**MACHINE LEARNING ASSIGNMENT-1**

**Group Members:**

**Aathira Satheesh Nair - AXN210057**

**Hari Prasad Tokala - HXT210020**

**No of free days used: 0**

**1. Project Selection**

**California Housing Dataset** [**https://www.kaggle.com/camnugent/california-housing-prices**](https://www.kaggle.com/camnugent/california-housing-prices)

**2. Features in the Dataset**

The feature attributes in the data set are as follows:

Independent features/attributes  
1. **longitude**: A measure of how far west a house is; a higher value is farther west

2. **latitude:** A measure of how far north a house is; a higher value is farther north

3. **housingMedianAge:** Median age of a house within a block; a lower number is a newer building

4. **totalRooms:** Total number of rooms within a block

5. **totalBedrooms:** Total number of bedrooms within a block

6. **population:** Total number of people residing within a block

7. **households:** Total number of households, a group of people residing within a home unit, for a block

8. **medianIncome:** Median income for households within a block of houses (measured in tens of thousands of US Dollars)

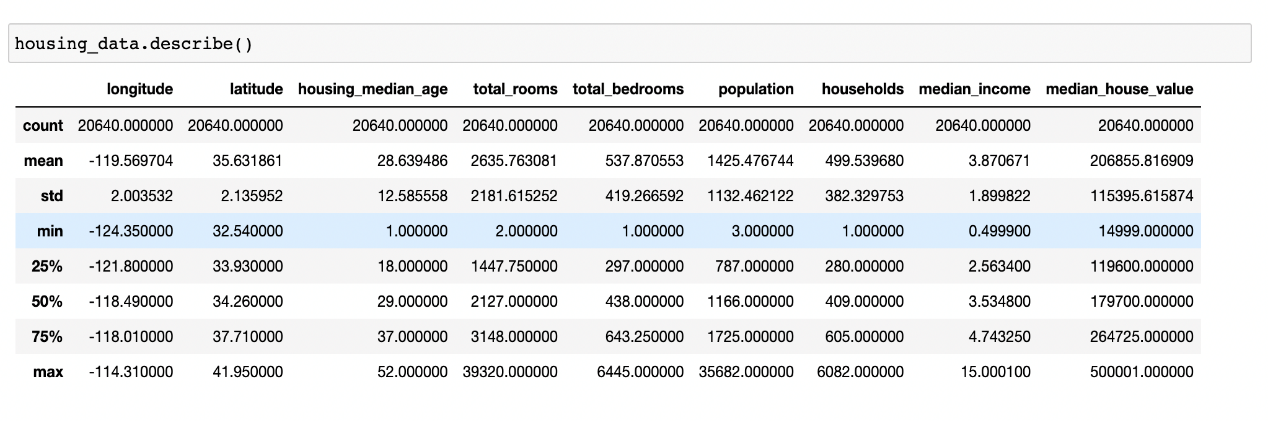
9. **oceanProximity:** Location of the house w.r.t ocean/sea

Dependent Feature

**medianHouseValue:** Median house value for households within a block (measured in US Dollars)

**2. Analyzing the Dataset**

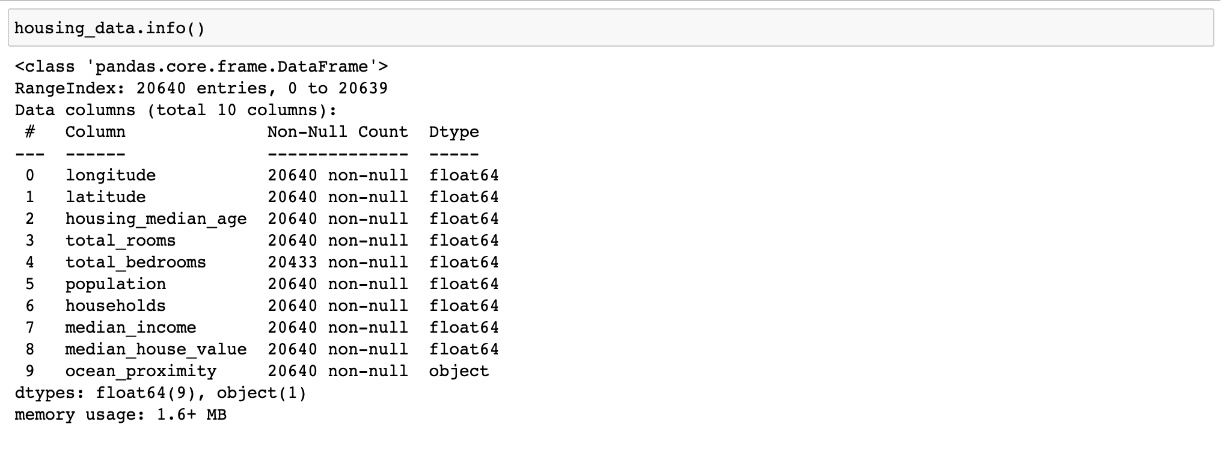
**Summary Statistics of Housing Data**



**Observations:**

We can see that for attributes like total\_rooms,population,household,total\_bedrooms, the range of values vary a lot.

