# Food Deserts and Literacy Rates:

# Educational Equity in NYS School Districts

DATA 705

Prof. Shortell

Brandon Max Flores

The Graduate Center

23 December 2025

**Abstract**

This study seeks to examine the relationship between food access, income, and literacy rates across race and sex demographics. The study uses NYSED school district ELA proficiency data (2023) to measure literacy, USDA Food Access Research Atlas (2019) data bound to 2010 Census Tract lines to measure food access, and U.S. Census Bureau American Community Survey (2023) bound to 2023 School District Census Tract lines to measure income. To account for different tract lines, this analysis uses an area-weighted spatial join to allocate data across tract levels to school district boundaries. Results showed that mean literacy was 43.84% (Standard Deviation = 14.73), and 60% of the population experienced low food access on average. Demographic disparities also arose in literacy, as Female students averaged 47%, Male (43%), Asian (39.9%), White (37.4%), Hispanic/Latino (26.5%), and Black (17.1%). Native American students averaged 3.8%; however, it is notable that only 100 districts reported Native American students, meaning the data may not be an accurate representation of the Native population. Regression analyses on all students revealed a statistically significant association between food access and literacy represented as ($\beta = -6.16$, $p < .001$) and accounting for 3.3% of variance. Poverty represented a larger effect, represented as ($\beta = -47.30$, $p < .001$). When controlling for poverty, food access was no longer statistically significant, with a p-value of 0.052 for all students. The model now represented 52.6% of the variance. Looking at race in particular, when controlling for all other variables except food access and demographics in the regression model, Asian students showed the strongest sensitivity ($\beta = -33.20$), followed by

White ($\beta = -30.88$), Hispanic/Latino ($\beta = -25.87$), and Black students ($\beta = -21.30$), all significant at $p < .001$. When controlling for all other variables than poverty rate and race in the regression model, Asian students showed the strongest sensitivity ($\beta = -166.918$), followed by White ($\beta = -155.747$), Hispanic/Latino ($\beta = -140.801$), and Black students ($\beta = -51.925$), all significant at $p < .001$. Findings suggest that food access has a mild association with literacy, with a varying impact by demographics. However, poverty is the dominant predictor explaining more variance and remaining statistically significant across all students.

**Introduction**

Food access and literacy gaps are frequently seen together in communities across the United States. A food desert, according to the USDA, is an area where residents have limited access to affordable and nutrient-rich food and is measured via distance to the nearest grocery store. This study seeks to explore if there is a relationship between food access and literacy rates across New York State while also examining the variability of this relationship by demographics such as race and sex. This study combines public datasets on food access, demographics, educational outcomes, and school zoning to identify patterns that may exist at the intersection of food security and educational achievement at the community level.

This exploration is guided by the Data Feminism principles of examining power and considering context through data. The data used and gathered during this study are not a reflection of the communities themselves, but rather a reflection of how they are measured, collected, and reported. This research does not intend to build or prove a narrative, but rather transform the professionally encoded datasets into more accessible representations, such as charts, graphs, non-technical writing, and an interactive map that can be utilized by the average community member at will.

**Data Sources Overview:**

*NYSED Report Card Data (2023-2024 School Year):* Provides proficiency rates based on English Language Arts (ELA) scores. Includes disaggregated demographic subgroups, including

Race and sex. The raw dataset included 649,933 records and was filtered down to district-level records, grades 3-8, and, where available, the subgroups of All Students, Female Students, Male Students, Black Students (Non-Hispanic), Asian Students (Non-Hispanic), White Students (Non-Hispanic), Native American Students (Non-Hispanic), and Hispanic Students. Data was retrieved from the NYSED Data webpage.

*USDA Food Access Research Atlas (FY 2019):* Provides census tract-level measures of food access, defining food access as living within ten miles from a supermarket for rural areas and one mile in urban areas. The dataset housed records for 4,858 New York State Census Tracts and included race categorical data, but omitted sex-categorical data. Data was from the USDA Food Access Research Atlas webpage.

*U.S. Census Bureau American Community Survey (FY 2023):* Provides demographic and economic data at the school district level. Includes median household income and poverty rates, and population counts for overall population and subgroups of Females, Males, Black individuals (Non-Hispanic), Asian individuals (Non-Hispanic), White individuals (Non-Hispanic), Native individuals (Non-Hispanic), and Hispanic individuals. Data was retrieved via the official census API for 664 unified school districts.

*TIGER/Line Geographic Boundary Files (FY 2010, FY 2023):* Provides 2023 school districts' geographic boundaries for the NYSED Report Card and ACS datasets, and 2010 Census Tracts' geographic boundaries for the 2019 USDA Food Access Research Atlas dataset. These boundary files enable spatial area-weighed analyses, allocating the USDA Food Access Research Atlas data from tracts to districts.

**Data Reflection**

Within *Data Feminism*, the authors bring forth the idea of examining power to draw attention to who is collecting data, who is represented in data, and whose interests are best served by data. This project uses the NYSED ELA scores to measure literacy as it has been defined as ELA proficient by the NYSED. It is crucial to note that these are standardized tests administered to grades 3-8 that reflect the creators of both the examination and the data capturers' choices as to what determines literacy and how it is measured.

Along with literacy, this project uses the USDA's Food Access Research Atlas to determine food security. It is important to note that the definition of low food access to the USDA is 1 mile from a supermarket in urban areas and 10 miles in rural areas. This definition does not account for transportation, what is considered a supermarket, and if near a supermarket, whether the food is attainable. Along with these oversights, edge cases such as people living .9 miles from a supermarket are considered food stable but are considered low food access if 1.1 miles away.

Another critical data issue in this study was the missing or suppressed records for ELA scores. Nearly all districts provided scores for all students and White students. Looking into other race categories saw far fewer reported numbers, with Native American students particularly low at only 100 reported districts. This systematically erases smaller demographics and not only renders these populations statistically invisible, but may also point to identity erasure, as many of these districts submitted "Mixed" populations.

Paired with examining power, a crucial component of my project is to reject deducing the results as a reflection of the communities themselves, but rather consider the context of the data. Black students averaged lower proficiency scores than other races. However, it is important to

consider the context of structural inequalities such as segregation and disinvestment that fostered different conditions, ultimately impacting the outcomes of students across generations.

New York State is a geographically diverse state and includes several different communities ranging from urban to rural. Due to these different communities, the concept of food access vastly differs across communities. Although urban communities may be more dependent on supermarkets, rural districts are home to several farming communities, which may impact food access scores as they are not counted as the categorical supermarket.

Government datasets are often challenging to access for the average person. Between the professionally coded information, along with some sources requiring API access through specific tools, my project seeks to employ the *Data Feminism* principle of "challenging power." This project aims to repurpose the difficult-to-access and interpret data into a comprehensible resource in the form of charts and an interactive map that community members and experts alike can utilize to understand how the factors analyzed impact their community at a glance.

## Methodology

This section is an abridged documentation for the full data processing and analytical pipeline conducted via Google Colab in a Jupyter Notebook.

### *Setup*

This analysis was conducted via Google Colab in a Jupyter Notebook using Python and the following packages:

- Pandas: Data manipulation
- Numpy: numerical operations
- Matplotlib: data visualizations

- Seaborn: data visualizations
- Scipy: statistical analysis
- Statsmodels: regression models
- Requests: ACS API calls
- Geopandas: Geospatial data handling
- Folium: interactive map creation
- Branca: unused

### *ACS Retrieval and Processing*

The US Census Bureau provides five-year and one-year estimates based on ACS results. To access the data, researchers must query the API using a particular syntax to return the needed variables for analysis.

