Food Security Analysis – Final Report

Context

Background

According to the U.S. Department of Agriculture's Economic Research Service (ERS), accessibility and affordability of food retailers, distance to food retailers, and the availability of high quality, healthy food options all impact people's decision making regarding food spending and eating habits. ERS highlights the impact of intersectionality, specifying that income level, location, and demographic factors all coalesce to influence consumers' diet and health outcomes.

Problem Statement

Using the ERS Food Environment Atlas data, I reduced the dataset to features containing the word "2016." Using the year 2016, I was able to limit my exploration to dig into the factors that impacted food insecurity more thoroughly. Through an exploratory examination of the Food Environment Atlas data for 2016, I visually explored opportunities for reducing food insecurity

Methodology

Data Wrangling

When I initially imported the State/County dataset, the data only had four columns with 852,810 rows. I recognized that the dataset's variables were stored in a column titled "Variable_Code." Furthermore, the Variable_Code column also contained data from multiple years. To combat these issues, I first limited the data to one year: 2016. My initial goal for the data wrangling process was to focus on reducing dimensionality by decreasing size through limiting my data to one year. From there, I removed all punctuation in the variables, made the words lowercase, and removed the word "County." I then converted the Variable_Code column to a series of columns, one for each variable. This contributed to my goal of reducing dimensionality by creating additional columns, while removing repetitive rows.

With my DataFrame shifted, I began the rest of the data wrangling process with 55 columns and 3328 rows. Next, I addressed null values. I dropped columns with greater than 50% missing values. I imputed state totals for each missing value for columns with missing data for state totals. I then imputed the median for each missing value of remaining columns with missing values. To further reduce the size of the data, I removed columns that were duplicative.

The existing data only had information regarding food stores and assistance programs, so I wanted to add demographic data to bolster the dataset. I loaded education and education data.

These datasets contained information by state. I altered the column names, and I merged them with my initial DataFrame.

I created some initial visualizations to address outliers. The final DataFrame had 31 columns and 3140 rows.

K-Means

K-Means forms clusters that are spherical and centered around a centroid. I started my clustering with K-Means because of its easy interpretability and efficiency.

To determine the best number of clusters to use for K-Means, I used the elbow method and silhouette scores. When using the elbow method, I aimed to choose a K that was not too large and minimized the sum of squares. This resulted in a K of 7.

Silhouette scores evaluate how well the clusters are divided. Through the use of silhouette scores, I was able to determine the best number of clusters to use. However, the silhouette scores for my data were relatively low for all amounts of clusters, with the highest silhouette score being for 2 clusters. Like with the elbow method, I needed to choose a number that was not too large but also had a high silhouette score.

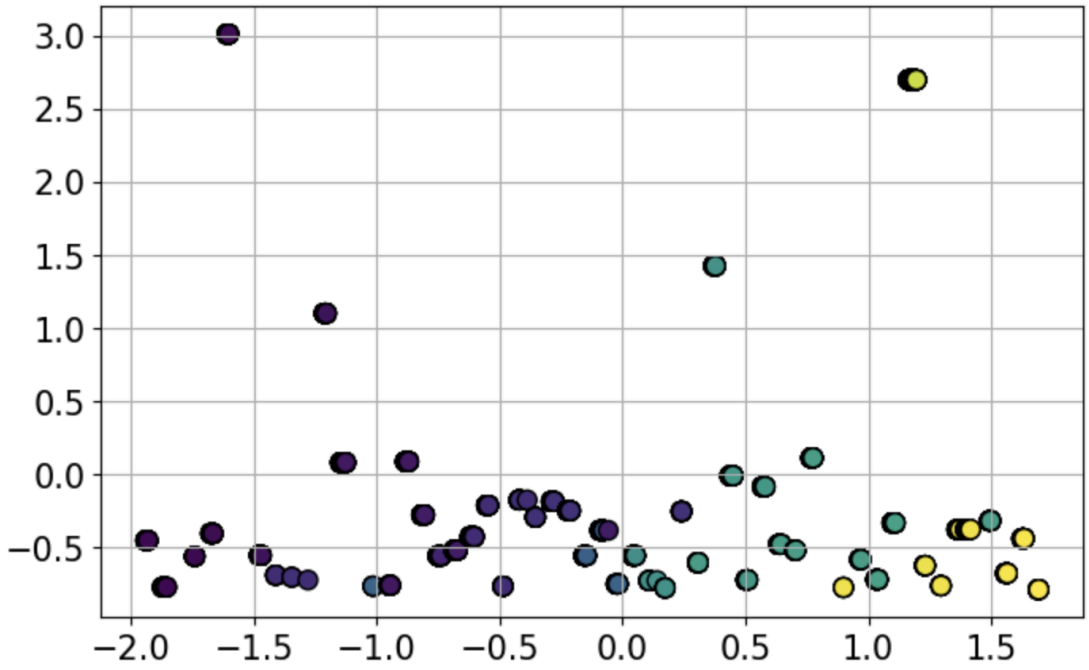
Next, I completed Principal Component Analysis (PCA), which allowed me to reduce dimensionality by converting the data into principal components. The principal components were a set of linearly uncorrelated variables. By reducing dimensionality, I was able to complete K-Means and more clearly define the clusters.

