Is Income correlated to Health?

Correlation between health and income finally explained!

October 30, 2018 | ***Editors:*** *Mahalakshmi, Prabhneet, Samyuktha and Vedant*

(Source: https://www.stlouis-mo.gov/government/departments/health/news/The-Health-Insurance-Marketplace-Is-Open.cfm)

Introduction

Correlation between income inequality and health has been a source of debate for decades now. It is generally assumed that income is strongly and positively associated with health and therefore investigators have tried to prove this intuition by conducting a myriad of experiments and tests, ultimately failing to prove their intuition. Majority of studies found a minor positive, statistically significant relationship between income and health, which was considerably reduced after controlling for unmeasured confounders and health selection parameters. Most of the studies investigated short-term associations between health and income, and therefore, did not account for long-term possibly stronger

associations between the two parameters of health and income. Nevertheless, true causal short-term relationship between health and income, estimated by research of income parameters and health factors that control for confounding can be matter of interest and has been explored in the resulting report.

Income is related to health in the following three ways: Gross National Product (GDP) of countries, individual income of a person and social inequality. A thought-provoking question arises as to which of these associations reflect a causal relationship among the two factors of health and income. If so, does the redistribution of income improve health of the individuals? Do the income disparities strongly affect the health of the individuals or is it just in our minds? We have tried to answer these questions using the atlas containing the health and well-being indicators.

Data pertaining to U.S. Department of Agriculture was used to assemble statistics related to food environment indicators and provide a spatial overview of healthy food and physical fitness of individuals using the well-being indicators such as diabetes, obesity, physical fitness, fast food restaurants, recreational facilities etc. The income level mapping is done using the Individual Income Tax Statistics provided by the Internal Revenue System for the years of 2013 and 2016. The spatial information of the Federal Information Processing Standards (FIPS) codes and ZIP codes is unified by going back and forth between FIPS and ZIP to combine the health and income together.

Linear Regression and Random Forest Regression models were employed to find the correlation between the factors and the important variables that can support or refuse our null hypothesis that income is not correlated to health. Before proceeding to the model results, let's walk-through the process of transformation of raw data to final model prediction to support or refute our null hypothesis and go with the alternate hypothesis (Income is correlated to health) represented by Figure 1.

Our hypothesis for the problem is:

Null Hypothesis H_0: *Income is not related to Health*

Alternate Hypothesis H_a: *Income is correlated with Health*

Data Exploration and Visualization

We cannot move forward with the model directly without considering the most important features that can be used for model building and finding the correlations of the features with other features.

As a starting point, 2013 income was mapped with health factors such as diabetes, obesity, physical fitness, fast food restaurants, recreational facilities. State-wise income for 2016 (which shall be used later to draw conclusions) was used as a criterion to analyze the income distributions for 2013 and 2016 and compute correlations with 2013 parameters likewise. The data exploration and building model visualizations was carried out using tools such as Tableau, Excel (Pivot Tables and 3-D Maps) and Python.

