U.S. Farmers Markets and County Income Analysis

Kristof Toth

July 2, 2020

Abstract: Nutrition and health in the United States are closely related in the study of public health. Local farmers markets have proliferated as a means to distribute fresh produce directly to consumers, skipping the costly distribution and packaging step. However, one criticism of farmers markets is that they are largely inaccessible to many Americans, due to location, especially to those of low income and socio-economic status. This project draws upon all the registered farmers markets in the United States in 2020 as well as economic statistics for each county to investigate the correlation of income and access to farmers markets. This preliminary study used multiple regression modeling to find a positive correlation of the density of farmers markets in a county to the per capita income. However, there was a negative correlation of the density of farmers markets to the county's median household/family income. Overall, there was no clear trend that greater income in a given county affected the density of farmers markets. Furthermore, county segmentation by K-means found an optimal of 3 clusters of counties by the elbow method. Four counties (Maricopa County, AZ; Los Angeles County, CA; Harris County, TX; and Cook County, IL) were identified as great potential for new farmers markets due to low market density, high population and relatively high income.

1. Introduction

1.1 Background

Over the past century, public health is the United States has achieved tremendous success in increasing the longevity and productivity of life. At the same time, due in part of the changes in lifestyle behaviors, the rates of non-communicable diseases, specifically, chronic diet related diseases have risen. Some of these most common diseases include cardiovascular disease, high blood pressure, type 2 diabetes, some cancers, and poor bone health. Alarmingly, roughly half of all American adults (117 million individuals) have one or more preventable chronic diseases. [1] The impact of chronic disease is also felt economically, as in 2008, the medical costs associated with obesity were estimated to be $147 billion a year. In 2017, the total estimated cost of diagnosed diabetes was $327 billion in medical costs and loss of productivity. [2-4]

1.2 Problem

Having access to healthy, safe, and affordable food choices is crucial for an individual to live a healthy lifestyle. Food access is influenced by diverse factors, including proximity to food retail outlets (e.g., distance to a grocery store/supermarket/market or overall density of markets), individual resources (e.g., income to dictate spending), and neighborhood-level resources (e.g., average income of the neighborhood/county to subsidize and incentivize agricultural infrastructure). Innovative approaches to food access such as farmers markets, mobile markets, shelters, food banks and community gardens/cooperatives have emerged as an alternative and improvement to grocery stores. However, the limited location of farmers markets makes their accessibility restricted. It should be noted that race/ethnicity, socioeconomic status, and the presence of a disability also may affect an individual's ability to access foods to support healthy eating patterns. However, the in depth study of these other contributing factors is outside the scope of this project.

1.3 Interest

This project seeks to visualize and communicate the relationship between the location and density of farmers markets and the economic status of U.S. counties. Knowing which regions and counties are lacking in farmers markets is critical both from a public health perspective and from an economic opportunity perspective. Local municipalities and city councils may use this information to organize and petition for additional farmers markets to meet the needs of their constituents. From a purely business perspective, new vendors and farmers can find target areas lacking in competition in order to sell to underserved communities.

2. Data acquisition and cleaning

2.1 Data Sources

The data sources were uploaded to Kaggle by Madeleine Ferguson. [5] The 2020 farmers market data is maintained by the USDA Agricultural Marketing Service. [6] The farmers market listings include market locations, directions, operating times, product offerings, accepted forms of payment, and more. A farmers market is defined as two or more farm vendors selling agricultural products directly to customers at a common, recurrent physical location.

The list of United States counties by per capita income is from the 2009-2013 American Community Survey 5-Year Estimates; data for Puerto Rico is from the 2013-2017 American Community Survey 5-Year estimates, and data for the other U.S. territories is from the 2010 U.S. Census. [7-10] The data contains the name of the county (or equivalent), state or territory, per capita income, median household income, median family income, population, and number of households.

An alternative method was considered to obtain farmers market information from the Foursquare API using a free developer user. The resulting searches are limited to 50 responses and the maximum radius of search is 100,000 meters. Foursquare API has advantages to tabulate all the venues near a particular position, but is cumbersome for acquiring clean data such as all farmers markets in the United States. Thus, farmers market data was taken from ams.usda.gov.

2.2 Data Cleaning

After initial analysis of the Farmers Market Data, there are 36 rows with NaN for "County" and 28 rows with NaN for the longitude and latitude coordinates. Looking into some of the NaN values reveals that they correspond to mobile farmers markets which may be located in multiple counties and cities depending on time. I exclude these in the analysis. The 28 instances of unavailable coordinates represents only 0.3% of the total dataset (8804 farmers markets) so it could be ignored without significant change in the overall analysis.

For the U.S. Economical County Data, I cleaned up the data by excluding two rows which don't contain any data. There also appeared to be duplicates of New Jersey, Maryland, and Puerto Rico as there are three rows which contains the sum values for the entire state/territory (NJ, MD, and PR had separate rows for each county data).

The 89 permanently-inhabited county-equivalents in the territories of the United States (such as the municipalities of Puerto Rico) are also listed (but are not ranked in the dataset). The dataset excludes the 8 county-equivalents in the U.S. territories that have zero people (Baker Island, Howland Island, Jarvis Island, Johnston Atoll, Kingman Reef, Navassa Island, Northern Islands Municipality and Rose Atoll). The 3 semi-populated county-equivalents in the U.S. Minor Outlying Islands (Midway Atoll, Palmyra Atoll and Wake Island) are also excluded.

2.3 Initial Economic Data Visualization

For the economic data for each U.S. county, initial boxplots were used to visualize the distribution of the per capita income, median household income, and median family income. Each state/territory had a single boxplot which highlighted the median, interquartile range, min (Q1-1.5*IQR) and max (Q3 + 1.5*IQR), and outliers counties. The states may be ordered in several ways, two of which are:

- In order of highest county in each state.
- In order of the highest median for each state.

I will show both for the dataset of per capita income only. The median household income and median family income visuals can be found in the notebook. A geographical map of the U.S. counties is also provided.

