Group 5:
111072134 駱櫻文 (Ying-Wen Luo)   111072138 廖芷萱 (Mevis Liao)
111072219 劉宣亭 (Fiona)   111048219 黃晟傑 (Ryder)

Food Environment and Health Outcomes: Evidence on Diabetes and Obesity

1. Introduction:

The relationship between the food environment and public health has become increasingly important in understanding the reasons for the prevalence of chronic diseases. This study examines how different dimensions of food access, from retail stores and restaurants to government assistance programs, relate to adult diabetes and obesity rates across the U.S. counties.

Our research focuses on two primary health issues: adult diabetes rates(2015 and 2019) and adult obesity rates(2017 and 2022). We analyzed whether the availability and distribution of various food sources contribute to these health outcomes and how this relationship changed over time.

According to previous research, the presence of a large number of stores is not a sign of healthy food availability. Research in the Mississippi Delta shows that the SNAP-authorized retailers in high-obesity counties are mainly convenience stores, where healthy choices are limited and much more expensive than at grocery stores. This setting essentially restricts consumers’options to cheap, high-calorie processed food, increasing risks for chronic disease.

These associations will be useful in informing food-related policy decisions, distribution of resources to the public health program, and urban planning. In finding the food environment factors that are most strongly related to diabetes and obesity, we seek to provide evidence to help guide modifications of the food environment to enhance public health.

2. Research Design:

2.1 Data Cleaning and Standardization

It is the utmost importance to ensure that the statistics in our data has the reliability and accuracy to perform an authentic analysis, therefore, we have to make sure that numbers without clarity to be removed and converted, whereas dataset should all be standardized to ensure that coefficients are accurately measured and not calculated based on pure value. To clearly illustrate our process, initially , we remove unavailable data such as -8888 and replace it with NA. Subsequently, we apply Z score standardization to ensure independent variables were all standardized and ready to be analyzed.

2.2 Linear Regression Model(OLS)

In our first part of analysis, we adopted the pooled ordinary least squares regression model (OLS), which its main function would be to identify significant relationships between food variables and health outcomes while measuring one specific variable when holding other factors constant, this establishes connections and relationships between the not so related factors and gives a specific outcome to prove the relationship of variables and outcomes. By checking adjusted R square and p-value, we could filter out the low explanatory data and leave significant related variables for our research. Simultaneously, we check for multicollinearity, to ensure a better quality of explanation power.

DIABETES15 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ GROCPTH16 + $\boldsymbol{\beta 2}$ SUPERCPTH16 + $\boldsymbol{\beta 3}$ SPECSPTH16 + $\boldsymbol{\beta 4}$ SNAPSPTH17 + $\boldsymbol{\varepsilon}$
(A demonstration of the model which discusses factors that influence diabetes)

2.3 PCA analysis

OLS is a great measurement in our research but still possesses a flaw that was previously mentioned, the multicollinearity and overlapping mechanisms issues. Therefore, we try to analyze the whole data in another perspective, providing a more complete understanding and analysis on the vast numbers. We apply the PCA analysis (which we learned in the previous class) as it stands out by grouping and combining the correlated variables into a single component, which then clears out the doubt and helps provide a better explanation power while validating the OLS findings, ensuring that our explanation could be trusted and consistent.

DIABETES15 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ PC1 + $\boldsymbol{\beta 2}$ PC2 + $\boldsymbol{\beta 3}$ PC3 + $\boldsymbol{\beta 4}$ PC4 + ε
(PCA demonstration to establish models on explanation)

2.4 Visualization

With the data and analysis provided on the previous two steps, we sure hold a better picture of the research we aim to discuss. We then summarize and create graphs and tables which would benefit in explaining research results and show clearer insights of the variables that were identified to be the most relevant ones with the relative impacts to obesity and diabetes, the main discussion and analysis of our research.

2.5 Suggestion

Finally, after our research, we tend to provide suggestions and recommendations based on our findings which include several policies and interventions that can help to improve health conditions and decrease obesity and diabetes rate by improving the food environment.

3. Data

The dataset originates from the United States Department of Agriculture (USDA), which has made a huge data base, named the Food Environment Atlas. All information and data are from true sources covering varying years and geographic levels. Spanning over 300 variables, the department has made the data useful in identifying factors which connect unforeseen relationships , also by providing an overview of a community's ability to access healthy food, then building up determinants on research of food and environment.

The biggest reason for choosing this vast dataset was purely based upon the wide range and variety of data it provides, with indicators on different types of stores, percentage rates of factors related to food and life. A little peek into the data has already brought up huge interests of our group to dig deeper into the dataset , finding the connections between various dimensions we would never have thought could have been related.

Mainly focusing on variables related to the local food environment (e.g. farmers' market density, direct farm sales, and CSA farms) and health status (such as adult obesity and diabetes rates), by collecting and analyzing the vast dataset, we have had the ability to summarize and analyze the most important data that we do believe deeply affect health outcomes, where the main issue to be discussed is by how an environment would affect an individual's health. By choosing the data below, also holding socioeconomics factors as control variables, we could form a complete analysis on the current question made and drive answers which explain the interesting insights found in our study.

Data Name	Explanation
GROCPTH16/20	● This data shows the density of grocery stores per 1000 population for the years 2016 & 2020, respectively.
SUPERCPTH16/20	● This data shows the density of supercenters and club stores per 1000 population for the years 2016 & 2020, respectively.
SPECSPTH16 / 20	● This data shows the density of specialized food stores per 1000 population for the years 2016 & 2020, respectively. (Stores that sell specific categories ex: Bakery, Fish Market, Butcher)
SNAPSPTH17 / 23	● This data shows the density of SNAP- authorized stores per 1000 population for the years 2017 & 2023, respectively. (Stores that accept food stamps)
DSPTH20	● This data shows the density of dollar stores per 1000 population for the years 2020, respectively. (discount shops)
FFRPTH16 / 20	● This data shows the density of fast food restaurants per 1000 population for the years 2016 & 2020, respectively.
FSRPTH16 / 20	● This data shows the density of full service restaurants per 1000 population for the years 2016 & 2020, respectively.

PCT_SNAP17 / 22	● This data shows the percentage of the country's population accepting SNAP programs of the years 2017 & 2022, respectively.
PCT_CACFP17 / 21	● This data shows the participation rate of the country's population with child and adult care food programs of the years 2017 & 2021, respectively.
FOOD_BANKS18 / 21	● This data shows the numbers of food banks of the years 2018 & 2021, respectively.
FOODINSEC_18_20	● This data shows the percentage of insecurity rate of the country's population in the years 2018 & 2020, respectively.
VLFOODSEC_18_20	● This data shows the percentage of the very low food security rate of the years 2018 & 2020, respectively. (with three year average, families that had to skip meals because of no money)
SODATAX_STORES14 / SODATAX_VENDM14	● This data shows the sales tax rate charged from soda bought in vending machines in 2014.
CHIPSTAX_STORES14 /CHIPSTAX_ VENDM14	● This data shows the sales tax rate charged from chips and snacks bought in vending machines in 2014.
PC_DIRSALES17	● The dollar amount of food sales that was sold directly from farmers to people(per person) in 2017.
FMRKTPTH/18	● This data shows the density of markets that farmers directly sell products to consumers per 1000 population for the year 2018.
FARM_TO_SCHOOL19	● Participation indicator of schools that have cooperated with programs to buy local food in 2019.
PCT_HSPA21	● This data shows the percentage of high school students who are physically active on all 7 days of the week in 2021.
RECFACPTH20	● This data shows the density of recreation & fitness facilities per 1000 population for the year 2020.
MEDHHINC21 — Median household income (USD)	● This data shows the median household income in 2021.
POVRATE21 — Poverty rate (%)	● This data shows the percentage of the poverty rate of the total population living below country standards in 2021.
CHILDPOVRATE21 — Child poverty rate (%)	● This data shows the percentage of poverty rate with people aged under 18 in 2021.

