

# Report: CDC X Yhills OPEN PROJECTS 2025-2026

## [Data Science PS]

Name: Abhinav Gupta  
Enrolment No.: 23322001  
Branch: Mathematics & Computing

---

### About Problem:

Traditional property valuation models rely mainly on tabular features such as location, size, and room counts, but they overlook important visual and environmental factors like greenery, road access, and nearby water bodies. These visual elements strongly influence buyer perception but are difficult to quantify using standard data.

Advances in satellite imagery and deep learning now make it possible to extract meaningful visual information and combine it with numerical features. This project addresses the challenge of integrating both data types into a single multimodal regression model to improve price prediction accuracy.

---

### Methodology overview:

The project follows a multimodal learning pipeline that combines tabular housing data with satellite imagery to predict property prices. Tabular features are first cleaned, selected, and scaled for use in traditional regression models. Satellite images are then fetched using property latitude and longitude to capture environmental and neighbourhood context.

A Convolutional Neural Network is used to extract high-level visual features from the satellite images. These image features are fused with the tabular data and passed to a regression model for final price prediction. Model performance is evaluated using RMSE and R<sup>2</sup> score, and Grad-CAM is used to provide visual explainability by highlighting image regions that influence predictions.

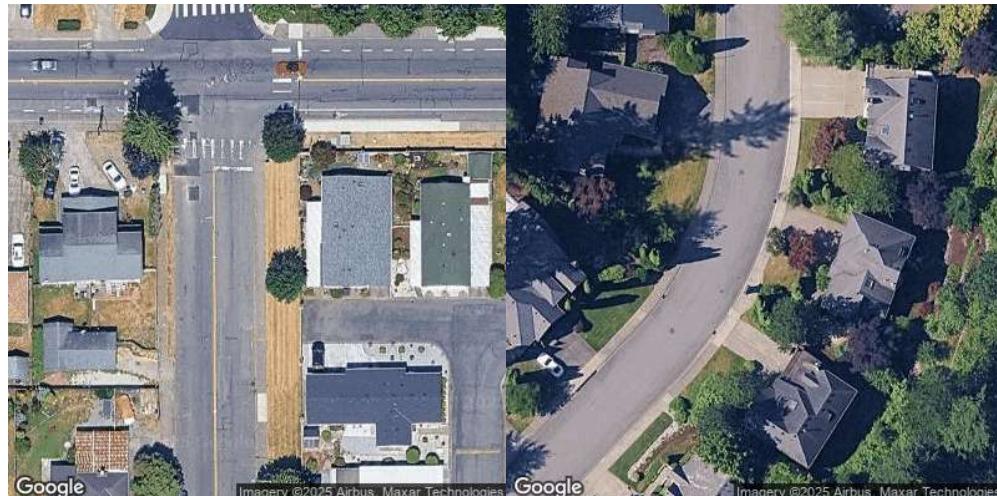
---

### Fetching Images:

Satellite images for each property were programmatically downloaded using the Google Maps Static API at zoom level 19, which provides high-resolution views of the immediate neighborhood and surrounding environment.

**Code:** data\_fetcher.ipynb

### Results:



# Exploratory Data Analysis:

## About the dataset:

```
[6]: train_df.columns  
  
[6]: Index(['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'],  
           dtype='object')
```

---

## Uniqueness of Data:

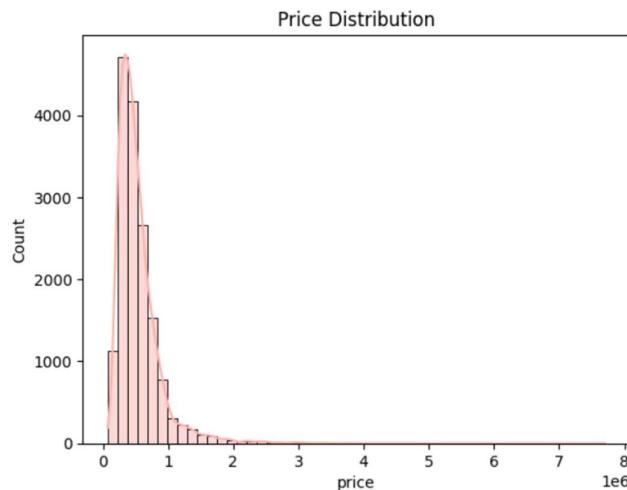
```
[10]: train_df["id"].nunique(), len(train_df)  
  
[10]: (16110, 16209)  
  
+ Code + Markdown
```

The training set contains 16,209 rows but only 16,110 unique house IDs, indicating a small number of duplicate property listings.

