The Effects of Socioeconomic Status and Fast Food on Individual Health in Counties

Citadel Datathon

Team 12

Patrick Li Rahul Patel Simon Suo Richard Zhang

September 15, 2018

1 Executive Summary

Our analysis reveals significant geographic differences in health characteristics across New York State counties. More specifically, upstate counties tend to suffer from poor health, and health issues resulting from high drinking, and high smoking. Counties surrounding Manhattan are less likely to suffer from drinking and smoking complications, but are of poorer health as compared to other regions.

From a logistic regression analysis of odds ratios, we infer that high density of fast food options (relative to healthier options), low household income, and high reliance on government assistance are the biggest predictor of health issues:

1. Higher household incomes in a county decrease the likelihood of belonging to an unhealthy cluster. (E.g. Increasing *mean_household_income* in a county by 1 increases the log odds of belonging to a healthy county by $2.57e^{-4}$)

2. Lower social security and higher reliance on government assistance increase the likelihood of belonging to a region with poorer health. (E.g. Decreasing *mean_soc_sec* benefits in a county by 1 increases the log odds of belonging to a healthy county by $-1.1e^{-3}$)

3. Having a higher percentage of non fast food restaurants in a county increases the likelihood of belonging to a healthy cluster group. (E.g. Increasing the percentage of *non_fast_food_restaurants* in a county by 1 increases the log odds of belonging to a healthy county by $4.87e^{-3}$)

The government should be more conscious of the existing demographic conditions in different counties. Our analysis has shown that these socioeconomic factors, which can be influenced by the government, lead to poor overall health. Contributing more assistance through welfare programs and promoting healthier eating options will lead to a healthier population and lower medical costs.

Contents

2 Topic Question

Over the years, large progress has been made to improve the general health condition of New York state residents. Though it's evident that residents do not enjoy the same level of health across counties.

We wish to understand the difference across the counties, and the cause for these large distinctions. We hypothesize that demographic and socioeconomic factors, and proximity of various categories of food establishments are the main factors that influence smoking and drinking habits, and cardiovascular, diabetic and obesity related diseases.

We believe this understanding is particularly critical for designing future government policies in food regulation. Base on our quantitative model, the government can better allocate resources to further improve equality and wellness.

3 Exploratory Data Analysis

We initially attempted to detect the change point in health indicators across time via time series analysis, but the time range proved to be insufficient for such an approach.

We begin by analyzing the distributions of the health indicators. From Fig. 1, 2, and 3, we see that the obesity, binge drinking and smoking rates roughly take on a normal distribution. There is no evident outlier county.

