Benjamin Harwood, Kevin Vogel, Kyle Wass, and Nina Williams

The United way of choosing… united way locations

# **Introduction:**

The United Way is a community chest organization that serves to improve the lives of the people who live in the communities in which they serve. The United Way was started over 125 years ago and has evolved over time to positively affect the lives of over 61 million people worldwide every single year. Their mission is simple and effective, “United Way improves lives by mobilizing the caring power of our community to advance the common good.” To that end they focus on three key areas - education, income, and health.

United Way operates over 1,200 branches in the United States alone; 57 of these are in the states of Kentucky and Tennessee, 23 and 34 respectively. In TN alone there are 95 counties; with only approximately one third of all counties having a United Way, difficult choices must be made. Who should they serve? Who shouldn’t they? The unfortunate downside of having such a vast country is that some of the neediest folks live in rural areas that must share resources amongst several counties. This creates an absolute disparity in utility amongst branches.

Each United Way branch is both a revenue center and a cost center; that is, they collect donations, pay for overhead, and allocate the remainder of the funds to the mission – furthering educational, income, and health initiatives. Each year, every United Way branch around the country campaigns to build funds. Most campaign funds are derived from payroll deductions from mid-size to large businesses. Staff size varies by location; there is a correlation between total dollars received and percent of funds that are distributed to forward the mission. In fact, some branches work at a deficit.

So, how does the United Way choose where to place their branches? There are many potential factors: Is it based on absolute utility – the greatest good for the greatest number of people? Are there counties that are underserved? Can we predict where the United Way thinks that they are best able to make and affect positive change in the world? If they opened a new branch somewhere, where would they open it and what would the revenue be?

Tennessee, Kentucky, and Ohio approximately represent the average state in terms of population, demographics of inhabitants, income and income disparity, and the divide between rural and city life. If we can predict where they place branches in these states, we should be able to predict where they put branches anywhere in the United States and can suggest where they may provide more utility than they currently provide.

# **Section 1: Analysis and Models**

This section will examine the data used for the study. A bulk of it comes from the US Census Bureau, though data regarding the United Way had to manually acquired.

# **Section 1.1: About the Data**

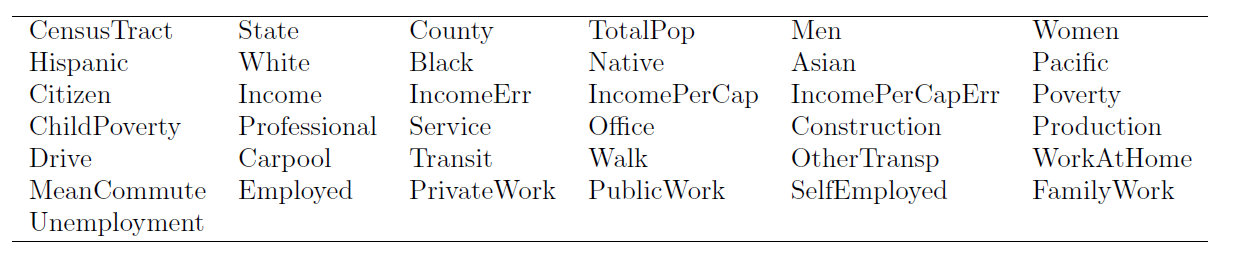
The bulk of the data comes from the previous census, taken in 2010. This data was subsequently cleaned and completed by the census bureau, with the ultimate final version coming in 2015. The country is broken into 74,001 different tracts, small subdivisions of each state and county comprising approximately 3,500 people. The data consists of 37 variables include information regarding location, race, income, population, gender, poverty levels, citizenship, commuting, employment, and job types as follows:

Table 1: Census data variable names

After screening for missing values, the two error variables were removed as they were determined to have no impact on United Way location placement.

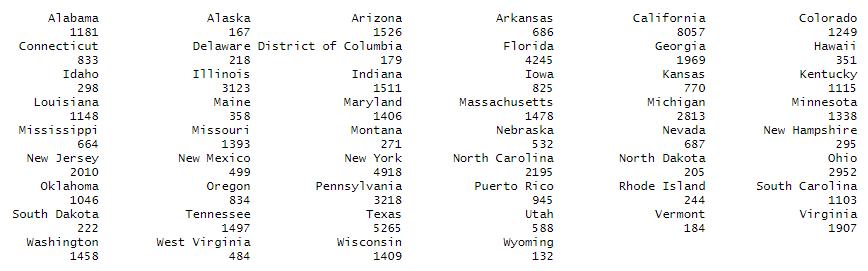
To get an idea of the real scope of the census data, he following table shows how many census track ID’s or lines of data there are for each State. Notice the range of values, with California having the largest number of census tracks of 8,057, while Wyoming has the least at 132. 

Table 1: State census tract counts

In order to reduce the scope of the task, the data was focused to two states, previously alluded to in the introduction, Kentucky and Tennessee[[1]](#footnote-2). In order to be able to better work with the values for each state, percentages were converted to the number of people represented. Below are summary statistics for each of the two states:

|  |  |
| --- | --- |
| Table 2: Summary statistics for Kentucky | Table 3: Summary statistics for Tennessee |

Next, United Way data was collected for each of the two states. Eight variables were included, but only the state, county, and annual gross was needed. Because the United Way data was broken by county, the state data sets were summarized by county as well. The United Way data was then merged with each of the two states, creating a new binary variable indicating whether the county had a location or not.