4. Methodology & Results:

Multi Regression

Under multi regression, we aim to examine the relationship between health status (diabetes in 2015/2019 and obesity in 2017/2022) and each category using multi-linear regression. First, we transformed values which are -9999 and -8888 into NA, and did the standardization. Later on, we decided to apply multiple linear regression within each category (such as food stores, assistance, socioeconomic factors, etc.) and check multicollinearity. Finally, we used multiple linear regression to analyze all variables together.
We set the significance level to $\alpha = 0.05$. In addition, we decide to ignore the categories whose adjusted R-squared values are below 0.09 due to their lower explanatory power.

4.1 Adult Diabetes Rate:

For diabetes-to-assistance, local, and health categories in 2015, they are not available in the dataset. For all other categories, the regression models are as follows.

4.1.1 Overview of Category Models - Adult Diabetes Rate in 2015

Diabetes to Food Stores

DIABETES15 = β_0 + β_1 GROCPTH16 + β_2 SUPERCPTH16 + β_3 SPECSPTH16 + β_4 SNAPSPTH17 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.56	0.0477	0
GROCPTH16	-0.485	0.108	7.58e- 6
SUPERCPTH16	-0.354	0.0437	1.05e-15
SPECSPTH16	-0.567	0.0743	3.98e-14
SNAPSPTH17	1.10	0.0540	1.78e-81

Adj. R^2 = 0.2746

Table 1

Diabetes to Restaurants

DIABETES15 = β_0 + β_1 FFRPTH16 + β_2 FSRPTH16 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.31	0.0272	0
FFRPTH	0.294	0.0312	9.16e-21
FSRPTH	-0.701	0.03379	7.28e-90

Adj. R^2 = 0.1273

Table 2

4.1.2 Overview of Category Models - Adult Diabetes Rate in 2019

Diabetes to Food Stores

DIABETES15 = $\beta 0$ + $\beta 1$ GROCPTH16 + $\beta 2$ SUPERCPTH16 + $\beta 3$ SPECSPTH16 + $\beta 4$ SNAPSPTH17 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	9.56	0.0821	0
GROCPTH	-0.0859	0.174	0.621
SUPERCPTH	-0.187	0.0615	0.00245
SPECSPTH	-0.804	0.131	0.0000000133
SNAPSPTH17	0.907	0.158	0.0000000159
DSPTH	1.01	0.184	0.0000000651

Adj. R^2 = 0.3657

Table 3

Diabetes to Restaurants

DIABETES19 = $\beta 0$ + $\beta 1$ FFRPTH16 + $\beta 2$ FSRPTH16 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.81	0.0303	0
FFRPTH	0.383	0.0357	2.37e-26
FSRPTH	-0.790	0.0403	8.55e-80

Adj. R^2 = 0.1291

Table 4

Diabetes to Assistance

DIABETES19 = $\beta 0$ + $\beta 1$ PCT_SNAP17 + $\beta 2$ PCT_CACFP17 + $\beta 3$ FOOD_BANKS + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	6.71	0.116	0
PCT_SNAP17	0.155	0.00871	3.72e-67
PCT_CACFP17	-0.0544	0.0276	4.89e- 2
FOOD_BANKS	0.378	0.09801	1.19e- 4

Adj. R^2 = 0.09711

Table 5

Diabetes to Local

DIABETES19 = $\beta 0$ + $\beta 1$ PC_DIRSALES17 + $\beta 2$ FMRKTPTH18 + $\beta 3$ FARM_TO_SCHOOL19 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.60	0.111	0
PC_DIRSALES17	-0.00381	0.000774	9.38e- 7
FMRKTPTH18	-0.329	0.0365	4.12e-19

FARM_TO_SCHOOL19	**0.261**	0.115	2.39e-2

Adj. R^2 = 0.1305

Table 6

Diabetes to Health

DIABETES19 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ RECFACPTH20 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.94	0.0427	0
RECFACPTH20	-0.702	0.0429	9.22e-55

Adj. R^2 = 0.1734

Table 7

4.1.3 Models that Adjusted R^2 < 0.09

In the category of regression models, some models exhibit very low explanatory power, with adjusted R-squared below 0.09. Which are:

2015

- Diabetes to Taxes, Adj, R^2 = 0.03591:

  DIABETES15 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ SODATAX_STORES14 + $\boldsymbol{\beta 2}$ SODATAX_VENDM14 + $\boldsymbol{\beta 3}$ CHIPSTAX_STORES14 + $\boldsymbol{\beta 4}$ CHIPSTAX_VENDM14 + ε

2019

- Diabetes to Insecurity, Adj, R^2 = 0.07823:

  DIABETES19 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ FOODINSEC_18_20 + $\boldsymbol{\beta 2}$ VLFOODSEC_18_20 + ε
- Diabetes to Taxes, Adj, R^2 = 0.044:

  DIABETES19 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ SODATAX_STORES14 + $\boldsymbol{\beta 2}$ SODATAX_VENDM14 + $\boldsymbol{\beta 3}$ CHIPSTAX_STORES14 + $\boldsymbol{\beta 4}$ CHIPSTAX_VENDM14 + ε

Therefore, variables from these models with adjusted R^2 < 0.09 are excluded from the regression model with all variables in Section 4.1.4, and only categories with relatively stronger explanatory performance are retained.

4.1.4 Diabetes to all of the variables

Note that these two models were built using only the variable categories whose individual models' adjusted R-squared are above 0.09.

2015

DIABETES15 = $\boldsymbol{\beta 0}$ + $\boldsymbol{\beta 1}$ GROCPTH16 + $\boldsymbol{\beta 2}$ SUPERCPTH16 + $\boldsymbol{\beta 3}$ SPECSPTH16 + $\boldsymbol{\beta 4}$ SNAPSPTH17 +

$\beta 5$ FFRPTH16 + $\beta 6$ FSRPTH16 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.59	0.0508	0
GROCPTH16	-0.198	0.109	6.91e- 2
SUPERCPTH16	-0.321	0.0422	5.23e-14
SPECSPTH16	-0.216	0.0788	6.16e- 3
SNAPSPTH17	1.02	0.0525	2.07e-75
FFRPTH16	0.2567	0.0510	5.54e- 7
FSRPTH16	-0.865	0.07825	2.21e-27

Table 8

Adj. $R^2 = 0.3286$

The regression model with all variables for 2015 has an adjusted R^2 of 0.3286, indicating a moderate explanatory power for county-level diabetes rates. Several food retail and restaurant variables show statistically significant associations. The model shows that food environments emphasizing diverse food retail and dining options are related to lower diabetes prevalence. In other words, counties with greater access to healthier food outlets and balanced dining structures tend to have lower diabetes rates.

