Unveiling Insights: Exploring Social Survey Data through Visualization

I. Introduction

This Digital Humanities (DH) workshop aims to 1) critically reexamine existing data through creative tools, 2) consider barriers and enablers of practicing data cleaning, manipulation, and visualization, and 3) understand societal perspectives by exploring the general social survey dataset and the stories that are embedded in this data moves beyond just numbers. Data reuse drives can shed light on previous research, ask new research questions, and engage in the research data lifecycle (Harper, 2023). This workshop attempts to discuss responsible data reuse with its opportunities and challenges throughout the data cleaning, manipulation, and visualization to DH beginners. By reusing the general social survey dataset conducted with the collective opinion of communities that shape the narration of society, this workshop also opens a window to understanding the experiences and beliefs of individuals and provides researchers with tools to navigate the complex windings in society.

1. Scope

1) Audience

The objective of this workshop is to invite beginners, including scholars, librarians, students, and people from various disciplines, to learn how to reuse and contextualize the big dataset and explore various kinds of visualizations for analysis and presentation with no prior experience with particular technologies or digital humanities.

2) How to critically (re)using research data

Research data, defined as "entities used as evidence of phenomena for the purposes of research or scholarship" (Borgman, 2015: 162), holds a unique position in the Digital Humanities (DH), where it is considered not only as a technical matter but also as a theoretical, methodological, and social concern aimed at unlocking humanistic inquiry (Schöch, 2014: 2). In the interpretation of data, the significance of contextual considerations is underscored to navigate challenges associated with power differentials and misaligned collection incentives (D'Ignazio & Klein, 2020). With its concerns, DH tries to uncover the structured arrangement of values which may conceal the nonscalable diversity of the data (Rawson & Muñoz, 2019). To navigate this challenge, a crucial step involves contextual considerations, encompassing an understanding of the data's

origin and environment, along with intentional framing of context in data communication, recognizing that numbers should not autonomously convey information in charts or spreadsheets. Additionally, analyzing social power dynamics within the data set is essential for the comprehensive integration of context into data analysis including exploring power imbalances, conflicts of interest, and suppressed knowledge.

The challenges of critical interpretation and contextualization become pronounced in data reuse[1] scenarios, where researchers encounter barriers such as uncurated, dirty, incomplete, or uncontextualized data, along with ethical concerns and privacy risks (Faniel et al., 2016: 1162; Frank et al., 2018). Despite these challenges, data reuse offers benefits, including opportunities for collaboration, transparent reproducibility, and self-taught and pedagogical possibilities (Sielemann, Hafner & Pucker, 2020). However, the process involves costs related to accessing, managing, licensing, sharing, and preserving data, making proper guidance crucial for DH practitioners seeking to improve barriers and maximize benefits. Contextualization remains key in the critical use and reuse of data, even though tracking the previous contexts of data may be challenging. To address this, the practice of data reuse can benefit from documenting the entire process, providing transparent contexts, and advocating for data co-authorship participation throughout the data lifecycle (CODATA, 2021). This workshop helps beginners to learn how to critically contextualize and reuse data.

3) Workshop as a pedagogy and community

Digital Humanities is “both humanities done digitally and the digital as scrutinized humanistically.” (Eichmann-Kalwara et al., 2018). These mirror-like aspects of DH call for the inclusion of diverse representatives and reinforcement of computational data analysis for all. However, there are two critical aspects that require attention. The first one is how to open the sphere in the context of diversity, access, and inclusion involving various people of different backgrounds and divergent levels of digital literacy. Although disciplines from universities provide computational courses, there is no standardized “curriculum” for all at the moment. There are breaches in digital literacy, accessibility, and different kinds of backgrounds that hinder the diversification of collaborations. In this regard, workshops can be an alternative and effective way to fill the gaps between academia and society, and professional and amateur by inviting people from various backgrounds and different levels of understanding asynchronously (Estill & Guiliano, 2023).

The second concern is how to avoid shallow software training and go beyond a critical and self-reflective understanding of data analysis in the short-term workshop (Russell & Hensley, 2017; LeBlanc & Walsh, 2019). In this regard, this workshop contains discussion between participants to evolve their critical thinking and communicate self-reflection including

[1] Data reuse is using data by people other than the original source (Faniel et al., 2016: 1162). In this workshop, data reuse is discussed in the context of knowledge production.

difficulties and anxieties throughout the process. This is because DH workshops can create a community sharing questions and interests of the humanities and blur the hierarchy and the tension between the educator and the educated by bringing a sense of transparency and safety to communicate with each other (Estill & Guiliano, 2023).

Conclusions drawn are majorly based on Tableau Visualizations, usage of which we shall not be demonstrating in our workshop, because of the lack of open-source software and prerequisite of data analytics knowledge along with SQL syntax. We will explore the usage of all other tools - OpenRefine, Flourish, Venngage, RawGraphs, and Palladio and demonstrate how to interpret a visualization through Tableau-generated visualization.

2. Goals

Learning Outcomes

The primary objective is to practice critical reuse and contextualization of data throughout the process of data extraction, refinement, cleaning, and visualization. Moreover, various DH analysis tools are introduced and compared with strengths and limitations. Participants will have the opportunity to apply the techniques learned to the real-world social survey dataset, ensuring practical application and a deeper understanding of the methodologies. By the end of the workshop, participants will not only have acquired a diverse set of skills in data refinement, attribute extraction, and visualization but will also begin to critically reuse and contextualize the data with the proper tools for specific analytical tasks. Tools are OpenRefine, Google Sheets, Flourish, Venngage, Tableau, Palladio, and RawGraphs to clean, explore, and understand the social survey data from 1972 to 2015.

The overarching goal is to empower participants to learn how to critically manage and contextualize data, and uncover meaningful stories within social survey data contributing to informed decision-making and a deeper understanding of societal dynamics by intervening data analysis and aesthetic critical speculation through visualization. Although this workshop mainly focuses on practical exercises of actual social data with various tools for data management and visual analysis, participants' critical self-reflection and sharing of authentic emotions and experiences will be discussed during the exercise to make the visualization a dynamic and ongoing process and promotes interdisciplinary collaboration discussed and reflected in problematizing the data and our own perspectives.

Session Outline

The general information and context of the dataset are introduced. Data filtering is demonstrated using OpenRefine and Google Sheets, and data visualization analysis using Flourish, Venngage, RawGraphs, Palladio, and Tableau is

demonstrated. Following the demonstration, discussion on the tools and critical reflections are conducted. Throughout the workshop, internet access, an email address for tool sign-ups, and downloading OpenRefine and Tableau are required.

II. Dataset and Tools

What is the dataset being studied?

