# Satellite Imagery-Based Property Valuation Report

Rishabh Singhal
23113127

## 1. Overview

1. **Data Preparation**
- Tabular features were standardized using StandardScaler.
- Satellite images were preprocessed to match CNN input requirements.

2. **Image Feature Extraction**
- A pretrained ResNet-50 model was used to extract 2048 deep visual features from satellite images.
- PCA was applied to reduce image features to 50 components, retaining essential visual information.

3. **Modeling Strategy**
- Two approaches were implemented:
  - Tabular-only baseline models
  - Multimodal models combining tabular and image features
- Multiple algorithms were trained, including Linear Regression, Random Forest, Gradient Boosting, and XGBoost.
-

4. **Model Evaluation**
- Models were evaluated using RMSE and $R^2$ metrics.
- Performance comparison was used to identify the best-performing model.
-

5. **Model Interpretability**
- Grad-CAM visualizations were generated to analyze the influence of satellite image features on property valuation.

## 2. Exploratory Data Analysis

### Dataset Overview :

- The dataset consists of 21 features, including:
  - 12 continuous variables
  - 4 categorical variables
  - 3 discrete variables
  - Date and ID fields

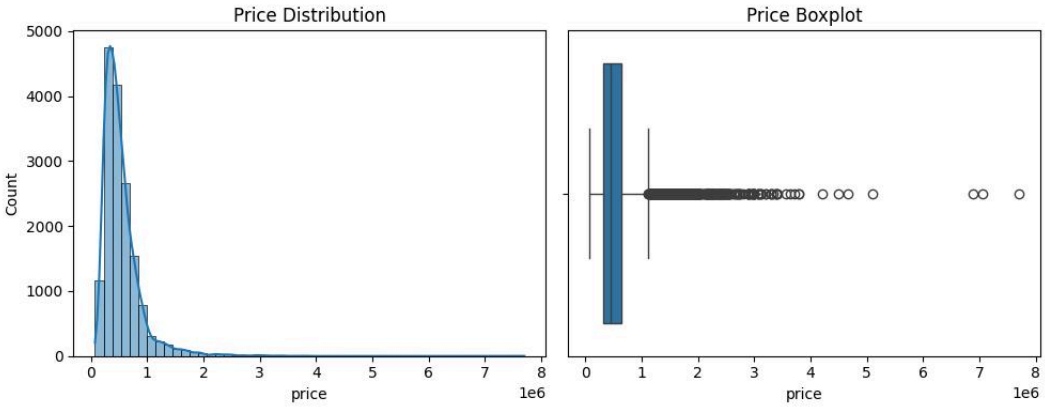
- The dataset contains 16,209 observations, out of which 16,110 have unique property IDs.

- Duplicate IDs were identified and removed to ensure a one-to-one mapping between tabular records and satellite images.

- No missing values were observed across any features, ensuring complete and reliable data for modeling.

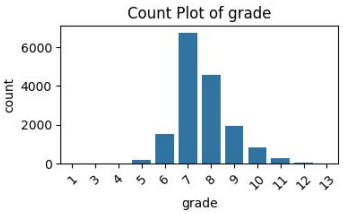
# Univariate Analysis

## 1. Price

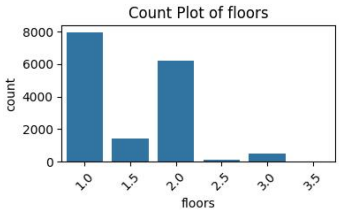



- The price distribution is right-skewed, meaning most properties are clustered at lower prices, while a few properties are extremely expensive.
- The mean price (~538,000) is higher than the median (likely lower based on the skew), which is typical for right-skewed distributions.
- The boxplot shows numerous high-end outliers. These are not errors; they represent high-value properties that are naturally rare.

