EMBRY-RIDDLE
Aeronautical University
SCHOLARLY COMMONS

Beyond: Undergraduate Research Journal

Volume 3 Article 3

April 2019

# Data Mining and Machine Learning to Improve Northern Florida’s Foster Care System

Daniel Oldham
*Embry-Riddle Aeronautical University, Daytona Beach*, oldhamd1@my.erau.edu

Nathan Foster
*Embry-Riddle Aeronautical University, Daytona Beach*, fostern2@my.erau.edu

Mihhail Berezovski
*Embry-Riddle Aeronautical University*, berezovm@erau.edu

UNDERGRADUATE RESEARCH JOURNAL
Beyond

# Data Mining and Machine Learning to Improve Northern Florida's Foster Care System

Daniel Oldham, Nathan Foster, & Mihhail Berezovski

## Abstract

The purpose of this research project is to use statistical analysis, data mining, and machine learning techniques to determine identifiable factors in child welfare service records that could lead to a child entering the foster care system multiple times. This would allow us the capability of accurately predicting a case's outcome based on these factors. We were provided with eight years of data in the form of multiple spreadsheets from Partnership for Strong Families (PSF), a child welfare services organization based in Gainesville, Florida, who is contracted by the Florida Department for Children and Families (DCF). This data contained a number of different aspects of the clients ("participants") who were entered into the system as part of PSF's record keeping. These aspects included dates, ages, removal types, disabilities, demographics, case details, and more of the parents, children, relatives, and caregivers involved. We analyzed and mined through this data using statistical analysis software (mostly R Studio), searching for correlations that could help us predict if a child is to be removed from their home and enter back into the foster care system. This research was overall a success, and we found significant insights into the cases that allowed us to predict their success or failure; we also built multiple machine learning models and prediction schemes that facilitated further understanding of statistically significant insights about the cases.

## Introduction

Partnership for Strong Families is a child welfare services organization based in Gainesville, Florida, which provides services for children and families in the Northern areas of the state. Partnership for Strong Families (PSF) provides an important role of connecting children to proper homes and families. PSF collects data regarding all of their clients who are involved in each "case", including, but not limited to, the child, their parents, relatives in the household, other siblings, and caregivers. This data also includes demographics, dates, disabilities, ethnicities, past placements (foster care, institutions, pre-adoption homes, etc.), and more. From the years 2010 to 2017, PSF recorded approximately 170,000 participants' data who were either newly entered into the system, or who had additional information updated to the associated child's case with a new record of involvement in the foster care system. All of these records in the datasets could prove to be extremely useful in benefitting PSF's operations and their ability to help these children by optimizing their efforts. However, the scale of the data itself, reaching approximately 300,000 rows of records over these years, makes it infeasible to analyze without proper statistical analysis and data mining software.

Embry-Riddle's involvement with Partnership for Strong Families is directly tied to this data; the researchers have been tasked to analyze and "mine" these data sets to find statistically significant insights about the participants and their associated case(s). Specifically, the research team is looking for any characteristics or identifiable factors that would lead to a child being removed from his or her family multiple times, and/or be re-entered into the system after being placed out of it. The majority of the children in the given data have multiple records, different placement types, and multiple caregivers, making this initiative a multi-faceted and complex one.

Overall, the research question is: Can the research team employ statistical analysis techniques and machine learning methods to pinpoint specific traits of a child and/or their case(s) that could be used to accurately predict if the child leaves the foster care system or is re-entered again at a later date?

## Literature Review

### Pertinent Research in this Field

Predictive analytics have been employed before for analysis on the child welfare services companies, and overall the use of new technology on older records is becoming widespread across academia, non-profits, and child welfare organizations. As authors like Scharenbroch & Park (2017) and Teixeira & Boyas (2017) show, using data science methods on records to find insight into optimizing child welfare operations can be in-depth, yet surprisingly simple and accessible for a wide range of providers. There are not only efforts being done to make this sort of analysis possible, but also smooth out the challenges that are present.

### Types of Machine Learning Models

For the purposes of this research, machine learning models were used. "Machine learning" itself is the process of training a computer to recognize some sort of correlation in data, and then predict further characteristics of the data set based on the correlations that it finds. The models, which were a neural network, decision tree, and random forest model, all attempted to "learn" how to predict a case's success based on the details and characteristics of a child's case. Once this was done, the prediction model could be based off of how the machine weighted each characteristic while it was trying to learn what sort of correlation each detail had with the success of the case.

## Methods

### Broad Statistical Overview

The first eight spreadsheets that were provided from Partnership for Strong Families were from each year of data collection: 2010 to 2017. They included "participants" data, mostly basic information like gender, ethnicity, birth dates, and "roles" (child, parent, etc.), however, they also provided both a *CaseID* and *IdentificationID*. Each case could have multiple participants, as a case normally involved a significant portion of a family, or household. The unique IDs of the actual people per case were given by *IdentificationID*, and most importantly, they could be cross-referenced through any of the spreadsheets. Along with these aspects of the data, a significant number of "flags" were also present: columns that had either a "Y" or "N", and "flagged" some sort of condition about the participant. Many of these were medical conditions, for example, Autism, but some weren't, for example, the Adoption flag indicated if the child had been adopted or not. Unfortunately, after looking closer into these "flags", the researchers realized that very few records did not have an "N", so these columns were mostly ignored for the analysis due to lack of a good sized sample of "Y" records for any of these columns. The PSF contact also made it clear that these flags in the data were mostly for statistical testing and not real insight. The research team briefly viewed them and then moved on to deeper analysis.

