London House Price Analysis

Using Python

Prepared By: Anasua Sarkar

Introduction

The London housing market is dynamic, with prices influenced by factors like location, property size, and the number of bedrooms and bathrooms. Understanding these factors provides valuable insights for buyers, investors, and policymakers. In this project, I'll analyze price distributions, house types, area size, and bedroom/bathroom numbers to explore London house prices, aiming to uncover trends and predict prices based on key features.

Problem Statement:

The main objective of this project is to examine the factors that influence house prices in London and to pinpoint the key variables that impact high or low prices. Through exploratory data analysis (EDA), I aim to address the following important questions:

- What is the distribution of house prices in London, and are there any significant outliers?
- How do property size (area in square feet), the number of bedrooms, and other features relate to house prices?
- Are there any discernible trends that indicate which areas or property types command the highest or loIst prices?
- Can normalization and outlier detection enhance the quality of data for subsequent predictive modeling?

OVERVIEW OF THE ANALYSIS:

1. DATA COLLECTION:

The dataset used for this analysis contains house price information for various properties in London, including details such as the property name, price, house type, area, and number of bedrooms and bathrooms.

2. EXPLORATORY DATA ANALYSIS (EDA):

I begin by exploring the dataset through summary statistics and visualisations to understand the distribution of key features such as price, area in square feet, and the number of bedrooms.

I analyse the correlation BETWEEN these features and house prices to identify which variables have the most influence on pricing.

3. OUTLIER DETECTION:

To detect outliers I used boxplots and statistical methods, as well as analysing the values with the outliers

NOTE:

THIS PRESENTATION INCLUDES ONLY THE VISUALISATIONS (CHARTS AND GRAPHS) ALONG WITH EXPLANATIONS OF THE RESULTS. FOR A DETAILED VIEW OF THE CODE AND THE UNDERLYING ANALYSIS, PLEASE REFER TO THE ACCOMPANYING PDF VERSION OF THE PROJECT, WHICH CONTAINS ALL THE CODE AND ADDITIONAL DETAILS.

INTRODUCTION TO THE DATASET

The dataset used in this project contains information about 3,480 residential properties in London, with various attributes that describe each property's features and characteristics. The dataset aims to provide a comprehensive overview of the housing market in London, capturing key factors that influence property prices. The columns in the dataset are:

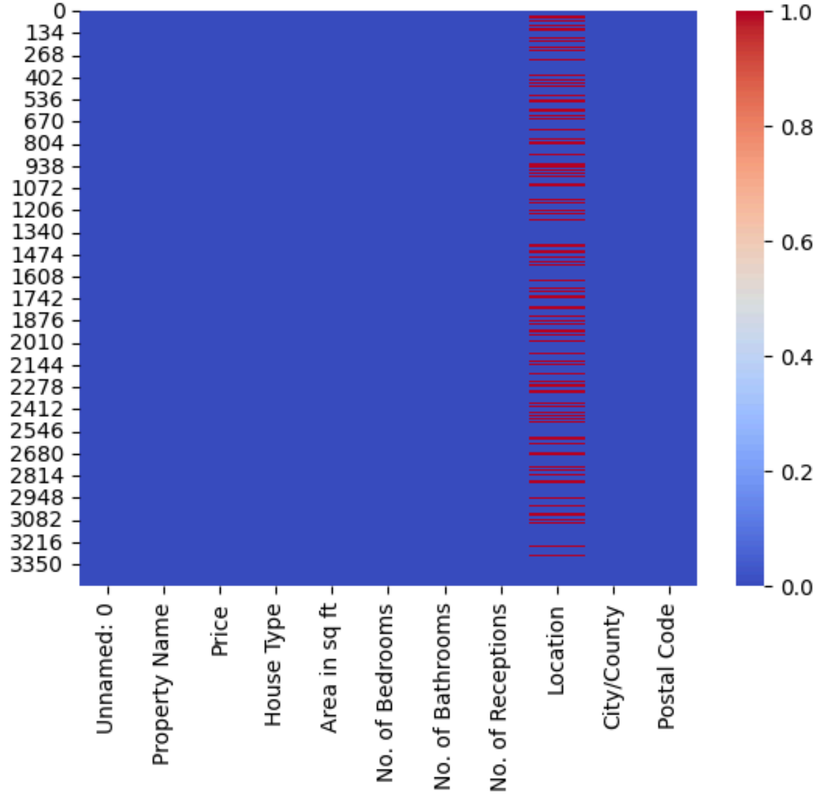
- **Unnamed:** Index column, which is a unique identifier for each property listing.
- **Property Name:** Name or description of the property, if available.
- **Price:** The listing price of the property.
- **House Type:** Type of the property (e.g., apartment, detached house, semidetached house, etc.).
- **Area in sq ft:** The total area of the property in square feet.