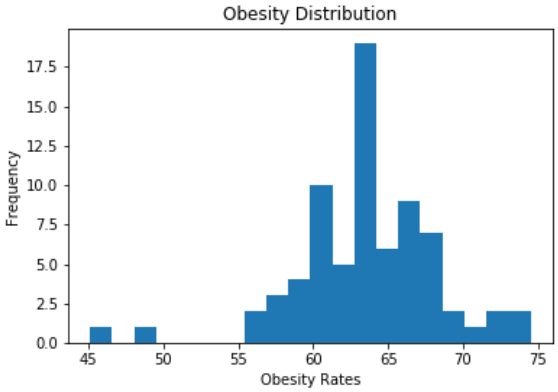

Figure 1: Obesity Rate Histogram

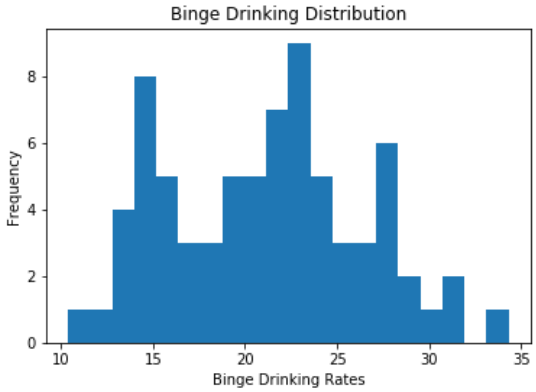

Figure 2: Binge Drinking Rate Histogram

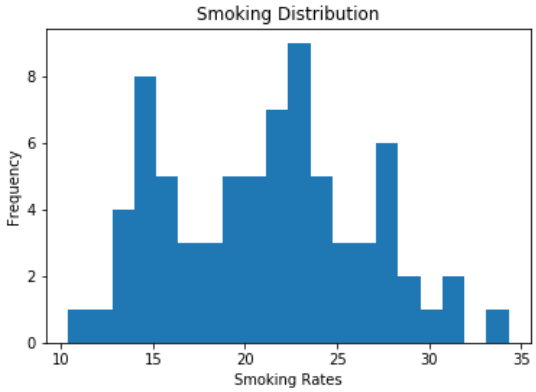

Figure 3: Smoking Rate Histogram

To better understand the myriad of health indicators, we began by visualizing a specific indicator, "Age-adjusted cerebrovascular disease (stroke) mortality rate per 100,000" as the specific", at the county level in Fig.4.

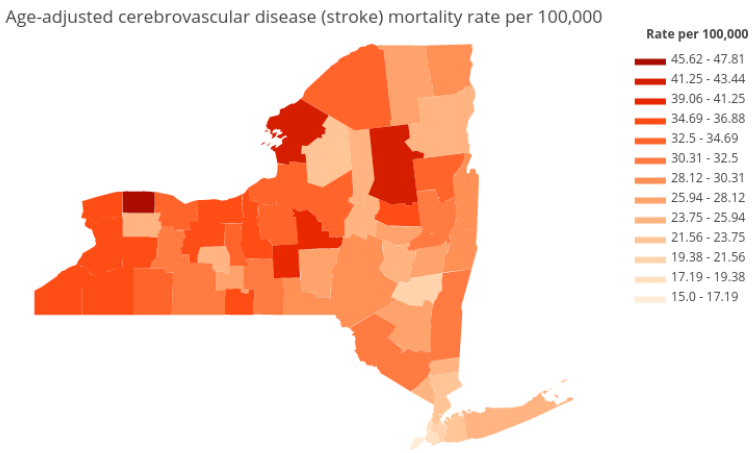

Figure 4: Obesity Related Stroke Mortality Rate by County

We picked this health indicator, under the category of "Obesity and Related Indicators", as we hypothesize that it will be highly correlated with poor diet choices and availaibility of fast food restaurants. The top 5 counties with stroke mortality rates are in Fig.5.

	County Name	Age-adjusted cerebrovascular disease (stroke) mortality rate per 100,000
3650	Orleans	47.7
3606	Hamilton	42.8
3610	Jefferson	42.5
3620	Madison	40.9
3586	Cortland	40.8

Figure 5: Counties with High Obesity Related Stroke Mortality Rate

To explore the correlation with diet habits, we visualize the density and type of food venues across New York State counties in Fig.6 and more specifically in New York City in Fig.7.

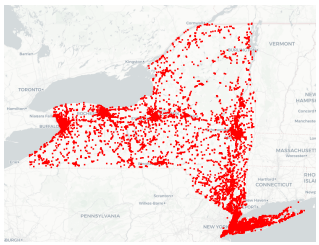

Figure 6: New York State Food Venues

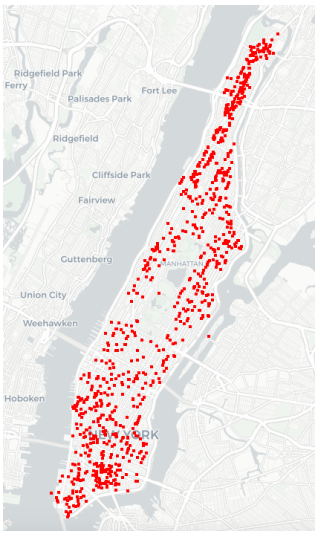

Figure 7: New York City Food Venues

To normalize for difference in population density across counties, We examined the per-capita number of fast food restaurant in each county. Fig.8 shows the food options in the top 5 most populous New York State counties.

