Digital Literacy and Privacy Protection

Project Paper

Hai Jing (Jane) Tu[a], Rahul Devajji[b]

[a]Data Science Graduate Online Certificate Program, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
(Email: tuhai@iu.edu)
[b]Data Science Graduate Residential Program, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
(Email: rdevajji@iu.edu)

1 Abstract

This project seeks to conduct secondary data analysis and visualization on the topic of digital literacy and privacy protection using American Trend Panel (ATP) Wave 49 results provided by Pew Internet Research. Americans are feeling less secure about their privacy online, as a majority of them (72%) report concerns about the prevalence of tracking and even more of them (81%) report lacking control over data collected about them [2].

Digital literacy, including knowledge and experiences with personalized information, data-driven ads, and privacy agreement on social media, is being studied to explain the skepticism or confidence about the usage of social media. The ATP wave 49 survey shows that while phishing scams and web cookies are familiar to the majority of US adults, subjects such as two-factor authentication and private browsing are not as well known.

When it comes to privacy protection, users reported frequently being asked to agree with privacy policies online, but only a small percentage of them actually read through the privacy policy. Data privacy laws are not well known among American adults, with 63% of them saying they understand very little or nothing at all about the laws and regulations that protect their privacy [2]. Ordinary social media users have little power against the platform and their control over privacy is flimsy.

By recreating the data visualization in the existing reports published by Pew Research based on the ATP Wave 49 dataset, the goal of this project is to reinforce the significance of digital literacy in empowering social media users. We will first study and critique data visualization used in the report "Americans and Privacy, Concerned, Confused, and Feeling Lack of Control Over Their Personal Information" [2]. Second, we will recreate the data visualizations on digital literacy, concerns about privacy, level of control over personal data, and plot them across age, gender, as well as demographic variables. Thirdly, we will conduct statistical analysis and plot out the association between digital literacy and other variables.

2 Introduction

How much should ordinary citizens know about technology, algorithm, and coding? The influence of digital technology is ubiquitous in our daily life, but the necessity of mastering knowledge on how digital technology works is still disputable. For example, coding is believed to foster creative and logical thinking, but it is still not a required skill for students even at college level[11]. Sedgewick wrote that "our technology-driven economy needs coding-literate citizens who are competitive, astute and discerning in the global marketplace of ideas, opportunity and commerce." Standing on the opposite side, Cuban[3] argues that "America's public schools should not exist to serve Silicon Valley CEO's need for programmers. Turning tax-supported schools into job-training sites for high-tech firms corrupts the very purpose of public education." Many people believe that coding skill is simply serving corporate interest that don't have to be catered by public schools, and there are more cognitive domains that provide training in creativity and logical thinking.

All of the visualizations in this paper, unless from an external source, can be recreated and accessed at:colab notebook and github

During recent years of news education on data breach, the general public have become more aware of the fact that personal data are being used for corporate decision-making and controlling data flow. There are growing concerns over how algorithms work, and about half of Facebook users say they are not comfortable when they see how the platform categorizes them[6]. In addition, worries arise over algorithms putting too much control in the hands of corporations and governments, perpetuating bias, creating filter bubbles, etc. [8]. Also arise is the urgency for citizens to be informed and educated on digital technology, such as how algorithms collect personal data and select personalized data for users. In order to battle for control and secure more power for protecting their privacy, digital literacy becomes essential.

There have been a great deal of research on the subject of digital literacy. In 2016, the Pew Research Center listed seven themes of the Algorithm era after concluding the seventh Future of the Internet study. The last of the seven themes recognizes the growing need for algorithmic literacy, transparency and oversight [8]. It is argued that "public education should instill literacy about how algorithms function in the general public", and "those who create and evolve algorithms should be held accountable to society" [8][p.15]. It's a common concern that human agency seems to be losing the decision-making power on key aspects of digital life to code-driven, black box tools[9]. A proper understanding of digital technology is obviously necessary for making educated decisions on social matters, because digital literacy contributes to awareness of power distribution across the social network.

Fear of privacy traps such as online phishing is common among social media users. Head, et al.'s research [5] show that as many research participants worry about "the 'creepiness' of algorithms that violated their privacy"(p.15), not all of them take action to keep personal information from prying algorithms, and these are often participants who trust algorithms doing more good than harm. Bartsch and Dienline [1] find that frequency of changing privacy settings on social media is positively related to online privacy literacy, and those who have more online privacy literacy feel safer when using SNSs.

Overall, digital literacy is a popular topic about not just literacy but also justice. Clearer graphs and charts will be greatly helpful to illustrate its necessity. Pew Research has previously released reports with a lot of visualizations on this subject. As shown in section 4, bar chart and pie charts are two of the most often used graph.

3 Dataset

The dataset is American Trends Panel Datasets Wave 49 available at Pew Research. There are in total 4272 individuals who completed the survey. The downloaded dataset is in `.sav` format. Each individual is asked over 120 questions, some multiple choice and some descriptive questions. All the multiple choice question responses have been recorded and stored in this dataset. It has 143 columns/attributes/questions which have detailed information about the respondent including their income, political affiliation, age and other demographic information other than the 118 questions they are asked as a part of the survey.

Pew datasets are known for their high quality and validity. We have chosen this dataset in particular because it offers a great resource for us to explore the underlying factors that contribute to the study of digital literacy and privacy protection. We wanted to improve the graphics and visualization techniques presented in the existing report. The idea is to explore the relationship between the demographic variables, digital literacy, levels of control and concern about privacy among participants.

Since our goal is to emphasize the need for digital literacy, we want to find exactly where the problem lies? The original report has highlighted what the problem is, so with our statistical analysis and visualizations, perhaps we can find out which demographics are lagging behind. We also experiment with a variety visualizations with this dataset to further utilize the richness of the data.

For this project, we select 17 questions measuring digital literacy, concerns about privacy, control of personal data, and demographic characters to explore how the level of digital literacy varies among people of different demographic characters, levels of control, and concerns for privacy.

4 Existing Visualizations

The researchers of the survey have also generated a report which provides an overview of the survey responses by exploring various attributes[2]. The report primarily explores the demographics of privacy and the overall feeling of concern and lack of control towards the ownership of personal data. This is done by visualizing the attributes (questions asked in the survey) individually, with the only correlation represented by graphs between race and age.

Some of the core findings explored in the report includes: How concerned are the respondents about who can access their personal data? What's their level of understanding about the current social media landscape? How often are they being subjected to data thefts and breaches? What is their knowledge level of privacy and data security on the internet?

4.1 Critique on Existing Visualizations

Visualizations included in the report show consistent style widely used in all Pew reports. They show a good representation of data with minimal ink and minimal colors. It's a unique style that combines fonts, color, and shapes that are easily identifiable to be Pew's. Every graph has been labeled appropriately and there are no instances where the information is lost in a sea of labels.

Given the already excellent graphs in the report, there is still room to improve the visualizations. As mentioned above, bars are the primary visualization technique used in the report (see Figure 1 and Figure 2), they represent nominal data and the labels defining the length of each sub-division in the bar is helpful as it makes it easier to infer the difference. The use of contrast while writing the labels on the bars is a good technique to improve readability.

Figure 1 is an example of uniform color usage, everything is represented in various shades of blue color. Figure 2 is an example where a second color, green, is used. However, the colors in both figures do not translate well into gray scale. After converting the figures to gray scale, the difference between various responses is lost even with the length of the columns being labeled.

The bar graphs, although clear in presenting the information, do not emphasize the key point instantly. We are exploring a different approach to plot the same information by using divergent stacked bar graphs to emphasize the trend/message in a better way.

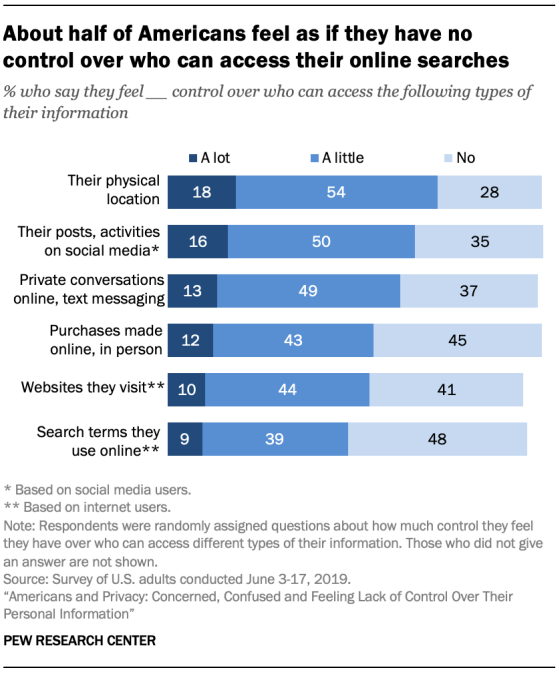


Figure 1: Visual encoding of length and point

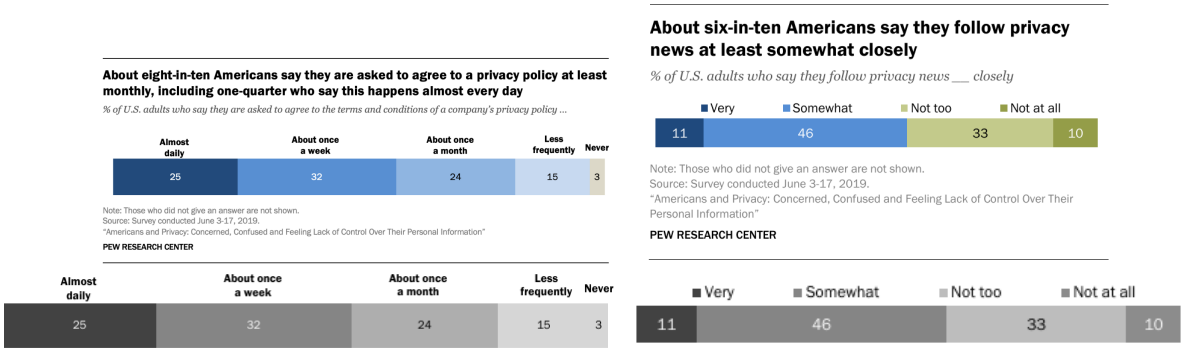


Figure 2: Color scheme does not translate well in gray scale.

