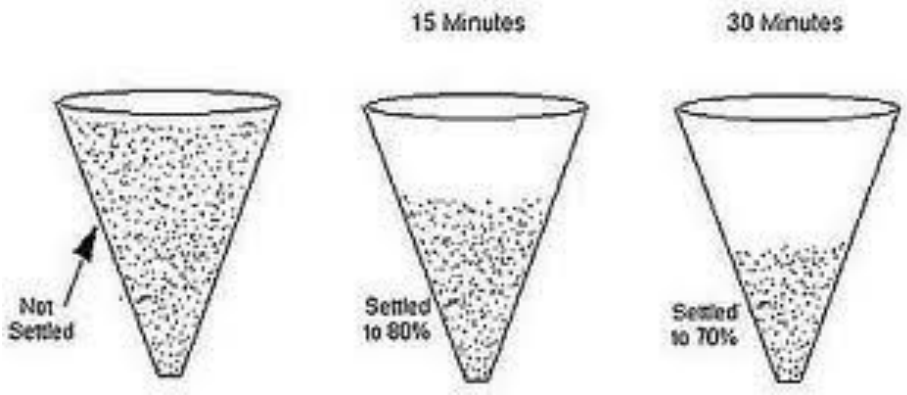
Predicting Sludge Volume Index at a Wastewater Treatment Plant

Nathan Jahns, M.Sc.

Data Science Intensive Capstone Project, Sept 23 Cohort

Background: Sludge Volume Index

- Sludge is removed from wastewater before discharge at wastewater treatment plants (WTPs)
- Sludge Volume Index (SVI) quantifies the separability of solids and liquids
  - Low values = optimal settling
  - High values = issues
    - Filamentous bacteria which impairs separation



- SVI can change in response to waste quality and operations
- Effective management ensures smooth WTP operation, conserves biomass, produces clear effluent, and meets regulatory compliance

Objective

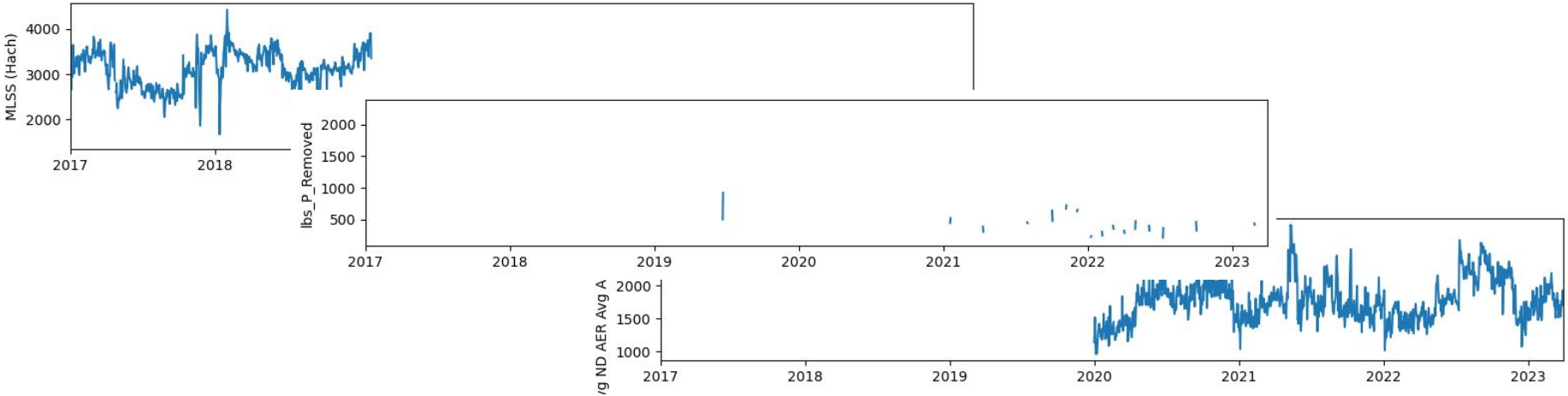
- Develop machine learning model capable of forecasting SVI using explanatory variables that can be modified by a WTP operator
- Dataset comprising daily wastewater measurements spanning five years at a fictitious WTP
- Provide valuable insights to achieve more efficient sludge settling