Affinity Propagation

When I began clustering, I had not chosen a number of clusters, so I used Affinity Propagation as one of my clustering methods because of its ability to determine the proper number of clusters based on the data.

The silhouette coefficient for Affinity Propagation was 0.541, which indicated that the clusters were not well-defined.

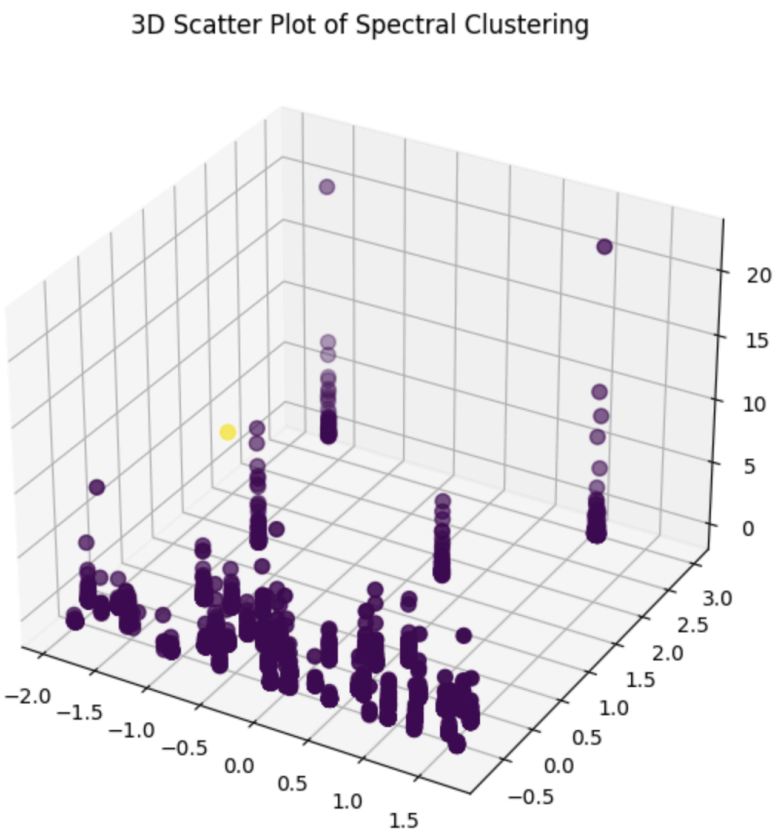
The Affinity Propagation clustering method determined that 443 clusters was the proper amount; however, this number of clusters was too high to be considered a good fit for the data.



Spectral Clustering

Spectral Clustering is useful when dealing with clusters that are not well-defined. With this in mind, I used Spectral Clustering because I suspected that my data would not be well-separated by traditional distance metrics.

The silhouette coefficient for Spectral Clustering was 0.966. This was the highest silhouette coefficient out of all the clustering methods I used. Despite the high silhouette score, the algorithm selected 443 clusters as the best number, which was an unreasonably high number of clusters.

