**Outlier Treatments**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable as Python File (.py) and R file as .r extension files.

Please ensure you update all the details:

**Name: Navya.P**

**Batch Id: Data Science WDEO 171220**

**Topic: Preliminaries for Data Analysis**

**Problem Statement:**

Most of the datasets have extreme values or exceptions in their observations. These values affect the predictions (Accuracy) of the model in one way or the other, removing these values is not a very good option. For these types of scenarios, we have the techniques for treating such values. Explore on various other techniques to treat these values, you can go through this link:

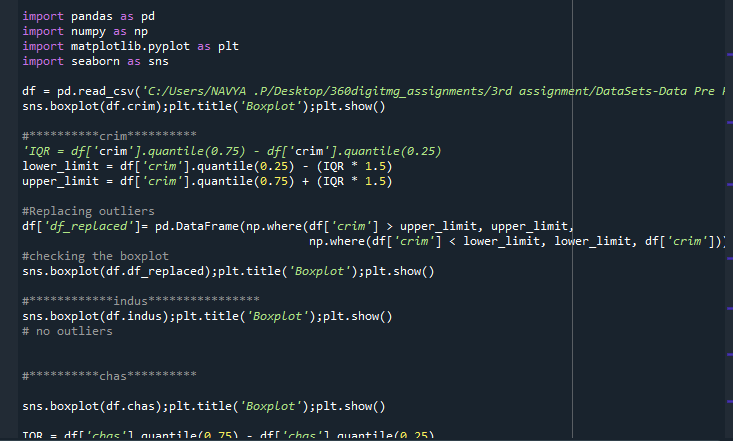
<https://360digitmg.com/mindmap-data-science>

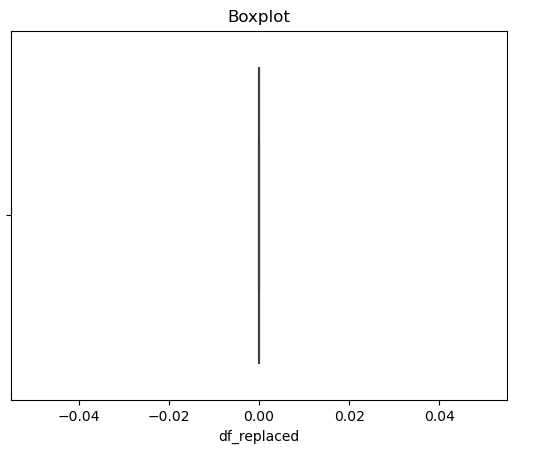
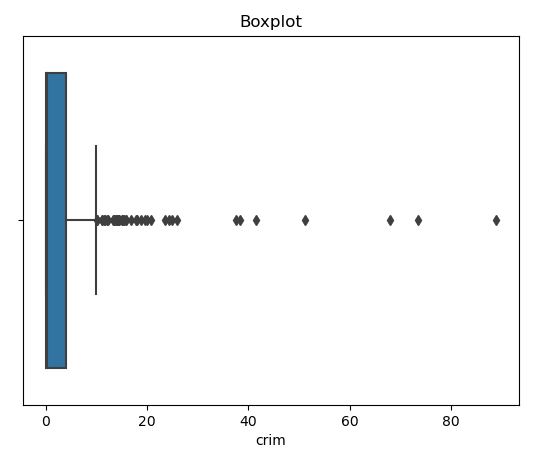
1. Prepare the dataset by performing the preprocessing techniques, to treat the outliers to improve the model prediction.

In R, the outliers are treated by the process of winsorization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Column name* | *No. of outliers* | *Q1,Q3* | *Winsorization limits*  *(5%, 95%)* | *Effect of Winsorization* |
| *Boston $ Crim* | *42* | *0.0823825,  4.0531575* | *15.33116,*  *41.43388* | *Solved outliers* |
| *Boston $ zn* | *49* | *0.0*  *12.5* | *0 1%*  *26.74 86%* | *Solved outliers* |
| *Boston $ Indus* | *0* |  |  |  |
| *Boston $ Chas* | *28* | *0*  *0* | *0*  *1* | *Solved outliers* |
| *Column name* | *No. of outliers* | *Q1,Q3* | *Winsorization limits*  *(5%, 95%)* | *Effect of Winsorization* |
| *Boston$nox* | *0* |  |  |  |
| *Boston$rm* | *21* | *5.90275 6.62925* | *5.39135*  *7.46250* | *Solved outliers* |
| *Boston$age* | *0* |  |  |  |
| *Boston$dis* | *5* | *2.087875 5.222125* | *1.470560*  *7.827605* | *Solved outliers* |
| *Boston$rad* | *0* |  |  |  |
| *Boston$tax* | *0* |  |  |  |
| *Boston$ptratio* | *12* | *17.375 20.200* | *14.7*  *21.0* | *Solved outliers* |
| *Boston$black* | *61* | *374.7100 396.0075* | *70.807*  *396.900* | *Solved outliers* |
| *Boston$lstat* | *2* | *7.135*  *16.910* | *3.9620*  *26.6115* | *Solved outliers* |
| *Boston$medv* | *27* | *17.1*  *25.0* | *10.200*  *30.416* | *Solved outliers* |

In python, the outliers are removed and replaced by the maximum and minimum limits.





**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Hint: Boston dataset is publicly available. Refer to Boston.csv file.
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function
7. Detailed explanation of your approach is mandatory