

Exploratory Data Analysis of Learner Engagement on Excelerate

1. Introduction

This exploratory data analysis (EDA) report aims to provide a comprehensive overview of learner engagement with opportunities available on Excelerate. The purpose of this report is to uncover patterns, trends, and key insights that may inform future strategies for user engagement and opportunity development.

The analysis is based on the Opportunity Sign Up and Completion Data and the User Data, which contains non-identifying information about learners who have signed up for specific opportunities. Each row in the dataset represents a unique sign-up event for a particular opportunity. As learners can enroll in multiple opportunities, there may be several rows corresponding to the same profile ID, representing different engagements.

This report will explore the relationships between sign-up activity, completion rates, and the characteristics of the opportunities, offering valuable insights into learner behavior and engagement patterns.

2. Data Overview- Opportunity sign up and completion data

The dataset analyzed in this report provides a snapshot of learner engagement with various opportunities on Excelerate. Below is a high-level summary of key statistics from the dataset:

2.1 Total Rows: 20,322

Each row represents a unique sign-up event for a specific opportunity, reflecting learner participation at the individual opportunity level.

2.2 Total Columns: 27

These columns capture a range of attributes related to each sign-up, such as the opportunity details, learner identifiers, dates of engagement, and completion statuses.

2.3 Unique Profile IDs: 20,322

Interestingly, the dataset contains 20,322 unique Profile IDs, indicating that each learner in the dataset has signed up for only one opportunity.

3. Data Overview- User data

The dataset analyzed in this report provides a snapshot of learner engagement with various opportunities on Excelerate. Below is a high-level summary of key statistics from the dataset:

3.1 Total Rows: 27,563

Each row represents a unique user data highlighting their preferred sponsors, country, sign up data and whether they are from social media

3.2 Total Columns: 8

These columns capture a range of attributes related to each sign-up, such as the preferred sponsors, sign up date and is from social media.

4. Column Analysis

4.1 Data Preparation

Separated Columns: All columns were separated using the 'Text to Columns' feature to ensure each attribute is in its own distinct column.

Formatted Dates: Columns related to dates, such as Opportunity End Date, Apply Date, and Opportunity Start Date, were formatted to maintain consistency in date representation.

4.2 Data Integrity Checks

Checked for Duplicates: An examination was performed to identify any duplicate entries in the dataset. No duplicates were found, indicating that each sign-up event is unique.

Categorical Data Standardization: Categorical columns, including Gender, State, and Current Student Status, were standardized to ensure consistency and avoid discrepancies due to variations in text formatting (e.g., capitalization, abbreviations).

4.3 Numerical Data Validation

Validated Numerical Data: For columns such as Reward Amount and Skill Points Earned, all blank values were replaced with 0 to ensure data completeness and accuracy.

4.4 Unique Values and Frequencies

Categorical Columns: Each categorical column was analyzed to summarize unique values and their frequencies. This helps in understanding the distribution of categories within the dataset (e.g., the distribution of different gender entries or states).

5. Profile ID Analysis

Uniqueness of Profile IDs: The uniqueness of Profile IDs was examined to ensure that each ID is distinct. This analysis checks for any instances of duplicate or missing Profile IDs, ensuring that each learner is properly represented in the dataset.

6. Opportunity Status Distribution

The Opportunity Sign Up and Completion Data captures the different statuses of learners as they progress through opportunities on Excelerate. The table below summarizes the distribution of various opportunity statuses and the number of learners associated with each status:

Status Description	Count of Profile Id
Applied	34
Dropped Out	24
Not Started	1324
Rejected	726
Rewards Award	2519
Started	810
Team Allocated	13994
Withdraw	622
Grand Total	20053

6.1 Summary of Opportunity Status Distribution

Team Allocated: The largest share of learners, with 13,994 learners (about 70% of the total), have been allocated to teams. This indicates a significant progression toward active participation.

Rewards Awarded: 2,519 learners (12.6%) have successfully completed their opportunities and received rewards.

Not Started: A notable 1,324 learners have not yet started their opportunity despite signing up.

Started: 810 learners (4%) have begun the opportunity but have not yet progressed to completion or team allocation.

Rejected: 726 learners were not able to progress due to rejection.

Withdrawn: 622 learners chose to withdraw after starting the opportunity.

Applied, Dropped Out: The counts for these statuses are relatively low, with 34 learners having applied and 24 having dropped out.

