Project Report: Multimodal Real Estate Price Prediction

Integrating Satellite Imagery with Tabular Data for Automated Valuation Models

1. Executive Summary

This project aims to enhance traditional Real Estate Automated Valuation Models (AVMs) by integrating visual data. While traditional models rely solely on tabular features (e.g., square footage, number of bedrooms), this research explores the hypothesis that visual context—such as neighbourhood density, greenery, and road connectivity—extracted from high-resolution satellite imagery can improve pricing accuracy or provide complementary insights.

We developed a Multimodal Neural Network utilizing a Late-Fusion architecture. The model was benchmarked against a strong gradient-boosting baseline (XGBoost). While the XGBoost baseline achieved a higher R2 score (0.88) compared to the deep learning model (0.706), the project successfully demonstrates the viability of extracting meaningful "curb appeal" features from aerial data.

2. Objective

The primary objectives of this study were:

1. **Data Integration:** To construct a hybrid dataset combining King County tabular housing data with corresponding 400 X 400 RGB satellite imagery fetched via the Esri World Imagery API.
2. **Architecture Design:** To design and implement a multimodal deep learning pipeline capable of processing heterogeneous data types (structured text and unstructured images).
3. **Explainability:** To utilize techniques like Grad-CAM to interpret what visual features the model deems important for price prediction.

