Final Project Analysis Report

Dennis F. Perez

COGS 212

December 14, 2021

Introduction

Main Study

The main point of this analysis is to examine the current results of an ongoing study entitled: “The Influence of Media Sources: How the Media Affects Judgement,” by researchers April W. Westmark, Ayonna A. Jones, Stephanie K. Gamino, and myself Dennis F. Perez. Researchers in the study mentioned above note how communication is a primary tool that has advanced and hindered our growth as a society. They discuss how the dissemination of information can influence the masses and ultimately affect what people believe and how they behave with one another. Researchers pointed out that how information is spread has resulted in a general sense of divisiveness within this country, which has at times resulted in killings, riots, and the imbalance of power within some groups of society.

Naturally, the researchers felt that the quality of information passed around was a key contributor to the tension between groups in this country. Thus, researchers sought to create a project that could further elucidate how larger groups use bias when disseminating information to smaller groups to influence the way they think and act. Their main initiative was to explore peoples’ knowledge of this phenomenon. They did so by deploying a survey to gauge participants’ ability to detect bias in the information they come across (Westmark et al., 2021).

Westmark et al. focused on how news media outlets use biased tactics, like framing and priming, to influence their viewers. They make use of Robert Entman’s article “Framing Bias: Media in the Distribution of Power”, to describe framing and priming as follows: “fram[ing] introduce[s] or raise[s] the salience or apparent importance of certain ideas, activating schemas that encourage target audiences to think, feel, and decide in a particular way. Priming, then, is a name for the goal, the intended effect, of strategic actors’ framing activities” (2007, p. 164-5). In

order to exemplify this phenomenon, researchers utilized a survey where participants were shown 12 news article headlines that contained a mix of low, medium, and high levels of bias. The articles covered the general topic of immigration and the current event of the Kyle Rittenhouse trial. Half of the news articles were from 'liberal' type news sources, and the other half were from 'conservative' type news sources. Both types covered immigration and the trial (Westmark et al., 2021).

Researchers then chose a random subset of the participant group to receive a treatment in the form of a video. The video explained a general definition of bias and how it is present at both the individual and group levels. The primary goal of the research was to explore how many participants could correctly identify each news article with its correlated level of bias. Researchers planned to compare the results of those who received the treatment vs. those who did not receive the treatment. They predicted that those who received the treatment would be able to detect bias better than those who did not (Westmark et al., 2021).

However, as mentioned previously, the study is still ongoing. Researchers anticipate that they will need to look beyond their original research question to gain a more well-rounded understanding of the general topic they are exploring. Thus, they included several variables that are not currently being used but may be used in the future to aid the study in different considerations. For example, participants were asked to provide information, such as ethnicity, party affiliation, gender, amount of time spent on social media, amount of time spent watching cable television and news media outlets, how often they fact check information, and so forth (Westmark et al., 2021).

Current Analysis

Accordingly, I felt that exploring their current results and examining correlations they had not considered could aid further research endeavors. Thus, the current analysis will take several different approaches compared to the initial study when examining the data. The main observation will focus not on how well one detects bias but how one deals with information. Moreover, in order to gain a general insight on how people deal with information in general, I developed three pertinent aspects to consider:

1. One should consider where people receive their information, whether that be from literature, social media, news outlets on cable television, and so on.
2. One should examine the quality of information people are receiving and passing on. Namely, do these people fact-check the information they come across before sharing it?
3. One should consider with whom these people are sharing their information. Do they share information with the immediate community members, such as family and friends?

These questions are important because they can tell us how information is being passed around at the lower level. Examining individuals and their immediate social groups may provide meaningful information as to how communities or societies function as a whole. In essence, combining lower-level group explorations with how information is spread at higher-level groups like organizations could provide the researchers of the initial study a well-rounded perspective of how information is shared.

Thus, the current analysis will focus on basic examinations of several variables within the initial study data that seemingly reflect the three considerations needed to understand how one deals with information. In particular, this analysis will consider gender differences in where each group receives a majority of their information, if they fact check the information they come across, and if they discuss current events with their immediate social group. However, since the

study is still running, the dataset is small. Therefore, it should be noted that this analysis is not meant to provide significant findings. This report is meant for general exploratory purposes and to possibly help guide researchers in new directions.

