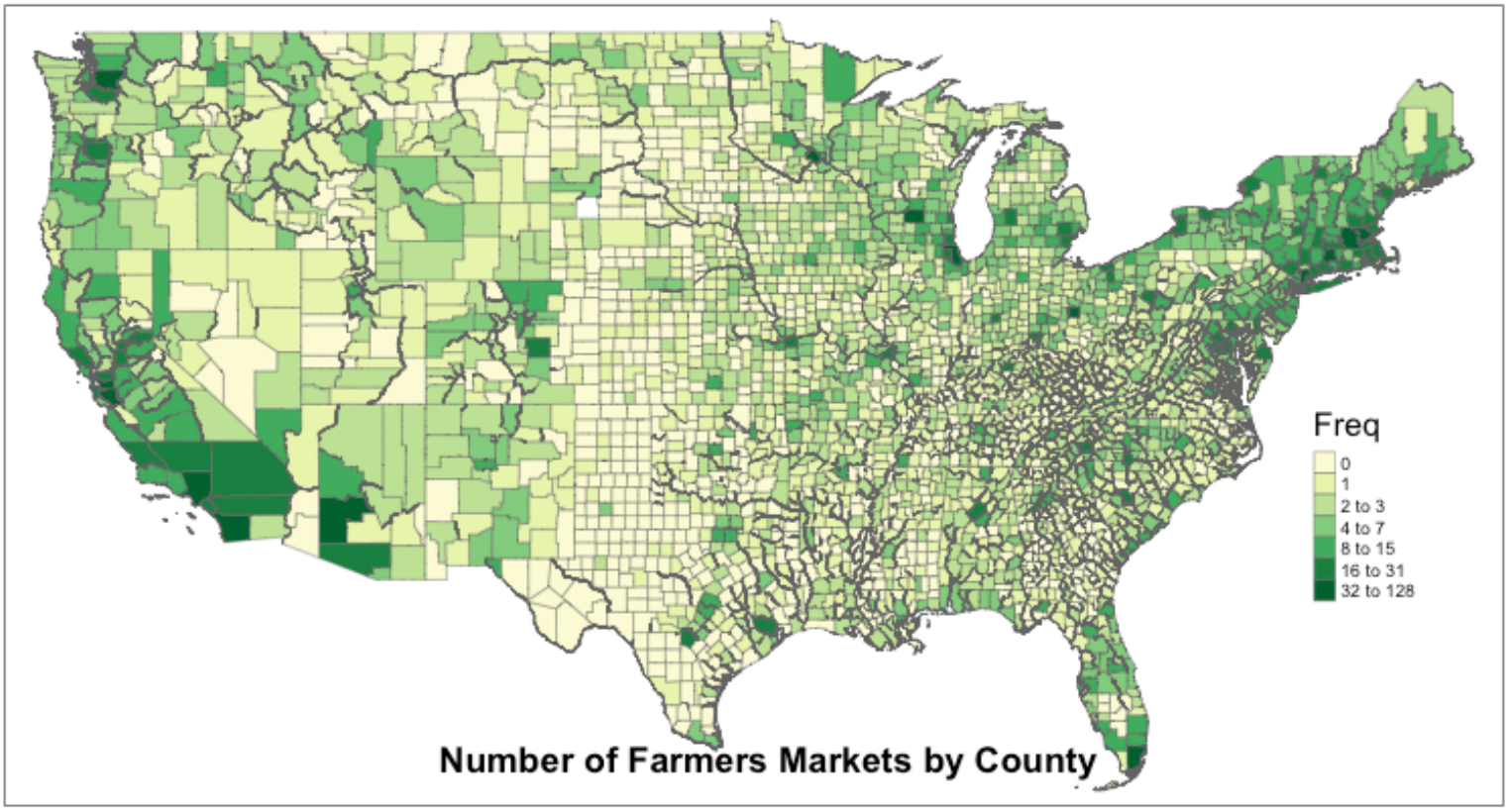
Negative Binomial Regression of Farmers Markets Occurrences



A Self-Directed Study by Michiel Besseling

Masters Capstone Project

Western Governors University

July 2020

Table of Contents:

I. Abstract

This study aims to determine which counties are underserved by farmers markets. In the context of this study, an underserved county shall be defined as a county whose observed number of farmers markets is less than the selected predicted model. Data from multiple sources is cleaned and joined together to create a comprehensive data set representing a multitude of relevant county level statistics. This data set was then explored to find relationships between variables using principle component analysis and hierarchical clustering, whose results were used to create a combination of variables that can explain the variation in the number of farmers markets. Poisson, negative binomial, and backwards eliminated zero inflated negative binomial models were constructed and compared to each other to build a regression model that best captures the variation in farmers market occurrences. The optimal model was selected and fitted to the existing observations and the underserved counties were discovered. Counties identified as underserved were selected based on their chi squared contributions.

II. Motivation

The number of farmers markets in the United States has been steadily increasing since the late 20th century. A farmers market has numerous benefits for both the farmers and the communities that they serve. They bring local farmers and producers to city grounds to sell local fruits, vegetables, honey, wine, nuts, and other locally produced items. The spirit of farmers markets aims to bring communities together in a place where consumers can purchase healthy, regional foods and socialize. McCarthy (2010) states that farmers markets have the profound ability to "grow the next generation of good eaters whose palates can discern the difference between a Roma and a beefsteak tomato" which leads to "these children [developing] the chance of selecting healthy foods because they like them." Clearly, farmers markets are more than a

luxury; they can have a significant impact on the quality of life in the communities that they serve.

While many prior studies have examined the individuals who patronize farmers markets, few studies have looked at the markets' geographical distribution. In a 2005 study by Wolf et al, researchers conducted two independent samples of customers from San Luis Obispo farmers markets and a local supermarket, and analyzed their results using chi squared tests for independence. The study determined that farmers market customers are significantly more likely to be female, be married with children, purchase local foods and wines, have higher levels of education, and have higher incomes than supermarket shoppers in the same region. While this study is useful in highlighting individuals' characteristics who frequent farmers markets, further research is needed to discover how these characteristics take shape at the county level. Therefore, this study seeks to compile many variables at the county level in order to predict farmers market frequency given a multivariate profile of a given county.

III. Research Question

Given a multivariate profile of a given region, how many farmers markets would one expect in that county?

IV. Software Selection

The R language via R Studio will be used to conduct this analysis. R can be installed on any Windows, Unix, and Mac platform. R is free and open source with many customized packages and functions written by users [Brittain et al 2018]. R can clean and join large data sets, create customizable and interactive visualizations, and build and diagnose many different types of models.

V. Data Gathering

The population of interest for this study is the lower 48 contiguous states, Hawaii, and the District of Columbia. Since Alaska has no traditional counties, it has been excluded from the study. Counties were chosen as the individual in this study since there is a reasonable amount of public data available at the county level. There are 3,114 counties in the population and this study will use each county's FIPS (Federal Information Processing Standards) code, which is a five-digit concatenation of the state and county codes, as a primary key.

This study aims to identify factors which can describe the geographic distribution of farmers markets, and as many independent variables were sought after as could explain the occurrences of farmers markets at the county level. Prior literature reviews of farmers markets have identified characteristics of farmers market customers which include age, ethnicity, education, family size, occupation, ethnicity, and number of family members (Rodriguez et al 2019). More specifically, prior research indicates that farmers market customers are more likely to be female, be married, and have higher levels of education than non-farmers market shoppers in the same region (Wolf et al 2005). Thus, this study aims to seek variables at the county level that can represent these traits and determine whether those factors have a significant impact on the frequency of famers markets.

The data for this study was gathered from reliable internet sources. The raw data files used were the farmers market table, 2016 election data set, multiple tables from the United States Department of Agricultural (USDA), files downloaded from the 2010 United States Census, and the zip code referential table downloaded from the United States Census Bureau (USCB). Two data sets available in R, the "zipcode" package and the "FIPS" function from the package "Tidycensus" were used to find zip codes or FIPS codes, given a city and state name.

The farmers market data was obtained through www.kaggle.com and the direct link to the posting can be found here. The table itself was likely scraped from the National Farmers Markets Directory released by the USDA whose link can be found here. The table contains the records of all 8,804 farmers markets registered with the USDA in 2020. There are 59 variables altogether, ranging from location data, social media accounts, and operating times to items sold. For the purpose of this study, the only information necessary from this table is the FIPS (Federal Information Processing Standards) code for each entry to be used to create a frequency table of the number of farmers markets per county. However, FIPS codes are not directly available on this table, and there are also many spelling inconsistencies and missing geographic data. Details on how these markets were matched to their respective counties will be covered in the next section of this report.

The USDA data was obtained through the QuickStat API (application programming interface) available from NASS (National Agricultural Statistics Service), a branch of the USDA. The data was selected to inspect the relationship between local agriculture output and farmers market occurrences. Variables selected at the county level include fruit and nut sales, vegetable sales, animal product sales, total crop sales, and agriculture land use. For each variable, the total estimated sales in US dollars were recorded as well as their respective coefficient of variance (CV). The CV is a dimensionless measure of the variance for each, with low values indicating that the estimates are reliable and higher values indicating they are less reliable.

The election data was located was on www.github.com and can be found here. This data set was scraped from various sources and contains 3,143 counties and 159 variables. The variables span such categories as votes for the 2016 presidential election by party,

meteorological data, ethnicity distribution, economic data, crime data, and health data at the county level. The above link has the details on how the user scraped the data.

