Association between Health and Wealth in the US

Aditya Lahiri, Rishi Laddha, Harinath, Anirudh Shaktawat Singh

I. Introduction

The objective of this project was to test the claim if a wealthy population tended to be healthy. To prove or refute this claim we used the Food Atlas dataset from USDA, income tax statistics data (2016) from IRS and life expectancy data from the Health Inequality Project [1]-[3]. To analyze the data, we integrated the data obtained from each of these different sources. Since all the datasets had a spatial component specifying the FIPS or county of origin for each sample point of data, they could be aggregated into a single dataset based upon the FIPS or county information. In the following sections we visualize the raw data, explore the trends of health and wealth, and then explore various other underlying parameters that might contribute towards proving or refuting the central claim.

II. Visualization of Raw Data

The Food Atlas dataset provides us with information regarding the percentage of population affected by diabetes and obesity across counties in 2013. We use these parameters to quantify "health" across the various counties and on the other hand we use the total income available along with other financial parameters available from the IRS dataset to measure "wealth". In Fig.1-3 we will visualize the health and wealth distribution across all the counties of the United States.

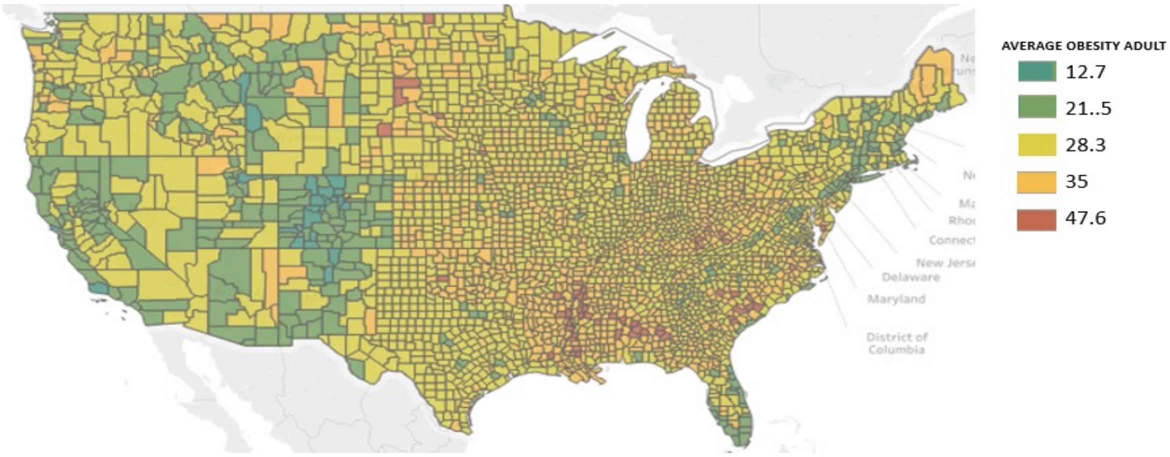


Figure 1 Obesity rate in Counties of The United States in 2013.

It can be observed from Fig.1 that the average percent of population with obesity varies from 11.7% to 47.6%. There are total 3144 Counties with highest obesity rate in the counties of Mississippi and lowest rate in Alaska.

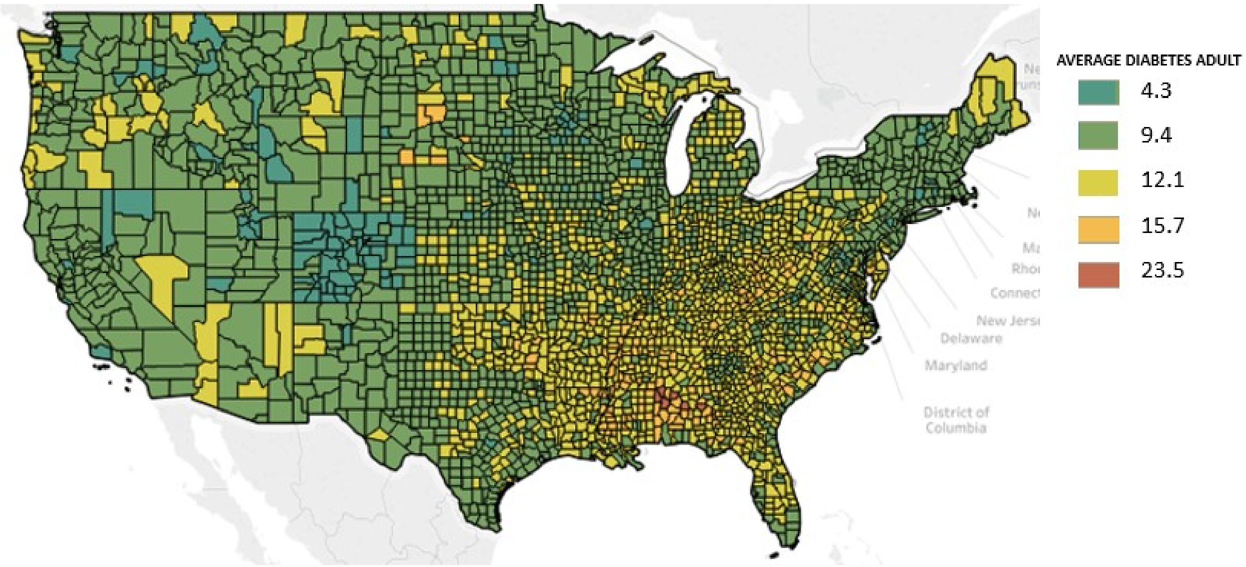


Figure 2 Diabetes rate in Counties of The United States in 2013.