Preferred Sponsors: There were 78,517 votes across five preferred sponsors. Excelerate leads with 16,084 votes, followed closely by Illinois Institute of Technology (15,856), Grant Thornton China (15,845), and Saint Louis University (15,677). GlobalShala has the fewest votes at 15,055.

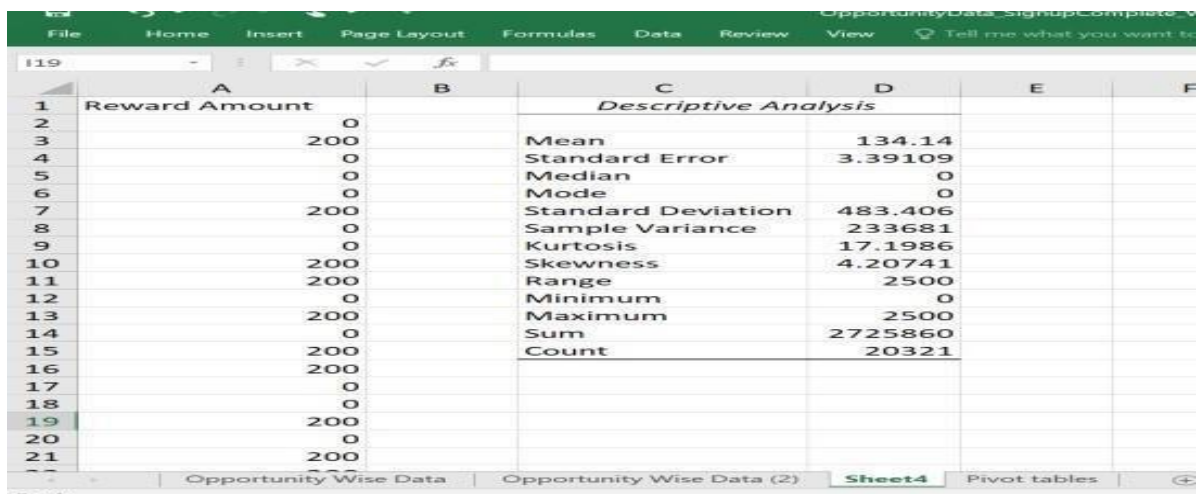
Is from Social Media: The is a distribution of 27,553 individuals based on social media usage. 13,811 are from social media, while 13,742 are not.

This distribution provides key insights into the progression of learners through various stages, helping to identify potential bottlenecks (e.g., high numbers in "Not Started") and successes (e.g., high rewards awarded and team allocations).

7. Basic Statistics

The analysis below provides a summary of key statistics for the two relevant numeric columns in the dataset: Reward Amount and Skill Points Earned. These statistics offer insights into the distribution, variability, and central tendencies of the data.

7.1 Reward Amount (in units)



The screenshot shows an Excel spreadsheet with a 'Descriptive Analysis' table. The table is located in columns C and D, starting from row 2. The first column of the table (C) lists statistical measures, and the second column (D) shows their corresponding values. The data is as follows:

	A	B	C	D	E	F
1	Reward Amount		Descriptive Analysis			
2	0		Mean	134.14		
3	200		Standard Error	3.39109		
4	0		Median	0		
5	0		Mode	0		
6	0		Standard Deviation	483.406		
7	200		Sample Variance	233681		
8	0		Kurtosis	17.1986		
9	0		Skewness	4.20741		
10	200		Range	2500		
11	200		Minimum	0		
12	0		Maximum	2500		
13	200		Sum	2725860		
14	0		Count	20321		
15	200					
16	200					
17	0					
18	0					
19	200					
20	0					
21	200					

Mean: 134.14

The average reward amount across all 20,321 learners is approximately 134.14 units. This relatively low mean suggests that most learners either received no reward or smaller rewards, with a few learners receiving higher rewards.

Median: 0 The median reward is 0, indicating that at least half of the learners in the dataset received no reward.

Minimum: 0

The minimum reward is 0, confirming that many learners did not receive any reward.

Maximum: 2,500

The maximum reward given to a learner was 2,500 units, which is the highest possible reward in the dataset.

Standard Deviation: 483.41

The standard deviation is quite high, reflecting significant variability in the reward amounts distributed, with some learners earning substantially more than others.