Figure 3 shows two horizontally stacked bar graphs. The bars are color coded to show percentage of American's who have different levels of understanding of existing data protection laws. "Vert little" is centered and most obvious. But the choice of color does not make sense because green and blue are not a good pair of colors to use together. According to Wong[14], "[c]olor can elicit size biases; some people find equal areas filled with vibrant colors seem to be more dissimilar than when less saturated colors are used." Therefore choosing appropriate color is definitely key to the effectiveness of colored charts.

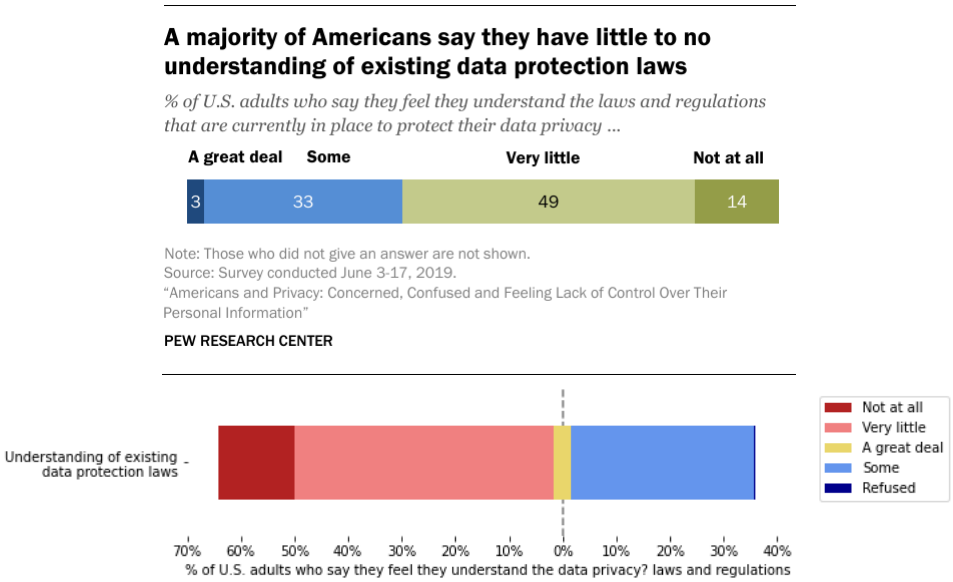


Figure 3: Stacked bar graph vs Horizontal bar graph

Figure 4 is a combination of pie chart and stacked bar charts to illustrate answers from three questions: levels of knowledge about profiling from the perspective of users, how many companies do they think rely on profiles, and how often do they see ads as based on profiles. The three questions are not easy to read together, and the numbers cannot be processed immediately by readers. There are 77 percent of users who heard about personal profiles, and these people are further divided by their answer to the two statements that follow. The entire graph is cumbersome and can exaggerate the percentage number on the stacked bars, given they are out of only 77% of the respondents.

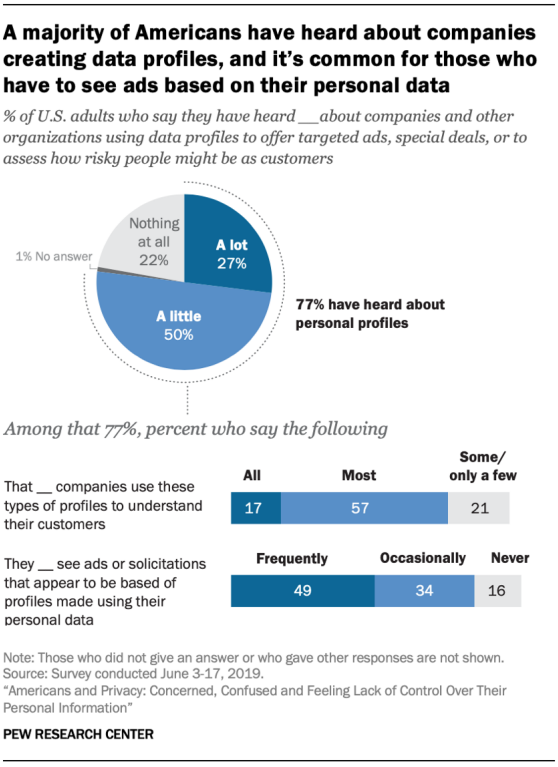


Figure 4: Two level correlation graph

The two charts in Figure 5 are created by dividing the respondents based on their race and age. This a simple graph which uses points to represent the data and the difference between various age groups. The color variation between the points in the first graph does not translate well into gray scale as the light blue and dark green look very similar (in gray scale).

The second graph describes the concern levels of various Americans about tracking based on age, race, income and education level.
We can observe here that visual encodings are used accurately, points where the data being shown is less and bars where the data being shown is more, thereby improving the readability of the graphs.

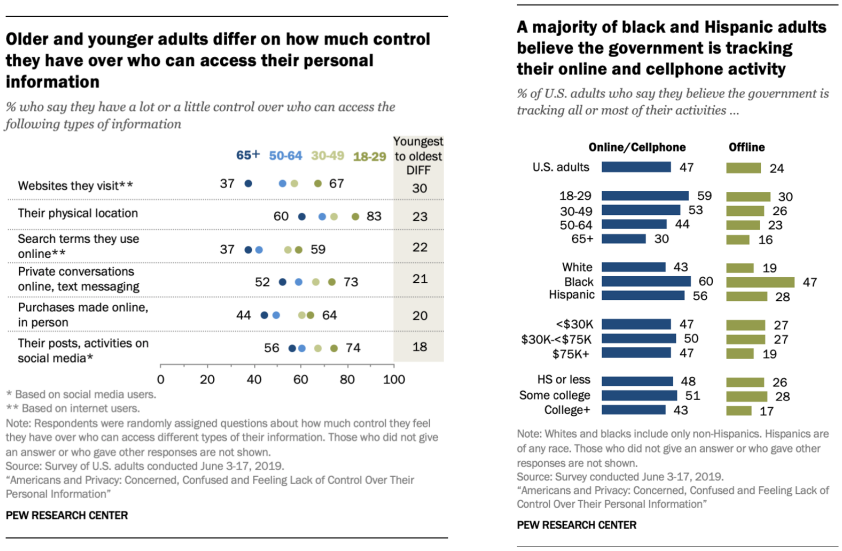


Figure 5: Demographic correlation graphs

Overall, the graphs are easy to look at and address the need of ordinary people without analytical background. Percentage is the most reported measures next to frequency. With a rich dataset that has over 4000 responses across many subjects, it's ideal to analyze association between digital literacy and the level of control on privacy, in addition to study these two variables separately.

5 Objectives

Our core objective of this project is to highlight the need for digital literacy. Based on the existing dataset, we will create visualizations that prove the lack of knowledge in digital technology and demonstrate the urgent need for digital literacy. We will recreate some of the visualization by Pew Research[2] to emphasize on the impact of demographic factors on digital literacy as well as control over privacy. Most importantly, we will create visualizations that reflect on the association between digital literacy, concerns about privacy online, and level of perceived control participants have over their privacy online.

The dataset is available in a format which can be readily imported by `Pandas` library for analysis. Since all the questions are in the likert-scale, we will explore other methods of visualizing this data, similar to the sample shown in figure 3. We plan to create inclusive charts by using a better color palette. Since the existing report uses a uniform visualization technique of bar graphs only, we will explore the effectiveness of data representation by other forms of charts like pie charts, box plots for comparison of means, and histogram to visualize the data. We will start with demographic data, then continue to variables that represent digital literacy and privacy control.

6 Methods

The major goal of this paper is to highlight the need of digital literacy among the general public using Pew research data. We will first visualize the demographic data to show the general picture of all participants in this survey. We will use bar charts to visualize age, income, education, and a pie chart for geological locations(metropolitan vs. non-metropolitan).

We will then select a group of questions to measure digital literacy, control and concern about privacy on social media. We will use statistical analysis to find out the relationship between these variables and digital literacy measurement based on the variable TOTALKNOW.

Based on a thorough study of the questionnaire[2], we select the following questions about knowledge of social media as measurement of digital literacy: UNDERSTANDCO, UNDERSTANGOV, PP4, KNOW1-KNOW10, TOTALKNOW.

TOTALKNOW is the score participants receive on the 10 questions on digital literacy, with minimum value 0 and maximum value 10. This will be a numerical variable that we will use to conduct ANOVA and t test with.

We select the following questions as measurement for how much control users reported to have over personal data: CONTROLCO and CONTROLGOV. These are both ordinal level variables. We will conduct ANOVA test to find out if there is significant difference between participants' levels of control and their digital literacy.

We select the following questions as measurement for how much participants are concerned about personal data being collected: CONCERNCO and CONCERNGOV, TRACKCO1a, TRACKCO1b,TRACKGOV01a, and TRACKGOV01b. These questions are also ordinal variables, therefore we will run ANOVA to find out whether the difference in digital literacy between levels of concern is significant or not.

