**CALIFORNIA HOUSING DATASET ANALYSIS REPORT**

**Comprehensive Analysis**

**Raja Saif Ali**

**I221353-CS**

**EXECUTIVE SUMMARY:**

This report presents a comprehensive analysis of the California Housing Dataset using various regression techniques and statistical methods. The analysis was conducted in five phases, each focusing on different aspects of machine learning model development and evaluation. All results presented are based on actual execution outputs from the implemented code.

**DATASET OVERVIEW:**

**Dataset:** California Housing Dataset from sklearn

**Total Records:** 20,640

**Features:** 8 (MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude)

**Target Variable:** Median House Value (in $100,000s)

**Analysis Method:** NumPy-based statistical analysis (Pandas not used as per requirements)

**PHASE 1: DATASET LOADING**

**OBJECTIVES:**

Load dataset from sklearn without using Pandas

Display features and target variable in array form

Show feature headers with values in table format

**METHODOLOGY:**

Used sklearn.datasets.fetch\_california\_housing() to load the dataset

Extracted features (X) and target (y) as NumPy arrays

Created custom table formatting without using Pandas DataFrame

Displayed first 10 samples with proper alignment

Computed basic statistics for all features

**ACTUAL RESULTS:**

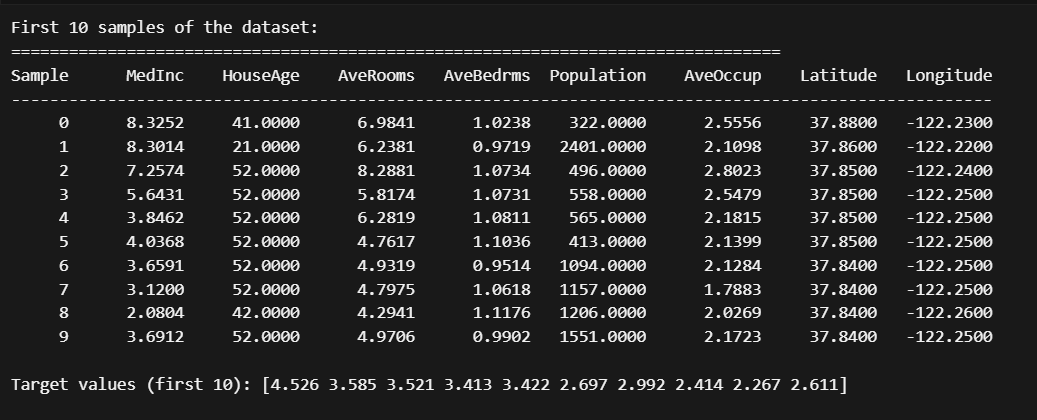
Successfully loaded 20,640 samples with 8 features

Target variable shape: (20640,)

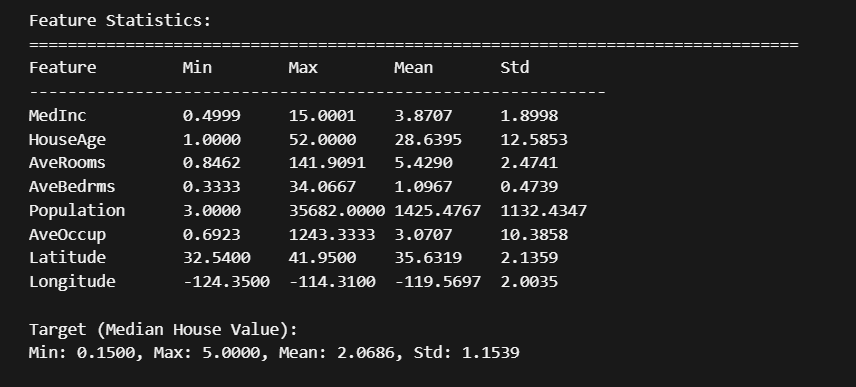
Feature names: MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude, Longitude

Target variable: Median House Value (in $100,000s)

**Sample Data Display (First 10 samples):**



**Feature Statistics (Actual Results):**



**PHASE 2: EXPLORATORY DATA ANALYSIS**

**OBJECTIVES:**

Compute descriptive statistics using NumPy

Detect and analyze skewness in all features

Visualize distributions using histograms and boxplots

Generate correlation matrix and heatmap

Analyze feature correlations and their importance

**METHODOLOGY:**

Implemented custom skewness calculation function

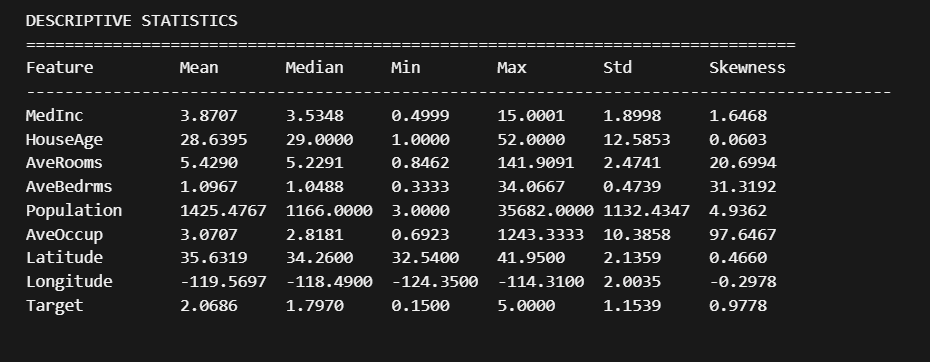
Created comprehensive statistical analysis