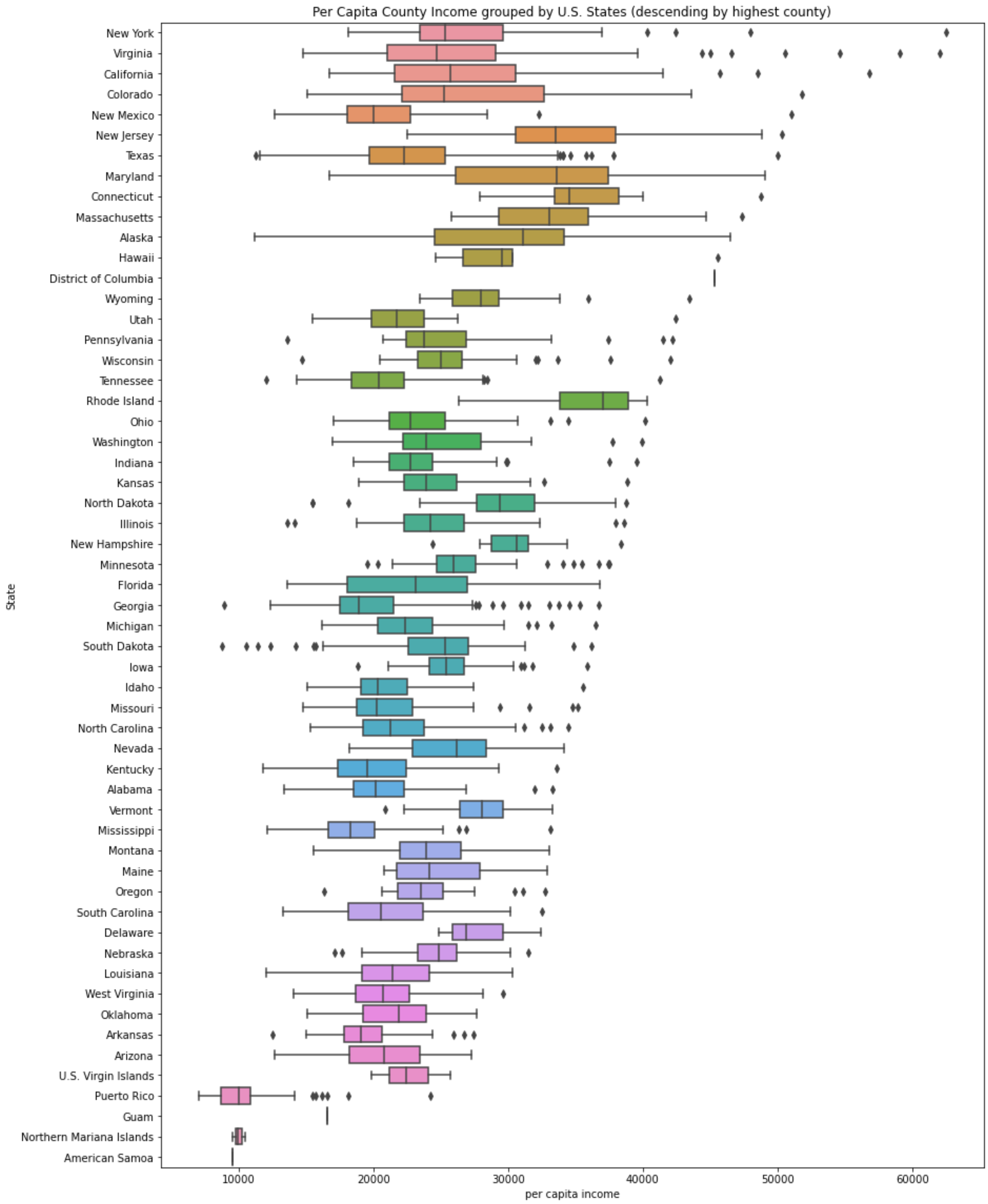


Figure 1: Boxplot of each U.S. state/territory based on Per Capita Income for each county. The order of the states is by descending of highest county.

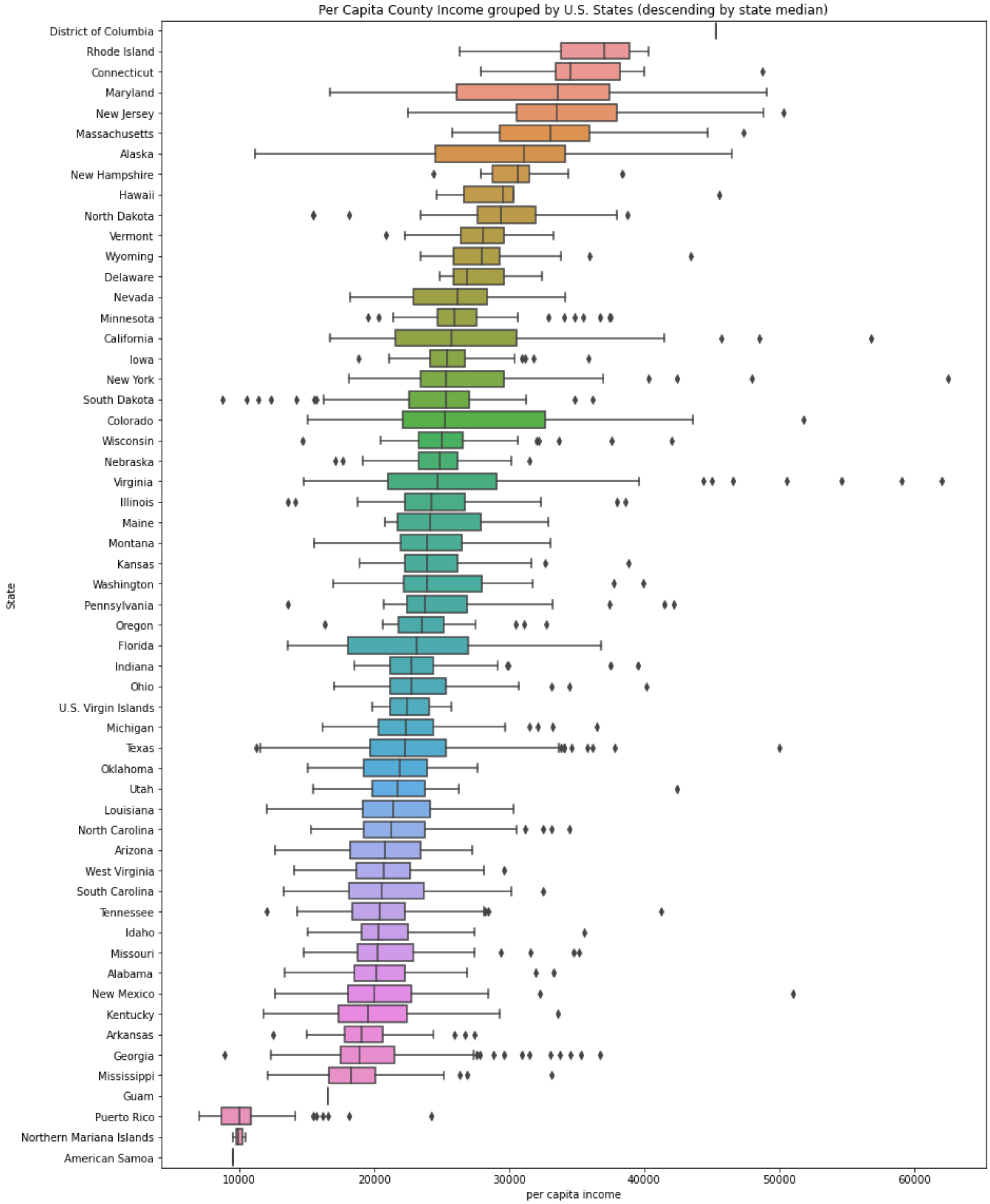


Figure 2: Boxplot of each U.S. state/territory based on Per Capita Income for each county. The order of the states is by descending of state median.

