Housing Price Prediction: Comprehensive Analysis Report

1. Executive Summary & Problem Understanding

Dataset Description:

The dataset contains **4,600 residential property records** from the state of Washington, USA. Key variables include structural attributes (bedrooms, bathrooms, square footage, floors), renovation history, and precise location details (city, state-zip).

Problem Statement:

Real-estate stakeholders face significant challenges in accurately valuing properties due to the complex interplay between physical features and location. A data-driven approach is required to minimize appraisal bias and provide reliable fair-market estimates.

Objective of the Analysis:

To perform thorough data cleaning and EDA, engineer predictive features, and build robust Machine Learning (ML) and Artificial Neural Network (ANN) models to accurately forecast house prices.

Summary of Key Findings:

- **sqft_living** and **Location (Target Encoded)** are the strongest predictors of value.
- **Waterfront** properties command a massive premium (~2x) despite being rare (<2%).
- Seasonality exists; Spring/Summer months (Apr–Jul) show higher average transaction prices.
- The **Tuned XGBoost** model outperformed all other architectures, including the deep ANN.

Final Model Recommendation:

The **Tuned XGBoost** regressor is recommended for production due to its high R^{2} score (0.7737), lowest error rates, and better interpretability through feature importance rankings.

2. Data Cleaning & Preprocessing

Cleaning Procedures:

- **Zero-Price Removal:** Removed records where price was $0 to eliminate erroneous listings.
- **Column Cleanup:** Dropped street, date, country, and yr_renovated to reduce dimensionality and noise.
- **Binary Flagging:** Converted yr_renovated into a binary is_renovated feature.

Outlier Treatment:

Applied **IQR-based Capping (Winsorization)** to continuous variables (price, sqft_living, sqft_lot, etc.). This retained data points while bounding the influence of extreme luxury properties that could skew the model.

Feature Engineering:

1. **House Age:** (2025 – yr_built) to capture property depreciation/vintage value.
2. **Basement Ratio:** (sqft_basement / sqft_living) to assess the impact of below-grade space.
3. **Living-to-Lot Ratio:** (sqft_living / sqft_lot) to measure density.

Categorical Encoding:

High-cardinality features (city and statezip) were processed using **Target Encoding**. This converted locations into their mean house price, preserving spatial value relationships without creating 100+ dummy variables.

3. Exploratory Data Analysis (EDA)

Univariate Analysis:

- **Price:** Highly right-skewed; most homes are under $800k. Log-transformation successfully normalized the target variable.
- **Structure:** Most homes feature 3 bedrooms and 1.5-2.5 bathrooms.

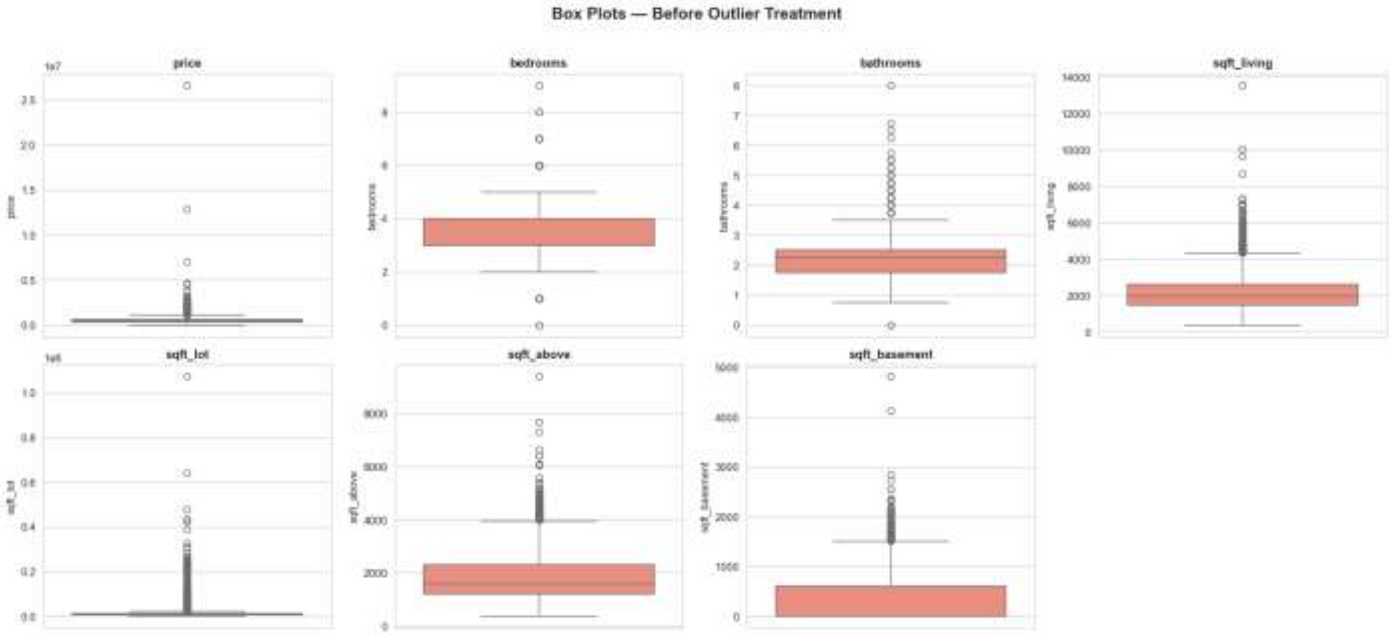
Bivariate Analysis:

- **Size vs. Price:** sqft_living shows the clearest positive linear trend.
- **View vs. Price:** Properties with a 'View' rating of 4 show significantly higher median prices compared to rating 0.

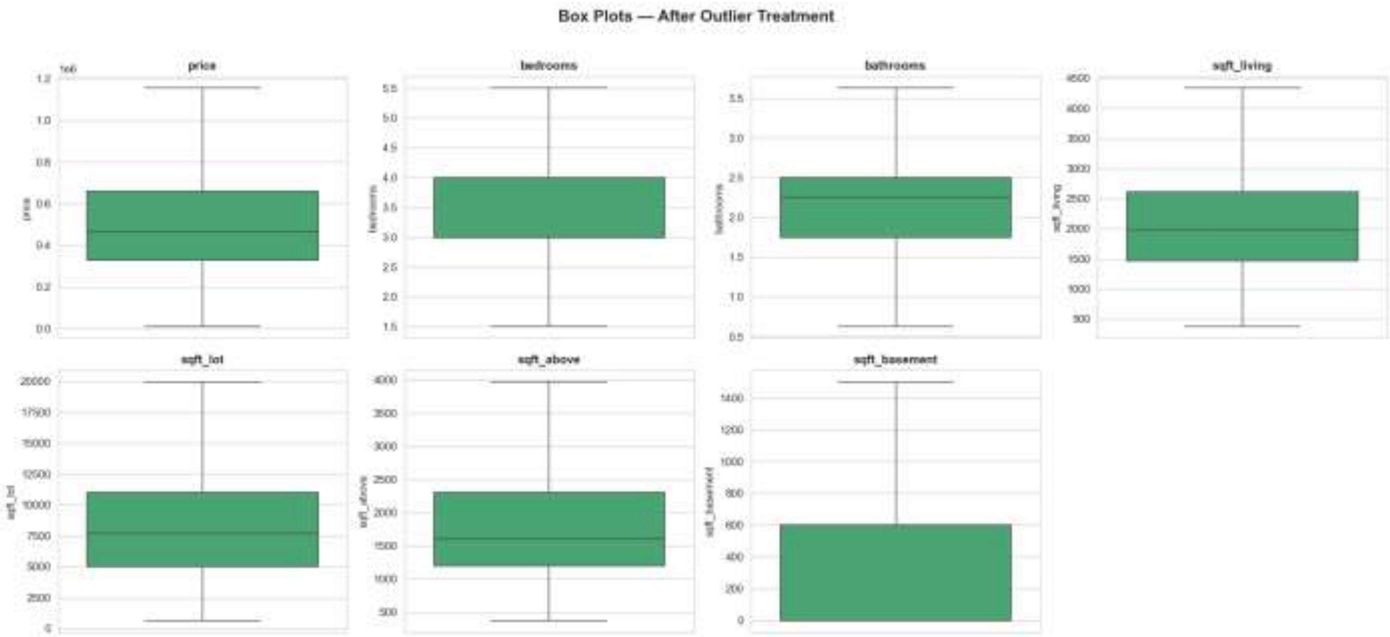
Correlation Insights:

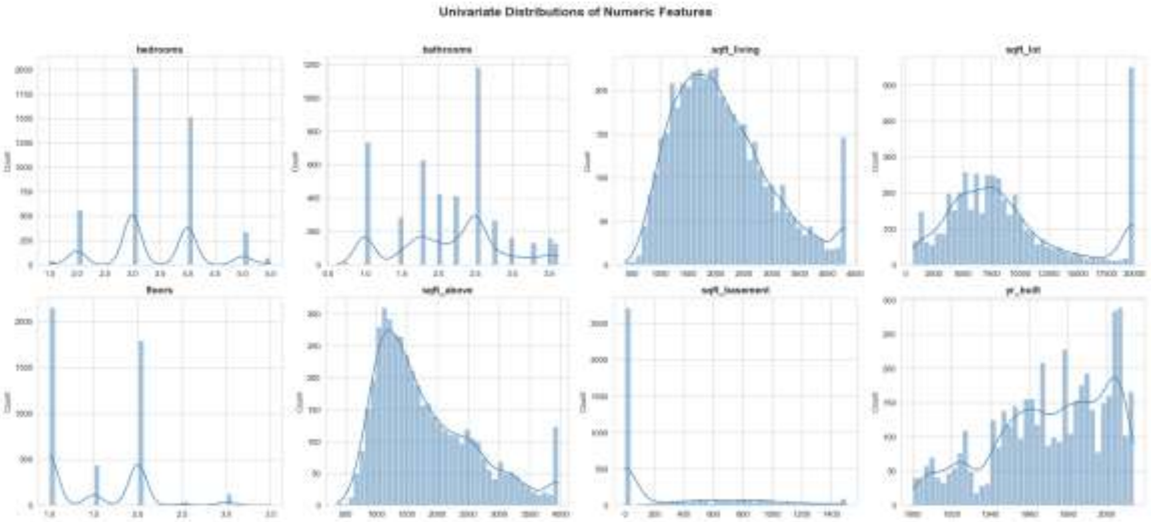
- Strongest positive correlation with price: sqft_living ($r \approx 0.70$).
- Multicollinearity identified between sqft_above and sqft_living ($r > 0.85$), requiring feature selection.

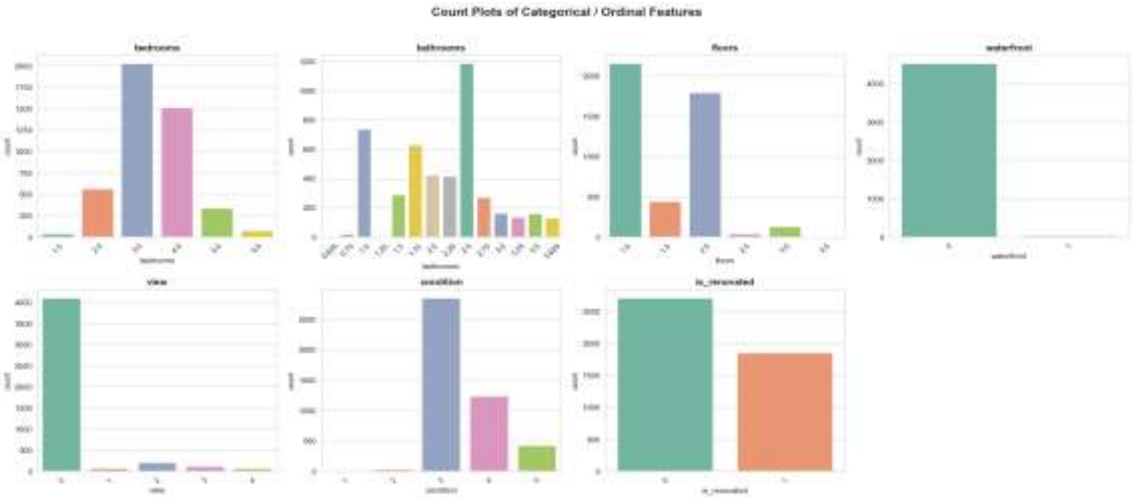
Box plot before outlier treatment:

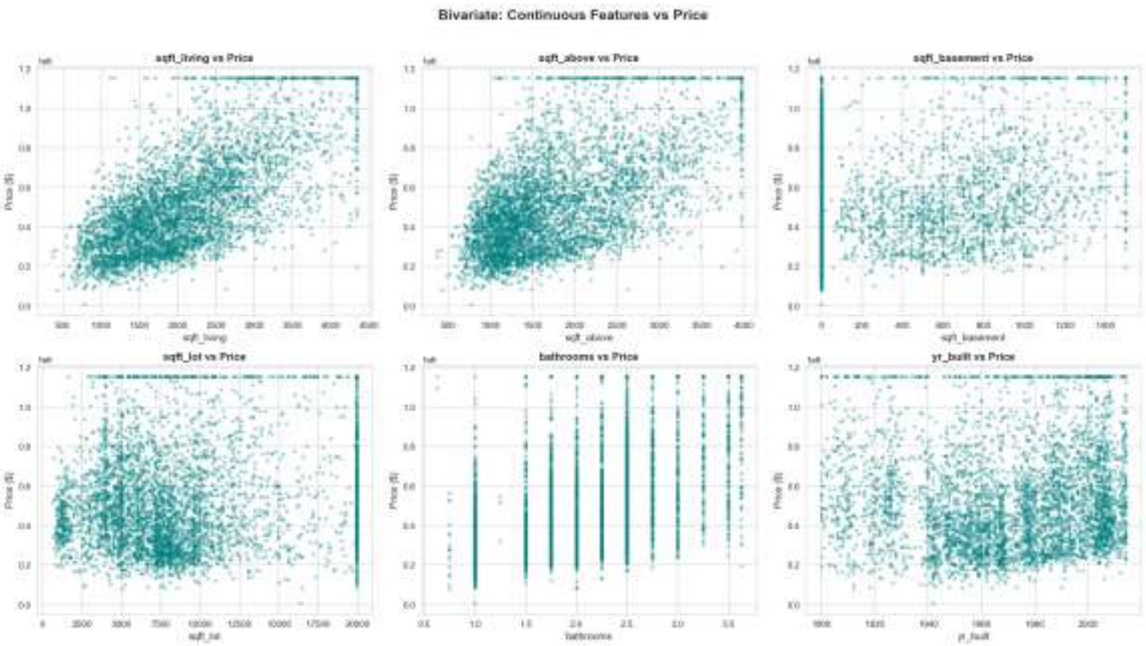


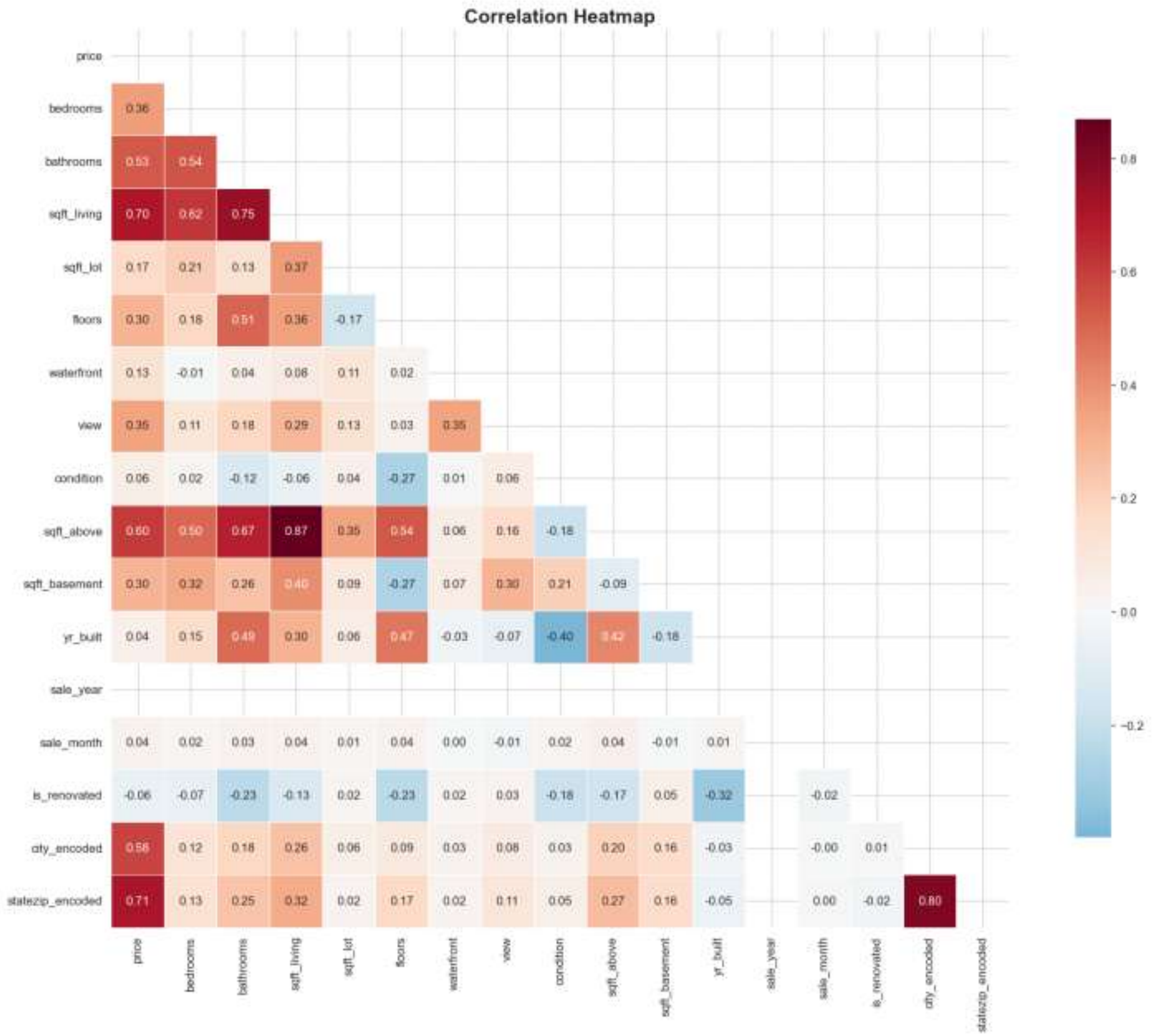
Box plot after outlier treatment:

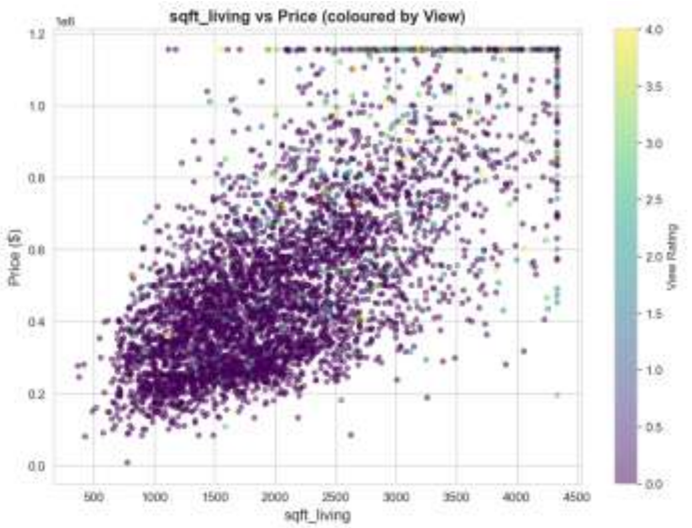



Univariate Distributions of Numeric Features
bedrooms
bathrooms
sqft_living
sqft_lot
floors
sqft_above
sqft_basement
yr_built


