Excelerate

DATA VISUALIZATION REMOTE INTERNSHIP

Week-1 Deliverable

INITIAL LEARNING ANALYTICS FOUNDATIONS REPORT

Team-8

Prepared By:

Sumaiya Tasnim
Faizan
Arop Kuol
Shri Vidya
Zainab Fatima
Dewpearl Gonsalves

CONTENTS

1. Introduction

The **Week-1 "Initial Learning Analytics Foundations"** activity focuses on building a foundational understanding of learner behaviour within the Excelerate platform. The primary goal is to explore user interaction data, become familiar with the structure and meaning of key learning metrics, and develop initial insights that will guide deeper analytical work in the coming weeks. Through early EDA, KPI interpretation, and learning path mapping, this week aims to equip you with the core analytical skills needed to evaluate engagement patterns and identify the factors that influence learner success.

Learning analytics is essential in EdTech because it transforms raw learner data into actionable insights that drive better learning outcomes. By analysing engagement time, progression patterns, and assessment performance, educators can identify where learners struggle and adjust content or pathways accordingly. This data-driven approach enables course optimisation, personalised learning experiences, and improved retention. Ultimately, learning analytics helps platforms make informed decisions that enhance both teaching effectiveness and learner success.

Our Week-1 Objectives are:

- ➔ To understand the structure and key attributes of learner interaction data.
- ➔ To explore initial engagement, progression, and performance patterns through EDA.
- ➔ To identify platform KPIs that indicate learner success.
- ➔ To develop a preliminary list of variables for future predictive modeling.
- ➔ To create a sample learning path map to visualize typical learner navigation.

Expected Learning Outcomes:

- We'll gain foundational understanding of how learner behaviour is captured and analysed in EdTech.
- We can be able to identify key engagement and performance metrics that influence learning success.
- We will be able to develop skills in performing basic exploratory data analysis.
- We can interpret early behavioural patterns to inform future modelling tasks.
- We can create a clear visual representation of learner progression through a learning path.

Together, we, the team, begin this internship's week-1 by grounding our work in a clear understanding of learner behaviour and core analytics principles. This foundational insight prepares us for deeper modelling and visualization in the coming weeks, ensuring that our future analyses are informed, meaningful, and aligned with how learners truly interact and succeed on the Excelerate platform.

2. Dataset Exploration

Assigned Dataset: “SLU Opportunity Wise Data.csv”

Dataset Overview

The dataset represents learner applications and interactions with various opportunities. It includes demographic details, educational background, and opportunity-related information. This dataset contains **8,558 records** (rows) and **16 columns**, capturing key attributes about learners and the opportunities they applied for.

Column Name	Meaning / Description	Datatype
Learner SignUp DateTime	Date and time when the learner registered in the system	Date/Time
Opportunity Id	Unique identifier for each opportunity (Primary key)	Categorical
Opportunity Name	Name of the opportunity applied for	Categorical
Opportunity Category	Type or category of the opportunity	Categorical
Opportunity End Date	Scheduled end date of the opportunity	Date/Time
First Name	Learner’s first name	Categorical
Date of Birth	Learner’s date of birth	Date/Time
Gender	Learner’s gender	Categorical
Country	Learner’s country of residence	Categorical
Institution Name	Name of the learner’s institution	Categorical
Current/Intended Major	Learner’s field of study or intended major	Categorical

Entry created at	Date and time when the entry was recorded in the system	Date/Time
Status Description	Description of the learner's application status	Categorical
Status Code	Numeric code representing the application status	Numerical
Apply Date	Date when the learner applied for the opportunity	Date/Time
Opportunity Start Date	Scheduled start date of the opportunity	Date/Time

Key Learner Attributes and Engagement Indicators

Key learner attributes, such as Date of Birth, Gender, Country, Institution Name, and Current/Intended Major, provide essential context for understanding patterns in engagement and performance. Engagement indicators, including Status Description, Status Code, Apply Date, Opportunity Start Date, Opportunity End Date, and Opportunity Id, help track how actively learners interact with the platform. Analysing these attributes and metrics together allows us to identify trends, predict outcomes, and highlight areas where learners may need additional support.

Data Cleaning Process

The dataset was cleaned using Python programming language to ensure accuracy, consistency, and reliability for further analysis. Various data cleaning tasks were performed, including handling missing values, removing duplicates, standardizing string formats, correcting typos, mapping ambiguous entries, and verifying data types. Each step was carefully executed to maintain data integrity and prepare the dataset for meaningful analysis.

Step No	Data Cleaning Task	Columns Affected	Method/Action Taken	Purpose/Reason
1	Converted date columns to datetime	Learner SignUp DateTime, Opportunity End Date, Date of Birth, Entry created at, Apply Date, Opportunity Start Date	pd.to_datetime() applied to all date columns	Ensures proper date format for calculations and analysis
2	Converted categorical columns	Opportunity Category, Gender, Country, Institution	Converted to category datatype	Optimizes memory usage, improves grouping,

		Name, Current/Intended Major, Status Description		aggregation, and dashboard performance
3	Standardized string case	Institution Name, Current/Intended Major	str.title() or custom mapping applied	Removes inconsistencies in capitalization, ensures uniformity
4	Handled missing values	All columns	Checked .isnull().sum(), filled or verified no nulls. Filled null values in date datatype using mode,mean,median etc & in categorical datatype with “Unknown”, “Other” value.	Ensures dataset completeness without removing rows unnecessarily
5	Removed duplicates	Checked Entire dataset	df.drop_duplicates()	Prevents duplicate records from skewing analysis
6	Corrected typos + invalid entries	Current/Intended Major, Institution Name	Fixed known typos like Computer And Infromation Sciences → Computer And Information Sciences, Could Computing → Cloud Computing - etc. & Many more	Standardizes data for accurate grouping and analysis
7	Standardized degree names	Current/Intended Major	Merged similar entries: e.g., MS / MSc, Msc Computer Science → MS/MSc	Reduces variation in entries for better aggregation & Important for analysis consistency
8	Mapped unclear single-word entries	Current/Intended Major	Invalid entries like Baf, Bsit, Ot, Te, Mmg, Bca, Pos Service → Other	Handles ambiguous or unclear entries to avoid misclassification
9	Standardized institution names	Institution Name	Mapped abbreviations and variations like short form to full proper form/ names. Replaced invalid values with “Not Applicable”	Ensures consistency and easier grouping/analysis

10	Verified data types	All columns	Checked df.dtypes, converted columns to appropriate types	Ensures correct data types for computations and analysis

Final Validation/Checking of Dataset

Final Checking of dataset

```
df.info()
```

[223] ✓ 0.0s Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8246 entries, 0 to 8245
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Learner SignUp DateTime  8246 non-null   datetime64[ns]
 1   Opportunity Id           8246 non-null   object
 2   Opportunity Name         8246 non-null   object
 3   Opportunity Category     8246 non-null   category
 4   Opportunity End Date     8246 non-null   datetime64[ns]
 5   First Name               8246 non-null   object
 6   Date of Birth            8246 non-null   datetime64[ns]
 7   Gender                   8246 non-null   category
 8   Country                  8246 non-null   category
 9   Institution Name         8246 non-null   category
 10  Current/Intended Major   8246 non-null   category
 11  Entry created at         8246 non-null   datetime64[ns]
 12  Status Description       8246 non-null   category
 13  Status Code              8246 non-null   int64
 14  Apply Date               8246 non-null   datetime64[ns]
 15  Opportunity Start Date   8246 non-null   datetime64[ns]
dtypes: category(6), datetime64[ns](6), int64(1), object(3)
memory usage: 800.6+ KB
```