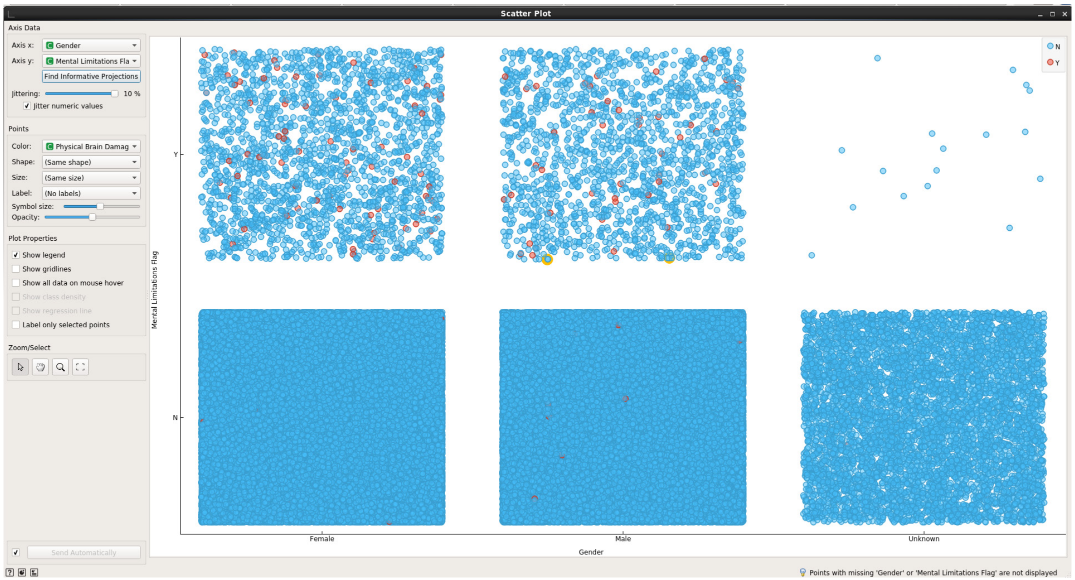
Although the data was anonymized for obvious reasons, researchers had a solid foundation to analyze a single case or specific people based on the IDs. The remaining few spreadsheets included more data regarding removals, cases, and placements. Included in these were zip codes, placement settings, types of services, end reasons, case begin and end dates, and more. In terms of size the spreadsheets for participants from 2010 to 2017, contained a total of nearly 230,000 records. The other three spreadsheets had about 70,000 total more rows of data entries. The participants' data was given an extra column ("Year") using dplyr's mutate() in R Studio, denoting which year the records came from. Then, all eight years' worth of records were combined to form one spreadsheet, which was imported into R Studio as a dataframe. A brief overview of the data is provided below (using approximate values):

- 230,000 total rows of participant data from 8 years of collection
- 180,000 of these records have ZIP codes
- 41,000 unique cases
- 170,000 different participants; ~60,000 of which are children
- 915 unique geographic areas (ZIP codes) ~7% of which have more than 200 cases
- 7 different Placement Settings
- 250 unique types of Services

### Attempt at Automated Machine Learning using ORANGE

One of the first aspects of this data that was noticeable to the researcher team was the absence of numeric values. The data itself is filled with "factors" - columns that have a certain set number of different entries, for example, ethnicity, gender, service role, and more. Most of the data was not numeric, so this posed an initial problem about basic data analysis: where to start if the research team can't even run a regression or try to plot a correlation? To start, the data was imported into a data mining software called "Orange", which had an automatic machine learning function that found correlations in the data, even in the presence of factors.

Visualizing the data was a great way to learn about it and explore the different features it had, however, the automatic data mining was largely unsuccessful (see Figure 1), due to the absence of "Y" records in the flags (in figure - "Y" entries are red dots, in a sea of blue). In addition to this, there was no clear way to tell what who had successful cases in terms of not being re-entered or pulled out of their families multiple



Axis x: Gender
Axis y: Mental Limitations Fla
Find Informative Projections
Jittering: 10 %
☑ Jitter numeric values
Color: Physical Brain Damag
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Symbol size:
Opacity:
☑ Show legend
☐ Show gridlines
☐ Show all data on mouse hover
☐ Show class density
☐ Show regression line
☐ Label only selected points
☑ Send Automatically

Points with missing 'Gender' or 'Mental Limitations Flag' are not displayed

***Figure 1***: *Plot of Gender, Mental Limitations Flag, and Physical Brain Damage Flag using Orange*

times, and who the system had failed. This plot below has genders on the X axis: Female, Male, Unknown; the "Mental Limitations" Flag on the Y axis (N on bottom and Y on top); and finally, the "Physical Brain Damage" flag is represented as a "Y" when the point is colored red. It would make sense that the children who are marked "Y" for Mental Limitations could possibly have had physical brain damage in their lives, in which case, they would be marked with a red dot.

## Basic Sub-setting Analysis using R Studio

Another exploratory method used to attempt to find insights out of the data was simply by creating subsets of the data based on characteristics that the child or case had which would place them in the group that was considered unsuccessful cases. Considering the researchers were looking for multiple entries into the system in the child's case span, as well as being removed from their home, one of the subsets created to compare with the rest of the data was children with two or more cases, who also had been reunited with their parents during these. This subset was done via dplyr in R Studio and had less than 1000 children, with the most prominent ethnicity being African American. It is worth noting, however, that over 700 of the children were missing data for their ethnicity. A number of children in the 0-5 year old age range had foster home cases, but less after that, in the 6-12 category. This 6-12 category was also dominated mostly by males: female children in this age range did not have many foster care. Interestingly, however, there was not a significant difference between the number of males (490) and females (473) in this subset. As a final note, almost all of these cases in the subset were *not* voluntarily removed from their homes; most had "Court Ordered" as their "Removal Manner".

All of these aspects in this subset were intriguing to consider, but still did not provide any hard correlation or defining factor about these cases or children. The insights were mostly scattered. Exploratory analysis was continued with the zip codes, before returning to the participants' data.

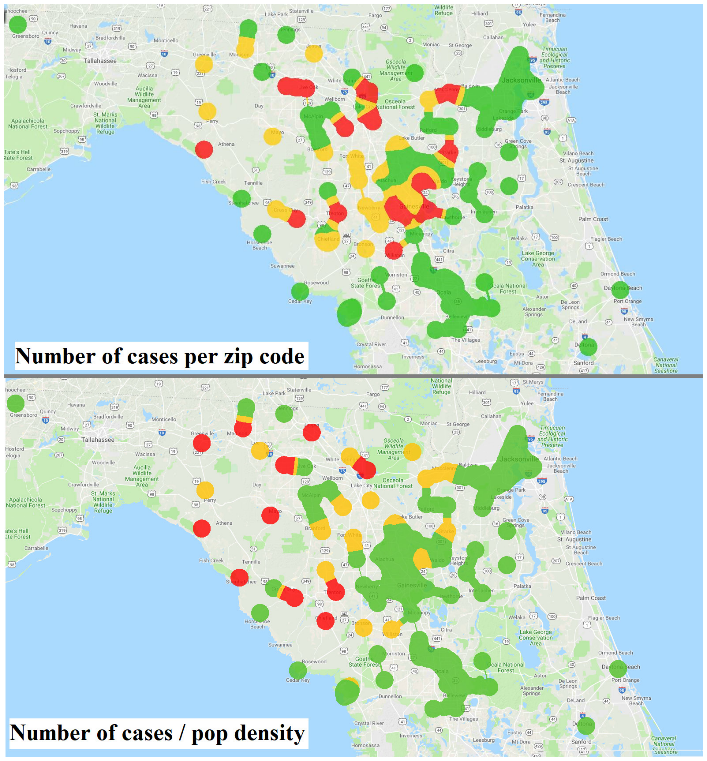


***Figure 2:*** *Heat mapping using zip codes in data.*

### Geographic Heat Mapping of Case Number and Density

The zip codes in the data were of significant interest from the start simply due to the capability of easily visualizing the massive amount of them that the data provided. Using eSpatial, a geographic heat mapping program, two maps were built (see Figure 2). The first was simply done based on the number of cases in a zip code. However, this was not very useful because it logically makes sense that the cities will have more cases due to population density. So, a second map was built using the same data, but this time, comparing the number of cases to the population density of the area (population / area). This allowed us to see "problematic" areas: zip codes where the number of cases per capita was significantly higher than the other areas. Most of these areas were out towards the west, away from the cities and more in rural areas. Areas with the highest number of cases per population density included Greenville, Steinhatchee, Live Oak, Old Town, Chiefland, Perry, and Lake City.

