Data Cleaning Report – Messy Data

**Project : Datahut QA Assignment**

Tools Used : Python, Jupyter Notebook, pandas, Visual Studio

This document outlines the steps taken to clean the dataset. The objective was to find and fix data issues to make the dataset clean, accurate, and ready for analysis.

Steps taken for data cleaning

1. Load the data

* Loaded the dataset into a Pandas DataFrame.
* Inspected the structure using .info(), .head(), and .describe().

1. Identify and Record QA Issues

* Created a separate Excel file QA\_Isues.xlsx to document all data quality issues.
* Included columns: Field, QA Issue, Issue Type, Data Line Number, Solution, and Suggestion to prevent the issue.

1. Handling Missing Values

* Filled missing values in:
* Salary - Replaced with the median salary.
* Department – Assigned as Unknown if missing.
* Email – Assigned placeholder [missing@example.com](mailto:missing@example.com).
* Dropped rows where name or Id was missing since they are crucial fields.

1. Removing Duplicates

* Identified duplicates using .duplicated().
* Removed duplicate rows using .drop\_duplicates().

1. Standardizing Email Formats

* Used regex to check for valid email formats ([username@doamin.com](mailto:username@doamin.com)).
* Removed non-professional email addresses (gmail.com, yahoo.com).

1. Cleaning Name Fields

* Trimmed spaces and removed extra characters using .str.strip().
* Standardized name capitalization using .str.title().

1. Standardizing Date Formats

* Converted Join Date to YYYY-MM-DD format using pd.to\_datetime().

1. Correcting Department Names

* Standardized department names (HR, Marketing) by mapping known variations.

1. Handling Salary Noise

* Identified outliers in Salary using .describe() and box plots.
* Crapped unreasonable salary values within an acceptable range.

**Key Assumptions**

1. Missing Salary values were filled with the median instead of mean to avoid skewed data.
2. If Department was missing, assigned it as Unknown.
3. Non-professional emails were assumed to be invalid for this dataset.

**Summary**

1. Cleaned dataset saved as cleaned\_dataset.csv.
2. QA issues documented in QA\_issues.xlsx.