Table of Content

List of Figures

List of Tables

Chapter 1
INTRODUCTION

This chapter introduces the project the author worked on alongside the author's team. It also includes the background of the project as well as the aims, vision, and mission of carrying out this project. It will also describe the structure and provide insights into the remaining chapters.

1.1 Background

There is no doubt that oil and gas are vital elements to the growth of the economy. There have been traces of oil trade ever since 1875 BC [1]. In this modern, technologically-advanced society, the demand for oil and gas has only continued to grow stronger. It is used for many modern inventions enjoyed by a vast majority of people, such as vehicles, fuels, medical equipment, agriculture, and many more [2]. Additionally, the oil and gas industry has also provided jobs to thousands of individuals [3].

There are many oil and gas reserves in different corners of the world. In Indonesia, in particular, the Energy Ministry has recorded that in January 2021, there is a total reserve of 2.44 billion barrels of oil and 43.6 trillion cubic feet of gas [4]. However, due to the rapidly increasing population and a growing economy, the demand for oil and gas in Indonesia is rising [5]. Furthermore, 50% of Indonesia's energy is derived from oil [5]. This reliance on oil results in Indonesia importing nearly 350,000 barrels per day (BPD) and 50,000 barrels of fuel per day from other countries [5].

Oil and gas have many uses and have a substantial impact on the economy of a country. Therefore, oil and gas industries often make use of dashboard-based software applications in order to help them manage it, such as a Virtual Data Room (VDR) application. A VDR is an online repository that can store data securely and can be accessed by multiple users simultaneously [6]. These kinds of applications can help the oil and gas industries discover which areas could have more oil and gas. It can also help clients visualize the oil and gas data. Lynx and INTViewer are examples of software applications capable of data visualization [7][8]. These applications are similar, yet they also have their differences. Lynx offers petroleum data services and geophysical and Geographical Information System (GIS) services [7]. It offers 2D and 3D seismic viewers and costs at least £250 per user per year [7]. On the other hand, INTViewer is a platform that allows users to check seismic data, geospatial integrity and also process datasets [8]. It can cost up to $4,000/person a year [9]. These types of applications can benefit oil and gas industries greatly; however, they tend to be expensive. Additionally, according to the product owner, some oil and gas companies in Indonesia often search for cheaper and more customizable software similar to Lynx and INTViewer. Therefore, the goal of this project is to develop a VDR application with similar features intended for the oil and gas industries in Indonesia at a lower cost. The proposed VDR application will be developed as a website application as part of the requirements from the author's customer, for example, PT Graha Teknologi Nusantara.

In order to enhance the VDR website application, data science could be incorporated. Data science is the method of obtaining meaningful insights from a large set of data [10]. The data will be analysed and processed so that high-level data analysis can be

performed [10]. The data analysis will then reveal patterns in the data, thus enabling users to draw conclusions regarding the data [10]. This is useful for the VDR as users will be able to understand the oil and gas data, thus gaining meaningful insights from it. Due to the potential of using data science in VDR, this thesis will focus mainly on that issue.

1.2 Scope

This section describes the scope of the author's group as well as the author's individual scope for this project.

1.2.1 Scope of Problem and Solution

The main goal of the project is to develop a VDR application that will benefit the oil and gas industry. As has been mentioned in section 1.1, there are existing applications for this purpose; however, these applications are expensive. Therefore, the author and the author's team will make use of open-source libraries and essential features. Additionally, the author and the author's team will also develop custom features requested by the customers. The application will consist of several features, such as uploading and storing files, deleting and downloading files, viewing maps as well as obtaining oil and gas production data with the use of a predictive model.

1.2.2 Scope of Work

In this project, the author and the author's team had different responsibilities, as shown in Table 1.1. The author's responsibility was to create a predictive model capable of predicting oil and gas production. The author had to collect and scrape valuable data in order to make a dataset. This dataset would then be processed and cleaned to train the machine learning model. The author will make use of machine learning algorithms

to predict the oil and gas production values. Furthermore, the author will conduct a comparative study of models to see which are capable of predicting oil and gas production. From this study, the author will choose the model which will be implemented in the VDR application. In order to choose the model, the author will evaluate the performances of each model using model evaluation techniques. After the model has been selected, trained, and evaluated, the author will connect the model to the website created by the other members of the author's team.

Student	Role
Kotrakona Harinatha Sreeya Reddy	Collecting and Processing Data Using the data collected to develop predictive models Visualizing data through diagrams, such as charts, as well as performing data analytics Acts as SCRUM Team
Elizabeth Chan	Design the frontend of the proposed VDR application that uses GIS Testing (e.g., unit test & integration test) Acts as SCRUM Team
Vicky Vanessa	Designing UI/UX of the frontend of the website application Visualizing the data of oil and gas Testing Act as the SCRUM Master

Table 1.1 : Scope of Activities

1.3 Aim and Benefits

The main aim of this VDR website application is to help the oil and gas industry discover more profitable areas of resources by visualizing oil and gas volume as well as visualizing reserve resources. The author and the team wish to increase the local services in the oil and gas industries as some local companies would like to have their own VDR that matches their needs. Furthermore, this application will help engineers understand complex data.

The author aims to build a predictive model capable of predicting oil and gas production values. This will help the oil and gas industry by showing areas that are more likely to contain more oil and gas. Therefore, the oil and gas industry can focus on the wells which contain more oil and gas, which would save time and money. The model the author builds will be integrated into the website application created by the author's team members. This will be further explained in Chapter 3.

1.4 Structure

This thesis consists of four chapters which will be briefly described in this section

1.4.1 Chapter 1 : Introduction

Chapter 1 introduces the author's topic, the scope, objectives, aims, vision, and mission of this project.

1.4.2 Chapter 2 : Theoretical Foundation

Chapter 2 describes the fundamental theories behind the predictive models designed by the author. It defines specific terms and provides further insights into the problem.

1.4.3 Chapter 3 : Problem Analysis

Chapter 3 will detail the problem even further and describe the works related to the author's project while also briefly describing the model the author intends to make.

1.4.4 Chapter 4 : Solution Design

Chapter 4 focuses on the design of the solution devised by the author; it includes data pre-processing as well as how the models will be manipulated.

Chapter 2

THEORETICAL FOUNDATION

This chapter will delve into the theories and techniques the author used while developing this project. It discusses the process of the Software Development Life Cycle (SDLC), which the author's team will follow while developing the project. Afterwards, it delves into the specifics of a VDR, which the author's team intends to develop. It will then probe into how oil and gas are produced in the reservoir. Additionally, it will discuss how the author intends to build the model to predict oil and gas production using machine learning. Afterwards, this chapter will discuss missing data, outliers, and feature correlation in the dataset used to train the model. It will also examine how to evaluate the performance of the model and how the model can be connected to a website.

2.1 Software Development Life Cycle

SDLC is the process that is made up of steps that a particular software can follow in order to develop in a proper manner [11]. This would make it more likely for the project to be accomplished on time whilst ensuring the quality of the product is suitable for the user [12]. The activities for a specific SDLC can be labelled as [13] :

1) understanding the case,
2) deciding solution scheme,
3) coding based on the solution decided,
4) testing.

However, these activities are quite broad; therefore, they can be broken down even further to illustrate the SDLC process better [13]. The phases of SDLC are

requirements analysis, design, development, testing, and deployment and maintenance [13].

Requirements analysis is the first phase of SDLC. In this phase, the business requirements of the project are gathered. The project managers and stakeholders will discuss to define the requirements of the software. These requirements could include answering questions such as “who will use the software” or “how will the system be used” [13]. After the discussion, a Software Requirement Specification (SRS) document will be created, which will contain the results of the discussion [13].

The main objective of the design phase is to turn the requirements specified in the first phase into an architecture [13]. In this phase, the hardware and system requirements are specified so that the architecture of the software can be defined [13]. Additionally, this phase is where testers are required to define what needs to be tested and how it should be tested [13].

In the development phase, the results of the design phase are converted into a system that meets the user requirements. A common name for this phase is the coding phase. All the developers and engineers play an active role in this phase, and they are required to follow the required guidelines defined beforehand [14]. It is the most extensive yet most crucial phase in the entire SDLC process. Additionally, the process of the development phase will be recorded in a document entitled Source Code Document (SCD) [14].

The next phase is the testing phase, where the software developed in the previous phase will be tested. There is usually a specific team whose purpose is solely for testing the software; their job is to conduct a series of tests on the software [14]. The testing team will document any errors they encounter and send this report to the development team so that the developers can attempt to remove the errors [13]. The testing phase is one of the most essential phases as it decides whether the software is eligible to be released to the users [14].

In the deployment and maintenance phase, the software has passed the testing phase and is bug-free; therefore, it is now deployed and useable by the client [14]. Additionally, in this phase, there are possibilities that the software needs to be updated due to technological advancements. Therefore, the developers need to maintain the software to ensure that its performance will not decline [11].

Over the years, the SDLC model has been adapted into different kinds of models. These models include the *Waterfall Model*, *V-shaped Model*, *Incremental Model*, *Agile Methodology*, and many more [13]. The agile methodology, in particular, is known for constant iterations for software testing and development [11]. In this methodology, it is typical for the development phase and the testing phase to occur concurrently [13]. The Agile methodology contains twelve core principles, which are [15] :

1) customer satisfaction,
2) adaptive to changing requirements,
3) regular software delivery, the faster the better,
4) productive collaboration between developers and stakeholders,

5) support developers by supplying an ideal work environment and believe that they will accomplish the project,
6) direct face to face communication for team discussion,
7) assess progress by checking on working software,
8) encourage maintainable development,
9) constant focus on technical quality and design,
10) simplicity is vital,
11) working units that can organize themselves will provide the ideal output (design, software architecture, requirements),
12) occasional reflection so that the team can improve.

