ECON 3115: Final Project
Mónica Martínez-Raga, Isabela Vieira

Food Inequality & Dietary Patterns: Effects of Supermarket Shortage in New York City's Boroughs

Abstract

According to the Center for Disease Control and Prevention (CDC), consuming the equivalent of two and a half cups of both fruits and vegetables daily is imperative for maintaining a healthful diet[1]. However, eating fresh fruits and vegetables on a daily basis is perceived as a privilege to many in the United States. Along with affordability, accessibility is viewed as a key determinant of American's eating habits[2]. The Supermarket Need Index devised by the New York City Department of City Planning reports a supermarket shortage in low-income areas within New York City – neighborhoods in Upper Manhattan, East Brooklyn, and the Bronx struggle the most with supermarket availability and accessibility[3]. The inability to easily access supermarkets may impact the dietary choices of individuals, affecting the overall quality of their health. For example, given a long commute to grocery shopping, individuals may opt for foods with longer shelf-life, and limit their consumption of fresh produce. While using data from the Behavioral Risk Factor Surveillance System (BRFSS) to access the eating habits of individuals within New York State, this paper attempts to access the relationship between the likeability of eating fruits and vegetables on a daily basis and an individual's location within one of the five New York City boroughs.

I. Introduction

As residents of Hamilton Heights in Upper Manhattan, we have struggled to find fresh and affordable fruits and vegetables in our neighborhood. In our personal observations, we noted that supermarkets in our region were filled with inexpensive processed foods, while a 16-ounce box of fresh strawberries would be priced at $6.99 without taxes included. Meanwhile, chain supermarkets such as Trader Joe's in 96th Street offer the same brand and portion of strawberries at $4.99[4]. As of now, there are no Trader Joe's stores north of West 96th street in contrast with a total of eight store locations in Downtown Manhattan. Since 2009, the New York City's Food

[1] U.S. Department of Health and Human Services and U.S. Department of Agriculture, "2015–2020 Dietary Guidelines for Americans," 8th Edition. December 2015.
[2] Hawkes, Corinna. "Dietary Implications of Supermarket Development: A Global Perspective." Development Policy Review 26, no. 6 (November 2008): 657–92.
[3] New York City Department of City Planning, New York City Health, New York City Economic Development Corporation, "Going to Market: New York City's Neighborhood Grocery Store and Supermarket Shortage," October 2008.
[4] "Best Price on Organic Strawberries at Trader Joe's." All-Natural Savings, February 18, 2017.

Retail Expansion to Support Health (FRESH) has offered tax incentives for grocery retailers based on zoning[5], but the issue of shortage persists in many neighborhoods, especially in Upper Manhattan and the Bronx.

Given these qualitative observations, we came to the assumption that the shortage of supermarkets in low-income areas, and thus a lack of competition altogether, was a key factor responsible for driving prices of fresh produce upward. Instinctively, we would assume that residents of these areas would rather do their grocery shopping locally, instead of employing time and capital into travelling to stores downtown, even if these stores have a greater availability of fresh produce at lower prices. Hypothetically, all factors mentioned above would influence the consumer choices of people living in low-income areas affected by a supermarket shortage. Assuming price and preparation time as key determinants of consumer choices in grocery shopping, we would expect people in low-income areas to eat less fruits and vegetables relative to residents of affluent areas.

The vast literature published by econometricians discussing the issue of food deserts in the United States has helped us to conceptualize our understanding of the topic. Recently, many works have expressed criticism against the concept of food deserts, some arguing that "this metaphor [food desert], which implies an absence of food, is misleading and potentially detrimental to the health of poor and racially diverse communities because it ignores the contribution of smaller stores, particularly that of so-called ethnic markets"[6]. Others criticize the one-sidedness of the literature produced, arguing that the concept of food deserts leads to inefficient supply-side policymaking[7]. Furthermore, many data-oriented works have failed to find statistically significant relationships between food deserts and people's health outcomes[8], some going as far as to reject any relationship between neighborhoods and nutritional inequality altogether[9]. The article "Food Deserts and Nutritional Inequality" analyzes the effect of supermarket entries in neighborhoods defined as food deserts, reaching the conclusion that a new supermarket does not significantly affect the nutrition of low-income individuals[10].

Somewhat discouraging, the literature reviewed for this paper led us to question our purpose in conducting this study. Nevertheless, we decided to move forward with our project, mainly because there is an absence of scholarly articles discussing the relationship between food

---

[5] "Food Retail Expansion to Support Health (FRESH)." NYCEDC

[6] Joassart-Marcelli, Pascale, Jaime S. Rossiter, and Fernando J. Bosco. 2017. "Ethnic Markets and Community Food Security in an Urban 'Food Desert.'" Environment and Planning A 49 (7): 1642–63.

[7] Wolf-Powers, Laura. "Food Deserts and Real-Estate-Led Social Policy." International Journal of Urban and Regional Research 41, no. 3 (May 2017): 414–25.

[8] Fitzpatrick, Katie, Nadia Greenhalgh-Stanley, and Michele Ver Ploeg. 2019. "Food Deserts and Diet-Related Health Outcomes of the Elderly." Food Policy 87 (August).

[9-10] Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dube, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell. 2019. "Food Deserts and the Causes of Nutritional Inequality." Quarterly Journal of Economics 134 (4): 1793–1844.

deserts and healthful eating in New York City specifically. The only article focused on food deserts in New York City—*Disparities in Neighborhood Food Environments: Implications of Measurement Strategies*—discusses parallel measures, such as vehicle ownership and crime, to address the supermarket shortage in diverse communities, without addressing the impact of supermarket availability in the overall nutrition of the population[11]. In view of that, we considered our discussion as relevant to the topic of food deserts and proceeded with our project.

Given the many assumptions we have made initially, we established a framework in which we could test these theories using publicly available data. Instead of analyzing the relationship between supermarket supply, prices, and consumer choices, we decided to focus on determining whether people's diets are, in fact, impacted by their location in New York City. In short, this paper attempts to identify whether supermarket shortages in low-income areas are correlated with fresh produce consumption through statistical software (R). The following is the theoretical framework we used to access the meaning of our empirical results:

I. Null Hypothesis: An individual's location within one of the five NYC's boroughs has no meaningful effect on the overall consumption pattern of fruits and vegetables.
II. Alternate Hypothesis 1.1: Living in a relatively more affluent borough (e.g., Manhattan) increases the likeability of a person consuming fruits and vegetables daily.
III. Alternate Hypothesis 1.2: Living in a relatively more affluent borough (e.g., Manhattan) decreases the likeability of a person consuming fruits and vegetables daily.

