# Exploratory Data Analysis (EDA) Report

# INDEX

# 1. INTRODUCTION

## 1.1 Purpose of Exploratory Data Report

The purpose of the Exploratory Data Report (EDR) is to uncover valuable insights, patterns, and trends within the User Data and Opportunity Sign Up and Completion Data. This report serves as a comprehensive exploration of the datasets, aiming to provide a deeper understanding of user behavior, demographic trends, and the factors influencing the success of opportunities on the platform. By leveraging data analysis techniques, the EDR aims to:

- **Inform Decision-Making:** The insights derived from the data can inform strategic decision-making processes. Understanding user behavior and preferences allows for more informed choices in designing and optimizing opportunities.
- **Enhance User Experience:** By identifying patterns in user engagement, the report can highlight areas for improvement in the user experience. This may include refining the platform interface, tailoring opportunities to user preferences, and optimizing the overall user journey.
- **Optimize Opportunity Offerings:** Analysis of the Opportunity Sign Up and Completion Data enables the identification of key success metrics. This information is crucial for optimizing the types of opportunities offered, ensuring they align with user interests and increasing the likelihood of successful engagement.
- **Identify Demographic Trends:** The exploration of demographic data can reveal important trends among different user segments. This information is valuable for targeted marketing, personalized user experiences, and addressing the unique needs of diverse user groups.
- **Improve Platform Performance:** Insights into user behavior, preferences, and success metrics contribute to overall platform optimization. By addressing pain points and capitalizing on strengths, the platform can be refined to enhance performance and user satisfaction.

The Exploratory Data Report serves as a bridge between raw data and actionable insights. It empowers stakeholders to make data-driven decisions, fostering continuous improvement and innovation within the platform based on a thorough understanding of user dynamics and opportunity success factors.

## 1.2 Datasets Overview

### 1.2.1 User Data

- This dataset encompasses non-identifying information about every user who has ever created an account on Excelerate.

- The data is comprehensive, covering all users, regardless of their engagement with specific opportunities.
- Each row represents a unique user, and the dataset provides a holistic view of the user base.

| Column Heading | Definition |
|---|---|
| **PreferredSponsors** | On the Excelerate Platform, learners can choose their sponsors i.e, who they want to see opportunities from. This column shows the different values selected by the learner who has signed up for the platform. Learners can choose one or more sponsors. |
| **Gender** | Shows the gender indicated by the user upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **Country** | Shows the country which the learner has indicated they live in upon sign up. |
| **Degree** | Shows the academic level indicated by the user upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **Sign up date** | Date on which they created their Excelerate account. |
| **city** | Shows the city which the learner has indicated they live in upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **zip** | Shows the zip code of the city which the learner has indicated they live in upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **isFromSocialMedia** | Shows whether the learner has signed up via a social media login. If True, they have signed up via Google Login. If False, they have manually signed up. |

### 1.2.2 Opportunity Sign Up and Completion Data

- This dataset focuses on non-identifying user information related to learners who have engaged with specific opportunities on Excelerate.
- Each row corresponds to a learner who has signed up for a particular opportunity.
- As learners may sign up for multiple opportunities, there could be multiple rows with the same profile ID.

| **Column Heading** | **Definition** |
|---|---|
| **Profile Id** | Unique AlphaNumeric Identifier on Excelerate for their profile. |
| **Opportunity Name** | Name of which opportunity (experience) they participated in. |
| **Opportunity Category** | The category of the experience they participated in- internship/event/competition/course. |
| **Gender** | Given Gender of the student on Excelerate. |
| **City** | Given City of the student on Excelerate. |
| **State** | Given State of the student on Excelerate. |
| **Country** | Given Country of the student on Excelerate. |
| **Current Student Status** | Have they identified themself as High School/Undergrad/ Graduate Student or Not in Education. |
| **Current/Intended Major** | The major they are currently pursuing or would want to pursue. |
| **Status Description** | What is the status of their application right now. There are 8 possible statuses:<br>1. APPLIED: The learner has made an application (applied) and shown interest in participating in that particular opportunity (experience) on Excelerate. Their application has not been evaluated as yet to be accepted or rejected.<br>2. TEAM ALLOCATED: The learner has been accepted to participate in the opportunity and has been allocated a |

| | |
|---|---|
| | start date for beginning the same.<br>3. DROPPED OUT: The learner has left the opportunity midway without completing it<br>4. NOT STARTED: The learner did not appear on the given start date, and hence did not start the opportunity<br>5. REJECTED: The learner did not meet the eligibility criteria for participating in the particular opportunity and hence was not accepted<br>6. REWARDS AWARD: The associated rewards for the opportunity (badge/scholarship) have been awarded to the learner. The opportunity is COMPLETED by the learner. |

# 2. DATA OVERVIEW

## 2.1 User Data

**Basic information about the dataset:**

- RangeIndex: 27562 entries, 0 to 27561
- Data columns: 8 columns

**Summary Statistics:**

| Column | Count | Unique | Freq |
|---|---|---|---|
| PreferredSponsors | 27562 | 94 | 22011 |
| Gender | 18027 | 4 | 11027 |
| Country | 27500 | 169 | 11893 |
| Degree | 16750 | 4 | 6527 |
| Sign Up Date | 27562 | 27561 | 2 |
| city | 18029 | 4728 | 743 |
| zip | 18028 | 7454 | 629 |
| isFromSocialMedia | 27553 | 2 | 13811 |

## 2.2 Opportunity Sign Up and Completion Data

**Basic information about the dataset:**

- RangeIndex: 20322 entries, 0 to 20321
- Data columns: 21 columns
- Unique identifier: Profile Id

**Summary Statistics:**

| Column | Count |
|---|---|
| Profile Id | 20322 |
| Opportunity Id | 20322 |
| Opportunity Name | 20322 |
| Opportunity Category | 20322 |
| Opportunity End Date | 20322 |

| | |
|---|---|
| Gender | 20321 |
| City | 20321 |
| State | 20316 |
| Country | 20322 |
| Zip Code | 20320 |
| Graduation Date(YYYY MM) | 20321 |
| CurrentStudentStatus | 20321 |
| Current/Intended Major | 20318 |
| Status Description | 20322 |
| Apply Date | 20322 |
| Opportunity Start Date | 19518 |
| Reward Amount | 2521 |
| Badge Id | 2521 |
| Badge Name | 2521 |
| Skill Points Earned | 2521 |
| Skills Earned | 2521 |

