*This document will summarize the extraction, transformation and loading of our dataset.*

### **BRFSS Introduction**

Every year since 1984, government employees have called around the country and asked the people who pick up a series of probing questions about their health and what they're doing about it. While it may sound annoying, the CDC's Behavioral Risk Factor Surveillance System (BRFSS) provides a wealth of information about health and health-related behaviours in the United States. It is the largest and longest-running health survey system in the world, and in its current incarnation, it covers over 400,000 adult interviews from all 50 states, the District of Columbia, and three territories. For more information about the survey itself, you should check out the [CDC BRFSS site](http://www.cdc.gov/brfss/).

The BRFSS is a rich source of information on how demographics, behaviours, and other risk factors can correlate with health. Many important population health studies and measures use the BRFSS as a key data source. For example, it is the source of the CDC's "Healthy Days" measurement, a key performance metric for the healthcare industry.

Unfortunately, BRFSS data isn't exactly easy to deal with. Its breadth and structure have changed considerably over the years, and there are important sampling considerations that must be taken into account when using the data to draw conclusions. Our goal is to demonstrate how to use BRFSS data and some of the interesting correlations and associations that can be drawn from this data set using machine learning and statistical techniques.

### **Getting BRFSS data**

To get started, you'll want to download the data from the [BRFSS Annual Survey Data](http://www.cdc.gov/brfss/annual_data/annual_data.htm) page. There you can find links to each year the survey has been conducted. The data is available in .XPT (SAS Transport Format) or in .ASCII files.

### 

### **Cleaning BRFSS data**

After reading the SAS file and creating our DataFrame, we note that the BRFSS data is not useful right out of the box. We have to do some heavy-duty cleaning.

The 2021 data set includes 438693 rows and 303 columns. *W*e must begin by defining which variables/features/indicators we want to keep.

According to the [CDC](https://www.cdc.gov/diabetes/basics/risk-factors.html), the following risk factors contribute to some developing diabetes:

[Type 1 diabetes](https://www.cdc.gov/diabetes/basics/type1.html) is thought to be caused by an immune reaction (the body attacks itself by mistake). Risk factors for type 1 diabetes are not as clear as for prediabetes and type 2 diabetes. Known risk factors include:

* Family history: Having a parent, brother, or sister with type 1 diabetes.
* Age: You can get type 1 diabetes at any age, but it usually develops in children, teens, or young adults.

In the United States, White people are more likely to develop type 1 diabetes than African American and Hispanic or Latino people. Currently, no one knows how to prevent type 1 diabetes.

You’re at risk for [type 2 diabetes](https://www.cdc.gov/diabetes/basics/type2.html) if you:

* Have prediabetes.
* Are overweight.
* Are 45 years or older.
* Have a parent, brother, or sister with type 2 diabetes.
* Are physically active less than 3 times a week.
* Have ever had gestational diabetes (diabetes during pregnancy) or given birth to a baby who weighed over 9 pounds.
* Are an African American, Hispanic or Latino, American Indian, or Alaska Native person. Some Pacific Islanders and Asian American people are also at higher risk.

If you have non-alcoholic fatty liver disease you may also be at risk for type 2 diabetes.

You can prevent or delay type 2 diabetes with proven lifestyle changes. These include losing weight if you’re overweight, [eating a healthy diet](https://www.cdc.gov/diabetes/managing/eat-well.html), and getting regular [physical activity](https://www.cdc.gov/diabetes/library/features/get-moving-to-manage-diabetes.html).

Using the [2021 BRFSS Codebook](https://www.cdc.gov/brfss/annual_data/2021/pdf/codebook21_llcp-v2-508.pdf) to identify the variable names, location, and frequency of values for all reporting areas combined for the combined data set.

Of the 303 features, we will narrow our data set to include:

| **Field** | **Description** | **Action** |
| --- | --- | --- |
| dispcode | 1100 - Completed Interview  1200 - Partial Complete Interview | Exclude records for partially completed surveys |
| medcost1 | Could Not Afford To See Doctor  1 = Yes 2 = No (change to 0) | Change NaN = 7  Exclude values = 7, 9 |
| exerany2 | Exercise in Past 30 Days  1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| cvdinfr4 | Diagnosed with Heart Attack  1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| cvdcrhd4 | Diagnosed with Angina or Coronary Heart Disease 1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| cvdstrk3 | Diagnosed with a Stroke  Disease 1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| chccopd3 | Diagnosed with C.O.P.D. emphysema or chronic bronchitis? 1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| chckdny2 | Diagnosed with Kidney Disease  1 = Yes 2 = No (change to 0) | Exclude values = 7, 9, blank |
| diabete4 | Diagnosed with Diabetes  1 = Yes  2 = Yes, but female told only during pregnancy (change to 0)  3 = No (change to 0)  4 = No, pre-diabetes or borderline diabetes (change to 0) | Exclude values = 7, 9,  Create a prediabetic column and exclude values = 1, 2, 7, 9, |
| diffwalk | Difficulty Walking or Climbing Stairs  1 = Yes 2 = No | Exclude values = 7, 9, blank |
| \_rfhlth | Adults with good or better health  1 = Good or Better Health  2 = Fair or Poor Health | Exclude values = 9 |
| \_imprace | Race/Ethnicty  1 = White  2 = Black  3 = Asian  4 = American Indian/Alaskan Native  5 = Hispanic  6 = Other |  |
| \_phys14d | Physical Health Status  1 = Zero days when physical health not good  2 = 1-13 days when physical health not good  3 = 14+ days when physical health not good  9 = Don’t know/Refused/Missing | Exclude values = 9 |
| \_ment14d | Mental Health Status  1 = Zero days when physical health not good  2 = 1-13 days when physical health not good  3 = 14+ days when physical health not good  9 = Don’t know/Refused/Missing | Exclude values = 9 |
| \_hlthpln | Have any health insurance  1 = Yes 2 = No | Exclude values = 9 |
| \_totinda | Physical Activity  1 = Yes 2 = No | Exclude values = 9 |
| \_rfhype6 | High Blood Pressure  1 = No 2 = Yes | Exclude values = 9 |
| \_sex | Gender  1 = Male 2 = Female |  |
| \_age\_g | Age  1 = Age 18 to 24  2 = Age 25 to 34  3 = Age 35 to 44  4 = Age 45 to 54  5 = Age 55 to 64  6 = Age 65 or older |  |
| \_rfbmi5 | Overweight or obese  1 = No  2 = Yes  9 = Don’t know/Refused/Missing | Exclude = 9 |
| \_educag | Level of education completed  1 = Did not graduate High School  2 = Graduated High School  3 = Attended College or Technical School  4 = Graduated from College or Technical School | Exclude = 9 |
| \_incomg1 | Income  1 = Less than $15,000  2 = $15,000 to < $25,000  3 = $25,000 to < $35,000  4 = $35,000 to < $50,000  5 = $50,000 to < $100,000  6 = $100,000 to < $200,000  7 = $200,000 or more | Exclude = 9 |
| \_smoker3 | Smoker  1 = Current smoker -now smokes every day 2 = Current smoker -now smokes some days 3 = Former smoker  4 = Never smoked | Exclude = 9 |
| \_rfbing5 | Binge drinkers (males having five or more drinks on one occasion, females having four or more drinks on one occasion)  1 = No 2= Yes | Exclude = 9 |
| \_rfdrhv7 | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)  1 = No 2 = Yes | Exclude = 9 |
| \_frtlt1a | Consume Fruit 1 or more times per day  1 = Consumed fruit one or more times per day  2 = Consumed fruit < one time per day | Exclude = 9 |
| \_vegete1a | Consume Vegetables 1 or more times per day  1 = Consumed fruit one or more times per day  2 = Consumed fruit < one time per day | Exclude = 9 |

| **Breakdown of cleaning** | | |
| --- | --- | --- |
| **Extraction** | **Step 1** | Working in Google Colab, we imported and read the raw SAS/.XPT file and created a DataFrame |
| **Transformation** | **Step 2** | At this point, our data consists of 438693 rows and 303 columns. We generate a list of column headings to better understand the data available, using the CDC Codebook to translate and understand the values used in each column. |
| **Step 3** | We removed any columns with null values. |
| **Step 4** | Our ***target*** variable, DIABETE4, includes responses where individuals surveyed indicated that they had either been diagnosed as diabetic, diabetic due to pregnancy, pre-diabetic, or not diabetic. As pre-diabetes is an indicator for diabetes, we wanted to isolate this indicator in its own column. We duplicated the DIABETE4. |
| **Step 5** | We selected the indicators that were of most interest to us based on research of the topic. |
| **Step 6** | Responses included incomplete survey results. As these responses would be largely incomplete, we deleted any records were the survey was noted as partially complete. |
| **Step 7** | As MEDCOST1 (whether individual did not see a doctor due to the cost) provides an interesting avenue of exploration. The entire dataset includes 3 blank results. In an effort to include the data in our analysis, we assigned any null values as non-responsive. |
| **Step 8** | We further transform the data to create some consistency in our labelling and in the nature of the data. This includes removing any data points that we unresponsive, altering binary classifiers to be consistent across all columns, renaming our columns to make them more readable, and changing our data type to integers. |
| **Load** | **Step 9** | Finally, we exported our clean dataset to CSV and saved it in a shared Google folder. Our final dataset includes 242888 rows and 27 columns. |