Exploratory Data Analysis (EDA) Report

Opportunity and User Data Analysis

DVA 0601 Team 1A

Project Overview

Analysis of two distinct datasets:

> **Opportunity Wise Data (20,322 records)**
> **User Data (27,563 records)**

Team Members

- **Nabiha Orchi (nabihaorch@gmail.com)**
- **Sai Lochan (lochanvana@gmail.com)**
- **Adewale Adeniji (Lxgwales@gmail.com)**
- **Shreya Wani (shreyawani2018@gmail.com)**
- **Aisha Hassan (aishamhmd5@gmail.com)**
- **Haruna Yusuf (harunaayusufrano@gmail.com)**
- **Derrick Addo (derrickaddo029@gmail.com)**

Team Lead

Sai Lochan (lochanvana@gmail.com)

Document Information

- **Document Type: Exploratory Data Analysis Report**
- **Project Code: DVA 0601**
- **Team: 1A**
- **Date: January 2025**

Document Contents

This report presents a comprehensive analysis of opportunity and user datasets, including data cleaning procedures, missing value handling, and categorical variable standardization and visualization.

1. Introduction:

The analysis encompasses two distinct datasets:

A. Opportunity Wise Data: This dataset contains detailed information about various opportunities, including user profiles, opportunity details, rewards, and skills. The data appears to be focused on tracking educational or professional opportunities, their rewards, and the skills participants can earn through them.

B. User Data: This dataset contains user demographic and engagement information, including geographical data, educational background, and sign-up information. It appears to be complementary to the opportunities dataset, providing additional context about the user base.

2. Data Examination:

A. Opportunity Wise Data Dataset:

- Number of rows: 20,322
- Number of columns: 21

Numerical Variables Statistics:

- Reward Amount:
    - Mean: $1,081.26
    - Median: $500.00
    - Standard Deviation: $927.07
    - Range: $50 to $2,500
- Skill Points Earned:
    - Mean: 1,186.96 points
    - Median: 1,182 points
    - Standard Deviation: 399.09 points
    - Range: 10 to 1,776 points

Key Categorical Variables:

- Profile Id
- Opportunity Category
- Gender
- Location (City, State, Country)
- Current Student Status
- Current/Intended Major
- Badge Name
- Skills Earned

B. User Data Dataset:

- Number of rows: 27,563
- Number of columns: 8

Key Variables:

- PreferredSponsors
- Gender
- Country
- Degree
- Sign Up Date
- City
- Zip
- isFromSocialMedia

The data types include a mix of:

- Categorical variables (Gender, Country, Degree)
- Temporal variables (Sign Up Date)
- Geographic variables (City, Zip)
- Boolean variables (isFromSocialMedia)

This comprehensive dataset structure allows for various analyses including:

- Opportunity engagement patterns
- Geographical distribution of users
- Reward distribution analysis
- Skill development tracking
- User demographics analysis
- Social media acquisition effectiveness

A. Opportunity Wise Data Dataset: Columns with Significant Null Values:

- Reward Amount: 17,801 nulls (87.59%)
- Badge Id: 17,801 nulls (87.59%)
- Badge Name: 17,801 nulls (87.59%)
- Skill Points Earned: 17,801 nulls (87.59%)
- Skills Earned: 17,801 nulls (87.59%)
- Opportunity Start Date: 804 nulls (3.96%)

Columns with Minimal Null Values (≤1):

- Gender: 1 null
- City: 1 null
- State: 1 null
- Zip Code: 1 null
- Graduation Date(YYYY MM): 1 null
- Current Student Status: 1 null
- Current/Intended Major: 1 null

Columns with No Null Values:

- Profile Id
- Opportunity Id
- Opportunity Name
- Opportunity Category
- Opportunity End Date
- Country
- Status Description
- Apply Date