### **Get DataTypes of Attributes**



**Observations:**

We Observe that except Ocean\_proximity , every other attribute has float values. Ocean proximity contains discrete values

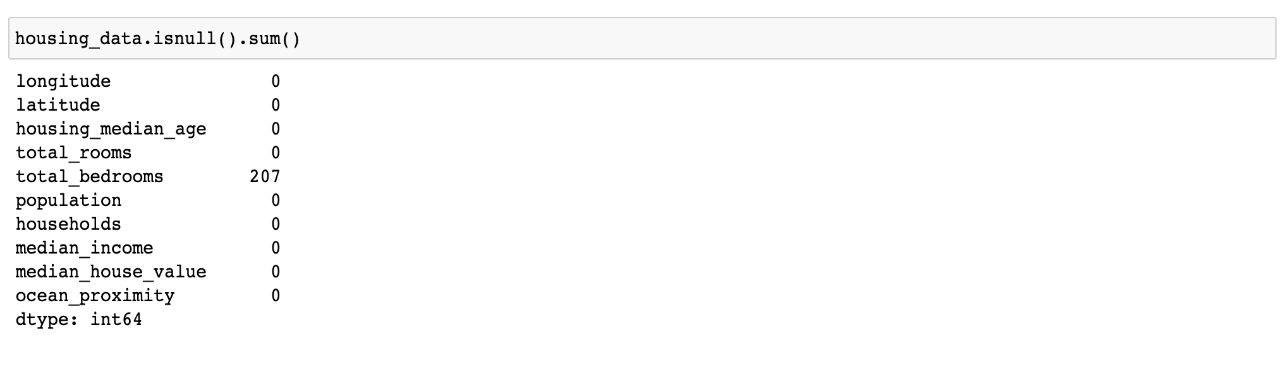
### **Find the count of each discrete value:**



**Observations:**

We Observe that except Ocean\_proximity contains the following Island, Nearbay, Near Ocean, Inland, <1H Ocean.

**Check for Null values if any:**



**Observations:**

We Observe that total Bedrooms is the only feature attribute that has 207 null values.

