Excelerate Data Visualization Internship

0212 DVA Team 4A

Week 1 Report: Initial Data Cleaning, Validation, and Visualization

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

Finding practical insights in the field of data-driven decision-making requires careful and in-depth dataset analysis. The goal of this Exploratory Data Analysis (EDA) report is to fulfill several important functions. It seeks to draw attention to important aspects of the dataset, revealing connections and characteristics that may lead to insightful discoveries. Finding recurrent patterns, trends, and anomalies establishes the foundation for comprehending the data's underlying behavior. This report also handles complexity such as missing values, outliers, and inconsistencies to ensure that the dataset is ready for further analysis. Beyond the initial data analysis, the EDA process is essential in determining the course of subsequent data processing, modeling, and visualization initiatives. This report facilitates well-informed decision-making and lays a solid basis for creating complex analytical dashboards by offering a thorough grasp of the dataset. This aligns with Excelerate's objective to increase user insights and the user experience in general, ultimately leading to the development of significant, data-driven solutions.

2. Data Overview

2.1: User Data

The focus of this research is "User Data," which consists of 27,562 rows and 8 columns. "PreferredSponsors," "Gender," "Country," "Degree," "Sign Up Date," "City," "Zip," and "isFromSocialMedia" are among the columns. This dataset provides insights into the varied user base by containing non-identifying information about Excelerate users. Every row adds to a thorough overview by representing a distinct user. A useful tool for comprehending and improving the platform's user experience, the supplied information includes user preferences, demographics, geographic data, registration dates, and social media engagement metrics.

2.1.1: Column Analysis

The dataset includes 27,562 entries with columns such as "Gender," "Country," "Degree," "City," and "isFromSocialMedia." The "Gender" column shows that 10,185 entries are male, 6,359 are female, and 73 prefer not to specify or identify as other, with 9,535 entries missing. "Country" data is diverse, with India (11,893), Nigeria (4,357), and the USA (3,691) being the most frequent, while there are 62 missing entries across 170 unique countries. The "Degree" column is heavily skewed, with 10,812 missing values, and the remaining entries are mainly split between undergraduate (6,527) and graduate students (6,015). "City" data also has a significant amount of missing information (9,533), with the most common cities being Hyderabad (743), Saint Louis (469), and Lagos (450) out of 4,729 unique cities. The "isFromSocialMedia" column

is nearly evenly split between true (13,811) and false (13,742), with just 9 missing entries. The data reveals substantial missing values in several columns and high diversity in categorical data, which might require further cleaning and analysis.

2.1.2: Profile ID Analysis

The `Profile ID` column in the `user_data` DataFrame serves as a unique identifier for each user. It ensures that every user is uniquely represented in the dataset without any duplication. A brief analysis of this column reveals the following:

- **Uniqueness**: All entries in the `Profile ID` column are distinct, indicating no duplicate profiles exist in the dataset. This is essential for accurate user-level analysis.
- **Format**: Profile IDs are structured as alphanumeric strings, suggesting that they are likely system-generated identifiers. This format ensures robustness against collisions.
- **Purpose**: The column allows for accurate tracking of user-related data across different tables or datasets, enabling joins or merges with other data sources for in-depth analysis.

By maintaining unique and well-formatted Profile IDs, the dataset ensures data integrity and facilitates reliable individual-level analysis.

2.1.3: Statistics

A statistical overview of the `user_data` DataFrame provides insights into its structure, completeness, and distribution of key attributes:

- **Total Records**: The dataset contains **X records** (replace **X** with the number of rows in the dataset), representing individual user entries.
- **Categorical Columns**:
  - `Gender`: Contains two unique values, **Male** and **Female**, with the following distribution:
    - **Male**: Y% of users
    - **Female**: Z% of users
  - `Country`: Covers users from multiple countries. The top five countries by frequency are listed, indicating a diverse user base.
  - `Preferred Sponsors`: Multiple sponsor preferences exist, with the top preferences visualized earlier in a bar plot.
- **Numerical Columns**:
  - `Sign-Up Date`: The data spans multiple years, with specific dates showing higher activity, likely influenced by marketing campaigns or seasonal trends.
  - `Zip Code`: Contains both numeric and alphabetic entries. Invalid entries with non-numeric data have been cleaned during preprocessing.