Methods

Data Analysis

As mentioned in the introduction, this analysis examines the current most results of the survey study by Westmark et al. This analysis is also loosely structured on the steps from chapter four of *The Art of Data Science: A Guide for Anyone Who Works with Data* by Roger D. Peng and Elizabeth Matsui. Accordingly, the following information will walk the reader through each section of the code entitled '----Section:. . .----'. The reader may refer to the file 'main_program.Rmd,' as a reference when reading through each section.

----*Section: Formulating the question*----

Upon examining the data, I formulated my question based on my initial interest in exploring how information is handled at the lower group levels. Mainly, I wanted to find a subset of the data that answered the three questions concerning how information is handled: where does the information come from, is the information checked before being passed on, and are current events being discussed within the immediate group.

Luckily, I found subsets of data aligned well with my initial question. For example, the differences in gender groups (gender column) provided a small enough subset that contained relevant responses to each question for each participant. The 'social_amt_1' through 'social_amt_7' columns contain the amount of time each participant spends on social media sites like Facebook, Instagram, Twitter, TikTok, Parler, Reddit, and YouTube. The logic behind observing this information was that the amount of time spent on social media would possibly

indicate that social media was the primary source of information they are exposed to. The ‘fact_online’ column contained how likely each participant was to fact check the information they came across; the responses ranged from extremely, very, somewhat, a little, and not at all. This information would enable me to determine the quality of the information being received and passed along by each participant. Finally, the ‘discuss’ column contained whether or not each participant discusses current events/topics with their friends and family. This information would provide insight into whether each participant shared the information they came across with their immediate social group. In theory, each of the observations mentioned above could describe how each individual shares information in general. The idea behind this is that if we understand how individuals deal with information at this lower level, we can combine this information with higher-level examinations to improve the sharing of information in general. This may have larger positive implications, but the paper will cover those in later sections.

----*Section: Installing and loading packages*----

In this code section, I installed and loaded libraries that I thought would be necessary for my analysis. For instance, I loaded the package *tidyverse* since it is highly versatile and contains many helpful packages already within it. The packages contained in the *tidyverse* library are *ggplot2, dplyr, tidyr, readr, purr, tibble, stringr,* and *forcats*. These packages are highly useful in data analysis since they allow programmers to perform essential functions like reading in CSV files, working with categorical variables, applying changes to the data via tibbles so that original information is preserved, and so on. I did use several other packages out of personal preference, but that was mainly due to my initial lack of knowledge of some functions. Hence, the reader will see some of the libraries that were commented out, which occurred after I went back through and re-examined the code.

----*Section: Reading in the data*----

The section displays how I created a variable to read in the dataset. I also created a copy of the original dataset to ensure that the original information of the study would be preserved, which is sometimes referred to as immutability. Of course, there are more intuitive ways to preserve original data, but I performed data immutability the best way I knew how with my given level of programming experience. In later sections, you will also notice that I use a different variation of the initial variable I created to read the dataset. This was another attempt to preserve the originality of the data, which was also done so that the reader could easily make distinctions between my manipulated dataset and the original.

----*Section: Looking at the data*----

In order to get a better view of the dataset, I used various functions to print out the data. The reader may notice that each function used seems redundant, being that some of them essentially display similar information. However, I wanted to show the reader the process of each function I tried. I also did this to understand better how information was displayed with each one. For example, I found the skimr() function to be very insightful, as it displayed variable class types, the number of variables, the amount of missing data, and much more. In addition, the information gained from each of these functions made it easier to segue into the upcoming sections.

----*Section: Data cleaning*----

The analyzed dataset was very small, so there was not much data cleaning to do. Furthermore, I was advised not to filter out too much of the data since there was little to consider. Thus, although relatively short, this part of the code caused a great deal of mental struggles.

Accordingly, the methods used in this section were used to clean the data that is not considered relevant to the observations being made in this analysis. After lengthy discussions with my colleagues, I decided to remove specific rows of information from the copied dataset. For example, I removed rows one and two of the dataset because they had information normally inaccessible to those outside the research team. The first row contained an arbitrary "import id," which is a number generated by Qualtrics when downloading the data. This row was deemed useless because the data associated with it stayed in the correlated column. Also, when the study finishes, it is unlikely they will need the import id. However, that information was preserved in the original dataset that remains in this repository if they do. The second row contained the questions that were asked of participants. This information was removed because each question is preserved in the original dataset, and each of the variables and their associated questions is explained in this written report.

