**DATA VISUALIZATION ASSOCIATE INTERNSHIP**

**GROUP 18A**

**WEEK 1 DELIVERABLE**

**Team Coordinator: Mr. Kolawole Oparinde**

**1. INTRODUCTION**

This report presents an in-depth Exploratory Data Analysis (EDA) of two datasets: **User Data** and **Opportunity Sign-Up Data**, sourced from the Excelerate platform. The primary goal is to extract meaningful insights that will aid in decision-making, optimization, and enhancing user engagement on the platform.

## Analysis Data Set

* User Data
* Opportunity Sign up Data

# 2. DATASET OVERVIEW

## 2.1 User Data

This analysis center on “User Data”, featuring 8 columns and 27,563 rows. The columns include “Preferred Sponsors”, “Gender”, “Country”, “Degree”, “Sign Up Date”, “City”, “Zip”, and “IsFromSocialMedia”. This dataset encompasses non-identifying information about every user who has ever created an account on Excelerate. The data is comprehensive, covering all users, regardless of their engagement with specific opportunities. Each row represents a unique user, and the dataset provides a holistic view of the user base. However, the data set featured users who registered more than one Preferred sponsor, with a sum of unique value for preferred sponsors to be 94, country with a unique value of 170, and 4015 for cities. Degree covered across Graduate Program Students, High School Student, Not in Education, Undergraduates, and several blank cell which was classified as unknown during the process of data cleaning.

## 2.2 Opportunity Sign Up Data

This analysis centers on the “Opportunity Sign Up Data”, a dataset comprising 21 columns and 20,322 rows. The column includes “Profile Id”, “Opportunity Id”, “Opportunity Name”, “Opportunity Category”, “Opportunity End Date”, “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”, “Skills Earned”. This dataset focuses on non-identifying user information related to learners who have engaged with specific opportunities on Excelerate. Each row corresponds to a learner who has signed up for a particular opportunity with having the possibility of a user signing up for more than one opportunity with the same profile Id. Effective data cleaning, descriptive statistics, visualization was done on these data set to ensure proper integrity on the Exploratory Data Analysis.

# 3. COLUMN ANALYSIS

Proper column analysis is carried out in the section for both the user data and the opportunity sign up data. The major focus is to identify unique values, missing values, errors were applied.

## User Data

**3.1.1 Preferred Sponsors:**

Data Type: Text

Description: this column indicates the user choice of preferred sponsor. In this column, each user can choose more than one preferred sponsor.

Unique Value: 94

**3.1.2 Gender:**

Data Type: Text (Categorical)

Description: This column reflects the demographic choice of the user in respects to gender with an option of whether or not to disclose their gender. Several missing values were found during the data cleaning process and was identified as “Unknown”

Missing Values: 9625

**3.1.3 Country:**

Data Type: Text

Description: This shows the geographical location of the user in terms of country.

Unique Value: 170

Missing Values: No

**3.1.4 Degree:**

Data Type: Categorical

Description: This column reflects the demographic choice of the user with respect to their academic level. This is not a mandatory field. Missing values were found and they were identified as “Unknown”

Missing Values: 10812

**3.1.5 Sign Up Date**

Date Type: Date

Description: This shows the date of the user signup. No missing values was identified Missing Values: No

**3.1.6 City**

Data Type: Text

Description: This column reflects the city of the user. This information was voluntarily provided by the user. Missing values were found and they were identified as Unknown

Unique Value: 4002

Missing Values: 9558

**3.1.7 Zip**

Data Type: Text

Description: This column shows the zip code of the city of every user. This is not a compulsory field. Missing values were identified as Unknown.

Missing Values: 9536

**3.1.8 isFromSocialMedia**

Data Type: Boolean

Description: this shows if the user signed through google logins which indicates True, or signed up manually which indicates False.

**3.2 Opportunity Sign Up Data**

**3.2.1 Profile ID:**

Data Type: Alphanumeric Unique Identifier: Yes Unique Values:

**3.2.2 Opportunity ID:**

Data Type: Alphanumeric

Unique Identifier: Yes No missing values identified.

**3.2.3 Opportunity Name:**

Data Type: Categorical

No missing values identified.

Unique Values: 33 opportunities with varying frequencies.

**3.2.4 Opportunity Category:**

Data Type: Categorical

No missing values identified.

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

**3.2.5 Opportunity End Date:**

Data Type: Datetime

No missing values identified.

Dates appear to follow a standardized format.

**3.2.6 Gender:**

Data Type: Categorical

One missing value identified.

