ISyE 6740 - Fall 2024
Final Report - Group 232

Team Member Names: Michael Southard and Mustafa Alsaeedi

Project Title: Where do healthy people live in the US?

Problem Statement

Understanding key factors correlating with positive health outcomes is important for driving public health policy. Here, we examine factors such as income, access to food (both "healthy" and "unhealthy" options), and access to recreational space and their relationship with health indicators such as diabetes rate, heart disease mortality, and life expectancy on the county level. We will also look for signs of racial/ethnic bias in access to health while controlling for potential differences in income. Our findings could function as an initial step towards government funding for public recreational areas, local farmers and markets, and/or grocery stores in counties lacking access.

Background

In a 2010 editorial titled "How Healthy is Your County?"[2], David Nash provides an overview of a set of public health reports called "County Health Rankings"[3] published by the Robert Wood Johnson Foundation, aimed to "aid public health and community leaders, policymakers, consumers, and others to see how healthy their county is, to compare it with others within their state, and to find ways to improve the health of their community."

Amongst many other relationships, these reports found that 80% of the least healthy counties were rural[4], and healthier counties tended to be better educated and have greater access to healthful foods and recreational areas. David Nash also noted that there are many complex social and cultural factors that must be understood and addressed "to truly improve the health of our counties."[2]

Our goals for this project were to further substantiate the above findings through statistical modeling, gain insight into the complex social and cultural aspects of health (particularly racial/ethnic bias), and possibly uncover other relationships not mentioned in the above reports.

Disclaimer: We acknowledge that we are not experts in public health and it is very likely that new studies have emerged since 2010 that discuss/expand upon the relationships we will examine in this project.

Data Sources

– *See References and Appendix for additional information.*

Our primary dataset comes from the USDA "Food Environment Atlas" (FEA)[5]. This dataset contains many useful variables at the county level, including socioeconomic factors, health factors, store access, restaurant availability, farms and markets, and recreational facilities.

We also used the following dataset for supplementary information:

> "Heart Disease Mortality Data Among US Adults (35+) by State/Territory and County": contains heart disease mortality rates by county, gender, and racial/ethnic group (2013-2015, 3-year average)[1]

Methodology

Data Preparation

Data Aggregation

To initiate our dataset, we selected 43 features from the FEA "StateAndCountyData" csv file and aggregated them by county. We then merged this dataset with the 2010 Census Population variable in "SupplementalDataCounty.csv" and the Data_Value variable from "Heart_Disease_Mortality_Data_Among_US_Adults__35___by_State_Territory_and_County.csv", incorporating only the overall heart disease mortality data. Future analysis could further examine heart disease mortality by gender and racial/ethnic group, which is also provided in this file.

Response Variable Selection

We used principal component analysis (sklearn PCA) to perform dimensionality reduction on adult diabetes rate and heart disease mortality, resulting in a response variable that had a correlation coefficient of 0.89 with both diabetes rate and heart disease mortality, and explained 79.9% of the variance in these features. We then scaled this response variable by –1 so positive values would indicate better health (lower diabetes rates and heart disease mortality), whereas negative values correspond to worse health outcomes at the county level. This variable will be referred to as "Health Score" throughout this paper.

Collinearity

Redundant collinear variables with correlation coefficients >0.8 were dropped from statistical modeling analysis or were combined with other variables to form the following ratios:

1. "healthy_store_ratio" = ratio between stores with "healthy" options (grocery + specialty + supercenters) : stores without healthy options (convenience).
2. "healthy_restaurant_ratio" = ratio between full service : fast food restaurants.
3. "recfac_pc" = recreational facilities per capita (RECFAC11 / 2010_Census_Population)
4. "fmrkt_pc" = farmers markets per capita (FMRKT13 / 2010_Census_Population)
5. "sum_farms" = sum(ORCHARD_FARMS12, BERRY_FARMS12, GHVEG_FARMS12, VEG_FARMS12, CSA12, AGRITRSM_OPS12)

Nan Analysis

We removed all variables that consisted of >5% nans besides features that related to 2010 Census Population. Then, we removed all counties with remaining nans, resulting in a dataset consisting of 2920 counties, 93.11% of the total number before nan removal.

We then compared the counties that were removed versus those retained and found a statistically significant difference in diabetes rates between the groups (t-test, p-value < 0.001), where counties with incomplete data had diabetes rates between 0.96-1.64 percentage points higher than counties with complete data (see Appendix). This bias warrants further examination, as it may indicate that less healthy counties may not have equal access to government and academic resources and attention.

Normality

Many of the variables in this dataset appeared to be exponentially or Poisson distributed, and others had bimodal distributions or lacked a clear pattern. To make this data suitable for linear regression, we converted bimodal features into dummy variables and used quantile transformations (sklearn QuantileTransformer) to coerce the remaining features into (approximately) normal distributions. Some applications in this project use the transformed dataset and others use the raw data, depending on the model assumptions and application purpose.

EDA

Our exploratory data analysis consisted mainly of visualizing linear trends for numerical variables (Seaborn regplot) and distribution differences for categorical variables (Seaborn boxplot).

Detrending

Based on our EDA, we found strong relationships between health and income/age structure at the county level. These relationships were expected, and our goal was to look for more complex interactions that could help inform beneficial policy or behavioral changes. To help uncover any such relationships, we detrended our response variable (Health Score) by subtracting the predictions from a linear model (sklearn LinearRegression) fit using the following income- and age-related variables: median household income, poverty rate, % households with low access to stores and have no vehicle, SNAP participants per 1000, % 65 and older, and % 18 and younger. This model had an R2 of 0.483 (see Appendix).

Data Partitioning

We standardized and partitioned the data with a 70-30 train-test split (sklearn StandardScaler and train_test_split) before fitting the following LassoCV, OLS, and RandomForestRegressor models. The training sets were used for alpha selection and all model fitting, and the OLS and RandomForestRegressor test R2 scores were calculated on the test set.

