**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: s

**Batch Id:**

**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.

A picture containing shape, arrow

Description automatically generated**

**ANSWER:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Crim | Per capita crime rate by town | Quantitative, Continuous, Ratio | Relevant |
| Zn | Proportion of residential land zoned for lots over 25,000 sq.ft | Quantitative, Discrete, Interval | Relevant |
| Indus | Proportion of non-retail business acres per town | Quantitative, Discrete, Interval | Irrelevant |
| Chas | Charles River dummy variable | Qualitative, Nominal | Irrelevant |
| Nox | Nitric oxides concentration(parts per 10million) | Quantitative ,Continuous, Ratio | Relevant |
| Rm | Average number of rooms per dwelling | Quantitative ,Continuous, Interval | Relevant |
| Age | Proportion of owner-occupied units built prior to 1940 | Quantitative ,Continuous, Ratio | Relevant |
| Dis | Weighted distances to five Boston employment centres | Quantitative ,Continuous, Ratio | Relevant |
| Rad | Index of accessibility to radial highways | Quantitative ,Discrete, Ordinal | Irrelevant |
| Tax | Full-value property tax rate per $10,000 | Quantitative ,Discrete, Ratio | Relevant |
| Ptratio | Pupil-teacher ratio by town | Quantitative ,Discrete, Interval | Relevant |
| Black | The proportion of blacks by town | Quantitative ,Continuous, Interval | Relevant |
| Istat | % lower status of the population | Quantitative ,Continuous, Ratio | Relevant |
| Medv | Median value of owner occupied homes | Quantitative , Continuous , Ratio | Relevant |

**BUSINESS PROBLEM**: Not able to predict accurately.

**CONSTRAINT**: Data loss is a big challenge.

**DATA UNDERSTANDING:**

1. All the columns have float values.
2. There are no missing values in the dataset.

**INSIGHTS FROM THE DATA:**

1. crim, zn, rm, dis, black, istat and medv columns have outliers.
2. Outliers detection by using boxplot.
3. Outliers detection by calculating upper limit and lower limit using Inter Quartile Range.
4. Used Trimming techniques to remove outliers(crim, zn & rm) but still outliers are there.
5. And finally outliers are removed by using Replacement technique.
6. Winsorization technique has been used to remove outliers(dis, black, istat & medv)

**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