KING COUNTY HOUSING SALES ANALYSIS

Group members

- Angela Kalelwa
- Brian Waweru
- Amos Kipngetich
- George Machoka
- Cecily Wahome

Introduction

- Real estate is a critical sector in any economy.
- This analysis aims to identify factors influencing housing prices in a northwestern county.

Business Understanding

The real estate agency assists homeowners in buying and selling homes.

They provide advice on home renovations to increase home value.

Goal: Develop a model to predict home value post-renovations based on type and cost of renovations.

Business Problem

- The agency needs accurate advice on how renovations affect home value.
- Currently, lacks a reliable method for predicting renovation impacts.
- Aim to identify pricing factors, analyze trends, and detect undervalued properties.

Problem Statement

The real estate agencies need a model to estimate house prices and provide advice to house owners about how renovations might increase the estimated values of their homes.

Inability to accurately identify pricing factors, analyze trends, and detect undervalued properties.

This results in unreliable information for buyers and sellers.

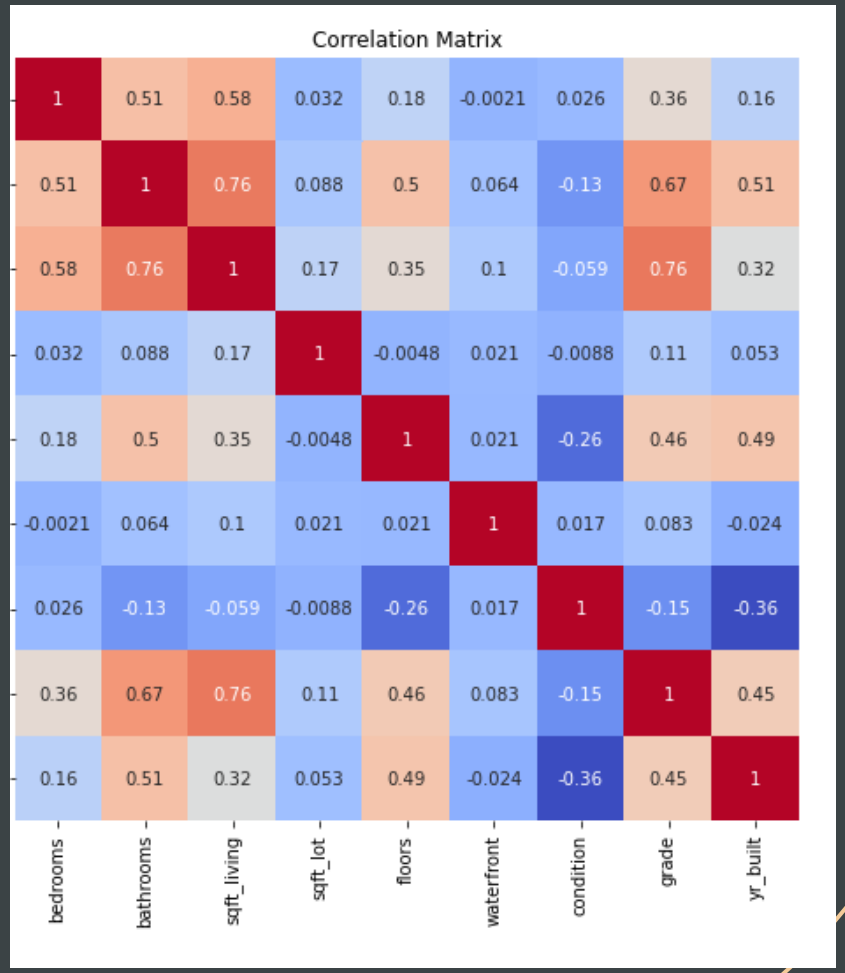
Data understanding

- **Data Sources**: Kings County housing dataset.
- **Features**: Includes target variables as `price and bedrooms`, `bathrooms`, `sqft_living`, condition, grade etc as the predictors.
- **Data Quality**: Assess completeness and accuracy of data. Identify any anomalies or inconsistencies.
- **Target Variable**: `price` – the housing price to be predicted.
- **Predictors**: Features such as `bedrooms`, `bathrooms`, `sqft_living`, etc.

Data Preprocessing

Data preparation involves cleaning and transforming the dataset for modeling:

- **Missing Values**: Handle missing values in features.
- **Data Types**: Convert data types as needed (e.g., encoding categorical variables).
- **Multicollinearity**: Check for and address correlated predictors.
- **Normalization**: Normalize numerical features for better model performance.
- **Encoding**: Convert categorical variables into numeric format using one-hot encoding if necessary.