Skewness: 4.21 and Kurtosis: 17.20

The positive skewness and high kurtosis suggest that the distribution of reward amounts is highly skewed, with a long tail on the right side. This indicates that while most learners received low or no rewards, a small group received significantly higher rewards.

7.2 Skill Points Earned

G	H	I	J	K
Skill Points Earned		Descriptive Analysis		
0		Mean	147.253	
10		Standard Error	2.91666	
0		Median	0	
0		Mode	0	
10		Standard Deviation	415.775	
0		Sample Variance	172869	
0		Kurtosis	5.6722	
10		Skewness	2.66038	
10		Range	1776	
0		Minimum	0	
10		Maximum	1776	
0		Sum	2992338	
10		Count	20321	
10				
0				
0				
0				
10				
0				
10				

Mean: 147.25

The average number of skill points earned is approximately 147.25, slightly higher than the average reward amount.

Median: 0

Similar to rewards, the median number of skill points earned is 0, meaning that at least half of the learners did not earn any skill points.

Minimum: 0

The minimum skill points earned is 0, consistent with the idea that many learners may not have progressed far enough to earn any points.

Maximum: 1,776

The maximum number of skill points earned by a learner is 1,776, indicating that some learners made significant progress and achieved higher skill points.

Standard Deviation: 415.78

The variability in skill points earned is also high, as reflected by the standard deviation of 415.78, showing a broad range of achievements across learners.

Skewness: 2.66 **and Kurtosis:** 5.67

The positive skewness and kurtosis values indicate that the distribution of skill points is also skewed, but less extremely than the reward amounts. There is still a concentration of learners earning few or no points, with a smaller number achieving high scores.

8. Initial Observations and Patterns from the Exploratory Data Analysis

8.1 Prevalence of Zero Values

A significant number of learners have both 0 reward amounts and 0 skill points earned. This could indicate that many learners did not complete the required actions to qualify for rewards or earn skill points, or perhaps they dropped out early. Investigating the reasons for this could provide insights into learner behavior and engagement with opportunities.

8.1.2 Highly Skewed Data Distributions

Both the reward amounts and skill points earned are highly skewed, with the majority of learners clustered at the lower end of the distribution and only a few achieving high values. This pattern suggests that only a select group of learners are receiving substantial rewards and skill points, which could be worth investigating in terms of learner characteristics (e.g., their demographics, engagement levels, or the types of opportunities they pursued).

8.1.3 Opportunity Status Distribution

The Opportunity Status Distribution indicates that a large proportion of learners have reached the "Team Allocated" stage, with 13,994 entries, while smaller groups either

dropped out, were rejected, or failed to start. It would be valuable to understand what factors contribute to success in reaching the team allocation stage versus those that lead to rejection or dropout.

8.1.4 Consistency and Data Formatting

Some columns, such as Opportunity End Date, Apply Date, and Opportunity Start Date, required formatting. Additionally, categorical columns like Gender, State, and Current Student Status were standardized for consistency. This suggests that some inconsistencies or formatting issues exist in the raw data, and further cleaning or validation may be necessary for accurate analysis.

8.1.5 No Duplicate Profile IDs

The data contains no duplicate Profile IDs, indicating that each learner is uniquely represented in the dataset for each opportunity they sign up for. This eliminates the potential issue of data duplication and suggests that any patterns or trends observed are distinct to individual learners.

8.2 Potential Areas for Deeper Investigation

8.2.1 Learner Engagement and Success Factors

Investigating the characteristics of learners who advance to stages like "Team Allocated" or "Rewards Awarded" versus those who drop out or do not start could reveal important insights into what drives learner success. This might include factors like the type of opportunities, learner demographics, or engagement levels.

8.2.1 Exploration of High Performers

A deeper analysis into the small group of learners with high rewards and skill points could help uncover what differentiates them from others. Understanding their behavior, opportunity choices, or background could provide valuable insights for improving outcomes for other learners.

8.2.2 Impact of Opportunity Categories

Examining whether certain opportunity categories (e.g., career development, technical skills) are associated with higher reward amounts or skill points earned could help in understanding which types of opportunities provide the most value to learners. It could also guide future opportunity design and curation.