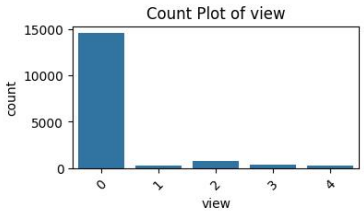
## 2. Categorical features



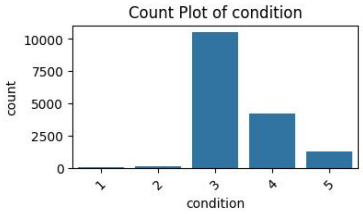

- Most homes fall within grades 7–9, representing average to above-average construction quality.
- Very low (≤3) and very high (≥12) grades are extremely rare, indicating limited extremes in build quality.




- Single-story (1.0) and double-story (2.0) houses dominate the dataset.
- Multi-level homes (≥3 floors) are uncommon, indicating a preference for low-rise residential buildings.

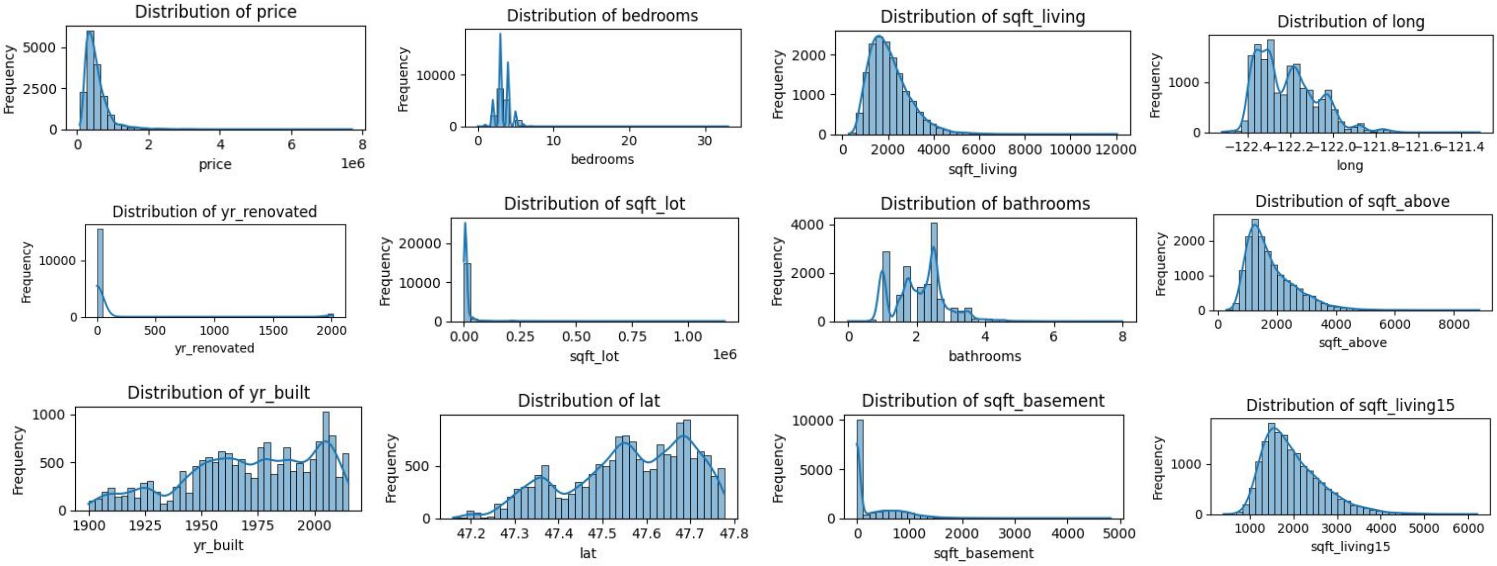



- The majority of properties do not have a special view (view = 0).
- Properties with higher view ratings are scarce, making them potentially premium features.




- Most properties are in average to good condition (ratings 3–4), indicating generally well-maintained housing stock.
- Very poor (1–2) and excellent (5) conditions are rare, suggesting limited extreme cases.

