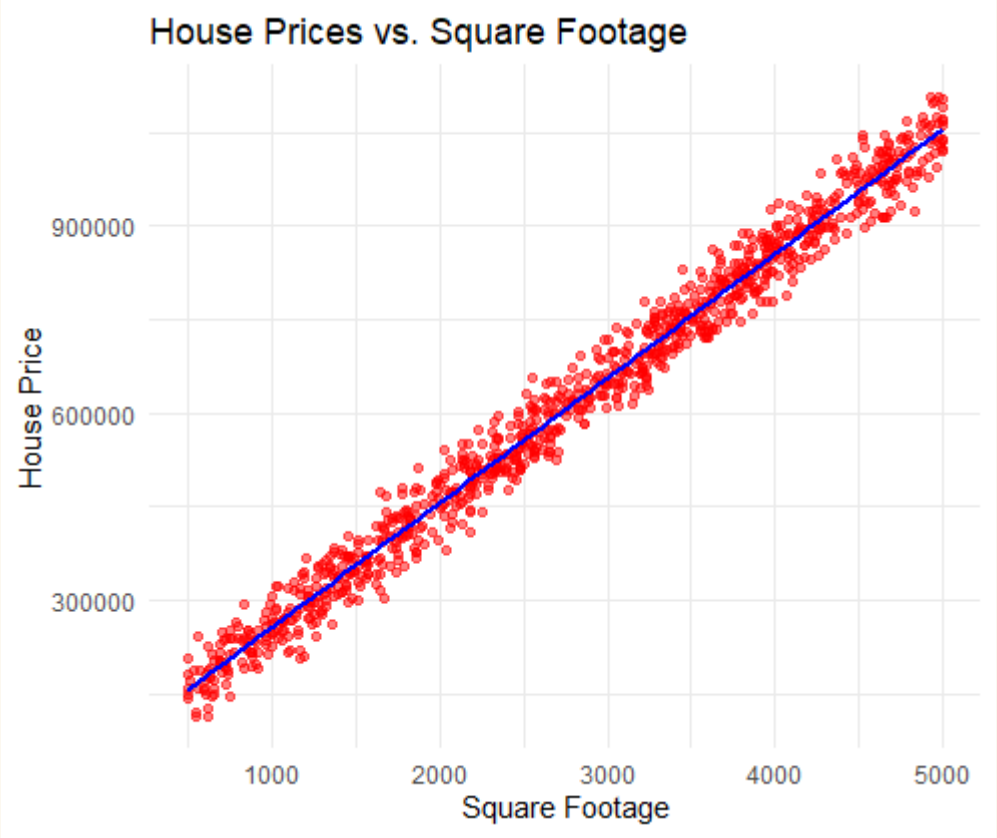
Introduction

This report examines the factors influencing house prices using a dataset of 1000 properties. It begins with descriptive statistics and correlation analysis to explore relationships between variables such as square footage, year built, and lot size. A multivariate linear regression model is then developed via forward stepwise selection, validated for assumptions like normality and multicollinearity, and evaluated for influential outliers using Cook's Distance. House prices for the remaining data are then predicted and recommendations for model refinement are made.