B. User Data Dataset: Columns with Significant Null Values:

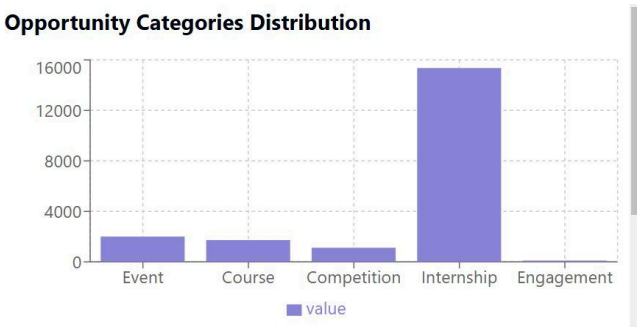
- isFromSocialMedia: 13,751 nulls (49.89%)
- zip: 9,820 nulls (35.63%)
- city: 9,533 nulls (34.59%)
- Degree: 9,370 nulls (34.00%)
- Gender: 8,253 nulls (29.94%)

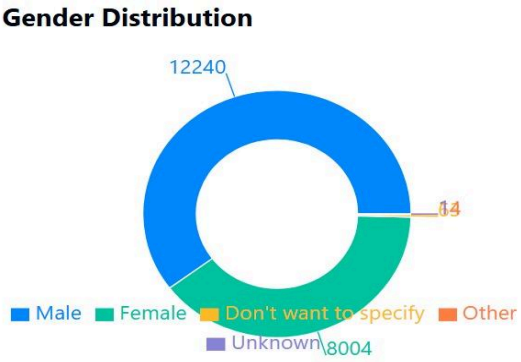
Columns with Minimal Null Values:

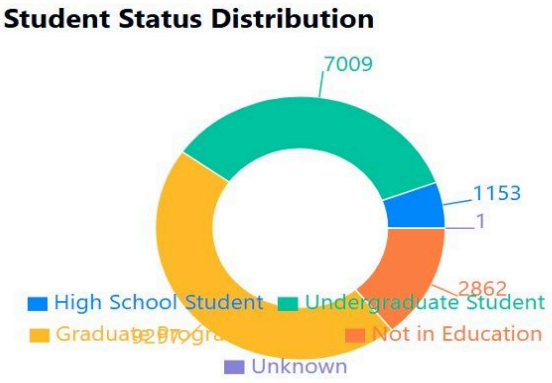
- Country: 36 nulls (0.13%)

Columns with No Null Values:

- PreferredSponsors
- Sign Up Date







Key Observations:

1. The Opportunity Wise Data shows a consistent pattern of nulls (87.59%) across reward and skill-related fields, suggesting these might be optional or conditional fields.
2. The User Data shows significant missing demographic information, with around 30-35% of records missing location and educational details.
3. The social media attribution data is missing for nearly half the users (49.89%).
4. Both datasets maintain complete records for their key identifier fields and temporal data.

3. Handling Missing Values

A. Missing Values Analysis

The dataset contained missing values across several columns that required attention:

- Degree column: Contained blank values
- Gender column: Had blank and null values
- Country column: Contained blank entries
- City column: Had null values and inconsistent entries

B. Missing Values Treatment Strategy

Degree Column

- **Issue**: Blank values in the degree field
- **Solution**:
    - Used filter functionality to identify blank entries
    - Imputed blank values with "Not Mentioned"
    - Applied auto-fill functionality for consistency
- **Justification**: This approach preserves data points while clearly indicating where information was not provided

Gender Column

- **Issue**: Multiple blank and null values
- **Solution**:
    - Applied filtering to identify missing entries
    - Filled out blank values systematically
- **Justification**: Maintaining complete gender information is crucial for demographic analysis

Location Data (Country and City)

- **Issue**:
    - Blank values in country field
    - Null values in city field
    - Inconsistent city entries (single letter 'A')
- **Solution**:
    - Country: Removed 36 rows with blank country values
    - City: Updated null values to 'Not mentioned'
    - City: Replaced single 'A' entries with 'Not mentioned'

