DATA PREPROCESSING AND TRANSFORMATION
Group 5B|0801-DV Associate internship

INTRODUCTION

Data preprocessing and transformation are crucial steps in the data analysis and machine learning workflow. These processes involve cleaning, organizing, and structuring raw data to make it suitable for analysis or training machine learning models. Proper data preprocessing can significantly impact the performance and reliability of the models.

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

USER DATA
The User data which includes every user that has ever created an account on the Excelerate platform involves datasets with 8 column/column headings; Preferred sponsors of the user, gender of the user, country of residence during signup, degree or academic level of user, signup date and time of the user, city of residence, zip or postal code and the last column showing whether the user got the information through social media or not. The data has a total of 27,563 rows which is imperative to the same number of users with an account on Excelerate.

OPPORTUNITY WISE DATA
The Opportunity Wise Data comprises 17 different columns and 20323 rows. The rows were later reduced to 11482 after duplicates were removed. The profile ID column was used to remove the duplication because it is the most unique field among other fields.

1. Handling Outliers and Anomalies

Opportunity Sign-Up and Completion Data:

1. select the rewards amount column
2. Copy and paste selected data into a new sheet and sort from A-Z from smallest to highest with 50 being the lowest and 2500 and null values(0) following suit.
3. use the Interquartile function to select cell A1:A:1449
 to discover an outlier is any data value that is
1.Less than First Quartile(Q1) -1.5*IQR
2. Greater than third quartile   +1.5*IQR
Interquartile Range (IQR) = Third Quartile(Q3) - First Quartile(Q1)
 And also manually check or glance through the figures of rewards points which were as follows:$2500,$1000,$500,$250,$150,$120,$100,$80,$50,$0 with the majority being null values.

REWARD AMOUNT VISUAL
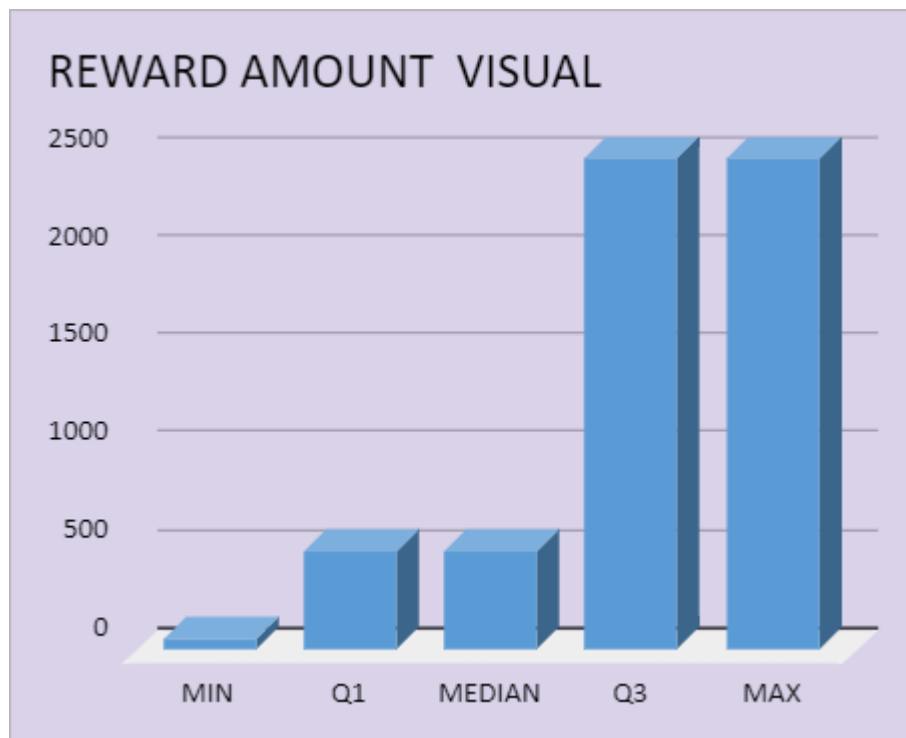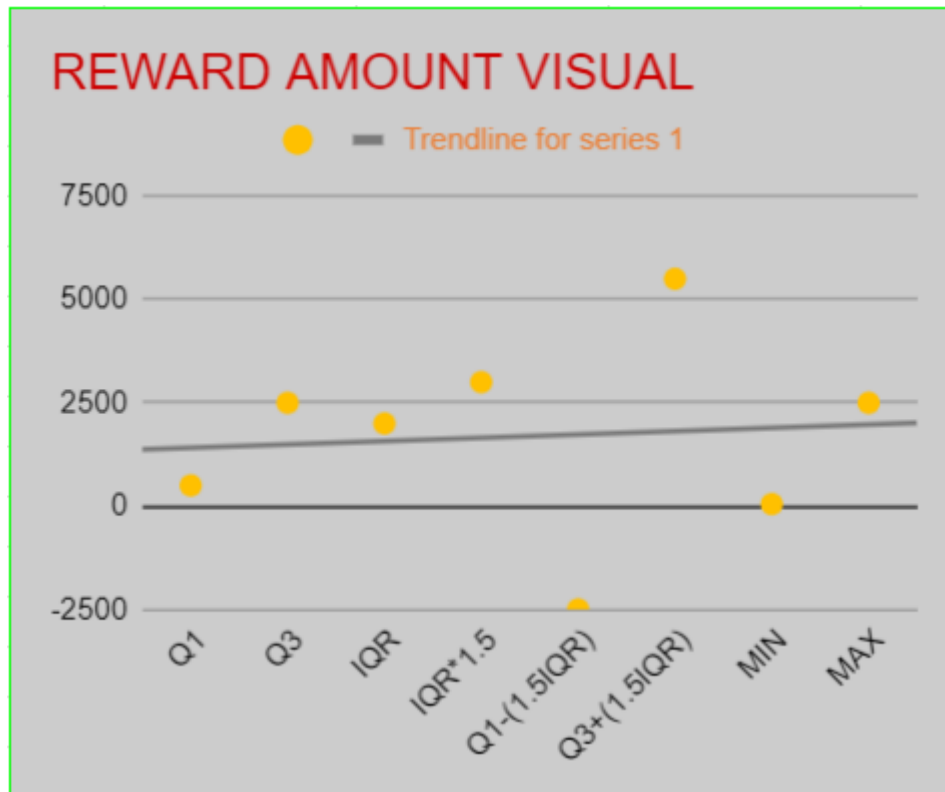
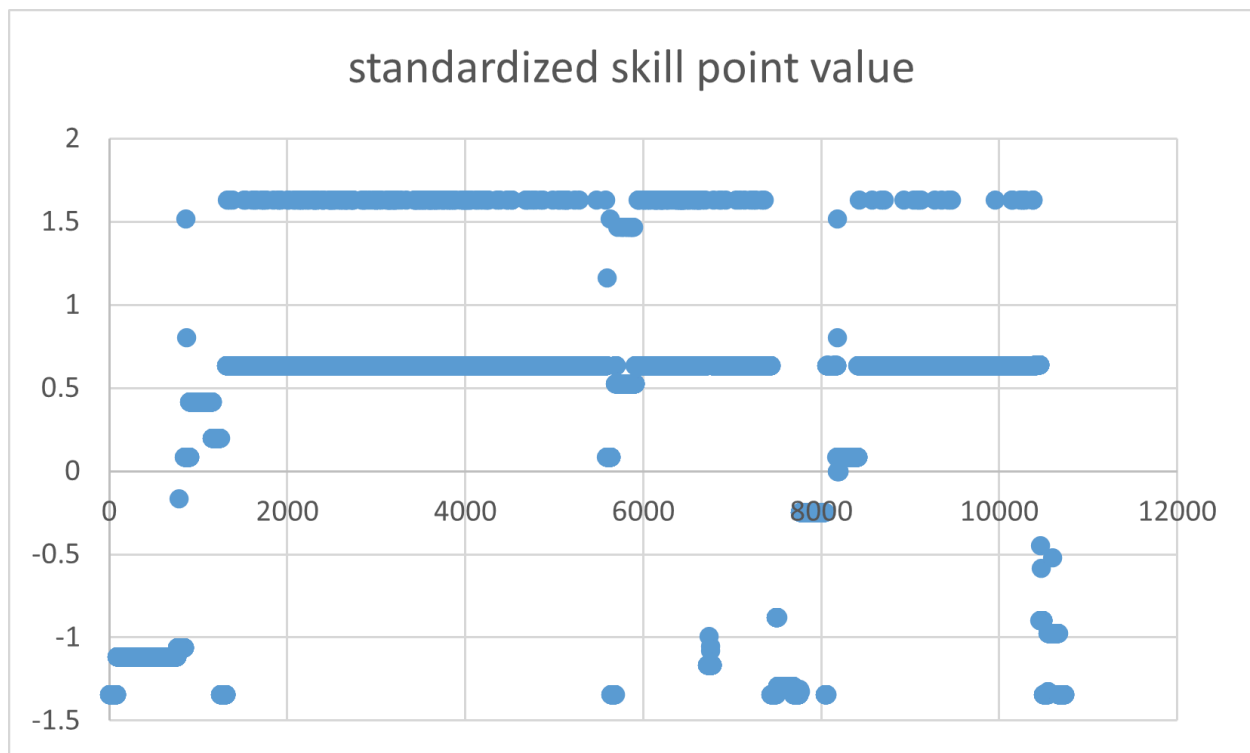**2. Normalize or Scale Relevant Features:**

USER DATA

The null cells in the gender column were filled with the modal gender i.e. the gender with the highest frequency which is the "male" because of the data type; (gender) which is nominal data. The columns were filled up to ensure consistency and reduce discrepancies.

