Data Analysis Intern - Assignment

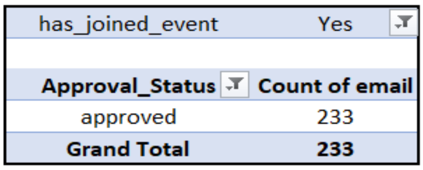
Step 1: Data Cleaning and Preparation Summary

- Imported necessary Python libraries (pandas, numpy)
- Loaded the dataset (shape = (648, 14))
- Conducted initial exploration using .head() and .shape
- Analyzed and summarized missing values across columns
- Removed rows with missing values in critical fields: email, Job Title, and LinkedIn profile
  → Resulting shape after removal: (636, 14)
- Standardized has_joined_event values to lowercase (yes, no)
- Converted amount, amount_tax, and amount_discount columns to numeric type, coercing invalid entries to 0
- Extracted created_at into two separate columns: created_date and created_time
- Final dataset shape: (637, 16) after transformations
- Exported the cleaned dataset as cleaned_data.csv for further analysis in Excel

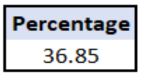
Step 2: Analyze Key Metrics

1. Conversion Funnel Metrics

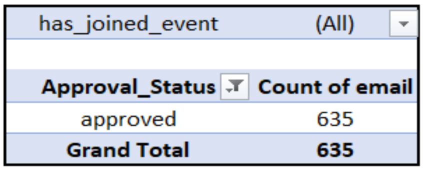
- Created a pivot table to analyze the approval_status and has_joined_event fields.
- Computed the total number of approved users by filtering for approval_status = Approved.
- Calculated the number and percentage of approved users who joined the event based on the normalized has_joined_event values.
- Derived show-up and no-show rates to evaluate event attendance effectiveness.



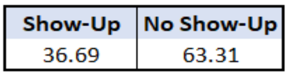
has_joined_event	Yes
Approval_Status	**Count of email**
approved	233
Grand Total	**233**



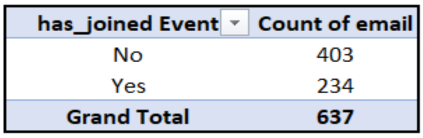
Percentage
36.85



has_joined_event	(All)
Approval_Status	**Count of email**
approved	635
Grand Total	**635**



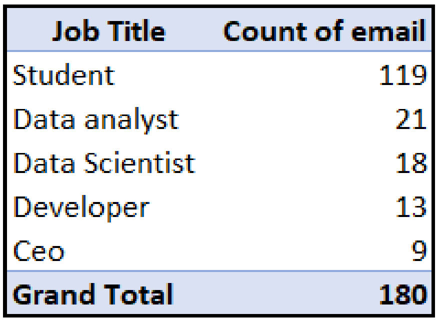
Show-Up	No Show-Up
36.69	63.31



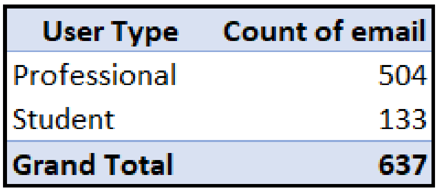
has_joined Event	Count of email
No	403
Yes	234
Grand Total	**637**

2. Job Role Insights

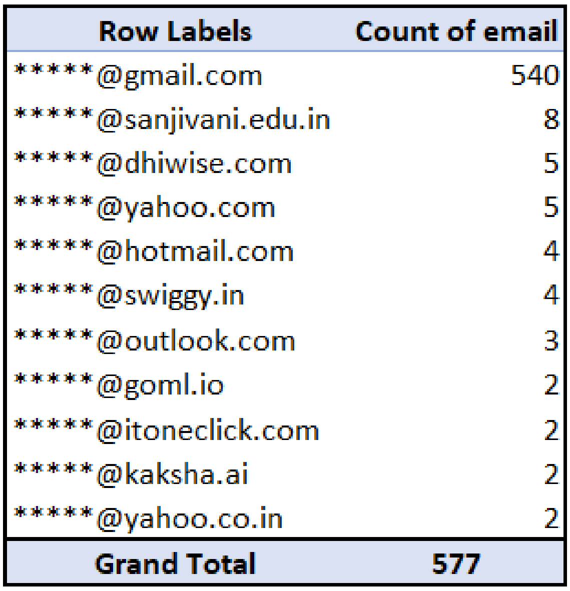
- Generated a frequency table of Job Title values using a pivot table to extract the top 5 most common roles.
- Introduced a helper column to classify users as "Student" or "Professional" based on keyword matching within job titles.
- Used pivot tables to determine the proportion of students versus working professionals.
- Identified duplicate or suspicious entries by applying conditional formatting and counting duplicates in the email column.



Job Title	Count of email
Student	119
Data analyst	21
Data Scientist	18
Developer	13
Ceo	9
Grand Total	**180**



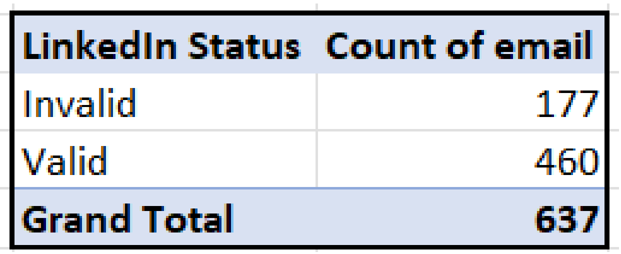
User Type	Count of email
Professional	504
Student	133
Grand Total	**637**



Row Labels	Count of email
*****@gmail.com	540
*****@sanjivani.edu.in	8
*****@dhiwise.com	5
*****@yahoo.com	5
*****@hotmail.com	4
*****@swiggy.in	4
*****@outlook.com	3
*****@goml.io	2
*****@itoneclick.com	2
*****@kaksha.ai	2
*****@yahoo.co.in	2
Grand Total	**577**

3. LinkedIn Presence

- Added a helper column with a formula to validate the LinkedIn URLs, checking for the presence of "linkedin.com/in/" and a minimum character length.
- Labeled each entry as "Valid" or "Invalid" based on this logic.
- Used a pivot table to count the number of valid LinkedIn profiles versus missing or broken ones.



LinkedIn Status	Count of email
Invalid	177
Valid	460
Grand Total	**637**