



Exploratory Data Analysis

USER DATA REPORT



Intro

Why Make Use of EDA?

- To Address outliers
- To detect and correct errors
- To understand patterns in data

EDA is a crucial first step in data analysis, as it assists in the process of looking at data before any assumptions are made.

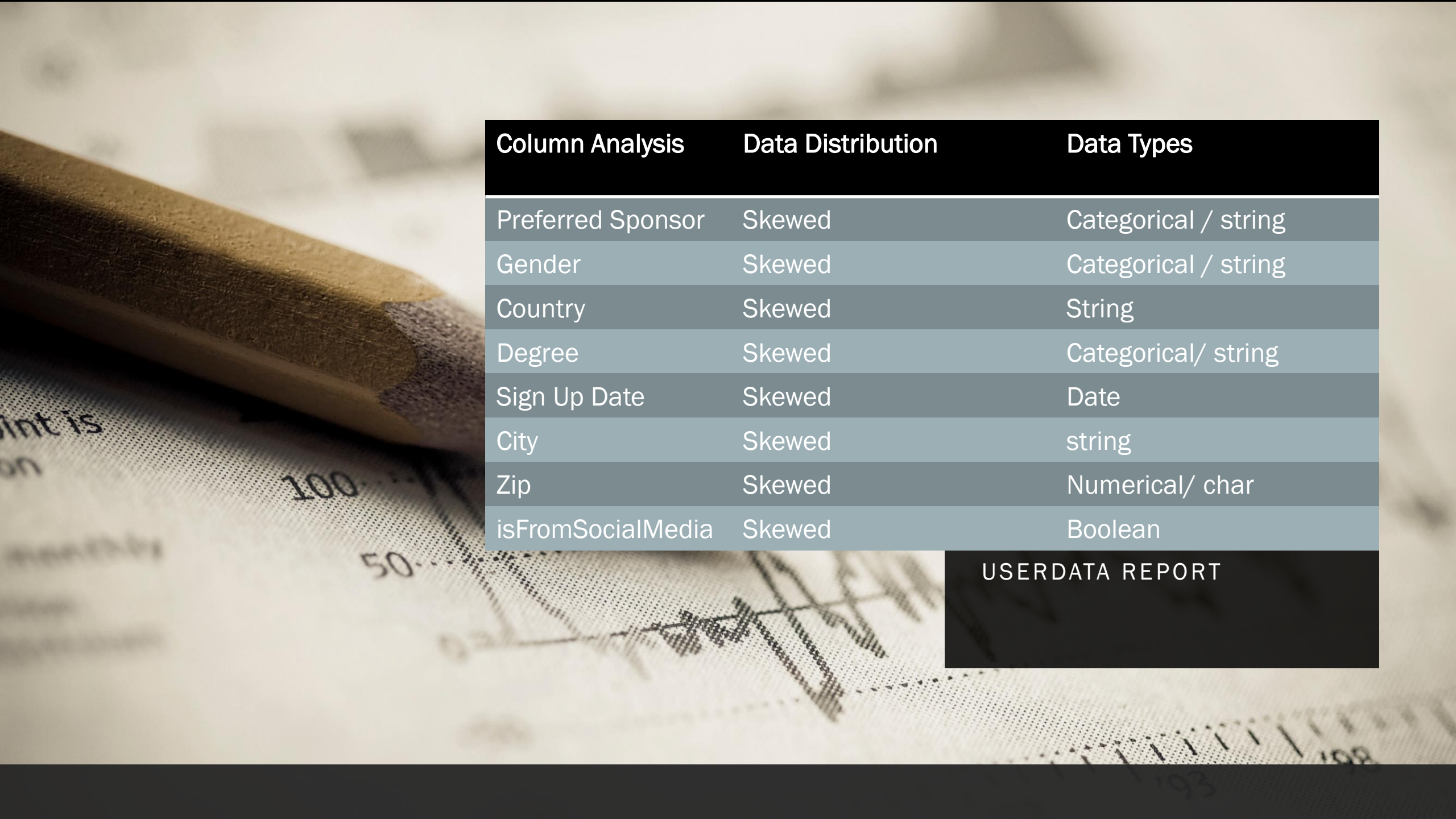


Overview

For the User Data set

The User Data set provides information on individuals who have/ has had an account with Excelerate. Each unique user is presented on each row with columns representing their details such as data on their gender, qualification, country, etc.

