Project Title: Machine Learning Sem 10 - Regression on California Housing Dataset

Student Name: Donat Gosalci
Course: Machine Learning - Semester 10
Instructor: Msc. Freda Dyrkaj
Presentation Date: 26-30/05/2025

I. Data Wrangling (Preparing the Data)

1. Data Acquisition

- **Dataset**: California Housing dataset from `sklearn.datasets.fetch_california_housing()`
- **Shape**: 20640 rows, 9 columns (8 features + 1 target)
- **Features**: MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude
- **Target**: MedHouseVal (Median House Value)

2. Data Inspection and Exploration

- Data type: All columns are float64
- No missing values or duplicates were found
- Descriptive stats include:
    - Mean of MedHouseVal: 2.07 ($207,000)
    - Min: 0.15 ($15,000), Max: 5.00 ($500,000 cap)

3. Data Cleaning

- No null or duplicate rows; no imputation required.

4. Data Transformation

- Features (`X`) and target (`y`) separated
- Feature scaling (StandardScaler) applied after train-test split

---

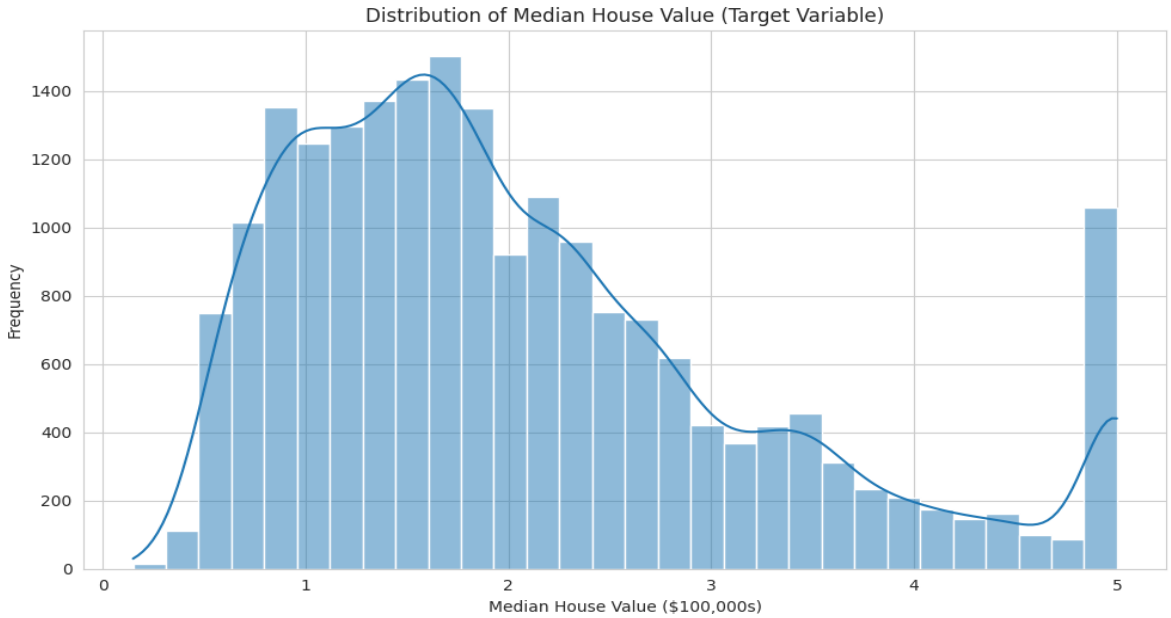
II. Data Analysis (Exploring and Understanding Patterns)