II. Methodology

Preliminary Data Observations

In the process of selecting the optimal dataset for addressing questions related to the issue of food inequality in the U.S., it became clear that a proper measure of healthy eating was needed. According to the Dietary Guidelines for Americans (2015–2020) published by the Centers for Disease Control and Prevention (CDC), individuals must consume 2 ½ cups of vegetables from all five subgroups -- dark green, red and orange, legumes, starchy and other -- and 2 ½ cups of whole fruits on a daily basis, preferably fresh, in order to keep a balanced diet[12].

[11] Bader, Michael D. M., Marnie Purciel, Paulette Yousefzadeh, and Kathryn M. Neckerman. 2010. "Disparities in Neighborhood Food Environments: Implications of Measurement Strategies." Economic Geography 86 (4): 409–30.
[12]U.S. Department of Health and Human Services and U.S. Department of Agriculture, "2015–2020 Dietary Guidelines for Americans," 8th Edition. December 2015.

We aimed to analyze the composition of individual diets, comparing the intake of fresh produce and processed foods since food deserts are characterized by the lack of fresh produce. The only publicly available dataset that contained detailed descriptions of all foods consumed was the CDC's National Health and Nutrition Examination Survey (NHANES). However, the survey's design consists of variables which represent food combinations as well as individual foods consumed for a total of two separate days per individual. This posed difficulties in the process of classifying all foods into either fresh or processed. For example, a single variable could include a meal of red meat, dark green vegetables and dessert. In addition, there was no way to identify whether any given person's fruit and veggie intake was isolated or part of a balanced or unbalanced diet. CDC emphasizes that what matters for a balanced diet is the pattern of consumption rather than intakes of individual nutrients[13]. Therefore, fruit and veggie intake data for two out of the 365 days a year is not representative of any dietary pattern.

Consequently, we opted for the 2019 Behavioral Risk Factor Surveillance System (BRFSS) instead. The BRFSS is administered by the CDC's Population Health Surveillance Branch and provides users with relevant demographics for each individual respondent with a focus on health, exercising and eating habits[14]. It includes streamlined variables representing the frequency in which respondents consume any fruits and veggies in a year. For our models, we used FVGREEN1, surveyed as "How many times did you eat dark green vegetables?" and FRUIT2, surveyed as "How many times did you eat fruit?" We are defining a healthy dietary pattern as daily consumption of fruits and veggies. We restricted each variable as such. In the Survey, 19.7% out of 385,188 respondents eat veggies on a daily basis, and 49.9% out of 387,216 respondents ate fruits on a daily basis.

Although the BRFSS is the most recent and comprehensive public dataset containing individual-level data on the consumption of fruits and vegetables, the dataset does not provide specific geographic locations for each individual. The highest level of geographic detail provided by this dataset identifies people living in metropolitan areas within New York State. With the BRFSS alone, we are not able to directly regress a healthy dietary pattern by New York City, much less by borough or neighborhood. Therefore, identifying dietary patterns of residents living within areas that struggle with a supermarket shortage as reported by the Supermarket Need Index by the New York City Department of City Planning would be unfeasible.

Acknowledging this constraint, we devised a machine learning model that enabled us to make an educated guess of the location of a BRFSS respondent within one of the five boroughs of NYC. The model was built to classify people in each borough by using demographic variables

[13] U.S. Department of Health and Human Services and U.S. Department of Agriculture, "2015–2020 Dietary Guidelines for Americans," 8th Edition. December 2015.

[14] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2019.

that are available both in the BRFSS and in a more geographically detailed dataset, the 2017 U.S. Census and American Community Survey (ACS - IPUMS Integrated Public Use Microdata Series)[15]. First, we chose a total of 10 independent variables that would best classify people from New York State in New York City, and people from New York City into each of the five boroughs. The independent variables include household income, age, race, gender, ethnicity, educational attainment, marital status, employment status, home ownership status, and health plan coverage.

In reference to Table 1 and 2, using IPUMS data we can see how household income is distributed among boroughs. This distribution is relevant to our hypotheses since we are trying to find a relationship for "affluent neighborhoods." The Bronx has the lowest average income with a median income of $48,300 yearly, while Manhattan has the highest average income with a median income of $100,000 yearly. According to our assumptions regarding the relationship of income and healthful eating, we would assume that less people in The Bronx eat fruits and veggies on a daily basis than in Manhattan, for example. It is worth noting that boroughs in themselves may be very demographically segregated, and consequently so can food availability. In Brooklyn for example, Trader Joe's and Whole Foods are more prominent in the West, where there are higher concentrations of wealth, than in the East. The same would go for Manhattan South and North of 125th, respectively. This segregation will affect our model, because different neighborhoods may be underrepresented by the boroughs' overall demographic metrics.

Table 1 - Trends in Household Income per NYC borough in Census Data

	Bronx (N=10577)	Manhattan (N=9792)	Staten Island (N=4097)	Brooklyn (N=24764)	Queens (N=21841)	Overall (N=196585)
Household Income						
10000	855 (8.1%)	478 (4.9%)	124 (3.0%)	1097 (4.4%)	561 (2.6%)	6274 (3.2%)
15000	576 (5.4%)	333 (3.4%)	95 (2.3%)	865 (3.5%)	442 (2.0%)	5191 (2.6%)
20000	652 (6.2%)	293 (3.0%)	87 (2.1%)	757 (3.1%)	573 (2.6%)	5678 (2.9%)
25000	583 (5.5%)	283 (2.9%)	82 (2.0%)	1010 (4.1%)	678 (3.1%)	6354 (3.2%)
35000	983 (9.3%)	531 (5.4%)	190 (4.6%)	1753 (7.1%)	1372 (6.3%)	12469 (6.3%)
50000	1207 (11.4%)	608 (6.2%)	282 (6.9%)	2464 (9.9%)	2250 (10.3%)	18641 (9.5%)
75000	1753 (16.6%)	942 (9.6%)	540 (13.2%)	3268 (13.2%)	3369 (15.4%)	29016 (14.8%)
2030000	3164 (29.9%)	5299 (54.1%)	2518 (61.5%)	12554 (50.7%)	11867 (54.3%)	100842 (51.3%)
Missing	804 (7.6%)	1025 (10.5%)	179 (4.4%)	996 (4.0%)	729 (3.3%)	12120 (6.2%)

Table 2 - Descriptive Statistics of Household Income per NYC Borough in Census Data

	Bronx (N=10577)	Manhattan (N=9792)	Staten Island (N=4097)	Brooklyn (N=24764)	Queens (N=21841)
Household Income					
Mean (SD)	69300 (79200)	179000 (229000)	123000 (112000)	112000 (127000)	106000 (97800)
Median [Min, Max]	49300 [0, 1550000]	100000 [-5900, 2030000]	100000 [-2800, 1380000]	79000 [-5900, 1550000]	84000 [-5900, 1280000]
Missing	667 (6.3%)	913 (9.3%)	127 (3.1%)	688 (2.8%)	582 (2.7%)

[15] Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2017.

