

**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: Vishvash C**

**Batch Id: 23012024**

**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 the Boston.csv file.
3. Research and perform all possible steps for obtaining the solutions.
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.

import pandas as pd

df = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/EDA/InClass\_DataPreprocessing\_datasets/Boston.csv")

import seaborn as sns

import matplotlib.pyplot as plt

from scipy import stats

import numpy as np

fig, axs = plt.subplots(ncols=7, nrows=2, figsize=(20, 10))

index = 0

axs = axs.flatten()

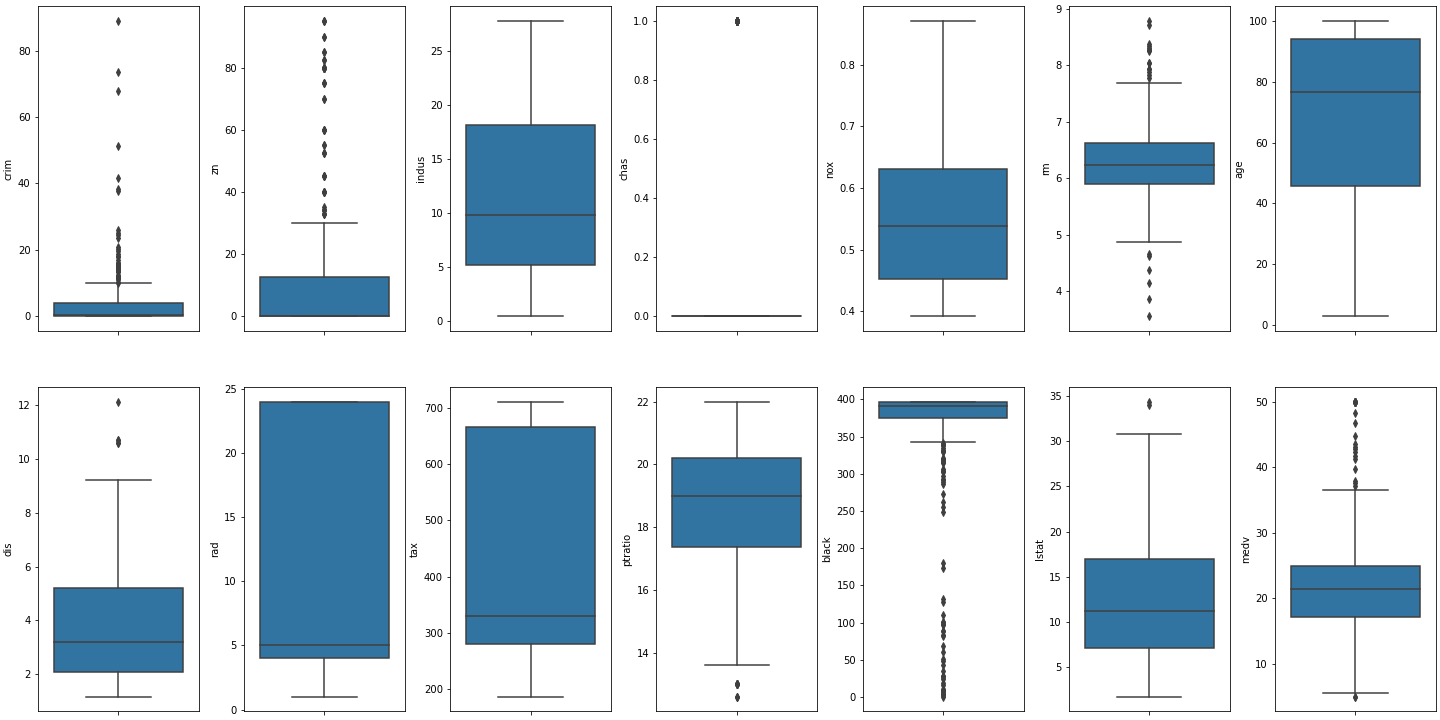
for k,v in df.items():

sns.boxplot(y=k, data=df, ax=axs[index])

index += 1

plt.tight\_layout(pad=0.4, w\_pad=0.5, h\_pad=5.0)

print(df.describe())



'''From get-go, two data coulmns show interesting summeries.