So, we will drop the duplicates and keep the most recent one.

---

## Price Distribution:



This implies we have prices in range roughly from 10000 to 8000000.

---

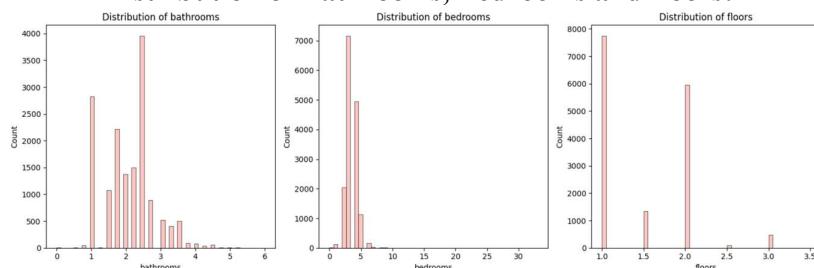
## Outliers:

We drop the data with very high value of sqft\_living and sqft\_lot as they will make our model worse.

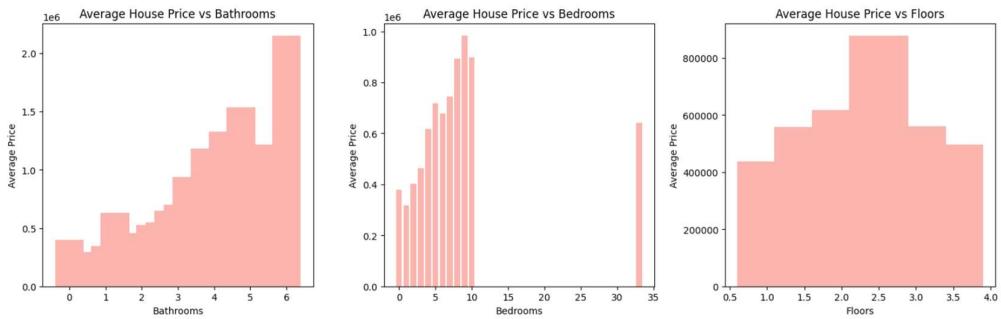
Dropped 474 rows (2.94%)

---

## Distribution of Bathrooms, Bedrooms and floors:

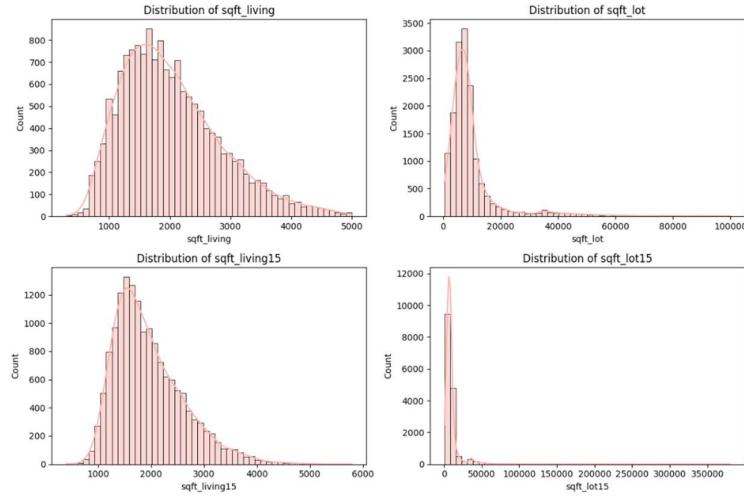


### Average House Price vs Bathrooms, Bedrooms and floors:

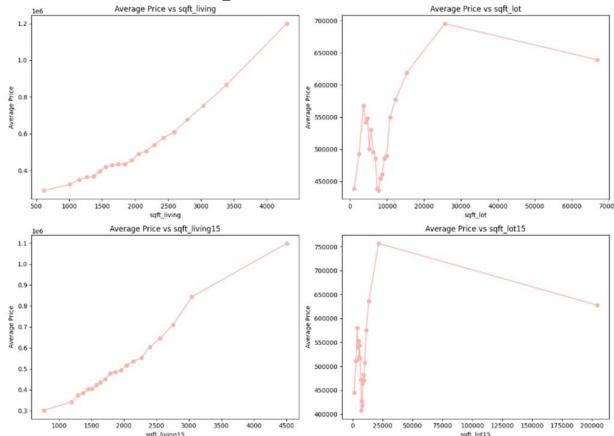


As the number of bedrooms and bathrooms increases, the average house price generally increases. However, the same trend is not observed for the number of floors, indicating that floor count alone does not strongly influence property value.

