🏠 PREDICTING HOUSE PRICES WITH MACHINE LEARNING & ADVANCED ENSEMBLES

A Data-Driven Approach to Real Estate Valuation

The team: Miquel, Gail, Helena and Julia

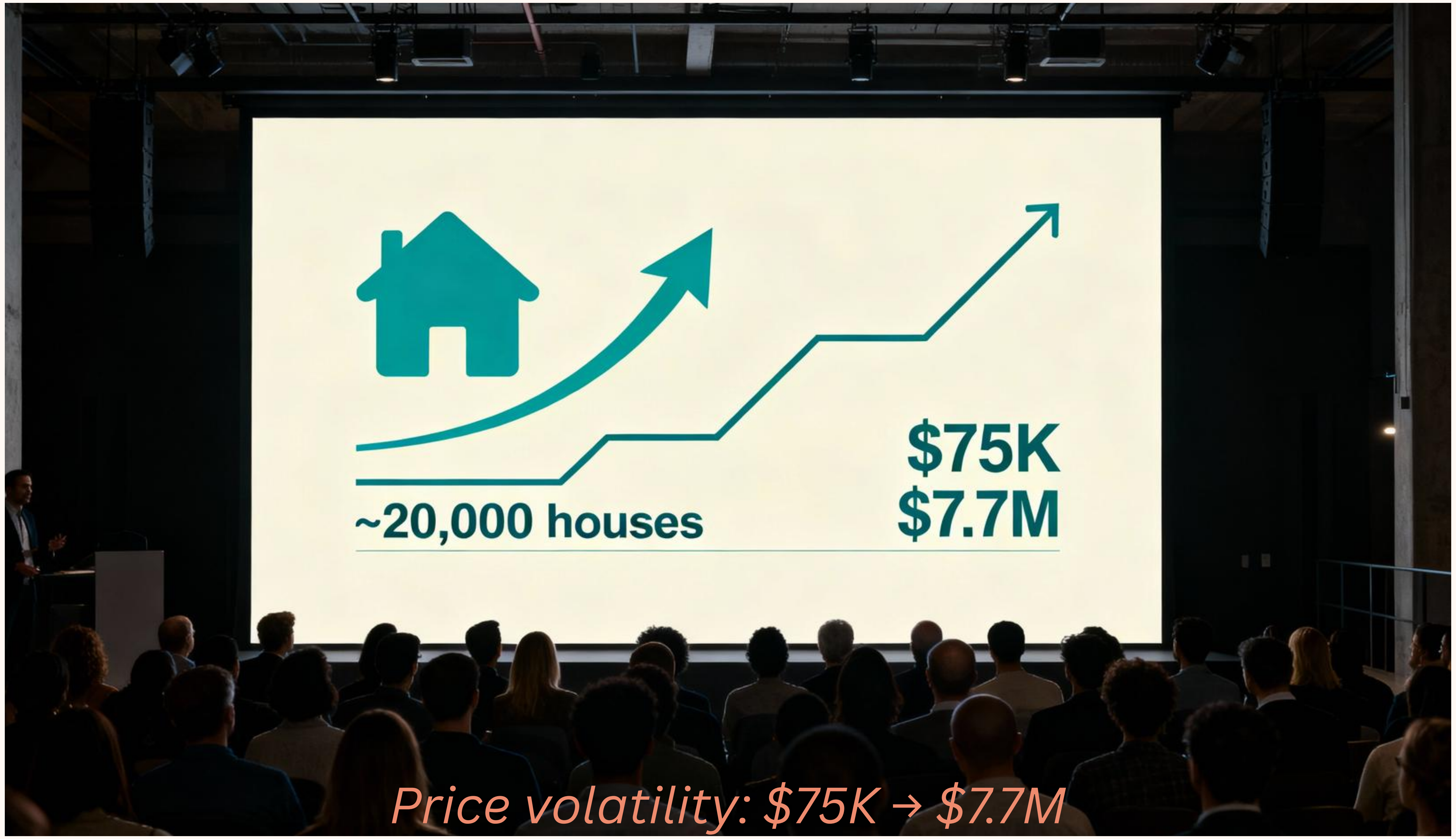
November 29th 2025

Ironhack

WHY PREDICT HOUSE PRICES?

The Challenge: Market Complexity Requires Data-Driven Solutions

- 📊 Real estate market is complex and non-linear
- ⏰ Current appraisal methods are slow and subjective
- 💡 ML can identify hidden patterns in data



