Exploratory Data Analysis and Machine Learning on California Housing Prices

May 2025

Presented by

Durgasi, Ranjitha
Ettam, Harshitha
Korapati Murali, Harshitha
Mengane, Dhawalshree Ashok
Nallanagula, Nikhitha Reddy

Content

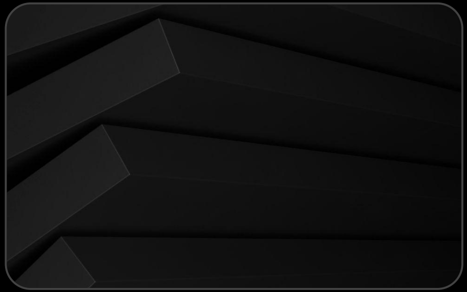
Abstract

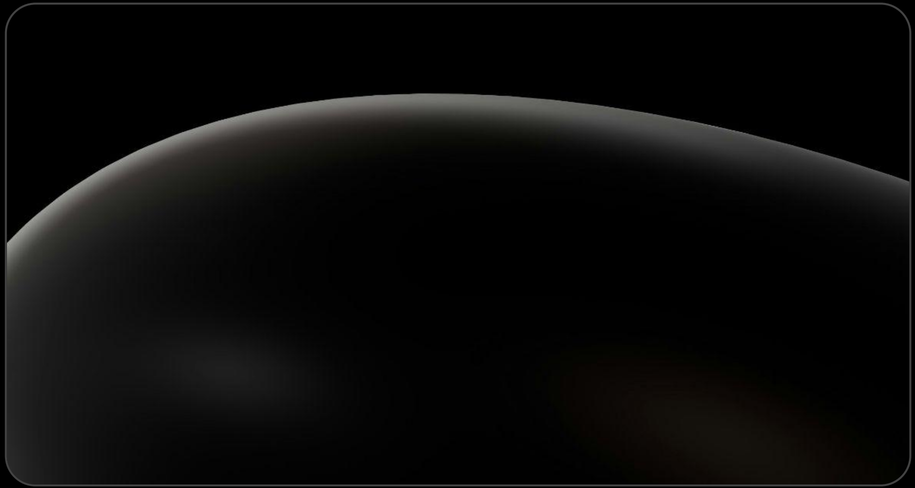
- This project presents an Exploratory Data Analysis (EDA) and Machine Learning (ML) approach to California housing data.

- Our objective is to uncover patterns and predict housing prices based on demographic and geographic features.

- The study uses supervised learning models to identify the most influential factors affecting house values.

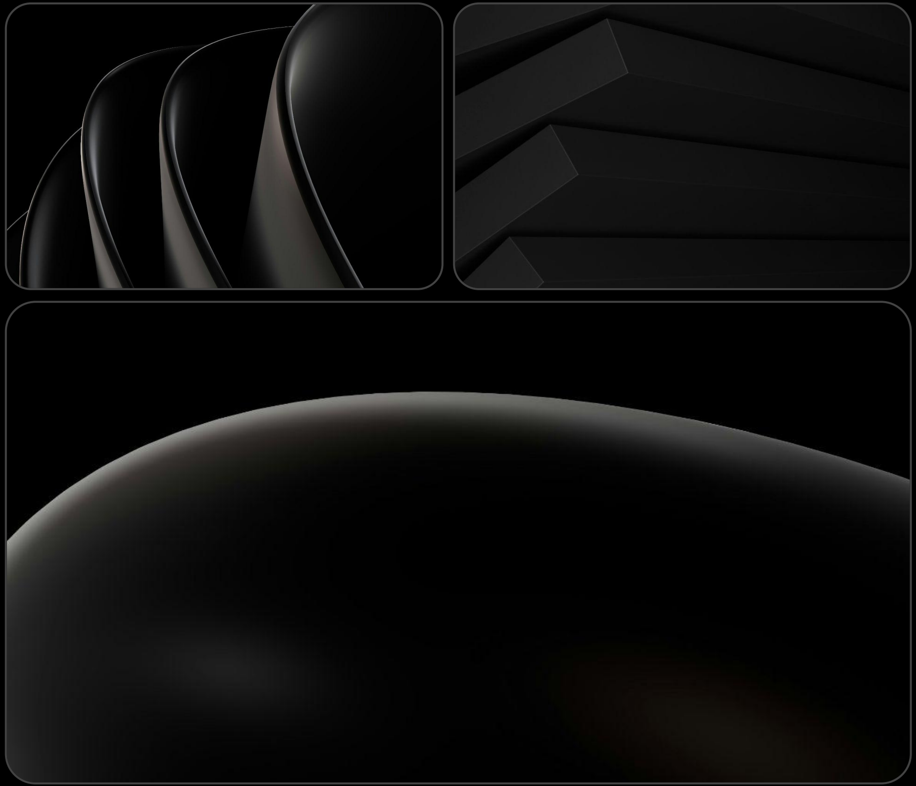
- The insights can support stakeholders like realtors, policymakers, and homebuyers in making informed decisions.





Introduction

- The California Housing dataset, derived from the 1990 Census, includes housing and population data across various districts.
- The primary aim is to explore the data, understand feature relationships, and predict the median house value.
- Machine Learning models such as Linear Regression, Decision Trees, and Random Forests are employed to enhance analytical depth.
- Classification is also performed to categorize houses into value groups (Low, Medium, High), aiding better market segmentation.



Dataset Description

Source: Provided as `housing.csv` – based on California census blocks.

Attributes:

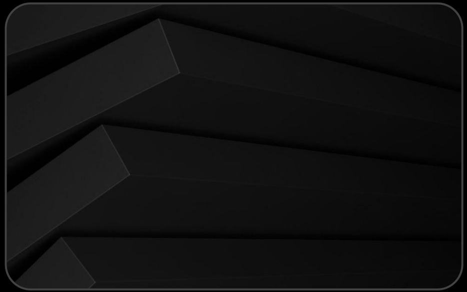
- **Numerical:** `median_income`, `housing_median_age`, `total_rooms`, `population`, etc.
- **Categorical:** `ocean_proximity` – indicates distance from the ocean.