Simple Regressions

Before running our classification models, we decided to test our assumption that income affects the ability of a person to maintain a healthful diet using the BRFSS data exclusively. We ran two Ordinary Least Squares (OLS) models to study the relationship between income and fruit/vegetable consumption. The OLS does not run with factors, therefore instead of using FVGREEN1 and FRUIT2, we used FRUTSU1 and VEGESU1 which are continuous numeric variables. FRUTSU1 and VEGESU1 return values between 0 - 17,600 and 0 -13,204, respectively, for total fruits/vegetables consumed in one day by each individual. BRFSS does not disclose the unit used to measure 1 fruit or vegetable. Below are the linear regressions using household income as a dependent variable and both FRUTSU1 (OLS Output 1) and VEGESU1 (OLS Output 2) as independent variables:

OLS Output 1

	Dependent variable:
	X_FRUTSU1
INCOME215000	-12.479
	(41.431)
INCOME220000	-19.539
	(40.829)
INCOME225000	6.158
	(40.611)
INCOME235000	40.717
	(41.595)
INCOME250000	-44.210
	(41.119)
INCOME275000	-15.287
	(36.196)
Constant	206.628***
	(32.461)
Observations	2,090
R^2	0.003
Adjusted R^2	0.0002
Residual Std. Error	405.432 (df = 2083)
F Statistic	1.069 (df = 6; 2083)
Note:	*p<0.1; **p<0.05; ***p<0.01

OLS Output 2

	Dependent variable:
	X_VEGESU1
INCOME215000	43.719
	(39.350)
INCOME220000	79.695**
	(38.945)
INCOME225000	55.514
	(38.784)
INCOME235000	52.854
	(39.473)
INCOME250000	107.993***
	(39.171)
INCOME275000	59.315*
	(34.517)
Constant	173.075***
	(31.101)
Observations	2,027
R^2	0.004
Adjusted R^2	0.002
Residual Std. Error	375.791 (df = 2020)
F Statistic	1.511 (df = 6; 2020)
Note:	*p<0.1; **p<0.05; ***p<0.01

As shown above, the relationship between the BRFSS household income variable and the consumption of fruits isn't statistically significant. With this data and model alone, we would

accept the null hypothesis. However, we see that some income levels are correlated with the consumption of vegetables with p-values that are low enough for us to potentially reject the null. A possible explanation is that fruits are easier to consume, since vegetables are less likely to be eaten raw if they need preparation and thus time. Following this theory, the consumption of vegetables would be correlated to income if individuals with lower incomes have less time to prepare these foods. Another possible explanation could be location and accessibility to affordable supermarkets with fresh produce. However, we test that assumption in the later stages of our analysis.

The BRFSS data for these variables is not as well surveyed as we'd like it to be. For instance, the description for the VEGSU1 variable in the BRFSS codebook reads as "Total vegetables consumed per day." A quick summary of the variable returns a maximum value reported of 13,204, which sounds humanly impossible or unlikely for any given respondent to record accurately. As mentioned before, we are limited because the unit of 1 fruit or vegetable is not disclosed. This raises skepticism about the accuracy of the data we are working with. Another issue is that BRFSS does not disclose whether the fruits and vegetables are consumed fresh. For example, people who report eating fruit may be counting fruit in syrup, canned fruit, dried fruit, fruit in dessert, and other variations. We cannot properly determine whether the type of consumption of fruit and vegetables is adequate in what the CDC describes as a "healthy eating pattern." The excessive generalization in these variables may be deterring statistical significance with household income in our model.

Predictive Models

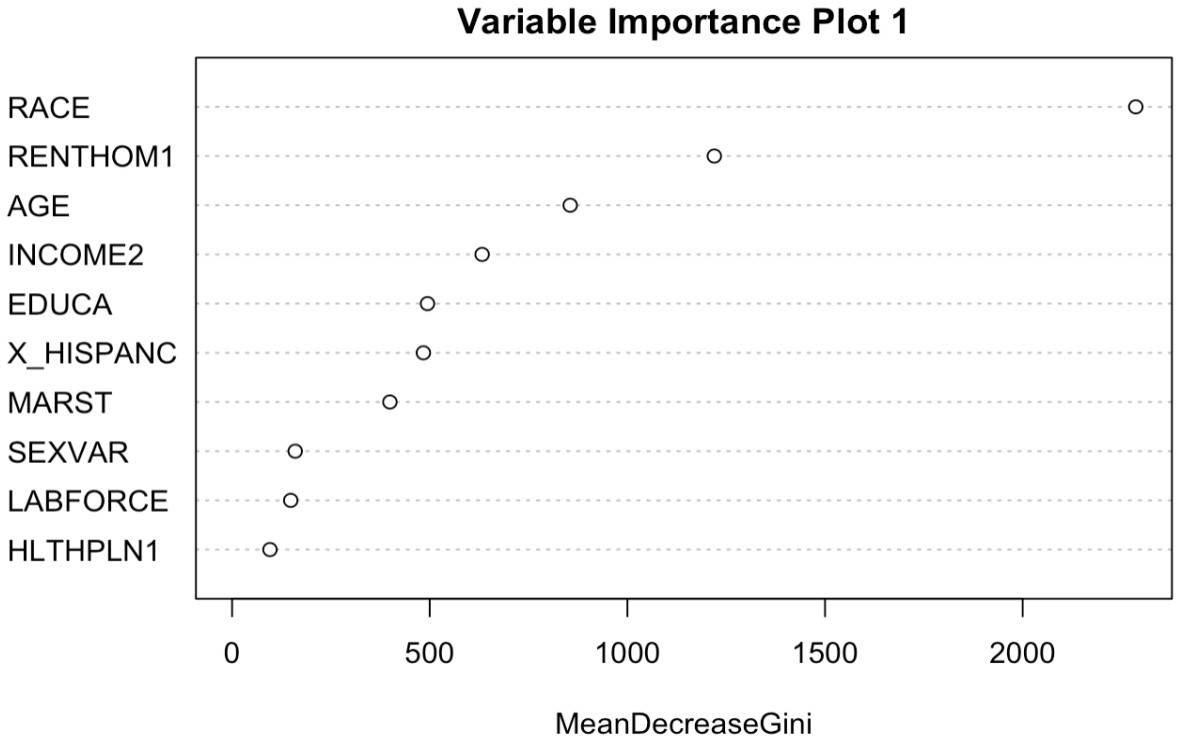
Model 1: Using Random Forest, we built a prediction model where the IPUMS demographic variables were independent, and our dependent variable was being in New York City (in_NYC). A key step in preparing our data for the classification of people in NYC was to convert numeric variables into factors, which allowed us to categorize our data identically in both datasets, this allows the model to use BRFSS variables to perform the same prediction. Random Forest produces 500 decision trees or iterations by default, and we kept it as such. At each "branch" of the decision trees, the model inputs a dummy variable, allowing it to perform a "Yes" or "No" classification, or to pick levels of categorical variables to define someone living in NYC. The model was fed 20% of randomly selected inputs from IPUMS as training data. We used this portion of data to train the branches to make classifications based on the pattern of demographic variables for most people in NYC.

