House Price Prediction using Tabular Data and Satellite Imagery

Akshay Vishwakarma
Enrollment No: 22119005

1. Project Overview: Approach and Modeling Strategy

1. Environment Setup

- Google Drive was first connected to the Google Colab environment to ensure persistent storage.
- This allowed seamless access to datasets, satellite images, and trained models throughout the project.

2. Project Organization

- A clean and reproducible folder structure was created to separate raw data, processed data, and images.
- Raw and processed datasets were stored separately to avoid data leakage and ensure reproducibility.
- Satellite images were organized into raw and processed folders for efficient handling during training.

3. Exploratory Data Analysis (EDA)

- Histograms were plotted to analyze feature distributions and detect skewness.
- Scatter plots were used to:
  - Identify outliers
  - Observe relationships between features and house prices
- Correlation analysis was performed to detect multicollinearity, and highly correlated features were addressed.
- Outliers identified through scatter plots were removed to improve model stability.

4. Feature Engineering

- New features such as house age were derived.
- Log transformation was applied to skewed numerical variables to normalize distributions.
- The final cleaned and engineered dataset was saved as a processed file for modeling.

5. Geospatial and Visual Analysis

- Geographic features (latitude and longitude) were analyzed using scatter plots to understand spatial price patterns.

- Satellite image regions were visually analyzed to study how surrounding environments influence house prices.
- Price distributions were compared across different image regions.
- Grad-CAM was later used to interpret which regions of satellite images influenced predictions the most.

6. Tabular Model Development

- Tabular features were divided into:
  - Binary features (e.g., waterfront, renovation status)
  - Ordinal features (e.g., bedrooms, condition, grade)
  - Continuous features (e.g., square footage, location coordinates)
- A preprocessing pipeline was created for scaling and transformation.
- Multiple classical models were trained and evaluated:
  - Linear Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosting

7. Neural Network for Tabular Data

- A standalone neural network was trained using only tabular features.
- This served as a benchmark to compare performance against multimodal learning.

8. Satellite Image Acquisition

- Satellite images were downloaded using an external API via a custom data fetching script.
- Images were saved in structured folders corresponding to training and validation sets.
- Images were copied from Google Drive to Colab's local storage to improve training speed.

9. Train–Validation Split

- The dataset was split into training and validation sets.
- Image paths and tabular features were aligned using unique house IDs.
- Separate arrays were prepared for tabular data and image inputs.

10. Multimodal Dataset Creation

- A custom dataset class was implemented to load:
  - Satellite images
  - Corresponding tabular features
  - Target house prices
- Images were resized, normalized, and transformed before being fed into the model.

11. Multimodal Model Architecture

- A pretrained ResNet-18 CNN was used to extract features from satellite images.
- A fully connected neural network (MLP) was used to process tabular data.
- Image features and tabular features were concatenated and passed to a regression head for price prediction.
- The CNN backbone was frozen to prevent overfitting and reduce training time.

12. Model Training Strategy

- The model was trained using Mean Squared Error (MSE) loss.
- Adam optimizer with a low learning rate was used for stable training.
- Training and validation performance were tracked using:
    - MSE
    - R^2 score
- Early stopping was applied to prevent overfitting.
- The best-performing model was saved automatically.

13. Model Interpretability

- Grad-CAM was applied to the trained CNN to visualize which regions of satellite images influenced price predictions.
- This improved interpretability and validated the importance of visual features.

```
RangeIndex: 16209 entries, 0 to 16208
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             16209 non-null  int64
 1   date           16209 non-null  object
 2   price          16209 non-null  int64
 3   bedrooms       16209 non-null  int64
 4   bathrooms      16209 non-null  float64
 5   sqft_living    16209 non-null  int64
 6   sqft_lot       16209 non-null  int64
 7   floors         16209 non-null  float64
 8   waterfront     16209 non-null  int64
 9   view           16209 non-null  int64
 10  condition      16209 non-null  int64
 11  grade          16209 non-null  int64
 12  sqft_above     16209 non-null  int64
 13  sqft_basement  16209 non-null  int64
 14  yr_built       16209 non-null  int64
 15  yr_renovated   16209 non-null  int64
 16  zipcode        16209 non-null  int64
 17  lat            16209 non-null  float64
 18  long           16209 non-null  float64
 19  sqft_living15  16209 non-null  int64
 20  sqft_lot15     16209 non-null  int64
dtypes: float64(4), int64(16), object(1)
```

```
os.listdir()
!ls /content/drive/MyDrive/satellite_project
```

```
best_model.pth
data
data_fetcher.py
gradcam_outputs
High_Price_Properties_-_Explainability.jpeg
images
Low_Price_Properties_-_Explainability.jpeg
Mid_Price_Properties_-_Explainability.jpeg
model_training.ipynb
multimodal_architecture
multimodal_architecture.jpeg
preprocessing.ipynb
price_predictions.csv
__pycache__
```

