BRFSS

Behavioral Risk Factor Surveillance System

Dated: 2/24/2023

# Dataset Details

* **Rows:** 2289902 (2.29M) **Columns:** 27
* **Data Link:** [BRFSS Data](https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu)
* **Year:** 2011 to 2021
* **Location:** 50 states, district of Columbia and 3 U.S. territories

|  |  |  |
| --- | --- | --- |
| Data Dictionary - BRFSS | | |
| **Sr No** | **Column Name** | **Description** |
| 1 | 'Year' | Year 2011 to 2021 |
| 2 | 'Locationabbr' | State Abbreviation |
| 3 | 'Locationdesc' | State Name |
| 4 | 'Class' | Health-related risk categories |
| 5 | 'Topic' | Reasons of risk factors |
| 6 | 'Question' | Behavioral questions |
| 7 | 'Response' | Response to the questions |
| 8 | 'Break\_Out' | Category wise breakout |
| 9 | 'Break\_Out\_Category' | Total 6 - 'Overall', 'Gender', 'Education Attained', 'Race/Ethnicity', 'Age Group', 'Household Income' |
| 10 | 'Sample\_Size' | No of people surveyed (0-31k) |
| 11 | 'Data\_value' |  |
| 12 | 'Confidence\_limit\_Low' |  |
| 13 | 'Confidence\_limit\_High' |  |
| 14 | 'Display\_order' |  |
| 15 | 'Data\_value\_unit' | % |
| 16 | 'Data\_value\_type' | The data value type, such as age-adjusted prevalence or crude prevalence |
| 17 | 'Data\_Value\_Footnote\_Symbol' | Blank, \*, \*\*, \*\*\* |
| 18 | 'Data\_Value\_Footnote' |  |
| 19 | 'DataSource' | BRFSS |
| 20 | 'ClassId' | 20 class ids are present |
| 21 | 'TopicId' | 71 topic ids are present |
| 22 | 'LocationID' | 1-99 |
| 23 | 'BreakoutID' | 5-6 ids |
| 24 | 'BreakOutCategoryID' | 6 |
| 25 | 'QuestionID' |  |
| 26 | 'ResponseID' | 1-229 |
| 27 | 'GeoLocation' |  |

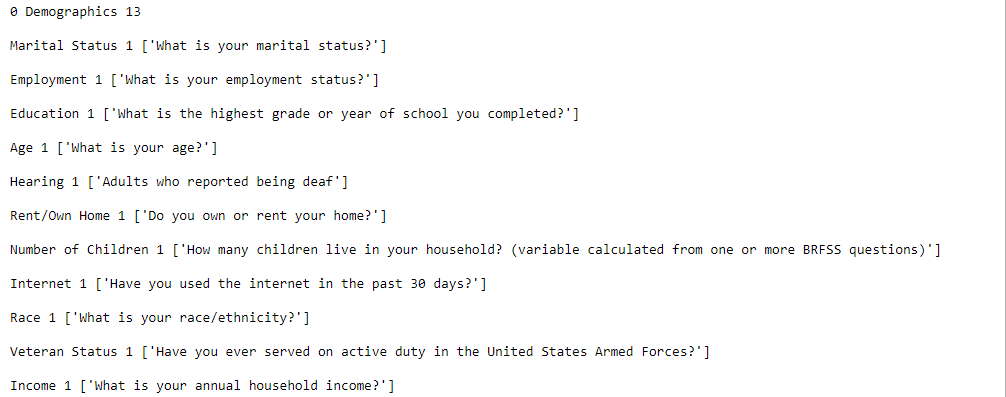
# Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a well-known public health survey conducted annually by the Centers for Disease Control and Prevention (CDC) in the United States. It is one of the largest ongoing telephone health surveys in the world and gathers data on a wide range of health-related topics, such as health conditions, risk behaviors, and preventive health practices. The BRFSS data is used by public health officials, researchers, and policymakers to identify emerging health issues and to design and evaluate public health interventions. The survey data provides valuable insights into the health status and behaviors of individuals across different states and territories in the US and is an important resource for advancing public health research and policy. From 2.29 million records there 43k. records of Indiana state.

**Phase-1 Data Analysis:**

Analyzing survey data can be a challenging task as it involves collecting responses from individuals in relation to a set of questions. Without prior knowledge of the questions asked, it can be difficult to begin any meaningful analysis of the responses. To gain a better understanding of the scope of the survey, we first examined the class feature of the data and found that it comprises of 18 distinct classes, with a total of 63 unique topics. On average, each topic corresponds to only one unique question, thereby providing a useful foundation for further analysis.

Example:



Disability status 7 ['Do you have difficulty doing errands alone?', 'Are you blind or do you have serious difficulty seeing, even when wearing glasses?', 'Adults with health problem(s) that requires the use of special equipment', 'Do you have difficulty dressing or bathing?', 'Do you have serious difficulty walking or climbing stairs?', 'Adults who are limited in any activities because of physical, mental, or emotional problems', 'Do you have serious difficulty concentrating, remembering, or making decisions?']

