**Data Cleaning and Preparation: A Comprehensive Guide**

**1. Objective**

The goal of this task is to clean and join datasets from two suppliers (Company A GmbH and Company B GmbH) into a single **inventory\_dataset**. The datasets provided are:

* task\_1/supplier\_data\_1.xlsx
* task\_1/supplier\_data\_2.xlsx

Additionally, the cleaned data and transformed data will be uploaded to Google BigQuery for further Querying.

**Data Cleaning and Processing Documentation**

**1. Importing Required Libraries**

* The necessary Python libraries such as Pandas, NumPy, were imported to facilitate data manipulations and Transformations.
* Ensured all required dependencies were installed and compatible with the Python environment.

**2. Loading the Dataset**

* Loaded supplier\_data\_1 and supplier\_data\_2 into Pandas DataFrames.
* Verified successful loading by displaying the first few rows of each dataset using df.head().
* Checked dataset dimensions using df.shape to understand the volume of data.

**3. Initial Data Exploration**

* Displayed summary statistics using df.describe() to identify potential outliers or anomalies.
* Used df.info () to examine data types, non-null counts, and memory usage.
* Checked column names for inconsistencies such as leading/trailing spaces or case sensitivity issues.

**4. Handling Missing Values**

* Used df.isnull().sum() to count missing values in each column.
* Identified columns with excessive null values and removed them

**'SURFACE\_COATING', 'HEIGHT\_MM', 'NUMBER\_OF\_COILS', 'DELIVERY\_EARLIEST' , 'DELIVERY\_LATEST', 'CO2\_PER\_TON\_MAX\_KG'**

* Implemented imputation techniques where applicable and fill the null values with Not\_Available .
* **PRODUCT\_TYPE** has null values filled those null values with their correct PRODUCT\_TYPE With the help of material\_name column
* **DEFECT\_NOTES** column null values were fixed the regex function by seeing the available datas and linked to the null values
* **MATERIAL\_QUALITY\_NORM** checked the previous data and fix accordingly

**5. Correcting the correct column names**

* Some column datas has been wrongly entered so fixed with their correct name Replaced ({'SHET': 'SHEET', 'COILS\_STRIP': 'COIL\_STRIP'}
* Renamed (columns={'Nenndicke NNN.NN mm mit Dezimalpunkt': 'Nenndicke'} for better viewing

5. **Handling Duplicate Data**

* Used df.duplicated().sum() to check for duplicate rows.
* Removed duplicate records using df.drop\_duplicates(inplace=True) to maintain data integrity.
* Verified successful removal by rechecking the dataset dimensions.

6. Data Consistency Checks

* Checked for inconsistent or unexpected values in categorical columns using df['column\_name'].unique().
* **Zugfestigkeit** column has some inconsistencies with the floating values so replaced 0.000 with NAN
* **'Si-Gehalt', 'Mn-Gehalt', 'P-Gehalt', 'S-Gehalt', 'Cr-Gehalt', 'Ni-Gehalt','Mo-Gehalt'** these columns has non-numeric or categorical values so replaced those with NAN values
* Converted **VALID\_UNTIL** to proper datetime format using pd.to\_datetime().
* Created a new column named **'SITE'** where the values are ‘**2 company gmbh'** throughout all the rows for easy inventory sorting
* Ensured numeric columns were of the correct data type to prevent computational errors.

7. Error Handling

* Checked for incorrect data types and applied necessary conversions.
* Addressed issues related to missing values and duplicates before proceeding with analysis.

8. Additional Data Cleaning Steps

* Further refined the dataset by renaming columns for better readability ie : {"Gewicht (kg)": "**Gewicht\_kg**"} and columns={"MIN/MAX\_BID\_EUR\_PER\_TON": "**MIN\_MAX\_BID\_EUR\_PER\_TON**"}
* Concatenate both the dataset and merged into one single dataset called **inventory\_df**

9. Assumptions

* Adjusted supplier\_data\_2 pricing values to maintain logical consistency by:

Setting **MIN/MAX\_BID\_EUR\_PER\_TON** to 570 where **BUY\_NOW\_EUR\_PER\_TON** was 600.

* Filling missing values in **BUY\_NOW\_EUR\_PER\_TON** by subtracting 30 from **MIN/MAX\_BID\_EUR\_PER\_TON**.

**5. Final Thoughts & Best Practices**

* **Always inspect datasets before cleaning** to understand structure and issues.
* **Use consistent column names** to avoid errors during merging and analysis.
* **Handle missing values appropriately** based on the nature of the data.
* **Ensure correct data types** before uploading to BigQuery.
* **Document steps taken** for future reference and reproducibility.

This guide provides a structured approach to data cleaning and preparation, ensuring high-quality data for analysis.