Is Income Positively Correlated with Health?

ECEN 689, Project 1

Brandon Thayer, Jatin Kamnani, Swati Ramachandran, Harish Kumar

Abstract

Does a higher income lead to better health? This report investigates this question by examining data relating to income, adult diabetes rates, adult obesity rates, and other metrics related to health and well-being, both mental and physical. This report shows that income and health do tend to be positively correlated, but since other factors are at play, higher income does not necessarily guarantee better health.

I. Introduction

PREVIOUS studies have concluded that the greater the income, the lower the likelihood of disease or premature death [1]. However, such broad conclusions are difficult to concretely establish as income and health are difficult to measure and compare, especially when the comparison is across wide variations in geography, societal conditions and several other demographic factors. With this report, we aim to present statistical evidence that backs the claim that income level is positively correlated to health. In order to evaluate this claim, we use data from different sources, analyze them along multiple pathways, and explore different types of correlation.

This report uses two primary indicators of health for a geographical region: rates of adult diabetes and obesity. While these metrics don't tell the entire story of human health, they can be used as an effective proxy. In addition, self-reported general health, other diseases, and feelings of sadness and their relationship with income are discussed. As for income information, we have data on the volume of tax returns filed in different tax brackets. However, additional substantiation from proxy metrics for income is often useful as well. A prime utility of such proxy metrics is to adjust for the ubiquitous disparities in cost of living across different counties. For example, people requiring food assistance from the government are likely to be poor. In counties where absolute income levels and cost of living are both high, such a proxy metric would adjust for the bias introduced by incorrectly assuming that higher absolute incomes fully translate to a better standard of living across all geographies.

The organization of this report is as follows: Section II discusses the data used to investigate the relationship between income and health, Sections III and IV present findings from different data sources, and Section V ties all the findings together and evaluates the claim that income is positively correlated to health.

II. Data: Sources, Mapping, and Correlation Assessment

We have extracted data from the Food Environment Atlas [2] (shortened to "Food Atlas" henceforth) to assemble statistics on food environment indicators. This includes food prices, food and nutrition assistance programs, and well-being indicators like obesity and diabetes rates. The second source of data is from the Internal Revenue Service (IRS) [3]. This dataset provides individual income tax statistics, aggregated by ZIP Code. The third source of data is a Federal Information Processing Standards (FIPS) [4] dataset that allows us to map FIPS codes that uniquely identify US counties to ZIP codes of areas that lie within the respective counties. This provides us a pathway to aggregate and match ZIP Code-level data from the IRS dataset to County-level data from the Food Atlas.

We have also used a cost of living index [5] that helps us place a state relative to others in terms of how expensive it is to live in that state. By means of this, we have some adjustment against geographical disparities in cost of living. Finally, we have used data from Summary Health Statistics for U.S. Adults: National Health Interview Survey, 2013 that is available publicly on the Center for Disease Control and Prevention website [6].

A. Exploration

The first step to begin statistical analysis is to explore the data set. For purpose of this study, we have used data from the year 2013. The reason for doing this is twofold:

- We have consistent information available for this year, which includes data on food assistance programs, obesity and diabetes rates, population estimates and cost of living indices.
- Most of the data in the Food Atlas is drawn as an estimate from the 2010 census data. The further away from 2010 we go, the greater the divergence introduced into the estimates. We want the data to have a low extrapolation error while still being representative of recent trends.

B. Data Mapping

The second step is to ensure that all data is on the same geographical scale. In this case, some information (Food Atlas) is aggregated at the county-level (uniquely identified by the FIPS codes), whereas IRS information is available ZIP code-wise. We created a tool that allows us to convert from FIPS to ZIP codes (and vice versa) effectively. We use this tool to aggregate IRS data up to the county level.

C. IRS Data Details and Derivations

As previously mentioned, we used IRS data from the 2013 tax year, and this data is geographically aggregated by ZIP code. We then agglomerate this data by counties. After aggregation, each county, uniquely identified by its FIPS code, has six rows affiliated with it - one for each adjusted gross income (AGI) bracket. The AGI brackets are as follows:

1) $1 - $24,999
2) $25,000 - $49,999
3) $50,000 - $74,999
4) $75,000 - $99,999
5) $100,000 - $199,999
6) $200,000 and above

It is important to note that tax returns are placed into an AGI bracket regardless of the number of individuals covered by a single tax return. For example, a family of three with an AGI of $60k will be placed in the third bracket, and an unmarried individual with no dependents and an AGI of $60k will also be placed in the third bracket. To avoid incorrect reporting arising from this unequal comparison, we estimated the number of individuals in each bracket using the following formula:

$$n_p = n_s + 2 \cdot n_j + n_d \tag{1}$$

Here, n_p is the estimated number of people in an AGI bracket, n_s is the number of single returns, n_j is the number of jointly-filed returns filed by married couples, and n_d is the total number of dependents in the bracket.

