California Housing Price Analysis Report

1. Introduction

This report presents an analysis of the California Housing dataset, which contains information about housing prices in California. The objective of this analysis is to understand the factors influencing housing prices and to build a predictive model using machine learning techniques.

2. Dataset Overview

The California Housing dataset, obtained from Scikit-learn's dataset repository, includes the following features:
- MedInc: Median income in block group
- HouseAge: Median house age in block group
- AveRooms: Average number of rooms per household
- AveBedrms: Average number of bedrooms per household
- Population: Block group population
- AveOccup: Average number of household members
- Latitude: Block group latitude
- Longitude: Block group longitude

The target variable is the median house value for California districts, expressed in hundreds of thousands of dollars.

3. Methodology

3.1 Data Cleaning and Preprocessing
- The dataset was loaded using Pandas.
- A check for missing values was performed, revealing no missing data.
- Column names were cleaned using string manipulation techniques, converting them to lowercase and removing spaces.

3.2 Exploratory Data Analysis
We analyzed the 'MedInc' (median income) feature using NumPy:
- Mean:3.87
- Median: 3.53

3.3 Model Development
- The dataset was split into training (80%) and testing (20%) sets.
- A Linear Regression model was chosen for its simplicity and interpretability.
- The model was trained on the training set and evaluated on the testing set.

4. Results

4.1 Model Performance
- Mean Squared Error: 0.56

- R-squared Score:0.58

The Mean Squared Error (MSE) indicates the average squared difference between predicted and actual house prices. A lower MSE suggests better predictions. The R-squared score represents the proportion of variance in the dependent variable that is predictable from the independent variables. An R-squared closer to 1 indicates a better fit.

4.2 Feature Importance

	feature	importance
3	avebedrms	0.783145
0	medinc	0.448675
1	houseage	0.009724
4	population	-0.000002
5	aveoccup	-0.003526
2	averooms	-0.123323
6	latitude	-0.419792
7	longitude	-0.433708

The feature importance analysis reveals which factors have the strongest influence on housing prices in our model.

5. Discussion

The model's performance, as indicated by the R-squared value, suggests that it explains a significant portion of the variance in housing prices. This aligns with common understanding that factors such as location and local income levels often play crucial roles in determining property values.

6. Limitations and Future Work

- Linear Regression assumes a linear relationship between features and the target variable, which may not capture all complexities of the housing market.
- The dataset is historical and may not reflect current market conditions.
- Future work could involve:
  1. Exploring non-linear models (e.g., Random Forests, Gradient Boosting) to potentially capture more complex relationships.
  2. Incorporating additional features such as proximity to amenities or school district ratings.
  3. Performing more advanced feature engineering to extract more information from the existing data.

7. Conclusion

This analysis provides insights into the factors influencing California housing prices and demonstrates the application of machine learning techniques to predict these prices. While the linear regression model shows promising results, there is room for improvement and further exploration. The findings can be valuable for real estate professionals and researchers interested in understanding and predicting housing market trends in California.