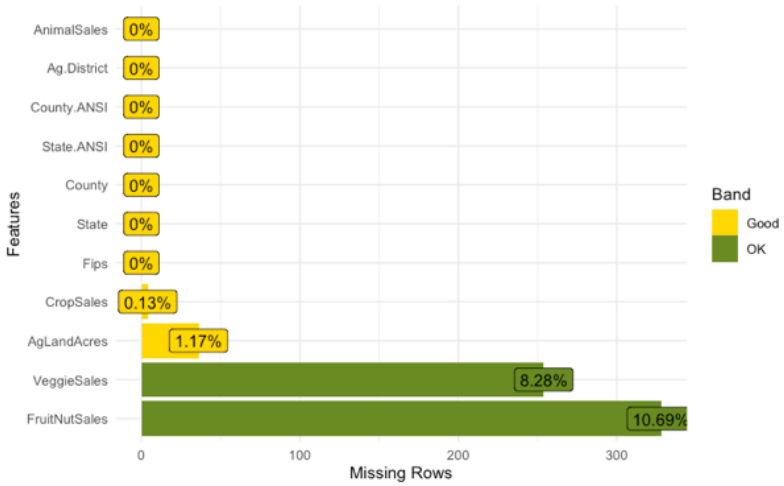
Since prior research has also indicated that age distribution and family size are significant factors in distinguishing farmers market shoppers and non-farmers market shoppers, the US Census was used to find possible relevant variables. The "Tidycensus" package in R allows the user to download selected variables directly from the US Census Bureau. Variables selected for this study included county level gender distribution, age distribution by gender, average family size, home ownership data, and marriage data. These variables were chosen to represent the results from prior research as well as possible.

The zip code referential table was obtained through the USCB at the link here. This table contains information on each zip code, such as the population and number of housing units in each county subdivision for each zip code in the 2010 census. Zip codes can span multiple cities and counties, in which case the same zip code can populate multiple rows– one for each county subdivision.

VI. Data Cleaning and Joining

This section describes how the data was cleaned, prepared, and joined. The R code for this section can be found in the "Data Cleaning" through the "Final Join" section of the accompanying R Notebook (see appendix for link).

The goal in preparing the data is to create a frequency table of the number of farmers markets per county that will be joined with the cleaned election data, USDA data, and US Census data. This section will also discuss dropped rows and columns, handling of missing values, and the construction of derived variables. The complete codes can be found in the R Notebook located in the appendix of this report.



Cleaning the USDA data required joining the five separate tables downloaded from NASS and imputing missing entries and entries with special codes. Each table includes state and county codes, agricultural regions (as defined by the USDA), the value of each statistic, and its coefficient of variance (CV). The first step is to join all five tables together by county, but only keeping state, county, agricultural district, and the given value of each statistic. The graph on the right shows the missingness of the joined USDA data. Although the missingness is relatively low for most variables, the USDA does disclose some of its values which is encoded as "(D)" in the spreadsheet. This is done to avoid disclosing values for individual farming operations. Since the size and agricultural profile of counties vary from state to state, a local approach was used to impute missing USDA data. Missing or disclosed values were imputed using the agriculture district medians. If the agricultural region data is missing, state medians were used. Medians were selected as opposed to means since medians are resistant to outliers, which abound in the data. After imputation, the USDA data contains the records of 3,068 counties and 11 variables, six of which are the independent numeric variables.

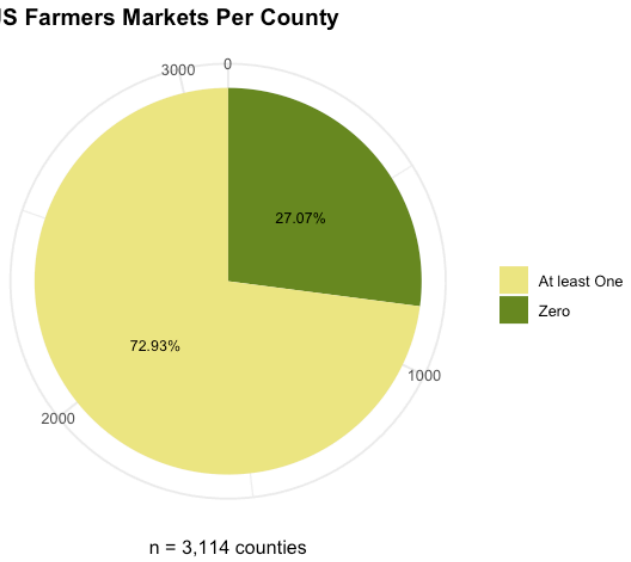
The election data contains a plethora of measures at the county level, including 2008, 2012, and 2016 presidential election votes, economic measures, ethnicity distributions, occupations, health statistics, and meteorological measures. The initial data set contained 3,143 rows (counties) with 159 variables. The data was checked for missing values. Some columns included the number of votes for non-major party candidates with very low total votes, so these

candidates' votes tallies were dropped. Thus, only votes for the major Democratic and Republican candidates for the 2008, 2012, and 2016 elections were kept, along with the Green and Libertarian parties for the 2016 elections. The homicide and infant mortality variables had over 50% missing values, so they were dropped from the study. The top four variables containing missing values included: temperature readings with 1,392 (44.3%) entries, HIV prevalence readings with 811 (25.6%) entries, Green Party with missing in 513 (16.3%) entries, and precipitation readings with 511 (16.3%) entries. In order to preserve data integrity and reduce variability, all these missing values were imputed with the corresponding state median values. There were still 511 Green Party vote values missing, implying some entire states had missing Green Party votes. These remaining 511 (16.3%) entries, along with 119 (3.8%) missing entries for HIV prevalence rates, were imputed with their overall population medians. Thus, any interpretation of Green Party votes in the study should include a mention on how the 513 missing values were imputed.

To obtain county level statistics that pertain to age, gender, marriage, renter occupied households, and family distributions, the R package "Tidycensus" was used to download Census statistics. First, a list of all available variables from the 2010 Census was obtained. Upon inspection of the available variables for download and considering the aforementioned findings of Rodriguez et al (2019), the following variables were derived from their corresponding 2010 Census value: family to household ratio, married household to family household ratio, urban ratio, renter occupied, household size categories, average household size, average family size, total female population, male age groups, female age groups, and husband-wife-children family ratio.

The entries in the initial farmers markets table needed to be matched to the FIPS code of their respective county. Markets outside of the population– the lower 48 states, DC, and Hawaii– were dropped. Since many county names were inconsistently spelled or missing, it was decided to use the zip codes to match the farmers market to its county. To do this, the zip code referential table was used acquired through the US Census Bureau. However, each zip code in this table does not always contain a unique row, since some zip codes span more than one county. So, it was decided that zip codes that spanned multiple counties would be assigned to the county where the greatest proportion of residents in that zip code live. However, there were also multiple missing zip codes and some zip codes that were not five digits in the farmers market table. In cross referencing, it was discovered that three- and four-digit zip codes contained missing leading zeroes, and ten-digit zip codes included a hyphen and four-digit extension. Those entries were adjusted. The remaining zip codes in the farmers market table were located using the "zipcode" package by city and state name. To matched the remaining unmatched counties, a few hand-tailored cleaning methods were employed, such as using "Parish" instead of county for Louisiana, capitalizing the third letter of counties that begin with "Mc" or "De." A few of the remaining unmatched county names could be easily identified by inspection were entered in manually. The final proper county names were matched to their FIPS code. After all counties that could be matched were matched, only 116 of the 8,804 (1.3%) were left unmatched and then dropped. Lastly a table containing each farmers market and FIPS code

was created for later visualizations, and a frequency table of the number of markets per county was created, with a total of 2,241 counties with at least one farmers market.

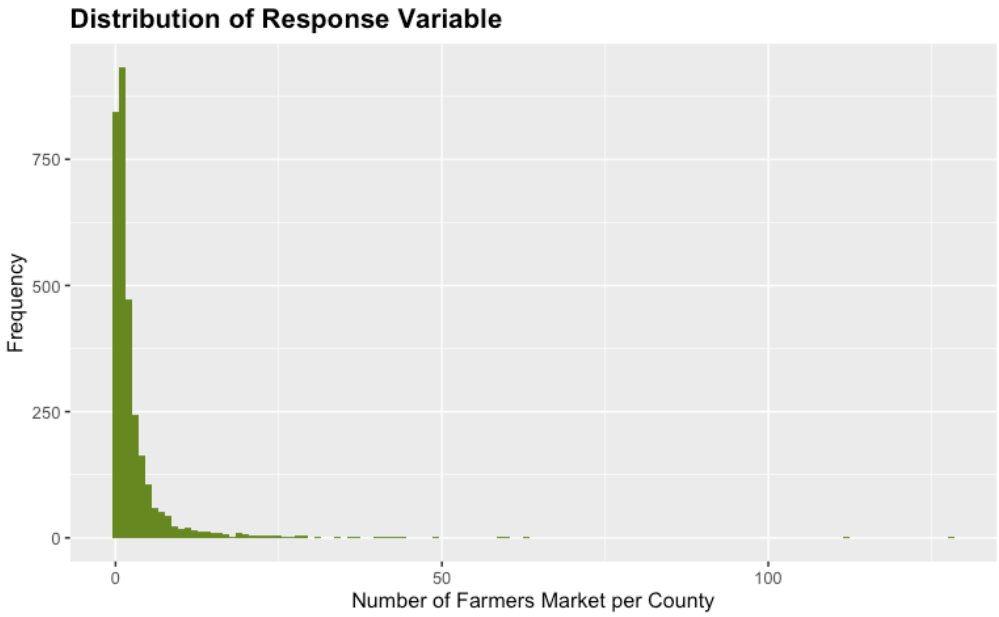


The last phase of data cleaning involved joining all the tables together and then imputing the remaining missing values. The cleaned farmers market frequency table, USDA data, election data, census data, and unique zip code table (for only county area and number of housing units) were joined together using each county's FIPS code as a primary key. There were at most 58 (1.8%) missing values for any column in the newly joined table. All missing values were imputed using state medians, except for the District of Columbia, which was missing USDA agricultural data. Considering that DC is an urban area and agriculture is likely minimal, the minimum value for the missing USDA data was imputed there. There were a few missing agricultural districts left, due to a few changes in FIPS codes in 2015, particularly in the state of Virginia. Thus, any missing agricultural district in the state of Virginia was assigned "Virginia (Unknown)," and all missing agriculture districts outside of Virginia were simply assigned "Unknown." Upon inspection, some variables were duplicated under different names, so the redundant variables were dropped. Lastly, a few derived variables were added, including the percent change in 2016 votes of both Democratic and Republican candidates from 2008 and 2012, and the range of annual temperatures.