1. Descriptive Statistics

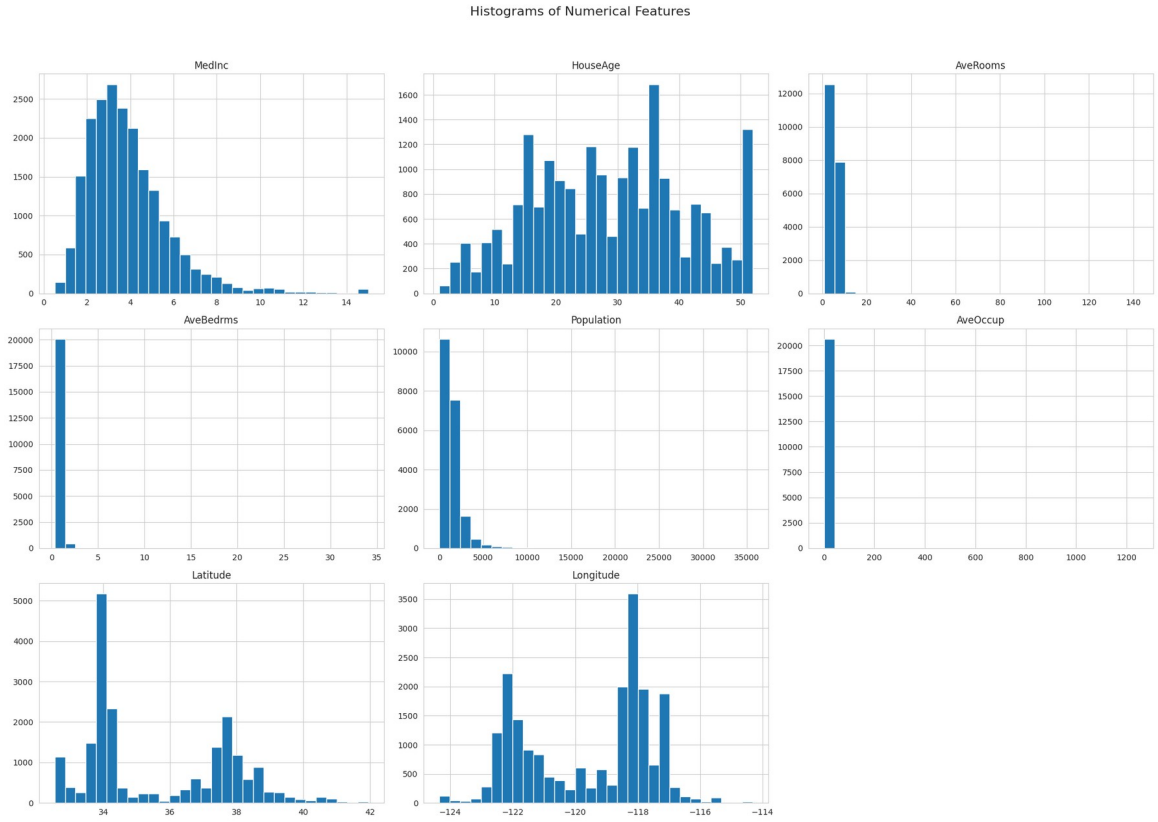
- Reviewed again to validate mean, std, and range of variables

