IST 652
Final Project

Mike DeMaria
mjdemari@syr.edu

Project:
Food prices are rising and farmer markets can be a good source for cheaper produce. I would like to see if the prevalence of farmers markets in New York State correlate with population density, income levels or if they correlate with farmland acreage.

Data Source:
Several sources of data was necessary for this project. Data on income, farmland, farmers markets and population density had to be retrieved and correlated together. None of this data was available from one source, let alone one agency. Instead, 4 distinct data sets and 4 organizations were involved.

The first source of data retrieved was a list of farmers markets in New York, published by the Division of Agricultural Development at https://data.ny.gov/Economic-Development/Farmers-Markets-in-New-York-State/qq4h-8p86. There are 725 rows and 20 columns in this data set. Each row represents one farmer's market.

The next source of data was the agricultural districts acreage total by county, published by the Division of Land and Water Resources at https://data.ny.gov/Economic-Development/Agricultural-Districts-Acreage-Totals-by-County/h2id-x25a. There are 53 rows and 3 columns in this data set. Each row represents one county.

The average income and tax liability of full year residents was obtained from the Department of Taxation and Finance at https://data.ny.gov/Government-Finance/Average-Income-and-Tax-Liability-of-Full-Year-Resi/2w9v-ejxd. This data set contains 1,072 rows and 8 columns. Each row represents one tax year for one county.

The final source of data was the population, land area and population density by county, obtained from the Department of Health at https://www.health.ny.gov/statistics/vital_statistics/2018/table02.htm. This data set contains 53 rows and 7 columns. Each row represents one county.

All data, except for population density, was obtained via a web service using the Socrata Open Data API in JSON format. The population density was copied off the webpage and placed into a comma separated file.

Data Cleanup:
The records obtained from New York's open data website are generally very clean, and this project has been no exception. The first thing I did was cast all the columns to the appropriate data types (string, float, etc) and standardized capitalizations. The counties were all in a mixture of lowercase, mixed case or uppercase. I converted all to lowercase so that they'll match up. Any rows with null values were discarded.

For land area, the data included total land that could be operated as farmland, and total land that was actually being used as farms. The total arable land is naturally greater than or equal to the total farmed land. As farmers markets are intended for the selling of produce, we only looked at farmed acres for this analysis.

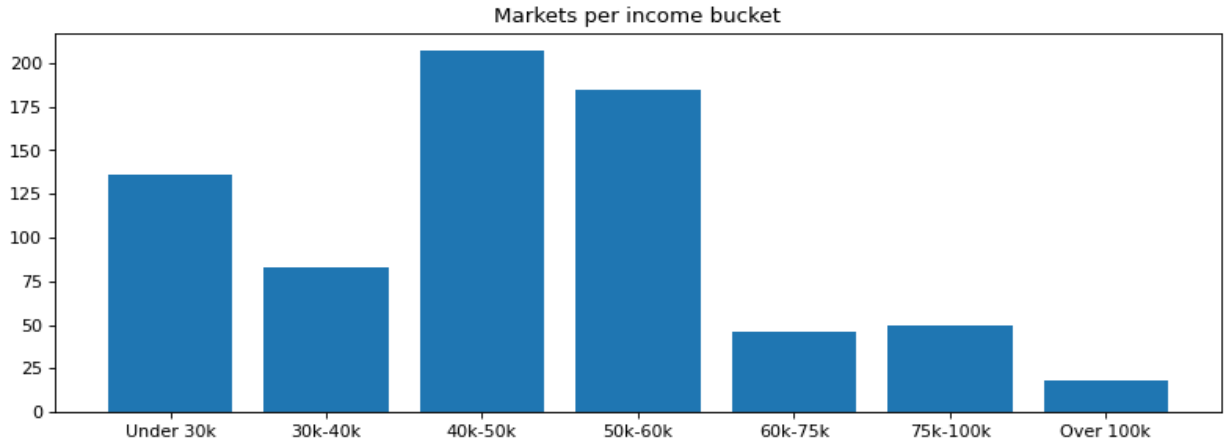
The tax data contained fields indicating the adjusted gross income of all returns and the adjusted gross income of taxable returns. The taxable returns only includes tax payers who pay New York income tax; low income earners will not fall into this category as they earn less than the minimum taxable amount. The 'all' returns will reflect a lower AGI than the 'taxable' returns. As we are trying to find a correlation between income and farmers market availability, the 'all' returns option was used.

