CDC X Yhills OPEN PROJECTS

Satellite Imagery-Based Property Valuation

-By Ansul
(23113029)

1. Overview

This project focuses on building a robust property valuation framework by integrating structured housing attributes with unstructured satellite imagery. The objective is to predict residential property prices using a multimodal regression approach that combines traditional real estate features with visual neighborhood context extracted from satellite images. The modeling pipeline was designed to remain interpretable, scalable, and computationally efficient, leading to the final selection of gradient-boosted decision trees (XGBoost) as the core regression model.

The workflow consists of four major stages: (i) extensive exploratory data analysis and feature engineering on tabular housing data, (ii) programmatic acquisition and preprocessing of satellite images using latitude and longitude coordinates, (iii) extraction of high-dimensional visual embeddings from a pretrained convolutional neural network followed by dimensionality reduction, and (iv) comparative modeling using XGBoost on tabular-only data and on fused tabular + image features. Model explainability is addressed through feature importance analysis and Grad-CAM visualizations applied to the image encoder.

2. Dataset & Preprocessing

The Tabular Dataset consists of 16209 house prices with 21 columns which are as follows:
['id', 'date', 'price', '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'].Columns like view,condition,grade and waterfront are categorical in nature.

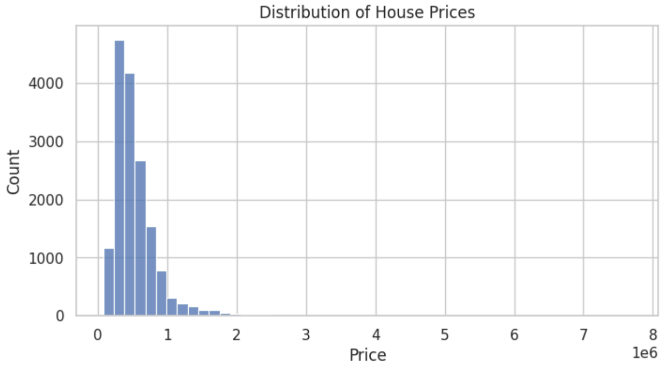
The tabular dataset, after cleaning and feature engineering, contains **16,209 training samples with 23 columns** and **5,404 test samples with 21 columns**. After aligning with available satellite imagery and removing rows without valid image embeddings, the final modeling dataset consists of **16,110 training rows** and **5,396 test rows**. This alignment step is critical to ensure consistency across modalities.

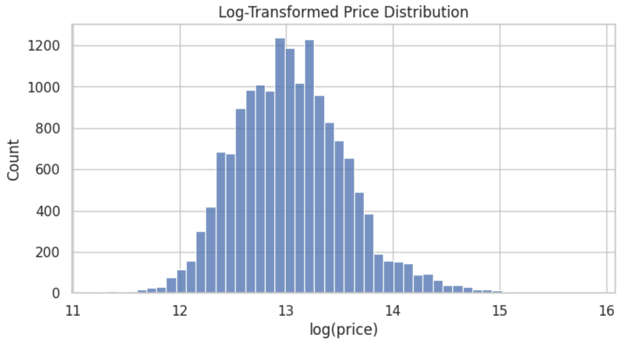
Satellite imagery was downloaded for each property using its geographic coordinates. Each image was resized to **224×224 RGB** and passed through a pretrained

ResNet-18 model. The output of the final convolutional block yields a **512-dimensional embedding** per image. These embeddings were stored as CSV files and later merged with the tabular dataset using the unique property identifier. Principal Component Analysis (PCA) was applied to the image embeddings, reducing the dimensionality from 512 to **150 components**, while retaining approximately **84.9% of the total variance**. This step mitigates overfitting and reduces computational complexity.

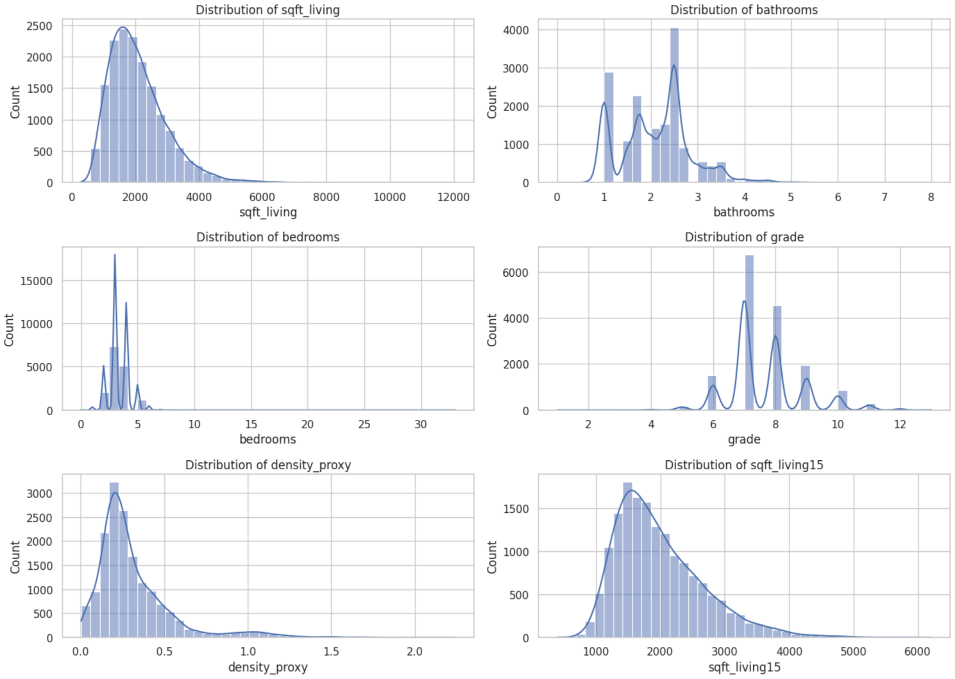
3. Exploratory Data Analysis

Univariate Analysis





The Price(Target Variable) is highly right skewed due to some high priced luxurious properties owing to which it was log-tranformed for the model training phase and finally during prediction the predictions were exponentialised.