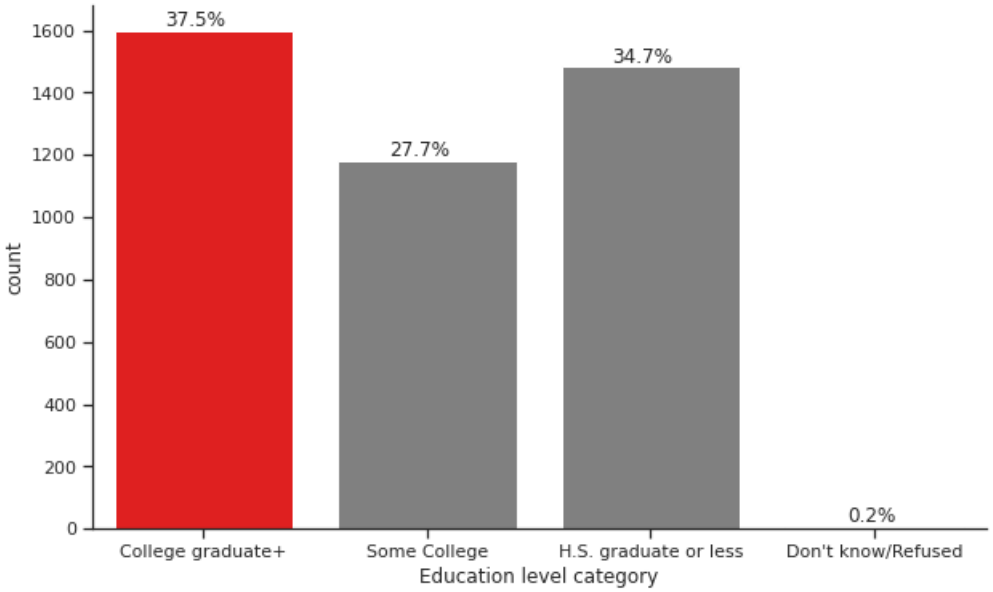
In the ANOVA tests, we will use "control" variables and "concern" variables as independent variables TOTALKNOW as dependent variable.

In addition to bar charts, we use box plots to display results of ANOVA tests. Specific test results will be reported and the visualization for each test will be explained in the paragraph before the graph.

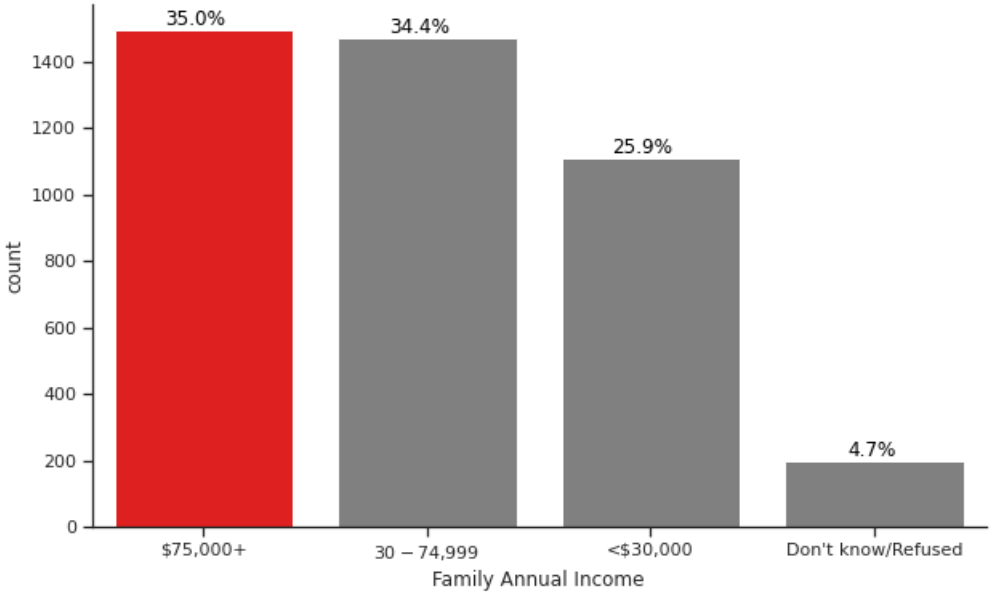
7 Findings and Visualizations

7.1 Demographics

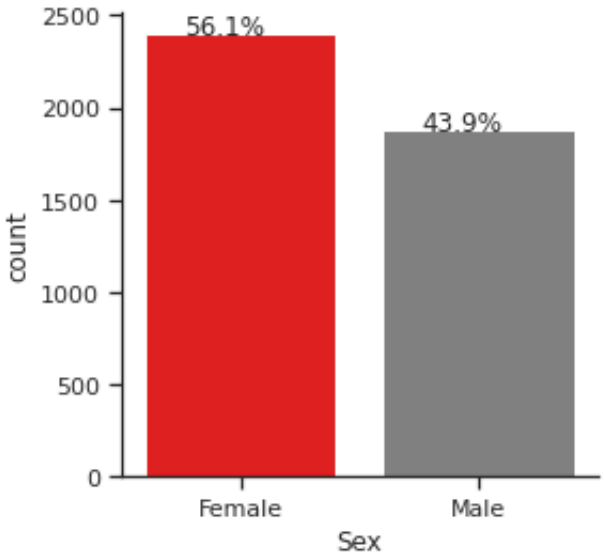
As can be seen from the graphs below, participants in this research are highly educated, with about 37% having college or post graduate degrees.



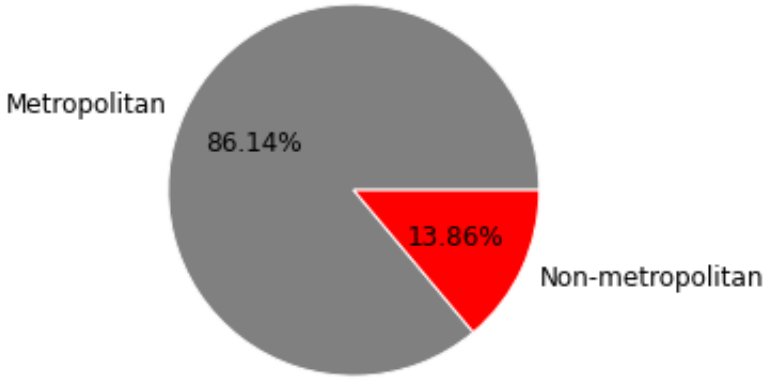
Participants also tend to have higher income, with the majority of the participants (69%) having income of 30k or above, and nearly 35% of the participants' income are above 75K a year.



Speaking of gender, there are more males than females included in the survey, with 43.9% males and 56.1% females.



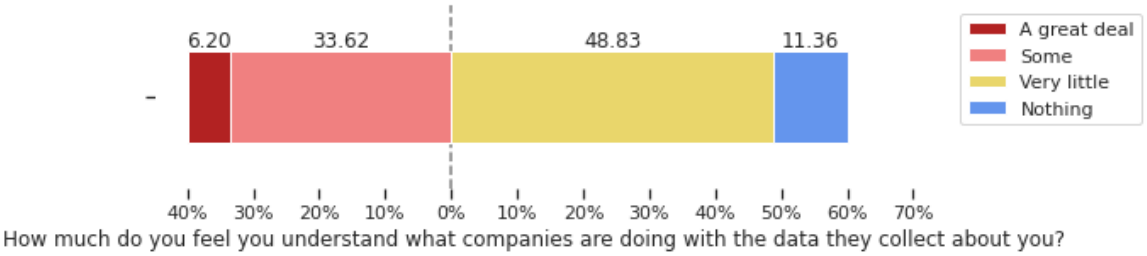
Based on the data, the vast majority of the participants (86.1%) live in a metropolitan area, while only 13.8% of them live in a non-metropolitan area.



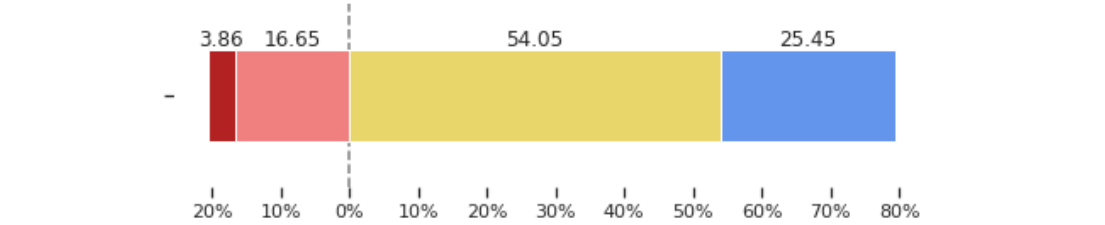
7.2 Digital Literacy

We are using the following questions to measure digital literacy among the general public: UNDERSTANDCO, UNDERSTANGOV, PP4, KNOW1-KNOW10, TOTALKNOW.

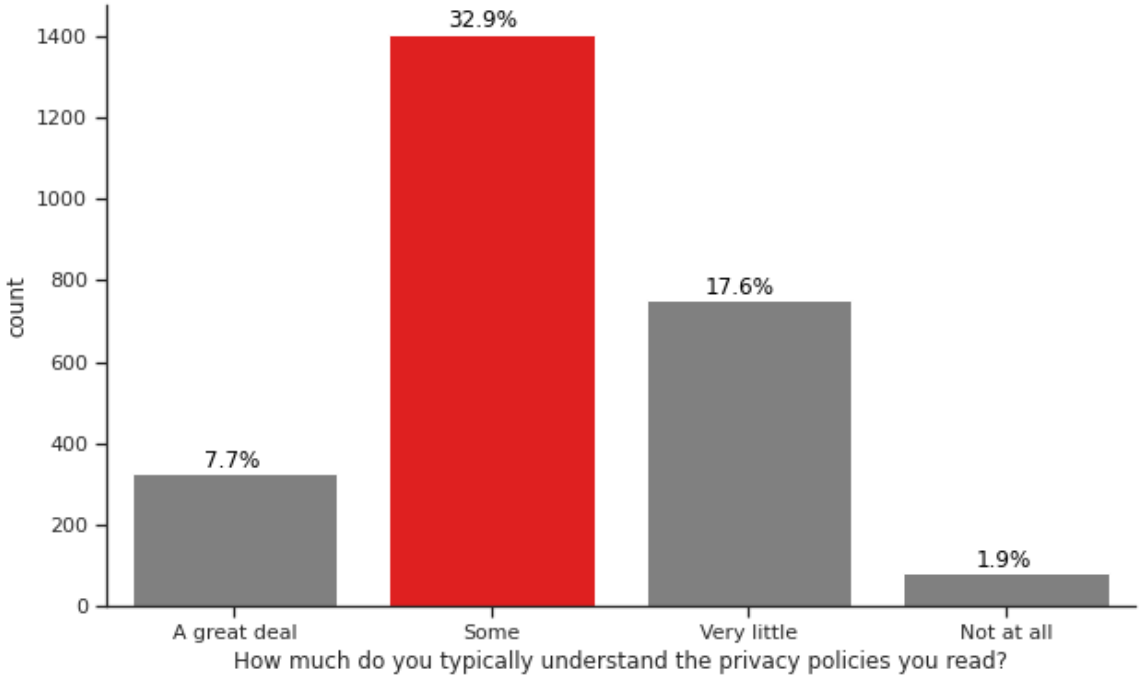
UNDERSTANDCO and UNDERSTANDGOV each asks whether participants understand what companies and the government are doing with the data they collect about them. The questions pp4 measures how much they typically understand the privacy policies they read online. All these variables use ordinal measurement that goes from "A great deal", "some", "very little", to "Not at all".
We use bar charts to visualize these variables that reflect the general understanding of privacy and the keeping of personal data by companies and governments:



How much do you feel you understand what companies are doing with the data they collect about you?