3. Dataset & Preprocessing

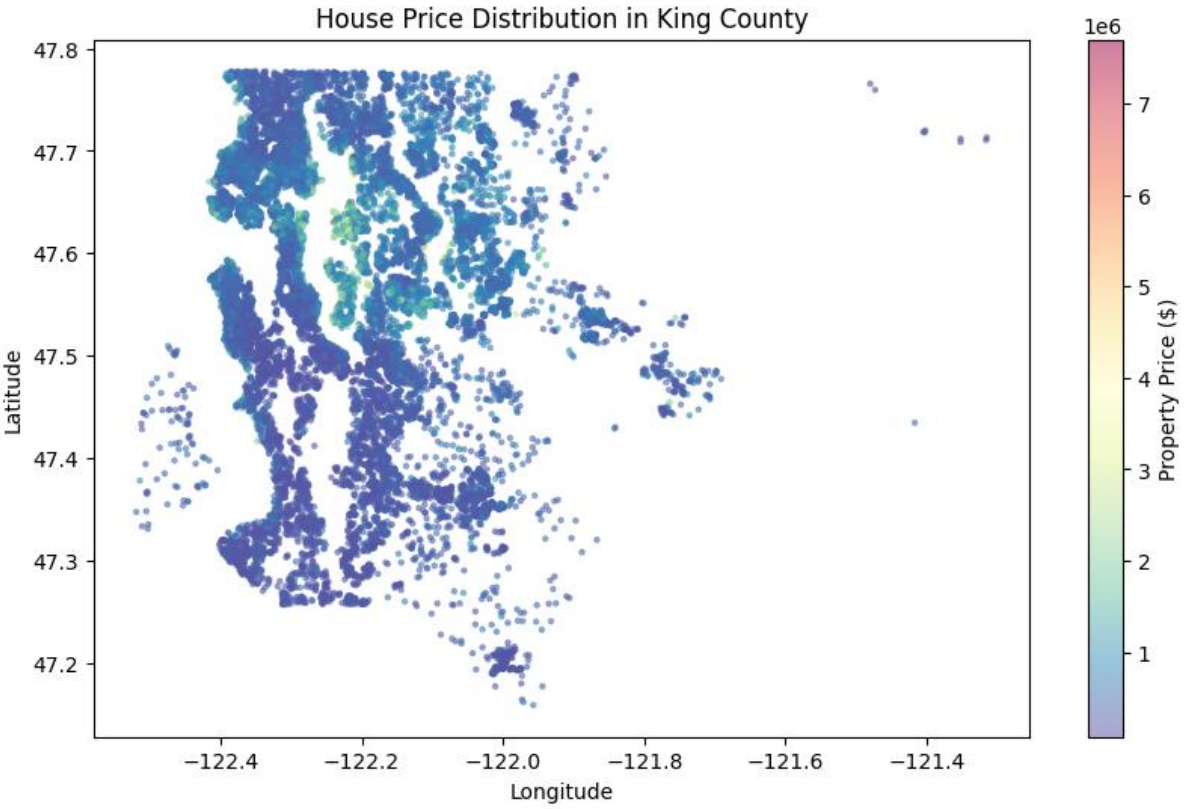
3.1 Tabular Data

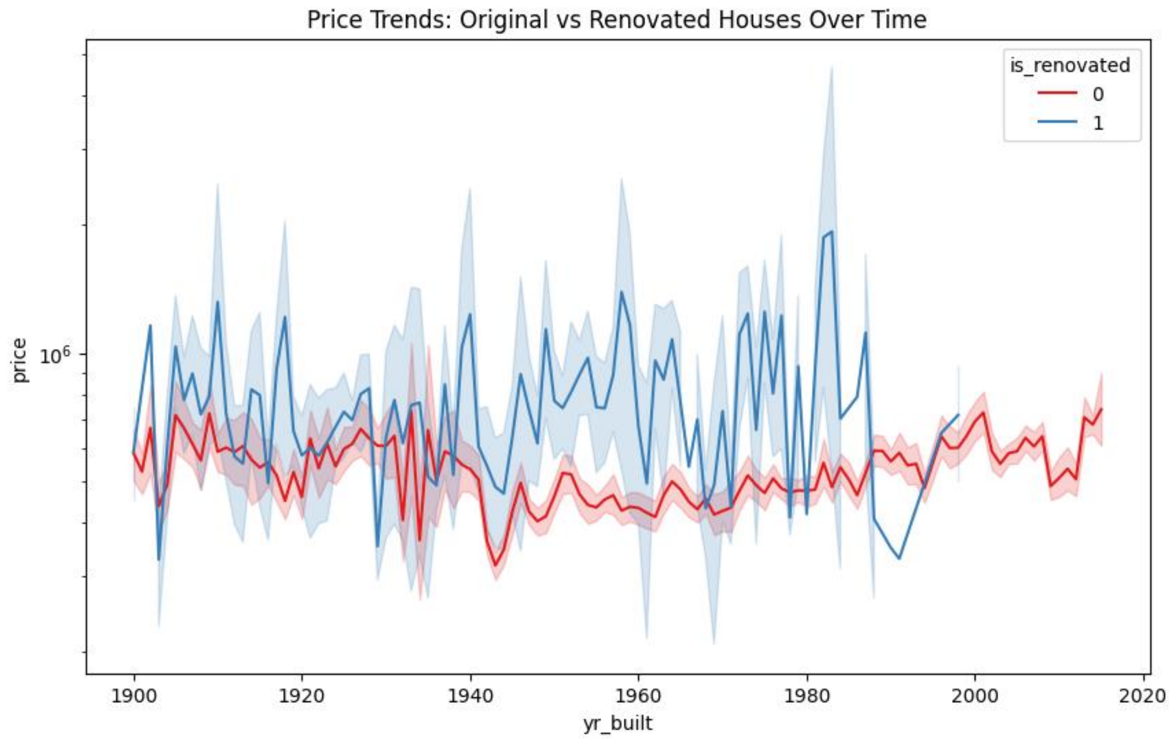
The tabular dataset consists of standard real estate features. Preprocessing ensured data stability and model convergence:

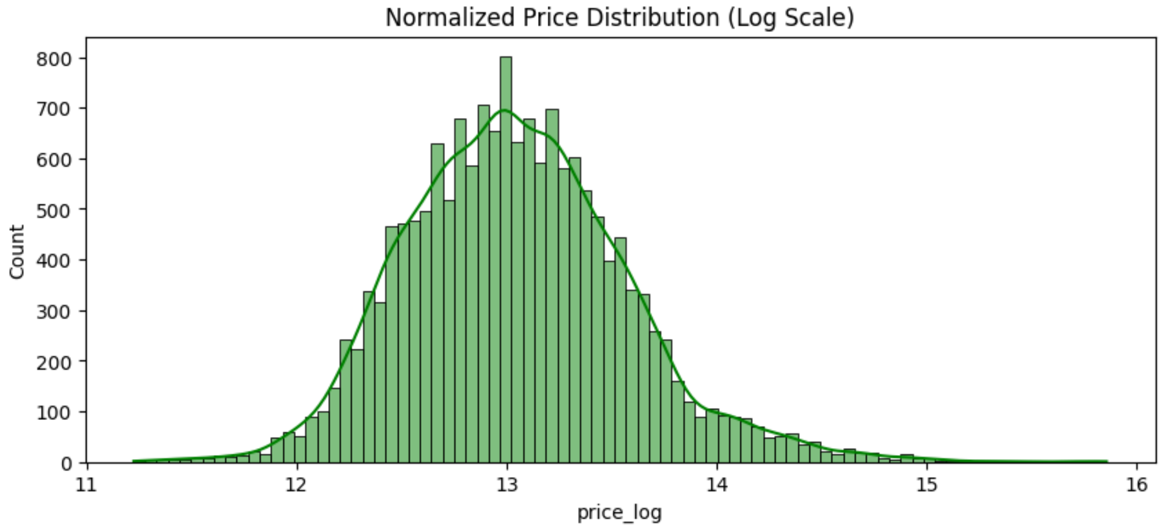
- **Feature Selection:** Key features included sqft_living, grade, lat, long, waterfront, and house_age.
- **Outlier Handling:** A RobustScaler was applied to handle outliers in square footage and lot size, ensuring that extreme luxury properties did not skew the feature distribution.