county	fast_food	grocery_store	non_fast_food_restaurant	population	fast_food_pc	non_fast_food_restaurant_pc	grocery_store_pc
New York	1052.0	1060.0	10258	1624084.0	0.000648	0.006316	0.000653
Suffolk	62.0	17.0	576	1483117.0	0.000042	0.000388	0.000011
Nassau	31.0	24.0	254	1346876.0	0.000023	0.000189	0.000018
Westchester	153.0	331.0	1401	956762.0	0.000160	0.001464	0.000346
Erie	314.0	379.0	1777	910418.0	0.000345	0.001952	0.000416

Figure 8: New York State County Food Options

This investigation, however, did not yield our expected result. Orleans, Hamilton, and Jefferson do not have the most number of fast food restaurants per capita. In fact, they rank 49th, 9th, and 15th.

Similar investigation into other specific health indicators did not yield consistent top counties for health issues. Even within one health category, correlation between health indicators, shown in Fig. 9 are not as consistent as we initially hoped.

To gain a more systematic understanding across counties and health indicators, we decide that the best approach is to separately analyze counties with similar health characteristics and develop prediction models based on clustered results.

4 Modeling Approach

We wished to segment the counties into groups based on health indicators, and use socioeconomic factors to model and predict group membership.

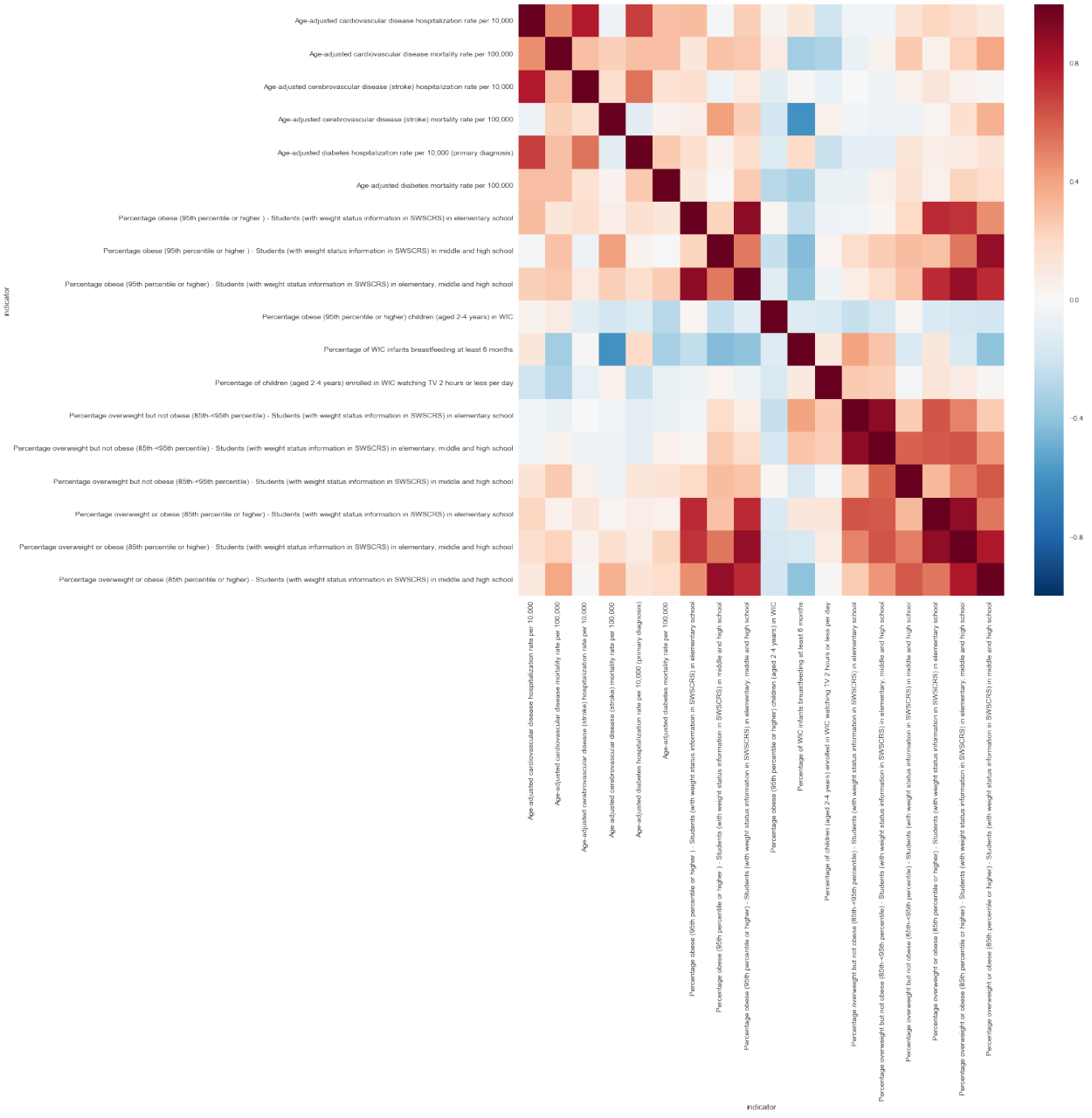

Figure 9: Pairwise Correlation of Obesity Related Indicators

4.1 K-Means Clustering