They are : ZN (proportion of residential land zoned for lots over 25,000 sq.ft.)

with 0 for 25th, 50th percentiles. Second, CHAS: Charles River dummy variable

(1 if tract bounds river; 0 otherwise) with 0 for 25th, 50th and 75th percentiles.

These summeries are understandable as both variables are conditional + categorical variables.

First assumption would be that these coulms may not be useful in regression task such as

predicting MEDV (Median value of owner-occupied homes).

'''

'''

Another interesing fact on the dataset is the max value of MEDV.

From the original data description, it says: Variable #14 seems to be censored at 50.00

(corresponding to a median price of $50,000). Based on that, values above 50.00 may not help to predict MEDV.

Let's plot the dataset and see interesting trends/stats.

'''

columns\_to\_drop = ['chas', 'zn']

df.drop(columns\_to\_drop, axis=1, inplace=True)

df = df[~(df['medv'] >= 50.0)]

fig, axs = plt.subplots(ncols=6, nrows=2, figsize=(20, 10))

index = 0

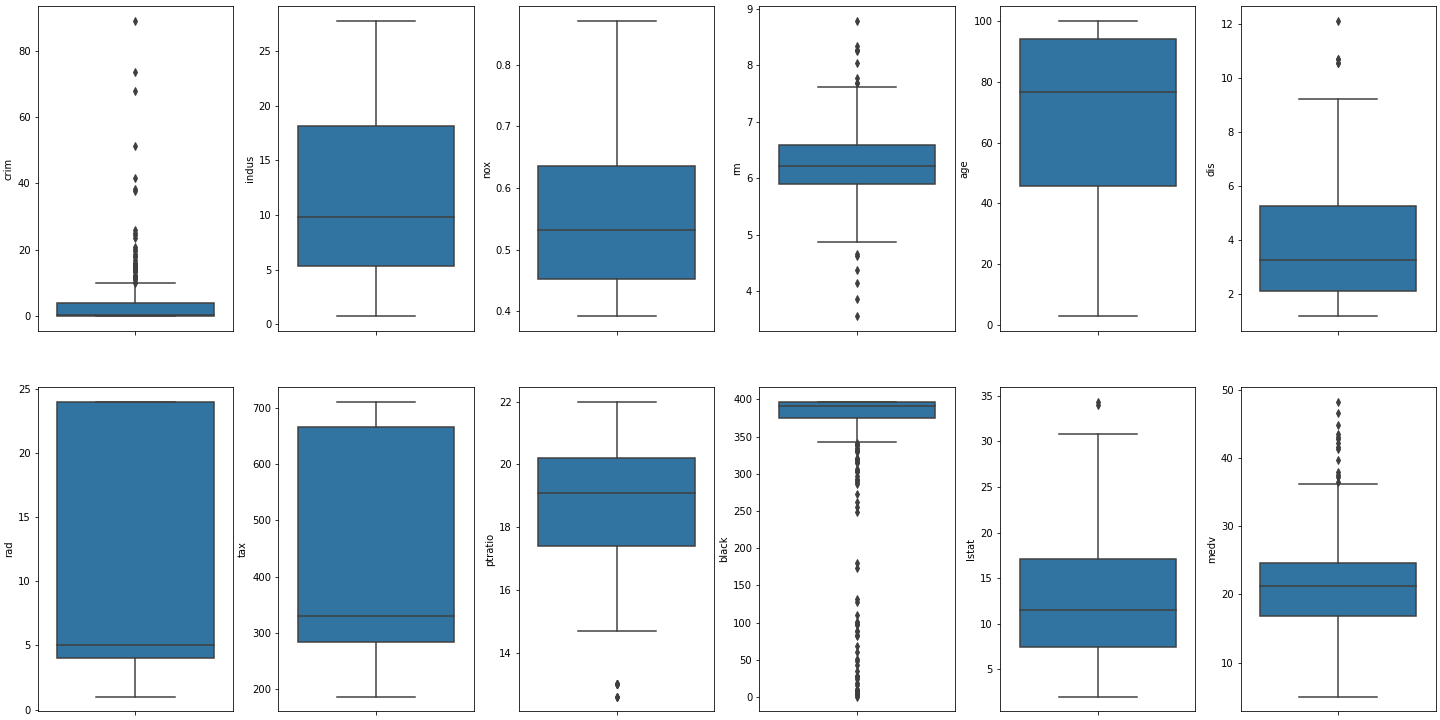
axs = axs.flatten()