```
df.isnull().sum()
```

[224] ✓ 0.0s

```
Learner SignUp DateTime    0
Opportunity Id             0
Opportunity Name           0
Opportunity Category       0
Opportunity End Date       0
First Name                 0
Date of Birth              0
Gender                     0
Country                    0
Institution Name           0
Current/Intended Major     0
Entry created at           0
Status Description         0
Status Code                0
Apply Date                 0
Opportunity Start Date     0
dtype: int64
```

```
df.duplicated().sum()
```

[258] ✓ 0.0s Python

```
np.int64(0)
```

Exported Dataset After Cleaning: “Cleaned_SLU_Opportunity_Wise_Dataset.csv”

Overall ,

- ➢ Handled null/missing values in all columns.
- ➢ Removed duplicate rows to avoid redundancy.
- ➢ Standardized string formats (e.g., institution names, majors).
- ➢ Corrected typos and spelling inconsistencies.
- ➢ Mapped ambiguous or unclear entries (e.g., single-word majors) to a standard category.
- ➢ Verified and ensured correct data types for each column.
- ➢ Ensured consistency across categorical values.
- ➢ Prepared the dataset for accurate analysis and reporting.

Finally, the dataset has been fully cleaned, standardized, and is now ready for analysis.

3. Key Platform KPIs & Success Metrics

Opportunity Completion & Status Rates

This metric tracks the funnel from application to completion. The data indicates a highly competitive or selective process, with a large portion of learners currently in the "Team Allocated" phase.

- **Completion Rate (Rewards Awarded):** 0.35%
  - Note: This represents learners who have reached the final "Rewards Award" status.
- **Active / In-Progress Rate:** 48.5%
  - Breakdown: Team Allocated (38.4%) + Started (8.8%) + Applied (1.2%)
- **Rejection Rate:** 41.8%
  - A significant portion of applicants are rejected, suggesting stringent selection criteria or a mismatch in applicant qualifications.
- **Drop-off / Withdrawal Rate:** 8.2%
  - Breakdown: Dropped Out (7.2%) + Withdraw (1.0%)

Average Time to Apply & Duration

These metrics highlight process efficiency and learner engagement speed.

- **Average Time to Apply:** 56.9 days
- **Median Time to Apply:** 4.1 days
    - Insight: The large gap between the average and median indicates a skewed distribution. Most learners (median) apply very quickly—within 4 days of signing up—showing high initial intent. The average is pulled up by a smaller group of "late bloomers" who apply months later.

- **Average Opportunity Duration:** 422.4 days
    - Insight: This likely represents the total window the opportunity is open (Start Date to End Date) rather than the time a learner spends on it.

Learner Demographics & Distribution

- This metric represents the percentage of total applications originating from specific countries. It measures the platform's market penetration and geographic reach

- Context for the 46% (United States): The data shows that nearly half of all platform activity is driven by learners in the US. This defines the US as the Primary Market, while India (34%) represents the Secondary Key Market. Together, these two regions account for 80% of all engagement

- Information Systems (25%) vs. Computer Science (12%): The high prevalence of "Information Systems" suggests that the platform attracts learners interested in the intersection of business and technology (e.g., Business Analysts, Product Managers) more than pure software engineering roles (Computer Science).

- Differentiation: Note that "Information Systems" (25%) and "Computer Information Systems" (3%) are listed separately. These are distinct but related fields; combined, they indicate a very strong trend toward IT management and data systems.

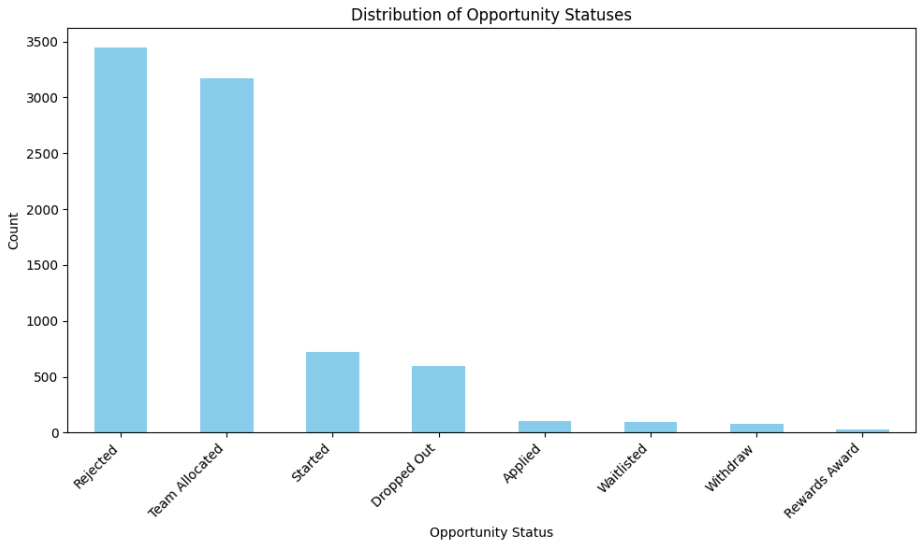
Country	Percentage
United States	46%
India	34%
Nigeria	9%
Ghana	3%
Pakistan	3

Major	Percentage
Information Systems	25%
Computer Science	12%
Health Data Science	4%
Cyber Security	4%
Computer Information Systems	3%

Engagement Patterns

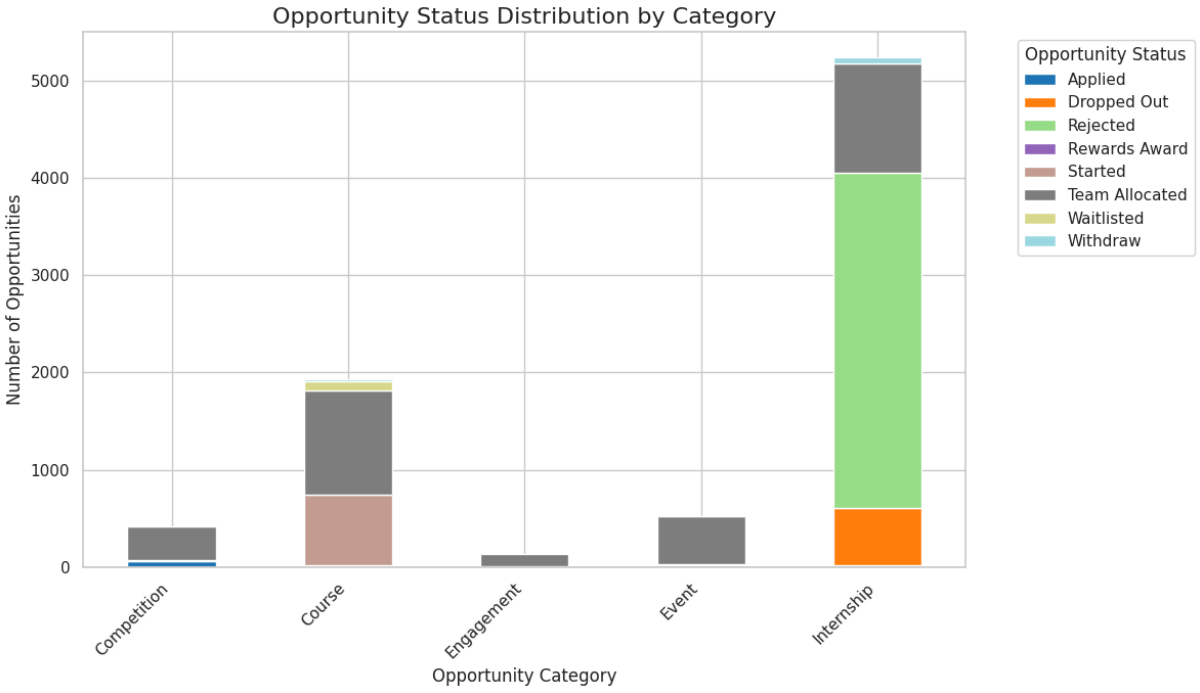
- Application Behavior: The histogram in the dashboard below shows that the vast majority of applications occur immediately after sign-up (the tall bar at the 0 mark), with a "long tail" of users applying much later.
- Seasonality: The trend line (bottom-right in the dashboard) tracks the volume of applications over time, allowing you to identify peak months for learner activity.

Status Analysis

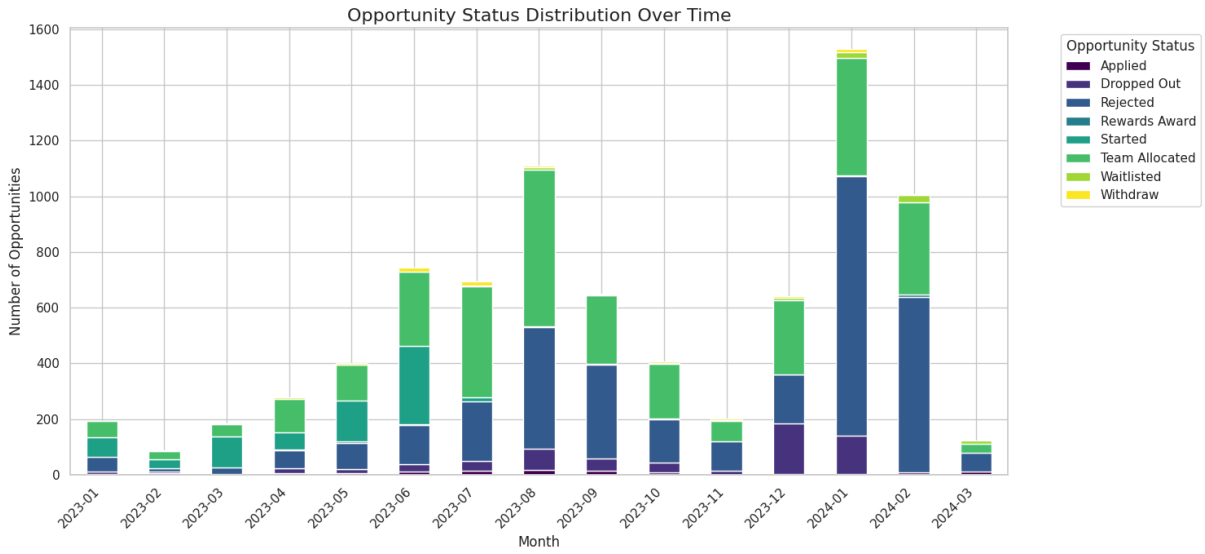


- "Rejected" is the most frequent opportunity status, with 3,447 occurrences, representing a significant portion of all opportunities.
- "Team Allocated" is the second most frequent status, with 3,169 occurrences, indicating a substantial number of opportunities to progress to this stage.

- "Started" opportunities number 724, followed by "Dropped Out" with 596, suggesting a high attrition rate after opportunities commence.
