sustainability

ANEXO 4

MDPI

Article

Big Data Analytics from a Wastewater Treatment Plant

Praewa Wongburi [1,*] and Jae K. Park [2]

[1] Faculty of Environment and Resource Studies, Mahidol University, Nakhon Pathom 73170, Thailand
[2] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA; jkpark@wisc.edu
* Correspondence: praewa.won@mahidol.ac.th

Abstract: Wastewater treatment plants (WWTPs) use considerable workforces and resources to meet the regulatory limits without mistakes. The advancement of information technology allowed for collecting large amounts of data from various sources using sophisticated sensors. Due to the lack of specialized tools and knowledge, operators and engineers cannot effectively extract meaningful and valuable information from large datasets. Unfortunately, the data are often stored digitally and then underutilized. Various data analytics techniques have been developed in the past few years. The methods are efficient for analyzing vast datasets. However, there is no wholly developed study in applying these techniques to assist wastewater treatment operation. Data analytics processes can immensely transform a large dataset into informative knowledge, such as hidden information, operational problems, or even a predictive model. The use of big data analytics will allow operators to have a much clear understanding of the operational status while saving the operation and maintenance costs and reducing the human resources required. Ultimately, the method can be applied to enhance the operational performance of the wastewater treatment infrastructure.

Keywords: big data; wastewater treatment plant; data analytics; statistical analysis

Citation: Wongburi, P.; Park, J.K. Big Data Analytics from a Wastewater Treatment Plant. *Sustainability* **2021**, *13*, 12383. https://doi.org/10.3390/su132212383

Academic Editor: Andreas N. Angelakis

Received: 20 September 2021
Accepted: 2 November 2021
Published: 10 November 2021

1. Introduction

Big data plays an essential role in many fields of our daily life. Data generation has been tremendously increasing since 2010. Of the world's data, 90% were created in the past two years [1]. Water quality sensors in various wastewater treatment operational processes have generated a large amount of data. However, operators cannot use the digital data collected from multiple sensors to identify the plant operation status [2]. Accordingly, most data collected are wasted. Wastewater treatment processes involve many operational systems. The objective of wastewater treatment is to remove contaminants from sewage to discharge into natural resources. Treated water must meet effluent discharge permits to protect public health and the environment [3]. A lack of access to safe water is a risk factor for infectious diseases such as cholera, diarrhea, dysentery, etc. In addition, 1.2 million people died prematurely in 2017 due to unsafe water [1].

Most wastewater treatment plants apply the SCADA system, a distributed computer system to aid wastewater treatment processes through monitoring and automation control. Wastewater treatment operators are constantly facing the traditional system challenge for controlling the plant including these three main areas [4]:

- Obtainability and reliability: Including old plant infrastructure, stability of the operation system, and reliability of the information coming into the system;
- Risk: Compliance concerns, plant security, reporting and errors of the systems, and experienced operators retiring;
- Cost: Operation and maintenance cost, chemicals cost, training new operators, and energy consumption costs.

WWTPs can address challenges in these three important areas in several ways. A modern big data analytics system supplies information anytime over a disaster recovery architecture for high availability and reliability [4]. Risk can be reduced by reliable data management, effective data analysis, consistent monitoring processes, and real-time prediction systems to detect a potential risk beforehand. The biggest challenge in data analytics from wastewater treatment is the dynamic behavior of the data [5]. The data are usually complicated and uncertain because of variations from the environmental conditions, changes in the process variables, and fluctuations in the flow rate and concentration of the influent composition [6]. Finding insight from historical and real-time data can allow for the better management of WWTPs and advanced operational decision support systems. In addition, big data from WWTPs were analyzed using statistical tools to visualize operating conditions. Efficient operations can reduce cost by thoroughly monitoring the plant, applying big data analysis, and finally, optimizing wastewater treatment operations.

2. Materials and Methods

The methodology of big data analytics for wastewater treatment operations is described below and shown in Figure 1.

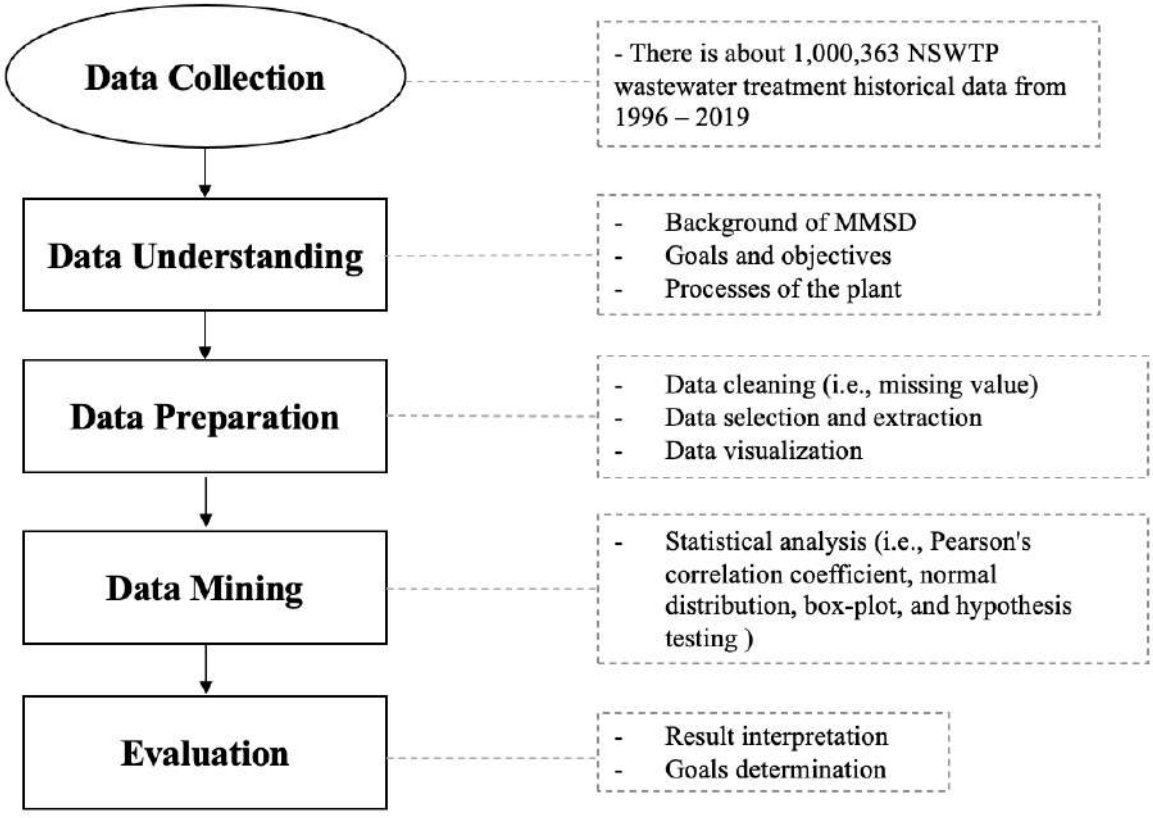


Figure 1. Flowchart of big data analytics in WWTPs.

2.1. Data Collection

Data were collected from the Nine Springs WWTP operated by the Madison Metropolitan Sewerage District (MMSD), Madison, WI, USA. The wastewater treatment plant is separated into four plants, and each plant involves the preliminary treatment, primary clarification, nitrifying activated sludge treatment incorporating biological phosphorus removal, ultraviolet disinfection, and effluent pumping [7].

The dataset contains 1,000,363 wastewater treatment historical data from 1996 to 2019. The dataset consists of a tremendous amount of data with a file size of about 311.6 MB. The data contain 13 columns, such as 'ResultDate', 'Result', 'MeasureCode', 'LocationCode',

etc. There are many parameters from many treatment processes, as shown in Figure 2a, which are from 'MeasureCode' column. The data were collected from several points in the wastewater treatment plant, which is shown in Figure 2b.

