HOUSE PRICE PREDICTION

Using Linear Regression

Problem statement :

House price prediction project helps to determine the selling price of a particular region and can help people get correct price for there land,in this project we will use
Data from california to create a machine learning model that will predict the prices of the house in the state.The data set includes population , median income and median house prices for each block in california.

We use block groups because they are the smallest geographic unit which typically depicts a population of 600 to 3000 people. This model will learn the data and we will predict median house prices in any given neighborhood provided all other mattresses .

Literature Review:

The trends in housing prices indicate the current economic situation and it is also a concern for the buyers and sellers. There are many factors that have an impact on house prices, such as the number of bedrooms. House price also depends upon its location as well. A house with more facilities like highways, schools, malls, employment opportunities, would have a greater price as compared to a house with no such accessibility in its surroundings. Predicting house prices manually is a difficult task and generally not very accurate. This system's aim was to make a model that can give us a good house price prediction based on other variables. They used the Linear Regression for Ames dataset and hence it gave good accuracy. The house price prediction project had two modules namely, Admin and the User. Admin can add location and view the location. Admin had the authority to add density on the basis of per unit area. Users can view the location and see the predicted housing price for that particular location.

Model Design :

- Libraries used pandas,matplotlib,sklearn,numpy.
- The data is imported on a jupyter base notebook using pandas. Each row represents a district and there are 10 attributes in the dataset. The info method is used to get quick metadata especially the no. rows , type of attributes and non zero values.

Table Visualize Statistics

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population
0	-122.23	37.88	41.0	880.0	129.0	322.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0