We constructed a matrix of county names and their smoking, drinking, cardiovascular, diabetic and obesity health indicators. We chose age-adjusted percentage and occurrence rates to remove the age effects (see appendix for list of indicators). Using these indicators, we performed a K-Means clustering of the counties with $k = 3$. The optimal k is discovered by the elbow-method (Figure 10) after iterating through k from 1 to 10. The elbow method measures the sum of Euclidean distances of data points from their centroids. Naturally, as we increase k, the total sum of Euclidean distances decreases, but after a certain point, the marginal decrease becomes lower. We wish to pick the largest k before the diminishing occurs. We also attempted Gaussian mixture clustering, which yielded similarly-shaped clusters. This was motivated by the quasi-normal distribution of the indicators.

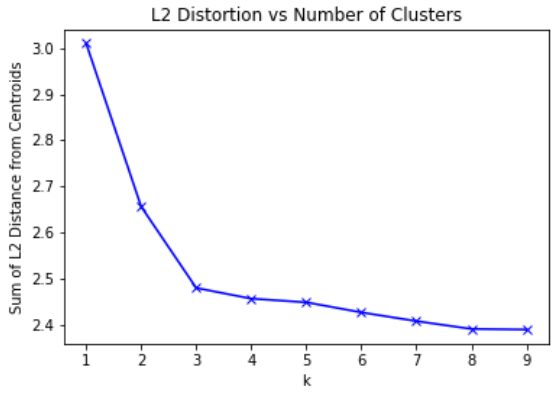

Figure 10: Elbow Method to Determine Optimal Number of Clusters

By inspecting the values of the cluster centroids, we label the 3 clusters as:

1. good health (cardiovascular & diabetes-wise), low drinking, low smoking
2. poor health, high drinking, high smoking
3. poor health, low drinking, low smoking

From Figure 11, we can see that although no location variables are used in clustering, there is clear geographical structure in the cluster labels, especially around the Manhattan area (green).

4.2 Gradient Boosting Classification

We joined the cluster labels onto the social-economical data, and performed a stratified train-test split to preserve the cluster proportions. We then modeled the cluster labels using a multiclass gradient boosting classifier.

