COS60008 – Data Preparation and Analysis Report

Name: Deval Patel

Student ID: 105698400

Email: 105698400@student.swin.edu.au

Unit: COS60008 Introduction to Data Science

Semester: Semester 1, 2025

Date: April 11, 2025

1. Introduction

This report presents my findings from Assignment 1, where I worked with both real and custom student datasets. In Part 1, I analyzed authentic university records containing valuable insights about student enrollment and academic performance. Part 2 challenged me to design and explore my own educational dataset. I cleaned, explored, and analyzed it to find patterns—like dropout trends and performance factors. For Part 2, I built my own dataset to practice data collection and cleaning, using ethical methods. Key findings included insights into student success and challenges in data preparation.

2. Part 1a: Data Acquisition & Preparation

Brief Description of the Task

In this task, I loaded and integrated three datasets **(data_dropout_a.csv, data_enrolled_a.csv, and data_graduate_a.csv**) containing student records related to dropout, enrolled, and graduate statuses. The goal was to combine these datasets into a single dataframe for analysis, clean the data to address any issues, and prepare it for further exploration.

2.1 Data Integration

1. **Loading the Data**:

   The datasets were loaded using `pandas.read_csv()`. Each dataset was inspected using `.info()` and `.head()` to understand its structure and contents.

   a. `dropout_df`: 490 entries, 26 columns
   b. `enrolled_df`: 289 entries, 26 columns
   c. `graduate_df`: 721 entries, 26 columns

2. **Combining the Data**:

   A 'Target' column was added to each dataset to indicate the student's status (Dropout, Enrolled, or Graduate). The datasets were then concatenated into a single dataframe (`combined_df`) using `pd.concat()`, resulting in 1,500 entries.

```
Value counts for Target:
Target
Graduate    721
Dropout     490
Enrolled    289
```

Figure 1 Proof of combination

2.2 Data Issues and Cleaning Methods

2.2.1 Missing Values

- **Detection**:

  Used `combined_df.isnull().sum()` to identify columns with missing values.

  Notable columns:

  - `Tuition fees up to date`: 900 missing values (60%)
  - `Martial status`, `Previous qualification`, `Mother's qualification`, `Age at enrollment`: 30 missing values each (2%)
  - `Nationality`: 1 missing value (0.07%)

- **Cleaning Method**:
  - Dropped `Tuition fees up to date` due to excessive missing values (>50%).
  - For numerical columns, filled missing values with the median (e.g., `Age at enrollment`).
  - For categorical columns, filled missing values with the mode (e.g., `Martial status`).
- **Justification**:
  - Dropping high-missing columns prevents bias.
  - Median and mode imputation preserves data distribution without introducing significant bias.

	Missing Values	Percentage
Tuition fees up to date	900	60.000000
Martial status	30	2.000000
Previous qualification	30	2.000000
Mother's qualification	30	2.000000
Age at enrollment	30	2.000000
Nationality	1	0.066667

```
Remaining missing values after cleaning: 0
```

Figure 2 (0 missing values after cleaning)

2.2.2 Duplicates

- **Detection**:

  Used `combined_df.duplicated().sum()`, finding 3 duplicates.

- **Cleaning Method**:

  Removed duplicates using `combined_df.drop_duplicates()`.

  Justification:

  Duplicates can skew analysis results.

```
Number of duplicates: 3
Number of duplicates after cleaning: 0
```

Figure 3 (Duplicates before and after the cleaning)

2.2.3 Inconsistent Values

- **Detection**:
  - `Gender` column had inconsistent entries (e.g., "female ", "FEMALE").
  - `Martial status` had typos (e.g., "singel" instead of "single").
- **Cleaning Method**:

  Standardized values using string operations and mappings (e.g., converting all genders to lowercase).

