🏠 Predicting House Prices Using Regression Models

An End-to-End Machine Learning Project with Real Estate Data.

by Harshal John Robson

Project Objective

Goal

Predict house prices from physical and locational features using regression.

Key Questions

- Which features impact price most?
- Which regression model offers top accuracy?
- Can regularization and polynomial transforms help?

Data Overview

Dataset: King County House Sales, Seattle

Source: IBM Developer Skills Network

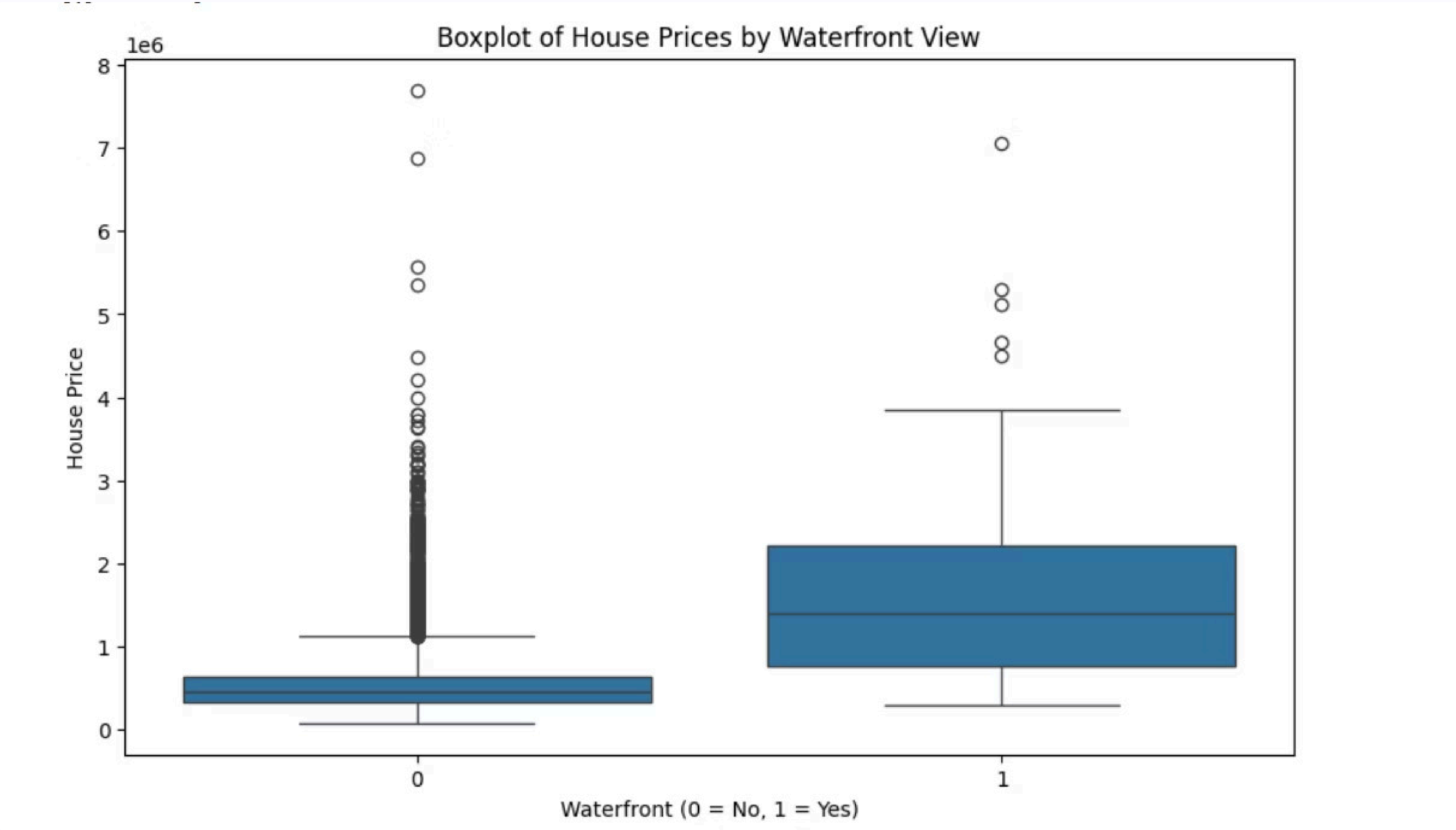
Size: ~21,600 entries, 21 features

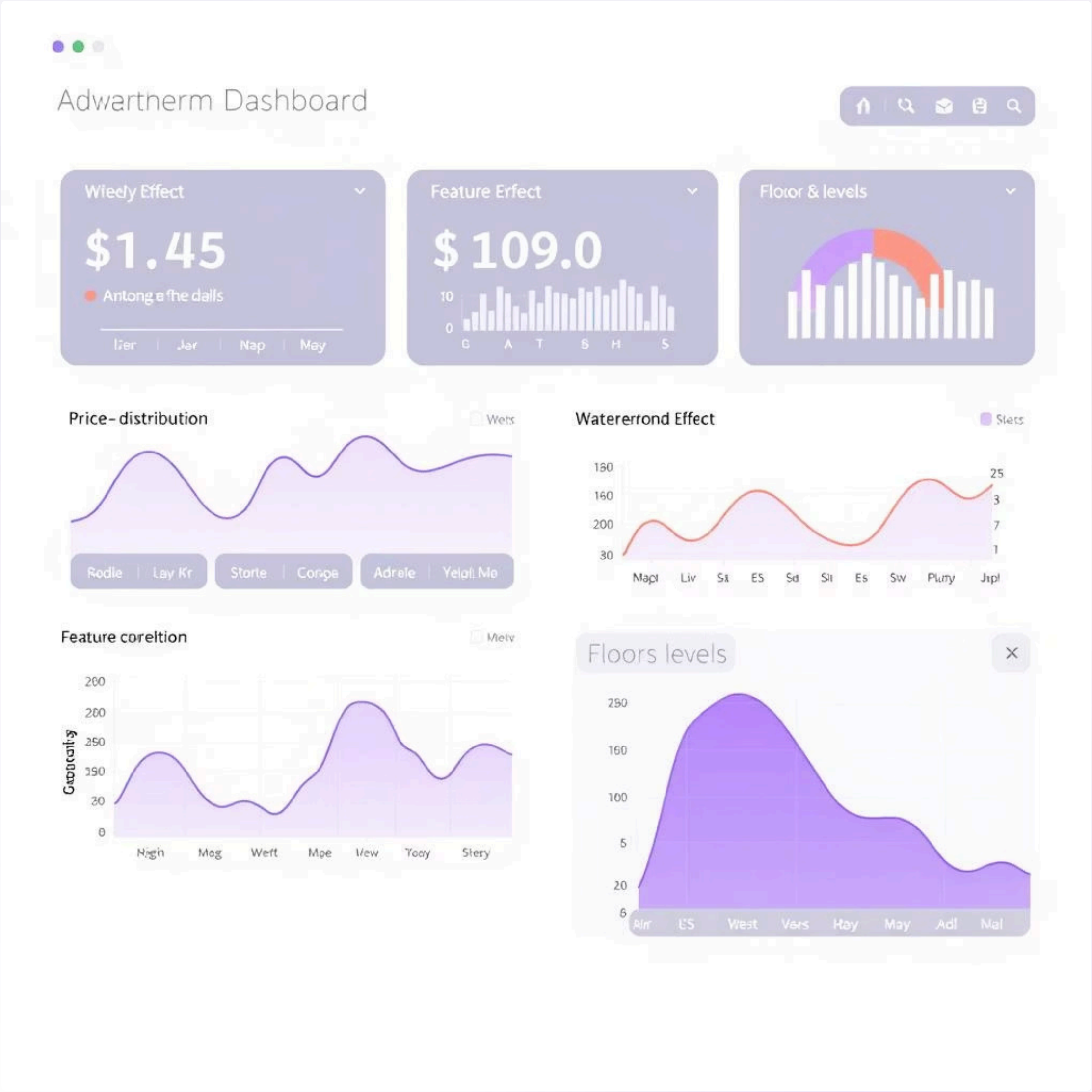
Target Variable: price

Features: sqft_living, bedrooms, bathrooms, lat, grade, waterfront, etc.

Missing Values Handling

bedrooms and bathrooms missing values replaced with mean.


Boxplot of House Prices by Waterfront View
1e6
8
7
6
5
4
3
2
1
0
House Price
0
1
Waterfront (0 = No, 1 = Yes)



Data Exploration

Price Distribution

Skewed distribution indicates possible outliers.

