GROUP 11 PRESENTATION

King County House Sales analysis

Welcome To Presentation

GROUP MEMEBERS

BRIAN KARIUKI
ISAAC WANG'ANG'A
FAITH MUTISYA
PAMELA CHEBII
JENIFFER GATHARIA

PROJECT OVERVIEW

- ✓ Understanding the **factors influencing home prices is crucial in King County's vibrant housing market**. This project analyzes the King County House Sales dataset using linear and multi-linear regression models to uncover key determinants of housing prices.
- ✓ The **goal is to provide actionable insights** for a real estate agency to aid in decision-making for real estate professionals, homeowners, and potential buyers.

Objectives

01

Analyze the Impact of Square Footage on Housing Prices

02

Identify Key Determinants of Housing Prices

03

Develop Predictive Model for House Pricing


Data Understanding

Data understanding

- Source of Data: **King County House Sales Dataset**
- **KEY FEATURES OF DATASET**
  - **'id': Unique identifier for a house.**
  - **'date': Date when the house was sold.**
  - **'price': Sale price of the house.**
  - **'bedrooms': Number of bedrooms in the house.**
  - **'bathrooms': Number of bathrooms in the house.**
  - **'sqft_living': Square footage of the interior living space.**
  - **'sqft_lot': Square footage of the lot.**
  - **'floors': Number of floors in the house.**
  - **'waterfront': Indicates if the house has a waterfront view.**
  - **'view': Quality of the view from the house.**
  - **'condition': Condition of the house.**
  - **'grade': Overall grade of the house.**
  - **'sqft_above': Square footage of the house excluding the basement.**

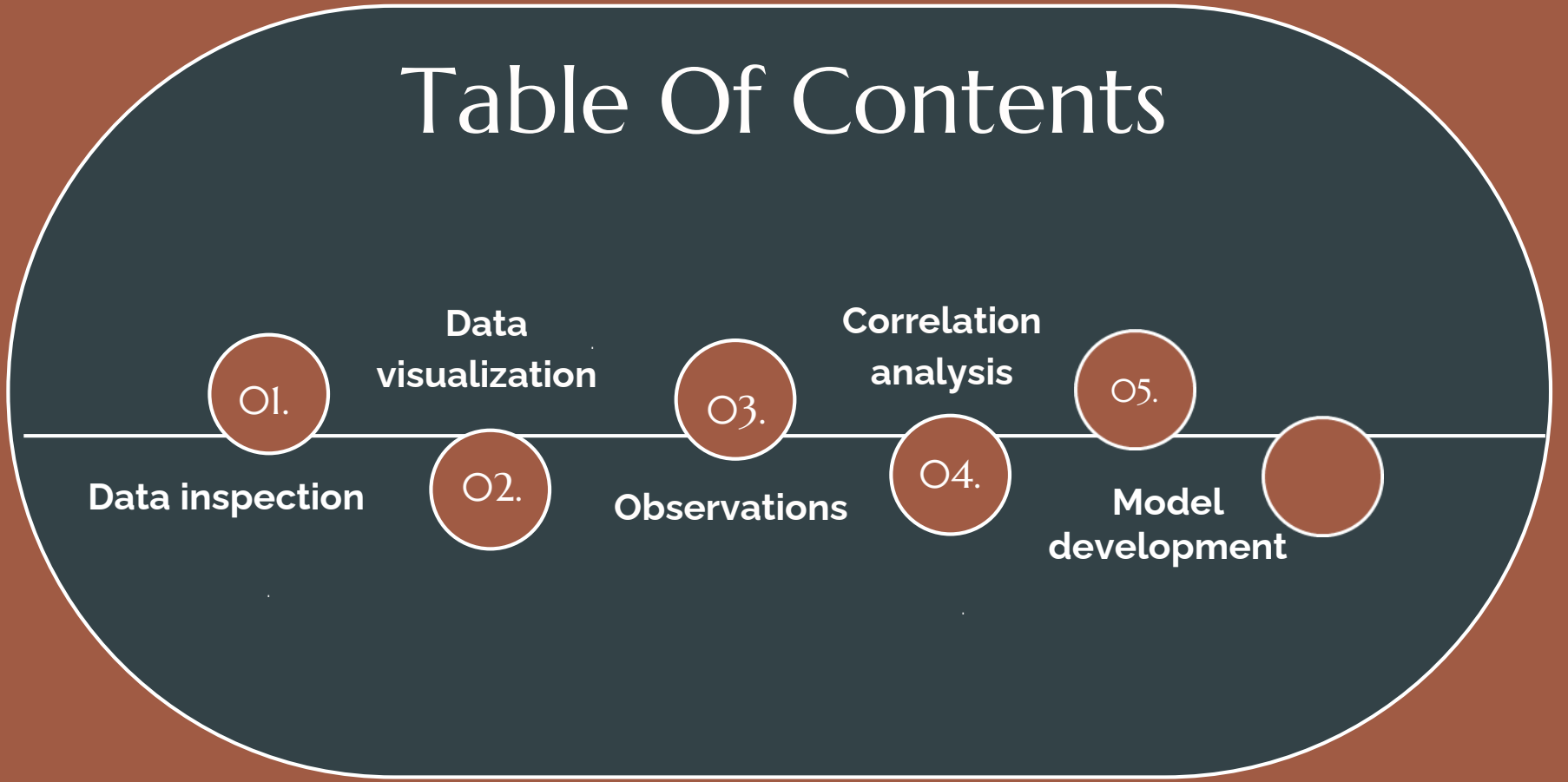
The above mentioned are just but a few datasets that were in the data. Number of Rows (Observations):There are 21597 rows in the DataFrame.-. Number of Columns (Variables)-There are 21 columns (or variables) in the DataFrame.

