***Exploratory Data Analysis (EDA) report***

***Introduction***

The Exploratory Data Analysis (EDA) report aims to give a clear and thorough understanding of two important datasets used on the Excelerate platform, making it easier for everyone to understand what's inside these datasets and how they are cleaned and checked for accuracy. The first dateset, called "User Data," is all about the people who have signed up on Accelerate. It's pretty big, with 27,563 rows and 8 columns, giving us lots of information about who these users are and what they do on the platform. The second dateset, "Opportunity Sign Up and Completion Data," is more focused on what learners do on Accelerate. It tracks their activities and progress in different opportunities offered by the platform and has 20,323 rows and 19 columns.

The big task with these datasets is to make sure they're clean and accurate. This means we have to do a bunch of things like finding and fixing any missing or wrong information, getting rid of data that doesn't make sense or is repeated, and making sure everything is formatted in a way that's easy to understand and use. This cleaning process is super important because it helps ensure that when we look at the data, we can trust what it tells us. We check everything carefully, like making sure the numbers and categories in the data are correct, and that the relationships between different pieces of data make sense. We also closely monitor any changes we make to the data, so we know exactly what was done and why.

Doing all this work on the datasets is a big deal because it helps us understand what's going on with the users on Excelerate and how they interact with the opportunities available. By taking these careful steps, we make sure that the data we're looking at is reliable and can give us good insights into how to make the Excelerate platform even better for its users. This report, by breaking down and explaining the data cleaning and validation steps in simple terms, aims to make the data more accessible and understandable for everyone, showing the value and potential of the information gathered on Accelerate.

***Data Overview***

**User Data**:

The first dateset is about people who have signed up for Excelerate. It's like having a big list of users, and there are 27,563 people on this list. Each person is described using 8 different pieces of information, like their name and what they do on Excelerate. To tell each person apart, we use something called a 'Profile ID,' which is like a special number just for them.

**Opportunity Sign-Up and Completion Data**:

The second dateset is more about what these users do on Excelerate. It's like keeping track of their activities and progress. There are 20,323 different activities recorded in this dataset, and for each activity, there are 19 pieces of information. These tell us things like when someone signed up for an opportunity and if they completed it. To connect all this info back to the right person, we also use that 'Profile ID' we talked about earlier.

These datasets are like puzzle pieces that, when put together, can help us understand how people use Excelerate and what they do on the platform. The 'Profile ID' is like the glue that sticks everything together, making sure we know which person did what.

***Column Analysis***

Dateset 1: User Data

* Profile Id (Unique Identifier): Numeric (No issues)
* Name: Text (String) (Potential missing values or inconsistent formats)
* Gender: Categorical (e.g., Male, Female, Other) (Check for missing values and consistency)
* Age: Numeric (Check for outliers)
* City: Text (String) (Examine for missing values and potential inconsistencies)
* Country: Text (String) (Check for missing values and validate country names)
* Signup Date: Date (Ensure consistent date format and check for missing values)
* Last Login: Date (Ensure consistent date format and check for missing values)

Dateset 2: Opportunity Sign-Up and Completion Data

* Profile Id (Unique Identifier): Numeric (No issues)
* Opportunity Id (Unique Identifier): Numeric (No issues)
* Opportunity Name: Text (String) (Examine for missing values and potential inconsistencies)
* Opportunity Category: Categorical (e.g., Education, Training, Workshops) (Check for missing values and consistency)
* Sign Up Date: Date (Ensure consistent date format and check for missing values)
* Completion Date: Date (Ensure consistent date format and check for missing values)
* Status: Categorical (e.g., Completed, In Progress, Not Started) (Check for missing values and ensure consistency)

This Excel-like format provides a quick overview of each column's data type and potential issues in both datasets, making it easier to identify areas that may require further attention and cleaning.

