**California Housing Price Prediction Project**

**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.

The data is read from the excel file. Below are the first 5 records.

Graphical user interface

Description automatically generated

The data is plotted on a histogram to understand the data better.

Chart, box and whisker chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

The above chart shows that median house value is more when location is near the ocean.

Text

Description automatically generated

There is a strong correlation between median\_house\_value and median\_income.

Chart, scatter chart

Description automatically generated

The data is cleaned by removing Nan values.

A picture containing text

Description automatically generated

The total\_bedrooms has 207 Null values. The null values are replaced by the mean value.

Ocean\_proximity is a category column. It is converted to numerical by using dummy variables.

Table

Description automatically generated

Table

Description automatically generated with medium confidence

The resulting dataframe now has 14 columns after concatenating the numerical columns with dummy variable columns.

The X (independent variables) and Y(dependent variable) are now extracted.

Y is median\_house\_value. X is taken as the remaining columns.

Graphical user interface, application

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Data is split into train (80 %) and test(20%)

Graphical user interface, text, application

Description automatically generated

Standardize the training and test datasets. Only the independent variables are standardized.

Graphical user interface, text

Description automatically generated

Now we perform Linear Regression.

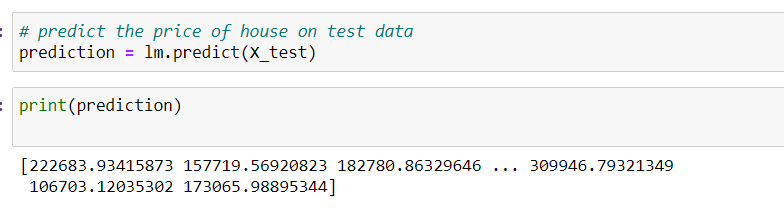
We get the coefficient (slope) and y -intercept values. The equations for straight line is y = mx + c. Since here we have multiple independent variables, we use multiple linear regression equation.

Y = m1x1 + m2x2 + …mnxn + c. Here m1,m2 …mn are coefficients. c is y -intercept.

Graphical user interface, text, application

Description automatically generated

Prediction can be done now since we know the coefficient and intercept.



Text

Description automatically generated

Text

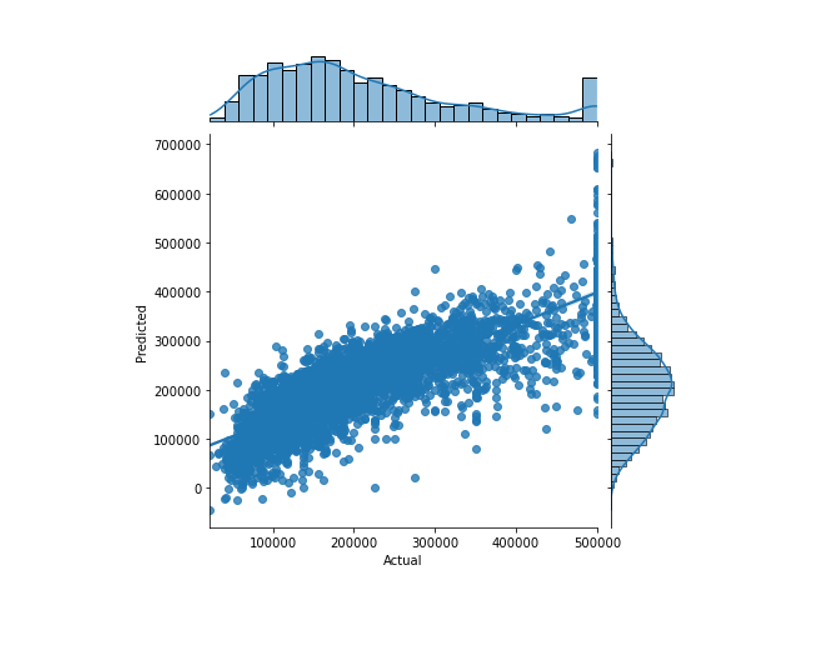
Description automatically generated with medium confidence

Text, letter

Description automatically generated

Chart, line chart, histogram

Description automatically generated



The above chart shows actual vs predicted values. We can see that the model does a decent job of prediction, though there is some margin of error.