Introduction

The housing market is one of the most important drivers of the economy and city planning. It affects how people invest, where they choose to live, and how communities grow. For this project we have selected North Vancouver, with its beautiful landscapes and high quality of life, is a great example of a fast-changing housing market.

This report examines housing data from North Vancouver to understand the key factors shaping property prices and market trends. Using a detailed dataset, the report provides valuable insights for buyers, sellers, investors, and policymakers.

The dataset includes important home details, such as their sold price, size, and lot area. It also contains unique features like pools or waterfront views and neighbourhood details. These help show how different factors add to a property's appeal and value. The analysis also identifies trends in how prices have changed over time by including the year properties were sold.

The first step in the analysis was to clean and prepare the data. This involved fixing missing information, handling extreme values, and ensuring all data consistency. For instance, townhouses missing lot sizes were marked as "NA," and unusually high sale prices were adjusted to match similar properties. This careful preparation made the data ready for accurate analysis.

The report examines many aspects of the North Vancouver housing market. It finds that housing prices have steadily increased over the last decade, especially in sought-after neighbourhoods like Deep Cove and Upper Lonsdale. Special features like pools or waterfront views also significantly increase property values. Regression models help identify which factors influence prices the most, showing that the year sold and building size are key drivers, while age has less impact.

This report provides helpful information for a wide range of people. Buyers can better understand what adds value to a property. Sellers and investors can use the findings to maximize returns. Policymakers and planners can use insights to create housing strategies for the community. By breaking down complex data into clear results, this report helps everyone involved in the North Vancouver housing market make better decisions.

Key Questions Addressed in the Project

This project answers key questions about what drives housing prices in North Vancouver. Here are the main questions explored:

1. **How have housing prices changed over the years?**
   The project looks at how average prices have grown or fluctuated each year, showing whether the market has been stable or unpredictable.
2. **What is the relationship between sold prices and key factors?**
   - **Neighborhoods:** How do prices differ between areas, and which neighbourhoods have the highest demand?
   - **Special Features:** Do unique features like waterfront views increase property values?
   - **Pools:** Does having a pool make homes sell for higher prices?
3. **How do neighbourhood and special features work together?**
   This examines if premium features, like waterfront views, add extra value in high-demand neighbourhoods.
4. **Which factors are most and least related to prices?**
   - Year Sold: Does the year of sale strongly affect the price?
   - Age: How much does a property's age matter?
   - Building Size: Does the size of the home directly relate to its value?
5. **Does the number of bathrooms influence prices?**
   The project calculates how adding bathrooms can increase a home's value.
6. **Does having a pool add value?**
   The analysis compares prices for homes with and without pools to determine whether pools significantly affect pricing.
7. **Do special features significantly impact property value?**
   The analysis highlights how features like waterfront or Waterview properties add value compared to homes without these features.

8. **Do property prices differ by neighbourhood?**
   This compares average prices across neighbourhoods to see which areas are most desirable.
9. **What is the best model for predicting prices?**
   The project uses a regression model to identify the strongest predictors of housing prices, such as year sold and building size.
10. **How is property age-related to the sold price?**
    While age might affect value due to modernization or wear, the project measures whether this relationship is strong or weak.
11. **What practical lessons can be learned from this study?**
    - **For Buyers:** Understand what features justify higher prices.
    - **For Sellers:** Focus on key attributes that increase property value.
    - **For Investors:** Spot trends and invest in high-performing areas.
    - **For Policymakers:** Use data to create housing policies that meet community needs.

This project provides a clear and thorough understanding of what drives housing prices, helping stakeholders make smart decisions based on solid data.

Analyzing the North Vancouver Housing Data

Housing data analysis is crucial for understanding market trends, identifying anomalies, and making informed decisions. This document provides an in-depth exploration of North Vancouver housing data, focusing on understanding the variables, cleaning and preparing the data, and addressing missing values and outliers. Here's a detailed analysis:

Understanding the Data

The dataset comprises various variables that describe properties in North Vancouver. Below is a breakdown of these variables:

1. **Year Sold**: A numerical variable representing the year a house was sold. This provides a temporal context for market trends.
2. **Property Type**: A categorical variable classifies properties into detached houses, townhouses, or apartments. This allows us to compare different property types.
3. **Age**: A numerical variable indicating the property's age helps understand its impact on depreciation or modernization.
4. **Building Size**: This variable quantifies the size of the house numerically, measured in square feet.
5. **Lot Size**: This is numerical, indicating the size of the property lot. Larger lot sizes often correlate with higher property values.
6. **Neighborhood**: A categorical variable specifying the house's location can significantly influence property prices due to varying desirability.
7. **# of Bedrooms**: Numerical, representing the number of bedrooms in a property.
8. **Number of Bathrooms: This is numerical, indicating the number** of bathrooms, an important feature in assessing property value.
9. **Special Feature**: A categorical variable that flags the presence of distinctive attributes, like waterfront views or luxury finishes.

10. **Pool**: A categorical variable marking the presence of a pool, often associated with higher-end properties.
11. **Sold Price**: This is numerical, representing the price at which the house was sold. It is the target variable for many analyses, including market trend studies.

Data Cleaning and Preparation

Data cleaning is critical in ensuring that analyses are accurate and reliable. The following issues were identified and addressed in the dataset:

1. Handling Missing Values

- **Issue**: ID 11 is missing in the *Lot Size* column.
- **Analysis**: As all townhouses in the dataset have missing *Lot Size* values (NA), we assumed the missing value for ID 11 was consistent with this pattern.
- **Solution**: Marked the missing value for ID 11 as "NA" to maintain dataset consistency.

2. Addressing Outliers

Outliers can significantly skew analyses, especially in datasets involving monetary values. Here are the outliers identified:

a. Row 16 (2011)

- **Details**: A detached house sold for $8,800,000, an unusually high value compared to similar properties sold in the same year in Central Lonsdale.
- **Comparison**:
    - Range of similar properties: $1,915,000 - $2,430,000.
    - Average value: $2,108,333.
- **Action Taken**: Adjusted the *Sold Price* to the average ($2,108,333) to align with comparable properties.

b. Row 57 (2015)

- **Details**: A detached house sold for $10,900,000 in Lynn Valley, far exceeding the typical range.
- **Comparison**:
    - Range of similar properties: $1,690,000 - $2,570,000.
    - Average value: $2,255,714.
- **Action Taken**: Adjusted the *Sold Price* to the average ($2,255,714) for consistency.

c. Row 95 (2019)

- **Details**: A detached house with unique characteristics (23,000-square-foot lot, 8 bedrooms) sold for $9,300,000. This price was vastly higher than a comparable property (ID 12), which sold for $2,800,000 despite having a similar size.
- **Comparison**:
    - ID 95: Lot Size = 6,500 sq. ft., *Sold Price* = $9,300,000.
    - ID 12: Lot Size = 6,220 sq. ft., *Sold Price* = $2,800,000.
- **Options Considered**:

1. **Retain the outlier**: The price might be valid if ID 95 represents a luxury property.
2. **Remove or adjust the outlier**: If the price discrepancy is due to an error or anomaly, treat it as an outlier.

