Summary Report

Tools Used

- **Python**: Utilized for data manipulation and cleaning.
- **Pandas**: Essential for handling data-frames and performing various data cleaning operations.
- **Excel**: Employed for loading and saving the dataset.

Data Cleaning Process

- The data cleaning process involved several meticulous steps to ensure the dataset's integrity and readiness for analysis:

1. **Loading Data**: The dataset 'Mentorship_Sessions.xlsx' was imported using the pandas library.
2. **Renaming Columns**: An unnamed column was identified and renamed to `Mentee_ID` based on its content.
3. **Handling Missing Values**: Missing values were identified and appropriately filled or flagged.
4. **Handling Duplicates**: No duplicates were found in the dataset.
5. **Correcting Inconsistent Data**: Inconsistent data types were corrected to ensure uniformity and validity.
6. **Saving the Cleaned Dataset**: The final cleaned dataset was saved for further analysis.

Findings

Upon inspection, the following findings were noted:

1. The first column was unnamed, but it appeared to represent mentee IDs and was therefore renamed `Mentee_ID`.
2. **Missing Values Analysis**

- The table below illustrates the missing values in various columns:

Variable	Missing values
Mentee_ID	1
Mentor_ID	1
Mentor_Name	0
Mentee_Name	2
Session_Number	1
Session_Duration_Min	2
Job_Info_Completed	1
Session_Date	1
Points_Awarded	109

Handling Missing Values:

- Missing values in `Session_Number` and `Session_Duration_Min` were filled using the mean of the respective columns.
- Other variables with missing values were flagged as `NaN`.
- `Points_Awarded` had numerous 109 missing values, which were initially filled with 0, to be updated later with the correct reward points in Task 2.
- The Python code used for handling missing values is shown below:

```python
# Handle the missing values by dropping or filling in rows/columns
df.dropna(subset=['Mentee_ID','Mentor_ID', 'Mentee_Name', 'Job_Info_Completed', 'Session_Date'])

# Fill NaN values in Session Number and duration with mean
Session_Number_Mean = df['Session_Number'].mean()
Session_Duration_Mean = df['Session_Duration_Min'].mean()
df['Session_Number'] = df['Session_Number'].fillna(Session_Number_Mean)
df['Session_Duration_Min'] = df['Session_Duration_Min'].fillna(Session_Duration_Mean)
df['Points_Awarded'] = df['Points_Awarded'].fillna(0)  # Replace with starting value
```

3. **No Duplicates Detected**: There were no duplicate entries found in the dataset.
4. **Inconsistent Data Types**: Several variables had inconsistent or invalid data types. For instance, the 'Session_Date' column was converted to a consistent datetime format.

Finally, the cleaned dataset was saved in Excel format for subsequent analysis.