

engige / king_county_modeling_project_g19

<> Code

Issues

Pull requests

Actions

Projects

Security

Insights

Settings

0 stars

0 forks

1 watching

1 Branch

0 Tags

Activity

Public repository

1 Branch

0 Tags

Go to file

Go to file

+

Add file

Code


engige

 readme update

637af3a · 5 minutes ago

data	first commit	last week
deliverables	further updates	4 days ago
images	further updates	4 days ago
notebooks	Updates	4 days ago
.gitignore	first commit	last week
README.md	readme update	5 minutes ago
index.ipynb	further updates	4 days ago
plan.txt	updated txt	last week

README



Overview

In today's competitive real estate market, homeowners seek ways to maximize the value of their properties through renovations. This project aims to leverage linear regression modeling to provide insights into how different types of home renovations can affect the estimated value of homes. By analyzing historical sales data and applying regression techniques, we will quantify the impact of specific renovation projects on home prices, helping homeowners make data-driven decisions and real estate agents provide expert advice to their clients.

Business Understanding

The primary objective of this project is to provide actionable insights to a real estate agency that assists homeowners in buying and selling properties. The agency's clients often inquire about the potential increase in home value resulting from various renovation projects. Therefore, this project aims to address the following key objectives:

1. Develop a Predictive Model: Create and validate a linear regression model that predicts the increase in home value based on the type and extent of renovations undertaken, ensuring its reliability and applicability to various property types and market conditions.
2. Quantify the Impact of Renovations: Determine how different types of home renovations contribute to the overall increase in property value by analyzing historical sales data and identifying the renovations that provide the highest return on investment.

