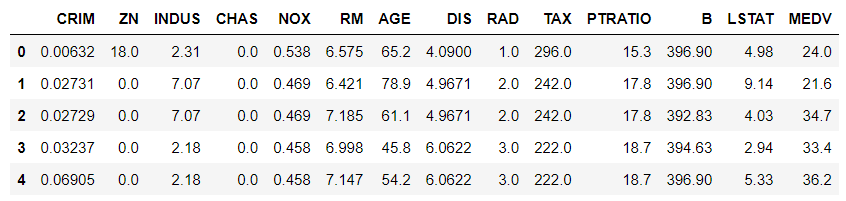
**Predicting Housing Prices**

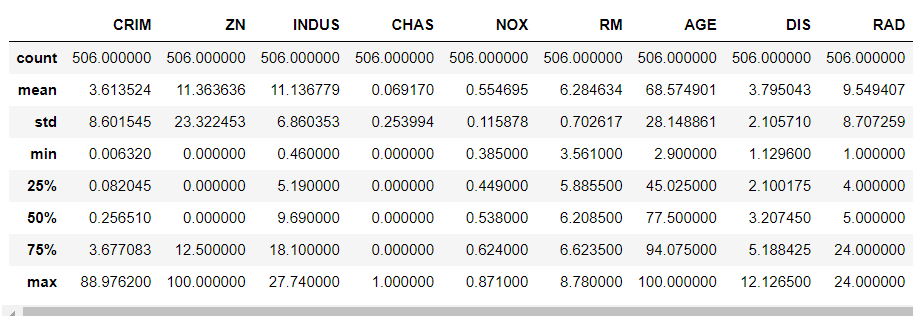
1. **Loading data and exploration**

For predicting the housing prices, we leverage the ‘Boston Housing Prices’ data set.

* Upon initial data description we notice that there are 14 numeric variables, with ***MEDV*** variable being the Median housing price and a few other variables like ***RM*** being the average number of rooms per dwelling; ***LSTAT*** referring to the percentage lower status of the population

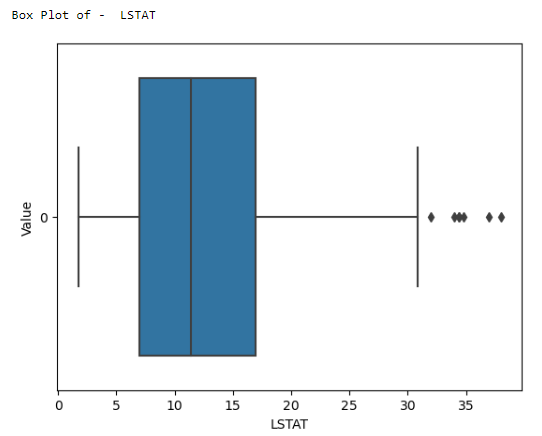
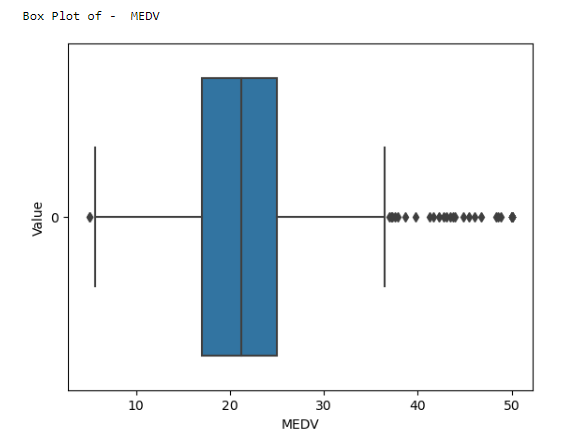


* We also do a descriptive analysis to look at the spread of the data distribution for all the numeric variables and find out that for the variables there is variation in the mean with the values ranging from ~0.07 to ~408. This highlights the need for normalizing the data.

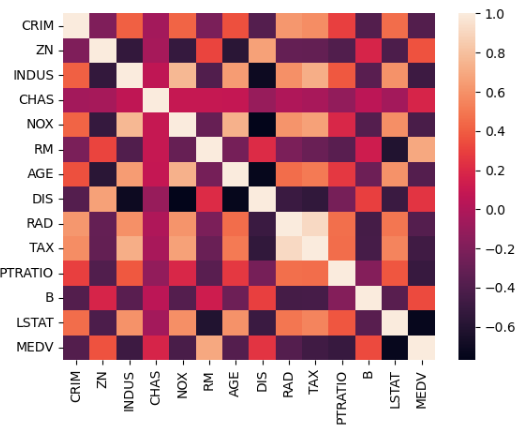
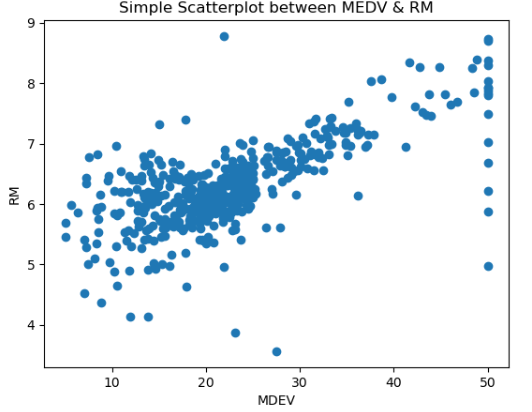


