**ABSTRACT**:

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. House prices increase every year, so there is a need for a system to predict house prices in the future. House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. There are three factors that influence the price of a house which include physical conditions, concept and location.

**PROBLEM STATEMENT**:

The purpose of the project is to predict median house values in California districts, given many features from these districts .The feature variables available can be used analyze and predict house prices.

The project also aims at building a model of housing prices in California using the California census data. The data has metrics for each block group in California, such as:

* Longitude
* Latitude
* housing\_median\_age
* total\_rooms
* total\_bedrooms
* population
* households
* median\_income
* ocean\_proximity
* median\_house\_value

This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

**APPROACH:**

We are going to break everything into logical steps that allow us to ensure the cleanest, most realistic data for our model to make accurate predictions from.

1. Load Data and Packages
2. Analyzing the Test Variable (Sale Price)
3. Multivariable Analysis
4. Impute Missing Data and Clean Data
5. Feature Transformation/Engineering
6. Modeling and Predictions

In this project*,* a test-driven approach to build a **Linear Regression** model using Python is used. The trained model will be further used to predict house sale prices.

**LINEAR REGRESSION:**

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.