Finally, the population density was measured in square miles, while farmland was measured in acres. There are 640 acres in a square mile. To normalize the units, the population density was multiplied by 640 to find the density per acre. Mathematically, this works out to the same ratio, just with different units.

Data Aggregation:
The farmers market data set is the largest of the group, showing the zip code and county for each market. However, all of the other data sets are aggregated at the county level only. Given the commonality of the county amongst the 4 sources, this formed a natural key value.

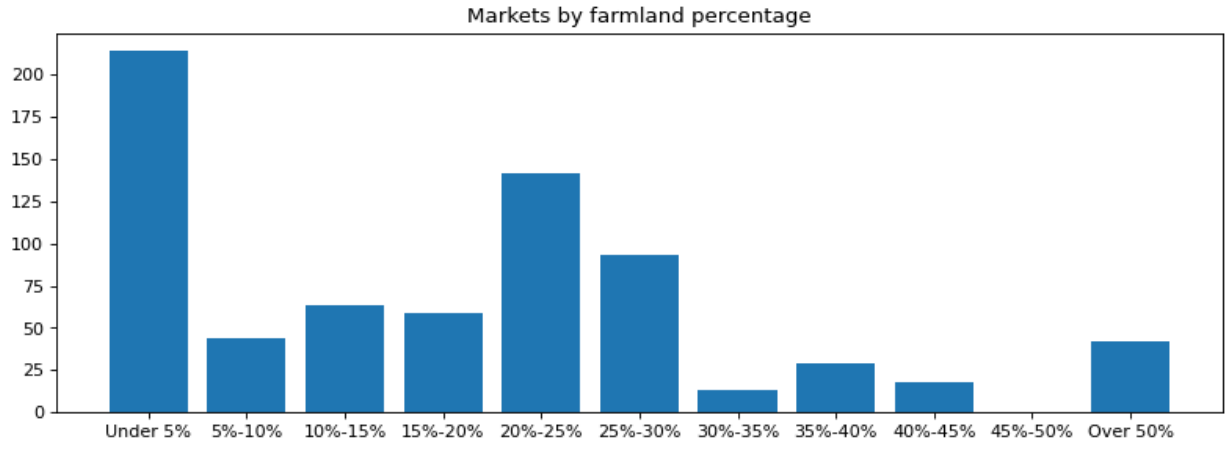
A large dataframe was formed, centered around the county, containing the average income level (AGI), farmland in acres, total county area, population size, population density, number of farmers markets, number of farmers markets that accept/do not accept SNAP (supplemental nutritional assistance program) and the percentage of the county that is farmland. Most of these values could be obtained from the various data sets, or derived by comparing the columns. Some columns did not have complete data. For example, the Bronx did not have any income statistics, but had population density. These incomplete rows were relatively small.

Initial Analysis
The primary hypothesis is that there is a correlation between income and farmers markets, so that is where the data analysis started. Income was grouped into buckets, mostly based around the buckets found in the Income Tax Components of Full-Year Residents by Size of Income and County at https://data.ny.gov/Government-Finance/Income-Tax-Components-of-Full-Year-Residents-by-Si/5kgr-h5g5. This produced a rather uneven bar chart. We could see larger amounts of markets for the $0-$30k range, and for $40k - $60k. There were very few at the higher levels. This inferred that, if there was a correlation, then farmers markets are more likely to skew towards lower income areas.



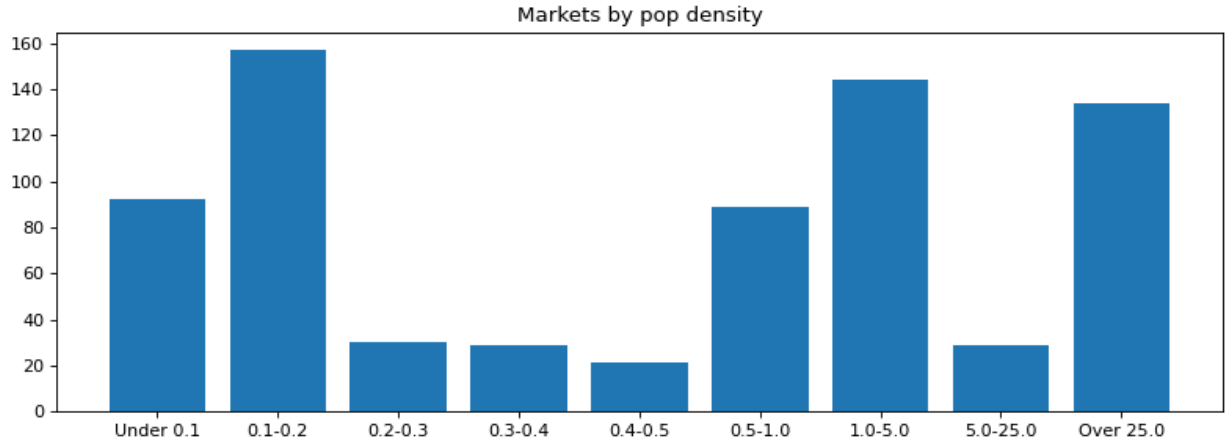
The next hypothesis is whether farmland influences the presence of farmers markets. If a region has a lot of farmland, are there more farmers markets? We grouped farmland into

