Spring 5-3-2023

A Probabilistic Exploration of Food Supplementation and Assistance

Logan Mattingly

Murray State University Honors College

HONORS THESIS

Certificate of Approval

A Probabilistic Exploration of Food Supplementation and Assistance

Logan Mattingly

May/2023

Jeffrey S. Young

Approved to fulfill the
requirements of HON 437

Dr. Jeffrey Young, Assistant Professor
Department of Agricultural Science

Approved to fulfill the
Honors Thesis requirement
of the Murray State Honors
Diploma

Dr. Warren Edminster, Executive Director
Honors College

Author: Logan Mattingly

Project Title: A Probabilistic Exploration of Food Supplementation and Assistance

Department: Department of Agricultural Science

Date of Defense: May 3rd 2023

Approval by Examining Committee:

________________________________ ________________

(Dr. Jeffrey Young, Advisor) (Date)

________________________________ ________________

(Dr. Naveen Musunuru, Committee Member) (Date)

________________________________ ________________

(Dr. David Gibson, Committee Member) (Date)

A Probabilistic Exploration of Food Supplementation and Assistance

Submitted in partial fulfillment
of the requirements
for the Murray State University Honors Diploma

Logan Mattingly

May/2023

Abstract

Food insecurity is a stark threat that grips our country and affects households throughout our country. Dietary insufficiency manifests itself in ways that affect health and public safety. According to researchers, individuals who suffer from food insecurity have a higher risk of aggression, anxiety, suicide ideation and depression. These problems tend to occur unequally distributed among those households with lower income. In this work, an exploratory analysis within these data sets will be performed to examine the socio-economic, biographical, nutritional, and geographical principal components of food insecurity among survey participants and how the US Supplemental Nutrition Assistance Program (SNAP) effects partakers of this study. Relevant statistical and algorithmic tools will be used such as Self organizing maps (SOMs) and hierarchical clustering will be used for cluster analysis in addition to logistic regression and random forests for propensity score matching. Final results show a positive effect on household wellbeing and increased food spending on SNAP participants.

Table Of Contents

1. Introduction.. 1

1.1 Studies Based on Nutritional Benefits of the SNAP Program 2

1.2 Studies on Applied Labor Effects of SNAP. 3

1.3 Purpose and Objective of Study 3

2. Methods 5

2.1 Database Acquisition, and Preparation 6

2.2 Cluster Analysis 10

2.2.1 Self Organizing Maps 10

2.2.2 Agglomerative Hierarchical Clustering 12

2.3 Supervised Classification and Propensity Analysis 14

2.3.1 Logistic Regression 14

2.3.2 Random Forest 15

2.3.3 Extreme Gradient Boosting 16

2.3.4 Performance Metrics 16

2.3.5 Propensity Score Matching 18

3 Results and Discussion 18

3.1 Unsupervised Cluster Analysis Results 19

3.2 Supervised Learning Analysis and Propensity Score Matching 24

3.3 Suggested Future Study 28

Table of Figures

Figure 1: SOM Heatmap of FoodAPS Households 19

Figure 2: FoodAPS Household Dendrogram 21

Figure 3: FoodAPS Household Clusters Gap Statistic 22

Figure 4:FoodAPS Households for k=3 23

Figure 5: ML Analysis ROC Curve 25

Figure 6: Propensity Score Effect on Cluster 1 26

Figure 7:Propensity Score Effect on Cluster 2 26

Figure 8: Propensity Score Effect on Cluster 3 27

Table of Tables

Table 1: SNAP Benefit Analysis Dataset 8

Table 2: Expected Values of Clustered Household Features 24

Table 3: Supervised ML Analysis Performance Metrics 25

Table of Equations

Equation 1: Sklearn MinMax Equation 10

Equation 2: SOM Weight Function 11

Equation 3: Agglomerative Hierarchical Clustering Function 12

Equation 4: Gap Statistic 13

Equation 5: Logistic Regression 14

Equation 6: Accuracy 17

Equation 7: Specificity 17

Equation 8: Sensitivity 17

1. Introduction

Food Insecurity is a far-reaching issue that has gripped American society throughout its country's history. According to the United Stated Department of Agriculture Economic Research Service, food insecurity has affected 10.4% of households over the last twenty years (Coleman-Jensen et al., 2022). This means that over the past two decades of American history, approximately thirty-six million households lack enough food to place on their table to adequately feed themselves. Additionally, 6.2% of these households contain children with 0.7% being considered to have very low security. Moreover, according to an article published by the Kentucky Department of Agriculture in 2016, 17% of Kentucky households were considered food insecure. (KDA, 2016) This equates to 743,000 people 222,000 of which were children. When such a high number of the population is facing food insecurity a solution must be applied. Naturally, this is in some part due to the forementioned statistics, the interpretation being low-income Americans simply aren't provided enough nutritional substance to thrive. Another flag that should insight change would be the latter mentioned effects that lack of food and economic purchasing power has on household health, mental wellbeing and labor incentives. While there are a variety of possible ways that one could seek to bridge the gap between households and proper nutrition most commonly the US government is looked to for aid. For the purposes of this study, USDA sanctioned assistance programs will be analyzed.

Nutritional supplemental assistance programs have been carefully sponsoring individuals and placing food on the tables of millions starting in the late 1930's. The first step in combatting food insecurity started in 1939 with the adoption of the Food Stamp Program (FSP) at the end of the Great Depression. This program was temporary until 1964 when President Johnson requested to pass legislation to make the FSP permanent. Most recently on 2008 the name Supplemental

Nutritional Assistance Program (SNAP) was coined. Over the years this program has received more funding as the United States need grows. But is the SNAP program successful in supporting the millions of hungry individuals in America and giving them a better quality of life.

1.1 Nutritional Benefits of the SNAP Program

Thus far much research has been done examining the underlying nutritional value in the SNAP Program. As stated by a study published by The Center on Budget and Policy Priorities in 2018 “The food assistance offered by SNAP helps with a modest benefit that may nevertheless make it easier for individuals and families to afford healthier food. SNAP benefits also free up resources that can be used on health-promoting activities and preventive health care by reducing what families must spend out of pocket on food.” (Carlson and Keith-Jennings, 2018) This fact is highlighted by the increased probability gained by low income families to be able to afford better healthcare and increased food expenditure shown in figures five and six in the aforementioned work. Additionally, Marianne Bitler writes in the University of Kentucky Center for Poverty Research Discussion Paper Series, “There is evidence suggesting SNAP recipients spend more on food than other similar families and that they have higher nutrient availability than others.” (Bitler, 2014) They continue this line of thinking later by stating, “Evidence discussed [above] about the introduction of the program as part of the War on Poverty however does find that introduction of SNAP in one’s county increases spending on food and decreases spending on food away from home.” (Bitler. 2014). Taking a more econometric and statistical approach by modeling the marginal utility of SNAP participants Gregory and Deb state, “Extra income frees up SNAP participants to participate in non-diet-related activities that nonetheless improve their wellbeing. Many kinds of recreation–including but not limited to exercise–might be the agent in this case. Or SNAP could help to relieve stress that includes but goes beyond that associated with food insecurity.” (Gregory and Deb, 2014)

1.2 Applied Labor Effects of SNAP.

An additional angle of study has been in the complementary relationship between SNAP and labor supply and employment decision making. Using empirical linear models as an estimator, Farkhad found that, “Contrary to the perception that SNAP significantly reduces incentives to work, we find that SNAP increases the likelihood of employment among low-income households. In addition, we find that SNAP increases the probability of working full time. It is likely that higher labor supply among SNAP households is driven by work requirements imposed in SNAP and the ability to afford job related expenses such as childcare” (Farkhad. 2018) In looking at the barriers to entry and exit in relation to the SNAP program and labor Gray, Leive, Prager, Pukelis and Zaki find, “SNAP work requirements dramatically reduce participation among affected adults, with point estimates suggesting a fifty three percent decline in participation by the completion of the roll-out. In practice, work requirements appear to screen out a many potential SNAP beneficiaries in exchange for possible earnings increase among a limited subset of individuals.” (Gray et al., 2021) Papers such as these illustrate the effects employment has on SNAP participation and the contrasting relationship therein.

1.3 Purpose and Objective of Study

However, among these writings it remains unclear the actual effects that SNAP participation has on key variables within household home economics. Few have investigated how food assistance may increase or decrease financial strain outside of food expenditure. It remains to be seen the actual financial impact SNAP has on impoverished households within the US. In addition, due to the non-homogeneity of socioeconomic standing among American SNAP participants, one could infer these findings would vary accordingly for various demographics. This is in part due to the large amount of expected variance among many independent variables present in day-to-day life. Truly these questions remain at large and largely valuable given the key statistics provided above.