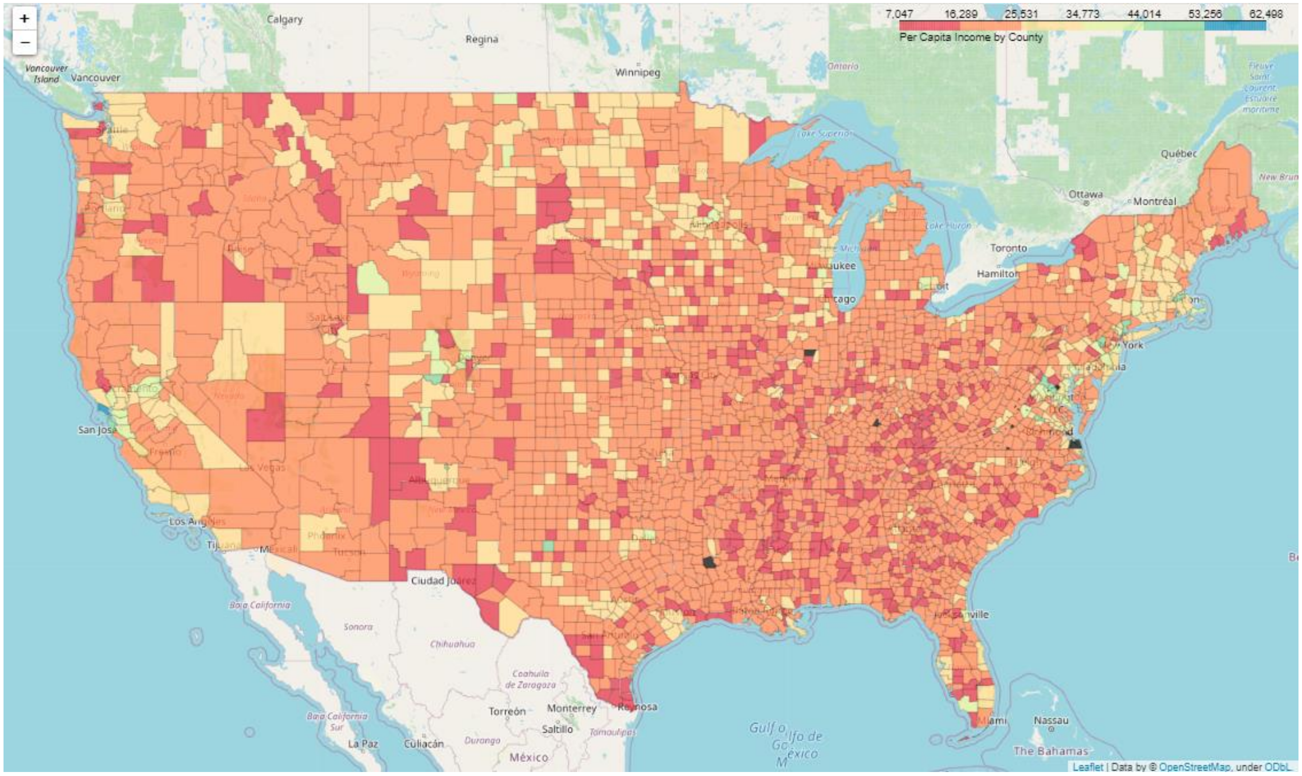


Figure 3: Geographical choropleth map of the U.S. counties color coded by the Per Capita Income. Low is red while high is blue.

2.4 Initial Farmers Market Data Visualization

The dataset of farmers markets was plotted using the python package Folium. The shapes of the U.S. states and counties were extracted from respective json files. Note that the Derwood Farmers Market location information was corrected to (39.126442, -77.150267) since the original entry of (1.7081209, -3.4606929) was incorrect.

In the dataset containing the names of U.S. counties, the phrase “County” or “Census Area” was removed to be compatible with the json file classifications. For the initial rough visualization, duplicates of county names in multiple states were largely ignored (since I used latitude and longitudinal value), but this was addressed in the analysis and regression modeling. As described in the methodology section, the “County” and “State” labels were combined to create a unique “County-State” identifier so that duplicate names can be handled. Otherwise, the summation across all farmers markets in a county would add counties with the same name but different state!

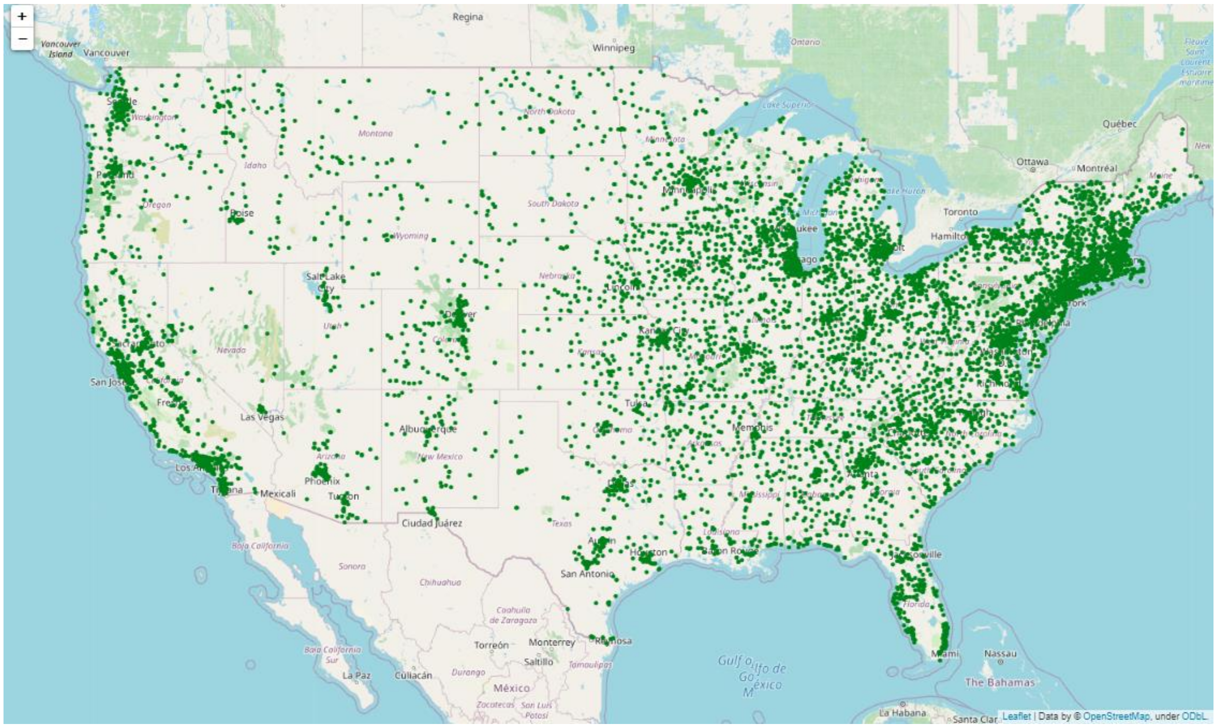


Figure 4: Visualization of each farmers market on the U.S. mainland.

