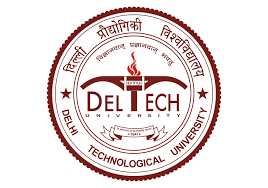
**DELHI TECHNOLOGICAL**

**UNIVERSITY**

****

**IT-106**

**PRACTICAL FILE**

**OPEN SOURCE**

**PROGRAMMING**

**Submitted By:** **Submitted To:**

Vaibhav Jain Ms. Lalita Luthra

23/IT/168 Department of Information & Technology

2nd Semester Delhi Technological University

Delhi Technological University

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **INDEX**   |  |  |  |  | | --- | --- | --- | --- | | **S.No** | **Experiment** | **Date** | **Sign** | | 1 | Familiarize with Python software (Fibonacci numbers, sorting a list of numbers) | 22-08-2024 |  | | 2 | Write a program to load a dataset from UCI repository into Python workspace and print its dimensions. Also load the target or class variable and print its dimensions. | 29-08-2024 |  | | 3 | Write a program to clean the data by removing noisy data or outliers and solving missing value problem. | 05-09-2024 |  | | 4 | Write a program to explore different data visualization techniques. | 12-09-2024 |  | | 5 | Write a program to perform statistical analysis of the data in a given dataset (mean, variance, standard deviation, median, mode). | 19-09-2024 |  | | 6 | Write a program to perform a classification experiment on a dataset and its target or class variable (Naïve Bayes, Random Forest). |  |  | | 7 | Write a program to perform a regression experiment on a dataset (linear regression). |  |  | | 8 | Write a program to perform a clustering experiment on a dataset (K-means, Hierarchical agglomerative clustering). |  |  | | 9 | Write a program to perform time series analysis for a given dataset. |  |  | | 10 | Write a program to perform association rule mining for a given dataset. |  |  | |

**PROGRAM 1**

**Objective:** To familiarize with Python software (Fibonacci numbers, sorting a list of numbers)

**Code:**

def fibonacci(n):

fib\_sequence = [0, 1]

while len(fib\_sequence) < n:

fib\_sequence.append(fib\_sequence[-1] + fib\_sequence[-2])

return fib\_sequence

n = 10

print(f"First {n} Fibonacci numbers: {fibonacci(n)}")

def bubble\_sort(arr):

n = len(arr)

for i in range(n):

for j in range(0, n-i-1):

if arr[j] > arr[j+1]:

arr[j], arr[j+1] = arr[j+1], arr[j]

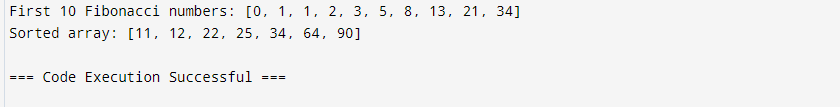
return arr

arr = [64, 34, 25, 12, 22, 11, 90]

sorted\_arr = bubble\_sort(arr)

print("Sorted array:", sorted\_arr)

**Output:**

****

**PROGRAM 2**

**Objective:** Write a program to load a dataset from UCI repository into Python workspace and print its dimensions. Also load the target or class variable and print its dimensions.

**Code:**

import pandas as pd

from ucimlrepo import fetch\_ucirepo

rice\_dataset = fetch\_ucirepo("Rice (Cammeo and Osmancik)")

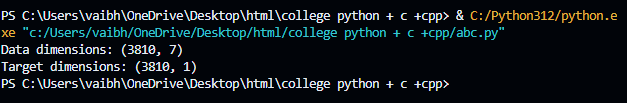
data = rice\_dataset.data.features

target = rice\_dataset.data.targets

print("Data dimensions:", data.shape)

print("Target dimensions:", target.shape)

**Output:**

****

**PROGRAM 3**

**Objective:** Write a program to clean the data by removing noisy data or outliers and solving missing value problem.

**Code:**

import pandas as pd

from ucimlrepo import fetch\_ucirepo

rice\_cammeo\_and\_osmancik = fetch\_ucirepo(id=545)

X = rice\_cammeo\_and\_osmancik.data.features

y = rice\_cammeo\_and\_osmancik.data.targets

df = X.copy()

df['Target'] = y

df=df.fillna(df.drop(columns=['Target']).median())

numeric\_columns = df.select\_dtypes(include=['float64', 'int64']).columns

def remove\_outliers\_iqr(df, numeric\_columns):

Q1 = df[numeric\_columns].quantile(0.25)

Q3 = df[numeric\_columns].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

df\_cleaned = df[~((df[numeric\_columns] < lower\_bound) | (df[numeric\_columns] > upper\_bound)).any(axis=1)]