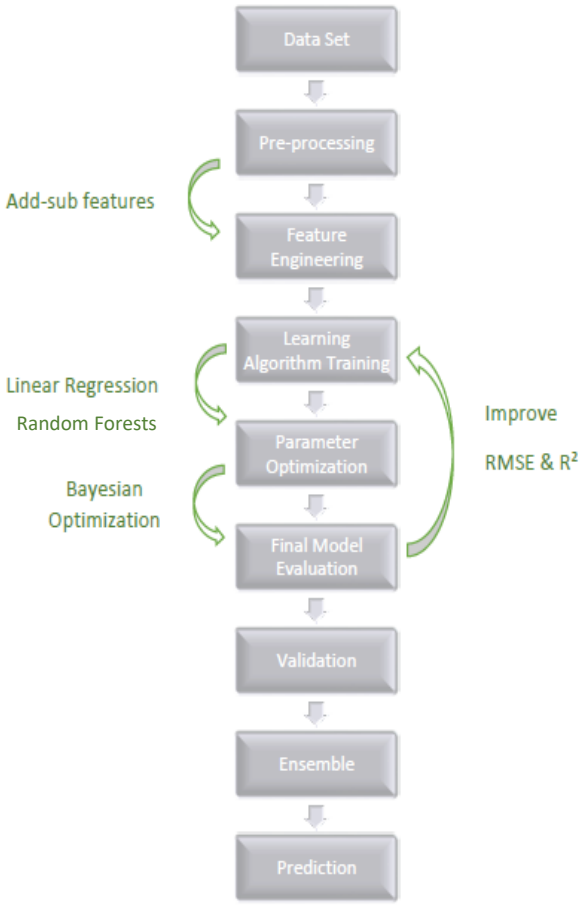


Figure 1: The process of transformation of raw data into formulation of final model for prediction

To understand the social inequality aspect, income distribution for the years 2013 and 2016 was plotted for all the states of USA represented by Figure 2 and Figure 3. It is clearly visible that states have high coefficient of social inequality. This status difference is protracted for the next three years to map a similar distribution for the year 2016.

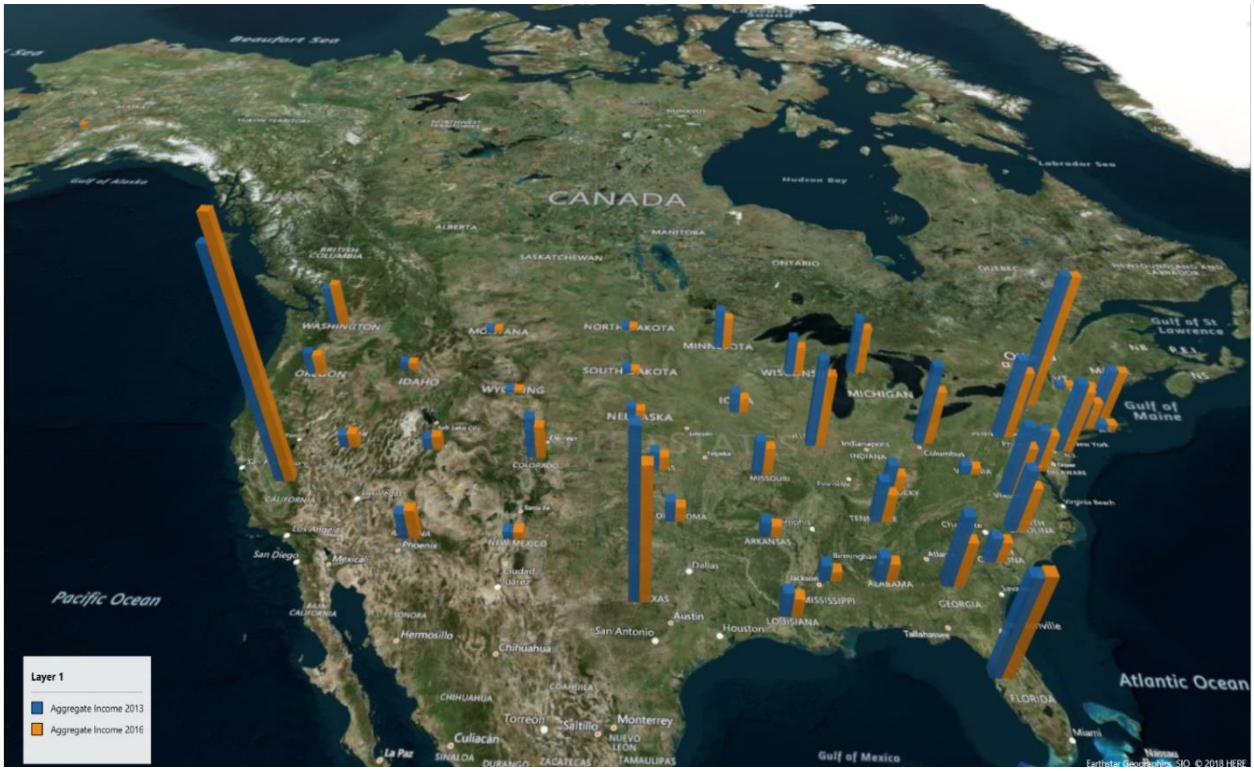


Figure 2: State-wise Income distributions for 2013 and 2016 as a 3D-representation.