# **Section 1.2: Models**

This and the coming subsections will examine each of the different models used to determine criteria to use in predicting United Way locations.

# **Section 1.2.1: Association Rules**

In order to use association rule mining, additional prep had to be performed on the data. Specifically, it had to be converted into market basket data. To do this, the state, county, and gross variables were removed. Then, each remaining variable (except the binary location variable) was discretized into its four quartiles. As an example, a county whose population was in the second quartile was recoded as: Q2(TotalPop). Additionally, the value range for each quartile was attached, with upper bound inclusion, allowing for quick identification of which variable was in question but also giving an idea for the values. Continuing the example, the ultimate identifier would have been: Q2(TotalPop) = (13591,24065]. The Pacific variable had to be re-coded differently, as 3 of the 4 quartiles were 0, so it was recoded as “No” if there were no individuals identifying as Pacific origin in that county, or “Yes” if there was more than 0.

Next, in order to tune the association rules for ideal support and confidence thresholds, rule counts were generated using all combinations of support and confidence in the range [0.1,0.9] for rules indicating there was a location and rules indicating there was not a location. Here are tables showing the results of these:

|  |  |
| --- | --- |
| Table 4: "Yes" rule counts by confidence and support | Table 5: "No" rule counts by confidence and support |

In the case of the “Yes rules” on the left, a minimum support above 0.1 yielded no rules. Here are the 20 most frequently occurring rules:

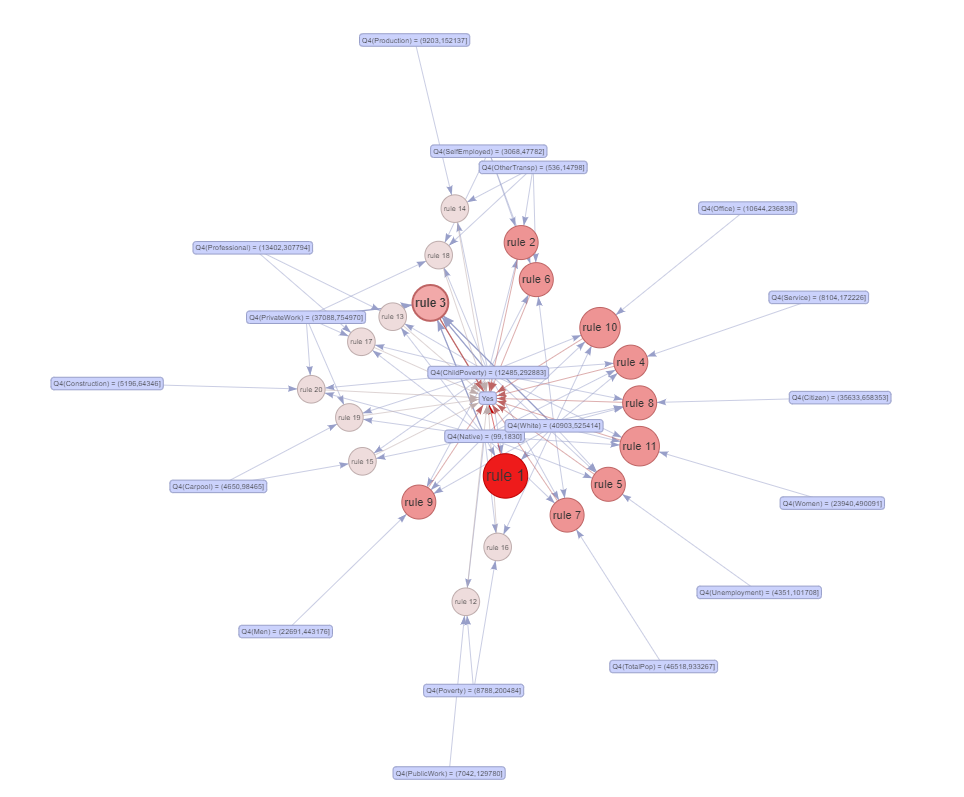


Figure 1: Yes rule display

Here, the intensity of the red color is indicative of the frequency in which that rule occurred.

There were 14 rules that appeared a total of 24 times, each with identical support (0.111), confidence (0.96) and lift (3.655). Surprisingly, in 12 of these 14 rules, the following set was present: {Q4(ChildPoverty), Q4(IncomePerCap), Q4(Native)}, so that one set occurred in at least 168 rules. Q4(ChildPoverty) appeared in other rules in the “top 20” as well, and while admittedly difficult to see in the previous graphic, this is indicated by the number of blue arrows stemming from that item (it’s toward the center).

