# Problem Statement

Which NJ county is potentially a food desert and can be picked for investing to open a supermarket with fresh produce section?

# Introduction

In this paper, we will look at USDA’s food desert dataset and will try to find a county (or counties) in New Jersey which has food desert and will be a good candidate for investment in fresh food/ retail grocery sector. The paper will also dive into the dataset to find meaningful correlations between existence of food deserts and other demographic data points. Finally, the paper will explore linear predictive model(s) to identify impacted population size for existence of food desert in a county and logistic regression model to identify potential food deserts based on census tract datapoints.

# Background

## What is a food desert?

According to Annie et al. (2021), Food deserts are geographic areas where residents have few to no con­ve­nient options for secur­ing affordable and healthy foods — especially fresh fruits and vegetables. Disproportionately found in high poverty areas, food deserts create extra, everyday hurdles that can make it harder for kids, families, and communities to grow healthy and strong.

## Where are food deserts located

According to Annie et al. (2021), food deserts are more common in areas with:

* smaller populations.
* higher rates of abandoned or vacant homes; and
* residents who have lower levels of education, lower incomes, and higher rates of unemployment.

Food deserts are also a disproportionate reality for Black communities, according to a 2014 study from Johns Hop­kins University (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3970577/>).

The study compared U.S. census tracts of similar poverty levels and found that, in urban areas, Black communities had the fewest supermarkets, white communities had the most, and multiracial communities fell in the middle of the supermarket count spectrum.

## How are food deserts identified?

According to Annie et al. (2021), researchers consider a variety of factors when identifying food deserts, including:

* Access to food, as measured by distance to a store or by the number of stores in an area.
* Household resources, including family income or vehicle availability.
* Neighborhood resources, such as the average income of the neighborhood and the availability of public transportation.

One way that the U.S. Department of Agriculture identifies food deserts is by searching for low-income, low access census tracts (<https://www.ers.usda.gov/webdocs/publications/93141/eib%20209%20summary.pdf?v=6115.7>).

In low-access census tracts, a significant share (33% or more) of residents must travel an inconvenient distance to reach the nearest supermarket or grocery store (at least 1 mile in urban areas and 10 miles in rur­al areas).

In low-income census tracts, the local poverty rate is at least 20% or the median family income is at most 80% of the statewide median family income.

## Why do food deserts exist?

According to Annie et al. (2021), there is no single cause of food deserts, but there are several contributing factors. Among them:

* Transportation challenges - Low-income families are less likely to have reliable transportation, which can prevent residents from traveling longer distances to buy groceries.
* Convenience food - Low-income families are more likely to live in communities populated by smaller corner stores, convenience markets and fast-food vendors with limited healthy food options.
* Added risks - Opening a supermarket or grocery store chain is an investment risk, and this risk can grow to prohibitive proportions in lower-income neighborhoods. For example: The purchasing power of customers in these communities - including families enrolled in the Supplemental Nutrition Assistance Program - can change dramatically over the course of a month. At the same time, the threat of higher crime rates, whether real or perceived, can raise a business’s insurance fees and security costs.
* Income inequality - Healthy food costs more. When researchers from Brown University and Harvard University studied diet patterns and costs (<https://bmjopen.bmj.com/content/3/12/e004277>), they found that the healthiest diets — meals rich in vegetables, fruits, fish and nuts — were, on average, $1.50 more expensive per day than diets rich in processed foods, meats and refined grains. For families living pay­check to pay­check, the higher cost of healthy food could make it inaccessible even when it’s readily available.

# Reason for this study

According to Annie et al. (2021), nearly 39.5 million people - 12.8% of the U.S. population - were living in low-income and low-access areas, accord­ing to the USDA’s most recent food access research report, pub­lished in 2017 (<https://www.ers.usda.gov/webdocs/publications/82101/eib-165.pdf?v=3395.3>).

Within this group, researchers estimated that 19 million people - or 6.2% of the nation’s total population - had limited access to a supermarket or grocery store.

Moreover, the coronavirus pandemic injected even more challenges — both logistical and financial — into the complex field of food access.

As COVID-19 cases rose across the country, restaurants, corner stores and food markets — among other businesses — closed their doors or reduced their operating hours. Residents who relied on public trans­porta­tion for fetching groceries faced additional hurdles, including new travel restrictions and scaled-back service schedules.

