DATA VISUALIZATION PROJECT

‘Visualizing Obesity in Massachusetts’

- **Chitra Pichaimuttu Kanickaraj**
- **Tabassum Kakar**
- **Suwodi Dutta Bordoloi**

Overview and Motivation:

Obesity is a major issue, more than 34% of US Adults are obese and it is dangerous as it causes severe diseases such as heart stroke, diabetes and various types of cancer. Through this project we want to highlight the some of the key factors that are known to cause obesity, how obesity is prevalent in the population and details about obesity programs available for the states - which would help the government in decision making to plan programs to fight obesity and allocate budgets to these programs.

Related Work:

We were inspired by the State obesity prevalence maps[1] provided by the Center of Disease Control and Prevention (Figure 1) and wanted to dig further in order to understand and investigate the factors such as socioeconomic status, access to grocery stores and restaurants etc. effecting or causing obesity in these states. Moreover we want to build a tool that provides insights that ultimately helps agencies in decision making processes in an effective way through visualizations to tackle obesity.

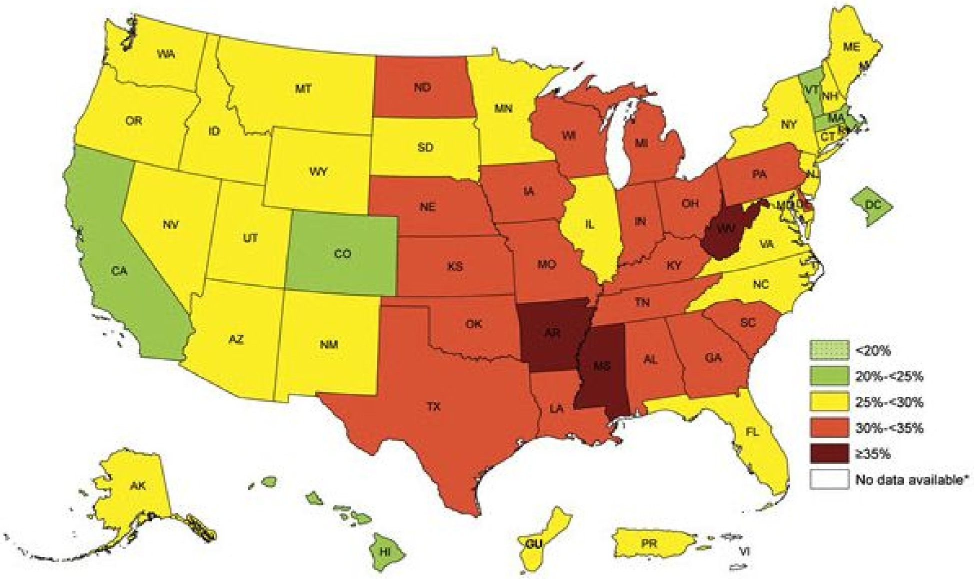


Figure 1

Questions:

The objective of our project is to understand and explore which counties within Massachusetts are mostly affected by obesity and what are the factors such as socioeconomic status, income and physical activity etc. contributing to obesity. We want to explore the factors affecting obesity using visualizations and showing some statistics of the distribution of obesity across different counties and people's access to different facilities such as grocery stores etc. The primary objective of the project is to build visualizations that will finally help state government to understand the counties and demography most prone to obesity in Massachusetts, how successful have the past programs been and to help in decision making process for budget allocation.

Data*:*

The dataset for this study has been chosen from the United States Department of Agriculture (USDA) Economic Research Service portal. This data is collected from a variety of sources and cover varying years and geographic levels. It comprises of about 168 predictors categorized across 9 different categories.

01. *Access and Proximity to Grocery Store*: It provides an overview of community's access to grocery store.
02. *Store Availability*: It provides numbers of supermarkets, grocery stores, supercenters and club stores, and various other establishments that are primarily engaged in retailing a general line of food.
03. *Restaurant Availability and Expenditures*: It provides numbers of limited-service restaurants and full-service restaurants along with their average expenditures on food purchased.
04. *Food Assistance*: It gives us statistics as to how providing nutrition assistance has helped people make healthy food choices and the effect of economic incentives on making informed decisions about diet quality.
05. *Food Insecurity*: It provides percentage of prevalence of household-level food insecurity which affects the food intake of household members and disrupted eating patterns because of insufficient money and other resources for food.
06. *Food Prices and Taxes*: It provides statistics on price and sales tax of few food items like milk, soda, chips and pretzels to help compare their prices with each other and other general food items.
07. *Local Foods*: It provides information on locally sourced and locally available foods and the programs used to enlighten the nutrition education of these farms.
08. *Health and Physical Activity*: It provides information on recreation and fitness facilities along with the diabetes and obesity rate among different age groups categorized by income .

