Data Preparation (Note: there are still missing values in the data due to stratification variables)

The Chronic Disease Indicators (CDI) dataset was downloaded as a .csv file from https://www.healthdata.gov/dataset/us-chronic-disease-indicators-cdi

This dataset contains information from a variety of data sources related to a number of chronic conditions. The full dataset contains 403,984 records. The metadata for this dataset are described in the Morbidity and Mortality Weekly Report, January 9, 2015.

The group of records used for analysis in this project are the Nutrition, Physical Activity, and Weight Status group (NPAWS) which consists of 29,512 records. The metadata for this subset of data are at page 148 – 190 of the above report.

These data are in long format with a separate row for every question. This makes direct numeric comparisons for each question result (mostly a %, or a median value) difficult (at least for me! If anyone has any suggestions for using the long format data for this purpose please let me know).

So I have used R code to reshape the data to wide format with a separate column for each question data value. The code can be found here. Any rows with all null data values have been excluded so that the final dataset as shown in the screenshot has 2,123 records. There are still null values in some of the question columns as a result of the stratification of the data (gender and race/ethnicity) and separate rows for Adults >= 18 and High School students.

In Tableau the data values for each question have come in as text (due to R producing “NA”s for null results. Tableau has been used to reset these as numeric and converted from a dimension to a measure for numeric analysis.

The dataset still has a large number of columns for questions as shown in this list. As a result of the exploratory analysis to follow, a number of these are likely to be excluded.

Data Preparation

The Chronic Disease Indicators (CDI) was downloaded as a .csv file from https://www.healthdata.gov/dataset/us-chronic-disease-indicators-cdi

This dataset contains information from a variety of data sources related to a number of chronic conditions. The full dataset contains 403,984 records. The metadata for this dataset are described in the Morbidity and Mortality Weekly Report, January 9, 2015.

The group of records used for analysis in this project are the Nutrition, Physical Activity, and Weight Status group (NPAWS) which consists of 29,512 records. The metadata for this subset of data are at page 148 – 190 of the above report.

These data are in long format with a separate row for every question. This makes direct numeric comparisons for each question result (mostly a %, or a median value) difficult (at least for me! If anyone has any suggestions for using the long format data for this purpose please let me know).

So I have used R code to reshape the data to wide format with a separate column for each question data value. The code can be found here. Any rows with all null data values have been excluded so that the final dataset as shown in the screenshot has 2,123 records. There are still null values in some of the question columns as a result of the stratification of the data (gender and race/ethnicity) and separate rows for Adults >= 18 and High School students.

In Tableau the data values for each question have come in as text (due to R producing “NA”s for null results. Tableau has been used to reset these as numeric and converted from a dimension to a measure for numeric analysis.

The dataset still has a large number of columns as shown in the following list. As a result of the exploratory analysis to follow, a number of these are likely to be excluded.