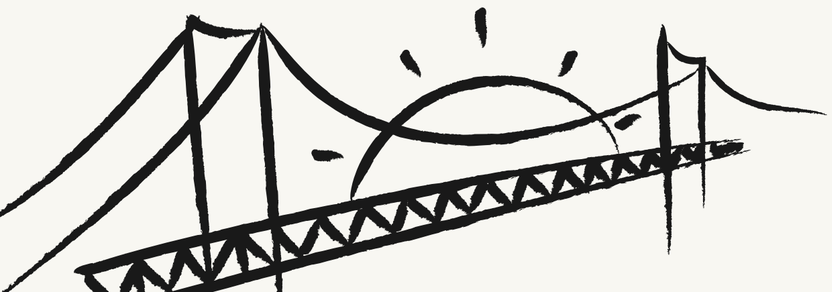
California Housing Price Prediction:

Machine Learning Regression for Real Estate Market Analysis & Investment Strategy

Medina Alifia Juniarto

Capstone Project 3 - Machine Learning

December, 2025



Discussion Flow

01
Dataset Overview

02
Exploratory Data Analysis (EDA)

03
Modeling

04
Conclusion

05
Recommendation

DATASET OVERVIEW

Executive Summary

1 Background

The dataset is limited to the **1990 California Census** information. To understand factors, influence California Housing Price.

2 Problem Statement

House prices are **hard to predict** because locations and conditions vary greatly. The market also changes quickly, **making traditional appraisal methods less reliable** during sudden price shifts.

3 Basic Statistics

Records
14.448

Features
10

Period
1990

4 Project Workflow

- **Business Problem Identification**
- **Data Understanding**
- **Data Cleaning**
- **Exploratory Data Analysis**
- **Feature Selection**
- **Feature Engineering**
- **Analytics Work (Deploy ML: Algorithm, Metrics)**
- **Limitations & Recommendations**

5 Stakeholders

Secondary:

- Real Estate Agents
- Governments

DATASET OVERVIEW

Data Preprocessing

1 Data Collection

data_california_house.csv
14448 rows, 10 columns

2 Data Preprocessing

Observe Data Types

Observe the statistics description
float(9), object(1)

Check & Handle Missing Value

total_bedrooms
137 missing values

Check & Handle Duplicated Value

0 Duplicated value
(No action)

Generate New Columns

Feature Engineering
(income_house_ratio, income_category, rooms_per_household, bedrooms_per_room, population_per_household)

Feature Selection

Drop multicollinearities columns
(corr. 0.8 - 0.9)