With that being said, the purpose of this study is to examine the economic effects of SNAP participation among impoverished households in a statistical fashion while accounting for the natural variances among their socioeconomic structures. This will be done by taking a multi-tiered approach that utilizes both supervised and unsupervised machine learning algorithms and relevant statistical models in tandem to prior data wrangling Extraction, Transfer and Loading (ETL) processes.

The data section process consists of a large amount of SQL and python database creation and querying. This was done to build a clean dataset that contains only the statistically important features[1] and to engineer other useful variables for further analysis. Unsupervised methods of understanding and clustering data to help understand and account for the non-normal distribution of key data. Tools such as Self Organizing Maps (SOMs) will be utilized to gather the general distribution of key variables such as income, USDA food security score and household size. Additionally Hierarchical Clustering will be used to form separate samples that contain normal distributions for independent variables. Once the unsupervised clustering is complete, a more traditional machine learning analysis will take place to determine the predicted effects of SNAP usage within households. This will be done using notable models such as logistic regression, random forest, and gradient boosted tree algorithms. Ultimately, these models will be used in propensity score matching and analysis by means of the logits and probability of class conformity within the newly formed clusters.

[1] Often within the data science domain, variables within a dataset are referred to as features. Due to the nature of being within said domain, the word feature is synonymous to the common term variable

The remainder of this paper will be organized as follows. Dataset, data cleaning, unsupervised clustering methods, supervised classification and analysis methods, analysis results and visualization. Finally, the paper will close with a brief discussion and takeaways from the research and suggestions for further study.

2. Methods

With such a large scope of research that this work could pertain, it was decided that this study would look at the financial and economic effects of SNAP participation within low income households. The mission behind such research is to find the components in which SNAP benefits users that later play into increased wellbeing rather that determining whether quality of life is improved one way or another

For the purposes of this study, a probabilistic machine learning approach was used. These mathematical and statistical methods can be used to determine unseen patterns and hidden correlations within the data provided. The data science pipeline will be followed to gain the most statistically significant and predictively powerful results. This is not to discount normally applied econometric methods, however it was decided that the utilization of the streamlined data science pipeline had the ability to yield increasingly more in depth and precise results by utilizing the strength found in the probabilistic approach of statistical analysis and the accuracy found in algorithmic computation.

Data will be selected from the USDA FoodAPS database and will be cleaned in a hybrid approach, combining relevant SQL, Pandas and NumPy functions alongside human interaction within the Dataframe. In order meet the statistical requirements for regression analysis, the data will be tested and transformed into a normalized, independent dataset, in which each feature has a constant mean and standard deviation.

At the beginning of the analysis, unsupervised machine learning algorithms such as Hierarchical Clustering, and SOMs will be applied to allow for relevant biographical clusters to be formed among participants. Additionally, once the clusters are formed and new features can be extracted via data engineering. Finally, traditional supervised ML procedures such as Logistic Regression and Random Forests will be used for food insecurity classification, prediction, and Propensity Score Matching to effectively explore the probabilistic effects for UDSA SNAP program enrollment and American food insecurity. Once these statistical analyses are completed, relevant data and results will be presented using customized visualizations, and relevant machine learning performance metrics.

2.1 Database Acquisition, and Preparation

Within this work The National Household Food Acquisition and Purchase Survey (FoodAPS) database published by the USDA would provide the must complete set information available. It must be noted that due to the Foundations for Evidence-Based Policymaking Act of 2018, only data set aside for public use could be utilized. This is for the protection of internal survey participants in relation to the personal information that is contained within the final data set. Nevertheless, FoodAPS is a collection of food purchases, prices, and nutrient characteristics of households across the country. Additionally, this dataset contains financial and biographical information on all households and individuals who took part in the USDA's purchase survey. Its creation is credited in the following way, "The FoodAPS data collection was sponsored by the U.S. Department of Agriculture (USDA) and managed by USDA's Economic Research Service (ERS) with support from USDA's Food and Nutrition Service (FNS). Due to special interest in the food acquisition patterns of households participating in these programs, the survey oversampled low-income households, both those receiving Supplemental Nutrition Assistance Program (SNAP) benefits and those not receiving SNAP benefits. The survey is weighted to be representative of all non-institutionalized households in the continental United States." (USDA ERS, 2016) In accordance with the expected distributions of each of the US racial and geographical groups, 4,826 households were chosen with proper weights set in place to receive a statistically significant sample for further analysis. To this survey and its analysis reflected in this study, the term household can be defined as all persons who live together and who were expected to attend the sampled address during at least part of the data collection period the USDA ran. Additionally, the randomized sample of households that utilize SNAP benefits was set at 1,500. This means that initially, this study starts with a sample size of 1,500 households who partake in SNAP benefits.

The FoodAPS database contains a few sub tables that hold a variety of informational datasets that contains individual and household food consumption, biographical and economical information. These tables are as follows, Household Survey Responses Individual Survey Responses, Food at Home Consumption (FAH), Food Away From Home Consumption (FAFH), Food at Home Meal Item, Food Away from Home Meal Item, Food at Home Nutrient Consumption, Food Away from Home Nutrient Consumption, Total Household Weights, and Daily Household Meal Count.

Once the full FoodAPS data was completely extracted in a .csv format. It was necessary to transfer the needed tables into a Microsoft SQL Server. This can be seen in the later appendix. The complete ETL process required sets of SQL queries that included creating necessary tables and engineering new features such as "average food expense" by taking quantitative meal count and comparing that to cumulative FAH and FAFH event costs. Once all relevant features were transferred and extracted the final step in the ETL process was loading the completed table into an external analysis software to begin the data cleaning and wrangling process. The preferred software suite for other data preprocessing would be a Python IDE named Spyder and RStudio.

The data validation process included three main steps. These were missing value detection, continued dimensionality reduction and finally data normalization. The first step in this process was finding all values that either contained no entry or held values that did not pertain to the problem at hand[2].. Self-designed python code was used to weed out these values to only analyze noteworthy data. Next, insignificant features were taken out of the dataset to produce more accurate results. This was done by both manually selecting features that pertain to the area of the study as well as using IBM SPSS Modeler significance testing. Features that had a statistical importance level lower than 0.9 were removed.

Once this process was completed the resulting dataset was reduced to 40 features and 738 instances. Table 2.1 highlights each feature below.

Table 1: SNAP Benefit Analysis Dataset

Feature Name	*Data Type*	*Description*
Household Number	*Categorical*	*Given Household Identifier*
Non-Metro	*Binary*	*Urban Identifier*
Rural	*Binary*	*Rural Identifier*
Region	*Categorical*	*US Geographic Region Identifier*
Household Weight	*Continuous*	*Physical sum of weight per household*
Residence Size	*Ordinal*	*Size of all inhabitants present*
Household Size	*Ordinal*	*Size of household*
Household Average Income	*Continuous*	*Moving average of income per month*
Self Employed	*Binary*	*Self-employed identifier*
Self Employed Food Prep	*Binary*	*Food prep identifier*
Job Change	*Binary*	*Flag for job change in six months*
Earn Less	*Binary*	*Decreased income flag*

[2] These values were often flagged by the data collection service as “-997” or “-996”

Earn More	*Binary*	*Increased income flag*
Earn Same	*Binary*	*No change in income flag*
Public Housing	*Binary*	*Public housing flag*
Subsidized Housing	*Binary*	*Subsidized housing flag*
Liquid Assets	*Ordinal*	*Amount of liquid assets identifier*
Any Vehicle	*Binary*	*Vehicle ownership flag*
Vehicle Number	*Ordinal*	*Amount of vehicles owned*
Car Access	*Binary*	*Car access flag*
Large Expenses	*Binary*	*Large expenses expected per month*
Mortgage Expense	*Continuous*	*Monthly mortgage expense*
Insurance Expense	*Continuous*	*Monthly insurance expense*
Property Tax Expense	*Continuous*	*Monthly property tax expense*
Public Transportation Expense	*Continuous*	*Monthly public transportation expense*
Electric Expense	*Continuous*	*Monthly electric expense*
HVAC Expense	*Continuous*	*Monthly heating and air expense*
Waste Management Expense	*Continuous*	*Monthly trash disposal expense*
Health Expense	*Continuous*	*Monthly health insurance expense*
Copay Expense	*Continuous*	*Monthly copay expense*
Doctor Expense	*Continuous*	*Monthly doctor or hospital expense*
RX Expense	*Continuous*	*Monthly medicine prescription expense*
Child Care Expense	*Continuous*	*Monthly childcare expense*
Child Support Expense	*Continuous*	*Monthly child support expense*
Average Food Expense	*Continuous*	*Moving average of monthly food cost*
Total Expense	*Continuous*	*Sum of all expenses*
Cashflow	*Continuous*	*Remaining liquid cash after expenses*
SNAP Benefits $ Amount	*Continuous*	*Value of SNAP benefits used*
Food Security Score	*Ordinal*	*USDA food security score identifier*