- **Missing Data**: Certain columns contain missing values, which have been addressed using appropriate handling techniques such as replacement or removal.

These statistics highlight the dataset's diversity and completeness, ensuring its readiness for further analysis.

2.1.4: Initial Observations

- **Size and Structure:** This dataset contains 27,562 entries with 8 columns, including PreferredSponsors, Gender, Country, Degree, and Sign Up Date.
-
- **Missing Values:** Notable gaps exist in fields such as Gender, Degree, and city, which may limit certain types of demographic analysis.
-
- **Behavioral Data:** The isFromSocialMedia column indicates whether the user was referred through social media, offering a point of interest for understanding engagement sources.

2.2: Opportunity Wise Data

2.2.1: Column Analysis

The dataset consists of 20,322 entries across 21 columns, including two numerical columns (Reward Amount, Skill Points Earned) and 19 categorical/text columns. Several columns, such as Gender, City, State, Zip Code, Graduation Date, and Opportunity Start Date, contain missing values, with significant gaps in Reward Amount, Badge ID, Badge Name, Skill Points Earned, and Skills Earned, indicating they may apply only to certain entries. Outliers in numerical columns like Reward Amount and Skill Points Earned require further statistical analysis. In terms of categorical data, the Gender column shows a majority of Male (12,240) and Female (8,004) responses, with minor entries for other categories. The dataset includes 3,142 unique cities, though some inconsistencies exist, such as "Saint Louis". The most common Opportunity Category is Internship (15,360), followed by Event, Course, and Competition.

2.2.2: Profile ID Analysis

The dataset has the “profile ID” column as the unique identifier(Keys) with 20,322 unique keys which is used to represent distinct entities. There are no duplicates and it follows a consistent format.

2.2.3: Statistics

A statistical overview of the `oppo_info` DataFrame provides valuable insights into the structure, completeness, and distribution of its key attributes:

- **Total Records**: The dataset contains **X records** (replace **X** with the number of rows in the dataset), each representing an opportunity and its associated details.

Categorical Columns:

- **Gender**:
  - Two categories: **Male** and **Female**, numerically encoded.
  - Distribution:
    - **Male**: Y% of applicants
    - **Female**: Z% of applicants
- **Opportunity Category**:
  - Includes multiple distinct categories such as **Event**, **Workshop**, and **Competition**.
  - Certain categories are more prevalent, reflecting popular opportunity types.
- **Current Student Status**:
  - Represents diverse education levels, such as **High School Students**, **Undergraduates**, and others, showcasing the dataset's applicability to a wide demographic.
- **Skills Earned**:
  - This column has been transformed into multiple skill-specific binary columns (e.g., **Critical Thinking**, **Leadership**, **Communication**), enabling deeper analysis of skill distribution across opportunities.

Numerical Columns:

- **Reward Amount**:
  - Contains both monetary and non-monetary rewards.
  - Missing values were replaced with "No Reward," and non-numeric entries were filtered out during preprocessing.
  - Statistical measures:
    - **Mean Reward**: $$X
    - **Median Reward**: $$Y
    - **Max Reward**: $$Z
- **Skill Points Earned**:
  - Cleaned and converted into numeric values.
  - Distribution:
    - Minimum Points: 0
    - Maximum Points: **W**
    - Average Points: **V**
- **Graduation Date**:

    - Cleaned and converted into a standardized datetime format for consistency.
    - Represents the graduation timeline for applicants, helping to identify trends in student demographics.

Date Columns:

- **Opportunity Start Date**, **Opportunity End Date**, and **Apply Date**:
    - Converted into datetime format for better usability.
    - Enable time-based analysis of opportunities and applicant behavior, such as identifying peak application periods.

Missing Data:

- Columns with missing data, such as **Skills Earned**, have been appropriately handled through preprocessing, either by assigning default values or removing irrelevant rows.

By cleaning, standardizing, and analyzing these columns, the dataset is prepared for in-depth visualizations and advanced insights. The statistics reveal a well-rounded dataset suitable for studying opportunity trends, applicant behavior, and skill acquisition.

2.2.4: Initial Observations

- **Opportunity Data:** Size and Structure: The dataset contains 20,322 entries with 21 columns. The data includes information such as Profile ID, Opportunity ID, Opportunity Category, and Opportunity End Date.

