Exploratory Data Analysis (EDA) Report

2024

Contents

List of Tables

List of Figures

Introduction

1.1 Purpose

The purpose of performing the EDA (Exploratory Data Analysis) is to conduct initial data set review before and after the data cleaning and validation process. We are able to understand the data structure, distribution, and patterns. This includes understanding the type of variables and relationships between them.

Understanding anomalies or outliers thus drawing attention to data entries that could affect the result of the data. Before performing real data analysis, EDA allows us to test initial hypotheses and refine them based on the observed data.

In summary, EDA is essential for making informed decisions about data processing, analysis, ensuring the reliability and validity of our conclusions drawn from the dashboard for the non-technical team.

1.2 Datasets

I am working with two datasets which are user data and opportunity sign up and completion data. **User data** covers all users regardless of their engagement with specific opportunities. Those users have created an account with Excellerate.

Opportunity sign up and completion data focuses on non-identifying user information related to learners who have engaged with specific opportunities on Excellerate. Learners can sign up to multiple opportunities.

The two datasets will be used throughout in performing Exploratory Data Analysis (EDA). In this report, I will be referencing the two datasets.

Data Overview

2.1 Data info

2.1.1 User data

User data have 27562 entries with 8 columns. The columns have the below information:

Column	Non-Null Count	Dtype
PreferredSponsors	27562	object
Gender	18027	object
Country	27500	object
Degree	16750	object
Sign Up Date	27562	object
City	18028	object
Zip	18018	object
IsFromSocialMedia	27553	object

Table 2.1: User Data Information

All the data items are string object data types, some of them having ‘NaN‘ values.

2.1.2 Opportunity sign up and completion data

Opportunity sign up and completion data have 20322 different entries with 21 columns. The table below has the columns information:

Column	Non-Null Count	Dtype
Profile Id	20322	object
Opportunity Id	20322	object
Opportunity Name	20322	object
Opportunity Category	20322	object
Opportunity End Date	20322	object
Gender	20321	object
City	20321	object
State	20308	object
Country	20322	object
Zip Code	20309	object
Graduation Date (YYYY MM)	20321	object
Current Student Status	20321	object
Current/Intended Major	20278	object
Status Description	20322	object
Apply Date	20322	object
Opportunity Start Date	19518	object
Reward Amount	2521	float64
Badge Id	2521	object
Badge Name	2521	object
Skill Points Earned	2521	float64
Skills Earned	2521	object

Table 2.2: Opportunity Data Information

As seen from the above table, reward amount and skills points earned are float object while the rest of the columns are string objects.

2.2 Summary statistics

2.2.1 User data

Statistic	Value
Preferred Sponsors count	27562
Preferred Sponsors unique	94
Preferred Sponsors top	["GlobalShala", "Grant Thornton China", "Saint L...]
Preferred Sponsors freq	22011
Gender count	18027
Gender unique	4
Gender top	Male
Gender freq	11027
Country count	27500
Country unique	169
Country top	India
Country freq	11893
Degree count	16750
Degree unique	4
Degree top	Undergraduate Student
Degree freq	6527
Sign Up Date count	27562
Sign Up Date unique	27561
Sign Up Date top	2022-10-30T17:25:54.072Z
Sign Up Date freq	2
City count	18028
City unique	4727
City top	Hyderabad
City freq	743
Zip count	18018
Zip unique	7453
Zip top	63108
Zip freq	629
Is From Social Media count	27553
Is From Social Media unique	2
Is From Social Media top	True
Is From Social Media freq	13811

Table 2.3: Summary Statistics for User Data

The above table shows a full summary statistics of the user data. Preferred sponsors have a count of **27562** and **94** unique sponsors. There are **18027** gender with males being on top. In the data there are 169 unique countries with India being at the top. There are more undergraduate users compared to any other level of education. In the user data there are 4727 different cities with **Hyderabad** being the top city with most users. Finally either a **True** or **False** that the user is from social media.

2.2.2 Opportunity sign up and completion data

	Reward Amount	**Skill Points Earned**
count	2521.000000	2521.000000
mean	1081.261404	1186.964697
std	927.251398	399.172150
min	50.000000	10.000000
25%	500.000000	1182.000000
50%	500.000000	1182.000000
75%	2500.000000	1182.000000
max	2500.000000	1776.000000

Table 2.4: Summary Statistics for Reward Amount and Skill Points Earned

Reward amount have a mean of **1081.26** with a standard deviation of **927.25**. Majority of the reward points lie between lower and upper quartile.Minimum reward point is **50** with a maximum of **2500**.

Skills points earned have a mean of **1186.96** with a standard deviation of **399.17**. Minimum skills points earned is **10** with a maximum of **1776**.

Column Analysis

3.1 User dataset

1. The first column is the preferred sponsors column. This column has no missing value with a string data type.

2. The gender column has 9535 missing values.We saw that there are 11027 males, 6910 females 75 of the people did not specify their gender while 15 choose others. We will drop those missing values(it is not easy to know the missing gender). The table below shows value counts of the gender column after dropping the 'NaN' values.

Gender	**Count**
Male	5826
Female	3755
Don't want to specify	69
Other	14

Table 3.1: Gender Distribution after dropping null values

3. The country column, we noted that some countries were not well formatted in accordance to standard country names. The table below shows a sample of the original name and corrected names:

No.	**Original Name**	**Corrected Name**
1	Tanzania, United Republic of Tanzania	Tanzania
2	Iran, Islamic Republic of Persian Gulf	Iran
3	Korea, Republic of South Korea	Korea
4	Libyan Arab Jamahiriya	Libya
5	Cote d'Ivoire	Côte d'Ivoire

Table 3.2: Original and Corrected Country Names

4. The degree column had 1395 missing values,this is after dropping the missing values in the gender column. It will be wise to drop them.

5. Changing the string **sign up date** into datetime object

6. In the from social media column there are only 4 null values to be dropped.There are two unique values in this column with **True** and **False**.

