INNOMATICS RESEARCH LABS

INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

# Data Understanding on California Housing Dataset

Presented By: YASWANTH G V N S

# About me

I am YASWANTH, a graduate with a Bachelor of Technology in Computer Science and Engineering (CSE) from SRM University.

# Why Data Analytics?

After I graduated, I became interested in Data Analytics and became very excited about this path. I like working with data, finding trends, and using it to solve problems in the real world. It's a field where I can keep learning, come up with new ideas, and make a difference in many fields.

# Analysis of the California Housing Dataset

➢ **Objective:** Analyze the 1990 California Census data to identify key drivers of housing prices and understand regional real estate dynamics.

➢ **Usefulness:** Provides insights into how factors like location, income, and housing density influence property values, supporting real estate valuation, investment decisions, and urban planning.

➢ **Dataset Overview:** Contains 20,640 rows and 10 columns, including 9 numerical (e.g., median income, housing age) and 1 categorical (ocean proximity).

**Key Features:**

- **Geographic:** longitude, latitude
- **Housing:** housing_median_age, total_rooms, total_bedrooms
- **Demographics:** population, households
- **Economic:** median_income
- **Target:** median_house_value
- **Categorical:** ocean_proximity

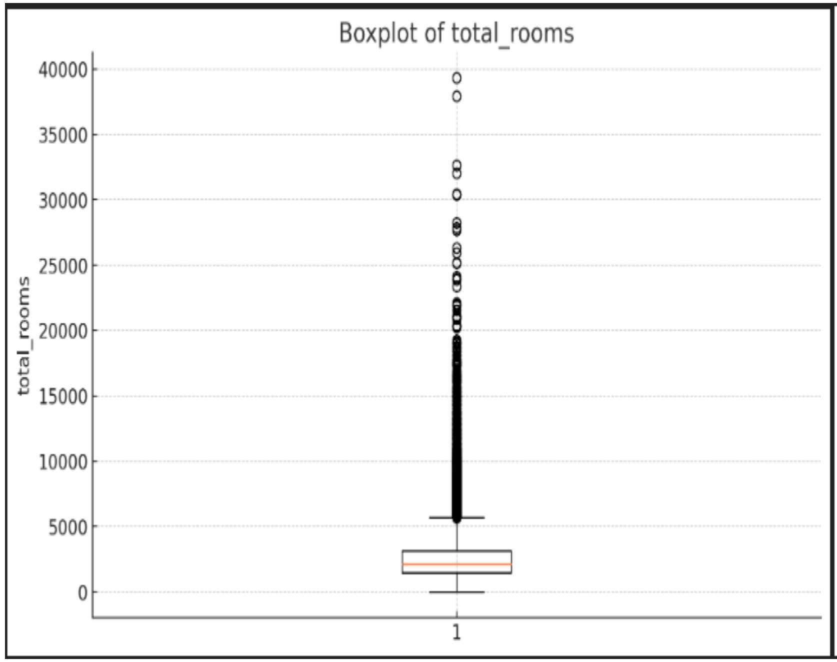
# Understanding the Columns

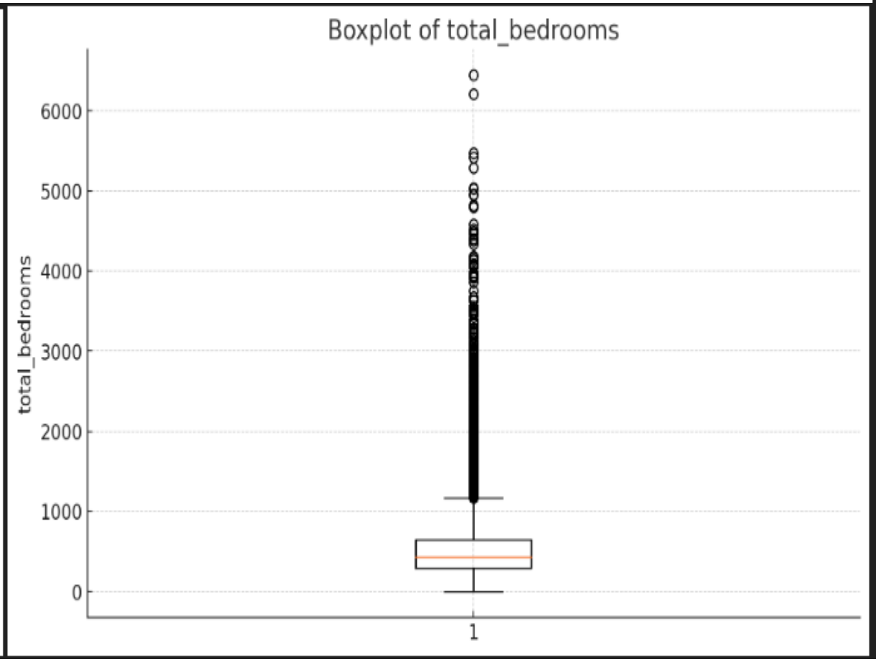
| Column Name | Missing Values | Data Type | Data Quality Issues | Cleaning / Preprocessing Required | Importance |
|---|---|---|---|---|---|
| **longitude** | 0 | float64 | No | No | Geographic location (x-axis) |