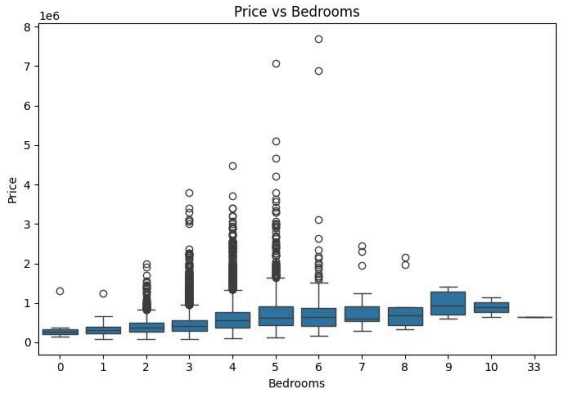
## 3. Numerical features



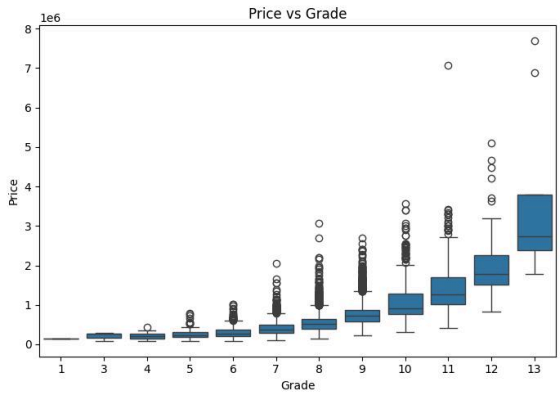
- **Size-related features** (sqft_living, sqft_above, sqft_living15) show right-skewed distributions, with most houses having moderate sizes and a few very large houses creating long tails, indicating high variability but rare extremes.
- **Lot size features** (sqft_lot, sqft_lot15) are extremely right-skewed with heavy tails, where most properties have small lots and only a few possess very large land areas.
- **The basement area** (sqft_basement) has a strong spike at zero, showing that many houses do not have basements, while the remaining values are right-skewed.
- **Location features** (latitude, longitude, zipcode) display multi-modal distributions, indicating geographical clustering across different neighborhoods and urban zones.
- **Time-related features** (yr_built, yr_renovated) show that newer houses are more common, while the renovation year has a large spike at zero, indicating most houses were never renovated.

## Price Vs Key Features

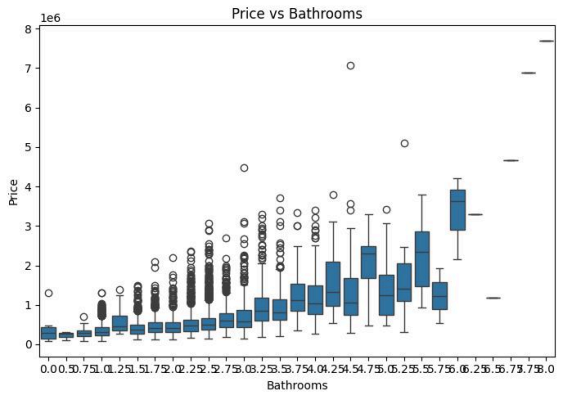

Price vs Sqft Living:
- Property prices increase steadily with larger living areas, indicating a strong positive correlation.
- Larger homes show greater price dispersion, suggesting that size alone does not fully determine value at higher ranges.
- Extremely large houses tend to act as outliers with very high prices.



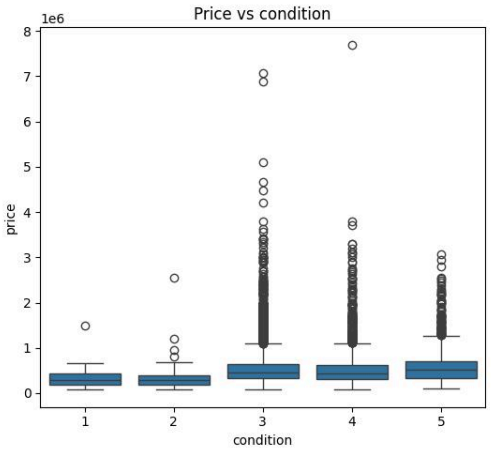

Price vs Bedrooms:
- Prices generally rise as the number of bedrooms increases, especially up to mid-range values.
- Beyond a certain number of bedrooms, the price increase slows, showing diminishing returns.
- Very high bedroom counts often display wide variability, indicating inefficient layouts or niche properties.

