Satellite Imagery-Based Property Valuation

Submitted by:
Aditi B R (23113009)

Overview

Data Preprocessing	Image feature extraction	Baseline Model	Multimodal
• Understood feature distributions and their relationships with the target variable • Applied Standardization on numerical features. • Target variable was prepared for regression modeling.	• Retrieved satellite images using coordinates. • Preprocessed images to fixed size and normalization. • Used pretrained ResNet50 to extract 2048-dimensional image embeddings.	• Trained regression models using tabular data only. • Evaluated Linear Regression, Random Forest, XGBoost, LightGBM, and HistGradientBoosting.	• Combined image embeddings with tabular features. • Trained identical regressors for fair comparison. • Applied Grad-CAM for visual interpretability of CNN features.

Compare Performance
(R^2 and MSE)

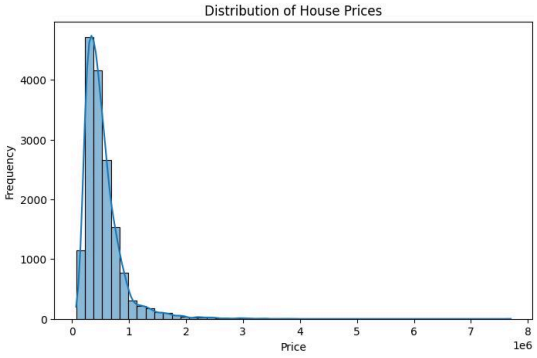
Exploratory data analysis

Basic Understanding of data

- The dataset consists of 21 features, including 12 continuous variables, 4 categorical variables, 3 discrete variables, along with date and ID fields.
- The dataset contains 16,209 observations, of which 16,110 have unique property IDs.
- No missing values were observed across any features, ensuring data completeness.

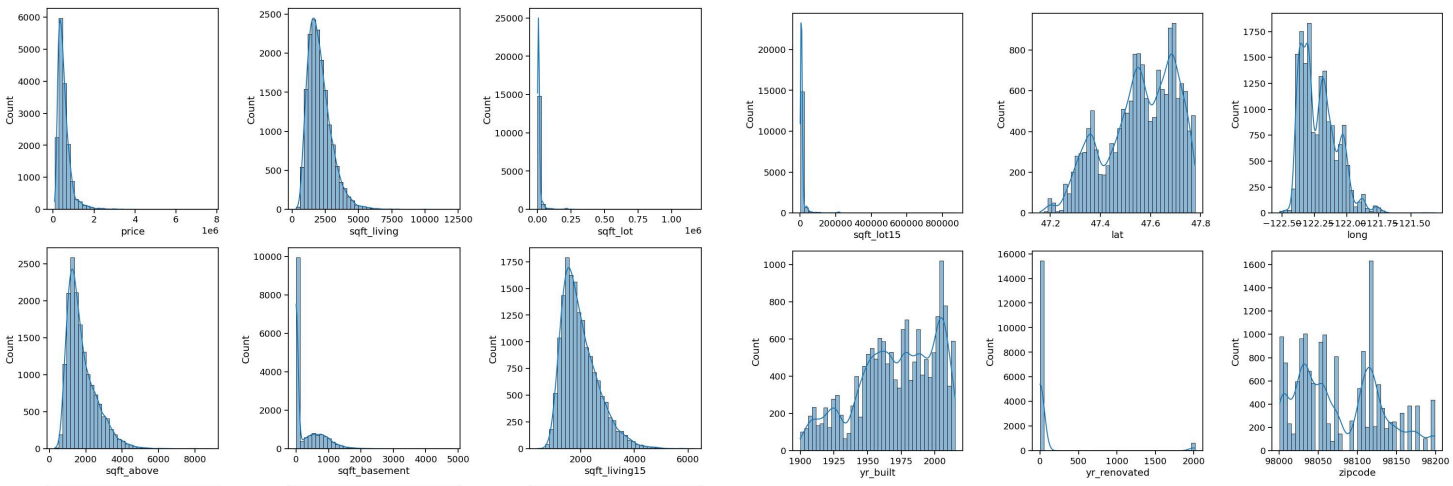
Univariate analysis

1. Price



- The price distribution is right-skewed, indicating a small number of very high-valued properties.
- The mean price is approximately 5,38,000.
- There are many outliers on the higher price end. These outliers represent high-value properties and are not necessarily errors.

2. Numerical Features



1. Size-related features (sqft_living, sqft_above, sqft_living15)
- Distributions are right-skewed, with most houses having moderate sizes.
- Few very large houses create long tails. Indicates size strongly varies but extreme values are rare.

