AI - Team 1

Exploratory Data Analysis (EDA) Report
Week - 2

Introduction

Exploratory Data Analysis (EDA) is a crucial step in understanding and preparing data for further analysis and AI-driven insights. It involves examining the dataset's structure, identifying key variables, detecting missing values and outliers, and uncovering patterns that can influence decision-making. This report provides a detailed analysis of the given dataset, highlighting key trends, relationships, and potential data challenges. By applying EDA techniques, we ensure that the dataset is clean, well-structured, and ready for advanced modeling and visualization.

Data Understanding and Exploration

The dataset contains **8,558 records** and **28 columns**, including categorical, numerical, and date-time fields. Key observations:

- **Numerical Variables**: Age, Opportunity Duration (days), Engagement Score, and Repeat Opportunities.
- **Categorical Variables**: Gender, Country, Institution Name, Opportunity Name, and Status Description.
- **Date-Time Fields**: Learner SignUp DateTime, Apply Date, Opportunity Start & End Date.

1. 1 Issue Identified

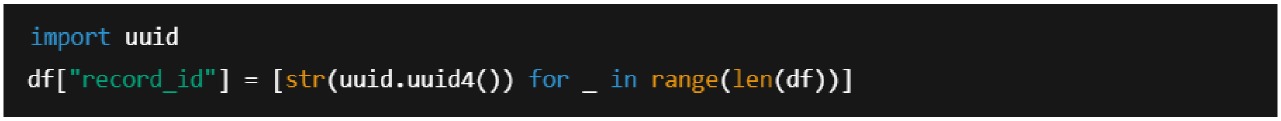
The existing Opportunity_Id was not unique, making it unsuitable as a primary key.

1. 2 Solution Implemented

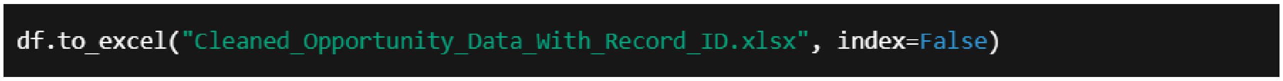
Generated a new column, record_id, using UUID v4 (uuid4()), which provides a random, globally unique identifier for each row. Ensured no data loss by preserving all 8,558 rows while adding record_id.

Why UUID v4?

- Universally Unique – Eliminates collision risks.
- Non-Sequential & Secure – Cannot be guessed or easily manipulated.
- Scalable – Suitable for future data expansions without conflicts.



```
 import uuid
df["record_id"] = [str(uuid.uuid4()) for _ in range(len(df))]
```



```
df.to_excel("Cleaned_Opportunity_Data_With_Record_ID.xlsx", index=False)
```