- **No. of Bedrooms:** Number of bedrooms in the property.
- **No. of Bathrooms:** Number of bathrooms in the property.
- **No. of Receptions:** Number of reception rooms (typically living or dining rooms).
- **Location:** The specific location (address or neighbourhood) of the property. This field contains missing values for some entries, as there are 2,518 valid entries.
- **City/County:** The city or county where the property is located (all properties are located within London or nearby areas).
- **Postal Code:** The postal code (ZIP code) of the property.

This dataset serves as the foundation for analysing house prices in London, exploring factors such as property size, location, and type to understand their impact on pricing.

Heatmap of Missing Values in London House Price Dataset

Findings:

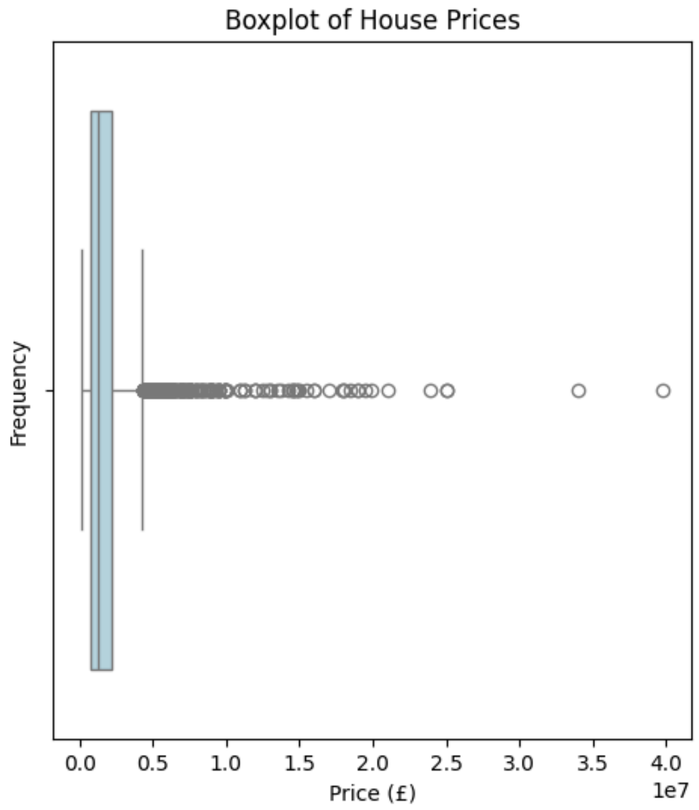


- **Location column contains missing values:** As per the heatmap, only the "Location" column shows missing data. The red and blue stripes in this column indicate the presence of both null and nonnull values.

- **No missing data in other columns:** All other columns, such as "Property Name," "Price," "House Type," "Area in sq ft," and so on, are fully populated, represented by the solid blue color across these variables.

- This heatmap confirms that the dataset is mostly complete except for some missing data in the "Location" column.