SNAP-authorized store density (SNAPSPTH17) is positively associated with diabetes, which may reflect underlying socioeconomic conditions rather than a direct effect of accessibility to SNAP retail stores. Overall, the 2015 results suggest that both food retail structure and restaurant composition are closely related to diabetes prevalence across counties.

2019

DIABETES19 = $\beta 0$ + $\beta 1$ GROCPTH + $\beta 2$ SUPERCPTH + $\beta 3$ SPECSPTH + $\beta 4$ SNAPSPTH17 + $\beta 5$ DSPTH + $\beta 6$ FFRPTH + $\beta 7$ FSRPTH + $\beta 8$ PCT_SNAP17 + $\beta 9$ PCT_CACFP17 + $\beta 10$ FOOD_BANKS + $\beta 11$ PC_DIRSALES17 + $\beta 12$ FMRKTPTH18 + $\beta 13$ ARM_TO_SCHOOL19 + $\beta 14$RECFACPTH + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	8.23	0.499	2.71e-50
GROCPTH	0.129	0.171	4.50e- 1
SUPERCPTH	-0.0402	0.0590	14.96e- 1
SPECSPTH	-0.369	0.136	6.72e- 3
SNAPSPTH17	0.829	0.153	9.48e- 8
DSPTH	0.442	0.188	1.90e- 2
FFRPTH	0.610	0.120	5.48e- 7
FSRPTH	-0.648	0.148	1.47e- 5
PCT_SNAP17	0.0303	0.0192	1.15e- 1
PCT_CACFP17	0.121	0.0514	1.92e- 2

FOOD_BANKS	**0.0899**	**0.106**	**3.98e- 1**	
PC_DIRSALES17	**0.00103**	**0.00120**	**3.91e- 1**	
FMRKTPTH18	**-1.02**	**0.244**	**3.62e- 5**	
FARM_TO_SCHOOL19	**0.339**	**0.397**	**3.93e- 1**	
RECFACPTH	**-0.585**	**0.0795**	**6.45e-13**	**Adj. R^2 = 0.5094**

Table 9

The regression model with all variables for 2019 has an adjusted R^2 of 0.5094, indicating that the included variables jointly explain over half of the variation in county-level diabetes rates.

In 2019. Dollar store density (DSPTH) is positively associated with diabetes, suggesting that areas with a higher concentration of discount retailers tend to induce higher diabetes rates. Farmers' market density (FMRKTPTH18) and recreation and fitness facility density (RECFACPTH) are both significantly negatively associated with diabetes, showing the importance of healthier food access and physical activity infrastructure. Variables related to food assistance and local food systems are not statistically significant. Overall, the 2019 model suggests a more comprehensive relationship between diabetes prevalence and multiple dimensions of the local food, socioeconomic, and physical environment.

4.1.5 Overview

Between 2015 and 2019, the variables with significant effects were specialized food stores, SNAP-authorized stores, fast-food restaurants, and full-service restaurants. Specialized stores and full-service restaurants have significant negative associations with diabetes, while SNAP stores and fast-food restaurants have significant positive associations.

The model of 2019 shows higher explanatory power than the model of 2015. Several patterns observed in 2015 remain consistent. Specialized food store density (SPECSPTH) and full-service restaurant density (FSRPTH) are significantly negatively associated with diabetes, while fast-food restaurant density (FFRPTH) continues to show a strong positive association.

In the category models, GROCPTH, PCT_CACFP17, and PC_DIRSALES17 all show negative associations with the diabetes rate. However, their signs turn positive in the full model. This shift might reflect overlapping among food access, assistance, and socioeconomic factors. Once multiple dimensions of the food environment are jointly controlled, the estimated marginal effects of individual variables may change direction or have lower significance.

These sign changes don't mean the contradictory relationships, but show that diabetes prevalence is affected by complex and interrelated structural variables. In other words, the results suggest that certain food access indicators might represent a broader effect of socioeconomic status.

The following results show that the significance of variables are consistent in both 2 years.

1. Diabetes-Decreasing Sources (Negative Association):
    - Food Stores - Specialized food stores
    - Restaurants - Full-service restaurants

2. Diabetes-Increasing Sources (Positive Association):
    - Food Stores - SNAP-authorized stores
    - Restaurants - Fast-food restaurants

4.2 Adult Obesity Rate:

For obesity-to-insecurity, local, health, and socioeconomic categories in 2017, no records are available in the dataset. For all other categories, the regression models are as follows.

4.2.1 Overview of Category Models - Adult Obesity Rate in 2017

Obesity to Food Stores

OBESITY17 = $\beta 0$ + $\beta 1$ GROCPTH16 + $\beta 2$ SUPERCPTH16 + $\beta 3$ SPECSPTH16 + $\beta 4$ SNAPSPTH17 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	31.0	0.106	0
GROCPTH16	-2.58	0.241	1.50e-22
SUPERCPTH16	0.380	0.0973	3.34e- 3
SPECSPTH16	-0.638	0.165	1.66e- 4
SNAPSPTH17	1.52	0.120	1.40e-34

Adj. R^2 = 0.1821

Table 10

Obesity to Assistance

OBESITY17 = $\beta 0$ + $\beta 1$ PCT_CACFP17 + $\beta 2$ PCT_SNAP17 + $\beta 3$ FOOD_BANKS + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	26.4	0.215	0
PCT_CACFP17	0.835	0.0514	1.50e-22
PCT_SNAP17	0.431	0.0162	1.67e-140
FOOD_BANKS	-1.54	0.182	3.34e- 3

Adj. R^2=0.2534

Table 11

4.2.2 Overview of Category Models - Adult Obesity Rate in 2022

Obesity to Food Stores

OBESITY22 = $\beta 0$ + $\beta 1$ GROCPTH20 + $\beta 2$ SUPERCPTH20 + $\beta 3$ SPECSPTH20 + $\beta 4$ SNAPSPTH23 + $\beta 5$ DSPTH20 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	34.0	0.199	0
GROCPTH20	-1.67	0.377	1.06e- 5
SUPERCPTH20	0.276	0.139	4.76e- 3
SPECSPTH20	-1.05	0.287	2.76e- 4
SNAPSPTH23	-0.606	0.413	1.43e- 1
DSPTH20	3.68	0.410	3.50e-18

Adj. R^2=0.3149

Table 12

Obesity to Restaurants

OBESITY22 = $\beta 0$ + $\beta 1$ FFRPTH + $\beta 2$ FSRPTH + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	35.1	0.0615	0
FFRPTH	0.264	0.0726	2.87e- 4
FSRPTH	-1.35	0.0819	8.24e-58

Adj. R^2=0.1045

-

Table 13

Obesity to Insecurity

OBESITY22 = $\beta 0$ + $\beta 1$ FOODINSEC_18_20 + $\beta 2$ VLFOODSEC_18_20 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	26.8	0.267	0
FOODINSEC_18_20	0.549	0.0428	9.03e-37
VLFOODSEC_18_20	0.530	0.102	2.10e- 7

