# Multimodal Real Estate Price Prediction

## 1. Overview

Real estate valuation is typically modeled using structured tabular data such as property size, location, and construction quality. While effective, these features do not explicitly encode neighborhood-level environmental context.

This project explores a **multimodal regression approach** that integrates structured housing data with **Sentinel-2 satellite imagery** to capture high-level spatial cues such as urban density, green cover, and proximity to water bodies. The objective is twofold:

1. evaluate whether satellite imagery improves prediction beyond a strong tabular baseline, and
2. enhance interpretability using **Grad-CAM**.

The project emphasizes **transparent evaluation** rather than forcing artificial accuracy gains.

---

## 2. Dataset

### 2.1 Tabular Data

The dataset is derived from the **King County Housing Dataset** and includes structural, locational, and neighborhood-level attributes.

Key features include:

- Property characteristics: bedrooms, bathrooms, living area, lot size, floors, grade, condition
- Location indicators: latitude, longitude, waterfront, view
- Neighborhood statistics: average nearby living and lot sizes
- Engineered features: house age, renovation indicator, size ratios, and relative neighborhood comparisons

The target variable is **sale price**, log-transformed during training for stability.

---

### 2.2 Visual Data

Satellite imagery is fetched using property latitude and longitude.

- Source: Sentinel-2

- Resolution: ~10 meters per pixel
- Bands: RGB (natural color)

Due to resolution constraints, images encode **neighborhood context** (water, vegetation, density) rather than individual property details.

### 2.3 Data Preparation

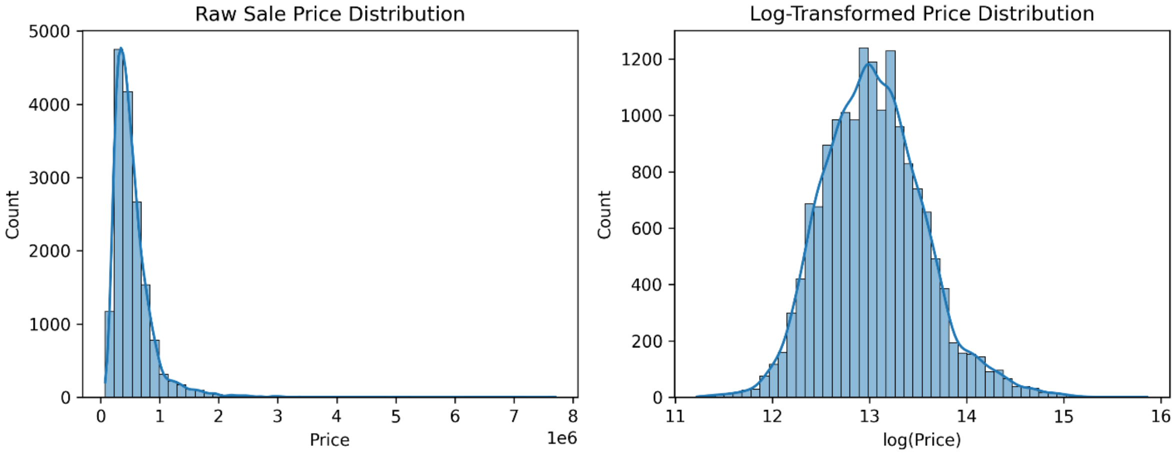
- Missing values handled during preprocessing
- Numerical features normalized
- Images resized and normalized using ImageNet statistics

Each training sample pairs tabular features with a corresponding satellite image.

