**Report Title**

**Week 1: Data Cleaning and Feature Engineering Report**

Intern’s Name: Marium Imran

Date of Submission: 22-06-2025

**Introduction**

**Purpose:**

To prepare the dataset for exploratory data analysis and modeling by performing thorough cleaning and creating meaningful features.

**Data Description:**

The dataset contains information about learners and their engagement

with opportunities.

Key columns include:

1. Learner Signup DateTime
2. Date of Birth
3. Opportunity Start Date
4. Opportunity Category
5. Status Description
6. Country
7. Gender
8. Major

**Data Cleaning Process**

### **Overview**

Data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in datasets to ensure the data is accurate, reliable, and suitable for analysis. This process improves data quality and enhances the effectiveness of any subsequent analytical work. For the SLU Opportunity Wise dataset, the data cleaning process included handling missing values, correcting formatting issues, standardizing categorical data, and removing errors and duplicates.

### **Dataset Overview**

| **Metric** | **Original Dataset** | **Cleaned Dataset** |
| --- | --- | --- |
| Number of Records (Rows) | 8,558 | 6,778 |
| Number of Columns | 16 | 17 *(Age column added)* |
| Columns with Missing Data | 2 | 0 |
| Total Null Values | 3,798+ | 0 |
| Duplicates | Present | Removed |

### **Key Aspects of Data Cleaning**

#### **1. Handling Missing Values**

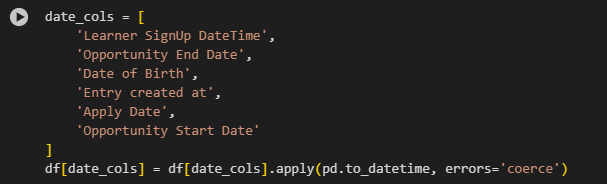
* **Original**: The Opportunity Start Date column had 3,794 missing values, and the Institution Name had 4.
* **Cleaned**: All missing values were handled by either imputing values or removing records with critical missing data.
* **Note**: Caution was exercised as some values were possibly omitted intentionally by users.

#### **2. Handling Outliers**

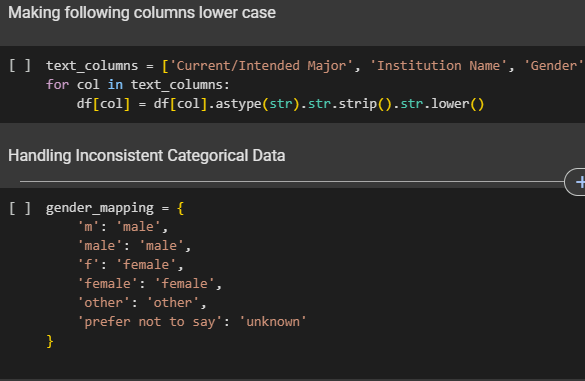
* Outliers and unrealistic values in fields such as Current/Intended Major were identified and removed.
* For example, gibberish entries like xxxhhyy were cleaned.
* Outlier removal helped bring the number of unique majors down from 407 to 355.

#### **3. Standardizing Formats**

* Standardization was applied to:
  + **City Names**: e.g., st. louis and Saint Louis unified.
  + **Date Formats**: Ensured consistency across Apply Date, Date of Birth, Opportunity Start Date, etc.



* + **Country Names and Institutions**: Variations in names due to typos or casing were corrected.



#### **4. Correcting Errors**

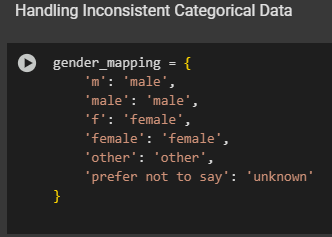
* Typographical errors and inconsistencies in textual entries were corrected.
* The number of unique values for fields such as Institution Name dropped from 2,090 to 1,618, indicating effective correction and consolidation.
* Categorical data entries were validated to align with expected standards.

#### **5. Dealing with Duplicates**

* Redundant records were removed, reducing the dataset size by **1,780 entries**.
* Ensured that each learner's entry was uniquely identified and represented.