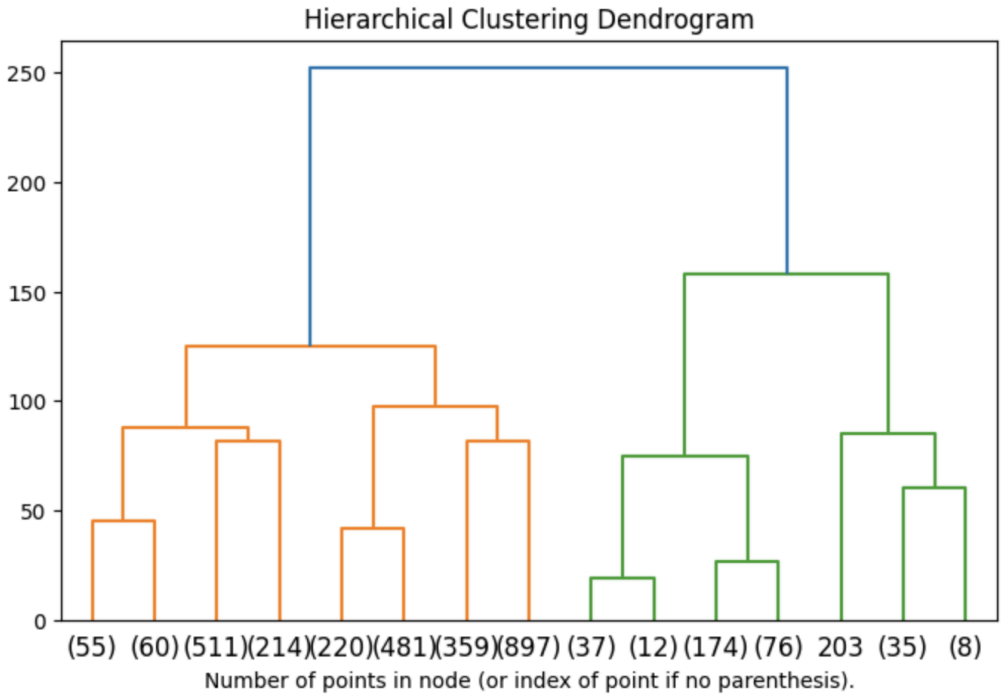


Agglomerative Clustering

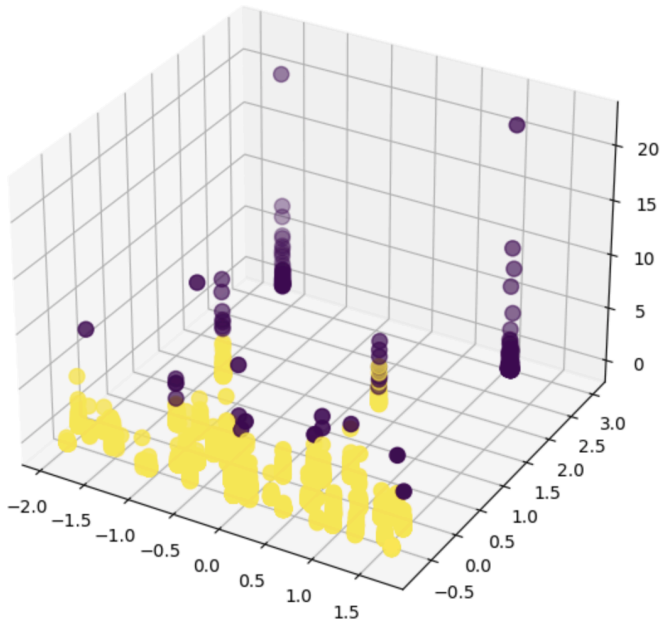
I used Agglomerative Clustering because it initially labeled each data point as a cluster, and as the algorithm continued, the clusters merged depending on their similarity until a certain distance threshold was met. I set the distance threshold to 1.0, and I allowed the algorithm to determine the proper number of clusters.

The silhouette coefficient for Agglomerative Clustering was 0.5, which indicated that the clusters were not well separated.

I used a dendrogram to map the Agglomerative Clustering, and based on the visualization, I determined that two clusters was the best number. From there, I applied this finding to a new Agglomerative Clustering algorithm with two specified clusters and created a scatter plot to visualize the clusters. The visualization showed that the clusters were not well separated.