#### *Defining Variables*

This study used the following ACS variable codes

- B19013_001E: Median Household Income
- C17002_001E, C17002_002E: Population for poverty status determination
- B03002_001E, B03002_003E-006E, B03002_012E: Race and ethnicity (Total, White Non-Hispanic, Black Non-Hispanic, Native Non-Hispanic, Asian Non-Hispanic, Hispanic/Latino)
- B01001_001E, B01001_002E, B01001_026E: Sex (Total, Male, Female)
- S1701_C01_001E, S1701_C02_001E, S1701_C03_001E Deeper poverty status determination

#### *API Requests*

*Call 1*

```
payload = {'get': 'NAME,{variables}', 'for': 'school district
(unified):*', 'in': 'state:36', 'key': censuskey}

r = requests.get('https://api.census.gov/data/2023/acs/acs5',
params=payload).json()
```

*Call 2*

```
#pulling more poverty data from ACS

base = "https://api.census.gov/data/2023/acs/acs5/subject"

vars_needed = ["NAME","S1701_C01_001E", # total population for whom
poverty status is determined "S1701_C02_001E", # count below poverty
level "S1701_C03_001E" # percent below poverty level]

params = {"get": ",".join(vars_needed), "for": "school district
(unified):*", "in": "state:36", "key": censuskey}

r = requests.get(base, params=params)
```

These API calls allowed me to request data for all unified school districts through the use of state code 36, which is NYS, and return all data based on the parameters I passed, which were the defined variables.

*Data Processing*

The ACS API returned data in JSON form, which needed to be converted to a more readable format. I passed the data into a dataframe using Pandas and dropped the first row to define

column names in place of the less comprehensible coded structure used by the ACS. Along with this, numeric columns were converted to integers from strings where applicable:

- District_Name, Median_Household_Income (numeric), Pop_Total_Poverty (numeric), Pop_Below_Poverty (numeric)
- Pop_Total_Race (numeric), Pop_White_NonHisp (numeric), Pop_Black_NonHisp (numeric), Pop_Native_NonHisp (numeric), Pop_Asian_NonHisp (numeric), Pop_Hisp_Latino (numeric)
- Pop_Total_Sex (numeric), Pop_Male (numeric), Pop_Female (numeric), State_Code, School_District_Code

*Poverty Data*

The initial poverty data pulled was not the proper data according to the ACS standard. I enhanced my dataset by calling S1701_C01_001E (Total population in poverty, S1701_C02_001E (Total population below poverty), and S1701_C03_001E (Poverty rate). With this data, I then converted the poverty rate, which was a percentage, into a proportion.

***USDA Food Atlas Data Processing***

The USDA Food Access Research Atlas was downloadable from the USDA website and came in a zip file containing a CSV file of census tract-level measures of food access, a codebook, and the applicable census tract lines.

*Data Processing*

Using the codebook, I selected categorical columns for my analysis and renamed them for readability:

- CensusTract was pulled, renamed to Census_Tract_FIPS to accurately reflect the 11-digit census tract identifier.
- State and county were pulled and renamed to State_Name and County_Name. This was done as a failsafe to help quickly identify geographies using name-matching across shapefiles if lines did not match up.
- LILATracts_1And10 was pulled and renamed to Flag_FoodDesert_1Mile. This is a binary flag for food desert status. For rural communities, it measures 10 miles from a supermarket, and for urban areas, it measures 1 mile.
- lapop1 was pulled and renamed to Count_LowAccess_Pop_1Mile. Importantly, this also encompasses 10 miles for rural areas. This is a count of the population with low food access.
- lalowi1 pulled and renamed to Count_LowAccess_LowIncome_1Mile. Importantly, this also encompasses 10 miles for rural areas. This is a count of the population with low food access and low income.
- lawhite1, lablack1, lahisp1, laasian1, and laaian1 were pulled and renamed using the same renaming syntax used for Count_LowAccess_Pop_1Mile and Count_LowAccess_LowIncome_1Mile.
- Pop2010 was pulled and renamed to Total_Tract_Pop_2010 and is the total population per tract from the 2010 Census.

- TractLOWI was pulled and renamed to Total_Tract_LowIncome and is the total population per tract from the 2010 Census who falls under low income
- TractWhite, TractBlack, TractHispanic, TractAsian, and TractAIAN were pulled and renamed using the same renaming syntax used for Total_Tract_Pop_2010 and Total_Tract_Pop_2010. These are the total populations per tract from the 2010 census, per racial identity

*Loading and filtering*

```
food_atlas_df = pd.read_csv('Food_Access_Research_Atlas.csv',
usecols=cols_to_keep)

food_atlas_df = food_atlas_df.query("State == 'New
York'").rename(columns=rename_mapping)
```

*Data Cleaning*

Auditing the data revealed that nearly 60% of values were missing for low access counts across the different low access count variables. This was not really missing data, however, and was rather null values representing that the population did not have low food access. These null values were treated as zeros during aggregation, as their null indicates a zero low-access population in the tract.

***NYSED Report Card Data Processing***

The NYS Education Department provides educational outcomes in a Microsoft Access database file. This database houses ELA assessment results for the 2023-2024 school year.

*Pre-Processing*

Before pulling the data into my notebook, I exported the data from the Access database to a CSV file. The file contained too many records to open, causing all filtering to take place in the notebook.

*Data Processing and Filtering*

I first loaded the data into a dataframe using pandas and pulled all the data in as strings.

```
report_cards_df = pd.read_csv('Annual_EM_ELA_CLEAN.csv', dtype='str')
```

After filtering, the columns pulled in were: ENTITY_CD, ENTITY_NAME, YEAR, ASSESSMENT_NAME, SUBGROUP_NAME, NUM_TESTED, PER_PROF, and NUM_PROF

After pulling in the data into the report_cards_df dataframe, I filtered the dataframe to include only ELA assessments for grades 3-8 for the 2023 school year:

```
target_assessments = ['ELA3', 'ELA4', 'ELA5', 'ELA6', 'ELA7', 'ELA8']
report_cards_df =
report_cards_df[(report_cards_df['ASSESSMENT_NAME'].isin(target_assessmen
ts)) & (report_cards_df['YEAR'] == '2023.0')]
```

Next, I converted numeric columns back to numbers from strings for analysis and to predetermine if any data was lost:

```
for col in ['NUM_TESTED', 'PER_PROF', 'NUM_PROF']: report_cards_df[col] =
pd.to_numeric(report_cards_df[col], errors='coerce')
```

| Variable | Suppressed/Missing | Percentage |
|---|---|---|
| NUM_TESTED | 0 | 0.0% |
| PER_PROF (Proficiency Rate) | 93,238 | 29.2% |
| NUM_PROF (Number Proficient) | 83,596 | 26.2% |

*Table 1. Data Suppression in NYSED Records*

Missing percentages are most likely a result of reporting/lack of data, not missing data. Values were nulled out to prevent errors in calculations.

