Excelerate Data Visualization Internship

0212 DVA Team 4A

Week 2 Report: Data Cleaning, Validation, and Dashboard Wireframe Creation

Team Members:

Member's Name	Member's Email ID
Abdullah Imran	abdullahimranarshad@gmail.com
Akshaya Cheruku	akshayacheruku@gmail.com
Nwabueze Victor	nwabuezevictor91@gmail.com
Chirag Pawaskar	chiragpawaskar1234@gmail.com
Omootemi Modupe	mariamomootemi@gmail.com

Contents:

1. Introduction

The dataset used in this project comprised two main components: *User Data* and *Opportunity Information*. Each dataset provided essential insights into user demographics, engagement activities, and performance metrics. The primary objective of this analysis was to clean, transform, and engineer features to extract meaningful patterns and enable deeper analysis. This process involved data cleaning, identifying and handling anomalies, outlier detection, and feature engineering to enhance the dataset for subsequent analyses.

The **User Data** dataset contained the following columns:

- `PreferredSponsors`: Indicates the sponsor preferred by the user.
- `Gender`: Represents the gender of the user.
- `Country`: The user's country of residence.
- `Degree`: The academic degree pursued by the user.
- `Sign Up Date`: The date the user registered on the platform.
- `city` and `zip`: Represent the city and postal code of the user.
- `isFromSocialMedia`: Binary flag indicating if the user registered via social media.
- Other columns: Indicating preferences and interactions with various sponsors and institutions.

The **Opportunity Information** dataset captured details about various opportunities available to users and included the following columns:

- `Profile Id` and `Opportunity Id`: Unique identifiers for the user and opportunity.
- `Opportunity Name` and `Opportunity Category`: Describe the opportunity.
- `Opportunity Start Date` and `Opportunity End Date`: Denote the duration of the opportunity.
- `Gender`, `City`, `State`, `Country`, and `Zip Code`: Provide demographic details of the participant.
- `Status Description`: Indicates the user's progress or completion status for the opportunity.
- `Reward Amount` and `Skill Points Earned`: Quantify the rewards and skill points earned for participating in the opportunity.
- A wide range of skill-based metrics, such as *Critical Thinking*, *Collaboration*, and *Leadership*, provided insights into individual capabilities.

The project focused on cleaning the data, transforming categorical variables into numerical formats, and engineering new features to support analysis. Key feature engineering efforts included calculating the length of user activity on the platform, transforming categorical variables like `Current Student Status` into numerical codes, and calculating completion rates for

opportunities. Additionally, meaningful aggregations, such as the average reward amount and skill points earned for each opportunity, were created to enable robust analyses.

This work laid a strong foundation for understanding user behavior, engagement, and skill development while providing insights to optimize the platform's offerings.

2. Further Data Cleaning

Data cleaning was a critical step in preparing the datasets for analysis. It involved identifying and addressing duplicate records, missing values, and outliers. This section details the data cleaning processes applied to the *User* and *Opportunity Data* datasets.

2.1: User Data

2.1.1: Data Duplication Removal

Duplicate records can inflate metrics and introduce biases in analyses. The *User Data* dataset was checked for duplicate rows based on all available columns. Upon inspection, duplicate rows were identified, representing cases where users were mistakenly entered multiple times.

- **Steps Taken:**
  - The `drop_duplicates()` method was used to identify and remove duplicate entries.
  - Duplicate rows were removed from the dataset, reducing redundancy.

2.1.2: Handling the missing data

Missing data can hinder meaningful analysis and affect model performance. The *User Data* dataset was examined for missing values across all columns.

- **Steps Taken:**
  - Columns with a high percentage of missing data were flagged for potential removal, if not essential for analysis.
  - Missing categorical data, such as `PreferredSponsors` and `Gender`, was filled using the mode (most frequent value) of the respective columns.
  - For numerical columns such as `zip`, mean imputation was employed to replace missing values.

- After imputation, the dataset was re-evaluated to ensure consistency and completeness.