After (Item 1) has been evaluated for every AGI bracket and county, we were able to compute several key county-level metrics that will be referenced throughout this report:

1) Mean AGI per person
2) Percentage of people in each AGI bracket
3) The median AGI bracket
4) Median mean AGI per person

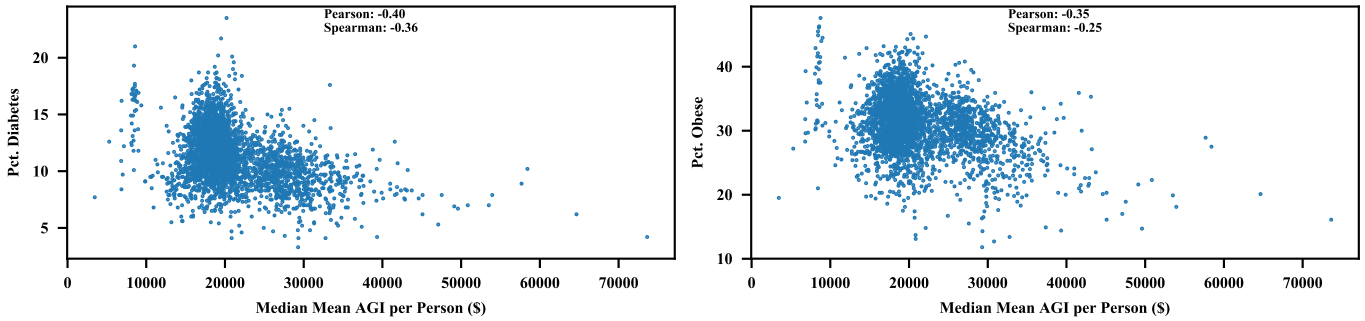
Item 4 in the list above is effectively a proxy for the true median AGI per person. This merits additional explanation: The IRS only provides data aggregated at the ZIP code level, and we've aggregated that up to the county level. For a given county, an estimate of the number of people (derived from Equation (1)) in an AGI bracket and the total AGI for the bracket enables us to compute the mean AGI per person (Item 1). The estimate for the number of people in each AGI bracket can be used to determine the median bracket. The median mean AGI per person is then the mean AGI per person (Item 1) of the median income bracket for a given county (Item 3).

D. Assessing Correlation

In this report, two methods of assessing correlation are used: Pearson's correlation coefficient and Spearman's correlation coefficient. Pearson's correlation coefficient is used for summarizing the strength of linear relationships between two features, while Spearman's correlation coefficient assesses the strength of a monotonic (not necessarily linear) relationship between features [7]–[9]. One disadvantage of Pearson's correlation coefficient is that it assumes the underlying data to be normally distributed. Spearman's correlation coefficient makes no such assumption of the underlying distribution, and this is advantageous for the analysis in this report: it is well known that the U.S. has significant wealth inequality [10], and the assumption of normally distributed incomes here would be gravely erroneous. We used [11] to compute both coefficients.

III. Findings from IRS and Food Environment Atlas Data

The IRS and Food Atlas data were *joined* by FIPS code, and we explored the relationships between income, diabetes, and obesity. While there are many factors that impact health, rates of adult diabetes and obesity are most readily available, and can provide critical insights into the correlation (or lack thereof) between income and health.



(a) (b)

Fig. 1. (a): Rate of adult diabetes (%) vs. median mean AGI for each county. (b): Rate of adult obesity (%) vs. median mean AGI for each county

A. Diabetes and Obesity vs. AGI

Fig. 1a and Fig. 1b show adult diabetes and obesity rates, respectively, versus the median mean AGI (see Section II-C) for each county. Both Fig. 1a and Fig. 1b have negative Pearson and Spearman correlation coefficients, indicating that as income increases, rates of diabetes and obesity decrease. However, none of the correlation coefficients in Fig. 1a and Fig. 1b are particularly close to -1, indicating the correlation is not very strong.

