Exploratory Data Analysis Report

Team 2F

Project Coordinator

Ayushi Gupta

Contents

1. Introduction

In the context of data-driven decision making, valuable insights are derived primarily from the examination and comprehension of datasets. This report's Exploratory Data Analysis (EDA) goal is to shed light on the dataset's salient characteristics, reveal recurring themes, and handle subtleties, laying the groundwork for further data processing and visualization projects. This EDA report is crucial for promoting informed decision-making and easing the creation of further analytical dashboards, which is in line with Excelerate's objective to improve user insights and overall user experience.

1.1 Dataset Overview (User Data)

The focus of this analysis centers on the "User Data" Comprising 8 columns and 27,563 rows. The 8 columns which include “PreferredSponsors”, “Gender”, “Country”, “Degree”, “Sign Up Date”, “City”, “Zip” and “isFromSocialMedia”. This dataset encompasses non-identifying information about every user who has created an account on Excelerate. Each row represents a unique user giving an overview of the user base.

1.2 Dataset Overview (Opportunity Sign Up Data)

The focus of this analysis centers on the "Opportunity Sign Up Data." Comprising 21 columns and 20,322 rows, this dataset encapsulates a wealth of information regarding user interactions with various opportunities on the Excelerate platform. The diverse nature of the columns encompasses both categorical and numerical variables, providing a comprehensive view of user engagement. The columns include "Profile ID," "Opportunity ID," "Opportunity Name," "Opportunity Category," "Gender," "City," "State," "Country'" "Zip Code," "Graduation Date(YYYY MM)," "Current Student Status," "Current/Intended Major," "Status Description," "Apply Date," "Opportunity Start Date," "Reward Amount," "Badge Id," "Badge Name," "Skill Points Earned," and "Skills Earned". These columns collectively contribute to understanding user behaviors, preferences, and the outcomes of their participation in different opportunities.

This EDA report serves as a crucial milestone, enabling a meticulous examination of the dataset's intricacies. By employing descriptive statistics, visualizations, and a systematic cleaning and validation process, the report aims to ensure data integrity, completeness, and accuracy. The insights derived from this exploration will lay the groundwork for subsequent stages of analysis, contributing to the overarching goal of optimizing user experiences on the Excelerate platform.

Through this EDA, we strive to unravel the hidden narratives within the Opportunity Sign Up Data, empowering Excelerate with actionable insights and facilitating a data-driven approach to decision-making. The following sections will delve into the specifics of each column, addressing challenges encountered, and showcasing the meticulous steps taken to cleanse and validate the data. As we progress, the report will unfold a narrative that not only illuminates the dataset's intricacies but also informs strategic decisions for future analytical endeavors. data processing and visualization tasks. As a part of Excelerate's mission to enhance user insights and improve overall user experience, this EDA report is instrumental in driving informed decision-making and facilitating the development of subsequent analytical dashboards.

2. Data Overview

2.1 User Data Overview

The dataset contains non-identifying details of 27,563 individuals who have registered on Excelerate. Each row corresponds to a distinct user, providing a comprehensive overview of the user base. With eight columns, the dataset includes information such as PreferredSponsors, indicating the sponsors selected by learners for opportunity recommendations. The Gender column represents user-provided gender information, which might be absent for some users. The Country, Degree, City, and Zip columns denote the user's location and academic level at the time of registration. The Sign-up date column logs the date when users established their Excelerate accounts, while the isFromSocialMedia column indicates whether users signed up through Google Login. The dataset comprises 94 distinct PreferredSponsors, 170 unique countries, and 4015 unique cities. In the Degree column, there are five distinct values: Undergraduate Student, High School Student, Graduate Program Student, Not in Education, and null. Specifically, there are 6,527 Undergraduate Students, 1,562 High School Students, and 6,015 Graduate Program Students. Furthermore, 2,646 learners fall into the Not in Education category, and 1,442 are categorized as null. Regarding login methods, 13809 learners registered via Google Login, while 13,753 opted for manual signup. Abundant in demographic data, the dataset enables in-depth analyses of user preferences, geographic dispersion, and academic backgrounds. This wealth of information offers valuable insights that can be harnessed for tailored opportunities and the formulation of effective user engagement strategies on the Excelerate platform. The dataset's richness extends beyond mere statistical details, providing a nuanced understanding of user characteristics that can inform

decision-making processes. Furthermore, the geographic distribution and academic backgrounds captured in the dataset contribute to a holistic view, enabling the creation of engagement strategies that consider diverse user contexts. In essence, the dataset's depth and breadth serve as a valuable resource for enhancing the precision and effectiveness of initiatives on the Excelerate platform.

2.2 Opportunity Sign Up Data Overview

The "Opportunity Sign Up Data" forms the focal point of our explorations. This dataset encapsulates a comprehensive snapshot of user engagement with various opportunities on the Excelerate platform. Below is a high-level summary outlining key statistics and characteristics of this dataset:

Number of Rows: The dataset comprises a total of 20,322 rows, each representing a unique entry or interaction within the Excelerate platform. Each row provides insights into the engagement
of a learner with specific opportunities.

Number of Columns: The dataset is structured with 21 columns, each offering distinct information about the participants, opportunities, and their interactions. These columns encompass a mix of numerical, categorical, and datetime variables, providing a diverse set of attributes for analysis.

Unique Identifier (Profile ID): The "Profile ID" serves as the alphanumeric unique identifier for each learner. Given that users can register for multiple opportunities, it is expected to observe multiple instances of a particular "Profile ID" within the dataset. This identifier plays a crucial role in tracking individual user journeys and linking them to various opportunities over time.

