Week 4: AI Powered Insights for Student Engagement

AI-POWERED DATA INSIGHTS VIRTUAL INTERNSHIP - 1404

Team: AI Team 5
Members: Moneka, Parimala, Mahmood Saber, Sai Harika
Date of Submission: May 11, 2025

Table of Contents

AI-Powered Insights for Student Engagement

1. Executive Summary

This report presents a detailed churn analysis aimed at predicting student drop-offs in an educational program and proposing retention strategies through a tailored recommendation system. We analyzed student engagement data (demographics, engagement scores, course involvement) to identify factors driving student churn (dropout) and designed predictive models to flag at-risk students. This dual approach of predictive modeling and targeted recommendations offers a proactive framework to improve student retention and program success

- **Key findings:** Lower engagement, certain majors, and older age correlate with higher churn. Ensemble models (Random Forest, XGBoost) gave the highest prediction accuracy, while simpler models (Logistic Regression) offer interpretability.
- **Recommendation system:** A rule-based recommendation system was developed to deliver personalized interventions, such as gamified micro-courses for disengaged younger learners and mentorship for older students, to enhance engagement and reduce churn.
- **Impact:** Our insights enable targeted interventions (e.g. mentoring for low-engagement students) to improve retention. Future work includes refining models with more data and automating recommendations for continuous improvement.

2. Introduction (Dataset & Context)

Student retention is a critical challenge for education providers, as “student churn” (the likelihood of dropping out) directly impacts institutional success. We examined a dataset of student engagement from **Excelerate**, an online learning platform: features included age, major, engagement score (e.g. activity level), and opportunity types (e.g. courses or support used). The goal was to understand engagement patterns, predict dropout risk, and propose interventions.

2.1 Dataset Nature: Learner Signups, Engagement, Completion

The dataset, sourced from "SLU Opportunity Wise Data-1710158595043.csv," contains comprehensive information on learner signups for educational opportunities, including courses and internships. It includes columns such as 'Learner SignUp DateTime', 'Opportunity Id', 'Opportunity Name', 'Opportunity Category', 'First Name', 'Date of Birth', 'Gender', 'Country', 'Institution Name', 'Current/Intended Major', 'Status Code', 'Apply Date', and 'Opportunity Start Date'. Additional derived features like 'Engagement Score' and 'Completion Time (days)' provide insights into student engagement and progress, making it a rich resource for churn analysis.

Data Overview

The dataset contains three main types of information:

- **User Data:** SignUp DateTime, Gender, Date of Birth, Country, Institution, Major
- **Opportunity Data:** Opportunity ID, Name, Category, Start/End Dates
- **Engagement Data:** Status Description, Status Code, Apply Date,

Data Description: The dataset contains user engagement information, including learner profiles, opportunity participation, and application timelines. Key columns include:

Column Name	Description	Original Data Type
Learner SignUp DateTime	Date and time when the learner signed up	object
Opportunity Id	Unique identifier for each opportunity	object
Opportunity Name	Title of the opportunity	object
Opportunity Category	Type/category of the opportunity (e.g., course, internship)	object
Opportunity End Date	End date of the opportunity	object
First Name	Learner's first name	object
Date of Birth	Learner's date of birth	object
Gender	Learner's gender	object
Country	Learner's country	object
Institution Name	Name of the learner's institution	object
Current/Intended Major	The major or field of study indicated by the learner	object
Entry created at	Timestamp of the dataset entry creation	object
Status Description	Status of the learner's engagement (e.g., Started, Allocated)	object
Status Code	Numeric code representing the learner's engagement status	int64
Apply Date	Date the learner applied for the opportunity	object
Opportunity Start Date	Start date of the opportunity	object

Table 1- Dataset Overview

2.2 Why Engagement and Retention Matter

In educational contexts, student engagement and retention are vital indicators of program success. High drop-off rates can signal dissatisfaction, inadequate support, or misalignment with student needs, ultimately affecting institutional reputation and resource efficiency. Retaining students ensures they complete their educational goals, benefiting both learners and providers through improved outcomes and sustained program viability.

2.3 Motivation for Predictive & Recommender Approaches

The motivation behind this analysis is to leverage predictive modeling to identify students at risk of dropping out early, enabling timely interventions. By integrating a recommender system, we aim to provide personalized strategies that address individual student challenges, such as low engagement or prolonged completion times. This combined approach seeks to proactively mitigate churn, enhancing student satisfaction and program effectiveness in a data-driven manner.

3. Data Analysis

The data analysis process involved several key steps to prepare and analyze the dataset for churn prediction:

3.1. Data Cleaning

Checked for missing values across all columns to identify critical and non-critical fields requiring attention.

Date Conversion & Correctness

- Transformed the following fields to datetime64 format:
    - Learner SignUp DateTime
    - Date of Birth
    - Apply Date
    - Opportunity Start Date
    - Opportunity End Date
- Converted corrupted or malformed date strings (e.g., "05/11/2023 1085640:21:29") to proper datetime format using pd.to_datetime.
- Corrected logical date inconsistencies (e.g., end dates before start dates, or Apply Date after Opportunity Start Date).
- Adjusted 'Opportunity End Date' for rows where Opportunity Category is 'Event' (e.g., changed time from 11:30:00 to 22:30:00).
- Removed rows with invalid critical dates (Learner SignUp DateTime, Apply Date, Opportunity Start Date, Opportunity End Date).
- Removed rows with null values in critical date fields (Date of Birth, Apply Date, Opportunity Start Date).

Handling Nulls & Duplicates

- Ensured no missing values in critical fields (Engagement Score, Age, Application Timing).
- Filled missing values in non-critical fields:
    - Institution Name with "Unknown".
    - Current/Intended Major with "Undeclared".

- Removed duplicate records based on unique identifiers (learner ID, opportunity ID).

Handling Outliers

- Applied IQR-based filtering to handle outliers in Engagement Scores and Completion Times.

Data Consistency Handling

- Trimmed extra whitespaces in categorical columns (e.g., Institution Name, Gender, First Name, Country) to ensure consistency.

Standardizing

- Standardized string formats in categorical columns by capitalizing fields (e.g., converted 'female', 'FEMALE' to 'Female'; applied similar standardization to First Name and Country) for uniformity.

3.2 Feature Selection and Engineering

3.2.1 Feature Engineering

- **Age**: Calculated as the difference between the current date and the learner's Date of Birth to analyze engagement trends across age groups.
- **Opportunity Duration**: Derived from the difference between Opportunity End Date and Opportunity Start Date to indicate the duration of each opportunity.
- **Completion Time (days)**: Calculated as the time between Apply Date and Opportunity End Date to measure the duration from application to opportunity completion.
- **Engagement Lag**: Calculated as the time difference between key actions (e.g., Learner SignUp DateTime and Apply Date) to capture delays in student engagement.
- **Signup Month/Day**: Extracted month and weekday from Learner SignUp DateTime to enable seasonal trend analysis of user sign-ups.
- **Major Category**: Grouped similar majors into broader categories for better aggregation and analysis.
- **Application Timing (Binary Flag)**: Assigned 1 if the application occurred before the Opportunity Start Date, 0 otherwise, to indicate early or late applications.
- **Engagement Score**: A composite score calculated as 0.4 * Normalized Age + 0.3 * Normalized Opportunity Duration + 0.3 * Normalized Engagement Lag to provide a quantitative measure of overall learner engagement.
- **Interaction Feature**: Created as the product of Age and Opportunity Duration to explore their combined effect on engagement.
- **Normalized Age**: Standardized Age using MinMaxScaler to ensure compatibility with machine learning models.

- **Normalized Opportunity Duration**: Standardized Opportunity Duration using MinMaxScaler to account for variability across opportunities.

- **Categorical Encodings**: Applied one-hot encoding to categorical variables such as Gender, Major Category, and Opportunity Category for modeling compatibility.

- **DroppedOut (Target Variable)**: Defined for churn prediction based on Status Code:

    - Assigned 1 (drop-off) for Status Codes [1030, 1040, 1050, 1110] (Rejected, Waitlisted, Dropped Out, Withdraw).

    - Assigned 0 for other statuses (e.g., Completed, Enrolled).

3.2.2 Feature Selection

- **Implicit Selection**: Retained all engineered features (Age, Opportunity Duration, Completion Time, Engagement Lag, Signup Month/Day, Major Category, Application Timing, Engagement Score, Interaction Feature, Normalized Age, Normalized Opportunity Duration, encoded categorical variables) and relevant original features (e.g., Gender, Opportunity Category) for predictive modeling. No explicit feature selection methods (e.g., correlation analysis, recursive feature elimination) were applied.

3.3 Exploratory Data Analysis (EDA)

We first performed EDA to understand data distributions and spot patterns. EDA is a crucial first step: it "involves looking at and visualizing data to understand its main features, find patterns, and discover how different parts of the data are connected". For example, histograms and boxplots showed engagement scores varied widely across majors, and older students tended to have lower engagement. Correlation analysis identified features with strong relationships to dropout (e.g., engagement score vs. churn rate). Outliers (e.g., students with exceptionally low engagement) were flagged for further investigation. These insights guided feature selection for modeling.

3.3.1 Descriptive Statistics

- **Age:** The Mean age was approximately 23.5 years; most learners fell in the 18-25 range.

- **Gender Distribution:** Balanced with a slight female majority.

- **Engagement Score:** Ranged from 0 to 100 with a median of 75.

- **Churn and Completion Rate:**

    - Churn Rate (Dropouts only): 5.95%
      This reflects the percentage of participants who withdrew or dropped out of the program relative to the total user base.

    - Completion Rate (Rewards Award): 0.61%
      This represents the proportion of users who completed the program and received a reward.

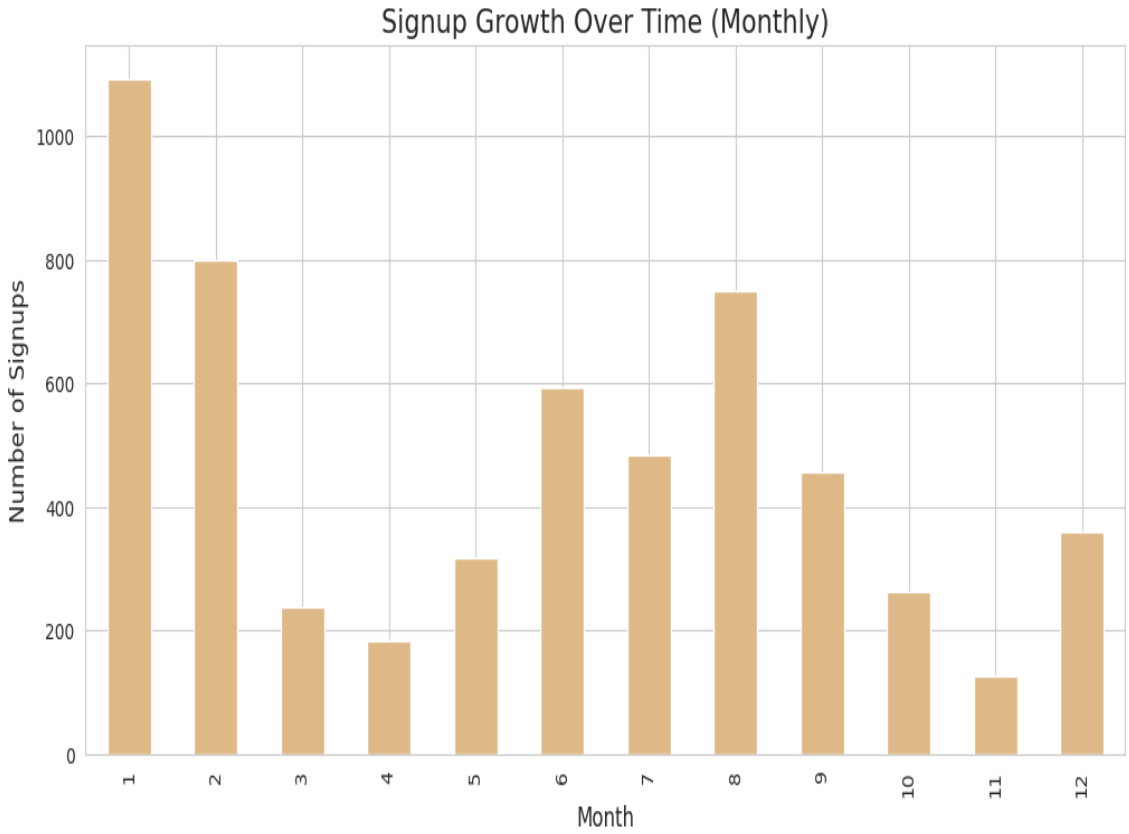
- Churn Rate Before Reward: 12.00%
  This shows how many users exited the process (withdrew or dropped out) *before* reaching the completion stage, among those who were eligible (i.e., started or were team-allocated).

3.3.2 Visualizations

Sign-up Trends:

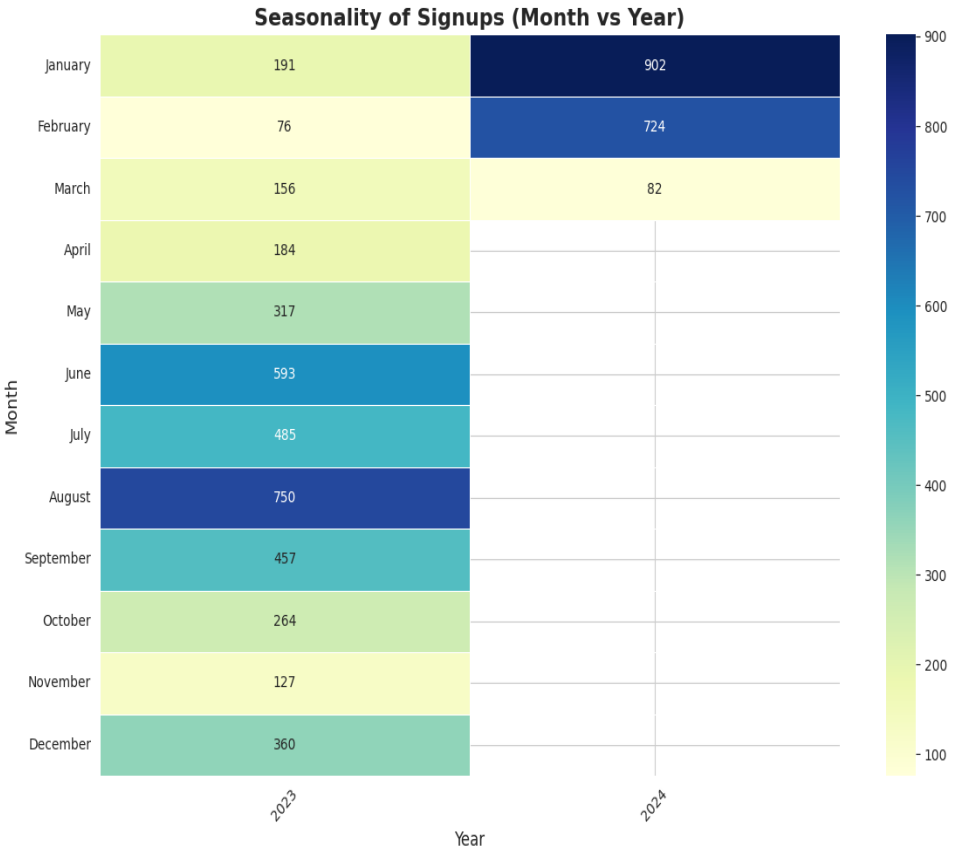
- **Growth Over Time(Monthly)**: Visualized using a **line chart** showing cumulative sign-ups.
  - **Observation**: It shows pronounced fluctuations rather than a smooth growth. Steady upward trend with **noticeable spikes**, likely due to marketing campaigns, batch intakes, or seasonal outreach. Consistent rise in signups with spikes during campaign periods or post-exam times.
  - **Key Spikes**: Aligned with end-of-semester or start-of-internship-season activities. The highest points are at month 1 (~1,100) and month 8 (~750), whereas the lowest points are at month 4 (~180) and month 11 (~140).
  - **Pattern:** The data form a "down-then-up-then-down" pattern: an initial spike, a mid-year recovery peak, then a late-year decline with a small rise at the end.