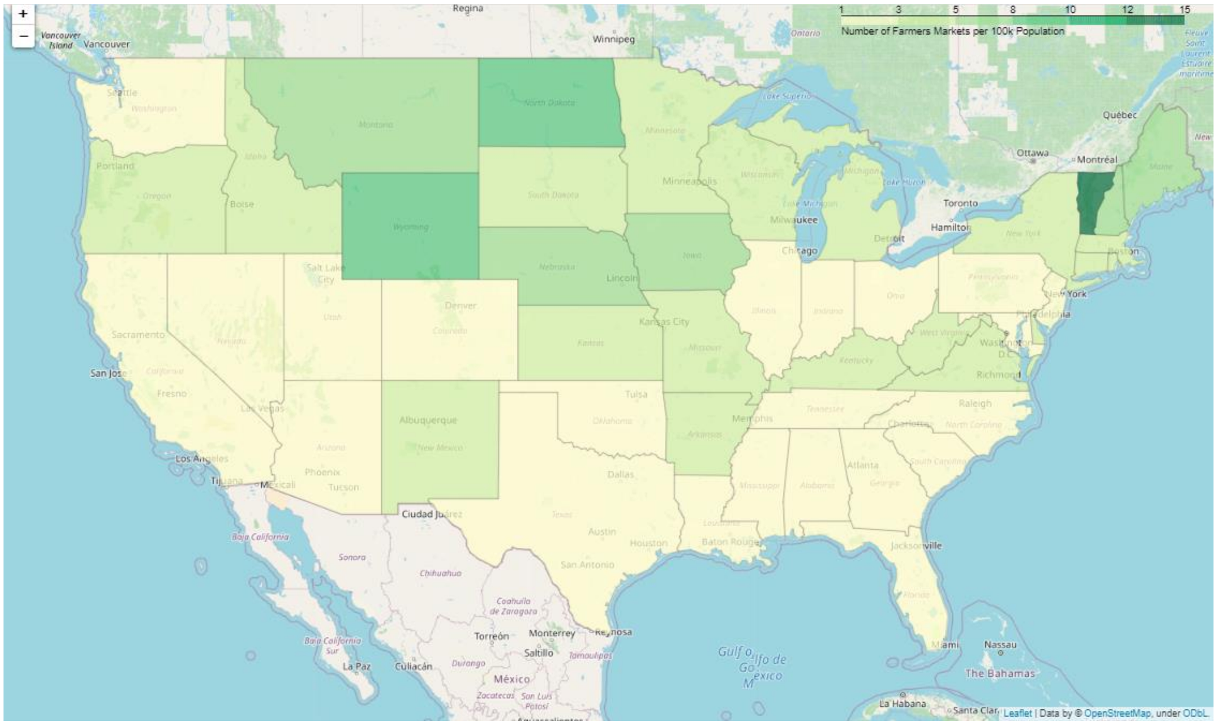

+
−
1
3
5
8
10
12
15
Number of Farmers Markets per 100k Population
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 5: Choropleth map visualization of the U.S. showing density of farmers markets in each state per 100k population.

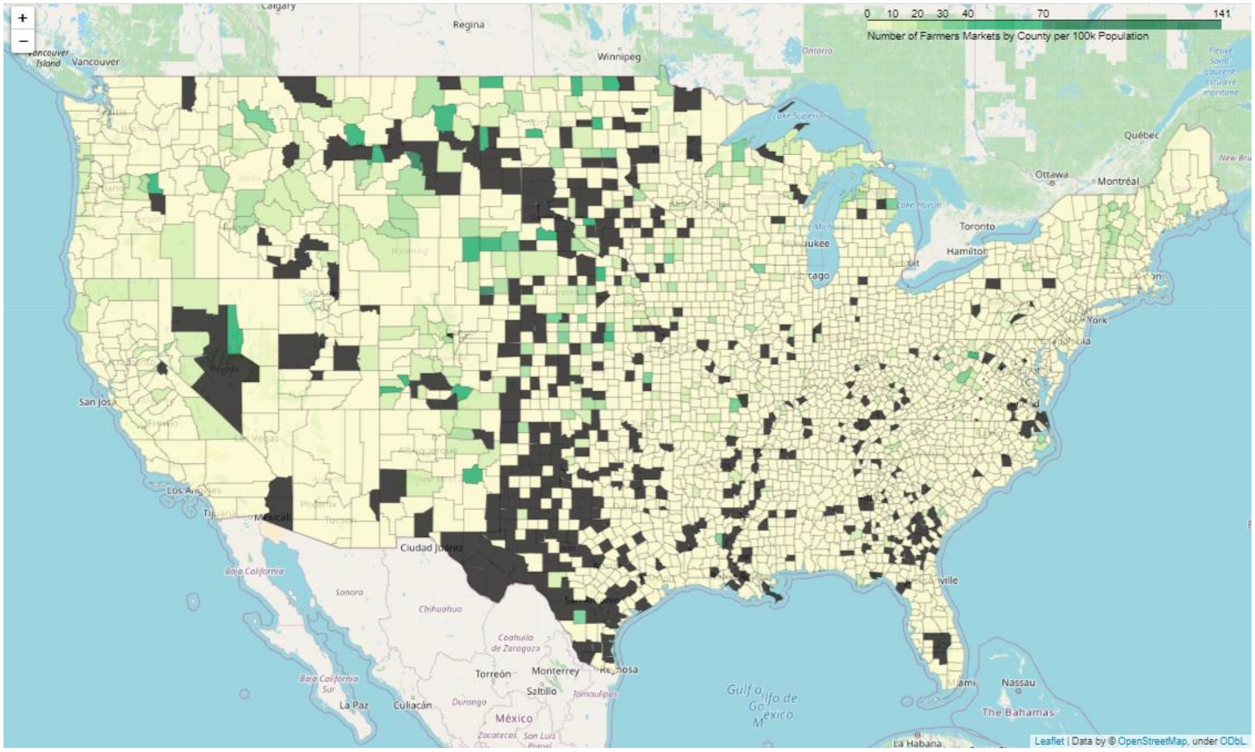
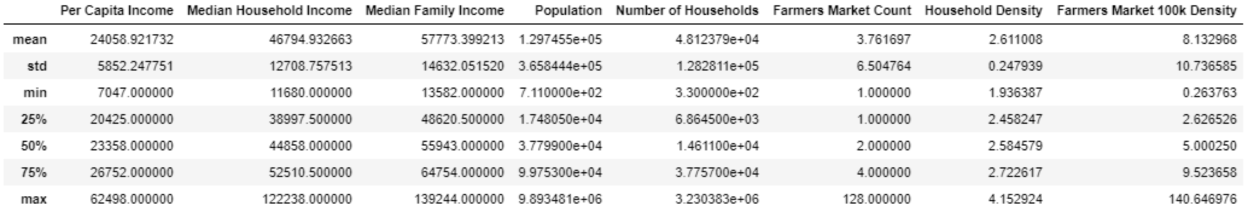


Figure 6: Choropleth map visualization of the U.S. showing density of farmers markets in each county per 100k population. Counties in dark did not have permanent farmers markets in the cleaned dataset. The color scale is skewed to show better distribution of farmers market density at the lower end of the range.

3. Methodology

The two datasets were combined to a single pandas dataframe by using the county names as a common key. Since generic county name can exist in multiple states, a "County-State" column was created in each dataset which contained the county and the state name. Washington county, for example, appears in 31 states! The combined and cleaned dataframe was then used for multiple regression modeling as well as county segmentation with K-Means method.



	Per Capita Income	Median Household Income	Median Family Income	Population	Number of Households	Farmers Market Count	Household Density	Farmers Market 100k Density
mean	24058.921732	46794.932663	57773.399213	1.297455e+05	4.812379e+04	3.761697	2.611008	8.132968
std	5852.247751	12708.757513	14632.051520	3.658444e+05	1.282811e+05	6.504764	0.247939	10.736585
min	7047.000000	11680.000000	13582.000000	7.110000e+02	3.300000e+02	1.000000	1.936387	0.263763
25%	20425.000000	38997.500000	48620.500000	1.748050e+04	6.864500e+03	1.000000	2.458247	2.626526
50%	23358.000000	44858.000000	55943.000000	3.779900e+04	1.461100e+04	2.000000	2.584579	5.000250
75%	26752.000000	52510.500000	64754.000000	9.975300e+04	3.775700e+04	4.000000	2.722617	9.523658
max	62498.000000	122238.000000	139244.000000	9.893481e+06	3.230383e+06	128.000000	4.152924	140.646976

Figure 7: Combined farmers market and economic data for each U.S. county.