09. *Socioeconomic Characteristics*: It provides a census of population based on race, age group, poverty rate etc.

Data cleanup -

One of our main problems with the dataset is the large number of predictors and many of them are highly correlated because there are multiple variables from the same category. Variables represent the attributes in percentage, percent change between two years as well in absolute numbers. We had to perform attribute selection (both apply statistical techniques) and manually to drop redundant variables. To address missing value problems, we have replaced them with mean or median in many cases. Also there are some data inconsistencies and repeated records that needed to be removed.

Exploratory Data Analysis:

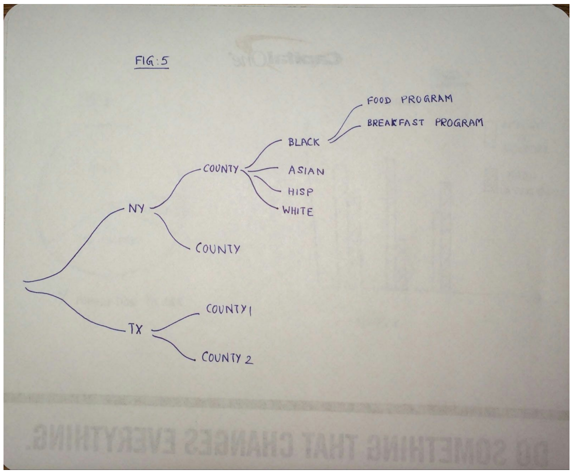
We used Weka and Orange (Python) to explore the data set and started with basic plots like pairwise scatter plots comparing each pair of predictors and explored the distribution of each of the attributes. It was very helpful to see the predictors which are highly correlated within themselves. We also found that some of the variables like accessibility to stores and local foods have a very skewed distribution.

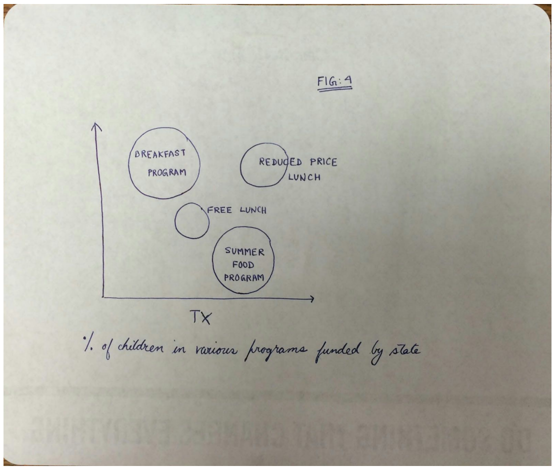
From the scatter plots, we saw child poverty, number of participants in free lunch, number of participants in free breakfast programs, accessibility to groceries, etc are all correlated with obesity. Another insight was that poverty rate is positively correlated with Black and Hispanic population (while it isn't true for while, asian or other races) as well as the benefit programs (like free lunch, snap participants, etc). Convenience stores have a positive correlation with adult obesity. We also found that snap benefits per capita is correlated with convenience stores - which is not very evident. But we saw that lower income households do participate more in the snap or other food assistance programs.

