**Name:** Ujwal Sahu

**Branch:**B.Voc-AIDS

**Experiment 3**

**Title:** Demonstrate Handling of Outliers and Missing Data using Python in Machine Learning.

**Tool:** Python libraries (e.g., Pandas, NumPy, Scikit-learn)

**Theory:** Outliers can skew results, and missing data can lead to inaccurate models; thus, proper handling is crucial for model performance.

- ## Outlier detection and Removal

Outliers are data points that deviate significantly from the rest of the dataset, potentially skewing results.

- ## Common methods for detecting outliers include:

**Interquartile Range (IQR):** Calculates Q1 and Q3 to determine the IQR, then identifies outliers as points outside the range defined by 1.5 * IQR.

**Z-Score:** Measures how many standard deviations a data point is from the mean; typically, a threshold of 2 or 3 is used to identify outliers.

**Visualization:** Boxplots and scatter plots can visually highlight outliers.

## Handling Missing Data:

Missing values can lead to biased results and reduced sample sizes.

Common strategies for handling missing data include:

**Imputation:** Filling missing values with mean, median, or mode.

**Forward/Backward Fill:** Using the last or next observed value to fill gaps.

**Interpolation:** Estimating missing values based on surrounding data points.Dropping Rows/Columns: Removing data points or features with excessive missing values.

**Code:**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import SimpleImputer

# Create a sample DataFrame
data = {
    'A': [1, 2, 3, 4, 5, 100],  # Outlier in column A
    'B': [5, np.nan, 7, 8, np.nan, 10],  # Missing values in column B
    'C': [10, 20, 30, 40, 50, 60]
}

df = pd.DataFrame(data)

# Display the original DataFrame
print("Original DataFrame:")
print(df)

# Visualize the data to identify outliers
plt.figure(figsize=(10, 5))
sns.boxplot(data=df)
plt.title("Boxplot to Identify Outliers")
plt.show()
```

```
# Handling Outliers using IQR
Q1 = df['A'].quantile(0.25)
Q3 = df['A'].quantile(0.75)
IQR = Q3 - Q1

# Define bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Remove outliers
df_no_outliers = df[(df['A'] >= lower_bound) & (df['A'] <= upper_bound)]

print("\nDataFrame after removing outliers in column A:")
print(df_no_outliers)

# Handling Missing Data
# Using SimpleImputer to fill missing values in column B with the mean
imputer = SimpleImputer(strategy='mean')
df_no_outliers['B'] = imputer.fit_transform(df_no_outliers[['B']])

print("\nDataFrame after imputing missing values in column B:")
print(df_no_outliers)

# Visualize the cleaned data
plt.figure(figsize=(10, 5))
sns.boxplot(data=df_no_outliers)
```

plt.title("Boxplot after Handling Outliers and Missing Data")

plt.show()

**Explanation of the Code:**

**Data Creation:** A sample DataFrame is created with some outliers and missing values.

**Visualization:** A boxplot is generated to visualize the presence of outliers in the dataset.

**Outlier Handling:**

The Interquartile Range (IQR) method is used to identify and remove outliers from column 'A'.

**Missing Data Handling:**

The Simple Imputer from Scikit-learn is used to fill missing values in column 'B' with the mean of the column.

Final Visualization: A boxplot is generated again to show the cleaned data after handling outliers and missing values.

Output:

- Original DataFrame:

```
   A    B   C
0   1  5.0  10
1   2  NaN  20
2   3  7.0  30
3   4  8.0  40
4   5  NaN  50
5 100 10.0  60
```

- DataFrame after removing outliers in column A:

```
   A   B   C
0  1   5.0   10
1  2   NaN   20
```

2 3 7.0 30

3 4 8.0 40

4 5 NaN 50

- DataFrame after imputing missing values in column B:

```
   A   B   C
0  1  5.0  10
1  2  7.5  20
2  3  7.0  30
3  4  8.0  40
4  5  7.5  50
```

**Conclusion:**

This code demonstrates how to effectively handle outliers and missing data in a dataset using Python, which is crucial for preparing data for machine learning

For Faculty Use

| Correction Parameters | Formative Assessment [40%] | Timely completion of Practical [ 40%] | Attendance / Learning Attitude [20%] | |
|---|---|---|---|---|
| Marks Obtained | | | | |