2. Lot size features (sqft_lot, sqft_lot15)
- Extremely right-skewed with heavy tails. Most properties have small lots, while a few have very large land areas.

3. Basement area (sqft_basement)
- Large spike at zero → many houses do not have basements. Remaining values are right-skewed.

4. Location features (lat, long, zipcode)
- Multi-modal distributions indicate geographical clustering. Reflects different neighborhoods and urban zones.

5. Time-related features (yr_built, yr_renovated)
- yr_built shows increasing density for newer houses. yr_renovated has a spike at zero most houses were never renovated.

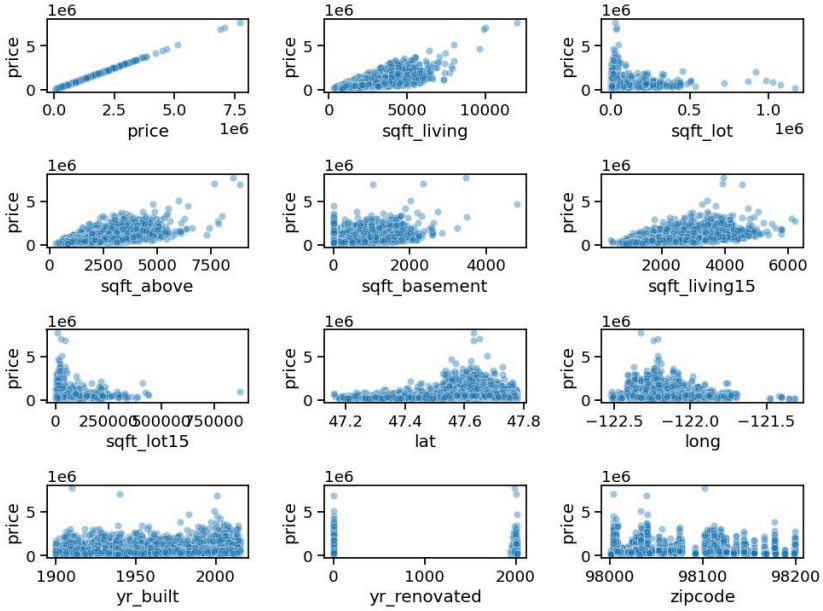
Scatter Plot Analysis (Price vs Features)

Strong positive relationships
- sqft_living, sqft_above, sqft_living15 show a clear upward trend with price.
- Larger living areas → higher property value.

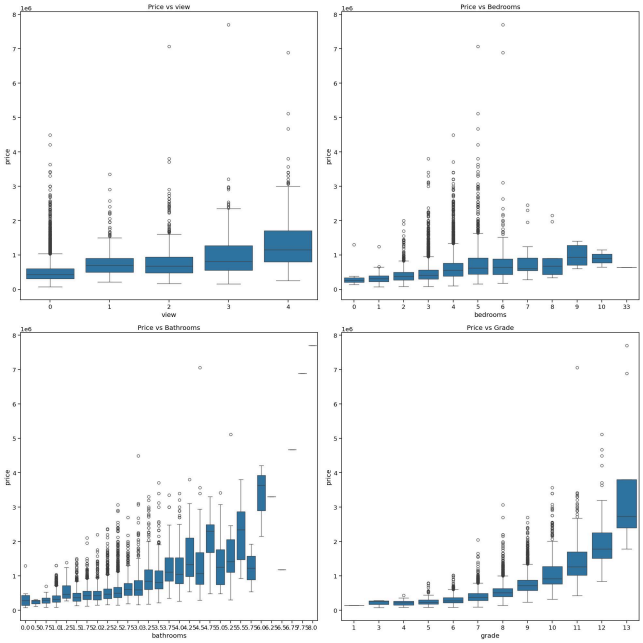
Weak or noisy relationships
- sqft_lot, sqft_lot15 show high spread and weak correlation with price.
- Large land area alone does not guarantee higher price.

Location effects
- Latitude and longitude show clustering, indicating location strongly influences price.
- Certain geographic bands correspond to higher-priced properties.

Temporal features
- yr_built shows mild trend: newer homes tend to be slightly more expensive.
- yr_renovated has weak direct correlation; renovation impact varies.