EDA AND DATA CLEANING

Objectives

1. Data inspection
2. Data Visualization
3. Observations from visualization
4. Correlation Analysis

Table Of Contents



Data Inspection

Checked for Missing Values:

- **Waterfront:** 2,376 missing values.
- **View:** 63 missing values.
- **Year Renovated:** 3,842 missing values.

Action Taken:

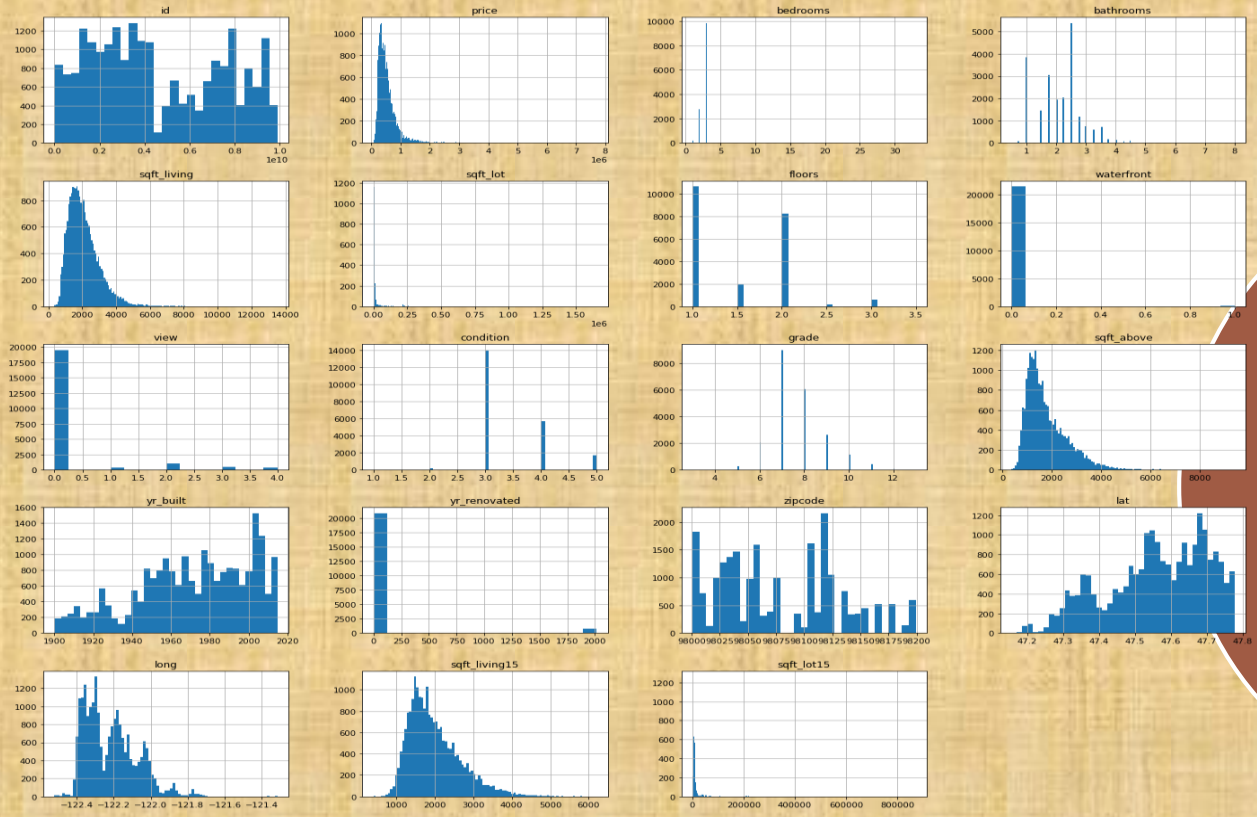
- We filled missing values with the mean of their respective columns to maintain dataset integrity.

Checked for Duplicates:

- **Result:** No duplicate entries found.