After all of the joining and cleaning was finished, there were 3,114 rows and 121 variables in the final cleaned data ready for analysis. 116 (1.3%) of the original 8,804 farmers markets which could not be matched had been dropped.

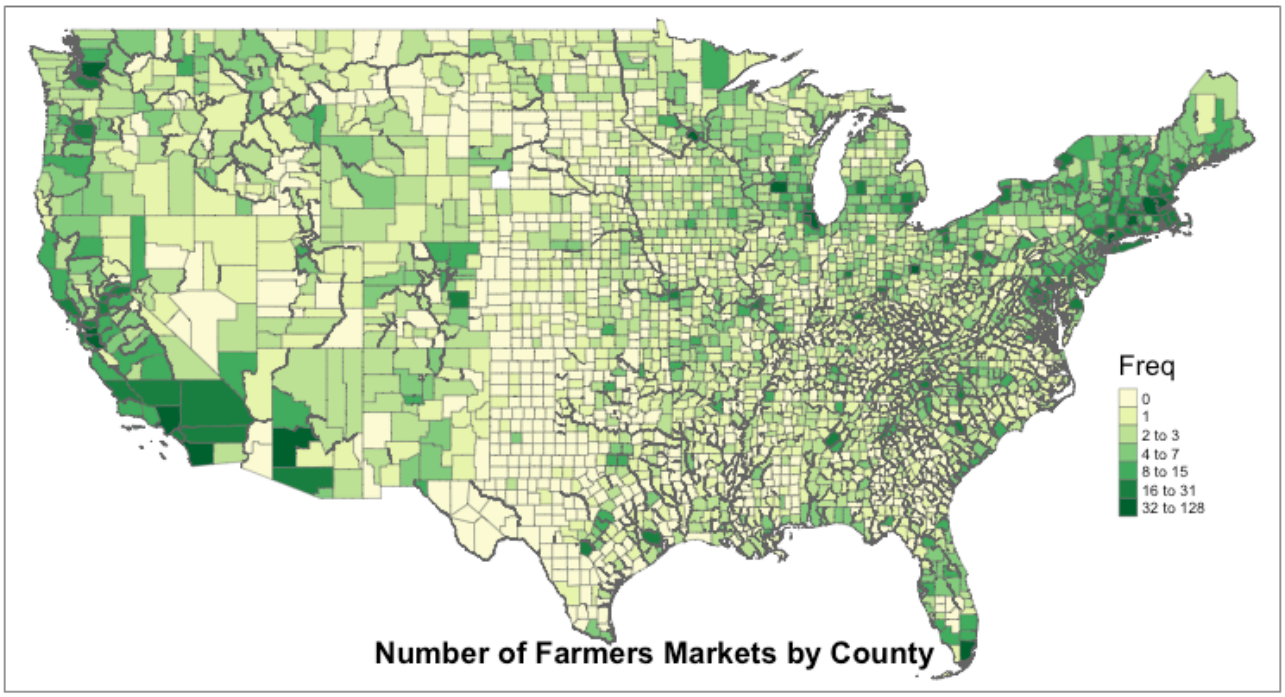
VII. Data Exploration

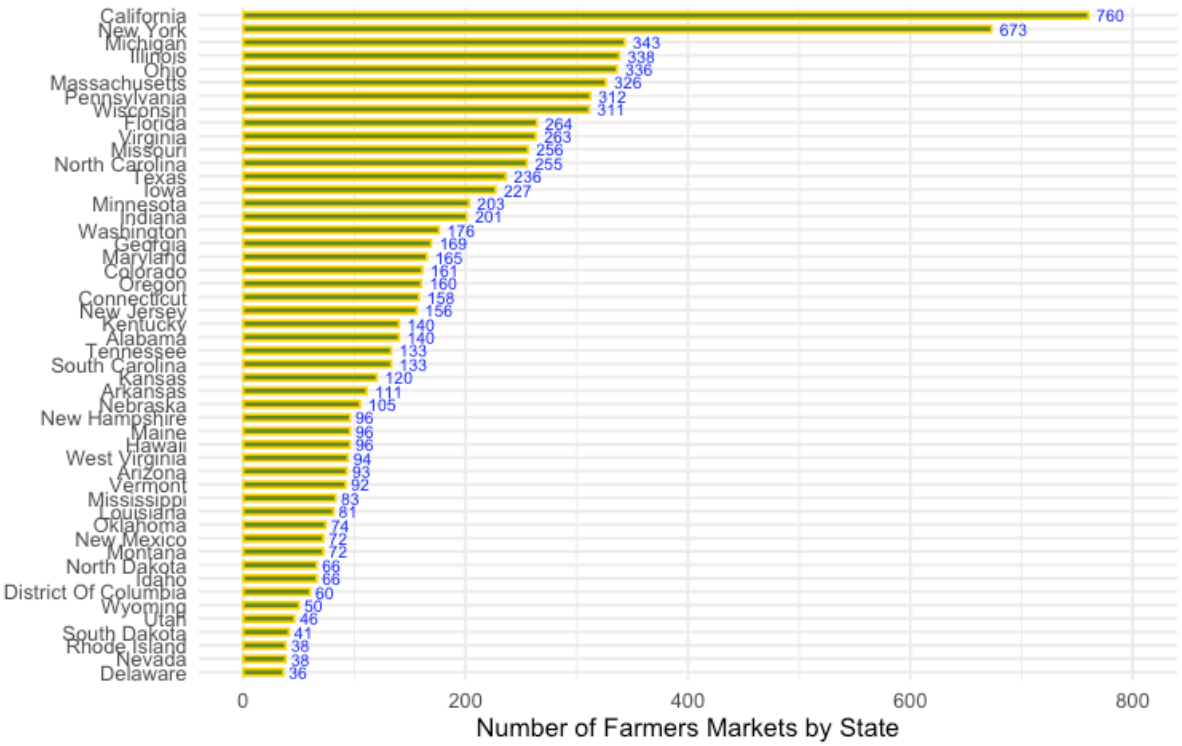
With such a large number of variables, many possible interactions exist. This section explores some relationships in the data. The R code for this section and additional graphs can be found in the "Data Visualization" section of the accompanying R Notebook (see appendix for link).



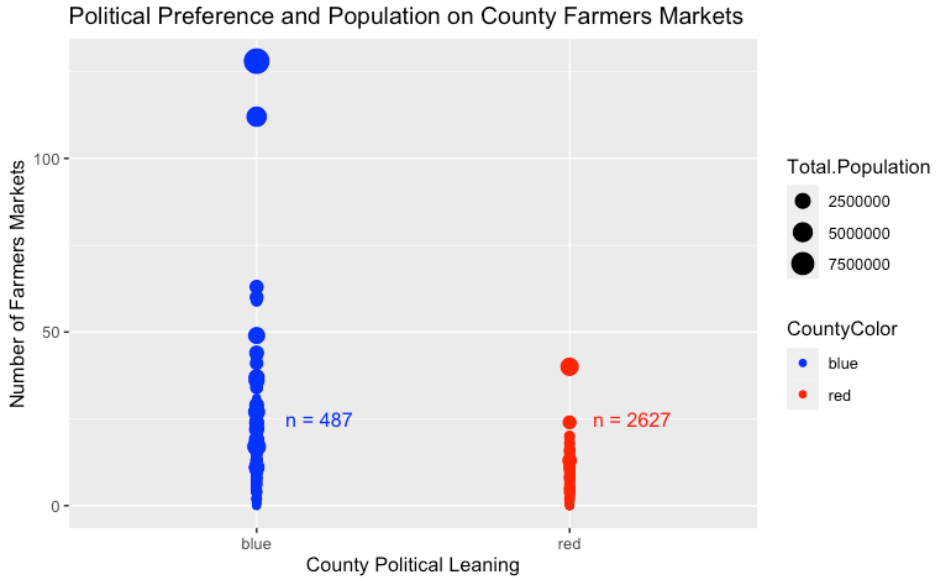
As illustrated by the above histogram, the distribution of the number of farmers markets per county is strongly skewed to the right with a few high outliers. The high frequency of outliers corresponds with the fact that this is a count variable, which is likely to follow a Poisson, or negative binomial, distribution. Thus, when making predictions based on this data, careful awareness for outliers should be considered. The median of the distribution is 1, and the mean is 2.7 farmers markets per county. Over 75% of counties have fewer than 3 farmers markets. So,

of the 3,114 counties, 843 have zero farmers markets, 932 have 1, 472 counties have 2, 243 have 3, and 624 counties have 4 or more farmers markets.

