Week 1: Data Cleaning and Feature Engineering Report

Pranav Verma

January 20, 2025

1. Introduction

The dataset used in this project contains information related to opportunities, learner signups, and related attributes. The primary objective of this phase was to clean the dataset by handling missing values, removing duplicates, and standardizing formats. Additionally, new features were engineered to facilitate further analysis.

2. Data Cleaning Process

- **Handled Missing Values**: Removed rows with missing critical fields such as *Date of Birth*, *Learner SignUp DateTime*, *Opportunity Start Date*, and *Opportunity End Date*.
- **Removed Duplicates**: Eliminated redundant records to ensure data uniqueness and accuracy.
- **Standardized Fields**:
  - Normalized city and major names for consistency.
  - Standardized date columns to a uniform datetime format.
- **Corrected Formatting Issues**:

  - Trimmed and capitalized text fields such as *Current/Intended Major* and *Country*.

3. Feature Engineering

New features were derived to enrich the dataset:

- **Age**: Computed based on *Date of Birth*.
- **Opportunity Duration**: Measured the number of days between *Opportunity Start Date* and *Opportunity End Date*.
- **Signup Month**: Extracted the month from *Learner SignUp DateTime*.

4. Data Validation

- **Missing Values**: Confirmed no missing values in critical fields after cleaning.
- **Date Formats**: Verified consistency across all date columns.
- **Outliers**: Checked for and resolved any irregularities in numeric or date-based fields.