Max Tjen

CSDS 234

Project Final Report

**Obesity: Comparing County Prevalence to Median Family Income**

**Introduction**

A problem that I wanted to investigate was the relationship between an area’s median family income and their obesity prevalence. When I first began my project, I was curious about the relationship between food deserts and obesity, but found that a lot of data regarding supermarket location was not the most accurate or usable without subject expertise. Along the way, I found that there appeared to be an inverse relationship between obesity and median family income, and so I decided to continue down that road. I was focused on the idea of relating obesity to easily usable data because of the growing obesity issue not only in the United States, but around the world. While obesity rates have been steadily growing with each passing year, there are many factors that have been identified as causes, such as poor diet, physical inactivity, and increased meal portions. Furthermore, rising obesity rates are problematic because of its residual effects, such as increased risk of high blood pressure, diabetes and heart disease. All of these effects essentially increase the chances of unhealthy living conditions and premature death.

**Related Work**

With the increasing issue of rising obesity prevalence around the globe, there have been many studies done involving various factors. Obesity’s relationship with income has been investigated on numerous occasions, with some study findings contradicting others. In a study by the Centers for Disease Control and Prevention (CDC), they investigated obesity prevalence as a whole, relative to income as well as prevalence by gender. Their results found that for men, income level did not have much of a relation with obesity, as they found that ‘obesity prevalence is generally similar at all income levels’. For women though, the study concluded that as income decreases, obesity prevalence increases, however most women who are obese are not from lower income households (Ogden). This study is relevant to my project because while I did not break down obesity prevalence by gender, it somewhat summarizes the relationship. It demonstrates how through their studies, the CDC seemingly found a relation between decreasing income and increasing obesity. This also brings up potential further research in breaking down my data by gender and looking at each one individually.

Similarly, an article by Science Daily cited a study by the CDC that found a correlation between income and obesity does exist. This is a different CDC study than before, as it was done nearly ten years afterwards and concluded that there was a ‘negative correlation between household income and both obesity and diabetes’. The rates of diabetes and obesity in the United States have been steadily rising in the poorest regions since 1990 and by 2015, it was concluded that ‘members of lower income households had a much greater chance of suffering from obesity and diabetes’ (R. Alexander Bentley). The findings directly relate to my project in that they correspond to what I am investigating in comparing the factors and its relation. Furthermore, it cites specific states and areas that strongly exhibit the study’s conclusions.

**Method**

I began my project by creating a new feature of average prevalence, which was simply the average obesity prevalence between males and females. After that, I used individual county maps of median family income and average obesity prevalence. In doing so, a major problem arose that would affect comparisons, as the scale of prevalence and income values were very different. Although they were on different scales though, the continuous scale of values for each of the counties allowed some comparison as to which counties were on the high and low side for median family income and obesity prevalence. With the varying scales, I had to figure out a way to look at the inverse relationship between them equally without losing the importance of individual values. In essence, I wanted to create a new feature that would easily visualize how inversely proportional the median family income of a county is to their obesity prevalence.

To accomplish this, I decided to create percentiles for median family income and obesity prevalence and then afterwards, get the absolute value of the difference. This would allow me to investigate the inverse relationship as if the percentiles of features were very different (ex. 5, 90), the difference would be high. Conversely, if the percentiles were very similar, their difference would be low, which would accurately reflect the inverse relation of the variables. To ensure that null values did not influence the percentile scores, I removed rows (individual counties) that were missing obesity prevalence values. I also made sure that when finding the percentiles, the scoring was strict, so only values that are less are counted, rather than the more typical method of also counting values that are the same. The percentile maps for both median family income and obesity prevalence were very similar, but now both of the data sets were on the same scale. As such, I could create a map of the percentile difference, which would easily show the inverse relationship.

Given that the United States as a whole is very big and includes many counties, I wanted to create a better way to visualize and compare areas of interest. To do so, I wanted to make my maps interactive, beginning with making each state and county hoverable. In doing so, if the user moved their mouse over a county or state, important data would pop up, such as the obesity prevalence by gender or the demographic breakdown. Furthermore, to increase ease of access to individual states, I created a dropdown menu on the county map which allows the user to select a particular state of interest and only have that state trace be visible. With that, it becomes easier to visualize certain areas and outside noise is blocked out, which introduces more potential research and functionality in the future.