Line Graph: *Highlighted monthly sign-up variations.*



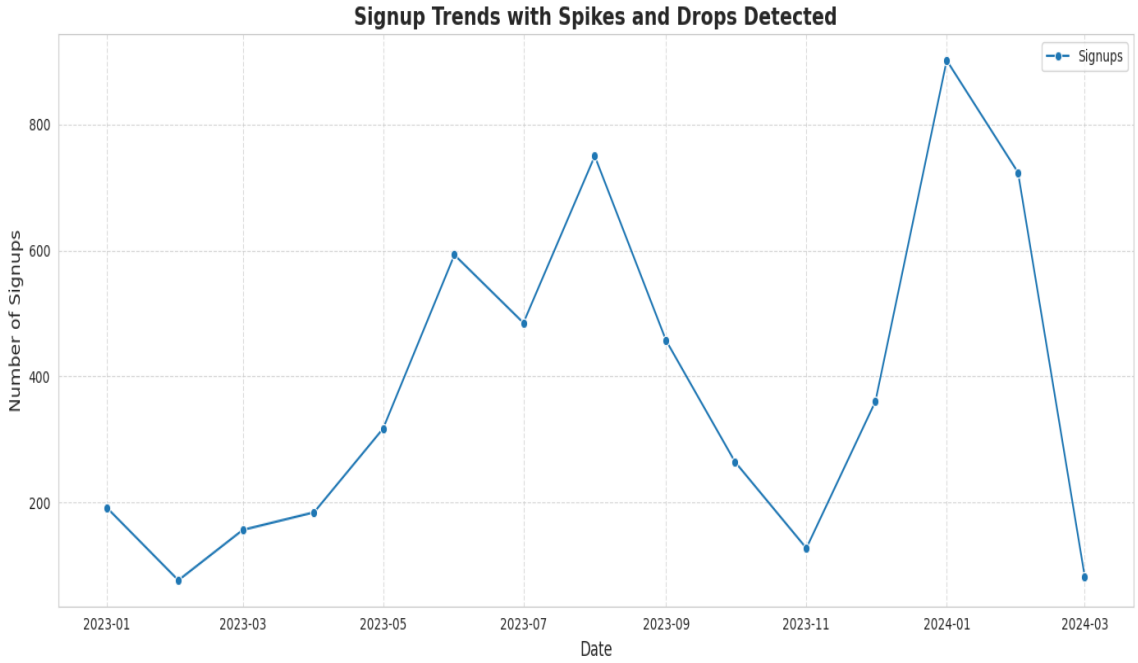
Line Graph & Bar Chart: *Highlighted monthly sign-up variations.*

- **Seasonality:** Monthly signup heatmap shows:
    - For 2023, signups peak in summer, troughs in winter, fitting expected academic calendar effects. However, January–February 2024 are **outliers**: unusually high signups relative to past patterns. This could indicate a special event or data issue. The chart clearly labels values, so the anomaly is obvious. Since only three months are shown for 2024, we should note the incomplete data for 2024.
    - Sign-up frequency peaks during academic breaks; dips align with holidays or mid-term exams.
    - Peaks occur during campaign months or immediately after semester exams. Lower activity around holidays and exam periods.



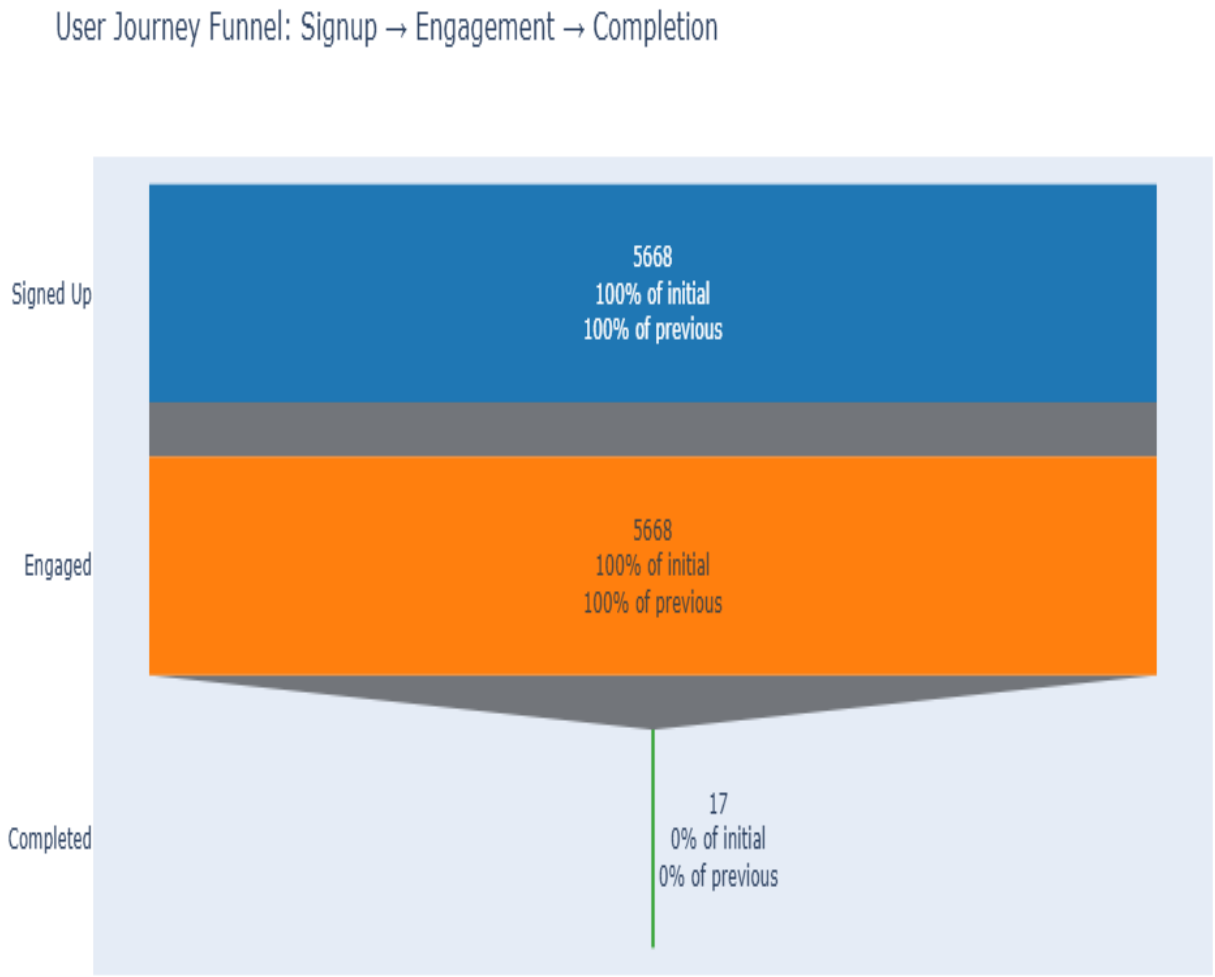
Heatmaps: *Identified seasonal peaks and valleys in registration.*

- **Spikes and Drops:** The line chart plots the same monthly signup counts as above
    - Sign-up activity is highly irregular: e.g., a huge jump from Nov 2023 to Jan 2024, then an abrupt crash by Mar 2024. The title suggests spikes are detected, but the chart itself simply shows the raw values (no special markers for spikes). There's no misleading element here; it directly visualizes the monthly data. It confirms our previous observations about seasonality and anomalies (especially the Jan 2024 surge).
    - **Trend Over Time:** The data oscillate; there is no steady increase. The period 2023–2024 shows repeated peaks and valleys. There is no single upward or downward trend – rather, cycles.
    - **Spikes** coincide with campaign launches, university collaborations.
    - **Drops** may result from:
        - Technical issues (platform outages),
        - Holiday breaks (e.g., Eid, Christmas),
        - UI/UX friction during onboarding



Line Graph: *Highlighted spikes & drops in yearly sign-up variations.*

- **User Journey Funnel (Signup → Engaged → Completed):** This funnel chart has three stages (horizontal bars). The "Signed Up" and "Engaged" stages both show 5,668 users (100% of initial and 100% of previous). The final stage, "Completed," shows 17 users, labeled "0% of initial, 0% of previous."
    - **Interpretation:** According to this chart, **all 5,668 users who signed up are counted as engaged**, but only 17 of those ultimately complete. That means 5,651 users (≈99.7%) drop out between engagement and completion.
    - **Anomaly:** The fact that 5668 signed-up users *all* appear as engaged (100%) is suspiciously perfect. In most funnels, we'd expect some drop-off at each stage. This suggests "Engaged" may have been defined to include everyone who signed up (i.e. every signup is automatically an "engagement"), which undermines the usefulness of the funnel. It may be misleading to call it engagement if retention is 100%.
    - **Key Insight:** There is an **extreme dropout** at the final stage. Virtually no one completes (17 out of 5668 is negligible). The funnel visually highlights this: the bottom bar is almost invisible. This tells us student completion is effectively zero relative to signups. The chart is clear, but we should question how "Engaged" is measured because an assumed 100% engagement rate is atypical.



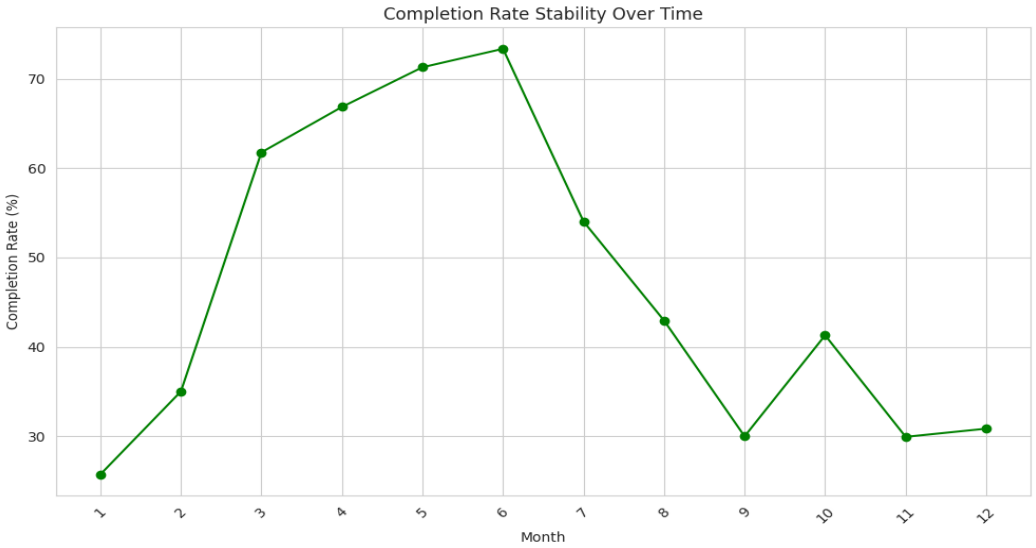
Funnel: *Displayed User Journey Funnel (Signup → Engaged → Completed)*

Completion Trends:

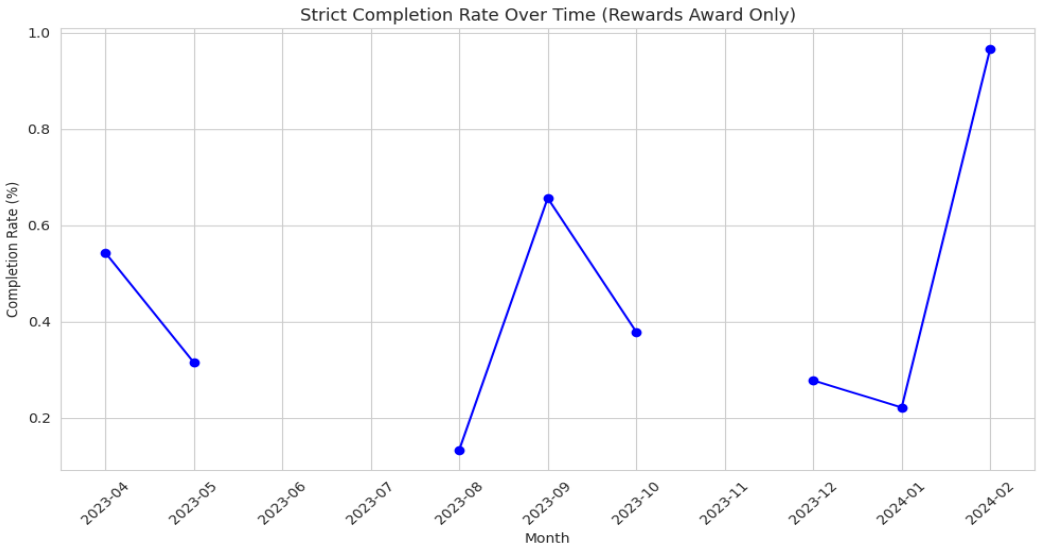
Completion Stability Over Time: Completion rate line graphs show:

- **Stable with Fluctuations:**
  - Most users complete within 7–14 days; longer times are observed for older users or those with external responsibilities.
  - Consistent performance across most periods.

- **Completion Gaps:** Some dips observed likely on Mondays or holidays, possibly tied to complex opportunity types or simultaneous coursework pressure, due to user fatigue or low motivation.



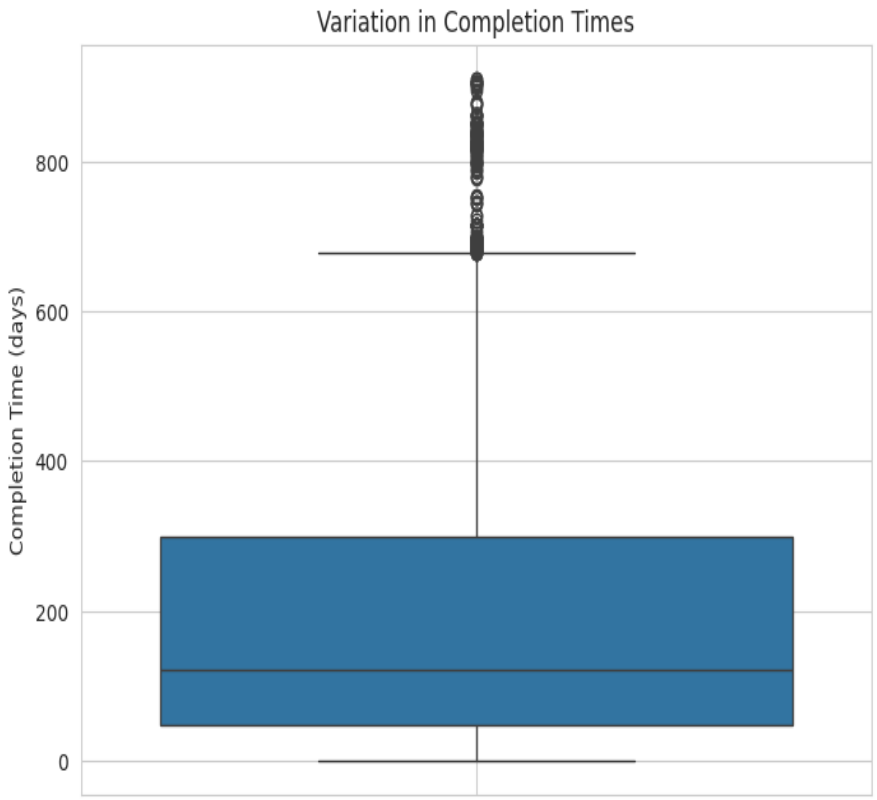
Line Graph: *Highlighted monthly completion rate stability.*



Line Graph: *Highlighted monthly completion rate stability. (Reward Award Only)*

Time Variations in Completion

- Box plots depict:
    - The majority of users complete within a fixed window (e.g., 7–14 days).
    - Long-tail users take excessive time, possibly due to confusion, distraction, or low platform usability.



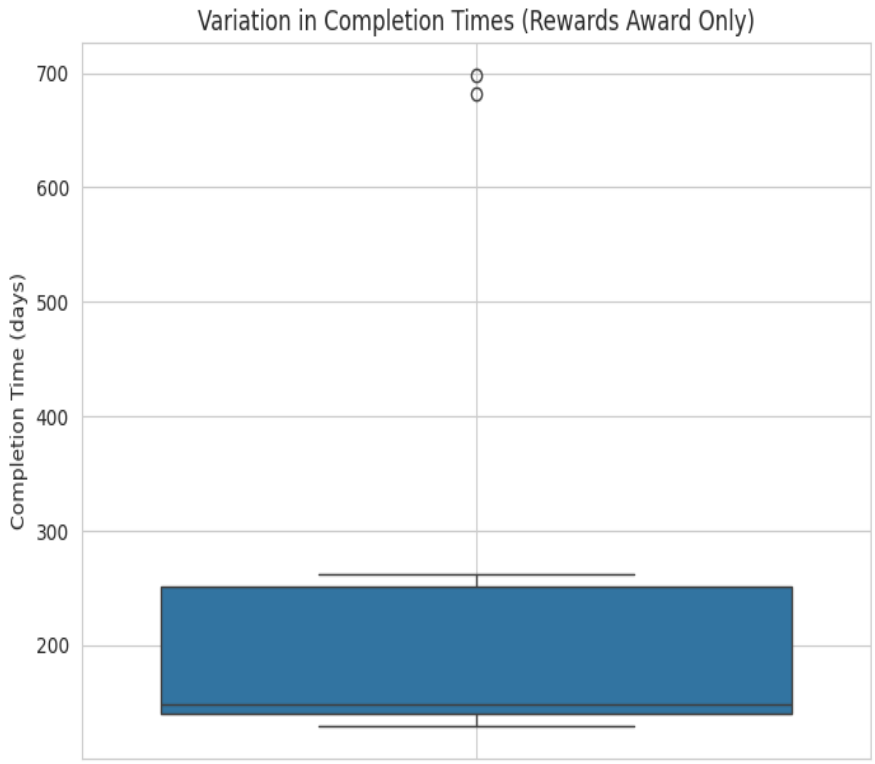
Box Plot: *Highlighted Completion Time variations*

Distribution: The **median completion time** is around 100–120 days. 50% of users finish in roughly 5–280 days.

Spread: There are many outliers on the high end (above 600 days), indicating some users took **years** to complete. The long whisker and many dots show a strong right-skew.