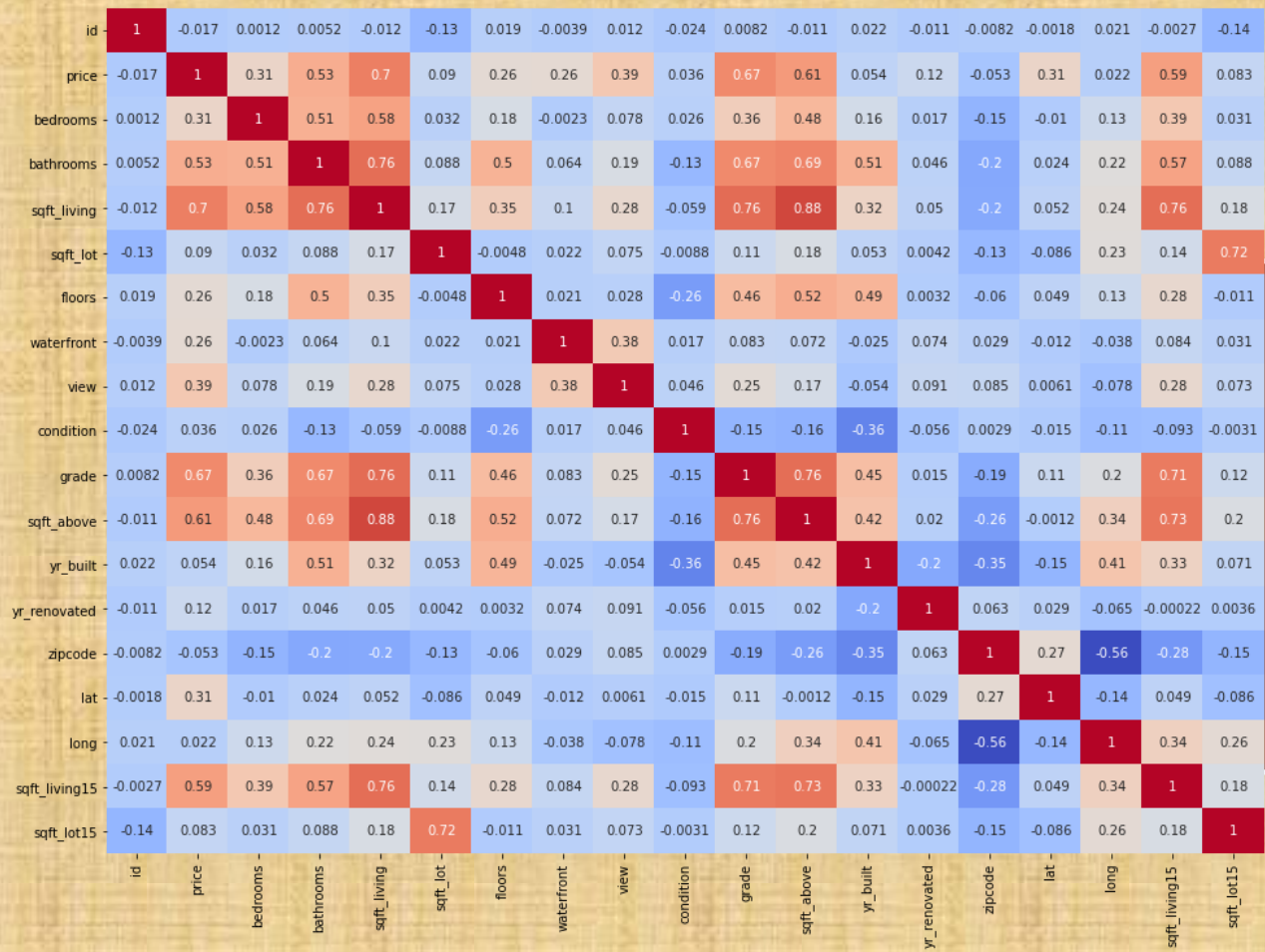
Data Visualization

Histograms:
Created histograms for all variables to understand their distribution.
Key continuous variables analyzed: price, sqft_living, sqft_lot, sqft_above, sqft_living15, sqft_lot15.
Pairplots:
Generated pairplots to examine relationships between variables.
Key relationships explored: price vs. sqft_living, price vs. grade, price vs. sqft_above.


id
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15

Observations

- Price, sqft_living, sqft_lot, sqft_above, sqft_living15 and sqft_lot15 are all continuous data
- Most values are bunched towards the lower end while there are a few very large values
- From the bedrooms feature it appears most houses have around 2 bedrrooms.
- From the bathrooms feature it appears most houses have between two and three bedrooms
- We can see that there is an increase in the number of houses built as time goes on.
- Most houses sold were built in the 2000s
- Most houses have only one floor