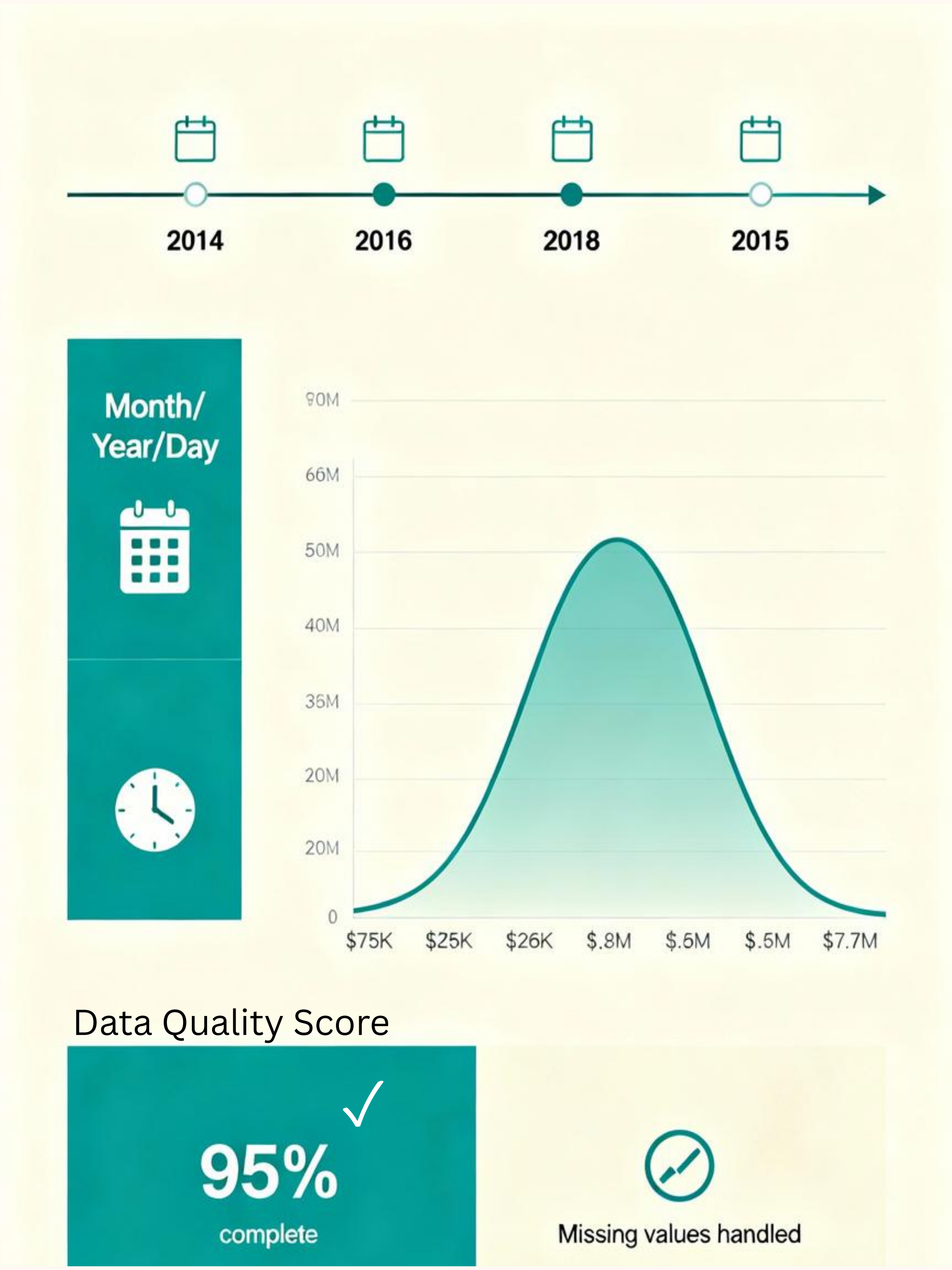
UNDERSTANDING KING COUNTY HOUSING DATA

DATASET SIZE

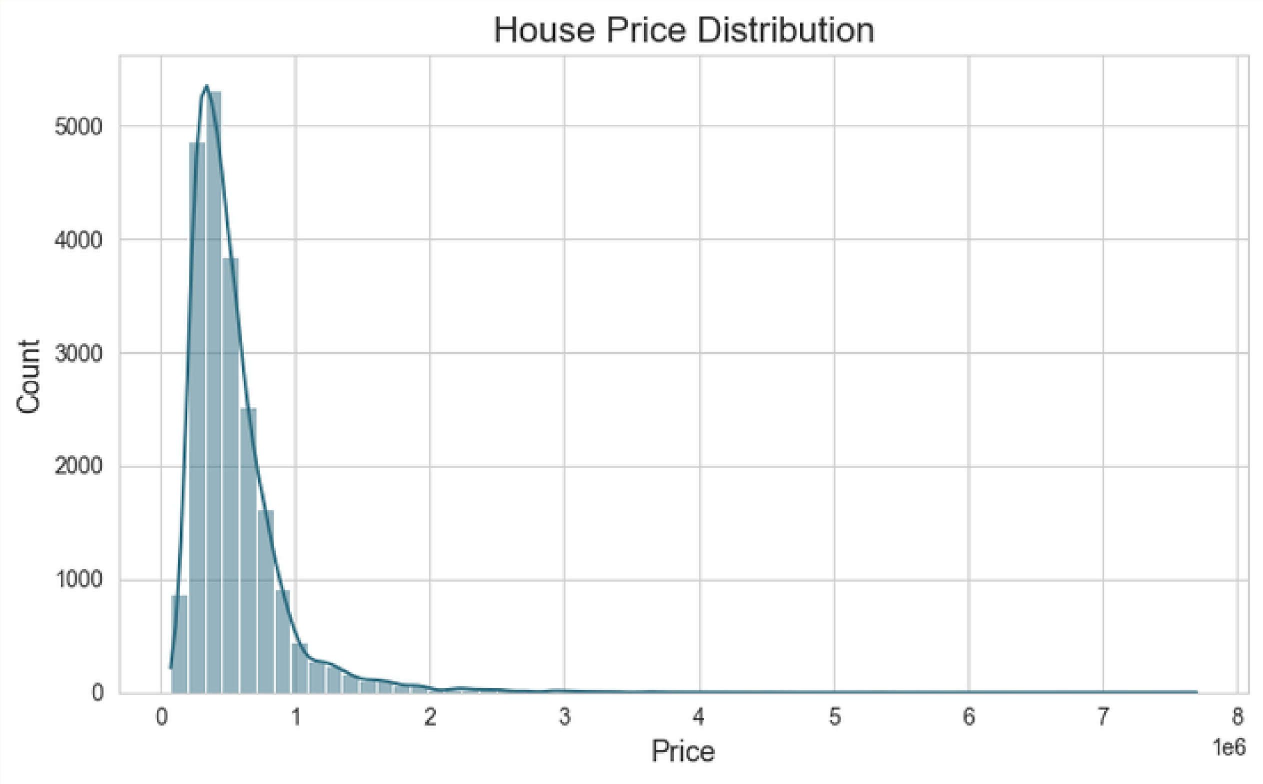
- 🏠 Dataset Size: ~20,000 houses in King County, WA (Seattle Metro Area)
- 💵 Price range: $75K–$7.7M
- 📍 Time Period: 2014–2015 Market Data
- Initial Features: 20 property/location attributes

DATA PREPARATION PART

- ✅ Handling missing values:
  - no missing values
  - meaningful zeroes
- ✅ Date transformations:
  - year_sold, month_sold, day_of_week
- ✅ Outlier detection and treatment
- ✅ Data scaling/normalization
  - during machine learning part
- 🔄 Initial feature engineering:
  - Created derived features for better model performance (detailed in modeling section)