Descriptive Analysis of Housing Dataset

1. Square_Footage

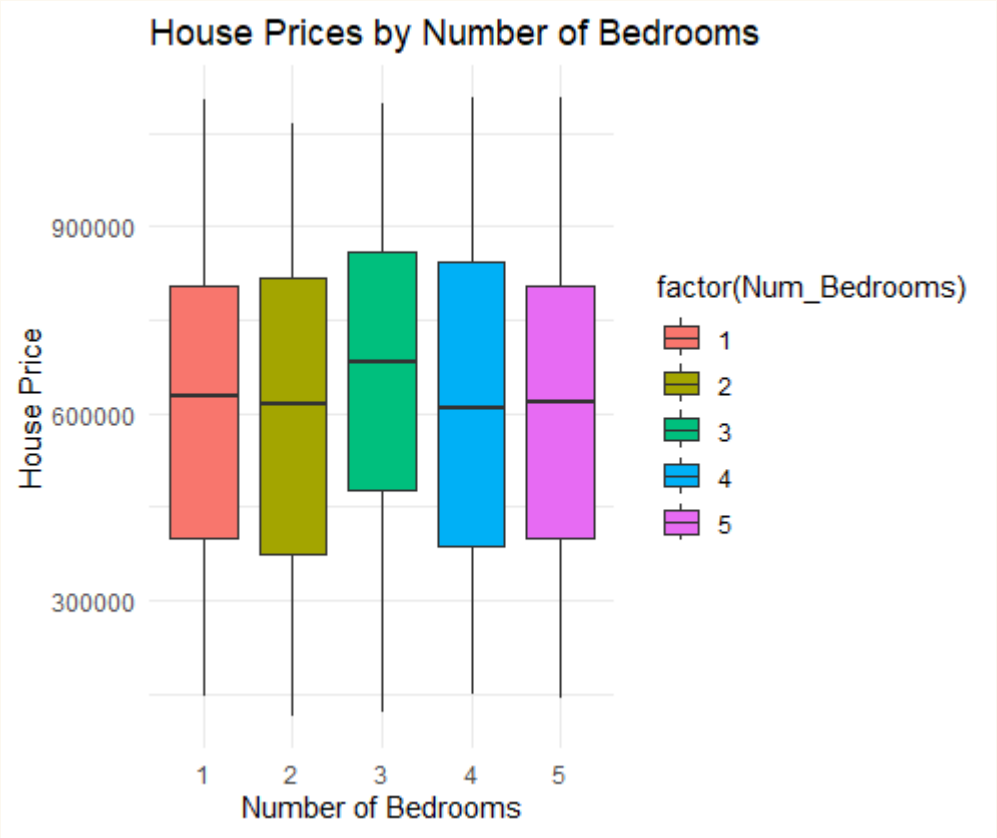


(House price increases almost perfectly with square footage)

The average property size is approximately 2,813 sq.ft, though the median of 2,856 sq.ft suggests a slight concentration of smaller homes that marginally reduce the mean. The large SD (1,334.51) indicates significant variability in property sizes, spanning from compact urban units (503 sq.ft) to expansive luxury homes (4,999

sq.ft). The IQR captures 50% of properties between 1,744 and 3,852 sq.ft, reflecting a mix of mid-sized and large homes.

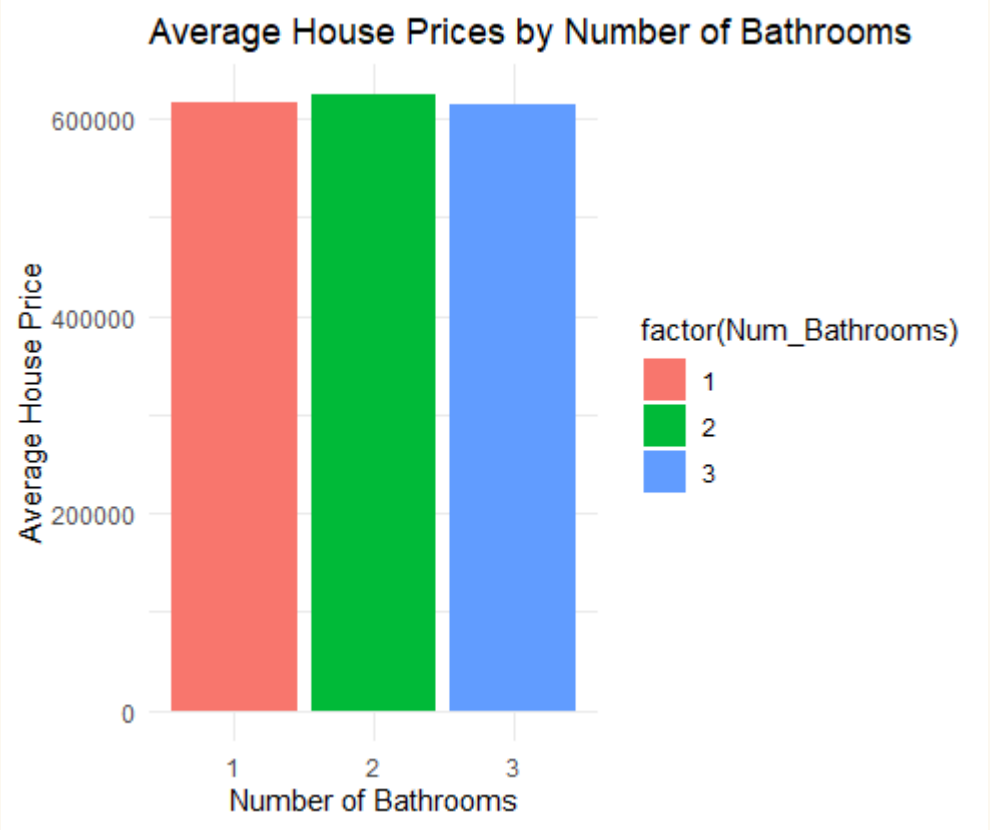
2. Num_Bedrooms



(3-bedroom houses have the highest median pricing)

A typical house has 3 bedrooms, aligning with the median. The narrow SD (1.256) and IQR (2–4) confirm consistency in bedroom configurations, though 25% of properties have 4–5 bedrooms. This suggests a market dominated by family-sized homes, with limited extreme outliers, such as 1-bedroom or 5-bedroom properties.

