use the Fisher-Jenks approach. Fisher-Jenks utilizes dynamic programming to group similar values into distinct classes or clusters(k).

Observations：
We see a clear trend of high prices, high density, and high crime rates in the city center, gradually decreasing as we move outward. Median income, however, exhibits a different pattern. Of particular interest is the polygon in the bottom right corner, which displays higher median income, lower prices, and crime rates, as well as low house density, making it an attractive location for living.


Med_priceper

House Point Density (units per sq. km.)

Med_Crimerate

Med_Inc

Figure 1: Choropleth mapping.

3.3 Spatial Autocorrelation

Is there a spatial clustering on price? - Calculate Moran's I

Here, we use K-nearest neighbors (KNN) weights to define the neighbors for target feature. We set k=4, meaning we count the four nearest neighbors of each feature when calculating the spatial weights.

Then calculate the spatial lag of our variable -house price, which is the average value of the variable for the feature's neighbors, weighted by the spatial weights. Then, we row-standardize the data for adjusting the weights in a spatial weights matrix, which ensures that each feature's contribution to the spatial autocorrelation measure is equal regardless of the number of neighbors it has.

If the prices are randomly distributed, there should not be clustering of similar values on the map below. We proceed with Spatial Autocorrelation which is kind of lack of spatial randomness after this part.

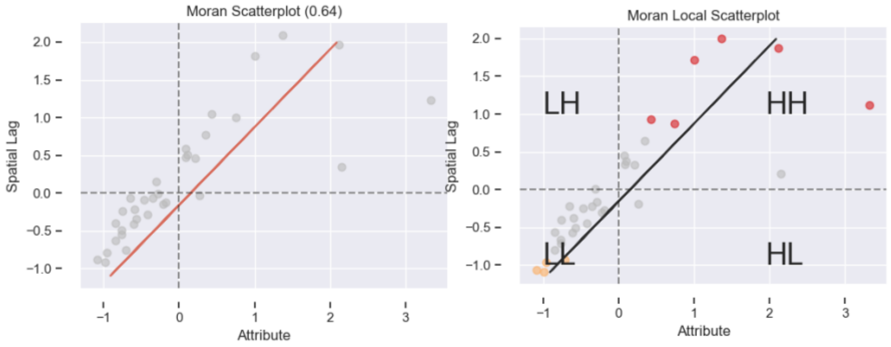


Figure 2: Spatial Autocorrelation.

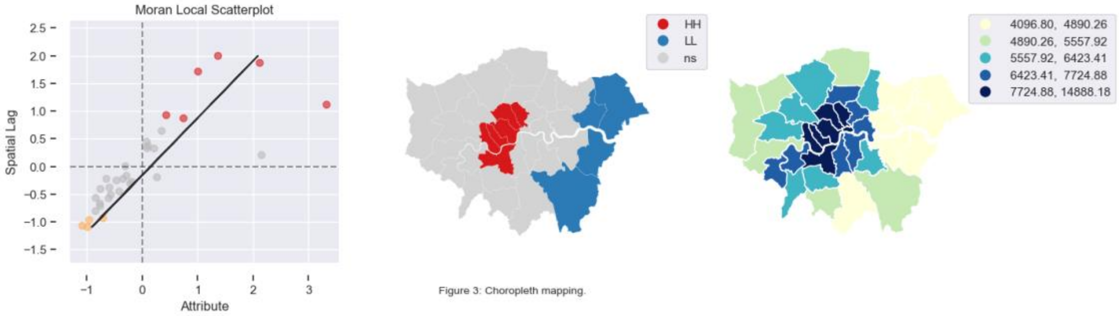


Figure 3: Choropleth mapping.