Feature Selection

After removing the income and age structure features, we had 16 features remaining for further exploration (see Appendix). Using the normalized data, we performed feature selection with LASSO regression, using 10-fold cross validation to select an appropriate alpha parameter (sklearn LassoCV, see appendix). LASSO regression did not remove any features from the model.

Health Models

We used linear regression (statsmodels OLS) and random forest (sklearn RandomForestRegressor) models to examine the relationship between detrended health score and race/ethnicity, access to healthy foods, food taxes, and a few other variables. Default parameters were used for both models, as increasing n_estimators did not appear to improve random forest OOB score. The model results, including training and testing R2, coefficients, and Gini importance can be found in Appendix. We also checked that the linear regression residuals were approximately normal and did not display heteroskedasticity.

For context, we also fit a random forest model using raw data (before nan removal, normalization, and detrending) to gauge the maximum proportion of variance in Health Score that could be explained by our dataset. This model used 150 estimators and had a test R2 of 0.737. Median household income was by far the most influential feature by Gini importance.

Race/ethnicity Clustering

To further explore the relationship between health and race/ethnicity, we used k-means clustering (sklearn KMeans) to group similar counties in terms of % white, % black, and % Hispanic populations. The highest silhouette score occurred with 3 clusters (see Appendix); and upon visual inspection, 3 clusters seemed to partition the data well (see Appendix). K-means clustering effectively separated the counties into 3 groups: the first group had relatively high average black populations (34.6%, 432 counties, "35PCT_BLACK"), the second group had relatively high average white populations (88.8%, 2190 counties, "89PCT_WHITE"), and the third group had relatively high average Hispanic populations (41.2%, 298, "41PCT_HISP"). Clearly, the vast majority of counties are predominately white.

Health Cluster KDE

> After identifying clusters, we used KDE (Seaborn kdeplot) to visualize the relationship between racial/ethnic cluster and health, income, and access to healthy foods. The same kdeplot parameters were used for all plots, with $levels = 5$ and $thresh = 0.1$ for the target/foreground cluster (light blue), while the default parameters were used for the background clusters.

Results and Discussion

EDA

In this section, our goal was to visualize simple linear relationships between features in our dataset. We begin exploring concepts mentioned in the "Background" section, including the correlations between health and (1) urban and rural counties; (2) access to healthy foods and recreational facilities; and (3) income, race/ethnicity, and age structure.

Urban versus rural counties

In our dataset, urban counties have better median health scores and are more populated than rural counties (see Appendix), which supports the findings from "How Healthy is Your County?"[2]

Healthy foods and recreational facilities

Farm access, healthy store ratio, healthy restaurant ratio, and recreational facility access are all positively correlated with health score in our dataset (1), also supporting the findings of "How Healthy is Your County?"[2]

Figure 1: Number of farms, healthy store ratio, healthy restaurant ratio, and recreational facility access are all positively correlated with health score. Data shown is quantile transformed.

Income and race/ethnicity

Regarding the relationship between health and wealth, counties with greater median income and lower poverty rates are healthier according to our dataset (2).

Additionally, predominately white counties tend to be slightly wealthier and predominately black counties slightly poorer. The relationship is unclear with predominately Hispanic counties, and it would be worth exploring potential nonlinearity (see Appendix).

Lastly, predominately black counties seem to have worse health outcomes, whereas Hispanic counties appear healthier (see Appendix).

This is evidence that income inequality between racial/ethnic groups could be generating the health disparities between predominately white, black, or Hispanic counties.

Age structure and race/ethnicity

Older counties have higher diabetes rates and younger counties may have higher heart disease mortality (see Appendix). This is expected because type 2 diabetes onset tends to occur later in life, and high heart disease mortality rates could skew the population towards younger individuals.

Predominately white counties tend to be relatively old and predominately Hispanic counties are

Figure 2: Health Score vs median household income and poverty rate. Health outcomes clearly improve with greater access to wealth, as expected.

relatively young. Predominately black counties have relatively few elderly people but no difference in young people (3).

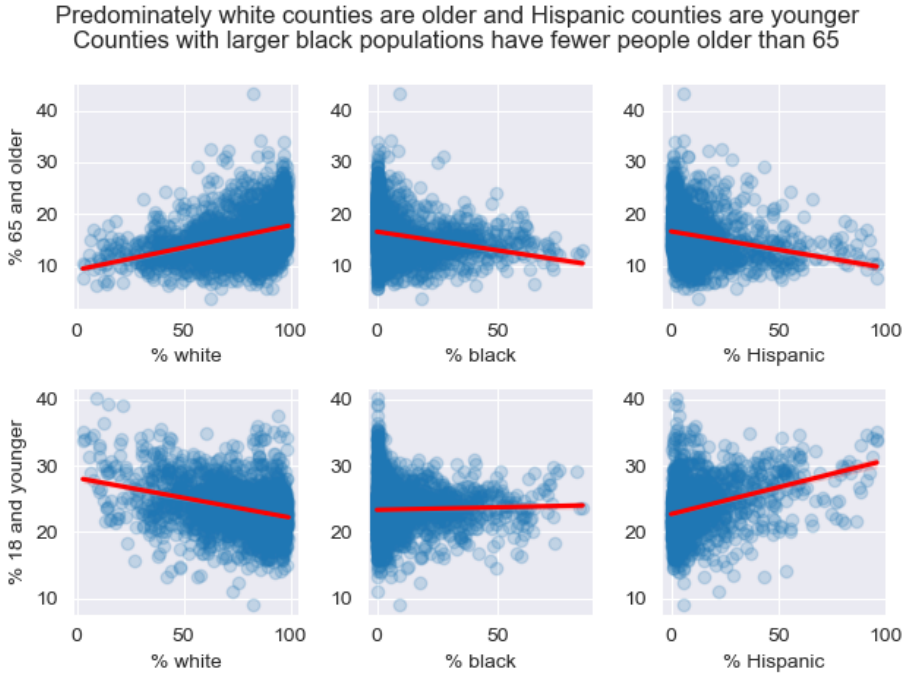

Figure 3: County age demographics by race/ethnicity

Differences in population age structure (along with income bias) between counties with different racial/ethnic composition could be causing perceived variation in health outcomes, which is why we controlled for these factors by detrending the health score in our upcoming models.

