Analysis of Physical Activity Survey(2022)

Jonathan Goodwin

27 April 2022

Abstract

Recreational activity and physical fitness are important factors in determining the health and safety of a society. We obtain data regarding the activity of individuals of different states and age groups as well as their respective distance to a park. We find that citizens who live closer to parks are more likely to be physically active, and that younger citizens are more likely to be physically active then older ones. These findings have implications for the future development of public recreational structures.

1 Keywords:

Health, Physical Activity, Recreation, United States, Survey

2 Introduction

Physical activity is an important indicator for the health of the population of the United States. As well living near to a public recreational service like a park allows residents a more convenient place to excercise. We see that decreasing the distance to a park as well as giving more opportunity for residents to participate in activities like biking is associated with more physically active residents who meet the CDC's guidelines for strength and aerobic health.

We analyzed the sources of water for two subsets of the Jordan population, urban and rural residents. From Table 2 and Table 3, we can see that both Colorado is the state with the lowest rate of obesity and that Mississipi is the highest.

This information is relevant to the further development and enhancement of the recreational and physical activity of citizens of the United States as well as leader of respective states when developing cities. The data can help to identify populaces with particularly low activity reletive to the rest of the country as well as identify states that are lacking in public recreational support for its citizens.

Overall we found that the state with the highest rates of Obesity was Mississipi, the lowest was Colorado and that there is a relationship between the proportion of state residents participating in recreational activities like muscle training or aerobics and the obesity rate of residents within that state.

In section 3 we talk about the process of gathering and analyzing the dataset and the variables. Then in section 4 we build a linear model for the rate of obesity. Then in section 5 we discuss the implications of the dataset presented. Finally in section 6 we discuss implications for the United States citizens, as well as the weakness of the study and the further investigation that may be useful to the topic of recreational activity support and physical activity in the United States.

3 Data

The dataset is from the National Health Interview Survey. The survey was conducted by the Centers for Disease Control and Prevention, (*Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System* 2022) and the data used was from the years 2010 to 2020.The survey was conducted and funded by the CDC but also has numerous sponsors found here.

The survey has been conducted every year for over 50 years, the sample is nationally representative, the intent of the survey is to monitor the health of the nation, as it collects information by interviewing American households. The content of the survey is updated roughly every 15-20 years with the most recent being in 2019. Of which 30,000 adults and 9,000 children were interviewed. Of this survey, this analysis focused on the information regarding the proportion of adults classified as obese or overweight, and how they responded to questions regarding aerobic activity and weight training. The data is collected from all 50 states including the District of Columbia, there is also a section included for the National averages of each question response.

To compile the data set, the R language was used(R Core Team 2020), along with the packages, Pointblank (Iannone and Vargas 2022), haven (Wickham and Miller 2021), the paper was compiled using Knitr (Xie 2021) and KableExtra(Zhu 2021) packages. Also made use of reshape2 (Wickham 2007) in manipulating the data to create plots.

Table 1 gives a small look at the dataset.

Table 1: Excerpt of dataset

Year	State	Question	Percent	Age
2011	Alabama	Percent of adults aged 18 years and older who have obesity	35.2	25 - 34
2011	National	Percent of adults who engage in no leisure-time physical activity	16.9	18 - 24
2016	Virginia	Percent of adults aged 18 years and older who have an overweight classification	40.1	35 - 44
2016	Washington	Percent of adults who engage in no leisure-time physical activity	18.8	55 - 64
2016	Alabama	Percent of adults aged 18 years and older who have an overweight classification	35.3	55 - 64
2011	National	Percent of adults who engage in no leisure-time physical activity	22.1	25 - 34

In total there are 6 characteristics from the survey being analyzed, they are:

1. "Percent of adults aged 18 years and older who have obesity"

2. "Percent of adults aged 18 years and older who have an overweight classification"

3. "Percent of adults who engage in no leisure-time physical activity"

4. "Percent of adults who achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)"

5. "Percent of adults who engage in muscle-strengthening activities on 2 or more days a week"

6. "Percent of adults who achieve at least 300 minutes a week of moderate-intensity aerobic physical activity or 150 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)"

These are stratified into 6 different age groups for adults, those 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and older.

