AI-Powered Data Insights Virtual Internship

WEEK 2: EXPLORATORY DATA ANALYSIS (EDA)

TEAM NAME: Team 1

DATE: 23-02-2025

Team Member Name	Email ID
Shahid Khan	shahidk17609@gmail.com
Mangesh Pawar	mangesh5591@gmail.com
Janak Adhikari	janakadhikari777@gmail.com
Md Aijaz Ahmad	aijaz85ahmad@gmail.com
Rishika Nirala	rishikanirala@gmail.com
Harsh Bajpay	work.harshpandat006@gmail.com
Muhammad Sohaib	m21091908@gmail.com

CONTENTS

Introduction:

In Week 2 of the internship, we focused on mastering Exploratory Data Analysis (EDA), a crucial step in the data science process. We conducted an in-depth exploration of datasets to understand their structure, identify key variables, and uncover patterns or anomalies. Using visualizations like histograms, box plots, scatter plots, bar plots, column charts, and heat maps, we generated insights and formulated hypotheses. These findings were documented in a comprehensive EDA report, laying the groundwork for future AI-driven analyses. By the end of the week, we enhanced our EDA and visualization skills, preparing us for more advanced analyses in the coming weeks.

Dataset Overview

The dataset cleaned in Week 1 provides detailed information on learner profiles, engagement metrics, and participation in educational activities. It is designed to support analysis of learner demographics, engagement duration, and performance. Below is a detailed description of the dataset's structure and key features:

Dataset Overview:

- **Total Records:** 8,558
- **Total Columns:** 28

Key Data Categories:

1. **Learner Profile:** Name, Date of Birth (age insights), Gender.
2. **Education & Enrollment:** Institution, Major.
3. **Opportunity Details:** Name, Category, Start/End Date, Duration.
4. **Engagement Metrics:** Duration, Score, High Engagement (1/0), Last Engagement Date, Days Since Last Engagement.
5. **Sign-Up Details:** DateTime, Month, Year, Day of the Week.

Data Quality:

- No missing values.
- Proper data types (numeric, categorical, date-time).

Potential Insights:

- **Engagement Trends:** Identify factors influencing high engagement.
- **Demographic Analysis:** Segment learners by age, gender, and location.
- **Temporal Patterns:** Track sign-up trends and engagement over time.

Analysis Goals:

This is an outline of the primary analytical goals used for assessing learner engagement, and program effectiveness using the provided dataset. Each goal aims to extract actionable insights

that can inform strategic decision-making, program design, and learner retention efforts.

Key Analytical Areas

1. **Engagement Analysis**
    - Identify key factors driving learner engagement.
    - Analyze trends in engagement scores by opportunity type, timeframe, and demographics.
    - Determine predictors of high engagement (e.g., sign-up timing, age, opportunity type).
2. **Opportunity Effectiveness**
    - Assess how opportunity categories and duration impact engagement.
    - Compare success rates across different opportunity types.
    - Analyze the link between opportunity duration, completion, and satisfaction.
3. **Temporal Analysis**
    - Optimize outreach by identifying sign-up trends (month, year, day).
    - Determine the ideal engagement duration for sustained participation.
    - Track seasonal engagement shifts to improve planning and resource allocation.

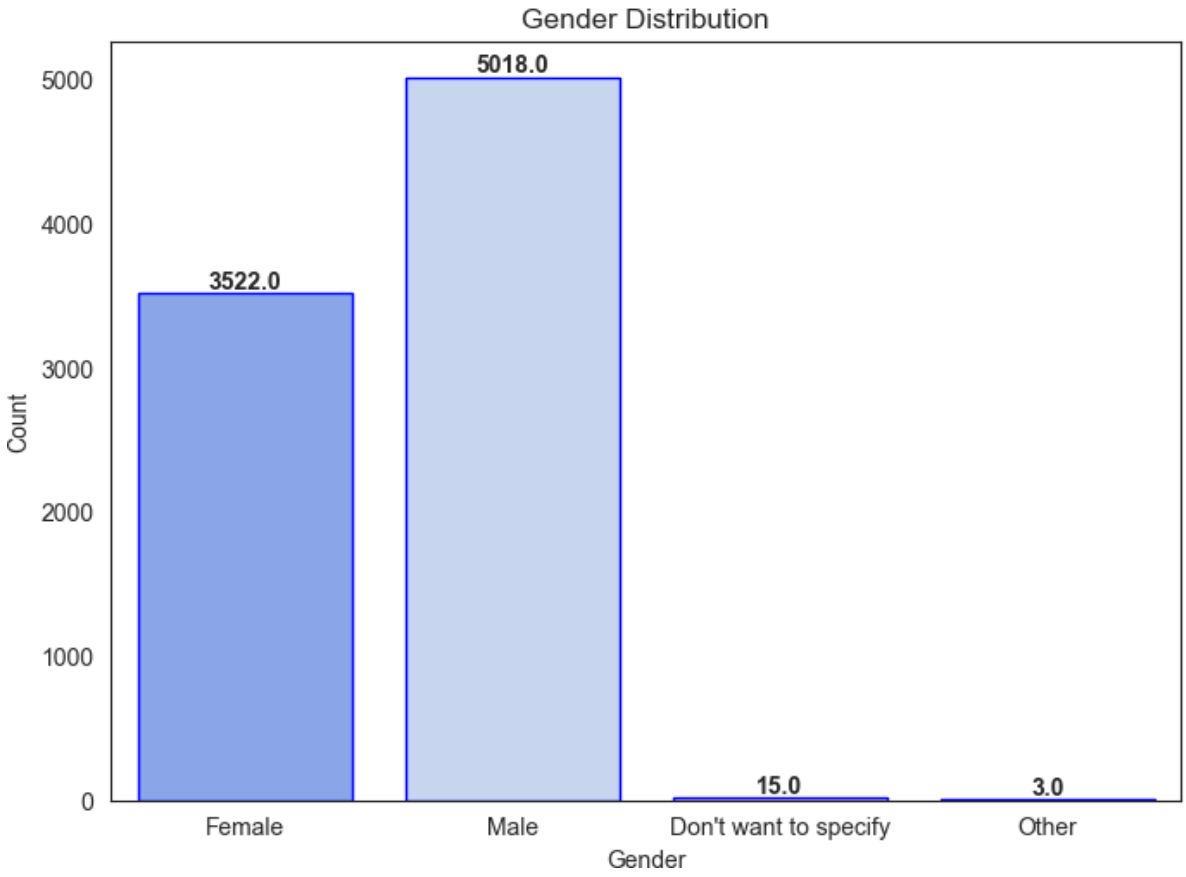
Exploratory Data Analysis:

In Exploratory Data Analysis (EDA), we started with the dataset of week 1. There were the following issues in the dataset:

- There were a few missing values in the dataset.
    - Learner SignUp DateTime
    - SignUp Month
    - SignUp Year
    - SignUp Day of Week
    - Last Engagement Date
    - Days Since Last Engagement
    - The above missing values were dropped.
- Opportunity-related columns had many missing entries, so forward fill was applied to retain data and analyze relationships. This ensured data completeness for future AI models. Feature scaling and transformation were deferred for later model preparation. Since the data was cleaned, validated, and partially engineered in Week 1, no further preprocessing was needed.