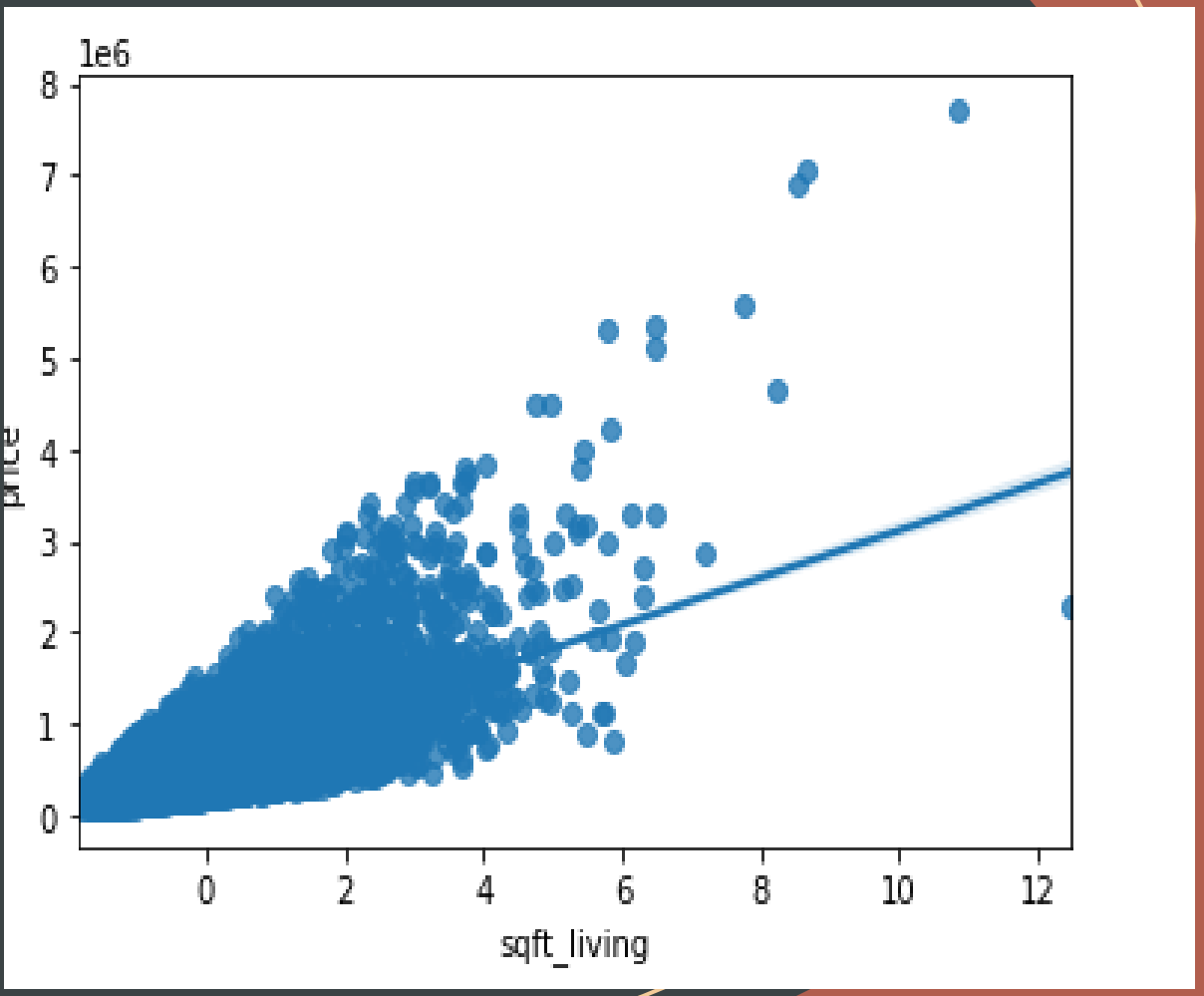
Correlation Analysis

Square Footage (sqft_living):

- **High Correlation with Price**: The feature 'sqft_living' has a strong positive correlation with price, indicating that larger living spaces tend to have higher prices. The correlation coefficient is around 0.76, suggesting a substantial impact on house prices.

- **Bathrooms**:
  - **Moderate to High Correlation with Price**: The number of bathrooms also shows a significant positive correlation with price, with a correlation coefficient of about 0.67. More bathrooms typically increase the value of a property.