New Data

14448 rows, 8 columns

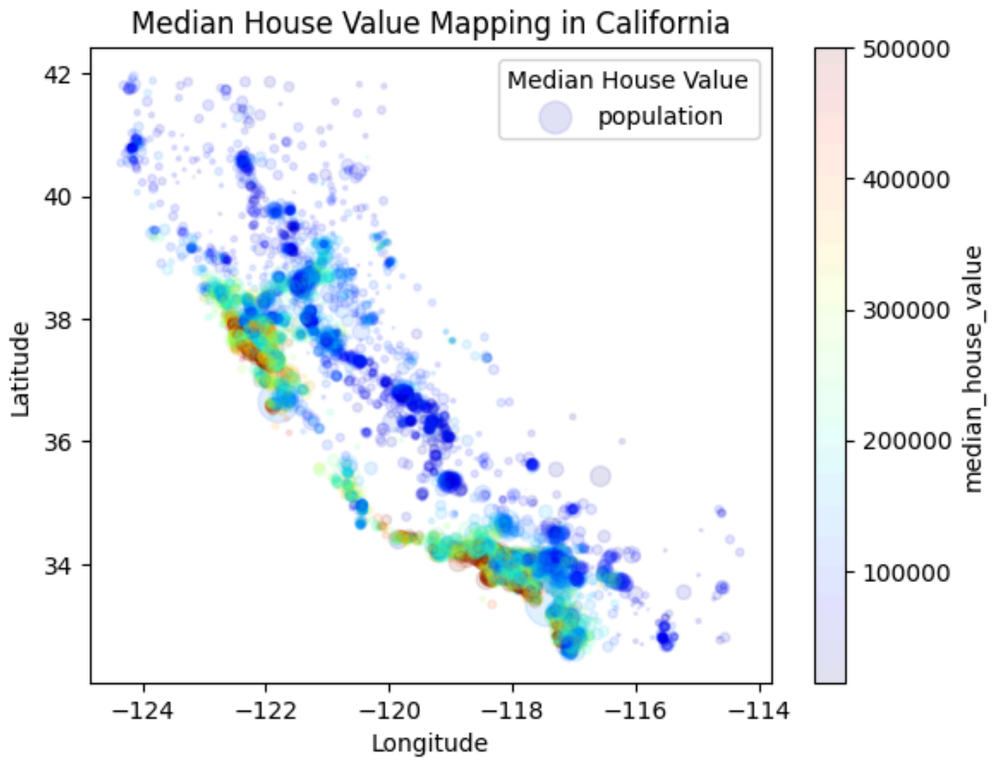
3 Data Analysis

Visual Studio Code matplotlib seaborn

EXPLORATORY DATA ANALYSIS

Target and Feature

Median House Value vs. People Density & Location:
Where People Pack, Prices Peak.



1. **Coastal Areas** (lower longitude, closer to the western coast) represent with (red/orange) colors have the highest median house value between $400,000 - $500,000 and has larger bubble sizes indicating **higher population density.**

2. **Value drops rapidly moving eastward** (higher longitude, shifting to inland areas where (blue/green) colors is <$200,000 with **scattered population density.**

Critical Takeaway:

Every degree of eastward movement = There are numbers of price reductions and vary densities. It's California's economic stratification mapped in coordinates.

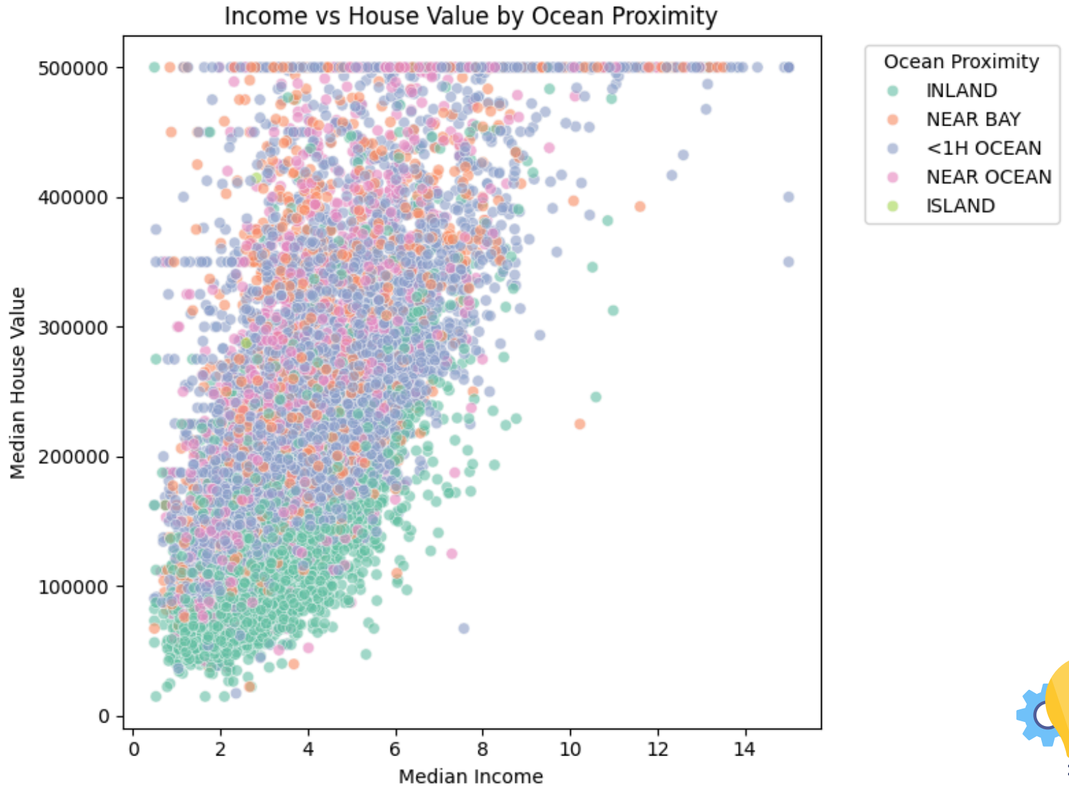
Property values in California clearly decline as distance from the coast increases, creating a gradient that shapes investment strategy, rental potential and development planning.