id 1 -0.017 0.0012 0.0052 -0.012 -0.13 0.019 -0.0039 0.012 -0.024 0.0082 -0.011 0.022 -0.011 -0.0082 -0.0018 0.021 -0.0027 -0.14
price -0.017 1 0.31 0.53 0.7 0.09 0.26 0.26 0.39 0.036 0.67 0.61 0.054 0.12 -0.053 0.31 0.022 0.59 0.083
bedrooms 0.0012 0.31 1 0.51 0.58 0.032 0.18 -0.0023 0.078 0.026 0.36 0.48 0.16 0.017 -0.15 -0.01 0.13 0.39 0.031
bathrooms 0.0052 0.53 0.51 1 0.76 0.088 0.5 0.064 0.19 -0.13 0.67 0.69 0.51 0.046 -0.2 0.024 0.22 0.57 0.088
sqft_living -0.012 0.7 0.58 0.76 1 0.17 0.35 0.1 0.28 -0.059 0.76 0.88 0.32 0.05 -0.2 0.052 0.24 0.76 0.18
sqft_lot -0.13 0.09 0.032 0.088 0.17 1 -0.0048 0.022 0.075 -0.0088 0.11 0.18 0.053 0.0042 -0.13 -0.086 0.23 0.14 0.72
floors 0.019 0.26 0.18 0.5 0.35 -0.0048 1 0.021 0.028 -0.26 0.46 0.52 0.49 0.0032 -0.06 0.049 0.13 0.28 -0.011
waterfront -0.0039 0.26 -0.0023 0.064 0.1 0.022 0.021 1 0.38 0.017 0.083 0.072 -0.025 0.074 0.029 -0.012 -0.038 0.084 0.031
view 0.012 0.39 0.078 0.19 0.28 0.075 0.028 0.38 1 0.046 0.25 0.17 -0.054 0.091 0.085 0.0061 -0.078 0.28 0.073
condition -0.024 0.036 0.026 -0.13 -0.059 -0.0088 -0.26 0.017 0.046 1 -0.15 -0.16 -0.36 -0.056 0.0029 -0.015 -0.11 -0.093 -0.0031
grade 0.0082 0.67 0.36 0.67 0.76 0.11 0.46 0.083 0.25 -0.15 1 0.76 0.45 0.015 -0.19 0.11 0.2 0.71 0.12
sqft_above -0.011 0.61 0.48 0.69 0.88 0.18 0.52 0.072 0.17 -0.16 0.76 1 0.42 0.02 -0.26 -0.0012 0.34 0.73 0.2
yr_built 0.022 0.054 0.16 0.51 0.32 0.053 0.49 -0.025 -0.054 -0.36 0.45 0.42 1 -0.2 -0.35 -0.15 0.41 0.33 0.071
yr_renovated -0.011 0.12 0.017 0.046 0.05 0.0042 0.0032 0.074 0.091 -0.056 0.015 0.02 -0.2 1 0.063 0.029 -0.065 -0.00022 0.0036
zipcode -0.0082 -0.053 -0.15 -0.2 -0.2 -0.13 -0.06 0.029 0.085 0.0029 -0.19 -0.26 -0.35 0.063 1 0.27 -0.56 -0.28 -0.15
lat -0.0018 0.31 -0.01 0.024 0.052 -0.086 0.049 -0.012 0.0061 -0.015 0.11 -0.0012 -0.15 0.029 0.27 1 -0.14 0.049 -0.086
long 0.021 0.022 0.13 0.22 0.24 0.23 0.13 -0.038 -0.078 -0.11 0.2 0.34 0.41 -0.065 -0.56 -0.14 1 0.34 0.26
sqft_living15 -0.0027 0.59 0.39 0.57 0.76 0.14 0.28 0.084 0.28 -0.093 0.71 0.73 0.33 -0.00022 -0.28 0.049 0.34 1 0.18
sqft_lot15 -0.14 0.083 0.031 0.088 0.18 0.72 -0.011 0.031 0.073 -0.0031 0.12 0.2 0.071 0.0036 -0.15 -0.086 0.26 0.18 1
id price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above yr_built yr_renovated zipcode lat long sqft_living15 sqft_lot15
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Correlation analysis

Done using the correlation heatmap Strong Correlations with Price:

- ‘bathrooms’ (0.53): Indicates a moderate positive correlation, suggesting that houses with more bathrooms tend to have higher prices.
- ‘sqft_living’ (0.70): Shows a strong positive correlation, meaning that larger living areas significantly increase house prices.
- ‘grade’ (0.67): Demonstrates a strong positive correlation, implying that higher quality and better-graded houses are priced higher.
- ‘sqft_above’ (0.61): Reflects a strong positive correlation, indicating that houses with more above-ground living space are more expensive.
- ‘sqft_lot15’ (0.59): Indicates a moderate to strong positive correlation, suggesting that larger lots in the vicinity (nearest 15 neighbors) tend to increase a house’s value

- These insights guide homeowners and real estate professionals in making strategic decisions regarding renovations and investments to enhance property value.


Feature selection

Selected Features:

Target: price

Predictors:- ‘bathrooms’, ‘sqft_living’, ‘grade’, ‘sqft_above’, ‘bedrooms’.

Factors that influenced feature selection

I. Correlation Analysis:- The Correlation heat map revealed key determinants influencing house prices.

II. Pairplots:-Visualized relationships between selected features and target variable.


Data Analysis

Impact of square footage on housing

Method used: **Simple linear regression**
We used ‘sqft_living’ as the predictor variable and analyzed the relationship between ‘sqft_living’ and house prices.

