**Exploring multi-attribute nonprofit survey data**

|  |  |  |  |
| --- | --- | --- | --- |
| Justin Chin | Bryan Costanza | Bret McSpadden | Ben Niu |
| University of Colorado Bolder | University of Colorado Bolder | University of Colorado Bolder | University of Colorado Bolder |

**Abstract**

Although many organizations are collecting data, few are analyzing it due to a lack of skill, time, or other factors. Using data from a nonprofit organization that had not been analyzed holistically before, we attempted to better understand the data through data visualization, build a tool that allowed the data to be presented clearly, and give feedback on the dataset to improve future analysis. As a final deliverable, we designed an interactive visualization that shows the correlation between many program factors and satisfaction of the program. Through this visualization we found that staff is the most highly correlated factor to satisfaction and were also able to display the general ratings for the variety of aspects of the organizations’ programming.

**Keywords**

Nonprofit, data visualization, Likert-scaled data, multi-attribute data

1. **Introduction**

The collection of data has become widespread throughout industry and a variety of organizations [1]. Although more and more prevalent, analyzing data presents a unique challenge for small nonprofit organizations. While demonstrating impact is a priority, and the need to utilize data to help make decisions is understood, often the resources, time, and expertise needed to collect and analyze data are not present in these organizations [2]. Therefore, although a lot of data is being generated and collected, it is rarely being analyzed as a whole [3]. The stakes for not doing so are high in an industry moving toward a data-informed model for decision making, donors, and other stakeholders.

For this project we looked at survey data from an international nonprofit which places volunteer teachers abroad for one-year contracts in partnership with local organizations or government ministries. During the volunteers’ service, the volunteers attend three trainings throughout the year: Orientation (OR), Mid-Service (MS), and End-of-Service (EOS). Following these trainings, the volunteers are given an opportunity to provide feedback on a variety of aspects of the program including feedback on training, support, and safety. The surveys are conducted online via Survey Monkey and include Likert-type data and open-ended responses.

As described in the research, the organization collects a lot of data, but has not analyzed the data as a whole, instead typically looking at individual cohorts’ responses following the trainings to see individual feedback on staff and any outliers regarding safety or other challenging factors to help staff address those challenges for the volunteers. Although the organization had a general sense of some of the feedback, it was not possible for them to look more broadly at the data to look for trends or correlations.

With these challenges in mind, we were interested in looking at how data visualization can help the organization better understand the feedback it receives. We were also interested in providing feedback on the collection of data and the unique challenges this dataset presented to help improve collection methods which may in turn help improve analysis and understanding.

1. **Related Work**

As part of the design process we looked at a variety of resources to better understand Likert-type data and how it is visualized. The foundations of this datatype are in behavioral science, but now used widely to attempt to measure sentiment and other qualitative metrics quantitatively [4]. The methodology for analyzing Likert data is vast, including a variety of statistical measures, and has been criticized, but research has shown that some of these criticisms are unfounded [5]. For the purposes of this project, we focused more on visualization techniques, but more robust statistical analysis of the data, beyond Pearson’s r, would be helpful in better understanding the data.

In researching visualization of Likert data, stacked bar charts, often diverging, appear to be a standard for visualizing individual questions and responses [6]–[8]. Less research was found in comparing Likert-scaled data to other questions, but scatterplots are a common technique when trying to compare two Likert-scaled questions [9].

1. **Description**

The goal of our project was to utilize the visualization techniques and ideas learned in class to analyze and understand a dataset that had not been looked at from a visualization perspective. Our partner organization provided the raw data after removing personally identifiable information. The data included responses from six program countries where the organization places volunteer teachers and represents survey data collected from volunteers at three points during their service: following Orientation, Mid-Service, and End-of-Service trainings. The three surveys are different but did have consistent metrics using a Likert scale. The surveys also included some consistent questions across the surveys, particularly MS and EOS.

Initial explorations focused on understanding the surveys and pulling out questions that seemed interesting to analyze. Our initial questions influencing our exploration were, “How are different aspects of the programming rated?” and "What contributes to the growth of volunteers?" In answering these questions, we were interested in how volunteer responses changed during their service from beginning to end.