Opportunity ID: The "Opportunity ID" is another alphanumeric identifier specific to each opportunity on the Excelerate platform. With 33 unique opportunities in the dataset, this column acts as a reference point for mapping opportunities during the backend processes. It provides a means to categorize user engagements based on the chosen opportunity.

The Opportunity Sign Up Data offers a rich and diverse landscape for exploration, comprising over 20,000 entries across 21 columns. The unique identifiers, such as "Profile ID" and "Opportunity ID," serve as crucial anchors for tracking individual user journeys and distinguishing between different opportunities. As we delve deeper into the dataset in the subsequent sections, we aim to uncover patterns, address data inconsistencies, and extract valuable insights that will contribute to the overarching goal of enhancing user experiences on the Excelerate platform.

3. Column Analysis

In this section, we conduct a detailed analysis of each column in the Opportunity Sign Up Data. We examine data types, identify potential issues such as missing values and outliers, and provide summaries for categorical columns.

3.1 Column Analysis (User Data)

Preferred Sponsors:

Data Type: Text

Description: This column shows the different sponsors selected by the learner who has signed up for the platform. On the Excelerate Platform, learners can choose their sponsors i.e, who they want to see opportunities from. Learners can choose one or more sponsors.

Gender:

Data Type: Categorical

Description: This column shows the gender indicated by the user upon sign up. This is not a mandatory field for signing up.

Country

Data Type: Text

Description: This column shows the country which the learner has indicated they live in 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 shows the city which the learner has indicated they live in upon sign up. This is not a mandatory field for signing up.

Zip

Data Type: Text

Description:This column 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.

isFromSocialMedia

Data Type: Boolean

Description:This column 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.

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

Profile ID (Alphanumeric):

Data Type: Alphanumeric

Unique Identifier: Yes

No missing values identified.

Opportunity ID (Alphanumeric):

Data Type: Alphanumeric

Unique Identifier: Yes

No missing values identified.

Opportunity Name (Categorical):

Data Type: Categorical

No missing values identified.

Unique Values: 33 opportunities with varying frequencies.

Opportunity Category (Categorical):

Data Type: Categorical

No missing values identified.

Unique Values: Event, Course, Competition, Internship, Engagement.

Opportunity End Date (Datetime):

Data Type: Datetime

No missing values identified.

Dates appear to follow a standardized format.

Gender (Categorical):

Data Type: Categorical

One missing value 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.

Key Observations:

Gender distribution is predominantly male.

"Graduate Program Student" is the most common status, followed by other student categories.

"Internship" is the most frequent opportunity category.

"Unknown" is the dominant category for Badge Name, indicating a lot of participants still enrolled but not finished.

Applied Date Sign Ups are distributed across months, with June, July and August having the highest counts.

This analysis provides a comprehensive overview of the Opportunity Sign Up Data, highlighting key characteristics, potential issues, and distributions within categorical columns. The information gleaned lays the groundwork for further exploration and decision-making in subsequent stages of data analysis and visualization.

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 11,481 unique Profile IDs.

4.2 Duplicate Profile IDs (Opportunity Sign Up Data)

Instances of duplicate Profile IDs were identified:
"c2245f7e-2e9d-42c9-b5af-550be9eae1c8" and "18e1e6bc-fada-4b09-bf52-ab45daf318f4" are tied for the highest count at 22.
Followed by "f8ee2854-73b4-4e75-9a11-62e4e9e52a5a."

4.3 Missing Profile IDs (Opportunity Sign Up Data)

No instances of missing Profile IDs were identified.

Key Observations:

The dataset contains a substantial number of unique Profile IDs, indicating diverse user engagement.
Some Profile IDs have multiple instances, suggesting that certain users have engaged with the platform multiple times.

The "Profile ID" column serves as a crucial identifier, and its analysis reveals the diversity of user engagement. The presence of duplicate Profile IDs suggests that some users have registered for multiple opportunities on the Excelerate platform. This insight is valuable for understanding user behavior and tailoring future interactions on the platform. The absence of missing Profile IDs ensures data integrity and completeness in this key identifier.

5. Opportunity Status Distribution

Here we concentrate on the "Status Description" column in the Opportunity Sign Up Data, providing a distribution summary of the different statuses and their occurrences.

Status Description	Count of Status Description
Applied	89
Dropped Out	24
Not Started	1,324
Rejected	726
Rewards Award	2,521
Started	810
Team Allocated	14,206
Withdraw	622

Key Observations:

"Team Allocated" has the highest count, indicating a significant number of participants have been allocated to teams.

"Rewards Award" also has a substantial count, signifying participants who have been awarded rewards.

"Not Started" and "Rejected" have notable counts, representing participants at different stages in their engagement.

The distribution of statuses in the "Status Description" column provides insights into the progression and outcomes of participants in various opportunities. The dominance of "Team Allocated" and "Rewards Award" suggests successful engagement and recognition for a significant number of participants. Understanding the distribution of statuses is crucial for evaluating the success and impact of the opportunities offered on the Excelerate platform.

6. Basic Statistics

The User Data dataset lacks numeric columns, presenting a hurdle for statistical computations. The absence of numerical data, including quantities or values, within this table constrains the capacity for quantitative analysis. This limitation inhibits the application of statistical methods and hampers the exploration of numerical relationships or patterns in the user data. Consequently, the absence of numeric information hinders the ability to derive meaningful quantitative insights from the dataset, impeding comprehensive statistical assessments.

In the following section, we calculate basic statistics (mean, median, min, max) for relevant numeric columns in the Opportunity Sign Up Data.

