Housing Price Prediction

End-to-End Machine Learning Project

Data Science with Python — Final Project

Student: **Nini Koiava**

Dataset: California Housing Dataset (sklearn)

ML Approach: Regression — Linear Regression vs Decision Tree

February 2026

Contents

Project Overview

Goal: The goal of this project is to predict the median house value of California census block groups using demographic and geographic features. Additionally, the project aims to identify which variables most strongly influence housing prices.

Dataset:

Source: California Housing Dataset (1990 census), built into sklearn.datasets
Records: 20,640
Features: 8 input + 1 target
Target: MedHouseVal ($100,000s)

Learning Objectives Demonstrated

This project demonstrates competency in four key areas:

1. Data Processing

- Reproducible preprocessing pipeline
- Clear documentation of transformations
- Proper train/test separation to prevent data leakage

2. Exploratory Data Analysis

- 8 visualizations
- Correlation analysis
- Distribution analysis and skewness measurement

3. Machine Learning

- Two regression models trained and evaluated
- 5-fold cross-validation
- Model comparison using multiple metrics

4. Code Quality

- Modular functions
- Docstrings and comments
- PEP 8 compliance
- Error handling

Data Processing Pipeline

Dataset Overview

Dataset Quality Assessment

- Duplicate rows: **0**
- Missing values: **0**
- Dataset is fully complete (20,640 observations)

Feature	Type	Missing	Min	Max
MedInc	float64	0	0.50	15.00
HouseAge	float64	0	1.00	52.00
AveRooms	float64	0	0.85	141.91
AveBedrms	float64	0	0.33	34.07
Population	float64	0	3.00	35,682
AveOccup	float64	0	0.69	1,243.33
Latitude	float64	0	32.54	41.95
Longitude	float64	0	-124.35	-114.31
MedHouseVal	float64	0	0.15	5.00

Table 3: Data quality report: all 20,640 records complete, no missing values

Cleaning Steps

The California Housing dataset contains no missing values; however, a robust and reusable preprocessing pipeline was implemented to ensure reproducibility and best practices for future datasets.

1. **Duplicate Removal**
   All rows were checked for duplication.
   - Result: **0 duplicate rows found and removed**
2. **Missing Value Flagging and Imputation**
   Each column is scanned for missing values. If missing values are detected, a binary indicator column is created and values are imputed.
   - Numerical features → Median imputation
   - Categorical features → Mode imputation
   - Result: **0 missing values in the dataset** (imputation logic retained for robustness)
3. **Data Type Conversion**
   All object-type columns are converted to categorical data types to ensure consistency and memory efficiency.
4. **Derived Feature Creation**
   A derived numerical feature is created using ratios of existing numeric columns to capture additional relational information between variables.
5. **Outlier Detection and Capping (IQR Method)**
   Outliers are handled using the Interquartile Range (IQR) method:

   - Lower bound = Q1 − 1.5 × IQR
   - Upper bound = Q3 + 1.5 × IQR
     All numeric features are clipped within these bounds.
   - Result: **0 extreme outliers detected; dataset already well-behaved**
6. **Feature Engineering (Modeling Stage)**
   Additional meaningful features are created to enhance predictive power:
   - RoomsPerPerson
   - BedroomsPerRoom
   - PopulationPerHH
   - NewHome (binary indicator: houses younger than 20 years)
7. **Feature Scaling**
   Numerical features used for Linear Regression are standardized using **StandardScaler**.
   - Scaler is fitted on the training set only
   - Same transformation applied to the test set
   - Prevents **data leakage**

Exploratory Data Analysis — Key Findings

Statistical Highlights

- **Median Income (MedInc)** shows the strongest correlation with house value (**r = +0.688**), making it the most influential single predictor.
- **Geographic location** (Latitude and Longitude) has a strong effect on prices, with higher values concentrated near coastal regions.
- Several features exhibit severe right-skewness, indicating the presence of extremely dense or unusual neighborhoods.
- The target variable (**MedHouseVal**) is **moderately right-skewed** (skewness = 0.978).
- House values are **capped at $500,000**, which creates artificial clustering at the upper bound.

Key Insights

1. **Income dominates:** MedInc has a correlation of **0.688** with house value — by far the strongest single predictor.
2. **Geography matters:** The geographic scatter plot clearly shows that coastal properties (Bay Area and Los Angeles) command significant price premiums.
3. **Severe skewness:** AveOccup and AveBedrms are extremely right-skewed (skewness **97.6** and **31.3**, respectively), indicating a small number of unusually dense blocks.
4. **Target distribution:** House values are right-skewed (skewness = **0.978**) with a hard upper cap that introduces artificial clustering.
5. **Income-group analysis:** Violin plots confirm that high-income neighborhoods ($60k+) exhibit both higher median values and wider price distributions.