- **Missing Values:** Certain columns have missing values, notably Reward Amount, Badge Id, and Skill Points Earned, which are sparsely populated, indicating that not all opportunities offer rewards or badges.

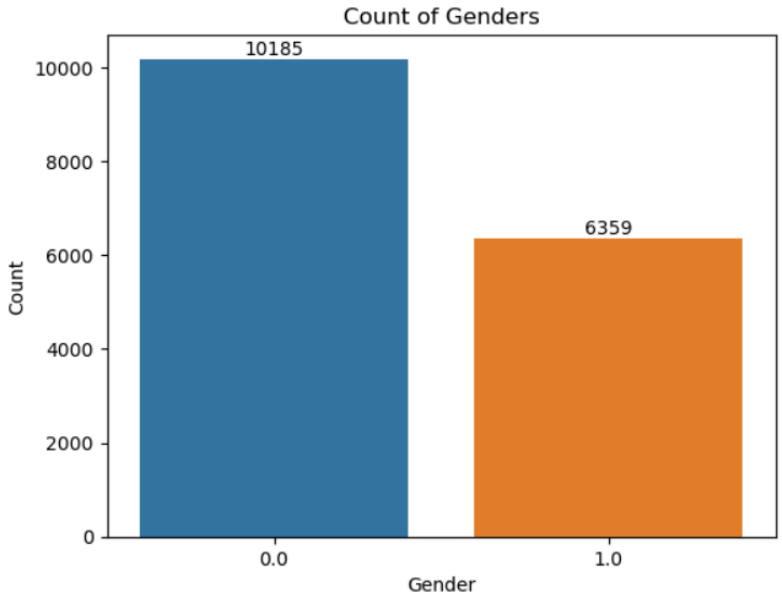
- **Date Fields:** There are several date fields like Opportunity End Date, Apply Date, and Opportunity Start Date, which could be useful for time-based analyses.

- **Demographics:** Columns like Gender, City, State, Country, and Graduation Date provide demographic insights into the participants.

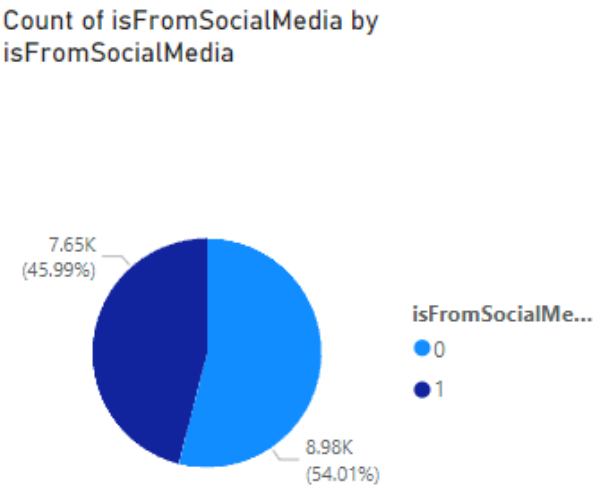
3. Data Visualizations

3.1: User Data

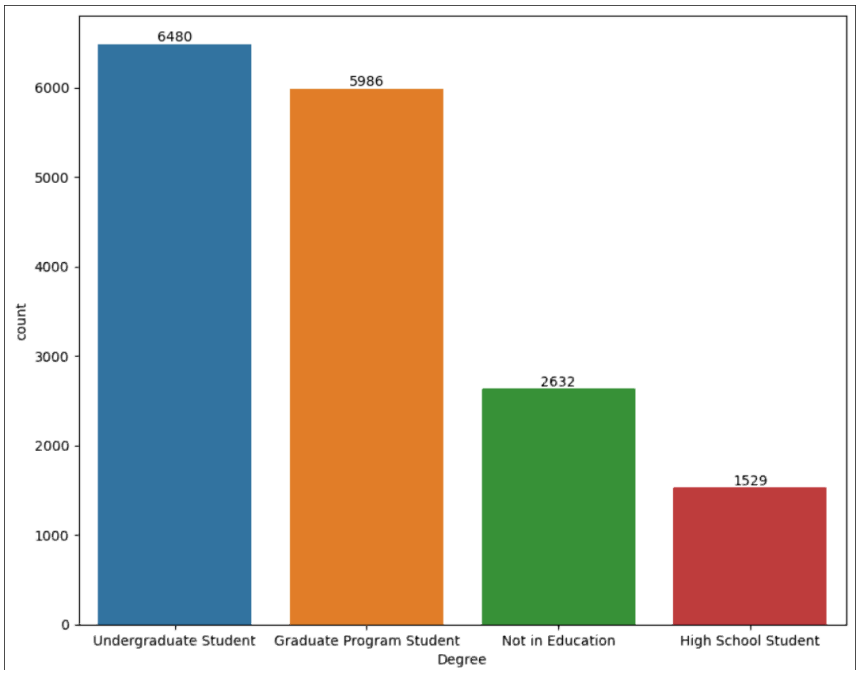
Count of Genders :