Similarly, in Fig.2 we visualize the percentage of population afflicted with diabetes in 2013. We find that the average percent of population with diabetes from 3% to 23.5%. Alabama had the highest share of population with diabetes whereas Colorado had the least percent of population affected by diabetes. In Fig.3 we plot the average income across the different states.

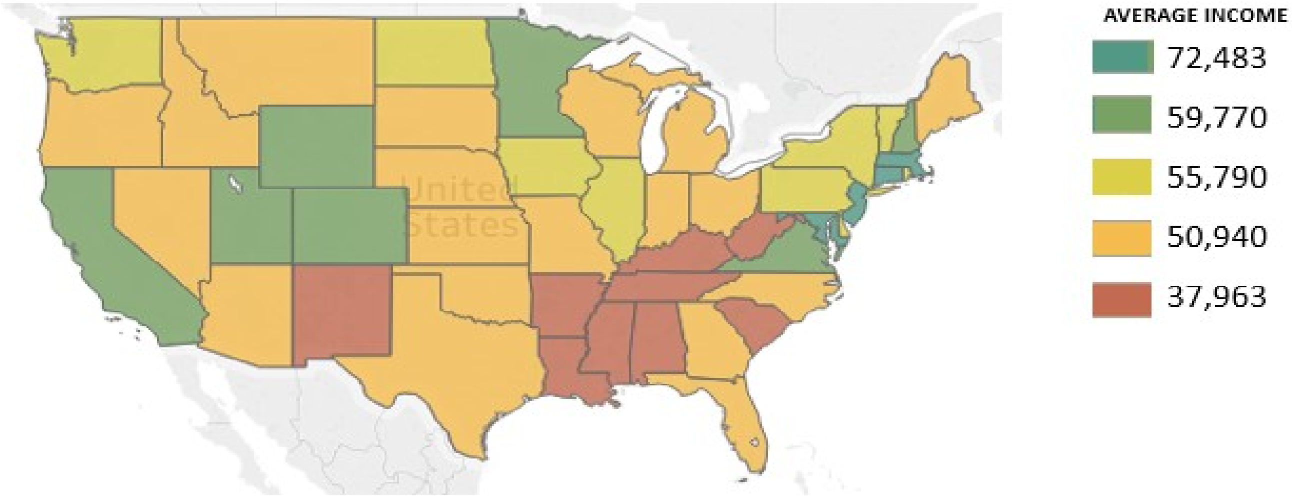


Figure 3 Average income across the states in 2013.

Based on Fig.1-3 it can be observed that the states identified with high rate of diabetes and obesity are the states with low average income. Hence, it can be hypothesized that there could be a direct correlation between the income and Health (Measured by Obesity and Diabetes rate). Although, there could be several other factors that may affect the obesity and diabetes rate in a region. To test our hypothesis, we will try to obtain statistical evidence of the correlation in the following sections.

III. Correlation analysis of Income with Health Indicators

First part of our analysis involves computing correlation between Health indicators (Diabetes rate, Obesity rate) with total income. After preprocessing the 'Food Atlas Data – Health sheet' and Income statistics from IRS data, datasets were merged based on the FIPS code. Now dependent variables were (Diabetes rate, Obesity rate) and independent variable was 'Total income', using which we performed a Simple Linear regression test whose outputs r value and p value you can see below. P vale of the test (<0.05) shows the value we get is significant and we can see that r-values (-0.27, -0.35) doesn't show a significant relation between the income and health indicators. As another experiment, we split the counties into two groups based on Total income of county (threshold - $100000). For rich counties (50% pf the data) we can see the relation between Obesity rate - Income has become stronger. Whereas for low income counties there is no significant relationship between Income and Health at all (r = 0.02,0.01).

Income Level	Dependent Variable	Pearson Correlation	P value
Overall	% Population Diabetes 2013	-0.27	1.70E-19
	% Population Obese 2013	-0.35	3.80E-58
>100000	% Population Diabetes 2013	-0.243	1.70E-19
	% Population Obese 2013	-0.42	3.80E-58
<=100000	% Population Diabetes 2013	0.021	0.06
	% Population Obese 2013	0.011	3.70E-01

We can see that for a segment of counties (Income > $100000) have correlation of -0.42 with Obesity rate indicating that counties with higher income would have less obese population than counties with lower income. While we are able to see some relationship between Health and income here r value being (<0.5) doesn't indicate a strong relationship.

Multiple Linear Regression model to predict Health:

From above section, we saw that income variable on its own doesn't explain Diabetes population % or Obesity population %. So, using the other segments of data from Food atlas sections which gives us county level statistics of 275 variables, including new indicators on access & proximity to a grocery store, socioeconomic spread, prices_taxes etc a MLR model was built to predict the Diabetes population %.

