**WEEK 2: EXPLORATORY DATA ANALYSIS REPORT**

**Prepared by Team D**

| Ansh Bugra | Faiza Bashir | Anubhav Rohilla |
| --- | --- | --- |
| Fouzia Ashfaq | Faridat Adewole | Fridaos Akorede |

1. **INTRODUCTION**

### **Objective**

The primary aim of an EDA report is to help in getting a deep understanding of a dataset and to ensure working with clean, accurate and consistent data. This report is set to examine the distribution of variables, checking and handling outliers and missing values to ensure the dataset is ready for more advanced AI-powered analysis. It also aims to identify patterns and relationships between variables by employing data visualizations. This helps in building an efficient AI model that can predict user behavior, engagement and seasonal trends that helps provide insights needed to make informed, data-driven decisions for refining user engagement strategies and optimizing programs.

### **Dataset overview**

#### **Data Source**

The dataset was obtained from Excelerate “SLU Opportunity-wise dataset” which consists of opportunity and learner/user details between the year 2022 and 2024.

#### **Data structure**

The dataset is a CSV file containing 2847 rows and 54 columns. It includes numerical data (age, engagement duration; days, standardized and normalized metrics, combined score), categorical data (gender, first name, Institution, current or intended major, country, status description), date fields (learner signup date, apply date, opportunity start and end date, entry created date), binary categorical variables ( one-hot encoded columns) representing the presence or absence of specific attributes for some of the categorical data entries with values of either True/False.

#### **Data Description**

The dataset includes sign-up and engagement information of learners applying to various opportunities. The key columns in the dataset includes:

* Learner details: Age, gender, country, institution name, current/intended major.
* Opportunity details: Opportunity name, Category, Start date and End date, Opportunity duration(days).
* Engagement tracking: Signup date, Apply date, Entry date, Status description and status code, Engagement duration(days).
* Standardized and normalized metrics: Age, Engagement duration, Opportunity duration.
* Encoded categorical variables: Gender (male, female, don’t want to specify), Status Description (team allocated, rewards awarded, started, withdrawn, dropped out), Opportunity category (event, course, engagement, internship, competition).
* Combined score: Composite metric derived from normalized features. .

1. **METHODOLOGY**

**Tools used**

Microsoft Excel was utilized for data cleaning, preprocessing and feature engineering tasks due to its flexibility with tabular data and built in functions for text, date and numeric transformations. Python (pandas, numpy, matplotlib and seaborn libraries) was used for data visualization for its simplicity, versatility, rich and extensive libraries and easy integration with data science and machine learning

**Data Cleaning and Preprocessing**

Cleaning and preprocessing data is a crucial step when analyzing to ensure data quality, consistency. and its reliability. Actions taken in the cleaning process are:

* Dropped irrelevant columns: 23 ‘Unnamed’ columns which contained mostly missing values.

**Sample of dropped column (Unnamed)**

| First name |
| --- |
| None |
| None |

* Standardized text fields: Converted cities, institutions, first name and majors to consistent formats.
* Parsed datetime fields: Applied datetime conversion to date fields (apply date, opportunity start date, signup date, entry date.
* Handled missing values: Imputed or cleaned field with limited missing data.
* Removed Duplicates: Duplicate entries were checked and removed.

**Sample of handled missing values**

| Date of birth | First name | Learner signup date | Opportunity start date |
| --- | --- | --- | --- |
| “Missing” | Faria | 2023-06-14 12:30:35 | 2022-03-11 18:30:39 |
| 2000-08-16 | Poojita | “Missing” | 2022-03-11 18:30:39 |
| 1991-01-11 | Amrutha Varshini | 2023-08-29 05:20:03 | “Missing” |

Feature Engineering

Derivation of new columns or features from the dataset to enhance its analytical capabilities. Features added include:

* Age: Calculated from DOB. This enabled the segmentation of learners by age group.
* Engagement Duration(days): highlighted time taken to apply for an opportunity from when the opportunity actually started.
* Opportunity Duration(days): provided insight into program duration lengths.
* Signup trends (Signup month and Signup year): It helped understand seasonal sign-up behavior.
* Normalization/Standardization: was applied to numerical fields.
* Combined score: It is aimed to be used to develop an aggregated metric for ranking user engagement or opportunity quality.

These engineered features formed the foundation for analysis and hypothesis generation.

Outlier Handling

Outliers are values in a dataset that are significantly different from most other values. These actions ensured logical integrity and usability of all time-based variables.

* Opportunity Duration: Negative values were detected and dropped, as a program cannot end before it begins.
* Engagement Duration: Presence of Negative values suggests application for an opportunity before the opportunity actual start date. So, these outliers were maintained.

Data Validation

Data validation is to ensure working with accurate, complete, consistent and relevant data. Validation checks carried out are:

* Missing Values: Confirmed no null entries post-cleaning.
* Duplicates: Verified removal of redundant rows.
* Date Time Consistency: Validated split columns for logical ranges (e.g., no future dates).
* Normalization Integrity: Checked z-score distributions for new features.

1. **RESULTS**

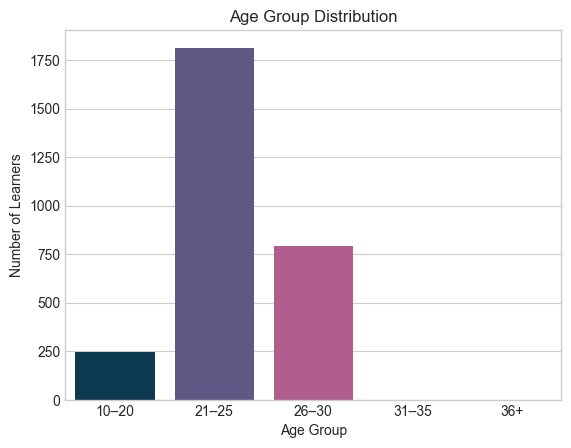
This section presents the key findings from the exploratory data analysis. It includes a combination of summary statistics, visualizations, and insights drawn from the relationships among variables. The analysis focuses on understanding data distributions, identifying patterns and trends, detecting outliers, and uncovering potential correlations that may inform further modeling or decision-making. All results are supported by relevant visualizations for clarity and interpretation**.**

Data Visualization

Data visualization is a crucial next step that allows us to represent our data graphically making it easier to identify trends and insights that are not immediately apparent from the data alone. Bar charts, line graphs and histograms are used to show the distribution for key columns.

#### **Univariate visual analysis**

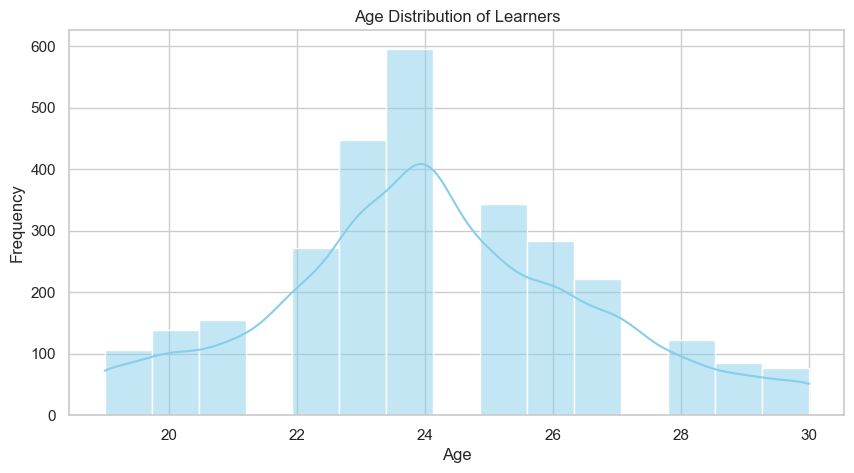
**Gender**: Bar graph illustrates the count of individuals categorized by gender with “male” having the highest count with approximately **1500** individuals. “Female” with slightly above **1000** individuals and “don’t want to specify” **very low**