### Distribution of SquareFeetLiving, SquareFeetLot, SquareFeetLiving15, SquareFeetLot15:

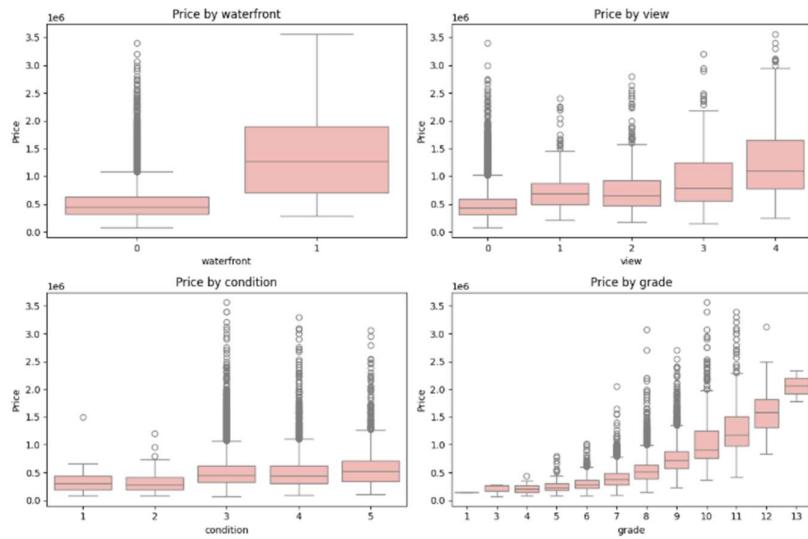


### Average Price vs SquareFeetLiving, SquareFeetLot, SquareFeetLiving15, SquareFeetLot15:



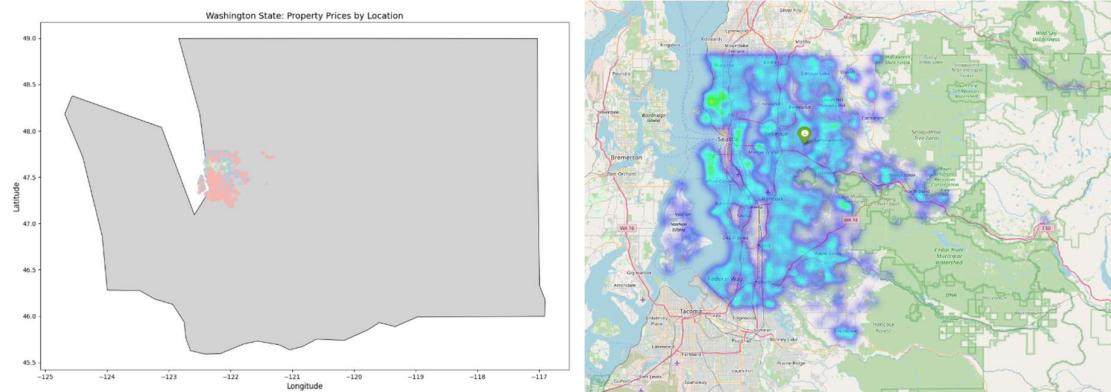
The average house price increases with higher values of sqft\_living, sqft\_lot, sqft\_living15, and sqft\_lot15, indicating that both property size and neighborhood density play an important role in valuation.

## Boxplot for Categorical Data WaterFront, View, Condition, Grade:

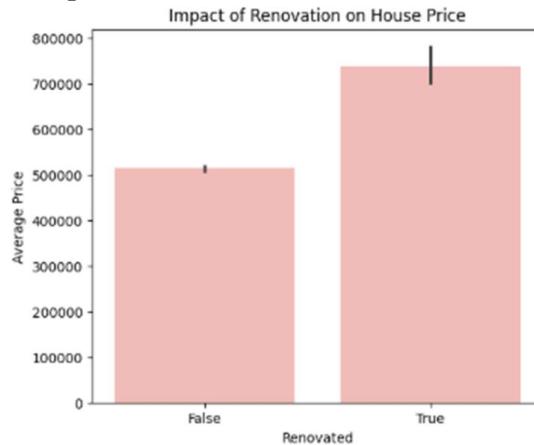


Properties with waterfront access, better views, higher condition ratings, and higher construction grades tend to have higher prices, which aligns with expected real estate valuation patterns.