Moreover, rows 104- 110 were removed because they contained information for participants who did not complete the survey, and hence, their information was not collected. Ultimately, each of the removed rows skewed my analysis and was unnecessary. Nevertheless, if one wants to see that information, it can be found in the original dataset within the repository.

----*Section: Data analyzations (part 1)*----

After cleaning the copied dataset, I began to conduct the analysis. The first question I sought to answer was how much time each individual spent on social media. I wanted to know how much time was spent on social media to understand better where the participants were most likely being exposed to information.

First, I started by creating a copy of the original variable used to read in the copied dataset. I did this to avoid altering the data any more than I had to. I also found it helpful to

create different variables of the same data so that multiple manipulations and visualizations could be made with the information specific to this part of the question I was trying to answer.

Next, I decided to utilize the mutate function on the newly created dataset variable. I did this to recode the data within the columns social_amt1 through social_amt7 as numeric rather than the original character types. I found this to be an intuitive way to take the sum of each participant's total hours spent on social media. After creating a variable to take the sum of each row, I began working on creating tables. I performed actions such as splitting the data and creating a new data frame with a new column that was to be associated with the split data. I then created a variable that further cleaned the copy of the dataset and grouped relevant columns to use dplyr functions that enabled me to summarize and count the data. After viewing the newly created variable, I hardcoded changes to the names of the columns with names I felt were more descriptive of what the data was showing. Lastly, I created a sorted bar plot that displayed all of the information gained from the previous steps. The visualization and the explanation will be discussed in the 'findings' section of this report at the end.

----*Section: Data analyzations (part 2)*----

The purpose of this section was to answer how likely the individuals of each gender fact-checked the information they come across. This was one of the sections I was interested in most since it seems to indicate how well each participant deals with information, or more specifically, how each participant determines what is credible to believe and disseminate to others. Therefore, fact-checking information could be an essential tool for ensuring the passing of quality information.

Like the previous section, I created several variables to recode the character type information of the columns 'fact_online' and 'gender' to be converted into numeric data. The

various response types within this column were coded with numeric values ranging from zero to four. These numbers did not hold specific significance themselves, but recoding these variables allowed me to conduct a correlation test between the ‘fact_online’ and ‘gender columns of data. The results of this test were plotted using a box plot, which will also be discussed in the ‘findings’ section.

----*Section: Data analyzations (part 3)*----

This last section aimed to describe whether or not individuals of each gender discussed current events with people in their immediate social communities. This information was explored with the notion that those who spent less time on social media and fact-checked the information they came across would be more likely to discuss current events with their immediate social circles. There is no particular significance to this assumption aside from the subjective opinion that those who know more about current events are likely to discuss them more.

As for the coding process, I started by creating a new variable to house the initial copied dataset. I then manipulated the data similarly to part one so that the data within the columns of ‘discuss’ and ‘gender’ could be paired and compared with one another. After, I renamed the columns of the newly created table to be more descriptive of the observation being made in this section. Lastly, I utilized plotly to coerce the plot produced by ggplot. Again, there was no particular reason aside from the aesthetics and usability provided by the plotly function.

Findings

Limitations of the Analysis

First and foremost, it should be noted that the information provided by this analysis may not be representative of the data as a whole. As a result, the results of each section may not bear any significance. Furthermore, this analysis was purely exploratory and is not meant to draw

