# Datasets

The two datasets used for this assignment are about heart disease. They both come from the American Centres for Disease Control and Prevention (CDC) and they are both about heart disease.

## Dataset #1

The first one is the National Health Interview Survey (NHIS) National Cardiovascular Disease Surveillance Data. [1] This tracks the rate of cardiovascular disease and any risk factors for cardiovascular diseases, and the data is organised based on location such as state, county etc. it has 4,455 rows and 27 columns. The table below contains a description of the 27 columns.

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Data Type** |
| Year | - | Temporal |
| Locationabbr | Location Abbreviation | Location |
| Locationdesc | Location Description | Location |
| Datasource | Where the information came from | Categorical |
| priorityarea1 | Priority Area (Million Hearts® or None) | Categorical |
| priorityarea2 | Priority Area (ABCS or None) | Categorical |
| priorityarea3 | Priority Area (Healthy People 2020 or None) | Categorical |
| priorityarea4 | Priority Area (AHA 2020 Goals: Cardiovascular Health Metrics or None) | Categorical |
| Category | - | Categorical |
| Topic | - | Categorical |
| Indicator | - | Categorical |
| break\_out\_category | How is the data being divided up (e.g. race, age) | Categorical |
| break\_out | Which group from the break\_out\_category (e.g. if break\_out\_category is race, break\_out could be ‘indian’) | Categorical |
| data\_value\_type | The data type e.g. percentage | Categorical |
| data\_value\_unit | Symbol for the datatype e.g. % | Categorical |
| data\_value | Value for the data type | Continuous |
| data\_value\_footnote\_symbol | Symbol that would be used for the flag footnotes | Categorical |
| data\_value\_footnote | Footnote description | Categorical |
| confidence\_limit\_low | 95% confidence interval lower bound | Continuous |
| confidence\_limit\_high | 95% confidence interval upper bound | Continuous |
| Categoryid | - | Categorical |
| Topicid | - | Categorical |
| Indicatorid | - | Categorical |
| Breakoutcategoryid | - | Categorical |
| Breakoutid | - | Categorical |
| Locationid | - | Location |
| Geolocation | Coordinates of location | Location |

## Dataset #2

The second dataset also comes from CDC and it contains all of the same columns as the first, so the table above also describes the columns in this dataset. [2] However, the difference is that the first one tracks the rate of cardiovascular disease from 2000 - 2014, this one tracks the rate of heart disease from 2011 – 2014. The other difference is that this dataset contains 35,004 rows, which is a lot more than the first dataset.

These datasets will be merged based on the year column.

# Datasets

[1] Centres for Disease Control and Prevention, (2016), National Health Interview Survey (NHIS) - National Cardiovascular Disease Surveillance Data, <https://data.world/cdc/nhis-national-cardiovascular>, ***Date Accessed:*** 3rd April 2020

[2] Centres for Disease Control and Prevention, (2016), Behavioral Risk Factor Heart, <https://data.world/cdc/behavioral-risk-factor-heart>, ***Date Accessed:*** 3rd April 2020