# DAT 325 Project Two

# Executive Summary Report

## Data Set Anomalies

| **Key Value** | **Description of Anomaly** | **Plan for Resolution** |
| --- | --- | --- |
| -43.19 | Negative value in Unit Price | Verify with source material to confirm if error, correct or remove. |
| -9999 | High value in COGS | Investigate using source material. Correct or remove if truly an error. |
| Empty Dates | Missing date | Find and correct. |
| Outlier high value in Quality | Value higher than other entries | Verify with source data, confirm, and note if valid, if not correct or exclude. |
| 4.1 Rating | Low outlier rating in comparison to others | Check data source and confirm if valid – correct or removed. |
| Bitcoin in Payment | Non listed payment method | Correct to standard Cash, Credit Card or Ewallet. |

## Data Types

| **Header Name From File** | **Data Types Note** |
| --- | --- |
| InvoiceID | varchar(25) - Unique ID of the order (Key Value), Not Null |
| Location | varchar(1) - See Lookup Table lkupLocation, Not Null |
| CustomerType | int - 0=normal, 1=member, Not Null |
| Gender | int - 0=Male, 1=Female, 2=N/A, Not Null |
| ProductLine | varchar(50) - The product line description, Not Null |
| UnitPrice | decimal - Cost of 1 unit of product, Not Null |
| Quantity | int - Number of items sold, Not Null |
| TaxApplied | decimal - Tax for the item = Quantity(.05UnitPrice), Not Null |
| TotalOrder | decimal - Total order amount = (Quantity \* Unit Price)+TaxApplied, Not Null |
| DateofPurchase | date - Date the item was purchased, Not Null |
| TimeofDay | int - 0 - 7am to 11am, 1- 11:01am - 3pm, 2 - for all other, Not Null |
| PaymentType | varchar(10) - Payment method used - valid values are Ewallet, Cash, CC, N/A for missing values, Not Null |
| CostofGoodsSold | decimal - Cost of the product to the company, Not Null |
| GrossMarginPercentage | decimal - Store as Decimal Percentage (percent/100), Not Null |
| GrossIncome | decimal - Gross Margin Percent \* Total, Not Null |

## Specific Transformations Needed to Join the Data

| **Header Name From File** | **Excel Function One** | **Excel Function Two** | **Excel Function Three** |
| --- | --- | --- | --- |
| InvoiceID | =TEXT(A2, "0") |  |  |
| Location | =IF(B2="Yangon", "A", IF(B2="Mandalay", "B", "C")) |  |  |
| CustomerType | =IF(C2="Normal", 0, 1) |  |  |
| Gender | =IF(D2="Male", 0, 1) | =IF(D2="N/A", 2, IF(D2="Male", 0, 1)) |  |
| ProductLine | =TEXT(E2, "0") |  |  |
| UnitPrice | =IF(F2<0, ABS(F2), F2) | =IF(ISNUMBER(F2), F2, "") |  |
| Quantity | =VALUE(G2) | =IF(G2>=0, G2, "") | =IF(ISNUMBER(G2), G2, "") |
| TaxApplied | =G2*0.05*F2 | =IF(G2*0.05*F2>=0, G2*0.05*F2, "") |  |
| TotalOrder | =(G2\*F2)+H2 | =IF((G2*F2)+H2>=0, (G2*F2)+H2, "") |  |
| DateofPurchase | =TEXT(I2, "YYYY-MM-DD") | =IF(ISDATE(I2), TEXT(I2, "YYYY-MM-DD"), "") | =IF(I2="", "", TEXT(I2, "YYYY-MM-DD")) |
| TimeofDay | =IF(TIMEVALUE(J2)<=TIME(11,0,0), 0, IF(TIMEVALUE(J2)<=TIME(15,0,0), 1, 2)) | =IF(J2="N/A", "", IF(TIMEVALUE(J2)<=TIME(11,0,0), 0, IF(TIMEVALUE(J2)<=TIME(15,0,0), 1, 2))) |  |
| PaymentType | =IF(K2="Bitcoin", "N/A", K2) | =IF(ISNUMBER(K2), "", K2) |  |
| CostofGoodsSold | =IF(L2<0, ABS(L2), L2) | =IF(ISNUMBER(L2), L2, "") | =IF(L2="", 0, L2) |
| GrossMarginPercentage | =M2 | =IF(ISNUMBER(M2), M2, "") |  |
| GrossIncome | =N2\*(O2) | =IF(N2*(O2)>=0, N2*(O2), "") | =IF(ISNUMBER(N2*(O2)), N2*(O2), "") |