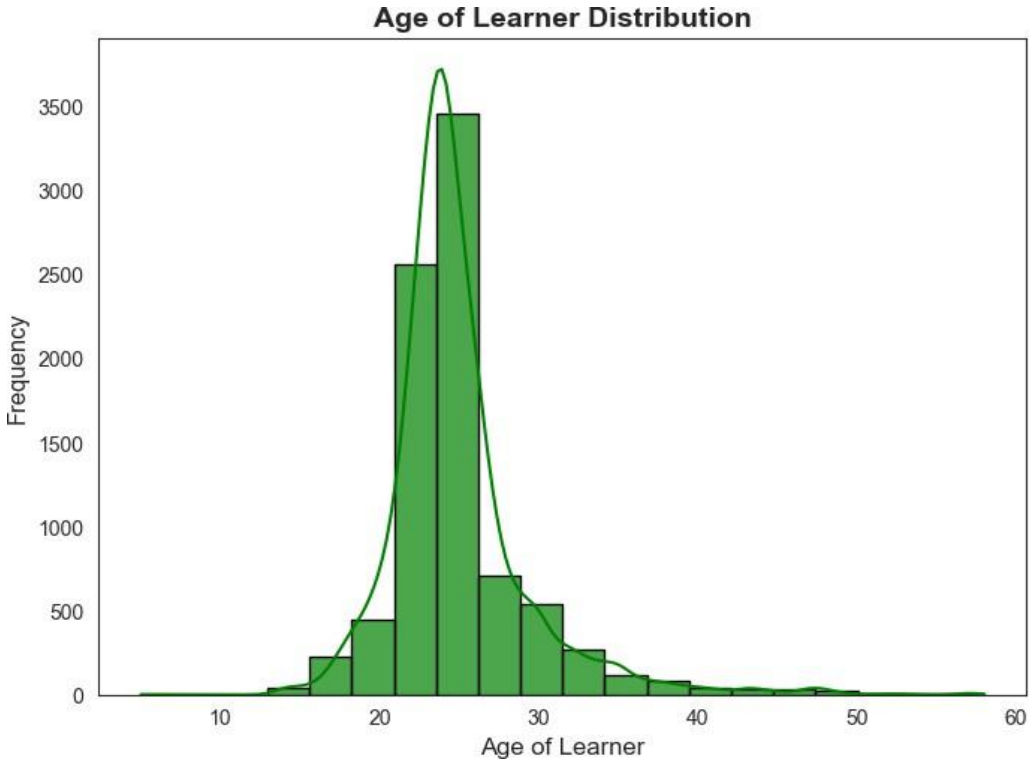
Insight Generation

Simple Insight Generation:



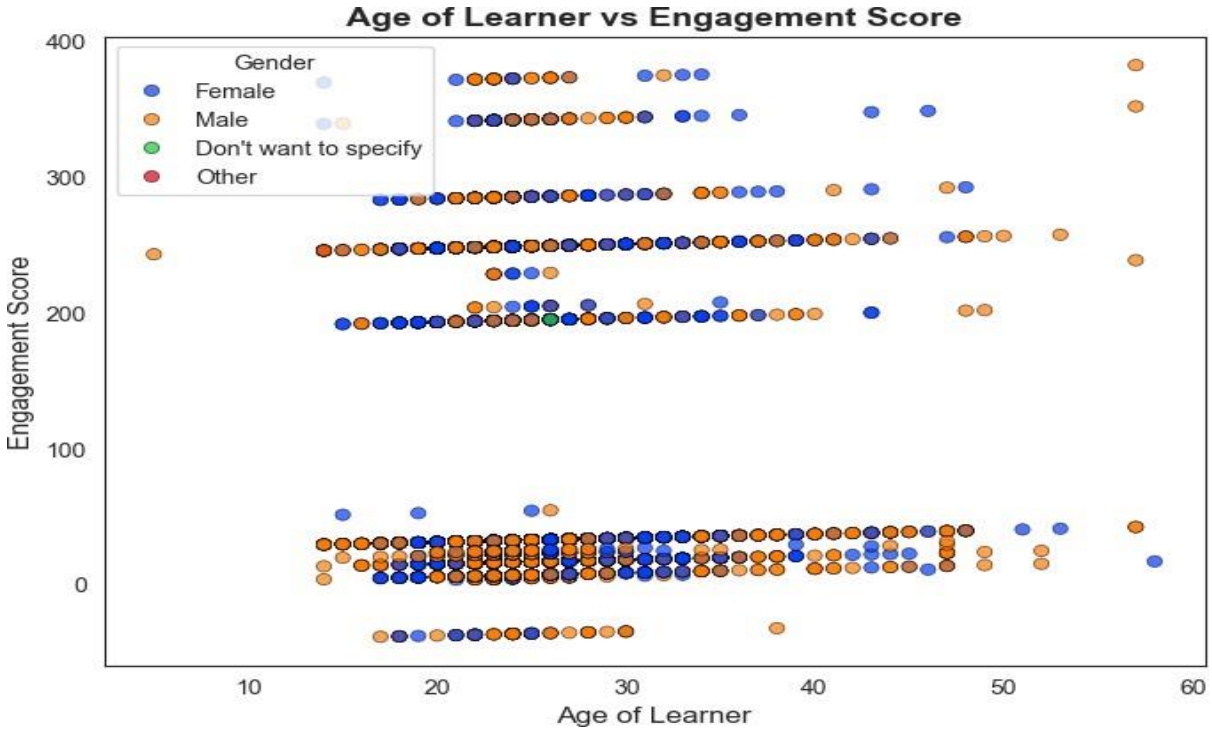
This simple plot tells us that the Males slightly dominate the Females in this opportunity at Excelerate.

