California Housing Prices

Introduction

The California Housing Prices dataset is widely used in machine learning and data science for regression tasks. It contains information about housing districts in California, collected during the 1990 U.S. Census. The dataset is often used to predict median house values based on various demographic and geographic features.

Context

Housing affordability and real estate market trends have always been important topics in public policy and urban planning. This dataset provides an opportunity to analyze how factors such as income, population, housing age, and location affect housing prices. It is a standard dataset for teaching regression techniques and evaluating machine learning models.

Content

The dataset (housing.csv) contains information about housing districts with the following columns:

- longitude → Longitude coordinate of the district
- latitude → Latitude coordinate of the district
- housing_median_age → Median age of houses in the district
- total_rooms → Total number of rooms within a district
- total_bedrooms → Total number of bedrooms within a district
- population → Population of the district
- households → Number of households in the district
- median_income → Median income of households in the district (in tens of thousands of dollars)
- median_house_value → Median house value for households within a district (target variable)
- ocean_proximity → Location of the district relative to the ocean (categorical variable)

Acknowledgements

This dataset originates from the 1990 U.S. Census and was first made popular through the book 'Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow' by Aurélien Géron. It is hosted by various machine learning repositories and is commonly used for regression analysis and educational purposes.

Source Reference:
- Aurélien Géron: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow
- UCI Machine Learning Repository (California Housing Dataset)