8.2.3 Correlation Between Rewards and Skill Points

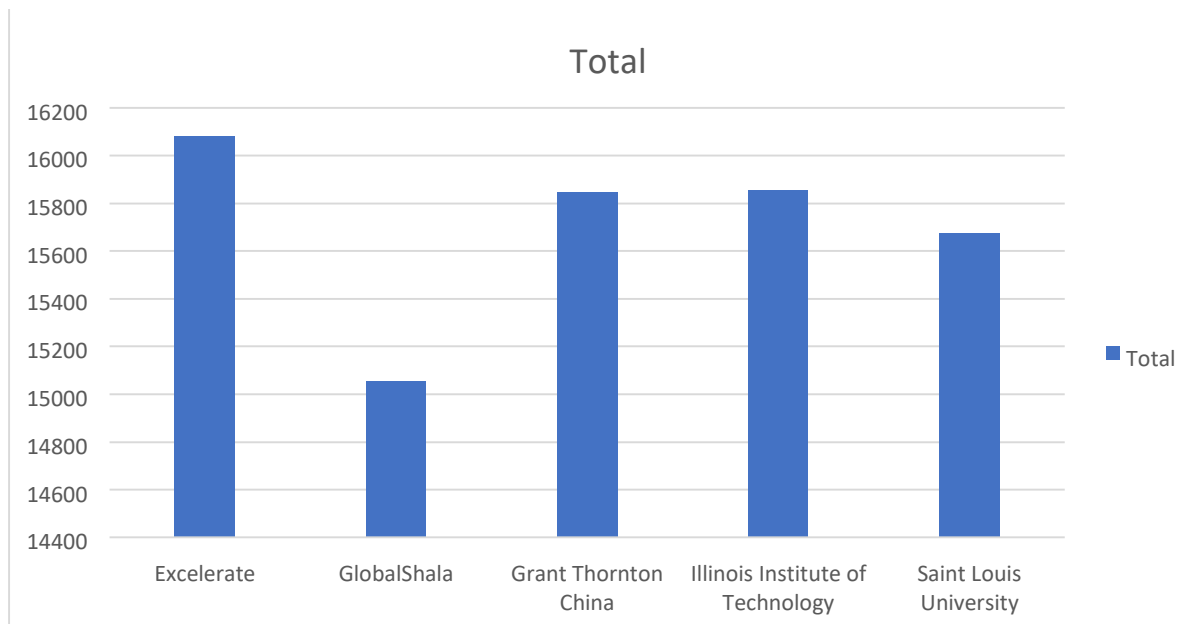
Analyzing the relationship between reward amounts and skill points earned could reveal if there is a direct correlation between the two metrics, i.e., do learners who earn more rewards also tend to earn more skill points? This could be important in understanding the overall engagement and performance of learners.

8.2.4 Categorical Insights

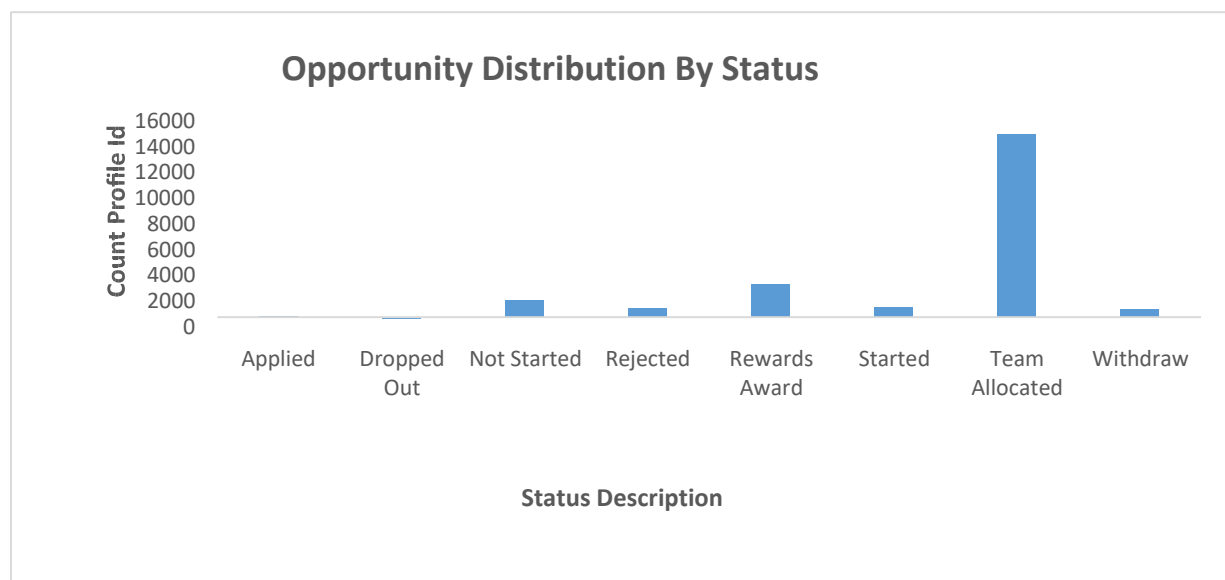
Investigating categorical columns like Gender, State, and Current Student Status may uncover potential disparities or trends in how different demographic groups engage with opportunities and achieve success. Identifying such trends could help in tailoring opportunities to better suit diverse learner needs.