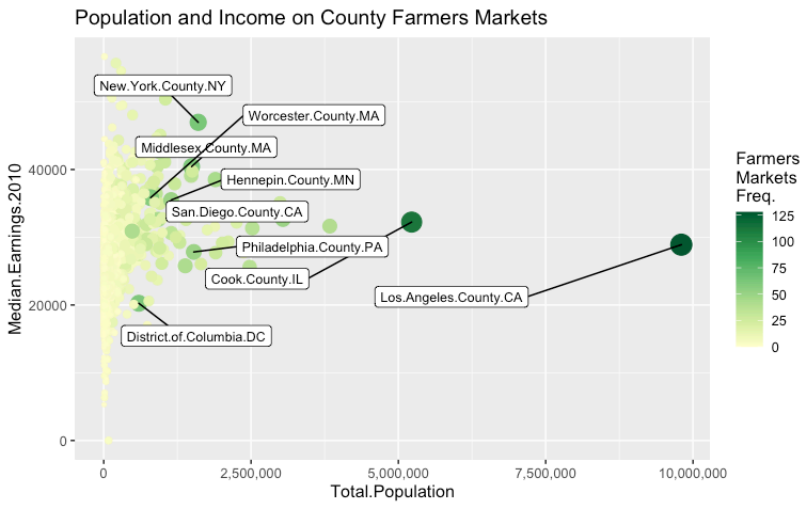




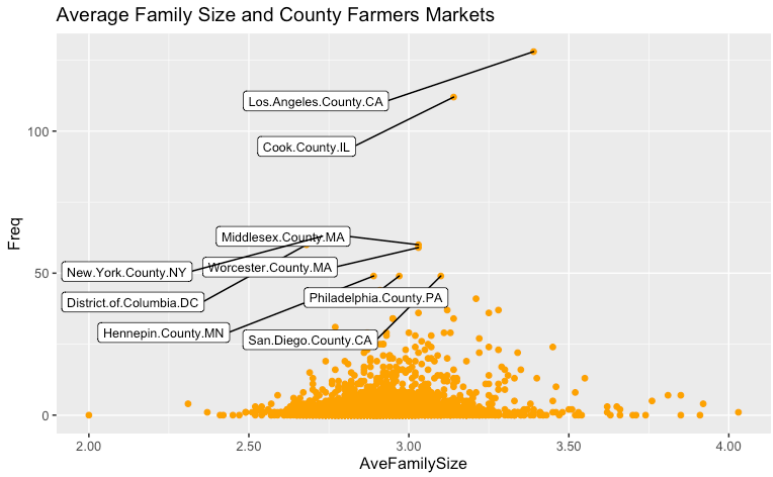
As shown by the above bar chart, California has the greatest number of farmers markets, but it also has the highest population, which must be considered as a potential cause of its high frequency of farmers markets. On the other hand, Texas and Florida have the next highest populations, yet do not rank as high in number of farmers markets. In fact, the states with high numbers of farmers markets tend to be "Blue States," or states that tend to vote Democrat. Below is that relationship at the county level, illustrated by the side-by-side dot plot.



So, while population size still ought to be considered, political preference may also have a strong effect. "Blue" counties do tend to have a higher number of farmers markets, but they also tend to have higher populations than "red" counties. However, it is still worthwhile to investigate political profiles in the modeling stage, as political stance appears to be a visual correlation on farmers market occurrences, so recognition of "red" vs "blue" counties could very well be linked to an explanation for the variance in farmers market occurrences.



The above scatterplot reveals that there is a positive relationship between the population size and the number of farmers markets, and the relationship between income and farmers market is weakly positive. So, population still needs to be investigated, while income should also still be considered as another potential explanatory variable.



In the above scatterplot, the relationship between average family size and farmers market is curved, with a peak farmers market frequency occurring somewhere around an average family size of 3.1. Therefore, counties with either small average family sizes or large family sizes tend to have fewer farmers markets.

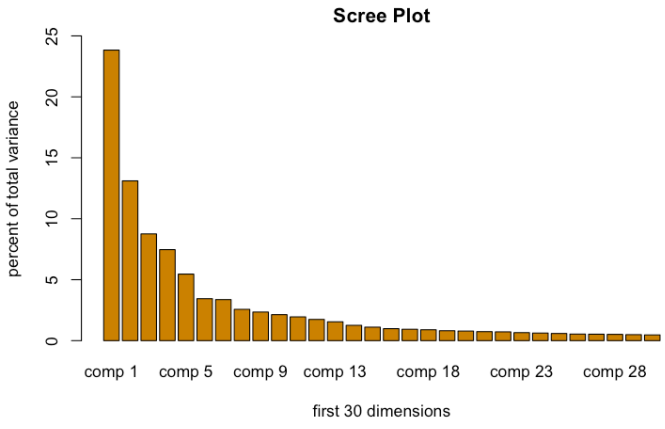
VIII. PCA and Hierarchical Clustering

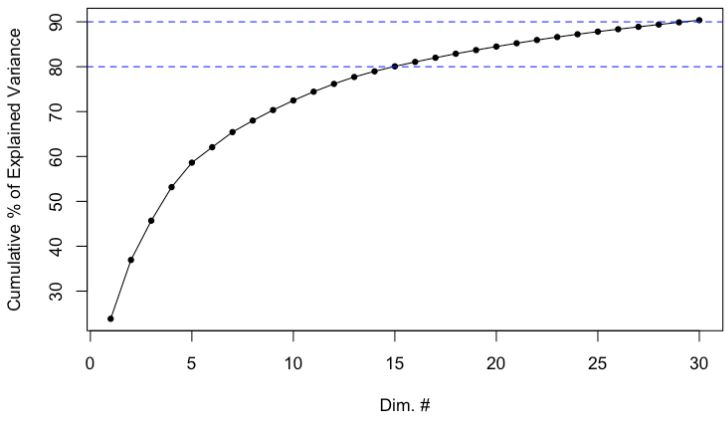
With so many variables and possible relationships, it would be helpful to discover relationships between variables (rows) and individuals (columns). Given that all of the independent variables are numeric, Principle Component Analysis (PCA) can explore these relationships. PCA assumes that the relationship between all pairs of variables is linear, which is assumed in this study. Although it is likely that not all relationships between variables are linear, Husson et al (2010) claim that most relationships can be assumed to be linear for initial approximations. (Subsequent data interpretation or feature selection algorithms may require further refining in order to be more accurate.) Additionally, PCA requires variables to be in the same units, so each variable is standardized during the execution of PCA and hierarchical

clustering. The R code for this section and additional graphs can be found in the "PCA" and "Hierarchical Clustering" sections of the accompanying R Notebook (see appendix for link).

The goal of PCA is to reduce the dimensions of the data by forming a basis for the higher dimensional vector space by decomposing it into its principle dimensions of variability. The dimensions, which are composed of a linear combination of the variables, can themselves can be used as new variables in the study. Using these new variables, one can typically reduce the number of dimensions to account for a given proportion of the total variation in the data set and to study and plot the relationship between both individuals and variables. The dimensions are constructed in descending order with respect to their proportion of variation, measured by their eigenvalue. Each dimension can be described by examining the types of variables (or individuals) that have the greatest contribution, or the smallest angle to the axes in the construction dimension. Individuals and variables that are near to each other on a PCA biplot share similar characteristics with respect to the two dimensions.

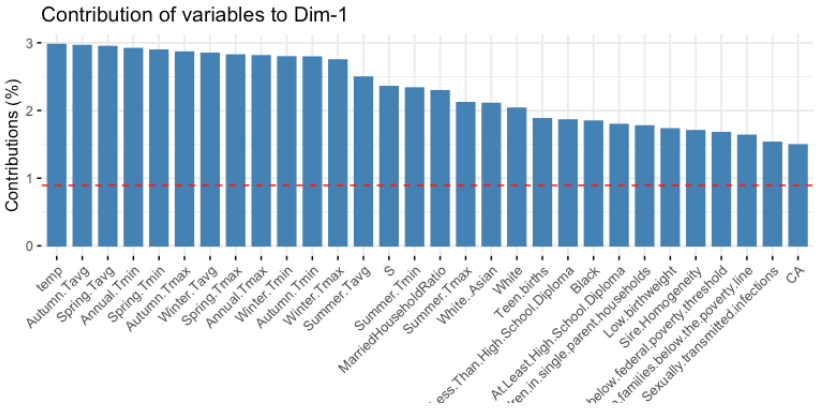
In this study, 113 quantitative variables will be used in the construction of the dimensions. The state names, county names, and farmers markets frequency are used as

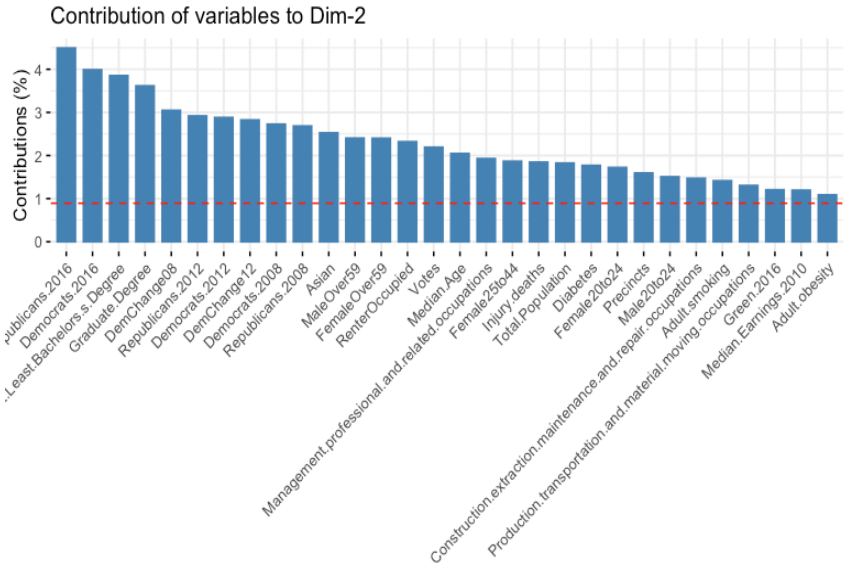




supplemental variables. Thus, analyzing the PCA results will be akin to studying the distribution of these 113 variables by geographic location.

The above scree plot shows that after running the PCA, the first dimension accounts for nearly 25% of the variability, and after the seventh dimension, each additional dimension is contributing roughly the same amount of variation. The cumulative scree plot (above right) shows that approximately 80% of the explained variance is captured by the first 15 dimensions and that approximately 90% of the explained variance is captured by the first 30 dimensions. These facts will be considered later when creating a model later in this report. The first four dimensions of the PCA will be examined.