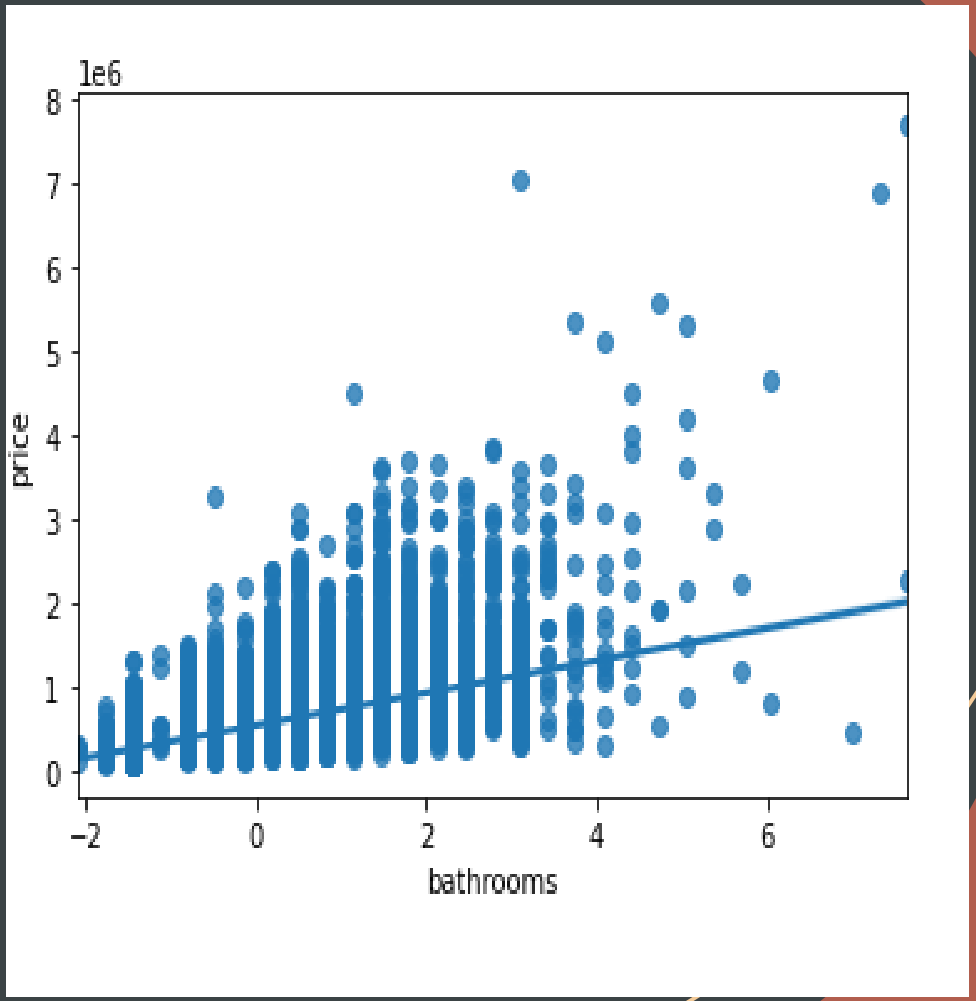
- **Grade**:
  - **Strong Correlation with Price**: The 'grade' of the house, which rates the construction and design quality, is strongly correlated with price (correlation coefficient around 0.75). Higher grades indicate higher house prices.



Key features analysis
Analysis of the Scatter Plot: `sqft_living` vs. `price`

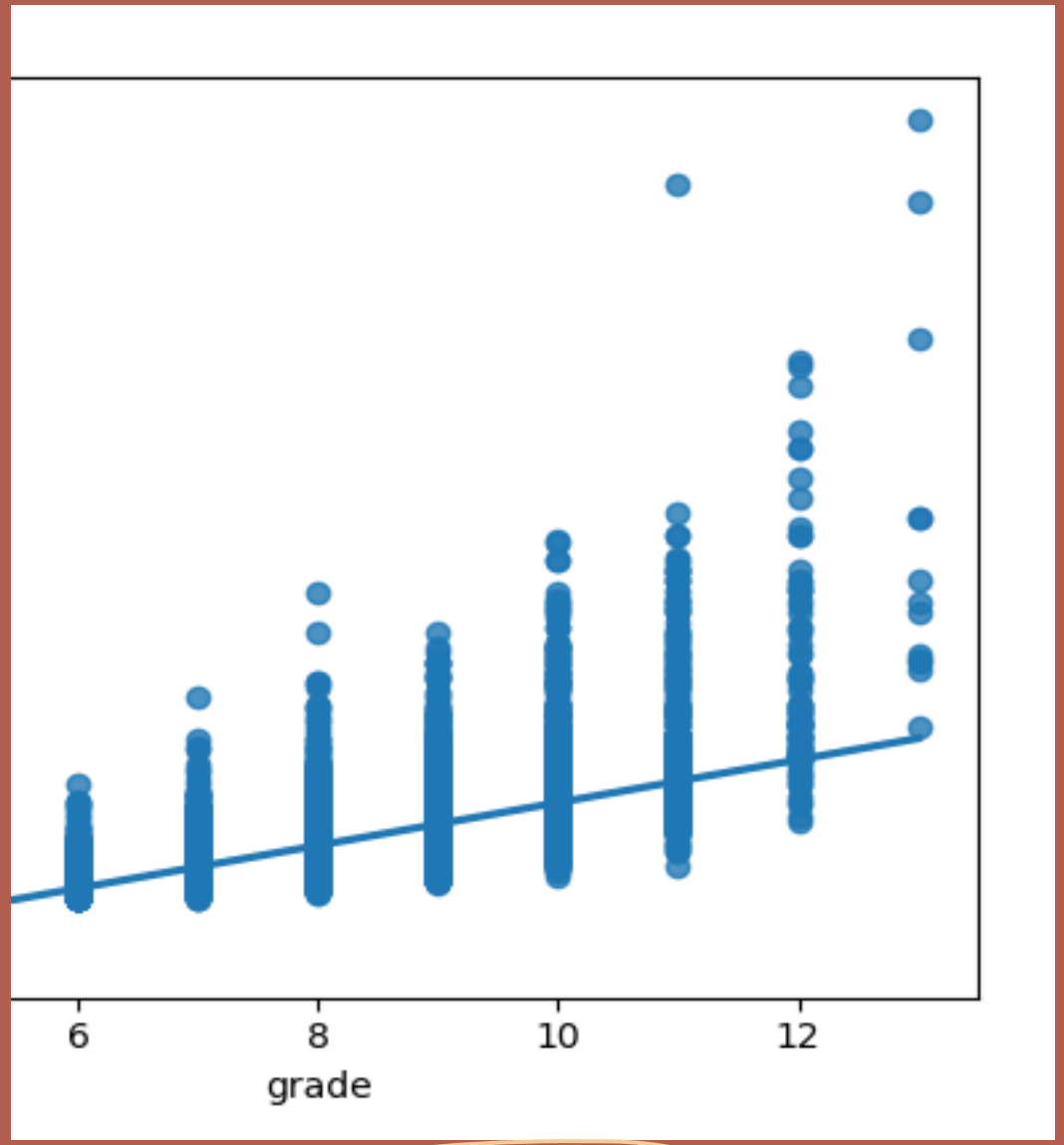
From the plot

- There is a clear positive correlation between the living area square footage and the house price.
- As the living area increases, the price of the house generally increases as well.
- This indicates that larger homes tend to be more expensive.



