**Final Report: Data Quality Analysis and Cleansing for SIS Faculty Dataset**

**Introduction**

This report presents a detailed analysis and data cleansing process for a Student Information System (SIS) faculty dataset. The dataset was analyzed to identify and resolve data quality issues before it could be used for machine learning purposes. The focus was on correcting missing data, ensuring consistency in formatting, and preparing the data for effective use in a machine learning pipeline.

**Data Quality Analysis**

The following data quality issues were identified in the initial dataset:

1. **Missing Values**:
   * The "ID" column had placeholder values of '0', indicating missing or invalid entries.
   * The "LWD" (Last Working Date) column contained numerous missing values.
2. **Inconsistent Data**:
   * The "Highest Qualification" column exhibited inconsistencies, with values like "Master of Science" and "[M.Sc](http://M.Sc)." used interchangeably.
   * The "Join Date" column contained newline characters, making it difficult to interpret the data consistently.
3. **Redundant or Irrelevant Columns**:
   * Categorical data in the "Location" and "Title" columns had already been encoded, with columns such as "Location\_London" and "Title\_Faculty - Business", which needed to be validated for encoding accuracy.
4. **Unstructured Text**:
   * Columns like "Courses Taught" contained long text values, which may require transformation for machine learning purposes.

**Data Cleansing Measures**

The following measures were applied to address the identified issues and prepare the dataset for machine learning:

1. **Handling Missing Values**:
   * Rows where the "ID" was '0' were removed.
   * Missing values in the "LWD" column were replaced with the string "Not Available" to maintain consistency.
2. Code:
3. **Data Standardization**:
   * The "Join\nDate" column was cleaned by removing newline characters, and inconsistent qualification titles were standardized.
4. Code:
5. **Encoding Categorical Data**:
   * Columns for "Location" and "Title" were already one-hot encoded, so no further transformation was required.
   * After validating the encoding, it was found to be correct, with no redundant transformations needed.

**Python Code Implementation**

The following Python code snippets were used during the data cleansing process:

# Drop rows with invalid IDs  
df\_cleaned = df[df["ID"] != "0"]  
  
# Fill missing values in the 'LWD' column  
df\_cleaned["LWD"] = df\_cleaned["LWD"].fillna("Not Available")  
  
# Clean 'Join Date' column and standardize qualifications  
df\_cleaned["Join Date"] = df\_cleaned["Join\nDate"].str.replace("\n", "")  
df\_cleaned["Highest Qualification"] = df\_cleaned["Highest Qualification"].str.lower().str.replace("master", "M.Sc.")  
  
# One-hot encoding was already applied for Location and Title columns

**Conclusions**

The data cleansing process successfully addressed the major data quality issues, including missing values, inconsistent formatting, and unstructured text. The dataset is now ready for machine learning, with consistent entries and proper handling of categorical columns through one-hot encoding. With these improvements, the dataset is now suitable for training models within the SIS management reporting system.