9. Visualizations

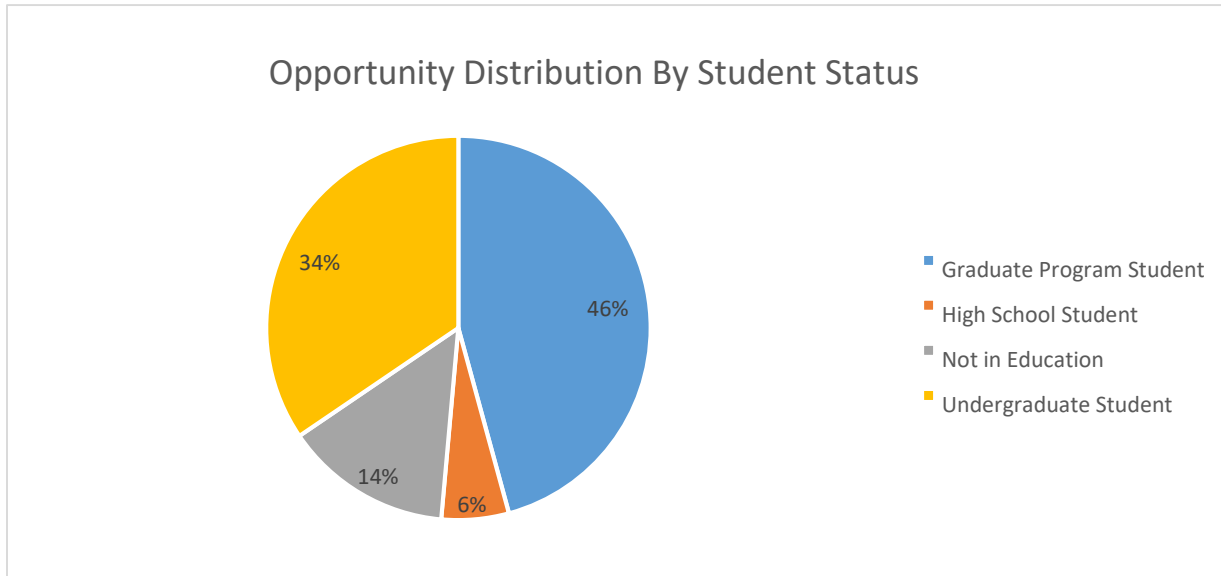
9.1 Status Distribution



The chart displays the distribution of preferred sponsors, with a total count of 78,517 across five sponsors. Excelerate leads with 16,084 votes, followed by Illinois Institute of Technology (15,856), Grant Thornton China (15,845), and Saint Louis University (15,677), while GlobalShala has the lowest preference at 15,055. The bar chart highlights Excelerate's slightly higher preference and GlobalShalala comparatively lower count, while the remaining sponsors show a similar distribution around 15,800 votes each.

