# EXPLORATORY DATA ANALYSIS REPORT

**BY**

**TEAM 14A DATA VISUALISATION EXCELERATE INTERN**

## Introduction

Excelerate provided us with two datasets namely “User Data” and “Opportunity Wise Data” which will be referred to as UD and OWD respectively. And this document is an exploratory data analysis report for the two datasets. All analysis and observations were done in Excel and PowerBI.

## Data overview

The dimensions for the given datasets are as follows; For UD, we have 27,562 records, 9 variables and for OWD, we have 20,322 records, 21 variables. Records refer to rows and variables refer to columns. This was obtained by looking at the row and column count displayed on the  status bar in Excel. It is important to note that the headers were not included in the numbers of records. A unique identifier (profile Id) was noticed in the OWD data set.

Below are pictorial sample of  the datasets.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Profile Id | Opportunity Id | Opportunity Name | Opportunity Category |
| 2 | 31ce84c2-2bd1-40ba-b2d8-f164fe125306 | 00000000-0G4F-19XB-EXPW-KS8F3N | Statement of Purpose (SOP) Writing Workshop | Event |
| 3 | 36814990-f854-4f76-8c63-91f27567d080 | 00000000-0G4F-19XB-EXPW-KS8F3N | Statement of Purpose (SOP) Writing Workshop | Event |
| 4 | 8154328c-f8fe-4bd1-af05-783e140f68b5 | 00000000-0G4F-19XB-EXPW-KS8F3N | Statement of Purpose (SOP) Writing Workshop | Event |
| 5 | a83abad6-db1e-44c4-a8f4-9e397e282d73 | 00000000-0G4F-19XB-EXPW-KS8F3N | Statement of Purpose (SOP) Writing Workshop | Event |
| 6 | c2b8a15f-2ba3-41e4-a553-7ca68b0d4a54 | 00000000-0G4F-19XB-EXPW-KS8F3N | Statement of Purpose (SOP) Writing Workshop | Event |
| 7 | 061389be-8094-47f3-a87b-8bffff80ed3a8 | 00000000-0GT8-HCVB-01AE-6QEP8Y | Life Beyond Saint Louis University's Campus | Event |
| 8 | 2b39f489-0bb7-4ea2-9a5f-de98868c4ec3 | 00000000-0GT8-HCVB-01AE-6QEP8Y | Life Beyond Saint Louis University's Campus | Event |
| 9 | 2f5a5a25-1c68-4c7b-abe6-9c0c8c71e7d8 | 00000000-0GT8-HCVB-01AE-6QEP8Y | Life Beyond Saint Louis University's Campus | Event |
| 10 | 31e6a5e7-ed26-4604-b12b-6bf6b37ac05e | 00000000-0GT8-HCVB-01AE-6QEP8Y | Life Beyond Saint Louis University's Campus | Event |

***The above image is for OWD***

| | A | B | C |
|---|---|---|---|
| 1 | PreferredSponsors | Gender | Country |
| 2 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | Male | Nigeria |
| 3 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | Male | India |
| 4 | ["GlobalShala","Illinois Institute of Technology","Saint Louis University","Grant Thornton China","Excelerate"] | No response | India |
| 5 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | No response | Albania |
| 6 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | Female | Ghana |
| 7 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | Female | India |
| 8 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | No response | Nigeria |
| 9 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | No response | United States |
| 10 | ["GlobalShala","Grant Thornton China","Saint Louis University","Illinois Institute of Technology","Excelerate"] | Male | Nigeria |

***The above image is for UD***

Below is a table showing  variable heading and definition.

| Column Heading | Definition |
|---|---|
| **PreferredSponsors** | On the Excelerate Platform, learners can choose their sponsors ie, who they want to see opportunities from. This column shows the different values selected by the learner who has signed up for the platform. Learners can choose one or more sponsors |
| **Gender** | Shows the gender indicated by the user upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **Country** | Shows the country which the learner has indicated they live in upon sign up. |
| **Degree** | Shows the academic level indicated by the user upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **Sign up date** | Date on which they created their Excelerate account |
| **city** | Shows the city which the learner has indicated they live in upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **zip** | Shows the zip code of the city which the learner has indicated they live in upon sign up. This is not a mandatory field for signing up, hence could be missing for some learners. |
| **isFromSocialMedia** | Shows whether the learner has signed up via a social media login. If True, they have signed up via Google Login. If False, they have manually signed up |

