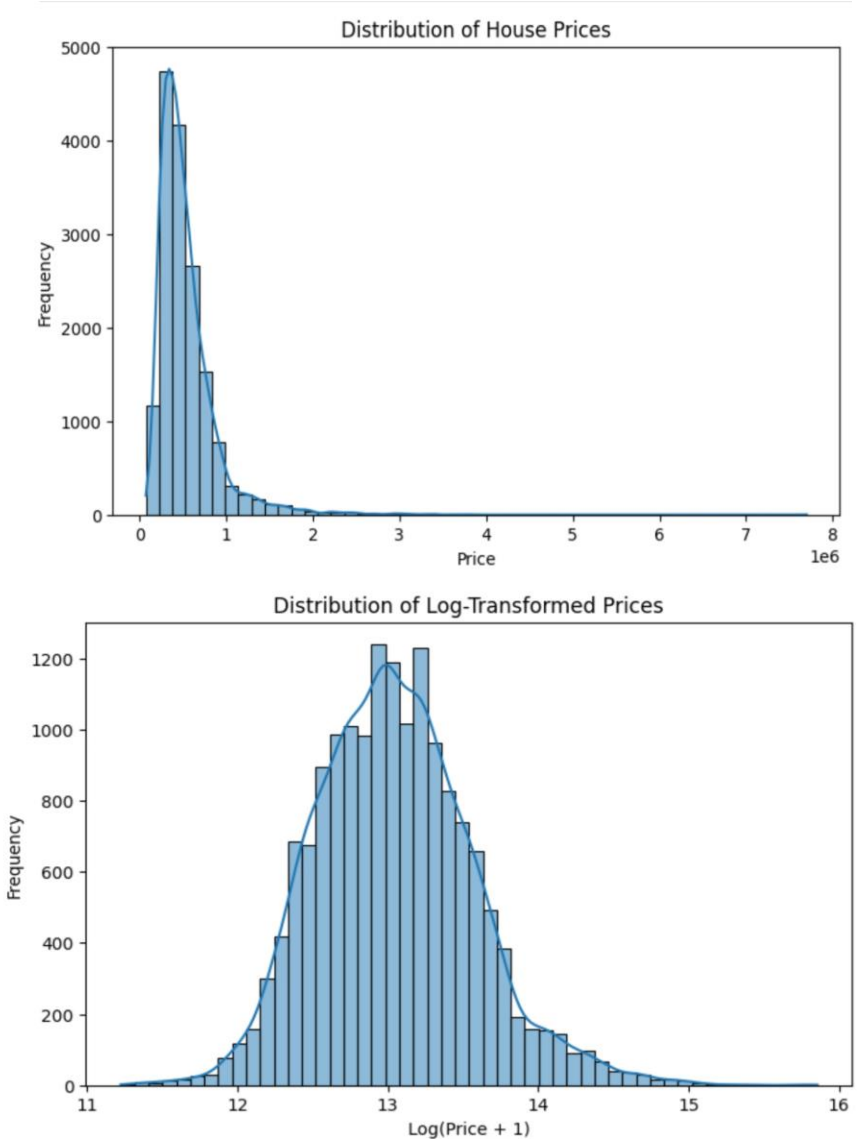
House Price Prediction Using Structured Data and Satellite Imagery

Ayush Girhay (23113038)