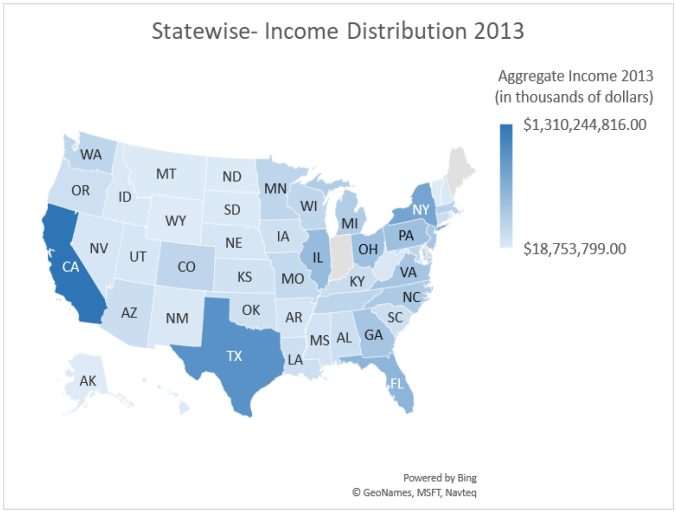


Figure 3: State-wise Income distribution for 2013

Also, Figure 3 and Figure 4 clearly state that there has been not much change in the income levels of the two years of 2013 and 2016 so our hypothesis for 2013 can be safely extended to 2016. This was required as the U.S. Department of Agriculture statistics provide features that have values pertaining to the years 2013 and 2014. Therefore, mapping income for the same year for visualization and model building seems to be a reasonable assumption. Moreover, very little variation between the 2013 and 2016 levels help us to extend the hypothesis results to 2016 income levels as well.

The next set of figures helps us to understand the correlation between the features that will be used for model building as can be seen clearly in Figure 4, 5 and 6.

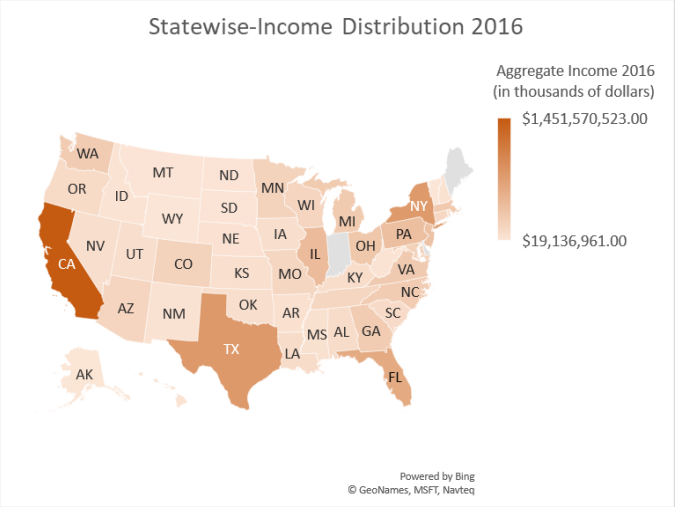


Figure 4: State-wise Income distribution for 2016

Income v/s diabetes and income v/s obesity do not represent ample correlation resulting in a conclusion that income is not directly correlated to either obesity and diabetes which is reasonable enough as diabetes is a genetic disease and both obesity and diabetes can only be correlated with income over a span of time (long-term effects) and do not exhibit short-term relationship with income. Although a lot research has been conducted in long-term effects of these parameters but no evidence regarding their effect on income was recorded. [1] On the other hand, Income v/s fast food and Income v/s recreational facilities have strong positive correlations and can be evident in Figure 6.

A similar analysis for correlation between obesity rate and recreational centers and fast food restaurants did not provide enough evidence for relationship between the two as shown in Figure 7.