EXPLORATORY DATA ANALYSIS

Target and Feature

Median Income vs Median House Value by Ocean Proximity

Where You Earn Matters, but Where You Live Matters More



1 Income Drives House Affordability

- Higher income → higher house price.
- **Income is a strong predictor but not the only one.**

2 Ocean Proximity Adds a Location Premium

- Homes near the ocean, bay, and <1H Ocean consistently have higher prices even with similar income levels.
- Dark green dots (INLAND) cluster in the lower price range.

Critical Takeaway:

- **Income determines buying power.**
- **Location acts as a price multiplier.**

Coastal proximity can significantly raise home values regardless of income

Differentiate the strategy for **Real Estate Agents and Governments.**

- Segmenting to the high-income level, near ocean > High ROI (Real Estate Agents)
- Governments : Focus on Urban Planning (Identify the Undervalued Residency)

FEATURES

TRANSFORMATION

Scaler

All features are numerical and cleaned. StandardScaler improves learning by equalizing feature scales. The 500,000 USD capped values are retained due to potential importance and limited information.

MODELING

First Step: Benchmarking

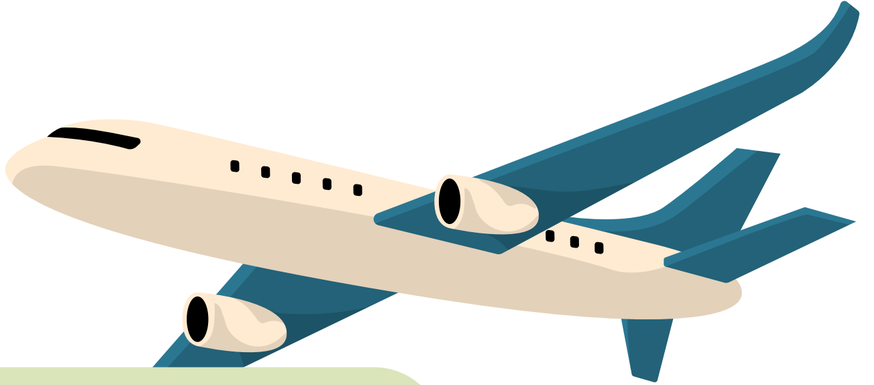
No.	Model	Avg MAPE	Std MAPE	Avg RMSE	Std RMSE	Avg R2 Score	Std R2 Score
1	LightGBM	18.58%	46%	$49.39	$293	0.8172	0.005
2	KNeighborsRegressor	22.51%	53%	$61.23	$773	0.7189	0.010
3	RandomForestRegressor	30.18%	94%	$67.21	$931	0.6613	0.013
4	LinearRegression	31.03%	88%	$73.90	$4.36	0.5883	0.060
5	DecisionTreeRegressor	32.07%	69%	$72.38	$1.19	0.6072	0.017
6	SVR	52.51%	159%	$118.46	$2.11	-0.0512	0.015

