## horizontal line



Diabetes Across the State Level

01.30.2022

Poseidon Team

Katejoe Akabogu

Chris Chaney

Liat Sayfan

Farrukh Sultani

# Overview

Diabetes is a chronic disease that affects how the body turns food into energy. There are three main types of diabetes - type 1, 2, and gestational diabetes (diabetes while pregnant). More than 133 million in the United States live with diabetes, and 96 million of 133 are US adults. Centers for Disease Control has categorized diabetes as the 7th leading cause of death in the United States. In the last 20 years, the number of adults diagnosed with diabetes has increased tremendously in the American population due to ageism and obesity. Yearly, an estimated cost of $327 million is spent on medical costs and lost work/wages for people with diagnosed diabetes; people with diabetes tend to have double the high medical cost in comparison with people who don’t have diabetes.

Furthermore, 90% to 95% of the population is diagnosed with type 2 diabetes, and the remaining 5% to 10% account for type 1 diabetes.

There isn’t a cure yet for diabetes, but weight management, good nutrition, and being active can improve the health outcome of a diabetic patient. Our study focus areas will highlight the relationship between physical activity, food nutrition, and weight can improve or degrade the diabetes prognosis at the state level of the United States.

# Goals - Questions to answer in this research study

1. What is the rate of diabetes in the United States?
2. Are there any differences between the four major regions of the United States in terms of diabetes, physical activity, obesity, and Nutrition?
3. How strong is the relationship between nutrition, physical activity, obesity, and diabetes at the state level?
4. Can we predict diabetes rates at the state level based on nutrition, physical activity, and obesity rates?

# Specifications - Datasets

## CDC NPAO.csv

The data has been extracted from the Mississippi Behavioral Risk Factor Surveillance System, the 2011 report. The dataset provides data for four years. The data collection started in 2011 and ended in 2015 across all states. The dataset classification is focused on Fruits and Vegetables, Obesity/weight, and physical activity. The main questions which were asked were:

- Percent of adults aged 18 years and older who have an overweight classification

- Percent of adults aged 18 years and older who have obesity

- 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)

- 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)

- Percent of adults who engage in muscle-strengthening activities on two or more days a week

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

- Percent of adults who report consuming fruit less than one time daily

- Percent of adults who report consuming vegetables less than one time daily

The dataset also provides figures on the sample sizes, ages, and education levels of the survey population groups. Age, Education, Gender, Income, Race, and Ethnicity are the values that not all the population groups were willing to respond to, so the dataset lacks that information for some of the classifications. Because the data is limited on these features, we decided to exclude this information from our analysis.

The team found the dataset from https://data.world; however, a combined report is available over the internet and on Mississippi Behavioral Risk Factor Surveillance System, 2011 report as a PDF file. Both links are available below:

<https://data.world/basilhayek/cdc-nutrition-physical-activity-and-obesity-by-state/workspace/file?filename=cdc_npao.csv>

<http://www.msdh.state.ms.us/brfss/brfss2011ar.pdf>

## 500\_Cities\_Diagnosed\_diabetes\_among\_adults\_aged\_18\_year.csv

The data provided in this dataset was sponsored by the Centers of Disease Control (CDC). Their data is an aggregate of the Behavioral Risk Factor Surveillance System (BRFSS) data (2017), Census Bureau 2010 census population data, and American Community Survey (ACS) 2013-2017 estimates. This dataset is limited to 2017 and uses over 500 city census tracts and 28,000 census tracts. This dataset highlights the percentage of census per city and state considered diabetic. This dataset does not provide specifics on other defining categories such as type of diabetes, dietary health, etc. With agreement from the CDC, it can be concluded that this open-source data source is viable and should be considered ethical. The data includes several superfluous columns that need to be removed and incomplete or missing data.

[www.cdc.gov/500cities](http://www.cdc.gov/500cities)

# Breakdown of Tasks

## Clean diabetes rates dataset -Chris

The data set includes rates per different cities. We need to combine the data by state for both the rates of diabetes and the population size. After cleaning, the dataset should include the population size in one column, and the diabetes rate in another, and each row would represent a different state.

## Clean nutrition, physical activity, and obesity dataset

The data set includes information gathered from multiple samples across different years for each state. The rates of the different variables of interest (nutrition, physical activity, and obesity) are located in rows. We need to reorganize the data set to create a data frame where the different variables represent an individual column, and each row represents a state.

## Merge the two cleaned datasets -Farrukh.

Once both datasets are cleaned and reorganized, we will combine both into a single data frame using the state and the indicator to merge. The final data frame will have five columns: population size, rates of diabetes, rates of obesity, nutrition level (defined as the percent of people eating less than one fruit or vegetable a day), and rates of physical activity (defined as “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).

## Describe the rate of diabetes in the US -Katejoe.

To answer our first question, we will create a chart (bar graph or map) to show the rates of diabetes in the different states.

Create the top and bottom state diabetes rate in the U.S

## Compare rates among US regions. Farrukh

We will group the data into four regions to answer the second question. We will then create bar charts comparing the rates of the different variables in the four areas.

## Inspect relationship between nutrition, physical activity, obesity, and diabetes - Katejoe.

We will calculate Pearson correlation coefficients between each pair of variables to answer the third question. We will determine the strength of the relationships based on the magnitude of the numbers we get.

## Predict diabetes rates based on nutrition, physical activity, obesity, and diabetes -Liat.

To answer the fourth question, we will create three scatterplots: obesity vs. diabetes, physical activity vs. diabetes, and nutrition vs. diabetes. We will analyze linear regression and add the regression line to our charts.