# 3. COLUMN ANALYSIS

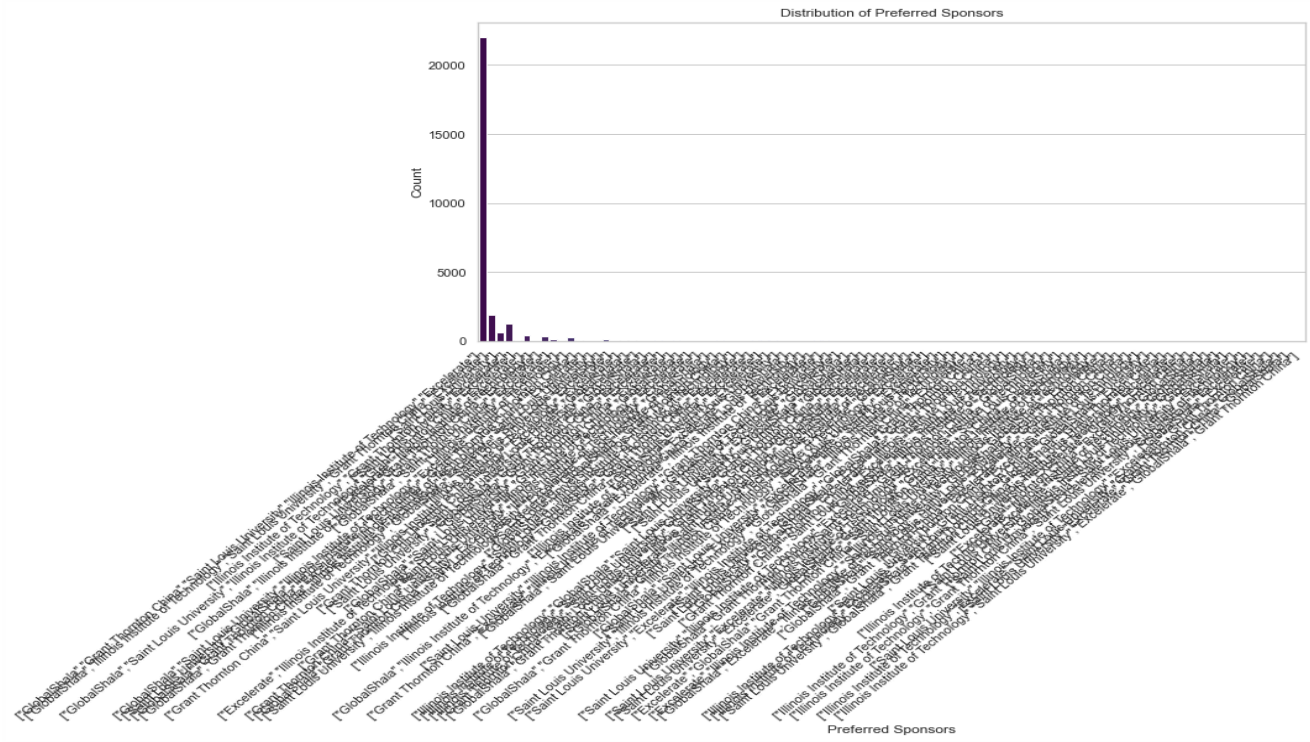
## 3.1 User Data

### 3.1.1 Data Types and Potential Issues:

| | Data Types | Missing Values |
|---|---|---|
| **PreferredSponsors** | object | 0 |
| **Gender** | object | 9535 |
| **Country** | object | 62 |
| **Degree** | object | 10812 |
| **Sign Up Date** | datetime64[ns, UTC] | 0 |
| **city** | object | 9533 |
| **zip** | object | 9534 |
| **isFromSocialMedia** | object | 9 |

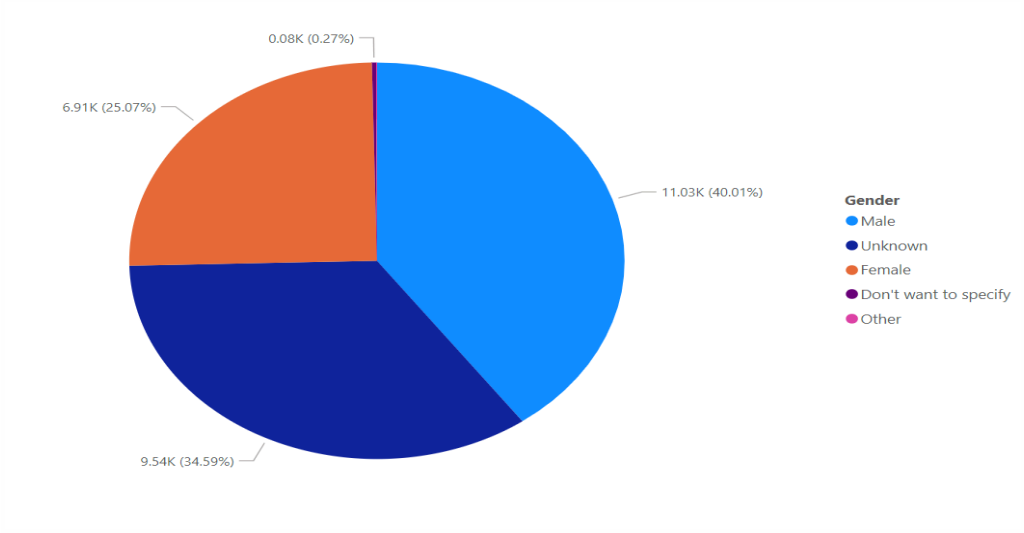
### 3.1.2 Categorical Columns Summary

#### a. Column Preferred sponsors



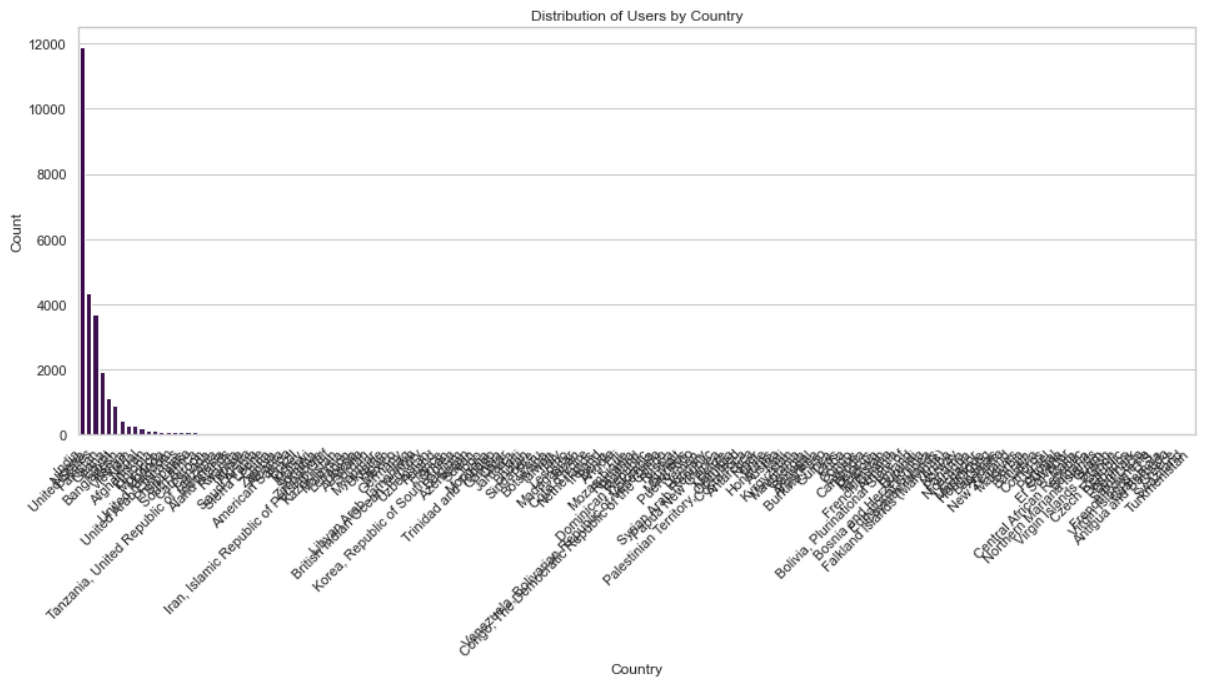
### b. Gender

| | |
|---|---|
| Male | 11027 |
| Female | 6910 |
| Don’t want to specify | 75 |
| Unknown | 9535 |
| Other | 15 |