CV = 5

Linear vs. Non Linear Performance

MODELING

Algorithm: LightGBM

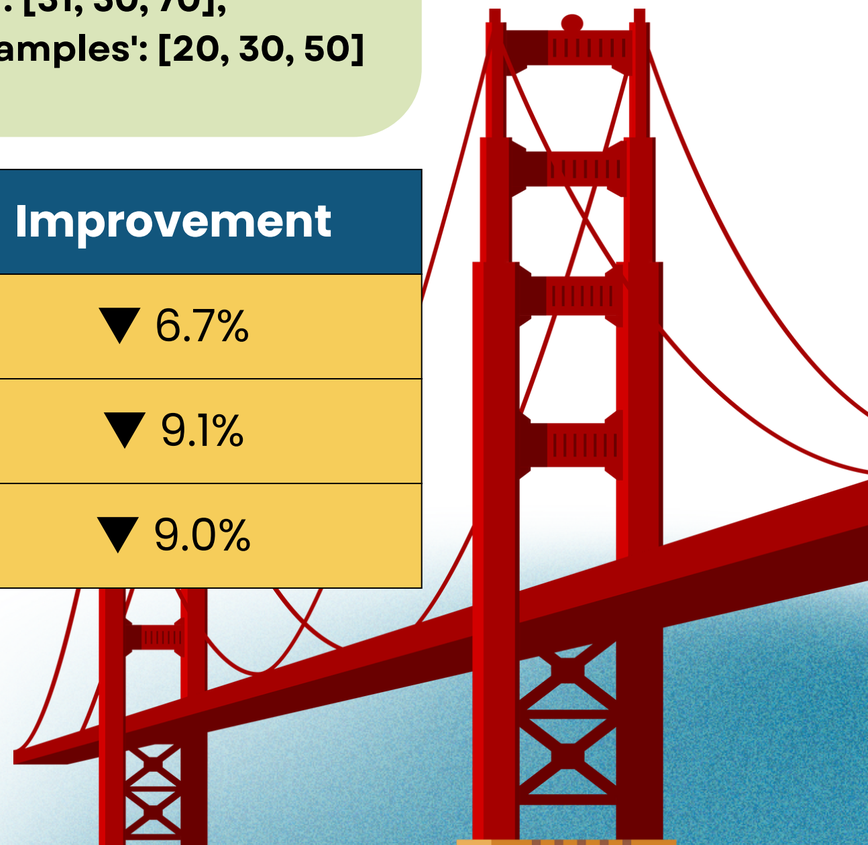


Model Summary: LightGBM Regressor
Evaluation method: 5-Fold Cross Validation
Scoring metric: RMSE
GridSearchCV explored 324 model combinations
(3 × 4 × 3 × 3 × 3 hyperparameter combination)

```
param_grid = {
 'n_estimators': [100, 200, 300],
 'max_depth': [5, 10, 20, None],
 'learning_rate': [0.01, 0.05, 0.1],
 'num_leaves': [31, 50, 70],
 'min_child_samples': [20, 30, 50]
}
```

CV = 5

Metric	Before Tuning	After Tuning	Improvement
RMSE	~47,600	$44,371	▼ 6.7%
MAE	~32,350	$29,435	▼ 9.1%
MAPE	~18.8%	17.06%	▼ 9.0%



MODELING

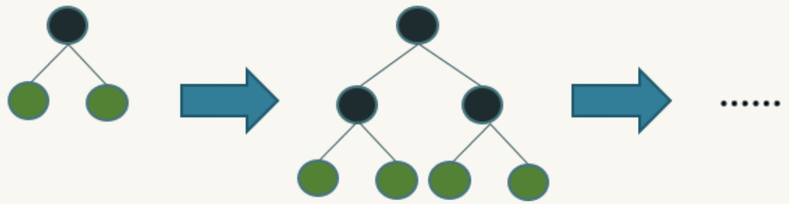
Algorithm: LightGBM

1 Core Workflow

- Start with a simple model (base learner)
- Build the next tree to correct previous errors
- Repeat for N boosting rounds (n_estimators)