The General Social Survey (GSS) Data set is 2.07 GB in size and exploration of it is important to the understanding of Social Research providing comprehensive and invaluable resources for understanding the dynamics of American society. This data set was initiated in 1972 by the National Opinion Research Centre (NORC) at the University of Chicago (*GSS General Social Survey | NORC*, n.d.). Ever since, annual data has been collected to capture evolving behaviors, attitudes, beliefs, and opinions of the American population. The data set is extensive and encompasses a wide range of topics which include social attitudes, perceptions, political views, familial ties, and demographic characteristics. By studying this data researchers, policymakers and sociologists can understand the multifaceted nature of the society and fathom its strengths and transformations. This data has been collected through meticulously designed survey instruments by asking a large expanse of questions trying to cover multiple aspects of living and has been consistently collected over 42 years. Research at AWS has produced statistical inference similar to ours (*Statistical Inference with the GSS Data*, n.d.), but while we move towards a more holistic approach, these conduct singular correlations which we think are limitations.

gss.csv (2.07 GB)

Detail Compact Column 10 of 10000 columns

GSS YEAR FOR THIS...	RESPONDNT ID NU...	LABOR FORCE STAT...	LABOR FORCE STAT...	NUMBER OF HOURS ...	NUMBER OF HOURS ...	NUM
2006.0 8% 1994.0 5% Other (52098) 87%	4511 unique values	1.0 49% 7.0 16% Other (20513) 34%	WORKING FULLTI... 49% KEEPING HOUSE 16% Other (20513) 34%	-1.0 42% 40.0 20% Other (22777) 38%	[null] 58% IAP 42% Other (217) 0%	-1.0 40.0 Other (1
year	id	wrkstat		hrs1		hrs2
1972.0	1.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	2.0	5.0	RETIRED	-1.0	IAP	-1.0
1972.0	3.0	2.0	WORKING PARTTIME	-1.0	IAP	-1.0
1972.0	4.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	5.0	7.0	KEEPING HOUSE	-1.0	IAP	-1.0
1972.0	6.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	7.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	8.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	9.0	2.0	WORKING PARTTIME	-1.0	IAP	-1.0
1972.0	10.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	11.0	7.0	KEEPING HOUSE	-1.0	IAP	-1.0
1972.0	12.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	13.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0
1972.0	14.0	1.0	WORKING FULLTIME	-1.0	IAP	-1.0

Manual Cleaning of the Dataset

There are almost 10000 attributes in this data set which is why we have manually cleaned it focusing on two specific aspects: Living conditions and Opinions. The output we generated has two manually cleaned datasets – first, having information such as year, identification number, age, gender, race, work status, employment details, parental employment details, self and parental educational level, geographic location, living arrangements, family income, immigration status and

information about people in the household to understand the living conditions of people; second having attributes such as power associated with the individual, their happiness, and people's perceptions of trust, manner, and weightage laid on success, honesty, sense of judgment, self-control, orientation, obedience, responsibility and others to understand the opinions of the people. The attribute coding done during data collection has also been modified by us so that we could better contextualize the attributes. Our coding strategy was to reduce jargon and incorporate the entire meaning of the attribute so that we could well understand the meaning of the attribute.

III. Demonstration

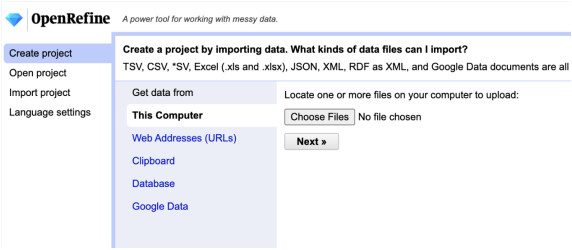
1. Research Objectives

For the first half of the workshop, the relationship between the evaluation of happiness, close relationships, and success will be compared between 1973 and 2014. To provide the context of this objective, this survey data is hard to contextualize with obscure information with its collection and backgrounds. Across cultures and time, the prevailing definition of happiness often centered on good luck and favorable external conditions, while American English has shifted to prioritize definitions emphasizing favorable internal feeling states (Oishi, Graham, Kesebir, & Galinha, 2013). Although the ethnographical context of the survey participants are unknown, happiness could be considered as the positive evaluation of one's life in most cultures and nations. Moreover, among its scales from 'very happy', 'happy', 'pretty happy', 'not too happy', 'IAP', and 'NA', 'very happy' is the most concrete with its evaluation by participants. We wanted to see how this variable changed during different time periods. We chose the 1973 and 2014 dataset because they were the oldest and newest data available with the dataset. We also chose personal relationships including marriage, relationship, family life, friendship, and success to explore how those close bonds of a person are related with self-evaluation of one's happiness and success.

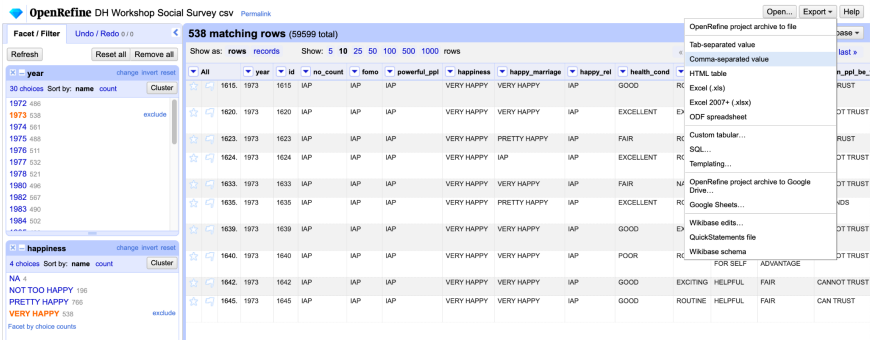
2. Data Filtering through Open Refine

The next step is to ensure that the dataset is in prime condition for analysis. While filtering out noise such as missing values and standardizing the data are to make them available for visualization according to our research objectives, it is important to clearly state the logic of the processes to contextualize the data reuse practiced here. OpenRefine is a free, powerful, and intuitive software for cleaning and transforming messy data into standardized formats. OpenRefine supports TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents.

To start with, uploading a CSV file and creating a project with the regarding button on the upper right.



Using facets, you can filter the data based on specific columns or values. Since most responses in 1972 are IAP, unanswered, so we will look at responses made in 1973 using text facet and sorting 1973 data by clicking the 'include' button. Similarly, you can sort 'VERY HAPPY' data in the happiness column using text facet and come up with 538 rows of data. Next, you can export sorted data as a CSV file, upload in Google drive.



3. Data Consolidation through Google Sheets

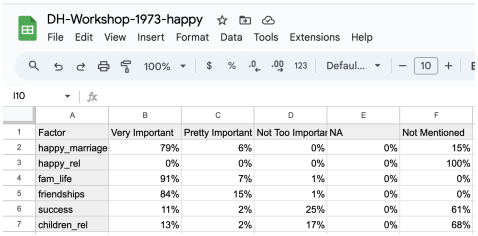
Google Sheets' connectivity with other Google apps and compatibility with various formats make it a strong tool for Digital Humanities. Sign-up with a gmail account is required for using Google Sheets which is a good way of data cleaning and manipulation with its intuitive design and functionality.