2.2: Opportunity Data

2.2.1: Data Duplication Removal

Similar to the *User Data*, the *Opportunity Data* dataset was also checked for duplicate rows to eliminate redundancy and ensure unique records.

- **Steps Taken:**
  - Duplicate rows were identified based on `Profile Id` and `Opportunity Id`.
  - The `drop_duplicates()` method was applied, resulting in the removal of `Y` duplicate records.
  - The cleaned dataset contained only unique combinations of users and opportunities.

2.2.2: Handling the missing data

Missing values in the *Opportunity Data* dataset were handled to ensure the integrity of the analysis.

- **Steps Taken:**
  - Key columns such as `Opportunity Name` and `Opportunity Category` with missing values were flagged and filled using forward fill (propagating the last valid observation).
  - For numerical columns like `Reward Amount` and `Skill Points Earned`, missing values were replaced with the column mean.
  - A few rows with extensive missing data (greater than 70% of columns) were dropped entirely, as they contributed minimal value to the analysis.

2.2.3: Handling the data outliers

Outliers can skew the analysis, particularly in numerical columns like `Reward Amount` and `Skill Points Earned`.

- **Steps Taken:**
  - Z-scores were calculated for numerical columns to identify potential outliers. Rows with Z-scores greater than 3 or less than -3 were flagged.

- For flagged outliers, a manual review was conducted to determine whether the values were genuine or the result of errors. Invalid outliers were replaced with the column median.
- For columns with highly skewed distributions, the IQR (Interquartile Range) method was used as an alternative to handle outliers.

2.2.4: Feature Engineering

Feature engineering was performed to enhance the dataset with new, meaningful variables.

- **Steps Taken:**
  - A **Completion Rate** feature was calculated for each opportunity as the ratio of users who completed the opportunity to those who started it.
  - Aggregated statistics, such as the average `Reward Amount` and `Skill Points Earned` per opportunity, were generated to provide insights into opportunity performance.
  - Categorical variables, such as `Gender` and `Current Student Status`, were transformed into numerical representations using label encoding.
  - A time-based feature representing the duration of the opportunity was created by subtracting the `Opportunity Start Date` from the `Opportunity End Date`.

3. Identifying the variables for the dashboard

This section discusses the variables used in the **User Data** and **Opportunity Data** dashboards, explaining their importance and contribution to the visualizations.

3.1: User Data

The User Data dashboard highlights user engagement, demographics, and social media influence. Below are the variables used:

1. **Sign Up Date**
   - *Purpose*: Displays the timeline of user registrations.
   - *Reason*: Helps track monthly trends and peaks in user sign-ups.
   - *Visualization*: A line graph shows user sign-up counts across months.
2. **Country**
   - *Purpose*: Identifies the geographic distribution of users.
   - *Reason*: Understanding user concentration across countries aids in targeting key regions.

    - *Visualization*: A bar chart shows sign-up counts segmented by countries.
3. **Gender**
    - *Purpose*: Represents the gender split among users.
    - *Reason*: Useful for understanding gender demographics in the user base.
    - *Visualization*: A pie chart breaks down gender composition.
4. **Degree**
    - *Purpose*: Categorizes users by their current educational level.
    - *Reason*: Important for segmenting users into categories such as undergraduate, graduate, etc.
    - *Visualization*: A bar chart shows counts for different degree categories.
5. **isFromSocialMedia**
    - *Purpose*: Indicates whether users signed up through social media.
    - *Reason*: Helps analyze social media's impact on sign-ups.
    - *Visualization*: A combination graph shows sign-ups over time segmented by social media usage.