Once the dataset was completed, the final step was to normalize the dataset to fit within the statistical rules of *iid0.* Thus, datasets that will be used for analysis consist of independent variables with a covariance of zero and must be identically distributed with a mean of zero. Independence was tested and determined within the feature selection process and normalization was performed using the python library Sklearn MinMaxScaler function. Once this operation was completed the final step of the data preprocessing stage was completed. Thus, the dataset was ready to be analyzed using cluster analysis.

$$x = \frac{(x_i - min_x)}{(max_x - min_x)}$$

Equation 1: Sklearn MinMax Equation

2.2 Cluster Analysis

Cluster analysis is the process of grouping relevant data points in a dataset to be able to better understand the data. In their 2001 publication entitled “Cluster Analysis”, Everitt, Landau, and Leese state, “Cluster Analysis is a generic term for a wide range of numerical methods for examining multivariate data with a view to uncovering or discovering groups or clusters of homogenous observations” (Everitt, Landau, and Leese, 2001) By using these processes, one is in a way able to shrink down a much larger data filled hyperplane into a two- or three-dimensional field for more understandable analysis. Within this study, SOMs and Agglomerative Hierarchical Clustering methods were chosen. Due to the large scope of the multivariate Dataframe, these methods will help explain the data and make further analysis more straightforward and accurately targeted.

2.2.1 Self Organizing Maps

Self-Organizing Maps (SOMs) are a type of neural network that is commonly used in data analysis and machine learning applications. It is an unsupervised clustering algorithm that is used to visualize and analyze complex and high-dimensional data. SOMs have many applications in data analysis, such as data compression, image recognition, and data visualization.

At a high level, SOMs are designed to mimic the way the brain works by creating a two-dimensional grid of neurons that can learn and adapt to the input data. The neurons in the grid are connected to each other and are organized in a way that allows them to identify patterns and similarities in the input data. While this act is like a traditional supervised Artificial Neural Network (ANN), SOMs use a different type of learning rule than traditional neural networks. SOMs use a competitive learning rule, where the neuron with the closest weight to the input data is activated and its weights are updated. Contrastingly, ANNs typically use a backpropagation algorithm to update the weights of the neurons in the network based on the error between the predicted output and the actual output.

When training a SOM, the algorithm begins by randomly initializing the neurons' weights. Then the algorithm presents the input data to the SOM and adjusts the weights of the neurons based on how well they match the output per epoch. The neurons that are closer to the input data are updated more than the neurons that are farther away, which allows the SOM to identify clusters and patterns in the input data. Accordingly, when discussing the weighting and ultimately the clustering technique of this algorithm, Everitt, Landau, and Leese state “Clustering occurs where an input vector is assigned to an output node. Operationally, each output node has a *p*-dimensional of synaptic weights **w**. The output node is initially assigned a random weight; as the network learns, the input cluster points are provisionally assigned to clusters and weights are modified”. (Everitt, Landau, & Leese. 2001) Overtime as the network trains itself, more accurate weights are generated and datapoints are placed into the “winning” cluster.

$$w_{new} = w_{old} + \alpha(x_i - w_{old})$$

Equation 2: SOM Weight Function

After training, the SOM creates a map of the input data in two dimensions. Each neuron in the SOM represents a specific area of the input data and is responsible for classifying new data into one of the categories it has learned. The map can be visualized as a topographical grid, where each neuron represents a specific location, and the colors or other visual elements represent the data's features.

2.2.2 Agglomerative Hierarchical Clustering

Hierarchical Clustering is a data analysis technique used to identify groups or clusters within a dataset. It is a type of unsupervised learning algorithm that does not require any prior knowledge or labels about the data. One advantage of hierarchical clustering is that it can reveal the hierarchical structure of the data, which can be useful for understanding the relationships between different groups or subgroups.

The algorithm starts by treating each data point as a separate cluster and then iteratively merges similar clusters together to form a hierarchical tree-like structure called a dendrogram. Agglomerative hierarchical clustering starts with each data point as a separate cluster and then merges the most similar pairs of clusters together, continuing until all data points are in the same cluster. The similarity between pairs of data points is typically defined by using a distance metric, in this case Euclidean distance, which measures the distance between two points in a multi-dimensional space. Once the dendrogram is formed, the algorithm can be used to identify clusters at different levels of granularity.

In reference to hierarchical clustering Everitt, Landau, and Leese discuss a so-called flexible clustering method that is used in this study that is defined by values of parameters of a general recurrence formula that gives the distance between groups for each set of points *i and j* in each dendrogram.

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|$$

Equation 3: Agglomerative Hierarchical Clustering Function

Accordingly, once the full tree has been computed, various statistical tests can be performed to determine the optimal number of clusters if not already specified. Some of these tests are the Elbow Method, The Silhouette Method, the Calinski-Harabasz Index, and the Gap Statistic. Within this study the Gap Statistic was used.

The Gap Statistic is a statistical method used to determine the optimal number of clusters in a dataset. It works by comparing the within-cluster sum of squares (WSS) for the original dataset to the WSS for artificially generated datasets with different numbers of clusters. The optimal number of clusters is the one that maximizes the gap between the two WSS values. The Gap Statistic method is based on the idea that a good clustering solution should have a WSS value that is much smaller than what would be expected by chance. Therefore, the method first generates a set of uniformly distributed random data points that cover the same range as the original dataset. It then computes the WSS for each number of clusters using both the original dataset and the random dataset.

$$gap(k) = E\left(log\left(WSS_{random}(k)\right)\right) - log\left(WSS_{x}(k)\right)$$

Equation 4: Gap Statistic

Unsurprisingly, the optimal number of clusters is the one that maximizes the Gap Statistic, where larger values of the Gap Statistic indicates a better clustering solution. The method also provides a measure of uncertainty in the form of a confidence interval, which can be used to determine whether the Gap Statistic is significantly different from what would be expected by chance. Henceforth, The Gap Statistic is a useful method for determining the optimal number of clusters in a dataset because it considers the size and complexity of the dataset and provides a measure of uncertainty and randomness that can be seen in the natural world.

As mentioned above, the exact applications of hierarchical clustering within this study will be discussed at length in a later section. However, it should be known that this unsupervised statistical tool was vital in understanding underlying relationships between each datapoint and allowed for clustering which benefits later analysis.

2.3 Supervised Classification and Propensity Analysis

Once the forementioned cluster analysis was completed, the next step was to perform a classification analysis that will investigate the probabilistic effects of SNAP benefits within the chosen sample. Within this study three models were chosen and compared in terms of performance in order create the strongest and most accurate results in this probabilistic exploration of SNAP benefits. These models were Logistic Regression, Random Forest Classification and XGBoost, a gradient boosting algorithm. Explanations of these algorithms can be seen below.

2.3.1 Logistic Regression

Logistic regression is a common adaptation of the well-known linear regression model used in regressive continuous estimation and the log-odds ratio used to determine the probability of a particular contentious random variable being a member of a discrete class K via the linear function as *x*. (Hastie et al., 2017) Due to the statistical assumption of logistic regression the dependent variable must be distributed in a logistics fashion. If said target is a continuous or normally distributed variable the algorithm loses its predictive power and is unable to perform in the expected way. This method of classification is widely used due to the binary nature of most diagnosis and can be computed using the function below.

$$P(y = k|X = x_i) \frac{1}{1 + \sum_{l}^{k-1} exp\left(\beta_{l0} + \beta_l^T x\right)} \text{ , } k = 1, \dots , K - 1$$

Equation 5: Logistic Regression

This function states that the probability of the given class *k* given the instances x_1 to x_i is equal to the exponentially transformed log odds ratio of all the given instances of a particular class. This allows for the proper classification of all data entries for their respective classes.

To fit a logistic regression model, the coefficients β_0 to β_{n-1} are estimated using maximum likelihood estimation, which involves finding the set of coefficients that maximize the likelihood of observing the data given the model. This involves iteratively updating the coefficients until convergence is reached. Once the coefficients are estimated, they can be used to make predictions about the probability of the dependent variable taking on the value of 1 given the values of the independent variables. These results are reported in log-odds which than then be solved for the probability of an instance belonging to a set class the predicted probability can then be used to make a binary classification by setting a threshold value, usually 0.5 or greater, the dependent variable is classified as 1 and below which it is classified as 0.