The agile methodology also consists of a framework entitled SCRUM. SCRUM is an agile development methodology that is based on an iterative as well as an incremental process [13]. One of the main features of SCRUM is that it focuses more on feedback, revisions, and frequent customer engagement rather than documenting procedures and predicting a plan of action for accomplishing the project [13]. In SCRUM, there are three prominent roles which are Product Owner (PO), Scrum Master (SM), and Scrum Team (ST). There is a lack of guidelines or descriptions for how the project should be accomplished in SCRUM; most of the decision-making is left to the team doing the project as the team knows best [13]. There are three constants in SCRUM, which are Product Backlog, Sprint Backlog, and Sprint Goal [13]. Product Backlog is the list of things that need to be done by the PO, Sprint Backlog is the list of things selected by the ST that needs to be done in the current sprint cycle, whereas Sprint Goal is the endgame of the current sprint [13]. SCRUM Methodology is beneficial for

complicated projects, and this methodology greatly helps the project progress efficiently [13].

2.2 Virtual Data Room

A VDR will be developed through a website in which the concept of SDLC will be used during the development process. A VDR is based on the concept of a data room. A data room is a valuable tool for the oil and gas industry as the companies can use it whenever they desire to dilute equity in assets [16]. The company places the data in the data room where it can be assessed [16]. If the company wishes to sell the data, buyers can visit the data room and inspect the data [16]. There are different types of data rooms, namely, Physical Data Room (PDR), VDR, and a PDR – VDR combination [16].

PDR is a physical secure room where the data is placed by the seller [16]. However, PDR has mostly been replaced by VDR as it has several drawbacks. A few of the drawbacks of a PDR are that it is expensive, burdensome, and time-consuming compared to VDR [16]. A VDR is a website where documents are uploaded that users can access and asses at ease [16]. A VDR is as secure as PDR and is also always available for as long as the client desires [16]. Furthermore, as a VDR can be accessed online, there is no need for the client or their representative to travel to access the data [16]. The information in the VDR can also be updated immediately to show new information and is immediately accessible [16]. Based on [17], the features of a VDR could include

1) Accessible anywhere and anytime, disregarding operating systems,
2) Downloading and generating documentations such as reports, and

3) Petrotechnical solutions such as reservoir analysis or exploration and production tools.

Figure 2.1 shows some of the features available in a VDR.

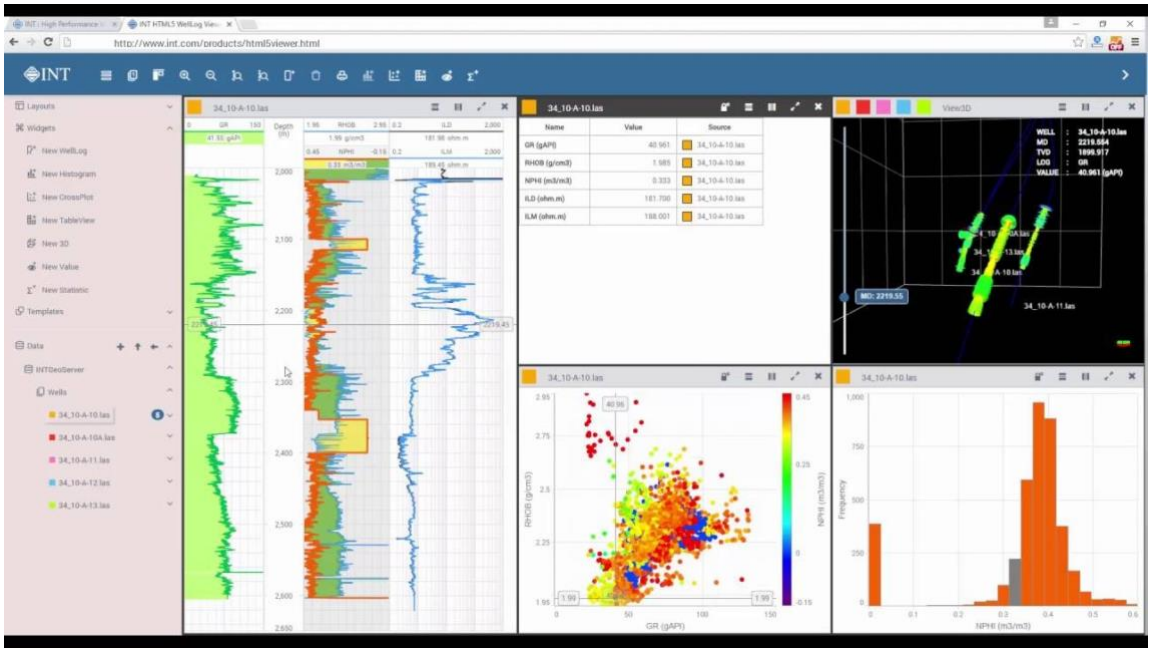


Figure 2.1: Dashboard in a Virtual Data Room

To summarize, a VDR is faster, cheaper, and more efficient compared to PDR [16].

2.3 Oil and Gas in a Reservoir

In order to build an oil and gas predictive model, it is vital to understand how oil and gas are formed in a reservoir and the factors that affect its formation.

2.3.1 Oil Formation

A formula that can be taken into account for oil formation is the oil formation volume factor (B_o). It is the ratio of oil volume and dissolved gas at a specific temperature and pressure that is needed to make one barrel of oil [18]. B_o is either greater than or equal to unity [19].

The equation for the oil formation volume factor is :

$$B_o = \frac{(V_o)pT}{(V_o)_{sv}}. \quad (2.1)$$

In Equation 2.1, B_o is the oil volume factor, V_o is the volume of oil, $(V_o)sc$ is the volume of oil measured under standard conditions, p is the pressure at the reservoir, whereas T is the temperature at the reservoir [18]. From Equation 2.1, it can be inferred that temperature and pressure are essential factors in the formation of oil. Once the oil reaches the surface, it loses the dissolved gas, which leads to changes in the reservoir oil obtained. First of all, the mass of the oil will reduce as it loses the dissolved gas, then the oil will also contract as temperature decreases on the surface [18]. Afterwards, the oil will again expand as the pressure increases [18]. Often the effect of the temperature and pressure changes when the oil reaches the surface is minimal and will cancel out each other [18].

2.3.2 Gas Formation

A formula that can be taken into account for gas formation is the gas formation volume factor (B_g). It is the ratio of the volume of gas at a specific temperature and pressure that is needed to manufacture one standard volume of gas [20]. This equation for gas formation volume factor can be expressed as :

$$B_g = \frac{V_{p,T}}{V_{sc}}. \quad (2.2)$$

In Equation 2.2, B_g is the gas formation volume, $V_{p,T}$ is the volume of gas at the reservoir pressure and temperature and V_{sc} is the volume of gas at standard conditions.

In real life, gases follow the real gas law, which can be expressed mathematically as :

$$pV = znRT, \qquad (2.3)$$

where p is the pressure, V is the volume, n is the number of moles of gas, R is the universal gas constant, T is the temperature, and z is the gas compressibility factor [21]. Variable z can be expressed as :

$$z = \frac{V_a}{V_i}, \qquad (2.4)$$

where V_a is the actual volume of n-moles of gas at a certain temperature and pressure, and V_i is the ideal volume of n-moles of gas at the same temperature and pressure [21]. Therefore, the equation for real gas law should be applied to Equation 2.2. Equation 2.3 is applied onto Equation 2.2 by substituting for the volume (V), which will result in Equation 2.5.

$$B_g = \frac{zTP_{sc}}{T_{sc}P}. \qquad (2.5)$$

In Equation 2.5, *Bg* is the gas formation volume, *P* is the pressure, *T* is the temperature, P_{sc} is 1 atm, T_s is 60°F, and *z* is the gas compressibility factor at standard conditions (1.0) [21]. With the assumption that the standard conditions are represented by $P_{sc} = 14.7\ psia$ and $T_{sc} = 520$, Equation 2.5 can be reduced to :

$$B_g = 0.0283\frac{zT}{P}. \qquad (2.6)$$

2.3.3 Pressure and Temperature in Oil and Gas Formation

Figure 2.2 shows the phase diagram of oil and gas in a reservoir. As stated previously in Section 2.3.1, when oil is drilled, it also contains dissolved gas. Therefore, in a reservoir, there exist 2 phases, namely liquid and gas. Based on the current pressure and temperature, the phase diagram shows that there is a region where the mixture will be either liquid or gas only and a region where both liquid and gas are at equilibria. The black line, known as the Bubble Point Line, denotes where both phases begin to appear [22]. Before the bubble point, the only phase that exists is liquid. However, at a constant temperature, as pressure decreases, the total volume of gas increases, whereas the volume of oil decreases [22]. This property is supported by Le Chatelier's Principle, which states that an increase in volume or decrease in pressure would increase the formation of the gaseous product.

