Paper 2*

Gavin Crooks Samarth Rajani

February 14, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

1 Introduction

The introduction is self-contained and tells a reader everything they need to know including: 1) broader context to motivate; 2) some detail about what the paper is about; 3) a clear gap that needs to be filled; 4) what was done; 5) what was found; 6) why it is important; 7) the structure of the paper. A reader should be able to read only the introduction and know what was done, why, and what was found. Likely 3 or 4 paragraphs, or 10 per cent of total.

2 Data

2.1 Data Introduction

The data used in this paper is from (cite og paper). The data used in the paper is extremely complicated as it combines numerous data sets to complete their analysis. In this paper, we wanted to simplify the data to determine if their paper had underlying biases within the data. To do this we focused on one of their eleven data sets, baseline dataset as it was the most comprehensive. This data was collected through a facebook ad. The respondents answered a number of questions including questions about income, ethnicity, family, political beliefs and political following. THis is a very useful data set outside the scope of the original paper as it can be analyzed to answer a number of questions. The data set was cleaned to focus on respondents income, race and how closely they follow politics. This provided a data set

*Code and data are available at: LINK.https://github.com/Crooksyyy/The-Effects-of-Social-Media

of approximately 6000 complete responses after removing unfinished responses. This was a substantial decrease from the orignal 24000 responses in the data.

2.2 Income Data

The variable within the data was the income variable. The questionnaire included a categorical value for the the household income of respondents. The data can be visualized in (**figure1?**). This graph illustrates that the least number of households make greater than 100,000 USD, below 20,000USD or preferred not to answer. This graph also shows what the categorical option were for the respondents to the questionnaire. (**figure1?**) closely resembles the expected distribution of USA household income. As expected in any income distribution the majority of responses fall within the average income ranges of the USA, between 20,000USD and 100,000USD. These factors indicate that the data set has an accurate representation of household income.

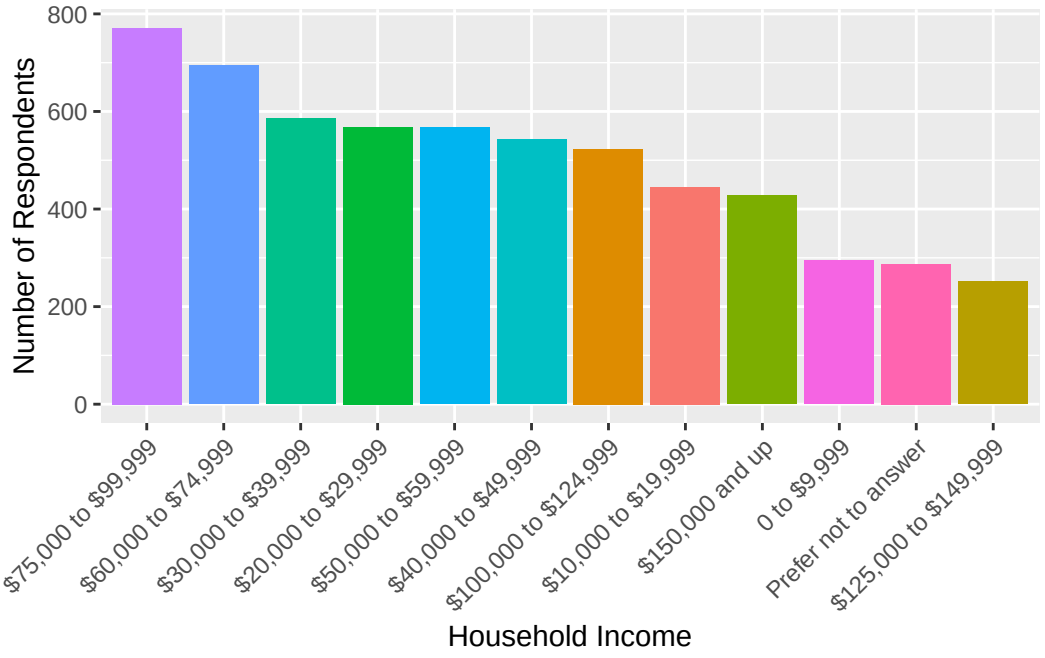


Figure 1: Distribution of Income from Responses in a facebook ad

2.3 Ethnicity Data

The second question of the data that we have included in our analysis is the ethnicity of the respondents. This again is a categorical variable, that the individual self identifies their own ethnicity. The response options included Asian or Pacific Islander, White / Caucasian,

Hispanic, Black or African American and other. (**race_dist?**) shows the percentage of respondents in each with the overwhelming majority of responses being Caucasian at nearly 70%. This is actually less than the most recent estimates by the United States government which estimate over 75% of the population is Caucasian (cite us gov). The data also has an over representation of Asian and Native Americans. This results in an under representation of Hispanic and African American populations.

https://www.census.gov/quickfacts/fact/table/US/PST045222

Table 1: Percentage of each Ethnicity from Responses in a facebook ad

Ethnicity	Percentage of Responses
American Indian or Alaskan Native	0.7554138
Asian or Pacific Islander	13.5806614
Black or African American	6.0936713
Hispanic	8.0577472
Other (please specify)	2.5851939
White / Caucasian	68.9273124

2.4 Politics Data

The third variable in the data that is included in our analysis is a variable of respondents self identifying how closely they follow politics. This is another categorical variable measured as Not at all closely, Somewhat closely, Rather closely and Very closely. This variable faces many problems as this categorical scale is not consistent across respondents. To be specific we mean someone who identifies as someone who does follows Not at all closely can be following politics more than someone who identifies as Somewhat closely. This is a measurement issue within to the questions asked in the survey and all self identifying variables in general. (**figure2?**) shows the quantity of respondents in each group. The most common response is that they follow somewhat closely and the other responses are relatively even.