Boxplot of House Prices



In this box plot for house prices:

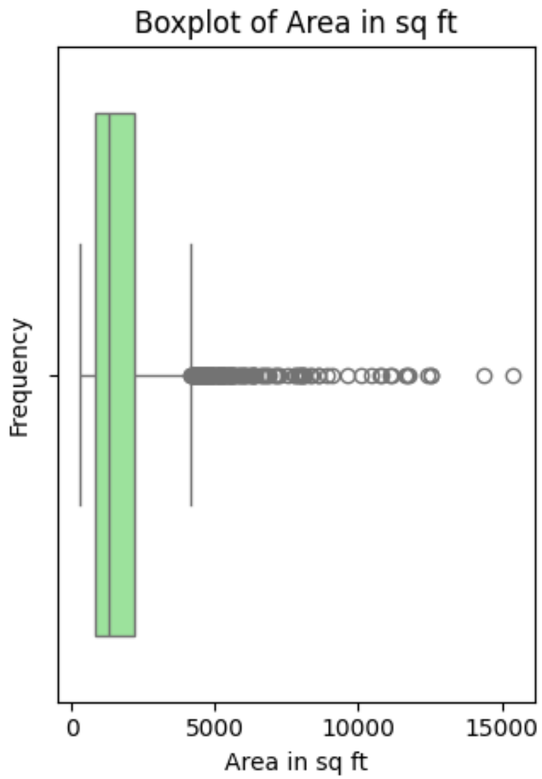
X-axis (Price): The house prices are represented on the x-axis in pounds (£). The scale shows a range up to around 40 million (£).

Outliers: There are a large number of outliers (represented by circles) that are located far from the main distribution of the data. These represent properties that are priced significantly higher than the majority of the dataset. In this case, anything beyond the whiskers (around the upper quartile) is considered an outlier.

Whiskers and the Box: The majority of the house prices fall within a very small range towards the left of the box plot. This indicates that most house prices are lower (likely clustered near the lower end). The box (between the first and third quartile) shows the inter quartile range (IQR), and the whiskers extend to the minimum and maximum non outlier values.

Right Skewness: The plot is highly rights-skewed, meaning a few properties are priced much higher than the majority. The presence of many outliers indicates that a few very expensive properties are pulling the mean price upwards.

Boxplot of Area in sq ft



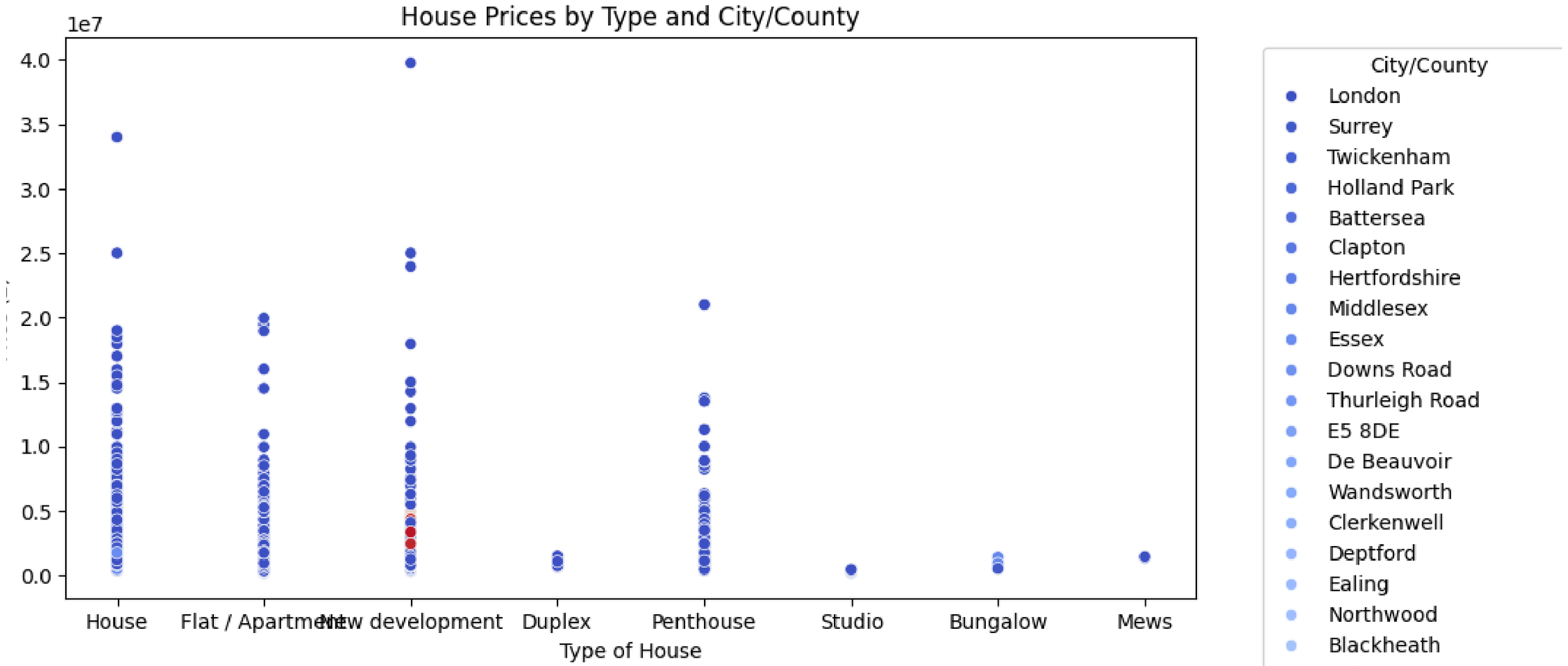
In this boxplot for the area in square feet:

