Analysing the property market for different property types in London.

Muhammad Sufyaan Shaikh
Department of Computer Science
City, University of London
London, England
180021953

***Abstract*: This research is about the property market as it explores the different property types and their impact on property prices and area size. I conducted geospatial analysis to examine if location influences property prices, property types, and area size within London, as well as assess the dataset's ability to accurately identify a property based on features provided.**

1. Introduction:

My project focuses on the London property market due to the intiguing nature of property prices constantly increasing, as it's likely to be the biggest financial transaction and commitment most individuals would ever make. That's because housing price in the UK, particularly in London have instaneously increased since 1970 (adjusted for inflation), with "house prices to income ratios more than doubling since the mid-1990's" as London is the most impacted city in the UK (Pettinger, 2021) [1]. It is getting harder to purchace any property in London, with property prices increasing. This could be due to an increase in income, amount of money allowed to loan (low interest rates) and population. These conditions have stimulated competition, resulting in a stronger demand for a limited supply of properties, causing property valuations to rise.

Futhermore, a buyer's demands are tailored to people's needs and wants, for instance the impact of coronavirus caused people to work at home, which resulted in an increased activity for a house with a garden and extra study, but once society returned to normality, there was a shift in demand from houses to flats (Osborne, 2021)[2]. Demenstroting, how any given event can influence property selection which includes number of rooms, area size of property and it's location. The extent to which property prices are increasing is detrimental to people wanting to buy a property at this time, to combat this the goverment have started projects to build new building (new developments) to improve on pre-existing properties (flats/houses) at affordable prices.

London is the reigon most affected by the volatility in property prices through it's global attraction and being a major financial centre, such that it is shown that you can predict property prices from New york property prices (Holly, Pesaran and Yamagata, 2010) [3]. Thus, showing valuation in property prices being impacted by location.

2. Analytical questions and data:

I aim to analyse the key characteristics of property types (Flat, House, and new development) using the features of the dataset such as price (£) and area size in square meters. I attempt to see whether we can see the difference between property type in terms of price and area size, in the hope to see how property prices is evaluated. I also attempt to see if geographical location has an impact on properties within London as we already know that geographical location does have an impact in the instance of New York being able to predict property prices in London and London being able to predict property prices in other regions within the UK from the works of (Holly, Pesaran and Yamagata, 2010) [3].

My analytical questions are:

1. What type of relationship (correlation) does the features have with property price?
2. Is there any difference in property types (Flat/ House/ New development) in terms of price and area? If yes, what are the differences for the property types?
3. How well can the dataset predict/classify a property type given the features (using a machine learning algorithm)?
4. Which property types give the most and least value for your money?

5. Does location influence property build, price, and area size with reference to London?

3. DATA (MATERIALS):

I intend to investigate the differences between property types like flats and houses. I will attempt to answer these queries using two datasets from Kaggle, a housing dataset called "House Prices in London" which contains features about homes including house type, area size in square meters and numbers of bedrooms. I used a postcode dataset which shows the location of each property in the UK, it has features including postcode area, postcode, longitude, and latitude, which is used to investigate the geospatial aspects of the housing market.

These two datasets were merged as the housing dataset didn't have verified observations of addresses for the property within London, whereas the postcode dataset did have verified addresses for all properties in the UK given that dataset has a UK open government licence. When both datasets are merged using the inner join, the false addresses/observations are removed, the postcode dataset also add longitude and latitude which will be used for geospatial analysis.

The merged dataset has 3395 rows with 10 columns after dropping unwanted columns from both housing and postcode dataset due to them being arbitrary or having unreliable observations from applying domain knowledge and adding a new derived feature. Figure 1 shows the count of different property types in London produced from the variable House Type which is a constant attribute used throughout the analysis stage, this includes assessing the difference in property types in terms of price and area size, the variable also acts as a response variable for a machine learning model. I use longitude and latitude variables to visualise the location of homes in the dataset. One assumption that I will be is that all property is in London, even though some locations may just be outside of London due to the postcode area ("direction" variable) shows that certain boroughs/towns overlap in and out of London.