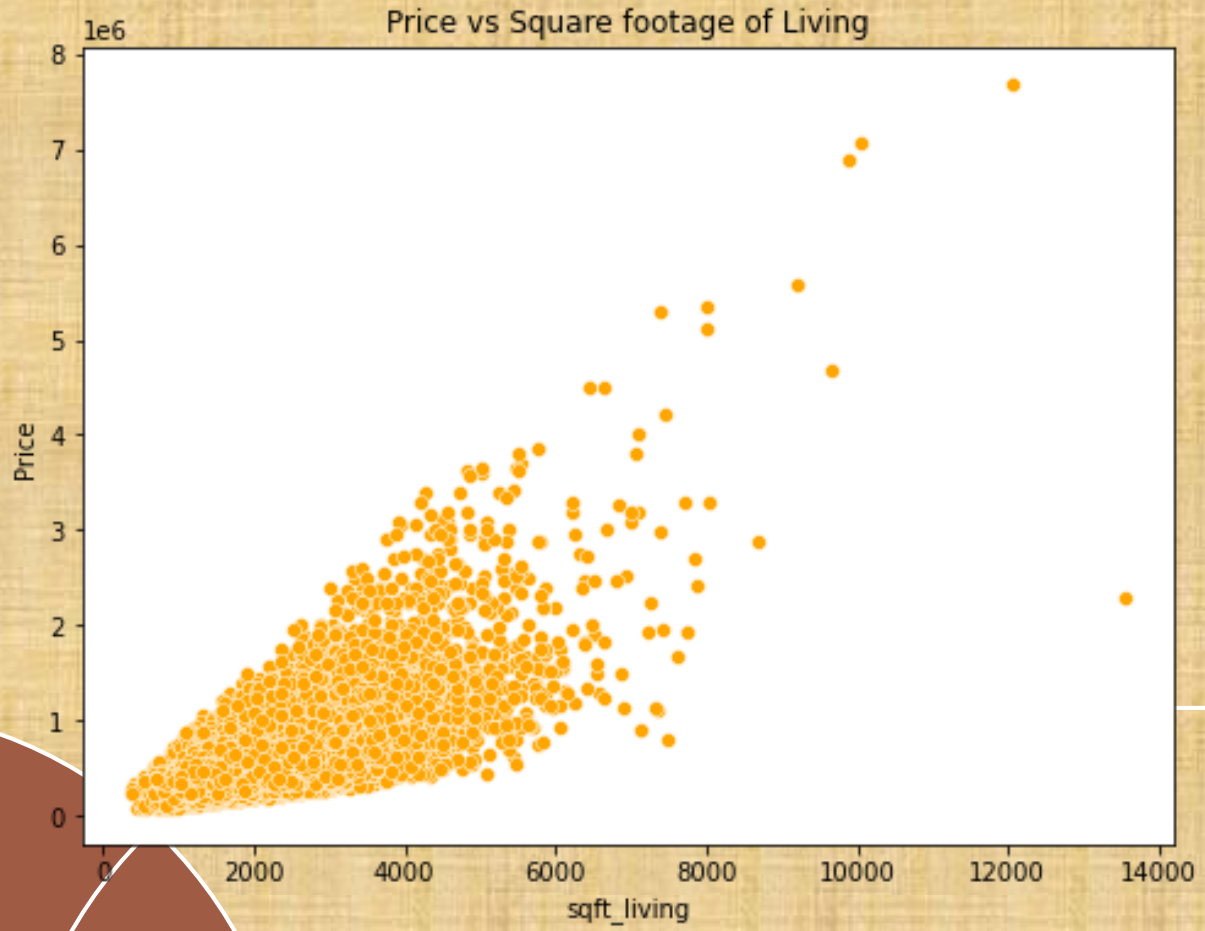
Findings: There was a Strong Positive Correlation.
Correlation coefficient (r) = 0.7019.
Larger living areas significantly increase house prices.

Model Performance: R-squared value: 0.493, indicating that approximately 49.3% of the variance in house prices is explained by ‘sqft_living’.

Scatter Plot and Regression Line
Shows a positive trend between sqft_living and price.
The visual representation confirms the strong positive correlation
The regression line demonstrates a clear positive trend between sqft_living and price.