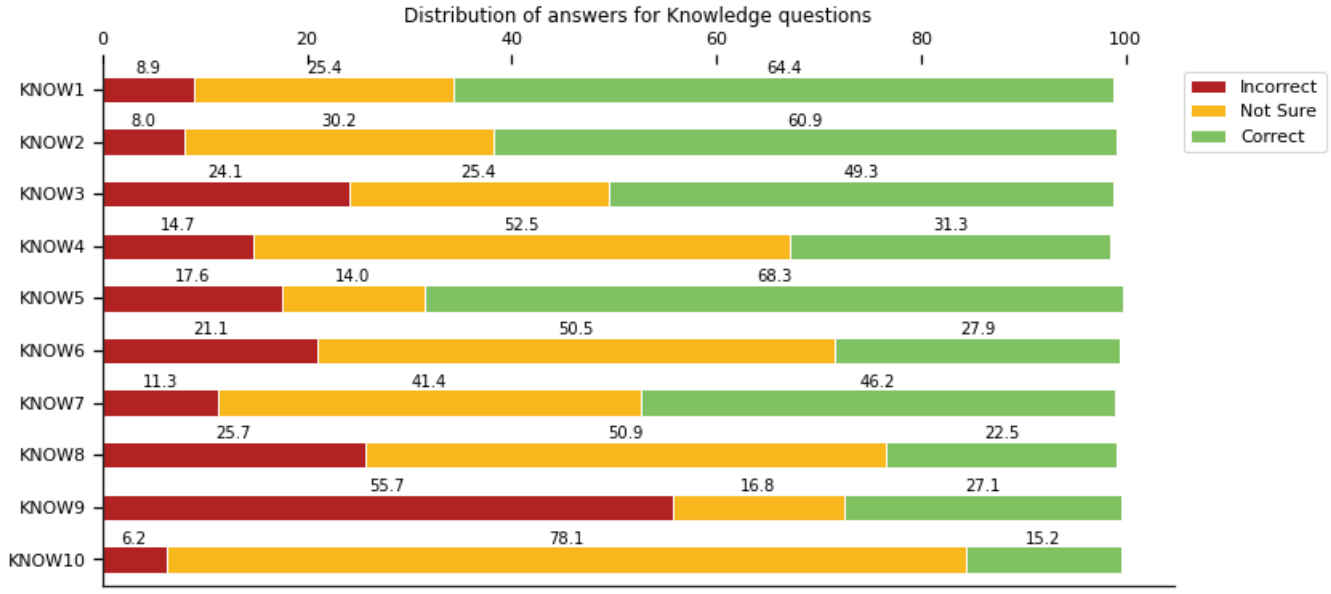


How much do you feel you understand what the government is doing with the data they collect about you?

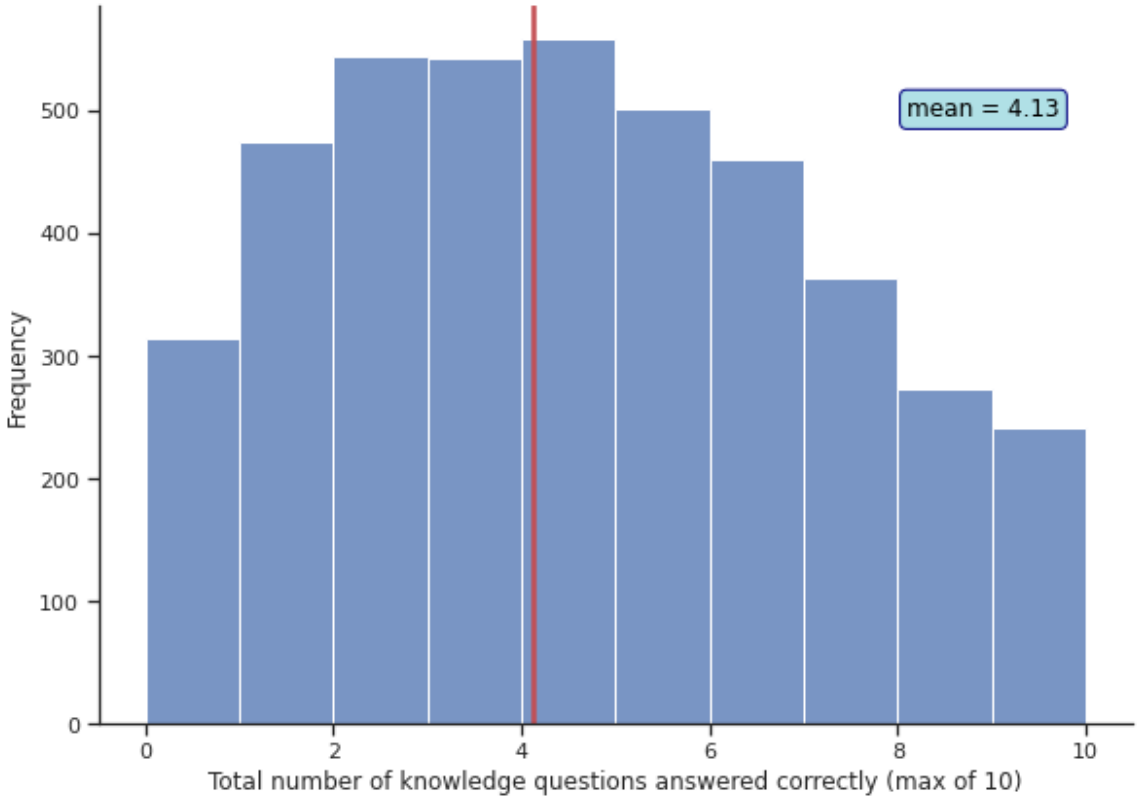


How much do you typically understand the privacy policies you read?

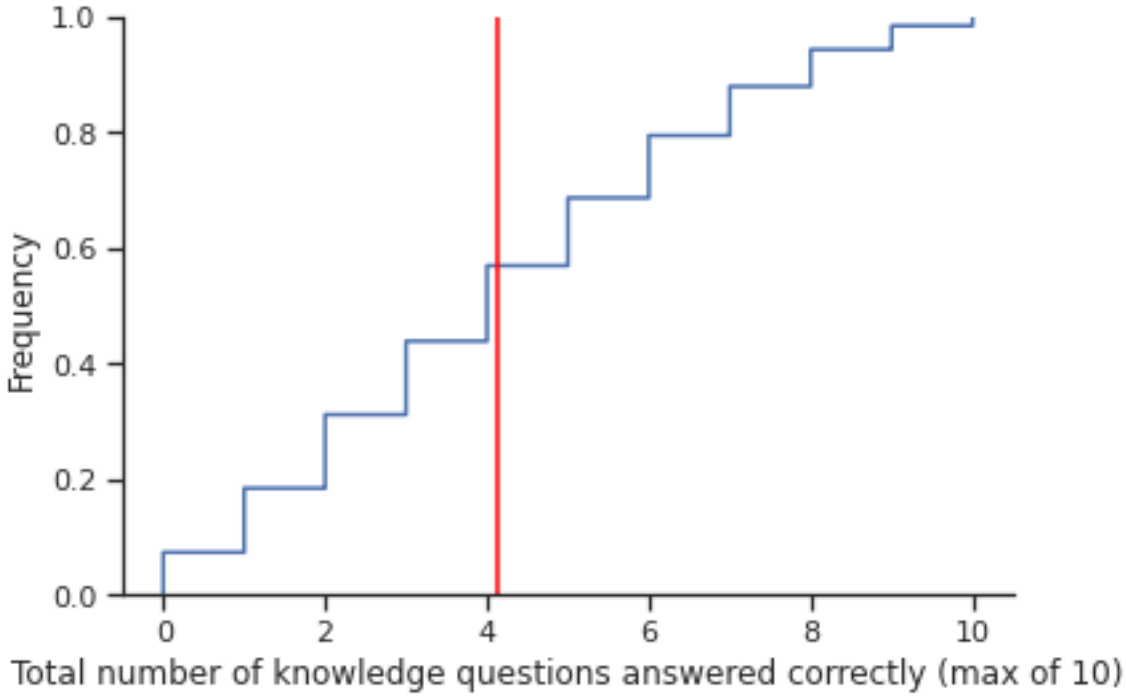
KNOW1-KNOW10 are ten questions that test their knowledge on digital technology, including cookies, sources of revenue for most major social media platform. For example, the question that has the most incorrect answers, KNOW9, asks "Some websites and online services use a security process known as two-step or two-factor authentication. Which of the following images is an example of two-factor authentication?" Only 27% of the participants got the correct answer. KNOW10 shows participants a picture of Jack Dorsey and ask who the man is, 78% of participants answered "not sure," while 15% got the answer correctly.KNOW1 to KNOW10 are all visualized in one graph below:



TOTALKNOW is the only numerical variable that record how many questions from KNOW1 to KNOW10 did each participants answer correctly, with a minimum of 0 and maximum of 10. A histogram as well as a cumulative histogram are constructed for the variable TOTALKNOW:



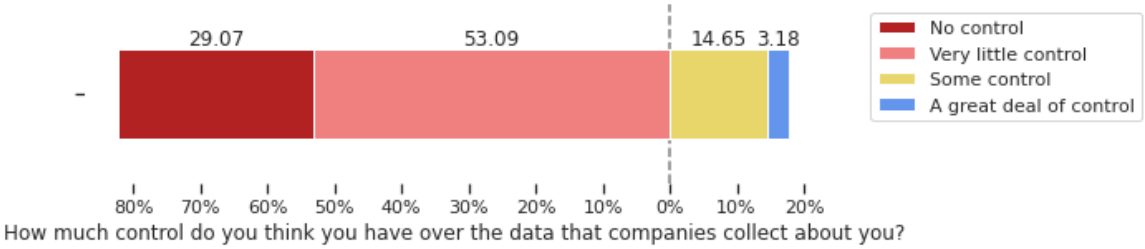
As shown in the graph TOTALKNOW, the mean is 4.13, and the distribution shows a positive skew, which means there are more participants with lower score in the range of 1 to 10.

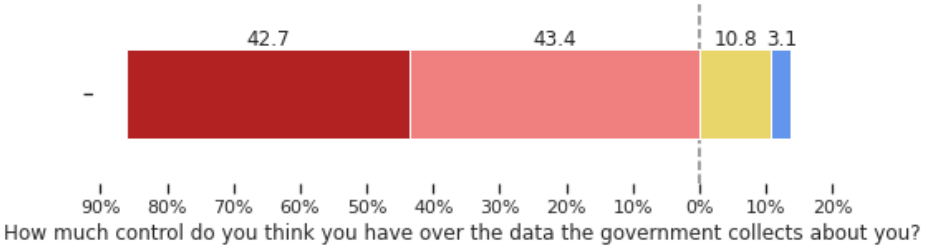


7.3 Control and Concern

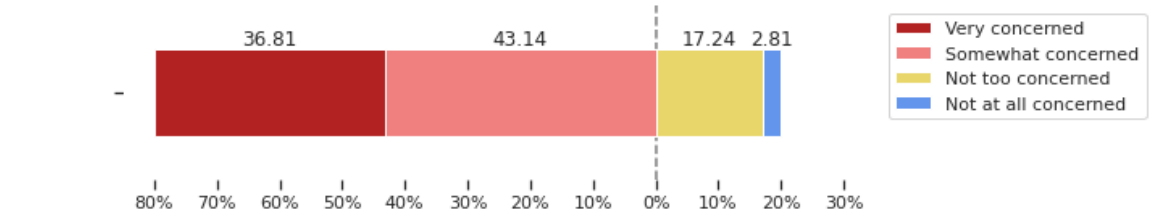
We measure control and concern about personal data protection using the following variables: CONTROLCO, CONTROLGOV, CONCERNCO, CONCERNGOV, TRACKCO1a, TRACKCO1b,TRACKGOV01a, and TRACKGOV01b.

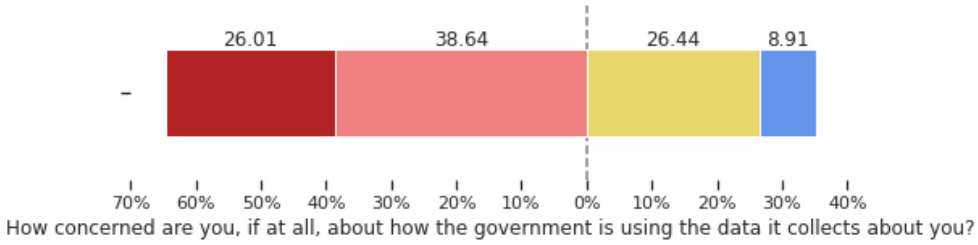
CONTROLGOV and CONTROLCO ask how much control do participants feel they have over the data the government or companies collect about them. The answers to control questions range from "no control," "very little control," "some control," and "a great deal of control."





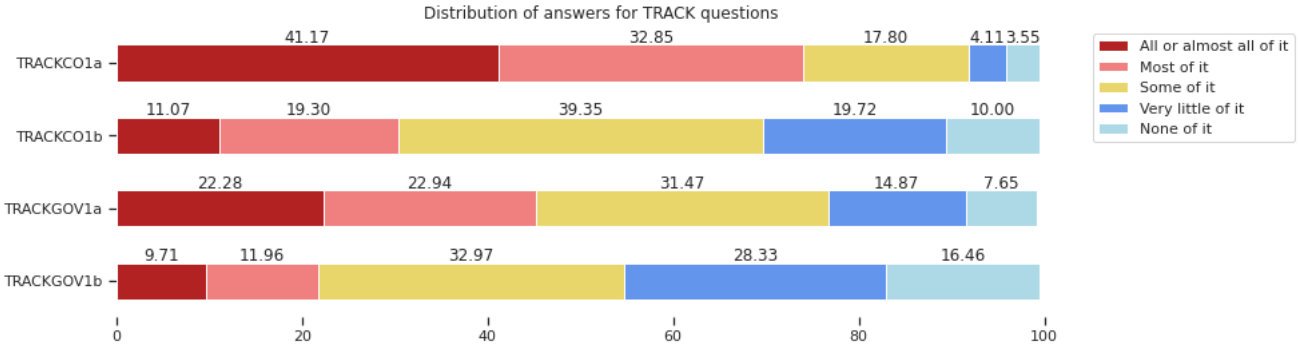
CONCERNGOV and CONCERNCO ask how concerned participants are about how companies and the government are using the data they collect on participants. The answers to "concern" questions range from "very concerned", "somewhat concerned", "not too concerned", to "not at all concerned."





The series of questions on tracking (TRACKCO1a, TRACKCO1b,TRACKGOV01a, and TRACKGOV01b.) ask how much of what participants did online as well as offline are being tracked by the government and advertisers, technology firms, or other companies. The answers use Likert type scale that range from "all or almost all of it", "most of it",""some of it, " very little of it", and "none of it."

The following graphs show the comparison of participants' answer to these questions. Most participants are aware that all or almost all of what they do online are tracked by companies and the government (41.17%, 22.28%), while very few of them think all they do offline are being tracked by companies and government(11.07%, 9.71%).

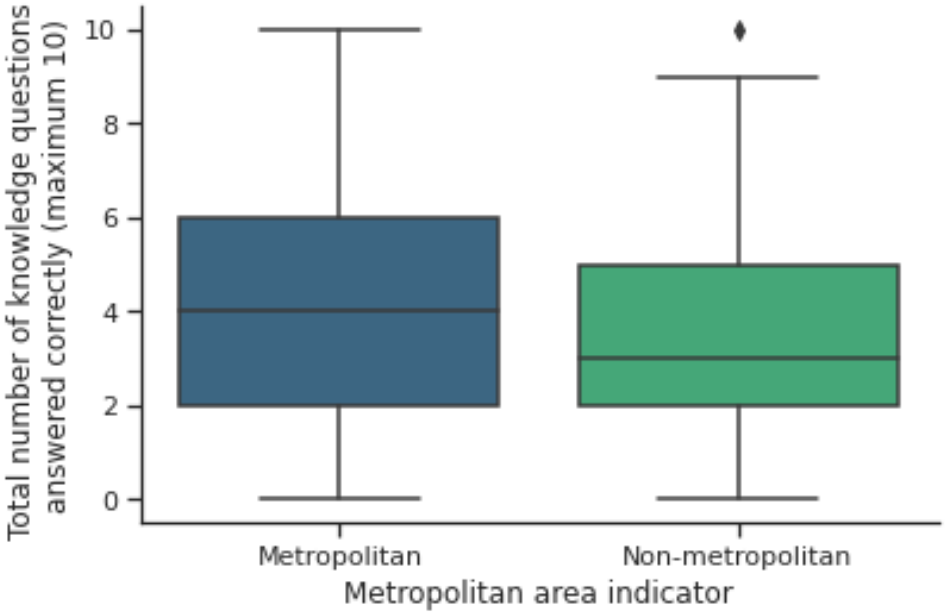


7.4 Comparison of Means

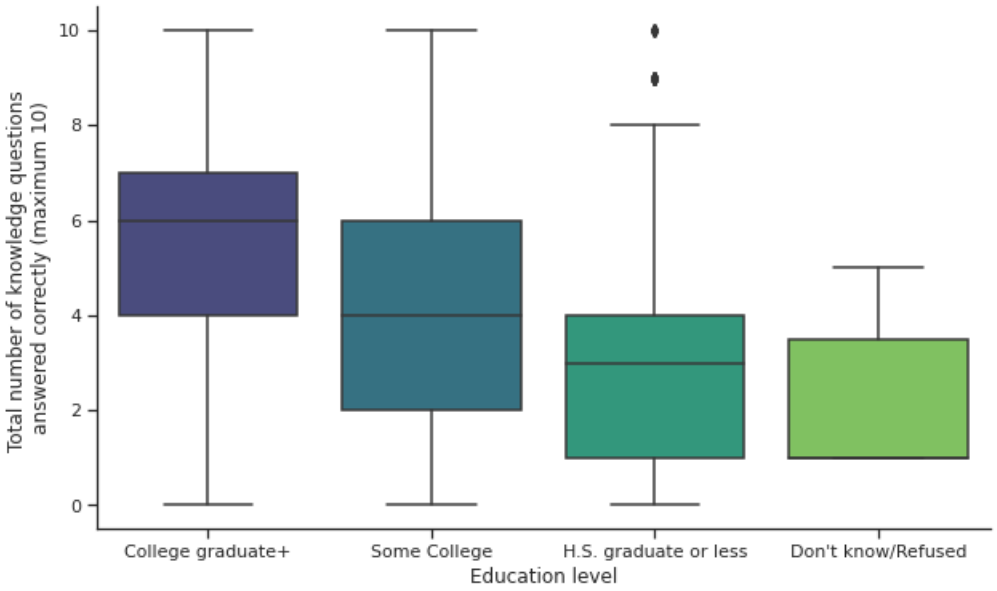
Boxplot is the best way to plot comparison of means. According to [13], "boxplots were invented by the statistician John Tukey in the early 1970s, and they quickly gained popularity because they were highly informative while being easy to draw by hand". Even today, results of t tests and F tests are still visualized using boxplots.

