Datasheet for 'Anxiety_or_Depression'*

Chay Park

April 19, 2024

The Household Pulse Survey, conducted by the U.S. Census Bureau in collaboration with several federal agencies, serves as a critical dataset to evaluate the socioeconomic and health impacts of the COVID-19 pandemic on American households. This dataset includes self-reported data collected through an internet questionnaire, focusing on areas such as mental health, economic status, and educational disruptions. The survey employs a probabilistic sampling strategy from the Census Bureau's Master Address File to ensure representativeness, with data being weighted according to demographic benchmarks. This datasheet provides detailed insights into the dataset's composition, collection process, usage, distribution, and maintenance, offering essential information for researchers and policymakers utilizing this resource to understand and respond to the pandemic's effects.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to complement the ability of the federal statistical system to respond and provide relavant information about how emergent issues are impacting American households (National Center for Health Statistics, 2024).
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Data was provided by NCHS/DHIS and the owner of the dataset is NCHS.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

*Code and data are available at: https://github.com/Chay-HyunminPark/Anxiety_depressive_disorder

- Although there was not significant information specified about the funding, generally in addition to annual appropriations from Congress, NCHS receives additional resources in the form of reimbursables from other Federal agencies. NCHS is reimbursed by other agencies to add specific questions of interest to our surveys (National Center for Health Statistics, 2023).

4. *Any other comments?*

    - Survey datas that are open to public is scarce and often not open to public. However, regarding health statistics for US research, NCHS is a good place to start with.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance in the dataset represents a response to the Household Pulse Survey, reflecting individual experiences and assessments concerning various impacts of COVID-19. The instances include responses about mental health, economic status, and other personal impacts due to the pandemic.

2. *How many instances are there in total (of each type, if appropriate)?*

    - The dataset is composed of multiple responses to the survey, each corresponding to an individual respondent. The total number of instances depends on the number of completed responses collected during the survey period.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset is a sample from a larger set, which would be all eligible households in the U.S. The sampling strategy involved random selection from the Census Bureau's Master Address File, aiming to create a representative sample of the U.S. population by using demographic weighting and adjustment for nonresponse.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance consists of structed data collected via the survey. This includes both "raw" data in the form of survey responses and derived data such as timestamps and

numerical values for responses. Each instance comprises responses to survey questions, which are formatted and categorized by various demographic and situational factors.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - While not a label in the machine learning sense, each instance includes several attributes that categorize the data, such as indicators of mental health status, demographic groupings, and time-related information. These can be considered as labels for the purpose of analyzing survey data across different subgroups and times.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

   - Missing information occurs in instances where respondents did not answer all survey questions, or where certain data, such as the 'Quartile Range', is applicable only to specific subgroups. Reasons for missing data typically include nonresponse or the inapplicability of certain questions to all respondents.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

   - Relationships in the dataset are not about interactions between instances but rather about the correlations and trends that can be observed across different groups (e.g., demographic categories) or over time periods.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

   - The dataset does not specify recommended splits for training and testing as it is primarily designed for analysis and reporting rather than predictive modeling.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - As with any survey data, errors and noise can occur through misreporting or misunderstanding of survey questions by respondents. Redundancies might exist in the form of repeated demographic categorizations across multiple instances.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for*

