CDC X Yhills OPEN PROJECTS 2025-2026

VARNIKA 22123047

Data Science PS

Satellite Imagery-Based Property Valuation

Overview

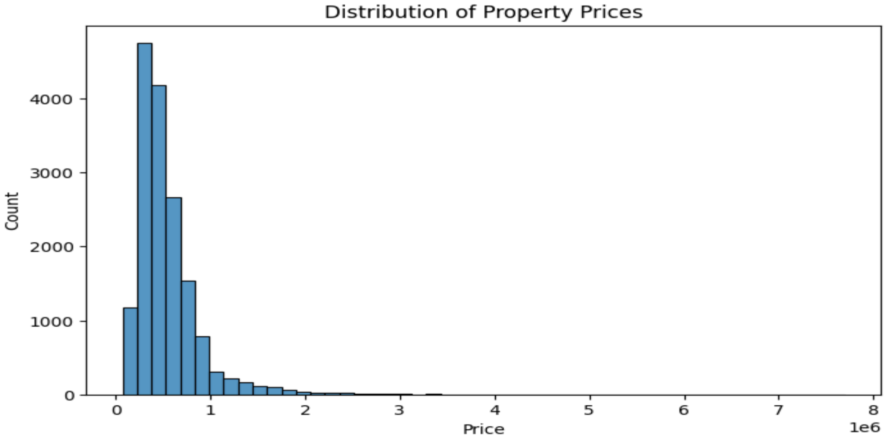
- In this project, the goal is to predict house prices using both traditional housing data and satellite images.
- Normally, house price prediction models use only tabular data such as number of bedrooms, living area, location, and other numeric features.
- However, such data does not fully capture the surrounding environment of a property.
- To address this limitation, I built a multimodal regression system that combines tabular data with satellite imagery.
- The satellite images provide visual information about the neighborhood, such as road connectivity, building density, and green areas, which can influence property value.
- The project was implemented in two main stages.
- First, a strong tabular-only model was developed using engineered features and machine learning techniques.
- This model serves as a baseline to understand how well prices can be predicted using only numeric data.
- In the second stage, satellite images were downloaded using latitude and longitude coordinates.
- A pretrained Convolutional Neural Network (ResNet18) was used to convert these images into numerical feature vectors.
- These image features were then combined with tabular features to build a multimodal regression model.
- To handle the high dimensionality of image features, Principal Component Analysis (PCA) was applied before feature fusion.
- This helped reduce overfitting and improved model stability.
- Finally, model performance was evaluated by comparing the tabular-only model with the multimodal model.
- Visual explainability was added using Grad-CAM to understand which regions of the satellite images influenced the model’s predictions.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to understand the distribution of house prices and the relationship between price and key property features.
The goal of this analysis was to identify important patterns and factors that influence property value.

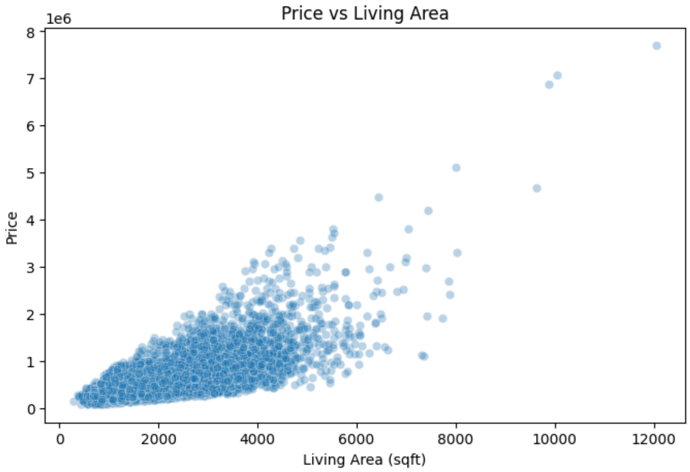
1. Price Distribution

The price distribution shows that most properties are priced in the lower to mid range, while a small number of properties have very high prices.
This indicates that the data is right-skewed, with a few luxury properties acting as outliers.
This observation justifies the use of logarithmic transformation on price during model training to stabilize variance and improve model performance.