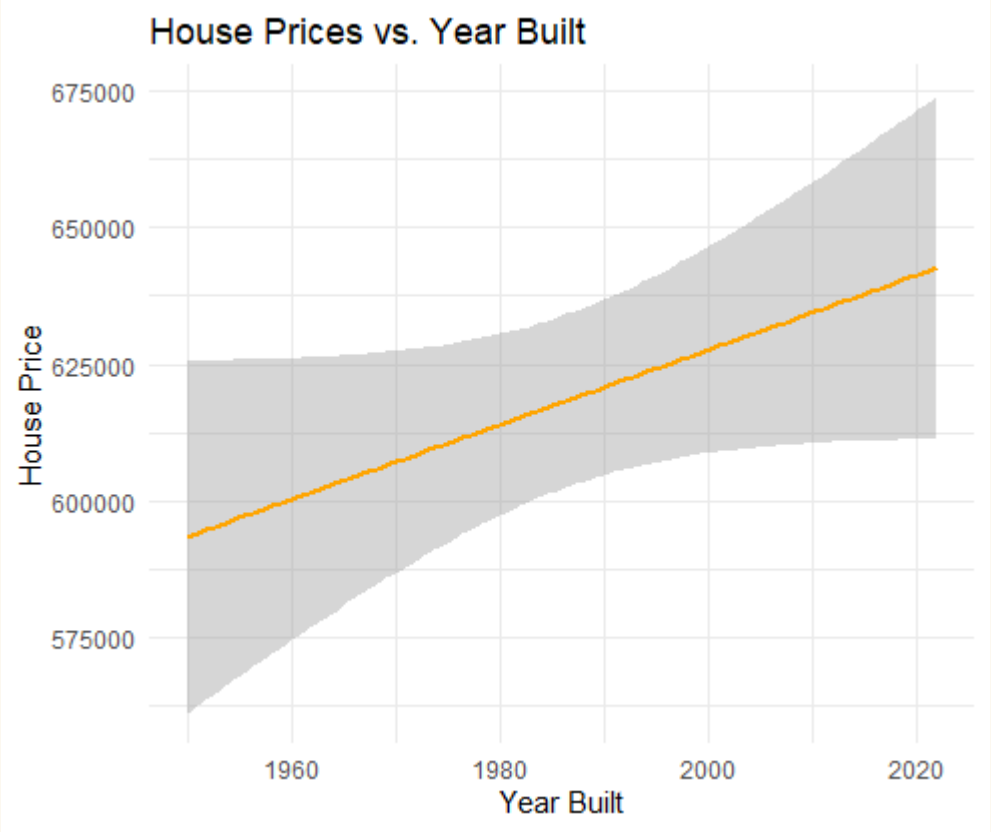
3. Num_Bathrooms



(Average house price in terms of number of bathrooms is very similar)

The median indicates that most properties have 2 bathrooms. The mean (1.975) is marginally lower, reflecting a slight prevalence of one-bathroom homes. The SD (0.857) and IQR (1–3) highlight a balanced distribution, with 50% of properties having 1–3 bathrooms. This suggests a mix of older, simpler homes and modern properties with probably upgraded amenities.