Distribution of Log-Transformed Prices



Graph: Histogram of `log(price + 1)`

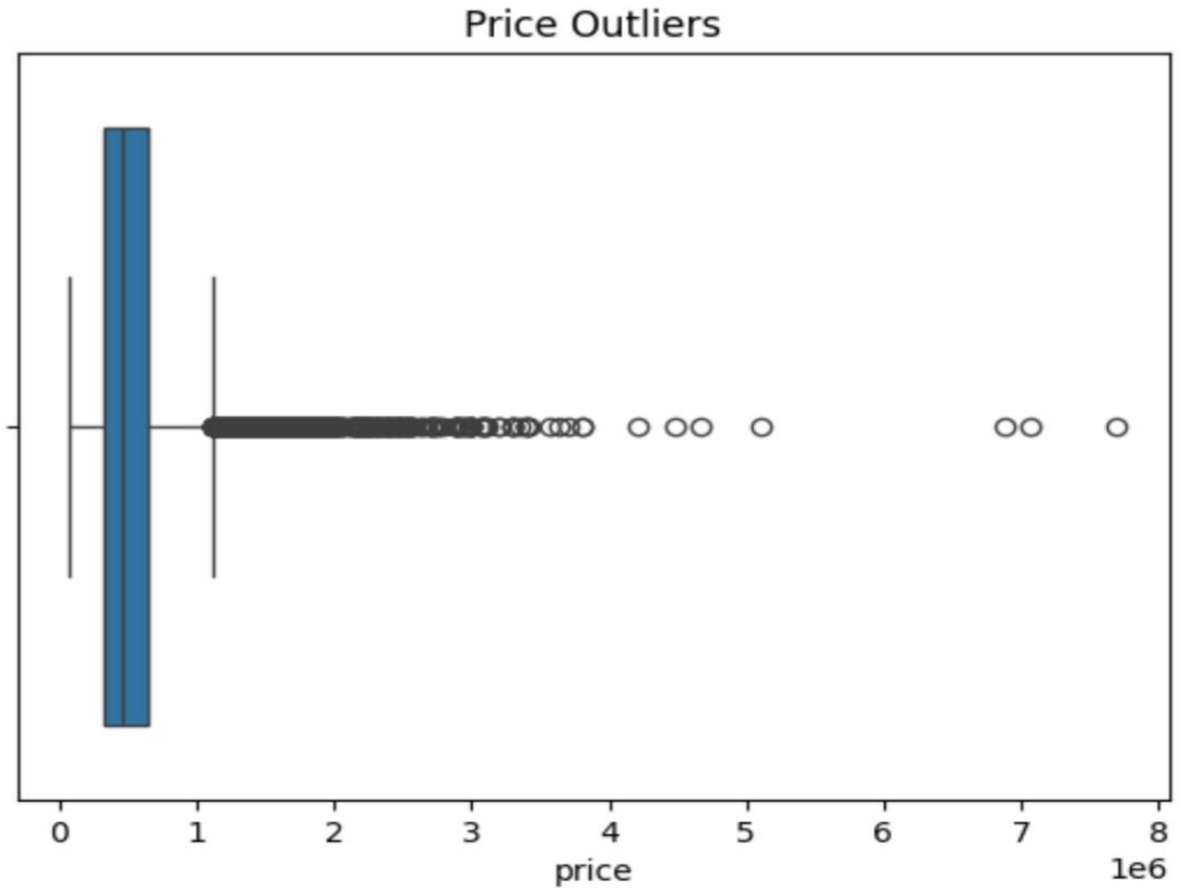
- This graph shows the distribution of house prices after applying a log transformation.
- The distribution becomes approximately normal (bell-shaped).
- Log transformation reduces the effect of extreme outliers.
- It stabilizes variance and improves linear model assumptions.

Price vs Living Area (sqft_living)

Graph: Scatter plot (`sqft_living` vs `price`)

- This plot shows the relationship between living area and house price.
- There is a positive correlation: as living area increases, price generally increases.
- Larger houses tend to be more expensive.
- Some high-priced outliers exist even at similar sizes.
- Living area is a strong predictor of house price.

