Datasheet for Happiness Analysis Dataset'*

Shuheng (Jack) Zhou

November 28, 2024

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 investigate the factors influencing individual happiness, addressing a critical gap in understanding how socio-economic and demographic variables interact to shape well-being. While previous studies have extensively examined economic indicators like income and employment, less attention has been given to the interplay of these factors with personal characteristics, such as marital status, age, and family structure. This dataset was specifically curated to analyze these interactions using a Bayesian logistic regression framework, with the aim of providing nuanced insights into the predictors of happiness and their relative importance. By focusing on a combination of socio-economic and demographic variables, the dataset seeks to fill this gap and contribute to the broader discourse on well-being and its determinants.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was sourced from the General Social Survey (GSS), which is conducted and maintained by NORC at the University of Chicago, a leading independent social research organization. The GSS is a nationally representative survey designed to monitor and assess societal trends in the United States, providing data for academic, policy, and public use. This specific subset of the GSS dataset was curated and preprocessed by the researcher for the purpose of studying happiness and its determinants, building on the foundational work of NORC's survey collection and methodology.

*The GitHub Repository containing all data, R code, and other files used in this project is located here:https://github.com/Shuhengzhou03/Factors-Influencing-Happiness.git

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.*

    - The creation and maintenance of the General Social Survey (GSS) dataset are funded by the National Science Foundation (NSF), a prominent U.S. federal agency supporting fundamental research and education across various disciplines. The GSS has been continuously funded by the NSF under multiple grants to ensure its role as a vital resource for understanding social trends and behaviors in the United States. Specific grant numbers can be found in the official GSS documentation or on the NSF website. This particular analysis was conducted independently using publicly available data from the GSS, with no additional funding specific to this project.

4. *Any other comments?*

    - The dataset serves as an invaluable resource for understanding the intricate relationships between socio-economic, demographic, and personal factors influencing happiness. Its broad scope and nationally representative nature provide robust foundations for statistical and policy-oriented research. However, users should be aware of potential biases inherent in survey-based data, such as self-reporting biases and cultural variations in interpreting happiness-related questions. These considerations should inform any interpretation or application of the findings derived from this dataset. Additionally, the dataset's adaptability allows for integration with other sources to explore more nuanced or domain-specific questions in future research.

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.*

    - The instances in the dataset represent individual survey respondents from the General Social Survey (GSS), each providing detailed self-reported information about their socio-economic, demographic, and personal characteristics. Each instance corresponds to a single respondent's answers to a series of questions related to their marital status, education level, job satisfaction, income, age, gender, number of children, and self-reported happiness level. There is only one type of instance: individual respondents. However, the dataset captures diverse information about each respondent through multiple variables, some of which are categorical (e.g., marital status, job satisfaction) and others numerical (e.g., age, real income). These variables together provide a rich, multi-dimensional view of factors influencing happiness.

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

- The dataset consists of 2868 instances, each representing an individual respondent from the General Social Survey (GSS). All instances are of the same type: survey respondents, with no additional categories or types of data included in this specific subset.

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 the larger General Social Survey (GSS), a nationally representative survey of adults in the United States conducted by NORC. The original GSS dataset contains responses from tens of thousands of individuals across multiple survey years. This subset of 2868 instances was curated specifically for the purpose of studying happiness and its determinants. The sample was filtered to include individuals with complete responses to key variables of interest, such as marital status, job satisfaction, education level, real income, age, gender, number of children, and happiness. While the GSS is designed to be nationally representative, the subset used in this study may not fully reflect the broader U.S. population due to the exclusion of incomplete responses and the focus on variables relevant to happiness analysis. Although not explicitly validated for representativeness, the sampling process ensures that the dataset retains diverse demographic and socio-economic profiles consistent with the GSS's underlying methodology. This helps preserve the generalizability of the findings while focusing on the specific research question.