2. Exploratory Data Analysis (EDA) & Visual Analysis

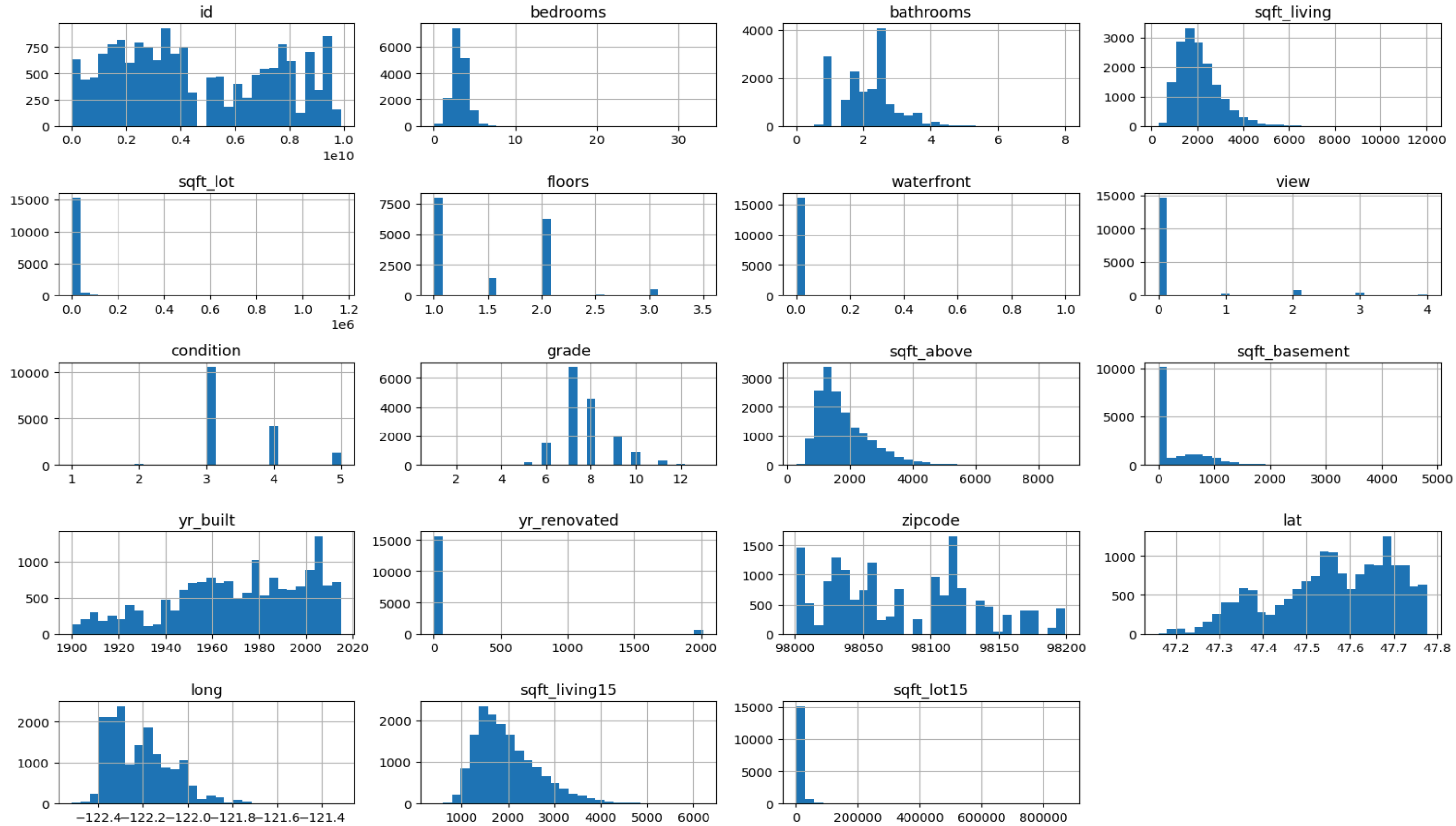


Figure 1: Distribution of House Prices (Before EDA) - *Histogram showing the original distribution of house