INDIVIDUAL SUBMISSION

Name: Christabel Danquah

Course: DATA QUALITY

The purpose of this report is to document the data quality checks and corrections made on an Excel database that was imported into an SQL database. The checks aimed to identify and address issues related to accuracy, completeness, consistency, reliability, and timeliness in the dataset. The report outlines the identified issues, the solutions, the SQL queries used for data quality checks, as well as views and procedures to enhance data visualization.

Process:

1. Data Import and Cleaning:
   The Excel database was imported into an SQL database system. The database was cleaned in Excel using formulas and GUI before being imported into SQL.

2. Data Quality Checks:
   SQL queries were developed to perform data quality checks on all tables within the database. These checks included validation of numeric values, identification of duplicates, and assessment of data completeness in non-nullable columns.

3. ZScore Calculation and Comparison:
   Stored procedures were created to calculate and compare imported ZScores results to check for accuracy.

4. Creation of Views
   Views were generated to provide additional insights into artist performances and venue outliers.

5. Analysis:
   The results of the data quality checks, and analysis were evaluated to identify areas requiring improvement and corrections in relation to accuracy, completeness, consistency, reliability, and timeliness.

6. Normal Forms:
   How first, second and third normal forms were applied when creating the database.

ABOUT THE DATABASE

I chose to make a database about a Music Festival Management System. With the help of ChatGPT, I created dummy data for the database and had eight tables namely:

- Artists
- Venues
- Tickets
- Concerts
- Sponsors
- Attendees
- Performances
- Merchandise

ISSUES IDENTIFIED IN EXCEL

Accuracy:

- In the ARTISTS TABLE, the Genre and Country columns display inconsistent capitalization.
- In the VENUES TABLE, there are inconsistencies in the Location values.
- In the SPONSORS TABLE, SponsorLevel values vary in capitalization.
- In the TICKETS TABLE, the Availability column contains the entry "thirty" which is inaccurate.
- In the SPONSORS TABLE, the NumEventsSponsored column contains the entry "two", which is inaccurate.
- In the CONCERTS TABLE, the Date column contains a string value of the date instead of in the format yyyy-mm-dd.
- The StartTime and EndTime columns in the PERFORMANCES TABLE contains redundant data since the dates can be found in another table.

Completeness:

- The VENUES TABLE has missing entries for the Capacity of "Domo Arena".

Consistency:

- The Country column in the ARTISTS TABLE had inconsistent capitalization.
- The CheckInStatus values in the ATTENDEES TABLE show inconsistencies.
- In the SPONSORS TABLE, the NumEventsSponsored column contains the entry "two", which is inaccurate.
- Some email addresses in the ATTENDEES TABLE contain unwanted characters affecting the consistency of the column.

Reliability:

- Inconsistent data formats could affect reliability, such as missing or corrupted values in the VENUES TABLE.
- Inaccurate entries, like the "thirty" entry in the TICKETS TABLE, reduces the data's reliability.
- Inconsistent SponsorLevel values in the SPONSORS TABLE affects the data's reliability.

SOLUTIONS FOR ISSUES FOUND

Accuracy:

- Standardize capitalization in Country column of the ARTISTS TABLE.
- Standardize Location values in the VENUES TABLE.
- Standardize SponsorLevel capitalization in the SPONSORS TABLE.
- Remove unnecessary spaces in the columns using TRIM function.
- Standardise date format in Date column of the CONCERTS TABLE.
- Extract the time from the values in the StartTime and EndTime columns of the PERFORMANCES TABLE.

Completeness:

- Provide missing entries for Location in the VENUES TABLE or if it is a nullable column leave it empty.

Consistency:

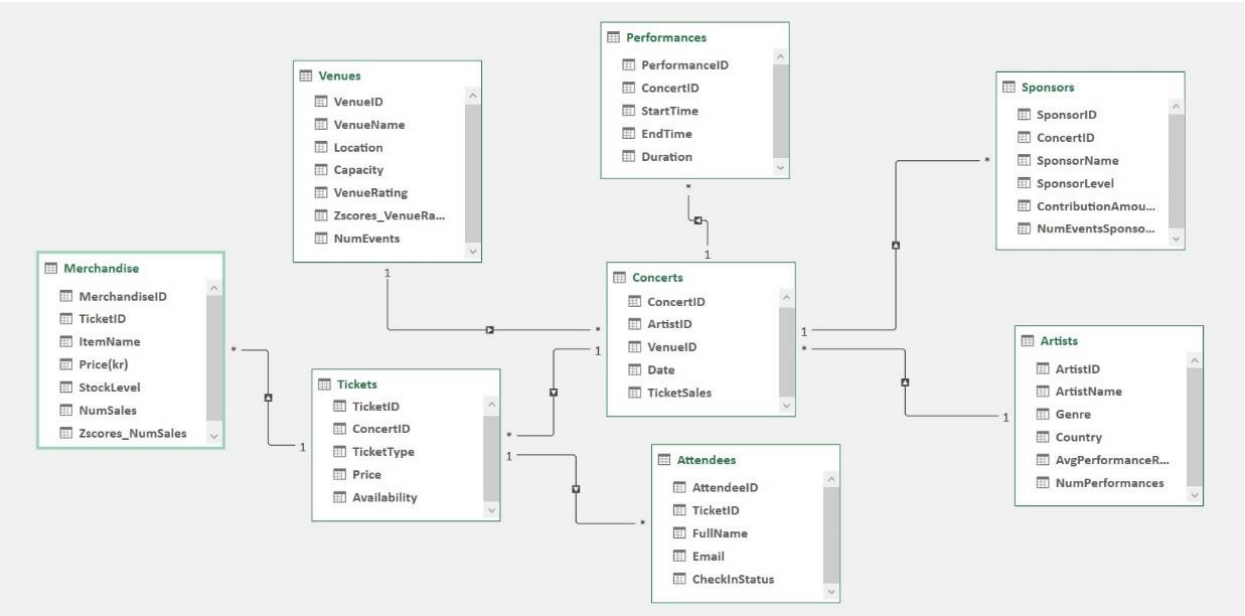
- Standardize capitalization in the STAGES TABLE.
- Standardize values in the CheckInStatus column of the ATTENDEES TABLE.
- Remove unwanted characters in the Email column in the ATTENDEES TABLE.
- Standardize the NumEventsSponsored in the SPONSORS TABLE.