Unique Values: Male, Female, Unknown

**3.2.7** City:

Data Type: Text

Unique values: No Missing Values:

**3.2.8** State:

Data Type: Text

Unique values: No

**3.2.9** Country:

Data Type: Text

Missing Values: No

**3.2.10** Zip Code:

Data Type: Text

Missing Values: Yes

**3.2.11** Graduation Date (YYYY MM):

Data Type: Date

Missing Values: No

**3.2.12** Current Student Status:

Data Type: Categorical

Missing Values: No

**3.2.13** Current/Intended Major:

Data Type: Text

Missing Value:

**3.2.14** Status Description:

Data Type: Categorical

Missing Value:

**3.2.15** Apply Date:

Data Type: Date

**3.2.16** Opportunity Start Date:

Data Type: Date

**3.2.17** Reward Amount:

Data Type: Numbers

Missing Values: Yes

**3.2.18** Badge Id:

Data Type: Alphanumeric

Missing Values: Yes

**3.2.19** Badge Name:

Data Type: Text

Missing Values: Yes

**3.2.20** Skill Points Earned:

Data Type: Numbers

Missing Values: Yes

**3.2.21** Skills Earned:

Data Type: Text

Missing Values: Yes

# 4. PROFILE ID ANALYSIS

Profile ID is only found in the Opportunity sign Up Data

|  |  |
| --- | --- |
| **Profile ID Analysis** |  |
| Column Name | Profile ID |
| Total Record | 20322 |
| Unique ID | 7583 |
| Duplicates ID | 12739 |
| Missing ID | 0 |

**Key observation:**

The profile ID shows the identification of users’ interaction on the platform. This shows the number of unique identification and also the number of times users tends to interact and signup to an opportunity.

# 5. OPPORTUNITY STATUS DISTRIBUTION

The status distribution column is found in the opportunity sign Data, this gives a summary analysis of the distribution of different statuses.

|  |  |
| --- | --- |
| Status Description | Count of Status Description |
| Team Allocated | 14206 |
| Reward Award | 2521 |
| Not Started | 1324 |
| Started | 810 |
| Rejected | 726 |
| Withdraw | 622 |
| Applied | 89 |
| Dropped Out | 24 |

# 6. BASIC STATISTICS

**Reward Amount (Opportunity Sign up data):**

|  |  |
| --- | --- |
| Mean | 134.1334514 |
| Standard Error | 3.390930529 |
| Median | 0 |
| Mode | 0 |
| Standard Deviation | 483.3949576 |
| Sample Variance | 233670.685 |
| Kurtosis | 17.19964004 |
| Skewness | 4.207529769 |
| Range | 2500 |
| Minimum | 0 |
| Maximum | 2500 |
| Sum | 2725860 |
| Count | 20322 |

**Skill Point Earned (Opportunity Sign Up Data):**

|  |  |
| --- | --- |
| Mean | 147.2462 |
| Standard Error | 2.9165228 |
| Median | 0 |
| Mode | 0 |
| Standard Deviation | 415.7664 |
| Sample Variance | 172861.7 |
| Kurtosis | 5.67271 |
| Skewness | 2.660474 |
| Range | 1776 |
| Minimum | 0 |
| Maximum | 1776 |
| Sum | 2992338 |
| Count | 20322 |

**Key Observations:**

# 7. INITIAL OBSERVATION

**7.1 User Data**

* Structure of Data: These Data set comprises of 8 column and 27,563 rows.

1. Gender Column

The gender column had a lot of empty cells which comprises of 8253 empty cells which was classified as unknown value.

1. The data type of the date column was not having the accurate data format.

**7.2 Opportunity Sign Up Data:**

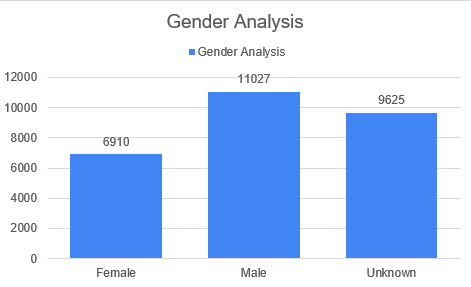
Few observations were made during the EDA process.