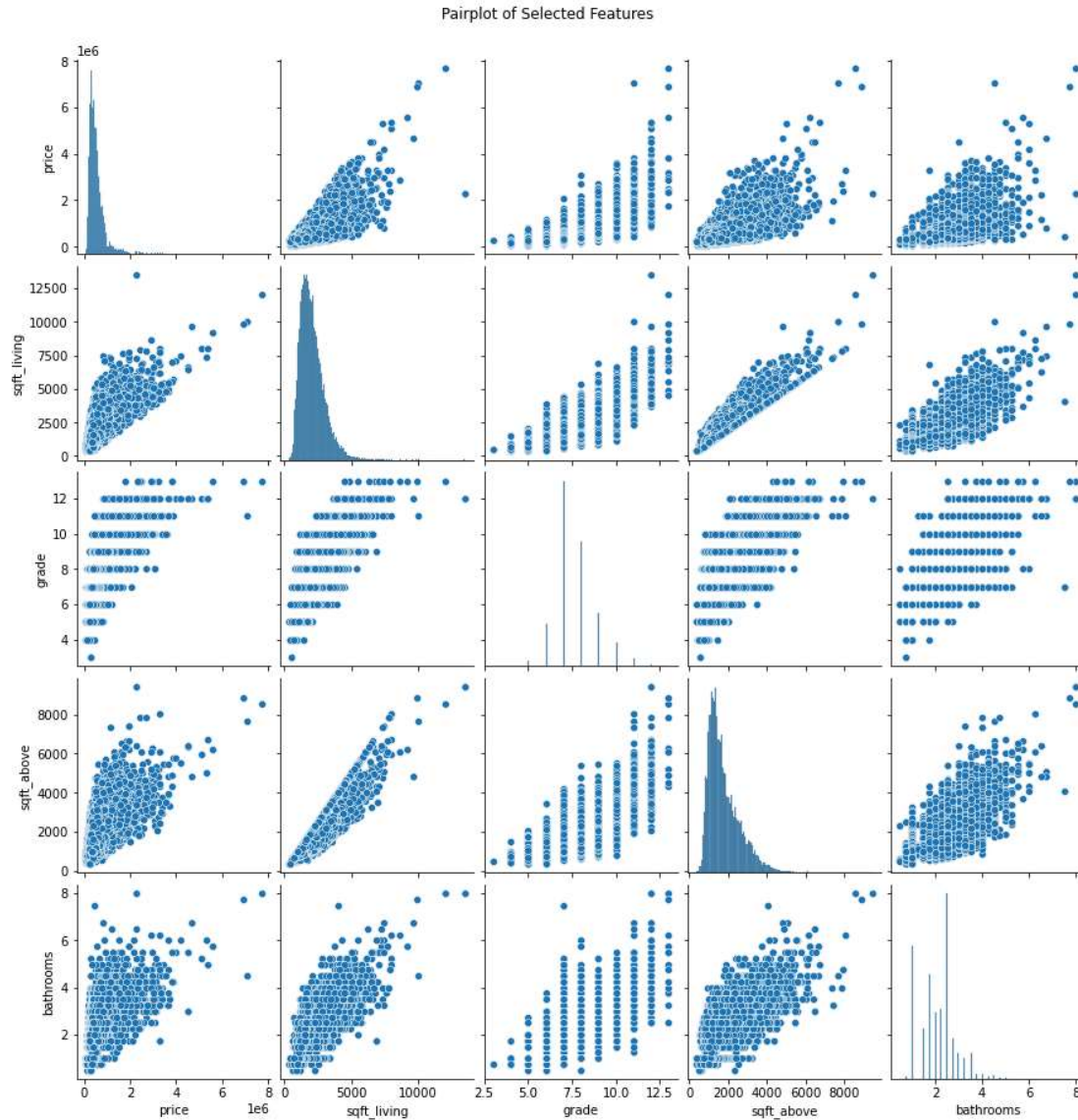
Data Understanding

This project utilizes the King County House Sales dataset, contained in the file named `kc_house_data.csv`. This dataset includes various features related to house sales, such as square footage, number of bedrooms and bathrooms, presence of a waterfront, view quality, year built, and renovation status. The dataset also provides the sale price of each property, which serves as the dependent variable in our regression modeling. Detailed descriptions of the column names can be found in the accompanying `column_names.md` file. The aim is to leverage this rich dataset to understand and quantify the impact of various home renovations on property values.

Data Preparation

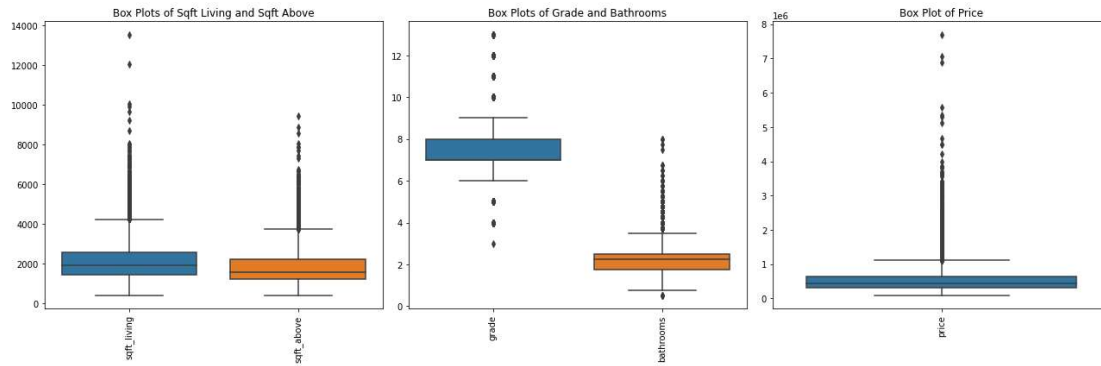
The data preparation process involved several critical steps to ensure data quality and relevance for modeling. Initially, the data was inspected for duplicates and missing values. No duplicates were found, and missing values were identified in three columns: `waterfront`, `view`, and `yr_renovated`. Missing values for categorical data (`waterfront` and `view`) were filled using the mode, while numerical data (`yr_renovated`) was filled using the mean. Features were selected based on correlation analysis and domain knowledge. The selected features for modeling included `sqft_living`, `grade`, `sqft_above`, and `bathrooms`, all of which showed significant positive correlations with the target variable (price) and were practically relevant in real estate valuation. To avoid clutter, pair plots and heatmaps visualization were generated for the selected features, further highlighting their relationships with price and identifying potential multicollinearity issues, particularly between `sqft_living` and `sqft_above`.

Pairplot (Correlation)



Outliers were examined using box plots, revealing their presence across all selected features. And since outliers, especially on the higher end, could skew data distribution and impact model accuracy, we capped them while iterating the models to observe their impact. The comprehensive data preparation ensures that the selected features are relevant and that any data anomalies are addressed, laying a solid foundation for the subsequent modeling process.

