Problem statement -

create an" Academic performance" dataset of students and perform the following operations using python.

1. scan all variables for missing values and inconsistance. ~~If~~ there are missing values and /or inconsistencies, use any of the suitable techniques to deal with them.

2. scan all numeric variables for outlier. 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 variables to convert a non-linear relation into a linear one or to decrease the skewness and convert the distribution into a normal distribution.

Objective -

Extract data from a source, and convert it into a usable format, and deliver it to a destination.

Theory -

Evaluating for missing Data -

- The missing values are converted to Python's default.
- Python's built-in functions is used to identify these missing values.
- There are two methods to detect missing data.
  1) .isnull ()
  2) .not null ()
- The output is a boolean value indicating whether the value that is passed into the argument is in fact missing data.
- "True" stands for missing value, while "false" stand for not missing value.

Count missing values in each column
by using
- data . isnull (). sum()

Fit-transform () -

is used on the training data so that we can scale the training data and also learn the scalling parameters of that data.

The fit method is calculating the mean and variance of each of the features present in our data.

pipeline() -

The purpose of the pipeline() is to assemble several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a '__', as in the example.

- sklearn.pipeline.Pipeline(steps, *, memory = None, verbose = false)

Result -

In this way we studied about data transformation and convert the distribution into a normal distribution.