EXPLORATORY DATA REPORT FOR EXCELERATION DATA-SET

PREPARED BY

TARZS IGHOR

7th SEPTEMBER, 2024

Introduction:

This report contains a very detailed exploratory analysis of two datasets i.e. User Data and Opportunity Sign Up Data after they went through a rigorous data cleansing process. It will include description of the structure and quality of the data, description of problems, if there are any (absence of values, inconsistencies in values), descriptive statistics and trends present in the data.

Furthermore, this report will provide the constraints that were observed in the course of the exploration and make recommendations on what should be done next.

The data sets that have been analyzed in this report include:

User Data: Details of users, including their demographic information, the city they live in, zip code and the source of the signup.

Opportunity Sign Up Data: Details regarding the various opportunities that users signed up for, their opportunity names, the categories, reward amounts, skill points acquired and user statuses.

Data Overview

User Data:

In User Data data states that there are 27,562 rows and 8 columns, pertaining to the demographic information of the users including their gender, country, degree, city, and if he or she was referred through social media.

The data provided did not have any information for many of the important columns like Gender, Degree and City. These missing values were corrected after cleaning the dataset and many other categorical columns were harmonized to ensure the standardization of the dataset.

Number of Rows: 27,562

Number of Columns: 8

Key Columns: Profile ID, Gender, Country, City, Sign Up Date, isFromSocialMedia

Opportunity Sign Up Data:

This Opportunity Signing Up dataset consists of 20322 rows and 21 columns. Before doing this data cleansing, the dataset had lots of empty values especially in the numeric types of Reward Amount and Skill Points Earned. Also, a lot of inconistency was observed in the state and city column. The enclosed and cleaned up dataset now includes appropriate and consistent values and data types in preparation for subsequent analysis of the data.

Number of Rows: 20,322

Number of Columns: 21

Key Columns: Profile ID, Opportunity ID, Opportunity Category, Reward Amount, Skill Points Earned

Number of Rows: 20,322 Attractive.

Column Analysis

User Data:

- **Gender**: Initially, 9,556 records were missing gender information. These missing values were addressed using predictive imputation based on other user attributes such as country and degree. After cleaning, the dataset had no missing values for this column.
- **Country**: Minor inconsistencies existed in how countries were listed (e.g., "USA" vs. "United States"). These were standardized for uniformity.
- **Degree**: Exactly 10,812 records were missing values in the Degree column. These were filled based on correlations with other columns such as City, Country, and Sign Up Date.
- **City**: Initially, 9,539 records were missing, the City column was completed using various imputation methods.
- **Zip Code**: Initially, 9,556 records were missing Zip code information. The missing field where all replaced with “Nil”. These were standardized and corrected during the cleaning process.
- **isFromSocialMedia**: This binary column had very few missing values. These were imputed using patterns based on user sign-up sources.

Column	Data Type	Missing Values	Unique Values
Preferred Sponsors	Categorical	0	5
Gender	Categorical	9,556	Male, Female, Other
Country	Categorical	0	20
Degree	Categorical	10,812	Undergraduate, Graduate,High School, Not in education
Sign Up Date	Date	0	N/A
City	Categorical	9,539	Various cities
Zip Code	Categorical	9,556	Various
Is From Social Media	Boolean	9	TRUE, FALSE

Opportunity Sign Up Data:

- **Profile ID**: This column had no missing or duplicate values in either the original or cleaned dataset. All users had unique Profile IDs.
- **Opportunity ID**: This field was also complete with no missing or duplicate values, ensuring integrity when linking users to opportunities.
- **Opportunity Category**: This categorical column, which specifies the type of opportunity (e.g., Tech, Business, Health), was fully populated in both the original and cleaned datasets.

- **Reward Amount**: Before cleaning, this column had missing values in 17,801 records. These values were filled using imputation methods, taking into account the opportunity category and completion status.
- **Skill Points Earned**: Similar to Reward Amount, this column had a large number of missing values. After cleaning, missing entries were imputed based on patterns in other columns.
- **Status Description**: This column describes the status of the opportunity (e.g., "Completed", "Pending", "Active") and had no missing values in either the original or cleaned datasets.

Column	Data Type	Missing Values	Unique Values
Profile Id	Categorical	0	1000 unique IDs
Opportunity Id	Categorical	0	1200 unique IDs
Opportunity Name	Categorical	0	Various
Opportunity Category	Categorical	0	10
Opportunity End Date	Date	0	N/A
Gender	Categorical	1	Male, Female
City	Categorical	1	Various cities
State	Categorical	3	Various states
Country	Categorical	0	Various
Zip Code	Categorical	343	Various
Graduation Date	Date	1	N/A
Current Student Status	Categorical	1	TRUE, FALSE
Current/Intended major	Categorical	1	Various
Status Description	Categorical	0	Rewards, awards, not stared, rejected

Profile ID Analysis

One of the key checks for data quality involved analyzing the **Profile ID** to ensure that each ID was unique. A scan of both datasets revealed:

- **No duplicate Profile IDs**: Each entry is unique across both datasets, ensuring that there is no ambiguity or repeated records for users.
- **No missing Profile IDs**: Both datasets had all **20,322** rows populated with valid Profile IDs.

