Predictive Modelling

# PREDICTION HOUSE

## Data Mining & Visualization

## INTRODUCTION

The goal of this project is to predict house prices in King County, USA, using data mining techniques. By analyzing features such as square footage, number of bathrooms, and lot size, the project aims to build a model that accurately estimates house prices and visualizes the key factors influencing these prices.

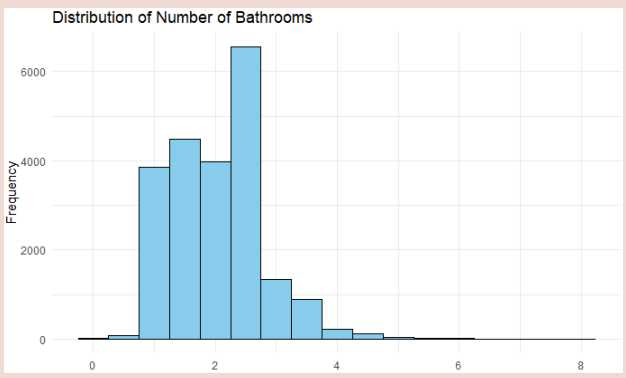
## DATASET

The dataset consists of house sales in King County, USA, which includes features such as price, square footage, number of bedrooms, and more.

## METHODOLOGY

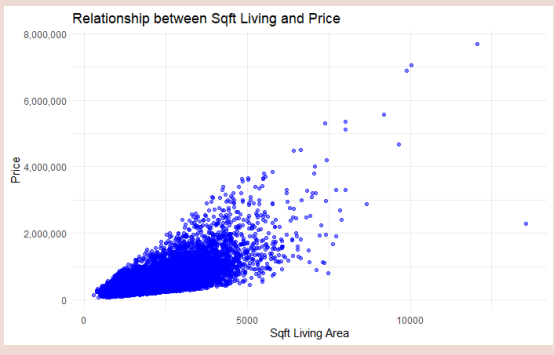
- Data Prepocessing
- Exploratory Data Analysis (EDA) : Distribution & Correlation Analysis
- Modelling : Linear Regression
- Model Diagnostics

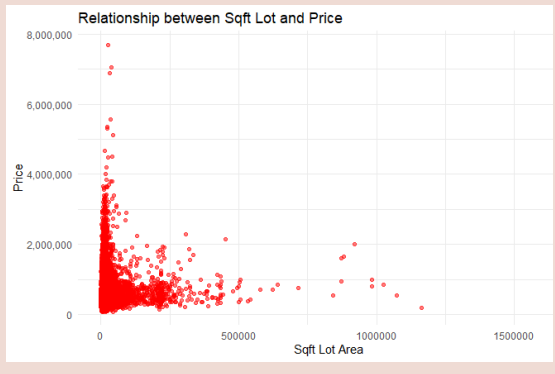
## HISTOGRAM




Distribution of the number of bathrooms showed a right-skewed distribution with most houses having 1-2 bathrooms

## SCATTERPLOTS

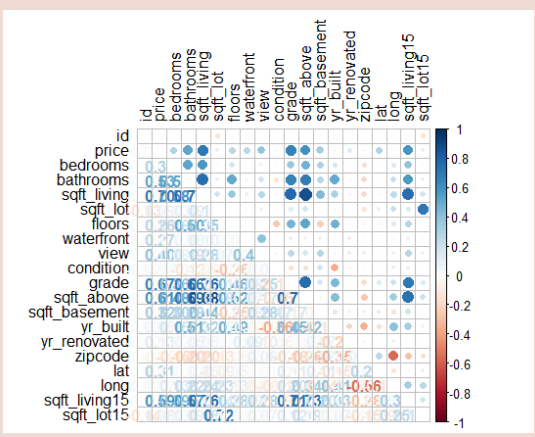



Revealed a positive correlation between sqft_living and price, but no significant correlation between sqft_lot and price.

## CORRELATION MATRIX

The correlation matrix highlighted the key features contributing to the price, reinforcing the significance of sqft_living.

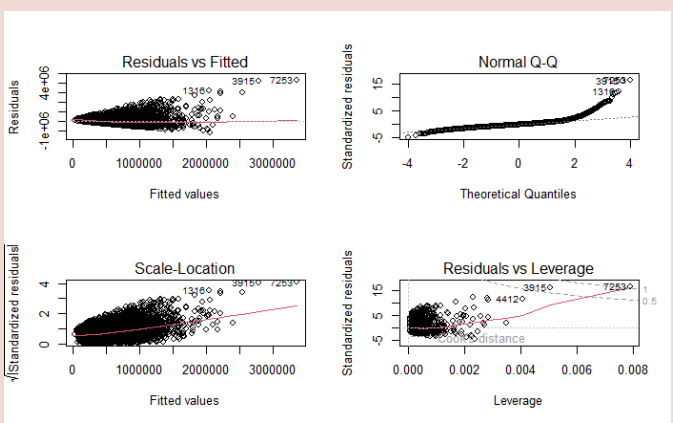



## LINEAR REGRESSION MODEL

```
Call:
lm(formula = price ~ sqft_living, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1263979  -147372   -23668   106145  4346863

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -46065.523   5307.196   -8.68 <0.0000000000000002 ***
sqft_living    282.092      2.338  120.67 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 263600 on 15201 degrees of freedom
Multiple R-squared:  0.4893,    Adjusted R-squared:  0.4892
F-statistic: 1.456e+04 on 1 and 15201 DF,  p-value: < 0.00000000000000022
```

The model indicated that sqft_living is a statistically significant predictor of house price.

## MODEL DIAGNOSTICS




The model diagnostics confirmed that the residuals followed a normal distribution, and there was no evidence of heteroskedasticity.

## PREDICTION TEST

| price <dbl> | predicted_price <dbl> |
|---|---|
| 538000 | 678909.7 |
| 604000 | 506833.9 |
| 510000 | 427848.2 |
| 291850 | 252951.5 |
| 662500 | 958180.3 |
| 650000 | 786104.5 |

While the model captures some trends in the data, there may still be room for improvement to enhance prediction accuracy.

## CONCLUSION

- ‘sqft_living’ is a strong predictor of house prices in King County.
- ‘sqft_living’ strongly influences price, other factors like lot size have less impact.
- The model performs well but needs improvement.