example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

- Yes, the dataset is self-contained and does not rely on external resources. It provides all necessary information within the data itself.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, the dataset does not contain data considered confidential as personal identifiers are removed to ensure anonymity of the respondents.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The dataset itself is not offensive; however, the content concerning individuals' mental health and economic impacts during the pandemic might be sensitive or cause discomfort due to the nature of the topics.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Yes, the dataset identifies sub-populations by demographic attributes such as age and gender. These are clearly marked, allowing for detailed subgroup analysis within the dataset.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, it is not possible to identify individuals directly from the dataset as the data is aggregated and anonymized, focusing on broader demographic and temporal trends.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes, the dataset contains sensitive data related to race, age, and other demographic factors. However, the presentation of this data is in an aggregated form to prevent identification of individual respondents and ensure privacy and ethical compliance.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data in the Household Pulse Survey was directly reported by subjects through an internet questionnaire. Given the nature of the survey, the responses were self-reported and not independently verified. However, the data has been weighted to adjust for nonresponse and to align with Census Bureau demographic estimates, which indirectly serves as a form of validation to ensure representativeness.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected via an internet questionnaire, with invitations sent by email and text message. The Census Bureau and collaborating federal agencies likely validated the digital collection mechanism through pre-launch testing and quality assurance checks to ensure functionality and data integrity.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The survey used a probabilistic sampling strategy where housing units from the Census Bureau's Master Address File Data were randomly selected. This approach ensures that the sample represents the general population.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - The survey was conducted by the U.S. Census Bureau in collaboration with other federal agencies. Specific details on individual involvement and compensation are not typically disclosed in public documents, but governmnet employees would have conducted and managed the survey.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The timeframe for data collection was during the COVID-19 pandemic, with the survey providing weekly estimates. The timeframe of data collection should match the creation timeframe of the data since the data relates to current experiences of households during the pandemic.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Federal surveys often undergo ethical review processes to ensure compliance with standards and regulations, including privacy protections. Details of these processes for the Household Pulse Survey specifically would be documented by the Census Bureau and related agencies.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected directly from individuals, not through third parties. Respondents provided information through the questionnaire based on personal and household experiences.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Participants were likely notified about the data collection at the point of contact - either through the email or text message invitaion. This notification would include information about the purpose of the survey and its voluntary nature.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Consent to participate in the survey was implied by the completion and submission of the online questionnaire. The invitation to participate likely included language regarding consent.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - As the data collection involves a one-time survey response, there typically isn't a mechanism to revoke consent post-submission, especially since the data is anonymized for analysis.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Given the scope and scale of federal surveys, an impact analysis, including considerations for data protection and potential effects on respondents, is likely conducted. Specific outcomes and details might be available through official documnetation or upon request from the Census Bureau.

12. *Any other comments?*

    - These responses are inferred from standard practices for federal surveys like the Household Pulse Survey; specific details would be provided by the Census Bureau or in official survey documentation.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Cleaning and Validation: The data collected through online surveys typically undergoes initial cleaning to remove any obviously erroneous entries (e.g., ages outside plausible human ranges, invalid characters in numeric fields).
    - Processing of Missing Values: Missing data is a common issue in survey responses. Techniques such as imputation (replacing missing data with statistical estimates), or simply marking data as missing, may be used depending on the nature of the question and the analysis needs. For weighted surveys like this one, the adjustment for nonresponse is a crucial aspect of ensuring that the results are representative, even with missing data.
    - Labeling of Categories: Questions with categorical responses (like "yes/no," "frequently/sometimes/never") are labeled for ease of analysis. This process involves coding textual responses into numeric codes that can be easily processed statistically.
    - Discretization or Bucketing: For continuous variables or ordinal data with a wide range of responses, discretization or bucketing might be used. For example, age might be grouped into categories like 18-24, 25-34, etc., to facilitate analysis and ensure privacy.
    - Weighting: As mentioned, the data is weighted to adjust for the probability of selection and nonresponse. Weighting ensures that the survey results are representative of the broader population according to key demographic variables like age, gender, race and ethnicity, and educational attainment.
    - Anonymization: To protect respondents' privacy, personally identifiable information (PII) is removed or obscured. This includes any direct identifiers or combinations of variables that could potentially be used to re-identify individuals.
    - Quality Assurance Checks: Repeated entries, consistency of responses across questions, and outlier checks are typical quality assurance measures in survey data processing.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - On the GitHub repository, follow inputs/data/Indicators_of_Anxiety_or_Depression_Based_on_R

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - None

4. *Any other comments?*

    - Aforementioned steps are generally documented in the methodological reports associated with the survey, which might be accessible through the Census Bureau's website or directly within the docomentation of the survey dataset. For exact specifics, however, reviewing these official documnets or contacting the managing federal agency would be necessary.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The Household Pulse Survey data has been widely used for tracking and analyzing the social and economic impacts of the COVID-19 pandemic. Researchers, policymakers, and public health officials have utilized the data to understand changes in employment, mental health, food security, and other areas affected by the pandemic.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - You can find studies and papers utilizing the Household Pulse Survey data through academic databases and resources such as Google Scholar. Additionally, the U.S. Census Bureau's website often features publications and analysis derived from the data.