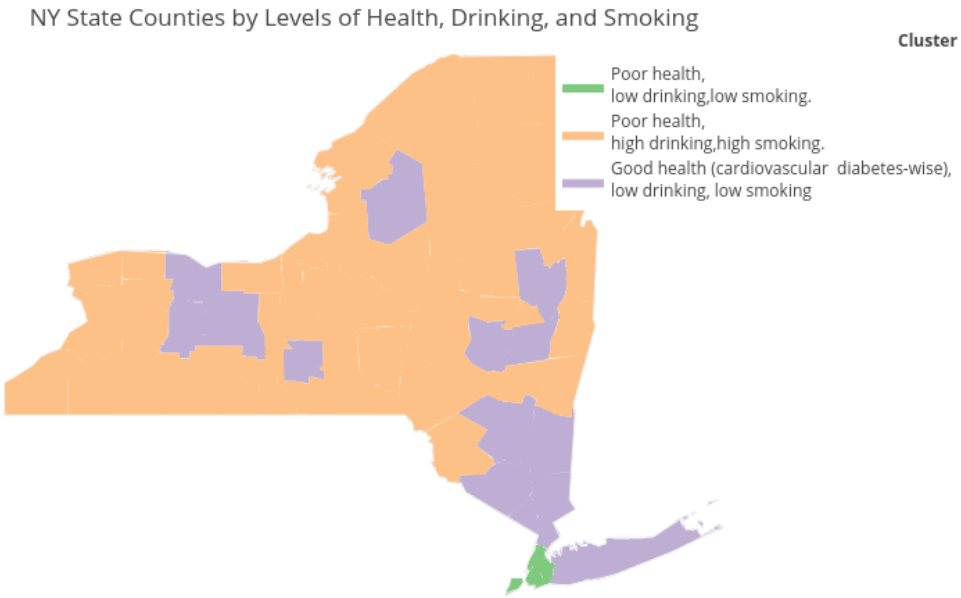

Figure 11: Counties by Health, Drinking, and Smoking

We added a Gaussian random noise variable and a uniform random noise variable to perform preliminary variable selection. Features whose importances (gini purity coefficient) are lower than those of the random variables are dropped. We then performed another round of feature selection by graphing the partial dependence plots of the remaining variables (Figure 12 and 13). Partial dependence measures the effect of one variable on the predicted response while holding all other variables constant. A large range in the average predicted response of the partial dependence plot is considered more important. Eight final variables are kept (Figure 14).

We optimized the final model using a Bayesian grid search on the hyperparameters (following a Gaussian Process). The final model achieved a cross-validated micro F1 score of 0.71.

4.3 Logistic Regression on Important Variables

To improve interpretability of the variables, we furthered explored them using a simpler model. Using these variables with high importance, we fit three one-vs-all logistic regression models: M_1, M_2, and M_3, where the response variable in M_i is set to 1 if an observation belongs to cluster i, and 0 otherwise. From the summaries, we see that most variates have relatively high p-values. This indicates that although a linear relationship may exist between the variates and the log odds ratio of belonging to a health-based cluster. Statistical significance is likely damaged by lack of data points, as there are only less than 100 counties in the state of New York. A logistic regression is also not able to sufficiently capture interaction and non-linearities compared to the more flexible tree based GBM model.