The responses are given as a proportion of the sample, rounded to 1 digit. So the value 35.2 means, 35.2% of respondents in that age group had been affirmative to that trait in the survey.

4 Model

The model being used is Obesity $= \beta_{\text{Intercept}} + \beta_{\text{150min of Aerobics}} \times$ (150min of Aerobics or more) $+ \beta_{\text{Muscle Training}} \times$ (Muscle Training). Where the beta coefficients represent the change in the predictor variables, 150min or more of Aerobics, and Muscle Training, have on the result, being the obesity rate. This model was chosen since other factors like no aerobic activity or 300min or more of aerobic activity were unsuprisingly highly correlated with the 150min of aerobics.

Both Muscle Training and 150min of Aerobic Training had significant p-values in this model of 0.0236 and 0.0027 respectively. Below Figure 1 shows the relationship between 150min of Aerobics training and Muscle Training compared to the rate of Obesity. There was strong positive correlation to people with 300min of aerobic training, and those who did muscle training or 150min of aerobic training so including that interaction was not beneficial. Additionally those that do no aerobic training at all are strongly negatively correlated with those that do 150min or more. Soi including that interaction was not beneficial to the model.

Our data does not have any particularly large outliers that would disturb the model present. And our outcome variable, the proportion of residents who are obese is a continuous variable, thus the bilinear model seems the best choice.

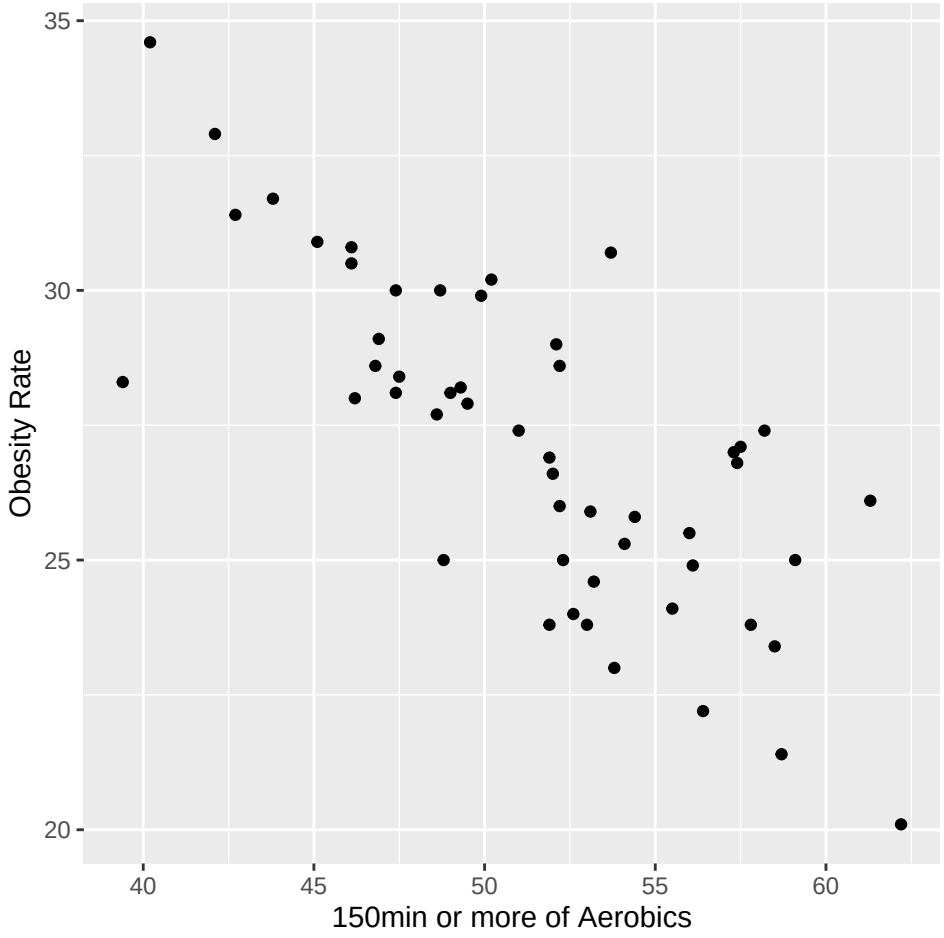

Figure 1: Obesity vs 150min of Aerobics

And below Figure 2 shows us the relationship between the rate of obesity and the proportion of a states residents participating in muscle training excercises.