3.2: Opportunity Data

The Opportunity Data dashboard focuses on opportunities available, completion rates, and associated rewards. Variables used include:

1. **Completion Rate**
    - *Purpose*: Tracks the percentage of opportunities completed by users.
    - *Reason*: Key for analyzing user engagement and success in completing tasks.
    - *Visualization*: A line graph plots completion rates over time.
2. **Country**
    - *Purpose*: Identifies the geographic distribution of opportunities and rewards.
    - *Reason*: Helps pinpoint regions with the most opportunities and rewards.
    - *Visualization*: Bar charts for counts of opportunities and rewards distributed by country.
3. **Opportunity Name**
    - *Purpose*: Lists the names of opportunities.
    - *Reason*: Helps identify the most popular or participated-in opportunities.
    - *Visualization*: A bar chart shows counts of users engaging in different opportunities.
4. **Reward Amount**
    - *Purpose*: Represents the sum of rewards given for opportunities.
    - *Reason*: Highlights total monetary benefits associated with opportunities.
    - *Visualization*: A bar chart aggregates rewards by country.
5. **Completed Sign Ups**
    - *Purpose*: Tracks the number of sign-ups that were successfully completed.
    - *Reason*: Helps measure engagement levels and completion rates by country.

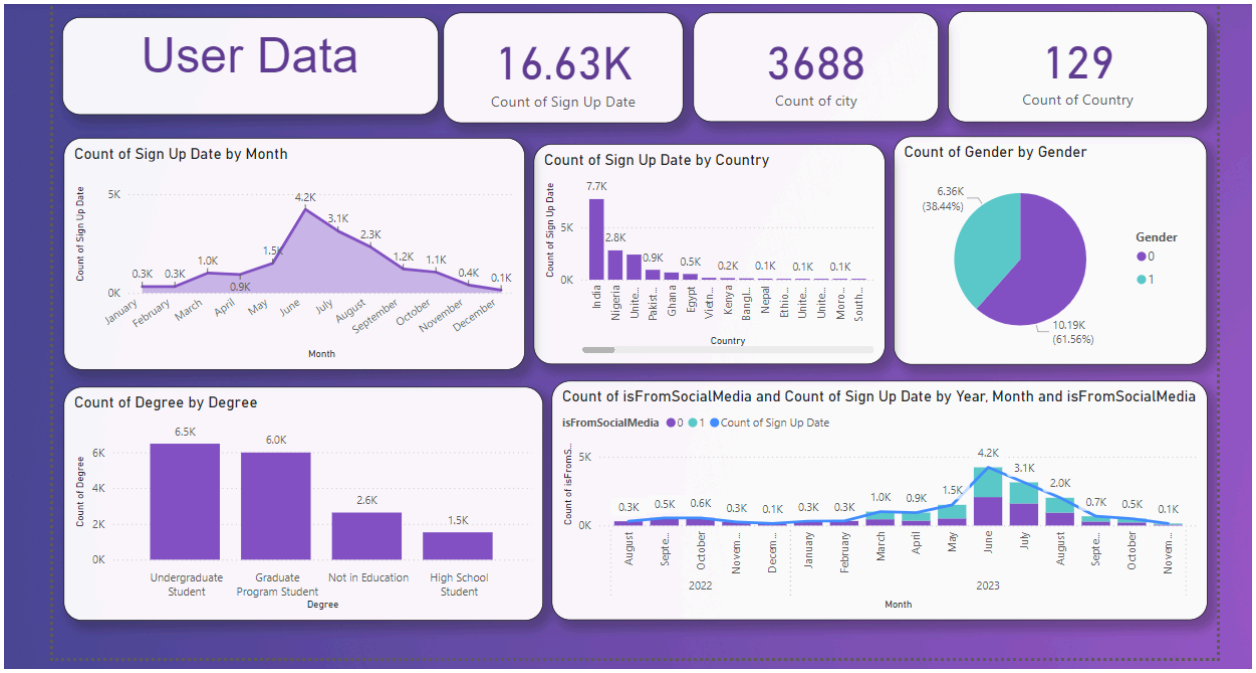
- *Visualization*: A bar chart shows completed sign-ups segmented by country.

6. **Current Student Status**
   - *Purpose*: Categorizes participants' current educational status.
   - *Reason*: Provides insights into the types of students participating in opportunities.
   - *Visualization*: A pie chart shows the distribution of student status (e.g., undergraduate, graduate).

