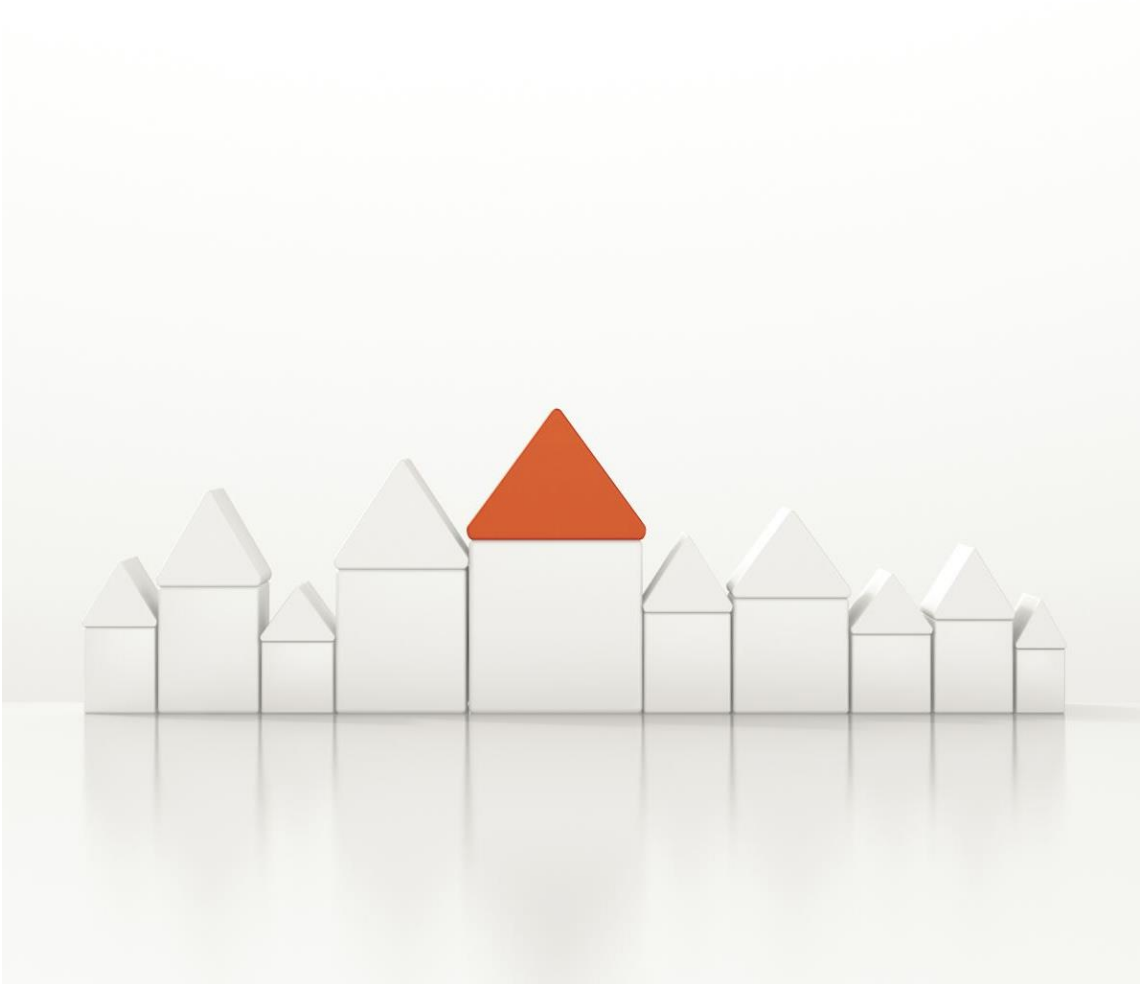
CS556-A

HOUSE PREDECTION USING LINEAR REGRESSION

Introduction

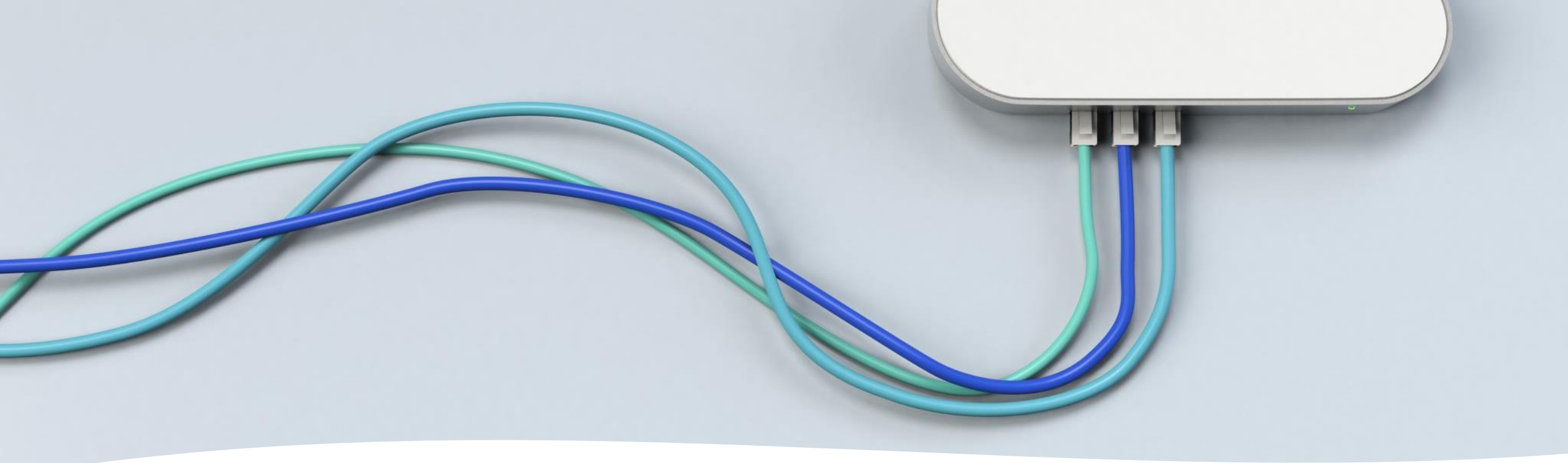
- Based on the information provided, we have estimated the prices of homes.
- We have calculated the median income of the houses using information from the data set, including latitude, longitude, the number of rooms, population, and median income.
- In order to forecast the median house prices, we choose linear regression model.



Introduction

- The data pertains to the houses found in each California district and some summary statistics about them based on 1900 census data.
- It contains one instance per district block group
- A block is the small geographical unit for which the U.S census bureau publishes the sample data.
- Our target variable in this project is median house value.

Imported Packages



Loading the Data

- Use pandas read_csv() function to load the dataset.