- Statuses like "Applied" (103), "Waitlisted" (96), and "Withdraw" (82) are less frequent, indicating fewer opportunities are stuck in the initial application or waiting phases.
- "Rewards Award" is the least frequent status, with only 29 occurrences, highlighting a very low completion rate for opportunities leading to a reward



- Internship opportunities show the highest counts of 'Rejected' and 'Dropped Out' statuses. This suggests that internships might be highly competitive or have stringent criteria leading to many rejections or withdrawals.
- Course opportunities have a significant number in 'Team Allocated' and 'Started' statuses, indicating higher engagement or progression rates within this category once learners begin.
- Competition and Engagement categories have a high proportion of 'Team Allocated' statuses, with fewer 'Rejected' or 'Dropped Out' instances compared to Internships. This implies a potentially smoother process or higher retention in these categories.
- 'Rewards Award' is primarily concentrated in 'Course' and 'Event' categories, suggesting these are the main avenues where successful completion or recognition is achieved.



- Opportunities show a steady upward trend throughout 2023, with strong growth from May to August and a sharp peak in January 2024, largely driven by high volumes in the Team Allocated and Started statuses, alongside a notable increase in Rejected opportunities toward late 2023.
- January 2024 represents the highest activity level across almost all opportunity statuses, indicating a major surge in intake or operational activity, after which there is a clear and rapid decline in February and March 2024, likely influenced by seasonality, reduced inflow, or process changes.
- Overall, while minor statuses such as Applied, Waitlisted, Withdrawn, Dropped Out, and Rewards Award remain consistently low and stable, the aggressive scaling in mid-to-late 2023 followed by an early 2024 cooldown highlights the importance of investigating the drivers behind the January spike and the subsequent drop.

4. Exploratory Data Analysis (EDA)

Introduction to the Dataset:

- **Source of the dataset:** Excelerate
- **Purpose of the dataset:** This dataset contains detailed records related to learners and the opportunities they engaged with, including opportunity details, learner demographics, key dates, and application status information. The data is structured at the learner opportunity level, meaning each record represents a learner's interaction with a specific opportunity.

The dataset is mainly used to understand participation behavior, timelines, and outcomes across different opportunities and learner backgrounds.

Dataset Overview:

Description	Messed up Dataset	Clean Dataset
Dataset Name	SLU Opportunity Wise Data	Clean SLU Opportunity Wise Data
Number of Rows	8,558	8,246
Number of Columns	16	16

Explanation:

- The cleaned dataset has fewer rows because duplicate, missing, or incorrect records were removed.
- Both datasets contain the same columns, but the clean dataset is easier to analyze.

1.3 Data Structure Overview:

The table below explains each column in simple words, along with its data type and quality.

Column Name	Data Type	Data Quality
Learner Sign-Up Date and Time	Date & Time	Mostly complete and consistent
Opportunity ID	Text	Clean and consistent, used as an identifier
Opportunity Name	Text	Well-formatted, repeated for many learners
Opportunity Category	Text	Limited and consistent categories
Opportunity End Date	Date	Few missing values, mostly correct
First Name	Text	Clean text, not unique
Date of Birth	Date	Some invalid values removed during cleaning
Gender	Text	Standardized values (e.g., Male/Female)
Country	Text	Spelling issues corrected
Institution Name	Text	Mostly consistent, repeated values

Column Name	Data Type	Data Quality
Current/Intended Major	Text	Some missing values present
Entry created at	Date & Time	Correct format after cleaning
Status Description	Text	Clean and easy to understand
Status Code	Text	Standardized codes
Apply Date	Date & Time	Some missing or unusual values
Opportunity Start Date	Date	Mostly complete

Data Preprocessing Methodology:

Before starting the analysis, the dataset went through a structured data cleaning process to improve accuracy, consistency, and usability. The goal of cleaning was to remove errors, handle missing values, and standardize formats so that meaningful analysis could be performed.

The following steps were taken during data cleaning:

- Duplicate learner opportunity records were identified and removed
- Date columns were converted into a proper and consistent date/time format
- Invalid or unrealistic dates of birth were filtered out
- Text fields such as country, gender, and institution names were standardized
- Rows with missing critical information (such as opportunity ID or key dates) were removed

Overall, these steps helped reduce noise in the data and made the dataset reliable for analysis.