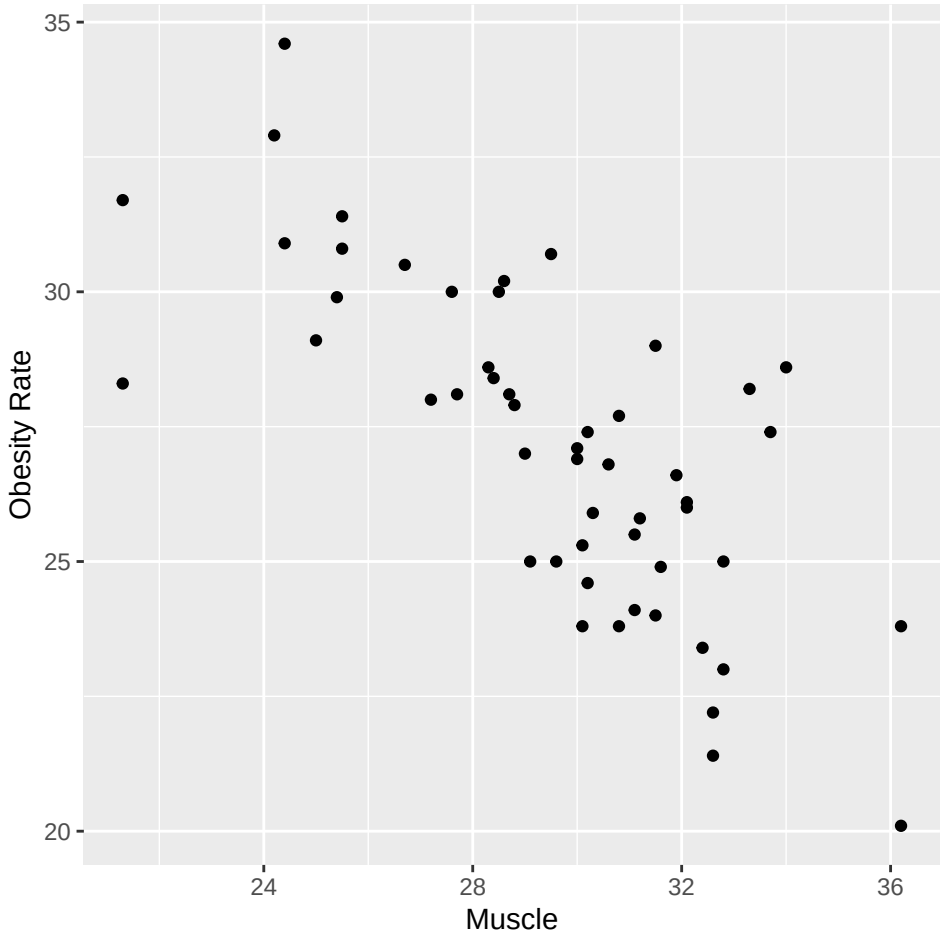

Figure 2: Obesity vs Muscle Training

5 Results

From Figure 3 we see a chart comparing the proportion of respondents who were obese based on the different states in the country. We can see from the chart that all states lie between a rate of 20% and 35% obese. The chart below is for 2011.

