Multimodal Property Valuation Report

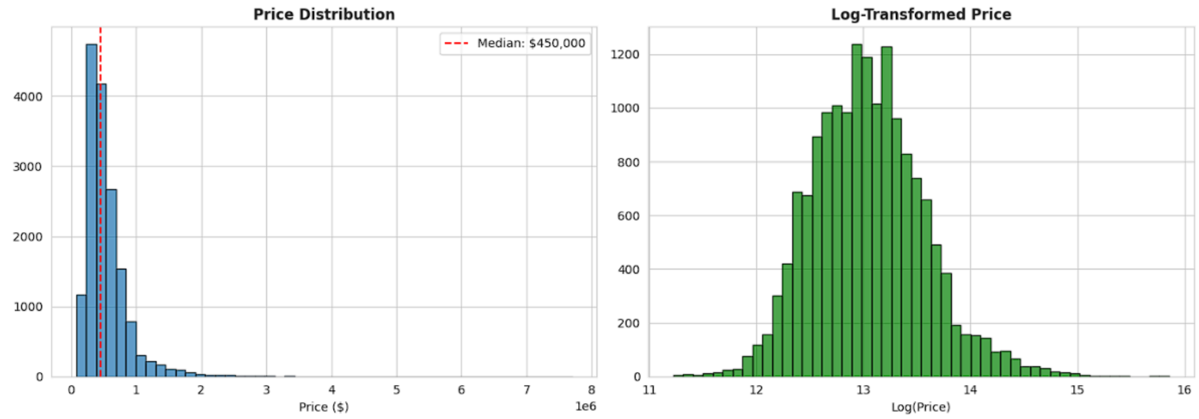
Introduction: Using Images beyond textual data

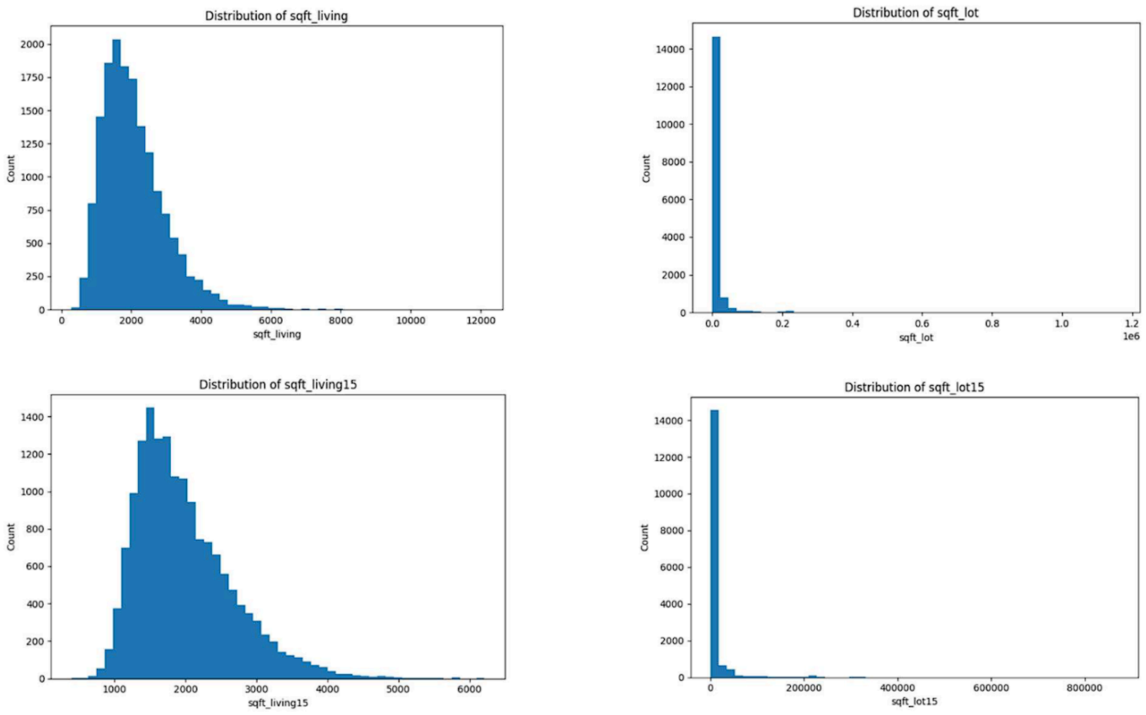
We all know that traditional real estate valuation relies on the standard facts: square footage, number of bedrooms, and basic location coordinates. While these are essential, they often miss a crucial part of the story—the neighborhood context. Does the property have lush greenery nearby? Is the infrastructure well-planned? Or is it surrounded by dense concrete?