- **Action Taken**: ID 95 was removed as an outlier due to its disproportionate impact on the dataset.

3. Data Formatting

- **Checks Performed**:
    - Ensured numerical variables were correctly classified (e.g., *Sold Price*, *Building Size*).
    - Verified categorical variables, such as *Property Type* and *Special Feature*, were properly encoded.

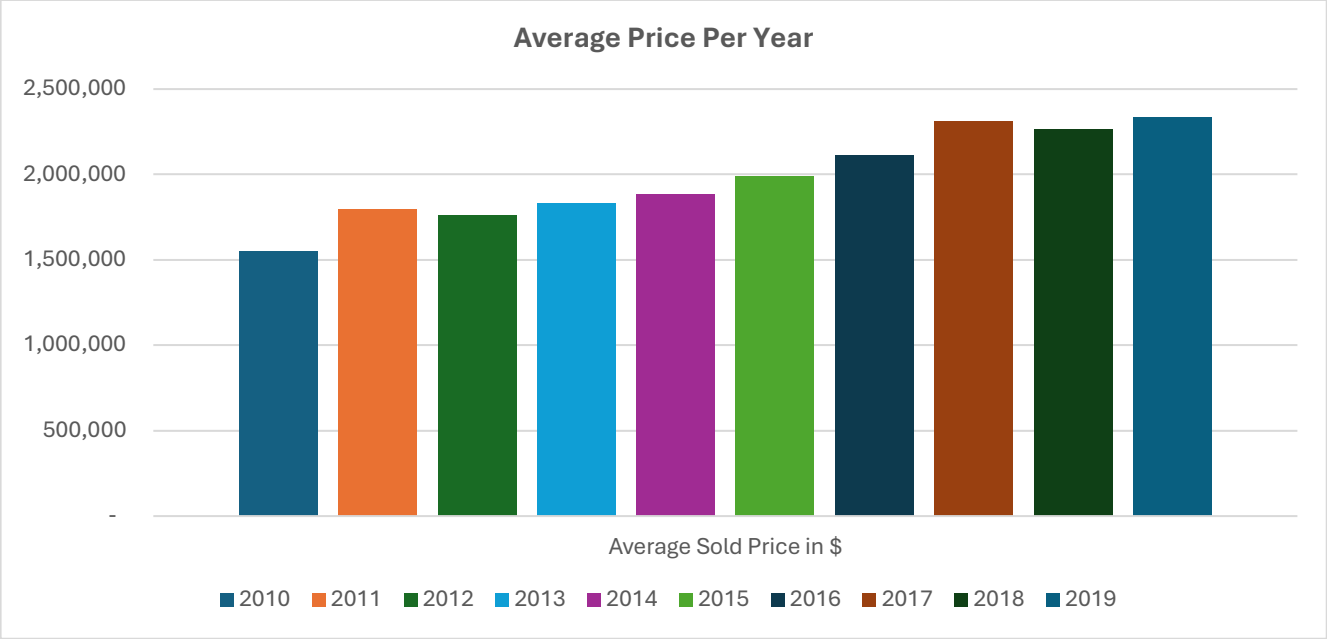
- **Action Taken**: Reformatted data types where necessary to ensure consistency.

4. Ensuring Data Consistency

- **Steps Taken**:
    - Cross-checked entries for errors (e.g., duplicate or inconsistent values).
    - Standardized data entry formats (e.g., categorical labels like "Yes/No" for *Pool* and *Special Feature*).
- **Outcome**: A uniform dataset was achieved that was ready for analysis.

Notes: The North Vancouver housing dataset was refined for accurate analysis through meticulous cleaning and preparation. Missing values were appropriately handled, outliers were identified and treated based on contextual comparisons, and data formatting and consistency checks ensured integrity. These steps are crucial for deriving meaningful insights into market trends and property valuations in North Vancouver.

Q1 **Has the mean price varied over the years?**



Steady Growth in Housing Prices

Over the past decade, the mean price of houses in North Vancouver has shown a consistent upward trend. In 2010, homes selling for an average of about **$1,548,500**, and by 2019, this figure had increased to about **$2,330,000**. This steady rise reflects the growing demand for housing in the area, driven by its nature , lifestyle appeal, and economic opportunities.