Entire rows containing null cells in the "Current student status" were removed completely as the larger percentage of the rows with null entries in the column has null entries for the majority of the columns. Keeping these cells blank or null might affect data analysis or give skewed results if filled using statistical methods because null entries showed a very high frequency.

The "City" column alphabetical case was adjusted as some were in uppercase while some were in lowercase. All null entries have been removed following the deletion of rows containing null current student status.

OPPORTUNITY WISE DATA

Duplicate entries were cleaned. In order to normalize and standardize the amount and skills of the rewards point numerical column, the column for status description was cleared of entries that were "rejected" or "withdrawn". The missing or null entries for reward amount and skill points were filled accordingly with the modal value of filled entries for each opportunity category type. Also, using the badge name column, the modal value also corresponds with the lowest possible points and reward amount for participation in each opportunity category type using the reward amount and skill point of cells with "completed" in the status description for each category type as reference. Categorical types of opportunities with no "completed" status in all entries were filled with 0 for the skill points and reward amount.
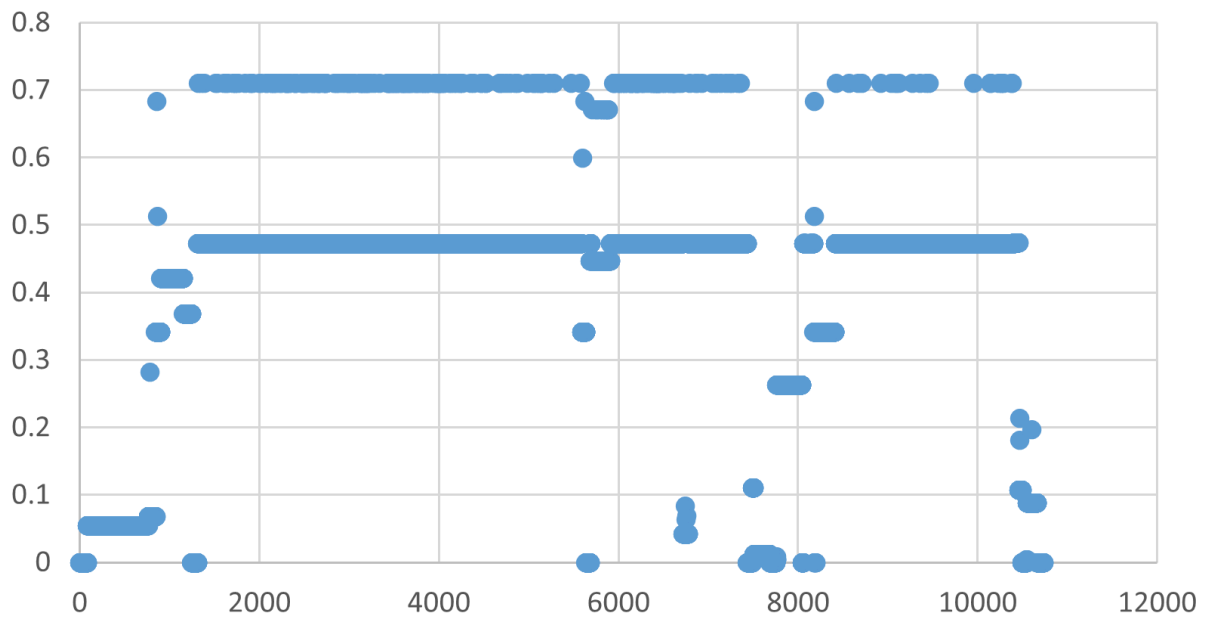
standardized reward amount value
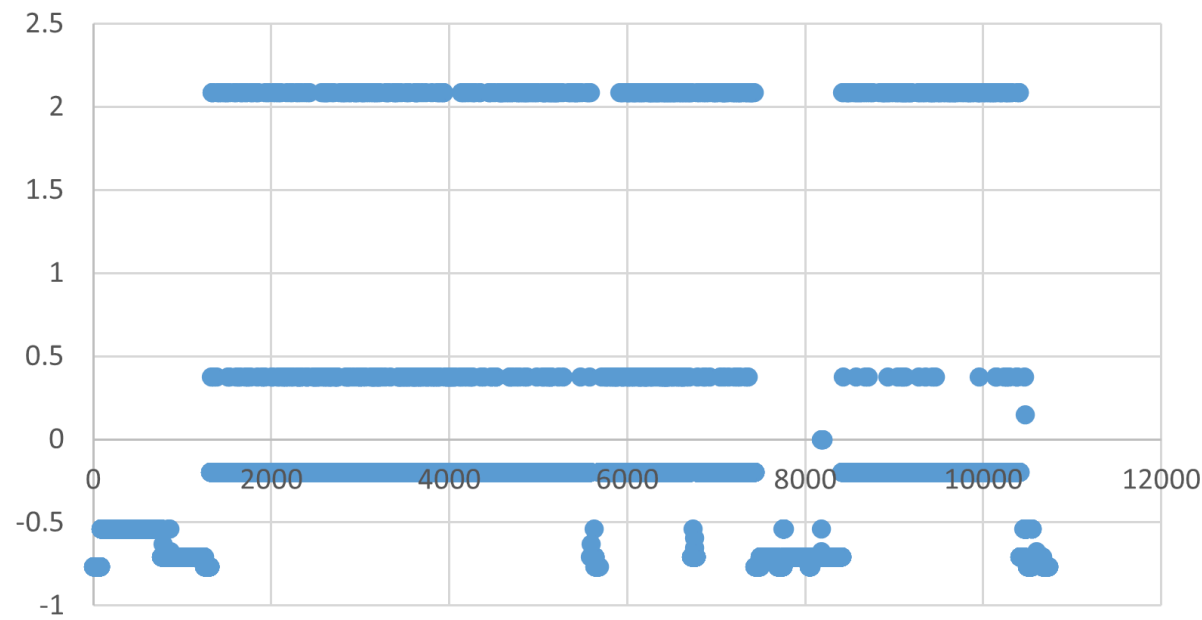
Normalized Reward Amount Value
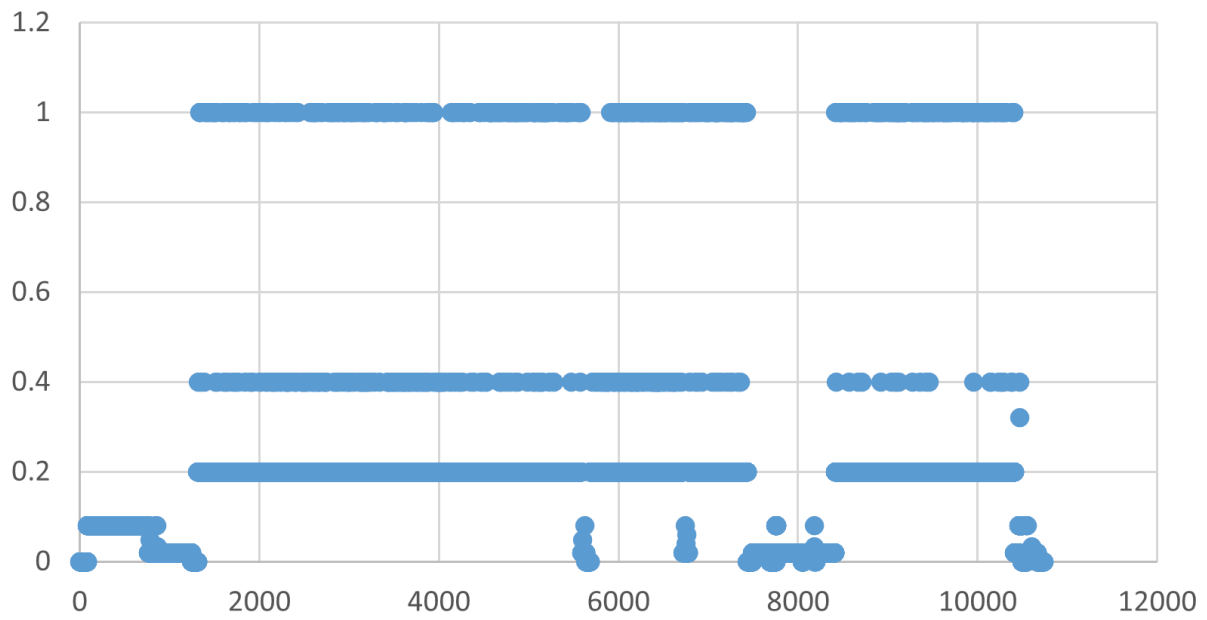
normalized skill points value

3. Addressing Data Quality Issues:

USER DATA

HANDLING NULLS

The null cells in the gender column were filled with the modal gender i.e the gender with the highest frequency which is the "male" because of the data type; (gender) which is a nominal data. The columns were filled up to ensure consistency and reduce discrepancies.

