**Ex-02-Outlier-Detection-and-Removal**

**AIM:**

To read the given data and perform data cleaning and save the cleaned data to a file.

**EXPLANATION:**

Outliers are extreme values that differ from most other data points in a dataset. They can have a big impact on your statistical analyses and skew the results of any hypothesis tests. It’s important to carefully identify potential outliers in your dataset and deal with them in an appropriate manner for accurate results.

# ALGORITHM:

## **STEP-1:**

Read the given data.

## **STEP-2:**

Get the information about the data.

## **STEP-3:**

Detect the outliers in the given datasets.

## **STEP-4:**

Remove the outliers from the given datasets by using outlier handling techniques.

## **STEP-5:**

Save the Cleaned data to the file.

# CODE:

# BHP dataset:

import pandas as pd

import numpy as np

import seaborn as sns

from google.colab import files

uploaded = files.upload()

df=pd.read\_csv("bhp.csv")

df

## **Remove outliers using IQR and Get a new Dataframe**:

q1=df['price\_per\_sqft'].quantile(0.25)

q3=df['price\_per\_sqft'].quantile(0.75)

IQR=q3-q1

print("First quantile:",q1," Third quantile:",q3," IQR: ",IQR,"\n")

lower=q1-1.5\*IQR

upper=q3+1.5\*IQR

outliers=df[(df['price\_per\_sqft']>=lower)&(df['price\_per\_sqft']<=upper)]

print("Outliers: \n")

print(outliers)

## **Using zscore of 3 to remove outliers:**

from scipy.stats import zscore

z=outliers[(zscore(outliers['price\_per\_sqft'])<3)]

print("Cleaned Data: \n")

print(z)

# Height\_Weight dataset:

import pandas as pd

import numpy as np

import seaborn as sns

from google.colab import files

uploaded = files.upload()

df=pd.read\_csv("height\_weight.csv")

df

## **Using IQR detect weight outliers and print them:**

q1=df['weight'].quantile(0.25)

q3=df['weight'].quantile(0.75)

IQR=q3-q1

print("First quantile:",q1," Third quantile:",q3," IQR: ",IQR,"\n")

lower=q1-1.5\*IQR

upper=q3+1.5\*IQR

outliers=df[(df['weight']>=lower)&(df['weight']<=upper)]

print("Outliers: \n")

print(outliers)

## **Using IQR, detect height outliers and print them:**

q1=df['height'].quantile(0.25)

q3=df['height'].quantile(0.75)

IQR=q3-q1

print("First quantile:",q1," Third quantile:",q3," IQR: ",IQR,"\n")

lower=q1-1.5\*IQR

upper=q3+1.5\*IQR

outliers=df[(df['height']>=lower)&(df['height']<=upper)]

print("Outliers: \n")

print(outliers)

# OUTPUT:

# 

# 

# 

# 

# 

# 

# RESULT:

Thus the outliers were detected and removed from the given datasets by using outlier handling techniques.