King County House Sales Analysis: Using Regression Modeling to Predict House Prices

Contributors:
- Maya Bororio
- Ruth Kitasi
- Ilyas Bourzat
- Nicholas Kirui
- Robert Kalafa
- Joseph Matheri

Project Overview

The project aims to use regression modeling to predict house prices in King County based on various features, including potential renovations. The goal is to identify key factors influencing house prices and advise homeowners on renovations that are likely to increase their home's value.

Business Problem

A real estate agency seeks a data-driven approach to estimate the impact of home features and renovations on house prices in King County. The goal is to provide homeowners with actionable insights on which renovations would yield the highest return on investment.

Objectives

1. Determine which features have the most significant impact on house prices.

2. Examine the correlations between different features and house prices to identify strong relationships.

3. Build a predictive model to estimate house prices based on features.

Data Understanding

This dataset includes house sale prices for King County, USA, covering transactions between May 2014 and May 2015.

Dataset Dimensions: It comprises 21 columns and 21,597 rows.There are 21 distinct variables, each recorded for 21,597 entries.

Data Cleaning

- Removed outliers.
- Imputed missing values in columns such as waterfront, view and sqft_basement.
- Created a subset of the main dataset consisting of features used in the analysis.

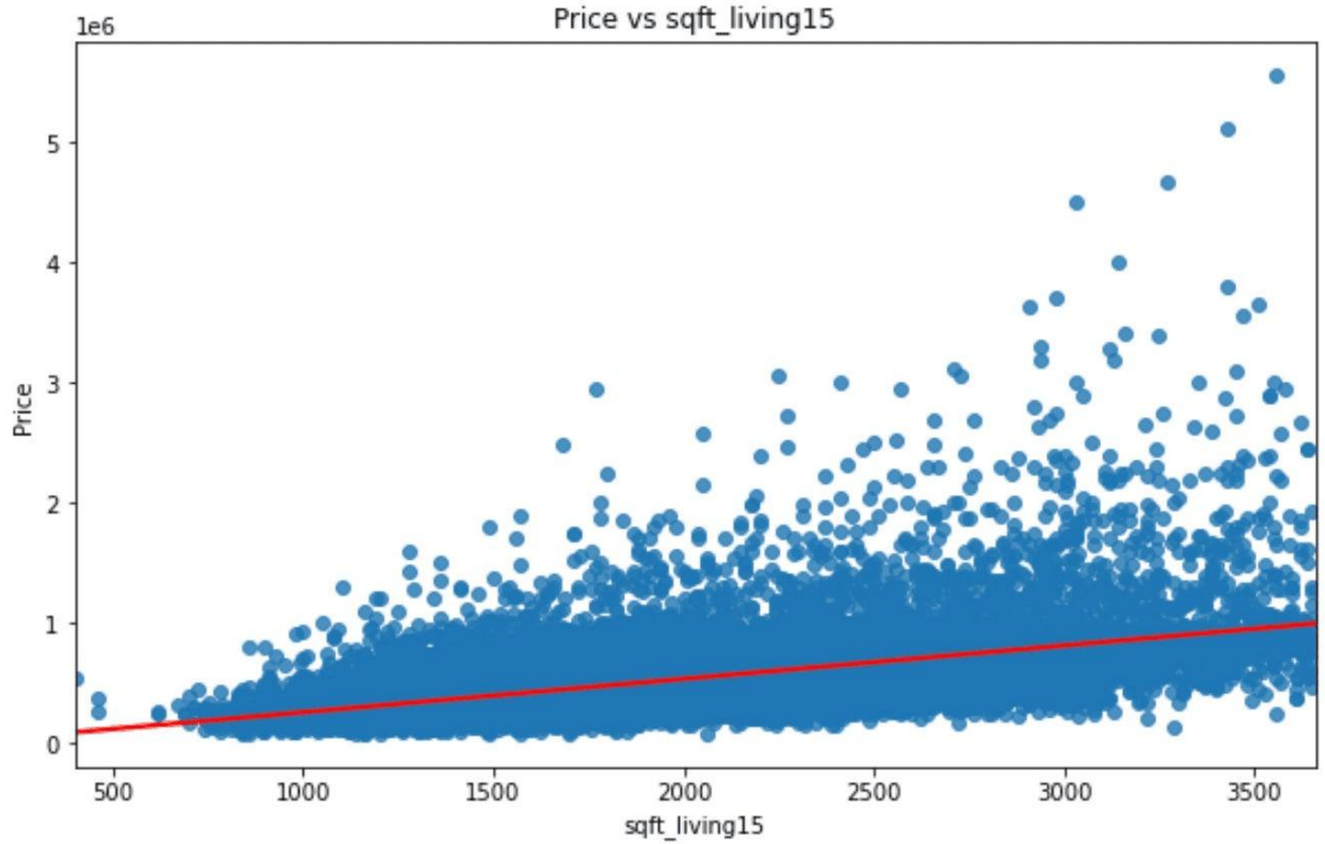
sqft_living	0.701917
grade	0.667951
sqft_above	0.605368
sqft_living15	0.585241
bathrooms	0.519628

Features with the strongest impact on house prices

The features with the highest impact on house prices show a strong correlation to house prices. This means that as these features increase, the house prices also increase. The correlation values closer to 1.0 indicate a stronger relationship.