4. Year_Built

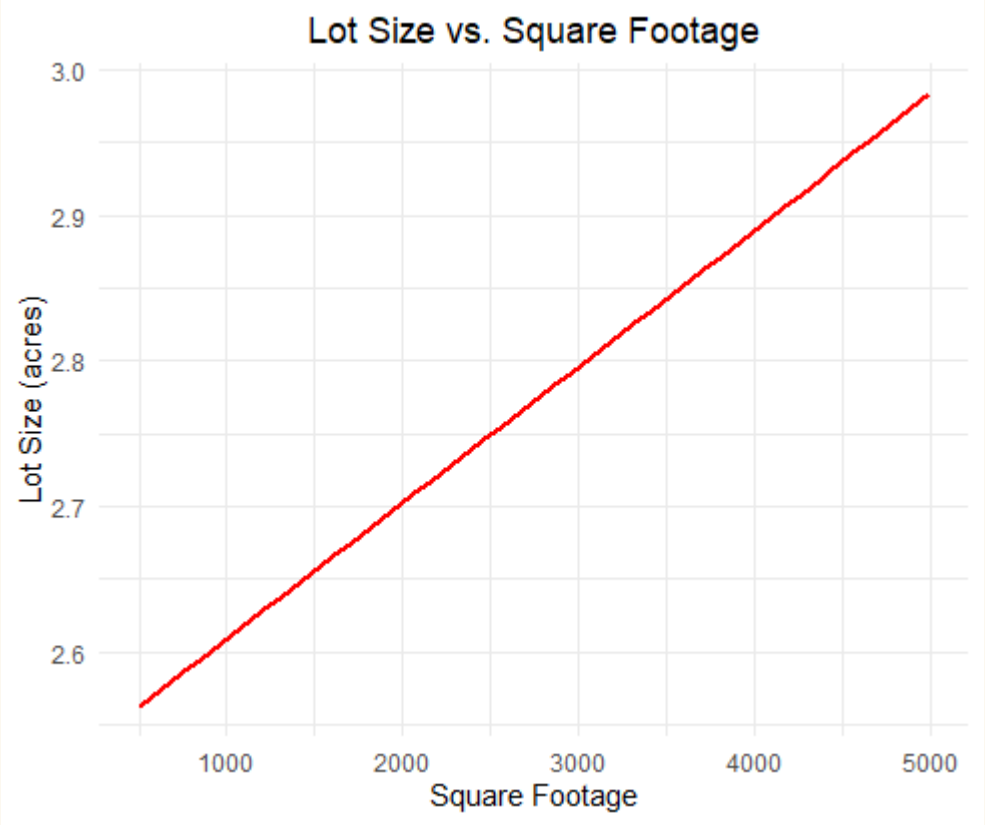


(There is a noticeable increase in price in newer homes)

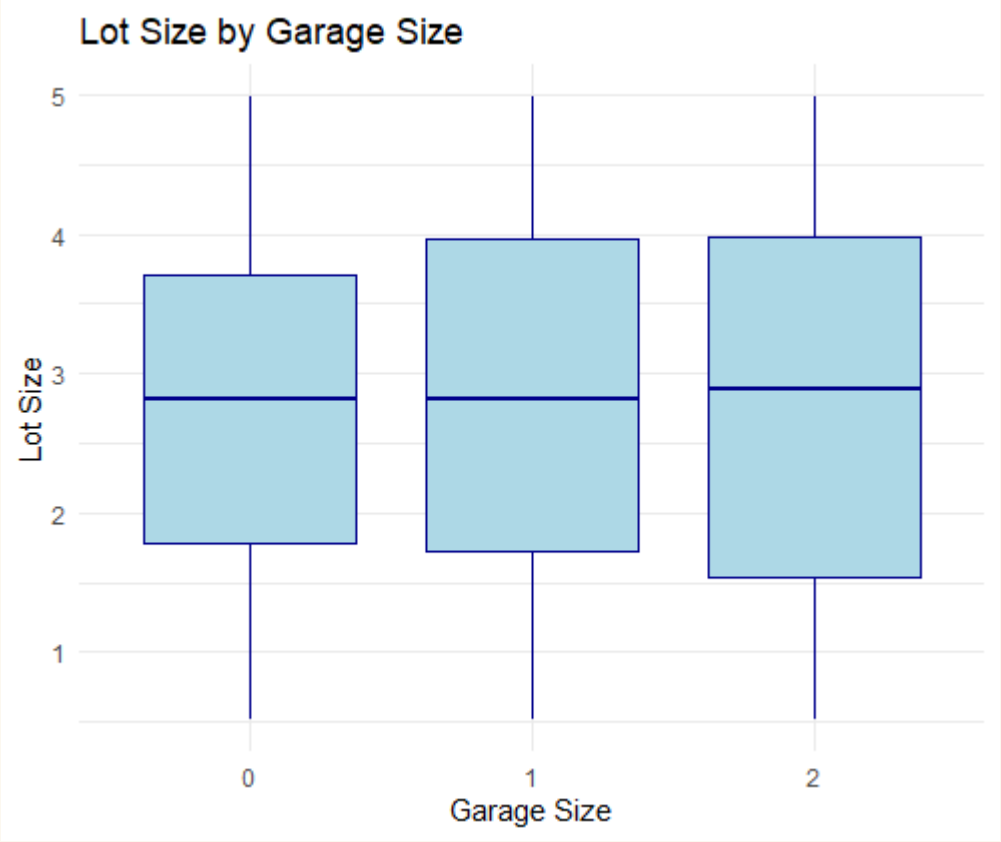
The dataset has properties built between 1950 and 2022, with the average construction year around 1987. The IQR (1969–2005) captures 50% of properties, indicating a focus on post-1970s housing stock. The SD (20.63) reflects a relatively even distribution of construction years, though 25% of homes are modern (post-2005), aligning with observed price premiums for newer constructions.

5. Lot_Size

The average lot size is 2.78 acres, with the median closely aligned (2.81 acres). The SD (1.382) and range (0.51–4.99 acres) reveal variability, likely due to high-end properties with larger lots (e.g., 4.99-acre estates). The IQR (1.66–3.93 acres) shows that half of all properties fall within this range, emphasising a mix of suburban and rural lots.
Lot size variability (SD = 1.38 acres) suggests a mix of urban, suburban, and rural properties.