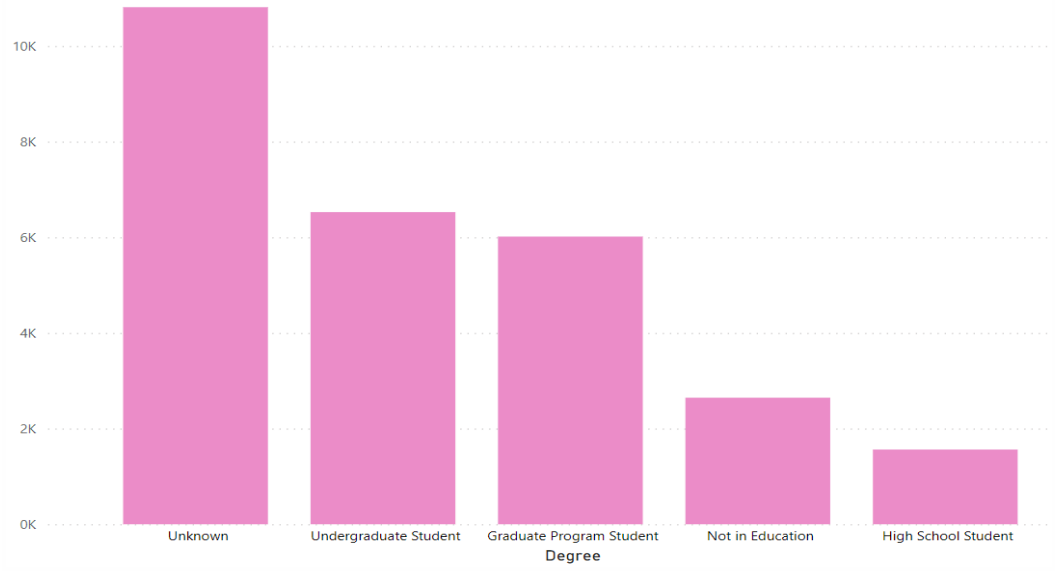

### c. Column Country

| | |
|---|---|
| India | 11893 |
| Nigeria | 4357 |
| United States | 3961 |
| Pakistan | 1928 |
| Ghana | 1124 |
| Egypt | 897 |
| Others | 3403 |

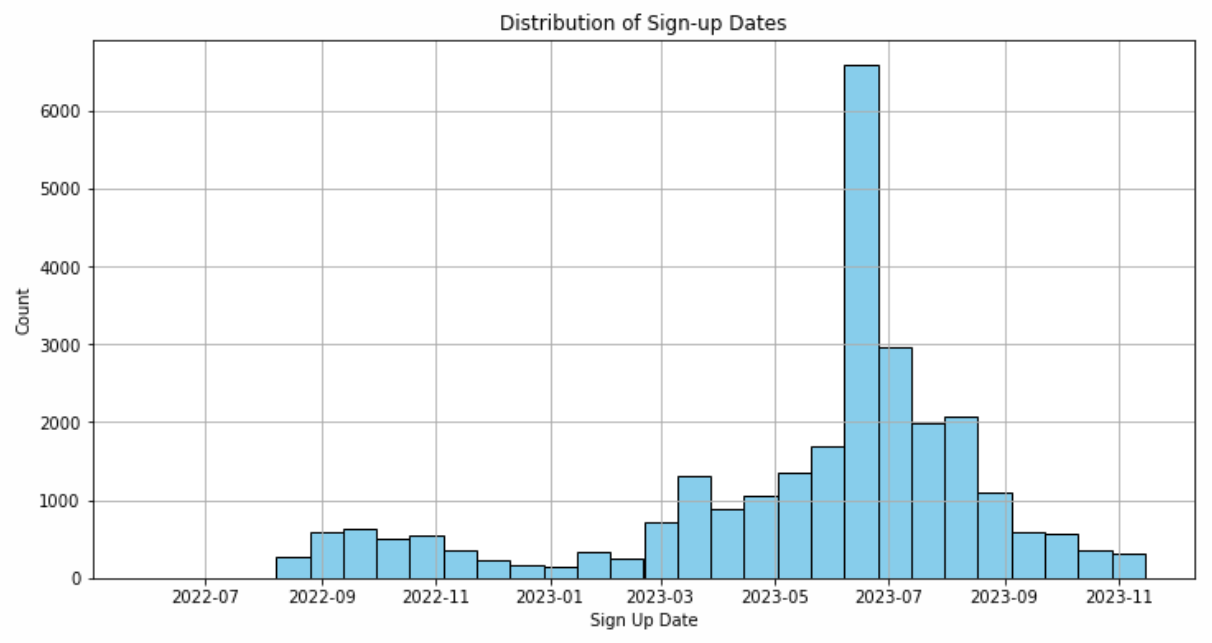



### d. Column Degree

| | |
|---|---|
| Undergraduate | 6527 |
| Graduate | 6015 |
| High school | 1562 |
| Not in education | 2646 |
| unknown | 10812 |

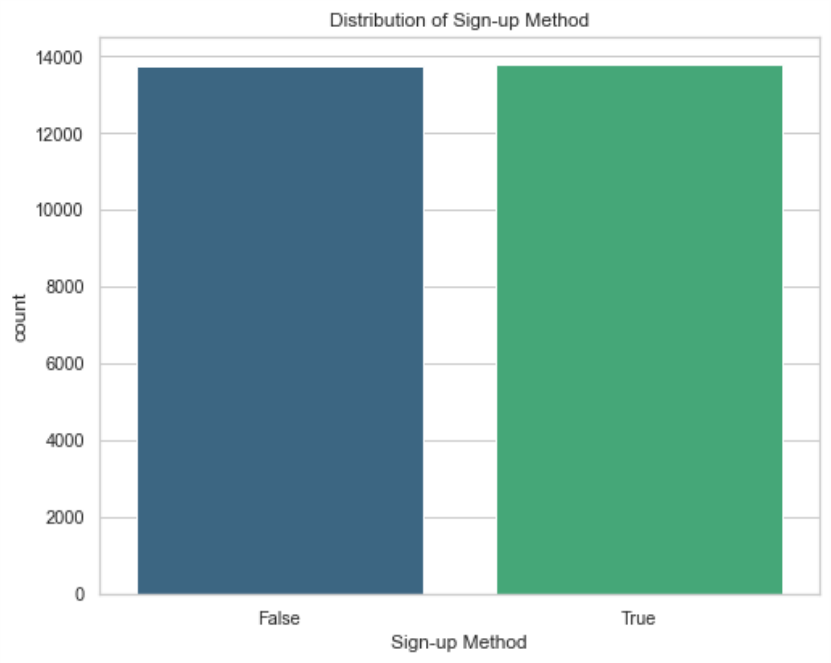


### e. Column Sign-Up Date




### f. Column isFromSocialMedia

| | |
|---|---|
| True | 13811 |
| False | 13742 |



### 3.2 Opportunity Sign Up and Completion Data

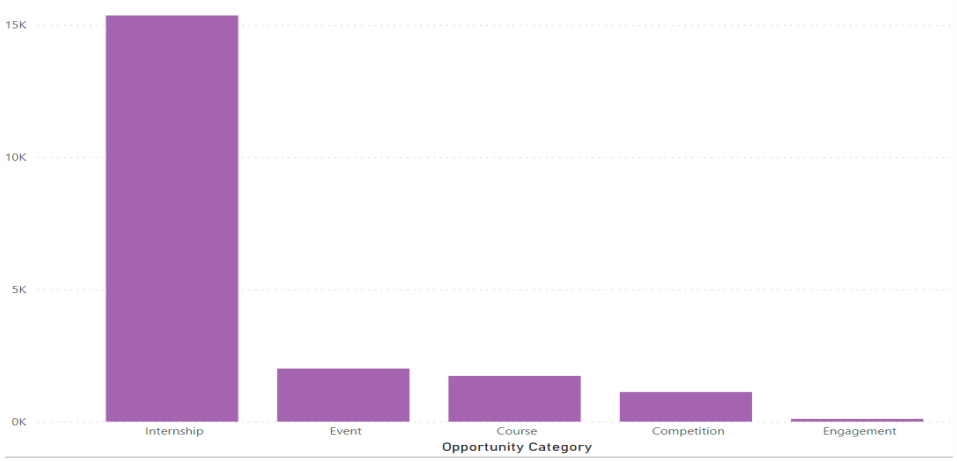
#### 3.2.1 Data Types and Potential Issues:

| Column | Data Types | Missing Values |
|---|---|---|
| Profile Id | object | 0 |
| Opportunity Id | object | 0 |
| Opportunity Name | object | 0 |
| Opportunity Category | object | 0 |
| Opportunity End Date | object | 0 |
| Gender | object | 1 |
| City | object | 1 |
| State | object | 6 |
| Country | object | 0 |
| Zip Code | object | 2 |
| Graduation Date(YYYY MM) | object | 1 |
| CurrentStudentStatus | object | 1 |
| Current/Intended Major | object | 4 |
| Status Description | object | 0 |
| Apply Date | object | 0 |
| Opportunity Start Date | datetime64[ns] | 804 |
| Reward Amount | object | 17801 |
| Badge Id | object | 17801 |
| Badge Name | object | 17801 |
| Skill Points Earned | float64 | 17801 |
| Skills Earned | object | 17801 |

### 3.2.2 Categorical Columns Summary

#### a. Opportunity category

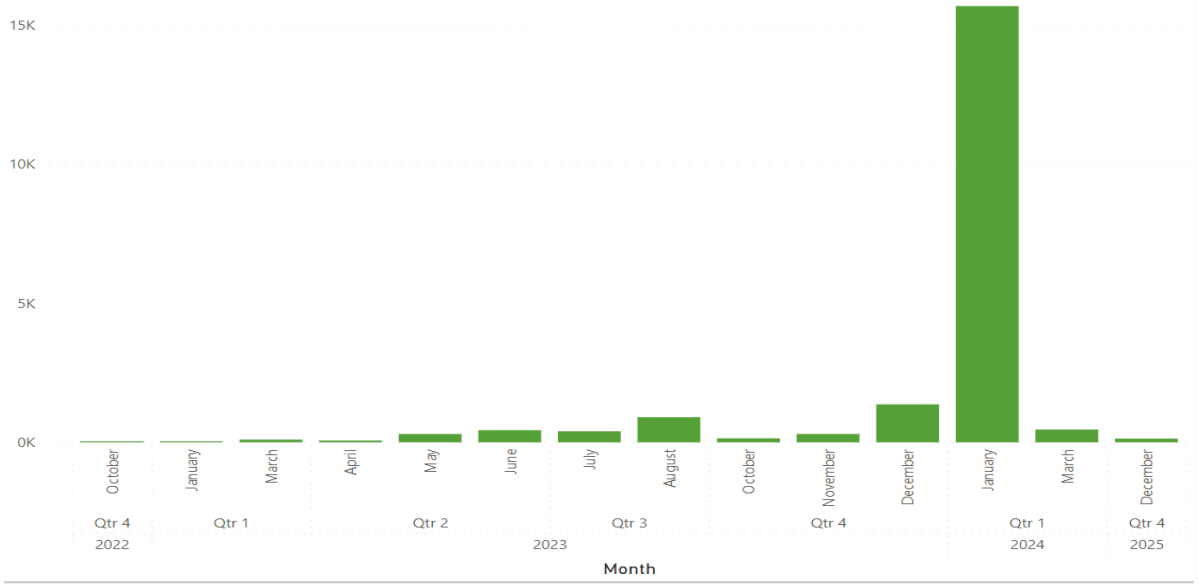
| | |
|---|---|
| Internship | 15360 |
| Event | 2007 |
| Course | 1730 |
| Competition | 1121 |
| Engagement | 104 |




#### b. Opportunity End Date

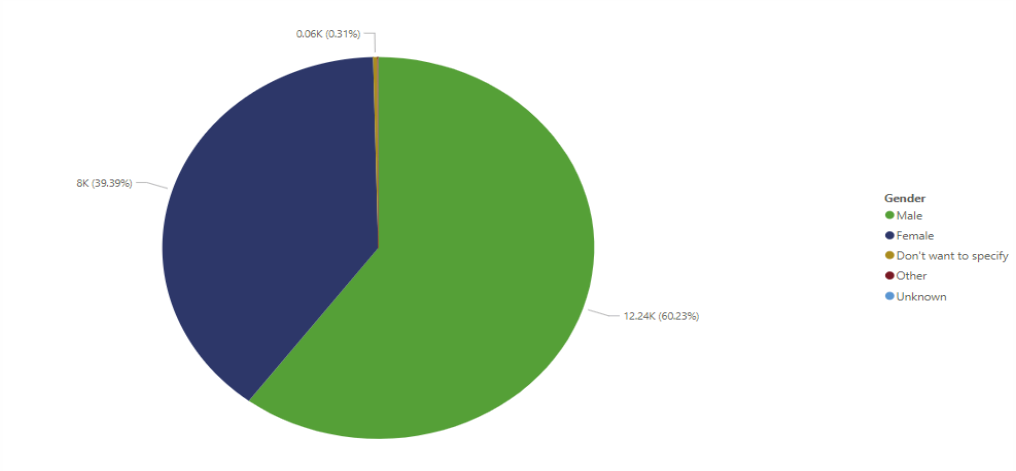
| | |
|---|---|
| March 2023 | 98 |
| April 2023 | 66 |
| May 2023 | 299 |
| June 2023 | 437 |
| July 2023 | 397 |
| August 2023 | 904 |
| October 2023 | 143 |
| November 2023 | 301 |
| December 2023 | 1365 |
| January 2024 | 15692 |

| | |
|---|---|
| March 2024 | 460 |
| December 2024 | 133 |



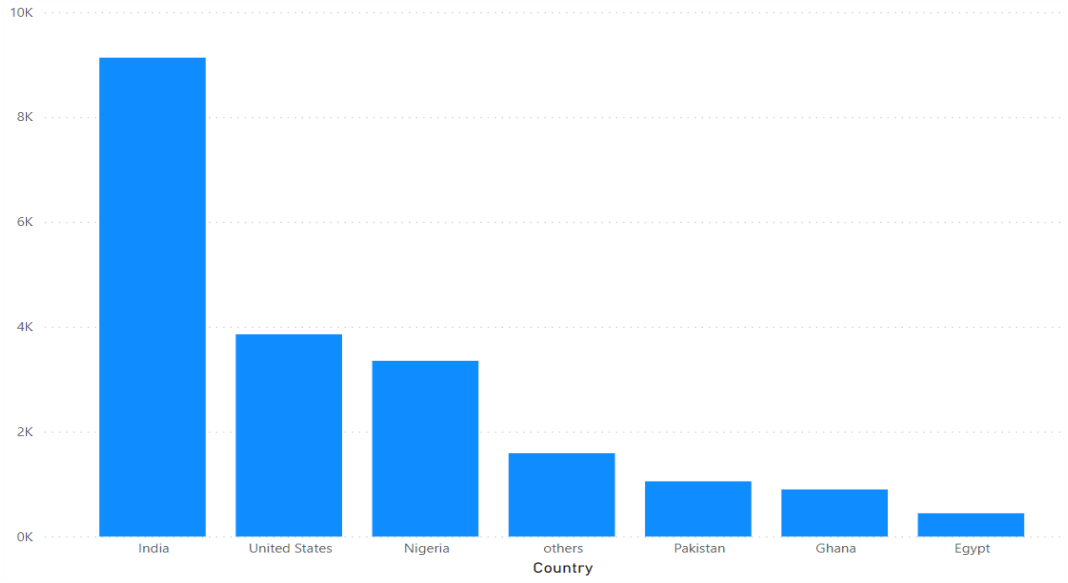

### c. Gender

| | |
|---|---|
| Male | 12240 |
| Female | 8004 |
| Prefer not to say | 63 |
| Other | 14 |
| unknown | 1 |