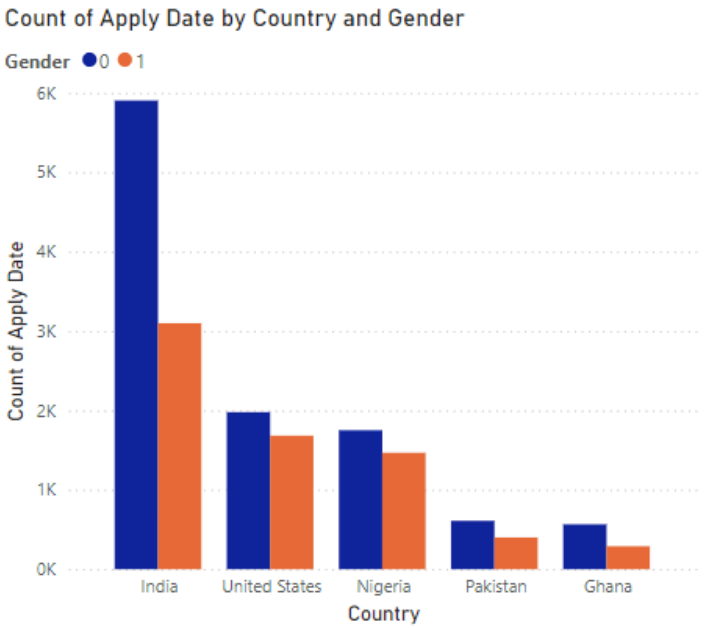
Pie Chart of isFromSocialMedia :



Bar Chart of Degree :