X-axis (Area in sq ft): The area in square feet is plotted along the xaxis, with values ranging from 0 to 15,000 sq ft.

Outliers: Similar to the price boxplot, there are several outliers (represented by circles) for larger properties, which are significantly bigger than the majority. These are homes with a much larger area, pulling the right side of the plot.

Whiskers and the Box: Most properties have a relatively smaller area, as shown by the narrow box and the position of the whiskers (the minimum and maximum nonoutlier values). The box and whiskers cover a tight range, with most homes having an area far smaller than 15,000 sq ft.

Right Skewness: This boxplot is also right skewed, indicating a small number of large properties (high area) are outliers compared to the majority of homes.


House Prices by Type and City/County
1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
House
Flat / Apartment
New development
Duplex
Penthouse
Studio
Bungalow
Mews
Type of House
City/County
London
Surrey
Twickenham
Holland Park
Battersea
Clapton
Hertfordshire
Middlesex
Essex
Downs Road
Thurleigh Road
E5 8DE
De Beauvoir
Wandsworth
Clerkenwell
Deptford
Ealing
Northwood
Blackheath
Fulham
Richmond
The Metal Works
N7 6QX
City Of London
Chessington
82-88 Fulham High Stre
Little Venice
Spitalfields

Scatter plot of house prices by type of house across different parts of London.

X-Axis: Type of House

The x-axis lists various types of houses, including:
House, Flat/Apartment, New Development, Duplex, Penthouse, Studio, Bungalow, Mews

Y-Axis: Price (£)

The y-axis represents house prices in pounds (£), ranging from 0 up to £40 million (4×10^7).
Most prices seem to be concentrated below £10 million, but there are a few high value properties.

Data Points:

Each circle on the plot represents a house price for a specific type of house in a certain part of the city. The vertical spread of dots within each house type category indicates the range of prices for that particular type of house.

Legend (on the right):

The colours of the data points correspond to different areas, with each location having its own colour. Example locations include London, Surrey, Twickenham, Richmond, etc. Each location has a distinctive colour, helping to identify how house prices vary by region.

Observations:

1. Houses and Flats/Apartments have the most price points, suggesting they are the most common property types.

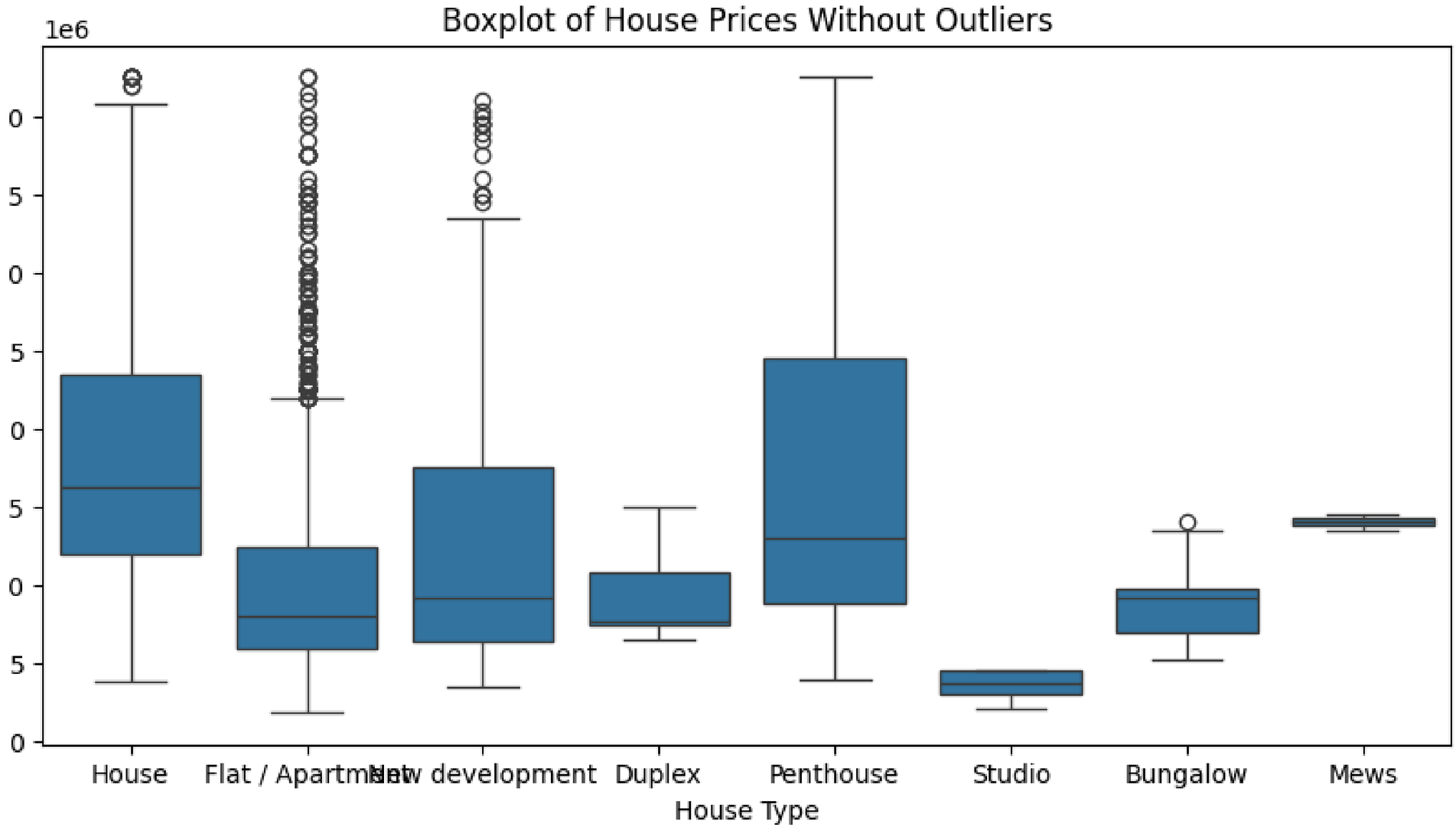
2. Penthouses have some of the highest prices, with a few exceeding £10 million.

3. New Developments also seem to have a wide price range, with a significant number of high value properties.

4. Other categories like Bungalows, Studios, and Mews show much lower price ranges, with fewer extreme values.

Key Takeaway:

This chart provides an overview of the price distribution of different types of housing across various parts of the city . We can see which areas have higher concentrations of expensive properties and how certain property types (like penthouses and new developments) tend to have higher prices compared to others (like studios and bungalows).


Boxplot of House Prices Without Outliers
1e6
0
5
0
5
0
5
0
5
0
House
Flat / Apartment
New development
Duplex
Penthouse
Studio
Bungalow
Mews
House Type

Comparison Between Boxplots (With and Without Outliers)

1. Impact of Outliers on the Distribution:

- **With outliers:** The price distribution is highly skewed due to the presence of extreme values. There are a significant number of outliers on the higher end, with some properties exceeding £10 million, and even going up to almost £40 million. These outliers stretch the plot horizontally, making it harder to analyse the majority of the data.

