Project 1 Report
Determining Correlations Between United States' Health and Income Data

ECEN 689: Applied Information Science Practicum - Team 7

Course Instructor: Prof. Jean-Francois Chamberland

Team 7 Members

1) Khaled Nakhleh
2) Shabarish Prasad
3) Fazia Batool
4) Venkata Pydimarri

Due Date: October 30th, 2018

Abstract

The goal of this project is to support the hypothesis ”Income level is positively correlated to health”. We have used the diabetes and obesity rates of people in a region as an indicator of the health of the people in that region. We have assessed the data for income, diabetes and obesity for the years 2008 and 2013 to evaluate this hypothesis.

Introduction

This project course involved gathering health and income data for the United States. The project's main objective was to determine if the health indicators considered (obesity, diabetes) were in any way correlated with income of people. The project's goal was also to better understand the relationship between health measurements within the broader US social classes, and the classes' income disparity proven in previous academic work by Bryan and Martinez [4].

Going through the literature, one can see an established truth about income and health: the higher an individual's income, the better his or her health [1]. However, a broader scope is needed when considering how those parameters are measured county-wise. Not to mention the perceived health evaluation individuals have about their life. We highlight the importance of rationally choosing measurement variables for county-level evaluation. Previous work recommends the variables should be dependent on the relevance of the variable to the model being studied, as L Shavers published [5]. We want to also acknowledge other factors that we believe affect a county's overall health. Education level was also seen to affect health quality in US social classes by Marmot [6]. More discussion is offered in the results and conclusion sections.

Life evaluation refers to the thoughts that people have about their life when they think about it [2]. While the literature did prove a correlation between perceived mental health and income on an individual level, it also proved no tangible progress beyond an annual income of ~$75,000 [2]. Mental health analysis and observations were not included in this work due to time constraints. We recommend future work to incorporate mental health evaluation into their findings.

We highlight how previous work by Wagstaff and Doorslaer [3] showed that it cannot be decisively concluded how income inequality affects individual health from population-level studies. Our used data is classified as a population-level data, hence, the results here cannot firmly prove a relationship between social class income inequality and health levels.

So, going into this project we would like to keep our view clear of all these previous studies and come up with a hypothesis of whether there is a correlation between health and income based on only on factual evidences from the data that we have in our hand. The next sections of this report provide graphs, data visualization, and sources that give better understanding of the health-income relationship in U.S. counties. References are provided at the end of this report.

Data Sets

We used data from three different sources in this project.

The first dataset is from "**Food Environment Atlas".** Food environment factors interact to influence food choices and diet quality. The objectives of the Food Atlas, made available by the U.S. Department of Agriculture, are to assemble statistics on food environment indicators and to provide a spatial overview of access to healthy food. The atlas contains health and well-being indicators such as diabetes and obesity rates.
The current version of the Food Environment Atlas has over 275 variables, including new indicators on access and proximity to a grocery store for sub populations; an indicator on the SNAP Combined Application Project for recipients of Supplemental Security Income (at the State level); and indicators on farmers' markets that report accepting credit cards or report selling baked and prepared food products. Data was used for the years 2008 and 2013.

The second source is from the **Internal Revenue Service (IRS)** of the United States Federal Government**.** The IRS is responsible for collecting taxes and administering the Internal Revenue Code. They publish their individual income tax statistics every year on their website. For the purpose of this project we are using the statistics they publish for every ZIP zone in the USA for the years of 2008 and 2013, since we have our health indicators data published by Food Environmental Atlas for those two years only.

The third dataset is "**US Zip Code to Country State to FIPS Lookup**", which was created to help users to go between County - State Name, State-County FIPS, City, or to ZIP Code. This dataset was the result of a project by a user named Nic Colley on data.world. He used the data from the Census Bureau to map each ZIP zone with the FIPS of that region. Most importantly, this dataset was created because we shouldn't have to pay for free & public data. So, basically we do the following steps:

1. Use a ZIP Code to find County, State string, or FIPS.
2. Find the ZIP Codes within a County or FIPS boundary.
3. Add City information to the FIPS data.

