Data Visualization Associate Internship
Group 2D
Week-1 Submission
Team coordinator: Madhesh

1. Introduction

In the sphere of data-driven decision-making, deriving valuable insights relies on a thorough examination and understanding of datasets. This Exploratory Data Analysis (EDA) report aims to highlight key characteristics, reveal recurring patterns, and address complexities. This foundational work not only sets the stage for further data processing and visualization efforts but also plays a crucial role in enabling informed decision-making. The importance of this EDA report is underscored by its essential role in advancing Excelerate's mission to improve user insights and overall user experience, ultimately aiding in the development of advanced analytical dashboards.

2. Data Overview

This analysis focuses on "User Data," which consists of 8 columns and 27,563 rows. The columns are "PreferredSponsors," "Gender," "Country," "Degree," "Sign Up Date," "City," "Zip," and "isFromSocialMedia." This dataset includes non-identifying information about Excelerate users, providing insights into the diverse user base. Each row represents an individual user, offering a comprehensive overview. The details provided cover user preferences, demographics, geographic information, registration dates, and social media engagement indicators, creating a valuable resource for understanding and enhancing the platform's user experience.

2.1 User data overview:

The dataset provides information on 27,563 individuals registered on Excelerate, offering a comprehensive view of the user base. It includes eight columns with details such as PreferredSponsors, indicating user choices for opportunity recommendations, and Gender, which may be unspecified. The dataset also captures user locations and academic levels through Country, Degree, City, and Zip. The sign-up date records when accounts were established, and isFromSocialMedia denotes sign-ups via Google Login. There are 94 PreferredSponsors, 170 countries, and 4,015 cities represented. The degrees include Undergraduate, High School, Graduate Program, Not in Education, and null, with respective counts. Google Login attracted 13,809 users, while manual sign-up drew 13,753 users. This extensive demographic data enables detailed analyses for tailored opportunities and effective user engagement strategies on Excelerate. It provides insights into user characteristics for informed decision-making. The dataset's depth supports the creation of precise and impactful initiatives, considering diverse geographic and academic contexts, thus enhancing Excelerate platform strategies.

2.2 Opportunity data overview:

Our investigation focuses on the "Opportunity Sign Up Data" dataset, which provides a detailed overview of user engagement on the Excelerate platform. This dataset contains 20,322 rows and 21 columns, each representing a unique interaction with a variety of numerical, categorical, and datetime variables. The "Profile ID" is a crucial alphanumeric identifier that allows us to track individual learner journeys across multiple opportunities.

Since users can register for multiple opportunities, the dataset includes repeated instances of certain "Profile IDs." Additionally, the "Opportunity ID," a unique alphanumeric identifier for each opportunity, helps categorize user engagements and serves as a reference during backend processes, with 33 distinct opportunities recorded. This dataset offers a rich and varied landscape for exploration, enabling us to uncover patterns, address data inconsistencies, and extract valuable insights. Our goal is to enhance user experiences on the Excelerate platform by leveraging identifiers like "Profile ID" and "Opportunity ID" to understand individual user journeys and differentiate between opportunities. Through this analysis, we aim to contribute to the overall objective of optimizing user interactions on the Excelerate platform.

3. Column Analysis

3.1 Column Analysis (User Data)

Preferred Sponsors:

Data Type: Text
Description: This column displays the diverse sponsors chosen by learners upon joining the Excelerate Platform. Here, learners have the flexibility to select sponsors of their choice, enabling them to receive tailored opportunities. They have the option to choose one or multiple sponsors based on their preferences for personalized engagement and relevant opportunities.

Gender:
Data Type: Categorical
Description: This column displays the voluntarily provided gender information from users during the sign-up process. It is not a compulsory field, allowing users the option to choose whether or not to disclose their gender.

Country:
Data Type: Text
Description: This column shows the country in which the learner has indicated they live upon sign-up.

Degree
Data Type: Categorical
Description: This column shows the academic level indicated by the user upon sign up. This is not a mandatory field for signing up.

Sign Up Date
Data Type: Date
Description: This column shows the date on which they created their Excelerate account.

City
Data Type: Text
Description: This column displays the city voluntarily provided by the learner during sign-up. It is optional, not mandatory.

Zip
Data Type: Text
Description: This column shows the zip code of the city in which the learner has indicated they live upon sign-up. This is not a mandatory field for signing up.

isFromSocialMedia Data
Type: Boolean
Description: This column indicates the method of learner registration. A value of True signifies sign-up through Google Login, while False indicates manual sign-up without using social media credentials.

3.2 Data Types and Potential Issues (Opportunity Sign Up Data)

Profile ID (Alphanumeric):
Data Type: Alphanumeric
Unique Identifier: Yes
No missing values were identified.

Opportunity ID (Alphanumeric):
Data Type: Alphanumeric
Unique Identifier: Yes
No missing values were identified.

Opportunity Name (Categorical):
Data Type: Categorical
No missing values were identified.
Unique Values: 33 opportunities with varying frequencies.

Opportunity Category (Categorical):
Data Type: Categorical
No missing values were identified.
Unique Values: Event, Course, Competition, Internship, Engagement.

Opportunity End Date (Datetime):
Data Type: Datetime
No missing values were identified.
Dates appear to follow a standardized format.

Gender (Categorical): Data
Type: Categorical
One missing value was identified.
Unique Values: Male, Female, Don't want to specify, Other.

3.3 Categorical Column Summaries

Gender:
Male: 60.23% (12,240)
Female: 39.39% (8,004)
Other categories: Don't want to specify, Other.

Current Student Status:
Graduate Program Student: 9,297 (45.75%)
High School Student, Undergraduate Student, Not in Education.

Opportunity Category:
Internship: 71.13%
Course: 11%
Event: 9.79%
Other categories.

Opportunity Name:
Data Visualization: 31.71%
Money Matters: A Personal Finance Workshop and others.

Status Description:
Team Allocated: 69.90%
Dropped Out, Rejected, Applied, and others.
Badge Name:
Unknown: 84.66%
Null: 7.07%
Data Visualization Internship Completed: 1.94%
Data Visualization Internship Star Performer: 1.34
Project Management: 1.07%
The other categories take up the remaining percentage.
Applied Date Sign Ups:
June: 23.73%

The months of June, July, and August are the top 3 months with the highest Applied Date Sign ups, ranging between 4823 and 3503.
December has the least Applied Date sign-ups of 131.

4. Profile ID Analysis

In this section, we delve into the analysis of the "Profile ID" column, examining the uniqueness of Profile IDs, and identifying instances of duplicates or missing values.

4.1 Uniqueness of Profile IDs (Opportunity Sign-up Data)

The dataset contains a total of 20,322 Profile IDs.
The number of unique Profile IDs is 11,481.

