Outlier Detection Model

Dataset: Housing data values about Boston suburbs.

Number of attributes: 14

All attributes: numerical

Attributes: CRIM, ZN, INDUS, CHAS, NOX, RM, AGE,DIS, RAD, TAX,PTRATIO, B, LSTAT, MEDV

MEDV - target

CHAS attribute is binomial (0 or 1 values) --> will be excluded from the detection model.

Data source file head:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
1	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
2	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
3	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2
4	0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222.0	18.7	394.12	5.21	28.7

506 rows

Exploratory Data Analysis

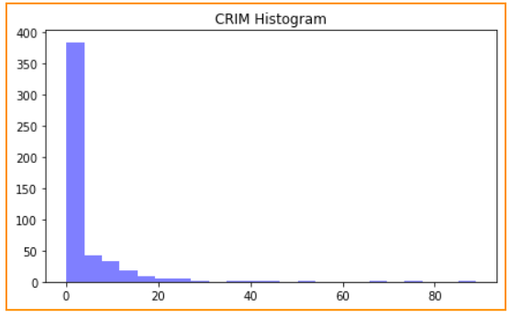
Descriptive statistics of the dataset:

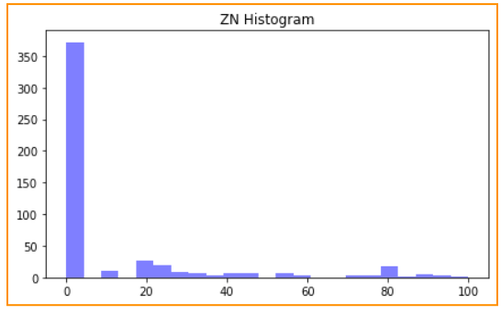
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000	505.000000
mean	3.620667	11.350495	11.154257	0.069307	0.554728	6.284059	68.581584	3.794459	9.566337	408.459406	18.461782	356.594376	12.668257	22.529901
std	8.608572	23.343704	6.855868	0.254227	0.115990	0.703195	28.176371	2.107757	8.707553	168.629992	2.162520	91.367787	7.139950	9.205991
min	0.009060	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
1%	0.013830	0.000000	1.252800	0.000000	0.398000	4.523360	6.608000	1.205712	1.000000	188.000000	13.000000	6.720000	2.882400	7.008000
25%	0.082210	0.000000	5.190000	0.000000	0.449000	5.885000	45.000000	2.100000	4.000000	279.000000	17.400000	375.330000	7.010000	17.000000
50%	0.259150	0.000000	9.690000	0.000000	0.538000	6.208000	77.700000	3.199200	5.000000	330.000000	19.100000	391.430000	11.380000	21.200000
75%	3.678220	12.500000	18.100000	0.000000	0.624000	6.625000	94.100000	5.211900	24.000000	666.000000	20.200000	396.210000	16.960000	25.000000
99%	41.402104	90.000000	25.650000	1.000000	0.871000	8.335400	100.000000	9.222796	24.000000	666.000000	21.200000	396.900000	33.938800	50.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

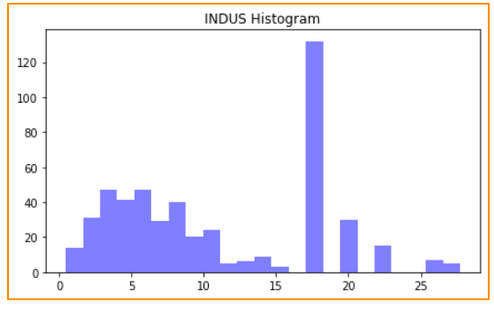
Conclusion: there are values in the data set which are less than 1-th percentile or greater than 99-th percentile in the distribution. It could be a sign that outliers are present. Further analysis and visualization of the dataset are needed.