Data Wrangling

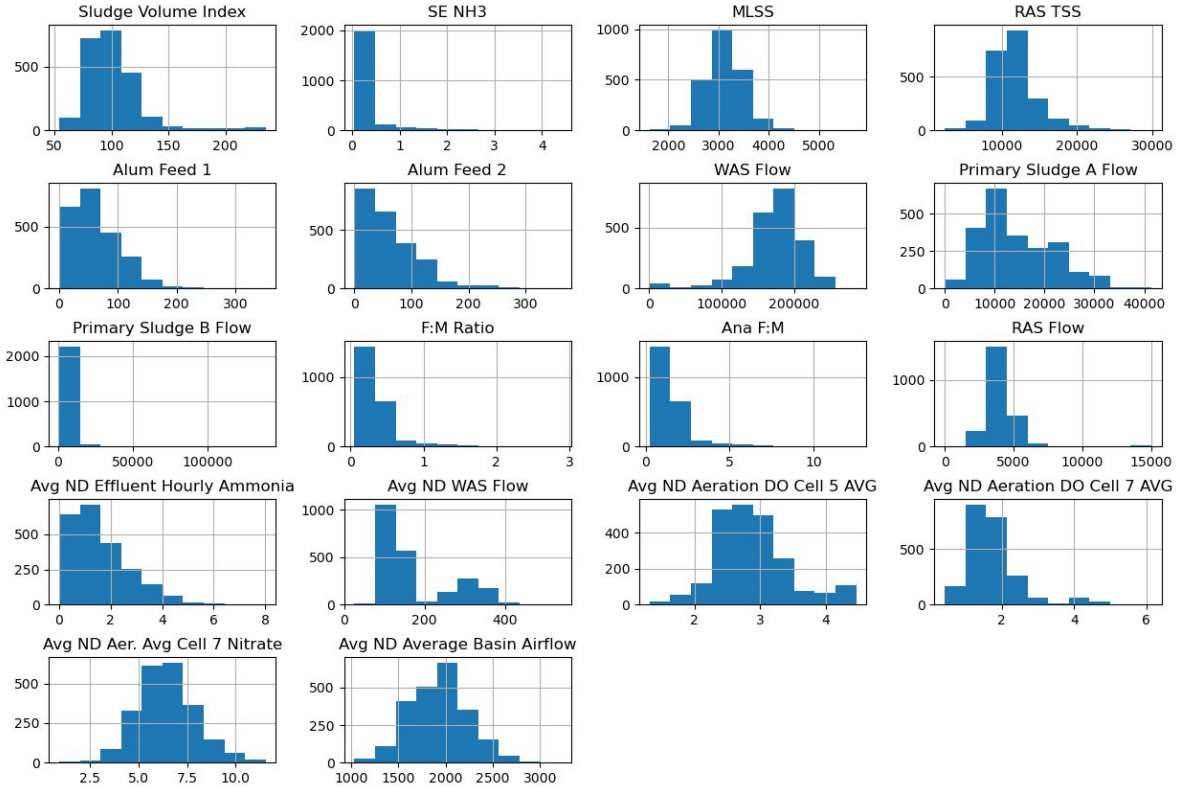
- Data provided by WTP manager, 2,282 rows and 76 columns
- Formatted data to floats
- Confirmed one day frequency
- No units - can’t confirming consistency

Data Cleaning

- Removed metrics with high percentage of NaNs and/or large temporal gaps
- Imputed missing values
  - Linear interpolation - calculates intermediate values between two known data points on a straight line
  - Both forward and backward directions from each NaN value