3.2 Opportunity sign up and completion data

1. There are zero null values in the opportunity name and opportunity category columns after removing duplicate values from profile id. Below are tables showing value count of the two columns:

Opportunity Name	Count
Data Visualization	3656
Project Management	2215
Digital Marketing	2083
Career Essentials: Getting Started with Your Professional Journey	1235
Health Care Management	948

Table 3.3: Opportunity Names and Counts

Opportunity Category	Count
Internship	9503
Course	1320
Event	464
Competition	192
Engagement	1

Table 3.4: Opportunity Categories and Counts

2. Changing the string Opportunity Start date data type into datetime object.

3. There are 365 opportunity start date instances which have missing values, We choose not to drop them because this individuals may have been in the system doing the course before it started and so they may not have a starting date. We filled the missing values with the most present date of entry or opportunity start date.

4. We also noted that current/intended major have some missing values so we can add them to "others" options, those which are missing.

5. Since there are many courses, some not valid courses in the dataset, we included the top ten most sought courses then put all the other together in ’Others’ group.

Profile ID Analysis

Profile id column in our dataset is a very important column because it will help us identify any repeated entry in the rows called duplicates. Each person should have a unique profile id.

- As seen, there are 11481 profile id and 33 opportunity id.
- There are 8841 duplicates in the profile id column.
- 20289 entries which are duplicate in the opportunity id column .
- remember that there are 11481 different profile ids and 33 unique opportunities seen. Keep in mind that we have 20322 total entries in our opportunity sign up and completion dataset.

Opportunity Status Distribution

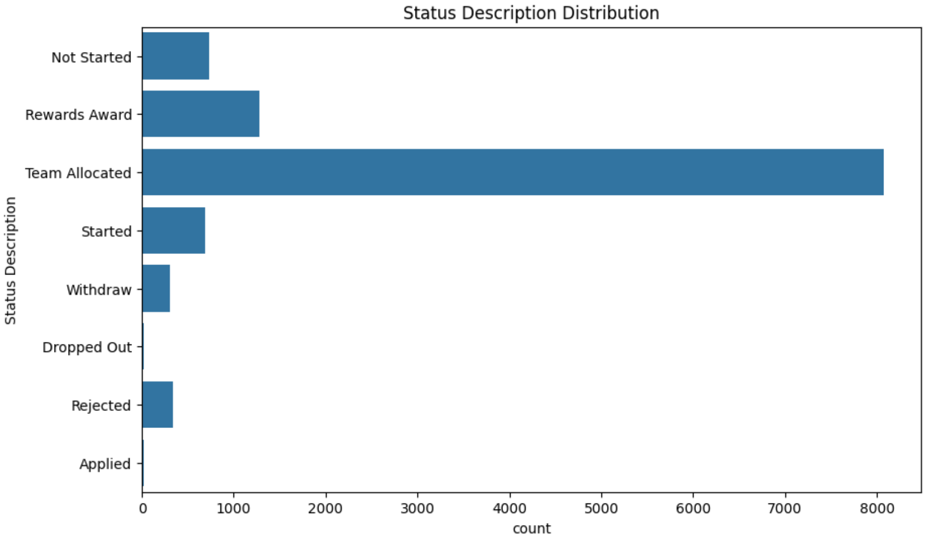

Figure 5.1: Opportunity status distribution

As seen from the above status description distribution figure, we can see that:

1. Team allocated status have more peple(**more than 8,000**).
2. There is a significant number of people who dropped out or rejected.
3. More than 1,000 people were rewarded after completing all the tasks.

Metric	Value
count	11480
unique	8
top	Team Allocated
freq	8077

Table 5.1: Summary Statistics

Status	**Count**
Team Allocated	8077
Rewards Award	1284
Not Started	732
Started	693
Rejected	340
Withdraw	311
Applied	26
Dropped Out	17

Table 5.2: Opportunity Status and Counts

Basic Statistics

let us understand the numerical data in our opportunity cleaned dataset. Reward amount has a mean of 116.3 with a deviation of 429 where the maximum reward amount is 2,500. Skill point earnded have a mean of 136.5 with a deviation of 412.72 and maximum skills point earned being 1776. This information is shown in the table below:

	Reward Amount	**Skill Points Earned**
count	11480.000000	11480.000000
mean	116.302265	136.514199
std	429.710641	412.729419
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	2500.000000	1776.000000

Table 6.1: Reward Amount and Skill Points Earned Statistics

Initial Observations

7.0.1 Initial Observations

In the initial review of the datasets, several key points were observed:

1. **Data Completeness**: The user data dataset has 27,562 entries, while the opportunity sign up and completion data dataset contains 20,322 entries. There are some columns with missing values, particularly in the user data's `Gender` and `Degree` columns.

2. **User Demographics**: The majority of users are from India, with a significant proportion being undergraduate students. Male users are the most represented gender in the dataset.

3. **Opportunity Engagement**: Users have engaged with a wide variety of opportunities, with reward amounts and skill points earned showing considerable variability. The majority of reward points lie between the lower and upper quartiles, with some high-value outliers.

4. **Sign-Up Dates**: The sign-up dates for users span a wide range, indicating continuous engagement with the platform over time. However, there is a notable spike on certain dates which may warrant further investigation.

5. **Data Types and Structures**: Both datasets primarily consist of categorical and string data types, with some numerical data in the opportunity dataset (e.g., reward amounts and skill points).

7.0.2 Area of Interest

Our primary areas of interest for further analysis include:

1. **User Engagement Trends**: Analyzing trends in user sign-ups and opportunity engagements over time to understand peak activity periods and potential factors driving these trends.

2. **Demographic Analysis**: Investigating the demographic distribution of users, including gender, country, and educational background, to identify any patterns or disparities.

3. **Opportunity Effectiveness**: Assessing the effectiveness of different opportunities by examining the distribution of reward amounts and skill points earned, and correlating these with user demographics.

4. **Social Media Influence**: Exploring the impact of social media on user engagement by analyzing the `IsFromSocialMedia` column.