*Variable description for UD*

| Column Heading | Definition |
|---|---|
| **Profile Id** | Unique Alpha Numeric Identifier on Excelerate for their profile |
| **Opportunity Name** | Name of which opportunity (experience) they participated in |
| **Opportunity Category** | The category of the experience they participated in- internship/event/competition/course |
| **Gender** | Given Gender of the student on Excelerate |
| **City** | Given City of the student on Excelerate |
| **State** | Given State of the student on Excelerate |
| **Country** | Given Country of the student on Excelerate |
| **Current Student Status** | Have they identified themself as High School/Undergrad/ Graduate Student or Not in Education |
| **Current/Intended Major** | The major they are currently pursuing or would want to pursue |
| **Status Description** | What is the status of their application right now. There are 8 possible statuses:<br><br>1. APPLIED: The learner has made an application (applied) and shown interest in participating in that particular opportunity (experience) on Excelerate. Their application has not been evaluated as yet to be accepted or rejected.<br>2. TEAM ALLOCATED: The learner has been accepted to participate in the opportunity and has been allocated a start date for begining the same.<br>3. DROPPED OUT : The learner has left the opportunity midway without completing it<br>4. NOT STARTED : The learner did not appear on the given start date, and hence did not start the opportunity<br>5. REJECTED : The learner did not meet the elgibility criteria for participating in the particular opportunity and hence was not accepted<br>6. REWARDS AWARD : The associated rewards for the opportunity (badge/scholarship) have been awarded to the learner. The opportunity is COMPLETED by the learner. |

*Variable description for OWD*

Next, we look at the datatype (number formats) of the datasets

First contact with the dataset presented all column datatype as 'General' on Excel on both datasets, however we changed the data type for a few shown below;

For UD, we have, after splitting Sign up Date into two separate column

| **Column** | **Datatype** |
|---|---|
| Sign up Date | Long Date |
| Sign up Time | Time |

For OWD, we have

| **Column** | **Datatype** |
|---|---|
| Opportunity End Date | Long Date |
| Apply Date | Long Date |
| Opportunity Start Date | Long Date |
| Reward Amount | Number |
| Skill Points Earned | Number |

## Profile Id analysis

From our observation, a unique identifier was only found in the OWD dataset, this is Profile ID column. This id is found to be unique to every single record and hence a check for duplicates was carried out. Using profile Id as a unique identifier, duplicates were removed from the datasets leaving 11, 481 rows. Removing missing values left us with 11,480 rows.

Other columns with missing values where;
Gender was filled with "No response".

