ASA's 2020 Fall Data Challenge

Get Out the Vote! Challenge Dataset FAQ

Contest Basics

What are the requirements of the Fall Data Challenge contest?

1. Identify a sponsor (instructor).
2. Form a team with 2 - 5 members.
3. Download the data and acknowledge the appropriate use.
4. Work as a team on the analysis.
5. Create a presentation with your results and insights (4 - 6 slides) and write a short technical description of your process.
6. Submit your presentation and technical essay by November 11, 2020.

Learn more on the Fall Data Challenge web page: (ThisIsStatistics.org/FallDataChallenge)

Where can I find the data?

Download this zipped file for the complete set of data, code, and documentation: bit.ly/FDC_Dataset.

What is this dataset about?

This is a particularly rich and interesting dataset about voting behavior in the U.S. over the past 14 years. It uses survey data from the Census Bureau and Department of Labor of Statistics that's collected on a sample of Americans every two years right after the November elections. The full dataset is large (28 variables on more than 640,000 cases) but it's also very clean, usable and well documented, thanks to the IPUMS organization (IPUMS.org).

The online version of this Dataset FAQ will be updated with additional questions from teams!

View it here: bit.ly/FDC_Dataset_FAQ_Living

Plenty of demographic information is available for each respondent: age, gender, race, ethnicity, marital status, military veteran status, size of community, employment status, and educational background.

There is also voting information for each respondent, including whether they were registered to vote (and if they were not registered, the reason why not), whether they voted in the most recent election (and if they didn't vote, the reason why not), whether they voted on Election day or before, in person or by mail.

Because the data on respondents' educational backgrounds go into some detail (including their current college enrollment status and number of years of college credit obtained), the dataset might be of particular interest for high school and college students interested in exploring voter turnout in young people ahead of this historic 2020 election.

Data Basics

What formats can I get the dataset in?

R/fixed-width

The zipped data file contains a subfolder with .xml and .dat files, plus R code that will help you read the data into R (and also do some preliminary processing). It makes use of the ipumsr package (cran.r-project.org/web/packages/ipumsr/ipumsr.pdf).

Comma-delimited

Comma-delimited files with full dataset, plus subsets with data from only the years 2016 and 2018, are available in the zipped data file. These .csv files can be imported into SPSS, SAS, STATA, Excel, CODAP and other software. The data dictionary is also available as a .csv file.

Excel

The zipped file () also contains a folder with .xlsx files, where the data and data dictionary are on separate sheets in the same file.

What are some questions I might think about in my data visualization and data analysis?

Remember that the focus of this Data Challenge is to look at voter behavior. "Voter behavior" includes whether or not a person voted in the last election, but there are also

other types of behavior that might be interesting to explore, such whether or not a person is registered to vote (and if they're not registered, why not); if a person didn't vote in the last election, the reason they gave for why they didn't vote; if a person did vote in the last election, whether they voted by mail or in person. Although not required, you may find it useful to consider other data as a part of your analysis.

Here are some other questions that might get you started in formulating your own specific questions to explore in the data:

- Does there appear to be a relationship between voting behavior and a specific personal demographic characteristic (such as gender or educational background)?
- Do voting patterns seem to differ by geography or the type of town people live in?
- Are there any trends in voting behavior over time?
- What insights do the data suggest about finding groups to be a focus of a "get out the vote" effort?
- What do the data suggest about the voting behaviors of young people, and what insights might that provide for an organization that wants to increase voter turnout among young people?

My software/computer can't handle large datasets. What should I do?

The entire dataset has almost 643,500 cases and 28 variables, which includes all 50 states and the District of Columbia, from 2004 through 2018. If you'd like to work with a smaller dataset, consider subsetting the data by time or geography. Three subsets are already provided: data from 2016 only (~80,000 cases), data from 2018 only (~73,000 cases), and data from 2016 - 2018 only. Another interesting option would be to focus on only your own state across time (for example, Virginia has ~13,000 cases from 2004 through 2018).

Data Background

Where did the data come from?