**Age**: This bar chart visually represents the number of learners within different age categories.

Here's a breakdown:

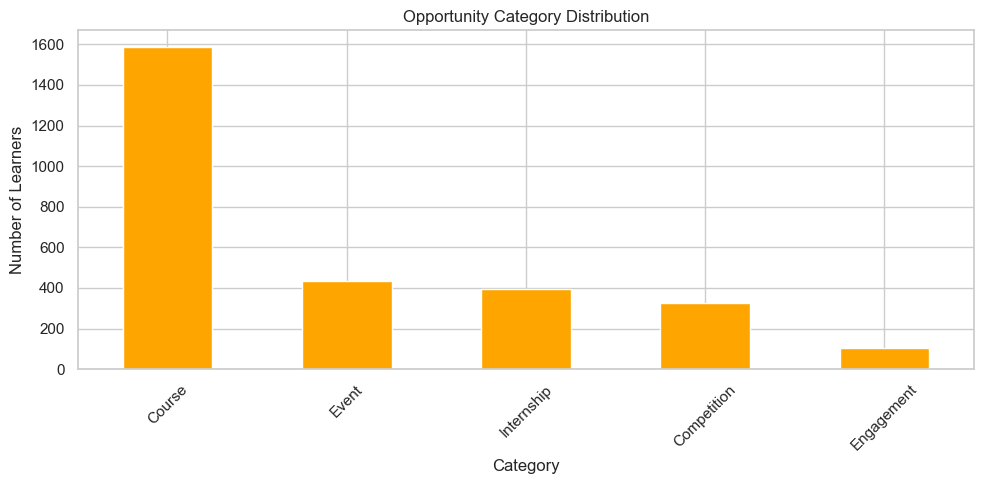
* Age group **21-25** has the highest number of learners, reaching approximately **1,750**.
* Age group **26-30** follows, with around **750** learners.
* Age group **10-20** has the least number of learners, with fewer than **250**.
* There are **no learners** recorded in the age groups **31-35** and **36+**



The above chart is an histogram with a density curve overlay that represents the age distribution of learners. Here's what it tells us:

* The **x-axis** shows the ages of learners, ranging from **20 to 30 years old**.
* The **y-axis** represents the frequency, reaching up to **600 learners**.
* The **histogram bars** depict the number of learners within each age group.
* The **density curve** provides a smooth representation of how the age distribution is spread.

The highest frequency is observed around **24 years old**, meaning that this age group has the largest number of learners. As the age moves away from **24**, the frequency gradually decreases, with fewer learners close to **20** and **30** years old.



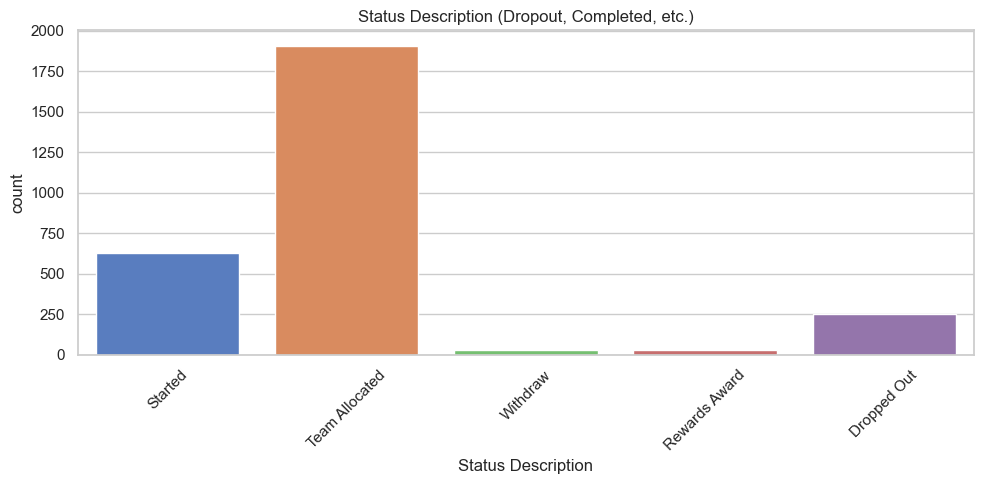
**Opportunity Category**: Bar plot revealed distribution of categories. This bar chart titled visually represents the number of learners engaged in different categories of opportunities.

Key Observations

* Course: This category has the highest participation, with around **1600** learners. This suggests that structured learning programs are the most popular among learners.
* Event: About **400** learners engage in events. This could include workshops, seminars, or networking gatherings.
* Internship: Around **300** learners are involved in internships, highlighting opportunities for hands-on professional experience.
* Competition: Another **300** learners participate in competitions, which might indicate interest in skill-based challenges or prize-driven learning.

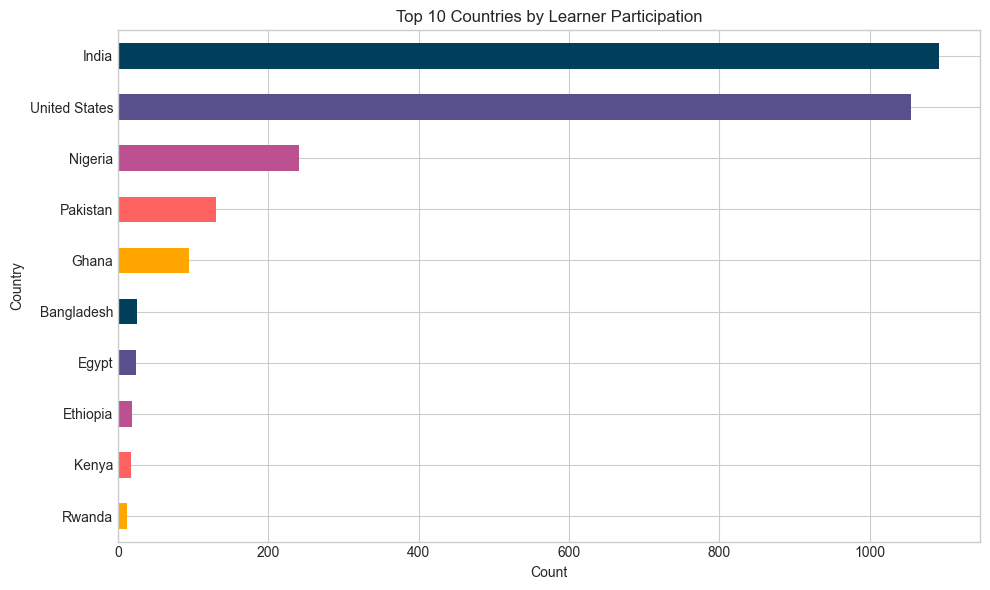
Engagement: This category has the lowest number, with about 100 learners, suggesting limited participation in interactive or community-driven opportunities.

Overall, the data indicates that “courses” dominate learner interest, while “engagement” opportunities have the least traction.



