King County

# House Price Prediction Multiple Linear Regression

Paul Mbbuitu Muriithi

# OVERVIEW

The real estate market is a complex ecosystem influenced by numerous factors, making it crucial to understand the dynamics that drive house prices.

In this project, the aim is to delve into the realm of house sales analysis in King County using multiple linear regression modeling. By leveraging the power of data analysis and machine learning techniques, the relationships between various attributes and the sale prices of houses in the region will be explored so that they can be used to make profitable decisions by a housing development company.

# BUSINESS PROBLEM

The stakeholders are searching for the qualities that lead to higher home sale prices. The aim is to develop an accurate and reliable multiple linear regression model by leveraging the King County House Sales dataset, a model that can predict house prices based on various independent variables.

# MAIN OBJECTIVE

To develop an accurate and reliable multiple linear regression model that can predict house prices based on various independent variables.

# Specific Objectives

## Objective 1

To find out which attributes have a significant impact on house prices in ing County.

## Objective 2

To find out the relationship between the independent variables and the sales prices of houses.

## Objective 3

To find out how accurately house prices can be predicted using the available attributes.

# Hypothesis

## Null hypothesis(Ho)

There is no relationship between our features and our target variable, price.

## Alternative hypothesis(Ha)

There is a relationship between our features and our target variable, price.

.

A significance level (alpha) of 0.05 will be used to decide if to reject or fail to reject the null hypothesis.