7.4.1 Comparison Using Demographic Characters

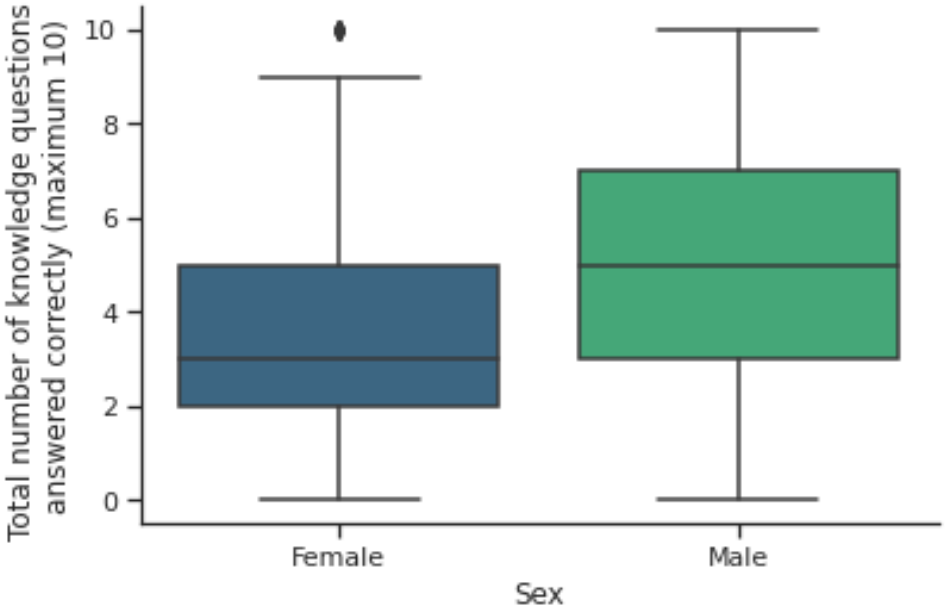
First, to find out if demographic factors affect digital literacy, we conducted a t test to to find out the difference in digital literacy between metropolitan and non-metropolitan residents $t = 6.517, p < .000, N = 4268$



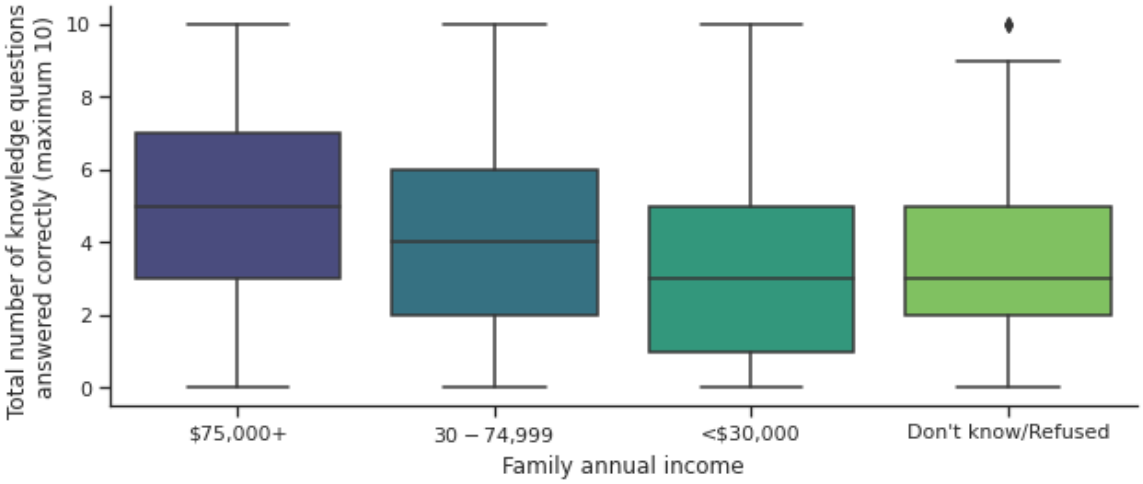
We also use education as the independent variable to find out the difference it makes in participants' digital literacy level. The result shows clear decrease in digital literacy as the education level decreases $F = 263.58, p < .000, N = 4272$.



Gender is used to find out if there are differences in digital literacy between males and females. The results shows females tend to have lower digital literacy score than males $t = 15.93, p > .000, N = 4270$



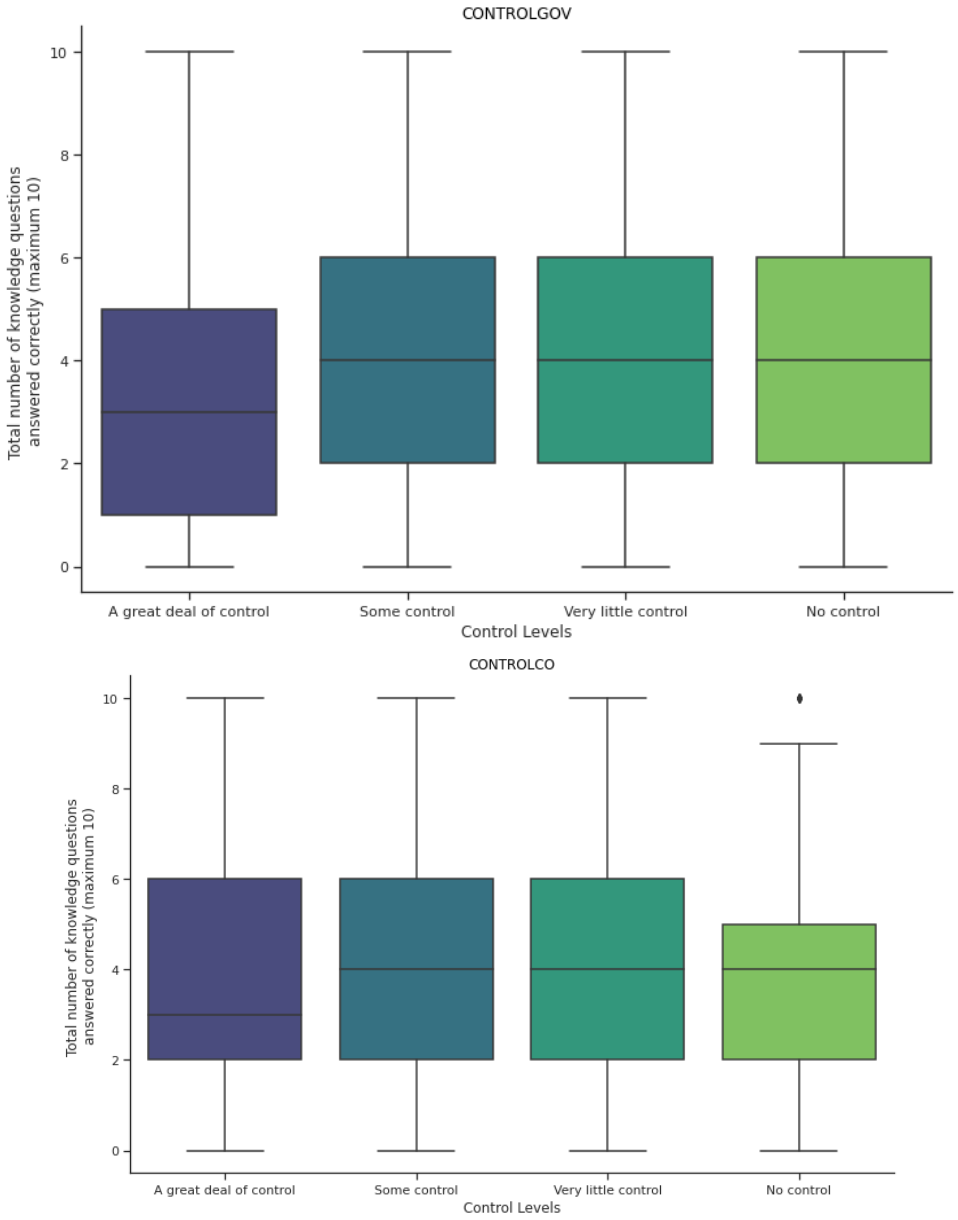
Income also indicates a difference in digital literacy $F = 159.05, p < .001, N = 4272$:



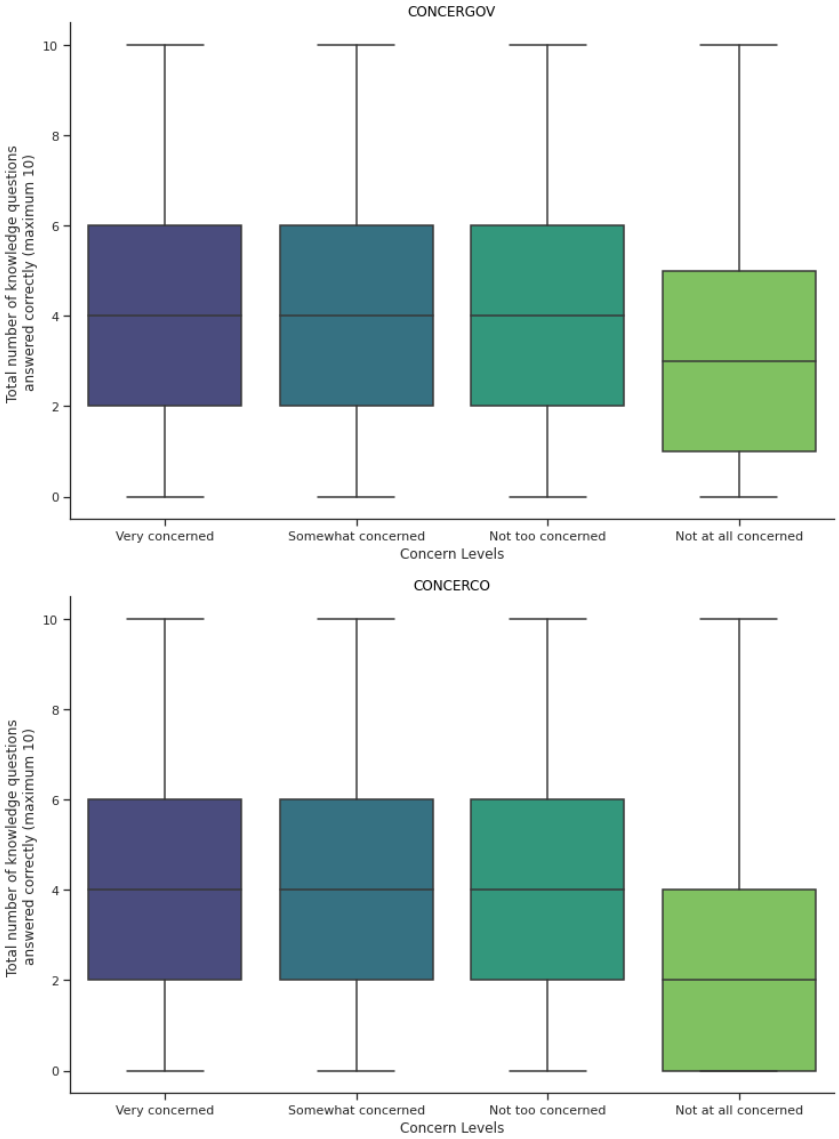
7.4.2 Comparison Between Control and Concern levels

To show the difference between levels of control and concerns on the participants' digital literacy, we conducted a series of ANOVA.

First, we use CONTROLGOV ($F = 6.017, p < .000, N = 2132$) and CONTROLCO ($F = 8.732, p < .000, N = 2140$) to find out if they make difference in the total knowledge of digital literacy:



Next, we use CONCERNGOV and CONCERNCO to run the same ANOVA test. It shows that both CONCERNGOV($F = 10.668, p < .000, N = 2132$) and CONCERNCO ($F = 8.732, p < .001, N = 2140$) make significant differences in participants' digital literacy.



7.5 Correlation

To find out about the relationship between the understanding of personal data collected by the government (UNDERSTANDGOV) and companies (UNDERSTANDCO), the reading of privacy contract (PP4), and digital literacy (TOTALKNOW), we conduct a correlation test between these variables. Recoding PP4, UNDERSTANDGOV and UNDERSTANDCO into numerical variables, "a great deal"=1, "some"=2, "very little" =3, "nothing"=4. The correlation table shows significant correlation between all three variables on understanding and TOTALKNOW.

Correlations

	Mean	Standard Deviation	PP4	UNDERSTANDCO	UNDERSTANDGOV	TOTALKNOW
PP4. How much do you typically understand the privacy policies you read?	2.42	4.33	1			
UNDERSTANDCO. How much do you feel you understand what companies are doing with the data they collect about you?	3.10	6.61	.001 (.97)	1		
UNDERSTANDGOV. How much do you feel you understand what the government is doing with the data it collects about you?	3.28	5.14	-.002 (.95)	.[b]	1	
TOTALKNOW. Total number of knowledge questions answered correctly (maximum of 10)	4.13	2.59	-.06** (.002)	-.07** (.001)	-.06** (.01)	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

b. Cannot be computed because at least one of the variables is constant.

8 Discussion

8.1 Demographic Characters and Digital Literacy

Based on the demographic characters of the participant,the majority of them are from a metropolitan area with a good education and a good income. By including demographic variables in the analysis, we find that geological location, education, gender, and income all make difference in participants' digital literacy. Obviously the mean for digital literacy scores is higher among those with higher income, better education, who are males, and from a metropolitan area.

These are not surprising findings, but should be taken into consideration how the society still lacks justice in providing equal opportunities for people from all backgrounds to obtain digital literacy and take control of their personal information. Awareness of information being used by the government and companies is key to individuals' digital welfare. It should be a right, not a privilege for individuals to take control of their personal information online.

8.2 Control, Concern, and Digital Literacy

It's obvious that people who reported high levels of concern show significantly higher grade in total score on digital literacy. However, it's interesting to see that people who reported that they have "a great deal of control" tend to have lower digital literacy score, while those who reported moderate control and very little control receive higher score on digital literacy test. Therefore people who report a lot of control have limited knowledge about digital technology, while they seem to have more confidence over control. For those who have more knowledge about digital literacy, they reported little confidence in the control they have over their personal data collected by the government and companies.

8.3 Personal Data Awareness and Digital Literacy

The relationships between UNDERSTANDGOV, UNDERSTANDCO, PP4, and TOTALKNOW show that the the more individuals understand about what companies and the government are doing about their personal data, the higher they score on the digital literacy tests. Those who actually understand privacy contract online tend to have a higher score on digital literacy.

The results tell us the trend, but not the details about individuals. For example, what makes certain individuals curious about privacy contract? What experiences did they have about privacy breach? What are the major limits to their efforts to achieve a better understanding of privacy online? These questions are not answered from the Pew survey.

8.4 Justification for Visualizations

We chose to use mostly bar charts and box plots to visualize the data generated by this Pew Survey[2]. It is shocking to see how box plots are missing from major news analysis and reports generated from a high quality surveys such as this one. [2] The most frequently used charts are pie charts and bar charts.

Based on the type of data (mostly categorical and ordinal), it is reasonable to use bar charts to present the findings, as was done in the Pew research report. However, the existing reports do not touch on significance test and did not make good use of box plots to show two dimensional data.

Box plots are very effective for comparison of means. The story told by box plots here reveal the impact of demographic characters on digital literacy. There are many more that can be compared, such as demographic characters and levels of control and concern.

According to [13], box plots can be drawn by hand, which made it popular before modern computing enters the playing field. Recently, with modern computing and visualization capabilities, we see box plots being replaced by violin plots. In our visual experiment, we tested a similar visualization graph– **beeswarm** plot, to reflect on density estimation. However, it did not turn out to be successful. An example of **beeswarm** is shown below:

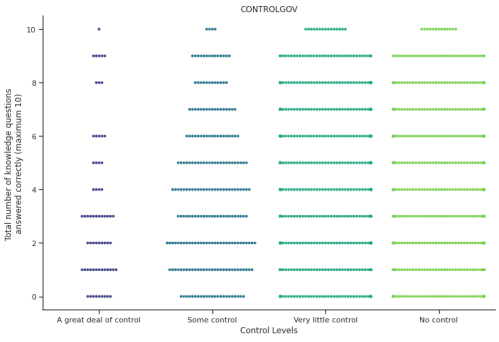


Figure 6: Swarm plot

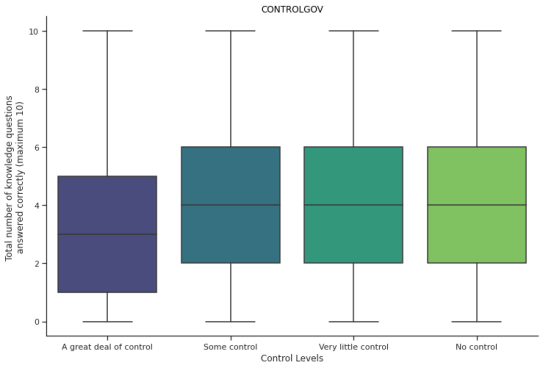


Figure 7: Box plot

Because of the nature of the data, the numerical data have only eleven values, therefore it's hard to visualize the density estimate across the y-axis.

There was a similar design choice made between vertical grouped bar chart [4], joyplots[12] and horizontal stacked bar chart while visualizing the responses to various knowledge and tracking questions.

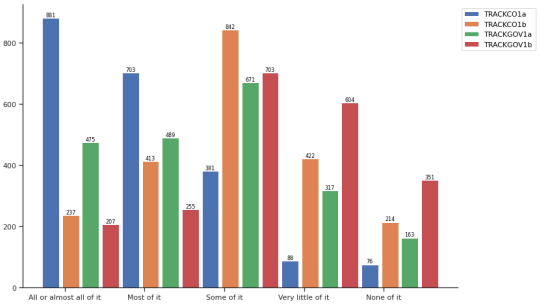


Figure 8: Vertical grouped bar chart

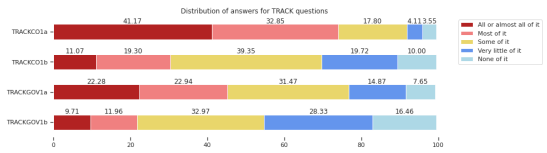


Figure 9: Horizontal stacked bar chart

As we observe from the above graphs, it is intuitively easier to interpret and understand the stacked horizontal bar charts as opposed to the other methods of visualization.