- Number of Rows: 8
- Number of Fields : 196
- There are no unique identifiers in the data set

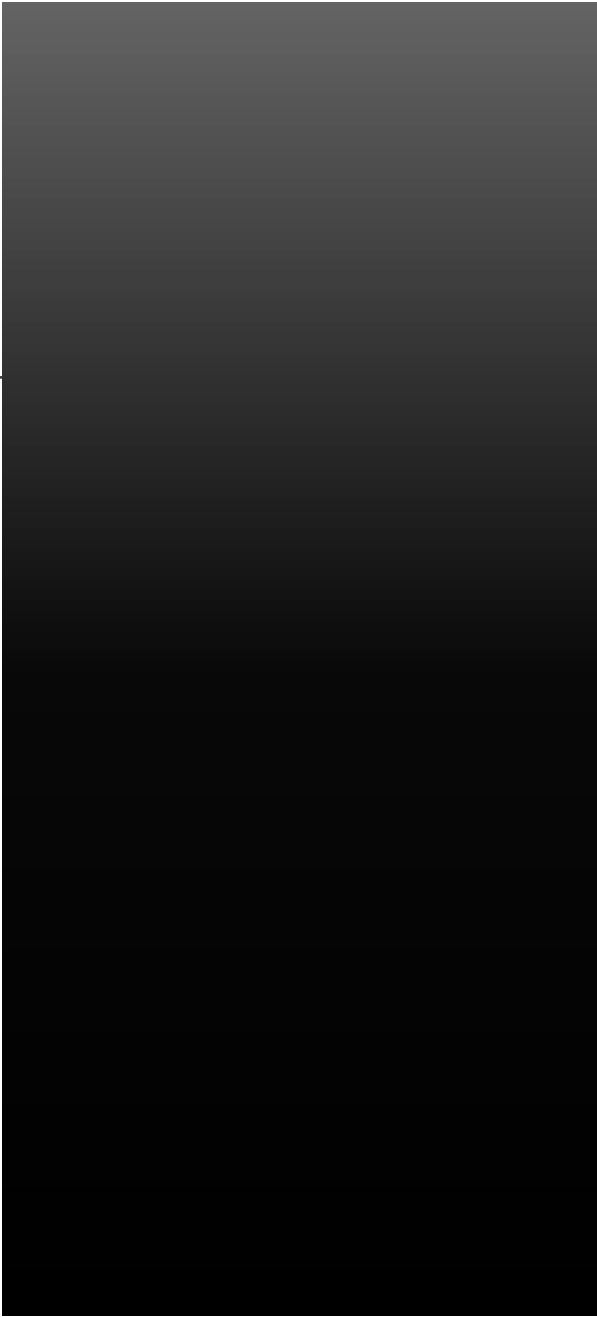


Column Analysis	Data Distribution	Data Types
Preferred Sponsor	Skewed	Categorical / string
Gender	Skewed	Categorical / string
Country	Skewed	String
Degree	Skewed	Categorical/ string
Sign Up Date	Skewed	Date
City	Skewed	string
Zip	Skewed	Numerical/ char
isFromSocialMedia	Skewed	Boolean

USERDATA REPORT

Categorical Classifications		
Preferred Sponsor	-Excelerate -Grant Thornton China -GlobalShala	-Illinois Institute of Technology -Saint Louis University
Gender	-Male -Female	
Degree	-Undergraduate Student -Graduate Program student -Not in Education -"open space"	

Profile ID Analysis for User Data	
Not Applicable: There is no private key that represents the uniqueness of Profile IDs. Values may therefore have duplicates in all columns.	
Opportunity Status Distribution	
Not Applicable: There is no private key that represents the uniqueness of Profile IDs. Values may therefore have duplicates in all columns.	