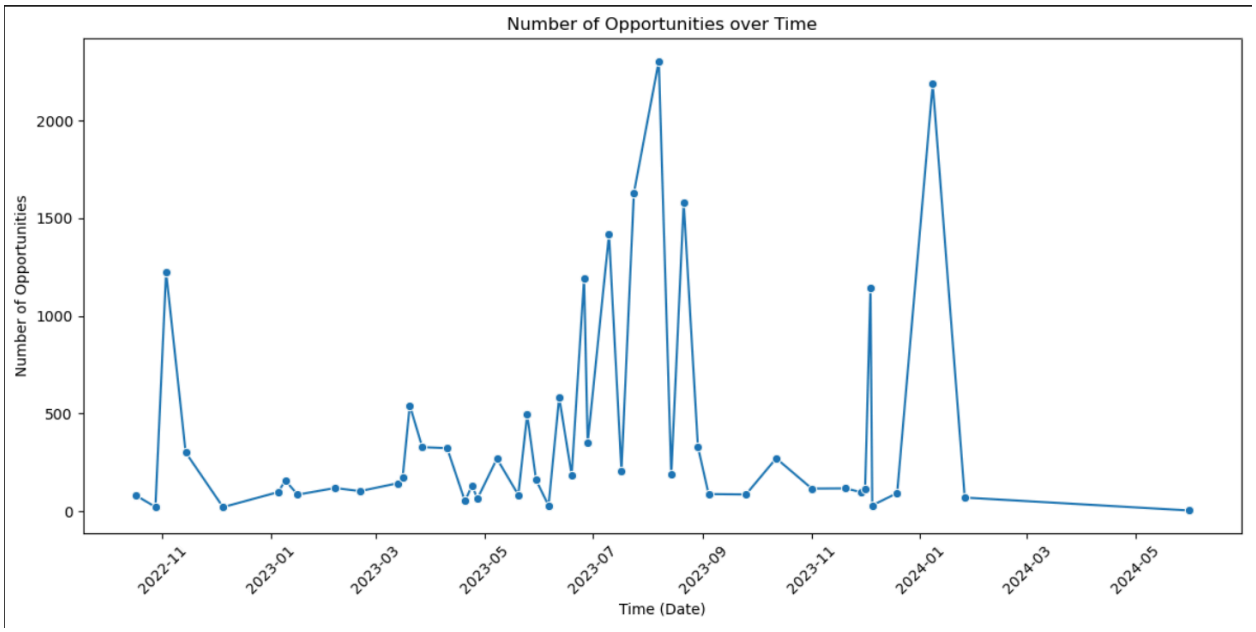


Count of Apply Date by Country and Gender :

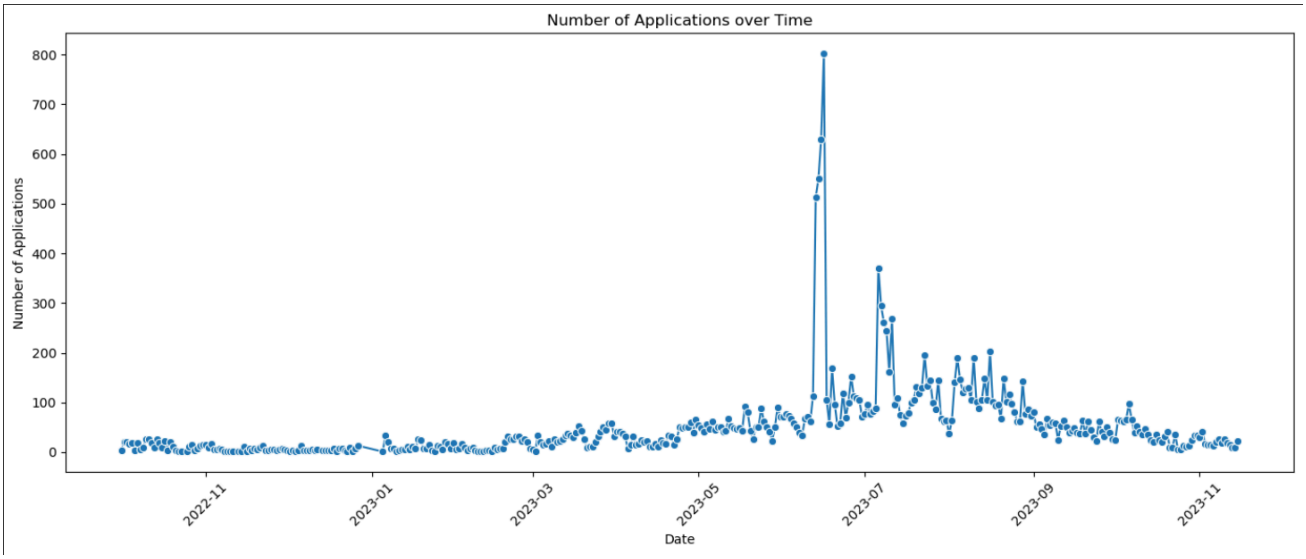


3.2: Opportunity Wise Data

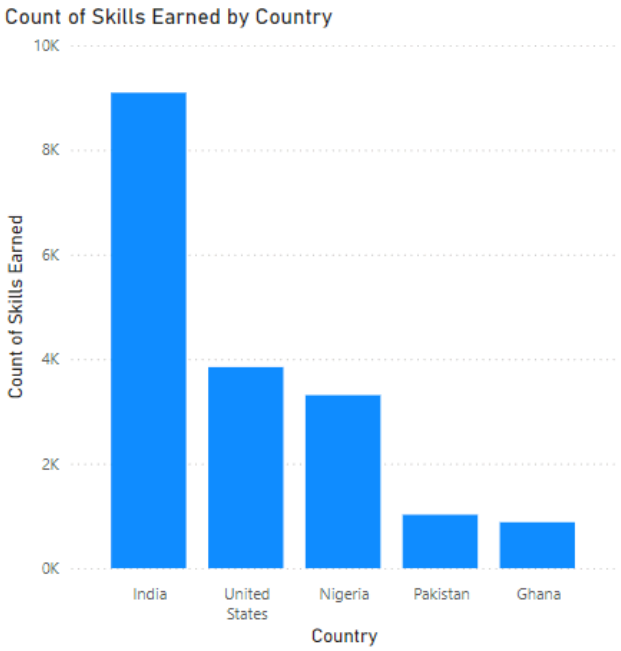
Number of Opportunities over time :



