

**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: Gunjan Kumar Gupta**

**Batch Id: DSWDMCSR 300522B**

**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**

**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

# -\*- coding: utf-8 -\*-

"""

Created on Tue Jul 5 14:59:16 2022

@author: Gunjan

"""

###### DataPreparation\_Outlier\_Treatment #########3

""" 1. Prepare the dataset by performing the preprocessing

techniques, to treat the outliers.

"""

import pandas as pd

import numpy as np

import seaborn as sns

df = pd.read\_csv(r"D:\DATA Science 360 DigiTMG\Assignment\Data Preprocessing\_Assignments\DataSets-Data Pre Processing\boston\_data.csv")

df.dtypes

df.columns # Cheking Columns of excel file

#finding ouliers

sns.boxplot(df.ptratio)

sns.boxplot(df.tax) # no outliers

IQR = df['ptratio'].quantile(0.75) - df['ptratio'].quantile(0.25)

lower\_limit = df['ptratio'].quantile(0.25) - (IQR \* 1.5)

upper\_limit = df['ptratio'].quantile(0.75) + (IQR \* 1.5)

############### 1. Remove (let's trim the dataset) ################

# Trimming Technique

# let's flag the outliers in the data set

outliers\_df = np.where(df['ptratio'] > upper\_limit, True, np.where(df['ptratio'] < lower\_limit, True, False))

sum(outliers\_df)

df\_trimmed = df.loc[~(outliers\_df), ]

df.shape, df\_trimmed.shape

# let's explore outliers in the trimmed dataset

sns.boxplot(df\_trimmed.tax) # now no outliers