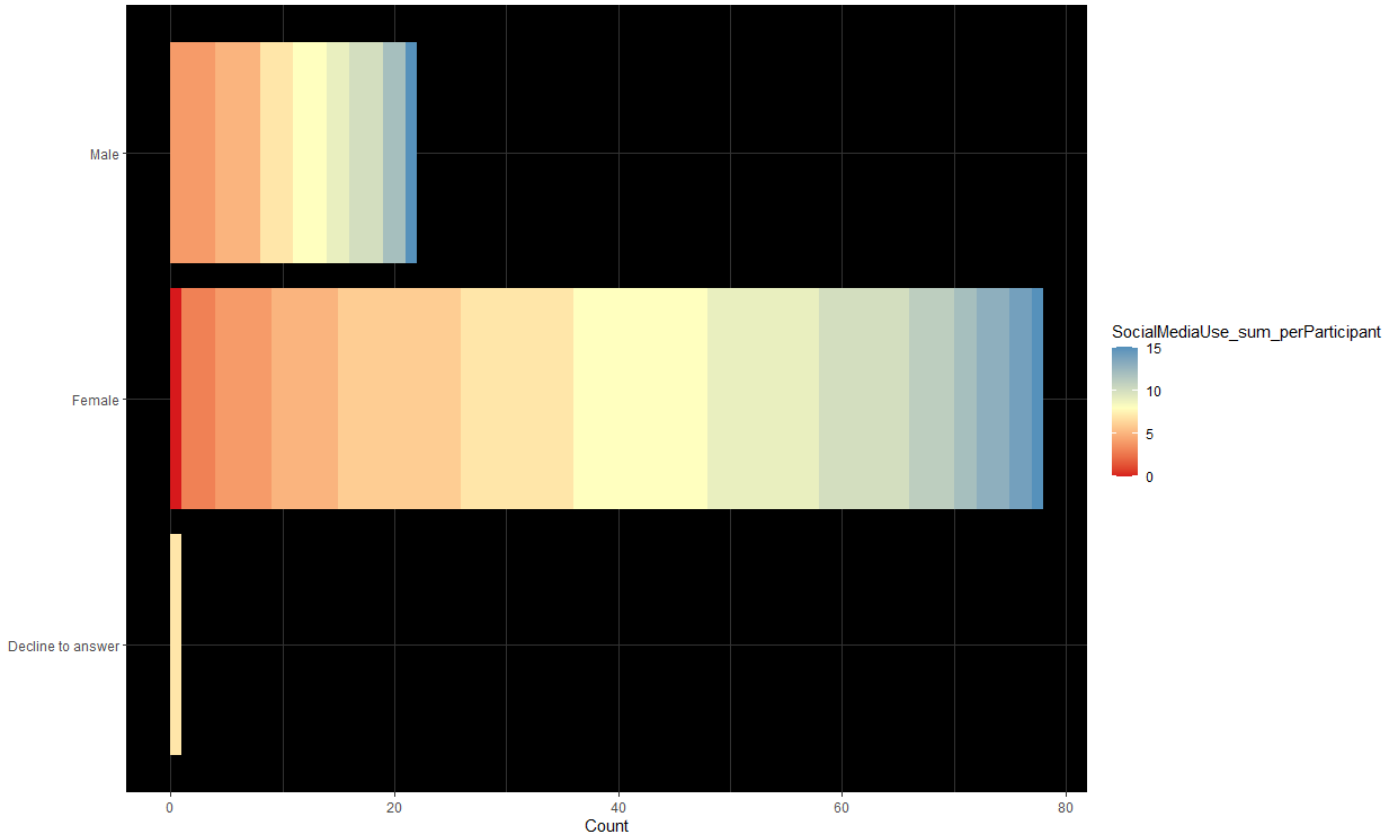
conclusions about the data. In fact, there were several limitations of this analysis that can affect how applicable its results may be.

Naturally, the biggest limitation was the size of the dataset itself. A lack of data may cause analyses results to be skewed in one direction or the other. Furthermore, since I used an even smaller subset of the data, the amount of information being examined was truncated even more, which could again reduce the significance of the results. Moreover, I could not remove arbitrary values like “-99” found in categorical type responses. This could have also interfered with how the data was interpreted. Lastly, the number of female participants outweighed the number of male participants tremendously, which may be a key contributor to the distributions in the upcoming data visualizations.

Data Visualizations

Now that the study's main limitations have been covered, I will describe the plots generated within parts one, two, and three of the data analyzations section of the code. Although these visualizations may display significant information or be the best type of plot, they may provide some low-level informative observations of relationships between variables of the original study.