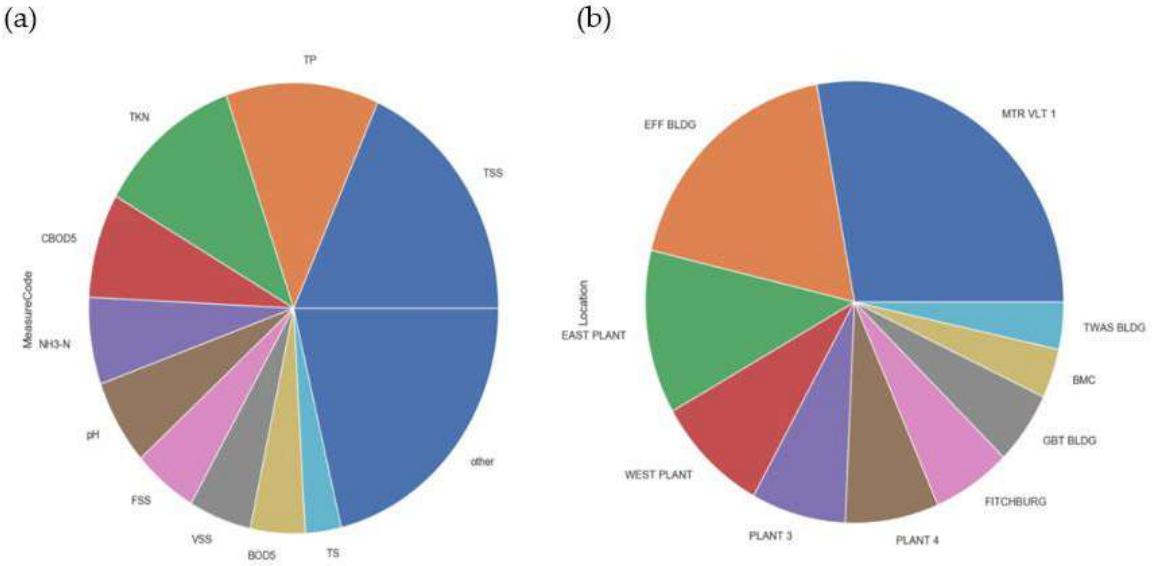


Figure 2. Pie charts of the two columns in the dataset: (**a**) The parameters in the 'MeasureCode' column, and (**b**) The collection points in the 'LocationCode' column.

2.2. Data Understanding

The collected data need to be studied and understood. The Madison Metropolitan Sewerage District 50-Year Master Plan was reviewed to research the background, goals, and WWTP processes.

2.2.1. Background and Goals

Madison Metropolitan Sewerage District (MMSD) is a municipal corporation created to collect and treat wastewater from the Madison metropolitan area. MMSD provides service to 43 municipal customers. The service area covers 177 square miles (458 km^2) and serves a current population of nearly 330,000 people. MMSD owns and operates the Nine Springs WWTP [7]. The Nine Springs WWTP averagely treats 41 million gallons of wastewater per day (155,000 m^3/day). The main objective of the plant is to provide exceptional service at a reasonable cost to customers while considering an appropriate balance between environmental, social, and economic impacts. This study analyzes data generated in the Nine Springs WWTP and finds insights and patterns to optimize WWTP operation.

2.2.2. The Liquid Treatment Processes

The liquid treatment processes at the Nine Springs WWTP include preliminary treatment, primary clarification, nitrifying activated sludge treatment incorporating biological phosphorus removal, ultraviolet disinfection, excess flow storage, and effluent pumping. Wastewater enters the plant through the influent meter vault, in which influent data were chosen, and the treated water is sent to the effluent building where that effluent data was selected.

2.3. Data Preparation

In this stage, the influent and effluent data were selected, and important parameters were determined. Data cleaning, visualizing, transforming, and obtaining feature selection and extraction are part of this stage. The data are now available in a form that is compatible with a modeling technique, which will be introduced in the next study.

Data Preprocessing

Data preprocessing is how the data are transformed or encoded to a state that a computer can easily parse. Data preprocessing helps the computer to understand data. Han et al. (2012) summarized the following steps involved in data preprocessing [8]:

1. Data cleaning

Data cleaning includes many tasks such as filling in missing values, smoothing noisy data, identifying or removing outliers, and correcting inconsistencies [9]. Unclean data can cause uncertainty for the mining process, resulting in inaccurate output. Thus, the data cleaning routine is one of the most important techniques of data preprocessing.

2. Data integration

The integration of multiple databases or data integration is often required in data mining processes. Data integration is the combining of data from various data stores. Thorough integration helps to reduce and avoid redundancies and inconsistencies in the dataset. Data integration can help increase the precision and speed of the mining process. The challenge of data integration is how we can match schema and objects from different sources. It is the essence of the entity identification problem. The techniques involve correlation tests, duplication recognition, and detection of data value conflicts.

3. Data transformation

The data are transformed or consolidated, so the result after the analytics process will be more efficient, and the patterns may be simpler and easier to understand. Strategies for data transformation include smoothing, aggregation, normalization, feature construction, and so on.

4. Data reduction

Data reduction reduces the size of the dataset that is much smaller in volume and still carefully maintains the integrity of the original data. Therefore, the valid data reduction will produce the same or almost the same analytical consequences.

2.4. Data Mining

Various statistical methods were employed to extract knowledge from the preprocessed data. Visualization and statistical analysis, such as Pearson's correlation coefficient, normal distribution, boxplot, and hypothesis testing, were implemented. Data pattern identification, statistical analytics, and normalization are performed in this step.

2.4.1. Correlation Coefficient for Numeric Data

For numeric attributes, the correlation between two attributes, A and B, can be evaluated by computing the correlation coefficient as in Equation (1) shown below [8]:

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{AB}}{n\sigma_A\sigma_B} \quad (1)$$

where n is the number of tuples, a_i and b_i are the respective values of A and B in tuple I, $\overline{A}$ and $\overline{B}$ are the respective mean values of A and B, σ_A and σ_B are the respective standard deviations of A and B, and $\Sigma(a_i b_i)$ is the sum of the $A.B.$ cross-product. Note that $-1 \leq r_{A,B} \leq +1$.

If $r_{A,B}$ is greater than 0, then A and B are positively correlated, meaning that the values of A increase as the values of B increase. The higher the value, the stronger the correlation. Therefore, a higher value may indicate that A (or B) may be removed as a redundancy. If the resulting value equals 0, then A and B are independent, and there is no correlation between them. If the resulting value is less than 0, then A and B are negatively correlated, where the values of one attribute increase as the values of the other attribute decrease. This means that each attribute discourages the other.

2.4.2. Normalization

In statistical analysis, D'Agostino's K^2 test measures a goodness-of-fit measure of departure from normality [10]. This test aims to determine whether the data sample comes from a normal distribution. The test is from the transformations of the example kurtosis and skewness and has power against the alternatives that the distribution is skewed or kurtic [11].

In the below equation, x_i denotes a sample of n observations, g_1 and g_2 are the sample skewness and kurtosis, respectively, the $m_j{}^{'s}$ are the jth sample central moments, and $\overline{x}$ is the sample mean.

The sample skewness and kurtosis are defined as follows [11]:

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}} \quad (2)$$

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3 \quad (3)$$

These quantities consistently estimate the theoretical skewness and kurtosis of the distribution, respectively.