**Status Description**: Dropout rates (dropped out and withdraw) vs Completed rates (Team allocated, rewards awarded, started) visualized. Here is a breakdown of what it reveals:

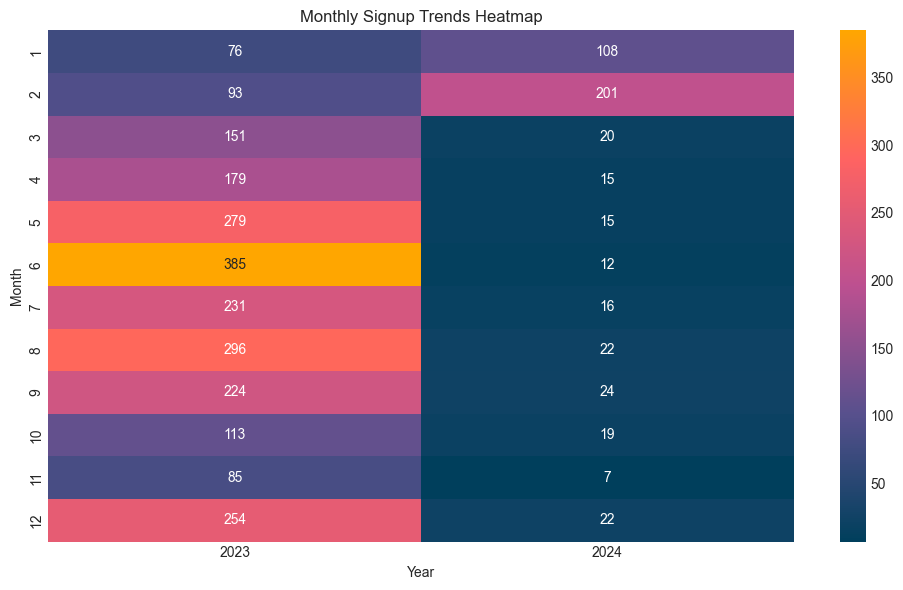
* **Started**: Approximately **500** individuals have initiated the process.
* **Team Allocated**: The largest category, with around **1800** participants being assigned to teams.
* **Withdraw & Rewards Award**: These categories have very low counts, close to zero.
* **Dropped Out**: Roughly **250** individuals have exit before completion.



**Top 10 countries**: Countries with the highest number of learner participation. This bar chart displays the number of learners from different countries. The horizontal bars represent the count of learners, with India leading the chart, followed by the United States and Nigeria.

Here's a breakdown:

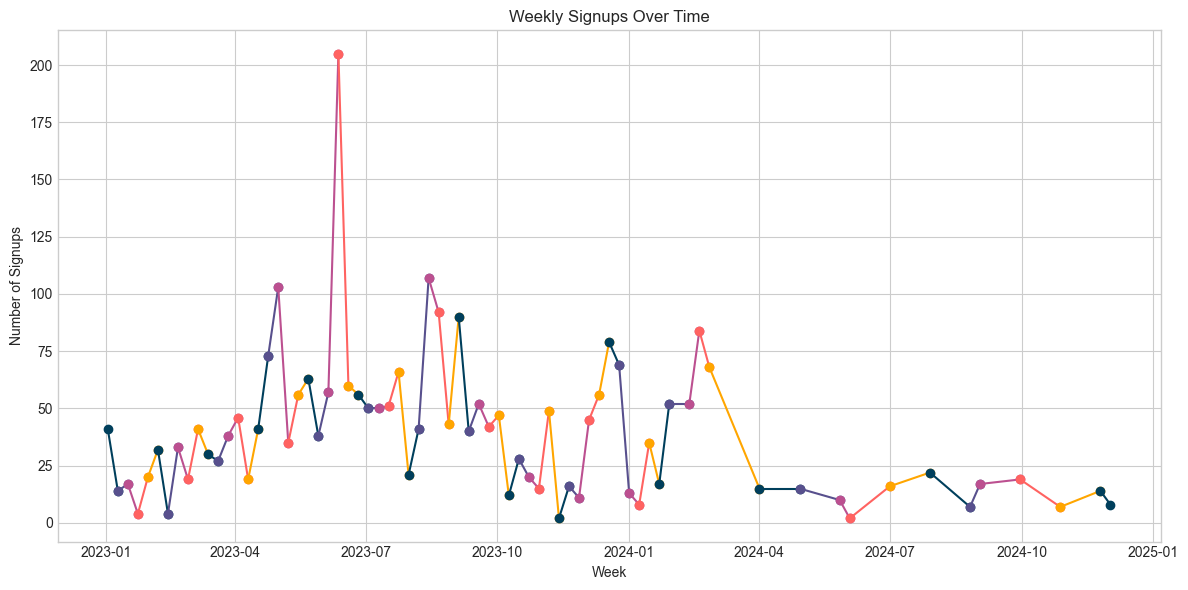
* **India** has the highest learner participation, with approximately **1,100** learners.
* **The United States** follows closely with around **1,000** learners.
* **Nigeria** ranks third with **400** learners.
* Other countries like **Pakistan** (**300 learners**) and **Ghana** (**200 learners**) also show strong engagement.
* The remaining countries—**Bangladesh, Egypt, Ethiopia, Kenya, and Rwanda**—have fewer learners, ranging from **100 to 20**.



**Seasonal trends:** Monthly trend in learner’s signup. Heat map showing the number of signups for years 2023 and 2024. The color intensity represents the number of signups, where dark blue signifies lower signups, while bright yellow indicates higher signups.

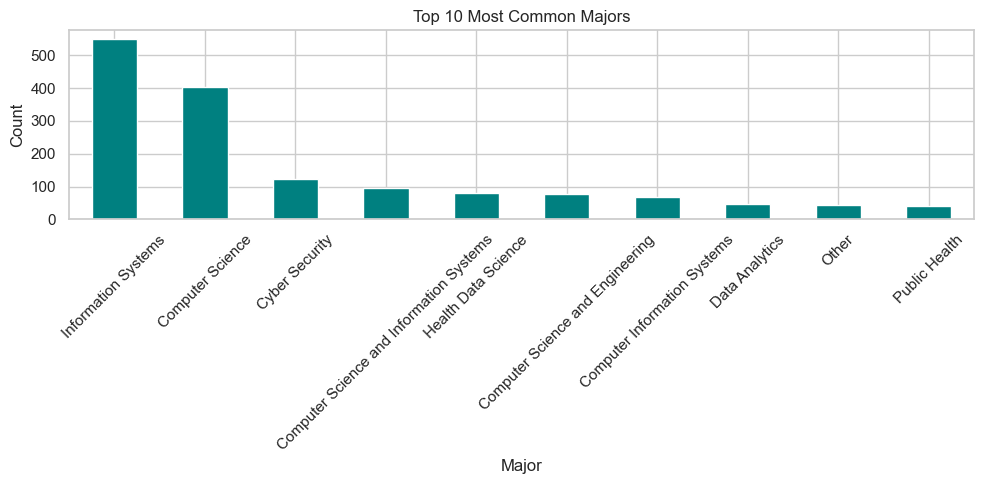
Observations

* **2023:** The signups steadily increase from January (76) to a peak in June (385), before fluctuating for the rest of the year.
* **2024:** Signups drop significantly from January (108) and February (201) to much lower numbers starting March (20) through December (22). This suggests a sharp decline in user engagement.