Health Models

Here, we removed the affect of wealth and age on health outcomes (detrending) to analyze any underlying and less prominent relationships that could potentially be targeted more easily through policy intervention or behavioral change, and to further explore complex cultural factors.

Our linear regression model had an R2 = 0.172 on the test data (after detrending), indicating that small but significant relationships still existed in the data after accounting for income and age structure (see Appendix). The top 4 features by coefficient absolute value were: 2010 Census Population (-0.24), % black population (-0.21), % white population (-0.17), and % Hispanic population (+0.17).

We also fit a Random Forest regression model to the same data (test R2 = 0.285). The top 3 features by Gini importance were: % black population, % Hispanic population, and % white population.

The combined results of these 2 models indicate that there may be some latent relationship between race/ethnicity and health independent of income or population age structure. Predominantly white and black counties may have poorer health outcomes than expected given their access to financial resources and relative age demographics, while the inverse is true for predominantly Hispanic counties. It may be worthwhile to further explore possible cultural explanations for these trends, including potential differences in diet, exercise, or social structure. It's also possible that the relationship between income and age structure was not entirely captured by our detrending model (e.g. nonlinear effects or missing features), in which case there may not be any significant cultural differences affecting health outcomes. Either way these findings warrant further analysis.

Another important finding from the linear regression model is that the ratio of healthy (full service) to unhealthy (fast food) restaurants; healthy (grocery, specialty, and supercenter/club) to unhealthy (convenience) stores; and the number of produce farms all had positive and significant (p-value $<$ 0.001) correlations with health, even after income detrending. Counties with farmers markets were also significantly healthier. This result suggests that government policy incentivizing healthy stores, restaurants, farms, and farmers markets could have a positive impact on public health.

Lastly, counties with general food taxes had significantly worse health (p-value $<$ 0.001), suggesting that food taxation might incentive citizens to purchase cheaper and lower quality foods, resulting in poor health outcomes.

Race/Ethnicity Clustering

Here, we expand on the previous modeling results by further examining the relationship between race/ethnicity and health, income, and access to healthy foods.

Health and Income

First, we revisited the correlation between health score and income, this time separating the counties by racial/ethnic clusters. In Figure 4, we once again see that counties with relatively large black populations tend to have below average health outcomes and median household incomes, whereas counties with relatively large Hispanic populations tend to have better health outcomes but a wide range of incomes. This is in line with previous findings.

Healthy Store Access

Next, we checked the relationship between racial/ethnic clusters and healthy store access, namely "Healthy Store Ratio" and "% population with low access to stores". We did not see clear distinctions between groups in this figure, although counties with relatively large black populations may have slightly below average access to healthy stores (see Appendix).

Detrended Health Score and Restaurant Access

Lastly, we looked at the relationship between detrended Health Score and healthy restaurant access 5. "Healthy Restaurant Ratio" was the fourth most important feature in our detrended random

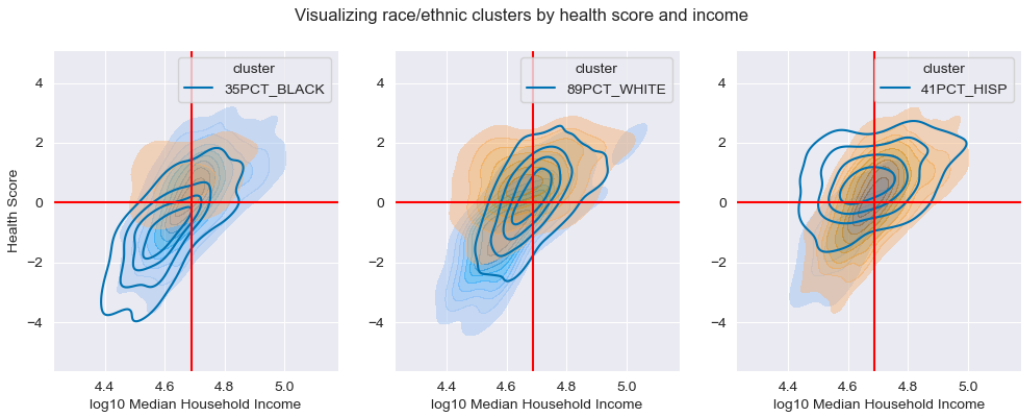

Figure 4: Counties with relatively large black populations have below average Health Scores and incomes. Counties with large Hispanic populations have above average Health Scores. Red lines denote average values.

forest model, only behind % black, % Hispanic, and % white. Here, we can see that counties with relatively large black and Hispanic populations both have below average access to healthy restaurants (more fast food), yet Hispanic counties have slightly better health outcomes than expected whereas black counties have slightly worse health outcomes. Additionally, predominately white counties have slightly worse health outcomes than expected, given above average access to healthy restaurants.

These results are fascinating, and add another layer to the complex relationship between health and culture. Future analysis should further categorize restaurants not just by the labels of "fast food" or "full service", but by the nutritional value of the food they sell. There is certainly a wide variety of cuisines sold at both full service and fast food restaurants, and the cultural influence on the types of restaurants present in different counties could be impacting the overall health of these communities. We should also contrast the dietary preferences of Hispanic Americans with those of black and white Americans to see whether there are significant differences in nutritional quality. It would also be worthwhile to add other racial/ethnic groups into this analysis to potentially find additional trends.