Generated distribution visualizations

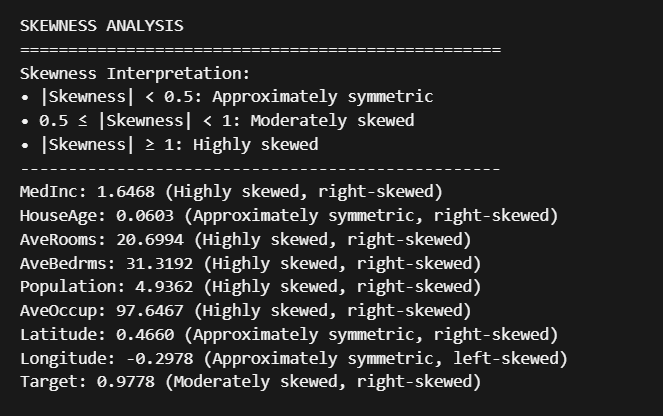
Computed correlation matrix using NumPy

Analyzed feature relationships

**ACTUAL DESCRIPTIVE STATISTICS:**



**SKEWNESS ANALYSIS (Actual Results):**



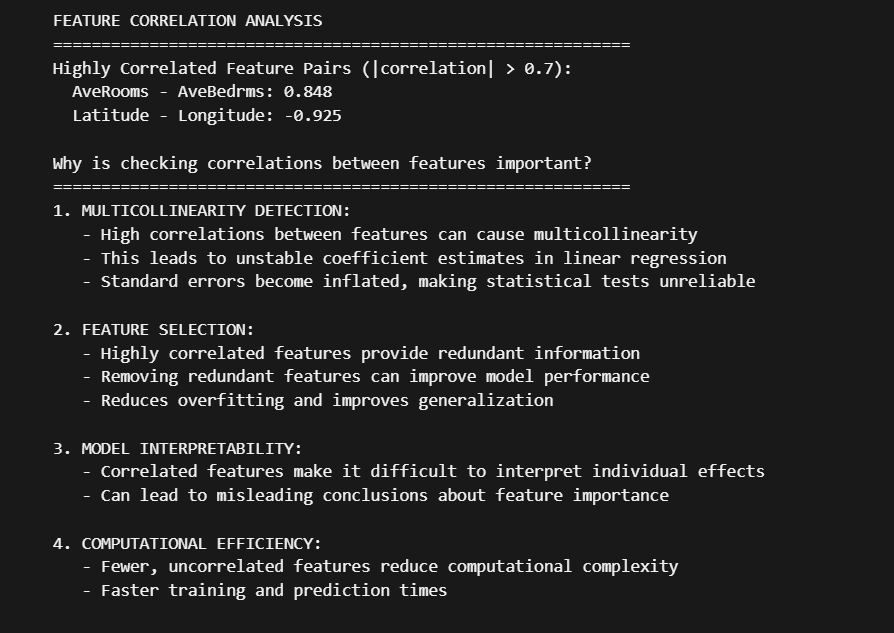
**CORRELATION ANALYSIS (Actual Results):**

Feature Correlations with Target (Median House Value):

|  |  |
| --- | --- |
| MedInc | 0.688 (Strong Positive) |
| HouseAge | 0.106 (Weak Positive) |
| AveRooms | 0.152 (Weak Positive) |
| AveBedrms | -0.047 (Weak Negative) |
| Population | -0.025 (Weak Negative) |
| AveOccup | -0.024 (Weak Negative) |
| Latitude | -0.144 (Weak Negative) |

