	Hope Foundation's **Finolex Academy of Management and Technology, Ratnagiri** **Department of Information Technology**		
Subject name	**Business Intelligence Lab**		**Subject Code: ITL602**
Class	**TE IT**	**Semester – VI (CBCGS)**	**Academic year: 2018-19 (FH 2019)**
Name of Student			**QUIZ Score :**
Roll No		**Assignment/Experiment No:**	**01**
Title:	**Tutorials on Data Exploration and Data pre-processing**		

1. Course objectives applicable: LO1, LO2

2. Course outcomes applicable: LO1, LO2

3. Learning Objectives:

1. To understand need of data preprocessing.
2. To understand use of MS Excel for statistical calculations.
3. To understand use of WEKA for preprocessing.

4. Practical applications of the assignment/experiment: Data analysis and pre processing

5. Prerequisites:

1. The basic statistics
2. Use of MS Excel

6. Hardware Requirements:

1. PC with minimum 2 GB RAM

7. Software Requirements:

1. MS Excel
2. WEKA

8. Viva Questions (if any): (Online Quiz will be taken separately batch-wise)

1. What is preprocessing?
2. Why the data can be dirty?
3. What is meant by data cleaning?
4. What is noise and outliers in the data and how can they affect data mining?

9. Experiment/Assignment Evaluation:

Sr. No.	Parameters	Marks obtained	Out of
1	Technical Understanding (Assessment may be done based on Q & A **or** any other relevant method.)		6
2	Neatness/presentation		2
3	Punctuality		2

Date of performance (DOP)		**Total marks obtained**		**10**
Date of checking (DOC)		**Signature of teacher**		

10. Theory: **<<handwritten work>>**

Why Data Preprocessing?

Data in the real world is dirty

- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=“ ”
- **noisy:** containing errors or outliers
  - e.g., Salary=“-10”
- **inconsistent:** containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data

Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

Q.1)
{3,4,6,5,7,9,8,1,3,2,9,4,7,9,6,8,5,9,2,6}

1. Calculate Mean, Mode, Median and std. deviation of the given data
2. Find the five number summary and Box plot diagram for the same.

Q.2)
{Y,N,Y,Y,N,Y,Y,N,Y,N,Y,N,N,Y,Y,Y,N,Y,N,Y,Y,Y,N,Y}

Calculate the Entropy of given data

11. Performance Steps:

Perform the data explorations and pre-processing with the given data collected via Google forms and shared as MS Excel files and find the correlations between different attributes

12. Results:

<< Add the hard-copy of output screen shots >>

13. Learning Outcomes Achieved

1. Understood the nature of data.
2. Understood the need of data pre-processing
3. Understood the use of MS EXCEL and WEKA in pre-processing

14. Conclusion:

1. **Applications of the studied technique in industry:** data analysis

2. **Engineering Relevance:** pre-processing for data mining.

3. **Skills Developed:** Understanding the MS EXCEL and WEKA operations for pre-processing.

15. References:

[1] Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3rd Edition.

[2] P. N. Tan, M. Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson Education.

[3] Michael Berry and Gordon Linoff, “Data Mining Techniques”, 2nd Edition Wiley Publications.

[4] https://en.wikipedia.org/wiki/Data_pre-processing