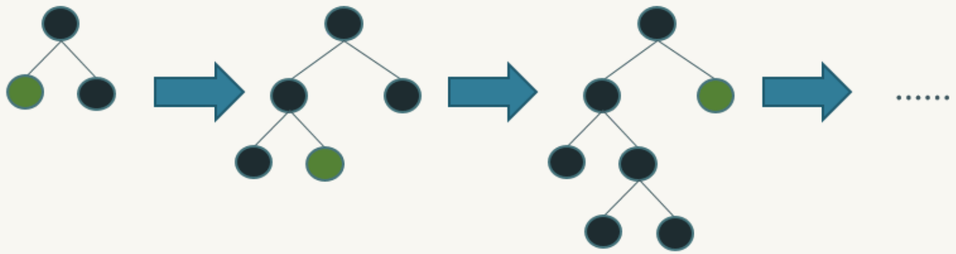
2 Unlike other, who is

Level-wise is expanding all nodes at a given depth,



Level-wise tree growth

LightGBM grows leaf-wise, choosing the leaf with the highest loss reduction.



Leaf-wise tree growth

3 Optimized Techniques

- Histogram-based learning → **Extremely fast**
- GOSS (Gradient-Based One-Side Sampling) → **Prioritizes high-error rows**

4 Limitations

- Housing Price Market **Fluctuations**
- Non-interpretable complexity **(Feature Importance)**
- **Limited Information** or Features, suggesting:
  - Proximity from the Ocean (in miles)
  - Recent Sales Price
  - Public Facilities Performance (Rating)
  - Crime Rates

5 Model Recommendations

- **LightGBM,** without drop the “ocean_proximity” features, encode: OHE
- **ElasticNet Regression,** Mix Lasso and Ridge, able to overcome multicollinearity, and easy to understand for business matters.

CONCLUSION

Cost Benefit Analysis

Aspect	Without Model	With Model
Cost	Expensive appraisal, slow process, human error	Relatively low
Speed	2–5 days per property	Thousands of predictions instantly
Accuracy	Depends on analyst, inconsistent	RMSE ~44.3k, ~MAE 29.43k, MAPE 17.6% (fair for a volatile housing market)
Scalability	Low	Very high
Risk	Prone to subjective judgement	Wrong prediction if data is biased
Benefit	Captures physical details	Automatically identifies undervalued & overvalued units
Net Result	Expensive & slow but legally recognized	Fast, scalable, cost-efficient, and data-driven

Critical Takeaways

Error metrics = "Cost"

$$\text{Percent Difference} = \frac{\text{Actual - Predicted}}{\text{Predicted}} \times 100\%$$

Generally,

- **Difference > 0 → Undervalued**
- **Difference < 0 → Overvalued**

RANGE OF RISK LEVEL (ERROR METRICS)