Entire rows containing null cells in the "Current student status" were removed completely as the larger percentage of the rows with null entries in the column has null entries for the majority of the columns. Keeping these cells blank or null might affect data analysis or give skewed results if filled using statistical methods because null entries showed a very high frequency.

The "City" column alphabetical case was adjusted as some were in uppercase while some were in lowercase. All null entries have been removed following the deletion of rows containing null current student status.

OPPORTUNITY WISE DATA

HANDLING NULL VALUES

Duplicate entries were cleaned. The column for status description was cleared of entries that were "rejected" or "withdrawn". The missing or null entries for reward amount and skill points were filled accordingly with the modal value of filled entries(completed opportunities) for each opportunity category type. The modal value also corresponds with the lowest possible points and reward amount for participation in each opportunity category type using the reward amount and skill point of cells with "completed" in the status description for each category type as reference. Categorical-type of opportunities with no "completed" status and no rewards awarded in all available entries were filled with 0 for the skill points and reward amount consecutively.

4. Feature Engineering:

Data feature engineering is the process of leveraging on data to create or manipulate new variables that are not in the training set

* Highlighted and copied out the columns opportunity name, skills point and reward amount.
*Then use the sort and replace function to filter out the opportunity names having null skill points and null reward...thus, set the null value to zero.
* Then sort and filter unique values this cleaned the rows from 10737 to 4749.....then start analysis
*Then use the count If formula to count each occurrence of each unique opportunity,
* Use the Average If formula to determine the average skills for each opportunity earned and the average reward for the opportunity as well.
*And divide reward by skill points to get the ratio of reward to skill points
 *Then use the IF function to grade opportunity categorization based on skill points and use the function for opportunity categorization based on reward amount
 *Use the SUM IF function to calculate the total points for each opportunity and the total reward for each opportunity
 *Then do the percentage analysis of opportunity by skills point grade using the COUNTA function

Number of opportunity with "Low" skill points	47
Number of opportunity with "Medium" skill points	9
Number of opportunity with "High" skill points	26

%age of opportunity with low skill points	57.32
%age of opportunity with Medium skill points	10.98
%age of opportunity with High skill points	31.71

percentage of opportunity by skill points grade
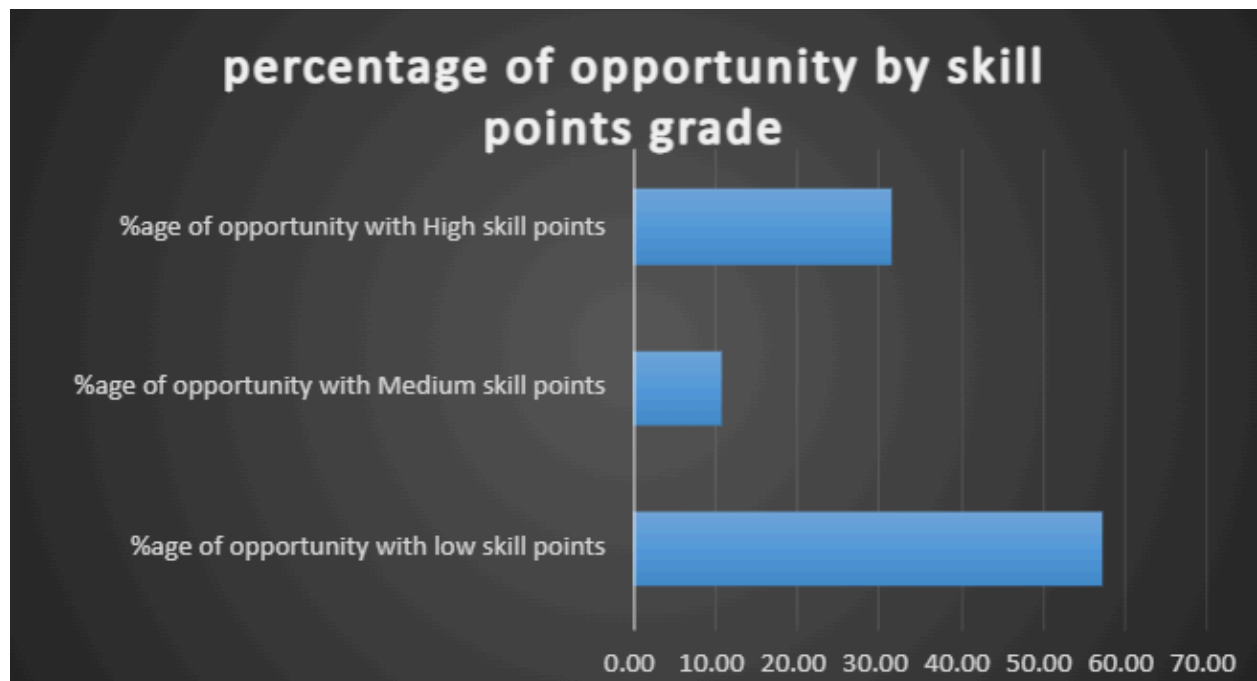
5. One Hot Encoding:

One-hot encoding is a technique used in machine learning and natural language processing to represent categorical variables as binary vectors. In this encoding scheme, each category or class is represented by a unique binary value, and all values are mutually exclusive.

Here's how one-hot encoding works:

Identification of Categories*:* Identify all the unique categories or classes in the categorical variable.

Assigning Binary Values: Assign a unique binary value to each category. Typically, these binary values are 0 or 1.

Creating Binary Vectors: Represent each data point by a binary vector of length equal to the number of unique categories. Each position in the vector corresponds to a category, and only one position will have the value 1, indicating the presence of that category.

Process: Datasheet 1

- Performed one-hot encoding for the **"Gender"** column and created separate columns for **"Male"** and **"Female,"** and created a new column indicating whether "True" or "False" is present in another column using a similar approach.

- To create one-hot encoding columns for "Male" and "Female" based on the given data structure in Sheets, the data starts from cell A1, we used the following formulas:

- The "Male" and "Female" columns start in column F, we used the following formulas:

- For the "Male" column ("Male" is the Nth column, starting in cell F2):

=IF(F2="Male", 1, 0)

For the "Female" column ("Female" is the Oth column, starting in cell F2):

=IF(F2="Female", 1, 0)

- To create one-hot encoding for the **"IsFromSocialMedia"** and created columns for **"TRUE"** and **"FALSE"** based on the given data structure in Sheets, the data starts from cell A1, we used the following formulas:

The "TRUE" and "FALSE" columns start in column M, we can use the following formulas:

For the "TRUE" column ("TRUE" is the Pth column, starting in cell M2):

=IF(M2="TRUE", 1, 0)

For the "Female" column (assuming "Female" is the Qth column, starting in cell M2):

=IF(M2="FALSE", 1, 0)

FINAL OUTPUT :

Gender	Male	Female	IsFromSocialMedia	FALSE	TRUE
Male	1	0	FALSE	1	0
Male	1	0	FALSE	1	0
Female	0	1	FALSE	1	0
Male	1	0	TRUE	0	1
Male	1	0	TRUE	0	1
Female	0	1	TRUE	0	1
Female	0	1	TRUE	0	1
Female	0	1	TRUE	0	1
Male	1	0	FALSE	1	0
Female	0	1	FALSE	1	0
Male	1	0	FALSE	1	0
Female	0	1	TRUE	0	1
Female	0	1	FALSE	1	0
Female	0	1	TRUE	0	1
Male	1	0	TRUE	0	1
Male	1	0	FALSE	1	0
Female	0	1	TRUE	0	1
Female	0	1	TRUE	0	1
Male	1	0	FALSE	1	0
Female	0	1	FALSE	1	0
Male	1	0	TRUE	0	1

Process : Datasheet 2

-- Performed one-hot encoding for the **"Gender"** column and created separate columns for **"Male"** and **"Female"**.