2.4.3. Box Plot

In statistical analysis, a box plot or boxplot is an approach for graphically representing groups of numerical data through the quartiles [12]. Box plots can have lines ranging from the boxes or whiskers, which demonstrate the variability of the upper and lower quartiles. Outliers may be plotted as individual points. In other words, box plots show variation in the dataset.

2.5. Evaluation

The results from the previous step were interpreted. The impact of new knowledge was evaluated to determine whether the goals have been met. Hypothesis testing was used for the evaluation part. Hypothesis testing in statistics is a method of testing the results to see if there is a meaning in a dataset. Statisticians have developed a way of drawing inferences from samples or finding through hypothesis testing. It can help interpret data, make decisions, and find errors in results. Hypothesis testing aims to determine the likelihood that a population parameter is likely to be true. Below are the four steps of hypothesis testing [13]:

1. Determine a null hypothesis;
2. State the null hypothesis;
3. Select an appropriate test;
4. Show to either support or reject the null hypothesis.

3. Results

Nine Springs WWTP's historical data were collected from 1996 to 2019 in SQL database. Python Jupyter Notebook, which is an open-source software containing live code, equations, and visualization, was used in this study. The program applications include data cleaning, data visualization, statistical modeling, machine learning, etc. The program was used to analyze, select, preprocess, visualize, and transform the large-size data into the appropriate dataset to develop a prediction model. After processing the first dataset, the influent parameters were selected from 'MTR VLT1', the influent meter vault where the wastewater has entered the plant. The effluent parameters were chosen from 'EFF BLDG,' which is the effluent building where the treated water is sent.

After selecting the locations, the data are separated into influent and effluent data. Table 1 shows the combination of the influent and effluent dataset, which is easier to understand and ready for the following process.

Table 1. Clean dataset from the Nine Springs WWTP big data.

Result Date	Influent_ TSS	Effluent_ TSS	Influent_ TP	Effluent_ TP	Influent_ TKN	Effluent_ TKN	Influent_ NH3N	Effluent_ NH3N	Influent_ BOD5	Effluent_ BOD5
1997-01-02	242	7	6.42	2.21	31.4	1.45	19.9	0.18	196	4
1997-01-02	242	7	6.42	2.21	31.4	1.45	19.9	0.18	196	4
1997-01-02	242	7	6.42	2.21	31.4	1.45	19.9	0.18	196	4
1997-01-02	242	7	6.42	2.21	31.4	1.45	19.9	0.18	196	4
1997-01-02	242	7	6.42	2.21	31.4	1.45	19.9	0.18	196	4
...	...	...	...	...	...	...	...	...	...	...
2019-01-02	320	4.7	5	0.26	41.8	2.08	27.7	0.11	257	6.3
2019-01-02	178	4.7	5	0.26	41.8	2.08	27.7	0.11	257	6.3
2019-01-02	210	4.7	5	0.26	41.8	2.08	27.7	0.11	257	6.3
2019-01-02	197	4.7	5	0.26	41.8	2.08	27.7	0.11	257	6.3
2019-01-02	252	4.7	5	0.26	41.8	2.08	27.7	0.11	257	6.3

3.1. Data Visualization

Data visualization means the graphic representation of data [14]. The relationships between the influent and effluent in each parameter are shown in Figures 3–7. The selected parameters include Total Suspended Solids (TSS), Total Phosphorus (TP), Total Kjeldahl Nitrogen (TKN), Ammonia-Nitrogen (NH_3-N), and the Biochemical Oxygen Demand (BOD_5), which are the essential parameters affecting wastewater treatment quality.

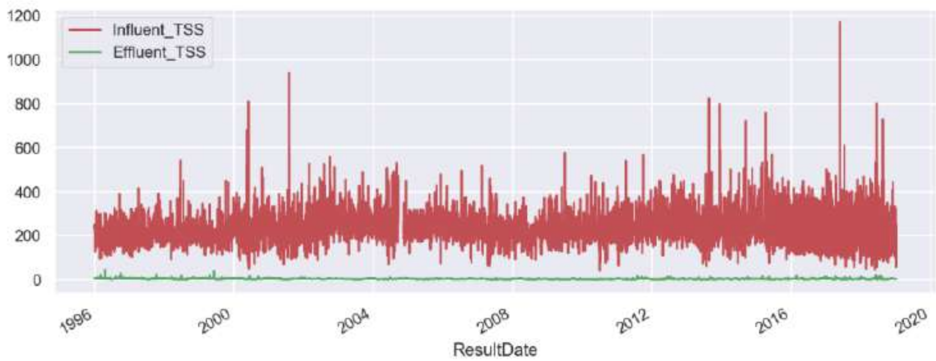


Figure 3. Relationship between influent and effluent in Total Suspended Solids (TSS).

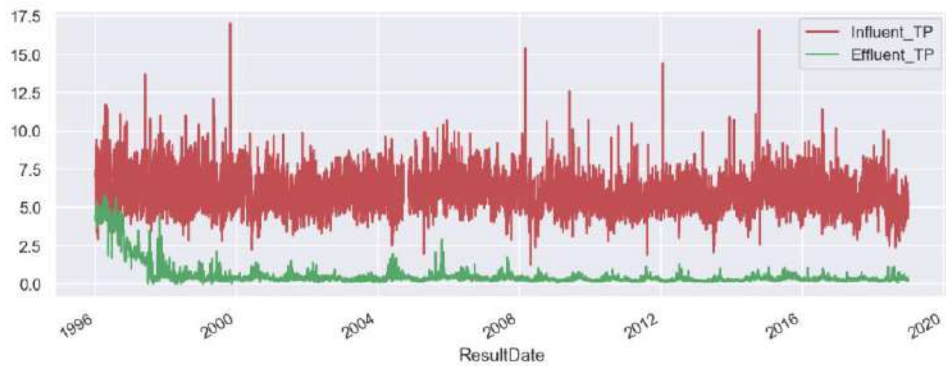


Figure 4. Relationship between influent and effluent in Total Phosphorus (TP).

Figure 5. Relationship between influent and effluent in Total Kjeldahl Nitrogen (TKN).

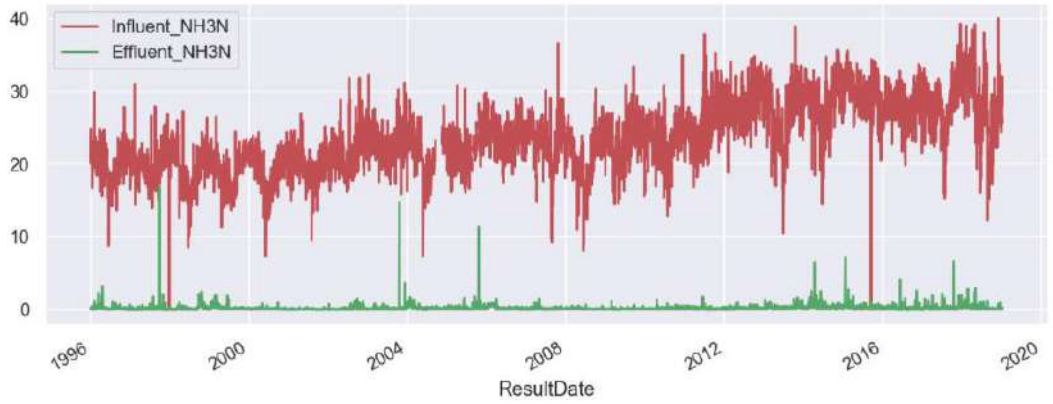


Figure 6. Relationship between influent and effluent in Ammonia-Nitrogen (NH_3N).

