Data Wrangling II
Outliers, Skewness

Problem Statement : Create an "Academic Performance" dataset of Students and perform the following operations using Python

1. Scan all variables for missing values and inconsistencies If there are missing values and or inconsistencies, use any of the suitable technique to deal with them
2. Scan all numeric variables for outliers. If there are outliers use any of the suitable techniques to deal with them
3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons : to change the scale for better understanding of the variable, to convert a non linear relation into a linear one or to decrease the skewness and convert the distribution into normal distribution

Introduction : Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. Data wrangling involves processing the data in various formats and analyze and get them to be used with another set of data and bringing them together into variable insights. Pandas is the main library which is used to perform data wrangling operation

Theory :

Outlier - In data analytics, outliers are values within a dataset that vary greatly from the others - they are either much larger or significantly smaller. Outliers may ~~include~~ indicate variabilites in a measurment, experimental errors or a novelty. There are two main reasons why giving outliers special attention is a neccessary aspect of the data analytics process.

1. Outliers may have a negative effective on the result of an analysis.
2. Outliers or their behavior - may be the information that a data analyst requires from the analysis.

Interquartile Range (IQR) - The interquartile range (IQR) formula is a measure of the middle 50% of the data set. The smallest of all the measures of dispersion in statistics is called the interquartile range. The difference between the upper and lower quartile is known as the interquartile range.

Interquartile range = Upper Quartile - Lower Quartile

$Q_2 = Q_3 - Q_1$

Q_1 is the median of the data points to the left of the median ~~for~~

Q_3 is the median of the data points to the right of the median