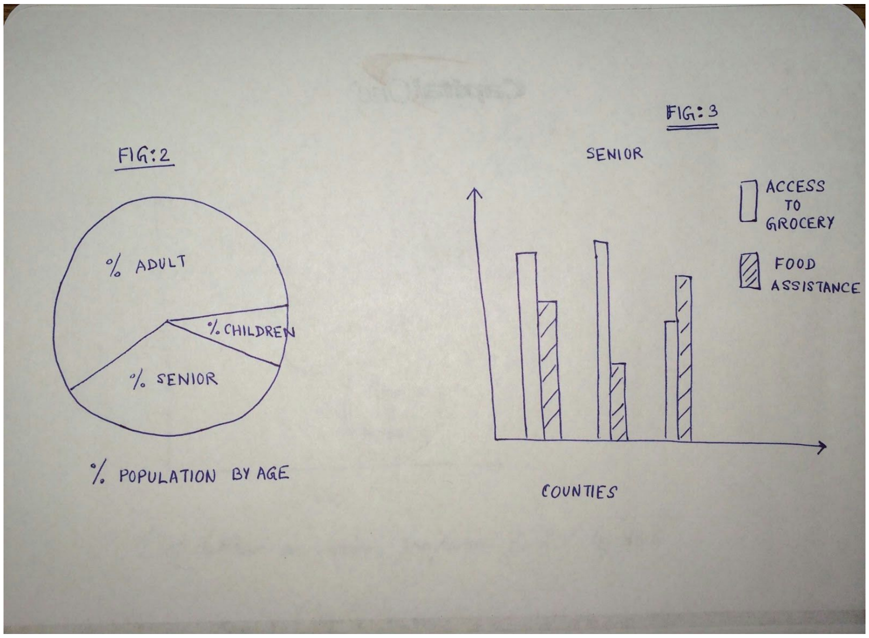
The insights gained were very crucial since we selected the variables correlated with obesity as some of our key factors for visualization.

Design Evolution:

We considered various visualizations for displaying our data such as maps, bar-charts, pie-charts, bubble-charts and treemaps, but we selected only maps, bar charts and pie charts as they are well suited for the type of analysis we intend to do. Below are the initial sketches of our design.







We chose map to show the obesity density distribution across Massachusetts across counties. The color density is mapped to the obesity density, i.e. darker color implies more obesity whereas lighter color implies less obesity. User can select a county and view the obesity distribution across different races and age groups through bar charts.

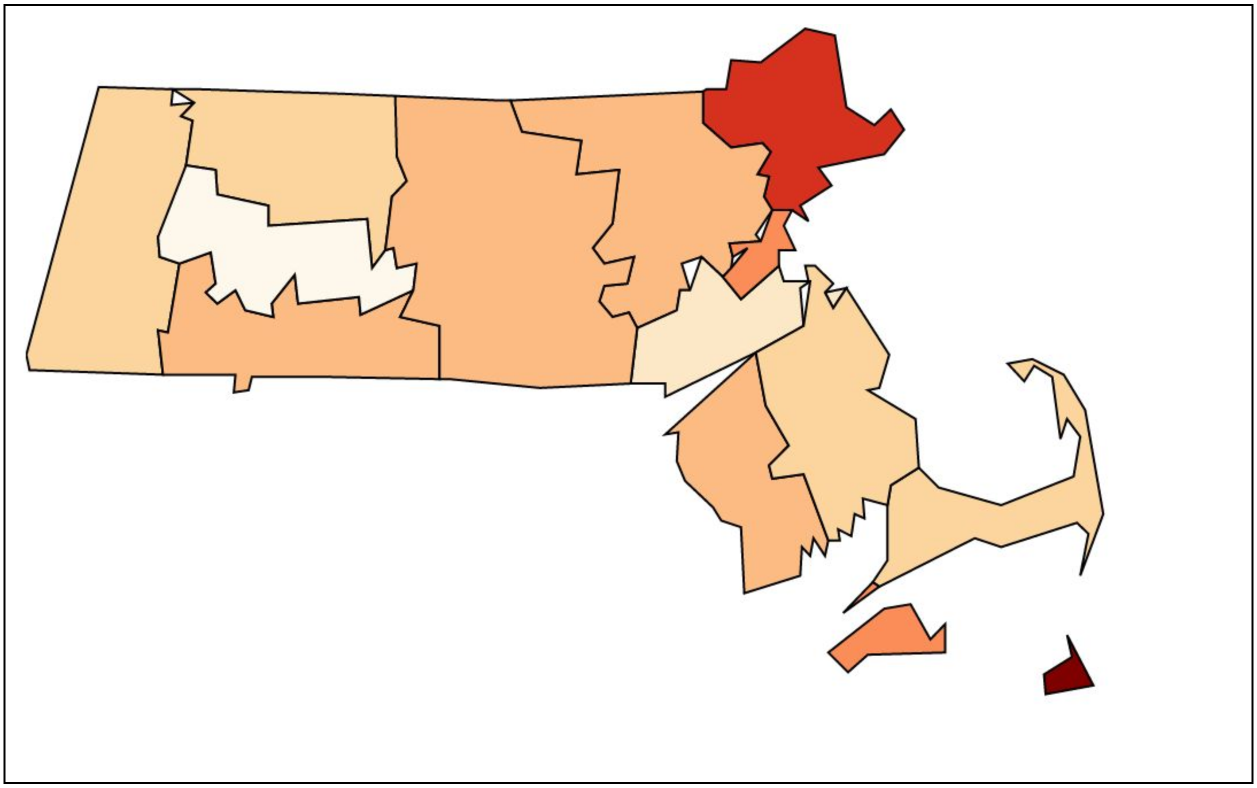
At first we planned to implement a US state-level overview of the obesity distribution. But as per Professor's suggestion and our thought of local health officials being more accountable for controlling obesity, we settled to show obesity per state across each of its counties.