EDA Insights:

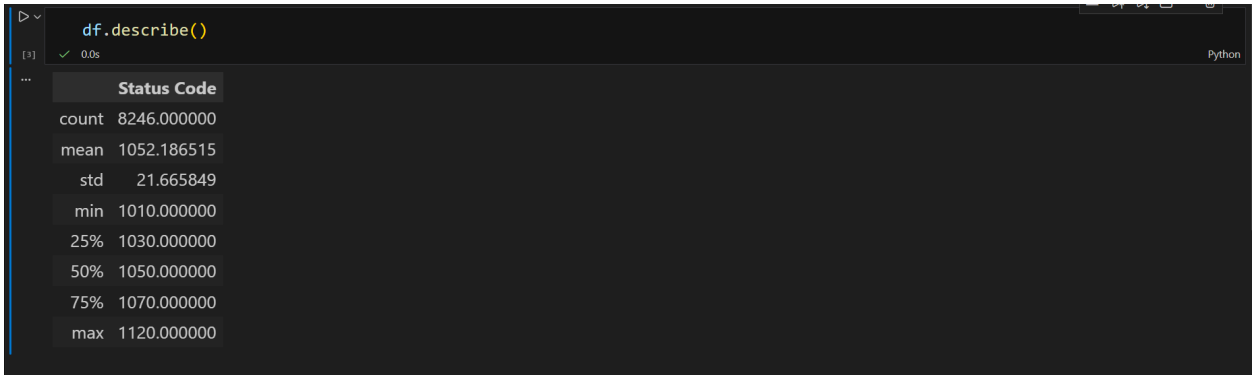
Summary Statistics:

The summary statistics provide a high level understanding of the dataset and help identify general trends and distributions.

- Learner participation varies significantly across opportunities, with a small number of opportunities attracting a large share of learners.
- Country wise analysis shows that learners are concentrated in a few countries, while many countries have relatively low participation.
- Institution wise distribution highlights that certain institutions consistently contribute a higher number of learners compared to others.
- Age analysis shows that most learners fall within the mid-20s age range, indicating that the dataset is dominated by early-career individuals.

- Time-based metrics, such as the time between sign-up and application, reveal that many learners apply quickly, while a smaller group takes a significantly longer time.

These statistics help set the foundation for deeper pattern and trend analysis.



In the dataset, **Status Code is the only column represented as a numerical data type**. It encodes different stages of a learner's application or opportunity status and therefore serves as an important indicator of progression rather than a continuous quantitative measure.

Exploratory analysis shows that the Status Code values range from **1010 to 1120**, with a **mean value of approximately 1052** and a **median of 1050**, indicating that most learners are concentrated around the mid-level status stages. The relatively **low standard deviation (≈21.7)** suggests limited dispersion, meaning learners tend to cluster within a narrow set of predefined status categories rather than being widely spread across all possible statuses.

The interquartile range (25th percentile: 1030, 75th percentile: 1070) further confirms that a large proportion of learners fall within a consistent status progression band. Extreme values at the lower and upper ends occur less frequently, indicating fewer learners at the earliest or final status stages. Overall, this distribution reflects structured and controlled status transitions within the platform.

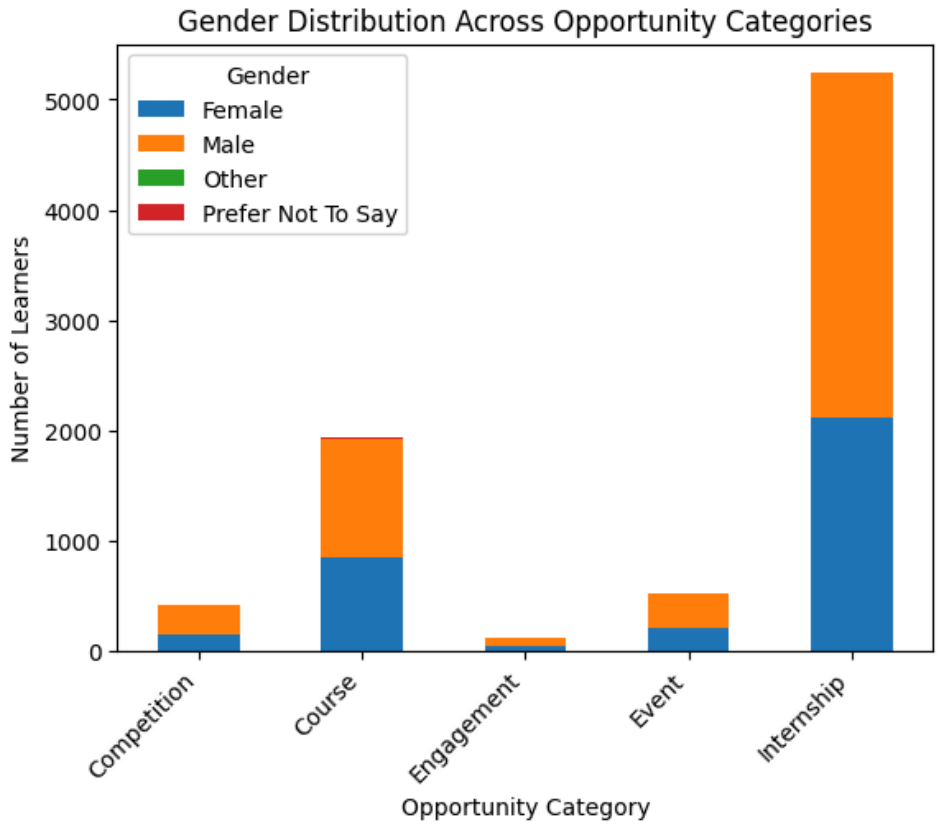
Emerging Patterns:

Several clear patterns emerge when exploring learner behavior and opportunity characteristics:

- Skill based and career oriented opportunities tend to receive higher engagement compared to niche or specialized categories.
- Learners from certain countries and institutions appear more active, suggesting stronger awareness or access to opportunities.
- Some majors are repeatedly observed across multiple opportunities, indicating popular academic or career interests among learners.
- Gender distribution across opportunities appears relatively balanced, with no extreme dominance in most cases.
- Faster application behavior is often associated with higher engagement levels.

These patterns provide insight into how different learner segments interact with opportunities.

Stacked Bar Chart – Gender vs Opportunity Category:



The stacked bar chart shows learner distribution by gender across five Opportunity Categories: Competition, Course, Engagement, Event, and Internship.

- **Total Participation:** Internship leads (~3,200 learners), followed by Course (~2,100), while Engagement is lowest (~500).
- **Gender Breakdown:** Female learners dominate most categories (Internship ~1,900; Course ~1,200), males are second (~Internship 1,200; Course 850). Other and Prefer Not to Say remain minimal (<100 per category).
- **Key Insight:** Internship and Course attract the highest learner engagement, with females representing the largest share.

Trends or Outliers:

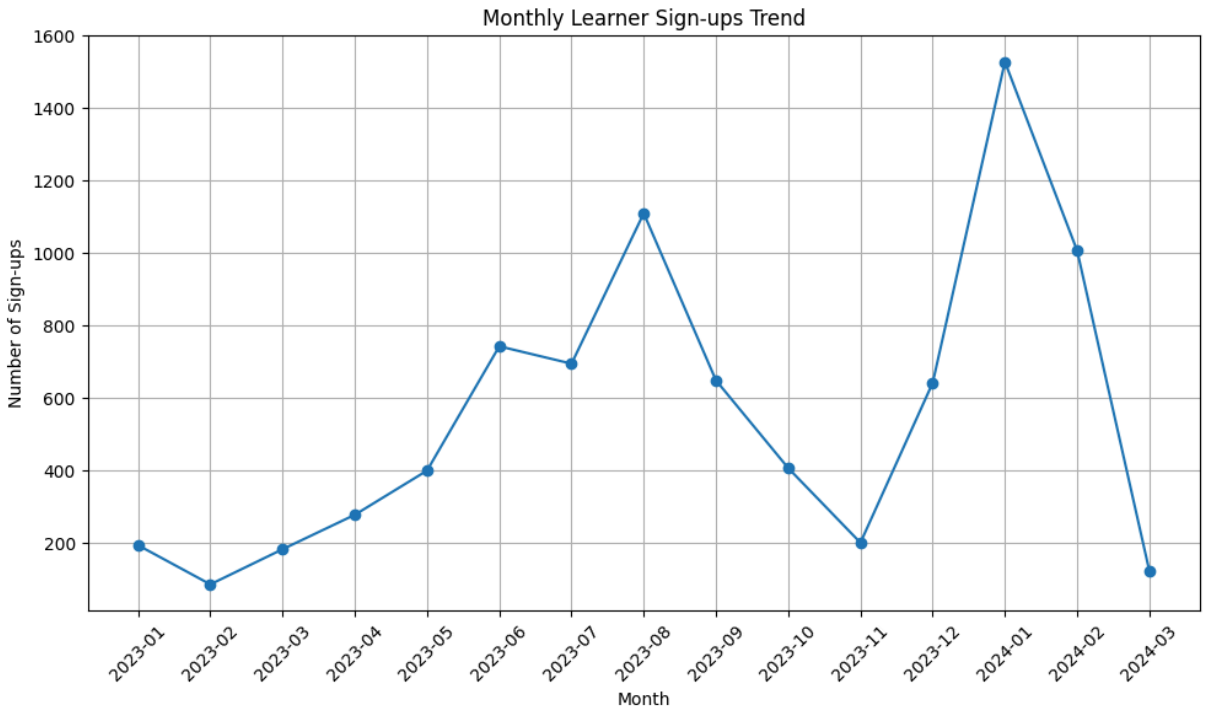
While most records follow expected trends, some unusual cases were identified:

- A small number of opportunities show unusually high or low engagement compared to the average.
- Certain learners have extremely short or extremely long application times, which may indicate late data entry or system delays.

- A few records show negative or inconsistent time gaps between dates, flagged as anomalies.
- Repeated entries for similar learner details suggest re-applications or data duplication before cleaning.

Identifying these outliers helps ensure that final insights are interpreted correctly.

Line Chart-Learner Enrollment Patterns by Month:

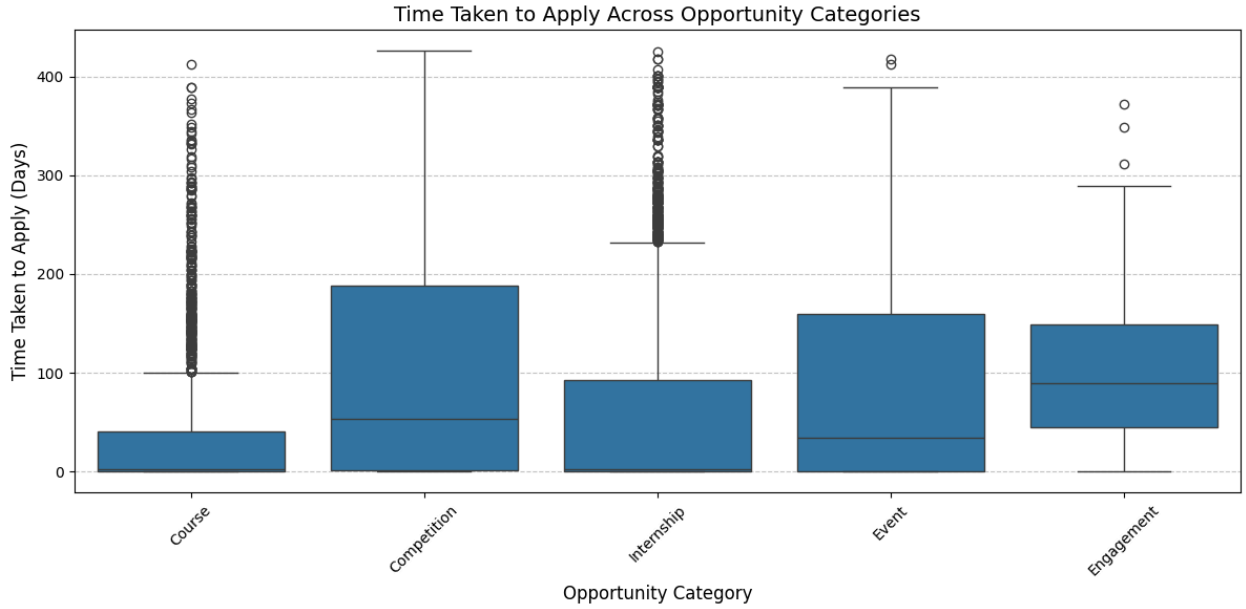


The line chart tracks the number of learner sign-ups from January 2023 (2023-01) to March 2024 (2024-03).The chart shows an overall fluctuating but generally upward trend in learner sign-ups, with distinct peaks and troughs over the period.

- **Initial Trend (Jan 2023 – Aug 2023):** Sign-ups started around 200 in January 2023, rising steadily with minor fluctuations. A mid-year peak occurred in August 2023 at approximately 1,100 sign-ups.
