Data Cleaning Summary Document

1. Loading the Data

- The dataset was loaded into a Jupyter Notebook using the pandas library.
- The initial structure of the dataset was inspected using .info(), .head(), and .describe() to identify inconsistencies, missing values, and data types.
- An overview of column names, data types, and summary statistics was recorded.

2. Inspection and Quality Assessment

- The dataset was examined for missing values, inconsistencies, incorrect formats, and duplicate entries using .isnull().sum(), .duplicated().sum(), and summary statistics.
- The presence of special characters, unnecessary spaces, and incorrect capitalization in categorical data was analyzed.
- Each identified issue was documented in an Excel file named **QA_Issues.xlsx**, including issue type and recommended fixes.

3. Data Cleaning Steps

3.1 Handling Missing Values

- Numerical fields with missing values were analyzed for distribution and filled using the mean or median where appropriate.
- Categorical missing values were replaced with the most frequently occurring value (mode) or labeled as "Unknown" if no logical replacement was available.
- Rows with excessive missing values (more than 50% missing fields) were removed to maintain data integrity.
- Specific columns with high importance but missing values were imputed based on related fields.

3.2 Removing Duplicates

- Duplicate records were identified using .duplicated() and removed to ensure uniqueness.
- Records with minor variations (e.g., case differences or extra spaces) were normalized before checking for duplicates.
- The primary key fields were used to verify duplication to avoid accidental removals.

3.3 Correcting Email Formats

- Email addresses were validated using a regex pattern to ensure correct structure (e.g., username@domain.com).
- Invalid emails (missing '@', incorrect domain formats, fake domains) were flagged and either corrected or removed.
- Only professional emails (excluding personal domains such as gmail.com, yahoo.com) were retained, assuming the dataset was for business purposes.

3.4 Cleaning Name Fields

- Names containing numbers, special characters, or unnecessary words (e.g., "Mr.", "Dr.") were cleaned.
- Extra spaces, leading/trailing spaces, and inconsistencies in capitalization were standardized.
- Entries with incomplete names (single-letter names or initials without a full name) were flagged for review.

3.5 Standardizing Date Formats

- Join Date formats were found to be inconsistent (mix of DD/MM/YYYY, MM-DD-YYYY, YYYY.MM.DD).
- Dates were converted to a uniform format: **YYYY-MM-DD** using pandas' to_datetime() function.
- Erroneous date values (e.g., future dates, unrealistic past dates, or non-date values) were flagged and corrected.

3.6 Correcting Department Names

- Department names were checked against a predefined valid department list: HR, Engineering, Marketing, Sales, Support.
- Variants, abbreviations, and typos such as "Engneering", "Mktg", and "SupportJ" were corrected to their standard forms.
- Inconsistent capitalization was standardized to title case.
- Departments with missing values were assigned based on other available employee attributes.

3.7 Handling Salary Noise

- Salary values were checked for outliers using summary statistics and box plots.
- Salaries below **20,000** and above **300,000** were flagged for manual review.
- Incorrect salary formats (e.g., textual values, negative numbers, missing currency symbols) were identified and corrected.
- Salaries were converted into a consistent currency format if multiple formats were found.

4. Documentation of Process

- The entire data cleaning process was documented within the Jupyter Notebook, including before-and-after transformations.
- Code snippets used for cleaning were commented to explain logic and methodologies applied.
- The final cleaned dataset was saved as **cleaned_dataset.csv** for further analysis.
- An Excel file (**QA_Issues.xlsx**) was created to track identified issues, solutions, and prevention methods.

5. Deliverables

- **cleaned_dataset.csv** (Final cleaned dataset with structured and corrected records)
- **QA_Issues.xlsx** (Issue tracking file documenting identified problems and resolutions)
- **Data Cleaning Jupyter Notebook** (Includes detailed steps, code implementations, and results)
- **This Summary Document outlining all steps taken**

6. Assumptions and Considerations

- Missing values were imputed based on logical assumptions (mean/median for numerical data, mode for categorical data).
- Only professional emails were retained, assuming the dataset was meant for business use.
- Salary thresholds were set based on reasonable industry standards.
- Department standardization assumed predefined valid department names.
- Duplicate removal was conducted only on exact matches to avoid unintentional data loss.
- Some corrections were made based on logical assumptions and available patterns in data.

7. Future Recommendations

- Implement validation rules at data entry points to minimize missing values and incorrect data types.
- Enforce a consistent date format across all input sources to avoid formatting issues.
- Utilize regex-based validation for email entries to ensure compliance with professional standards.
- Regularly audit and clean the dataset to maintain high data quality for analysis.
- Set up automated checks for detecting salary anomalies and incorrect department names.
- Implement standardization tools to prevent inconsistent formatting in names and emails.

This data cleaning process ensures that the dataset is structured, error-free, and ready for further analysis, providing reliable insights for decision-making.