Methodology

We are trying to analyze the income vs health data for the years 2008 and 2013. The data containing the health indicators contains the average rate of obesity and diabetes per county for all the counties in the USA, whereas the data we collected from the IRS publication contains the average income of every ZIP zone and the total number of people who have filed their tax reports in that zone.

To specify the indicators of income, we have the average adjusted gross income (AGI) of every ZIP zone as a measure. But since we have obesity and diabetes rates as average over the population of every county, looking on population belonging to different income groups may be more efficient than looking at the AGI vs the health indicators. We make this assertion in the assumption that people may put their personal hygiene first, and if they earn enough to support their own personal hygiene, they would do it.

The IRS publication for each ZIP zone contains the information about the number of people from different income classes (AGI classes) who have filed their tax reports. Their report consists of eight different income classes and these classes include:

1. Under $1
2. $1 under $10,000
3. $10,000 under $25,000
4. $25,000 under $50,000
5. $50,000 under $75,000
6. $75,000 under $100,000
7. $100,000 under $200,000
8. $200,000 or more

In order to have much more meaningful interpretation, we would like to further bin these groups into only three groups, the low, medium and high income groups. Since we find in the data that a few people have their annual income under $1 and still managed to file a tax report, we assume they should be dependent of some people from one of the other three income groups. And so we group these people as dependents. And to bin the other seven classes into the low, medium and high income groups, we considered a source from Business insider[7], which says that the average income of an American is $58,000. In the dataset have, we have the upper bound of the fifth class as $75,000, and so have decided to consider this class as the upper bound of the medium income group. We consider only the second class as the low income group and the upper three classes are considered high income group. So we have decided to come up with only four groups in the end, and these groups are:

1. Dependents: Under $1
2. Low Income : $1-$10,000

3. Moderate Income : $10,000- $75,000
4. High Income : >$75,000

Now that we have decided on the indicators of health and income. We can analyse how they stand against each other. But the only problem is that we have the income information in every ZIP code region and the health indicators are averaged over every county. We have to translate the data in ZIP code region to county. Since we are considering the number of people in the different groups, the translated data will just be the total of total of these numbers over all the ZIP regions in a county. We can use the crosswalk dataset to map each ZIP location with its corresponding FIPS id of the county and sum up all the information under one FIPS id.

After mapping the income data to the health indicators, we have considered the ratio of the low income and high income groups over the total number of tax reports filed in every county and tried to come up with meaningful correlations between these ratios and the health indicators. We have discussed our findings in the next section.

Discussion

The following results show both the visual relationship between obesity, diabetes and the Adjusted Gross Income for the year 2013. The visual graph, showcasing data projected onto the United States map, was made using Plot.ly. Polt.ly is a data visualization software with packages available for Python. The package was utilized for generating the maps based on FIPS codes. Several graphs were generated for the different parameters we gathered.