<= $11.092 = Low Risk

$11.093 ~ $44.371 = Medium Risk

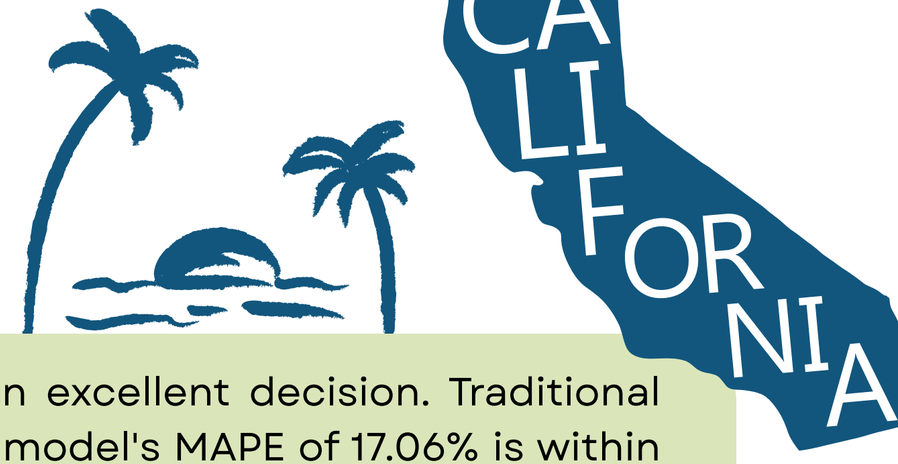
> $44.371= High Risk

Conclusion

The ML model enables real estate agents and governments to estimate property values more quickly, cheaply, and consistently than manual appraisals. Helps identify undervalued or overvalued houses, leading to wiser purchase decisions and more equitable government initiation.

RECOMMENDATION

Descriptive and Predictive



Implementing Machine Learning to forecast California house prices is an excellent decision. Traditional evaluation methods have error rates of 25-30% **(Case & Shiller, 1989),** Our model's MAPE of 17.06% is within the “Good forecasting” range **(Lewis, 1982)**, while R^2 of 0.83 meets academic norms **(Malpezzi, 2003).** Location explains 39.65% of price variation, which is consistent with coastal premium effects ranging from 15 to 40% **(Glaeser & Gottlieb, 2009).** LightGBM outperformed other models (RMSE = $44,300) in delivering accurate and generalizable price projections as a predictive measurement.

Predictive Analysis: Affecting Decision

For Real Estate Agency,

- Actual price: $450,000
- Predicted price: $520,000

➡ Differences: - 70,000 (overvalued) (Over $44,300)

➡ Property cost higher than the ML estimation → **Not worth the price or renegotiation**

For Government,

- Overvalued **(Implement Market Fairness)**
- Undervalued **(Urban Planning, Market Revitalization)**

For both parties,
able to identify future gentrification. ML models can predict gentrification 12-18 months in advance with 75% accuracy.
(Been et al., 2019)