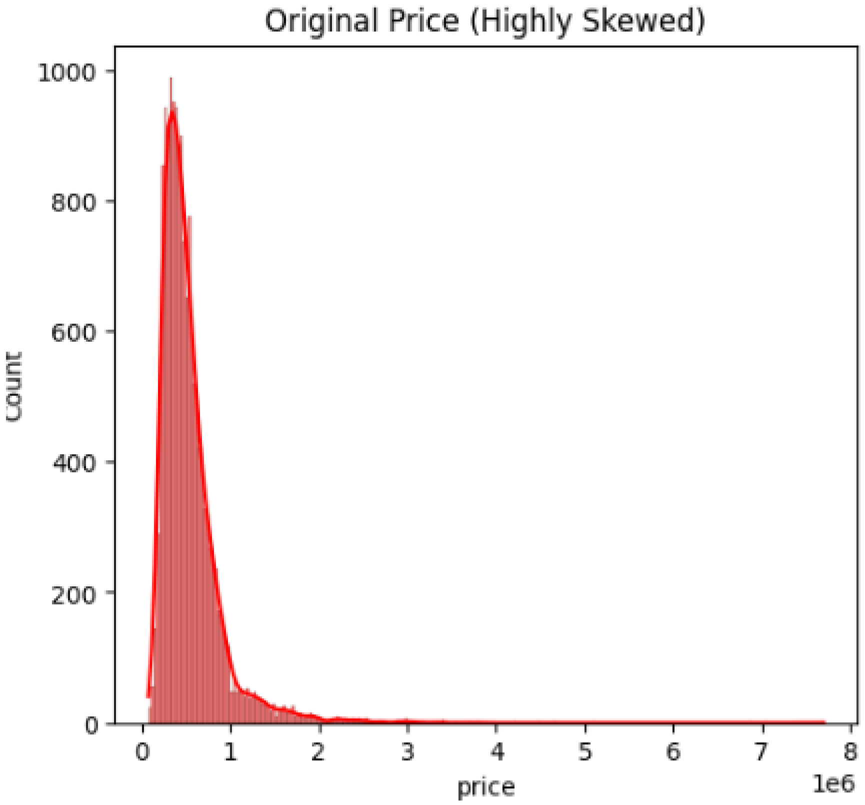
- **Target Variable:** The target variable (Price) exhibited a right-skewed distribution. We applied a Log-transformation (np.log1p) to normalize the target, stabilizing the variance for the regression loss function.