Boxplot Analysis (Price vs Features)



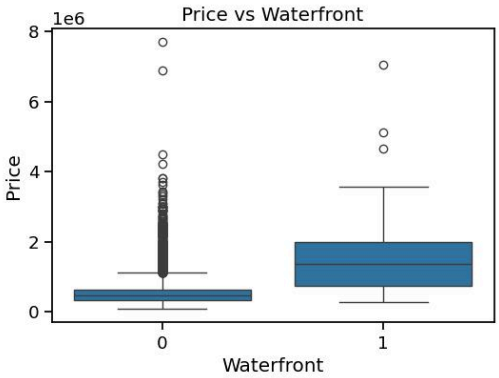
1. Price vs View
- Median house price increases consistently with better view ratings.
- Properties with view = 4 have significantly higher prices and wider spread.
- View quality is a strong value-adding categorical feature.

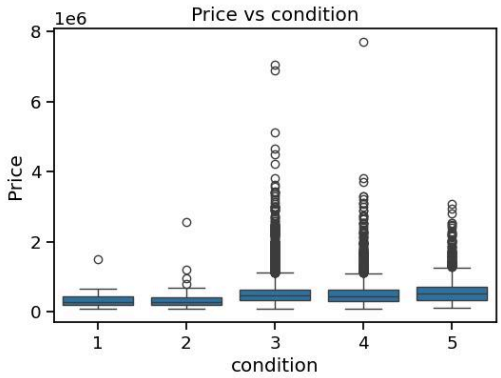
2. Price vs Bedrooms
- Price generally increases from 1 to ~4–5 bedrooms.
- After ~5 bedrooms, gains plateau or become noisy, with many outliers.
- Bedroom count alone is not a strong linear driver of price.

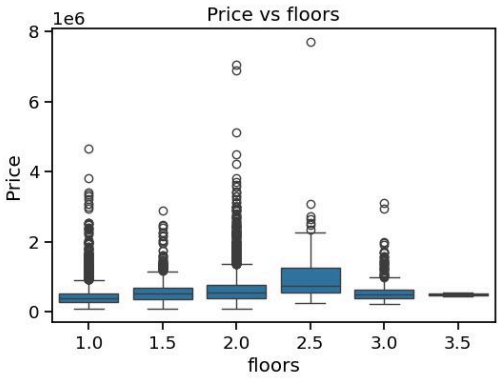
3. Price vs Bathrooms
- Clear monotonic increase in price as number of bathrooms increases.
- Homes with 3+ bathrooms command substantially higher prices.
- Bathrooms are a stronger predictor than bedrooms.

4. Price vs Grade
- Strongest relationship among all variables.
- Higher construction/finish grade leads to exponential increase in price.
- Grade captures overall quality, making it a key driver of valuation.

Boxplots of categorical variables







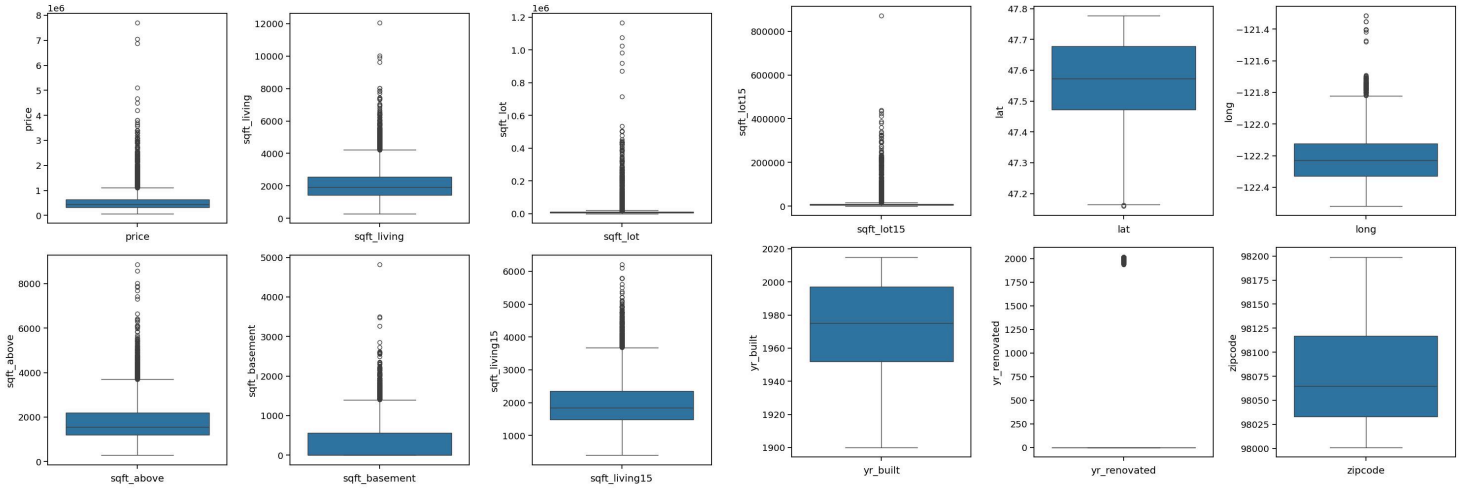
1. Price vs Waterfront
   - Waterfront properties are significantly more expensive than non-waterfront homes.
   - Median price for waterfront homes is much higher, with more high-value outliers.
   - Indicates strong positive impact of waterfront location on house value.