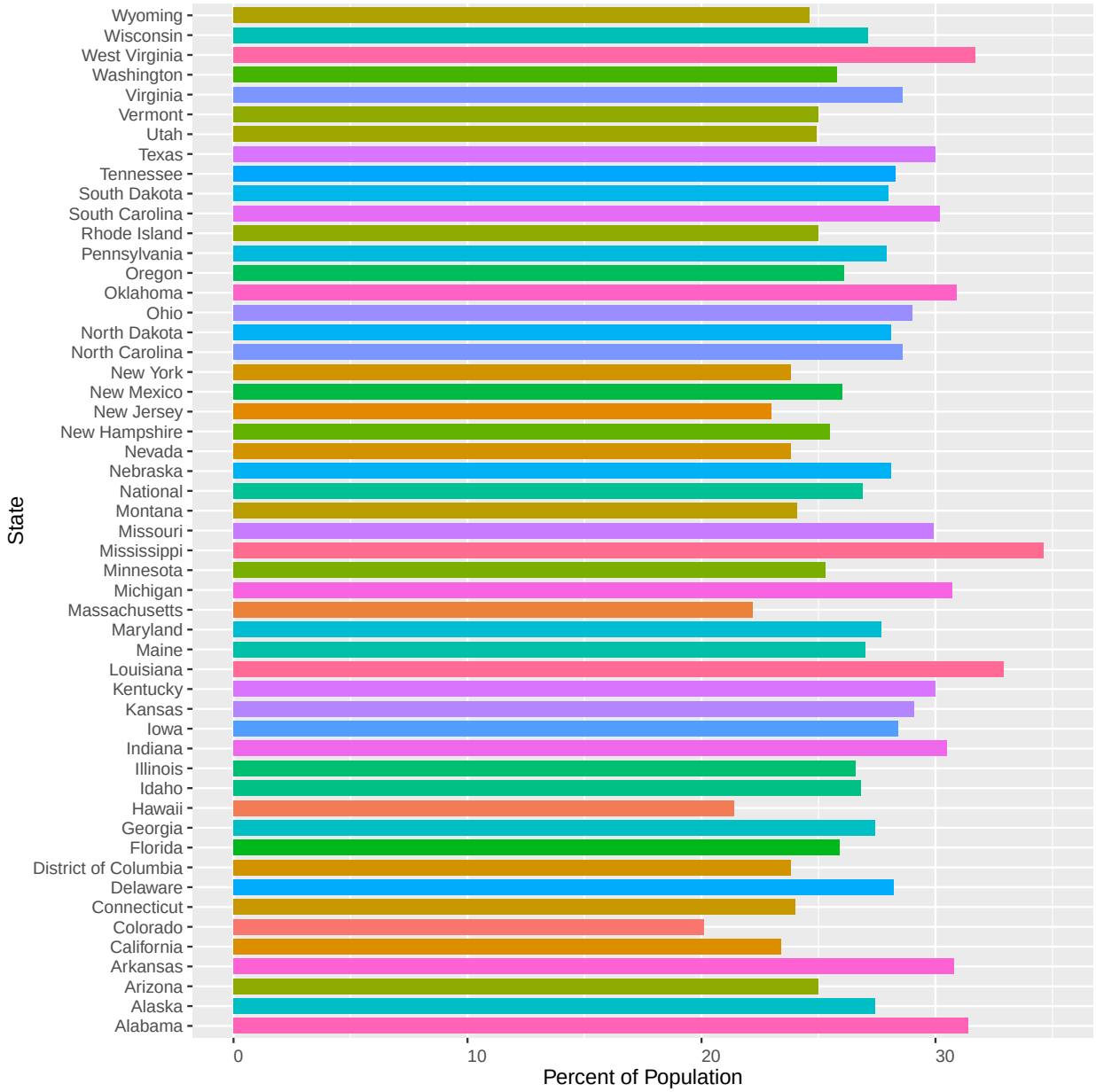

Figure 3: Bar plot of Obesity by State

Below Table 2, shows the states with the lowest proportions of obesity.

Table 2: Excerpt of dataset

Colorado	20.1
Hawaii	21.4
Massachusetts	22.2
New Jersey	23
California	23.4
District of Columbia	23.8

And here Table 3 gives the states with the highest obesity rates.