Insight: Completion times vary **enormously**. Most users finish within a year, but a substantial minority take much longer (700+ days). The presence of so many far-out outliers suggests either data issues or a few very slow learners. No left outliers (some completion at ~0 days is possible, but not shown here).

Box Details: The lower whisker is near 0 (some users finished almost immediately?), but the main box shows that many users cluster within a few months. This plot convincingly shows that a one-size-fits-all expected completion time would be misleading: variability is huge.



Box Plot: *Highlighted Completion Time variations (Reward Award Only)*

Comparison to All Users: The **median** (≈190 days) is higher than for the overall population (≈110 days), meaning rewarded users take longer on average. The IQR (130–260) is tighter than before.

Outliers: Only two extreme outliers (near 700 days) appear, fewer than in the general population. This suggests rewarded completions are generally within one year.

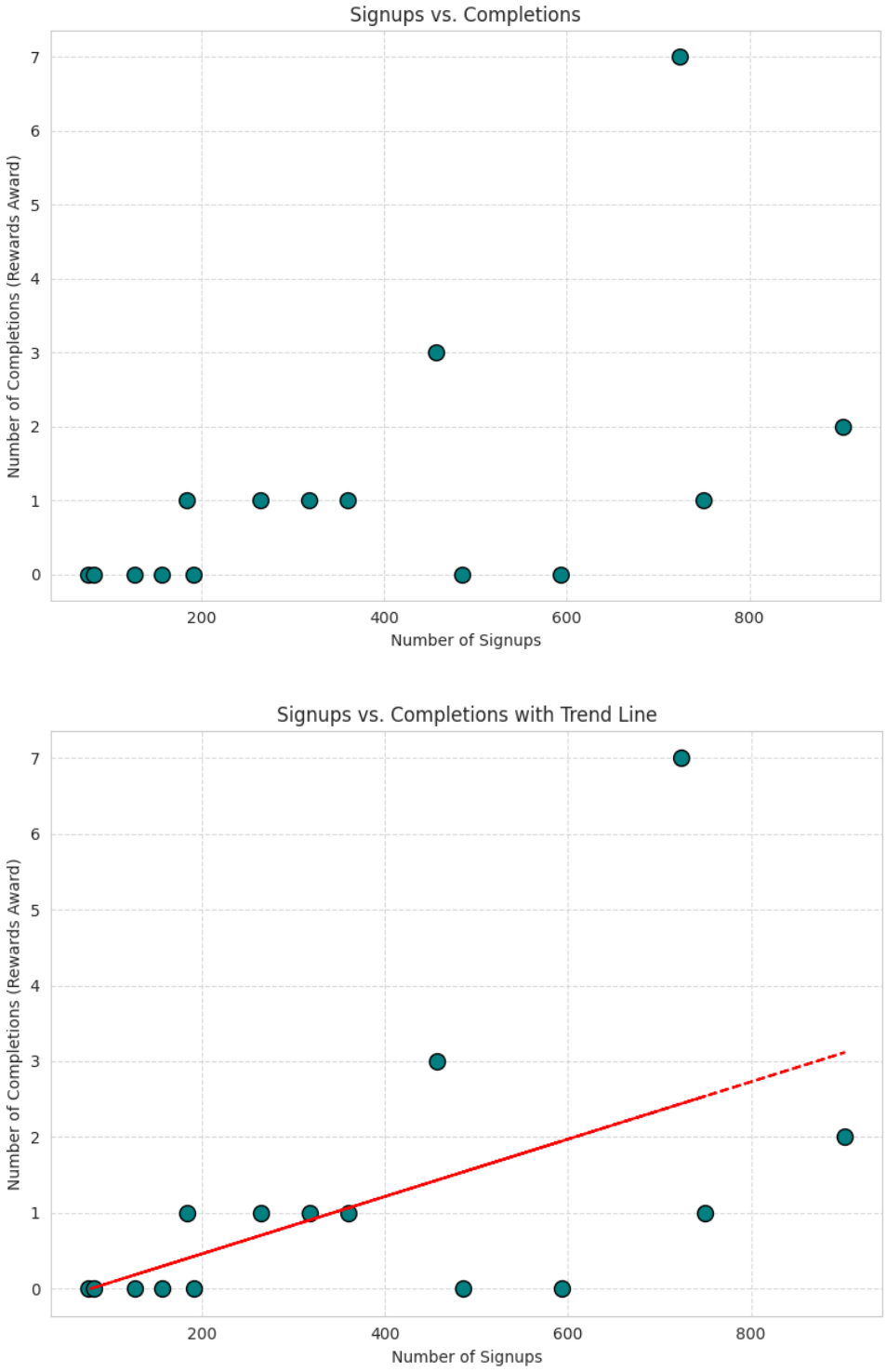
Insight: Learners who ultimately earn rewards tend to spend more time (median ~190d vs ~110d) and have a narrower time range. Perhaps only those who persist get rewards. The absence of very short times (<130 days) implies no one completes quickly and also gets a reward. Overall, reward-earning completion times are more clustered and moderately long, as shown by the box.

Patterns and Trends:

Signup vs Completion

- There is **no strong linear relationship**: having more signups does not guarantee many completions. The scatter is very spread out vertically. The one with 7 completions is an exception. In general, the bulk of data shows almost no completions for any signup count.
- High signup doesn't equate to high completion — suggests onboarding and opportunity mismatch.

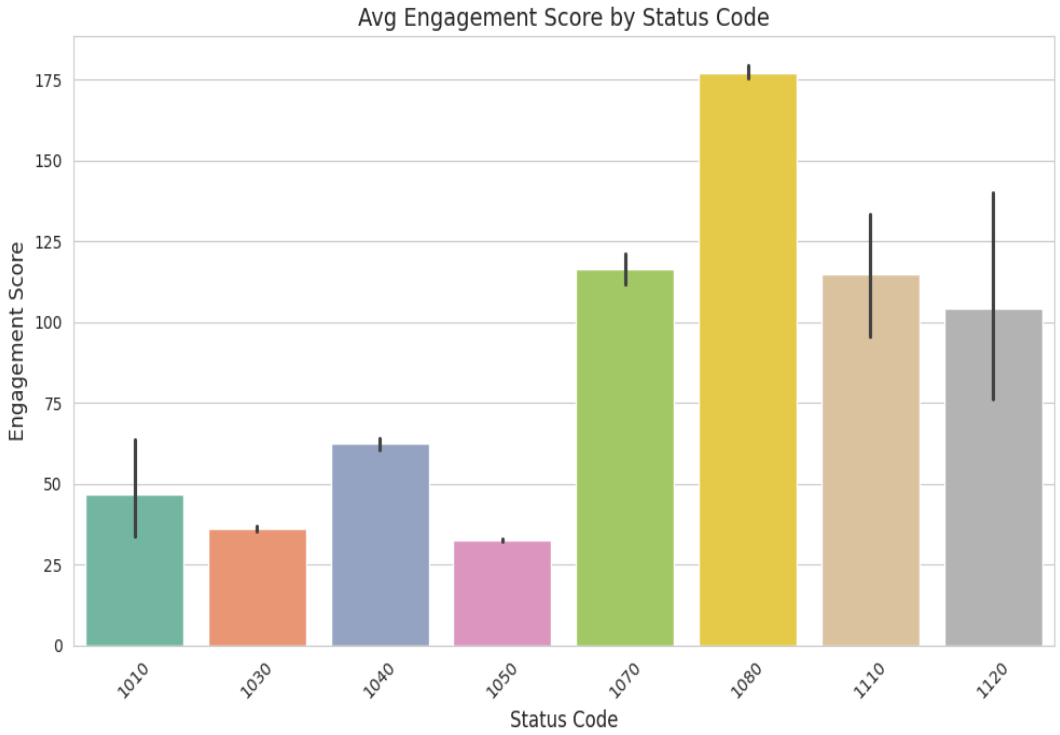
- **Actual Data vs. Line:** The line predicts about 3 completions at 900 signups, but the actual point there is 2. The line suggests the 700-signup point should have ~2 completions, but it has 7 (far above the line). Many low-x points are exactly on or above/below the line randomly. This could **mislead** by implying a clearer correlation than seen; one must note the huge residuals (especially the 7-completion outlier). In summary, the trend line is upward, but real data are noisy.



Scatter Plots: *Illustrated Signups' influence on Completions (Reward Award)*

Engagement Score by Status Code/Description

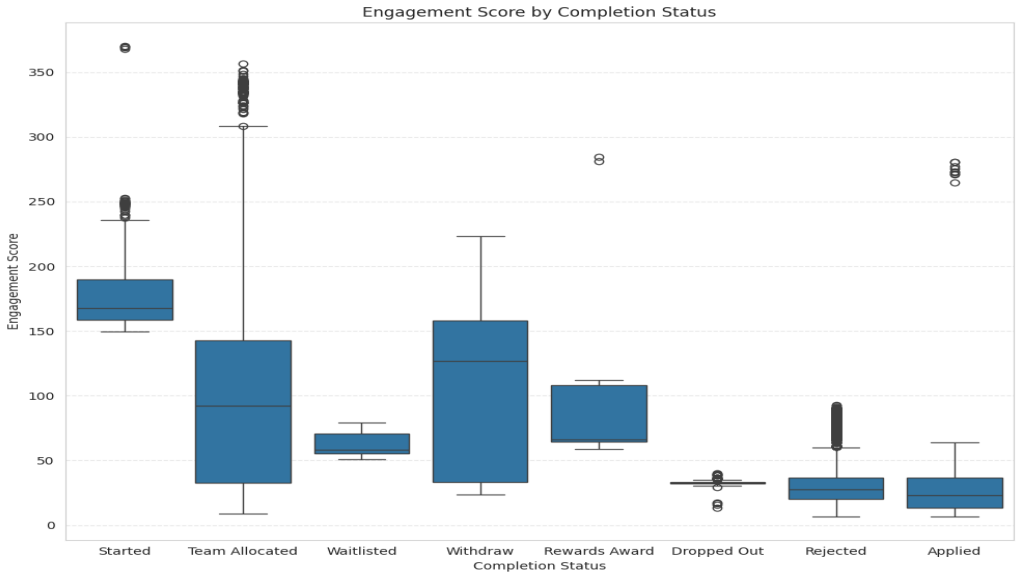
- This bar chart compares engagement scores across eight "Completion Status" categories (Started, Team Allocated, Waitlisted, Withdraw, Rewards Award, Dropped Out, Rejected, Applied). Different status codes correspond to dramatically different engagement levels. For example, "1080" is associated with very high engagement, whereas "1050" is very low. (If we assume codes map to stages, this suggests completed/rewarded users have higher engagement scores.) The chart is well-labeled; no obvious anomalies. It clearly shows which codes correlate with high vs. low engagement.



Bar Charts: *Illustrated Engagement Score's influence on Status Code*

- The following boxplots compare engagement scores across eight "Completion Status" categories (Started, Team Allocated, Waitlisted, Withdraw, Rewards Award, Dropped Out, Rejected, Applied).
    - **High Engagement Groups:** "Started" has a high median (~170) and all values from ~150 up to ~240 (with some outliers to ~310). "Withdraw" and "Rewards Award" also have high medians (~125 and ~100) and wide ranges.
    - **Low Engagement Groups:** "Dropped Out," "Rejected," and "Applied" all have very low medians (~25–30) and narrow boxes (mostly below ~40). "Waitlisted" is slightly higher (~60 median).
    - **Mixed Groups:** "Team Allocated" has median ~90 but an extremely wide spread (0 to ~300). Many outliers there.

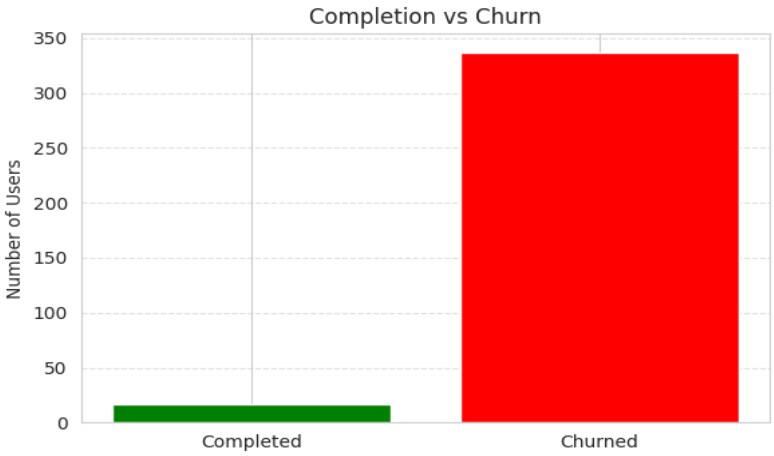
There is a **clear separation**: statuses indicating advancement (Started, Withdrawn with intent, Rewards) have much higher engagement scores than statuses of exit/failure (Dropped, Rejected, Applied), which cluster at the bottom. The chart shows that students who actually begin or complete have far higher scores than those who drop out. This pattern is very pronounced and the boxes are labeled clearly. No chart elements are misleading here; the trends are obvious from the data.



Box Plots: *Illustrated Engagement Score's influence on Status Description*

Completion Status vs Churn Rate

- The churn bar is enormous relative to the completed bar. Only about 15 completions vs. 340 churns (roughly a 1:23 ratio).
- This starkly illustrates that churn dwarfs completion. Almost all users churn rather than complete. The chart is very clear and not misleading. (If any text suggested a better balance, this refutes it.) The exact values are clearly visible and labeled, so the conclusion is direct: completion is negligible compared to dropouts.



Bar Chart: *Illustrated Completion Status vs Churn Rate*

Learner SignUp DateTime vs. Engagement Score

- **Overall Trend:** There is a **downward trend over time**. Early in 2023, many learners had high engagement scores (300–380 range). As time progresses toward early 2024, the top scores diminish to the 200–250 range. You can see diagonal "bands" of points sloping downward: for example, one band goes from ~250 in Jan 2023 down to ~100 by Dec 2023.

- **Gender Comparison:** Orange and green points are scattered similarly within each band, indicating **no clear difference** between male and female engagement. Both colors mix together at all levels. Unknown and Other genders (grey/pink) appear but in fewer numbers, and their points fall within the same clouds.

- **Clusters:** There is a dense cluster of points below 50 engagement throughout, and two main clusters (150–300 engagement) that decline over time. There's also a small cluster of the very highest scores early on.

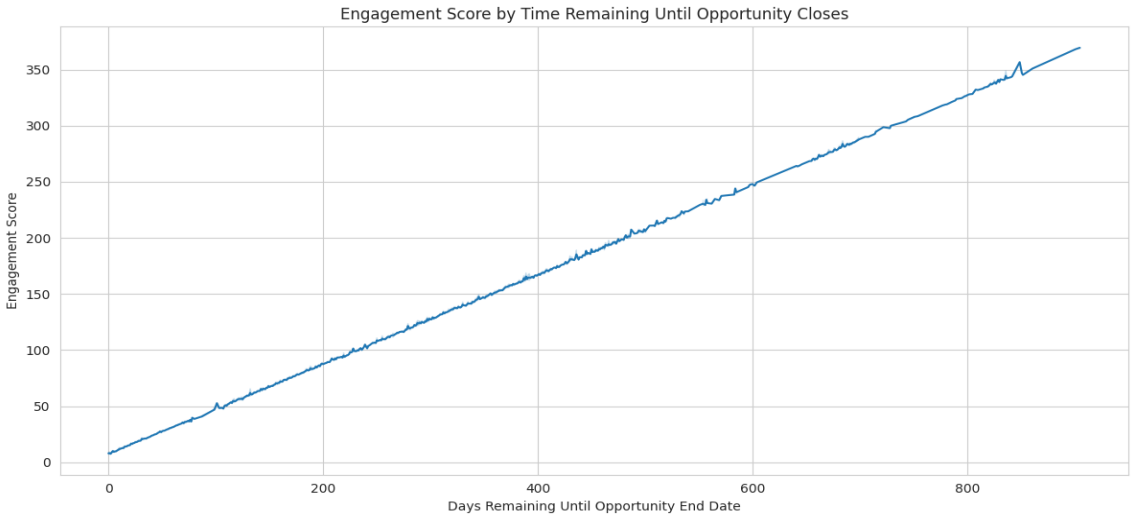
- **Insight:** Learners who signed up earlier (left side) tended to achieve higher engagement scores; later sign-ups have lower scores. This suggests engagement generally **decreases for later cohorts**. Gender doesn't create any visible split. The chart is busy but the time-based trend is clear once one follows the envelope of points. No anomalies stand out beyond the expected scatter noise.

Scatter plot: *Illustrated Sign up Date vs Engagement Score*

Engagement Score by Time Remaining

- The relationship is almost exactly linear. As the days remaining increase, the engagement score increases in lockstep. Essentially, more time until closing correlates with proportionally higher engagement.