    - Total updates: 9,502 city values modified
- **Justification**:
    - Country data was deemed critical enough to warrant row removal when missing
    - City standardization improves data quality while preserving records

And for the other dataset -

A. Missing Values Analysis

The dataset contained missing values across multiple columns that required attention:

- **Degree Column**: Contained blank values
- **Gender Column**: Had blank and null values
- **Country Column**: Contained blank entries
- **City Column**: Had null values and inconsistent entries
- **Skill Points Earned and Reward Amount Columns**: Missing values were treated as 0 due to their numeric data type
- **Badge ID and Badge Name Columns**: Missing values were treated as "Not Mentioned" as they primarily consist of string data

B. Missing Values Treatment Strategy

Degree Column

- **Issue**: Blank values in the degree field
- **Solution**:
    - Used filter functionality to identify blank entries
    - Imputed blank values with "Not Mentioned"
    - Applied auto-fill functionality for consistency
- **Justification**: This approach ensures data completeness while clearly indicating when information was unavailable

Gender Column

- **Issue**: Multiple blank and null values
- **Solution**:
    - Applied filtering to identify missing entries
    - Filled out blank values systematically
- **Justification**: Maintaining complete gender information is crucial for demographic analysis

Location Data (Country and City)

- **Issue**:
    - Blank values in country field
    - Null values in city field
    - Inconsistent city entries (e.g., a single letter "A")
- **Solution**:
    - Country: Removed 36 rows with blank country values
    - City: Updated null values to "Not Mentioned"
    - City: Replaced inconsistent entries ("A") with "Not Mentioned"

Skill Points Earned and Reward Amount Columns

- **Issue**: Missing values due to numeric data type
- **Solution**: Missing values were treated as 0
- **Justification**: This maintains the numerical integrity of the dataset

Badge ID and Badge Name Columns

- **Issue**: Missing string values
- **Solution**: Imputed blank values with "Not Mentioned"

Skills Earned Column

- **Issue**: Multiple responses in a single column
- **Solution**: Split into separate columns — 1st Skill Earned, 2nd Skill Earned, 3rd Skill Earned, and 4th Skill Earned for better data interpretation

Deleted Rows

- **Issue**: One row had missing values for Gender, City, Graduation Date, and Current Date
- **Solution**: Deleted the row to maintain data accuracy and consistency

4. Cleaning Categorical Variables

A. Preferred Sponsor Analysis

- **Issue**: Multiple response question requiring separation for analysis
- **Solution**:
    - Utilized "Text to Column" function to split responses into separate columns
    - Created distinct columns for sponsor preferences
- **Justification**: This transformation enables:

- Analysis of first-choice vs. last-choice sponsors
- Clear visualization of sponsor preference distribution
- Better understanding of sponsorship priorities

