"Week 1: Data Cleaning and Feature Engineering Report"

Intern's Name: Preta Kumar Binda

Team 08

Date of Submission: 19 January, 2026

Introduction:

Purpose:

The purpose of this project is to clean and prepare a user–opportunity engagement dataset for analysis by correcting inconsistencies, handling missing values, and engineering meaningful features. The cleaned dataset enables accurate analysis of user demographics, participation patterns, and engagement outcomes across different opportunities over time.

Data Description:

This dataset represents user engagement with Excelerate opportunities at the user–opportunity level. Each record combines user demographic information with opportunity details such as category, application date, duration, and participation status. The data enables analysis of user characteristics, enrollment trends, and engagement outcomes across different opportunities over time.

The dataset consists of **16** columns (features), and **8558** entries and the features names and types as follows:

Learner Sign Up Date Time	Opportunity Id	Opportunity Name	Opportunity Category	Opportunity End Date	First Name	Date of Birth	Gender
object	**object**	**object**	**object**	**object**	**object**	**object**	**object**

Country	Institution Name	Current/Intended Major	Entry created at	Status Description	Status Code	Apply Date	Opportunity Start Date
object	**object**	**object**	**datetime64[ns]**	**object**	**Int64**	**object**	**object**

```
df1.head(5)
```

	Learner SignUp DateTime	Opportunity Id	Opportunity Name	Opportunity Category	Opportunity End Date	First Name	Date of Birth	Gender	Country	Institution Name	Current/Intended Major	Entry created at	Status Description	Status Code	Apply Date	Opportunity Start Date
0	06/14/2023	00000000-0GN2-A0AY-7XK8-C5FZPP	Career Essentials: Getting Started with Your P...	Course	06/29/2024	Faria	01/12/2001	Female	Pakistan	Nwihs	Radiology	03/11/2024	Started	1080	06/14/2023	11/03/2022
1	05/01/2023	00000000-0GN2-A0AY-7XK8-C5FZPP	Career Essentials: Getting Started with Your P...	Course	06/29/2024	Poojitha	08/16/2000	Female	India	SAINT LOUIS	Information Systems	03/11/2024	Started	1080	05/01/2023	11/03/2022
2	04/09/2023	00000000-0GN2-A0AY-7XK8-C5FZPP	Career Essentials: Getting Started with Your P...	Course	06/29/2024	Emmanuel	01/27/2002	Male	United States	Illinois Institute of Technology	Computer Science	03/11/2024	Started	1080	05/11/2023	11/03/2022
3	08/29/2023	00000000-0GN2-A0AY-7XK8-C5FZPP	Career Essentials: Getting Started with Your P...	Course	06/29/2024	Amrutha Varshini	11/01/1999	Female	United States	Saint Louis University	Information Systems	03/11/2024	Team Allocated	1070	10/09/2023	11/03/2022
4	01/06/2023	00000000-0GN2-A0AY-7XK8-C5FZPP	Career Essentials: Getting Started with Your P...	Course	06/29/2024	Vinay Varshith	04/19/2000	Male	United States	Saint Louis University	Computer Science	03/11/2024	Started	1080	01/06/2023	11/03/2022

Data Cleaning Process:

Cleaning Steps:

A comprehensive data cleaning process was applied to improve data quality, consistency, and usability for analysis. The following steps were performed:

- **Invalid and noisy values were corrected** across multiple columns. In the ***Current/Intended Major*** field, numeric and symbolic entries were identified as invalid and standardized. In the ***Institution Name*** column, multilingual and inconsistent text entries were detected and handled to reduce categorical noise.
- **Redundant columns were identified and removed**. A strong correlation was observed between ***Status Code*** and ***Status Description***, and between ***Opportunity Id*** and

Opportunity Name. To avoid redundancy and multicollinearity, one column from each highly correlated pair was removed.

- **Date and time attributes were standardized** by removing unnecessary time components, retaining only date values. This ensured consistency across temporal columns and simplified time-based analysis.
- **Text formatting inconsistencies were resolved** by converting all categorical text fields to lowercase, eliminating variations caused by mixed casing.
- **Location-based variations in institution names were normalized**. Place names appended to institution names were identified and conditionally removed to prevent multiple representations of the same institution while preserving official institution identities.
- **Invalid values in personal name fields were cleaned**. Numeric characters and special symbols in the *First Name* and *Current/Intended Major* column were removed, and missing or unusable entries were replaced with a standardized "**Unknown**" value.