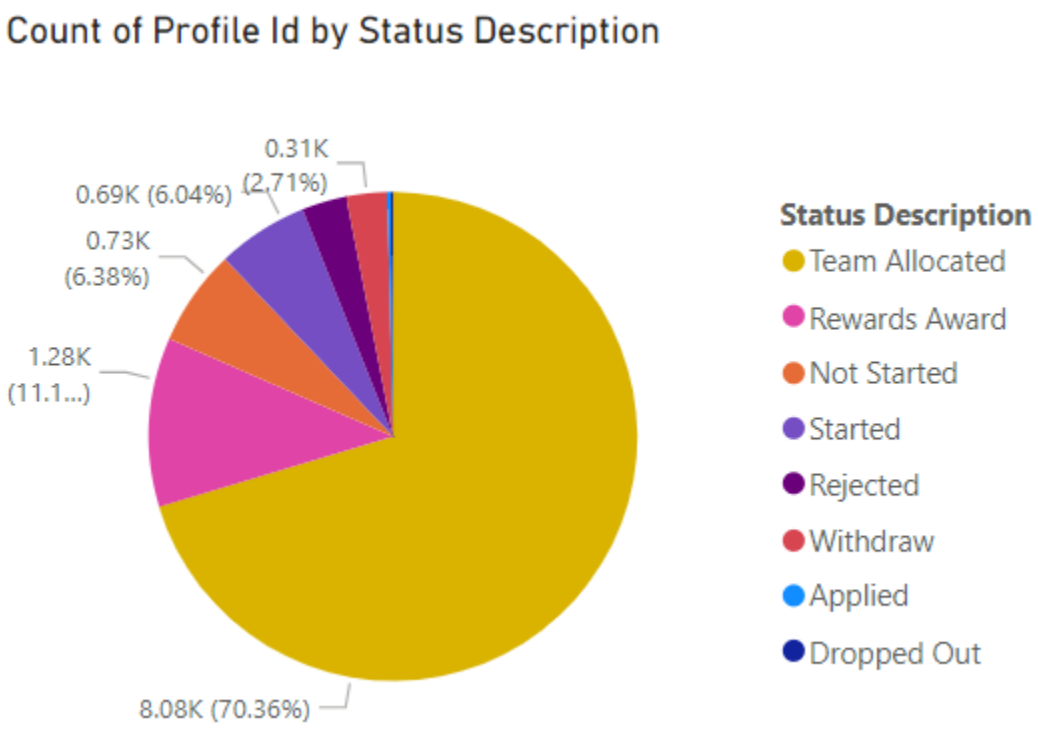
Opportunity start date, Badge id, badge name,Award amount, Skill point earned etc.

## Opportunity status distribution

This column presented us with insight about the current state of the opportunity as at the time the data within data set was collected.

| Status Description | Count of Profile Id |
|---|---|
| Team Allocated | 8077 |
| Rewards Award | 1284 |
| Not Started | 732 |
| Started | 693 |
| Rejected | 340 |
| Withdraw | 311 |
| Applied | 26 |
| Dropped Out | 17 |

**Status Description Distrubution for OWD**




**Graphical representation for Status Description Distrubution for OWD**

## Basic statistics

we calculated the basic statistics (mean, median, min, max) for Reward Amount, Skill Points Earned because of number datatype.

| Column Name | Mean | Median | Sum | Min | Mode | Max |
|---|---|---|---|---|---|---|
| Reward Amount | 119 | 0 | 502700 | 0 | 0 | 2500 |
| Skill Points Earned | 103 | 0 | 431896 | 0 | 0 | 1776 |

Worthy of note is how mode for both datasets is 0, the reason for this is all empty cells were replaced with 0, making it the highest occurring value within the dataset.

## Initial Observations

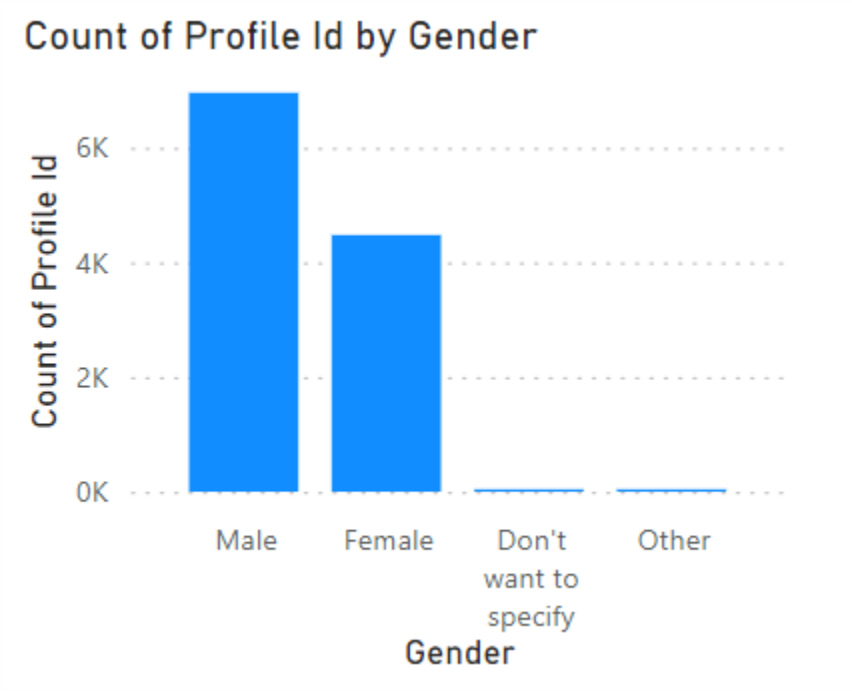
For OWD, we observed there were only 6 unique opportunities. We would love to observe further into what these specific ID's are connected to.  We also noticed the opportunity category and suggest deeper examination be carried on that column along with the country, state and perhaps gender to see if there's any correlation between the opportunities chosen and the demography characteristics. Possibly create a marketing focus based on insight found therein.