Fig. 2 and Fig. 3 plot rates of adult diabetes and obesity, respectively, vs. the approximate percentage of people in each income bracket. Each scatter plot has Pearson and Spearman correlation coefficients labeled. Note that for both Fig. 2 and Fig. 3, the Pearson and Spearman correlation coefficients start positive with moderate correlation strength, and decrease for each successively higher AGI bracket. The exception is that in Fig. 2 the correlation coefficients increase between the penultimate and ultimate income brackets. These correlation coefficients are a numerical way to analyze the shape of each scatter plot. Visually, we can see the scatter plot shape change from income bracket to income bracket. In the lowest bracket, especially in Fig. 2, the scatter plot has a generally positive slope; in general, as the percentage of people in the lowest income bracket increases, so does the county's adult diabetes rate. Moving to the right (to higher income brackets), the scatter plots in Fig. 2 and Fig. 3 seem to rotate counter-clockwise. This visual rotation indicates that the lowest and highest incomes are most strongly correlated to rates of diabetes and obesity.

Another finding from Fig. 2 and Fig. 3 is that diabetes tends to be more strongly correlated with income than obesity. With the exception of the highest AGI bracket, both the Pearson and Spearman correlation coefficients have a higher magnitude in Fig. 2 than in Fig. 3.

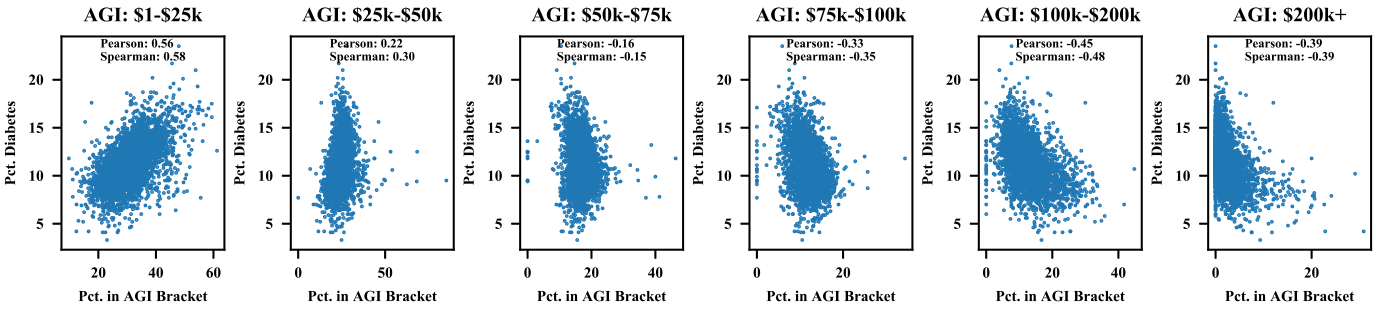


Fig. 2. Percentage of adults with diabetes versus percentage of people in AGI bracket. Each data point represents a county.

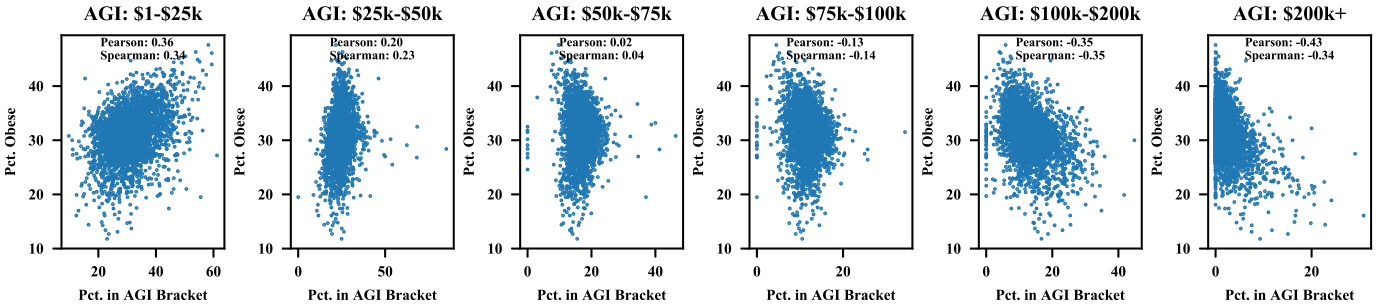
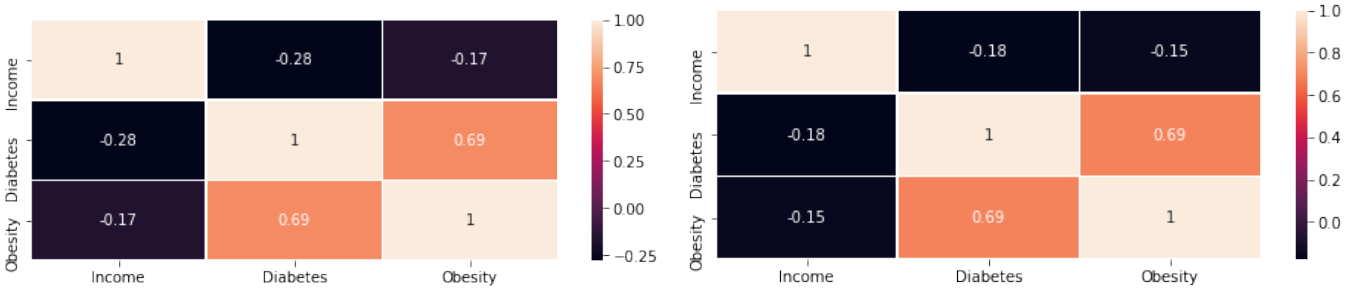


