



# ***Unlocking Insights in Real Estate:***

*A Regression Analysis of House  
Sales Trends in King County*

Presented by;

Elsie Ochieng,  
Richard Taracha,  
Cindy King'ori,  
Peter Muthoma

# Contents

- 01. Introduction**
- 02. Data Preparation and Preprocessing**
- 03. Exploratory Data Analysis**
- 04. Regression Modeling**
- 05. Results and Findings**
- 06. Conclusion**
- 07. References**





# ***Introduction***



## ***Project Overview***

In this data science project, we embark on a journey to analyze house sales data in King County. Our goal is to uncover valuable insights that can aid stakeholders in making informed decisions in the real estate market.

## ***Business Problem***

The real estate agency we are partnering with faces a critical challenge: how to provide homeowners with guidance on home renovations that can enhance the estimated value of their properties. To address this problem, we will leverage multiple linear regression modeling to explore the relationships between various features and house sale prices.





# ***Data Preparation and Preprocessing***



In this phase, we focus on ensuring that our dataset is well-prepared and suitable for analysis by doing the following:

## ***01. Data Loading***

We began by loading the dataset from the provided `kc_house_data.csv` file using Python's Pandas library.

## ***02. Handling Missing Data***

Identified and addressed missing values in certain columns, such as "waterfront," "view," and "yr\_renovated." to ensure data quality.

## ***03. Feature Engineering***

Created new features, including date-related attributes and a binary "is\_renovated" column to indicate whether a property has been renovated.

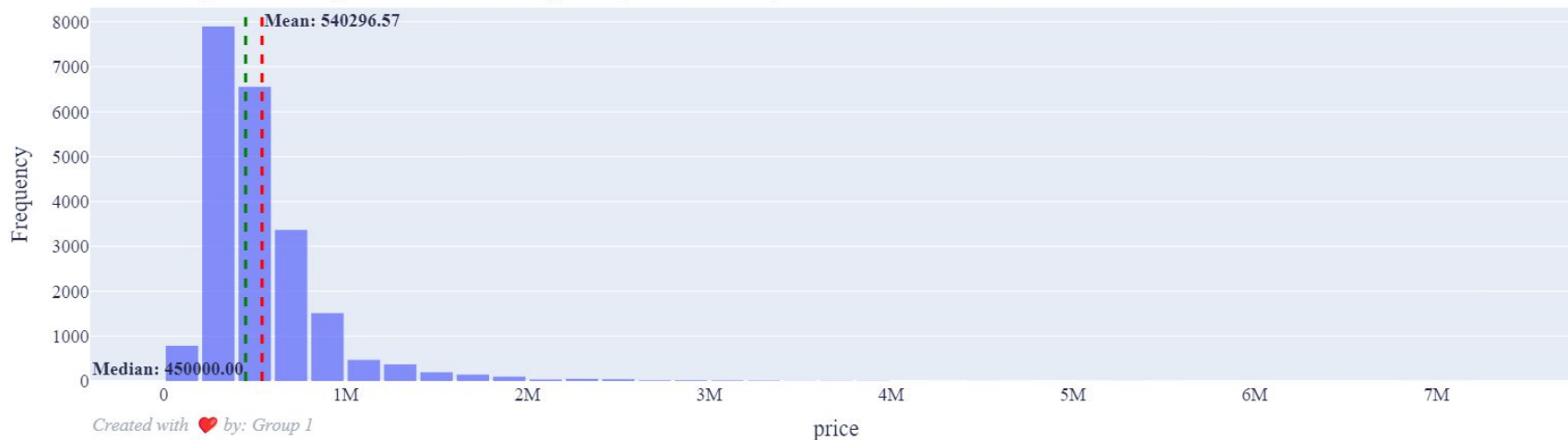


# Exploratory Data Analysis

## The "Price" Histogram Visualisation.

A histogram distribution of the 'Price' variable showing the mean, median and skew of the data.

Histogram Skewness: 4.02



The column of interest was price. It had a right skewed distribution with a positive skewness value of 4.02 indicating that a small number of houses have significantly higher prices in comparison to the majority.