**Task Identified:**

Following the initial examination of the survey data and its corresponding responses, we proceeded to undertake a new task focused on identifying the percentage of individuals who reported health issues within specific age groups. To achieve this, it directed our attention to the "Breakout\_category" column, which contained four distinct breakout categories, namely Race/Ethnicity, Age Group, Household Income, Education Attained, Gender, and Overall. To obtain data for the desired age groups, we filtered the dataset to only include records in the "Age\_Breakout" category. This preprocessing step resulted in a reduction of the number of records from the original 1,220,800 to 109,670 records, representing only those individuals whose age had been recorded. It was also noted that a small number of null values existed in the dataset, which were subsequently removed to ensure the accuracy and reliability of the data.

**Findings/Conclusion:**

While in this phase we found that there is this same data prepared by CDC and is been preprocessed for further analysis. In this preprocessed data every row containing a set of values are converted to a column and values are encoded in the numeric form. This data was stored in ASCII format which then had to be converted to csv. SAS programming was used to convert this data into readable characters.

# Preprocessed Dataset Details

* **Rows:** 4,38,693  **Columns:** 304
* **Data Link:** [BRFSS Preprocessed Data](https://www.cdc.gov/brfss/annual_data/annual_2021.html)
* **Year:** 2021 to 2022
  + 2021: 9439 records
  + 2022: 490 records
* **Location:** 50 states, district of Columbia and 3 U.S. territories

**Data Preprocessing:**

As we have to focus on specific state we filter data to Indiana state, which gives us following numbers:

* Rows: 9929

While looking at the data we found there are few columns containing all values as null, so we drop these columns, which gives us following numbers

* Columns: 221

There are few columns where more than 90% of the values are null. Let us drop more columns which are not related to health and prepare a data dictionary.

|  |  |  |
| --- | --- | --- |
| Data Dictionary – BRFSS Preprocessed | | |
| **Sr No** | **Column Name** | **Description** |
| **1** |  | Full form  VALUE: Description |
| **2** | X.STATE\* | States  1-51, 53-56, 66, 72, 78 |
| **3** | FMONTH\* | File month  1-12: Jan to Dec |
| **4** | IDATE\* | Interview date  HIDDEN  Available - MMDDYYYY |
| **5** | IMONTH\* | Interview month  1-12: Jan to Dec |
| **6** | IDAY\* | Interview day  1-31 |
| **7** | IYEAR\* | Interview year  2021, 2022 |
| **8** | DISPCODE | Final disposition:  1100 – completed interview  1200 – partial complete interview |
| **9** | SEQNO | Annual sequence number  HIDDEN – data not displayed  Available, 10 digit number starts with 2021000001 |
| **10** | PSU | Primary sampling unit  HIDDEN – data not displayed  Available, 10 digit number starts with 2021000001 |
| **11** | CTELENM1 | Correct telephone number?  1 - Yes—Go to LL.02, PVTRESD1  2 - No—Terminate Phone Call  Blank – Missing Notes: QSTVER > = 20 |
| **12** | PVTRESD1 | Private residence?  1 – Yes  2 – No |
| **13** | COLGHOUS | Do you live in college housing?  1 – Yes  Blank - Missing |
| **14** | STATERE1 | Residence of state?  1 – Yes  Blank - Missing |
| **15** | CELPHON1 | Cellular number?  2 – not a cell phone  Blank – missing |
| **16** | LADULT1 | Are you 18 years of age or older?  1 – yes  2 – no  Blank – not asked or missing |
| **17** | COLGSEX | Are you male or female?  1 – male  2 – female  Blank – not asked or missing |
| **18** | NUMADULT | Number of adults in household  1  2  3  4  5  6-99: 6 or more  Blank - missing |
| **19** | GENHLTH\* | General health  1 - excellent  2 – very good  3 - good  4 - fair  5 - poor  7 - don’t know / not sure  9 – refused  Blank – not asked or missing |
| **20** | PHYSHLTH\* | Number of Days Physical Health Not Good  1-30: number of days  88 – none  77 - don’t know/not sure  99 – refused  Blank – not asked or missing |
| **21** |  |  |
| **22** |  |  |
| **23** |  |  |
| **24** |  |  |
| **25** |  |  |
| **26** |  |  |
| **27** |  |  |

Columns

17

'income3', - Income Level (1-11)

'pregnant', - Pregnancy Status

'weight2', - Reported Weight in Pounds

'height3', - Reported Height in Feet and Inches

'deaf', - Are you deaf or do you have serious difficulty hearing?

'blind', - Blind or Difficulty seeing

'decide', - Difficulty Concentrating or Remembering

'diffwalk', - Difficulty Walking or Climbing Stairs

'diffdres', - Difficulty Dressing or Bathing

'diffalon', - Difficulty Doing Errands Alone

'smoke100', - Smoked at Least 100 Cigarettes

'smokday2', - Frequency of Days Now Smoking (every day, some days, not at all)

'usenow3', - Use of Smokeless Tobacco Products (every day, some days, not at all)

'ecignow1', - Do you now use e-cigarettes, every day, some days, or not at all?

'alcday5', - Days in past 30 had alcoholic beverage

'avedrnk3', - Avg alcoholic drinks per day in past 30

'drnk3ge5’ - Binge Drinking