Adj. R^2=0.2459

Table 14

Obesity to Taxes

OBESITY22 = $\beta 0$ + $\beta 1$ SODATAX_STORES14 + $\beta 2$ SODATAX_VENDM14 + $\beta 3$ CHIPSTAX_STORES14 + $\beta 4$ CHIPSTAX_VENDM14 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	33.8	0.131	0
SODATAX_STORES14	-0.131	0.0343	1.42e- 4

SODATAX_VENDM14	0.224	0.0516	1.43e- 5
CHIPSTAX_STORES14	0.491	0.0286	3.40e-63
CHIPSTAX_VENDM14	0.0869	0.0308	4.79e- 3

Adj. R^2=0.125

Table 15

Obesity to Socioeconomic

OBESITY22 = $\beta 0$ + $\beta 1$ MEDHHINC21 + $\beta 2$ POVRATE21 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	5.71	2.54	2.44e- 2
MEDHHINC21	2.48	0.230	1.22e-26
POVRATE21	0.164	0.00980	1.70e-60

Adj. R^2=0.1293

Table 16

4.2.3 Models that Adjusted $R^2 < 0.09$

In the category of regression models, some models exhibit very low explanatory power, with adjusted R-squared below 0.09. Which are:

2017

- Obesity to Restaurants, Adj, $R^2 = 0.0796$:

  $OBESITY17 = \beta 0 + \beta 1 * FFRPTH16 + \beta 2 * FSRPTH16 + \varepsilon$
- Obesity to Taxes, Adj, $R^2 = 0.08795$:

  $OBESITY17 = \beta 0 + \beta 1 * SODATAX\_STORES14 + \beta 2 * SODATAX\_VENDM14 + \beta 3 * CHIPSTAX\_STORES14 + \beta 4 * CHIPSTAX\_VENDM14 + \varepsilon$

2022

- Obesity to Assistance, Adj, $R^2 = 0.05871$:

  $OBESITY22 = \beta 0 + \beta 1 * PCT\_CACFP21 + \beta 2 * PCT\_SNAP22 + \beta 3 * FOOD\_BANKS21 + \varepsilon$
- Obesity to Local, Adj, $R^2 = 0.05005$:

  $OBESITY22 = \beta 0 + \beta 1 * PC\_DIRSALES17 + \beta 2 * FMRKTPTH18 + \beta 3 * FARM\_TO\_SCHOOL19 + \varepsilon$
- Obesity to Health, Adj, $R^2 = 0.08005$:

  $OBESITY22 = \beta 0 + \beta 1 * PCT\_HSPA21 + \beta 2 * RECFACPTH20 + \varepsilon$

4.2.4 Obesity to all of the variables

Note that these two models were built using only the variable categories whose individual models' adjusted R-squared are above 0.09.

2017

OBESITY17 = $\beta 0 + \beta 1$ GROCPTH16 + $\beta 2$ SUPERCPTH16 + $\beta 3$ SPECSPTH16 + $\beta 4$ SNAPSPTH17 + $\beta 5$ PCT_CACFP17 + $\beta 6$ FOOD_BANKS18 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	31.3	0.112	0
GROCPTH16	-2.34	0.236	1.50e-22
SUPERCPTH16	0.283	0.0963	3.34e- 3
SPECSPTH16	-0.609	0.0161	1.66e- 4
SNAPSPTH17	1.48	0.118	1.40e-34
PCT_CACFP17	0.708	0.0807	4.69e-18
FOOD_BANKS18	-0.615	0.206	2.82e- 3

Adj. R^2=0.2244

Table 17

In the all variable regression models for 2017, the variable PCT_SNAP17 was removed due to multicollinearity issues. All the included variables were highly statistically significant, with p-value much less than 0.05, and all variables were consistent across categorical variables and all other variables.

The model shows that most food environments along with the existence of food banks are related to a lower obesity rate. In other words, counties with various and accessible food choices face a lower risk of being obese.However, government assistance programs, such as SNAP-authorized stores and the participation of Child & Adult Food Programs, show a significant positive relationship with the obesity rate. This suggests that the food options provided through these government offerings may not be sufficiently nutritious.

2022

OBESITY22 = $\beta 0 + \beta 1$ GROCPTH20 + $\beta 2$ SUPERCPTH20 + $\beta 3$ SPECSPTH20 + $\beta 4$ SNAPSPTH23 + $\beta 5$ DSPTH20 + $\beta 6$ FFRPTH20 + $\beta 7$ FSRPTH20 + $\beta 8$ VLFOODSEC_18_20 + $\beta 9$ SODATAX_STORES14 + $\beta 10$ CHIPSTAX_STORES14 + $\beta 11$ CHIPSTAX_VENDM14 + $\beta 12$ MEDHHINC21 + $\beta 13$ POVRATE21 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	-15.8	4.69	8.24e- 4
GROCPTH20	-0.942	0.325	3.93e- 3
SUPERCPTH20	0.300	0.119	1.20e- 2
SPECSPTH20	-0.767	0.263	3.75e- 3
SNAPSPTH23	-0.0883	0.407	8.28e- 1
DSPTH20	1.84	0.370	8.29e- 7
FFRPTH20	1.27	0.216	7.40e- 9
FSRPTH20	-0.758	0.250	2.51e- 3
VLFOODSEC_18_20	1.46	0.141	3.16e-23

SODATAX_STORES14	0.0569	0.0429	1.85e- 1	
CHIPSTAX_STORES14	0.246	0.0715	6.10e- 4	
CHIPSTAX_VENDM14	-0.0148	0.0403	7.14e- 1	
MEDHHINC21	3.96	0.417	4.86e-20	
POVRATE21	-0.0350	0.0335	2.96e- 1	Adj. R^2=0.5355

Table 18

In the all variable regression models for 2022 (Table 9), the variables FOODINSEC_18_20 and SODATAX_VENDM14 were removed due to multicollinearity issues. Except for the variables SNAPSPTH23, SODATAX_STORES14, CHIPSTAX_VENDM14 and POVRATE21, all the included variables were highly statistically significant, with p-value much less than 0.05.

In the category-specific regression model, CHIPSTAX_VENDM14 (Chips and snacks tax rate for vending machines) and POVRATE21 (poverty rate) showed a positive correlation coefficient. However, in the full combined model, its sign reversed to negative. While SODATAX_STORES14 (Soda tax rate for stores) reversed from a negative association to a positive correlation. These suggest that after controlling for the influence of other variables, the effects' direction of the Chips and snacks tax rate for vending machines, Poverty rate and Soda tax rate for stores on the obesity rate have changed.

Overall, the 2022 model indicates that adult obesity rates are driven by complex conditions and environmental factors.

4.2.5 Overview

Model performance improved, with the Adjusted R-squared rising from 0.2244 to 0.5355. This suggests that the 2022 model explains substantially more of the variance in obesity rates.

Between 2017 and 2022, the variables with significant effects are grocery stores, supercenters & club stores, specialized food stores, dollar stores, child & adult care food program participation rate, number of food banks, fast-food restaurants, full-service restaurants, very low food security, chips and snacks tax rate (stores) and median household income (USD).