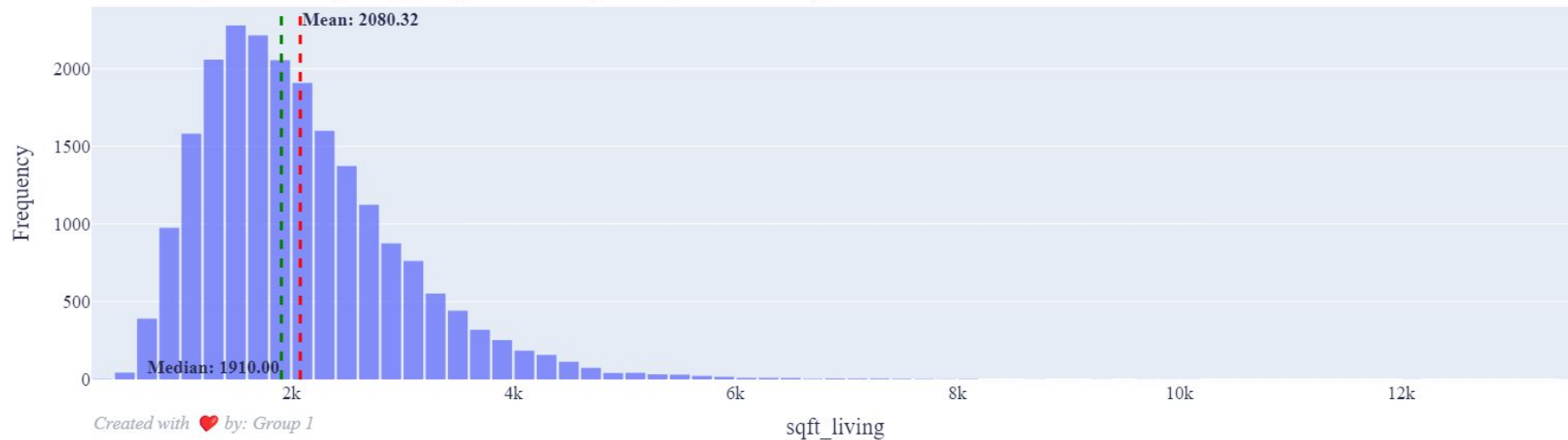
# Exploratory Data Analysis



## The "Sqft\_Living" Histogram Visualisation.

A histogram distribution of the 'Sqft\_Living' variable showing the mean, median and skew of the data.

Histogram Skewness: 1.47

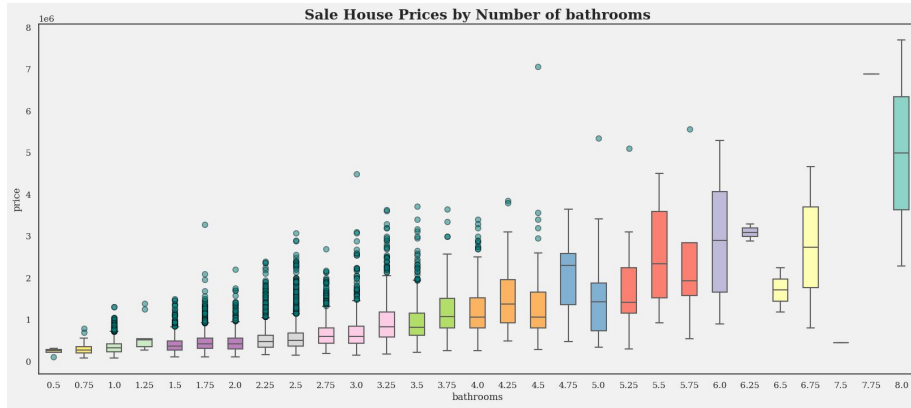


The variable with the highest correlation to the prices of houses was sqft living which referred to the square footage of living space in the home.

The distribution was right skewed distribution with a positive skewness value of 1.47. This means that there are a few houses with very spacious living areas.

# Exploratory Data Analysis

Other variables that we considered key to the pricing of homes were the number of bathrooms and number of bedrooms.



For bedrooms the trend was that houses with more bedrooms tend to fetch higher prices than those with few bedrooms.



Houses with more bathrooms generally command higher prices. The trend is clear until around 6-7 bedrooms beyond which price variations become more scattered






# **Regression Modelling – Multiple Linear Regression using Ordinary Least Squares (OLS)**

## **Model Evaluation**

### **Interpreting the Model Metrics**

- **R-squared:** This is 0.695, which means that about 69.5% of the variance in the dependent variable (price) can be explained by the independent variables in the model.
  - **F-statistic and Prob (F-statistic):** The F-statistic tests whether at least one predictor variable has a non-zero coefficient. A low p-value (here, 0.00) rejects the null hypothesis that all predictor coefficients are zero, meaning at least some predictors are significant.
  - **P>|t|:** This is the p-value associated with each predictor. A small p-value (typically  $\leq 0.05$ ) indicates strong evidence that the predictor is a meaningful addition to the model. For example, bedrooms, bathrooms, sqft\_living, floors, waterfront, view, condition, grade, yr\_built, lat, long, and year are all statistically significant predictors. However day and day\_of\_week have p-values greater than the alpha value making them statistically insignificant predictors
  - **Condition Number:** The condition number is large ( $3.11e+08$ ), indicating potential issues with multicollinearity or other numerical problems.
- 