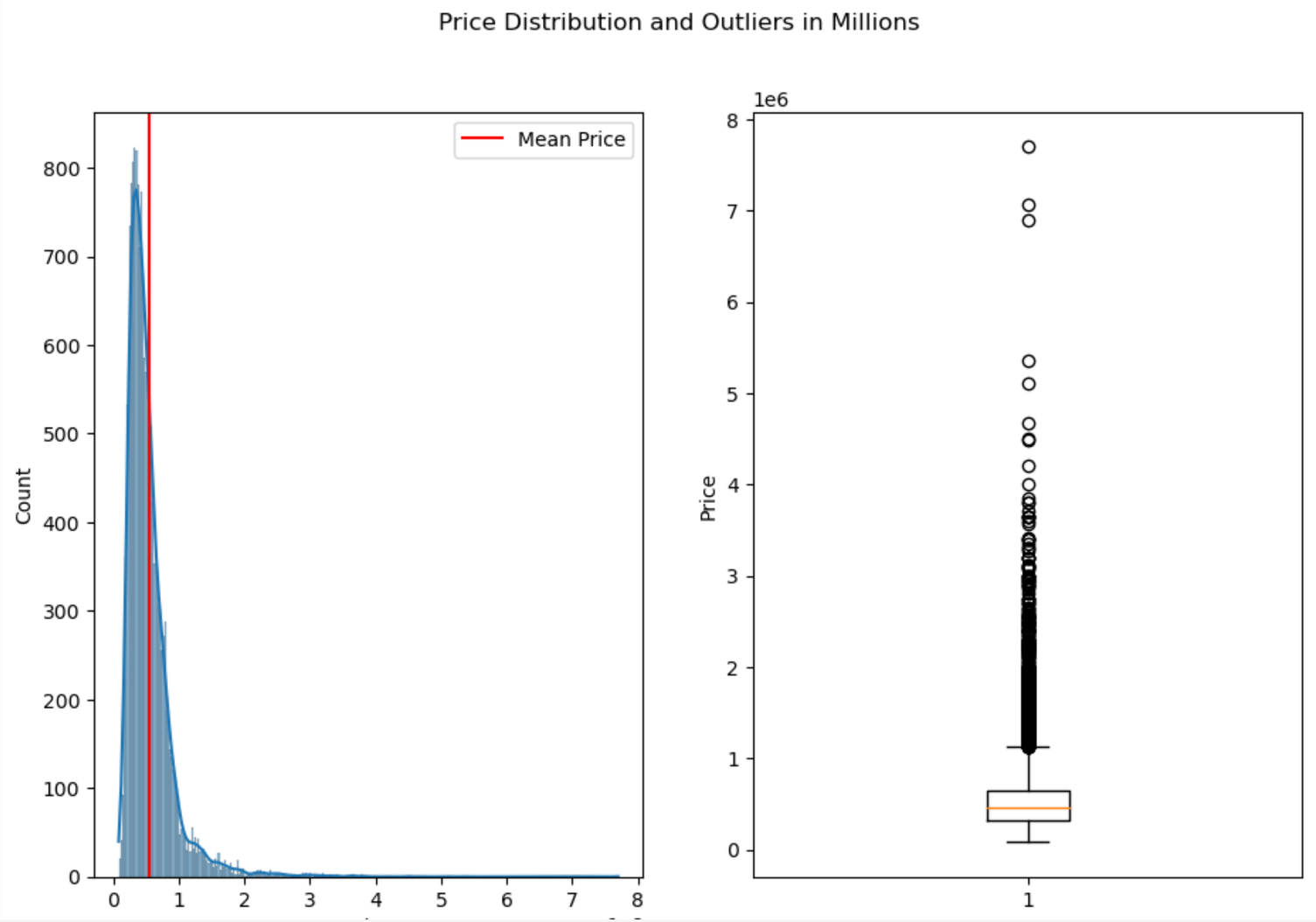

# DATA UNDERSTANDING

King County House Data will be used. The file contains data for 21,597 homes built in King County from 1900 to 2015.

Each house in the set contains information regarding features such as the number of bedrooms/bathrooms, number of