**BDA LAB ASSIGNMENT**

**Name - Umang Patel**

**Roll No – 2193271**

**Assignment 4: Preprocessing Using Python, Powerbi**

**Theory:**

Basic steps

Step 1 : Import the libraries

Step 2 : Import the data-set

Step 3 : Check out the missing values

Step 4 : See the Categorical Values

**Data cleaning:**

* The main aim of Data Cleaning is to identify and remove errors & duplicate data, in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making.
* Needless to say, data cleansing is a time-consuming process and most data scientists spend an enormous amount of time in enhancing the quality of the data. However, there are various methods to identify and classify data for data cleansing.
* There are mainly two distinct techniques, namely Qualitative and Quantitative techniques to classify data errors. Qualitative techniques involve rules, constraints, and patterns to identify errors.

**Normalisation**:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here’s the formula for normalization:



Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

• When the value of X is the minimum value in the column, the numerator will be 0, and hence X’ is 0

• On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the

denominator and thus the value of X’ is 1

• If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

**Standardisation:**

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:



u is the mean of the feature values and is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.