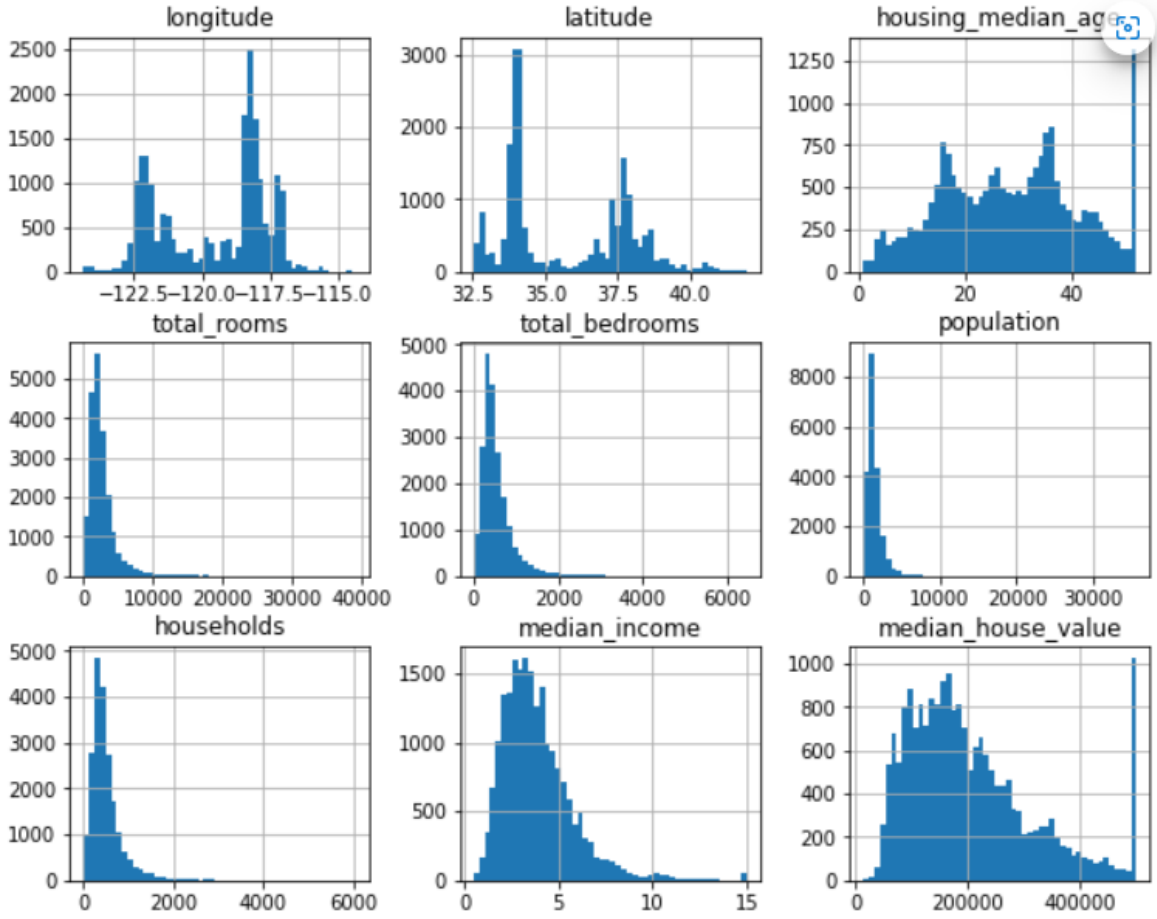
5 rows × 10 columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 16512 entries, 12655 to 19773
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           16512 non-null  float64
 1   latitude            16512 non-null  float64
 2   housing_median_age  16512 non-null  float64
 3   total_rooms         16512 non-null  float64
 4   total_bedrooms      16512 non-null  float64
 5   population          16512 non-null  float64
 6   households          16512 non-null  float64
 7   median_income       16512 non-null  float64
 8   ocean_proximity     16512 non-null  object 
dtypes: float64(8), object(1)
memory usage: 1.3+ MB
```

There are 20,650 instances. The total bedroom attribute has 20433 non zero values. which means there are 207 districts which have uncertain data.

All attributes are numeric except ocean proximity. Its type is object , so it can contain any type of python object.

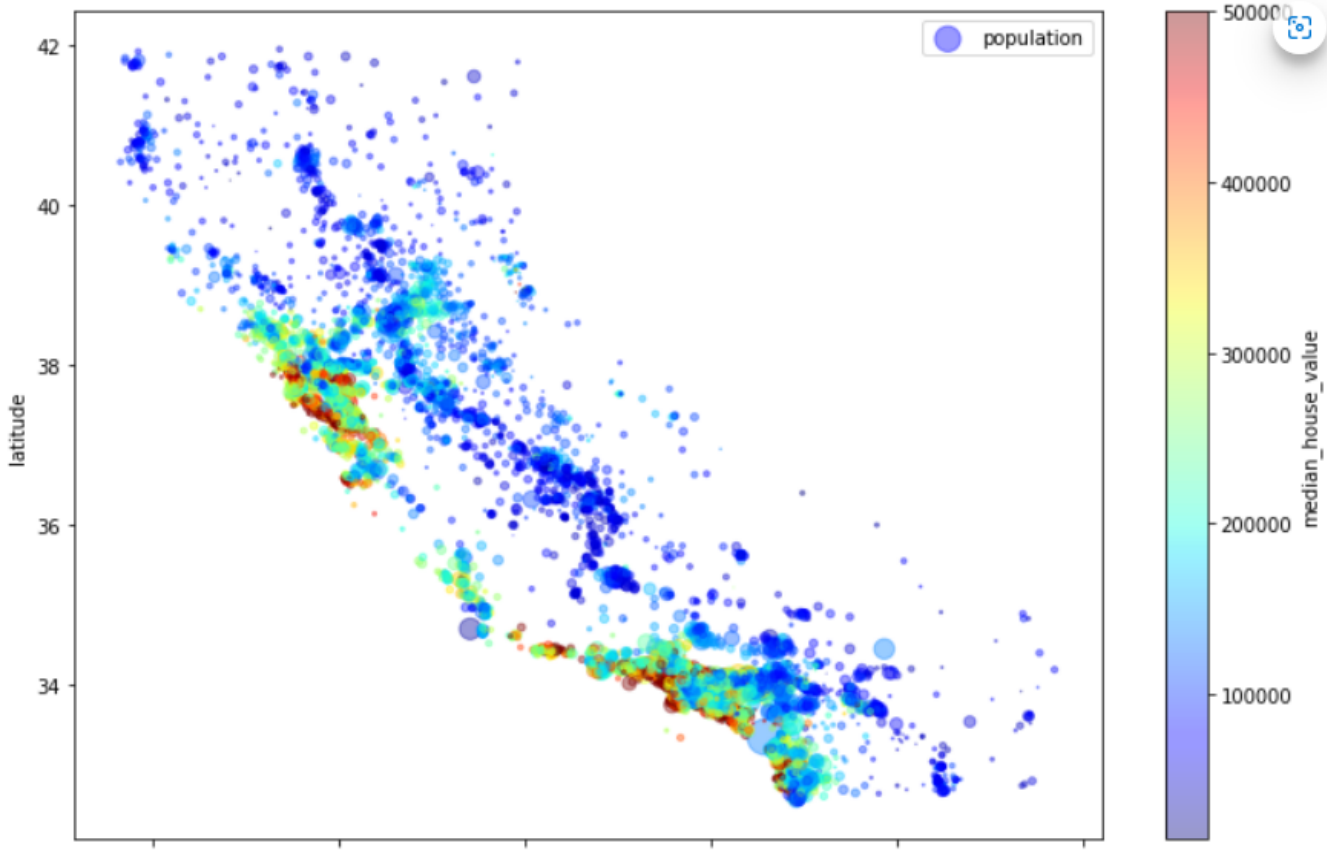
To elaborate data be use matplotlib to plot its histogram with each numerical value.



The next part of the analysis is to split data into training. For test data , we take 20% off the data set and set them aside. It is very important to have sufficient no. of instances in the data to avoid biased analysis. The next step is to perform stratified sampling to get homogenous data and increase the precision of the data.