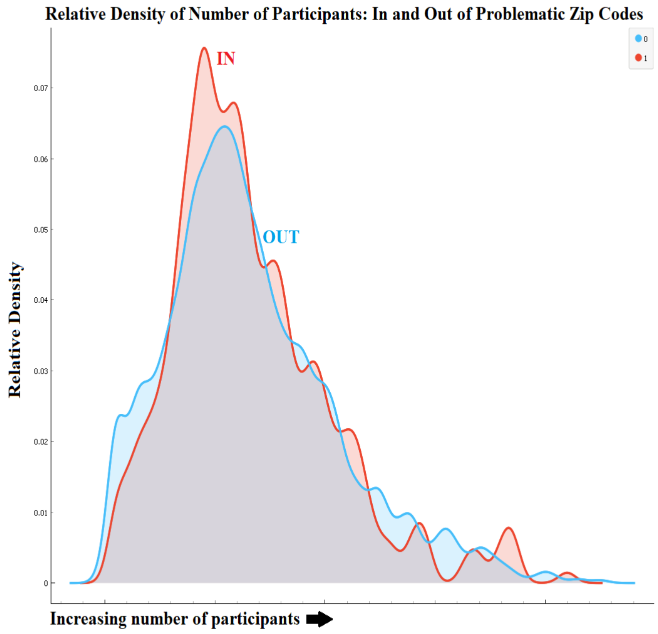


***Figure 3:*** *Plot of relative density of cases in and out of "problematic" zip codes.*

After sub-setting the data and analyzing the cases from just these locations, no significant difference could be seen with the data. This data was loaded into Orange and compared side-by-side with the data that was not from these areas. No correlations or differences in distribution could be deciphered, except a very small difference in the number of participants per each case. In these zip codes, the cases had slightly higher zip codes (see figure below), however, the overall distribution was surprisingly unchanged, despite this higher "per capita" rate of cases (See Figure 3). It became obvious then that the researchers needed to mine the data even deeper for correlations behind the surface of the data.

### Tracking a Child's Placements through Multiple Placements

The simplest way to see how well or badly children are doing in the foster care system is to look at their cases individually. When the data was subset for a single child, it becomes clear that some children jump back and forth from different placement settings like foster care, institutions, or group homes, for many years in some cases. It seems clear that the system is failing them: they are not being placed where they need to be to leave the system through adoption or exiting somehow else.

One of the objectives of this research was to track all of the children individually as they moved around different placements, and then analyze what sorts of "paths" they take to determine who is being provided for properly. This could lead the researchers to a different paths that are always more successful (move to pre-adoption or leave the system), or always less successful (further cases or looping back around to the same placement setting), thus giving the team at the very least subsets of children to analyze further and see what characteristics they had, including their "path" through the system.

