AReL: Action for Refugee Life

Data Wrangling in Excel Assignment

Course Name:	Data Analytics for Beginners
Date:	14th February 2025
Submission Deadline:	17th February 2025
Tutor:	Samuel Mati

Objective:

You are a Data Analyst for a Logistics Company. Your task is to help the company answer key business questions such as:

- Who are the best suppliers?
- Who are the top customers?
- What are the most efficient shipment routes?

However, the dataset provided is messy and needs significant data wrangling before any analysis can be conducted. This assignment focuses on cleaning and preparing the data using **Excel**.

Learning Outcomes:

By completing this assignment, you will learn how to:

- Identify and handle missing values.
- Detect and remove duplicates.
- Resolve inconsistencies in data entries.
- Identify and manage outliers and errors.
- Correct misclassified data.
- Document the data cleaning process.

Dataset Description:

The dataset contains the following columns:

- **Order ID** - Unique identifier for each order
- **Customer Name** - Name of the customer
- **Supplier Name** - Name of the supplier
- **Shipment Route** - Route taken by shipment
- **Order Date** - Date when the order was placed
- **Delivery Date** - Date when the order was delivered
- **Order Amount** - Total amount of the order
- **Shipment Cost** - Cost incurred for shipment
- **Payment Status** - Status of payment (Paid, Unpaid, Pending)

Instructions:

- Use **Microsoft Excel** to perform all data wrangling tasks.
- Document each step and provide explanations for your approach.
- After cleaning the data, summarise the changes made (e.g., how many duplicates were removed, how many missing values were filled, etc.).
- Submit the cleaned Excel file along with a short report (1-2 pages) explaining your data cleaning process.

Assignment Questions:

1. Handling Missing Values

- Identify columns with missing values.
- Use the following methods to handle them:
    - **Order Amount** and **Shipment Cost**: If missing, calculate and fill with the average of the respective column using the **AVERAGE** function.

    - **Customer Name** or **Supplier Name**: If missing, fill with “Unknown.”
- Document the number of missing values found and the method used for filling them.

2. Detecting and Removing Duplicates

- Check for duplicate records in the dataset.
    - Consider a duplicate as a record with the same **Order ID**.
- Use **Remove Duplicates** in Excel to eliminate duplicates.
- Document how many duplicates were found and removed.

3. Resolving Inconsistencies

- Standardize inconsistent entries in **Payment Status** (e.g., 'paid', 'Paid', 'PAID' should all be standardized to 'Paid').
    - Use **Find and Replace** for standardization.
- Ensure consistent date formats in **Order Date** and **Delivery Date** columns (e.g., DD/MM/YYYY).
    - Use **Format Cells** to standardize the date format.
- Document the inconsistencies found and how they were resolved.

4. Identifying and Handling Outliers and Errors

- Check for outliers in the **Order Amount** and **Shipment Cost** columns.
    - Sort the values from smallest to largest to inspect for unusually high or low values.
- Correct or remove erroneous data points (e.g., negative values for **Order Amount** or **Shipment Cost**).
    - For negative values, change to positive or mark as “Error” if unsure of the correction.

- Document the outliers found and the changes made.

5. Correcting Misclassified Data

- Identify and correct misclassified data, such as:
    - **Order Date** occurring after **Delivery Date**.
        - Swap the dates if necessary or mark as “Error.”
    - Invalid entries in **Payment Status** (e.g., numeric values or irrelevant text).
        - Change invalid entries to “Pending.”
- Document the misclassifications found and how they were corrected.

6. Summary and Documentation

- Provide a summary of all changes made to the dataset, including:
    - The number of missing values filled.
    - The number of duplicates removed.
    - Inconsistencies standardized.
    - Outliers and errors corrected.
    - Misclassified data fixed.
- Write a brief report (1-2 pages) explaining:
    - The data wrangling process.
    - Challenges faced.
    - Decisions made and why.

Submission Requirements:

- **Cleaned Excel file** with all changes saved.
- **Report (1-2 pages)** detailing the data cleaning process.
    - The report should be in PDF format.

Grading Criteria:

- Completeness and accuracy of data cleaning.
- Justification of methods used for handling data issues.
- Clarity and structure of the report.
- Overall presentation and organization of the cleaned dataset.