2. Exploratory Data Analysis (EDA)

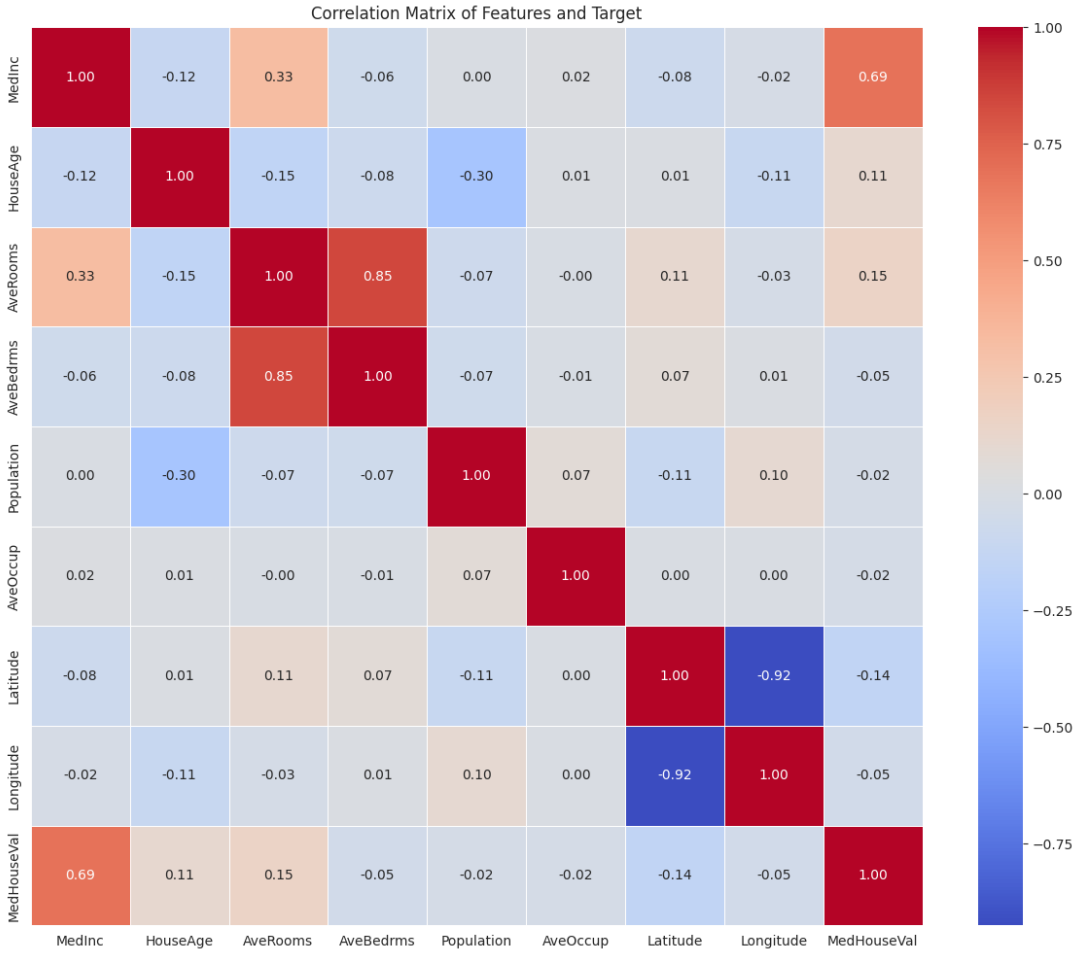
- **Visualizations Generated**:
  - Distribution of Median House Value:



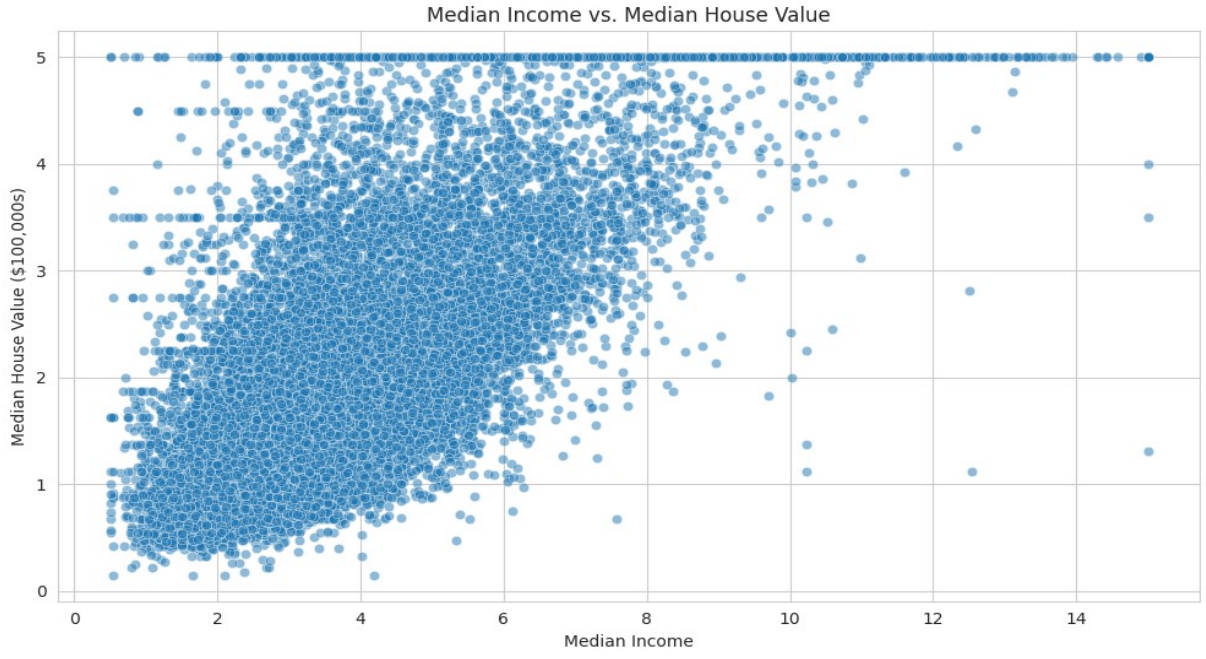
  - Histograms of all features:



- Correlation heatmap:



- Scatter plot of MedInc vs MedHouseVal:



- **Key Correlations with MedHouseVal**:
    - MedInc: **+0.69** (strongest positive correlation)
    - Latitude: -0.14, Longitude: -0.05

III. Regression Model Execution

1. Model Selection

- Linear Regression selected as baseline model

2. Train/Test Split

- 80% training, 20% test
- Shapes: X_train (16512, 8), y_train (16512), X_test (4128, 8), y_test (4128)

3. Feature Scaling

- StandardScaler applied

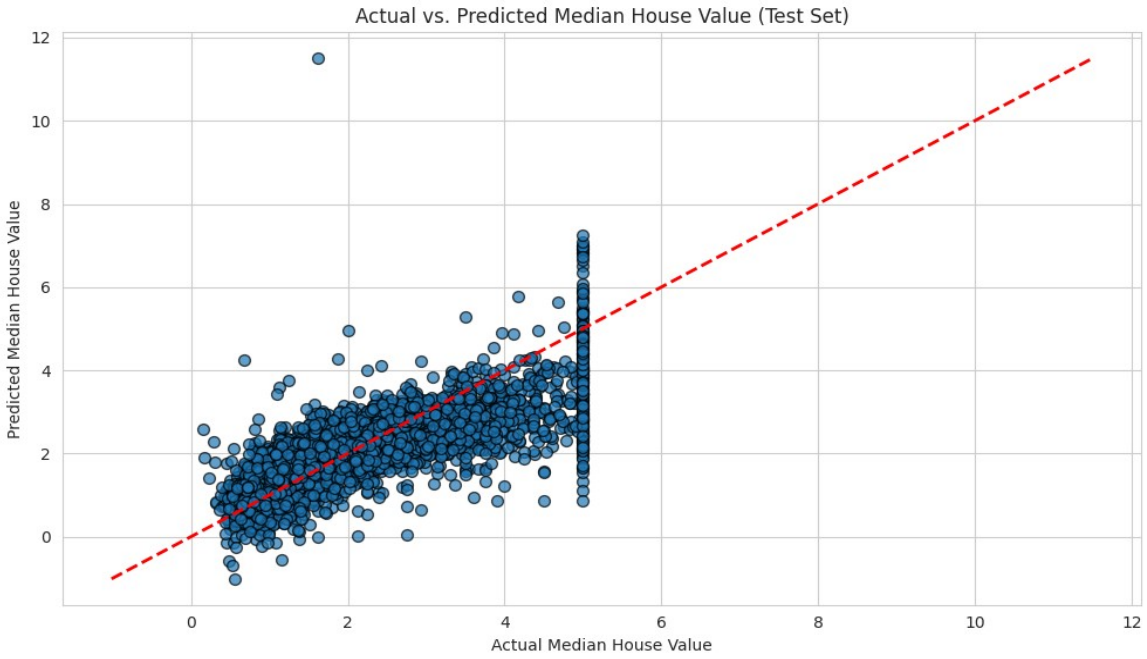
4. Model Training

- Trained using `LinearRegression()`
- **Coefficients:**
    - MedInc: +0.85
    - HouseAge: +0.12
    - AveRooms: -0.29
    - AveBedrms: +0.34
    - Latitude: -0.89
    - Longitude: -0.87