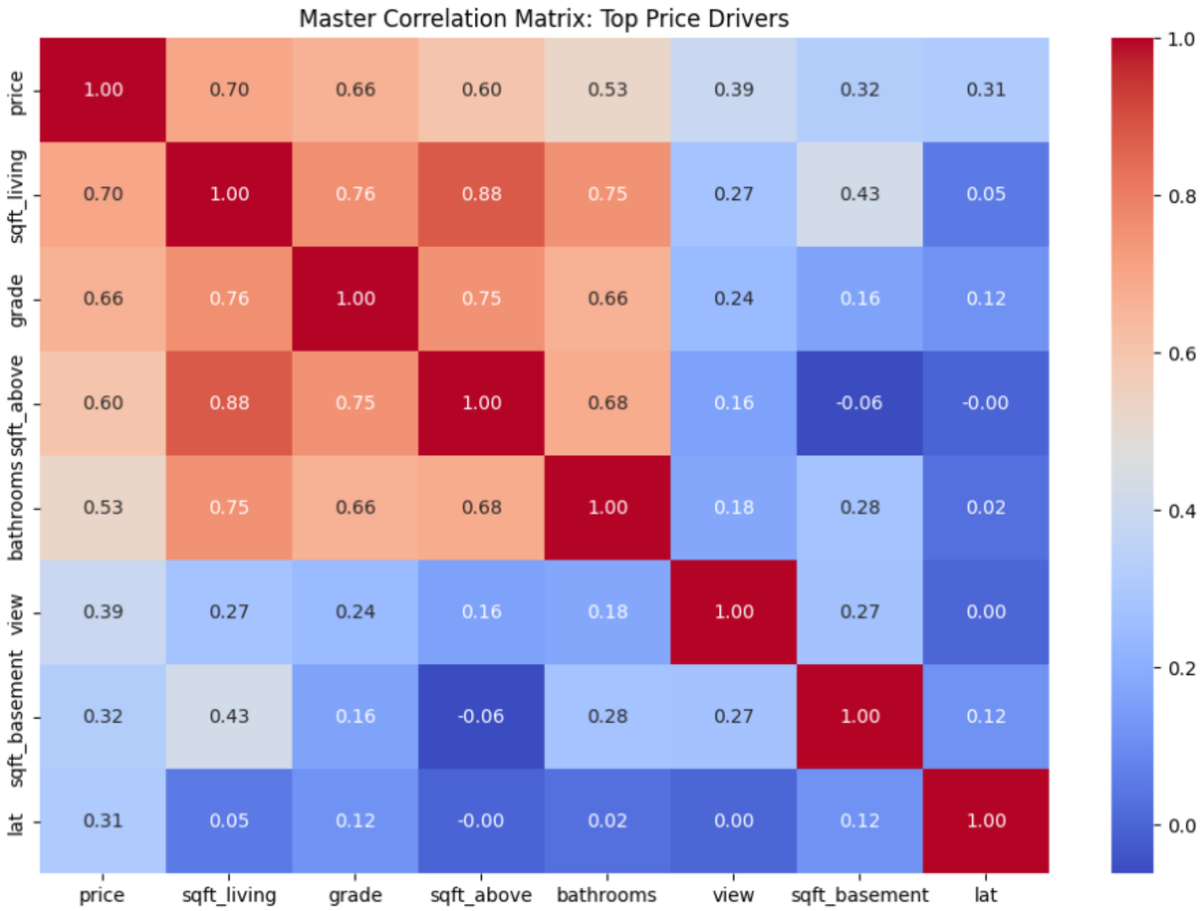
All The Images of EDA-



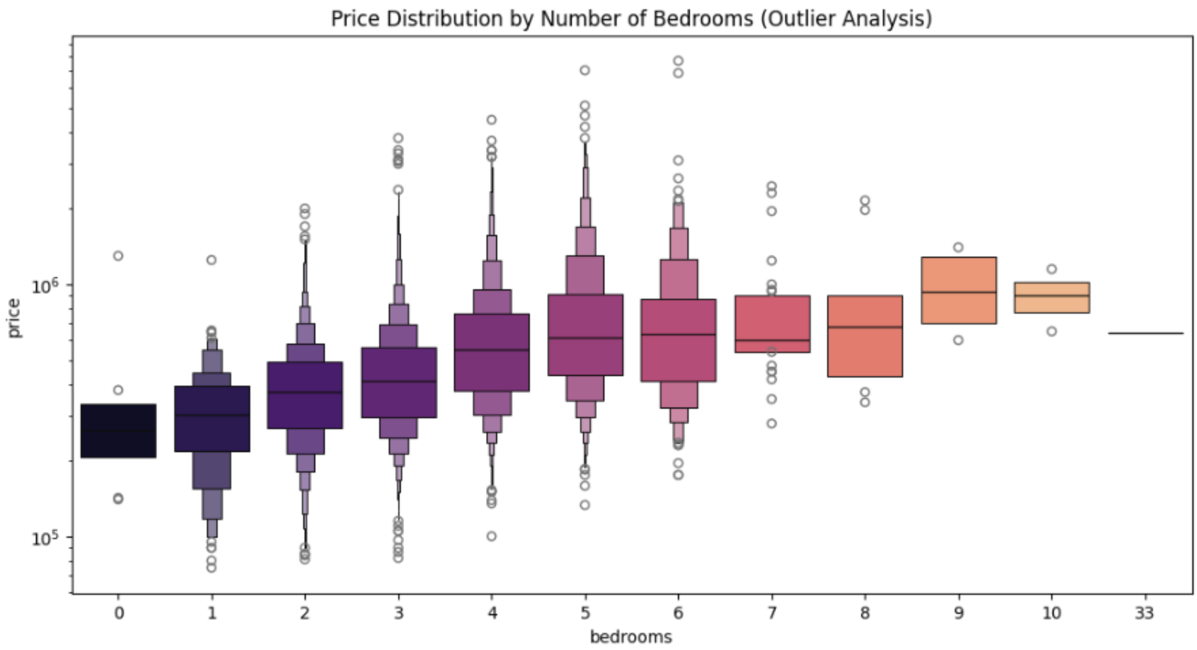

Price Trends: Original vs Renovated Houses Over Time
is_renovated
0
1
price
10^6
1900
1920
1940
1960
1980
2000
2020
yr_built