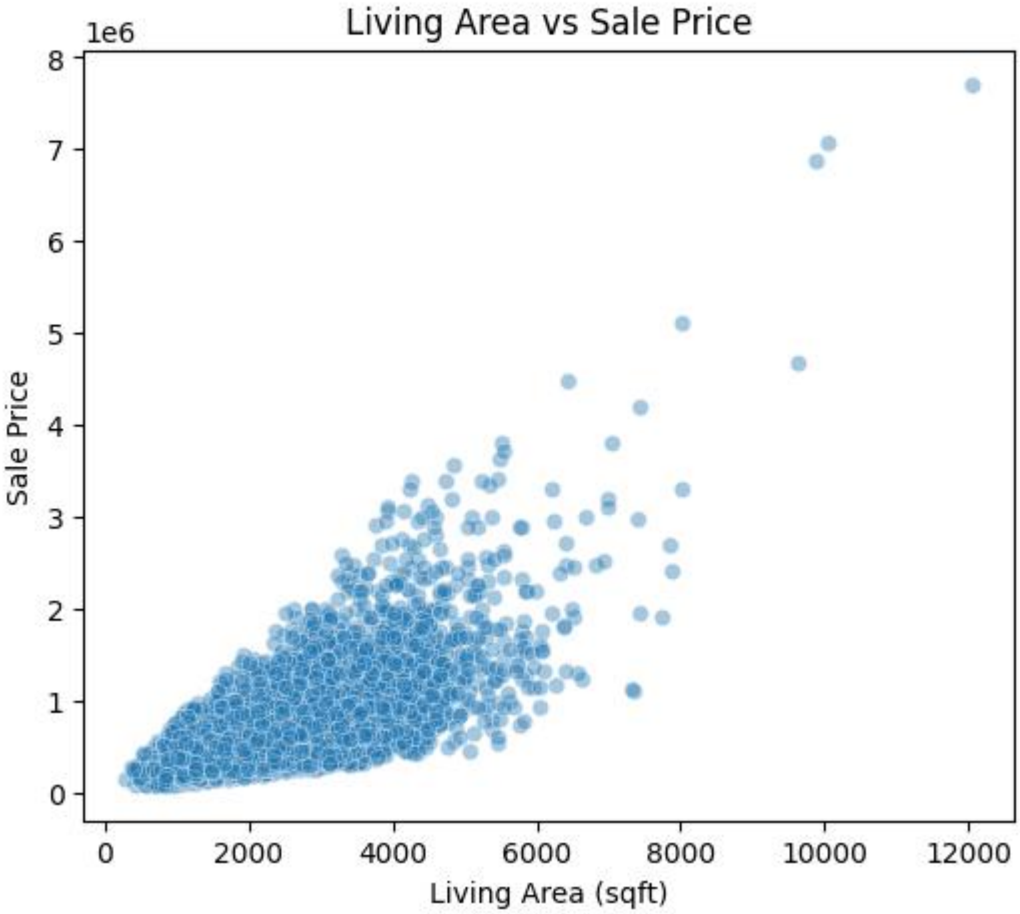
## 3. Exploratory Data Analysis

EDA shows that:

- Sale prices are right-skewed and benefit from log transformation
- Figure 1: Distribution of property sale prices before and after log transformation. Log scaling reduces skewness and improves regression stability:----



- Living area, construction grade, waterfront presence, and neighborhood statistics strongly correlate with price
- Figure 2: Relationship between living area and sale price, showing a strong positive correlation:----



- High-value properties cluster near waterfronts and dense urban area
- Satellite images visibly distinguish urban, suburban, and green regions, supporting their use for contextual analysis rather than precise valuation.
- Figure 3: Example Sentinel-2 satellite tile illustrating neighborhood-level spatial context:----

## 4. Methodology

The task is formulated as a **supervised regression problem** predicting log-price.

Three models are evaluated:

- **Tabular-only model:** MLP trained on engineered features (baseline)
- **Image-only model:** ResNet-18 with regression head
- **Multimodal model:** Late fusion of tabular and visual embeddings

Training uses Smooth L1 loss, AdamW optimization, gradient clipping, and a fixed train-validation split.