1. In The opportunity category column, it shows that users were more interested in internship
2. The applied date was not in the right date format
3. Not every participant got the badge
4. "City Names” contained certain typos and inconsistencies, which was all Cleaned

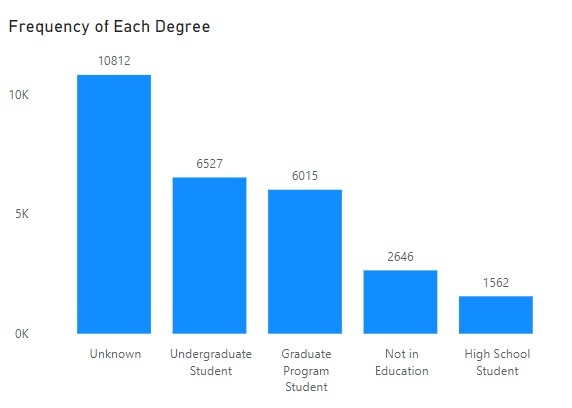
**8. VISUALIZATIONS**

**8.1 User Data**

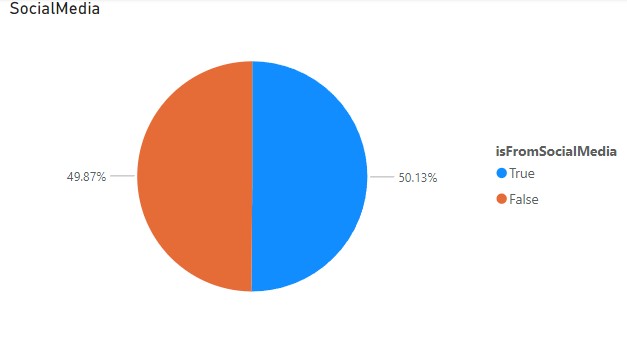
# Visualization for the Gender Column



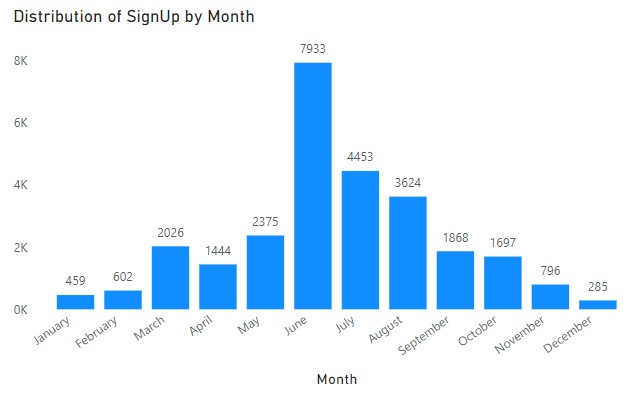
# Visualization for Degree Column



# Visualization For isFromSocialMeadia

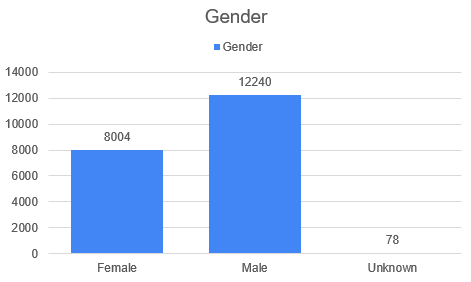


* **Visualization For Signup by Month**

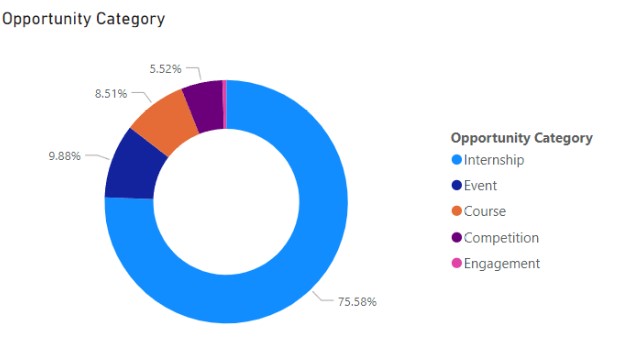


**8.2 Opportunity Data**

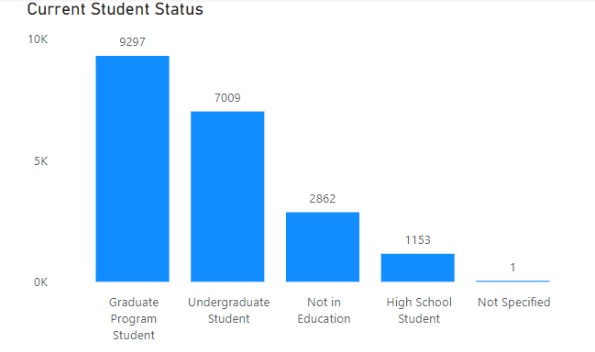
# Visualization For Gender



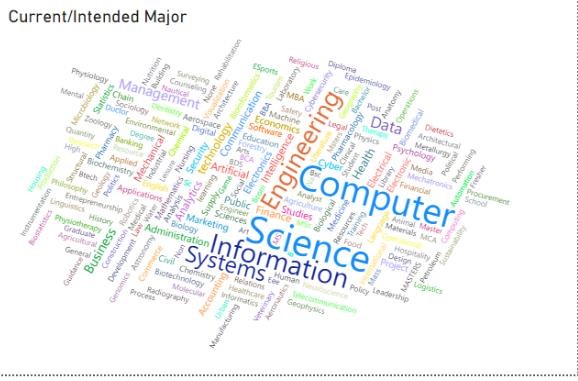
# Visualization For Opportunity Category