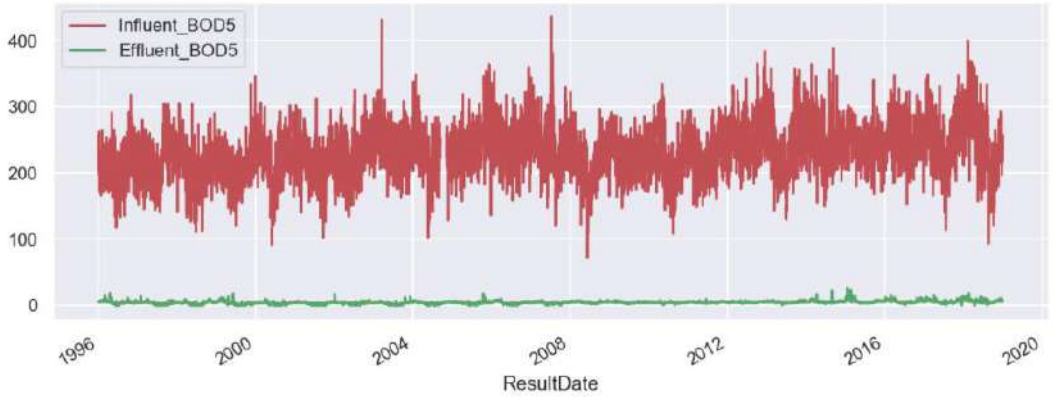


Figure 7. Relationship between influent and effluent in the Biochemical Oxygen Demand (BOD_5).

3.2. Statistical Analysis

In the Python Jupyter notebook, a STAT module can help summarize data using the 'describe a function', describe (). The describe () function is used to generate descriptive statistics that summarize the central tendency and dispersion of the numerical values in the dataset. The values of parameters—mean, standard deviation, percentile, and the interquartile range—of the data are shown in Table 2.

Table 2. Summary of descriptive statistics.

	Count	Mean	Std	Min	25%	50%	75%	Max
Influent_TSS	3,713,349.00	231.84244	52.45347	43.00	200.00	225.00	258.00	1170.00
Effluent_TSS	3,713,349.00	4.81817	1.62729	0.00	3.90	4.60	5.50	46.00
Influent_TP	3,713,349.00	5.95551	1.04525	1.22	5.25	5.91	6.58	17.00
Effluent_TP	3,713,349.00	0.56240	0.85971	0.00	0.23	0.31	0.43	5.78
Influent_TKN	3,713,349.00	35.92247	6.94565	4.43	31.20	35.40	40.10	95.30
Effluent_TKN	3,713,349.00	1.47726	0.41980	0.00	1.25	1.41	1.64	19.30
Influent_NH3N	3,713,349.00	22.50772	4.05514	0.00	19.90	22.10	24.90	40.10
Effluent_NH3N	3,713,349.00	0.16773	0.28697	0.00	0.05	0.08	0.18	16.90
Influent_BOD5	3,713,349.00	230.06485	37.16475	72.00	207.00	230.00	252.00	436.00
Effluent_BOD5	3,713,349.00	3.65499	1.70947	0.00	2.80	3.60	4.40	27.00

3.3. Correlation Coefficient

The relationship between factors can be analyzed by using the correlation coefficient calculation. It can be used to determine the strength of the relationship of each parameter and select the input and output parameters for our developing model in the next chapter.

When the correlation coefficient between the two parameters is greater than 0, they are positively correlated. This means that when one value increases, another value also increases. The higher the correlation coefficient value, the stronger the correlation. Figure 8 shows the heatmap colors of correlation. The lighter color means that they have a stronger relationship. However, if the color turns black, it shows a negative value, indicating that each attribute discourages another. Effluent BOD_5 has all positive values with higher correlations than other parameters. Therefore, the effluent BOD_5 was used as an output parameter for the first model. After that, other parameters will be used as inputs. Figure 9 shows the correlations of effluent BOD_5 to other parameters such as flow rate, pH, temperature, and other effluent parameters.

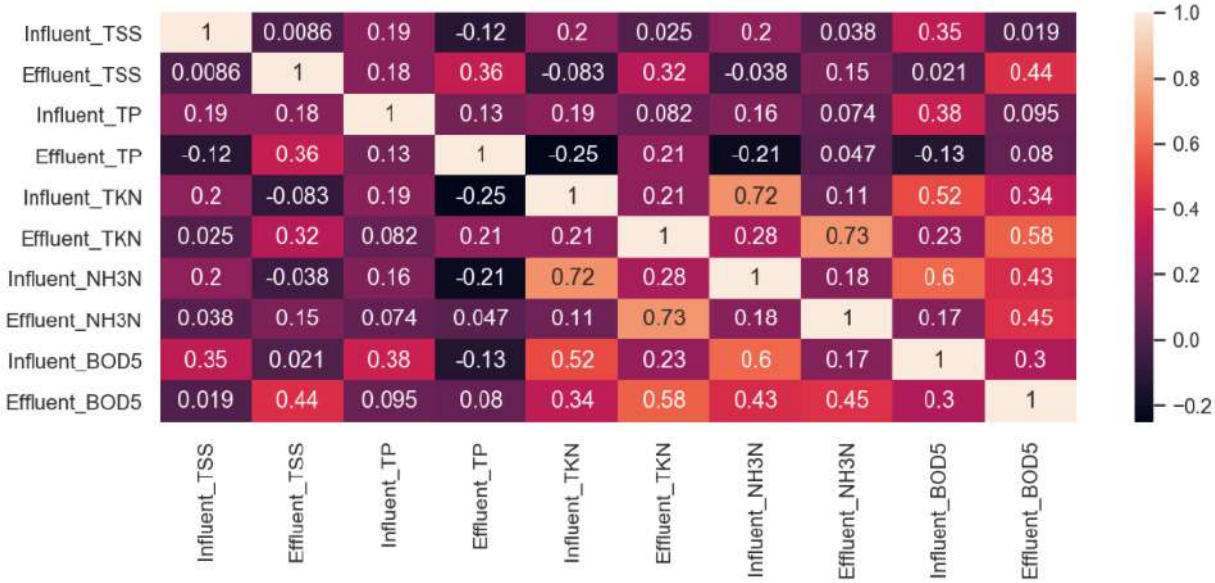


Figure 8. Heatmap of the correlation coefficient.