1.sqft_living

Distribution is **positively skewed** with a long right tail, indicating presence of large luxury houses.
Majority of observations lie between **1000–2500 sqft**, representing typical residential homes.
Outliers beyond **5000 sqft** may disproportionately influence mean-based models.

2.bathrooms

Discrete, **multi-modal distribution**, with peaks around **1, 2, and 2.5 bathrooms**.
Right-skewed with sparse high-end values (>4 bathrooms).
Indicates categorical-like behavior despite being numeric.
Small decimal increments suggest engineered or averaged values.

3.bedrooms

Highly **concentrated around 3–4 bedrooms**, showing low variance.
Extreme right-side outliers (≥10 bedrooms) are rare and possibly anomalous.
Distribution is **right-skewed and discrete**.
Outliers may need capping or validation during preprocessing.

4.grade

Ordinal variable with **distinct peaks**, reflecting discrete quality ratings.
Most houses fall in **grades 6–8**, indicating mid-range construction quality.
Slight right skew toward higher grades.
Should be treated as an **ordered categorical feature**, not purely continuous.

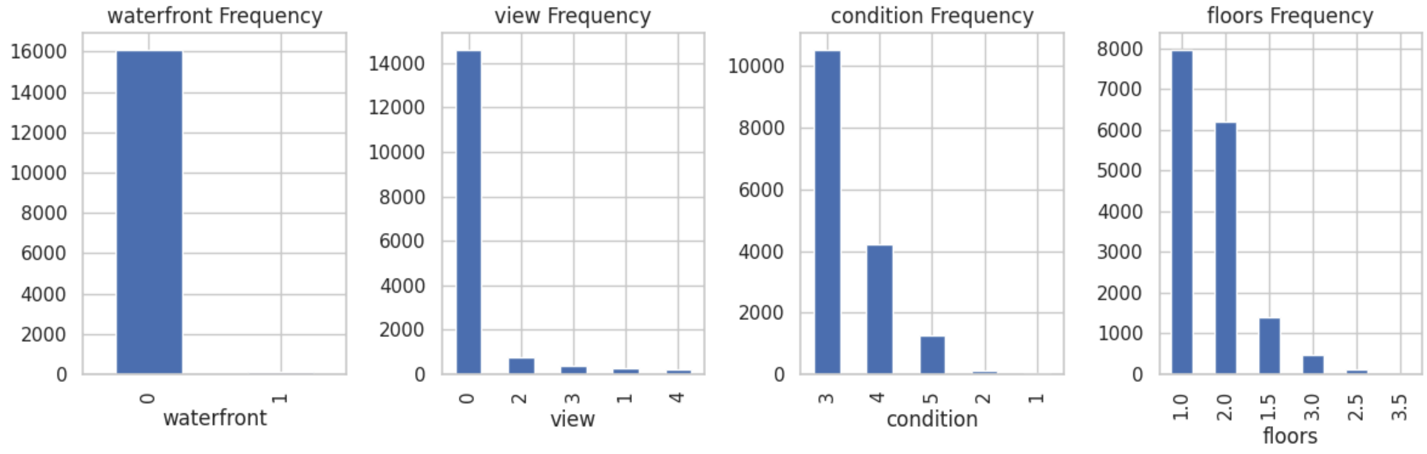
5.density_proxy

Strong **positive skew**, with most values clustered near zero.
Long tail indicates areas of high local density.
Non-normal distribution suggests **log or Box-Cox transformation** may improve model performance.
Likely sensitive to outliers and noise.

6.sqft_living

Similar shape to `sqft_living`, indicating **spatial correlation** with neighboring houses.
Positively skewed with most values between **1000–3000 sqft**.
Fewer extreme outliers compared to `sqft_living`, implying smoother neighborhood averages.
Useful contextual feature for capturing **neighborhood effects**.

Categorical Variable Analysis:



1.waterfront

Binary variable with **severe class imbalance**, where the majority of properties are non-waterfront.
Waterfront homes constitute a **very small fraction** of the dataset.
Indicates high potential predictive power despite low frequency.

2.view

Discrete ordinal variable with values ranging from **0 to 4**.
Highly **right-skewed**, with most properties having **no special view (0)**.
Higher view ratings are increasingly rare, indicating exclusivity.