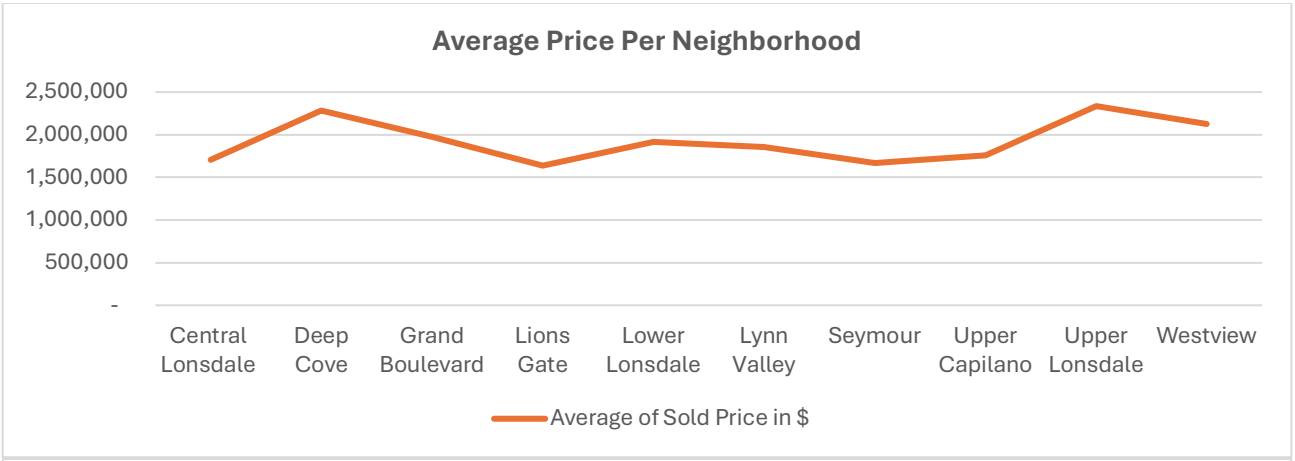
Year-to-Year Variations

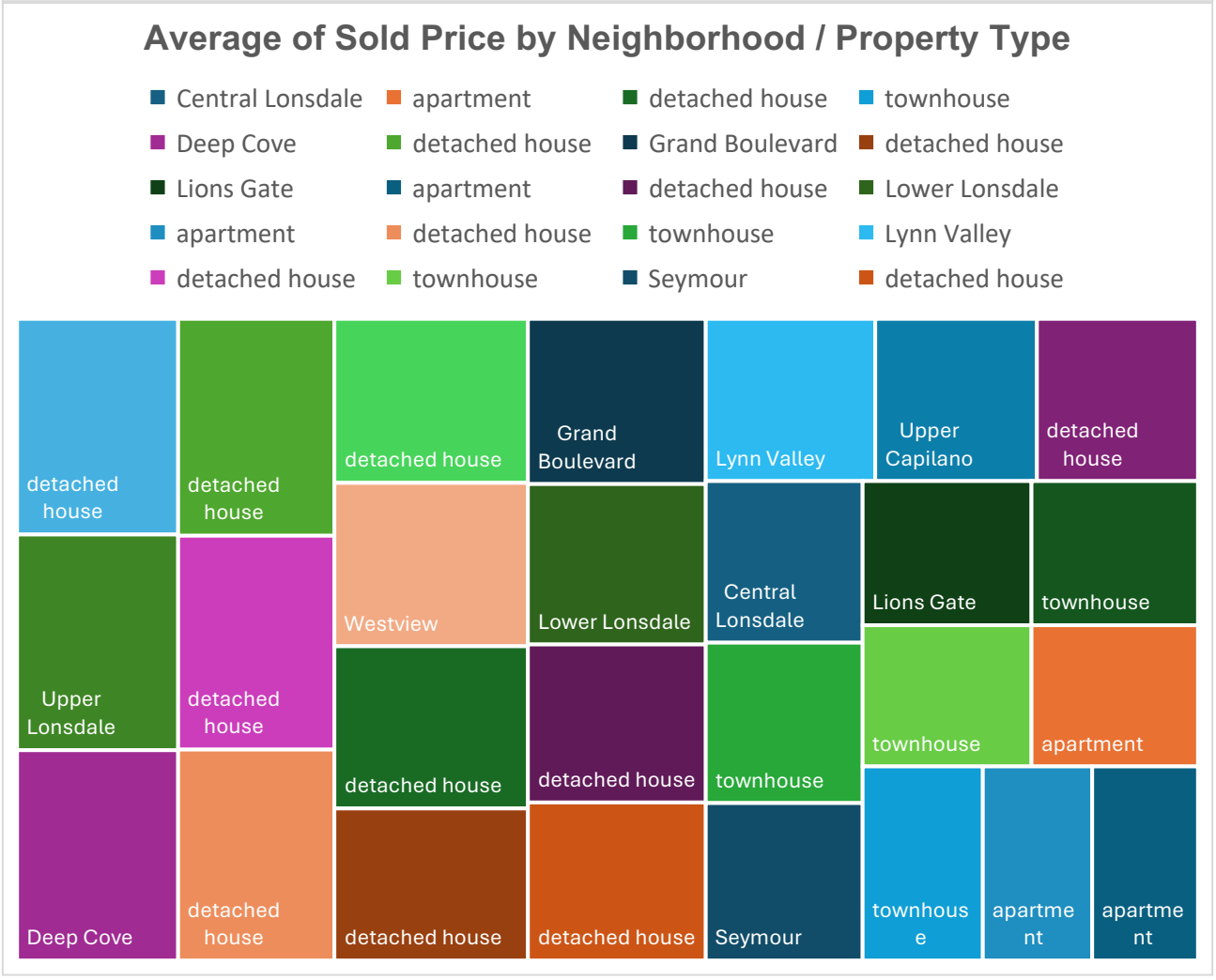
While the overall trend has been upward, there were slight fluctuations along the way. For example, the average price dipped slightly between **2011 and 2012** and again in **2018** before rebounding. Despite these minor setbacks, the market has remained strong, particularly after **2015**, when prices consistently exceeded **$2,000,000**. These changes suggest a resilient housing market with long-term growth potential.

Takeaways for Buyers

For someone like David, considering a move to North Vancouver, this data highlights the importance of timing. With prices steadily increasing, entering the market sooner could prove beneficial. While the cost of housing is undoubtedly high, the investment aligns with the region's strong demand and exceptional quality of life, offering both a home and long-term value.

Q2 **Use appropriate charts to describe individually the relationship between the key response variable (price sold) and the following key explanatory variables (neighbourhood, special feature, pool). Comment on the result.**