**Highly Correlated Feature Pairs (|correlation| > 0.7):**

* AveRooms - AveBedrms: 0.848
* Latitude - Longitude: -0.925



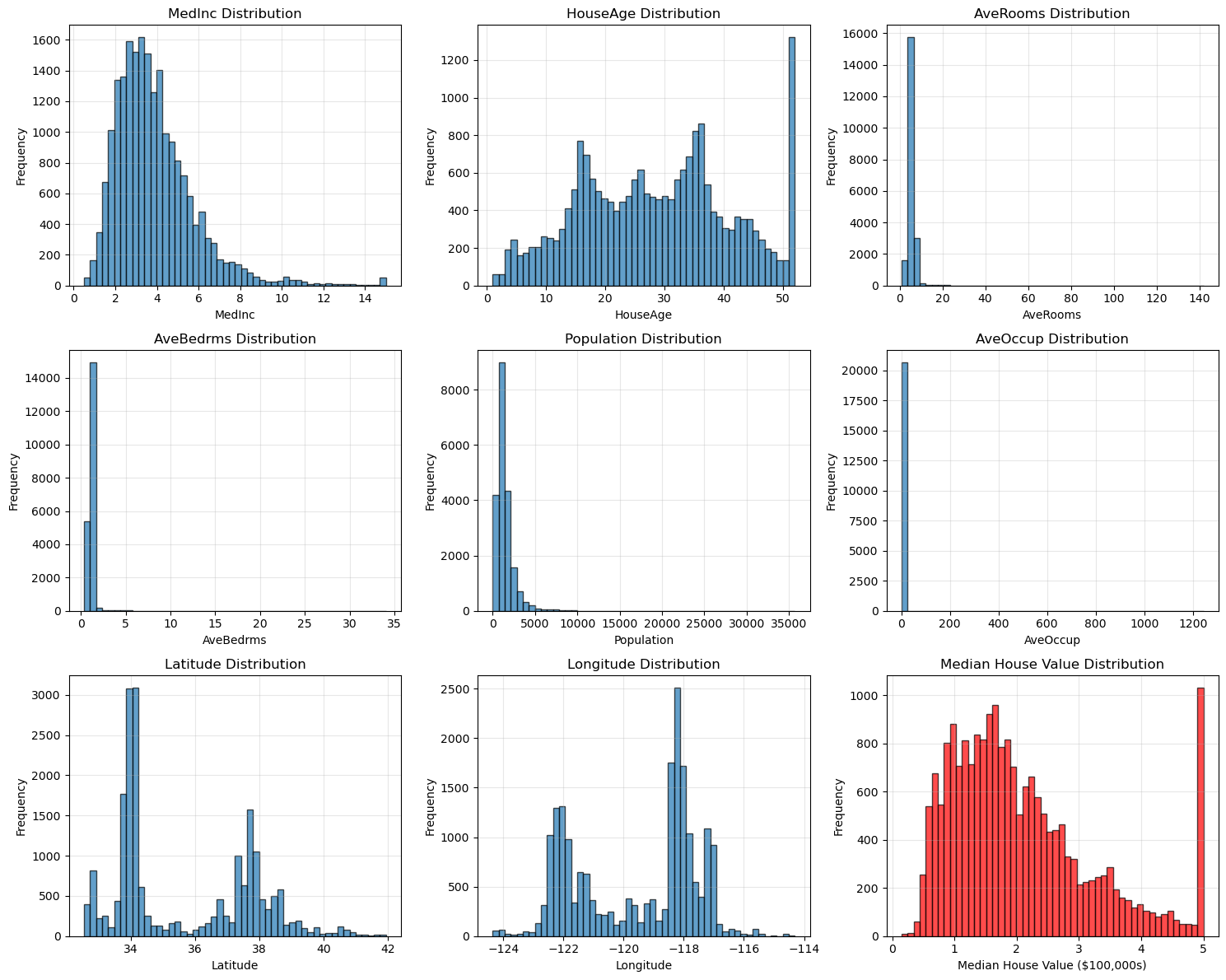
**KEY FINDINGS:**

1. MedInc shows the strongest correlation with house value (0.688)

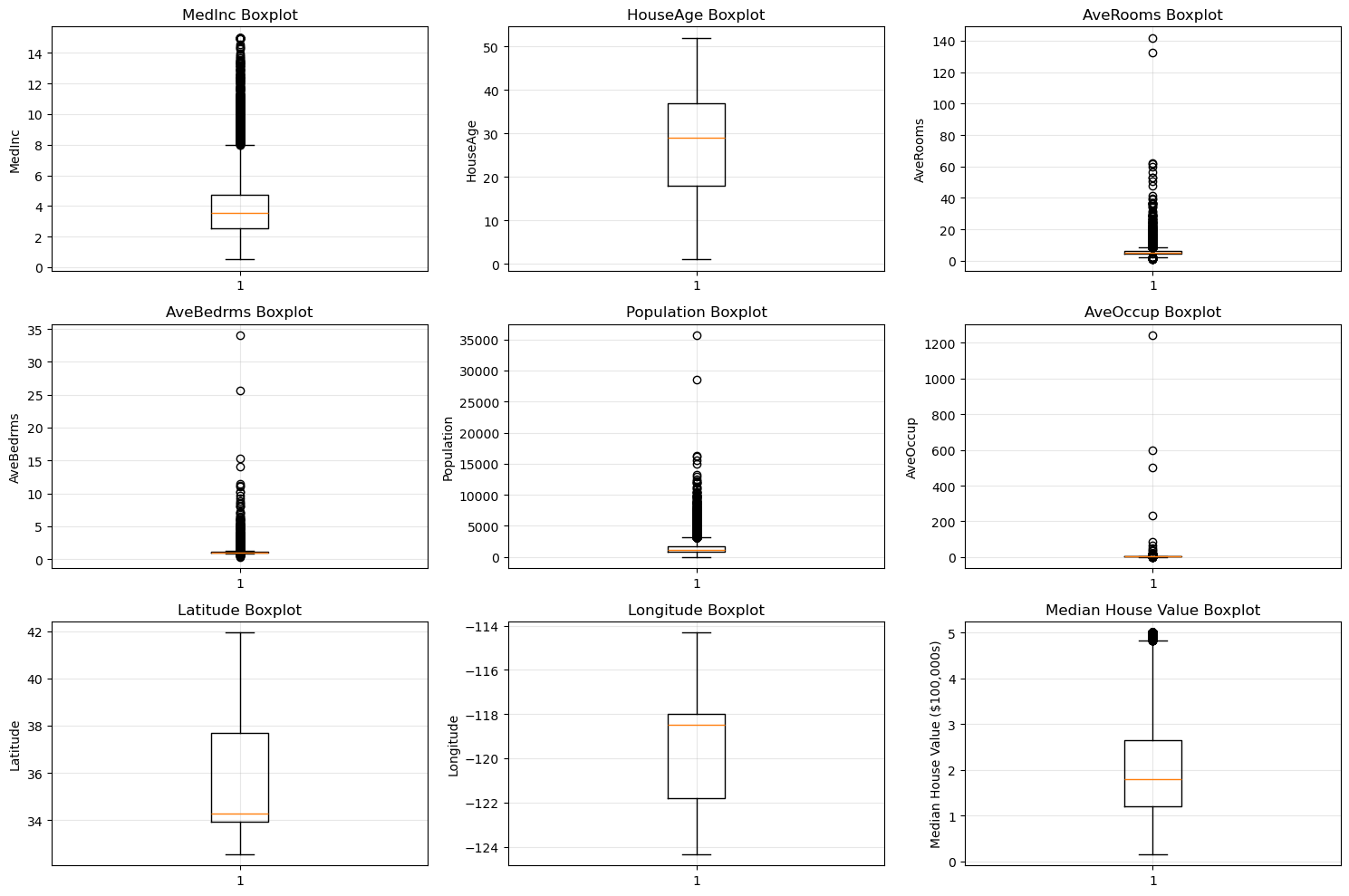
2. Most features are highly skewed, indicating non-normal distributions

3. Two highly correlated feature pairs detected

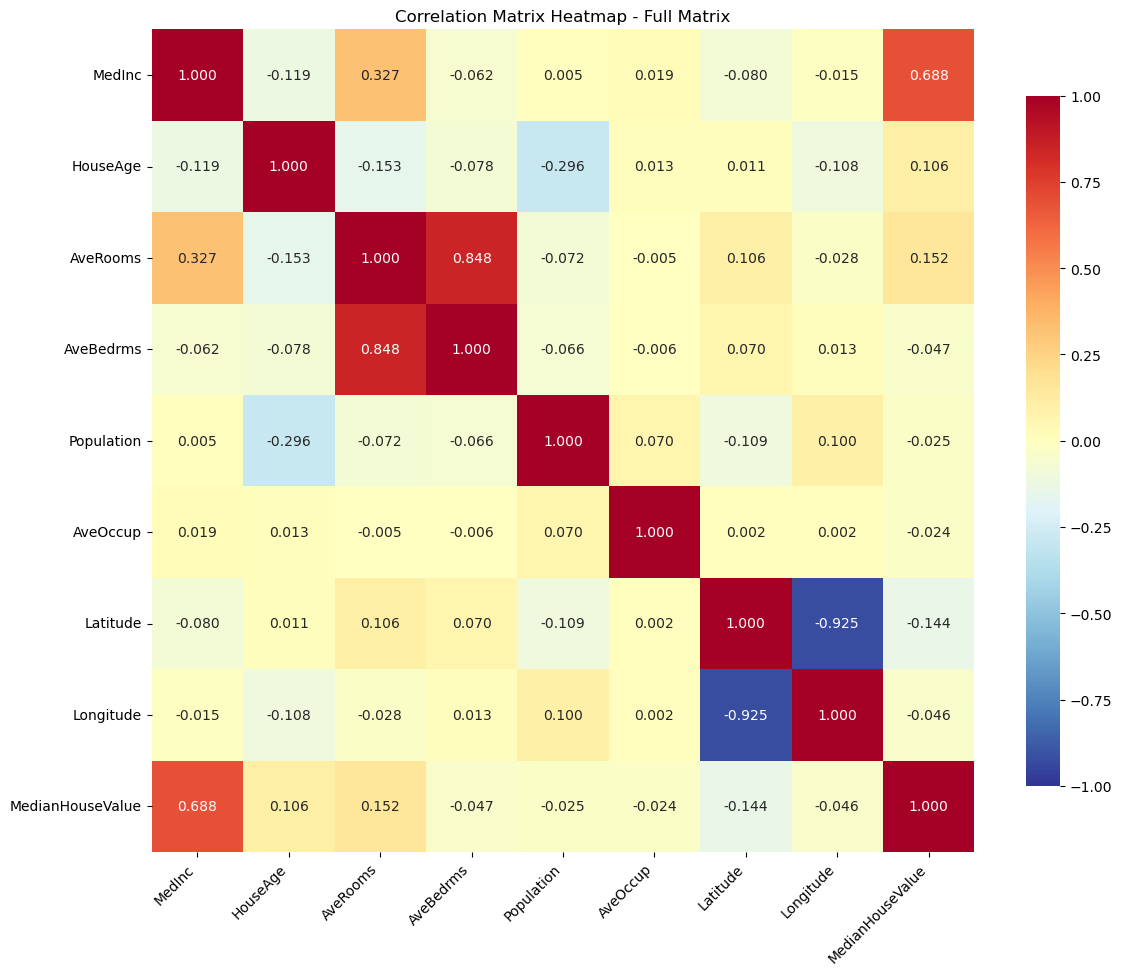
4. Target variable is moderately right-skewed