prices, highlighting strong right skewness and the presence of extreme values.*

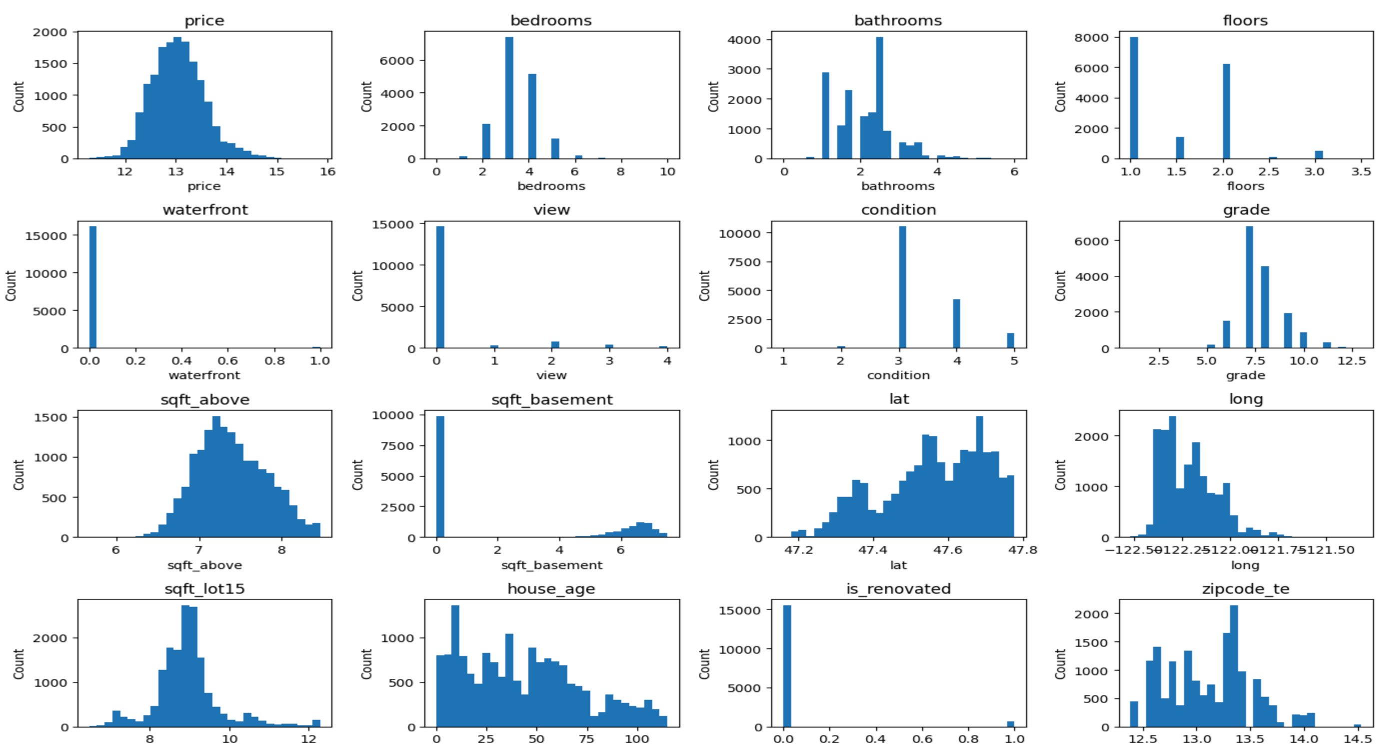
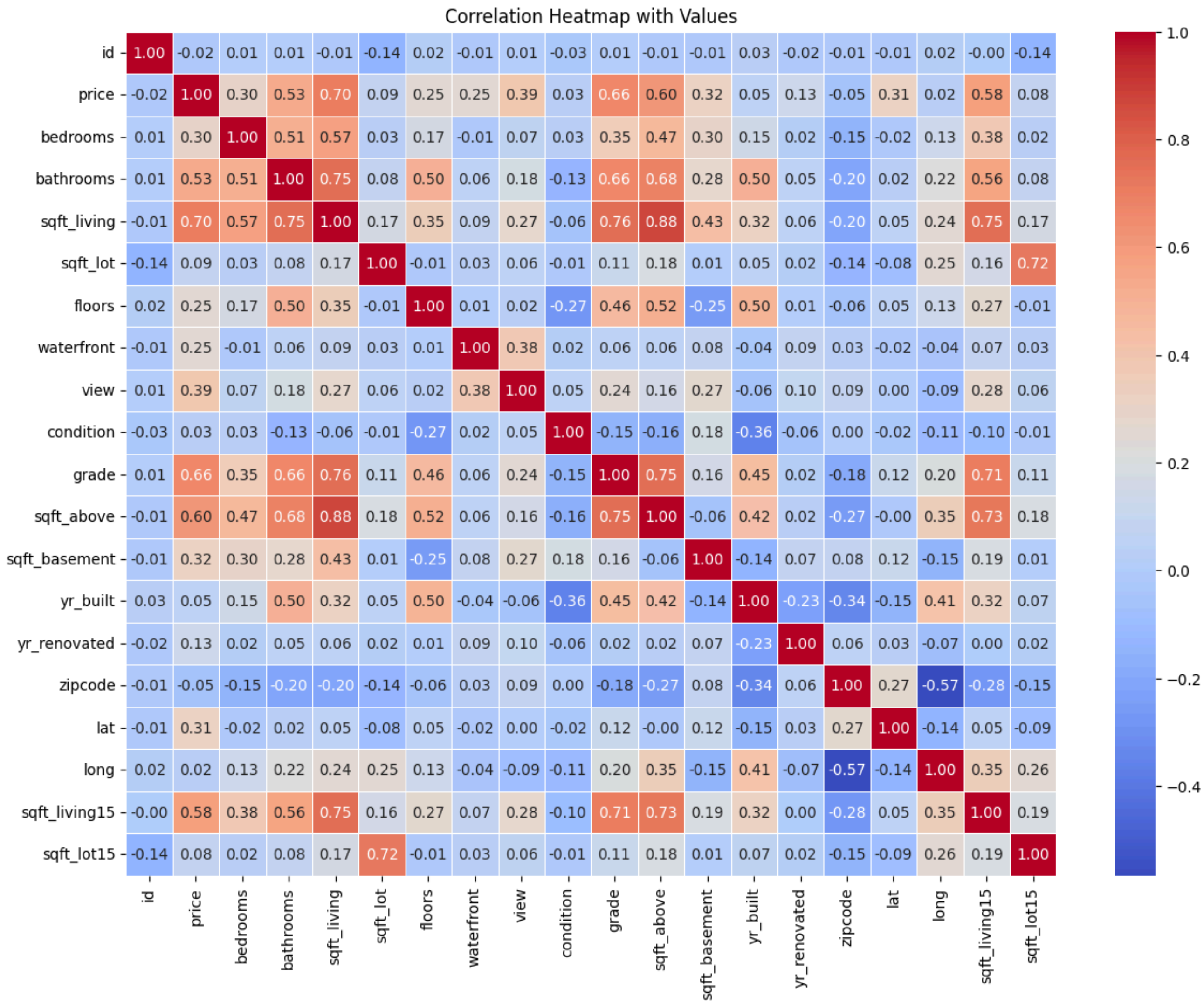
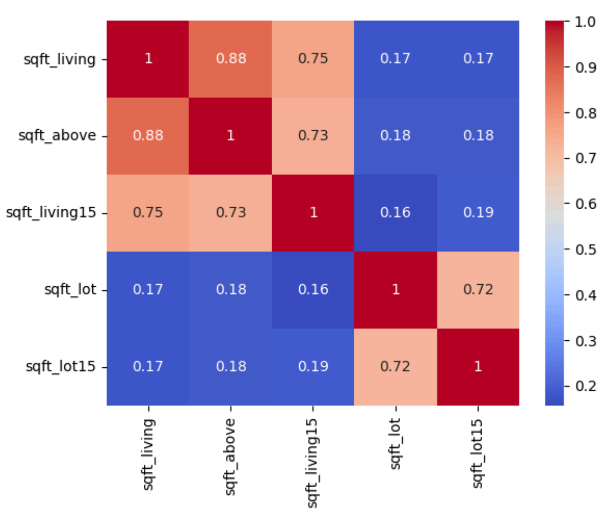