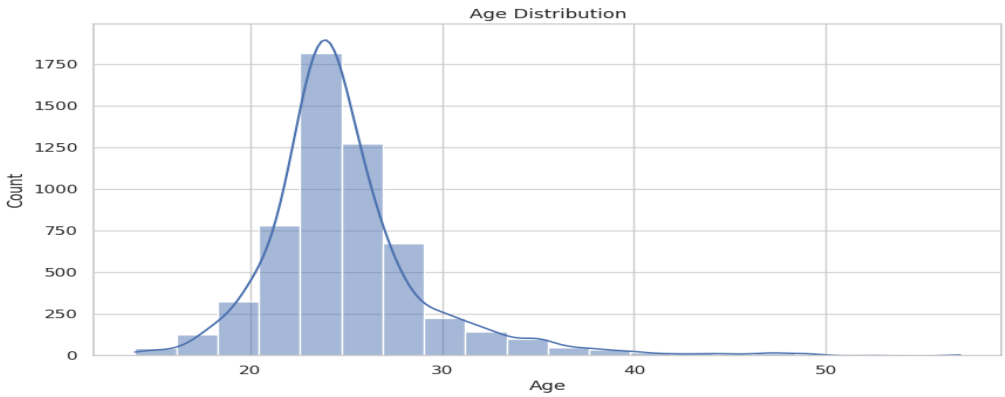
- This suggests a very strong positive relationship: if a learner has more days left, they have a higher engagement score. The plot is suspiciously smooth – real data rarely forms such a perfect line. It may indicate some deterministic or sorted process. Nevertheless, reading it as given: every ~100 additional days yields a ~35–40 point increase in score. If any claims are made about the influence of "time remaining," this chart asserts a nearly perfect linear effect.



Line Plot: *Illustrated Time Remaining vs Engagement Score*

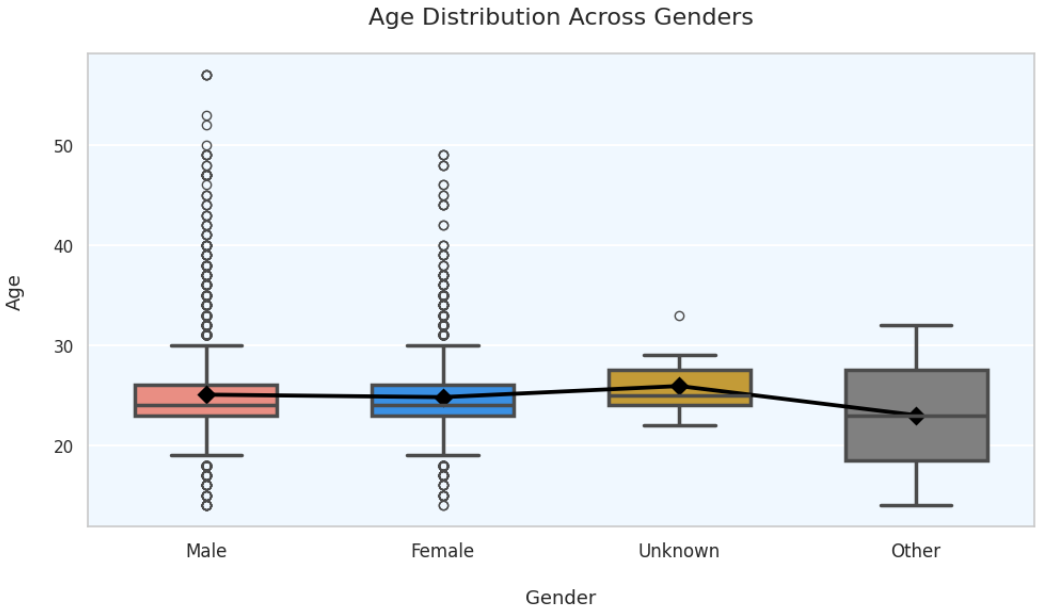
Demographic Distributions:

- **Age Distribution:** Most learners are young (late teens to mid-20s), between 20–24 years old, confirming a primarily university student demographic. The long tail indicates a small number of older learners, but they are rare. The chart is clear and marked, accurately showing that around 70–80% of users are under 30.



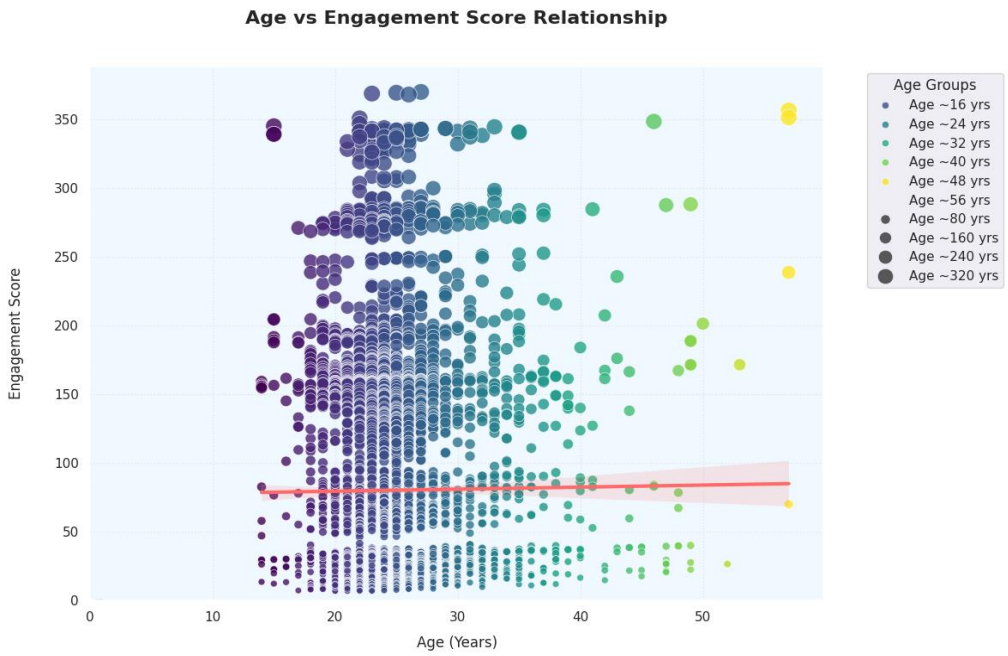
Histogram: *Highlighted Age distribution*

- **Age by Gender:** Across gender categories, the central age is around the mid-20s. There is no drastic difference: males, females, and others all hover in the 20–25 range. The "Unknown" gender group is a bit older on average. All groups show some outliers into the 40–50 range, but those are few. The boxplots clearly show that age distributions largely overlap; there's no special age anomaly by gender.



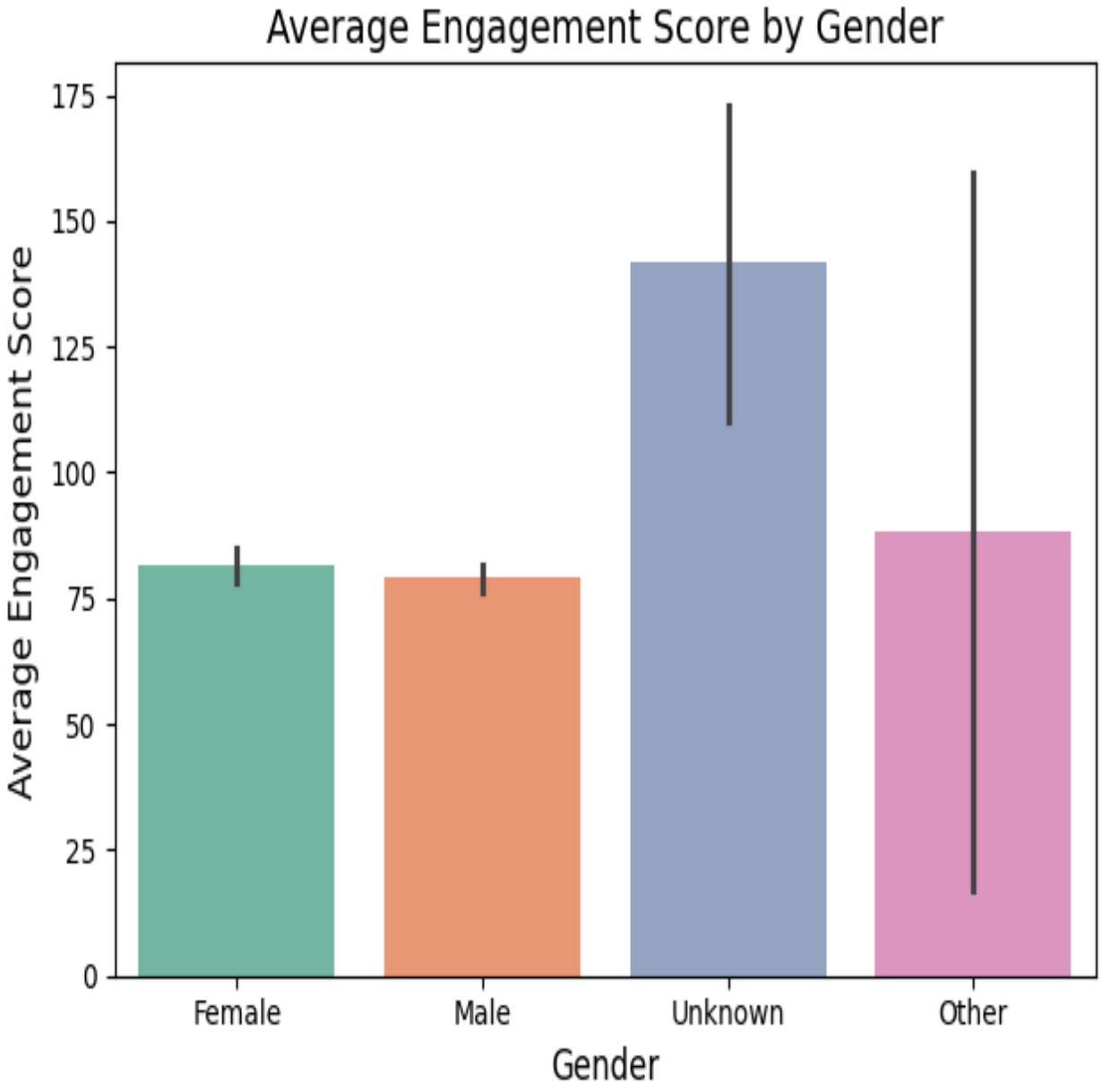
Point plot: *Illustrated Age distribution across Genders*

- **Engagement Score vs Age:** Peak engagement between ages 22–25, dips before 20 and after 26.



Scatter/Reg Plot: *Illustrated age influence on engagement.*

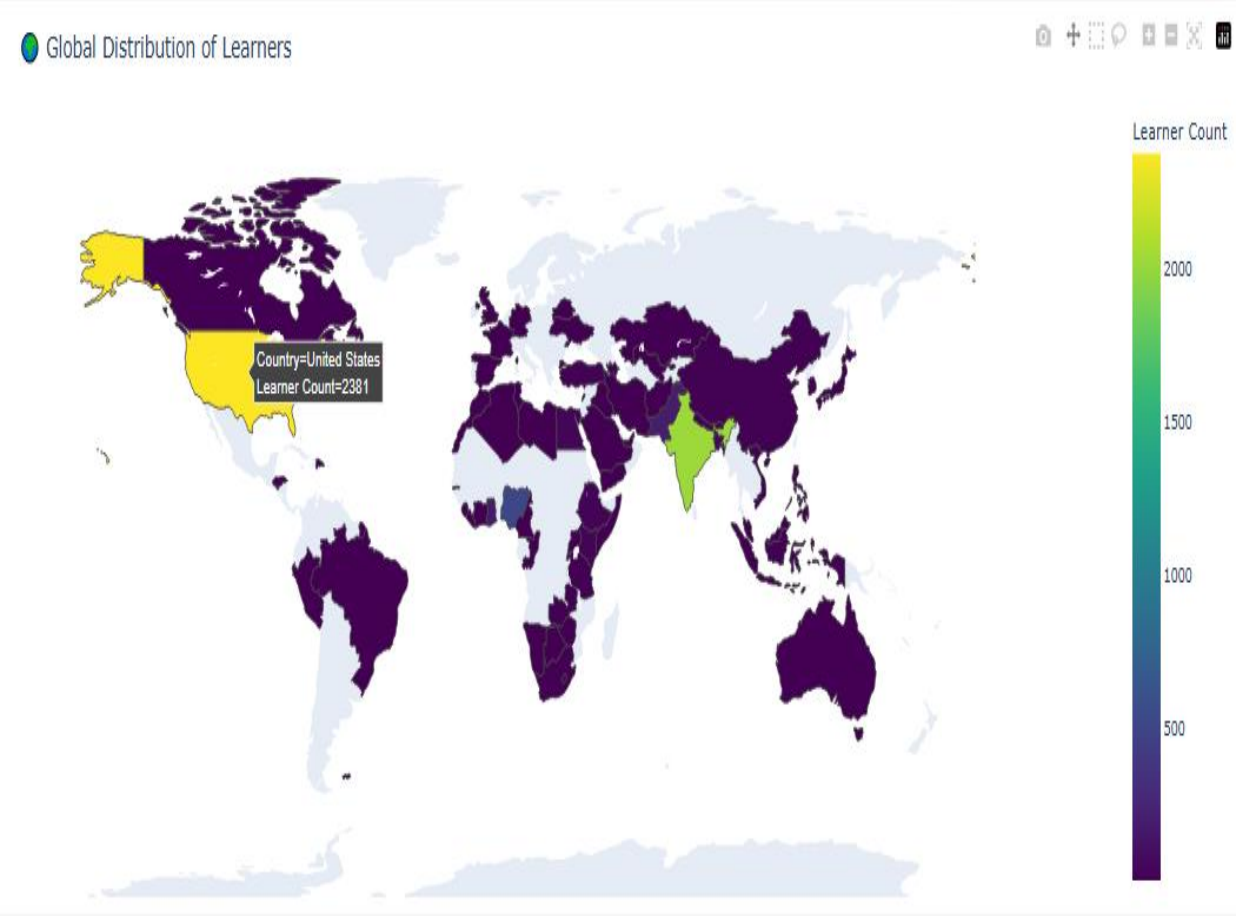
- **Average Engagement Score by Gender:** There is no strong evidence of a gender effect on average engagement: males and females have nearly identical scores. The seemingly higher scores for “Other” and “Unknown” genders come with huge variance bars, so those differences are not definitive. We should be cautious about over-interpreting the Unknown category due to its wide error. No anomalies beyond that; the chart is straightforward.



Scatter/Reg Plot: *Illustrated age influence on engagement.*

- **Global Distribution of Learner - Geo Heatmap:** This world map shades countries by the number of learners. Yellow indicates the highest counts, green moderate, purple low, and white no data. Hover text is shown for the U.S. (2,381).

- **Top Countries:** The **United States** (bright yellow) has the most learners (2,381). **India** (greenish) has the next highest (~1,920).
- **Other Regions:** Much of Europe (UK, Germany, etc.) and South America (Brazil) are shaded dark purple, indicating counts in the few hundreds. African nations, shown (e.g. Nigeria, South Africa, also have a few hundred each. Most Asian countries (China, etc.) appear light (very few or no learners). Australia is purple (~ a few hundred).
- **Insight:** The platform's reach is heavily skewed to the **USA and India**, with the U.S. having the single largest user base. All other countries have significantly fewer learners (typically under 500). The map is clear: only these two countries have more than ~1500 learners. This reveals a major geographic imbalance. No data issues apparent (except that many countries are blank/white, implying zero or missing data).