for k,v in df.items():

sns.boxplot(y=k, data=df, ax=axs[index])

index += 1

plt.tight\_layout(pad=0.4, w\_pad=0.5, h\_pad=5.0)



fig, axs = plt.subplots(ncols=6, nrows=2, figsize=(20, 10))

index = 0

axs = axs.flatten()

for k,v in df.items():

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

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

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

df[k] = df[k].clip(lower=lower\_limit, upper=upper\_limit)

sns.boxplot(y=k, data=df, ax=axs[index])

index += 1

plt.tight\_layout(pad=0.4, w\_pad=0.5, h\_pad=5.0)

from feature\_engine.outliers import Winsorizer

winsor\_iqr = Winsorizer(capping\_method = 'iqr',

# choose IQR rule boundaries or gaussian for mean and std

tail = 'both', # cap left, right or both tails

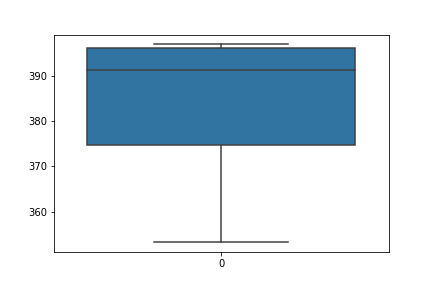
fold = 1,

variables = ['black'])

df1 = winsor\_iqr.fit\_transform(df[['black']])

sns.boxplot(df1.black)

#iqr works for removing outliers in column "black"



from feature\_engine.outliers import Winsorizer

winsor\_iqr = Winsorizer(capping\_method = 'quantiles',

# choose IQR rule boundaries or gaussian for mean and std

tail = 'both', # cap left, right or both tails

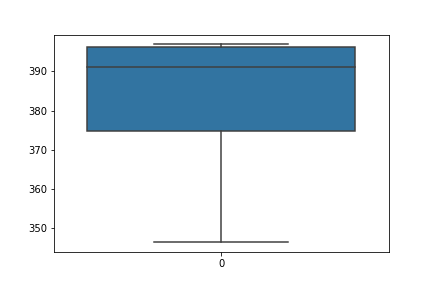
fold = 0.16,

variables = ['black'])

df1 = winsor\_iqr.fit\_transform(df[['black']])

sns.boxplot(df1.black)

# quantiles work for removing outliers in column "black"



from feature\_engine.outliers import Winsorizer

winsor\_iqr = Winsorizer(capping\_method = 'gaussian',

# choose IQR rule boundaries or gaussian for mean and std

tail = 'both', # cap left, right or both tails

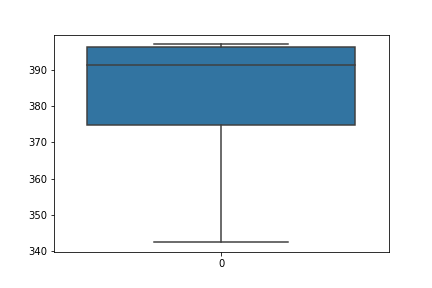
fold = 2,

variables = ['black'])