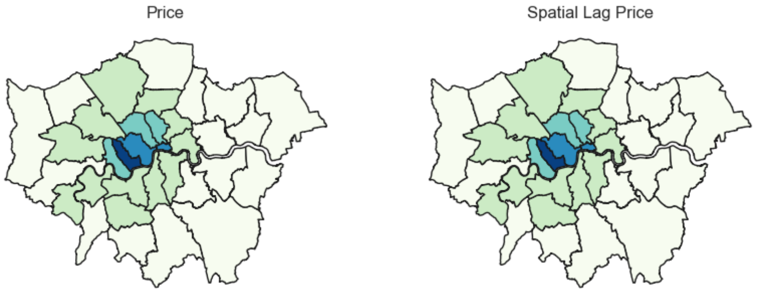
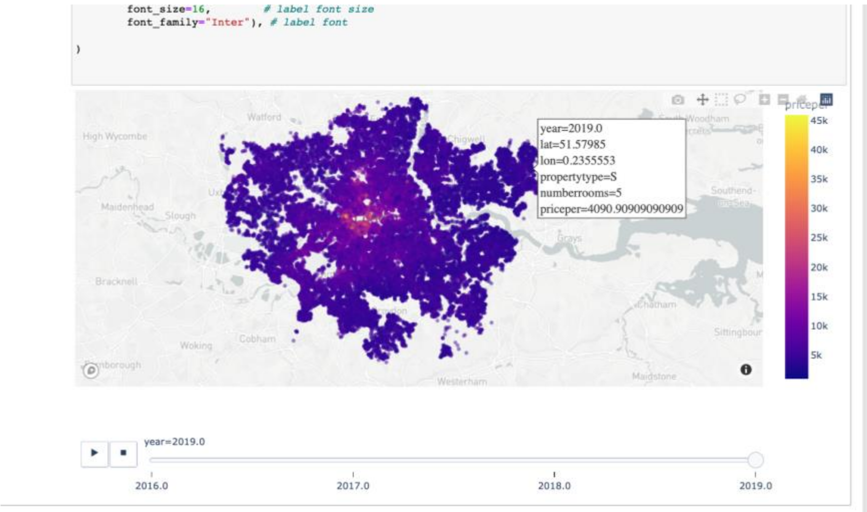


Figure 4: Compare Price and Lag_Price.

3.4 Interactive Points clustering on Map

Since our main target is house price. Here look at a time series of house price (per square foot) for all regions where data is available using an interactive points map. We map the points from 2016-2019 in our data so we can observe the changes.

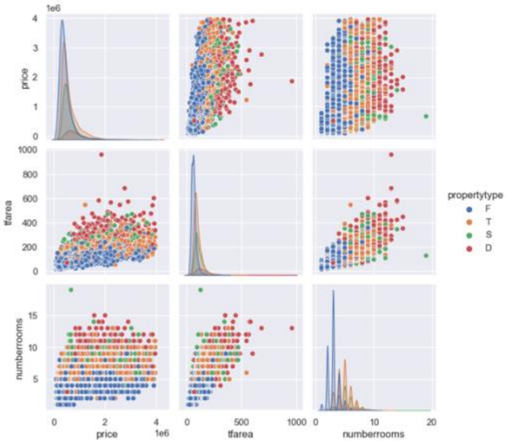


```
        font_size=16,         # label font size
        font_family="Inter"), # label font
)
```

```
year=2019.0
lat=51.57985
lon=0.2355553
propertytype=S
numberrooms=5
priceper=4090.90909090909
```

3.4 ML

Prior to utilizing machine learning algorithms, we generate a pairplot and correlation matrix to gain insight into the relationships between the variables in the dataset.



