Water Quality in the Niobrara River - Using Dimensionality Reduction as a WQI

Data Source:
https://catalog.data.gov/dataset/niobrara-national-scenic-river-2001-2023-water-quality-data-from-the-niob0001-project-wat-

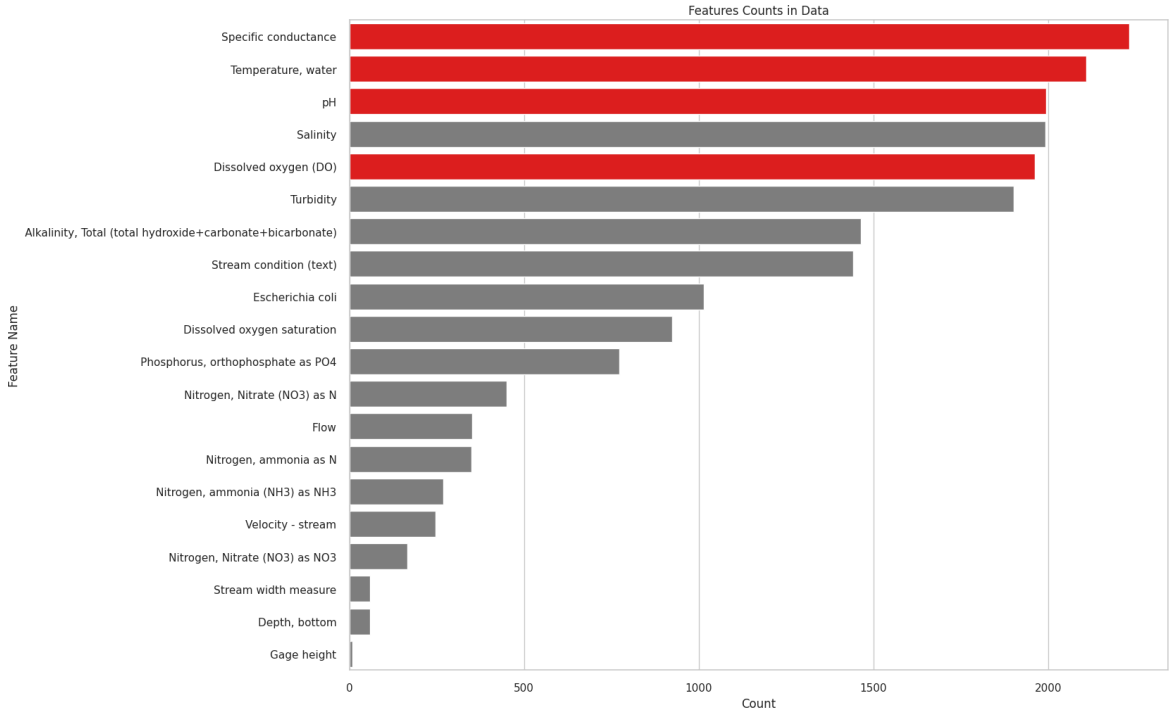
Introduction -

The dataset chosen was data collected by the Department of the Interior (DOI) on behalf of the National Parks Service (NPS) on different water quality indices found in multiple rivers across the US. Originally, the plan was to use a dependent variable within the data as a categorical "Water Quality" measure, but as it turns out through both research and analyzing all the data, there is no all-encompassing measure of water quality. Therefore, the project scope was reduced to look at one of the datasets included, the Niobrara River. The Niobrara River is a national park in North Nebraska that claims that it has excellent water quality (taken from the NPS website for the Niobrara River). However, there are no quantitative reasons to back this claim up, so this project will aim to address the water quality in the river.

Upon analyzing the data, it is clear that there is no target variable that can be chosen for traditional classification or regression machine learning models. Of the unsupervised machine learning techniques we are aware of, clustering and dimensionality reduction, it would be easiest to capture the data using dimensionality reduction. Using dimensionality reduction, we can measure and visualize the distance between higher-dimensional data over time using the DOI dataset.