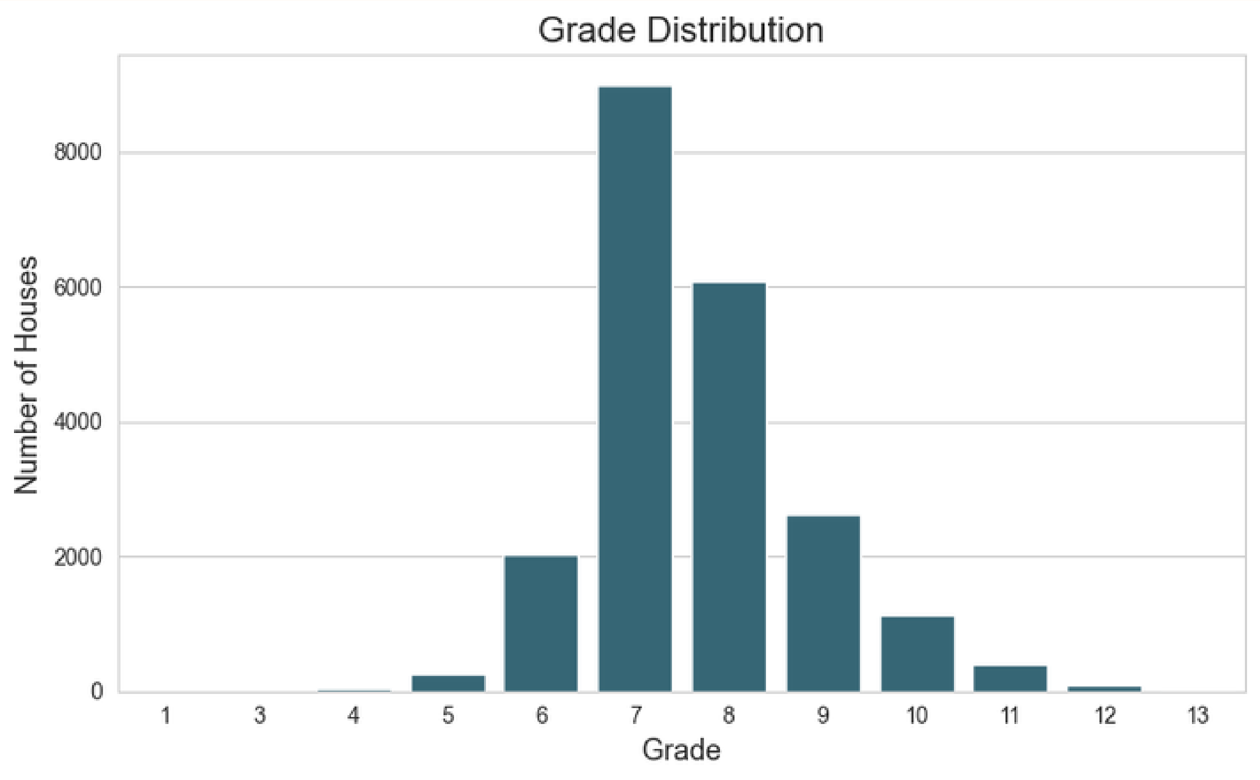


INITIAL FEATURE INSIGHTS: WHAT THE DATA REVEALS

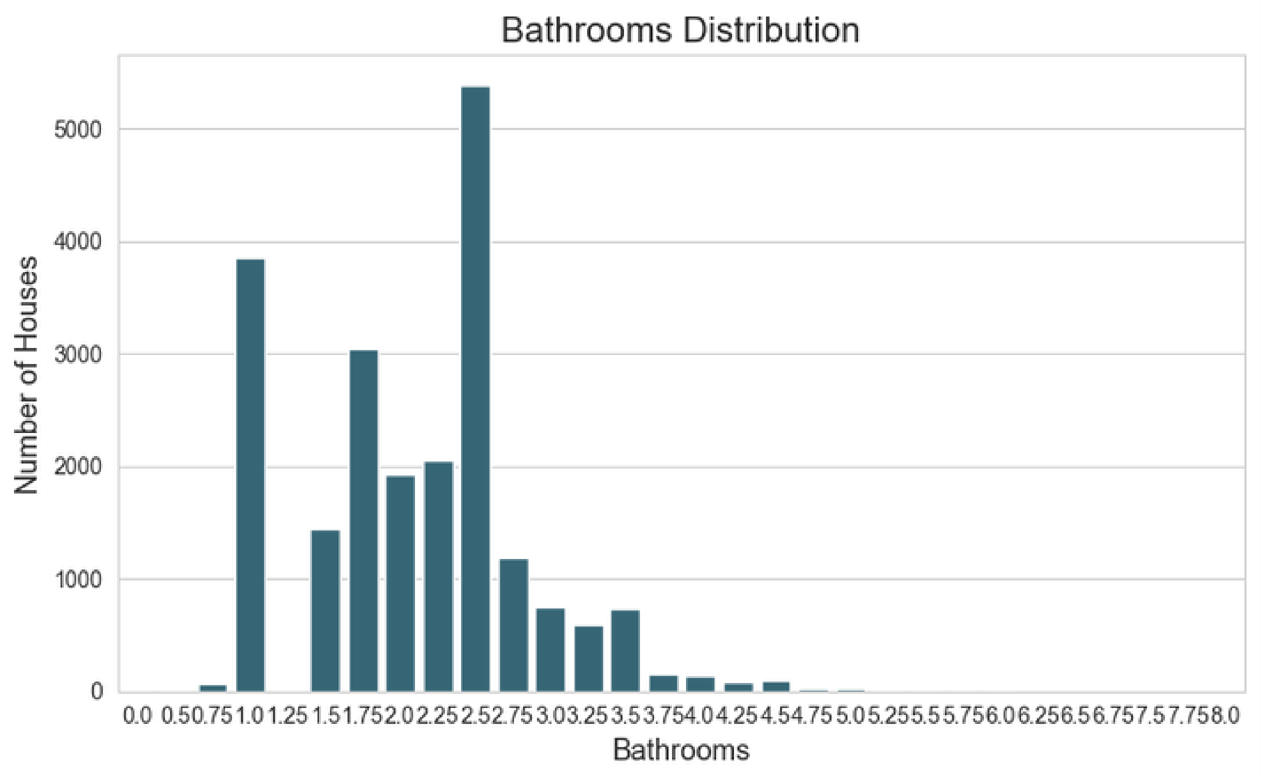


Right-skewed: Most homes \$300K–\$800K, luxury outliers >\$2M

Concentration: 70% of homes 1,500–4,000 , right-skewed

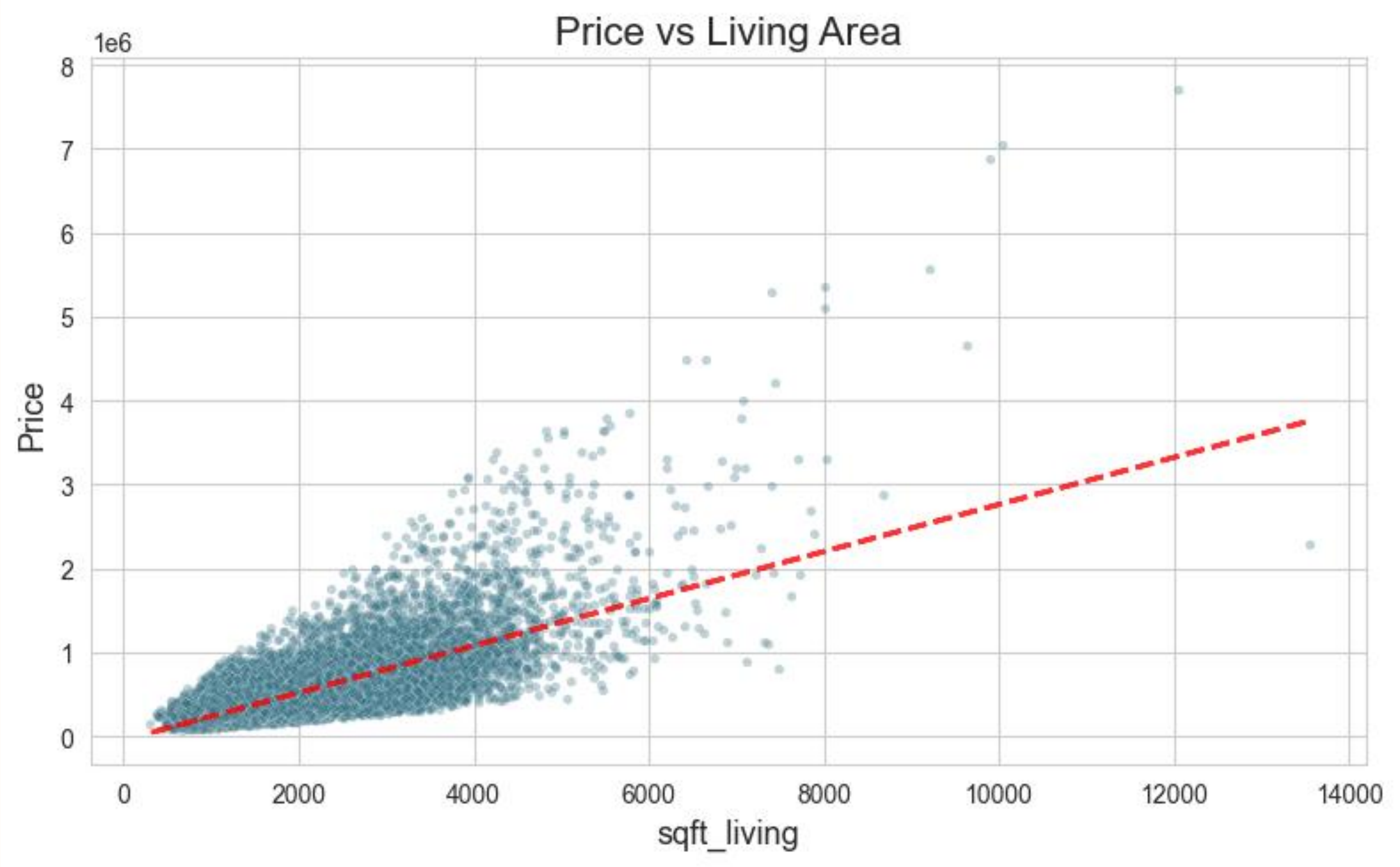