# **Regression Modelling (Cont.)**

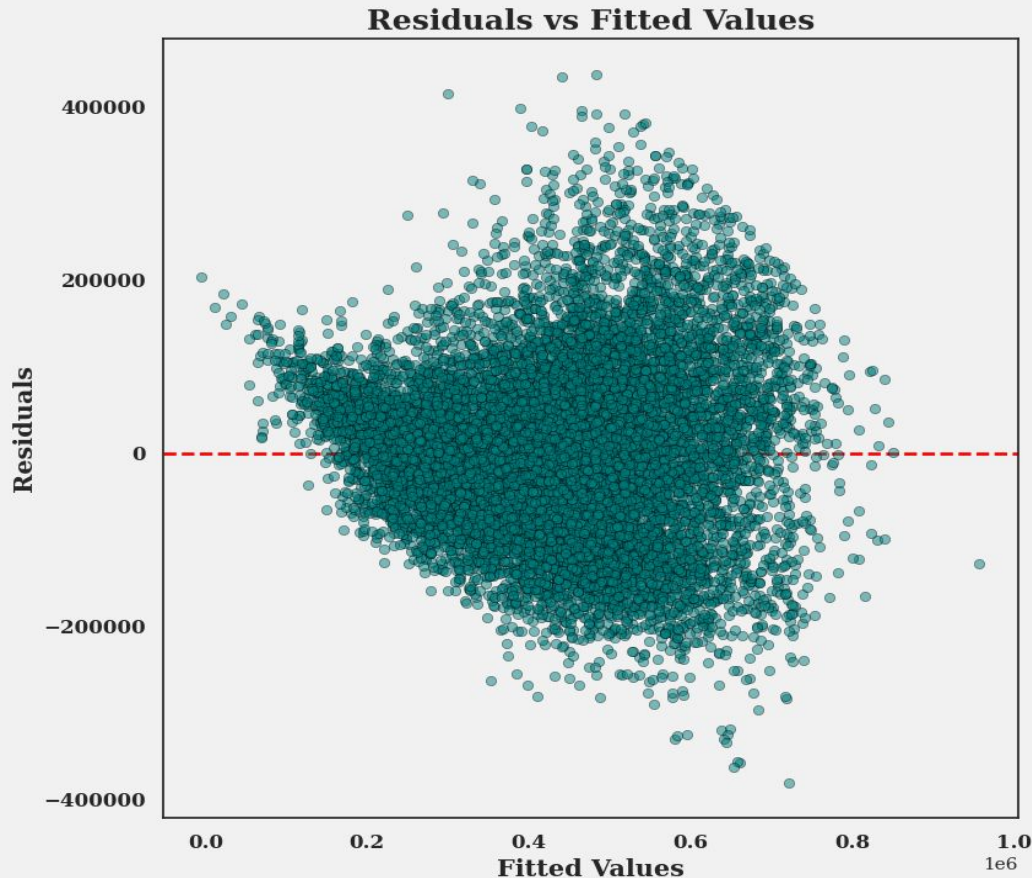
## **Model Evaluation**

### **Interpreting the Model Parameters**

- **Coefficients:** These values represent the change in the dependent variable for a one-unit change in the respective predictor, assuming all other predictors are held constant. For instance, for every additional bedroom (bedrooms), the price decreases by approximately \$34,710.
- **Intercept:**  $-1.04e+08$ , which means that if all other predictors (like bedrooms, bathrooms, sqft\_living, etc.) are zero, the predicted price of a house would be  $-1.04e+08$ .
- However, in practice, the intercept often doesn't have a meaningful interpretation, especially when it doesn't make sense to have all predictors be zero (like in this case, a house cannot have zero bedrooms or zero square footage). It's more useful in adjusting the model's predictions to the scale of the dependent variable. It's also worth noting that the p-value for the intercept is less than 0.05, indicating that it is statistically significant in this model.