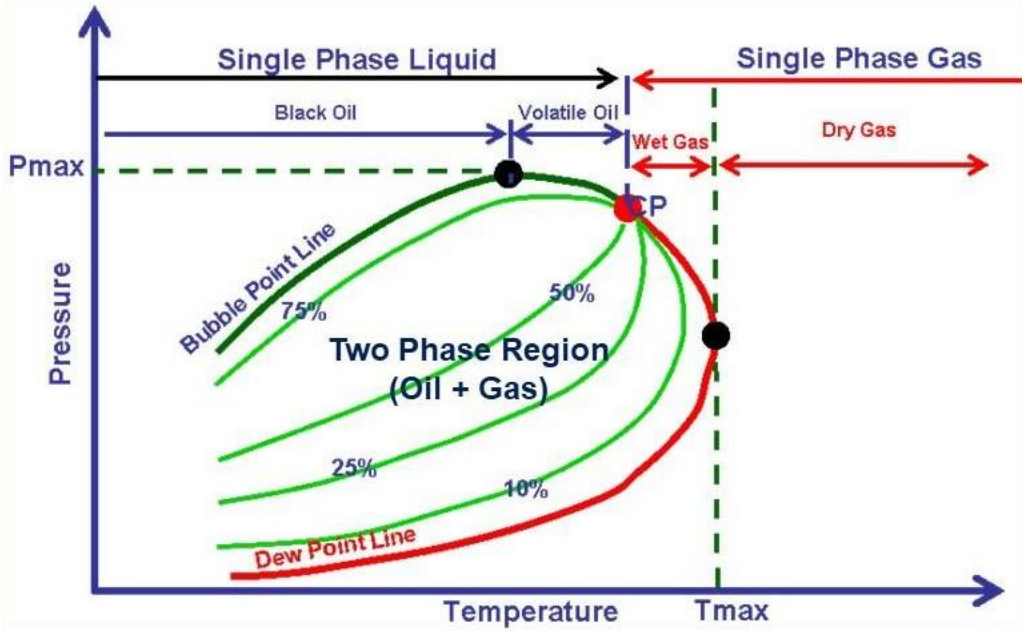


Figure 2.2: Phase Diagram of Oil and Gas [22]

As the pressure continues to decrease, more heavier molecules become gaseous, increasing the density and viscosity of the gas [22]. Subsequently, there will be a point where only a small portion of liquid remains; this is called the Dew Point [22]. If the pressure drops below this point, the only phase that exists is gas [22].

2.4 Machine Learning

Upon briefly explaining the oil and gas formation process, this section will delve into machine learning. Machine learning is the central machinery in building a prediction model of the oil and gas production data. It is defined as *the capability of a system to be able to learn from data and algorithms to automate the process of solving certain tasks* [23]. It is a branch of Artificial Intelligence (AI) that centers on using data and algorithms to echo the way humans act and learn [24]. Machine learning helps uncover insights, make classifications, and predictions from the data given in order to aid users [24]. Machine learning depends on a dataset, which is a collection of data that will be regarded as one unit by the machine [25]. This dataset will act as the "training data" for the machine to learn. It is preferable to have large amounts of data as this means the machines would learn more efficiently and be able to solve problems with better accuracy. However, the quantity of the dataset is not the only significant factor in machine learning; the quality of the dataset is also a notable factor. A machine would perform significantly better with a high-quality dataset in contrast to a poor-quality dataset. In the context of the author's problem, a high-quality dataset would mean a dataset that provides the oil production value, the gas production value, and data on sensors such as pressure and temperature.

Machine learns in different ways, namely, supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is a part of machine learning and artificial intelligence; it is learning by means of mapping between a set of input variables and output variables [26]. The input variables are fed into the machine learning model, and after the training phase, it will apply what it learned to unknown data [27]. This type of machine learning is one of the most common methods and is usually used for classification and regression problems [27]. There are several types of supervised machine learning models, namely Naive Bayes, linear regression, support vector machine (SVM), KNN, and others [28]. On the other hand, unsupervised learning aims to derive meaningful information from unlabelled data [29]. It is not as commonly used as supervised learning [29]. Reinforcement learning is another complex part of AI where the model is trained to make decisions sequentially [30]. The output is dependent on the state of the current input, and the following input would then be reliant on the output of the previous output [30].

2.5 Predictive Models

Predictive modelling is a part of machine learning. It is the process of predicting future outcomes from data gathered beforehand. There are different types of predictive models; in order to decide which type of model should be used, it is necessary to understand what type of variable the model needs to predict. If the model aims to predict discrete variables, then classification machine algorithms should be used. However, regression machine algorithms should be used if the model aims to predict continuous variables. A part of machine learning that can be used for predictive modelling is deep learning. Deep learning consists of multiple layers of algorithms known as an artificial neural network (ANN). An ANN is designed to behave similarly

to a human brain. The simplest ANN consists of a single neuron, also known as a perceptron [31]. These neurons will be stacked on top of one another, which will create layers [31]. Each layer will learn something new and pass it on to the next layer that will learn something else. There are different types of ANN, such as recurrent neural network (RNN) and convolution neural network (CNN). RNN is a kind of neural network that works well with sequential data, whereas CNN works well for image and video data [31] [32]. In a study [33], a researcher compared the use of deep learning algorithms and tree-based machine algorithms on a variety of datasets for prediction. The research showed that deep learning tends to perform better on unstructured data, such as images or voice [33]. On the other hand, tree-based algorithms function better with tabular structured data compared to deep learning [33]. Tabular data is a dataset that consists of a set of rows and columns; it is one of the most common types of datasets.

Tree-based algorithms are a well-known part of machine learning, more specifically, predictive modelling. Tree-based regression algorithms are commonly used for predictive analysis of numerical values [34]. This regression model works by investigating the connection between variables [34]. It will determine the value of one variable based on the other variables present [34].

A commonly used algorithm for predictive models is the random forest algorithm [34]. This is a supervised learning algorithm that is based on the ensemble learning method [34]. Ensemble learning is the process of combing the prediction results of several machine learning algorithms [34]. The goal of this is to make the prediction results more accurate. The random forest algorithm combines the predictive results of several

decision trees [34]. The respective decision trees do not interfere with one another [34]. There are two steps for the random forest algorithm; the first step is building *n* decision tree regressors, where *n* is the number of decision tree regressors [34]. These trees can be modified by specified hyperparameters, such as the strategy best used to split the node into sub-nodes or the function used to measure the quality of the split [35]. The final step would be to take the average prediction values of the decision tree regressors; this average will serve as the final output of the model [34].

Another algorithm for predictive models is the gradient boosting algorithm. This algorithm is based on the concept of boosting [36]. In terms of regression, boosting is a procedure of building strong regressors by combing weak learners [36]. This algorithm has three requirements, namely loss function, weak learners, and additive model.

A loss function would measure how similar the values predicted by the algorithm are to the actual values. In terms of regression problems, the loss function used could be Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Coefficient of Determinant (R^2). [36]. Additionally, this algorithm is based on the idea that combing multiple weak learners would result in an accurate result. The weak learners used in gradient boosting are typically decision trees [36]. Gradient boosting is also an additive model as it adds the weak learners one by one. Every new predictor would gain new knowledge from the error of the previous predictor, and it would work to correct the error, which would result in a better model [36].

2.6 Data Analytics Pipeline

The predictive models have to be trained on a dataset so that they can learn; however, before training, it is vital to understand and clean the dataset used. The steps that can be taken to understand and clean the dataset are shown in Figure 2.3. These steps will be explained further in the upcoming sections.

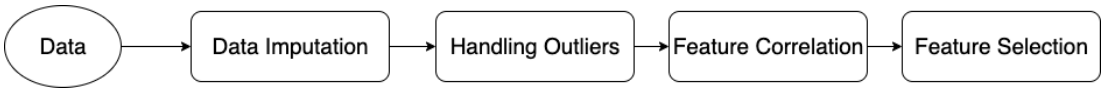


Figure 2.3 : Data Analytics Pipeline

2.7 Data Imputation

An essential part of model training is the quality of the dataset. A possible problem in a dataset is missing data. Missing data in a dataset could prove to be problematic as it could affect the model's ability to perform well.

2.7.1 Mechanism of Missingness

There are three possible mechanisms for missing data in a dataset; these mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

In the MCAR mechanism, the missing values are unrelated to the other values in the dataset, both missing and present; therefore, these missing values are random. In this situation, the missing values are considered negligible as they would not significantly impact the model performance [37].

In the MAR mechanism, the missing values are also random such as in MCAR; however, there are possibilities of the data in question being dependent on other values in the dataset. In this situation, the missing values should be considered as they could affect the model's performance. However, the effect is not extreme [37].

In the MNAR mechanism, the missing values are strongly dependent on the other values in the dataset, both missing and present. MNAR is the most serious reason mechanism for missing data as it cannot be ignored and could affect the model's performance [37]. In these cases, it is recommended to validate the data collection process [37].

In order to counteract the effects of the missing values on the dataset, data imputation methods could be used. Data imputation methods include central value imputation and forward filling.

2.7.2 Central Value Imputation

Central value imputation is the process of filling in the missing data in the dataset with their central tendencies [38]. These central tendencies could either be the mean, median, or mode. The mode is typically used to fill in the missing data for categorical variables, whilst the mean and median are often used to fill in for numerical variables [39]. The central tendencies are deemed as reasonable estimates for filling in the missing data. However, this method would not yield ideal results if the missing data follows the MNAR mechanism, and it could also introduce bias in the dataset [40]. Additionally, filling in the missing values with the mean could reduce the variance in the data set [38].

2.7.3 Forward Filling

Forward filling is the process of filling in the missing data with the value observed before the missing value [41]. For instance, in a dataset such as Table 2.1, the forward filling method could be used to fill in the missing data. Using this method would change the dataset, as shown in Table 2.2. This method is generally used for time series datasets and is one of the easiest ways to deal with missing values. However, a disadvantage of this method is that it will not be able to fill in the missing value if there is no value prior to the missing value.