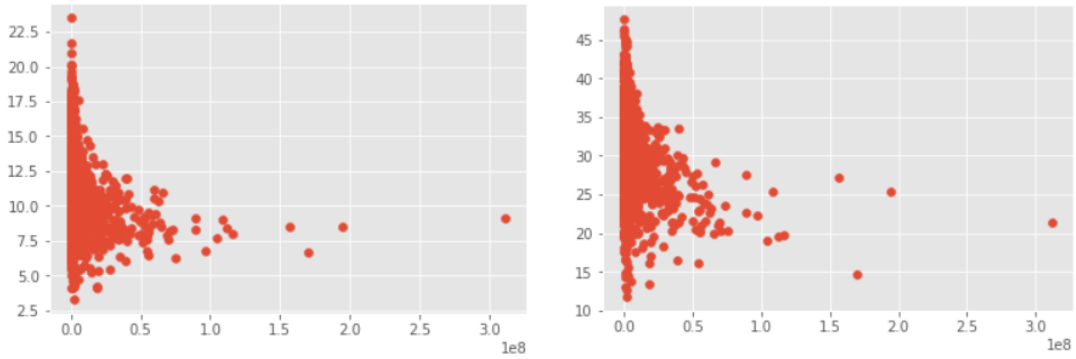


Figure 5: Correlation between Income v/s Diabetes(left) and income v/s obesity (right) is negligible

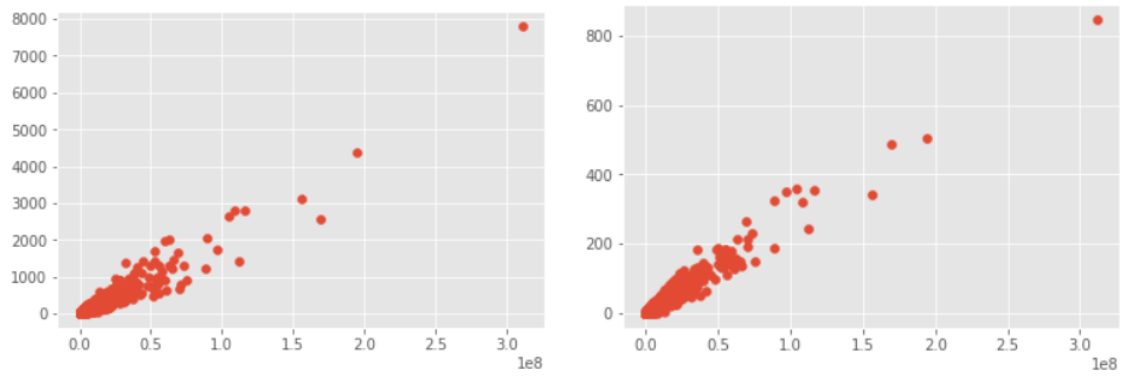


Figure 6: Strong positive correlation between Income v/s Fast food restaurants (left) and income v/s recreational activities(right)

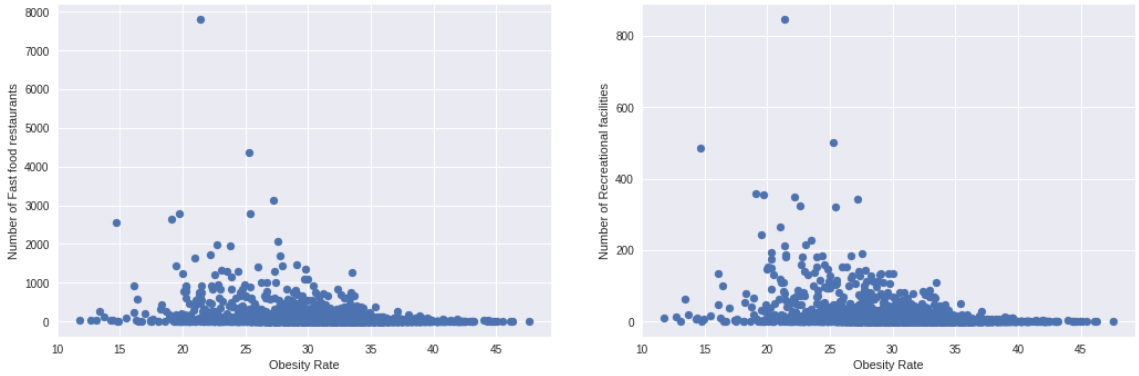


Figure 7: Correlation between obesity rate and Fast food Restaurants (left) and Recreational facilities (right)