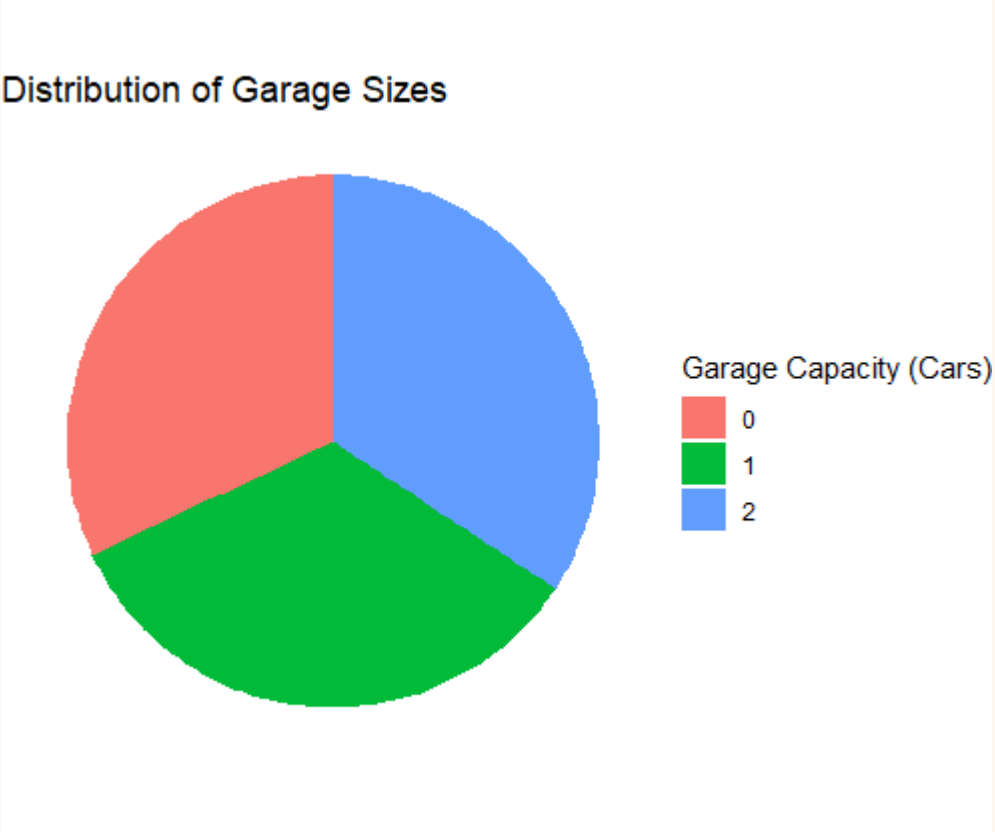


(More square footage means a bigger lot size)



(The median lot size for each garage size is similar)

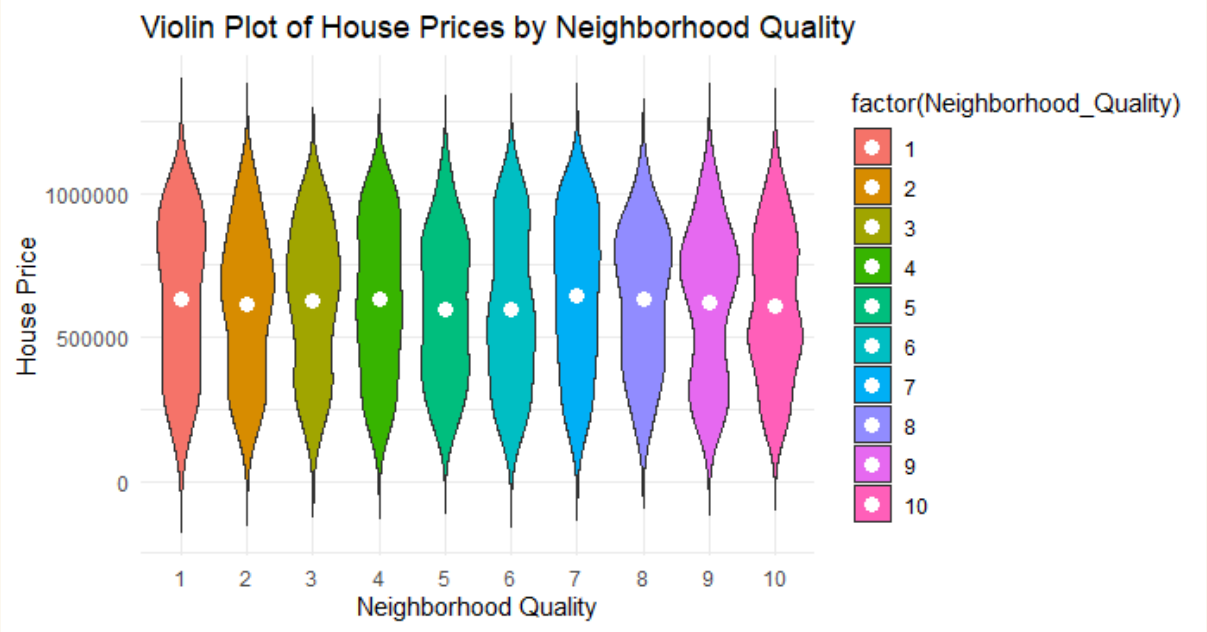
6. Garage_Size



(Garage size is almost perfectly distributed throughout the data set)

The median garage size is a 1-car garage. The mean (1.023) is marginally higher due to a subset of 2-car garages that skew it upwards. Garages are probably not a strong predictive factor in house pricing.