buckets of 5%, and plotted the total number of markets in each. We found two major spikes. Farmers markets were most common in areas with very little farms (under 5%). The next highest spike was at 20-25%.



Upon analysis of the farmland dataset, none of the New York City counties were included. We found references to urban farms at https://www.nycfoodpolicy.org/15-urban-farms-and-gardens-bringing-fresh-produce-and-food-education-to-new-yorkers/ showing less than 66 acres of NYC is urban farmland. This is orders of magnitude smaller than even the smallest county (Putnam at 3,942 acres) and is essentially a statistical anomaly. NYC is approximately 193,700 acres, giving it about 0.034% farmland. For all intents and purposes, farmland in NYC is virtually nonexistent. Aside from that spike, there appears to be a sweet spot of 20-25%.

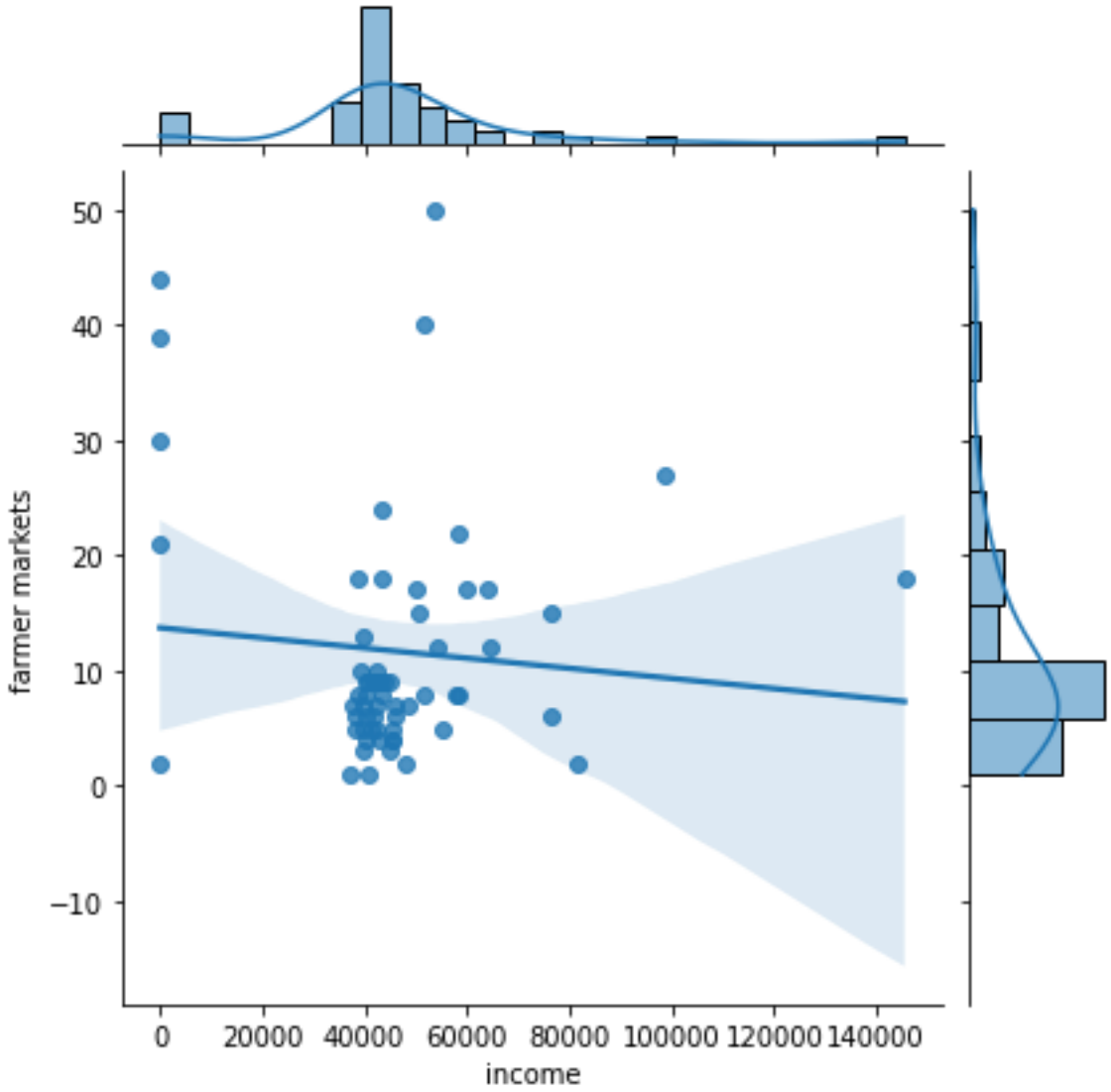
Finally, we looked at whether or not population density made a difference in the number of markets. We bucketed the density into several groupings. Note that this is population density per acre, not square mile (1 acre is 0.0015625 square miles), so these density numbers will appear low.