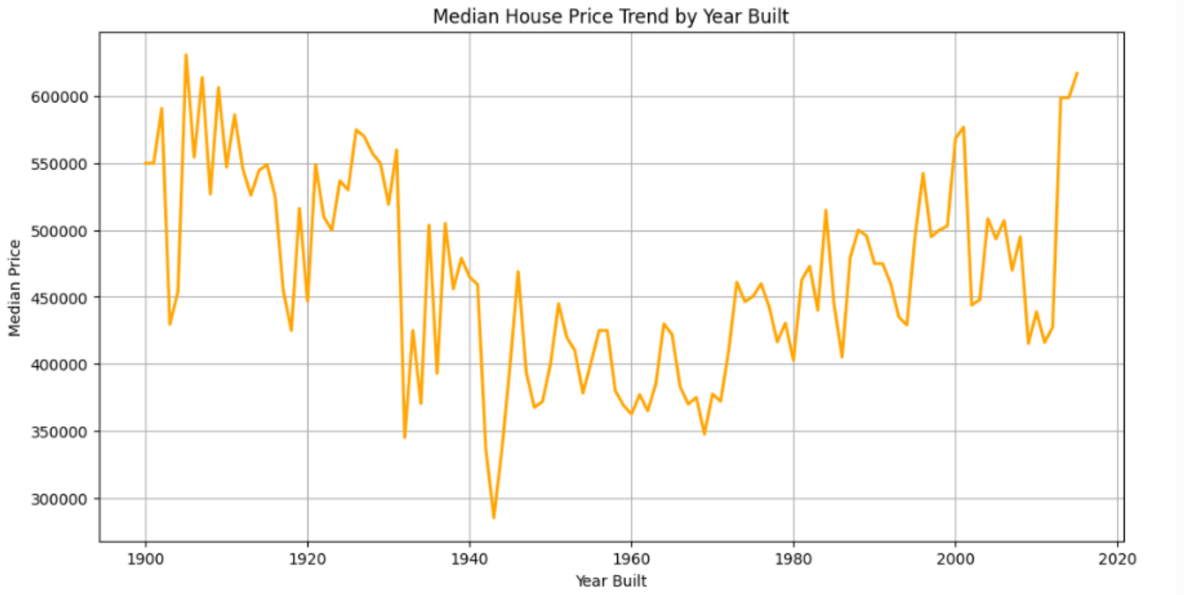

Normalized Price Distribution (Log Scale)
Count
0
100
200
300
400
500
600
700
800
11
12
13
14
15
16
price_log

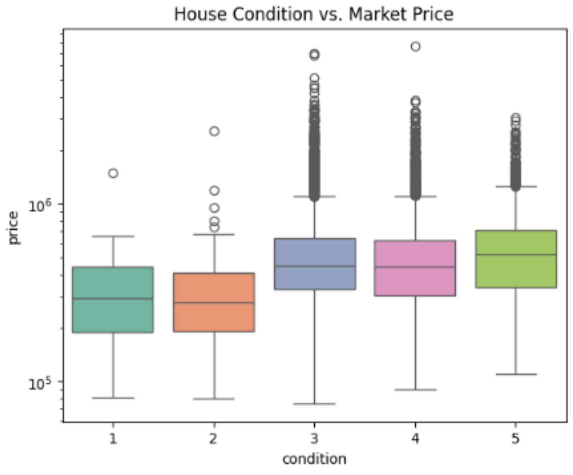

Original Price (Highly Skewed)