Appendix

APPENDIX

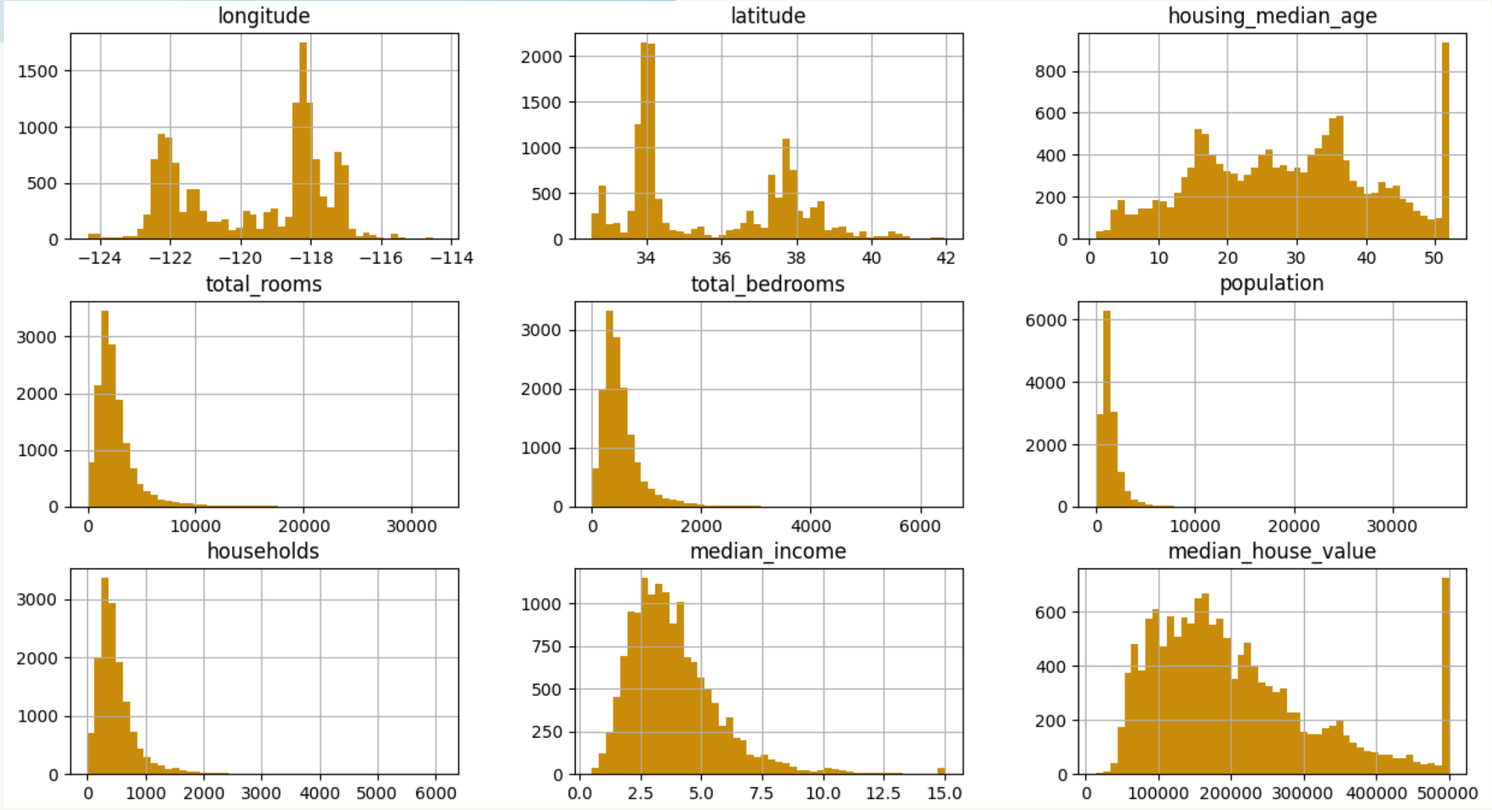
Preprocessing

A. Data Columns

Feature	Description
longitude / latitude	Geography
housing_median_age	Median House Age
total_rooms	Total Rooms
total_bedrooms	Total of Bedrooms
population	Population in Aera
households	Households Number
median_income	Median Income (×$10.000)
ocean_proximity	Distance to Ocean
median_house_value	Median House Value in USD

The dataset includes 9 numerical features and 1 categorical feature (ocean_proximity)