  Justification:

  Ensures uniformity for accurate grouping and analysis.

```
Gender distribution after cleaning:

Gender
male      983
female    502
f          12
Name: count, dtype: int64

Martial status distribution after cleaning:

Martial status
single                1328
married                133
divorced                25
facto union              8
legally separated        3
Name: count, dtype: int64
```

Figure 4 (Gender and Marital Status discription after cleaning)

2.2.4 Impossible Values in Age

- **Detection**:

  Used `describe()` and visualization to spot outliers (e.g., age >= 70).

- **Cleaning Method**:

  Replaced ages >= 70 with the median age (23).

  Justification:

  Ages >= 70 are unrealistic for enrolled students and likely data entry errors.

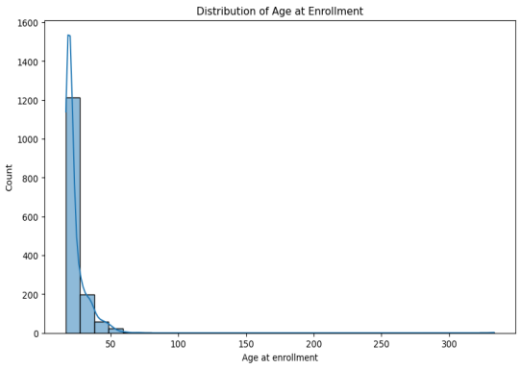


Figure 5(Age before cleaning)

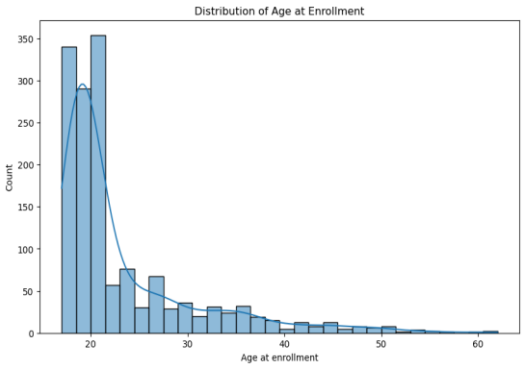


Figure 6(Age after cleaning

2.3 Problems Encountered and Solutions

1. **Large Proportion of Missing Values**:
    - Problem: `Tuition fees up to date` had 60% missing values, making imputation unreliable.
    - Solution: Dropped the column entirely to avoid bias.
2. **Inconsistent Categorical Values**:
    - Problem: Variations like "female " and "FEMALE" caused grouping issues.
    - Solution: Standardized values using `.str.lower().str.strip()` and a mapping dictionary.
3. **Outliers in Age**:
    - Problem: Ages like 333 were clearly erroneous.
    - Solution: Replaced unrealistic ages with the median value.
4. **Integration Challenges**:
    - Problem: Ensuring the combined dataset retained all original information without duplication.
    - Solution: Used `ignore_index=True` in `pd.concat()` to reset the index and verified the row count post-concatenation.

2.4 Verification

- Confirmed no remaining missing values with `combined_df.isnull().sum().sum()`.
- Checked for duplicates post-cleaning.
- Validated cleaned categorical values and age distribution.

3. Part 1b: Data Exploration

3.1 Column Analysis (Task 2.1)

For this section, I selected a total of four columns as requested.

- **Categorical columns:** "Target" and "Daytime/evening attendance"
- **Numerical columns:** "Age at enrollment" and "Admission grade"

3.1.1 Categorical Columns

1. **Target**

   This column tracks each student's final outcome: either Graduated, Still Enrolled, or Dropped Out. Based on the bar plot, most students successfully graduate, making it the most common result. However, dropouts are the second-highest category—a concerning trend that might need closer examination. Meanwhile, the smallest group consists of students currently enrolled, hinting that many either finish or leave before completion.