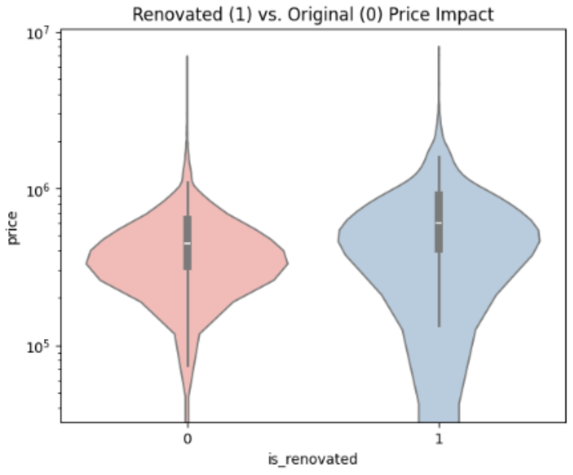

Master Correlation Matrix: Top Price Drivers

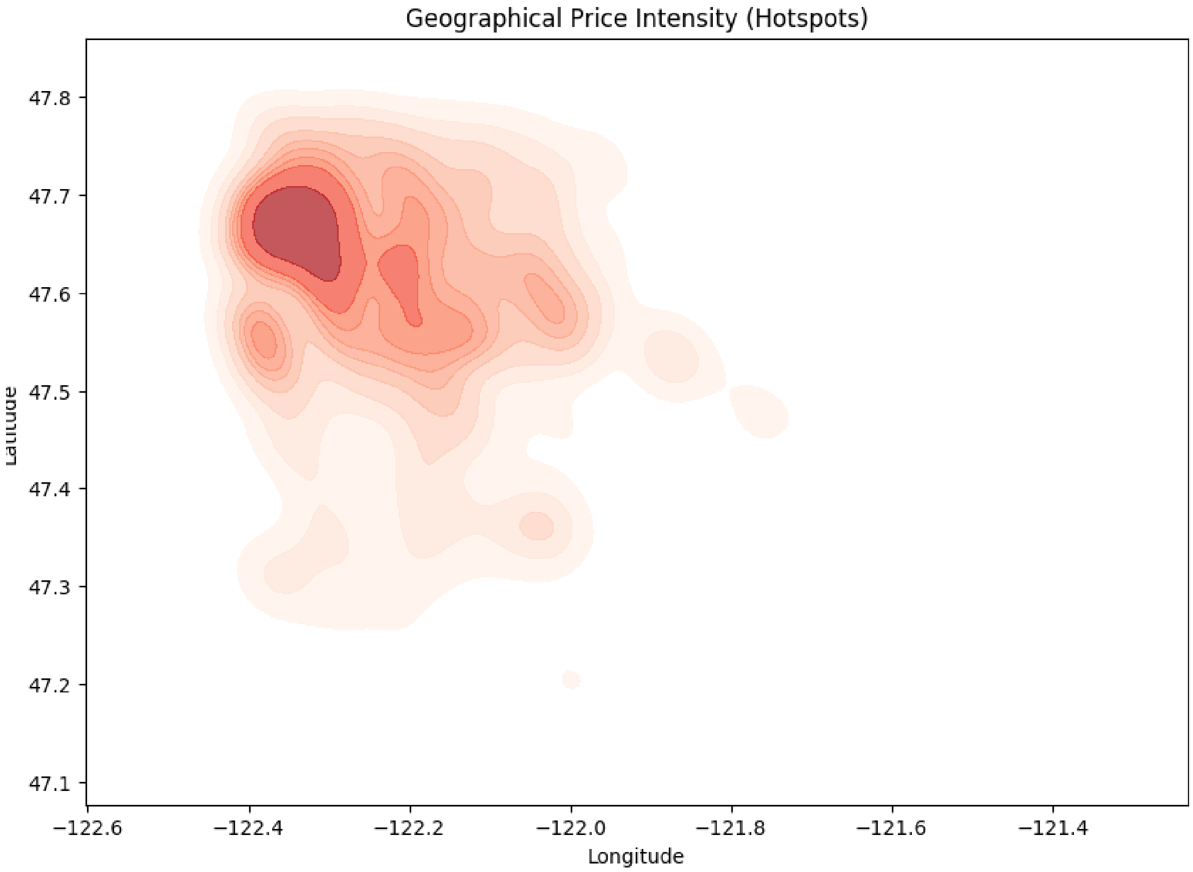
	price	sqft_living	grade	sqft_above	bathrooms	view	sqft_basement	lat
price	1.00	0.70	0.66	0.60	0.53	0.39	0.32	0.31
sqft_living	0.70	1.00	0.76	0.88	0.75	0.27	0.43	0.05
grade	0.66	0.76	1.00	0.75	0.66	0.24	0.16	0.12
sqft_above	0.60	0.88	0.75	1.00	0.68	0.16	-0.06	-0.00
bathrooms	0.53	0.75	0.66	0.68	1.00	0.18	0.28	0.02
view	0.39	0.27	0.24	0.16	0.18	1.00	0.27	0.00
sqft_basement	0.32	0.43	0.16	-0.06	0.28	0.27	1.00	0.12
lat	0.31	0.05	0.12	-0.00	0.02	0.00	0.12	1.00