B. Standardization of Categorical Values

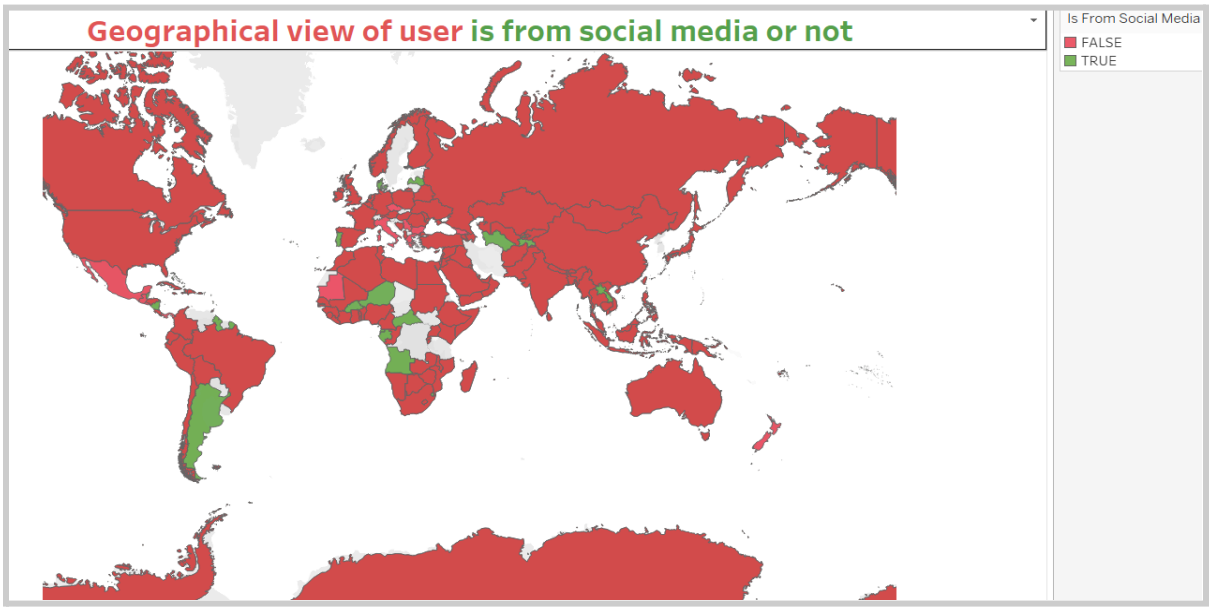
- Implemented consistent value representation:
  - Used "Not mentioned" as standard placeholder for missing categorical data
  - Standardized city names by replacing inconsistent entries
  - Maintained systematic approach to missing value representation across categories

C. Impact on Dataset

- Final dataset maintains integrity while addressing quality issues
- Categorical variables now have consistent formatting and naming conventions
- Missing value handling preserves maximum possible data points while clearly indicating where information was unavailable

5. Data Visualization:

Social Media Analysis:



Looking at the world map visualization, we can observe that certain countries show a notably higher proportion of users coming from social media platforms. These countries, including

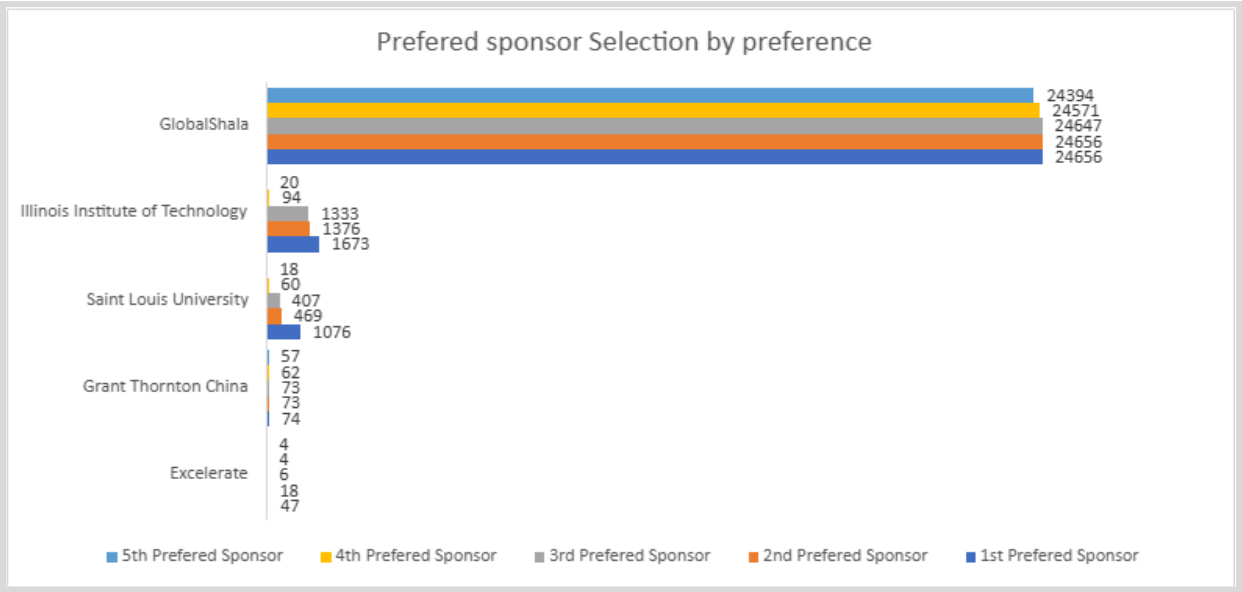
Argentina, Niger, Burkina Faso, Portugal, Central African Republic, Congo, Angola, Turkmenistan, Latvia, Denmark, Brazil, and Laos (shown in green on the map), stand out against the predominantly red coloring of other nations where users typically come from non-social media sources. This distribution pattern spans across multiple continents, suggesting that social media effectiveness as a user acquisition channel varies significantly by country rather than by geographic region.