**Weekly signups overtime**: Scattered plots showing number of signups per week between January 2023 and January. The x-axis marks the weeks, while the y-axis indicates the number of signups, ranging from **0 to 200**. Here’s what stands out:

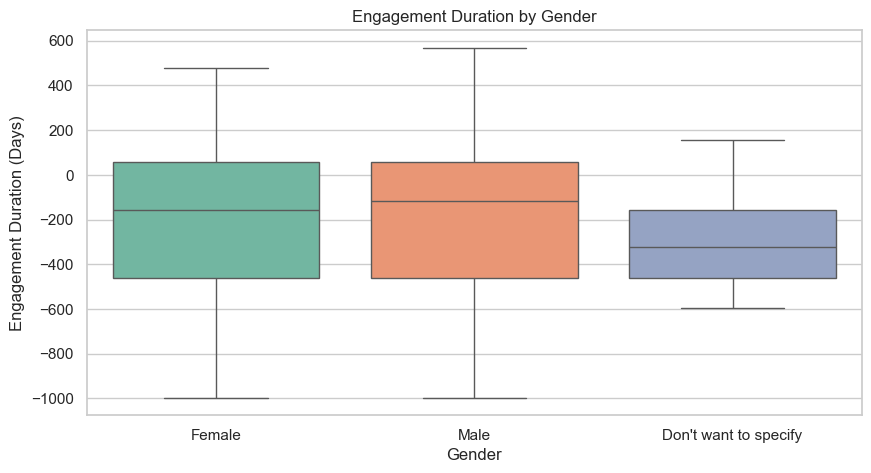
* **Mid-2023 Peak:** There's a sharp increase around mid-2023, reaching **200 signups**, suggesting a major event or promotion boosted signups.
* **Fluctuations:** Signups are not constant; they rise and fall, showing variations in user engagement over time.
* **Trend Over Time:** Some periods show growth, while others reflect a decline, which might indicate seasonal effects, marketing campaigns, or external factors influencing signups.



**Top 10 Majors**: Bar chart of most popular academic specialization of users. The top 10 most common majors were visualized, showing which academic interests are most prevalent among the learners. The key insights are:

* **Information Systems** is the most common major, with more than **500** students.
* **Computer Science** follows closely behind, with around **400** students.
* **Cyber Security** comes in third.
* Other majors such as **Health Data Science, Computer Science and Engineering, and Public Health** have lower student counts.
* Most of the majors listed are related to **technology and data sciences**.

#### **Bivariate & Multivariate Exploration**



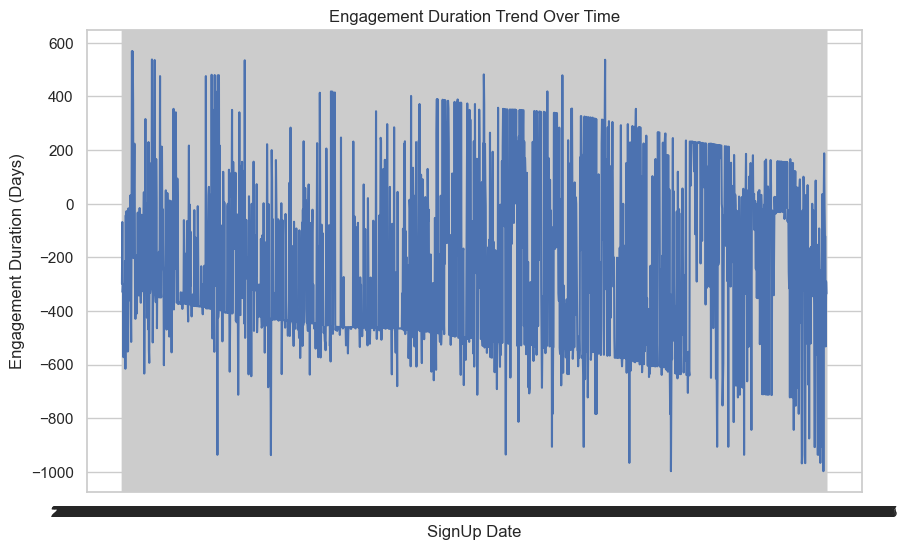
**Engagement Duration by Gender**: Boxplot exposed engagement duration variation by gender. This box plot titled visualizes the distribution of engagement durations in days across the three gender categories (Female, Male, and Don't want to specify)

**Breakdown**

* **Female:** The median engagement duration is around **-200 days**, meaning half of the data points fall below this value and half above. The interquartile range (IQR) spans approximately **-400 to 0 days**, showing where the middle 50% of engagement durations lie. The whiskers extend from about **-1000 to 500 days**, indicating the full range of engagement durations recorded.
* **Male:** The median engagement duration is slightly lower at around -**250** days, suggesting slightly shorter engagement compared to females. The IQR ranges from -**400** to 0 days, similar to females. The whiskers extend from -**1000** to **600** days, showing slightly higher variability.
* **Don't want to specify:** The median engagement duration is around **-200 days**, with an IQR between **-300** and **-100 days**, indicating less spread compared to the other categories. The whiskers extend from about -**800 to 300 days**, meaning engagement durations in this group tend to have slightly less extreme values than the other categories.

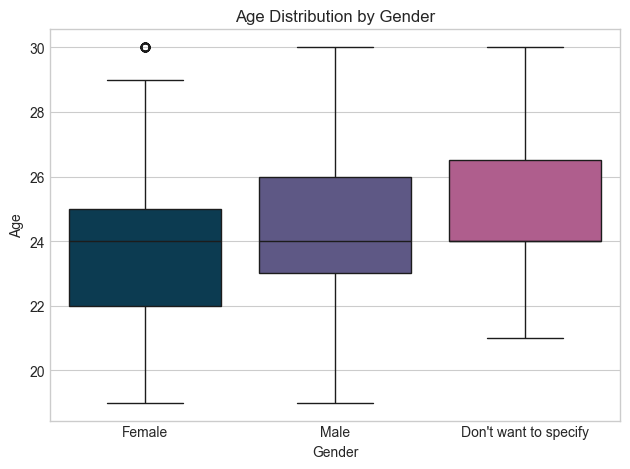
**Interpretation**

* The engagement duration distribution for **"**Don't want to specify**"** appears less variable, meaning this group shows more consistency in engagement duration compared to the Male and Female categories.
* Both Male and Female categories have similar engagement distribution patterns, but Male participants show slightly more variability in engagement duration.



**Engagement duration trend over time:** The graph plots user engagement duration in days against their signup dates. It illustrates how long users take to sign up for an opportunity, measured in days (y-axis), plotted against their sign-up date (x-axis). Based on the data distribution:

* The engagement durations vary widely, with some users showing prolonged activity while others engage quickly.
* The trend appears slightly downward, implying that more recent users tend to have shorter engagement durations compared to earlier ones.
* There are fluctuations, suggesting external factors may influence how long users remain active.
* Certain peaks indicate that some users maintain significantly longer engagement durations.

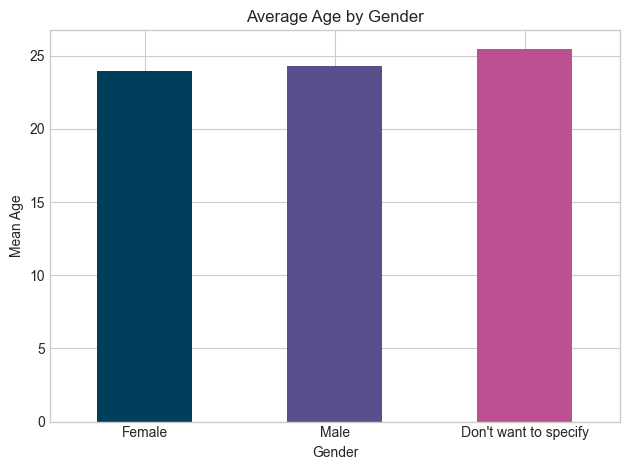


**Age distribution by gender**: The plot is useful for understanding age distribution patterns among different gender identities. The y-axis represents age, ranging from **20 to 30 years**.