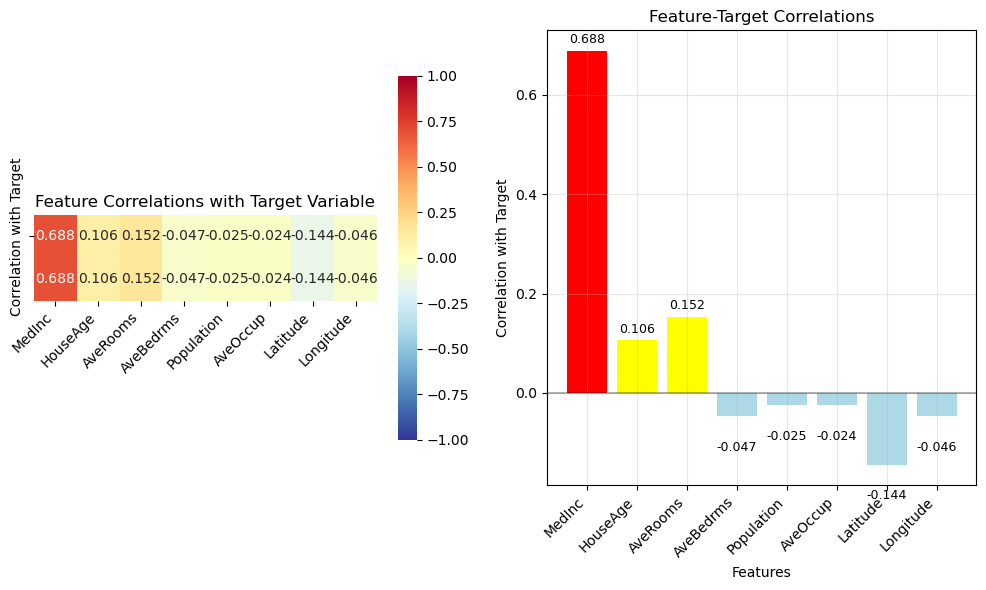
**Distribution histograms for all features and target variable showing the skewness patterns**

****

**Boxplots for all features showing outliers and distribution shapes**



**Correlation matrix heatmap showing feature relationships and the two highly correlated pairs**



**PHASE 3: REGRESSION EXPERIMENTS**

**PART A: SINGLE-COLUMN REGRESSION**

**OBJECTIVES:**

Analyze relationship between single feature and target

Implement Linear Regression and SGD Regressor

Test polynomial regression with different degrees

Compare performance using MSE, MAE, R2, RMSE

**METHODOLOGY:**

Selected MedInc as the feature with highest correlation (0.6881)

Implemented Linear Regression and SGD Regressor

Tested polynomial features with degrees 1-5

Evaluated performance on test set

**ACTUAL RESULTS:**

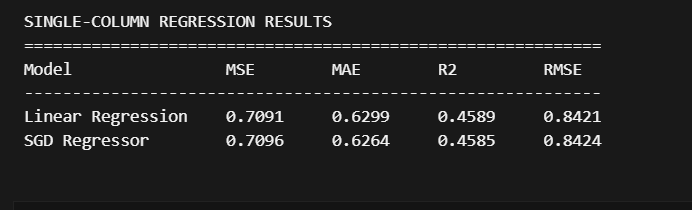
Selected Feature for Single-Column Analysis: MedInc