1. Obesity-Decreasing Sources (Negative Association):
    - Food Stores - Grocery Stores, Specialized food stores
    - Restaurants - Full-service restaurants
    - Assistance - Number of food banks
2. Obesity-Increasing Sources (Positive Association):
    - Food Stores - Supercenters & club stores, Dollar stores
    - Restaurants - Fast-food restaurants
    - Assistance - Child & Adult Care Food Program participation rate
    - Insecurity - Very low food security
    - Taxes - Chips and snacks tax rate (stores)
    - Socioeconomic - Median household income (USD)

In summary, our research demonstrates that the obesity rate is not driven by a single factor but by a complex interaction of socioeconomic status, the accessibility of government assistance programs, and the structural diversity of local food stores. Specifically, we can conclude that socioeconomic status, government assistance, and specific food store environments are the critical factors influencing higher obesity rates.

In addition, the coefficient sign for SNAP-authorized stores changed from positive in 2017 to negative in 2022. This shift, however, is superseded by the fact that the variable's P-value became statistically insignificant in the 2022 model. This indicates that, in the most recent combined model, the density of SNAP-authorized stores does not have an influential impact on the obesity rate.

4.3 Principal components analysis, PCA

Under PCA, we aim to reweight the importance of each variable to reduce the dimensionality of the data. First, values of −9999 and −8888 were recorded as missing (NA), and all variables were standardized. Subsequently, missing values were dropped.
To ensure an accurate representation of the PCA results, we also removed the variables SUPERCPTH16, SUPERCPTH20, SPECSPTH16, and SPECSPTH20 due to excessive missing values.

4.3.1 PCA of Diabetes-Related Variables

2015

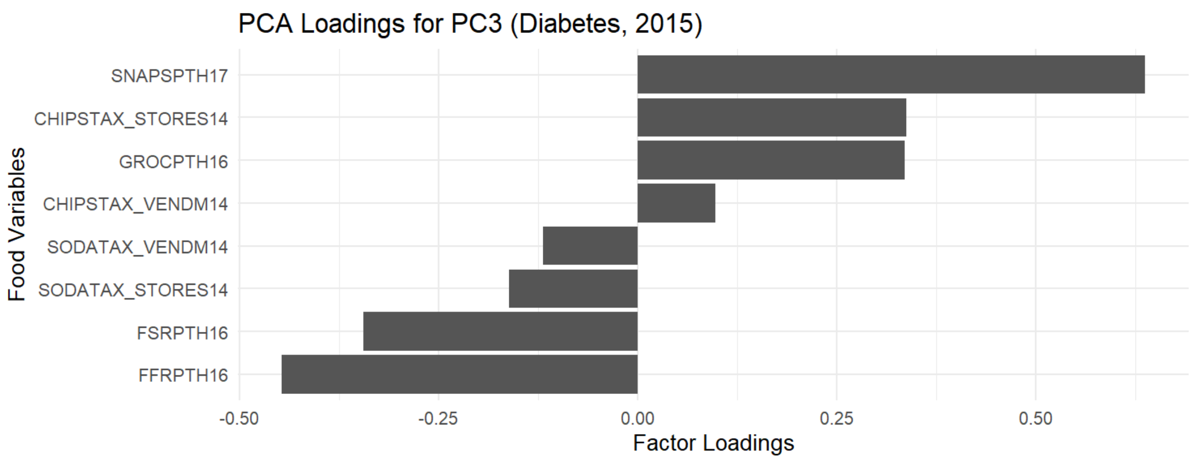
$$DIABETES15 = \beta_0 + \beta_1 PC1 + \beta_2 PC2 + \beta_3 PC3 + \beta_4 PC4 + \varepsilon$$

Variable	Estimate	Std. Error	P-value
(Intercept)	8.34	0.0277	0
PC1	**-0.117**	**0.0177**	**3.76e-11**
PC2	**0.161**	**0.0212**	**4.79e-14**
PC3	**0.313**	**0.0248**	**1.61e-35**
PC4	**-0.221**	**0.0296**	**1.01e-13**

Adj. R^2 = 0.09667

Table 19

In 2015, PC3 had the smallest p-value, with β = 0.313 and p = 3.76e-11. This indicates that it had the strongest association with the diabetes outcome among the following components:



PC3 represents a high-retail, low-dining food environment characterized by a high density of SNAP-authorized stores and grocery stores, and a relatively low number of fast-food restaurants and full-service restaurants. Therefore, a higher PC3 score indicates that these counties rely more on SNAP and retail food access rather than restaurant dining.

This food environment dimension is positively associated with diabetes prevalence. Counties with higher PC3 scores tended to exhibit higher diabetes prevalence, suggesting that even with easy access to groceries, SNAP and retail-oriented food environments are associated with worse diabetes outcomes. This finding is consistent with prior literature documenting adverse health associations in SNAP-related retail environments.

*Although these components do not always correspond to the first principal component, they capture the dimension most strongly associated with diabetes outcomes in the data.

2019

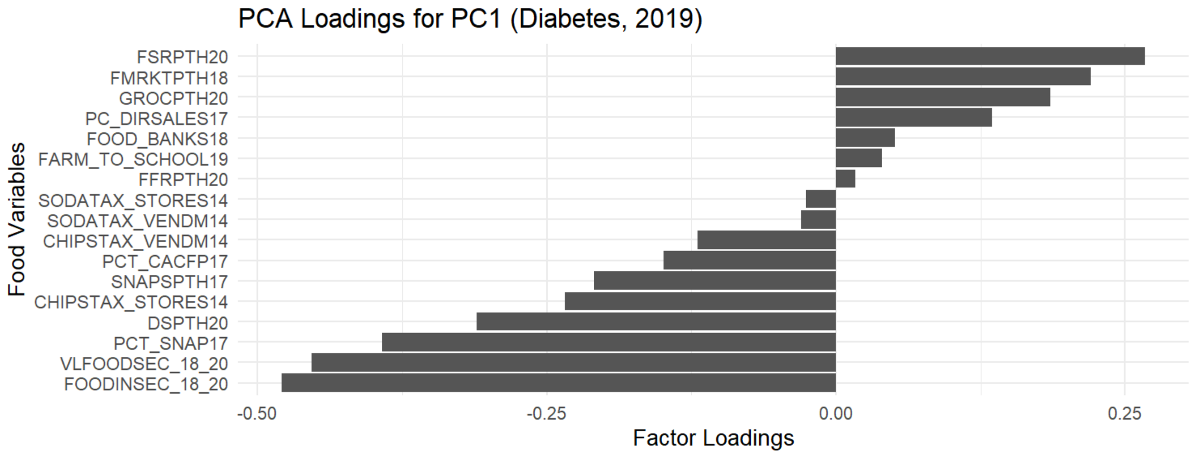
$$\text{DIABETES19} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\,\text{PC1} + \boldsymbol{\beta_2}\,\text{PC2} + \boldsymbol{\beta_3}\,\text{PC3} + \boldsymbol{\beta_4}\,\text{PC4} + \boldsymbol{\beta_5}\,\text{PC5} + \boldsymbol{\beta_6}\,\text{PC6} + \boldsymbol{\beta_7}\,\text{PC7} + \boldsymbol{\beta_8}\,\text{PC8} + \varepsilon$$