## Washington State: Property Prices by Location:



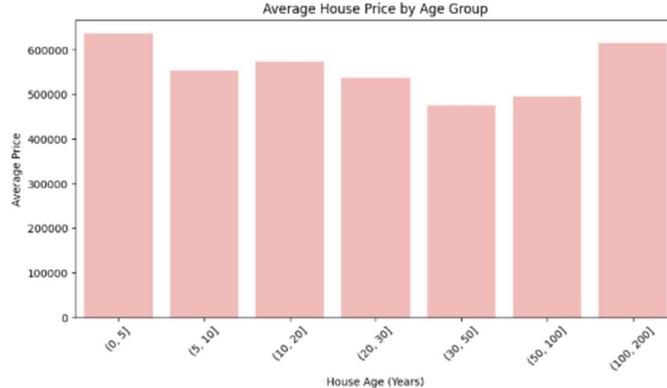
## Impact of Renovation on House Prices:



The average property price is higher for houses that have been renovated, indicating that renovations have a positive impact on market value.

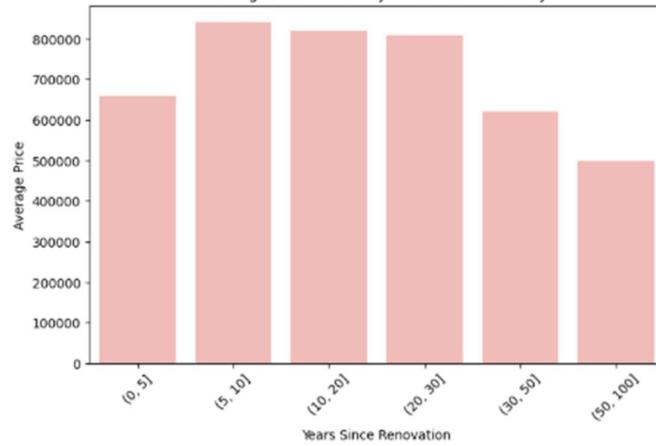
## **Impact of House Age on House Prices:**

Reference Year : 2017



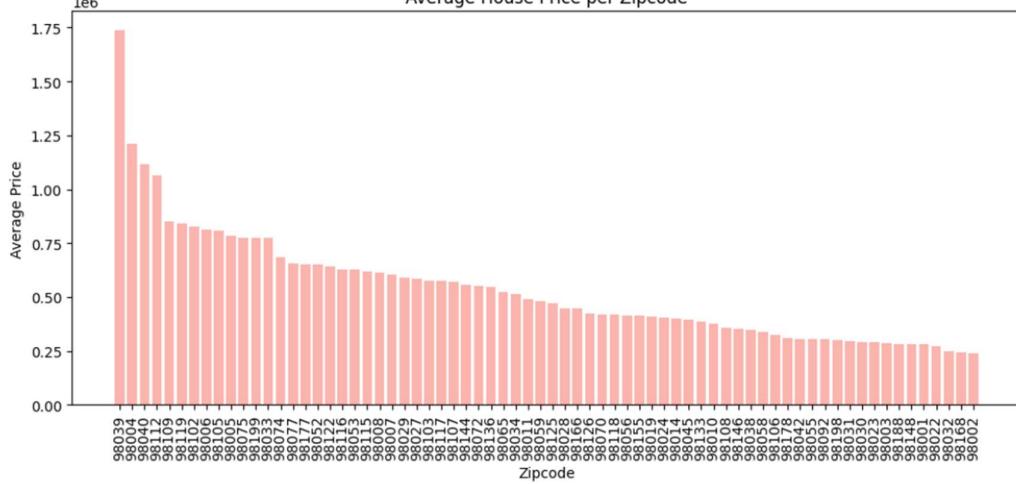
#### Average House Price by Renovation Recency:

### Average House Price by Renovation Recency



## Average House Price per Zipcode:

### Average House Price per Zipcode



## Why are prices higher in these zipcodes?

These zipcodes cover premium suburbs and some of the best localities in Washington (WA), such as Medina and Bellevue, known for waterfront views, excellent infrastructure, and proximity to major employment hubs. High demand and limited housing supply in these areas lead to significantly higher average house prices.