6.1 Reward Amount (Opportunity Sign Up Data):

Mean: 133.96
Median: 0
Minimum: -1
Maximum: 2500
Standard Deviation: 483.39
Range: 2501
Sum: 2,722,238
Count: 20,322

6.2 Skill Points Earned (Opportunity Sign Up Data):

Mean: 147.18
Median: 0
Minimum: -1
Maximum: 1776
Standard Deviation: 415.79
Range: 1777
Sum: 2,990,966
Count: 20,322

Key Observations:

For "Reward Amount," the mean is influenced by the presence of negative values, indicating missing rewards for certain participants.
Both "Reward Amount" and "Skill Points Earned" have a wide range, with some participants receiving high rewards or earning a significant number of skill points.
The presence of negative values in both columns suggests cases where participants did not meet the requirements for rewards or skill points.

Calculating basic statistics for numeric columns provides a quantitative understanding of the distribution and variability of rewards and skill points earned by participants. These insights are

valuable for assessing the impact of opportunities on the Excelerate platform and identifying patterns in participant performance.

7. Initial Observation

7.1 User Data

1. **Structure of Data:**
   Dataset consists of 8 columns and 27.563 rows.

2. **Date Column:**
   The date column requires correction because of inaccuracies and inconsistencies in both the format and datatype.

3. **Diversity in Data:**
   Within the provided dataset, certain countries are represented by a single individual, while others have representations exceeding 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 "Rewards Award" status, which shows a low completion rate in the program.
   A significant portion of participants has either not started or has 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 have multiple instances, suggesting that certain users have registered for multiple opportunities.

3. **Reward and Skill Points**
   The "Reward Amount" and "Skill Points Earned" columns exhibit wide ranges, with some participants receiving high rewards or earning a significant number of skill points.

Negative values in the "Reward Amount" column suggest cases where participants was Withdrawn, Dropped Out, and 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.

Areas of Interest for Deeper Investigation:

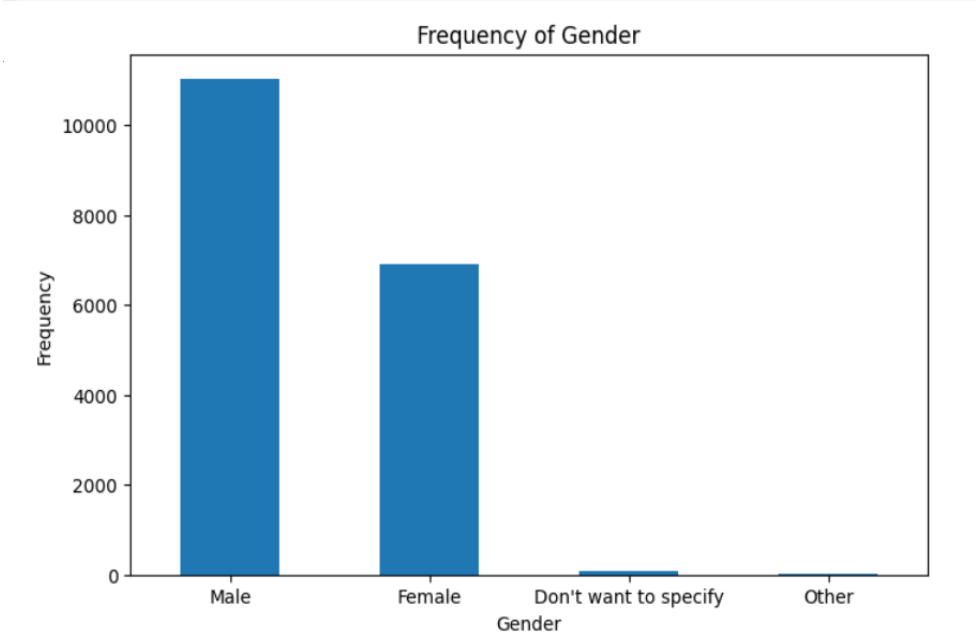
1. **Status Progression Analysis**
   Explore the progression of participants through different status categories.
   Investigate factors contributing to participants not starting or being rejected.
2. **User Engagement Patterns**
   We could analyze the engagement patterns of users with multiple instances of Profile IDs.
   Investigate the characteristics of users who have registered for multiple opportunities.
3. **Impact of Rewards and Skill Points**
   We could also explore the impact of rewards and skill points on participant engagement and success.
   Investigate cases where participants did not receive rewards despite engagement.
4. **Badge Achievements**
   Investigate the significance and impact of specific badges.
   Explore patterns in badge achievements and their correlation with other participant characteristics.
5. **Data Quality and Missing Values**
   Conduct a thorough assessment of missing values and their potential impact on analyses.
   Explore reasons for missing values in specific columns and assess their implications.

These initial observations provide a foundation for deeper investigations in the upcoming weeks, aiming to gain more insights into participant engagement, achievements, and the overall effectiveness of opportunities on the Excelerate platform.