After converting, I used name matching to find matching districts with my ACS dataset. I then aggregated the records to create a one-per-row district combination by summing my NUM_TESTED and NUM_PROF across each grade level:

```
NUM_TESTED and NUM_PROF across all grade levels:

combined_reportcards_df = report_cards_df.groupby(['ENTITY_CD',
'ENTITY_NAME', 'SUBGROUP_NAME'], as_index=False)[['NUM_TESTED',
'NUM_PROF']].sum(min_count=1)
```

I then recalculated the proficiency rate from the aggregated counts:

```
combined_reportcards_df['PER_PROF'] =
np.where(combined_reportcards_df['NUM_TESTED'] > 0,
(combined_reportcards_df['NUM_PROF'] /
combined_reportcards_df['NUM_TESTED']) * 100, np.nan)
```

*Proficiency Rate Distribution*

The distribution of proficiency rates across all records with valid data showed that there was a mean proficiency of 37.61% with a Standard Deviation of 22.17%. Full descriptive statistics showed:

| Statistic | Value |
|---|---|
| Count | 72,469 |
| Mean | 37.61% |
| Standard Deviation | 22.17 |
| Minimum | 0.00% |
| 25th Percentile | 21.18% |
| Median (50th) | 36.51% |

| | |
|---|---|
| 75th Percentile | 52.86% |
| Maximum | 100.00% |

*Table 2. Proficiency Rate Descriptive Statistics*

***Shape Files Processing***

Two TIGER/Line were loaded into GeoDataFrames using the GeoPandas package to enable spatial analysis:

*Census Tract Boundaries (2010)*

The 2010 Census Tract boundaries were pulled into a GeoDataFrame using:

```
census_geo_df = gpd.read_file('tl_2010_36_tract10.zip')
```

These boundaries align with the USDA Food Access Research Atlas and were provided in the zip file. This tract was specifically for NYS and held 4,919 tracts.

*School District Boundaries (2023)*

The 2023 Unified School District boundaries were pulled into a GeoDataFrame using:

```
school_geo_df = gpd.read_file('tl_2023_36_unsd.zip')
```

These boundaries directly align with the ACS dataset and match names with the NYSED Report Card dataset, and were suggested to use via the U.S. Census Developer Documentation. This tract was specifically for NYS and held 665 Unique School Districts.

***Subgroup Standardization***

The datasets used in the study had different naming structures for subgroups, so I created canonical subgroups to keep across datasets, which were defined within the notebook as:

```
CANON_SUBGROUPS = ['All Students', 'Male', 'Female', 'White
(Non-Hispanic)', 'Black (Non-Hispanic)', 'Asian (Non-Hispanic)', 'Native
(Non-Hispanic)', 'Hispanic/Latino']
```

The NYSED dataset required a special mapping dictionary to be created:

| NYSED Original Category | Standardized Label |
|---|---|
| All Students | All Students |
| Male | Male |
| Female | Female |
| White | White (Non-Hispanic) |
| Black or African American | Black (Non-Hispanic) |
| Asian or Native Hawaiian/Other Pacific Islander | Asian (Non-Hispanic) |
| American Indian or Alaska Native | Native (Non-Hispanic) |

| | |
|---|---|
| Hispanic or Latino | Hispanic/Latino |

*Table 3. NYSED Category Mapping*

*Unmapped Categories*

After applying the mapping to the NYSED dataset, I was left with 34% of the original rows, as the following categories were filtered out as they are out of scope of this study:

- Parent Not in Armed Forces
- Not Migrant
- Not in Foster Care
- Not Homeless
- Non-English Language Learner
- Students with Disabilities
- Economically Disadvantaged
- General Education Students
- Not Economically Disadvantaged
- Small Group Total: Race & Ethnicity

***Spatial Join: Food Atlas to School Districts***

As mentioned earlier, the USDA Food Atlas uses different geography boundaries than the ACS and NYSED datasets. To account for these different boundaries, an area-weighted spatial join was performed, allocating the USDA Food Atlas data to school districts.

*Creating Geographic Identifiers*

```
food_atlas_df['TRACT_GEOID'] =
food_atlas_df['Census_Tract_FIPS'].astype(str).str.zfill(11)

census_geo_df['TRACT_GEOID'] =
census_geo_df['GEOID10'].astype(str).str.zfill(11)

school_geo_df['GEOID_DISTRICT'] =
school_geo_df['GEOID'].astype(str).str.zfill(7)
```

I created columns within my dataframes to tie geographies with their original datasets based on their associated TIGER/Shape files.

*Merging Food Data with Tract Geometry*

I created a new DataFrame that takes the food atlas and its TIGER/Line shapefile and merges it with the Food Atlas dataframe

```
tracts_food = census_geo_df[['TRACT_GEOID',
'geometry']].merge(food_atlas_df, on='TRACT_GEOID', how='inner')
```

*Geometry & Spatial Merge*

To complete the geometric overlay, all intersections were computed between the Census Tract and Unified School District geometries:

```
tracts_food['TRACT_AREA'] = tracts_food.geometry.area

inter = gpd.overlay(tracts_food, districts, how='intersection',
keep_geom_type=False)

inter['INTER_AREA'] = inter.geometry.area

inter['AREA_FRAC'] = inter['INTER_AREA'] / inter['TRACT_AREA']
```

This computed an AREA_FRAC, which was a representation of how much of the tract falls within each district. For example, if 70% of a tract fell in the district, AREA_FRAC would equal .70

After computing the intersections between the geometries, population counts were then multiplied by area fractions to allocate proportionally to districts:

```
alloc_cols = la_cols + ['Total_Tract_Pop_2010', 'Total_Tract_LowIncome',
'Total_Tract_White', 'Total_Tract_Black', 'Total_Tract_Asian',
'Total_Tract_Native', 'Total_Tract_Hispanic']

for c in alloc_cols: inter[c + '_w'] = inter[c] * inter['AREA_FRAC'
```

```
district_food = inter.groupby(['GEOID_DISTRICT', 'NAME'],
as_index=False)[[c + '_w' for c in alloc_cols]].sum()
```

Food access shares were then calculated using the proportions from the area fractions to determine the proportion of the population with low food access:

```
district_food['LA_SHARE_1MI'] = np.where(district_food['POP_2010'] > 0,
district_food['LA_POP_1MI'] / district_food['POP_2010'], np.nan)
```

Race Specific:

```
district_food['LA_WHITE_SHARE_1MI'] =
np.where(district_food['Total_Tract_White_w'] > 0,
district_food['Count_LowAccess_White_1Mile_w'] /
district_food['Total_Tract_White_w'], np.nan)
```

***NYSED to ACS Crosswalk***

The NYSED used different codes to identify school districts than the ACS. To rectify this, I performed a name matching syntax on the name from the ACS and the NYSED district name columns. I added “SD”, “School District”, “UNSD”, and “Unified School District” to the ACS name column to find any matching names in the NYSED data. Once matches were found, I added a column to the NYSED data with the GEO_ID from the ACS dataframe to use in matching across dataframes when building my master dataframe. After completing this, I exported the file as a CSV.