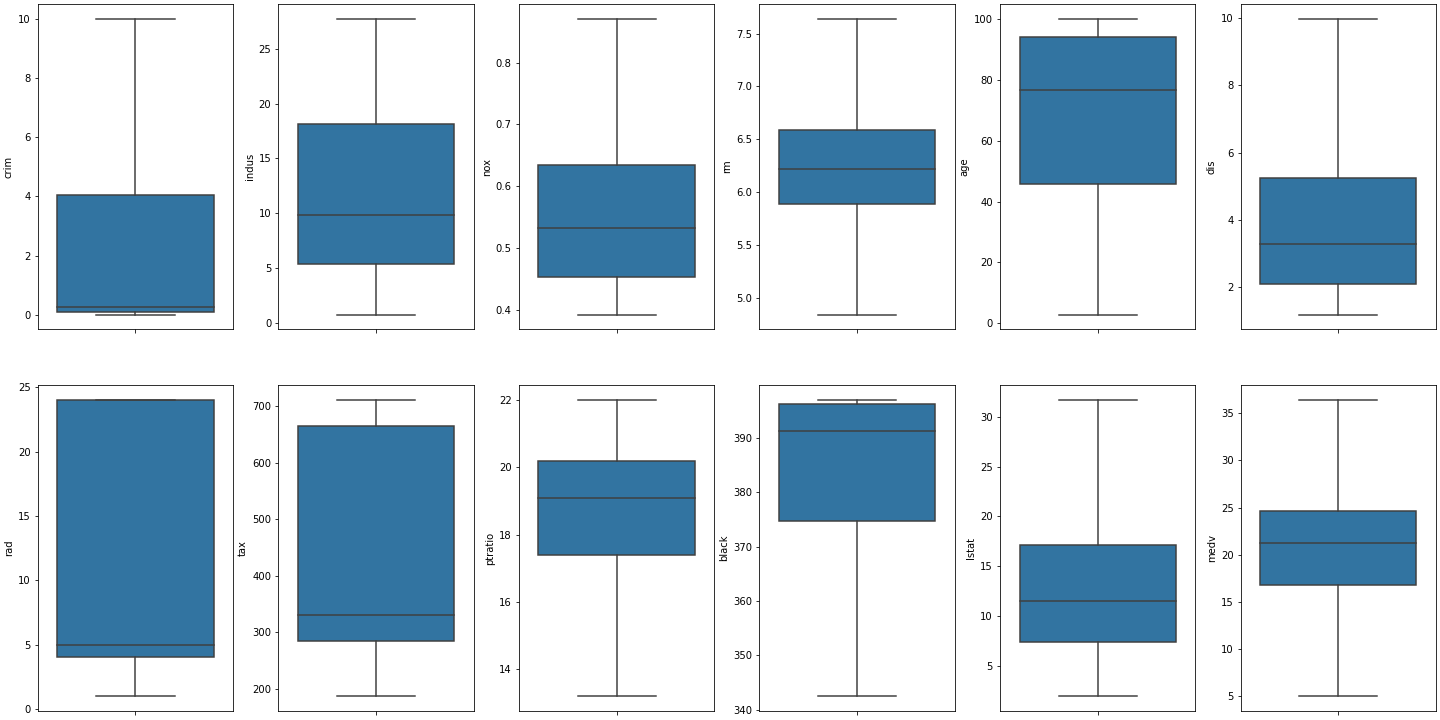
df1 = winsor\_iqr.fit\_transform(df[['black']])

sns.boxplot(df1.black)

# gaussian also worls for removing outliers in column "black"



**Final output as boxplots after handling outliers:**



|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| **ID** |  | **Quantitative/ Nominal** | **Irrelevant (ID does not provide useful information)** |
| crim | per capita crime rate by town | Quantitative | Relevant |
| zn | proportion of residential land zoned for lots over 25,000 sq.ft. | Nominal | Irrelevant (0 for 25th and 50th percentiles) |
| indus | proportion of non-retail business acres per town | Quantitative | Relevant |
| chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) | Nominal | Irrelevant (0 for 25th, 50th and 75th percentiles) |
| nox | nitric oxides concentration (parts per 10 million) | Quantitative | Relevant |
| rm | average number of rooms per dwelling | Quantitative | Relevant |
| age | proportion of owner-occupied units built prior to 1940 | Quantitative | Relevant |
| dis | weighted distances to five Boston employment centres | Quantitative | Relevant |
| rad | index of accessibility to radial highways | Quantitative | Relevant |
| tax | full-value property-tax rate per $10,000 | Quantitative | Relevant |
| ptratio | pupil-teacher ratio by town | Quantitative | Relevant |
| black | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town | Quantitative | Relevant |
| lstat | % lower status of the population | Quantitative | Relevant |
| medv | median value of owner-occupied homes in $1000's | Quantitative | Relevant |