# Week 1: Data Cleaning and Feature Engineering Report

## Title Page

**Title:** Week 1: Data Cleaning and Feature Engineering Report
**Intern's Name: Shahid Khan**
**Date of Submission: 2/17/2025**

---

## Introduction

### Purpose

This report outlines the data cleaning and feature engineering process undertaken during Week 1 of the internship. The primary goal was to preprocess the dataset for further analysis by handling missing values, standardizing data formats, and creating meaningful features.

### Data Description

The dataset used in this project contains records related to student engagement and churn analysis. It includes various attributes such as:

- **Learner SignUp DateTime** – Timestamp of when a learner signed up.
- **Opportunity Details** – Name, Category, Start and End Date.
- **Personal Information** – First Name, Date of Birth, Gender, Country.
- **Institution Details** – Name, Current/Intended Major.
- **Status Details** – Application Date, Status Code, and Description.

The dataset consists of **8,558 entries** and **16 columns**, with missing values in some fields.

---

## Data Cleaning Process

### Cleaning Steps

1. **Handling Missing Values:**

    - Filled missing values in *Institution Name* and *Current/Intended Major* with the mode.

    - Forward-filled missing values in *Opportunity Start Date* and *Opportunity End Date*.
    - Replaced missing values in *Engagement Duration* with the mean.
2. **Standardizing Formats:**

    - Converted all text fields (e.g., *First Name, Institution Name*) to title case.
    - Standardized date columns to *datetime* format.
3. **Removing Inconsistencies:**

    - Identified and fixed entries containing non-alphabetic characters in *First Name* and *Institution Name*.
    - Stripped unnecessary spaces in categorical fields.
4. **Duplicate Handling:**

    - No duplicate entries were found; hence, no removal was necessary.

### Issues Encountered

- **Formatting Errors:** Some date values were incorrectly formatted, requiring conversion.
- **Missing Data in Date Columns:** Forward filling was applied to ensure consistency.

---

# Feature Engineering

### New Features Created

1. **Age of Learner –** Computed from *Date of Birth*.
2. **Engagement Duration –** Difference between *Apply Date* and *Opportunity Start Date*.
3. **SignUp Month & Year –** Extracted from *Learner SignUp DateTime*.
4. **Time in Opportunity –** Difference between *Opportunity Start Date* and *Opportunity End Date*.
5. **Engagement Score –** Weighted score using *Time in Opportunity, Age of Learner*, and *Opportunity Category*.
6. **SignUp Day of Week –** Extracted from *Learner SignUp DateTime*.

### Feature Examples

**Age Calculation:**

```
 SLU['Age of Learner'] = (pd.to_datetime(‘today’) - SLU['Date of Birth']).dt.days // 365
```

-

**Engagement Score Computation:**

```
 w1, w2, w3 = 0.5, 0.3, 0.2
SLU['Engagement Score'] = (w1 * SLU['Time in Opportunity'] + w2 * SLU['Age of Learner'] +
w3 * SLU['Opportunity Category Encoded'])
```

-

---

# Data Validation

## Validation Summary

- **Checked for missing values** after preprocessing and ensured none remained.
- **Ensured consistency** by verifying the proper format of dates and categorical values.
- **Confirmed data integrity** by ensuring new features aligned logically with existing ones.

## Validation Checks

- **Missing Values:** No null values post-cleaning.
- **Data Type Verification:** Ensured numeric fields were correctly formatted.
- **Outlier Detection:** No extreme inconsistencies detected in key features.

---

# Conclusion

## Summary

During Week 1, the dataset was successfully cleaned, formatted, and enhanced with additional features to improve analytical insights. The processed dataset is now structured for exploratory data analysis and modeling.