floors, square footage, zipcode, condition of the house, the year when the house was built, and more.

# DATA ANALYSIS

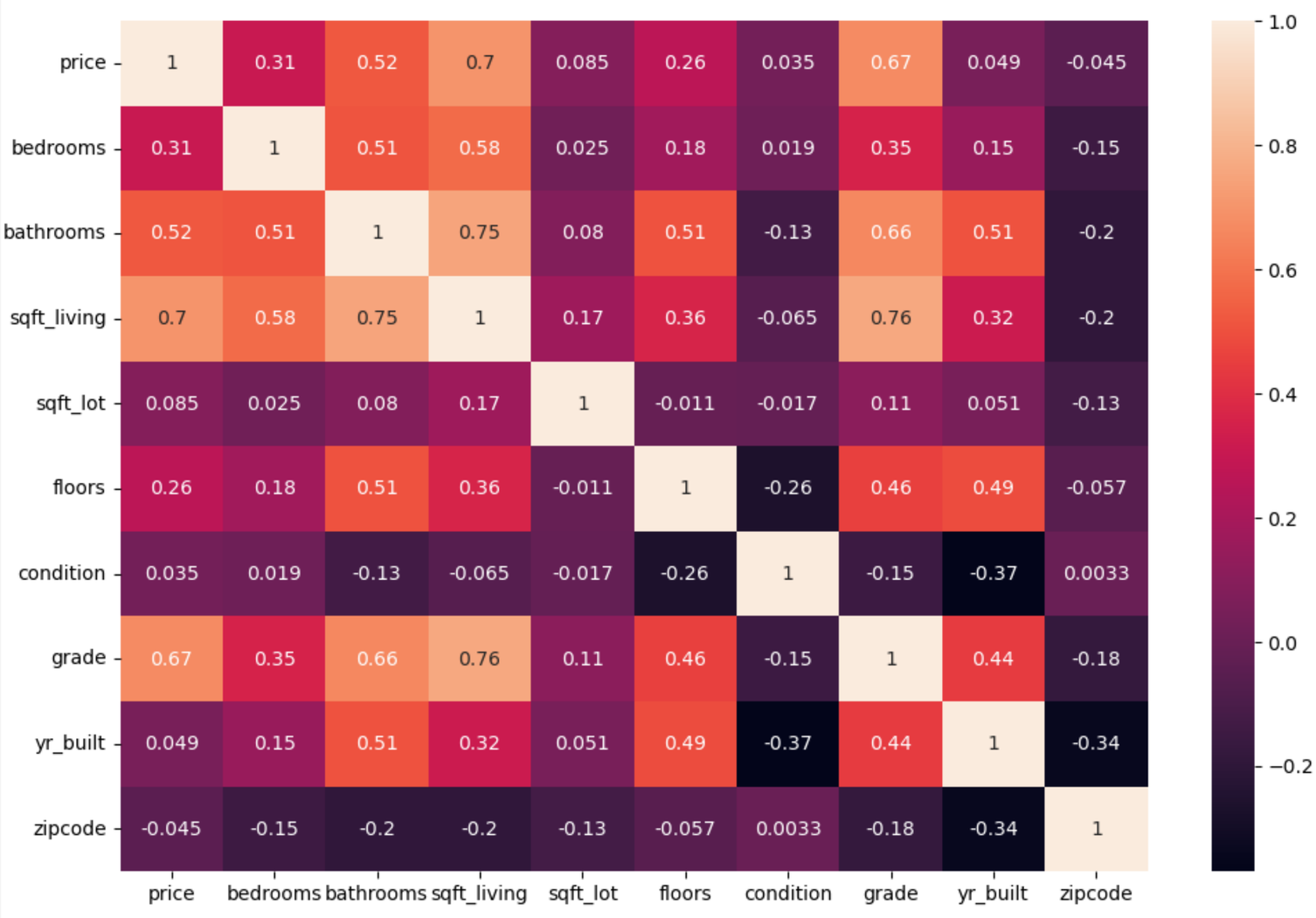
Outliers in our price for houses was checked. Absolute outlires was considered to be any price above 5 million and they were dropped.



