In-class Lab II Report

Team members: Mahsa Aghazadeh, Hong Youjin, Zahra Bayramli, Yungeun Song

The purpose of this exercise is to compare the results of non-probability sample and probability sample. We already have the results of the probability sample (https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use) and we want to know how much our non-probability sample results are close to the probability sample.
So, our team conducted a survey on social media usage. We designed a Google Form, deployed it via Prolific, and included attention checks. After collecting and cleaning the data, we analyzed it using Python, comparing our results with Pew Research Center's probability sample.

1. Sampling method

In this report, results from a non-probability sample and probability sample have been compared. Existing results of probability sample from a Pew Research Center survey of 5,733 U.S. adults conducted May 19-Sept. 5, 2023. In this assignment a non-probability sampling was conducted through a survey with 35 people with the Prolific website.

2. Analysis

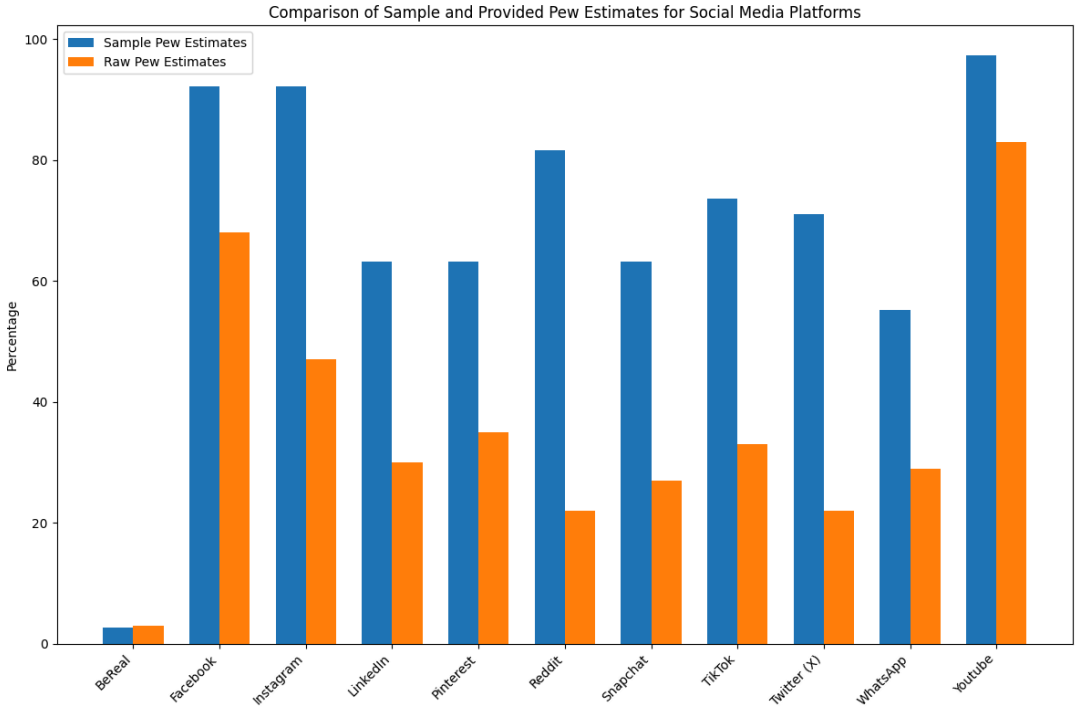

Figure 1. Comparison of Sample (our) and provided PEW estimates

In Figure 1, a visual comparison of our sample Pew estimates for various social media platforms with those provided by the Pew Research Center was provided, highlighting significant disparities in perceived usage rates. When comparing our sample Pew estimates with those provided by the Pew Research Center, noticeable disparities emerge across various social media platforms. For instance, our sample Pew estimate for WhatsApp stands at 55.26%, significantly higher than the Pew Research Center's estimate of 29%. Similarly, our estimate for Pinterest is notably higher at 63.16% compared to Pew's 35%. TikTok also shows a substantial variance with our estimate of 73.68% contrasting starkly with Pew's 33%. Instagram's estimate from our sample data is considerably higher at 92.11%, while Pew's is 47%. Furthermore, our sample estimates for Snapchat, Facebook, LinkedIn, Youtube, Twitter (X), Reddit, and BeReal all demonstrate notable discrepancies compared to Pew's estimates, underscoring the variability in perceptions of social media usage between our sample and Pew's broader research. In conclusion, our collected data which used a non-probability sample is overestimating the usage of social media.