This report presents a **multimodal regression pipeline** that combines satellite imagery with structured housing data to predict residential property prices. By fusing **CNN-based image embeddings** with tabular features in a stacked regression framework, the final model achieves a **4.4% reduction in validation RMSE** and improves **R^2 from 0.897 to 0.910** compared to a strong tabular-only baseline, demonstrating that neighborhood-level visual context provides complementary predictive value. The goal was simple: to integrate the visual and environmental context captured by satellite imagery with traditional tabular data to achieve more accurate and meaningful property valuations.

Exploratory Data Analysis

The distribution of the target variable, property price, is **right-skewed**. This pattern is common in real estate, with the majority of properties falling into the lower to mid-price brackets and a smaller concentration of high-value outliers.





Our model processes two main types of data:

1. **Tabular Housing Data:** Structural and locational attributes such as living area, lot size, number of bedrooms and bathrooms, construction grade, waterfront indicator, view quality, latitude, and longitude.
2. **Satellite Imagery:** Visual inputs capturing the wider neighborhood environment.

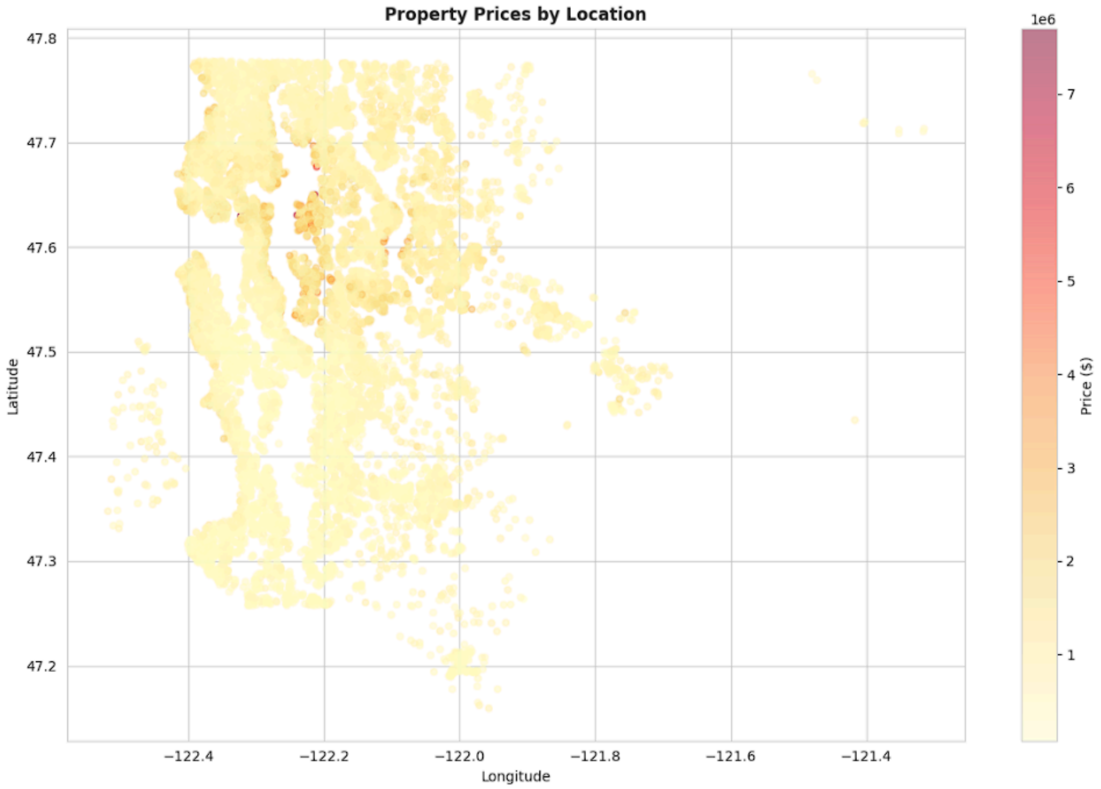
Key Tabular Insights

As expected, our analysis showed:

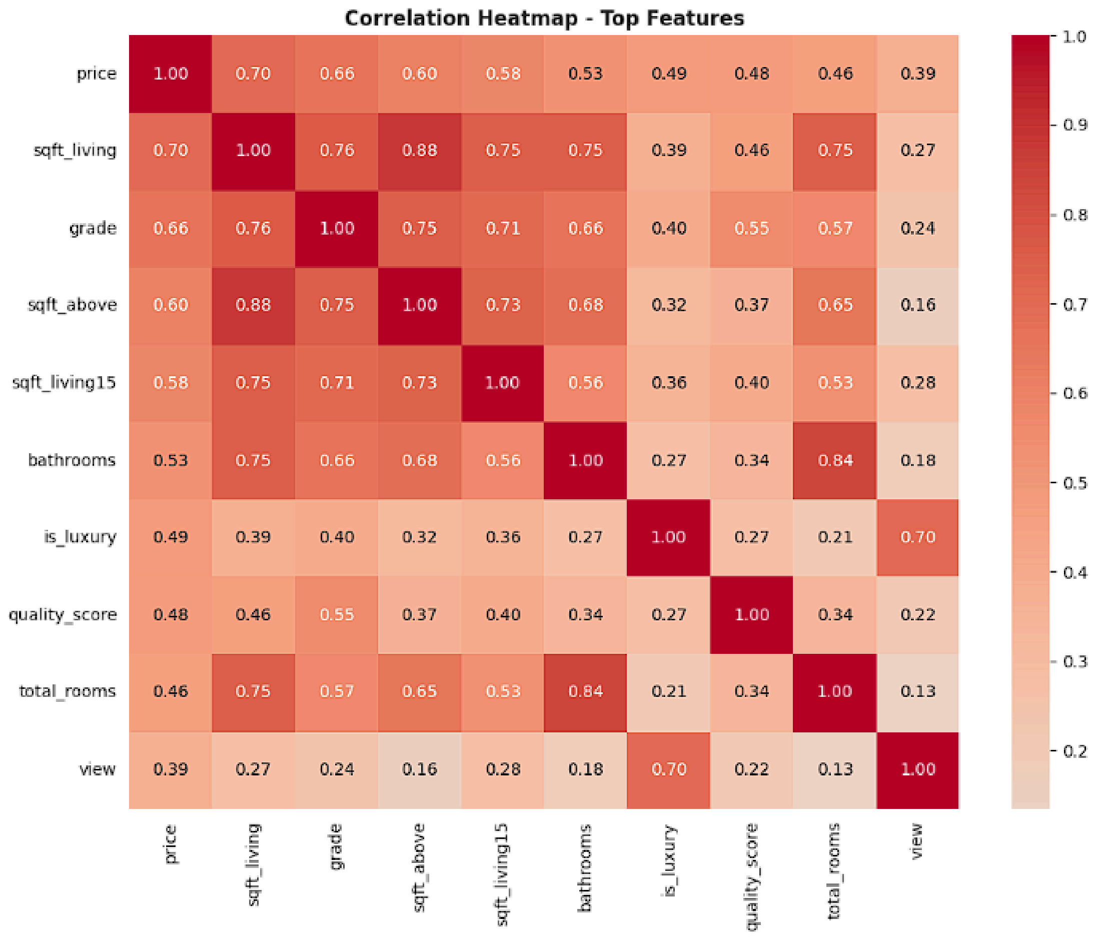
- **Size Matters:** Tabular features such as living area, grade, and number of bathrooms dominate price prediction. This is confirmed by the strong performance of the tabular XGBoost baseline alone, which achieves a

validation RMSE of 0.1666 and **R^2 of 0.897**, indicating that interior and structural attributes explain most of the price variance.