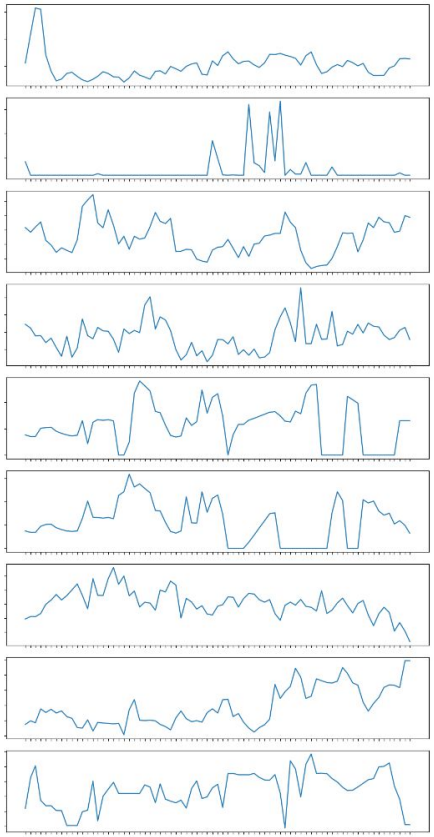


Exploratory Data Analysis: Distribution

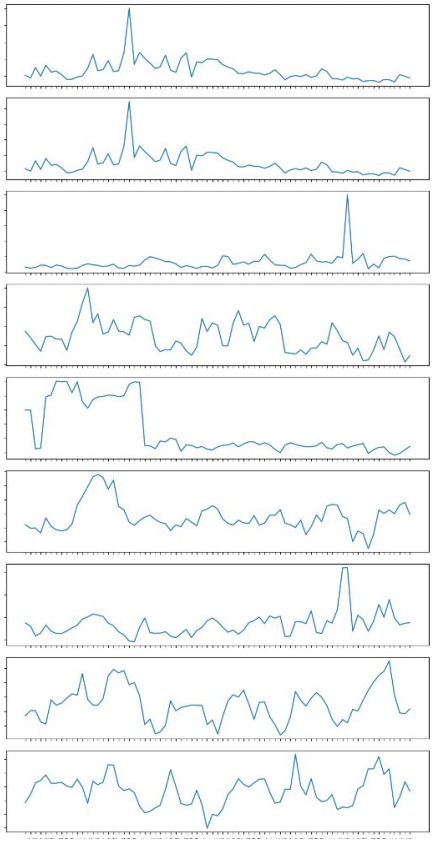


- Examining distribution of metrics
  - Min, max, mean
  - Plotted histograms
  - Shapiro-Wilk tests
- Not normally distributed

Exploratory Data Analysis: Visualization

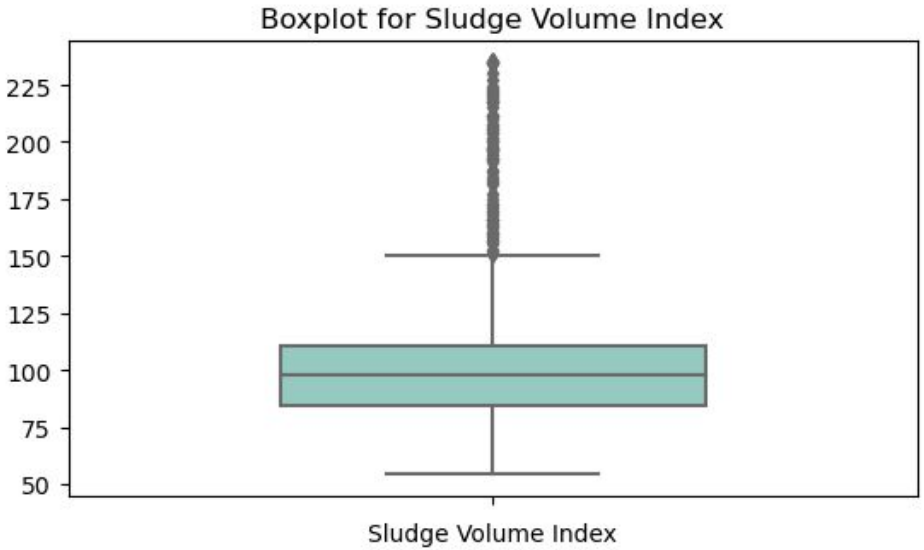


- Plotted line plots of median monthly metrics
- Visually assessed for trends (directional, seasonal) and stationarity
- ‘ANA F:M’ and ‘F:M Ratio’ are the same
- Decreasing in ‘WAS flow’ and increase in ‘Primary Sludge A Flow’ from mid 2018 to 2023
- Many metrics relatively large in 2017/2018
- Metrics appeared to be stationary



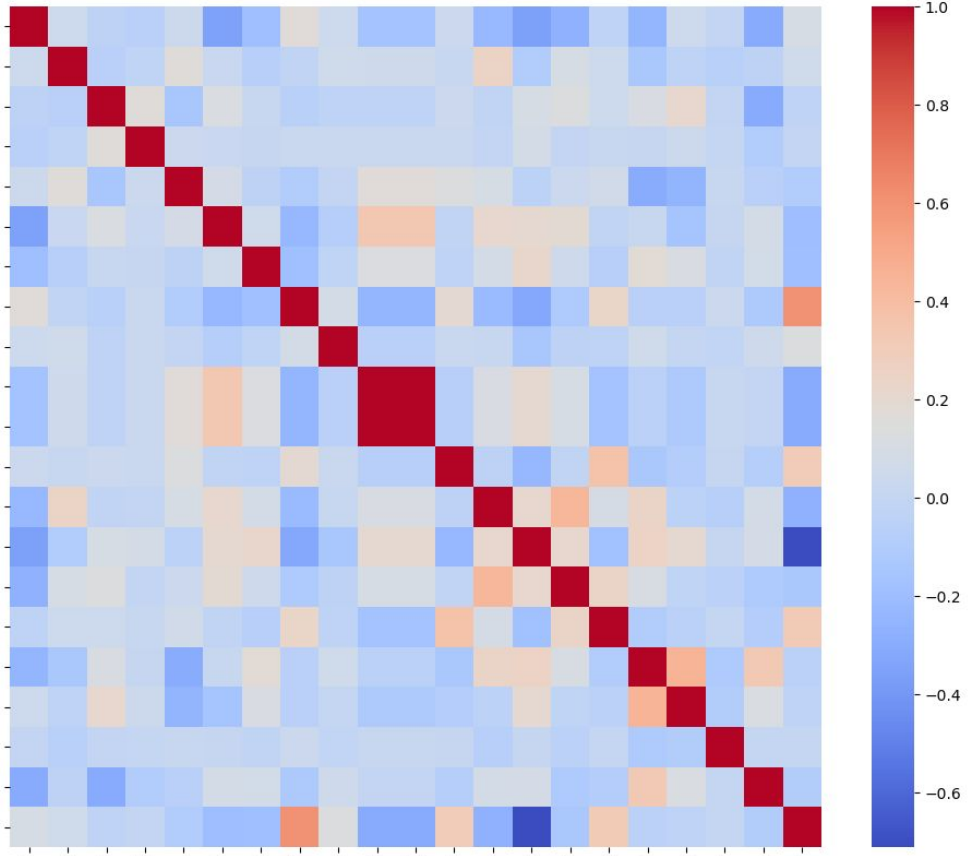
Exploratory Data Analysis: Visualization

