Multimodal Property Valuation using Satellite Imagery

1. Approach to the Problem Statement

The core objective of this project was to determine if unstructured visual data (satellite imagery) could provide additive predictive signal to a robust tabular property valuation model.

Our approach followed a **"Three-Stage Evolution"** strategy to ensure scientific rigor:

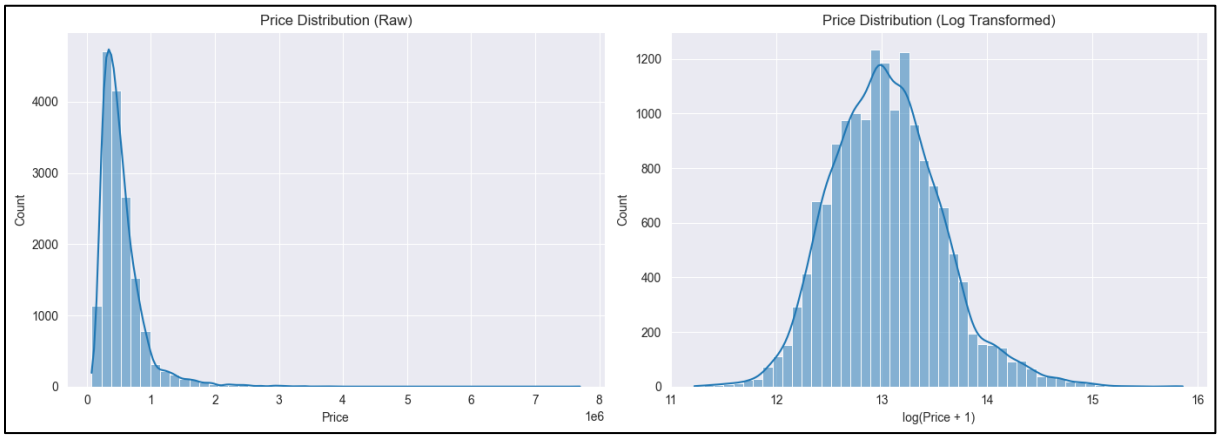
1. **Baseline Establishment (Tabular Only):** We first trained a Gradient Boosting (XGBoost) model to establish a "performance floor," quantifying how well structural features (e.g., sqft_living, bedrooms) explain price variance.
2. **Multimodal Integration (Late Fusion):** We developed a deep learning architecture that fuses tabular features with visual embeddings from a ResNet-50 backbone. This tested the hypothesis that the model could learn joint interactions (e.g., "a small house on a waterfront lot is worth more than a small house on a highway").
3. **Explainability-First Modeling (Residual Learning & ViT):** In addition to the primary Late Fusion model, we explored two diagnostic architectures - a Residual CNN and a Vision Transformer (ViT). These models were not intended to outperform the fusion model, but to isolate and analyze the marginal visual signal contributed by satellite imagery and to probe whether global (ViT) or local (CNN) representations better captured environmental context.

(Late Fusion is the main model I trained for this problem statement. The Residual learning model and ViT model are just two other approaches that I explored while doing this PS. Later it was selected as final model due to its superior performance and stable interpretability.)

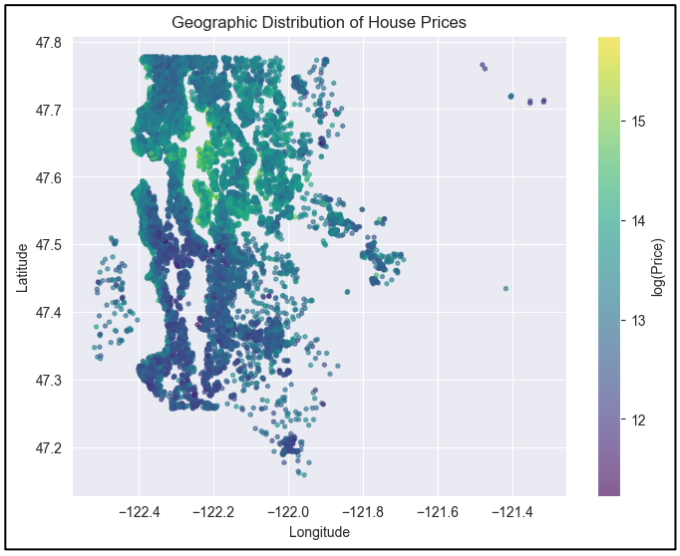
2. Findings of Exploratory Data Analysis (EDA)