Dep.Variable	PCT_DIABETES_ADULTS08	**R-squared:**	0.986
Model:	OLS	**Adj.Rsq**	0.985
Method:	Least Squares	**F-statistic:**	1.19E+04
Prob F Statistic	0		
Log-Likelihood:	-4795.9		
Observations #:	2973	**AIC:**	9626
Df Residuals:	2956	**BIC:**	9728
Df Model:	17		

Feature type	Features	coef	std err	t	P>|t|	Comments
Access	PCT_LACCESS_HHNV10	0.0851	0.018	4.606	0	Households, no car, low access
Access	PCT_LACCESS_HHNV15	0.1485	0.019	7.92	0	Households, no car, low access
Food insecurities	FOODINSEC_13_15	0.1419	0.01	13.549	0	Food insecurities
Food insecurities	FOODINSEC_CHILD_03_11	0.0551	0.018	3.125	0.002	Food insecurities
Local	CSA12	-0.0215	0.005	-4.225	0	Local food outlets
Local	DIRSALES_FARMS07	0.0048	0.002	3.187	0.001	Farmer's markets direct sales
Local	DIRSALES_FARMS12	-0.0038	0.002	-2.449	0.014	Farmer's markets direct sales
Local	FMRKTPTH09	-0.5913	0.397	-1.49	0.136	Farmer's markets count

Local	FMRKTPTH16	-0.7648	0.291	-2.627	0.009	Farmer's markets count
Median Household Income	MEDHHINC15	-4.26E-05	2.27E-06	-18.77	0	Median Household Income
Prices_taxes	CHIPSTAX_STORES14	0.0813	0.012	6.563	0	Price of chips
Prices_taxes	CHIPSTAX_VENDM14	0.0712	0.01	6.977	0	Price of chips
Prices_taxes	MILK_PRICE10	-2.999	0.498	-6.017	0	Milk price
Prices_taxes	MILK_SODA_PRICE10	8.4048	0.484	17.362	0	Soda price
Socio Economic	PCT_18YOUNGER10	0.1167	0.007	16.515	0	% Population age <18
Socio Economic	PCT_65OLDER10	0.082	0.006	14.787	0	% Population age >65
Socio Economic	PCT_HISP10	-0.0381	0.002	-17.87	0	% Hispanic population

Table 1: Statistics of multiple linear regression

Best subsets regression was run on the dataset to identify the model with max adjusted r sq and their statistics along with set of features was extracted. From the results table.1 above we can see that model is robust owing to high R Sq (0.986) and high F statistic. Moreover all the features in the model has p value (<0.1) indicating their importance in the model. We can see that model features are from different types of datasets : 2 features from Access,2 features from Food insecurity data, 2 from 'Local' data, an 'Income' indicator, 4 features from 'Prices-Taxes' and 3 features from 'Socio Economic' dataset. Comments about the features are given in the table.1 itself.

Based upon these results we can see that health is influenced from multiple factors among which 'Median Household Income' is one of them. But it is unclear whether the relationship (coefficient) is negative because increased income allows individuals to purchase more health inputs that improve their health, because healthy individuals are more productive and thus can earn higher wages in the labor market, or because a third factor is improving health and increasing income. For further understanding of effects of these features we perform detailed analysis which evaluates how relationship changes with respect to demographics, geography, prices & taxes etc.

IV. Correlation analysis with 'Local' & 'Prices_taxes' dataset

In this section we explore to find additional features relating to health and income from datasets like 'Prices_taxes' and 'Local Food'. 'Prices_taxes' sheet from food atlas data contains the average prices of essentials like Milk, Soda, their sales tax values for each county in 2010. 'Local food' data contains features like direct sales of farms, farmers markets count and farms & crop acerage etc. This also contains 'Physical activity' indicators like recreation & fitness facilities. Using this data we computed the pearson correlation coefficient of all combinations of features.

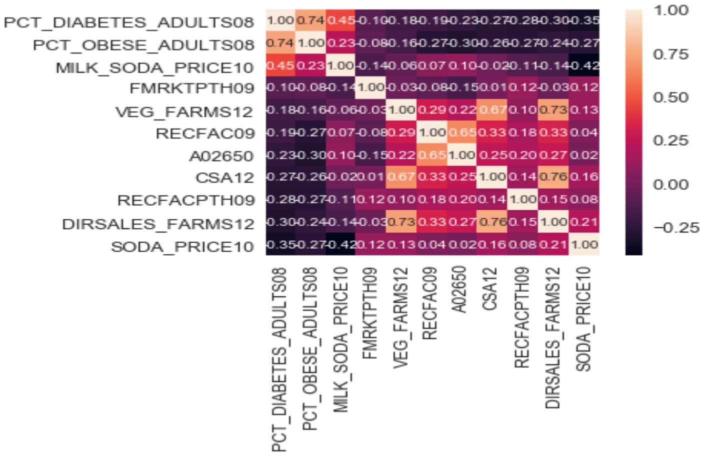


Figure 4: Correlation Matrix of various parameters

Table.1 above contains important features from three datasets and their corresponding correlation value sorted based on correlation with % Diabetes population 2008. From above, we see that Milk_soda_price, soda_price have significant positive relationship with diabetes rate (r =0.45, -0.35). Milk_soda_price is ratio of price of low fat milk with soda price. Seeing that it has positive correlation with diabetes and soda_price has negative correlation with diabetes rate, we see that when soda price increases diabetes rate is low because soda consumption is less. In addition, we see that total income ('A02650') has high positive correlation (r=0.65) with Recreation & Fitness facilities ('RFCFACPTH09') which is because more rich the county is, more number of people are ready to spend on recreation activities. Recreation & fitness facilities have negative correlation with Diabetes (-0.28) & obesity (-0.27) and thus has positive effect on health.