Implementation:

We implement the visualizations such that with each additional visualization, the user has an additional layer of information about a county that ultimately helps in learning the prevalence of obesity, how the factors affecting obesity are prevailing in the counties, and the government funded programs. At each level, we also provide a reference to compare to (for example either the state average or the most/lease obese county).

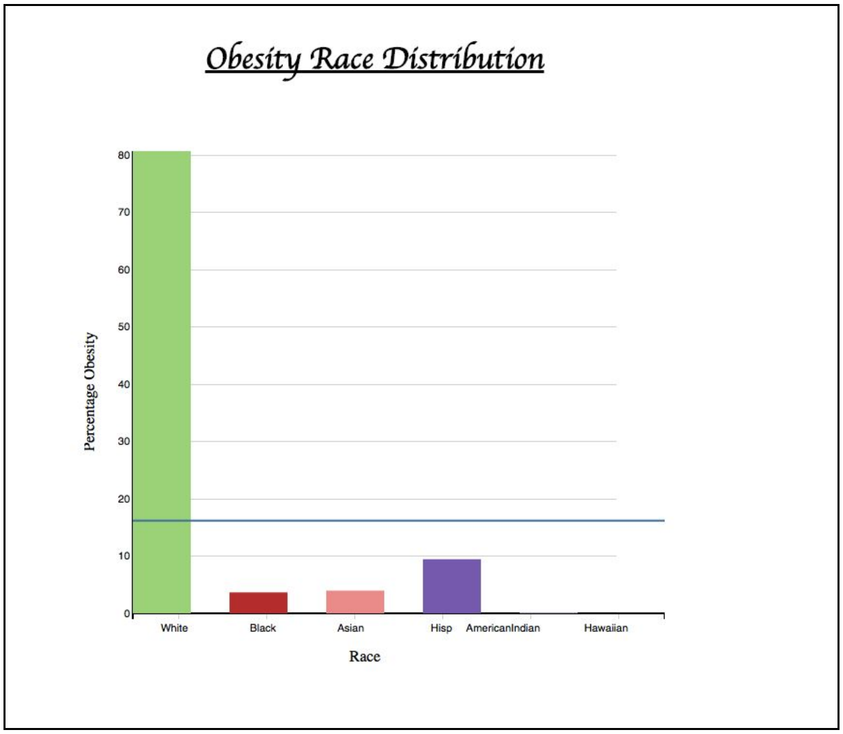
Visualization 1: MA Interactive Obesity Map

For the year of 2014, we show the obesity map for MA counties with low to high obesity percentages represented by the color gradient. The user will be able to select any county, and the following visualizations show details for that county only.