- **Mid-Period Fluctuation (Sep 2023 – Dec 2023):** Following the August peak, sign-ups declined sharply, reaching a low of around 200 in November 2023, before partially recovering in December.
- **Peak Period (Jan 2024):** The highest sign-ups were recorded in January 2024, reaching roughly 1,500.
- **Recent Trend (Feb 2024 – Mar 2024):** Post-January, sign-ups declined steeply to about 1,000 in February and dropped further to around 100 in March 2024, the lowest in the observed period.

This trend highlights seasonal or periodic variations in learner engagement and identifies critical months of high and low activity.

Boxplot of "Time Taken to Apply" across Opportunity Categories:



The box plot shows the distribution of application times (in days) across five opportunity categories: Course, Competition, Internship, Event, and Engagement. It visualizes the minimum, maximum, median (line inside the box), and interquartile range (box).

Key Insights:

- **Longest Typical Application Time:** Engagement and Competition have the highest median times, indicating these applications generally take longer to complete.
- **Shortest Typical Application Time:** Course and Internship have the lowest median times, suggesting most learners complete these applications quickly.
- **Most Variable Application Times:** Competition shows the largest interquartile range, reflecting wide variation in the middle 50% of applicants.
- **Extreme Outliers:** Internship and Course categories exhibit the most extreme outliers, with some learners taking over 400 days to apply, far beyond the typical range.

This visualization highlights differences in application behavior across categories and helps identify trends and unusual patterns in learner engagement.

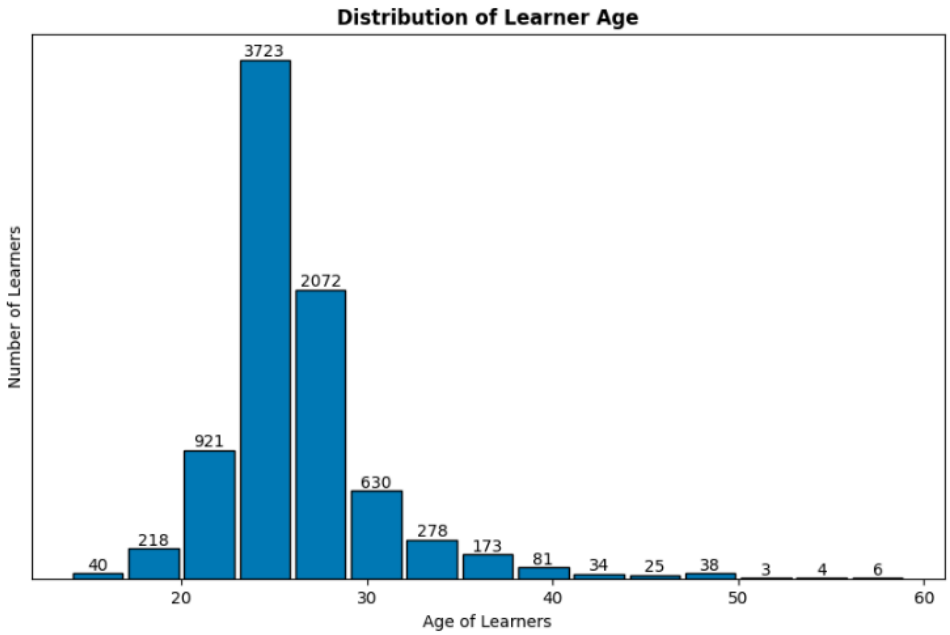
Potential Insights:

Based on the exploratory analysis, several early insights can be drawn:

- Opportunity category and structure may influence how quickly learners apply and complete applications.
- Learner background factors such as country, institution, and major appear to affect engagement levels.
- Opportunities with clear timelines and relevance may encourage faster learner action.
- Understanding these factors can help improve opportunity design and learner outreach strategies.

Visualizations:

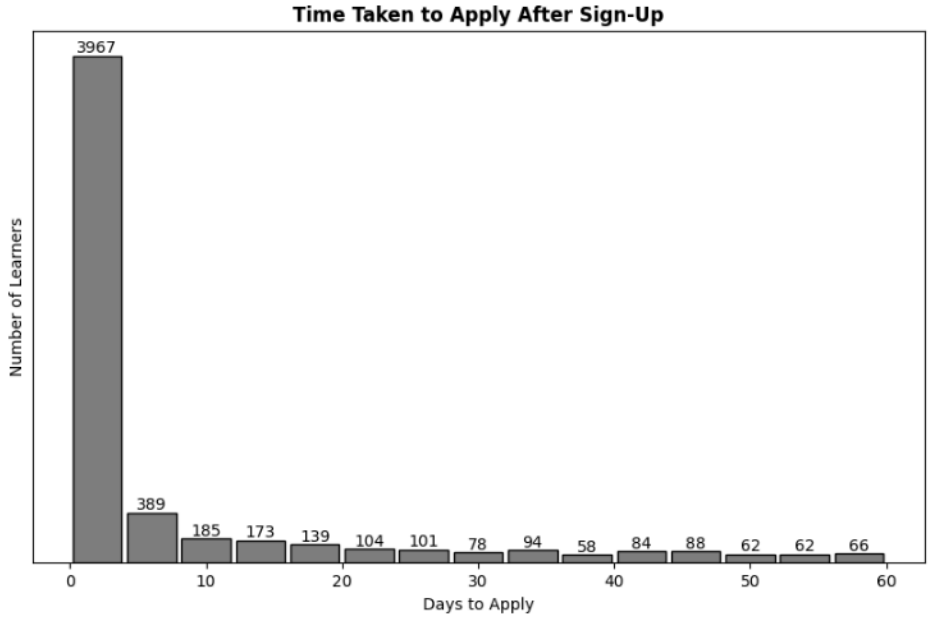
Histogram (Age of Learners):



The Insight: The data reveals a demographic that is heavily concentrated in early adulthood. The distribution is "right-skewed," meaning the majority of learners are bunched up on the left side (younger ages), with a long "tail" stretching out to the right (older ages).

- **The Peak:** The highest engagement comes from learners in their mid-20s (specifically the bin around 25 years old), with **3,723 learners**.
- **The Drop-off:** Participation drops significantly after age 30. While there are learners in their 40s and 50s, they represent a very small niche compared to the core 20–30 demographic.

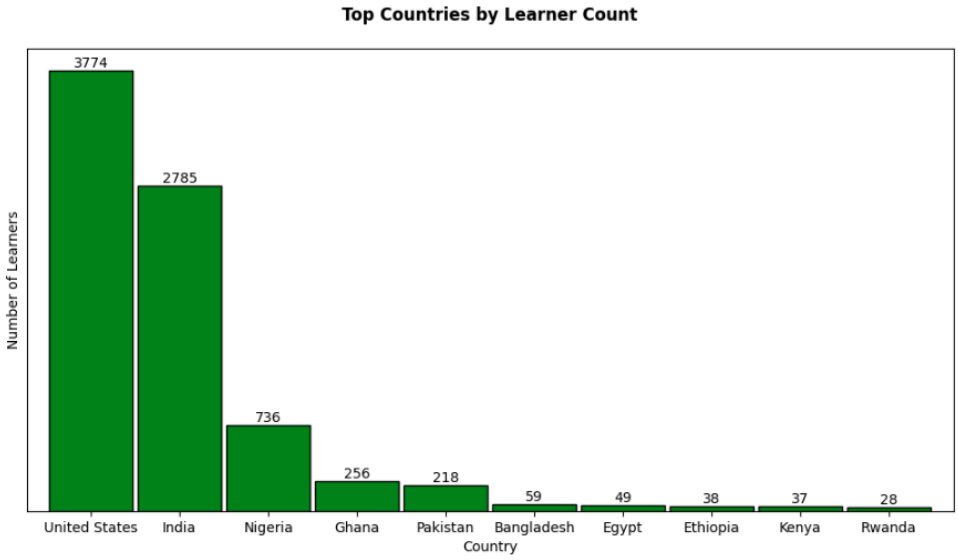
Histogram (Time Taken to Apply):



The Insight: This chart reveals a critical behavior

- **The Cliff:** There is a massive spike in the first bin (0–4 days), where **3,967 learners** applied. This represents the vast majority of the dataset.
- **The Drop-off:** Engagement plummets immediately after the first few days. The second bin drops to just 389 learners.

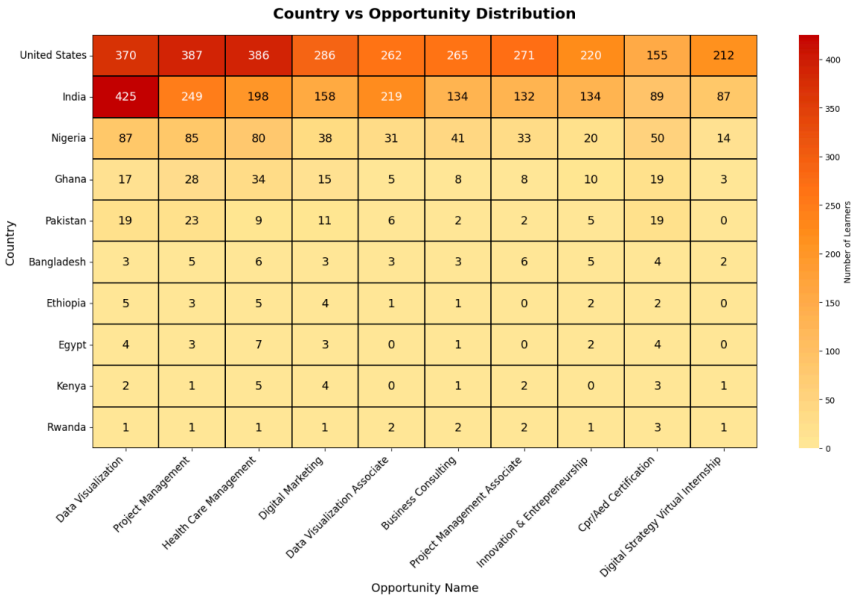
Bar Chart (Top Countries or Institutions):



The Insight: This chart provides a clear hierarchy of your geographic markets.

- **Market Dominance:** The **United States (3,774)** and **India (2,785)** are outliers compared to the rest of the world. Combined, they likely make up over 70-80% of your total learner base.
- **The "Middle" Market:** There is a steep drop to the third place, **Nigeria (736)**, which is still a significant market but much smaller than the top two.
- **The Long Tail:** Countries like Bangladesh, Egypt, and Kenya represent very small segments (under 100 learners each).

Heatmap (Country vs Opportunity):



The Insight: This chart identifies the "sweet spots" where specific countries interact with specific course topics.

- **The Power Players:** The darkest red cells confirm that **India** and the **United States** are the primary drivers of volume.
- **Topic Preferences:**

1. **India** shows the highest single interest in Data Visualization (425 learners), indicating a strong demand for technical analytics skills in that region.
2. **The US** has a more balanced spread but leads slightly in Project Management (387 learners) and Healthcare Management (386 learners).

5. Proposed Variable List

This section covers a set of explanatory variables derived from learner demographics, opportunity details, and learner engagement metrics to support the analysis and prediction of learner outcomes. The selected variables are designed to capture both static background information and dynamic engagement-related behaviors, which prior research has shown to be critical in modeling learner performance and application success.