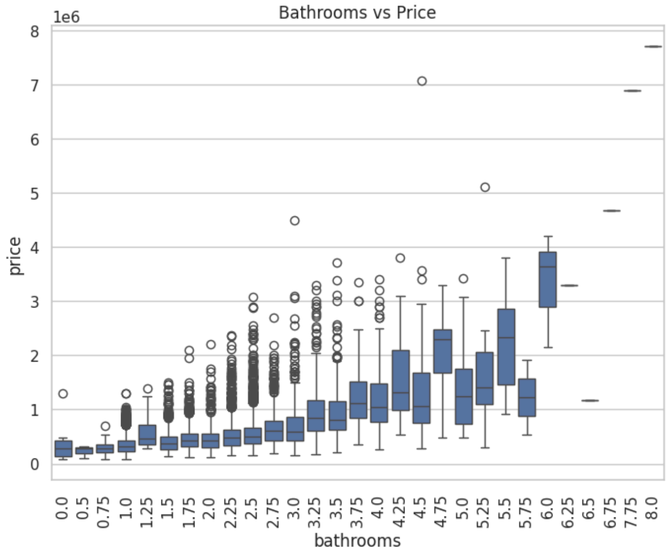
3.condition

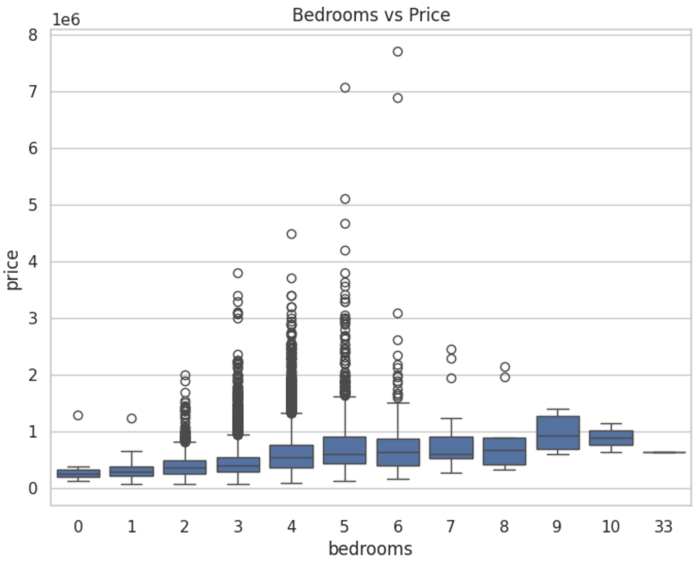
Ordinal variable with most observations clustered around **condition = 3 (average)**.
Gradual decline in frequency toward higher condition ratings.
Very few properties in poor (1–2) or excellent (5) condition.
Suggests limited variance but strong interpretability as a quality indicator.

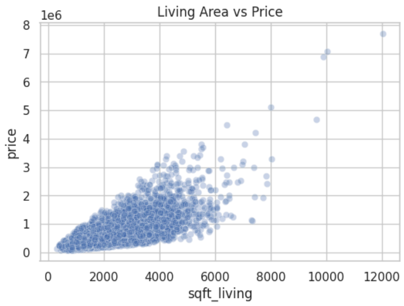
4.floors

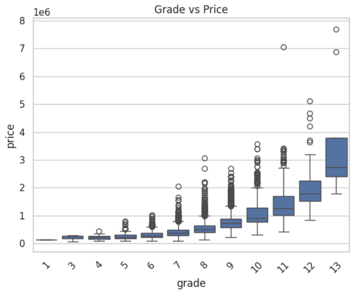
Discrete numeric variable with dominant peaks at **1 and 2 floors**.
Fractional values (1.5, 2.5) indicate split-level or partial floors.
Right-skewed distribution with very few high-rise residential units.
Should be treated as **numeric but non-continuous**.

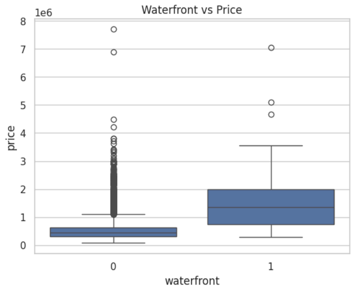
Bivariate Analysis

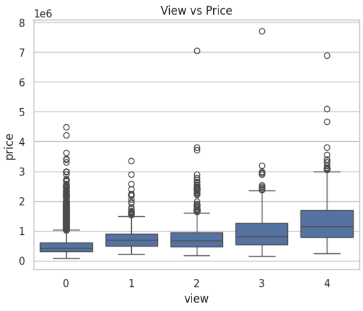


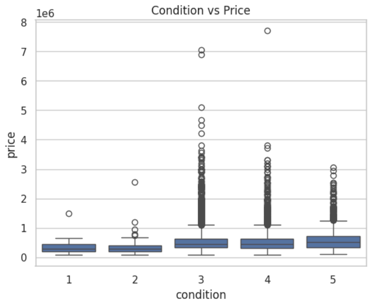


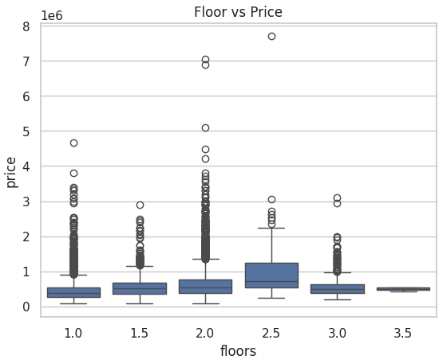












1. Bedrooms vs Price

- Median house price **increases with number of bedrooms** up to around 5–6 bedrooms.
- Price variability increases as bedroom count increases, indicating **heteroscedasticity**.
- Beyond 6 bedrooms, the median price does not increase significantly, suggesting **diminishing returns**.
- Extremely high bedroom counts (e.g., 10+, 33) show irregular pricing, likely reflecting **non-residential or anomalous properties**.
- High-priced outliers for 4–6 bedroom houses may correspond to luxury homes with superior location or amenities.
- Very large bedroom counts with low prices suggest **data entry errors or special-purpose buildings**.