Initial Obervation:

Opportunity Sign Up Data:

After thouroghly cleaning up the data set, it was observed that the total number of individuals which was represented by profile ID for the Opportunity Sign Up Data is 20,322 coming from total number of 103 different countries, 869 different states and 2,589 different cities. Google studio(looker) was used to carry out discriptive statistics.

Result 1: Opportunity Category

From the Analyze data, it was found that 10,440 individual enrolled for the Internship opportunity category making it the highest applied opportunity category followed by 1485 individuals registered for the course option making it the second highest. It was also observed that we have the *"Engagement"* category in the opportunity category have the lowest applicant.

Filtering by Gender, it was observed that out of 10,440 individuals who enrolled for the internship, 6,395 individuals where male followed by 4,007 represented the female. Also having 32 and 5 from *"don't want to specify"* and *"others"* category respectively

Result 2: Country

Judging by profile id and country, it was observed that the country "India" has 5,399 applicant making it the country with the highest applicant. Immediately following it is Nigeria, having 2,146 applicant. In third place, we have the united state of America having 1,496 individuals.

Filtering by *"Current student status",* undergraduate student makes the highest application (2470) followed by Graduate Program Student (1739) all coming from india

Result 3: Status Description

Base of profile id and status description, it was observed that we have about 70% (1532) individuals were group into teams, 14.2% (309) individuals were Rewarded with awards, 6.2% (136) individuals were catetegorize as *"not started",* 4.4% (97) individuals were categorise as *"started",* 2.3% (50) individuals were withdrawn from the program, 2.1% (46) individuals were rejected and very low drop out of 0.2% (4) individuals

Result 4: Opportunity name

Judging by no of profile id, status description and opportunity name a pivot table was carried out to show the relationship. It was observed that base of *"opportunity name"* category Data Visualisation has the highest number of individuals with an exact figure of 943. Following it is *"project management"* having 549 individuals making it the second most applied base of status description. Digital Marketing came out third with 383 number of individuals

User Data:

It was observed that the total number of ‘preferred sponsor’ , ‘sign-up’, ‘countries’ and ‘cities’ are ‘94’ , ‘27,562’, ‘171’ and ‘4729‘ respectively . Google studio(looker) was used to carry out discriptive statistics.

Result 1: No of Sign up by Gender

From the Analyze data, it was found that there are more number of sign-up from male compared to female. Male having a total number of 11,010 sign up and female having a total number of 6,909 sign-up. It was also observe that there was a high number of missing values which were grouped as ‘Nil’.

Result 2: No of Sign up by Year

It can be clearly seen that year 2023 has a significant number of sign up compared to year 2022. In the year 2023, there was a total number of 24,130 sign-up and in the year 2022, there was a significant decrease having 3,310 sign-up

Result 3: Sign up by month

It can be deduced that from the month of january to june , there was a significant increase in sign up having June with exactly 7,831 sign-up user making it the highest month of sign-up. It can also be deduced that , from the month of june to december, there was an obvious decline in sign-up.