Visualizations

8.1 User data

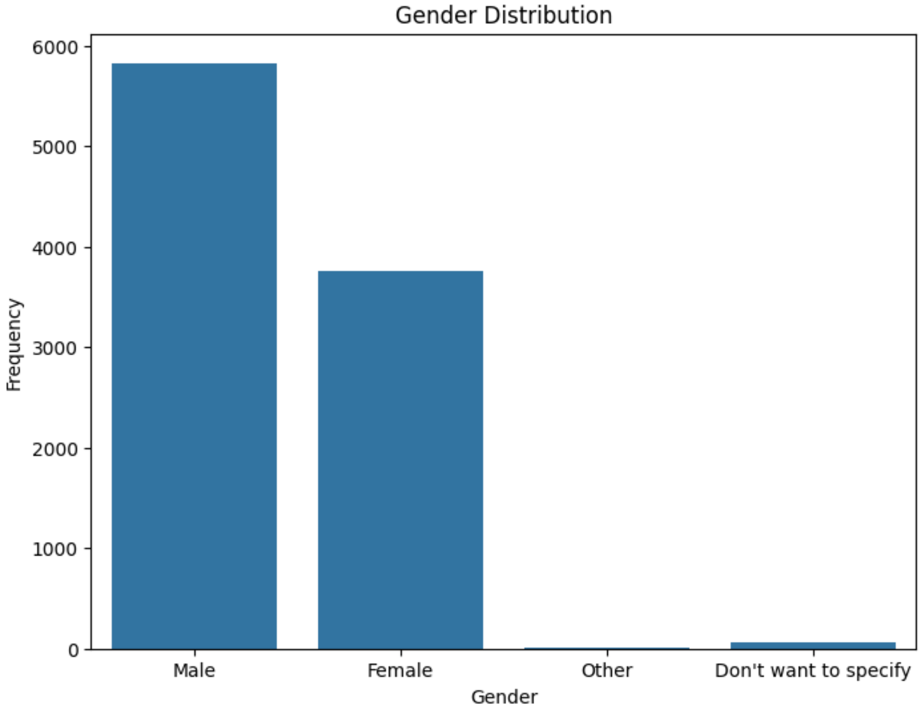

Figure 8.1: Gender distribution figure

As seen from the chat above there are more male users compared to females, only a few who did not want to specify their gender

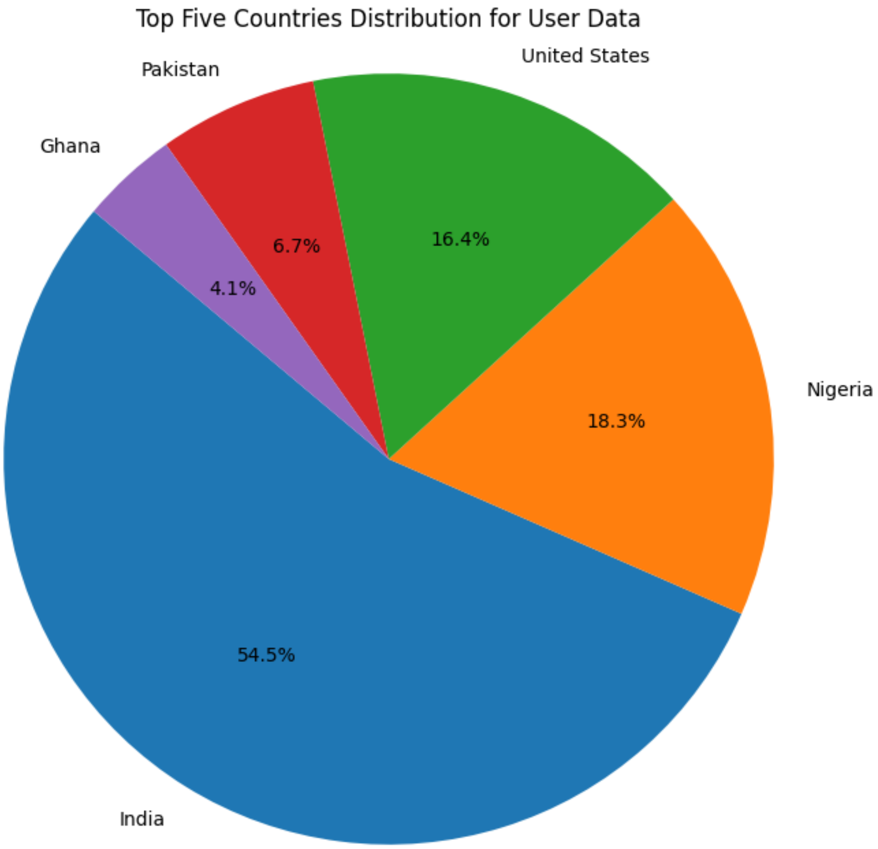

Figure 8.2: top countries distribution

As seen more than half of the users are from India.

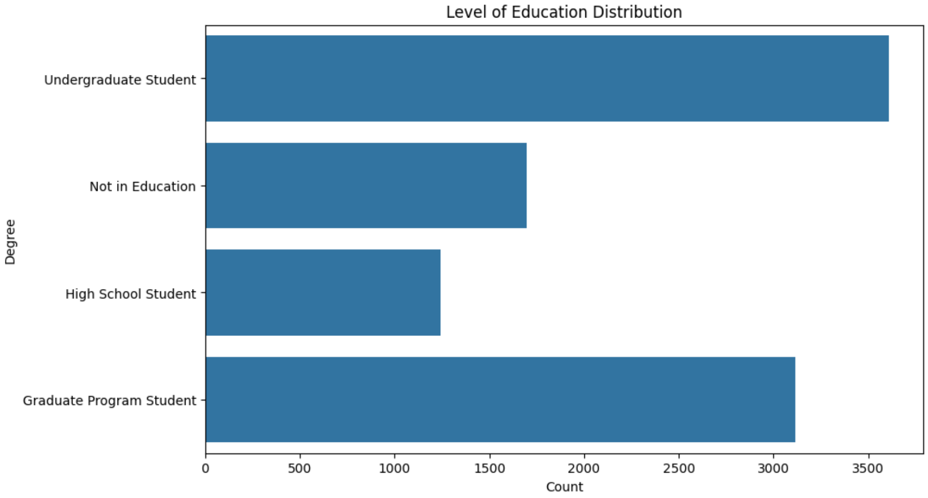

Figure 8.3: level of education

Majority of the user are undergraduate students while the least are in high school.