2.Bathrooms vs Price

- Strong positive relationship between **number of bathrooms and price**.
- Median price rises almost monotonically as bathrooms increase, showing higher predictive power than bedrooms.
- Larger interquartile ranges for higher bathroom counts indicate increased price dispersion.
- High-price outliers at moderate bathroom counts (3–4) indicate **premium properties**.
- Extremely high bathroom counts with limited observations may represent **luxury estates or recording inconsistencies**.

3.Living Area(sqft_living) vs Price

- Clear **positive linear trend** between living area and price.
- Price variance increases with square footage, indicating **non-constant variance**.
- Larger homes (>8000 sqft) show steep price increases, emphasizing size as a dominant driver.
- Very large houses with disproportionately high prices represent **luxury market extremes**.
- Large sqft homes with relatively low prices may reflect **location disadvantages or aging properties**.

4.Waterfront vs Price

- Waterfront properties have a **substantially higher median price** than non-waterfront homes.
- Distribution for waterfront homes shows wider spread, reflecting diverse luxury pricing.
- Clear separation between the two categories suggests **strong binary influence** on price.
- Extremely high-priced waterfront homes represent **exclusive premium locations**.
- High-priced non-waterfront outliers may indicate **indirect water views or elite neighborhoods**.

5.View vs Price

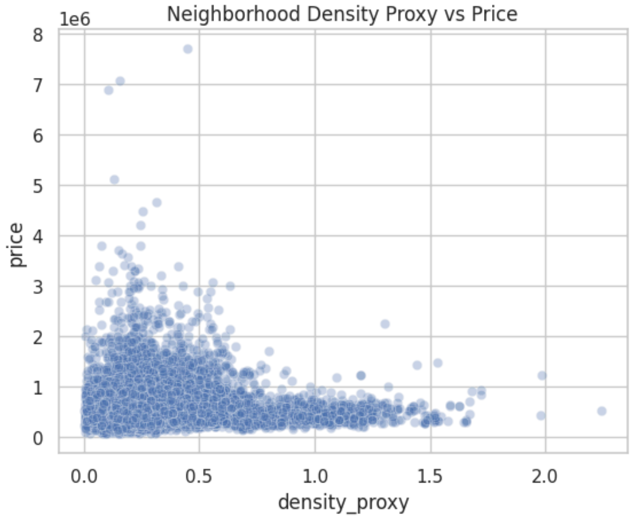
- Median price increases with **higher view ratings**, confirming view quality as a value enhancer.
- Price variability increases with view score, especially for ratings 3–4.
- View has a **graded ordinal impact** rather than a binary one.
- High-priced outliers even at low view levels indicate **other dominant factors (location, size)**.
- Very high prices at top view ratings reflect **scarcity-driven premiums**.

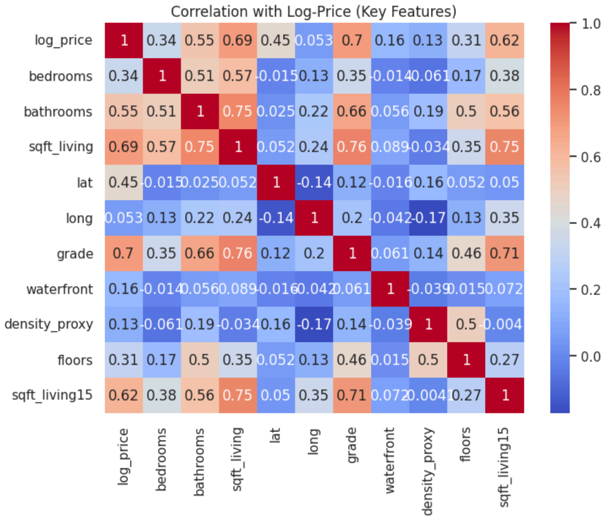
6.Conditions vs Price

- Median price increases slightly with better condition, but overlap between categories is high.
- Condition alone is a **weaker predictor** compared to size or location-related features.
- Most properties cluster around average condition (3–4).
- High-priced outliers across all condition levels suggest **renovations, location, or historical value**.
- Low-condition but high-price homes may represent **land value dominance**.