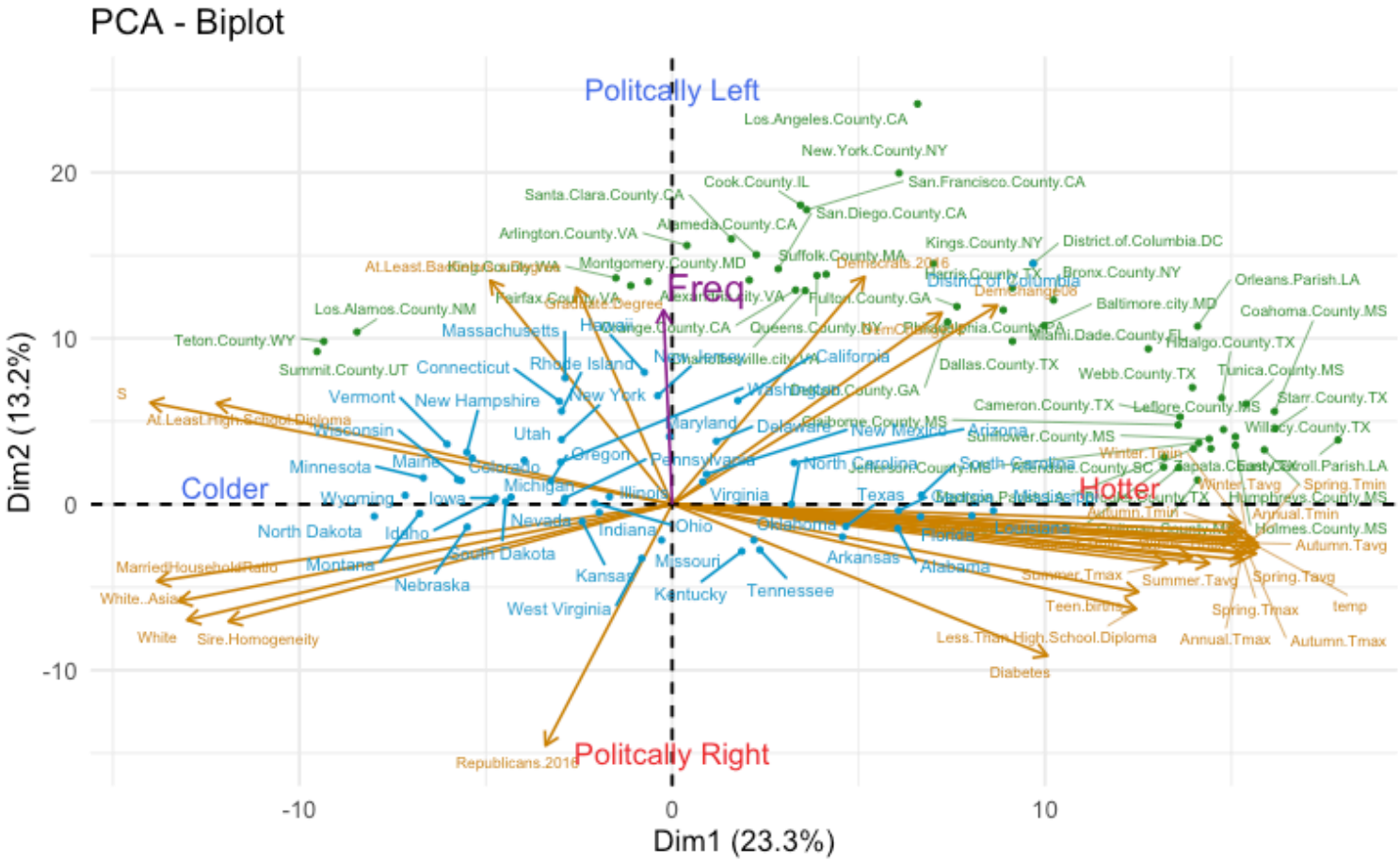




The first dimension of the PCA accounts for 23.3% of the total variability in the data and is mostly defined by temperature data. All of the different temperature statistics are likely to be collinear with each other. Since the "temp" variable, which represents the average annual temperature, has the strongest contribution to the first dimension, it would be recommended as the strongest single variable to represent temperature in future models. Other contributors to the first dimension include

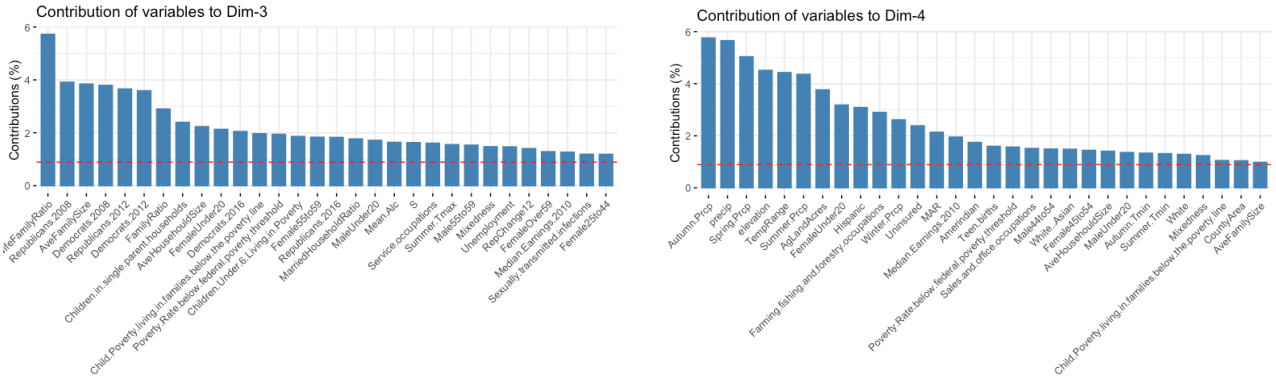
ethnicity and education. The second dimension is strongly represented by political parties and education, with Republicans 2016 votes being the single largest contributor.

Interpretation of these dimensions by themselves is limited, as they represent a bulk of the data contained in the cleaned table, not necessarily the most significant variables. However, one can examine the relationship of both individuals and variables with respect to these two dimensions on a PCA biplot. The plots should be interpreted with the understanding that the farther an individual or variable is from the origin, the more significant that factor is in



separating the data. If the variable is on the positive side of either the x or y axis, the more positive the correlation with respect to that dimension. The converse is true for negative sides of the axes. Individuals or variables near to each other share similar traits with respect to that dimension.

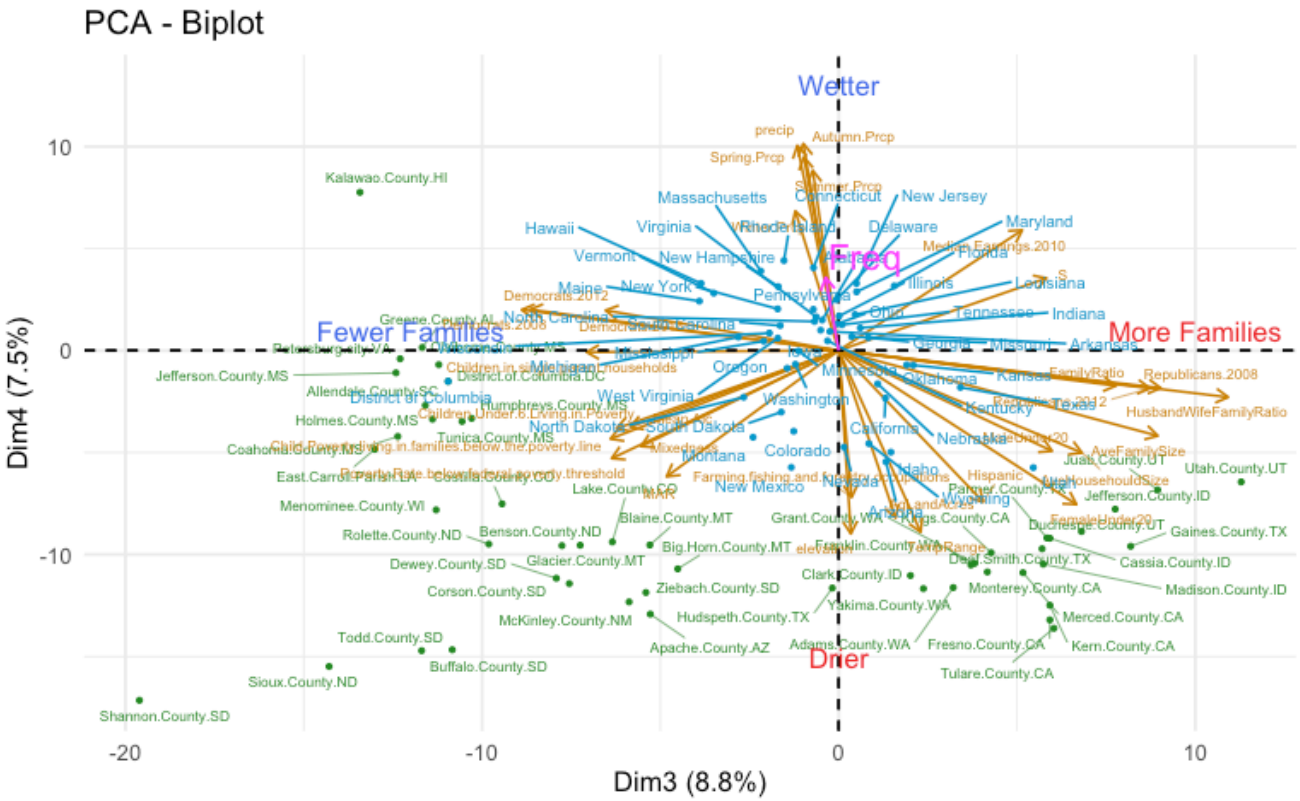
The first dimension mainly divides hotter areas on the right with cooler areas on the left. The second dimension separates politically left-leaning areas on the top half of the graph with politically right-leaning regions at the bottom of the graph. To avoid cluttering on the biplot, it should be noted that only the top contributing 50 counties and 30 variables were plotted. The variable labeled "Freq" is the frequency of farmers markets. This graph suggests that farmers markets are positively correlated with variables such as graduate degrees and proportion of Democratic votes. Based on their closeness to these criteria, regions positively correlated with farmers markets include Hawaii, New Jersey, and Montgomery County, MD.



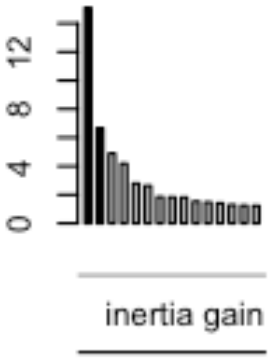
The third dimension can best be described by family-oriented characteristics, such as the husband/wife family ratio, or the proportion of households with married couple and children. The fourth dimension can be described by precipitation statistics.