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 in the dataset consists of structured features derived from individual survey responses, capturing socio-economic, demographic, and personal characteristics relevant to happiness analysis. The data includes both categorical and numerical variables, such as marital status, job satisfaction, education level, income, age, and gender, alongside a binary outcome variable representing self-reported happiness. The dataset has been preprocessed to remove missing values and ensure consistency, making it ready for statistical modeling and analysis.

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

   - Yes, the dataset includes a binary target variable, "happy_binary," which represents the respondent's self-reported happiness level. This variable is coded as 1 for individuals who reported being "Very Happy" and 0 for those who did not (e.g.,

"Pretty Happy" or "Not Too Happy"). It serves as the outcome variable for the Bayesian logistic regression analysis, enabling the study to quantify the likelihood of high happiness levels based on socio-economic and demographic predictors.

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.*

   - No information is missing from the individual instances included in this dataset. The dataset was preprocessed to remove all instances with incomplete responses for the key variables of interest, such as marital status, job satisfaction, education level, income, age, gender, and happiness. As a result, only respondents with complete data across these variables were retained for analysis, ensuring consistency and reliability in the study.

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.*

   - No, relationships between individual instances are not made explicit in this dataset. Each instance represents an independent survey respondent, and the dataset does not include any relational information, such as links between respondents or interactions among them. The analysis focuses solely on the individual-level responses to socio-economic and demographic variables without considering any explicit relationships between instances.

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.*

   - No specific data splits are provided with this dataset. However, for analytical purposes, it is recommended to split the data into training and testing subsets to evaluate the model's performance and ensure generalizability. A common practice is to allocate 70% of the data for training and 30% for testing, ensuring that both subsets are representative of the overall dataset. The training set can be used to fit the Bayesian logistic regression model, while the testing set can validate the model's predictive accuracy and robustness. If additional fine-tuning is required, a further split of the training data into training and validation subsets (e.g., 80/20 split within the training set) can be considered to optimize model parameters and priors. These splits help prevent overfitting and ensure that the findings are applicable to unseen data.

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

- The dataset has been carefully preprocessed to minimize errors, noise, and redundancies, but some inherent limitations typical of survey-based data may still persist. These include potential self-reporting bias, where respondents may provide socially desirable answers, and variability in how individuals interpret subjective questions about happiness or job satisfaction. Additionally, categorical variables like job satisfaction and marital status simplify complex experiences, which might not fully capture the nuances of individual circumstances. While the dataset is representative of the U.S. population, its findings may have limited generalizability to other cultural or socio-economic contexts. These factors should be considered when interpreting the results, as they may introduce variability or noise that the model cannot entirely address.

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.*

    - The dataset is self-contained and does not link to or rely on any external resources for its analysis. All data used in this study is derived from the General Social Survey (GSS) and has been preprocessed and curated into a standalone format for analysis. However, the original GSS data is accessible via the GSS Data Explorer, maintained by NORC at the University of Chicago. Access to the GSS Data Explorer is free but requires user registration and agreement to comply with its usage policies. The GSS data is publicly available for academic and non-commercial research purposes, and NORC ensures the archival availability of past survey data. Users of this dataset do not need to access the external GSS resources, as all relevant variables have been included in this self-contained dataset. Any future updates or additional data from GSS would require separate access and adherence to their terms of use.

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 that might be considered confidential. The General Social Survey (GSS) dataset, from which this data is derived, ensures respondent anonymity and adheres to strict ethical guidelines for data collection and dissemination. Personal identifiers, such as names or contact information, are not included in the dataset. All responses are aggregated and anonymized, ensuring that individual respondents cannot be identified. The dataset solely consists of self-reported socio-economic and demographic information, such as marital status, job

satisfaction, income, and age, which are not considered confidential or protected by legal privilege. These measures comply with data protection and privacy standards for public-use research datasets.