For the purposes of this project, we focused solely on the Likert-scaled questions and ignored all open-ended responses. We first looked through all the surveys and grouped questions by themes: self-assessment, education office, staff, safety, and training. We also found questions relating to satisfaction and whether the volunteer would recommend the program. Overall, only one set of questions, the self-assessment, were asked across all three surveys, but items about staff support, safety, and other program features were asked across MS and EOS surveys.

The raw data was somewhat challenging to work with due to the formatting from the Survey Monkey export. After identifying the sets of questions to analyze, we cleaned up the sheets manually to standardize the header rows and processed the data using Python in Jupiter Notebook to calculate the average responses across countries for the three surveys. We combined that data into one CSV file. Although we would have liked to look at individual responses, this was not possible because individual respondents could only be identified by their name or email address which were removed by the organization to protect the respondents’ privacy.

After cleaning the data, we used Tableau to better understand the cleaned data and prototype potential visualizations. We explored visualizing the data over time across surveys as well as individual metrics and categories of questions. During these initial explorations it was clear that there were interesting trends and unexpected outcomes in the data. Some examples such as the self-reported ratings trends, low ratings for educational support tools, and the difference between satisfaction and recommendation are included below (Figure 1, Figure 2, Figure 3).

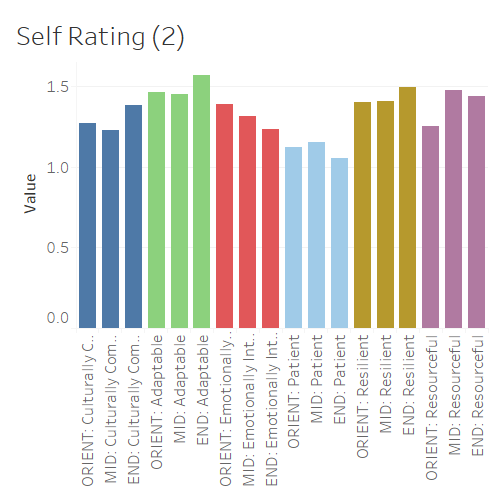


Figure 1: Self rating metrics over time

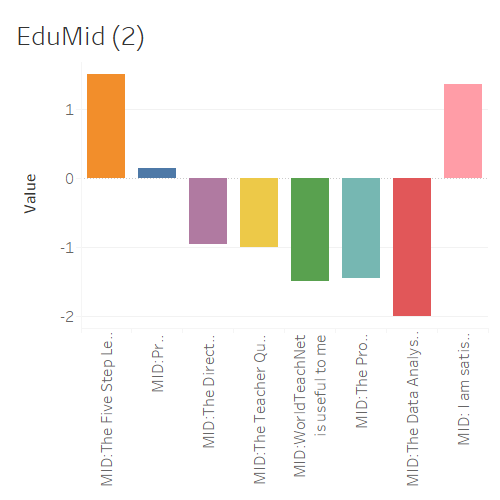


Figure 2: Education support tools, MS

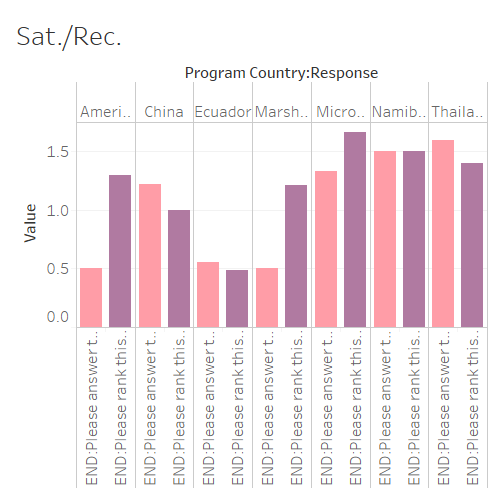


Figure 3: Satisfaction and Recommendation across programs

In prototyping it was also clear that the number and metrics across the different countries made it difficult to compare different metrics and countries easily due to having too much information. Additionally, although the static bar graphs were interesting, because of the challenges in comparing data, it was hard to further probe the data beyond simple charts. We had a better understanding of the data but had more questions around how the various questions affected satisfaction and recommendation of the various programs, as well as how the data compared among programs.