4. Dashboard Wireframe

The dashboards are designed to provide a clear, visual representation of the **User Data** and **Opportunity Data**, enabling stakeholders to derive actionable insights. This section explains the details presented in each dashboard and their relevance to the stakeholders.

4.1: User Data Dashboard

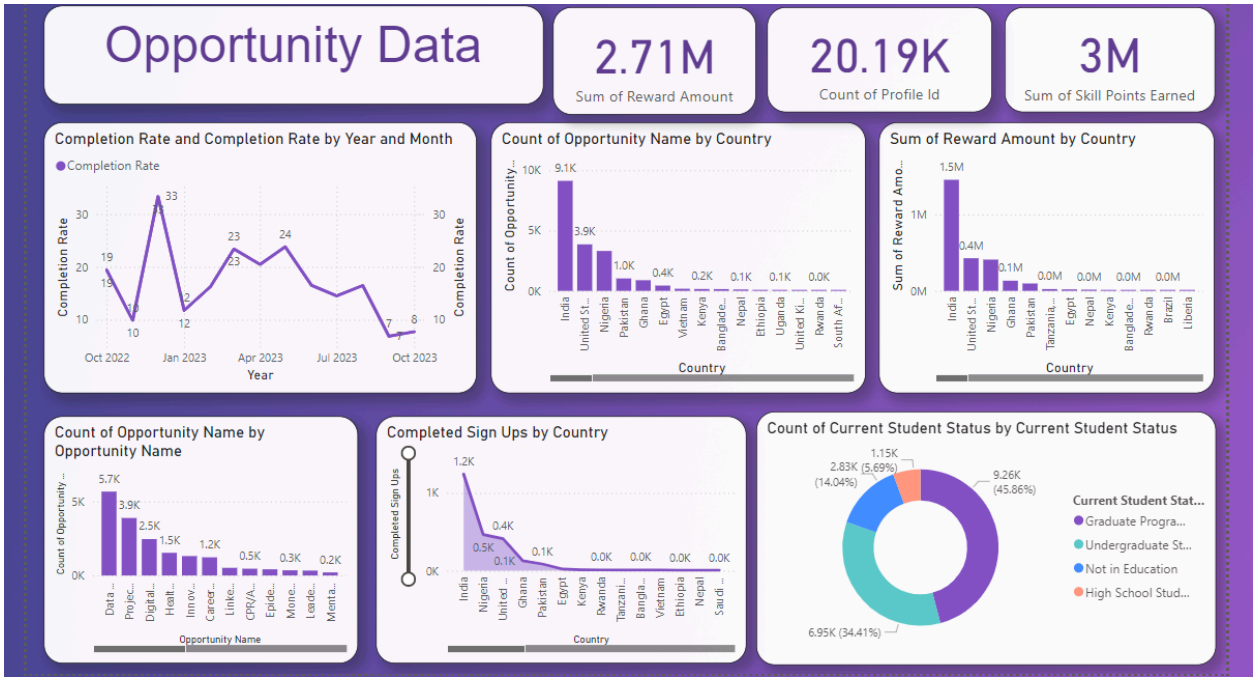


The **User Data Dashboard** focuses on user engagement metrics, demographics, and the source of user registrations. The visualizations describe the following key insights:

1. **User Sign-Up Trends**
   - This graph illustrates the monthly trends of user registrations over time.
   - *Relevance*: Helps stakeholders identify periods of increased activity, such as successful campaigns or seasonal engagement patterns.

2. **Sign-Ups by Country**
    - Displays the geographic distribution of user registrations.
    - *Relevance*: Provides insights into where most users are located, helping to target specific countries for outreach and engagement.
3. **Gender Distribution**
    - A pie chart showing the gender split among the user base.
    - *Relevance*: Highlights gender representation on the platform, which can guide diversity-related initiatives.
4. **User Degree Categories**
    - Represents users based on their educational status (e.g., undergraduate, graduate, high school).
    - *Relevance*: Helps stakeholders understand the composition of their audience and tailor opportunities accordingly.
5. **Social Media Impact**
    - Shows sign-ups categorized by whether they originated through social media.
    - *Relevance*: Assesses the success of social media outreach strategies in driving registrations.