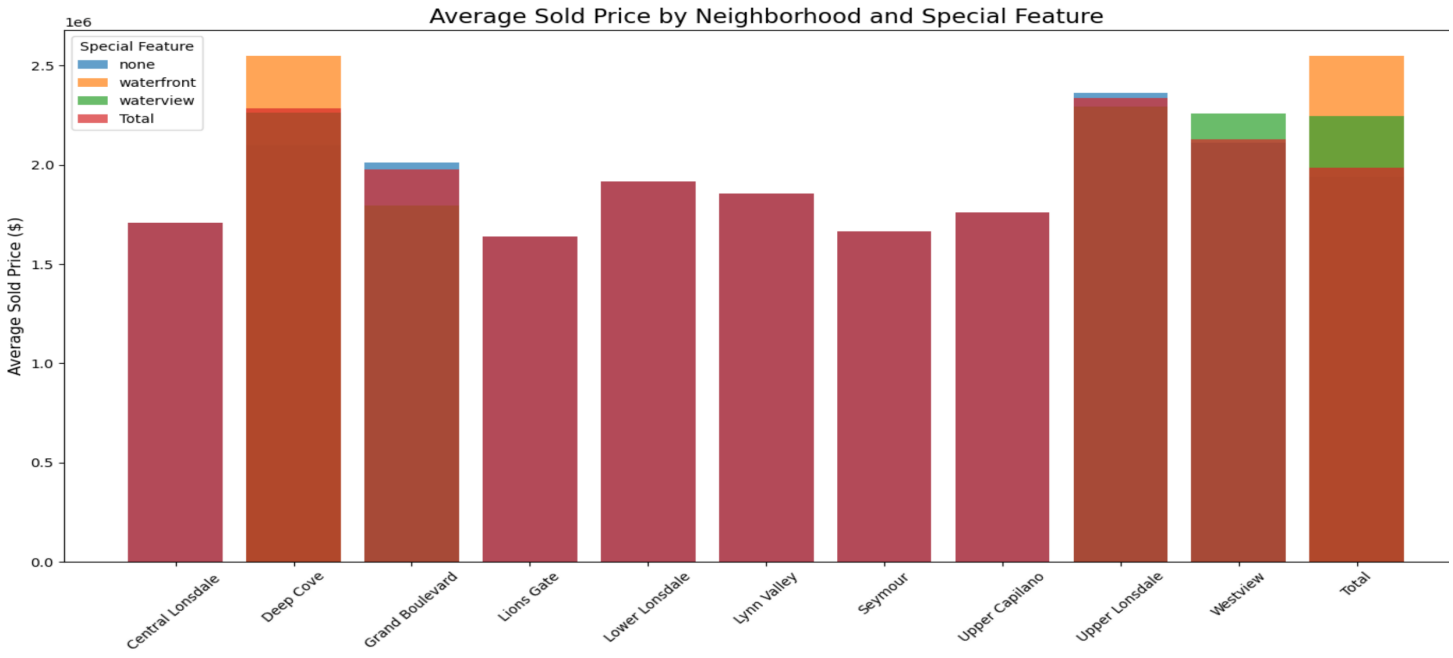
Relationship Between Sold Prices and Neighborhoods

The analysis of sold prices across neighborhoods highlights significant differences in property values. Upper Lonsdale and Deep Cove stand out as the most expensive areas, with average sold prices exceeding $2.2 million. Also, neighborhoods such as Lions Gate and Central Lonsdale exhibit lower average prices, around $1.6 to $1.7 million. This suggests that property values are closely tied to location and neighborhood desirability, with premium areas like Deep Cove benefiting from unique features such as waterfront properties.

Influence of Pools on Sold Prices

The presence of a pool also contributes to higher property values, though its impact is less pronounced compared to other factors like location or special features. Properties with pools, such as those in Upper Lonsdale and Grand Boulevard, often sell at a premium due to their added luxury appeal. However, the data suggests that other factors, such as lot size and neighborhood, often play a more significant role in determining overall prices. This highlights how various elements, including pools, location, and special features, work together to shape property values in the market.

Q3 **Construct a pivot table to group the following variables: Neighborhood vs. special feature. Use various statistics tools to describe the relationship between variables. Comment on the result.**



Impact of Special Features on Home Prices

The data highlights that special features such as waterfront and waterview significantly increase property values in North Vancouver. Homes with waterfront access average the highest prices at $2,550,000, followed by waterview properties at $2,245,714. In contrast, homes without these features average $1,935,798, indicating a strong premium for properties with scenic advantages.

Neighborhood Trends

Neighborhood desirability further influences pricing. In the "none" category, Upper Lonsdale ($2,360,714) and Westview ($2,110,714) lead in average prices, reflecting their popularity. Meanwhile, neighborhoods such as Lions Gate ($1,637,500) and Seymour ($1,665,000) have lower averages, likely due to reduced demand or fewer premium homes. For homes with waterfront access, Deep Cove commands the highest average price, reinforcing its appeal for buyers seeking premium locations.

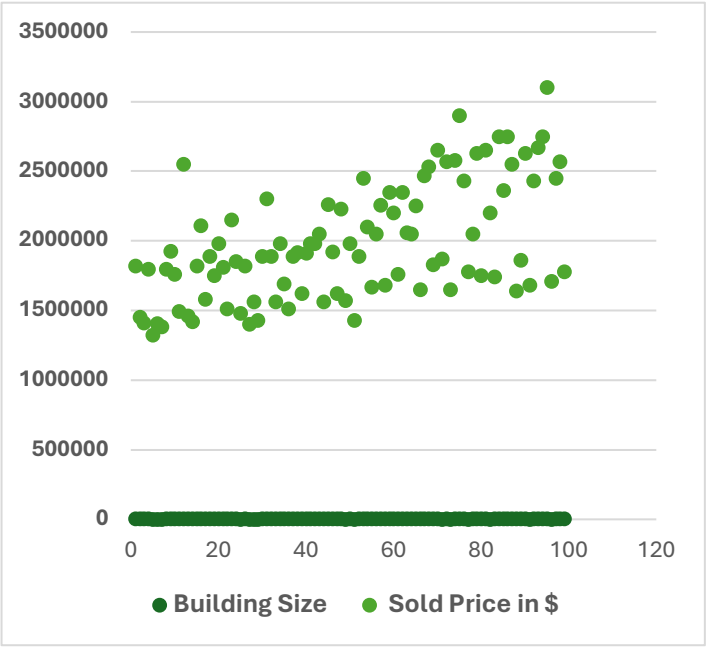
Understanding

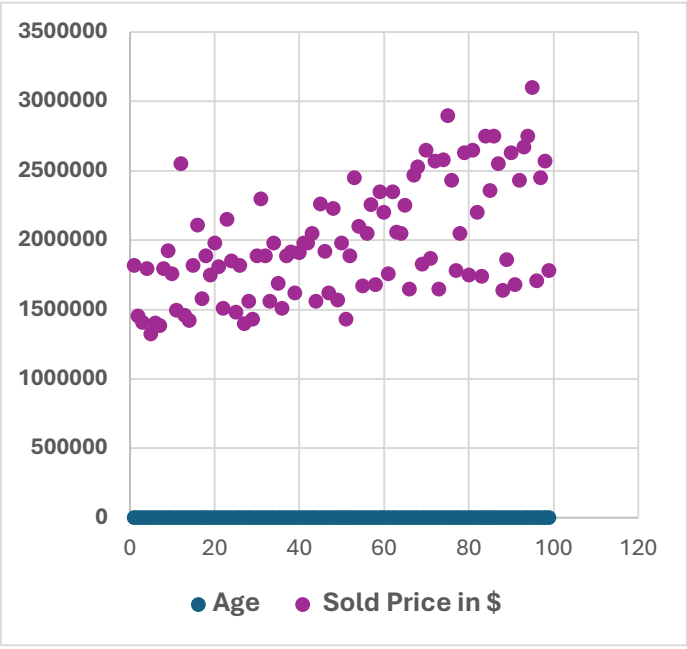
The data clearly shows that special features and neighborhood choice are key drivers of housing prices. Properties with waterfront or waterview features, particularly in neighborhoods like Deep Cove and Upper Lonsdale, consistently fetch higher prices, making them attractive for buyers willing to invest in premium real estate.