Variable	Type	Relevence	Notes
age_at_sign_up	Numerical	Represents learner maturity and eligibility, which may influence preparedness and commitment	Computed from Date of Birth and Learner Sign-up Date
gender	Categorical	Enables demographic segmentation and fairness	
country	Categorical	Reflects geographic and socioeconomic context	Group low-frequency countries
institution_name	Categorical	Indicates academic background and preparedness	
current_major	Categorical	Measures alignment with opportunity requirements	Can be grouped by discipline
opportunity_category	Categorical	Strong indicator of opportunity difficulty and fit	Key contextual variable
opportunity_duration	Numerical	Longer durations may increase dropout risk and resource demands	Opportunity End Date - Opportunity Start Date
time_from_signup_to_application	Numerical	Reflects learner motivation and engagement	Highly predictive feature
application_lead_time	Numerical	Early applications may indicate higher motivation	Apply Date - Opportunity Start Date
time_before_deadline	Numerical	Late applications may indicate lower preparedness	Opportunity End Date - Apply Date
n_of_opportunities_applied	Numerical	Serves as a proxy for learner persistence and motivation	Aggregated at learner level

6. Sample Learning Path Map

This section visualizes a typical learner's journey through an opportunity on the Excelerate platform. The learning path map highlights the sequence of engagement stages, key drop-off points, rapid progression stages, and distinguishes between required and optional learner actions.

Typical Learner Journey
A learner generally progresses through the following stages:
Sign-Up → Application → Status Progression → Opportunity Completion

After signing up, learners apply to opportunities and move through different status stages such as *Applied, Team Allocated, Started, Rejected,* or *Dropped Out* until they either complete the opportunity or exit the process.

Drop-Off Points

- Learners who sign up but do not submit an application
- Learners who apply but do not progress during status updates
- Learners who drop out after allocation or starting an opportunity

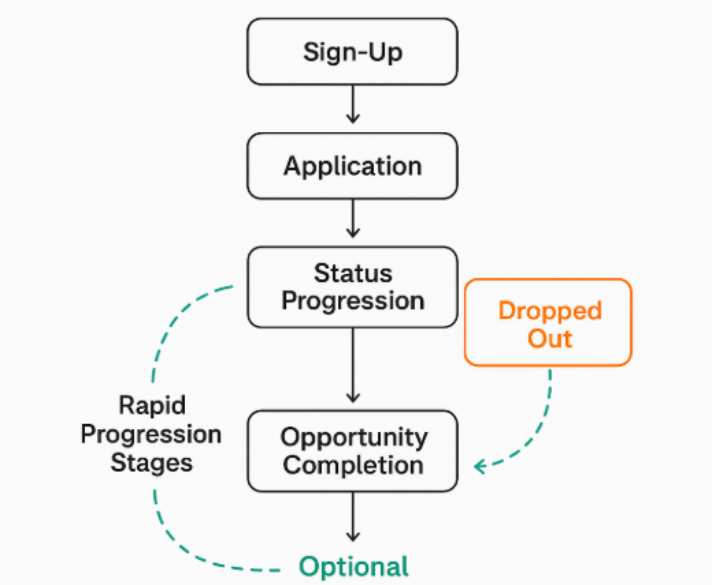
Rapid Progression Stages

- Many learners apply shortly after sign-up, indicating high initial intent
- Faster transitions from *Applied* → *Team Allocated* are observed for suitable learners

Optional vs Required Steps

- **Required:** Sign-Up, Application, Status Progression, Opportunity Completion
- **Optional:** Browsing multiple opportunities, profile updates, repeated visits

The flowchart below visually represents the learner journey using arrows to illustrate progression, drop-offs, and optional engagement paths.



7. Observations & Reflection

1. Learner Progress & The Engagement Funnel

The learner journey on the platform is characterized by high initial intent but significant filtering at the application stage.

- The Funnel:
    - Application Stage: 100% of the active dataset applied.
    - Selection/Allocation: ~42% of applications are Rejected, while ~38% are Team Allocated.
    - Active Participation: Only ~9% of applications reach the Started status.
    - Attrition: Approximately 7.2% of applicants explicitly Dropped Out.

Speed of Engagement:

- Immediate Action: The median time from Sign-up to Application is just ~4 days, indicating that most learners join the platform with a specific opportunity in mind.
- The Long Tail: However, the mean time to apply is ~56 days, revealing a segment of users who "lurk" or browse for months before committing to an opportunity.

2. Patterns Influencing Engagement & Drop-offs

The type of opportunity is the strongest predictor of learner outcomes.

- Category Dynamics:
    - Internships (High Stakes, High Friction): These are the most competitive, with a 66% rejection rate and the highest drop-out rate (11.3%). Only ~22% of applicants are allocated to teams.
    - Courses (High Accessibility): In contrast, Courses have a near-zero rejection rate. Learners are either Started (37%) or Team Allocated (55%). This category drives the bulk of "successful" engagement.
    - Competitions & Events: These are almost exclusively "Team Allocated" (>84%). The lack of a "Started" status here suggests that being allocated to a team is the primary form of participation.
- The "Team Allocated" Anomaly:
    - This is the dominant status (38% overall). For Internships, it likely represents passing the initial screen. For Courses, it is surprisingly high (55%), potentially indicating cohort-based learning where users are grouped but tracking individual "Start" status is secondary.

3. Factors Affecting Performance (Hypotheses)

Demographics and educational background heavily influence the *choice* of opportunity, which in turn dictates success rates.