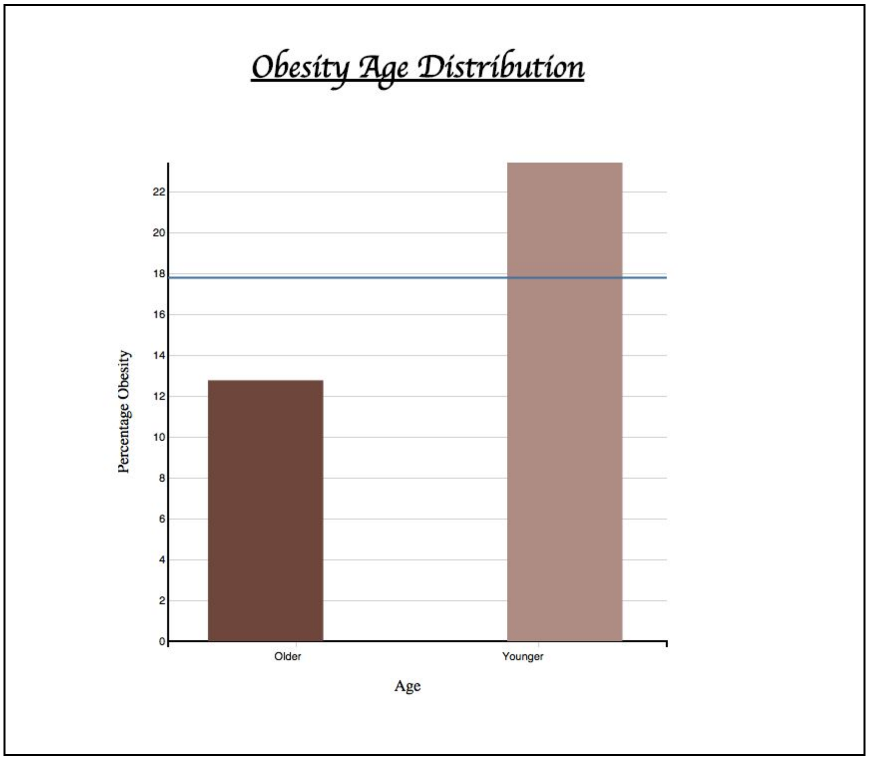


Visualization 2: Obesity Race Distribution

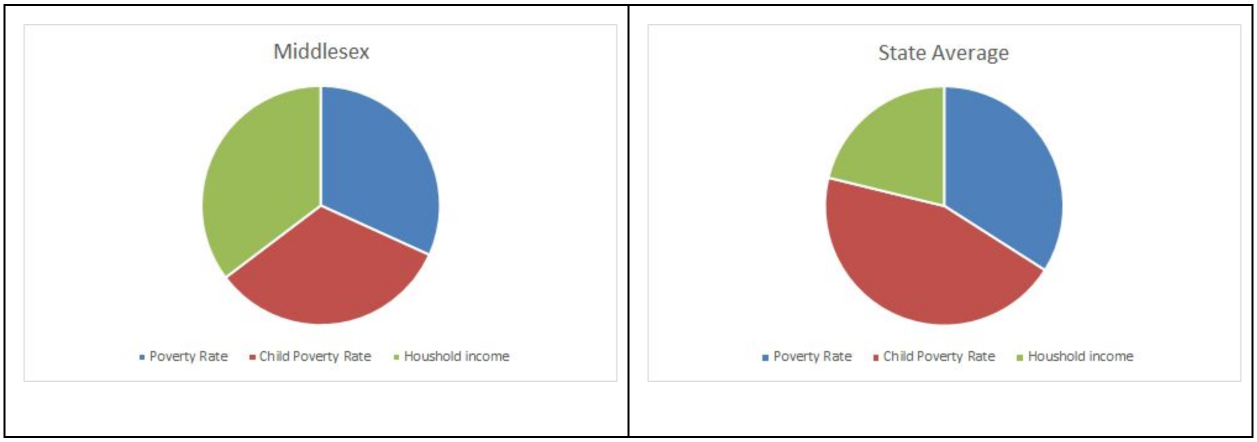
For the county selected by the user, we show the demographics of the county in the barcharts (Obesity Distribution by race and age) for that specific county. We will also show the average of the obesity for race and age respectively for entire state of massachusetts for comparison. Our intent is that for each selected county, the user gets an overview of the demography and socio-economic conditions of the county and is able to compare with the state average.