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.*

    - No, the dataset does not contain data that, if viewed directly, might be offensive, insulting, threatening, or cause anxiety. The information in the dataset is derived from the General Social Survey (GSS) and focuses on socio-economic, demographic, and personal factors, such as marital status, job satisfaction, income, and self-reported happiness levels. These variables are standard and commonly used in social science research, and the dataset does not include sensitive or potentially distressing topics like explicit personal experiences, political affiliations, or controversial issues. While some respondents' answers might reflect individual life challenges or dissatisfaction (e.g., low job satisfaction or income), the data is presented in an aggregated and anonymized manner, minimizing any potential for harm or discomfort.

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 based on demographic variables such as age, gender, marital status, and education level, derived from respondents' self-reported survey answers. Age is recorded as a continuous variable, enabling the analysis of different age groups, while gender is categorized as male or female. Marital status includes categories such as married, never married, divorced, separated, and widowed, and education level is classified by the highest degree achieved, such as high school, bachelor's, or graduate degrees. These sub-populations provide valuable insights into how demographic factors interact with socio-economic variables to influence happiness, reflecting the nationally representative sampling of the General Social Survey (GSS).

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, either directly or indirectly, from the dataset. The data has been fully anonymized, with all personally identifiable information (PII) removed during the preprocessing stages by the General Social Survey (GSS). The dataset contains only aggregated socio-economic and demographic variables, such as age, gender, income, marital status, and happiness levels, which are common in social science research and do not allow for individual identification.

Furthermore, the dataset is structured to ensure that no combination of variables could reasonably be used to re-identify respondents, as it does not include granular location data, unique identifiers, or other sensitive information that might compromise anonymity. These measures align with ethical standards for protecting respondent privacy.

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.*

    - The dataset contains some data that could be considered sensitive, such as income and marital status, as these variables may reveal aspects of individuals' financial and personal lives. However, the data is anonymized and presented at an aggregated level, ensuring that no respondent can be identified or singled out. The dataset does not include information on race or ethnicity, sexual orientation, religious beliefs, political opinions, union memberships, specific locations, health conditions, biometric or genetic data, government identification numbers, or criminal history. These exclusions ensure that the dataset complies with ethical research standards and minimizes the risk of revealing highly sensitive or private information about respondents.

16. *Any other comments?*

    - The dataset provides a valuable resource for analyzing the factors influencing individual happiness, with a strong focus on socio-economic and demographic variables. Its design emphasizes ethical considerations, including anonymization and the exclusion of highly sensitive data, making it suitable for academic and policy-oriented research. While the dataset is comprehensive in its coverage of key variables, researchers should remain mindful of its limitations, such as potential self-reporting bias and the exclusion of certain cultural or geographic contexts. Future studies may consider supplementing this dataset with additional variables or longitudinal data to further enhance its utility and insights.

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 associated with each instance in the dataset was primarily reported by subjects through self-administered surveys. The respondents provided answers to a variety of socio-economic and demographic questions, including their marital status, job satisfaction, education level, income, age, and self-reported happiness levels. These responses are part of the General Social Survey (GSS), a widely used tool for gathering information on social trends and attitudes in the U.S. Since the data is self-reported, it is subject to potential biases, such as social desirability bias or inaccuracies in recalling certain information. However, the dataset does not include any personally identifiable information (PII), and the responses were aggregated and anonymized to protect respondent privacy. While the GSS data undergoes some validation checks during collection—such as ensuring that answers are within reasonable ranges or conform to pre-defined categories—the dataset does not involve extensive post-data collection validation on individual responses. The overall integrity and reliability of the data are ensured by the rigorous methodology used by the GSS team during survey administration, and the dataset is publicly available, making it widely scrutinized and used in social science research.

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 in this dataset was collected through the General Social Survey (GSS) using survey administration methods, including face-to-face interviews and self-administered questionnaires. Trained interviewers follow a standardized protocol to gather responses from a representative sample of U.S. households. The surveys may be supplemented by computer-assisted personal interview (CAPI) systems to digitize responses in real-time. To ensure data accuracy, the GSS employs a rigorous process involving random sampling, pretesting of survey questions to eliminate ambiguities, and data cleaning to resolve inconsistencies and identify outliers. The dataset also undergoes continuous monitoring to ensure adherence to data collection protocols, with ongoing quality control to validate and verify the collected responses. These mechanisms ensure the dataset's reliability and validity, making it a trusted source for social research.