-- To create one-hot encoding columns for "Male" and "Female"

For the "Male" column ("Male" is the Kth column, starting in cell B2):

=IF(B2="Male", 1, 0)

For the "Female" column ("Female" is the Lth column, starting in cell B2):

=IF(B2="Female", 1, 0)

-- Performed one-hot encoding for the **"Opportunity Category"** column and created separate columns for 5 different categories: **Internship, Event, Course, Competition, Engagement.**

-- For the "Internship" column ("Internship" is the Mth column, starting in cell H2):

=IF(H2="Internship", 1, 0)

-- For the "Event" column ("Event" is the Nth column, starting in cell H2):

=IF(H2="Event", 1, 0)

-- For the "Course" column ("Course" is the Oth column, starting in cell H2):

=IF(H2="Course", 1, 0)

-- For the "Competition" column ("Competition" is the Pth column, starting in cell H2):

=IF(H2="Competition", 1, 0)

-- For the "Engagement" column ("Engagement" is the Qth column, starting in cell H2):

=IF(H2="Engagement", 1, 0)

FINAL OUTPUT :

GENDER	Male	Female	Opputunity Category	Internship	Event	Course	Competition	Engagement
Male	1	0	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Female	0	1	Event	0	1	0	0	0
Female	0	1	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Male	1	0	Course	0	0	1	0	0
Female	0	1	Course	0	0	1	0	0
Female	0	1	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Female	0	1	Internship	1	0	0	0	0
Male	1	0	Internship	1	0	0	0	0

6. Documentation:

1.	Handling Outliers and Anomalies - Identify outliers and anomalies	- Outliers can distort analysis and negatively affect model performance. Identifying and handling them ensures a more accurate representation of the data.
2.	Normalize or Scale Relevant Features - Normalize or scale numerical features	- Features with different scales can impact model training differently. Normalising or scaling ensures consistency and improves the model convergence.
3.	Addressing Data Quality Issues - Identify and address missing values	- Missing values can lead to biased or inaccurate results. Addressing data quality issues ensures the reliability of the analysis.
4.	Feature Engineering - Create new features	- Feature engineering introduces new information to enhance the dataset, potentially improving model performance.
5.	Data Transformation - Perform additional transformations - One-hot Coding	- Data transformations, such as log transformations or encoding, help meet model assumptions, handle skewed distributions, and capture non-linear relationships. Transformations contribute to improved model fit and capturing complex patterns. -Many machine learning algorithms require | numerical input. One-hot encoding is a technique to convert categorical variables into a format suitable for numerical analysis. - Transform categorical variables into a binary format, creating binary columns for each category. This

		enables the algorithm to consider each category independently.

-Handling Outliers and Anomalies:

Outliers in sales data might distort the overall trend.
Identifying anomalies helps ensure the accuracy of customer demographics analysis (Q2).
Modified Rationale:
Handling outliers and anomalies is crucial to maintaining the integrity of the data, ensuring that the overall sales trend and customer demographics insights are not skewed by extreme values.

-Normalize or Scale Relevant Features:

Scaling features is essential for accurate analysis of different product categories' contributions.
Modified Rationale:
Scaling relevant features ensures that the contributions of different product categories are comparable, aiding in the analysis of their impact on overall sales.

-Addressing Data Quality Issues:

Addressing data quality issues ensures the reliability of insights related to factors influencing sales.
Modified Rationale:
Ensuring data quality is critical to reliably identify and analyze factors like seasonality or promotional events that might influence sales.

-Feature Engineering:

Creating new features can enhance the dataset for better predictions of sales trends.
Modified Rationale:
Feature engineering contributes to a more nuanced understanding of the factors contributing to the overall sales performance trend.

-Data Transformation (One-Hot Encoding):

One-hot encoding enables the analysis of customer demographics, aligning with.
Modified Rationale:
The one-hot encoding of categorical variables, such as customer demographics, facilitates a detailed analysis of patterns and insights that may affect sales.

CONDUCT INITIAL ANALYSIS

Trends and Comparative Analysis
Trend Analysis

Trend analysis involves examining data over time to identify patterns, tendencies, or movements that may indicate a consistent direction or tendency in a particular variable. In the context of a sales performance dashboard, trend analysis is crucial for understanding how sales have been evolving.

Trend analysis is a quantitative review of what happens over time. It entails the collection of data from multiple periods and plotting the information on a line graph for further analysis. There are three types of trend analysis: geographic, temporal, and intuitive.
1) Trend Analysis > Using User Data

 a) User Growth Overtime: Utilized timestamp information to visualize user sign-ups and identify periods of Growth.

  Outcome:
    • The period of Growth started on January 2023 where the highest Growth in signup
Is in July 2023
    • The period of decline in signup is from July 2023 - November 2023

b) Demographic Shift: Using the categorical variables (gender, IsFromSocailMedia) to observe the changes over Time.

Outcome:
    • Users who indicated their Gender as "other" have the highest signup, followed by "Male"
    • Majority of users sign up using Social media platforms.

Sum of Count

Total

Series "Total" Point "2023-06"
Value: 1546

— Total

Data ▼

Sum of Count

isFromSocialMedia ▼

— TRUE
— FALSE
— Null

Data ▼

The chart shows "Sum of Count" over time (2022-05 to 2023-11) by Gender (Other, Male, Female).

Grand Total column values:
1, 116, 212, 185, 107, 66, 97, 121, 376, 268, 475, 1546, 918, 580, 196, 137, 56
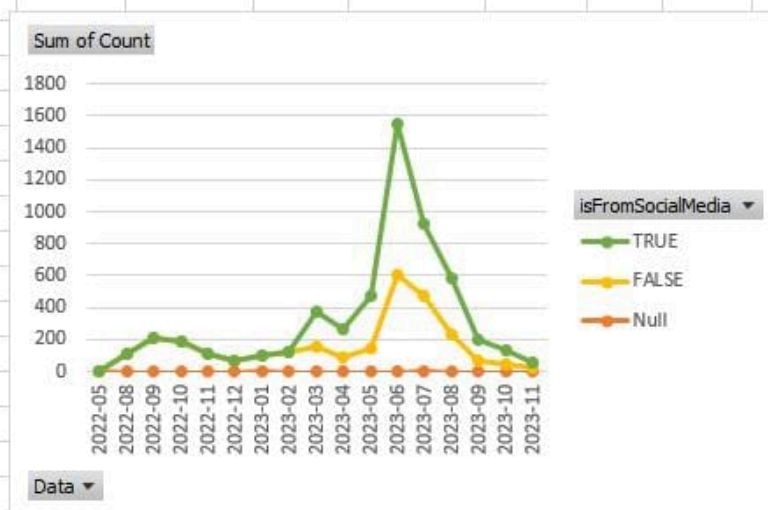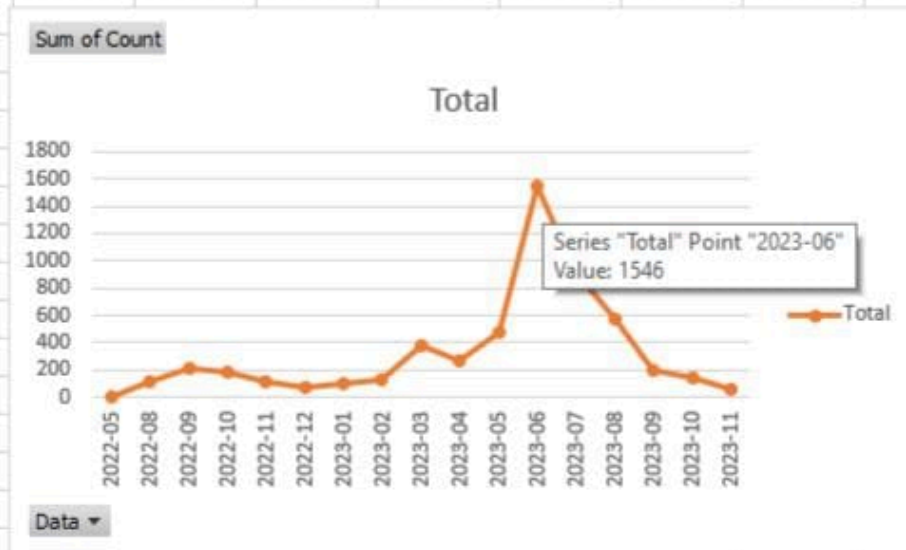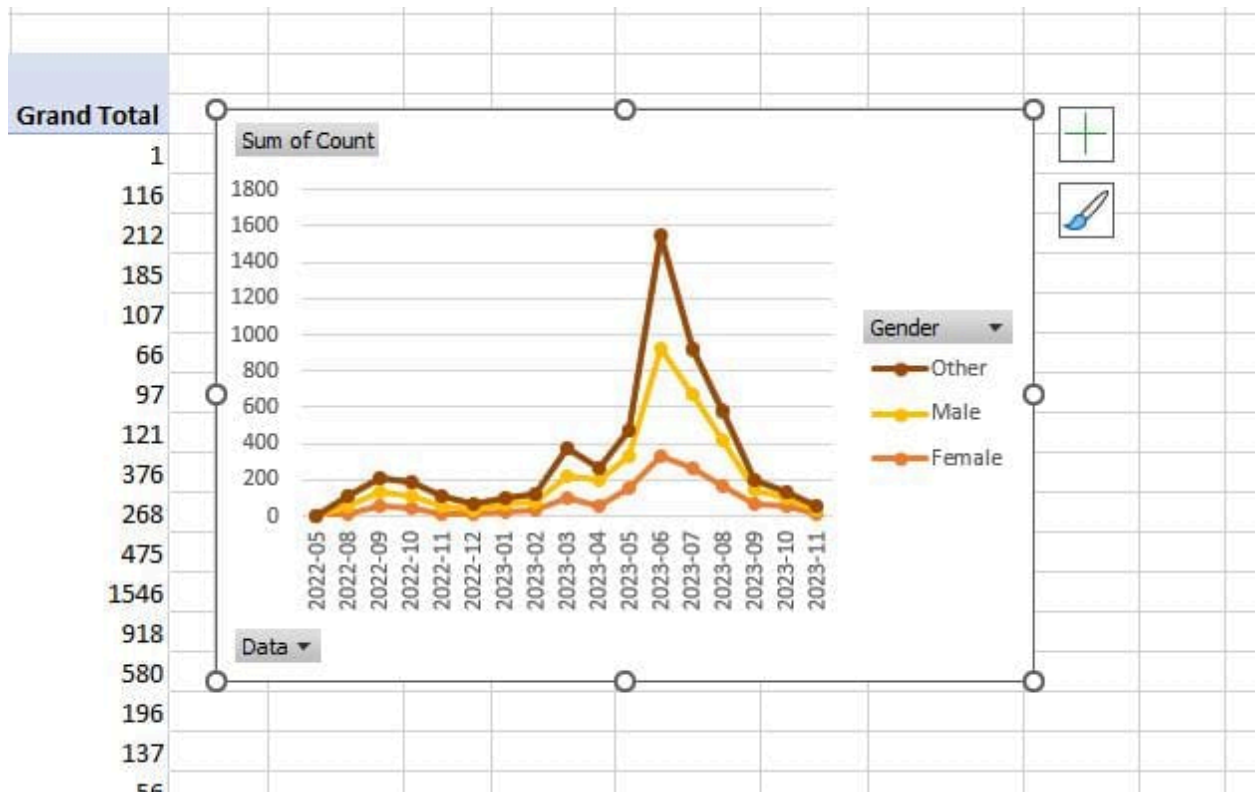
2) Trend Analysis > Using Opportunity Dataset
a) Opportunity participation Trend: Analyze the trend (growth or decline) in the number of learners signing up for Opportunity over time