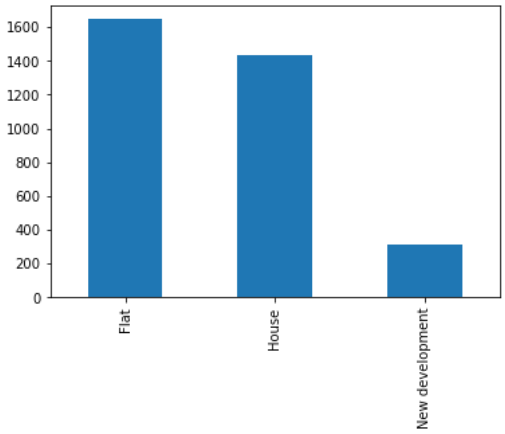


Figure 1: Bar plot for the counts of property type (before outlier removal)

4. ANALYSIS

4.1 Data Preparation:

The first step I took was merging the housing dataset with the postcode dataset (after importing both datasets), the dataset transformation from the housing dataset was rows reduced from 3480 to 3395 observations, meaning that there were 85 false addresses and dataset increased from 11 columns to 27 columns, however I had to remove certain columns because they weren't needed for the analysis such as country, different postcode formats and property names etc. This included removing number of bathrooms and receptions as I suspect that the values were copied from the number of bedroom column since all columns had the same value for every row (checked through inspection), for example: it's unlikely for number of bedrooms to have 6 bathrooms and receptions but it is normal to have 6 bedrooms in London.

After removal of columns, my merged dataset currently had 10 columns, however I dropped a further column which was the Location column due to the column having all the missing values in the dataset, which was 28% of the observations, to many to replace/remove and this variable is arbitrary since variables such as longitude and latitude can give me the same information. Therefore, my cleaned merged dataset as stated in the previous section is 3395 rows with 10 columns after adding a new variable through data derivation: area in sqm. I made sure each variable had the correct datatypes.

4.2.1 Data Derivation:

I converted the variable Area in sqft to Area in sqm, via converting the unit for property area size, from square feet to square metres, because square metres unit more recognised. Furthermore, I derived a new column Price per sqm, calculated through dividing the columns Price by Area in sqm. This derived column can be used to answer the question of which property has the best value for money.

4.2.2 Explanatory data analysis (EDA) with data derivation:

I plotted categorical data by first checking the unique values for variable House type and found 8 different classes of property type, Which I converted to 3 predefined classes of property type: 'Flat', 'House', and 'New development'. By first simplifying the class value 'Flat/Apartment' to 'Flat', as flat is the most common term used in Britain. Then I converted {'Penthouse', 'studio'} to 'Flat', 'Bungalow' to 'House' and {'duplex', 'mew'} to 'New development'. These extra classes in the original

dataset were converted to the 3 predefined property type to simplify the classification case. For the direction variable, I converted postcode area to general directions with reference to London: 'North', 'West', 'East', 'Central', 'South', 'North-West', 'North-East', 'South-West' and 'South-East'; some locations may be outside of London. Finally, I used the Area in sqm variable to define whether a property is classed as 'small', 'medium' or 'large' using a for-1f conditional statement iterating through the rows: 'small' is defined if your property is less than 120 square metres, 'medium' is defined if your property is between 120 and 200 square metres and large defined when your property is larger than 200 square metres. These values were picked based upon the quantiles for the variable associated.

Figure 2: Count of the categorical variables after dealing with the outliers

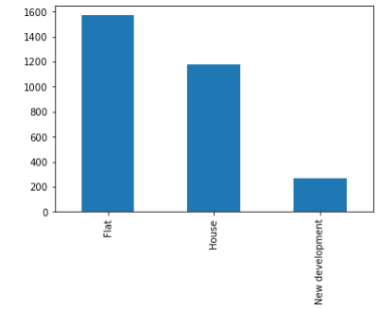


Figure 2 (A) Property types

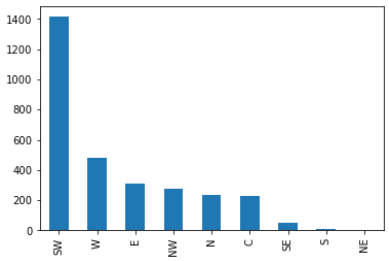


Figure 2 (B) Direction with reference to London.