V. Health and Wealth across Socioeconomic Groups, Food Assistance and food Insecurity

In this section we explore how health and wealth vary across different socioeconomic groups, food assistance and food insecurity. To analyze the trends of health and wealth across these various parameters we use the dataset pertaining to socioeconomics, assistance and insecurity available in the food atlas dataset. These three sub-datasets data are aggregated with the health data (in food atlas dataset) and the 2016 IRS tax dataset. The datasets are merged based on the FIPS and then compressed by averaging the parameters (columns) of counties belonging to same states. The final averaged dataset gives us information regarding health, wealth, socioeconomics, food assistance and food insecurity across every state. The first step in our analysis involved analyzing the trends wealth and health across the fifty states. So, we created a scatter plot (fig.5) where wealth was measured by the median household income in 2015, and health was measured by the percentage of population who had diabetes and were obese.

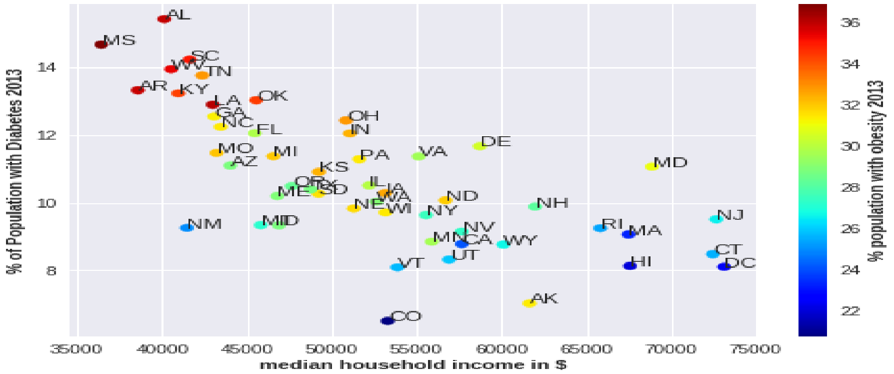


Figure 5: Scatter plot for obesity, diabetes and median household trends across the states.

The scatter plot in fig.5 reveals that as the median house hold increase the percentage of population affected by diabetes and obesity also decrease. However, these are observations and to test the claim whether wealthier people tend to be healthier, we performed single linear regression using by keeping the independent variable as median household income in 2015 and using the percent of population affected by diabetes in 2013 and the percent of population affected by obesity in 2013 as dependent variables. We have summarized the results of linear regression in Table 2 below.

Dependent Variable	Pearson Correlation	P value
% Population Diabetes 2013	-0.685	2.833e-08
% Population Obese 2013	-0.659	1.447e-07

Table 2: Linear regression of obesity and diabetes with respect to median house hold income.

The results in the table.2 show that diabetes and obesity are negative correlated with median household income which means wealthier the population the healthier it is. These results have a p-value less than 0.05 which means that these correlations are significant and not due to chance. Now we will look into how different races within the populations are affected by this health-wealth trend. We only consider diabetes for measuring health in this context as we have seen from the scatter plot in fig.5 that diabetes and obesity seem to be highly correlated with each other and closely follow each other. We carry out linear regression with population of each race for 2010 as independent variable and median household income (MHI) (2015) and percent of population with diabetes (2013) as the dependent variables to study the correlation of health and wealth with race. The results of our analysis is summarized in table 3.

#	Independent Variable	Dependent Variable	Pearson Correlation	P value
1	White Population	Diabetes 2013	0.0555	0.698
2	White Population	MHI 2015	-0.1361	0.3406
3	African American Population	Diabetes 2013	0.4594	0.00069
4	African American Population	MHI 2015	-0.097	0.4976
5	Hispanic Population	Diabetes 2013	-0.2973	0.034
6	Hispanic Population	MHI 2015	0.0269	0.8508
7	Asian Population	Diabetes 2013	-0.301	0.031
8	Asian Population	MHI 2015	0.433	0.00149
9	Native American Population	Diabetes 2013	-0.255	0.0705
10	Native American Population	MHI 2015	0.0036	0.979
11	Hawaiian or Pacific Islander	Diabetes 2013	-0.2008	0.157
12	Hawaiian or Pacific Islander	MHI 2015	0.2400	0.0897

Table 3: Linear regression analysis across socioeconomic groups.

The results in table.3 indicate that African Americans are mostly affected by diabetes whereas diabetes seem to be negative correlated with the Asian and Hispanic populations. We also see that the Asian population has positive correlation with median house hold income. We do not see any significant trend for white or Hawaiian/ pacific islander populations, we see that native Americans do seem to have negative correlation with diabetes. In fig.6 below we plot the population distribution in some of the lowest and highest income states.

