❖ DATA SCIENCE

EXPERIMENT NO.1

AIM: Write a python program to open Comma Separated Value (CSV) and perform given statistical operations.

THEORY:

What is CSV?

CSV stands for Comma Separated Value and it is a commonly used extension that has a specific format which allows data to store in a table in a structured format. In CSV each lines store a data or record, where each record consists of one or more columns or fields which is separated by commas. CSV also provide the feature to import and export of the spreadsheets and allows o store data that acts like a database.

What are categorical variables and types?

Categorical variables are also known as qualitative variable. It is one that has two or more categories, where the particular ordering of the categories may or may not be considered.

Examples of categorical variables are: race, sex, age group, and educational level.

There are 2 types of Categorical variables:

- Nominal variable:
  A categorical variable where the categories do not have a natural ordering. (e.g., gender, ethnicity, country).

- Ordinal variable:
  A categorical variable where the categories have a natural ordering. (e.g., age group, Income level, educational status).

What are numerical variables and types?

Numerical variables are known as Quantitative variable. A numerical variable is a data variable that takes on any value within a finite or infinite interval.

Examples of categorical variables are: length, test scores,

There are 2 types of numerical variables:

- Continuous variable:
  A numerical variable that can take values on a continuous scale.
  (e.g., age, weight).

- Discrete variable:
  A numerical variable that only takes on whole numbers.
  (e.g., number of visits)

PERFORM OPERATION USING PYTHON ON CSV FILE:

1. Identification of Categorical and Numerical Variables.

```
name=input('Enter name of Variable: ')
print('The entered variable is',name)

value=list(data_set[name])
result = []
for i in value:
    if i not in result:
        result.append(i)

if len(result)>15:
    print('The variable is numerical')
else:
    print('The variable is categorical')
```

2. Contingency table of at most 2 different categorical variables.

```
from prettytable import PrettyTable

Contigency_Table = PrettyTable(["Species","SepalLengthCm","PetalLengthCm"])
Contigency_Table.add_row(["Iris-setosa", "4.3-5.8", "1.0-1.9"])
Contigency_Table.add_row(["Iris-versicolor", "4.9-7.0", "3.0-5.1"])
Contigency_Table.add_row(["Iris-virginica", "4.9-7.9", "4.5-6.9"])

print("Contigency Table of Two Different Categorical Variables: \n",Contigency_Table)
```

3. Mean, Median, Mode, Variance, Standard Deviation, Quartile Range

```
import numpy as np
import pandas as pd
data_set = pd.read_csv("Data Set.csv")
```

```
data1 = list(data_set['SepalLengthCm'])
l_ds = len(data1)

# Mean
get_mean=sum(data1)/l_ds
print("Mean of SepalLengthCm: ", get_mean)

#Median
data1.sort()

if l_ds % 2 == 0:

value1 = data1[l_ds // 2 ]
    value2 = data1[l_ds // 2 - 1]
    Median = (value1+ value2) / 2
else:
    Median = data1[l_ds // 2]

print("Median is: ", Median)

# Mode
mode = max(data1, key= data1.count)
print("Mode is: ",mode)

# Variance
import statistics
Variance = statistics.variance(data1)
print("Variance is: ", Variance)

# Standard Deviation
SD = Variance ** 0.5
print("Standard Deviation is: ", SD)


# Quartile Range
Q1 = np.median(data1[:50])
Q2 = np.median(data1[50:])

IQR = Q2 - Q1          # Interquartile range (IQR)

print("Quartile Range: ",IQR)
```

4. Show categorical variables.

a. Show Binary data.

```
# Nominal Data
ND=list(data_set['Species'])
species = []

for i in ND:
    if i not in species:
        species.append(i)


print("Nominal Data is: ",species)
```

b. Show Binary data.

Binary Data doesn't exist in a data set.

c. Show Ordinal data.

Ordinal Data doesn't exist in a data set.

CONCLUSION:

Hence, in the assignment we have learned how to perform mathematical concepts like mean, median, mode, standard deviation, variance, etc. on the csv dataset and implement different function like how to identify Categorical and Numerical variable. Also, we learned how to implement contingency table.

Name: Rishab Jha
PRN: 20190802072