The model classified individuals living in New York City with an accuracy level of 77.47% (OOB 22.53%), represented by the grey values in Confusion Matrix 1.

Confusion Matrix 1

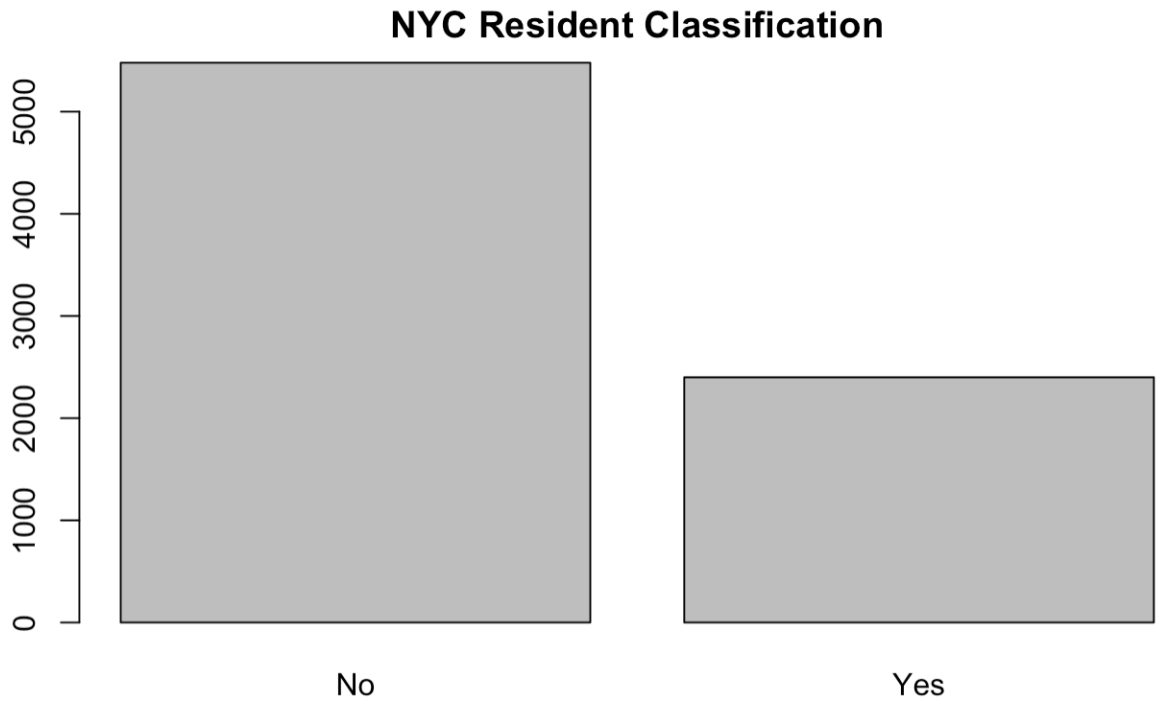
	Yes	No	% Class. Error
Yes	7,193	2,845	15.4%
No	3,786	15,614	34.4%

According to Variable Importance Plot 1, the most statistically significant variables to classify people living in NYC are race, home ownership type, age, and income respectively.



Model 1 - Prediction Output: When applied to a subset of the BRFSS data composed of NYS residents within metropolitan areas, the model classifies a total of 2,400 individuals as living in New York City, representing 30.46% of our total NYS metropolitan residents. It is worth mentioning that our model was trained using 2017 Census data, which may have affected our predictive figures given that BRFSS data is from 2019. The NA's are a result of our model not being able to classify into NYC inputs that had NA's in any of the 10 demographics variables. We tried to impute data into the NA's using the Mice package in R. We discarded this method because it only created 9 imputations. Another way to "fix" our NA's would have been to impute

the mode for each variable. However, we also discarded it because this would only introduce bias into our model. We simply kept the NA's at the cost of having a smaller sample size. Forcing the model to classify an individual without all the variables available would only increase our error. Below is a graphic illustration of NYS residents classified into NYC in the BRFSS dataset.



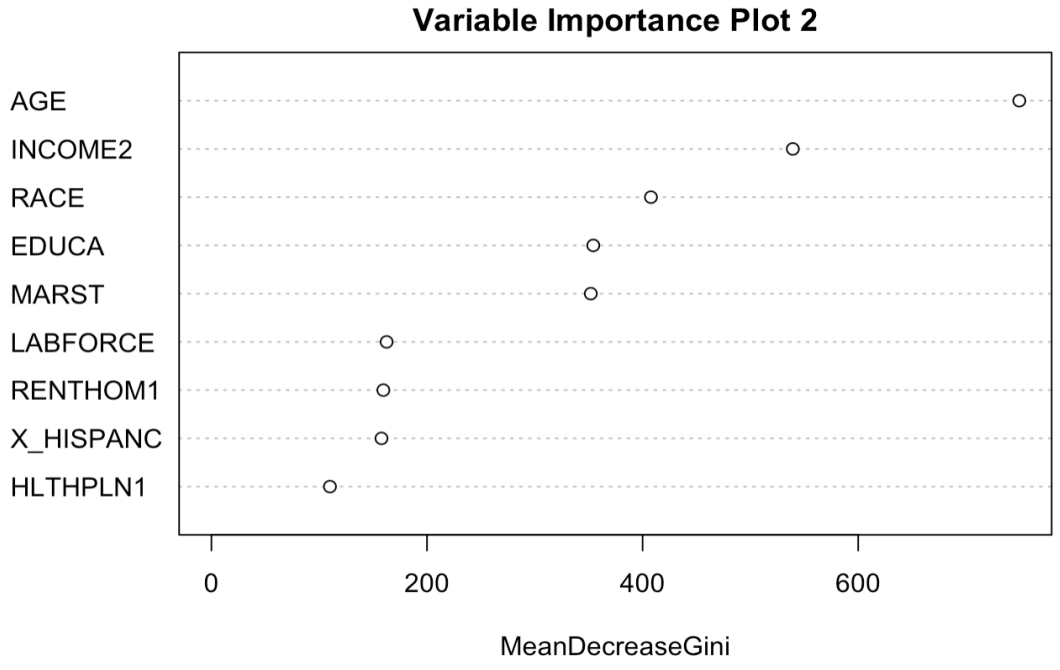
Model 2: Using the same demographic variables, we produced a second prediction model to classify into boroughs BRFSS respondents who were assorted into New York City in the prediction output of Model 1. For this model, we created a borough factorial variable with 5 levels (borough) from the IPUMS dataset. Each "branch" of the random forest is categorizing inputs into either of the 5 boroughs (instead of "Yes" or "No" like for in_NYC) given the 10 independent variables. The model was fed 75% of randomly selected training data from IPUMS, which we increased compared to our NYC classification training data because our sample size is much smaller.