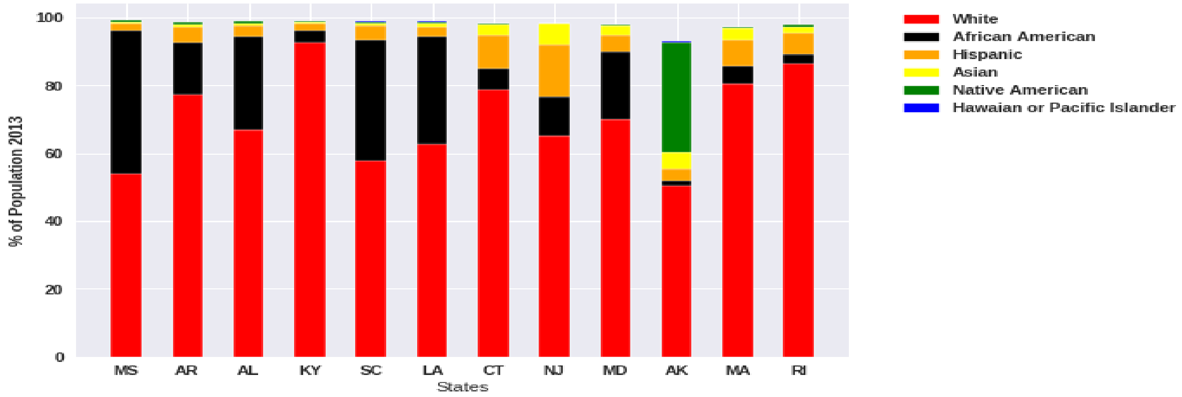
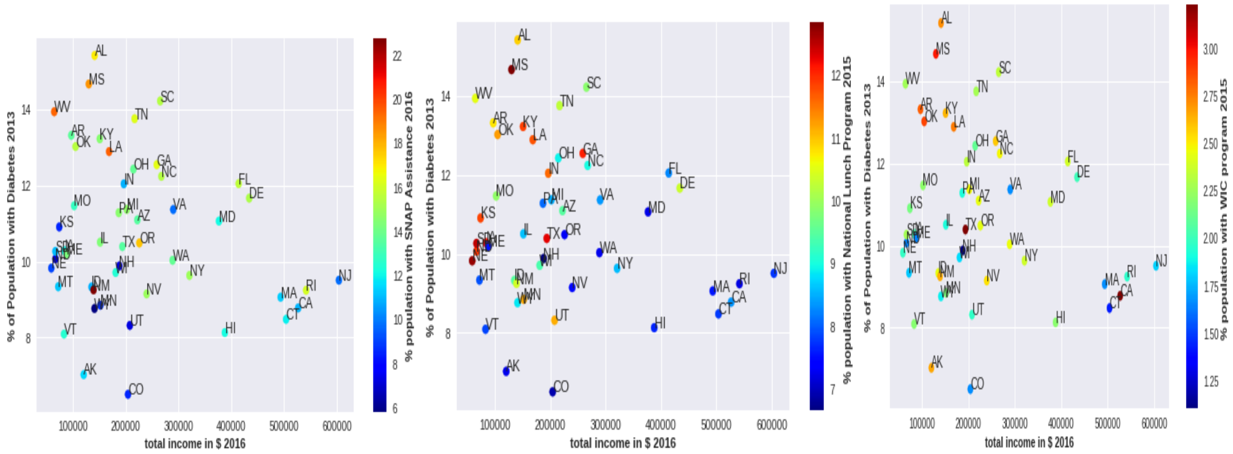


Figure 6: Race distribution across low and high income states.

We indeed see that most of the low-income states have high African American population and we also found positive correlation of diabetes for the African American population in our regression analysis. This further reinforces the statement that wealthier population tend to be healthier. We do not observe any significant trends for white population as they seem to be well distributed across the states. We see that Asian population is much higher in the wealthier states and we also found from our regression analysis than the Asian population was negatively correlated with diabetes. This further supports the statement that wealthier population seem to be healthier. Now we will consider the trends of food assistance statewide how it may be affected by the observed trend of wealthier population tending to be healthier. In fig.7a-c we plot how the participation in food assistance program such as Supplemental Nutrition Assistance Program (SNAP), National School Lunch Program (NSLP) and Woman, Infants and Children (WIC).



7.a SNAP

7.b NSLP

7.c WIC

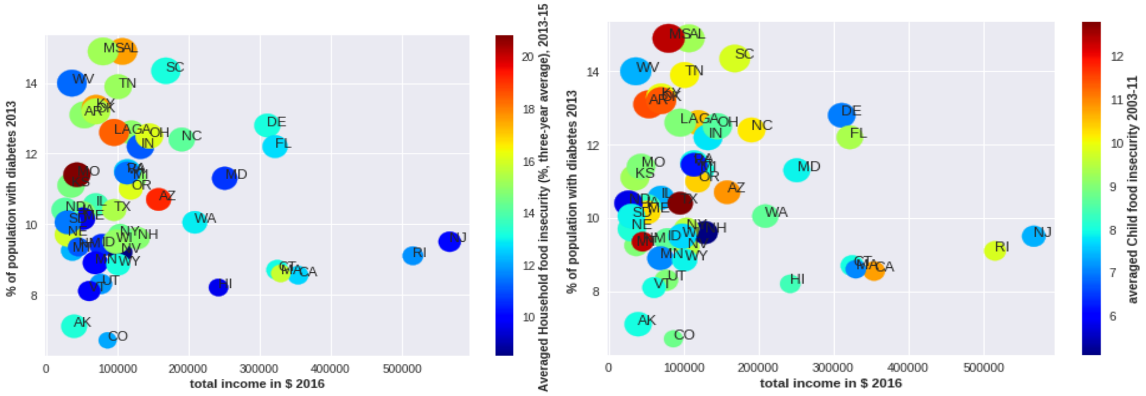
Figure 7: Variation in food assistance program participation with respect to income and diabetes across the states.