Statistical Analysis

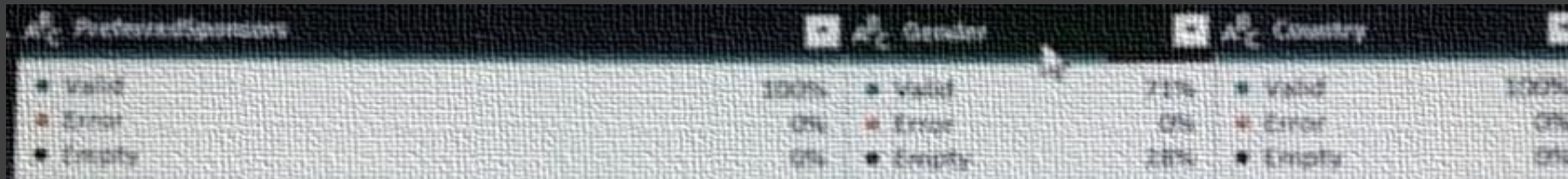
For the User Data set

Since the Distribution of the Data Set is Skewed, we used the Mode to fill in missing values in each suitable column. The mode for each Column's Data Set is:

- Preferred Sponsor – No missing Values recorded,
- Gender – Females(52), Males(74), null (. Not mandatory for sign-up. Therefore can be empty
- Country – Mandatory for sign-up. Mode, most people are from India. Therefore each value in an empty cell was assigned to “India”
- Degree – Not Inserted (62, null(16), Undergrad(31), Graduate(50). Not mandatory
- Sign Up Date – Creation of account with Excelerate
- City – Not Mandatory
- Zip – Not applicable
- isFromSocialMedia -

Initial Observations

For the User Data set



Preferred Sponsor	Gender	Country
100% Valid	100% Valid	71% Valid
0% Error	0% Error	0% Error
0% Empty	0% Empty	28% Empty

In this instance: We Observe the Accuracy rate of each data Field in the Database

- Preferred Sponsor – No missing Values recorded,
- Gender – Females(52), Males(74), null (. Not mandatory to sign-up process. Therefore can be empty
- Country – Mandatory for sign-up. Mode, most people are from India. Therefore each value in an empty cell was assigned to “India”
- Degree – Not Inserted (62, null(16), Undergrad(31), Graduate(50). Not mandatory
- Sign Up Date – Creation of account with Excelerate
- City – Not Mandatory
- Zip – numerical values, but are in character form, since they can’t be calculator like decimals and integers.
- isFromSocialMedia - Boolean values with True or False results

Challenges !!

For the first dataset (UserData), a more primitive and hands-on method is used to address the dataset. This approach is more tedious and time-consuming. The process of cleaning the data was extensive as it involved observing data distribution from each database field/column when addressing missing values, using the replace button to substitute empty values with the mode.

Future process: For the succeeding dataset, Power Query was used, this stirred up the process and saved more time for the team. This enhanced our process of data cleaning, validation, and analysis.





Exploratory Data Report (EDA) on

Opportunity Sign Up and Completion Data

by

DV Associate Team 3A

Introduction:

This Exploratory Data Report aims to provide an overview of the Opportunity Sign Up Completion Data, a dataset containing information on opportunities signed up and completed by individuals. The report will provide a high-level summary of the dataset, identify potential issues, and analyze each column's data type to ensure data quality.

Dataset:

The dataset used for this report is the Opportunity Sign Up Completion Data, which contains information on opportunity sign-ups and completions.

High-Level Summary:

The dataset consists of [23] rows and [20,323] columns, with a unique identifier being the Profile Id. The dataset includes information on various aspects of the opportunities, including demographics, opportunity details, and completion status.