| **latitude** | 0 | float64 | No | No | Geographic location (y-axis) |
| **housing_median_age** | 0 | float64 | No | No | Represents age of houses in the block |
| **total_rooms** | 0 | float64 | Outliers possible | Scaling / outlier handling | Indicator of housing size |
| **total_bedrooms** | 207 | float64 | Missing values (~1%) | Imputation required | Indicator of living capacity |
| **population** | 0 | float64 | Outliers possible | Outlier detection / scaling | Measures population density |
| **households** | 0 | float64 | Outliers possible | Outlier handling / scaling | Represents family size or households |
| **median_income** | 0 | float64 | No | Normalization | Strongest predictor of housing prices |
| **median_house_value** | 0 | float64 | Values capped at $500,000 | Target variable – no changes | Target variable for prediction |
| **ocean_proximity** | 0 | object | Categorical variable | Encoding categories | Major factor impacting housing price |

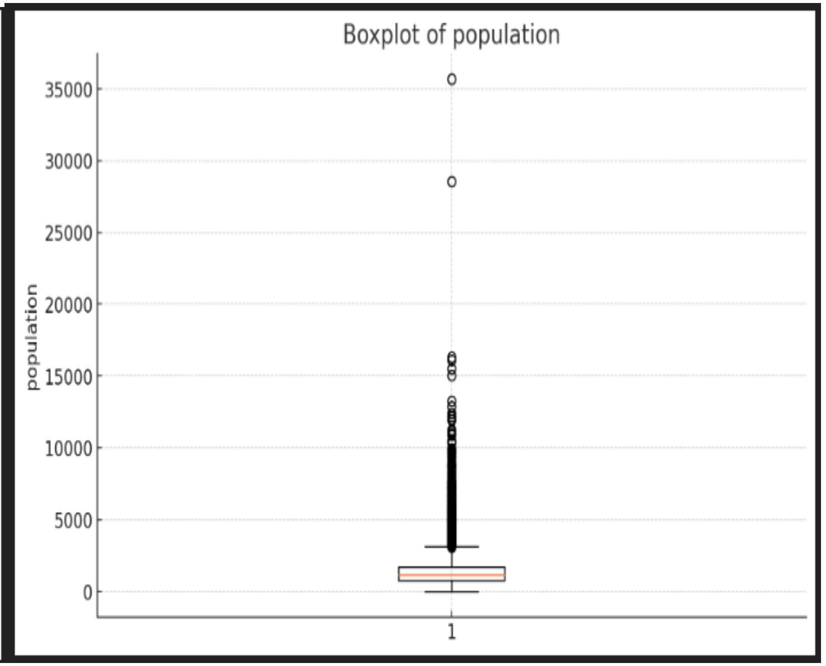
# Handling Outliers

**Outliers** are data points in the California Housing dataset that deviate significantly from the majority of observations and can distort statistical summaries and model performance.

- Outliers are extreme values that can skew statistical summaries and affect EDA accuracy.
- In this dataset, extreme values were observed in features like total_rooms, total_bedrooms, and population.
- Boxplots were used to detect these outliers.
- They were treated using IQR method and value capping to reduce their impact.



# Handling Missing Data

Missing data in the California Housing dataset occurs when values in certain columns are absent, reducing dataset completeness and potentially biasing analysis.