*Loading, Standardizing, and Merging the Crosswalk*

Loading the crosswalk:

```
xwalk = pd.read_csv('matched_districts_rebase.csv')

xwalk['GEOID_DISTRICT'] = xwalk['GEOID'].astype(str).str.zfill(7)

xwalk['ENTITY_CD'] = xwalk['ENTITY_CD'].astype(str).str.replace(r'\.0$',
'', regex=True).str.zfill(12)
```

Standardizing the crosswalk:

```
school_geo_df['GEOID_DISTRICT'] =
school_geo_df['GEOID'].astype(str).str.zfill(7)

acs_df['GEOID_DISTRICT'] = (acs_df['State_Code'] +
acs_df['School_District_Code']).astype(str).str.zfill(7)

district_food['GEOID_DISTRICT'] =
district_food['GEOID_DISTRICT'].astype(str).str[-7:]

district_subgroup_ela['ENTITY_CD'] =
district_subgroup_ela['ENTITY_CD'].astype(str).str.strip().str.zfill(12)
```

Merging the crosswalk with the NYSED file:

```
nysed_ready = district_subgroup_ela.merge(xwalk[['ENTITY_CD',
'GEOID_DISTRICT']], on='ENTITY_CD', how='left')
```

***Master Dataset***

*Extracting All Student Data*

The NYSED dataset was apportioned as rows per subgroup (eg, racial identity, sex) rather than columns per subgroup, as seen in the ACS and USDA datasets. Due to this, each school district's subgroup data was first extracted and then put into a new dataframe, turning the disjointed row per subgroup data into columns merged into a single school district. I began by extracting the data for all students and pushing it to a new DataFrame using Pandas:

```
nysed_all = nysed_ready[nysed_ready['SUBGROUP_STD'] == 'All
Students'].copy()

nysed_all = nysed_all.rename(columns={'NUM_TESTED': 'ELA_NUM_TESTED_ALL',
'NUM_PROF': 'ELA_NUM_PROF_ALL', 'PER_PROF': 'ELA_PER_PROF_ALL'})
```

*Merging with USDA and ACS Data:*

Following the extraction of the "All Students" data, I then combined this data with the USDA and ACS data into a singular or master dataframe, which had one row per school district with all the data from the three sources per row in column format:

```
district_master = school_geo_df[['GEOID_DISTRICT', 'NAME',
'geometry']].merge(acs_df, on='GEOID_DISTRICT',
how='left').merge(district_food, on='GEOID_DISTRICT',
how='left').merge(nysed_all, on='GEOID_DISTRICT', how='left')
```

After merging the datasets, some data did not line up perfectly. Two of the ACS districts and nine of the NYSED school districts were not matchable across the three files. This may be due to the way school districts are grouped, as some school districts serve multiple townships. Additional considerations may suggest that the tract line file from 2010 had seen changes, eliminating or adding the missing districts post-spatial join.

*Extracting NYSED Subgroup Data*

After merging all student data, I did not pull the individual rows per subgroup into a new dataframe, but rather accessed this data during modeling, pulling the "SUBGROUP_STD" column from my NYSED dataset. This calls back to my standardization made from my canonical categories to measure each subgroup.

## ***Findings***

*Baseline*

I began by creating an initial sample for all students across all school districts and pulling a summary of the baseline statistics seen from the NYSED, USDA, and ACS DataFrames:

| Variable | Mean | SD | Min | 25% | Median | Max |
|---|---|---|---|---|---|---|
| ELA Proficiency (%) | 43.84 | 14.73 | 0.0 | 33.32 | 41.59 | 86.39 |
| Low Food Access Share | 0.60 | 0.32 | 0.0 | 0.36 | 0.67 | 1.00 |
| Poverty Rate | 0.10 | 0.06 | 0.0 | 0.06 | 0.09 | 0.40 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Median HH Income ($) | 92,913 | 39,332 | 37,569 | 66,045 | 78,937 | 250,001 |

*Table 4. NYSED Category Mapping*

These baseline statistics show that the average NYSED ELA proficiency rate was 43.84% across all districts, and as the median fell below the mean, it suggests that there was a slight skew towards the right of the data.

Food access showed that around 60% of the population of each district falls into low food access. However, it is notable to consider how access is reported, meaning the numbers may be skewed towards the left of the data for this reason.

Poverty rate tells a different story compared to food access, showing that the average poverty rate is 10%. Notably, the right tail showed 40% poverty. Median household income showed that half of the districts fell between $66k and $109k, a stark difference. Along with this, the maximum value was capped at $250,001, and may reflect that districts falling under this bracket saw higher incomes than the cap.

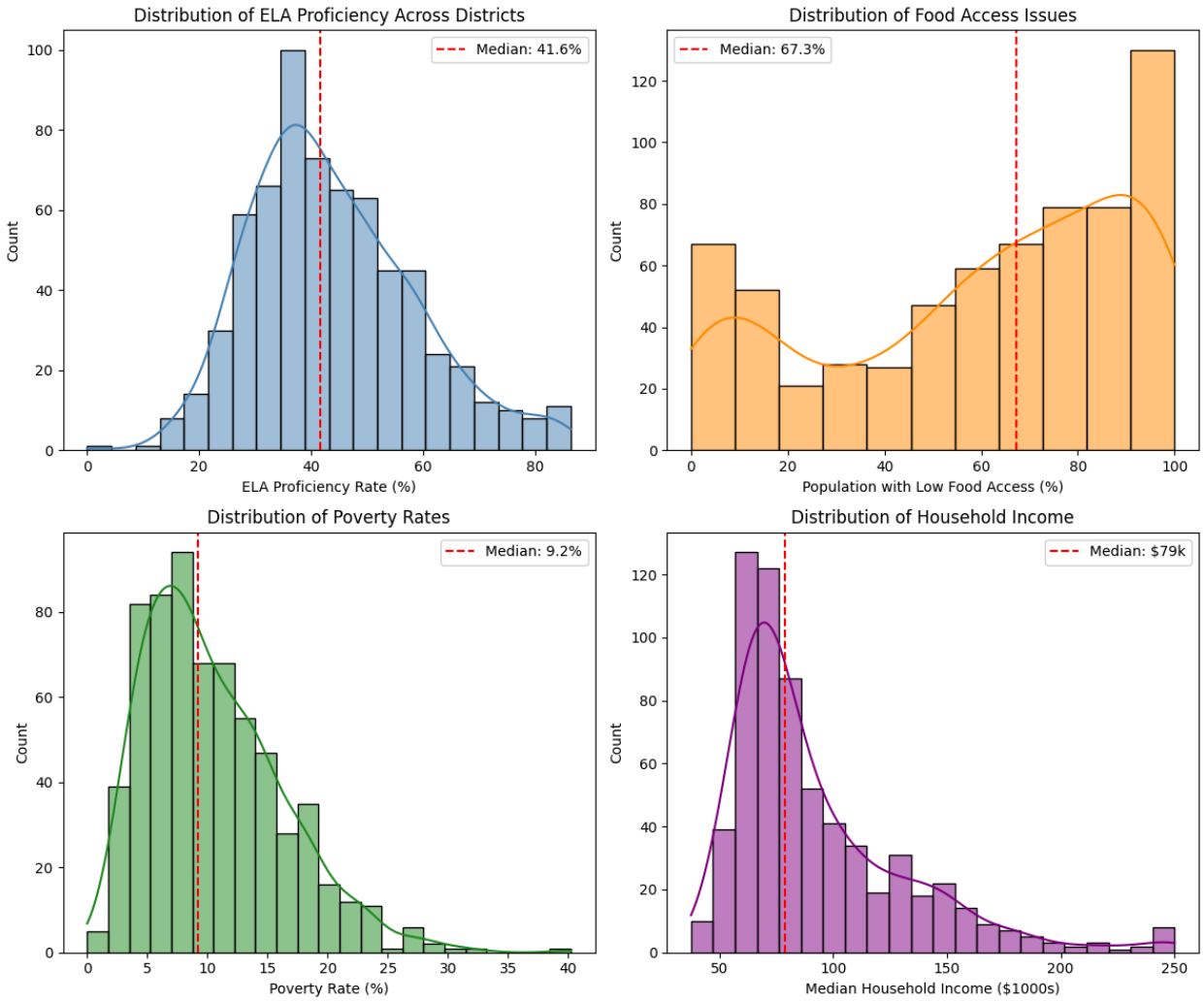


*Figure 1. NYSED Category Mapping*

## *Correlation: Food Access vs. Literacy*

Following my initial summary statistics, I ran a Pearson correlation test:

```
r, p_val = stats.pearsonr(analytic['LA_SHARE_1MI'],

analytic['ELA_PER_PROF_ALL'])
```

Result: Pearson correlation (Food Access vs Literacy): r = -0.181, p = 2.99e-06.