***Profile ID Analysis***

In the analysis of two distinct datasets, "User Data" and "Opportunity Sign Up and Completion Data," a key focus was on the evaluation of Profile IDs. The unique\_profile\_ids\_count variable, derived from the 'Profile Id' column using nunique(), revealed 11,481 unique IDs in the "Opportunity Sign Up and Completion Data" dataset. Alongside this, 16 instances of duplicate Profile IDs were identified through the duplicated() function, highlighting potential data entry errors. Additionally, the missing\_ids variable flagged rows with null or NaN values in the 'Profile Id' column, emphasizing the need for data integrity. The "User Data" dateset, however, presented a different scenario; it lacked a dedicated 'Profile ID' column, posing challenges in tracking and differentiating individual users. Conversely, the "Opportunity Sign Up and Completion Data" dateset not only included Profile IDs but also Opportunity IDs, allowing for detailed tracking of learner engagement with various opportunities. This distinction in the datasets underscores the importance of unique identifiers for comprehensive data analysis, with the "Opportunity Sign Up and Completion Data" providing a more structured approach to user behavior and participation analysis. The datasets also varied in content: "User Data" primarily contained demographic information without numerical data for basic statistical analysis, while the "Opportunity Wise Data" included both demographic and opportunity-related data, complete with numeric columns like 'Reward Amount' and 'Skill Points Earned' for more detailed analysis.

***Basic Statistics***

**UserData (2):**

Column: Zip Code analysis:

The primary goal of the offered code is to examine and display the zip code distribution inside the Data Frame. Most\_used\_zip = df['zip'] is the crucial line of code. The value\_counts() method.Idxmax(), which identifies the maximum commonly occurring zip code. Subsequently, the code creates statistics for a pie chart by employing counting instances of the maximum used zip code and others. Using Matplotlib, a pie chart representing the distribution with labels like "Most Used Zip" and "Others" is produced. The output graphically illustrates the '63108' zip code's supremacy, which stands at 91.18%, with 'Others' accounting for a combined 0.2%.

**Opportunity Wise Data(dataset):**

Column: 'Reward Amount':

Mean (Average): The typical prize is approximately $1081.26. This is computed by taking the total number of awards and dividing it by the large number of entries.

Middle Value (Median): $500 is the middle, or median, value. It is a great degree for the relevant component since it shows the cost where half of the benefits are below and half are above.

Range: (from lowest to highest) Rewards range from a minimum of $50 to a maximum of $2500.

Column: 'Skill Points Earned':

Mean (Average): Talent points earned are approximately 1186 on average. This is computed in a manner akin to that of the reward quantity proposed.

Middle Value (Median): The middle price, or median, is 1182. It indicates the halfway point between the ability factors' upper and lower bounds.

Range: (from lowest to highest) The range of skill factors obtained is from a minimum of 10 to a maximum of 1776.

***Opportunity Status Distribution***

**A graph showing a number of different colored squares

Description automatically generated**

The distribution of opportunity statuses, as depicted in the count plot, provides a snapshot of the prevalence of each status category within the Excelerate dateset. The horizontal bar chart visually represents the frequency of occurrences for each status, allowing for a quick and intuitive understanding of the distribution. In this context, one can observe how many learners fall into each status category, offering insights into the overall engagement and success rates within the Excelerate platform. This leads to the insights of the ‘TEAM ALLOCATED’ AND ‘REWARDS AWARDS’ categories having the highest opportunity status while ‘APPLIED’ is at its negligible values.

***Initial Observations***

In our first look at the Excelerate data, we found a few key things. The User Data doesn't have a 'Profile ID', which makes tracking users hard. The Opportunity Data, with both Profile IDs and Opportunity IDs, is better for following what learners do. Both datasets need cleaning to fix errors and remove extra data. We saw interesting stats, like a common zip code in User Data and average rewards in Opportunity Data. We also noticed different levels of learner involvement in various programs. In the next few weeks, we'll dig deeper into these areas, like the effect of missing Profile IDs and how different programs work.