- The primary column with missing values is total_bedrooms, containing a small number of null entries.
- Other columns such as median_income, total_rooms, and population have no missing values.
- Although the proportion of missing data is small, ignoring it can affect measures of central tendency and model performance.
- Missing values were handled using median imputation, which helps maintain the distribution without introducing bias.

# Handling Duplicates:

Duplicates are repeated records in the dataset that can inflate counts and bias analysis.

- Checked for duplicate rows across all columns.
- No duplicate records were found in the dataset.
- This ensures the dataset remains clean, unique, and reliable for further analysis.

# Fixing Inconsistencies in Categorical Data:

Inconsistencies occur when the same category is represented in multiple forms due to formatting differences.

•The dataset contains one categorical column: ocean_proximity.
•Standardized category values to ensure consistency (e.g., removed trailing spaces or unusual string entries if any existed).
•Ensured all values were in consistent text format and checked for unique categories to avoid irregular values.

# Data Type Conversion:

Data type conversion ensures columns are stored in the correct format for accurate analysis.

•Verified that numerical columns (total_rooms, total_bedrooms, population, median_income, median_house_value, etc.) were stored in numeric format for proper statistical calculations.

•Confirmed that geographical coordinates (latitude, longitude) were numeric to support spatial visualizations.

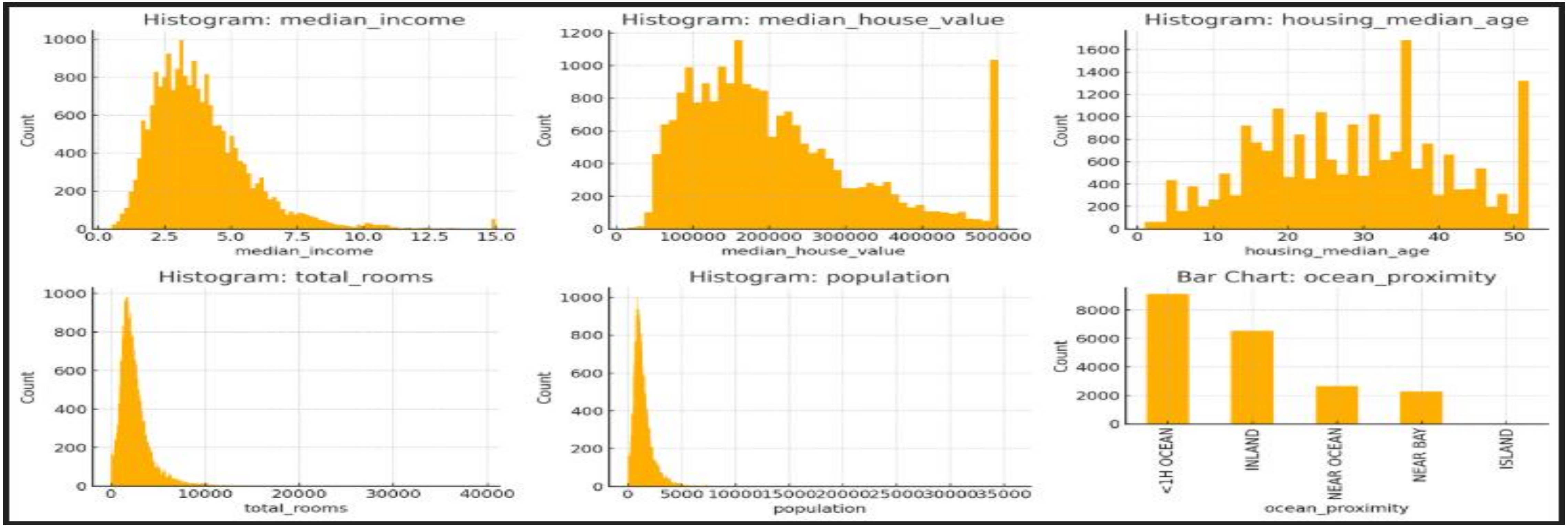
•Ensured ocean_proximity remained in categorical format (object type).

•Applied conversions where needed to fix any inconsistencies.

# Univariate Analysis

Univariate analysis examines a single variable at a time to understand its distribution and frequency.

- Analyzed key numerical features: median_income, median_house_value, housing_median_age, total_rooms, population.
- Median_income is right-skewed with few high-income outliers.
- Median_house_value shows a capped distribution.
- Housing_median_age is more evenly spread, with older houses concentrated.
- Ocean_proximity shows most districts are inland, fewer are coastal.