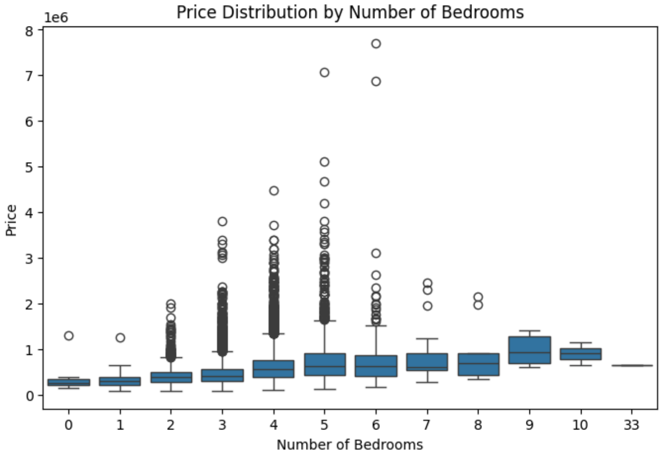
2. Living Area vs Price

The scatter plot between living area and price shows a clear positive relationship.
As the living area of a house increases, the price generally increases as well.
However, for very large houses, prices vary significantly.
This suggests that while size is important, other factors such as location, view, and neighborhood also influence property value.



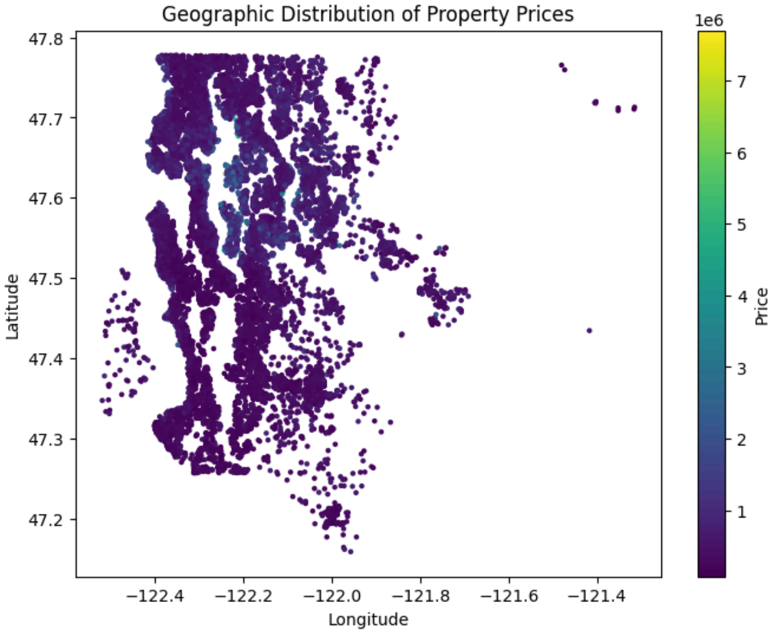
3. Number of Bedrooms vs Price

The box plot of price by number of bedrooms shows that houses with more bedrooms tend to have higher prices.
However, the increase in price is not strictly linear.
Some houses with fewer bedrooms are priced higher than houses with more bedrooms.
This indicates that bedroom count alone does not determine price and must be considered along with other features.



4. Geographic Distribution of Prices

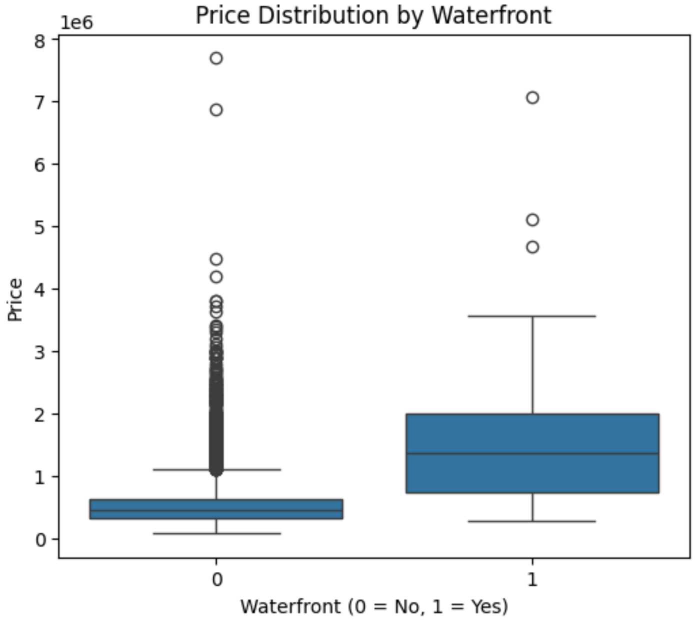
The geographic plot shows that property prices vary significantly across different locations.
Clusters of higher-priced properties are observed in specific geographic regions.
This confirms that location plays a major role in property valuation.
Latitude and longitude are therefore important features for the prediction model.