Model 2 predicted the borough in which a person *would* live with an accuracy of 43.46% (OOB 56.54%). This is quite unsatisfactory. Far from being ideal, the out-of-bag error is unavoidably high due to similarities between people living across boroughs, and the differences amongst people living within different neighborhoods of the same borough. For instance, individuals living in Upper Manhattan may have more in common with individuals living in the Bronx than with individuals living in Chelsea. Therefore, our model misclassifies people into the boroughs with highest diversity, as is the case with the unusually large amount of people classified into Brooklyn. See Confusion Matrix 2 for misclassification details.

Confusion Matrix 2

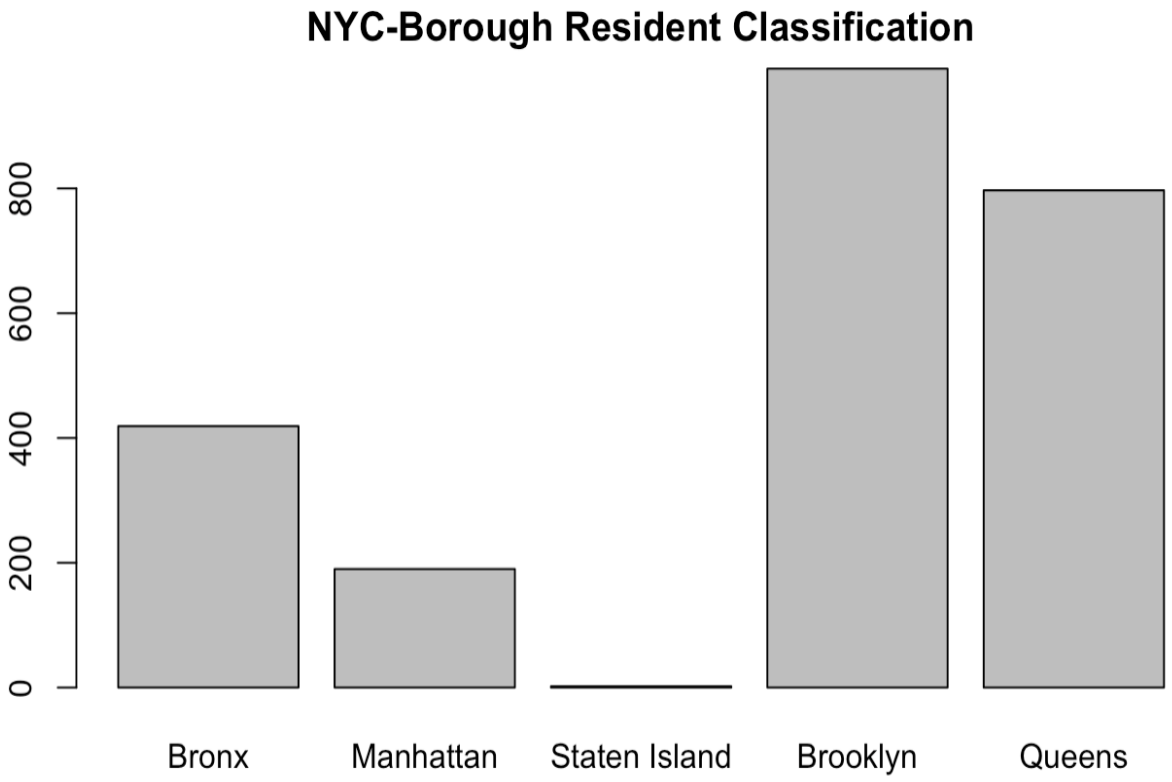
	Bronx	Manhattan	Staten Island	Brooklyn	Queens	%Class. Error
Bronx	439	60	8	482	461	69.7%
Manhattan	155	285	2	763	310	81.1%
Staten Island	22	32	25	241	277	95.81%
Brooklyn	256	186	24	2032	1167	44.5%
Queens	223	186	29	1120	1898	45.0%

According to Variable Importance Plot 2, the most statistically significant variables to classify people living in NYC into each single borough are age, household income, race, and educational attainment respectively.



Model 2 - Prediction Output: Ideally, with a larger subset of BRFSS respondents living in NYS metropolitan areas we would seek a higher level of detail by defining neighborhoods within the boroughs. Not only would independent variables better correlate with demographic differences neighborhood by neighborhood, but we would also have a better count of produce supply and supermarkets. In our case, seeking more detailed classification for the individuals in this research would constrain the number of observations within our working dataset, which would negatively impact our regression analysis down the line.

The below graph illustrates a distribution of BRFSS people potentially living in the different NYC boroughs according to the predictions generated by our model. No respondents were classified in Staten Island, arguably because it is less diverse in several demographic aspects and it also has the smallest population out of the 5 boroughs.



Refer to the table below for a detailed demographic description of our BRFSS sample of NYC population by predicted borough location.