# Bivariate Analysis

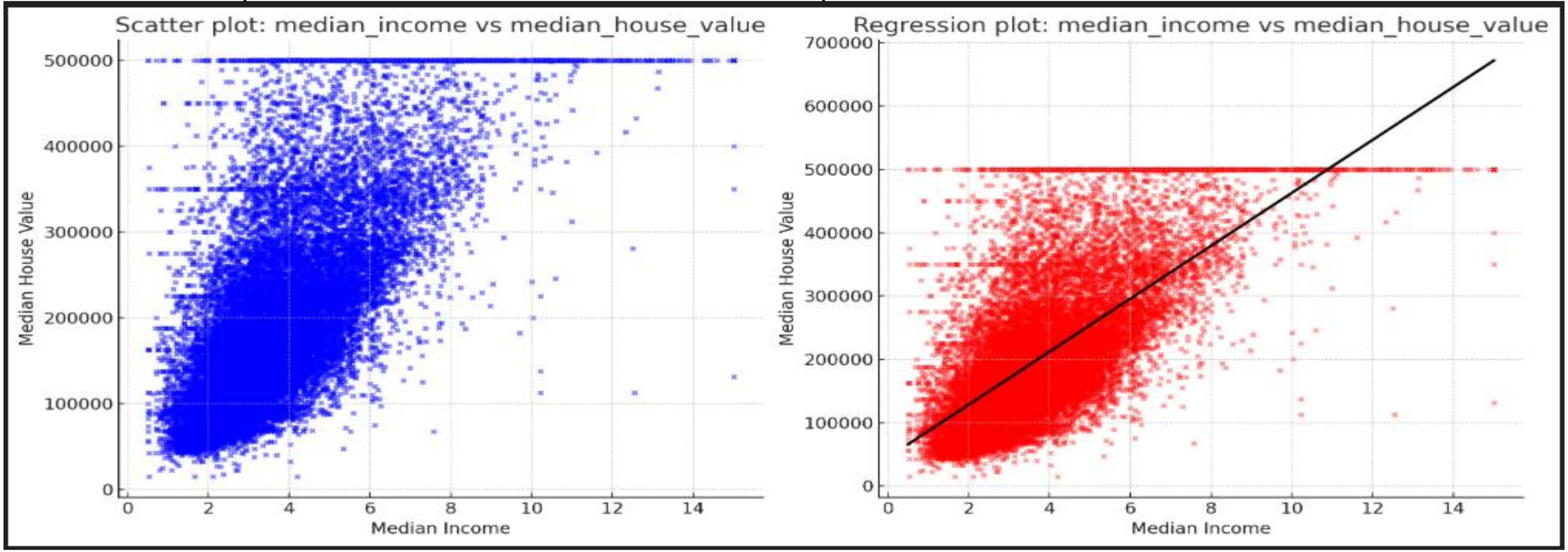
Bivariate analysis studies the relationship between two variables to identify associations, trends, or dependencies.

## 1. Numerical vs Numerical:

•Median_income vs Median_house_value shows a strong positive linear relationship — higher income areas have higher house values.