Dropping Missing Values

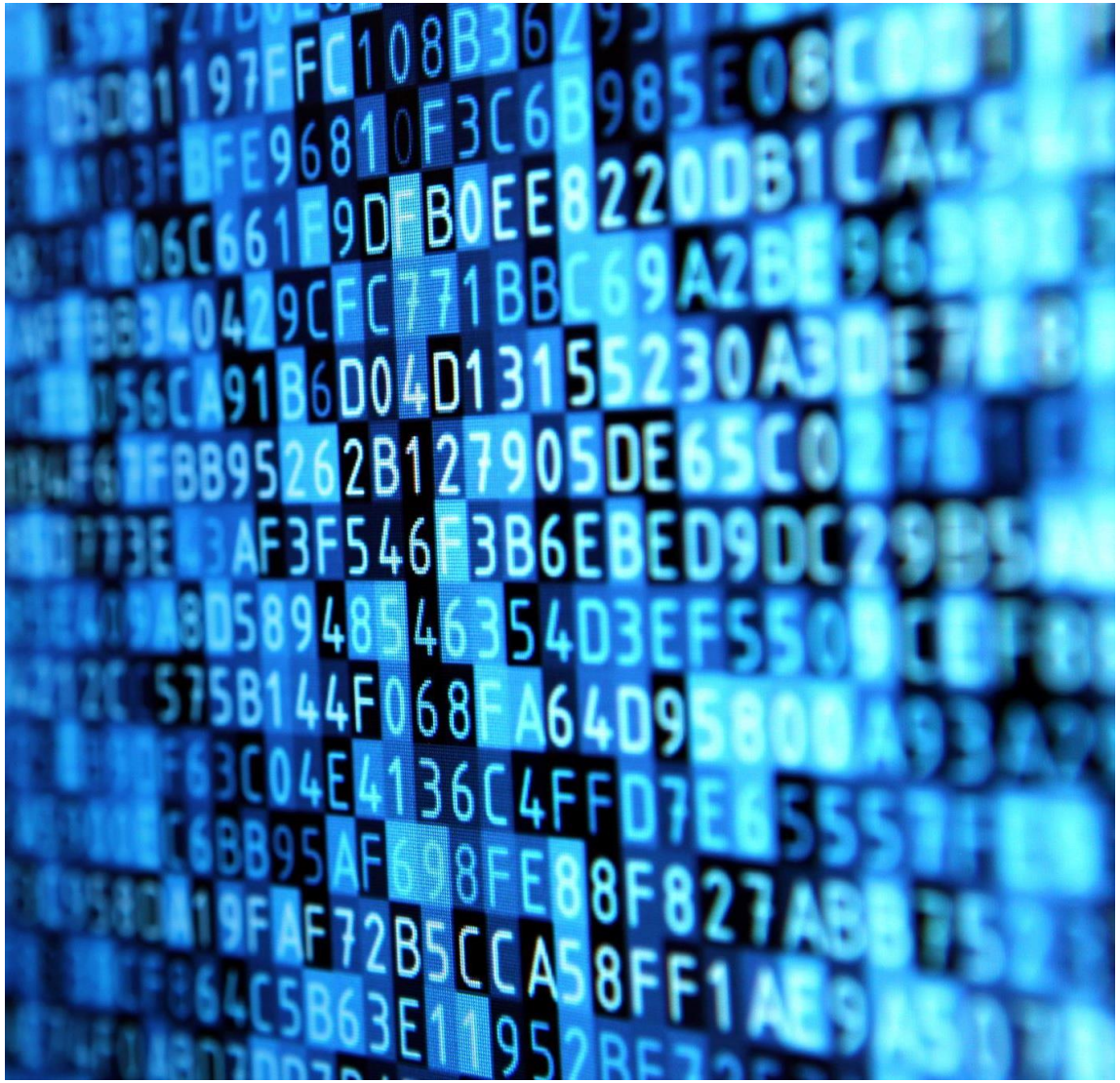
- Importance of handling missing values.
- Use dropna() function to drop NA values

Correlation

- What is correlation and its importance in data analysis.

One-Hot Encoding

- What is categorical variables and their importance in machine learning.
- Why one-hot encoding is necessary to handle categorical variables.



Data Visualization

- Use hist() function to visualize the distribution of each feature in the dataset.
- Plot each histogram as a separate subplot.
- Use of describe() function to find the mean, median, and standard deviations for each feature.

Data Splitting

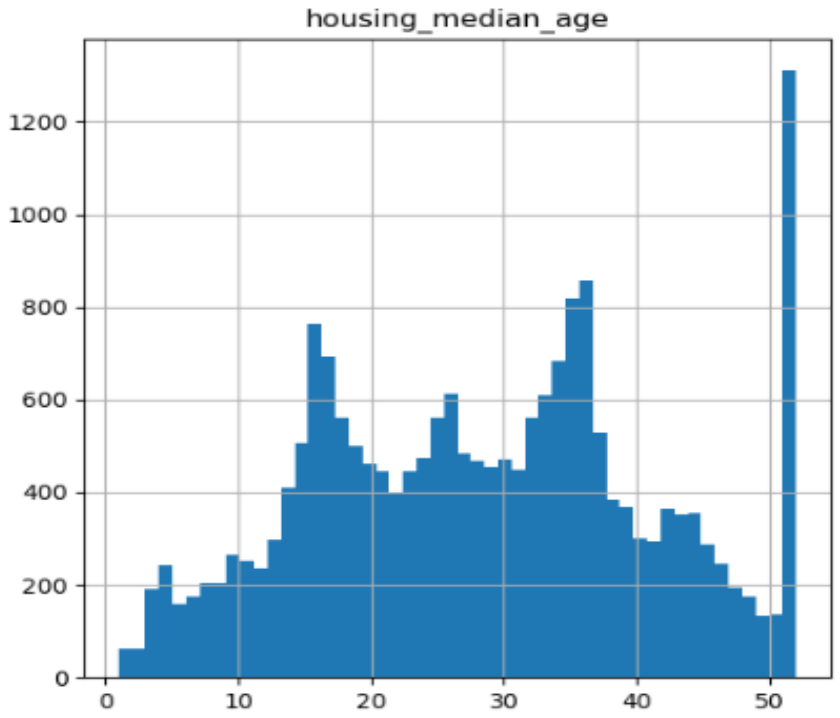
- For linear Regression
- We start by splitting the data in different attributes.
- According to question we split the data in 70% training and 30% test data.
- X_test, y_test, X_train, y_train.