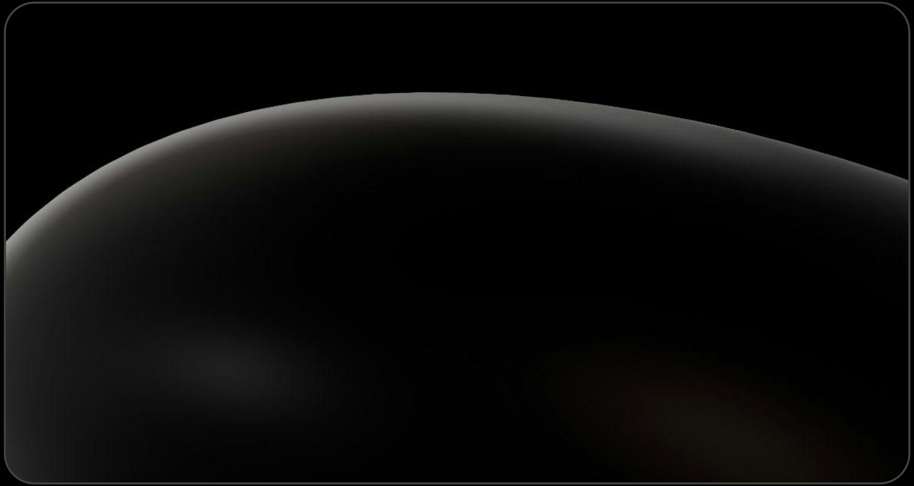
Target Variable: `median_house_value` – used for regression and classification tasks.

Summary:

- No major data-type inconsistencies.
- Descriptive stats show varying income levels and house values across regions.

Data Preprocessing





Missing Values: Rows with missing entries (mainly `total_bedrooms`) were removed to maintain data integrity.

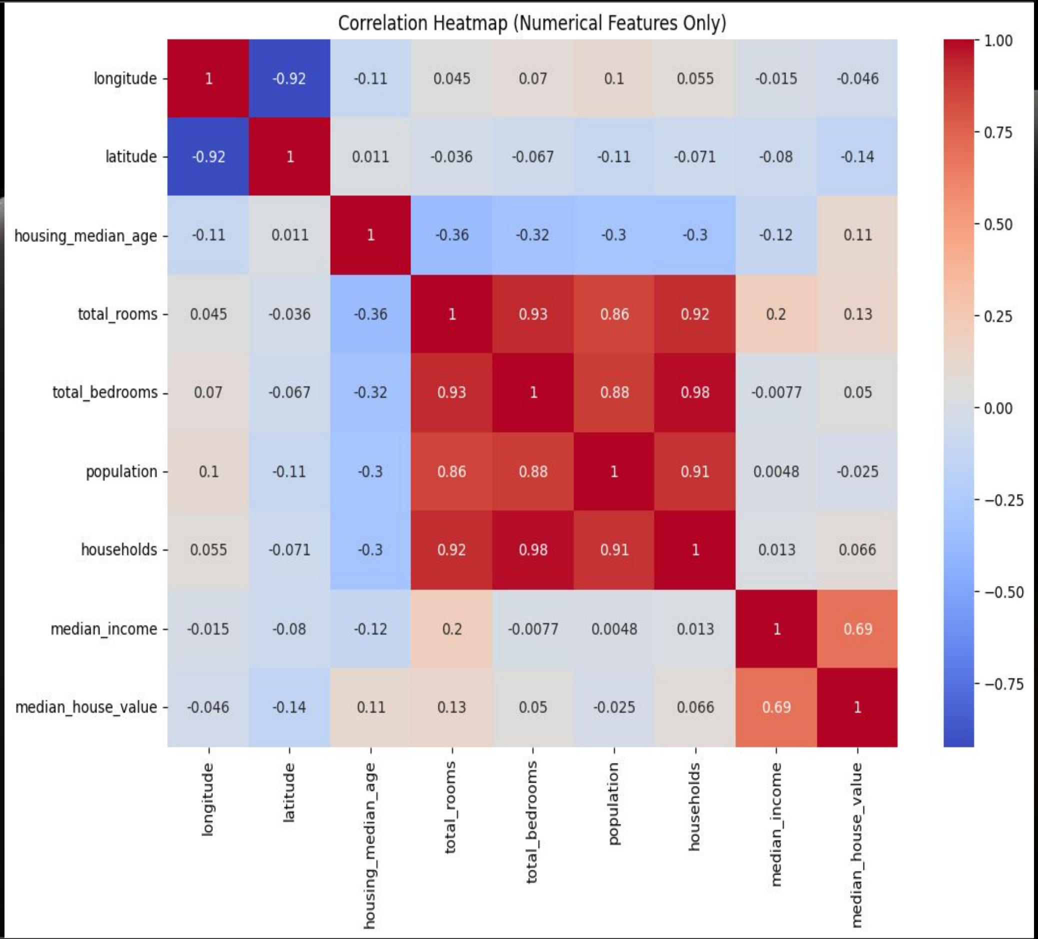
Categorical Encoding:

- `ocean_proximity` was converted into dummy variables using one-hot encoding.
- This avoids introducing bias from ordinal assumptions.

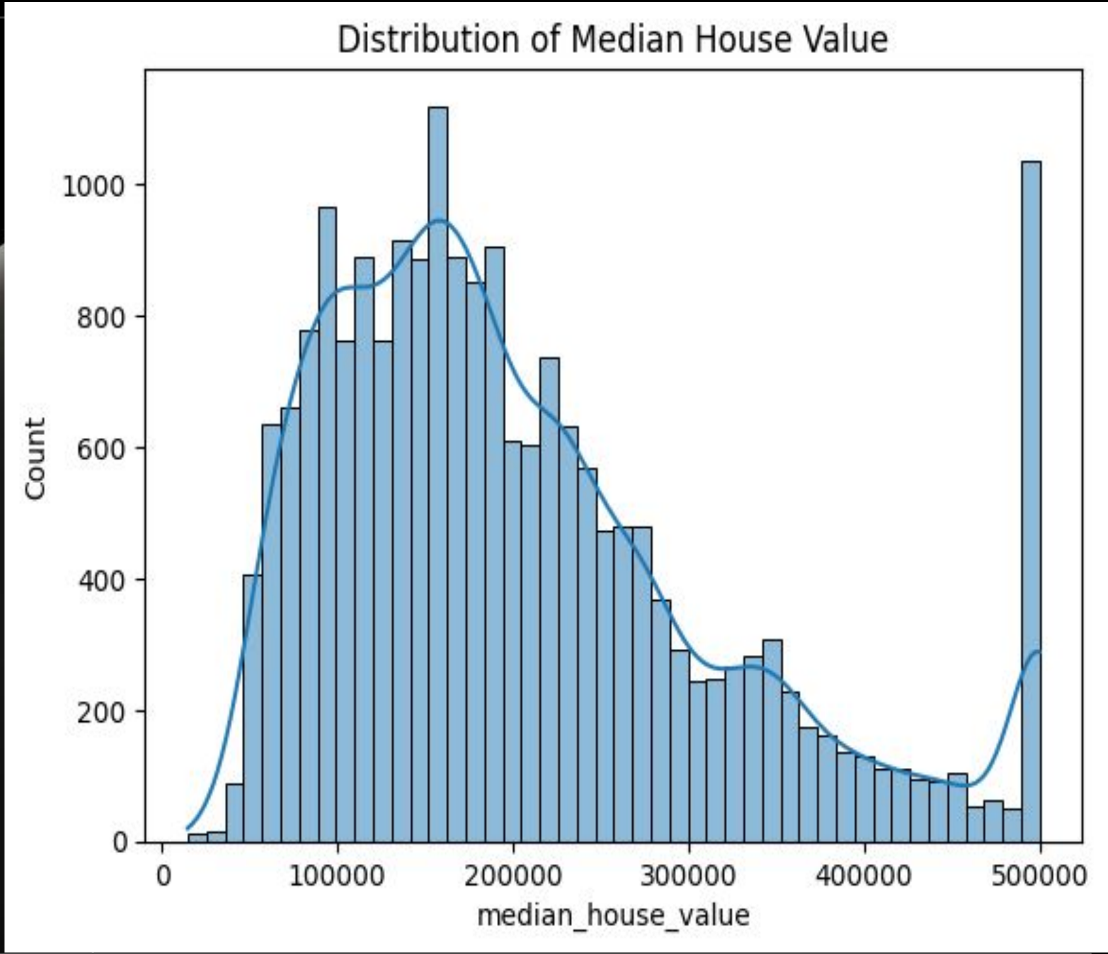
Feature Scaling:

- StandardScaler was applied to normalize numerical features.
- This ensures uniformity in scale, especially important for distance-based algorithms and gradient-based optimization.

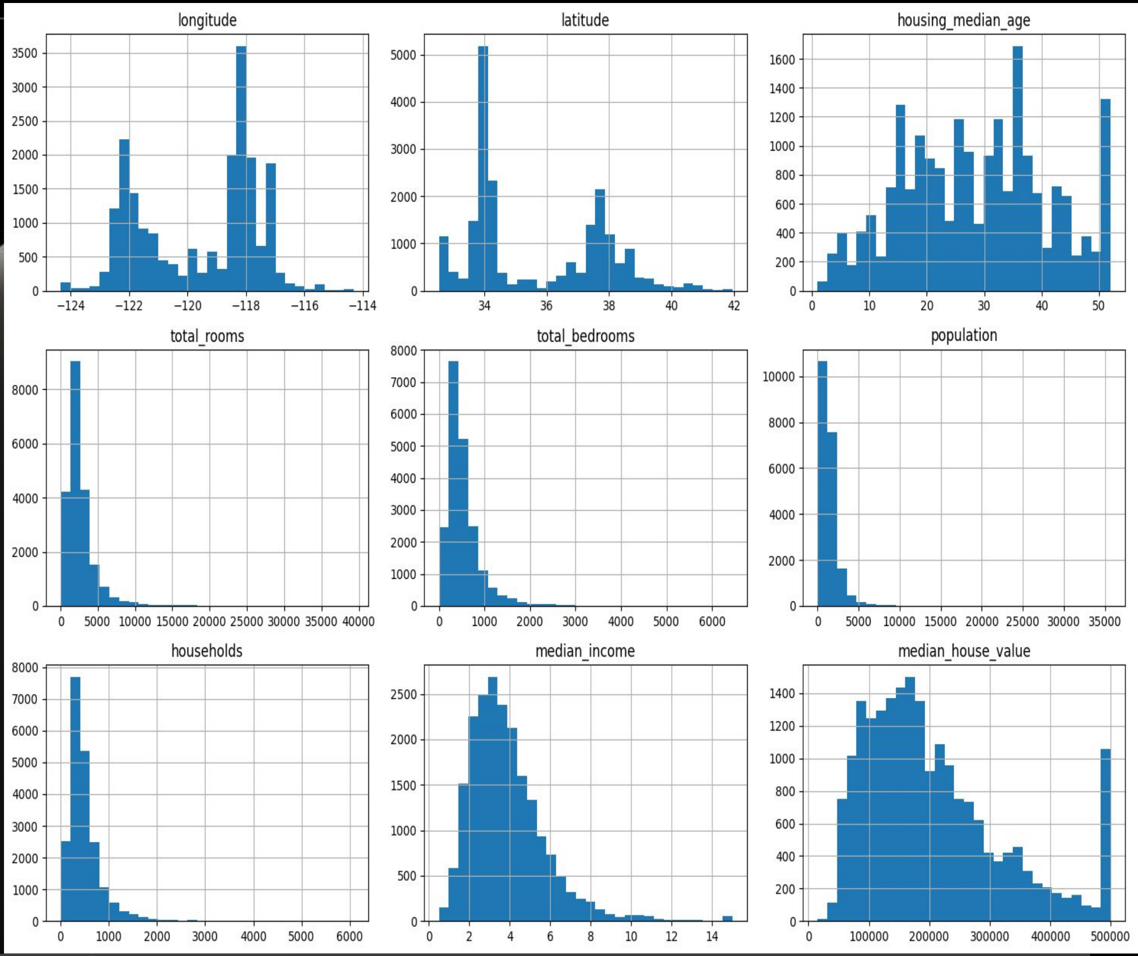
Correlation Heatmap



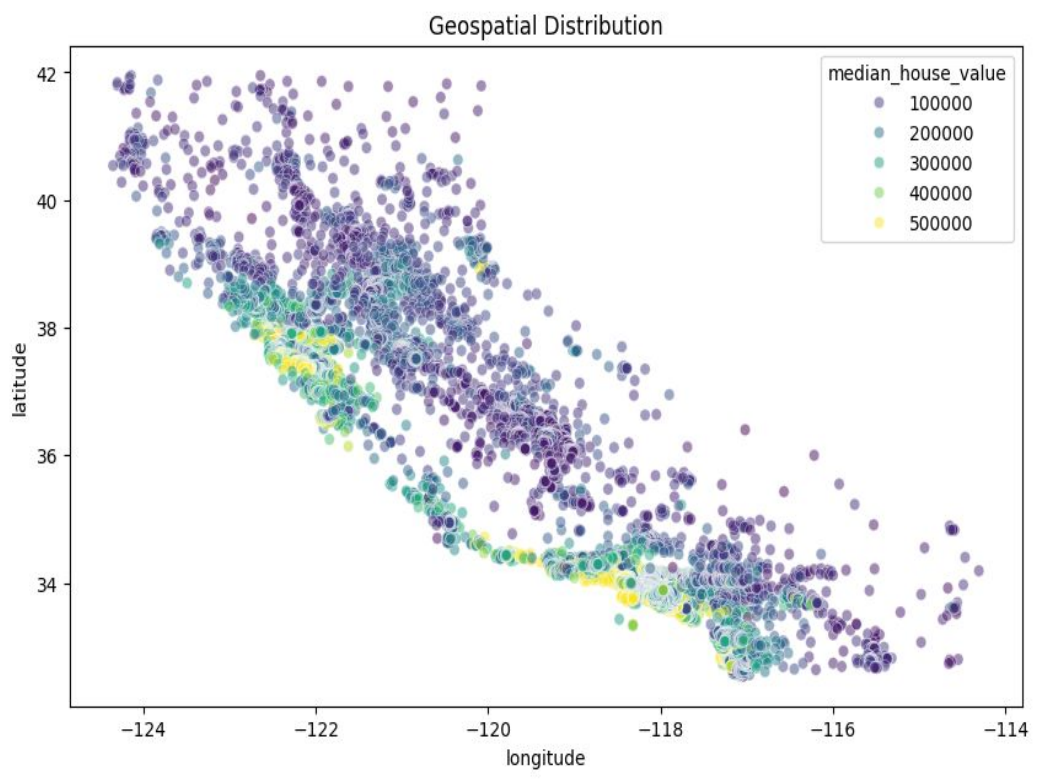
Target Variable Distribution



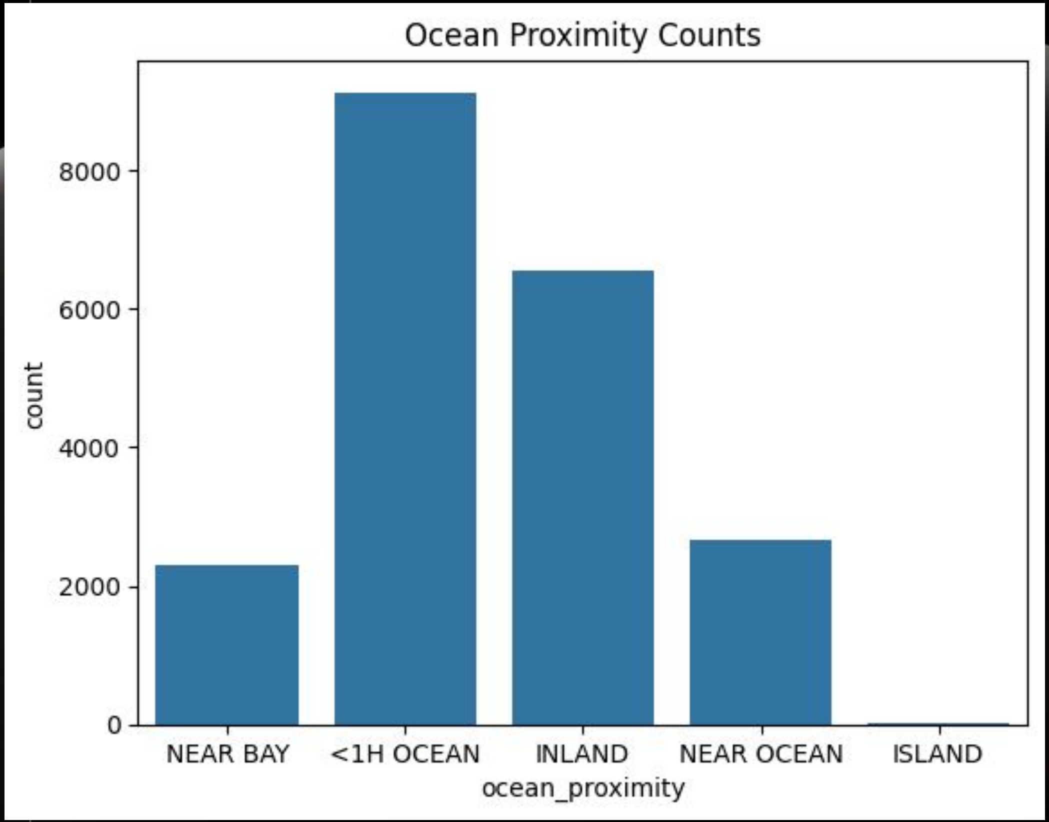
Histogram of All Features



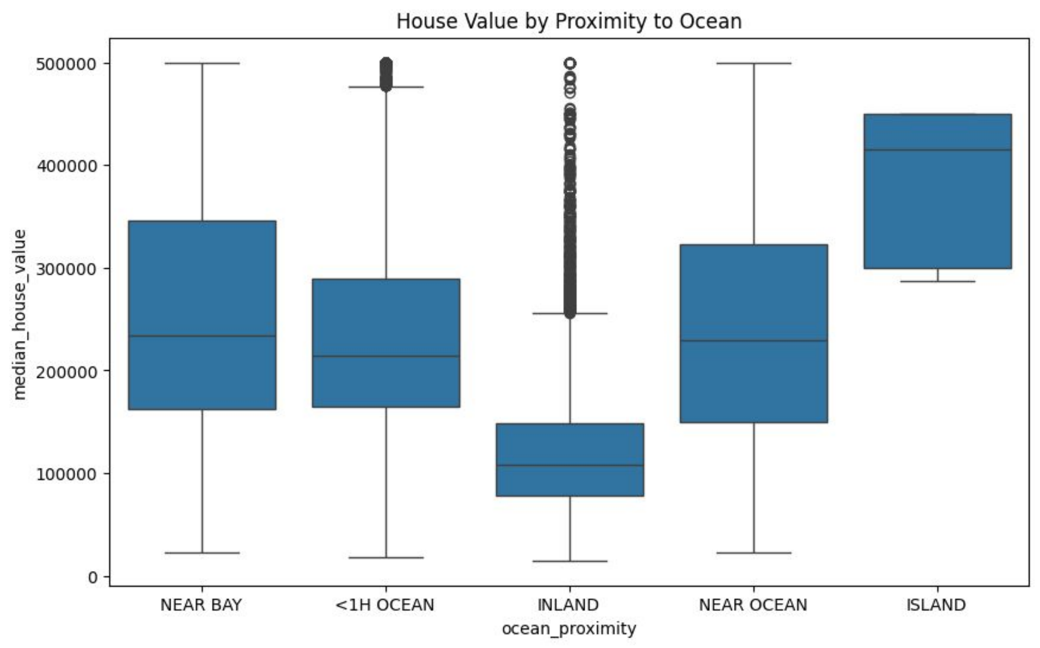
Geospatial Scatter Plot



