Real Estate Price Prediction Using Machine Learning

CATHERINE MAINA

MORINGA SCHOOL, DATA SCIENCE PROGRAM

EXECUTIVE SUMMARY

Predicting house prices is essential for buyers, sellers, and investors.

This project builds machine learning models to estimate house prices based on historical housing data.

Goal: Provide accurate, data-driven property valuations.

PROBLEM STATEMENT

- Pricing homes is complex due to market variability.
- Stakeholders seek reliable estimates based on tangible features.
- Objective: Build regression models that predict house prices using features like size, location, and amenities.

OBJECTIVES

- Provide accurate, data-driven property price estimates to support better decision-making for agents, developers, buyers, and sellers.
- Analyze and identify key drivers of house prices, such as square footage, location, and amenities.
- Automate the valuation process to reduce manual effort and pricing subjectivity.
- Detect market trends and pricing anomalies for investment insights.
- Simulate how property improvements (like renovations) affect value.
- Build a predictive tool that can be integrated into real estate platforms.

DATASET OVERVIEW

- Source: Kaggle — House Price Prediction Dataset.
- 21,000+ property listings.
- Features include bedrooms, bathrooms, sqft, city, ZIP, etc.
- Region: King County, WA.

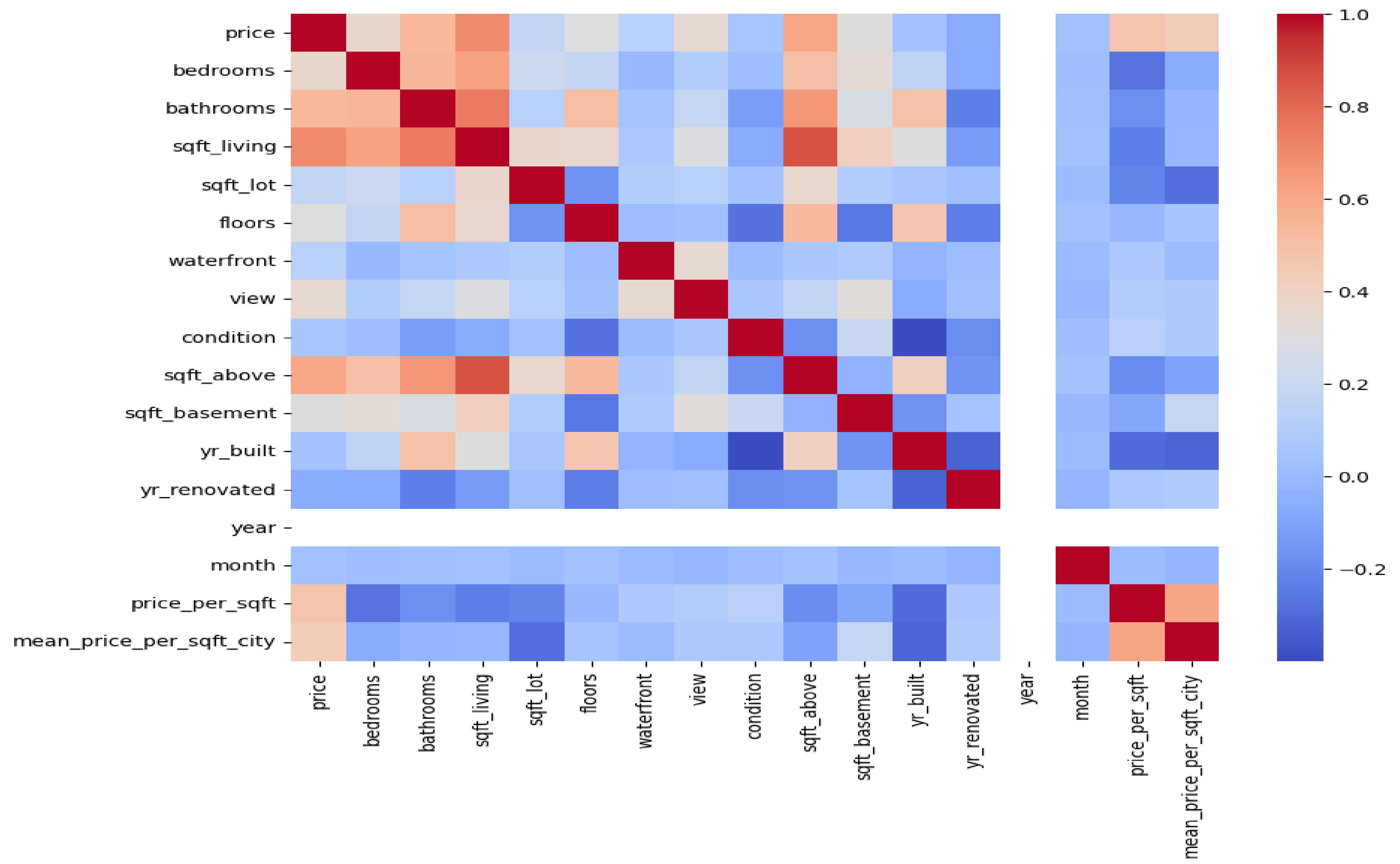
DATA PREPROCESSING

- Removed nulls, irrelevant fields (IDs, exact dates).
- Categorical encoding (e.g., city, ZIP code).
- Feature engineering: year built, year renovated.
- Outlier analysis — excluded 0-price entries only.