Data Scaling

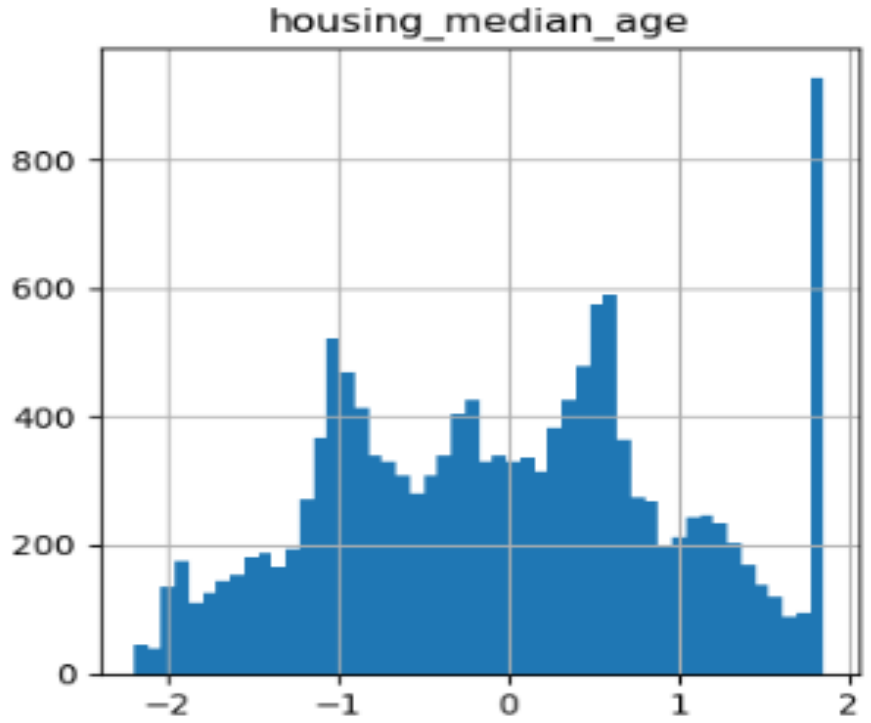
- Centering and scaling the data
- The fit method :-
  - Calculate two values :-
    - Mean and Standard deviation.
- The Transform method :-
  - This is the method that is used for scaling the data.

$$x_{scaled} = \frac{x - mean}{std.dev.}$$

Scaling

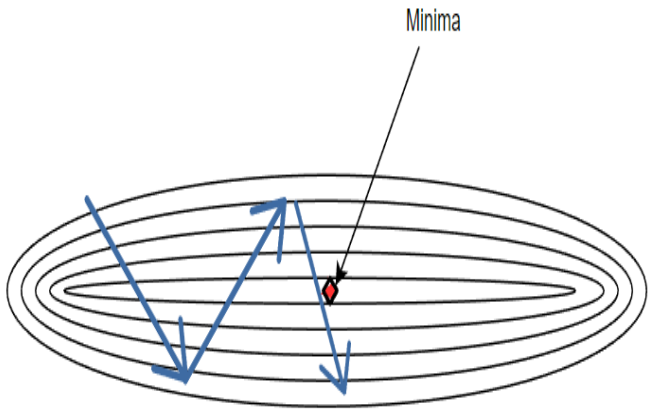


Unscaled Data



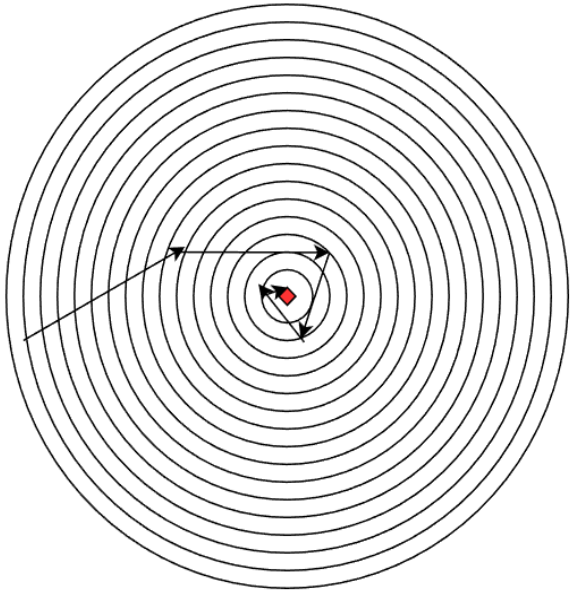
Scaled Data

Why Scale ?