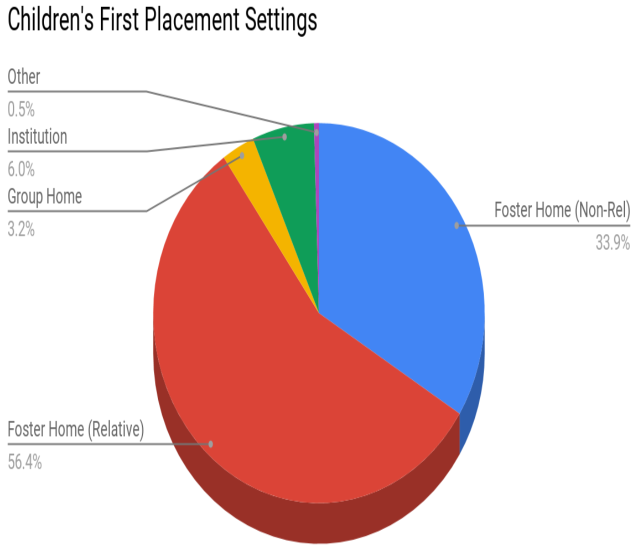


***Figure 4:*** *Pie chart of placement settings.*

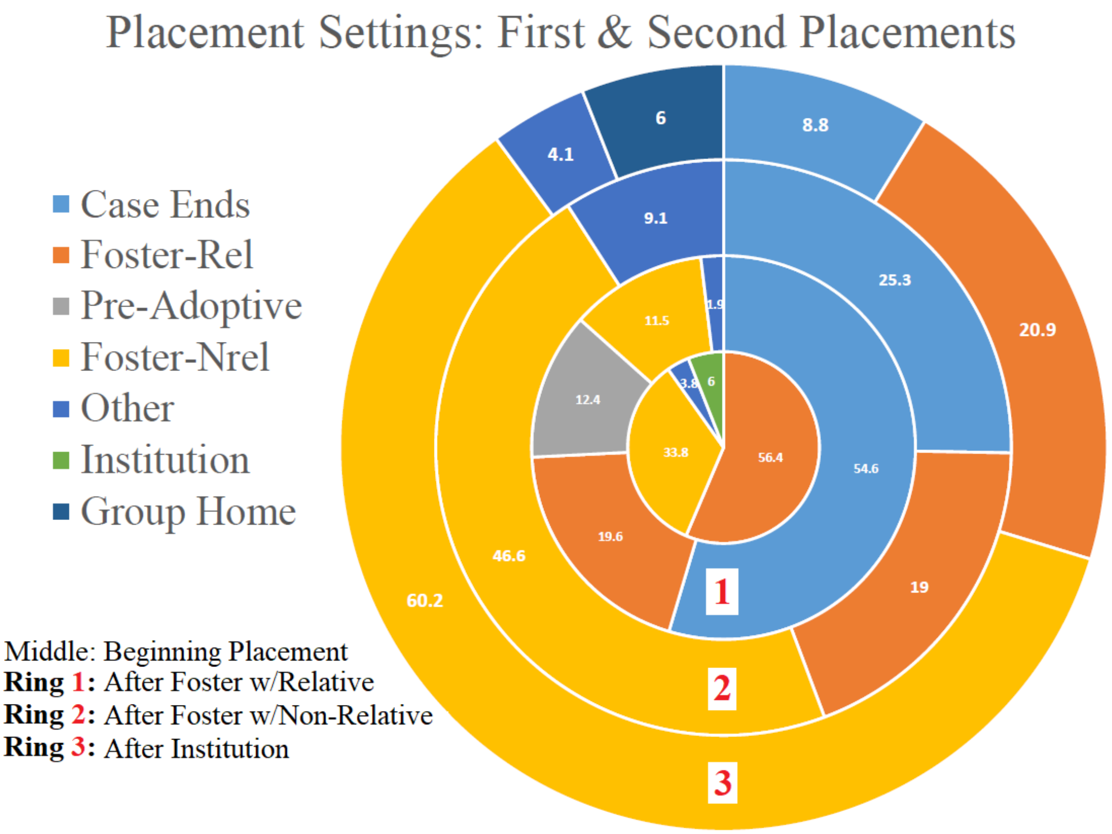


***Figure 5:*** *Expanded pie chart of children's placement settings.*

Only looking at the percentages of each placement setting for a certain point in each child's case (say, second, or third), was not very useful (see pie chart: 90% go to foster homes to begin with). What the researchers *really* needed to see was where the child moved to *after* being placed to start, say, in a foster home with a relative; then what? Do they stay in that same placement setting with a different provider, or are they moved into a group home? Such questions are extremely important when considering the overall case itself.

So, R Studio was utilized to design and code programs that would output the percentages of where children are placed, after a set first placement (that we choose). A multi-layered pie chart visualization was built (Figure 5) that includes the placement settings of the child after the three top first settings (from Figure 4): Foster Home: Relative, Foster Home: NonRelative, and Institution. The outer rings of the pie show the percentage of children that went to the indicated placement setting, after this first setting. To be clear, this only visualizes the first and second placements. The R programs can track any number of cases, but after three or four the amount of data starts to dwindle significantly, so tracking beyond that point has so far been fairly unfruitful.

Having "No Case" is a significant insight for the research team because it means that when the program was run, there was an "NA" in the data for the next case. This means that there is no additional record for this child after moving out of a placement setting. This is good because it can be assumed that the child does not have a need to be re-entered into the system and was provided for properly by the system. One of the original goals of this research was to find factors that influenced this sort of behavior, so, from this visualization, we can conclude the following:

- After Foster Relative - highest chance (55%) of leaving system
- After Foster-Not Relative - high chance (75%) of staying in system
- After Institution - lowest chance (9%) of leaving system
- Highest chance of going to Pre-Adoptive Home after Foster-Rel (12%)

Moving to a Pre-Adoptive home is a huge deal; using these same R programs, the researchers found out that **98%** of children who then move from Foster Homes (Relative) into Pre-Adoptive Homes leave the foster care system. This is absolutely astonishing, so a child getting into this placement would be a success. However, the research team kept in mind that adoption situations are limited in the real world due to the capacity of adoption with regards to adopting families.

Ranking each Child's "Success" in Foster Care System

Considering a child's last record, what was just found were the most successful ways to leave the system; therefore, a numeric "rank" for a child's most recent placement can be created to form a scale of higher/lower success:

| Rank | Details |
|---|---|
| 5. | 100% success: Left system through adoption |
| 4. | Foster Relative setting, 55% chance of leaving |
| 3. | Foster Non-Relative setting, 25% chance of leaving |
| 2. | Institutional placement, < 9% chance of leaving system |
| 1. | All other children who don't fit in ranks 2-5 |

These ranks are inherently limited because they are only considering a child's last placement, however, they do provide a good basis to see the rate of success in a certain subset in terms of leaving the system. It made sense to immediately apply this to the zip code analysis which had been put on hold to dive into the data deeper.

Interestingly enough, this "Rank" factor added into the data proved that the initial impressions about the population density were correct: there is still no significant change between the higher cases per capita areas versus the rest of the data (see Figure 6). In fact, there is a slight spike in the number of cases with Rank 5 (Adoption) within what was initially denoted as the "problematic" areas. Now, this does not discount that there could be other reasons why these areas are so riddled with cases, but it may be one of economics that had not been considered yet. This seems like the most likely case, considering that no significant differences have come out of subsetting these areas out of the data for comparison, or even now with this much more complicated approach of a ranking system.

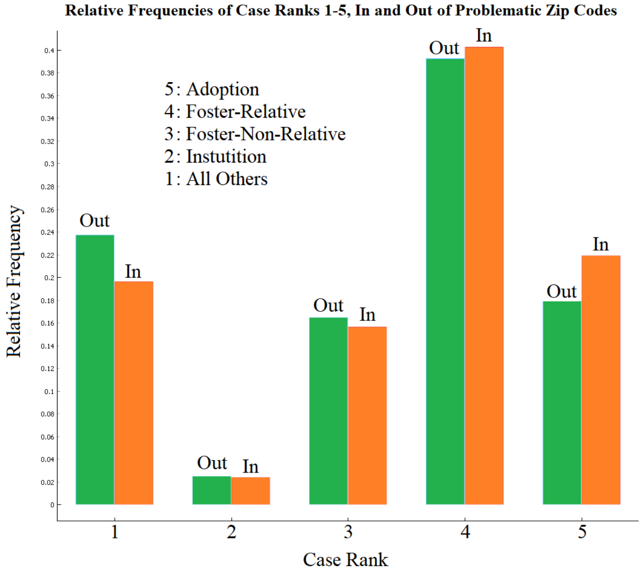


***Figure 6:*** *Bar chart of Ranks in and out of "problematic" zip codes.*

## Weighing the Entirety of a Child's Case

As mentioned above, the ranking system is inherently flawed for deeper analysis due to the fact that it's based solely on last placement for each child in their case. The researchers realized that there was a fairly simple way to get a numerical factor that was directly related to not only the types of placement settings that the child had, but also how many. All that was needed to be done was to assign a number to each type of placement setting, and then for each time it came up in a child's records, add that number to this new factor called "Weight".

The researchers set up this algorithm in R so that a higher weight was less of a success than a lower case; therefore, inversely, higher Ranks like 5 (Adoption), had a lower Weight number that was added for each time a record with Adoption came up in a child's cases. This way, if a child has a high number of cases, this also influences the weight; for example, a child's situation where they were jumping back and forth from many different placements will show up as a fairly high weight, signifying that their case was not particularly successful. This is *precisely* what is missing in the "Rank" factor: the inability to look at the progression of records as a whole for the child.