Variable	Estimate	Std. Error	P-value
(Intercept)	8.96	0.0326	0
PC1	**-0.461**	**0.0182**	**7.33e-122**
PC2	**0.0152**	**0.0203**	**4.56e- 1**
PC3	**0.0301**	**0.0231**	**1.93e- 1**
PC4	**-0.0502**	**0.0276**	**6.89e- 2**
PC5	**0.0597**	**0.03100**	**5.47e- 21**
PC6	**0.302**	**0.0322**	**1.85e- 20**
PC7	**0.0871**	**0.0328**	**7.95e- 3**
PC8	**0.160**	**0.0342**	**2.92e- 6**

Table 20

Adj. R^2 = 0.2823

In 2019, PC1 had the smallest p-value, with β = -0.461 and p = 7.33e-122. This indicates that it had the strongest association with the diabetes outcome among the following components:



PC1 represents the dimensions of food security and socioeconomic status. It is characterized by a strong negative correlation between food insecurity, very low food security, SNAP participation rates, and dollar store density, while full-service restaurants and farmers' markets load positively on this component. Therefore, a higher PC1 score reflects a worse food insecurity situation in these counties, a lower dependence on food assistance, and a more favorable food environment.

This dimension is negatively associated with diabetes prevalence. Counties with lower PC1 scores, which indicate higher food insecurity and greater socioeconomic disadvantage, tended to exhibit higher diabetes rates. Overall, the results suggest that structural food insecurity and socioeconomic conditions are closely linked to the occurrence of diabetes.

Overview

Comparing the PCA regression results for 2015 and 2019, we come to the conclusion.

1. **Retail and restaurant density:**
   Areas with lower restaurant density reflect a more limited food environment, where residents face fewer dining options and less access to diverse food offerings. This suggests that limited restaurant access is associated with a narrow range of dietary choices and a greater reliance on easily accessible retail food.
2. **Food access structure:**
   Even with grocery stores, regions in which food access is primarily provided through SNAP-authorized stores and general retail outlets are more strongly associated with higher diabetes rates. Research from the CDC study shows that in high obesity rural areas, convenience stores, which make up the largest share of SNAP-authorized outlets, have healthy food availability scores up to 70% lower than supermarkets, while nutritional food prices are higher, indicating limited access to healthy foods in these locations.

In other words, reliance on retail-based food access may limit dietary diversity and is associated with less favorable dietary environments that correlate with higher diabetes prevalence.

3. **Socioeconomic and food security conditions:**
   When socioeconomic and food security indicators are available, higher levels of food insecurity and greater reliance on food assistance are closely associated with higher diabetes rates. Research on SNAP retailer formats in Virginia found that SNAP-authorized convenience, dollar, and non-traditional store formats are associated with higher county-level obesity prevalence, highlighting how specific SNAP store types can be linked to adverse health outcomes.

This suggests that constrained food choices due to socioeconomic disadvantage and the prevalence of unhealthy food retail formats are an important factor contributing to regional differences in diabetes prevalence.

4.3.2 PCA of Obesity-Related Variables

2017

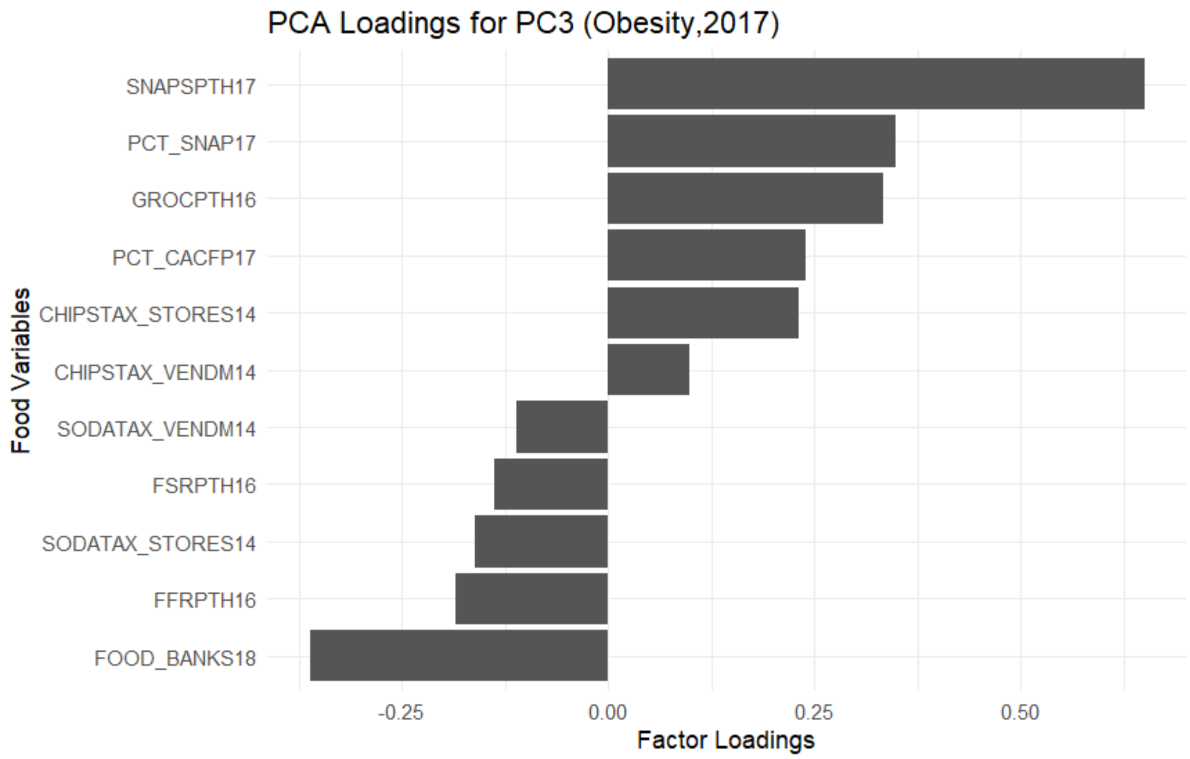
OBESITY17 = $\beta 0 + \beta 1$ PC1 + $\beta 2$ PC2 + $\beta 3$ PC3 + $\beta 4$ PC4 + $\beta 5$ PC5 + $\beta 6$ PC6 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	31.9	0.0506	0
PC1	-0.595	0.0319	1.97e- 73
PC2	0.573	0.0379	1.35e- 49
PC3	1.05	0.0414	8.24e-129
PC4	0.622	0.0473	1.88e- 38
PC5	0.213	0.0499	1.99e- 5
PC6	-0.167	0.0558	2.87e- 3

Adj. R^2 = 0.3266

Table 21

In 2017, PC3 had the smallest p-value, with β = 1.05 and p = 8.24e-129. This indicates that it had the strongest association with the obesity outcome among the following components:



The structure of PC3 is characterized by a high positive loading of SNAP-authorized stores, SNAP participation rates and grocery store density, contrasted by a significant negative loading of fast-food and food banks. Consequently, a higher PC3 score represents counties with extensive SNAP food access infrastructure but a very low density of established dining facilities and food banks.