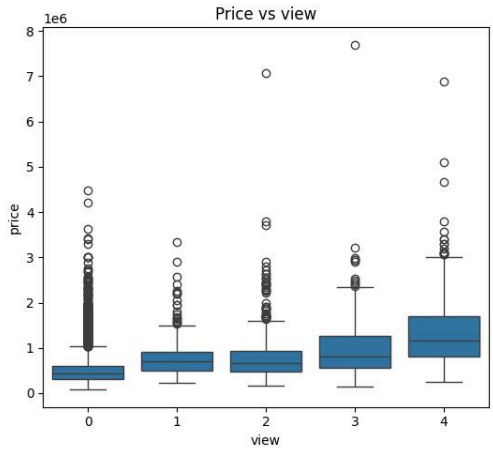


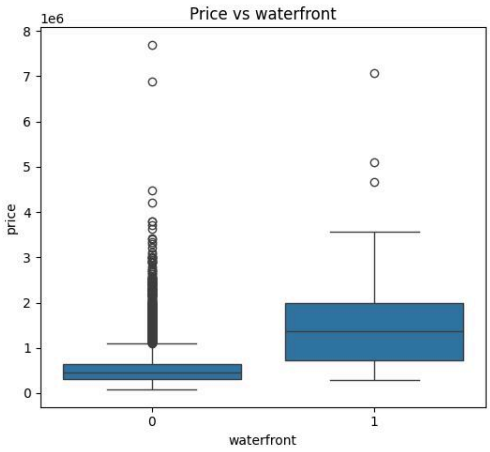

Price vs Grade:
- House grade shows a strong and consistent positive relationship with price.
- Higher grades are associated with sharp increases in median property value.
- Price variability increases at higher grades, reflecting differences in premium features and overall quality




Price vs Bathrooms:
- Properties with more bathrooms consistently achieve higher prices.
- The median price rises with bathroom count, highlighting bathrooms as a strong value driver.
- Higher bathroom numbers also introduce more price variability, reflecting luxury and design differences.




Price vs Condition:

- Property prices show a mild upward trend as condition improves, indicating better-maintained houses generally sell for more.
- Differences between lower condition levels are relatively small compared to other features.
- Higher condition ratings exhibit greater price variability, suggesting maintenance quality interacts with other premium factors.

Price vs View:

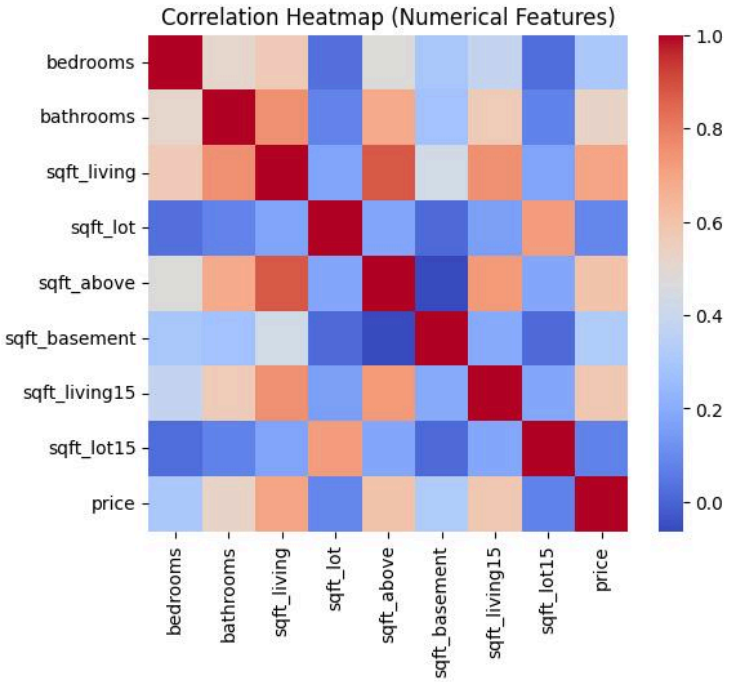
- Prices increase consistently with better view ratings, showing a clear positive relationship.
- Properties with higher view scores have noticeably higher median prices.
- Price dispersion grows with view quality, reflecting the premium placed on scenic or desirable surroundings.