Beyond making it harder to get to the grocery store, the pandemic also kicked off an economic crisis that made it harder for some families to afford groceries. In fact, nearly 10% of parents with only young children — kids ages five and under — reported having insufficient food for their families and insufficient resources to purchase more, according to a fall 2020 food insecurity update from Brookings (<https://www.brookings.edu/blog/up-front/2020/11/23/hungry-at-thanksgiving-a-fall-2020-update-on-food-insecurity-in-the-u-s/>).

This case study will help investors identify potential ***NJ counties*** in which to invest. Both **Federal and State Governments** should also look into these areas, implementing [strategies](https://www.aecf.org/blog/exploring-americas-food-deserts), like:

* Incentivizing grocery stores and supermarkets in underserved areas.
* Funding city-wide programs to encourage healthier eating.
* Extending support for small, corner-type stores and neighborhood-based farmers markets.

***Reference:***

The Annie E. Casey Foundation (2021). From Food Deserts in the United States. <https://www.aecf.org/blog/exploring-americas-food-deserts>

# Dataset

This data is pulled from the Food Access Research Atlas site. Documentation can be found here:

https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/

This data measures access by the Census-Tract, and as such provides a fairly granular overview.

# Methodology

1. Gather dataset from USDA’s [site](https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/),
2. Make sure the date is clean by performing simple data profiling (null checks, spell check for States and counties); convert all cell containing ‘NULL’ to ‘’ using built- in replace function in excel,
3. Filter on just New Jersey’s data,
4. Copy the data over to a new excel file as the source file is huge and any operation on it will have performance issue,
5. Make sure the data type for ‘State’ and ‘County’ is set to “Geography”,
6. Aggregate the data at county grain using pivot table; choose ‘County’ as the Rows for the pivot table,
7. Add the sum of population, occupied housing unit, and average of poverty rate and median family income as values in the pivot table,
8. Also add in sum of low-income population, housing units without a vehicle, housing units receiving SNAP benefits as value in the pivot table, flags for low access tracts at-
   * 1 mile for urban areas and 10 miles for rural areas
   * 0.5 mile for urban areas and 10 miles for rural areas
   * 1 mile for urban areas and 20 miles for rural areas
9. Finally add sum of low-access population count for different mile radiuses as well as low-access low-income population count for different mile radiuses as values in the pivot table,
10. Use data bars from conditional formatting on all the values column in the pivot table; this will help us identify hard hitters in our data set,
    * Use green gradient fill for population, occupied housing unit and median income
      + Bigger the bar, better place to invest for lesser risk
    * Use red gradient fill for low-income population count, SNAP benefit receivers count, low-access plus low-income population counts in different radiuses
      + Bigger the bar, more risk to invest
    * Use yellow gradient file for the remaining ones (housing units without a vehicle, low-access population counts in different radiuses)
      + Bigger the bar, more necessary to invest (Higher impact)
    * Try to find county (or counties) that would have bigger green & yellow bars and smaller red bars if possible
11. Copy the content of the pivot table to a blank sheet for visualization purpose,
12. Create a dashboard with maps for population, housing units, poverty & median family income and bar charts for low access population living outside 0.5 mile in urban / 10 mile in rural, low-income population, SNAP benefit receivers & housing unit with no vehicle
    * Try to find county (or counties) that would have low to moderate poverty, SNAP benefit receivers, housing units with no vehicle and moderate to high population, housing units, median income as well as people living in low access area.
13. Compare findings from step 10 & 14 and reach final conclusion.

# Findings

First, the source dataset was filtered on ‘State’ column to just keep rows for ‘New Jersey’. There were 2002 rows with 147 columns/ attributes. Then the first pivot table was created where the below listed attributes were aggregated at ‘County’ level:

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Description** | **Operation type** |
| POP2010 | Population counts from 2010 census | Sum |
| OHU2010 | Occupied housing unit count from 2010 census | Sum |
| PovertyRate | Share of the tract population living with income at or below the Federal poverty thresholds for family size | Average |
| MedianFamilyIncome | Tract median family income | Average |
| LILATracts\_1And10 | Flag for food desert when considering low accessibility at 1 and 10 miles | Sum |
| LILATracts\_halfAnd10 | Flag for food desert when considering low accessibility at 1/2 and 10 miles | Sum |
| LILATracts\_1And20 | Flag for food desert when considering low accessibility at 1 and 20 miles | Sum |
| TractLOWI | Total count of low-income population in tract | Sum |
| TractHUNV | Total count of housing units without a vehicle in tract | Sum |
| TractSNAP | Total count of housing units receiving SNAP benefits in tract | Sum |
| LAPOP05\_10 | Population counts beyond 1/2 mile for urban areas or 10 miles for rural areas from supermarket | Sum |
| LAPOP1\_10 | Population counts beyond 1 mile for urban areas or 10 miles for rural areas from supermarket | Sum |
| LAPOP1\_20 | Population counts beyond 1 mile for urban areas or 20 miles for rural areas from supermarket | Sum |
| LALOWI05\_10 | Low-income population count beyond 1/2 mile for urban areas or 10 miles for rural areas from supermarket | Sum |
| LALOWI1\_10 | Low-income population count beyond 1 mile for urban areas or 10 miles for rural areas from supermarket | Sum |
| LALOWI1\_20 | Low-income population count beyond 1 mile for urban areas or 20 miles for rural areas from supermarket | Sum |