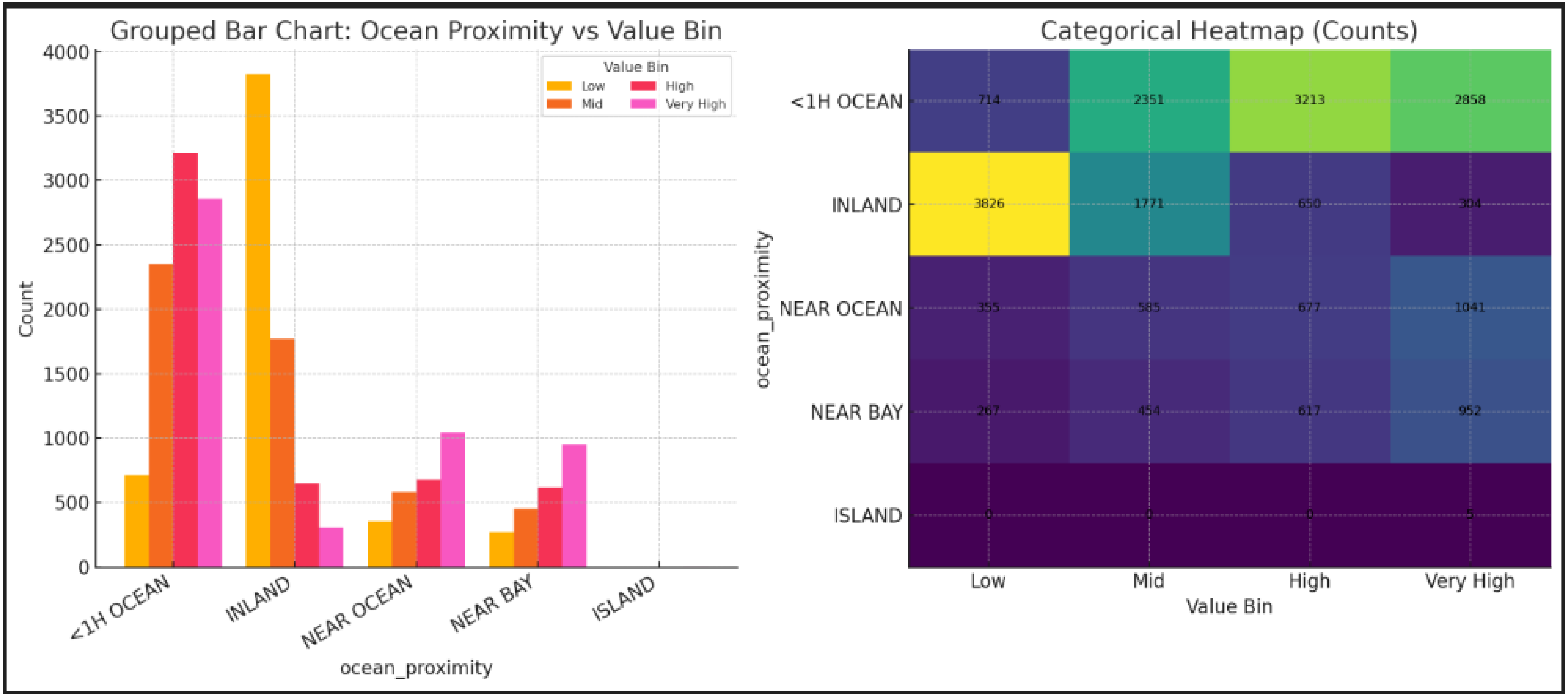
•Housing_median_age vs Median_house_value shows moderate variation with some clustering in mid-age ranges.

•Other numerical pairs show weaker or non-linear relationships.



# 2. Categorical vs Categorical

- Compared ocean_proximity with house value bins (Low / Mid / High / Very High).
- Coastal categories ("<1H OCEAN", "NEAR OCEAN", "NEAR BAY") have higher shares of High/Very High values.
- INLAND tracts are concentrated in Low/Mid value bins.