Outcome:
   • The highest participation is on August 7 2023 over time.

b) Completion Rate: Calculate and analyze the trend in the Completion rate of various opportunities(to identify which opportunities consistently show a high Completion rate over time)

Outcome:
    • Created a new column from the status description, as "Completion Status) using the the IF Formula (where "Reward   Awards" = Completed else Not Completed). After which we calculated the Completion rate = Count of completed users/ Total number of User (1285 ÷ 11481 = 11.19%)
   • Internship has the highest Completion rate over time. On July 13th 2022 the internship opportunity has 100% Completion.

Count of Profile Id

Opportunity participation over time

Opportunity Start Date ▾

Total
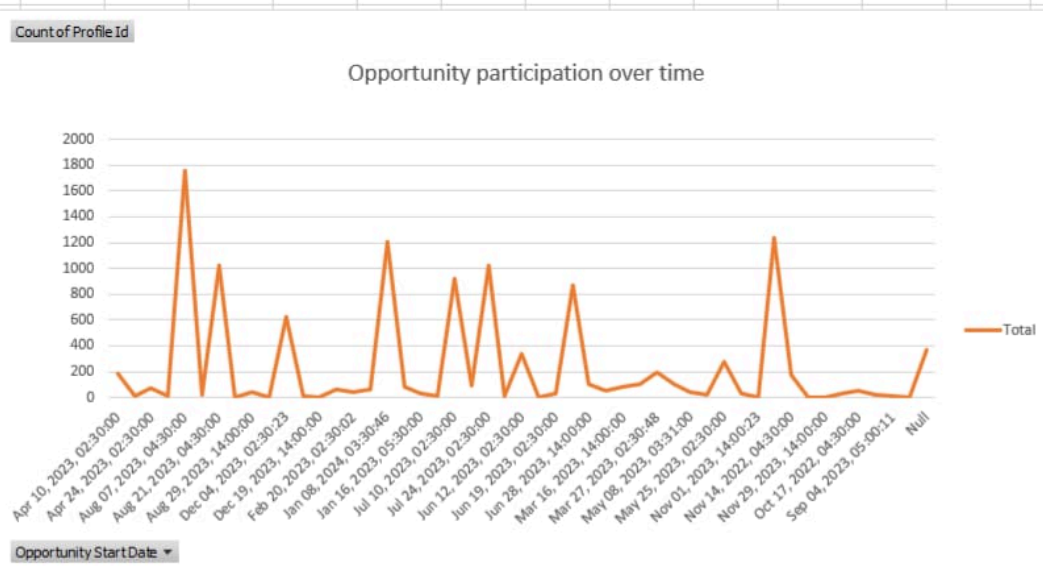
=IF(OR([@[Status Description]]= "Rewards Award"),"Completed", "Not Started")