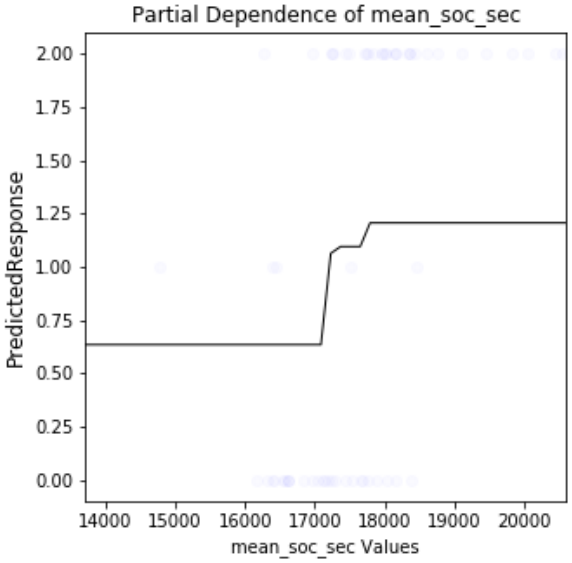

Figure 12: mean_soc_sec PDP

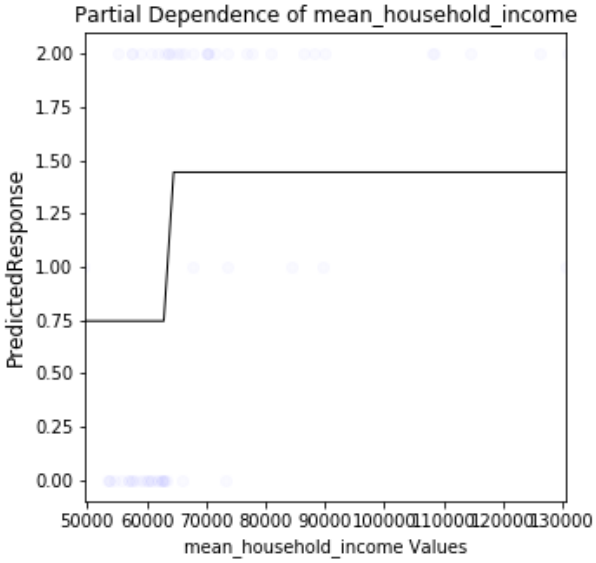

Figure 13: mean_household_income PDP

	range	variable
0	0.698413	mean_household_income
1	0.571429	mean_soc_sec
2	0.126984	pop_no_health_insurance
3	0.111111	mean_cash_pub_income
4	0.095238	non_fast_food_restaurant
5	0.095238	food_stamp_benefits
6	0.079365	mean_ret_income
7	0.079365	median_household_income

Figure 14: Top 8 Features ordered by Feature Importance from lightGBM model

5 Analysis of Results

We can interpret a coefficient β_i from M_i as follows: holding all other variables constant, one unit increase in β_i increases the likelihood of an observation appearing in Cluster 2 over all other clusters (i.e. that a county is, on average, less healthy than most other New York counties) increases by β_i percent. From the extracted coefficients of M_2 and M_1 (less healthy counties) relative to M_3 (healthier counties), (Figure 15) we notice the following trends:

1. Higher household incomes in a county decrease the likelihood of belonging to an unhealthy cluster. (E.g. Increasing *mean_household_income* in a county by 1 increases the log odds of belonging to a healthy county by $2.57e^{-4}$)

2. Lower social security and higher reliance on government assistance increases the likelihood of belonging to a region with poorer health. (E.g. Decreasing *mean_soc_sec* benefits in a county by 1 increases the log odds of belonging to a healthy county by $-1.1e^{-3}$)

3. Having a higher percentage of non fast food restaurants in a county decreases the likelihood of belonging to an unhealthy cluster. (E.g. Increasing the percentage of $non\_fast\_food\_restaurants$ in a county by 1 increases the log odds of belonging to a healthy county by $4.87e^{-3}$)

Log odds lr can be defined as follows, where p is the probability of an event occurring:

$$lr = \log(\frac{p}{1-p})$$

In aggregate, we see that less privileged communities, which are also the primary targets of fast food companies are more likely to suffer from drinking, smoking, and cardiovascular health related issues.

(Intercept)	mean_household_income	mean_soc_sec
6.577525e+00	2.527008e-04	-1.104494e-03
pop_no_health_insurance	**mean_cash_pub_income**	**non_fast_food_restaurant**
8.658367e-05	1.486288e-04	4.867068e-03
food_stamp_benefits	**mean_ret_income**	**median_household_income**
-1.722208e-04	-3.273428e-04	-5.374201e-04