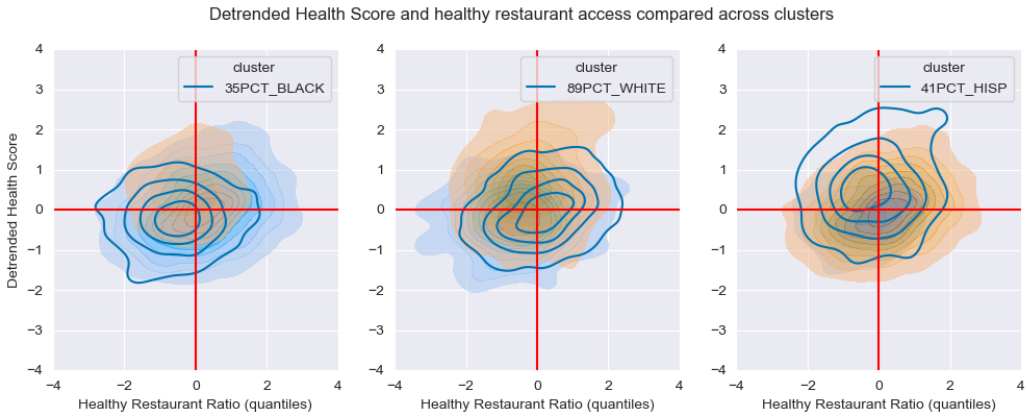

Figure 5: Counties with relatively large black populations have below average health scores and access to full service restaurants. Predominately Hispanic counties tend to be healthier despite below average access to full service restaurants, while the inverse is true for white counties. Red lines denote average values.

Conclusion

The main finding of this report is that there is a clear correlation between health and wealth, and that black Americans seemingly have below average access to both. Long-term policy solutions must continue to prioritize creating equal quality of life for all people in the US. Secondly, we found that access to healthy farms, markets, restaurants, and stores does have a positive correlation with health outcomes, even after accounting for the impact of wealth. Incentivizing businesses that provide healthy foods and changing individual shopping habits could be an easy way to improve overall community health. And finally, we presented evidence of possible cultural influence on health outcomes, where predominately Hispanic counties had above average health outcomes despite the prominence of fast food restaurants, and predominately white counties had greater access to full service restaurants yet still suffered below average health outcomes. The complex relationship between culture and health warrants further exploration.

References

[1] Centers for Disease Control and Prevention (CDC). *Heart Disease Mortality Data Among U.S. Adults (35+) by State, Territory, and County*. Accessed: December 7, 2024. 2024. URL: https://catalog.data.gov/dataset/heart-disease-mortality-data-among-us-adults-35-by-state-territory-and-county.

[2] David B. Nash. "How healthy is your county?" In: *P & T: A Peer-Reviewed Journal for Formulary Management* 35.8 (2010), p. 427.

[3] Robert Wood Johnson Foundation and University of Wisconsin. *How Healthy is Your County? New County Health Rankings Give First County-by-County Snapshot of Health in Each State*. Available at: www.countyhealthrankings.org/latest-news/press-release-how-healthy-your-county-new-county-health-rankings-give-first-county-count. Accessed: June 30, 2010. Feb. 2010.

[4] Doug Trapp. *Health Status Varies by County: Where Patients Live Matters*. Available at: www.ama-assn.org/amednews/2010/03/01/gvl10201.htm. Accessed: June 30, 2010. Mar. 2010.

[5] United States Department of Agriculture, Economic Research Service. *Food Environment Atlas: Data Access and Documentation Downloads*. Accessed: December 7, 2024. 2024. URL: https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/.

Appendix

1. ***Food Environmental Atlas documentation and variable selection***

   Interactive atlas link:
   https://www.ers.usda.gov/data-products/food-environment-atlas/go-to-the-atlas/

   Initial variables screened:
   All the following variables are recorded at the county level unless otherwise specified.
   More recent data was available for some variables, but we chose to use data between 2010-2015 to align in time with our health indicator variables: adult diabetes rate (2013) and heart disease mortality (2013-2015).

   Socioeconomic

   "PCT_65OLDER10" - % 65 and older in 2010

   "PCT_18YOUNGER10" - % 18 and younger in 2010

   "MEDHHINC15" - median household income 2015

   "POVRATE15" - poverty rate 2015

   "METRO13" - metro/non-metro counties 2010

   "PCT_NHWHITE10" - % population identifying as white 2010

   "PCT_NHBLACK10" - % population identifying as black 2010

   "PCT_HISP10" - % population identifying as Hispanic 2010

   Store access and food assistance

   "PCT_LACCESS_POP15" - % population with low access to store 2015

   "PCT_LACCESS_LOWI15" - % low income and low access to store 2015

   "PCT_LACCESS_HHNV15" - % households w no car and low access to store 2015

   "PCT_LACCESS_SNAP15" - % SNAP households w low access to store 2015

   "PCT_LACCESS_CHILD15" - % children w low access to store 2015

   "PCT_LACCESS_SENIORS15" - % seniors w low access to store 2015

   "GROC11" - grocery stores 2011

   "SUPERC11" - supercenters and club stores 2011

   "CONVS11" - convenience stores 2011

   "SPECS11" - specialized food stores 2011

   "SNAPSPTH12" - SNAP authorized stores per 1000 people 2012

   "PC_SNAPBEN12" - SNAP benefits per capita 2012

   Restaurant availability

   "FFR11" - fast food restaurants 2011

   "FSR11" - full service restaurants 2011

   Food tax (by state)

   "SODATAX_STORES14" - soda sales tax retail stores 2014

"SODATAX_VENDM14" - soda sales tax vending machines 2014

"CHIPSTAX_STORES14" - chips and pretzels sales tax retail stores 2014

"CHIPSTAX_VENDM14" - chips and pretzels sales tax vending machines 2014

"FOOD_TAX14" - general food sales tax 2014

Local foods

"DIRSALES_FARMS12" - farms with direct sales 2012

"DIRSALES12" - direct farm sales 2012 ($1000s)

"PC_DIRSALES12" - direct farm sales per capita 2012 ($1s)

"FMRKT13" - farmers markets 2013

"FMRKT_SNAP13" - farmers markets that report accepting SNAP 2013

"PCT_FMRKT_FRVEG13" - % farmers markets that report selling fruits and vegetables 2013

"PCT_FMRKT_ANMLPROD13" - % farmers markets that report selling animal products 2013