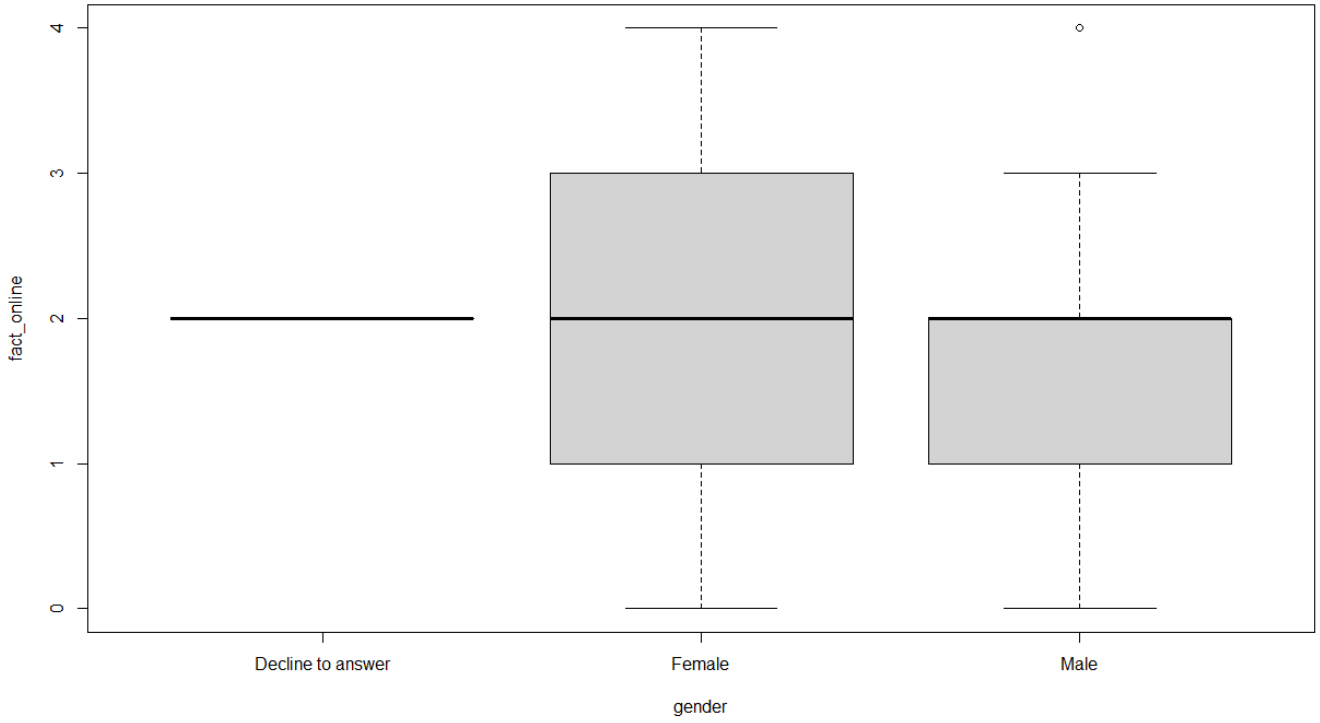
Plot 1



Plot one is a sorted bar chart that displays the gender types of 'Male,' 'Female,' and 'Decline to answer' on the y-axis and the number of participants in each gender type on the x-axis. The legend on the right-hand side describes the gradient of colors meant to display the number of hours spent on social media for each participant. I decided to use a sorted bar chart because I felt it would be an intuitive and aesthetically pleasing way to display the cumulative hours spent on social media by each group.