5	NaN	4
NaN	3	2
3	2	NaN

Table 2.1 : Sample Dataset

5	NaN	4
5	3	2
3	2	2

Table 2.2 : Sample Dataset after Forward Filling

2.8 Outliers

Besides missing data, another problem possible in a dataset is the presence of outliers. Outliers can be defined as *a data in a dataset that strays from the other data* [35]. It is necessary to detect these outliers as they could skew the model's training which would

reduce the accuracy of the model [42]. The removal of outliers is usually one of the earliest steps in a machine learning problem [42]. There are several methods that can be utilized in order to identify these outliers. One of those methods is to use Tukey's method. The Tukey's Method is based on statistics where data is expected to follow a distribution model such as normal distribution [43]. A data is considered an outlier if it deviates from the model [43]. The Tukey's Method divides the dataset into quartiles; the quartiles commonly used are the lower quartile (Q_1), median (Q_2), and upper quartile (Q_3) [43]. The equation for a quartile is :

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_2 - l_1), \tag{2.7}$$

where Q_r is the r^{th} quartile, l_1 is the lower limit, l_2 is the upper limit, f is the frequency, and c is the cumulative frequency of the class preceding the quartile class [43]. The Tukey's Method involves calculating the Interquartile Range (IQR) between the lower quartile and the upper quartile in a boxplot [43]. The equation for the IQR is

$$IQR = Q_3 - Q_1. \tag{2.8}$$

In order to accurately determine which data is an outlier, the Tukey's Method calculates the upper limit and lower limit of the data distribution. The equation for the upper limit is

$$Upper\ Limit = Q_3 + (1.5 * IQR). \tag{2.9}$$

On the other hand, the equation for the lower limit is

$$Lower\ Limit = Q_1 - (1.5 * IQR). \quad (2.10)$$

The Tukey's Method will remove any data that does not fall between the upper limit and lower limit [43].

2.9 Correlations

In order to better understand a dataset, the correlation between features in the dataset could be considered. Correlation is known as a statistical measure that describes how one feature is related to another feature [44]. It is often used during Exploratory Data Analysis (EDA) to gain a better understanding of how a feature affects other features in the dataset. There are different types of correlations, namely positive correlation, negative correlation, and no correlation [44].

A positive correlation denotes that when the value of one feature increases, the value of the other feature increases as well [44]. In a graph format, a strong positive correlation would have a positive gradient, as shown in Figure 2.4.

A negative correlation indicates that when the value of one feature increases, the value of the other feature decreases [44]. A negative correlation would have a negative gradient, as shown in Figure 2.5.

No correlation indicates that the features being assessed are not related; therefore, a change in one feature would not impact the other feature [44]. In a graph format, features with no correlation would look like Figure 2.6.

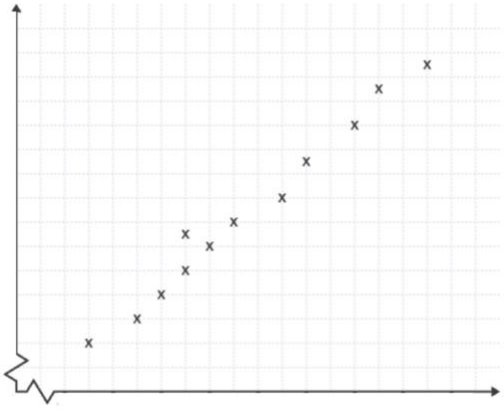


Figure 2.4 : Positive Correlation [45]

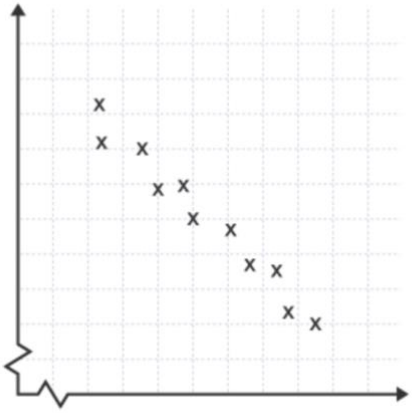


Figure 2.5 : Negative Correlation [45]

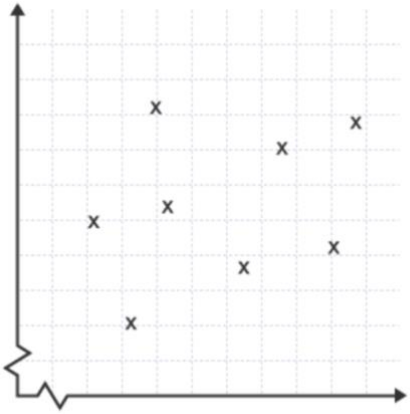


Figure 2.6 : No Correlation [45]

For numeric features, the commonly used methods for measuring the correlation between features are Pearson Correlation and Spearman Correlation.

2.9.1 Pearson Correlation

In Pearson correlation, the features being compared get assigned a value between -1 and 1 [46]. A correlation value of 1 or – 1 would mean that the features being compared are strongly related to one another. A correlation value of 1 expresses that if one feature is present, then the other feature will unquestionably be present as well [46]. In

addition to this, a correlation value of -1 would mean that if one feature is present, then the other feature will undeniably be absent [46]. There are also possibilities of having a correlation value of <1 or >-1. This means that the correlation is almost exactly positive or negative; however, there exists a small number of records that behave differently [46]. On the other hand, a correlation value of 0 would mean that the absence or presence of a feature is in no way related to the presence or absence of another feature [46].

The equation for Pearson correlation is :

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}, \qquad (2.11)$$

where *r* is the Pearson correlation coefficient, *x* is the values in the first set of data, *y* is the values in the second set of data, and *n* is the total number of values.

2.9.2 Spearman Correlation

The Spearman correlation is a method that measures the strength and direction of the relationship between two features in a dataset [47]. Spearman correlation requires continuous data, which has a monotonic relationship. This means that when one feature increases, the other feature could either increase or decrease [47]. However, the relationship between the features does not have to be linear [47]. The correlation values in Spearman correlation follow the same principle as those in Pearson correlation. The values range from – 1 to 1 as well. If the correlation value is – 1, then as one variable increases, the other variable would decrease [47]. If the correlation value is 0, then a change in a variable would not affect the other variable [47]. On the other hand, if the

correlation value is 1, then as one variable increases, the other variable would increase as well [47].

There are two equations that can be used to calculate Spearman's correlation. The first equation is :

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}, \qquad (2.12)$$

where ρ is the spearman correlation, d_i is the difference between the features, and n is the total number of values [48]. Equation 2.12 can only be used if there are no duplicates in the dataset. If duplicates exist in the dataset, then the second equation will be used. The second equation is :

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \qquad (2.13)$$

where ρ is the spearman correlation, x is the value of feature x, $\bar{x}$ is the mean of feature x, y is the value of feature y, and $\bar{y}$ is the mean of feature y [48].

2.10 Feature Selection

After understating the dataset, the process of feature selection could be implemented. Feature selection is the process of cutting down the input variables which will be fed into the models [49]. This is useful as it gets rid of the noise in the dataset so that the model can focus on valuable information [49]. In order to determine which features are ideal to be used in the dataset, the Pearson correlation of the features should be taken into consideration as the values in the dataset are numerical [50]. It is ideal to add highly correlated features for the model's training. However, highly correlated parameters should not be the only features added to the model as they could reduce the

model's accuracy [50]. It would lead to a lack of variation in the data or even result in data leakage, which would make the model perform unrealistically well [50].

2.11 Model Evaluation

After the dataset has been cleaned and the model has been trained, it is time to evaluate the performance of the model. Model evaluation is vital as it allows researchers to determine whether or not the model made is accurate. In order to evaluate models, researchers make use of metrics; the metrics for regression models are MAE, MSE, RMSE and, R^2. The MAE, MSE, and RMSE metrics greatly penalize outliers as their' value increases significantly in the presence of outliers [51]. For these metrics, a higher value indicates poor performance. However, RMSE is generally preferred over MAE and MSE as RMSE uses the same units as the variable in the y-axis [51]. The R^2 metric is also another ideal metric to consider as it is able to explain how well the model can predict the value compared to the original value [51].

2.11.1 Root Mean Square Error

This metric is the root squared average difference between the actual value and the predicted value [51]. RMSE is the square root of the MSE metric. The lower this value, the lower the deviations between the actual and predicted values [51].

The formula for RMSE is,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_p - y)^2}{n}}, \qquad (2.14)$$

where y_p is the predicted value, *y* is the actual value, and *n* is the number of values [51].

2.11.2 Coefficient of Determinant

This metric is the measure of how well the regression model has predicted the value based on the actual value [51]. R^2 generally ranges from 0 to 1; however, there are instances when the value could be negative [51]. A R^2 value closer to 1 would mean that the model gives an accurate prediction. The formula for R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_p - y\right)^2}{\sum_{i=1}^{n}(\bar{y} - y)^2}, \tag{2.15}$$

where y_p is the predicted value, y is the actual value, and $\bar{y}$ is the average of the actual values [51].

2.12 REST API

Sometimes the models developed might be used by external applications such as websites. In these cases, REST API could be used to connect the models to the external application. Representational State Transfer (REST) is a type of architectural style that specifies principles that will act as a guide for website architecture design [52]. The REST API allows users to access web services in a simple manner. Users use HTTP methods, namely GET, POST, DELETE, PUT, and PATCH, to operate the resources such as websites [52]. The GET method is mainly used to read information; this method does not allow information modification [52]. The POST method is used to create new resources which are subordinate to another parent resource [52]. The DELETE method is used to delete an existing resource [52]. The PUT method is used to update a resource that is present; if the resource specified is not present, then a new resource could be generated [52]. The PATCH method is also used to update resources,

similar to the PUT method [52]. However, the PATCH method only performs partial updates; it will not wholly change the resource [52]. Unlike the PUT method, the PATCH method is not capable of creating a new resource [52].