Ocean Proximity Counts

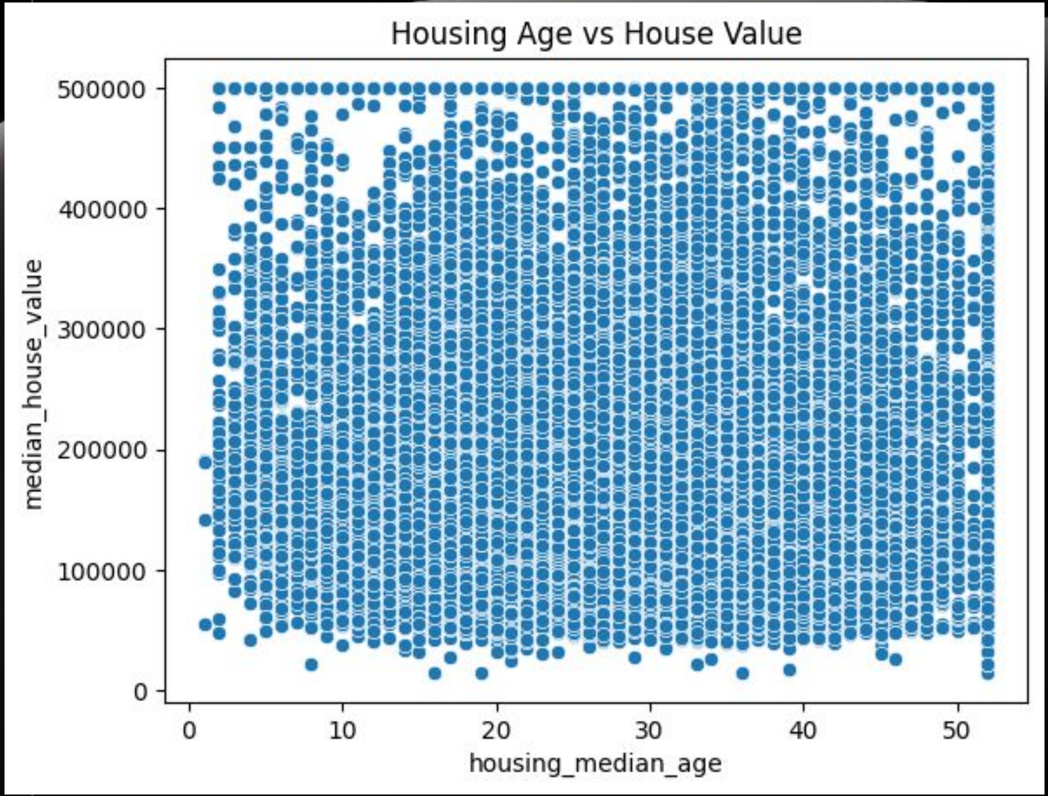


Boxplot by Ocean Proximity

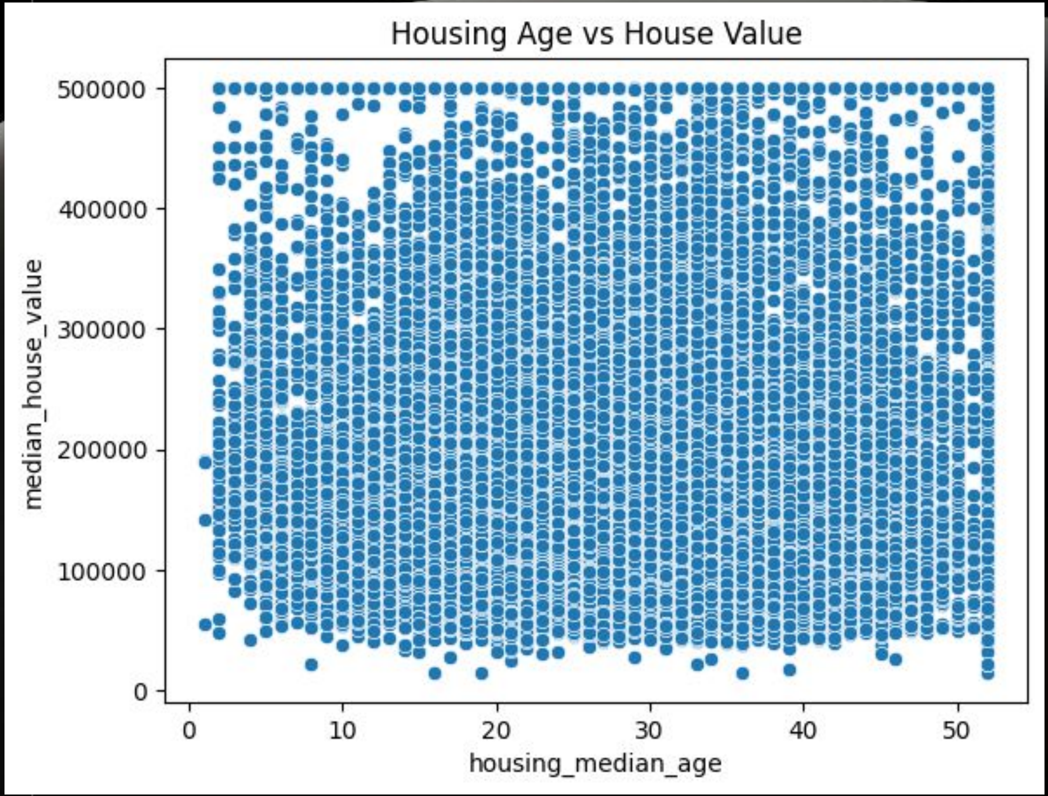


Income vs House Value Scatter Plot

Housing Age vs House Value Scatter Plot

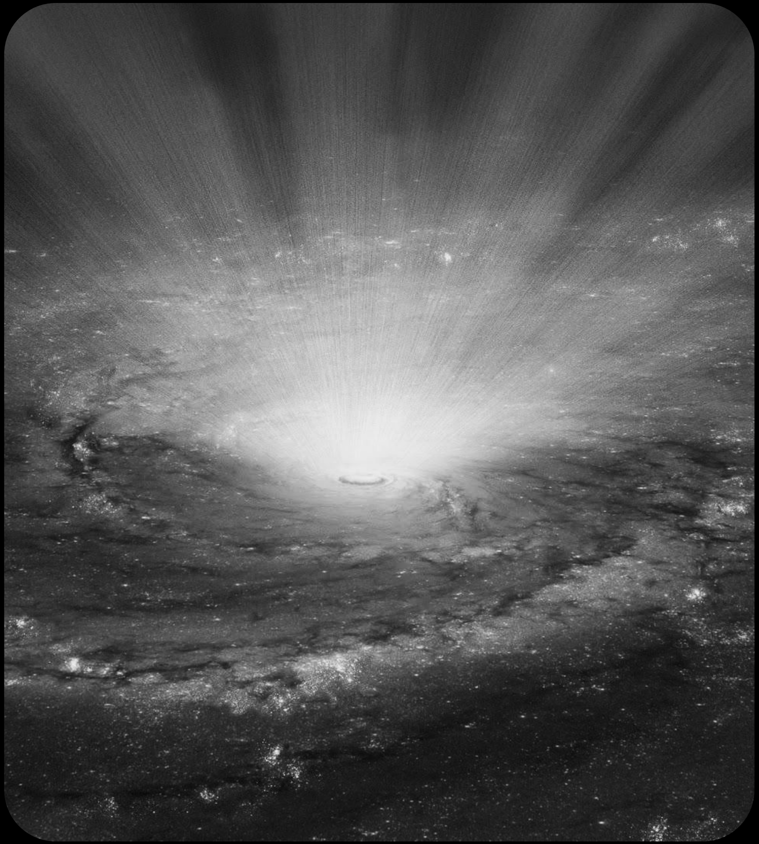


Missing Value Check



Literature Review

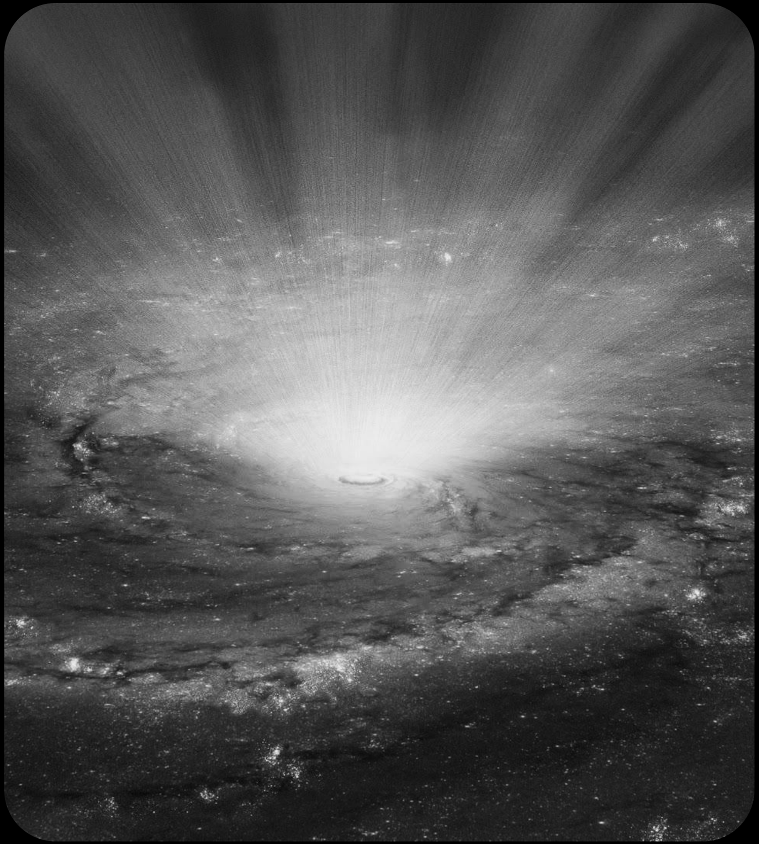
- **Previous Work:**
  The California Housing dataset is widely used in academic and industrial settings for housing price prediction.
