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.

3. **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

4. **Prerequisite:** Introduction to different types of attributes.

5. **Requirements:** Personal Computer, Windows XP operating system, Internet Connection, Microsoft Word, WEKA tool.

6. **Theory:**
   - Types of attributes
   - Define the following terms and give formulae
     - Mean, Median, Mode, Variance, Standard deviation, Five number summary, Box plot, Range, Quartile, Interquartile range

7. **Laboratory Exercise: Write Java / Python code for any problem and attach** Printout of code along with output Snapshots

8. **Post-Experiments Exercise**
   A. **Solve following problems:**
   1. 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

    i.   What is mean of data? What is median of data?

    i.   What is mode of data? Comment on data's modality (bimodal/trimodal etc.)

    i.   What is mid range of data?(smallest+ largest value/2)

    i.   Can you find roughly the first quartile Q1, and the third quartile Q3 of the data.

    i.   Give the five point summary of the data.

    ii.   Show a box plot of the data.

9. 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.

Importing Required Libraries

```
In [ ]:
```

```python
import statistics as st
import numpy as np
from tabulate import tabulate
import seaborn as sns
import matplotlib.pyplot as plt
```

Creating a Array

```
In [ ]:
```

```python
arr = np.array([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]
```

Statistical Measures

```
In [ ]:
```

```python
min= np.min(arr)
max= np.max(arr)
mean = np.mean(arr)
median = np.median(arr)
mode = st.multimode(arr)
range = max-min
variance = np.var(arr)
std_deviation = np.std(arr)
Q1 = np.percentile(arr,25)
Q2 = np.percentile(arr,50)
Q3 = np.percentile(arr,75)
IQR = Q3-Q1
five_number_summary = np.array([min, Q1, Q2, Q3, max])
```

```
In [ ]:
```

```python
all_data = [
    ['Minimum', min],
    ['Maximum', max],
    ['Mean', "{:.2f}".format(mean)],
    ['Median', "{:.2f}".format(median)],
    ['Mode', mode],
    ['Range', "{:.2f}".format(range)],
    ['Variance', "{:.2f}".format(variance)],
    ['Standard Deviation', "{:.2f}".format(std_deviation)],
    ['Q1', Q1],
    ['Q2', Q2],
    ['Q3', Q3],
    ['IQR', IQR],
    ['5-Point Summary', [min, Q1, Q2, Q3, max]]
]

print(tabulate(all_data, headers=['Measure', 'Result'], tablefmt='pretty',stralign="left"))
```

```
+--------------------+----------------------------+
| Measure            | Result                     |
+--------------------+----------------------------+
| Minimum            | 13                         |
| Maximum            | 70                         |
| Mean               | 29.96                      |
| Median             | 25.00                      |
| Mode               | [25, 35]                   |
| Range              | 57.00                      |
| Variance           | 161.29                     |
| Standard Deviation | 12.70                      |
| Q1                 | 20.5                       |
| Q2                 | 25.0                       |
| Q3                 | 35.0                       |
| IQR                | 14.5                       |
| 5-Point Summary    | [13, 20.5, 25.0, 35.0, 70] |
+--------------------+----------------------------+
```

Creating a Boxplot of the declared array

In [ ]:

```python
plt.figure(figsize=(8, 6))
sns.boxplot(data=arr, color='skyblue', width=0.5)
plt.title('Box Plot of Data', fontsize=16)
plt.ylabel('Value', fontsize=14)
plt.xlabel('Data Set', fontsize=14)
plt.grid(True, linestyle='--', alpha=0.7)
plt.show()
```