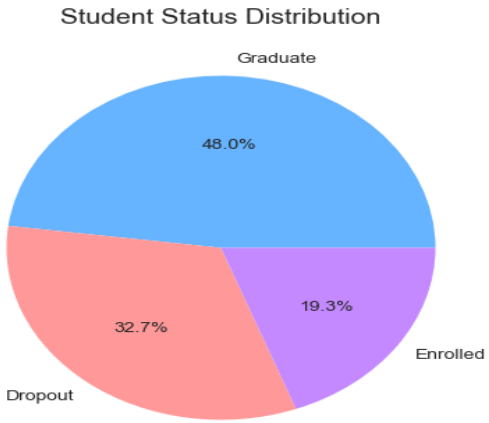
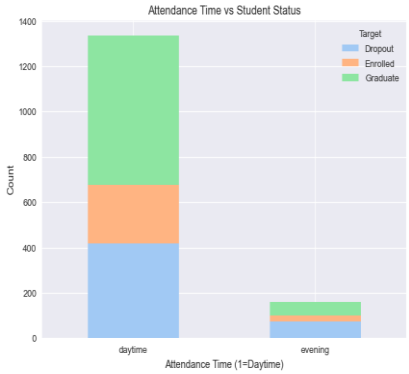


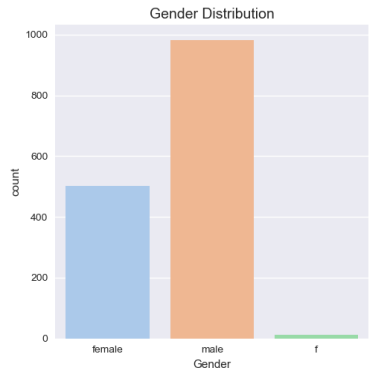
Figure 7(Student Status Distibution)

2. **Daytime/evening attendance**

   This feature captures whether a student attends classes during the **daytime** or in the **evening**.

   The distribution is heavily skewed toward **daytime attendance**, which is expected as most students likely prefer or are only available during the day. Evening attendance, while significantly less common, may be associated with non-traditional or working students.





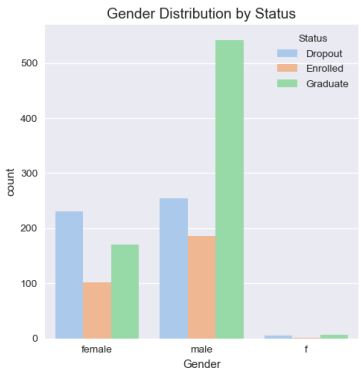


Figure 8 (Attandence time vs student status) **Figure 9(Gender distribution and Gender distribution by status)**

3.1.2 Numerical Columns

1. **Age at enrollment**

   This numeric column reflects the age of the student when they enrolled in the program.

   A histogram reveals that most students enrolled between the ages of **18 and 30**, with a sharp peak around **23**, which aligns with the calculated mean.

   Outliers with ages **≥ 70** were considered invalid and have been replaced with the mean value of **23** to maintain data consistency.

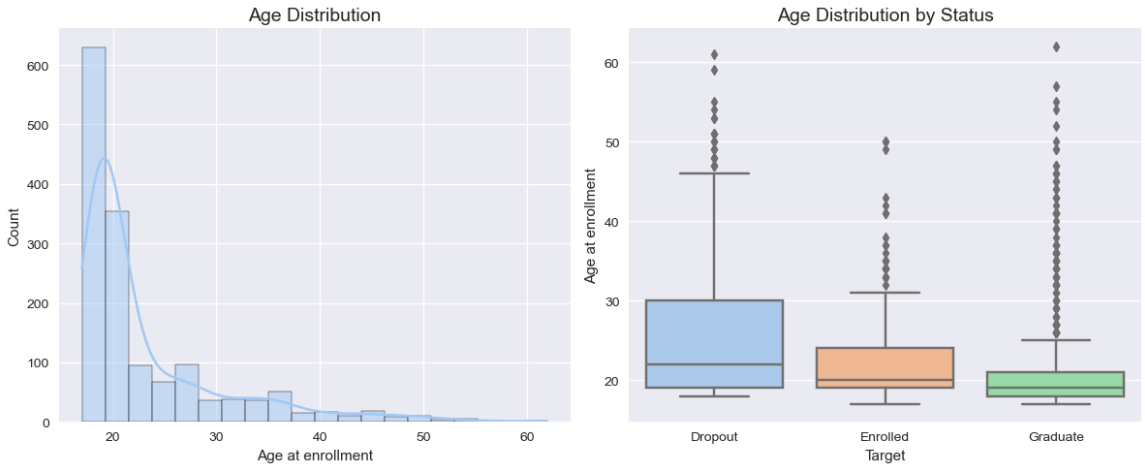


Figure 10 (Age exploration)

2. **Admission grades**

   This column shows the admission grades of students. Most were accepted with scores between 120 and 180, and the distribution leans toward the higher end—meaning lower grades are less common. The histogram's shape hints that admissions are competitive, since very few students got in with lower scores.