Significance

This distribution pattern provides valuable insights for:

- Understanding user acquisition channels across different regions
- Identifying successful social media marketing strategies in specific countries
- Planning future user acquisition strategies based on current success patterns
- Recognizing potential opportunities for expanding social media presence in surrounding regions.

Preferred Sponsor Analysis:

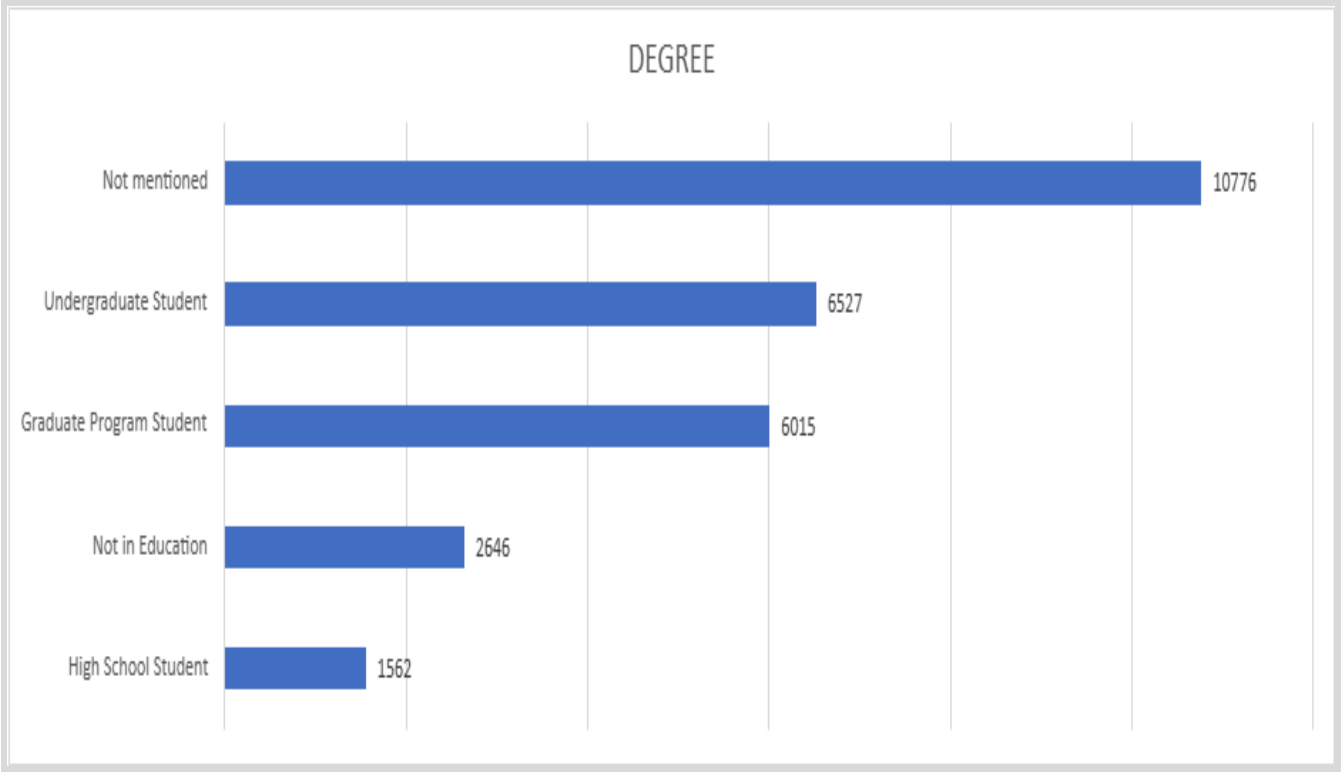


The bar chart shows the sponsor preference rankings across different organizations. GlobalShala emerges as the overwhelmingly preferred sponsor, consistently receiving around 24,000-25,000 selections across all preference levels (1st through 5th). Following distantly is the Illinois Institute of Technology, with approximately 1,300-1,600 selections, while Saint Louis University ranks third in popularity with roughly 400-1,000 selections. Grant Thornton China and Excelerate received notably fewer preferences, with selections ranging from around 50-70 and 4-47 respectively. This visualization clearly demonstrates GlobalShala's dominant position as the most favored sponsor by a significant margin.

Significance:

- The unmatched dominance of GlobalShala's selection rates (24,000-25,000 across all preference levels) demonstrates exceptional brand strength and reveals successful engagement strategies that could be replicated.