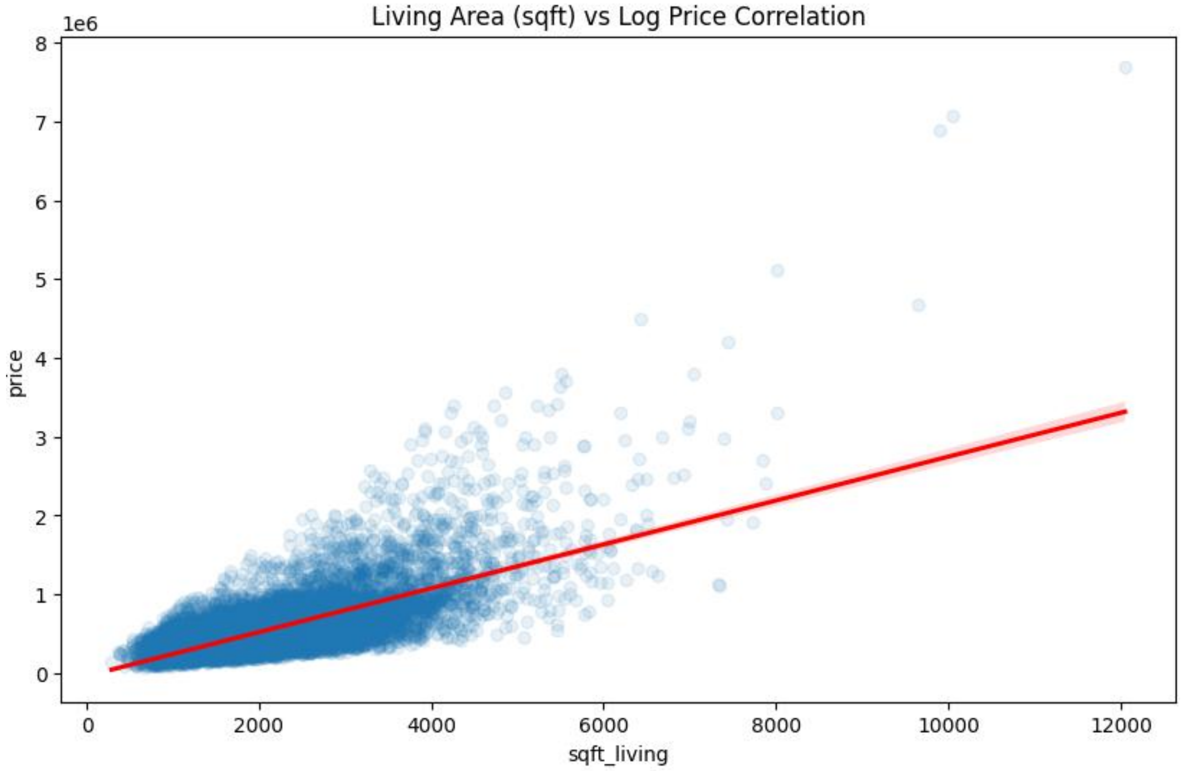

Price Distribution by Number of Bedrooms (Outlier Analysis)


Median House Price Trend by Year Built


House Condition vs. Market Price


Renovated (1) vs. Original (0) Price Impact


Geographical Price Intensity (Hotspots)
Latitude
47.8
47.7
47.6
47.5
47.4
47.3
47.2
47.1
−122.6
−122.4
−122.2
−122.0
−121.8
−121.6
−121.4
Longitude


Living Area (sqft) vs Log Price Correlation
1e6
price
8
7
6
5
4
3
2
1
0
0
2000
4000
6000
8000
10000
12000
sqft_living

3.2 Image Data

- **Source:** Satellite images were programmatically fetched for every house coordinate in the dataset using the Esri World Imagery API.
- **Dimensions:** Images were standardized to 400 X 400 pixels with 3 RGB channels.
- **Augmentation:** To prevent overfitting, standard augmentations (random horizontal flips, slight rotations) were applied during the training phase.

4. Methodology: Multimodal Architecture

We implemented a **Late-Fusion** strategy, where visual and tabular features are processed in separate "branches" before being merged.

4.1 The Vision Branch

We utilized a **ResNet-18** backbone pre-trained on ImageNet.

- **Feature Extraction:** The final fully connected layer of ResNet-18 was removed.
- **Fine-tuning:** The convolutional layers were frozen initially and then gradually unfrozen to allow the model to learn specific real-estate spatial patterns (e.g., roof size, road width).
- **Output:** This branch outputs a 1D feature vector representing the visual embedding of the property.

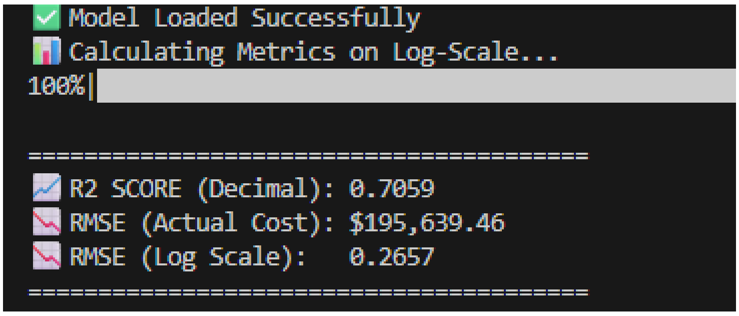
4.2 The Tabular Branch (MLP)

A standard Multi-Layer Perceptron (MLP) was designed for the structured data.

- **Structure:** 3 Dense layers with decreasing neuron counts (e.g., 128 -> 64 -> 32).
- **Regularization:** Batch Normalization and Dropout layers (rate=0.3) were interspersed to prevent overfitting.
- **Activation:** ReLU activation functions were used for non-linearity.