## Modelling:

### Data Preparation:

```
[50]: train_df["house_age"] = REFERENCE_YEAR - train_df["yr_built"]  
train_df["years_since_reno"] = np.where(train_df["yr_renovated"] > 0, REFERENCE_YEAR - train_df["yr_renovated"], 0)
```

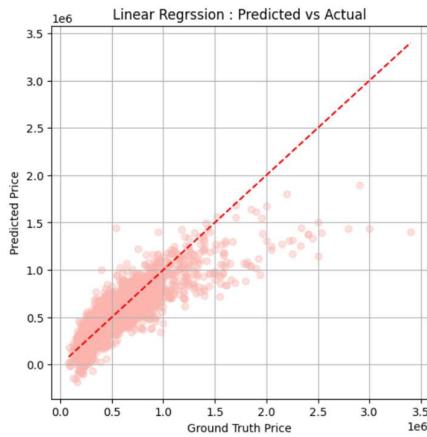
```
[28]: train_df['dist_to_seattle'] = np.sqrt((train_df['lat'] - 47.6062)**2 + (train_df['long'] + 122.3321)**2)
```

```
[59]: TARGET = "price"  
  
TABULAR_FEATURES_tab = [  
    "bedrooms",  
    "bathrooms",  
    "sqft_living",  
    "sqft_lot",  
    "floors",  
    "waterfront",  
    "view",  
    "condition",  
    "grade",  
    "sqft_above",  
    "sqft_basement",  
    "house_age",  
    "years_since_reno",  
    "lat",  
    "long",  
    "sqft_living15",  
    "sqft_lot15",  
    "dist_to_seattle"  
]
```

### Tabular-Only Baseline Models:

#### 1. Linear Regression

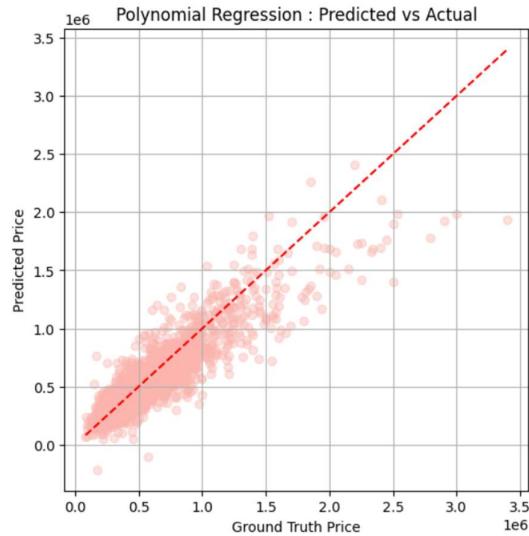
Linear Regression RMSE : 176737.09592000337, R2 : 0.6974419177662354



---

## 2. Polynomial Regression

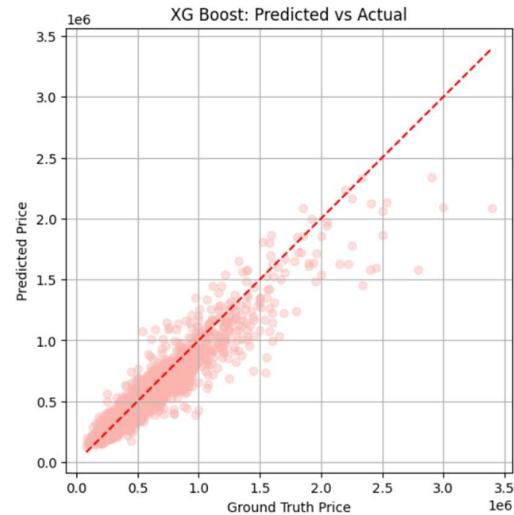
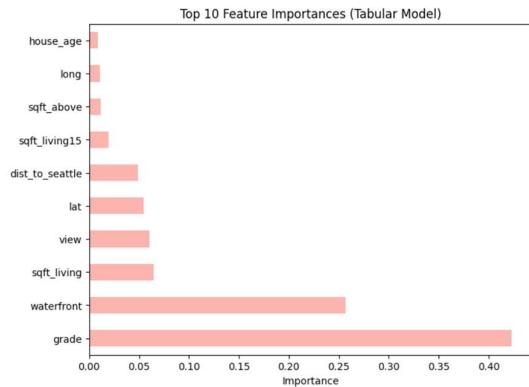
Polynomial Regression RMSE : 146190.66649299485, R2 : 0.7929893820380935