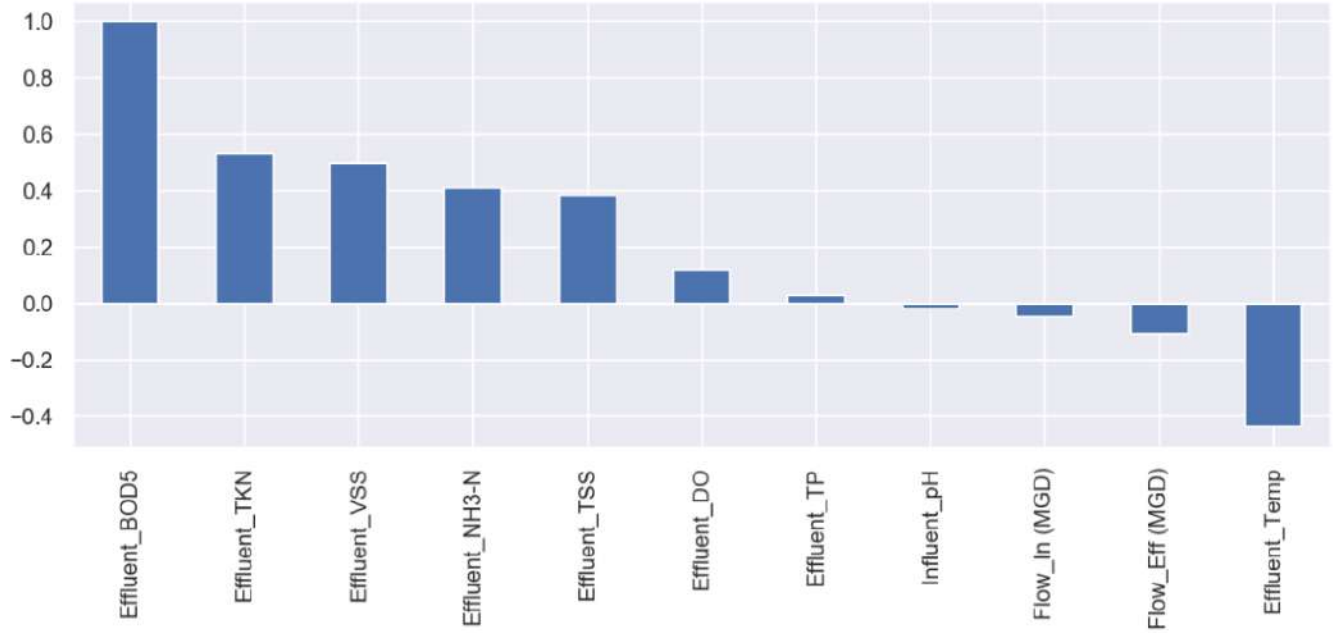


Figure 9. Correlation of effluent BOD_5 to other parameters.

3.4. Removal of Missing Value and Merging of Date and Time Data

The data preprocessing steps are the following [15]: (1) merge date and time into one column and change to DateTime type, (2) convert all data to numeric, (3) remove missing values, and (4) create year, quarter, month, and day features. After removing the missing values, the data contain 3,713,349 measurements collected between January 1996 and January 2019. The initial data include several variables. However, the statistical analysis will focus on a single value: Historical Effluent BOD_5 data.

3.5. Normalization

3.5.1. Statistical Normality Test

Several statistical tests can be applied to quantify whether the data were drawn from a Gaussian distribution. D'Agostino's K^2 statistical test will be implemented in Python Jupyter [16]. The *p*-value is interpreted as follows:

$p \leq$ alpha: reject H0, not normal;
$p >$ alpha: fail to reject H0, normal.

The result of the statistical analysis shows that effluent BOD_5 data reject H_0, which means that the data are not a Gaussian distribution (not normal).

The kurtosis and skewness can also determine if the data distribution departs from the normal distribution [15]. Kurtosis describes the heaviness of the tails of a distribution. If the kurtosis is more than 3, the dataset has heavier tails than a normal distribution. In other words, there are more data in the tails on either side. If kurtosis is less than 3, the dataset has lighter on tails or less in the tails [17]. Figure 10 shows that our kurtosis is 5.205, meaning the heaviness of the tails of a distribution. It means that the dataset has large outliers.

Skewness measures the asymmetry of the distribution. If the skewness is between -0.5 and 0.5, the data are relatively symmetrical. If the skewness is between -1 and -0.5 or between 0.5 and 1, the data are moderately skewed. If the skewness is less than -1 or greater than 1, the data are highly skewed. Figure 10 shows that the skewness of normal distribution is 0.779, implying the data are moderately skewed.

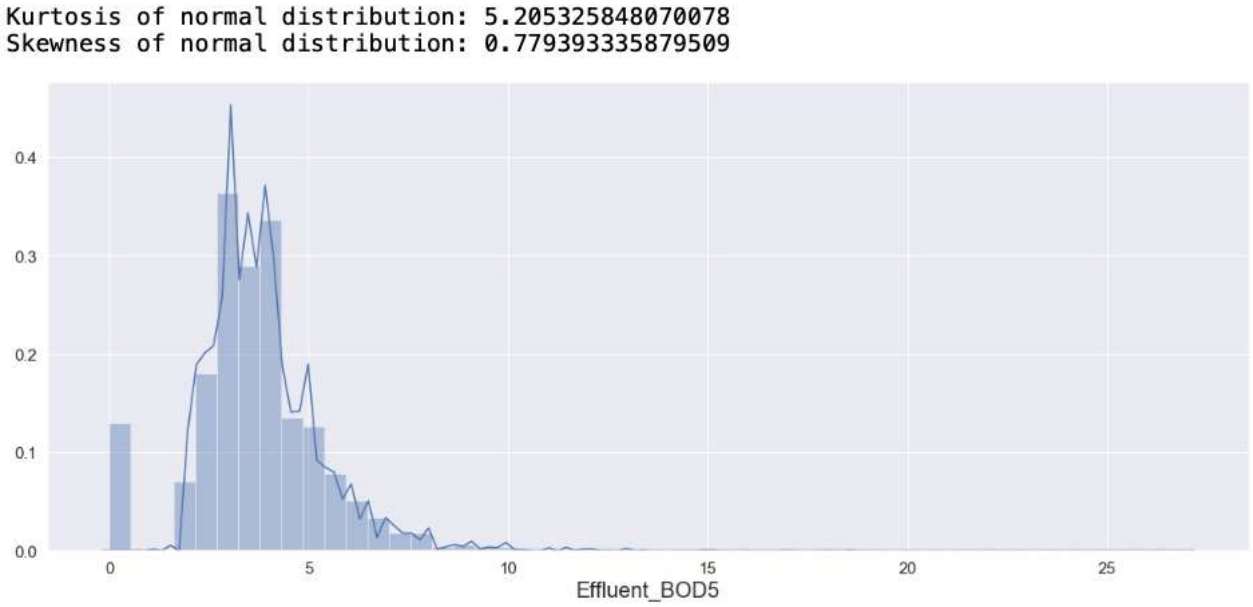


Figure 10. Kurtosis and skewness of the normal distribution.

3.5.2. Box Plot

A box plot is another tool for visualizing data. Figure 11 shows the yearly box plot, noticing that the median effluent BOD_5 in 2014 is higher than the other years. The median effluent BOD_5 values were higher in the first and fourth quarters (winter) and the lowest in the third quarter (summer).

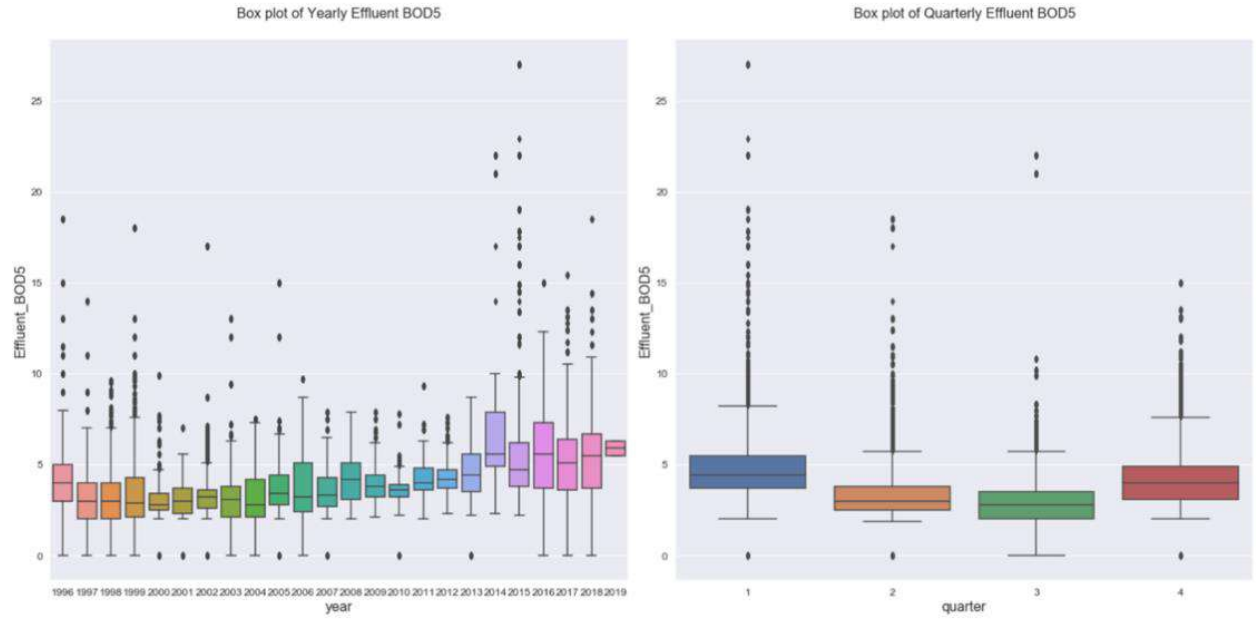


Figure 11. Box plots of yearly and quarterly effluent BOD_5.