While exploring methods to visualize the responses to the CONTROL, CONCERN, PP and UNDERSTAND questions, we realized that likert-plots are better than horizontal stacked bar charts. Likert-plots are effective in visualizing the overall sentiment of the responses to the questions. For example, in the

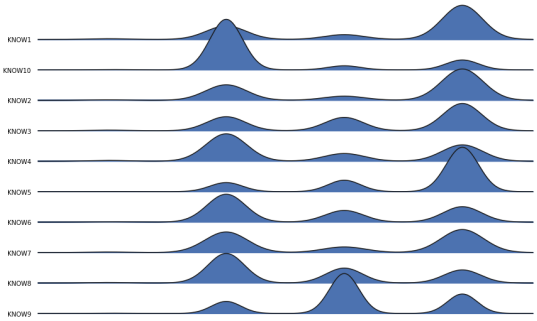


Figure 10: Joy Plot

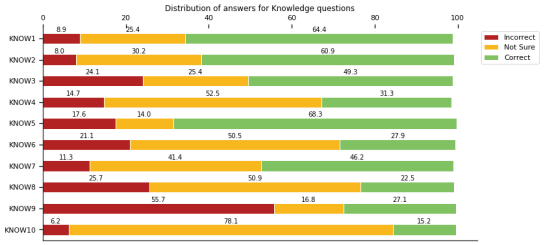


Figure 11: Horizontal bar chart

following graph, we can interpret that the respondents feel they have "very little" to "No" (62%) understanding about the privacy policies they read on the internet as opposed to (34%) feeling they have "some" understanding.

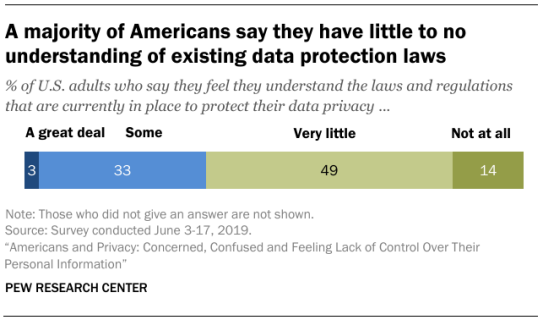


Figure 12: Stacked bar chart

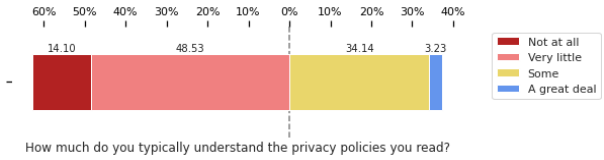


Figure 13: Likert chart

8.5 Visualization Process

All the processing, manipulation and visualization of the data was done using `pandas` and `numpy` packages, `matplotlib` and `seaborn` are used to plot the graphs, the seaborn package was useful in making the graphs look neater by removing the top bar and right bar that enclose the graph in a box. We also used the statistic packages named `statsmodel`[10] and `researchpy` to perform the t-test and F-tests along with summary statistics.

We started by exploring the demographics of the survey respondents. For each of the demographic variables, we plot a bar chart, a boxplot and perfrom t-test and F-test. We experimented with absolute and percentage values for the bar-labels in the bar chart, along with the color-palette and decided to use percentage values on top of the bars and a red-grey color theme for the bars to highlight the maximum value and the perceptually uniform `viridis` palette for the boxplots.

Visualizing the responses to the survey questions was the challenging, we explored scatter-plots of two kinds (Categorical Vs Quantitative and Categorical Vs Categorical[7]) and mosaic plots. None of which were as effective and clear at interpreting than the boxplots. The scatter plots were not effective since our range was small and all the points tend to be distributed fairly uniformly and the mosaic used area as its encoding method which is not easy to differentiate between the various categories.

Since the dataset contained categorical data, we had to convert the categorical data from string data type to ordinal data in integer data type. This was done to create the stacked and grouped bar charts and perform the t-test and F-test using the python `statsmodel` library.

All the visualizations follow the following rules :

1. The visual encodings used, bar charts and boxplots, rank higher in the effectiveness ranking.
2. The colors chosen are grey-scale friendly.
3. The graphics do not attempt to mislead the reader.

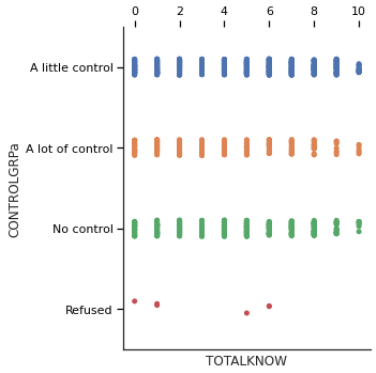


Figure 14: Scatter plot

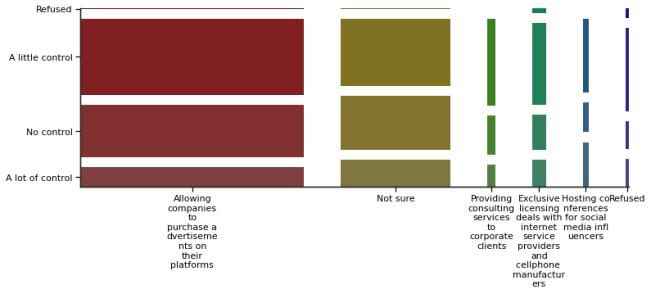


Figure 15: Box plot

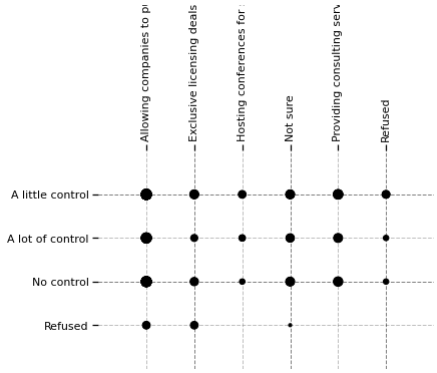


Figure 16: 2D Categorical scatter plot

8.6 Learning Outcomes

Our core learning from this project is that, there is no standard method of visualizing a particular form of data. We start off the visualization task by implementing the standard methods like bar graphs and scatter plots. Based on the nature of the data, we use augmented visualizations to better tell the story we want to tell.

The improvement in visual effectiveness is evident when choosing horizontal stacked bar graphs over vertical grouped bar graphs. However, horizontal bar graphs were not so effective while visualizing the overall sentiment of the answers, so we chose likert-plots.

Another learning outcome is that we now have better understanding of the `pandas` library and various methods to manipulation methods. Importing SPSS data as dataframe in `pandas` offers a lot more power and convenience in our efforts to complete this project.

8.7 Limitations and Future Work

There are lots of other variables for which we weren't able to explore in this paper. For example, speaking of privacy protection, what is the context that people are more likely to share their personal information? There are three questions asked about this, one of them is "Do you belong to any loyalty programs of a grocery store or a supermarket that you frequent?" Another one is "Have you ever used a mail-in DNA testing service from a company such as AncestryDNA or 23andMe?" Another is "Do you regularly wear a smart watch or a wearable fitness tracker?"

Obviously from sharing personal information with grocery store and sharing DNA information with companies doing DNA tests are two different levels of sharing. It'd be interesting to see to what degree people are willing to trade their personal information for convenience ranging from promotions in grocery store to checking information on their ancestors. If we have more time, it'd be our next goal to analyze

different privacy sharing activities and their digital literacy.

Another limitation is that the test for digital literacy in this survey is made of 10 questions that do not have a clear structure. The validity of the questions should be tested. It seems these are random questions and the answer can change at anytime. For example, just before this article was written, Twitter CEO Jack Dorsey resigned from his job, which means KNOW10 in the test will have to change it's answer if the question is to be reused in a future test.

It's a great challenge to test the concept digital literacy. Is it the same with coding literacy? Should everyone in our society be aware of data structure and algorithms in order to be an informed citizen and protect their privacy effectively? Will computer science be like English 101 in a college curriculum so every college student is required to learn coding? We look forward to revisiting this problem in the near future.

References

[1] Miriam Bartsch and Tobias Dienlin. "Control your Facebook: An analysis of online privacy literacy". In: *Computers in Human Behavior* 56 (2015), pp. 147–154.

[2] Pew Research Center. *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information*. https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/. Nov. 2019.

[3] Larry Cuban. *Public Schools shouldn't cater to CEO's needs*. https://www.wsj.com/articles/should-all-children-learn-to-code-by-the-end-of-high-school-11582513441. pp. R1, R2. Feb. 2020.

[4] Andres Fernandez. *Grouped Bar Plot*. https://stackoverflow.com/a/69170270/6826264. Sept. 2021.

[5] Alison J. Head, Barbara Fister, and Margy Macmillan. "Information literacy in the age of algorithms: Student experiences with news and information and the need for change". In: 2 (2020). Project Information Literacy., pp. 17, 29. URL: https://projectinfolit.org/pubs/algorithm-study/pil_algorithm-study_2020-01-15.pdf.

[6] P. Hitlin and L. Rainie. *Facebook algorithms and personal data*. Tech. rep. Pew Research Center, Jan. 2019. URL: https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/.

[7] Myriam. *Visualize Categorical Relationships With Catscatter*. 2020. URL: https://en.wikipedia.org/wiki/Likert_scale.

[8] L. Rainie and J. Anderson. *Code-dependent: Pros and cons of the algorithm age*. Tech. rep. Pew Research Center, Feb. 2017. URL: https://www.pewresearch.org/internet/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/.

[9] L. Rainie, J. Anderson, and A. Luchsinger. *Artificial Intelligence and the Future of Humans*. Tech. rep. Pew Research Center, Dec. 2018, p. 18. URL: https://www.pewresearch.org/internet/2018/12/10/artificial-intelligence-and-the-future-of-humans/.

[10] Skipper Seabold and Josef Perktold. "statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[11] Robert Sedgewick. *Coding fosters creative and logical thinking*. https://www.wsj.com/articles/should-all-children-learn-to-code-by-the-end-of-high-school-11582513441. pp. R1, R2. Jan. 2020.

[12] Leonardo Taccari. *Joyplot python library*. 2021. URL: https://github.com/leotac/joypy.

[13] Claus Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 1st ed. O'Reilly Media, 2019.

[14] Bang Wong. "Avoiding Color". In: *Nature Methods* 8 (2011), p. 525.