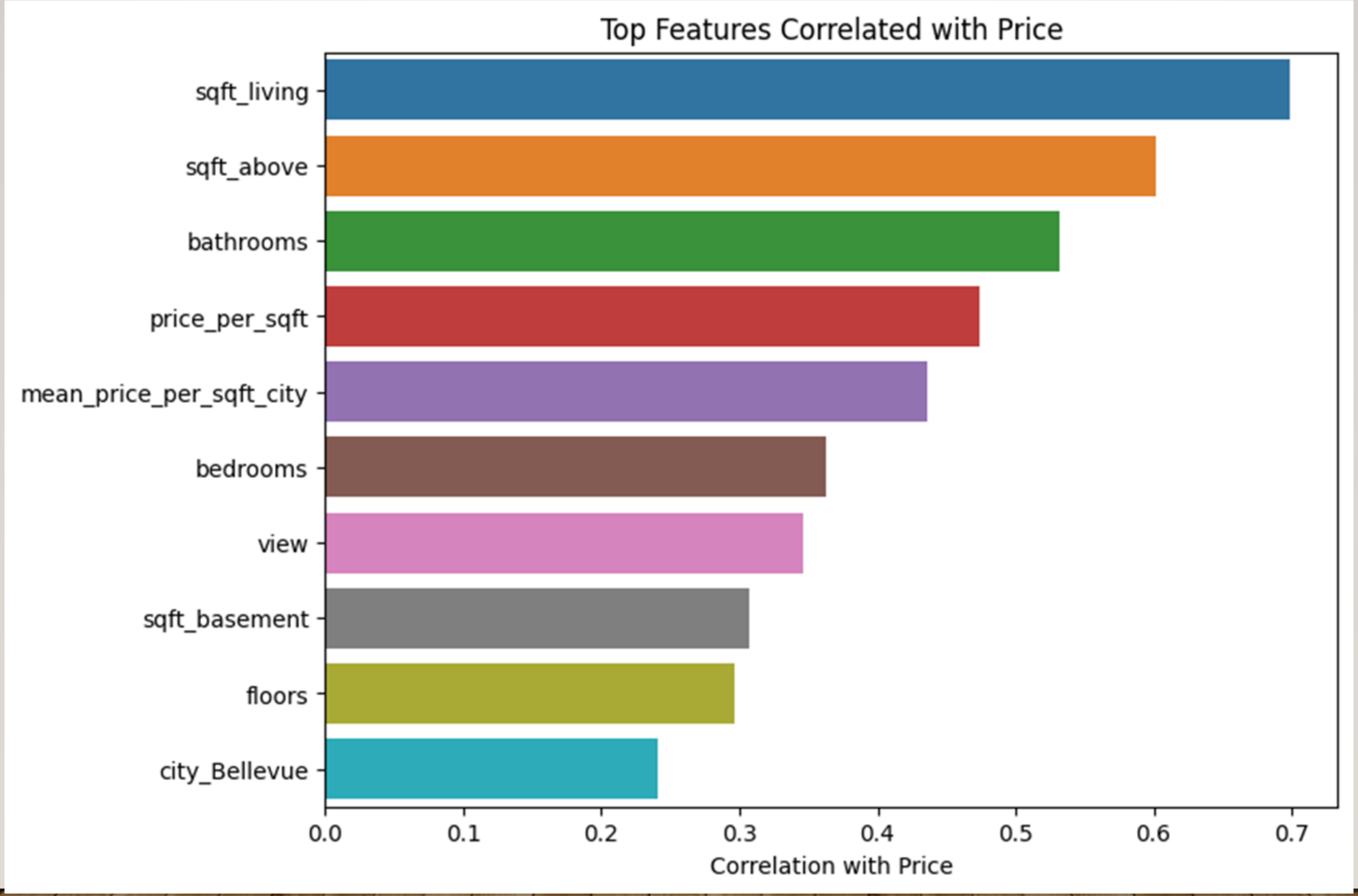
EXPLORATORY DATA ANALYSIS

- Correlation heatmap revealed strong price correlations with:
  - Square footage
  - Bathrooms
  - Location (city, ZIP)
- Visualizations: scatter plots, box plots