#### **6. Handling Inconsistent Categorical Data**

* Categories were normalized. For instance:
  + Gender was restricted to 4 recognized values.
  + Status Description and Opportunity Category were made consistent



### 

### **Summary of Improvements**

| Feature | Before Cleaning | After Cleaning |
| --- | --- | --- |
| Data Completeness | Missing in 2 columns | 100% complete |
| Data Consistency | Many inconsistencies | Standardized |
| Format Uniformity | Mixed formats | Fully uniform |
| Erroneous Data | Present | Corrected |
| Duplicate Records | Present | Removed |
| Categorical Uniformity | Inconsistent | Normalized |

### **Conclusion**

The data cleaning process significantly enhanced the quality of the SLU Opportunity Wise dataset. By addressing issues related to missing values, outliers, formatting inconsistencies, and categorical discrepancies, the dataset is now ready for accurate, robust, and meaningful analysis. These cleaning steps ensure improved reliability and validity for any data-driven insights or machine learning models derived from this dataset.

**Feature Engineering**

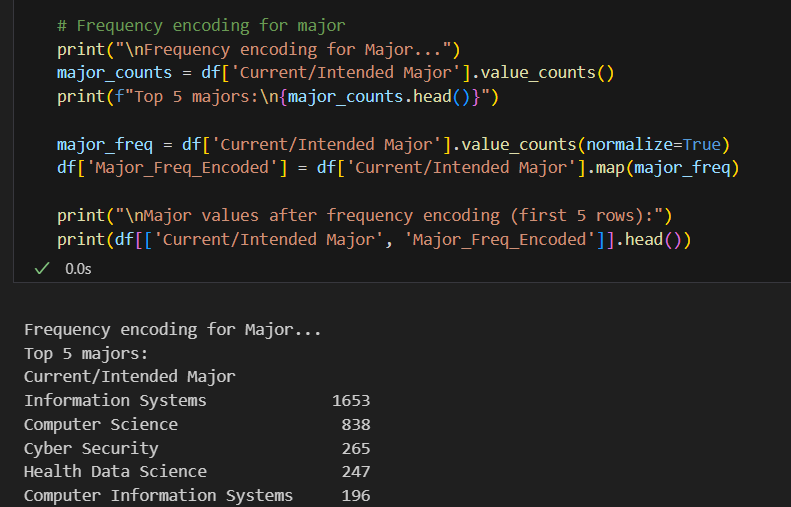
During the feature engineering process, several new features were created better to capture patterns in learner engagement and opportunity participation. These features aim to enhance the dataset's predictive power and prepare it for meaningful exploratory data analysis and modeling.

List of new features:

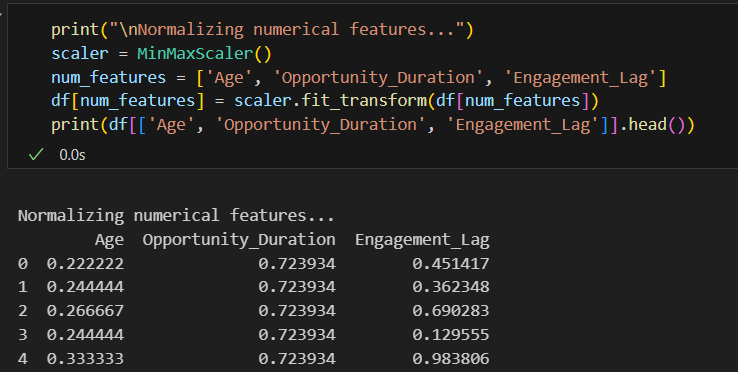
1. Age: calculated using the Date of Birth column. Learner age plays a crucial role in understanding patterns of engagement across different demographics.
2. Opportunity Duration: Measures the length(in days) of each opportunity, which could affect engagement.
3. Engagement Lag: Captures the time gap between application and opportunity start, indicating how early or late a learner applies.
4. Normalized Features: Scales numerical features to [0,1] range for better model performance and comparability.
5. Status Encoded: Converts categorical status into numerical codes for modeling.
6. One-hot Encoded Gender and Country: Converts categorical variables into binary columns for each unique value.
7. Major Frequency Encoded: Encodes the frequency of each major, capturing its popularity.
8. Temporal Analysis: Extract temporal patterns from dates for seasonality or trend analysis
9. Behavioral Feature: Categorize learners based on how quickly they sign up and end engagement timings.

Feature Examples:

1. **Frequency encoding** replaces each major with how often it appears in the dataset (as a percentage). This creates a numeric feature (Major\_Freq\_Encoded) that captures the popularity of each major for use in modeling.



1. **Normalization:** Scale numerical features to a range to ensure equal weightings in models.



**Data Validation**

**Summary:**

The dataset was validated to ensure accuracy and consistency through the following checks:

**1. Missing Value Check**

Each column in the dataset was examined for null or missing values. This helped identify incomplete records, which may affect the reliability of the analysis.