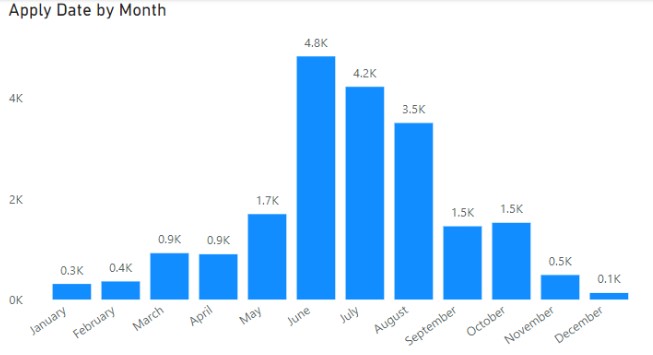
# Visualization For Current Student Status



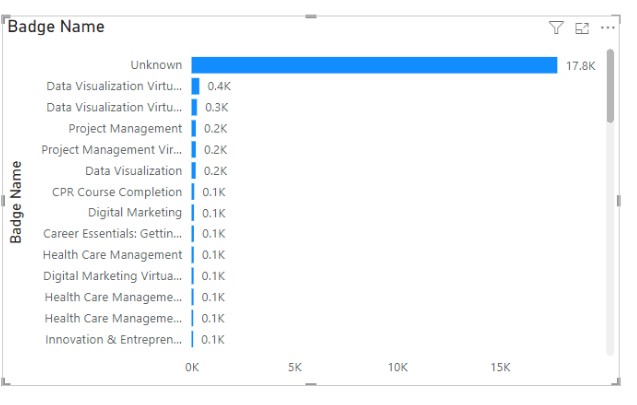
# Visualization For Current/Intended Major