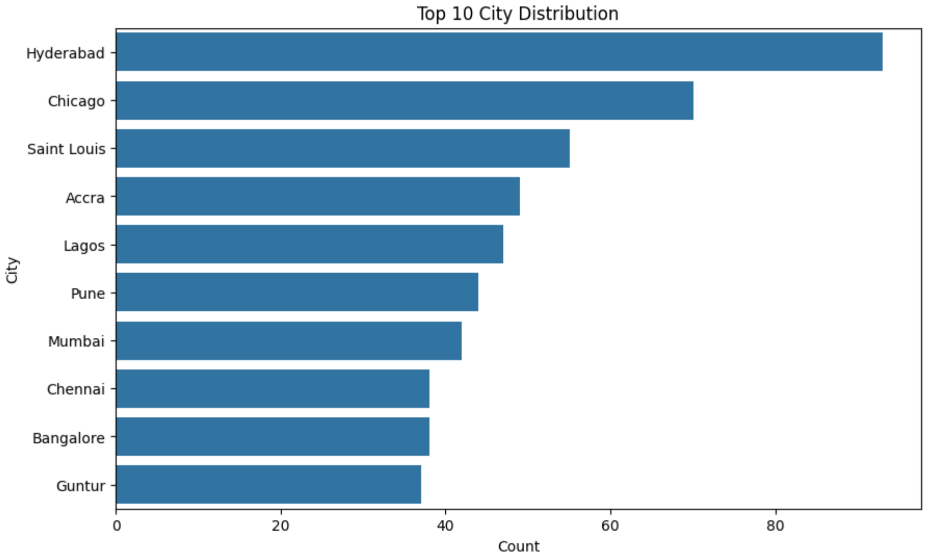

Figure 8.4: top cities

Hydrabad has more than 80 users and thus holding more users as seen from the above chart