Opening the previously sorted 1973-very happy file with Google Sheets. Among columns, we focus on the data related to respondents' perspectives on intimate relationships and success with happiness. We delete every column except 'happy_marriage' (happy marriage), 'happy_rel' (happy relationship), 'fam_life' (family life), 'friendships', 'success', and 'children_rel' (relationship with children).

Before visualizing the data, we need to manipulate and transpose this data. Some tools or to be specific, some templates require a certain arrangement of data with its rows and columns to run. When you learn how to transpose (switching rows and columns in datasets for data analysis) the data, you can easily merge or manipulate the data for use. A survey scale, comprising a set of numeric or verbal answer options, encompassing a spectrum of opinions on a given topic, is an integral component of closed-ended questions, which present respondents with predetermined answer choices.

First, different scales need to be categorized to standardize the data to be tool-readable for visualization. When you manipulate scales, critical deliberation needs to be done with the decision making. In this workshop, we will focus on the degrees of evaluation to deploy scales between the highest to the lowest. Also, not answered or mentioned responses will be put into other categories. Followingly, scales are 'very important', 'pretty important', 'not too important', 'NA' (not answered), and 'IAP' (not mentioned). Now we need to count the responses according to the scales. You can count them with the filter column by column, but we will use the 'countif' function here for efficiency. When it comes to handling a huge amount of data, it is easier if you know some functions. You can google them or use Chat GPT to learn about them. To count the number of cells that meet a criterion of "very happy", you can type "=countif(A2:A529, "very happy")". You can change the range (it's from A2 to A529 here) and the criterion "very happy" according to your data. When we write the function and set the same range, you can type $ mark in front of the alphabet and number to avoid the auto change of the range. However, 'countif' function doesn't work for multiple criteria with one scale. Still, you can use the 'sumproduct' function such as "=SUMPRODUCT((C2:C539 = "very great deal")+(C2:C539 = "great deal"))". This is counting "very great deal" and "great deal" criteria from the range C2 to C539.

After counting the data with standardized scales, you can transpose the dataset using the 'transpose' function. In a new sheet, copy and paste values only to avoid missing values out of its previous arrangement with the evaluated function, and type "=transpose(A1:G6)" to transpose the column into row, row into column. To go further, now you can calculate the percentages of each cell by counting the sum ("=sum(B2:F2)") and dividing each cell with it ("=B2/sum"). To make the data more intuitive and comparable for understanding, you can decrease decimal points and add the percentage symbol. The 'decrease decimal places' button helps you to remove decimal places. You can use the "% (format as percent)" button to make it more percentage-like. Now we are ready for the visualization. You can download the dataset as a CSV file by clicking 'file' and 'download' buttons. Same process can be done with the dataset surveyed in 2014.



	A	B	C	D	E	F
1	Factor	Very Important	Pretty Important	Not Too Importar	NA	Not Mentioned
2	happy_marriage	79%	6%	0%	0%	15%
3	happy_rel	0%	0%	0%	0%	100%
4	fam_life	91%	7%	1%	0%	0%
5	friendships	84%	15%	1%	0%	0%
6	success	11%	2%	25%	0%	61%
7	children_rel	13%	2%	17%	0%	68%

4. Data Visualization

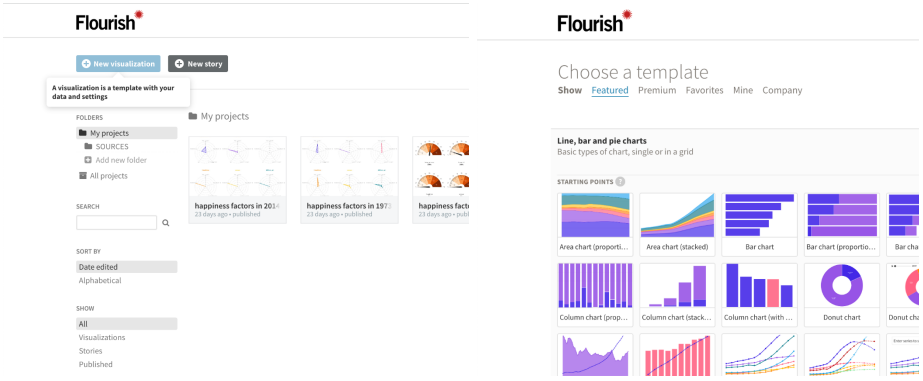
Visualization is an aesthetic method to elicit critical speculation and re-interpretation, and a mediating and collaborating way to provoke different research perspectives to discuss further thus combining and intersecting different viewpoints and

disciplines (Hinrichs et al., 2019). One of the goals of our workshop was to be able to help understand how one might begin with the visualization of DH dataset. This has been done through an introduction to several tools that may be used for this purpose, their benefits and a clear demonstration of the tool might be used. Due to the time limitations of the workshop, it was not possible to conduct a complete walkthrough of the tool, but the basics have been demonstrated. Further as aforementioned, since interpreting visualizations was also an aim of the workshop, we have done so through Tableau generated visualizations.

1) Flourish

Flourish is a web-based tool for Data chart visualization and map storytelling which provides various interactive templates and tutorials with unlimited free public views. It is partially premium but mostly free templates and a great option to explore and try visualization for your own purpose. It also has collaboration functionality to share the ongoing projects. Since this is a web-based tool, you don't need to download any software but log-in with your email address.

In the live demonstration on Flourish, radar with filter and gauge charts are visualized with the 1973 and 2014 dataset we cleaned previously. In Flourish, you can click 'New visualization' to choose a template and upload the data with Excel, CSV, TSC, JSON, and GeoJSON formats. Every template has its own example so we can set a starting point to clean our data for visualization. In your profile page, you can see and continue to work on your previous projects.

