Data Analytics Project

Group 16

Author(s)

Ramya Movva: UTD Ramya.Movva@utdallas.edu Business Analytics Masters

Sujith Siddi :UTD Sujith.Siddi@utdallas.edu Business Analytics Masters

Manas Rahul Katragadda: UTD Manas.Katragadda@utdallas.edu Business Analytics Masters

Faculty Advisor

Ravishankar Narayan, rln130030@utdallas.edu

Summary

Group 16 strives to explore, analyze and make strong recommendations on taproot's nonprofit data involving columns that explain how users respond to different types of nonprofit recruitment, projects, and more. Given the amount of time and our other priorities; however, we have transformed that goal into a much more realistic one. With our limited expertise and our strengths and weaknesses, running against time, we still hope to find something in this data.

1) Problem and Motivation

We decided to try and tackle these 4 questions for the following reasons:

1. Please analyze Open rates, click thru, bounce by email type and user type (volunteer and nonprofit)

Non-profit organizations' have a harder time keeping a user than other organizations and this is most of the time because the product they sell would not benefit a user directly unlike many profit businesses. Because of this, click through rates (the number of times a user clicks on a notification or ad per the number times it was displayed) are low and bounces (or the number of times users click off a notification or ad) high. Maybe if we can figure out where are efforts are either strong or weaker based on user type or email type, we can help non-profits reduce cost while splitting their costs efficiently and effectively.

We planned to do more but sadly, we could not find the right data relationships in the short:

2. What % of nonprofit and volunteers users dropoff prior to completing registration.
3. How many sessions does a nonprofit create before starting a project?
4. How many nonprofits are conducting sessions on the same or related subject?

Because we couldn't find sufficient insights to answer the above three questions, we decided to tackle these two questions instead:

5. Please analyze Comprehensive breakdown of Project Recommendation email by day sent, day of month, time of year etc.
Breakdown of project recommendation email by day of month and time of year helps the foundation to keep track of all the project recommendations provided by different users from their mail ids. Solution to this problem mainly helps the Taproot foundation to look into project recommendations whenever they want to take up a related project and also can choose a project based on the type of user and the value that the project adds to the organization.

6. Which project types produce the best outcomes in the data?
Choosing the project based on the outcomes is better than choosing a project based on opinions of the management. So, analysis on the outcomes produced by different projects enable the foundation to look deeply on what projects to handle to maximize the value and decrease expenses. The outcome is based on the state of the project such as completed, cancelled etc.

2) Approach

Our ideal approach was to explore all of the columns in all of the tables of this data. Then we would finalize key relationships. After knowing exactly what each column contains and how it relates to the others, we would turn to working with joins to join all of the data together. Then using backend visualization software like tableau's calculated fields feature or PowerBI's data design interface or even Alteryx, if possible, we would create more calculated columns and split any necessary columns that needed to be split. Maybe even using python would help in our cleaning and preparing of this data. After all that preparation and cleaning, we would use a supervised learning approach to solve all 4 of our questions. Coupled with Tableau visualizations, we would make a possible dashboard that could help us in gathering more insight into our data. After all that, we would start on choosing the possible models for our dataset, split them into training and testing datasets with a 10:80 ratio, run them through the chosen models, then validate our findings through our validation dataset of the rest of the 10% ratio.

That would be our ideal approach, yes. Fortunately, though, because of time constraints and our own priorities getting in the way, as well as the fact it was a project done without any personal contact, we could accomplish only a few things based off this approach. Both team members Ramy and Sujith approached this constraint differently. The approach we have finally taken was Sujith's. The main idea of this approach was to first implement a data science pipeline with data extraction as the first step. During the extraction phase, we categorized the data columns based on the data type and the columns that were useful for the analysis of the problems. In the second step, we consolidated the User analysis data by concatenating all the years' spreadsheets into a single spreadsheet. Then we preprocessed the dataset by removing all the irrelevant columns and garbage values from the data using pandas. If a column contained more than 70% missing values, we dropped those columns and output new spreadsheets for analysis. In the last step, we put together all the data by creating relationships on the user ids and session ids between the datasets using Tableau. We used different visualizations such as bar chart, bubble graph, cross tab, Highlight tables. Some of our analysis results summarized different metrics for comparison between different user types and email types etc.

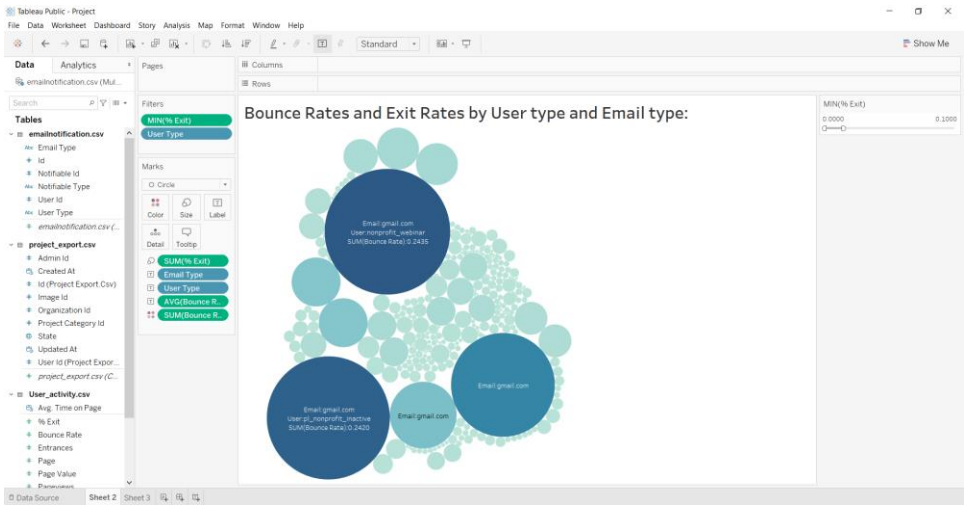
Team member Ramya had her own approach as well, but we decided Sujith's approach was much more effective, so we resolved to remove the details of her work from the rest of this paper.