7.Floors vs Price

- Property price shows a non-linear increase with number of floors, with median prices rising sharply from 1 to 2–2.5 floors and plateauing thereafter.
- 2 and 2.5 floor houses exhibit the highest price variability and extreme outliers, indicating that premium and luxury properties are concentrated in these categories.
- 3+ floor properties do not consistently command higher prices, suggesting diminishing marginal valuation for additional floors.
- The non-monotonic trend confirms that floors is a strong but non-linear predictor, well suited for tree-based models like XGBoost.





	log_price	bedrooms	bathrooms	sqft_living	lat	long	grade	waterfront	density_proxy	floors	sqft_living15
log_price	1	0.34	0.55	0.69	0.45	0.053	0.7	0.16	0.13	0.31	0.62
bedrooms	0.34	1	0.51	0.57	-0.015	0.13	0.35	-0.014	0.061	0.17	0.38
bathrooms	0.55	0.51	1	0.75	0.025	0.22	0.66	0.056	0.19	0.5	0.56
sqft_living	0.69	0.57	0.75	1	0.052	0.24	0.76	0.089	-0.034	0.35	0.75
lat	0.45	-0.015	0.025	0.052	1	-0.14	0.12	-0.016	0.16	0.052	0.05
long	0.053	0.13	0.22	0.24	-0.14	1	0.2	-0.042	-0.17	0.13	0.35
grade	0.7	0.35	0.66	0.76	0.12	0.2	1	0.061	0.14	0.46	0.71
waterfront	0.16	-0.014	0.056	0.089	-0.016	0.042	0.061	1	-0.039	0.015	0.072
density_proxy	0.13	-0.061	0.19	-0.034	0.16	-0.17	0.14	-0.039	1	0.5	-0.004
floors	0.31	0.17	0.5	0.35	0.052	0.13	0.46	0.015	0.5	1	0.27
sqft_living15	0.62	0.38	0.56	0.75	0.05	0.35	0.71	0.072	0.004	0.27	1

Feature Extraction

Density Proxy

- `density_proxy` is defined as the **ratio of built-up area to land area**.
- It represents **land-use intensity** rather than population density.
- Higher values indicate **more construction on a given land parcel** (compact, dense development).
- Lower values correspond to **larger plots with relatively smaller houses** (spacious, low-density areas).
- The feature captures **urbanization and zoning effects** not directly expressed by size variables alone.

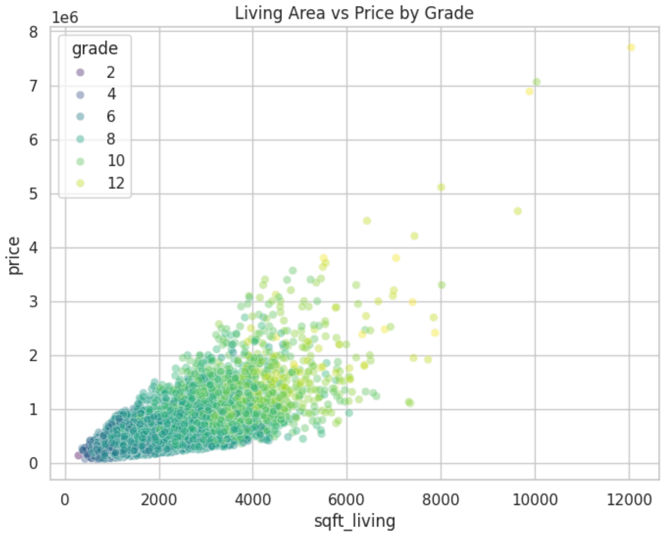
Density Proxy and Price Variation

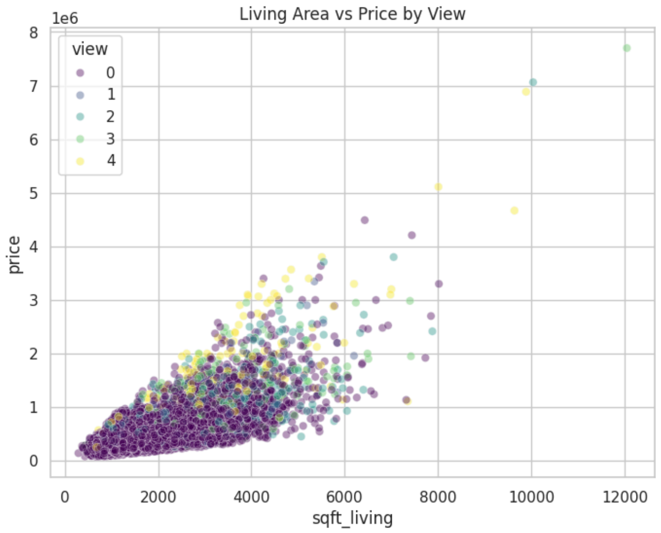
- Properties with **low density proxy** exhibit **large price dispersion**:
    - Includes both affordable suburban homes and premium large-lot properties.
- **Moderate density values** are associated with **stable mid-range pricing**.
- Very **high density proxy** values show **limited price variation**:
    - Indicates compact developments where price is constrained by space.
- This implies a **nonlinear relationship**:
    - Increasing land-use intensity does not proportionally increase property price.
- Density acts as a **contextual modifier**, not a primary price determinant.

Correlation Matrix

- `log_price` shows **strong positive correlation** with `sqft_living`, `grade`, and `sqft_living15`, identifying **house size and construction quality** as primary price drivers.
- **Moderate correlations** between `log_price` and `bathrooms` / `bedrooms` indicate functional utility affects price but is secondary to size and quality.
- Strong **inter-correlation among size-related variables** (`sqft_living`, `bathrooms`, `bedrooms`, `floors`) reflects structural dependency and potential multicollinearity.
- `sqft_living15` correlates with both `log_price` and `grade`, highlighting **neighborhood quality and peer effects** in pricing.
- Geographic variables (`lat`, `long`) exhibit **weak-to-moderate correlation** with price, implying location effects are not fully captured linearly.
- `density_proxy` shows **weak linear correlation** with `log_price` and low correlation with dominant predictors, indicating it contributes **independent contextual information** rather than acting as a primary driver.