3D Scatter Plot of Agglomerative Clustering

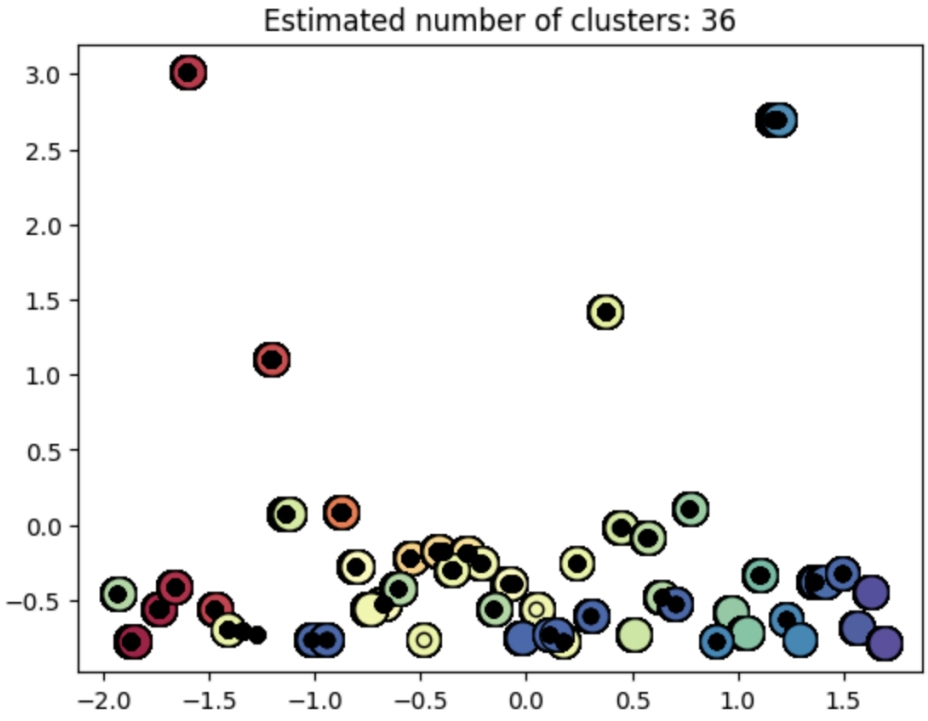


DBSCAN

I chose DBSCAN as one of my clustering methods because of its ability to identify clusters of varying shapes and sizes, while not being affected by noise. I defined the distance at which points could be considered neighbors at two and the minimum number of data points at three. The outliers were given a label of -1.

The silhouette coefficient for DBSCAN was 0.469, which indicated that the clusters were not well-defined.