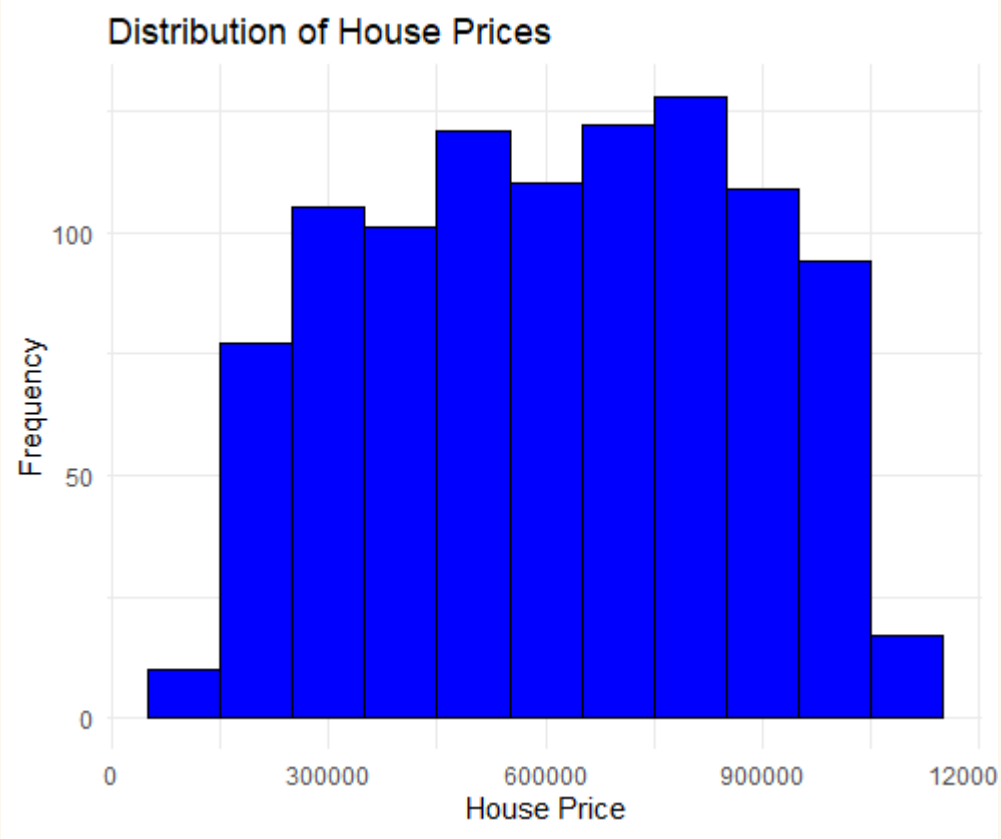
7. Neighborhood_Quality



(House prices are similar across all neighbourhood types with a slight increase in price for the 8 and 9-rated neighbourhoods. Mean prices(white dots) are similar as well)

Neighbourhood quality ratings average 5.6/10, but the median (6.0) suggests a concentration of mid-to-high-quality areas. The SD (2.75) and IQR (3–8) reflect variability, with 25% of properties in top-tier neighborhoods (8–10). This variable likely captures locational premiums which are critical for price differentiation.

8. House_Price (Target Variable)



(Most houses are priced between £300,00 to £900,000)

The median price (£627,624) exceeds the mean (£618,467), signalling the presence of lower-end outliers. The SD (£268k) and range (£111k–£1.1M) highlight significant price variability, driven by size, location and modernity. The IQR (£400k–£827k) captures the core market segment.

Correlation Analysis and Key Observations

Why Spearman

Given that your Shapiro-Wilk test results show extremely low p-values ($p < 1e-10$) for all variables, the data deviate from normality. This justifies the use of a non-parametric method.

Although some variables—like the number of bedrooms or bathrooms—may have tied values, the impact is mitigated by the large sample size. Spearman's correlation, which assesses monotonic relationships, is robust enough thus its selection over alternatives like Kendall's.