Based on these results, we conclude that counties with high PC3 scores likely represent low-socioeconomic regions or rural areas. The high reliance on government assistance combined with the lack of restaurant infrastructure and emergency food bank support suggests a structural limitation in food choices.

Crucially, the regression analysis reveals a significant positive relationship between PC3 and obesity rates. This suggests that although these counties have high SNAP-authorized store density, the available food options in these states are likely nutritively poor. These stores typically stock highly processed foods high in carbohydrates, oils, and fats, which are affordable for low-income populations but significantly increase the risk of obesity.

*Although these components do not always correspond to the first principal component, they capture the dimension most strongly associated with diabetes outcomes in the data.

<u>2022</u>

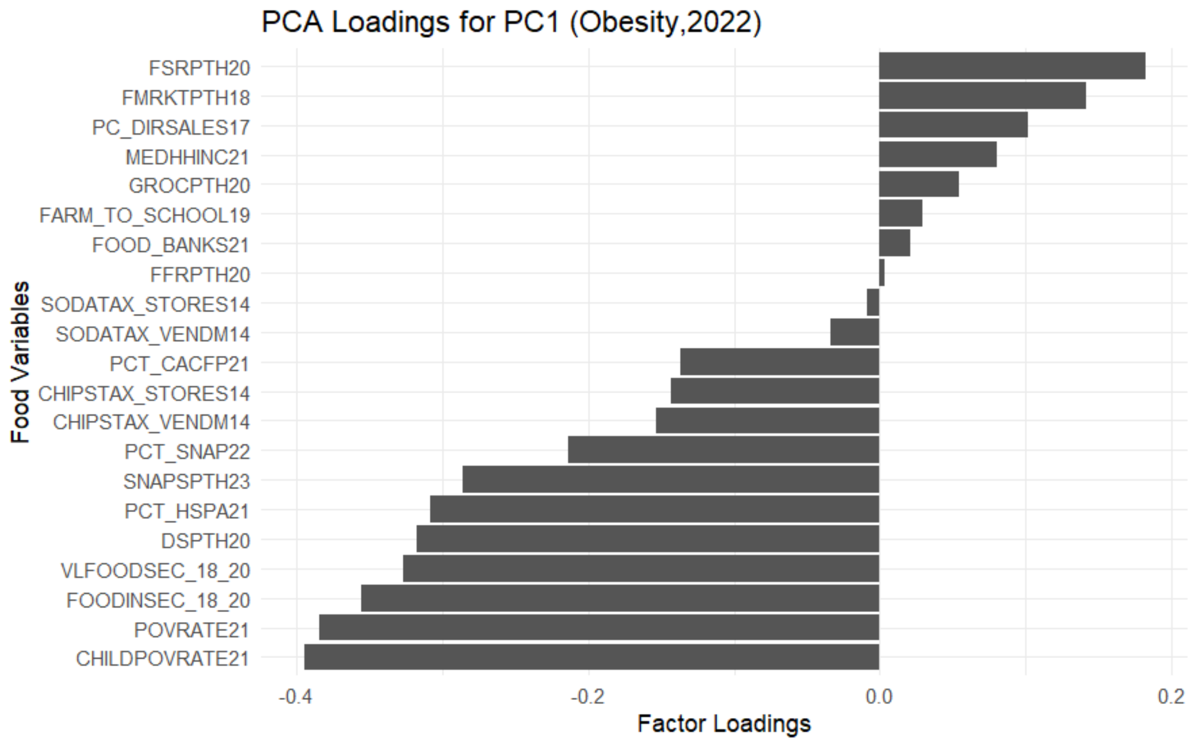
OBESITY22 = $\beta_0 + \beta_1$ PC1 + β_2 PC2 + β_3 PC3 + β_4 PC4 + β_5 PC5 + β_6 PC6 + β_7 PC7 + β_8 PC8 + β_9 PC9 + β_{10} PC10 + ε

Variable	Estimate	Std. Error	P-value
(Intercept)	35.5	0.0616	0
PC1	-0.725	0.0295	1.13e-114
PC2	-0.454	0.0376	2.84e-32
PC3	-0.0684	0.0422	1.05e-1
PC4	0.0993	0.0504	4.90e-2
PC5	0.425	0.0550	1.74e-14
PC6	-0.281	0.0565	7.35e-7
PC7	-0.185	0.0593	1.83e-3
PC8	0.0274	0.0616	6.56e-1
PC9	-0.507	0.0650	1.14e-14
PC10	0.319	0.0155	3.34e-6

Adj. $R^2 = 0.3476$

Table 22

In 2022, PC1 had the smallest p-value, with β = -0.725 and p = 1.13e-114. This indicates that it had the strongest association with the obesity outcome among the following components:



The structure of PC1 is defined by positive loadings for full-service restaurants, farmers' markets and direct farm sales, while having strong negative loadings for poverty rates and food insecurity. Therefore, a higher PC1 score characterizes counties with robust local food economies and high socioeconomic stability.

Based on these results, we conclude that counties with high PC1 scores likely represent higher-income or urbanized areas. These regions offer residents a high opportunity environment to reach a diverse range of food sources—from local markets to a complete dining experience in full-service restaurants.

The regression analysis reveals a significant negative relationship between PC1 and obesity rates in 2022. This indicates extensive food accessibility, coupled with the presence of local farms, serves as a protective factor that decreases the risk of obesity. In contrast, populations in low-income backgrounds often lack consistent access to sufficient nutritious food. This socioeconomic disadvantage exposes them to a higher risk of obesity due to limited dietary choices and the higher affordability of calorie-dense, low-nutrient alternatives.

Overview

Compared to the results in 2017 and 2022, we come to the conclusion.

1. **Restaurants density:**
   A higher density of restaurants indicates a diverse food environment. This suggests restaurant access correlates with a wider variety of dietary choices, offering a healthier alternative to a forced reliance on poor-quality processed foods.
2. **Local production:**
   Counties closer to agriculture's production and farms provide residents with easier access to fresh, nutrient-dense whole foods, which serves as a protective factor against obesity.
3. **Socioeconomic status:**
   Higher standards of living are associated with lower food insecurity, and fewer forced choices in their diet. In other words, individuals with higher socioeconomic status can afford more nutritious food, which leads to a lower risk of obesity.

In summary, these findings indicate that limited food choice, driven by socioeconomic status and structural environment, is the most critical determinant of obesity.

5. Recommendation:

Based on our findings, we suggest that there are several policies and interventions that can decrease the prevalence of diabetes and obesity by improving the food environment.

1. **Expand access to Specialized Food Stores and Full-Service Restaurants**
   Our results consistently show that specialized food stores and full-service restaurants are negatively associated with both diabetes and obesity. Because they offer more diverse, fresher, and less processed food options compared to convenience stores or fast-food chains. Policymakers should consider incentive programs, such as tax

credits, reduced licensing fees, or infrastructure grants, to encourage these businesses to establish operations in underserved areas.