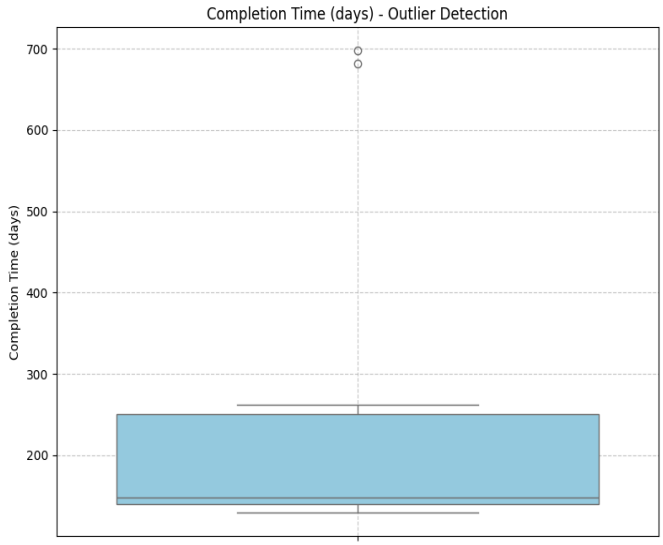


Geo Heatmap: *Illustrated engagement across different countries.*

Outliers and Anomalies

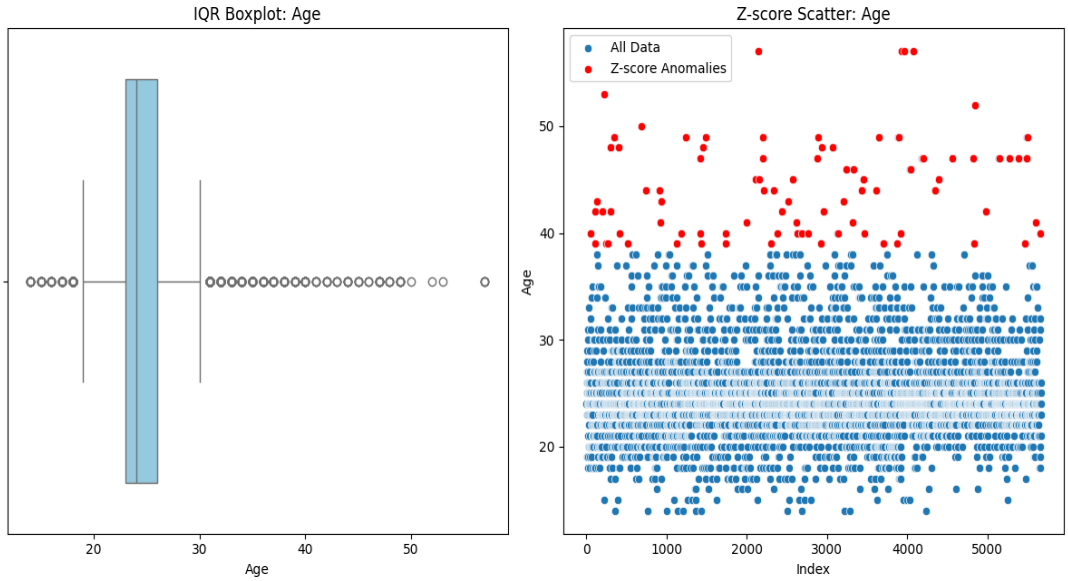
Completion Time Outliers

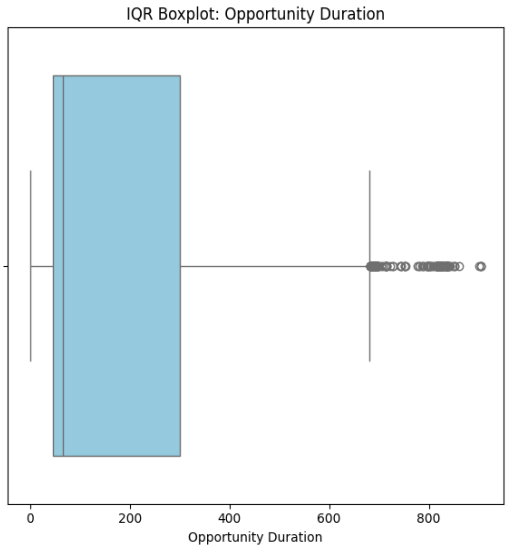
- Identified users taking excessively long. Causes may include: Leaving tabs open, Technical issues, Difficulty understanding tasks
- **Action:** Add session timeout logic or check-ins.



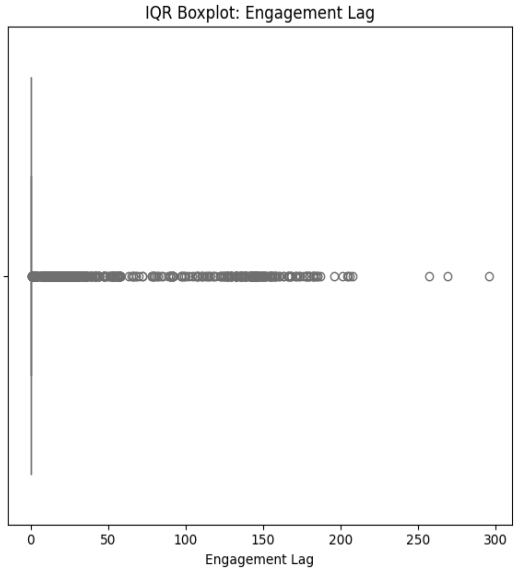
Boxplot: *Explored Completion Days Outliers*

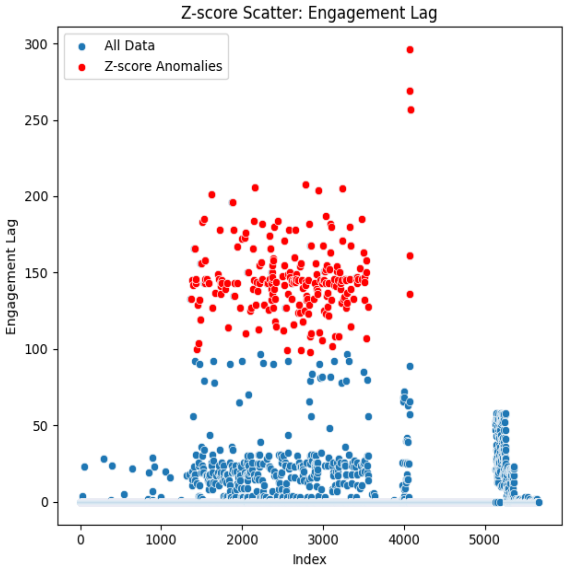
Major columns' Outliers

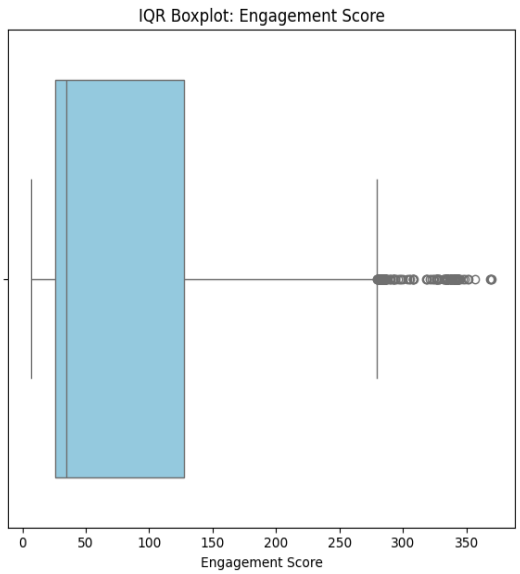



IQR Boxplot: Opportunity Duration
0
200
400
600
800
Opportunity Duration

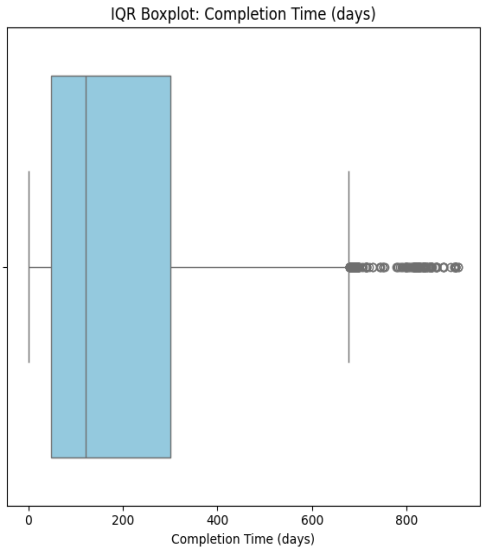

Z-score Scatter: Opportunity Duration
All Data
Z-score Anomalies
Opportunity Duration
0
200
400
600
800
0
1000
2000
3000
4000
5000
Index

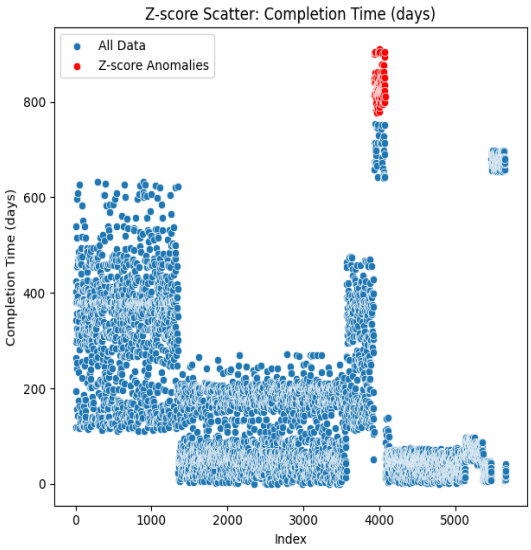

IQR Boxplot: Engagement Lag
0
50
100
150
200
250
300
Engagement Lag

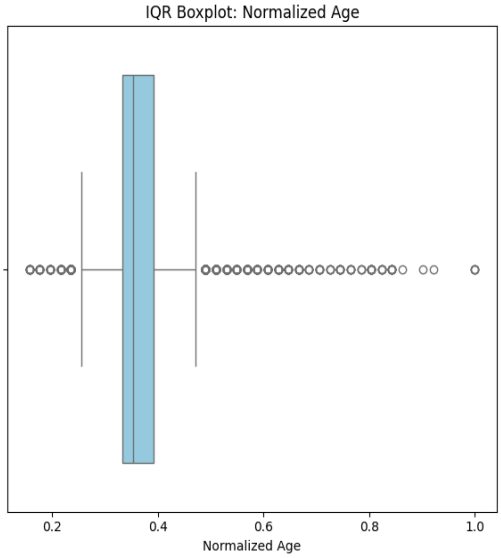

Z-score Scatter: Engagement Lag
All Data
Z-score Anomalies
Engagement Lag
0
50
100
150
200
250
300
0
1000
2000
3000
4000
5000
Index

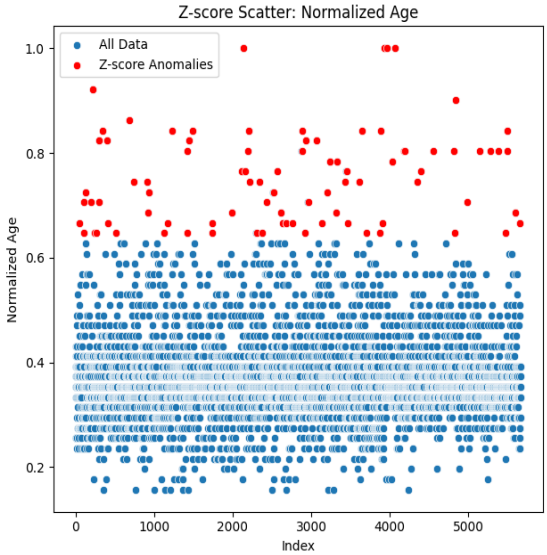

IQR Boxplot: Engagement Score
0
50
100
150
200
250
300
350
Engagement Score

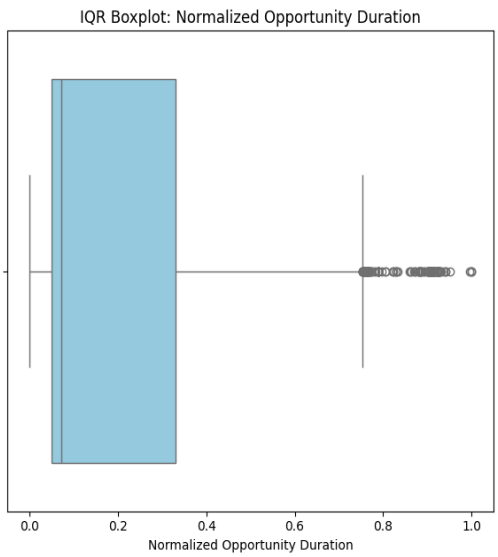

Z-score Scatter: Engagement Score
All Data
Z-score Anomalies
Engagement Score
0
50
100
150
200
250
300
350
0
1000
2000
3000
4000
5000
Index

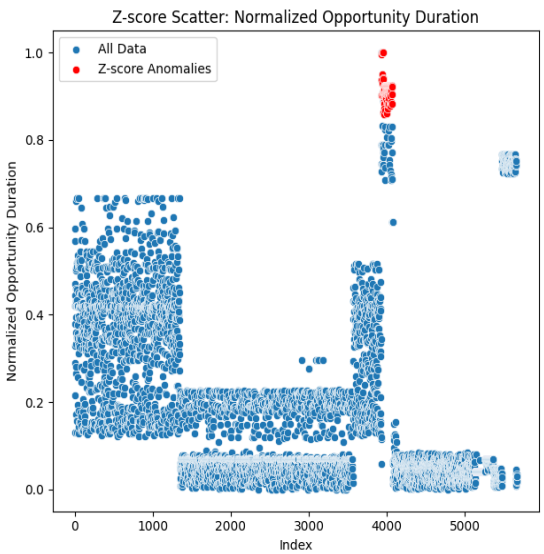

IQR Boxplot: Completion Time (days)
0
200
400
600
800
Completion Time (days)


Z-score Scatter: Completion Time (days)
All Data
Z-score Anomalies
Completion Time (days)
0
200
400
600
800
0
1000
2000
3000
4000
5000
Index


IQR Boxplot: Normalized Age
0.2
0.4
0.6
0.8
1.0
Normalized Age


Z-score Scatter: Normalized Age
All Data
Z-score Anomalies
Normalized Age
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Index


IQR Boxplot: Normalized Opportunity Duration
0.0
0.2
0.4
0.6
0.8
1.0
Normalized Opportunity Duration

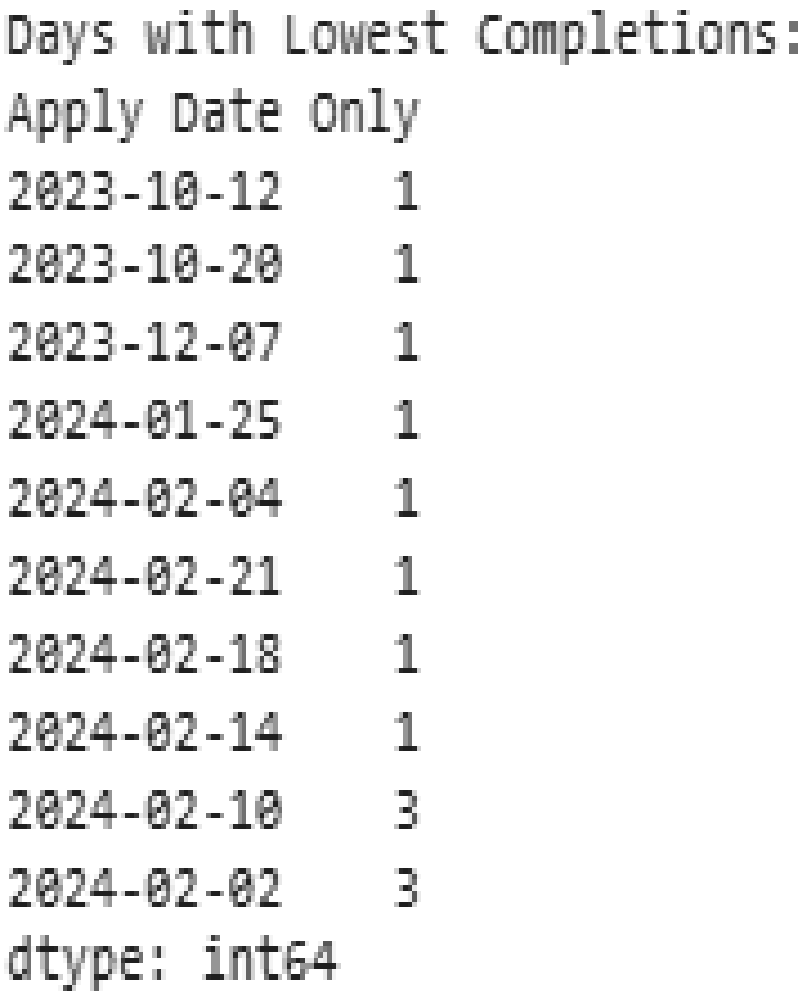

Z-score Scatter: Normalized Opportunity Duration
All Data
Z-score Anomalies
Normalized Opportunity Duration
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000
Index

Low Completion Days Detection

- Several days had **only 1 completion**, indicating possible **operational issues**, **low engagement**, or **external factors** (e.g., holidays, server downtimes).

- May be due to any public holidays, exams, or organizational events, technical issues (e.g., form errors, page downtime) occurred on these days. It can coincide with changes in campaigns, communication strategy, or UI modifications

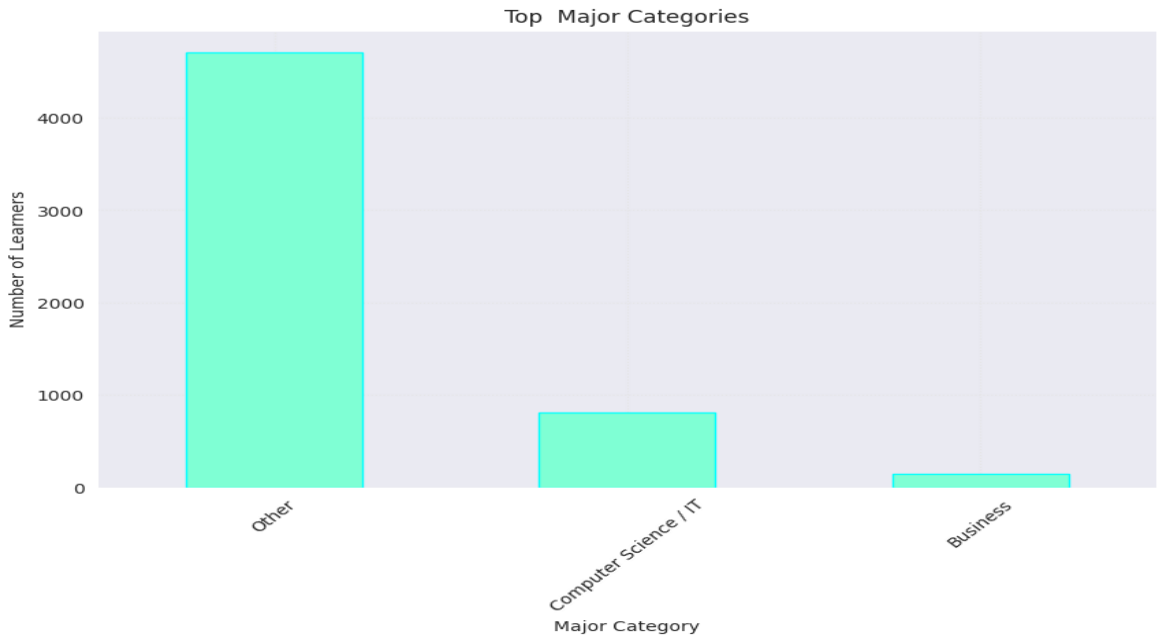
- These dates are spread out, but notably, **mid-February 2024 (Feb 14, 18, 21)** and **early February (Feb 4)** had multiple such low-activity days, suggesting a **pattern of inactivity during that period**. **Feb 2 and Feb 10, 2024**, show **3 completions**—slightly better but still among the lowest. These might represent weekends or specific low-traffic days. Dates like **Oct 12, Oct 20**, and **Dec 7, 2023**, also had minimal completions. This might indicate **seasonal dips**, possibly due to **festive seasons or academic schedules**.