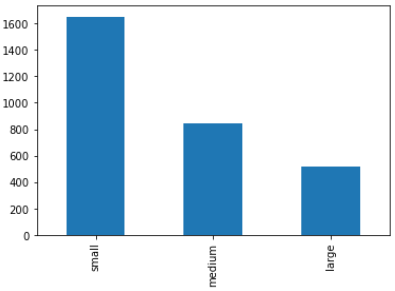


Figure 2 (C) Size of property determined by size of property in square metres

4.2.3 Explanatory data analysis (EDA) detecting outliers through boxplots:

I detected outliers using boxplots and removed the outliers instead of replacing except for the number of bedrooms column, because I suspect the extreme values to be true as property prices in London are volatile and random. I assume removing outliers in Price and area in sqm columns will also remove outliers for Price per sqm column since it's derived from those variables. Outliers are removed to improve the fit of the data and remove any values that could be an error, this resulted in my dataset reducing from 3395 rows to 3013 rows. I changed the minimum number of bedrooms from 0 to 1 since it made no sense for a property to have no bedroom.

Figure 3: Boxplots after dealing with the outliers and grouped by property type for the numerical variables

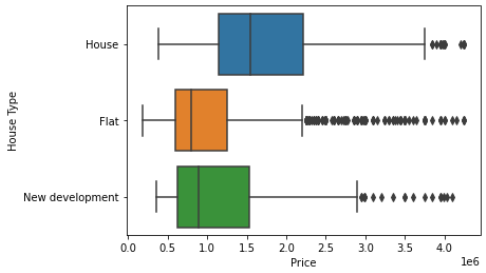


Figure 3 (A) Property price

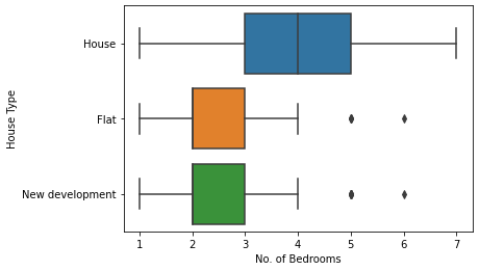


Figure 3 (B) Number of bedrooms

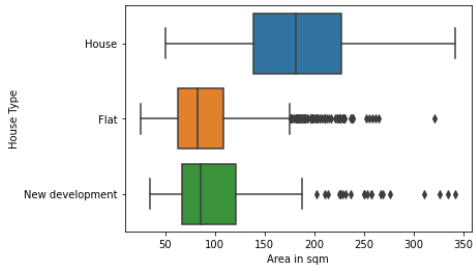


Figure 3 (C) Property area size

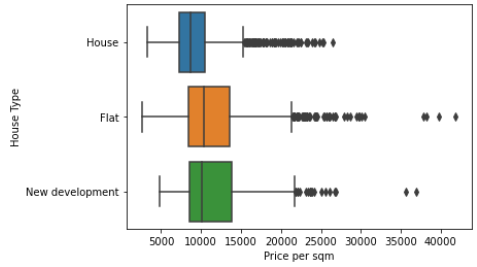


Figure 3 (D) Price per square metres

4.2.4 Geospatial Analysis – plotting geographical maps:

I plotted a London map embedded with points of my data frame, using a shapefile from a government file website. Longitude and latitude column from the dataset was used to plot the points of location for each property onto the London map. I used coordinate reference system (CRS) to make sure that the London map and points have the same reference and scalability such that the points will be able to be plotted onto the map. I had to create a geometry point from the longitude and latitude columns where the datatypes were floats, using this I created a geodata frame: all the features from dataset were added with a geometry which has the same CRS so points from dataset can be mapped on the map. I then plotted multiple maps where points were coloured by house type, price, and area in sqm columns. These plots can state if location within London has an impact on property prices and verify counts of the direction variable.