Waterfront Effect

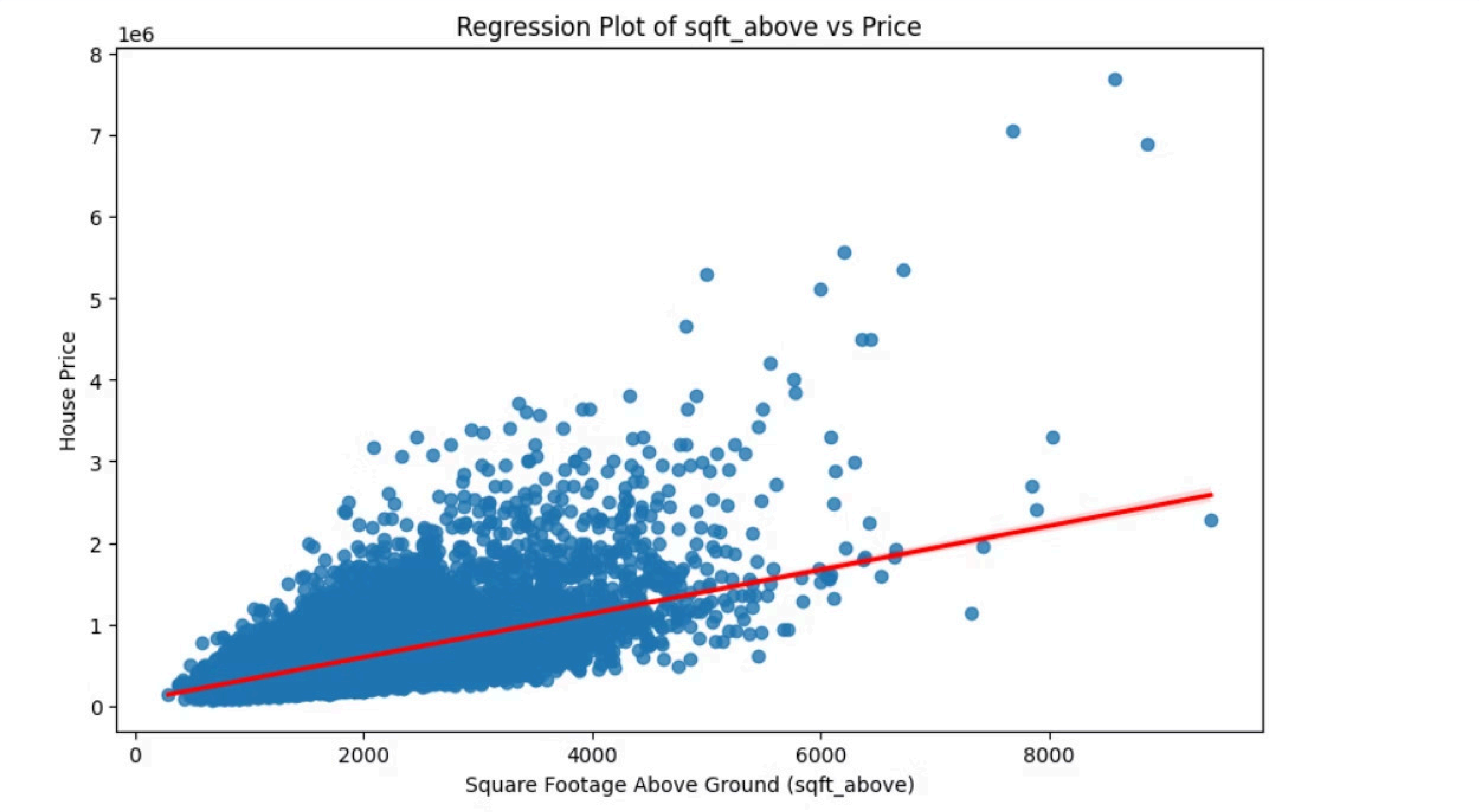
Boxplot reveals higher prices and more outliers for waterfront houses.

Feature Correlation

sqft_above shows a positive correlation with price.