Count Plots of Categorical / Ordinal Features
bedrooms
bathrooms
floors
waterfront
view
condition
is_renovated


Bivariate: Continuous Features vs Price
sqft_living vs Price
sqft_above vs Price
sqft_basement vs Price
sqft_lot vs Price
bathrooms vs Price
yr_built vs Price


Correlation Heatmap
price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
sqft_above
sqft_basement
yr_built
sale_year
sale_month
is_renovated
city_encoded
statezip_encoded

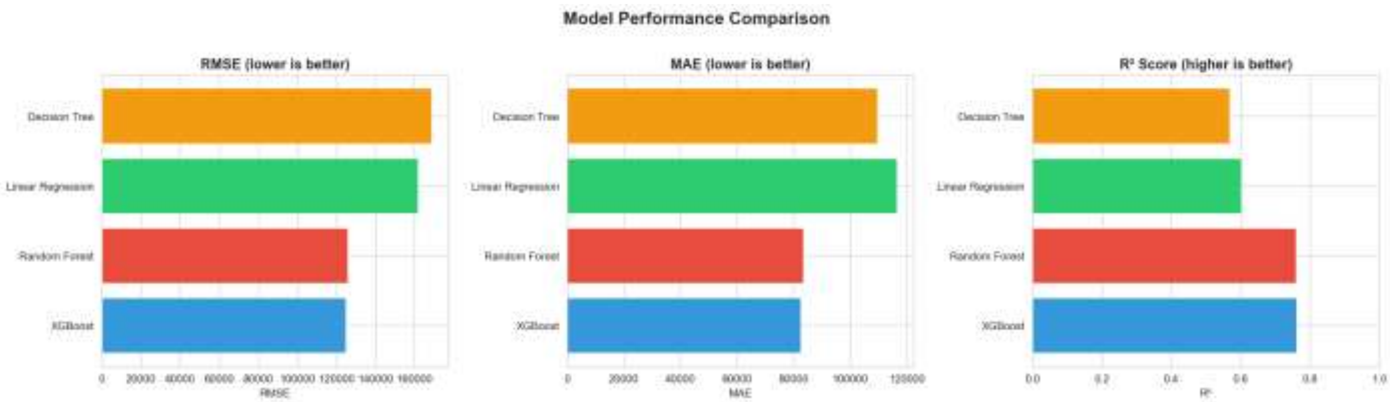

sqft_living vs Price (coloured by View)