- The substantial preference gap between GlobalShala and other sponsors like Illinois Institute of Technology (only 6% of GlobalShala's numbers) highlights important market imbalances that need addressing.
- The steep preference decline pattern seen in Saint Louis University's numbers (from 1,076 to 18) indicates clear user prioritization behaviors that can inform future sponsor positioning strategies.
- The consistently low selection rates for Grant Thornton China and Excelerate (under 100 selections) identify critical areas needing improvement in brand awareness and engagement tactics.

Degree Analysis:



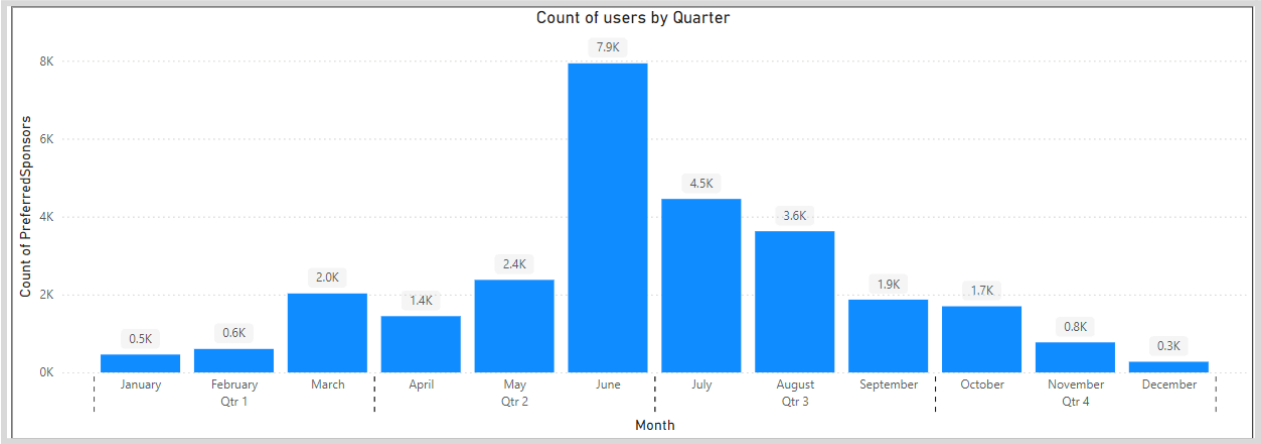
The bar chart titled **"Degree Analysis"** provides a distribution of individuals based on their education level. The x-axis represents the count of individuals, while the y-axis shows the different categories of educational attainment. The categories include:

- **Not Mentioned**: Represents the largest group, with 10,776 entries.
- **Undergraduate Students**: The second-largest group, with 6,527 entries.
- **Graduate Program Students**: Close to undergraduate figures, with 6,015 entries.
- **Not in Education**: Includes 2,646 individuals.
- **High School Students**: The smallest group, with 1,562 individuals.