---

## 2. XG Boost

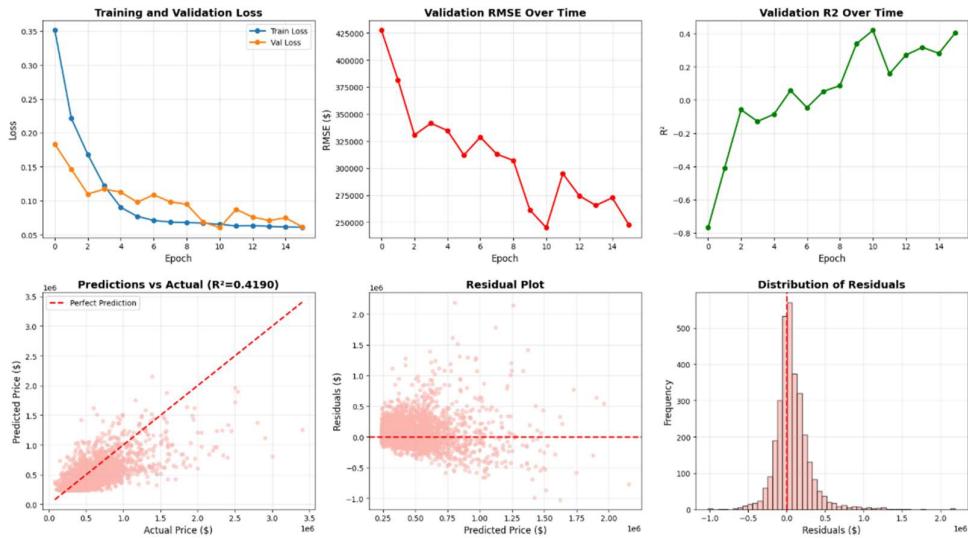
XG Boost RMSE : 112144.50563447145, R2 : 0.878182590007782



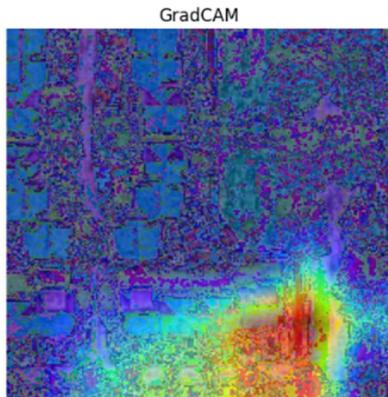
## Image-only CNN model

We use **ResNet-18** as the backbone architecture.

The model is trained for a maximum of **100 epochs** with **early stopping** enabled (patience = 5), meaning training stops if the validation metric does not improve for 10 consecutive epochs. Epochs ran = 16.



Results and GradCam:



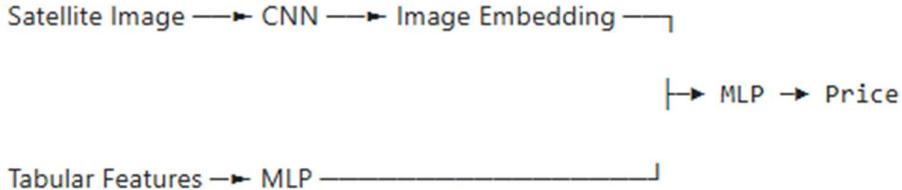
### PREDICTION STATISTICS

```
=====
Mean Actual Price:      $541,727
Mean Predicted Price:   $471,167
Std Actual Price:       $321,309
Std Predicted Price:    $230,619
Max Residual:           $2,193,301
Min Residual:            $7
=====
```

