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:**

**1. Types of attributes**

**ANS:** Attributes can be classified into different types based on their characteristics and usage. The main types include:

1. **Single-valued vs. Multi-valued Attributes**

* *Single-valued*: Holds a single value (e.g., Age).
* *Multi-valued*: Can have multiple values (e.g., Phone Numbers).

1. **Simple vs. Composite Attributes**

* *Simple (Atomic)*: Cannot be divided further (e.g., Employee ID).
* *Composite*: Can be broken down into smaller parts (e.g., Full Name → First Name, Last Name).

1. **Derived vs. Stored Attributes**

* *Derived*: Computed from other attributes (e.g., Age derived from Date of Birth).
* *Stored*: Stored directly in the database (e.g., Date of Birth).

1. **Key vs. Non-Key Attributes**

* *Key*: Used to uniquely identify records (e.g., Primary Key).
* *Non-Key*: Does not uniquely identify records (e.g., Address).

1. **Nominal, Ordinal, Discrete, and Continuous Attributes**

* *Nominal*: Categorical, without any order (e.g., Gender, Color).
* *Ordinal*: Categorical, but with an inherent order (e.g., Education Level).
* *Discrete*: Finite, countable values (e.g., Number of Employees).
* *Continuous*: Infinite values within a range (e.g., Temperature, Salary).

**2. Define the following terms and give formulae**

Mean, Median, Mode, Variance, Standard deviation, Five number summary, Box plot, Range, Quartile, Interquartile range

**ANS:**

### Mean: The mean, or average, is the sum of all values divided by the total number of values. It represents the central value of a dataset. The formula for the mean is:

### Median: The median is the middle value of an ordered dataset. If the number of observations is odd, the median is the middle value. If it is even, the median is the average of the two middle values. It is used to measure the central tendency, especially when dealing with skewed data.

### Mode: The mode is the most frequently occurring value in a dataset. A dataset may have one mode (unimodal), more than one mode (bimodal or multimodal), or no mode if all values occur with the same frequency.

### Variance: Variance measures the dispersion of data points from the mean. A higher variance indicates a wider spread of data. The formula for variance is:

### Standard Deviation: Standard deviation is the square root of variance and represents the average distance of each data point from the mean. It is useful for understanding how data deviates from the central tendency. The formula for standard deviation is:

### Five Number Summary: The five-number summary consists of the minimum, first quartile, median, third quartile , and maximum values of a dataset. It provides a quick overview of data distribution and is useful for identifying outliers.

### Box Plot: A box plot, or whisker plot, graphically represents the five-number summary. The box represents the interquartile range (IQR) with the median inside, while the whiskers extend to the minimum and maximum values, helping to visualize the data distribution and detect outliers.

### Range: Range is the difference between the maximum and minimum values of a dataset. It is a simple measure of spread and is calculated as:

### Range=Maximum−Minimum

1. **Quartiles:** Quartiles divide the dataset into four equal parts. The first quartile (Q1Q\_1Q1​) is the 25th percentile, the second quartile (Q2Q\_2Q2​) is the median (50th percentile), and the third quartile (Q3Q\_3Q3​) is the 75th percentile. Quartiles help in understanding data spread and identifying outliers.

### Interquartile Range (IQR): The interquartile range measures data spread by considering the middle 50% of values. It is the difference between the third quartile and the first quartile, given by the formula:

### 

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.

2.  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.

**Write Java / Python code for any problem and attach** Printout of code  along with output Snapshots

**CODE**:

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

# Given data

ages = [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]

# Mean

mean\_age = np.mean(ages)

# Median

median\_age = np.median(ages)

# Mode: Find all modes by counting frequencies

unique\_ages, counts = np.unique(ages, return\_counts=True)

max\_count = np.max(counts)

mode\_values = unique\_ages[counts == max\_count]

mode\_counts = counts[counts == max\_count]

# Modality

num\_modes = len(mode\_values)

modality = "Unimodal" if num\_modes == 1 else "Bimodal" if num\_modes == 2 else "Trimodal" if num\_modes == 3 else "Multimodal"

# Mid-range

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

# Quartiles

Q1 = np.percentile(ages, 25, interpolation='nearest')

Q3 = np.percentile(ages, 75)

# Five-point summary

five\_point\_summary = (min(ages), Q1, median\_age, Q3, max(ages))

# Box plot (horizontal orientation)

plt.figure(figsize=(6,4)) # Adjust the size to make it smaller

sns.boxplot(x=ages) # Note: 'x' for horizontal axis

plt.title("Box Plot of Age Data (Horizontal)")

plt.show()

# Print results

print(f"Mean: {mean\_age}")

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

# Print the modes with their counts

if len(mode\_values) > 1:

modes\_text = f"Modes: {', '.join(map(str, mode\_values))} (each appears {', '.join(map(str, mode\_counts))} times)"

else:

modes\_text = f"Mode: {mode\_values[0]} (appears {mode\_counts[0]} times)"

print(modes\_text)

print(f"Modality: {modality}")

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

print(f"First Quartile (Q1): {Q1}")

print(f"Third Quartile (Q3): {Q3}")

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

**OUTPUT:**

|  | Mean: 29.962962962962962  Median: 25.0  Modes: 25, 35 (each appears 4, 4 times)  Modality: Bimodal  Mid-range: 41.5  First Quartile (Q1): 20  Third Quartile (Q3): 35.0  Five-point Summary: (13, 20, 25.0, 35.0, 70) |
| --- | --- |