Here are the “No” rules, using support of 0.2 and confidence of 0.9:

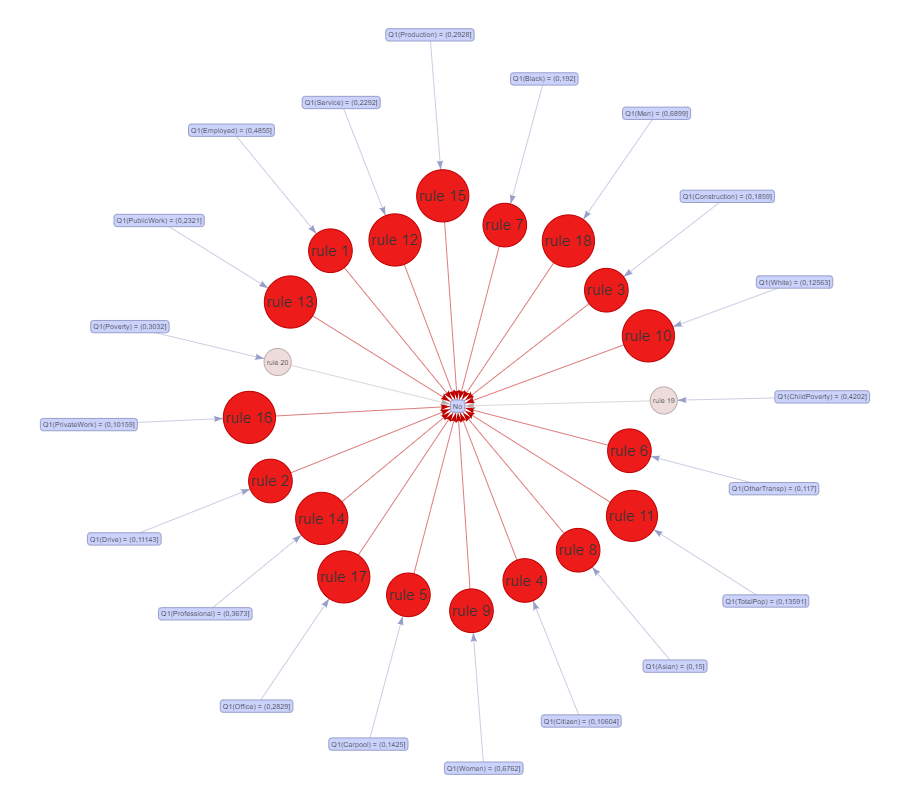


Figure 2: "No" rules display

In this case, the most frequent rules involve ONLY the bottom quartile ranges for each variable. Compare this to the “Yes” rules that involved all the fourth quartile ranges.

# **Section 1.2.2: Clustering**

There were minimal additional steps that had to be taken to ensure that data was cleansed and usable for the sake of clustering. One step that must be taken whenever clustering is removing the data label before the cluster was created, so identifying information was removed from the dataset before clustering to include revenue and whether there was a United Way there or not.

Before clustering methods such as k-means and hierarchical clustering, the optimal number of clusters was determined. There are several different methods by which to determine the optimal number of clusters; the results of three common methods (gap statistic, silhouette, and elbow) follow. Gap statistic and the silhouette method both agreed that two was the optimal number of clusters. A case could be made via the elbow method for a number higher than 2. Ultimately, 4 clusters were chosen because this allowed for a clear distinction between several of the clusters.