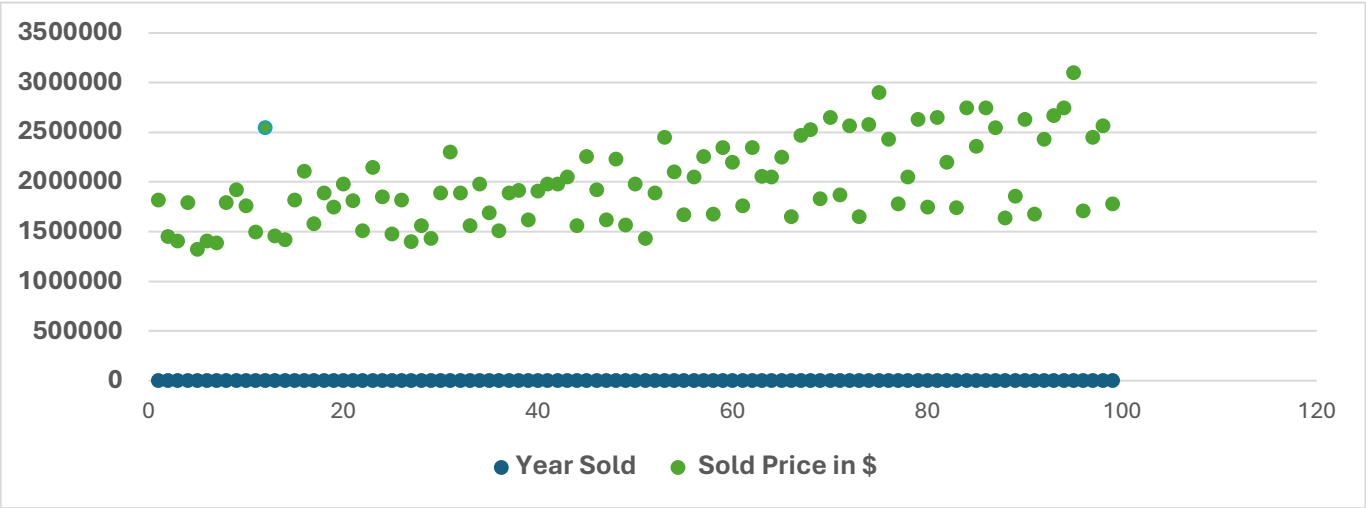
Q4 Which of the variables (year sold, age, building size) is the most and least highly correlated with price?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Sold Price in $   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.739
Method:                 Least Squares   F-statistic:                     93.67
Date:                Sun, 05 Jan 2025   Prob (F-statistic):           2.83e-28
Time:                        16:39:14   Log-Likelihood:                -1352.5
No. Observations:                  99   AIC:                             2713.
Df Residuals:                      95   BIC:                             2723.
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const        -1.691e+08   1.51e+07    -11.184      0.000   -1.99e+08   -1.39e+08
Year Sold     8.454e+04   7507.216     11.261      0.000    6.96e+04    9.94e+04
Age           1015.0271   1273.449      0.797      0.427   -1513.089    3543.143
Building Size  294.1597     25.363     11.598      0.000     243.808     344.512
==============================================================================
Omnibus:                        2.084   Durbin-Watson:                   2.118
Prob(Omnibus):                  0.353   Jarque-Bera (JB):                1.489
Skew:                           0.212   Prob(JB):                        0.475
Kurtosis:                       3.426   Cond. No.                     2.39e+06
==============================================================================
```

Descriptive analysis:







Most

Highly Correlated Variable:

The variable "Year Sold" has the highest correlation with price, indicated by its high t-statistic (11.26) and a P-value close to zero (3.45E-19). This shows that "Year Sold" is highly significant in explaining variations in price, suggesting that house prices have generally increased over the years.

Least Highly Correlated Variable:

The variable "Age" is the least correlated with price. It has the lowest t-statistic (0.797) and the highest P-value (0.427), which indicates it is not statistically significant. This means that the age of the property does not have a meaningful impact on price in this model.

Building Size:

The "Building Size" variable also shows a strong correlation with price, with a high t-statistic (11.60) and a very low P-value (6.74E-20). It is a significant predictor of price, reflecting that larger homes tend to sell for higher prices.

Summary:

Year Sold is the most significant predictor of price, followed closely by **Building Size**, while **Age** having the weakest relationship with price.

Q5 **How strongly is age related to price sold?**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Sold Price in $   R-squared:                       0.099
Model:                            OLS   Adj. R-squared:                  0.089
Method:                 Least Squares   F-statistic:                     10.61
Date:                Sun, 05 Jan 2025   Prob (F-statistic):           0.00155
Time:                        16:45:27   Log-Likelihood:                -1415.5
No. Observations:                  99   AIC:                             2835.
Df Residuals:                      97   BIC:                             2840.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.758e+06   8.04e+04     21.854      0.000     1.6e+06    1.92e+06
Age         7339.5409   2253.157      3.257      0.002    2867.648    1.18e+04
==============================================================================
Omnibus:                       10.688   Durbin-Watson:                   1.567
Prob(Omnibus):                  0.005   Jarque-Bera (JB):               11.316
Skew:                           0.823   Prob(JB):                      0.00349
Kurtosis:                       3.176   Cond. No.                         72.2
==============================================================================
```

The variable "**Age**" has a weak relationship with the price sold:

The R Square value:

0.099, meaning only 9.9% of the variation in the price sold is explained by the age of the property. This indicates a relatively weak relationship.

Statistical Significance:

The P-value for Age is 0.002, which is significant at a 95% C-I level. While Age has a statistically significant effect on price, the overall impact is minimal.

Positive Coefficient:

The coefficient for Age is 7,339.54, suggesting a slight positive relationship- on average, price increases by $7,340 for each additional year of property age. 6, this is counterintuitive and could reflect other confounding factors in the dataset.

While Age is statistically significant, its explanatory power is very low and its practical impact on price is minimal compared to other factors like year sold or building size.