"VEG_FARMS12" - vegetable farms 2012

"VEG_ACRES12" - vegetable acres harvested 2012

"ORCHARD_FARMS12" - orchard farms 2012

"ORCHARD_ACRES12" - orchard acres 2012

"BERRY_FARMS12" - berry farms 2012

"BERRY_ACRES12" - berry acres 2012

"SLHOUSE12" - small slaughterhouse facilities 2012

"GHVEG_FARMS12" - greenhouse veg and herb farms 2012

"CSA12" - CSA farms 2012

"AGRITRSM_OPS12" - agritourism operations 2012

"AGRITRSM_RCT12" - agritourism receipts 2012 ($1s)

Recreation and fitness

"RECFAC11" - recreation and fitness facilities 2011

Supplemental data

2010 census population

2. ***Nan bias***

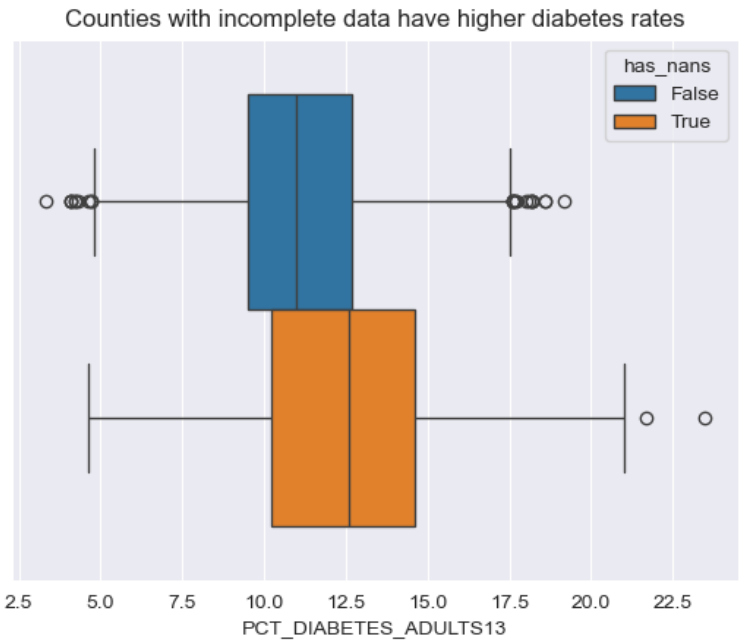

Figure 6: Diabetes rates are higher in counties with incomplete data

3. ***Detrending model***

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    PCA   R-squared:                       0.483
Model:                            OLS   Adj. R-squared:                  0.482
Method:                 Least Squares   F-statistic:                     454.3
Date:                Sun, 08 Dec 2024   Prob (F-statistic):               0.00
Time:                        00:23:59   Log-Likelihood:                -3198.2
No. Observations:                2920   AIC:                             6410.
Df Residuals:                    2913   BIC:                             6452.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             -0.0032      0.013     -0.242      0.809      -0.030       0.023
PCT_18YOUNGER10   -0.1574      0.016     -9.577      0.000      -0.190      -0.125
PCT_65OLDER10     -0.1223      0.020     -6.251      0.000      -0.161      -0.084
MEDHHINC15         0.2654      0.037      7.197      0.000       0.193       0.338
PCT_LACCESS_HHNV15 -0.2465     0.016    -15.128      0.000      -0.278      -0.215
POVRATE15         -0.1525      0.034     -4.440      0.000      -0.220      -0.085
SNAPSPTH12        -0.1269      0.016     -7.895      0.000      -0.158      -0.095
==============================================================================
Omnibus:                       88.399   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              162.142
Skew:                           0.232   Prob(JB):                     6.18e-36
Kurtosis:                       4.057   Cond. No.                         6.52
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 7: Health Score vs all income and age structure features. Predictions from this model were used to detrend Health Score for future models.

4. ***LassoCV alpha***

   Graphic code from: https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_model_selection.html#sphx-glr-auto-examples-linear-model-plot-lasso-model-selection-py