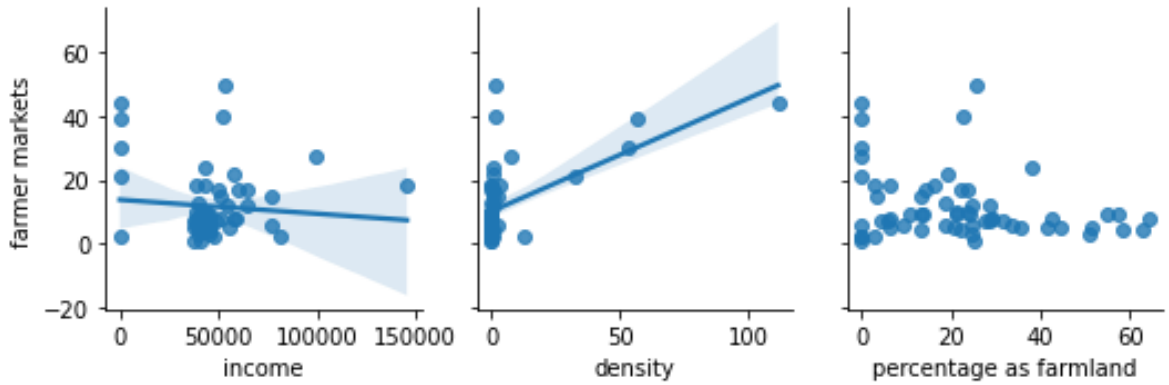


We can see a U shaped distribution. Farmers markets appear to congregate in low density areas and high density areas, and less so in midsized regions.

Linear Modeling

We felt that the bar charts were insufficient to illustrate the trends, and that linear modeling would be beneficial. The Seaborn package was installed to generate modeling charts. Linear models were created for income, density and farmland percentage.





We can see a downward correlation with income, an upward with density, and a clustering of markets around the 0% and 25% farmland usage.

Conclusion

Based on our research and linear models, we can draw a few conclusions. First, it appears that there is a trend for farmers markets to be less prevalent as income increases, and more prevalent at lower income levels. Secondly, we can see that farmers markets are more prevalent in areas with higher population densities. Finally, it appears that farmers markets show up most in areas with virtually no farmland, or a moderate amount of farmland, but not as much in-between. My current working theory is that farmers markets are prevalent the most in large cities with no farmland or in smaller cities that are surrounded with nearby farmlands. Additional analysis would need to be conducted to confirm that hypothesis or to see if it holds true outside of New York.

I want to acknowledge that there are certainly a lot of socioeconomic circumstances and nuances that would go into drawing definitive conclusions from this analysis. While we may show a correlation, there may not be enough evidence to suggest either a causation or recommend a course of action. For this project, we were looking at correlations. It should also be noted that just because there are markets in a county, they may not be easily accessible to all the citizens of said county. It is possible that lower income citizens may live in different parts of the county than higher income residents, yet the farmers markets are all located physically closer to the higher income areas. Further studies should be conducted into 'food deserts' and more granularity of income to distance (such as distance from the nearest farmers market). It is also possible that the markets offer different value propositions, depending on location. Farmers markets in large cities may sell produce at costs lower than the grocery store market rate, while smaller cities may have the reverse. Conversely, large city farmers markets may be less of a bargain as compared to their smaller counterparts. Further analysis of the price of goods in each region would be beneficial in drawing more complete conclusions and correlations.