2.3.2 Random Forest

Additionally, Random Forests (RF) were another form of algorithmic classification used in this study. Traditionally RF is known for its high capacity as a classifier and high performance. This widely used model is derived from decision tree models. However instead of a simple one tree algorithm, RF creates *n* trees and the classification label with the most "votes" among the trees built by the model is chosen as the predicted output. Specifically, RF trees use a selection of features that have low to no correlation to preformed non biased calculations. As trees are created the algorithm can learn the data and create and adapt the proper variable coefficients to fit a more accurate model over time. Due to this powerful design RF maintains its usage among large datasets and is commonly used as a baseline model for a verity of classification applications.

Random forests have several advantages over single decision trees. They are more robust to noise and overfitting, and they can handle many features without requiring feature selection. Additionally, random forests can provide estimates of feature importance, which can help identify the most relevant features for making predictions. Overall, Random Forests ability to handle complex data and provide accurate predictions makes them a powerful tool for a wide range of applications.

2.3.3 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is a powerful and popular machine learning algorithm that is used for supervised learning tasks, such as classification and regression. It is an ensemble learning algorithm that combines multiple weak learners, decision trees, to form a strong learner that can make accurate predictions on new data. The XGBoost algorithm works by iteratively adding decision trees to the ensemble model while minimizing the error between the predicted and actual outcomes. Each decision tree is constructed to predict the residual errors, the difference between the predicted and actual outcomes, of the previous tree. This process is known as boosting.

In addition to boosting, XGBoost also includes regularization techniques to prevent overfitting, such as L1 and L2 regularization, and it uses a gradient descent algorithm to optimize the objective function. The XGBoost algorithm is known for its speed and accuracy, and it is widely used in a variety of applications. Its ability to handle large datasets with many features and to automatically handle missing data and outliers make it a versatile tool for different types of data analysis.

2.3.4 Performance Metrics

Once the models produced a classification prediction, performance metrics were used to interpret the reliability and accuracy of the given predictions. Due to the nature of the situation within this present study, four metrics were chosen accuracy, specificity, sensitivity, and AUC score. In these formula TN, TP, FN, FP are the true positive, false positive, true negative and false negative classifications that each model makes the predicted classifications ate testes against the testing data. The former can be defined as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 6: Accuracy

$$specificity = \frac{TN}{TN + FP}$$

Equation 7: Specificity

$$sensitivity = \frac{TP}{TP + TN}$$

Equation 8: Sensitivity

Accuracy considers all true classification versus all classification, both true and false. Specificity tracts the cumulative accuracy of negative classifications while resistivity accounts similarly for positive classifications.

AUC is a score that ranges from 0 – 1 which determines the area under the linear model of the function that is created given TP and FP rate. Traditionally, these models are shown on a two-dimensional plane. Overall, AUC serves as another form of observing the predictive nature of the model predictions. This metric is usually displayed visually using an ROC curve and can compare models graphically as well as numerically by recording an AUC score. Due to its ability to understand the underlying model formulation, AUC will be a key index to assess model performance. However, other performance results will be presented to support these statistics.

2.3.5 Propensity Score Matching

This statistical method allows for the probabilistic understanding of the effect certain features in a data set have on others in a tangible way (Rosenbaum & Rubin,1983). A propensity score is the estimated probability of being assigned to the event group based on their observed characteristics. Propensity Score Matching relies on the assumption that, given some observable characteristics, units untouched by a certain event can be compared to units that are effected by said event, as if the event was fully randomized. In this way, Propensity Score Matching seeks to mimic randomization to overcome issues of selection bias that plague non-experimental methods.

Propensity score matching involves creating a matched comparison group by matching individuals in the treatment or event group those in the control group who have a similar propensity score, which is the estimated probability of being assigned to the event group based on their observed characteristics. The propensity score is estimated using a logistic regression model or other models that have a probabilistic output, where the dependent variable is treatment assignment, and the independent variables are the observed characteristics of the individuals. By using this test, one can determine the effects of the event in question on chosen variables. Additionally, this can be visualized for further analysis. In this study, propensity score matching will be used to determine the effect SNAP usage has on households and chosen variables such as expenses, household weight and other quality of life factors.

3 Results and Discussion

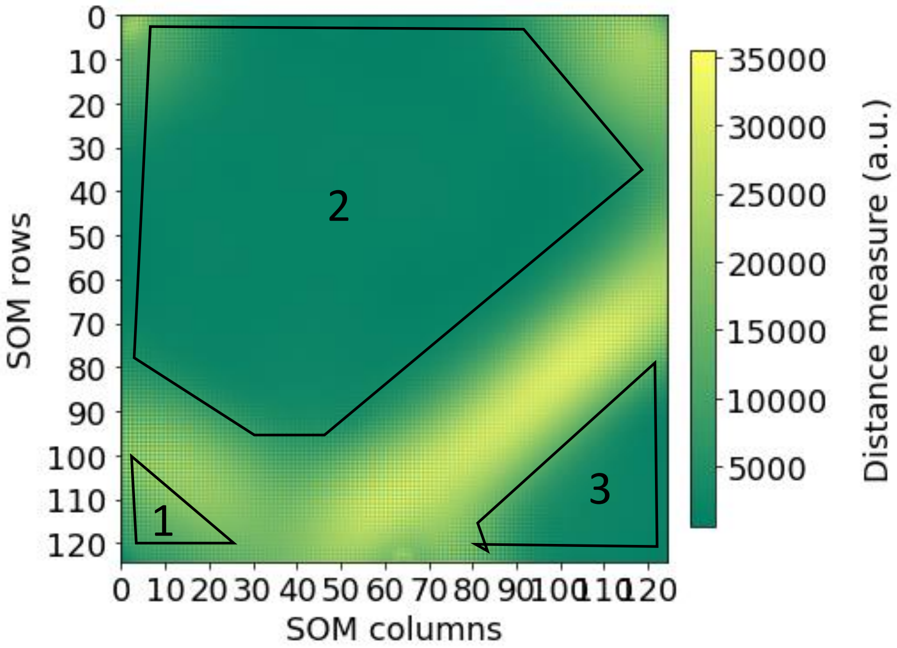
Within this section, the results of the forementioned analysis will be discussed. Both the findings of the unsupervised cluster analysis as well as the supervised classification and propensity score matching will be documented. In doing this analysis both the statistical software

RStudio and Python were used, and the full code will be cited in the appendix. Additionally, graphical results will be published as visualizations created by these software tools.

3.1 Unsupervised Cluster Analysis Results

The targeted goal of the SOM analysis was both a rough cluster of all data points to see the hidden relationships between instances as well as targeted clustering based off key variable to understand feature distribution topographically. Figure 3.1 shows a cluster of the given SNAP analysis dataset at 16% resolution. Due to computational constraints, a 120x120 network was trained rather than a 700x700. This would simply be impractical and would require substantial computational power that was not available during the time of testing.

Figure 1: SOM Heatmap of FoodAPS Households



This visualization shines a light on the similarities and differences of households within the chosen survey. According to the distance measurement spectrum on the right, there appear to be 3 distinct clusters of households present on the dataset. One large cluster directly in the center of the heatmap, a smaller one in the right corner and an even smaller group formed on the left.

Additionally, it seems that the distance between clusters 3 and 2 is greater than 1 and 2 while 1 and 3 are the most divorced from one another. While further quantitative analysis will be shown later it should be noted that cluster 2 can be seen as the statistical median for feature distribution within this study. Meaning that the expected values of all features in cluster 2 create a type of "middle class" economic structure within the chosen sample. In the same way, clusters 1 and 3 pose as outliers from cluster 2. Cluster 1 is on the higher end of expected variable values while cluster 1 is on the lower end. This creates a economic hierarchy with that shows an decreased economic state provides increased SNAP benefit usage.

A hypothesis could then be created stating within this dataset there lies three distinct groups of SNAP benefit partakers. This means that there could be three distinct distributions within the class variables each with different ranges and means and variances. However due to the more visual approach that SOMs take it is difficult to distinguish which cluster an individual household falls into and which numerical distribution each household's independent variables fall into. Therefore, a more quantitative analysis must take place to better validate and pinpoint accurate results. Thus is the necessity of hieratical clustering in this problem.

By using this unsupervised model in conjunction with RStudio, one can build a complete dendrogram and surmise which instances belong to a given cluster within the set *k* clusters parameter. In this case *k* can be set manually but as mentioned above the Gap Statistic was used to determine the optimal number of clusters due to the unknown amount within the data set. In addition, this protects the results of the clustering model from confirmation bias due to the previous analysis. However, first the dendrogram must be constructed before any analysis can take place. Figure 3.2 shows the completed dendrogram.