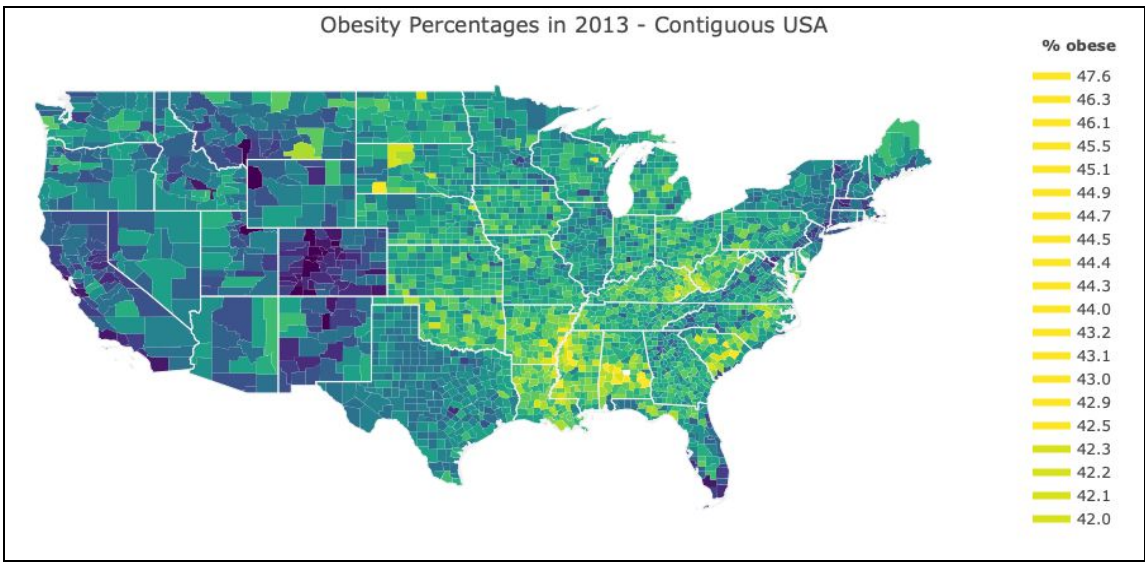


Figure 1: Obesity percentages for 2013

.