5. Waterfront vs Non-Waterfront Properties

The comparison between waterfront and non-waterfront properties shows a clear difference in price.

Waterfront houses have significantly higher median prices compared to non-waterfront houses. This indicates that proximity to water is a strong value-driving factor and should be included as an important feature in the model.



Sample Satellite Images

To understand the visual context of different properties, sample satellite images were analyzed.

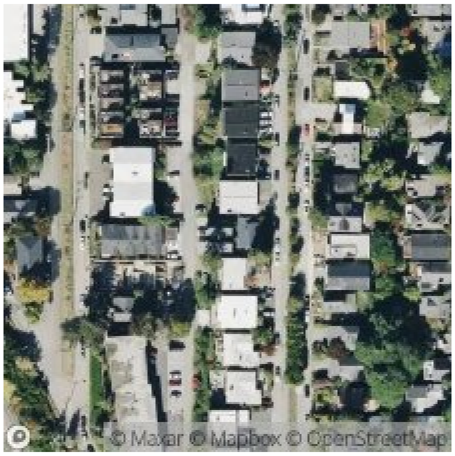
These images show variations in neighborhood structure, road layout, and surrounding green areas.

© Maxar © Mapbox © OpenStreetMap

The first image shows a property located in a greener and more open area, with fewer surrounding buildings and more vegetation.
Such environments are often associated with better living conditions and can positively influence property value.

© Maxar © Mapbox © OpenStreetMap

The second image represents a moderately dense residential neighborhood, where houses are closely spaced but still surrounded by trees and open areas.
This type of locality reflects typical suburban housing patterns.



The third image shows a highly dense urban area with tightly packed buildings and limited green spaces.
Properties in such regions may benefit from better connectivity and access to amenities, but may lack open or green surroundings.

These visual differences highlight aspects of neighborhood quality that are not fully captured by tabular data alone.
This observation supports the use of satellite imagery as an additional input in the multimodal property valuation model.

Financial / Visual Insights

1. Financial Insights from Tabular Data (EDA-based)

From the exploratory data analysis of the tabular dataset, the following financial patterns were observed:(Graphs above attached)

Price Distribution

Property prices are **right-skewed**, meaning:

- Most houses are in the **low to mid-price range**
- A small number of very expensive houses create a long tail

This indicates that predicting extreme high prices is harder and requires more information.

Living Area vs Price

A strong positive relationship is observed between **sqft_living** and price:

- Larger houses generally have higher prices
- However, price does not increase linearly for very large houses, showing diminishing returns

This confirms that living area is one of the **most important financial drivers** of property value.

Bedrooms

- Price increases as the number of bedrooms increases up to a point
- Very high bedroom counts show high variance, meaning bedrooms alone are not sufficient to explain price

This suggests that **house quality and location matter more than just room count**.

Waterfront Effect

Properties with **waterfront access** have significantly higher median prices compared to non-waterfront properties.

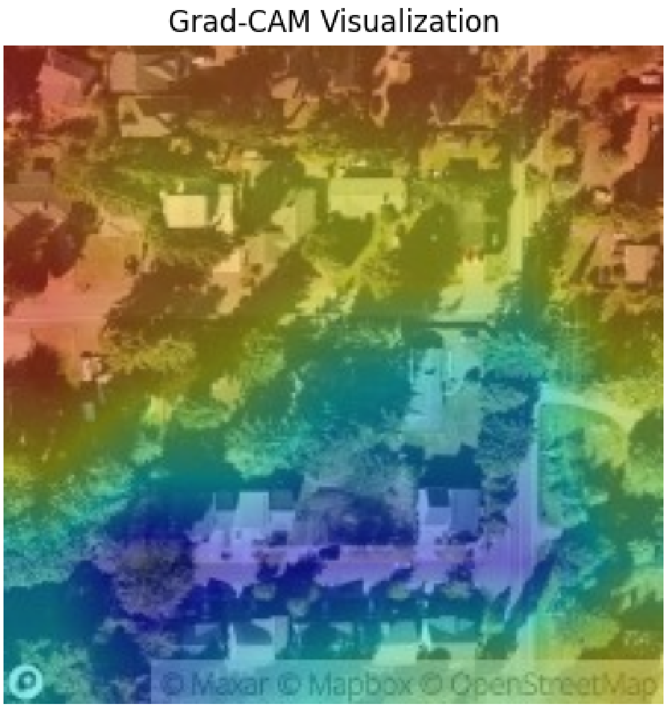
This clearly shows that **location-based premium features** strongly affect property value.