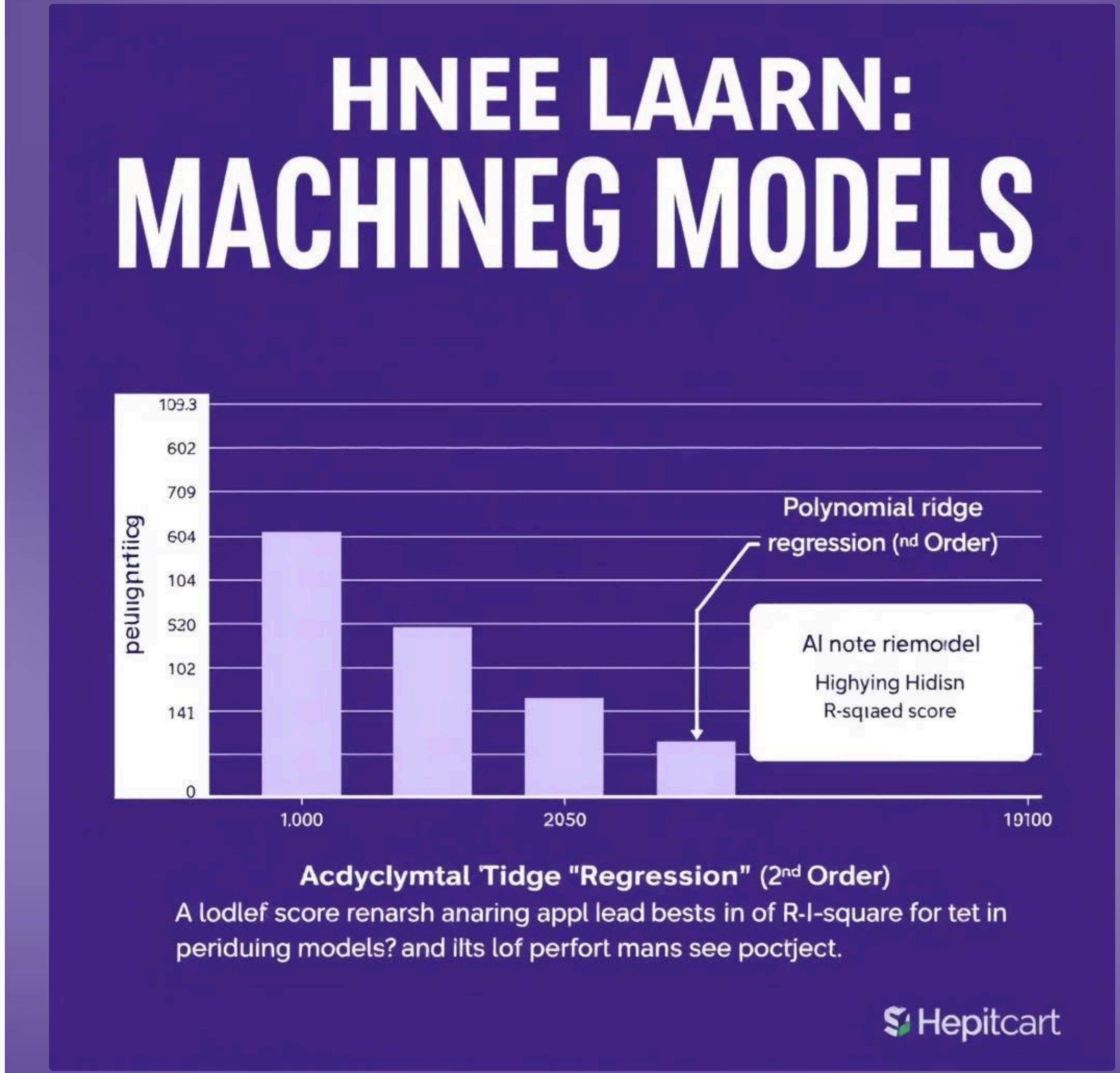
Floors

Value counts analyzed for unique floor levels in dataset.


Regression Plot of sqft_above vs Price
1e6
8
7
6
5
4
3
2
1
0
House Price
0
2000
4000
6000
8000
Square Footage Above Ground (sqft_above)