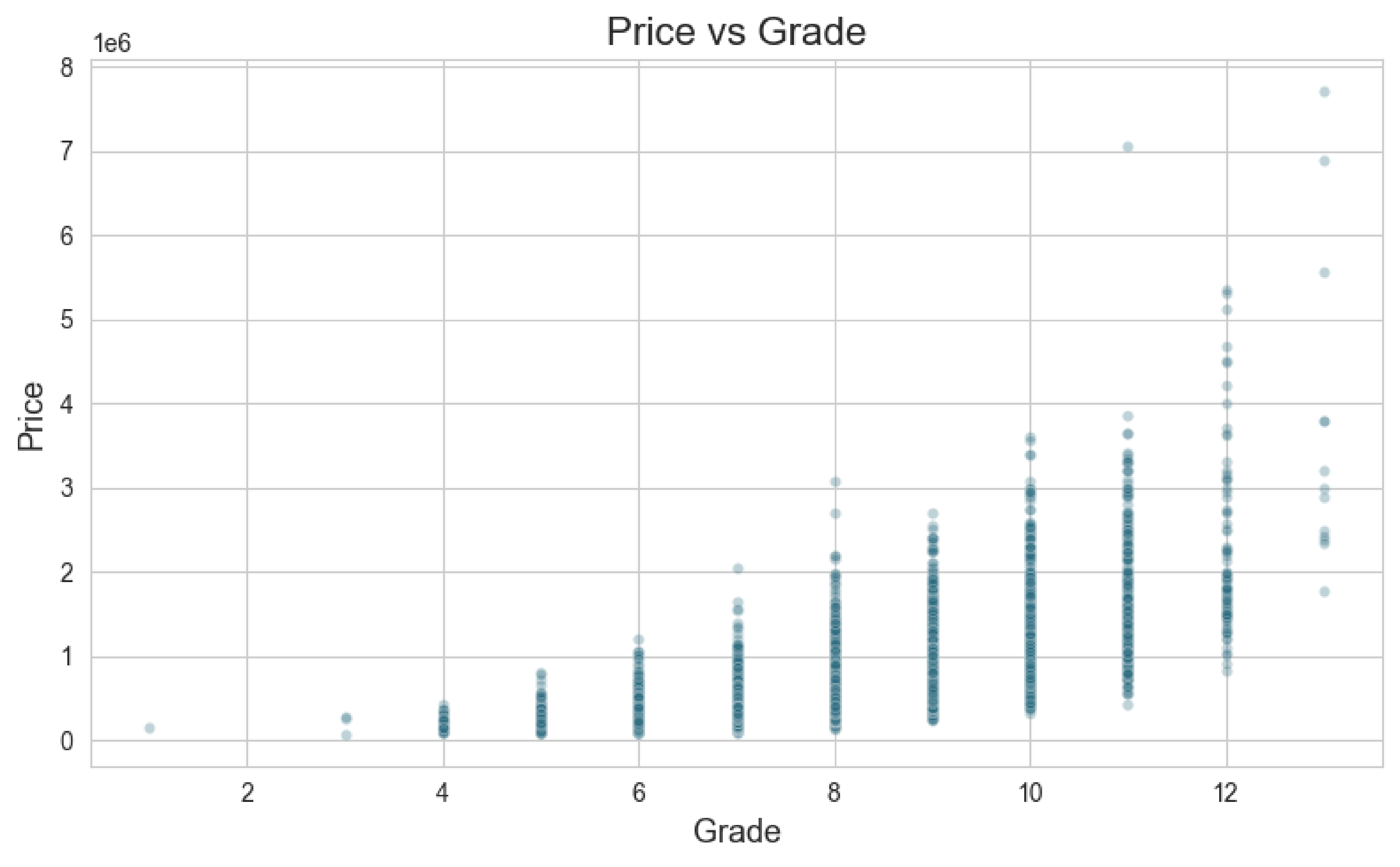
Concentration: ~70% of homes grade 7-8



Concentration: most homes have 1-2.5 bathrooms

RELATIONSHIP ANALYSIS: FEATURES VS. PRICE

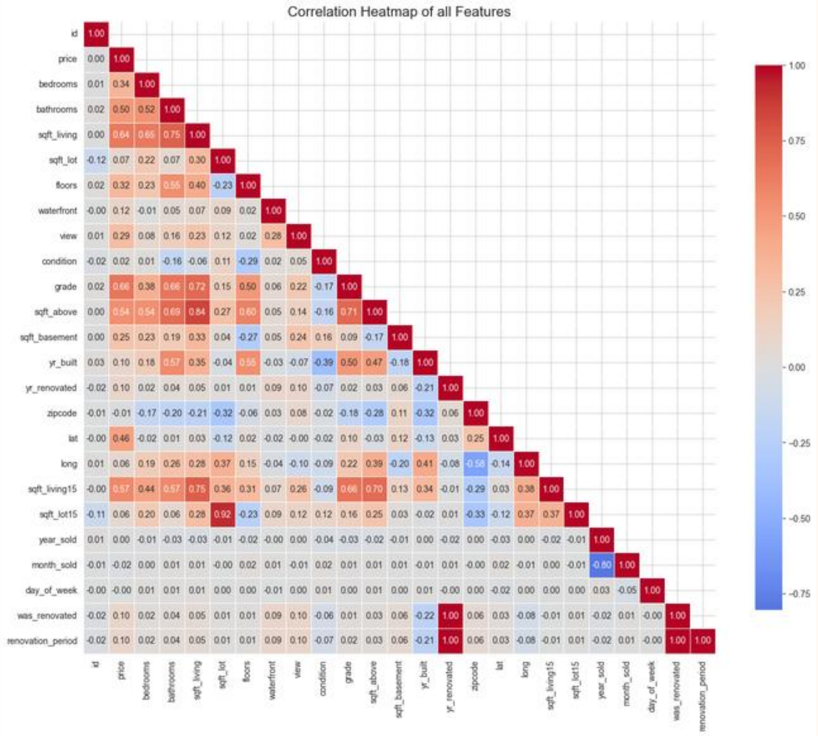




- *Living Area → Strong linear relationship (r = 0.64)*
- *Grade → Clear price tiers by quality level*