Data Bar formatting was used with different color scales to identify different factor.

Table

Description automatically generated

Looking at the data bars, some county candidates were selected those would have less risk (bigger green bar, smaller red bar) and moderate to high impact (moderate to big yellow bars):

1. Ocean
2. Middlesex
3. Monmouth
4. Burlington

The pivot table was copied over to another spreadsheet and pasted as a data table for visualization purposes.

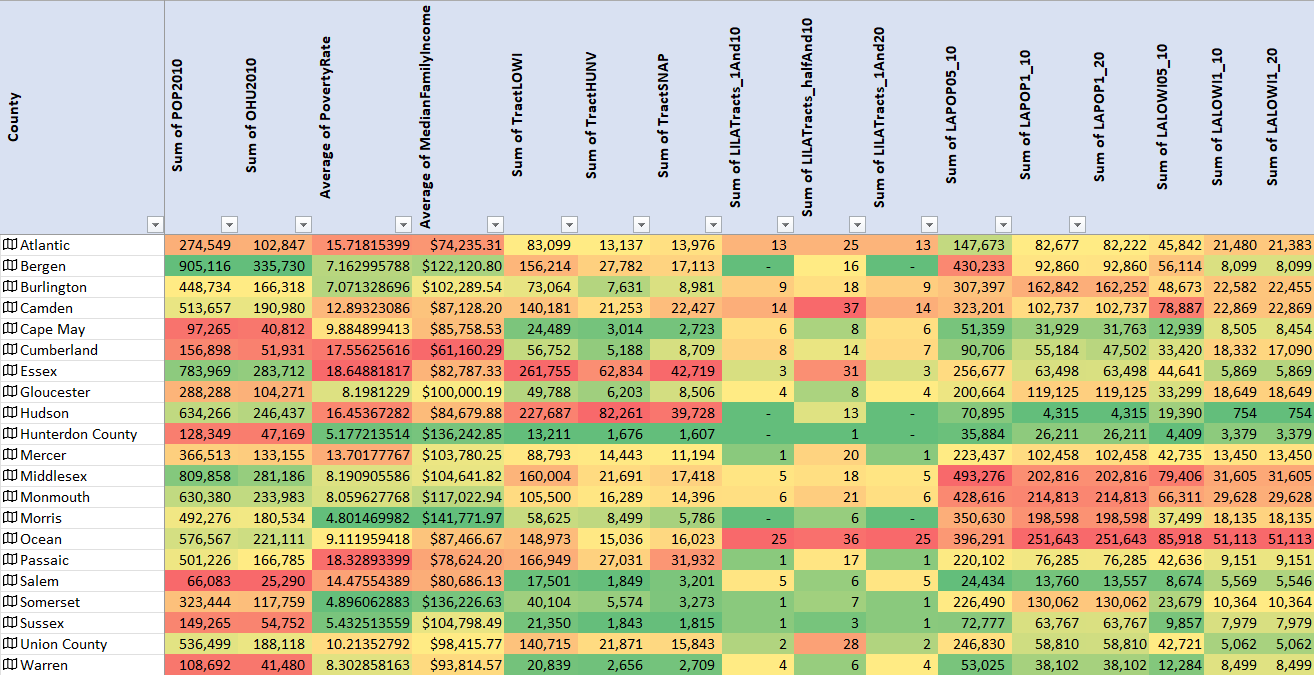


Figure : Data Table

Again, some county candidates were selected based on the following constraints:

1. Counties with red-ish cells in right most 9 columns are preferred as it would have bigger impact
2. Counties with green-ish cells in the left most 7 columns are preferred as it would have less risk

So, looking at the color scales on the data table, below counties were selected:

1. Ocean
2. Monmouth
3. Morris
4. Burlington

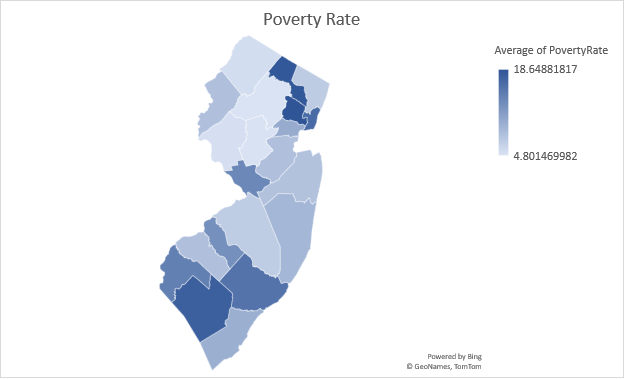
So, both analyses have produced same list of county candidates. Let’s dive deeper using some dashboards to make sure no other potential candidates have been left out.

Counties with high count of low access population are preferred so that new grocery stores would impact larger population. So, potential candidates would be (considering 300,000k cutoff) –

1. Bergen
2. Burlington
3. Camden
4. Middlesex
5. Monmouth
6. Morris
7. Ocean

Thus, a list of seven counties as potential candidates to invest in was generated.

Let us look at the poverty rates and low-income populations counts per county:



But counties with higher poverty rate and low-income population should be avoided to lower the risk, as these regions suffer from low purchasing power as well as perceived high crime rates. This is how Camden was eliminated and even though Bergen, Middlesex & Ocean counties have higher low-income population, they were not eliminated just because their overall poverty rates were lower. Thus, we are left with 6 counties from the initial list.

Now, let us look at the SNAP receivers count:

We would like to avoid counties with higher SNAP benefit receivers as it correlates to low purchasing power. No further counties were eliminated from the remaining 6.

Let us take a look at the Figure 1: Data table again and check the Sum of LILATracts\_1And10 & Sum of LILATracts\_1And20 for all those 6 counties. Bergen and Morris do not have any –

1. Low income and low access tract measured at 1 mile for urban areas and 10 miles for rural areas
2. Low income and low access tract measured at 1 mile for urban areas and 20 miles for rural areas

Also, these 2 counties do not have high housing units without vehicles

Which means these counties do not require immediate intervention as majority of the people can travel a little bit to get to the nearest food /grocery stores. So, after eliminating these 2, the final list of counties is as follows:

1. Ocean – Moderate risks due to moderate poverty rate and lower than average median household income, but highest impact in terms of population helped by investment.

2. Middlesex – Low risk due to lower than average poverty rate and higher than average median household income, great impact in terms of population helped by investment.

3. Monmouth - Low risk due to lower than average poverty rate and higher than average median household income, moderate impact in terms of population helped by investment as it has lower low-income population than Middlesex.

4. Burlington - Low risk due to lower than average poverty rate and higher than average median household income, least impact among these 4 in terms of population helped by investment

# Exploratory Data Analysis (Bonus)

## Correlation Matrix

Table

Description automatically generated with low confidence

Above is the correlation matrix for all the attributes in data table (Figure 1: Data table). Conditional formatting was used to color high correlation value (Pearson’s P - avlue) with red. High correlations were seen between:

1. Population & housing unit counts (positive)
2. Population & low-income population count (positive)
3. Population & Low access, population at 1/2 mile for urban areas and 10 miles for rural areas, number (positive)
4. Population & Low access, low-income population at 1/2 mile for urban areas and 10 miles for rural areas, number (positive)
5. Occupied housing unit counts & low-income population count (positive)
6. Occupied housing unit counts & Low access, population at 1/2 mile for urban areas and 10 miles for rural areas, number (positive)
7. Occupied housing unit counts & Low access, low-income population at 1/2 mile for urban areas and 10 miles for rural areas, number (positive)
8. Poverty Rate & median household income (negative)
9. Low-income population & households without vehicle (positive)
10. Low-income population & SNAP benefit receivers (positive)
11. Households without vehicle & SNAP benefit receivers (positive)

Apparently, none of these attributes mentioned above have high correlations with either food desert tract counts or population living in those tracts.

## Inhabitants of Food Deserts by Race

For this analysis, we just looked at Low income and low access tract measured at 1 mile for urban areas and 10 miles for rural areas (LILATracts\_1And10 = 1)

According to the chart, it is inferred that most of the food deserts are inhabited by Caucasians. Other races tend avoid living in those areas.