Figure 4: Geospatial plot of London with markers plotted of the properties coloured by property type.

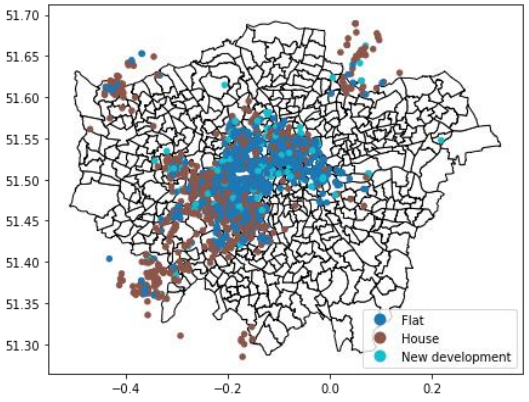


Figure 5: Geospatial plot of London with markers plotted of the properties coloured by property price.

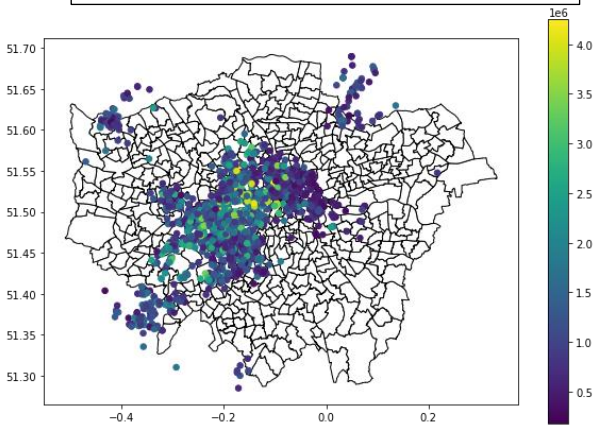
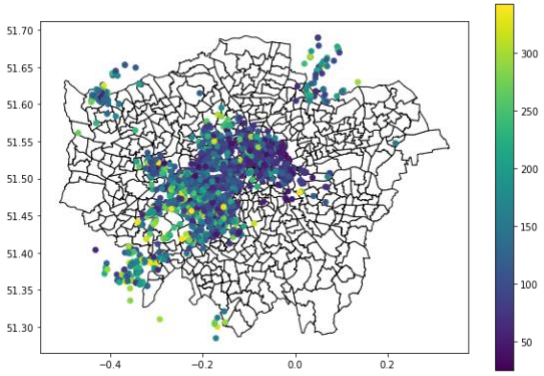
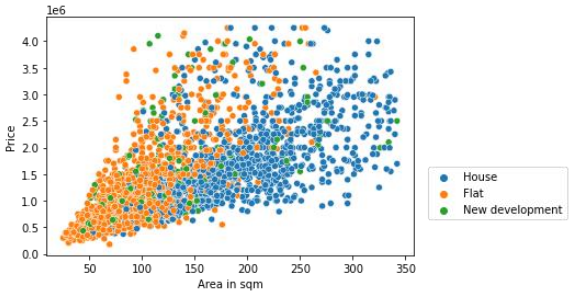


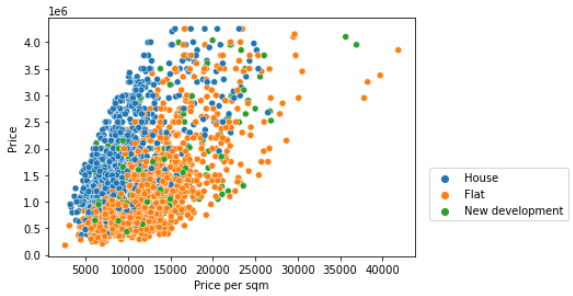
Figure 6: Geospatial plot of London with markers plotted of the properties coloured by area size.