Figure 2: Distribution of House Prices (After EDA & Log Transformation) - *Histogram after preprocessing, demonstrating reduced skewness and a more normalized price distribution suitable for modeling.*





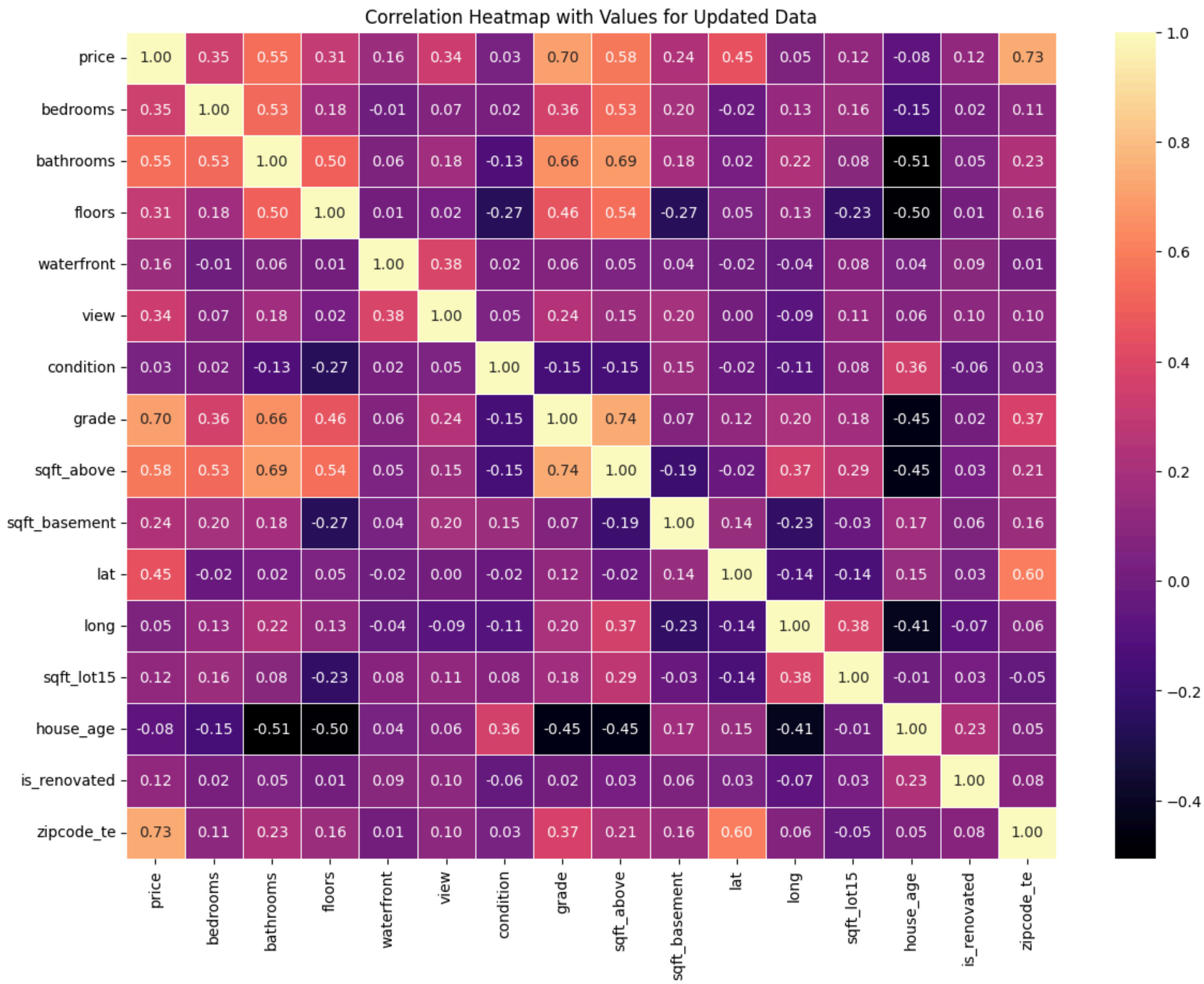
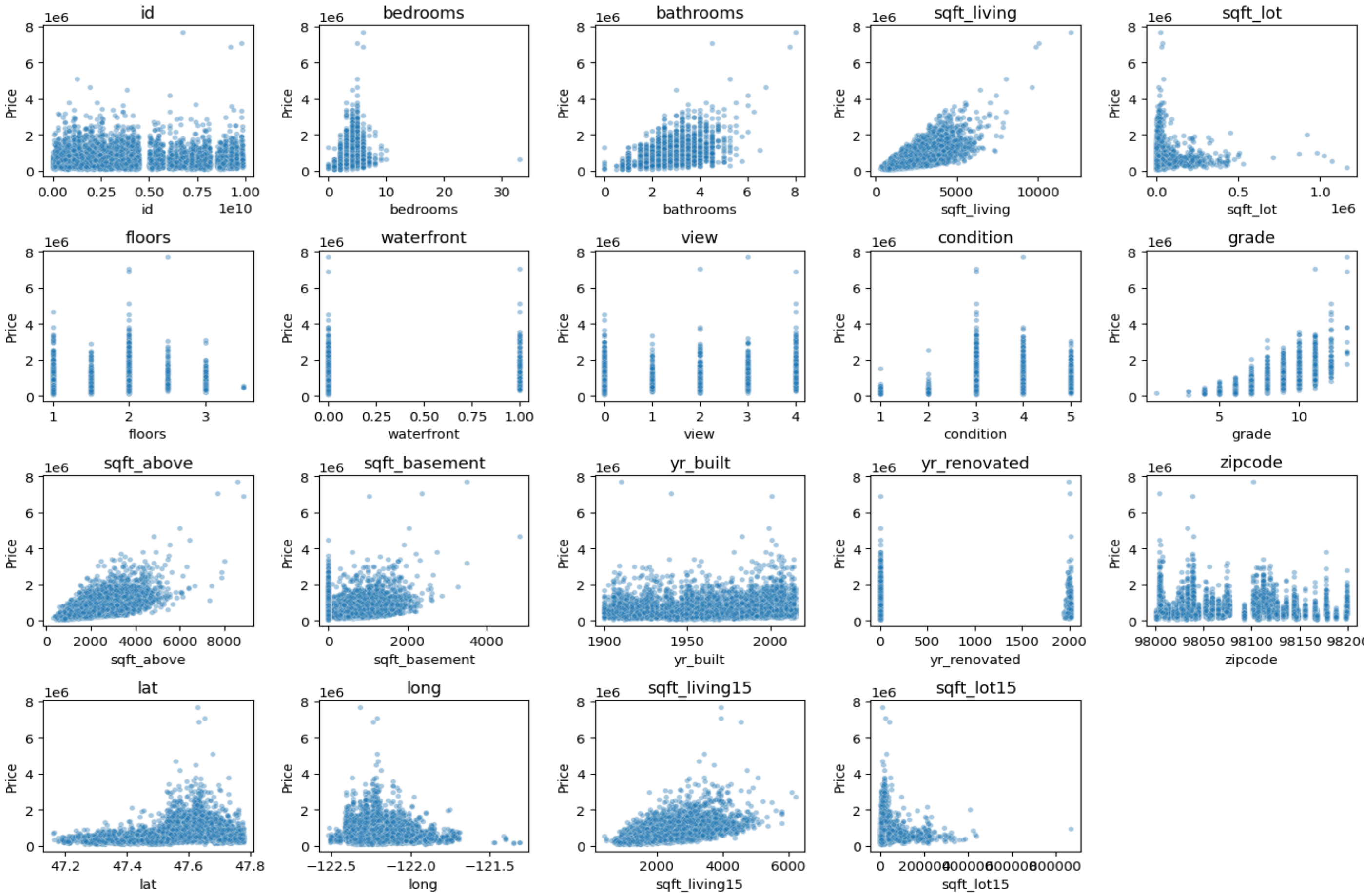


Figure 3: Correlation Heatmap of Selected Features (After Multicollinearity Reduction)

Correlation heatmap after removing highly correlated variables, resulting in a more stable and interpretable feature set.