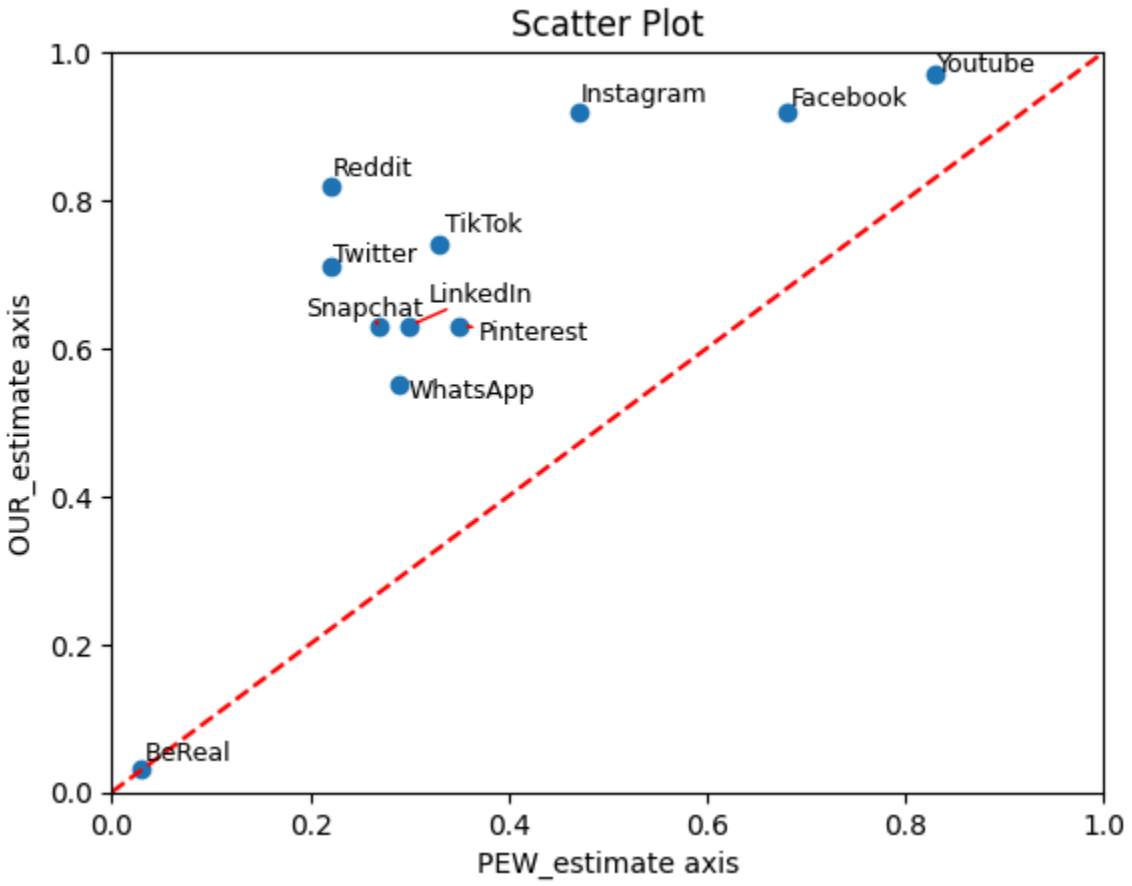

Figure 2. Scatter plot of sample (our) and provided PEW estimates

Scatter plots are utilized to examine the relationships between variables. In this plot (Figure 2), we leverage a scatter plot to compare our survey data against the Pew Research Center's estimates. It's observable that respondents from our Prolific study used the platforms more

frequently than those identified by the Pew Research Center. And by comparing this scatter plot to the y=x line reveals that the relationship depicted is not linear.