## Extrapolating Numerical Values from the Datasets

Now with these Rank and Weight values, it became necessary to start accumulating numerical columns of data that may possibly correlate to these, into a single dataframe. A dataframe was built entirely of numerical values including the Identification ID, Case ID, Rank, Zip, Zip Density, Number of Cases in Zip (ZipCount) and Weight to start. However, this did not seem to be substantial enough for the amount of data that was given, so more numerical values were sought out that could be programmed into R to find and record as part of this new dataframe. A few more obvious ones were added in, like Number of Participants in a case, and Number of Cases per child, along with other less obvious ones, like number of caregivers per child, the total duration of the case from start to finish, and even the exact age of the child. Their ages were not directly provided, but instead a birth date. By running R loops using this value to the year of the case, the age for each child at the time of the case was calculated. A similar loop was programmed to find the average age of the caregivers of a child. In the end, there were hundreds more lines of R code programmed to extrapolate as much numerical data as possible from the limited amount of numerical columns that the data started as.

## Deeper Statistical Analysis: Plotting of Extrapolated Numerical Values

One of the first insights that was clearly able to be seen based on these numerical factors was the relationship between age and rank. As the ranks become more successful, the average age of the child decreases. This is significant because it gives the research team a deeper look into who is actually filling up the higher ranks: in this case, younger children. To add onto this, one of the initial assumptions that was made before running through the statistical analysis was that the children who were younger would have more cases, due to the fact that they had more time to get placed around the system, switch caregivers, and more. However, when the relationship between Number of Cases and Rank was plotted (Figure 8), the researchers found that those children who were young and in the Adoption Rank did not have a significant number of cases higher than the other ranks; in fact, the children who ended up in less successful ranks were both older, and had longer cases overall. Bearing in mind that the length of the case was for their entire history of records in the system (so an older child does not necessarily mean longer case duration), this is particularly interesting to see because it would make more logical sense to the researchers if the adoption children had in fact been bounced around the system a lot and then finally moved to a Pre-Adoptive home. But this is not the case; even the Foster-Relative placement (which filters a good portion of its cases into Pre-Adoptive homes) has the *lowest* average number of cases of all the ranks.

In addition to these plots, the researchers also were interested in seeing how the number of caregivers affected any of the factors about the child's case. The researcher team found out that an increase in the number of caregivers synced up well with an increase in the "weight" of the case (See Figure 9). Initially, this made perfect sense because the weight was directly

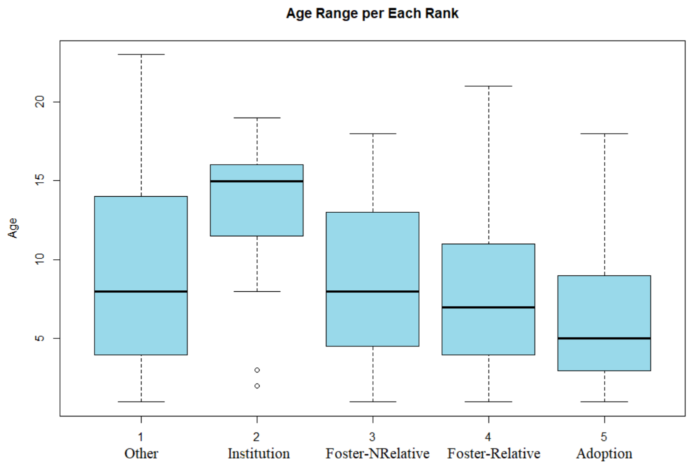


***Figure 7:*** *Box-and-whisker plots of Age range per Rank.*

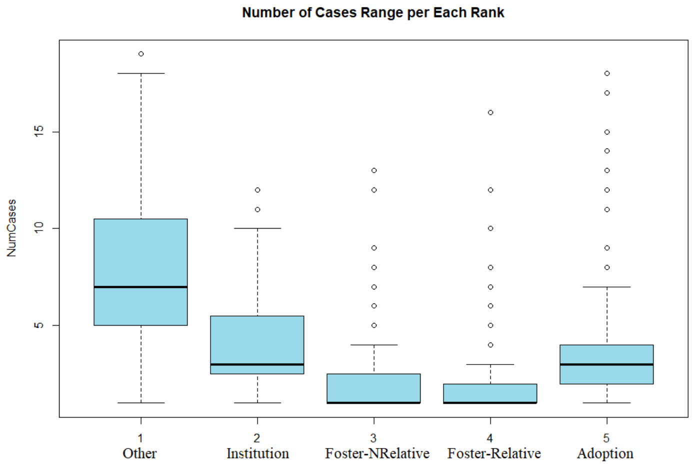


***Figure 8:*** *Box-and-whisker plots of Case range per Rank.*

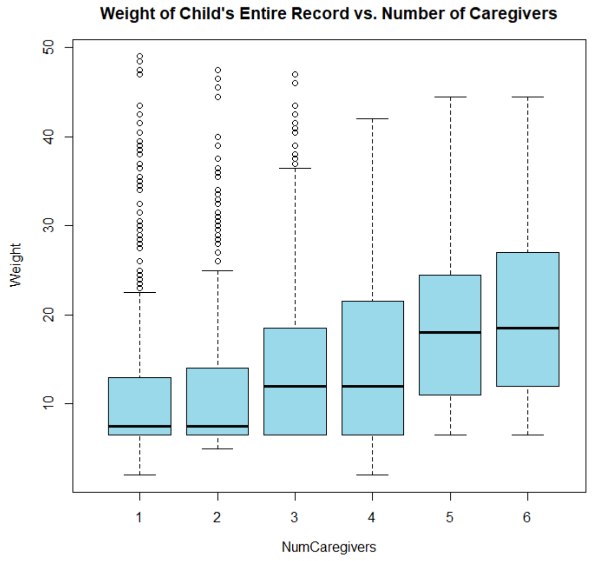


***Figure 9:*** *Box-and-whisker plots of Cases and Caregivers.*

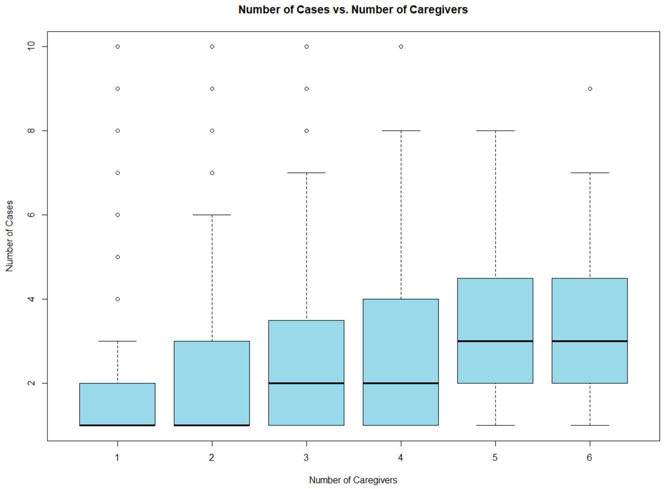


***Figure 10:*** *Box-and-whisker plots of Weight and Caregivers.*

connected to the ranks of the child's placements, as that was how it was calculated in the dataframe of all numerical values. However, there was not as strong of a correlation between the number of cases and the number of caregivers (See Figure 10). This could imply that if the caregivers are constantly switched around in a case, the child's ranks tend to be less successful, but further analysis was needed to prove or disprove this theory. But, as shown, there is not a significant change in the number of cases (Y) as the number of caregivers (X) increases: only about 1-2 more cases for those reaching higher numbers of caregivers.