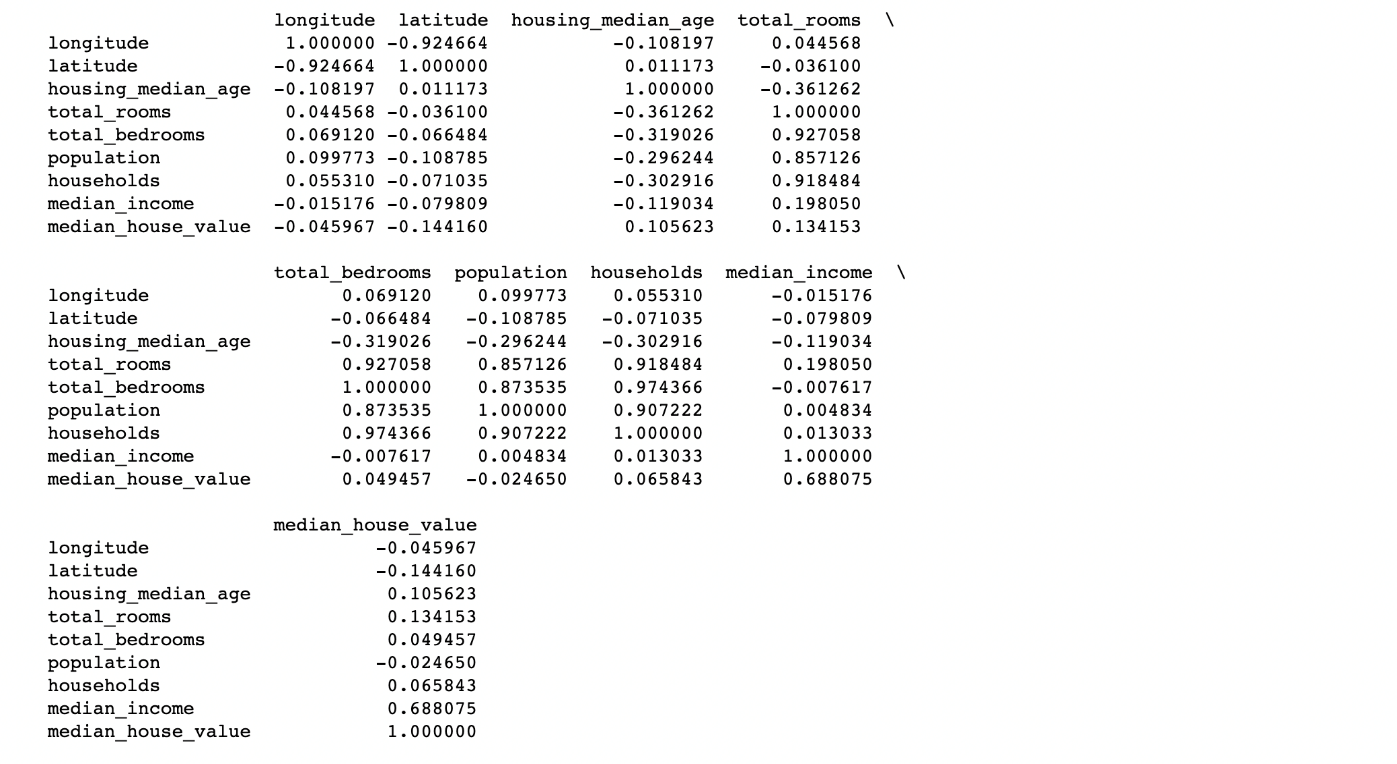
**Eliminate Null values:**

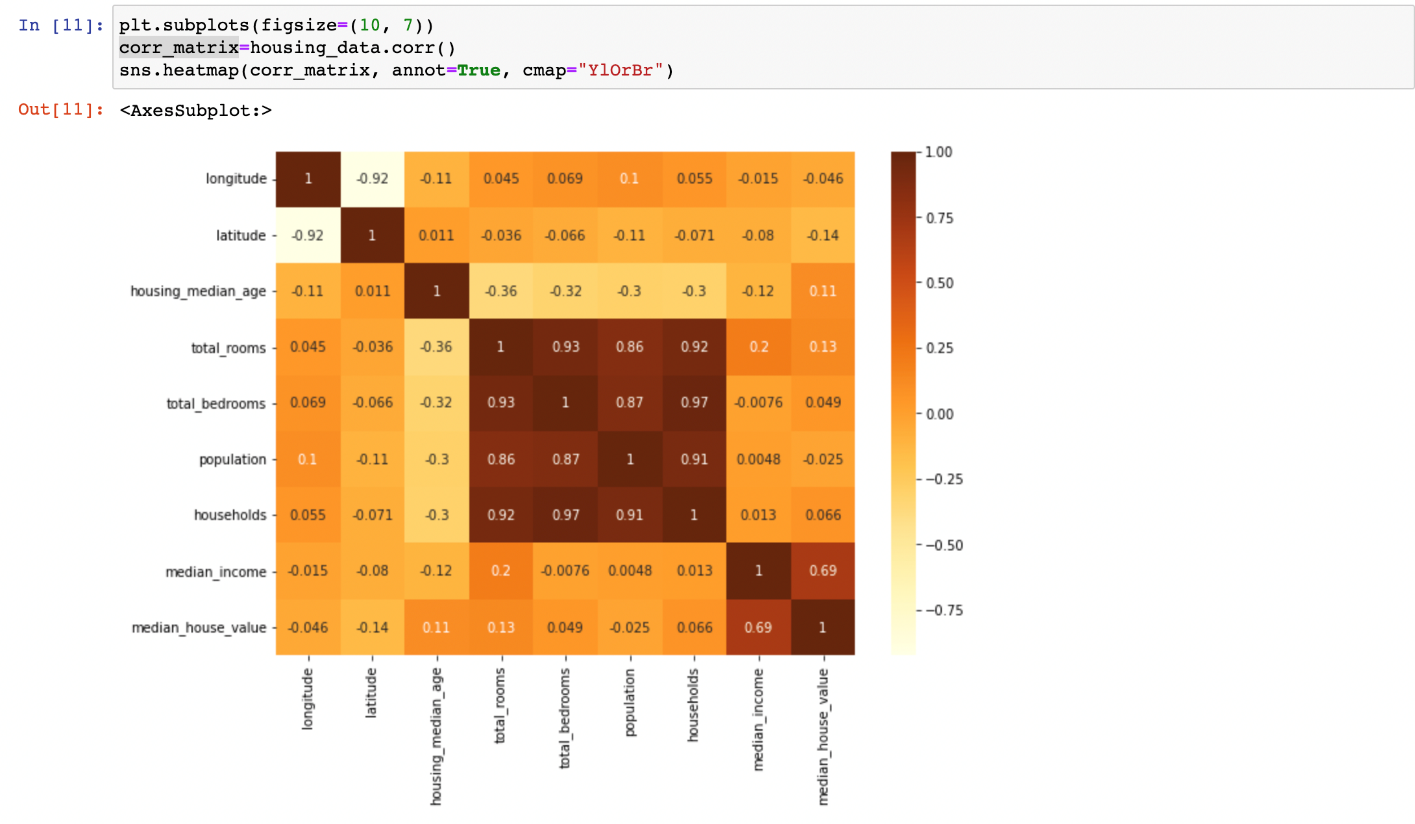
We can eliminate in many ways, here we chose to eliminate null values present by replacing them with median value of that attribute.