3) Tools and Analytics

The Team's members, Sujith and Ramya, in charge of the rushed analytics used python Jupyter Notebook, Excel and Tableau for data preparation and data visualization. Sujith used python in Jupyter notebook to extract the session ids of nonprofit users then check if those ids are present in the main data as well. Then he used tableau to create visualizations that would be used to draw insights from.

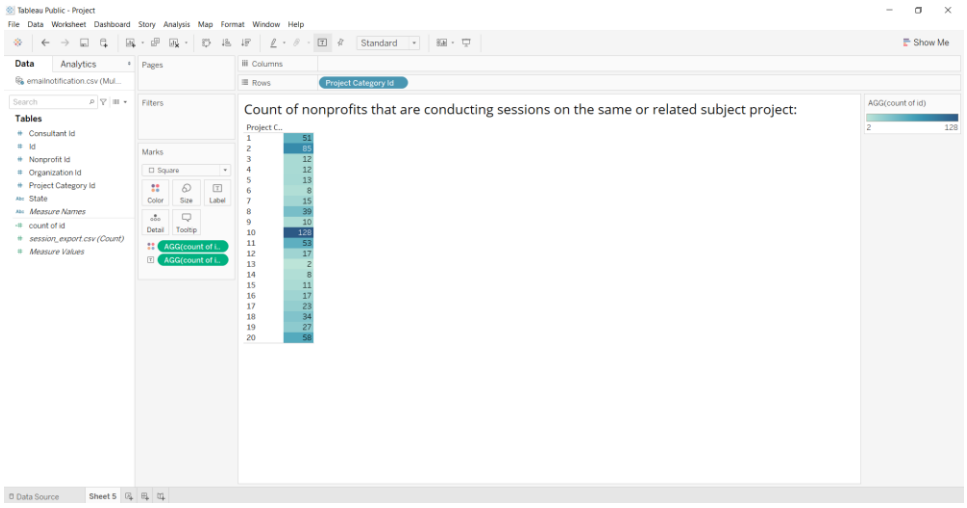
4) Results===

The below graph represents bounce rates and exit rates by email type and user type. We extracted the non-profit user ids data and summarized the average of bounce rates by email type. The email types include Gmail.com, yahoo.com etc., We extracted the average bounce rates for user type with less than 10% exit rates and found that user type nonprofit webinar has the highest average bounce rate with 24.35% with the corresponding email type being gmail.com.

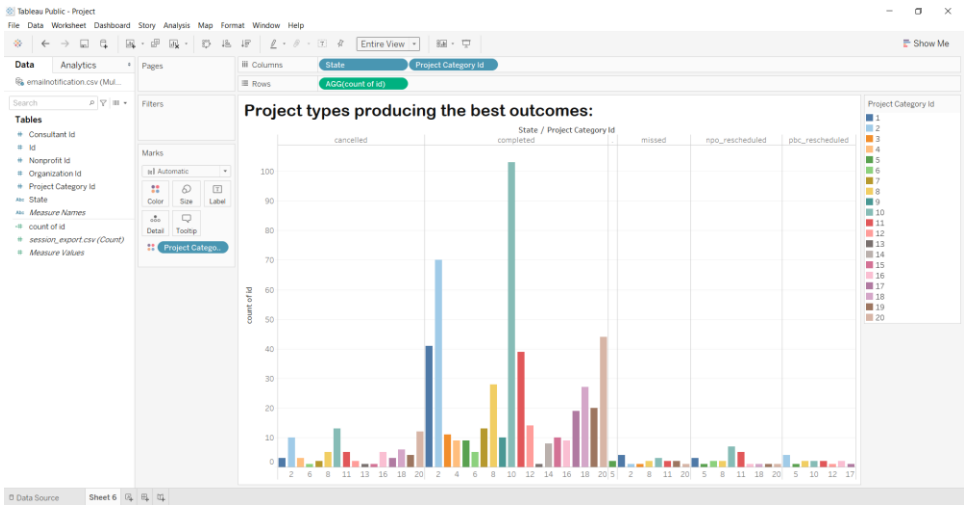


When looking at the non-profits that are conducting sessions on the same subject area as project, we grouped the project type with the number of non-profit session ids that are being generated.

The project type with id 10 had the highest number of sessions whereas the project type with id 13 had the least number of non-profit sessions.



The bar graph below depicts which project produces the best outcomes. The outcome is based on the state of the project. Given that a completed project produces the best outcome for the organization, project with id 10 had the most number of sessions completed with more than 100 sessions being completed. On the other hand, project id 13 had the least number of project sessions completed. Therefore, we can infer that project 10 produced the best outcomes.



5) References:

Teradata University

Taproot foundation

Also, Team Member Ramya has minimal in-field experience as she volunteers for a Non-Profit Organization and is working with similar but more SEO related data