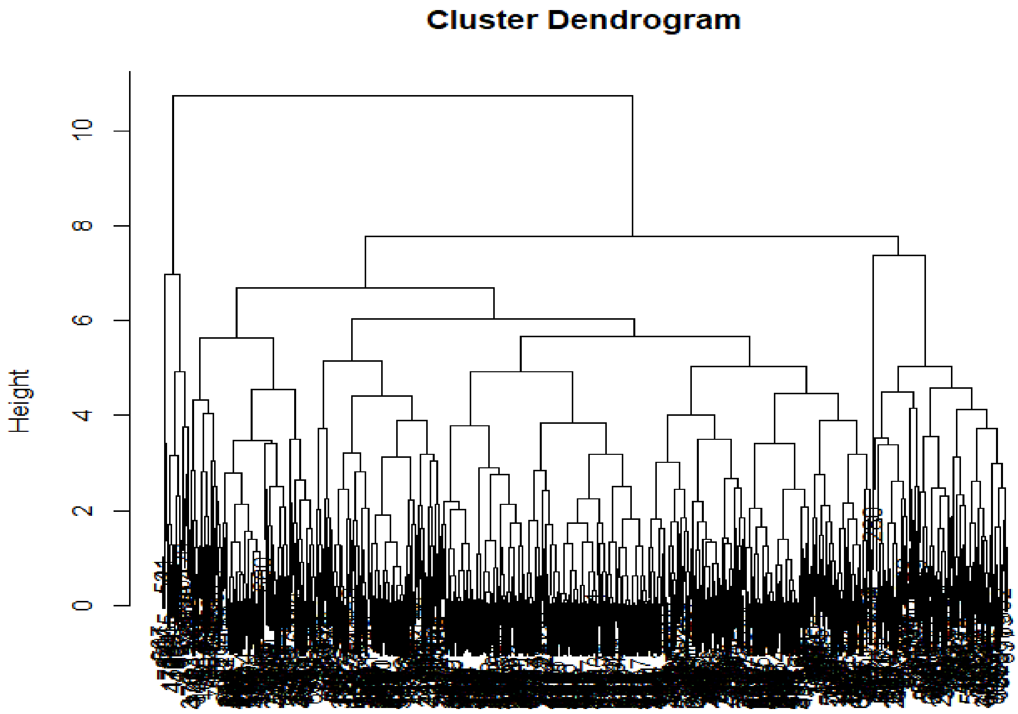


Figure 2: FoodAPS Household Dendrogram

When looking at the dendrogram several trees and branches are formed in a way that makes distinguishing the optimized number of clusters very difficult. Due to this the Gap Statistic Test was ran. As mentioned above the actual WSS is plotted against an artificial WSS formed from a randomized dataset synthesized using the mean and variance of the original dataset. This plot can be seen in Figure 3.3

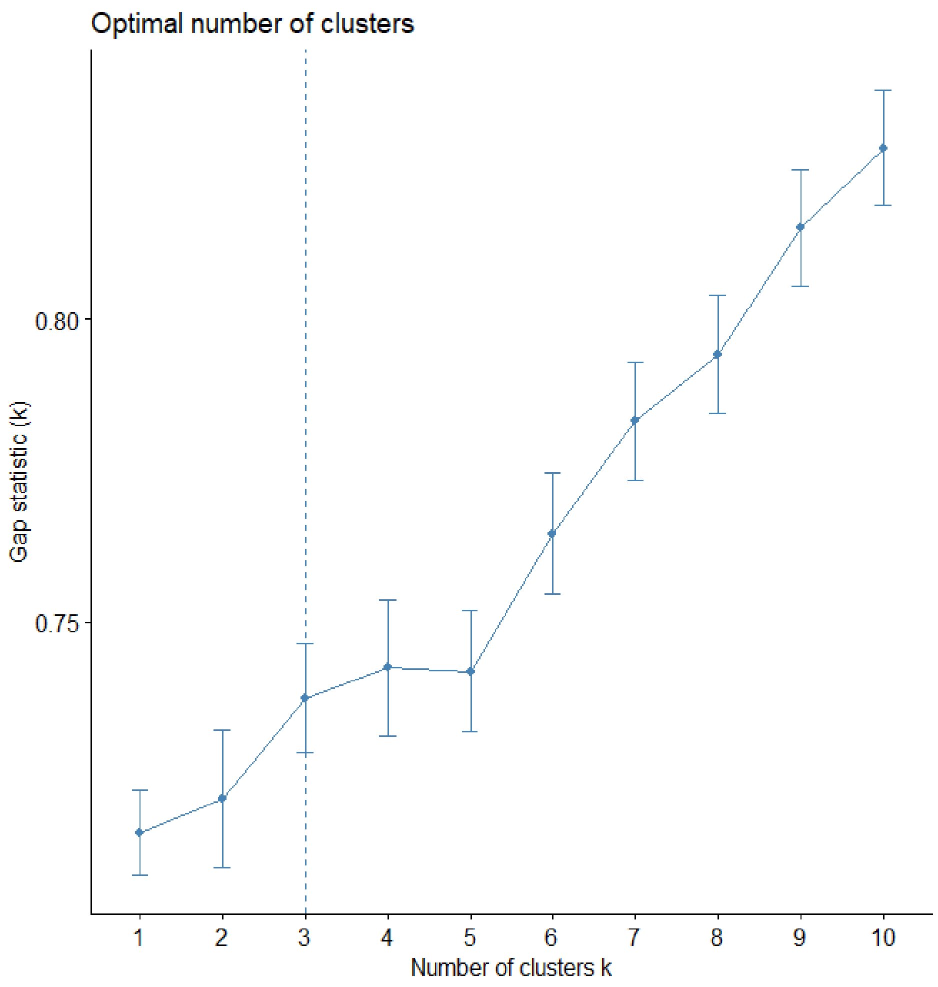


Figure 3: FoodAPS Household Clusters Gap Statistic

In accordance with the figure above, it was determined that when k = 3 the optimal number of clusters where formed. Once this number of clusters was chosen, the dendrogram could be properly divided into clusters for easier visualization and analysis. Additionally, using statistical software the expected values and other key informational statistics can be calculated for each unique distribution within the data.

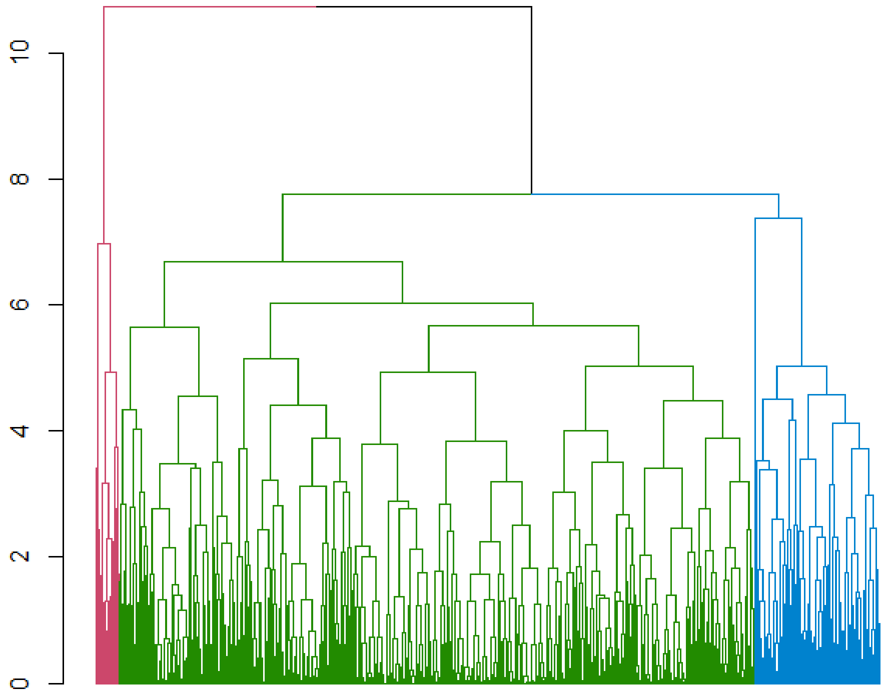


Figure 4:FoodAPS Households for k=3

Once each cluster was formed, and household were accordingly grouped, the mean values of each feature were calculated. The data shown in Table 3.1 highlights the key features in this study and the according calculations and allows for the variance between each cluster to be determined. As can be seen, cluster 1 records the highest income and the highest weekly expenses within the study. This group also contains the household with the highest food security thus the need for SNAP benefits is lowered. Cluster two contains the lowest income households with the larger need for SNAP benefits. Interestingly, this is the largest sample presented, taking up 80% of the overall study population. Finally, the third cluster claims the highest usage of SNAP benefits, but records double the monthly income when compared to cluster 2. While no educational or racial records are provided within the FoodAPS dataset, it could be suggested that these factors play a part in this finding due to existing data provided by the USDA ERS. With these results of the cluster analysis finalized and analyzed, these findings can be applied to the supervised machine learning analysis and propensity score matching testing.