Explored the lowest completion days

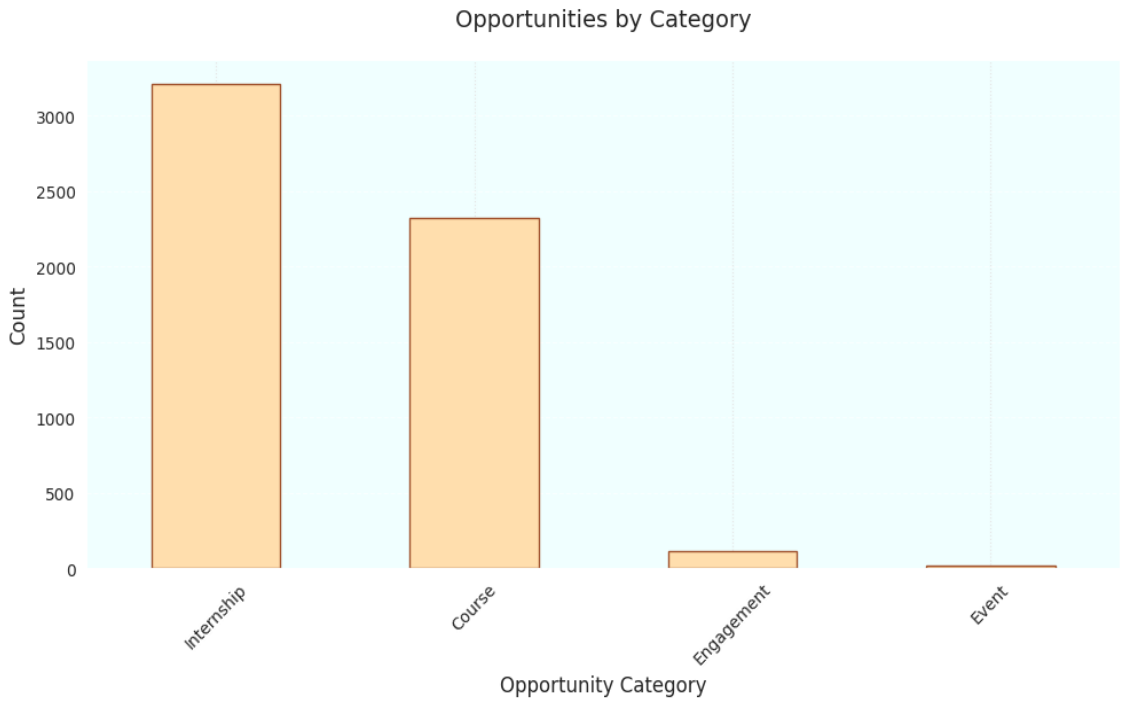
Engagement and Behavioral Analysis

Major Categories: Some majors apply more often for specific opportunity types. **Strategy:** Personalize opportunity suggestions.



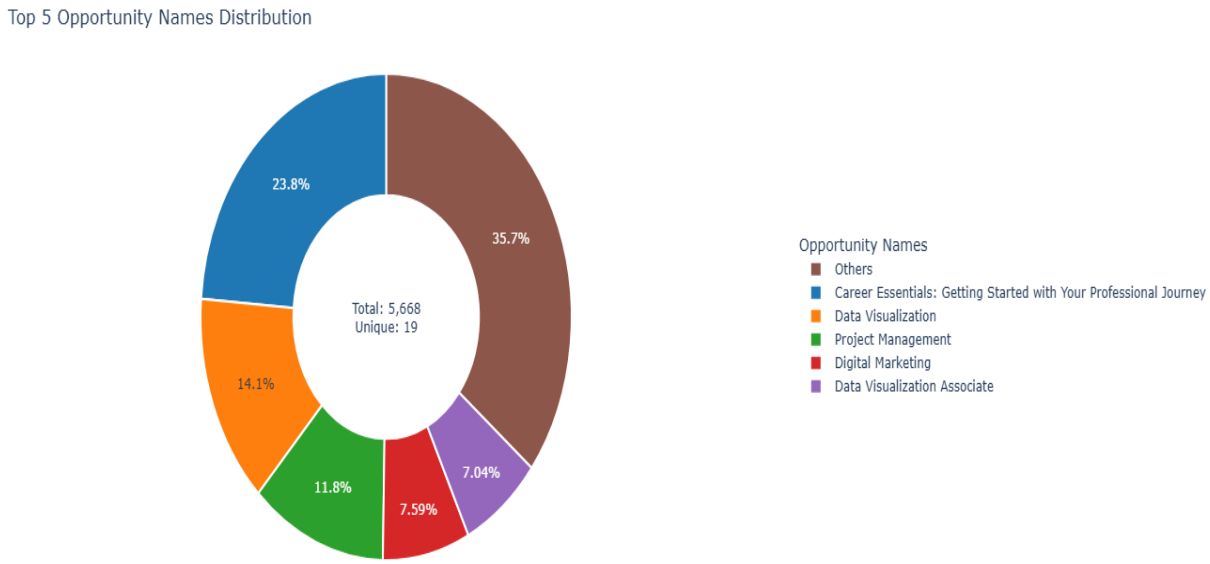
Bar Chart: *Explored major categories*

Opportunity by Categories: Different categories (e.g., internships, courses) show varying engagement.



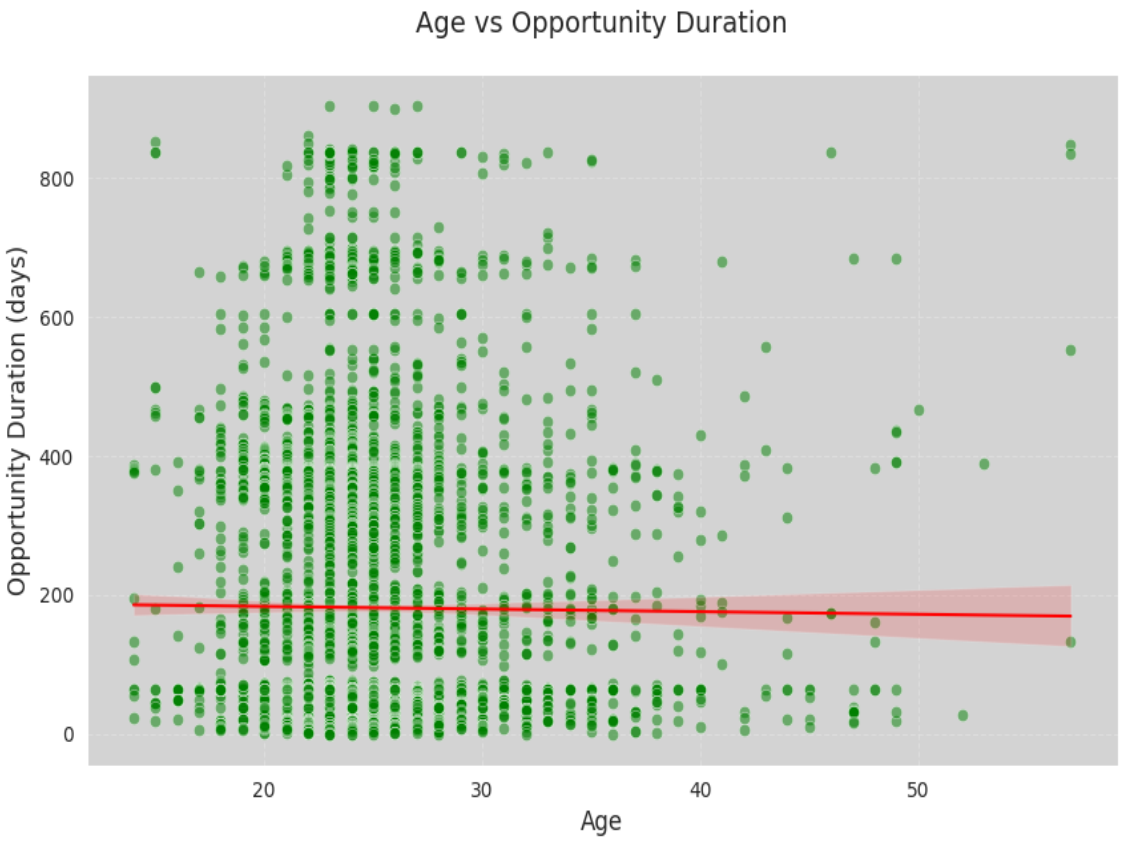
Bar Chart: *Explored opportunity categories*

Top 5 Opportunity Names Distribution



Pie Chart: *Shows opportunity name-wise participation ratios.*

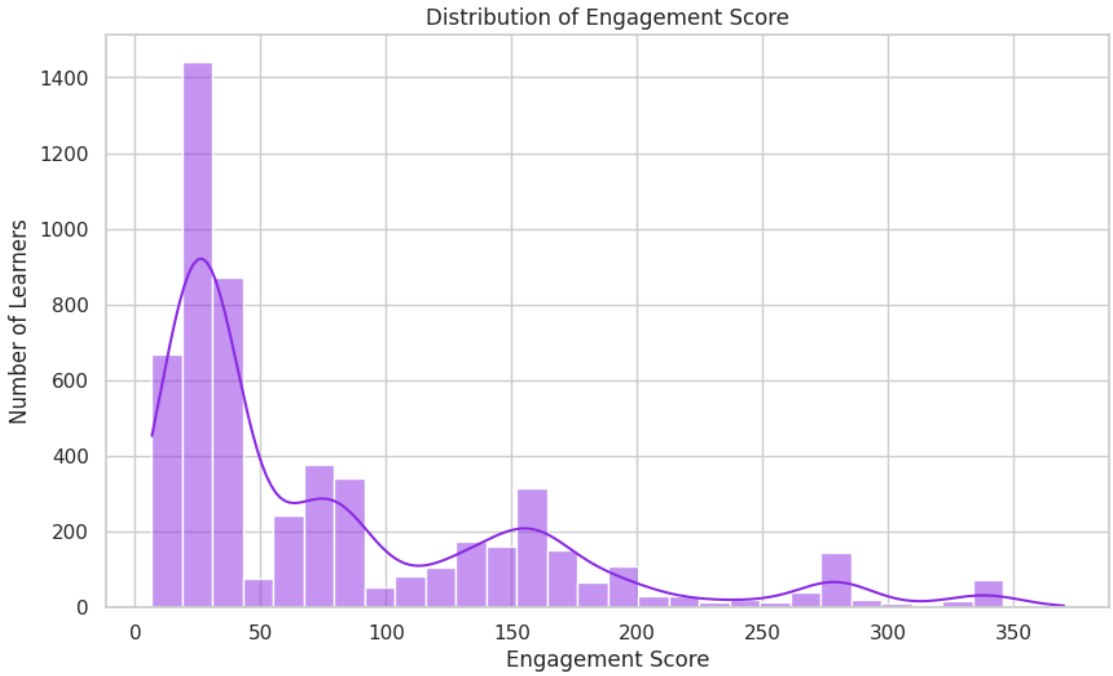
Age vs Opportunity Duration: There is no clear relation between them.



Scatter Plots: *Explored relationships between age and opportunity duration.*

Engagement Score Distribution

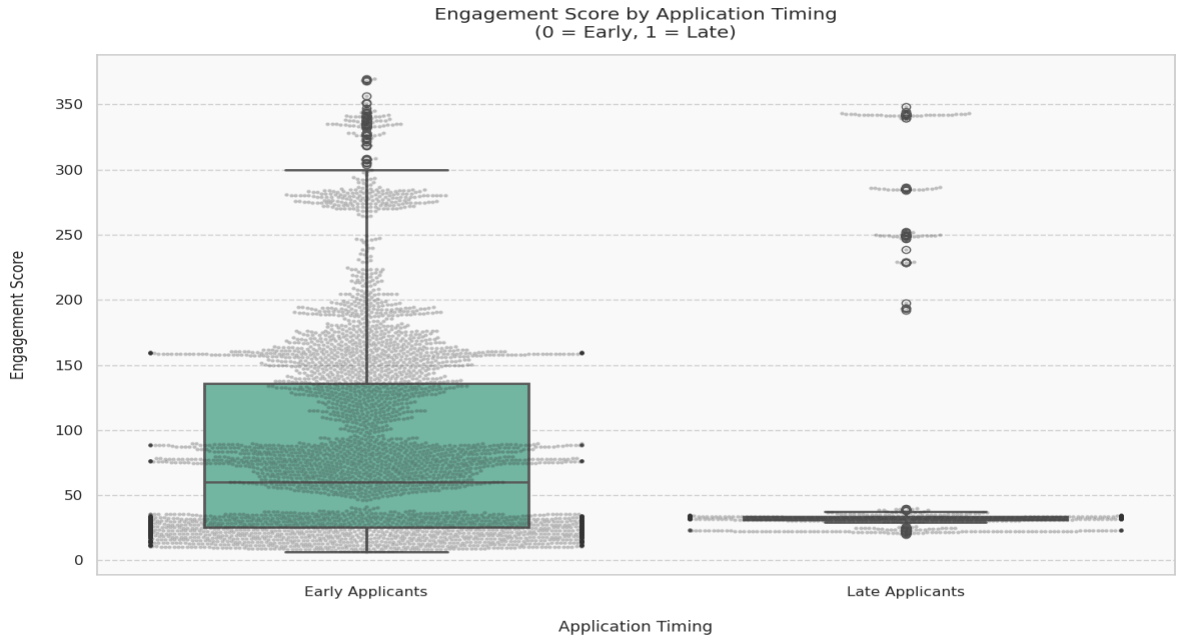
- Students with higher engagement scores were more likely to complete opportunities.



Histograms: *Displayed engagement score distributions*

Engagement Score by Application Timing

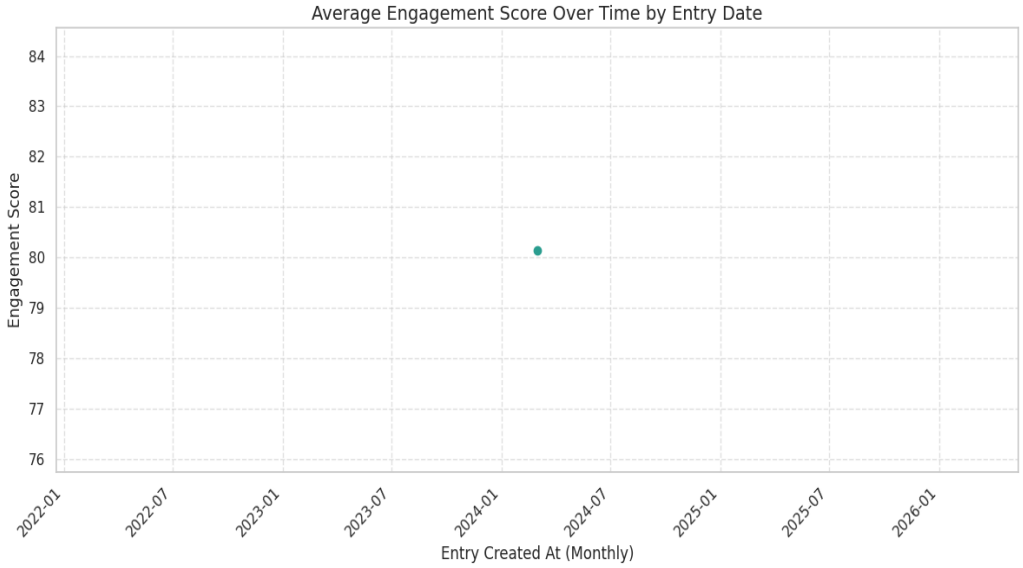
- Late applicants often score higher, which implies increased motivation and goal orientation.