4.2.5 Explanatory data analysis (EDA) – Scatter plots and correlation matrix:

I plotted scatter plots and a heatmap for corelation plot to see the relationship between each numerical features and property prices.





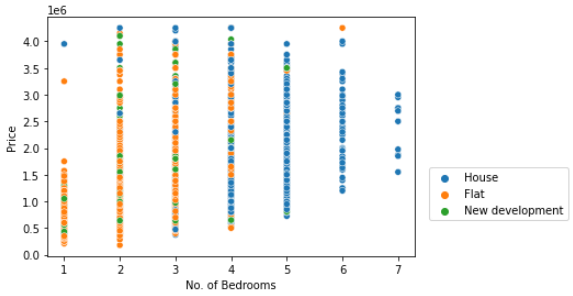


Figure 7: Scatter plots between numerical features and price.

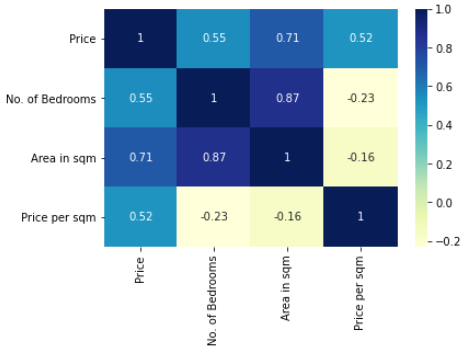


Figure 8: Corelation heatmap plot of the numerical features

4.3 Construction of models – Logistic regression:

I used logistic regression (machine learning algorithm model) to classify the property type of the dataset, making the response variable House type. I drop the property type 'New development' as it had a small count of properties compared to the other property types, thus I converted the model from a multinomial into a binary classification model. New development properties would have a smaller count than flat and house due to the latter being more established property types. Next, I dropped most location columns as they're not needed for the model. I checked to see the frequency of the class 'Flat' and 'House' from the response variable, showed in figure 9.

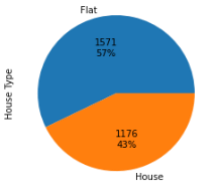


Figure 9: Pie chart of the counts and percentage for the property types.

From figure 9, we can see that flat (57%) has slightly more class frequency than House (43%), meaning that the model has a slight imbalance problem but only slightly that we don't have to rebalance. I converted the response variable classes binary value Flat = 1 and House = 0, and I changed the categorical variables into dummy variables, this is to fit the model. From this I concatenated the dummy variables with the dataset and dropped the categorical variables. This dataset has 70:30 split into training and test set, as logistic regression model was fitted using the training data, I used p-values to remove features which was less than 0.05 and significant, this was done until model was optimal through feature selection as I predicted the labels for test set.

4.4 Validation of results:

I validated the results of my model through confusion matrix plot, calculation of performance metric like accuracy and a ROC plot with AUC value. I also validated my results from cross-referencing the different graphs as they are interlinked. I further compared with the works of others from articles in terms of the general trends they find in evaluating a property price in London.