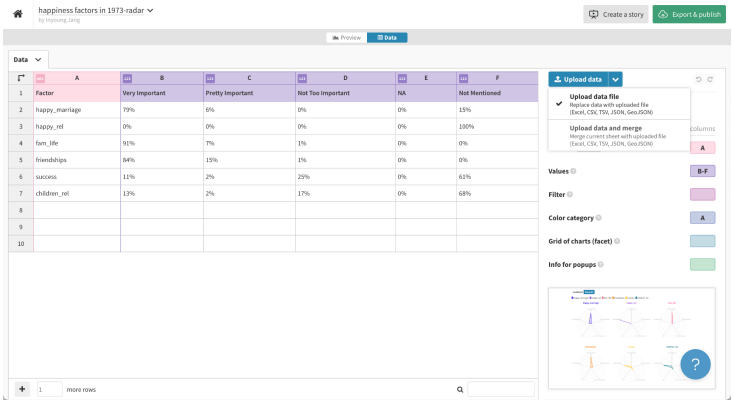


Radar with filter charts

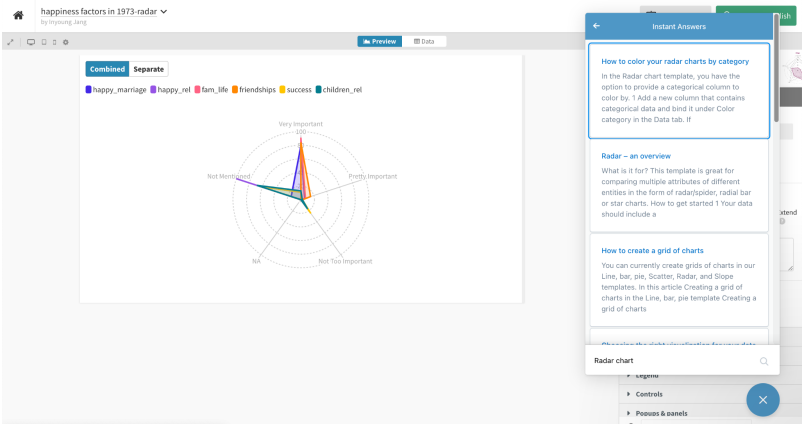
First, we will make Radar charts with a template 'Radar with Filter.' Radar charts, also known as spider charts, help us to compare entities across multiple metrics. They are widely used for visualizing the attributes of people, places and other

entities. Using Flourish, we can make radar charts with filters which provide interactive and comparative understanding of multivariate data with different categories.

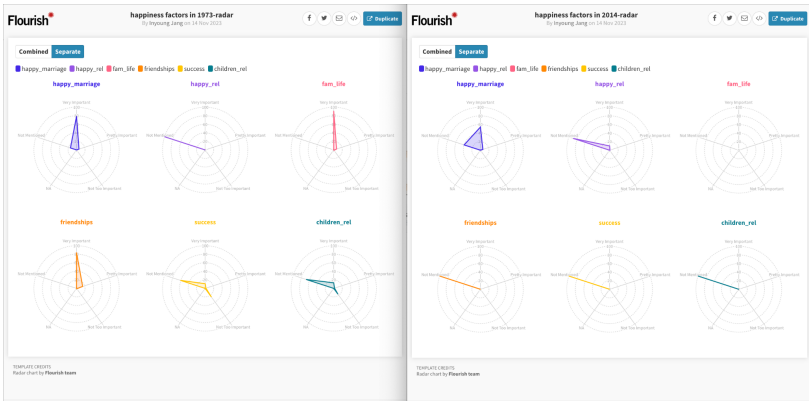
You can upload the CSV file we made with OpenRefine and Google Sheets by clicking ‘Upload data’, and check the data tap, set name and value. You can edit the data in the rows from the ‘Data’ tap.



Now, you can click the ‘Preview’ tap to customize the charts on the right. This template offers both combined and separate interactive labels to explore the data. You can export and publish the visualization with the ‘Export & Publish’ button. For the free version, a script or iframe embed code is available. You can also publish the chart with the online link or download the chart as an image. Also, along the process, whenever you come up with problems, you can ask the chatbot with the “?” button on the right bottom of the website.

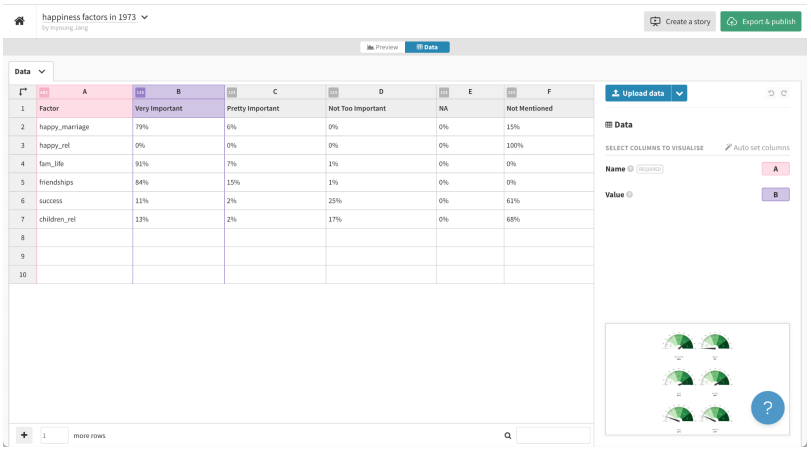


By visualizing two dataset from 1973 and 2014, you can compare the difference and rethink about the factors and implications of the differences.

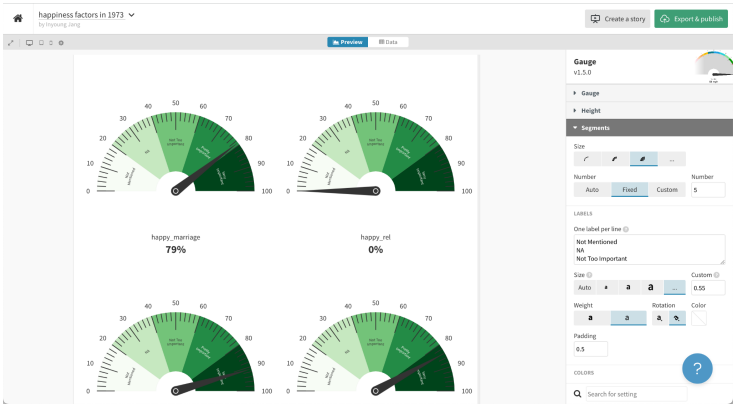


Gauge charts

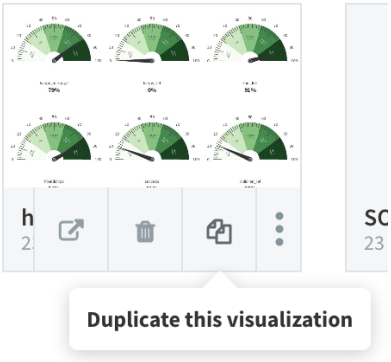
Second visualization with Flourish are Gauge charts, also known as speedometers or dials, often used in data journalism to highlight a key metric. This is a strong way of communicating a certain range of values whether something is within, below, or above a range. From the homepage, you can create a new visualization and choose the template ‘gauge-simple.’ To upload the data, you can click the data tap, and upload the dataset we made.



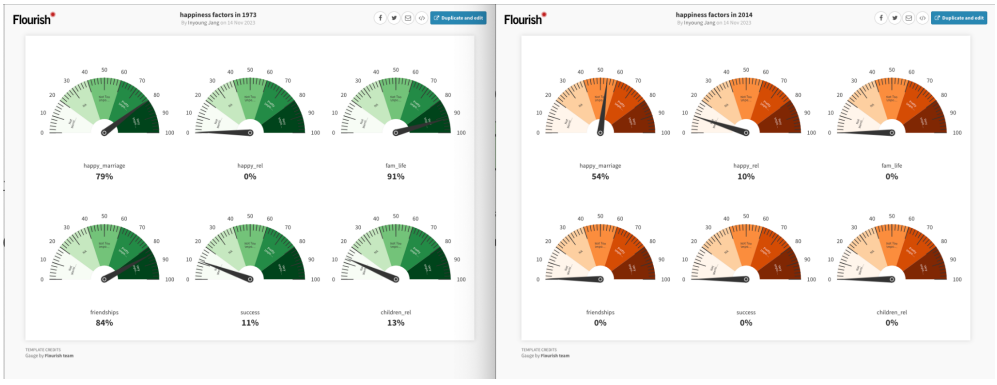
After setting column names and the value column, you can custom segments, add labels and colors for your own visualization strategies in Preview tap.