Key Statistics:

- Number of rows: 23
- Number of columns: 20,323
- Unique identifiers: Profile Id, Opportunity Id . Opportunity Name, Badge Id, Badge Name

Column Data Types:

The dataset contains a mix of numeric, categorical, and date/time columns. The following columns have been analyzed for data types:

- Profile Id & Opportunity Id(unique identifier): integer
- Reward Amount: decimal
- Skill Points Earned: integer
- Apply Date, Opportunity Start Date, Opportunity End Date, Graduation Date: date

Opportunity Category:

* Competition * Course * Engagement * Event * Internship

Counties (Total 108):

Afghanistan, Albania, Algeria, American Samoa, Andorra, Angola, Aruba, Australia, Azerbaijan, Bangladesh, Belarus, Belgium, Belize, Benin, Bhutan, Botswana, Brazil, British Indian Ocean Territory, Burkina Faso, Burundi, Cameroon, Canada, China, Colombia, Congo, Congo (The Democratic Republic of the Congo), Costa Rica, Cote d'Ivoire, Dominican Republic, Egypt, Ethiopia, Falkland Islands (Malvinas), Fiji, France, Gambia, Germany, Ghana, Guam, Haiti, Honduras, Hungary, India, Indonesia, Iran (Islamic Republic of Persian Gulf), Iraq, Ireland, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya (Republic of South Korea), Kuwait, Lebanon, Lesotho, Liberia, Libyan Arab Jamahiriya (Libya), Malawi, Malaysia, Mauritius, Mexico Mongolia Morocco Mozambique Myanmar Namibia Nepal Netherlands Nigeria Norway Oman Pakistan Peru Philippines Qatar Russia Rwanda Saint Lucia Saudi Arabia Senegal Sierra Leone Singapore Somalia South Africa Spain Sri Lanka Sudan Swaziland Tanzania (United Republic of Tanzania) Togo Trinidad and Tobago Tunisia Turkey Uganda Ukraine United Arab Emirates United Kingdom United States Uzbekistan Venezuela (Bolivarian Republic of) Vietnam Virgin Islands (U.S.) Yemen Zambia Zimbabwe

Categorical Column Analysis:

The following categorical columns were analyzed for unique values and frequencies:

Gender:

- Male: 12,240 (60.23%)
- Female: 8,004 (39.39%)
- Don't want to say: 63 (0.31%)
- Other: 14 (0.07%)
- Blank: 1

Status Description:

- Applied
- Dropped Out
- Not Started
- Rejected
- Rewards Award
- Started
- Team Allocated
- Withdraw

Current Student Status:

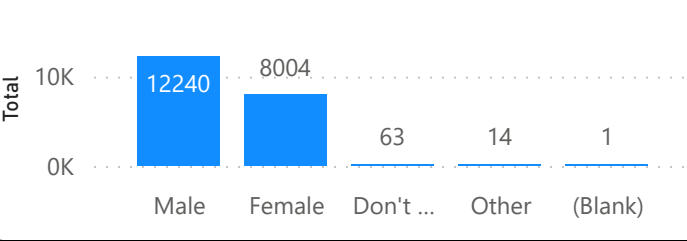
- Graduate
- Program Student
- High School Student
- Not in Education
- Undergraduate Student

Potential Issues:

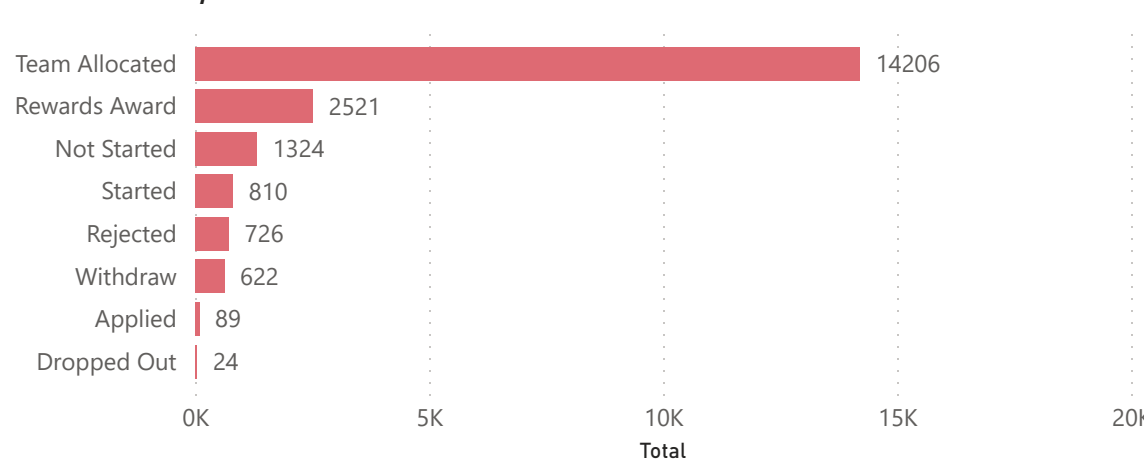
Upon reviewing the dataset, some potential issues were identified:

- Missing values
- Outliers in the Skill Points Earned column (three records with values greater than 100)

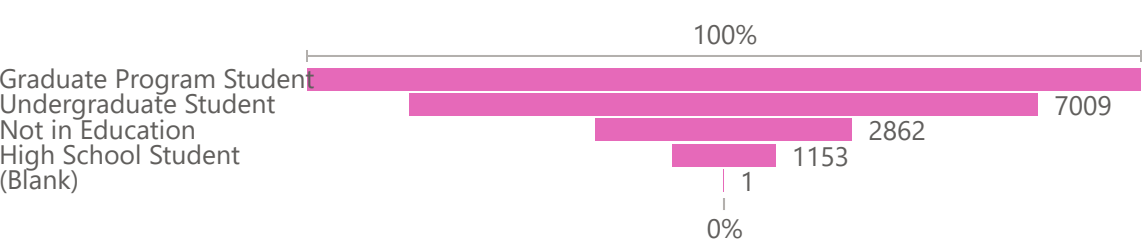
Count of Gender



Status Description



Current Student Status



Exploring the Opportunities with the Highest Number of Graduates timeline

Introduction:

With the increasing importance of higher education, *Opportunity* programs have become a vital stepping stone for individuals seeking to advance their careers. However, not all graduate programs are created equal. Some programs offer more opportunities for graduates than others. In this analysis, we will examine the top *Opportunity* programs that offer the highest number of graduates.

Methodology:

To identify the top *Opportunity* graduate, we used a dataset of over [20,323] *Opportunity* programs from various fields, including data visualization, marketing business, engineering, healthcare, and more. We analyzed the number of student who graduated of each *Opportunity*.

Opportunity Name	2023	2024	2025	2026	2027	2028	Total
Data Visualization	1109	1217	1065	130	50	7	3578 ↑
Project Management	667	768	894	95	52	7	2483 →
Digital Marketing	424	443	491	87	51	8	1504 →
Health Care Management	254	254	323	50	70	8	959 →
Innovation & Entrepreneurship	208	265	399	49	35	3	959 →
Career Essentials: Getting Started with Your Professional Journey	242	216	293	44	44	5	844 →
Linked Up: The LinkedIn Makeover Workshop	92	92	139	18	12	2	355 →
CPR/AED Certification	66	81	107	18	64	3	339 →
Epidemiology Training Internship	75	87	85	16	24	1	288 →
Money Matters: A Personal Finance Workshop	72	44	76	14	18	2	226 →
Leadership Launchpad	70	44	69	11	7	1	202 →
Mental Health First Aid Workshop	24	57	57	13	6	4	161 ↓
AI Ethics Challenge	16	36	48	6	2		108 ↓
Cracking the Interview Code Workshop	31	25	43	3	3		105 ↓
The Brand Booster Challenge	34	24	37	2	4	1	102 ↓
Changemakers Challenge	23	37	29	4	5		98 ↓
Lens Masters: A Photography Contest	16	22	44	4	4		90 ↓
Verses: A Poetry Writing Competition	15	23	44	3	4		89 ↓
Crafting Your Personal Brand Workshop	13	20	41	6	3	1	84 ↓
Startup Mastery Workshop	18	17	31	4	3		73 ↓
Join a Student Organisation	14	6	51				71 ↓
Info Innovators Challenge: An Infographic Design Contest	8	21	36	3	2		70 ↓
Digital Palette: A Global T-shirt Design Competition	19	26	13	4	2		64 ↓
Empowered: A Mindfulness and Emotional Intelligence Workshop	11	15	32	2	3	1	64 ↓
Cook a Tale	10	14	22	2	5	1	54 ↓
Slide Geeks: A Presentation Design Competition	21	17	10	2	1	1	52 ↓
Resume Booster Workshop	8	8	22	1			39 ↓
Million Dollar Idea	8	9	14		1		32 ↓
Mental and Physical Health Session	6	1	13				20 ↓
Jump Start: Developing your Emotional Intelligence	5	1	13				19 ↓
Life Beyond Saint Louis University's Campus		8	2				10 ↓
Statement of Purpose (SOP) Writing Workshop	2	1					3 ↓
Upload Your First Year Transcript	1	1	1				3 ↓
Total	2210 →	2233 →	1992 →	369 →	231 →	29 ↓	7064 ↑

