**Due: 21 November 2024**

**Brief for Level 5 Data Engineer - Formative Piece 1**

**"Data quality and performance in action"**

**Word count:** 500 words (+/- 10%)

**Objective:** *Demonstrate understanding of data quality and performance in action by applying core concepts and practical skills learned so far. This formative assessment aims to showcase your ability to assess and document data quality improvements and create a database schema using SQL as a data definition language (DDL).*

**Deliverables:***Submit* ***two items of evidence****to support your work, including* ***screenshots****and brief explanations for each. Submit via the Hub by the deadline at the top of the document.*

**Task Breakdown:**

1. **Data Quality Assessment and Improvement:**
   * Select an **Excel spreadsheet** or a small dataset that may contain potential quality issues.
   * Perform a **data quality assessment**, documenting **all data quality dimensions** (such as accuracy, completeness, consistency, timeliness, and uniqueness). Even if you focus on improving only one or two dimensions, still provide a brief assessment of each.
   * Where possible, apply **metrics** (e.g., percentage of missing values, average data entry lag) to quantify the current state of each dimension.
   * Focus on improving at least one or two dimensions, documenting your **assumptions** and any challenges encountered.
   * Apply necessary **transformations or data cleaning** steps (e.g., merging cells, handling missing data).
   * Include **Screenshot(s)** of the dataset before and after improvements, with annotations or brief descriptions explaining each step and the metrics used.
2. **Database Schema Design with SQL as a DDL:**
   * Using **SQL as a DDL** (Data Definition Language), create a **schema** based on the cleaned data, incorporating your knowledge of star schemas where possible.
   * Define **tables, fields, data types, and relationships** using relevant SQL commands.
   * Include your own **field explainers** to document each field’s purpose, assumptions (such as the range of values), and data type. Provide sample values.
   * Provide a sample of the **SQL (DDL) commands** used to create the schema in an SQL environment, such as **sqlfiddle.com or an equivalent sandbox environment**.
   * Include **Screenshot(s)** of your schema creation in SQL, along with the DDL statements used to define tables and relationships. Include brief explanations for each element, explaining why this schema structure supports performance and quality.

**Key Points:**

* Document all data quality dimensions and support assessments with metrics where possible.
* Demonstrate critical thinking by documenting assumptions and quality improvements.
* Structure explanations clearly to show your understanding of **data quality dimensions** and **schema design principles**.

**Submission Requirements:**

* Submit your work in a single PDF or Word document.
* Caption and number each screenshot (For example: Fig. 1 - My SQL commands) and then refer to your numbered screenshots to explain the actions taken (for example "As seen in Fig. 1, I have...").
* Save this document both in your learning journal and on the Hub.