Q6 Is there a relationship between the No. of bathrooms and price sold?

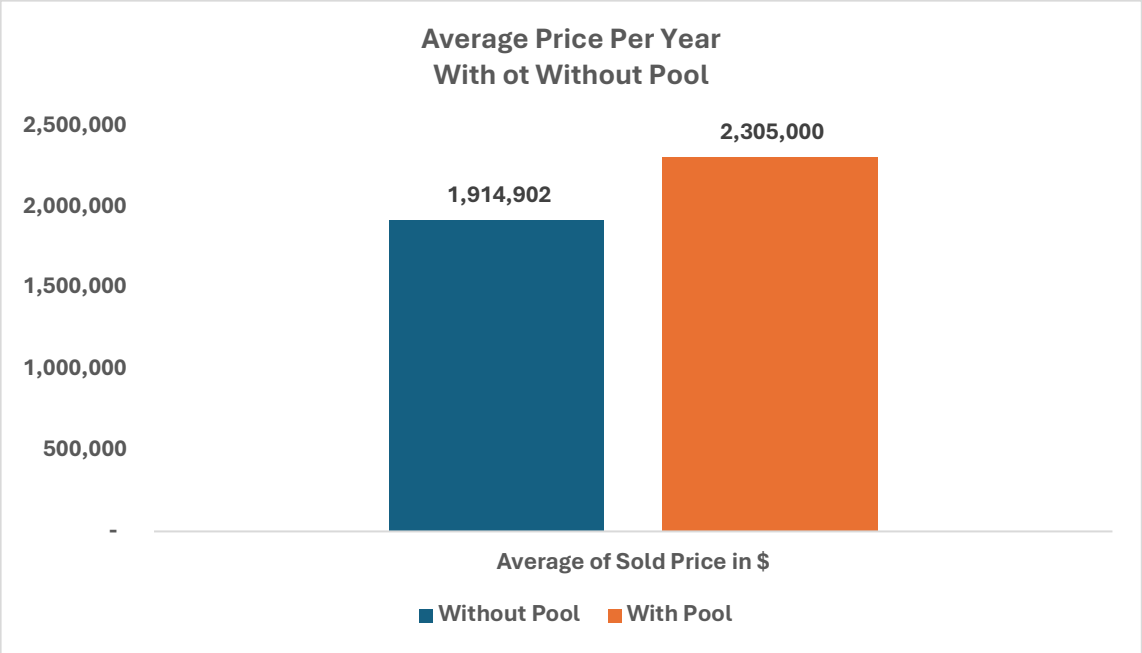
```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Sold Price in $   R-squared:                       0.246
Model:                            OLS   Adj. R-squared:                  0.238
Method:                 Least Squares   F-statistic:                     31.62
Date:                Sun, 05 Jan 2025   Prob (F-statistic):           1.81e-07
Time:                        17:23:17   Log-Likelihood:                -1406.7
No. Observations:                  99   AIC:                             2817.
Df Residuals:                      97   BIC:                             2822.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.213e+06   1.42e+05      8.532      0.000    9.31e+05     1.5e+06
bedrooms    2.046e+05   3.64e+04      5.623      0.000    1.32e+05    2.77e+05
==============================================================================
Omnibus:                        4.728   Durbin-Watson:                   1.090
Prob(Omnibus):                  0.094   Jarque-Bera (JB):                4.484
Skew:                           0.461   Prob(JB):                        0.106
Kurtosis:                       2.515   Cond. No.                         16.2
==============================================================================
```

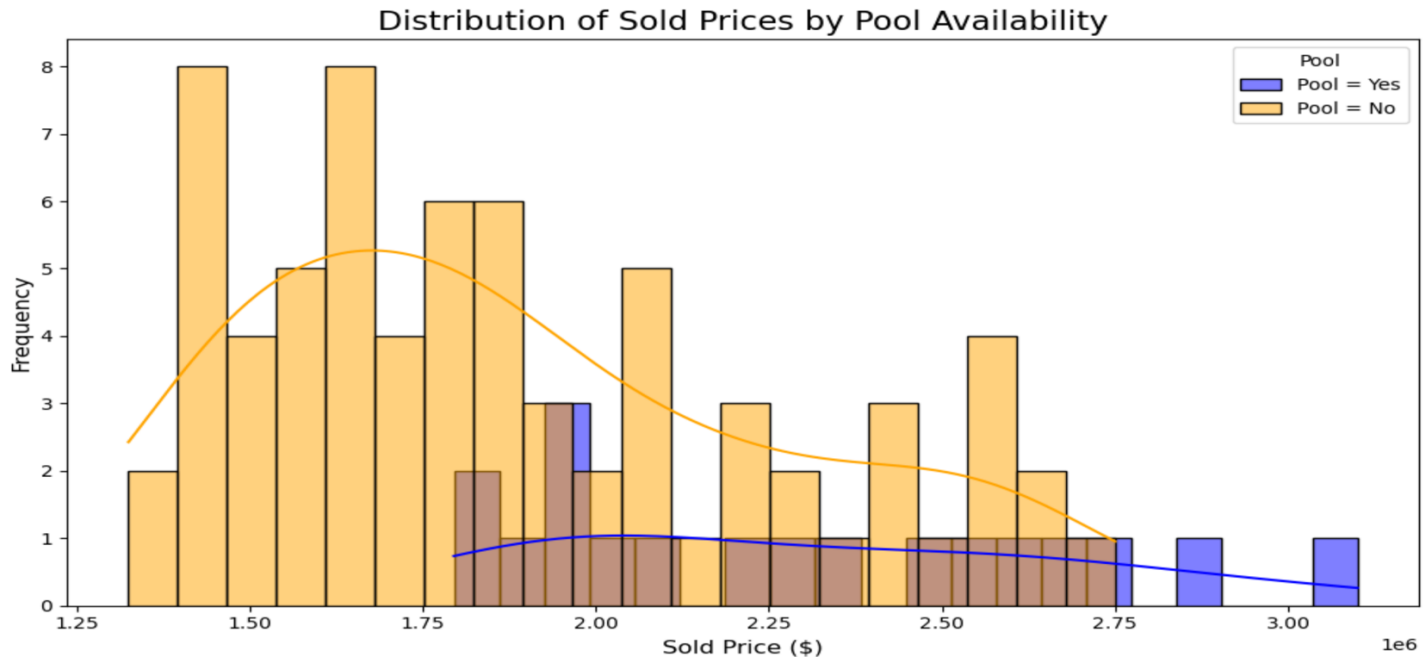
The regression output shows a **moderate positive relationship** between the **number of bathrooms** and the price sold:

1. **R Square Value**: The R Square value is 0.246, meaning 24.6% of the variation in the price sold can be explained by the number of bathrooms. While this indicates a moderate relationship, other factors likely influence price as well.
2. **Statistical Significance**: The P-value is almost 0, which is highly significant. This confirms that the number of bathrooms has a statistically significant effect on the price sold.
3. **Positive Coefficient**: The coefficient for the number of bathrooms is about 204,570**,** indicating that, on average, each additional bathroom increases the price by approximately $204,570**.** This suggests that bathrooms are valued features in a property and contribute meaningfully to higher prices.