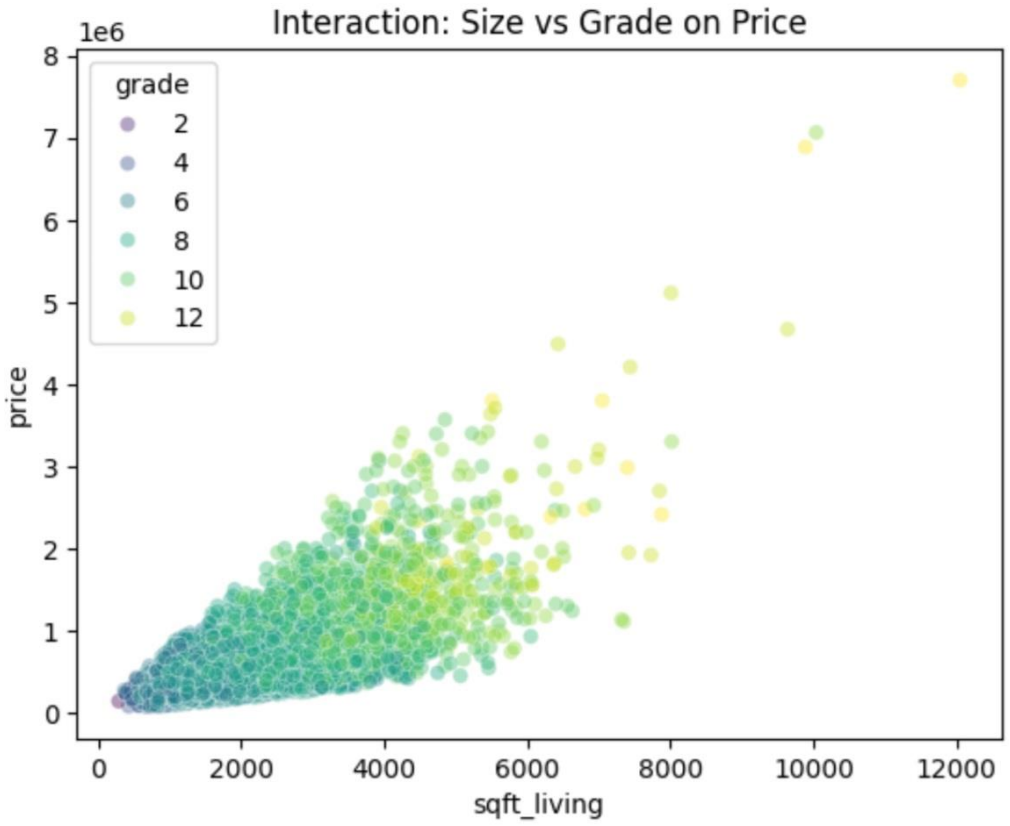
Price Outliers



Graph: Boxplot of `price`

- This plot highlights outliers in house prices.
- Most properties fall within a narrow price range.
- A few houses have extremely high prices compared to the majority.
- Outliers can distort model training and may require transformation or special handling.

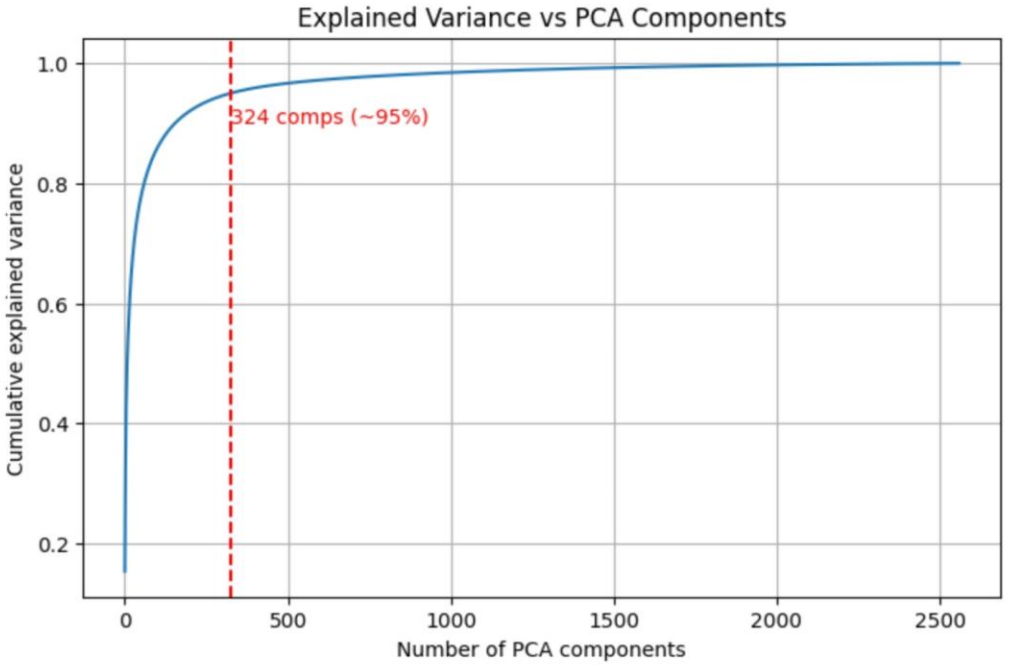
Interaction: Size vs Grade on Price



Graph: Scatter plot (`sqft_living` vs `price`, colored by `grade`)

- This graph shows how house grade interacts with size to influence price.
- Higher-grade houses generally have higher prices for the same living area.
- Large houses with high grades command the highest prices.
- House grade amplifies the effect of size on price.

Explained Variance vs PCA Components (and Metrics)