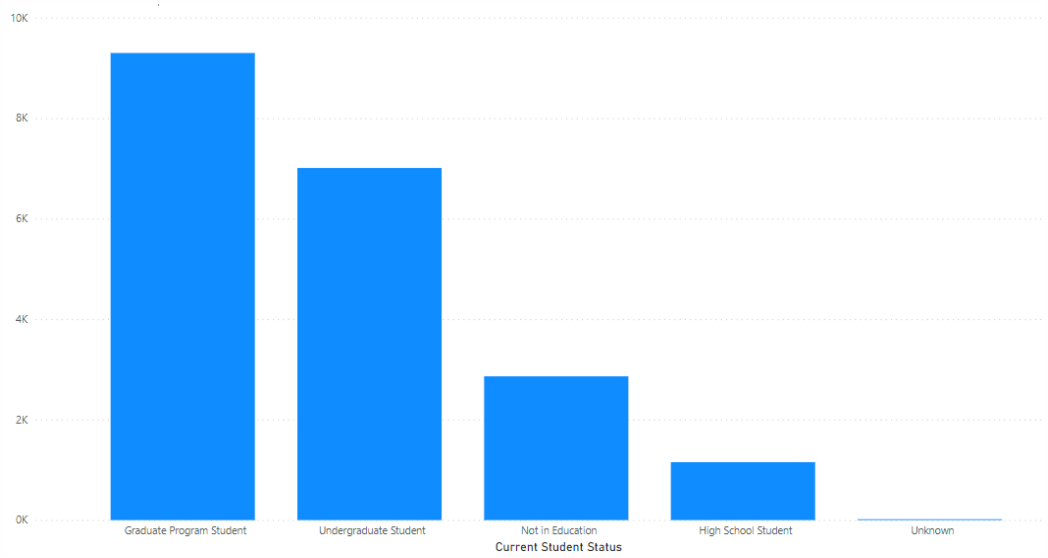
**d. Country**

| | |
|---|---|
| India | 9131 |
| United states | 3456 |
| Nigeria | 3351 |
| Pakistan | 1053 |
| Ghana | 898 |
| Egypt | 444 |
| Others | 1589 |



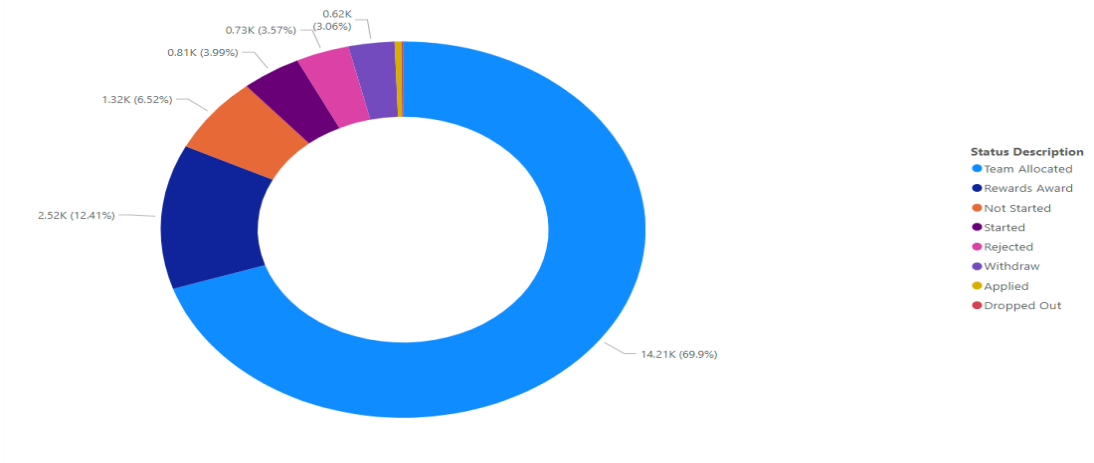

**e. Current Student Status**

| | |
|---|---|
| Graduate | 9297 |
| Undergraduate | 7009 |
| High school | 1153 |
| Not in education | 2862 |
| unknown | 1 |



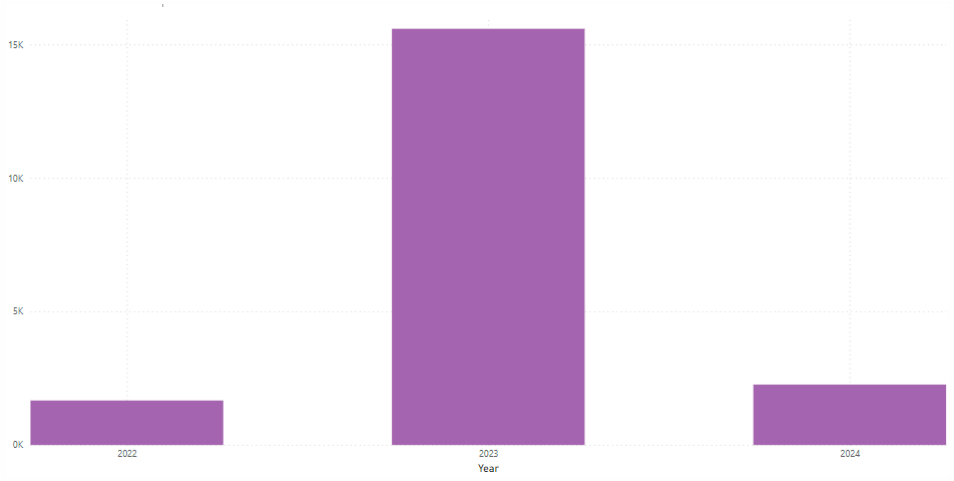

### f. Status Description

| | |
|---|---|
| Teams allocated | 14206 |
| Rewards award | 2521 |
| Not started | 1324 |
| Started | 810 |
| Rejected | 726 |
| Withdraw | 622 |
| Applied | 89 |
| Dropped out | 24 |