Before modeling, extensive EDA was conducted (check the preprocessing.ipynb for all the findings) and these were the major observations:

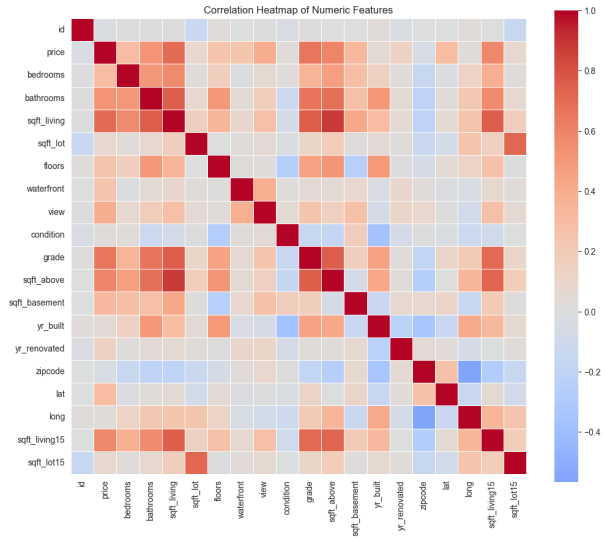
- **Target Variable Skew:** The property price distribution was highly right-skewed, confirming the need for a Log-Transformation (np.log1p) to stabilize training gradients and prevent the model from over-prioritizing the outliers.



- **Spatial Clustering:** Plotting properties by Latitude/Longitude revealed distinct high-value geographical pockets. We capitalized on this by generating a “geo_cluster” feature using K-Means, which served as a powerful proxy for neighbourhood level spatial segmentation in the tabular model.



- **Feature Correlations:** Some Features (for eg - sqft_living and sqft_above) showed strong correlation with each other and with price. However, significant variance remained unexplained, suggesting that some price variance may be associated with environmental context not explicitly encoded in tabular features.



3. Image Acquisition & Preprocessing Pipeline

A. Fetching Imagery Data

A pipeline was designed to handle batch requests efficiently while maintaining spatial consistency across all properties:

- **Coordinate Extraction:** The Latitude and Longitude for each property were extracted from the tabular CSV.
- **API Parameterization:** We used a standardized **Zoom Level of 18** (as it gives a balance between parcel-level detail and neighborhood context, ensuring that both property boundaries and surrounding environmental cues were visible.
- **Dimension Standardization:** Images were requested at a resolution of **400x400 pixels** from the **Mapbox Static Images API**.

B. Preprocessing & Normalization

Before being fed into the Multimodal Model, the raw Mapbox imagery underwent a secondary transformation pipeline:

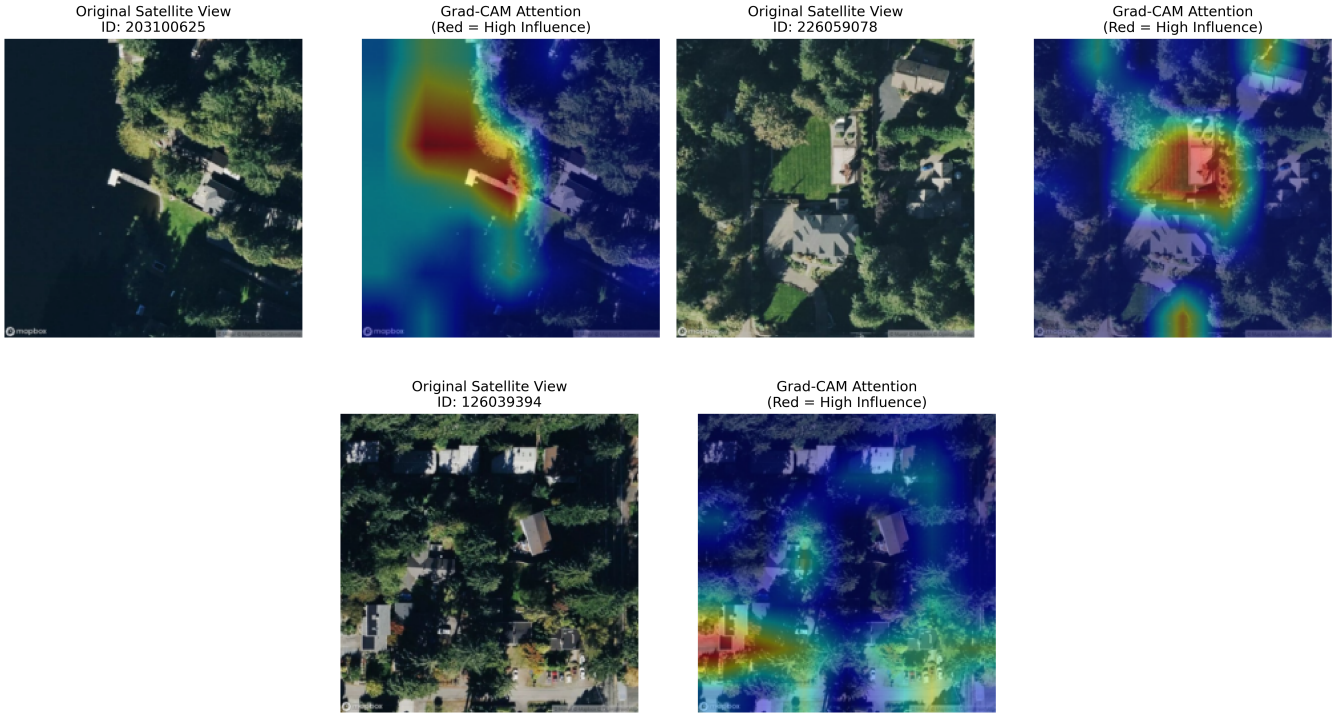
- **Resizing:** All images were downsampled to **224x224 pixels** to match the input requirements.
- **Tensor Conversion:** Images were converted from PIL format to PyTorch Tensors and normalized using the **ImageNet Mean and Standard Deviation** (μ=[0.485, 0.456, 0.406], σ=[0.229, 0.224, 0.225]).

4. Model Descriptions & Visual Analysis

A. Late Fusion Multimodal Model (Best Performer)

This architecture processed the two modalities in parallel branches. A Multi-Layer Perceptron (MLP) encoded the 64-dimensional tabular vector, while a pre-trained **ResNet-50** extracted a 2048-dimensional embedding from the satellite image. These features were concatenated and passed through a fusion head to predict the log-price.

- **Why it worked:** It allowed the model to weigh visual and structural evidence simultaneously, achieving the highest overall accuracy.
- **Visual Analysis:** Grad-CAM heatmaps suggest that the ResNet-50 branch prioritize environmental context over primary structures, specifically targeting waterfronts and tree canopies. This allows the model to identify luxury or detractive visual markers, such as shoreline access or commercial density, that are absent from tabular data.