Analysis of the Scatter Plot: `bathrooms` vs. `price`

- There is a noticeable positive correlation between the number of bathrooms and the house price.
- As the number of bathrooms increases, the price of the house generally increases.
- This suggests that houses with more bathrooms tend to be priced higher.



Analysis of the Scatter Plot: grade vs. price

- There is a clear positive correlation between house grade and price. As the grade increases, the price of the house generally increases.
- Higher grades are associated with higher prices, which makes sense as better quality and more desirable features contribute to increased property value.
- The trend line in the scatter plot reinforces the positive relationship between grade and price.

Analysis of OLS Regression Results for `sqft_living`

Coefficient for `sqft_living`:

- The coefficient of 280.863 indicates that for every additional square foot of living space, the house price is expected to increase by approximately $280.86, holding all other factors constant.
- The high t-value and low p-value indicate that `sqft_living` is a statistically significant predictor of house prices.

Model Fit:

- The R-squared value of 0.493 suggests that nearly half of the variation in house prices can be explained by the square footage of living space alone.
- This implies that while `sqft_living` is a strong predictor, other factors also contribute to house price variability.

Analysis of OLS Regression Results for `bathrooms`

Coefficient for `bathrooms`:

- The coefficient of 193,200 indicates that for every additional bathroom, the house price is expected to increase by approximately $193,200, holding all other factors constant.
- The statistical significance (P-value < 0.05) confirms that the number of bathrooms is a significant predictor of house prices.

Model Performance:

- The R-squared value of 0.277 suggests that 27.7% of the variation in house prices can be explained by the number of bathrooms. This implies that while the number of bathrooms is a significant predictor, other factors also contribute to house price variability.

Diagnostic Metrics:

- The diagnostic metrics indicate potential issues with normality (high Skew and Kurtosis) and the presence of outliers, which should be addressed for more robust modeling.

Analysis of OLS Regression Results for `grade_11`

Coefficient for `grade_11`:

- The coefficient of 975,500 indicates that houses with a grade 11 are expected to have prices approximately $975,500 higher than houses that are not grade 11, holding all other factors constant.
- The statistical significance (P-value < 0.05) confirms that grade 11 is a significant predictor of house prices.

Model Performance:

- The R-squared value of 0.128 suggests that 12.8% of the variation in house prices can be explained by whether a house is grade 11. This implies that while grade 11 is a significant predictor, other factors also contribute to house price variability.

Diagnostic Metrics:

- The diagnostic metrics indicate potential issues with normality (high Skew and Kurtosis) and the presence of outliers, which should be addressed for more robust modeling.

Insights

Impact of Living Space (`sqft_living`):

- The OLS regression model indicates that an increase in living space significantly raises the house price, with an estimated increase of approximately $257.90 for each additional square foot.
- The model explains about 49.3% of the variability in house prices, suggesting that while living space is a strong predictor, other factors also play a role.

Impact of Number of Bathrooms (`bathrooms`)