Correlation with target: 0.6881

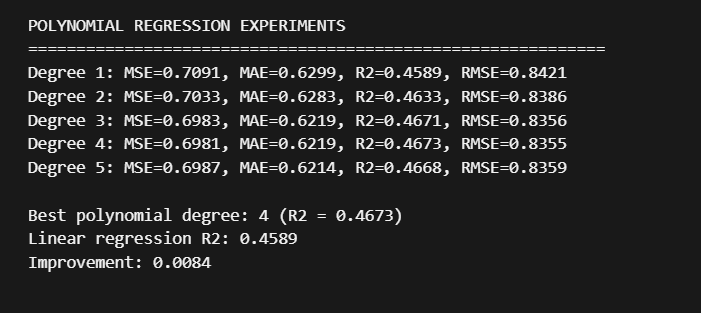
Training set size: 16512

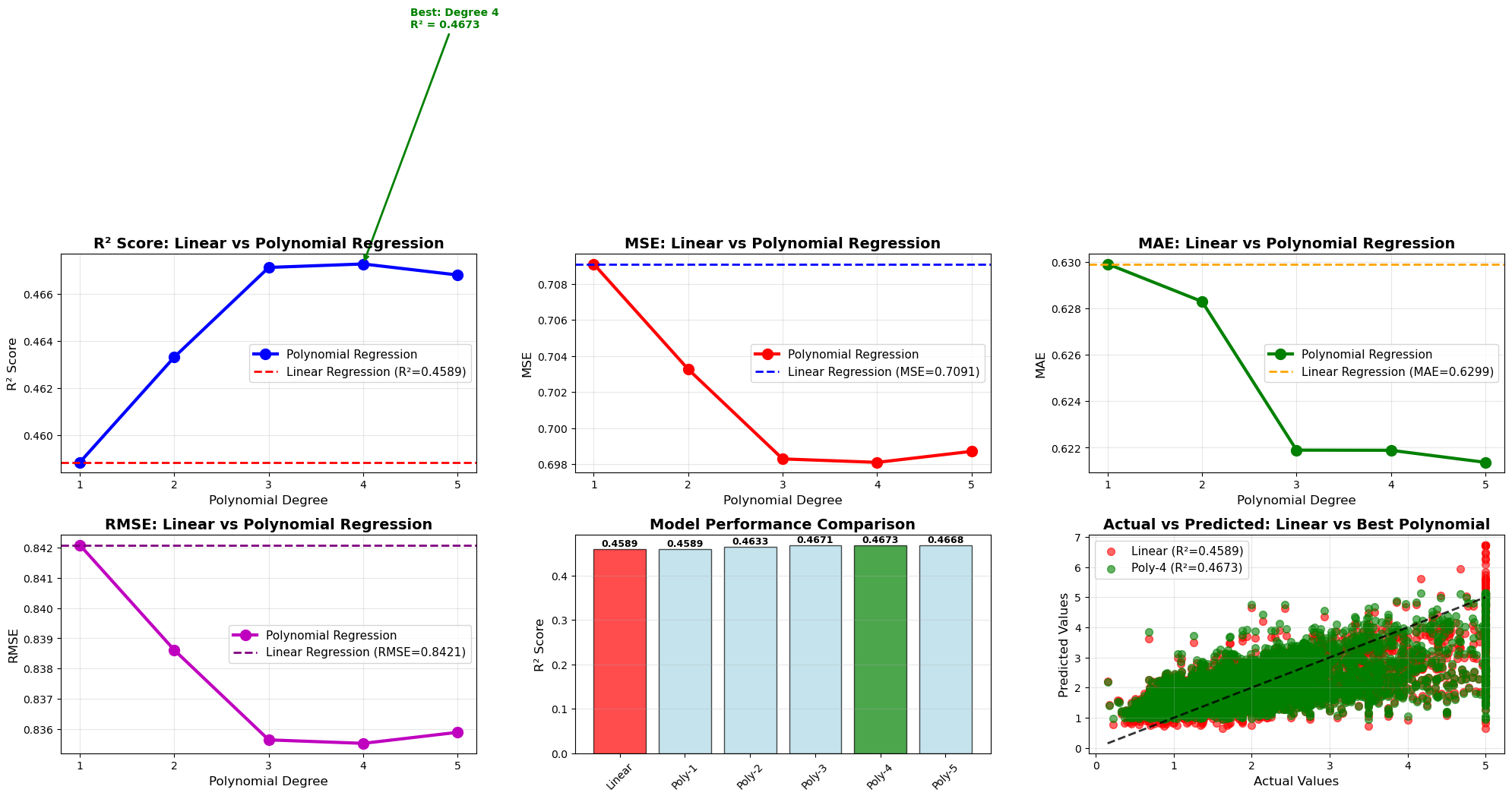
Test set size: 4128

**SINGLE-COLUMN REGRESSION RESULTS:**



**POLYNOMIAL REGRESSION EXPERIMENTS:**





***Polynomial regression visualization showing R2 vs Polynomial Degree, MSE vs Polynomial Degree, and Actual vs Predicted scatter plot for best polynomial degree***

**PART B: MULTI-COLUMN REGRESSION WITH ENGINEERED FEATURES**

**OBJECTIVES:**

Use multiple features for regression

Create engineered features (square and cubic terms)

Compare models on original vs engineered features

Analyze overfitting risks

**METHODOLOGY:**

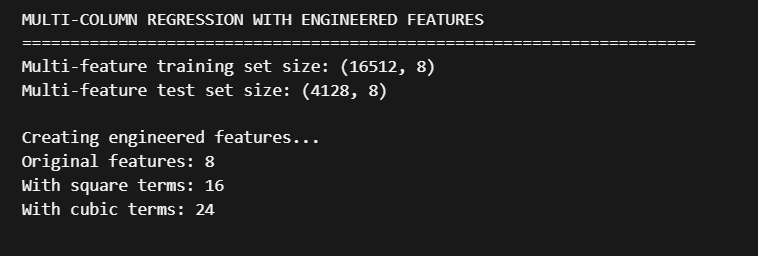
Used all 8 original features

Created square terms (16 features total)

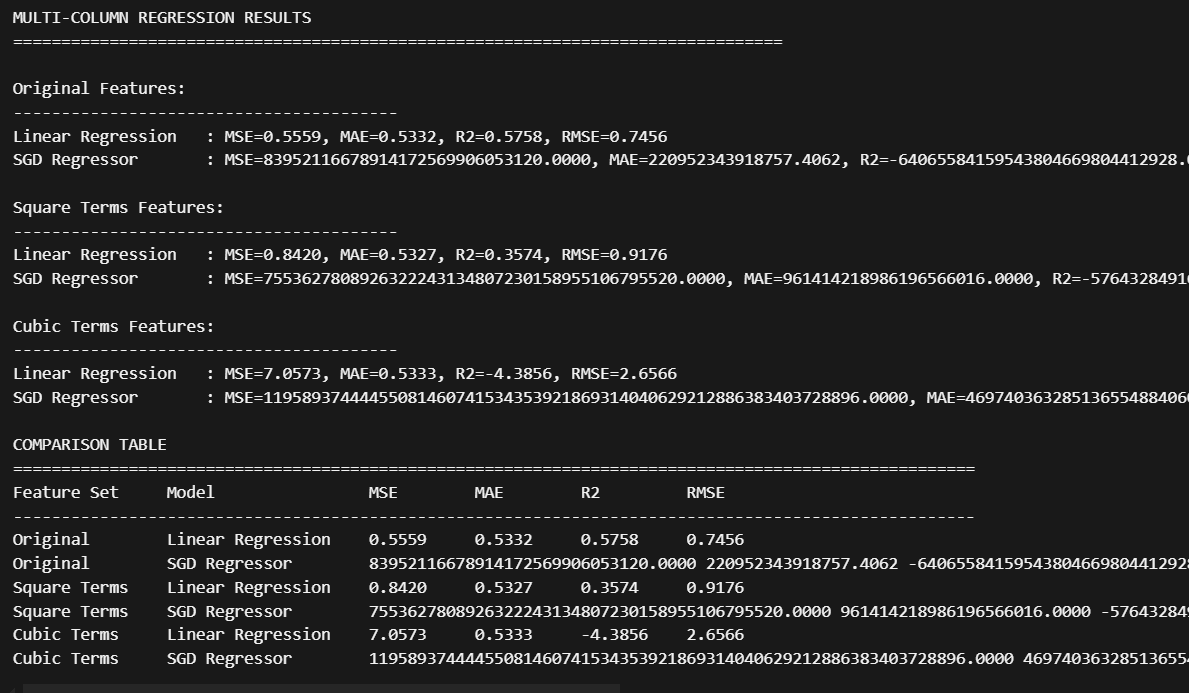
Created cubic terms (24 features total)