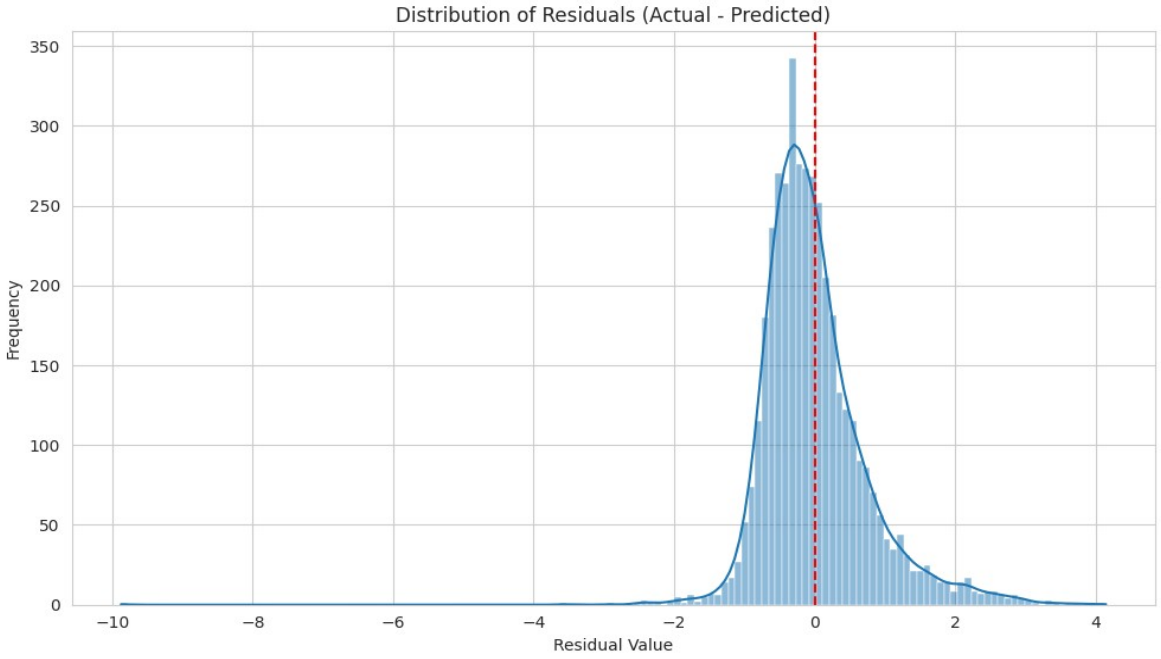
5. Model Evaluation

- **Mean Squared Error (MSE)**: 0.5559
- **Root Mean Squared Error (RMSE)**: 0.7456
- **R-squared (R^2)**: 0.5758 (moderate explanatory power)
- **Plots Generated**:

- Actual vs Predicted:



- Residuals Distribution:



IV. Documentation and Reporting

Summary of Findings:

- Dataset has 20640 samples and 8 numerical features
- Median Income has the strongest correlation with house value
- Linear Regression model provides a reasonable baseline ($R^2 \sim 0.58$)

- Visualizations helped understand feature relationships

Tools Used:

- **Python Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn
- **Notebook**: Python script `ml_project_sem10.py`

Final Thoughts:

- More complex models (e.g., Random Forest, Gradient Boosting) may improve R^2
- Outliers and capped values (at $500,000) likely affect performance
- Feature engineering or regional segmentation could enhance prediction accuracy

---

Repository/Reference: GitHub Repo (Sample Reference)