Undergraduate Student
\$979,160
Sum of Reward Amount
Not in Education
\$398,390
Sum of Reward Amount
High School Student
\$156,800
Sum of Reward Amount
Graduate Program Student
\$1,189,010
Sum of Reward Amount
(Blank)
\$2,500
Sum of Reward Amount

This visualization shows the top students with the highest total reward amounts by their current student status. The goal is to identify which student groups are most successful in earning rewards.

Insights:

- The "Current Student" group has earned the most rewards, with a total sum of [reward amounts].
- The "Undergraduate" group has earned significant rewards than the other groups.
- The "Graduate Program Student" has earned the lower amount of rewards.
- The "Student who did not provided their status" has the lowest amount of rewards .

Potential Actions:

- Develop targeted marketing campaigns to reach out to prospective students and encourage them to enroll.
- Offer additional incentives or rewards to current students to maintain their engagement and retention.

Distribution of Students by Country

Objective:

To visualize the distribution of students from different countries and gain insights into the diversity of our student population.

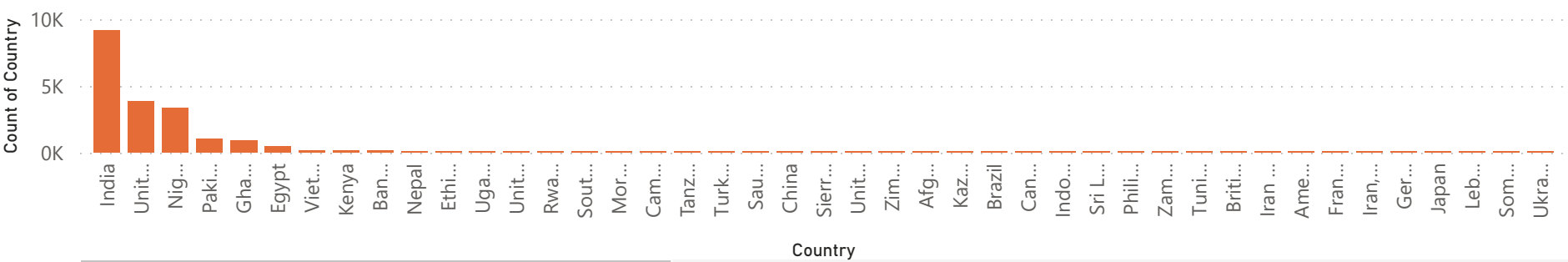
Data:

We have a dataset containing information on students, including their country of origin. The data is collected from (Opportunity Wise Data)

Findings:

- **Top 5 Countries with the most students:** [India, United States, Nigeria, Pakistan, Ghana]
- **Countries with fewest students:** [Oman, Senegal, Burkina Faso, Norway, Dominican Republic, e.t.c]
- **Distribution:** The distribution of students by country appears to be skewed, with a few countries having a large number of students and many countries having only a few or no students at all.

Count of Students from their respective Country



Country India United States Nigeria Pakistan Ghana Egypt Vietnam Kenya Bangladesh Nepal Ethiopia Uganda United Kingdom...