Table 3: Excerpt of dataset

Mississippi	34.6
Louisiana	32.9
West Virginia	31.7
Alabama	31.4
Oklahoma	30.9
Arkansas	30.8

From these tables and the chart above we see the lowest obesity rate in the United States is Colorado with 20.1% and the highest is Mississippi with 34.6%. For this analysis we want to see whether physical activity factors such as aerobic excercise. Below in 4 however we will briefly see how these obesity rates have changed by the end of the decade.

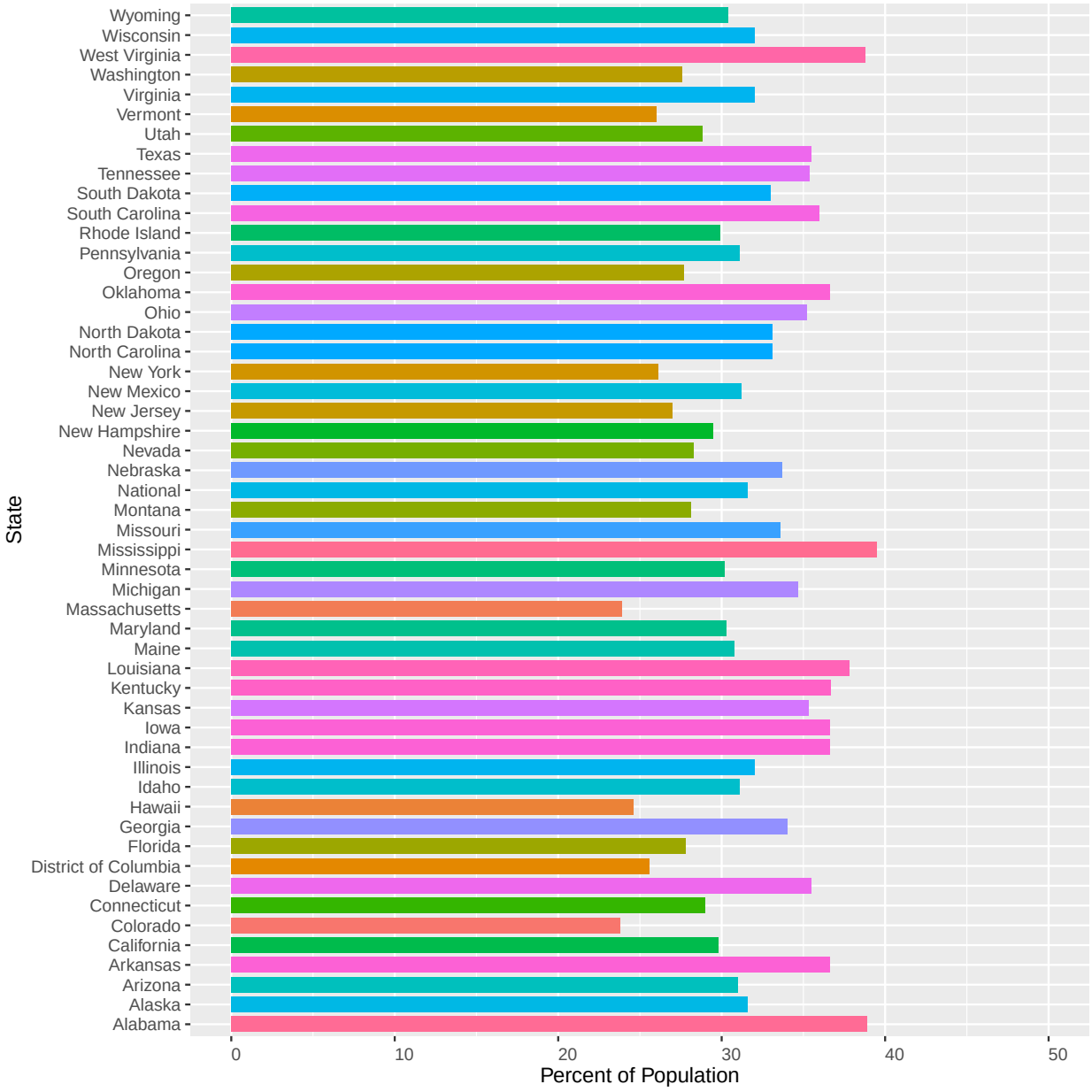

Figure 4: Bar plot of Obesity by State

We see from 4 that the overall rate of obesity has clearly increased, however the states with the highest and lowest rates remain largely the same as we can see from Table 4 and Table 5 below.