# DATA ANALYSIS

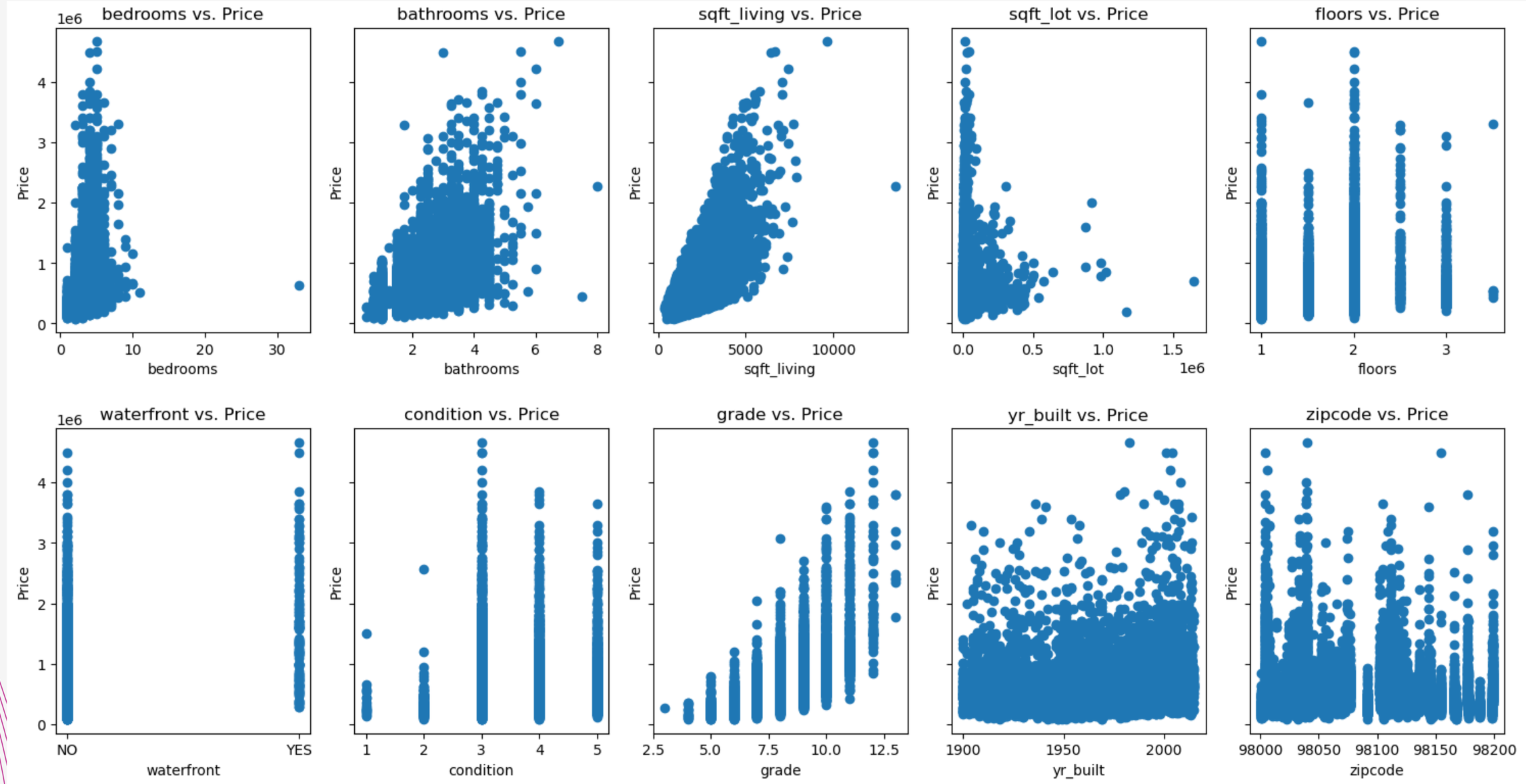
Corrrelations between our feautures was found. A heatmap was generated to display correlations DataFrame.

This will help to determine which pairs of features should not be used together in our model to avoid multicollinearity.



