**A Project Report on**

# Data Management

## Under supervision of Prof. Dr. Ajinkya Prabhune,

Professor, SRH university

## Prof. Dr. Barbara Sprick

Professor, SRH university


## Submitted by

**Medisetti Narendra, Matrikel Number: 11012094**

M.Sc in Big data & Business Analytics, SRH University, Heidelberg

## GOAL – B

**Title:** Student's Performance

**Source:** Kaggle.com

### Description:

This dataset tells student's Performance in academic.

### Features:

| Features Name | Data Type | Description |
|---|---|---|
| Gender | String | Student's Gender |
| race/ethnicity | String | Race |
| parental level of education | String | Parent's Education Qualification |
| Lunch | String | Type of Lunch |
| test preparation course | String | Status of test preparation course |
| math score | Integer | Math score gained by student |
| reading score | Integer | Reading score gained by student |
| Writing Score | Integer | Writing score gained by student |

### Tools required to perform uncleaning data: Talend Data Preparation

### Step:

1. **Clearing cell on matching values:**
    - This function filters all values in parental level of education, math score, reading score, Writing Score features.
    - After completion of filtering these records, select columns which need to be changed, selected cells are made empty values by using a function called **fill cells with value**. By this we can say that missing data have removed **"completeness dimension"** of the dataset because mandatory columns like student score are empty.

2. **Changed the case of certain cell entries in columns:**
    - First, analyse the dataset which columns has strings or characters. Now consider to be so after looking at many rows in Student's Performance dataset.

. I filter some cells value in parental level of education, lunch columns. In this case the string needs to changed using a function "change to upper case". Remaining columns attributes are in lower case.

. This ensures that that the dataset lack of "consistency dimension" because entities in columns are in both upper case and lower case.

3. **Creating negative value:**
    - First, analyse the dataset for which columns has integers. Now consider, after looking at many rows in Student's Performance dataset.
    - select a column which has integer are math score, reading score columns. I filter some cells in these columns. By using function Negate the positive integer which we are filter are changed into negative values.
    - This ensures that that the dataset lack of "validity dimension" because entities in columns are in both positive and negative entities.

4. **Swapping:**
    - By looking Student's Performance dataset. I filter some cells in Lunch, race/ethnicity, test preparation course fields. Now select these columns and swaps the filter cells by using a function called swap columns. Some cells were adding a junk value in parental level of education, reading score by using a function called Add Extra characters.
    - This illustrate that that the dataset lack of Accuracy dimension because columns has swapped values and junk values.

5. **Search and Replace:**

First, analyse the dataset for which columns has integers. Now consider, after looking at many rows in Student's Performance dataset. Check the schema of the column it is integer. In the writing score which has about integer and string. This illustrate that that the dataset lack of Validity dimension.

## 6. Remove Uniqueness

- Since it is not possible to duplicate rows in Talend, once the dataset has been uncleaned and exported, the dataset can be opened in Excel and duplicate a few records so that it violates the UNIQUENESS dimension in the dataset.

## 7. Check for Currency

- The dataset about the Student's Performance contain list of marks score obtained does not meet the requirements to fulfil the CURRENCY dimension.