2.5 Missing Data

```
# Parts of this code from R-charts
#https://r-charts.com/part-whole/stacked-bar-chart-ggplot2/
ggplot(cleaned_data, aes(x = follow_trump, fill = race)) +
  geom_bar(position = 'stack', color = 'black', stat = 'count') +
  labs(title = "Number of People Following Trump by Race",
       x = "Follows Trump",
       y = "Count") +
```

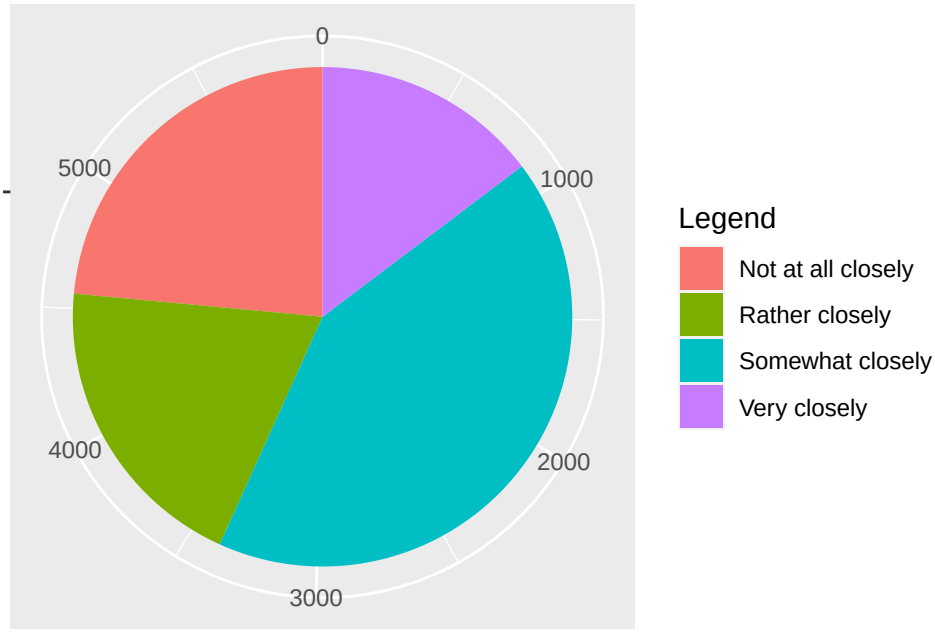
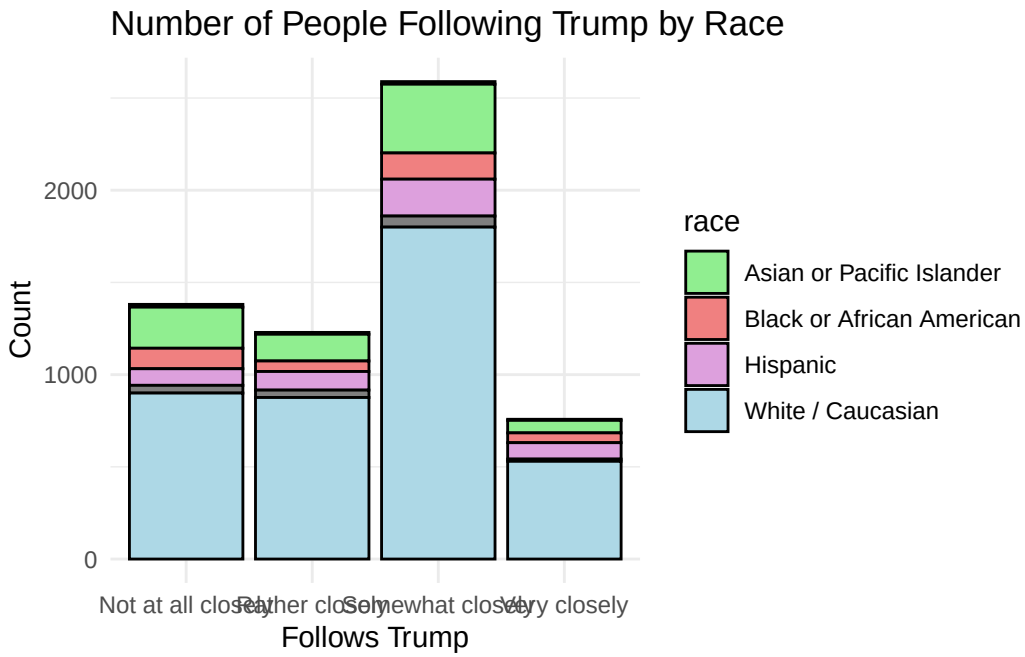


Figure 2: How Closely People Follow Politics from Responses in a facebook ad

```
  scale_fill_manual(values = c('White / Caucasian' = 'lightblue', 'Black or African American
                               'Other  (please specify)' = 'lightgoldenrodyellow')) +
  theme_minimal()
```



3 Model

simple regression of income and politics data controlling for race

4 Discussion

like to incluse variables like state etc data set too small/ incomplete for that ## First discussion point {#sec-first-point}

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

4.1 Second discussion point

4.2 Third discussion point

4.3 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

References

Friendly, Michael, Chris Dalzell, Martin Monkman, and Dennis Murphy. 2020. *Lahman: Sean "Lahman" Baseball Database*. https://CRAN.R-project.org/package=Lahman.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.