**Assignment 8**

**Question and Dataset detail**

The dataset used in this assignment is California housing dataset which describes the occupancy of housing and population and location of houses in California. This data is taken from sklearn datasets package in python. The fields in this dataset are MedInc, HouseAge, AveRooms, AveBedrms, Population, AveOccup, Latitude and Longitude. Each column represents the house occupancy, population, and type of houses in California.

House Age column describes the age of the house in California.

Population column describes the number of people living in the given houses in California.

AveRooms columns describes the average of rooms of the houses in california.

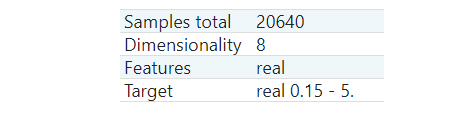
AveBedrms columns describes the average bedrooms of the houses in california.

AveOccup column represents the average occupancy of houses in california.

Latitude and Longitude columns shows the location of the houses in california.

This dataset has total of 20640 samples and 8 dimensions with real features.

For this assignment we are predicting the average occupancy of the houses in california.



**EDA & Data cleaning**

EDA is explanatory data analysis used to find the features which influences the target and clean the features which do not have effect on the target. We have created different forms of EDA like heatmaps and histograms for the cleaned data.

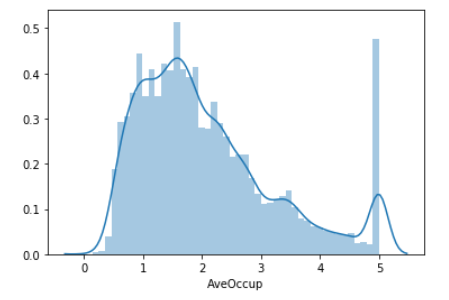


Fig. AveOccup histogram

Figure shows the histogram of AveOccup of the dataset which has normal distribution with some outliers.

**Correlation Matrix**

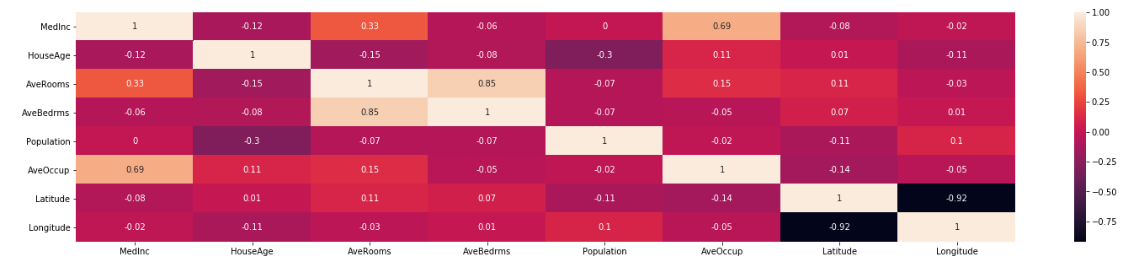


Fig. Correlation Matrix

The above figure shows the correlation matrix of the california housing dataset showing the relations between different features and ‘AveOccup’ column.

The ‘AveOccup’ influences the MedInc by 0.69, HouseAge by 0.11, AveRooms by 0.15, AveBedrms by -0.05, Population by -0.02, AveOccup by 100 percent (1) as shown in the figure.

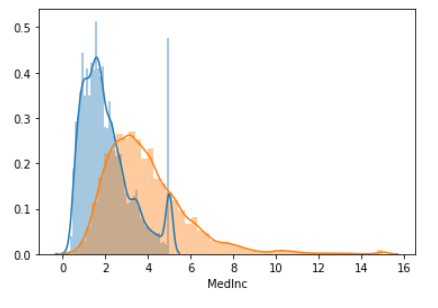
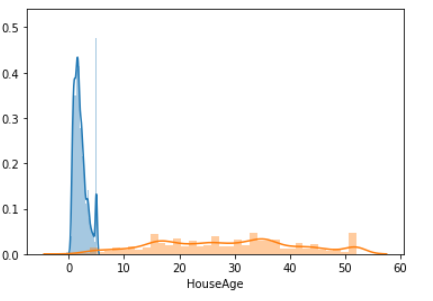
 

Fig. AveOccup and MedInc histogram Fig. AveOccup and HouseAge histogram

The above figures shows the histograms which represents the relation between AveOccup and MedInc and HouseAge which has correlation factor morethan 0.1 with AveOccup. The blue color represents AveOccup and orange represents other feature.