Overshooting the minima

If one dimension is much greater than the other, we would have the problem of over-shooting the minima, and our model will fail to converge.



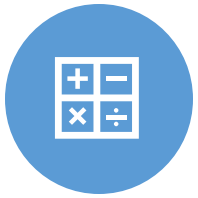
In comparison, gradient descent converges better when the dimensions have been scaled to have a standard deviation of 1.

Modelling

Model used Linear Regression.

Using the split data.

Training the Model using X_train and y_train.

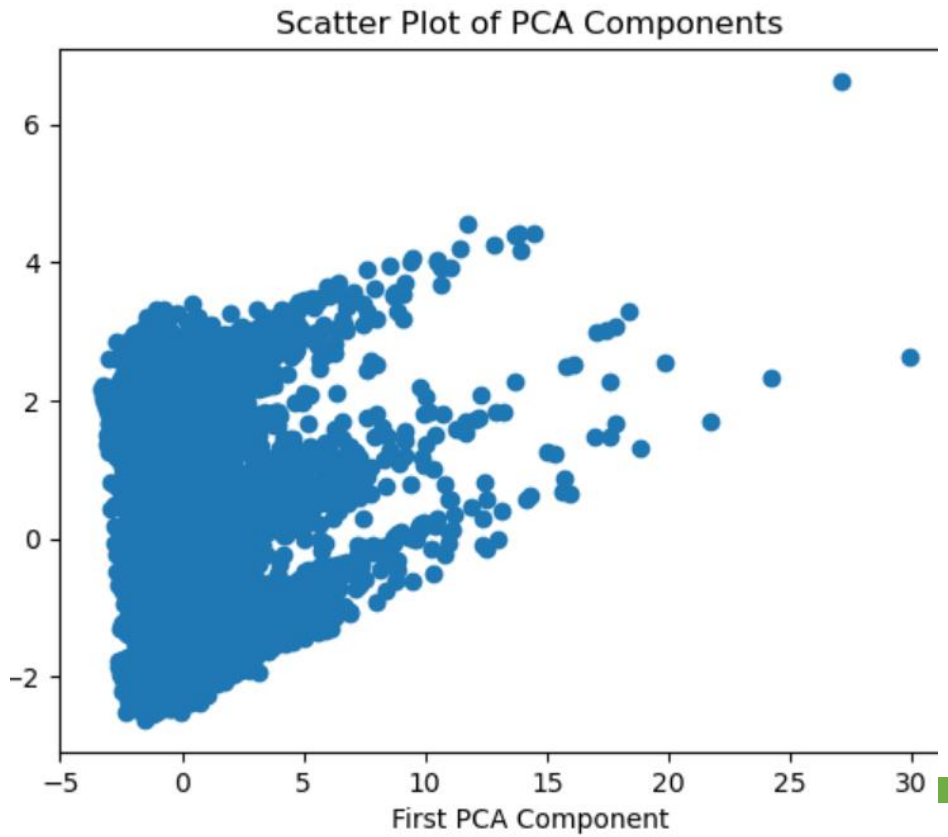


Calculating Predictions using the test cases X_test.

Finally, Inversely transforming the predictions.