# DATA ANALYSIS

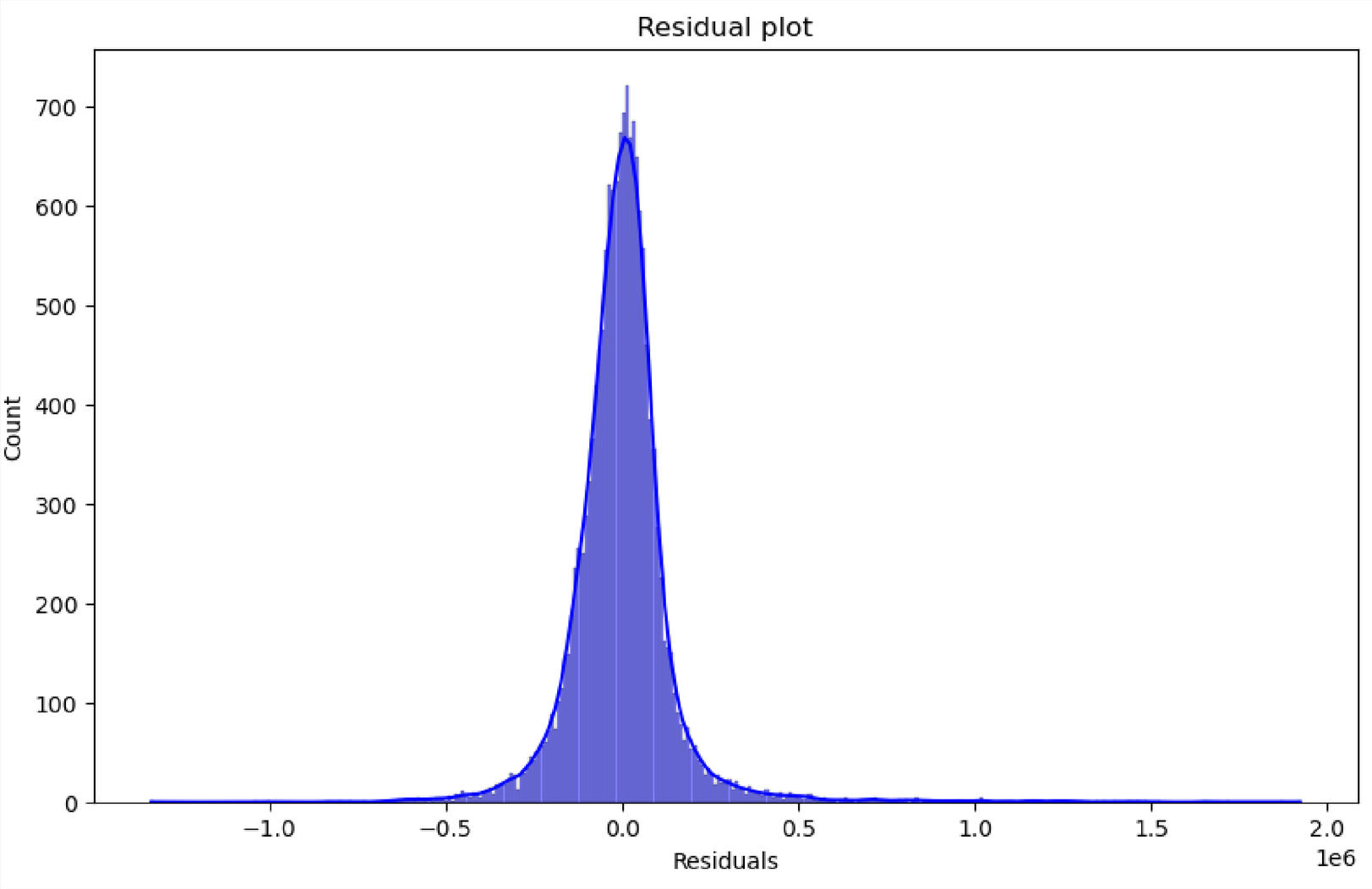
The relationship between out target variable, price, and the predictors, independent variables, was determined.



# DATA ANALYSIS

Differences between the predicted values and the actual values in the final model were found.

This plot provides insights into the distribution of the residuals, helping to assess if they follow a normal distribution and detect any patterns or outliers.



# RESULTS

The multiple linear regression model built has an R-squared value of 0.825, which indicates that the model can explain 83% of the variance of the market-house sale price which is a good sign that the model is effective in predicting the prices.

The model is off by about 91,226 dollars

All of the available features are impactful for inferring and predicting house sale prices and can be considered by home developers in order to increase selling prices.

# RECOMMENDATIONS

- Increase the square footage of living space.
- Attain the highest possible building condition.
- Attain the highest possible building grade.

By following the above recommendations, a housing development company in King County can increase its chances of selling higher-priced homes.

# CONCLUSIONS

The prob(F-statistic) of 0.00 tells us that there is an extremely low probability of achieving these results with the null hypothesis being true, and tells us that our regression is meaningful. Our p-values for our features are well below our alpha or significance level, showing that they are each contributing to the model significantly. With an alpha of 0.05, at a confidence level of 95%, we reject the null hypothesis that there is no relationship between our features and our target variable, price.

# NEXT STEPS

- In the future, reducing noise in the data to improve the accuracy of our model is needed.
- Additionally, it is good to investigate certain features, such as proximity to a top school and coffee shop to see what trends could be discerned from that.

# Thank you for your  Attention!

Incase of any questions or clarifications, feel free to contact me.

## CONTACT

**Paul  Mbuitu Muriithi**

**paulmbuitu@gmail.com**

**+254-704-606-930**