The data for the 2020 Fall Data Challenge come from IPUMS-CPS, which is an integrated set of data from the monthly Current Population Survey (CPS) by the U.S. Census Bureau and Bureau of Labor Statistics. The voter data comes from a special Voting and Registration Supplement of CPS that's conducted in November every two years. The focus of the survey is on voter behavior. The IPUMS website has more information about the IPUMS-CPS (cps.ipums.org/cps/about.shtml) and its Voter Supplement (cps.ipums.org/cps/voter_sample_notes.shtml).

What do the variable codes mean?

YEAR

YEAR reports the year in which the survey was conducted.

STATEFIP

STATEFIP identifies the household's state of residence, using the Federal Information Processing Standards (FIPS) coding scheme, which orders the states alphabetically. See the codebook for individual numeric codes.

METRO

METRO indicates whether a household was located in a metropolitan area. For households within metropolitan areas, METRO specifies whether the housing unit was inside or outside the central city of the metropolitan area. Information on metropolitan status was added by the Census Bureau, rather than being directly collected from respondents.

Value	Label
0	Not identifiable
1	Not in metro area
2	Central city
3	Outside central city
4	Central city status unknown
9	Missing/Unknown

AGE

AGE gives each person's age at last birthday.

SEX

SEX gives each person's sex.

Value	Label
9	NIU
1	Male
2	Female

RACE

Racial categories in the CPS have been more consistent than racial categories in the census. Up through 2002, the number of race categories ranged from 3 (white, negro, and other) to 5 (white, black, American Indian/Eskimo/Aleut, Asian or Pacific Islander, and other). Beginning in 2003, respondents could report more than one race, and the number of codes rose to 21, and then up to 26 codes in 2013. See the codebook for specific codes.

MARST

MARST gives each person's current marital status, including whether the spouse was currently living in the same household. See the codebook for specific codes.

VETSTAT

VETSTAT is a dichotomous variable identifying veterans, that is, persons who served in the military forces of the United States (Army, Navy, Air Force, Marine Corps, or Coast Guard) in time of war or peace, but who were not in the armed forces at the time of the survey.

Value	Label
0	NIU
1	No service
2	Yes
9	Unknown

CITIZEN

CITIZEN reports the citizenship status of foreign-born persons. In IPUMS-CPS, people born in the U.S., Puerto Rico, or U.S. outlying areas were excluded from the question universe. Respondents were identified as belonging to one of three groups: citizens by virtue of being born abroad to American parents; naturalized citizens; and non-citizens.

Value	Label
1	Born in U.S
2	Born in U.S. outlying
3	Born abroad of American parents
4	Naturalized citizen
5	Not a citizen
9	NIU

HISPAN

HISPAN identifies and classifies persons of Hispanic/Spanish/Latino origin. Origin is ancestry, lineage, heritage, national group, or country of birth. See codebook for individual codes.

LABFORCE

LABFORCE is a dichotomous variable indicating whether the respondent participated in the labor force during the preceding week. Those coded as "yes" in LABFORCE were either: were at work; held a job but were temporarily absent from work due to factors like vacation or illness; were seeking work; or were temporarily laid off from a job during the reference period.

Value	Label
0	NIU
1	No, not in the labor force
2	Yes, in the labor force

EDUC99

EDUC99 reports the respondent's highest level of educational attainment. Respondents without high school diplomas were to indicate the highest school grade they had completed, while those with high school diplomas were to indicate the highest diploma or degree they had obtained. See codebook for details.

EDCYC

EDCYC indicates the number of years of college credit earned by those with at least some college education but less than a bachelor's degree.

Value	Label
99	NIU
04	The third of junior year
05	Four or more years
03	The second or sophmore year
02	The first or freshman year
01	Less than one year (includes 0 years completed)

EDDIPGED

EDDIPGED indicates whether an individual completed high school by means of an equivalency test, such as the GED.

Value	Label
01	Graduated from high school
02	GED or other equivalent
99	NIU

EDHGCGED

EDHGCGED identifies the level of formal schooling those with no higher than a GED obtained before dropping out. Because of the limited universe, the highest grade (K-12) attended by GED earners who achieve higher educational attainment cannot be identified with this variable. The related variable, EDDIPGED, identifies GED earners from those with a traditional high school diploma. See codebook for details.