Here we can see that we have eliminated all the null values present in total bedrooms and replaced them with the median value of that attribute.

**Correlation Matrix and Heatmap**

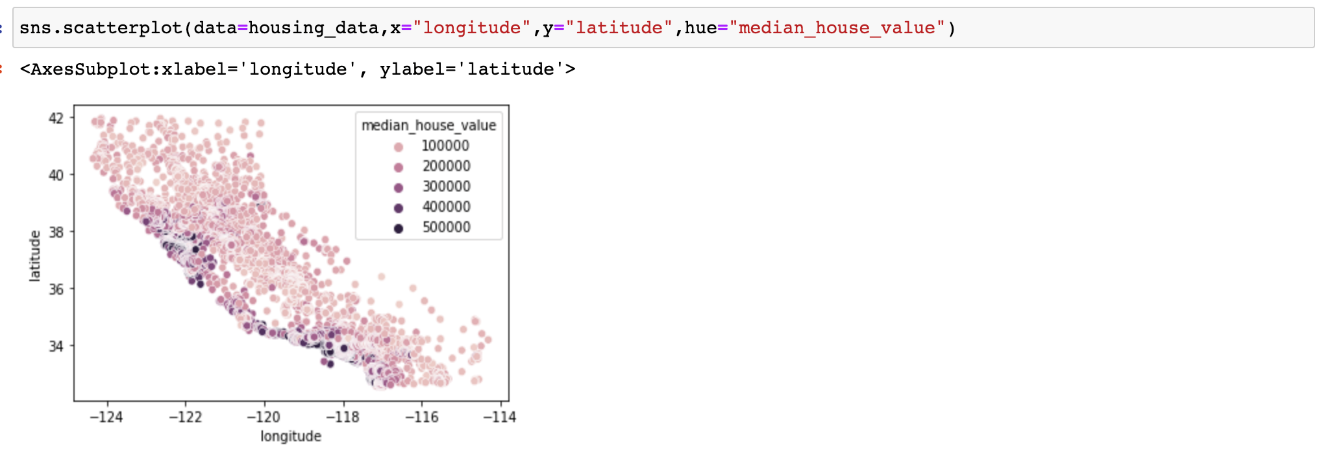




**Observations:**

We can see that median\_income is the attribute with highest correlation with median house value. The correlated attributes are longitude and longitude (negatively correlated), total bedrooms and total rooms, total bedrooms and households, population, and households and so on.