Fig. 3. Percentage of obese adults versus percentage of people in AGI bracket. Each data point represents a county.

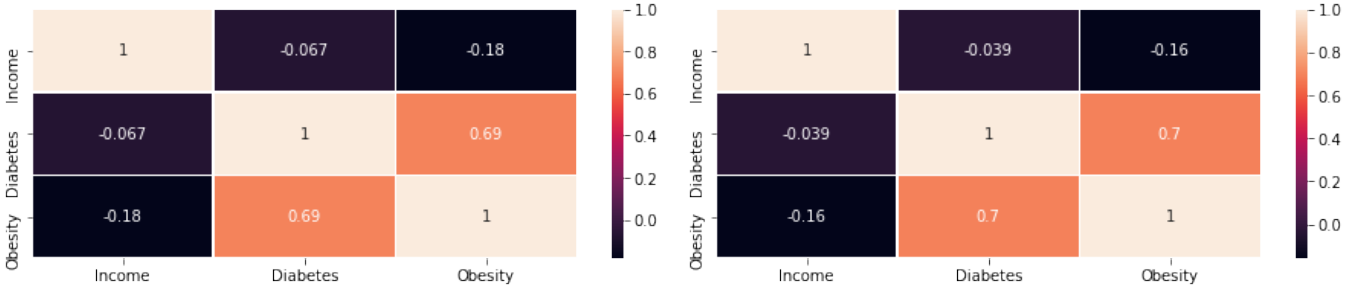
B. Correlation Between Diabetes, Income, Obesity

To evaluate whether income and health are correlated, it is imperative that we check if a correlation exists among income and health factors (diabetes, obesity). To do that, we first segregated the data into different AGI brackets and then created heat maps representing the correlations among the mean AGI per person (Section II-C), the percentage of people suffering from diabetes and obesity. These heat maps are shown in Figs. 4a, 4b, 5a, 5b, 6a, and 6b. The heat maps show an interesting relationship in AGI brackets 1, 2, 3, and 4: the correlation coefficients between income and health metrics are negative, implying that within these AGI brackets, higher income likely correlates with lower rates of diabetes and obesity. It is worth noting that the correlation coefficients approach zero for the people in higher income brackets (AGI brackets 5 and 6). Collectively, these heat maps tell us that the correlation between income and diabetes/obesity in general becomes weaker with increased income. On a side note, from Figs. 4a-6b, we discern that diabetes and obesity are positively and strongly correlated, meaning that if an area has a relatively high incidence rate for one of these diseases, it is likely to have a relatively high incidence rate for the other as well.



