House Price Prediction Using Decision Tree Model

Introduction

The real estate market is influenced by various factors and accurately predicting the sale price of a house is crucial for buyers, sellers, and real estate professionals. In this report, we aim to develop a decision tree model to predict the sale price of a house using predictor variables such as the number of bedrooms, number of bathrooms, home square footage, lot square footage, and age of the house.

Objective

Our goal is to develop accurate prediction models and understand the regional differences in real estate values. By analyzing the relationships between the predictor variables and the sale price, we can gain insights into the factors driving real estate prices in different regions. This information can be valuable for buyers, sellers, and real estate professionals looking to make informed decisions in the housing market.

Dataset Overview

The dataset contains 10,659 rows with 12 columns containing information on house sales. The dataset has no missing values in any of the columns. Below is a summary of the columns present in the dataset:

- Record: A unique identifier for each record.
- Sale_amount: The sale price of the house.
- Sale_date: The date of the sale.
- Beds: The number of bedrooms in the house.
- Baths: The number of bathrooms in the house.
- Sqft_home: The square footage of the house.
- Sqft_lot: The square footage of the lot.
- Type: The type of house (e.g., single-family, multi-family, etc.).
- Build_year: The year the house was built.
- Age: The age of the house at the time of sale (calculated as the sale year minus the build year).
- Town: The town where the house is located.
- University: The university campus associated with the town.

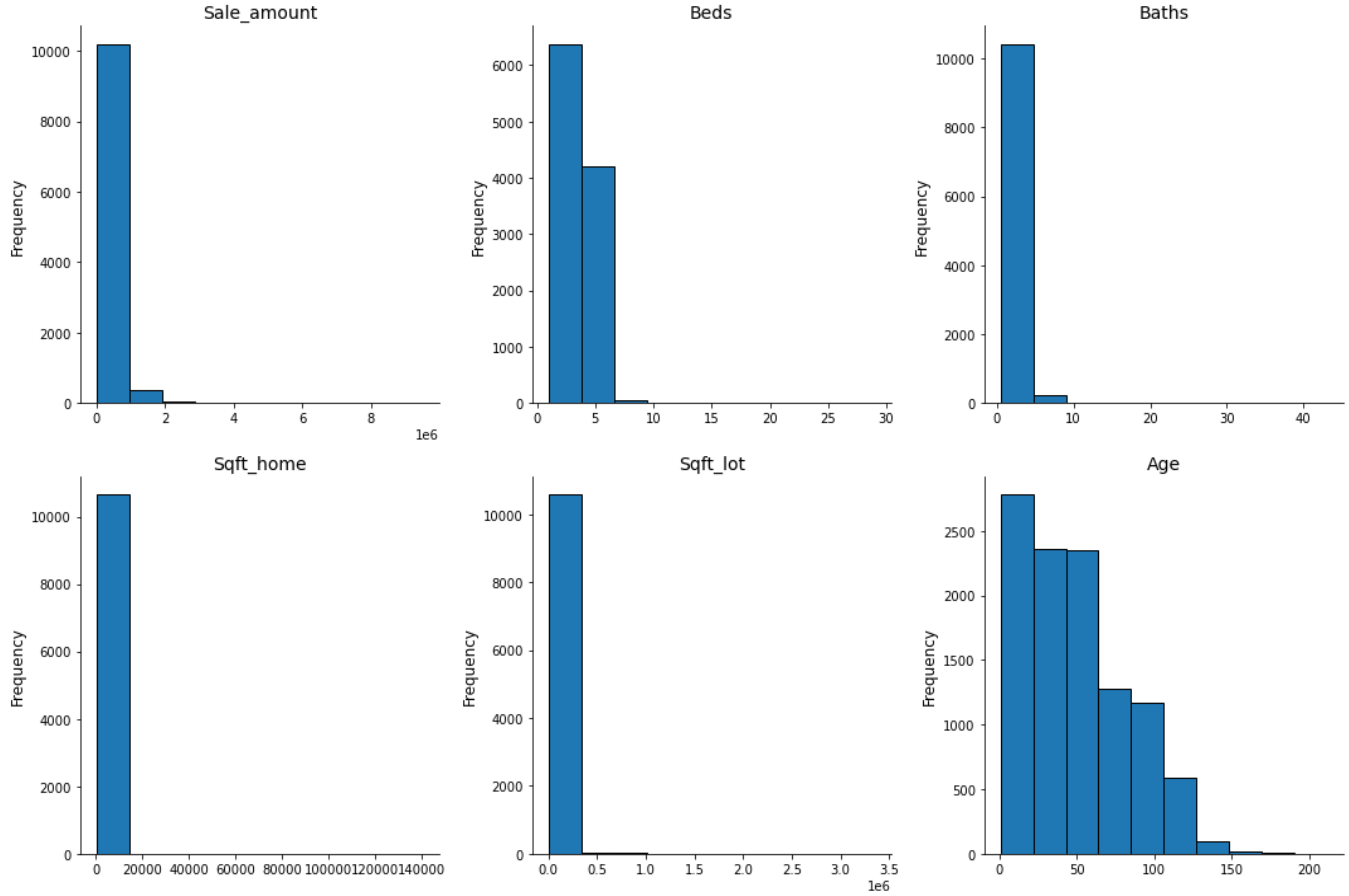
Exploratory Data Analysis

Descriptive Statistics:

Variable	N	Mean	Std Dev	Min	25%	Median	75%	Max
Sale Amount	10,659	337,556	341,301	17,000	165,975	249,000	375,000	9,500,000
Beds	10,659	3	1	1	3	3	4	29
Baths	10,659	2	1	0	2	2	3	43
Sqft Home	10,659	2,073	1,669	400	1,389	1,826	2,480	140,328

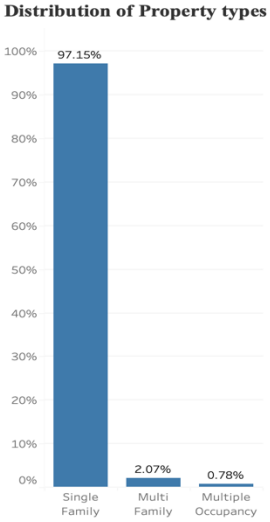
Sqft Lot	10,659	20,469	73,360	562	6,098	8,712	13,504	3,350,635
Age	10,659	49	33	1	21	44	68	211

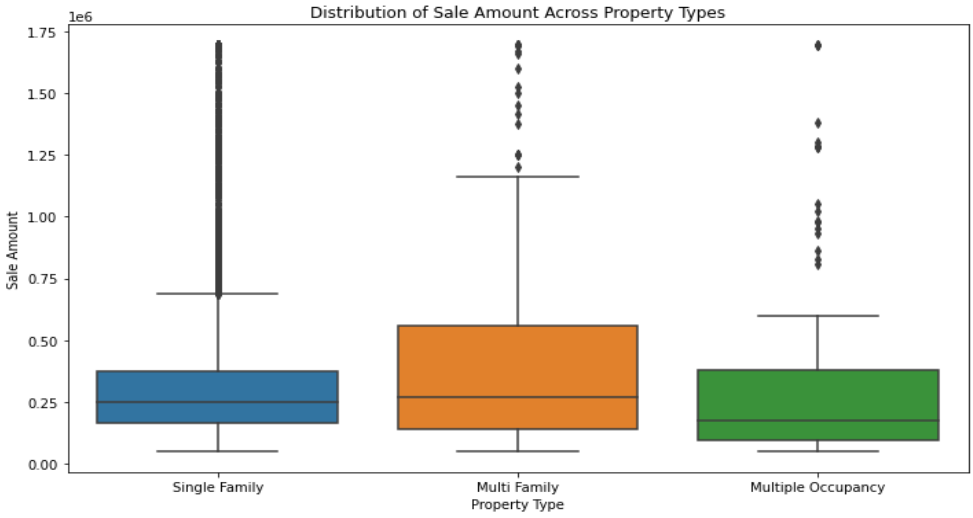
Univariate Analysis – Quantitative Variables:



The histogram charts above provide a snapshot of the distribution of quantitative variables in the dataset of property sales. The sale prices of houses vary widely, with an average of $337,556. The distribution is likely right-skewed, indicating that a few houses are sold for significantly higher prices ranging up to $9,500,000. Most properties typically have 3 to 4 bedrooms and 2 to 3 bathrooms, with some houses having unusually large numbers of beds and baths. The distribution of square footage for both home and lot size are right-skewed, varying widely in size with most properties having smaller square footage. Additionally, while most properties are relatively new, with an average age of 49 years, there are some older properties in the dataset, with an overall age range of 1 to 211 years.

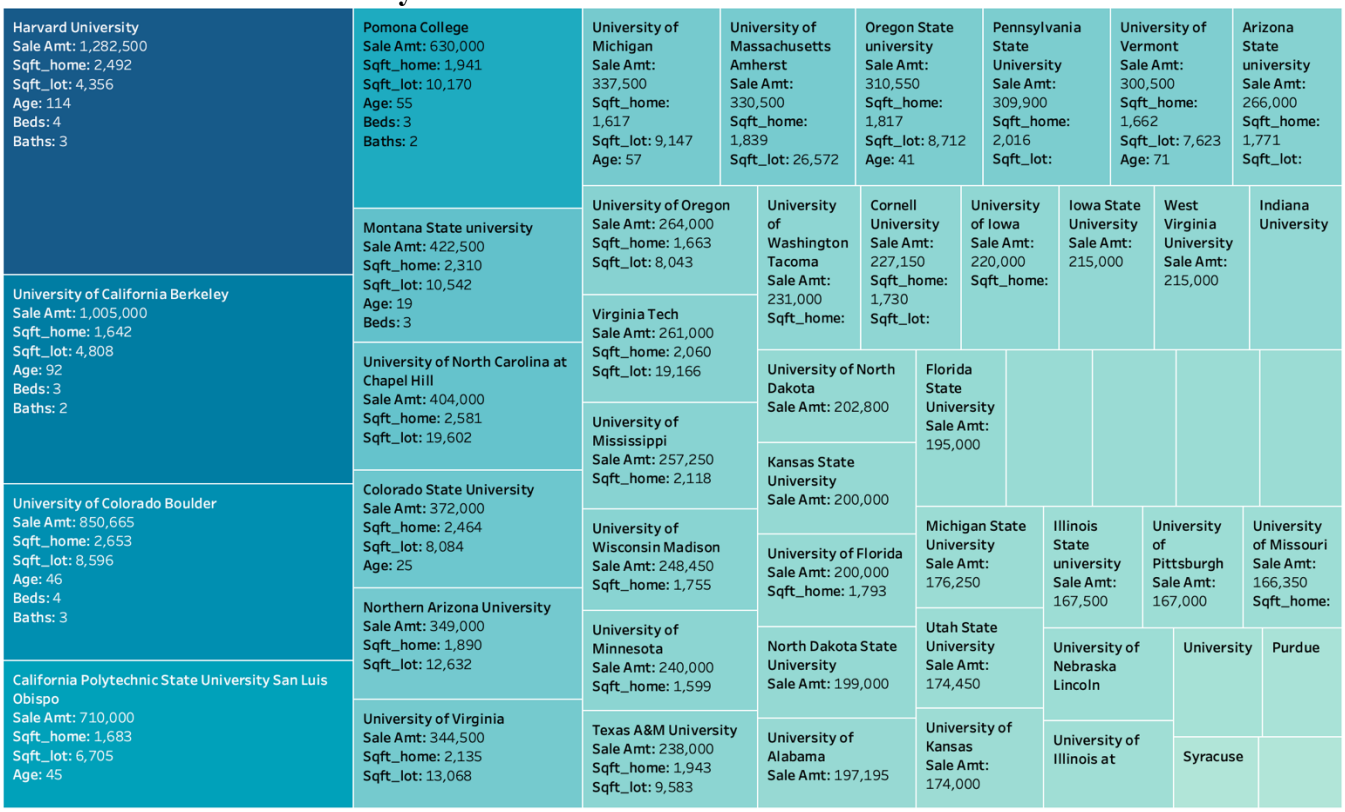
Univariate Analysis – Categorical Variables:





Single-family houses dominate the dataset, accounting for approximately 97.15% of all records. “Multi-Family" and "Multiple Occupancy" houses comprise only a small portion, with percentages of approximately 2.07% and 0.78%, respectively. The distribution of sale prices by property type indicates that while the median sale prices of single-family and multi-family properties are similar, multi-family properties have the highest median sale price and exhibit a broader spread in sale prices compared to single-family and multiple-occupancy properties.

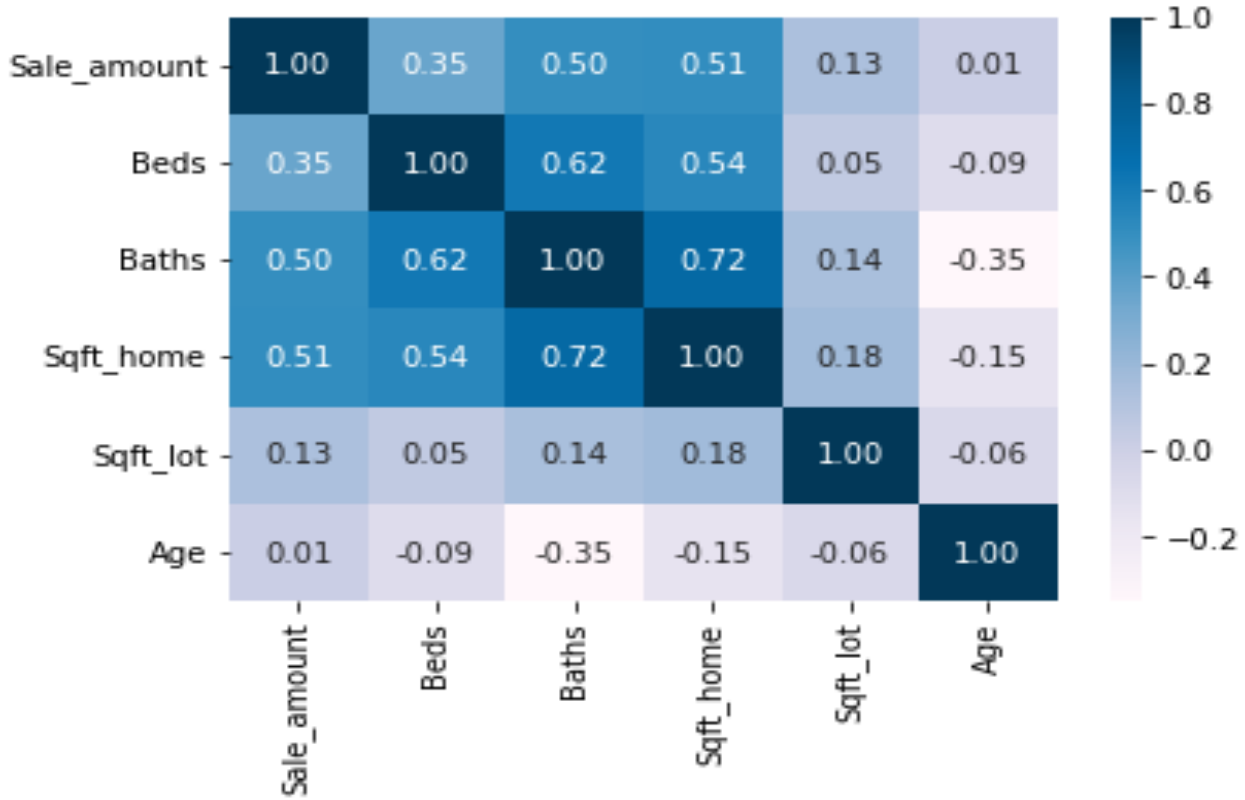
Distribution of Universities by Sale Price:



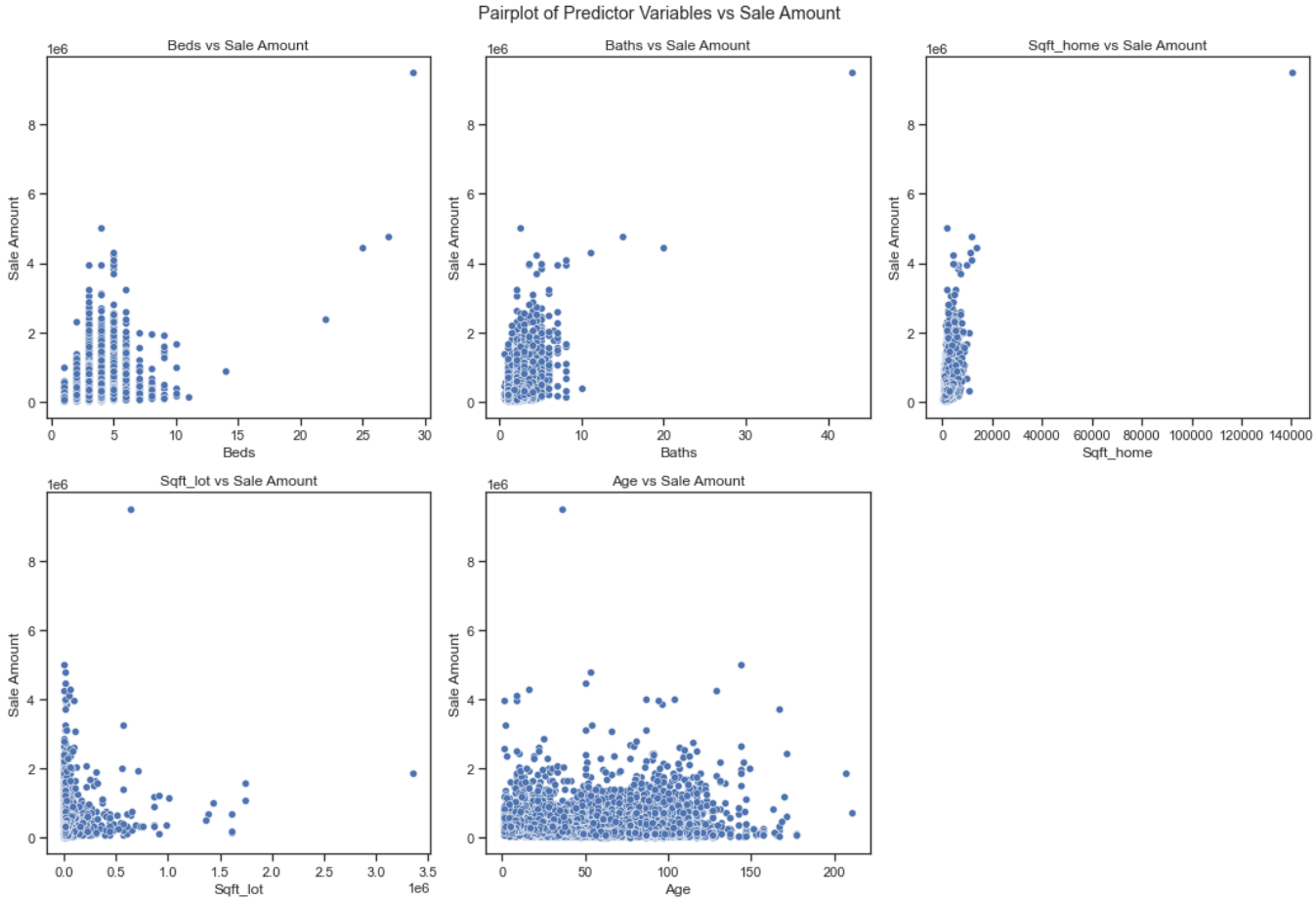
The above tree diagram shows the distribution of universities based on the median sale price. To understand the variations in the sale price, among the top universities identified – Harvard University, University of California Berkeley, University of Colorado Boulder, California Polytechnic State University, and Pomona College – it is crucial to explore the diverse house characteristics associated with each institution.

Harvard University stands out with its historically significant properties, featuring the highest median sale price and older properties with a median age of 114 years. Meanwhile, the University of California Berkeley offers relatively smaller median square footage homes, reflecting potentially more compact urban housing options. On the other hand, University of Colorado Boulder presents the largest median square footage homes, providing spacious living areas. California Polytechnic State University, with the youngest properties at a median age of 45 years, indicates recent development and modern housing options. Lastly, Pomona College distinguishes itself with the largest median lot size, offering expansive outdoor spaces. These diverse house characteristics contribute to the variations in the sale prices across the top universities.

Correlation Analysis:
The correlation plot reveals moderate positive correlations between the sale price and variables such as bedrooms, bathrooms, home square footage, and lot square footage. This indicates that larger homes with more bedrooms and bathrooms tend to command higher prices. The age of the house shows a weaker correlation, indicating that age alone may not significantly impact the sale price. Instead, factors such as amenities and overall condition might have a more significant role in determining the sale price.



Multivariate Analysis:



The pair plots above illustrate relationships between predictor variables and sale amount. It reveals positive correlations between sale amount and bedrooms, bathrooms, home square footage, and lot size. However, the age of the house shows no strong linear relationship with the sale amount, suggesting other factors have more influence on pricing.

