Library Survey Model: How to Better Serve Library Patrons Through Feedback Analysis

Jeremy Zimmett

2018

About

This project aims to understand how survey data can be used to better serve library patrons and programming. By considering the different ways survey questions can be designed and analyzed, it offers a commentary on the importance of weighted thinking and learning R Studio.

Introduction

How can we engage with our communities in more meaningful ways?

As mission-driven organizations, libraries are invested in the well-being of their communities. However, the needs of each community vary across time, region, and income. In order to address these needs, libraries offer programs that are free and open to the public. These programs include workshops in technology, self-improvement, and adult literacy. While these programs have great benefit, we must consider the scope of this benefit. How do libraries decide what programs to offer? For whom are these programs designed? On what rationale, and with what criteria in mind?

Qualitative studies try to capture and model the diverse needs of communities. One such example is the Harwood Institute Model, a structure interested in educating "library professionals to bridge divides by cultivating shared values and strengthening civic culture; helping them become recognized as local leaders who are relevant to the sustainability of community life."

The Harwood Model promotes diversity, inclusion, participation, and equity through community voices and roundtable discussions. The model asserts that program development is successful only insofar as it considers ongoing feedback from community members. The model, however, offers only the theoretical and ethical framework for program development. Once

community members have had their say, how do we utilize the responses? How do we collect the data? How do we eliminate bias in the analysis of that data?

In order to pose a solution, the following material offers a data model to address this limitation.

Data Model

Current Data

Before collecting new data, libraries must consider the information that they already possess. Most, if not all, libraries collect/assign the following information when issuing library cards: dates of birth, addresses, and library card numbers. Here, we have a unique identifier (card barcode), a geospatial location (address), and a useful demographic (age). With these elements, an effective relational analysis can be constructed in conjunction with community feedback.

New Data

In order to build a data model, a survey sampling method needs to be created. The most concise format for surveying library program feedback consists of the following four elements: origin, time, type, and awareness.

1. *Origin*. Have a point that grounds the analysis. The factors that follow will be in relation to this element. How is the survey-taker connected to the survey?
2. *Time.* Consider the availability of your community. What times/days are they available? The more granular your response options, the more useful it will be for analysis. However, with granularity comes plurality. Too many options can overwhelm survey-takers, skewing results.

3. *Type*. Offer a list of possible program offerings. I suggest making the wording broad (e.g. "Arts/Culture" or "Science/Technology").
4. *Awareness.* List the way you currently (and could potentially) market programs. What is the best way to get in touch with your community? How do they hear about events? How do they want to be informed?

The questions that you generate from these elements should remain constant over time. In order to perform a longitudinal analysis of your data, consistency is key. The possible responses to these questions should come from a pick-list. Once again, consistency from a controlled vocabulary is essential. However, responses can be modified over time to meet changing community interests. Design your questions and responses based off the feedback from roundtable discussions.

Example Model:

- *Origin.* What library event did you attend today?
- *Time.* What three days do you prefer for library events (in order of preference)?
- *Type.* What other type of events would you like to see?
- *Awareness.* How did you hear about this event?

Aggregation

The "current data" and "new data" work together to bring about rich insights. In particular, libraries will be able to geolocate which populations desire what events, at what times. The combined data also has the potential to show how regional communities and neighborhoods learn about events. You can thus give individual communities the programs they want and

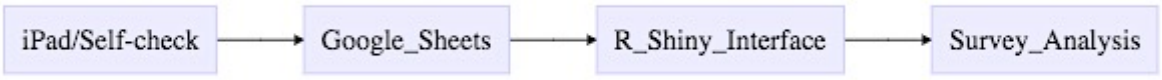
reinforce your marketing and outreach decisions. Depending on how you design your survey questions (broad or specific), patron age can be used as a helpful indicator for program planning.

Methodology

The following is a fictional walkthrough of how to apply the data model. I will present a scenario and unpack the mechanics necessary for analysis.