## Machine Learning: Building Neural Networks

The researchers were striving for under 10% error with the neural network machine learning scheme, and finally achieved this with the addition of the numerical column, Number of Caregivers. After that, another column for the ages of the caregivers was added in, and this sent the error of the neural network down into the 5-6% range. This network below (Figure 11) in particular trained on 70% of the data and then was tested on the other 30%, proving that it is extremely accurate.

The real use of this neural network is not the graphic itself but instead is the equation that the neural network computation achieves in R. Any full entry of the data with those eight numeric values filled in could be put into these program, and then the neural network would compute a number that was its prediction for the Weight of the case. Even if the case details were created with a bunch of random numbers, it would still predict based off of the data that it had been trained on, in this case, 70% of the dataframe with just numeric values. Therefore, it could predict whichever factor was wanted; the answer would be in a numeric format as well. The R

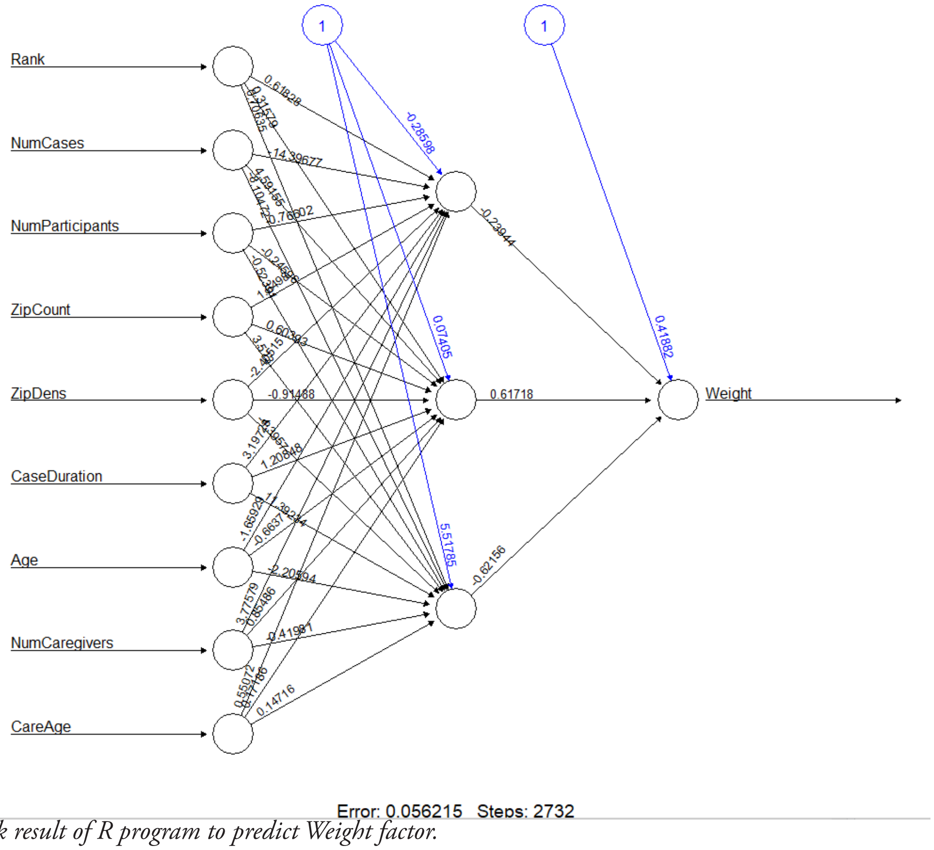


***Figure 11:*** *Neural network result of R program to predict Weight factor.*

package used ("neuralnet") also plotted the accuracy of the computation's prediction for each result in the 30% of the data which it tried to predict. It plotted the line y=x in order to display a fit line for a perfect prediction (Figure 12). Even while leaving the extreme outlier in the data (seen on upper right), the prediction was surprisingly accurate, as shown.

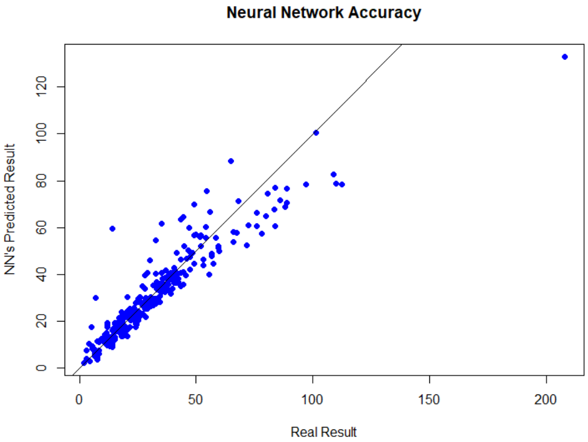


***Figure 12***: *Neural network accuracy plot.*

## Machine Learning: Building Decision Trees

Modeling the numerics dataframe with decision trees gave the researchers a deeper insight into correlations with the numerical columns, however, they were slightly limited in reach and scope due to their creation of more basic, discrete distinctions between cases. For example, here it is obvious that more cases increases the average weight of the case. And, more participants appears to also correlate with higher weights. This makes logical sense, however, the actual numbers are very rigid and the researchers have to consider that they are only getting the average weight for each of these leaves on the diagram. The decision trees still provide good insight and the researchers can see correlation, but they were generally not as insightful as the other programs. Nonetheless, it was yet another way to visualize the data that from the beginning looked very difficult to look at on a broad scale without giant diagrams and borderline confusing subsets.

Each level of these diagrams adds to 100% (the percentages are the bottom number on the leaves), however, keep in mind that the program, in order to fit the percentages on the leaves, does round up. The diagram below (Figure 13) was designed to display the average weight of a sector of the data that had a certain number of participants or cases. The number of cases had a stronger correlation with the weight than the number of participants did, so those were used as the uppermost distinguishing leaves and then number of participants was worked into the ones closer towards the bottom (shown in Figure 13 below).