3.5.3. Normal Probability Distribution

The normal probability plot, Figure 12, also shows that the effluent BOD_5 data are far from normally distributed.

	count	mean	std	min	25%	50%	75%	max
Effluent_BOD5	3713349.000	3.655	1.709	0.000	2.800	3.600	4.400	27.000

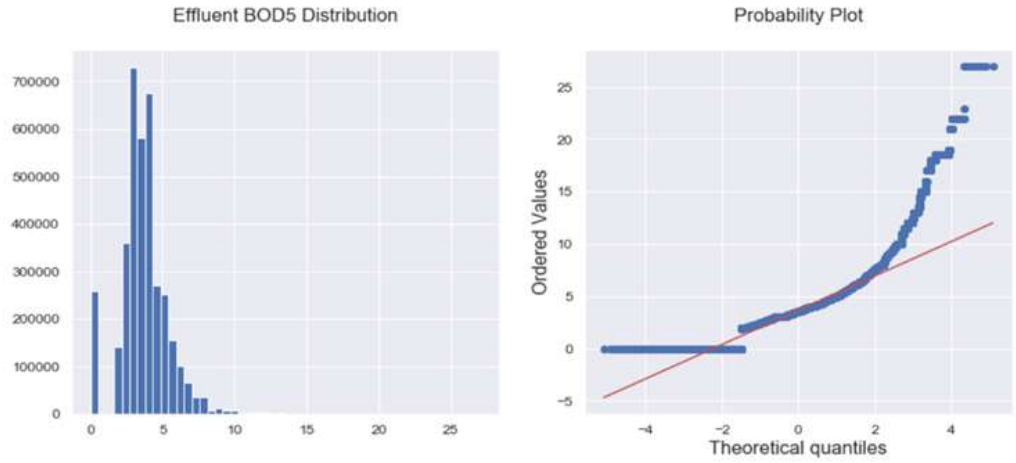


Figure 12. Normal probability distribution.

Figure 13 represents the means of effluent BOD_5 over days, weeks, months, quarters, and years. The highest year of effluent BOD_5 is in 2014, where data are missing in a certain period. Further into the investigation, the plant was modified in 2014. Therefore, the big data analytics results can show when the data are not normal.

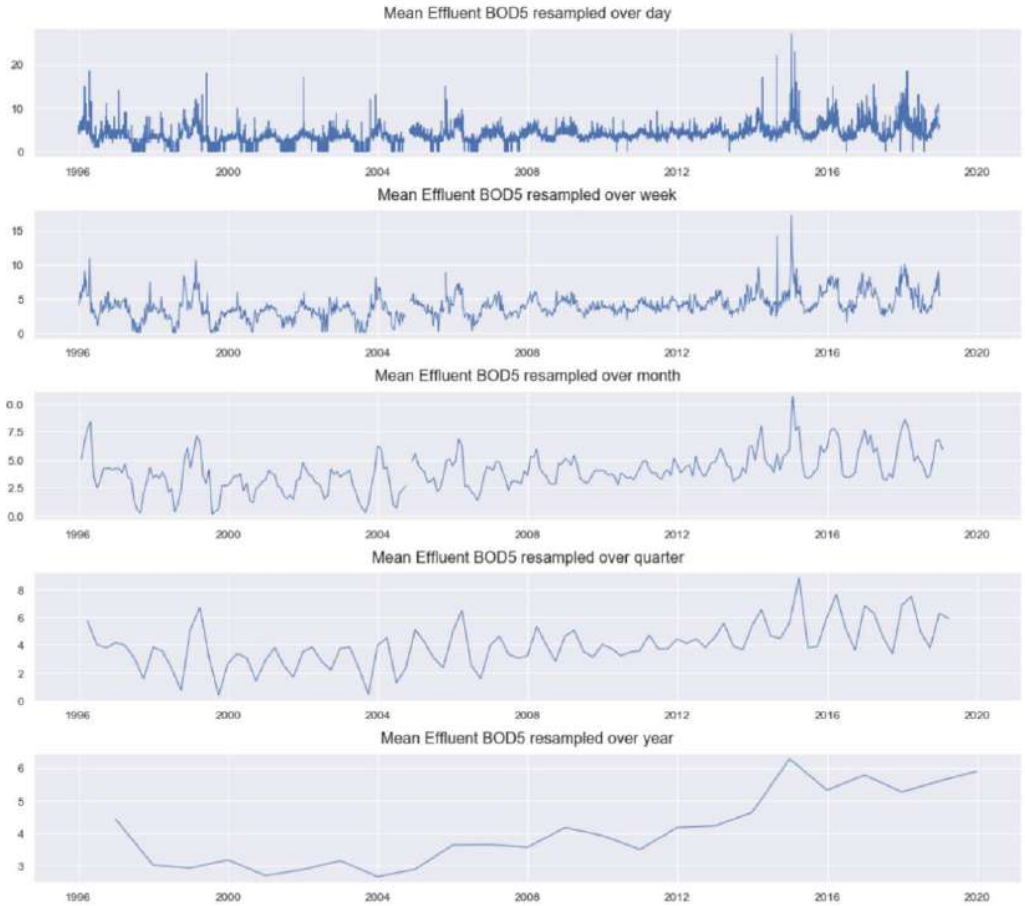


Figure 13. Mean effluent BOD_5 over a day, week, month, quarter, and year.

Figure 14 confirms the previous analysis shown in Figure 12. The highest yearly effluent BOD_5 was in 2014. The lowest quarterly average effluent BOD_5 was in the third quarter. The lowest monthly effluent BOD_5 was in August. The lowest daily average effluent BOD_5 was around the 30th of the month. This analysis can assist the plant in closely monitoring the plant and managing the operations before a system failure happens, such as by visualizing the historical peak load by quarter, month, and day. In addition, big data analysis can reduce the operation and maintenance cost and find hidden information inside the dataset.

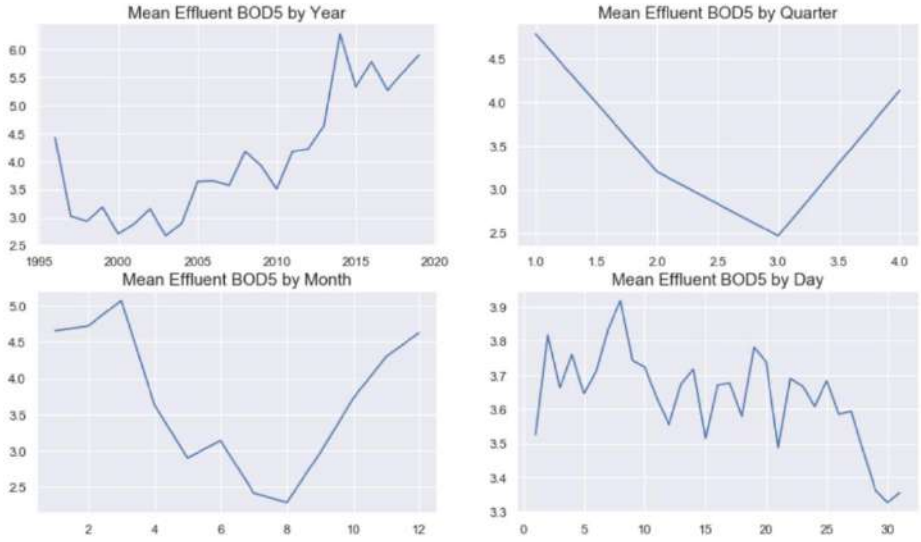


Figure 14. Mean effluent BOD_5 by year, quarter, month, and day.