Figure 4: Scatter Plot of Continuous Features vs House Price (After and Before Outlier Removal and Log Transformation)

Scatter plot showing improved linear relationships and reduced variance after removing outliers and applying log transformation to continuous features.


id
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
yr_built
yr_renovated
zipcode
lat
long
sqft_living15
sqft_lot15
Price
1e6
1e10
1e6

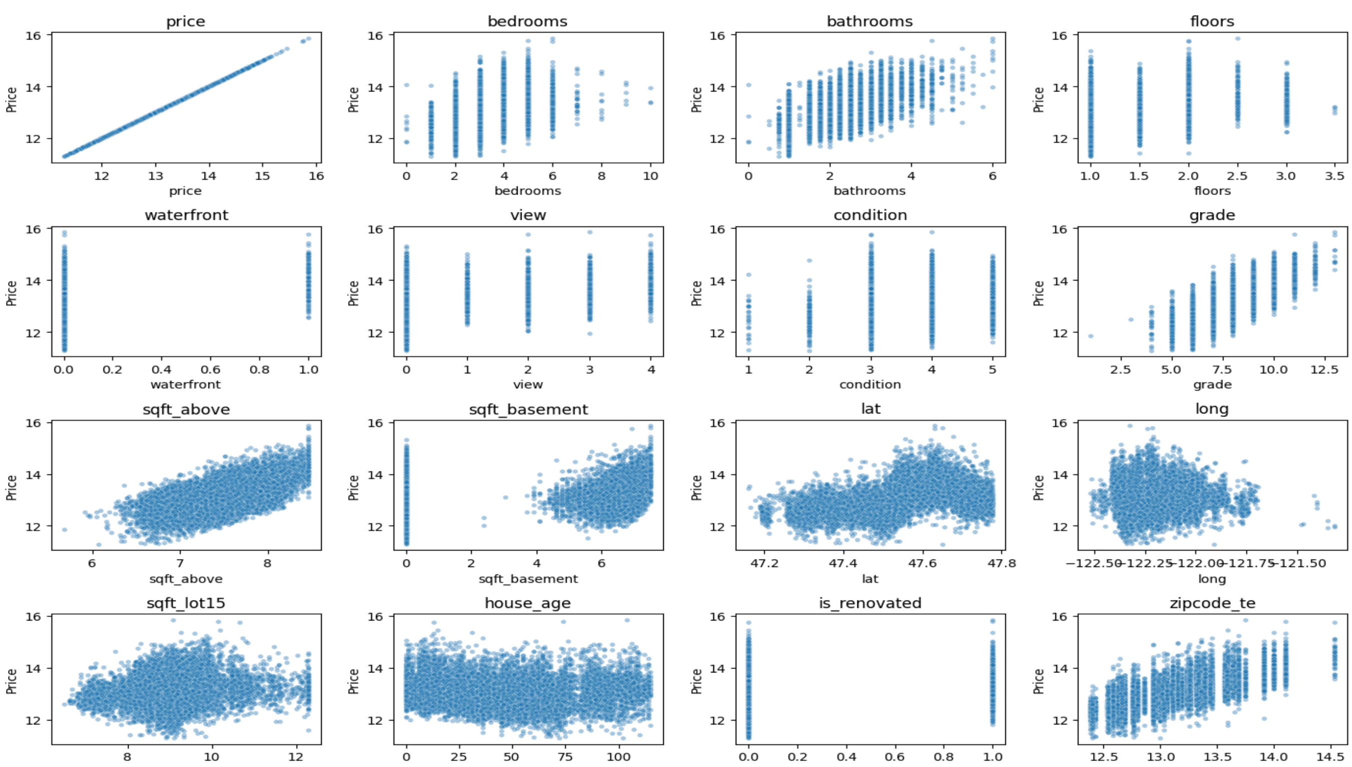
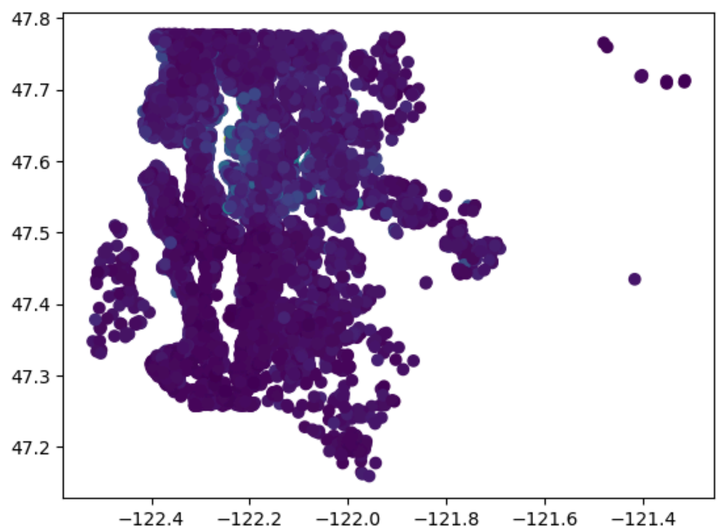

price
bedrooms
bathrooms
floors
waterfront
view
condition
grade
sqft_above
sqft_basement
lat
long
sqft_lot15
house_age
is_renovated
zipcode_te
Price

Figure 5: Geospatial Scatter Plot of House Prices (Latitude vs Longitude)

Geospatial scatter plot showing how house prices vary spatially, revealing clear location-based clustering of high- and low-priced properties.



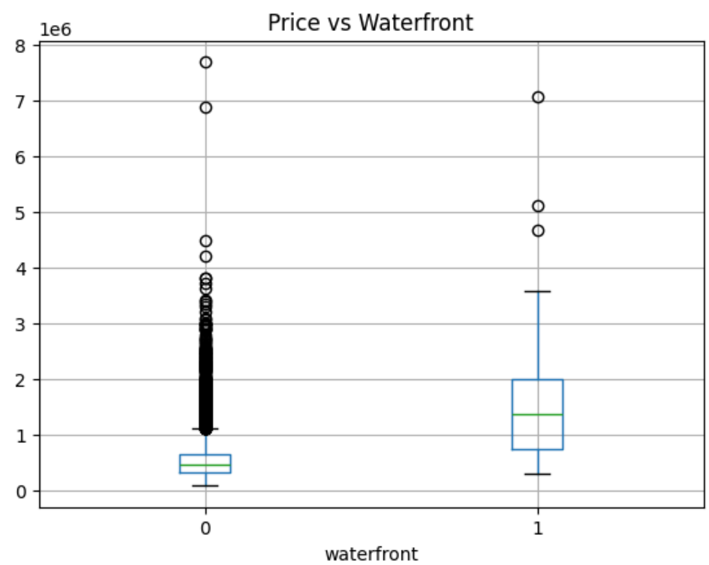
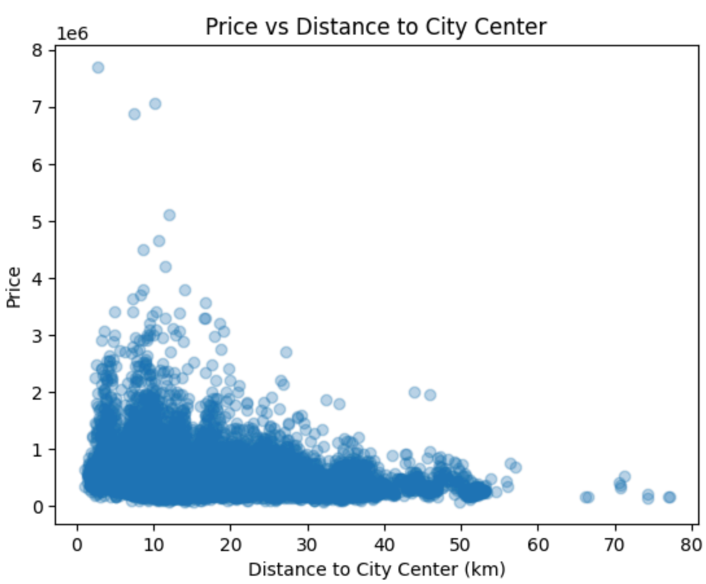