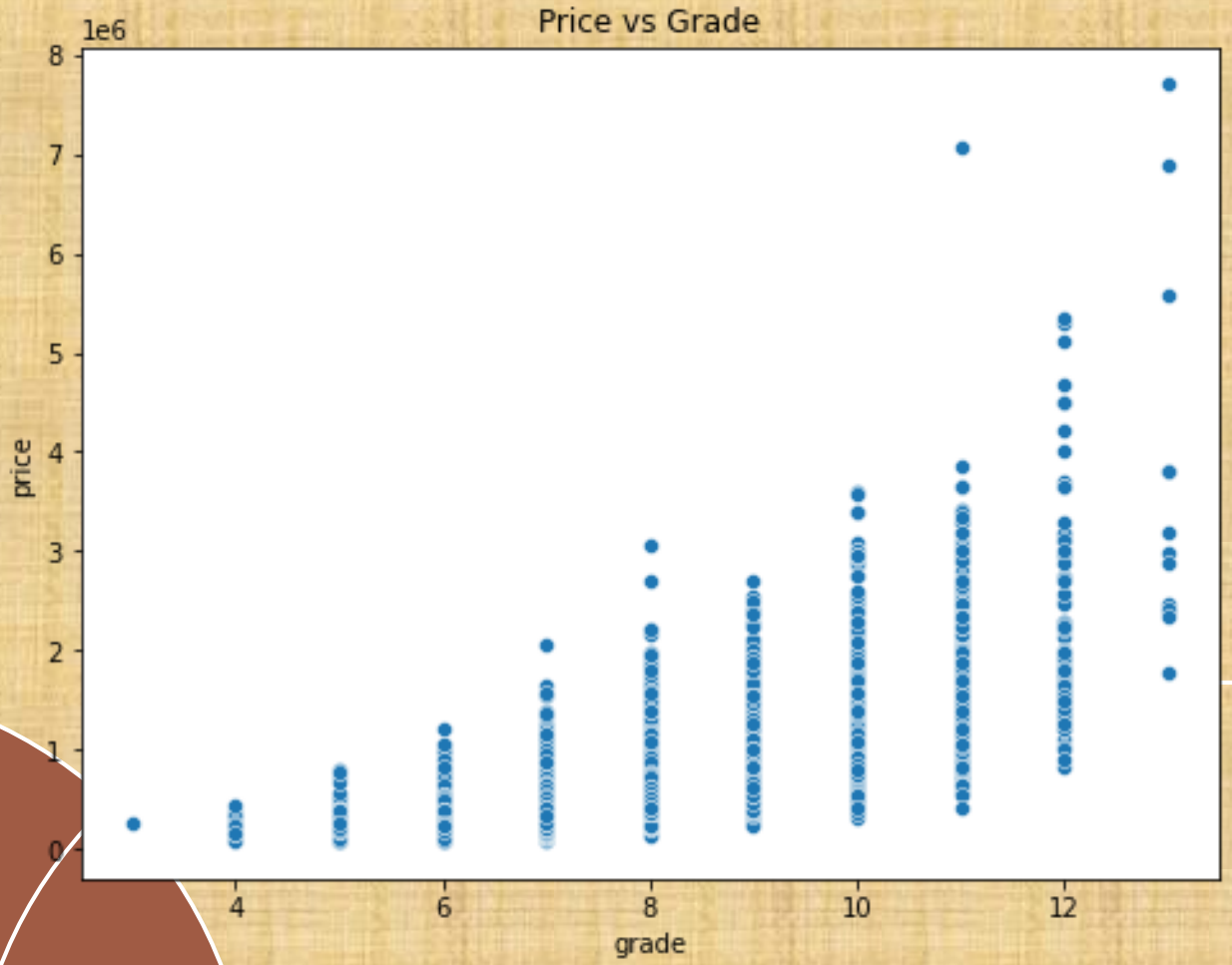
Implications
Expand Living Areas- Investing in expanding living areas is a key strategy to enhance property value.
Homeowners and real estate professionals should consider this when planning renovations or property improvements.

VISUALIZATIONS



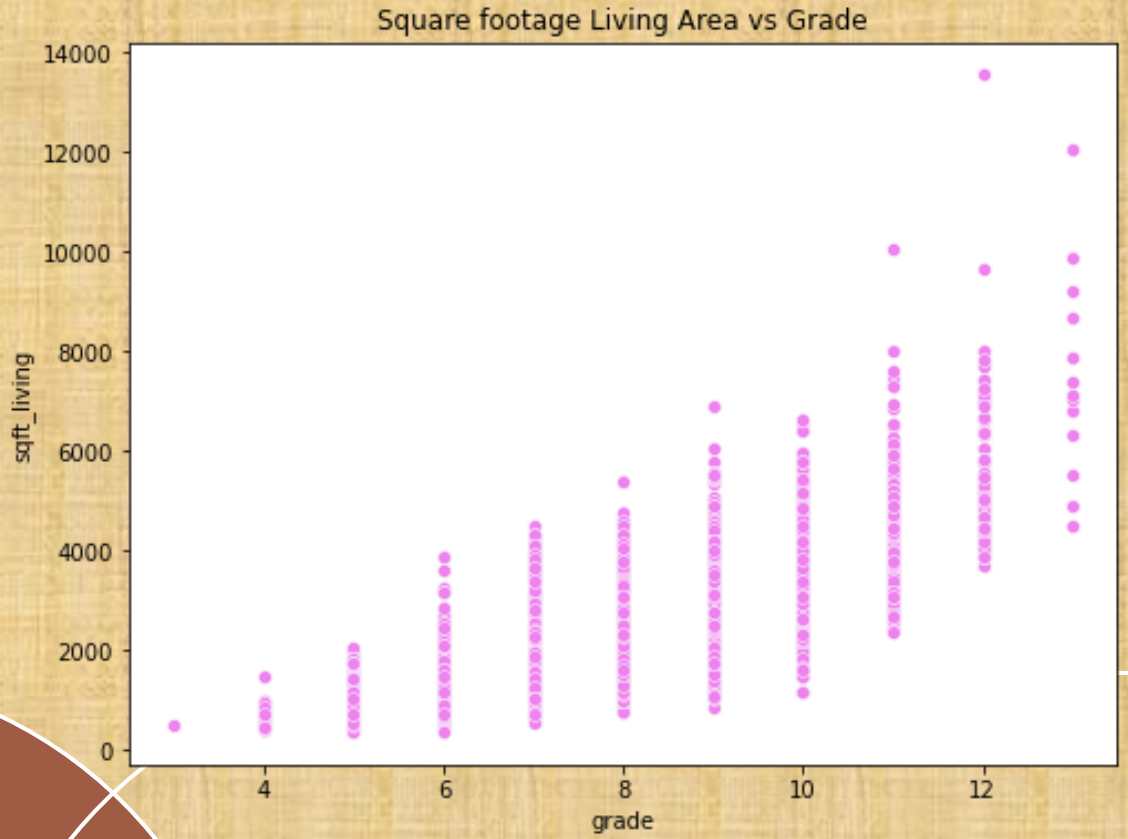
From the Scatter plot we observe that the price and square foot living area have a strong co-relation with a positive trend . As the Square foot living area increases the price of the house also increases.