Key Variable Correlation Analysis:

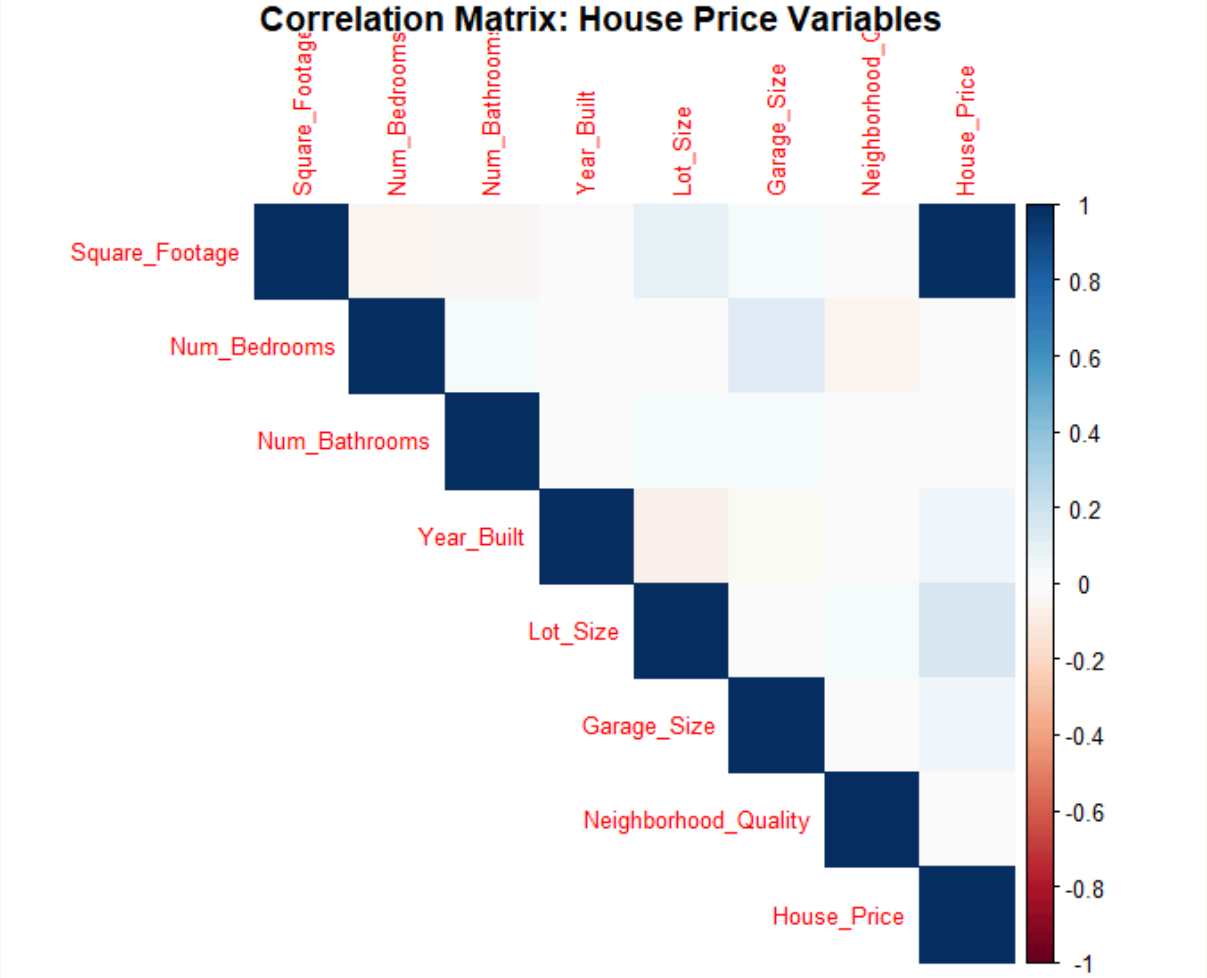
1. Square_Footage and House_Price ($r = 0.99$, $p < 0.001$)
The near-perfect positive correlation between Square_Footage and House_Price indicates that property size overwhelmingly dominates pricing. However, if there is extreme collinearity that will pose a statistical concern, as it could destabilise regression coefficients and inflate standard errors.

2. Neighborhood_Quality and House_Price ($r = -0.01$, $p = 0.93$)
The absence of a correlation between Neighborhood Quality and house price contradicts theoretical expectations. High-quality neighbourhoods typically command premium prices, suggesting potential data integrity issues. This raises a concern whether the variable was mislabeled.

3. Year_Built and House_Price ($r = 0.06$, $p = 0.04$)
The weak positive correlation between Year_Built and House_Price conflicts with prior analysis showing stronger effects for newer homes. This discrepancy may arise from nonlinear trends for example post-2000 constructions having disproportionately higher prices or confounding variables for instance modern homes in lower-quality neighbourhoods.