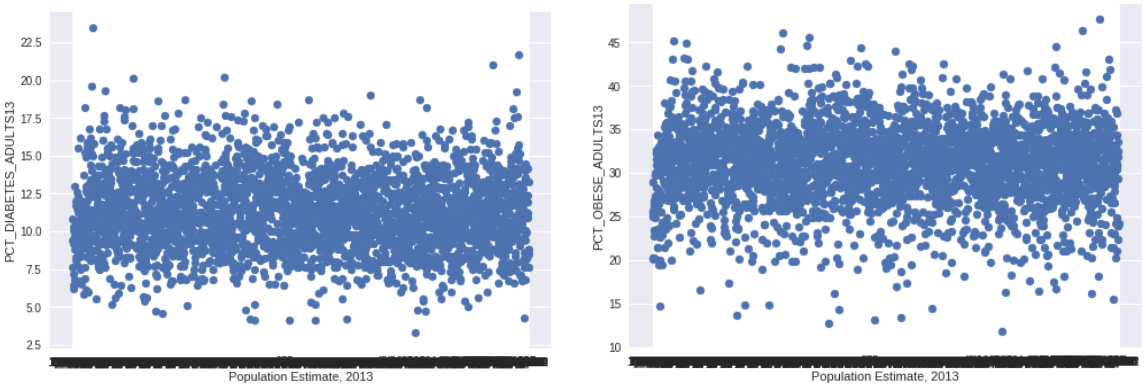


Figure 8: Data distribution for population estimate and obesity rate and diabetes rate. The mean value for diabetes rate is 11.25 and 31.10 for obesity rate.

Data distribution for population estimate v/s obesity rate and diabetes rate resulted in mean value of 11.25 and 31.10 for the two respectively as shown in figure 8.

Also, diabetes and obesity exhibit strong correlation with each other when plotted state-wise for all states of USA for the year 2013 as can be seen in Figure 7. This relationship is valid as one of the major reasons for type-2 diabetes is obesity. [2]

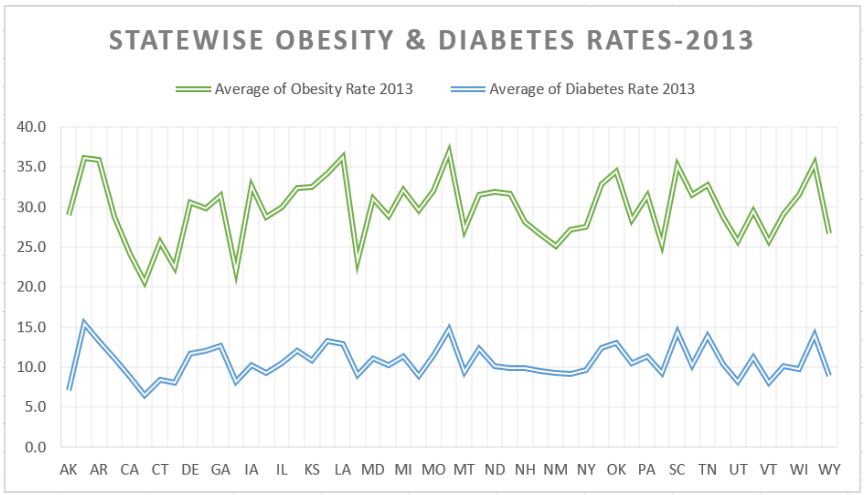


Figure 7: State-wise Obesity and Diabetes Rates 2013. The graph shows strong positive correlation between the two.