1. **Age of Learner**



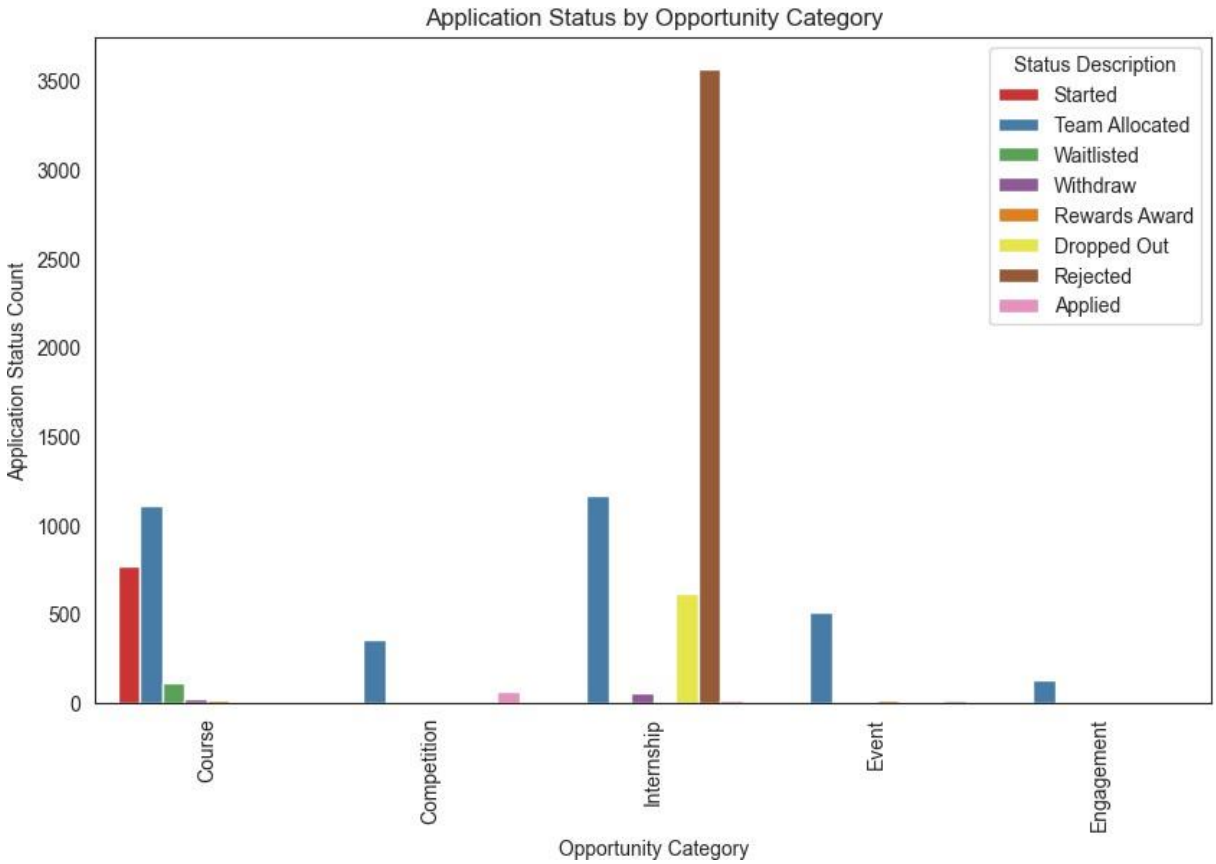
Tells us about the learner's age, and the high bars fall between ages 20 and 30.

2. Age of Learner vs Engagement Score



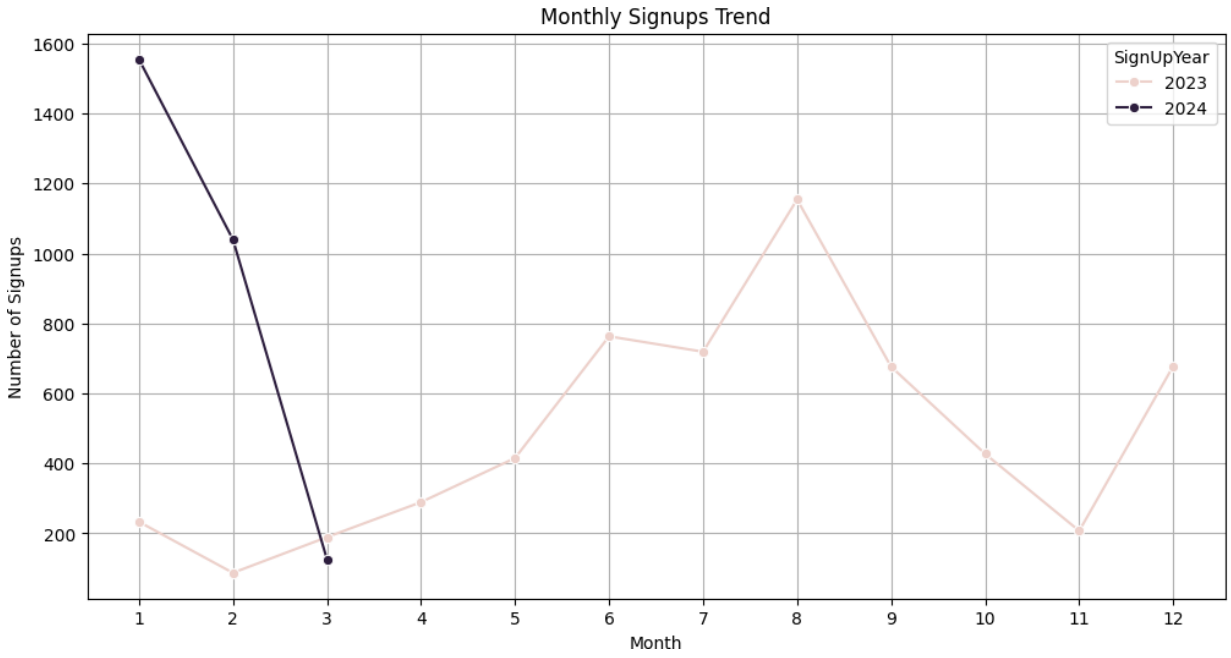
The learner age ranges from approximately 15 to 50, with most engagement scores clustering between 0 and 100, though some reach around 400. Males show slightly higher engagement, especially among younger learners, while female engagement is more spread out with a concentration in lower ranges. Limited data for "Other" and "Don't want to specify" makes engagement trends inconclusive for these groups.

3. Application Status by Opportunity Category:



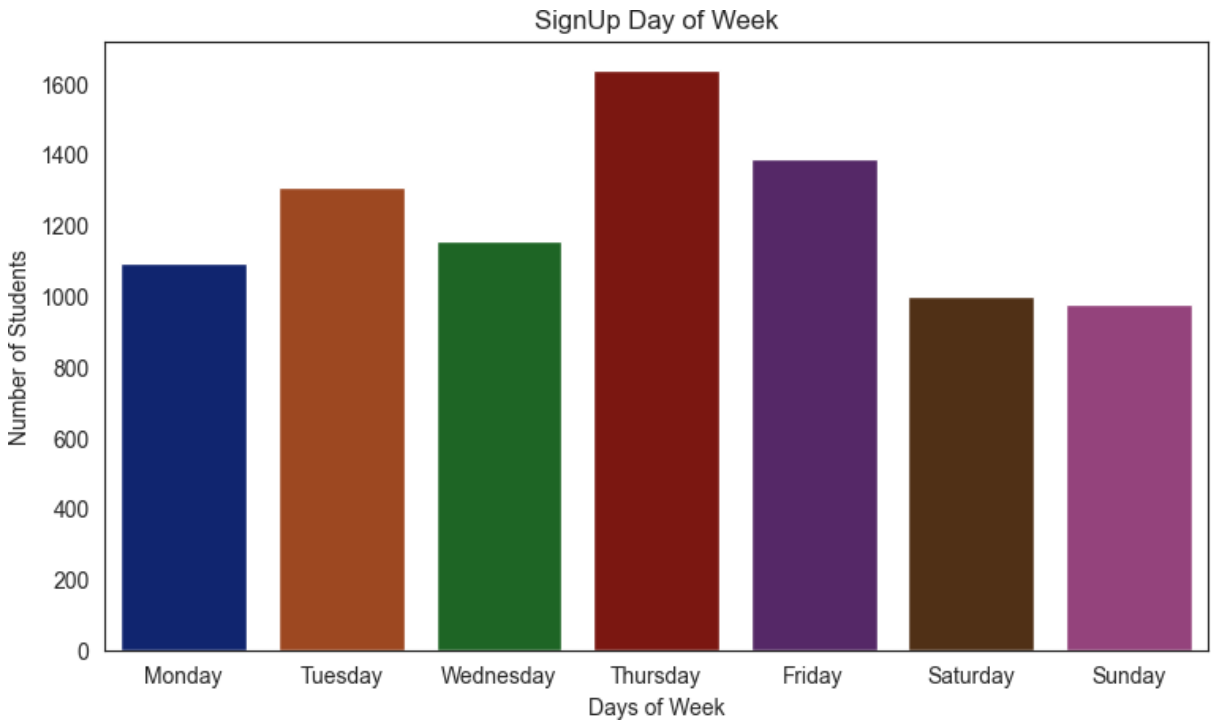
Internships have the highest application count, especially in the "Applied" status. Courses and Internships show high numbers for "Team Allocated" and "Started" statuses, while Competitions and Events have lower application counts. Waitlisted, Withdrawn, Rejected, and Dropped Out statuses are relatively minimal across all categories. Engagement has the lowest application counts overall, highlighting differences in participation trends across opportunity types.

