

**Outlier Treatments**

**Instructions**:

Please share your answers filled inline in the word document. Submit code files wherever applicable.

Please ensure you update all the details:

**Name: RAKESH SETHU NP**

**Batch Id: 11052022\_7.30PM**

**Topic: Data Pre-Processing**

**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 various techniques to treat such values.

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

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

A picture containing shape, arrow

Description automatically generated**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Crim | per capita crime rate by town | Continuous, Ratio | It is a useful information |
| Zn | proportion of residential land zoned for lots over 25,000 sq. ft. | Continuous, Ratio | It is a useful information |
| Indus | proportion of non-retail business acres per town. | Continuous, Ratio | It is a useful information |
| Chas | Charles River dummy variable (1 if tract bounds river; 0 otherwise) | Discrete, Ordinal | It is a useful information |
| Nox | nitric oxides concentration (parts per 10 million) | Continuous, Ratio | It is a useful information |
| Rm | average number of rooms per dwelling | Continuous, Ratio | It is a useful information |
| age | proportion of owner-occupied units built prior to 1940 | Continuous, Ratio | It is a useful information |
| Dis | weighted distances to five Boston employment centers | Continuous, Ratio | It is a useful information |
| Rad | index of accessibility to radial highways | Continuous, Ratio | It is a useful information |
| Tax | full-value property-tax rate per $10,000 | Continuous, Ratio | It is a useful information |
| ptrati  black | pupil-teacher ratio by town | Continuous, Ratio | It is a useful information |
| Lstat | % lower status of the population | Continuous, Ratio | It is a useful information |
| medv | Median value of owner-occupied homes in $1000's | Continuous, Ratio | It is a useful information |

**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