- The "Country-Category" Effect:
    - United States Anomaly: US learners have a very low "Started" rate (~3%) and high Rejection rate (~49%). Hypothesis: This is selection bias, not underperformance. US learners overwhelmingly apply to Internships (70%), which are highly competitive. They rarely apply to Courses (16%).
    - Pakistan Success Story: Pakistan has the highest "Started" rate (~42%). Hypothesis: Pakistan-based learners prefer Courses (57%), which have high acceptance rates, over Internships (35%).
    - Takeaway: "Performance" differences between countries are actually differences in product preference (Internships vs. Courses).
- Major-Specific Behavior:
    - Tech & Data Majors (CIS, Data Science): These students aggressively target Internships (>75%), leading to higher rejection rates (>50%) and higher drop-off rates (~11-12%). They are likely seeking career experience over learning content.
    - Computer Science (General): These students have a more balanced portfolio, applying to Courses (24%) and Internships (58%), resulting in better overall engagement metrics.

4. Emerging Trends & Anomalies

- Temporal Surge: There was a massive spike in applications in January-February 2024 (over 4,000 applications), significantly higher than the previous year. This suggests a successful marketing campaign or the launch of a highly desirable program cycle during that period.
- Implicit Completion: The lack of a "Completed" tag is a data gap. Since all "Started" users have passed their end dates, the platform likely lacks a feedback loop to close the status to "Completed," or "Started" is being used as a catch-all for "Participated."
- Gender Parity: Engagement metrics are surprisingly balanced. Females have a slightly higher "Started" rate (9.3% vs 8.3%) and slightly lower rejection rate than males, suggesting the platform is effectively reaching and retaining female learners in tech.

8. Conclusion

During Week 1, our team conducted a thorough analysis of the SLU Opportunity-Wise dataset, focusing on understanding learner engagement patterns and platform performance. The following summarizes our key findings and contributions:

1. Understanding the Dataset Structure

Our team mapped the architecture of 8,246 learner records across 16 critical attributes. This dataset provides a comprehensive view of each learner, linking demographic information (Age, Gender, Country) with institutional details (University, Major) and engagement metrics (Application Dates, Status Codes). Through this analysis, we confirmed that the dataset is robust and suitable for studying both learner demographics and program performance.

2. Key KPIs & Success Metrics Identified

Based on the available data, we defined three primary metrics to assess program health:

- **Conversion Rate:** Measured by the transition from Applied → Team Allocated. The overall allocation rate currently stands at ~38%, though it varies significantly by opportunity category.

- **Attrition/Drop-off:** Tracked via Dropped Out and Withdrawn statuses. Internships show a higher attrition risk (~50% of allocated learners) compared to other categories.

- **Inclusivity Index:** Evaluated through gender and geographic distribution. The current male-to-female ratio is 59:41, providing a baseline for future diversity and inclusion initiatives.

3. Exploratory Data Analysis (EDA) Highlights

Our team performed EDA to uncover operational patterns and anomalies:

- **Geographic Focus:** Learners are concentrated primarily in the United States (46%) and India (34%), representing 80% of the total population.

- **Opportunity Mix:** Internships dominate the dataset (64% of records), followed by Courses (23%).

- **Friction Points:** Internships account for all "Rejected" statuses (3,447 rejections), highlighting high competitiveness or resource constraints.

4. Learner Journey Mapping

We analyzed typical learner paths and identified two distinct journeys:

- **High-Stakes Journey (Internships):** Characterized by multiple barriers. Applicants face a ~65% probability of rejection. Even after selection, high dropout rates are observed, suggesting intense program demands.

- **Nurture Journey (Courses):** Characterized by low friction. Learners move smoothly from Application → Team Allocation → Started. Courses act as an effective engagement entry point, potentially serving as a feeder for higher-stakes opportunities.

Overall, our team successfully cleaned, analyzed, and visualized the dataset, identified key KPIs, mapped learner journeys, and highlighted trends and friction points. These insights confirm that the dataset is reliable and ready for deeper analysis, modeling, and future strategic recommendations in subsequent weeks.

End of Report

Thank You !

Data Visualization Remote Internship - Excelerate

Sub-Group 8 Team Charter

