## Data Quality Inspector Model

Model aims to detect the quality of a given dataset by using Python libraries.

1. **Introduction**

This notebook aims to detect the quality of a given text dataset. There are couple of metrics need to considered in terms of data quality. These are;

1. Completeness-This measures whether all the necessary data is present in a specific dataset. You can think about completeness in one of two ways: at the record level or at the attribute level. Measuring completeness at the attribute level is a little more complex however, as not all fields will be mandatory.
2. Accuracy-How accurately does your data reflect the real-world object? In the financial sector, data accuracy is usually black or white – it either is or isn’t accurate. That’s because the number of pounds and pennies in an account is a precise number. Data accuracy is critical in large organizations, where the penalties for failure are high.
3. Consistency-Maintaining synchronicity between different databases is essential. To ensure data remains consistent on a daily basis, software systems are often the answer.
4. Validity-Validity is a measure of how well data conforms to required value attributes. For example, ensuring dates conform to the same format, i.e., date/month/year or month/date/year.
5. Timeliness-Timeliness reflects the accuracy of data at a specific point in time. An example of this is when a customer moves to a new house, how timely are they in informing their bank of their new address? Few people do this immediately, so there will be a negative impact on the timeliness of their data.
6. Integrity-To ensure data integrity, it’s important to maintain all the data quality metrics we’ve mentioned above as your data moves between different systems. Typically, data stored in multiple systems breaks data integrity.

In this notebook, we only looked into missing values, duplicated values, multicolinearity and erroneous values for a given dataset in order to analyze the metrics mentioned above.

1. **Criteria Used for Data Quality**

High Quality Data Criteria

1. Overall Missing Value percentage less than %5 and,
2. Overall Duplicated Value percentage less than %2 and,
3. No Multicolinearity (Correlation between columns NOT higher than %90) and,
4. Overall Erroneous Data percentage is less than %2.
5. **Technologies Used**

Python Libraries

1. **Ydata-Quality**- ydata\_quality is an open-source **python** library for assessing **Data Quality** throughout the multiple stages of a dataset.
2. **Klib-** **klib** is a Python library for importing, cleaning, analyzing and preprocessing data.
3. **Illustration**

The model evaluates the high quality data criteria and gives whether the given data set is high quality or low quality. Model returns the result as follows for a given dataset.