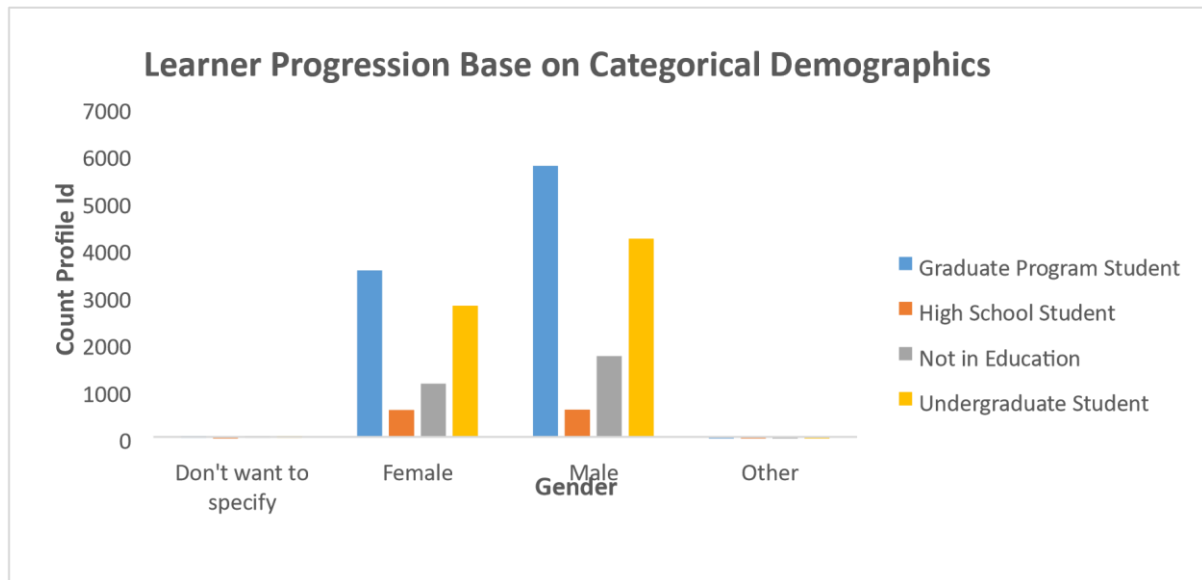


The status distribution shows that most learners (14,205) are in the Team Allocated stage, indicating strong participation. Rewards Awarded follows with 2,521 learners, while 1,324 have Not Started, suggesting some drop-off. Smaller numbers are in Rejected (726), Withdrawn (622), and Applied (89) statuses, highlighting varying engagement levels.



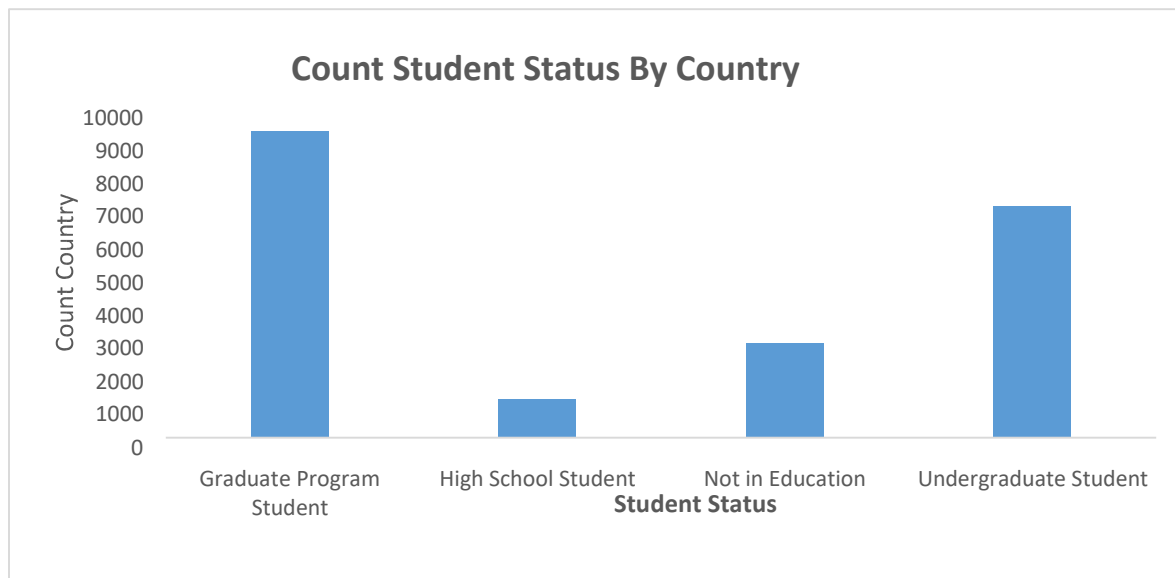
The student status pie chart shows that the majority of learners are Graduate Program Students (9,297), making up nearly half of the dataset. Undergraduate Students follow with 7,009 learners. A smaller portion, 2,862 learners, are Not in Education, while High School Students make up the smallest group with 1,152 learners. This distribution highlights a strong representation of higher education students engaging with opportunities.