Issues Encountered and Resolution

Several challenges were encountered during the cleaning process, including free-text input errors, multilingual entries, inconsistent formatting, and structural redundancy between columns. These issues were resolved using rule-based text normalization, regex-based validation, conditional entity normalization, and careful feature selection.

Feature Engineering:

New Features Created:

Several new features were engineered to enhance the analytical value of the dataset and enable deeper insights into user behavior and opportunity engagement.

- **Age_at_Application_Int**

  This feature represents the applicant's age (in years) at the time of application. It was calculated by subtracting the date of birth from the application date and converting the difference into years. The code used there is:

```
df['Age_at_Application_Int'] = (
    (df10['Apply Date'] - df10['Date of Birth']).dt.days // 365
)
```

- **Institution_Participation_Count**

  This aggregated feature represents the number of applicants from each institution. It was created to analyze institutional engagement and participation trends. The code used there is:

```
df10['Institution_Participation_Count'] = (
    df10.groupby('Institution Name')['Institution Name']
        .transform('count')
)
```

Feature Engineering Example: Age Calculation

The applicant's age at the time of application was derived using date difference computation. This temporal transformation converts raw date values into a meaningful demographic feature.

Code:

```
df10['Age_at_Application_Int'] = (
    (df10['Apply Date'] - df10['Date of Birth']).dt.days // 365
)
```

Validation Summary

After completing data cleaning and feature engineering, the dataset was systematically validated to ensure accuracy, consistency, and analytical reliability. Multiple validation checks were performed, and their outcomes confirmed that the dataset is suitable for further analysis.

First, **schema validation** was conducted using data type inspection. All columns were found to have appropriate and consistent data types: date-related attributes were stored as datetime64[ns], engineered numerical features such as ***Age_at_Application_Int***, and ***Institution_Participation_Count*** were stored as integer values, and categorical attributes were stored as object types. This confirms correct structural formatting of the dataset.

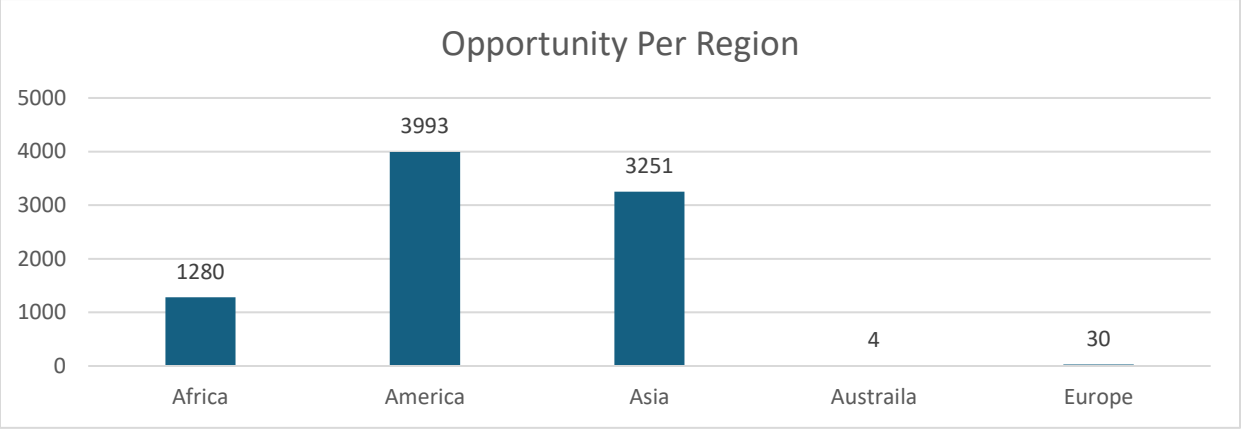
Second, **missing value validation** was performed to identify any remaining null values. The results showed that all columns contained complete data except for *Opportunity Start Date*, which had missing values. This missingness was previously identified as structural and related to opportunity lifecycle stages (e.g., applied or waitlisted opportunities) rather than data quality issues. Therefore, these values were retained as valid missing data.

Third, **record completeness and consistency** were verified by confirming that all other critical fields—including demographic attributes, opportunity details, and engineered features—contained no missing values. This indicates that the data cleaning process successfully resolved incomplete and inconsistent entries.