Analysis and Visualization -

Figure 1: Key features in the DOI dataset, and feature selection



The dataset was split by which spot in the river the water quality metrics were measured in. Ideally, the “model” for the dimensionality reduction should be as generalizable as possible, but in the interest of including as many data points as possible, we chose the most characteristic variables that appeared the most times (across all data collection locations). The chosen variables were:

1) Specific Conductance (µS/cm) - a measure of conductivity of the water. Additionally, it can measure human pollution and is highly correlated with salt/ion content.
2) Temperature (degrees Celsius) - measures the average temperature of the water at the testing site. Very variable by time of year measured, but long-term temperature changes are linked to habitat changes.
3) pH (0-14) - measures hydrogen ion content (or acidity), is also variable but long-term changes can be linked to habitat changes and potentially harmful to organisms in the water body.
4) Dissolved Oxygen (mg/L) - measures the productivity of producers, such as algae and submerged vegetation. Increased DO is linked to the water being a carbon sink and high water quality, while lower DO is linked to being a carbon source

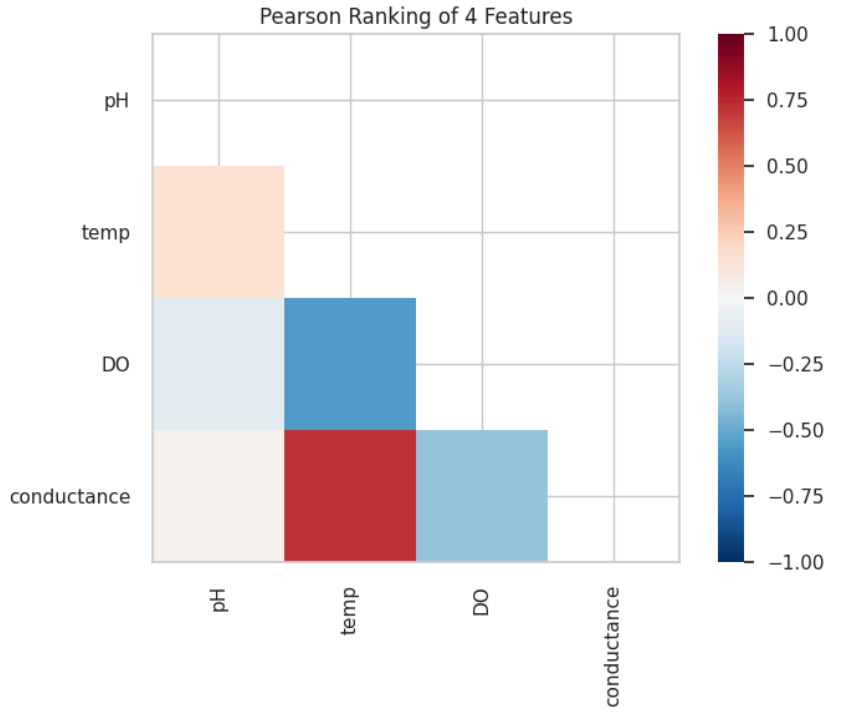
Variables that were cut were based on a couple reasons. For one, features like salinity and alkalinity are features dependent on which part of the stream they are looking at. Secondly, variables such as “stream condition” and “turbidity” were either text-based or observational, without concrete data behind it. For example, turbidity is a feature that’s recorded by handwritten notes about what the researcher sees in the water, instead of a quantifiable measure like total suspended solids per liter or secchi depth. Finally, many of the features that could be useful (eutrophication measures like nitrogen and phosphorus or e.coli measure) were cut because they just weren’t measured enough times and at enough of the locations for it to be worth looking into.

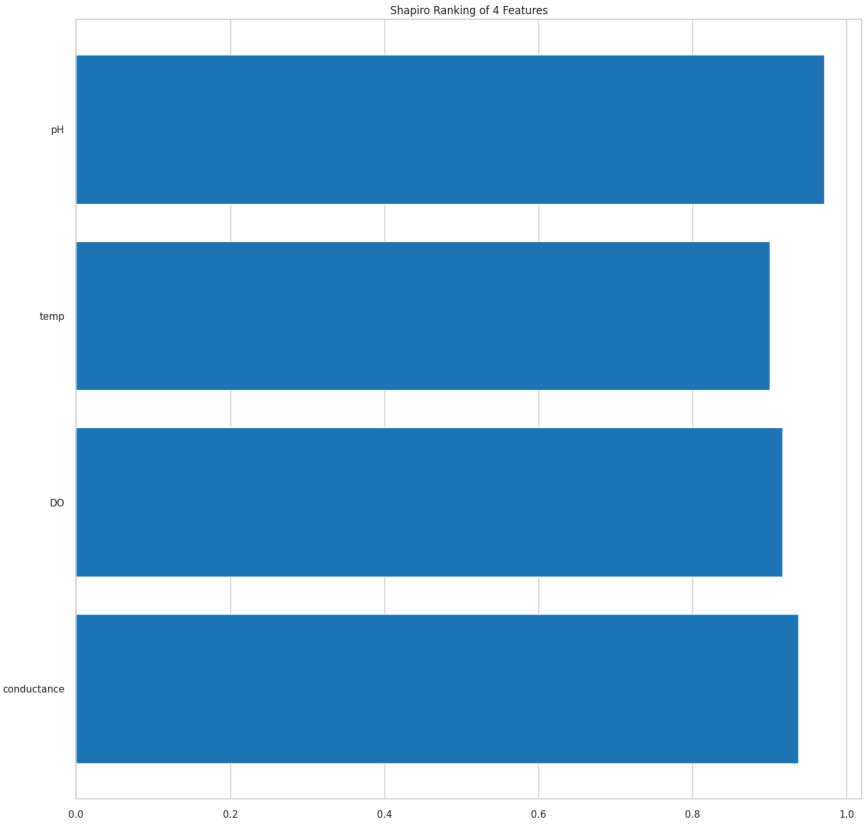
Figure 2/3: Time Series Analysis of the four chosen features over time for the Berry Bridge