- **Without outliers (the previous plot):** By removing outliers, the boxplot shows a much clearer and more compact distribution. This helps focus on the central tendency and variability of property prices, making it easier to analyse trends for the typical range of house prices.

2. Median and Quartile Comparisons:

- **With outliers:** The boxplot shows a narrower box (the interquartile range, or IQR), suggesting that the majority of properties are priced within a lower range (below £1 million), but the overall visual is dominated by the extreme values on the higher end.

- **Without outliers:** The medians are clearer for each house type, and the IQR is wider, giving more insight into how prices vary for the bulk of properties. This helps to avoid the distortion that the extreme luxury properties cause.

3. Range of Prices:

- **With outliers:** The total range of house prices extends dramatically, from below £100,000 to close to £40 million. This extreme range distorts the visual representation, making it difficult to discern pricing trends for the majority of properties.
- **Without outliers:** The range is significantly narrower, focusing on properties that fall within a more typical range. This makes the price trends much more interpretable.

4. Price Anomalies:

- **With outliers:** The outliers likely represent luxury properties, which distort the overall view of the housing market. These outliers can be interesting to analyse separately but tend to obscure the general trends.

- **Without outliers:** By removing these anomalies, we can see a clearer picture of the pricing patterns for most properties, giving a better sense of what buyers can typically expect in the London housing market.

Key Findings from the Comparison:

- **Without outliers,** the boxplot is easier to interpret, focusing on the majority of properties that fall within the typical price range. This gives a more accurate picture of price trends and variation across house types.

- **With outliers,** the luxury property market skews the data, making it difficult to analyze the core trends. While it's valuable to acknowledge that such highpriced properties exist, they should be analyzed separately from the bulk of the data.

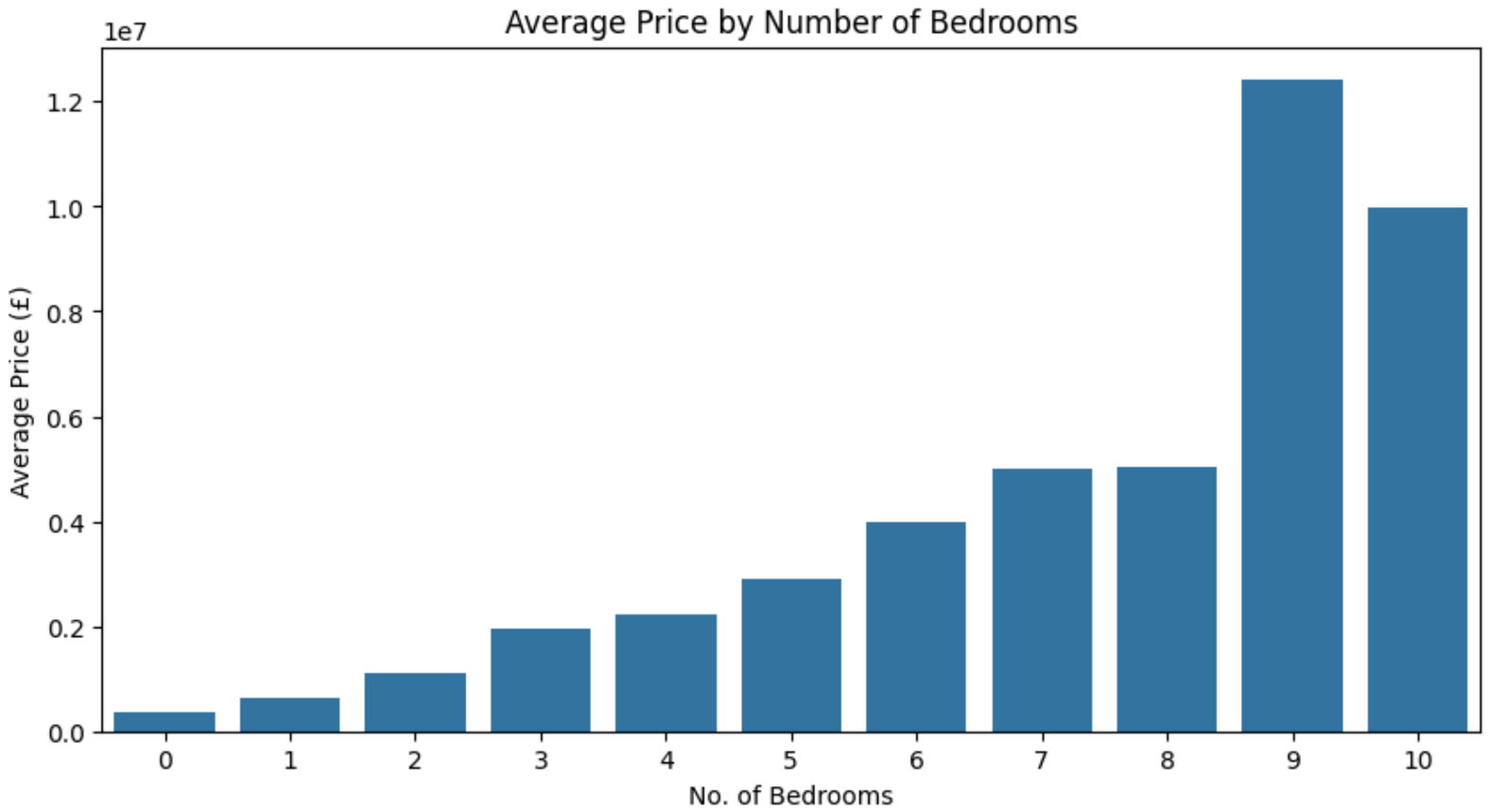
1. Price vs. Area (Price vs. "Area in sq ft")