After conducting the exploratory analysis for the above factors, we are clear that Fast food restaurants and recreational activities have strong positive correlation with Income whereas the intuitively correlated factors that is obesity rate (2013) and diabetes rate (2013) do not relate well with the income estimates of 2013.

Model Evaluation and Validation

Moving on to the model building to validate the results we got from data exploration, we employed two models namely, Multiple Linear Regression and Random Forests using Income as our predictor and the other health variables as our features.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.960
Model:	OLS	Adj. R-squared:	0.960
Method:	Least Squares	F-statistic:	1.890e+04
Date:	Tue, 30 Oct 2018	Prob (F-statistic):	0.00
Time:	01:37:20	Log-Likelihood:	-50167.
No. Observations:	3118	AIC:	1.003e+05
Df Residuals:	3113	BIC:	1.004e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.891e+05	3.13e+05	2.842	0.005	2.76e+05	1.5e+06
x1	2.357e+04	2.39e+04	0.985	0.325	-2.33e+04	7.05e+04
x2	-2.48e+04	1.34e+04	-1.852	0.064	-5.11e+04	1460.166
x3	1.465e+04	540.790	27.086	0.000	1.36e+04	1.57e+04
x4	2.359e+05	4170.106	56.564	0.000	2.28e+05	2.44e+05

Omnibus:	2323.479	Durbin-Watson:	2.051
Prob(Omnibus):	0.000	Jarque-Bera (JB):	248979.193
Skew:	2.771	Prob(JB):	0.00
Kurtosis:	46.425	Cond. No.	1.99e+03

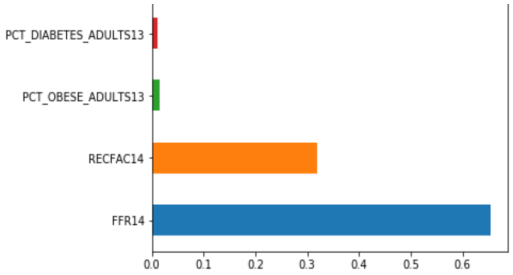
Figure 9: Summary Table for OLS Regression

The results of Multiple Linear Regression are listed in Figure 9. Linear Regression resulted in the p-values shown in Table 1. The model was then cross validated. The same dataset was used for employing the Random Forest Model and resulted in the variable importance plot as shown in Figure 10.

Table 1: Summary Table with p-values of features

Label in the Summary table	Feature Name	P Values
x1	Diabetes	0.325
x2	Obesity	0.064
x3	Fast food restaurants	0.000
x4	Recreational Facilities	0.000

Figure 10: Variable Importance Plot for Random Forest Design. Most important factor being Recreational activity



The above tables and figures clearly signify that health parameters obesity rate and diabetes rate do not affect our income parameter which is evident by the corresponding p-values (>0.05) and variable importance plots.

Results

The models and exploratory data analysis quite fall together in place to prove that the income levels are not strongly correlated to health parameters of obesity and diabetes. Although, there is some correlation between income and other parameters like fast food restaurants and recreational activities. But our analysis proves that obesity which is correlated to diabetes, does not correlate to these factors (recreational and fast food restaurants) resulting in lack of evidence between direct relationship between health and these two parameters. When the models are extended to 2016 income levels too, similar results were obtained.

Therefore, there is no evidence of strong positive or negative correlation between the income levels and health parameters as shown by our analysis, we would accept our Null Hypothesis. Also, if this analysis was extended for computing long-term effects, it might as well have proved to be a better representation of mapping relationship between health and income as evident from previous researches. [1]

References

[1] Gunasekara, F. I., Carter, K., & Blakely, T. (2011). Change in income and change in self-rated health: Systematic review of studies using repeated measures to control for confounding bias. *Social science & medicine*, *72*(2), 193-201

[2] Weyer, C., Funahashi, T., Tanaka, S., Hotta, K., Matsuzawa, Y., Pratley, R. E., & Tataranni, P. A. (2001). Hypoadiponectinemia in obesity and type 2 diabetes: close association with insulin resistance and hyperinsulinemia. *The Journal of Clinical Endocrinology & Metabolism*, *86*(5), 1930-1935.