Figure 15 shows the patterns of the effluent BOD_5 for each year. As a result, the 2004 and 2019 data were removed because the data were missing, and the 2014 data were removed because the values were too high.

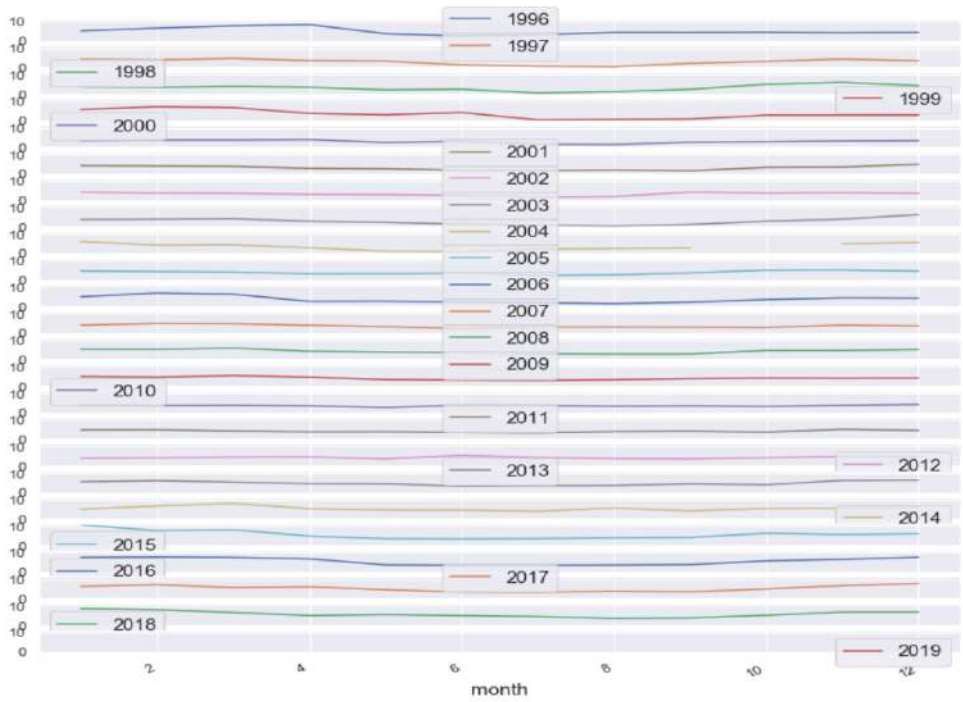


Figure 15. Yearly effluent BOD_5 pattens.

Therefore, the data of 2005–2013 or 2015–2018 were selected for a more accurate prediction model. In the next section, the data of 2015–2018 were evaluated for their accuracy and trend.

4. Discussion

The data of 2015–2018 are further evaluated. In Figure 16, the data are plotted to determine the pattern. The box plot in Figure 17 displays how close the data are in each year and quarter.

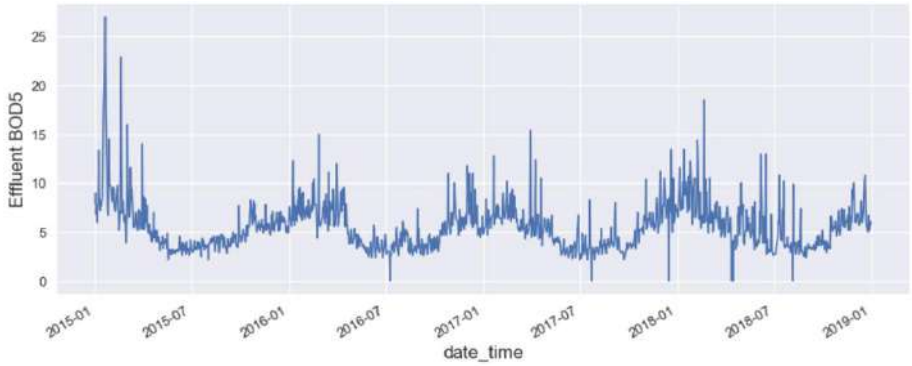


Figure 16. Time series plot of effluent BOD_5 from 2015–2018.

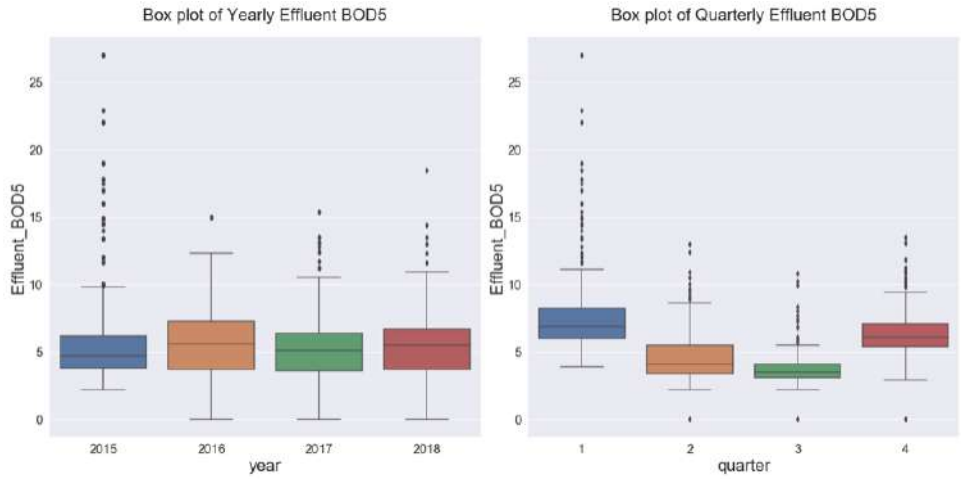
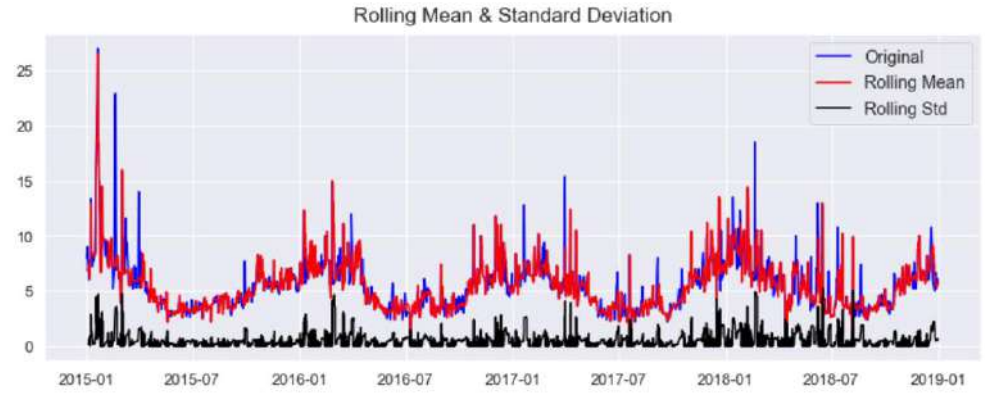


Figure 17. Box plots of effluent BOD_5 by year and quarter from 2015–2018.