Table 2: Expected Values of Clustered Household Features

Cluster	*n*	*% of Sample*	*Income*	*Total Expenses*	*Food Expense*	*Cashflow*	*SNAP Benefits*	*Food Security*
1	23	3%	\$12,526[3]	\$2,250	\$128.33	\$10,276.0	\$58.09	1
2	595	80%	\$2,404	\$1,055.1	\$110.68	\$1,349.4	\$132.7	2
3	119	16%.	\$4,769	\$1,515.50	\$120.26	\$3,254	\$157	2

3.2 Supervised Learning Analysis and Propensity Score Matching

Using python, a Logistic Regression, Random Forest and XGBoost model was trained to the dataset and tasked with predicting if a given household would revive SNAP benefits. Once the training was complete, each model was tested against a accordingly partitioned testing dataset, and these findings were recorded using ML performance metrics, Accuracy, Sensitivity Specificity and AUC. These metrics are recorded below in Table 3.2. The resulting ROC curve will also be provided.

Model	*Accuracy*	*Specificity*	*Sensitivity*	*AUC*
Logistic Regression	**0.77**	0.74	**0.8**	**0.88**
Random Forest	0.77	**0.82**	0.73	0.78
XGBoost	0.75	0.81	0.71	0.75

[3] Notice that expected income and cashflow seem large according to SNAP eligibility standards. No indication on why the recorded income of SNAP participants within this assigned cluster is shown within the FoodAps database documentation. However, it is hypothesized to be caused by clerical data entry errors or a miss interpretation of the variable time frame by the survey participants.

Table 3: Supervised ML Analysis Performance Metrics

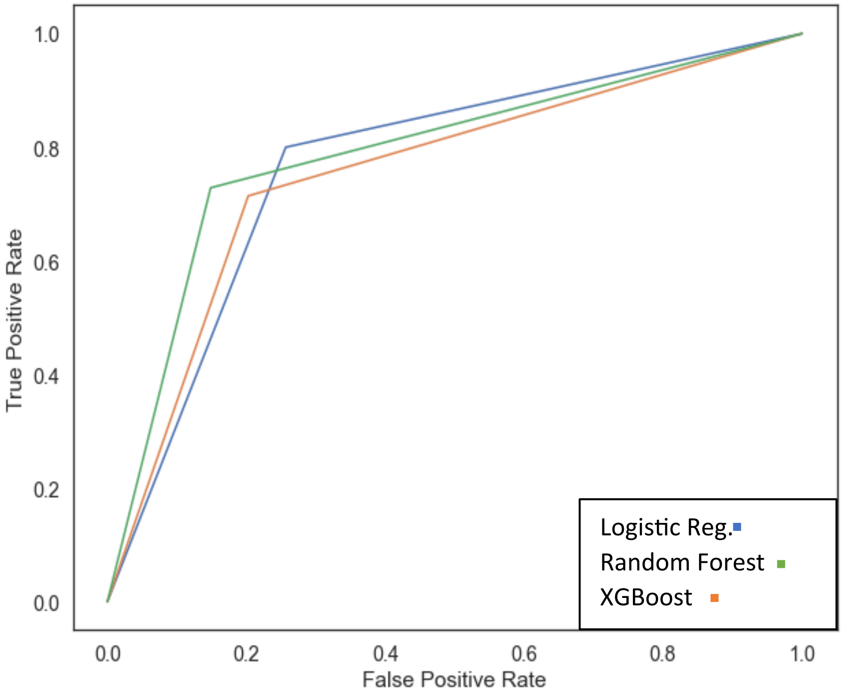


Figure 5: ML Analysis ROC Curve

Given the results above, Logistic Regression proved to be the best fit for the dataset at hand. It had the highest accuracy, AUC, and sensitivity scores, which proves that it is best fit to handle the complexity of the FoodAps data. Therefore, Logistic Regression was chosen as the model used in the propensity score matching.

Given that propensity score matching determines the effects that an event has on a set of dependent variables, this was the prime tool in answering the question this study poses. Additionally, being that three distinct populations were uncovered in past analysis; each cluster was tested and scored separately. Each group resulted in three distinct sets of effects that SNAP has on the survey participants. The figures below display the effects that SNAP has on each clustered group. Within each visualization below, many variables have a decreased propensity effect which translates to an optimized outcome effect of SNAP Benefits in relationship to overall cashflow.