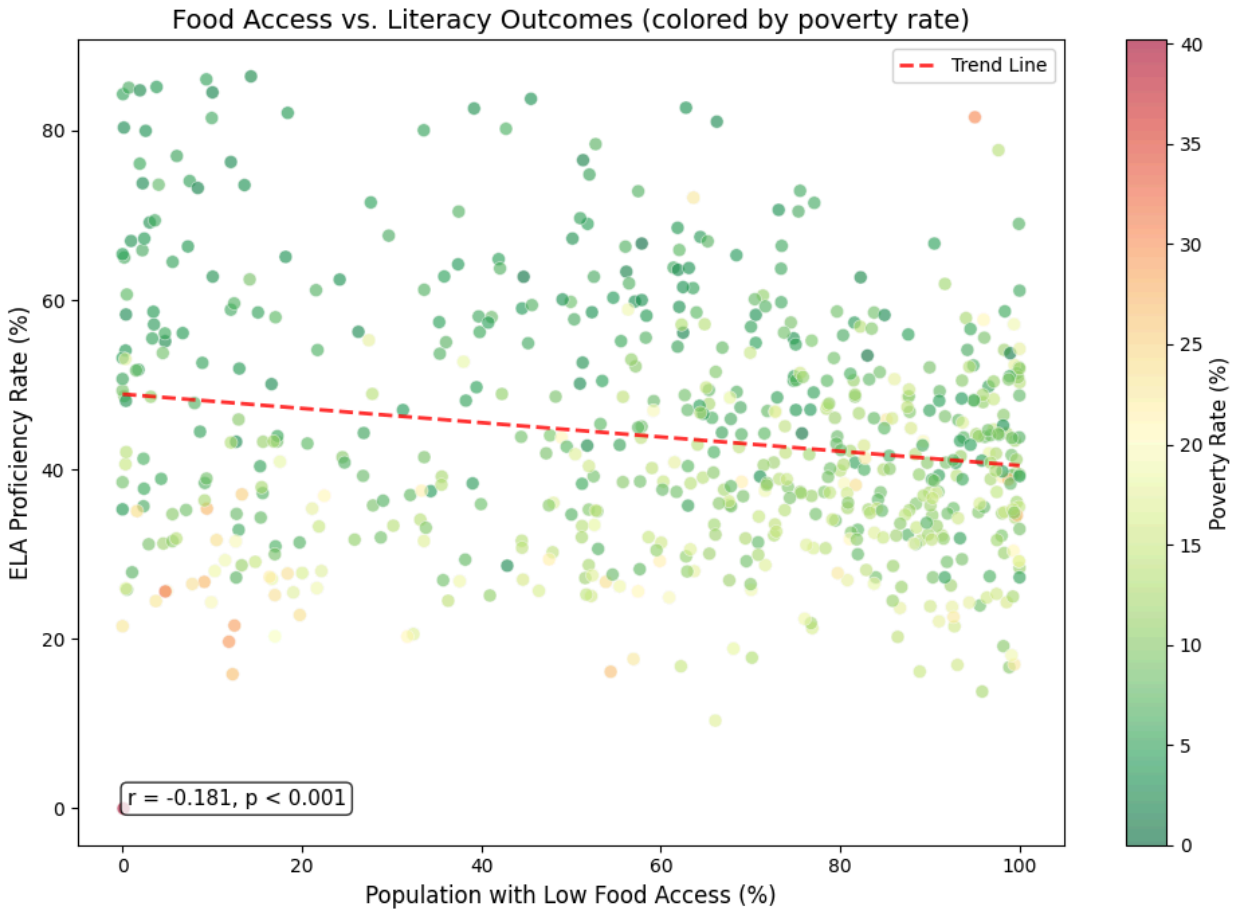


*Figure 2. Food Access vs. Literacy Outcomes*

Results showed that there is a statistically significant correlation between food access and literacy as measured by proficiency. Despite this relationship, it was weak, and looking at the line, there was not much clustering, and literacy ranged widely across all levels of food access.

*Correlation Across All Variables*

Although the relationship was weak between food access and literacy, Figure 2 showed a larger portion of literacy rates below the median shaded red, suggesting there may be a link between poverty and literacy.

To test this, I created a correlation matrix:

```
corr_vars = ['ELA_PER_PROF_ALL', 'LA_SHARE_1MI', 'LA_LOWCOME_SHARE_1MI',
'Poverty_Rate', 'Median_Household_Income']

corr_labels = ['ELA Proficiency', 'Low Food Access', 'Low Food Access
(Low Income)', 'Poverty Rate', 'Median Income']

corr_matrix = analytic[corr_vars].corr()

corr_matrix.index = corr_labels

corr_matrix.columns = corr_labels display(corr_matrix.round(3))
```

| | **ELA Prof** | **Food Access** | **Poverty** | **Income** |
|---|---|---|---|---|
| ELA Proficiency | 1.000 | -0.181 | -0.566 | 0.707 |
| Low Food Access | -0.181 | 1.000 | 0.088 | -0.364 |
| Poverty Rate | -0.566 | 0.088 | 1.000 | -0.630 |
| Median Income | 0.707 | -0.364 | -0.630 | 1.000 |