Here’s what the data reveals:

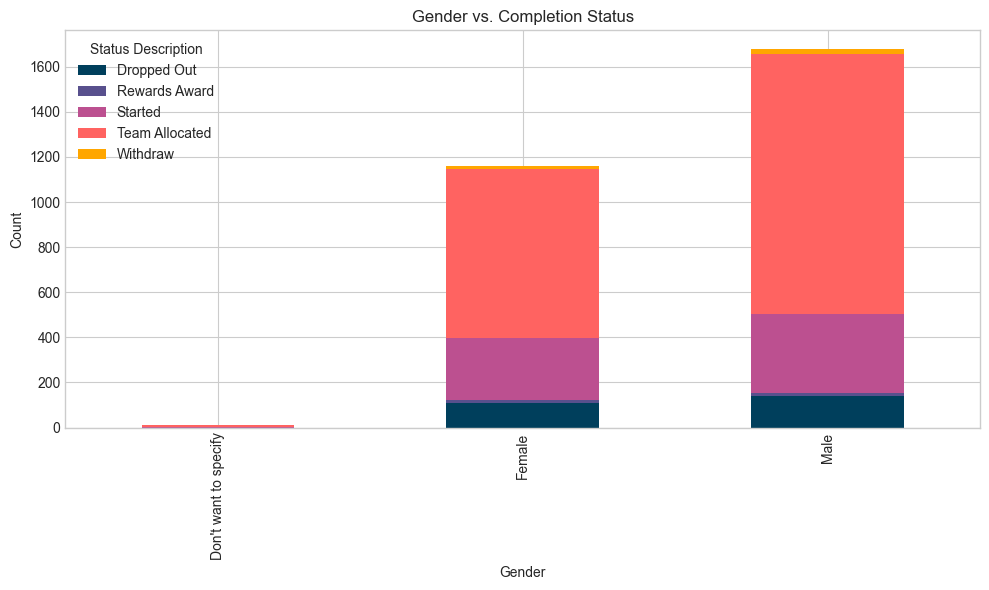
* **Female:** The median age is about **24 years**, with most ages falling between **22 and 26 years**. The youngest age recorded is **20 years**, and the oldest is **28 years**, with an outlier at **30 years**.
* **Male:** The median is also **24 years**, but with a slightly narrower spread, mostly between **23 and 26 years**. Ages range from **21 to 29 years**.
* **Don't want to specify:** The median age here is **25 years**, with ages clustering between **23 and 27 years**. The youngest age recorded is **21 years**, and the oldest is **29 years**.

Overall, the distributions are quite similar across categories, but the **Female** group has the widest spread and an outlier, while the **Don't want to specify** group has a slightly higher median. The plot is useful for understanding age distribution patterns among different gender identities.



**Average age by gender**: This bar chart presents the mean age of individuals across different gender categories. Here's a breakdown of the interpretation:

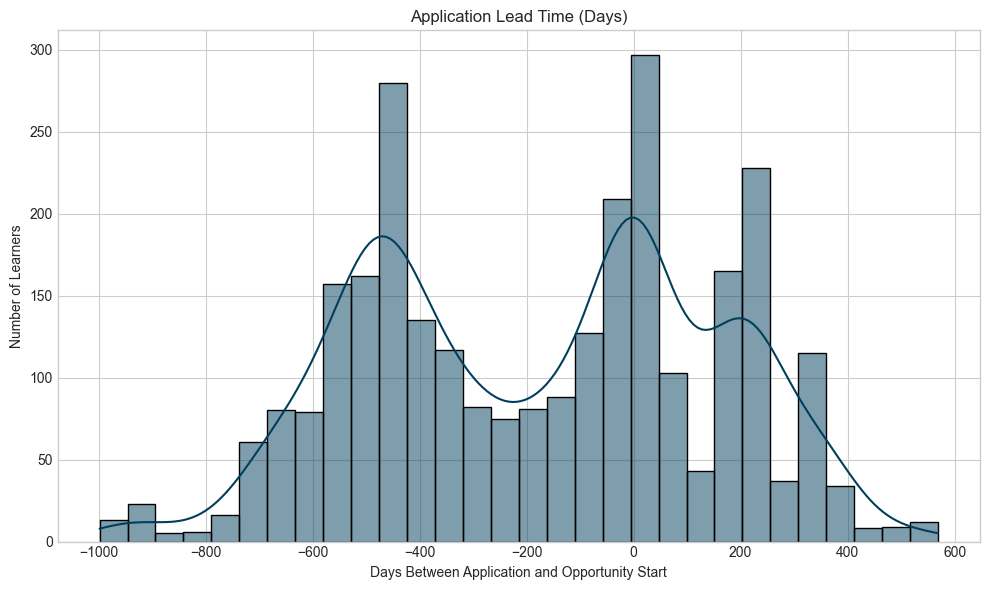
* Female participants have an average age of about 24 years.
* Male participants have a slightly higher average age, around 24.5 years.
* Individuals who chose "Don't want to specify" their gender have the highest average age, which is a little over 25 years.



**Gender vs Completion status:** It presents the count of individuals based on their gender and their completion status in a certain activity or program.It is a stacked bar chart, meaning each bar represents the total count for a gender, with segments showing how many individuals fall into each status.

Interpretation by Gender

* **Male**: Highest total participation. Majority are in the "Team Allocated" status (red segment). Significant numbers in "Started", "Dropped Out", and "Rewards Award". A few have withdrawn.
* Female: Fewer participants compared to males. Most also fall under "Team Allocated", with a notable number having started. Visible presence in "Dropped Out" and "Rewards Award". Very few have withdrawn.
* Don't Want to Specify: Extremely low participation. Only a small count in "Team Allocated"; negligible in other categories.

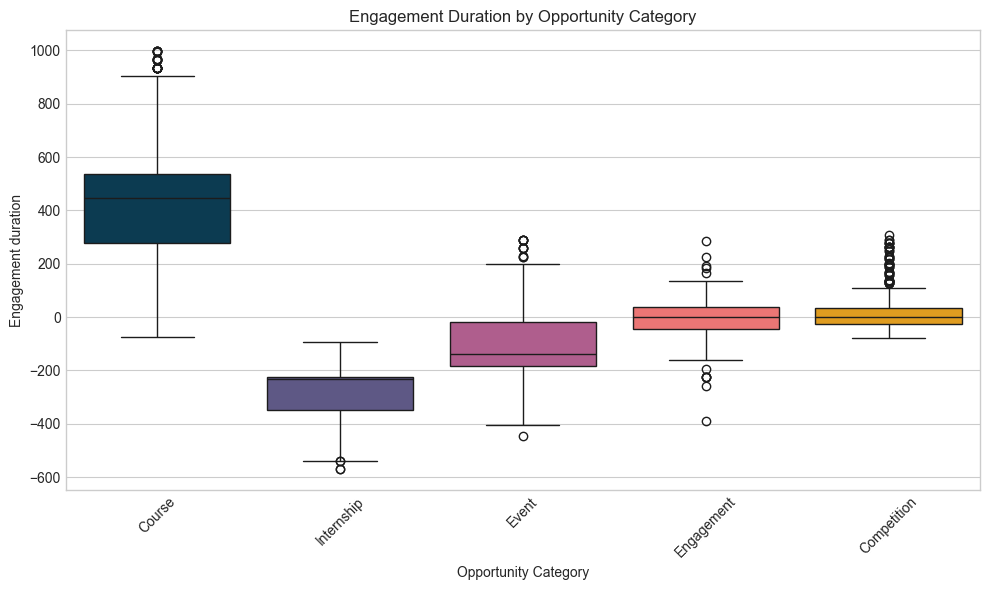


**Application Lead Time (Days):** This histogram with a density plot represents the distribution of application lead times in days—essentially showing when learners applied relative to the start date of opportunities.