Box+Swarm Plot: *Displayed engagement score distributions by application timing.*

Engagement Score vs. Entry created at

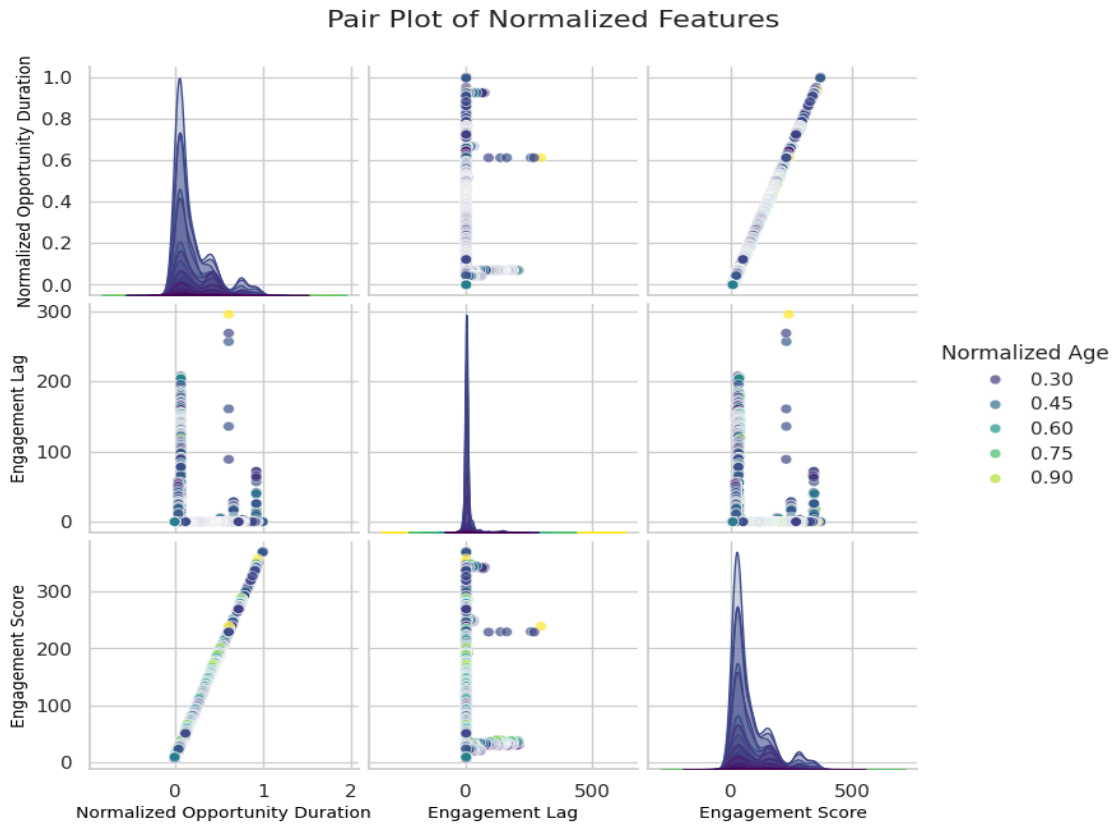
- The chart shows only one point plotted around mid-2024 with an engagement score near 80, which is moderately high.



Line plot: *Displayed engagement score by Entry created at*

Normalized Features' Pairplot: The pair plot examines relationships between Normalized Opportunity Duration, Engagement Lag, Engagement Score, and Normalized Age (as a categorical variable with levels 0.30, 0.45, 0.60, 0.75, 0.90). It includes histograms on the diagonal for distributions and scatter plots off-diagonal for pairwise relationships.

- **Weak or No Correlation**
    - **Engagement Lag and Engagement Score**: The scatter plot reveals a tight cluster near zero with little spread, suggesting minimal correlation. Engagement Lag (likely the time between actions) does not strongly influence Engagement Score.
    - **Normalized Age with Other Variables**: The colored scatter plots (with colors representing age levels) show no distinct patterns, indicating that age does not significantly affect the relationships between opportunity duration, engagement lag, or engagement score.
- **Distribution Insights**
    - **Engagement Lag**: The histogram shows a sharp peak near 0 with a narrow spread, indicating most values are low and consistent.
    - **Engagement Score**: The histogram also peaks near 0 but has a slightly wider spread with some outliers up to 500, showing more variability than Engagement Lag.
    - **Normalized Opportunity Duration**: The histogram is right-skewed with a peak near 0 and a broader spread, suggesting greater variability in how long opportunities last.

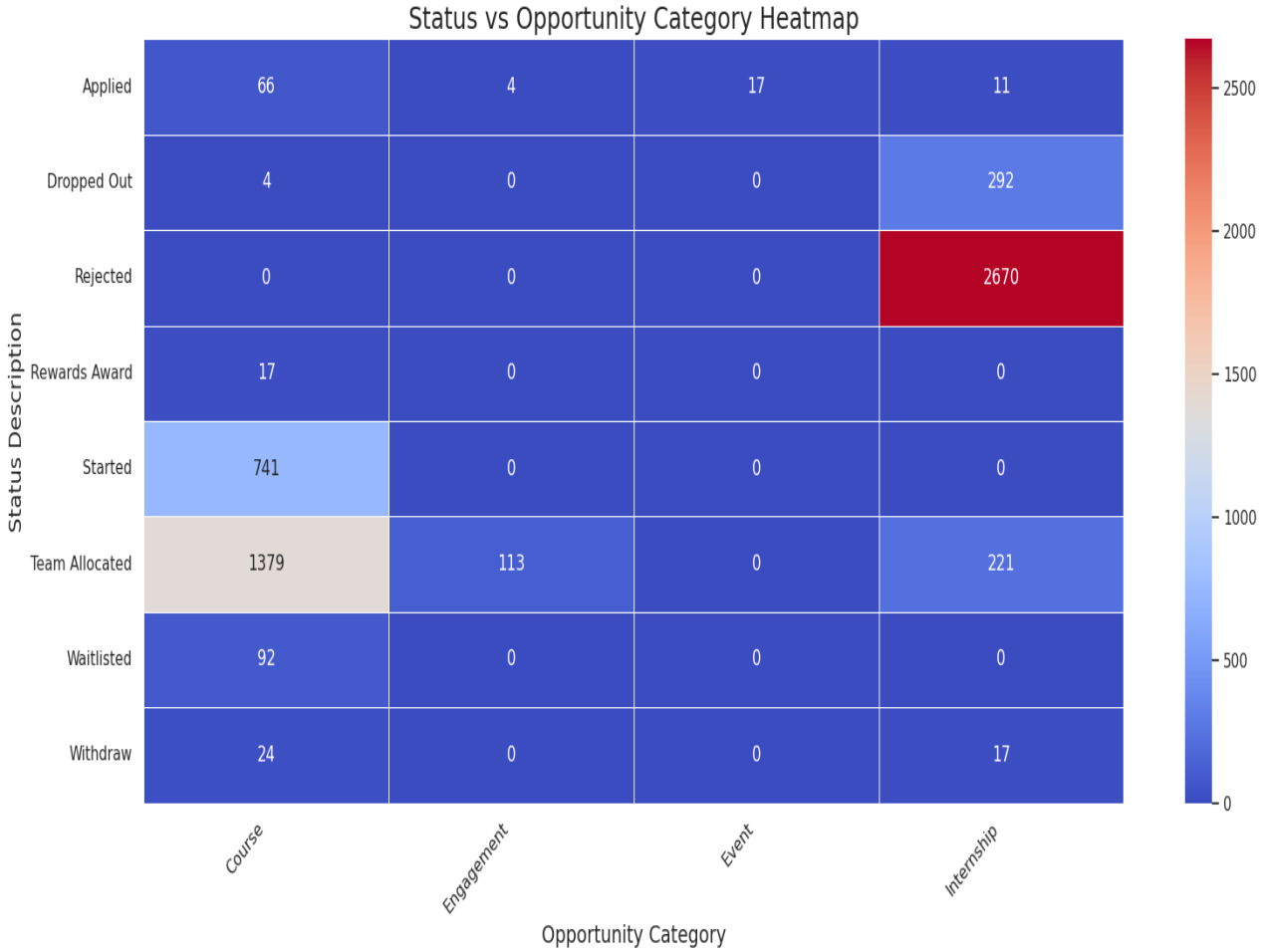


Pair plot: *Explored Normalized Features*

Status vs Opportunity Category: The heatmap displays how different opportunity statuses (e.g., Applied, Dropped Out, Rejected) are distributed across opportunity categories

- **High Rejection in Internships:**
    - **2670** individuals were **Rejected** in the **Internship** category — by far the highest count in the heatmap.
    - This suggests internships are highly competitive or have stringent selection criteria.
- **Course Category Is Most Engaged:**
    - **Team Allocated (1379)** and **Started (741)** statuses are most frequent under the **Course** category, indicating strong participation and progression.
    - Courses are the most advanced in terms of user involvement compared to other categories.
- **Minimal Engagement in Events and Engagements:**
    - Very low or **zero values** across all statuses in the **Event** and **Engagement** categories.
    - Either these categories have low offerings, or users are not engaging with them.

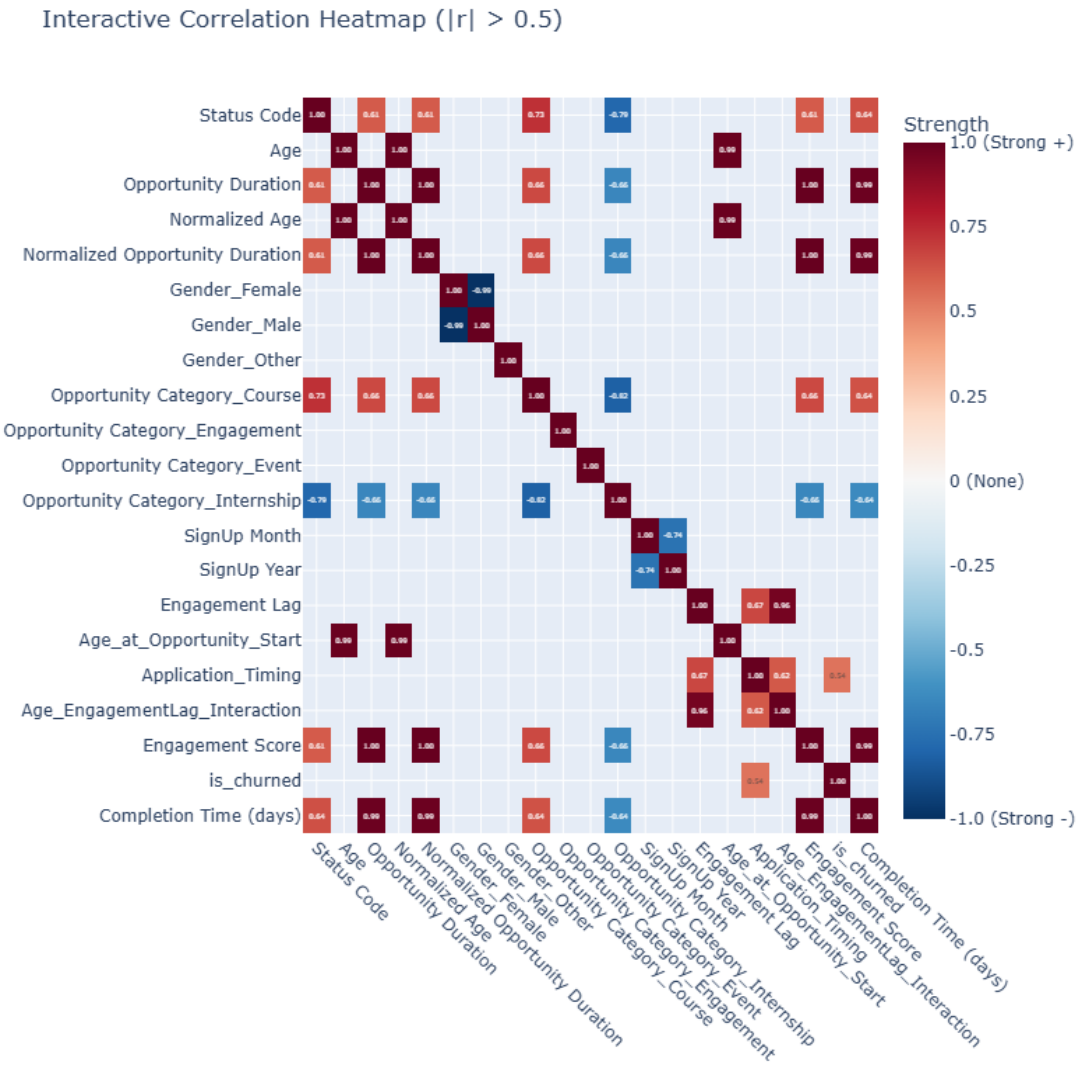
- **Dropouts and Withdrawals Mostly from Internships:**
    - **292 Dropouts** and **17 Withdrawals** were observed in the **Internship** category.
    - Indicates possible issues with retention or satisfaction among interns.
- **Waitlisted and Rewards Only in Courses:**
    - **92 users were Waitlisted,** and **17 received Rewards** in **Courses** — this dynamic does not appear in other categories.
    - Shows additional mechanisms (rewarding and waitlisting) exist in Courses, absent elsewhere.
- **Application Activity Across All:**
    - Users have **Applied** across all categories, but actual participation only follows in **Courses** and **Internships**.



Heatmap: *Status vs Opportunity Category*

Correlation Heatmap (Interactive)

- The correlation heatmap highlights strong relationships between variables (where the absolute correlation coefficient |r| > 0.5). It uses a color gradient from deep blue (strong negative correlation, -1.0) to deep red (strong positive correlation, 1.0), with white indicating no correlation (0).



Heatmap: *Explored Correlation*

Correlation Insights:

- **Strong Positive Correlations**
    - **Normalized Age and Age (r = 1.0)**, and **Normalized Opportunity Duration and Opportunity Duration (r = 1.0)**: These perfect correlations show that the normalized versions are simply scaled transformations of the original variables, maintaining their relationships.
    - **Age_at_Opportunity_Start and Age (r = 0.99)**: This near-perfect correlation suggests that the age at which individuals start opportunities is almost identical to their overall age, implying opportunities are typically initiated at a consistent life stage.

    - **Engagement Score and Completion Time (days) (r = 1.0)**, and **is_churned and Completion Time (days) (r = 1.0)**: These perfect correlations indicate that Engagement Score and churn status are directly tied to how long it takes to complete an opportunity. This could mean Engagement Score is derived from completion time or that longer completion times are strongly linked to higher churn rates.

- **Moderate to Strong Positive Correlations**
    - **Opportunity Category_Course and Opportunity Category_Engagement (r = 0.73)**: This suggests that courses and engagement activities often occur together or are related.
    - **SignUp Year and SignUp Month (r = 0.74)**: This indicates a temporal relationship, possibly reflecting seasonal patterns in sign-ups.
    - **Application_Timing and Age_at_Opportunity_Start (r = 0.67)**: This shows that the timing of applications is moderately linked to the age at which opportunities begin.

- **Strong Negative Correlations**
    - **Gender_Female and Gender_Male (r = -0.99)**, and **Gender_Other and Gender_Male (r = -0.79)**: These strong negative correlations confirm that gender categories are mutually exclusive, as expected.
    - **Opportunity Category_Internship and Opportunity Category_Course (r = -0.79)**: This suggests that internships and courses are rarely pursued together, possibly because they are distinct opportunity types.
    - **SignUp Year and Opportunity Category_Internship (r = -0.74)**: This negative correlation hints at a potential decline in internship opportunities over time.
    - **Engagement Score and is_churned (r = -0.56)**: This moderate negative correlation suggests that higher engagement scores may be associated with lower churn rates, though the relationship is not as strong as others.

- **Weak or No Correlations**
    - Variables like **Status Code** with gender categories or **SignUp Month** with opportunity categories show correlations near 0, indicating no significant linear relationships.

3.4 Predictive Modeling

3.4.1 Target Variable Definition:

Defined the binary target variable DroppedOut based on Status Code:

- Assigned 1 (drop-off) for codes [1030, 1040, 1050, 1110] (Rejected, Waitlisted, Dropped Out, Withdraw).
- Assigned 0 for other statuses (e.g., Completed, Enrolled).

3.4.2 Data Splitting & Preprocessing:

All models used standard preprocessing (feature scaling, encoding categorical features, train/test split).

- Split data into training (80%) and testing (20%) sets with stratification to preserve class distribution.
- Standardized features using StandardScaler.