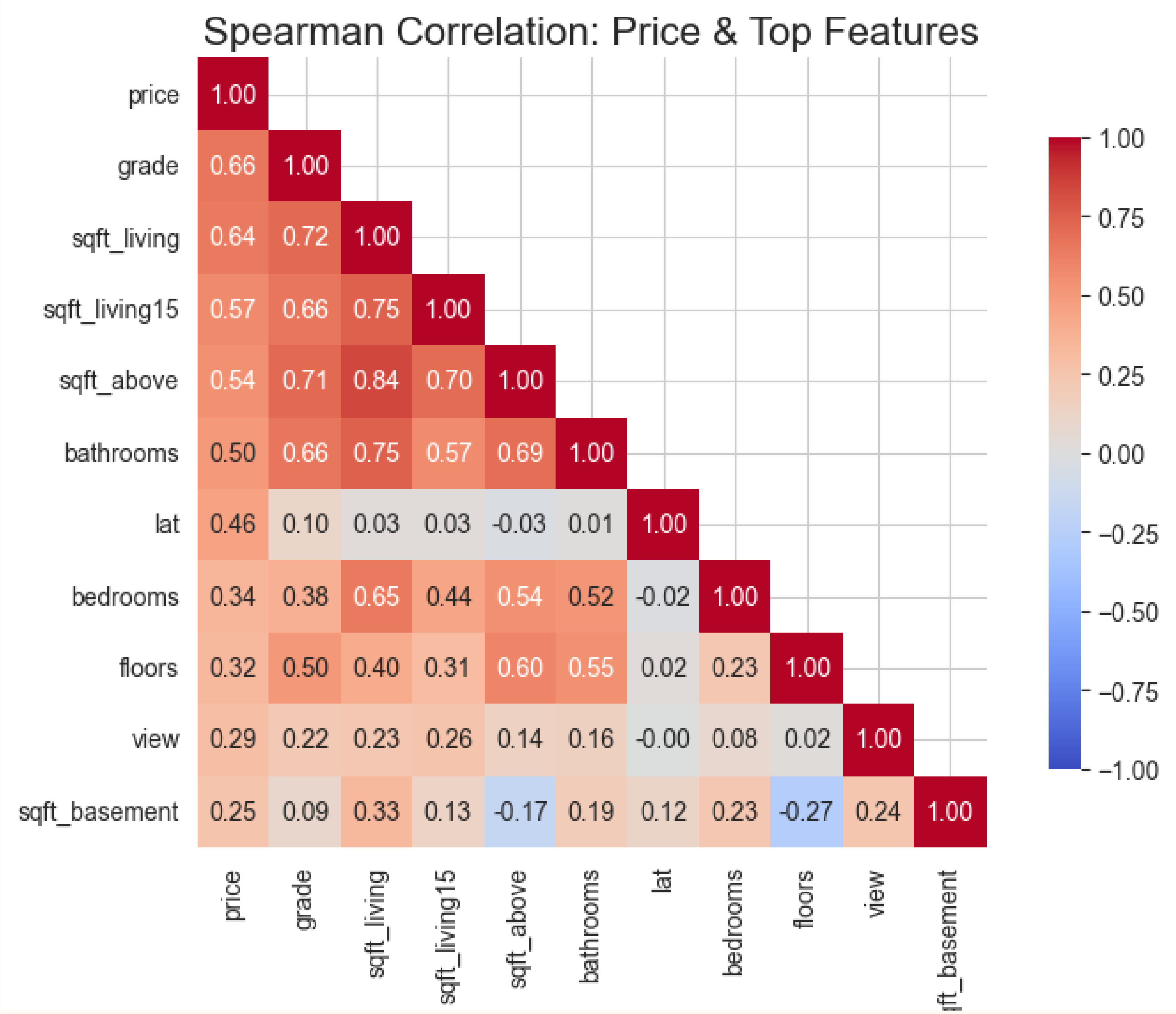
SPEARMAN CORRELATION: IDENTIFYING PRICE DRIVERS

Full Correlation Matrix (All Features)