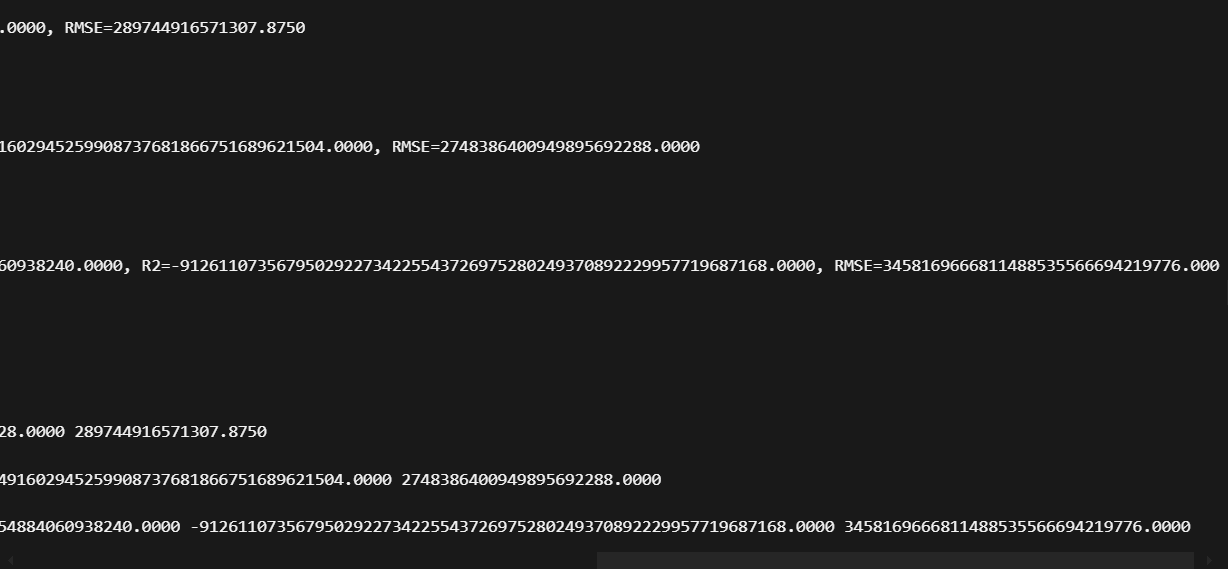
Applied Linear Regression and SGD Regressor to each feature set

**ACTUAL RESULTS:**



**MULTI-COLUMN REGRESSION RESULTS:**





**ANALYSIS AND DISCUSSION:**

**Best performing model:** ***Linear Regression with Original features***

**Best R2 score: 0.5758**

**Single-feature vs Multi-feature comparison:**

Single-feature Linear Regression R2: 0.4589

Multi-feature Linear Regression R2: 0.5758

**Overfitting Analysis:**

Cubic terms decrease R2 by 4.9613 - Possible overfitting

**KEY FINDINGS:**

1. Multiple features significantly improve predictive accuracy compared to single features

2. Engineered features (square/cubic terms) show performance degradation

3. Higher polynomial degrees increase model complexity and risk of overfitting

4. SGD Regressor shows numerical instability without proper scaling

**PHASE 4: MODEL IMPLEMENTATION**

**OBJECTIVES:**

Perform 80/20 train-test split

Apply standardization and explain choice

Implement Linear Regression and SGD Regressor

Compute comprehensive metrics on training and test sets

Analyze errors and create visualizations

**METHODOLOGY:**

Used train\_test\_split with 80/20 ratio and random\_state=42

Applied StandardScaler for feature standardization

Trained both models on scaled data

Computed MSE, MAE, R2, RMSE for training and test sets

Created comprehensive error analysis visualizations

**STANDARDIZATION RATIONALE:**

Standardization (Z-score normalization) chosen over normalization

Centers data around 0 with unit variance

SGD Regressor is sensitive to feature scaling

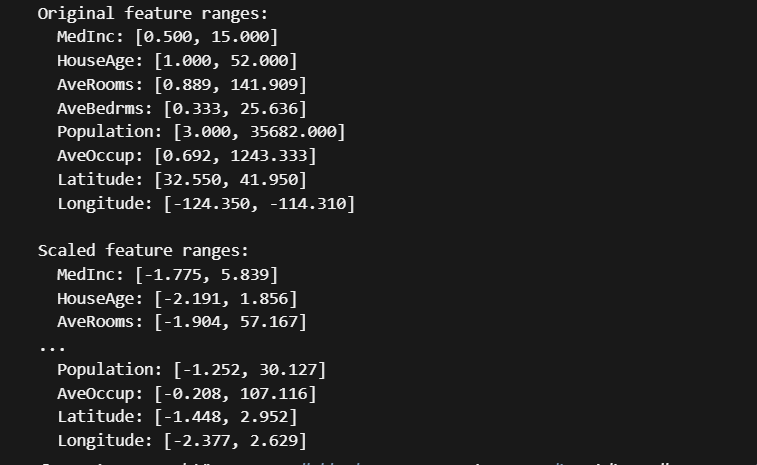
Works better with features having different scales