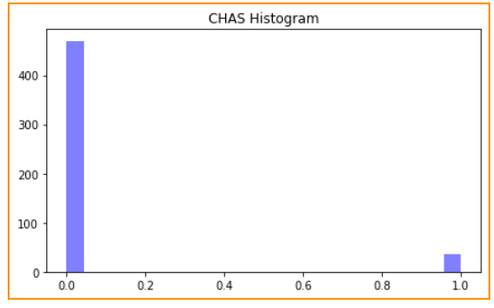
Histograms of the signals

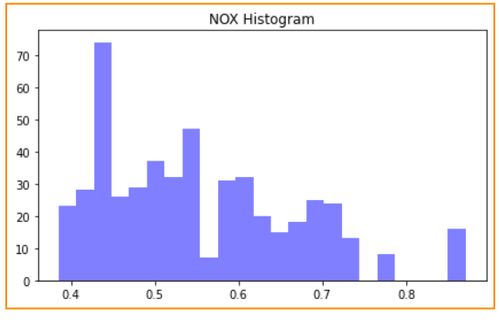
Help to visualize the distribution of the data points and check the outliers.

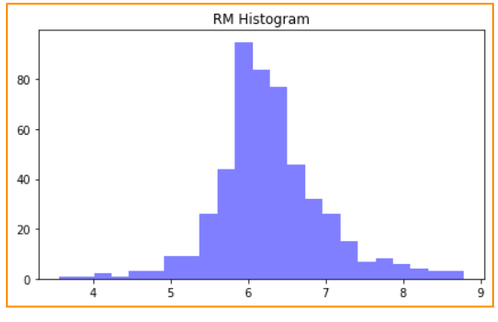


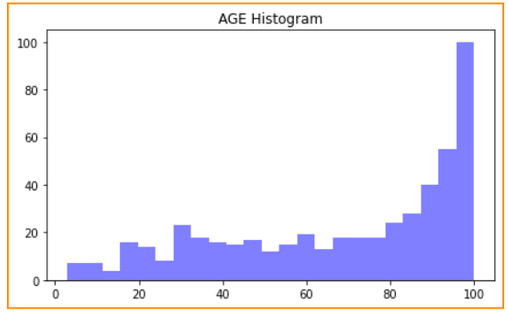


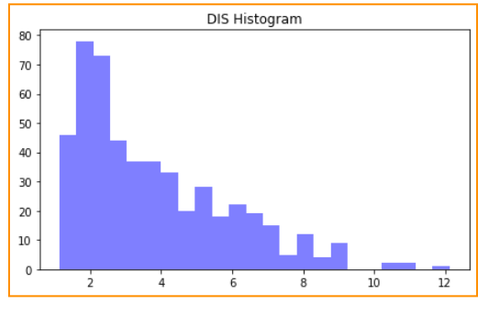


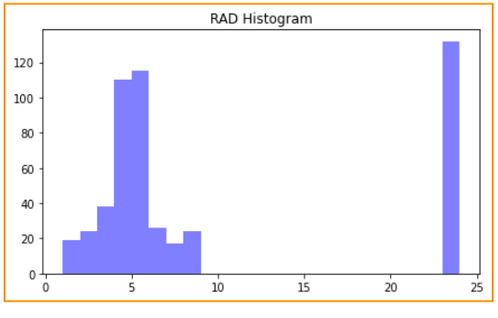


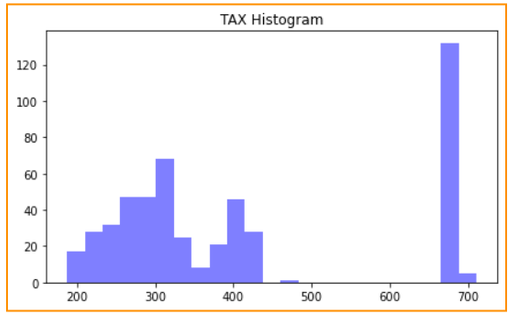


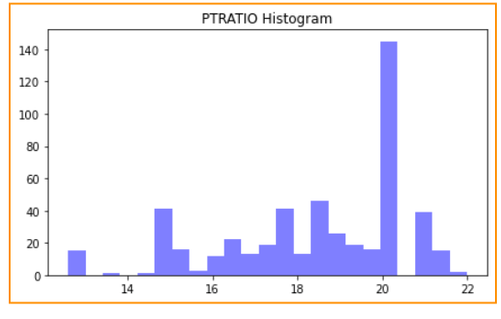


