PICT, PUNE

Name: Aarkin Karnik
Roll no: 31140

DSBDA Lab

Assignment - 2

- Title: Data Wrangling II

- Problem statement:
Create an "Academic Performance" dataset of students & perform the following operation using Python.

1) Find & solve inconsistencies & missing values
2) Find & deal with outliers
3) Apply data transformation on at least one of the variable.

- Learning objectives:
1) To learn & understand data wrangling in pandas
2) To deal with missing values / inconsistencies
3) To deal with outliers in the dataset
4) To learn & perform transformation methods.

- Learning Outcomes:
Student will be able to:
1) Perform handling of outliers
2) Perform data transformation for better understanding of variable.

- H/W & S/W Requirements:
Windows 10 64 bit, 8 GB RAM, 256 GB SSD, Intel i5 - 8300H processor, Jupyter Notebook, Python 3.9

PICT, PUNE

31140

- Theory:
An outlier is an observation in a given dataset that lies far the rest of observation

- Mean is accurate measure to describe data when we do not have outliers present.
- Median is used when outlier is present in dataset.
- Mode is used if there is outlier & greater than or equal to $\frac{1}{2}$ of data is same.

- Some techniques to detect outliers:
- Box Plot
- Z-score
- Inter Quantile Range

- Some techniques to treat the outliers:
- Trimming / Removing
- quantile based flooring in capping.
- Mean / Median interpretation

Normalization is a technique with the goal to change the values of numeric columns to a column scale without differences in the ranges of values or losing information.

- Conclusion:
Missing values, outliers detected & normalization applied.