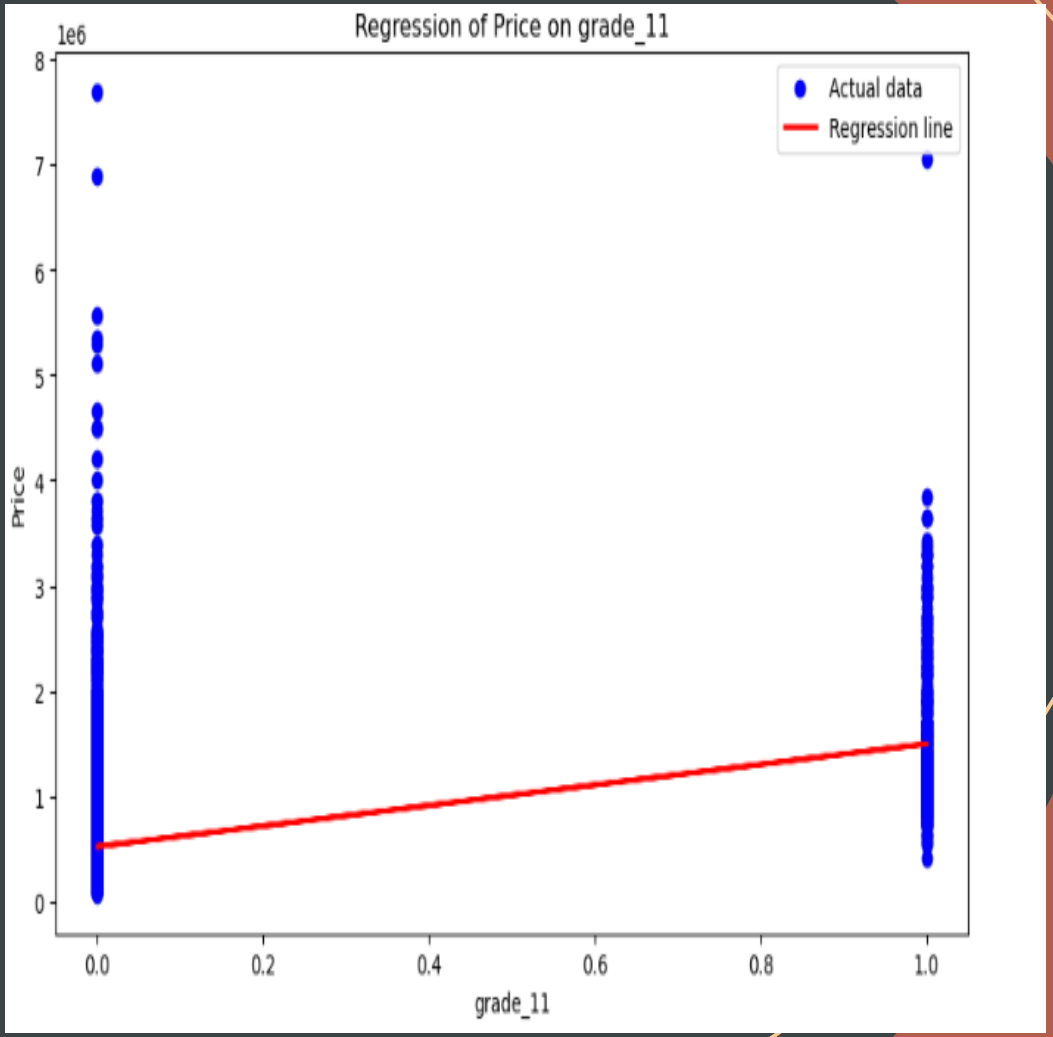
- The OLS regression results show that each additional bathroom increases the house price by approximately $193,200.
- The model explains about 27.7% of the variability in house prices, indicating that the number of bathrooms is a significant predictor but not as strong as living space.
- **Impact of House Grade (`grade_11`)**:
- The regression model shows that being classified as grade 11 increases the house price by approximately $975,500.
- The model explains about 12.8% of the variability in house prices, indicating that while grade 11 is a significant predictor, many other factors contribute to house price variability.

Model development

we build and refine models to predict housing prices:

- **Model Selection**: Evaluate different regression models such as Linear Regression, Decision Trees, and Random Forests.
- **Overfitting and Regularization**: Implement techniques to avoid overfitting and improve model generalization.
- **Validation**: Use cross-validation to assess model performance on unseen data.
- **Loss Functions**: Employ appropriate loss functions like Mean Squared Error (MSE) for regression tasks.
- **Performance Metrics**: Measure model performance using metrics such as R-squared and Root Mean Squared Error (RMSE).

Regression for grade_11



Positive Correlation:

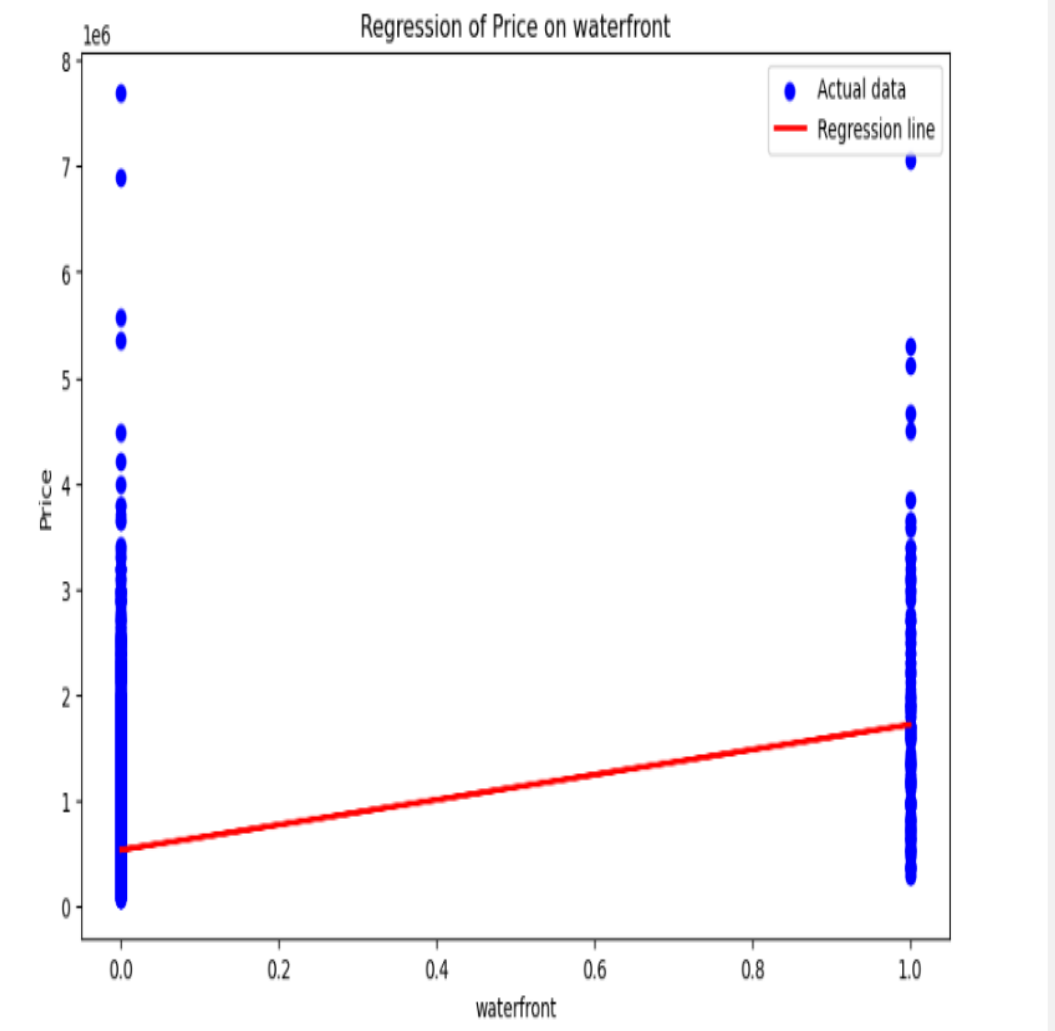
- The plot indicates a positive correlation between `grade_11` and house prices.
- **Wide Spread of Prices**:
- There is a wide spread of prices within each category (both for grade_11 = 0 and grade_11 = 1).
- This suggests that while being graded as "Excellent" has a positive impact on house prices, it is not the sole determinant. Other features also play a significant role in determining house prices.