Significance:

- Undergraduate and Graduate students dominate the user base.
- "Not in Education" group shows potential for alternative opportunities.

Monthly User Distribution by Quarter :



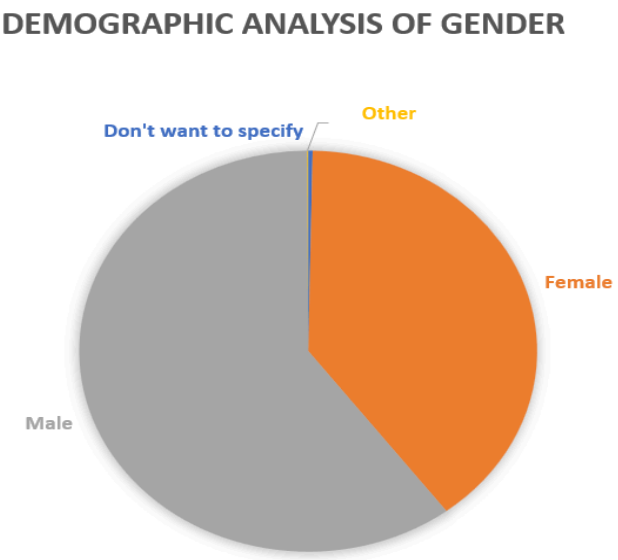
The bar chart shows the number of users per month, grouped by quarters, based on the count of preferred sponsors.

- **Highest Activity**: June has the highest user count (7.9K), followed by July (4.5K).
- **Quarterly Peaks**:
    - Q2 sees the highest engagement, with a sharp rise in user count during May and June.
    - Q3 shows a steady decline after July.
- **Lowest Activity**: December has the lowest user count (0.3K), indicating minimal engagement towards the year's end.

Significance

1. **Seasonal Trends**: Q2 is the most active period, suggesting the importance of campaigns during this time.
2. **Engagement Drops**: Significant drops in Q4 highlight a need to analyze and improve year-end engagement.
3. **June's Dominance**: High user activity in June presents an opportunity to capitalize on peak user engagement.
4. **Quarterly Insights**: Quarterly patterns help prioritize resource allocation for sponsor-driven campaigns.

6. Demographic Analysis:



Gender	Count of Gender Number
Don't want to specify	63
Female	8004
Male	12240
Other	14
Grand Total	**20321**