**2. Duplicate Records Check**

Duplicate rows were identified and reviewed to prevent data redundancy and ensure each entry is unique.

**3. Gender Validation**

The 'Gender' column was validated against acceptable values — 'male', 'female', and 'other'. Records containing any other values were classified as invalid and flagged for correction.

**4. Age Range Validation**

The 'Age' column was checked to identify unrealistic values. Entries with age less than 10 or greater than 100 were considered invalid and listed for review.

**5. General Consistency**

Additional checks were performed to ensure consistent capitalization, spacing, and formatting in string fields such as names and gender.

**Outcome:**

Several rows had unrealistic age values outside the accepted range.

Missing and duplicate values were reported, though no major issues were found. These validation steps ensured that the dataset was clean, reliable, and ready for further analysis.

**Conclusion**

During Week 1, the dataset underwent extensive cleaning and preprocessing to ensure accuracy, consistency, and readiness for analysis. Invalid dates were corrected, missing values were handled, and outliers were removed. Following this, feature engineering was applied to enrich the dataset by introducing meaningful variables such as **Age** and **Opportunity Duration**, derived from date fields. These features will help capture user engagement patterns and improve the predictive capacity of future models.

**Next Steps**

In Week 2, we will move forward with:

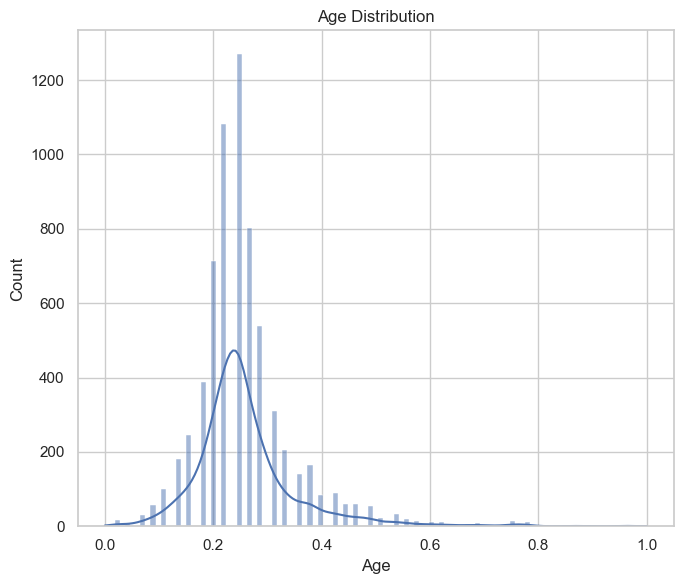
Exploratory Data Analysis (EDA): Identifying trends, distributions, and correlations in the dataset using visual and statistical methods.

The cleaned and feature-enhanced dataset from Week 1 will serve as the foundation for generating insights and building predictive models on learner engagement and opportunity outcomes.

**Appendix**

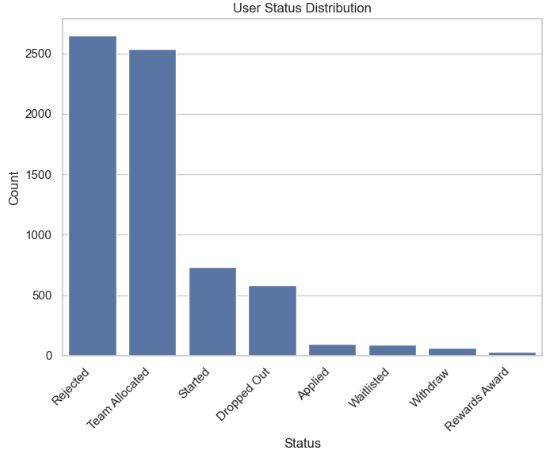
Age Distribution:

Shows a smoothed (KDE) and histogram view of age distribution, making trends and peaks easier to spot.