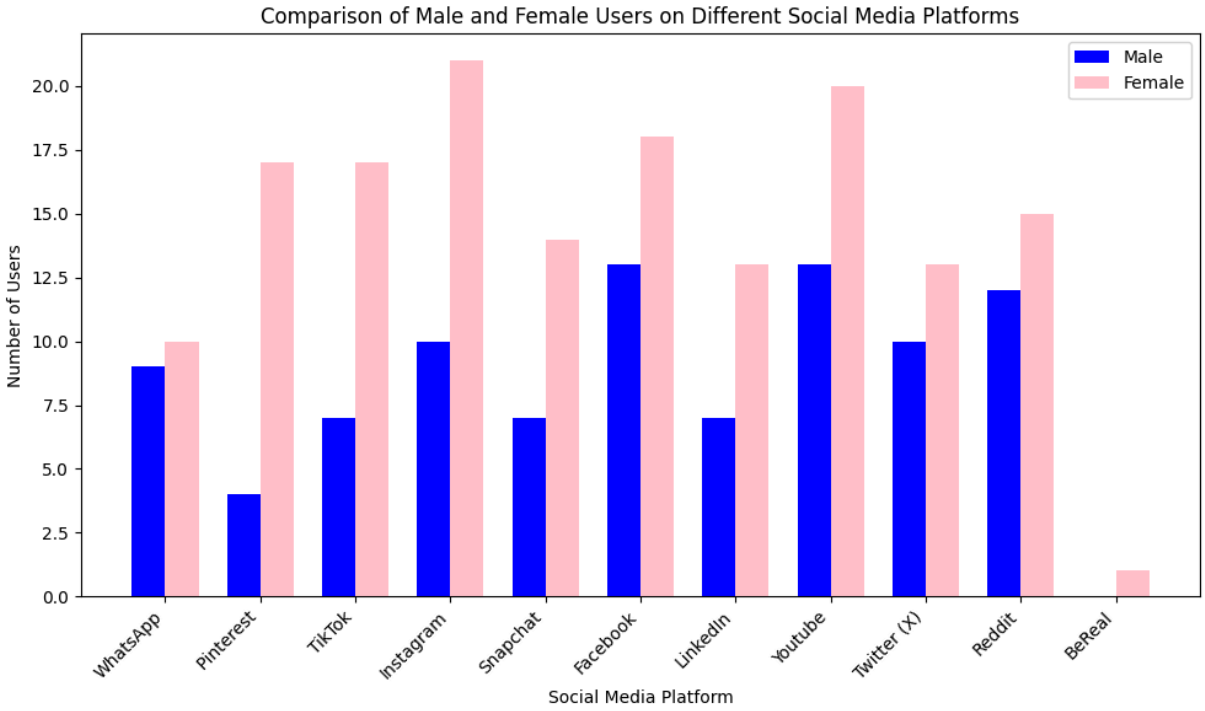

Figure 3. Gender distribution of participants for each social media platform

This graph displays the gender distribution of users across various platforms, indicating a higher proportion of female users compared to male users. However, it's important to consider the original composition of respondents' gender, with females constituting 61% and males 38%, when interpreting these results. According to the Pew Research website(probability sample), in the case of TikTok, the usage rate of women is higher than that of men using the platform. And we have more female participants(21) than male(14), the difference with proportion of sex may have heightened the differences in results.