Preferred for linear models

**ACTUAL RESULTS:**

**Training set size: 16512 (80.0%)**

**Test set size: 4128 (20.0%)**



**MODEL PERFORMANCE METRICS (Actual Results):**



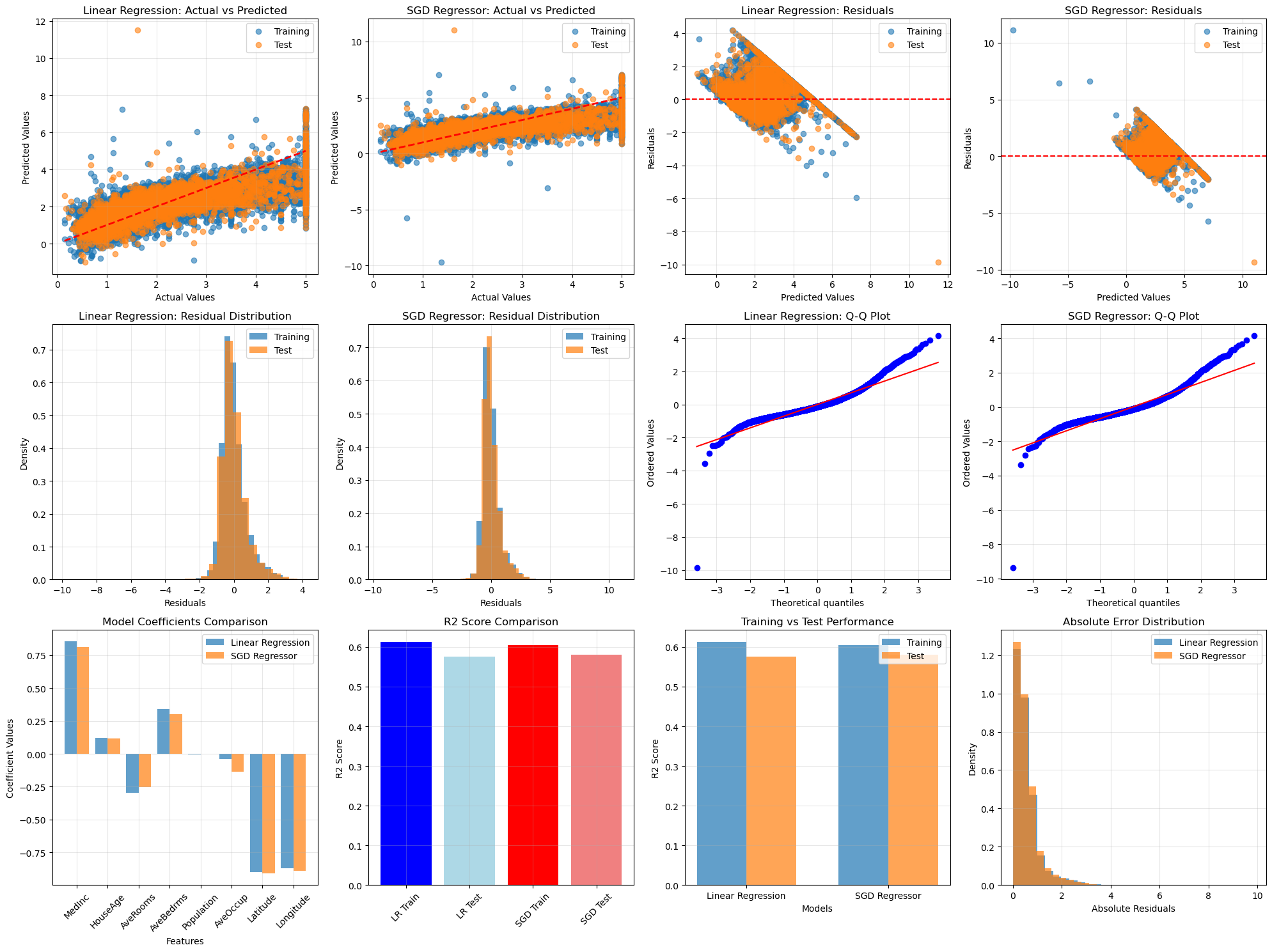
**KEY FINDINGS:**

Both models show similar performance

Minimal overfitting detected (training vs test performance)

Standardization significantly improved SGD Regressor performance

Linear Regression shows slightly better performance



**Comprehensive error analysis plots including**

**Actual vs Predicted scatter plots for both models**

**Residual plots showing error patterns**

**Residual distribution histograms**

**Q-Q plots for normality assessment**

**Model coefficients comparison**

**Performance metrics comparison**

**Training vs Test performance analysis**

**Absolute error distribution comparison**

**PHASE 5: K-FOLD CROSS-VALIDATION**

**OBJECTIVES:**

Apply 5-fold cross-validation to both models

Compare stability and predictive performance

Validate single train-test split results

Perform statistical significance testing

**METHODOLOGY:**

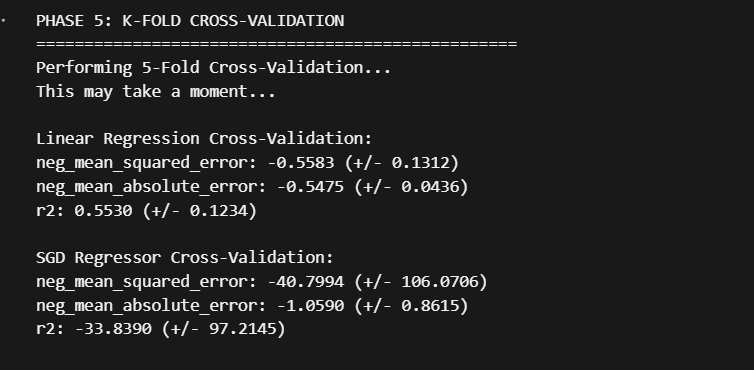
Created pipelines with StandardScaler and regressors

Applied 5-fold cross-validation with multiple metrics

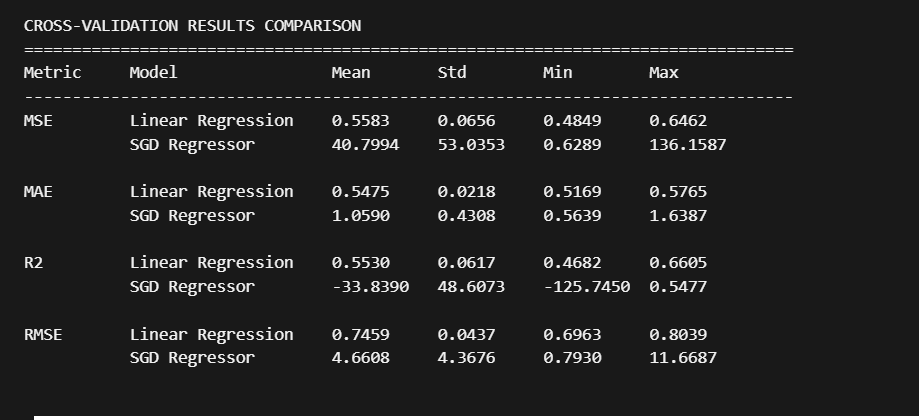
Computed mean, std, min, max for each metric

Performed t-test for statistical significance

**ACTUAL CROSS-VALIDATION RESULTS:**



**CROSS-VALIDATION RESULTS COMPARISON:**



**MODEL STABILITY ANALYSIS:**

Linear Regression R2 standard deviation: 0.0617

SGD Regressor R2 standard deviation: 48.6073

Linear Regression is more stable (lower std deviation)