	Bronx (N=369)	Manhattan (N=45)	Brooklyn (N=530)	Queens (N=200)	Overall (N=1144)
Household Income					
0	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10000	46 (12.5%)	3 (6.7%)	31 (5.8%)	6 (3.0%)	86 (7.5%)
15000	87 (23.6%)	2 (4.4%)	36 (6.8%)	14 (7.0%)	139 (12.2%)
20000	92 (24.9%)	2 (4.4%)	46 (8.7%)	16 (8.0%)	156 (13.6%)
25000	57 (15.4%)	4 (8.9%)	57 (10.8%)	12 (6.0%)	130 (11.4%)
35000	30 (8.1%)	6 (13.3%)	66 (12.5%)	30 (15.0%)	132 (11.5%)
50000	19 (5.1%)	1 (2.2%)	82 (15.5%)	21 (10.5%)	123 (10.8%)
75000	38 (10.3%)	27 (60.0%)	212 (40.0%)	101 (50.5%)	378 (33.0%)
2030000	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Home Ownership Type					
Owned	23 (6.2%)	10 (22.2%)	114 (21.5%)	173 (86.5%)	320 (28.0%)
Rent	346 (93.8%)	35 (77.8%)	416 (78.5%)	27 (13.5%)	824 (72.0%)
Race					
White	0 (0%)	21 (46.7%)	142 (26.8%)	2 (1.0%)	165 (14.4%)
Black	16 (4.3%)	1 (2.2%)	318 (60.0%)	66 (33.0%)	401 (35.1%)
American Indian/Alaskan Native	0 (0%)	0 (0%)	26 (4.9%)	1 (0.5%)	27 (2.4%)
Asian	353 (95.7%)	23 (51.1%)	44 (8.3%)	131 (65.5%)	551 (48.2%)
Other Race	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Multiracial	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Hispanic (Yes/No)					
Yes	353 (95.7%)	21 (46.7%)	41 (7.7%)	108 (54.0%)	523 (45.7%)
No	16 (4.3%)	24 (53.3%)	489 (92.3%)	92 (46.0%)	621 (54.3%)
Educational Attainment					
Elementary	78 (21.1%)	2 (4.4%)	20 (3.8%)	6 (3.0%)	106 (9.3%)
SomeHS	39 (10.6%)	0 (0%)	21 (4.0%)	6 (3.0%)	66 (5.8%)
HS	110 (29.8%)	1 (2.2%)	74 (14.0%)	39 (19.5%)	224 (19.6%)
SomeC	70 (19.0%)	0 (0%)	85 (16.0%)	54 (27.0%)	209 (18.3%)
College	72 (19.5%)	42 (93.3%)	330 (62.3%)	95 (47.5%)	539 (47.1%)
Employment Status					
in.LF	259 (70.2%)	35 (77.8%)	367 (69.2%)	118 (59.0%)	779 (68.1%)
Not.in.LF	110 (29.8%)	10 (22.2%)	163 (30.8%)	82 (41.0%)	365 (31.9%)
Healthplan Coverage					
Yes	290 (78.6%)	40 (88.9%)	469 (88.5%)	184 (92.0%)	983 (85.9%)
No	79 (21.4%)	5 (11.1%)	61 (11.5%)	16 (8.0%)	161 (14.1%)
Age Group					
18	7 (1.9%)	0 (0%)	2 (0.4%)	1 (0.5%)	10 (0.9%)
25	20 (5.4%)	2 (4.4%)	12 (2.3%)	7 (3.5%)	41 (3.6%)
30	31 (8.4%)	2 (4.4%)	53 (10.0%)	6 (3.0%)	92 (8.0%)
35	49 (13.3%)	0 (0%)	70 (13.2%)	14 (7.0%)	133 (11.6%)
40	47 (12.7%)	10 (22.2%)	39 (7.4%)	16 (8.0%)	112 (9.8%)
45	46 (12.5%)	0 (0%)	68 (12.8%)	29 (14.5%)	143 (12.5%)
50	39 (10.6%)	6 (13.3%)	59 (11.1%)	34 (17.0%)	138 (12.1%)
55	37 (10.0%)	9 (20.0%)	53 (10.0%)	11 (5.5%)	110 (9.6%)
60	41 (11.1%)	0 (0%)	52 (9.8%)	29 (14.5%)	122 (10.7%)
65	19 (5.1%)	4 (8.9%)	54 (10.2%)	15 (7.5%)	92 (8.0%)
70	18 (4.9%)	5 (11.1%)	34 (6.4%)	26 (13.0%)	83 (7.3%)
75	9 (2.4%)	5 (11.1%)	18 (3.4%)	3 (1.5%)	35 (3.1%)
80	6 (1.6%)	2 (4.4%)	16 (3.0%)	9 (4.5%)	33 (2.9%)
Marital Status					
married	219 (59.3%)	18 (40.0%)	318 (60.0%)	147 (73.5%)	702 (61.4%)
divorced	91 (24.7%)	25 (55.6%)	161 (30.4%)	45 (22.5%)	322 (28.1%)
widowed	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
separated	59 (16.0%)	2 (4.4%)	51 (9.6%)	8 (4.0%)	120 (10.5%)
never.married	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

III. Results

Logistic Regressions

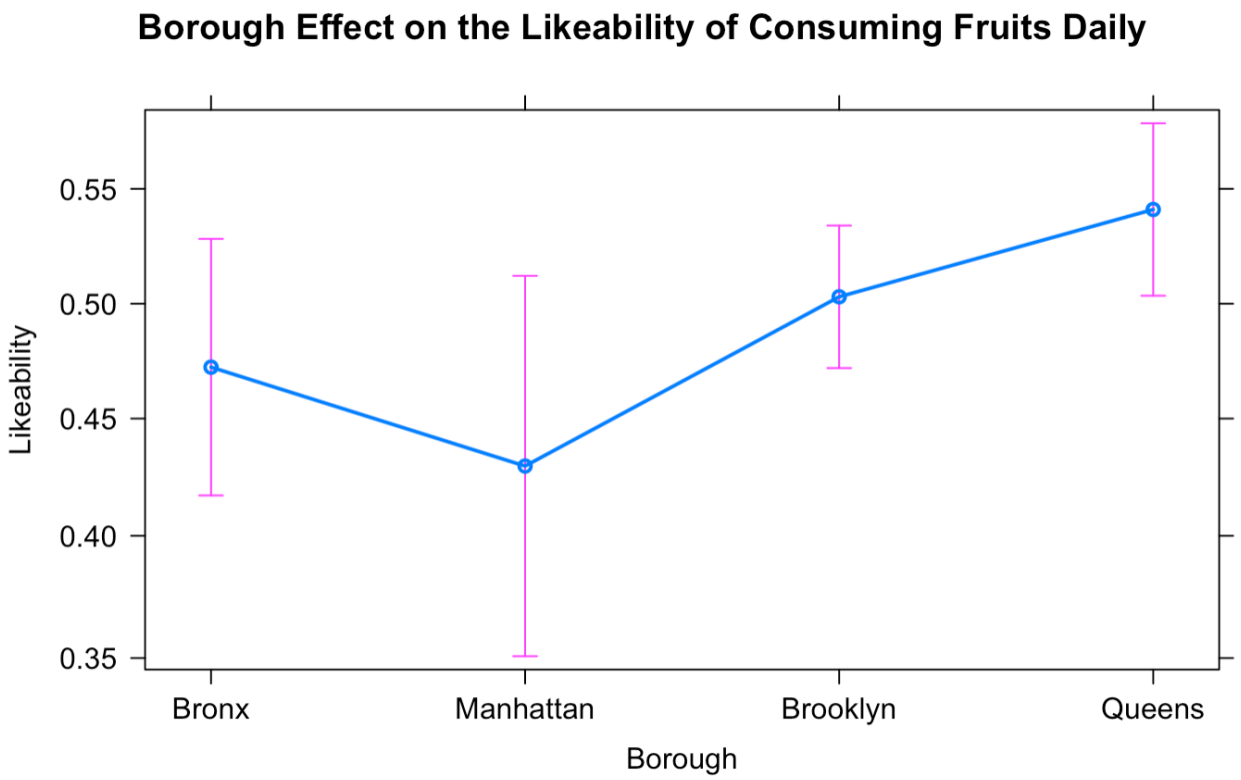
The next step in our research was to build a model in which we could showcase the relationship between a person's location within one of the five boroughs and its likeability of consuming fruits and vegetables daily.