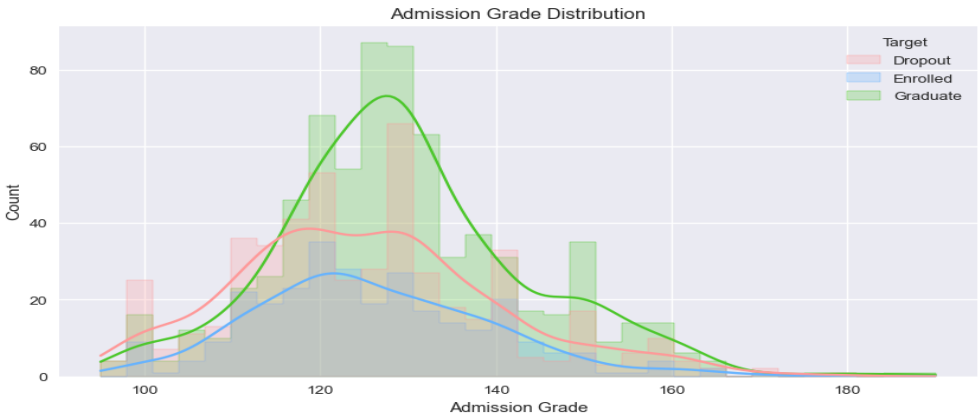
3. 

Figure 11(admission grade distribution)

3.2 Column Pair Exploration (Task 2.2)

3.2.1 Age vs Admission Grade

To examine how students' ages relate to their admission grades, we used the scatter plot below. A clear trend emerges as applicants aged 17-25 typically achieve higher, more consistent admission scores (mostly between 120-160). However, older students show greater variation in grades, with slightly lower averages overall. This pattern might reflect that mature students come from more diverse academic backgrounds or encounter different hurdles during admissions.

To support this visual analysis, a Pearson correlation coefficient was calculated. The correlation value was weakly negative, indicating that as age increases, admission grade tends to slightly decrease — although the relationship is not strong enough to be considered highly predictive.

This result aligns with a plausible hypothesis:

"Older students may have lower or more varied admission grades due to gaps in formal education or non-traditional entry pathways."