The third dimension separates higher family ratios to the right, with smaller families and ones with higher rates of children living in poverty. The fourth dimension separates wetter climates on the top half, with drier and higher counties on the bottom half. With respect to the third dimension, farmers market frequency does not seem to correlate in either direction, supporting an observation noted previously in plotting average family size and farmers market

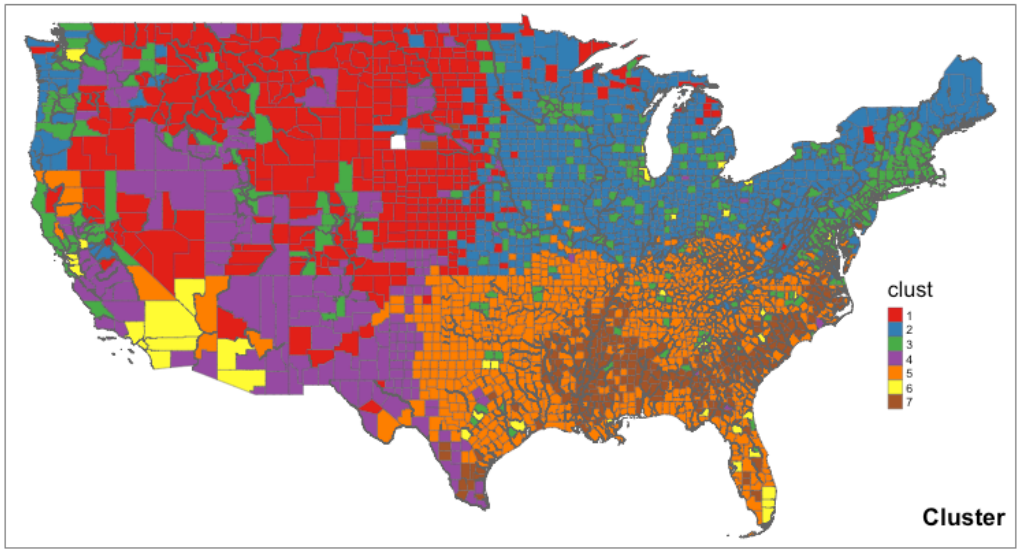
frequency. The fourth dimension suggests that famers markets are positively correlated with wetter climates.

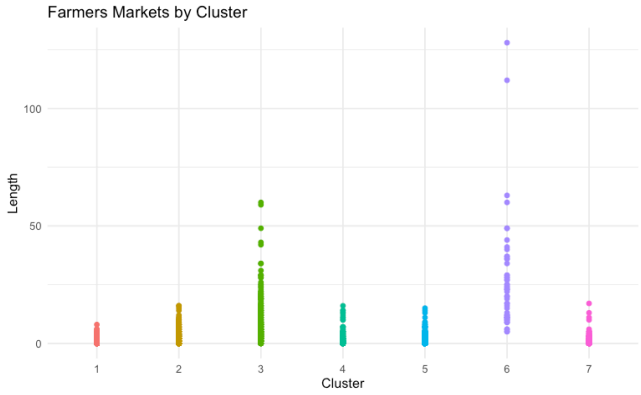


PCA was used in hierarchical clustering, used to group individuals based on common characteristics. Seven clusters were chosen for this study as the inertia gain dropped off significantly after the seventh cluster. The preceding list of characteristics for each cluster are all statistically significant at a 2% significance level. Variables are deemed significant when the mean value for the cluster is significantly different from the population mean. Exact details for the variables that separate each category can be found in R Notebook attached with this study.



Counties in Cluster 1 are characterized by older population, high elevation, more likely to vote Libertarian, and smaller household size. Counties in Cluster 2 are characterized by colder climates, white, married, and a large increase for Republican votes in 2016. Counties in Cluster 3 are characterized by higher levels of education, higher incomes, higher Democratic votes, higher average family size, and larger populations. Counties in Cluster 4 are characterized by large counties with agriculture, Hispanic, large families, young population, farming jobs, high male population, and low education levels. Counties in Cluster 5 are characterized by warmer temperatures, poor health metrics, construction jobs, small populations, and low third-party votes. Counties in Cluster 6 are characterized by large populations, higher HIV rates, more renters, younger population, and children living in single parent households. Counties in Cluster 7 are characterized by high African American populations, poor health metrics, warmer temperatures, unemployment, older populations, and low married households. Farmers markets frequencies





are positively correlated with clusters 3 and 6, and negatively correlated with clusters 1, 2, 4, 5, and 7.

Selected Variables by Cluster

Cluster	Variable	Cluster mean	Overall mean	p value
1	elevation	991.271	399.052	0
1	Median.Age	45.207	39.903	0
1	MarriedHouseholdRatio	0.835	0.763	0
1	AgLandAcres	215715.49	72353.224	0
1	Farming.fishing.and.forestry.occupations	4.321	2.11	0
1	Libertarians.2016	4.185	3.163	0
1	At.Least.High.School.Diploma	88.038	83.009	0
1	Republicans.2016	72.543	63.597	0
1	Freq	1.054	2.79	0
1	AveFamilySize	2.802	2.923	0
1	AveHousehouldSize	2.283	2.479	0
2	Sire.Homogeneity	0.859	0.719	0
2	White	92.252	79.035	0
2	RepChange12	15.657	7.502	0
2	At.Least.High.School.Diploma	86.493	83.009	0
2	Production.transportation.and.material.moving.occupations	19.002	16.252	0
2	MarriedHouseholdRatio	0.79	0.763	0
2	Libertarians.2016	3.667	3.163	0
2	Freq	2.361	2.79	0.013
2	Uninsured	0.139	0.179	0
2	Winter.Tmax	357.047	449.02	0
3	Graduate.Degree	13.212	6.445	0
3	Management.professional.and.related.occupations	38.357	29.827	0
3	Asian	3.598	1.071	0
3	Median.Earnings.2010	30944.886	25437.655	0
3	Democrats.2016	47.587	31.69	0
3	HusbandWifeFamilyRatio	0.329	0.279	0
3	School.Enrollment	79.487	74.986	0
3	Freq	7.897	2.79	0
3	Total.Population	262081.426	97754.035	0
3	AveFamilySize	2.989	2.923	0
3	Children.in.single.parent.households	0.273	0.316	0
3	Republicans.2016	46.009	63.597	0
4	Hispanic	36.309	7.94	0
4	AveFamilySize	3.182	2.923	0
4	AgLandAcres	292620.602	72353.224	0
4	HusbandWifeFamilyRatio	0.333	0.279	0
4	Farming.fishing.and.forestry.occupations	4.82	2.11	0
4	Median.Earnings.2010	23548.701	25437.655	0
4	TotalFemale	0.492	0.501	0
4	Freq	1.787	2.79	0.014
4	At.Least.High.School.Diploma	76.792	83.009	0
5	temp	16.05	12.623	0
5	Poor.physical.health.days	4.538	3.807	0
5	Republicans.2016	73.802	63.597	0
5	Uninsured	0.211	0.179	0
5	Construction.extraction.maintenance.and.repair.occupation	13.182	11.519	0
5	Children.in.single.parent.households	0.329	0.316	0
5	Freq	1.232	2.79	0
5	Total.Population	47497.851	97754.035	0
5	Green.2016	0.555	0.85	0
5	Libertarians.2016	2.303	3.163	0
6	Total.Population	1547092.58	97754.035	0
6	Freq	26.2	2.79	0
6	HIV.prevalence.rate	815.847	149.193	0
6	RenterOccupied	0.468	0.277	0
6	At.Least.Bachelors.s.Degree	33.278	18.995	0
6	Children.in.single.parent.households	0.404	0.316	0
6	Median.Age	34.5	39.903	0
7	Black	42.011	8.83	0
7	Poverty.Rate.below.federal.poverty.threshold	24.855	15.478	0
7	Adult.obesity	0.357	0.306	0
7	Democrats.2016	49.551	31.69	0
7	Less.Than.High.School.Diploma	25.085	16.911	0
7	temp	17.32	12.623	0
7	Unemployment	0.107	0.077	0
7	Freq	1.089	2.79	0
7	Median.Age	37.358	39.903	0
7	MarriedHouseholdRatio	0.627	0.763	0

The "Selected Variables by Cluster" table shows some selected significant variables that represent each cluster. The occurrences of farmers markets are highlighted in green. Both the mean of the cluster and overall mean are represented in the table. Each variable is put through Kuiper's Test, whose null hypothesis in this study is that the cluster mean is equal to the overall mean.

Overall, the resulting dimensions of the PCA may or may not be useful directly in the modeling

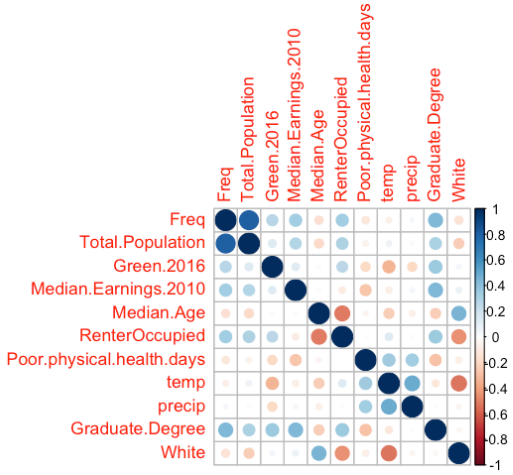
stage, but should be investigated further. The PCA is helpful in discovering potential relationships among groups of variables and can assist in identifying and selecting features that may represent multiple variables in order to reduce overall numbers of variables and collinearity in a given model.

IX. Model Building

In this section, a regression model will be created using selected variables. Since the response is a count variable, Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models are analyzed. Poisson models are checked for overdispersion, or when the conditional variance of the model is greater than the conditional mean. Where overdispersion is present, negative binomial models will be selected that produce the lowest AIC while all factors remain statistically significant. Zero inflated negative binomial models will next be analyzed. Lastly, backwards elimination will be run n the zero inflated model to build a model with fewer parameters. AIC is chosen as a selection criterion since it penalizes for the inclusion of extra parameters, as the goal will be to keep the models as parsimonious as possible. The R code for this section and additional graphs can be found in the "Model Selection" section of the accompanying R Notebook (see appendix for link).

