**Datahut QA Assignment Summary Document**

**1. Introduction**

This document summarizes the data quality assessment and cleaning process performed on the provided dataset for the Datahut QA Assignment. The objective was to ensure the dataset is accurate, consistent, and ready for analysis.

**2. Data Inspection**

* **Initial Inspection:** Used pandas to inspect data structure, identify missing values, duplicates, formatting errors, and outliers.
* **Key Issues Identified:** Missing values in Age and Salary, invalid Email formats, inconsistent Join Date formats, typos in Department names, and salary outliers.

**3. QA Issues Recording**

* Documented field-wise QA issues in an Excel sheet with columns for Field, QA issue, Issue type, Data line number, Solution, and Suggestions.

**4. Data Cleaning Steps**

* **Missing Values:** Imputed missing values using median.
* **Duplicates:** Removed duplicate rows.
* **Email Formats:** Corrected email formats using regex and retained only professional emails.
* **Name Fields:** Cleaned noise (extra spaces, symbols) from names.
* **Date Formats:** Standardized all dates to YYYY-MM-DD.
* **Department Names:** Corrected typos and ensured consistency.
* **Salary Noise:** Handled outliers by capping salary values within a reasonable range.

**5. Final Output**

* Cleaned dataset saved as cleaned\_data.csv.
* QA issues documented in Datahut\_QA\_Issues.xlsx.

**6. Conclusion**

This document outlines the complete QA process followed for the Datahut QA Assignment. The cleaned dataset is now accurate, consistent, and ready for further analysis or reporting.