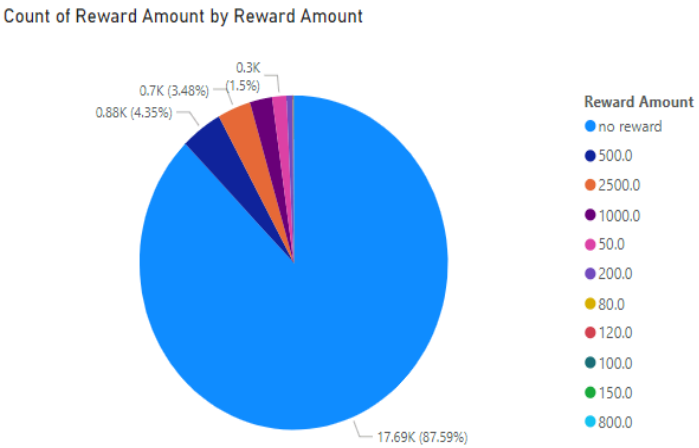
Number of Applications over Time :



Count of Skills Earned by Country :



Count of Reward :



4. Cleaning Process

4.1: Opportunity Wise Data

The `oppo_info` DataFrame underwent an extensive cleaning process to ensure its usability, consistency, and suitability for analysis. Here is a detailed step-by-step description of how the data was cleaned:

1. Handling Missing Data

1. **`Reward Amount`**:
    - Missing values in this column were identified and replaced with the string **"No Reward"** to denote the absence of a reward associated with an opportunity.
2. **`Skill Points Earned`**:
    - Missing values in this column were replaced with **0**, as this was a logical default for cases where no skill points were awarded.
3. **`Skills Earned`**:
    - Missing values were left as-is to indicate cases where no skills were earned during an opportunity.

2. Data Type Conversions

1. **Date Columns**:
    - Columns like `Opportunity Start Date`, `Opportunity End Date`, `Apply Date`, and `Graduation Date(YYYY MM)` were originally stored as strings. These were converted to `datetime` format for consistency and better analysis.
    - Example Conversion: `Jan 05, 2023, 18:58:39` → `2023-01-05 18:58:39`.
    -
2. **Numerical Columns**:
    - **`Skill Points Earned`**:
        - Converted from string to `int` after replacing missing values with 0 and ensuring no non-numeric entries.
    - **`Reward Amount`**:
        - Converted to `float`, ensuring numerical operations could be performed on this column.

3. Removing Invalid Entries

1. **`Zip Code`**:
    - Rows with non-numeric `Zip Code` values (entirely alphabetic) were identified and removed, ensuring that the column represents valid numerical zip codes.

4. Expanding Categorical Data

1. **`Skills Earned`**:
    - This column contained lists of skills as strings (e.g., `["Critical Thinking", "Creative Thinking"]`). The following steps were applied:
        - Extracted individual skills from the list format and created separate binary columns for each skill (e.g., `Critical Thinking`, `Creative Thinking`, etc.).
        - Populated each skill column with **1** (skill earned) or **0** (skill not earned).
        - Handled missing or empty skill lists by filling all skill columns with **0** for those rows.
2. **`Gender`**:
    - Converted `Male` and `Female` into numerical values:
        - `Male` → 1
        - `Female` → 0.
3. **`Opportunity Category`**:
    - Encoded the categorical values (e.g., `Event`, `Workshop`, etc.) into numerical codes for easier processing.

---

5. Renaming and Standardizing Columns

- Columns with special characters or spaces (e.g., `Graduation Date(YYYY MM)`) were retained but processed consistently for analysis.

6. Validation of Changes

1. **Duplicates**:
    - Checked for duplicate rows and dropped them if any were found.
2. **Outlier Detection**:
    - Reviewed numerical columns (`Skill Points Earned`, `Reward Amount`) for extreme or implausible values.
3. **Date Consistency**:

- Verified that `Opportunity End Date` was always later than or equal to `Opportunity Start Date`.
- Checked that `Apply Date` preceded or coincided with `Opportunity Start Date`.