# Regression Modelling (Cont.)

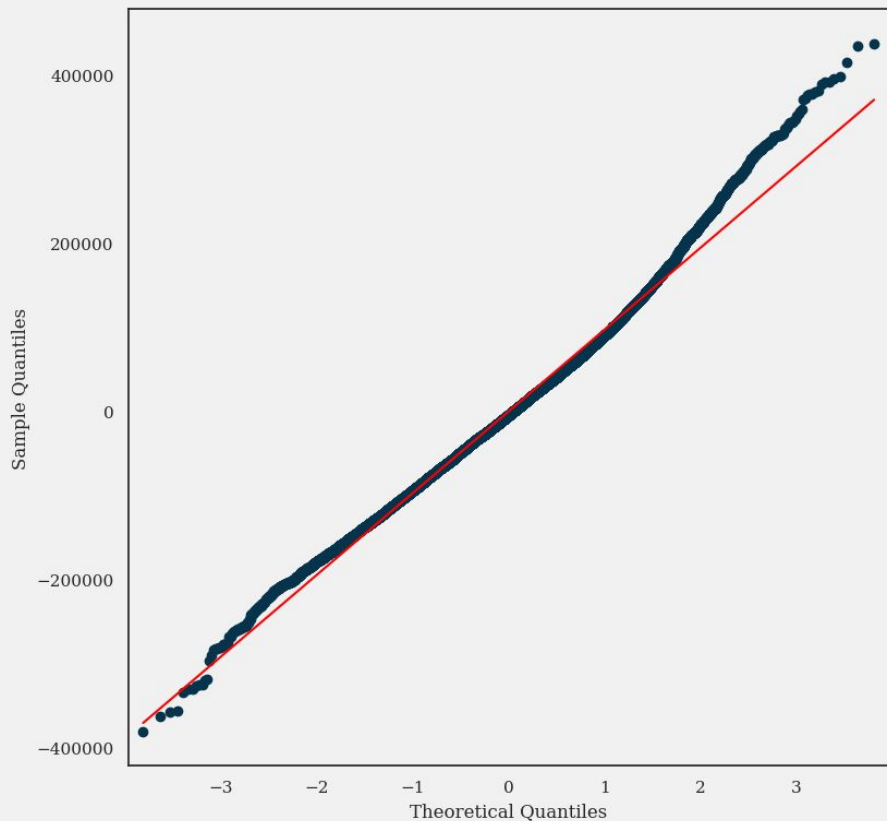
## Communicating Results – Residual Plots



The points in the residual plot are randomly dispersed around the horizontal axis, but a granular pattern is not obvious hence a nonlinear model maybe more appropriate for the research question.

# Regression Modelling (Cont.)

## Communicating Results – QQ Plot



The data points fall above or below the line, meaning the data is not normally distributed.

# **Regression Modelling (Cont.)**

## **Communicating Results – Test for Heteroskedasticity**

- We see that a lot of values are scattered around the mean while a fairly large amount are spread further apart from the mean, meaning that there are no obvious patterns. We performed a test for heteroscedasticity to be certain. We will use Bartlett's Test to test the null hypothesis that the variances in this dataset are homogeneous (equal).
- This is a hypothesis test that establishes a null hypothesis that the variance is equal for all our data points, and the alternative hypothesis is that at least one of the variances is different.
- The test uses the chi-squared distribution to calculate the test statistic and make a decision about the null hypothesis.


**\*\*Test statistic = 1538.5317150354024\*\***

**\*\*p-value = 0.0\*\***

**\*\*We reject the null hypothesis because the variances are unequal\*\***



# ***Results and Findings***

1. The price (target) has many outliers, and it is positively skewed, which makes it hard to generate a proper model to predict the price.
  2. From EDA, we understand several key findings:
    - The living square footage is highly correlated with the price.
    - The grade is highly correlated with the price
    - The number of bathrooms positively correlated with the price.
    - The view also determines the price.
    - Usually, the neighborhood has a similar size of the living space.
    - Houses that have been renovated can sell slightly higher than the houses that are yet to be renovated.
  3. Our model has room for improvement. Its  $R^2$  score is 0.63. This is not a poor result, considering that we are dealing with real world data that has a lot of noise. However, this also implies that the selected input features, cannot account for more than 40% of the variation in housing prices. The linear model would not be suitable model for the specific question as the residuals are not normally distributed.
- 



## ***Conclusion***

In conclusion, All 3 models offer valuable insights into house price determinants but Model 2 stands out with the highest R-squared value, indicating superior explanatory power

Furthermore multicollinearity is a concern in all models, particularly 2 and 3, cautioning against interpreting individual variable effects.

The normality assumptions of residuals are not fully met, suggesting a need for further model refinement and exploration potentially through variable transformations or additional feature engineering to enhance predictive accuracy.



## ***Recommendations***

Based on the analysis of the three regression models and their respective coefficients we recommend that seeing as the factors that significantly affect house prices encompass square footage, grade, waterfront views, geographic location, year built, bathrooms, floors, condition and the number of bedrooms, prospective buyers, sellers should take each into account so as to make informed decisions and maximize on the value of their Real Estate investments



## ***Future Plans***

The future plans and steps for buyers, sellers and investors should always be tailored to their specific goals and circumstances.

Some of the key factors to consider are Budgeting, Property Condition and financing for buyers, Pricing strategy and Marketing for sellers, while investors should consider Diversification, Risk Management and Tax Planning while considering Real Estate

Whether buying a home, selling a property, or investing in real estate, informed decisions and a well-thought-out strategy are essential for success in the dynamic real estate market.





**THANK YOU**