2. Price vs Condition
   - House price generally increases with better condition, but the effect is moderate.
   - Large overlap across condition levels → condition alone does not fully explain price.
   - High-priced outliers exist at most condition levels, suggesting other factors matter more.

- 3. Price vs Floors
   - Homes with 2–2.5 floors tend to have higher median prices.
   - Very high floor counts are rare and show less consistent pricing.
   - Relationship is non-linear: more floors help up to a point, then benefits flatten.

Overall Insight
- Waterfront is a strong value driver.
- Condition and floors influence price but are secondary compared to location and size.

Outlier Analysis



1. Price
   - Strong presence of high-end outliers.
   - Majority of properties fall in a narrow lower range.

Implication: Luxury homes significantly increase variance.

2. Size-related features (sqft_living, sqft_above, sqft_living15)
   - Clear upper-end outliers representing very large houses.
   - Central mass remains compact.

Implication: Extreme sizes exist but are rare.

3. Lot size features (sqft_lot, sqft_lot15)
   - Extremely heavy-tailed distributions.
   - Many extreme land-size outliers.

Implication: Lot size varies drastically across properties.

4. Basement area (sqft_basement)
   - Large number of zeros with some extreme positive values.
   - Outliers correspond to unusually large basements.

Implication: Basement contributes as a sparse feature.

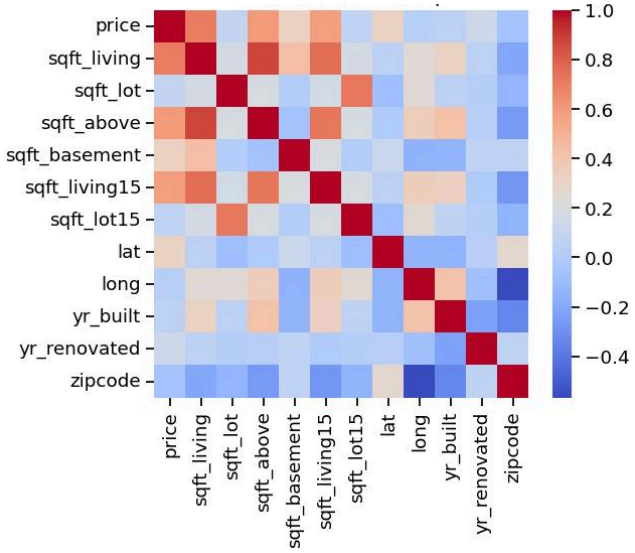
5. Location features (lat, long, zipcode)
   - Very few mild outliers.
   - Most values tightly clustered.

Implication: Geographic data is stable and reliable.

6. Year features (yr_built, yr_renovated)
   - yr_built shows moderate spread with few old-property outliers.
   - yr_renovated dominated by zeros → very few renovated homes.

Implication: Renovation is rare but potentially impactful.