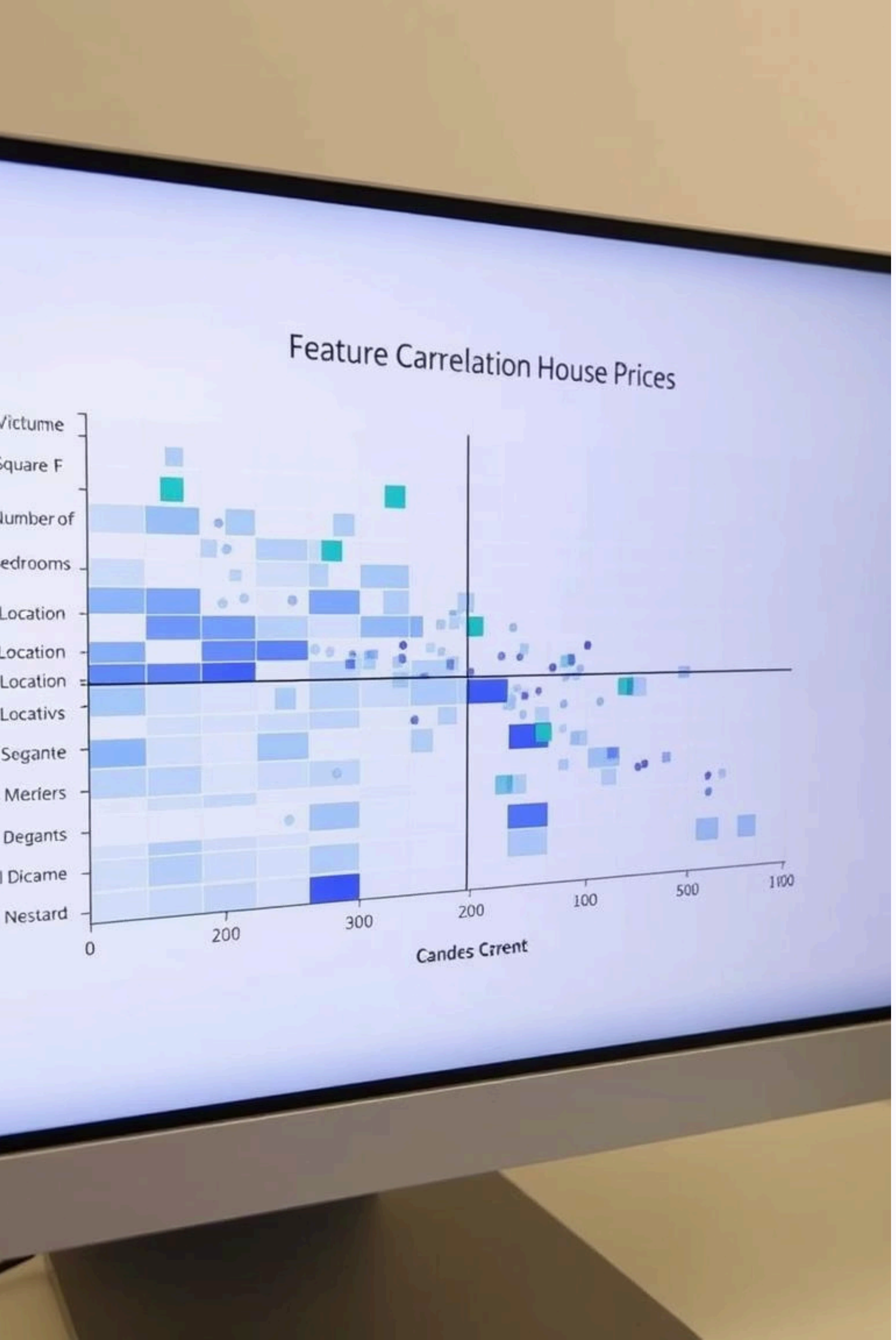
Model Building & Evaluation

Model	Features Used	R^2 Score
Simple Linear Regression	sqft_living	~0.49
Multiple Linear Regression	11 selected features	~0.70
Pipeline (Polynomial + Scaler + Linear Regression)	11 features	~0.79
Ridge Regression (α=0.1)	11 features	~0.70
Polynomial Ridge Regression (2nd Order)	11 features	**~0.84** ✅



Acdyclymtal 'Tidge "Regression" (2nd Order)

A lodlef score renarsh anaring appl lead bests in of R-I-square for tet in periduing models? and ilts lof perfort mans see poctject.



Insights & Findings

Key Features

sqft_living, grade, and latitude strongly correlate with price.

Best Model

Polynomial Ridge Regression achieved highest R^2 score (~0.84).

Regularisation Benefits

Ridge reduces overfitting and improves generalization.

Pipeline Advantage

Simplifies feature scaling and model training steps.

Conclusion & Learnings

Model Comparison

Built multiple regression models and evaluated their performance.

Data Handling

Handled missing data and leveraged pipelines effectively.

Feature & Preprocessing

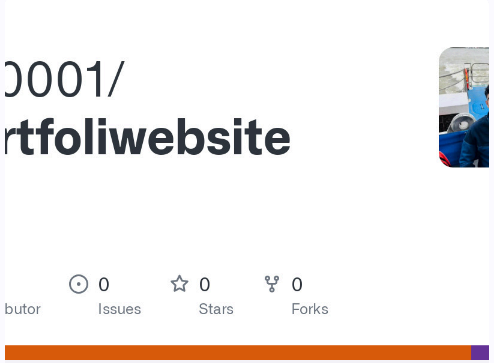
Importance of feature selection and preprocessing confirmed.

Future Work

Approach ready for scaling to other housing markets and platforms.

Project Files & Links

Code Notebook



)0001/
rtfoliwebsite

butor 0 Issues 0 Stars 0 Forks

GitHub

Portfoliwebsite/Housi...

Contribute to hjr0001/Portfoliwebsite...

Dataset Source

Access original dataset here: [KC House Data CSV]()

Portfolio Page

hjr0001.github.io

Harshal John Robson | Portfolio

I'm Harshal John Robson, a dedicated and passionate Master's candidate in Business Analytics at the University of Leeds. My journe...