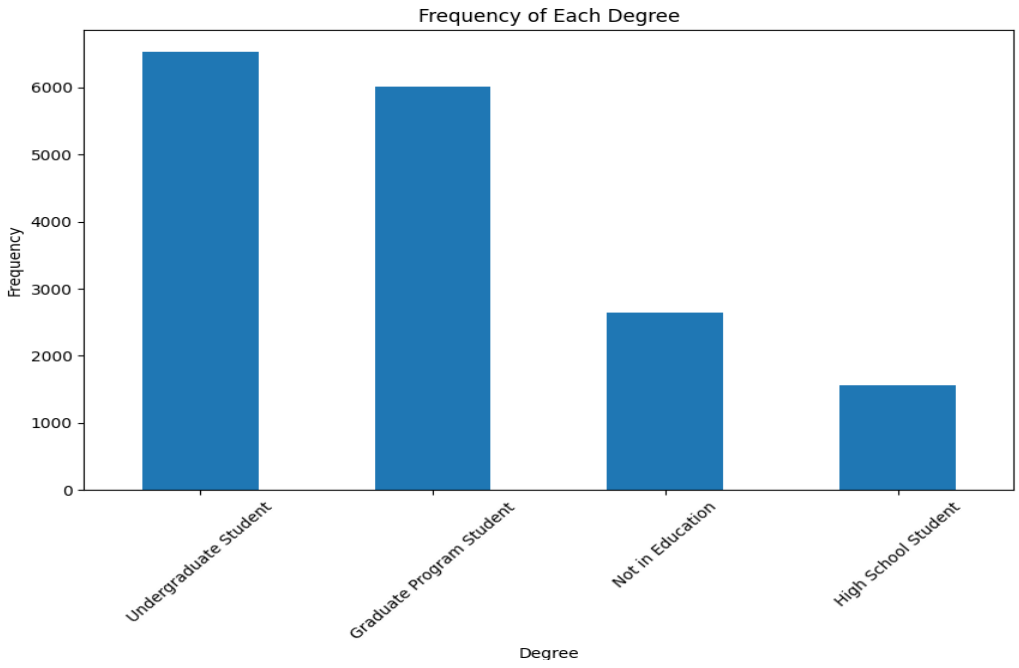
8. Visualizations

8.1 User Data

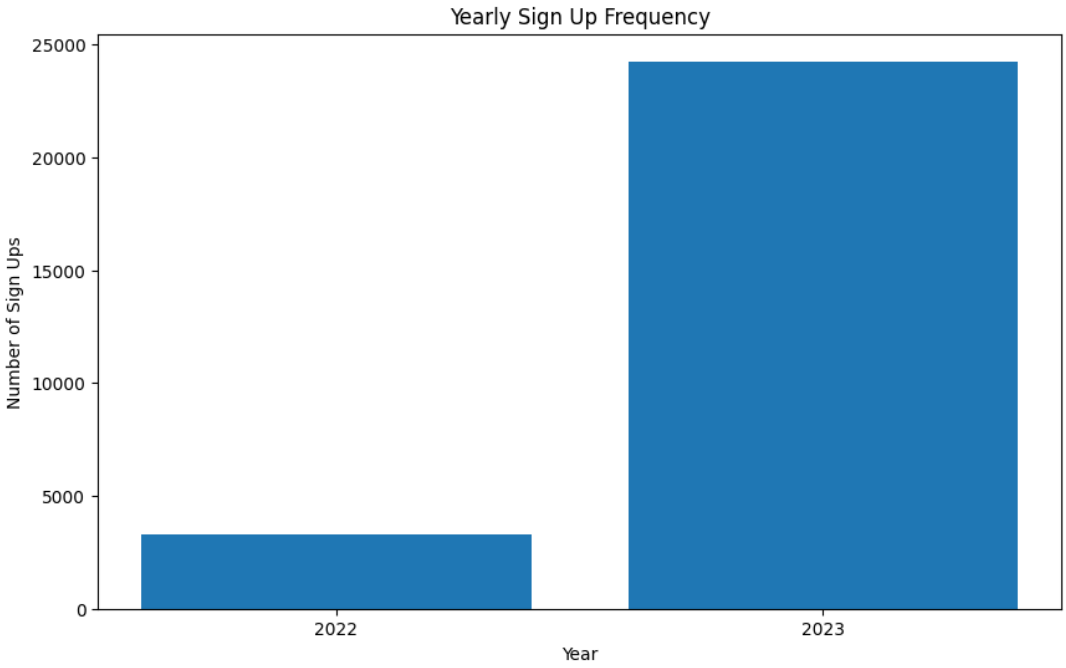
- **Visualization for the Gender Column.**



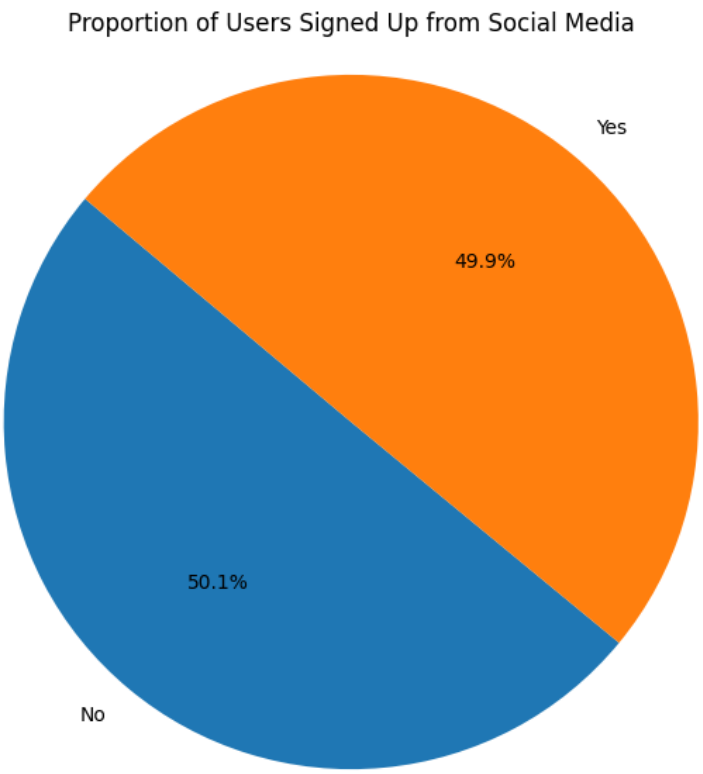
- **Visualization for the Degree Column.**