This implies that houses with larger square foot living areas, command higher prices



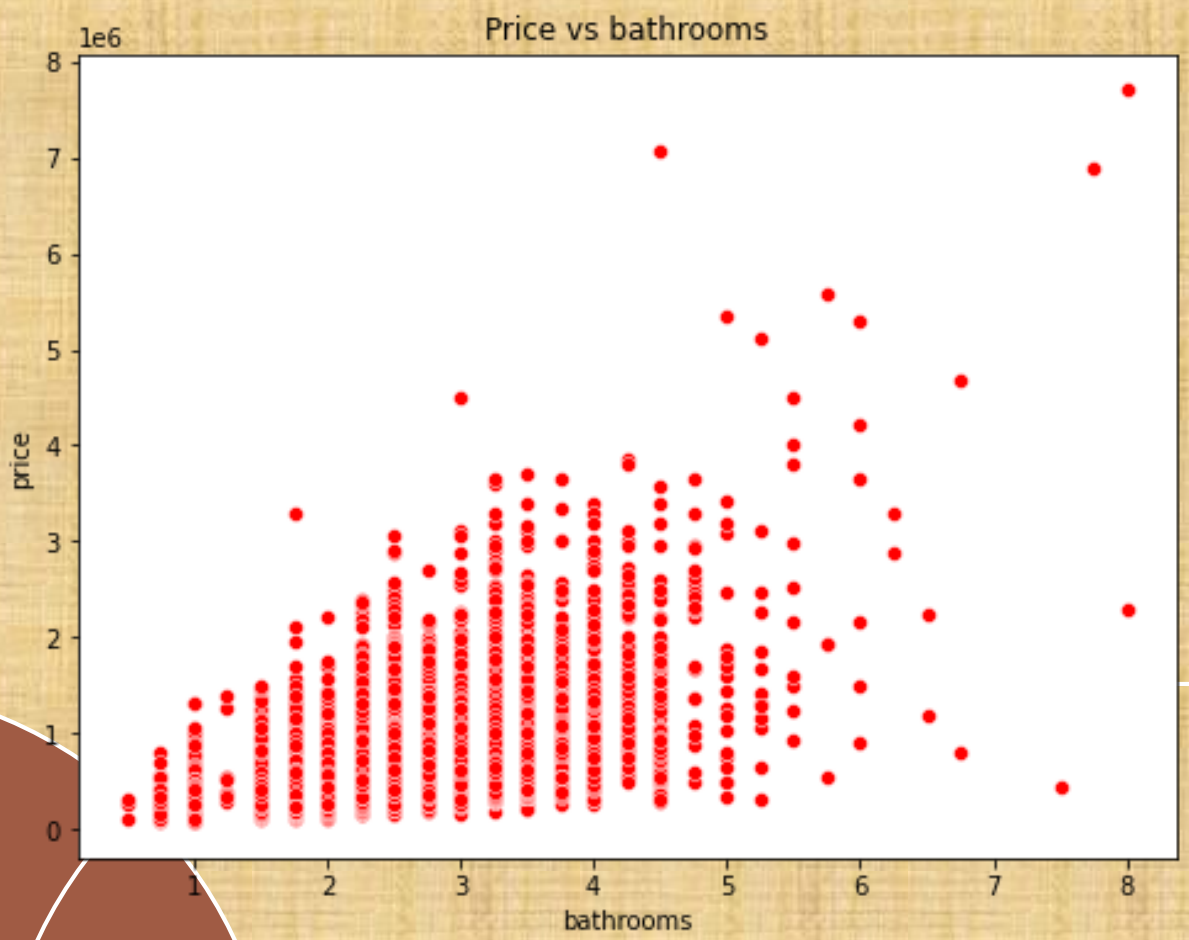
The scatter plot indicates that there is a positive correlation between the quality grade of houses and their prices.

Higher grade houses tend to have higher prices, but there is variation within each grade level. This trend is consistent across the entire range of grades, showing a strong relationship between house quality and market value.