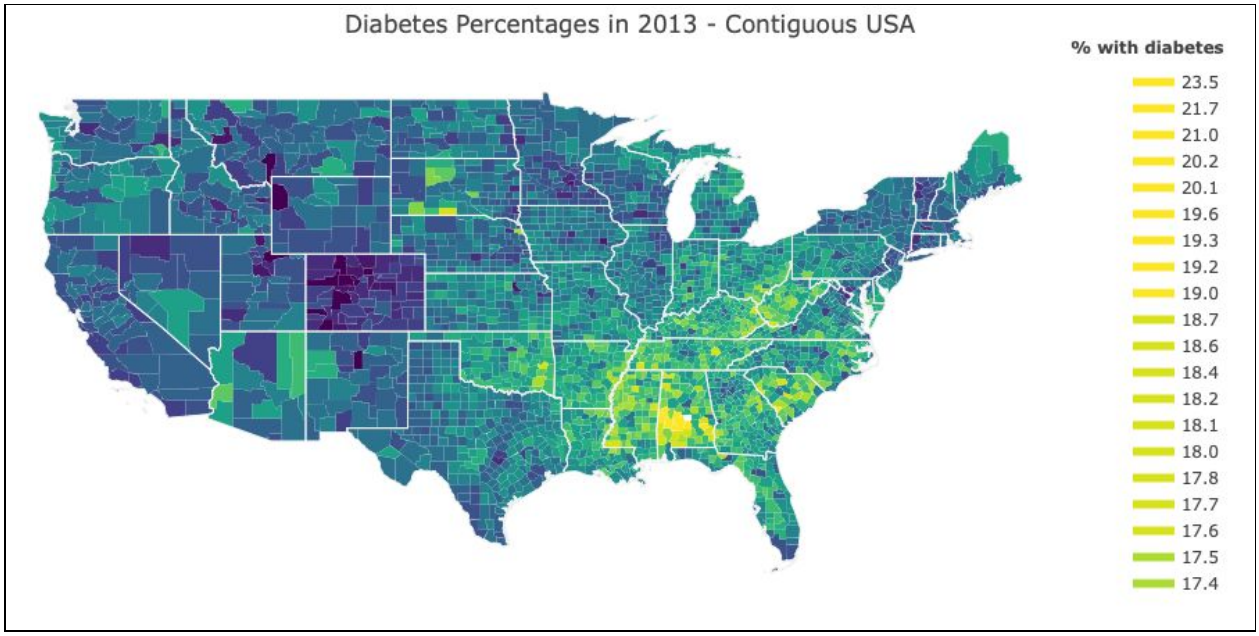


Figure 2: Diabetes percentages for 2013

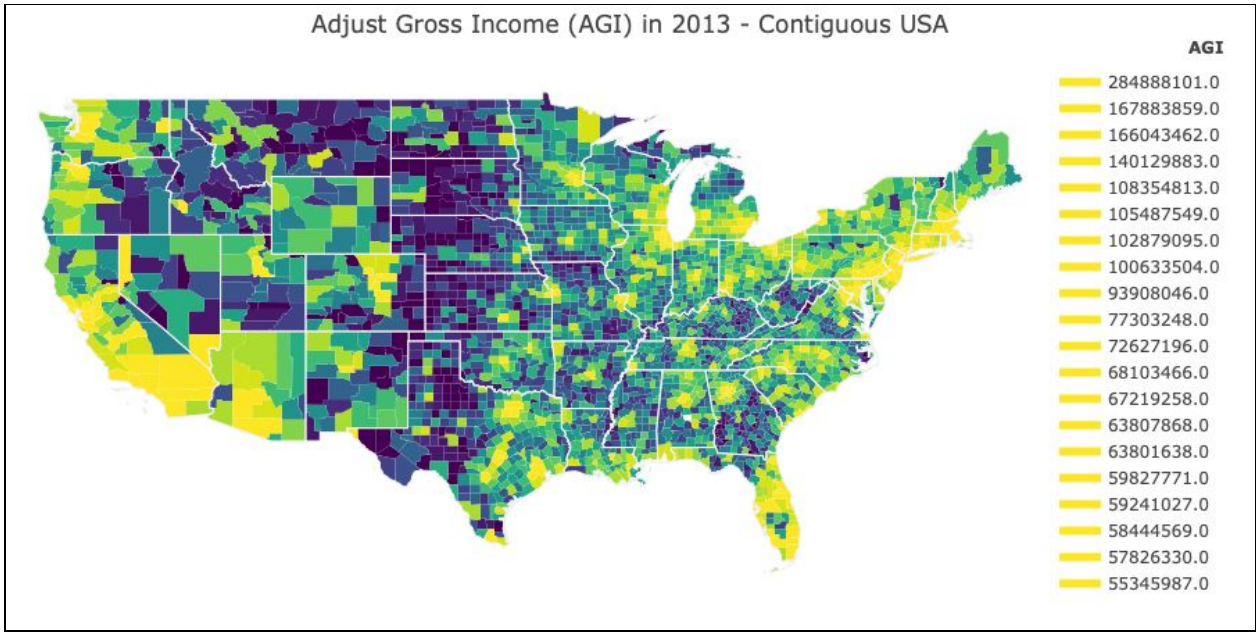
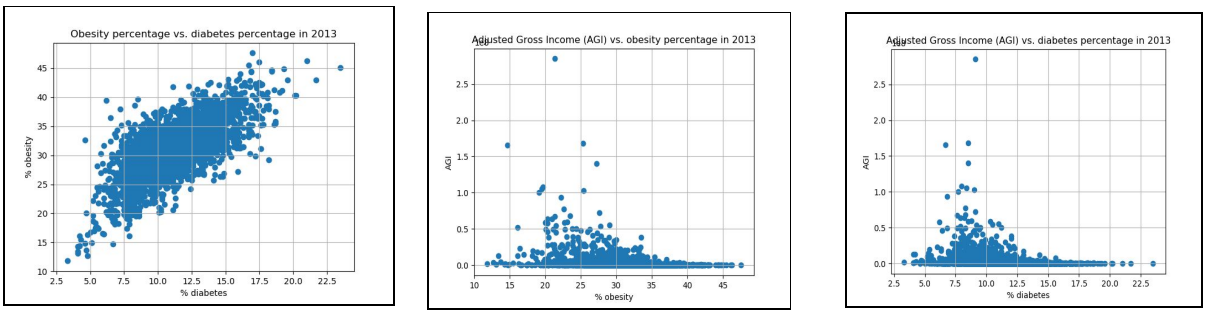