- **Visualization for Yearly Sign Up.**

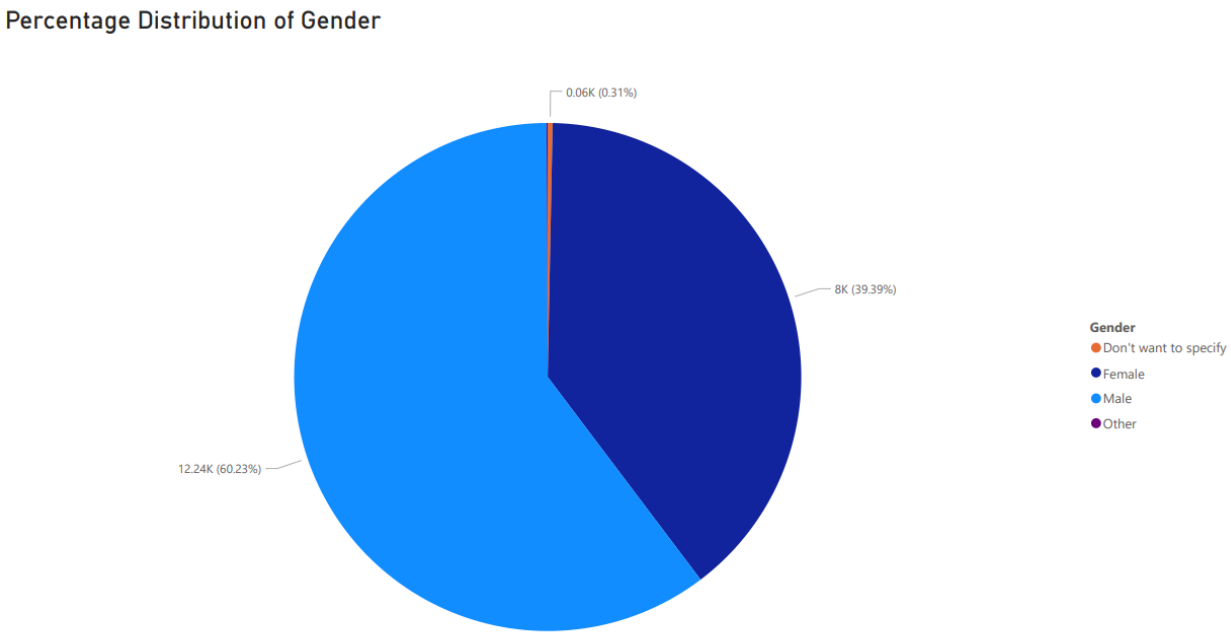


- **Visualization for Column isFromSocialMedia.**

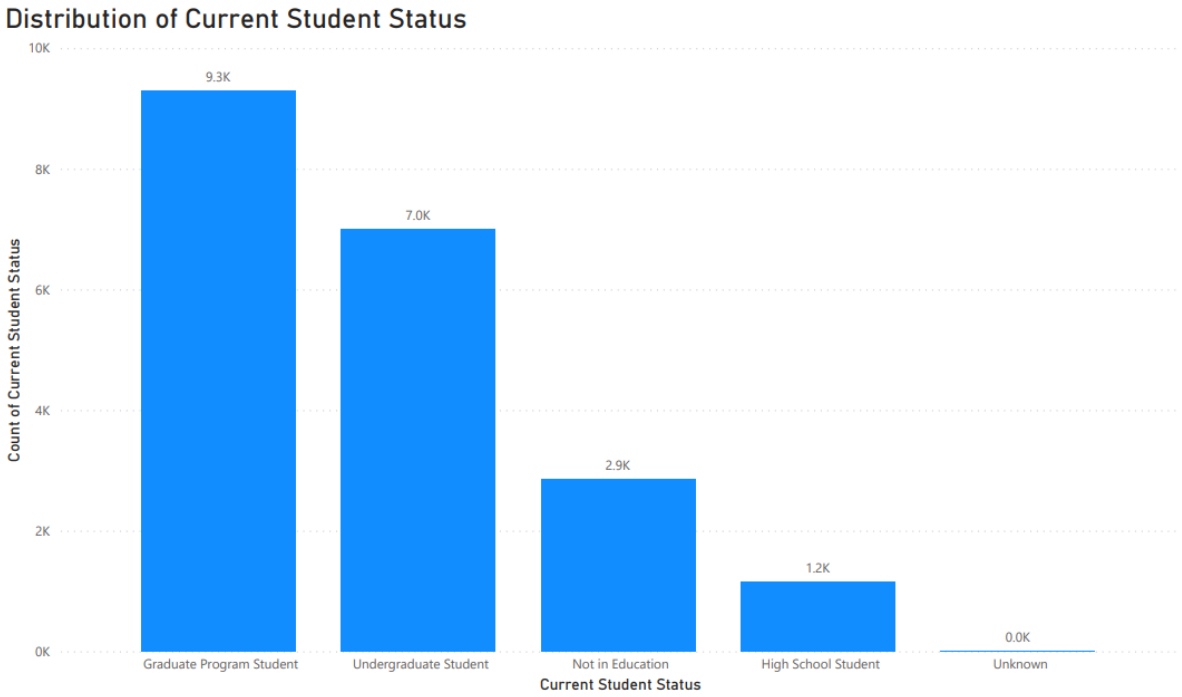


8.2 Opportunity Sign Up Data