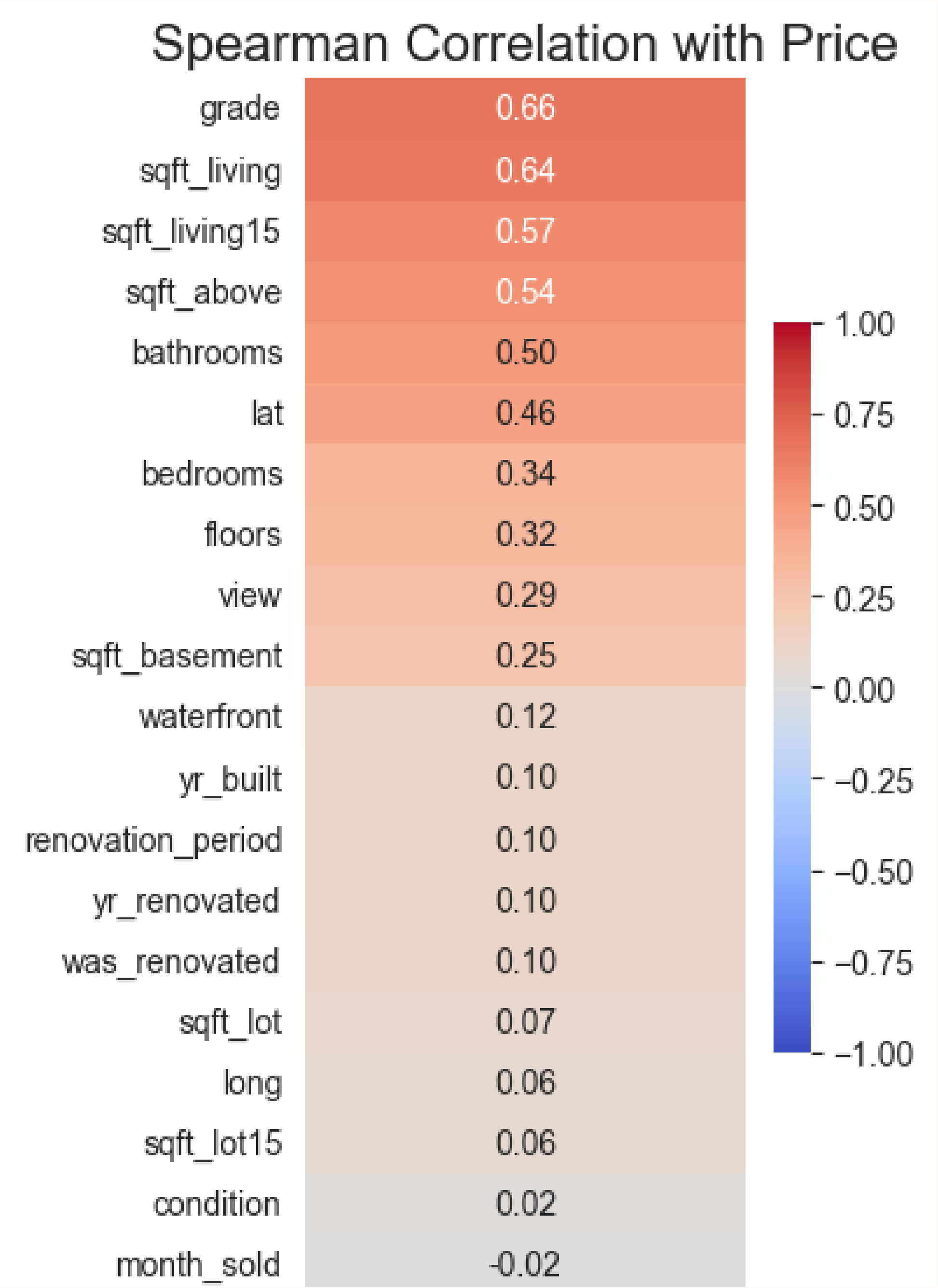


Key Finding: *Building quality and size dominate price; location adds nuance*

Spearman Correlations: Price & Top 10 Features



Feature Importance Ranking by Correlation



Grade, sqft_living, location (lat/long) are strongest predictors

ENGINEERED FEATURES: BUILDING SMARTER PREDICTORS

Why Feature Engineering?
Raw features alone miss important patterns. We combine existing features to capture business logic—living-area ratios reveal value density, age features capture depreciation, and log transforms handle extreme price ranges. These engineered signals dramatically improve model accuracy.

LIVING AREA

- ***total_sqft*** – Combined living + basement
- ***living_to_lot_ratio*** – Space intensity
- ***bath_per_bed*** – Comfort indicator
- ***living15_diff*** – Relative size vs neighborhood
- ***basement_share*** – Basement proportion
- ***has_basement*** – Binary basement flag

DENSITY

- ***lot_per_living*** – Land-to-building ratio

✓ *Reveals whether land adds value (suburban vs urban)*

AGE & RENOVATION

- ***house_age*** – Years since construction
- ***since_renovation*** – Years since last update
- ***was_renovated*** – Renovation status flag

✓ *Captures depreciation & modernization effects*

LOG TRANSFORMS

- ***log_price*** – Handle price skew
- **log_sqft_living** – Transform area
- **log_sqft_lot** – Compress outliers

✓ *Linearize relationships with target*

MODEL DEVELOPMENT STRATEGY

Testing Strategy

Linear Models
Baseline: Linear; Ridge, Lasso

✓ Simple interpretable models | Understand baseline performance

Tree Ensemble
Random Forest (tuned)

✓ Captures non-linear patterns | Handles interactions well

Boosting Methods
Gradient Boosting, XGBoost, AdaBoost

✓ Advanced ensemble learning | Best accuracy potential

Instance-Based
KNN Regression

✓ Memory-based approach | Test algorithmic diversity

Evaluation Metrics

R^2 (Coefficient of Determination)
% of variance explained by model (0–1 scale)