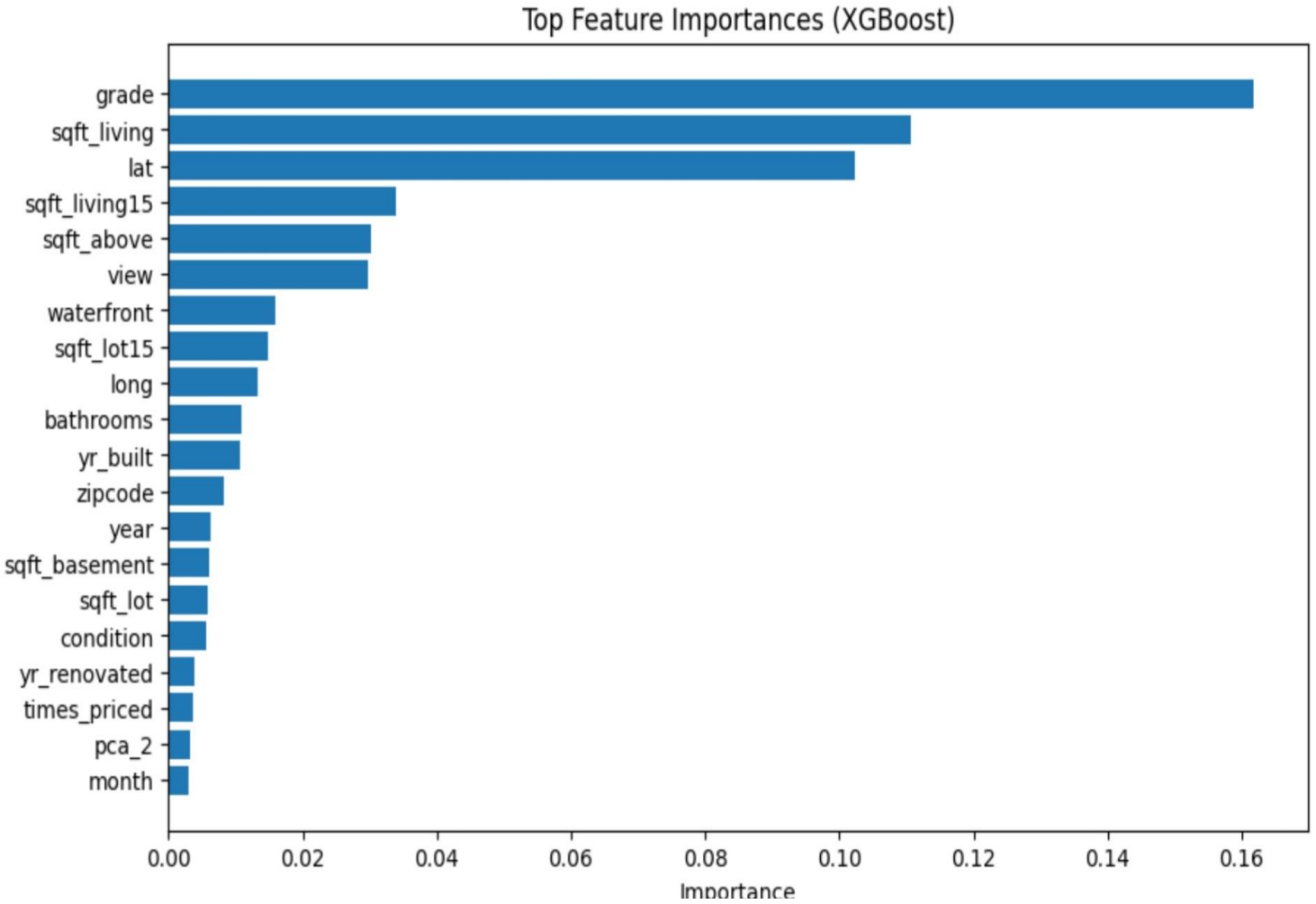


- The plot shows how **cumulative variance** increases with more PCA components.
- About **324 components capture ~95% of the total variance**.
- Beyond this point, adding more components provides minimal improvement.

Model	R-Squared
• **XGBoost (with extracted features) (with PCA) (Without Log Transformation)**	0.884
• **XGBoost (with extracted features) (with PCA) (Log Transformation)**	0.89
• **XGBoost (without extracted features) (Without Log Transformation)**	0.85
• **XGBoost (without extracted features) (Log Transformation)**	0.86

The results show that XGBoost with extracted features and PCA performs best overall, with a slight improvement when log transformation is applied (R^2 = 0.89 vs 0.884). Models without extracted features achieve lower R^2 scores, though log transformation still provides a modest gain (0.86 vs 0.85). Overall, feature extraction + PCA contributes more to performance improvement than log transformation alone.

Feature Importance (XGBoost)



- This chart shows the **most important features** used by the XGBoost model for house price prediction.
- **Grade** is the strongest predictor, followed by **living area (sqft_living)** and **location (latitude)**.
- Size-related features, neighborhood averages, and premium attributes like **view** and **waterfront** also influence prices.
- Temporal and PCA features contribute relatively less.

Conclusion

- House prices show strong dependence on structural features such as living area and grade, with higher-grade and larger houses commanding significantly higher prices.
- Log transformation of prices helps reduce the impact of outliers, improves data normality, and slightly enhances model performance.
- Outliers exist in house prices and require careful handling to avoid skewing predictions.
- Interaction between house size and grade reveals that quality amplifies the effect of size on price.
- Feature extraction from satellite imagery combined with PCA captures rich spatial information and significantly boosts model accuracy.
- Among all models tested, XGBoost with extracted features, PCA, and log-transformed prices performs best, achieving the highest R^2 score (~0.89).
- Feature importance analysis confirms that grade, living area, and location (latitude) are the most influential predictors.
- Overall, integrating structured data with visual features leads to more accurate and robust house price predictions.

Thank You