Many different models were run, including using models using the first 30 dimensions of the PCA. Models were compared to each other using AIC and checking model diagnostics, particularly by examining QQ plots. All of the models created int his study assume that the residual error is normally distributed, which can be verified by inspecting a QQ plot of the residual, which should be in a straight line if the assumption holds. The regression models using the dimensions of the PCA turned out be sensitive to outliers and were not considered in the final model selection.

The final models to be evaluated used variables that were selected based on their relevance to prior research, their representation from the PCA, their ability to reduce collinearity (when applicable), ability to reduce AIC compared to similar variables, and their significance in the model itself. Variables thus chosen were the logarithm of the total population to linearize the skewed populations, Green Party votes to represent political mindsets, 2010 median earnings to represent income, median age, renter occupied rate for households to represent family distribution, poor physical health days to represent health trends, average temperature for meteorological data, graduate degree to represent education levels, and the proportion of white residents to represent ethnicity. Again, the particular variables chosen to represent a field were selected to ensure they were significant in the model and reduced AIC over other variables in that field. The variables were also chosen to reduce collinearity. The correlation plot shows the correlation between the response variable (Freq) and the selected independent variables. Although the graduate degree variable shows some correlation among other variables, its inclusion and exclusion were explored, and the model performed better with its inclusion.



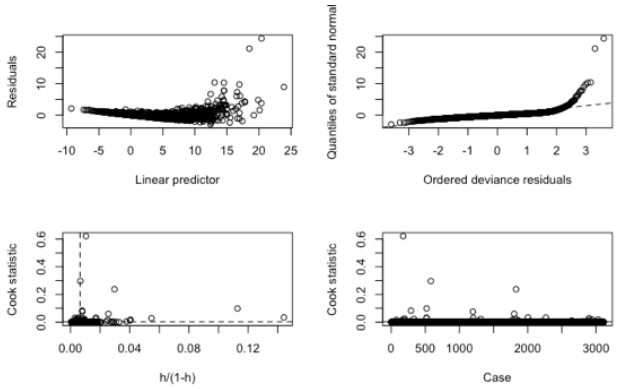
For the Poisson and negative binomial regression models, a chi-squared goodness of fit test at a 5% significance level is testing the hypotheses:

H_0: The distribution of farmers market occurrences follows the given model.

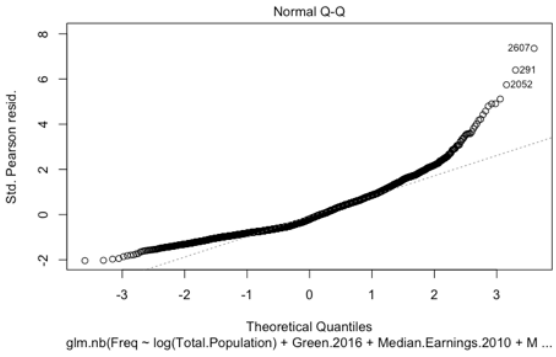
H_1: The distribution of farmers market occurrences does not follow the given model.

The deviance residuals are compared to the critical chi-squared value with the appropriate degrees freedom (sample size - number of parameters).

First, a Poisson regression model is fitted and checked for overdispersion. Using a goodness of fit test, the first Poisson models rejects the null hypothesis with an observed residual deviance of 61,118 on 3104 degrees freedom, far greater than the cut-off value is $\chi^2_{0.95}$ = 3234.8 under 3,104 degrees freedom. Additionally, a review of diagnostic plots shows unequal variance, and the residuals show non-normality based on the QQ-plot, the plot on the top right. Since there is evidence of overdispersion, a negative binomial model, which do not suffer from overdispersion, is investigated next.



After ruling out a Poisson model, the next model uses a negative binomial regression on the same set of variables. The goodness of fit test this time fails to reject the null hypothesis with a residual deviance of 2,779.2, below the 3234.8 critical value, on 3,104 degrees. The AIC of the model is 9,491, far below the Poisson with an AIC of 18,129. Thus, there is evidence to suggests that this model fits the data reasonably well. However, inspection of the QQ plot of the residuals shows some departures from normality. Although this model fits it could be improved upon by considering zero inflated models.



The next model to be considered is a zero inflated negative binomial (ZINB) model using the same set of selected variables. A zero inflated model assumes that certain members of the population will always be a zero. In the context of this study, using a zero inflated model suggests that there are counties with a certain set of characteristics that will never have a farmers

market. The zero inflated model separates the model into two components. The first component includes a binomial response, whether a county will have zero or at least one count. The second component is the specified distribution, which in this case is a negative binomial distribution, conditioned on the preface that a county has at least one count.

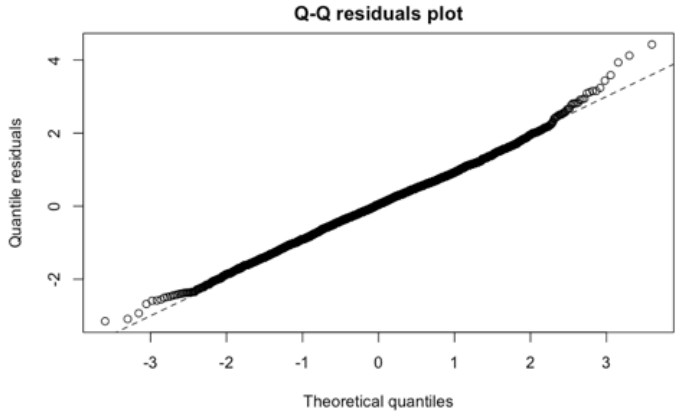
The zero inflated negative binomial model reduced the AIC to 9459.2 and the distribution of the residuals showed a better fit with respect to departures to normality, given a sample size of 3,114. To test whether this model is an improvement on the previous model, Vuong's Closeness Test, used to compare non-nested models, was run to compare this model to the previous negative binomial model. The hypotheses are:

H_0: The models are indistinguishable

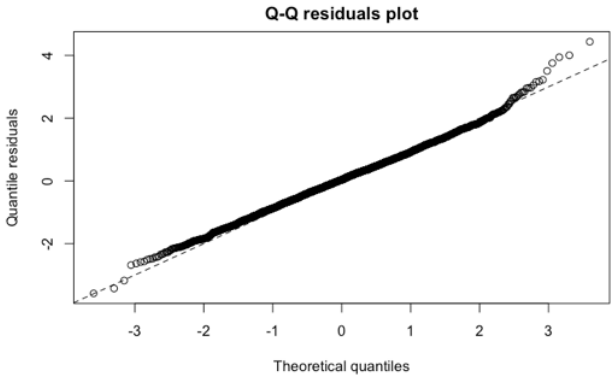
H_1: The models are distinguishable.

The zero inflated model was shown to be significantly lower using the AIC criteria, with a p-value of 0.0047. The parameter estimates are given in the printout and show that not all predictors are significant.

```
Call:
zeroinfl(formula = m5.nb, data = All_Final_Data_Cleaned2 %>%
mutate(Median.Earnings.2010 = Median.Earnings.2010/1000),
    dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.0366 -0.6724 -0.1497  0.4987  7.0929

Count model coefficients (negbin with log link):
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -8.0684944  0.2706574 -29.811  < 2e-16 ***
log(Total.Population)      0.6963074  0.0134850  51.636  < 2e-16 ***
Green.2016                 0.2280674  0.0218751  10.426  < 2e-16 ***
Median.Earnings.2010      -0.0112761  0.0032293  -3.492  0.00048 ***
Median.Age                 0.0471821  0.0036955  12.767  < 2e-16 ***
RenterOccupied             1.2524146  0.2323436   5.390 7.03e-08 ***
Poor.physical.health.days -0.0933566  0.0181369  -5.147 2.64e-07 ***
temp                      -0.0544433  0.0043058 -12.644  < 2e-16 ***
Graduate.Degree            0.0231387  0.0041089   5.631 1.79e-08 ***
White                      0.0005677  0.0010870   0.522  0.60151
Log(theta)                 2.7299611  0.1232024  22.158  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               26.85199    8.44812   3.178  0.00148 **
log(Total.Population)     -3.00209    0.73249  -4.098 4.16e-05 ***
Green.2016                -0.13218    1.74015  -0.076  0.93945
Median.Earnings.2010       0.13694    0.14960   0.915  0.35999
Median.Age                -0.26807    0.11504  -2.330  0.01980 *
RenterOccupied            -4.87373    7.58918  -0.642  0.52075
Poor.physical.health.days -1.58466    0.58225  -2.722  0.00650 **
temp                       1.00484    0.21988   4.570 4.88e-06 ***
Graduate.Degree           -0.24048    0.33457  -0.719  0.47229
White                     -0.03005    0.02414  -1.245  0.21311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 15.3323
Number of iterations in BFGS optimization: 77
Log-likelihood: -4709 on 21
```



To further improve the model, backwards stepwise selection using the zero inflated negative binomial model was employed, using a 5% significance level in variable selection. The reduced zero inflated model reduced the AIC slightly from the last model to 9453.0. A Vuong's Closeness Test was again run between this model and the previous model, resulting in a p-value of 0.1251. The results indicate that the latter model is not a significant improvement on the previous model. Like before, the residuals show no major violations of normality. By design, all variables are significant, as is shown by the model's printout below.



