Data Science Coding Challenge
Anomaly Detection
Version 1.0

Adam Stewart

ExtraHop

February 11, 2019

In the Oxford Dictionary, an *anomaly* is defined as "something that deviates from what is standard, normal, or expected [1]." In this coding challenge, a 'real world' dataset was given with the intent of having all the anomalies detected and recorded. Additionally, a cleaned version of that dataset had to be created. One of the obstacles in this challenge was that very little context was given to explain what the data represents. Though detecting the anomalies was relatively straightforward, some assumptions had to be made to define a "cleaned dataset" for the sake of this challenge.

I began the challenge with Exploratory Data Analysis (EDA). EDA is generally defined as the phase of data analysis for finding anomalies, determining potentially appropriate models, and finding relationships amid explanatory variables [2, pp. 75]. I kicked-off the EDA phase by opening the dataset and creating a plot using the Matplotlib Python library:

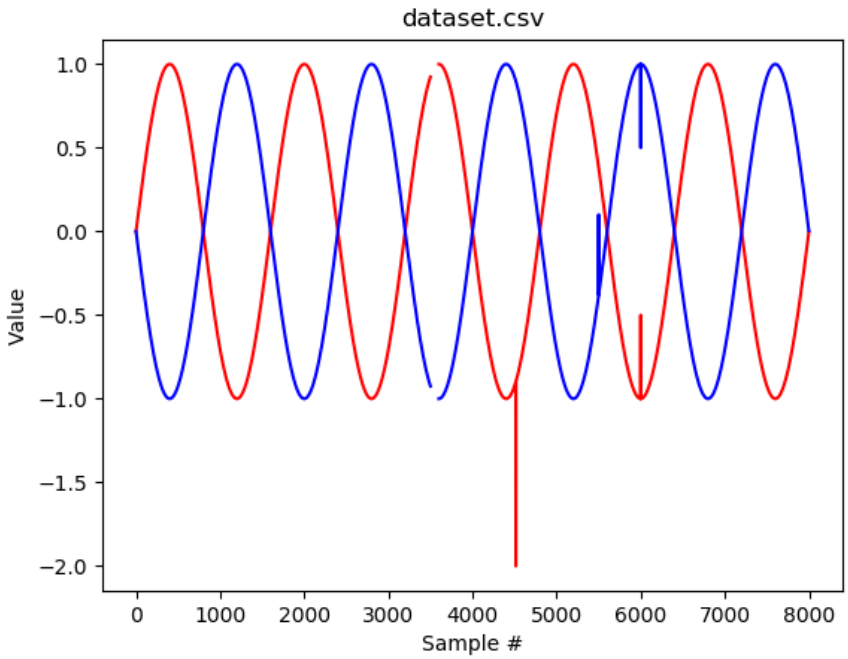

Figure 1: This is a plot of the original "dataset.csv" before any modifications were made.

Immediately, it was clear that the two sensors formed intersecting, sinusoidal patterns. It also appeared that there were at least 4 anomalies. I proceeded to open the "dataset.csv" in Microsoft Excel to get a closer look at the anomalous values. The values from sample number 3501 to 3599 were not simultaneously empty as the plot might suggest. Instead, only one sensor had a value at a time, alternating back-and-forth throughout the anomaly. Sample numbers 4515 and 5500 both had singular anomalous values as the graph suggested. However, the anomaly found at sample number 6000 contained mirrored, anomalous values, which could indicate a detection rather than an error to be omitted.

Since the anomaly found at sample number 6000 could have been either a detection or an error, I decided to create two cleaned versions of the dataset:

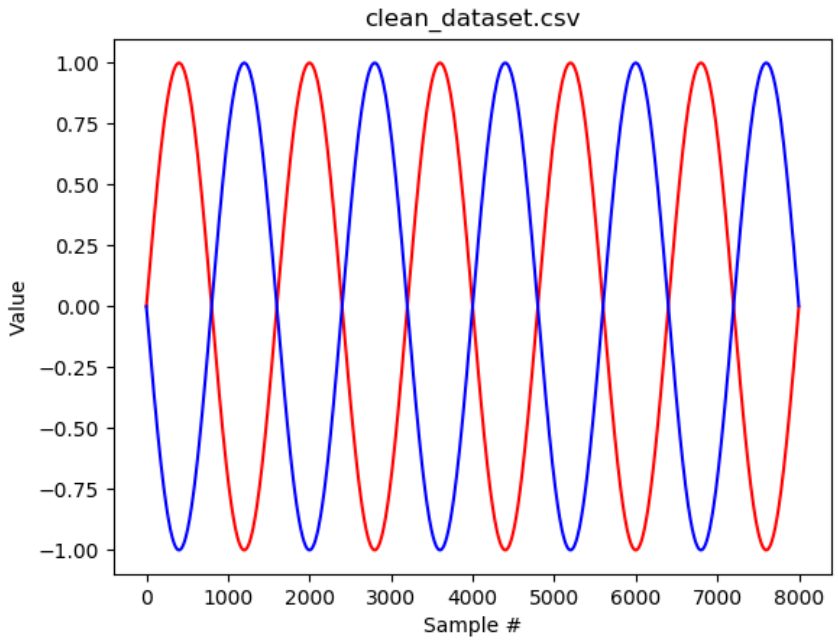

Figure 2: This is a plot of "clean_dataset.csv", which is the dataset produced once the anomalies were removed. (Assuming that all anomalies were intended to be removed.)

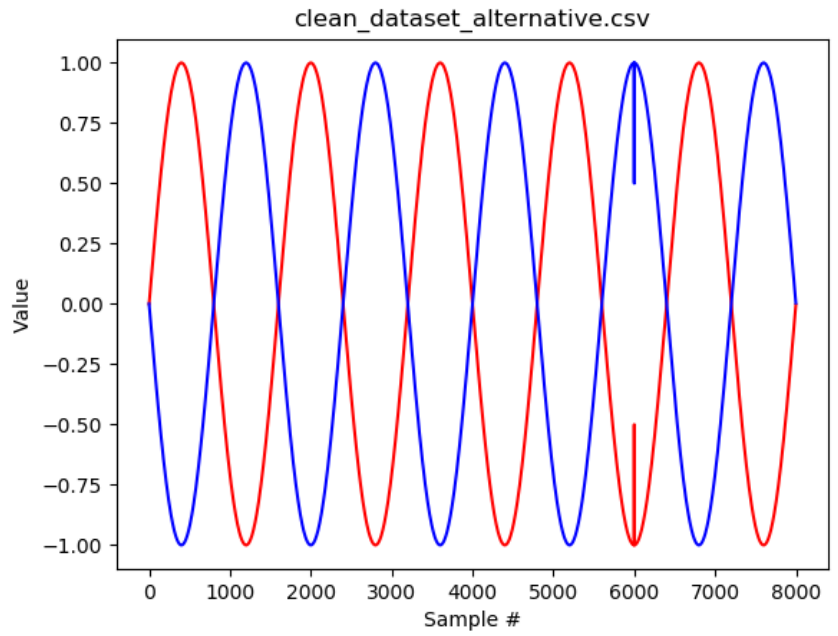

Figure 3: This is a plot of "clean_dataset_alternative.csv", which is the alternative to the "clean_dataset.csv" shown in figure 2. (Assuming that the anomaly at sample number 6000 was a detection.)

All of the anomalies in the dataset were detected by checking each sample for missing values and values outside of an expected variance. Each value found outside of the expected variance was then deleted. Finally, all samples with missing values were filled in using linear interpolation. The dataset was cleaned through two different methods, both of which can be found within this ZIP file. If the code in this ZIP file is manually executed for the sake of verifying reproducibility, it will produce a cleaned dataset identical to the one plotted in figure 2.

References

1. "Anomaly | Definition of Anomaly in English by Oxford Dictionaries," *Oxford Dictionaries*, 2019. [Online]. Available: https://en.oxforddictionaries.com/definition/anomaly. [Accessed January 30, 2019].
2. Seltman, Howard. "Experimental Design and Analysis." *Carnegie Mellon University*, July 11, 2018. [Online]. Available: https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf. [Accessed February 11, 2019].