This indicates that there are duplicate Profile IDs in the dataset, as the number of unique Profile IDs is less than the total number of Profile IDs.

4.2 Duplicate Profile IDs (Opportunity Sign-up Data)

The number of Duplicate Profile IDs is 3,898.
Some Profile IDs appear multiple times, with the highest number of duplicates for a single Profile ID being 22 occurrences and 21 occurrences being the least. Some most frequent duplicates with 22 occurrences are :

- “18e1e6bc-fada-4b09-bf52-ab45daf318f4”
- “c2245f7e-2e9d-42c9-b5af-550be9eae1c8”

Followed by the least frequent duplicates with 21 occurrences such as

- “f8ee2854-73b4-4e75-9a11-62e4e9e52a5a”
- “19ce6f4c-215c-412e-9e29-f09d6f5a4422”
- “d01a98bd-90af-4314-91fc-f31e0a0b6b5d”

4.3 Number of Missing Profile IDs (Opportunity Sign-up Data)

There are no missing Profile IDs in the dataset.

Key Observations:

The dataset contains a high number of duplicate Profile IDs, suggesting that there may be repeated entries for some individuals. No missing Profile IDs were found, ensuring that all entries have associated Profile IDs.The presence of duplicates needs to be addressed to ensure accurate analysis and reporting.

5. Opportunity Status Distribution

1. High Allocation to Teams:

- The majority of opportunities (14,206) are in the "Team Allocated" status. This suggests that many participants are grouped into teams for their activities.

2. Rewards Awarded:

- A significant number of opportunities (2,521) have reached the "Rewards Award" status, indicating a successful completion and reward distribution for these opportunities.

3. Not Started Status:

- A notable number of opportunities (1,324) are in the "Not Started" status. This could imply that a substantial portion of participants haven't yet begun their assigned tasks.

4. Engagement Level:

- The "Started" status has 810 occurrences, showing a moderate level of initial engagement by participants in their tasks.

5. Rejections and Withdrawals:

- There are 726 opportunities marked as "Rejected" and 622 as "Withdraw," highlighting potential areas where participants may face challenges or lose interest.

6. Low Application Rate:

- Only 89 opportunities are in the "Applied" status, indicating either a streamlined application process or that most applications progress quickly to other statuses.

7. Minimal Dropouts:

- The "Dropped Out" status has the lowest count at 24, suggesting that once participants start, few drop out entirely.

These findings can help in understanding the overall flow and engagement of participants in the opportunities provided, as well as identifying areas for improvement in participant retention and initiation.

6. Basic Statistics

The User Data dataset presents a challenge for statistical analysis due to the absence of numeric columns, which restricts the ability to perform quantitative assessments. This lack of numerical values impedes the exploration of relationships and patterns within the user data. Conversely, the following section examines the Opportunity Sign Up Data, where we calculate key statistics (mean, median, min, max) for relevant numeric columns, allowing for a more thorough statistical analysis. The dataset also provides numeric data for "Reward Amount" and "Skill Points Earned." The basic statistics for these columns are as follows:

6.1 Reward Amount (Opportunity Sign Up Data)

Rows	**Reward Amount (Opportunity Sign Up Data)**	**Skill Points Earned (Opportunity Sign Up Data)**
Mean:	1081.26	1186.96
Median:	500.00	1182.00
Minimum:	50.00	10.00
Maximum:	2500.00	1776.00
Standard Deviation:	483.39	399.172150
Range:	2501	
Sum:	2,722,238	
Count:		20,322

These statistics offer insights into the distribution and range of rewards and skill points earned by participants. The mean reward amount indicates a relatively high average reward, with a notable median suggesting that half of the rewards are below 500. Skill points earned show a high level of skills being recognized, with a tight clustering around the median of 1182 points.

7. Initial Observations

7.1 User Data

1. Data Structure:

The dataset comprises 8 columns and 27,563 rows.

2. Date Column:

The date column contains inaccuracies and inconsistencies in both format and data type, necessitating correction.

3. Data Diversity:

The dataset shows significant diversity in representation, with some countries represented by only one individual and others by over a thousand individuals.

7.2 Opportunity Sign-Up Data

During the exploratory analysis of the Opportunity Sign-Up Data, several initial observations and patterns emerged:

1. Status Distribution:

- The majority of participants are in the "Team Allocated" status, indicating successful team assignments.
- A notable number of participants did not receive rewards, as indicated by the low completion rate in the "Rewards Award" status.
- A significant portion of participants has either not started or been rejected.

2. Profile IDs:

- The dataset includes a diverse range of Profile IDs, reflecting varied user engagement with the Excelerate platform.
- Some Profile IDs appear multiple times, suggesting that certain users have registered for multiple opportunities.

3. Reward and Skill Points:

- The "Reward Amount" and "Skill Points Earned" columns show a wide range, with some participants receiving high rewards or earning a significant number of skill points.
- Negative values in the "Reward Amount" column indicate cases where participants were withdrawn, dropped out or rejected.

4. Badge Information:

- The "Badge ID" column includes repetitions, indicating instances where multiple participants earned the same type of badge.

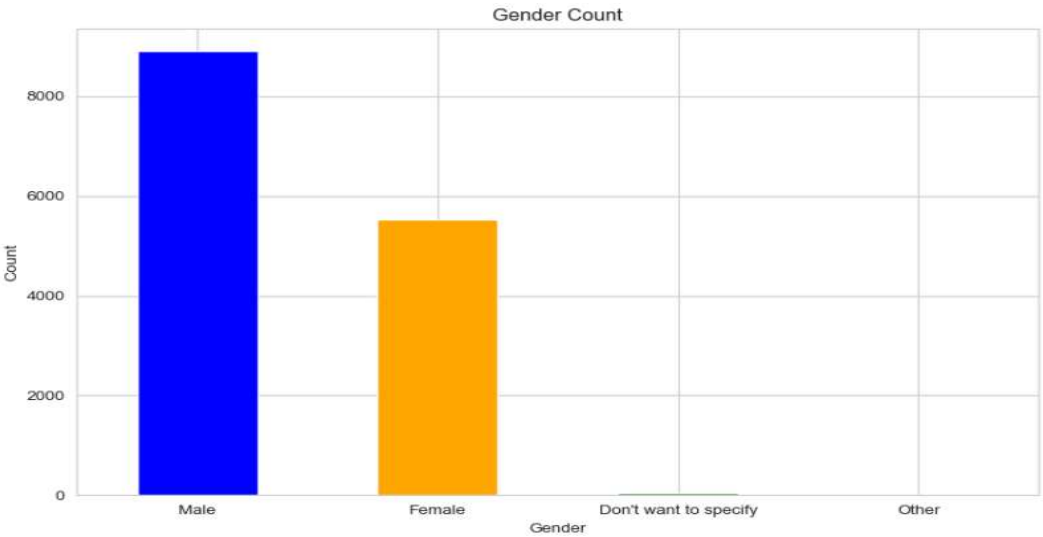
5. Missing Values:

- Missing values were observed in various columns and were handled differently based on the nature of the data and the participants' status descriptions.