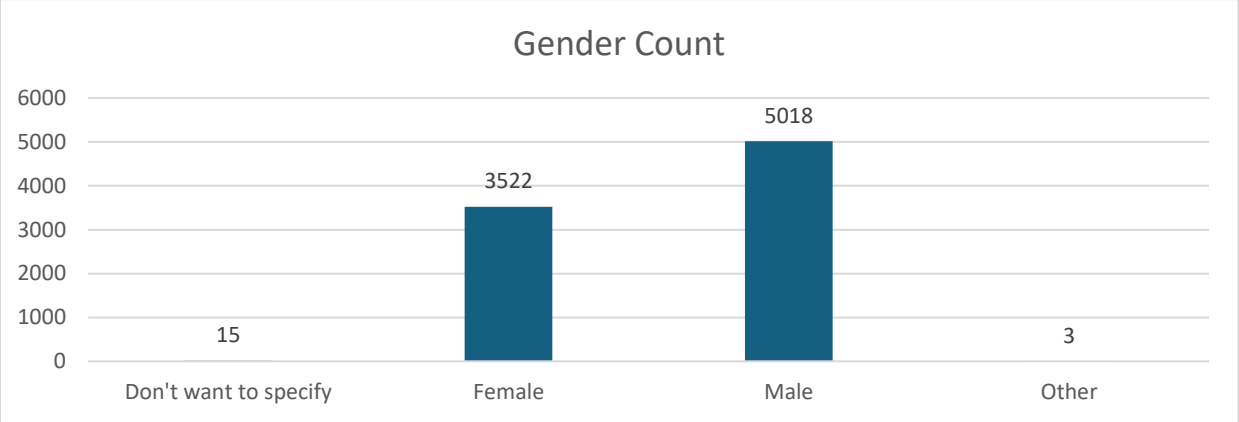
Finally, **feature-level validation** confirmed that engineered features were correctly populated for all records. The age feature contained valid numeric values, binary features contained only valid categories (0 or 1), and aggregated participation counts were consistently applied across institutional records.

Overall, the dataset passed all validation checks and was confirmed to be clean, consistent, and ready for reliable exploratory analysis and reporting.

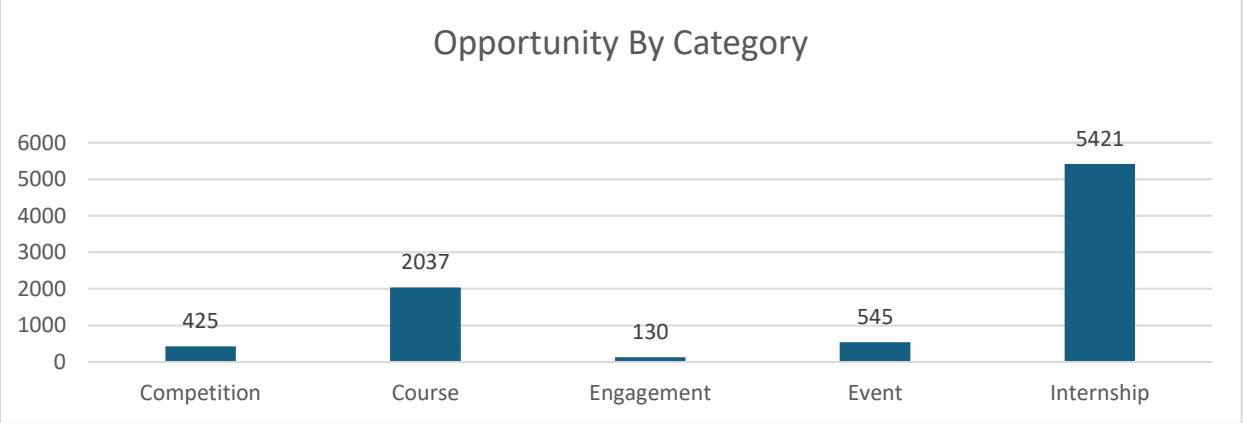
Observations:



This graph shows that most opportunities are concentrated in America and Asia, with moderate participation in Africa and minimal representation in Europe and Australia.



The gender distribution indicates higher participation among males, followed by females, while very few participants selected other or unspecified gender options.



Internships are the most common opportunity type, followed by courses and events, indicating a strong focus on skill-based programs.

Conclusion

Summary

This week's work focused on preparing the dataset for reliable analysis through comprehensive data cleaning, feature engineering, and validation. Major issues such as inconsistent text entries, redundant columns, malformed dates, and structural missing values were resolved. Key features including applicant age, opportunity duration, completion status, and institutional participation counts were successfully engineered, resulting in a clean and analysis-ready dataset.