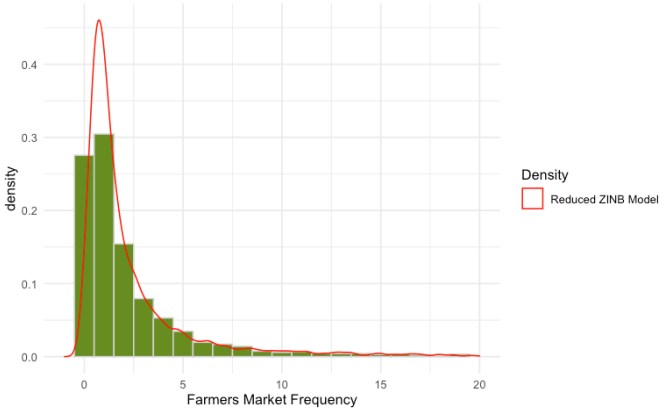

```
Call:
zeroinfl(formula = eval(parse(text = out)), data = data, dist =
dist)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-2.0311 -0.6783 -0.1510  0.4999  9.0286

Count model coefficients (negbin with log link):
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.036678   0.226183 -35.532  < 2e-16 ***
log(Total.Population)     0.696213   0.013156  52.919  < 2e-16 ***
Green.2016                0.225709   0.021608  10.445  < 2e-16 ***
Median.Earnings.2010     -0.011898   0.003160  -3.765 0.000167 ***
Median.Age                0.048289   0.003623  13.327  < 2e-16 ***
RenterOccupied            1.224023   0.208236   5.878 4.15e-09 ***
Poor.physical.health.days -0.090773   0.018035  -5.033 4.82e-07 ***
temp                     -0.056381   0.003798 -14.845  < 2e-16 ***
Graduate.Degree           0.024352   0.003981   6.116 9.57e-10 ***
Log(theta)                2.729628   0.123151  22.165  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              28.48974    7.16739   3.975 7.04e-05 ***
log(Total.Population)    -3.77411    0.82621  -4.568 4.92e-06 ***
Median.Age               -0.25837    0.08145  -3.172  0.00151 **
Poor.physical.health.days -1.30748    0.48117  -2.717  0.00658 **
temp                      1.16139    0.24598   4.721 2.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 15.3272
Number of iterations in BFGS optimization: 32
Log-likelihood: -4712 on 15 Df
```

The reduced zero inflated negative binomial was selected as the final model on the grounds that it has the lowest AIC and has the fewest number of variables, while the residuals do not violate the normal assumption.

X. Model Analysis

Interpretation of the parameter estimates is on the whole uninformative, as the model is split in two parts: the "zero" part of the model and the "count" part of the model. Additionally, the link between the two has to be considered, making

interpretation even more difficult and uninformative. Parameter estimates as written in the printout are in terms of log-odds. The count portion suggests once a threshold for a county having at least one farmers market has been surpassed, then population, proportion of Green Party votes, age, renter occupied household rates, and graduate degree rates are positively correlated with farmers market frequency, while median earnings, poor physical health days, and temperature have a negative correlation with farmers markets. The histogram to the right shows the density curve of the model fitted over the observed values. Note that the outliers have been truncated from the graph in order to show the majority of the observations more clearly.



The model was then fitted to the existing data and the chi squared component, $((observed - expected)^2/expected)$, was calculated for each observation was recorded. Chi squared contributions are chosen to measure the extent of a county's adherence to the model as it divides the squared deviance by the expected value. Thus, large and small counties alike can be top contenders. The sign for each observation was preserved, where a negative value indicates the model predicts more farmers markets than were actually observed. The table on the following page shows the lowest 30 ranked counties in terms of chi-squared contribution, indicating that they are the most underserved counties under the assumption of the model.

The top 30 counties using the selected variables under the zero inflated negative binomial model are listed in the table "Most Extreme Expected Observations Using Reduced ZINB Model." The list identifies the candidates for counties that are underserved according to the model.

Most Extreme Expected Observations Using Reduced ZINB Model

County	Frequency	Expected	Residual	Signed χ^2 Component
Queens.County.NY	19	45.0382	-26.0382	-15.0537
Arapahoe.County.CO	4	15.0605	-11.0605	-8.1229
Boulder.County.CO	6	17.9278	-11.9278	-7.9358
Richmond.County.NY	2	10.602	-8.602	-6.9792
Macomb.County.MI	8	18.5632	-10.5632	-6.0109
Whatcom.County.WA	3	10.8934	-7.8934	-5.7196
Middlesex.County.NJ	7	16.4347	-9.4347	-5.4162
Ramsey.County.MN	9	18.6041	-9.6041	-4.958
Benton.County.OR	2	8.3577	-6.3577	-4.8363
Harris.County.TX	17	28.735	-11.735	-4.7924
Kanawha.County.WV	1	6.3922	-5.3922	-4.5486
Deschutes.County.OR	2	7.945	-5.945	-4.4485
Dallas.County.TX	11	20.3913	-9.3913	-4.3252
Solano.County.CA	2	7.7959	-5.7959	-4.309
Bergen.County.NJ	12	21.1782	-9.1782	-3.9776
Otter.Tail.County.MN	0	3.974	-3.974	-3.974
St..Louis.city.MO	4	10.3761	-6.3761	-3.9181
Lane.County.OR	11	19.7148	-8.7148	-3.8524
Luzerne.County.PA	4	10.1961	-6.1961	-3.7653
Lucas.County.OH	5	11.4355	-6.4355	-3.6217
Union.County.NJ	6	12.7984	-6.7984	-3.6113
Jackson.County.OR	6	12.7859	-6.7859	-3.6015
Jefferson.County.WA	2	6.962	-4.962	-3.5365
Lewis.and.Clark.County.MT	1	5.2219	-4.2219	-3.4134
Essex.County.NJ	12	20.2681	-8.2681	-3.3729
Pinal.County.AZ	1	5.1318	-4.1318	-3.3267
Genesee.County.MI	5	11.0104	-6.0104	-3.281
Kings.County.NY	37	49.7468	-12.7468	-3.2661
Bernalillo.County.NM	8	14.9036	-6.9036	-3.1978
Weber.County.UT	1	4.9704	-3.9704	-3.1716

In cross referencing some of these counties, the model proves its worth. With a population over 2.25 million, Queen's County, NY should have many more farmers markets than 19. Arapahoe County, CO includes many suburbs of Denver and hosts a population over half a million and appears to be legitimately underserved as well. Kanawha County, WV contains the capitol city of Charleston, a metropolitan area of over 200,000 people, and a politically left profile, yet it has only one farmers market. A complete table of all counties named "FarmersMarketPrediction.csv" is available in the "Cleaned Data" folder accompanying this report.

XI. Model Interpretation

The purpose of this study is to determine which counties, given the current conditions, are underserved by farmers markets. There are a few drawbacks to this approach. First, since this is an observational study, there is no evidence of causation. Additionally, the concept of being "underserved" is relative to the conditions already in place when the model was constructed. That is, any current biases present in the data that influence farmers markets are likely captured by the model, passing that bias on to the expected frequencies. It could be entirely plausible that certain regions are underserved because of, for instance, socioeconomic factors which constrain the occurrences in farmers markets. This in itself may provide an opportunity to expand farmers markets beyond the current client base and identifies the variables as underserved, rather than the individual counties being underserved.

XII. Conclusion

This study aims to provide county recommendations for successful opening of new farmers markets, as well as resources to evaluate a given county, should there be interest in opening a new farmers market there. This is meant only as a starting point. It is recommended, should an interested analyst hope to start a farmers market without a location in mind, to look through the list to find a favorable location and use the available data for that county for cross reference. If one should already have a location in mind to begin a new farmers market, an analyst can look up the county on the table and determine whether the county is a productive environment for a new farmers market.

This study can be improved upon or studied further in several ways. First, new columns can easily be joined to the existing data table if one should further explore variables not already in the study. For example, one can join more economic data from the US Department of Labor

that may compliment the study. Another benefit of this study is that the structure of this project easily allows for the analysis of not only farmers markets, but nearly any variable that is a count. For example, one can use the columns of the data and the structure of the regression models to analyze the distribution of number of shopping malls by county given a representative sample.

Another area for potential study stemming from this project includes a more in-depth study of the factors that support farmers markets. For example, this study did discover a negative correlation between farmers markets and public health measures. These areas may benefit from more farmers markets and access to local and healthy foods. Therefore, a sociologist would be well-poised to investigate the social and political reasons for a saturation or scarcity of farmers markets.

IX. Appendix

Sources:

Wolf, Marianne & Spittler, Arianne & Ahern, James. (2005). *A Profile of Farmers' Market Consumers and the Perceived Advantages of Produce Sold at Farmers' Markets.* Journal of Food Distribution Research. 36.

Figueroa-Rodríguez, Katia A. & Álvarez-Ávila, María & Castillo, Fabiola & Rindermann, Rita & Figueroa, Benjamín. (2019). *Farmers' Market Actors, Dynamics, and Attributes: A Bibliometric Study*. Sustainability, no. 111, 3: 7.

Brittain, Jim; Cendon, Mariana; Nizzi, Jennifer; and Pleis, John (2018) *Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance.* SMU Data Science Review*:* Vol. 1: No. 2, Article 7.

Husson, F., Lê, S., & Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. Boca Raton: CRC Press Taylor & Francis Group. 12:18.

Hu MC, Pavlicova M, Nunes EV. *Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial.* American Journal of Drug Alcohol Abuse. 2011;37(5):367-375. doi:10.3109/00952990.2011.597280

McCarthy, R. (2010). *Evaluating the Social, Financial and Human Capital Impacts of Farmers Markets:* Market Umbrella. 1:10

Links to Data and R Notebook

Raw Data Files, including Farmers Market table, Election data, and USDA data

https://github.com/michiel1134/Capstone-Project/tree/master/Raw%20Data

Cleaned Data Sets

https://github.com/michiel1134/Capstone-Project/tree/master/Cleaned%20Data

Final Results, Predicted number of farmers markets along with all other county level features:

GitHub: https://github.com/michiel1134/Capstone-Project/blob/master/Cleaned%20Data/FarmersMarketPredictions.csv

Or try the Google Drive version:

https://drive.google.com/file/d/1tPlwEIDicgjfht7VXQUxFfU77Ze50XkZ/view?usp=sharing

R Notebook

Code:

https://drive.google.com/file/d/1RfMqBIaQHdhiLIULQdUyimzwpsw9B8IJ/view?usp=sharing

HTML version (download and use a web browser to open):

https://drive.google.com/file/d/1Gvr0GojxfwH2bgi8yckaSf5Oh5E0kLA4/view?usp=sharing