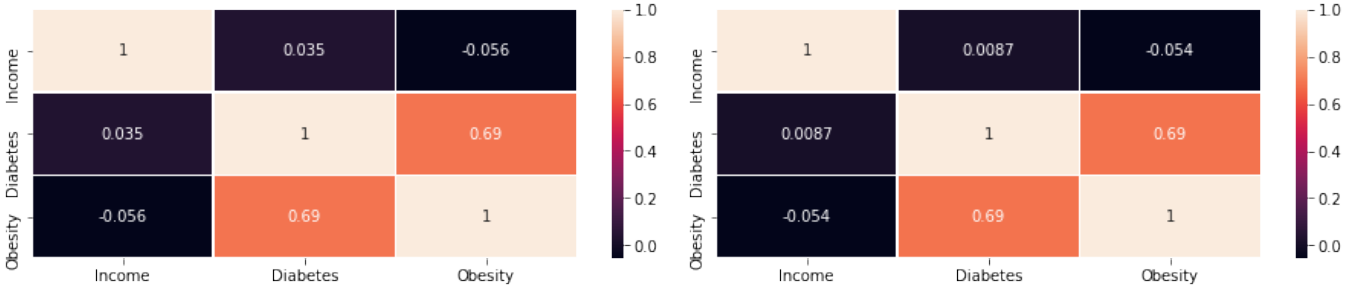
(a) (b)

Fig. 4. Correlation coefficients between mean AGI per person ("Income"), adult diabetes rate ("Diabetes") and adult obesity rate ("Obesity"). (a): AGI bracket 1, (b): AGI bracket 2



(a) (b)

Fig. 5. Correlation coefficients between mean AGI per person ("Income"), adult diabetes rate ("Diabetes") and adult obesity rate ("Obesity"). (a): AGI bracket 3, (b): AGI bracket 4



(a) (b)

Fig. 6. Correlation coefficients between mean AGI per person ("Income"), adult diabetes rate ("Diabetes") and adult obesity rate ("Obesity"). (a): AGI bracket 5, (b): AGI bracket 6

C. Geographic Visualization of Income, Diabetes, and Obesity

Figs. 7, 8, and 9 show a geographical representation of the percentage of people in the first AGI bracket, adult diabetes rate, and adult obesity rate, respectively. Maps were created with [12]. In each of these figures, the ranges of the color bins

have been chosen to contain an equal number of counties for each color - they represent percentiles of the underlying data, ranging from 0% to 100% in increments of 20%. The legends indicate the range of values that fall into each percentile bin. It can be seen from the figures that there are areas of the U.S. with relatively large percentages of people in the lowest AGI bracket and relatively high rates of both obesity and diabetes. These trends can specifically be seen in southern states such as Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Georgia, Florida, and South Carolina. However, this trend is not true across the country. The starkest counter-example is Colorado. Despite having several counties with percentages of people in AGI bracket 1 in the highest percentile bins, the entire state is in the lowest diabetes and obesity bins. Similarly, California and New Mexico have several counties in higher AGI bracket percentile bins, but most counties are in the lower diabetes and obesity percentile bins. The maps in Figs. 7, 8, and 9 indicate that there are likely other regional factors that impact health and may be independent of income.

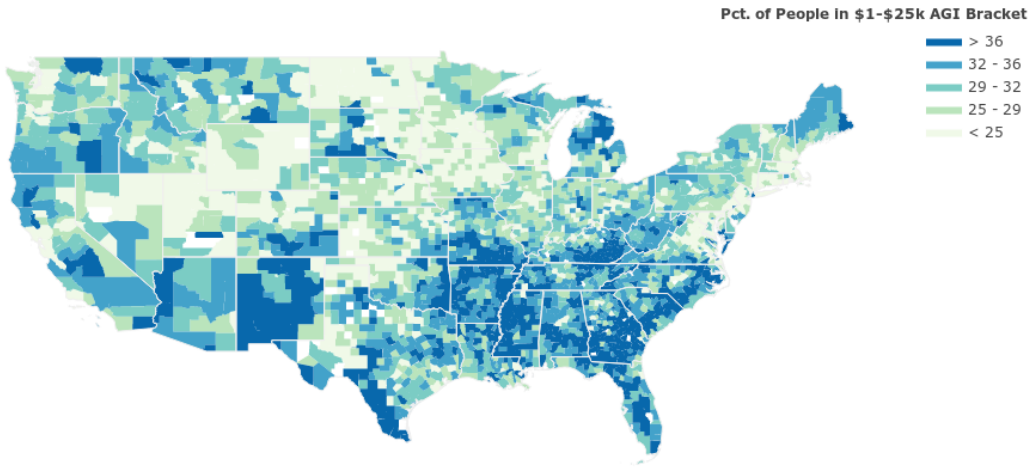


Fig. 7. Percentage of people in AGI bracket 1 by county. Color bins represent data percentiles 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%.