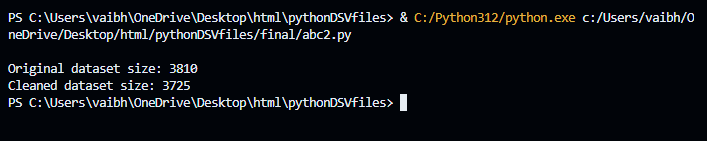
return df\_cleaned

df\_cleaned = remove\_outliers\_iqr(df, numeric\_columns)

print(f"\nOriginal dataset size: {df.shape[0]}")

print(f"Cleaned dataset size: {df\_cleaned.shape[0]}")

**Output:**

****

**PROGRAM 4**

**Objective:** Write a program to explore different data visualization techniques

**Code:**

import pandas as pd

from matplotlib import pyplot as plt

from ucimlrepo import fetch\_ucirepo

rice\_cammeo\_and\_osmancik = fetch\_ucirepo(id=545)

X = rice\_cammeo\_and\_osmancik.data.features

y = rice\_cammeo\_and\_osmancik.data.targets

df = X.copy()

df['Target'] = y

def pie\_chart(df):

features = ['Area', 'Perimeter', 'Major\_Axis\_Length']

for feature in features:

sums\_by\_target = df.groupby('Target').sum()

sums\_by\_target[feature].plot(kind='pie', fontsize=20)

plt.ylabel(feature, horizontalalignment='left')

plt.title('Breakdown for ' + feature, fontsize=25)

plt.savefig(f'rice\_pie\_for\_{feature}.jpg')

plt.close()

def bar\_chart(df):

sums\_by\_Target = df.groupby('Target').sum()

var = 'Area'

sums\_by\_Target[var].plot(kind='bar', fontsize=15, rot=30)

plt.title('Breakdown for ' + var, fontsize=20)

plt.savefig('rice\_bar\_for\_one\_variable.jpg')

plt.close()

sums\_by\_Target.plot(kind='bar', subplots=True, fontsize=12)

plt.suptitle('Total Measurements, by Target')

plt.savefig('rice\_bar\_for\_each\_variable.jpg')

plt.close()

def histogram(df):

df.drop(columns='Target').plot(kind='hist', subplots=True, layout=(3, 3), bins=20, figsize=(10, 8))

plt.suptitle('Rice Histograms', fontsize=20)

plt.tight\_layout(rect=[0, 0.03, 1, 0.95])

plt.show()

def mean\_mediam\_mode(df):

col = df['Area']

Average = col.mean()

Std = col.std()

Median = col.median()

Perc25 = col.quantile(0.25)

Perc75 = col.quantile(0.75)

Clean\_Avg = col[(col > Perc25) & (col < Perc75)].mean()

print(f"Average: {Average}")

print(f"Standard Deviation: {Std}")

print(f"Median: {Median}")

print(f"25th Percentile: {Perc25}")

print(f"75th Percentile: {Perc75}")

print(f"Clean Average (excluding outliers): {Clean\_Avg}")

**Output:**

**PROGRAM 5**

**Objective:** Write a program to perform statistical analysis of the data in a given dataset (mean, variance, standard deviation, median, mode).**Code:**

import pandas as pd

from ucimlrepo import fetch\_ucirepo

rice\_cammeo\_and\_osmancik = fetch\_ucirepo(id=545)

X = rice\_cammeo\_and\_osmancik.data.features

y = rice\_cammeo\_and\_osmancik.data.targets

df = pd.DataFrame(X)

df['Target'] = y

def statistical\_analysis(dataframe, feature):

analysis = {}

analysis['Mean'] = dataframe[feature].mean()

analysis['Variance'] = dataframe[feature].var()

analysis['Standard Deviation'] = dataframe[feature].std()

analysis['Median'] = dataframe[feature].median()

analysis['Mode'] = dataframe[feature].mode()[0]

return analysis

feature\_to\_analyze = 'Area'

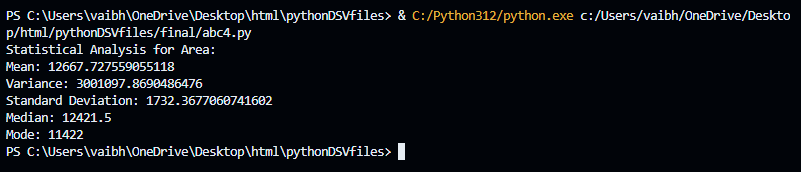
stats = statistical\_analysis(df, feature\_to\_analyze)

print(f'Statistical Analysis for {feature\_to\_analyze}:')

for stat, value in stats.items():

print(f'{stat}: {value}')

**Output:**

****