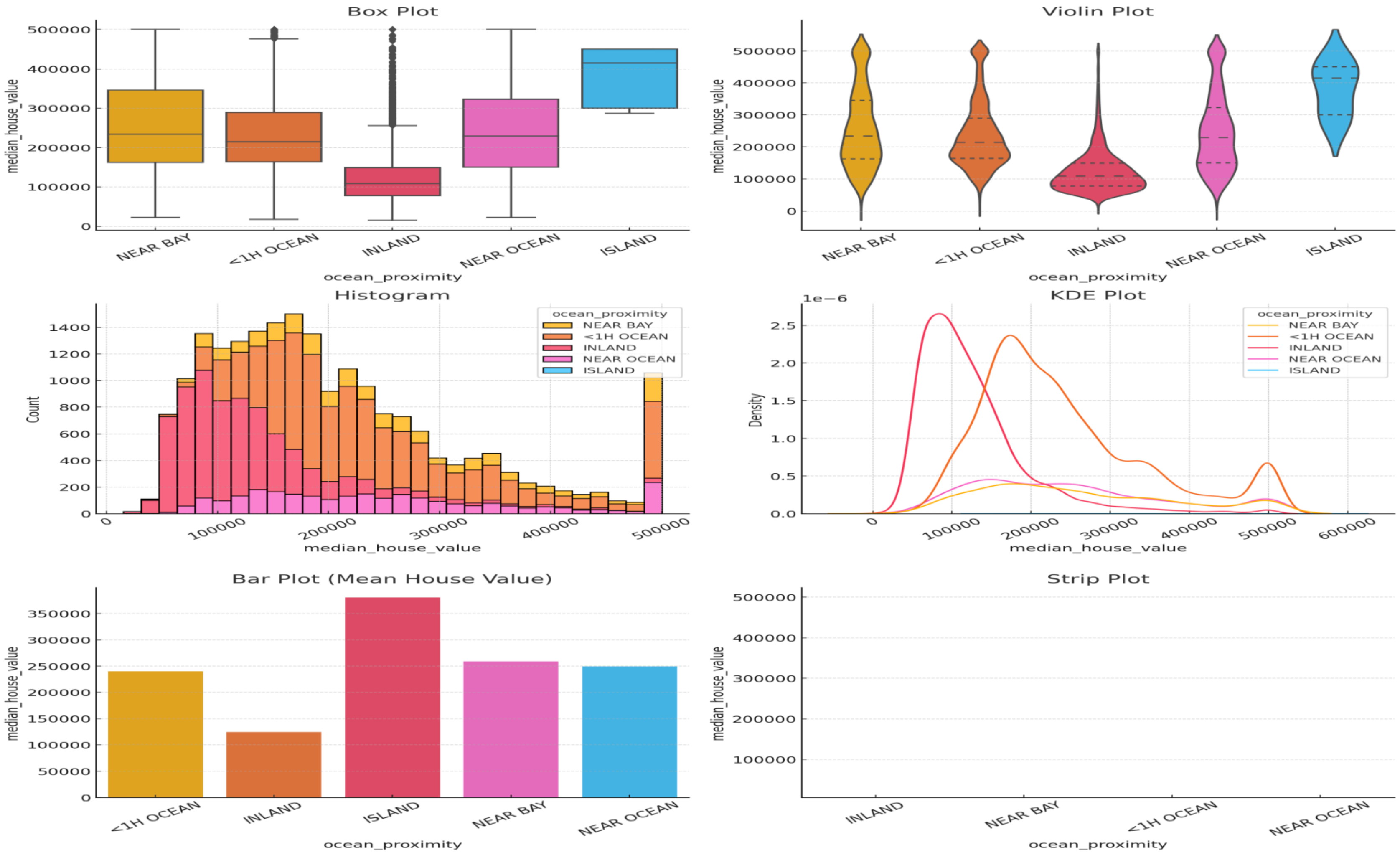


# 3. Numerical vs Categorical:

•**Median house values** vary significantly across different **ocean proximity** categories.

•**Coastal and bay areas** show **higher house values**, while **inland areas** have lower values on average.

•Clear distinctions in distribution shape and median across categories.