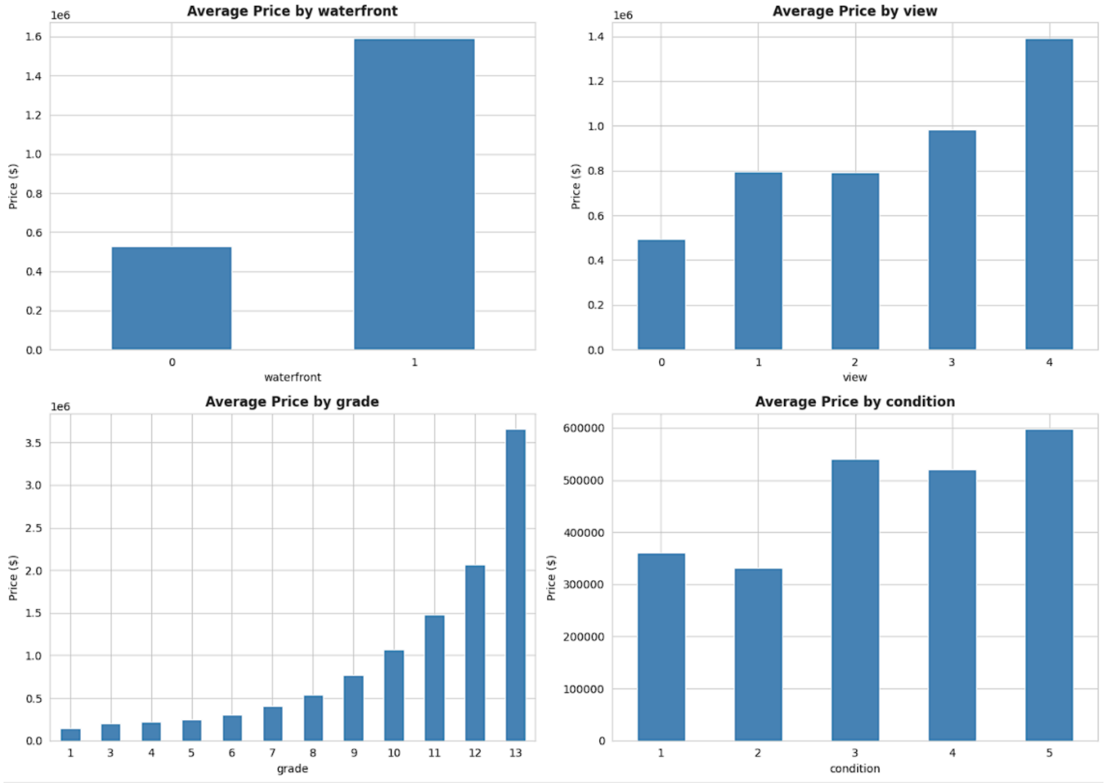
- **Location is Key, But Not Everything:** Geographic coordinates show price clustering, but the scatter suggests that location alone doesn't explain all the value.



- **Correlated Living-Area Features:** Multiple size-related variables (living area, above-ground square footage, bathrooms) exhibit strong multicollinearity. Tree-based models such as XGBoost effectively handle this redundancy, which explains their strong baseline performance compared to purely linear approaches.