Price ($)
sqft_living
View Rating

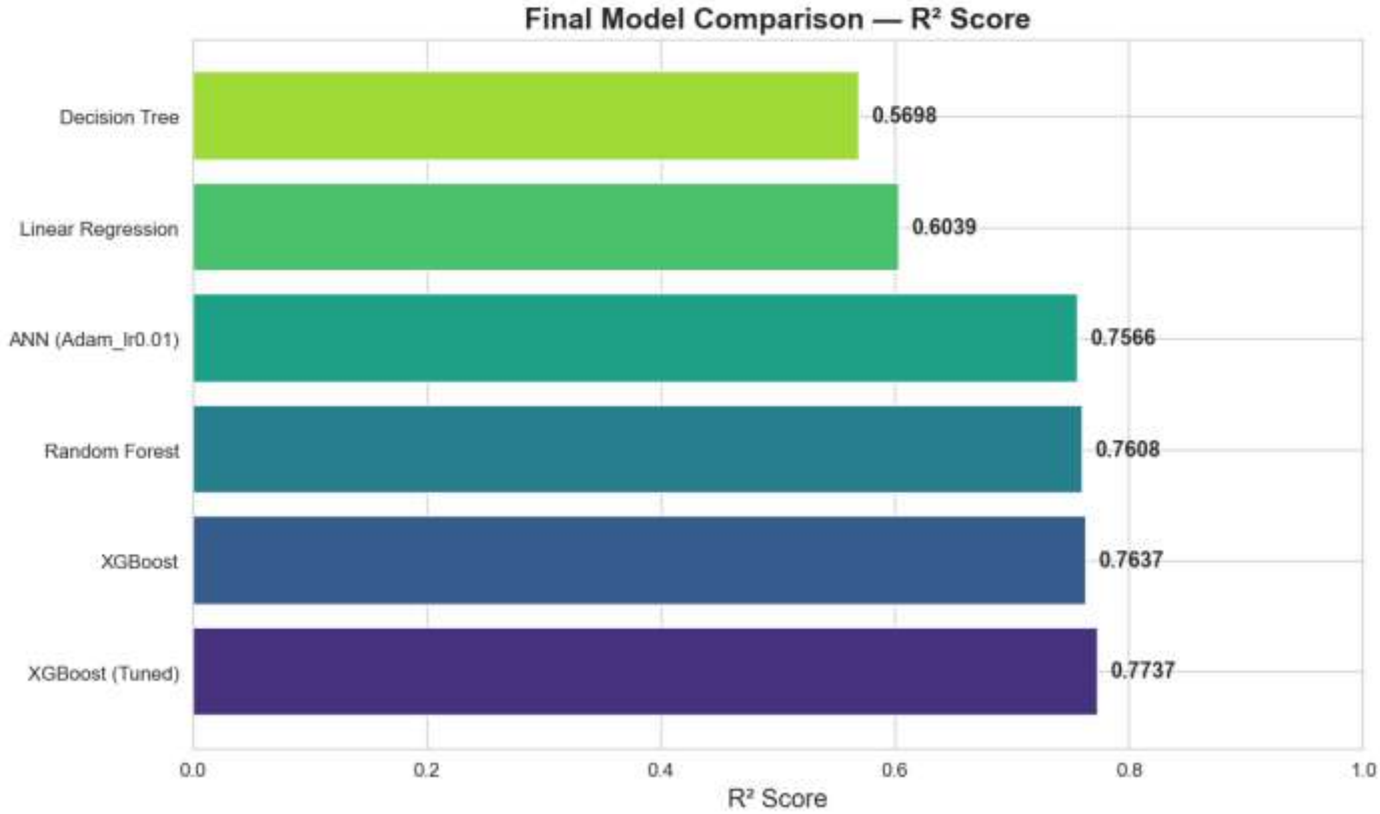

sqft_living vs Price (coloured by Condition)
Price ($)
sqft_living
Condition

4. Model Training & Comparison

We evaluated four Machine Learning models and one Deep Learning architecture.

Model	RMSE	MAE	R2
XGBoost (Tuned)	**122,433.12**	**79,825.02**	**0.7737**
Random Forest	125,867.71	83,570.80	0.7608
ANN (Adam_lr0.001)	126,629.00	84,865.10	0.7579
Linear Regression	161,993.61	116,642.97	0.6039
Decision Tree	168,811.50	109,698.90	0.5698