Multi-Variate Analysis





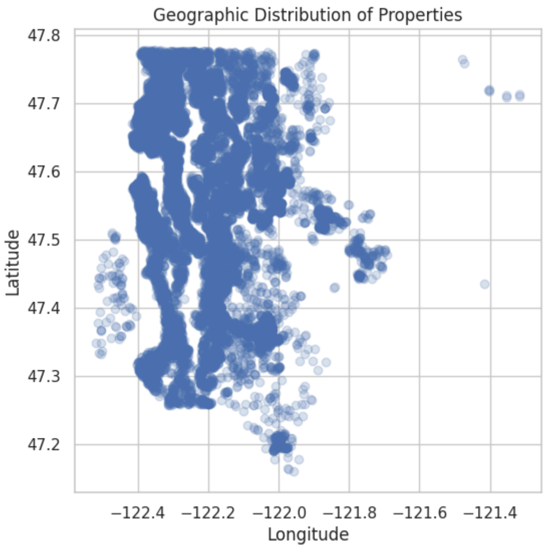
Living Area Vs Price by Grade

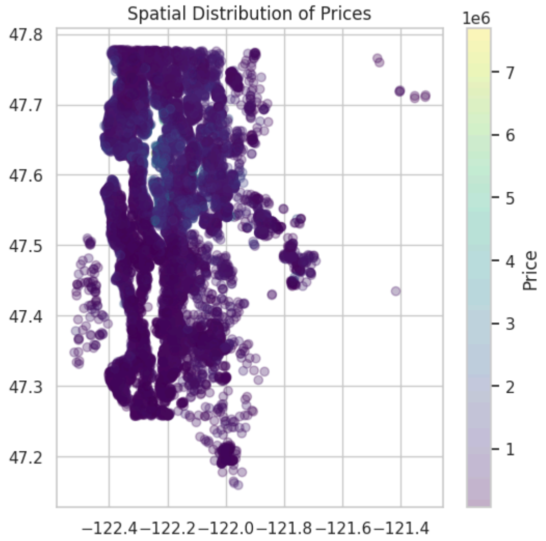
- `sqft_living` exhibits a **strong positive relationship** with price, confirming living area as a primary price driver.
- At a fixed living area, **higher grade homes consistently command higher prices**, indicating grade has an **independent and multiplicative effect** on price.
- Price dispersion **increases with both size and grade**, showing interaction effects rather than a simple linear trend.
- High-grade properties dominate the **upper price envelope**, especially for large living areas.
- Grade effectively **segments the market**, separating low- and high-value homes even at similar sizes.

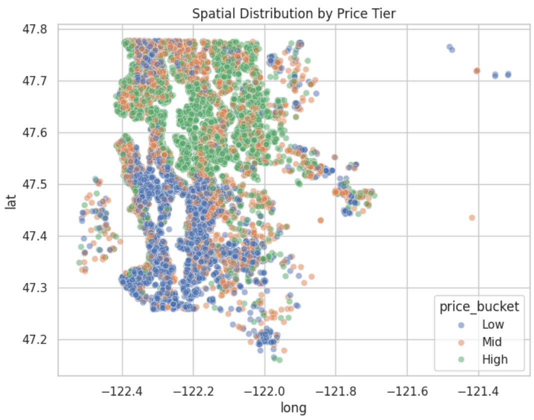
Living Area Vs Price by View

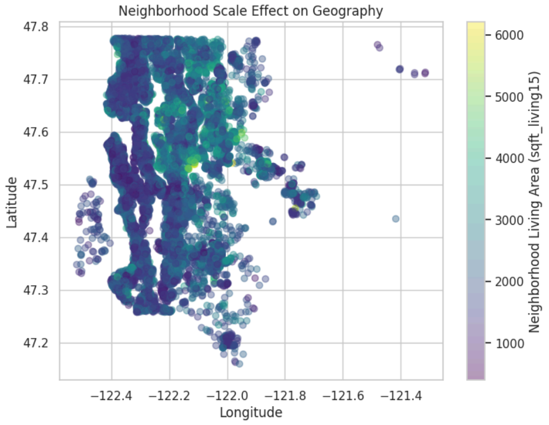
- Living area remains the **baseline predictor** of price across all view categories.
- Properties with better views exhibit a **systematic upward price shift** at comparable living areas.
- View acts as a **premium feature**, amplifying price rather than replacing the effect of size.
- The premium associated with view is **more pronounced at higher living areas**, indicating interaction between size and view.
- Lower-view categories show **greater price overlap**, suggesting view contributes to price differentiation mainly at higher market segments.