A final thing that was very important for my data preparation and contributed to my percentile calculations was my way of correctly matching counties. Because I had two datasets from different sources, numerous counties were labelled differently between the two in terms of their suffixes. In order to do so, I had to create a function that checked many aspects between the two county values. It began by comparing state values, splitting the county into individual words, and then keeping track of the last word. With that, I could see if the last word was a suffix that was noted in a list of unique words that were following the counties (ex. Borough or Municipality). Once all of those checks were completed and the values passed, I would change the county value without the suffix to be the same as the value with the suffix so that the two dataframes could be merged based on the county and state values being the same. Although my function worked well and matched all of the appropriate counties, in the future I’d like to rework it so it is more efficient as I used two for loops that iterated through each dataset, so it took a long time to complete.

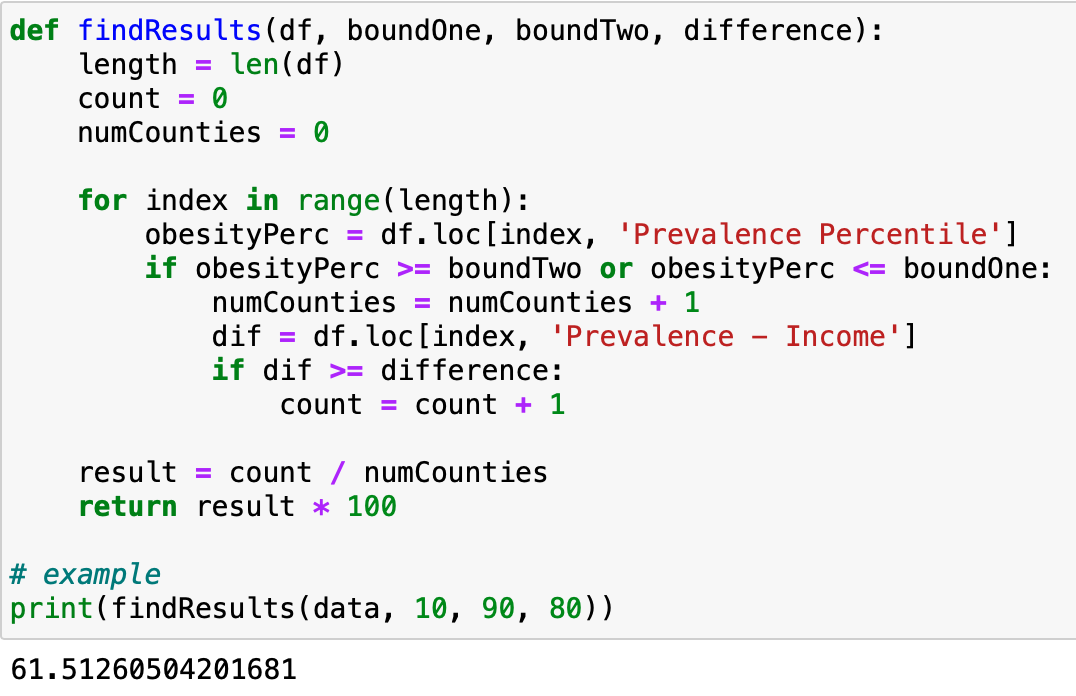
**Results and Findings**

In the beginning of my project, I intended to only compare the percentile maps and the percentile difference map. By doing that, I could eyeball certain areas with very high or low percentiles in one feature and look at the percentile difference map to see if it was very different from the other feature’s value. However, this method was not very good as I would have to manually check nearly every county and I would not have any real mathematical results to back up my project goal of identifying and proving an inverse relationship.