## Machine Learning: Random Forest Analysis

One of the most useful machine learning schemes used was the Random Forest model, specifically because it gave the *influence* per each characteristic that the model used to predict off of. This would directly give the factors that had the most influence on a case's success or not. In Figure 14, the weight is being predicted for a case, and the most influential factors were Rank and CaseDuration. This is logically obvious because the

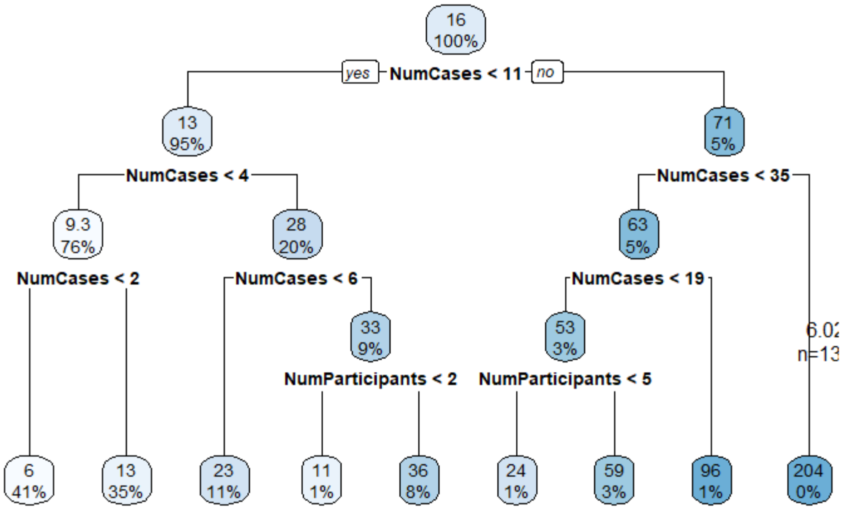


***Figure 13:*** *Decision Tree result from R program: predicting Weight.*

weight is built off of rank: if the rank is higher, and the case is longer, then the weight adds up all those ranks to a sum value that will be much higher. So those influences can be ignored to look at the next most influential. There are really three present here: Zip Density, Age, and Number of Participants.

These are all fairly intriguing. First of all, there is still something deep about the zip code density that is being overlooked because so far the researchers still have not been able to see what was particularly different about the "problematic" zip codes. Age was surprising, however, the researchers did see that younger children have better outcomes. The number of participants the researchers also did see effect the weight in the decision tree. What is most interesting is that all three of these affect the random forest model in nearly the same intensity, or level of influence.

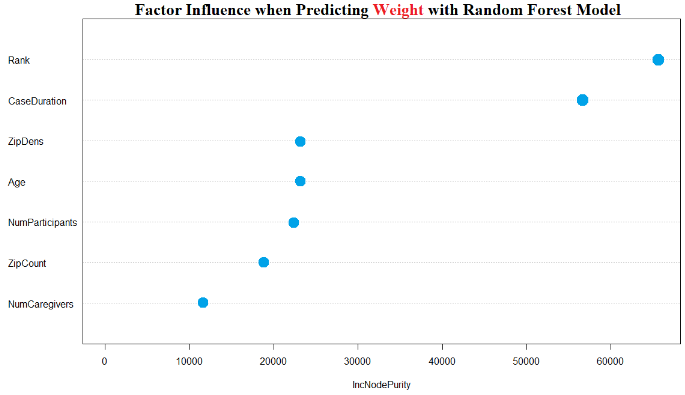


***Figure 14:*** *Random Forest model result: predicting Weight.*

Predicting the duration of the cases was also attempted, and the results with the Random Forest model were just as surprising. Of course, a younger child has more time for additional cases, which was expected, so "Age" came out on top. However, again, the Number of Participants involved, and then zip density showed up as the second and third most influential factors in the data, this time all within close proximity to each other on the chart (See Figure 15).

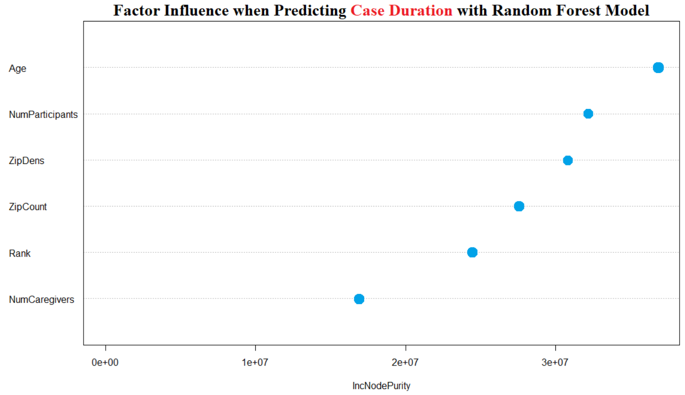


***Figure 15:*** *Random Forest model result: predicting Case Duration.*

### Final Random Forest Model to Predict Case Success

After seeing the capabilities of the random forest model in terms of it ranking how important each detail of a child's case was, the researchers wanted to robustly use it to determine which of these details were most important, as shown, and then determine if these details had *positive* or *negative* correlations with the weight. This is important because the above charts only tell what is important in terms of the detail of a child's case. What is *really* the objective of using this model is to figure out *how* these characteristics affect the result of a child's case in the foster care system. More details about the cases were extrapolated, and the random forest model was run once more.

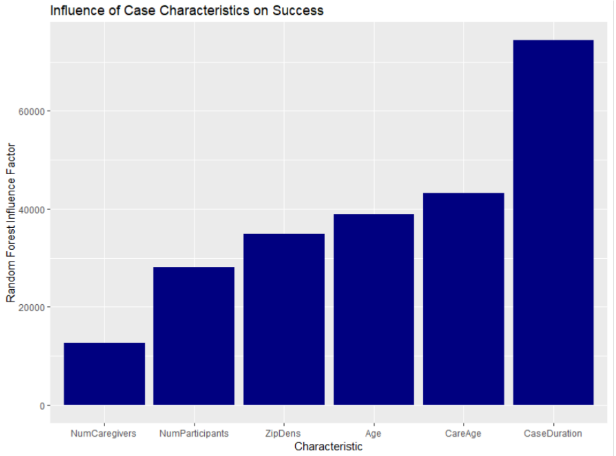


***Figure 16:*** *Random Forest model result: predicting Weight with more extrapolated characteristics (this is a small portion of the graph outputted).*

Now the researchers have the top influences on a child's success in the foster care system. These were taken from 15 different characteristics that were conglomerated and ran through the random forest model. The top five were used for analysis:

1. Duration of their case
2. Age of the caregiver
3. Age of the child
4. Population density of the zip code
5. Number of participants in their case

These are concrete results, but, are they positively or negatively correlated with the weight? Is a longer or shorter case more successful? Older or younger child/caregiver? City or rural population density? The answers to these questions would be the final results.