In the BRFSS data, the frequency with which a person consumes fruits and veggies is given by the variables FRUIT2 and FVGREEN1 respectively. Those variables have five different levels (excluding missing values). The levels are daily, weekly, monthly, yearly, and never. Since we are following the CDC's guideline that sustains that a person needs to consume at least 2 ½ cups of fruits/vegetables at a daily basis, we have chosen to create dummy variables from these two variables, which allows us to capture whether a person consumes *at least* 1 fruit or vegetable daily, or not. Since this makes our dependent variable a binomial, we decided to use a logistic regression to access the relationship between the likeability of a person consuming fruits or vegetables daily, and its location within one of the five boroughs. To reiterate, we can do this now because we imputed a borough variable into BRFSS and each individual in our subset has been assigned a borough.

Model 3: In this model we have ran a logistic regression using glm() function specifying the family as binomial to determine the effect of borough location on whether a person consumes fruits daily or not. The following table showcases our results:

	Dependent variable:
	Daily.fruit
boroughManhattan	-0.173
	(0.204)
boroughBrooklyn	0.123
	(0.131)
boroughQueens	0.275**
	(0.138)
Constant	-0.111
	(0.114)
Observations	2,123
Log Likelihood	-1,467.348
Akaike Inf. Crit.	2,942.695
Note:	*p<0.1; **p<0.05; ***p<0.01

As demonstrated above, the only borough location that has a statistically significant correlation with the likeability of a person eating fruits daily is Queens. The results returned by the model contradict our initial assumptions that a person in the Bronx would consume less fruits and vegetables than a person in Manhattan. Recall that earlier in this paper, we discussed that this assumption was made while considering the data from the Supermarket Need Index. In fact, according to this logistic regression model, a person in the Bronx could consume more veggies than a person in Manhattan. However, the variability of the population in Manhattan is very great, as shown in the below graph.

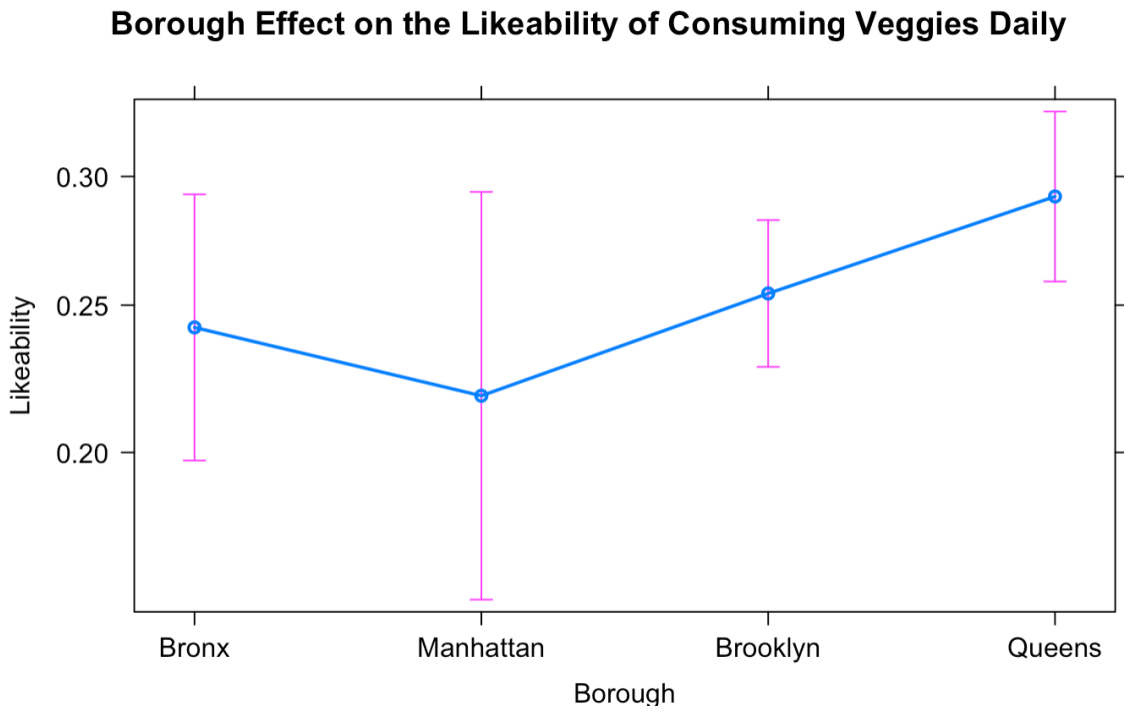


The variability of Manhattan respondents expressed by the pink line suggests that the daily fruit intake of people within the borough differs significantly, which might be explained by the fact that the demographics of people living north of 125th street and below that differ considerably. Moreover, the availability of supermarkets within the borough of Manhattan differs a lot as well, as demonstrated by the Supermarket Need Index. Another reason why Manhattan would have a higher error margin is that the sample size is significantly smaller than the other boroughs, with only 45 individuals. Overall, the results are very linear, ranging from more than 0.4 likeability to less than 0.55.