Figure 3: Adjusted Gross Income (AGI) values per county for 2013

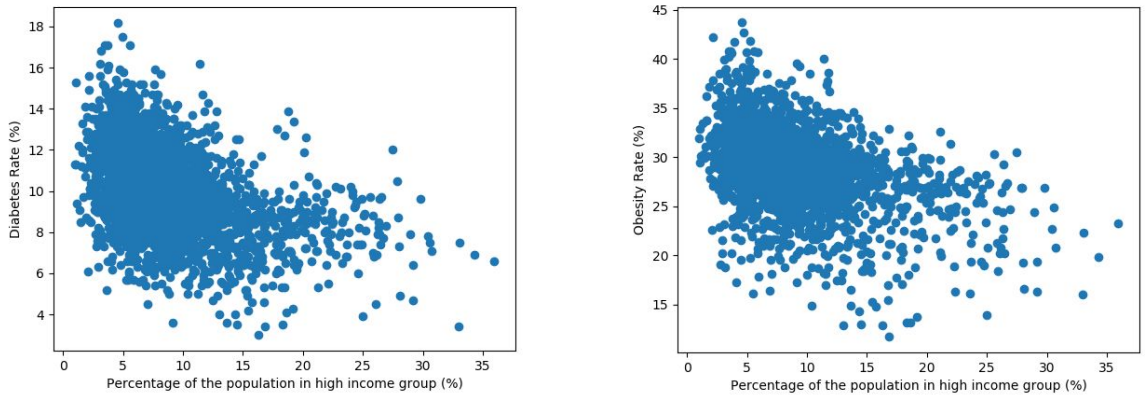
From these maps, we can find a little to no correlation between either of the health indicators and the AGI. But we can find good correlation between obesity and diabetes. The correlations are much more evident when we plot the values in a scatter plot as shown in the following graphs.



(a) (b) (c)

Figure 4: (a) Obesity vs. diabetes percentages, (b) Adjusted Gross Income (AGI) versus obesity percentages, (c) Adjusted Gross Income (AGI) versus diabetes percentages for 2013

So, we are not able to make any meaningful interpretations by looking at the AGI vs health indicators, as expected. On the other hand, making a scatter plot between the ratio of the high income group to the population in each county and the indicators of health for the year 2008 we see some interesting trends.



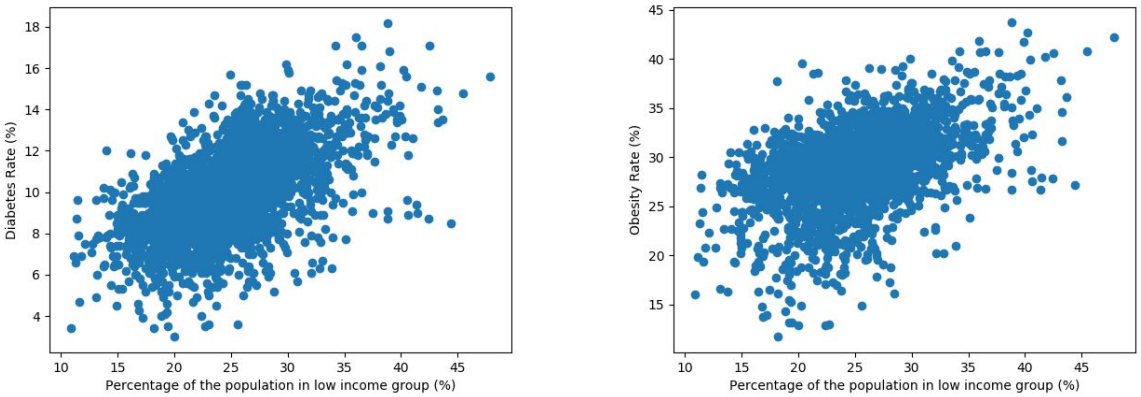
(a) (b)

Figure 5: (a) Diabetes Rate (%) versus Percentage of population in high income group (%) in 2008. (b) Obesity Rate (%) versus Percentage of population in high income group (%) in 2008.

We can visibly see a negative trend on both the graphs. The correlation coefficient between the variable in the graph in the Figure 5 (a) is -0.4129, whereas for Figure 5 (b) it is -0.4152. This indicates that as the ratio of the number of higher income people in a county

to the population of the county increases, both diabetes and obesity rates in the county tend to decrease.