*Table 5. Correlation Matrix*

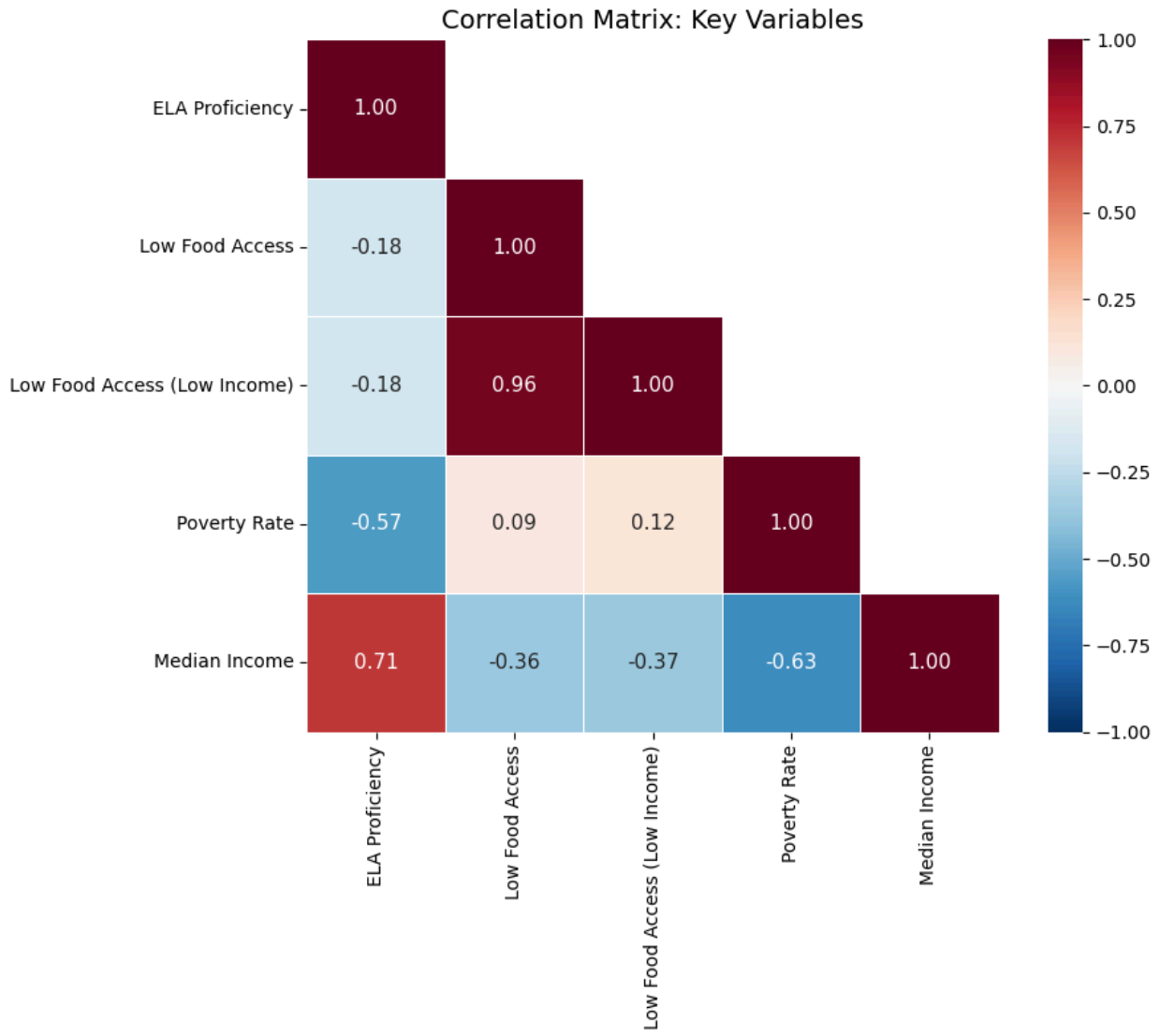


*Figure 2: Correlation Matrix*

Results showed that there was a potentially higher link between income and poverty with literacy, and suggest that higher poverty districts have lower literacy, while the inverse may be observed in higher income districts.

*Linear Regression: Food Access & Literacy*

Following my correlation matrix, I then tested my food access compared to literacy using an OLS model:

```
food_literacy_ols = smf.ols('ELA_PER_PROF_ALL ~ LA_SHARE_1MI',
data=analytic).fit()

print(food_literacy_ols.summary())
```

| Variable | Coefficient | Std Error | t-value | p-value | 95% CI |
|---|---|---|---|---|---|
| Intercept | 48.8983 | 1.213 | 40.328 | <0.001 | [46.52, 51.28] |
| LA_SHARE_1MI | -8.4276 | 1.788 | -4.713 | <0.001 | [-11.94, -4.92] |

Model Statistics: $R^2 = 0.033$, Adjusted $R^2 = 0.031$, $F(1, 654) = 22.21$, $p = 2.99e-06$, $n = 656$

*Table 6. Food Access and Literacy Regression Model*

Results from this model confirmed that there is a statistically significant relationship between food access and literacy. Although statistically significant, in observation, it would cause a mere decrease of 8.43 percentage points in literacy per one unit (100%) increase in food access score, or more realistically, a decrease of 0.84 percentage points per 10 percentage point increase in food access. Results also show that food access only explained 3.3% of the variance, meaning that food access alone is not the culprit for differences in literacy.

*Multiple Regression*

As the results of the one-variable OLS model showed that food access alone did not explain a high amount of variance, I then ran a multiple regression model to determine the impact of the ACS data:

```
model_controls = smf.ols('ELA_PER_PROF_ALL ~ LA_SHARE_1MI + Poverty_Rate
+ Median_Household_Income', data=analytic).fit()
```

| Variable | Coefficient | Std Err | t-value | p-value | 95% CI |
|---|---|---|---|---|---|
| Intercept | 25.9773 | 2.495 | 10.411 | <0.001 | [21.08, 30.88] |
| LA_SHARE_1MI | 2.6677 | 1.373 | 1.943 | 0.052 | [-0.03, 5.36] |
| Poverty_Rate | -47.3024 | 8.988 | -5.263 | <0.001 | [-64.95, -29.65] |
| Median_Household_Income | 0.0002 | 1.42e-05 | 16.113 | <0.001 | [0.0002, 0.0003] |

Model Statistics: $R^2 = 0.526$, Adjusted $R^2 = 0.524$, $F(3, 652) = 241.2$, $p = 2.91e\text{-}105$, $n = 656$

*Table 7. Multiple Regression Model*

Note: The condition number is large, 2.34e+06. This might indicate that there is strong multicollinearity or other numerical problems.

The output of this model showed that when controlling for poverty and median household income, the relationship between low food access and literacy is no longer statistically significant. The model also showed that poverty and income are strongly related to literacy, with this model explaining 52.4% of the variance. Notably, the model also showed in the output that poverty and income level showed strong multicollinearity, telling us that these variables were closely related as expected.

*Subgroup Summary Statistics*

Following the results of my multivariate model, I then printed the summary statistics of each subgroup's ELA performance:

```
subgroup_summary =
subgroup_analytic.groupby('SUBGROUP_STD')['ELA_PER_PROF'].agg(['count',
'mean', 'median', 'std']).round(1)
subgroup_summary.columns = ['N Districts', 'Mean', 'Median', 'Std Dev']
display(subgroup_summary.sort_values('Mean', ascending=False))
```

| Subgroup | N Districts | Mean (%) | Median (%) | Std Dev | Range |
|---|---|---|---|---|---|
| Female | 645 | 47.3 | 45.4 | 15.8 | 0-100 |
| All Students | 656 | 43.8 | 41.6 | 14.7 | 0-86 |

| | | | | | |
|---|---|---|---|---|---|
| Asian (Non-Hispanic) | 260 | 39.9 | 38.3 | 27.4 | 0-100 |
| Male | 645 | 37.8 | 35.5 | 14.5 | 0-100 |
| White (Non-Hispanic) | 629 | 37.4 | 36.6 | 21.3 | 0-100 |
| Hispanic/Latino | 428 | 26.5 | 25.7 | 18.8 | 0-100 |
| Black (Non-Hispanic) | 325 | 17.1 | 15.2 | 15.7 | 0-100 |
| Native (Non-Hispanic) | 100 | 3.8 | 0.0 | 9.9 | 0-67 |