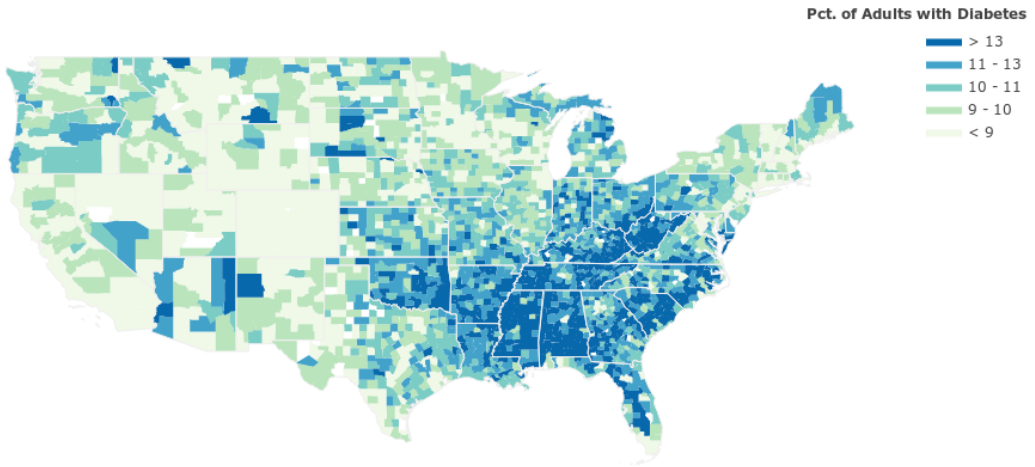


Fig. 8. Adult diabetes rate (%) by county. Color bins represent data percentiles 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%.

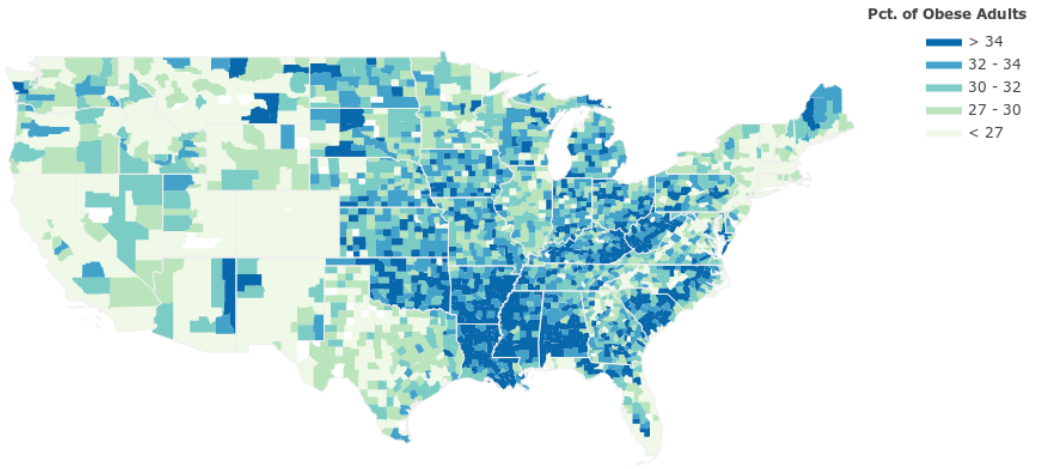


Fig. 9. Adult obesity rate (%) by county. Color bins represent data percentiles 0-20%, 20-40%, 40-60%, 60-80%, and 80-100%.

D. Health insurance

People with low income usually tend to have restricted health care access and are thus likely to be uninsured. They might face financial barriers in terms of co-payments and deductibles and hence not receive all the recommended treatment or preventive care [13]. Additionally, data on the percentage of people requiring food assistance can be used as a proxy to model low-income groups. Thus, a positive correlation between percentage of people with no health insurance for any given state and number of people requiring food assistance indicates that access to health care is linked to income. Fig. 10 shows the correlation coefficients between the proportion of uninsured population and some indicators of poverty. This correlation is calculated at the state level. It can be observed that there is a moderate correlation between the percentage of people with no access to health insurance and people requiring food assistance. This supports the claim that people with lower income tend to have little or no access to affordable health care and thus, have worse health as compared to their high-income counterparts.