CORRELATION HEATMAP



TOP 10 FEATURES CORRELATED WITH PRICE



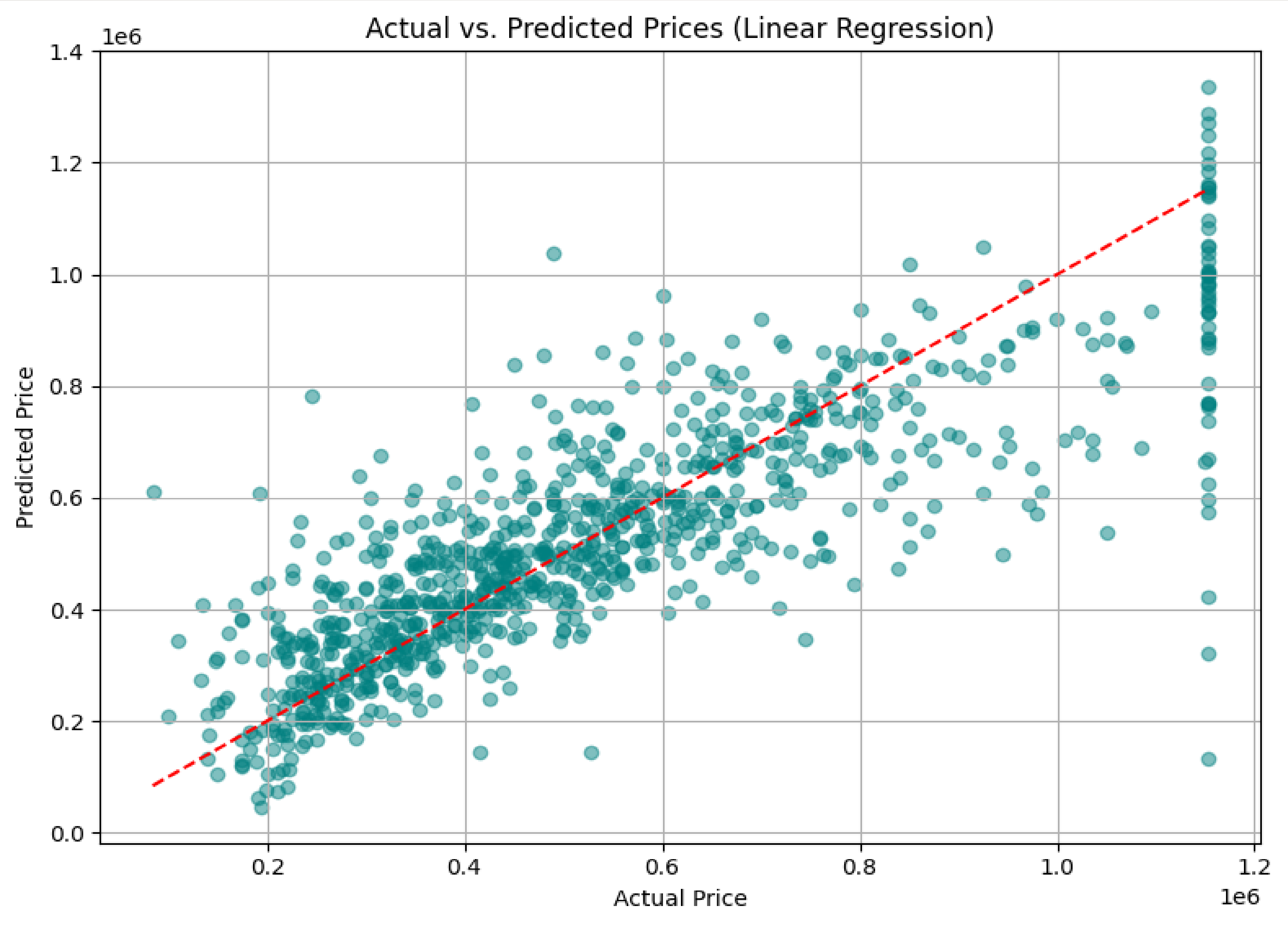
MODELING APPROACH

- Regression Models Used:
  - Linear Regression
  - OLS Regression
  - Decision Tree Regressor
  - Random Forest
  - Gradient Boosting
- Used GridSearchCV for hyperparameter tuning.

EVALUATION METRICS

- R^2 Score — model fit
- MAE — average error in dollars
- RMSE — penalizes large errors
- Baseline model: Linear Regression

