In this project component you will be working with the same dataset you have already been using in class: existing survey data

Links to an external site.

from Data & Society, a non-profit organization that “studies the social implications of data-centric technologies & automation”.

As you hopefully already know, the data you will work with focuses on inequities in security and privacy among U.S. internet users. As the full report

Links to an external site.

on the survey describes, “the nationally representative survey was fielded in November and December of 2015 among 3,000 American adults, including an oversample of adults with annual household incomes of less than $40,000. The survey provides new insights into the privacy and security experiences of low-socioeconomic status (low-SES) populations and aims to contribute to a deeper understanding of their technology-related behaviors and beliefs.”

Using the public data files available for the survey, you will engage in an analysis of inequities in people’s smartphone behavior to answer the following research questions:

- RQ1: Are there gender (RQ1a) or income (RQ1b) inequities in people’s use of smartphones as their primary means of accessing the Internet (`q4` in the dataset/survey questionnaire)?
- RQ2: Are there sociodemographic (income, gender, race, age, and/or education) inequities in people’s smartphone security & privacy behaviors (`SM1`)?
    - Alternative option: you can pitch your own question to answer with the dataset. The research question should be one that is possible to answer using regression analysis.
- [Extra Credit: 5pts] Pitch a third research question of your choice to answer using the provided dataset.

To answer these questions, complete the following steps:

Step 0: Download the data files

Links to an external site.

from Data & Society. Load the survey data CSV into BlueSkyStatistics (see section 2.3.1 of the BlueSky user guide

Links to an external site.

if needed).

Alternative: You can use R to analyze the data instead of BlueSky if you prefer.

Step 1 [3pts]: To answer RQ1 you will need to use one of the testing methods we discussed in class. You will need to modify the variables, as we did in class, to create variables relevant for your analysis. There is no one right way to do this analysis. Some hints are below.

Note that non-response to survey questions may be coded as '99', '98', or '9' depending on the question. You will need to recode these responses to `NA` in BlueSky so that they are analyzed correctly.

Hint: For RQ1b you can handle the income variable in many different ways. You can transform the `inc` variable into a binomial (boolean) by making a variable that is `true` if income is less than $40,000 but false otherwise; you could use income as an ordinal variable as it is already coded; you could treat income as a linear variable by taking the midpoint of each of the income categories; or you could choose to make income a categorical variable (e.g., convert it to three categories <50k, 50-100k, 100k+). Your choice will determine which analysis method you use

Links to an external site.

and what findings you can generate.

Step 2 [3pts]: To answer RQ2 use a regression analysis. You will need to modify the sociodemographic variables to have fewer categories (for example, you could examine inequities of women vs. non-women, income below or above 40K; people who have no college vs. some college education). Please use the `raceethn` variable for race and ethnicity. It is coded into 4 categories (1 is white, 2 is Black, 3 is Hispanic, 4 is Other, and 9 is those who refused the question or said they did not know).

There are many different ways you could approach this analysis. I encourage you to thoughtfully explore as if this was a real research problem in which you were engaged. For example, you can run four regressions, one for each of the sub questions in `sm1` or you can run one regression where the dependent variable is whether someone did at least 1 of the behaviors in the `sm1` question or where the

dependent variable is a count of the number of behaviors they reported doing in `sm1`. You can also add additional variables to your analysis, if you think this would better inform your understanding of inequity in privacy & security in the U.S. (e.g., you could control for whether people felt knowledgeable about privacy and security [`q12`] or whether they perceive themselves as having control over their private information [`q14`]).

Step 3 [4pts]: In 250 words or less, briefly justify your analysis choices (e.g., how you transformed your variables for each RQ and why you made the choices you did.

Step 4 [5pts]: In 400 words or less, report your results answering each research question; include statistical results (e.g., the p-value and other test statistics, a table of regression results) integrated into your narrative.

Here are a few examples (note these are longer than 400 words) of how to report statistical results along with your own commentary:

- http://varianceexplained.org/r/trump-tweets/
- Links to an external site.
-
- https://blogs.scientificamerican.com/observations/will-americans-be-willing-to-install-covid-19-tracking-apps/
- Links to an external site.
-

Submission format:

- Export your BlueSky output and submit the export as a PDF.
    - If using R, submit both your R script and output as PDFs.
- Submit steps 3 and 4 as a single PDF.
-