Geospatial Analysis


Geographic Distribution of Properties


Spatial Distribution of Prices

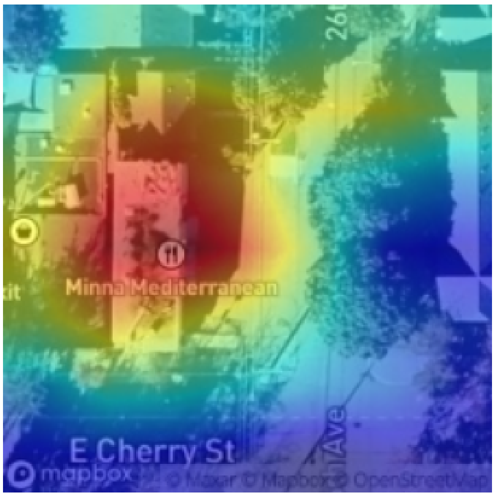

Spatial Distribution by Price Tier


Neighborhood Scale Effect on Geography

- Property prices exhibit **strong spatial clustering**, with high-value homes concentrated in specific geographic pockets rather than uniformly distributed.
- Distinct **price tiers align with neighborhood boundaries**, indicating location-driven market segmentation.
- Central and well-developed regions command **consistent price premiums**, while peripheral areas show lower valuations.
- Neighborhood-scale attributes (average living area, development density) **reinforce spatial price patterns**.
- Visually greener, open, and well-planned areas correlate with **higher property values**, whereas dense concrete regions show price compression.
- Overall, geospatial and visual context acts as a **critical secondary driver of price**, amplifying the effects of structural property features.

Financial & Visual Analysis from Grad-cam analysis

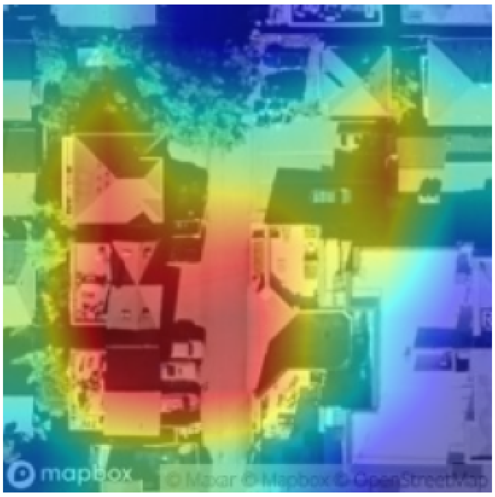
Grad-CAM ID: 6840700165
Price: $202,000



1.Low Value Property- Dense Built up Environment

Grad-CAM activation is **highly concentrated on the building footprint**, with minimal attention to surroundings.
The surrounding area shows **dense housing, narrow roads, and limited or no green cover**.
Financial interpretation:
1.Value is driven mainly by **structure presence**, not environmental quality.
2.Lack of open space and greenery reduces land premium and lifestyle appeal.
Model insight:
The CNN implicitly penalizes **congestion and spatial crowding**, leading to lower predicted prices.

Grad-CAM ID: 7011201470
Price: $625,000



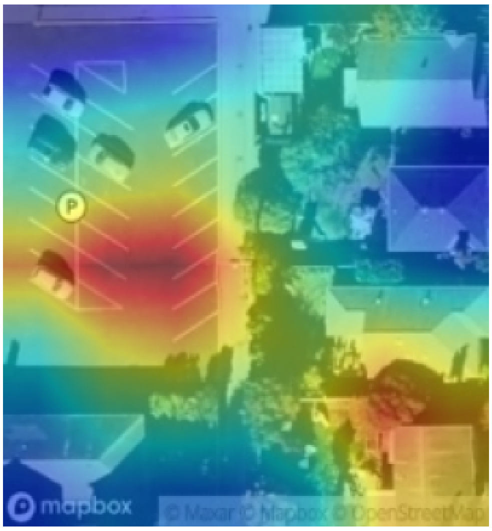
2.Mid Value Property - Balanced Residential Layout

Grad-CAM highlights both the **main structure and nearby open spaces**, including yards and access roads.
Partial greenery and moderate spacing between buildings are visible.

Financial interpretation:
1.Represents a balance between **built-up value and neighborhood quality**.
2.Moderate green cover and organized layout contribute positively but not dominantly.

Model insight:
The model captures **neighborhood quality as a secondary multiplier**, complementing structural features.

Grad-CAM ID: 6021503570
Price: $525,000



3.High Value Property-Open Space & Green Dominated Area

Grad-Cam highlights strong activation over-large building footprint and surrounding open space,high greenery density with well-separated structures and clear road access and low local congestion.
Financial Interpretation
1.Indicates premium valuation driven by spatial exclusivity and environmental quality.

2.Green cover and low-density layout significantly enhance perceived value.

Model Insight
The model associates open space and greenery with high-end properties and environmental features act as strong positive signals, not just complements.

Cross-Interpretation Across Price Segments

Observed Pattern

- Low-value properties emphasize built density and immediate structures.
- Mid-value properties show balanced attention to structure and surroundings.
- High-value properties prioritize open space, greenery, and spatial separation.