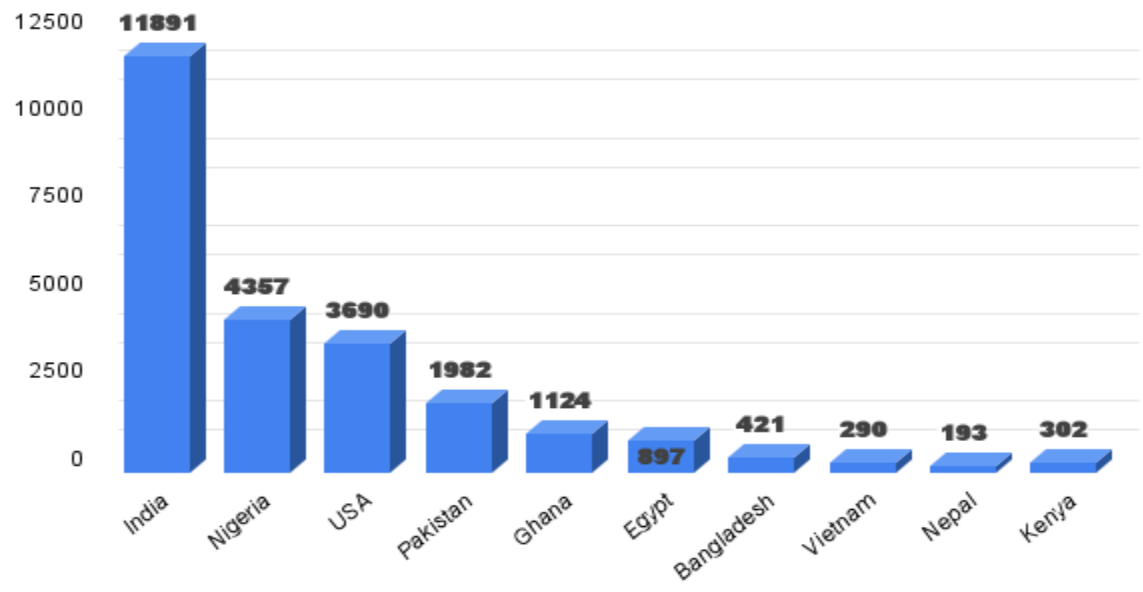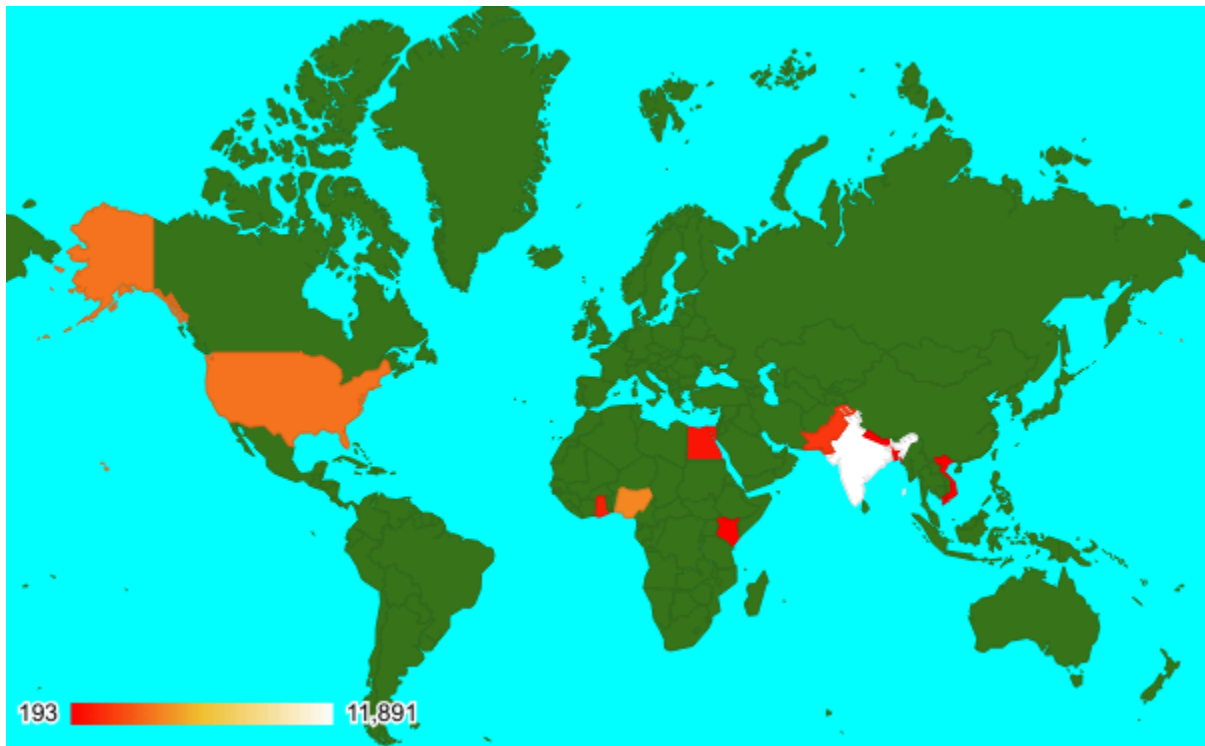|	18	19	20	21	
	tended Major ▾	Status Description ▾	Completion Status ▾	Apply Date ▾	Oppo
	ce	Not Started	Not Started	Oct 03, 2022, 17:29:17	Jan 05
	s	Rewards Award	Completed	Jan 05, 2023, 18:58:37	Jan 05
	Science and Engi	Not Started	Not Started	Oct 04, 2022, 08:10:07	Jan 05
	Science	Not Started	Not Started	Oct 05, 2022, 13:41:47	Jan 05
		Not Started	Not Started	Oct 05, 2022, 11:07:17	Jan 05
	Science and Infor	Rewards Award	Completed	Oct 27, 2022, 15:05:04	Oct 28
	Engineering	Not Started	Not Started	Oct 26, 2022, 05:20:31	Oct 28
	guage and Litera	Team Allocated	Not Started	Oct 16, 2022, 03:29:18	Oct 28
	nagement	Rewards Award	Completed	Oct 27, 2022, 06:34:05	Oct 28
	ngineering	Rewards Award	Completed	Oct 26, 2022, 17:51:37	Oct 28
	Science and Engi	Team Allocated	Not Started	Oct 10, 2022, 09:33:30	Oct 28
	rity	Rewards Award	Completed	Oct 27, 2022, 15:06:43	Oct 28
	Information Syst	Not Started	Not Started	Oct 27, 2022, 15:03:15	Oct 28
	Science and Infor	Rewards Award	Completed	Oct 27, 2022, 15:05:40	Oct 28
	n technology	Rewards Award	Completed	Oct 26, 2022, 05:28:19	Oct 28
	Science and Infor	Not Started	Not Started	Oct 11, 2022, 06:50:30	Oct 28
	Science and Infor	Rewards Award	Completed	Oct 27, 2022, 06:12:48	Oct 28
	puter Science	Not Started	Not Started	Oct 11, 2022, 17:00:10	Oct 28
	Science	Rewards Award	Completed	Oct 26, 2022, 05:12:43	Oct 28
	Information Syst	Rewards Award	Completed	Oct 27, 2022, 15:06:56	Oct 28
	Science	Not Started	Not Started	Oct 13, 2022, 20:50:16	Oct 28
	Science	Not Started	Not Started	Oct 17, 2022, 14:29:59	Oct 28
	Science and Infor	Rewards Award	Completed	Oct 26, 2022, 09:22:38	Oct 28