Data Preprocessing

- **Removal of Irrelevant Data/Columns:** We removed the Record, Sale date, and Build Year columns from the dataset and streamlined the analysis by eliminating redundant or irrelevant information. The Record column, serving merely as an index, does not offer any insights for our analysis. Meanwhile, the Sale Date and Build Year columns have been replaced with a more informative attribute: the Age of the house, which directly contributes to our analysis of housing dynamics.
- **Normalization:** We utilized z-score normalization as a data transformation technique to standardize the scale of numerical features in the dataset. By doing so, we ensured that numerical features were on a comparable scale, which is essential for preventing any particular feature from dominating the learning process during model training. However, we excluded dummy variables from this normalization process to preserve their binary nature and the inherent information they provide.
- **Feature Engineering:** We converted Categorical variables (Type, Town, and University) into binary dummy variables by dropping the first level of each categorical variable as the reference

category. Additionally, given that most of our dataset consists of single-family residences (97%), we merged the records related to multi-family and multiple occupancy properties into an 'Other' category. The consolidation facilitates clearer analysis and data interpretation, focusing on the predominant property type while acknowledging alternative housing options.

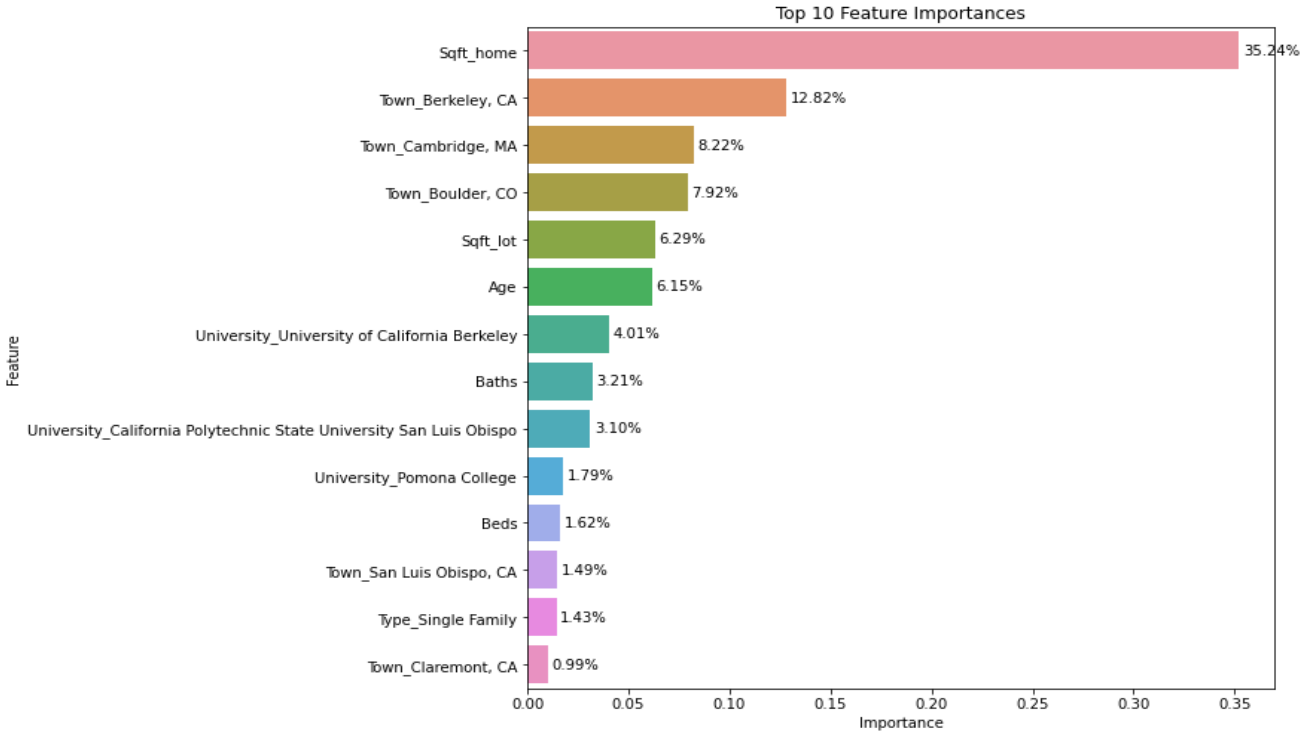
By incorporating these transformations, our dataset is now structured and compatible with machine learning algorithms, allowing for more meaningful and compelling analysis of the underlying patterns and relationships within the data.

Model Development and Evaluation

We divided our dataset into training and testing sets using a 70-30 split in the model development and evaluation phase. Initially, our model demonstrated excellent performance on the training set, achieving an R^2 score of 99%. However, this high performance did not generalize well to the testing set, where the R^2 score was only 53%, indicating overfitting. To improve the model's generalization ability, we employed GridSearchCV to fine-tune its parameters. As a result of this optimization, our model R^2 score dropped to 78% on the training dataset and improved to 70% on the testing dataset. These enhanced performance metrics indicate that our tuned model effectively balances accuracy and generalization, making it more robust and reliable for predicting house sale prices.