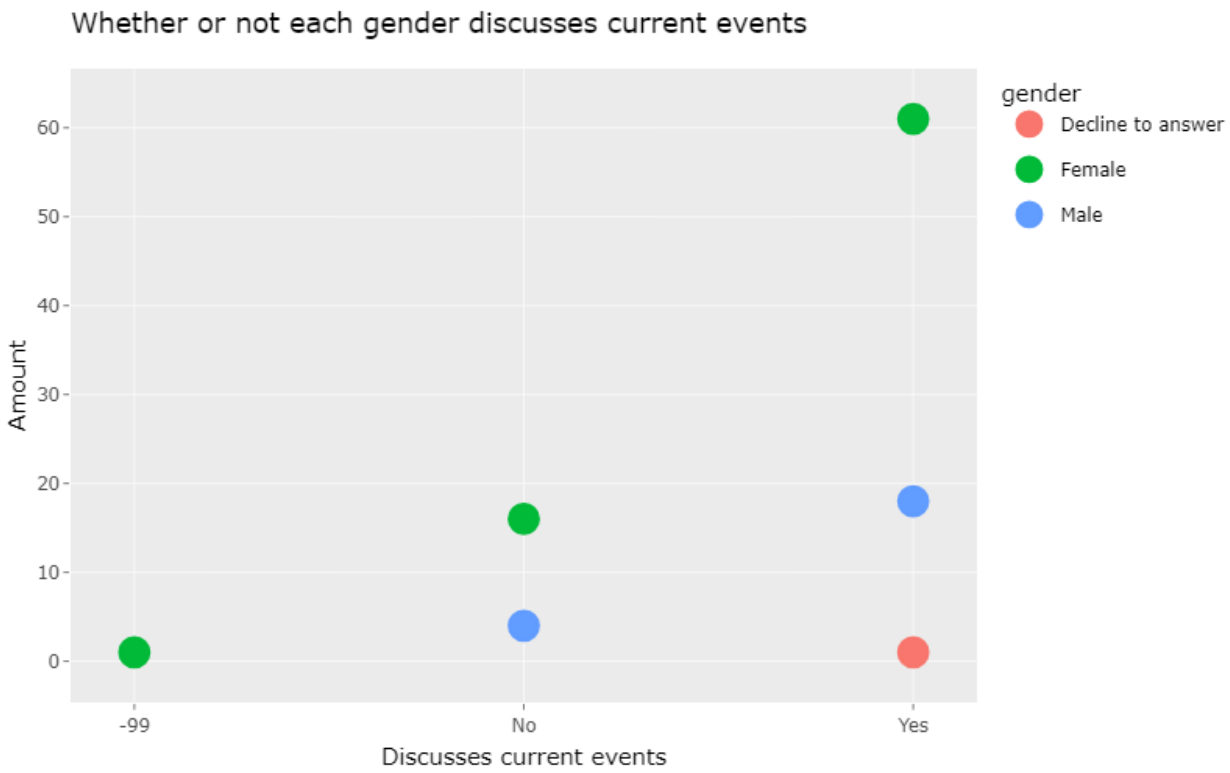
Given the number of higher-gradient colors in each bar, it seems as though females spend the most time on social media. However, since the number of females greatly outweighed the number of male participants and those who declined to answer, these distributions may not represent the question it was attempting to answer. Lastly, there may have been a better way to measure the number of hours spent on social media. It is unclear whether this information would be a reliable indicator of social media as a primary source of information for each participant.

Plot 2



Plot two displays a box plot with the likelihood that each participant fact-checked information they came across on the y-axis and the gender type on the x-axis. Although this plot did not produce a significant correlation in terms of a box plot, it seemed to be the most informative of the heavily weighted data for each gender. Regardless of the difference in group size, it seems as though there is more variation in the female group, which could indicate they are more likely to fact-check their sources.

Plot 3



Plot three displays the number of participants for each answer type on the y-axis and whether or not each participant discussed current events on the x-axis. The legend on the right-hand side displays the color associated with each gender so the viewer can easily see the type of answer for the question and the number of participants who answered that way.

At first glance, I thought this may have been the easiest graph to read. However, since this plot is in picture form, it loses the interactive feature that displays the number of participants when hovering over each dot. Thus, it may not be as readable as I anticipated.

Furthermore, I was unable to get rid of the arbitrary "-99" value, which may also confuse the viewer. Also, since only one participant declined to answer about their gender, there are no results for those who chose not to discuss current events. Consequently, it may be difficult to

correlate with the information from the last two plots, which would make it challenging to use in the general understanding of how each participant deals with information.

Discussion

As one can see, the current analysis had many limitations and did not produce significant results. However, both the report and code were written so that others may have a detailed explanation of the entire process of an exploratory data analysis, which I believe is an important aspect of research in general. Furthermore, the dataset may have been small, but the analyses made on the variables within this study could guide further explorations. At the very least, this analysis and the questions it attempted to answer can be used as a template of where to start or where not to start when considering alternative examinations of bias phenomena and how it affects how information is distributed.

Bibliography

Entman, R. M. (2008). *Framing bias: Media in the distribution of power*. Journal of Communication, 57(1), 163-173. https://doi.org/10.1111/j1460-2466.2006.00336.x

Peng, Roger D., and Elizabeth Matsui. 2016. *The Art of Data Science: A Guide for Anyone Who Works with Data*. Leanpub.

Westmark, A., et al. (2021). *The Influence of Media Sources: How the Media Affects Judgement*. University of California, Merced