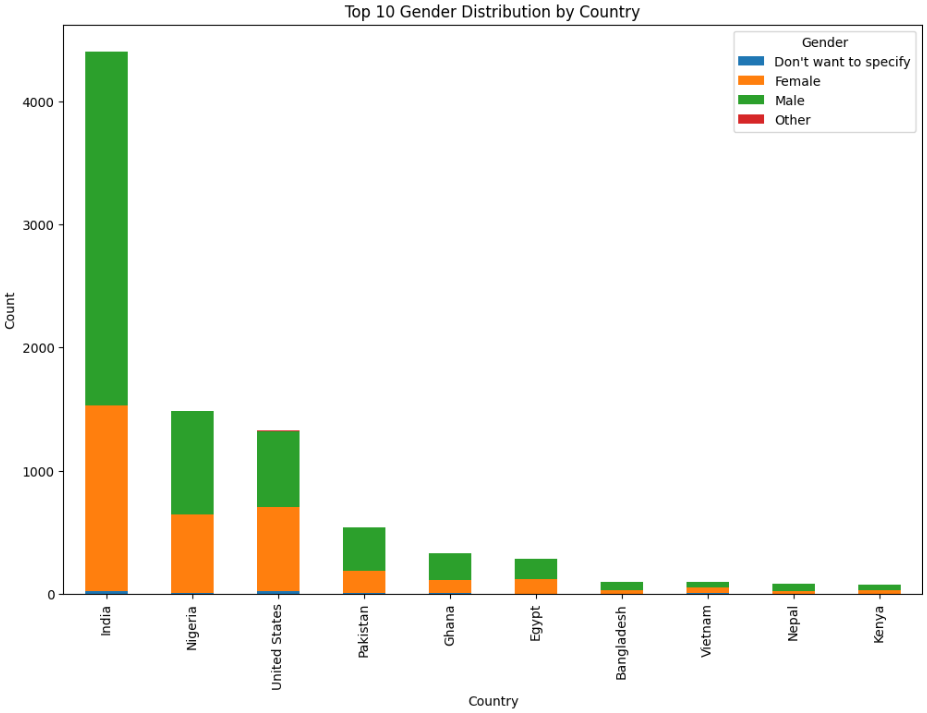

Figure 8.5: Gender distribution by county

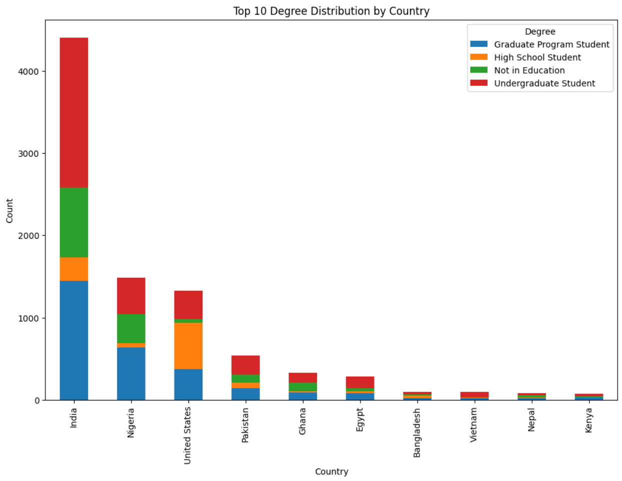

Figure 8.6: degree distribution by country

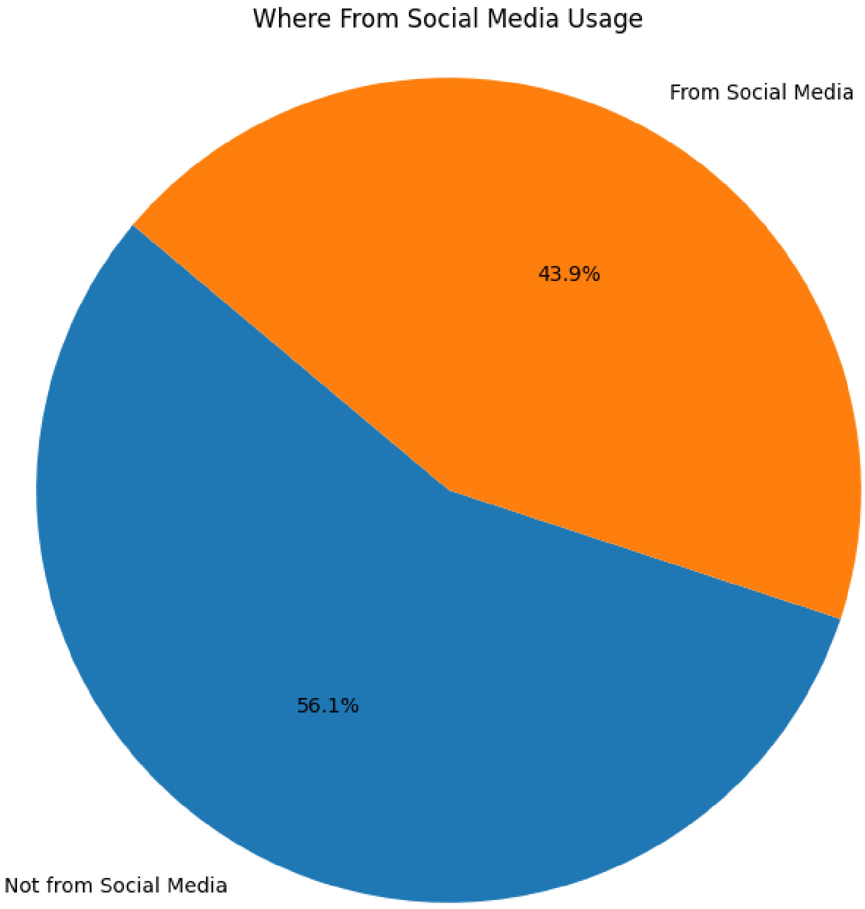

Figure 8.7: is from social media

We can see that 56.1% of the users are not from social media.

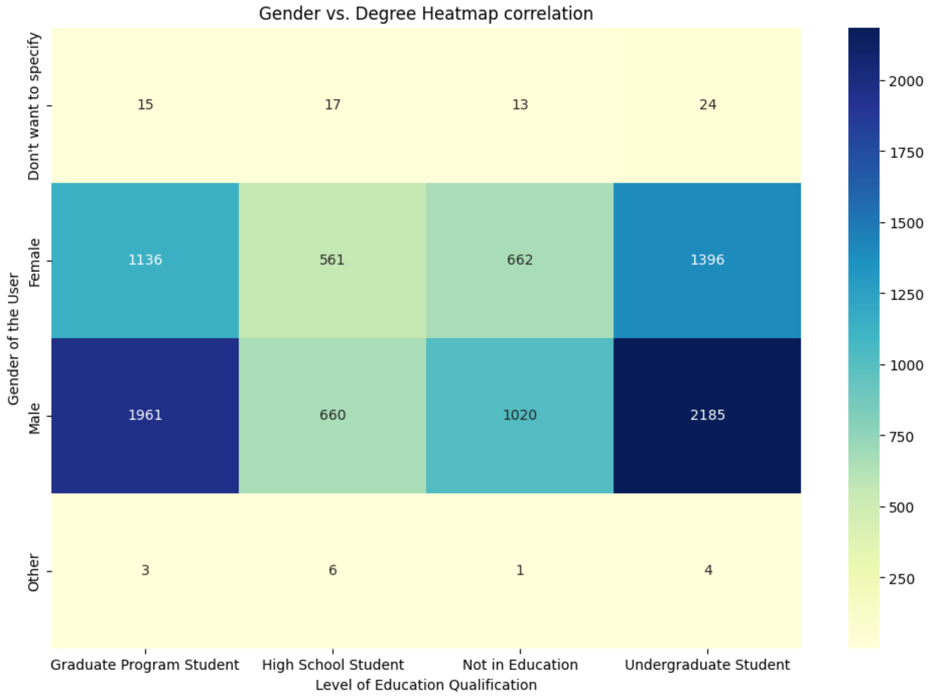

Figure 8.8: Degree and gender correlation

There is a very high correlation between male users and both graduate and undergraduate qualifications or level of education they are at.