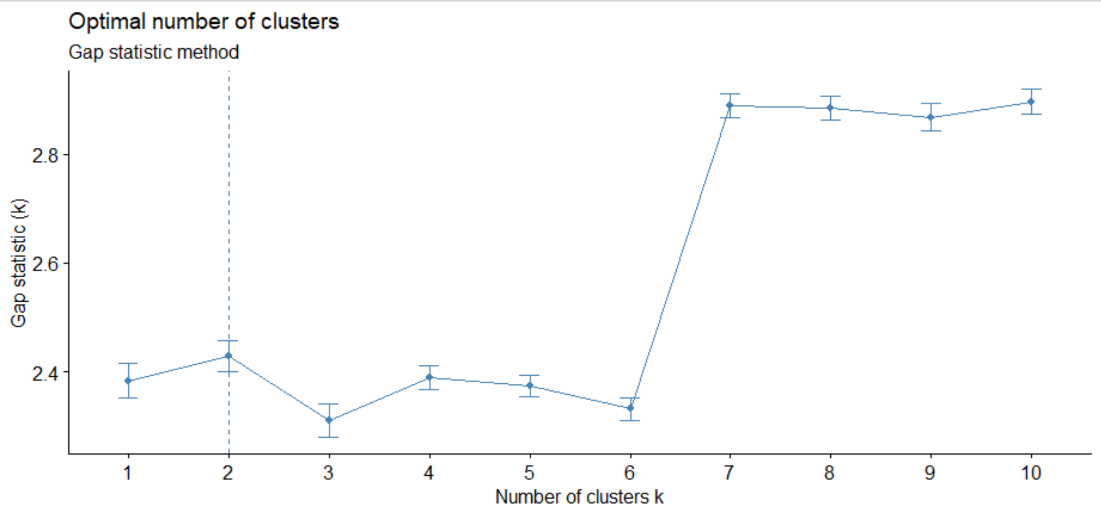


Figure 3: GAP statistic for clusters

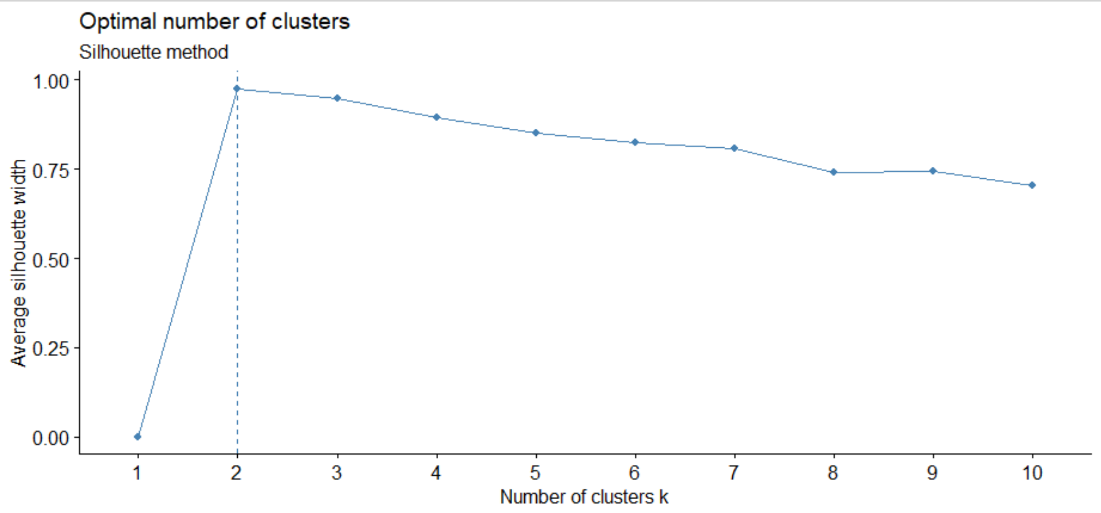


Figure 4: Silhouette cluster method

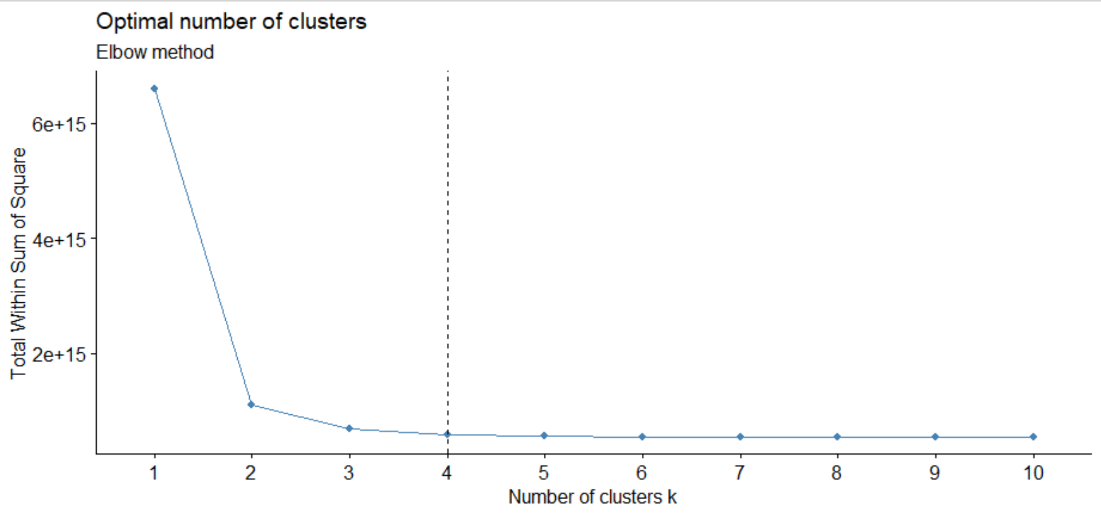


Figure 5: Elbow method for cluster

There was not a lot to glean from the hierarchical clustering. Different methods were used to determine distance, namely Euclidian distance and cosine similarity. Ward.D and complete methods were used for hierarchical clustering. In order to get more appropriately sized groups, k-means clustering was chosen to proceed with. The following dendrogram reflects cosine and complete.

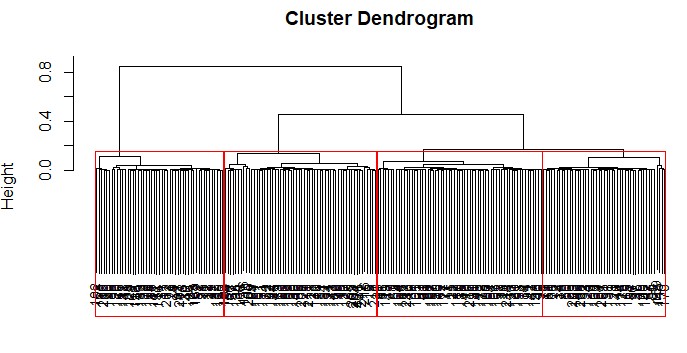


Figure 6: 4 cluster dendrogram

As you can see, four clusters are evident. With that in mind, *k*-means clustering was performed searching for four clusters. The resulting clustering is show here.