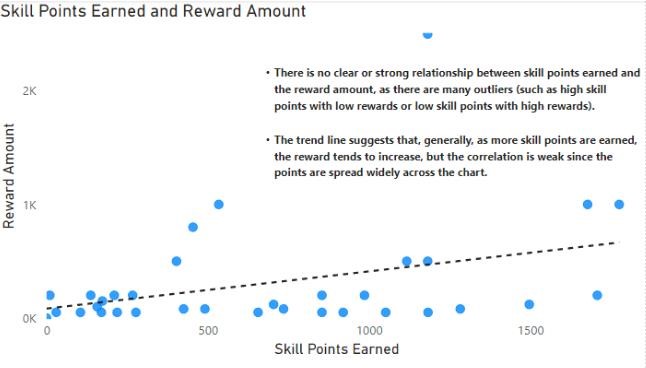
# Apply Date by Month



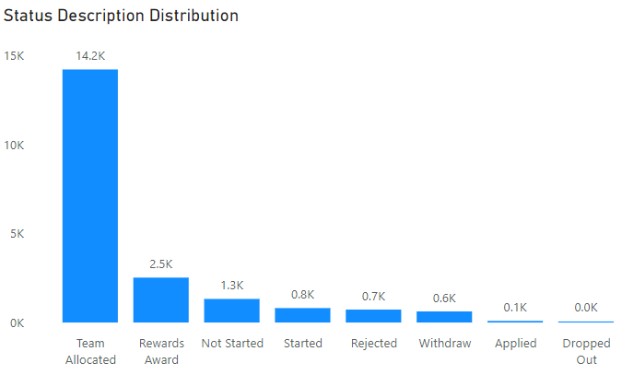
# Badge Name



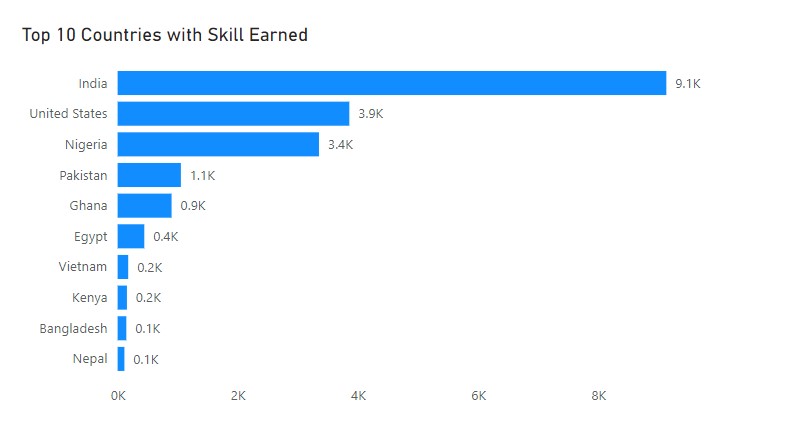
# Skill Points Earned and Reward Amount



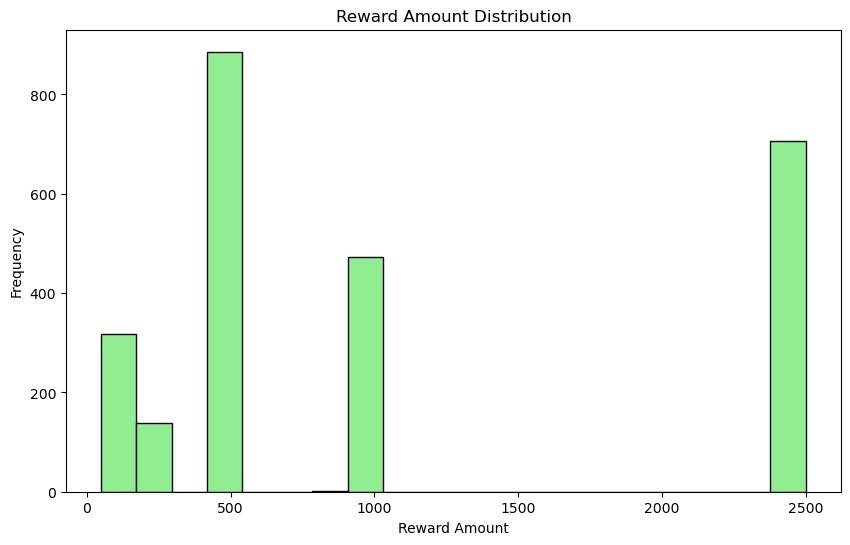
# • Status Description Distribution



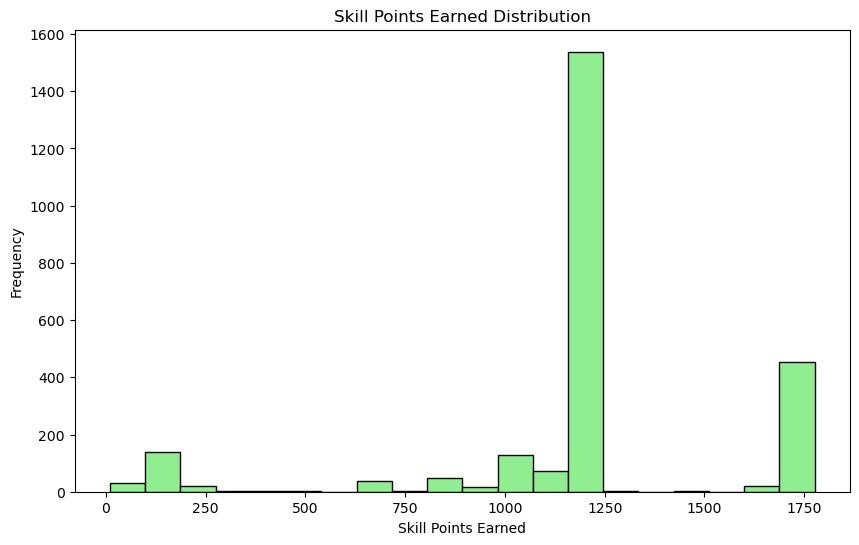
# Top 10 Countries with skilled earned



# Reward Amount Distribution



* **Skill Points Earned Distribution**



**9. CHALLENGES FACED**

* **Incomplete Dataset**

Several columns contained empty cells or null, making the data inaccurate for proper analysis.

* **Data Types**

The datatypes of some columns where in an inaccurate format.

# 10. NEXT STEP

Creation of insightful dashboard that will serve as a pointer toward decision making and optimization where necessary indicating important indicators such as Platform Activities, Global Reach, Opportunity Popularity, Completion Trend, Demographic Analysis, Skill Development, Scholarship Impact.