REST API is the ideal method to connect to an external application as, based on the adoption trend, REST API is widely accepted by many developers [53]. Furthermore, REST separates the client side and server side, which is advantageous as if one component fails, it would not impact the other components [54]. In addition to this, REST is capable of adapting to any type of schema or platform [54].

Chapter 3

PROBLEM ANALYSIS

This chapter will discuss the problem statement of this project as well as the proposed solution for the problem. It will also discuss existing works done in this field.

3.1 Related Works

In a study [55], the researcher made use of 2 models to predict the concentration of gas. The researcher used a long and short term memory (LSTM) model and a random forest model and compared the results. An LSTM model is a variant of RNN, which is capable of remembering past information. This makes it suitable for predicting features. However, as mentioned in Section 2.5, deep learning models do not perform as well as tree-based algorithms when it comes to structured tabular data. On the other hand, the random forest algorithm is a tree-based algorithm that performs exceptionally well on tabular data. However, there are possibilities that random forests models would overfit [56]. In this study, the models were evaluated with the R-squared score, RMSE, and MAE. The LSTM has an R-squared value of 0.31, an RMSE value of 0.45, and a MAE value of 0.56. On the other hand, the random forest model has a R-squared value of 0.95, RMSE value of 0.23, and MAE value of 0.34. From the values of these evaluation metrics, the researcher concluded that the random forest model was simpler and gave better results than the LSTM model.

In an article [57], the researcher made use of Facebook's Prophet model in order to predict gas production. The dataset used by the researcher was Canadian's natural gas production; the dataset contained two columns which were the date and the volume of gas. The model was evaluated with the R-squared score and the Mean Absolute Error

(MAE). The R-squared value was 0.911, whereas the MAE score was 7782. An advantage of using the Prophet model is that the results are easy to understand [58]. However, it requires a large dataset as it is recommended to have at least two or three years of historic data [58]. The Prophet model works better if the dataset contains daily and weekly observations [58]. Furthermore, though this model performs quickly, the results are often less accurate compared to when other algorithms are used [59].

Another article [60] shows a researcher using linear regression to predict oil production. The dataset used was the Volve dataset which is located in the North Sea and was updated on a daily basis from 2005 to 2016. The linear regression model was evaluated with the R-squared score; the value was 0.55. An advantage of linear regression is that there are low chances of the model overfitting [61]. Additionally, it is easy to implement and works exceptionally well on variables that have a linear relationship [61]. On the other hand, linear regression models perform poorly when the variables are dependent on one another [61]. Furthermore, there are possibilities for linear regression models to underfit [61].

The same article [60] also showed the researcher using polynomial regression to predict oil production. The same dataset used for the linear regression model was used for this polynomial regression. The dataset features were converted into their higher orders, and the linear regression algorithm was applied to it. The model was evaluated with the R-squared score; the value was 0.95. An advantage of polynomial regression is that it works well even if the variables do not have a linear relationship [62]. Furthermore, the dataset size does not matter, as polynomial regression works well regardless of dataset size [62]. On the other hand, polynomial regression is extremely

sensitive to outliers; the results could change drastically with the presence of one outlier [62].

The information obtained from this research is summarized in Table 3.1.

Table 3.1: Summary of Research

Title	Model Used	Metrics / Performance	Predicted Feature
"Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway"	LSTM	R-squared : 0.31 RMSE : 0.45 MAE: 0.56	Gas Concentration
"Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway"	Random Forest	R-squared : 0.95 RMSE : 0.23 MAE: 0.34	Gas Concentration
"Using Facebook Prophet for Forecasting Natural Gas Production"	Facebook's Prophet	R-squared : 0.911 MAE : 7782	Gas Production Value
"Prediction of Oil Production by applying Machine Learning on Volve Field Production Data."	Linear Regression	R-squared : 0.55	Oil Production Value
"Prediction of Oil Production by applying Machine Learning on Volve Field Production Data."	Polynomial Regression	R-squared : 0.95	Oil Production Value

3.2 Problem Statement

As has been mentioned in Chapter 1, the main goal of this project is to create a VDR website application to help the oil and gas industry. The author wants to enhance the VDR website application by providing a way for users to see which areas contain more

oil and gas. Oil and gas companies have to drill exploratory wells to discover whether or not there is a presence of oil and gas [63]. If the presence of oil and gas is detected, then the company would drill more wells, known as development wells, to obtain the oil and gas [63]. However, development wells do not always contain a large amount of oil and gas; therefore these wells end up being abandoned [63].

3.3 Proposed Solution

Machine learning can help determine how likely the well would contain oil and gas. Through machine learning predictions, users can focus on the wells which contain more oil and gas which would save time and money as they would not waste time on wells that contain less oil or gas. The author plans to develop a predictive model that can predict oil and gas production. This section describes the model the author intends to use as well as how the model will be integrated into the website created by the author's team. It will also discuss the assisting datasets that the model will train with.

3.3.1 Model Selection

As discussed in Section 3.1, there are several models that have been used in the field of oil and gas production prediction. Amongst these models, the random forest algorithm was shown to achieve one of the best results. Section 2.5 discussed how tree-based algorithms are more ideal for tabular datasets. Therefore, as the author will use a tabular dataset as described in Section 3.3.3, this project will use tree-based algorithms. The author will compare the random forest algorithm and gradient boosting algorithms to determine which one performs better. The gradient boosting algorithm has the capability of giving a more accurate result compared to the random forest algorithm. This is because in the gradient boosting algorithm, the trees are trained one by one; thus, the current tree is capable of correcting the error of the

previous one [64]. The author will test the chosen algorithm on different hyperparameters to determine which hyperparameter would give a better result.

3.3.2 Model Integration

Figure 3.1 shows the flowchart of the model integration. The user will first upload the file containing the data, such as the pressure and temperature of the oil and gas wells. The model will then predict oil and gas production from this file, and the user will have the option to download the prediction results.

3.3.3 Assisting Datasets

For the project, the author utilized two open-sourced datasets. The first dataset is entitled Volve, whilst the second dataset is entitled Kyle Master. The Volve dataset contained 15,634 rows of data and was obtained from *Kaggle*. On the other hand, the Kyle Master dataset contained 27,324 rows of data and was obtained from the online data centre of the *Oil and Gas Authority*. It is ideal to use a large dataset as it would lead to lower estimation variance, which means the model will be able to predict more accurately. Both Volve and Kyle Master datasets contain valuable information. However, in order to ensure that the data in these datasets are in better shape for a machine learning model, data cleaning and pre-processing must be done.

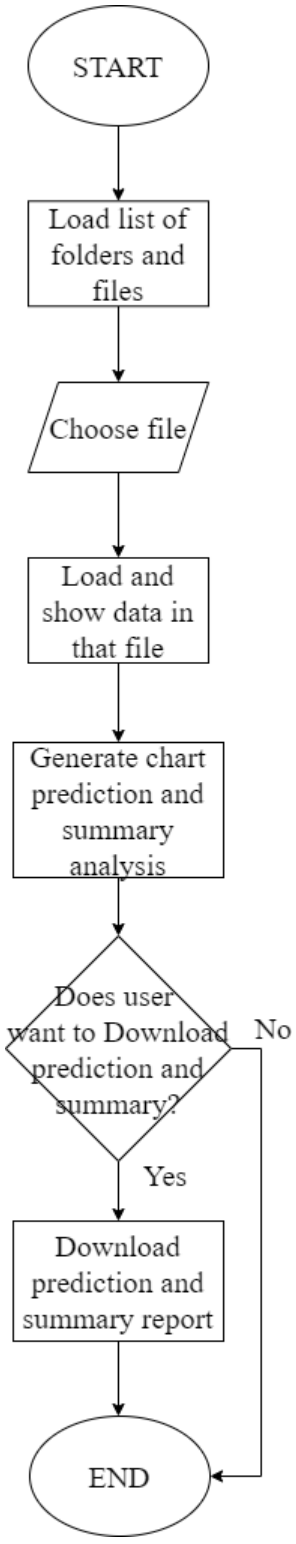


Figure 3.1 : Flowchart for Integration

Chapter 4

SOLUTION DESIGN

Figure 4.1 shows the overall process that will be explained in this chapter. The processes are data cleaning and pre-processing, model training, and model evaluation.

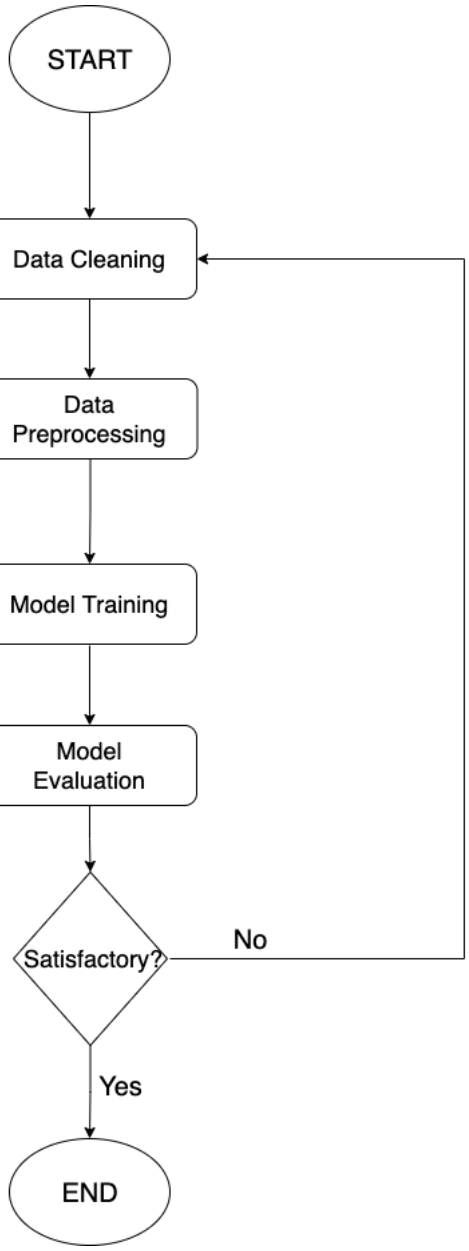


Figure 4.1: Flowchart

4.1 Data Cleaning and Pre-processing

This section will highlight the steps taken in order to clean and process the data for model training.