For UD, further analysis can be carried on the gender, country and "isfromsoialmedia" columns to inspect our clients social media presence and what can be leveraged with that to boost client's reputation on the internet.

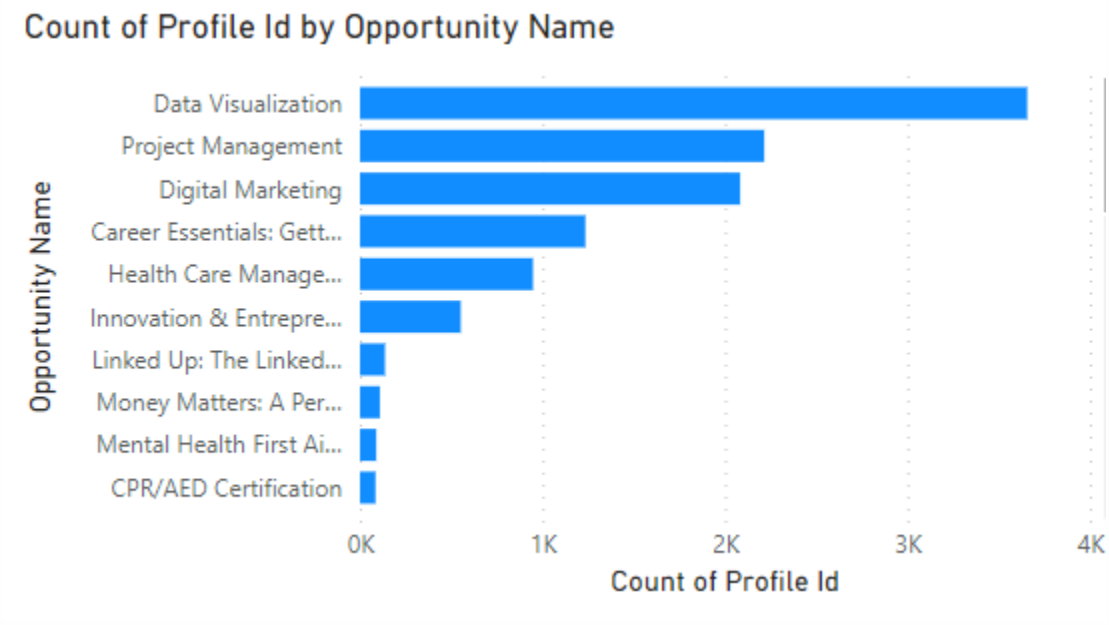
## Visualizations

Below are basic visualizations about insights on gender and oppurtunity name obtained from opportunity wise dataset.

| Gender | Count of Profile Id |
|---|---|
| Male | 6957 |
| Female | 4477 |
| Don't want | 38 |
| Other | 8 |



| Opportunity Name | Count of Profile Id |
|---|---|
| Data Visualization | 3656 |
| Project Management | 2215 |
| Digital Marketing | 2083 |
| Career Essentials: Getting Started with Your Professional Journey | 1235 |
| Health Care Management | 948 |
| Innovation & Entrepreneurship | 552 |
| Linked Up: The LinkedIn Makeover Workshop | 137 |
| Money Matters: A Personal Finance Workshop | 109 |
| Mental Health First Aid Workshop | 89 |
| CPR/AED Certification | 85 |



## Challenges faced

Dealing with missing values, misspellings and inconsistent inputs from users. This happened mainly in the geographically related features. A possible solution is to validate the data received by including dropdown options to select specific locations from a list of countries and states.