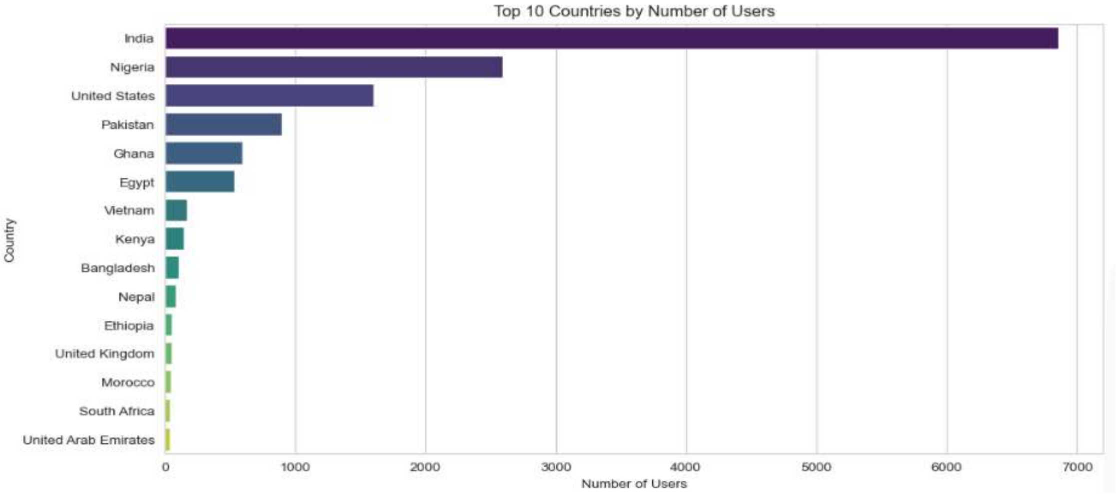
8. Visualisations

User data :

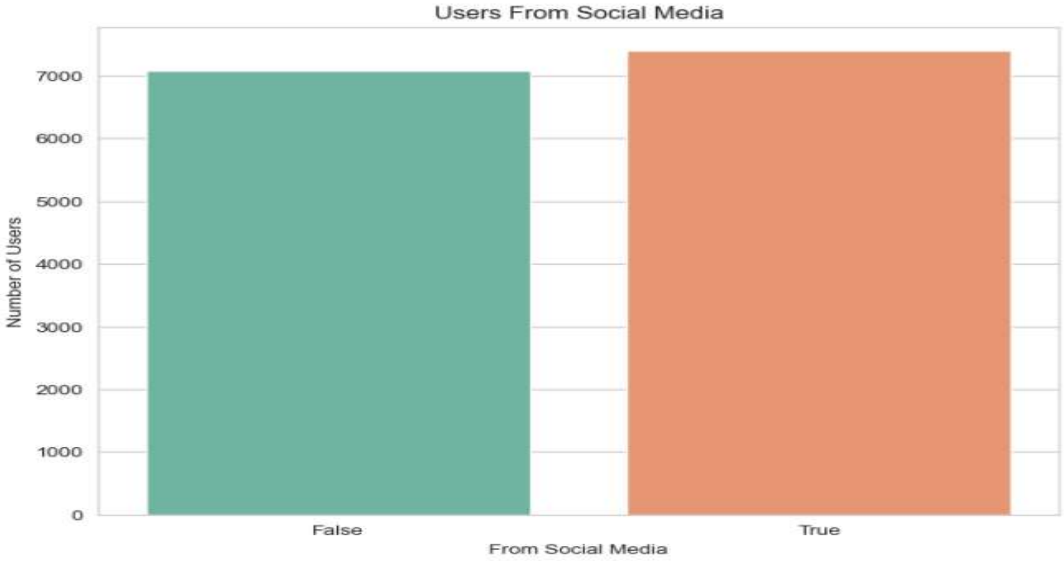
1. **Gender count**



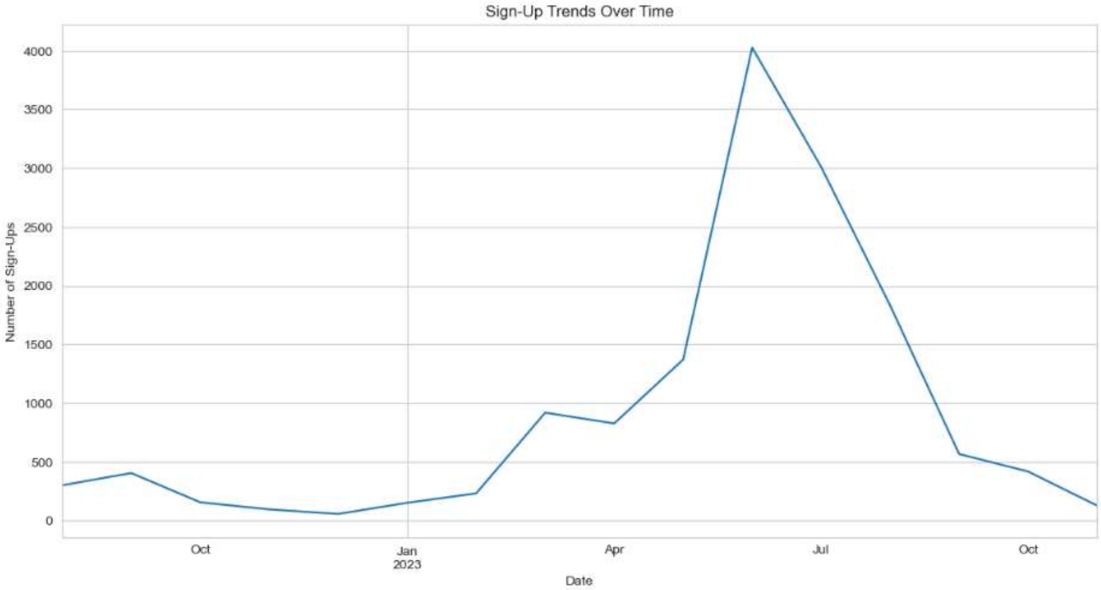
2. **Top 10 countries by number of users**