Scatter Plot with Regression Line: Visualising how price changes with square footage. And a regression line to see if there's a linear relationship (i.e., higher area leads to higher prices)

The **regression line** shows the correlation between price and area, showing whether prices generally increase as square footage increases.

2. Price by Number of Bedrooms/Bathrooms

Column Chart: Each category (e.g., 1 bedroom, 2 bedrooms, 3 bedrooms) is displayed on the x-axis, and the average price is plotted on the y-axis.



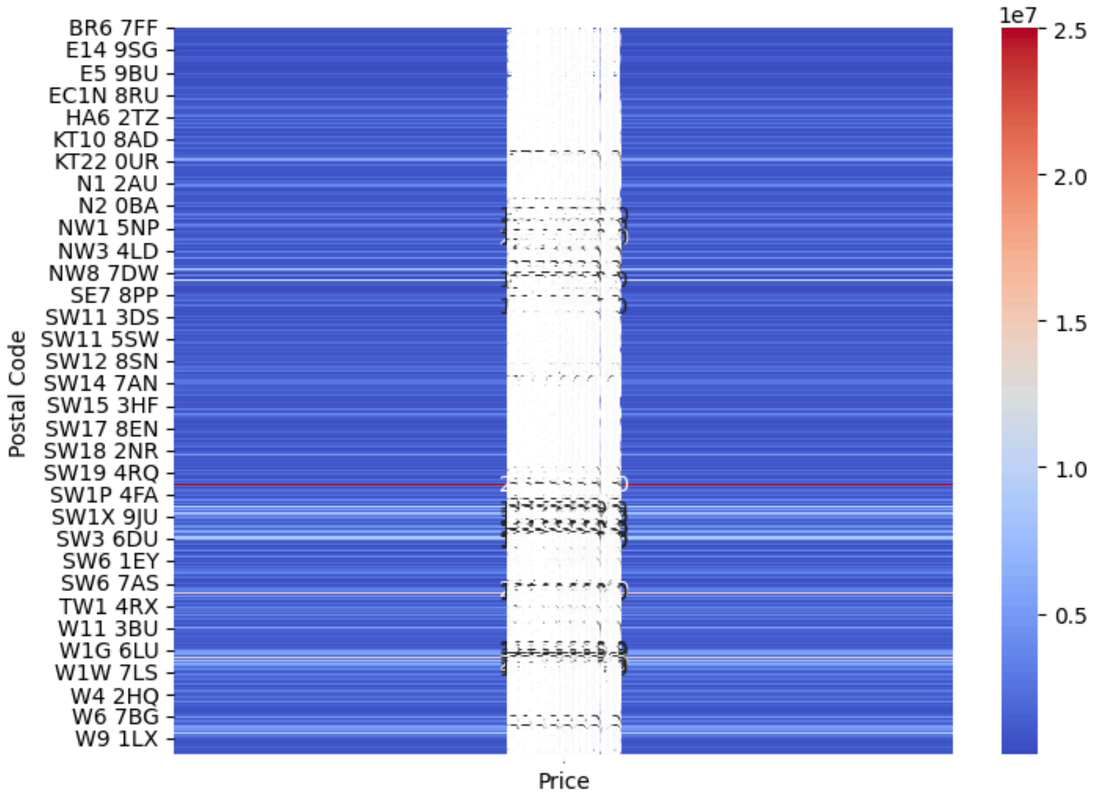
Heatmap for Price Distribution by Postcode

