**DATA PREPROCESSING**

**ASSIGNMENT NO: 1**

**DATA MINING**

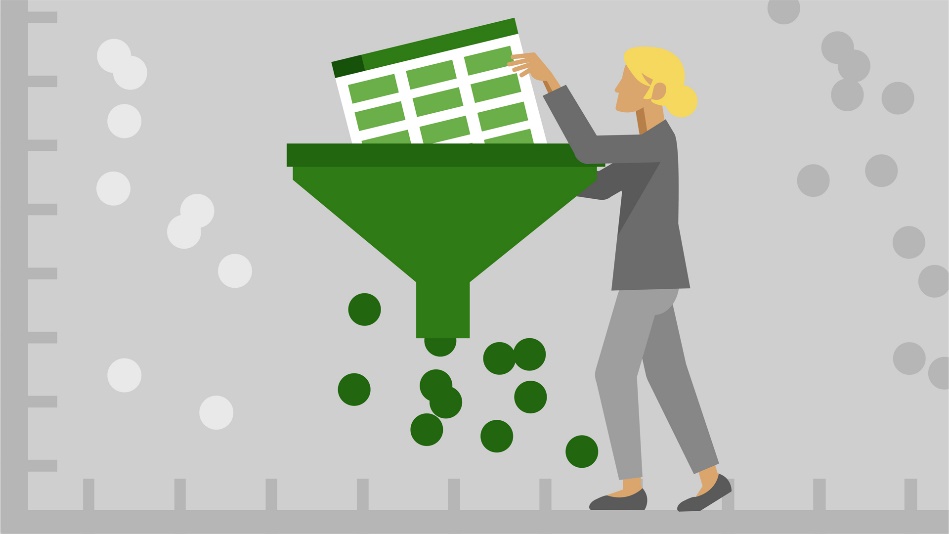
**MEMBERS NAME \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**MEMBER REGISTRATION NO \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**TITLE: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Please read the following and below instructions:

1. Please provide a complete solution.
2. Plagiarism or Cheating is not allowed.  **Even if you cheat in only one part, you will be marked zero in the *entire* assignment.**
3. Maximum member allowed in group is 5 and minimum 1
4. Please do this assignment by yourself and participation of each member is compulsory .As this is a research based assignment try to research and do this assignment for learning not just for marks. If you face any difficulty regarding understanding or doing assignment you may email me your problem.
5. viva and 1 quiz will be expected from this assignment
6. Submit your assignment on LMS session 4.
7. If you are new to python I have compile and attached a pdf file contain all python commands with their syntax for your help.



**QUESTION NO 1:**

Find text data set of your own choice from any free web resources (e.g. kaggle etc.) these repositories are covered in first lecture. All related links and sites are available.

After choosing data set analyze your data set (e.g. features, instances, type of data and their data types) after analyzing do preprocessing of that data set. Do this task using python language.

**1 Data cleaning:**

<https://realpython.com/python-data-cleaning-numpy-pandas/>

* Find any missing or null values are present
* Find outliers using matplotlib or use any data visualization technique

<https://www.geeksforgeeks.org/finding-the-outlier-points-from-matplotlib/>

* Fill those missing values

**2 Data integration**: if required

**3 Data Normalization [0 1]**

<https://www.journaldev.com/45109/normalize-data-in-python/>

<https://www.educative.io/edpresso/data-normalization-in-python/>

**4. Feature selection**

Chose feature selection techniques according to your data set and study any feature selection and understand its working .Given below link contain different feature selection techniques for reference. Chose and apply any feature selection technique on your data set.

<https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>

**DECLARATION:**

I am aware of and understand the University’s policy on plagiarism and I certify that this assignment is my own work**,**except where indicated by referencing, and that I have followed the good academic practices noted above

Signed \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_