Team Members	**Active Members:** Sumaiya Tasnim (sumaiyaa.tasnim.18@gmail.com) Faizan (faizanlal563@gmail.com) Arop Kuol (Kuolbdka@gmail.com) Shri Vidya (vidyashri750@gmail.com) Zainab Fatima (malikzainab1901@gmail.com) Dewpearl Gonsalves (gonsalvesdewpearl@gmail.com) **Inactive Members:** Farooq Zaman (farooqzaman891@gmail.com)
Team Lead	Sumaiya Tasnim (sumaiyaa.tasnim.18@gmail.com)
Department Head Associate	Muhammad Sulait Sufian (msufian@vempower.org)
Associates	Palak Gunani (palak@vempower.org) Kritika (kritika@vempower.org) Shruti Mishra (shrutimishra@gmail.com)
Team Members Roles and Responsibilities	Team Members Roles and Responsibilities:

	Sumaiya Tasnim (sumaiyaa.tasnim.18@gmail.com) - **Team Lead**, represents team to sponsor, via email and on calls, to minimize communication errors. Faizan (faizanlal563@gmail.com) - **Project Manager**, provides guidance and draws out insight from other team members, ensures that the project execution remains on track. Shri Vidya (vidyashri750@gmail.com) - **Project Scribe**, responsible to taking meeting minutes and distributing notes/assignments. Can assist Team Lead in drafting emails and communication between sponsor and group. Arop Kuol (Kuolbdka@gmail.com) - **Project Lead**, responsible for holding the group accountable for meeting deadlines and ensures that the project deliverables are being met.
Mission, Vision Objectives & Core Values	**Mission:** To analyze learner interaction data, uncover actionable insights, and build predictive models that optimize learning pathways. Our team aims to enhance learner engagement and platform performance through data-driven decision-making, visual storytelling, and collaborative analytical rigor. **Vision:** To become a high-impact analytics team that transforms raw learner data into clear, actionable insights, delivering optimized learning experiences and strategic recommendations that advance the platform's educational effectiveness and scalability. **Core Values***:* ● **Integrity** – We uphold transparency and accuracy in every task and report. ● **Accountability** – We take ownership of our responsibilities and deliver on time.

	● **Collaboration** – We support one another through clear communication and teamwork. ● **Excellence** – We aim for continuous improvement and high-quality results. ● **Responsiveness** – We stay proactive and ensure timely communication across all channels. ● **Growth Mindset** – We view every challenge as a learning opportunity to refine our skills and strengthen our impact.
Internal Checks, Balances, and Reviews	● Team check-ins will occur daily or every 2 days unless cancelled in agreement with all team members. ● The tasks leading up to the final deliverable submission will be split between team members in an even manner. ● Team members are expected to help others with their work to ensure the completion and growth of other members. ● Daily or alternate-day check-in meetings will be used to ensure all team members or on track and ensure all doubts are getting clarified. ● Team lead and Project manager are responsible for leading meetings unless they are unable to attend and must notify so and delegate their responsibilities to someone else. ● Then, team members will work to set an agenda for the next meeting and the appropriate task deadlines. ● Project Scribe will keep a record of group discussions, everyone's tasks, and inactive members. They will provide the documentation after the meeting in Google Chat. ● The team will have a final review of each team member's contributions to reduce errors, and they will collaboratively create the final submission for the week.

Operations: **● Assignments** **● Meetings** **●Communication Guidelines** **● Status Updates** **● Deadlines**	**Assignments:** Tasks are distributed to members present during meetings. Absent members can be assigned tasks if they respond within 24 hours. Everyone should check the **Task Tab** regularly for updates and assigned roles. Must acknowledge assigned tasks promptly and meet deliverables. **Meetings:** Regular meetings **every Wednesdays** are required for alignment and updates. First absence is excused; consistent attendance afterward is expected. Must attend all remaining meetings and actively participate. **Communication Guidelines:** Open, respectful, and consistent communication across channels. Reply within **24 hours** and check group chats **every hour** for updates. Stay responsive and maintain professionalism in all group communications. **Status Updates:** Progress must be visible and shared for transparency. Members provide regular updates in the **chat or task tracker**. Non-reporting may affect task reassignment or accountability evaluation. **Deadlines:** All deliverables must meet agreed timelines. Weekly deliverable deadline is on every **Saturday at 11:00 PM IST**. Submit outputs on or before the set deadline. Late submissions require valid reasons.