There is a statistically significant and moderately strong positive relationship between the number of bathrooms and price sold, additional bathrooms will increase property values.

Q7 **Does having a pool matter? Is there a difference in the mean and distribution of price between the two groups (pool versus no pool)?**



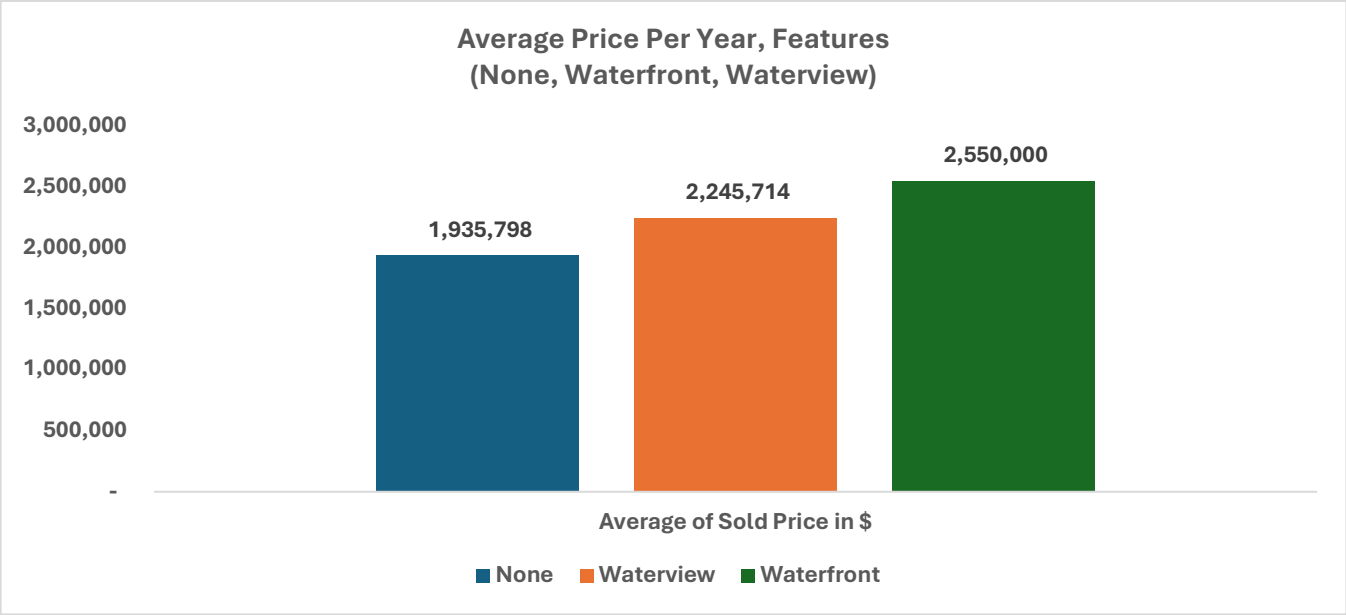


Homes with pools have higher average prices than those without. The mean price for homes with a pool is $2,305,000, compared to $1,914,902 for those without, a difference of approximately $390,000. This trend is also reflected in the median prices, with homes with pools at $2,265,000 and those without at $1,830,000. While the price range for both groups is significant, homes without pools show a slightly larger price variation, indicating more diverse pricing for this category.

The standard deviations for both groups are nearly identical ($388,867 with pool vs. $388,262 without), but homes with pools tend to have higher minimum and maximum prices. The minimum price for homes with pools is $1,795,000, while those without are priced from $1,325,000, and the maximum for homes with pools is $3,100,000. This further confirms that pools have a positive impact on property values, especially at the higher end of the market.

Having a pool significantly increases a home's value. The data shows that homes with pools sell for higher prices and are more likely to reach the upper end of the price range.

Q8 **Does having a special feature matter? Is there a difference in the mean and distribution of price between the two groups (special feature versus no special feature)?**



Impact of Special Features on Property Price

Homes with special features, such as waterfront or waterview properties, have significantly higher average prices compared to those without special features. The average price for homes with these features is around $2,245,714 for waterview and $2,550,000 for waterfront properties, while homes without special features average $1,935,798. This difference is reflected in other statistics, such as the median and mode, with median prices for homes with special features exceeding $2,100,000, compared to $1,855,000 for those without.