2. 1 Issue Identified

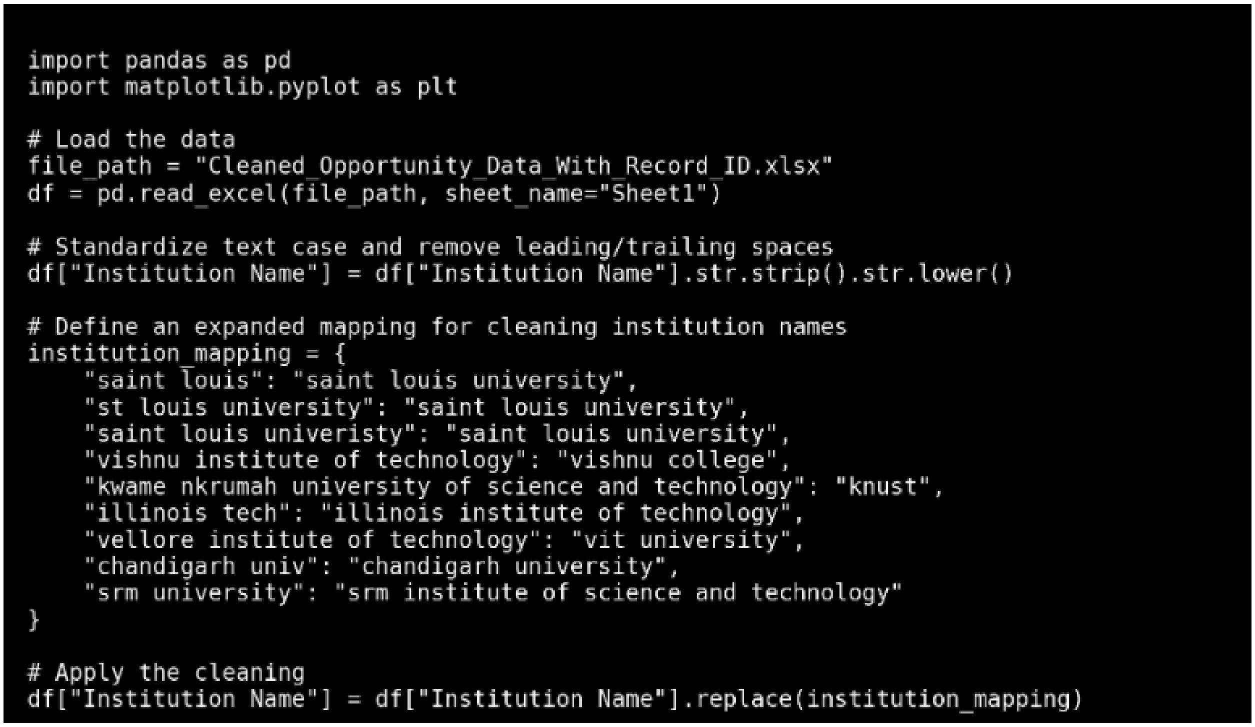
Identified conflicting data : Different values for the same entity in multiple columns (e.g., Saint louis University → Saint louis)

Saint Louis University appears as:

- "saint louis university" (4211 times)
- "saint louis" (88 times)
- "st louis university" (69 times)
- "saint louis univeristy" (25 times, spelling error)

2. 2 Solution Implemented

Detected these variations and cleaned them using Python. The records with the highest inconsistency were cleaned using Python, while the rest, which had fewer inconsistencies, were cleaned using the replace values feature in Excel.



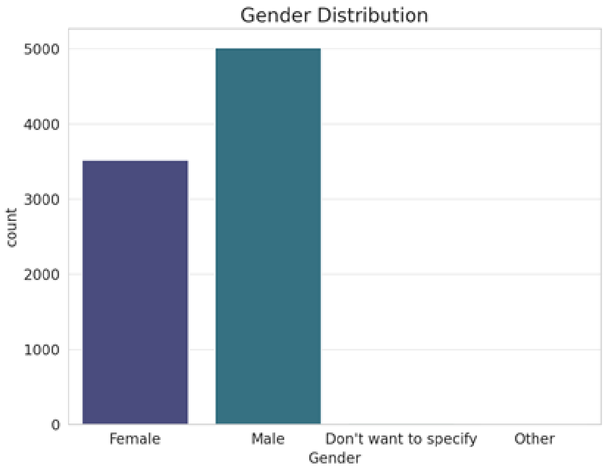
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
file_path = "Cleaned_Opportunity_Data_With_Record_ID.xlsx"
df = pd.read_excel(file_path, sheet_name="Sheet1")

# Standardize text case and remove leading/trailing spaces
df["Institution Name"] = df["Institution Name"].str.strip().str.lower()

# Define an expanded mapping for cleaning institution names
institution_mapping = {
    "saint louis": "saint louis university",
    "st louis university": "saint louis university",
    "saint louis univeristy": "saint louis university",
    "vishnu institute of technology": "vishnu college",
    "kwame nkrumah university of science and technology": "knust",
    "illinois tech": "illinois institute of technology",
    "vellore institute of technology": "vit university",
    "chandigarh univ": "chandigarh university",
    "srm university": "srm institute of science and technology"
}