In table.4 we conduct single linear regression analysis to study correlation between percent of population participating in food assistance programs with health (diabetic population) and with wealth (total income).

#	Independent Variable	Dependent Variable	Pearson Correlation	P value
1	Total income 2015	SNAP 2016	0.0111	0.938
2	Diabetic Population 2013	SNAP 2016	0.5456	4.16e-05
3	Total income 2015	NSLP 2015	-0.480	0.0004
4	Diabetic Population 2013	NSLP 2015	0.4813	0.000402
5	Total income 2015	WIC 2015	-0.096	0.5034
6	Diabetic Population 2013	WIC 2015	0.384	0.00581

Table 4: Linear regression of food assistance programs with respect to diabetes and total income.

From the scatter plots in fig.7 and the regression analysis in table.4 we see that the food assistance participation is higher mainly in the low-income state, with an exception of California which is a high- income state and has a high participation in the WIC program. We find significant positive correlation between diabetes and each of the three food assistance program and diabetes. Total income had a significant negative correlation with participation in the NSLP program. In general, we can conclude that low-income states have high food assistance program participation and have higher percent of the population afflicted by diabetes. We will now study how food insecurity affects the trends of health and wealth across the states. In fig.8a-b, we see how the average house hold food security (2013-2015) and averaged child food insecurity (2003-2011) vary with the total income (2016), percentage of population with diabetes (2013) and obesity (2013).



8a. Averaged Food Insecurity.

8a. Child Food Insecurity.

Figure 8 Variation in Food security with respect to diabetes, total income and obesity (size of markers).

Fig.8a-b. Please note the size of the markers represents the percentage of population with obesity in that state. Hence larger the state marker larger the population with obesity in that state.

We see from the scatter plots that average household food insecurity is high only for a few low-income states and is relative low across the various income levels, child food insecurity on the other hand seems slightly more prevalent in the low income states. We see that states with high averaged household food insecurity (HFI) and child food insecurity (CFI) seem to have higher percent of the population affected by diabetes. We carry out a regression analysis to establish these possible correlations and the results are summarized in table.5.

#	Independent Variable	Dependent Variable	Pearson Correlation	P value
1	Total income 2015	HFI	-0.193	0.178
2	Diabetic Population 2013	HFI	0.446	0.0025
3	Obese Population 2013	HFI	0.4467	0.0011
4	Total income 2015	CFI	-0.083	0.5626
5	Diabetic Population 2013	CFI	0.2763	0.052
6	Obese Population 2013	CFI	0.087	0.544

Table 5: Linear regression of food insecurity indices with respect to obesity, diabetes ,total income.

Indeed, we see that there are indication of significant positive correlation of diabetes with averaged household food insecurity and child food insecurity. Fig.9 below we plot the average house hold food security across some of the lowest and highest income states and we see that food insecurity is indeed higher in low income state and from our earlier analysis we know the percentage of population with diabetes in low income states is higher. This supports the general trend that wealthier population seems to be healthier.

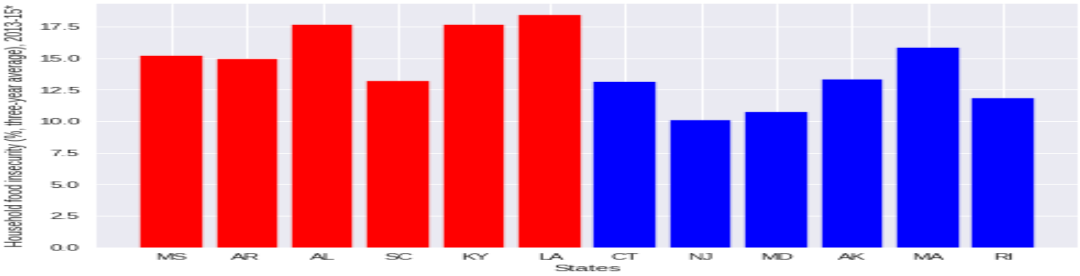


Figure 9: Averaged Food Insecurity across low(red) and high(blue) income states.

Finally, we carry out a k-means cluster analysis of the states across their income and the percent of their population with diabetes (2013). We found that the optimal cluster was 3 and it divided states across economic lines of low, mid and high-income states. Our results for the cluster analysis is summarized in fig.10 below. Diabetic rate declined with increasing income.