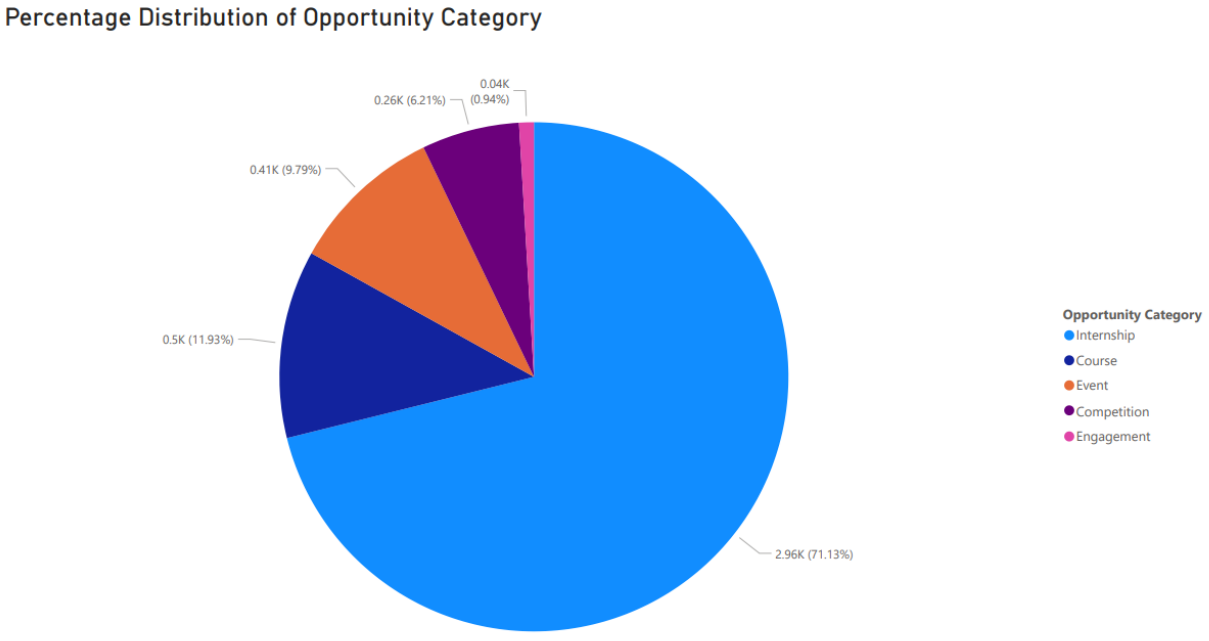
- **A pie chart of the gender distribution in the dataset.**



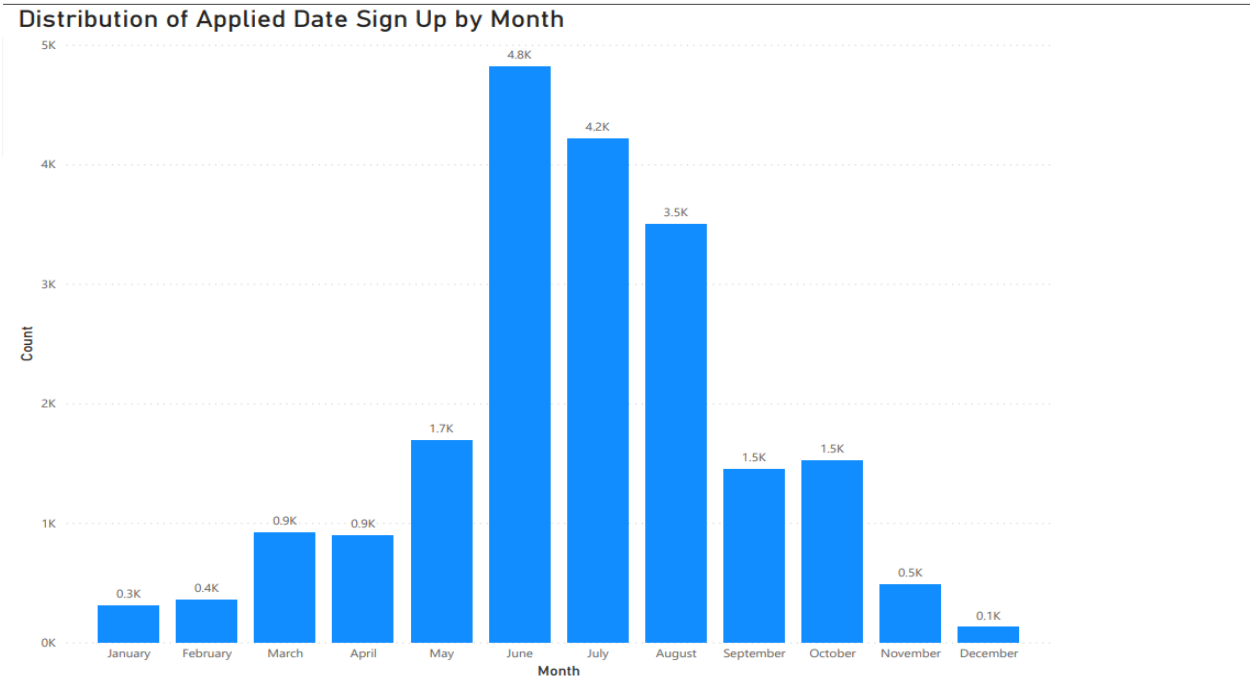
- **A column plot of the Current Student Status in the Dataset.**