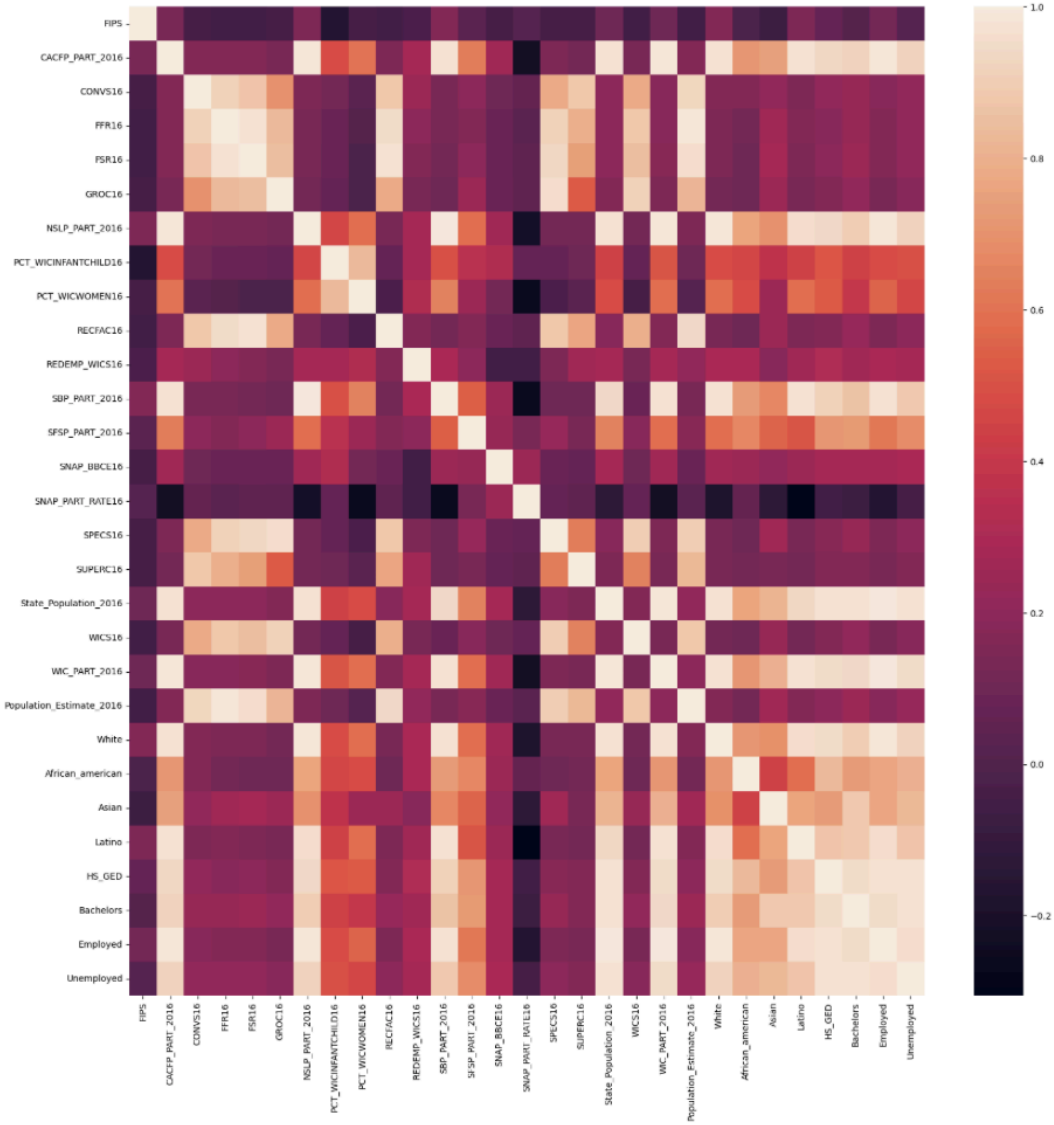
I plotted the DSCAN results on a scatter plot. The noise was identified with black dots on the plot. Based on the results of the DBSCAN, the ideal number of clusters was 36.