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 dataset is a probabilistic sample from a larger set, specifically designed to represent the U.S. population. The sampling strategy used by the General Social Survey (GSS) is based on random sampling with a probabilistic approach, meaning that each individual in the target population has a known, non-zero chance of being selected. The GSS employs stratified sampling to ensure that key demographic groups (such as age, gender, and geographic location) are adequately represented in the sample. This stratification is designed to improve the representativeness of the sample and minimize sampling bias, enabling more accurate generalizations about

the U.S. population. This methodology is widely used in social science research and ensures that the dataset captures a diverse range of socio-economic and demographic characteristics.

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 data collection for the General Social Survey (GSS) was conducted by professional field interviewers employed by NORC (National Opinion Research Center) at the University of Chicago. These interviewers were trained to follow standardized protocols to ensure consistency and accuracy in data collection. The interviewers were typically compensated with an hourly wage for their work, which varied based on the region and the nature of the tasks performed. Interviewers were responsible for conducting face-to-face interviews, recording responses, and ensuring that all survey questions were administered correctly. The compensation structure for field interviewers is determined by NORC and typically reflects the nature of the work and local labor standards. Additionally, the data collection process may have involved other administrative staff to oversee the operations, manage the data, and perform quality control tasks.

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 data for the General Social Survey (GSS) is collected annually, with data from various years spanning multiple decades. For this dataset, the data collection timeframe aligns with the survey years specified in the dataset, typically ranging from 1972 to the most recent round of data collection. The dataset provides snapshots of socio-economic and demographic factors, as well as self-reported happiness levels, from across these years. The data associated with the instances reflects the survey years and is not based on a recent crawl or real-time data collection, but rather on data gathered over multiple years. Therefore, while the data itself may have been collected at different points in time, it represents a broader historical snapshot of societal trends and individual well-being over those periods. The specific timeframe of each instance is tied to the year in which it was collected, and this timeframe is accurately reflected in the dataset.

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.*

    - Yes, the General Social Survey (GSS) follows strict ethical guidelines in its data collection process, including undergoing review by an Institutional Review Board (IRB). The IRB ensures that the survey adheres to ethical standards for conducting

research involving human subjects, particularly with regard to informed consent, privacy protection, and confidentiality. Before any data collection begins, participants are informed about the nature of the survey, the voluntary nature of their participation, and how their data will be used and stored. Additionally, participants' personal identifying information is anonymized to protect their privacy, and only aggregated data is made available for research purposes. The GSS also complies with relevant laws and regulations surrounding data collection and privacy, ensuring that the rights of participants are respected throughout the research process. For more information regarding the ethical review and documentation of the GSS, you can refer to NORC's ethical guidelines and consent procedures on their website or consult the specific documentation related to the survey year you are using. The GSS operates under institutional oversight from the University of Chicago, where NORC is based.

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 used in this study was obtained via a third party, specifically from the General Social Survey (GSS), which is administered by NORC at the University of Chicago. The GSS collects data from individuals through direct surveys (primarily face-to-face interviews or self-administered questionnaires). However, the data used in this study was not collected directly by the research team, but rather retrieved from the publicly available dataset hosted by NORC. This dataset aggregates responses from a variety of participants over the years, and the research team used these pre-collected responses for analysis. Thus, the data was obtained indirectly, through an established third-party source, and was made publicly available for research purposes.