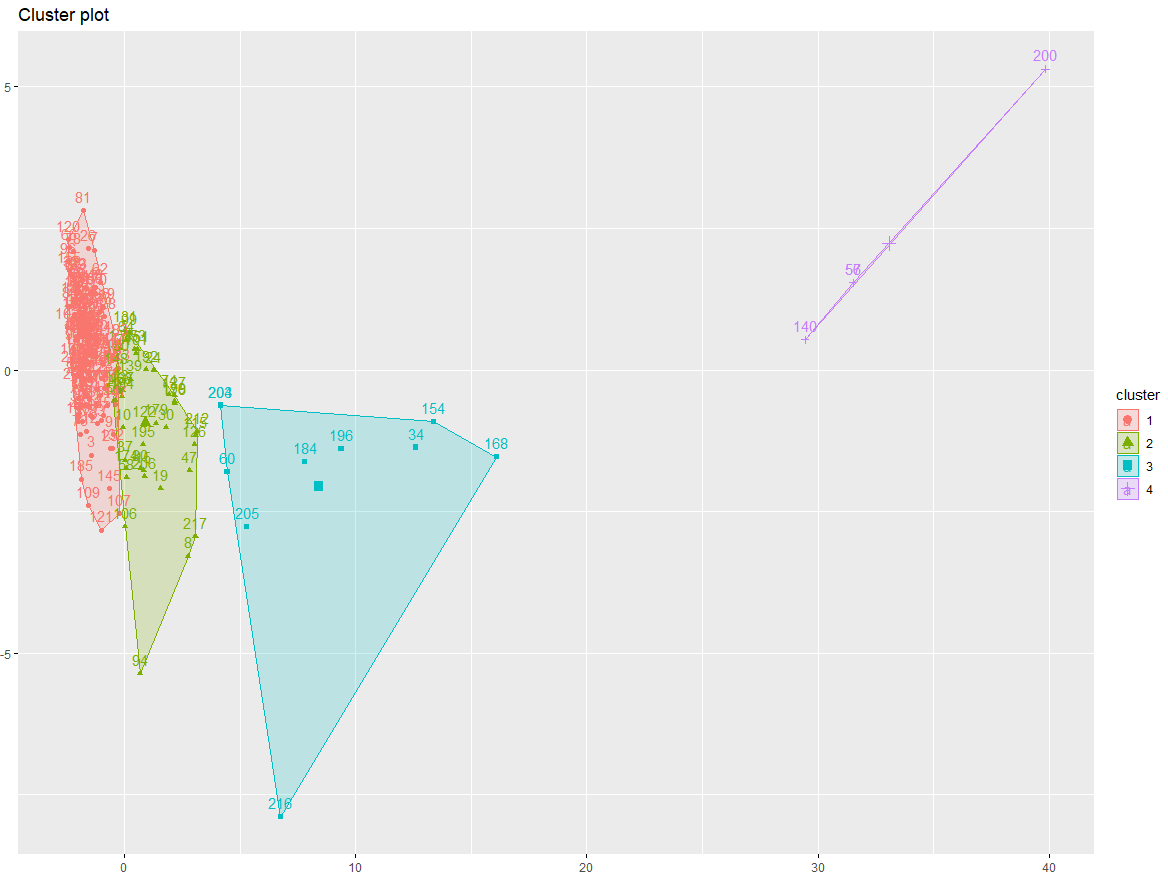


Figure 7: k-means clusters

Notice that the cluster with most data points, cluster 1 (the red one), seemed to have the most data points. This cluster housed the vast majority of “no” and a small minority of “yes.” This table shows the distribution of locations between the different clusters.

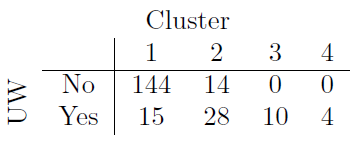


Table 6: k-means cluster table

What is curious here is that clusters 3 and 4 contained only counties that had a United Way location. Here are combined summary statistics for clusters 3 and 4.

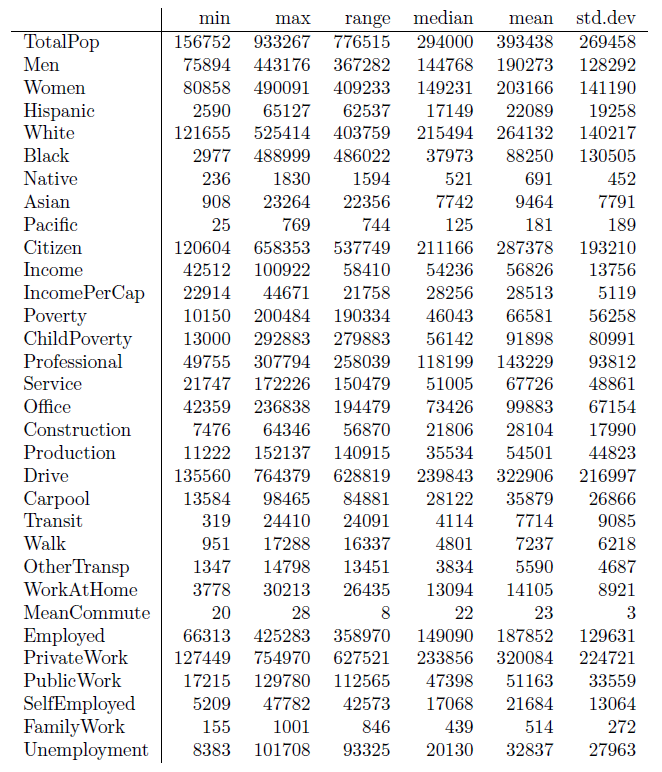


Table 7: Summary statistics for clusters 3 and 4

Viewing the boxplots[[2]](#footnote-3) below, it’s evident that population is a differentiator between the different clusters. Population is a defining characteristic of cluster 4.