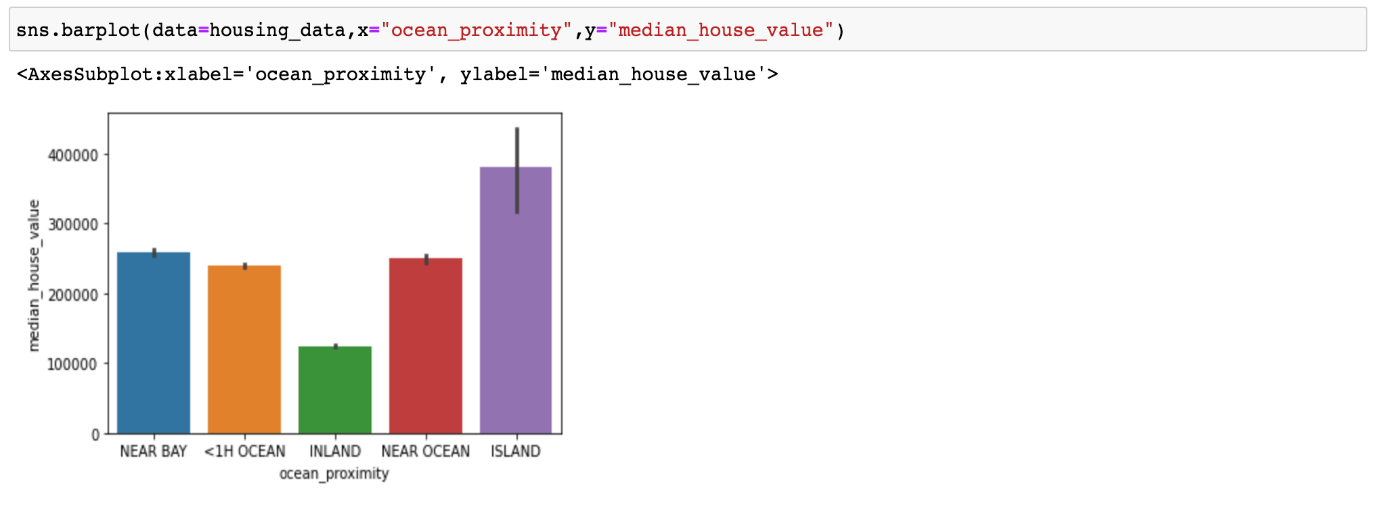
**Scatterplot to analyze the relation between latitude,** **longitude, and median house value:**



**Observations:**

From the above scatterplot we can see that, lesser the longitude and latitude value, Higher the median house value. This implies that house value depends on the location of that house.