Similarly, after customizing charts, you can export them with the 'Export & Publish' button. For the 2014 dataset visualization, you can also copy and paste the chart from your profile page for using the same customized design.



Now, we have two different visualizations of 1973 and 2014 happiness factors to compare. You can also compare this set of visualizations with the different template visualizations. What specific analysis or visualization is possible here? What are the limitations? More questions can emerge.

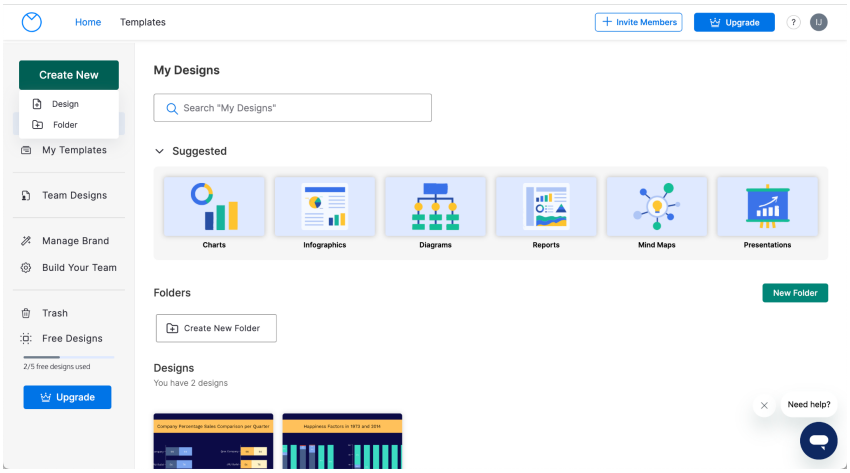


2) Venngage

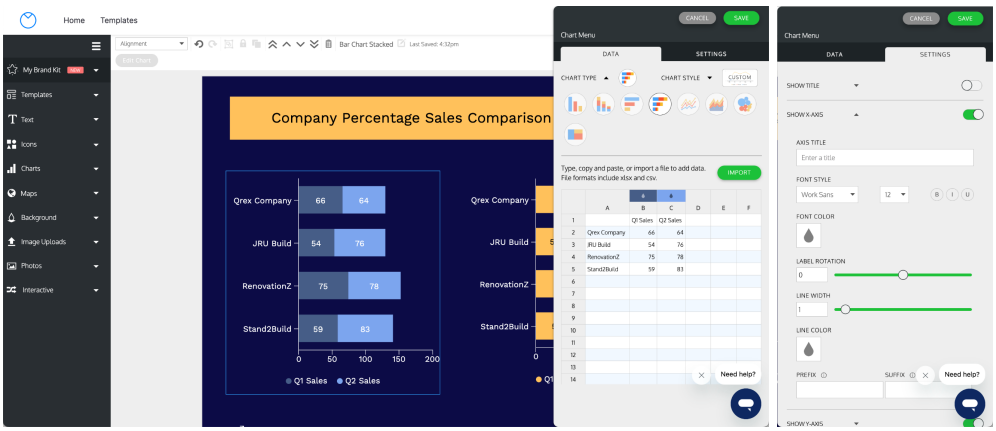
Venngage is another web-based visualization tool ideal for beginners. It has a simple interface with various chart templates to design and build your own interactive visualization even with no previous experience. This tool is especially concentrated on survey data visualization so you can easily build engaging reports with Venngage. Images, icons, colors, fonts, and data are customizable. Although some are premium templates, you can filter free templates to use. You also need to sign up with your email address.

You can start by clicking the 'Create New' and 'design' button for a new project.

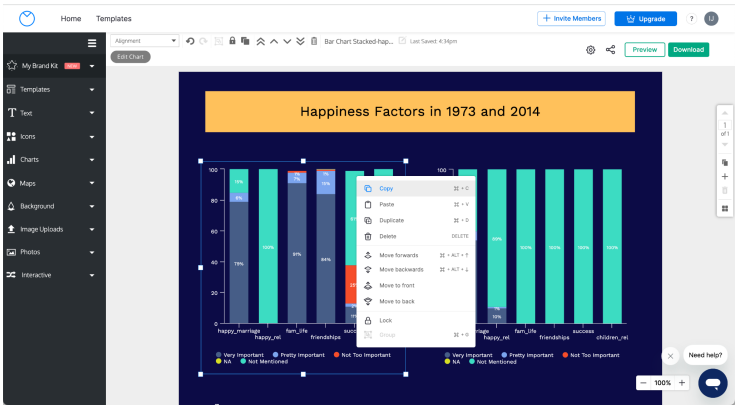
We will use the 'Bar Chart Stacked' template. You can find it in the 'Charts' menu, or search it from the search bar.



When you click the template, the example data and visualization are shown. You can click any part of the visualization or 'Edit Chart' button on the left top corner to start editing with the side chart menu. The Chart Menu has 'data' tap for importing, managing the data, and selecting chart type and style, and 'settings' tap for visual customization. You can type, copy and paste, or import a file to add data. XLSX and CSV are available formats. Similar to Flourish, there is a chatting button on the right bottom for asking support from chatbot.



You can choose different colors and sizes for the font as you want. But one thing suggested here is to make the X-axis to be from 0 to 100. Since our data is calculated as 100 percent, the X-axis of 100 can more precisely visualize the data. Although the value stays the same, if X-axis is 200, it will seem to be half of it and people might mislead it as 50%. When you practice visualization, not only data themselves, but also the audience and their reception need to be accounted for. You can import and customize the chart for the 1973 dataset first, and copy and paste the chart with a few clicks for comparative visualization of the 2014 dataset.



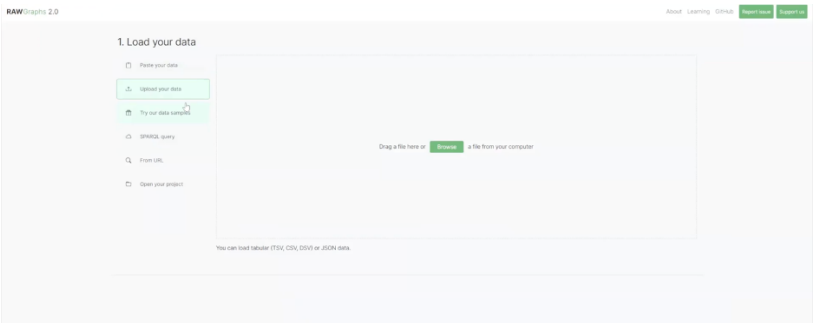
For the free account, it is only available to publish the chart and share the link with the ‘download’ button. However, you can also save the image by screenshot or print out. Similar to the Flourish charts, this visualization also provides interactive labels to actively interpret and compare the data. You may also duplicate, edit, and review your visualizations in your profile.