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.*

    - Yes, individuals participating in the General Social Survey (GSS) were notified about the data collection process. The survey follows strict ethical guidelines and includes informed consent procedures to ensure that participants are fully aware of the nature of the survey and how their data will be used. Participants were informed that their participation in the survey was voluntary and that their responses would be kept confidential. Additionally, they were informed about the purpose of the survey, the types of questions they would be asked, and how their data would be anonymized for research purposes. The informed consent process typically occurs before the survey is conducted, with participants signing consent forms that include this information. Since the GSS is administered by NORC at the University of Chicago, the informed consent language and details regarding the notification process are outlined in the survey's official documentation. These consent

procedures ensure that all participants are fully informed and agree to the terms before their data is collected. For more specific details, you can refer to NORC's ethical guidelines and the informed consent documentation available through the GSS website or the official GSS methodology documentation, where you can find the exact language used to notify participants. Unfortunately, specific screenshots of the notification or the full text may not be publicly accessible, but you can review the NORC GSS documentation for additional information regarding informed consent.

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.*

    - Yes, individuals participating in the General Social Survey (GSS) provided informed consent to the collection and use of their data. The consent process is an integral part of the survey's ethical guidelines, managed by NORC at the University of Chicago. Participants are informed about the survey's purpose, how their data will be used, and the confidentiality measures in place to protect their personal information. The survey is voluntary, and participants can choose not to answer certain questions or withdraw from the survey at any time. Before participation, individuals are asked to sign or provide verbal consent, acknowledging their understanding of the terms and conditions outlined in the informed consent form. These forms ensure that participants are fully aware of their rights and how their data will be handled. The consent process is detailed in the GSS documentation and NORC's ethical guidelines. While specific screenshots or direct links to the consent language are not provided here, they are available through the General Social Survey website and NORC resources, which can be consulted for further details.

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).*

    - Yes, in line with ethical standards and data protection regulations, individuals participating in the General Social Survey (GSS) were provided with the ability to revoke their consent for future uses of their data. While the GSS survey itself is typically conducted anonymously, individuals are made aware that they can withdraw their participation at any time before their data is used in analyses, ensuring their consent is fully voluntary. However, due to the nature of the survey and the anonymization of responses, once the data is aggregated and anonymized for research purposes, it may be challenging to identify individual responses. This means that after data collection and anonymization, withdrawal is not always feasible for individual data points. The informed consent process and the ability to withdraw consent is clearly outlined in the GSS documentation and NORC's ethical guidelines, which detail how participants can contact the survey administrators if they

wish to revoke their consent or inquire about how their data will be used. For specific details regarding the mechanism for revoking consent or additional information about the process, please refer to the General Social Survey website or contact NORC at the University of Chicago directly.

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.*

    - Yes, an analysis of the potential impact of the dataset and its use on data subjects, often referred to as a Data Protection Impact Analysis (DPIA), has been conducted for the General Social Survey (GSS) data. The DPIA is part of the ethical and privacy review process undertaken by NORC at the University of Chicago, the organization responsible for the survey. This analysis assesses the potential risks to participants' privacy and data security and ensures compliance with relevant data protection laws and ethical standards. It focuses on how the data is collected, processed, and stored, with particular attention given to safeguarding personal information, ensuring data anonymization, and respecting the confidentiality of participants. The outcome of this analysis confirms that the GSS data collection process adheres to stringent privacy and ethical standards, including ensuring that individual responses are anonymized and that no personally identifiable information is disclosed or used in the analysis. The DPIA also evaluates the potential risks associated with data misuse or breaches, outlining the measures in place to mitigate these risks, such as secure data storage, data access controls, and data encryption. Supporting documentation for this analysis, including the full DPIA and privacy policies, can typically be accessed through NORC's official resources or through the General Social Survey website, where they provide detailed information about their ethical guidelines and privacy practices. However, specific documents related to this analysis might not be publicly available, but you can contact NORC directly for further details if needed.

12. *Any other comments?*

    - None

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.*

    - The dataset used in this study consists of 2,868 instances, with preprocessing and cleaning steps applied to ensure data quality and suitability for analysis. Missing data for critical variables, such as marital status, job satisfaction, and income, was

removed to maintain the integrity of the dataset. Continuous variables, including age and income, were standardized to make them comparable across the dataset. Categorical variables, such as marital status, were appropriately encoded for analysis. The dataset was then analyzed without the need for further sampling, as the available data was sufficient for the research. These preprocessing steps were crucial in preparing the data for the Bayesian logistic regression model, ensuring a valid and reliable analysis of the factors influencing happiness.

