Housing Price

Prediction

Data Science with Python — Final Project

Student: Nini Koiava

Dataset: California Housing (sklearn) | February 2026

Regression: Linear Regression vs Decision Tree Regressor

Project Overview

20,640
Records

8
Features

2
ML Models

Goal: Predict median house values for California census block groups and identify which factors most influence housing prices.

📊 Data Processing — reproducible cleaning pipeline

🔍 EDA — 8 visualizations, statistical summaries, correlations

💻 Code Quality — docstrings, modular functions, PEP 8

🤖 Machine Learning — 2 regression models trained & evaluated

Data Processing & Cleaning

Feature Overview

MedInc	Median income ($10k)	0.69
HouseAge	Median house age (yrs)	0.11
AveRooms	Avg rooms per household	0.15
AveBedrms	Avg bedrooms/household	—
Population	Block group population	—
AveOccup	Avg household occupancy	—
Latitude	Block group latitude	-0.14
Longitude	Block group longitude	—

Corr = target correlation

Cleaning Pipeline

1. **Duplicate Removal**

   0 duplicates found. Pipeline ready for any dataset.

2. **Missing Value Handling**

   0 missing values. Median imputation strategy built-in for robustness.

3. **Outlier Detection**

   IQR × 3 method on target. 0 extreme outliers removed.

4. **Feature Engineering**

   4 new features: RoomsPerPerson, BedroomsPerRoom, PopulationPerHH, NewHome

5. **Feature Scaling**

   StandardScaler fit on train set only — no data leakage.

✅ 0 missing values | ✅ 0 duplicates | ✅ 0 outliers removed | ✅ 4 new features engineered | ✅ No data leakage

Exploratory Data Analysis — Key Findings

💰 Income is #1 Predictor

MedInc has correlation of 0.688 with house value — by far the strongest single signal. Linear and non-linear models agree.

🗺️ Location Drives Premium

Coastal areas (Bay Area, LA) show dramatically higher values. Geographic scatter map confirms clear spatial clustering.

📈 Right-Skewed Features

AveOccup (skew = 97.6) and AveBedrms (31.3) are extremely skewed — a few unusually dense blocks distort distributions.

🎻 Income Group Effect

Violin plots show High Income groups (6+) have higher AND more variable house values than low-income areas.

📊 8 visualizations created (required: 5) — Histogram, KDE, Box Plot, Heatmap, Scatter, Geo Map, Violin, Actual-vs-Predicted

Live Demo

Step 1	**Run import_data & data_preprocessing**	Load data, show quality report, run preprocess_data()
Step 2	**Run visualization**	Show all 8 visualizations — geographic map, violin plots, correlation heatmap
Step 3	**Run machine_learning**	Feature engineering — 4 new derived features
Step 4	**Run machine_learning**	Train Linear Regression — show R^2, RMSE
Step 5	**Run machine_learning**	Train Decision Tree — show R^2, RMSE

Machine Learning Results

Linear Regression

0.6755

R^2 Score

RMSE	0.6405
MSE	0.4103
MAE	0.4687
CV R^2	0.6947 ± 0.080

🏆 Decision Tree

0.6742

R^2 Score

RMSE	0.6718
MSE	0.4489
MAE	0.4489
CV R^2	0.6954 ± 0.0158

🏆 Linear Regression achieves the highest R^2 and lowest RMSE, while both models perform very similarly overall

Conclusions & Future Work

Key Takeaways

- Income is the single strongest driver of house value (r = 0.688)
- Coastal geography commands a measurable price premium
- Linear Regression achieves the highest R^2 and lowest RMSE, while both models perform very similarly overall.
- Cross-validation confirms both models generalize well

⚠ Limitations

- 1990 census — not current market
- Hard cap at $500k distorts top end

Future Work

- 🌲 Random Forest / Gradient Boosting — expected $R^2 > 0.80$
- 🏫 Add school ratings, crime index, walkability scores as features
- 🌐 Deploy as Streamlit web app for real-time price prediction

Final Result: Linear Regression $R^2 = 0.6755$