↑ Higher = Better fit

RMSE (Root Mean Squarred Error)
Average prediction error in dollars

↓ Lower = Better accuracy

Train vs. Test Gap
Difference between training and test R^2

↓ Lower = Balanced model

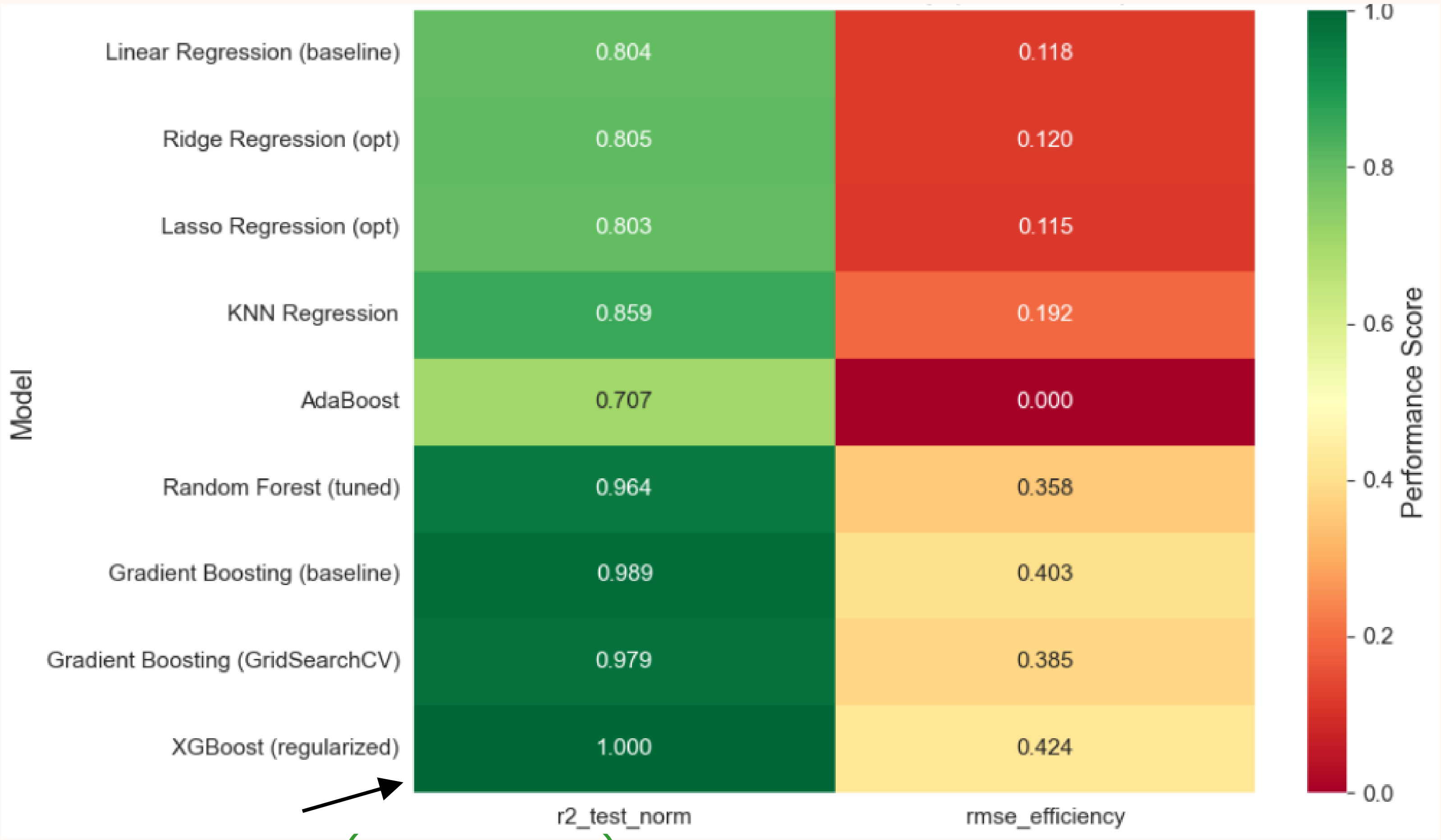
9 models tested across 4 families

Goal: Find best balance of accuracy, efficiency & generalization

MODEL PERFORMANCE COMPARISON

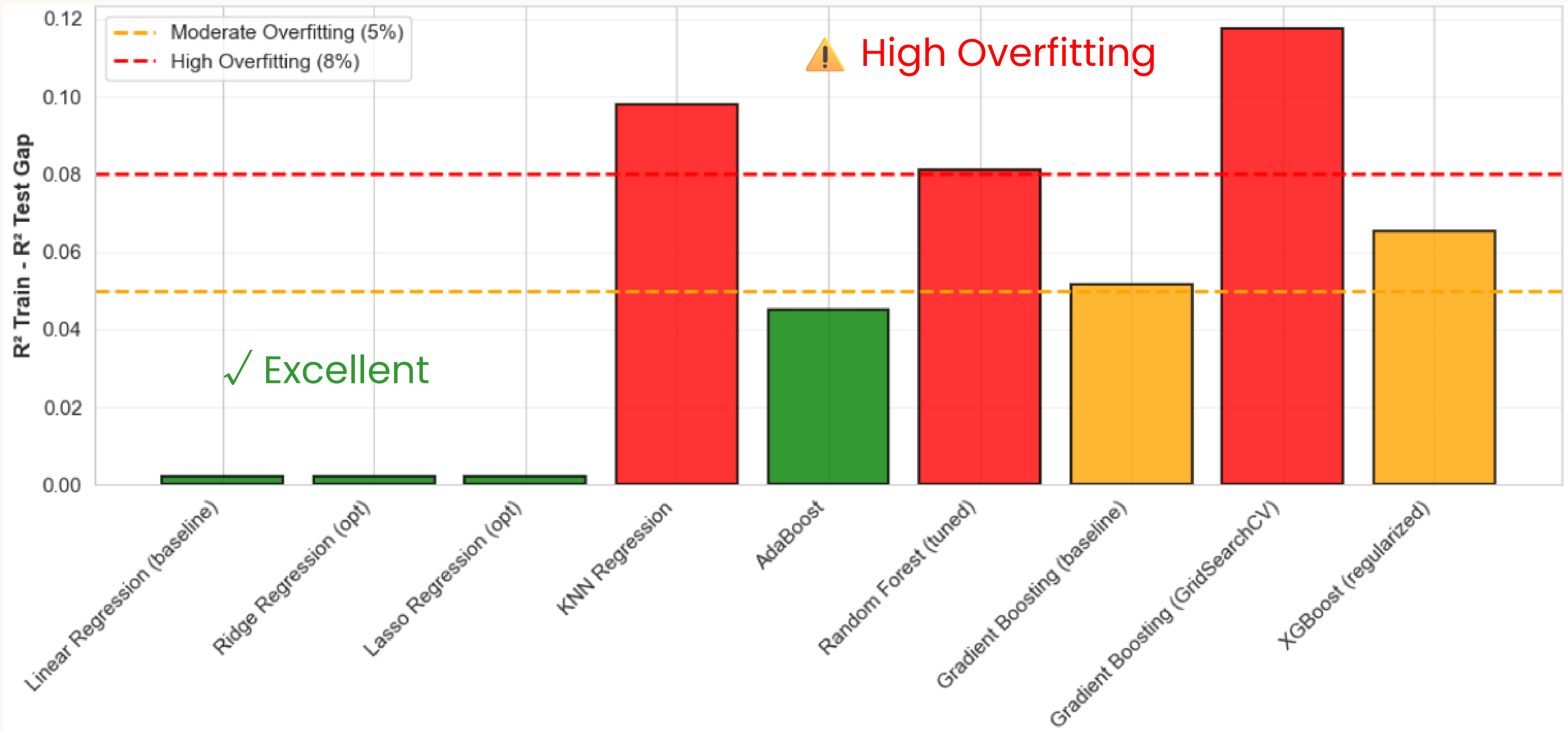
📊 Test Dataset Price Range: $75K – $7.7M | Middle 50% of homes: $322K – $645K | Median: $450K

Model Performance Heatmap (Normalized)



Best score = 1.0 (normalized), color: dark green

Overfitting Analysis



🏆 Winner: XGBoost

R^2 = 0.878 (explains 87.8% of variance)

