Project Name: California Housing Price Prediction

Objective: The purpose of the project is to predict median house values in Californian districts, given many features from these districts.

The project also aims at building a model of housing prices in California using the California census data. The data has metrics such as the population, median income, median housing price, and so on for each block group in California. This model should learn from the data and be able to predict the median housing price in any district, given all the other metrics.

Districts or block groups are the smallest geographical units for which the US Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people). There are 20,640 districts in the project dataset.

Solution: To solve this problems we performed following steps

step1: Data preprocessing

First we have to check if there is any column with null value in it; if So we need to replace that value with mean value. We load data into pandas dataframe as follows

```
housing_df=pd.read_csv("housing.csv")
```

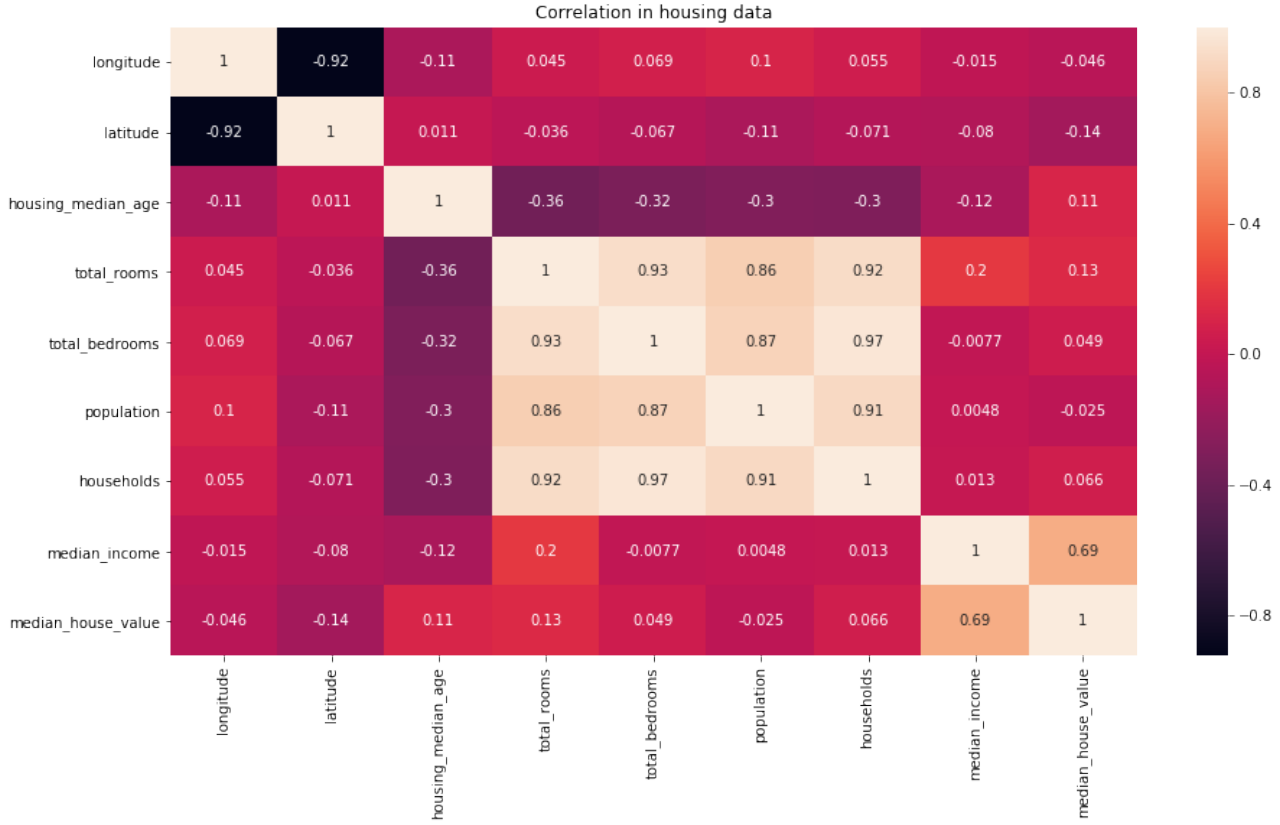
in order to check if there is any column with null value, we use following statement

```
housing_df.isnull().any()
```

we found `'total_bedrooms'` column has null value . We replace null value with mean of that column.

Step2: Exploratory data analysis

In this step we try to find relation among different features in our data. We also plot heatmap for correlation in housing data as follows