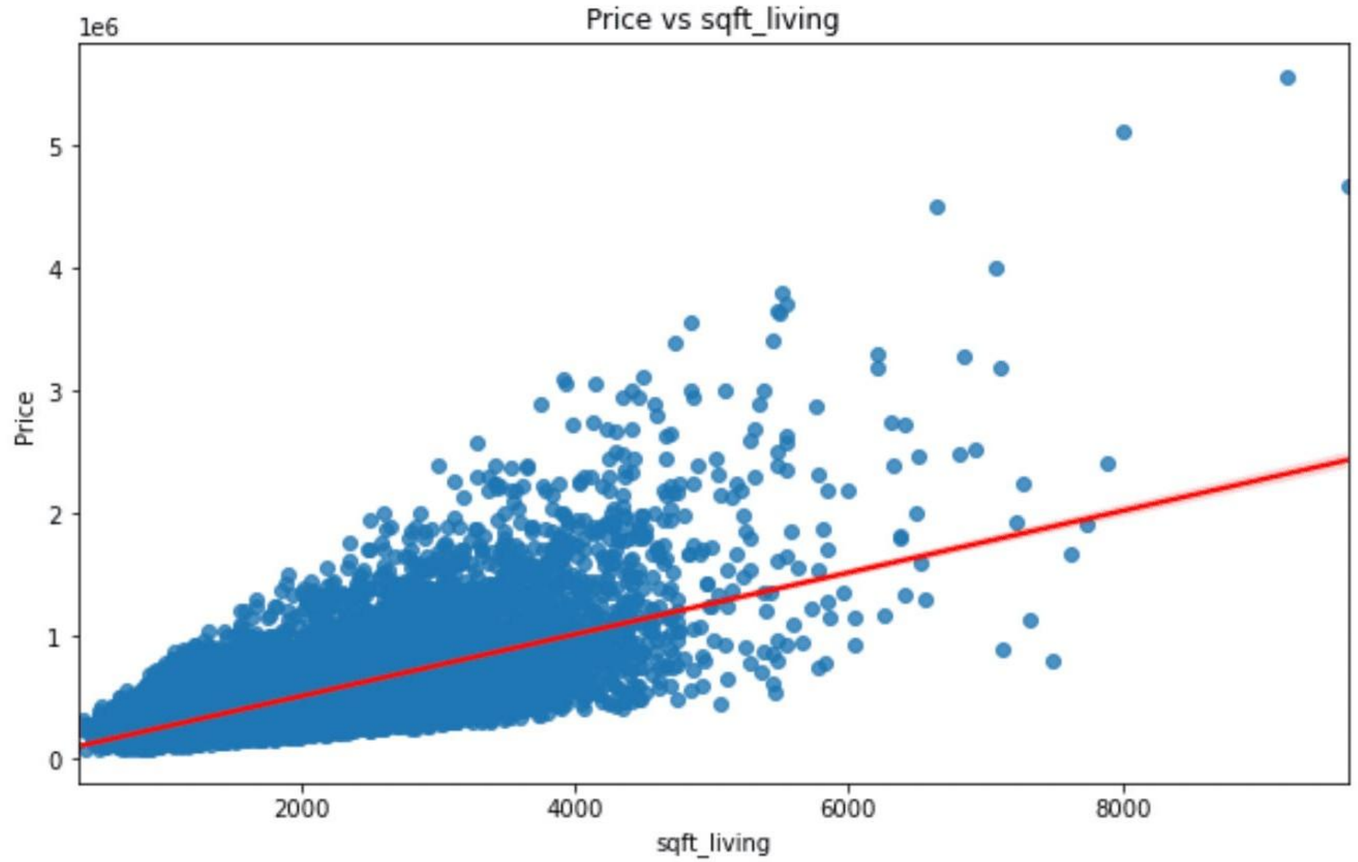
Correlation between features

This visualization shows that larger interior living spaces generally lead to higher house prices. The red trend line highlights this positive correlation.



Correlation between features

This visualization shows that house prices generally increase with larger interior living spaces. The red trend line highlights this positive correlation, indicating that more square footage leads to higher prices.



Modelling

We started off with a baseline model consisting of the features selected above and improved the model through:

- Log transformation on the price column
- Log transformation on the other numerical columns

Baseline Model

For our baseline model, we used features that are highly correlated with price those are: bathrooms, sqft_living, grade and sqft_above.

Model Summary:
R-squared: 0.56

Final Model

From the baseline model, we transformed the price column as well as the sqft_above and sqft_living columns.

Model Summary:
R-squared: 0.61

Predictive model to determine house price based on features

This visualization compares true house prices (x-axis) with predicted prices (y-axis). The red dashed line represents perfect predictions. Points closer to this line indicate more accurate model predictions, helping assess the model's accuracy and identify deviations.

Conclusion

Our analysis identifies living space size, construction quality, and number of bathrooms as key factors impacting house prices. The predictive model using these factors provides reliable price estimates for better decision-making.

Reccomendations

1. **Investment in Key Features:** Enhance living space, construction quality, and add bathrooms to increase property values.

2. **Market Positioning:** Emphasize these key features in marketing to attract buyers and justify higher prices.

3. **Data-Driven Decisions:** Refine the predictive model regularly with new data for accurate pricing and strategic investment decisions.

Thanks!

DO YOU HAVE ANY QUESTIONS?