## Results

| Negative Influence | Positive Influence |
|---|---|
| Shorter case duration<br>Lowest success: 360-380 days<br>15% more placements within case | Longer case duration<br>Highest success: 410-430 days |
| Younger caregiver (≤ 49)<br>26% less successful case | Older caregiver ( > 49) |
| Older child ( > 7)<br>Median age in Institution: 15 | Younger child (≤ 7)<br>71% higher chance of adoption<br>Median adopted child age: 5 |
| Higher population density<br>11% shorter case duration | Lower population density |
| Higher number of participants ( > 13)<br>33% less successful case | Lower number of participants (≤ 13) |

***Figure 17:*** *Overall insights into positive and negative influences on a child's success in the system. T.*

### Analysis of these Influences

These five influences were pulled from the random forest model's results. Taking the mean of each influence (average caregiver age was 49, child age was 7), and sub-setting the records for the participants who were above or below these averages, and comparing the weights, resulted in the above table. The best case scenario for a child would be to have been entered into the system early, have a longer case with an older caregiver in a more rural area with less people involved in the case. Many of these details are available at the beginning of a case, so from the start they can be used to "red flag" children who display more or less negative or positive characteristics.

The original research question was: Is it possible to employ statistical analysis techniques and machine learning methods to pinpoint specific traits of a child and/or their case(s) that could be used to accurately predict if the child leaves the foster care system or is re-entered again at a later date? These specific traits were clearly found, and this research was extremely successful.

### Further Results

This research paper is a lot of charts, statistics, analysis, and information to take in at face value, however, the objective of this research has essentially been fulfilled: the researchers now know identifiable factors that lead to children being re-entered into the foster care system, and have a fairly solid grasp on how to get them to leave once entered. To clearly understand these results, let's look at what clearly works out best for a child.

First of all, it is clear that Foster Homes with a Relative provide superior service and a higher chance of a better overall case than Foster Homes with a Non-Relative. Proof of this is in the pie chart: those children who move into a Foster Home with a Relative have a 55% chance of leaving the system after that placement (compared to ~25% of Foster-Non Relative placements), and the highest chance of moving to a Pre-Adoptive home, which the researchers already determined was extraordinarily significant, with 98% leaving. Beyond these metrics from the pie chart, Foster-Relative placements also have the lowest average number of cases, signifying that those who end up there are not bouncing around. To add onto this, the pie chart also confirms this because the Foster-Non Relative children have a much higher chance of staying in a Foster Home (66%) than those with a relative (31%). With all of these metrics, the researchers can definitely say that Foster Homes with a Relative are a much better choice than Foster Homes with a Non-Relative, and that welfare service providers should emphasize these placements with a relative.

Another insight already mentioned was the prevalence of children who left the system after being put into a Pre-Adoptive home (98%). A suggestion for child welfare services providers would be to get children into this placement setting in any way possible. This, of course, could be limited due to Adoption capacity of an area (there are only so many families ready to adopt), or other real-world logical barriers. However, this placement was what the researchers considered the most successful. Outside of just simple logic assuming that these were good cases, when the research team ranked the children's last cases, they found that Rank 5 (Adoption) had a relatively low number of cases. This is a good thing and shows that for many of the children, they were provided appropriate service being put into this placement setting.

Alongside of this insight, the ages of adopted children were the lowest of the entire dataset, and this shows that these children were given this service earlier in their lives instead of later; the researchers found that as the cases became less and less successful, the ages of the children increased. An older child is much more likely to be put into an Institution or with a non-relative (see box-plot for Age vs. Rank) instead of being adopted, so early action is another key piece to this data.

As a final note, the researcher team did notice correlation between the weight of a child's entire case history and the number of caregivers that they had, even if it was not clearly seen via the random forest machine learning model. As their weight increases, they also are more likely to have more and more caregivers associated with their case. The last suggestion would be to emphasize keeping caregivers instead of assigning new ones. Now, the weight is calculated based on number of cases too, however, if the child had what the researchers consider "more successful" records, even in the presence of multiple caregivers, their weight would still be low. Seeing this correlation shows that the cases were not as successful, and the caregivers changing multiple times may have affected this. The obvious logic behind this was "more cases mean more caregivers", however, this was disproved when box-plotted this relationship for Number of Caregivers vs. Number of Cases, and saw no significant upwards increase with the cases as caregivers increased.

These results come with a caveat: correlation does not mean causation. The researchers worked through a lot of the "real world logic" that could cause these results to happen, and the insights appear robust, however, it is worth pointing out that these correlations may not be the causes of the factors changing in the end (ex. weight). However, these were the strongest insights that the researchers got from the data set.

## Discussion

All in all, predicting a child's path or end result in the foster care system is still very complex and the few distinct insights that the researchers came up with are still "watered down" simply by the sheer number and prevalence of basic foster care home placements without much additional detail, regardless of what number of records the research team look at. This is to be expected from the data, however, and proves that the system itself is still working by placing children into foster care homes, albeit not perfectly. Doing predictive analytics on the datasets that the researchers were given from Partnership for Strong Families is an ongoing effort at Embry-Riddle Aeronautical University, with multiple students taking different approaches to do so. The researchers approaches here were fairly deep and complex, but also required them to create a lot of the numerical values that are in the results (number of cases, caregivers, case durations, etc.), and a significant portion of the data had to be removed due to missing entries. This opens up some degree of error in the calculations, assumptions, and insights, however, the researchers showed that, given the absence of this data, many of these models are very accurate for the factors and data sets. Missing data has always been a problem in data analytics, and these datasets were no exception.

In terms of future research, the researchers would definitely recommend the approach that they took involving "extrapolating" more data out of the existing data: the number of participants, the case duration, and more. This worked wonders on the ability to analyze the data more deeply. Without doing this, the research team members were completely lost as to where to look and what to even look for. With the numerical factors present, the analysis sped up and became more insightful. Another recommendation would be to consider the individual factors of the parents or caregivers more closely; in the analysis the researchers only looked at their ages, but more information about them (background, salary, demographics), might correlate to more or less successful cases. The researchers put a lot of emphasis on the child themselves and their individual placements, but this is obviously only one piece of the puzzle. Overall, however, the entire team were equally impressed at the capabilities that R gave to analyze this data set, and were surprised at many of the insights that the researchers were able to come up with.

## Works Cited

Scharenbroch, C., & Park, K. (2017). Principles for Predictive Analytics in Child Welfare. *National Council on Crime & Delinquency.* Retrieved February 5, 2019, from https://www.nccdglobal.org/sites/default/files/inline-files/Principles for Predictive Analytics in Child Welfare-1.pdf.

Teixeira, C., & Boyas, M. (2017). Predictive Analytics in Child Welfare: An Assessment of Current Efforts, Challenges and Opportunities. *US Department of Health and Human Services*. Retrieved February 1, 2019, from https://aspe.hhs.gov/system/files/pdf/257841/ PACWAnAssessmentCurrentEffortsChallengesOpportunities.pdf.