Table 4.1 : Observations for Missing Data

Dataset	**Volve**	**Kyle Master**
Observation	Contains mainly "<1" and "1" feature correlation values, meaning the features are highly dependent on one another. A value of "<1" denotes that the correlation is almost exactly 1.	Feature correlation values are mostly 0.1, and some features have a correlation value of 1, meaning most of the features do not show much correlation, however, few features are highly correlated.
Missing Data Mechanism	Missing Not at Random (MNAR)	Missing at Random (MAR)

4.1.1 Empty Data Analysis

Volve and Kyle Master contained missing data; therefore, it is imperative to check the relationship between the features in the dataset. This is done so that it can be determined whether or not the presence of the missing value is correlated to other values in the dataset. In order to check this, a heatmap was used to see the correlation values on both datasets. Table 4.1 describes the observations derived from the heatmaps. As stated in Table 4.1, the Volve dataset follows the MNAR mechanism, whereas the Kyle Master dataset follows the MAR mechanism. Section 2.7.1 states that these missing mechanisms imply that the missing values are dependent on one another. Thus, it should not be ignored and should either be deleted or filled in using data imputation methods.

4.1.2 Data Imputation

As has been mentioned in section 4.1.1, both Volve and Kyle Master dataset contains missing values. Additionally, the missing data mechanisms are not MCAR as the dataset contains missing values that are dependent on one another. Therefore action

should be taken to ensure the model performance will not be affected. For this project, the author will use two methods and compare the feature correlation to see which method would make the model perform better. The first method the author will use is forward filling, where the empty value is replaced by the last observed record. The second method used is central value imputation, where the author will fill in the missing values with the median value of the feature.

4.1.3 Correlation in Dataset

In this section, this paper will explore the correlations between the features in the dataset. The author made use of the concept of Pearson's correlation, which was describes in Section 2.9.1, to calculate the correlation between the features. Table 4.2 describes the features with the highest correlation values in Volve and Kyle Master datasets, respectively, when the respective data imputation methods are used. The feature correlation for Volve using median imputation is not as strong as the feature correlation when forward filling is used. On the other hand, the feature correlation for Kyle Master using forward filling is similar to when median imputation is used.

Table 4.2 : Features with High Pearson Correlation in Volve and Kyle Master

Volve				**Kyle Master**			
Forward Filling		**Median Imputation**		**Forward Filling**		**Median Imputation**	
Features	Correlation	Features	Correlation	Features	Correlation	Features	Correlation
BORE_OIL_VOL and BORE_GAS_VOL	0.999	BORE_OIL_VOL and BORE_GAS_VOL	0.999	Oil (m3) and Gas(m3)	0.428	Oil (m3) and Gas(m3)	0.430
AVG_DOWNHOLE_PRESSURE and AVG_DOWNHOLE_TEMPERATURE	-0.844	AVG_DOWNHOLE_PRESSURE and AVG_DP_TUBING	0.697	Av. DHT (Deg C) and Av. DHP (bar)	-0.596	Av. DHT (Deg C) and Av. DHP (bar)	-0.567
AVG_WHT_P and AVG_WHP_P	0.677	AVG_WHT_P and BORE_WAT_VAL	0.674	Av. WHT (Deg C) and Oil (m3)	0.500	Av. WHT (Deg C) and Oil (m3)	0.503

4.1.4 Feature Selection and Conversion

This section will explain and justify which features will be used for the model's training and discuss feature conversion. Table 4.3 shows the columns of each dataset that have the same meaning displayed side by side. For instance, *DATEPRD* in Volve is the same as *Date* in Kyle Master. As the goal is to create a model that can predict oil and gas production, it is essential to include their production values. In Volve, the first two features selected for model training are *BORE_OIL_VOL* and *BORE_GAS_VOL*. As mentioned in Section 2.3, oil and gas formation are also reliant on pressure and temperature. Therefore, *AVG_DOWNHOLE_PRESSURE* and *AVG_DOWNHOLE_TEMPERATURE* are also included. Oil and gas production can also be improved by water injection [65]; therefore, *AVG_WHP_P* and *AVG_WHT_P* are added to the model's training. *ON_STREAM_HRS* will also be added as this column shows how long the machine operates. In Kyle Master, the first two features selected are *Oil (m3)* and *Gas (m3),* as these features contain the production value of oil and gas. *Av. WHT (Deg C)* and *Av. WHP (bar)* are also included as water injection improves oil and gas production. Furthermore, these features have a decent correlation with *Oil (m3)* and *Gas (m3).* Additionally, as oil and gas production is reliant on the pressure and temperature of the reservoir, the features *Av. DHT (Deg C)* and *Av. DHP (bar)* are added for the model's training. Lastly, *Hours Online* will also be added for training the model.

These datasets both have similar columns even though the names are different. For instance, *Av. DHT (Deg C)* and *Av. DHP (bar)* in the Kyle Master dataset has the same meaning as *AVG_DOWNHOLE_PRESSURE* and *AVG_DOWNHOLE_TEMPERATURE* in the Volve dataset. Additionally, *Oil (m3)*

and *Gas (m3)* in the Kyle Master dataset have the same meaning as *BORE_OIL_VOL* and *BORE_GAS_VOL*. However, the unit of measurement in each dataset is different. Therefore, it needs to be standardized so that the model will perform better. Hence, the temperatures will be standardized into ^{o}C (degrees Celsius), while the pressures will be standardized into *bar*, and the volumes will be standardized into m^3 (meter cubic).

Table 4.3 : Column Headings in Volve and Kyle Dataset

Volve Dataset	Kyle Master Dataset	Unit of Measurement	Selected for Model Training
DATEPRD	Date	-	No
WELL_BORE_CODE	Wellbore ID	-	No
NPD_WELL_BORE_CODE		-	No
NPD_WELL_BORE_NAME		-	No
NPD_FIELD_CODE		-	No
NPD_FIELD_NAME		-	No
NPD_FACILITY_CODE		-	No
NPD_FACILITY_NAME		-	No
ON_STREAM_HRS	Hours Online	hours	Yes
AVG_DOWNHOLE_PRESSURE	Av. DHP (bar)	bar	Yes
AVG_DOWNHOLE_TEMPERATURE	Av. DHT (Deg C)	^{o}C	Yes
AVG_DP_TUBING		-	No
AVG_ANNULUS_PRESS		bar	No
AVG_CHOKE_SIZE_P	Platform Choke %	%	No
AVG_CHOKE_UOM		%	No
AVG_WHP_P	Av. WHP (bar)	bar	Yes
AVG_WHT_P	Av. WHT (Deg C)	^{o}C	Yes
DP_CHOKE_SIZE		-	No
BORE_OIL_VOL	Oil (m3)	m^3	Yes
BORE_GAS_VOL	Gas (m3)	m^3	Yes
BORE_WAT_VOL	Produced Water (m3)	m^3	No
BORE_WI_VOL		-	No
FLOW__KIND		-	No
WELL_TYPE		-	No

4.1.5 Feature Statistics

In order to better understand the selected features in the dataset, several techniques were employed to understand how the data is distributed. Table 4.4 describes the selected features of the Volve dataset, whereas Table 4.5 describes the selected features for the Kyle Master dataset.

Table 4.4 : Feature Statistics for Volve Dataset

Feature	Range	Outlier Count	Mean	Standard Deviation	Variance
ON_STREAM_HRS	25 hours	715	23 hours	3 hours	9 hours
AVG_DOWNHOLE_PRESSURE	307 bar	144	242 bar	27 bar	729 hours
AVG_DOWNHOLE_TEMPERATURE	107.7 ºC	156	104 ºC	4 ºC	16 ºC
BORE_OIL_VOL	5900 m^3	283	1458 m^3	1463 m^3	2 140 369 m^3
BORE_GAS_VOL	86863 m^3	182	212937 m^3	207073 m^3	42 879 227 329 m^3
AVG_WHP_P	120 bar	44	48 bar	20 bar	400 bar
AVG_WHT_P	86 ºC	352	73 ºC	18 ºC	324 ºC

Table 4.5 : Feature Statistics for Kyle Master Dataset

Feature	Range	Outlier Count	Mean	Standard Deviation	Variance
Hours Online	1912 hours	1326	23 hours	27 hours	729 hours
Av. DHP (bar)	1122 bar	3	111 bar	39 bar	1521 bar
Av. DHT (Deg C)	245 ºC	645	94 ºC	9 ºC	81 ºC
Oil (m3)	3509 m^3	447	380 m^3	328 m^3	107.584 m^3
Gas (m3)	1 304 298 362 420 m^3	226	178 525 800 000 m^3	175 599 300 000 m^3	30 835 114 160 490 000 000 000 m^3
Av. WHP (bar)	325 bar	48	57 bar	35 bar	875 bar
Av. WHT (Deg C)	228 ºC	597	62 ºC	19 ºC	361 ºC

In Table 4.4 and Table 4.5, range denotes the range of the specified feature; more specifically, it is the difference between the lowest value up to the highest value of the feature. Outlier count is the number of outliers in the feature. Mean is the center point of the feature. It is the mathematical average of the feature. Standard deviation is the measure of how varied the feature is relative to the mean.