SCHLCOLL

SCHLCOLL indicates whether respondents age 16 to 24 (or 16 to 54 for ASEC 2013 onward) were enrolled in high school or college during the previous week, and, if so, whether they were enrolled full- or part-time. College or high school students who were currently on holiday or seasonal vacation were to answer yes, but those not taking classes during summer vacation were to answer no.

Value	Label
0	NIU
1	High school full time
2	High school part time
3	College or university full time
4	College or university part time
5	Does not attend school, college or university

VOWHYNOT

If an individual reported that he/she was registered to vote, but did not vote in the most recent November election, they were asked for the reason why they did not vote. See codebook for details.

VOYNOTREG

For those who were not registered to vote in the election, VOYNOTREG gives the reported primary reason for not registering.

Value	Label
99	NIU
98	No Response
96	Refused
97	Don't know
02	By mail
01	In person

VOTEWHEN

VOTEWHEN reports whether the respondent voted on election day or before election day in the most recent November election.

Value	Label
98	No Response
99	NIU
96	Refused
97	Don't know
01	On election day
02	Before election day

VOREGHOW

VOREGHOW reports the method that registered voters used to register to vote in the most recent November election. See codebook for details.

VOTED

VOTED identifies whether the person voted during the most recent November election.

Value	Label
01	Did not vote
02	Voted
96	Refused
97	Don't know
98	No Response
99	Not in universe

VOREGHOW

VOREG identifies whether the person registered to vote for the most recent November election.

The CPS assumes that individuals that voted during the most recent November elections were already registered (no double-check question for this group of observations). Thus, the universe for VOREG comprises only those persons who did not vote.

Value	Label
01	Did not register
02	Registered
96	Refused
97	Don't know
98	Not reported/Not available
99	Not in universe

VOSUPPWT

VOSUPPWT is the weight specific to the Voter Supplement.

The full codebook (cps.ipums.org/cps/resources/codebooks/cpsnov18.pdf) has more detailed technical information on all the variables. A less technical codebook is also in .csv, .xlsx, and .xml formats in the zipped data file.

What is the “Voter Supplement Weight” variable?

Only a sample of U.S. residents is interviewed for the CPS, not the entire U.S. population. So to correct for this, and to make sure the survey results better reflect the entire U.S.

population, CPS statisticians have included weights in the data, to show that some data rows represent more cases than others.

You can read more about IPUMS sample weights at the IPUMS blog (blog.popdata.org/ipums-faqs-sample-weights/) and at the IPUMS-CPS FAQ (cps.ipums.org/cps-action/faq). For R users, the "survey" package (r-survey.r-forge.r-project.org/survey/) by Thomas Lumley is a good resource.

Note: For your analyses in this Data Challenge, working with unweighted data is permissible, ***but results from analyses of unweighted data may not accurately extrapolate to the population.***

Other Resources

Where can I learn more about analyzing CPS data?

Data Training Exercises | IPUMS
ipums.org/support/exercises

Current Population Survey (CPS)
census.gov/programs-surveys/cps.html

Where can I learn more about the US electoral system?

United States Elections Project
electproject.org/

Where can I learn more about the ipumsr package?

Introduction to ipumsr - IPUMS Data in R
cran.r-project.org/web/packages/ipumsr/vignettes/ipums.html

Where can I learn more about weighted data?

Grinnell College Stat2Lab resources
stat2labs.sites.grinnell.edu/weights.html

Where can I learn more about analyzing weighted data in R?

Complex sampling and R at the University of Washington
faculty.washington.edu/tlumley/tutorials/survey-user.pdf

Survey Weights in R by Anthony B. Masters
medium.com/@theintersectuk/survey-weights-in-r-a2346273e2cf

Where can I read about good data pipelines?

Simply Statistics: How Data Scientists Think – A Mini Case Study
simplystatistics.org/2019/01/09/how-data-scientists-think-a-mini-case-study/

Is there a way I can analyze data with R in the cloud?

RStudio Cloud
rstudio.cloud