**Analyzing Ocean proximity with median house value:**

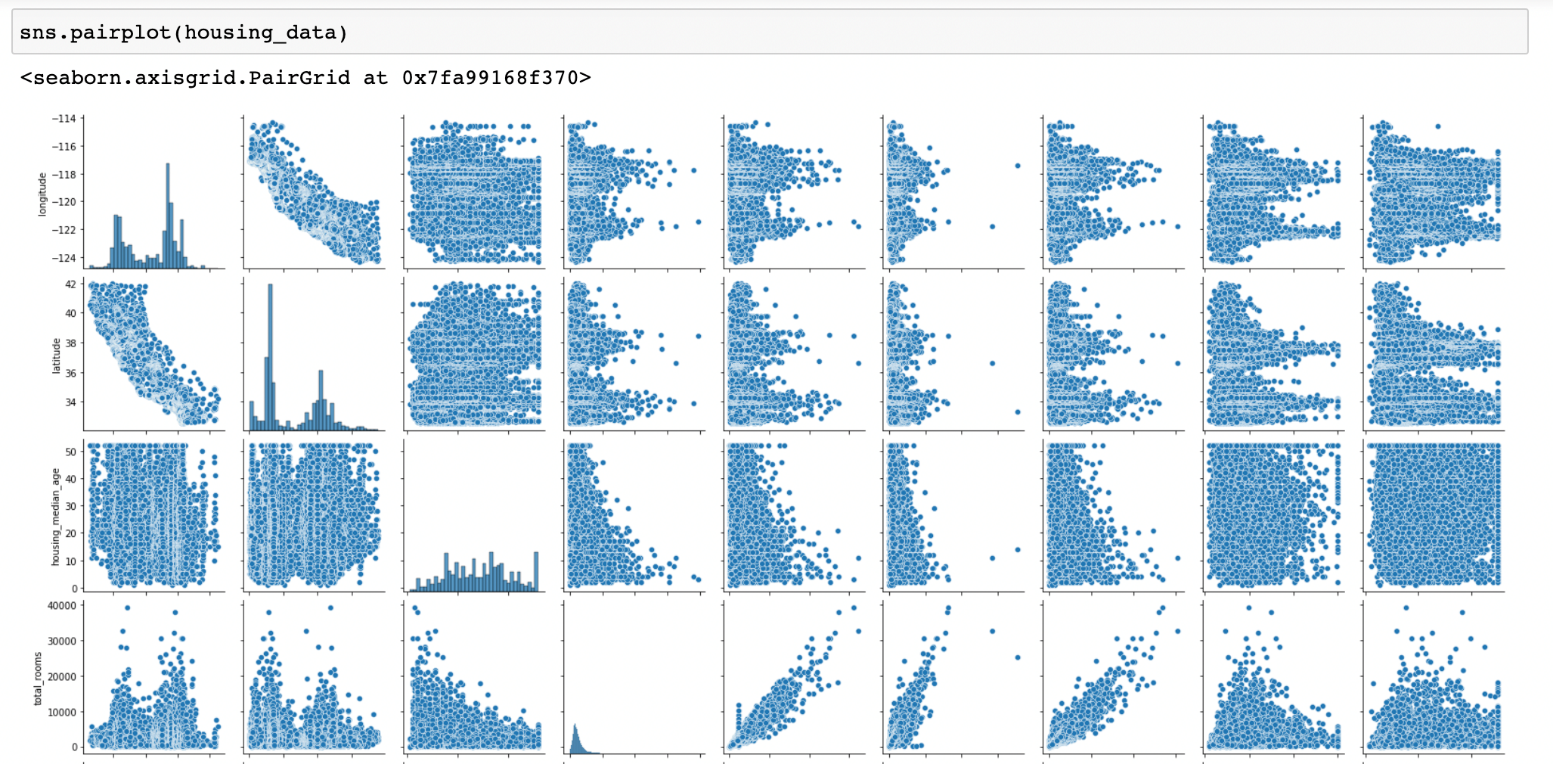


**Observations:**

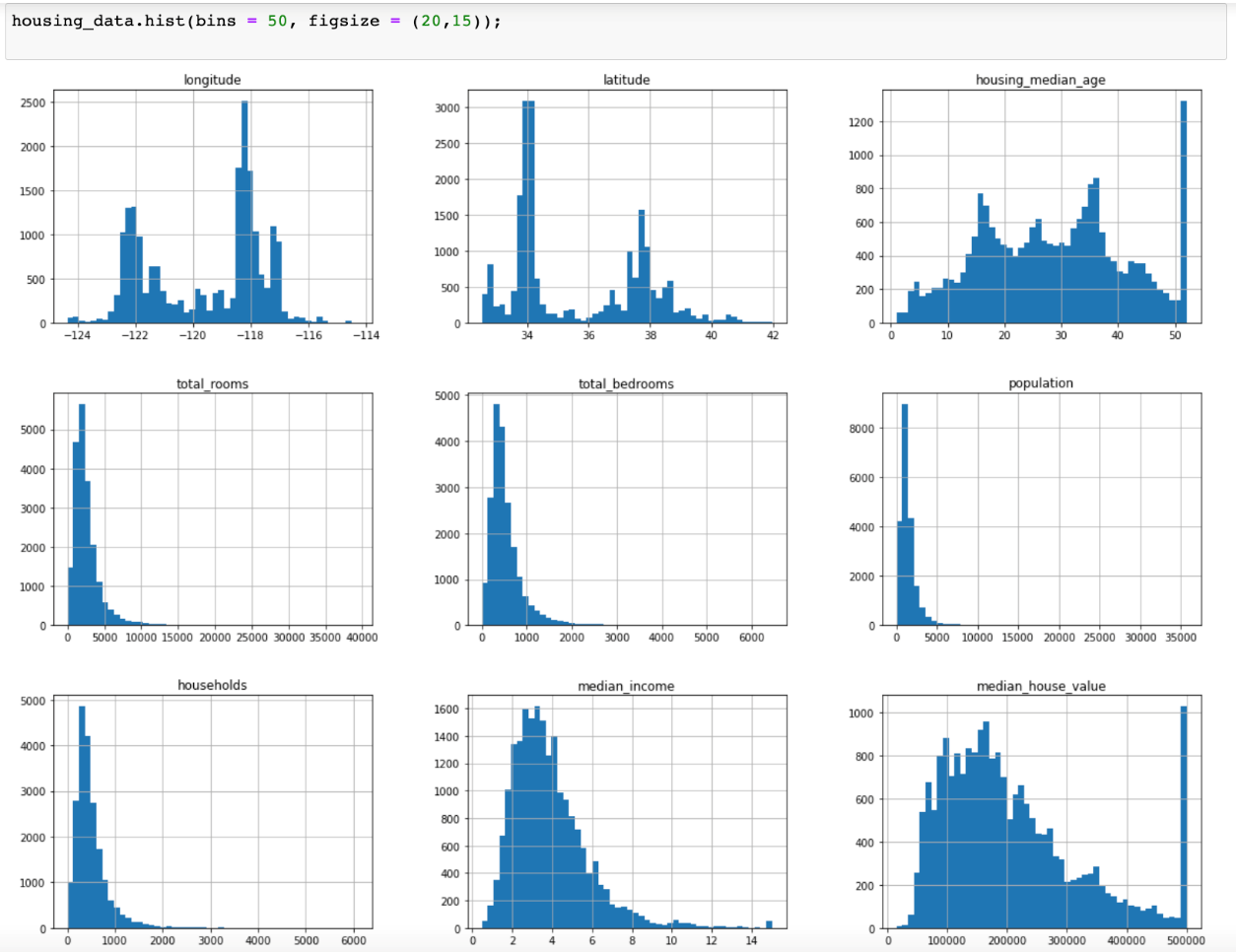
Here we can see that Houses in islands have a higher median house value The houses that are far away from ocean(inland) has lesser median house value, i.e. its cheaper at the inland.