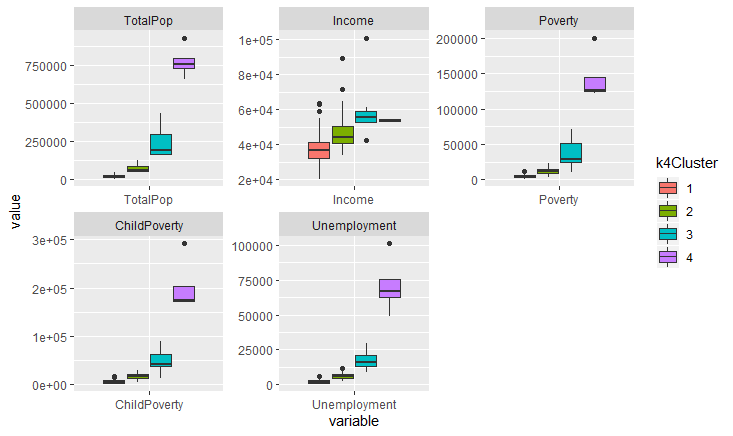


Figure 8: Box plots by cluster

**Section 1.2.3: Naïve Bayes**

Each of the remaining models require normalized data, so before continuing the original data (translated: prior to discretization for association rule mining) was normalized. Also, the remaining models all require training with a subset of the data, and the resulting model is then applied to the remainder of the data, called the testing set. For the purposes of this study, the training set in each case was 80% of the population, but randomly sampled every time. For consistency, the same training set was used for each model.

Although the Naïve Bayes classifier assumes independence between the predictor variables, it is very unlikely that this is the case with the census data. It is known that there is a relationship between poverty levels and income, and demographics and income, among other factors. Despite this, the Naïve Bayes model showed high accuracy in its predictions on the Tennessee and Kentucky test data. What follows are the prediction tables and receiver operating characteristic (ROC) curve[[3]](#footnote-4) for each of the three models. Boldface entries in the tables indicated accurate predictions. Performance metrics will be reported at the end of this section for all techniques.

|  |  |
| --- | --- |
| Model 1    Accuracy: 86% |  |
| Model 2    Accuracy: 81.4% |  |
| Model 3    Accuracy: 88.3% |  |

# **Section 1.2.4: Decision Trees**

A Decision Tree is essentially a flowchart structure that contains internal nodes representing a feature or a test. Each of the nodes represent a label which is the decision resulting of the feature or test. The branches of the trees represent conjunctions of the features that ultimately lead to the label. Overall, the Decision Tree is an algorithm identifying ways to split the data based on specific conditions.

Among the three Decision Tree models there is a commonality in that poverty is present in all three at least once. Communities with higher poverty rates have a higher likelihood of having a United Way location. Below are the decision trees generated from each of the three testing sets, along with their prediction tables and accuracies.

|  |  |
| --- | --- |
| Model 1    Accuracy: 81.4$ |  |
| Model 2    Accuracy: 76.7% |  |
| Model 3    Accuracy: 88.4% |  |

# **Section 1.2.5: Support Vector Machines**

As with the decision tree and Naïve Bayes classifier, multiple support vector machines were developed for the three training sets. Each was tuned to ensure the proper gamma, base coefficient, and cost were utilized. Each model used a polynomial kernel function. These svm’s were then applied to the test set to predict United Way location. For brevity, here is the prediction for the most accurate of the three models (which was, incidentally, model 1, with 88.3% accuracy):

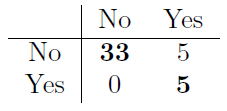


Table 8: Support vector machine predictions

Notice that the support vector machine was more accurate than the Naïve Bayes model on the same training set. Section 1.3 will address the performance measures for each model and method.

**Section 1.2.6: *k*-nearest neighbors**

The idea of the *k*-nearest neighbor algorithm is, in a sense, similar to *k*-means clustering in that it is assumed that the data points are grouped together with those most similar. Think of the saying “birds of a feather flock together”. The algorithm is a supervised learning algorithm[[4]](#footnote-5) that uses a distance metric to find the*k* closest neighbors to each data point. Basically, it seeks to capture the “similarity” between various points, so that when new data is introduced it can see which group of points the new data is most like and assign it a label. Where this is different from *k*-means clustering is that with *k*-means one must specific the number of clusters, while *k*nn requires the number of neighbors desired. With this in mind determining the correct number for *k* is crucial before training.

Continuing with training set 1 (though *k*nn was more accurate on training set 3, but by less than 2%), here is a graph showing how the accuracy varies based on the value of *k*. Notice that the accuracy is maximized at *k*= 11 and *k*=13. Because of this, *k*=11 was used, which yielded an accuracy of 88.4%.

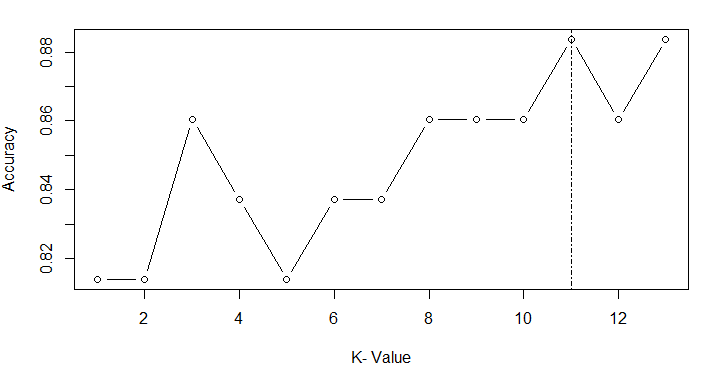


Figure 9: k-nearest neighbor optimization

**Section 1.2.7: Random Forest**

Section 1.2.3 discussed the results of different decision tree models for the different training sets. Taking this to the extreme with a random forest can yield more interesting results. The general idea is to generate a few thousand decision trees and have them vote on which one is the “best” so to speak.