The comparative charts show how personal relationships have changed over the years and that success itself has not been a crucial factor for happiness.

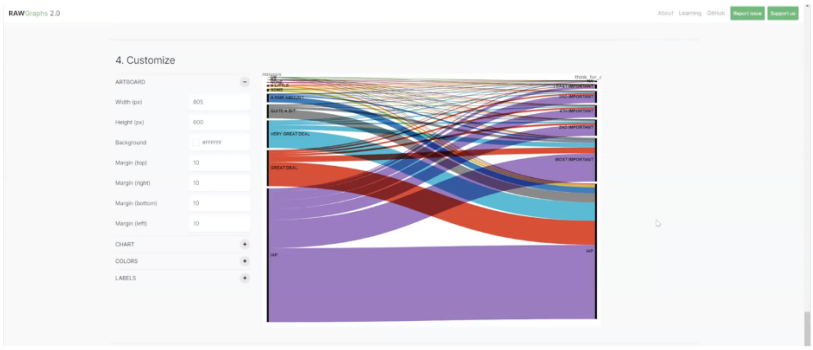
3) RawGraphs

One of RawGraphs' strengths is its ability to unveil hidden patterns and outliers through unconventional visualizations. Some of the advantages of Rawgraphs are - One of the primary strengths of RawGraphs is its flexibility in creating custom and unique visualizations. Users have the freedom to experiment with various chart types, layouts, and designs, enabling the expression of creativity in data representation; RawGraphs supports a wide range of chart types, including scatter plots, bar charts, network diagrams, and more. This versatility allows users to choose the visualization that best suits the characteristics of their data, fostering a more nuanced understanding; Users have extensive customization options, allowing them to tailor visualizations to their specific needs. From color schemes to axis labels, RawGraphs provides granular control over the appearance of visual elements, ensuring the output aligns with the user's preferences; RawGraphs is an open-source tool, making it freely available to users. This accessibility fosters a collaborative environment and encourages the sharing of knowledge and techniques within the data visualization community.

A live demonstration of the tool was conducted during the workshop, noted below: First, you will load the dataset into the web application. Note, that this needs to be a csv, tsv, dsv or json file.



Once this file has been loaded, you will be presented with the option to create a visualization. We created an Alluvial Graph to see the correlation between friendship, and the need to think for oneself. We first dragged and dropped the friendship attribute from this dimension column into the steps section and did the same for the think_for_oneself attribute.

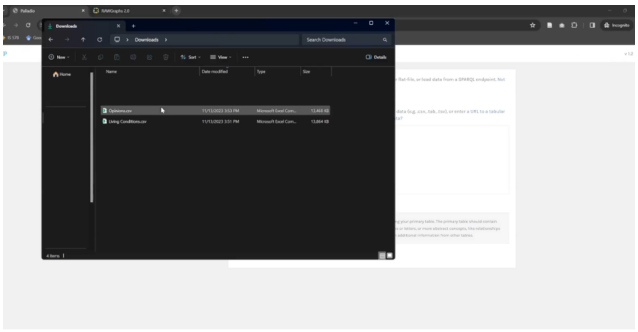


Through this graph, one interesting aspect of this visualization is that while many chose to not answer how important friendship is to them, almost half of them chose to think for themselves. This might be an interesting observation for studying egoism in individuals. There are several other graphs you could choose to explore based on your dataset and goal. We strongly encourage participants to do that. Keep in mind that creativity in data visualization is not just about aesthetics; it's about enhancing our understanding of the data and communicating complex ideas in a more engaging manner.

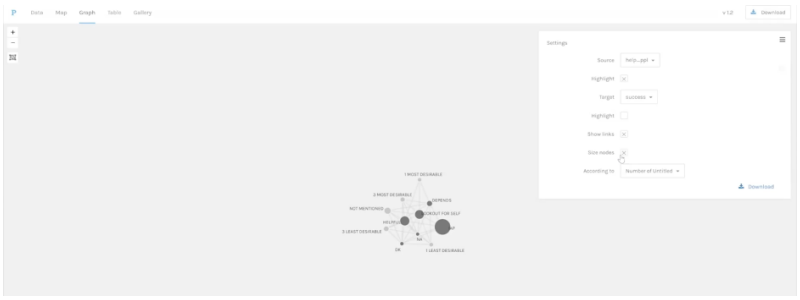
4) Palladio

Social interactions, affiliations, and dependencies are often hidden beneath the surface. Palladio allows researchers to visualize these connections, offering a unique perspective on the relational aspects of our dataset, through network generation. Some of the advantages of Palladio include: The offerance of an interactive environment for exploration. Users can interact with maps, networks, and other visualizations, gaining a deeper understanding of the data. This interactivity enhances the user experience and facilitates a more dynamic analysis; Palladio incorporates temporal elements into its analyses, allowing users to explore changes and patterns over time. This temporal dimension is crucial in social survey data, where responses may evolve, or shift based on societal trends and events; Palladio allows users to export visualizations for further analysis or to share findings with others. This feature enhances collaboration and facilitates the integration of Palladio's results into broader research or reporting efforts; It is built on an open-source framework, fostering a community of users who can contribute to its development.

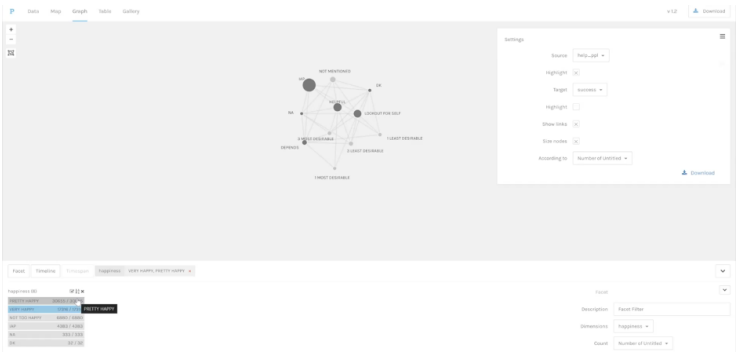
Live demonstration of Palladio began with loading our attributes.



After that, the source is set to help_ppl and the target is set to success. Due to the nature of our dataset, the network doesn’t provide us with a lot of insight as it would do to a geographical dataset.



In this visualization we now see how many people responded to two of the responses together. This means that, for example, that most people thought that they should be helpful, but did not mention whether success was the most important to them. We can also add certain facets to the dataset which would act as a filter. Adding happiness as the facet filter, we choose to look at ‘pretty happy’ people. The change we now observe is that the number of people who think they must be helpful and those who think they must look out for themselves have become the same for those who are pretty happy.