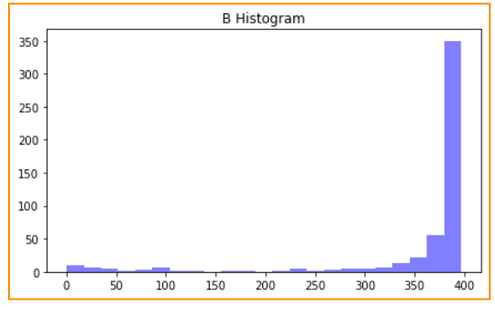


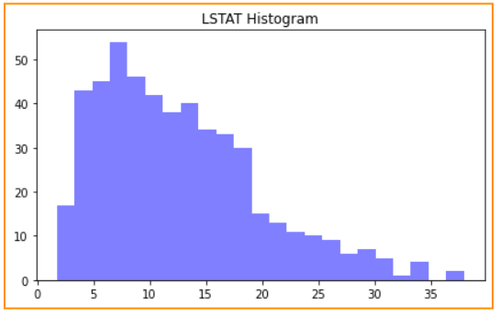








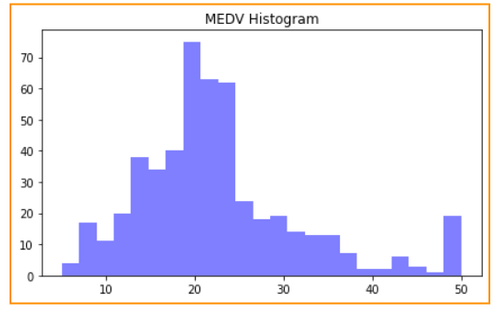




Conclusion:

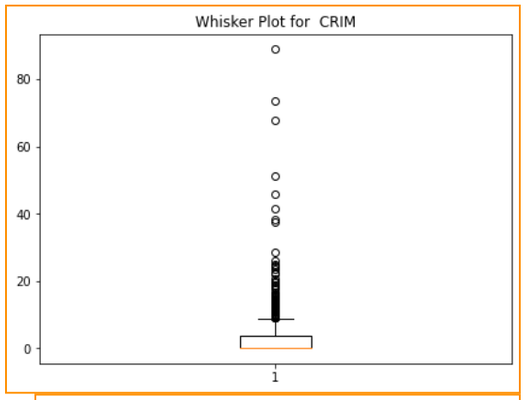
All signals (except binomial CHAS) are normally distributed, some of them have skewness to the right or to the left.

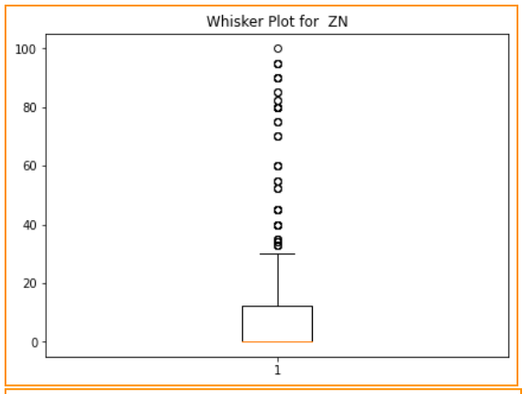
The histograms of some signals demonstrate and visualize that outliers are present.