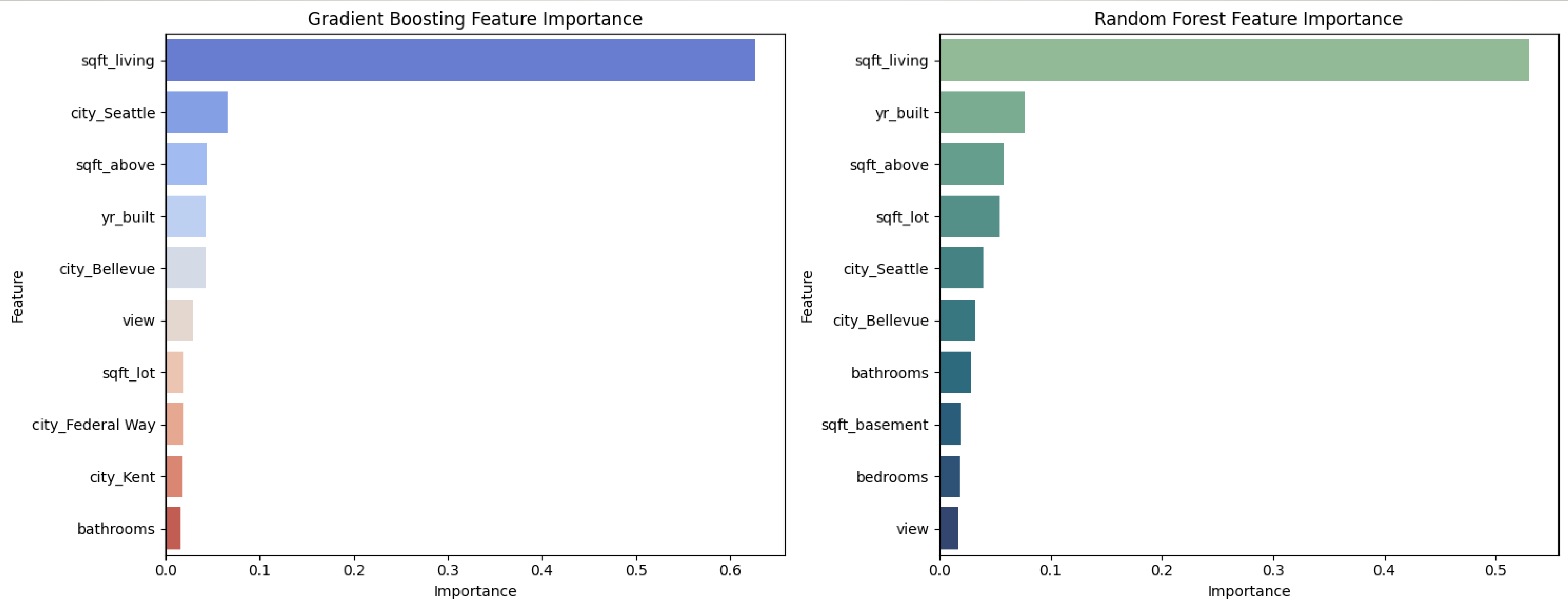
LINEAR REGRESSION RESULTS



MODEL COMPARISON

Model	R^2	MAE	RMSE
Linear Regression	0.6763	$101K	$146K
OLS Regression	0.712	~$99K	~$146K
Decision Tree	0.5670	~	$169K
Random Forest	0.6415	$105K	$153K
Gradient Boosting (Tuned)	**0.6799**	**$99K**	**$145K**

Top 10 Feature Importances



Both models agree that square footage (sqft_living) is the most important feature. Location-based features like city_Seattle and city_Bellevue also rank highly, affirming the significance of geography in real estate pricing.

INSIGHTS & FINDINGS

- Best model: Optimized Gradient Boosting Regressor.
- Linear & Gradient models outperformed tree-based models.
- Square footage and location were the strongest predictors.
- Outlier filtering did not significantly improve performance.

RECOMMENDATIONS

- Use ML-based tools in real estate platforms for price estimation.
- Regularly retrain models with updated market data.
- Use model insights to guide renovation investments.
- Deploy as a web tool or API for user access.

LIMITATIONS & FUTURE WORK

- **Limitations:**
- Geographic bias: limited to King County, WA.
- Sensitive to missing or inaccurate features.
- **Future Work:**
- Add economic indicators (e.g., interest rates).
- Test deep learning models.
- Incorporate maps/geospatial data.
- Build full-stack web app.

CONCLUSIONS

- The project aimed to build a predictive model for house prices using various regression models.
- Among all the models tested, **Optimized Gradient Boosting Regressor** performed best, with:
  - **R^2 Score: 0.6799**
  - **MAE: $99,444.53**
  - **RMSE: $145,521.95**
- However, performance across all models remained below an R^2 of 0.70, suggesting:
  - There is **still unexplained variance**, possibly due to omitted features (e.g. interior design, crime rate, school district).
  - The dataset may benefit from **more granular or external data**.
- **Linear models (OLS, Linear Regression)** performed surprisingly well, indicating that the relationship between features and price is mostly linear with minor non-linearity

RECOMMENDATIONS

- **Feature Engineering**: Add more relevant features (e.g. proximity to amenities, schools, crime rates) to improve model performance.
- **Outlier Handling**: Consider early-stage outlier analysis to improve data quality.
- **Ensemble Methods**: Continue using tree-based models like Gradient Boosting with tuning—they consistently outperform others.
- **Regular Validation**: Ensure you revalidate performance using data from different time periods or locations for generalizability.
- **Deployment Readiness**: While model performance is reasonable, it is not production-grade. A/B testing with real users could provide insights before full deployment.

Thank you!