**Key Observations**

* The highest peak appears **at 0 days**, meaning the largest group of learners applied on the exact day an opportunity started. This suggests some learners waited until the last possible moment.
* Another notable peak occurs around **-400 days**, showing that a significant number of applicants applied well in advance.
* There's also a small rise around **200 days**, indicating some late applications.
* The long tails extend to both negative and positivea days, meaning some applicants submitted extremely early while others were significantly late.

This graph helps in understanding applicant behavior and optimizing application timelines for better engagement.



**Engagement Duration by Opportunity Category**: A box plot, which helps visualize the distribution of engagement durations for different types of opportunities. The X-axis (Horizontal) shows Opportunity categories while the Y-axis (Vertical) shows Engagement Duration (likely measured in days).

**Box Plot Components**

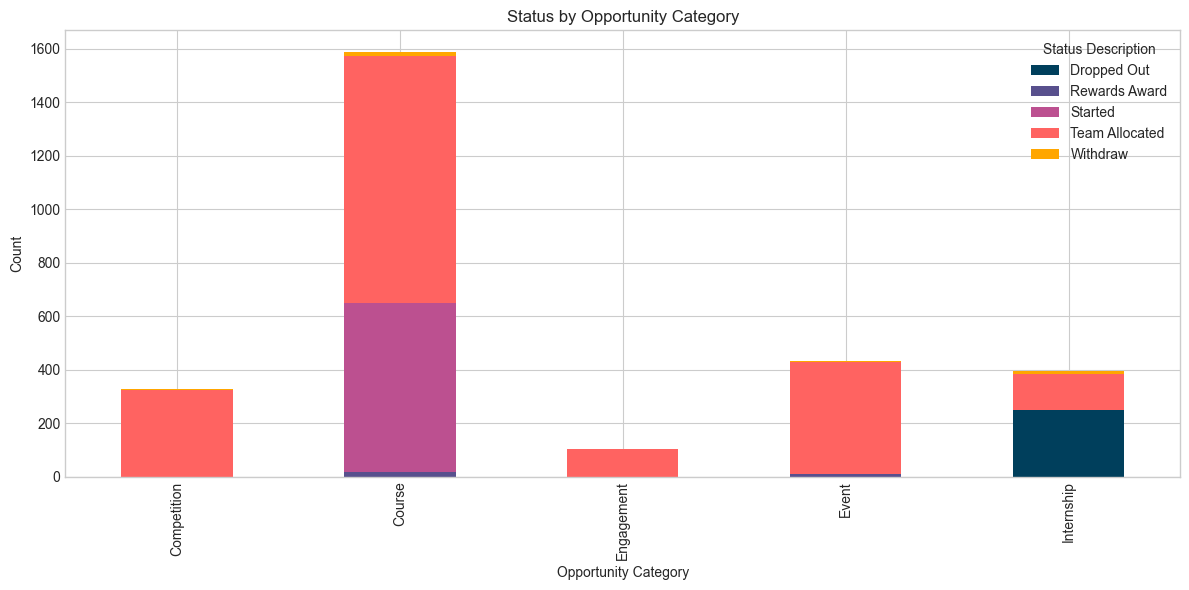
* Box: Shows the interquartile range (IQR) — middle 50% of the data.
* Horizontal line inside the box: The median (middle value).
* Whiskers: Extend to show the range of the data, excluding outliers.
* Dots outside whiskers: Outliers — values far from the rest of the data.

**Interpretation by Category**

* Course: Longest engagement duration overall. Median around 450–500. Wide range, with some outliers above 900. Suggests that courses involve long and varied time commitments.
* Internship: Negative engagement durations are prominent which might indicate interest of users in opportunity before its actual start date.
* Event: Median engagement duration is slightly negative, similar to internships. Contains both positive and negative values with many outliers. Likely short-term engagements with inconsistent logging.
* Engagement: Compact spread with a median close to 0, indicating short, one-off or minimal-time engagements.
* Competition: Similar to Engagement: median near 0, small IQR. Some high outliers, but mostly short durations.

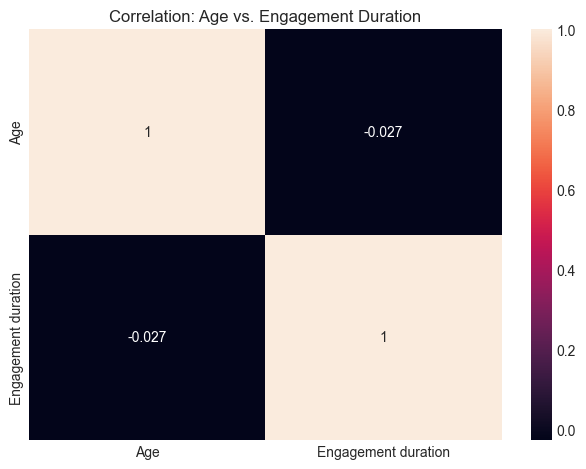
**Key Takeaways**

* Courses have the longest and most structured durations.
* Engagements and competitions tend to have short, concentrated participation periods.
* Internships have the shortest engagement periods as some users even Signup before the opportunity starts.



**Status by Opportunity Category**: A bar graph showcasing the count of different statuses within various opportunity categories. Here are key observations:

* **Course** category has the highest count, especially in Started and Team Allocated statuses.
* **Internship** shows a notable number in the Dropout status.
* **Event** category has a strong presence in the Team Allocated status.
* **Competition and Engagement** have lower overall counts, primarily appearing in the Team Allocated status**.**

****

**Correlation of age and Engagement duration**: The chart is a correlation matrix displaying the correlation between "Age" and "Engagement Duration". Here's an interpretation:

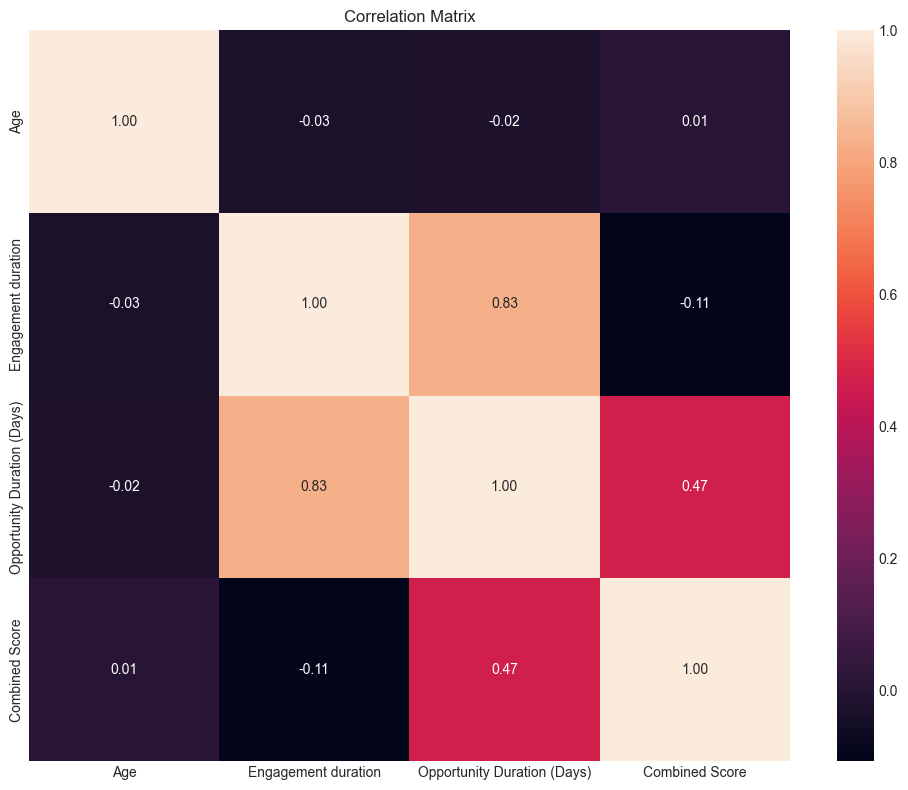
* A correlation matrix visualizing the correlation coefficients between multiple variables. In this case, we have two variables; "Age" and "Engagement Duration".
* The diagonal elements (top-left and bottom-right) always show a correlation of 1. This is because a variable is perfectly correlated with itself. So, "Age" with "Age" is 1, and "Engagement Duration" with "Engagement Duration" is 1.
* The off-diagonal elements (top-right and bottom-left) show the correlation between the two different variables.