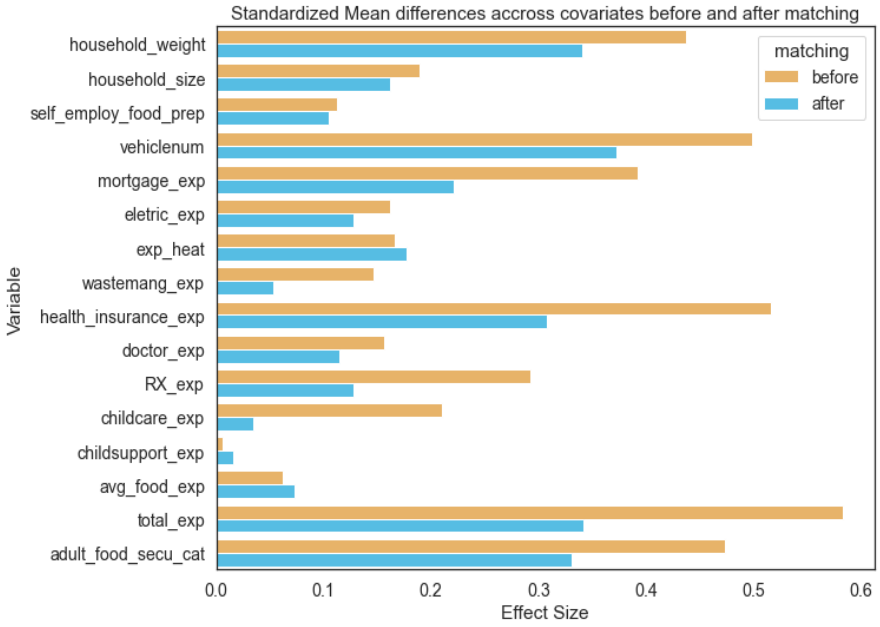


Figure 6: Propensity Score Effect on Cluster 1

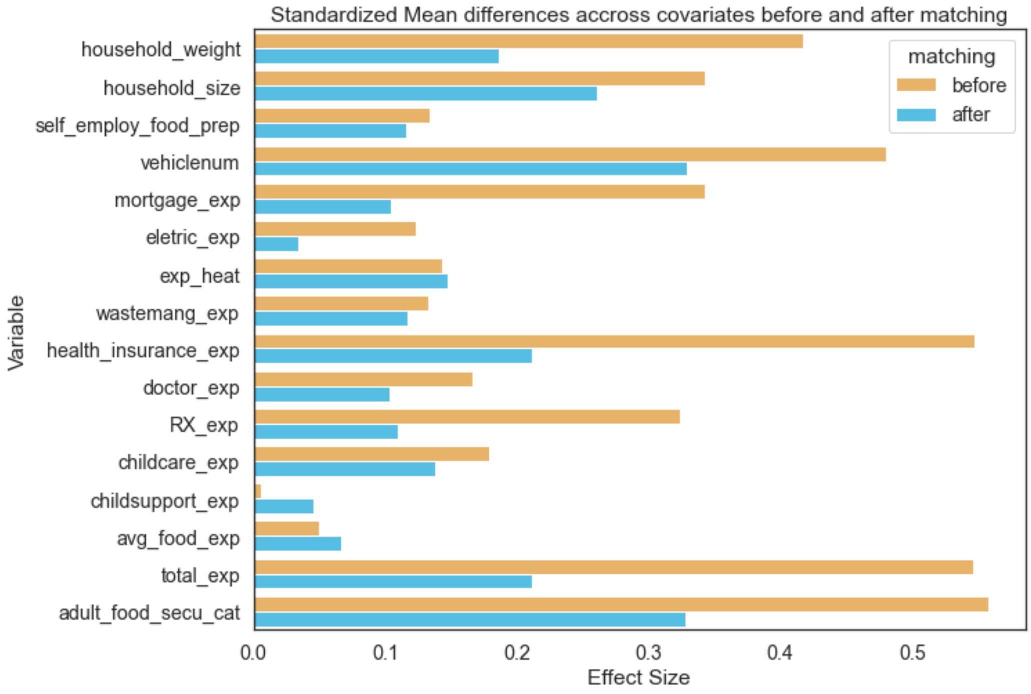


Figure 7:Propensity Score Effect on Cluster 2

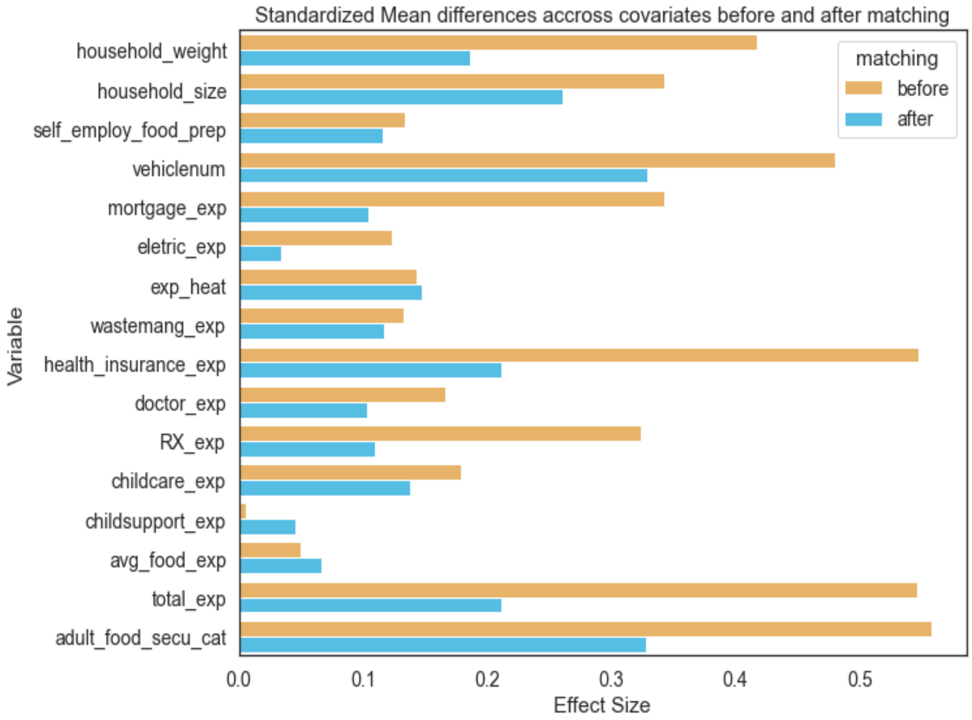


Figure 8: Propensity Score Effect on Cluster 3

As can be seen, each cluster has a unique response to SNAP benefits. Cluster three shows the most change overall while SNAP benefits in clusters one and two have lower yet still significant effects. Being that group there is considered the most at risk of food insecurity and therefore partakers the most in SNAP benefits, having higher effects seems reasonable. It should be noted that there also seems to be an effect on household weight which leads to the assumption that more nutritional food can now be purchased and consumed according to the previously reviewed literature.

Therefore, it can be concluded that SNAP benefits have a static positive relationship with the economic variables in individual's lives. This is highlighted in increased cashflow for external expenses, a more sustainable pathway for food consumption and decreased household weight. It is important to note that this analysis is only able to mesure the static conomic effects of rough;y one month on SNAP and its effect. Due to the constraints of the data chosen, it would be statistically insignificant to claim that these results can be forecasted due to the nature of this analysis. Still, the Snap program implemented by the US government by proxy of the USDA is in fact a strong supporter to the impoverished and less fortunate households in America. Therefore, this study concludes on a positive outlook on this program and its effect on American food consumption and economic activity.

3.3 Suggested Future Study

While this study investigates the economic effects of SNAP benefits on households, much work is left to be done. One area could be the added nutritional access that SNAP partakers now have available. Weight loss is a strong effect of SNAP benefits however this study did not cover the method in which this was done within this work. An additional path of research could be using a similar analysis pipeline on the effects of SNAP benefits on individuals in the households in this study. Additionally, looking into the long-term findings of the effects of SNAP cited in this study could also prove fruitful. This could be possible by finding more available data and regressing the long-term effects on SNAP using households.

Additionally, there is a large opening for policy related qualitative analysis on how SNAP regulations and current food costs play a role in the wellbeing of American households. Recent studies (Young,2021) show a significant benefit for the taxation of luxury food items to subside and then increase the availability of more nutritional food goods. Likewise, much time and effort could be put into underdoing the barriers of entry and exit within the SNAP program and what mental and physical tole are placed on Americans.

Work Cited

Bitler, M. (2014) The Health and Nutrition Effects of SNAP: Selection Into the Program and a Review of the Literature on Its Effects. University of Kentucky Center for Poverty Research Discussion Paper Series, DP2014- 02. Retrieved from http://www.ukcpr.org/Publications/DP2014-02.pdf.

Carlson, S., & Keith-Jennings, B. (2018, January 17). *SNAP is linked with improved nutritional outcomes and lower health care costs*. Center on Budget and Policy Priorities. Retrieved April 27, 2023, from https://www.cbpp.org/research/food-assistance/snap-is-linked-with-improved-nutritional-outcomes-and-lower-health-care

Coleman-Jensen, Alisha, Matthew P. Rabbitt, Christian A. Gregory, Anita Singh, (2022). *Household Food Security in the United States in* 2021, ERR-309. U.S. Department of Agriculture, Economic Research Service

Everitt, B., Landau, S., Leese, M. (2001) *Cluster Analysis.* Hodder Headline Group.

Farkhad, F. B. (2018). The Impact of Participation In SNAP on Labor Force Decisions. Lehigh University.

Gray, C., Leive, A., Prager, E., Pukelis, K., Zaki, M. (2021). Employed in a SNAP? The Impact of Work Requirements on Program Participation and Labor Supply. *NBER Working Paper Series.* National Bureau of Economic Research

Gregory, C. A., & Deb, P. (2015). Does snap improve your health? *Food Policy*, *50*, 11–19. https://doi.org/10.1016/j.foodpol.2014.09.010.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction*. Springer.

Kentucky Department of Agriculture. (2016). *Kentucky AG News*. Hunger-study-finds-food-insecurity-levels-remain-historically-high. Retrieved from https://www.kyagr.com/Kentucky-AGNEWS/2016/Hunger-study-finds-food-insecurity-levels-remain-historically-high.html. Kentucky Department of Agriculture.

ROSENBAUM, P. R., & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

United States Department of Agriculture. (2016). National Household Food Acquisition and Purchase Survey (FoodAPS): User’s Guide to Survey Design, Data Collection, and Overview of Datasets. United States Department of Agriculture.

Young, Jeffrey S. 2021. “Measuring Palatability as a Linear Combination of Nutrient Levels in Food Items.” *Food Policy*, 104, 102146

Appendix

Dataset SQL ETL Pipeline

```
-- data type evaluation
SELECT
TABLE_CATALOG,
TABLE_SCHEMA,
TABLE_NAME,
COLUMN_NAME,
DATA_TYPE
FROM INFORMATION_SCHEMA.COLUMNS
where TABLE_NAME = 'Household_Data'

-- create new table for demographic data
CREATE TABLE Demographics(
   Household_Number INT,
   Household_Size INT,
   Region INT,
   Rural INT,
   Income FLOAT,
   Target_Group INT,
   SNAP INT,
   Food_Sufficient INT
);

-- fill new table with data from main table only if snap program was used
INSERT INTO Demographics
(Household_Number,Household_Size,Region,Rural,Income,Target_Group,SNAP,Food_Sufficient
)
SELECT household_num, household_size, region, rural, Income_avg_mon, target_group,
snap_ever, food_sufficient_score
FROM Household_Data
WHERE snap_ever = 1

--Run test join inorder to prevent errors
SELECT Demographics.Household_Number, Clusters.Household
FROM  dbo.Demographics
INNER JOIN Clusters ON Demographics.Household_Number = Clusters.Household

-- create new table for detrmined clusters found in R code
CREATE TABLE Demographics2 (
   Household_Number INT,
   Household_Size INT,
   Region INT,
    Rural INT,
    Income FLOAT,
    Target_Group INT,
    SNAP INT,
    Food_Sufficient INT,
    Cluster INT
);

--Actual data merege
INSERT INTO Demographics2
(Household_Number,Household_Size,Region,Rural,Income,Target_Group,SNAP,Food_Sufficient
,Cluster)
SELECT DISTINCT Demographics.Household_Number, Demographics.Household_Size,
Demographics.Region, Demographics.Rural,
Demographics.Income,Demographics.Target_Group, Demographics.SNAP,
Demographics.Food_Sufficient,Clusters.Cluster
FROM  dbo.Demographics
INNER JOIN Clusters ON Demographics.Household_Number = Clusters.Household

-- Insure no duplicates are within the new table
SELECT Count(Household_Number) AS DuplicateRanks
FROM Demographics2
GROUP BY Household_Number
HAVING COUNT(Household_Number)>1;

--Test new table
SELECT *
FROM Demographics2
ORDER BY Household_Number

--copy main foodaps data conditional on snap benifits
SELECT * INTO Household_Data_SO
FROM Household_Data
WHERE snap_ever = 1
ORDER BY household_num

--Test join
ALTER TABLE Household_Data_SO
ADD CLuster INT

--Add clusters to snap table
SELECT * INTO Household_DataSO
FROM Household_Data_SO
LEFT JOIN Clusters2
```

```
ON Household_Data_SO.household_num = Clusters2.labels1

-- Remove dummy table
DROP TABLE Household_Data_SO

-- Remove unwanted column
ALTER TABLE Household_DataSO
DROP COLUMN labels1

-- Test new table
SELECT household_num, Income_avg_mon, household_poverty_guideline
FROM Household_DataSO
WHERE Cluster = 1
```