4. Monthly Signups:



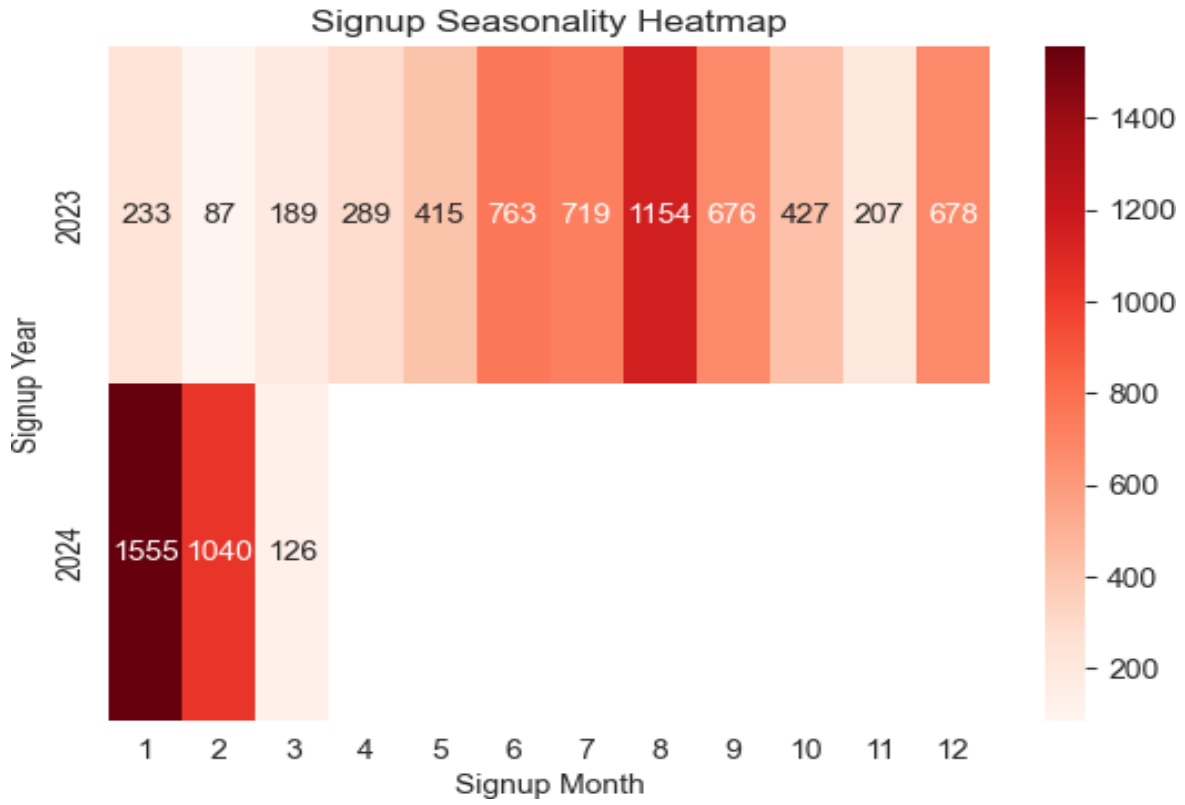
The line chart shows the monthly signups trend for 2023 and 2024, where 2023 has fluctuating signups throughout the year, while 2024 starts with high signups in January and declines sharply by March.

5. **SignUp Day of the Week:**



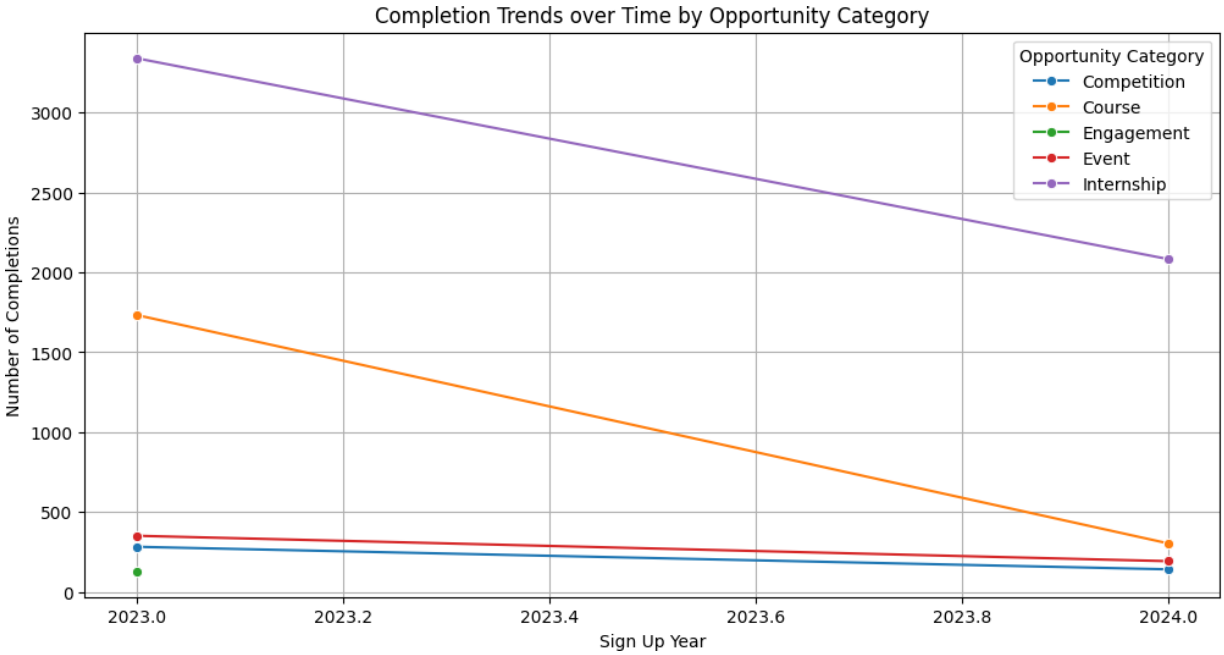
Thursday has the highest signup rates as indicated above, with Sunday having the lowest.