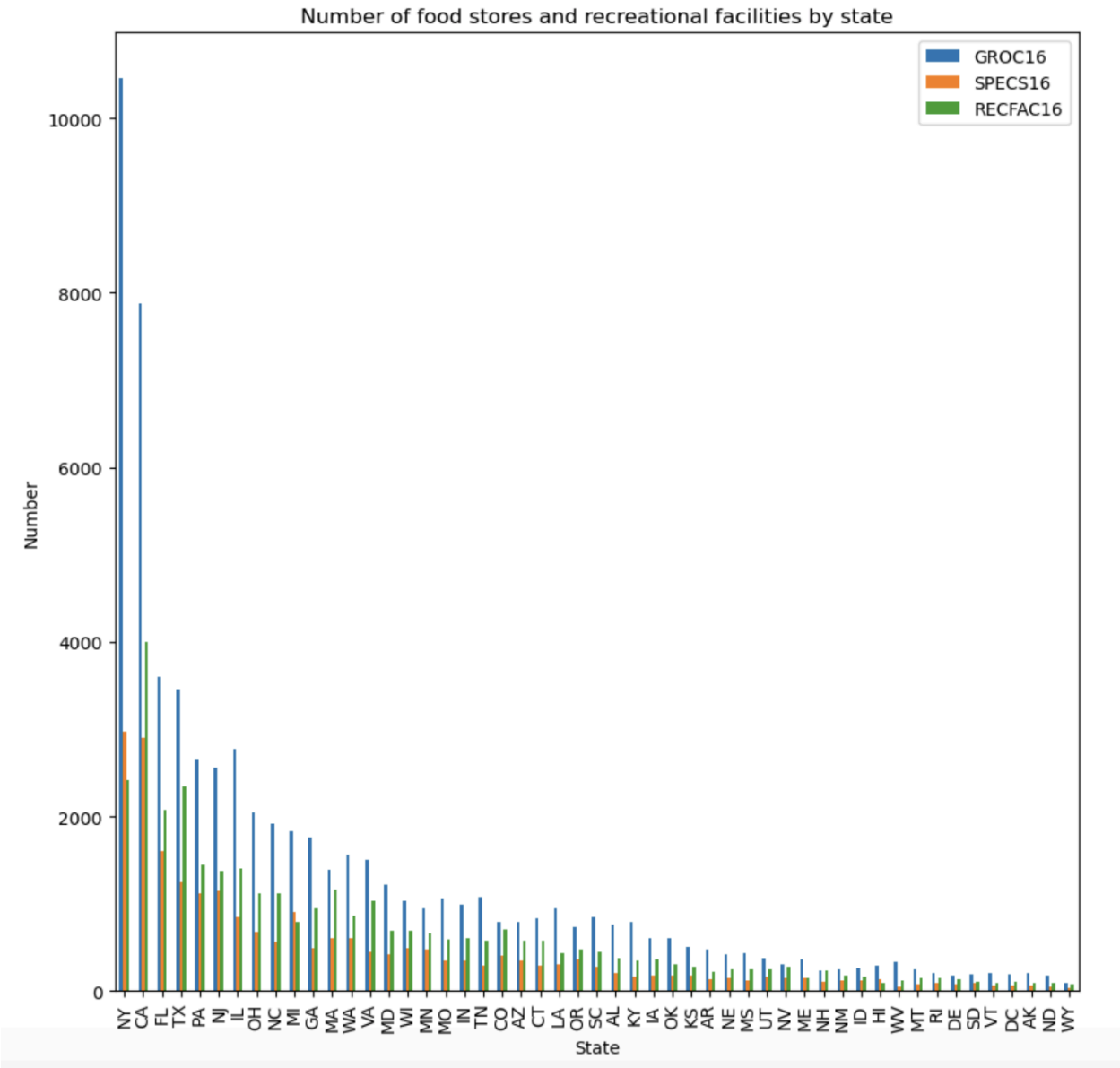
Results

Exploratory Data Analysis

Although my clustering methods did not provide strong answers for next steps, I discovered a few strong findings through Exploratory Data Analysis (EDA).



I began EDA by examining the correlation between all the variables in my dataset. From there, I plotted the number of grocery stores, number of specialty food stores, and number of fast food restaurants by state. These plots revealed that state population skews the visualizations. Each of these plots showed that states with larger populations have more stores/restaurants. This suggests that the availability of these facilities is driven, in part, by the population demands of each state.