Price Distribution: The majority of the price points appear to be centered around a specific range (likely between 0 and 1 million), as shown by the denser regions (dark blue areas) in the centre of the plot.

Postal Code Variations: The price distribution seems fairly consistent across most postal codes. No clear outliers or extreme differences are visually evident between the postal codes listed on the y-axis.

-

Heatmap Colors: The colour scale indicates prices ranging up to 25 million (in red), though it seems that most prices are concentrated in the lower range (blues), with fewer higher priced properties reaching above 10 million.



The white zone in the middle of the graph represents a price gap, where there are few or no properties within that price range across the postal codes. Most properties either fall in the lower price range (shown by dark blue) or in higher price ranges (shades of red), with very few in between. This gap likely indicates that the dataset has mostly affordable or expensive properties but lacks representation in the middle price tier.

Conclusion

In this project, I conducted an in-depth analysis of house prices in London. I examined various factors that influence property values through exploratory data analysis, statistical methods, and visualisations (such as box plots). I uncovered several important trends and insights regarding the London housing market.

Key Findings:

House Type: The type of property significantly influences the price. For example, penthouses and houses generally have higher median prices, while studios and flats tend to be more affordable.

Price Variation: There is considerable price variation within certain property types, especially in flats and new developments, as shown by the range of values in the boxplots. These variations may be influenced by location, property size, and other amenities.

Outliers: The presence of outliers, such as luxury homes or penthouses, heavily skews the price distribution, especially in high-end areas of London. After removing outliers, the majority of properties fell within a more reasonable price range, providing a clearer picture of the typical housing market.

Location: Although not fully explored in this particular analysis, it is evident from the data that location significantly impacts prices. Central and premium neighborhoods tend to command higher property prices, further skewing the data when considering outliers.

Challenges and Limitations:

Missing Data: Some properties lacked specific details, such as location data, which may have limited the accuracy of certain analyses.
Future work could benefit from a complete dataset to better understand the role of neighbourhoods in price determination.

Outliers: While removing outliers helped in understanding the general trends, the presence of these extreme values suggests a need for
a separate analysis of the luxury property segment.

Future Work:

This project presents several opportunities for further analysis. Conducting a more detailed examination of the geographical distribution of property prices across London boroughs or neighbourhoods would yield additional insights. Additionally, expanding the dataset to include factors such as proximity to transport hubs, school districts, and other amenities could provide a more comprehensive understanding of the factors influencing house prices in London.

In conclusion, this project has effectively explored the key factors influencing house prices in London. It establishes a strong foundation for further research and can be utilised to inform property investment decisions, policy planning, or more detailed market analysis. By continuing to expand the scope of data and analytical methods, future work can illuminate the complex and dynamic nature of the London housing market.

Thank you for taking the time to view my project. I hope you found the analysis insightful, and I appreciate your interest in the London house price market analysis. Your feedback is valuable, and I look forward to any thoughts or suggestions you might have.