Focusing on training set 1 again, the random forest model was only 79.1% accurate (the random forest was most accurate for training set 3, as with *k*nn, at 88.4%). The plot below grades the variables in terms of their performance. Notice that, once again, ChildPoverty is at the top.

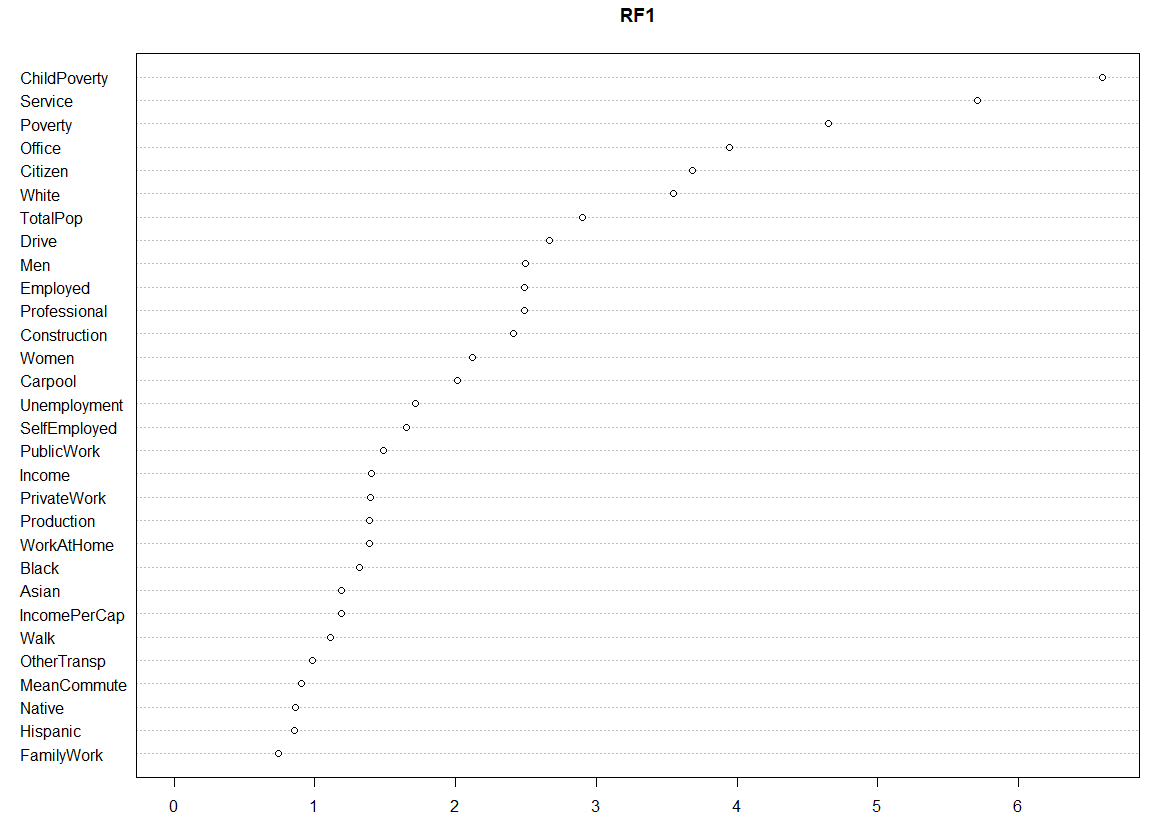


Figure 10: Random forest variable importance plot

**Section 1.3: Performance**

As mentioned through Section 1.3, each model had varying levels of accuracy or “success” in predicting United Way Locations. This section is just a table showing how each model performed across the various training sets, for reference if nothing else.