Finally, the Dickey-Fuller test with hypothesis testing was performed to determine if the data were stationary. When the data are stationary, it will make a model easier and faster to learn and predict. The null hypothesis is that a unit root is present in an autoregressive model. The alternative hypothesis is usually stationarity or trend-stationarity. The stationary series has a constant mean and variance over time. The rolling average and rolling standard deviation of time series do not change over time:

- Null Hypothesis (H_0): It suggests that the time series has a unit root, meaning it is non-stationary. It has some time-dependent structure;
- Alternate Hypothesis (H_1): It suggests that the time series does not have a unit root, meaning it is stationary. It does not have a time-dependent structure.
- p-value > 0.05: Accept the null hypothesis (H_0); the data have a unit root and are non-stationary;
- p-value $\leq$ 0.05: Reject the null hypothesis (H_0); the data do not have a unit root and are stationary.

Figure 18 shows that *p*-value is less than 0.05. After running the data, the test statistics value was −12.1262. The more negative the test statistics means are, the more the null hypothesis is rejected. Therefore, the data reject the null hypothesis H_0 with a significance level of less than 1%, which means the data are a stationary dataset and do not have a unit root.




```
<Results of Dickey-Fuller Test>
Test Statistic                  -12.1262
p-value                           0.0000
#Lags Used                       12.0000
Number of Observations Used   93812.0000
Critical Value (1%)              -3.4304
Critical Value (5%)              -2.8616
Critical Value (10%)             -2.5668
dtype: float64
```

Figure 18. Dickey-Fuller test with hypothesis testing.

5. Conclusions

Big data analytics was performed to analyze data collected in the Nine Springs Wastewater Treatment Plant from 1996 to 2019. The methods include data collection, data understanding, data preparation, data mining, and evaluation. The following conclusions can be drawn:

The background, goals, and unit processes of a WWTP must be studied to understand the dataset clearly.

In data preparation, the first step was to select the data collection locations. The selected datasets should be the influent and effluent data. The second step is to clean datasets by filling or deleting missing values. In the Nine Springs WWTP, the significant parameters were TSS, TP, TKN, NH_3N, and BOD_5, which are the essential parameters affecting wastewater treatment quality. Data visualization helped assess the relationship between influent and effluent.

In the data preprocessing or data mining stage, the descriptive statistics function was applied to measure the average and distribution of the numerical values in the dataset. The correlation coefficient helped calculate the relationship among parameters. The effluent BOD_5 closely correlated to other parameters. The correlation values show that TSS, NH_3N, and TKN highly correlate to effluent BOD_5, which are 0.443, 0.428, and 0.342, respectively. TP has less correlation, which is 0.095. However, TP is one of the critical regulatory parameters, so it recommends applying as an input for model development.

The normality test showed that effluent BOD_5 data rejected the null hypothesis, which means that the data are not a Gaussian distribution nor a normal distribution. In addition, kurtosis and skewness testing can help determine normal distribution. The result showed that the kurtosis is 5.205, implying the heaviness of the tails of the distribution, and the skewness is 0.779, meaning the data are moderately skewed.

Visualizing the data using a box plot and graphical representation showed that the median effluent BOD_5 in 2014 was higher than the other years, which is not normal.

Therefore, the data in 2014 should be removed as well as the data in 2004 and 2019 because of the missing data.

Finally, the Dickey-Fuller test was performed in the evaluation step to assess the data from 2015 to 2018. The result showed that the data rejected the null hypothesis H_0, implying that the data are stationary. Therefore, it can be used for a predictive model when the data have a clear trend and seasonality.

In conclusion, data analytics with statistical analysis is essential for analyzing and interpreting data, especially big data. This method will help find insights, remove unnecessary information, obtain a suitable dataset, and develop a precise predictive model. In addition, this step is vital for applying artificial intelligence (AI) to wastewater treatment plant operations and diagnoses of probable upset leading to violation of effluent limits. The data analytics method developed in this study was the first step in developing a predictable AI model for wastewater treatment plant operation.

Author Contributions: Conceptualization, J.K.P.; data curation, P.W.; formal analysis, P.W.; methodology, P.W.; software, P.W.; supervision, J.K.P.; validation, P.W.; visualization, P.W.; writing—original draft, P.W.; writing—review and editing, J.K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ritchie, H.; Roser, M. Water Use and Stress. Available online: https://ourworldindata.org/water-use-stress (accessed on 28 May 2020).
2. Maiza, M.; Beltrán, S.; Westling, K.; Carlsson, B.; Mulas, M.; Bergström, P.-H.; Hyyryläinen, S.-M.; Urchegui, G. DIAMOND: AdvanceD Data Management and InformAtics for the OptimuM OperatiON anD Control of WWTPs. In Proceedings of the 11th IWA Conference on Instrumentation Control and Automation, Narbonne, France, 18–20 September 2013.
3. Siegrist, R.L. Introduction to Decentralized Infrastructure for Wastewater Treatment and Water Reclamation. In *Decentralized Water Reclamation Engineering: A Curriculum Workbook*; Siegrist, R.L., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–37, ISBN 978-3-319-40472-1.
4. Millinger, A. The Modernization of SCADA and HMI. Available online: https://www.wwdmag.com/scada-systems/modernization-scada-and-hmi (accessed on 14 June 2021).
5. Newhart, K.B.; Holloway, R.W.; Hering, A.S.; Cath, T.Y. Data-Driven Performance Analyses of Wastewater Treatment Plants: A Review. *Water Res.* **2019**, *157*, 498–513. [CrossRef] [PubMed]
6. Harrou, F.; Dairi, A.; Sun, Y.; Senouci, M. Statistical Monitoring of a Wastewater Treatment Plant: A Case Study. *J. Environ. Manag.* **2018**, *223*, 807–814. [CrossRef] [PubMed]
7. McGowan, S.; Wang, E. *50-Year Master Plan Review of Existing Treatment Facilities*; Malcolm Pirnie: New York, NY, USA, 2008.
8. Han, J.; Kamber, M.; Pei, J. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann Publishers: Waltham, MA, USA, 2011; ISBN 978-0-12-381479-1.
9. Just into Data, L.& J.@ J. into Data Cleaning in Python: The Ultimate Guide (2020). Available online: https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d (accessed on 28 October 2021).
10. Rahman, M.; Wu, H. A Note on Normality Tests Based on Moments. *Far East J. Math. Sci.* **2013**, *2*, 9.
11. D'agostino, R.B.; Belanger, A.; D'agostino, R.B., Jr. A Suggestion for Using Powerful and Informative Tests of Normality. *Am. Stat.* **1990**, *44*, 316–321. [CrossRef]
12. Box Plot. Available online: https://en.wikipedia.org/w/index.php?title=Box_plot&oldid=1049990947 (accessed on 15 October 2021).
13. Privitera, G.J. Introduction to hypothesis testing. In *Statistics for the Behavioral Sciences*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2017.
14. Few, S. Data Visualization for Human Perception. Available online: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception (accessed on 28 May 2020).
15. Li, S. Time Series Analysis, Visualization & Forecasting with LSTM. Available online: https://towardsdatascience.com/time-series-analysis-visualization-forecasting-with-lstm-77a905180eba (accessed on 28 May 2020).

16. Mudugandla, S.Y. 10 Normality Tests in Python (Step-By-Step Guide 2020). Available online: https://towardsdatascience.com/normality-tests-in-python-31e04aa4f411 (accessed on 28 October 2021).
17. McNeese, B. Are the Skewness and Kurtosis Useful Statistics? Available online: https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics (accessed on 28 October 2021).