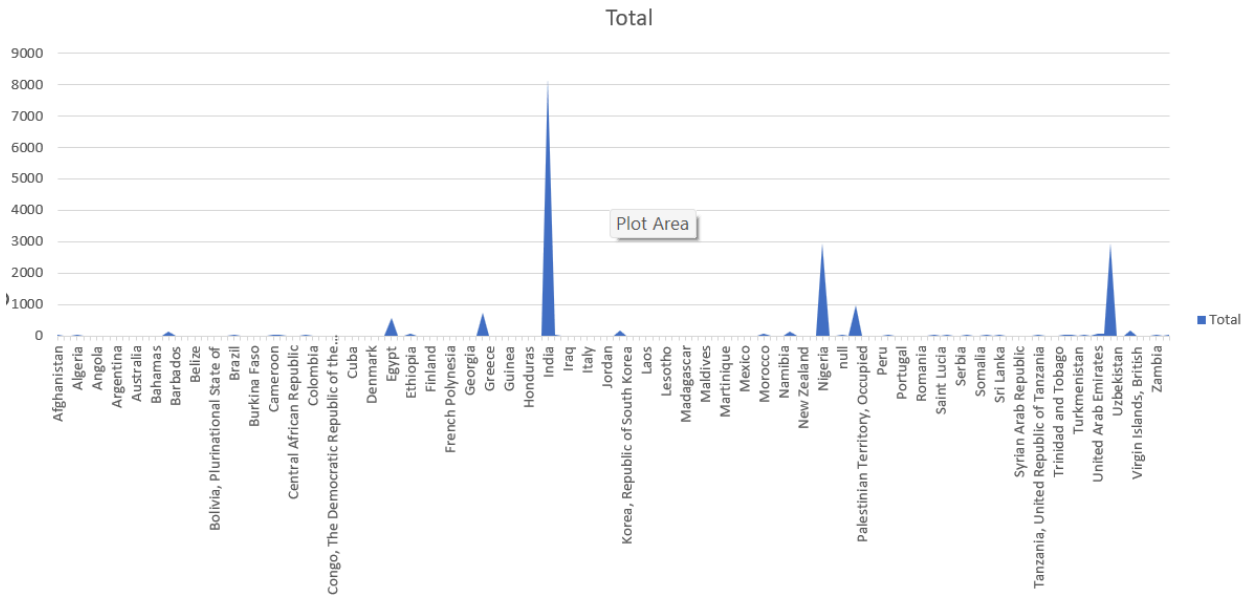
The pie chart titled **"Demographic Analysis of Gender"** shows the distribution of gender among 20,321 individuals:

- **Male**: Largest group, with 12,240 individuals (60%).
- **Female**: Second largest, with 8,004 individuals (39%).
- **Others**: 14 individuals (0.07%).
- **"Don't want to specify"**: 63 individuals (0.3%).

This highlights a male-dominated demographic with moderate female representation.

7. Correlation Analysis:

Distribution of Data by Country



This bar graph presents data labeled as "Total" across various countries worldwide. The x-axis lists country names, while the y-axis quantifies corresponding degree ranging up. Notable spikes are visible for specific countries such as Honduras and Nigeria, indicating significantly higher totals for these locations compared to others.

Significance:

- The graph provides insight into the geographical distribution of a particular dataset.
- Peaks highlight regions of interest or concern, possibly indicating higher rates of a variable such as population, incidents, or resource allocation.
- Understanding such distributions aids in making data-driven decisions, resource planning, and identifying patterns or anomalies.

8. Insights and Findings

1. **Demographic Patterns**:
    - Male users dominate the user base (60%), followed by female users (39%). A small percentage includes other genders and unspecified entries.
    - Undergraduate and graduate students are the most prominent groups, forming the primary user segment.
2. **Trends in Education**:
    - A significant proportion of users fall into "Not Mentioned" for education levels, highlighting data gaps.
    - Non-students and high school participants represent smaller segments, suggesting an opportunity to diversify the target audience.
3. **Social Media Impact**:
    - Countries like Argentina, Niger, and Brazil show strong social media acquisition success, offering insights for geographically targeted campaigns.
4. **Sponsor Preferences**:
    - GlobalShala is the overwhelmingly preferred sponsor, indicating effective engagement and brand positioning.
    - Other sponsors like Illinois Institute of Technology lag significantly, suggesting a need to explore new engagement strategies.
5. **Reward and Skill Data Gaps**:
    - Missing values in reward and skill-related fields (87.59%) suggest conditional or optional data collection, which can limit analysis depth.

Relation to Dashboard Objectives

- Insights on gender and education segments help design tailored filters and content.
- Highlighted geographic patterns support region-specific dashboard customizations.
- Sponsor preferences can inform feature prioritization for sponsor-related campaigns.
- Identified data gaps suggest adding tools for better user data capture and management.

9. Conclusion

The Exploratory Data Analysis (EDA) revealed key demographic patterns, trends in education, and insights into sponsor preferences. Male users and students dominate the user base, while geographic and social media trends highlight areas for targeted growth. Significant data gaps in rewards and skills emphasize the need for improved data collection. These findings provide actionable insights to refine dashboard features, enhance user engagement, and support data-driven decision-making.