In conclusion, Palladio shows a network dimension of our social survey dataset. However, this might not be the best observed tool for the analysis of this dataset.

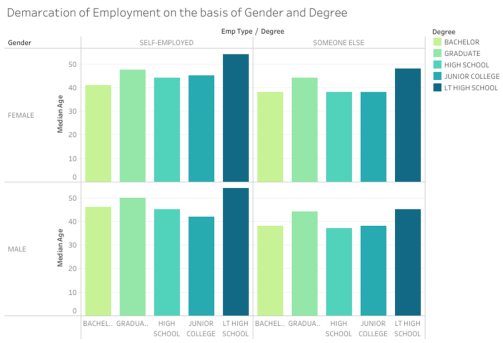
5) Tableau

Tableau, a popular data visualization tool, allows us to create compelling visualizations with ease. From basic charts to complex dashboards, Tableau empowers us to tell a visual story about our social survey data. Advantages of using Tableau include: Advanced analytics and calculations, enabling users to perform complex calculations within the tool. From creating custom calculations to conducting statistical analyses, Tableau provides a robust environment for in-depth data exploration; It allows users to connect to real-time data sources, ensuring that visualizations and dashboards can be updated as data changes. This real-time connectivity is crucial for businesses that require up-to-the-minute insights; Tableau is highly scalable, making it suitable for both small teams and large enterprises. It can handle large datasets and complex analyses, ensuring that it can grow alongside the needs of the organization.

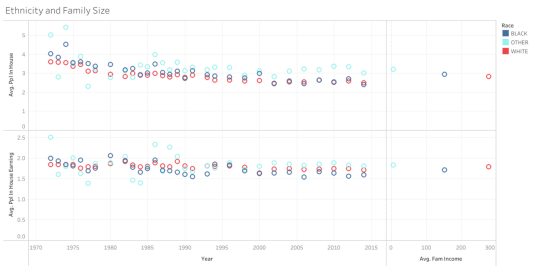
When interpreting a visualization, there are two routes a person can take. First, they can either begin looking for anomalies, or they can begin searching for similarities. This usually depends on the hypothesis generated by the researcher, and they can choose to refute a negative hypothesis or support a positive one. A good data visualization can be generated based on whether it supports or disproves a hypothesis and provides insight into correlations within the data. A hypothesis has been created only here, because we want to produce a clear distinction between the exploration and interpretation section of the workshop. We are not aiming at evaluating every tool for the same hypothesis; instead we want to showcase how each tool can be leveraged based on its own strengths on a single data source.

Let us suppose that for the Living Conditions dataset, we hypothesize that there is a distinction in employment types based on gender and degree, and we clump together the attributes to create a visualization. This visualization shows that employment trends do not differ much based on gender and there are minor fluctuations in self-employment based on whether males and females have been to highschool and junior college.

This is nothing major that can be flagged; hence, this hypothesis cannot be supported by the data.

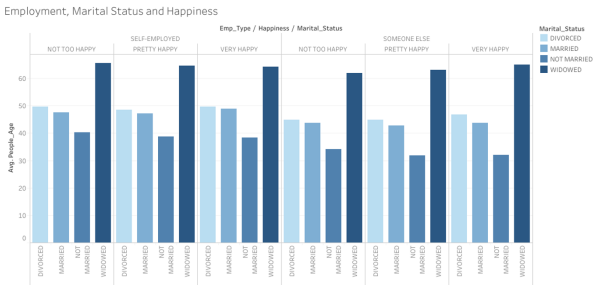


A second approach to visualization is to find the correlation between multiple attributes; this essentially does not require a hypothesis. Looking at the second visualization, the number of people in a particular household has been mapped against the ethnicity of individuals in a temporal analysis. This helps more so, visualize the data itself, rather than forming multiple conclusions. Therefore, it depends on what you wish to achieve and what the purpose of your visualization is, that you can decide which approach to take.



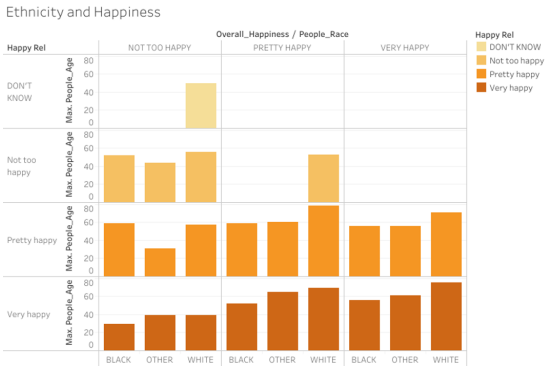
The third visualization we have here is an amalgamation of the two methods we described before. While it does not initially have a hypothesis, its goal is exploration as well as insight development. This approach is usually taken when one does not have a clear idea of what the data could present, nor do they wish to support a pre-notion.

In this visualization example, we can straightaway observe that widowed individuals are older to the other population that has been surveyed, and responses are well distributed. Yet, there is no particular insight that can be drawn from this visualization that is completely supported by data.



However, the next visualization uses the same underlying principle, but has a much stronger result. You can observe that white individuals are more likely to be confused about relative happiness. The x axis measures overall happiness while the y axis shows relative satisfaction.

While white individuals seem pretty happy overall, they are not too satisfied with their lives.



We understand that there are biases inherent to analyzing any humanistic dataset, and quantitative analysis of numeric and text responses might not be the most accurate. However, this is the best method to conduct a temporal analysis of a large representative sample, hence we have chosen to analyze this data. These biases include errors in self-reporting, exclusion of key factors and assumptions that participants and researchers make while answering surveys and analyzing visualizations. These biases produce skews in the visualizations produced, which subsequently end in erroneous conclusions.

