Report on Mentorship Sessions Data Cleaning and Analysis

Introduction

This report outlines the process of cleaning and analyzing a dataset from mentorship sessions aimed at providing insights for optimizing a rewards program. The dataset included various inconsistencies and duplicates, necessitating a comprehensive cleaning process to ensure data quality. This step was crucial for accurate reporting and analysis of the mentorship program.

Tools Used

The following tools were employed throughout the data cleaning process:

- Python: The programming language used to manage the data cleaning tasks.

- Pandas: A Python library crucial for handling large datasets, data manipulation, and cleaning.

- Google Sheets: The source of the original mentorship sessions dataset.

Data Cleaning Process

1. Data Loading

The dataset was imported directly from Google Sheets into a Pandas DataFrame using the `read_csv()` function. This allowed for easy manipulation and analysis in Python.

2. Initial Data Inspection

We inspected the dataset's structure using methods like `.head()` and `.info()` to get a sense of the data types, missing values, and duplicate entries. The key columns examined included `Mentor_ID`, `Session_Number`, `Session_Date`, `Mentee_Name`, `Session_Duration_Min`, and `Job_Info_Completed`.

3. Handling Missing Values

- Rows where critical data such as `Mentor_ID`, `Session_Number`, and `Session_Date` were missing were dropped as these were necessary for accurately identifying sessions.

- Missing values in the `Mentee_Name` column were left intact for now, as this information was considered non-essential for the immediate task. However, further analysis is recommended to investigate why some mentee names were missing.

- Missing session durations were filled using the median session duration. This approach ensured a consistent range of values and prevented skewed analysis from incomplete data.

4. Data Standardization

The `Job_Info_Completed` column had inconsistent values, such as `YES` and `NO` in different formats. These were standardized to `Yes` and `No`, ensuring uniformity across the dataset and easier future filtering.

5. Duplicate Record Removal

We identified and removed all duplicate rows from the dataset using the `.duplicated()` function. This step was critical in ensuring that the analysis wouldn't be distorted by repeated records.

Findings and Observations

- A total of several duplicate records were removed from the dataset, ensuring that only unique mentorship sessions were considered for further analysis.

- Missing values in critical fields were handled efficiently, allowing for a clean dataset ready for reporting.

- The standardization of columns like `Job_Info_Completed` helped ensure uniformity and consistency across the data.

- The cleaned dataset is now fit for further analysis, such as evaluating session durations, mentor performance, and mentee participation, leading to deeper insights into the mentorship program.

Conclusion

Through this data cleaning process, the dataset was refined for accurate reporting and analysis. The next steps involve leveraging the clean dataset for exploratory data analysis to derive actionable insights for the rewards program, particularly focusing on user engagement and satisfaction.