The final columns in the combined farmers market and economic dataframe were “Per Capita Income”, “Median Household Income”, “Median Family Income”, “Population”, “Number of Households”, “Farmers Market Count”, “Household Density” (Population/Number of Households), “Farmers Market 100k Density” (Farmers Markets per 100k population). The statistical breakdown in shown in **Figure 7**.

The Farmers Market Density per 100k population was chosen as the main metric of analysis, because a simple count of the number of farmers markets in a county is not a clear representation of accessibility. Rather, the per capita value of farmers markets is a more accurate measure of how many an average person in a county had access to.

3.1 Multiple Regression Modeling

The Scikit-learn library was used to perform multiple regression modeling on the combined farmers market and economic data for each U.S. county. Scikit-learn uses plain Ordinary Least Squares method. The independent variables were ‘Per Capita Income’, ‘Median Household Income’, ‘Median Family Income’, ‘Population’, and ‘Number of Households’. The dependent variable (one to predict), was ‘Farmers Market 100k Density’.

3.2 County Segmentation with K-Means

Unsupervised Machine Learning was performed on the combined farmers market and economic dataframe to cluster the counties. I dropped the columns with discrete variables as the Euclidean distance function isn't meaningful when clustering. I also dropped the ‘Number of Households’, ‘Household Density’, and ‘Farmers Market Count’ as they are duplicates of already presented columns in the dataframe. I used StandardScaler function from Scikit-learn to normalize the dataset.

The K-Means algorithm need an input for number of clusters (K). The KElbowVisualizer from Yellowbrick was used to implement the “Elbow” method to select the optimal number of clusters by fitting the model with a range of values for K. The optimal K value occurs at the inflection on the curve and is shown with a dashed line.

4. Results

4.1 Multiple Regression Modeling

Multiple regression modeling plain Ordinary Least Squares method was performed to determine the impact of various independent variables on the ‘Farmers Market 100k Density’.

Independent Variables	Regression Coefficient
Per Capita Income	1.5e-03
Median Household Income	-5.7e-04
Median Family Income	-7.3e-05
Population	6.9 e-05
Number of Households	-2.1e-04

Residual sum of squares was 98.42 and Variance score was 0.15.

These results for the coefficients imply that increasing the per capita income increases the farmers’ market density. However, surprisingly, the farmers market density decreases for larger median household or family income. This suggests that economic indicators are not trivial for how many farmers’ markets are in a county (normalized by population).

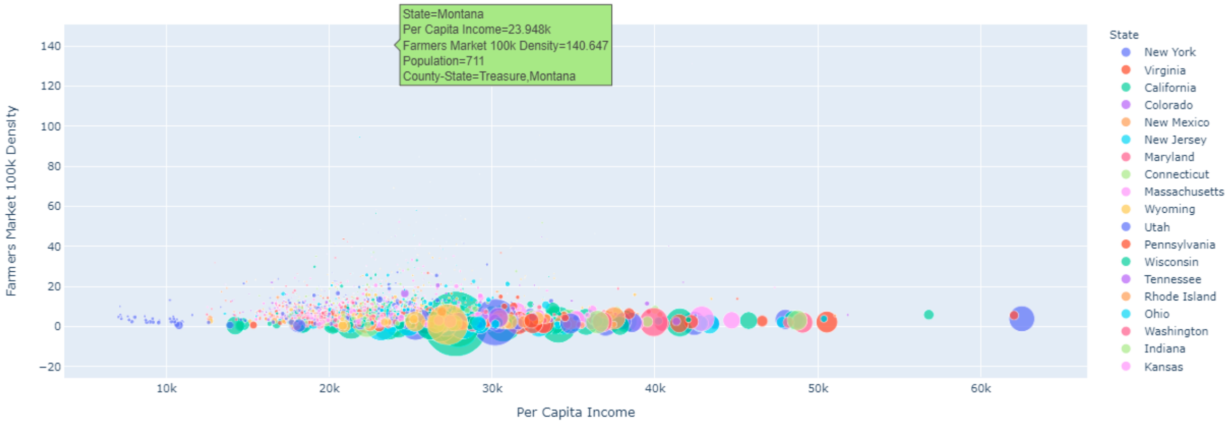


Figure 8: Farmers Market 100k Density as a function of Per Capita Income. Treasure County, Montana is an outlier at a density of 140 since it has 1 farmers market but a population of only 711. The size of each county bubble shows population.

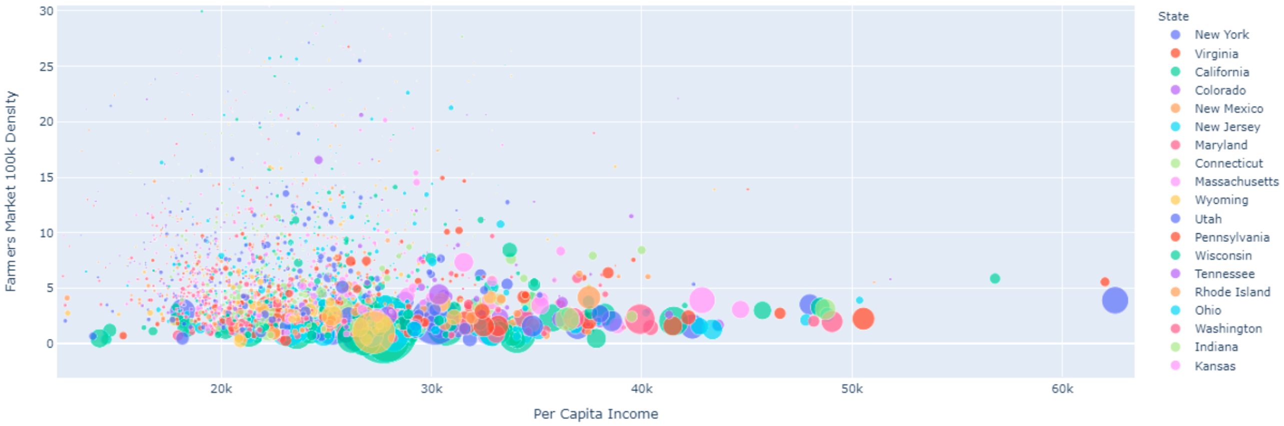


Figure 9: Farmers Market 100k Density as a function of Per Capita Income. The size of each bubble is population.

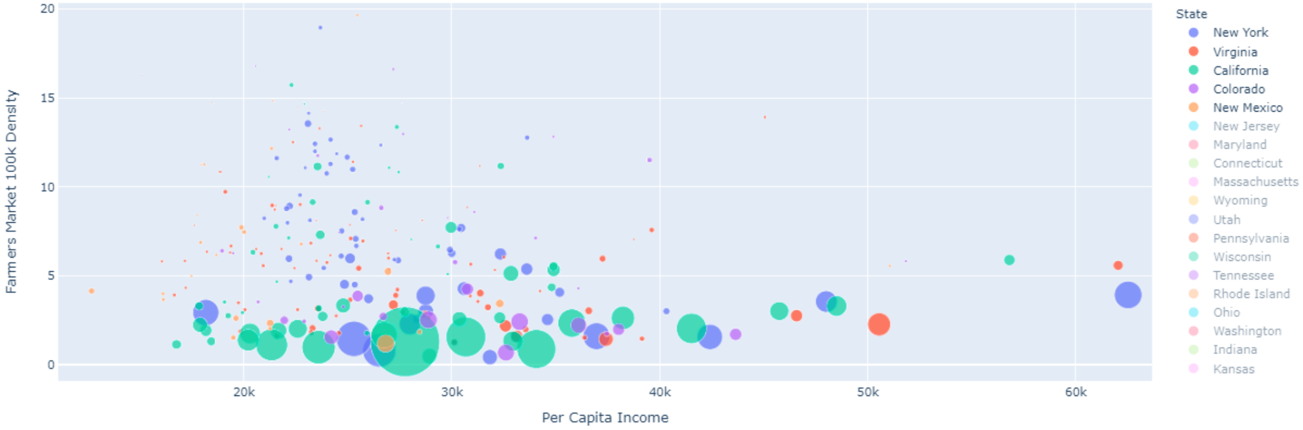


Figure 10: Farmers Market 100k Density as a function of Per Capita Income. The size of each bubble is population. Plot shows only 5 states: NY, VA, CA, CO, and NM.