  - Studies show that median_income is the strongest predictor of median_house_value.
  - Common references include Kaggle kernels, Scikit-learn tutorials, and research papers on socioeconomic housing analysis.
- **Modeling Strategies:**
  - Linear models are useful for establishing performance baselines.
  - Decision Trees and Random Forests are preferred for capturing complex, nonlinear interactions.
  - Classification approaches are used when converting price prediction into value segmentation (e.g., Low, Medium, High).



Methodology / Architecture

Regression Models:

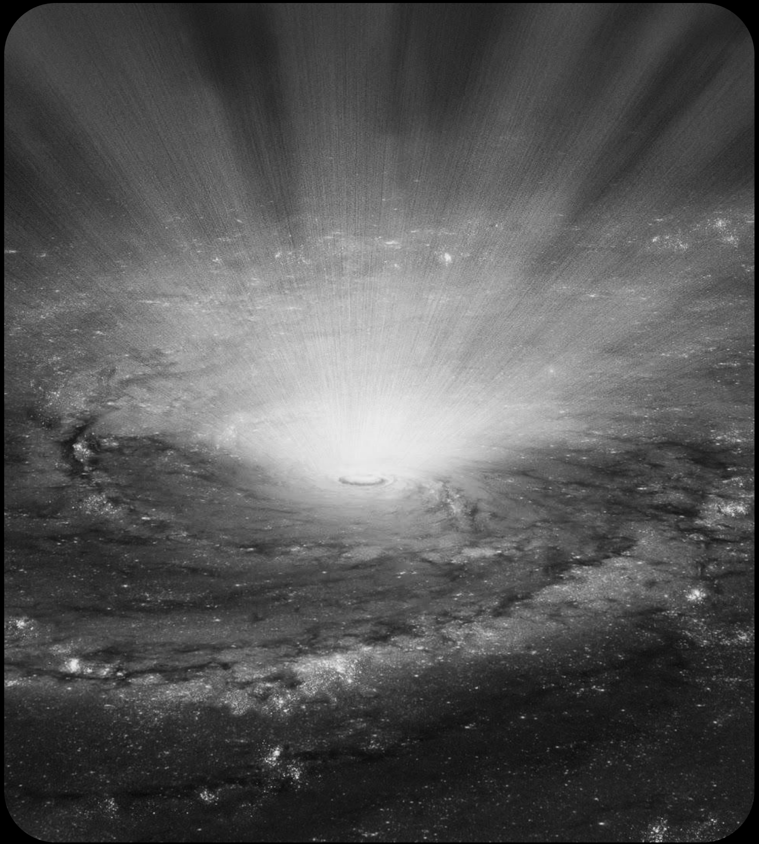
- **Linear Regression:** Fits a straight line; interpretable but limited in handling complex patterns.
- **Decision Tree Regressor:** Splits data into regions to model nonlinearity; can overfit on small data.
- **Random Forest Regressor:** Aggregates multiple decision trees to reduce overfitting and improve accuracy.