Figure 15: Coefficients from M_3

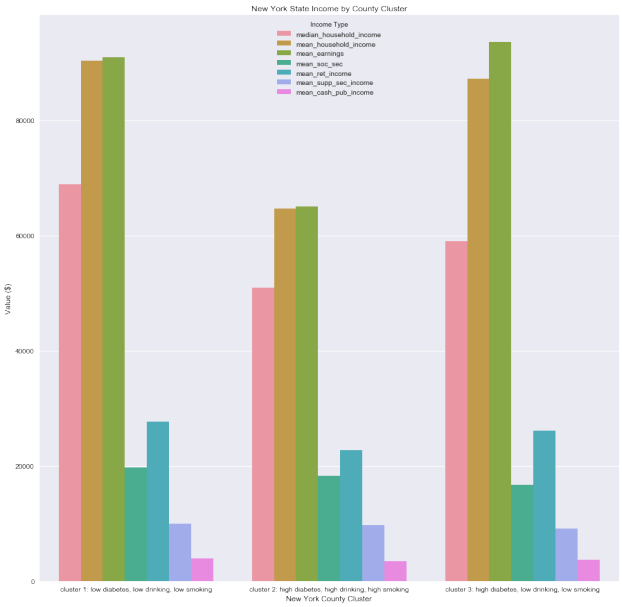

Figure 16: New York State County Income by Cluster

```
Call:
glm(formula = y ~ ., family = "binomial", data = complete)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96856  -0.39037  -0.00277   0.50650   1.93353

Coefficients:
                               Estimate Std. Error z value Pr(>|z|)
(Intercept)                   3.328e+01  1.809e+01   1.839   0.0659 .
mean_soc_sec                 -1.393e-03  1.034e-03  -1.347   0.1781
mean_household_income        -2.204e-04  3.911e-04  -0.563   0.5731
fast_food_pc                 -3.489e+03  1.687e+04  -0.207   0.8362
mean_cash_pub_income          1.871e-05  9.588e-04   0.020   0.9844
X.150.000_to_.199.999        -7.175e-04  1.049e-03  -0.684   0.4938
median_household_income       7.943e-05  3.539e-04   0.224   0.8224
non_fast_food_restaurant_pc -3.116e+02  1.418e+03  -0.220   0.8261
grocery_store                -1.503e-02  4.909e-02  -0.306   0.7594
mean_ret_income               4.004e-05  2.979e-04   0.134   0.8931
population_y                  1.145e-05  2.661e-05   0.430   0.6670
non_fast_food_restaurant     -7.294e-03  2.175e-02  -0.335   0.7374
fast_food                     1.458e-01  2.335e-01   0.624   0.5324
pop_w_public_coverage         1.162e-04  1.835e-04   0.633   0.5266
X.9.999_or_less              -7.862e-04  1.810e-03  -0.434   0.6640
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 17: Summary of Coefficients for Logistic Regression Model M_1. M_2 and M_3 had similarly insignificant p-values

6 Appendix

6.1 Variables Used in K-Means clustering