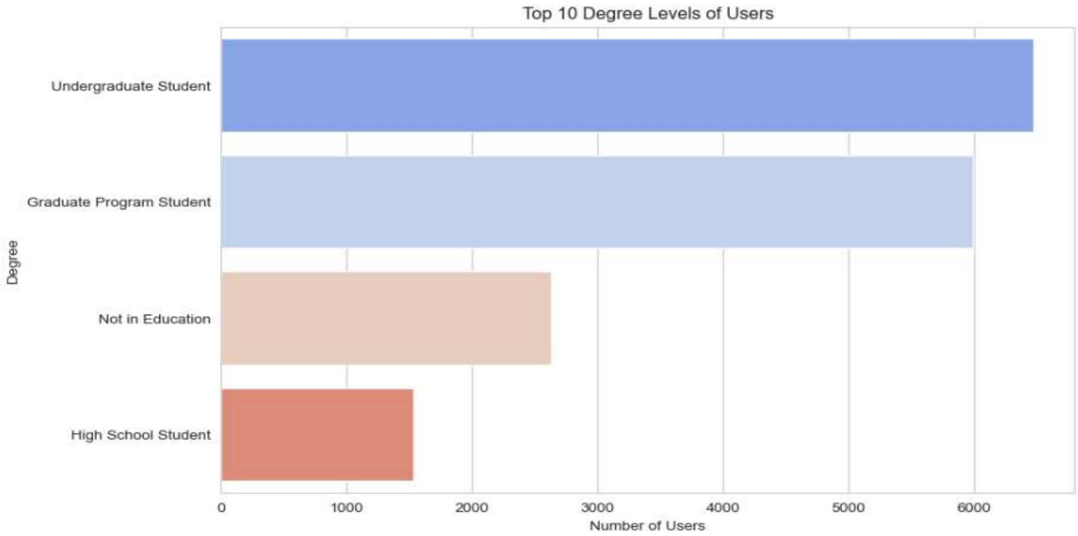
3. Users from Social Media



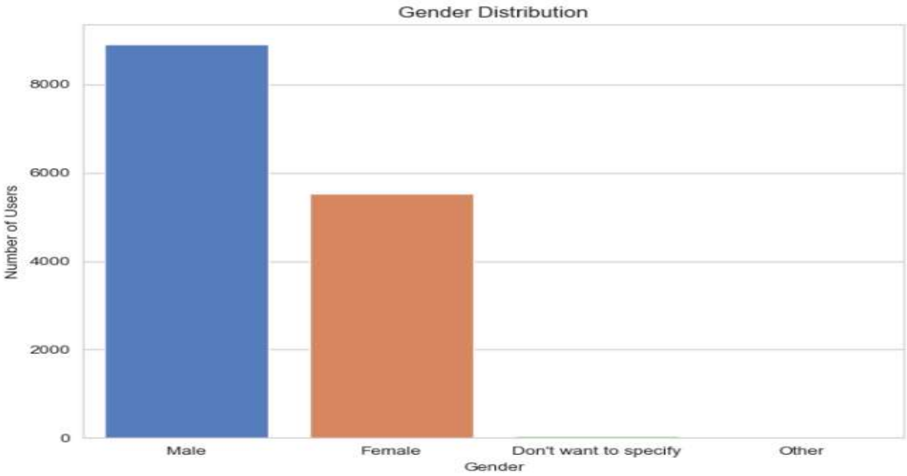
4. Sign up trends Over time



5. **Top 10 Degree level of users**

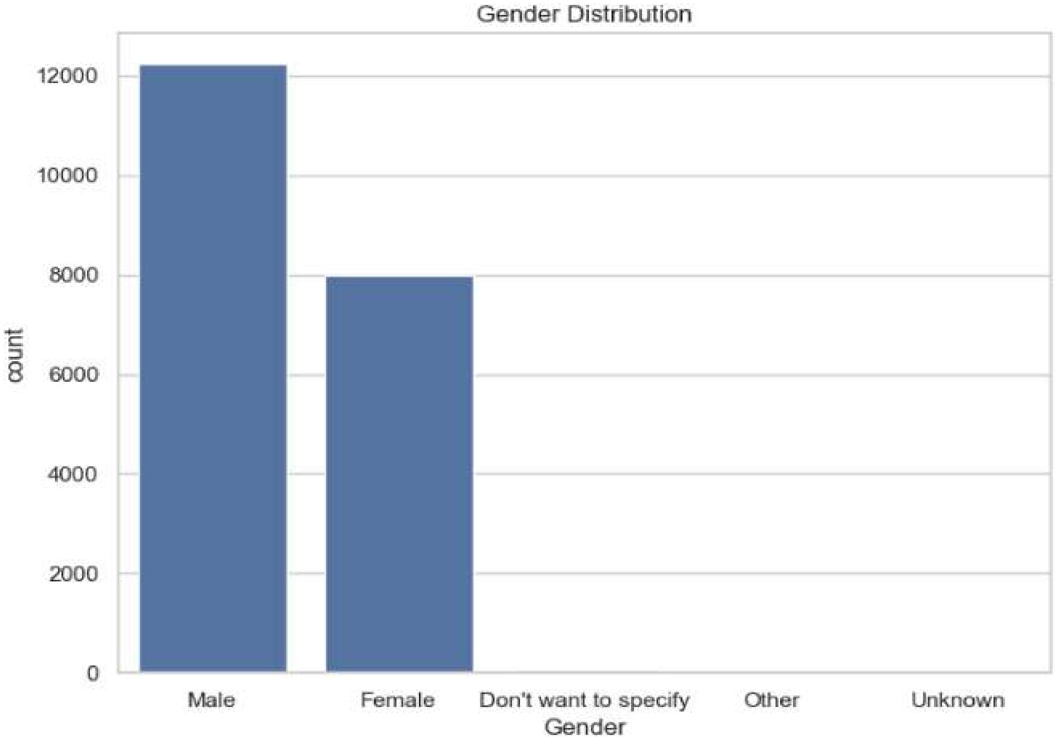


6. **Gender distribution**

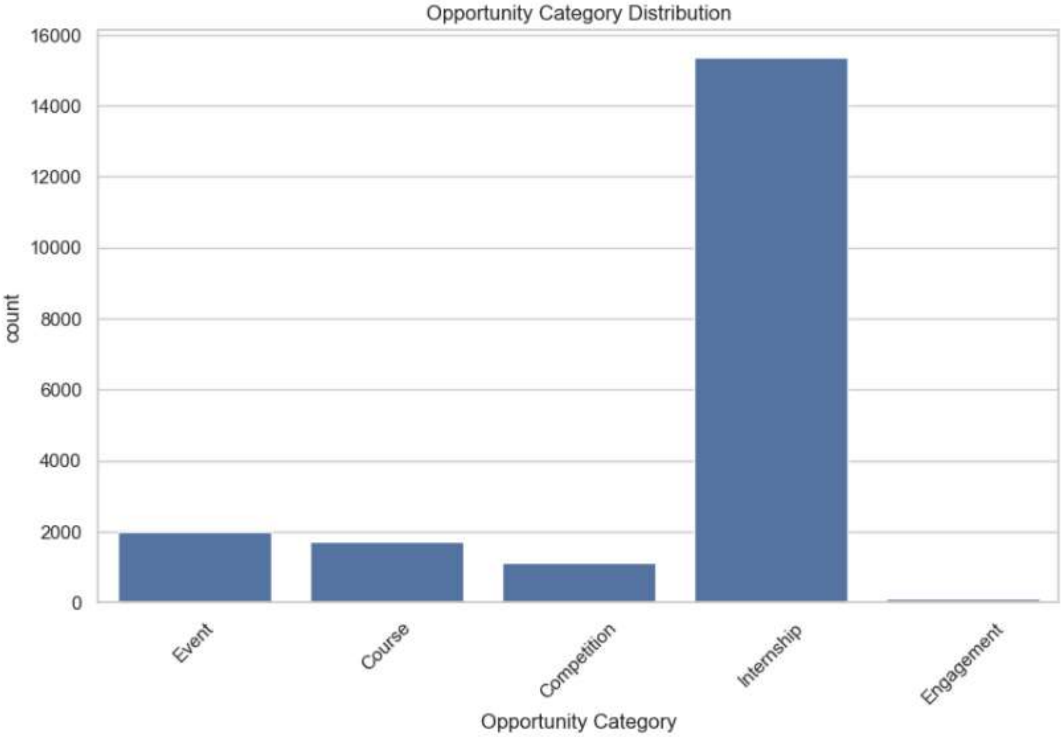


Opportunity wise data:

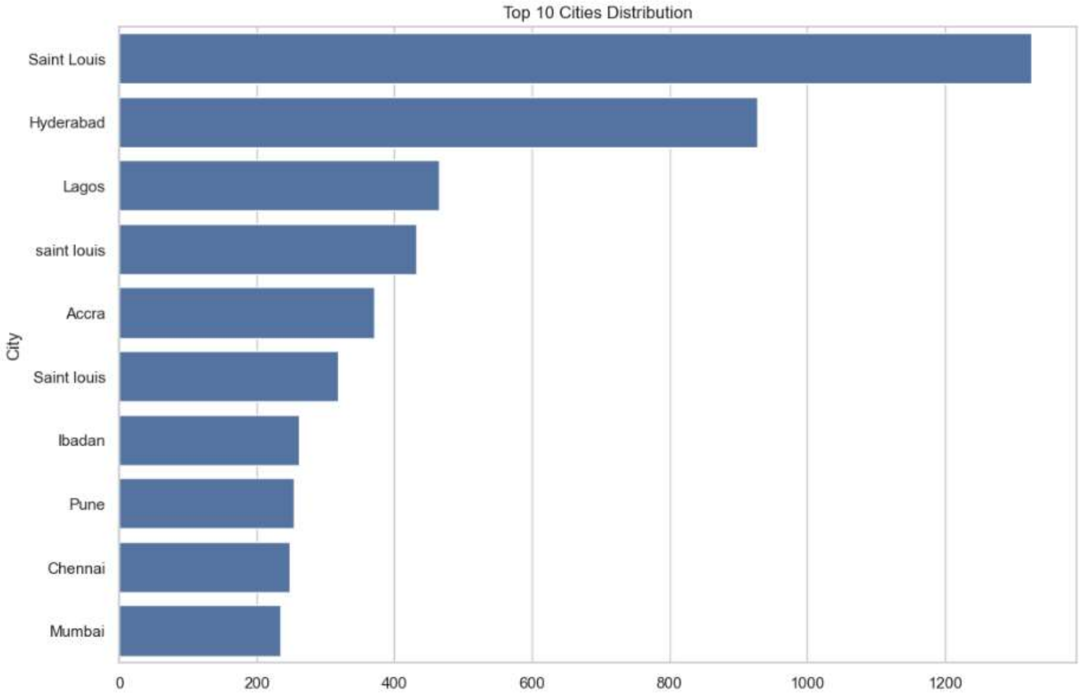
a. Gender analysis



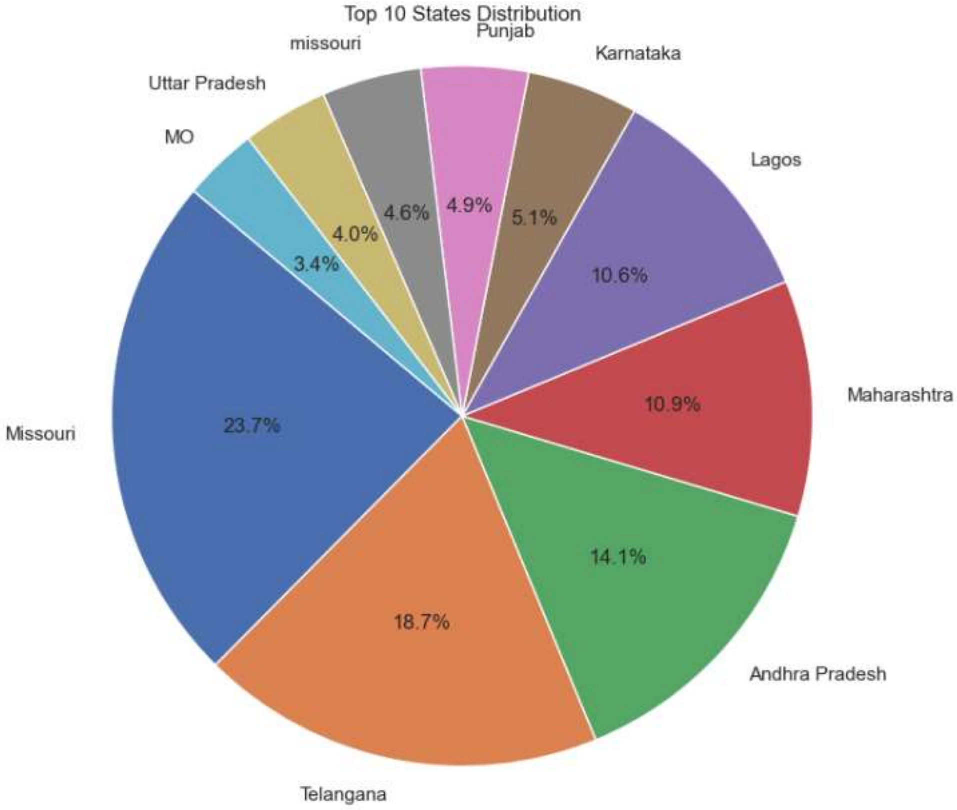
b. Opportunity category distribution



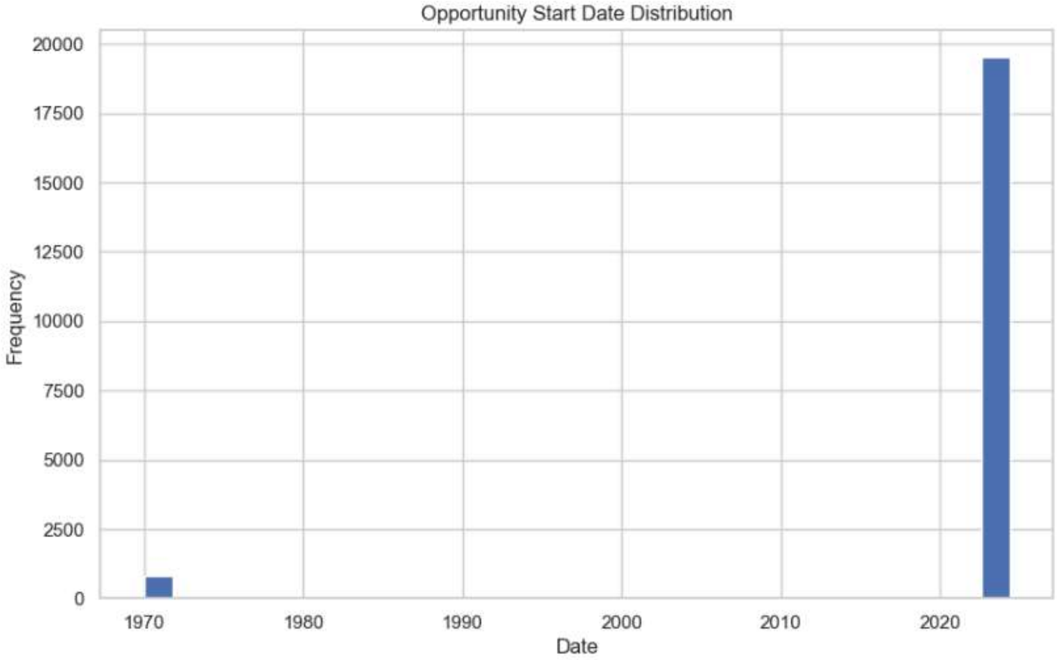
c. Top 10 cities



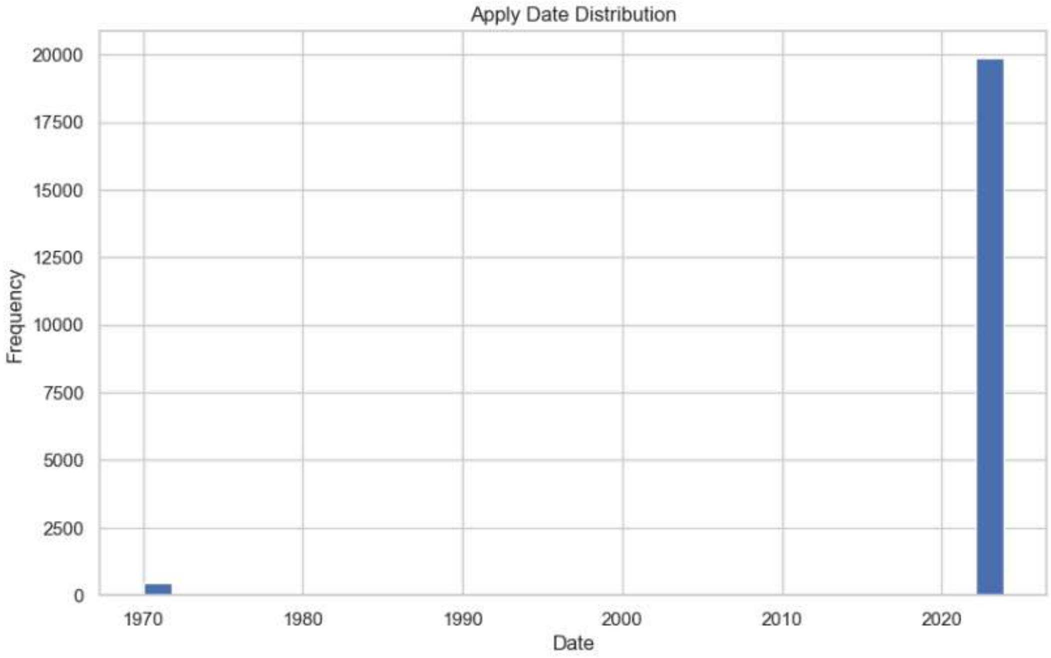
d. Top 10 states distribution



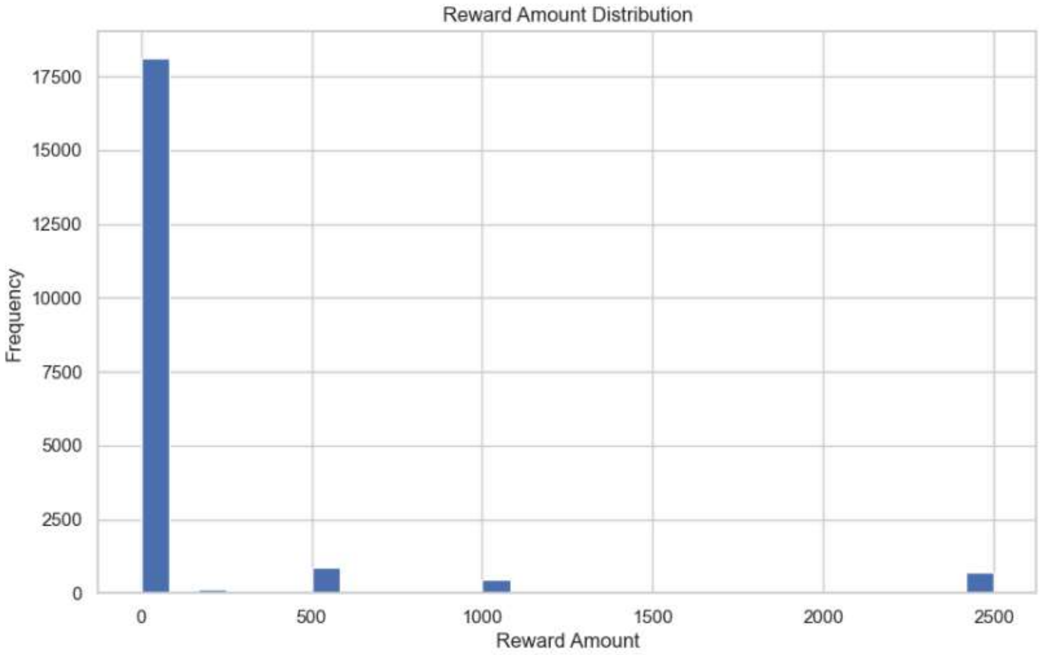
e. Opportunity start date distribution



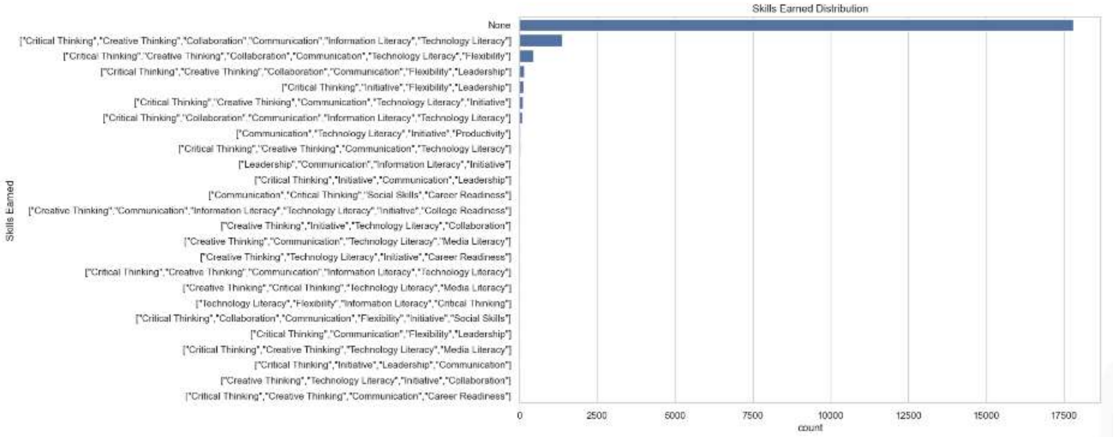
f. Apply Data distribution



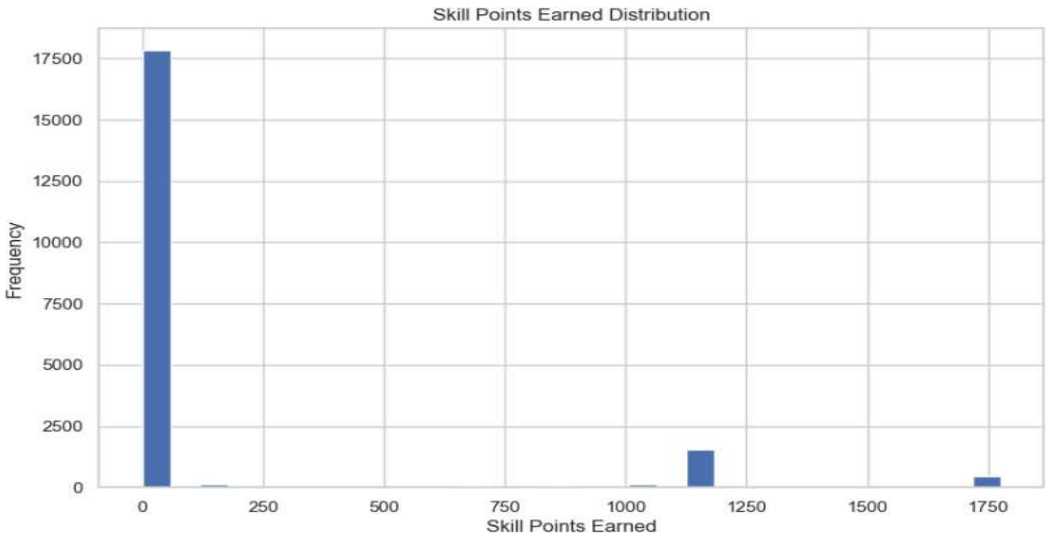
g. Reward Amount distribution



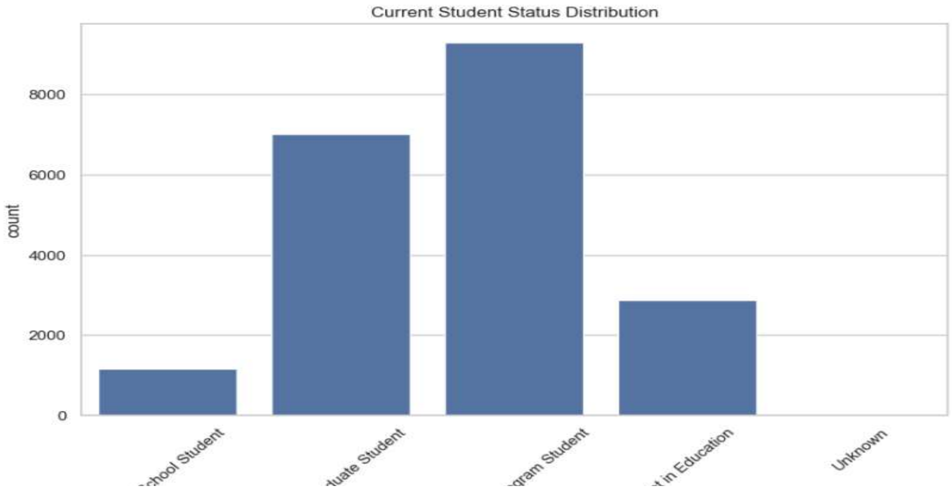