It gets even more interesting, when we look at the scatter plot plotted between the ratio of the number of people in the low income group in every county to the population of that county and diabetes and obesity of the same year. These graphs are shown in the figure below.

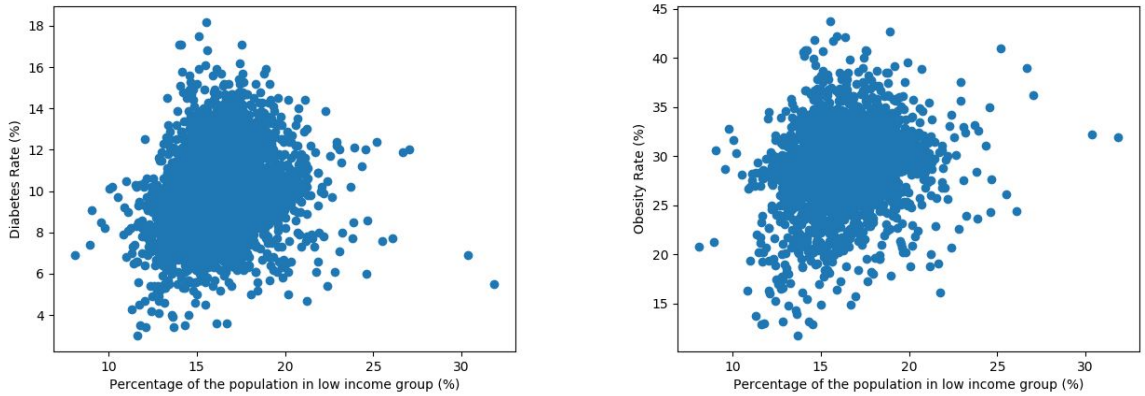


(a) (b)

Figure 6: (a) Diabetes Rate (%) versus Percentage of population in low income group (%) in 2008. (b) Obesity Rate (%) versus Percentage of population in low income group (%) in 2008.

The graphs in the figueres, Figure 6(a) and Figure 6(b) show a positive trend. The correlation coefficient of these graphs are 0.5883 and 0.4540 respectively. This would mean that this ratio is more positively correlated to the health indicators than the negative correlation shown by the ratio of high income group people in a county towards these variables. So these graphs imply not only the diabetes and obesity rates tend to take a smaller value when a county relatively has higher number of high income group people, but also that the counties tend to show a higher obesity and diabetes rates, when there are relatively higher number of low income group people live in that county for the year 2008.

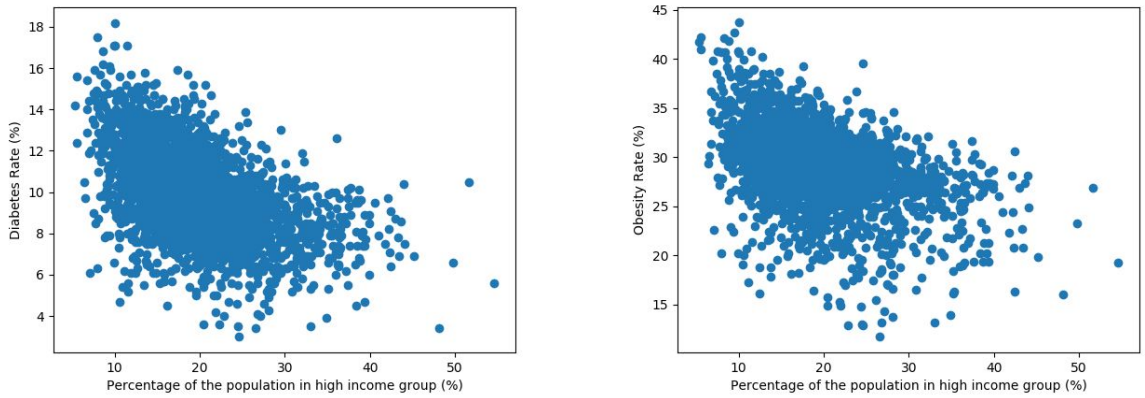
While we look on the same graphs for the year 2013, these statistics have drastically changed. The heavy positive correlation seen between the ratio of low income group people in a county and the health indicators has dropped drastically. While we were able to visually see the trend in the scatter plot for 2008, the case is not the same for 2013. These graphs can be seen in Figure 7, shown below.