Geographic Patterns

The geographic scatter plot shows price clustering:

- Certain latitude–longitude regions consistently have higher prices
- This confirms that **location plays a critical role** in real estate valuation

Visual Insights from Satellite Images (Grad-CAM based)


Grad-CAM Visualization

The Grad-CAM visualization highlights the regions of the satellite image that most influenced the model's prediction.

The warm colors (red, yellow, green) indicate areas that the model considers important, while cooler colors (blue) indicate less important regions.

Key Observations

1. The model focuses strongly on clusters of buildings and nearby open areas.
   This suggests that neighborhood structure and building layout influence price prediction.

2. Roads and connectivity patterns are partially highlighted.
   This indicates that accessibility and infrastructure are important visual signals for the model.

3. Green and open regions receive moderate attention.
   This shows that environmental quality, such as greenery and open space, contributes to property value.

4. The model does not focus on empty or irrelevant areas.
   This means the CNN is learning meaningful visual patterns instead of random noise.

Simple Architecture Diagram

1. Tabular Data Processing

The tabular housing data is first prepared for modeling using the following steps:

- Non-informative identifiers (such as house ID) are removed
- Date information is converted into numeric features like:
    - Sale year
    - Sale month
- Raw date strings are dropped to avoid model errors
- New meaningful features are created (for example, living area to lot size ratio)
- The target variable (house price) is log-transformed to handle skewed distribution

This processing ensures that tabular data is clean, numerical, and model-ready.

2. Image Data Processing

Satellite images of properties are processed using a deep learning model:

- Images are resized to a fixed size (224 × 224)
- Pixel values are normalized and converted into tensors
- Images are passed through a pretrained ResNet-based CNN
- The final classification layer of the CNN is removed
- Each image is converted into a high-level feature vector

These image features capture visual and neighborhood information useful for price prediction.

3. Dimensionality Reduction using PCA

The extracted image features are high-dimensional and may contain redundant information.
To address this, PCA is applied:

- Image feature vectors are standardized
- PCA reduces feature dimensions (e.g., from 2048 to a smaller size)
- Important visual information is preserved while noise is reduced

Advantages of PCA:

- Faster model training
- Lower risk of overfitting
- Better generalization performance

4. Multimodal Feature Fusion

After preprocessing both data types:

- Tabular features describe property-specific attributes
- PCA-reduced image features represent surrounding context
- Both feature sets are concatenated to form a single feature vector

This multimodal representation allows the model to learn from **structured data and visual data together**, leading to improved prediction performance.

Results

In this project, two models were developed and evaluated:

1. **Tabular Data Only Model**
2. **Multimodal Model (Tabular + Satellite Images)**

Both models were evaluated using **RMSE** and **R² score**.

Tabular Data Only Model

The tabular model was trained using only structured numerical features such as:

- Bedrooms, bathrooms
- Living area and lot size
- Location (latitude and longitude)
- Waterfront, view, condition, and grade

Performance:

- **RMSE:** 113767.24251084936
- **R² Score:** 0.896

Multimodal Model (Tabular + Satellite Images)

In the multimodal model:

- Satellite images were generated using latitude and longitude.
- A pretrained CNN (ResNet18) was used to extract visual features from images.
- Image features were combined with tabular features and used for regression.

Performance:

- **RMSE:**187896.76836732912
- **R² Score:**0.72

Key Takeaway

- The **tabular model** provides highly accurate price predictions.
- The **multimodal model** adds **visual interpretability** and contextual understanding.
- Satellite images help explain **environmental and neighborhood effects**, even if numeric accuracy is slightly lower.
- Although Grad-CAM highlighted meaningful spatial regions, image features alone could not outperform tabular data due to limited image resolution, lack of temporal or neighborhood metadata, and a simple concatenation-based fusion approach.

- PCA helped stabilize the model by reducing high-dimensional image embeddings, but it may have removed some fine-grained visual details.