St. Francis Institute of Technology, Mumbai-400 103

**Department Of Information Technology**

A.Y. 2024-2025

Class: TE-ITA/B, Semester: VI

Subject: **Business Intelligence Lab**

**Experiment – 2: Exercise on Data Exploration**

1. **Aim:** Exercise on Data Exploration.
2. **Objectives:** After study of this experiment, the student will be able to

* Understand different types of attributes.

1. **Outcomes:** After study of this experiment, the student will be able to

**CO1:** Understand the importance of data mining along with identification of issues and technologies associated with it.

**CO2:** Organize and prepare the data needed for data mining using pre preprocessing techniques and Perform exploratory analysis of the data to be used for mining

1. **Prerequisite:** Introduction to different types of attributes.
2. **Requirements:** Personal Computer, Windows XP operating system, Internet Connection,

Microsoft Word, WEKA tool.

1. **Theory:**

* Types of attributes
* Define the following terms and give formulae
* Mean, Median, Mode, Variance, Standard deviation, Five number summary, Box plot, midrange, Quartile, Interquartile range

1. **Laboratory Exercise: Write Java / Python code for any problem and attach** Printout of code  along with output Snapshots
2. **Post-Experiments Exercise** 
   1. **Solve following problems:**
3. Suppose that value for given set of data are grouped into intervals. The

     intervals and corresponding frequencies are as follows:

| Age | Frequency |
| --- | --- |
| 1-5 | 200 |
| 6-15 | 450 |
| 16-20 | 300 |
| 21-50 | 1500 |
| 51-80 | 700 |
| 81-110 | 44 |

Compute an approx. median value for the data.

1. Suppose that data for analysis includes the attribute age. The age values for data tuples are(in increasing order):

13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70

1. What is mean of data? What is median of data?
2. What is mode of data? Comment on data's modality (bimodal/trimodal etc.)
3. What is mid range of data?(smallest+ largest value/2)
4. Can you find roughly the first quartile Q1, and the third quartile Q3 of the data.
5. Give the five point summary of the data.
6. Show a box plot of the data.
7. **Conclusion:**

1. Summary of Experiment

2. Importance of Experiment

3. Application of Experiment

**Reference:** Data Mining: Concept & Techniques, 3rd Edition, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier.

code :

import numpy as np

from scipy import stats

from collections import Counter

# Taking input from the user

data = list(map(int, input("Enter numbers separated by spaces: ").split()))

# Mean

mean\_value = np.mean(data)

# Median

median\_value = np.median(data)

# Mode using Counter to find all modes

freq\_count = Counter(data)

max\_count = max(freq\_count.values())

modes = [key for key, value in freq\_count.items() if value == max\_count]

# Modality detection

if len(modes) == 1:

modality = "Unimodal"

elif len(modes) == 2:

modality = "Bimodal"

else:

modality = "Trimodal or more"

# Mid-range

mid\_range = (min(data) + max(data)) / 2

# First Quartile (Q1) and Third Quartile (Q3)

q1 = np.percentile(data, 25)

q3 = np.percentile(data, 75)

# Five-point summary

min\_value = min(data)

max\_value = max(data)

five\_point\_summary = (min\_value, q1, median\_value, q3, max\_value)

# Printing results

print(f"Mean: {round(mean\_value, 2)}")

print(f"Median: {median\_value}")

print(f"Mode(s): {modes} (appears {max\_count} times each)")

print(f"Modality: {modality}")

print(f"Mid-range: {round(mid\_range, 2)}")

print(f"First Quartile (Q1): {round(q1, 2)}")

print(f"Third Quartile (Q3): {round(q3, 2)}")

print(f"Five-Point Summary: {five\_point\_summary}")