- Created boxplots
- Some metrics had extreme values but most were not greater than three standard deviations from the median
- No reason to believe eronious
- Did not remove any values as outliers



Exploratory Data Analysis: Correlations

- Correlation heatmap
- Weak correlation between SVI and other metrics
- ‘ANA F:M’ and ‘F:M Ratio’ are perfectly correlated
    - Removed ‘F:M Ratio’ from the dataset
- Few other moderate correlations
- Pairplot agreed



Feature Engineering

- Used existing data to calculate additional WTP metrics
- Performed differencing
  - Makes stationary and remove seasonality trends
- Created lagged versions of data
  - Incorporates temporal relationships

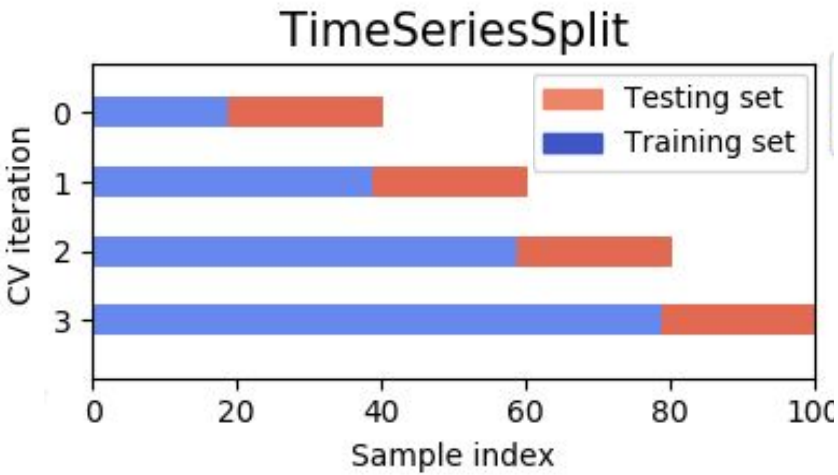
New Metric	**Calculation**
Total Alum Feed	Alum Feed 1 + Alum Feed 2
Total Primary Sludge Flow	Primary Sludge A Flow + Primary Sludge B Flow
Ave ND Aeration DO in Cells	(Avg ND Aeration DO Cell 5 AVG + Avg ND Aeration DO Cell 7 AVG) / 2
MVLSS	MLSS x 0.75
MCRT	MVLSS / (Activated Sludge Flow + RAS Flow)
Sludge Age	MCRT / (1 + F:M Ratio)

Pre-processing

- Defined multiple versions of predictor variables, X
  - Try different combinations of independent variables
  - Reduced size of dataset to prevented the curse of dimensionality and overfitting
  - Original data, differenced, and lagged data containing and not containing SVI
- Defined multiple versions of target variables, y
  - Original SVI
  - SVI shifted forward one day
    - Avoids potential data leakage

Modeling

- ARIMAX, linear regression, and random forest models
  - Cross-validation with TimeSeriesSplit
    - Specific for time series and avoids overfitting
  - Squared error (RMSE) and adjusted R squared (R2adj)
  - ARAMEX
    - Included exogenous features
    - Auto ARIMA to choose the most appropriate p, d, and q hyperparameters



- Run with combinations of my different X and y versions
  - ARIMAX model used the Xs that did not include differenced and lagged SVI

Results

ARIMAX

- Perform poorly with all data

Linear Regression

- Perform moderately poor with shifted y and a lag of 1 day data
  - RMSE 8.76
  - R2adj 0.16
- Lag of 4 days was slightly worse followed by lag of 7 days
  - Ran models with lag of 2 days and lag of 3 days
  - Not as good as 1 day

Results: Random Forest

- Perform very well with shifted y and a lag of 1 day data
- Error between the model's predictions and the actual values was very small
  - RMSE 0.54
- The model explains almost 97% of the variance in the target variable
  - R2adj 0.97