Reliability:

- Ensure consistent data formats and correct inaccuracies to enhance data reliability.

Timeliness:

- The database is timely since the dates for concerts to be held are from 2024 to 2025.



ERD Diagram in Excel after all corrections have been made.

SQL REPORT

1. **Accuracy:**

- The SQL queries checks for duplicates and if numeric columns contains string values.

2. **Completeness:**

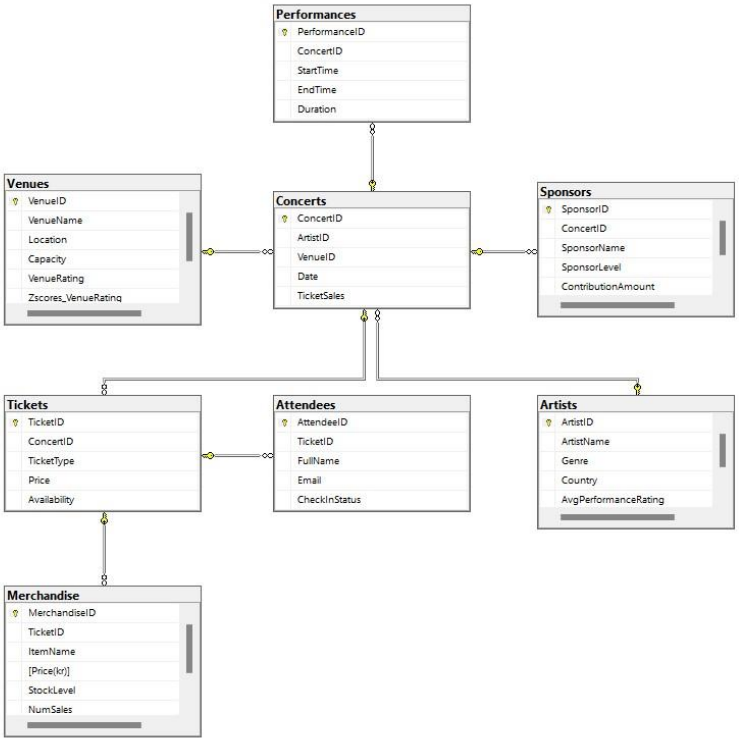
- The SQL queries checks if non nullable columns contains null values.

3. **Consistency:**

- The SQL queries checks if columns have standardized formats, for example if the date column contains date values and if a column containing emails follow a specific format.

4. **Reliability:**

- The database is reliable since it is accurate, complete, and consistent.



ERD in SQL after creating relationships.

DATABASE NORMALIZATION

1. **First Normal Form (1NF):**

    - Each data in every column had single values.
    - Every column contained data of the same data type.
    - Every table had a primary key.
    - Every table had unique names.

2. **Second Normal Form (2NF):**

    - Meeting the requirements of 1NF.
    - No partial dependencies in any of the tables.

3. **Third Normal Form (3NF):**

    - Meeting the requirements of 2NF.
    - There are no transitive dependencies, meaning there are no columns that are not fully dependent on the primary key.

In conclusion, this report documents the process of creating, checking, cleaning, and performing data quality checks on an Excel database that was later imported into an SQL database. Various issues related to accuracy, completeness, consistency, reliability, and timeliness were identified and addressed through solutions such as standardization of data formats and removal of inconsistencies or redundant data.

The SQL queries developed for data quality checks proved effective in identifying and addressing the accuracy, completeness, consistency, and reliability of the database. Stored procedures and views were used to calculate ZScores, compare the ZScore imported from Excel and the ZScore calculated in SQL, and provide additional insights into artist performances and venue outliers.

Furthermore, the database was normalized up to the third normal form (3NF), ensuring data integrity and minimizing redundancy.

By addressing the identified issues and ensuring adherence to normalization principles, the database has been optimized for reliability and usability, laying a solid foundation for future data-driven initiatives within the Music Festival Management System.

REFERENCES:

1. PowerPoints by Sir Richard Chalk
2. Relational Database Design by Ben Brumm on Udemy
3. https://youtu.be/_jmiEGZ6PIY?si=t_5NSV8H9Noxgx3z
4. Data Management Masterclass by George Smarts on Udemy
5. https://youtu.be/EwmuaqnoaKs?si=VSlMM7jjdtQgju2k
6. ChatGPT