# London House Price Analysis Project Report

---

## 1. Introduction

### 1.1 Background

This project focuses on analyzing London housing market data to understand property price patterns, key influencing factors, and location-based trends. The London housing market is one of the most dynamic and expensive real estate markets, making it an ideal case study for data-driven price analysis.

### 1.2 Business Questions

The analysis was guided by the following key business questions. Each question is directly supported by the SQL queries used in the project.

1. **What is the overall size of the dataset and how many properties are available for analysis?**
   *SQL Query Used:*
   SELECT COUNT(*) AS total_properties FROM houses;
2. **What does the raw housing data look like before detailed analysis?**
   *SQL Query Used:*
   SELECT * FROM houses LIMIT 10;
3. **What is the overall price range of properties in London (average, minimum, and maximum)?**
   *SQL Query Used:*
   SELECT AVG(price) AS avg_price, MIN(price) AS min_price, MAX(price) AS max_price FROM houses;
4. **How do average property prices vary by house type?**
   *SQL Query Used:*
   SELECT house_type, AVG(price) AS avg_price FROM houses GROUP BY house_type ORDER BY avg_price DESC;
5. **Which are the top 10 most expensive properties in the dataset?**
   *SQL Query Used:*
   SELECT property_name, price, location, house_type FROM houses ORDER BY price DESC LIMIT 10;
6. **How do average property prices differ across London locations?**
   *SQL Query Used:*
   SELECT location, AVG(price) AS avg_price, COUNT(*) AS property_count FROM houses WHERE location IS NOT NULL GROUP BY location ORDER BY avg_price DESC;
7. **Which properties and locations are the most expensive when evaluated by price per square foot?**
   *SQL Query Used:*
   SELECT property_name, price, area_sq_ft, ROUND(price * 1.0 / area_sq_ft, 2) AS price_per_sq_ft FROM houses WHERE area_sq_ft > 0 ORDER BY price_per_sq_ft DESC;

8. **How does the number of bedrooms influence average property prices?**
   *SQL Query Used:*
   SELECT bedrooms, AVG(price) AS avg_price FROM houses GROUP BY bedrooms ORDER BY bedrooms;
9. **How do average property prices vary by postal code?**
   *SQL Query Used:*
   SELECT postal_code, AVG(price) AS avg_price FROM houses GROUP BY postal_code ORDER BY avg_price DESC;

These mapped queries ensured that each business question was answered using structured, reproducible SQL analysis, forming the foundation for insights presented in Python visualizations and the Power BI dashboard.

---

# 2. Dataset Overview

## 2.1 Data Source and Description

The dataset contains information on residential properties in London, including:

- Property name
- Location and postal code
- House type (e.g., house, flat)
- Price
- Area in square feet
- Number of bedrooms, bathrooms, and receptions

The raw dataset was initially stored as a CSV file (londonhouses.csv) and then processed through Python and SQL for analysis and visualization.

---

# 3. Methodology: Data Preparation and Analysis Using Python

Python was used for data loading, cleaning, feature engineering, exploratory data analysis (EDA), and database integration.

## 3.1 Data Loading and Initial Inspection

- The dataset was loaded using **Pandas**.
- Initial inspection was performed using head(), info(), and describe() to understand structure, data types, and summary statistics.

### 3.2 Data Cleaning

- Column names were standardized by converting them to lowercase and replacing spaces with underscores.
- Unnecessary whitespaces were removed to ensure consistency.
- A cleaned version of the dataset was saved as londonhouses_clean.csv for further use.

### 3.3 Feature Engineering

Additional features were created to enhance analysis:

- **Price per square foot**: calculated as price divided by area in square feet.
- **Price bands**: properties were categorized into price ranges to support segmented analysis.

### 3.4 Exploratory Data Analysis (EDA)

Several analyses and visualizations were created using **Matplotlib** and **Seaborn**:

- Distribution of property prices in London.
- Average price comparison by house type.
- Top 10 most expensive locations by average price.
- Relationship between property area and price.
- Price distribution by number of bedrooms.
- Identification of best-value locations based on average price per square foot.
- Correlation analysis between price and numerical features such as area, bedrooms, bathrooms, and receptions.

These analyses helped identify trends such as higher prices for larger properties, strong correlation between area and price, and significant variation across locations.

---

# 4. Database Design and SQL Analysis

### 4.1 Database Setup

- The cleaned dataset was uploaded to a **MySQL** database using **SQLAlchemy** and **PyMySQL**.
- A table named houses was created to store the property data.

### 4.2 SQL Queries and Analysis

SQL was used to perform structured analysis directly on the database:

- **Total number of properties** to understand dataset size.
- **Descriptive statistics** including average, minimum, and maximum property prices.
- **Average price by house type** to compare houses, flats, and other property categories.
- **Top 10 most expensive properties** in London.
- **Average price by location**, along with property counts per location.
- **Price per square foot analysis** to identify premium and value areas.
- **Average price by number of bedrooms** to evaluate how size affects pricing.
- **Average price by postal code** to capture regional pricing differences.