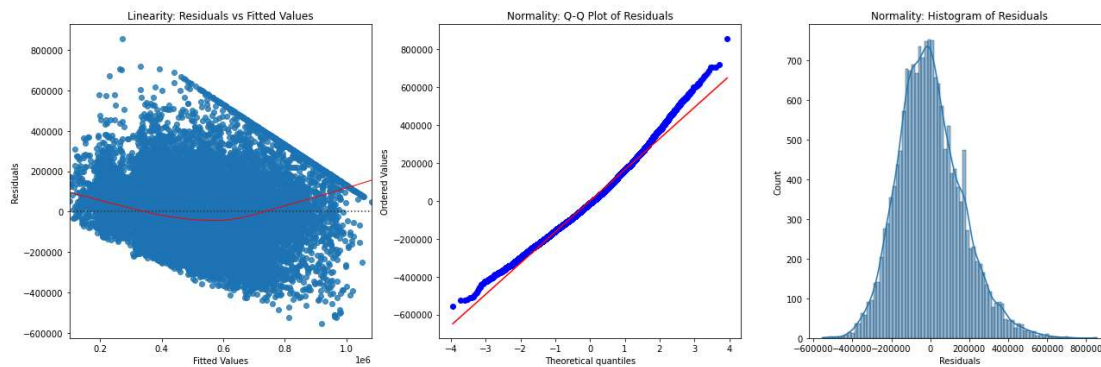
🔗 Box Plots (Outliers)



Modeling & Validation

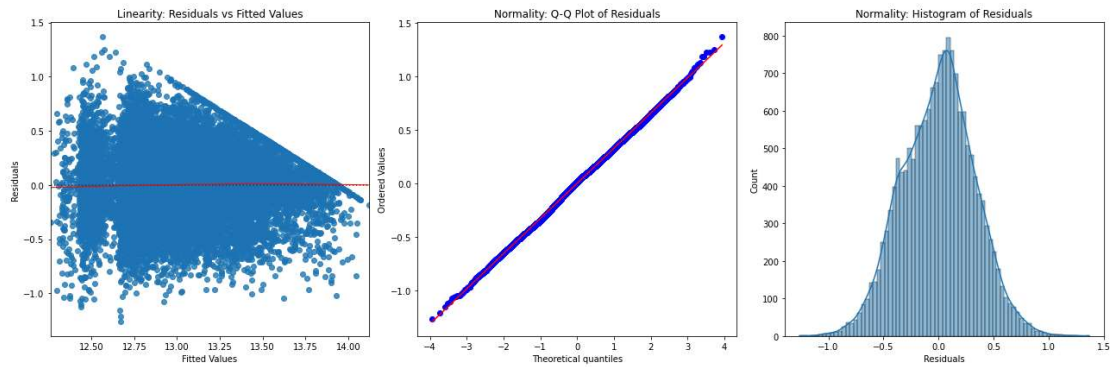
The modeling process involved an iterative approach to evaluate and improve model performance based on key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-Squared (R^2). Initially, four models were built without addressing outliers, multicollinearity, or scaling. Among these, Model 4 (with features `sqft_living`, `grade`, `bathrooms`, and `sqft_above`) performed best with the lowest errors and highest explanatory power (**Model 4 - MAE: 129,161, MSE: 27,242,617,862, R^2 : 0.560**). Initial modelling also indicated that including additional relevant features improves model performance. Subsequent iterations involved addressing outliers by capping them, which improved the model's metrics significantly. However, addressing multicollinearity by dropping `sqft_above` did not enhance the performance, indicating the importance of this feature. Scaling predictors also did not impact the model's performance, as linear regression inherently adjusts for the scales of input features.

Model 4 Diagnosis



Further refinement included diagnosing the best-performing model (Model 4, under Iteration 2) for conformity to linear regression assumptions. The initial diagnosis indicated violations of linearity and normality (see Model 4 Diagnosis above). To address this and based on previous observations, further model iteration was performed with outliers removed, predictors scaled, price log transformed and an additional feature included (`bedrooms`). The refined Model 5 showed improved adherence to linear regression assumptions, with better linearity and normality (See Model 5 Diagnosis below). When predictions were transformed back to the original scale, **Model 5 achieved the best metrics: MAE of 123,967, MSE of 26,571,843,532, and R^2 of 0.568**, indicating the highest accuracy and explanatory power achieved yet. This iterative approach ensured the development of a robust and reliable linear regression model for predicting house prices.

Model 5 Diagnosis



Conclusion

The iterative modeling process culminated in the development of a robust predictive model (Model 5) that estimates the increase in home value based on `sqft_living`, `grade`, `sqft_above`, `bathrooms`, and `bedrooms` as key features. The model's R^2 of 0.57 on the test set indicates that 57% of the variance in housing prices can be explained by these key features. This strong explanatory power between key features and home prices suggest that strategic renovations can substantially increase a property's value.

Recommendations:

- 1. Focus on High-Impact Renovations:** The agency should advise homeowners to prioritize renovations that significantly increase the living area, improve the quality of construction, and add more functional spaces such as bathrooms and bedrooms. These improvements are shown to have the most substantial impact on home value.
- 2. Use the Predictive Model for Client Consultations:** Incorporate the predictive model into client consultations to provide data-driven estimates of potential home value increases from specific

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

Contributors 5



Languages

● Jupyter Notebook 100.0%