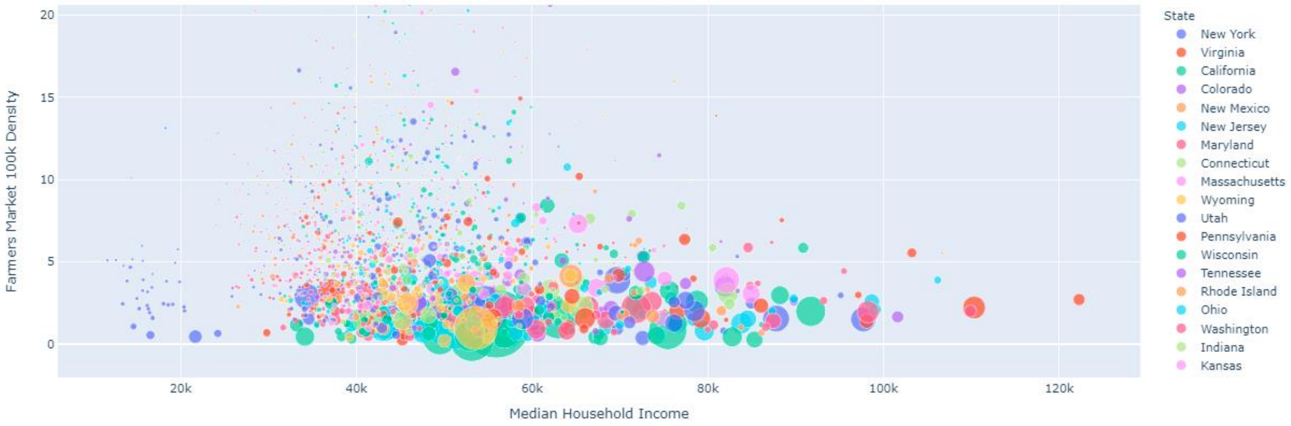


Figure 11: Farmers Market 100k Density as a function of Median Household Income. The size of each bubble is population.

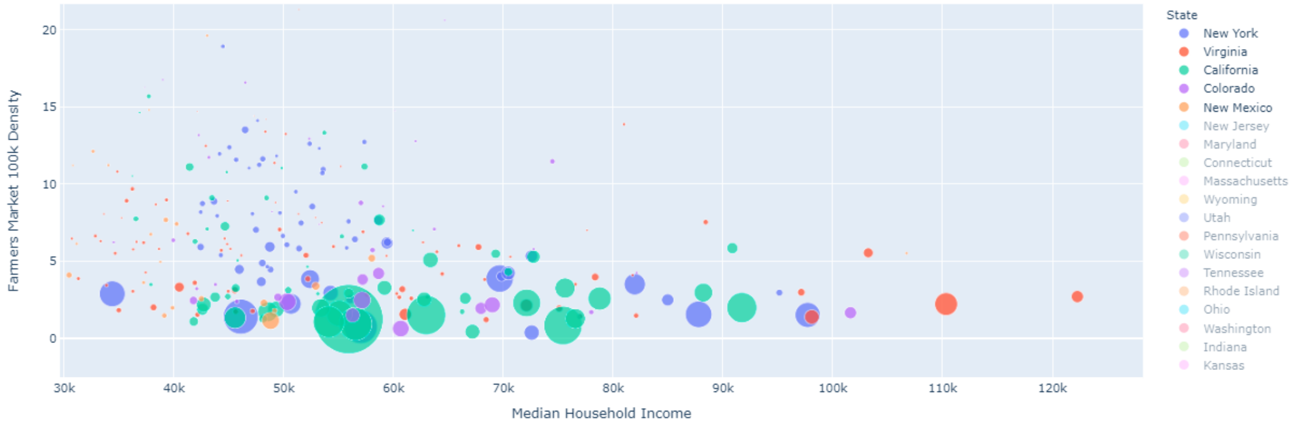


Figure 12: Farmers Market 100k Density as a function of Median Household Income. The size of each bubble is population. Plot shows only 5 states: NY, VA, CA, CO, and NM.

From an initial visual inspection, the density of farmers markets does not seem to be influenced by the income level of the county. Rather, the distribution was fairly linear across both per capita income and median household income.

4.2 County Segmentation with K-Means

Unsupervised Machine Learning was performed on the combined farmers market and economic dataframe to cluster the counties. The KElbowVisualizer from Yellowbrick was used to implement the "Elbow" method to select the optimal number of clusters by fitting the model with a range of values for K. The optimal K value occurs at the inflection on the curve and is shown with a line.

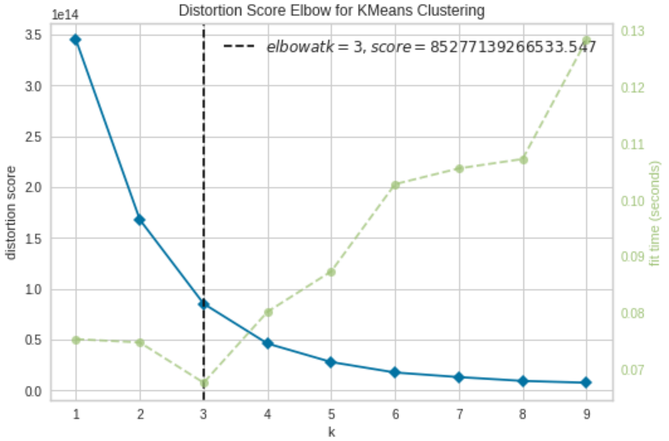
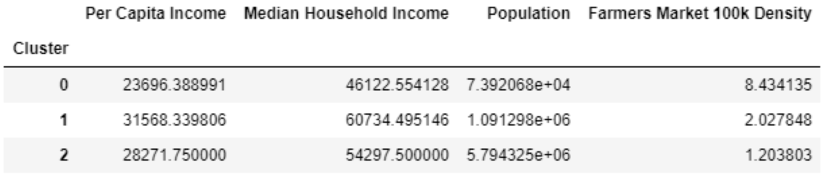


Figure 13: Distortion Score Elbow for K-Means Clustering

The Elbow method showed that 3 should be used for the K-Means clustering. The three clusters obtained by the method yielded the following statistics:



Cluster	Per Capita Income	Median Household Income	Population	Farmers Market 100k Density
0	23696.388991	46122.554128	7.392068e+04	8.434135
1	31568.339806	60734.495146	1.091298e+06	2.027848
2	28271.750000	54297.500000	5.794325e+06	1.203803

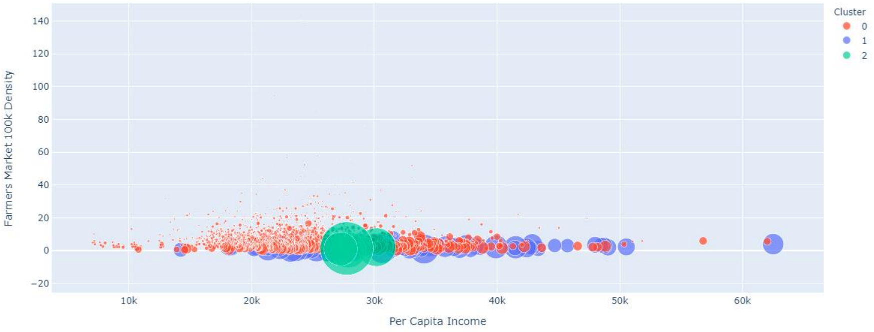


Figure 14: Scatter plot for K-Means clusters for K=3. Farmers Market 100k Density as a function of Per Capita Income data. Size is population.