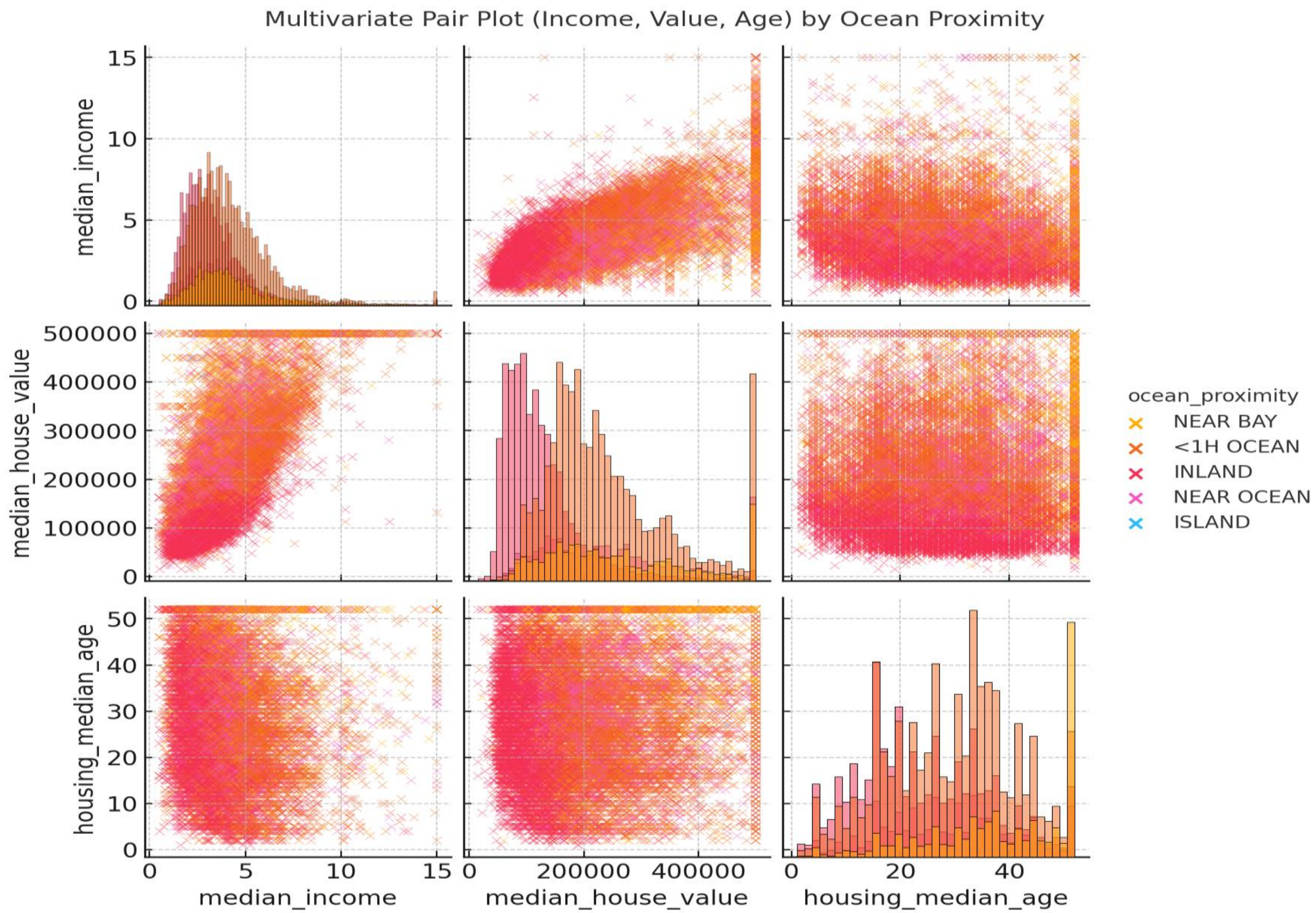
# Multivariate Analysis

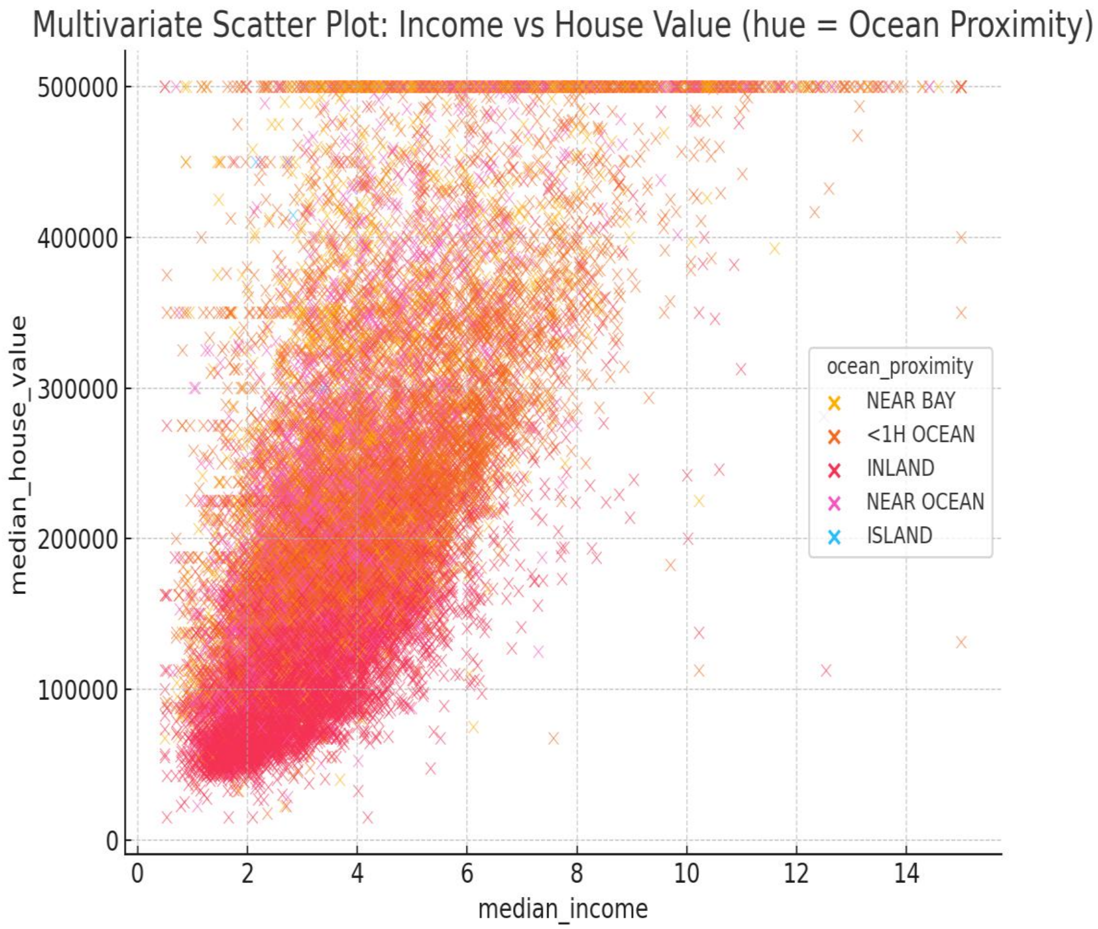
Multivariate analysis explores relationships among three or more variables simultaneously.