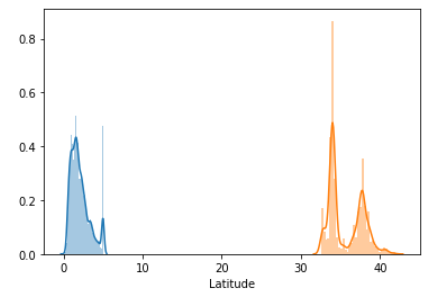
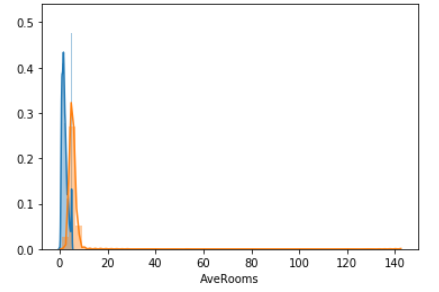


Fig. AveOccup and AveRooms Histogram Fig. AveOccup and HouseAge Histogram

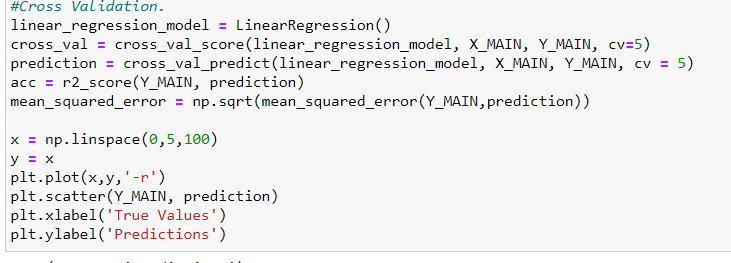
The above figures show the histograms of AveOccup and AveRooms and HouseAge which shows the correlation factor morthan 0.1 with AveOccup.

We have to remove the columns which do not have correlation more than 0.1 from the dataset and perform the regression analysis. All the rows with nulls are also cleaned from the data.

**Linear Regression**

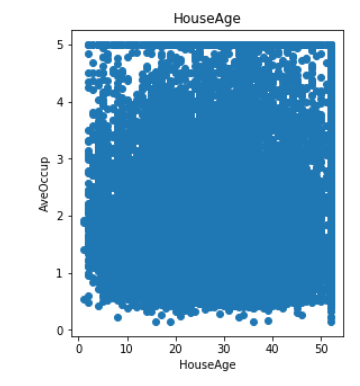
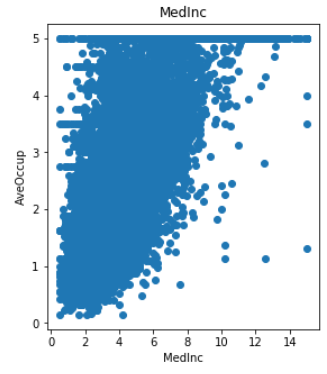
We are using built in linear regression method from sklearn package. Te package has many methods like linear regression, metrics etc. For creating any data analysis model, we need test and training sets. The training tests needs to be random and does not belong to any specific group.

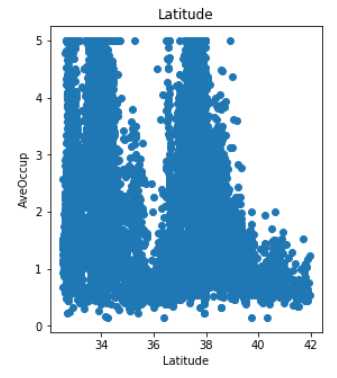
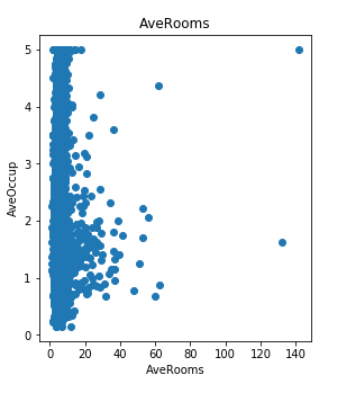
We use cross validation to implement the test and training sets and for that we can import cross\_val\_score and cross\_val\_predict from sklearn model package. To initiate the regression model, we need to import sklearn linear model package and parameter is cross\_val\_sccore.



The Cross\_val\_score method returns the scores for each iteration of cross validation in an array which gives training and test datasets for every iteration. The both methods cross\_val\_score and cross\_val\_predict calls the same parameters and the prediction method returns the Y values which are predicted values for each of the cross validation iterations.

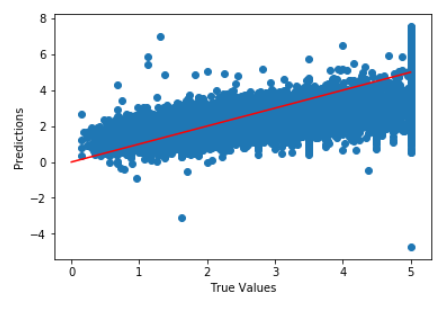
**Visualization**

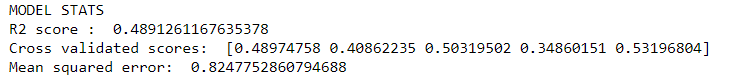




The above figures shows the scatter plots of different fields with relation to the AveOccup which are used to analyze the linear regression.

The below figure shows the scatter plot of predicted and actual values. The points which are closer to the red line represents the less error.





The mean squared error and cross validation scores are shown in above fig.