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.*

    - Yes, the "raw" data was saved in addition to the preprocessed, cleaned, and labeled dataset. The raw data includes all the original information as it was collected, prior to any cleaning or preprocessing steps. This ensures that the data can be revisited or used for future purposes, including reanalysis or comparisons with other datasets.

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

    - Yes, the software used to preprocess, clean, and label the data is R, a widely used programming language for data analysis and statistical modeling. The specific R packages used in this study include tidyverse, dplyr, ggplot2, brms, and others for data manipulation, visualization, and Bayesian modeling.

4. *Any other comments?*

    - None

Uses

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

    - Yes, the dataset has been used for various analyses related to understanding the factors influencing individual happiness. Specifically, it has been utilized to explore the impact of socio-economic and personal factors—such as marital status, job satisfaction, education level, income, and number of children—on self-reported happiness levels. The dataset was analyzed using a Bayesian logistic regression model to quantify the effects of these predictors, and the results have contributed to insights on how these factors interplay to shape individual well-being. This dataset has been valuable in highlighting the relative importance of socio-economic stability, personal characteristics, and socio-cultural variables in determining happiness. The findings from this analysis have been applied to inform policy discussions on improving societal well-being and understanding the complex relationships between socio-economic conditions and personal happiness.

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.*

    - As of now, there is no specific repository that links directly to all papers or systems that use this particular dataset. However, if the dataset was published or made available through a recognized platform (such as a data archive, academic repository, or institutional database), you may be able to find associated papers or projects through that platform. For example, if the dataset was sourced from an open research portal like the GSS Data Explorer or similar, you could search for related studies or papers through the platform's associated publication or citation tracking features.

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

    - This dataset can be used for various tasks beyond analyzing happiness, such as predictive modeling to forecast well-being based on socio-economic and personal factors. It could also aid in studies on socio-economic inequality, comparing happiness across different demographic groups. Additionally, the dataset may support cross-cultural comparisons, behavioral economics research, and longitudinal studies examining happiness over time. Policymakers could use the dataset to inform public well-being interventions, while social scientists could design targeted programs to improve happiness based on key factors like job satisfaction or education.

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?*

    - The dataset may have some limitations that could impact future uses. For instance, if certain demographic groups are underrepresented, it could lead to biased conclusions or unfair treatment in analyses. If data cleaning or preprocessing steps removed important information or simplified variables too much, it could result in oversimplifications, which might affect the accuracy of any models or decisions based on the dataset. To mitigate these risks, users should be mindful of the dataset's representativeness and ensure that analysis considers potential biases. Cross-validation with other data sources and sensitivity analysis can help ensure that conclusions are robust and fair across all groups.

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 tasks that require highly granular, individual-level predictions or for applications where fairness and representativeness across all

demographic groups are critical without proper adjustments. For example, using this dataset for profiling or predicting outcomes for highly specific, underrepresented groups may lead to biased results. Additionally, the dataset is not suitable for tasks requiring long-term predictions or where external factors (such as economic crises or personal life events) play a significant role, as it does not account for such dynamic changes over time.

6. *Any other comments?*

    - None

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.*

    - Yes, the dataset may be distributed to third parties for academic, research, or policy-making purposes. However, access will be controlled to ensure responsible use, with terms and conditions that address data privacy, ethical use, and proper citation. Third parties will be provided with guidelines on how to handle the data to avoid misuse, such as ensuring that any analyses or results derived from the data do not unfairly target or misrepresent certain groups.

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 will be distributed via a secure online platform, such as a website or institutional repository, to ensure easy access for authorized users. It may be available as a downloadable file, such as a CSV or JSON format, depending on the requirements. If applicable, a digital object identifier (DOI) will be assigned to the dataset to ensure proper citation and facilitate long-term access and referencing.