9.2 Demographic Distribution:

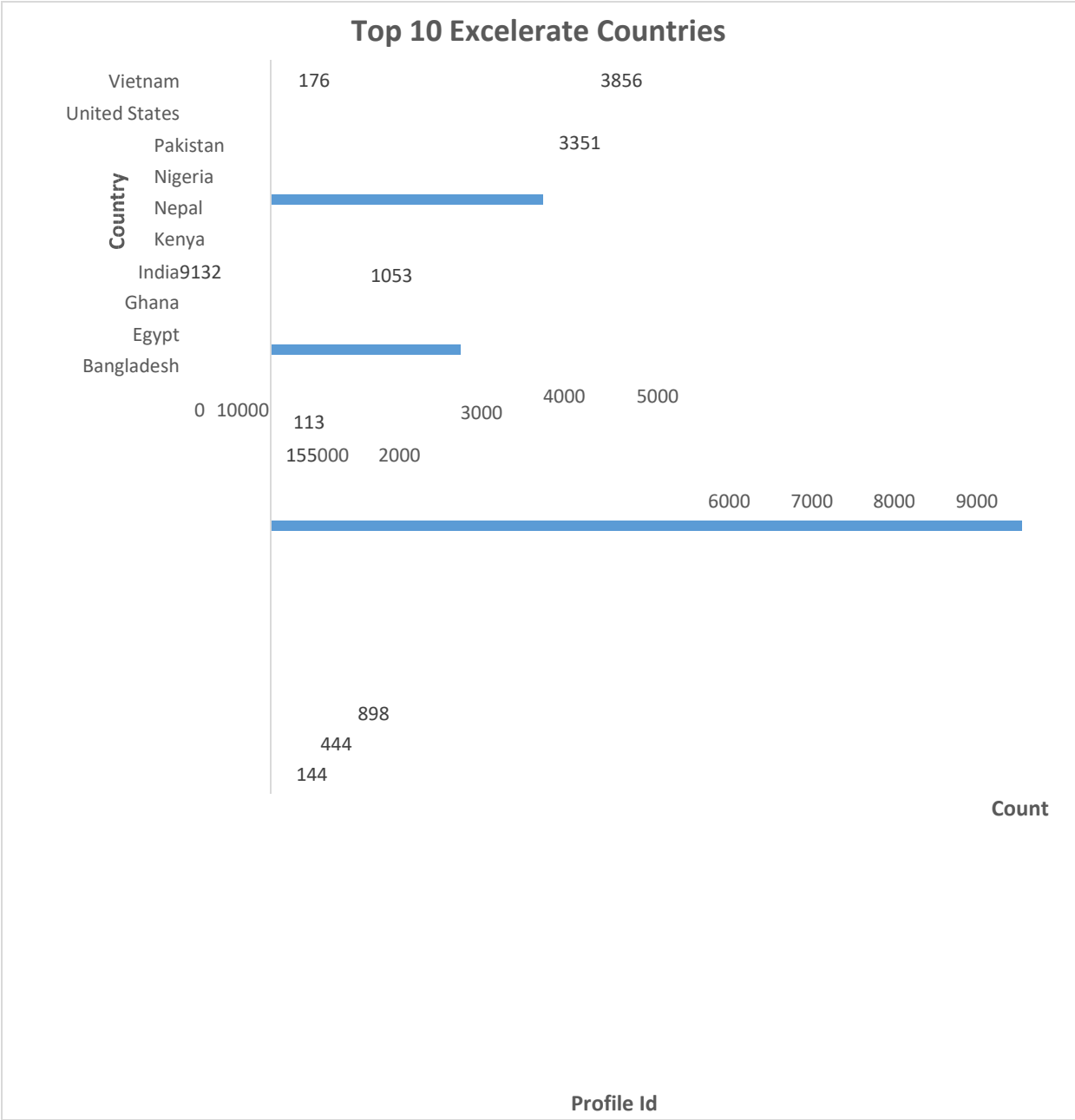


The demographic distribution chart shows more male learners (12,239), especially in Graduate and Undergraduate programs. Female learners total 8,004, also primarily in these categories. A small number of learners identify as Other (14) or prefer not to specify (63). Male participation is higher, but female engagement remains strong in higher education.