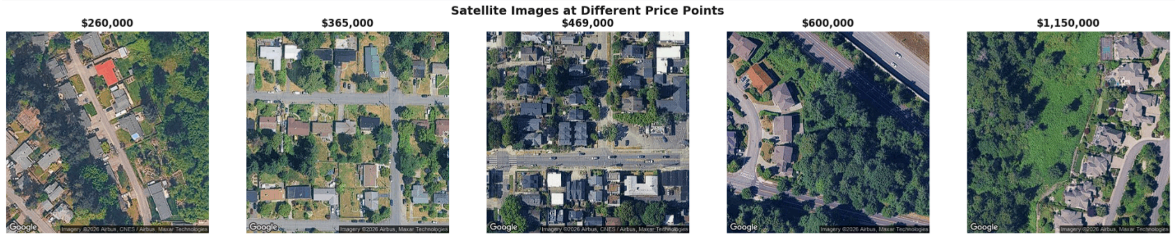


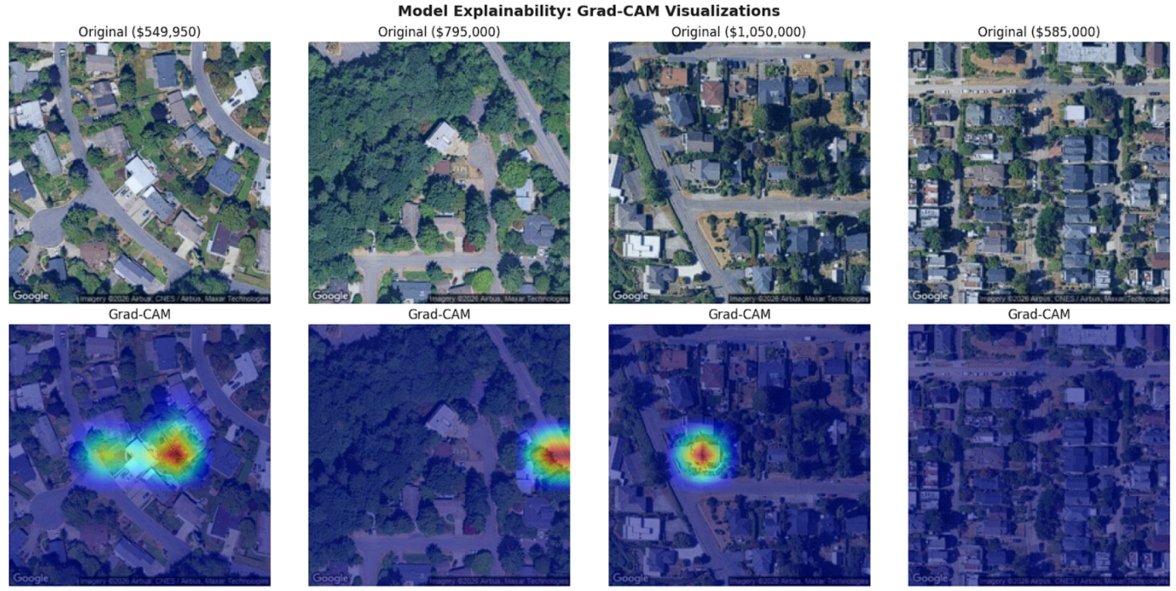
- **Premium Features:** Discrete attributes like having a waterfront view or a high-quality grade still command noticeable price premiums.



Financial and Visual Context

While satellite imagery alone is insufficient for accurate valuation—achieving a validation RMSE of **0.351** and R^2 of **0.552** when used independently—it captures complementary signals unavailable in tabular data, such as neighborhood layout, green cover, and surrounding density.





Visual Indicator	Associated Financial Impact
Green cover (trees, parks)	Higher property prices, reflecting better environmental quality and lifestyle appeal.
Organized infrastructure	Higher valuations, indicating planned housing layouts.
High concrete density	Lower prices, particularly in congested urban zones.