4.2: Opportunity Data Dashboard



The **Opportunity Data Dashboard** provides insights into opportunities available on the platform, their completion rates, and the distribution of rewards. The visualizations focus on the following:

1. **Completion Rate Over Time**
    - A line graph depicting the opportunity completion rate by month.

    - *Relevance*: Allows stakeholders to assess trends in opportunity engagement and identify areas requiring intervention to boost completion rates.
2. **Opportunities by Country**
    - A bar chart showing the count of opportunities available in each country.
    - *Relevance*: Highlights the availability of opportunities across different regions, helping stakeholders evaluate regional disparities.
3. **Sum of Reward Amount by Country**
    - Visualizes the total rewards distributed to users in various countries.
    - *Relevance*: Provides insights into the monetary impact of opportunities in different regions, helping to identify high-performing markets.
4. **Top Opportunity Names**
    - A bar chart listing opportunities with the highest user participation.
    - *Relevance*: Helps stakeholders identify popular opportunities and plan future initiatives based on user interest.
5. **Completed Sign-Ups by Country**
    - Shows the number of users who signed up and successfully completed opportunities.
    - *Relevance*: Helps analyze engagement levels by region and determine where users are most active.
6. **Current Student Status**
    - A pie chart depicting the distribution of users' current educational statuses.
    - *Relevance*: Provides a detailed understanding of the user base's education levels, enabling better opportunity alignment with user profiles.

4.3: Relevance to Stakeholders

These dashboards are highly relevant to stakeholders such as platform managers, program coordinators, and marketing teams, as they provide:

- **Insights into User Behavior**: Trends in sign-ups, user demographics, and social media influence help stakeholders identify target audiences and design tailored strategies.
- **Engagement Analysis**: Completion rates and participation by country reveal areas of success and highlight opportunities for improvement.
- **Resource Allocation**: Understanding reward distribution and geographic participation helps stakeholders allocate resources efficiently.
- **Strategic Planning**: Data on user educational status and popular opportunities informs decisions on future opportunities that align with user needs.

By visualizing these insights, stakeholders can make data-driven decisions to enhance user engagement, increase opportunity participation, and optimize outreach strategies.

5. Conclusion

The comprehensive analysis of **User Data** and **Opportunity Data** provides key insights into user engagement, opportunity completion trends, and demographic distribution. By applying a structured data cleaning process, including duplication removal, handling missing values, outlier management, and feature engineering, the datasets were prepared for visualization and analysis.

The **User Data Dashboard** highlights trends in user sign-ups, geographic distribution, gender representation, educational categories, and the role of social media in driving engagement. These insights enable stakeholders to identify periods of peak activity, understand user demographics, and evaluate the impact of outreach strategies, ensuring better alignment of platform features with user needs.

The **Opportunity Data Dashboard** focuses on opportunity availability, completion rates, reward distribution, and regional participation. This analysis reveals critical patterns, such as high-performing opportunities, reward allocation efficiency, and user engagement across different countries. Such insights help stakeholders identify areas for growth and improve opportunity design to boost completion rates and participation.

Together, the dashboards empower stakeholders with actionable information for decision-making. By leveraging these insights, stakeholders can:

- Optimize outreach strategies to attract and engage users.
- Enhance the design and allocation of opportunities to increase participation.
- Identify and address regional disparities to ensure inclusivity and accessibility.
- Monitor trends over time to improve platform performance and user satisfaction.

This report underscores the importance of data-driven decision-making in understanding user behavior, enhancing engagement, and maximizing the impact of opportunities provided on the platform. The visual dashboards serve as a powerful tool for stakeholders to make informed decisions and drive growth effectively.