Model 2: For model 2, we use the same logistic regression except the dependent variable now represents daily vegetable intake instead of fruit. The results of the regression are shown below:

	Dependent variable:
	Daily.greens
boroughManhattan	-0.133
	(0.243)
boroughBrooklyn	0.066
	(0.151)
boroughQueens	0.256
	(0.157)
Constant	-1.142***
	(0.133)
Observations	2,114
Log Likelihood	-1,213.250
Akaike Inf. Crit.	2,434.501
Note:	*p<0.1; **p<0.05; ***p<0.01

Again, we haven't had results that are statistically significant for us to prove that borough has an effect on the dietary choices of people in New York City. The results are even more linear, and Manhattan again a large variability. Below is a graph for this model:



IV. Conclusion and Points of Improvement

Given the results of our logistic regressions, our study does not provide enough evidence for us to reject the null-hypothesis previously stated. Nevertheless, it is important to acknowledge that we have chosen to work with NYC boroughs instead of neighborhoods because of our data constraints, a problem in itself because boroughs in New York are highly diverse. Additionally, the city assessment of the supermarket shortage in New York City reports shortages in places like East Brooklyn, while a neighborhood like Williamsburg is very well-supplied with supermarkets.

The issue with zooming into neighborhoods with imputed data is that we increase our chances of errors caused by inaccurate predictions at the NYC and borough levels. Any error caused in classifying BRFSS NYS metropolitan area residents into NYC will create a bias within the borough classification. The borough classification errors similarly create a bias when classifying that subset into neighborhoods. In other words, when zooming into neighborhoods, our type 2 error will increase due to the endogenous nature of the in_NYC and borough variables, even though, in theory, type 2 error should decrease because we would account for specific locations that suffer from supermarket shortages. Ideally, we would be dealing with data that has a level of detail high enough to describe the neighborhood in which respondents are located. This way, we wouldn't have to account for type 2 error as a byproduct of machine-learning classification methods that are subjective to inaccuracies by nature.

Another limitation we encountered was that BRFSS doesn't explicitly define what units are being considered for some of the variables we used. For instance, in the 2019 codebook for BRFSS we see that FRUTDA2, which contains the response for the survey question "Fruit intake in times per day", does not specify what unit constitutes as 1 time of fruit intake. When running a quick summary of the variable, we see that the mean intake is 112.5 and the median is 100.0. If "times per day" is to be interpreted as whole fruits, an average response of 112.5 is unrealistic.

Like other studies previously conducted, our statistical analysis has not found enough evidence of a relationship between location and dietary habits. In the article "Food Deserts and the Causes of Nutritional Inequality", the authors cluster their population samples utilizing zip codes to delimitate neighborhoods, which is far more detailed than our borough approach[16]. The study did find evidence that supermarkets are concentrated in zip codes with higher income levels and generally offer healthier foods in comparison to grocery stores in lower-income neighborhoods[17]. Nevertheless, when analyzing the effect of a supermarket entry in a region considered a food desert, the study did not find any statistically significant evidence of change in

[16 -17-18] Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dube, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell. 2019. "Food Deserts and the Causes of Nutritional Inequality." Quarterly Journal of Economics 134 (4): 1793–1844.

the demand patterns of individuals within that area[18]. It is important to acknowledge that, as any other habit, eating patterns previously acquired may be resistant to change, and that the effects of greater availability and accessibility of healthful foods may be seen only in the longer-term.

Even though we were not able to quantitatively prove that there is a relationship between income, location and daily intake of fruits or vegetables, it cannot be denied (from a qualitative standpoint) that the supply of supermarkets and fresh produce in New York City is targeted and condensed by area. Low-income residents may still find ways to purchase and consume fruits and vegetables. However, if it takes more time, effort or a larger percentage of their income to do so, access to a healthy dietary pattern is unequal. Our study cannot conclude such a claim, but the data available to us is not specifically being surveyed to tackle these inequalities. Moving forward, to address inequalities in food access and fresh produce intake in NYC, surveys should aim to answer the question, *why did you not eat fruits and vegetables today?*

[18 -17-18] Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dube, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell. 2019. "Food Deserts and the Causes of Nutritional Inequality." Quarterly Journal of Economics 134 (4): 1793–1844.

Bibliography

Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dube, Jessie Handbury, Ilya Rahkovsky, and Molly Schnell. 2019. "Food Deserts and the Causes of Nutritional Inequality." Quarterly Journal of Economics 134 (4): 1793–1844.

Bader, Michael D. M., Marnie Purciel, Paulette Yousefzadeh, and Kathryn M. Neckerman. 2010. "Disparities in Neighborhood Food Environments: Implications of Measurement Strategies." Economic Geography 86 (4): 409–30.

"Best Price on Organic Strawberries at Trader Joe's." All-Natural Savings, 18 Feb. 2017. Available at: http://www.allnaturalsavings.com/best-price-on-organic-strawberries-at-trader-joes/.

Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System Survey Data. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2019.

Fitzpatrick, Katie, Nadia Greenhalgh-Stanley, and Michele Ver Ploeg. 2019. "Food Deserts and Diet-Related Health Outcomes of the Elderly." Food Policy 87 (August).

"Food Retail Expansion to Support Health (FRESH)." NYCEDC. Accessed December 17, 2020. Available at: https://edc.nyc/program/food-retail-expansion-support-health-fresh.

Hawkes, Corinna. "Dietary Implications of Supermarket Development: A Global Perspective." Development Policy Review 26, no. 6 (November 2008): 657–92.

Joassart-Marcelli, Pascale, Jaime S. Rossiter, and Fernando J. Bosco. 2017. "Ethnic Markets and Community Food Security in an Urban 'Food Desert.'" Environment and Planning A 49 (7): 1642–63.

New York City Department of City Planning, New York City Health, New York City Economic Development Corporation, "Going to Market: New York City's Neighborhood Grocery Store and Supermarket Shortage," October 2008. Available at: https://www1.nyc.gov/assets/planning/download/pdf/plans/supermarket/presentation_2008_10_29.pdf

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

U.S. Department of Health and Human Services and U.S. Department of Agriculture, "2015–2020 Dietary Guidelines for Americans," 8th Edition. December 2015. Available at: http://health.gov/dietaryguidelines/2015/guidelines/

Wolf-Powers, Laura. “Food Deserts and Real-Estate-Led Social Policy.” International Journal of Urban and Regional Research 41, no. 3 (May 2017): 414–25.