Regression Line:

- The red regression line shows the trend, indicating that houses with `grade_11 = 1` have higher predicted prices than those with `grade_11 = 0`.

Cont...

- **Model Metrics**:
- **Mean Squared Error (MSE)**: 117696639050.11594
  - This value measures the average squared difference between the observed actual outcomes and the outcomes predicted by the regression model.
- **R-squared (R^2)**: 0.1287280865311878
  - This value indicates that approximately 12.87% of the variability in house prices can be explained by the `grade_11` feature.
  - A low R-squared value suggests that while `grade_11` is a statistically significant predictor, other factors contribute to house price variability.



Regression for waterfront

- **Positive Correlation**:
- The plot indicates a positive correlation between `waterfront` and house prices.
- Houses with waterfront access (waterfront = 1) generally command higher prices compared to those without waterfront access (waterfront = 0).

Wide Spread of Prices:

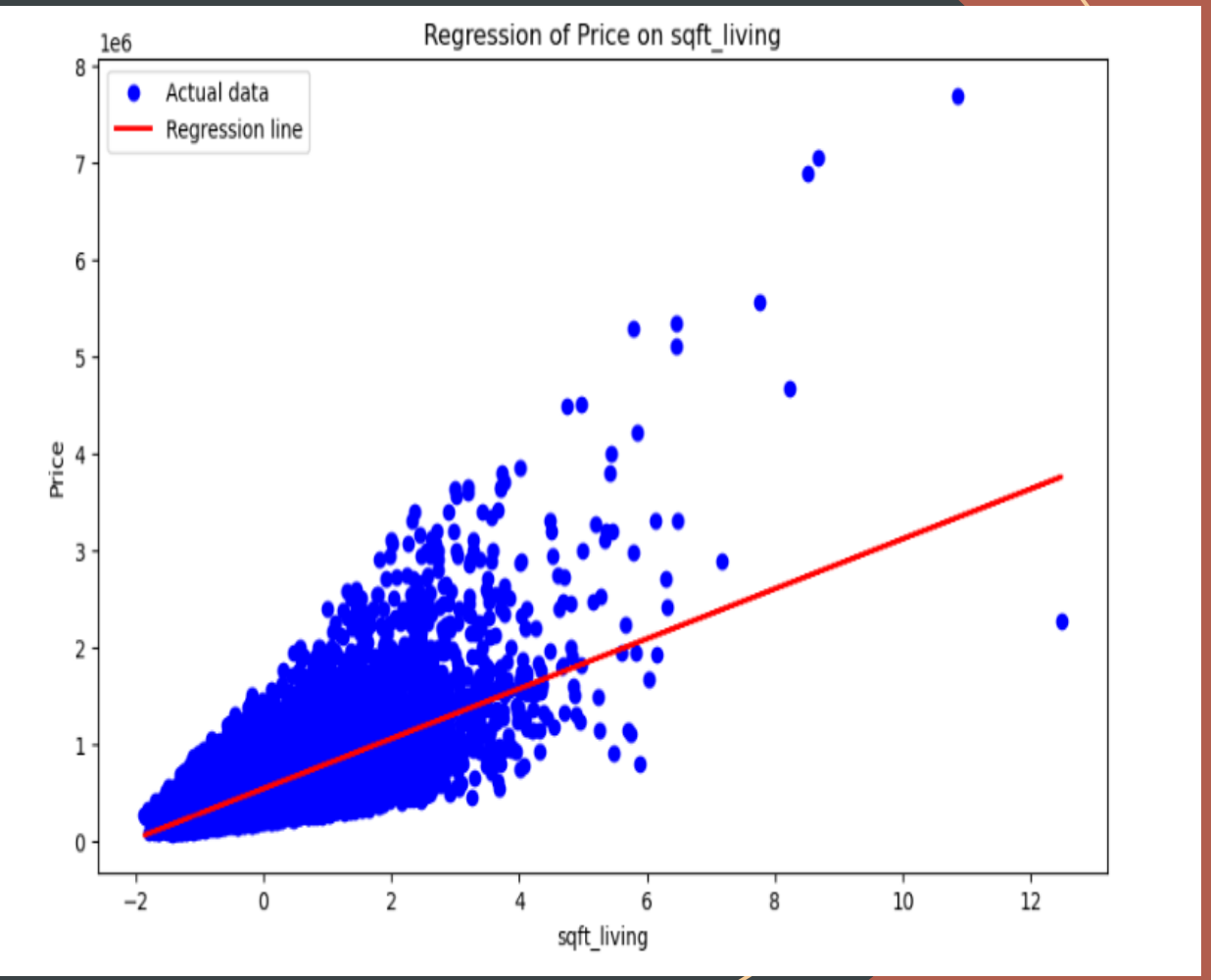
- There is a wide spread of prices within each category (both for waterfront = 0 and waterfront = 1).
- This suggests that while having waterfront access has a positive impact on house prices, it is not the sole determinant. Other features also play a significant role in determining house prices.