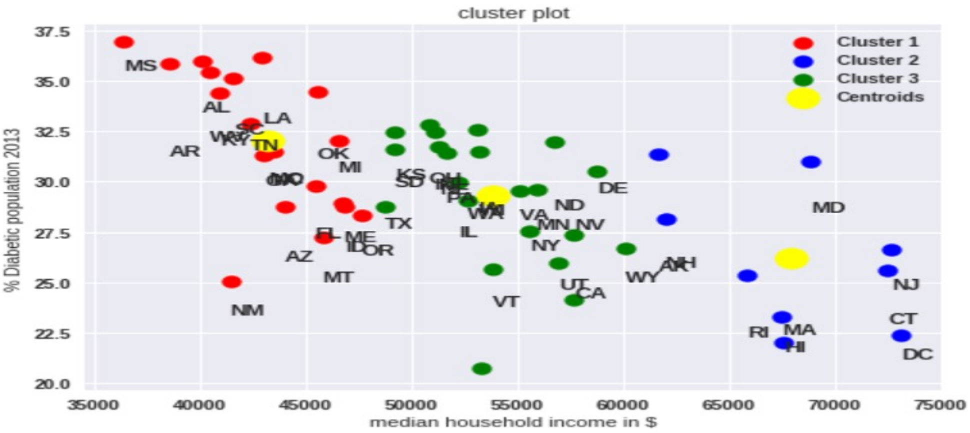


Figure 10: Clustering of high, mid and low income states.

VI. HEALTH AND WEALTH TRENDS ACROSS ACCESSIBILITY

In this section, we observe the relationship between total income of the households with respect to the utilities that they hold, which in turn affects their health. In the given dataset, we have a variable which tells that in a state, what percentage of the population have no car and live in an area which has low accessibility to stores (includes: specialized food stores, grocery stores, recreational centers, clubs etc.). Based on these parameter and average total income across the states we create the scatter following plot to observe the trends.

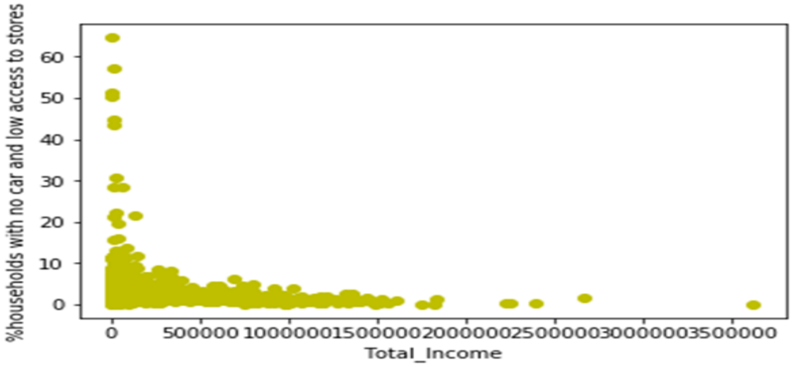
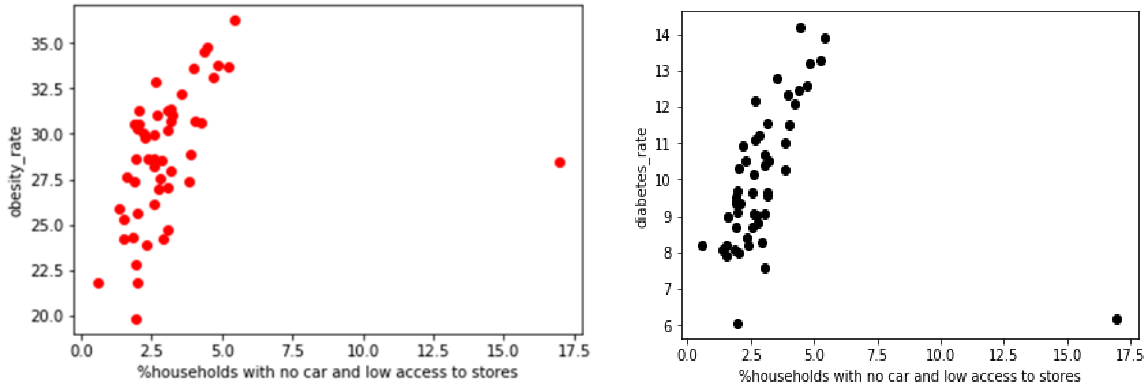


Figure 11: Accessibility vs Total income

Fig.11 clearly indicates that the states which have low average income tend to have higher percentage of the households with no car and low access to stores. This also implies that people with low income tend to have low access to grocery stores, gyms, recreational centers etc. The above observation is intriguing and leads us to investigate, how people with no car and low access to stores perform on the health criterion. We created scatter plots of the percentage of these households with respect to the obesity and diabetes rate in fig.12a-b.



12a. Obesity

12b. Diabetes

Figure 12: Obesity and Diabetes rate with respect to access

From the above graphs, we cannot observe significant relationship between health and these low income households but there is still some positive slope that can be seen indicating some kind of relationship between these variables. So, we put this through a statistical test (linear regression) and the results show pvalue = 0.02, rvalue = 0.35 (indicating that it is a significant test) and the slope is positive indicating that as the percentage of households with no cars and low access to store increases, obesity and diabetes rate also increases. The possible explanation for this might be that the people who have low access to stores probably tend to buy more non-perishable food, and the food which have preservatives to make it long lasting, which is unhealthy for long term. Since we have already seen that as the income increases, the percentage of these households with low access to stores decreases. So, all these observations, when combined are indicating that high income might lead to good health.