h. Skills Earned distribution



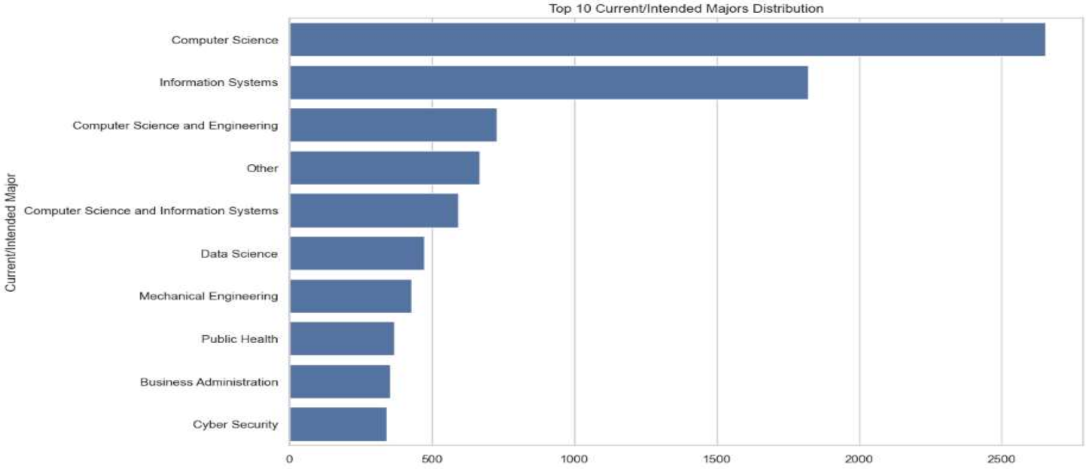
i. Skills point earned distribution



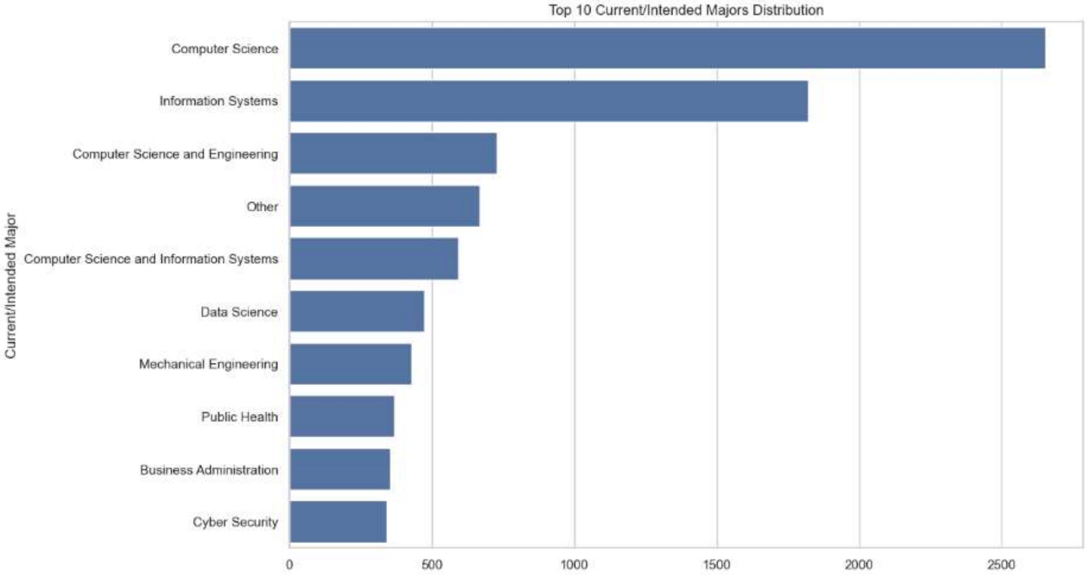
j. Current student status distribution



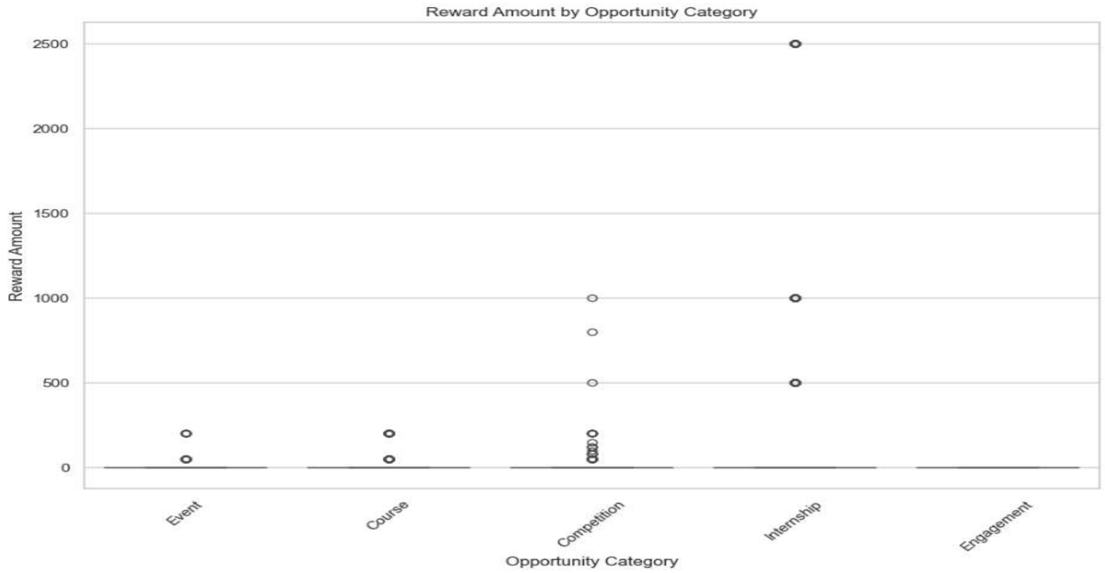
k. Top 10 current/intended distribution



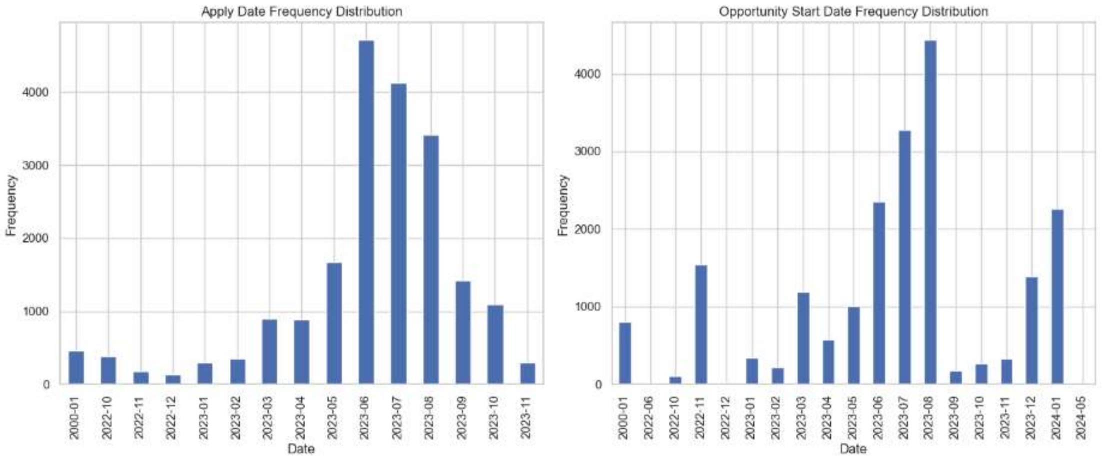
l. Opportunity end date distribution



m. Reward amount by category



n. Frequency distribution



9. Challenges Faced

Data Validation: Some countries in the dataset are represented by only one or two individuals, which are considered outliers and suggest inaccurate data.

Incomplete and Unclear Dataset: At least five out of eight columns contain null, blank, or missing values, affecting the accuracy of the data analysis.

Continuous Data Monitoring: Data quality can deteriorate over time due to factors like changes in data sources, updates in data collection methods, or evolving user behaviours. Thus, implementing continuous data monitoring and maintenance mechanisms is essential to ensure the dataset's ongoing reliability and relevance for future analysis.

10. Next Steps

Creating a comprehensive dashboard for Excelerate's leadership requires a systematic approach to address key questions. The goal is to develop a dashboard that offers insights into platform activity, global reach, engagement, opportunity popularity, completion trends, demographic analysis, skill development impact, and scholarship distribution. In the coming weeks, a thorough examination of feature analysis and data processing will take place, uncovering valuable insights that will be seamlessly incorporated into future documentation.