Best accuracy across all models

💰 Error Magnitude

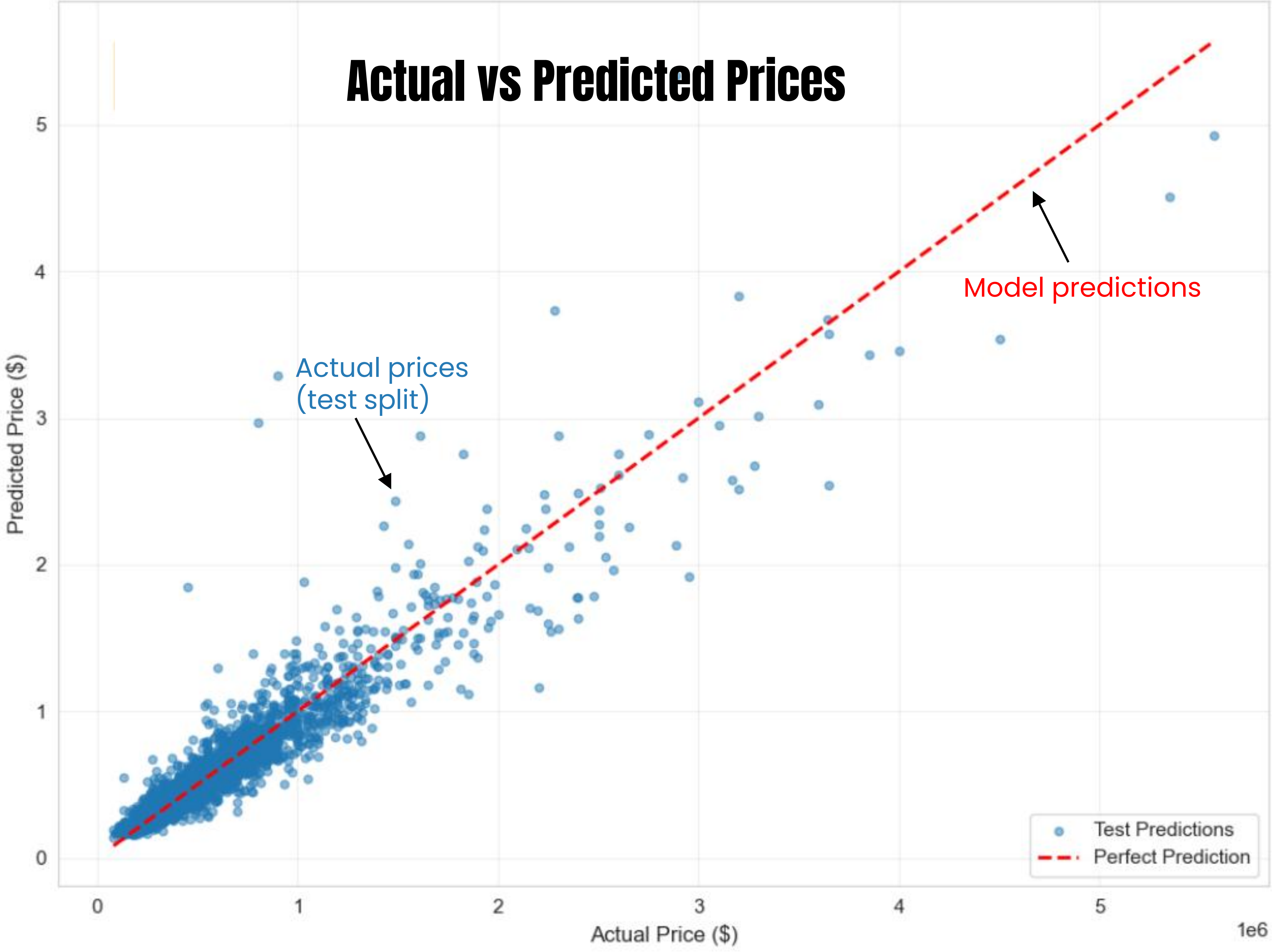
RMSE = $135,693

2.47% of the price range, ~ 30% of median price

⚖️ Generalization

Low overfitting gap (6.05%) shows balanced train-test performance

MODEL PREDICTIONS: ACCURACY ANALYSIS



R^2 Score

0.878

Explains 87.8 % of variance

RMSE

$135,694

Root mean squared error

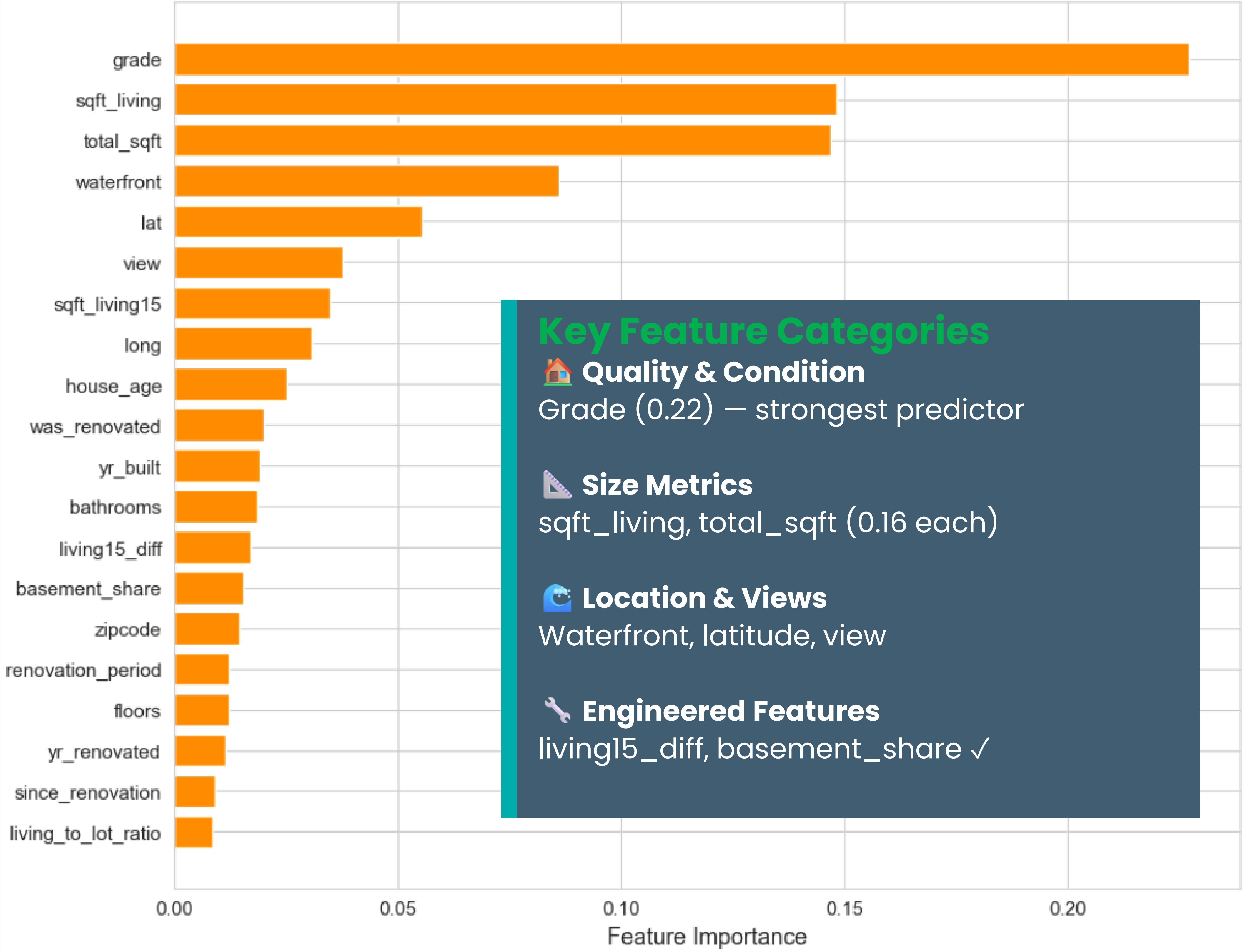
MAE

$70,787

Average prediction error

WHAT DRIVES PRICES? TOP FEATURES

Top 20 Feature Importances



Key Feature Categories

🏠 **Quality & Condition**
Grade (0.22) — strongest predictor

📐 **Size Metrics**
sqft_living, total_sqft (0.16 each)

🌊 **Location & Views**
Waterfront, latitude, view

🔧 **Engineered Features**
living15_diff, basement_share ✓

Business Insights

1. **Grade matters most**

   Home quality condition is the primary price driver

2. **Size is crucial**

   Living area and total square footage heavily influence pricing

3. **Location & features count**

   Waterfront, views and neighborhood locations add significant value

4. **Renovation history matter**

   Age and recent updates are meaningful price signals

LIMITATIONS & FUTURE WORK

⚠️ Current Limitations

Outlier predictions
Model underestimates ultra-luxury homes (>$2M). Rare high-priced outliers have fewer training examples.

Temporal Blind Spot
Model ignores market trends, economic cycles, and seasonal variations in housing demand.

Location Proxy
Lat/long serve as proxies. True neighborhood effects (schools, crime, amenities) not explicitly captured.

🚀 Future Improvements

Temporal Features
Add year-over-year trends, seasonal dummies, market indices. Track price evolution over time.

Granular Neighborhoods
Incorporate zip codes, school districts, crime data. Replace lat/long proxies with explicit features.

Ensemble + Deep Learning
Combine XGBoost with neural networks. Explore stacking and meta-learners for edge cases.

Current Model Strength: Accurate for typical properties in King County.

Next Phase: Refine for edge cases and market dynamics.