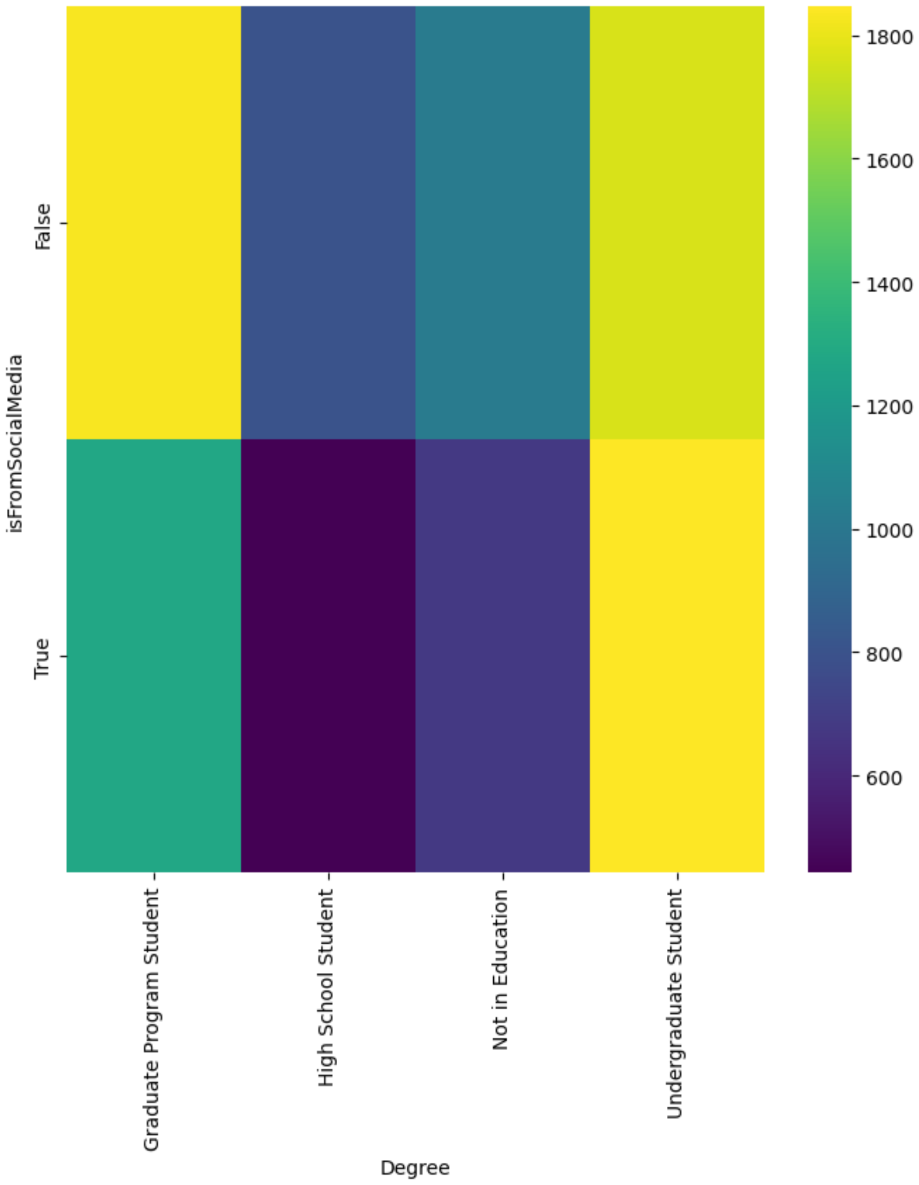

Figure 8.9: degree and level of education correlation

Most undergraduate students have a higher correlation to have come from the social media.

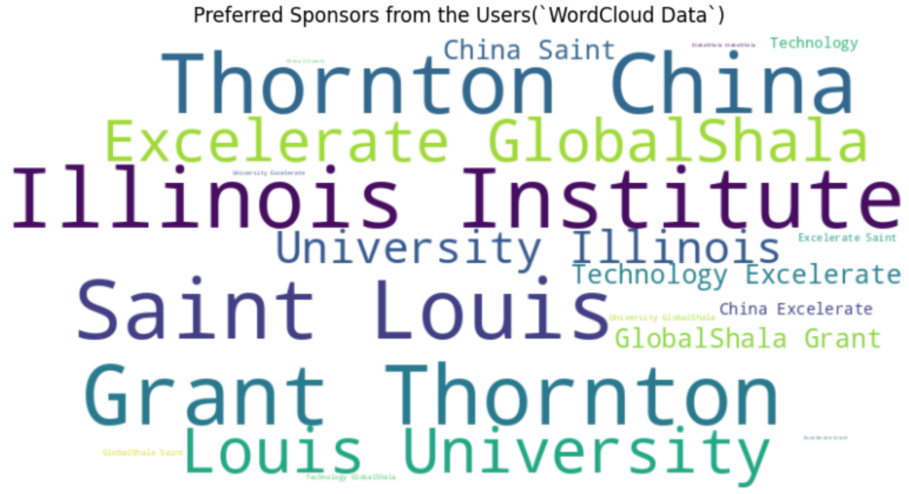

Figure 8.10: preferred sponsors

The above is a list of all preferred sponsors as seen in the user data.