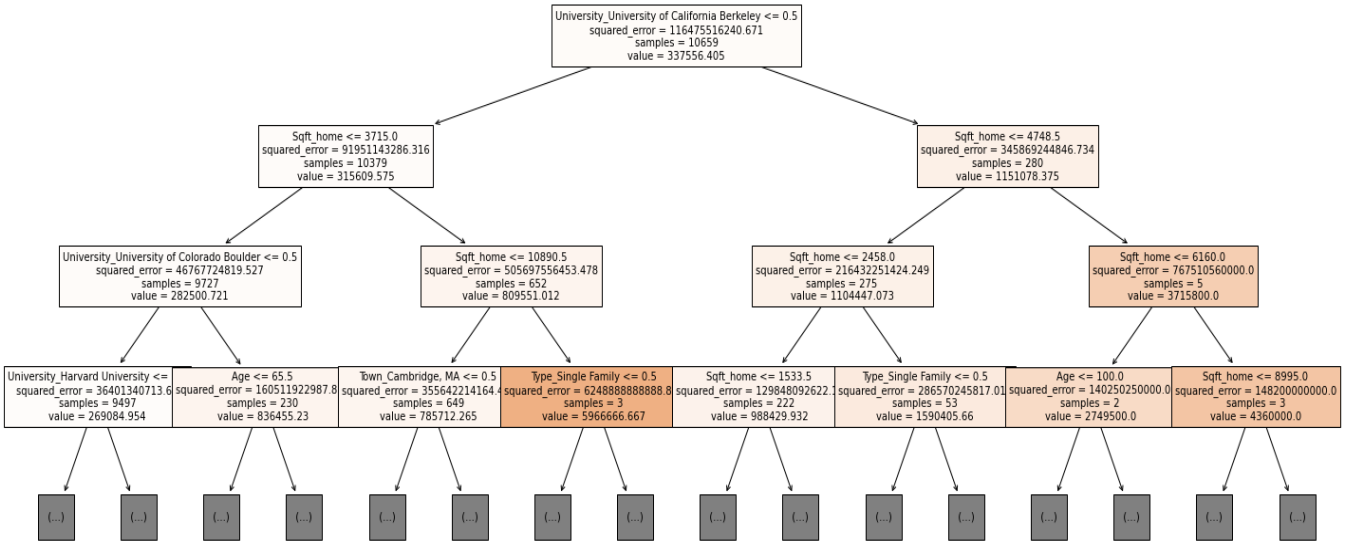
Decision Tree Regression	Initial Model	Tuned Model
R-Squared (R^2) • Training data • Testing data	 • 0.9999 • 0.5339	 • 0.7775 • 0.6969

Feature Importances:



The top 10 feature importances highlight the factors that most strongly influence the prediction of sale amounts. Among these, the size of the home stands out as the most significant predictor, indicating that larger homes tend to command higher prices. Location is also crucial, with properties in desirable areas like Berkeley, CA, and Cambridge, MA, showing higher sale prices. The lot's size and the house's age are also important considerations, suggesting that buyers value spacious properties and newer constructions. Additionally, the number of bathrooms and proximity to universities play notable roles in predicting sale prices. These findings underscore the importance of location, size, and amenities in the real estate market.

Tree Plot:



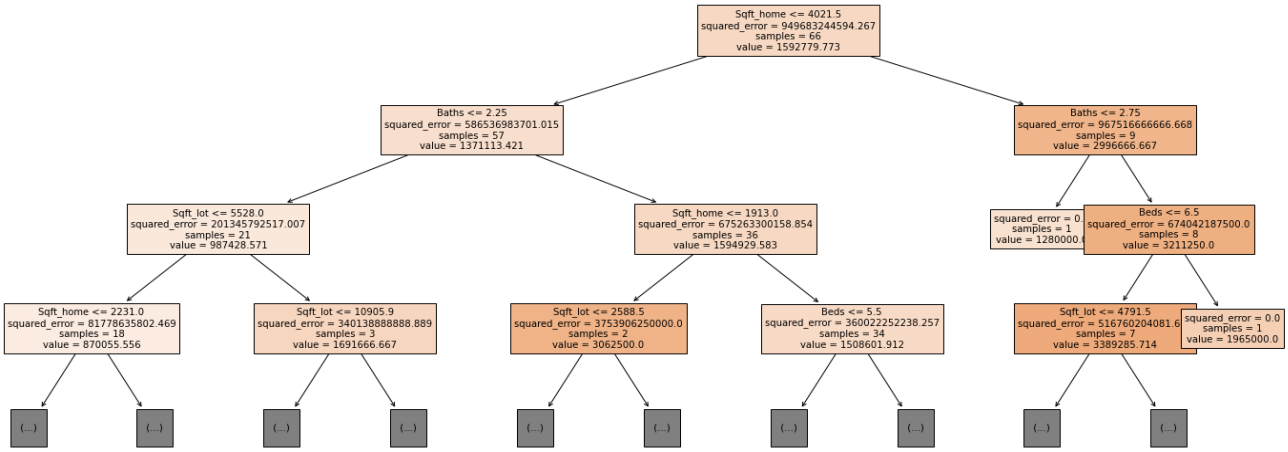
The decision tree analysis provides a clear and intuitive framework for understanding the factors influencing the sale price of houses. The tree reveals that the size of the home is the most influential factor, with larger homes generally commanding higher prices. The location together with proximity to prestigious universities also plays a significant role, with properties in certain towns, such as Berkeley CA, Cambridge MA, Boulder CO.