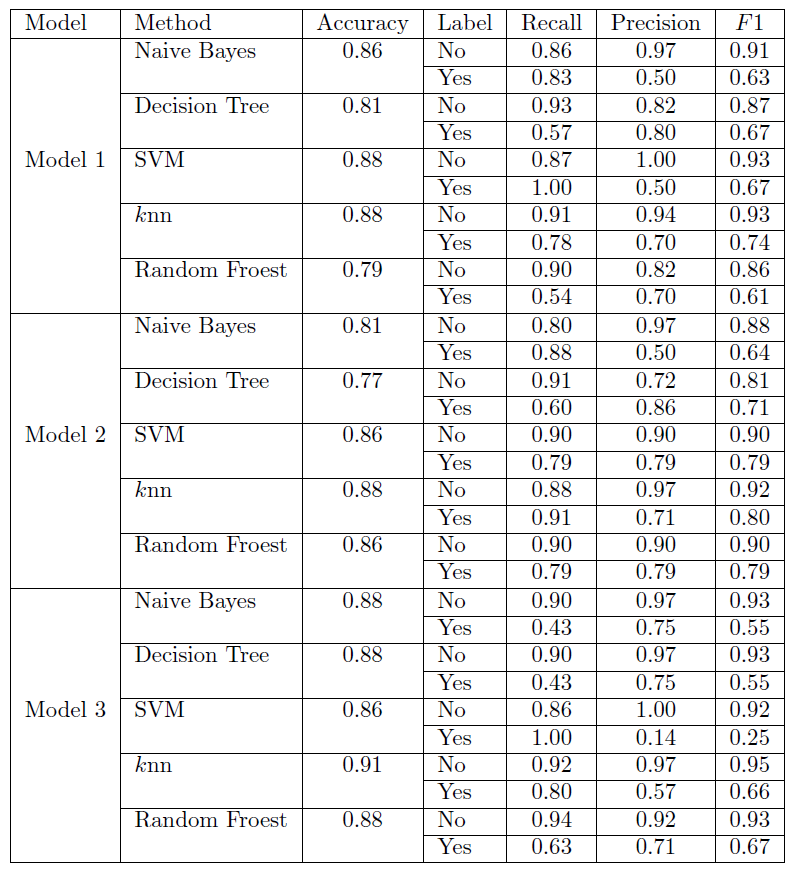


Table 9: Model performance by technique

Notice that Model 3 was the most accurate model for four of the five tests. This is really due to the fact that training set 3 had a much larger percentage of the “Yes” counties than it did the “No” counties.

# **Section 2: Results**

Having considered how well different techniques could predict United Way locations, the models deemed most accurate for each testing method above were applied to data for Ohio, that was prepared in the same way as the Kentucky and Tennessee data. Here are the results:

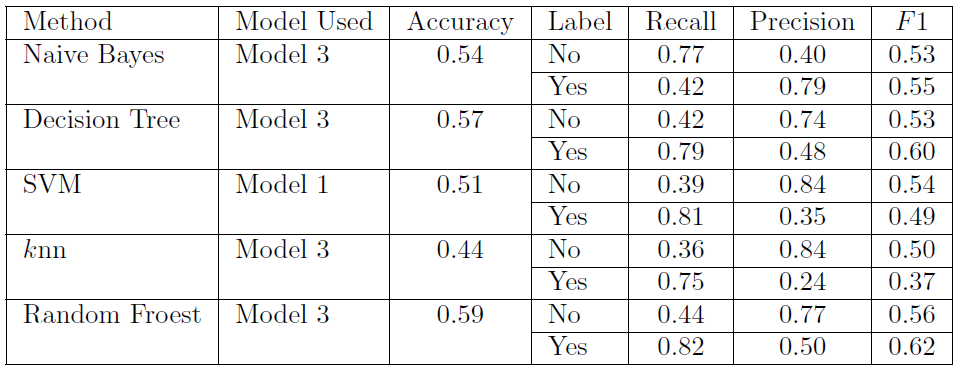


Table 10: Model performance applied to Ohio

Although models were generated with high accuracy on predicting Tennessee and Kentucky data, these same models were noticeably less accurate when used on Ohio test data. Indeed, even when using the entire data set for KY/TN dataset and not just the training set with the highest accuracy, only the support vector machine model was over 55% (specifically 87.1%). This suggests that there may be other factors behind where United Way chooses its locations other than demographics. For example, they may choose locations in areas that are convenient to reach for a larger number of people, or that are along distribution routes for shipping.

# **Conclusion:**

Initially, it was believed that census data alone would be able to predict where United Ways would place their branches. While it is possible to predict within similar regions of the country at a rate of 87.1 %, more data is necessary to predict at a higher rate than ~87.1% for any given state. The data that would be most useful is lat/long data and the number of people that are served by each of the branches.

Being that each United Way branch is both a revenue and a cost center, the United Way must be incredibly picky with where they choose to place branches. Donations are much easier to come by in populated areas. Less densely populated areas need more assistance per capita but are less able to donate. Also, less densely populated areas share United Ways; this increases the distance that someone who needs help must travel to get help, or someone who wishes to donate has to travel to donate.

There is a nonlinear relationship between population and the number of branches that exist in a given area; that is more populated areas (states, counties, etc.) get more United Ways per area whereas less populated areas get less United Ways or even no United Way. United Ways that exist in lesser populated counties are essentially the geographical center of need for several counties that are clustered together. A cursory Google search shows that United Way locations are typically at the center of each county; for United Ways that serve several contiguous counties, folks that live near the edge of adjacent counties are affected the most by distance.

The goal of the United Way is to positively impact the lives of the greatest number of people in need. As such, the United Way Worldwide should take a small cut of each donation and pool those funds in a corporate coffer. These funds should be disbursed and fund branches that otherwise would operate at a loss. Put simply, there are branches that should exist that don’t[[5]](#footnote-6). This will minimize the number of miles that those in the neediest of areas must travel to get help, thereby extending the reach of this organization.

1. These states were chosen for various demographic reasons, [↑](#footnote-ref-2)
2. For a small collection of seemingly important variables [↑](#footnote-ref-3)
3. Remember that a good model will have the bend in the ROC curve as far away from the diagonal as possible [↑](#footnote-ref-4)
4. Supervised learning algorithms rely on labeled input data with which to learn a function that produces appropriate outcomes when provided new data that is unlabeled. Compare this to unsupervised learning algorithms that learn the structure of un-labeled data. [↑](#footnote-ref-5)
5. Note: this was ultimately not part of the study, but seems obvious given the results [↑](#footnote-ref-6)