8.2 Opportunity sign up and completion data

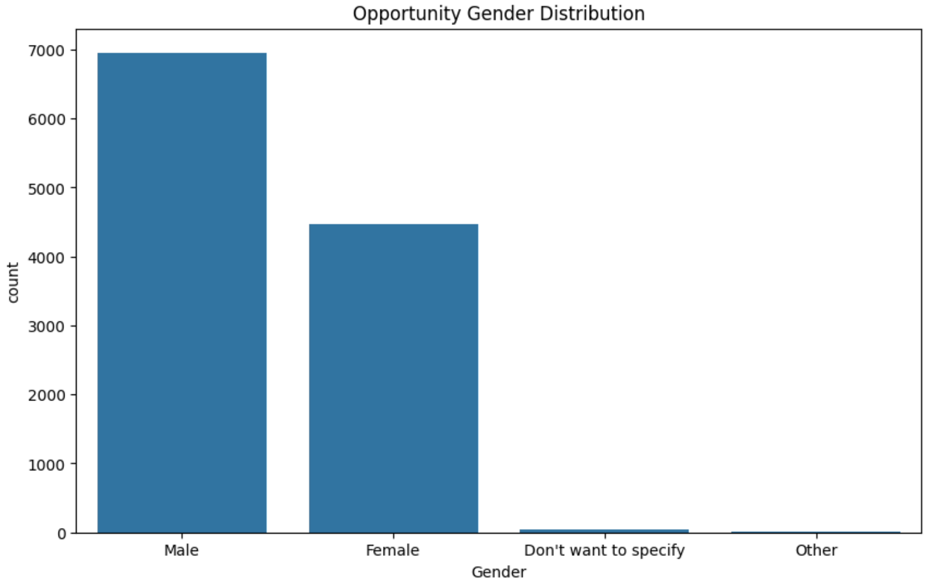

Figure 8.11: Opportunity gender distribution

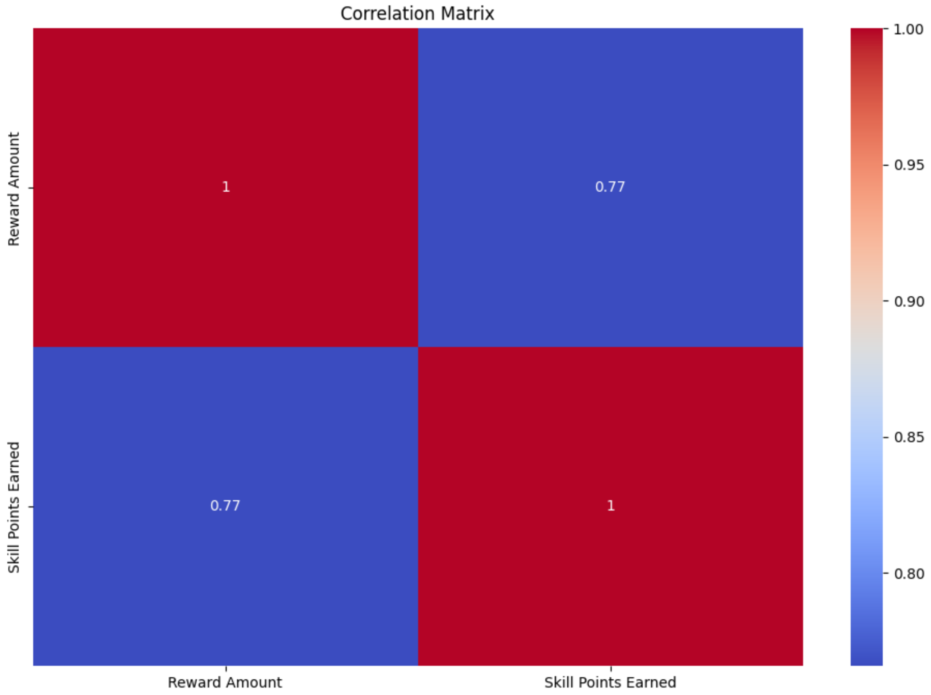

Figure 8.13: Reward amount distribution

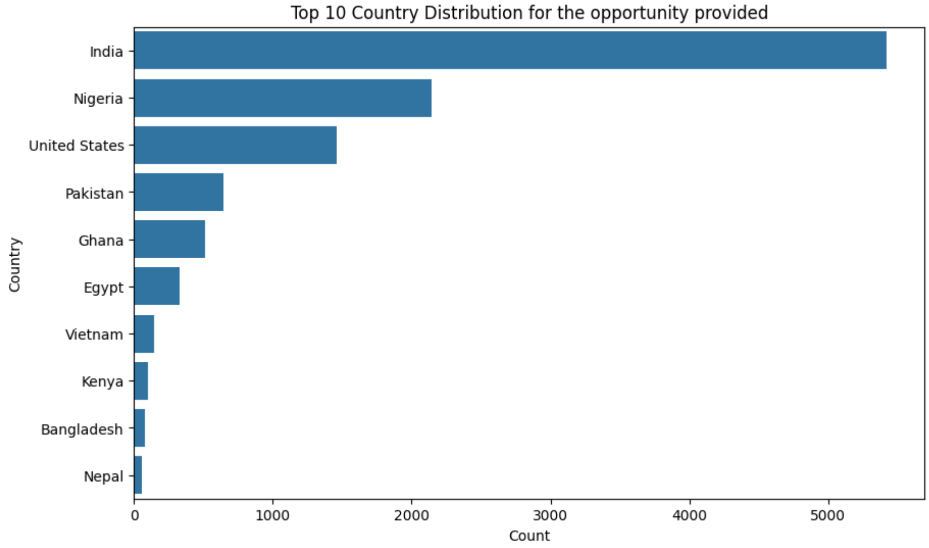

Figure 8.12: Top country by opportunity provided

More than 1000 people who got the opportunity got zero reward amounts.

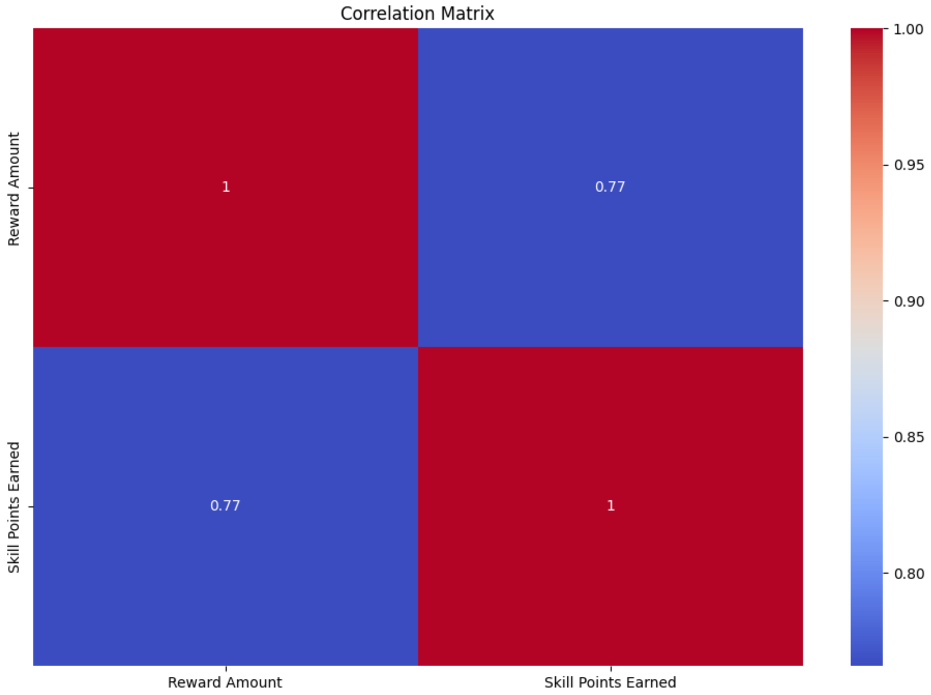

Figure 8.14: correlation matrix

There is a very high positive correlation between reward amount and skills point earned, which means that those who get high rewards also earn more skills points.

