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: ULLI VENKATA SAI KUMAR

Batch Id: 04072024HYD10AM

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.

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

```
import pandas as pd

import numpy as np


data = pd.read_csv(r"Boston.csv")


# Treating outliers with Winsorization

from scipy.stats.mstats import winsorize


# Apply Winsorization to 'crim' as an example

data['crim'] = winsorize(data['crim'], limits=[0.05, 0.05])


# Example of a log transformation on 'tax'

data['tax'] = np.log(data['tax'])


# Capping 'ptratio' at 1st and 99th percentiles

lower_bound = data['ptratio'].quantile(0.01)

upper_bound = data['ptratio'].quantile(0.99)
```

data['ptratio'] = np.clip(data['ptratio'], lower_bound, upper_bound)

Check the transformed dataset

print(data)

```
In [10]: print(data)
         crim    zn  indus  chas    nox  ...       tax  ptratio   black  lstat  medv
0     0.15876   0.0  10.81   0.0  0.413  ...  5.720312     19.2  376.94   9.88  21.7
1     0.10328  25.0   5.13   0.0  0.453  ...  5.648974     19.7  396.90   9.22  19.6
2     0.34940   0.0   9.90   0.0  0.544  ...  5.717028     18.4  396.24   9.97  20.3
3     2.73397   0.0  19.58   0.0  0.871  ...  5.998937     14.7  351.85  21.45  15.4
4     0.04337  21.0   5.64   0.0  0.439  ...  5.493061     16.8  393.97   9.43  20.5
..        ...   ...    ...   ...    ...  ...       ...      ...     ...    ...   ...
399   9.32909   0.0  18.10   0.0  0.713  ...  6.501290     20.2  396.90  18.13  14.1
400  15.57570   0.0  18.10   0.0  0.597  ...  6.501290     20.2    2.60  10.11  15.0
401   0.02875  90.0   1.21   1.0  0.401  ...  5.288267     13.6  395.52   3.16  50.0
402   0.02875  85.0   0.74   0.0  0.410  ...  5.746203     17.3  396.90   5.77  24.7
403   0.08244  30.0   4.93   0.0  0.428  ...  5.703782     16.6  379.41   6.36  23.7

[404 rows x 14 columns]
```