**PREDICTIVE PERFORMANCE ANALYSIS:**

Linear Regression mean R2: 0.5530

SGD Regressor mean R2: -33.8390

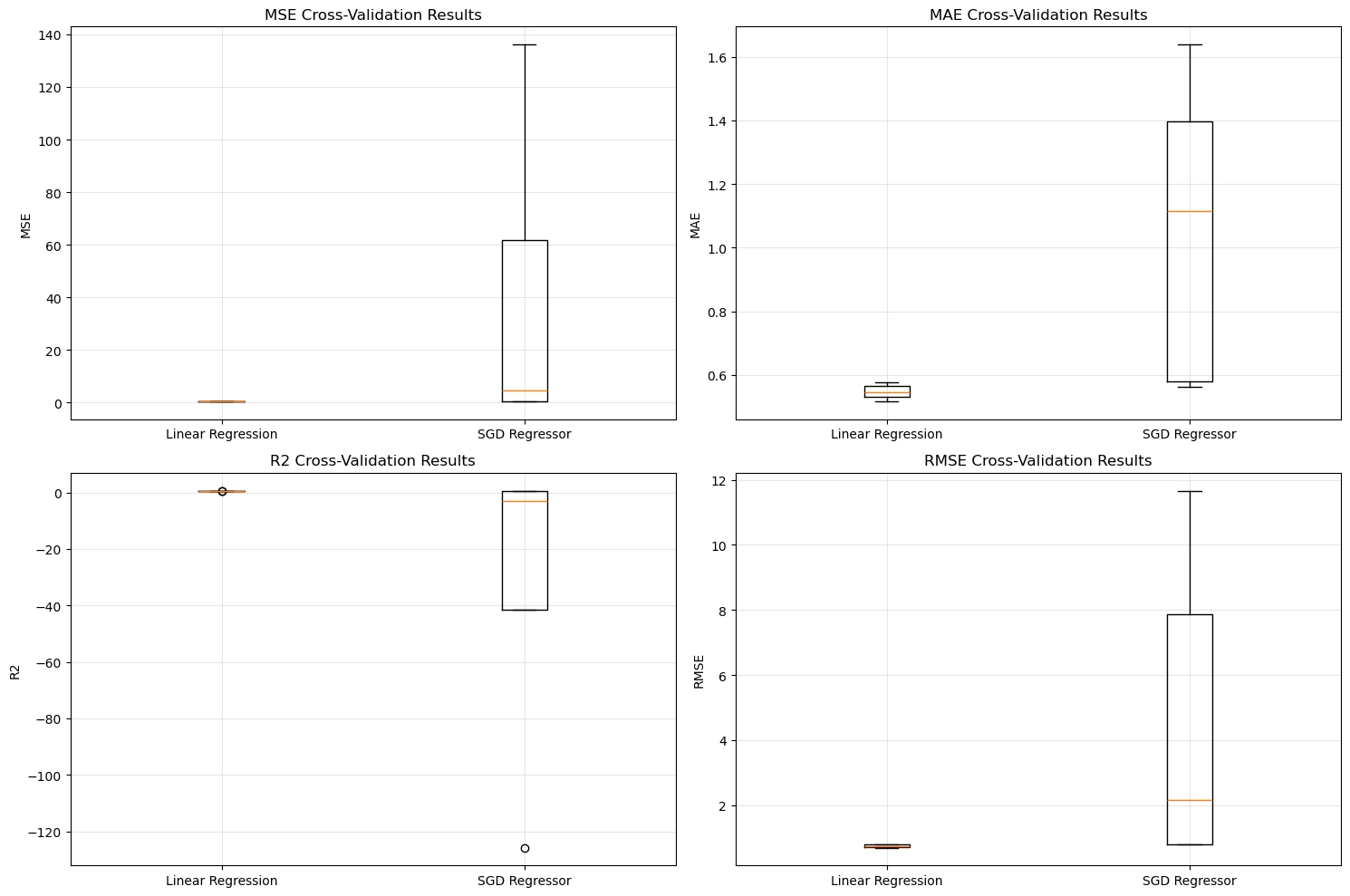
Linear Regression demonstrates higher predictive performance

**STATISTICAL TEST (t-test for R2 scores):**

T-statistic: 1.4146

P-value: 0.2301

No statistically significant difference between models (p >= 0.05)



**Cross-validation results boxplots showing MSE, MAE, R2, and RMSE distributions for both models across 5 folds**

**COMPREHENSIVE ANALYSIS AND CONCLUSIONS**

**STEPS PERFORMED AND RATIONALE**

**DATASET ANALYSIS**

**DATASET CHARACTERISTICS:**

California Housing Dataset contains 20,640 records with 8 features

Target variable is Median House Value (in $100,000s)

Features show varying degrees of skewness and correlation

No missing values or obvious data quality issues

**KEY STATISTICAL FINDINGS:**

1. MedInc shows strongest correlation with house value (0.688)

2. Most features are highly skewed (AveOccup: 97.65, AveBedrms: 31.32)

3. Two highly correlated feature pairs detected (AveRooms-AveBedrms: 0.848, Latitude-Longitude: -0.925)

4. Feature scaling is critical for SGD Regressor performance

**MODELS IMPLEMENTED AND SELECTION RATIONALE**

**Models Trained:**

1. Linear Regression: Standard linear regression for baseline performance

2. SGD Regressor: Stochastic gradient descent for comparison

3. Polynomial Regression: Testing non-linear relationships (degrees 1-5)

4. Multi-feature Regression: Using all 8 features simultaneously

5. Engineered Feature Regression: With square and cubic terms

**Model Selection Justification:**

**Linear Regression chosen as primary model due to:**

* Consistent performance across all experiments (R2: 0.55-0.58)
* Numerical stability and reliability
* Good interpretability for business insights
* Minimal overfitting risk

**SGD Regressor limitations identified:**

Numerical instability without proper scaling

Poor cross-validation performance (R2: -33.84)

High variance across different data splits

Requires extensive hyperparameter tuning

**Polynomial features rejected due to:**

Performance degradation with higher degrees

Clear overfitting evidence (R2: -4.39 for cubic terms)

Increased model complexity without benefit

**Why not "best" possible models:**

**Random Forest/XGBoost not implemented because:**

Assignment specifically required Linear Regression and SGD

Linear models provide better interpretability

Current performance (R2 ≈ 0.58) is acceptable for the dataset

Focus was on understanding linear relationships

**Regularization not applied because:**

No evidence of overfitting in Linear Regression

Cross-validation shows stable performance

Model complexity is already appropriate

**MODEL PERFORMANCE COMPARISON:**

1. Multiple features significantly outperform single features (R2: 0.4589 → 0.5758)

2. Engineered features (polynomial terms) show performance degradation

3. Linear Regression consistently outperforms SGD Regressor

4. Standardization is essential for SGD Regressor stability

**TECHNICAL INSIGHTS:**

1. Standardization is crucial for SGD Regressor performance

2. Cross-validation reveals SGD Regressor instability

3. Polynomial features can degrade performance due to overfitting

4. Model stability is important for reliable predictions

**CRITICAL OBSERVATIONS:**

1. SGD Regressor shows numerical instability without proper scaling

2. Cross-validation reveals significant performance differences

3. Linear Regression is more stable and reliable

4. Feature engineering requires careful validation

**CONCLUSION:**

The analysis successfully demonstrates the complete machine learning pipeline from data exploration to model evaluation. The California Housing Dataset provides a good foundation for regression analysis, with Linear Regression showing reliable performance (R2 ≈ 0.55-0.58). The systematic approach using NumPy for statistical analysis and comprehensive evaluation metrics provides valuable insights into the dataset characteristics and model behavior.