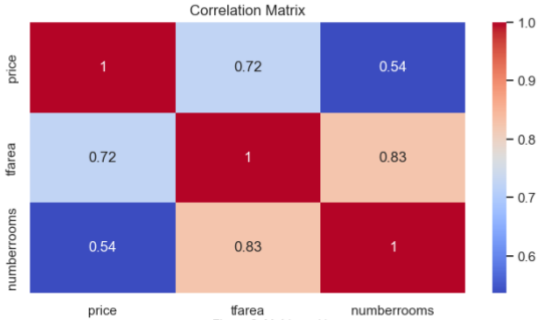
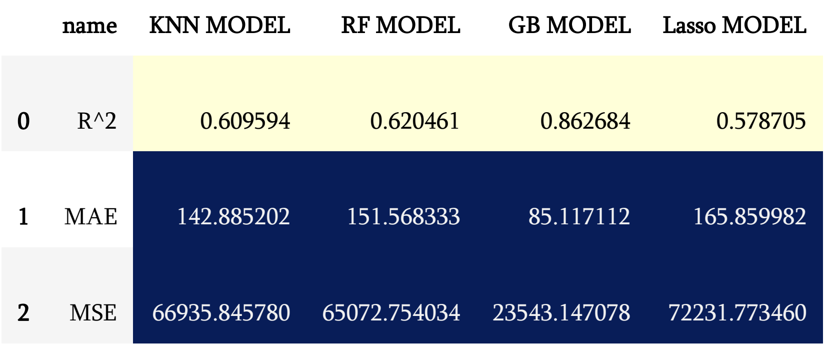


Figure 6: Matrix and heapmap

Table 1: ML Evaluation Metrics



	name	KNN MODEL	RF MODEL	GB MODEL	Lasso MODEL
0	R^2	0.609594	0.620461	0.862684	0.578705
1	MAE	142.885202	151.568333	85.117112	165.859982
2	MSE	66935.845780	65072.754034	23543.147078	72231.773460

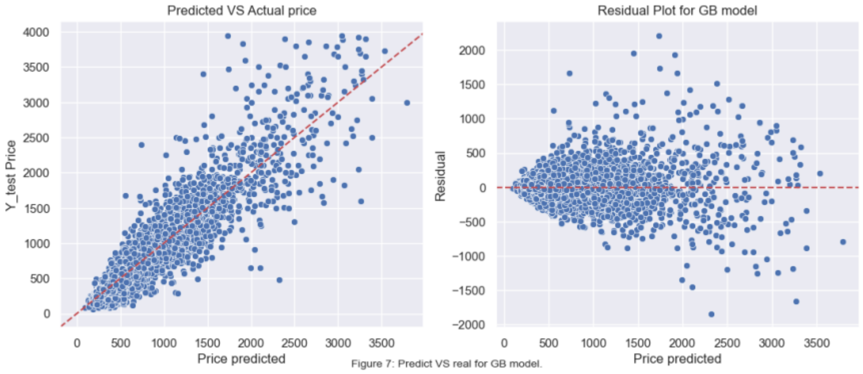


Figure 7: Predict VS real for GB model.

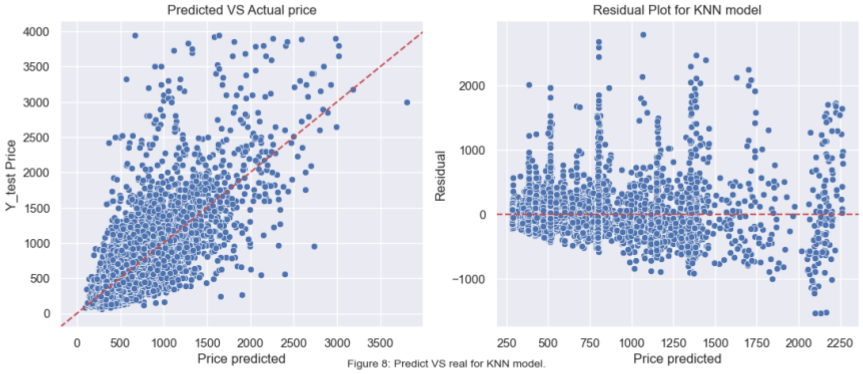


Figure 8: Predict VS real for KNN model.