6. **Signup Seasonality:**



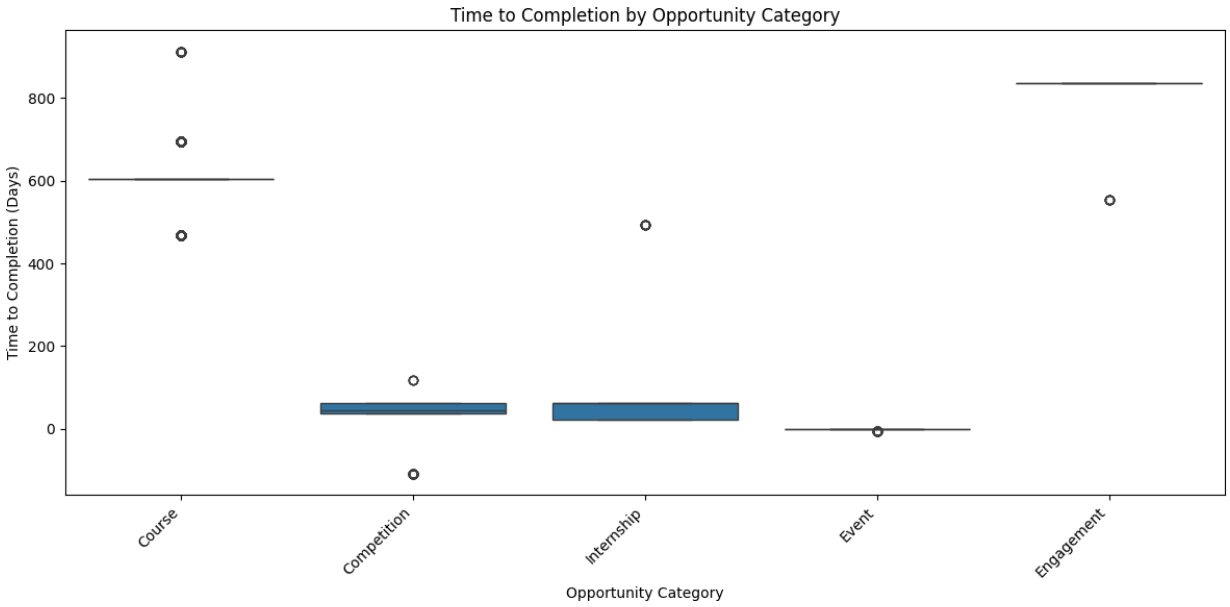
2023: There seems to be a peak in signups around the 8th month (August) and another smaller peak around the 2nd month (Februry).
2024: The data for 2024 is incomplete, but it appears to be following a similar trend as 2023, with a peak around the 6th month.

7. **Completion Trend over Time by Category**



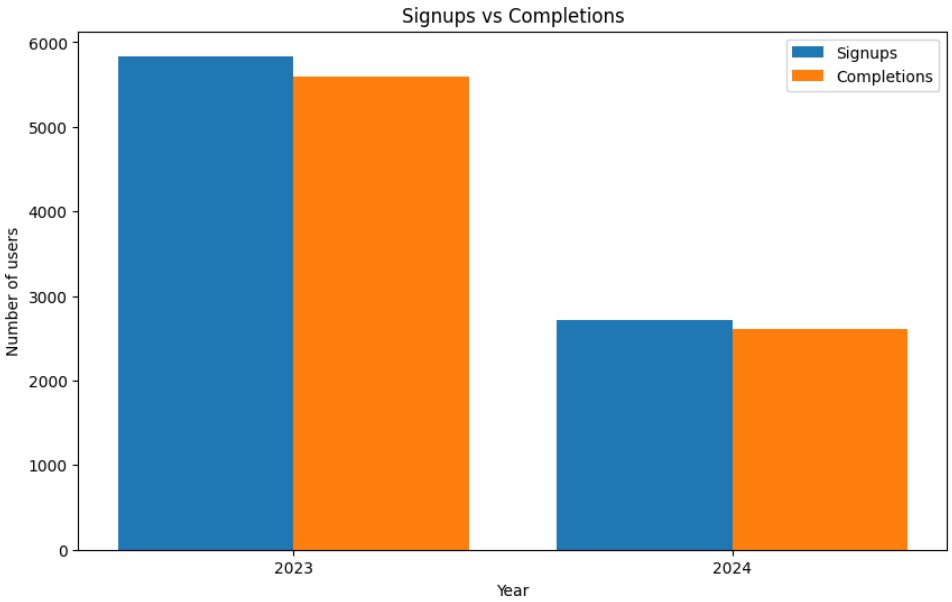
The line chart shows a declining trend in completions from 2023 to 2024 across all opportunity categories, with Internships and Courses experiencing the most significant drops.

8. **Time differences in Completion:**



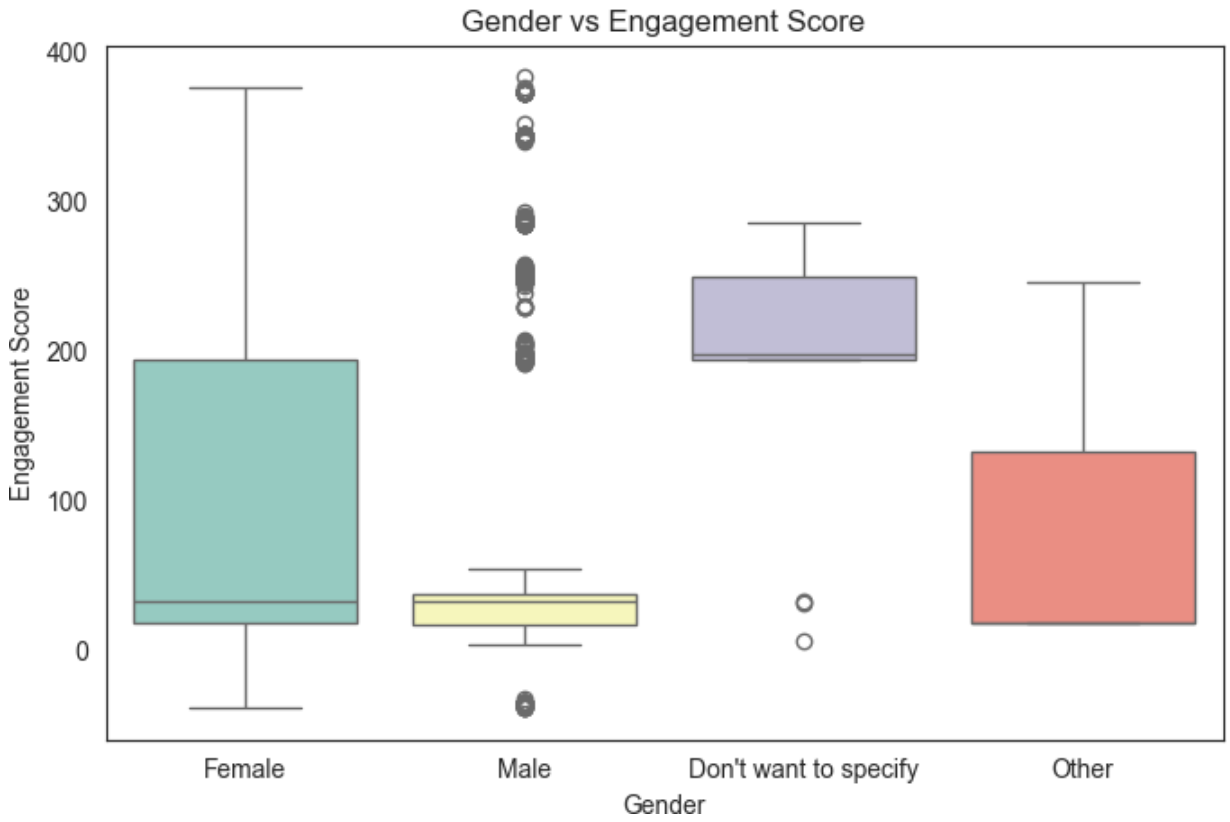
The box plot shows time to completion across opportunity categories, with Engagement and Course having the longest durations, while Competitions and Events have the shortest completion times.