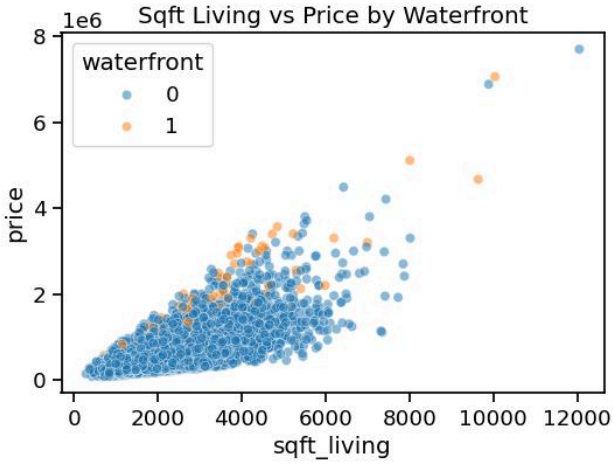
Correlation Heatmap

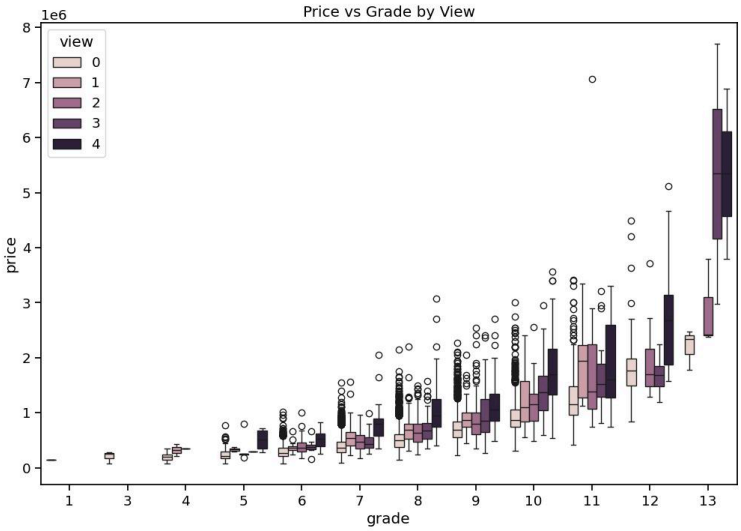


- As observed earlier, sqft_living, sqft_above, sqft_basement, sqft_living15, and latitude show strong influence on house price.
- sqft_living is highly correlated with sqft_above and sqft_basement, since sqft_living = sqft_above + sqft_basement, leading to multicollinearity.
- To reduce redundancy, sqft_basement is removed, as it has a comparatively weaker relationship with price.

Multivariate Analysis

- Living area is the primary driver of house prices, with prices increasing strongly as sqft_living increases for both waterfront and non-waterfront properties.
- Waterfront properties command a clear price premium for the same living area, indicating an added geospatial value beyond size alone.
- Price dispersion increases for larger homes, suggesting that location and premium features become more influential at higher square footage levels.

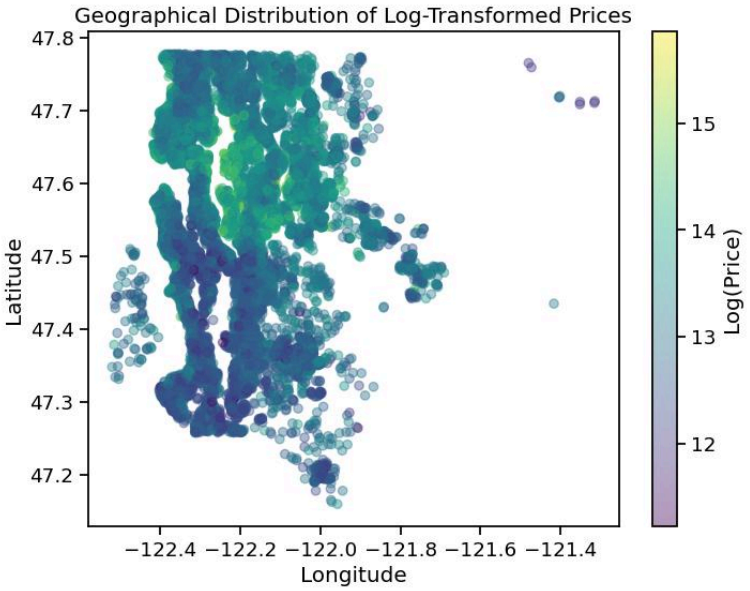


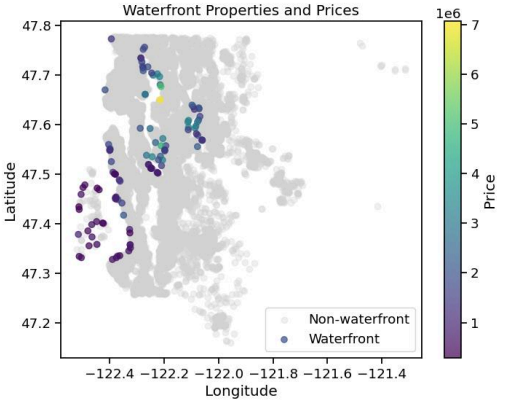


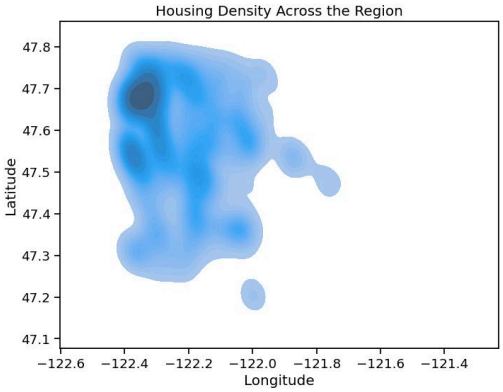
- For the same grade, properties with a better view have higher median prices.
- Higher-grade properties often have higher prices even with poorer views, compared to lower-grade properties with better views.
- The impact of view becomes stronger at higher grade levels.
- Price variability increases with grade, indicating greater diversity among high-end properties.
- Overall, grade is the main driver of price, while view adds an additional premium, especially for premium homes.