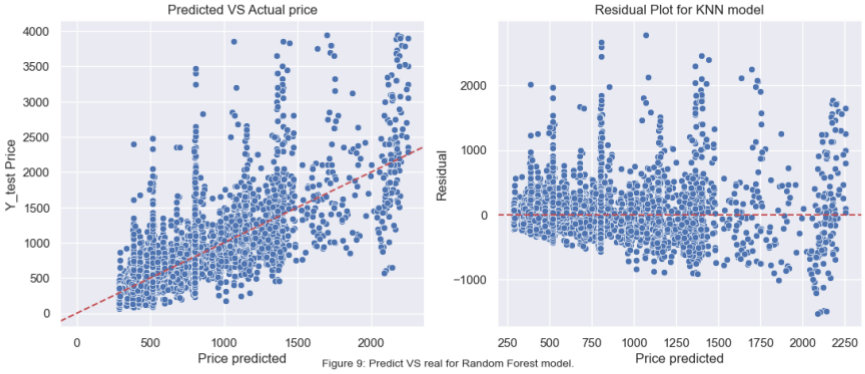


Figure 9: Predict VS real for Random Forest model.

Part4: Description of results and significance

From choropleth mapping in Figure 1, we observed a clear trend of high prices, high density, and high crime rates in the centre of the city, with these values gradually decreasing as we move outward. Median income, however, exhibits a different pattern. Of particular interest is the polygon in the bottom right corner, which displays higher median income, lower prices, and crime rates, as well as low house density, making it an attractive location for living.

Then we proceed with Spatial Autocorrelation analysis to investigate further spatial pattern. The calculation of Moran's I value for the house price data shows a moderate positive spatial autocorrelation, with a value of approximately 0.642. Additionally, Local Indicators of Spatial Association (LISAs) tell us whether the local association between polygons and its neighbors is positive/similar (HH/LL) or negative/ dissimilar (HL/LH). This suggests a spatial clustering of similar house prices in the study area. From Figure 2, we can see most of them are in HH and LL area. As shown in Figure 3 more specifically, there are about 11 locations significant when setting p=0.05.

The features we have chosen for house price prediction with machine learning includes total floor area, number of rooms, property type, and old-new status. To do this, we utilized four different supervised machine learning regression models, including GB Model, Random Forest, KNN, and LASSO, and compared their evaluation metrics. During machine learning process, we used GridSearch-CV to optimize hyperparameters in each model. The evaluation results of each of the four machine learning models on the test set are obtained by specifying three evaluation metrics, R2, MAE, and MSE, under the optimal parameters selected based on grid search.

Table 1 presents the ML evaluation metrics, allowing us to easily discern the performance of the various models. It is apparent that the GB Model outperformed the others, achieving an accuracy score (R^2) of 0.86, followed by Random Forest at 0.62, KNN at 0.6, and LASSO at 0.57. This indicates that the GB Model was the most effective at predicting house prices based on the selected features for our data. In figure 7-9, we compared price predicted and real price in test data and calculated residuals for three main models. Residual values (Observed – Predicted) can capture the difference. It's evident that GB model is performing very well for most of the regular data, except for data points with very high house prices. I assume those points are influenced more by social and environmental factors, because in this project here we only have features about location and house itself. That's what we are going to do next.

For further analysis we can use hedonic home price model to better understand the factors that influence housing prices, including both internal characteristics of the house and external neighbourhood amenities and public services, such as crime exposure. Our findings suggest that spatial pattern plays an important role in housing prices, with systematic spatial patterns emerging from individual real estate transactions.

Overall, the results of our study have important implications for understanding the factors that influence house prices in our study area. By identifying dynamic features that are associated with house prices, we aim to help decision making relevant with housing affordability, investment, and urban planning. Furthermore, our findings highlight the importance of considering spatial autocorrelation in house price data, as this can help identify areas that are particularly desirable or undesirable for living based on the characteristics of their neighbouring regions.