Total Number of users	Count of Completed User	CompletionRate
11481	1285	11.19%

Row Labels	Sum of Completion Rate
Apr 10, 2023, 02:30:00	11.89%

Sum of Completion Rate

Completion Rate of Opportunity Category Over Time

Series "Total" Point "Jun 13, 2022, 04:41:02 Internship"
Value: 100.00%

Total

Opportunity Start Date   Opportunity Category

Activate Windows

Demographic comparative analysis

was carried out using statistical chart representation for visualization to understand demographic differences among user groups. The distribution of student statuses according to regions and their completion rates. The demographic distribution of users spans 170 countries from all continents. Chart 1 shows the top 10 countries with the highest frequency of users, from India to Nepal and chart 2 shows the map distribution
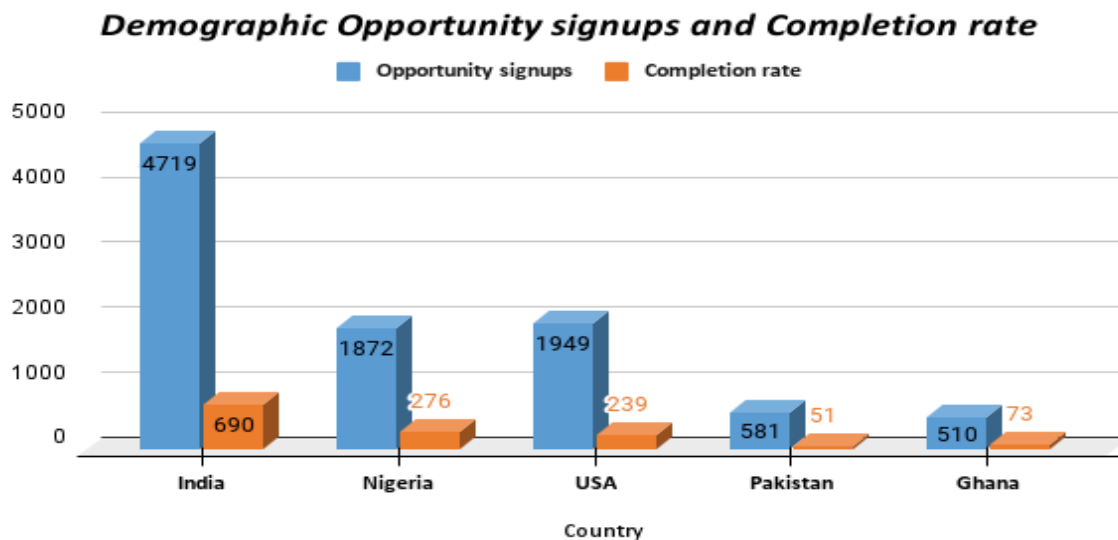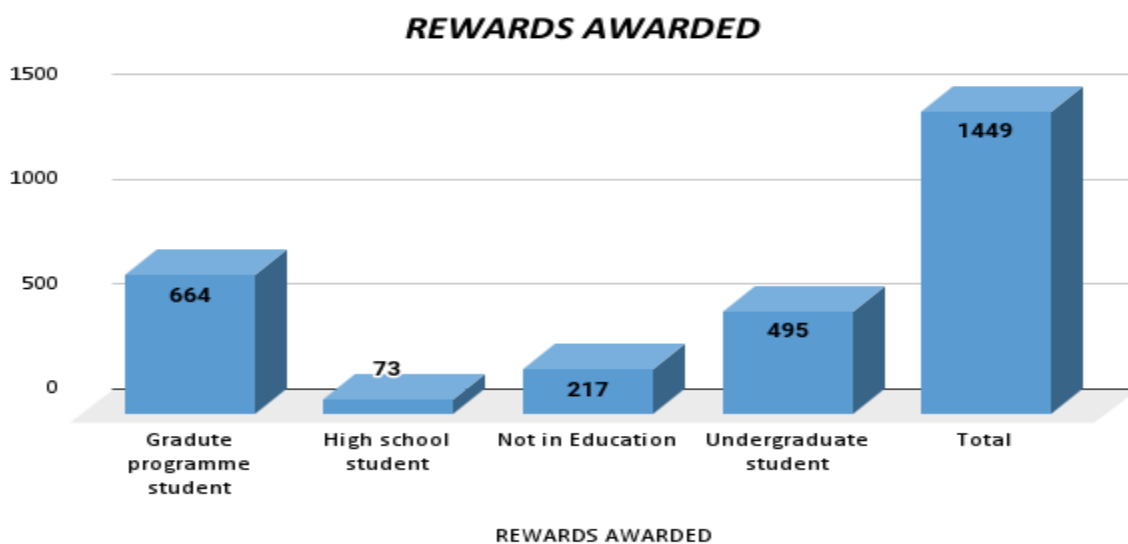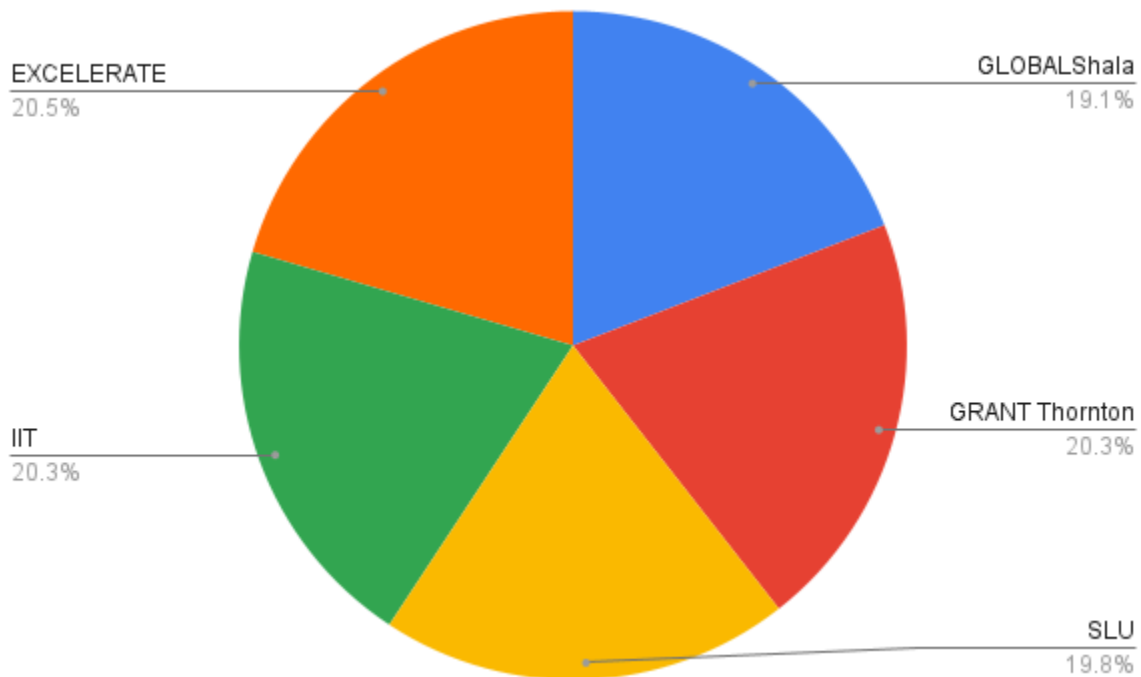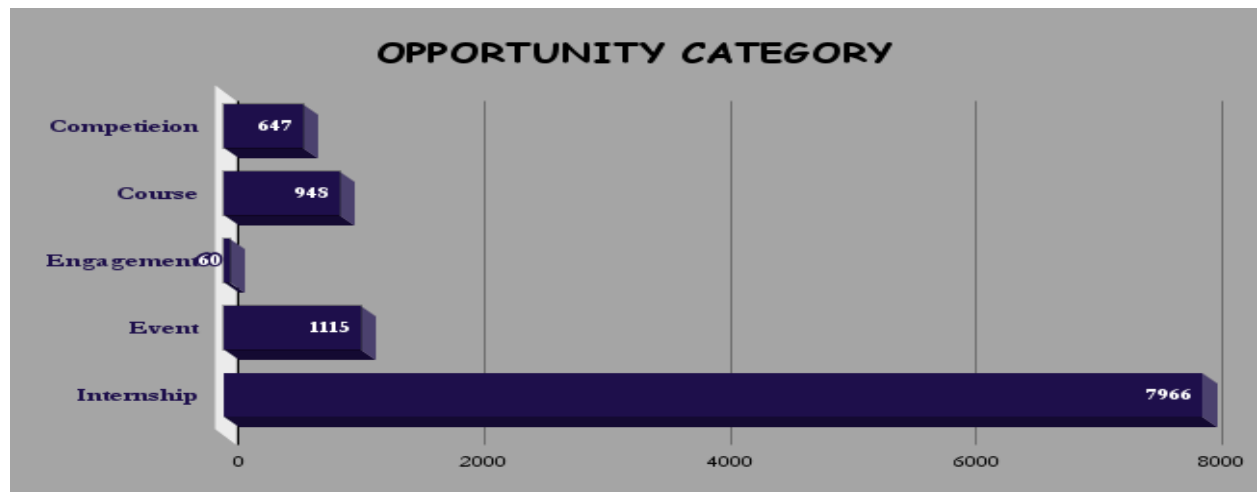
CHART 1

CHART 2

The chart below shows the demographic distribution of opportunity signups of the top five countries including India, which is the highest, Nigeria, USA, Pakistan and Ghana and also the completion rates per region.

Demographic Opportunity signups and Completion rate

Opportunity signups ▮   Completion rate ▮

India: 4719 (signups), 690 (completion)
Nigeria: 1872 (signups), 276 (completion)
USA: 1949 (signups), 239 (completion)
Pakistan: 581 (signups), 51 (completion)
Ghana: 510 (signups), 73 (completion)

Country

The rewards awarded chart below shows student statuses for completed opportunities which are also with their rewards awarded. With 1449 as the total number of completed opportunities, the graduate programme student showed the highest value with the high school students in the low with just 73 opportunities completed.

REWARDS AWARDED

Gradute programme student: 664
High school student: 73
Not in Education: 217
Undergraduate student: 495
Total: 1449

REWARDS AWARDED

The opportunity category chart shows the most preferred opportunity type using opportunity signup data. Internship opportunities showed the highest frequency by far with 7966 entries and Engagement with the lowest having 60 entries. The pie chart below shows the percentage of users preferred sponsors for each sponsor.

OPPORTUNITY CATEGORY

Category	Value
Competieion	647
Course	948
Engagement	60
Event	1115
Internship	7966

EXCELERATE 20.5%
GLOBALShala 19.1%
GRANT Thornton 20.3%
SLU 19.8%
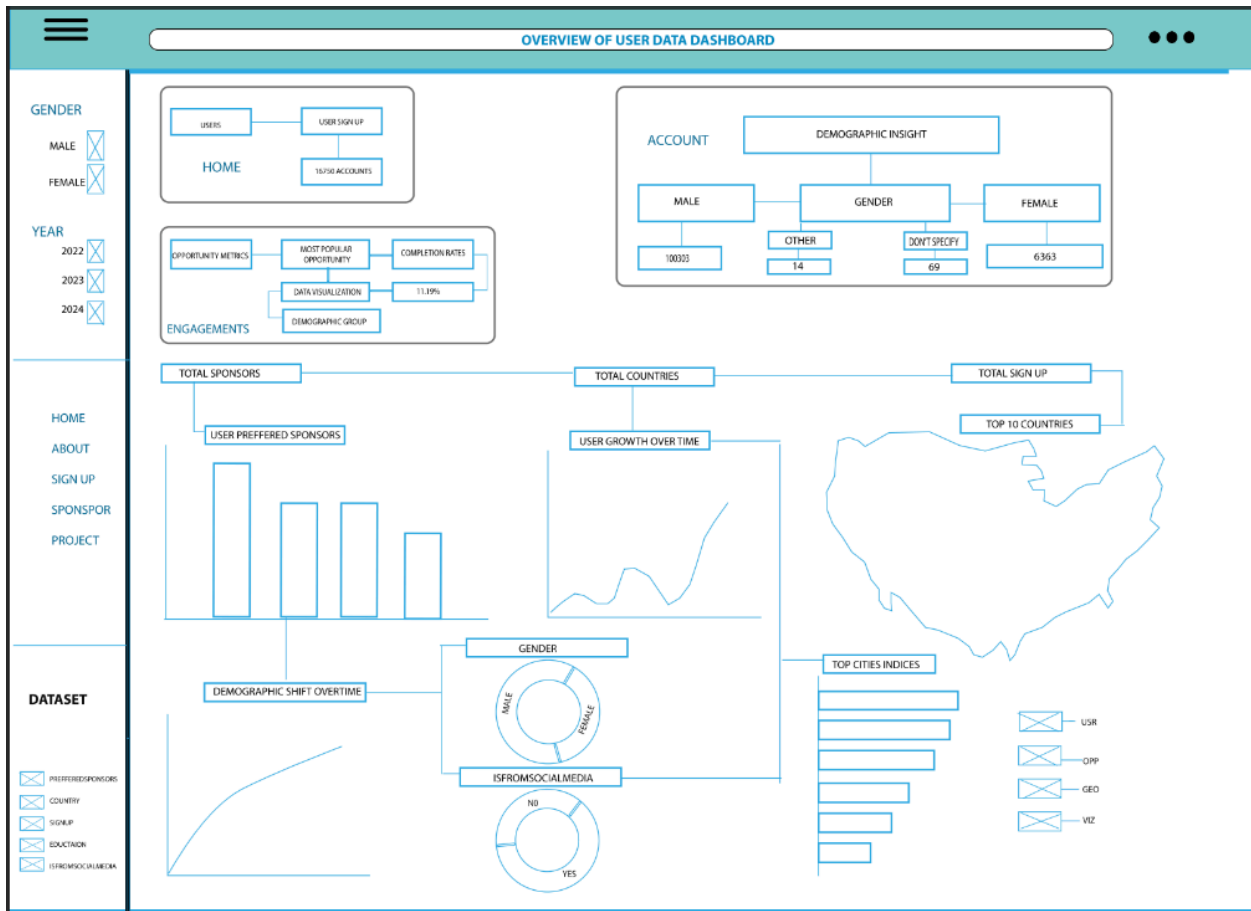IIT 20.3%