Figure 6: House Price Distribution by Waterfront Status

Comparison of house prices for waterfront and non-waterfront properties, showing a significant premium for waterfront locations



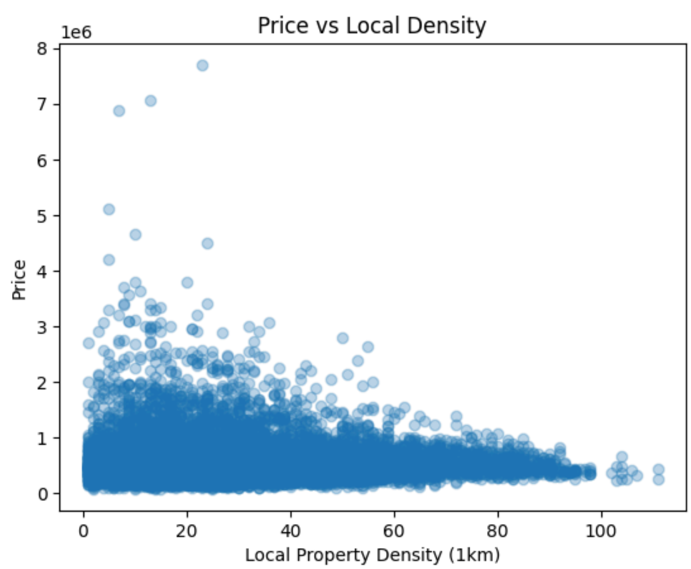


Figure 7: Satellite Image Samples from ESRI (Train and Test Data)- *Example satellite images downloaded using the ESRI, demonstrating consistent image quality and coverage across training and testing datasets.*

Figure 8: Price Buckets Distribution- *Visualization of houses grouped into discrete price buckets to analyze patterns across different price segments.*

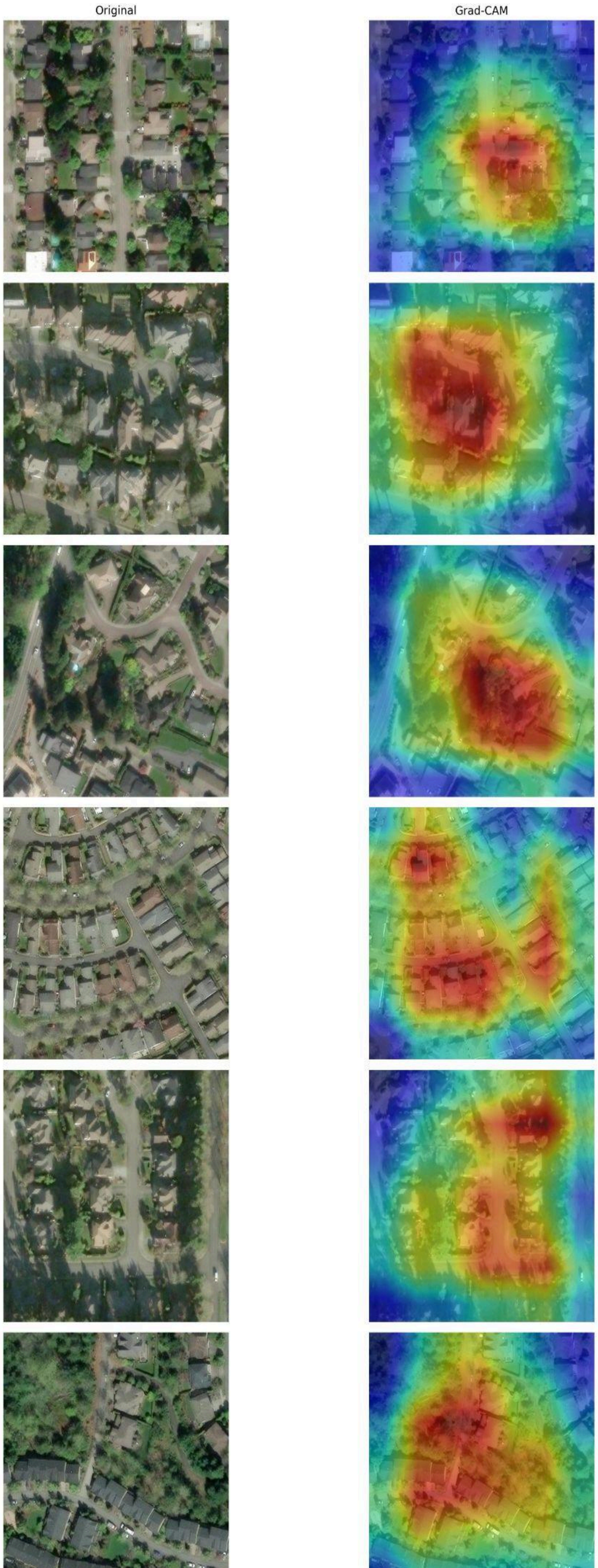
Figure 9: Sample Test Satellite Images per Price Bucket

Three representative satellite images from each price bucket, illustrating how surrounding infrastructure and land usage vary with price in dataset.