Feature	Mean	Std	Median	Skewness	Corr w/ Target
MedInc	3.871	1.900	3.535	1.647	**+0.688**
HouseAge	28.64	12.59	29.00	0.060	+0.106
AveRooms	5.429	2.474	5.229	20.70	+0.152
AveBedrms	1.097	0.474	1.049	31.32	—
Population	1425	1132	1166	4.936	—
AveOccup	3.071	10.39	2.818	97.64	—
Latitude	35.63	2.136	34.26	0.466	-0.144
Longitude	-119.6	2.004	-118.5	-0.298	—
MedHouseVal	2.069	1.154	1.797	0.978	1.000

Table 4: Descriptive statistics and target correlations

Machine Learning Models

Model Configuration

Parameter	Linear Regression	Decision Tree Regressor
Train/Test split	80/20	80/20
Random state	42	42
Feature scaling	StandardScaler	Not required
Max depth	N/A	8 (prevent overfitting)
Min samples split	N/A	20
Min samples leaf	N/A	10
Cross-validation	5-fold	5-fold

Table 5: Model hyperparameters

Results

Model	R^2 Score	MSE	RMSE	MAE
Linear Regression	0.6755	0.4103	0.6405	0.4687
Decision Tree Regressor	**0.6742**	**0.4119**	**0.6418**	**0.4488**

Table 6: Test set performance

Model Interpretation

Linear Regression ($R^2 = 0.6755$)
The Linear Regression model explains approximately **67.6% of the variance** in median house values. It achieves an RMSE of **0.6405**, meaning the average prediction error is about **$64,000**. Cross-validation results (**$R^2 = 0.6947 \pm 0.0080$**) indicate strong and stable generalization performance.

The model's coefficients show that **median income (MedInc)** is the most influential predictor, confirming the strong relationship between income levels and housing prices. Positive effects from **AveRooms**, **RoomsPerPerson**, and **NewHome** suggest that larger and newer homes tend to be more valuable, while geographic features (**Latitude and Longitude**) capture the coastal price premium.

Decision Tree Regressor ($R^2 = 0.6742$)
The Decision Tree Regressor explains approximately **67.4% of the variance**, with an RMSE of **0.6418** and MAE of **0.4489**. Its cross-validation performance (**$R^2 = 0.6954 \pm 0.0158$**) is comparable to Linear Regression, indicating good generalization.

The Decision Tree captures **non-linear relationships** between features and house prices, allowing it to model complex interactions. Feature importance rankings highlight **MedInc**, **Latitude**, and **Longitude** as dominant predictors, consistent with patterns observed during exploratory analysis.

Overall Comparison
Both models perform similarly; however, **Linear Regression achieves a slightly higher R^2 score and lower RMSE**, making it the **best-performing model** in this project. Additionally, Linear Regression offers greater interpretability, making it more suitable for explaining how individual features influence house prices.

Feature Engineering

Four new features were derived from existing columns to improve model expressiveness:

Feature	**Formula**	**Rationale**
RoomsPerPerson	AveRooms / AveOccup	Captures housing spaciousness per resident
BedroomsPerRoom	AveBedrms / AveRooms	Ratio reflecting property type
PopulationPerHH	Population / AveOccup	Neighborhood density measure
NewHome	HouseAge < 20 (binary)	Separates modern from older housing stock

Table 8: Engineered features and their rationale

Conclusions

Summary

This project developed an end-to-end machine learning pipeline for predicting California housing prices. A robust preprocessing workflow, extensive exploratory data analysis, and two regression models were implemented and evaluated. Linear Regression achieved the best overall performance (**R^2 = 0.6755**), with Decision Tree Regressor producing comparable results. The findings confirm that **income and geographic location are the strongest determinants of housing value**, while engineered features further improved model expressiveness.

Limitations

Data is from 1990 and may not reflect modern market dynamics

Neither model includes hyperlocal factors such as school quality or crime rates

The $500,000 price cap in the original data creates artificial clustering at the upper end

Future Work

Implement Random Forest or Gradient Boosting (expected $R^2 > 0.80$)

Enrich dataset with school ratings, walkability scores, and crime indices

Deploy best-performing model as a Streamlit web application