4. Garage_Size and Num_Bedrooms ($r = 0.11$, $p = 0.03$)
The moderate correlation between Garage_Size and Num_Bedrooms suggests a weak but statistically significant association between larger homes and garage capacity. However, **Garage_Size's correlation with House_Price** ($r = 0.05$, $p = 0.05$) is marginal, implying it has minimal practical relevance for pricing. This aligns with the dataset's skew toward properties with 0–2 garages.

5. Lot_Size and House_Price ($r = 0.16$, $p = 0.01$)
The modest positive correlation between Lot_Size and House_Price indicates that larger lots contribute to higher prices, though the effect is secondary to square footage. This relationship may be non-linear for example premium pricing for lots exceeding 4 acres or confounded by geographic clustering for instance rural vs. urban properties.



The correlation matrix above shows that square footage, lot size, year built and garage size have the highest correlation with house price respectively.

Analysis of the Regression Model for House Price Determination.

The regression model ***(House_Price ~ Square_Footage + Year_Built + Lot_Size + Num_Bedrooms + Num_Bathrooms + Garage_Size)*** was developed using forward stepwise direction. Starting with an intercept-only baseline model (House_Price ~ 1, AIC = 24,742). **Square_Footage was the first variable included**, reducing the AIC to 20,717. Subsequent **additions of Year_Built and Lot_Size** further lowered the AIC to 19,667, **followed by Num_Bedrooms**, Num_Bathrooms, and Garage_Size, leading to a final AIC of 18,277. **Neighborhood_Quality was excluded** at all stages due to its minimal contribution to AIC reduction, suggesting it lacks predictive utility in this dataset.

The model's assumptions were validated through VIF tests, which yielded values between 1.01 and 1.02 for all variables. These values are well below the threshold of 5, confirming no multicollinearity contrary to what was speculated in the correlation analysis. The Shapiro-Wilk test for residual normality returned a p-value of 0.5893, failing to reject the null hypothesis of normally distributed residuals. This validates the model's assumptions.

Interestingly, all model development methods retained the same variables across iterations, reinforcing the robustness of the developed model. However, the exclusion of Neighborhood_Quality—despite theoretical relevance to property valuation—suggests potential data mislabeling or confounding effects.

Cook's Distance analysis identified 51 influential observations exceeding the threshold of approximately 0.14. These observations disproportionately skew coefficient estimates. For example, House 23 has a very small square footage of 577 with 5 bedrooms and 3 bathrooms on a 2-acre lot in a 10-rated neighbourhood at only £174,629. Such a record represents either a very unique house or a case of data inconsistency. ***But since the R^2 (0.9985) is near perfect omitting these records will not improve the model significantly thus retaining them in the model development process.***

Interpretation of the Multivariate Linear Regression Model

Final model with coefficients;

House_Price = -2004397.10
*+ 199.80 * Square_Footage*
*+ 990.30 * Year_Built*
*+ 14932.80 * Lot_Size*
*+ 10163.60 * Num_Bedrooms*
*+ 8280.80 * Num_Bathrooms*
*+ 5139.30 * Garage_Size*

The model quantifies the marginal effects of variables on house prices, with coefficients representing the change in price per unit increase in each variable while holding others constant. The **intercept** (-£2,004,397.10) is statistically derived but non-interpretable since no house can cost a narrative value. **Square_Footage** dominates the model (£199.80/sq.ft), aligning with real-world intuition that size is the primary price driver. **Year_Built** adds £990.30/year, suggesting an increment in the price for newer constructions, though this may capture unmeasured factors like modern amenities. **Lot_Size** contributes £14,932.80/acre, while **Num_Bedrooms** (£10,163.60/bedroom) and **Num_Bathrooms** (£8,280.80/bathroom) reflect demand

for family-oriented layouts. **Garage_Size** has the smallest effect (£5,139.30/space) despite statistical significance ($p < 2.2e-16$), likely due to low variability. **Neighborhood_Quality** was excluded in stepwise methods, implying its influence is confounded by other variables. The model explains 99.85% of price variability ($R^2 = 0.9985$) with all variables statistically significant ($p < 2.2e-16$), though the near-perfect fit and exclusion of a relevant variable like neighbourhood quality warrants scrutiny for overfitting.

Predicted House Prices using the Developed Model

No	Predicted_House_Price
995	423771
996	697582
997	686741
998	577093
999	966543
1000	726250

Conclusion

This analysis identifies Square_Footage as the dominant driver of house prices, with a near-perfect correlation and a strong coefficient. Secondary factors include Year_Built and Lot_Size, reflecting premium prices for modernization and land value. The model explains 99.85% of price variability and satisfies parametric assumptions. However, 51 influential observations that suggest luxury properties or data anomalies—disproportionately affect coefficient estimates. The exclusion of neighbourhood quality and near-perfect fit suggest potential overfitting or omitted variable bias. Future work could explore the non-linear effects of variables such as Year_Built and subgroup analyses for high-leverage cases.