*Table 8. NYSED ELA Subgroup Summary Statistics*

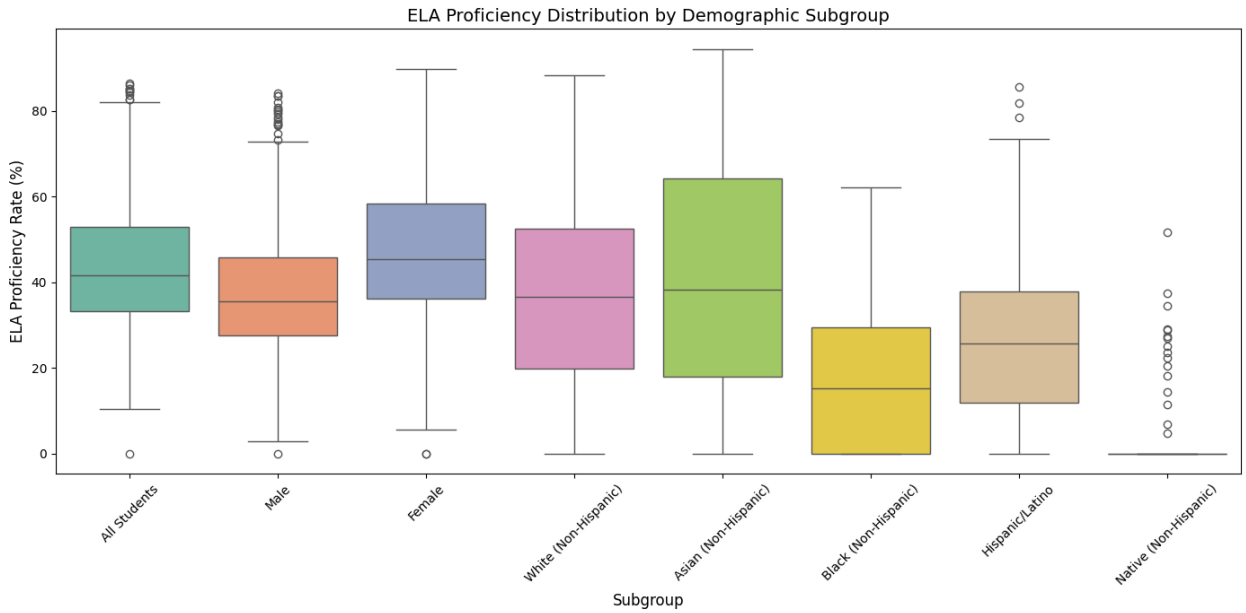


*Figure 3. NYSED ELA Subgroup Summary Box Plot*

Table 8 and Figure 3 show that the literacy rates vary by subgroup, with Asian and White students showing higher medians than other students. Notably, many districts did not report every subgroup or suppressed the data, meaning that there was a data availability constraint that may have impacted the outcomes.

*Regression by Subgroup*

Separate models were fit for each subgroup as long as there were more than 50 observations:

```
subgroup_results = []

for subgroup in subgroup_analytic['SUBGROUP_STD'].unique():

    subset = subgroup_analytic[subgroup_analytic['SUBGROUP_STD'] ==
subgroup].copy()

    if len(subset) >= 50:  # Minimum sample size for reliable regression

        Try:

            # Model using district-level food access

            model = smf.ols('ELA_PER_PROF ~ LA_SHARE_1MI + Poverty_Rate',
data=subset).fit()

            # Also try race-specific food access if available for this
subgroup

            race_specific_coef = np.nan

            race_specific_pval = np.nan

            if subset['LA_SUBGROUP_SHARE_1MI'].notna().sum() >= 30:
```

```
                try:
                    model_race = smf.ols('ELA_PER_PROF ~
LA_SUBGROUP_SHARE_1MI + Poverty_Rate',
data=subset.dropna(subset=['LA_SUBGROUP_SHARE_1MI'])).fit()
                    race_specific_coef =
model_race.params.get('LA_SUBGROUP_SHARE_1MI', np.nan)
                    race_specific_pval =
model_race.pvalues.get('LA_SUBGROUP_SHARE_1MI', np.nan)
                except:
                    pass
            subgroup_results.append({
                'Subgroup': subgroup,
                'N': len(subset),
                'Food Access Coef': model.params['LA_SHARE_1MI'],
                'Food Access p-value': model.pvalues['LA_SHARE_1MI'],
                'Poverty Coef': model.params['Poverty_Rate'],
                'Poverty p-value': model.pvalues['Poverty_Rate'],
                'R-squared': model.rsquared,
                'Race-Specific Food Coef': race_specific_coef,
                'Race-Specific Food p-value': race_specific_pval
            })
```

```
        except Exception as e:

            print(f"Could not fit model for {subgroup}: {e}")

results_df = pd.DataFrame(subgroup_results)

results_df = results_df.sort_values('Food Access Coef')

display(results_df.round(3))
```

| Subgroup | N | Food β | Food p | Poverty β | Pov p | $R^2$ |
|---|---|---|---|---|---|---|
| Asian (Non-Hispanic) | 260 | -33.195 | <.001 | -166.918 | <.001 | .247 |
| White (Non-Hispanic) | 629 | -30.878 | <.001 | -155.747 | <.001 | .419 |
| Hispanic/Latino | 428 | -25.866 | <.001 | -140.801 | <.001 | .401 |
| Black (Non-Hispanic) | 325 | -21.302 | <.001 | -51.925 | <.001 | .197 |
| Male | 645 | -8.108 | <.001 | -131.533 | <.001 | .315 |
| Female | 645 | -6.326 | <.001 | -152.154 | <.001 | .339 |
| All Students | 656 | -6.161 | <.001 | -140.636 | <.001 | .337 |
| Native (Non-Hispanic) | 100 | -0.743 | .810 | 57.714 | <.001 | .135 |