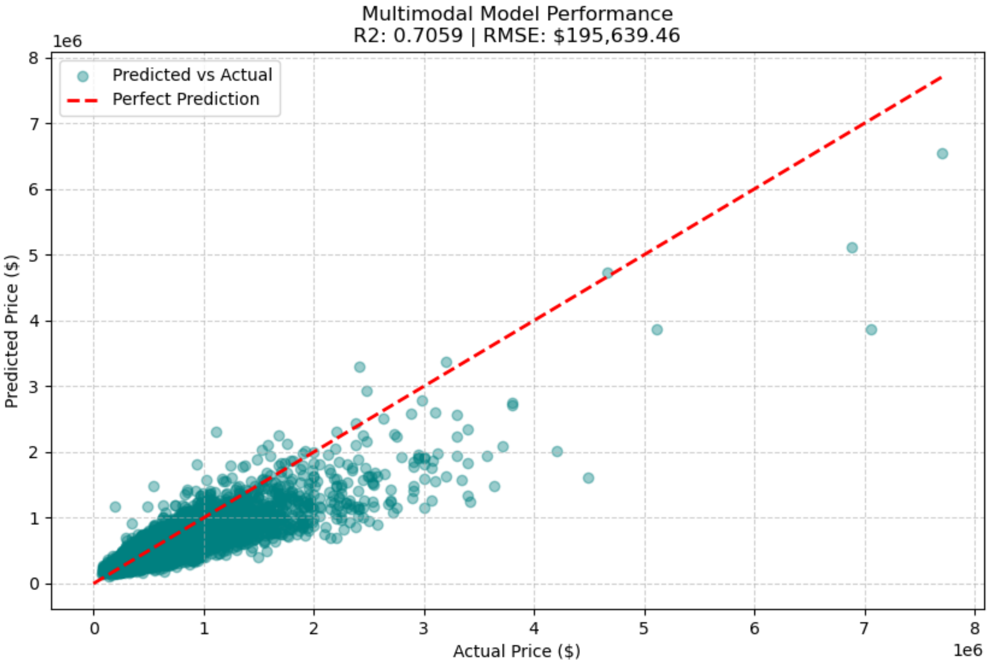
4.3 Fusion & Regression Head

- The outputs from the Vision Branch and Tabular Branch were concatenated. This combined vector was passed through a final sequence of dense layers to predict the single continuous output: the log-price.



```
✅ Model Loaded Successfully
📊 Calculating Metrics on Log-Scale...
100%|

========================================
📈 R2 SCORE (Decimal): 0.7059
📉 RMSE (Actual Cost): $195,639.46
📉 RMSE (Log Scale):   0.2657
========================================
```



5. Comparative Analysis & Results

We compared our deep learning approach against an industry-standard gradient boosting model.

5.1 Performance Metrics

Model Architecture	**R2 Score**	**RMSE (Log Scale)**
XGBoost (Baseline)	**0.880**	**0.16**
Multimodal DL (ResNet + MLP)	0.706	0.26

5.2 Discussion: The Performance Gap

The XGBoost baseline outperformed the multimodal network by approximately 10%. Several factors contributed to this:

1. **Tabular Dominance:** Real estate prices are overwhelmingly driven by square footage and location grade—features that tree-based models like XGBoost handle exceptionally well with minimal tuning.
2. **Data Volume:** Deep learning models, especially multimodal ones, typically require massive datasets to generalize better than gradient boosting on tabular tasks.
3. **Image Quality/Relevance:** While satellite images provide neighborhood context, they do not capture the *interior* condition of the house, which is often a stronger predictor of price than the roof or street view.

However, the 0.706 R2 score is significant. It proves that the model successfully learned to correlate visual patterns (e.g., large houses, well-planned streets) with higher prices, serving as a successful Proof of Concept.

6. Visual Explainability (Grad-CAM)

To ensure the "Black Box" nature of the neural network did not obscure insights, we applied **Gradient-weighted Class Activation Mapping (Grad-CAM)** to the final convolutional layer of the ResNet branch.

Key Findings:

- **Road Connectivity:** The model frequently activated (showed "hot spots") around major roads and intersections, indicating accessibility is a learned value driver.
- **Green Spaces:** For higher-priced properties, the heatmaps showed strong activation over parks, lawns, and water bodies.
- **Density:** In lower-priced areas, the model focused on the density of neighboring structures.

7. Conclusion & Future Scope

7.1 Conclusion

This project successfully demonstrated the end-to-end pipeline of a multimodal real estate pricer. While the XGBoost baseline remains the superior pure-prediction tool for this specific dataset, the Multimodal model offers a layer of "visual intelligence" that tabular models lack.