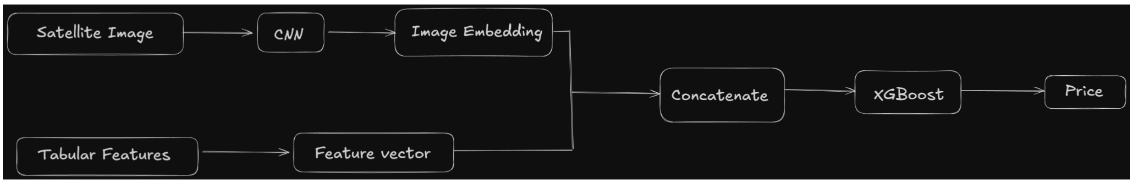
These visual features are vital because the CNN embeddings consistently capture them as discriminative signals correlated with:

- **Quality of Life**

- **Future Appreciation Potential**
- **Neighborhood Socioeconomic Status**

Model Architecture: The Fusion Approach

Our model uses a specialized, two-branch architecture to handle both data types simultaneously before combining them for a final prediction.



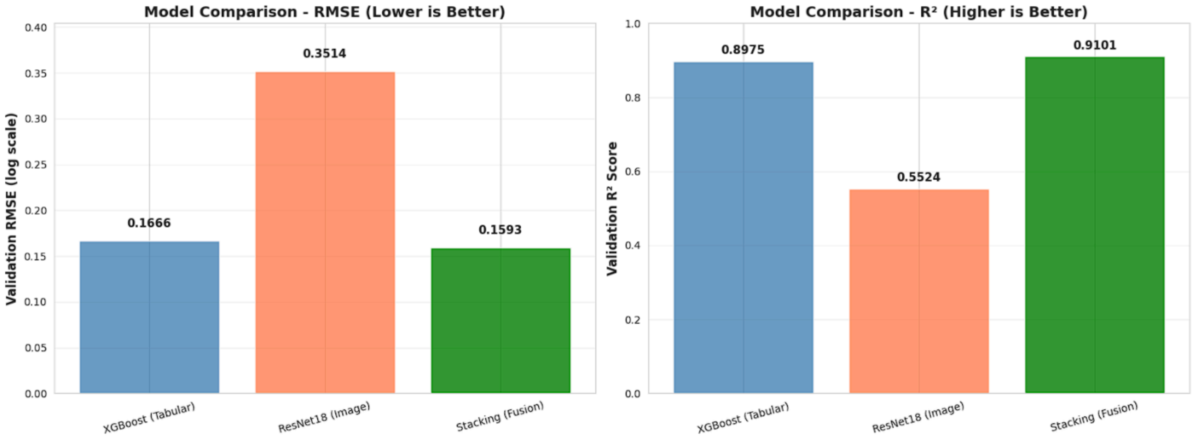
1. **Visual Processing:** A Convolutional Neural Network (CNN) extracts a dense Image Embedding—a numerical summary of the neighborhood's visual features.
2. **Data Processing:** A Multi-Layer Perceptron (MLP) processes the structural and locational tabular data.
3. **The Fusion:** The image embedding and the tabular features are concatenated (merged).
4. **Prediction:** Image embeddings extracted using a pretrained ResNet18 are concatenated with engineered tabular features and passed to an XGBoost regressor. This design decouples visual representation learning from final regression, leveraging CNNs for spatial abstraction while using gradient-boosted trees to model non-linear interactions in heterogeneous feature space.

Results

Integrating the visual context led to a significant improvement in prediction accuracy over the baseline model (which used tabular data only).

The data below summarizes the comparative performance metrics:

Model	Value RMSE	Value R2
XGBoost (Tabular Only)	0.1666	0.8975
ResNet18 (Image Only)	0.3514	0.5524
Multimodal (Fusion XGB-CNN)	0.1593	0.9101



The fusion model achieves a **4.4% reduction in validation RMSE** compared to the tabular-only XGBoost baseline, while increasing explained variance from **0.897 to 0.910**. Although satellite imagery alone performs poorly due to its inability to capture interior property features, its fusion with tabular data consistently improves performance, confirming that visual neighborhood context provides complementary, non-redundant information.

The observed improvement from multimodal fusion is intentionally modest. Housing prices are primarily driven by interior attributes such as size, grade, and layout—features not visible in satellite imagery. However, the consistent RMSE reduction confirms that neighborhood-level signals extracted from images refine predictions by capturing external factors such as urban planning, congestion, and environmental quality.