Visualization 3: Obesity Age Distribution

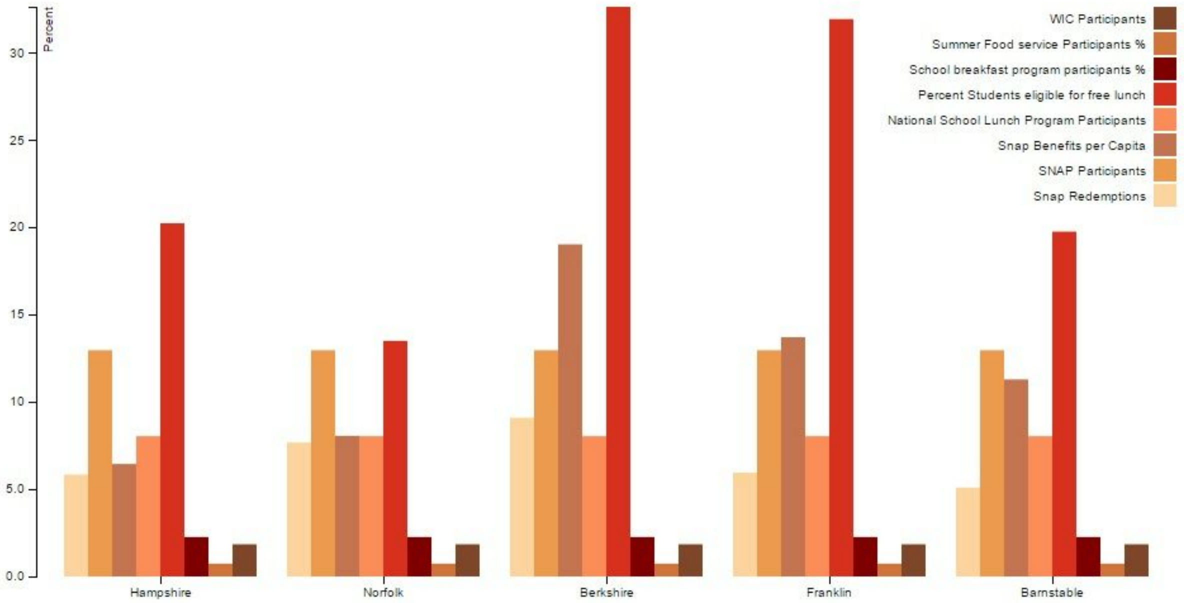


Visualization 4: Pie charts showing poverty rates and income of state versus county



Visualization 5: Interactive grouped bar chart

To learn about how the government funded programs are distributed across counties, we display an interactive grouped bar chart. The user will have an option to compare the selected county (from the map) with either the top four obese or bottom four obese counties. This will provide a definite view of how the programs are being funded in other counties as opposed to the particular county.



Visualization 6: Interactive grouped bar chart showing obesity factors

Visualization 6 will be very similar to visualization 5, but it will compare the factors (strong predictors of obesity) such as food insecurity, health (highschool activity), access to grocery stores and available fast-food or full restaurants in the county.

Evaluation: Looking at the distribution of obesity across different races in counties within Massachusetts we have learnt that majority of the obese people are white, but we have to consider that the majority of the population is also white. The average line across different races in Massachusetts is helpful in getting insights about the obesity within races and how it compares to each county. On the other hand the obesity across the age groups tells a sad story, that is, except for only one county (Barnstable), majority of the obese people are younger, i.e. less than 18 in rest of all counties. This can drive the state government's attention towards planning programs targeting younger people in order to help them tackle obesity in younger generation.

The grouped bar chart showing programs across counties show that most of the highly obese counties have lower percent of students who are eligible for programs like free lunch. Same is true for other programs like SNAP redemption and other free breakfast school programs, i.e., there is a lower eligibility and number of participants in more obese counties. This highlights a very serious problem that the obesity programs are somehow falling short. We need to investigate why the affected population is unable to participate in these programs and how they can be customized to target obese people from all backgrounds and sections.

In the future, our design can be improved by including state map, so that user can select any US state, and drill down to the counties within that state and get insights about the obesity distribution for each county. In addition to state map we can extend this work to multiple years and provide the user an option of selecting data from a particular year and comparing across different years.

References:

1. State Obesity Prevlance Maps: http://www.cdc.gov/obesity/data/prevalence-maps.html
2. http://bl.ocks.org/mbostock/5872848
3. http://bl.ocks.org/mbostock/3887051