•Scatterplots and pairplots revealed patterns between **median_income**, **median_house_value**, and **ocean_proximity**.

•**High house values** cluster in **coastal areas** with **higher incomes**.

•Inland tracts generally have lower house values despite varying incomes.



# Problem Statement

- **Problem 1:** Housing prices vary drastically across locations, especially between coastal and inland regions.

  Goal: Identify the key features (income, house age, population density) that drive these variations.
- **Problem 2:** Income level may be directly linked to housing affordability.

  Goal: Analyze the relationship between median income and house prices to assess affordability gaps.
- **Problem 3:** Proximity to the ocean and geographical location affect housing values.

  Goal: Evaluate how longitude, latitude, and ocean proximity influence price differences.
- **Problem 4:** Population density and household size can impact housing affordability.

  Goal: Study how demand, crowding, and household size explain variations in housing prices.

# Key Insights – California Housing

- Median income is the strongest predictor of median house value — higher income areas have significantly higher property prices.
- Coastal regions (Near Ocean / Near Bay) show higher house values, while inland regions mostly fall in lower value bins.
- House value distribution is right-skewed with a clear upper cap.
- Outliers in features like total_rooms, total_bedrooms, and population were detected and treated using statistical methods.
- Data cleaning ensured correct data types, handled missing values (total_bedrooms), and removed inconsistencies.
- Bivariate and multivariate analysis highlighted geographic and income-driven clusters within the dataset.

# Handling Outliers

- Identified Outliers: Detected in numerical features such as total_rooms, population, and households.
- Techniques Used: Boxplots and histograms were applied to detect extreme values deviating from the central distribution.
- Findings: A few block groups showed very high population or room counts. These were legitimate cases from densely populated urban regions, not data errors.
- Why It Matters: These outliers reflect real-world variations in the housing market. Removing them could lead to biased insights, especially for high-density areas.
- Action Taken: No outliers were removed to maintain the dataset's representativeness.
- Impact: Preserving these data points provides a more accurate picture of California's housing diversity — from suburban neighborhoods to dense urban zones.

# Conclusion

- The California Housing Dataset is clean, structured, and well-suited for real-world data analysis.
- Minimal missing values (~1%) make it high-quality and reliable.
- Outliers reflect actual market variations, such as luxury coastal homes vs. affordable inland housing, offering valuable insights rather than errors.
- Median income and location (ocean proximity, latitude/longitude) are the strongest predictors of house prices.
- The dataset is ideal for exploratory data analysis and serves as a solid foundation for regression modeling and predictive analytics.

# Future Scope:

- Develop predictive models using regression and machine learning techniques to estimate housing prices more accurately.
- Apply feature engineering (e.g., rooms per household, population density, geographic clustering) to enhance model performance and insights.
- Perform geospatial analysis with mapping tools to visualize housing trends and spatial patterns across California.
- Extend the study to affordability research and urban planning applications, supporting real estate development and data-driven policy decisions.

# THANK YOU