## 5. Results

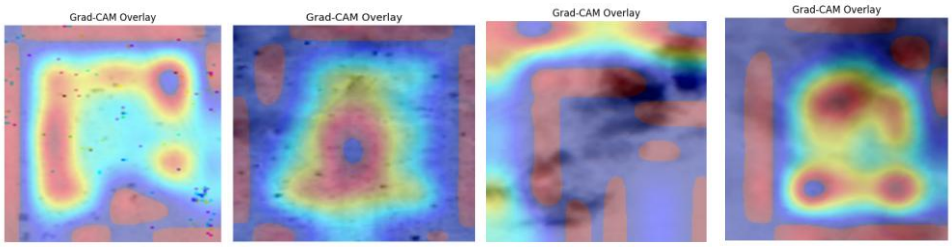
| Model | Validation RMSE |
|---|---|
| Tabular-only | ~0.31 |
| Multimodal | ~0.45 |
| Image-only | ~1.32 |

**Key insight:** Tabular features dominate predictive accuracy. Satellite imagery alone is insufficient for pricing and slightly degrades RMSE when fused, due to added variance.

## 6. Explainability (Grad-CAM)

Grad-CAM is applied to the multimodal model's convolutional backbone to visualize image regions influencing predictions.

Figure 4: Grad-CAM visualizations highlighting neighborhood-scale regions influencing price predictions. High attention is observed near water bodies and dense urban areas, consistent with contextual valuation signals.



### Observed patterns:

- Strong attention around water bodies
- Emphasis on dense urban layouts
- Low attention in homogeneous regions

Attention is distributed across **neighborhood-scale regions**, consistent with Sentinel-2 resolution. Grad-CAM confirms that imagery captures **contextual environmental signals**, not parcel-level details.

---

## 7. Conclusion and Limitations

### Conclusion

The tabular-only model provides the best predictive performance, while satellite imagery primarily enhances **interpretability**. Grad-CAM analysis demonstrates that visual features encode meaningful neighborhood context such as water proximity and urban density.

This project highlights that multimodal learning can add value through explanation, even when it does not improve numerical accuracy.

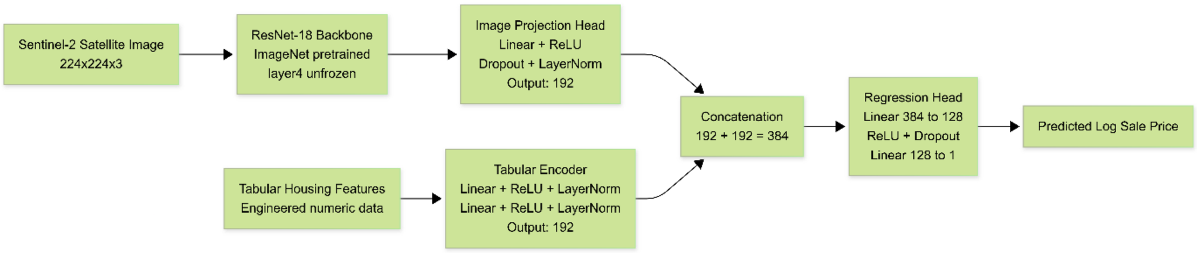
### Limitations

- Coarse satellite resolution limits property-level detail
- Visual features introduce noise when tabular signals are strong
- Dataset is geographically limited
- Multimodal models increase complexity without proportional gains

### Future Work

- Higher-resolution or street-level imagery
- Attention-based fusion architectures
- Temporal satellite data
- Broader geographic coverage

## Architecture:-


Sentinel-2 Satellite Image
224x224x3
ResNet-18 Backbone
ImageNet pretrained
layer4 unfrozen
Image Projection Head
Linear + ReLU
Dropout + LayerNorm
Output: 192
Tabular Housing Features
Engineered numeric data
Tabular Encoder
Linear + ReLU + LayerNorm
Linear + ReLU + LayerNorm
Output: 192
Concatenation
192 + 192 = 384
Regression Head
Linear 384 to 128
ReLU + Dropout
Linear 128 to 1
Predicted Log Sale Price