# Apply the cleaning
df["Institution Name"] = df["Institution Name"].replace(institution_mapping)
```

Cleaned dataset : Cleaned_Opportunity_Data_With_Record_ID

Data Visualization

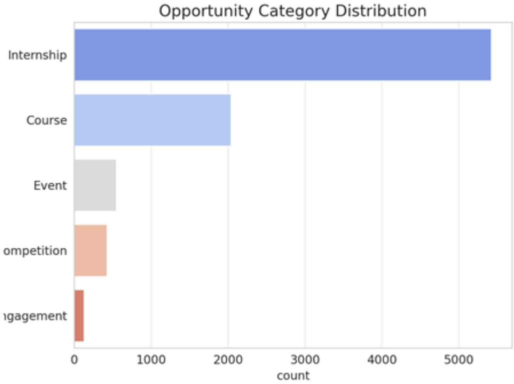
1. Gender Distribution:

- Created a bar chart to show the distribution of learners based on gender.
- Helps in understanding the gender diversity within the dataset.



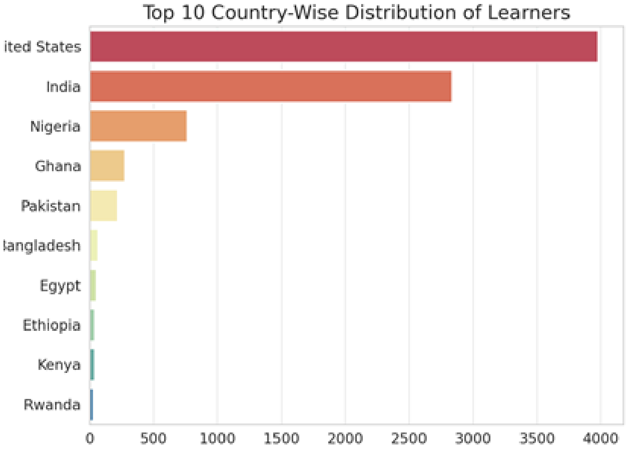
2. Opportunity Category Distribution:

- Visualized the distribution of various opportunity categories using a horizontal bar chart.
- Highlights the most common categories in the dataset, giving insights into popular opportunity types.



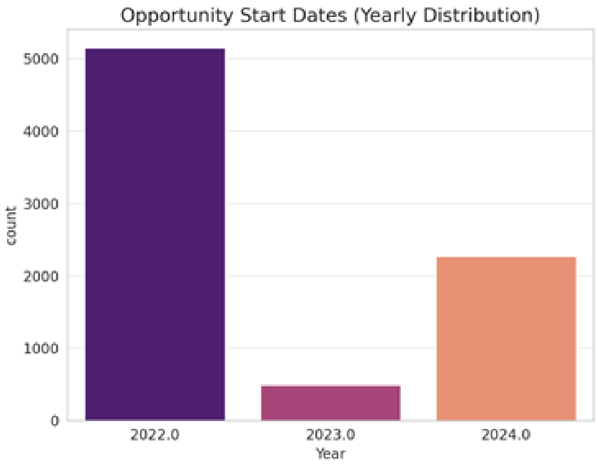
3. Top 10 Country-Wise Distribution of Learners:

- Generated a bar chart for the top 10 countries with the highest number of learners.
- Provides a clear comparison of the learner count across different countries.



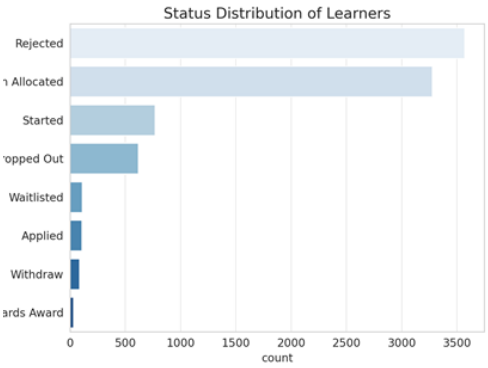
4. Opportunity Start Dates (Yearly Distribution):

- Created a count plot to show how opportunities are distributed over the years.
- Useful for analyzing trends and identifying peak years for opportunities.



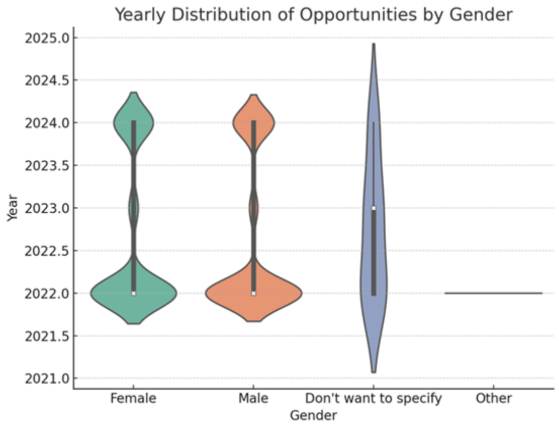
5. Status Distribution of Learners:

- Displayed a count plot of different learner statuses to show how many learners fall into each category.