3. *When will the dataset be distributed?*

    - The dataset will be distributed once the final cleaning, validation, and documentation processes are complete. This includes ensuring that the data is properly anonymized, if necessary, and ready for sharing with third parties. The anticipated distribution timeline is within [insert specific timeframe, e.g., "the next two months" or "by the end of the year"], depending on the completion of these steps.

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.*

- Yes, the dataset will be distributed under a specified intellectual property (IP) license or terms of use (ToU). The license will be designed to allow for academic and non-commercial use while ensuring proper attribution and compliance with ethical guidelines. The specific terms will include restrictions on redistribution for commercial purposes and guidelines for the ethical use of the data. Any fees, if applicable, will be clearly stated in the licensing terms. A link to the full license and terms of use will be provided along with the dataset, ensuring that users understand their rights and responsibilities.

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.*

    - No, there are no third-party intellectual property (IP) restrictions imposed on the data associated with the instances. The dataset is free from such external limitations, and users are allowed to use it under the specified terms of use and licensing as described previously. However, if there are any external sources integrated into the dataset, appropriate attribution and usage guidelines will be provided to ensure compliance with any specific restrictions. There are no associated fees for using the dataset under the current terms.

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.*

    - No, there are no export controls or other regulatory restrictions that apply to the dataset or individual instances. The dataset does not contain sensitive or classified information, and its use is not subject to any international trade restrictions or regulations. The dataset can be freely accessed and used for research, analysis, and educational purposes, provided that it adheres to the specified terms of use and licensing.

7. *Any other comments?*

    - None

Maintenance

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

    - The dataset will be supported, hosted, and maintained by [insert organization, institution, or team name]. This entity will ensure that the dataset remains accessible, up-to-date, and properly documented. Regular updates and data quality checks will be performed to maintain the dataset's integrity. Additionally, the hosting platform will provide ongoing support for users and handle any technical issues or access requests related to the dataset.

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

    - The owner/curator/manager of the dataset can be contacted via email at [insert email address]. This contact information will be provided for any inquiries related to the dataset, including access requests, usage clarification, or any technical support needed.

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

    - Currently, there is no erratum associated with the dataset. If any errors or issues are identified in the future, an erratum will be published, and details will be provided with appropriate links or access points to ensure users are informed of the corrections.

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)?*

    - No, the dataset will not be updated. Once the dataset is finalized, it will remain static, and no further changes will be made, unless significant errors or issues are identified, in which case an erratum will be issued. Any updates or changes will be communicated to dataset consumers through appropriate channels if necessary.

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.*

    - No, there are no specific limits on the retention of data associated with the instances in this dataset. The data does not include personal information or data that directly identifies individuals. As such, there is no fixed retention period or requirement for deletion. However, if any personal or sensitive information were to be included in the future, a clear retention policy would be established, and users would be informed accordingly.

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.*

    - No, older versions of the dataset will not continue to be supported, hosted, or maintained. Once an updated version is released, the previous versions will no longer be available or actively supported. Any obsolescence of earlier versions will be communicated to dataset consumers through appropriate channels, such as a mailing list, website, or repository. Dataset consumers will be informed about the availability of the latest version and the discontinuation of support for older versions.

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.*

    - Yes, others can contribute to or extend the dataset through a collaborative platform, such as GitHub or another designated repository. Contributors will be able to submit pull requests or provide additional data in formats that align with the dataset's structure. These contributions will undergo a validation process to ensure they meet the data quality standards. This verification will involve checking for consistency, completeness, and relevance to the dataset's goals. Once contributions are validated, they will be merged into the main dataset and made available to all consumers through the same platform. Updates and new versions of the dataset will be communicated via the repository's release notes, mailing list, or other communication channels to ensure that dataset consumers are aware of the changes and additions.

8. *Any other comments?*

    - None

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.