```
3    0.350533
2    0.318798
4    0.176357
5    0.114341
1    0.039971
Name: income_cat, dtype: float64
```

Before creating a machine learning model , We visualized data in terms of longitude and lattitude.



Considering the size of the dataset We calculate the standard correlation coefficient between age peer of attribute. Corr()

```
median_house_value    1.000000
median_income         0.687151
total_rooms           0.135140
housing_median_age    0.114146
households            0.064590
total_bedrooms        0.047781
population           -0.026882
longitude            -0.047466
latitude             -0.142673
Name: median_house_value, dtype: float64
```

Correlation ranges between -1 and 1. When it is close to 1 , it means there is a positive correlation when it's close to -1 , it means that there is a negative correlation.

Now we add three new columns in dataset:rooms per household, bedrooms per room and population per household

```
median_house_value          1.000000
median_income               0.687160
rooms_per_household         0.146285
total_rooms                 0.135097
housing_median_age          0.114110
households                  0.064506
total_bedrooms              0.047689
population_per_household   -0.021985
population                 -0.026920
longitude                  -0.047432
latitude                   -0.142724
bedrooms_per_room          -0.259984
Name: median_house_value, dtype: float64
```

Linear Regression for House Price Prediction with Python :

Now I will use the linear regression algorithm for the task of house price prediction with Python:

In this phase we perform various transformations in data for data preparation so that the data can be prepared with help of Scikit-Learn. We give the pipeline class for sequences of transformation .
In the end a regression algorithm is used to predict the prices.

```
Predictions:  [ 85657.90192014 305492.60737488 152056.46122456 186095.70946094
 244550.67966089]
```

Conclusion:

1. System will provide the approx rate charges of land according to the size and shape and geological factors.

2. People who migrate from their hometown in different cities to work or learn would get an approx knowledge of pricing for land in the particular area which will reduce the tendency of fraud, paying off artificially inflated rates of land.

3. Beneficial from Governmental aspects and as well as corporate entities, as whenever any new infrastructure planning is initiated it requires land, so organization has to compensate the owner of land which could be done by analyzing the data and could pay an average cost to them which leads to time saving as well as earlier finishing of project.

4. This study will support the policy makers to relook the movement of the identified factors to have control on rise in the land price and stabilize it. The current urbanization from rural areas is growing at a higher rate which will lead to increase of land requirement which will result in higher pricing of land.

Project Work File:

https://github.com/harshchauhan17/House-Price-Prediction-using-linear-Regression.git

==========