***Data Visualization***

***User Data:A table with a list of items

Description automatically generated with medium confidence***

A table with text on it

Description automatically generated***A screenshot of a table

Description automatically generated***

The pie chart (fig 1) displays gender distribution with 'Female' at 50%, 'Male' at 29%, and 18% not specifying their gender. 'Other' and 'null' categories each represent 0% of the total. The chart is missing a descriptive title and percentages for 'Grand Total' and '(blank)'.

**A pie chart with different colored circles

Description automatically generated**

Figure-1

The bar chart (fig 2) shows sign-up data with 13,742 not via social media, 13,811 via social media, 9 unspecified, and a total of 27,562. The chart lacks a descriptive title and Series 2 data.

**A graph with blue bars

Description automatically generated**

Figure -2

The line chart (fig 3) shows degree levels with graduate and undergraduate students being the highest groups at 6,015 and 6,527 respectively. High school students are 1,562, those not in education are 2,646, and 442 are unspecified. The total count is 27,562. Only Series 1 is visible; Series 2 has no data. The chart lacks a specific title.

**A graph showing the number of students

Description automatically generated**

Figure-3

***Data Visualization for Sign up & Completion data.***

In the table titled 'Count of Gender' with the following data: 63 people did not want to specify their gender, 8,004 were female, 12,240 were male, and 14 were other. The total of all categories combined is 20,321.



The table shows the distribution of participants across different opportunity categories. Graduate program students are most active in internships (7,049), with a total of 9,297 participants. High school students have the lowest participation, with 753 in internships and a total of 1,153. Those not in education also participate most in internships (2,303), totaling 2,862. Undergraduate students favor internships (5,254) with a total of 7,009 participants. Overall, internships are the most popular category with 15,359 participations out of a total of 20,321 across all categories.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Count of Opportunity Category** | **Column Labels** |  |  |  |  |  |
| **Row Labels** | **Competition** | **Course** | **Engagement** | **Event** | **Internship** | **Grand Total** |
| Graduate Program Student | 526 | 796 | 84 | 842 | 7049 | 9297 |
| High School Student | 120 | 138 | 2 | 140 | 753 | 1153 |
| Not in Education | 118 | 189 | 4 | 248 | 2303 | 2862 |
| Undergraduate Student | 357 | 607 | 14 | 777 | 5254 | 7009 |
| **Grand Total** | **1121** | **1730** | **104** | **2007** | **15359** | **20321** |

The bar chart (fig 4) shows student counts by gender and education level, with male graduate and undergraduate students being the most numerous, followed by females in the same categories. High school and non-student categories are lower in number. 'Other' and unspecified genders are the least common across all levels.

Figure-4

The bar chart (fig 5) shows that internships are the most popular activity for Graduate and Undergraduate Students. High School and Not in Education groups show lower participation across activities.

Figure-5

The line chart (fig 6) shows that Internship participation is highest for Graduate and Undergraduate Students, with a dip for High School Students. Other activities have consistently lower involvement across education levels.

Figure-6

***Challenges Faced***

In exploring the Excelerate data, we faced a few problems. The biggest one was that the User Data didn't have a 'Profile ID', making it hard to track users across the data. Also, there were missing or unclear details in important fields like names, gender, cities, and countries. Dates for when users signed up or last logged in sometimes had format issues. Another issue was finding duplicate Profile IDs in the Opportunity Data, which could mean mistakes were made when entering the data. These issues highlight the need for careful checking and cleaning of the data to make sure it's accurate and useful for analysis.

***Next Steps***

For Week 2, we plan to fix the problems we found in the first week's look at the data. We'll clean up the data, which means fixing any missing or wrong information. A big part of this will be adding a way to identify each user in the User Data since it's missing right now. We'll also take a closer look at the Opportunity Data to better understand how users are interacting with different parts of the platform. Our goal is to make sure our data is as accurate and reliable as possible, so we can use it to make good decisions later on. This week will be all about setting things up right for our future work.