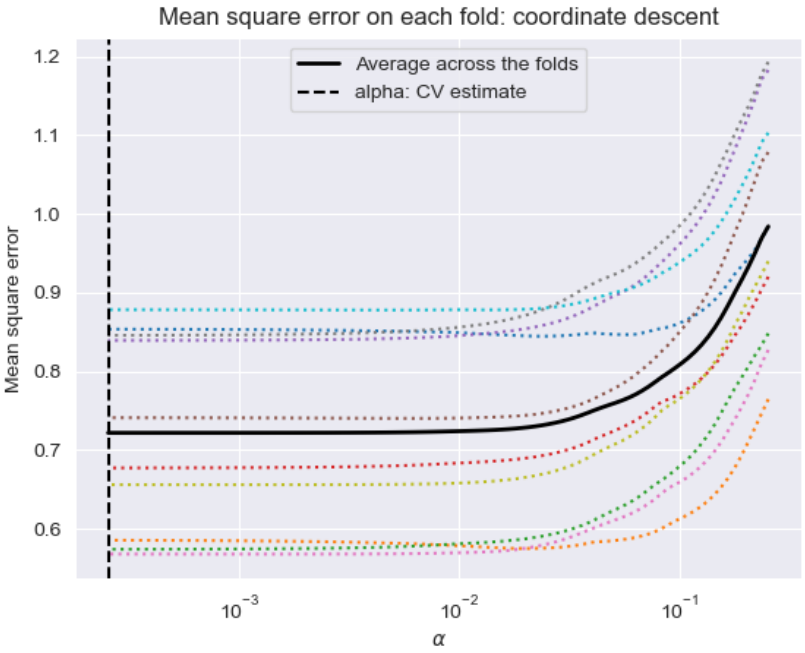

Figure 8: LASSO 10-fold cross validation alpha selection

5. ***Statsmodels OLS results***

```
                            OLS Regression Results
==============================================================================================
Dep. Variable:     PCA_dt_allinc_popstr   R-squared (uncentered):                   0.279
Model:                             OLS   Adj. R-squared (uncentered):              0.273
Method:                  Least Squares   F-statistic:                              48.96
Date:                 Sat, 07 Dec 2024   Prob (F-statistic):                   4.32e-131
Time:                         17:42:18   Log-Likelihood:                         -2549.9
No. Observations:                2044   AIC:                                      5132.
Df Residuals:                    2028   BIC:                                      5222.
Df Model:                          16
Covariance Type:              nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
DIRSALES_FARMS12             0.1145      0.034      3.329      0.001       0.047       0.182
METRO13                     -0.0466      0.024     -1.949      0.051      -0.093       0.000
PCT_HISP10                   0.1716      0.031      5.540      0.000       0.111       0.232
PCT_LACCESS_POP15            0.0635      0.034      1.892      0.059      -0.002       0.129