2. **Address the issue with Dollar Stores**
   There are strong positive associations between dollar store density and both diabetes and obesity rate in 2019 and 2022. Although dollar stores provide affordable options for families with financial constraints, they basically sell high-processed foods with limited nutrition. Rather than simply restricting dollar store expansion, policies should focus on improving the nutritional quality of products they offer. This could include requiring minimum standards for fresh food availability or providing subsidies that make it economically viable for dollar stores to stock healthier options.

3. **Reallocate SNAP-Authorized Store network**
   One of the findings of the analysis indicates that convenience stores, which make up a large proportion of SNAP-authorized retailers, provide far fewer healthy food options than supermarkets. The SNAP program should set higher nutrition standards for approved retailers and favor supermarkets and grocery stores over convenience stores when it can. Even in communities where convenience stores are the only practical choice, requirements could be put in place for minimum levels of fresh produce sales.

4. **Strengthen local food systems**
   Farmers' markets and local food systems were positively correlated with health outcomes. To support these systems and increase access to fresh, nutrient dense foods, support through farmers' market incentive programs, farm-to-institution programs, and direct subsidies to local producers is needed.

5. **Socioeconomic factors in food policy**
   Effective interventions require consideration between food access initiatives, economic development programs, and poverty reduction strategies. Minimum wage increases, earned income tax credits, and other policies that improve household purchasing power complement food environment improvements.

In summary, food environment factors are becoming more important determinants of health outcomes over time. This makes addressing food access increasingly urgent as a public health priority. However, our findings also make clear that no single intervention will solve these problems. Comprehensive strategies that simultaneously address food retail diversity, restaurant availability, assistance program quality, and socioeconomic conditions offer the best prospects for meaningful improvements in population health.

6. Limitations and Improvement:

6.1 Methodologies

Due to data limitations, we were unable to implement the Difference-in-Differences method (DID), as the tax change occurred in 2014, prior to the years for which health outcome data are available. Instead, we applied Principal Components Analysis, which we learned in the previous class, to conduct a more structural analysis of the food environment. To further strengthen this study, acquiring data on tax policy changes around 2018 would allow for a more experimental evaluation of treatment effects on adults diabetes and obesity.

While our analyses focus on differences in health outcomes (diabetes and obesity) and time periods, future research can refine this project by exploring heterogeneity among counties with different structural and socioeconomic status. In particular, maybe we can cluster counties based on institutional, demographic, and geographical factors to build a more detailed model and find where the impact of the food environment is most significant.

For example, counties could be divided into areas with varying levels of SNAP participation, poverty rates, or urban/rural areas. These kinds of comparisons between groups would allow us to examine whether the association between food environment variables and health outcomes varies across these factors.

Incorporating these dimensions of exploratory heterogeneity will deepen the interpretation of the results and help identify which populations are most sensitive to changes in food access, assistance programs, and retail structures.

6.2 Results

An unexpected finding in our research was the positive correlation between SNAP-authorized stores and higher rates of adult diabetes and obesity. While our current analysis interprets this through the CDC study, further research is needed. In particular, additional datasets on government funding programs or SNAP participation would enhance both the explanatory power and depth of this research. This would allow for a more understanding of how food assistance infrastructure affects public health, and enable more precise recommendations for the government.

7. Reference:

Dataset from the Government

Food Environment Atlas from Economic Research Service
https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads

Essay

Caspi, C. E., Lenk, K., Pelletier, J. E., Barnes, T. L., Harnack, L., Erickson, D. J., & Laska, M. N. (2021), A market basket assessment: Prices and availability of healthy foods across SNAP-authorized food outlets in counties with high obesity rates in Mississippi, Preventing Chronic Disease, 18, E38.
https://www.cdc.gov/pcd/issues/2021/21_0173.htm

Zhang, M., Wang, H., & Zhang, Q. (2020), Supplemental Nutrition Assistance Program (SNAP)-authorized grocery, convenience, dollar, and restaurant or delivery service settings are associated with increased obesity prevalence in Virginia, Journal of Nutrition Education and Behavior, 52(10), 980–988.
https://pubmed.ncbi.nlm.nih.gov/32567325/

U.S. Department of Agriculture, Economic Research Service. (2018), Eligibility requirements for SNAP retailers: Balancing access, nutrition, and integrity, Amber Waves.
https://www.ers.usda.gov/amber-waves/2018/january-february/eligibility-requirements-for-snap-retailers-balancing-access-nutrition-and-integrity

8. Appendix:

GitHub

https://github.com/mevisliao/Food-Environment-Effects-on-Diabetes-and-Obesity

PCA of Diabetes-Related Variables

2015

GROCPTH16	SNAPSPTH17	FFRPTH16	FSRPTH16
0.3349496	**0.6373219**	**-0.4470979**	**-0.3438438**
SODATAX_STORES14	SODATAX_VENDM14	CHIPSTAX_STORES14	CHIPSTAX_VENDM14
-0.1617035	**-0.1189683**	**0.3371059**	**0.0977883**

2019

GROCPTH20	SNAPSPTH17	DSPTH20	FFRPTH20
0.1855281	**-0.2088617**	**-0.3109274**	**0.0171110**
FSRPTH20	PCT_SNAP17	PCT_CACFP17	FOOD_BANKS18

0.2675849	**-0.3925809**	**-0.1489481**	**0.0511385**
FOODINSEC_18_20	VLFOODSEC_18_20	SODATAX_STORES14	SODATAX_VENDM14
-0.4793698	**-0.4531232**	**-0.0258815**	**0.2206358**
CHIPSTAX_STORES14	CHIPSTAX_VENDM14	PC_DIRSALES17	FMRKTPTH18
-0.2342380	**-0.1197547**	**0.1352928**	**0.0977883**
FARM_TO_SCHOOL19			
0.0396548			

PCA of Obesity-Related Variables

2017

SNAPSPTH17	PCT_SNAP17	GROCPTH16	PCT_CACFP17
0.65091733	0.34896444	0.33399991	0.23866255
CHIPSTAX_STORES14	CHIPSTAX_VENDM14	SODATAX_VENDM14	FSRPTH16
0.23062370	0.09821049	-0.11062173	-0.13902587
SODATAX_STORES14	FFRPTH16	FOOD_BANKS18	
-0.16147602	-0.18530947	-0.36221905	

2022

FSRPTH20	FMRKTPTH18	PC_DIRSALES17	MEDHHINC21
0.18294142	0.14223493	0.10240017	0.08061979
GROCPTH20	FARM_TO_SCHOOL19	FOOD_BANKS21	FFRPTH20
0.05457218	0.02966315	0.02148585	0.00377769
SODATAX_STORES14	SODATAX_VENDM14	PCT_CACFP21	CHIPSTAX_STORES14
-0.00888665	-0.03349589	-0.13647495	-0.14285089
CHIPSTAX_VENDM14	PCT_SNAP22	SNAPSPTH23	PCT_HSPA21
-0.15366589	-0.21346620	-0.28644233	-0.30816227
DSPTH20	VLFOODSEC_18_20	FOODINSEC_18_20	POVRATE21
-0.31728444	-0.32728866	-0.35593093	-0.38454367
CHILDPOVRATE21			
-0.39523092			