5. Neural Network Implementation

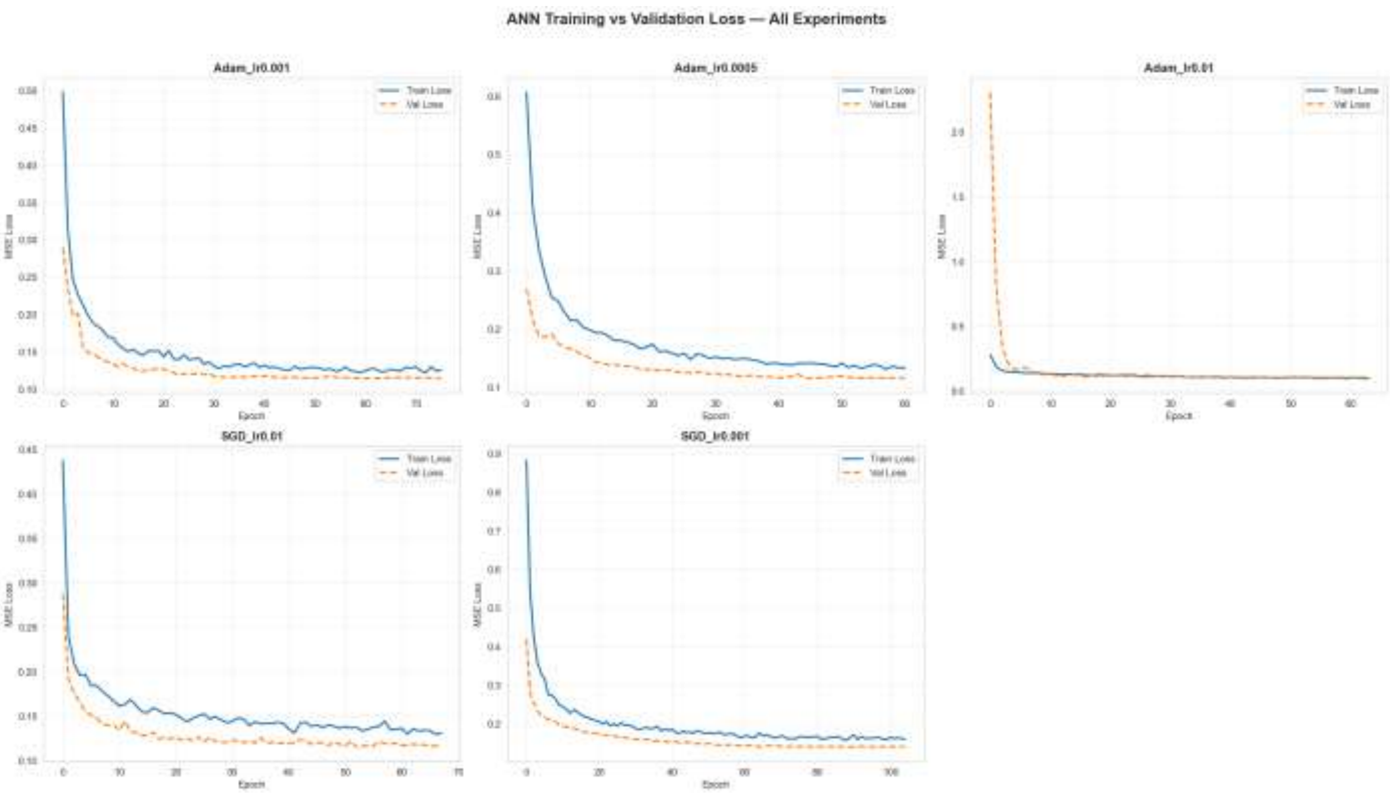
Architecture:

A deep **6-Hidden-Layer** ANN was built using:

- **Input Layer:** 10 neurons (based on RFE selection).
- **Layers:** 256 → 128 → 128 → 64 → 64 → 32 neurons.
- **Activation:** ReLU for hidden layers, Linear for output.

Regularization & Optimization:

- **Techniques:** Batch Normalization and Dropout (10–20%) were used to prevent overfitting.
- **Loss Function:** Huber Loss (delta=1.0) to handle remaining residuals robustly.
- **Optimizer Experiment:** Adam with LR=0.001 converged significantly faster and more accurately than SGD.



6. Business Interpretation & Conclusion

Business & Real-World Implications:

- **Investment Strategy:** Identify "undervalued" properties where listed price < predicted price for high-yield flips.
- **Risk Management:** Mortgage lenders can use the model as an independent valuation check to prevent over-leveraging.
- **Pricing Strategy:** Real-estate agents can justify listing prices to sellers using objective data-driven metrics.

Limitations:

- **Temporal Limits:** The data reflects a specific time period; it does not account for modern economic shifts like current interest rate hikes.
- **Regional Specificity:** The model is optimized for Washington; applying it to other states would require retraining.

Future Improvements:

- **External Data:** Incorporating school ratings, crime rates, and proximity to transit would likely bridge the error gap.
- **Time-Series:** Using LSTM layers to capture long-term price appreciation trends.