*Table 9. Multiple Regression NYSED Subgroup Model*

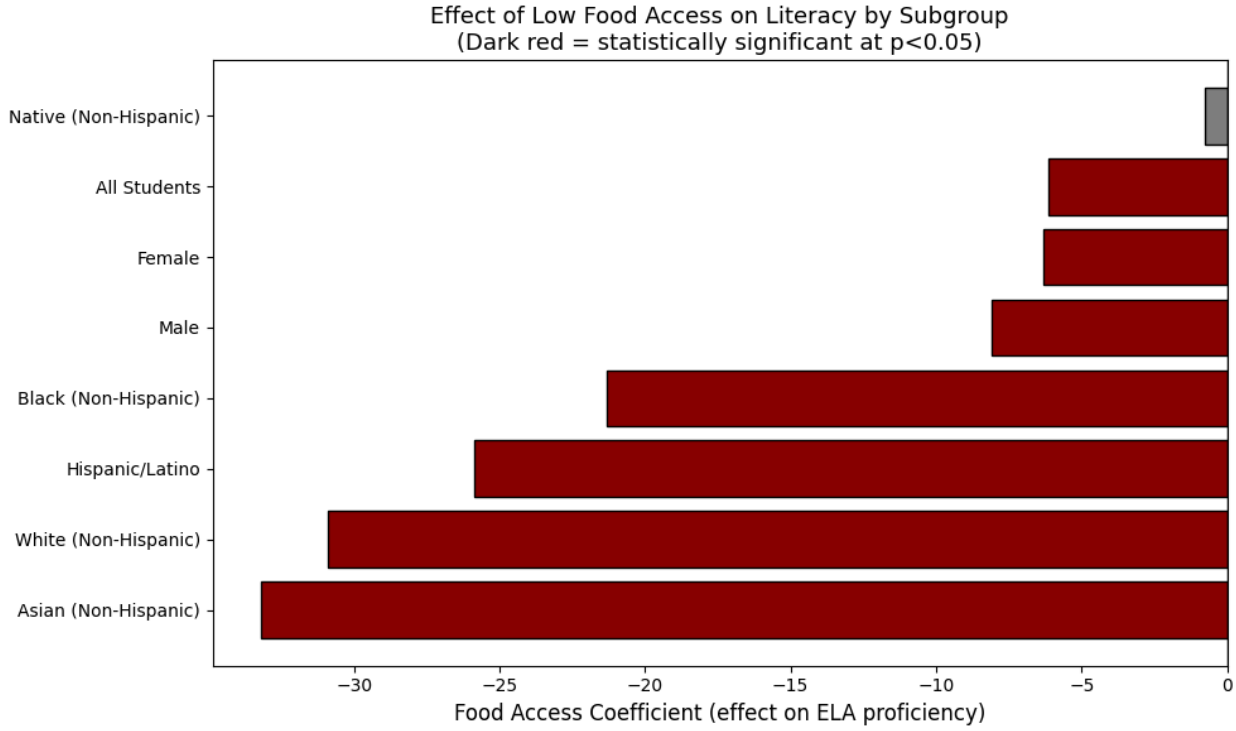


*Figure 4. Food Access Impact on Literacy by Subgroup*


*Figure 5. Food Access Impact on Literacy by Subgroup*

Subgroup-specific regression showed differences across subgroups for low food access and literacy after controlling for poverty rates. Both Figure 4, Figure 5, and Table 9 show that the impact of food access on literacy when controlling for poverty was stronger for certain subgroups compared to larger roll-up groups, such as all students or by sex. Asian students saw a 33.2 percentage point decrease in literacy per full unit increase in food access, or a more realistic 3.3 percentage point decrease with a 10 percentage point increase in low food access. Similar effects were seen for White, Hispanic, and Black students. Although subgroup-specific models revealed larger associations than all students, it is important to note that strong associations that may have been formed due to associations within particular subgroups can be lost when rolled up into one group.

The regression model also showed that across all subgroup models, poverty rate emerged as a consistent predictor of literacy measured by ELA proficiency. Most subgroups showed that a 10 percentage point increase in poverty per district is associated with a 12 to 17 percentage point decrease in literacy. This pattern held across race and sex-based subgroups, indicating that although food access is associated with literacy, poverty has stronger correlations.

*Interactive Map*

A major component of this study was transforming the datasets into an interpretable-at-a-glance resource for professionals and non-professionals alike. I created an interactive map after exporting my master datasets to a JSON file in JS. This map is hosted at fleurhes.com/food-literacy-map

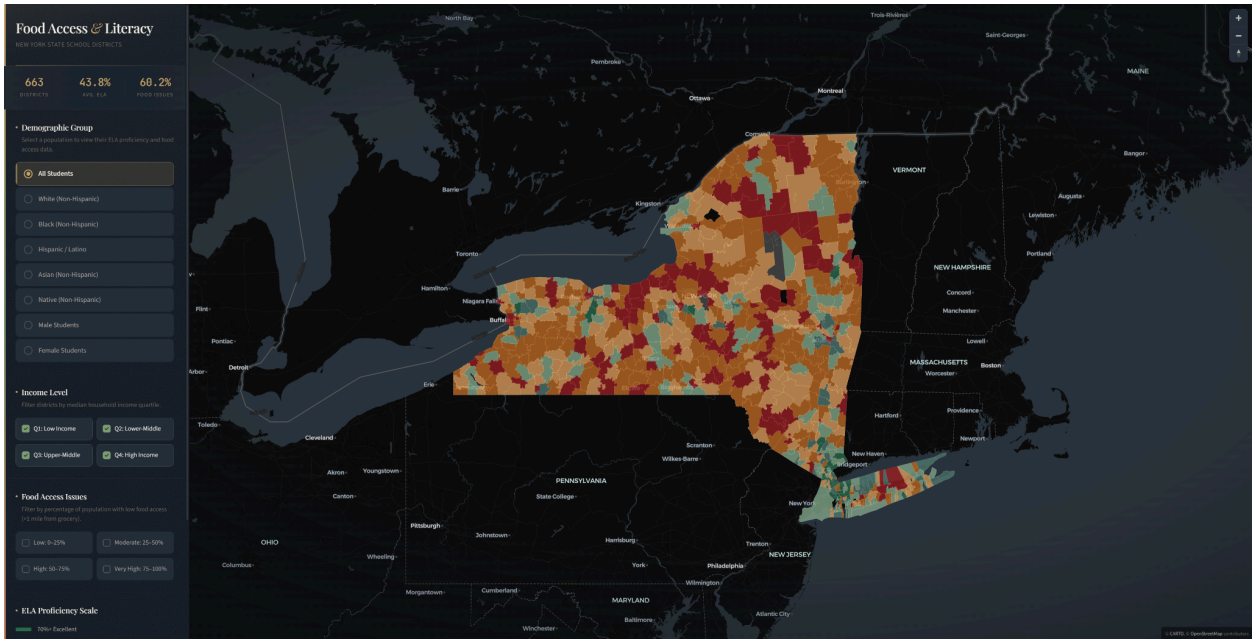


*Figure 6. Interactive*

*Overall Findings*

The results of this study suggest that there is a higher link between poverty and literacy than between food access and literacy. It is important to note that the data provided was not all-inclusive across populations and may have had significant impacts on the results. This impact was strongly felt when mapping, as some school districts showed substantially higher "All Student" literacy rates compared to race or sex-based literacy rates, which should be what drives the "All Students" rate.

### *Reflection on Ethical Approaches*

#### *Privacy & Data Suppression*

Although suppression of data is an important protection that allows student data to remain private when a subgroup is at a short length that a student may be identifiable, this suppression also systematically erases smaller demographic groups from the dataset. Native American students were invisible in the majority of districts, causing an ethical dilemma in that the same privacy protections meant to protect students ultimately made them invisible, harming proper analysis.

#### *Historical Inequality*

When considering the results of this study, it is important to consider the disparities that exist across communities. Framing or attributing lower literacy rates as a direct result of race, sex, or income can reinforce harmful stereotypes and detract from the underlying systemic causes.

This study attempted to avoid framing deficits by identifying systemic causes in the introduction, acknowledge that the combinations of all factors tested did not account for over 55% of total variance, and drawing attention to the limitations of the public datasets when

possible and necessary. Most importantly, however, this study seeks to present patterns to investigate rather than draw conclusions about communities. It is crucial to note that the datasets provided were created, encoded, and measured in ways that do not directly benefit the sample in which they are tested. Literacy and the variables examined are not a one-size-fits-all solution, and the results should not be treated as such.

*Limitations of Quantitative Analysis*

Ultimately, this study cannot capture the lived experiences of students facing food insecurity or make visible families who face food insecurity despite not being defined as such by the USDA dataset.

***Additional Resources***

*Limitations*

- The USDA Food Access Research Atlas predates the ACS and NYSED datasets by four years.
- Nearly 30% of proficiency records are either missing or suppressed, disproportionately impacting the analysis of demographic groups.
- The USDA Food Access Research Atlas measures food access near a supermarket. This does not accurately reflect or capture the ability to afford food, alternative solutions in rural communities, or edge cases falling right under or above the thresholds defined as 1 mile for urban and 10 miles for rural.
- The NYSED ELA results may not fully capture literacy skills by design.

*Useful Resources for Further Analysis*

- SNAP Data: SNAP data may be a more meaningful indicator of food insecurity.

- Individual-Level Data: Datasets linking individuals' demographics, food access, income, and educational outcomes would improve analysis efforts.
- Community-Defined Measures: This study seeks to transform professional data into accessible community data in the form of an interactive map. Partnering with community voices to develop relevant measures within communities may improve experience and relevancy.