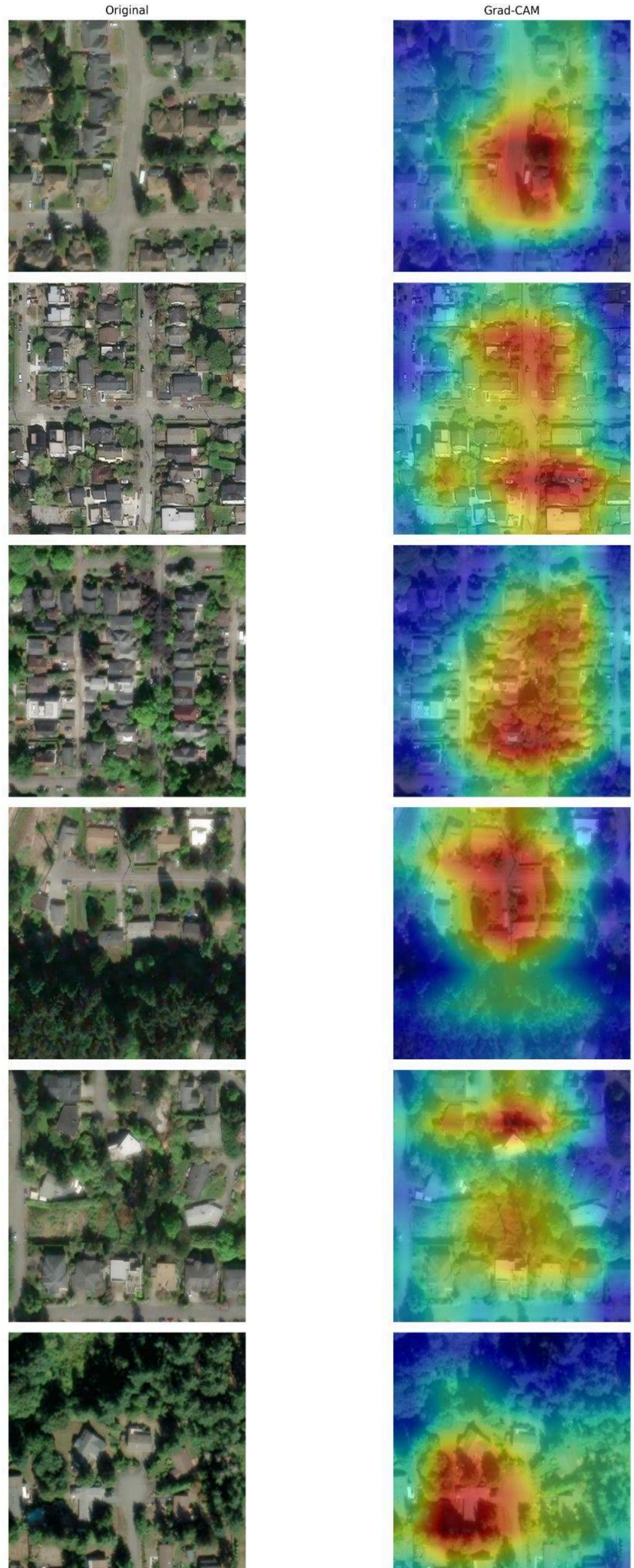
3. Financial/Visual Insights : Grad-CAM Visualization on Test Data -

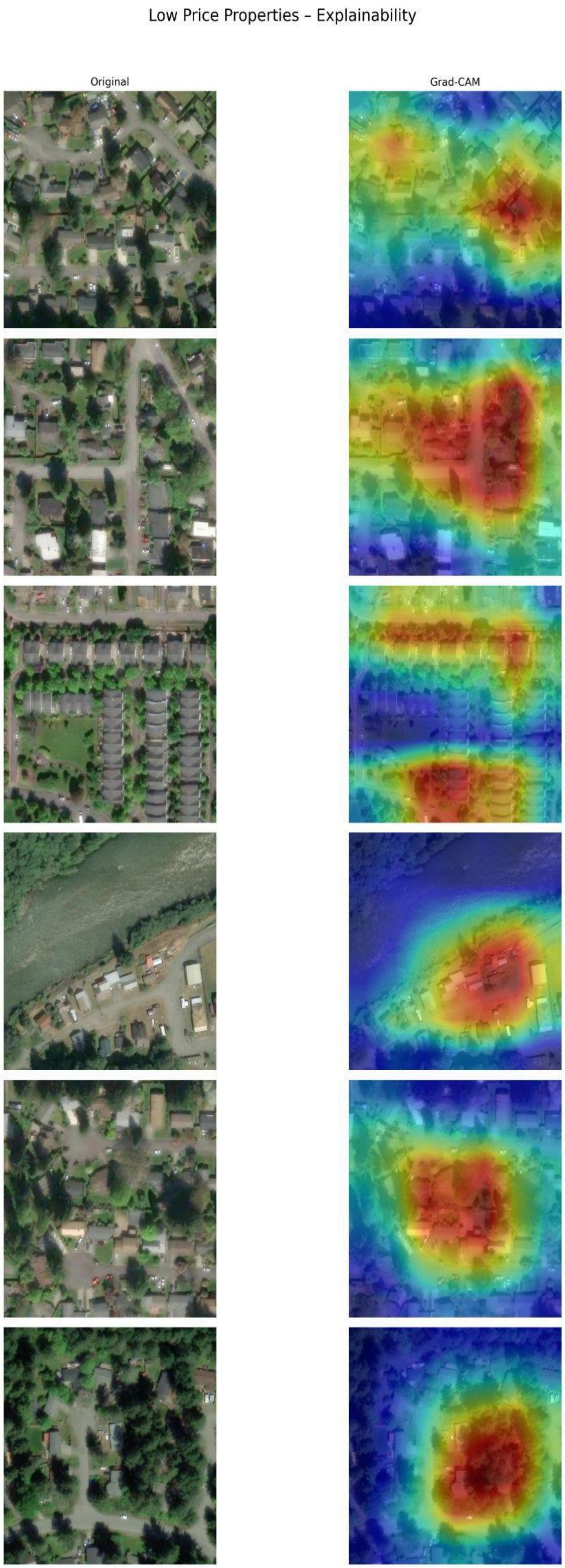
Grad-CAM heatmaps showing which regions of satellite images most influenced the model's price predictions, improving interpretability of the multimodal model.

High Price Properties – Explainability



Mid Price Properties – Explainability





Grad-CAM Based Comparison of Price Categories

1. Low Price Properties

- ➢ *Sparse or irregular housing*
- ➢ *Small, congested structures*
- ➢ *Poor road connectivity*
- ➢ *Limited or no greenery*
- ➢ *Focus on edges, roofs, random area*
- ➢ *Indicates lower visual quality signals*

2. Mid Price Properties

- ➢ *Moderately dense housing*
- ➢ *Semi-planned layouts*
- ➢ *Some greenery present*
- ➢ *Roads partially highlighted*
- ➢ *Mixed activation patterns*
- ➢ *Balance between buildings & surroundings*
- ➢ *Transitional urban development*