Scenario

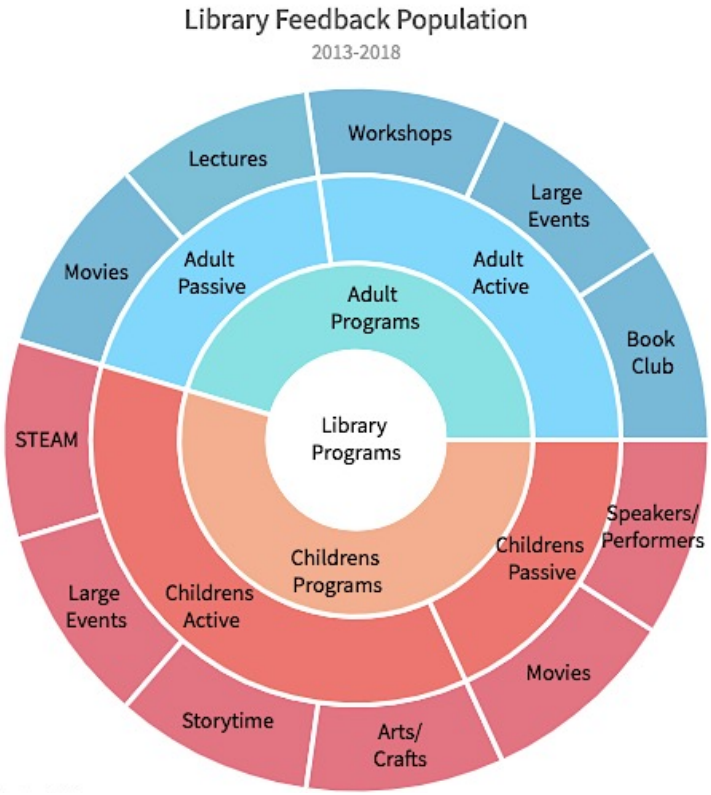
Library X is a public library that provides a variety of programs at no cost to the community. For the last five years, Library X has been collecting feedback information from the program participants in the form of an electronic survey. Surveys are made available on iPads during community outreach events and on self-check machines at the library. When participants fill out the survey, it is tied to their library card number in order to prevent duplication. The surveys are ingested into a Google Sheet by the iPads and self-check machines. The Google Sheet is actively tied to an R Shiny Application, automatically updating itself as new surveys are recorded. The R Shiny Application completes the data pipeline with an easy-to-use interface from which library staff can ask questions of the data.



In order to mimic "potential" survey results, I have generated a fictional database of 5,632 surveys in R Studio. The survey questions that produce this data are worded as in the

example model above. This database consists of 11 different programs, each of which contributes 512 surveys to the total population. This will generate results with a fairly equal distribution. For greater cross-tab analysis, each of the 11 programs has been classified as either an Adult or Children's program and as either Active or Passive. By Active/Passive, I mean whether or not the program was hands-on and involved participation.

This pie chart shows the sample population:



R Shiny Application

Each R Shiny Application will need different levels of construction in order to function as a relational database for your data structure. Each application, however, should consist of three components: input functions (user interface), output functions (server), and a command that

knits the two together. More information on how to design an R Shiny Application can be found here. The following are the three basic components written as code:

```
library(shiny)
#Input Functions
ui <- fluidPage()
#Output Functions
server <- function(input, output) {}
#Knit Components Together
shinyApp(ui = ui, server = server)
```

The code used for the R Shiny Application in this model will not be disclosed. Please contact the author for licensing.

Analysis

Input

The R Shiny Application is the interface that performs analysis. In order for the output to be meaningful, the relationships between your data elements must be hard-coded into the application. These relationships are derived from Boolean operators, namely AND and OR. These operators limit the range of data from which the application calculates. As a result of these relationships, the R Shiny Application provides a logical statement as its output. This logical statement is provided as text and also rendered visually.

For simplicity, this example demonstrates relationships built solely on AND operators. The image to the right is the input interface.

Text Output

The input interface allows you to ask a question from your data. You can create endless relationships from selecting elements from the dropdown; however, not all relationships are meaningful. You can also leave fields blank or select more than one option from each field.

The following data table is generated from choosing the option "All" from *Times/Days* and "Children's Programs" from *Adult/Children's.* All other fields are left blank:

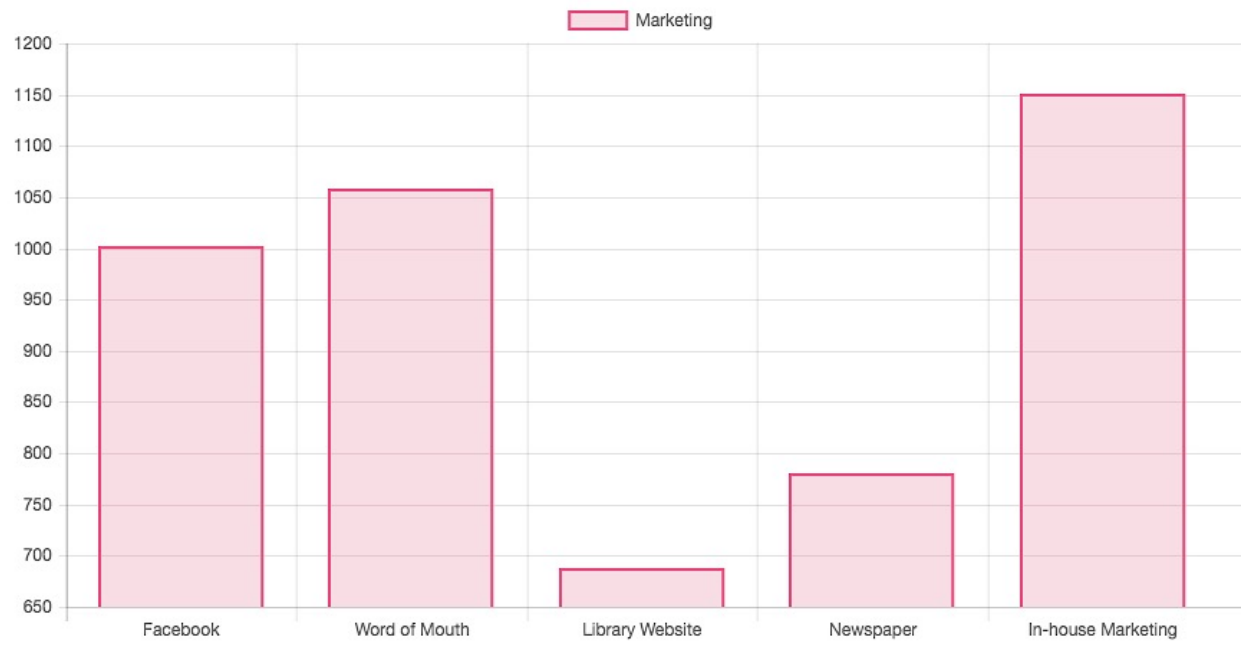
	Day/Time	Adult/Children's	Responses	Percentage
1	Monday Morning	Children's Programs	512 of 942	(54.3%)
2	Monday Afternoon	Children's Programs	508 of 940	(54.0%)
3	Monday Evening	Children's Programs	508 of 893	(56.8%)
4	Tuesday Morning	Children's Programs	513 of 930	(55.1%)
5	Tuesday Afternoon	Children's Programs	530 of 970	(54.6%)
6	Tuesday Evening	Children's Programs	502 of 944	(53.1%)
7	Wednesday Morning	Children's Programs	518 of 949	(54.5%)
8	Wednesday Afternoon	Children's Programs	489 of 878	(55.6%)
9	Wednesday Evening	Children's Programs	494 of 919	(53.7%)
10	Thursday Morning	Children's Programs	486 of 918	(52.9%)
11	Thursday Afternoon	Children's Programs	526 of 963	(54.6%)
12	Thursday Evening	Children's Programs	526 of 958	(54.9%)
13	Friday Morning	Children's Programs	528 of 982	(53.7%)
14	Friday Afternoon	Children's Programs	505 of 909	(55.5%)
15	Friday Evening	Children's Programs	568 of 995	(57.0%)
16	Saturday Morning	Children's Programs	491 of 914	(53.7%)
17	Saturday Afternoon	Children's Programs	490 of 921	(53.2%)
18	Sunday	Children's Programs	518 of 971	(53.3%)

The fictional dataframe has a uniform distribution for Times/Days. The data shows that there is no statistical significance attached to Times/Days, even when filtered to include only Children's Programs. Depending on your survey results, certain days will be preferred over others. Use this information to better plan programs.

Visual Output

Depending on how you design the elements of your R Shiny Application, you can render multiple output features. These output features can be tied to multiple or single input values. For instance, you can render two bar charts for side-by-side comparison. Always consider what question(s) you are asking of your data, and the best way to represent that question.

Below is an example of a bar chart generated from selecting "All" from *Marketing:*



"In-house Marketing" received the most responses, but "Word of Mouth" is not far behind. How do we make use of this information? Well, we need to combine more data elements together in order to construct a richer narrative. Add more data elements, and you'll be pleased at how the information becomes faceted.

GIS Mapping

This data model requires that survey-takers provide their library card number as a part of their feedback. Upon the axis of the library card number, patron addresses (or converted geospatial coordinates) can be mapped on top of survey results. You will be able to visualize what communities desire what programs, at what times. You can also use date of birth as an additional component to aid in constructing a narrative from your data.

Palladio is an application developed by researchers at Stanford. With Palladio, you can map your addresses (as coordinates) in conjunction with other data elements. In choosing what data elements to map, remember to consider what relationships are useful, make sense, and will answer your question(s). The section below details how to model your data for mapping.

In order to use GIS mapping in your analysis, your data pipeline will need to be organized in the following way:

Data Cleanup

The following information addresses just a few ways in which you may need to clean up your data.

Mapping terms

As survey questions and responses evolve, new data will need to be mapped onto the old data, where possible. This requires additional coding within R Studio to align results. If you want

to reduce granularity in your data, you can redefine data objects. For example, you can rename "Monday Morning," "Monday Afternoon," and "Monday Evening" simply as "Monday" in the following way:

```
dataframe [dataframe == "Monday Morning"] <- "Monday"
dataframe [dataframe == "Monday Afternoon"] <- "Monday"
dataframe [dataframe == "Monday Evening"] <- "Monday"
```

You can go from granular to broad definitions quite simply. The inverse is much more difficult to achieve.

