**DATA WRANGLING REPORT**

Name: Mubarak Daud Mosmur

Course: Data Analytics

Email: mubarakdaud@gmail.com

**1. Dataset Overview**

* **Rows and Columns:**
* There are 211 rows and 9 columns in the dataset.
* **Features:**
* The Dataset has the following features: Order ID, Customer Name, Supplier Name, Shipment Route, Order Date, Delivery Date, Order Amount, Shipment Cost and Payment Status.

**2. Assumptions made**

* Presence of missing values.
* possible inconsistencies.
* Duplicates
* Outliers
* Misclassifications

**3. Data Wrangling Process**

**3.1 Handling Missing Values**

* **Identification:** Customer names and Supplier names in the Dataset has missing values.
* **Method:** All the missing values were filled with “Unknown”.
* **Outcome:** There were 39 missing values and they had been handled by filtering.

**3.2 Detecting and Removing Duplicates**

* **Identification:** The Duplicates were detected by Order ID.
* **Method:** The Duplicates were removed by Excel’s inbuild "Remove Duplicates" feature.
* **Outcome:** There were 10 Duplicates detected and removed from the Dataset.

**3.3 Resolving Inconsistencies**

* **Identification:** Order Dates and Delivery Dates did not follow the proper date formats. And Payment status formats were inconsistent e.g paid and PAID, unPaid and UnPaid, Pending and PENDING.
* **Method:** Standardizing payment statuses and Reformatting dates correctly.
* **Outcome:** 167 Inconsistencies were detected and corrected.

**3.4 Identifying and Handling Outliers and Errors**

* **Identification:** Order Amount and Shipment cost Columns were sorted from the smallest to the largest and outlies were spotted where there was a Negative value.
* **Method:** All the negative values were corrected.
* **Outcome:** 24 Outliers were corrected and the column that contains the negative values(outliers) were sorted from the smallest to the largest then the negative values were multiplied by -1 using paste special.

**3.5 Correcting Misclassified Data**

* **Identification:** Order Date after Delivery Date and invalid payment status entries.
* **Method:** The misclassified Dates were identified using “=IF(E2>F2, “Misclassified”, “Correct”) and Corrected using “=IF(E2>F2,F2,E2) and =IF(E2>F2,E2,F2) in different cells columns and copy pasted as values in the initial cells columns. The Dates were swapped into the correct forms and Invalid entries were changed to “I i.e unPaid for UnPaid, PAID, paid for Paid, PENDING FOR Pending.
* **Outcome:** 41 misclassifications were corrected.

**4. Summary of Changes**

* Missing values were filled.
* Duplicates were removed.
* Inconsistencies were fixed.
* Outliers were handled.
* misclassified data corrected.

**5. Challenges Encountered**

* Difficulty in identifying errors
* Handling multiple missing values.

**6. Decisions Made**

* Choice of method for handling missing values.
* standardization approach.