3. Discussion

The data collected through the non-probability sampling method is highly biased in pew estimation compared to existing work. One of the possible reasons can be, the fact that people who answered this survey knew about Prolific website which means they are more familiar with websites and possibly social media (i.e they are web smart people). Also, the ratio of female to male respondents is 1.5 which further biases our findings. Moreover, from an economic

standpoint, it can be said that individuals with lower income levels are more likely to participate in surveys on the Prolific site compared to those with higher incomes.

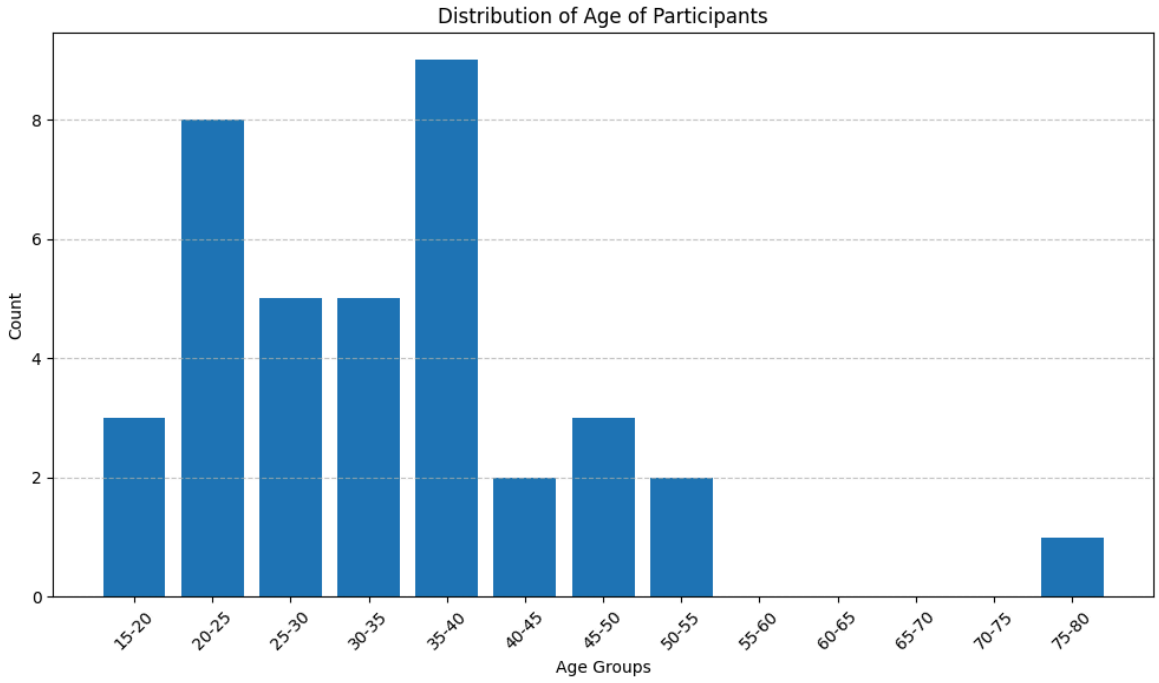

Figure 4. Age distribution of participants

The distribution of age groups among our participants might also be a potential contributing factor to the observed disparities in our sample Pew estimates compared to those provided by the Pew Research Center. As depicted in the bar graph (Figure 4), the majority of our participants, comprising 27 out of 35 individuals, fall within the age range of 0-40 years. This concentration of participants within a relatively younger demographic could have influenced our sample Pew estimates, as younger individuals might exhibit different social media usage patterns compared to older age groups. Consequently, the discrepancies between our sample estimates and Pew's estimates may, in part, be attributed to the age composition of our sample.

We confirmed through statistics that non-probability sampling does not may not accurately represent the entire population and replace probability sampling. When using non-probability sampling, it was necessary to understand what the subjects of the survey were and what their characteristics were. However, despite our biased results, the results were the same regarding which platforms were used the most and which platforms were used the least. Looking at these results, when the population is hard to be randomly sampled, the results can be skimmed with non-probability with relatively little time and effort (of course, special care must be taken regarding bias).