VII. Other Health Factors and their Relationship to wealth

In this section, we discuss about how the income affects other health factors such as life expectancy and mortality rate. Based on the central claim we should expect higher the income, higher the life expectancy and lower mortality rate. This claim is based on the fact that higher income means people have more financial resources to get access to better healthcare, spend on high quality food and possess accessibility to gyms and recreational centers. All these factors contribute to high life expectancy and low mortality rate according to World Health Organization. To test the central claim, we needed data that contains the average income of a household and the life expectancy and the mortality rate in that household. Since the data given to us did not contain this information, we went beyond the given dataset and acquired the dataset from different source (life expectancy data [3]) which consists of all the required information to test our hypothesis. We construct scatter plots of life expectancy and mortality rate vs. average household income in fig.13 and fig.14.

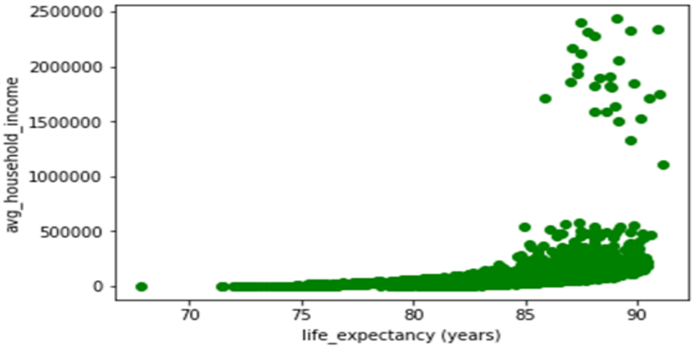


Figure 13: Average household income vs life expectancy.

The fig.13 indicates that as we go right in the graph, which means as we go towards higher life expectancy, the graph is moving upwards (the household income is increasing). To confirm this result statistically, we performed a regression test. The pvalue = 2.37e-82 and rvalue = 0.37 indicating that the test is significant, and the slope comes out to be positive and its value is 1.9e03 indicating that as the income increases, the life expectancy also increases.

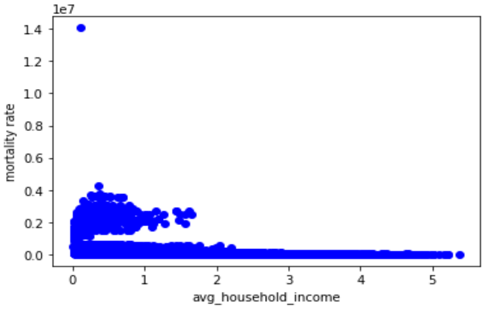


Figure 14: Mortality rate vs averaged household income.

The graph in the fig.14 is normalized. In this graph, we can observe the cluster of high mortality rate towards the lower average household income indicating that low income means high mortality rate. As we move towards the right of the graph, the mortality rate is decreasing with the increasing household income.

VIII. Conclusion

In order to establish the relationship between health and income we used 'Obesity rate' and 'Diabetes rate' as indicators of health and 'Total income', 'Median household income' as income indicators. Correlation analysis between 'Obesity rate' and 'diabetes rate' with 'Total income' indicated a weak negative relationship between them and stronger relationship was found in richer counties case. When a Multiple Linear regression model was built to predict 'Diabetes rate' using all features from the aggregated dataset, we found that 'Median Household Income' was a significant feature in model with negative coefficient value. Model results also indicated that parameters such as 'Access', 'Food insecurity', 'Prices & Taxes', 'Local Food', 'Socio economic' etc. had significant influence on the health indicators.

Further analysis was performed on different subsets of data containing parameter specific information to understand how the relationship between health and income varies across these parameters. Correlation analysis between health and 'Prices Taxes' showed how income effects soda price which has negative correlation with health indicator. We also observed that Recreation & fitness facilities have negative correlation with diabetes and obesity and thus has positive effect on health. Our analysis also revealed that states with low income had high participation in food assistance programs, high food insecurity, low accessibility to stores, higher percent of low-income groups such as African Americans. The populations of these states tended to have higher rates of diabetes and obesity along with high mortality rates. These trends were reversed in high income states where the population had low mortality rate and tended to be healthier and had low participations in food assistance programs, low food insecurity and high accessibility to stores. Thus, the various correlation analysis across various parameters supported the central claim that a wealthier population tended to be healthier.

REFERENCES

[1]USDA ERS - Data Access and Documentation Downloads", Ers.usda.gov, 2018. [Online]. Available:https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads. [Accessed: 31- Oct- 2018].

[2]SOI Tax Stats Individual Income Tax Statistics ZIP Code Data (SOI) | Internal Revenue Service", Irs.gov, 2018. [Online].
Available: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi. [Accessed: 31- Oct- 2018].

[3]"The Health Inequality Project", Healthinequality.org, 2018. [Online]. Available: https://healthinequality.org/data/. [Accessed: 31- Oct- 2018].

[4]"HUD USPS ZIP Code Crosswalk Files | HUD USER", Huduser.gov, 2018. [Online].
Available: https://www.huduser.gov/portal/datasets/usps_crosswalk.html. [Accessed: 31- Oct- 2018].