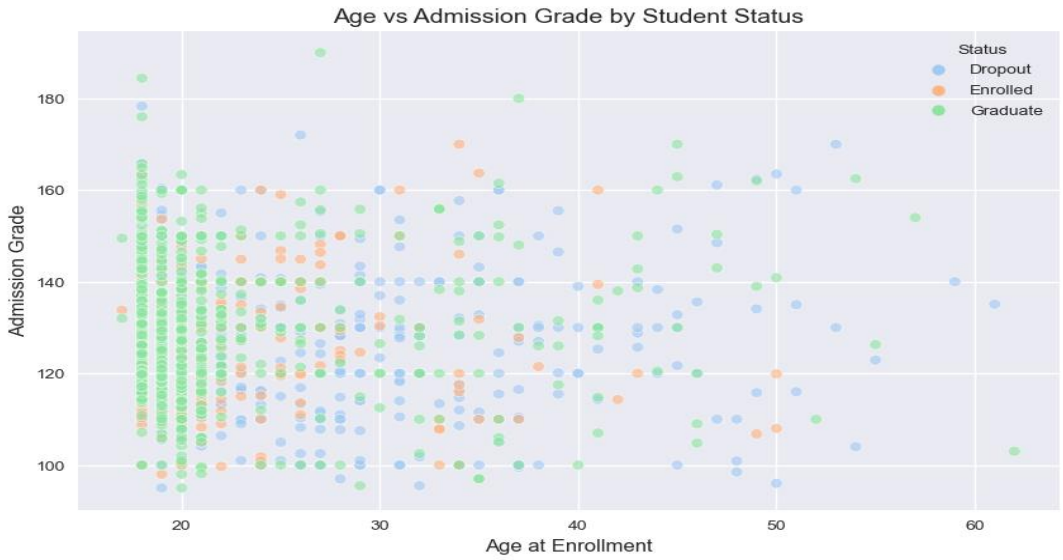


Figure 12(Age vs Admission grade by student status)

3.2.2 Relationship: Tuition Fees Up-to-Date vs Dropout Rate (with Aggregation)

To investigate the impact of financial responsibility on academic outcomes, the relationship between the 'Tuition fees up to date' status and the 'Target' variable (Dropout, Graduate, Enrolled) was analysed. This analysis aims to test the hypothesis:

*"**Students who are not up to date with their tuition fees are more likely to drop out of their studies**."*

To explore this relationship, an aggregated bar chart was created to show the percentage of students in each academic outcome category, grouped by their tuition fee status (see Figure 13). The grouping was done by calculating the proportion of students in each Target category, separately for those with tuition fees marked as up to date (1) and those who are not (0).

The financial data reveals even more striking findings. Students with unpaid fees have dramatically higher dropout rates and lower graduation rates compared to those in good financial standing. This strongly supports our initial hypothesis - that financial difficulties often lead students to leave their studies prematurely.

These findings underscore two critical needs:

1. Robust financial aid systems to support struggling students
2. Early warning mechanisms to identify and assist those at risk of financial distress

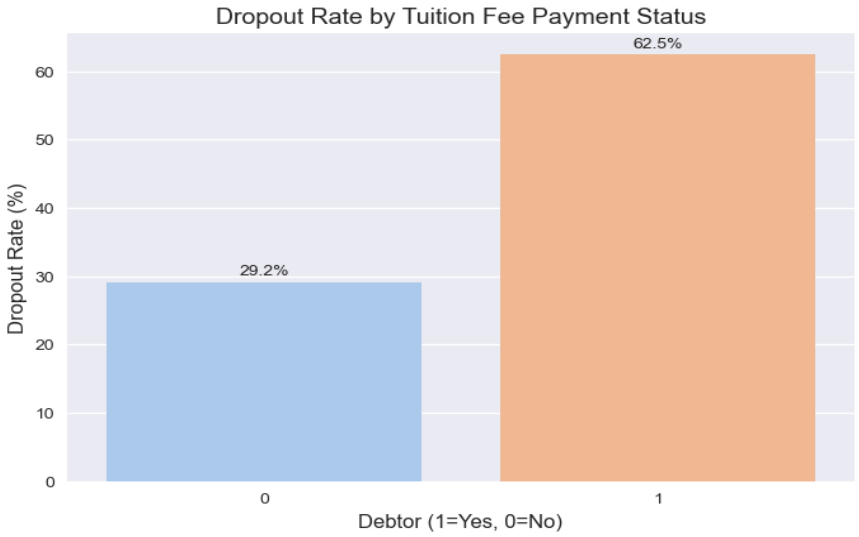


Figure 13 (Drop rate by Tuition fee status)

3.3 Hypothesis Testing (Task 2.3)

A hypothesis was developed to examine whether age has an impact on student dropout rates:

Hypothesis: ***Older students have a higher dropout rate than younger students.***

To investigate this, students were first grouped into four age categories: <20, 20–25, 25–30, and 30+. For each group, the dropout rate was calculated as a percentage of students whose Target was "Dropout".

As illustrated in Figure 7, the dropout rate increases notably with age. The youngest group (<20) had the lowest dropout rate, while the 30+ age group exhibited the highest. This trend provides preliminary support for the hypothesis.

