Having completed the Data Analytics and Visualization Bootcamp, I have acquired a number of new skills, and an awareness of the value and opportunity data can provide to businesses and organizations. Prior to attending this course, I would not have described myself as a programmer, but am now confident in planning, implementing and testing code in a number of languages – with an emphasis on Python. The importance of preparing data for modeling and visualizations was well-established, and I understand the complexity and time-commitment required to do this well.

This course was primarily based on self-study, with weekly information sessions to review each module. There was little time for in-depth discussions, or for the instructional team to spend time on 'how' and 'when' the topic of the week is applied.

For my final project, I was not part of a team, and chose to work alone.

I chose to contact a wastewater treatment facility in Ontario, who had an interesting challenge, and who was willing to share data for my final project. The challenge involved detecting the level of total phosphorus (TP) in treated wastewater before it was released back into the environment. The objective was to determine when TP would be greater than 0.35mg/L. I was provided data in 24 Excel spreadsheets, and started by documenting and understanding how the data was organized and identified data inter-dependencies. I developed an ETL process, which required a number of steps to 'cleanse' and convert the data in preparation for modeling and visualization.

In the effort to detect the objective (TP > 0.35mg/L), I prepared a wide variety of regression and classification models (15 in total), and unsupervised models. In addition, I evaluated each model on 6 of the 10 'trains' as defined by the plant topology (a train is an independent treatment path through the plant). To effectively experiment with model settings and options, I created functions which allowed me to automate the execution of several models with a single invocation.

My final visualization presented an overview of the project and process, as well as pages with charts and metrics to review the performance of each model/train combination, as well as observations in the form of a carousel.

None of the models were successful in achieving what I hoped for. However, I was able to identify possible reasons why the models failed, and believe with additional data, the is a reasonable chance of success in the future. I will be meeting with the plant supervisor who provided the data and hope to continue building on this foundation to achieve usable results.