**Pairplot:**

We can see from the pairplot that median\_house\_value has hardly any correlation with all the features except median\_income .



**Plotting Histogram:**



**Observations:**

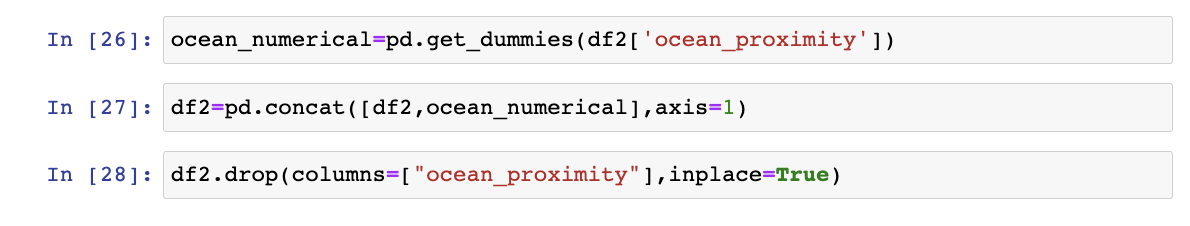
Here we can see that from the histograms that the formed bell shaped curve is a little right skewed, it happens because of the outliers.

**Eliminate Outliers:**

1. We have plotted Box plot to get outliers
2. Next, we plotted scatterplot to observe outliers present in households, median income, total rooms, total bedrooms, population.
3. Based on the scatter plot we observed , we remove all those values above which it appears to be an outlier.

**Change Categorical values to numerical**

As we need to make all categorical variables to numerical, The only categorical variable we have is ocean\_proximity so convert it into numerical.



As we can see that the data has a highly varying magnitudes, we need to standardize the independent features present in the data in a fixed range. Hence we also performed standard scaling of the data.

**3)Two Promising Features**

As we can see from the heatmap and correlation matrix , Other than median\_income , no other feature has high correlation with target variable.

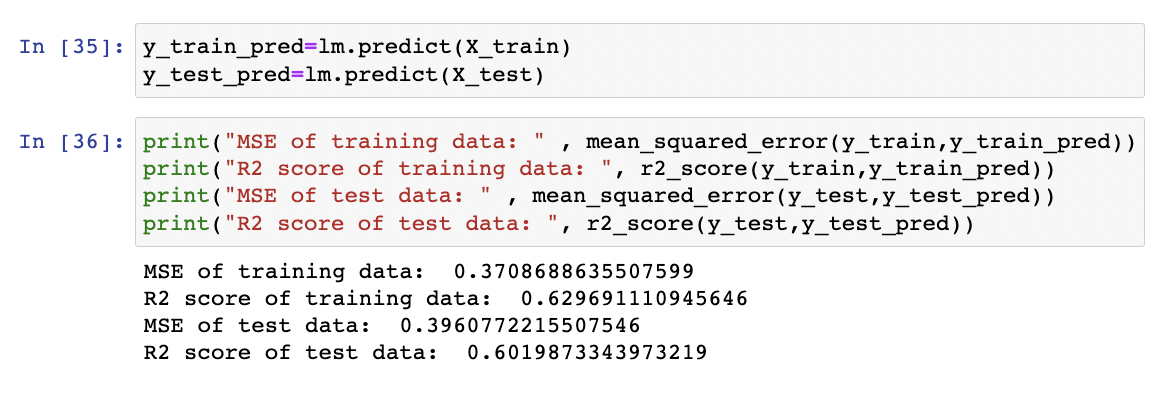
The only promising feature is median\_income.

**Linear Regression:**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Here we are splitting the data as 80% training data and 20% test data and performed Ordinary least squares linear regression.

**Observations:**



**SGD Regressor:**

SGD stands for Stochastic Gradient Descent: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing learning rate.