Price vs Waterfront:

- Waterfront properties command significantly higher prices than non-waterfront properties.
- The median price for waterfront homes is substantially higher, indicating a strong location premium.
- Waterfront homes also show wider price variation, reflecting differences in size, luxury, and exact location.

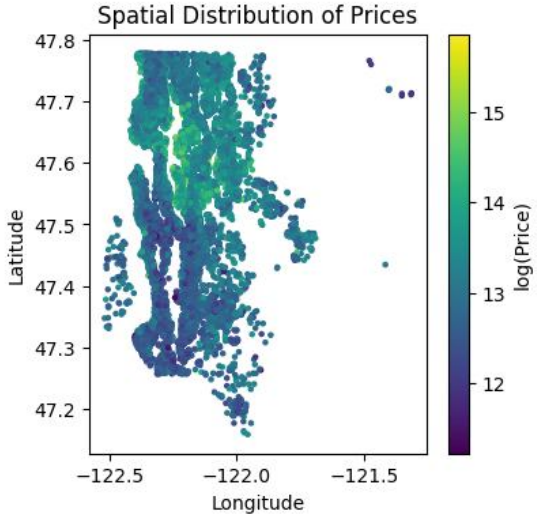
## Correlation Analysis

- sqft_living has the strongest positive correlation with price, making it the most influential numerical feature.
- Bathrooms correlate more strongly with price than bedrooms, highlighting the importance of functionality over room count.
- Lot size variables (sqft_lot, sqft_lot15) show weak correlation with price, indicating limited direct impact.
- sqft_basement has a mild positive correlation, adding value but not being a key driver.
- Size-related features are highly correlated with each other, indicating multicollinearity and the need for careful feature handling in modeling.




## Geospatial Analysis

- House prices exhibit clear spatial clustering, with high-value properties concentrated in specific geographic pockets.
- Premium-priced houses are primarily located around central and waterfront-adjacent regions, as indicated by higher log(price) values.
- Lower-priced properties are more dispersed and dominate peripheral areas.
- The visualization highlights strong location dependence of property prices, justifying the inclusion of latitude and longitude as important predictive features.



- Property prices increase with higher neighborhood average living area, showing a clear positive relationship.
- Neighborhood-level features capture locality effects better than individual property size alone.
- Price variability increases for larger neighborhoods, indicating influence of additional factors such as amenities and views.
- Extremely high-priced outliers suggest the presence of luxury properties beyond what size alone explains.


### Feature Engineering

- The tabular data was cleaned by handling missing values, removing duplicates, and standardizing data types to ensure consistency across all features.
- Numerical features were scaled so that they could be effectively used by machine learning models.
- Satellite images were processed using a pretrained CNN to extract high-level visual embeddings, which were then reduced using PCA to retain the most informative components.
- The reduced image features were merged with the engineered tabular features to form a multimodal dataset, with identical preprocessing applied to training, validation, and test data to avoid data leakage.

## 3. Financial / Visual Insights from Grad-CAM

To understand how visual cues from satellite imagery influence property valuation, Grad-CAM was applied to the convolutional neural network used for image feature extraction. Grad-CAM highlights spatial regions in the satellite images that contributed most strongly to the model's price prediction, enabling an interpretable link between visual features and real estate value.

The following analysis compares Grad-CAM outputs for low-priced, mid-priced, and high-priced properties.

### 1. Low-Priced Property (~298k)

- Grad-CAM highlights are concentrated on building roofs and paved surfaces, with minimal attention on surrounding greenery.
- Tree cover is sparse and fragmented, indicating limited environmental quality.
- This suggests lower valuation is associated with dense construction, reduced green space, and compact neighborhood layouts.
- Financially, such environments are often linked to higher congestion and lower perceived livability.