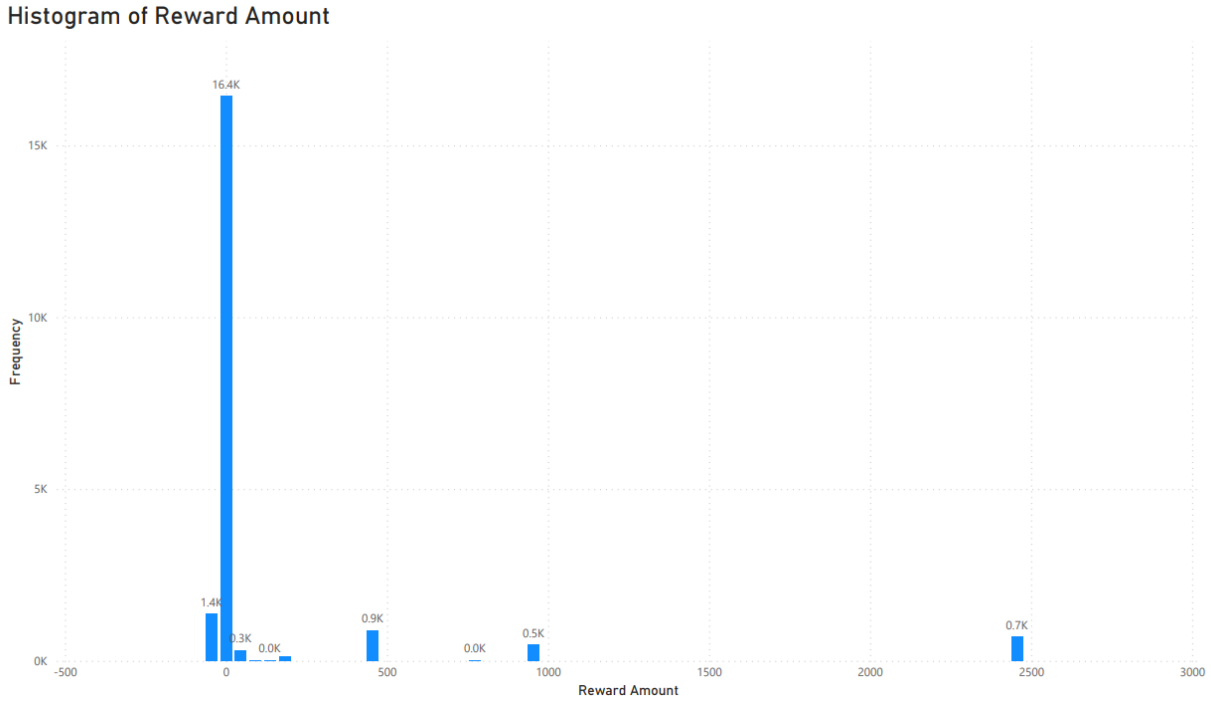
- **A pie chart of the distribution of Opportunity Category in the dataset**



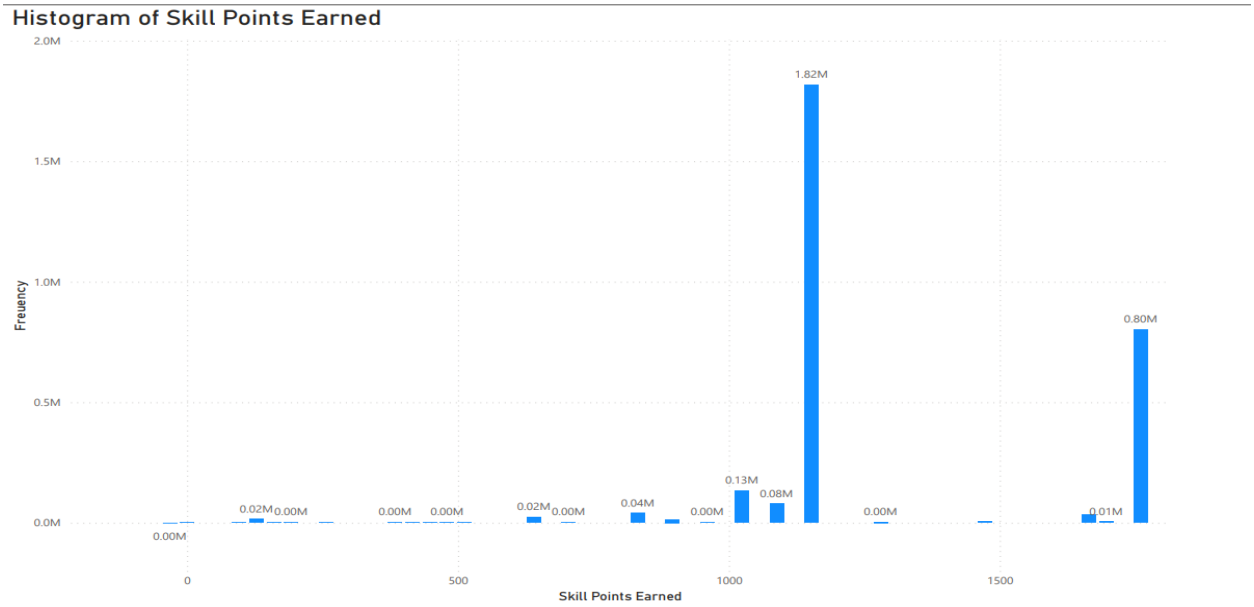
- **The distribution of the number of Applied Date Sign Ups by the month throughout the year.**