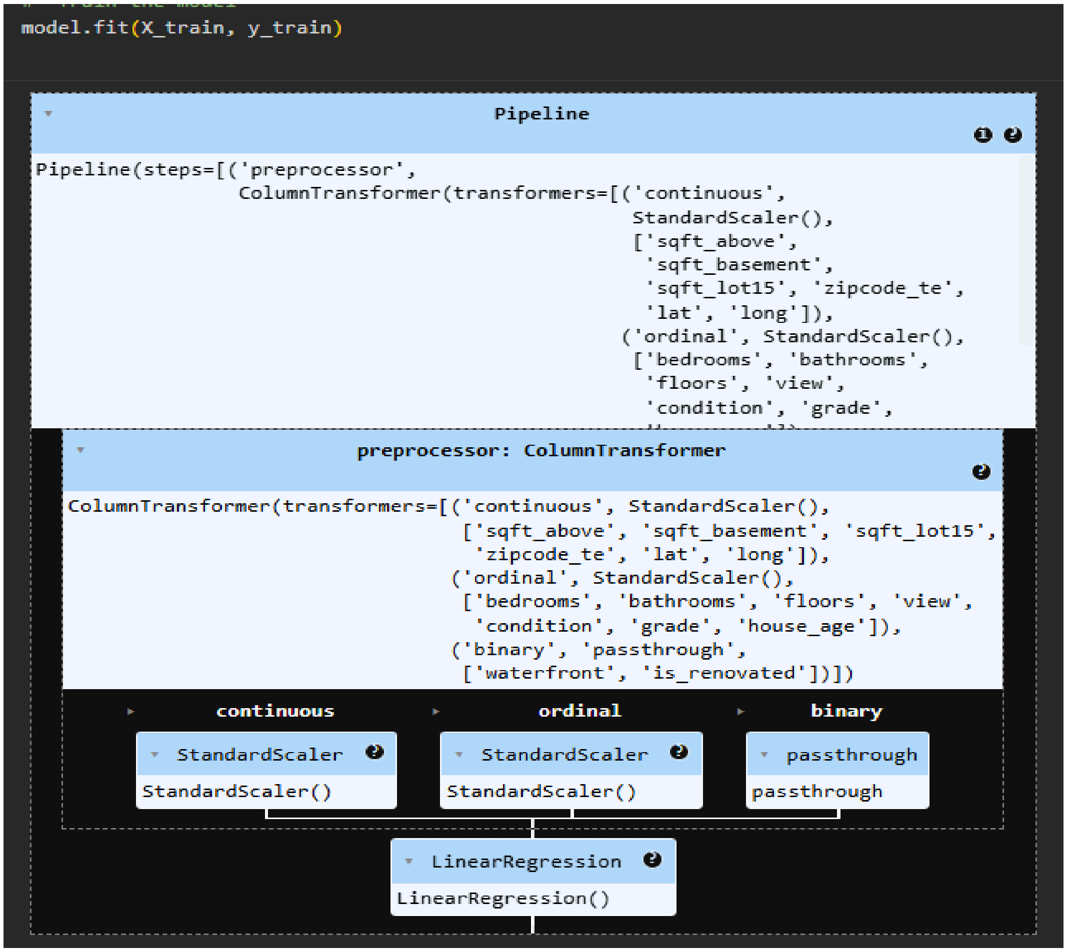
3. High Price Properties

- ➢ *Dense but well-organized housing*
- ➢ *Larger, clearly separated houses*
- ➢ *Strong road network structure*
- ➢ *High vegetation coverage*
- ➢ *Strong, centralized activations*
- ➢ *Focus on neighborhood core*
- ➢ *Clear urban planning signals*

Key Visual Differentiators
Planning quality ↑ → Price ↑
Green cover ↑ → Price ↑
Activation clarity ↑ → Model confidence ↑
Random patterns → Lower valuation

4. Architecture Diagram

Figure 1: Tabular Data Preprocessing and Regression Pipeline



➢ Tabular features are grouped into continuous, ordinal, and binary categories.
➢ Continuous and ordinal features are standardized using StandardScaler.
➢ The processed features are passed to a regression model for price prediction.

Figure 2: (Tabular + Satellite image) The model uses a dual-input architecture where satellite images are processed using a CNN and tabular property attributes are processed using an MLP. The extracted feature vectors from both modalities are concatenated and passed through fully connected layers to predict property price.

Satellite Image

|

▼

CNN (ResNet-18)

─ Feature Extractor

─ Output: Image Embeddings (e.g. 512)

|

▼

Image Feature Vector

\

─► Feature Concatenation ──► Fully Connected Layer ──► Price Prediction

/

Tabular Features

|

▼

MLP (Dense Network)

─ Input: 15 features

─ Output: Tabular Embeddings (e.g. 64)

Multimodal House Price Prediction Model

1. Data Handling

- Each house is represented using two data sources:
    - Satellite image of the property and surrounding area
    - Tabular features such as size, location, and house attributes

- A custom dataset loads the image, tabular data, and price together for each sample.

2. Image Processing (CNN Branch)

- Satellite images are resized to 224 × 224 and normalized.
- A pretrained ResNet-18 model is used to extract visual features.
- The final classification layer is removed so the model outputs 512 visual features.
- CNN weights are frozen to reduce training time and prevent overfitting.

3. Tabular Data Processing (MLP Branch)

- Tabular features are passed through a fully connected neural network:
  - Input features → 128 neurons → 64 neurons
- This network learns meaningful representations from numerical house attributes.

4. Feature Fusion

- Image features (512) and tabular features (64) are concatenated.
- The combined feature vector represents both visual context and property details.

5. Price Prediction

- The fused features are passed to a regression head.
- The model outputs a single continuous value representing the predicted house price.

6. Model Training

- The model is trained using Mean Squared Error (RMSE) loss.
- Performance is monitored using R^2 score on training and validation data.
- Early stopping is applied to avoid overfitting.
- The best-performing model is saved automatically.

5. Results and Performance Comparison

Model Type	Input Features	R^2 Score	RMSE	Remarks
Random Forest	Tabular only	**0.8780**	**0.1786**	Best predictive performance on tabular data
Tabular Neural Network (MLP)	Tabular only	**0.8624**	**0.1959**	Learns feature interactions well
Combined CNN + MLP	Tabular + Satellite Images	**0.8587**	**0.1915**	Adds spatial context & visual explainability

Key Observations

- Tree-based models perform best on structured tabular data.
- MLP provides competitive performance with neural feature learning.
- Combined model:
    - Slightly higher error due to frozen CNN and limited fine-tuning
    - Captures **neighborhood-level spatial patterns**
    - Enables **Grad-CAM visual explainability**, which tabular models lack

Early Stopping Statement

Early stopping was triggered at Epoch 38 to prevent overfitting, selecting the model with the lowest validation RMSE.

Conclusion

While tabular-only models achieve lower RMSE, the combined CNN–MLP model enhances interpretability by leveraging satellite imagery without retraining from scratch.