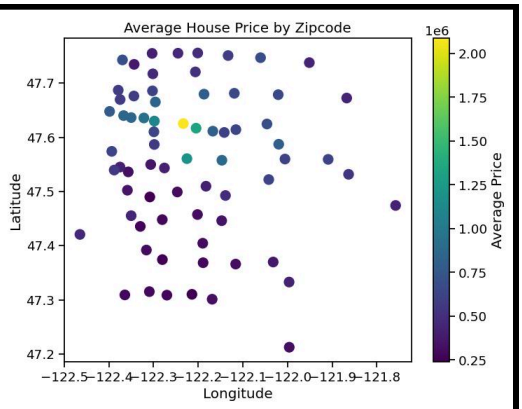
Geospatial Analysis

- House prices exhibit clear spatial clustering, indicating that prices are strongly influenced by geographic location.
- Nearby properties tend to have similar prices, reflecting neighborhood-level effects.
- High-priced properties are concentrated in specific pockets, while lower-priced homes occupy broader regions.
- These patterns justify the inclusion of spatial features such as latitude–longitude interactions or location-based clustering in the model.









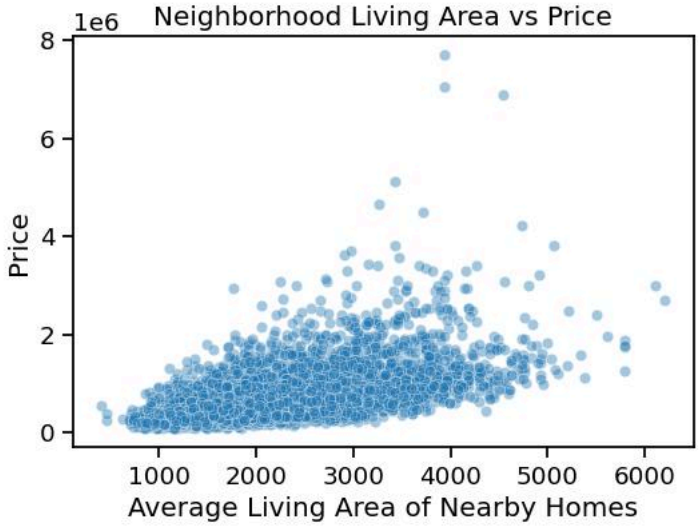
Waterfront houses are clustered near water bodies and generally show higher prices than non-waterfront homes, indicating a strong location-based price premium driven by environmental and visual appeal.

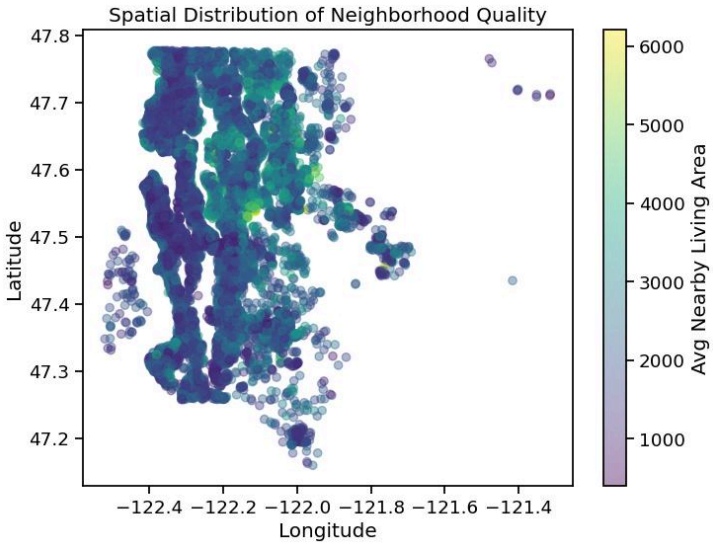
Areas with higher housing density represent more developed and urbanized regions, while low-density areas indicate suburban or less developed zones. These patterns influence demand, accessibility, and pricing.

Average prices vary significantly across zipcodes, capturing neighborhood-level effects such as amenities, infrastructure, and socio-economic factors that impact property values.

Overall Relevance to the Project: These spatial patterns highlight the importance of location and surroundings in property valuation and justify using satellite imagery to capture environmental cues (e.g., proximity to water, greenery, urban density) that tabular data alone may miss.