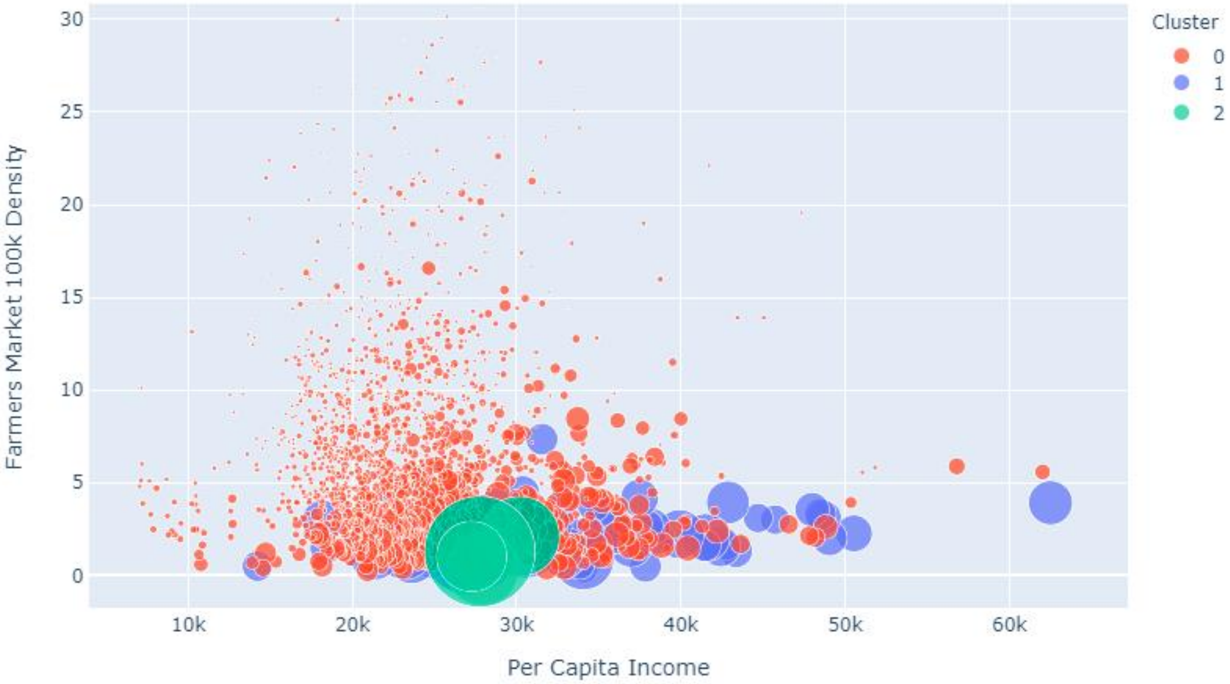


Figure 15: Scatter plot for K-Means clusters for K=3. Zoomed in region of the scatter plot excluding outliers. Farmers Market 100k Density as a function of Per Capita Income data. Bubble size ~ population.

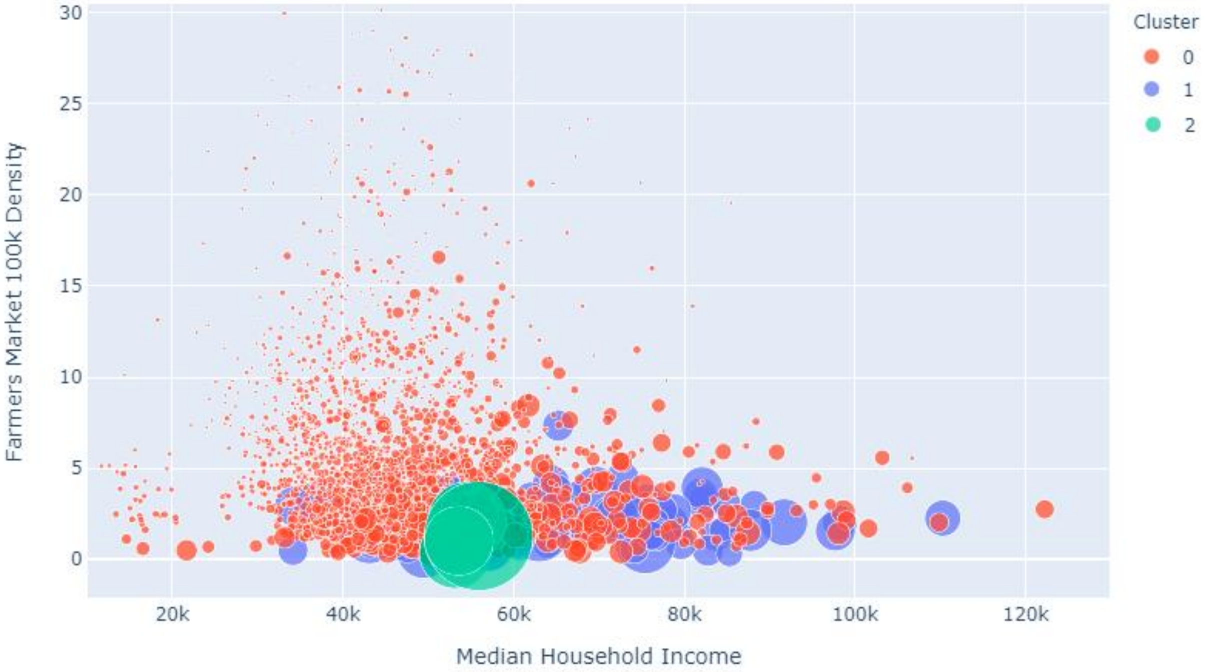


Figure 16: Scatter plot for K-Means clusters for K=3. Farmers Market 100k Density as a function of Median Household Income data. Bubble size ~ population.

5. Discussion

Overall, per capita income in the U.S. territories tends to be lower than per capita income in the 50 states and District of Columbia. Excluding the uninhabited county-equivalents, the county or county-equivalent with the highest per capita income is New York County, New York (Manhattan) ($62,498), and the county or county-equivalent with the lowest per capita income is Manu'a District, American Samoa ($5,441). Puerto Rico is also much poorer in comparison to the 50 states.