Table 4: Excerpt of Figure 1

Colorado	23.8
Massachusetts	23.9
Hawaii	24.6
District of Columbia	25.6
Vermont	26
New York	26.1

Table 5: Excerpt of dataset

Mississippi	39.5
Alabama	38.9
West Virginia	38.8
Louisiana	37.8
Kentucky	36.7
Arkansas	36.6

Using our model from the

6 Discussion

6.1 FirstPoint

6.2 Weaknesses and next steps

Appendix

DataSheet

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 analyze the responses from National Health Interview Survey regarding physical activity.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Jonathan Goodwin
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.*
    - No one.
4. *Any other comments?*
    - No.

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 row of the dataset corresponds to a the proportion of intervewees that responded affirmatively to a question by year and stratified by age group.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 6 questions, 6 age categories, over the course of 10 years for all 53 states, resulting in 14,196 different entries.
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 National Health Interview Survey conducted by the CDC.
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 a numeric value for the proportion of respondents which responded affirmatively to that question in the survey.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - No.
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.

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.

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.

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

   - No.

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 does not rely on any external sources.

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 data is all publicly available from the Department of statistics Jordan and required no special permissions to access.

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.

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.*

    - The data is split into 6 age groups, 18-24, 24-35, 35-45, 45-55, 55-64, and 65 and older.

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 individuals can be identified from the dataset.

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.*

    - None of the data is of a sensitive nature.

16. *Any other comments?*

    - No.

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 was originally acquired by the Center for Disease Control and Prevention, (*Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System* 2022) via the National Health Interview Survey.
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?*
    - Subjects for the survey were interviewed in person.
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 sample was from the 50 states of the United States and the District of Columbia as well as Peurto Rico and the sampling strategy of cross-sectional household interview survey.
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 Center for Disease Control and Prevention conducted the survey along with various sponsors.
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 was collected from 2010 to 2020.
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.*
    - NHIS surveys are approved and review by the ICF Institutional Review Board(IRB).(**citeDHS__Ethics?**).
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - Individuals of the survey were contacted directly for the survey.
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 are notified of all aspects of the survey.
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.*
    - All participants gave consent for their data.
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).*
    - Participants of the survey are told they may terminate participation at any time.

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.*
    - NHIS is approved by the Research Ethics Review Board of the National Center for Health Statistics and the U.S. Office of Management and Budget. All NHIS respondents provided oral consent prior to participation.
12. *Any other comments?*
    - No

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 was reduced from the original sample data provided through the National Health Interview Survey.
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, both the raw data acquired through the survey, and the cleaned version of the dataset is available in the repository associated with this analysis.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - The R language and packages associated with the cleaning process are all freely available.
4. *Any other comments?*
    - No.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Prior to this analysis the dataset was only used as part of the original Survey.
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.*
    - Yes, it is available on github here
3. *What (other) tasks could the dataset be used for?*
    - The dataset has been minimized for this analysis but the full raw data includes other factors regarding water security that may be of interest.
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?*
    - No.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No.

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.*
    - No.
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 via Github.
3. *When will the dataset be distributed?*
    - The dataset is currently available via Github.
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.*
    - No.
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.
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.
7. *Any other comments?*
    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be available on Github
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - No.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
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.

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.

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.

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.*

   - No.

8. *Any other comments?*

   - No.

Code

Repository associated with this analysis is available at github

References

Iannone, Richard, and Mauricio Vargas. 2022. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*. https://CRAN.R-project.org/package=pointblank.

Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. 2022.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

Wickham, Hadley, and Evan Miller. 2021. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. https://CRAN.R-project.org/package=haven.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*.