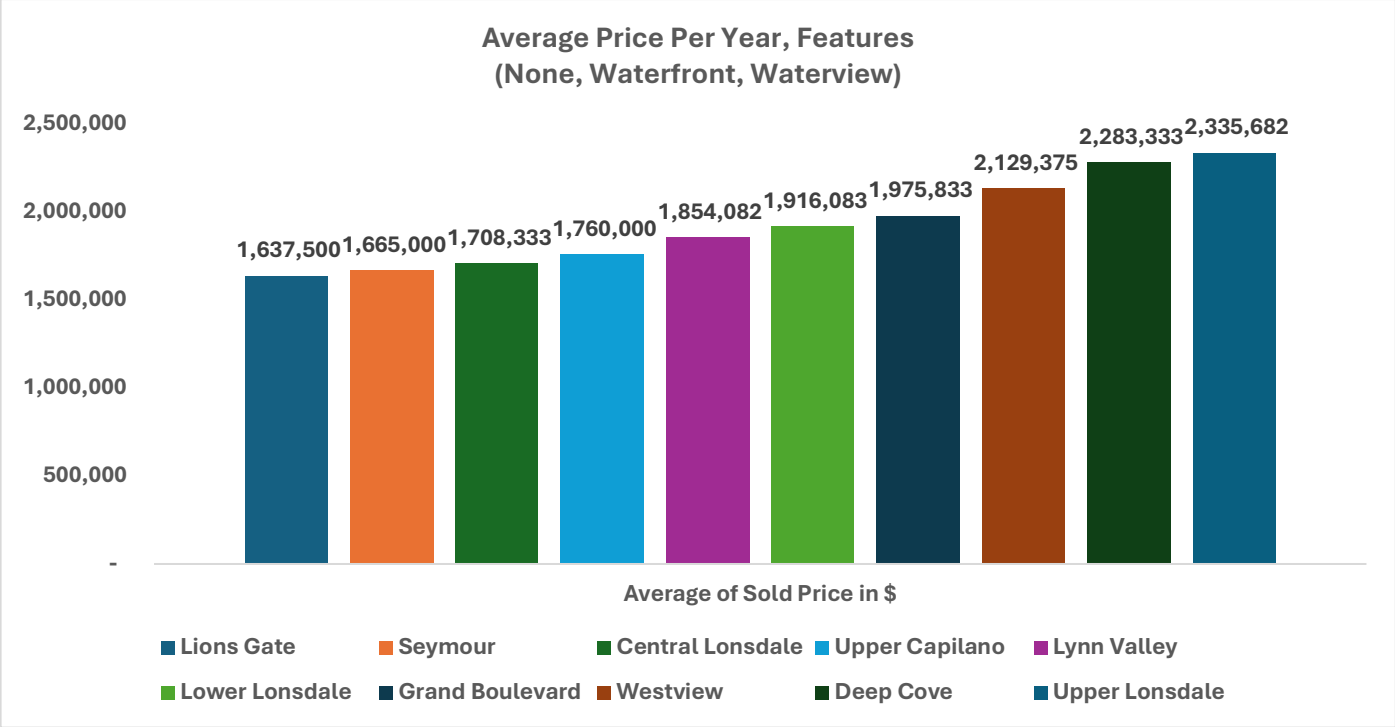
Price Distribution and Variability

The price distribution for homes with special features also shows more variability. Waterfront properties range from $1,795,000 to $3,100,000, and waterview properties from $1,795,000 to $2,750,000, while homes without special features range from $1,325,000 to $2,750,000. Homes with special features exhibit slightly higher standard deviations, indicating greater price variability. This confirms that special features like waterfront and waterview attributes significantly enhance property value.

Summary

The data clearly shows that having a special feature, such as a waterfront or waterview, adds considerable value to a property, both in terms of average price and price distribution, highlighting the importance of these features in determining market value.

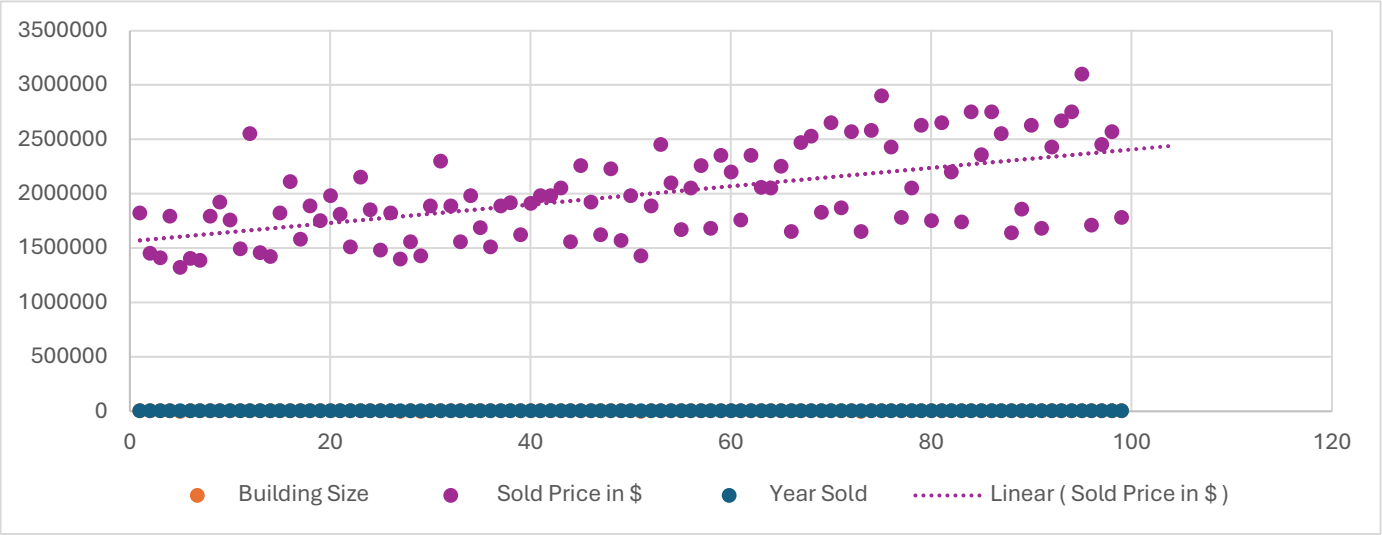
Q9 Is there a difference in price for the different neighbourhoods? Compare the means for the different neighbourhoods.



There is a clear price gradient, with neighborhoods like **Upper Lonsdale** and **Deep Cove** achieving higher prices, reflecting their desirability or other influencing factors. Conversely, neighborhoods such as **Lions Gate** and **Seymour** have lower average prices, indicating possible differences in amenities, location, or market demand.

Q10 Create the best multiple regression model you can to predict price.

Regression Statistics	
Multiple R	0.863519685
R Square	0.745666247
Adjusted R Square	0.740367627
Standard Error	211419.5783
Observations	99



Regression Model Summary

The multiple regression model with Year Sold and Building Size as the predictors shows a strong fit with an R-squared value of 0.746, meaning that approximately 74.6% of the variance in the sale price can be explained by these two variables. The model has a high Multiple R of 0.864, indicating a strong positive relationship between the predictors and the price. The Standard Error of 211,420 indicates that the predictions may deviate from actual prices by about this amount.

Key Variables and Significance

Year Sold: This variable has a coefficient of 85,683, indicating that for each additional year, the price of a home increases by this amount on average. It is highly significant with a p-value of 4.49e-20, well below the typical significance threshold of 0.05. Building Size: With a coefficient of 299.67, this suggests that for each additional square foot, the price increases by $299.67. This variable also shows a high level of significance, with a p-value of 1.91e-21.

This regression model is robust, with significant predictors and strong explanatory power. The most influential variables in predicting price are Building Size and Year Sold, which are both statistically significant.