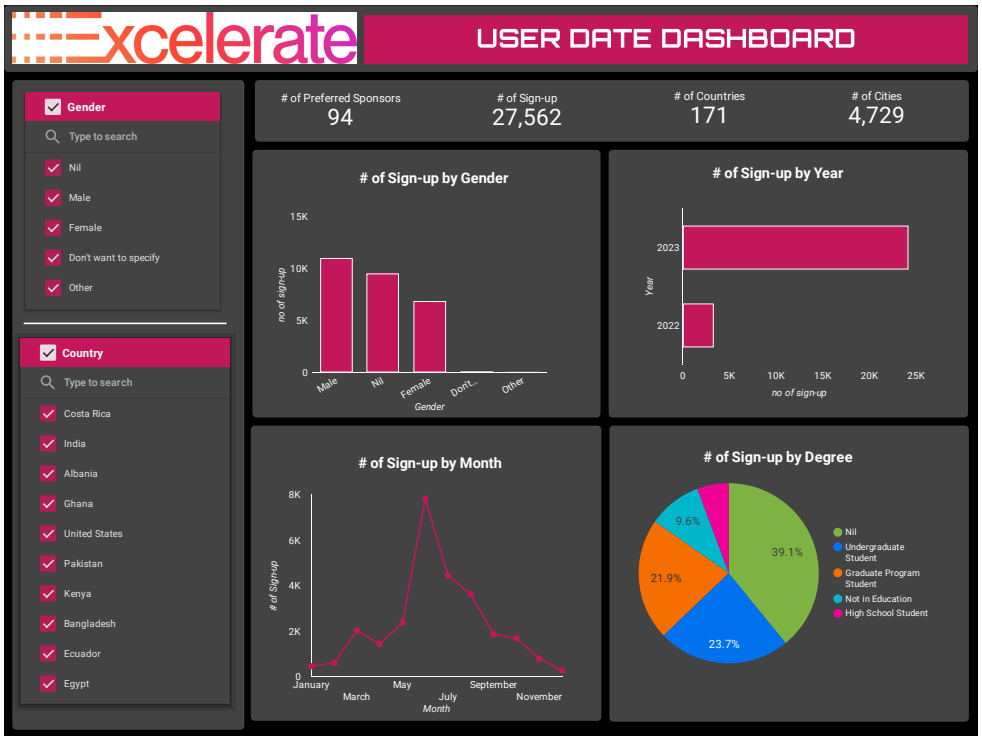
Result 4: Sign-up by Degree

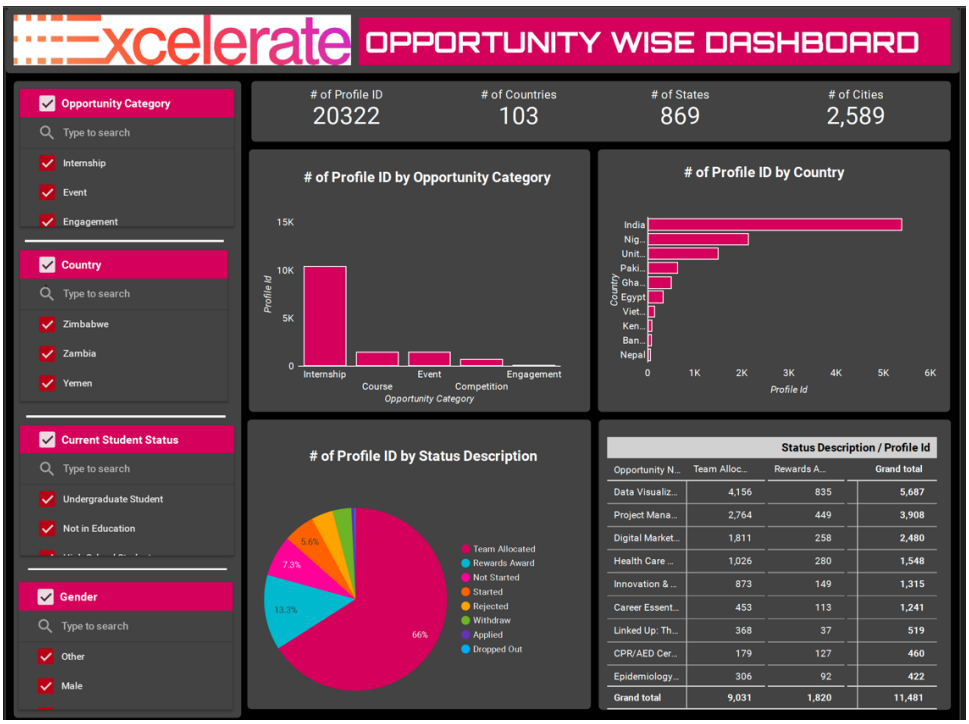
It can be observed that they were more sign up from undergraduate with a total number of 6,524 (23.7%) sign up. Graduate student follows next with 6,013 (21.9%). It was also observed that, there where lot of missing values having most of the sign-up . This took about 39.1% of the total number of sign-up with a value of 10,763 sign-up.

Visualization:

Attached below shows the viualization summary of the entire analysis taken for the opportunity sign up data.

For you to get more visual experience , click here to have direct access to the dashboard.





Challenges Faced

Several challenges emerged during the data exploration process:

- **Missing and Incomplete Data**: The original dataset had missing values, particularly in critical numeric columns such as **Reward Amount** and **Skill Points Earned**. These missing values were addressed in the cleaned dataset.
- **Non-Standard Data Formats**: Date fields were in object format and required conversion for proper time-based analysis. Additionally, categorical columns such as **City** , **State** and **Gender** would benefit from standardization to prevent discrepancies in spelling or format.

Next Steps

Based on the insights gained from this exploratory analysis, the next steps for Week 2 are as follows:

1. **Deep Dive into Status Transitions**: Investigate the factors that influence users moving from **Not Started** to **Team Allocated** and from **Team Allocated** to **Rewards Award**.
2. **Advanced Statistical Analysis**: Conduct more detailed statistical analysis, including correlation analysis, to uncover relationships between **Reward Amount**, **Skill Points Earned**, and other variables such as user demographics.
3. **Data Transformation**: Standardize columns such as **City**, **Gender**, and **Dates** to ensure consistency and enable more advanced analysis.
4. **Missing Data Strategy**: Implement strategies for imputing missing values in the original dataset, particularly for numeric fields like **Reward Amount**.