These figures are meant to analyze the overall trend over time for the four chosen variables. The data was specifically chosen from the most commonly measured location, the Berry Bridge. Some key takeaways are that when you zoom in (looking at figure 3), there is a clear decrease in water quality. That is: pH trends to be more acidic, temperature trends higher, DO trends lower, conductivity trends higher. Figure 2 is meant to illustrate that there is some nuance to the data, and clear evidence of periodicity within both a single year, and between multiple periods of years (notice how pH lowers but increases again, and how temperature increases and then decreases).

All bodies of water everywhere will experience periodicity within a year, with higher temperatures/conductivity and lower DO in the summer and vice versa in the winter. Focusing on this might be interesting, but it doesn't add or detract from the point of the project.

Figure 4/5: Correlation Matrix and Shapiro Ranking

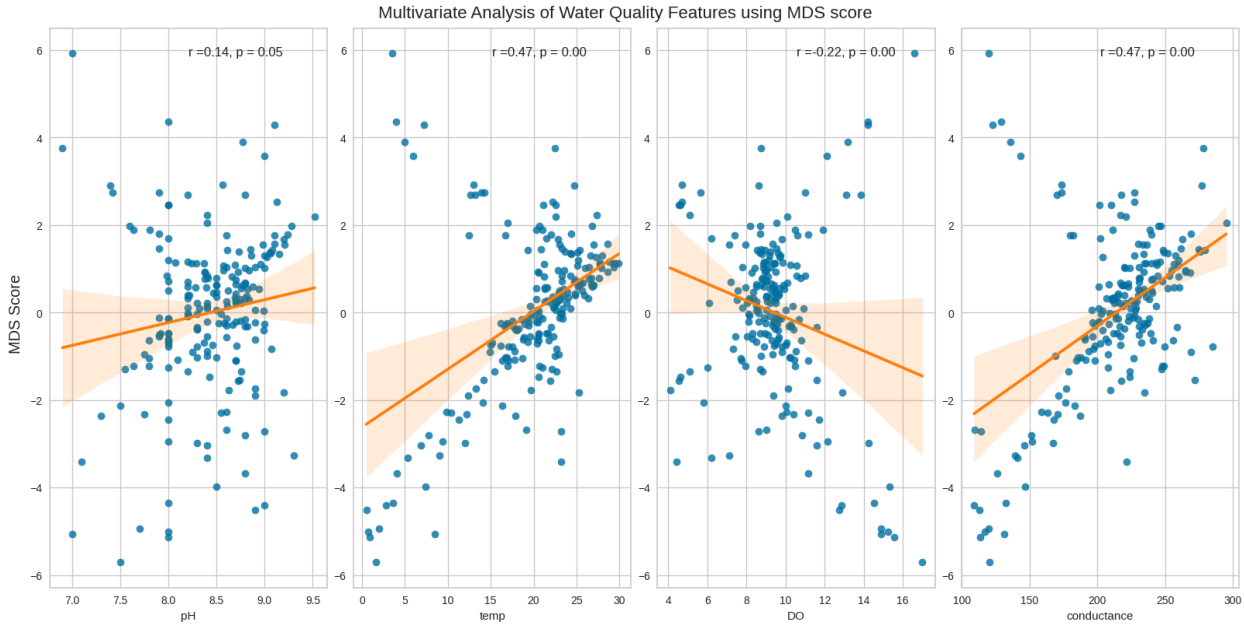


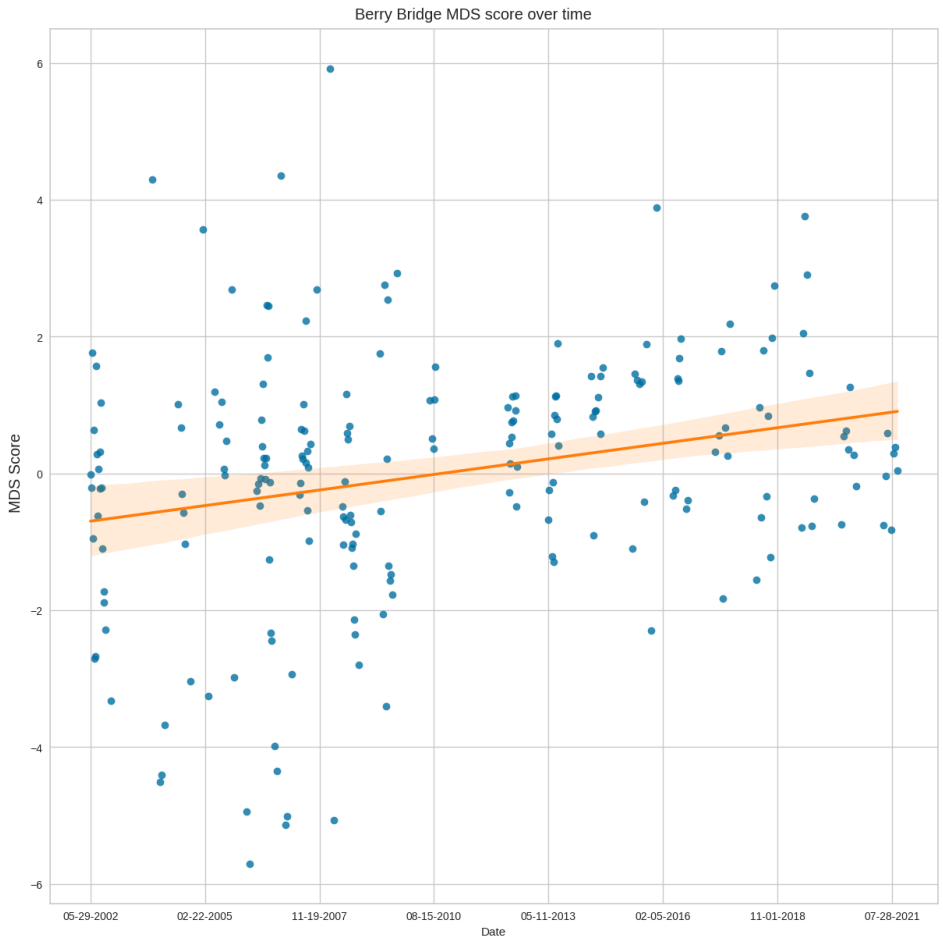


These plots are precursors to choosing which dimensionality reduction technique will be used. The first plot is a heatmap of the correlation matrix between the four chosen variables. The main takeaways from this plot is that although there are some co-correlated variables (temp/conductivity and DO/temp, which make sense biologically), there is not enough correlation to warrant using PCA. If these variables were strongly correlated, PCA would result in a first principle component with a majority explained variance, but without that, we can use MDS instead to measure distances between points.