In order to quantify the inverse relationship between median family income and obesity prevalence, I created a function that would return numerical results. The function takes inputs of percentile bounds and lowest difference acceptable. The percentile bounds signify the levels of obesity prevalence percentile the user wants to look at. For instance, if the values were 10 and 90, it would find counties whose obesity prevalence percentile was less than or equal to 10 or higher than or equal to 90. To keep track, for each county that applies, a variable *numCounties* is incremented, keeping a count of the amount of counties that fit the criteria. The last input is the lowest difference acceptable, which is the key part of the quantification function. For instance, if the input value was 80, then for each of the counties that passed the percentile bounds, the absolute difference with income percentile would be found. If it was above 80, then a variable *count* is incremented, which keeps track of how many counties clearly show the inverse relationship between income and obesity. Once all of the counties have been checked, then the function returns a value that is *count/numCounties \* 100*. This value represents the percentage of

counties that demonstrate the inverse relationship based on the given parameters. With this, you can see the percentage of how many counties actually have an inverse relationship between median family income and obesity prevalence given an acceptable level.

With this function, I could easily test varying acceptable levels for both bounds and acceptable differences. Below are some of the tests that I ran along with the function code:



| **Low Ceiling** | **High Floor** | **Lowest Acceptable Difference** | **Result** |
| --- | --- | --- | --- |
| 20 | 80 | 70 | **52.36%** |
| 20 | 80 | 80 | **35.65%** |
| 20 | 80 | 90 | **16.13%** |
| 20 | 80 | 95 | **6.78%** |
|  | | | |
| 10 | 90 | 70 | **75.63%** |
| 10 | 90 | 80 | **61.51%** |
| 10 | 90 | 90 | **32.77%** |
| 10 | 90 | 95 | **13.78%** |
|  | | | |
| 5 | 95 | 70 | **87.29%** |
| 5 | 95 | 80 | **79.60%** |
| 5 | 95 | 90 | **54.52%** |
| 5 | 95 | 95 | **27.42%** |
|  | | | |
| 1 | 99 | 70 | **94.92%** |
| 1 | 99 | 80 | **89.83%** |
| 1 | 99 | 90 | **79.66%** |
| 1 | 99 | 95 | **59.32%** |

From these experiments, it can be concluded that an inverse relationship between obesity prevalence percentile and median family income percentile does in fact exist. Acceptable differences may vary between people and studies, but a difference value that I believed is acceptable is 80, as it seems like it represents a big difference between obesity and income percentiles. This difference level helps to prove the relation with the four various bound levels, returning results of 35.65%, 61.51%, 79.60%, and 89.83%. These results mean that out of all counties with very high or low obesity prevalence percentiles, *x%* of them demonstrate an inverse relationship depending on the acceptance value.

An interesting finding from these results is that it appears that the relationship is much more pronounced in more extreme instances of high or low obesity percentiles. I ran four tests with different bounds, making them more restrictive each time so that less counties would qualify. For each of the lowest acceptable differences in percentiles, the results actually increased, signifying that the inverse relationship greatly applies when a county has extremely high or low obesity prevalence. For instance, making an extremely selective criteria of a given lowest acceptable difference of 95, the result values increased from 6.78% to 13.78% to 27.42% to 59.32%.

All of the tests that I ran proved that an inverse relationship exists between the percentiles of county obesity prevalence and median family income. Furthermore, it became apparent that the relationship became much stronger the more extreme a county’s obesity prevalence was, whether it was extremely high or low. With these results, many actions can be taken to try and reduce obesity prevalence in the United States, primarily for counties with extremely high obesity prevalence and low income. For example, one solution may be to set up a free or cheap local recreation center that can promote more active and healthy lifestyles. Another action that may be taken is to set up supermarkets in those counties that provide healthy food options at affordable prices. Although the findings do demonstrate a clear inverse relationship, what my project did not focus on or do was identify the key factors that lead to obesity, which may be the most important aspect. While median family income is definitely a factor in obesity prevalence, there are many more important factors to investigate that would help reduce obesity prevalence at a faster rate.

Works Cited

Ogden CL, Lamb MM, Carroll MD, Flegal KM. Obesity and socioeconomic status in adults:

United States 1988–1994 and 2005–2008. NCHS data brief no 50. Hyattsville, MD:

National Center for Health Statistics. 2010.

R. Alexander Bentley, Paul Ormerod, Damian J. Ruck. **Recent origin and evolution of**

**obesity-income correlation across the United States**. *Palgrave Communications*, 2018;

4 (1) DOI: [10.1057/s41599-018-0201-x](http://dx.doi.org/10.1057/s41599-018-0201-x)