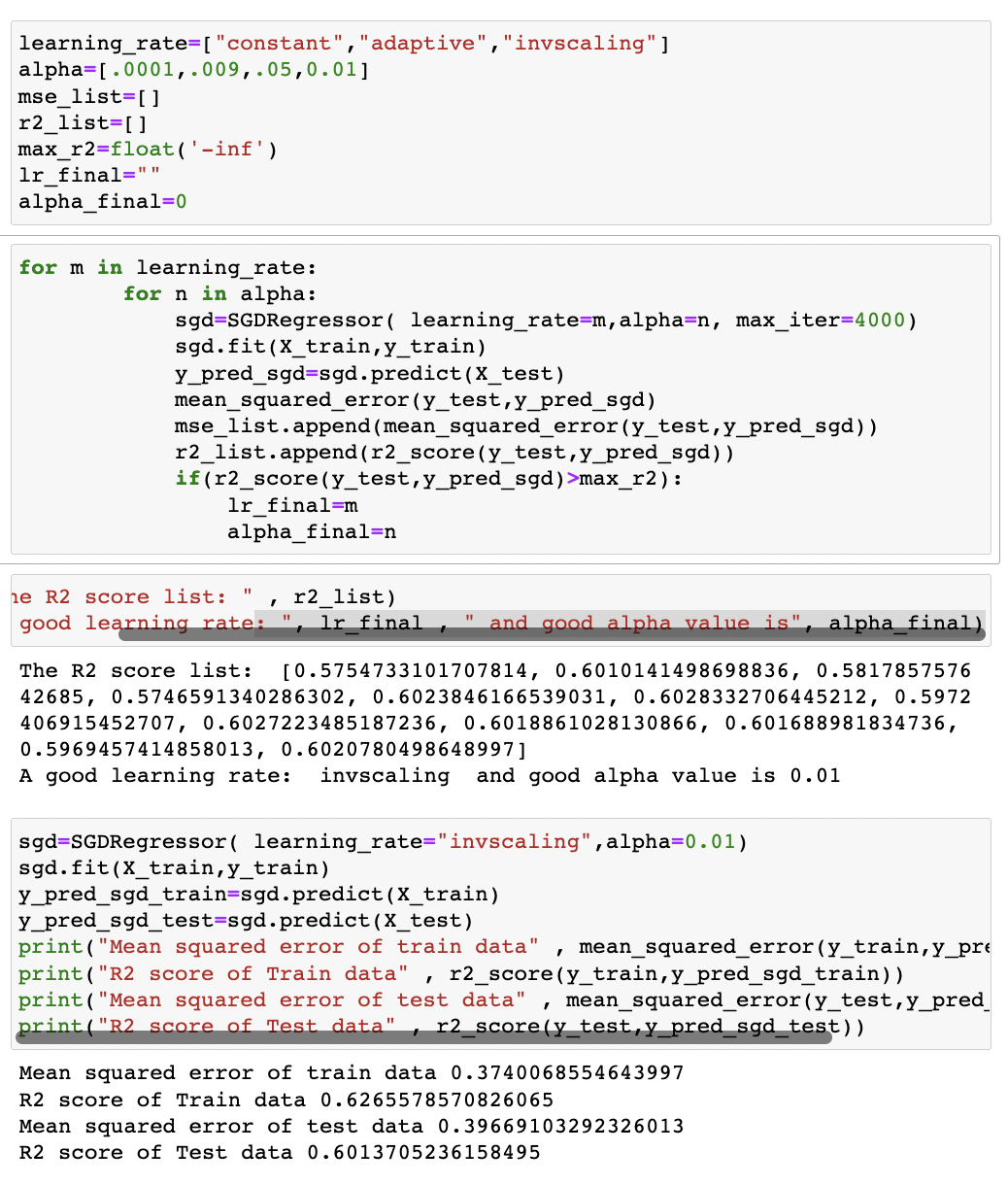
SGD Regressor has many hyperparameters that can be tuned like alpha, learning rate etc.

We made a list of these and tune it and create the model and see the MSE and R2\_score and finally select the hyperparameters with least MSE and high r2\_score.

In our case the better learning\_rate is invscaling

And better alpha value is 0.01.

**Observations:**



**Comapring OLS Linear Regression and SGD Regression:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Evaluation Method** | **OLS Linear Regression** | | **SGD Regression** | |
| **Training** | **Test** | **Training** | **Test** |
| **Mean Squared Error** | 0.3708688635507599 | 0.3960772215507546 | 0.3740068554643997 | 0.39669103292326013 |
| **R2 Score** | 0.629691110945646 | 0.6019873343973219 | 0.6265578570826065 | 0.6013705236158495 |

As we can see, for our dataset, OLS and SGD give equivalent results for both training and testing data.

Both methods try to minimize the residual errors.

It's just that SGD performs faster than OLS, but they arrive at the same answer.

One possible reason for this less r2\_score could be that The data we used had features with high correlation

For example :- longitude and latitude is highly negatively correlated

Similarly households, population, total rooms and total bedrooms are highly correlated as well.

Removing the correlated features can possibly give better results.

The data isn’t overfitting as the performance on training and testing set has slight difference only.