Model Evaluation -

Figure 6/7: Results of the MDS index for the Berry Bridge





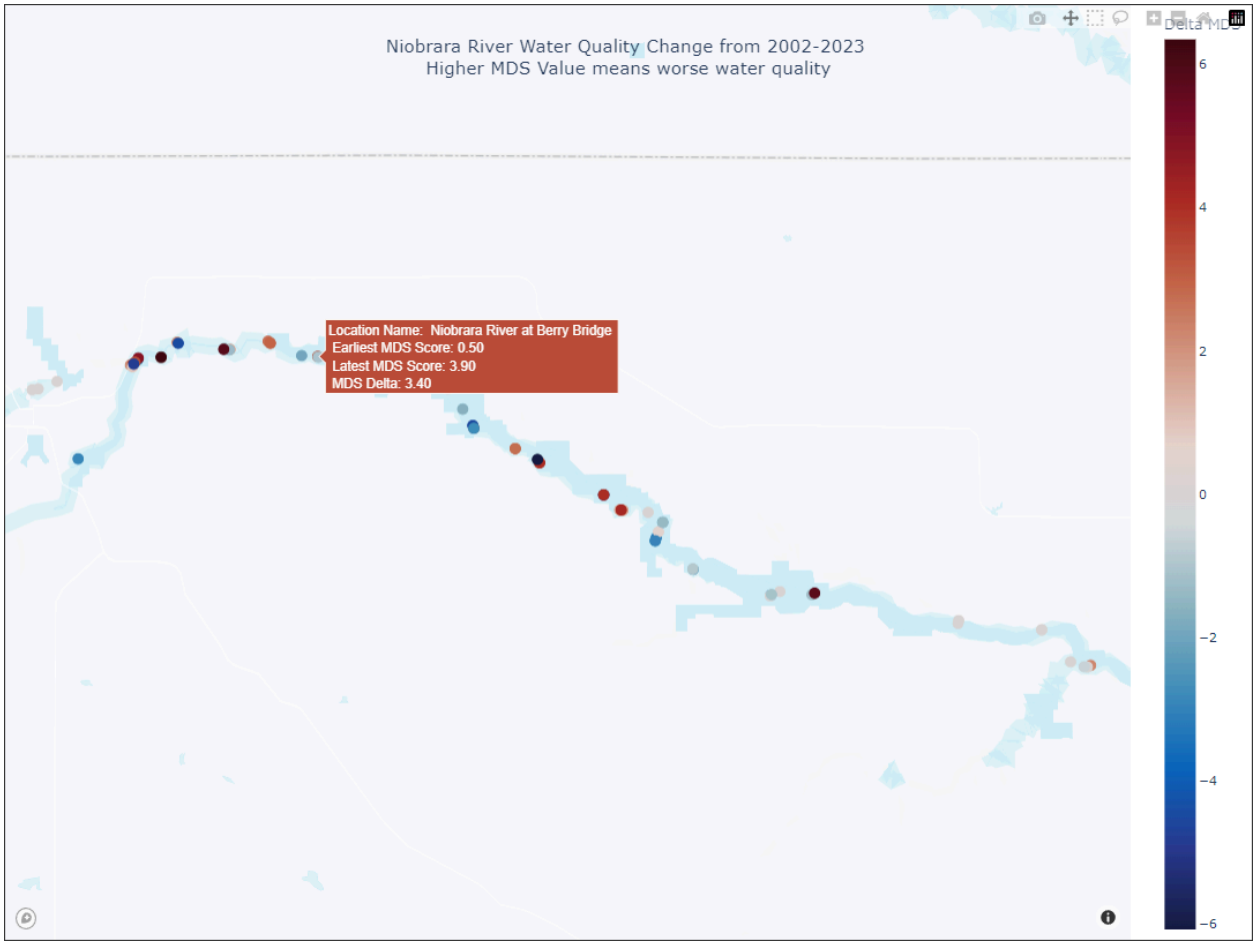
These figures outline the results of running the MDS model on data from the Berry Bridge. The first figure is a multiplot of the four variables and the MDS score correlated with it. In the top right are Pearson correlation scores and their significance, important in showing the relation between the variables and the MDS score. According to both the graphs and the correlation scores, an increasing MDS is correlated with increasing pH, temperature, conductance, and a lower DO. As discussed at the top, these are all potential measures of lowering water quality, so we can correlate an increasing MDS score (so a delta of above 0) with a decrease in water quality. However, this is just a correlation, that is while significant, based on quite low correlation values.

The second figure is showing the same data plotted over time, so the MDS score can be shown changing over time.

For both of these plots, the y-axis is centered at 0 specifically because 0 is what the MDS score considers as “average” - as the different scores will be the distance from 0. Centering the graph at zero MDS also makes it easy to see that the regression lines are crossing, for example, negative MDS to positive MDS.

Conclusion -

Figure 8: Interactive Water Quality for the Niobrara River Plot (2002-2023)



Interactive Plot link:
https://colab.research.google.com/drive/1oc_75LbI0vEgJj0nnIt2MNFJ4egOuXIg?usp=sharing

This is the final plot that our model created when trained on each individual data collection location, finding MDS scores for the first and most recent data points, and finding the difference between them. According to our model, if the MDS delta is above 0, that means the MDS has increased over time, so the water quality has dropped. If the MDS delta is below 0, the MDS has decreased over time, and the water quality has increased. While not a perfect model, this opens the door for much more quantitative analysis on water quality, and the model can be quickly improved by way of introducing more features, and more frequent data recording. The model

isn’t as concrete as it could be, though. Currently, the deltaMDS value only takes into account two data points, when something more rigorous could be chosen (although more ecological background knowledge would be needed). Additionally, not all locations are equal, as some spots only have two data points, and many of them are weighted more in the summer times. Overall, many points of the graph show red and dark red spots, but there are lots of the graph that shows gray. The gray points indicate (ignoring the lack of much data) the water quality staying the same. Looking at the map subjectively, it looks like there is slightly more red than there is blue, but that’s not a cause to say that the entire river is decreasing in water quality. However, a particular strength of this interactive plot is you can specifically see which locations are lowering in water quality, and target them for more research or restorative actions. For example, Berry Bridge, Brewer Bridge, and other popular hotspots for human activity seem to have higher MDS delta scores, so these could be locations that researchers or NPS workers can target.