B. Histogram for Handling Outliers & Cleaning



C. Findings (Dataset and Histogram)

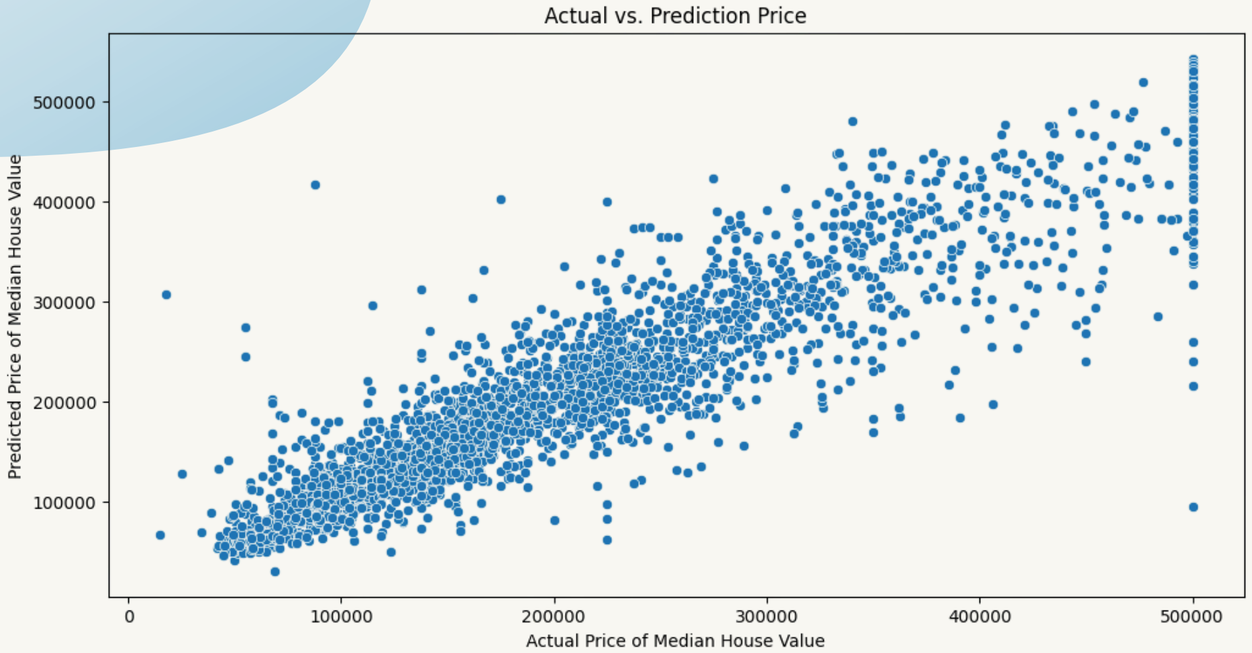
- **“total_bedrooms”** has 137 missing values
→ impute mean (numerical value)
- **No duplicated data**
- **median_house_age shows a sharp peak around age 50–52.** No cleaning is applied, but this requires verification.
- **median_house_value shows a significant spike at 500,000 USD,** likely due to a capped value or missing higher variance. No cleaning is performed, but needs further review.
- **Right-skewed are too big, further action must be taken of (Feature Engineering)**

APPENDIX

Prediction

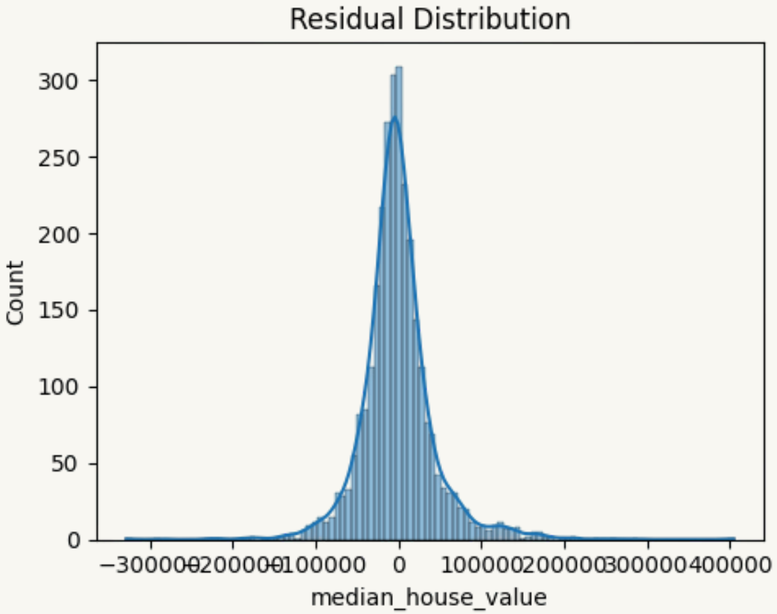
(1) Actual vs. Predicted Price

- Shows how well the model predicts median house values.
- Points close to the diagonal line indicate accurate predictions. **The outliers: capped value**



(2) Residual Distribution

- Displays the frequency of **prediction errors (residuals).**
- Ideally, residuals should be **normally distributed around zero**.



(3) Residuals vs. Predictions

- Reveals patterns in prediction errors across different predicted values.
- **High variances prone to error.**