ID: 3629960550 | Price: 450,000


## 2. Mid-Priced Property (~450k)

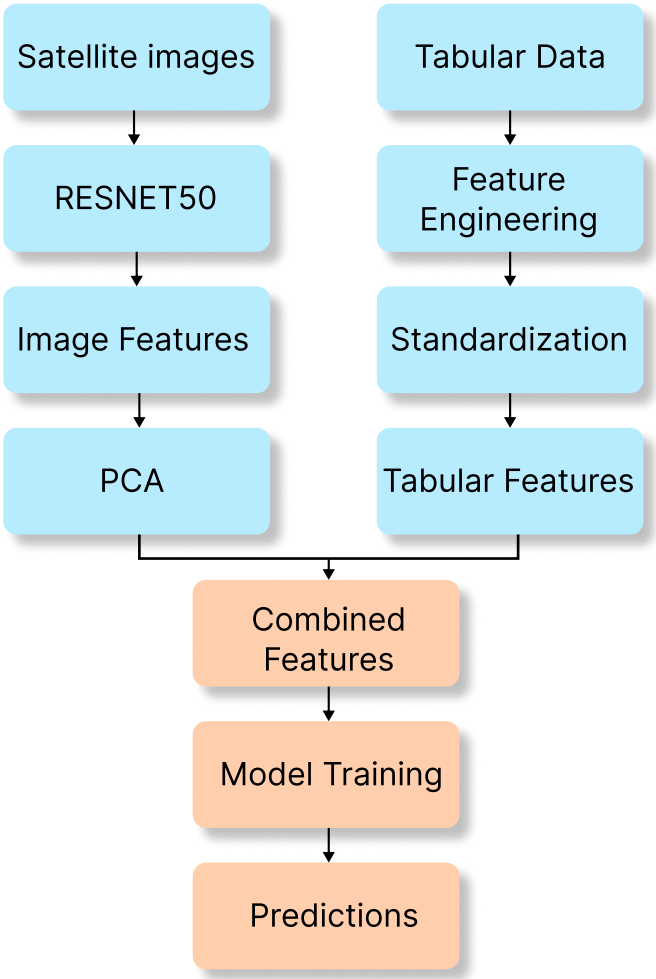
- Attention is split between partial tree cover and built structures.
- The presence of moderate greenery improves visual appeal, but it is interspersed with concrete and roads.
- This balance reflects a transitional neighborhood, offering reasonable living quality but not premium conditions.
- Property value benefits from greenery, though gains are limited by surrounding density.