These SQL queries allowed efficient aggregation and filtering, complementing the Python-based analysis.

---

# 5. Power BI Dashboard

## 5.1 Dashboard Purpose

Power BI was used to design an interactive dashboard that communicates the results of the analysis in a clear and intuitive manner for non-technical users. The dashboard allows stakeholders to explore housing prices across London using filters and visual comparisons.

## 5.2 Dashboard Pages and Visuals

**Figure 1: Overall KPIs**

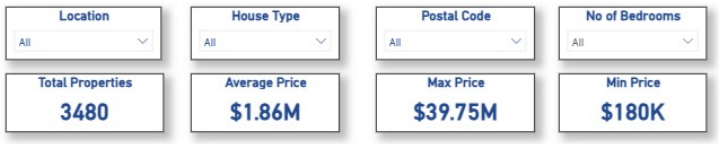


**Figure 2: Average Price by House Type**

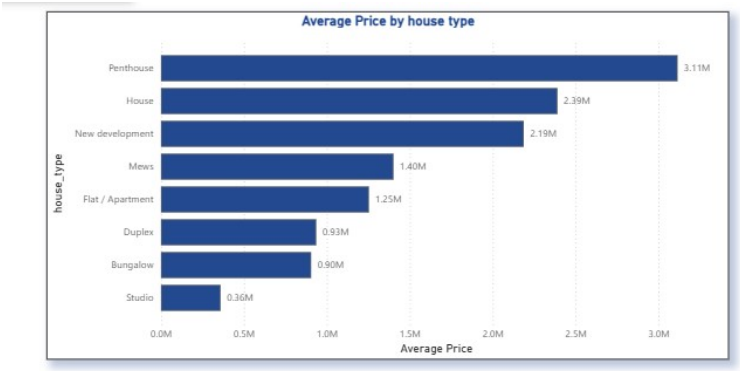

**Figure 3: Average Price by Location**

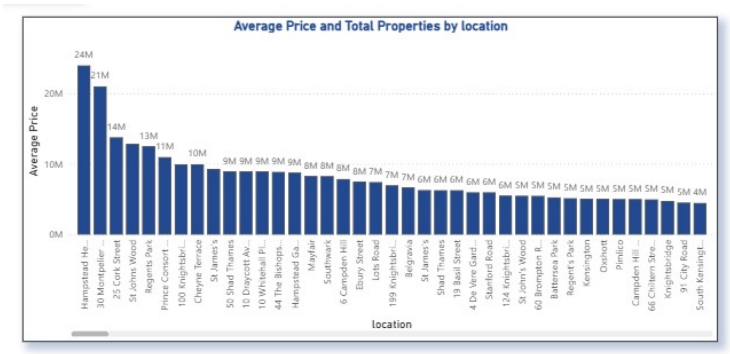


**Figure 4: Price  by No of Bedrooms**

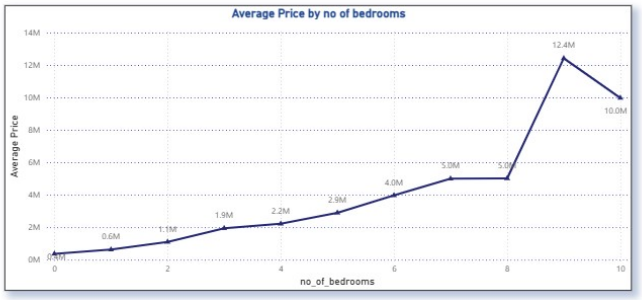


**Figure 5: Price by Postal Code**


## 5.3 Insights from the Dashboard

- Certain London locations consistently show significantly higher average property prices.
- Houses tend to have higher average prices compared to flats.
- When evaluated using price per square foot, some locations provide better value despite lower total prices.
- Property size and number of bedrooms strongly influence pricing trends.

---

# 6. Results and Key Findings

- Property prices in London vary widely by location and property type.
- Area (in square feet) is one of the strongest predictors of price.
- High total price does not always mean poor value; price per square foot provides deeper insight.
- SQL and Python together enable both deep analysis and scalable querying.
- Power BI enhances understanding through interactive, business-friendly visualizations.

---

# 7. Conclusion and Future Work

## 7.1 Conclusion

This project demonstrates an end-to-end data analytics workflow using Python for data preparation and exploratory data analysis, SQL for structured querying and aggregation, and Power BI for interactive data visualization. By integrating these tools, the project delivers clear insights into the London housing market and showcases practical data analytics skills applicable to real-world business problems.

---

## 7.2 Future Work

Future enhancements to this project could include:

- Incorporating time-series data to analyze price trends over time.
- Adding socio-economic or transport accessibility data for deeper insights.
- Deploying the dashboard using Power BI Service for real-time access.

---

# 8. Tools and Technologies Used

- **Python**: Pandas, Matplotlib, Seaborn, SQLAlchemy
- **SQL**: MySQL
- **Power BI**: Interactive dashboard creation
- **Jupyter Notebook**: Analysis and documentation

---