Categories	Correlation
WIC participants	0.349029
National School Lunch Program participants	0.356203
School Breakfast Program participants	0.438156
Child and Adult Care participants	0.306012

Fig. 10. Correlation between lack of insurance and poverty indicators

IV. Findings from CDC Survey

In order to corroborate our claim that income is a positive indicator of health, we have used publicly available data obtained from the Center for Disease Control and Prevention [6]. CDC Data for 2013 has grouped annual family income into the following categories (USD):

1) Less than $35,000
2) $35,000 to $49,999
3) $50,000 to $74,999
4) $75,000 to $99,999
5) More than $100,000

We graph and analyze some common indicators of the lack of physical and mental health against the percentage of people in the above income brackets.

A. Risk of diseases versus income

Fig. 11 represents the percentage prevalence of some common diseases with respect to income brackets. With the exception of arthritis, these diseases monotonically decrease as income increases. Additionally, the CDC Survey states that in 2013, 56.0% people who were reported as “Not Poor” got sufficient physical activity whereas only around 37% people reported as “Poor” were sufficiently active. This serves as further evidence to the previous findings where it was seen that lower income has a correlation to higher risk of diseases like diabetes and obesity.

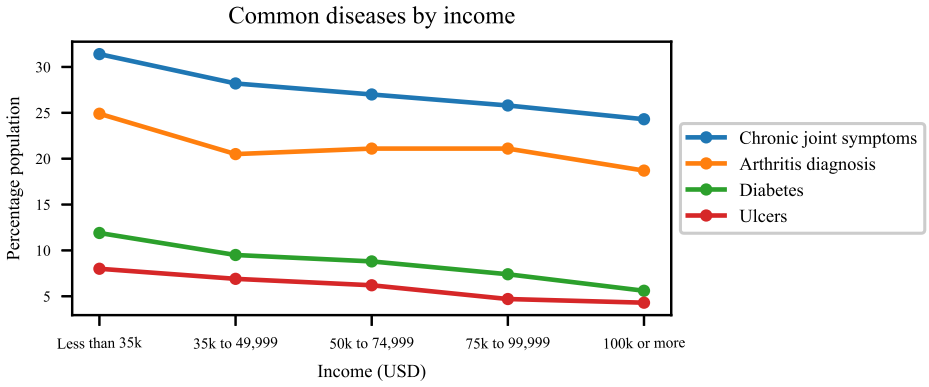


Fig. 11. Percentage of disease prevalence with respect to income

B. Mental health versus income

Fig. 12 represents the Feelings of Sadness, Hopelessness or Worthlessness with respect to income. Compared with people

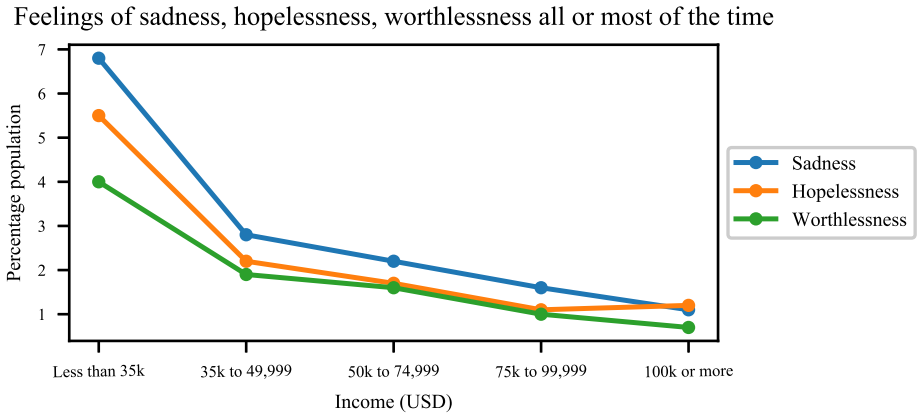


Fig. 12. Feelings of Sadness, Hopelessness or Worthlessness with respect to income

from families earning $100,000 or more per year, people from families earning less than $35,000 a year are almost four times as likely to report feelings of sadness or hopelessness all or most of the time [6]. This information also directly corresponds to self assessed reports of health. Respondents were asked to assess their own health as Fair/Poor, Good or Excellent/Very good. Fig. 13 shows self assessed health reports with respect to income. From the findings above, it can be seen that an individuals self assessment of health also supports the statement that more income probably means better health.

