Firmable - Assessment

Prepared by: **Antony Rohan R**

1. Overview

This assessment focuses on evaluating and improving the quality of the “news events” dataset using Python, SQL, and Tableau.
The goal was to identify and remediate data quality issues and to visualize improvements between the raw and cleaned datasets.

2. SQL Data Setup and Cleaning

SQL scripts were created to design the database structure, store cleaned data, and track data quality metrics.
The following main tables were defined:

Table Name	Purpose
news_events_cleaned	Stores the cleaned and standardized event data with metadata.
dq_metrics_run	Captures metrics such as duplicates removed, invalid URLs, and missing counts before and after cleaning.

3. Python (Jupyter Notebook) Data Cleaning

The Jupyter Notebook (Firmable_assesment.ipynb) handled profiling, cleaning, and validation.
The key tasks included:

- Profiling the dataset for missing values, duplicates, and invalid URLs.
- Standardizing fields (company names, URLs, phone numbers).
- Handling nulls and data type conversions.
- Removing duplicate entries and invalid records.
- Exporting the cleaned dataset for visualization and quality comparison.

4. Tableau Visualization

In Tableau, a bar chart was created to compare data quality metrics between the Raw Data and Cleaned Data versions.
The following indicators were visualized for both datasets:

- Total count
- Article URL missing flag
- Company name missing flag
- Duplicate flag
- Invalid URL flag
- Product missing flag

This visualization provided a clear view of improvements across various quality dimensions, highlighting the impact of the cleaning process.

5. Key Observations

- Significant reduction in missing and invalid values after cleaning.
- Duplicate records were successfully identified and removed.
- All key columns standardized and validated.
- Cleaned dataset is ready for downstream analytics and reporting.

6. Conclusion

The data quality assessment successfully demonstrated an end-to-end cleaning and validation workflow.
By combining SQL for structure, Python for automation, and Tableau for visualization,
the overall quality and reliability of the *news events* dataset were significantly improved.