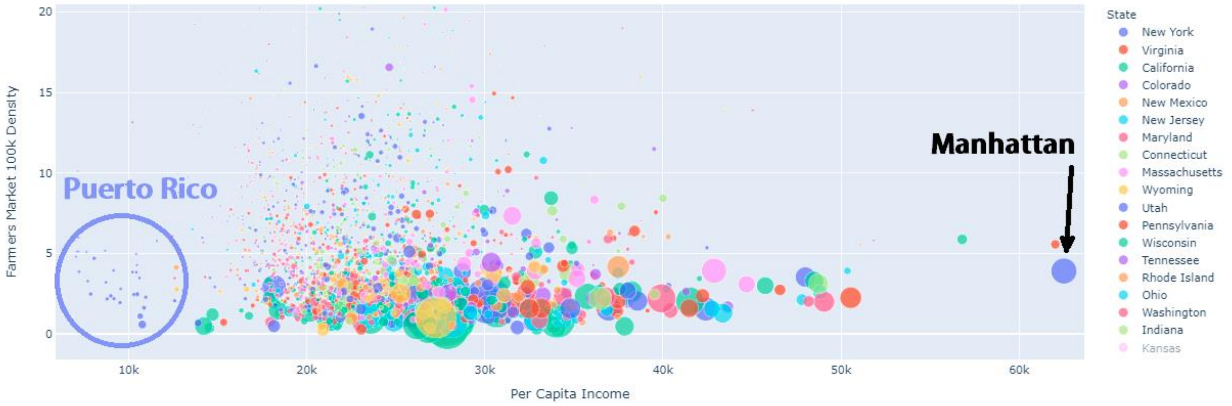


Figure 17: Farmers Market 100k Density as a function of Per Capita Income. The size of each bubble is population. Highest per capita income is New York County (Manhattan). Lowest is Manu'a District and Puerto Rico as a whole territory has less per capita income as the 50 states.

The U.S. Census Bureau defines a family as two or more people related by birth, marriage, or adoption residing in the same housing unit. A household consists of all people who occupy a housing unit regardless of relationship. A household may consist of a person living alone or multiple unrelated individuals or families living together. Median family income is typically higher than median household income because of the composition of households. Family households tend to have more people, and more of those members are in their prime earning years; as contrasted with members who have lesser incomes because they are very young or elderly. Areas with a wide disparity between the two measures have an excess of nonfamily households: single persons or otherwise.

K-means partitioned the counties into 3 mutually exclusive groups. The counties in each cluster are similar to each other demographically and economically. Now I can create a profile for each group, considering the common characteristics of each cluster. The 3 clusters are:

- Less dense population counties (~73,900), with a median household income of $46,122 and Per Capita Income of $23,696. High density of farmers markets (8.4/100k population). These are rural areas in the U.S., which while lower in economic status, are high in farmers markets.
- Medium dense population counties (~109,000), with a median household income of $60,734 and Per Capita Income of $31,568. Medium density of farmers markets (2/100k population).
- Highly dense population counties (~579,000), with a median household income of $54,297 and Per Capita Income of $28,271. Low density of farmers markets (1.2/100k population).

There are only 4 counties designated as #2 cluster: Maricopa County, Arizona; Los Angeles County, California; Harris County, Texas; and Cook County, Illinois. These would be the best places for new farmers markets to open because of the low density of existing markets but relatively high income and population!

When looking at the scatter plot of Per Capita Income as a function of Median Household Income, there is a clear linear trend as expected. For a richer county, if Per Capita Income is higher, I would expect also a higher Median Household (or Family) Income. In the following plot the bubble size is the farmers market density. If this density was strongly correlated to the income, I would see the bubbles increase or decrease linearly as well.

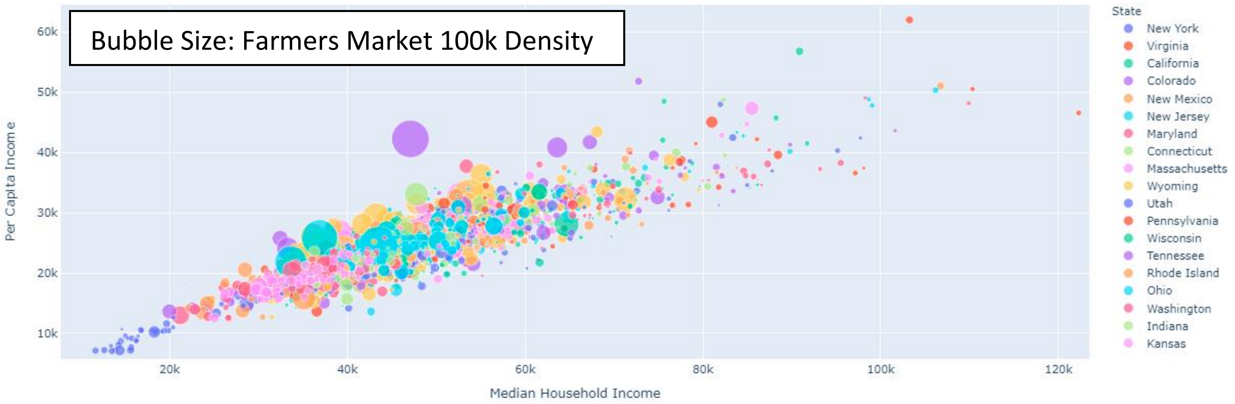


Figure 18: Scatter plot of Per Capita Income as a function of Median Houshold Income. The bubble size here is Farmers Market 100k Density.

6. Conclusion

Using multiple regression modeling and unsupervised machine learning of K-Means clustering, I analyzed the relationship of the density of farmers markets and economic income indicators in the United States by county. Four counties (Maricopa County, AZ; Los Angeles County, CA; Harris County, TX; and Cook County, IL) were identified as serving great potential for new farmers markets because of the low farmers market density, high population and relatively high median income. It should be mentioned that an individual’s access to a farmers market depends on more factors, such as individual income and means of transportation. While a county may be above average in economic terms, multiple cities and towns are located in each county. Even individual cities have income inequality between neighborhoods. Also, this report did not access the size of the farmers markets and availability of specific dietary foods. Yet, it is reassuring, nevertheless, to find no substantial evidence on the county level that farmers markets are disproportionately distributed by per capita and median income.

7. References

[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. Available at https://health.gov/our-work/food-and-nutrition/2015-2020-dietary-guidelines/.

[2] Centers for Disease Control and Prevention (CDC). About Chronic Diseases. October 23, 2019. Available at https://www.cdc.gov/chronicdisease/about/index.htm. Accessed June 23, 2018.

[3] American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. Diabetes Care 2018;41(5):917-928. PubMed abstract

[4] Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer- and service-specific estimates. Health Aff 2009;28(5):w822-31. PubMed abstract

[5] https://www.kaggle.com/madeleineferguson/farmers-markets-in-the-united-states

[6] https://www.ams.usda.gov/local-food-directories/farmersmarkets

[7] "SELECTED ECONOMIC CHARACTERISTICS 2009-2013 American Community Survey 5-Year Estimates". U.S. Census Bureau. Archived from the original on 2015-01-17. Retrieved 2015-01-12.

[8] "ACS DEMOGRAPHIC AND HOUSING ESTIMATES 2009-2013 American Community Survey 5-Year Estimates". U.S. Census Bureau. Archived from the original on 2015-01-05. Retrieved 2015-01-12.

[9] "HOUSEHOLDS AND FAMILIES 2009-2013 American Community Survey 5-Year Estimates". U.S. Census Bureau. Archived from the original on 2020-02-12. Retrieved 2015-01-12.

[10] U.S. Census Bureau: American FactFinder. 2013-2017 American Community Survey 5-Year Estimates (Puerto Rico) and "Profile of selected economic characteristics: 2010" (American Samoa / Guam / Northern Mariana Islands / U.S. Virgin Islands).