3. *What (other) tasks could the dataset be used for?*

    - Beyond studying the impact of COVID-19, the dataset could be used for longitudinal studies on public health trends, economic analysis, and social science research. It can serve as a baseline for future crises management studies, or for comparative studies on how different demographics cope with societal stressors.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues)*

or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

- Given the self-reported nature of the data, researchers should be cautious about biases inherent in self-report surveys, such as social desirability bias or recall bias. Additionally, because the dataset represents a specific period (COVID-19 pandemic), care should be taken when generalizing findings to other times or situations. Mitigating these risks involves robust statistical techniques, clear reporting of the survey's scope and limitations, and, where possible, validation with other data sources.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset should not be used for predicting individual behaviors or outcomes due to its aggregated and anonymized nature. It is also not suitable for clinical or diagnostic purposes, as the data reflects self-reported indicators rather than medically verified conditions.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - The Household Pulse Survey data is publicly available through the U.S. Census Bureau and can be accessed by researchers, policymakers, and the general public, promoting transparency and accessibility.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset is available for download in various formats (such as CSV) from the Census Bureau's website. It does not typically have a DOI as it is a governmental statistical data release rather than a publication in a scientific archive.

3. *When will the dataset be distributed?*

    - The dataset is released in phases and updates, often aligned with survey waves. The Census Bureau's website provides timelines and updates regarding data release schedules.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- As a governmental dataset, the Household Pulse Survey data is typically in the public domain, but users are encouraged to cite the source appropriately. Specific terms of use can be found on the Census Bureau's website.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - There are no export controls or regulatory restrictions on the dataset, given its nature as publicly available governmental data.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - As the Household Pulse Survey data is collected and distributed by the U.S. Census Bureau, a federal agency, it is typically not subject to export controls or specific regulatory restrictions. The data is publicly available and designed to be accessible by the general public, researchers, and policymakers both domestically and internationally. This openness aligns with the mission of the Census Bureau to provide useful data to help government officials, businesses, and the public make informed decisions. However, users of the data should be aware of the general legal responsibilities that come with handling data, especially when used in research and publication. These responsibilities may include ensuring privacy protections and ethical use of the data, particularly when integrating this data with other datasets that might have their own restrictions or sensitive information.

7. *Any other comments?*

    - For specific details about any applicable terms of use, users can refer to the Census Bureau's website or the direct Household Pulse Survey page, which provide guidelines and updates on data usage, distribution, and any pertinent legal notices. These resources ensure users are fully informed of their rights and obligations when accessing and using the dataset.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The U.S. Census Bureau, along with collaborating federal agencies, is responsible for maintaining, updating, and hosting the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Contact Name : National Center for Health Statistics and Contact Email : cdcinfo@cdc.gov

3. *Is there an erratum? If so, please provide a link or other access point.*

    - https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Repor/8pt5-q6wp/about_data

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - The dataset is updated periodically with new survey waves. The Census Bureau communicates these updates through its website and via press releases to registered users and the public.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - While specific retention limits are not typically imposed on public domain data like this, the data is managed in accordance with federal records management laws and guidelines.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the Household Pulse Survey data are typically archived and remain accessible through the U.S. Census Bureau's website. This ensures that historical data can be reviewed and utilized for longitudinal studies or comparative analyses. Each release is clearly dated and labeled, allowing researchers to access specific data waves or versions according to their needs. The Census Bureau typically does not remove older data versions but instead adds new survey data as additional files or entries within their data portal. In the event of any significant changes to the dataset or its availability, the Census Bureau communicates these updates through announcements on their website and potentially via email notifications to users who have registered for updates. Researchers and data users are encouraged to regularly check the Census Bureau's website for the most current data and for any notices regarding updates or changes to the dataset's maintenance and availability. This proactive approach helps ensure that all users have access to the latest information and data resources provided by the agency.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- Generally, external contributions to governmental datasets like this are not standard practice. However, researchers can use the data to build secondary datasets or augment it with other data sources for specific studies, which they can then make available independently.

References

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.