Cont:

Regression Line:

- The red regression line shows the trend, indicating that houses with waterfront access have higher predicted prices than those without waterfront access.
- The slope of the line suggests a moderate increase in price for houses with waterfront access.
- **Model Metrics**:
- **Mean Squared Error (MSE)**: 125525573273.88926
    - This value measures the average squared difference between the observed actual outcomes and the outcomes predicted by the regression model.
- **R-squared (R^2)**: 0.0698578899281942
    - This value indicates that approximately 6.99% of the variability in house prices can be explained by the `waterfront` feature.
    - A low R-squared value suggests that while `waterfront` is a statistically significant predictor, other factors contribute to house price variability.

Regression for sqft_living



Positive Correlation:

- The plot indicates a strong positive correlation between `sqft_living` and house prices.
- As the square footage of the living area increases, the house price generally increases.

Trend Line:

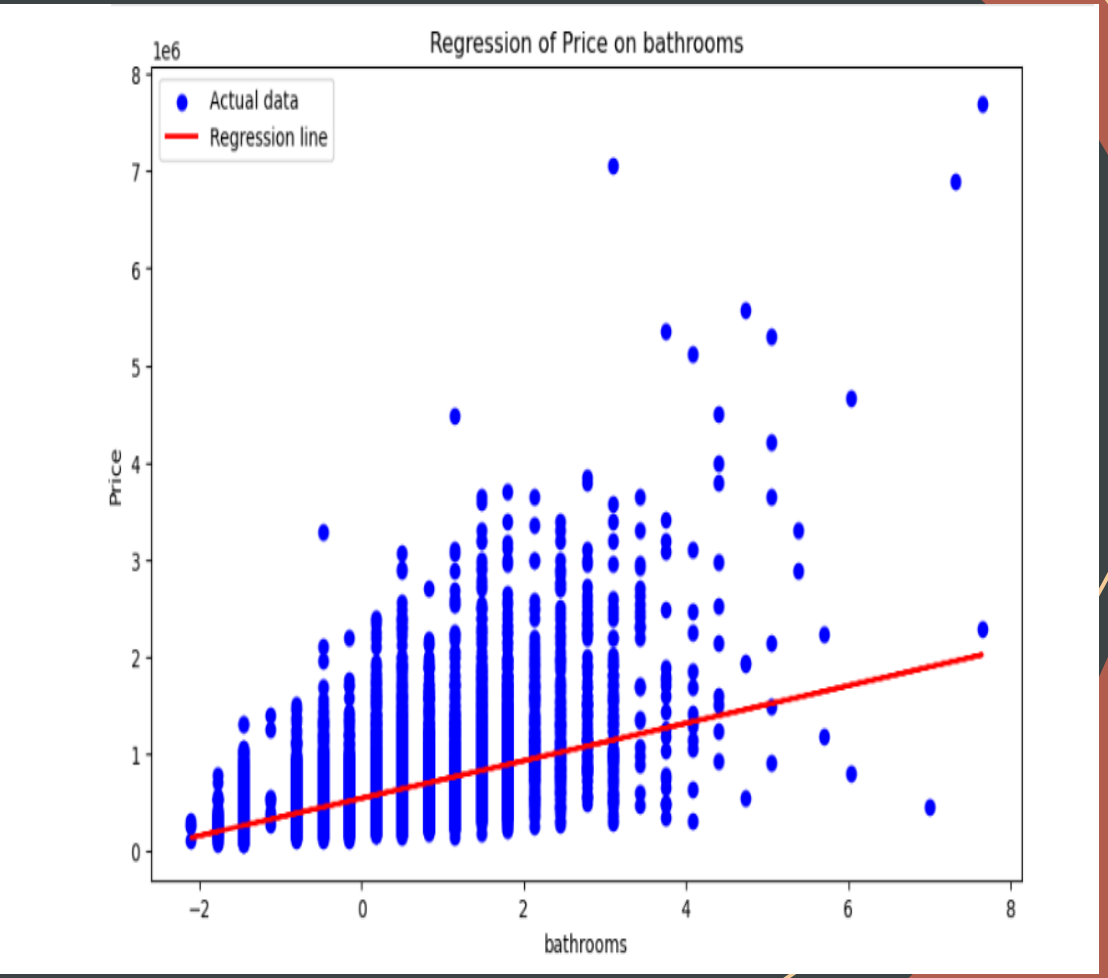
- The red regression line shows the trend, indicating that larger living areas are associated with higher house prices.
- The slope of the line suggests a significant increase in price for each additional square foot of living space.