# Data Mining / Prediction (Bonus)

## Linear Predictive models

4 Linear predictive models were created:

1. Predicting LAPOP05\_10 - Population count beyond 1/2 mile for urban areas or 10 miles for rural areas from supermarket
2. Predicting LAPOP1\_20 - Population count beyond 1 mile for urban areas or 10 miles for rural areas from supermarket
3. Predicting LALOWI05\_10 - Low income population count beyond 1/2 mile for urban areas or 10 miles for rural areas from supermarket
4. Predicting LALOWI1\_20 - Low income population count beyond 1 mile for urban areas or 20 miles for rural areas from supermarket

### Predicting LAPOP05\_10

Below is the summary output of the linear model form Excel:

Table

Description automatically generated

As we can see the model has 97.58% accuracy in predicting LAPOP05\_10 based on the R-squared value.

### Predicting LAPOP1\_20

Below is the summary output of the linear model form Excel:

Graphical user interface, application, table, Excel

Description automatically generated

As we can see the model has 82.38% accuracy in predicting LAPOP1\_10 based on the R-squared value.

### Predicting LALOWI05\_10

Below is the summary output of the linear model form Excel:

Graphical user interface, application, table, Excel

Description automatically generated

As we can see the model has 96.33% accuracy in predicting LALOWI05\_10 based on the R-squared value.

### Predicting LALOWI1\_20

Below is the summary output of the linear model form Excel:

Table

Description automatically generated

As we can see the model has 89.11% accuracy in predicting LALOWI05\_10 based on the R-squared value.

So, comparing all R-squared values from these models, and based on the data at hand, the best linear predictive is LAPOP05\_10 prediction model. And the equation for this model is:

Sum of LAPOP05\_10 = -138720.709974656 + 2.41971464195797\*Sum of POP2010 -4.18181883670726\*Sum of OHU2010 + 1602.61335362862\*Average of PovertyRate + 0.93723377754756\*Average of MedianFamilyIncome -2.61567781571532\*Sum of TractLOWI + 0.0642154197094879\*Sum of TractHUNV + 3.75352621558695\*Sum of TractSNAP + 8121.76602932826\*Sum of LILATracts\_1And10 + 2547.67996996631\*Sum of LILATracts\_halfAnd10 -533.242631823472\*Sum of LILATracts\_1And20

## Logistic regression model

For this modeling I have used the following columns as explanatory variable:

Urban, Pop2010, OHU2010, GroupQuartersFlag, NUMGQTRS, PCTGQTRS, HUNVFlag, LowIncomeTracts, PovertyRate, MedianFamilyIncome, TractLOWI, TractKids, TractSeniors, TractWhite, TractBlack, TractAsian, TractNHOPI,TractAIAN, TractOMultir, TractHispanic, TractHUNV, TractSNAP

And I used LILATracts\_1And10 as the response variable.

*\*\*Metadata can be found in the excel file.*

Below is the summary output of regression statistics from Excel:

Application, table, Excel

Description automatically generated

Using these coefficients, we can predict whether LILATracts\_1And10 will be equal 0 or 1 after rounding the result to 0th place in the following formula:

P(X) = e^( -27.1747573037723 + 4.11761056773327\*Urban - 0.000704960539063267\*Pop2010+ 0.00015653342909736\*OHU2010 - 26.6673110310635\*GroupQuartersFlag + 9.44036181288358E-05\*NUMGQTRS + 0.0504626470640226\*PCTGQTRS + 2.38067307466838\*HUNVFlag + 24.5787670693794\*LowIncomeTracts - 0.0783237931206215\*PovertyRate - 3.66572000796622E-05\*MedianFamilyIncome + 0.00056680033787382\*TractLOWI - 0.0008201356482283\*TractKids + 0.000995828551259705\*TractSeniors + 0.0010017465834465\*TractWhite + 0.000645223609282556\*TractBlack -2.12881922468186E-05\*TractAsian - 0.013825269710104\*TractNHOPI -0.0036749704961802\*TractAIAN + 0.00344110577041204\*TractOMultir - 0.0020574475578762\*TractHispanic - 0.00921012621165106\*TractHUNV + 0.00354615880992393\*TractSNAP)

So, if P(X) > 0.5, the predicted value will be 1 which will identify that census tract has a food desert. On the other hand, if P(X) < 0.5, the predicted value will be 0 which will identify that census tract does not have a food desert.