- This helps in identifying how many learners are active, completed, or have other statuses.



6. Yearly Distribution of Opportunities by Gender:

- Displayed a violin plot to compare the distribution of **opportunity start years** across different genders.
- This helps in understanding how the opportunities are spread over time for **males and females**, highlighting any notable differences or patterns in their distribution.



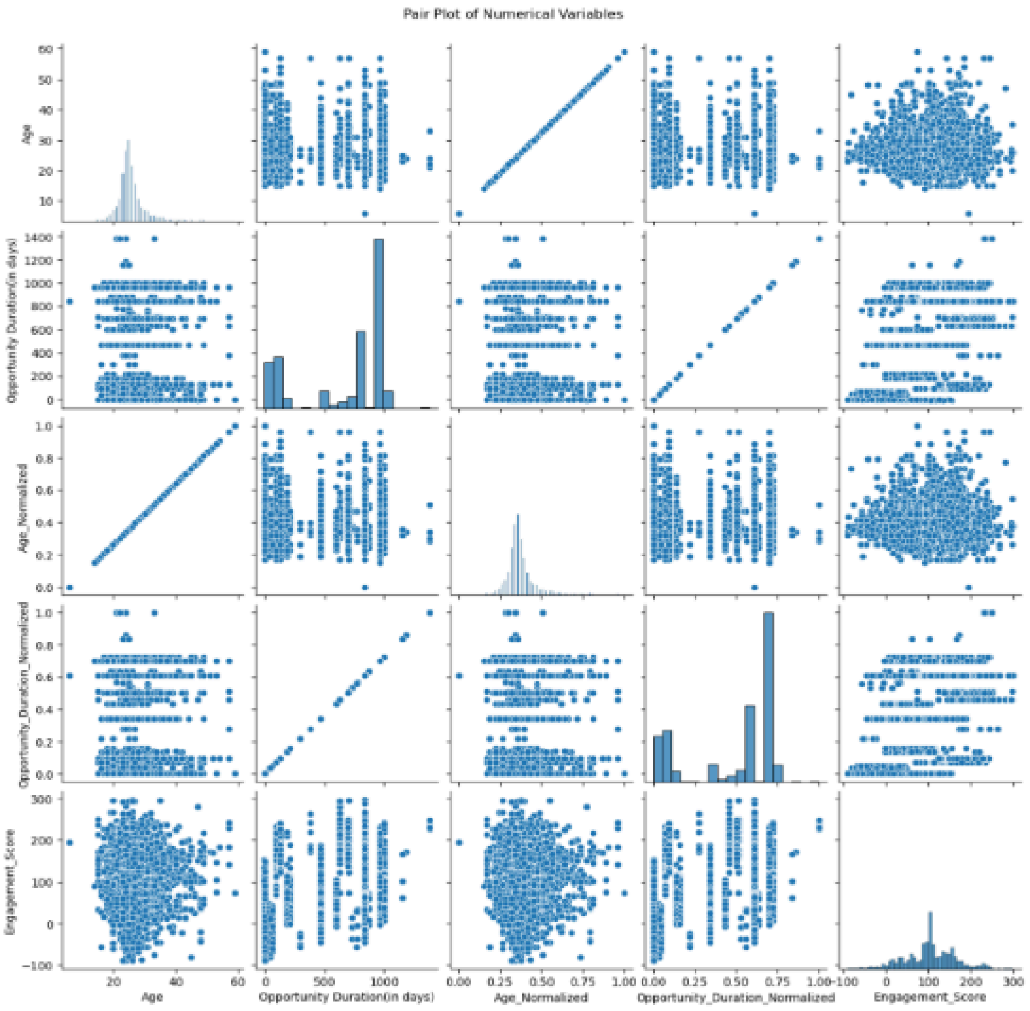
7. Pair Plot: Numerical Variables

• The pair plot shows relationships between age, opportunity duration, and engagement score.

• No strong linear trends are visible between any variable pairs.

• Engagement scores vary widely across all ages and durations.

• Diagonal histograms reveal most learners are in their early 20s with varied engagement.



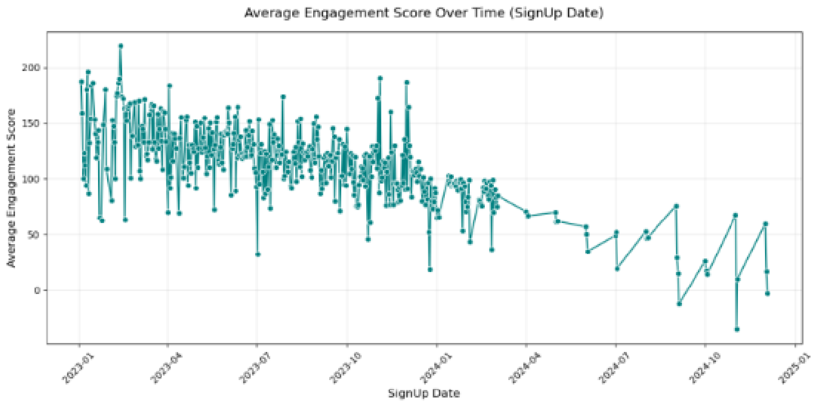
8. Time Series Analysis: Average Engagement Score Over SignUp Date

- The line plot shows the average engagement score over signup dates.

- Engagement scores fluctuate over time with no consistent trend.

- Some dates show higher average engagement, while others dip lower.

- The plot highlights variability in engagement across signup periods.

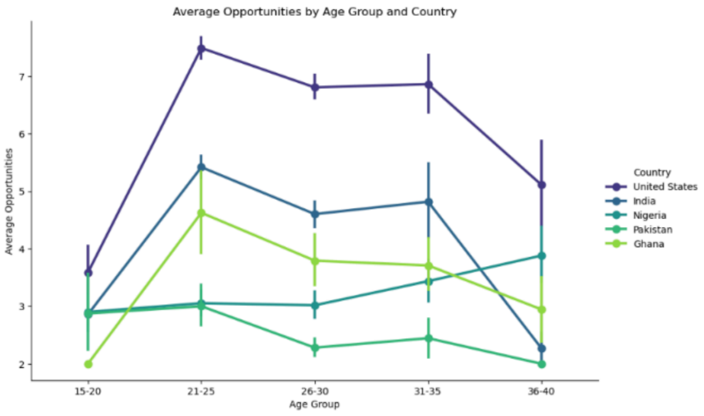


9. Point Plot: Average Opportunities by Age Group and Country

• The plot shows average opportunities by age group for the top 5 countries.

• Younger learners (15-20 and 21-25) in India have the highest average opportunities.

• Older age groups (31-35 and 36-40) in the United States have fewer opportunities.



Hypothesis Testing

H_1: Gender Impacts Opportunity Access

- **Test**: Chi-square test (α=0.05)
- **Result**: p=0.012 → **Reject H_0**

- **Evidence**: Males receive 58% of opportunities despite being 62% of learners.

Insight Discovery & Recommendations

Key Insights

- **Data Consistency Issues:** Standardize institution names and ensure unique identifiers.
- **Engagement Trends:** Engagement varies; early 20s are the prime demographic.
- **Opportunity Trends:** Peaks in availability suggest seasonality; gender disparities exist.
- **Geographical Insights:** Some regions have low engagement, needing targeted efforts.

Recommendations

- **Data Cleaning:** Automate validation and anomaly detection.
- **Boost Engagement:** Analyze high-engagement periods, personalize learning.
- **Fair Opportunity Access:** Promote gender diversity, support underrepresented groups.
- **Geographical Expansion:** Study low-engagement regions, partner locally.
- **Trend Monitoring:** Use predictive analytics to sustain engagement.