Methodology / Architecture

Classification Models:

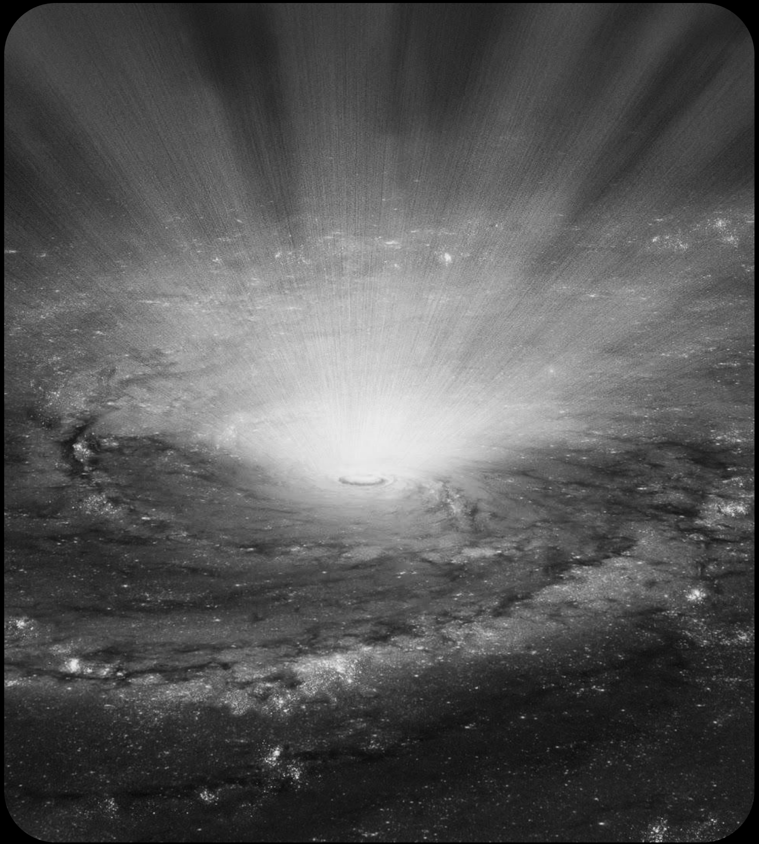
- **Logistic Regression:** Probabilistic linear model used to classify houses into value categories.
- **Decision Tree Classifier:** Makes decisions by creating a tree-like structure based on feature values.
- **Random Forest Classifier:** Uses an ensemble of decision trees for better generalization.



Methodology / Architecture

Train-Test Splitting & Evaluation:

- Used 80/20 train-test split to evaluate models.
- Applied **StandardScaler** to normalize numerical features.
- Metrics:
  - Regression: R^2 Score and scatter plots of actual vs. predicted.
  - Classification: Accuracy, precision, recall, F1-score via classification reports.



Results & Model Evaluation

Regression Model Performance (R^2 Score)

Model	R^2 Score
Linear Regression	0.6488
Decision Tree Regressor	0.7219
Random Forest Regressor	**0.7860** ⭐ *(Best)*

- **Insight:** Random Forest outperforms others in capturing complex feature interactions.

Results & Model Evaluation

Classification Model Performance (Accuracy & F1 Score)

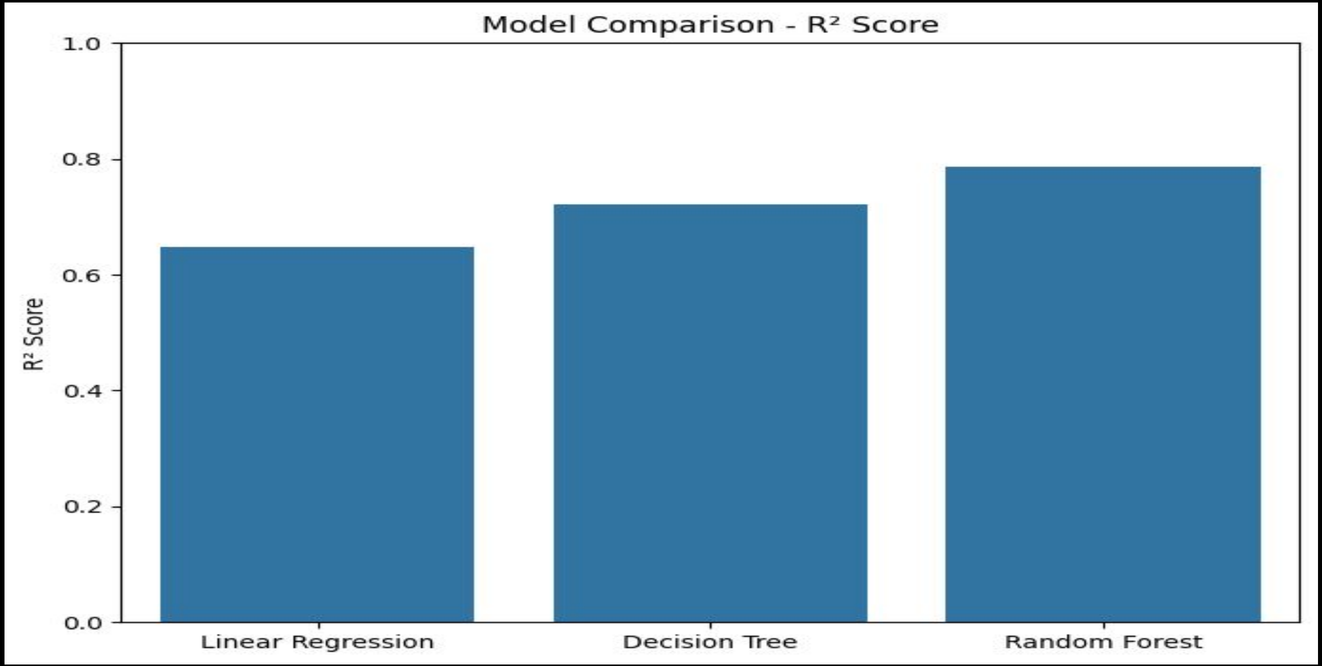
Model	Accuracy	Weighted Avg F1 Score
Logistic Regression	75%	0.75
Decision Tree Classifier	79%	0.79
Random Forest Classifier	**80%**	**0.79** ⭐ *(Most balanced)*

Class-wise Insight:

- **High value homes:** Precision highest in Random Forest (91%)
- **Medium value homes:** Best recall in Random Forest (87%)
- **Low value homes:** Consistently high across all models