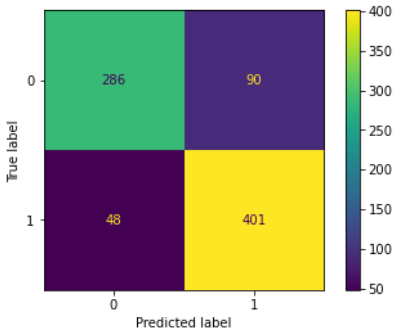


Figure 10 (a): . Confusion chart for classification for classification model

	precision	recall	f1-score
0	0.86	0.76	0.81
1	0.82	0.89	0.85

Figure 10 (b): performance metrics for classification model

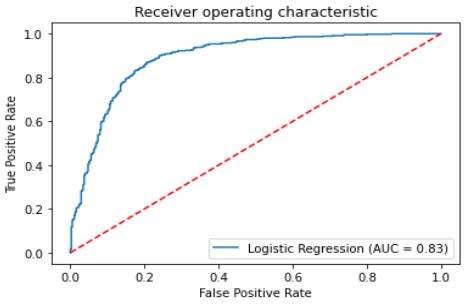


Figure 10 (c): ROC plot and AUC value for the classification model

5. Findings, Reflections & Future work:

5.1 Findings:

From figure 3 (boxplot), we answer if property types have any difference in price and area. Where houses have the highest and most stable property prices through having fewer extreme values, whilst flat has the lowest price and most extreme values, with 'New development' having a slightly higher valuation than flat as it's built to improve the living standards from the more established property types: flats and houses.

Houses have a higher price than flat and new development, this could be explained by houses having a higher possible range of bedrooms (from 1 to 7 bedrooms) whereas both flat and new development have a range from (1 to 4 bedrooms) with extreme cases of having 5 and 6 bedrooms which is still lower than the maximum number of

bedrooms in houses. This trend is also true for area size in property sizes. Concluding that Price column has a strong corelation between Area in sqm, Price per sqm and No. of bedrooms columns, which was corroborated with the use of scatter plots (figure 7) and correlation heatmap plot (figure 8), this finding answered another analytical question of correlation between features and property price. I further found that there are differences in property type in terms of features through scatterplots and boxplots as these plots are grouped by house type column, its findings suggests that houses are more expensive and have more room but surprisingly it's the best value for money in terms of price per sqm column value being lower than flats and new developments, new developments have the worst value for money.

When looking at the geospatial plots (figure 4,5,6) we can see that the majority property build 'New development' and 'Flat' is in southwest & central London. From the same plots we can see that property prices increases as you go more into London, the proportion of houses tend to increase when compared to flat and new developments as you move out of London. The property prices also tend to increase in the inner parts of London (central) and decrease as you move in outer parts as areas like East, Northeast, Northwest London which have the lowest property price. Area size also increases as you move out of London and decreases when you move towards central London as this is the most desired location and so will get overcrowded by overpopulation. The plots show that location plays a major factor on price within London.

I used logistic regression model to test the accuracy where I found that. It is better at predicting flats than houses at 85% compared to 81% (F1-score) as ROC plot shows a good trade-off between the true positive rate and false positive rate with AUC value of 0.83. Accuracy can improve with further hyperparameter tuning and adding a validation set to further validate the results.

5.2 Reflections and future work:

My limitations are that there is not the same property count across the direction column as southwest London had half the count of properties such that the geospatial analysis is not a balanced analysis within London. I also could have had more features to do with property such as if it property had outdoor space, energy efficiency of property and how close property is to school, work, tube stations and the year property was sold. For future work I could have performed geographical weighted models to test the dynamics of a location to another and performed multiple linear regression to predict house prices.

Section	Word count
Abstract	**57**
Introduction	**306**
Analytical questions and data	**215**
Data (materials)	**307**
Analysis	**1109**
Findings	**573**
TOTAL WORD COUNT	**2,567**

References:

[1] (Pettinger, 2021)
Pettinger, T. (2021) *Why are UK house prices so high? - Economics Help*, *Economics Help*. Available at: https://www.economicshelp.org/blog/8733/housing/uk-house-prices-high/ (Accessed: 23 December 2021).

[2] (Osborne, 2021)
Pettinger, T. (2021) *Why are UK house prices so high? - Economics Help*, *Economics Help*. Available at: https://www.economicshelp.org/blog/8733/housing/uk-house-prices-high/ (Accessed: 23 December 2021).

[3] (Holly, Pesaran and Yamagata, 2010)
Holly, S., Pesaran, M. and Yamagata, T. (2010) "Spatial and Temporal Diffusion of House Prices in the UK", *SSRN Electronic Journal*. doi: 10.2139/ssrn.1545683.

Datasets uses:

[4] https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london

[5] https://www.kaggle.com/arnavkulkarni/housing-prices-in-london

[6] https://www.kaggle.com/danwinchester/open-postcode-geo