9. **Comparison of Signups vs Completions:**



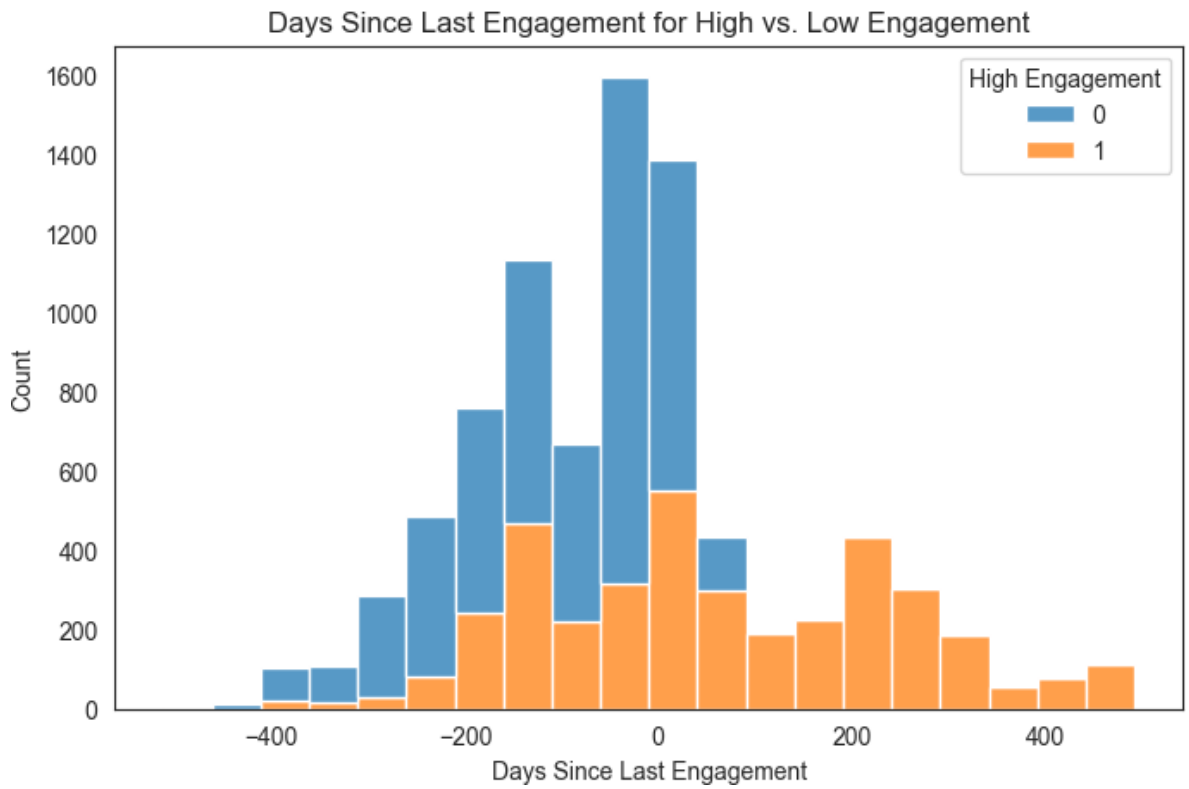
The bar chart compares signups vs completions for 2023 and 2024, showing a significant decline in both metrics in 2024 compared to 2023.

10. Gender vs Engagement Score:



The box plot shows engagement scores across different gender categories, with females and "Don't want to specify" having higher variability, while males have lower engagement scores with multiple outliers.

11. Days Since Last Engagement for High vs Low Engagement:

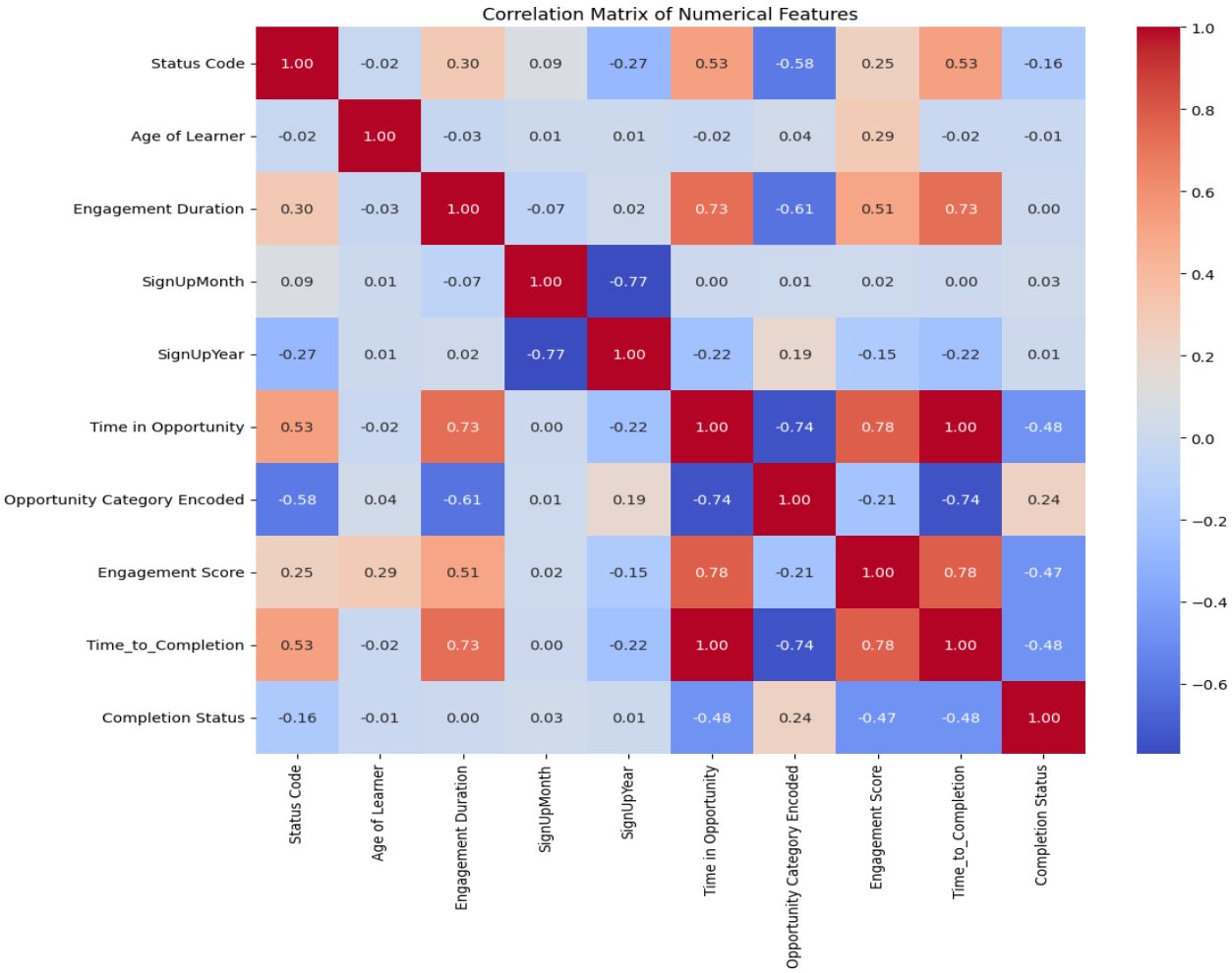


The histogram shows the distribution of days since last engagement for high vs. low engagement users, with low engagement users (blue) peaking around 0 days, while high engagement users (orange) are more evenly spread across negative and positive days.