To further explore the dataset, we generated a prototype in Tableau comparing averages of individual metrics to the average satisfaction and recommendation ratings. Here it was clear there was some correlation but overdraw and general clutter made it difficult to understand exactly what was going on among all the data (Figure 4).

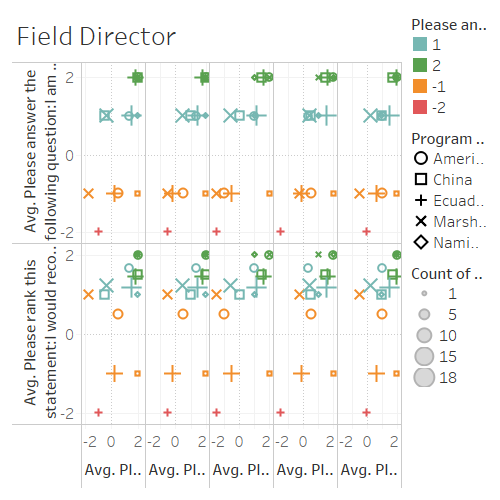


Figure 4: Staff metrics compared to recommend & satisfaction

From this prototype, we generated a new set of questions to answer to help narrow down or final design: “How is WT rated overall? By country?”, “Across programs, what factors are rated highly?”, “What factors are rated poorly?”, “Of the highly rated factors, which ones are mostly correlated with satisfaction?”, “Of the poorly rated factors, which ones are mostly correlated with dissatisfaction?”, “Outliers? Possible why?”

To accomplish the more complex task of looking at what factors correlate with satisfaction, we designed an interactive visualization to display the r value between questions and satisfaction. We chose satisfaction over recommendation because there was more variance in the satisfaction ratings, and the recommendation was generally high for all programs. Questions were manually divided up into six categories: education, impact, safety, self-reported, staff, and support. The visualization was built using the Bokeh library in Jupiter Notebook and put into a HTML webpage. In addition to using examples online for certain tasks and interactions, we used code from project two [10]. Certain programming challenges led to some design choices such as the buttons over a dropdown menu.

The scatterplot on the top left of the page is displayed throughout the visualization and displays each questions r correlation on the x-axis. The y-axis is the rank, with more highly correlated items (closer to +1) ranked higher. The y-axis is somewhat confusing because we could not find a way to invert the axis to show “1” as the highest rank at the top, so the rank is not intuitive in how we generally think of it numerically.

To further explore the correlation, the user can choose a subset of questions and view the top three correlated questions, as well as all staff questions, since those questions were the highest correlated group. The individual graphs showing particular questions are displayed alongside the r value chart and are ordered from highest to lowest correlation. In these, charts countries are represented by color and the number of respondents are represented by circle size. The tooltip can be shown by mousing over the individual circles, displaying the charted values. Overdraw was accounted for by displaying multiple tooltips at the same time when points were overlapping. We chose to only show three of each question, besides staff, since the correlation was generally weak for most categories except staff and support. Safety also only had three questions, so displaying 3 kept the visualizations consistent. This design choice does prevent the user from exploring each question individually.

In addition to exploring the correlation between different questions and volunteer satisfaction, the tool can also be used to generally see how different questions are rated by only looking at the x-axis scale and individual countries. Because the country data is divided by response, this observation is only a generalization.

1. **Findings**

In exploring this dataset, we had quite a few interesting findings. These can be broken into three categories: data, from prototyping, and from our final visualization.

From working with the data, a few things stood out. First, although the data was collected in Survey Monkey, the export from the service did not work well with Tableau without some cleaning, and needed further cleaning to display the data appropriately with readable titles for the different datasets. We also noticed that some of the questions were not scaled the same, with the majority being a 4-point scale, but some with 5-point scales or fewer than 4. Additionally, some questions were not formatted the same, creating unnecessarily messy data that forced added cleaning. Fixing these issues would make further analysis easier, particularly for less skilled programmers. Additionally, the literature shows that a 5-point Likert scale is the norm and helps create equal distance between the ratings. We compensated for this by using -1, -.5, .5, and 1 to substitute for Strongly Disagree, Disagree, Agree, and Strongly Agree.