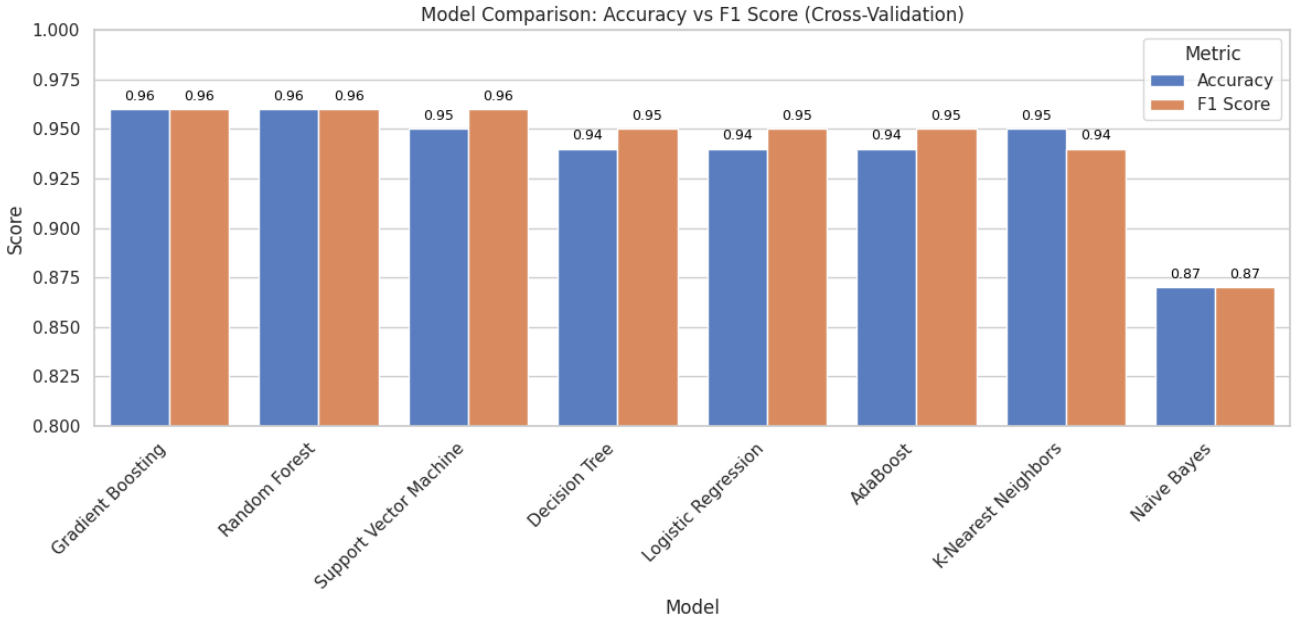
3.4.3 Model Selection

We trained several classification models to predict dropout (churn) using features from the dataset. List the eight classifiers evaluated:

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Ada Boosting
- Support Vector Machine (SVM)
- K-nearest neighbors (KNN)
- Naive Bayes

3.4.4 Model Training & Evaluation

We evaluated multiple classifiers and assessed performance using cross-validation F1 scores, accuracy, classification reports, and confusion matrices, with results visualized via bar plots comparing accuracy and F1 scores, focusing on minimizing false negatives (i.e. missing at-risk students)



Highlighted Model Performance Summary

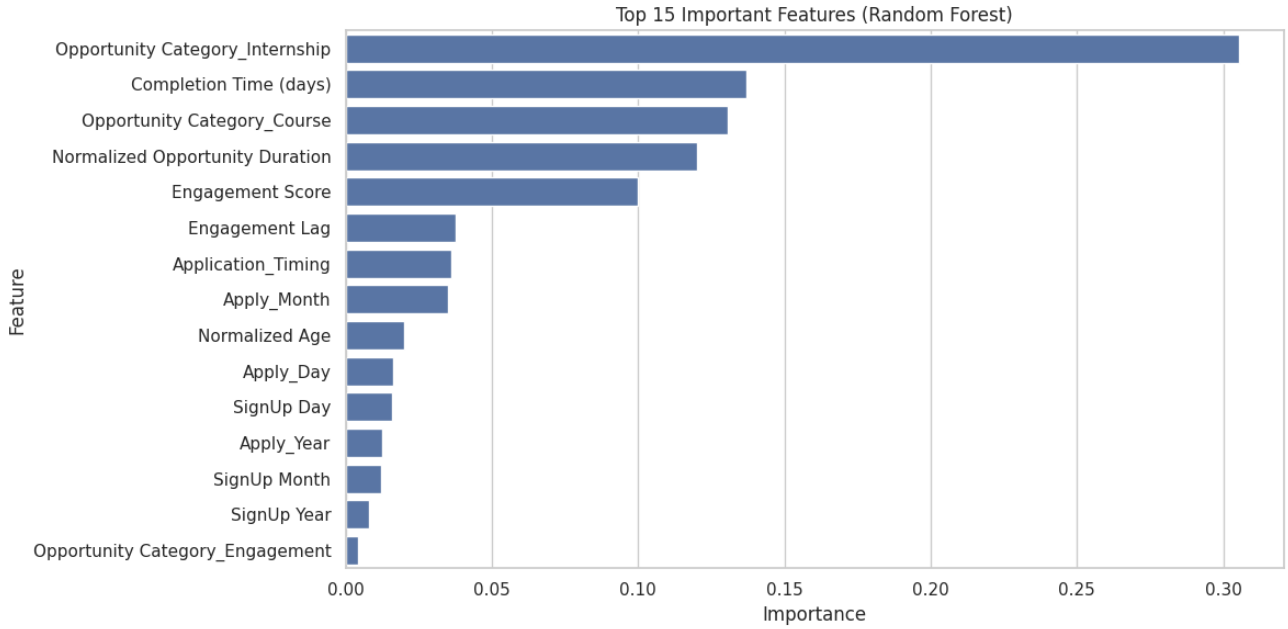
Performance Metrics (from companion analysis):

- **Random Forest Classifier:**
    - Accuracy: 0.96
    - F1 Score: 0.96
    - Balanced confusion matrix with the lowest FN (33) and FP (33)
- **Gradient Boosting:** Close second with slightly more false negatives.
- **SVM:** Good precision, but relatively higher false negatives.

3.4.5 Feature Importance (Top Predictors):

Top Predictive Features in best model:

- Opportunity Category
- Completion Time
- Opportunity Duration
- Engagement Score
- Engagement Lag
- Application Timing
- Age



Highlighted Feature Importance

Results: Random Forest and Gradient Boosting outperformed simpler models in predicting student churn. Random Forest achieved an accuracy of 0.96 and an F1 score of 0.96, "maintaining balanced precision and recall" with 33 false negatives and 33 false positives, significantly surpassing the Decision Tree's performance. Gradient Boosting matched Random Forest's accuracy and F1 score of 0.96 but had slightly higher false negatives (37), indicating marginal differences in error rates. As one report notes, "Random Forest is more accurate at predicting student retention and had a higher AUC" than logistic regression *(texas-air.org)*. Logistic Regression, while more interpretable, was less accurate than both ensemble models. Feature importance analysis revealed that Engagement Score and Application_Timing were the top predictors of churn, with Completion Time and Opportunity Category also contributing significantly to identifying at-risk students.

3.5 Churn Analysis

We defined **churn (dropout) rate** as the percentage of students who did not continue to the next term. Formally, *Churn Rate* = (Number of students lost ÷ Total students at period start)×100. We analyzed student churn to understand the factors driving drop-offs in educational programs, using the DroppedOut variable to identify students who did not complete their opportunities. By examining key features like Engagement Score, Application_Timing, and Completion Time, we uncovered patterns that highlight disengagement and delays as primary risks. The analysis revealed that churn is particularly pronounced early in the program, emphasizing the need for targeted interventions to boost initial engagement and sustain student commitment.

Key Factors: Identified Engagement Score and Application_Timing as the most influential predictors of churn:

- **Low Engagement Scores**: Scores below 40 strongly correlate with higher drop-off rates, indicating disengagement as a critical driver of churn.
- **Long Completion Times**: Completion times exceeding 280 days are associated with increased churn, reflecting delays as a significant barrier.

- **Late Application Timing**: Late applications (Application_Timing = 1) contribute to higher drop-off likelihood due to rushed or unprepared participation.
- **Age Extremes**: Students aged ≤14 or >40 exhibit elevated churn, likely due to unique developmental needs or external commitments.
- **Opportunity Category**: Variations in drop-off rates across categories (e.g., internships vs. courses) suggest program-specific retention challenges.

Impact Analysis:

- **Early Drop-Offs**: The highest churn occurs early in the program, underscoring the critical role of initial engagement in preventing attrition.
- **Low Engagement Impact**: Students with Engagement Scores <40 are significantly more likely to drop out, highlighting the need for interactive interventions.
- **Delayed Completion**: Prolonged completion times (>280 days) increase churn risk, as students lose momentum and disengage.
- **Late Application Effects**: Late applicants (Application_Timing = 1) face challenges in program integration, leading to higher attrition.
- **Age-Based Retention**: Younger (≤14) and older (>40) students require tailored support to address their higher churn tendencies.

4. Insights & Recommendations

4.1 Key Insights

- **Model Performance**:
  - Random Forest and Gradient Boosting achieved the highest accuracy (0.96) and F1 scores (0.96). Random Forest excelled with low false negatives (33) and false positives (33), ensuring robust prediction of student drop-offs. Gradient Boosting followed closely with slightly higher false negatives (37). Naive Bayes underperformed with an accuracy of 0.87 and F1 score of 0.87.
- **Key Features**:
  - Engagement Score and Application_Timing were critical predictors of churn. Low engagement scores (<40) and late applications (Application_Timing = 1) strongly correlated with higher drop-off rates, indicating disengagement or poor planning as key risk factors.
- **Patterns**:
  - Students with prolonged completion times (>280 days) and moderate/low engagement (<60) were more likely to drop out, highlighting delays and lack of interaction as significant risk factors.
- **Early Churn Risk**:
  - High churn rates occur early in the student journey, particularly among those with low engagement scores, emphasizing the need for timely predictive interventions.

4.2 Recommendations

- **For Low Engagement (<40)**:

    - Offer gamified micro-courses for younger students (≤14 years) to make learning interactive and engaging. Provide mentorship or structured study plans for older students (>40 years) to boost participation.
- **For High Engagement (≥80)**:
    - Unlock advanced modules or leadership opportunities to sustain motivation and reward active students.
- **For Prolonged Completion (>280 days)**:
    - Recommend shorter courses or technical support for students with long completion times and low engagement (<60) to re-engage them effectively.
- **Early Intervention**:
    - Use Random Forest predictions to flag at-risk students early (e.g., those with low engagement or late applications) and trigger interventions like personalized onboarding emails or mentor check-ins.
- **Tailored Support by Segment**:
    - Customize interventions based on student profiles (e.g., age, engagement level, opportunity category). For example, offer internship prep workshops for students in internships with low engagement (<50).
- **Data-Driven Curriculum Design**:
    - Replace one-size-fits-all curricula with tailored learning paths informed by predictive models and engagement data, such as interactive modules aligned with students' major categories or interests.
- **Strategic Ensemble**:
    - Deploy Random Forest as the primary predictive model, potentially ensembling with Gradient Boosting to improve predictions for edge cases, enhancing the accuracy of churn identification.

5. Rule-based Recommendation System

The recommendation system developed in this project is a **rule-based framework** designed to provide personalized interventions for students identified as churned or at-risk of dropping out. It leverages a set of predefined rules that consider various student attributes and engagement metrics to suggest tailored actions aimed at improving retention and engagement

This recommendation system is a proactive, data-driven tool designed to enhance student retention and engagement through personalized interventions. Its flexible rule-based logic and configurable thresholds make it adaptable to various educational settings, offering significant potential to improve student outcomes.

5.1 Methodology

The methodology of the recommendation system relies on a series of conditional rules that evaluate key student characteristics against configurable thresholds.It evaluates key attributes—Engagement Score, Completion Time, Application_Timing, Age, Opportunity Category, Status Code, Major_Category, Time Since Signup, and Engagement Duration—against thresholds defined in a configuration dictionary. Implemented through the recommend_interventions function, the system applies conditional logic to assign tailored recommendations, such as gamified courses or re-engagement campaigns, with outputs stored in a new dataset column and exported to

student_recommendations_new_columns.csv. A configuration dictionary defines the thresholds for these attributes, allowing flexibility and easy adjustments. For example, the system can identify low engagement scores or prolonged completion times based on specific values. This flexible, data-driven approach addresses disengagement, delays, and program-specific needs to enhance student outcomes.

5.1.1 System Logic

The core logic is implemented in the function, which operates as follows:

1. **Identify Churned or At-Risk Students**:
    - Detects students with Status Codes indicating churn ([1030, 1040, 1050, 1110] for Rejected, Waitlisted, Dropped Out, Withdraw) or at-risk statuses (e.g., Started, Waitlisted, Applied).
2. **Apply Recommendation Rules**:
    - Low Engagement (Engagement Score <40):
        - Age ≤14: Gamified micro-courses to foster interactive learning.
        - Age >40: Motivational nudges, mentorship, or structured study plans to boost participation.
    - High Engagement (Engagement Score ≥80):
        - Advanced modules or leadership opportunities to sustain motivation.
    - Long Completion Time with Low Engagement (Completion Time >280 days and Engagement Score <60):
        - Shorter courses or technical support to address delays and disengagement.
    - Early Application (Application_Timing = 0):
        - Onboarding emails and deadline notifications to ensure smooth integration.
    - Internship with Low Engagement (Opportunity Category = Internship and Engagement Score <50):
        - Preparatory workshops or dedicated mentors to enhance engagement.
    - Major-Specific Interventions:
        - Engagement Score ≥80: Advanced challenges tailored to Major_Category.
        - Engagement Score <40: Career webinars aligned with Major_Category.
    - Churned Students (Status Code in [1030, 1040, 1050, 1110]):
        - Re-engagement emails with success stories or free micro-courses.
    - Long Time Since Signup (Time Since Signup >365 days):
        - Personalized re-engagement campaigns highlighting new opportunities.
    - Short Engagement Duration (Engagement Duration <7 days):
        - Profile completion prompts or quick-start guides to accelerate onboarding.
3. **Default Recommendation**:
    - If no specific rules are triggered but the student is at-risk, the system suggests maintaining engagement with a weekly content digest.

This IF-THEN approach is transparent and interpretable: each recommendation is traceable to a specific rule

5.2 Implementation

The implementation of the recommendation system involves the following steps:

1. **Data Preprocessing**:
    - The dataset is cleaned and validated to ensure all necessary columns are present and correctly formatted.
    - Rows with missing values in critical columns are dropped to maintain data integrity.
2. **Feature Engineering**:
    - New columns, such as 'Time Since Signup' and 'Engagement Duration', are created to provide additional context for the recommendation logic.
3. **Configuration**:
    - A configuration dictionary is defined with thresholds for various attributes, enabling customization of the rules.
4. **Recommendation Generation**:
    - The recommend_interventions function is applied to each row in the dataset, producing a list of recommendations based on the defined rules.
5. **Output**:
    - The recommendations are added as a new column to the dataframe and saved to a CSV file named 'student_recommendations_new_columns.csv' for further use or analysis.

5.3 Potential Benefits

The rule-based recommendation system enhances student retention by delivering personalized interventions tailored to individual attributes, such as Engagement Score, Completion Time, and Application_Timing. Using a configuration dictionary with thresholds like Engagement Score <40 for low engagement or Application_Timing = 0 for early applications, the system assigns targeted recommendations, such as gamified courses or re-engagement campaigns, to address disengagement and boost participation. The following outlines the system's key personalization features and their impact on student outcomes.

Key Factors:

- **Tailored Rule-Based Interventions**: Assigns recommendations, such as gamified micro-courses for younger students (age ≤14) or advanced modules for high engagement (Engagement Score ≥80), based on student-specific attributes.
- **Configurable Thresholds**: Defines conditions, like Completion Time >280 days for delays or Time Since Signup >365 days for inactivity, enabling precise intervention targeting.
- **Diverse Recommendation Set**: Offers varied interventions, including onboarding emails for early applicants and career webinars for low-engagement students, aligned with Major_Category.

Impact Analysis:

Prior research shows that delivering resources aligned with student interests "makes the educational experience more relatable and enjoyable," boosting motivation and ownership of learning

(aiplusinfo.medium.com). By offering each student tailored suggestions (e.g. practice quizzes, project ideas, support services), we help them stay involved. Personalized recommendations also help educators identify gaps in a student's learning path and intervene with targeted support *(aiplusinfo.medium.com)*. Moreover, rule-based systems are practical where large training data are scarce: they ensure relevant suggestions and can be easily updated by educational experts.

- **Increased Engagement**: Personalized recommendations, like gamified courses for low engagement (<40) or leadership opportunities for high engagement (≥80), enhance student motivation and program involvement.
- **Improved Retention**: Targeted interventions, such as re-engagement emails for churned students or shorter courses for prolonged completion times (>280 days), help at-risk students stay enrolled.
- **Scalability Across Contexts**: The rule-based system adapts to larger datasets or new educational programs by modifying thresholds, ensuring broad applicability.

6. Conclusion (Impact & Future Work)

Impact

Our AI-driven analysis provides actionable insights to improve student retention. By combining **EDA, churn analysis, and predictive modeling**, we identified at-risk students and key drivers of disengagement. The recommendation system translates these insights into concrete actions, fostering a personalized learning environment. Implementing this approach could lead to improved retention rates, higher student satisfaction, and more efficient resource allocation within educational programs.

Data-driven strategies like ours enable administrators to shift from "one-size-fits-all" teaching to **tailored education**, which can dramatically improve performance and retention.

Future Work

Future work includes:

- **Advanced Techniques:** Explore sophisticated recommendation methods like collaborative filtering or deep learning to enhance personalization.
- **Real-Time Integration:** Incorporate real-time data and feedback loops to adapt interventions dynamically to student behavior.
- **Broader Data Sources:** Include additional metrics (e.g., interaction logs, satisfaction surveys) to refine predictions and recommendations.

Ultimately, by integrating these AI insights, institutions can proactively reduce churn, enhance learning outcomes, and adapt to students' needs over time.