Results & Model Evaluation

Model R^2 comparison



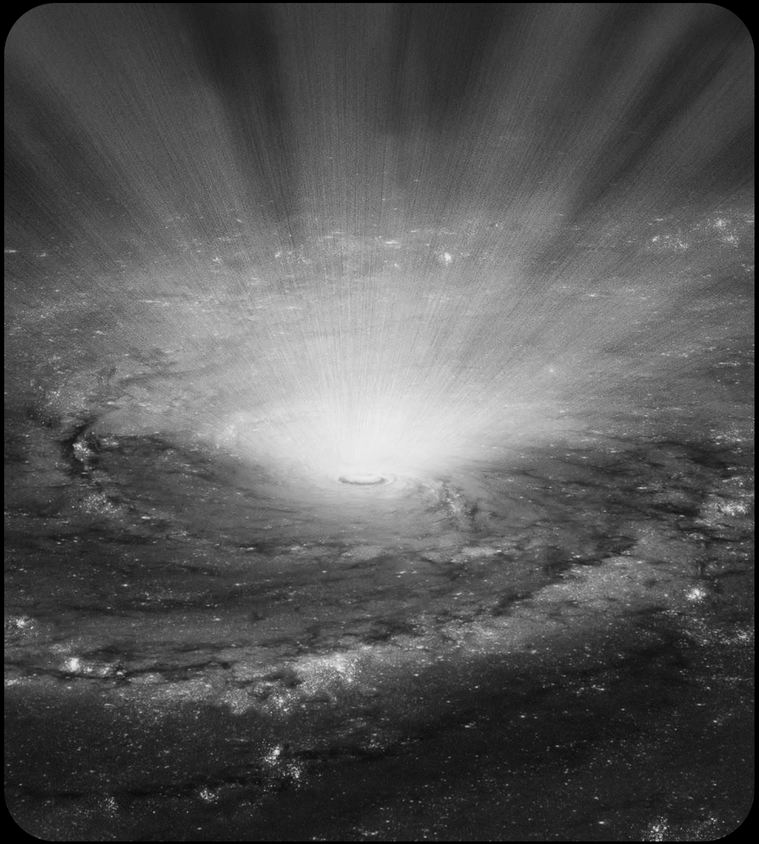
Conclusion

Key Findings:

- Median_income is the strongest predictor of house value.
- Models capture nonlinearities better than linear approaches.
- Random Forest performed best across both regression and classification tasks.

Limitations:

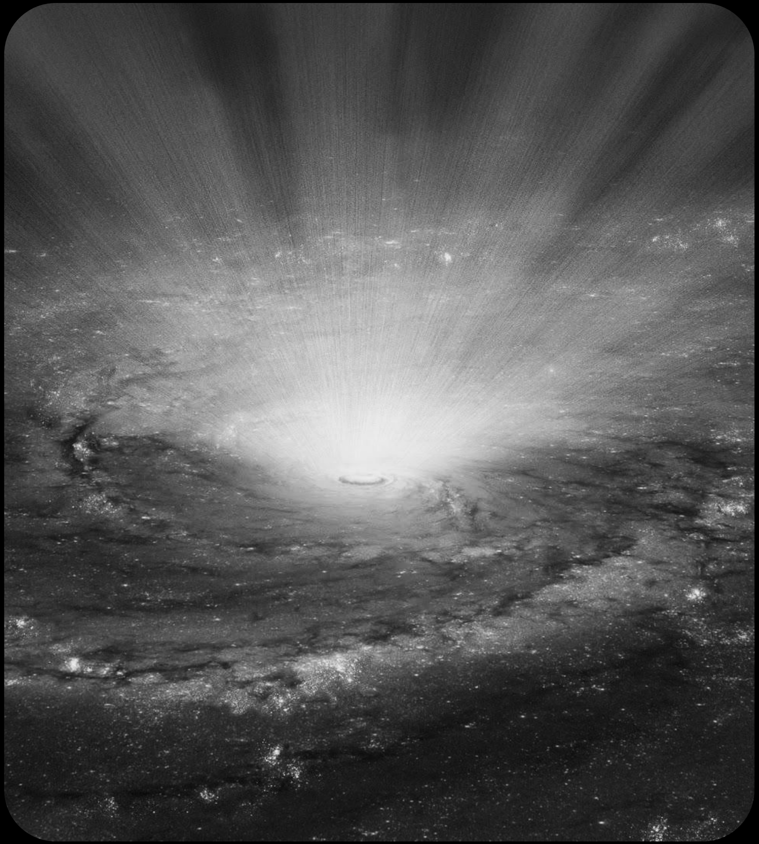
- Dataset only covers California and may not generalize.
- No feature for real-time or temporal housing trends.



Recommendations & Future Work

- **Improvement Suggestions:**
  - Include temporal variables like year of sale or economic indicators
  - Use ensemble stacking or XGBoost for potentially higher accuracy

- **Future Work:**
  - Explore deep learning methods (e.g., neural networks)
  - Combine with geospatial or map-based APIs for location intelligence



References

Dataset

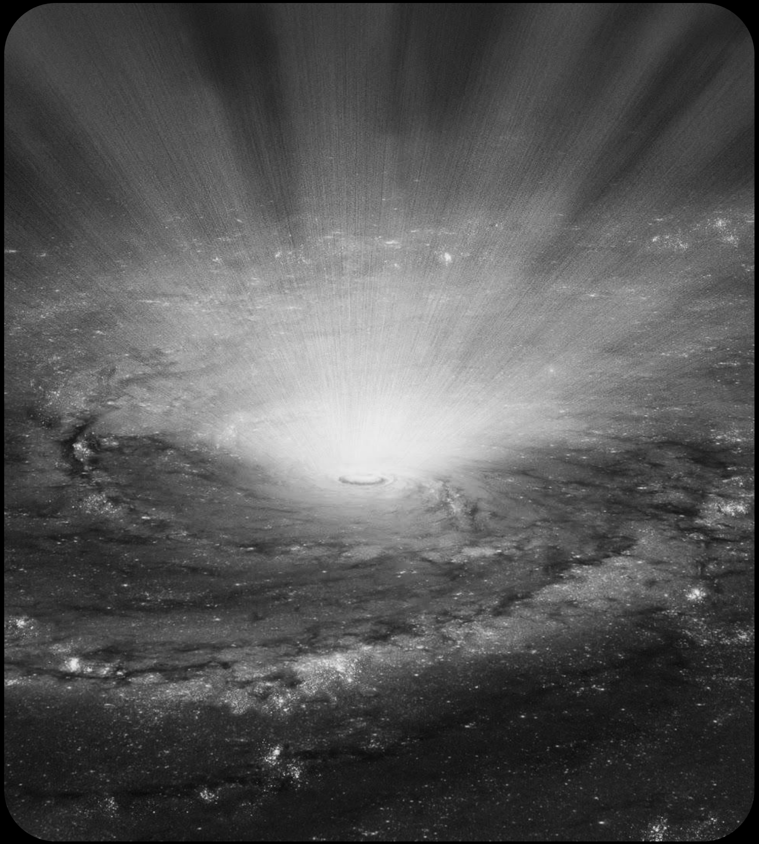
- Nugent, C. (2017). *California Housing Prices Dataset*. Retrieved from https://www.kaggle.com/datasets/camnugent/california-housing-prices (File used: housing.csv, based on the 1990 California Census)

Libraries and Tools

- Pandas – Data manipulation and analysis
- NumPy – Numerical computations
- Matplotlib and Seaborn – Data visualization
- Scikit-learn – Machine learning models and preprocessing
- Google Colab – Cloud-based Python execution environment

Documentation and Resources

- Scikit-learn Official Documentation: https://scikit-learn.org
- Kaggle Notebooks related to California Housing Price predictions
- Research papers and blogs on housing value modeling and regression/classification techniques



GitHub

https://github.com/rd39257n/Python-Final