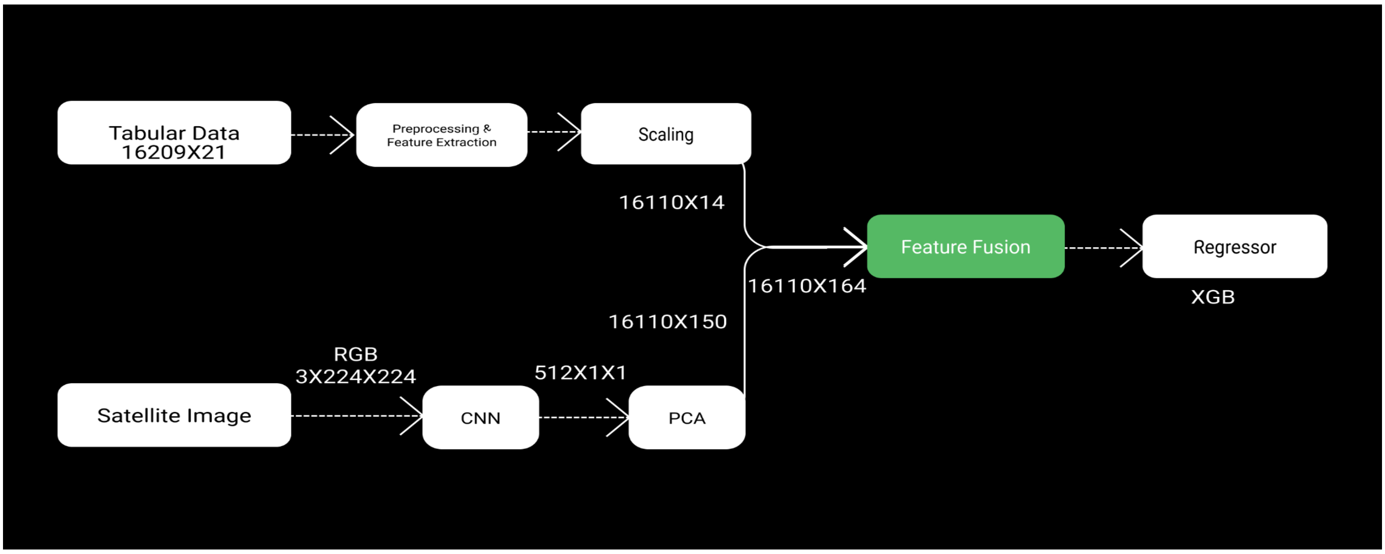
Financial Insight

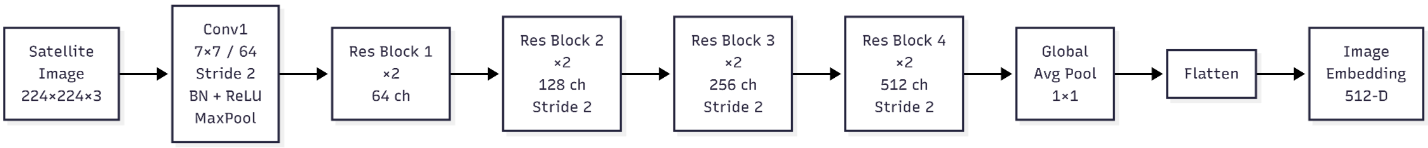
- Visual features increasingly influence price as property value rises.
- Environmental quality acts as a differentiator in higher price segments.

Explainability Conclusion

- Grad-CAM confirms that the model learns economically meaningful visual cues.
- Satellite imagery enhances valuation by capturing neighborhood-level signals beyond tabular data

Architecture Diagram





CNN Architecture

Results & Model Selection

Multiple regression models were evaluated using **Tabular Data Only** and **Tabular + Satellite Image embeddings**, with performance assessed via **log-RMSE** and **R² (Train / Validation)** as seen in the table below. Tree-based ensemble models consistently outperformed linear baselines, indicating strong non-linear relationships in property valuation.

Among all multimodal models, **XGBoost (Tabular + Image)** achieved strong validation performance (**RMSE ≈ 0.172, Val R² ≈ 0.89**) while explicitly incorporating satellite imagery, thereby satisfying the project's multimodal learning objective. Although a train–validation gap is observed (Train R² ≈ 0.96), this level of overfitting is expected given the high-dimensional image embeddings and remains comparable or lower than other multimodal tree-based models.

While **LightGBM (Tabular-only)** marginally outperformed XGBoost in terms of RMSE, it does not leverage visual information and therefore does not align with the project's goal of integrating satellite imagery. Additionally, **LightGBM (Tabular + Image)** showed stronger overfitting tendencies, with near-perfect training performance and reduced validation stability.

Consequently, **XGBoost (Tabular + Satellite Images)** was selected as the final model, as it offers a strong balance between predictive accuracy, robustness, and multimodal capability. Importantly, this model enables **Grad-CAM–based visual explainability**, allowing inspection of which spatial and environmental features (such as greenery, open space, building density, and road structure) influence price predictions. These visual explanations directly support the **Financial and Visual Insights** analysis presented in the following section.

	Model	RMSE (log)	R2 (Train)	R2 (Val)
12	LightGBM (Tabular)	0.167400	0.977047	0.899848
10	XGBoost (Tabular)	0.167783	0.941563	0.899391
13	LightGBM (Tabular + Image)	0.171680	0.997382	0.894663
11	XGBoost (Tabular + Image)	0.172087	0.964417	0.894162
8	Gradient Boosting (Tabular)	0.174393	0.911304	0.891307
9	Gradient Boosting (Tabular + Image)	0.179155	0.916921	0.885289
6	Random Forest (Tabular)	0.179390	0.970574	0.884988
7	Random Forest (Tabular + Image)	0.194160	0.973653	0.865270
5	Lasso Regression (Tabular + Image)	0.249660	0.781935	0.777238
3	Ridge Regression (Tabular + Image)	0.249676	0.782327	0.777209
1	Linear Regression (Tabular + Image)	0.249682	0.782327	0.777199
2	Ridge Regression (Tabular)	0.266552	0.750180	0.746074
0	Linear Regression (Tabular)	0.266555	0.750180	0.746067
4	Lasso Regression (Tabular)	0.266604	0.750108	0.745974