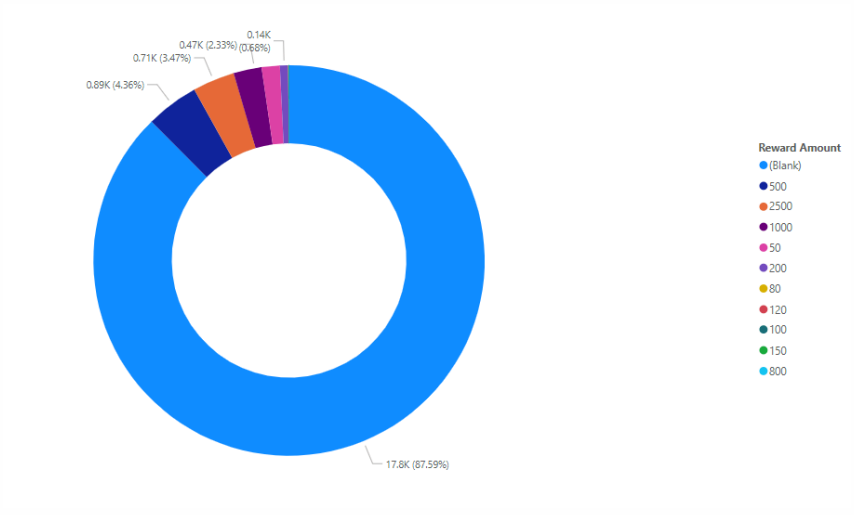
**g. Opportunity Start Date**

| | |
|---|---|
| 2022 | 1666 |
| 2023 | 15589 |
| 2024 | 2263 |




**h. Reward Amount**

| **Rewards** | |
|---|---|
| 500 | 886 |
| 2500 | 706 |
| 1000 | 473 |
| 50 | 305 |
| 200 | 138 |
| 120 | 8 |
| 80 | 6 |
| 800 | 1 |
| Blank | 17801 |


0.14K
0.47K (2.33%)
(0.68%)
0.71K (3.47%)
0.89K (4.36%)
17.8K (87.59%)
Reward Amount
(Blank)
500
2500
1000
50
200
80
120
100
150
800

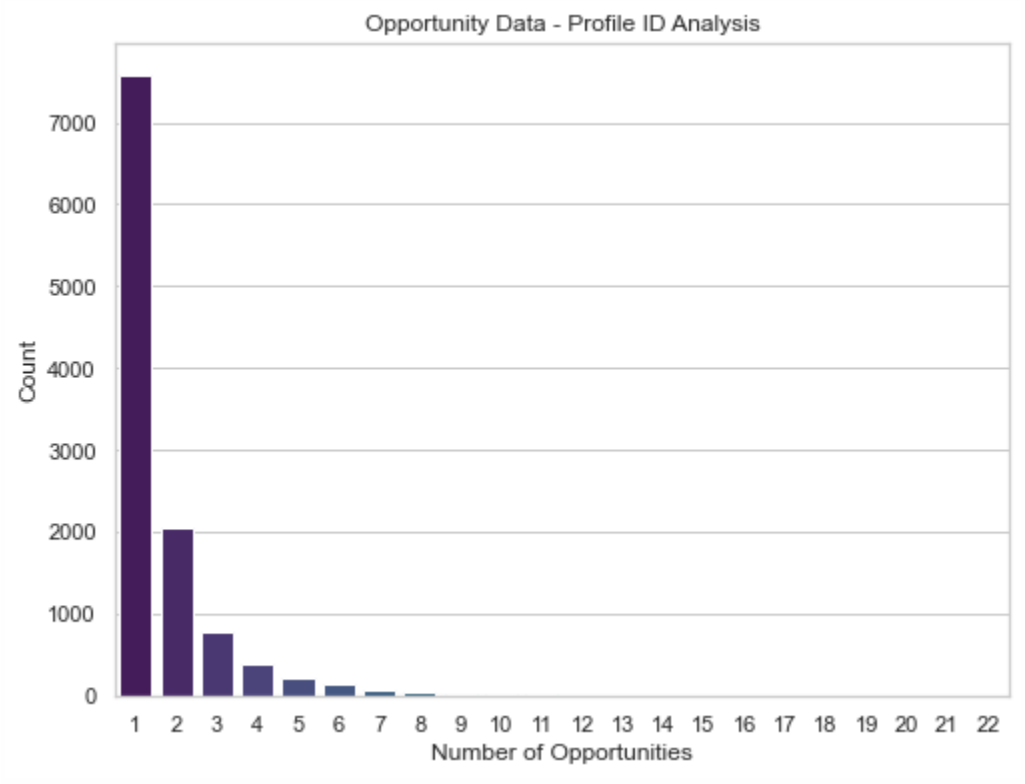
# 4. PROFILE ID ANALYSIS

## 4.1 User Data

User Data does not contain a 'Profile ID' column.

## 4.2 Opportunity Sign Up and Completion Data

| | |
|---|---|
| Total number of rows | 20322 |
| Number of unique Profile IDs | 11481 |
| Number of duplicate IDs | 8841 |
| Number of missing Profile IDs | 0 |



# 5. OPPORTUNITY STATUS ANALYSIS

In the context of opportunity management, the "Status Description" column typically represents the different stages or states an opportunity can be in during its lifecycle. Each status reflects the progress or outcome of the opportunity. Understanding the distribution of these statuses is crucial for assessing the overall performance and efficiency of the opportunity management process.

**1. Team Allocated:** Opportunities in this status have had teams assigned to them. This could indicate that the necessary resources or personnel have been identified and are working on the opportunity.

**2. Rewards Award**: This status suggests that the opportunity has been successfully completed or achieved, and rewards or recognitions have been granted. It signifies a positive outcome.

**3. Not Started**: Opportunities in this status have not yet commenced. They may be in the planning phase or awaiting initiation.

**4. Rejected:** Opportunities that have been rejected were likely considered but were not deemed feasible or suitable, resulting in a decision not to pursue them further.

**5. Started:** Opportunities in progress fall into this category. They have been initiated and are currently undergoing implementation or execution.

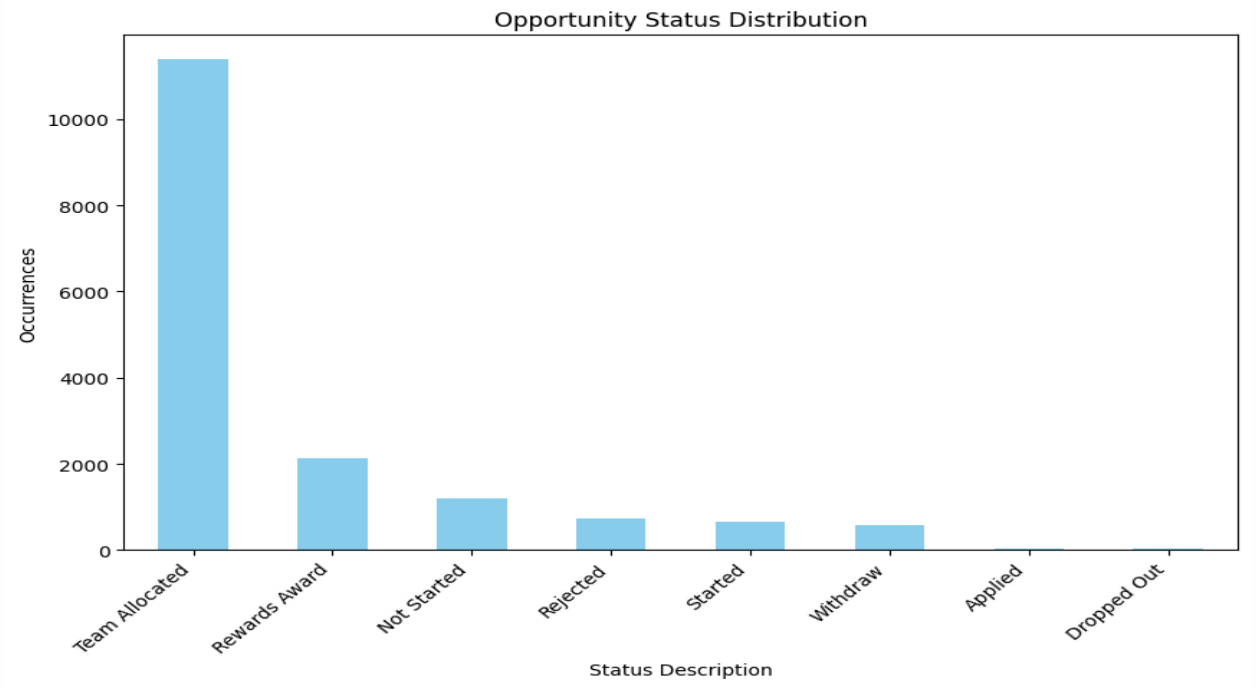
**6. Withdraw:** Opportunities that were initially considered but later withdrawn or canceled for various reasons are categorized as "Withdraw."