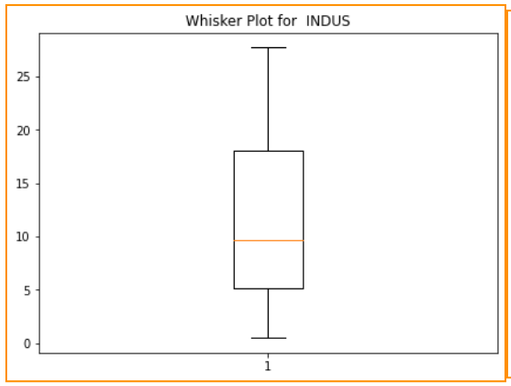


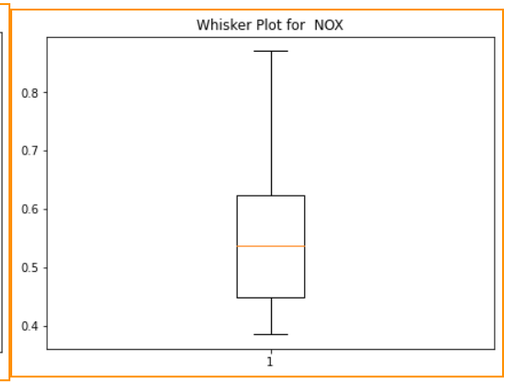
Box_and_Whisker Plots

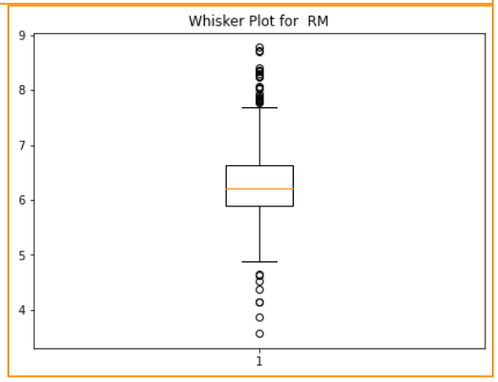
Help to visualize the outliers - the values away from max and min data points.