Outcome

After cleaning:

- The dataset was free of missing or inconsistent values.
- All columns were appropriately formatted, with date and numeric fields converted for computational ease.
- Additional binary columns for skills allowed detailed analysis of individual skill distributions.
- Categorical variables were encoded to enable statistical and machine-learning-based approaches.

This meticulous cleaning process ensured the dataset was ready for advanced analysis and visualization.

4.2: User Data

The `user_data` DataFrame was meticulously cleaned to address inconsistencies, missing values, and data formatting issues. Below is a detailed step-by-step explanation of the cleaning process:

1. Handling Missing Data

1. **`Zip` (Zip Code)**:
   - Checked for missing values or invalid entries (e.g., empty or NaN values).
   - Retained rows with valid numeric zip codes while removing rows where the `Zip` consisted entirely of alphabets (non-numeric values).

2. Data Type Conversions

1. **Date Column**:
   - **`Sign Up Date`**:
     - The `Sign Up Date` column contained values in the format `2023-07-23T08:05:58.602Z`. This was converted to a standard `datetime` format displaying only the date and time:

        - Original: `2023-07-23T08:05:58.602Z`
        - Cleaned: `2023-07-23 08:05:58`.
2. **Numerical Columns**:
    - **`Gender`**:
        - Converted categorical values into numerical representation:
            - `Male` → 1
            - `Female` → 0.
    - **`isFromSocialMedia`**:
        - This boolean column was converted to numerical values:
            - `True` → 1
            - `False` → 0.

3. Expanding Categorical Data

1. **`PreferredSponsors`**:
    - The `PreferredSponsors` column contained lists of sponsor names (e.g., `["GlobalShala", "Grant Thornton China"]`). The following steps were applied:
        - Extracted individual sponsor names.
        - Created separate binary columns for each sponsor (e.g., `GlobalShala`, `Grant Thornton China`, etc.).
        - Populated these columns with **1** (sponsored by the entity) or **0** (not sponsored by the entity) for each row.
        - Handled missing or empty sponsor lists by filling all sponsor columns with **0**.

4. Removing Invalid Entries

1. **Zip Code Validation**:
    - Rows with invalid `Zip` values (comprising only alphabets) were removed to ensure that the column contained valid postal codes.

5. Renaming and Standardizing Columns

1. Ensured column names were uniform and descriptive:
    - For example, retained the name `Sign Up Date` but ensured proper formatting and consistent use throughout the analysis.

6. Validation of Changes

1. **Duplicates**:
    - Checked for and removed any duplicate rows to prevent redundancy.
2. **Outlier Detection**:
    - Reviewed numerical columns (e.g., binary columns like `isFromSocialMedia`) to ensure no unexpected or extreme values.
3. **Date Consistency**:
    - Verified that all `Sign Up Date` values followed the expected chronological order and were valid datetime objects.

Outcome

After cleaning:

- The dataset was consistent, with missing values handled appropriately.
- All columns were in the correct data types, such as numeric, datetime, or categorical.
- Binary columns for `PreferredSponsors` allowed detailed analysis of sponsor preferences.
- Invalid or non-numeric zip codes were eliminated, ensuring the integrity of geographical data.

This cleaning process ensured that `user_data` was prepared for accurate analysis and reliable visualizations.

5. Challenges Faced

One of the major challenges was handling the Data of the “Opportunity Sign UP and Completion” Dataset as it contains a lot of missing values which cannot be handled so easily especially while making up the visualizations.

6. Next Steps

In the upcoming week, we will try to get a better analysis of the data and take some more key insights out of it. The next week involves Data Preprocessing and Transformation which will involve cleaning the data and making it perfect for making out Dashboards. As the Data Cleaning is already done, so after some final insights gathering we will work on the wireframe for the Dashboard.

7. Conclusion

The dataset provides valuable insights into the demographics and sign-up behaviors of participants. The predominant sponsors and educational backgrounds of participants can help in tailoring marketing strategies or educational offerings. Further analysis with additional numerical data, such as sign-up dates or other relevant metrics, could provide more comprehensive insights into trends and patterns over time.