✓ Image model variables saved: RMSE=\$244,907, R<sup>2</sup>=0.4190

## Fusion Models

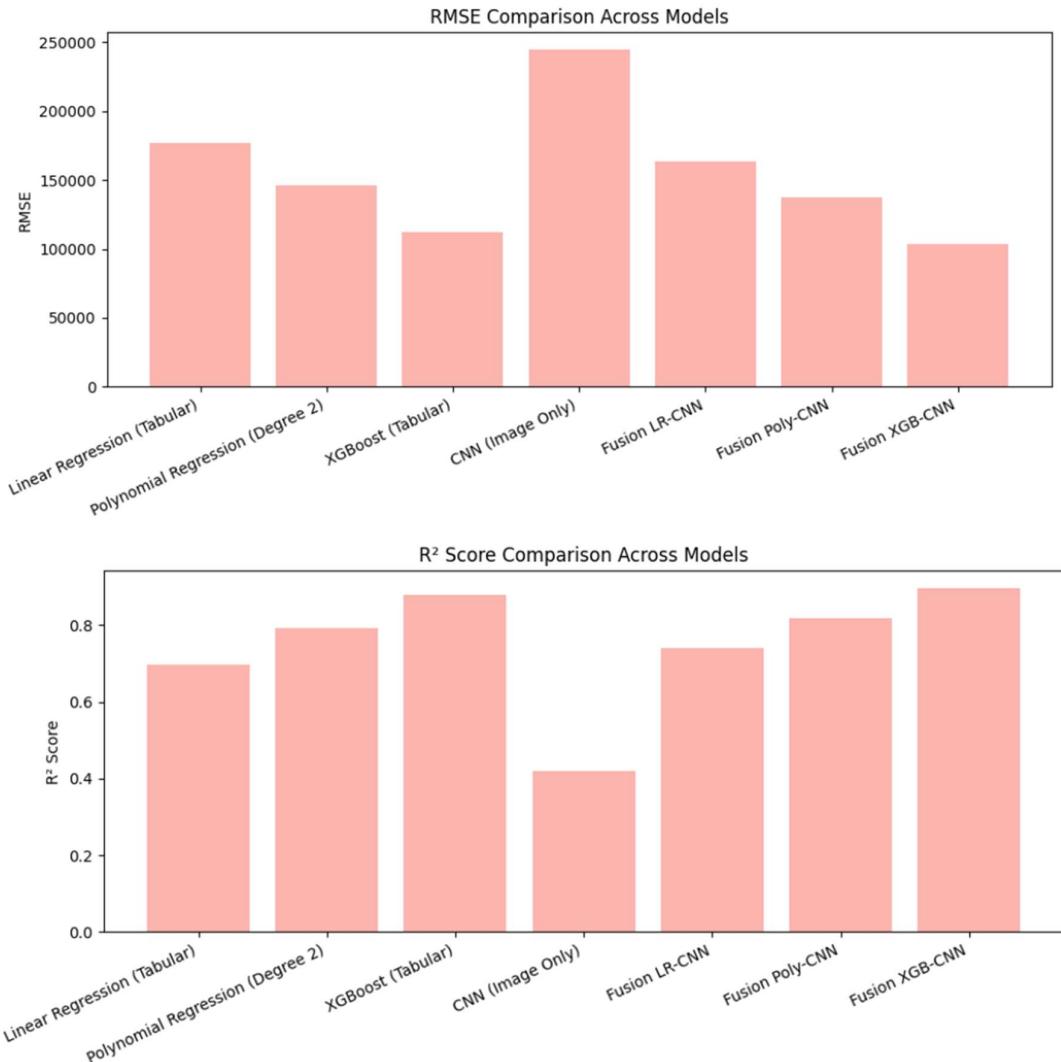
Architecture:



Results:



## Model Comparisons:



As expected, fusion models outperform single-modality models in the Satellite-Image-Based Property Valuation task.

Image-only CNN performs the worst, since satellite images alone cannot capture many critical property attributes.

Important factors such as number of floors, bathrooms, interior size, condition, grade, view, and waterfront access are not reliably inferable from images.

Tabular features like sqft\_living, sqft\_above, sqft\_basement, sqft\_lot, neighborhood statistics, and quality indicators carry strong predictive signals.

Fusion models (Tabular + Image) combine visual context with structured housing data, leading to lower RMSE and higher R<sup>2</sup>.

Among all models, XGBoost + CNN fusion performs the best, showing that strong tabular learners benefit most from complementary image features.

Overall, these results confirm that images add value only when fused with rich tabular data, rather than used in isolation.

## Test Results:

**Using Best Model (Fusion XGB-CNN) we predict on the Train Dataset.**

**Saved to : 23322001.csv**