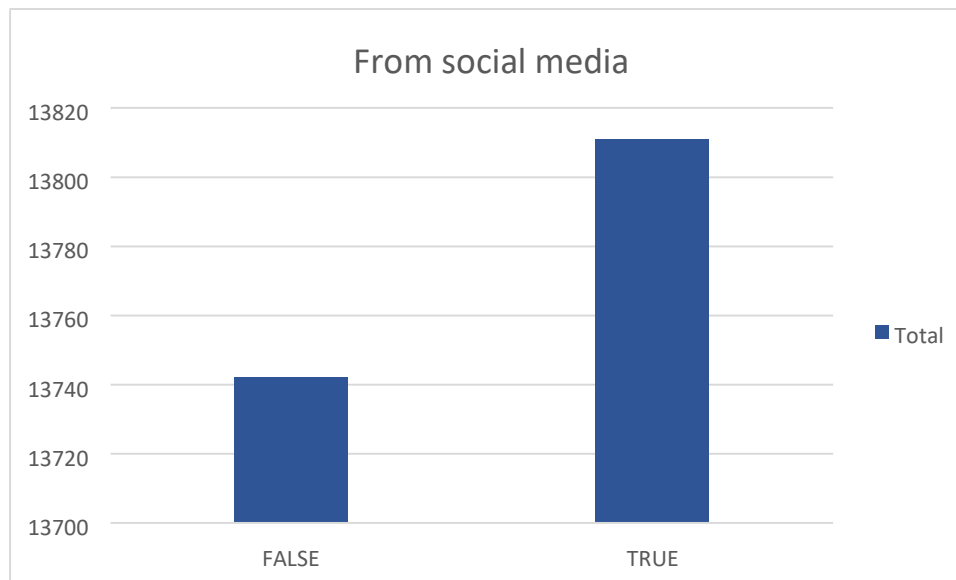
9.3 Demographic distribution by country



The demographic distribution by country column chart reveals that Graduate Program Students (9,297) form the largest group, followed by Undergraduate Students (7,009). Not in Education learners account for 2,862, while High School Students are the smallest group with 1,152. This indicates that higher education students dominate across countries.



The chart shows a high concentration of profiles in India (9,132), followed by the United States (3,856) and Nigeria (3,351). Countries like Kenya, Nepal, and Bangladesh have much smaller numbers, indicating the platform's strongest presence in India, the US, and Nigeria. This suggests these regions are key markets, while other countries show lower engagement, representing opportunities for growth.



The chart presents the distribution of individuals based on whether they are from social media, with a total count of 27,553. Out of this, 13,811 are from social media (marked as TRUE), while 13,742 are not (marked as FALSE). The bar chart visually shows a slight difference, with a nearly equal split between those from social media and those who are not, though individuals from social media account for a marginally higher count.

10. Challenges Encountered

Handling Missing Data:

Reward Amount and Skill Points Earned had missing values that were replaced with 0. This might mask distinctions between non-participation and non-reporting, requiring further clarification.

Data Formatting Issues:

Columns such as Opportunity End Date, Apply Date, and Start Date required formatting adjustments, which slowed the exploration process.

Standardization of Categorical Data:

Ensuring consistency in columns like Gender, State, and Current Student Status was necessary due to variations in data entry, leading to extra steps in preprocessing.

Highly Skewed Data:

Both the reward amounts and skill points earned were skewed, limiting the use of certain statistical analyses without transformation.

10.1 Missing or Unclear Data Points

Unclear Definitions of Statuses:

Certain statuses like "Team Allocated" and "Rewards Awarded" need clearer definitions to interpret their impact.

Incomplete or Missing Demographic Information:

Some categorical columns like City and Opportunity Category lack clarity in their unique values, affecting their analysis.

11. Next Steps (Week 2)

Analyze Learner Outcomes: Investigate the impact of different statuses (e.g., "Team Allocated," "Rewards Awarded") on learner progression.

Demographic Analysis: Explore patterns in gender, state, and current student status to understand their influence on learner success.

Address Data Skewness: Apply transformations to correct skewed reward and skill point data for more accurate analysis.

Examine Opportunity Categories: Determine if certain opportunity types lead to higher completion rates or better performance.

Refine Data Quality: Continue resolving missing values and inconsistencies to improve overall data reliability.