PCT_LACCESS_SNAP15           0.0646      0.034      1.923      0.055      -0.001       0.130
PCT_NHBLACK10               -0.2086      0.031     -6.652      0.000      -0.270      -0.147
PCT_NHWHITE10               -0.1731      0.039     -4.459      0.000      -0.249      -0.097
2010_Census_Population      -0.2393      0.041     -5.878      0.000      -0.319      -0.159
healthy_store_ratio          0.1541      0.021      7.356      0.000       0.113       0.195
healthy_restaurant_ratio     0.1252      0.021      5.884      0.000       0.083       0.167
sum_farms                    0.1186      0.032      3.674      0.000       0.055       0.182
has_food_tax                -0.1293      0.021     -6.147      0.000      -0.171      -0.088
has_soda_tax                -0.0529      0.020     -2.591      0.010      -0.093      -0.013
has_recfac                   0.0342      0.025      1.369      0.171      -0.015       0.083
has_slaughter                0.0156      0.020      0.793      0.428      -0.023       0.054
has_fmrkt                    0.0848      0.023      3.710      0.000       0.040       0.130
==============================================================================
Omnibus:                      129.062   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              477.613
Skew:                          -0.198   Prob(JB):                    1.94e-104
Kurtosis:                       5.335   Cond. No.                         5.41