1. Age-adjusted cerebrovascular disease (stroke) mortality rate per 100,000
2. Age-adjusted cardiovascular disease mortality rate per 100,000
3. Age-adjusted cirrhosis mortality rate per 100,000
4. Age-adjusted diabetes mortality rate per 100,000
5. Age-adjusted percentage of adults overweight or obese (BMI 25 or higher)
6. Age-adjusted percentage of adults who smoke cigarettes
7. Age-adjusted percentage of adults who binge drink
8. Age-adjusted diabetes hospitalization rate per 10,000 (primary diagnosis)
9. Age-adjusted cirrhosis hospitalization rate per 10,000
10. Age-adjusted percentage of adults who did not participate in leisure time physical activity in last 30 days
11. Age-adjusted percentage of adults eating 5 or more fruits or vegetables per day

(Intercept)	mean_household_income	mean_soc_sec
2.312355e+01	-8.151555e-05	-8.381925e-04
pop_no_health_insurance	mean_cash_pub_income	non_fast_food_restaurant
-3.080975e-04	5.421822e-05	-6.750396e-03
food_stamp_benefits	mean_ret_income	median_household_income
6.931459e-04	2.574699e-05	-7.068460e-05

Figure 18: Coefficients from M_1. Other coefficients can be found in the appendix

(Intercept)	mean_household_income	mean_soc_sec
-2.899894e+01	1.613144e-04	1.122690e-03
pop_no_health_insurance	mean_cash_pub_income	non_fast_food_restaurant
4.232980e-05	-1.836645e-04	-9.954664e-04
food_stamp_benefits	mean_ret_income	median_household_income
-8.589546e-05	1.994226e-05	-5.692865e-06

Figure 19: Coefficients from M_2

6.2 Pairwise Correlation of All Health Indicators

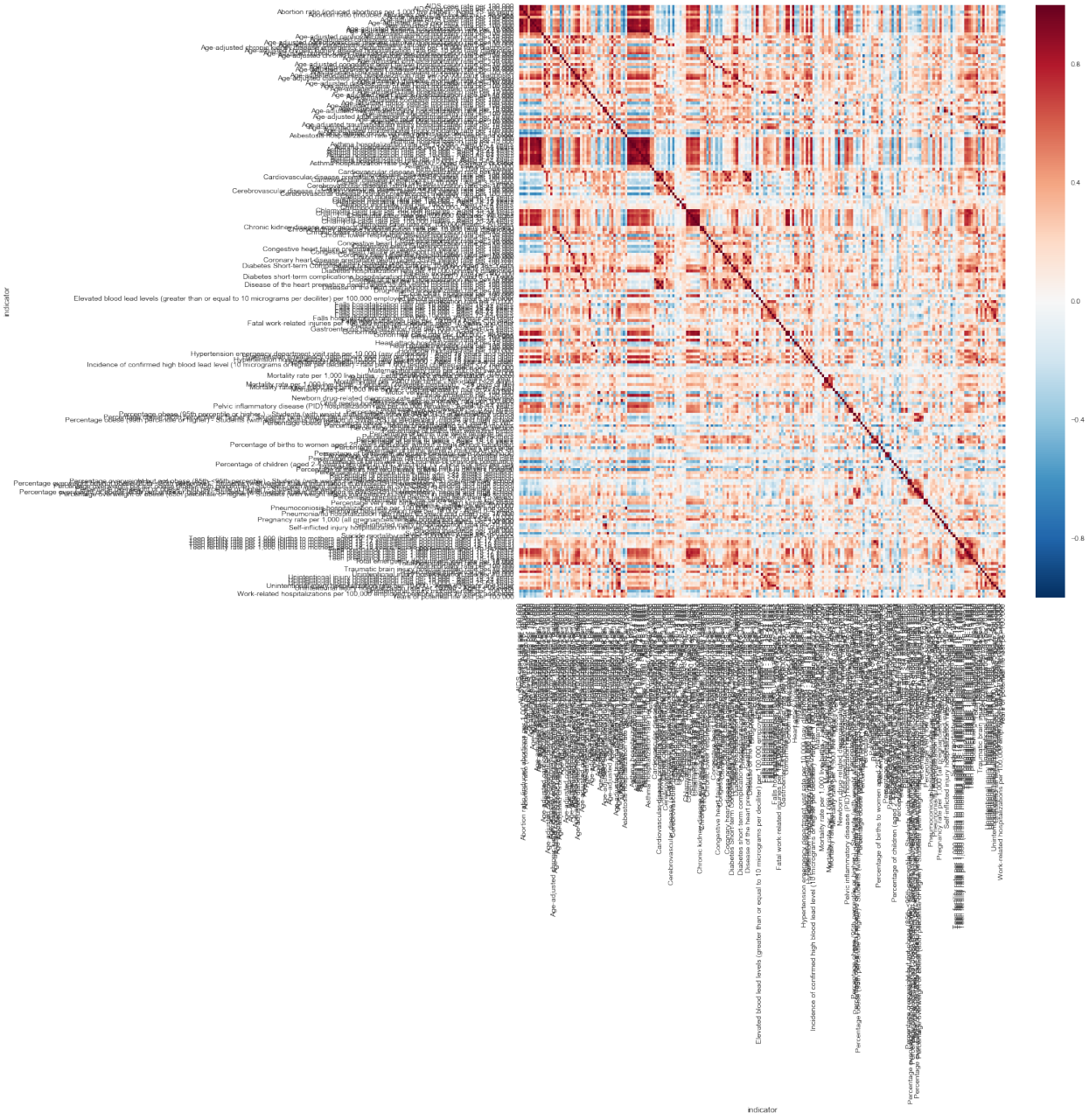

Figure 20: Pairwise Correlation of All Health Indicators