Upon analyzing feature importances, we observed significantly higher importance values for the same five universities/locations identified in our previous analysis. Recognizing the significance of these universities in influencing house sale prices, we made the strategic decision to develop a separate decision tree model focused exclusively on these locations.

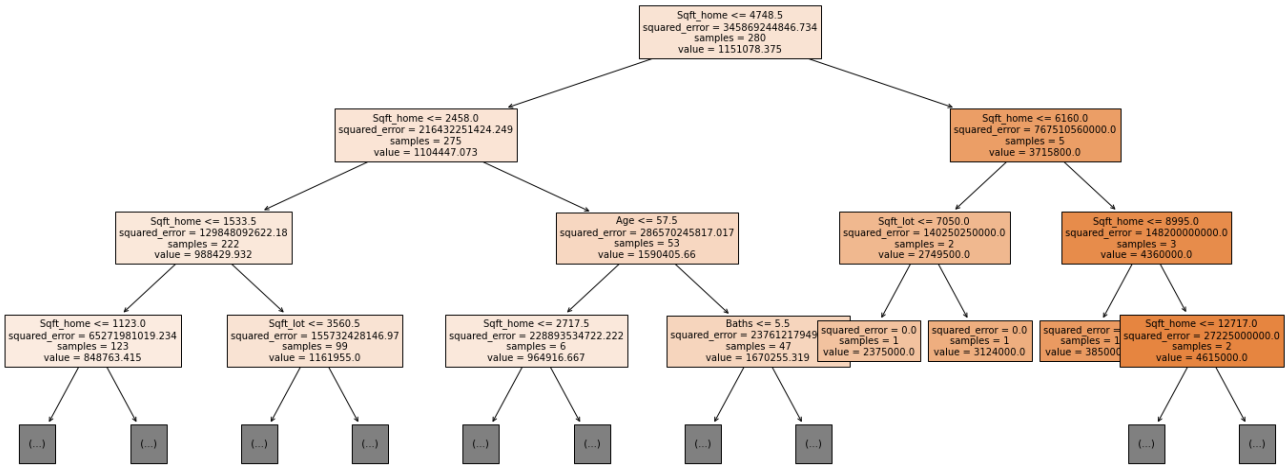
University-Specific Decision Tree Analysis:

By focusing our analysis on individual universities, we aim to gain deeper insights into the variability of sale prices across different educational institutions. This targeted approach will allow us to uncover unique factors driving housing market dynamics in proximity to these prominent universities, ultimately enhancing our understanding of regional real estate trends.

