****  
  
**Quality Control Form**

| **Team Name** | ORDER\_BY |
| --- | --- |
| **Team Members** | Hodan, Svetlana, Graciela, Sian |
| **Client Name** | Mystic Manuscripts |
| **Date of file received** | 09/06/2025 |
| **Name of file received** | Books Sales Records – V2 |
| **File Format received** | csv |
| **Size of file (KB) received** | 946KB |
| **Encoding of the file** | UTF-8 |
| **Recorded Number of Columns** | 36 |
| **Recorded Number of Rows** | 3,480 |
| **Name of Schema of Destination Table** | **Mystic\_manuscript** |
| **Name of Destination Table** | **Order\_by\_team** |

| **ID of SQL or Checked Rows** | **Description of Check** | **Result in SOURCE** | **Result in DESTINATION** | **OUTCOME** |
| --- | --- | --- | --- | --- |
| *DD-P1-1 (Example Row – DELETE)* | *Count of rows* | *5* | *5* | *PASS* |
| *DD-P1-2 (Example Row – DELETE)* | *Count of columns* | *5* | *6* | *FAIL* |
| SQL-1 | **Count of rows** | 3,480 | 3,480 | *PASS* |
| SQL-2 | **Count of Distinct Rows** | 3,480 | 3,480 | *PASS* |
| SQL-3 | **Count of Columns** | 36 | 36 | *PASS* |
| SQL-4 | **Sum of column Sums** | 11680962.79 | 11676304.79 | *FAIL* |
| SQL-5 | **Sum of Row Sums** | 11680962.79 | 11676304.79 | *FAIL* |
| SQL-6 | **Date format Check**  *Compare 5 randomly chosen values from Source, check date matches in Destination.* | | 9/1/2019 | | --- | | 8/25/2019 | | 4/1/2019 | | 1/11/2019 | | 6/2/2018 | | | 9/1/2019 | | --- | | 8/25/2019 | | 4/1/2019 | | 1/11/2019 | | 6/2/2018 | | *PASS* |
| SQL-7 | **Eyeball Check**  *Inspect 5 randomly chosen records from Source* | | 2246 | | --- | | 214 | | 2735 | | 2671 | | 3269 | | | 2246 | | --- | | 214 | | 2735 | | 2671 | | 3269 | | *PASS* |
| SQL-8 | **Max String Length**  *Check for max string length in product name column* | 154 | 154 | *PASS* |
| SQL-9 | **Null count in 5 random columns**  *Check nulls in cost price column* | Date 0  Date\_1 0  Quantity 0  Time 0  Item Price 0 | 0 Nulls in whole document | *PASS* |
| SQL-10 | **Min Value of Random Column**  *Check the minimum value of a random column (between 7 column)* | | col\_name | min\_v | | --- | --- | | Profit (INR) | 5.2 | | 5.2 | *PASS* |
| SQL-11 | **Max Value of Random Column**  *Check the maximum value of a random column((between 7 column)* | | col\_name | max\_v | | --- | --- | | Profit Percentage (%) | 18 | | 18 | *PASS* |

**Notes about discoveries and quality checks**

When carrying out query six we found the format of entries in the ‘Date’ column to differ between source and destination. On Excel, the format appeared as ‘DD/MM/YYYY’, whereas on SQL the format was ‘YYYY-MM-DD’. We were able to resolve this quickly using ‘to\_char()’, so that the column matched exactly across source and destination. We were aware however, that this change was not permanent, as it applied only to the query and not the actual table. If the date format were to be altered in the actual table itself, further work would be needed.

We also struggled with issues relating to column names. In the source, many column names involved capitals and spaces, both of which are to be avoided for best practice in SQL. Additionally, the Boolean columns did not map correctly on the table. We tried to solve this using multiple methods however due to time constraints were unable to solve the issue.

A further discovery made involved the repetition of information. Multiple columns contained the same data, or combined other columns. For example, ‘Purchase-date’ was a combination of ‘Date’ and ‘Time’, and ‘Quarter’ was a combination of ‘Qtr’ and ‘Year’. There were two columns, ‘Gender’ and ‘Gender F=1665 M=1815’, to identify gender. If we were to go about further cleaning of the data, information such as this would likely be refined, through the removal or combining of repeated columns, to save future computational cost and improve data quality.

During quality checks we spotted potential inaccuracies with regards to ‘Author’ and ‘Publication’. For some rows, these two columns contained the same information. Notably these were ID (871, 879, 895, 935, 943, 959) which had ‘MTG PUBLICATION’ in both columns, and ID 2651 which has ‘Collins’ in both. We questioned from this whether there was missing ‘Author’ information, which had then been filled in to avoid missing data. There were instances of author names being repeated in different formats. One in particular appeared as both ‘p r yadav’ and ‘PR Yadav’. This would lead to incorrect counts for distinct author names, and would need to be fixed in data cleaning.

Through carrying out various queries we were able to verify information. The count of rows, distinct rows, and columns, matched between source and destination, confirming that the data shape remained the same. An eyeball check of five random rows ensured that data was the same. We counted nulls in five random columns and found these to be the same. We also checked for maximum string length in the ‘Product-Name’ column, and these matched in source and destination. This verified that strings were not truncated. The final two extra queries involved finding minimum and maximum values in two random columns. By ensuring these values matched, we could make sure that the same result was obtained.