Data Distribution:

- There is a concentration of data points around the lower to mid-range values of `sqft_living`, with prices increasing more steeply for larger homes.

Model Metrics

- **Mean Sqared Error (MSE)**: 68463341389.86132
  - This value measures the average squared difference between the observed actual outcomes and the outcomes predicted by the regression model.
- **R-squared (R^2)**: 0.4926878994035093
  - This value indicates that approximately 49.27% of the variability in house prices can be explained by the `sqft_living` feature.
  - A moderate R-squared value suggests that `sqft_living` is a significant predictor of house prices, but other factors also contribute to price variability.



Regression for bathrooms

- The regression plot analysis demonstrates that the number of bathrooms is a positive predictor of house prices. Houses with more bathrooms generally command higher prices, making this an important factor in property valuation. This insight can guide strategic decisions in home renovations, real estate marketing, and investment strategies. The model metrics further support the significance of the number of bathrooms as a key determinant of house prices, while also indicating the importance of considering additional features for accurate property valuation.

Model analysis

Model Performance:

- *Training Score*: 68.05%
- *Test Score*: 64.87%
- *R-squared*: 0.681 (indicating that approximately 68.1% of the variance in the housing prices can be explained by the model)

Coefficients Analysis:

- Several variables are statistically significant with p-values less than 0.05, indicating that these variables have a meaningful impact on the housing prices.
- The waterfront feature has a large positive coefficient, suggesting that houses with waterfront views significantly increase the price.
- Negative coefficients for bedrooms and sqft_lot could suggest that, all else being equal, having more bedrooms or a larger lot size might not always correspond to higher prices, possibly due to the influence of other features.

Multicollinearity:

- There could be potential multicollinearity problems, as indicated by the smallest eigenvalue. This can affect the stability and interpretation of the coefficient estimates.

Residual Analysis:

- The skewness and kurtosis indicate that the residuals might not be normally distributed, which could suggest issues with model assumptions.

Recommendations

- Focus on improving key features like bathrooms, square footage, and waterfront views. These have significant positive impacts on house prices. Consider renovations or upgrades that improve the condition and grade of houses, as these are also significant predictors of price.
- The model shows that having a waterfront view significantly increases house prices. Real estate agents should highlight this feature when marketing properties with waterfront access. Invest in or develop waterfront properties to maximize returns. Analyze the impact of location variables further (e.g., neighborhood analysis) to target high-value areas.
- Use the model to set competitive prices by considering the significant predictors. Ensure that pricing strategies reflect the contributions of these key features.
- Highlight features with the highest impact on price in marketing materials (e.g., newly renovated bathrooms, high-grade finishes, and waterfront views).Tailor marketing strategies based on the model’s insights to emphasize the most valued aspects of properties.

Recommendations

- Higher grades (e.g., grade_11) are associated with significantly higher prices. Consider investing in property upgrades and renovations to improve the overall grade of the property.
- Older houses (yr_built) have a negative impact on price. This suggests that newer properties tend to have higher values, so renovation and updating older properties could be beneficial.
- Properties in better condition (condition_5) command higher prices. Regular maintenance and improvements are essential to maximize property value.
- The coefficients suggest that having more bedrooms or bathrooms might not always increase price as expected. This could indicate that other factors, such as location or property condition, may play a more significant role. Evaluate whether adding more rooms or focusing on improving existing features offers better returns.

Next Steps

- Investigate and address multicollinearity. This could involve removing or combining highly correlated features to improve model stability.
- The residuals show signs of non-normality. Revisit the data transformation methods to address this issue and ensure the model meets regression assumptions.
- Explore creating interaction terms or polynomial features to capture more complex relationships between features and house prices.
- Try alternative regression models, such as Ridge or Lasso regression, which can handle multicollinearity better.
- Consider applying transformations to the dependent variable (price) or independent variables to meet the assumptions of linear regression.

General Conclusions

Living Space: Increasing the square footage of a house is a highly effective way to raise its market value. Larger homes are consistently priced higher, making living space a crucial factor for both buyers and sellers.

Number of Bathrooms: Adding more bathrooms to a house also contributes significantly to its value. Homes with more bathrooms are valued higher, suggesting that functionality and convenience are important to buyers.

Waterfront: Having a waterfront view significantly increases houseprices .Agents should invest in or develop waterfront properties to maximize returns.

Grade: Consider renovations or upgrades the improve the condition and grade pf houses as they are significant predictors

Overall Model Insights: While each of these factors independently contributes to house prices, the combined model indicates that there are additional variables influencing property values that were not captured in these simple models. Therefore, a more comprehensive model that includes

THANK YOU