4th March 2021

**Data Mining Assessment**

**Group Members**

Ivy Odametey

André Afful

Kingsley Abru

Otema Yirenkyi

Alfred Ametepey

**Data Quality Assessment Project**

The aim of this project is to help appreciate the principles of data quality.

It is divided into these 3 tasks:

* Data profiling
* Defining quality rules
* Creating quality checks

In tackling this assignment, the first thing the team did was to discuss how well to present the deliverables in the expected time frame.

Members of the team set up MSSQL for analysis of the .bak file. This file contains the data which will profiled, and checked for its quality.

Since the data columns were provided in Spanish and not English, the language the team is conversant with, the team was furnished with a Data dictionary. This dictionary had the English translation of the column names to allow the members appreciate the data better.

The first task was Data Profiling. Here, the team worked together to write queries which check the quality of data in MSSQL.

The written queries answered the following questions:

* How many columns are present? What are their data types? Number of values?
* Are they missing data?

For better visualization of the output, the team created a dashboard with Microsoft PowerBI to present the data in a more readable form with charts and tree maps.

In the end, the team presented an SQL script that profiles the data and a screenshot of the dashboard that displays information about the profile of the data.

The next step for the team, after the output of the Profiling task has been approved, is to **Define Data Quality Rules**.

Here, the team will draft data quality rules that will govern how the data should be handled. To demonstrate this, the team will present a documentation on how quality of this data should be measured.

The team has slated a week for the completion of the second task.

The third and final task will be the **Creation of Quality Checks**.

After the delivered output of task two has been approved, the team will generate an SQL script that logs problems identified into an audit\_log table. This is slated to be completed in a week.