Name : Suvidhi V. Pareek

Class / Div : TE / 03

Roll No : 09

| Experiment No.4 |
| --- |
| Perform data pre-processing using Outlier Detection |
| Date of Performance: 29-07-2024 |
| Date of Submission: 05-08-2024 |

**AIM:** To implement a data pre-processing algorithm for Outlier Detection

**Objective:-** Develop a program to detect and handle outliers during the data pre-processing phase using the IQR method.

**Theory:**

Why preprocess the data?

In the real world, data is often messy, incomplete, and noisy. Preprocessing the data is crucial to handle these issues and improve the quality of the dataset before further analysis. One key issue in data is the presence of outliers, which are data points that deviate significantly from the majority of the dataset.

* What are Outliers?

Outliers are extreme values in the dataset that can skew results and affect the performance of data mining algorithms. They may arise due to data entry errors, equipment malfunctions, or natural variations in the data.

Outliers need to be identified and either removed or adjusted to improve the quality of the dataset. Outlier detection methods, such as the Interquartile Range (IQR) method, are designed to identify these extreme data points.

**Interquartile Range (IQR) Method:**

The Interquartile Range (IQR) method is a widely used technique for detecting outliers. It is based on the spread of the middle 50% of the data.

1. **Calculate Q1 (First Quartile)** – The 25th percentile of the data.
2. **Calculate Q3 (Third Quartile)** – The 75th percentile of the data.
3. **Calculate IQR** – The difference between Q3 and Q1:   
    IQR=Q3−Q1
4. **Define Outlier Boundaries**:
   * Lower Boundary: Q1−1.5×IQR
   * Upper Boundary: Q3+1.5×IQR

Any data point falling below the lower boundary or above the upper boundary is considered an outlier.

**Code and output:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = np.random.randn(100)

data = np.append(data, [8, 10, -8, -10])

df = pd.DataFrame(data, columns=['Value'])

Q1 = df['Value'].quantile(0.25)

Q3 = df['Value'].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

outliers = df[(df['Value'] < lower\_bound) | (df['Value'] > upper\_bound)]

plt.figure(figsize=(8, 6))

sns.boxplot(x=df['Value'])

plt.scatter(outliers.index, outliers['Value'], color='red', label='Outliers')

plt.title('Boxplot with Outlier Detection')

plt.xlabel('Value')

plt.legend()

plt.show()

print("Outliers detected:")

print(outliers)



# prompt: z score normalization

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from scipy import stats

!pip install -q matplotlib-venn

# Generating normally distributed data

data = np.random.randn(500)

# Calculate IQR

Q1 = np.percentile(data, 25)

Q3 = np.percentile(data, 75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

outliers = data[(data < lower\_bound) | (data > upper\_bound)]

plt.figure(figsize=(10, 6))

plt.hist(data, bins=30, edgecolor='black')

plt.scatter(outliers, np.zeros\_like(outliers), color='red', marker='o', label='Outliers')

plt.axvline(lower\_bound, color='green', linestyle='--', label='Lower Bound')

plt.axvline(upper\_bound, color='green', linestyle='--', label='Upper Bound')

plt.xlabel('Data')

plt.ylabel('Frequency')

plt.title('IQR Outlier Detection')

plt.legend()

plt.show()

data\_mean = np.mean(data)

data\_std = np.std(data)

normalized\_data = (data - data\_mean) / data\_std

plt.figure(figsize=(10, 6))

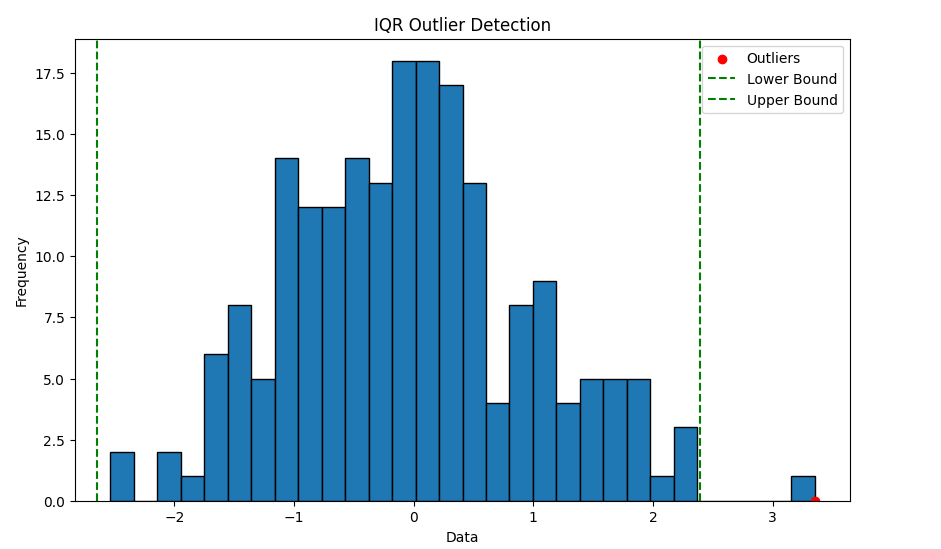
plt.hist(normalized\_data, bins=30, edgecolor='black')

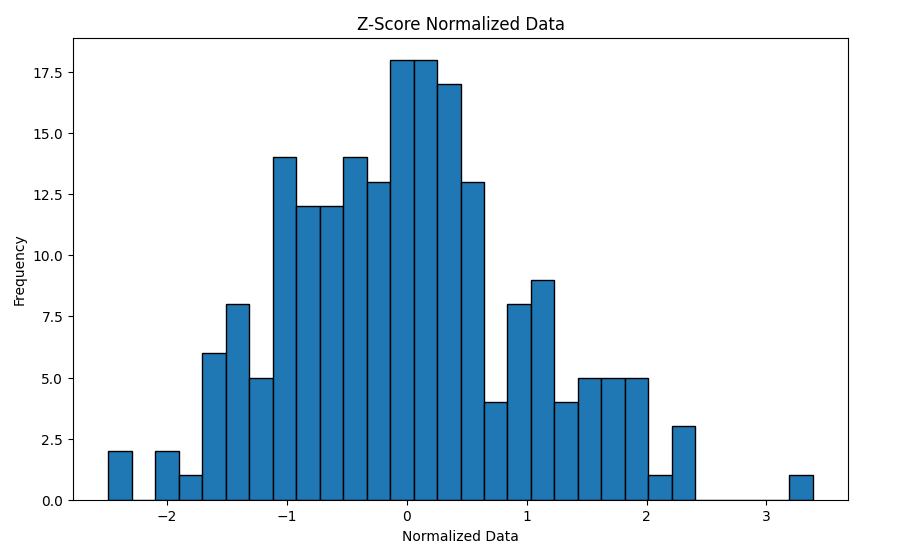
plt.xlabel('Normalized Data')

plt.ylabel('Frequency')

plt.title('Z-Score Normalized Data')

plt.show()





**Conclusion**: Comment on importance of pre-processing and what happens if the data is not pre-processed. Pre-processing is crucial for outlier detection as it cleans and normalizes data, and without it, unprocessed data can lead to inaccurate models and poor predictive performance.