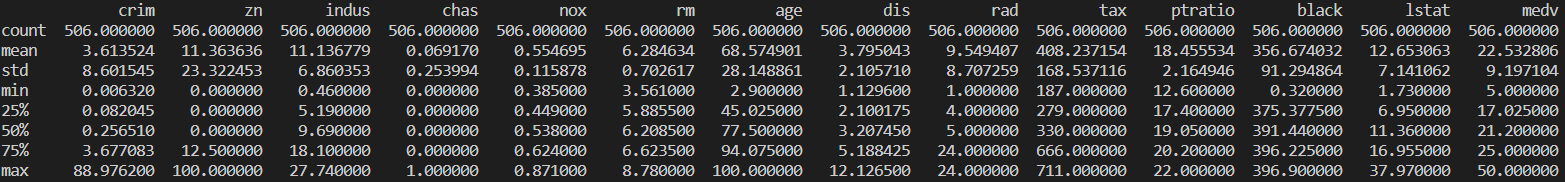
Multiple Linear Regression Lab

Questions:

1. Using the Boston dataset and the EDA, I can see that the candidate features for the model are CRIM(per capita crime rate by town) , ZN(proportion of residential land zoned for lots over 25,000 sq.ft.) , INDUS(proportion of non-retail business acres per town) , CHAS (Charles river dummy variable) , NOX(nitric oxides concentration ppm) , RM(average number of rooms per dwelling) , AGE(proportion of owner-occupied units built prior to 1940) , DIS(weighted distances to five Boston employment centers) , RAD(index of accessibility to radical highways) , TAX(full-value property-tax rate per $10,000) , PTRATIO(pupil-teacher ratio by town) , BLACK(black is the proportion of black residents by town) , LSTAT(percent of lower status of the population) , MEDV (median value of owner-occupied homes in $1000s)
2. After calculating the descriptive statistics for the dataset:

The information from this printout tells us the mean, standard deviation, minimum, 25th percentile, 50th percentile, 75th percentile, and max. This gives us more information about the data and gives us a deeper understanding of each of the columns and features we will be using in the model.

1. After plotting a histogram of each of the columns, and seeing the distribution plots, I am able to see a lot of the plots are skewed and not normal. The only roughly normal plots I am able to see are the MEDV and RM.

A screenshot of a graph

Description automatically generated with low confidence

1. I started off with 14 out of 14 features which gave a 74.06% score. This continued with the numbers of features of 13 and 12. The score then changed for 11 features dropping the score to 73.4267%. The score continued dropping slightly until I got to 6 features. After changing the number of features to 5, the score dropped drastically. Because of this, I am certain that the appropriate number of features is 6 features.
   1. These are the features to keep:
      1. CHAS
      2. NOX
      3. RM
      4. DIS
      5. PTRATIO
      6. MEDV

A screenshot of a computer code

Description automatically generated with low confidence

1. The score dropped drastically after I changed the number of features from 6 to 5. The score went from 71.577% to 63.36%. Because of this, the appropriate number of features is 6, with the following features: CHAS, NOX, RM, DIS, PTRATIO, MEDV.
2. By using the StandardScalar(), I was able to standardize all of the values from all of the features. The screenshot below shows some of these standardized values from some of the rows of data:

A screenshot of a computer program

Description automatically generated with medium confidence

1. I performed Normalization on the data using the Normalizer() and scaled the data to a range between 0 and 1. I have provided a screenshot below on what code I wrote, and how my data turned out:

A screen shot of a computer program

Description automatically generated with low confidence

I also tried another version of normalization using the MinMaxScaler(), This is what the output showed:

A screenshot of a computer program

Description automatically generated with medium confidence

1. With the standardized data, I got the descriptive statistics in the screenshot below:

A screenshot of a computer program

Description automatically generated with medium confidence

I have also got the histograms in the screenshot below:

A screenshot of a graph

Description automatically generated with low confidence

I also used RFE to see the number of features needed. The score dropped drastically after going lower than 5 features. This shows that 5 features are required to make predictions using the Standardized data.

A picture containing text, screenshot, software, multimedia software

Description automatically generated

1. With the normalized data, I got the descriptive statistics below:

A screenshot of a computer screen

Description automatically generated with medium confidence

I then got histograms such as the following in the screenshot below:

A screenshot of a graph

Description automatically generated with low confidence

I also used RFE to see the number of features needed. The score dropped drastically after going lower than 10 features. This shows that 10 features are required to make predictions using the Normalized data.

A picture containing text, screenshot, software, font

Description automatically generated