## Executive Summary

The above data profiling highlighted several data quality issues within the dataset. Identifying and resolving these issues is crucial for ensuring data integrity and preparing for the upcoming organizational merge. A summary of the key data quality issues identified and our comprehensive plan to address them.

**Key Data Quality Issues:**

* **Negative or Unusually High Values:**
  + Examples: Negative unit prices like -43.19 and extremely high values in the Cost of Goods Sold column, such as -9999.
  + Impact: These anomalies can distort financial calculations and analyses.
* **Missing Dates:**
  + Examples: Entries without a purchase date.
  + Impact: Missing dates affect the accuracy of time-based analyses and reporting.
* **Inconsistent Payment Methods:**
  + Examples: Unstandardized payment method entries, such as Bitcoin instead of CC.
  + Impact: Inconsistent payment data can lead to misinterpretation and challenges in financial reconciliation.
* **Outliers in Quantitative Data:**
  + Examples: Unusually high quantities, unit prices, or ratings that deviate significantly from the norm.
  + Impact: Outliers can skew statistical analyses and misrepresent business performance.

**Plan for Data Cleaning and Preparation:**

* **Data Validation and Correction:**
  + Verify Data: Cross-reference with original sources to confirm anomalies are genuine errors.
  + Correct Invalid Entries: Correct negative or unusually high values in UnitPrice and CostOfGoodsSold. For instance, negative values will be converted to their absolute values.
  + Impute Missing Dates: Use logical imputation methods, such as the most common or average date for entries with missing dates.
* **Standardization:**
  + Payment Methods: Standardize payment method entries to Ewallet, Cash, and CC. Replace any unrecognized methods with N/A.
  + Time Slots: Convert Time entries into predefined slots (0: 7am-11am, 1: 11:01am-3pm, 2: all other times).
* **Handling Outliers:**
  + Review and Assess: Identify and review outliers in Quantity, UnitPrice, and Rating to determine if they are valid outliers or errors.
  + Adjustments: Adjust or exclude outliers if they are confirmed as errors, ensuring they do not skew analyses.
* **Consistent Formatting:**
  + Date Formatting: Ensure all dates are in the YYYY-MM-DD format.
  + Data Types: Confirm all numerical values are correctly typed as integers or decimals as required.
* **Documentation:**
  + Maintain a detailed log of all corrections and transformations made, including the reasons and methods used for transparency and future reference.

By systematically addressing these data quality issues through thorough validation, precise correction, and rigorous standardization, we will significantly enhance the accuracy and reliability of our dataset.

Implementing these meticulous steps, we will transform our dataset into a robust, reliable foundation for the organizational merge. This enhanced dataset will support precise data-driven decision-making and strategic planning, ultimately contributing to the seamless integration of our organizational data.

We greatly appreciate your attention to these crucial data quality issues and remain steadfast in our commitment to maintaining the highest standards of data integrity. Thank you for your continued support and collaboration.

Resources

Xiao, D., Chen, B., Fu, X., Li, Z., Wei, C., & Lu, D. (2023). *Control, operation and trading strategies of intermittent renewable energy in smart grids*. Frontiers Media SA.

‌

*Introduction to Data Mining*. (n.d.). YouTube. https://www.youtube.com/playlist?list=PL8eNk\_zTBST-gN6Y5E-5FZdARXjglYpyT

‌