1. **Data processing**

* Next step is to look for **missing values**, and we notice that there are **no** missing values in the data
* Based on the **box plot** – we see that for a few variables there are **outlier values** present in the data set. Using the IQR (Inter Quartile Range) approach we find that there ***are around 13% outlier data points which have been removed before further processing***



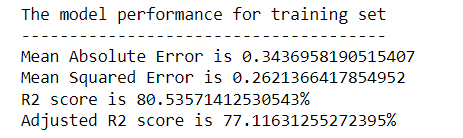
* We also perform the correlation analysis and find that the outcome variable ***MEDV*** (which is our outcome variable) has a moderate positive correlation of ~0.695 with ***RM*** variable and moderate negative correlation of ~ -0.73 with ***LSTAT*** variable. The same conclusion was also generated when we **performed scatter plot** between these variables and we observed positive and negative linear relationship between these variables.

****

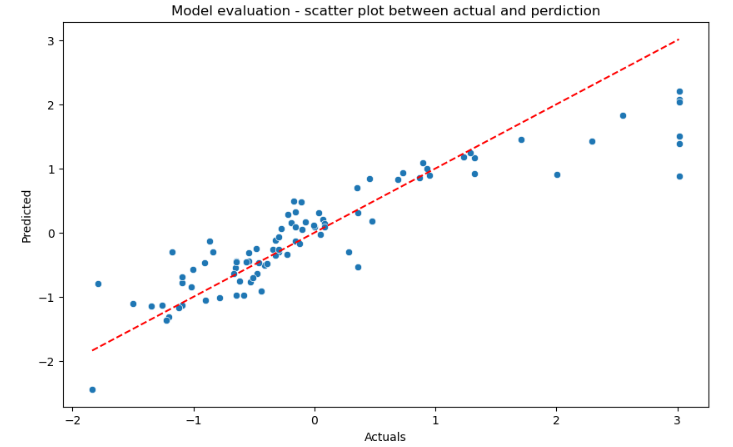
* Since we observed that there is difference in the mean values across the data, **Data Scaling** using ***Standard scaler*** was performed before model creation

1. **Model training and evaluation**

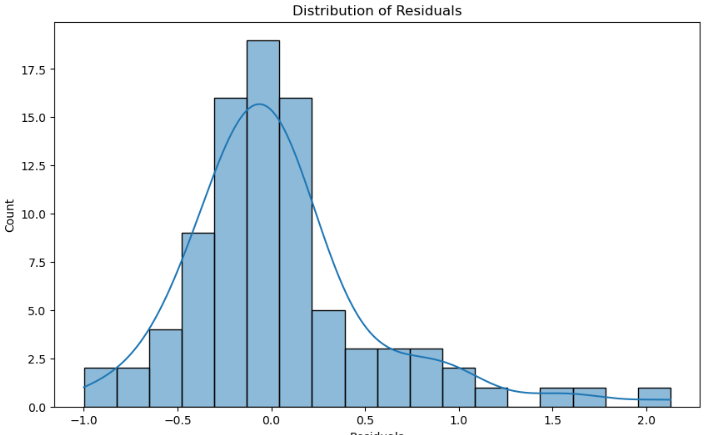
* In order the *predict the housing prices* a **Multiple Linear Regression** was built with the ***MDEV variable as the outcome*** variable and all other variables as the independent or regressor variables
* The data was split into 80% training data and 20% test data. The model was built on the 80% training set and the prediction was made on the 20% test data. The actual values from the test set were compared with the predicted values from the model and the following error metrics (MAE, MSE, R square and Adjusted R square) were calculated. Summarising the values below -



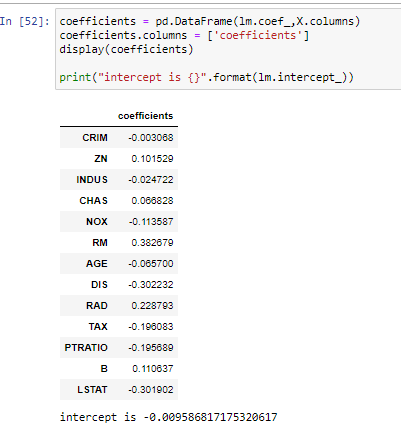
* We also look at the model performance by looking that distribution of the actuals and predicted values using a scatter plot and we see that actuals and predictions closely follow the 45 degree line highlighting that the prediction is very close to the actual value

****

* We also look at the distribution of the residuals and see that it closely resembles a normal distribution



* **Feature Importance –** To understand the importance of the features used in the model creation phase we look at the coefficient values from the regression model, as shown below -

****

In the above we can see coefficients of all the features in our data. Based on the above data we see that the variables -

* RM has a high coefficient value of 0.38 and DIS & LSTAT also have high negative coefficient values of -0.30. This indicates that these variables have a very strong correlation with the outcome variable

1. **Conclusion**

Using a multiple linear regression, we were able to make the prediction for estimating the housing prices using 35% error (based on MAE) with an almost normal distribution of residuals.

But due to presence of few outliers the data is not exactly normal distribution indicating scope of improvement for the prediction model.

There is a scope for improvement using k-fold cross validation or even trying decision tree based prediction.