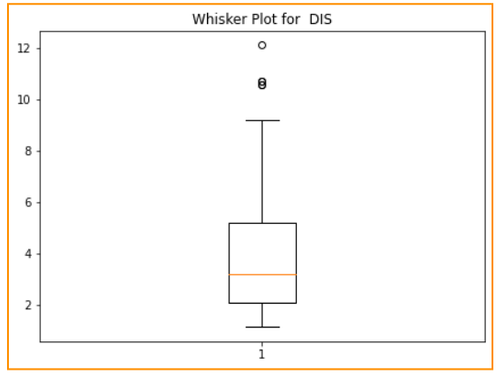


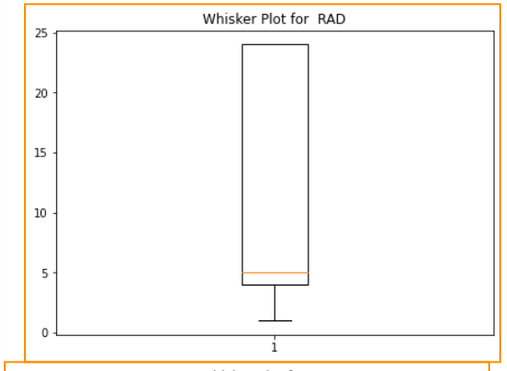


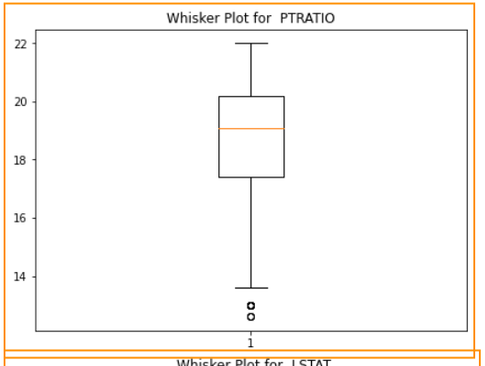


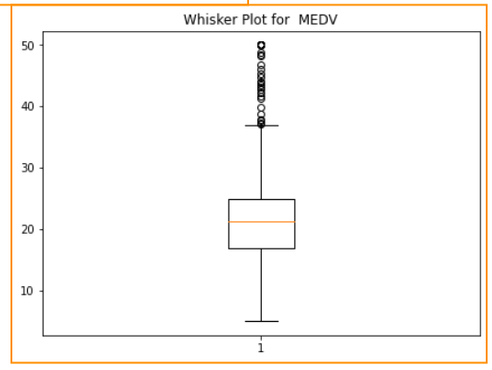


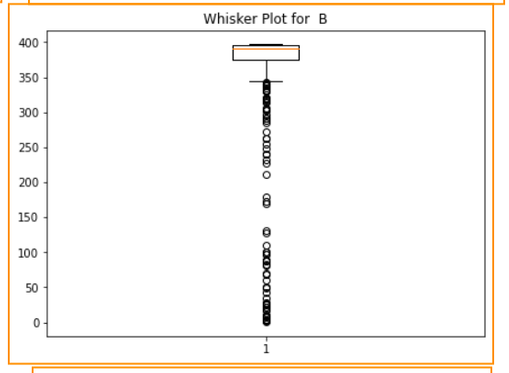


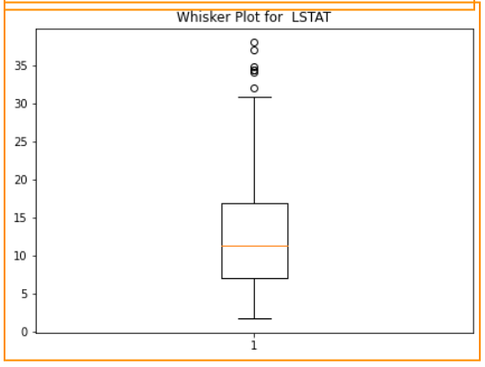


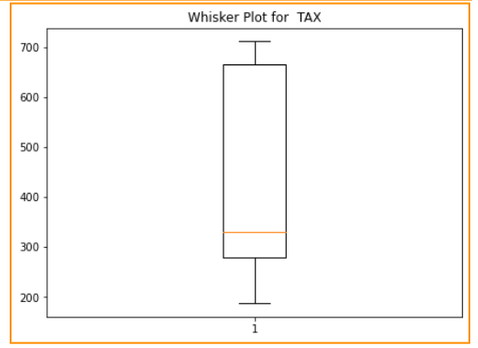












CONCLUSION

Whisker plots prove that signals MEDV, B, LSTAT, PTRATIO, RM, DIS, CRIM, ZN have the outliers.

Outlier Detection model is based on the IQR (Inter Quartile Range) technique

IQR = Q3 - Q1

Q3 is a value that represents 75-th percentile, Q1 is a value that represents 25-th percentile

According to the IQR approach outliers are values below lower bound and above the upper bound

Lower bound = Q1 - 1.5 * IQR

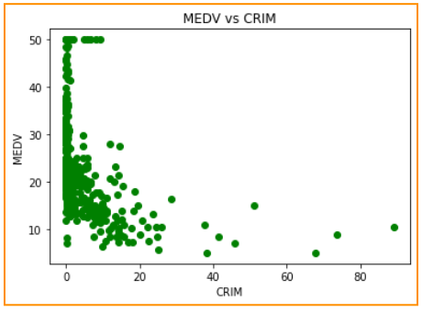
Upper bound = Q3 + 1.5 * IQR

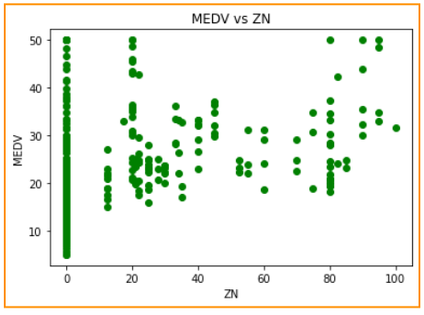
Outlier Detection Model:

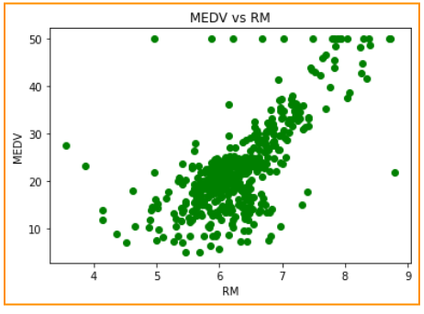
1. Calculates the bounds for all signals' distribution

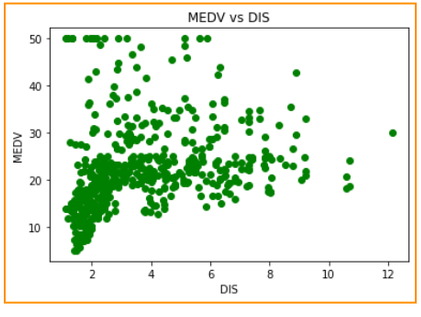
2. Marks and visualizes the outliers on the scatter plots shown the correlation between the target and other signals.

Before using the model

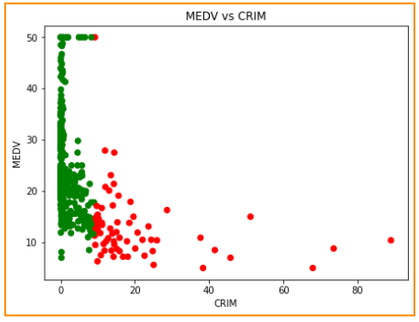


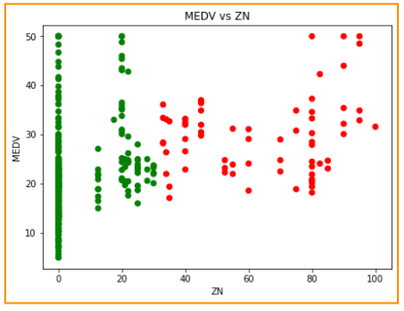


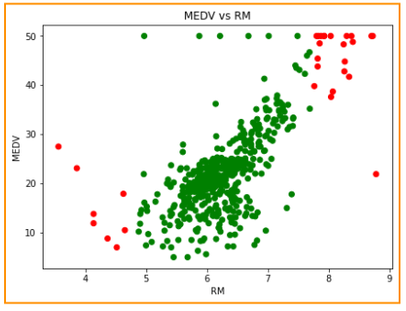


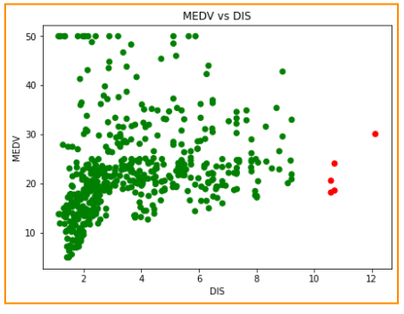


After using the model

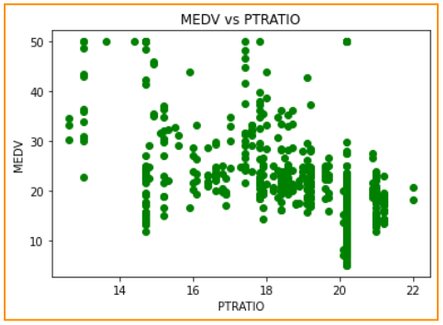


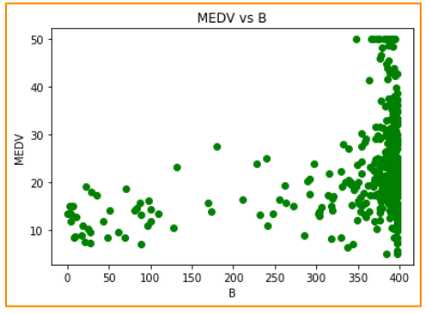


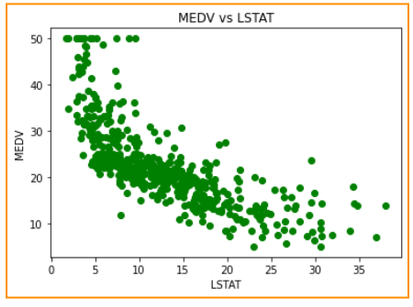


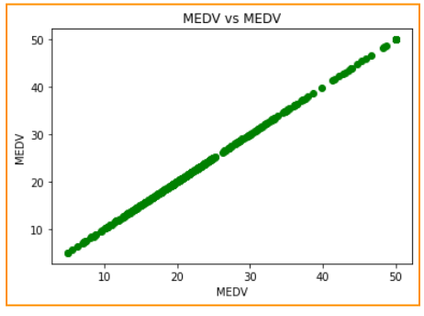


Before using the model

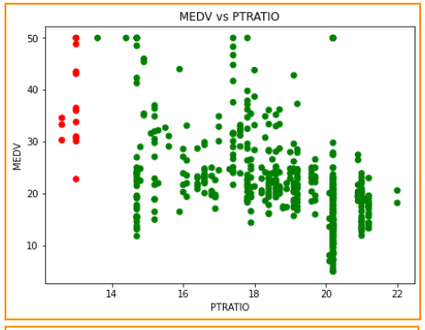


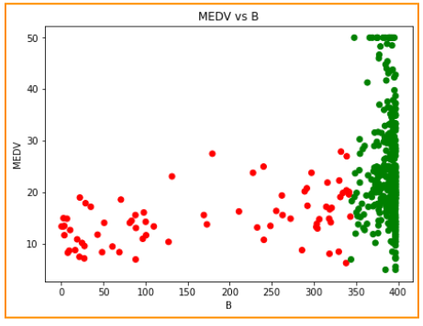


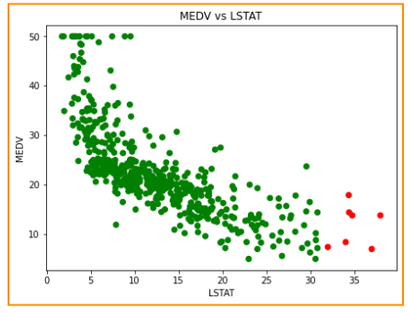


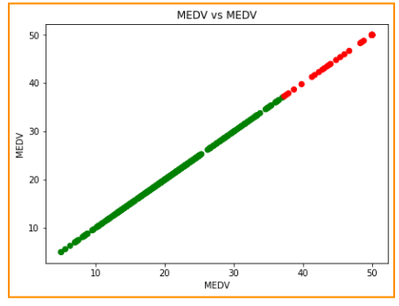


After using the model









CONCLUSIONS

The model detected the outliers in 8 signals of the dataset.

The Outlier Detection Model's code is re-usable and could be implemented on any other dataset with the numerical features and normal distribution of their data points.

The detection of the outliers is an important part of the data cleaning and pre-processing:

- shows the skewness or anomaly of the data points in the distribution

- the outliers must be treated to prevent underperforming of the future prediction model and enhance the accuracy of the model

Possible outliers' treatment:

1. Removing the outliers from the data set

2. Filling the outliers with the median values

3. Filling the outliers with the values of the 10-th(5-th) or 90-th(95-th) percentile

4. Log Transformation of the skewed signals

Next steps after the outliers detection and treatment:

OLS (Ordinary Least Squares) analysis to find significant features

Evaluating the scatter plots and OLS, Feature Engineering

Model training and testing (multiple regression model, polynomial regression model)

Predicting of the target - MEDV in this case.