**Interpretation of the values**

* Correlation Coefficient: The value shown is -0.027.

**Nature of Correlation**

A correlation coefficient close to 0 indicates a very weak (almost non-existent) linear relationship between the variables. The chart shows that there is a very weak negative linear correlation between "Age" and "Engagement Duration". In practical terms, this suggests that there is almost no discernible linear relationship between a person's age and how long they engage with the opportunities. Changes in age do not reliably predict changes in engagement duration in a linear fashion, and vice-versa.

****

**Correlation Matrix heatmap**: Let's break down what it shows and how to interpret it:

The colors in a heatmap typically represent the strength and direction of the correlation: warmer colors (like orange/pink) often indicate stronger positive correlations, while cooler colors (like dark purple/blue, though not prominently seen here) might indicate stronger negative correlations. The diagonal cells always show a correlation of 1.00 because a variable is perfectly correlated with itself. The matrix includes four variables:

* Age
* Engagement duration (Days)
* Opportunity Duration (Days)
* Combined Score

**Interpretation of the Chart**

* Age vs. Engagement duration

Correlation of -0.03. This interpretes a very weak negative correlation, almost negligible. It suggests that there is practically no linear relationship between a person's age and their engagement duration.

* Age vs. Opportunity Duration (Days)

Correlation of -0.02. Again, a very weak negative correlation. Age has almost no linear relationship with the duration of the opportunity.

* Age vs. Combined Score.

Correlation of 0.01. A very weak positive correlation, close to zero. Age has virtually no linear impact on the combined score.

* Engagement duration vs. Opportunity Duration (Days)

Correlation of 0.83. This is a strong positive correlation. It indicates that as the opportunity duration increases, the engagement duration in days also tends to increase significantly. There's a strong linear relationship between these two variables.

* Engagement duration vs. Combined Score

Correlation Of -0.11. This is a weak negative correlation. While there's a slight tendency for the combined score to decrease as engagement duration increases, the relationship is not strong.

* Opportunity Duration (Days) vs. Combined Score

Correlation of 0.47. This is a moderate positive correlation. It suggests that as the opportunity duration in days increases, the combined score tends to increase as well, but not as strongly as the relationship between engagement duration and opportunity duration.

**Summary of Key Relationships**

* Strongest Positive Correlation: "Engagement duration" and "Opportunity Duration (Days)" (0.83). This is the most notable relationship, implying that longer engagement periods are highly associated with longer opportunity durations.
* Moderate Positive Correlation: "Opportunity Duration (Days)" and "Combined Score" (0.47).
* Weak/Negligible Correlations: "Age" shows very little linear correlation with any of the other variables. "Engagement duration" and "Combined Score" also have a very weak negative relationship.

In essence, this heatmap helps quickly visualize which variables move together (positively or negatively) and the strength of those linear relationships. The most impactful relationship here seems to be between engagement duration and the length of the opportunity.

Outlier analysis

* For Age, Engagement Duration (Days), and Opportunity Duration (Days), there are no statistical outliers based on the interquartile range (IQR) method. All values fall within the expected range for these variables.
* For Combined Score, there are 19 outliers. The normal range is from -232 to 848, but the actual data ranges from -268 to 895. This means a small number of learners have unusually low or high combined scores, which could be due to negative engagements (signups before opportunity starts) or longer engagement duration.

**Summary of Outlier analysis**

|  | Age | Engagement duration (days) | Opportunity duration (days) | Combined score |
| --- | --- | --- | --- | --- |
| Normal range | 18.50 to 30.50 | -1240.50 to 835.50 | -1104.00 to 2008.00 | -232.00 to 848.00 |
| Actual range | 19.00 to 30.00 | -998.00 to 569.00 | -7.00 to 841.00 | -268.00 to 895.00 |
| No of Outliers | 0 | 0 | 0 | 19 |

Relevant statistics for numerical variables

The count, mean, standard deviation, minimum value, maximum value, 25%, 50% and 75% value of numerical variables.

**Summary Statistics for Numerical Variables**

Age Engagement Duration

count 2847.000000 2847.000000

mean 24.144714 -175.404636

std 2.518127 324.315961

min 19.000000 -998.000000

25% 23.000000 -462.000000

50% 24.000000 -127.000000

75% 26.000000 57.000000

max 30.000000 569.000000

Combined Score Opportunity Duration

count 2847.000000 2847.000000

mean 314.314015 465.573937

std 205.560449 364.868741

min -268.000000 -7.000000

25% 173.000000 -63.000000

50% 341.000000 468.000000

75% 443.000000 841.000000

max 895.000000 841.000000

1. **DISCUSSION**

This section interprets the key findings, explores potential reasons behind observed trends, considers the implications of these insights and suggests actionable solutions to improve users engagement, reduce drop offs and increase overall efficiency. It discusses engagement and drop off trends and behaviours across time, age, gender and opportunity categories.

Engagement Across Opportunity Categories

* Courses: dominate in terms of participation, with high numbers in Started and Team Allocated statuses. This suggests strong initial engagement.
* Internships: show a notable number in the Dropped Out category, meaning retention may be a challenge in this area.
* Competitions: have relatively low overall participation, potentially indicating limited interest or entry barriers.
* Event category has a strong presence in the Team Allocated status, meaning structured teamwork may play a vital role in participation.
* Engagement: This category has the lowest number, with about 100 learners overall, suggesting limited participation in interactive or community-driven opportunities.

**Possible Insights**

* Courses and Events seem to foster more participation and team involvement.
* Internship dropout rates may need attention—perhaps introducing more engagement strategies or better support systems.
* Competitions could benefit from initiatives to increase interest or accessibility.
* Rewards systems may need improvement if motivation and completion rates are low.

Completion rate by Opportunity duration (Dropout Analysis)

**High Dropout Category** : Internship category has the highest dropout rate, which may suggest:

* Challenges with long-term engagement: students or participants may start but struggle to complete the program.
* Possible mismatch between expectations and reality: some individuals may realize the internship doesn’t align with their interests or goals.
* Workload or commitment difficulties: balancing an internship with other responsibilities might lead to dropouts.

**Moderate Dropout Category**: Courses show a relatively high dropout rate but not as extreme as internships. Potential reasons include:

* Difficulty level: some participants may struggle with the subject matter.
* Lack of motivation: if rewards or outcomes aren’t clear, participants may lose interest.
* Time constraints: courses require consistent effort, which some individuals may find overwhelming.