(a) (b)

Figure 7: (a) Diabetes Rate (%) versus Percentage of population in low income group (%) in 2013. (b) Obesity Rate (%) versus Percentage of population in low income group (%) in 2013.

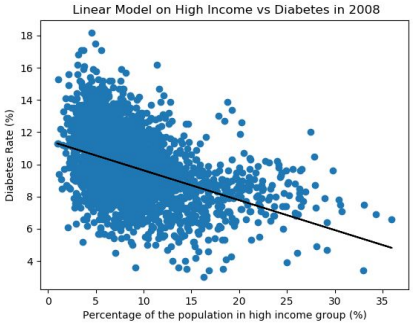
The correlation coefficient of these graphs are 0.2098 for diabetes and 0.2011 for obesity. We can interpret from these values that, even though the graphs do not visually show it very well, a positive trend is still seen between these variables. The graphs for the ratio of high income group people and diabetes and obesity is shown in Figure 8.

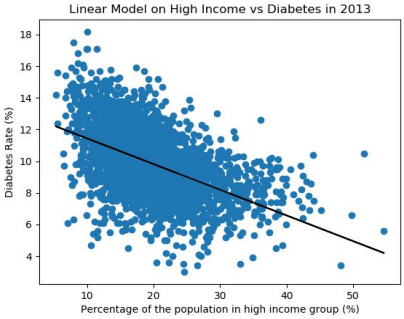


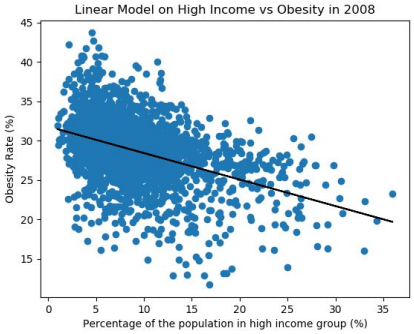
(a) (b)

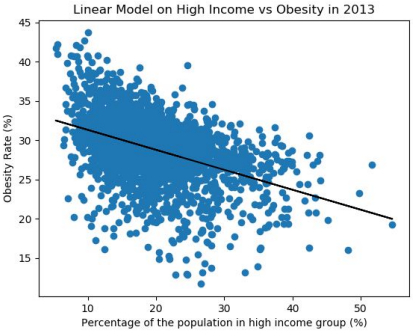
Figure 8: (a) Diabetes Rate (%) versus Percentage of population in high income group (%) in 2013. (b) Diabetes Rate (%) versus Percentage of population in low income group (%) in 2013.

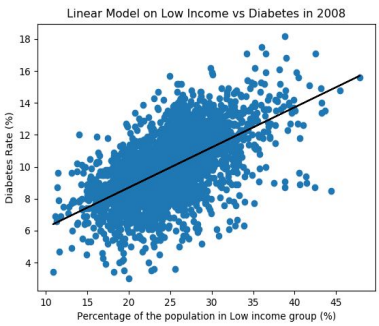
Figure 8 shows that the strong negative correlation seen between the ratio of the high income group people in a county to the health indicators in 2008 still exist and now even show higher correlation coefficients (-0.5032 for Figure 8(a) and -0.4368 for Figure 8(b)) in 2013.