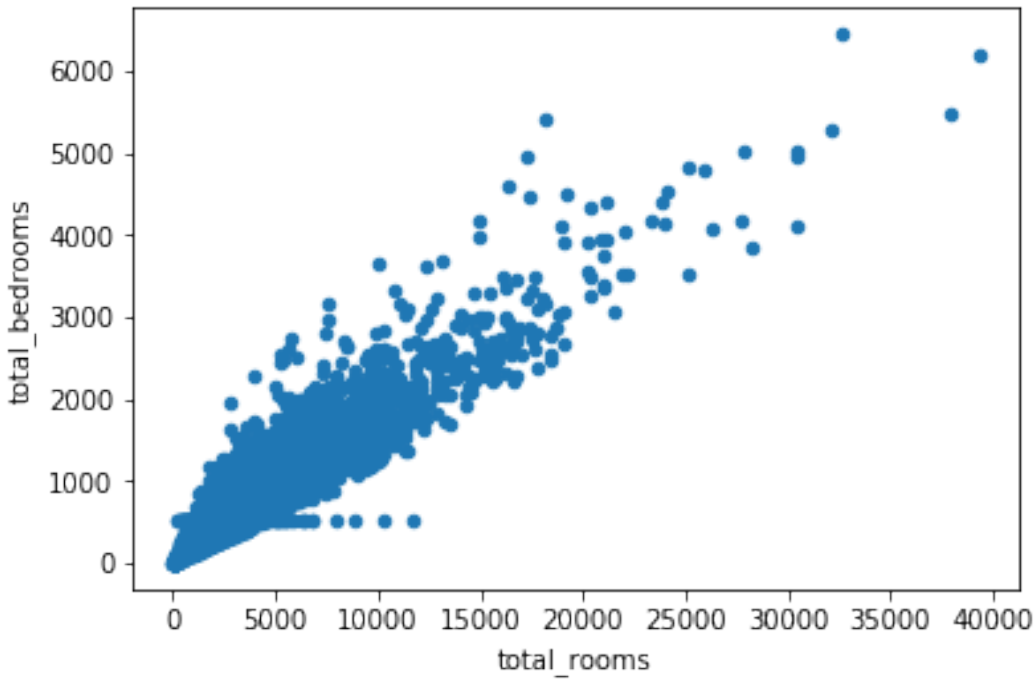


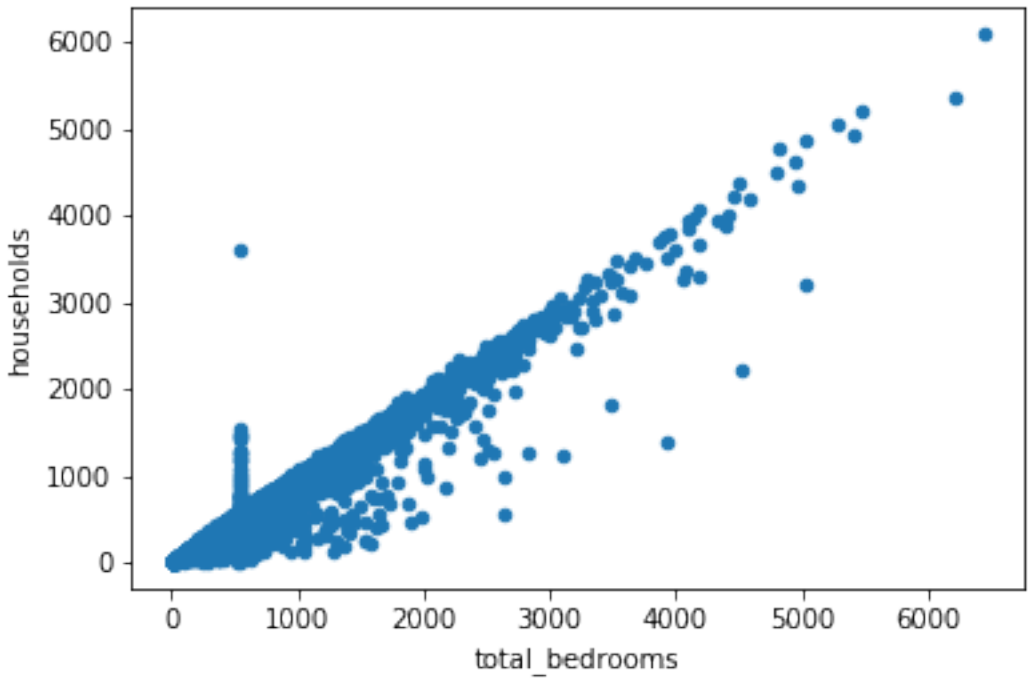
from above diagram we found that there is high correlation between

- total_bedrooms and households
- total_bedrooms and total_rooms
- households and total_rooms
- population and households

At this point we can drop households and total_rooms columns while creating model; since other columns are present which can convey similar information

scatterplot for total_bedrooms vs households and total bedrooms vs total rooms are shown as follows


total_bedrooms
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
40000
total_rooms


households
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
total_bedrooms

step 3: Feature Engineering

Instead of having longitude and latitude as separate attribute we will put an attribute distance_from_california
california is at
Latitude 36.778259
Longitude -119.41793
we use the 'haversine' formula to calculate the great-circle distance between two points – that is, the shortest distance over the earth's surface – giving an 'as-the-crow-flies' distance between the points (ignoring any hills they fly over, of course!).

```
Haversine formula:
  a = sin²(Δφ/2) + cos φ1 · cos φ2 · sin²(Δλ/2)
  c = 2 · atan2( √a, √(1−a) )
  d = R · c
where    φ is latitude, λ is longitude, R is earth's radius (mean radius =
6,371km);
note that angles need to be in radians to pass to trig functions!
```

Step 4: Standerdising data and performing machine learning.

In final step we standerdise our data and divided it into train and test data. Different regression techniques, mean squared error and score of models are as given in following table.

Regression Technique	Mean squared error	Score of model
Linear regression	72.33141914038299	0.6512403943232119
Decision tree regression	67.19112486979536	0.6990488345070018
Random forest regression	59.624760503580426	0.7630124664153124

Conclusion: Thus we have developed models for predicting house price. And we found that random forest regression gives minimum mean squared error and maximum score.