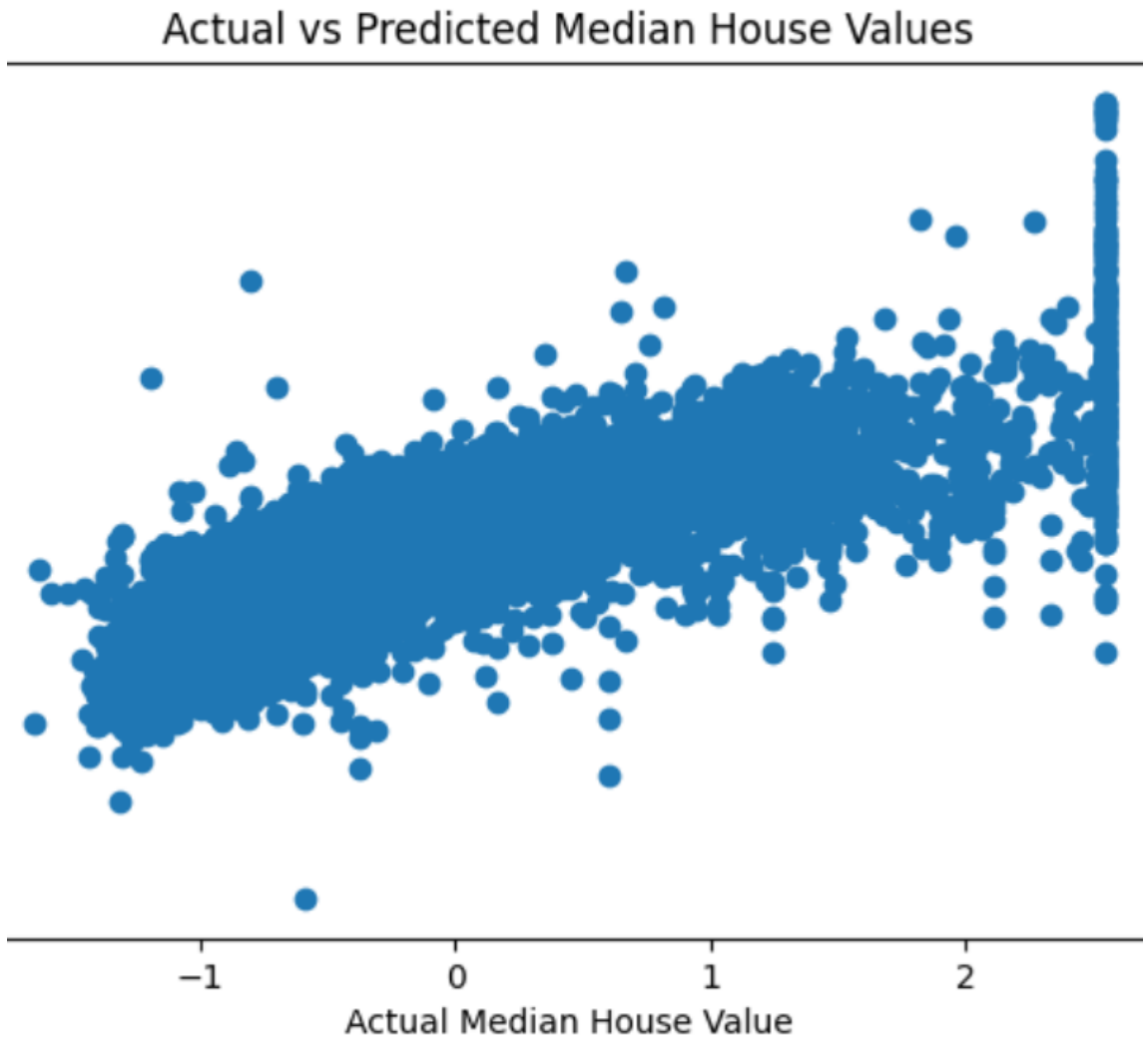
Principle Component Analysis

- PCA analysis done with 2 components.
- Strength of PCA components
  - Strength of first components is : 236.62.
  - Strength of Second components is: 191.60.
- Explained Variance Ratio:
  - For first component is: 0.3011.
  - For second component is: 0.1974.



Analysis

- We predicted the median of the house prices our of test data.
- We compared the predicted results with the actual results, and we plotted a graph for that, and the graph is shown below.
- The graph has actual median values on the X-axis and predicted median house values on the Y-axis.



Evaluation

We calculated the mean absolute percentage error.

The mean absolute percentage error is the measure of the average percentage difference between actual and predicted values.

It gives an idea of the magnitude of percentage error between the actual housing prices and the predicted housing prices.

The lower mean absolute percentage value indicates that the accuracy of our model is good.

Evaluation Metrices

```
Evaluation Metrics:
--------------------
MAPE: 0.29%
RMSE: 67922.85
R2 Score: 0.65
```

Thank You