Mapping language

Communities are often multilingual. In order to offer everyone equal opportunity, surveys should be administered in as many languages as possible. However, in doing so, results must be mapped onto a primary language. This is done in the same way as with the "Monday" example:

```
dataframe_spanish [dataframe_spanish == "Lunes por la mañana"] <- "Monday Morning"
dataframe_spanish [dataframe_spanish == "Lunes por la tarde"] <- "Monday Afternoon"
dataframe_spanish [dataframe_spanish == "Lunes por la noche"] <- "Monday Evening"
```

While you can map non-English terms onto English results, it may be more fruitful to do a separate analysis based on the language of survey-taker. But if you are trying to perform a holistic analysis, consistency is crucial. The non-English survey questions should appear in the

same order as the English survey questions so that the data export is mirrored in Google Sheets. If the columns line up, you can combine the dataframe in the following way:

```
cbind(dataframe, dataframe_spanish)
```

You can do this with as many dataframes as you would like, each separated by a comma in the parentheses.

GIS mapping

In order to achieve a useful dataframe for Palladio, you must first export your patron data. Clean up the data using R Studio or OpenRefine so that you are left with only the addresses and library card numbers. Upon the axis of the library card numbers, merge the patron data with the survey results. Next, geocode the addresses. This can be done through online generators, or in R Studio. You can find a tutorial here. Once you have your aggregated data, upload the dataframe to Palladio and adjust any facets as needed. Keep in mind that Palladio does not have a useful export feature.

Why not Excel?

As your corpus of data grows, Excel becomes less useful for analysis. When data-saturated, the application can slow down. Beyond this, Excel must be purchased.

The Excel environment is useful for pivot tables and visuals, yes. However, the syntax language is less intuitive than R and cell-dependent. R Studio has a much cleaner workspace and broader statistical application. Remember, Excel's primary purpose is to *store* data, while R Studio's purpose is to *analyze* it.

It is also bad practice to work out of your storage site. You can make copies of your Excel books, but this creates issues with file confusion, version control, and data loss.

Sustainability

This data model is dependent on a variety of software, hardware, skill, and financial access. Issues of obsolescence and hosting should always be a consideration for digital projects. Google Sheets is dependent on Google, whereas R Shiny is dependent on R, whereas Palladio is dependent on Stanford. These dependencies create a fragile structure.

R Shiny has certain limitations, namely that free hosting is supported at only 25 hours a month. Depending how often staff calls upon the data, this amount of time may be insufficient. However, R Shiny is merely a user interface; all of the same results can be generated through scripted queries. Even so, this requires a certain level of ability.

As time progresses, file formats change or are no longer supported by applications. Live ingestion can fail. Library staff will need to periodically confirm that survey results are being uploaded into the Google Sheet. Additionally, staff should make copies of the survey data once per month, storing the file separately as a redundant backup.

Staff will also need to update survey input options, so that the public-facing choices reflect current program offerings. This requires superficial code changes to the R Shiny Application.

In considering user experience, companies that provide self-check machines, such as Bibliotheca, may be unwilling to support survey ingestion as a feature of their product (although it is certainly possible). Self-check software, Google applications, and R Shiny also constantly

update. Their updates may one day no longer support the data pipeline set forth in this model. Staff time and export features should also be considered.

Security and Access

Patron confidentiality is a huge concern for libraries. If you're going to use patron data, make that use explicit in the library card application. Period.

Whenever data is moved around, there is the risk of unwanted exposure. This model suggests using multiple web-based applications for data analysis. In doing so, you put your data at risk. Create internal policies that will help mitigate risk and promote security. Additionally, take the time to review each software application's Terms of Use to see what the parent companies will do with your data. Terms of Use change all the time, so it is essential to review these updates.

Only certain staff members should perform data aggregation involving patron data. In limiting access, you heighten security. The public trusts libraries, and we want to keep it that way.

Conclusion

The purpose of this project is to present a data model that can be used to analyze community voices and feedback. Given that human rights constantly intersect with data violence, the onus is on library professionals to humanize data and serve people. In order to ensure informational integrity, we have to listen to community voices and shrewdly design data systems with them in mind. It is not enough, however, to hold roundtable discussions and collect

feedback. We must analyze these voices through a tangible application. The data will show us how to serve the public—if only we listen and apply the results.

Note: visit https://jeremy-zimmett.webnode.com/library-survey-model/ for a more interactive experience.

References

Harwood Institute: https://theharwoodinstitute.org/

OpenRefine: https://openrefine.org/

Palladio: https://hdlab.stanford.edu/palladio/

R Studio: https://rstudio.com/