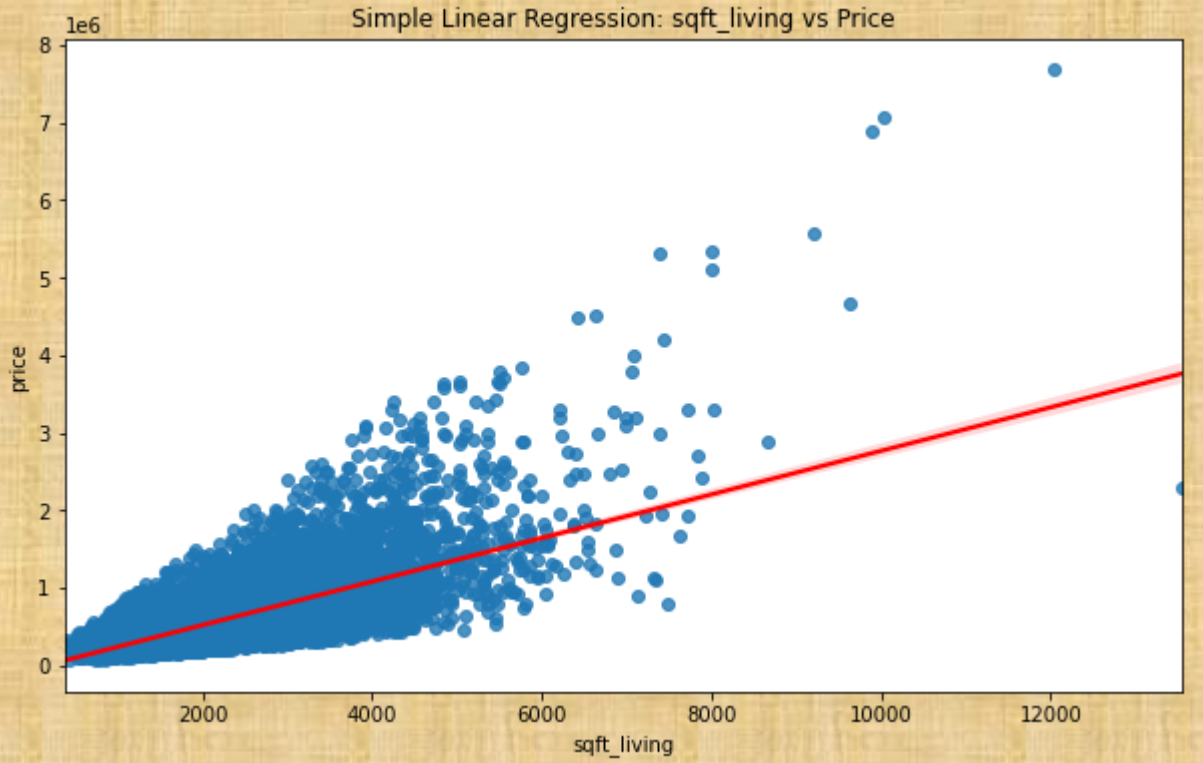
From the scatterplot , we observe that there is a clear positive correlation between the grade of the house and the living area. Higher-grade houses tend to have larger living areas.

This implies that Higher grade houses can justify their higher prices with one of the factors being having a larger living area



From the scatterplot , we observe that there is a clear positive correlation between the no of bathrooms and the price.
This implies that houses with more bathrooms fetch higher prices in the market.

MODEL DEVELOPMENT



- The visualization shows the actual data estimates of the living area square footage

- We note from the above that the price of the house increases as the living room square footage increases.

Key Determinants of Housing Prices

Method used: **Multiple Linear Regression**
Analyzed the combined effect of multiple features on house prices. The features included are:- ‘sqft_living’, ‘grade’, ‘sqft_above’, ‘bathrooms’, and ‘bedrooms’.
Findings:- These features collectively explain 53.77% of the variance in house prices.
Multiple Linear Regression model performance: R-squared value of 0.5377.

- Implications:Focus on Key Features:Renovations and improvements in sqft_living, grade, sqft_above, and bathrooms can significantly enhance property value.
- Real estate professionals should prioritize these areas for maximum return on investment.
- Strategic planning in these key areas will guide homeowners in making informed decisions to optimize property value.


Conclusion

Impact of Square Footage on Housing Prices

The analysis shows a strong positive correlation (r =0.7019r=0.7019) between living area size and housing prices. Larger living areas generally lead to higher prices in the King County market, highlighting the significance of square footage in determining property values.

Key Determinants of Housing Prices

Key factors affecting housing prices include living space size (r=0.7019r=0.7019), housing grade (r=0.6680r=0.6680), square footage above ground (r=0.6054r=0.6054), number of bathrooms (r=0.5259r=0.5259), and number of bedrooms (r=0.3088r=0.3088). Improving these attributes, especially living space and housing grade, can significantly boost property values. This insight helps homeowners and real estate professionals make strategic renovation and investment decisions.

Predictive Models for House Pricing

1. Simple Linear Regression

Using sqft_living as the predictor, the model achieves an R-squared score of $R2$=0.493R 2 =0.493. This indicates that 49.3% of the variance in housing prices can be explained by living area alone, suggesting that increasing living space can notably enhance property value.

Predictive Models for House Pricing

2. Multiple Linear Regression

The model incorporating bathrooms, sqft_living, grade, sqft_above, and bedrooms achieves an R-squared score of $R2=0.5377R\ 2\ =0.5377$. This indicates that 53.77% of the variance in housing prices is explained by these predictors, providing a more accurate estimate of property values and aiding in informed pricing and investment decisions.

Recommendations

1. Invest in Increasing Living Space:
Homeowners should consider expanding their living areas, as the analysis shows a strong positive correlation between square footage and housing prices. This investment can significantly enhance property value.

2. Focus on Key Features for Renovations:
Real estate professionals and homeowners should prioritize improvements in key determinants like housing grade, number of bathrooms, and square footage above ground to maximize property value and attract higher prices.

3. Utilize Predictive Modeling for Pricing Strategies:
Implementing predictive models that incorporate multiple key features will help real estate professionals and homeowners make more informed pricing and investment decisions, optimizing returns and efficiency in the housing market.

Q&A