Two other major items stood out about the dataset itself. First, because the data is collected as individual responses across the three surveys, the only individually identifiable information is a respondent’s name or email, but this information was removed at the organizations request to protect the privacy of individual responses. Although we ended up not comparing data across surveys for the final visualization, in our explorations we were forced to aggregate the data. This aggregation and lack of identifiable individual respondents made it challenging to dial down into interesting patters during the prototyping and added to uncertainty knowing that different respondents may be answering different surveys. Knowing who will process the data, and considering privacy concerns in advance, may help allow for deeper analysis of the data in the future.

Second, the survey contained a lot of open ended responses, and although these may be helpful in better understanding the data, we decided to not attempt to work with those questions, leaving out a lot of valuable information. Better capturing and mapping those responses may be helpful in further analyzing the dataset.

During prototyping it was interesting to be able to quickly look at the data visually. During these explorations it was useful to quickly show unexpected aspects of the data which helped us question our assumptions and identify interesting aspects of the data to explore further. During this process, we saw that satisfaction and recommendation did not always get rated the same (Figure 2). The self-reported data measuring resilience, patience, etc. did not follow a linear trend, and even showed some decrease in rating, which was unexpected and not an ideal result (Figure 1). We also noticed that the education support was rated significantly lower than any other factor (Figure 3).

Finally, in the designing and building of the final visualization, we were able to show that staff and support training are the two aspects of the organizations’ programming that mostly correlate with satisfaction. Staff support factors make up eight of the top ten most correlated factors. Orientation training is the most correlated support training, providing a worthwhile experience is the most correlated impact factor, and the five-step lesson plan is the most correlated education support factor. Other items were not very correlated with satisfaction.

In looking specifically at the highest correlated group, staff, it is also interesting to see that administrative items are less correlated than emotional and general support.

In exploring the individual questions, one can see that the highly correlated items follow the expected linear trend of poor rating to poor satisfaction or high rating to high satisfaction. One also can see that many items are generally rated high, regardless of correlation, except for the education items, which are generally rated poorly.

From these observations, one can infer that the organization should focus on hiring good staff members and ensuring its trainings, particularly Orientation, are well organized as these factors are the most correlated to satisfaction. It is also interesting to note that it seems that emotional support is more correlated than administrative support. Therefore, the organization should focus on supporting its staff in supporting volunteers through their service.

**References**

[1] S. Lohr, “Opinion | Big Data’s Impact in the World,” *The New York Times*, 11-Feb-2012.

[2] C. Bopp, E. Harmon, and A. Voida, “Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work,” in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 3608–3619.

[3] N. L. Maxwell, D. Rotz, and C. Garcia, “Data and Decision Making: Same Organization, Different Perceptions; Different Organizations, Different Perceptions,” *Am. J. Eval.*, vol. 37, no. 4, pp. 463–485, Dec. 2016.

[4] H. N. B. Jr and D. A. Boone, “Analyzing Likert Data,” *J. Ext.*, vol. 50, no. 2, Apr. 2012.

[5] G. Norman, “Likert scales, levels of measurement and the ‘laws’ of statistics,” *Adv. Health Sci. Educ.*, vol. 15, no. 5, pp. 625–632, Dec. 2010.

[6] R. Heiberger and N. Robbins, “Design of Diverging Stacked Bar Charts for Likert Scales and Other Applications | Heiberger | Journal of Statistical Software.”

[7] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual Analysis of Multi-Attribute Rankings,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2277–2286, Dec. 2013.

[8] S. Wexler, “How to Visualize Sentiment and Inclination,” *Tableau Software*. [Online]. Available: https://www.tableau.com/about/blog/2016/1/how-visualize-sentiment-and-inclination-48534. [Accessed: 03-May-2018].

[9] swexler, “Likert vs. Likert on a Scatterplot,” *Data Revelations*, 16-Sep-2014. [Online]. Available: http://www.datarevelations.com/likert-vs-likert-on-a-scatterplot.html. [Accessed: 03-May-2018].

[10] *project-2-ncwit-team-13 created by GitHub Classroom*. INFO 4602/5602: Information Visualization, 2018.