Project 1 - California Housing Price Prediction .

Writeup

**Background of Problem Statement :**

The US Census Bureau has published California Census Data which has 10 types of metrics such as the population, median income, median housing price, and so on for each block group in California. The dataset also serves as an input for project scoping and tries to specify the functional and nonfunctional requirements for it.

**Problem Objective :**

The project aims at building a model of housing prices to predict median house values in California using the provided dataset. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

Districts or block groups are the smallest geographical units for which the US Census Bureau  
publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset.

**Domain**: Finance and Housing

**Analysis Tasks performed:**

1. Explore the data set to understand the relevant features that could be used to build the linear regression model
2. Clean up the dataset to remove and
3. Convert categorical data to dummy variables
4. Split and standardize the input features
5. Perform linear regression model using the relevant attributes
6. Perform linear regression on a single variable to compare against the previous model and interpret the outcome

**Conclusion:**

Although the single dependent variable model may not perform as well as the multi-linear regression model, the root mean square error is quite close to each other. This indicates that median income is one of the key features required in linear regression models.