4. Discussion

1) Comparison between the tools and reasons for choosing them

OpenRefine was chosen for its prowess in data cleaning and transformation. When dealing with the intricacies of social survey data, where inconsistencies and discrepancies are commonplace, OpenRefine emerged as the ideal choice. Its facet feature, allowing for nuanced data exploration and filtering, paved the way for a clean and reliable dataset. Google Sheets provided excellent functionality for manipulating the data. We were going through counting, transposing, and calculating with Google Sheets. There are a lot of functions to explore further. One of the most brilliant strengths of Google Sheets is its collaborative workflow. You can share the project with your teammates to work with. You can also see the history of every edited version to cooperate effectively. Flourish is a fantastic tool especially for beginners. There are a lot of free templates and import/export options for visualization. You can explore all the different types of charts and graphs with sample data and try your own project immediately. Whenever you get lost, you can just ask the chatbot for inquiries throughout the whole process–from choosing a template to designing one. Importing datasets and exporting the results are also very handy. Venngage, although it provides limited options for free templates and exporting options, it is easier to search for a template with its search tool and filter. Also, Venngage is mentally more accessible with an intuitive interface and template design. It seems to be more useful for daily uses when you write a report paper or design a presentation page or website. Vennage also provides chatbot service which shows more interactive and intimate reactions. The decision to integrate Tableau was rooted in its user-friendly interface and expansive visualization capabilities. It transformed our refined data into dynamic and interactive dashboards, making complex information accessible. We did not demonstrate how to use the software, but we hope that the visualizations generated from it, were helpful in understanding the dataset and provided a clear insight into the multiple ways one can approach data visualization. RawGraphs' strength lies in the ability to create visually striking and unconventional representations, opening avenues for deeper insights and sparking new questions. For this dataset however, this software seemed restrictive, as opposed to Tableau which paved way for advanced data analytics; RawGraphs works much better when the proportion of numeric data is larger than textual data, and if one wants to create plots such as stacked bar plots, they find the number of attributes they can visualize together are limited. Social survey data often carries inherent spatial and network components, and Palladio's strengths lie in transforming numerical data into interactive maps and network visualizations. It provided the lens through which we could understand the contextual factors influencing survey responses. For this dataset, since there was no geographical data to map, Palladio was limited in the insights it could provide. The use of these tools is towards the goal of helping participants understand

how the exploration of a tool brings out its nuances and abilities, and based on the purpose of research, tools can be well selected to solve that purpose.

2) Critical re-speculation on the data reuse and tools, and feelings

This workshop provides different kinds of tools for reusing data and practice visualizations which may cause overwhelming experience for the beginners or overly focusing on tools more than critical understanding of data analysis. However, in order to broaden beginner's options and open up a sphere for multiple practices and experiences, this workshop has chosen to bring various tools and exercises for participants. Also, while practicing various processes, participants can acquire examples of critical contextualization with given situations while the logics and related issues provided. To avoid the shallow practices of tools, questions on limitations and strengths of different visualizations have been asked throughout the process. Also, personal feelings and experiences on the process can be communicated before, during, and after the demonstration. Once again, this workshop is to unlock the dynamic dimensions of data analysis and visualization that are continuous and always in the process, and the workshop itself is required to be renewed with the interactions made throughout the sessions.

IV. Conclusion

This workshop might be challenging but for that very reason, this workshop gives vivid examples for data reuse with processes of data collection, cleaning, visualization, and contextualization. Throughout the workshop, every step of data manipulation and visualization attempts to be transparent with its context, and provide critical parts to think about in conducting critical data analysis and visualization. The orchestration of these tools was not accidental; it was a strategic ensemble, each tool playing its unique role in creating a comprehensive analysis. Understanding the strengths of each tool allowed us to navigate the diverse landscape of social survey data, ensuring that our exploration was thorough and insightful. In this workshop, we witnessed the power of data refinement, attribute extraction, and visualization using OpenRefine, Google Sheets, Flourish, Venngage, Tableau, RawGraphs, and Palladio. Armed with these tools, participants can now confidently navigate the intricate landscape of social survey data, uncovering stories that were once hidden in the numbers and text.

However, due to the time limitations, the profound discussions on critical speculation of data analysis and visualization, self-reflection and communication among participants for building the community might be challenging. Also, since the data was already collected by the others, data collection cannot be covered in this workshop. To catch up on the gaps and

improve difficulties, facilitators need to sincerely and actively engage and converse with participants while archiving the interplays and non-verbal communication happening in the workshop.

Responsibility Distribution

We would like to mention that Sanchita was responsible for data cleaning and attribute extraction, visualization on RawGraphs, Palladio and Tableau while Inyoung worked on data filtering, numeric analysis of a particular attribute and visualization on Flourish and Venngage.

References

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in The Networked World.* The MIT Press.

CODATA (2021). "Data curation" URL: https://codata.org/rdm-terminology/data-curation/

D'Ignazio, C., & Klein, L. (2020). 6. The Numbers Don't Speak for Themselves. In *Data Feminism*. https://data-feminism.mitpress.mit.edu/pub/czq9dfs5

Duke, C. S., & Porter, J. H. (2013). The Ethics of Data Sharing and Reuse in Biology. *BioScience*, *63*(6), 483–489. DOI: 10.1525/bio.2013.63.6.10

Eichmann-Kalwara, N., Jorgensen, J., & Weingart, S. B. (2018). Representation at digital humanities conferences (2000–2015). *Bodies of Information: Intersectional feminism and digital humanities*, 72-92.

Estill, L., & Guiliano, J. (Eds.). (2023). *Digital Humanities Workshops: Lessons Learned*. Taylor & Francis.

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. DOI: 10.1002/asi.23480

Frank, R. D., Tyler, A. R. B., Gault, A., Suzuka, K., & Yakel, E. (1970). Issues of Privacy in Qualitative Video Data Reuse. *International Journal of Digital Curation*, *13*(1), 47–72. DOI: 10.2218/ijdc.v13i1.492

GSS General Social Survey | NORC. (n.d.). Retrieved 15 December 2023, from https://gss.norc.org/*Statistical inference with the GSS data*. (n.d.). Retrieved 13 December 2023, from https://rstudio-pubs-static.s3.amazonaws.com/609489_845dec65eb9c4c84b9dfc3294045c4a7.html

Harper, L. M. (2023). *Data Reuse Among Digital Humanities Scholars: a Qualitative Study of Practices, Challenges and Opportunities* (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. A. (2022). How do properties of data, their curation, and their funding relate to reuse? *Journal of the Association for Information Science and Technology*, *73*(10), 1432–1444. DOI: 10.1002/asi.24646

Hinrichs, U., Forlini, S., & Moynihan, B. (2019). In defense of sandcastles: Research thinking through visualization in digital humanities. *Digital Scholarship in the Humanities*, *34*(Supplement_1), i80-i99.

LeBlanc, Z., & Walsh, B. (2019). Workshopping the Workshop: Moving Your Sessions Beyond Buttonology. *# DLFTeach Toolkit: Lesson Plans for Digital Library Instruction*.

Oishi, S., Graham, J., Kesebir, S., & Galinha, I. C. (2013). Concepts of happiness across time and cultures. *Personality and social psychology bulletin*, *39*(5), 559-577.

Poole, A., A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science*, *15*(2), 101–139. DOI: 10.1007/s10502-014-9236-y

Rawson, K., & Muñoz, T. (2019). Against Cleaning. In M. K. Gold & L. F. Klein (Eds.), *Debates in the Digital Humanities 2019* (pp. 279–292). University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk.26

Russell, J. R., & Hensley, M. K. (2017). Beyond buttonology: Digital humanities, digital pedagogy, and the ACRL Framework.

Schöch, C. (2014). *Big? Smart? Clean? Messy? Data In the Humanities*. DOI: 10.5281/ZENODO.8432

Sielemann, K., Hafner, A., & Pucker, B. (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ*, *8*, e9954. DOI: 10.7717/peerj.9954