CROSS DATA SET ANALYSIS

A major challenge was faced in combining both datasets for combined analysis using different tools. Some were due to contrast in the number of rows and so on. The observation made using charts from both datasets however follows:

- The user data shows that there are 27,563 users with an account while about just below average of the figure proceeded to sign up for opportunities(11482).
- The demographic signup follows the same trend with the same top five countries for both datasets. India, USA, Pakistan, Ghana and Nigeria.
- The current student status also showed a level of consistency with the graduate programme students having the highest population for both datasets.
- The male gender has the highest level of gender for both datasets after the null entries were handled.
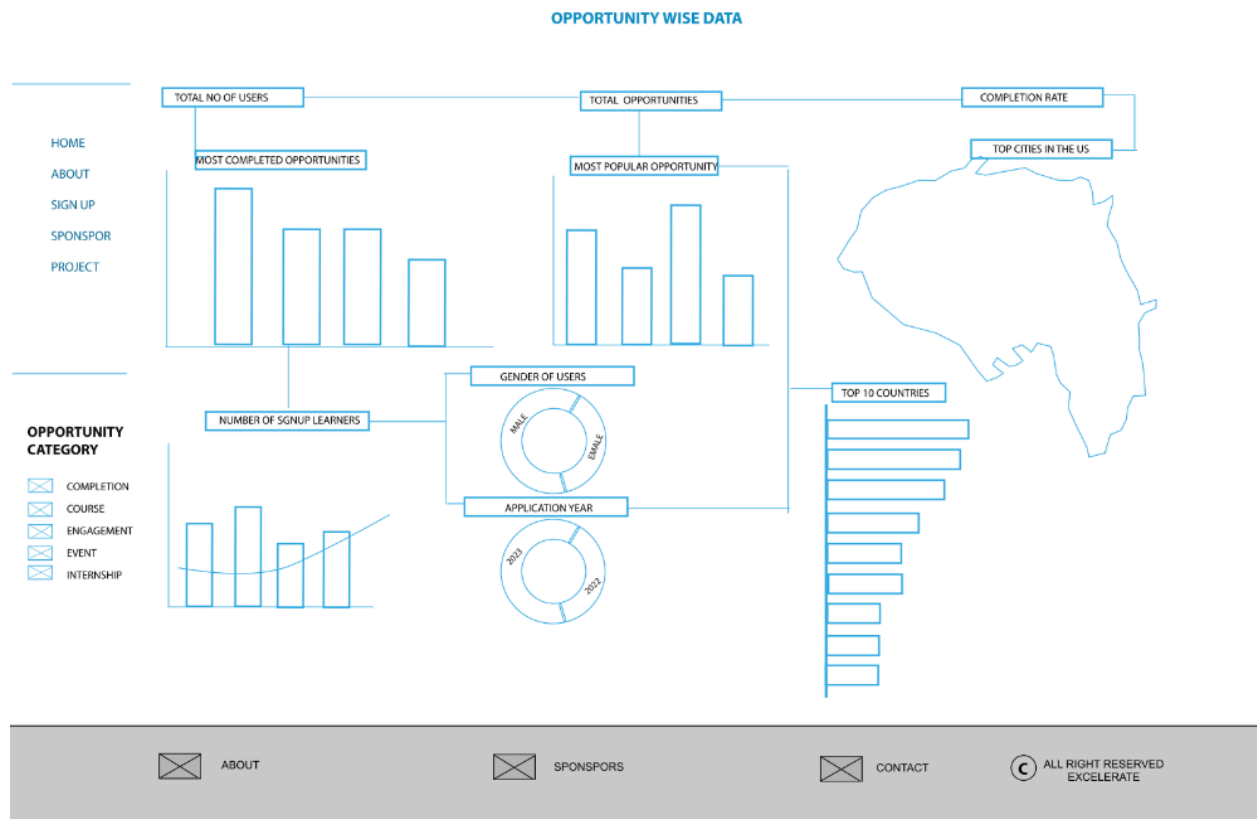
CREATING WIREFRAME

To create a wireframe, begin by defining the purpose and target audience of your project. Gather relevant information and sketch initial ideas on paper, focusing on the layout and structure. Choose a digital tool like Balsamiq, Adobe XD, or Figma for more detailed work. Establish the basic page structure, including the header, footer, navigation, and content areas. Add key elements such as buttons, forms, images, and text boxes, prioritizing functionality over visual design. Define user interactions and navigation, considering how users will move through the interface. Seek feedback from team members or stakeholders and refine the wireframe accordingly. Iterate through this process, making adjustments until you have a solid foundation. Annotate the wireframe to explain specific features and functionalities. Present the wireframe to stakeholders for approval, and once finalized, it can serve as a guide for designers and developers during the visual design and development phases.

USER DATA DASHBOARD:

Overview of user data dashboard wireframe showing sidebar filters (Gender, Year, navigation links, Dataset), Home and Account sections, Demographic Insight, and multiple data visualizations including Total Sponsors, Total Countries, Total Sign Up, User Preferred Sponsors, User Growth Over Time, Top 10 Countries, Demographic Shift Overtime, Gender, IsFromSocialMedia, and Top Cities Indices.

OPPORTUNITY WISE DATA:

OPPORTUNITY WISE DATA

ANNOTATIONS AND EXPLANATION

USER DATA DASHBOARD:

Data visualization: The dashboard uses a variety of charts and graphs to visualize the data, including bar charts, pie charts, and line graphs. This is a good way to present a variety of data in a way that is easy to understand.

Data selection: The dashboard allows users to select a specific year from a dropdown menu. This is a good way to focus on the data that is most relevant to the user.

User information: The dashboard shows some basic information about the user, such as their gender and the country they are from. This can be helpful for understanding the context of the data.

OPPORTUNITY WISE DATA:

The dashboard seems to be focused on opportunity data, with sections for the total number of users, total opportunities, completion rate, most completed opportunities, most popular opportunities, and top cities in the US.
There is also a section for user gender and top 10 countries.
The layout is fairly simple, with a sidebar for navigation and a main content area.
The data is visualized using a variety of charts and graphs.

FLEXIBILITY AND ITERATIONS

*Embrace low-fidelity: Start with low-fidelity wireframes – simple sketches or diagrams focusing on layout and information hierarchy. This makes it easier and faster to iterate and adapt as new information emerges.

*Utilize modular components: Design key elements like buttons, cards, and charts as independent, reusable components. This allows you to easily swap or rearrange them as needed without affecting the overall structure.

*Focus on information architecture: Prioritize the organization and flow of information over visual details. This ensures the core structure remains adaptable even if the visual design evolves.

*Leverage white space: Leave ample white space on your wireframes. This provides room for incorporating additional data points or functionalities without feeling cluttered.

*Document assumptions and considerations: Clearly document any assumptions you make or design decisions you take. This transparency helps everyone understand the rationale behind the wireframe and facilitates smoother future adjustments.

*Use digital wireframing tools: Consider using digital wireframing tools that allow easy editing and version control. This makes it simple to track changes, revert to previous iterations, and collaborate with stakeholders.