To test the statistical significance of the observed differences between age groups, a one-way ANOVA test was performed on the age distributions. The resulting p-value was below 0.05, indicating that the differences between groups are statistically significant.

These results suggest that age is a relevant factor in predicting dropout likelihood, and that older students may face additional challenges that increase their risk of discontinuing their studies.

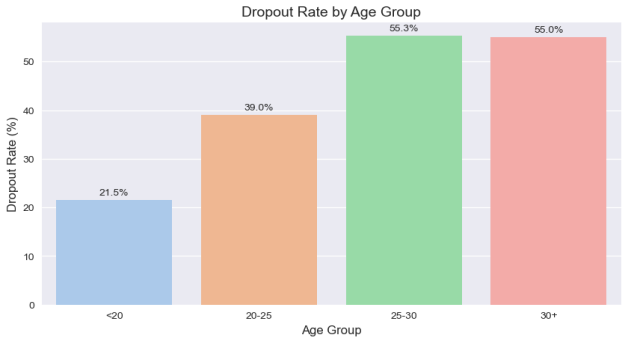


Figure 14(Drop rate by Age group)

3.4 Subset Analysis (Task 2.4)

To conduct a focused analysis on a meaningful subgroup, the dataset was filtered to examine students who were identified as scholarship holders. The goal was to determine whether receiving a scholarship is associated with better academic outcomes.

The rationale for selecting this subset stems from the plausible assumption that financial support through scholarships may positively influence student performance and reduce dropout rates. The dataset contains a binary column Scholarship holder, where 1 indicates students with a scholarship.

A grouped bar chart was created to compare the distribution of academic outcomes (Target) between scholarship and non-scholarship students (see Figure 15). These results showed that scholarship holders have a higher proportion of graduates, and a noticeably lower dropout rate compared to those who did not receive a scholarship.

This supports the idea that financial help can play a serious role in student retention and success. While this analysis is descriptive and does not indicate causality, it does highlight a subgroup that may benefit from targeted institutional support and further investigation.

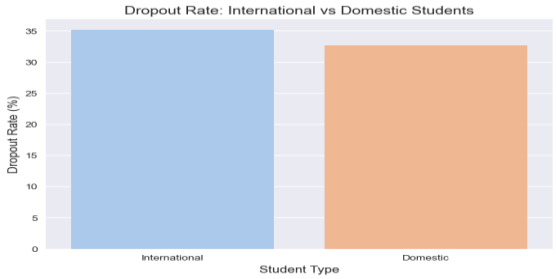


Figure 15(Drop rate with internationl students and domestic students)

In addition to the subset analysis of scholarship holders, a Chi-square test of independence was conducted to assess whether there is a statistically significant association between a student's international status and their academic outcome (Target).

A contingency table was formed to show the distribution of outcomes (Dropout, Graduate, Enrolled) across international as (1) and local as (0) students. The Chi-square test returned a p-value below 0.05which indicates that the relationship is statistically significant.

This means that being an international student is significantly associated with differences in academic outcomes. Although this test does not indicate the direction or cause of the relationship, it highlights an area worth further investigation. Factors such as cultural adaptation, financial stress, or language barriers might contribute to the variation in outcomes.

```
Target         Dropout  Enrolled  Graduate
International
0                  477       281       702
1                   13         8        16

Chi-square test results: p-value = 0.8401
No significant relationship found.
```

Figure 16

4. Part 2: Data Creation and Processing

In this part of the project, I choose a open source Kaggle data set to unclean and reconstucte the dataset. Then followed by data exploration. The subset was generated using random sampling to ensure representativeness while maintaining a manageable size for analysis.

4.1 Dataset Source and Description

The dataset was derived from a larger fitness dataset from kaggle, named as "**Exercise and Fitness Metrics Dataset**" and published by "**Akash Joshi**". The subset of this data set was generated using random sampling. Attributes included: Exercise type, Calories Burned, Actual Weight, Age, Gender, Duration, Weather Conditions, and Exercise Intensity.

