I started with a dataset that had too many columns, where each center had multiple columns instead of rows. This made it difficult to work with and analyze because the centers were defined by columns rather than rows, leading to a very wide dataset. To make the data easier to handle, I transformed it into a long format where each row represented a center, and the variables were more organized. This change allowed for easier analysis and a clearer understanding of the data.

It is worth looking at an overview of the dataset since the quality of the analysis that's coming up is heavily dependent on the quality of the data. I noticed several discrepancies in the data. For example, there is a center that doesn’t record their counseling hours, and most centers don’t record the hours spent on other program activities. One center even recorded the number of clients using decimal places. So, if they had three full-time consultants and one part-time consultant, they entered that as 3.5. Each center records information differently, but I did my best to standardize the data so it’s consistent across all centers. This ensures the analysis reflects a clearer picture of the actual

In this analysis, we used a statistical model to estimate the number of counseling hours each center should be spending based on various factors, such as staffing levels, the population they serve, and the hours spent on other program-related activities. These estimates are designed to help us understand if a center is operating as expected in terms of client support or if there are discrepancies. For example, the model takes into account how much time consultants are spending on activities other than direct counseling, which gives us a more comprehensive view of how efficiently a center is operating.

The actual counseling hours column reflects the hours that centers have reported, while the estimated hours are what the model predicts based on the input data. By comparing these two, we can see whether a center is meeting expectations or if there are gaps in their reported hours. This helps us understand how Nebraska SBDC centers compare to others in terms of efficiency and whether they may be understaffed or overstaffed.

In terms of staffing levels and client hours, this approach allows us to see how Nebraska SBDC centers perform relative to others across different states. If a center is spending significantly fewer hours with clients than expected, this might indicate an issue with staffing or data reporting. Conversely, if a center exceeds expectations, it might suggest they are either more efficient or have more resources than their counterparts in other states.

Now that we’ve discussed the predictions and the differences between reported and estimated hours, let's transition to a key part of the analysis: understanding which factors most influence these predictions. To do this, I created a variable importance plot. This plot shows us which variables have the most significant impact on the model’s predictions of counseling hours. The higher the importance score, the more influence that variable has in determining the predicted hours. For example, if staffing levels rank high, it means that the number of staff plays a large role in how the model estimates the time spent with clients.

Next, I’ve included partial dependency plots for the top four most important variables. These plots help us visualize how each of these variables affects the predicted counseling hours while keeping all other factors constant. Essentially, they show us the direct relationship between each variable and the predicted hours. For example, we can see how changes in staff size or time spent on other program-related activities affect the expected counseling hours. These plots help us understand the nuances of the model and give insight into how specific factors contribute to the overall counseling time estimates.