**PROJECT OVERVIEW: DATA CLEANING**

**TITLE: NIGERIAN RAILWAY DATA**

**OUTLINE**

* Introduction
* Objectives
* Data Overview
* Tool used for Data cleaning
* Data Cleaning
* Summary (Key Finding and suggestions)
* Conclusion

1. **INTRODUCTION**

Data Cleaning in Data Analysis is the process of standardizing data(fixing/correcting) by removing duplicates, errors (wrong values) and inconsistencies (incorrectly formatted) to ensure the integrity of Data so as to yield accurate results. There are different tools that can be used for data cleaning, e.g., Microsoft excel, Power Query (found in excel/Power Bi). For this Dataset, **Nigerian Railway Data,** we would be using **Microsoft Excel** for Data Cleaning.

1. **OBJECTIVES**

There are three main objectives to achieve from cleaning this Data

* Data Standardization
* Improve Data Quality
* Ensure Data Integrity for future analysis

1. **DATA OVERVIEW (NIGERIAN RAILWAY DATA)**

The dataset contains the following key fields:

* Date: The date of maintenance
* Vehicle Registration Number: Unique identifier for each vehicle (Primary Key).
* Make / Model: The brand and model of the vehicle.
* Item Description: Description of the maintenance services performed.
* Assigned To: The person or department responsible for the vehicle.
* Service KM: The recorded kilometers at the time of service.
* Workshop Name: The name of the maintenance workshop.
* Invoice No.: The invoice number for the service.
* Amount (NGN): The cost of the maintenance service.
* Remarks: Additional comments on the maintenance.

1. **TOOL USED FOR DATA CLEANING**

* Microsoft Excel

1. **DATA CLEANING**

Phase 1- Text Overflow

Issue: Some cells contain text that is too long and gets cut off or hidden

Data in range format which makes sorting data tasking

Action: Used **Alt+H+O+I** to adjust column without manual resizing

Used **Ctrl+T** to create a table

Phase 2- Validating and Standardizing Data Types

Issue: Incorrect date format (Date was stored in general format causing dates to appear as numbers)

Incorrect format on Amount

Action: Used **Excel Number Formatting** to change the numerical values to short date format

**Currency formatting** on Amount (NGN)

Phase 3- Removing Duplicates

Issue: Duplicate Vehicle Registration Numbers found in unique identifier but there were no duplicates as some vehicles had multiple maintenance records (different dates).

Action: Used the **Remove Duplicates** feature under the Data tab

Phase 4- Standardizing Make/Model Names and Item Description

Issue: Inconsistent capitalization and spacing.

Action: Used (**TRIM(PROPER(C9))** function to standardize names and **(TRIM(PROPER(E9))** item description and correct spacing.

Phase 5- Cleaning Numerical Data (Service KM, Amount & Invoice)

Issue: Missing Values in Service KM/ Inconsistent Values in amount

Action: The missing values in the Service KM column were filled by finding the **Median** of the values in the column. Since the data is not perfectly normally distributed, the **median** is a better choice than the mean as it is more robust to outliers and skewed data. There are no outliers as confirmed using the **Interquartile range(Any Service KM value below (Q1 - 1.5IQR) or above (Q3 + 1.5IQR) was considered an outlier, but none were found)**.

**Find and replace** was used to fill in the blanks.

The inconsistent values in the amount column were rectified by applying ROUND () to standardize decimal places

The missing values in the invoice were replaced with **N/A,** the number format was also adjusted under the **Number Formatting**

Phase 6- Correcting Errors: Typos and Incorrect Spelling.

Issue: Inconsistent workshop names and department assignments.

Action: Used **Find and Replace** to correct typos and inconsistencies

Phase 7- **Final Validation**

**Issue: Checked for blanks after cleaning**

**Checked for data type format**

Action: Blank column was removed (Remarks). Proper alignment on all columns.

All columns were in the right datatype

1. **SUMMARY**

### **Key Findings**

* The dataset contained **missing values**, **inconsistencies**, **typos, incorrect datatypes.**
* Some missing values (e.g., invoice numbers) couldn’t be resolved.
* Data cleaning techniques significantly improved the dataset’s quality.
* Further validation may be needed for workshop names to ensure uniformity and accuracy.

**Recommendations**

To maintain data integrity, data validation rules were applied in Microsoft Excel to prevent errors in future entries. These include:

* Restricting Invalid Date Entries: A date validation rule was applied to ensure all entries in the Date column follow the YYYY-MM-DD format.
* Ensuring Consistency in Vehicle Registration Numbers: Applied a unique constraint to prevent duplicate entries, reducing tracking errors and improving data accuracy.
* Standardizing Workshop Names: Created a drop-down list of verified workshop names to ensure uniformity.
* Preventing Incorrect Currency Entries in the ‘Amount (NGN)’ Column: Applied currency formatting (NGN) to ensure all amounts are stored as numbers.

1. **CONCLUSION**

The data cleaning process enhanced accuracy and consistency, resolving errors and ensuring reliability. Issues like incorrect formats, missing values, duplicates, and inconsistencies were resolved, ensuring the data is ready for analysis. Validation rules were added to prevent future issues, making the dataset ready for analysis and better decision-making

[Nigerian Railway Data- Cleaned](Nigerian_Railway_Data%20Cleaned.xlsx)