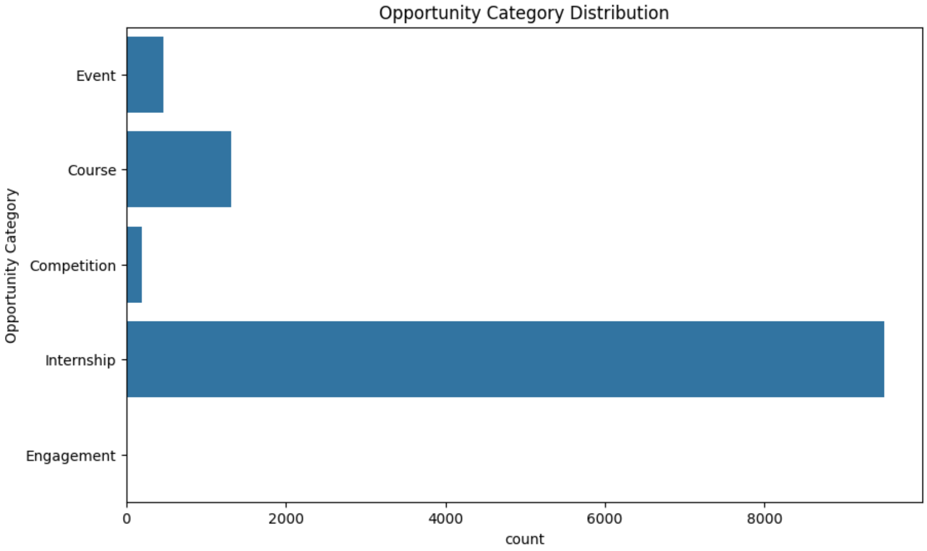

Figure 8.15: Opportunity category

More internship opportunities are sought out by most applicants as an opportunity for them.

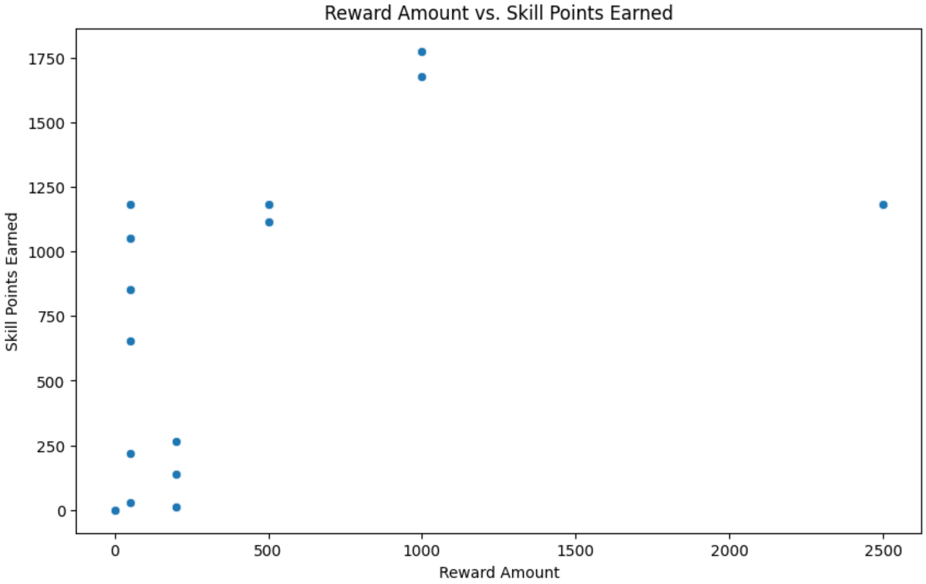

Figure 8.16: scatter plot

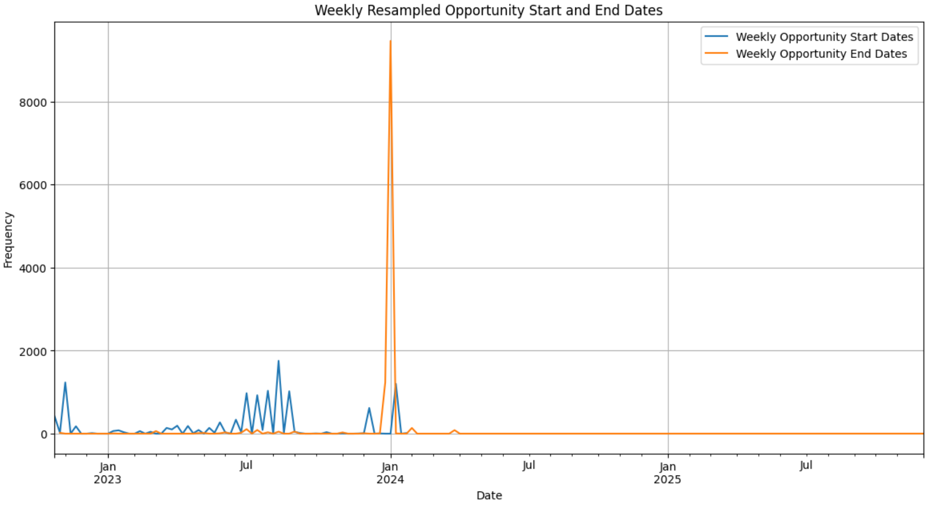

Figure 8.17: Opportunity start and end date

There is a higher variability between start and end date as seen from the plot.

Challenges Face

During the exploratory data analysis process, several challenges were encountered:

1. **Missing Values**: Significant portions of the datasets contain missing values, particularly in key columns such as `Gender` and `Degree`. This required careful handling to avoid biased analysis results.
2. **Data Inconsistencies**: There were inconsistencies in data entries, such as varying formats for the same country names and different date formats. These needed to be standardized for accurate analysis.
3. **High Cardinality**: Some columns, like `City` and `PreferredSponsors`, had a high number of unique values, making it challenging to draw meaningful insights without appropriate grouping or filtering.
4. **Data Integration**: Combining insights from the user data and the opportunity sign up and completion data required careful alignment of columns and ensuring that the merged data maintained its integrity.
5. **Computational Constraints**: Processing large datasets with many entries and columns required substantial computational resources and efficient coding practices to ensure timely analysis.

Next Steps

After we have successfully cleaned the data, the next steps include:

1. **Handling Outliers and Anomalies**: Identifying and addressing any outliers and anomalies in the data to ensure accurate analysis.
2. **Normalizing or Scaling Relevant Features**: Normalizing or scaling the relevant features to improve the performance of machine learning algorithms.
3. **Addressing Data Quality**: Ensuring the quality of data by addressing any remaining issues such as inconsistencies or inaccuracies.
4. **Feature Engineering**: Creating new features from the existing data to better capture the underlying patterns and improve model performance.
5. **Data Transformation**: Transforming the data as necessary for analysis, which may include encoding categorical variables or aggregating data at different levels.