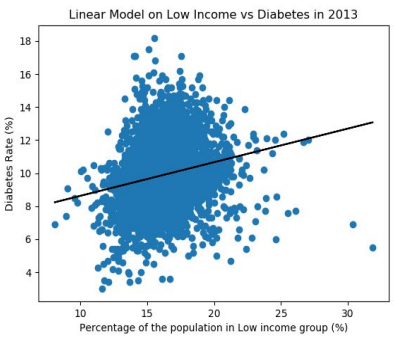


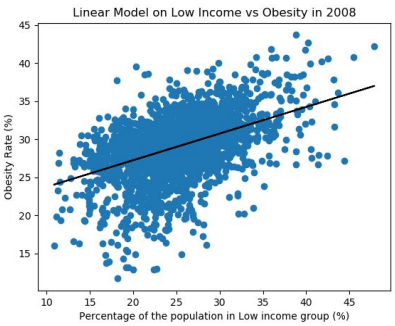












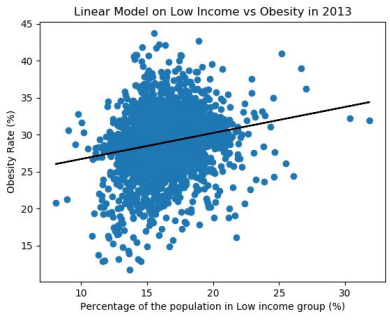


Figure 9: Linear model fitted between the variable in each scatter plot.

We also fitted a linear model on the health indicators over the income groups for both the years, and those results can be seen in the Figure 9.

The conclusion section offers more insight on these graphs, as well as recommended future areas of work

Results & Conclusion

Based on the generated graphs, and analyzed datasets, although no measurable correlation was found between the two health parameters (Obesity, Diabetes) and the Adjusted Income Level (AGI), a considerable correlation was found between the income groups and these parameters. These correlations indicate as if there are a relatively higher number of high income group people in a county, the diabetes and obesity rates tend to stay lower in that county and the vice versa is also true. These trends are seen in both the years of 2008 and 2013.

The statistical models we fitted on these variables also support this statement. Although there is a noticeable change in the scatterplot between ratio of low income group vs the health indicators moving from 2008 to 2013, the correlations between the health indicators and ratio of high income groups still holds good. In addition, this graph even shows a more stronger correlation as compared to 2008.

Instead of incorporating the literature review in the scenario and based on the factual evidence alone that we gathered from the above mentioned sources, we can conclude that a person's health tends to get better if there is a raise in his/her income.

One recommended area of improvement is to include statistics of minorities population percentages in the United States that affect the health levels in a county. Observation of race/minorities income levels and population percentages is also a good recommendation. While this report did not list the race population percentages, we conclude that a deeper analysis is required to determine whether race affects the health level or income level. Other factors to consider are: Gender, Immigration status, Education level, Number of family members, and Age.

Another improvement area is to include more health and income parameters in the study. In this study, we only analyzed obesity, diabetes, and Adjusted Gross Income (AGI) data parameters, and we only described their correlation on a county level.

According to our study and results, for the high income group, there is a negative correlation with the health indicators (Obesity rate and Diabetes rate). And for the low income group, there is a positive correlation.
Also, the correlation was more evident in 2008 which changed in 2013 for low income groups. A possible explanation behind this change would be that people in high income group can afford healthier lifestyle and are aware about the impact it can have on their life. They can afford to spend extra dollars in organic produce which happens to be a bit more expensive.

References

[1] Kawachi, I., & Kennedy, B. P. (1999). Income inequality and health: pathways and mechanisms. *Health service research*, 34(1 Pt 2), 215-27.

[2] Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 107(38), 16489-93.

[3] Wagstaff, Adam & Doorslaer, Eddy. (2000). Income Inequality and Health: What Does the Literature Tell Us?. *Annual review of public health,* 543-67.

[4] Bryan, Kevin A. and Martinez, Leonardo, On the Evolution of Income Inequality in the United States (2008). *FRB Richmond Economic Quarterly*, Vol. 94, No. 2, Spring 2008, 97-120.

[5] L Shavers, Vickie. (2007). Measurement of Socioeconomic Status in Health Disparities Research. *Journal of the National Medical Association*. 99, 1013-23.

[6] Marmot, Michael. (2002). The Influence Of Income On Health: Views Of An Epidemiologist. *Health affairs (Project Hope)*. 21, 31-46.

[7] Tanza Loudenback. (2017). Middle-class Americans made more money last year than ever before. Article by Business insider.
https://www.businessinsider.com/us-census-median-income-2017-9