**7. Applied:** This status may represent opportunities where an application process is involved, and applicants have submitted their interest or proposals.

**8. Dropped Out:** Opportunities that were started but later abandoned or discontinued are classified as "Dropped Out."

**Status Distribution Summary:**

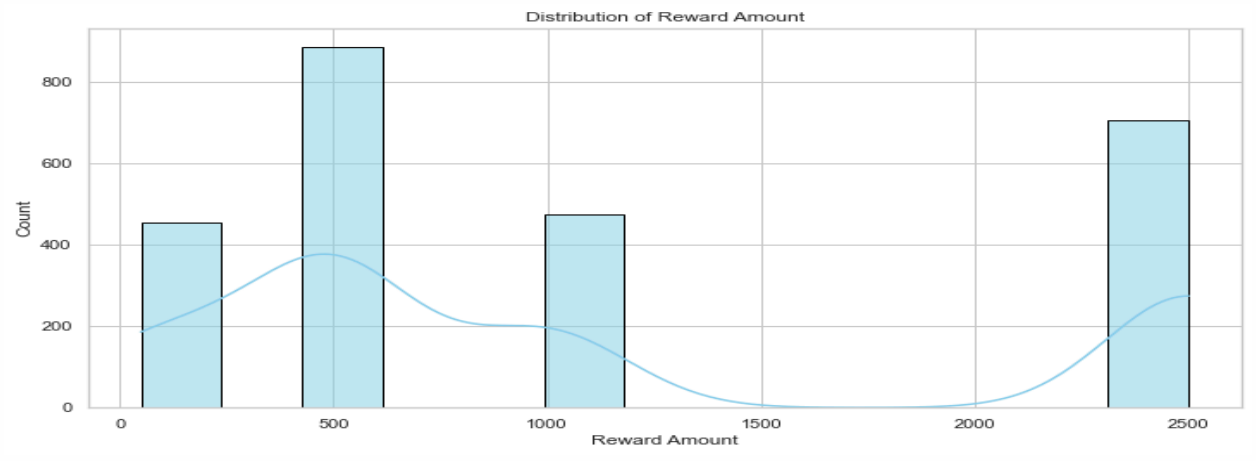
| | |
|---|---|
| Teams Allocated | 11384 |
| Rewards Awards | 2136 |
| Not Started | 1195 |
| Rejected | 726 |
| Started | 657 |
| Withdraw | 575 |
| Applied | 34 |
| Dropped Out | 24 |

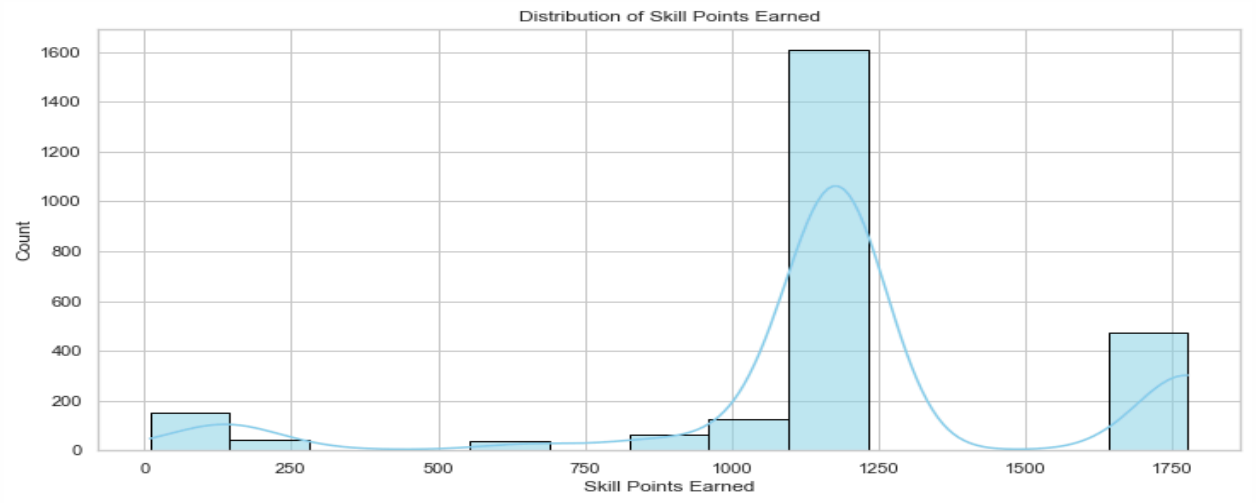



This distribution summary provides a quantitative overview of the occurrence of each status in the "Status Description" column. It helps stakeholders and decision-makers understand the proportion of opportunities at different stages, identify potential bottlenecks, and assess the overall health of the opportunity management process. Adjustments and actions can then be taken based on these insights to optimize the management of opportunities and improve outcomes.

# 6. BASIC STATISTICS

| | Reward Amount | Skill Points Earned |
|---|---|---|
| **count** | 2521.000000 | 2521.000000 |
| **mean** | 1081.261404 | 1186.964697 |
| **std** | 927.251398 | 399.172150 |
| **min** | 50.000000 | 10.000000 |
| **25%** | 500.000000 | 1182.000000 |
| **50%** | 500.000000 | 1182.000000 |
| **75%** | 2500.000000 | 1182.000000 |
| **max** | 2500.000000 | 1776.000000 |



# 7. INITIAL OBSERVATIONS

**Initial Observations or Patterns:**

**1. Profile ID Duplicates:** In the Opportunity Sign Up and Completion Data, there are 8,841 duplicate Profile IDs, suggesting that some learners have engaged with multiple opportunities. Further exploration is needed to understand the implications of these duplicates.

**2. Opportunity Status Distribution:** Teams Allocated and Rewards Award are the most frequent statuses in the Opportunity Status Distribution, indicating successful engagement in opportunities. Understanding the factors contributing to these outcomes could provide valuable insights.

**3. Challenges Faced:** Identified challenges include missing information, data inconsistencies, outliers, and limited context. Addressing these challenges is crucial for ensuring the reliability and accuracy of the analysis.

**Areas of Interest for Deeper Investigation:**

**1. Duplicate Profile IDs:** Investigated the reasons for duplicate Profile IDs in the Opportunity Sign Up and Completion Data. Examine whether learners participating in multiple opportunities contribute differently to the success metrics.

**2. Factors Influencing Success:** Explored the factors contributing to successful outcomes (Teams Allocated, Rewards Award) in opportunities. Analyze the characteristics of learners and opportunities associated with positive results.