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 9: Statsmodels OLS results. Detrended health score vs normalized features.

6. ***RandomForestRegressor Gini importance***

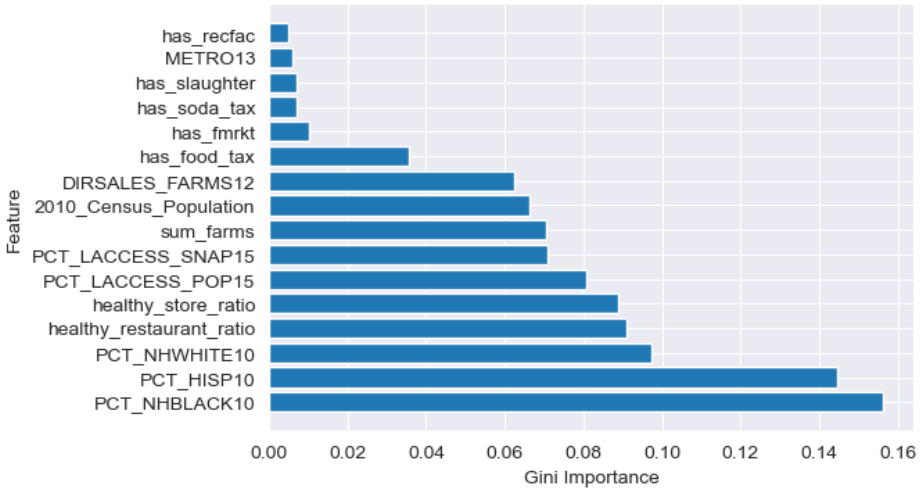

Figure 10: Random Forest Gini importance. Detrended health score vs normalized features.

7. ***Race/ethnicity k-means scores***

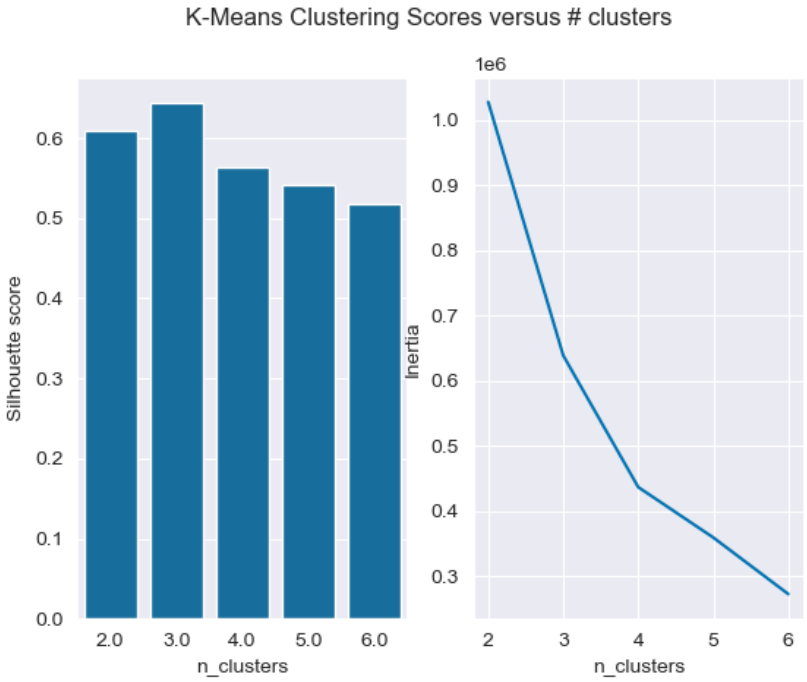

Figure 11: K-means clustering scores by number of clusters

8. ***Race/ethnicity cluster distributions***

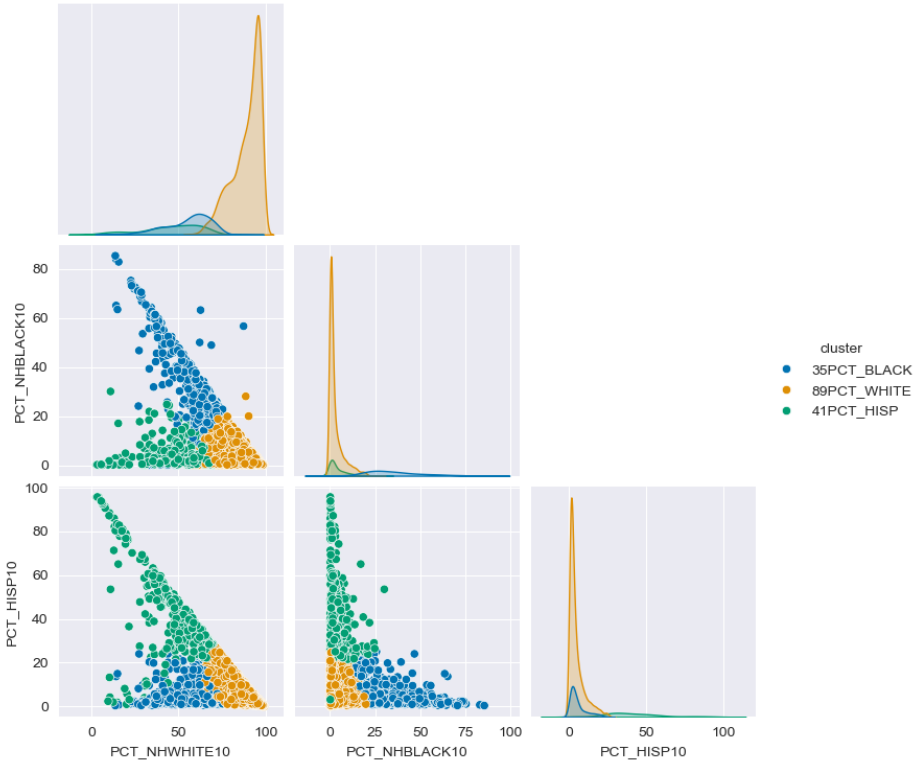

Figure 12: K-means clustering with 3 clusters seems to partition the data effectively

9. ***Urban vs rural counties***

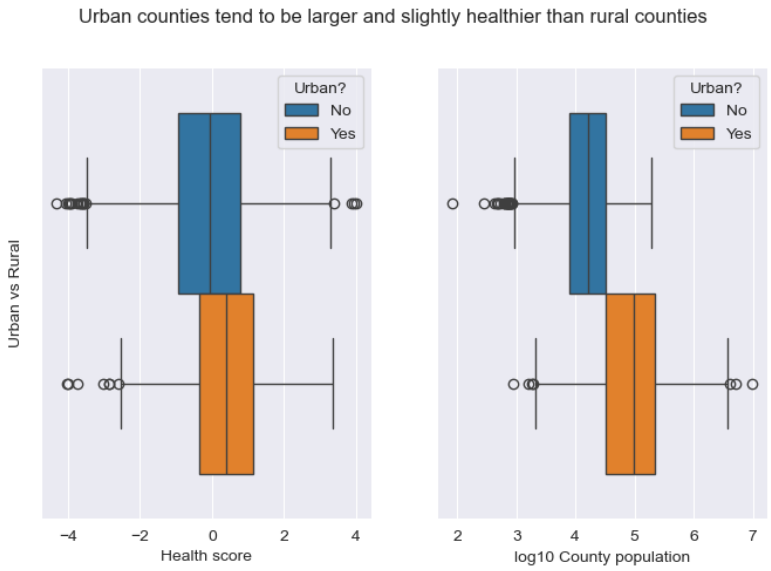

Figure 13: Urban counties are healthier and more populated (by median)

10. ***Income and race/ethnicity***

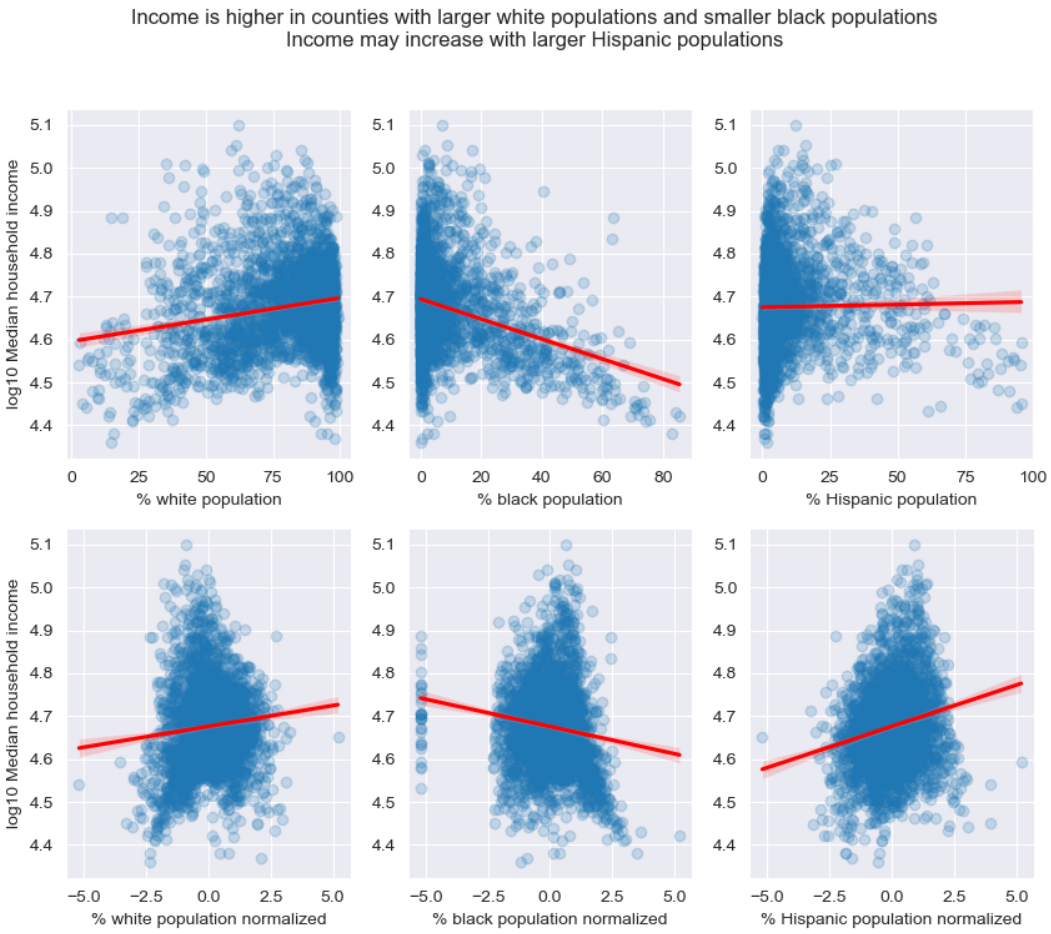

Figure 14: Median household income by race/ethnicity

11. ***Health and race/ethnicity***

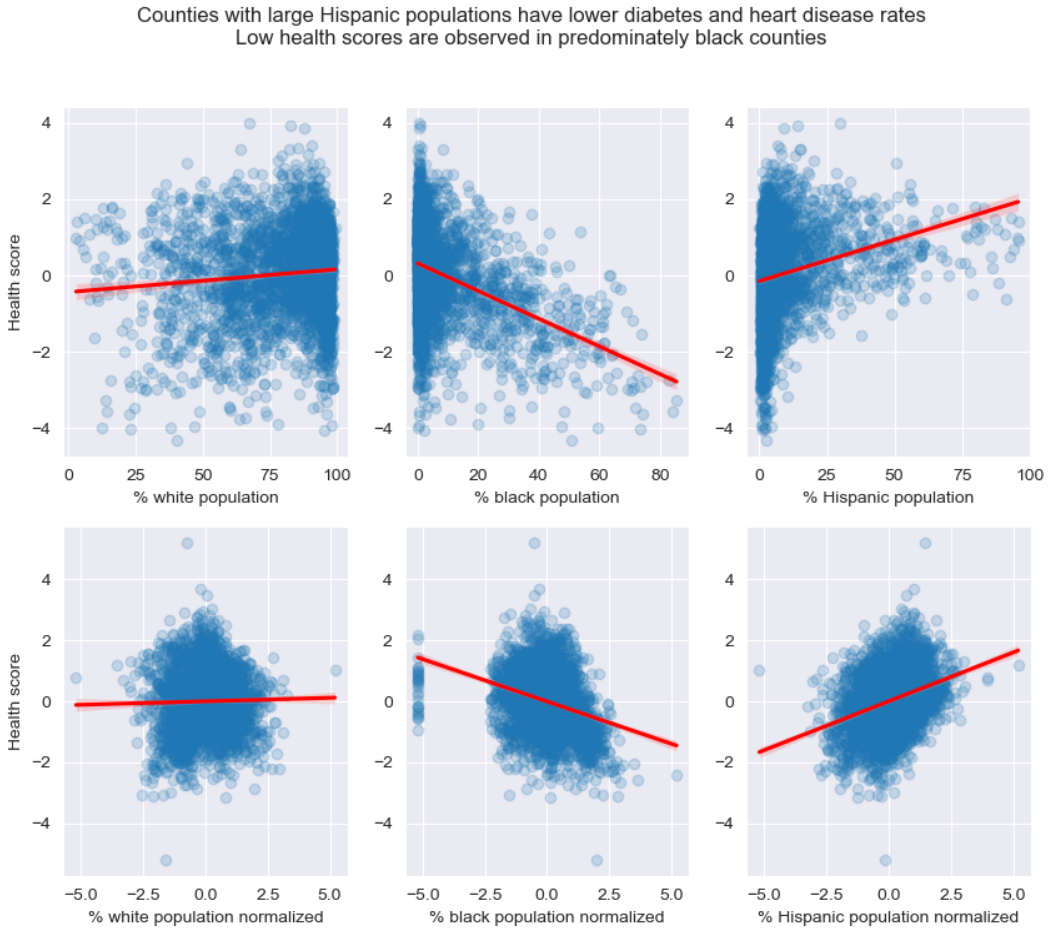

Figure 15: Health Score by race/ethnicity

12. ***Age structure and chronic disease***

Figure 16: Counties with older populations have higher diabetes rates and counties with younger populations may have higher heart disease mortality

13. ***Healthy store access by racial/ethnic cluster***

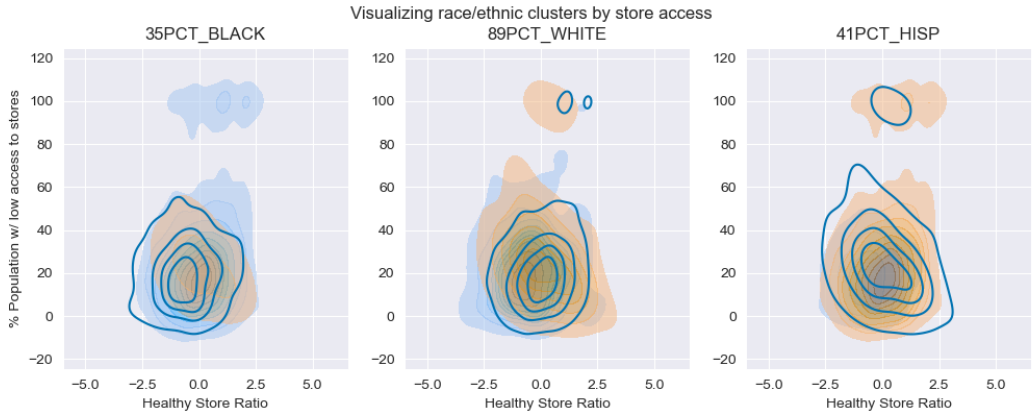

Figure 17: There are no clear trends regarding healthy store access and racial/ethnic clusters.