**DATA WRANGLING REPORT**

**Name: Sarah Achok Deu**

**Course Name: Data Analytics**

**Date: 16/02/2025**

1. **Data overview**

The number of columns in this data set is 9 while the number of rows were 210.

The features are: order ID, customer name, supplier name, shipment route, order date, delivery date, order amount, shipment amount and payment status.

1. **Asssumptions**

In this data set, there were several mistakes noticed. They were: the presence of missing values in the customer and supplier name columns. The supplier name had 24 missing contents and the customer name had 16 missing names. Duplicates were also noticed. 10 duplicates were found and removed and 200 were the unique values that remained. There were inconsistencies in the payment status, order date and delivery date columns. There were errors in the data too and lastly, there was misclassification of data in the order date and delivery date columns.

1. **Data wrangling process.**
   1. **Handling missing values.**

First of all, I used the count blank formula to count the number of missing values for each column.

After identifying the missing values, I highlighted the columns I wanted and then I used the find and replace method. This is where I click on the find and replace or click control H. then, for this missing values I left the find box empty then filled the replace with box with “Unknown” and clicked replace all. This filled all the blank spaces with the word Unknown.

The number of missing values handled were 16 in the customer’s name and 24 in the supplier’s name columns. The rest of the columns had not missing values.

3.2**. Detecting and removing duplicates**.

For this issue, I highlighted the order ID column. I clicked on the data tab and went to remove duplicates with all their content and clicked ok.

After doing the above, 10 duplicates were removed and 200 unique values remained.

* 1. **Resolving inconsistencies**

To identify inconsistencies, I filtered the column and see the type of inconsistencies and used the find and replace with icon to make sure everything was right.

The inconsistencies that were changed were: PAID, PENDING, paid, unPaid to Paid and Pending. The date Format was also changed to DD/MM/YYYY using the formatting icon and selected cell formatting where I selected the date format and clicked ok.

* 1. **Identifying and handling outliers and errors**

I identified errors by sorting the values in the order amount and shipment amount from the smallest to the largest to see which values were unrealistic.

The method I used to make them positive was =ABS(CELL) which gave me 0.

* 1. **Correcting misclassified data**

I used this formula in a new column, =IF(A2>B2, “MISCLASSIFIED”, “CORRECT”) to identify misclassified data.

I tried using the same formula to correct the dates but it was giving me values like 4250 and 4251 in the next column. The method used was =IF(A2>B2, B2, A2). This was corrected by correcting the date format

1. **SUMMERY OF CHANGES**

Missing values filled, duplicates removed, inconsistencies fixed, outliers handled, misclassified data identified.

1. **CHALLENGES ENCOUNTERD**

The challenges faced were on the correction of misclassified data, where the date format was incorrect and I did not know it.

**6. Decisions Made**

1. Choice of method for handling missing values.

2. Standardization approach.