Neighborhood Context and Its Influence on Housing Prices





1. Spatial Distribution of Neighborhood Quality
- Higher average nearby living areas cluster in specific locations, indicating premium neighborhoods.
- Neighborhood quality shows strong spatial patterns rather than being evenly distributed.

2. Neighborhood Living Area vs Price
- Prices generally increase with the average size of nearby homes.
- Neighborhood context amplifies property value beyond individual house features.

Key Takeaway
- House prices depend on both local neighborhood characteristics and individual property attributes.

Feature Engineering

1. Removal of Multicollinear Features:
   The feature sqft_basement was removed due to high correlation with sqft_living and sqft_above, which could introduce multicollinearity and negatively impact model stability.
2. Handling of Date Feature:
   The date and id variables were discarded as it did not provide meaningful predictive value in its raw form and could introduce noise without proper temporal feature extraction.
3. Feature Scaling:
   Numerical features were standardized using StandardScaler to ensure all variables operate on a comparable scale.
4. Prevention of Data Leakage:
   The scaler was fitted only on the training dataset and then applied to validation/test and prediction datasets to maintain strict separation between training and evaluation data.

Financial / Visual Insights from Grad-CAM

Grad-CAM visualizations were generated for three representative price levels—low (20th percentile), moderate (50th percentile), and high (80th percentile)—to analyze how the CNN's visual attention varies across different property value ranges. Grad-CAM highlights the regions of an image that most strongly contribute to the model's price prediction.

ID: 301401410 | Price: 298,000

1. Low-priced property (Price ≈ 298,000)
   - Activation is weak and scattered
   - Focus is mainly on:
     - Roof structures only
     - Dense housing blocks
     - Limited greenery
   - Little attention to surrounding amenities

Inference:
Lower prices are associated with high density, limited open space, and fewer visible natural or premium features.

2. Mid-priced property (Price ≈ 450,000)
   - Moderate activation around:
     - Houses and road layout
     - Some greenery and trees
   - Mixed focus between built structures and surroundings

Inference:
Mid-range prices reflect a balance between built area and neighborhood quality, with moderate greenery contributing positively.

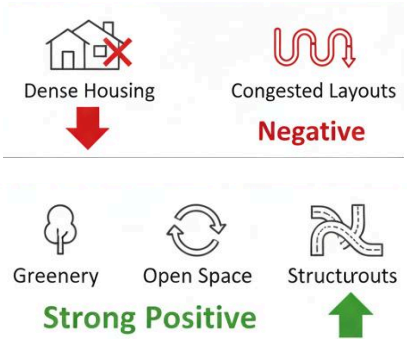
ID: 3629960550 | Price: 450,000

ID: 8901500178 | Price: 700,000

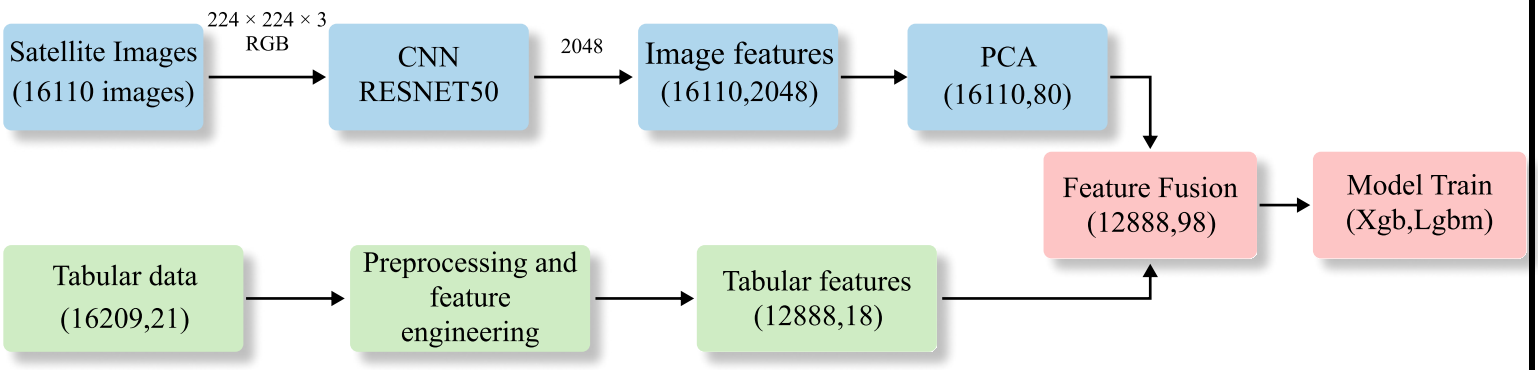
3. High-priced property (Price ≈ 700,000)
   - Strong, concentrated activation on:
     - Green spaces and trees
     - Spacious layouts
     - Well-defined roads and open areas
   - Less emphasis on dense rooftops