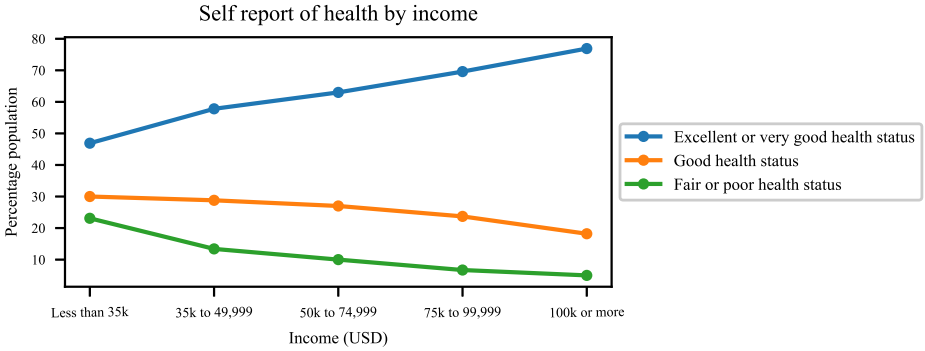


Fig. 13. Self assessment of health with respect to income

V. Conclusions

We presented an analysis of health and income data from various sources and the conclusions that we could infer from it. In summary, it is observed that higher the income, the better is the likelihood of good health. The correlative analysis done in Sections III-A and III-B establish this relationship numerically. Interestingly, we also noted correlation between the incidence of diabetes and obesity, and this is only expected as medical studies show the significant effect of obesity upon the risk of diabetes [14].

We also showed in Section IV how the prevalence of other debilitating diseases and negative mental states negatively correlates with income. We believe that there are several factors that may be contributing to this trend and offer supporting data that backs these assertions:

- As noted in Section III-D, high-income groups have access to better insurance coverage. Considering the high healthcare costs in the US, this should play a significant role in causing the correlation of health with wealth.
- More affluent people can afford nutritious meals which are potentially more expensive than processed, calorie-dense "junk food". This is evident from the negative correlation of obesity with income levels that we discussed in III-A and III-B.
- The strong negative correlation(IV-B) between the self-declared presence of poor mental state and increasing income can also be a contributing factor in this relationship. [15] demonstrates how mental health contributes to physical health, functional ability and mortality in adults, and it is possible that these two correlations jointly strengthen the relationship between higher income and better physical well-being.

During the course of this study, we did not fully account for factors like geographical variations in cost of living, access to grocery stores/farmers' markets or education to analyze the trends between income and health. These could be interesting ideas for further study in this area. Overall, it is seen that to some extent, income is indeed positively correlated with health.

References

[1] "Health, united states, 2011," National Center for Health Statistics, Tech. Rep., 2012.
[2] "Food environment atlas," https://www.ers.usda.gov/data-products/food-environment-atlas, accessed: 2018-10-28.
[3] "Individual income tax zip code data," https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi, accessed: 2018-10-28.
[4] "Federal information processing standards publications," https://www.nist.gov/itl/itl-publications/federal-information-processing-standards-fips, accessed: 2018-10-28.
[5] "Cost of living by state with detailed u.s. maps," https://www.uslearning.net/cost-of-living-by-state.html, accessed: 2018-10-28.
[6] "Tables of summary health statistics," https://www.cdc.gov/nchs/nhis/shs/tables.htm, accessed: 2018-10-28.
[7] J. Brownlee, "How to use correlation to understand the relationship between variables," https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/, accessed: 2018-10-28.
[8] "Pearson correlation coefficient," https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, accessed: 2018-10-28.
[9] "Spearman's rank correlation coefficient," https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient, accessed: 2018-10-28.
[10] "Wealth inequality in the united states," https://inequality.org/facts/wealth-inequality/, accessed: 2018-10-30.
[11] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2018-10-28]. [Online]. Available: http://www.scipy.org/
[12] P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly
[13] "Health insurance in the united states: 2013 - tables," https://www.census.gov/data/tables/2014/demo/health-insurance/p60-250.html, accessed: 2018-10-28.
[14] R. H. Eckel, S. E. Kahn, E. Ferrannini, A. B. Goldfine, D. M. Nathan, M. W. Schwartz, R. J. Smith, and S. R. Smith, "Obesity and type 2 diabetes: what can be unified and what needs to be individualized?" *The Journal of Clinical Endocrinology & Metabolism*, vol. 96, no. 6, pp. 1654–1663, 2011.
[15] Y. Lee, "The predictive value of self assessed general, physical, and mental health on functional decline and mortality in older adults," *Journal of Epidemiology & Community Health*, vol. 54, no. 2, pp. 123–129, 2000.