**3. Data Inconsistencies:** Addressed data inconsistencies in both datasets, focusing on standardizing formats and structures. Understand the impact of these inconsistencies on the analysis and decision-making.

**4. Missing Information:** Investigated the impact of missing information in User Data, especially in columns like Gender and Degree. Assess whether imputing missing values or exploring patterns in missing data is necessary.

**5. Trend Analysis:** Conducted trend analysis on relevant variables to identify patterns and changes over time. This can provide insights into the evolution of user behavior and opportunity engagement.

**6. Communication Complexity:** Addressed the challenge of communication complexity by refining visualizations to effectively convey findings to diverse stakeholders.

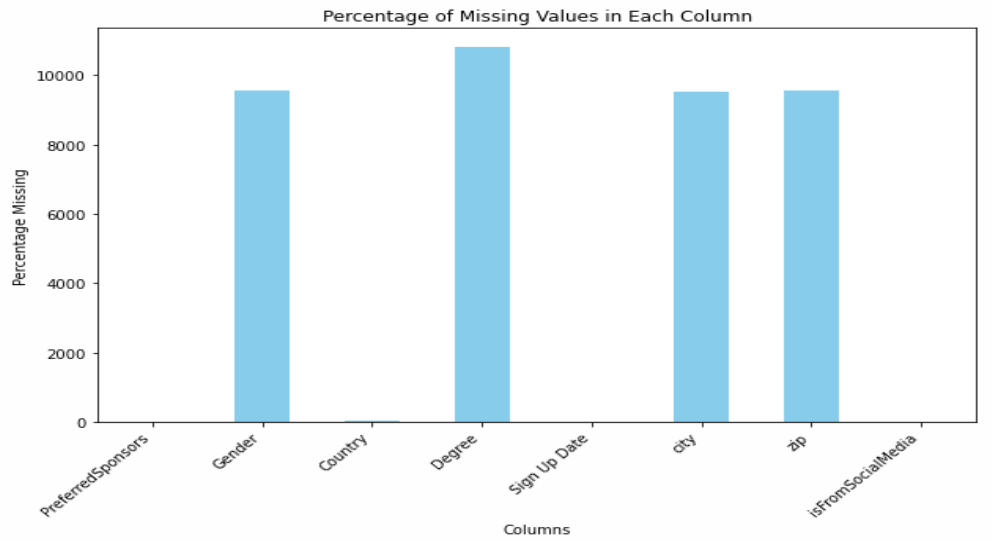
These areas of interest will guide deeper investigation and contribute to a more comprehensive understanding of the datasets, user behavior, and opportunity success factors.

# 8. CHALLENGES FACED

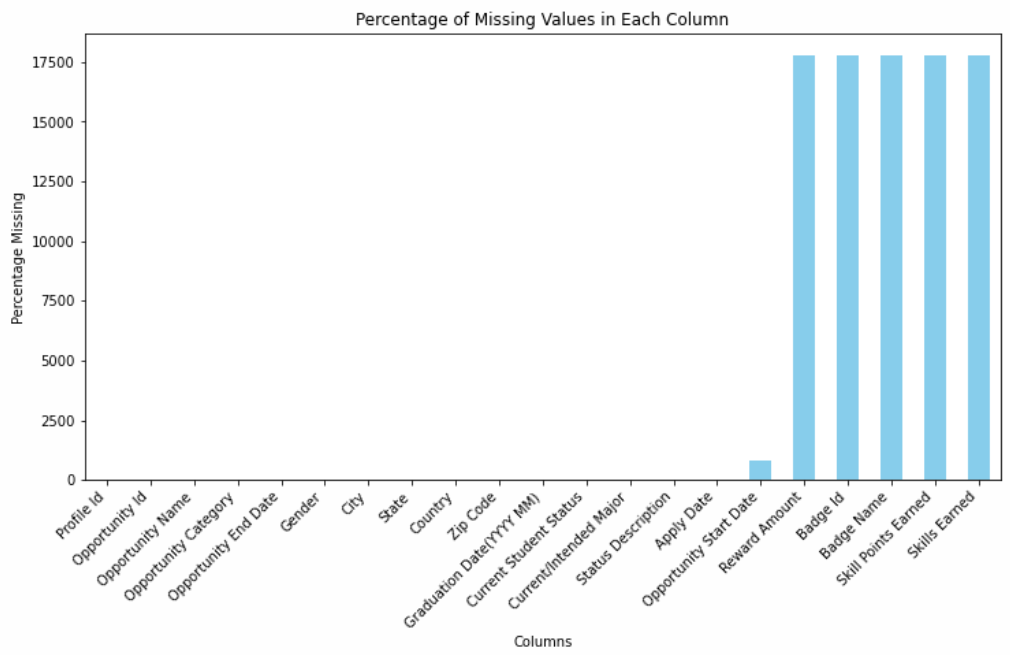
- **Missing Information:**

  Understanding the dataset fully can be challenging when there are gaps or missing pieces of information. Incomplete records or variables with missing values may hinder a comprehensive analysis.

  - **User Data**




  - **Opportunity Sign Up and Completion Data**



- **Data Inconsistencies:**

  Inconsistencies in the format, structure, or coding of variables can complicate the EDA process. Handling variations in data representation can be time-consuming and may require standardization.

- **Outliers and Anomalies:**

  The presence of outliers or anomalies can impact the statistical measures and distort the overall interpretation of the dataset. Identifying and deciding how to handle extreme values is a common challenge.

- **Limited Context:**

  EDA reports may lack sufficient context about the variables or the data generation process. This lack of background knowledge can make it challenging to draw meaningful insights or make informed decisions.

- **Data Quality Issues:**

  Unforeseen issues related to data quality, such as duplicates, inaccuracies, or discrepancies, can create uncertainty and pose challenges during the analysis. Addressing these issues may require collaboration with data owners or domain experts.

- **Domain Knowledge Gaps:**

  Lack of domain-specific knowledge can hinder the interpretation of variables and relationships within the dataset. Understanding the context of the data is crucial for accurate analysis and meaningful insights.

- **Unclear Variable Definitions:**

  Ambiguous or poorly defined variables can lead to misinterpretation. When variable meanings are unclear, it becomes challenging to draw accurate conclusions or make informed decisions based on the data.

- **Scale and Unit Differences:**

  Differences in scales or units across variables may pose challenges in comparing and interpreting data. Addressing these discrepancies is essential for a more accurate understanding of the relationships within the dataset.

- **Time Constraints:**

  Limited time for conducting a thorough EDA may restrict the depth of analysis. Balancing the need for a comprehensive exploration with time constraints is a common challenge.

- **Communication Complexity:**

  Presenting the findings of the EDA in a clear and accessible manner for diverse stakeholders with varying levels of technical expertise can be challenging. Ensuring effective communication is essential for the report's impact

# 9. NEXT STEPS

I. **Data preprocessing:** Data preprocessing is a crucial step in the data analysis pipeline that involves cleaning and transforming raw data into a format suitable for analysis and modeling. It encompasses various tasks aimed at improving the quality and structure of the dataset, making it more conducive to extracting meaningful insights. Key aspects of data preprocessing include handling outliers and anomalies, normalization or scaling of features, and addressing issues related to data quality.

II. **Trend analysis**: Trend analysis involves examining data points over time to identify patterns, tendencies, or changes in a particular variable. This is crucial for understanding the direction and magnitude of changes and making predictions based on historical data. In the context of dashboard creation, trend analysis helps visualize and communicate how certain metrics or aspects of the data evolve over time.