From Table 4.4 and Table 4.5, it can be seen that the range of values for all the selected features in the Kyle dataset is larger than the features in the Volve dataset. This denotes that the data in the Kyle dataset is dispersed compared to the Volve dataset. In addition to this, the standard deviation and variance of the features in the Kyle dataset are much larger than the features in the Volve dataset. This observation further supports the fact that the features in the Kyle dataset are more spread out than the features in the Volve dataset.

For the model's training, both these datasets will be combined. As shown in Table 4.3, the selected columns in the Volve dataset and the Kyle Master dataset have the same meaning. Therefore when combining the datasets, the columns in the Volve dataset were renamed to match the columns in the Kyle Master dataset. Table 4.6 describes the selected features of the combined dataset after forward filling is used. On the other hand, Table 4.7 describes the selected features of the combined dataset after median imputation is used.

Table 4.6 : Feature Statistics for Volve and Kyle Master Dataset after forward filling

Feature	**Range**	**Mean**	**Standard Deviation**	**Variance**
Hours Online	1912 hours	23 hours	22 hours	484 hours
Av. DHP (bar)	162 bar	102 bar	19 bar	361 bar
Av. DHT (Deg C)	344 °C	158 °C	74 °C	5476 °C
Oil (m3)	5900 m^3	801 m^3	1117 m^3	1.247.689 m^3
Gas (m3)	1164213.36 m^3	176375.2 m^3	178282.6 m^3	31 784 685 462 76m^3
Av. WHP (bar)	325 bar	49 bar	29 bar	841 bar
Av. WHT (Deg C)	228 °C	64 °C	18 °C	324 °C

Table 4.7: Feature Statistics for Volve and Kyle Master Dataset after median imputation

Feature	Range	Mean	Standard Deviation	Variance
Hours Online	1912 hours	22 hours	19 hours	361hours
Av. DHP (bar)	307 bar	182 bar	69 bar	4761 bar
Av. DHT (Deg C)	108 °C	100 °C	4 °C	16 °C
Oil (m3)	5900 m^3	733 m^3	927 m^3	859 329 m^3
Gas (m3)	13044298.36 m^3	155145.4 m^3	162085.2 m^3	2 627 161 205 904 m^3
Av. WHP (bar)	325 bar	45 bar	23 bar	529 bar
Av. WHT (Deg C)	228 °C	69 °C	17 °C	289 °C

Table 4.6 and Table 4.7 show that the mean and standard deviation of the datasets after data imputation has changed slightly. Most of the mean and standard deviation of the features in Table 4.6 are greater than the features in Table 4.7. This shows that the distribution of the dataset after forward filling is more dispersed compared to the dataset after median imputation.

4.2 Model Training

This section will discuss the steps taken by the author to train the model. Figure 4.2 describes the overall methodology for data preparation, model training, and evaluation.

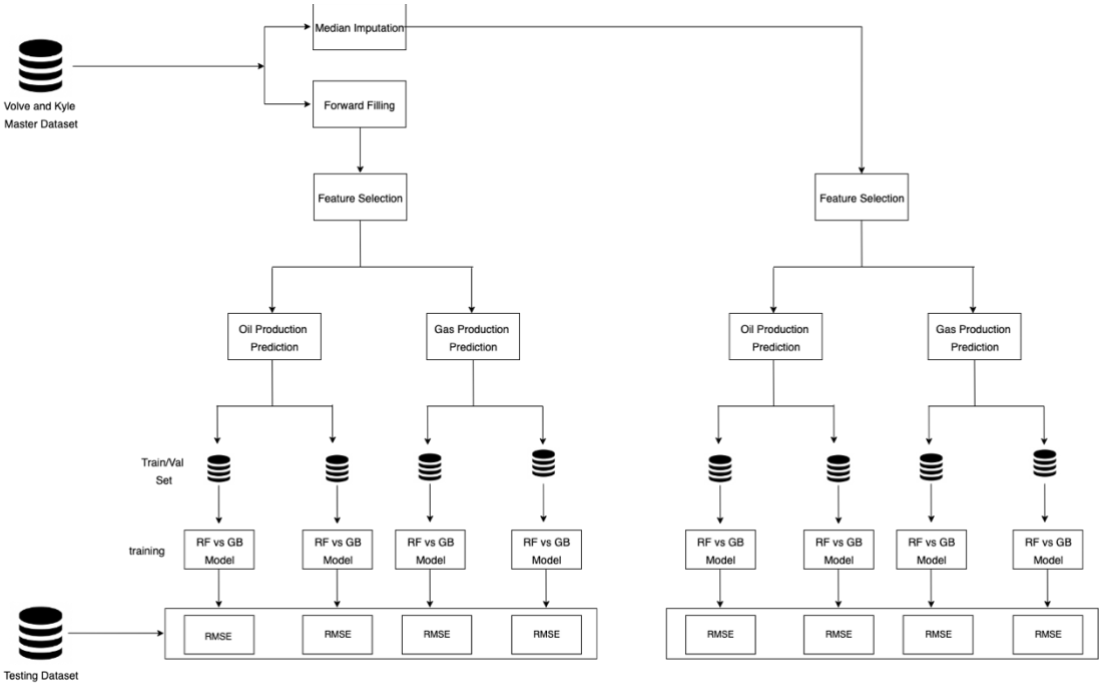


Figure 4.2: Methodology

4.2.1 Forward Filling and Median Imputation

As shown in Figure 4.2**,** two methods were used to fill in the missing values in the dataset, namely forward filling and median imputation. A gradient boosting model and random forest baseline models were trained with both forward filling and median imputation datasets. The RMSE metric was used to evaluate the performances of the model on the test dataset.

Table 4.8: Forward filling method used on gradient boosting and random forest models

	Gradient Boosting	Random Forest
	RMSE (m^3)	
Oil prediction	175	261
Gas prediction with oil predicted value	95601	72488
Gas prediction without oil predicted value	53121	76189

Table 4.9: Median imputation method used on gradient boosting and random forest models

	Gradient Boosting	Random Forest
	RMSE (m^3)	
Oil prediction	972	461
Gas prediction with oil predicted value	86441	83969
Gas prediction without oil predicted value	67436	82702

Table 4.8 and Table 4.9 showed that the models performed worse if the values of the predicted oil production were included in the training. Therefore, it would be better not to include the predicted oil production value for model training. In terms of oil prediction and gas prediction without the oil predicted values, Table 4.8 and Table 4.9 showed that the forward filling method is better. This is because the forward filling method gave a lower RMSE. A high RMSE value indicates a poor model. Therefore, forward filling is the better method for data imputation.

In the forward filling method, Table 4.8 showed that the gradient boosting model performed better than the random forest model. The gradient boosting model generally gave a lower RMSE value compared to the random forest model. Therefore, this makes the gradient boosting model a better choice compared to the random forest model.

REFERENCES

[1] M. S. Vassiliou, Historical dictionary of the petroleum industry, 2018.

[2] International Association of Oil & Gas Producers, "Oil and gas in Everyday Life," International Association of Oil & Gas Producers, [Online]. Available: https://www.iogp.org/oil-natgas-in-everyday-life/. .

[3] W. P. Council, "Why are oil and gas important?," [Online]. Available: https://www.world-petroleum.org/edu/221-why-are-oil-and-gas-important#:~:text=Oil%20is%20one%20of%20the%20most%20important%20raw,about%20two%20million%20tonnes%20of%20oil%20and%20gas. . [Accessed March 2022].

[4] R. Ranggasari, "Oil and gas reserves potential in eastern Indonesia reaches 9.8bn barrels," Tempo, [Online]. Available: https://en.tempo.co/read/1536679/oil-and-gas-reserves-potential-in-eastern-indonesia-reaches-9-8bn-barrels#:~:text=Overall%2C%20the%20Energy%20Ministry%20recorded%20there%20are%2070,2.44%20billion%20barrels%20and%20gas%20of%2043.6%20TCF. .

[5] Indonesia Investment, "Crude Oil Indonesia," [Online]. Available: https://www.indonesia-investments.com/business/commodities/crude-oil/item267..

[6] W. Kenton, "Virtual Data Room (VDR)," 23 June 2021. [Online]. Available: https://www.investopedia.com/terms/v/virtual-data-room-vdr.asp#:~:text=Virtual%20Data%20Rooms%2C%20or%20VDRs%2C%20exist%20as%20a,joint%20venture%20that%20requires%20access%20to%20shared%20data.. [Accessed 19 03 2022].

[7] Lynx, "License Pricing - Lynx Information Systems," Lynx Information System, [Online]. Available: http://www.lynxinfo.co.uk/download-pricing.html.

[8] Intviewer, "Intviewer - Fast Geoscience Visualization, Analysis & QC,," Intviewer, 02 August 2021. [Online]. Available: https://www.int.com/products/intviewer/#:~:text=INTViewer%20is%20a%20platform%20and%20application%20that%20allows,to%20a%20desktop%20or%20remotely%20via%20the%20cloud..

[9] INTViewer, "INTViewer. Geoscience Analysis and QC, Simplied.," INTViewer, [Online]. Available: https://www.int.com/products/intviewer/.

[10] IBM, "Data Science," IBM, 15 May 2020. [Online]. Available: https://www.ibm.com/cloud/learn/data-science-introduction.