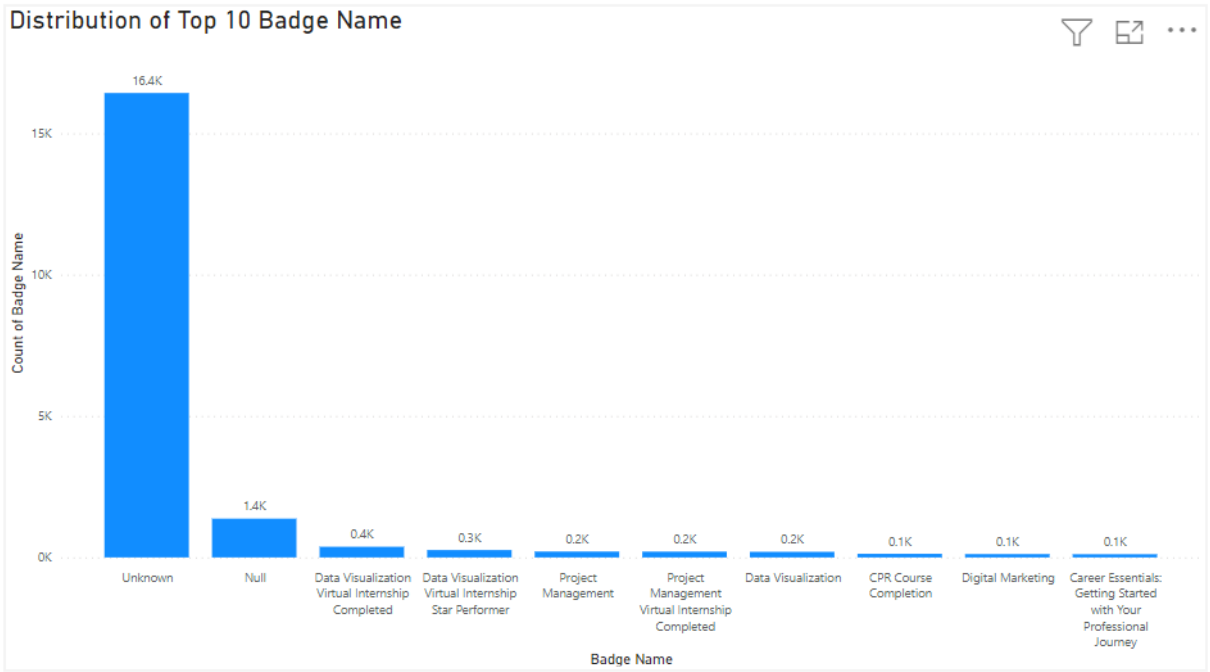
- **A histogram (Right-skewed) showing the distribution of Reward Amount received by the participants.**



- **A histogram (Left-skewed) showing the distribution of Skill Points Earned by the participants.**



- **A count plot showing the distribution of the top 10 Badge Names earned.**



9. Challenges Faced

9.1 User Data

Data Validation:

There are countries in the dataset represented by one or two individuals, which are considered outliers indicative of inaccurate data.

Incomplete and Unclear Dataset:

A minimum of five out of eight columns contain null, blank, or missing values, impacting the precision of the data analysis.

9.2 Challenges Faced

During the exploration process of the Opportunity Sign Up Data, several challenges were encountered:

1. **Inconsistencies in City and State Columns**

   The "City" and "State" columns exhibited numerous inconsistencies, including typos, abbreviations, and misspelled entries. The large number of unique instances made it challenging to thoroughly clean and standardize the data for all entries.

2. **Handling Missing Values in Gender Column**

   The "Gender" column had a missing value (row: 15841) without clear guidance on how to impute it. The absence of additional information such as a name field for that observation limited the ability to make informed imputation decisions.

3. **Zip Code Column Inconsistencies**

   The "Zip Code" column contained a high level of inconsistency, and the format varied widely. Given the lack of standardization and the potential privacy concerns associated with zip code data, a decision was made to suggest dropping the column from the analysis.

4. **Analysis of Duplicate Profile IDs**

   Analyzing duplicate Profile IDs posed a challenge in determining the reason behind multiple instances for certain users. Understanding the nature of these duplications and their impact on the analysis required more detailed information about user behavior and engagement.

5. **Handling Missing Values in Date Columns**

   The "Opportunity Start Date" and "Opportunity End Date" columns had missing values, especially for participants with statuses like Rejected, Dropped Out, or Withdrawn. Decisions were made to fill missing values based on available information, but this process could benefit from additional context.

Addressing these challenges will require collaboration with stakeholders, obtaining additional context or information about the data collection process, and making informed decisions based on the available data. Further discussions and clarifications are recommended to enhance the accuracy and depth of the analysis.

10. Next Steps

In Week 1, the primary focus was on initiating the exploratory data analysis (EDA) by conducting a comprehensive review of the Opportunity Sign Up Data. The tasks included addressing missing values, validating and cleaning categorical variables, and generating an initial EDA report with descriptive statistics and visualizations. The intention was to set the stage for subsequent data processing and visualization.

While some data cleaning and validation were performed in Week 1, it appears that there are additional refinements and deeper investigations needed, as highlighted in the insights gained. The outlined next steps for Week 2 reflect a continuation of the data refinement process and data preprocessing, including addressing specific challenges, clarifying certain data points, and conducting more detailed analyses. The iterative nature of the data exploration process allows for ongoing improvements and insights.

Furthermore, crafting an exhaustive dashboard for Excelerate's leadership necessitates a systematic strategy to address pivotal questions. The aim is to fashion a dashboard providing insights on platform activity, global reach, engagement, opportunity popularity, completion trends, demographic analysis, skill development impact, and scholarship distribution. In the upcoming weeks, a meticulous exploration of feature analysis and data processing will unfold, unraveling profound insights to be seamlessly integrated into subsequent documentation.