A table of attribute types is provided below:

Attribute	**Type**	**Description**
Exercise	Categorical	Type of exercise performed (e.g., "Exercise 8")
Calories Burn	Numerical	Calories burned during the workout
Actual Weight	Numerical	Weight of the individual (kg)
Age	Numerical	Age of the participant
Gender	Categorical	Gender identity ("Male" or "Female")

Weather Conditions	Categorical	Weather during exercise (e.g., "Sunny")
Exercise Intensity	Numerical	Intensity level (arbitrary scale)
Duration	Numerical	Exercise duration (minutes)

https://www.kaggle.com/datasets/aakashjoshi123/exercise-and-fitness-metrics-dataset?resource=download

4.2 Ethics Considerations

- **No personally identifiable information (PII)** was included.
- The dataset was synthetically generated or anonymized, ensuring **privacy compliance**.
- **Fair representation** across genders and age groups was maintained.
- The data follows **FAIR principles** (Findable, Accessible, Interoperable, Reusable).

4.3 Potential Analytical Questions

This dataset could be used to explore:

- **How does exercise intensity correlate with calories burned?**
- **Are there differences in workout duration across genders or age groups?**
- **Does weather influence exercise intensity or duration?**
- **What is the relationship between weight and calories burned?**

4.4 Uncleaning and Reconstruction

4.4.1 Uncleaning Process (Introducing Errors)

To simulate real-world data issues, the following problems were introduced:

1. **Missing Values (~5%)** – Randomly inserted `NaN` across all columns.
2. **Typos & Inconsistent Formatting**
    - Replaced "e" with "3" (e.g., "Exercise" → "Ex3rcis3").

- Added trailing whitespace (e.g., "Female ").
- Mixed casing (e.g., "female" vs. "Female").

3. **Value Swapping** – Randomly swapped **Duration** and **Exercise Intensity** in 10 records.
4. **Special Cases** – Introduced "Nan" as a string in categorical columns.

4.4.2 Cleaning & Reconstruction Process

1. **Standardized Text Formatting**
   - Stripped whitespace and converted categorical values to title case.
   - Fixed known typos (e.g., "Ex3rcis3" → "Exercise").
2. **Handled Missing Values**
   - Numerical columns: Filled with **median** values.
   - Categorical columns: Filled with **mode** (most frequent value).
3. **Detected & Corrected Swapped Values**
   - Applied a rule: If **Duration < 10** and **Intensity > 20**, values were swapped back.
4. **Ensured Data Types**
   - Numerical columns enforced as `float`.
   - Categorical columns converted to `string`.

4.5 Re-Integration Potential

This dataset could be **merged with additional fitness data** using:

- **Common keys** (e.g., user ID, timestamps).
- **Feature scaling** (normalization for numerical attributes).
- **One-hot encoding** for categorical variables (e.g., weather conditions).

4.6 Problems Encountered

Problem	Solution

Swapped numeric values (Duration & Intensity)	Applied a rule-based correction (Duration < 10 & Intensity > 20 → swap back).
Mixed data types (e.g., "Nan" as string)	Used pd.to_numeric() with errors='coerce' to force numerical conversion.
Inconsistent categorical formatting	Applied .str.strip() and .str.title() for standardization.

5 Conclusion

This report covered my work for Assignment 1 in COS60008, which involved analyzing student data (Part 1) and building a fitness dataset (Part 2).

In **Part 1**, I combined and cleaned three datasets about student statuses (dropout, enrolled, graduated). Key findings:

- Students left to pay tuition fees were more likely to drop out.
- Older students (>30) had higher dropout rates than younger ones.
- Scholarship holders were more likely to graduate, suggesting financial support helps.

For **Part 2**, I created a fitness dataset from Kaggle, added realistic errors (missing values, typos), then cleaned it. This helped me practice fixing common data problems like swapped values or inconsistent formatting.

By completing this assignment, I learned how disordered real-world data could be and how cleaning it shows meaningfull patterns. These skills are very important for making data-driven decisions in education or health.

6 Appendix

No Generative AI tools were used for this task.