SQL Food Expense Feature Synthesis

```sql
USE [Thesis Data]

--Join food expences in FAH with Foodaps household databses
SELECT Fah_exp.["hhnum"], sum(Fah_exp.["fah_exp_total"])as 'total_fah_exp',
COUNT(["fah_exp_total"]) as 'num_meals_fah'
FROM Fah_exp
Right JOIN dbo.Household_DataSO ON dbo.Household_DataSO.household_num = Fah_exp.["hhnum"]
WHERE Fah_exp.["fah_exp_total"] > 0
GROUP BY ["hhnum"]
ORDER BY ["hhnum"]

--Join food expences in FAFH with Foodaps household databses
SELECT fafh_exp.["hhnum"], sum(fafh_exp.["fafh_exp_total"]) as 'total_fafh_exp',
COUNT(["fafh_exp_total"]) as 'num_meals_fafh'
FROM fafh_exp
Right JOIN dbo.Household_DataSO ON dbo.Household_DataSO.household_num = Fafh_exp.["hhnum"]
WHERE Fafh_exp.["fafh_exp_total"] > 0
GROUP BY ["hhnum"]
ORDER BY ["hhnum"]

-- Removing missing values from both databases
DELETE FROM FAH_totals WHERE num_meals_fah = 0;
DELETE FROM FAFH_totals WHERE num_meals_fafh = 0;

--FAH data validation
SELECT *
FROM FAH_totals
ORDER BY ["hhnum"]

--FAFH data validation
SELECT *
FROM FAFH_totals
ORDER BY ["hhnum"]

--Sum and joung of FAH and FAFH food expense totals
SELECT DISTINCT FAH_totals.["hhnum"], FAH_totals.total_fah_exp, FAFH_totals.total_fafh_exp,
FAH_totals.num_meals_fah, FAFH_totals.num_meals_fafh
FROM FAH_totals
INNER JOIN dbo.FAFH_totals on dbo.FAFH_totals.["hhnum"] = FAH_totals.["hhnum"]
ORDER BY FAH_totals.["hhnum"]
```

Python SOM Model

```
#Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import susi
from susi.SOMPlots import plot_nbh_dist_weight_matrix, plot_umatrix

#data loading
X = pd.read_excel('C:\\Demographics 1.xlsx')

#parameter selection
rows = 120
columns = 120

#Model Training
som = susi.SOMClustering(n_rows = rows, n_columns=columns)
som.fit(X)
print("SOM Fitted")

#Model deployment
u_matrix = som.get_u_matrix()
plot_umatrix(u_matrix, rows, columns,cmap='summer')
plt.show()

#Ploting commands
plot_nbh_dist_weight_matrix(som)
plt.show()
```

R Hierarchical Clustering Model

```
#Import packages
library(dendextend)
(library(dplyr))
(library(ggplot2))

#Data Extraction
set.seed(569)
df <- as.data.frame(read.csv('C:\\ Demographics 1.csv'))
print(df)
any(is.na(df))
summary(df)
count(df)

#Data Cleaning and Standarizing
labels1 <- df$Household_Number
df$Household_Number <- NULL
df <- as.data.frame(scale(df))
df <- subset(df, select = -c(SNAP))
#df <- df[df$Food_Sufficient==3,]
summary(df)
str(df)
print(df)
count(df)


#First CLustering
dist_matrix <- dist(df, method = 'euclidean')
hclust_1 <- hclust(dist_matrix, method = 'complete')
plot(hclust_1)


#Cluster recolor
plot(hclust_1,labels = labels1)
rect.hclust(hclust_1, k=3, border=2:6)
avg_dend_obj <- as.dendrogram(hclust_1)
avg_col_dend <- color_branches(avg_dend_obj, k =8 )
plot(avg_col_dend)

#Cluster Maping
cut_avg <- cutree(hclust_1, k = 3)
df_cl <- mutate(df, cluster = cut_avg)
count(df_cl,cluster)
```

Python Supervised ML Models

```
#import packages
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import xgboost as xgb
from sklearn.ensemble import RandomForestClassifier
from sklearn import linear_model
from xgboost import plot_importance
import tensorflow as tf
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn import metrics
import matplotlib.pyplot as plt

#data loading, scaling and cleaning
df = pd.read_excel('C:\\ Household_DataSO_withtotalexp_demreduction.xlsx')
scaler = MinMaxScaler()
df=df.dropna()
df['benefits'] = df['benefits'].apply(
   lambda x: 1 if x > 0  else (0 if x == 0 else None))

#target selection
y = df['benefits']
df.drop(columns=['benefits'],inplace=True)
x = scaler.fit_transform(df)

#train test split
X_train, X_test, y_train, y_test = train_test_split(x,y,test_size=0.3,random_state=1234)

#XGBoost model training and testing
xgb = xgb.XGBClassifier()
xgb.fit(X_train,y_train)
y_hat = xgb.predict(X_test)
y_prob = xgb.predict_proba(X_test)

#Logistic Regression model training and testing
logr = linear_model.LogisticRegression(random_state=1234, max_iter=100)
logr.fit(X_train,y_train)
y_pred_logr = logr.predict(X_test)
```

```
#Random Forest Training and Testing
rf = RandomForestClassifier(n_estimators=1000).fit(X_train,y_train)
y_pred_rf = rf.predict(X_test)


#XGBoost Model Evaluation
print('XGBoost')
test2 = (accuracy_score(y_test,y_hat))
auc = metrics.roc_auc_score(y_test,y_hat)
print('The accuracy of Testing is '+str(test2))
print('\n')
print(classification_report(y_test, y_hat))
print('The AUC is '+str(auc))
cm = confusion_matrix(y_test,y_hat)
print(cm)

#Logistic Regression Model Evaluation
print('Logistic Regression')
test2 = (accuracy_score(y_test,y_pred_logr)*100)
y_pred_proba = logr.predict_proba(X_test)[::,1]
auc = metrics.roc_auc_score(y_test, y_pred_proba)
print('The accuracy of Testing is '+str(test2))
print('\n')
print(classification_report(y_test, y_pred_logr))
print('The AUC is '+str(auc))
cm = confusion_matrix(y_test,y_pred_logr)
print(cm)

#Random Forest Model Evaluation
print('Random Forest')
test2 = (accuracy_score(y_test,y_pred_rf)*100)
auc = metrics.roc_auc_score(y_test, y_pred_rf)
print('The accuracy of Testing is '+str(test2))
print('\n')
print(classification_report(y_test, y_pred_rf))
print('The AUC is '+str(auc))
cm = confusion_matrix(y_test,y_pred_rf)
print(cm)

#ROC Curve Plotting
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_logr)
fpr1,tpr1, _ = metrics.roc_curve(y_test, y_hat)
fpr2,tpr2, _ = metrics.roc_curve(y_test,y_pred_rf)
plt.plot(fpr,tpr)
plt.plot(fpr1,tpr1)
plt.plot(fpr2,tpr2)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.show()
```

Python Propensity Score Matching

```
#import packages
import pandas as pd
from psmpy import PsmPy
from psmpy.functions import cohenD
from psmpy.plotting import *

#cluster selection
cluster = 1

#data loading and cleaning
df = pd.read_excel('C:\\propensity score vars.xlsx')
df.loc[df['CLuster']==cluster]
df=df.dropna()
df['benefits'] = df['benefits'].apply(
   lambda x: 1 if x > 0  else (0 if x == 0 else None))

#mathcing
psm = PsmPy(df, treatment='benefits', indx='household_num', exclude = [])
psm.logistic_ps(balance=True)
print(psm.predicted)
psm.knn_matched(matcher='propensity_score', replacement=False, caliper=None,
drop_unmatched=True)

#results plot
psm.effect_size_plot(save=False)
```