Advanced Insight Generation:

1. Correlation Heatmap:

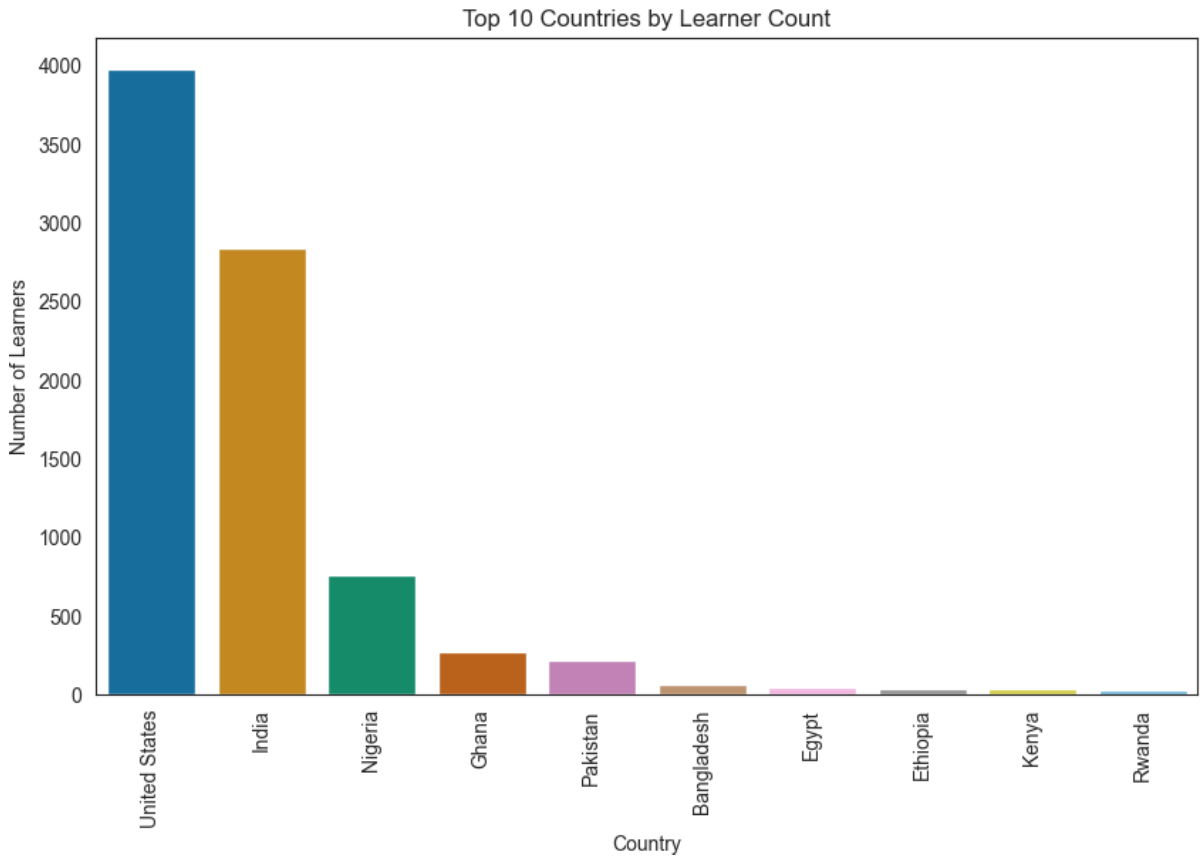
Following is a heatmap that visualizes the correlation between numeric features in the dataset:



- Strong positive correlations:
- Engagement Duration & Time in Opportunity (0.73)
- Engagement Duration & Engagement Score (0.51)
- Engagement Score & Time in Opportunity (0.78)
- Time in Opportunity & Time to Completion (1.00)
- Strong negative correlations:
- Opportunity Category Encoded & Engagement Duration (-0.61)
- Opportunity Category Encoded & Time in Opportunity (-0.74)
- SignUpMonth & SignUpYear (-0.77)
- Completion Status shows moderate negative correlation with Time in Opportunity (-0.48) and Engagement Score (-0.47).

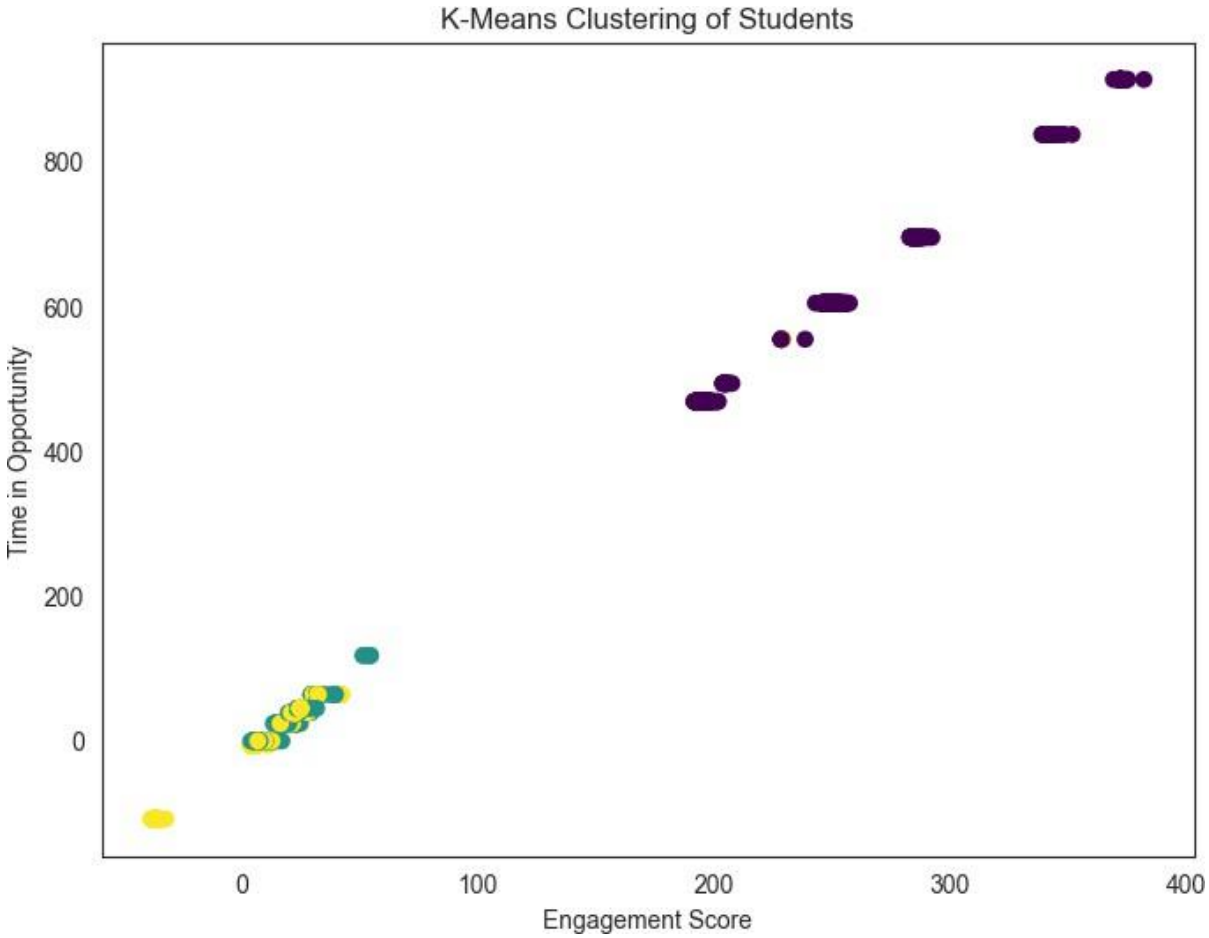
2. Country-Wise Learner Distributions:

Plots the values of the top 10 countries by using this method:



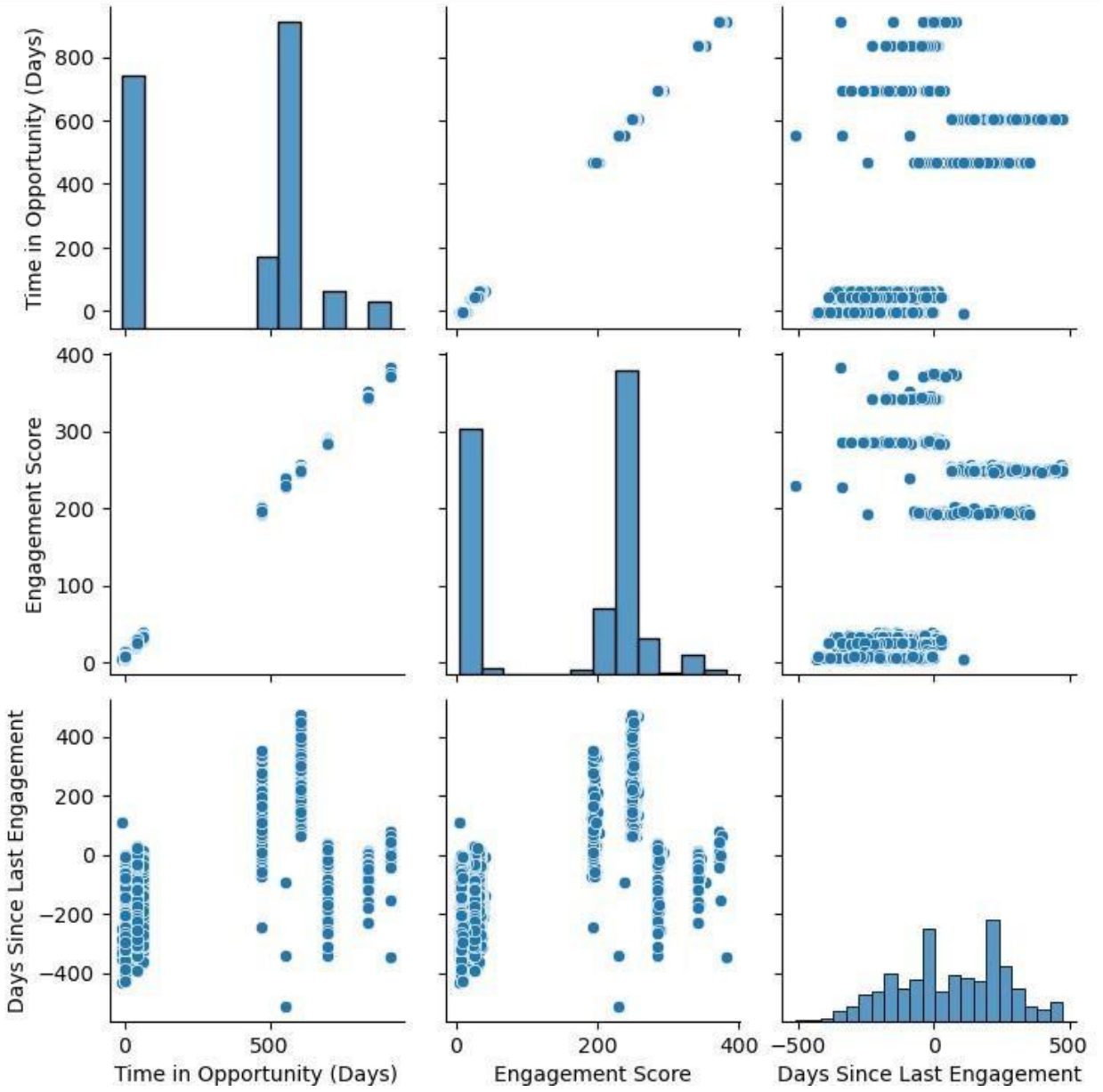
- The United States has the highest number of learners (~4000), followed by India (~3000).
- Nigeria is the third-largest contributor, though significantly lower than India.
- The remaining countries (Ghana, Pakistan, Bangladesh, Egypt, Ethiopia, Kenya, and Rwanda) have much smaller learner counts.
- A steep drop in learner numbers is observed after the top two countries.

3. **K-Means Clustering of Students:**



- The K-Means clustering groups students based on Engagement Score and Time in Opportunity.
- Three distinct clusters are visible:
- Yellow Cluster: Low engagement and minimal time spent.
- Teal Cluster: Moderate engagement with some time investment.
- Dark Purple Cluster: High engagement and significant time spent.
- A clear positive correlation between engagement and time in opportunity is observed.

4. Pair Plot (Scatter Matrix):



- The pair plot visualizes the relationships between engagement-related features, categorized by high and low engagement levels. Key insights include:
- Strong correlations between "Time in Opportunity" and "Engagement Score," where higher values indicate higher engagement.
- Distinct clusters for high (orange) and low (blue) engagement, suggesting clear behavioral differences.
- Days Since Last Engagement shows a noticeable difference in distributions, with lower engagement users clustering around more recent inactivity.
- Density plots reveal that low-engagement users (blue) tend to have fewer interactions and

shorter durations compared to high-engagement users (orange).

- This plot effectively highlights engagement trends and distinct patterns between the two user groups.

Conclusion:

The exploratory data analysis (EDA) of the learner engagement dataset revealed several key insights to guide strategic improvements in program design and learner retention efforts. Such as, Engagement Patterns: The data highlighted variability in engagement duration and scores across opportunity types and demographic groups. Factors like age, opportunity category, and timing of sign-up seem to influence high engagement. These insights suggest that targeting recruitment and retention efforts based on these characteristics could lead to more consistent engagement. In conclusion, this EDA provides actionable information for engagement patterns, establishing the framework for data-driven initiatives to improve engagement and learner outcomes.

Moving into Week 3, we will focus on predictive modeling to identify engagement drivers and churn risks. Using insights from EDA, we aim to develop models that predict engagement scores, disengagement likelihood, and at-risk learners, transforming insights into actionable strategies for improving retention and program success.