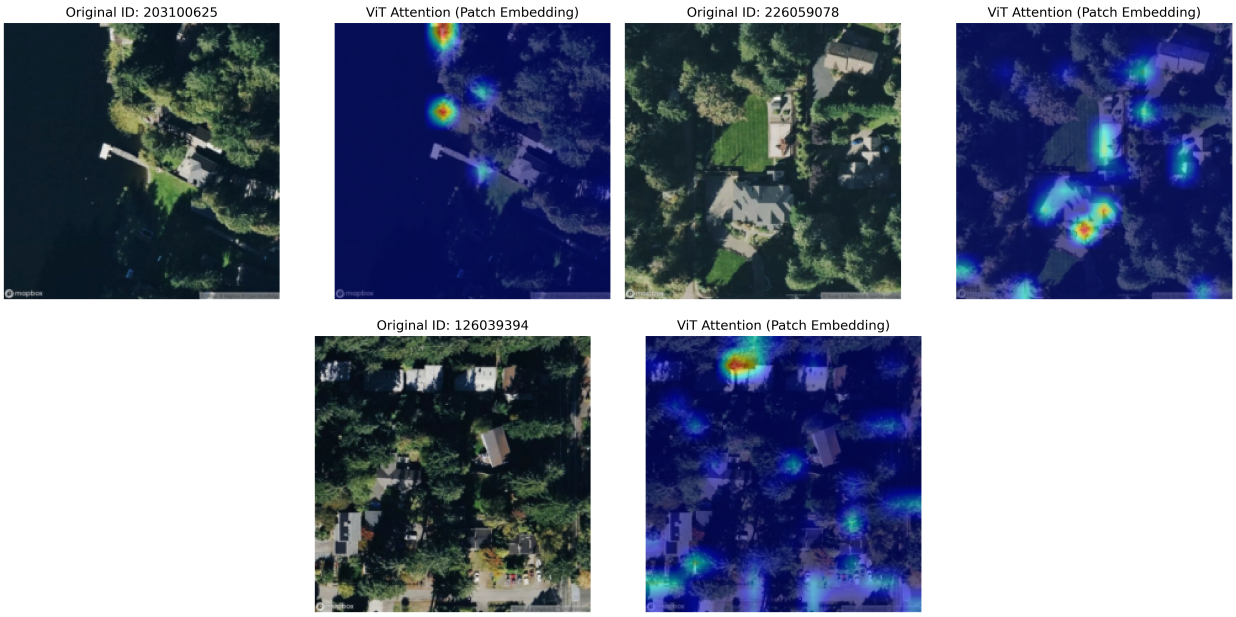


B. Vision Transformer (ViT-B/16)

We implemented a ViT to analyze the image as a sequence of 16x16 patches. This "Token-based" approach allowed us to visualize global attention without the local bias of Convolutions.

- **Visual Analysis:** The ViT utilizes patch-based tokenization to detect high-frequency features and geometric boundaries like parcel edges. By performing global texture

composition rather than local spatial analysis, it aggregates specific value-driving patches across the entire image to quantify environmental influence.



- While the ViT achieved competitive performance, its gains were marginal compared to the Late Fusion CNN, suggesting that local spatial cues captured by convolutions may be more relevant than global patch interactions for satellite-based property valuation.

C. Residual CNN

This model was trained to predict the **residuals** of the XGBoost baseline. By removing the structural signal, the CNN was forced to focus purely on visual context that "corrected" the tabular prediction.

Although the Residual CNN improved over the tabular baseline, its constrained additive formulation limited its ability to model interactions between structural and visual features, resulting in slightly lower performance than full multimodal fusion.

4. Comparative Results

We evaluated all models on the final RMSE and R2 score during training on the training and validation dataset. The Multimodal approach demonstrated a consistent improvement over the tabular baseline.

Architecture	RMSE (log)	R2 score
XGBoost (Baseline)	0.1847	0.8743
Late Fusion (ResNet-50)	0.1790	0.8819
Residual CNN	0.1823	0.8775
Vision Transformer (ViT)	0.1798	0.8808

Analysis: The **Late Fusion model** provided a lift in over the baseline. Although the numerical gains appear modest, such improvements are meaningful in real estate valuation, where pricing variance is dominated by location-specific factors and noise.

Comparing all the models: **Late Fusion > Vision Transformer > Residual CNN > XGBoost**

5. Summary of Final Predictions

The final predictions were made using the Late Fusion model on the unseen test dataset.

- **Transformation:** Predictions were converted from Log-Scale back to USD (exp(pred)).
- **Distribution:**
    - **Mean Price:** $475,696
    - **Min / Max:** $113k / $4.84M

6. Conclusion

This project successfully demonstrated that **satellite imagery contains quantifiable predictive signal** for property valuation. By integrating ResNet-50 embeddings with tabular data, we reduced the prediction error (RMSE) compared to a strong tabular baseline.

Crucially, our **Grad-CAM analysis** provided qualitative evidence that the learned visual representations correspond to meaningful environmental features. The models learned to identify semantic environmental features—specifically **waterfront proximity, vegetation density, and neighborhood layout**—acting as an automated appraisal of "curb appeal" and location quality. This confirms that multimodal AI can effectively bridge the gap between structured property data and the visual reality of the physical world.