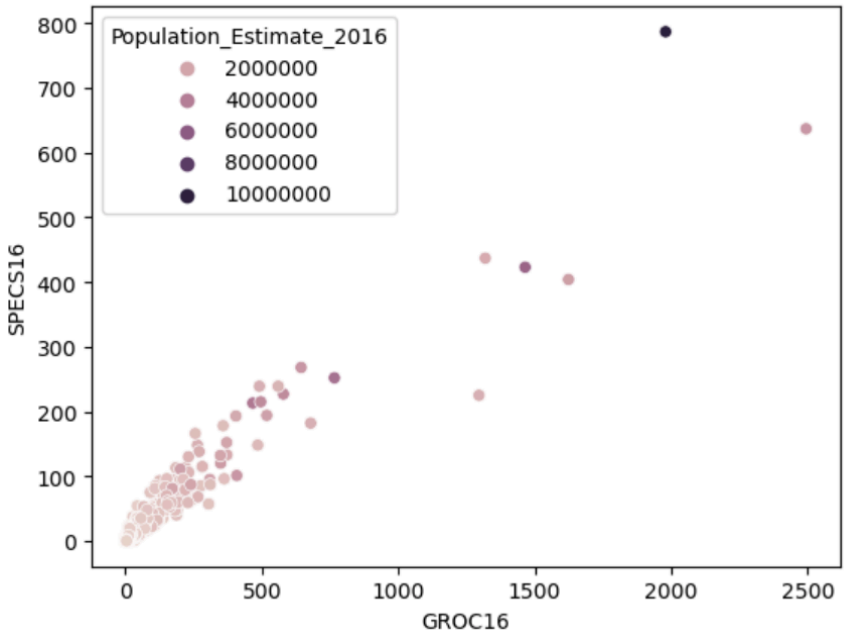


Then, I layered bar graphs into one plot to create a comparison between grocery store counts, specialty food store counts, and recreation facility counts by state. In every state, the number of grocery stores exceeds the number of specialty food stores and recreation centers. This trend is particularly pronounced in highly populated states like New York (NY) and California (CA), where the number of grocery stores far surpasses the other two categories.

While most states follow a pattern of grocery stores outnumbering both specialty stores and rec centers, some states such as Colorado (CO), Nevada (NV), New Hampshire (NH), and Wyoming (WY) show unique behavior. In these states, the number of recreation centers is close to the count of grocery stores, suggesting a higher emphasis on recreational infrastructure. However, this proximity does not extend to specialty food stores, which remain significantly fewer in number.

Three states—Hawaii (HI), Michigan (MI), and New York (NY)—are outliers, with more specialty food stores than recreation centers. Meanwhile, in Maine (ME) and South Dakota (SD), the counts of recreation centers and specialty food stores are relatively close, reflecting a more balanced distribution of these types of facilities.


```
<Axes: xlabel='GROC16', ylabel='SPECS16'>
```



The scatter plots I created all exhibited positive relationships between the variables, indicating that as one variable increased, the others tended to increase as well. Initially, I focused on

visualizing the relationships between different types of facilities (grocery stores, specialty food stores, and recreational centers). However, these plots did not provide particularly useful insights because the variables were likely highly correlated due to their shared association with state population. In other words, the raw counts of facilities were naturally influenced by population size, which masks any meaningful differences in infrastructure distribution.

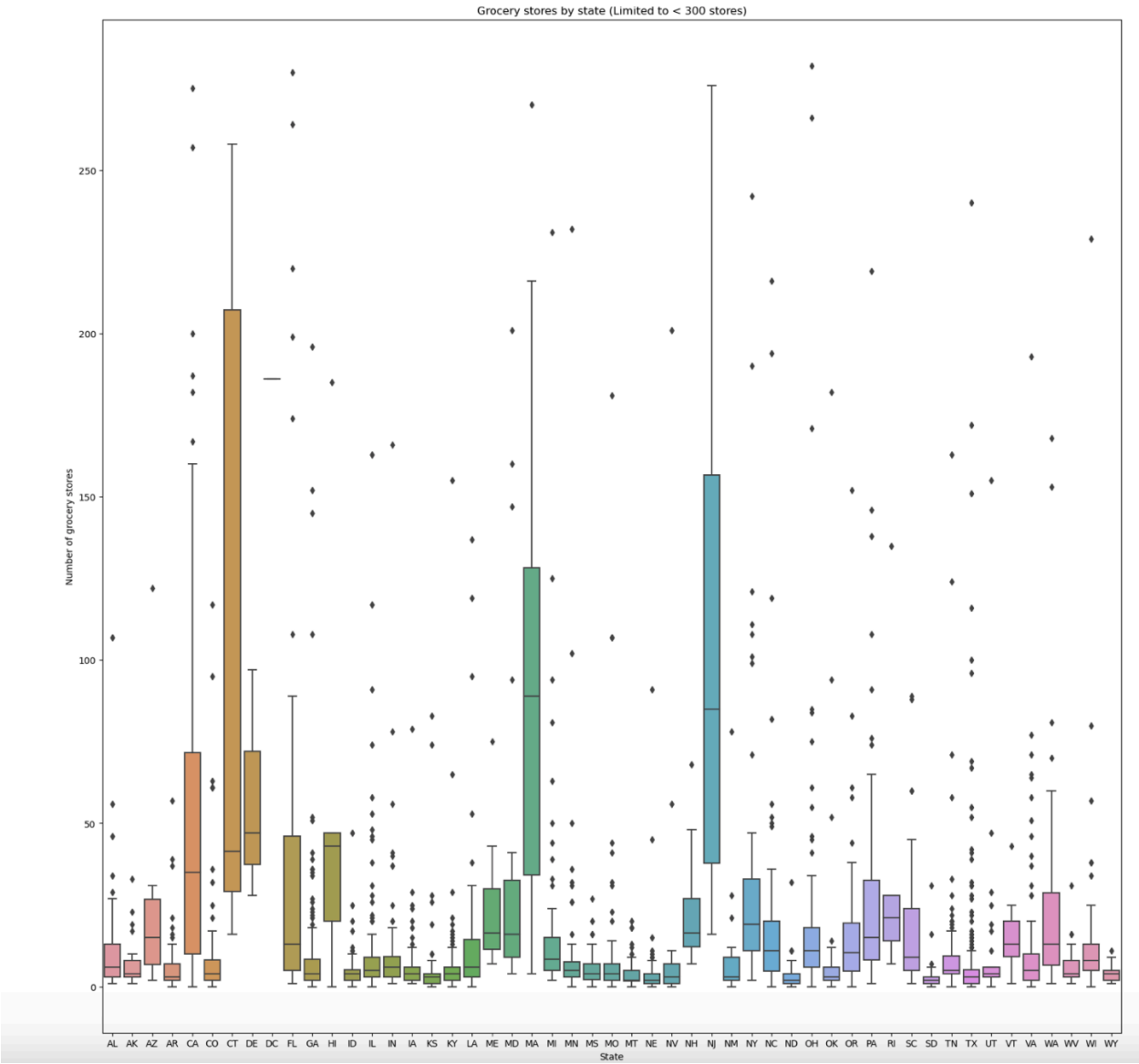
To address this, I introduced population as a key factor in the final scatter plot. By accounting for population, I was able to test and confirm my hypothesis: population size significantly influences the correlation between the counts of food stores and restaurants. This population-adjusted analysis reveals a clearer picture of how the density of facilities per state varies, not simply based on the number of stores, but relative to how large or small the state's population is.