ID: 8901500178 | Price: 700,000

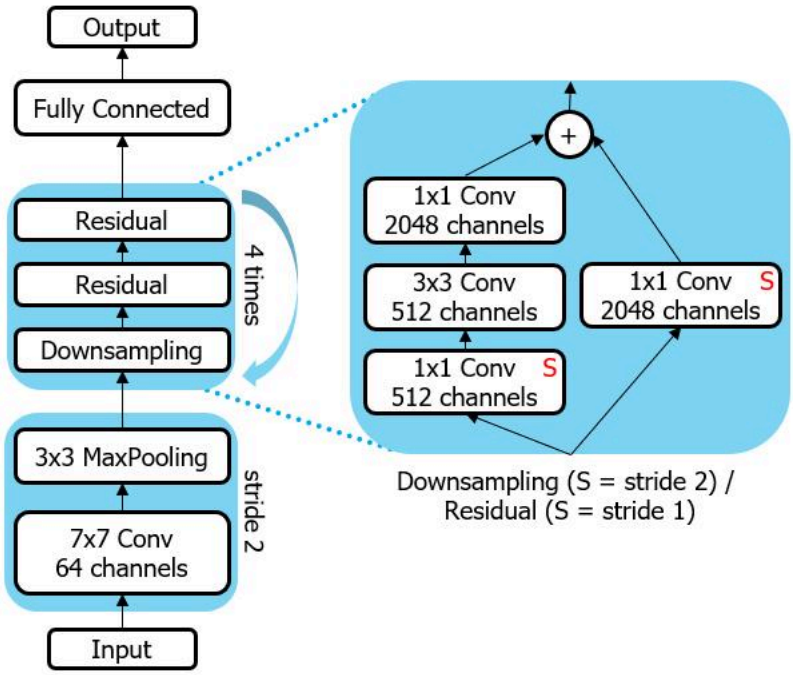

## 3. High-Priced Property (~700k)

- Strong Grad-CAM activation appears over dense tree canopies and landscaped areas.
- Buildings are visually embedded within green surroundings, indicating low-density, high-quality residential zones.
- The model clearly associates extensive greenery and spatial openness with premium pricing.
- Financially, such environments command higher prices due to improved aesthetics, privacy, and long-term desirability.

## 4. Architecture Diagram




CNN Architecture



# 5. Result

## Tabular Model Performance

| | Type | Model | RMSE | MAE | R2 |
|---|---|---|---|---|---|
| 5 | Tabular | XGBoost | 105691.756500 | 63053.230469 | 0.908617 |
| 4 | Tabular | GradientBoosting | 114692.877876 | 69813.168566 | 0.892390 |
| 3 | Tabular | RandomForest | 119709.961346 | 69373.057726 | 0.882769 |
| 1 | Tabular | Ridge | 190439.052755 | 126733.233410 | 0.703317 |
| 2 | Tabular | Lasso | 191866.302768 | 128400.219100 | 0.698853 |
| 0 | Tabular | LinearRegression | 191867.111203 | 128405.707391 | 0.698850 |

- Multiple regression models were evaluated using tabular features to predict property prices.
- XGBoost achieved the best performance with an $R^2$ score of 0.9086, along with the lowest RMSE (105,691) and MAE (63,053).
- Gradient Boosting and Random Forest also showed strong predictive performance but were slightly inferior to XGBoost.
- Linear models (Linear Regression, Ridge, and Lasso) performed comparatively worse, indicating the presence of strong non-linear relationships in the data.

## Multimodal Model Performance

| | Type | Model | RMSE | MAE | R2 |
|---|---|---|---|---|---|
| 8 | Multimodal | XGBoost | 109395.813302 | 65800.578125 | 0.902100 |
| 7 | Multimodal | GradientBoosting | 118642.788692 | 71827.790550 | 0.884850 |
| 6 | Multimodal | RandomForest | 127324.660908 | 74275.886886 | 0.867381 |

- Multimodal models combining tabular data and satellite image features were evaluated to assess the impact of visual information.
- Among these, Multimodal XGBoost achieved the best performance with an $R^2$ score of 0.9021.
- The inclusion of satellite imagery provided additional visual context related to surroundings and land characteristics.
- However, the multimodal approach did not surpass the best tabular-only model, indicating that tabular features captured most of the predictive signal for price estimation.

## Comparison: Tabular vs Multimodal

- Overall, tabular-only models outperformed multimodal models across all evaluation metrics.
- This indicates that engineered numerical features (such as size, location, and property attributes) capture the primary drivers of property prices.
- Satellite imagery added complementary contextual information, such as surroundings and land characteristics.
- However, the visual features played a secondary role, improving interpretability more than predictive accuracy.

## Best Model Selection

- Tabular XGBoost was selected as the final model based on overall performance across all evaluation metrics.
- It achieved the highest $R^2$ score (0.9086), indicating strong explanatory power.
- The model also recorded the lowest RMSE and MAE, reflecting superior prediction accuracy.
- These results confirm that engineered tabular features capture the dominant factors influencing property prices more effectively than the multimodal approach.

## Key Observation

- Satellite imagery contributes valuable visual and environmental context, enhancing qualitative understanding of property surroundings.
- Tabular features remain the strongest predictors of property prices in this dataset, as reflected by higher quantitative performance metrics.
- The multimodal approach improves interpretability, even when it does not outperform tabular-only models.
- Grad-CAM visualizations provide explainability, highlighting which visual regions (e.g., greenery or built-up areas) influence model predictions.