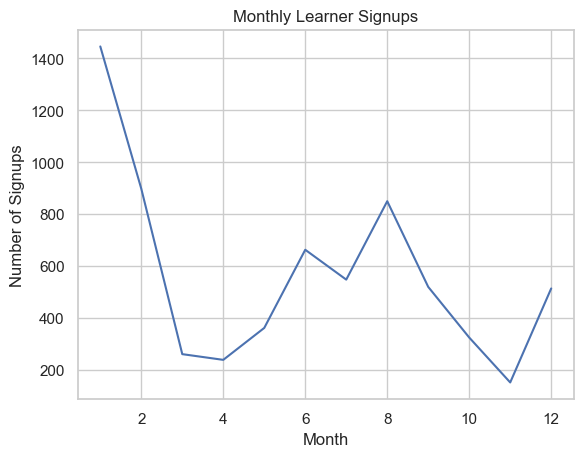
User Status Distribution shows:

The count of users in each status (e.g., Started, Waitlisted, Team Allocated)

****

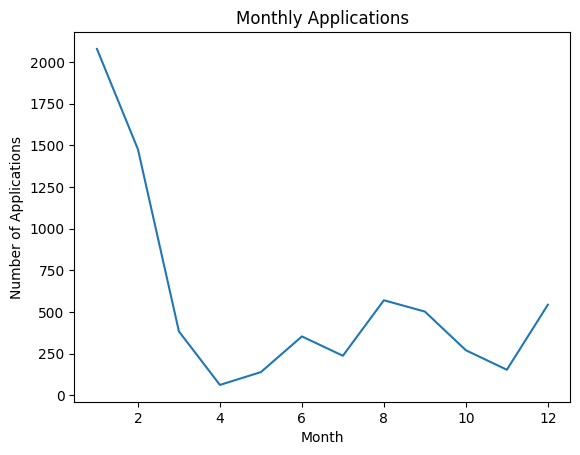
Monthly Learner SignUps Trend:

Peaks may indicate popular signup periods (e.g., start of academic terms). Dips may show off-seasons or less interest

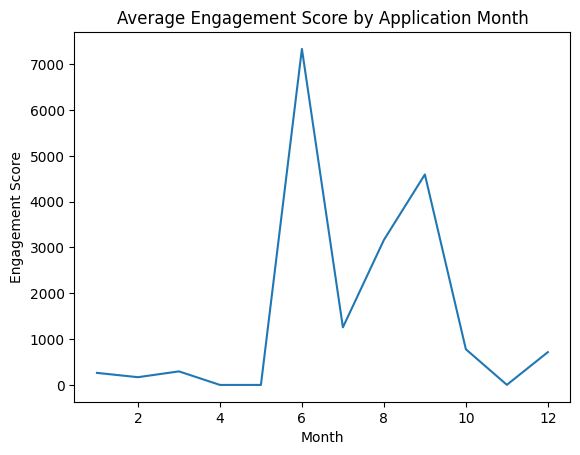


Monthly Application Trend:

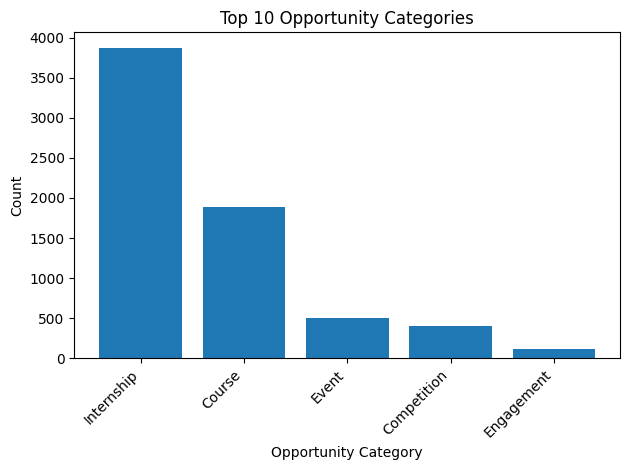
Compare with signups to see if applications follow the same pattern. Spikes may align with application deadlines or new opportunity postings.



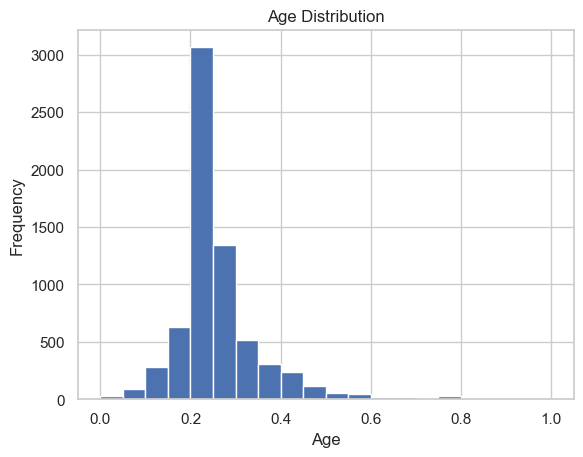
The trend shows the engagement scores of applicants per month



Opportunity Category Trend shows the categories of most Applicants

****

Distribution of Learner Age

****