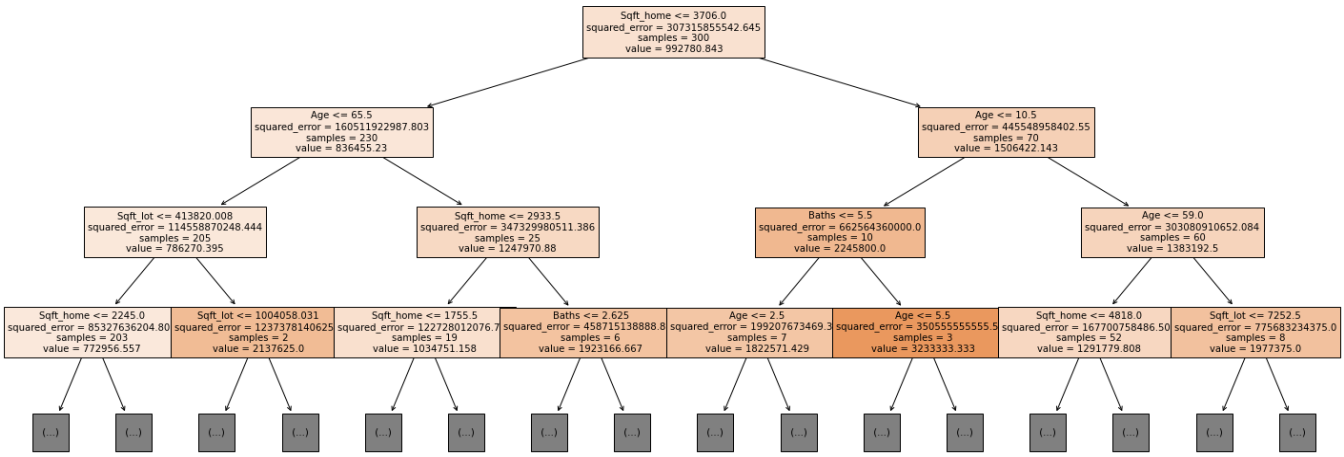
Decision Tree for Harvard University

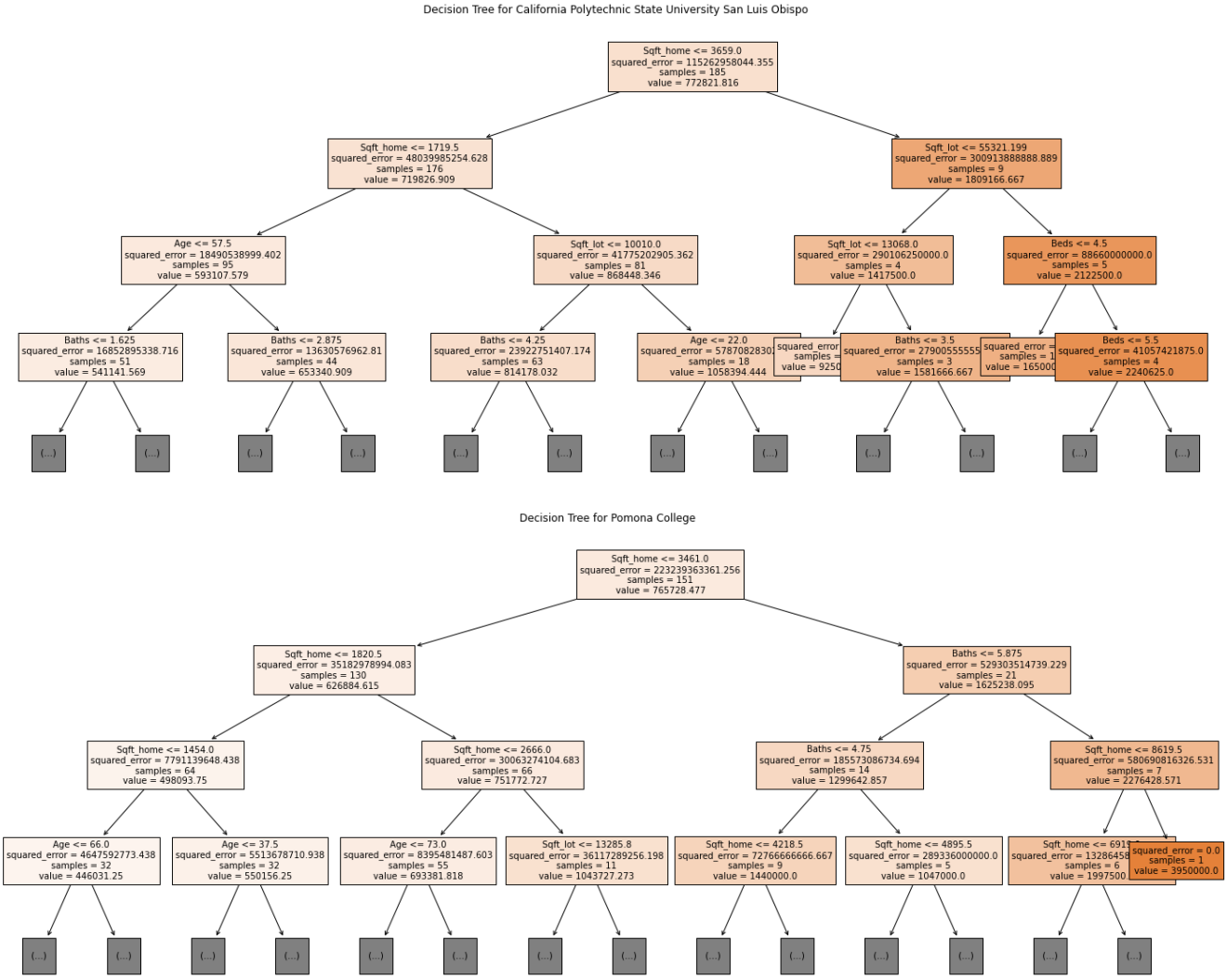


Decision Tree for University of California Berkeley



Decision Tree for University of Colorado Boulder





The above tree plots highlight that the home's square footage (Sqft_home) emerges as the most influential predictor variable across all universities studied. This suggests that the size of the property significantly impacts its sale price regardless of the university's location. However, a distinct variation is observed regarding the role of the property's age as a predictor. While age significantly determines sale prices in most university areas, its importance diminishes notably for Harvard University. Additionally, other key predictors observed consistently across the universities include the number of bedrooms (beds), number of bathrooms (baths), and square footage of the lot (sqft_lot).

Actionable Insights and Recommendations

In conclusion, the decision tree analysis provides valuable insights into the factors influencing the sale price of houses. Based on the outcome of the analysis of the most important features of the decision tree model, we arrived at the following conclusions:

- House square footage (Sqft_home) is a key indicator of house price. Hence, we highly recommend that buyers prioritize properties with larger square footage, as these tend to have higher resale values. Sellers should consider investing in expanding or renovating existing properties to increase square footage and potentially command higher prices.

- Location matters. Properties near prestigious universities or in desirable neighborhoods tend to command higher prices. Therefore, buyers and sellers should carefully consider the location when making real estate decisions.
- Age doesn't tell the whole story. Newer properties may not always fetch higher prices if they lack desirable features or are in less desirable locations. Conversely, older properties may still command high prices if well-maintained and located in prime areas.
- Consider Lot Size. The square footage of the lot (Sqft_lot) also plays a role in determining sale prices, albeit to a lesser extent than home size. Properties with larger lots may appeal to buyers seeking outdoor space or potential for expansion, thus influencing their perceived value.
- There is a strong preference for Single-family residences among buyers, likely due to factors such as privacy, space, and ownership autonomy.

Our analysis also reveals potential investment opportunities in areas with high demand for housing, such as towns with renowned universities. Investors may consider targeting properties in these areas for rental or resale, leveraging the consistent demand and potentially higher sale prices associated with proximity to educational institutions. Lastly, it is essential to consider current market trends and dynamics when making real estate investment decisions. Factors such as supply and demand, interest rates, and economic conditions can influence property values and should be carefully monitored to make informed investment choices.