**Low Dropout Categories:** Competition, Event, Engagement have lower dropout rates meaning participants tend to stick with them. Possible reasons for lower dropout rates:

* Short-term nature: events and competitions typically have a clear start and end date making them easier to commit to.
* Higher engagement levels: group participation and structured challenges may keep people motivated.
* Less demanding commitments: these categories often require less long-term involvement than internships or courses.

**Key Insights & Recommendations**

* Internships may need better onboarding strategies to ensure participants understand the expectations before committing.
* Courses could benefit from additional support mechanisms like mentorship, flexible schedules, or incentives.
* Refined rewards and motivation strategies could help decrease dropout rates further in all categories.
* For competitions and events, leveraging their engagement models in other categories might improve retention rates.

Engagement trends across age groups

The correlation matrix for Engagement duration and age groups shows no significant relationship between the variables. This invariably suggests that how long a learner engages with opportunities is independent of age or age groups.

**Key Insights and Recommendation**

Age, as shown, is not a significant predictor or factor influencing engagement duration.

Recommendation: Do not focus marketing or campaigns on age as a primary differentiator for engagement duration.

**Engagement and Completion trends across gender categories**

**Engagement duration:** The key insights across gender groups includes**:**

* Both Male and Female categories have similar engagement distribution patterns, but Male participants show slightly more variability in engagement duration.
* The engagement duration distribution for "Don't want to specify" appears less variable, meaning this group shows more consistency in engagement duration compared to the Male and Female categories.

Considering these deductions from the visualisation, gender category seems to have no influence on how long/short it takes a user to engage. Data of longer duration might however show a relationship.

**Completion status**

* Male: Highest total participation. Majority are in the "Team Allocated" status. Significant numbers in "Started", "Dropped Out", and "Rewards Award". A few have withdrawn.
* Female: Fewer participants compared to males. Most also fall under "Team Allocated", with a notable number having started. Visible presence in "Dropped Out" and "Rewards Award". Very few have withdrawn.
* Don't Want to Specify: Extremely low participation. Only a small count in "Team Allocated"; negligible in other categories.

**Key Insights**

* Males have the highest participation, mostly "Team Allocated", but with notable dropouts and some withdrawals.
* Females are fewer, but show similar trends—good start, but also dropouts.
* "Don't Want to Specify" group has very low participation and minimal progress.

**Recommendations**

* Improve retention across all gender groups.
* Promote inclusivity for non-specified gender to boost engagements.
* Track data regularly to adjust strategies and support all groups better.

Application duration trend

The“Engagement duration trend over time” graph and “Application lead time” histograms informed us that:

* The engagement durations vary widely, with some users showing prolonged activity while others engage quickly.
* More recent users tend to have shorter engagement durations compared to earlier ones.
* The largest group of learners applied on the exact day an opportunity started. This suggests some learners showed interest in an opportunity before it started.

The most notable insight is the longer duration it takes for existing users to engage with opportunities compared to new users. Here are some recommendations to optimise engagement duration for existing users:

* Personalized Recommendations: Implement or enhance a recommendation engine that suggests relevant opportunities to existing users based on their past engagement, major, stated interests, and profile data. This reduces the time they spend searching.
* Dashboard Enhancements: Design the user dashboard for existing learners to prominently feature new or highly relevant opportunities. Consider "Recommended for You" sections or "Opportunities You Might Like" even before the opportunity starts.
* Filter and Search Optimization: Ensure that search and filter options are highly intuitive and efficient for existing users who may be looking for very specific types of opportunities.
* Targeted Email Campaigns: Send personalized email digests to existing users highlighting new opportunities that match their profiles, rather than generic blasts.
* In-Platform Notifications: Utilize in-platform notifications for new opportunities, deadlines, or relevant content that could prompt quicker engagement.
* Community Building: Foster a sense of community among existing users. Engaged communities can lead to peer recommendations and increased awareness of new opportunities.
* "What's New" Section/Digest: Create a dedicated section or send a regular "What's New” digest to existing users, focusing specifically on newly added opportunities.
* Address Potential User Fatigue/Overwhelm (Opportunity Pacing): If opportunities are launched too frequently or are too numerous, existing users might feel overwhelmed. Explore if a more curated or staggered release of opportunities for returning users could be beneficial.
* Feedback Loops: Implement surveys or feedback mechanisms specifically for existing users to understand why they might be taking longer to engage. Are they finding the opportunities less appealing, too demanding, or simply missing them?
* Clear Value Proposition for Continued Engagement: Clearly articulate the benefits of continued engagement for existing users (e.g., skill development, career progression, networking).
* Early Bird Access/Perks: Consider offering existing users "early bird" access to highly sought-after opportunities or small incentives for applying within a certain timeframe. This should however be used cautiously to avoid devaluing the opportunities themselves).
* Badges/Recognition for Consistent Engagement: Gamification elements like badges for consistent engagement or quick application could encourage faster action.

**Seasonal trends in learner Signup**

According to the monthly and weekly trend plots, here are some key Insights:

* Mid-Year Peak in 2023 (May - August): This suggests a potential alignment with academic breaks (e.g., summer holidays in many parts of the world) when students might have more time to explore and engage with online learning opportunities.
* Early and Late-Year Dips in 2023(January, February and November). This could correlate with the start and end of academic semesters, exam periods, or other commitments that reduce free time.
* Significant Drop in 2024 (Early Months): The most striking insight is the dramatic drop in sign-ups in early 2024 compared to 2023. While January 2023 had 76 sign-ups, January 2024 had 108, which is an increase. However, February 2024 has 201 sign-ups, significantly higher than February 2023's 93. This represents a substantial positive trend for the very beginning of 2024, indicating a potentially different pattern or successful initiatives.
* Correction/Refined Insight: Looking at the color intensity, February 2024 (201) is clearly higher than January 2024 (108). Compared to 2023, January 2024 is higher than January 2023. February 2024 is significantly higher than February 2023. March 2024 and April 2024 show very low numbers (20 and 15 respectively), which is a steep decline from earlier months in 2024 and significantly lower than the corresponding months in 2023.
* Root Cause Analysis: What changed on the platform, marketing efforts, or external factors (e.g., economic conditions, new competitors, academic calendars shifted) that led to such low numbers compared to 2023 and even compared to the early months of 2024?

The trend overly suggests external influences such as marketing pushes, seasonal factors, or platform updates affecting signup activity.

**Recommendations**

* Optimize Campaign Timing with Academic Cycles: Recognizing that the majority of learners are academically active, (age factor) strategically aligning major recruitment campaigns with their academic holiday periods is recommended. This approach leverages their increased availability during breaks, fostering higher engagement and sign-up rates for experiential learning opportunities.
* Diversify Learner Acquisition Strategies: The observed decline in 2024 sign-ups, suggested to be potentially influenced by a focus on existing users acquisition strategies, highlights the need to broaden outreach. There's a proposed need to actively explore and implement new advertising channels and marketing initiatives to effectively reach and acquire a wider audience of prospective learners, ensuring sustained growth and platform visibility.

1. **CONCLUSION**

The EDA report offers a comprehensive view of user engagement, sign-up behavior, and opportunity dynamics. The findings highlight clear areas for intervention such as internship retention and marketing timing. Leveraging these insights will improve user experience, increase retention, and drive more effective program delivery.

**Next Steps:**

In Week 3, we will:

* Predictive modelling: Use features like”Combined score” to forecast engagement outcomes.
* Churn analysis: Use churn analysis to analyse learner drop offs.