1. **Understanding the Problem:**

    - How would you design a model to predict the SVI using the dataset?

My first step to design a model for SVI prediction was to clean the data. Specifically, I formatted metrics to the correct data types, investigated missing dates and intervals, identified outliers, dropped unnecessary metrics and those with a high proportion of missing values and/or very large gaps in data, investigated duplicate rows, and imputed missing data. Second, I performed exploratory data analysis (EDA). Here, I examined data distribution of metrics, identified temporal trends and patterns, and investigated the relationship between metrics. I created multiple lagged versions of metrics where the values were shifted back in time in order to be used for predicting non-lagged SVI values. Lastly, based on information gathered during EDA, I developed a random forest model using these lagged metrics to predict SVI. I prioritize clarity over complexity for this first version of the model. Ideas for increasing performance of future versions are noted below.

2. **Data Understanding and Preparation:**

    - What data types and sources are essential?

SVI and other controllable metrics were essential for this model development. Non-controllable metrics can have an impact on SVI, but the operator cannot control them and their inclusion in a model is of less value to an operator. However, these metrics can easily be included in future versions of the model to possibly increase performance.

    - Outline your data collection strategy.

Data was provided for this project.

3. **Exploratory Data Analysis (EDA):**

    - Which EDA techniques would you employ?

I examined data distribution of metrics through the development of histograms and testing for normality. Through this process I discovered that all metrics are not normally distributed.

Therefore, I plotted median, rather than mean, monthly metric values to identify temporal trends and patterns. A decreasing trend in ‘WAS flow’ and increase in ‘Primary Sludge Flow’ from mid 2018 to 2023 was identified. In addition, many metrics had relatively large values in 2017 and 2018.

Lastly, I investigated the relationship between metrics by creating a correlation heatmap and scatter plots. I included the lagged data for these assessments as their relationships between lag times may differ with SVI. The heatmap showed little correlation between SVI and other metrics other than strong positive autocorrelation. In addition, ‘Primary sludge flow A’ is positively correlated with year and ‘Avg ND WAS Flow’ is negatively correlated with ‘year’. The scatter plots don’t indicate strong relationships between SVI or lagged SVIs with other metrics.

While checking for stationarity and seasonal decomposition of metrics before creating a random forest model is not necessary, doing so can be informative for metrics engineering and model

design. Performing these steps can be incorporated into future versions of the model to increase performance.

4. Metrics Engineering:

- What additional metrics might be beneficial?

I created lagged versions of metrics at one, two, and three weeks, since SVI tends to change over the course of weeks. Trying different lag values or determining lag values based on data analysis could be performed for future versions of the model in order to determine the time period that is most predictive. Creating rolling statistics could also be explored.

For future versions of the model, I would also discuss the metrics with an expert in wastewater treatment. Missing values in existing metrics may possibly be calculated from other metrics or additional metrics may be calculated from existing metrics.

The dataset does not include units unfortunately. Confirming uniformity of units for a metric is important for data integrity. Inconsistent units and associated values can be converted in python to make them uniform through the dataset. Units combined with knowledge of the metric are also helpful when determining outliers.

- Strategies for handling missing data and outliers?

I handled missing values by performing linear interpolation in both directions. This method was more appropriate than others, such as imputing mean or median, as this is time series data and values are related to each other. Experimenting with different strategies for handling NaN values can be performed for future versions of the model.

The plotted monthly metric medians were also used to identify potential outliers. Very high values for ‘Primary Sludge B Flow’ and ‘Ras Flow’ were identified but not yet removed as I would want to discuss the metrics with an expert in wastewater treatment.

Related to missing data, I noted that ‘SVI’, ‘MLSS’, and ‘RAS TSS’ were collected on weekends less frequently than weekdays. In addition, ‘SE NH3’ was collected on only Sun-Thur. For future versions of the model, I would investigate if this data had been shifted up one row in the database at some point and should be corrected.

5. Model Development:

- Which machine learning algorithms are suitable, and why?

A model capable of making predictions from time series data was essential for this project. I chose to use the random forest model because it meets these requirements, can adapt to non-linear patterns, resistant to overfitting, can handle seasonal trends, and is simpler than other options. In addition, I used TimeSeriesSplit as a cross-validator because it is specifically designed for time series data. The dataset is split into consecutive folds, where each fold is a superset of the previous one, ensuring that the model is trained on data that retains the temporal nature of time series data. In addition, these methods can also check for autocorrelation. For my model, the lack of correlations identified in the heatmap and scatterplots

indicates that no variables are substantially more correlated with SVI than other metrics. Therefore, all metrics were retained for the model.

The model was run with the one week, two week, and three week lagged data.

- Detail your model optimization steps.

Optimization was not thoroughly explored for this first version of the model. Performance can be increased in future versions by experimenting with different hyperparameter values for the RandomForestRegressor (e.g., n_estimators, max_depth, min_samples_split) and by adjusting the number of splits in TimeSeriesSplit. Trying different lag values or determining lag values based on data analysis could be performed for future versions of the model in order to determine the time periods that are most predictive.

6. Model Evaluation:

- How will you measure your model's success?

Model evaluation was performed by calculating the mean squared error (MSE) of each fold, plotting actual versus predicted values, and calculating the mean squared error for each lagged dataset.

- What performance metrics are pivotal?

MSE for each lagged dataset is the most telling metric of model success. The MSEs from the three lagged datasets indicate that the model is better at predicting the SVI one week into the future than two or three weeks.

7. Challenges and Limitations:

- Identify potential hurdles in this prediction endeavor.

As suggested many times above, more wastewater treatment knowledge would be beneficial in creating this model. Information about metrics and their importance could be very informative to the model. In addition, knowledge about the specific treatment plant would also be informative. For instance, collection of some metrics stopped halfway through the dataset while others began halfway. Understanding the reasons for this could help determine what action to take with these metrics which were dropped for this version of the model.