Inference:
Higher prices are driven by environmental quality, such as:
   - More greenery
   - Lower density
   - Better spatial organization

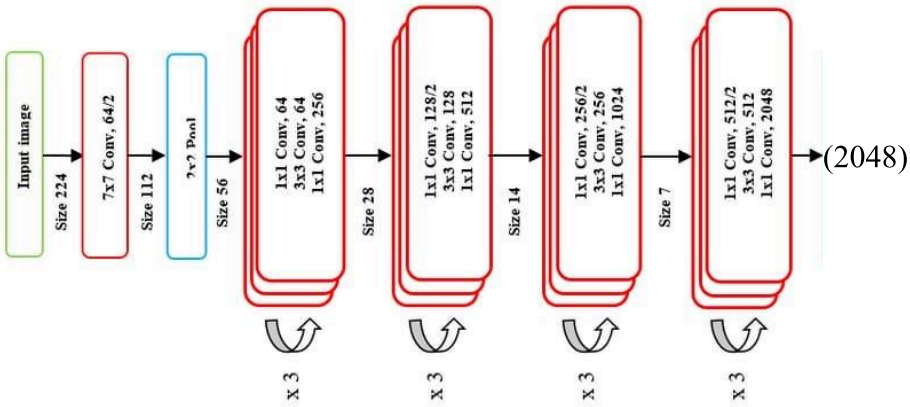
Overall Conclusion
Grad-CAM analysis shows that the CNN does not rely only on house roofs, but strongly considers neighborhood-level visual cues. Properties surrounded by greenery, open space, and structured layouts are consistently associated with higher predicted prices, while dense and congested areas correspond to lower values.
This demonstrates that satellite imagery provides meaningful financial insight into real-estate valuation, complementing traditional tabular features.



Architecture Diagram



CNN Architecture



Source: https://miro.medium.com/v2/resize:fit:1400/1*rPktw9-nz-dy9CFcddMBdQ.jpeg

Architecture overview

- Input: 224 × 224 × 3 RGB satellite image
- Backbone: ResNet-50
  - Convolution + BatchNorm + ReLU
  - Residual blocks (skip connections)
- Global Average Pooling
- Fully Connected (FC) layer removed
- Output: 2048-dimensional feature vector

Results

Model	Train R²	Test R²	MSE	RMSE
XGBoost	0.986701	0.910395	1.095344e+10	104658.679678
LightGBM	0.913385	0.897165	1.257074e+10	112119.318202
HistGradientBoosting	0.908778	0.893164	1.305984e+10	114279.676821
Random Forest	0.969950	0.881877	1.443952e+10	120164.558631
Linear Regression	0.696012	0.704371	3.613817e+10	190100.429108

Tabular Data Only

Model	Train R²	Test R²	MSE	RMSE
XGBoost	0.995793	0.905618	1.153740e+10	107412.293673
LightGBM	0.929234	0.895257	1.280400e+10	113154.753193
HistGradientBoosting	0.916228	0.885140	1.404064e+10	118493.198883
Random Forest	0.971733	0.863885	1.663891e+10	128991.911190
Linear Regression	0.719395	0.726388	3.344682e+10	182884.709582

Tabular + Satellite Images

Model Selection

- XGBoost, despite its higher accuracy, shows clear signs of overfitting, as the training accuracy is nearly one and there is a larger gap between the training and test R2 scores, making it less reliable for deployment.
- LightGBM offers the best generalization trade-off, with consistent performance and lower variance between training and test sets.
- Therefore, LightGBM was selected as the final model due to its robustness and stability.

Model Performance Comparison and Interpretation

Baseline (Tabular-only) Models

- Tree-based ensemble models (XGBoost, LightGBM, Random Forest) significantly outperform Linear Regression, indicating strong non-linear relationships in the data.
- XGBoost achieves the highest train and test R^2, but the large gap between them suggests overfitting.
- LightGBM shows a better train–test balance, indicating more stable generalization.

Multimodal (Tabular + Image Features) Models

- Incorporating satellite image features improves training performance across all models, confirming that image features contain additional predictive information.
- However, test performance slightly decreases for most models, especially for highly flexible learners such as XGBoost and Random Forest.
- This indicates that while image features add signal, they also introduce high-dimensional noise, increasing the risk of overfitting.