The final visualizations I created were box plots, which allowed me to analyze the distribution of food store counts across states. Specifically, I wanted to examine how the number of stores—both grocery stores and specialty food stores—varied within each state. Box plots are particularly useful for visualizing the spread, median, and potential outliers in the data, offering a deeper understanding of the range and central tendency of store counts.

Upon analyzing the box plots, I observed that states on the East Coast, along with California, exhibit the largest spreads in terms of food store counts. This indicated significant variability in the number of stores across these states, suggesting that certain areas within these states may have had much higher concentrations of food stores than others. These wide ranges could be attributed to population density, economic factors, or geographic differences, such as urban versus rural areas, which affect the distribution of stores.

The larger spreads in states like California and those on the East Coast also pointed to potential disparities in food accessibility. For example, some regions may be densely packed with grocery stores and restaurants, while others may be underserved. Understanding these variations is critical for policymakers and businesses, as it highlights areas that may require additional infrastructure or support for food accessibility.

In contrast, states with smaller spreads tend to show more consistency in the availability of food stores across their regions. However, these ranges hovered closer to zero than states with wider interquartile ranges. While this might suggest more uniform development or less pronounced urban-rural divides, this could also suggest food insecurity is more widespread in these states.



Recommendations and Next Steps

Demographic analysis

I tried to incorporate demographic data, but I would recommend a greater emphasis on analyzing the impact of demographic variables on food security. I would also recommend incorporating data of different variable types for future research. It would be beneficial to incorporate categorical variables such as race, education level, and employment status. In my study, these variables were included as numerical variables, but they would be more beneficial as categorical data for future research.

County-level analysis

While I incorporated some demographic data, all the included variables were state-level. State-level data is great for country-wide comparisons; however, county-level comparisons are necessary to see variations within states. Since state-level data is affected by big cities (outliers), county-wide comparisons would provide a more accurate understanding of food insecurity within each state.

Infrastructure analysis

In future analyses, exploring the differences in infrastructure in conjunction with demographic data—such as income levels or population density—could offer valuable insights into the socioeconomic factors influencing food store distribution. Additionally, mapping out these trends spatially could highlight specific regions or communities that may be food deserts or areas of opportunity for expanding food-related infrastructure.