[11] G. Gurung, R. Shah and D. P. Jaiswal, "Software development life cycle models-A comparative study," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 33-27, 2020.

[12] A. Mishra and D. Dubey, "A Comparative Study of Different Software Development Life Cycle Models in Different Scenarios," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 1, no. 5, pp. 1-6, 2013.

[13] J. Shah, "A Comparative Study of Software Development Life Cycle Models".

[14] P. Pedamkar, "What is SDLC: Different phases and models of SDLC," *EDUCBA*, 2022.

[15] "Principles behind the Agile Manifesto," Agile, [Online]. Available: https://agilemanifesto.org/principles.html..

[16] Harrison, Bob, "The data room," 2020, pp. 21-26.

[17] Schlumberger, "Secure, Remote Access to Field Datasets Enables Potential Investors to Complete Asset Evaluations," Schlumberger, [Online]. Available: https://www.slb.com/resource-library/case-study/dss/delfi-virtual-data-room-generic-asia-pacific-cs.

[18] A. El-Banbi, A. Ahmed and E.-M. Ahmed , "Black Oils," in *PVT Property Correlations*, Elsevier, 2018, p. 147–182.

[19] S. Mokhatab, W. A. Poe and J. Y. Mak, "Natural Gas Fundamentals," in *Handbook of Natural Gas Transmission and Processing*, Elsevier, 2019.

[20] A. El-Banbi, A. Ahmed and E.-M. Ahmed, "Dry Gases," in *PVT Property Correlations*, Elsevier, 2018.

[21] T. Ahmed, "Reservoir-Fluid Properties," in *Reservoir-Fluid Properties* , Elsevier, 2010, pp. 29-135.

[22] I. Fetoui, "Hydrocarbon Phase Behavior," [Online]. Available: https://production-technology.org/category/pvt/.

[23] P. Z. a. K. H. C. Janiesch, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685-695, 2021.

[24] IBM Cloud Education, "What is machine learning," IBM, 15 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning..

[25] I. Sydorenko, "What is dataset in Machine Learning," High quality data annotation for Machine Learning, 5 April 2021. [Online]. Available: https://labelyourdata.com/articles/what-is-dataset-in-machine-learning.

[26] W. X. P. C. M. C. a. S. D. Y. Gong, "Supervised Learning," *Machine learning techniques for multimedia*, p. 21–49., 2008.

[27] IBM Cloud Education, "What is Supervised Learning," IBM, 2022. [Online]. Available: https://www.ibm.com/cloud/learn/supervised-learning. .

[28] A. M. J. A. V. M. A. A. L. a. A. A. S. A. R. van Loon, "Understanding supervised, unsupervised, and reinforcement learning," *Big Data Made Simple*, 2019.

[29] D. N. Dimid, "Unsupervised learning algorithms cheat sheet,," 17 February 2022. [Online]. Available: https://towardsdatascience.com/unsupervised-learning-algorithms-cheat-sheet-d391a39de44a. .

[30] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *Cambridge, MA*, 2020.

[31] z_ai, "Deep Learning for NLP: ANNs, RNNs and LSTMs explained!," 8 July 2019. [Online]. Available: https://towardsdatascience.com/deep-learning-for-nlp-anns-rnns-and-lstms-explained-95866c1db2e4.

[32] aravindpai, "CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning download Share," 17 February 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/.

[33] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep Learning is not all you need," *Information Fusion,* vol. 81, pp. 84-90, 2022.

[34] Elzain, Hussam Eldin, Chung, Sang Yong, Senapathi, Venkatramanan, Sekar, Selvam, Lee, Seung Yeop, Roy, Priyadarsi D., Hassan, Amjed and Sabarathinam, Chidambaram, "Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination," *Ecotoxicology and Environmental Safety,* vol. 229, pp. 61-113, 2022.

[35] Scikit, "ScikitLearn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

[36] A. Kumar, "Introduction to the Gradient Boosting Algorithm," 20 May 2020. [Online]. Available: https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b.

[37] S. Rosenthal, "Data Imputation," in *The International Encyclopedia of Communication Research Methods*, Wiley, 2017, pp. 1-12.

[38] A. Swalin, "How to handle missing data," Medium, 19 March 2018. [Online]. Available: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4..

[39] H. Kang, "The prevention and handling of the missing data," in *Korean Journal of Anesthesiology,* vol. 64, 2013, p. 402.

[40] W. Badr, "“6 different ways to compensate for missing data (data imputation with examples)," 12 January 2019. [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779..

[41] K. N, "Part-1 : Data Preparation Made Easy with python!!," Medium, 09 May 2020. [Online]. Available: https://medium.com/analytics-vidhya/part-1-data-preparation-made-easy-with-python-e2c024402327..

[42] Á. Fernández, J. R. Dorronsoro and J. Bella, "Supervised outlier detection for classification and regression," *Neurocomputing,* vol. 486, pp. 77-92, 2022.

[43] C. M. Salgado, C. Azevedo, H. Proença and S. M. Vieira, "Noise Versus Outliers," in *Secondary Analysis of Electronic Health Records*, Cham, Springer International Publishing, 2016, pp. 163-183.

[44] F. Malik, "Understanding value of correlations in data science projects," Medium, 10 June 2019. [Online]. Available: https://medium.com/fintechexplained/did-you-know-the-importance-of-finding-correlations-in-data-science-1fa3943debc2#:~:text=Correlation%20is%20a%20statistical%20measure.%20Correlation%20explains%20how,%28variables%29%20can%20be%20positively%20correlated%.

[45] BBC Bitesize, "Types of correlation - scattergraphs - national 4 application of Maths Revision," BBC News, [Online]. Available: https://www.bbc.co.uk/bitesize/guides/zmt9q6f/revision/2..

[46] Nettleton, David, "Selection of Variables and Factor Derivation," in *Commercial Data Mining*, Elsevier, 2014, pp. 79-104.

[47] "Spearman correlation coefficient: Definition, formula and calculation with example," QuestionPro, 15 January 2020. [Online]. Available: https://www.questionpro.com/blog/spearmans-rank-coefficient-of-correlation/. .

[48] Swapnilbobe, "Spearman's correlation," 13 April 2021. [Online]. Available: https://medium.com/analytics-vidhya/spearmans-correlation-f34c094d99d8#:~:text=Here%2C%20we%20are%20calculating%20spearman%E2%80%99s%20correlation%20using%20the,of%20relationship%20between%20ranks%20of%20two%20individual%20features. .

[49] V. Kumar and S. Minz, "Feature Selection: A literature Review," *Smart Computing Review*, vol. 4, pp. 1-19, 2014.

[50] K. Menon, "Feature selection in machine learning," Simplilearn, 16 September 2021. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#what_is_feature_selection..

[51] A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?," 8 December 2020. [Online]. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

[52] Zhao, Yunwei and Wan, Xin, "The Design of Embedded Web System based on REST Architecture," in *IEEE*, 2019.

[53] D. Bryant, "GraphQL-ultimate-guide," [Online]. Available: https://www.infoq.com/articles/GraphQL-ultimate-guide/.

[54] BBVA API Market, "REST API: What is it, and what are its advantages in project development?," 2016. [Online]. Available: https://www.bbvaapimarket.com/en/api-world/rest-api-what-it-and-what-are-its-advantages-project-development/.

[55] C. Xie, L. Chao, Y. Qin, J. Cao and Y. Li, "Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway," *AIP Advances*, vol. 10, no. 11, 2020.

[56] C. Teng, D. Garreau and U. v. Luxburg, "When do random forests fail?".

[57] R. Sharma, "Using Facebook Prophet for Forecasting Natural Gas Production," Medium, 13 March 2021. [Online]. Available: https://medium.com/mlearning-ai/forecast-using-prophet-canadian-natural-gas-production-dataset-b1f9c57548d8.

[58] S. Goled, "Why Are People Bashing Facebook Prophet," 18 October 2021. [Online]. Available: https://analyticsindiamag.com/why-are-people-bashing-facebook-prophet/.

[59] L. Menculini, A. Marini, M. Proietti, A. Garinei, A. Bozza, C. Moretti and M. Marconi, "Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices," *Forecasting*, vol. 3, no. 3, pp. 644-662, 2021.

[60] J. Chahar, "Prediction of Oil Production by applying Machine Learning on Volve Field Production Data.," 17 December 2020. [Online]. Available: https://www.linkedin.com/pulse/prediction-oil-production-applying-machine-learning-volve-chahar/.

[61] "Advantages and Disadvantages of Linear Regression," [Online]. Available: https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/.

[62] A. Pant, "Introduction to Linear Regression and Polynomial Regression," 13 January 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb#:~:text=Disadvantages%20of%20using%20Polynomial%20Regression%20The%20presence%20of,analysis.%20These%20are%20too%20sensitive%20to%20the%20outliers..

[63] M. Taylor, "Machine Learning in the Oil and Gas Industry," 21 January 2021. [Online]. Available: https://newengineer.com/blog/machine-learning-in-the-oil-and-gas-industry-1507752.

[64] EDUCBA, "Difference Between Random forest vs Gradient boosting," [Online]. Available: https://www.educba.com/random-forest-vs-gradient-boosting/.

[65] D. P. Nolan, "Overview of Oil, Gas, and Petrochemical Facilities," in *Handbook of Fire and Explosion Protection Engineering Principles for Oil, Gas, Chemical, and Related Facilities*, Elsevier, 2019, pp. 33-50.

APPENDICES

Appendix A

I. Feature correlation in Volve and Kyle Master datasets when forward filling is used

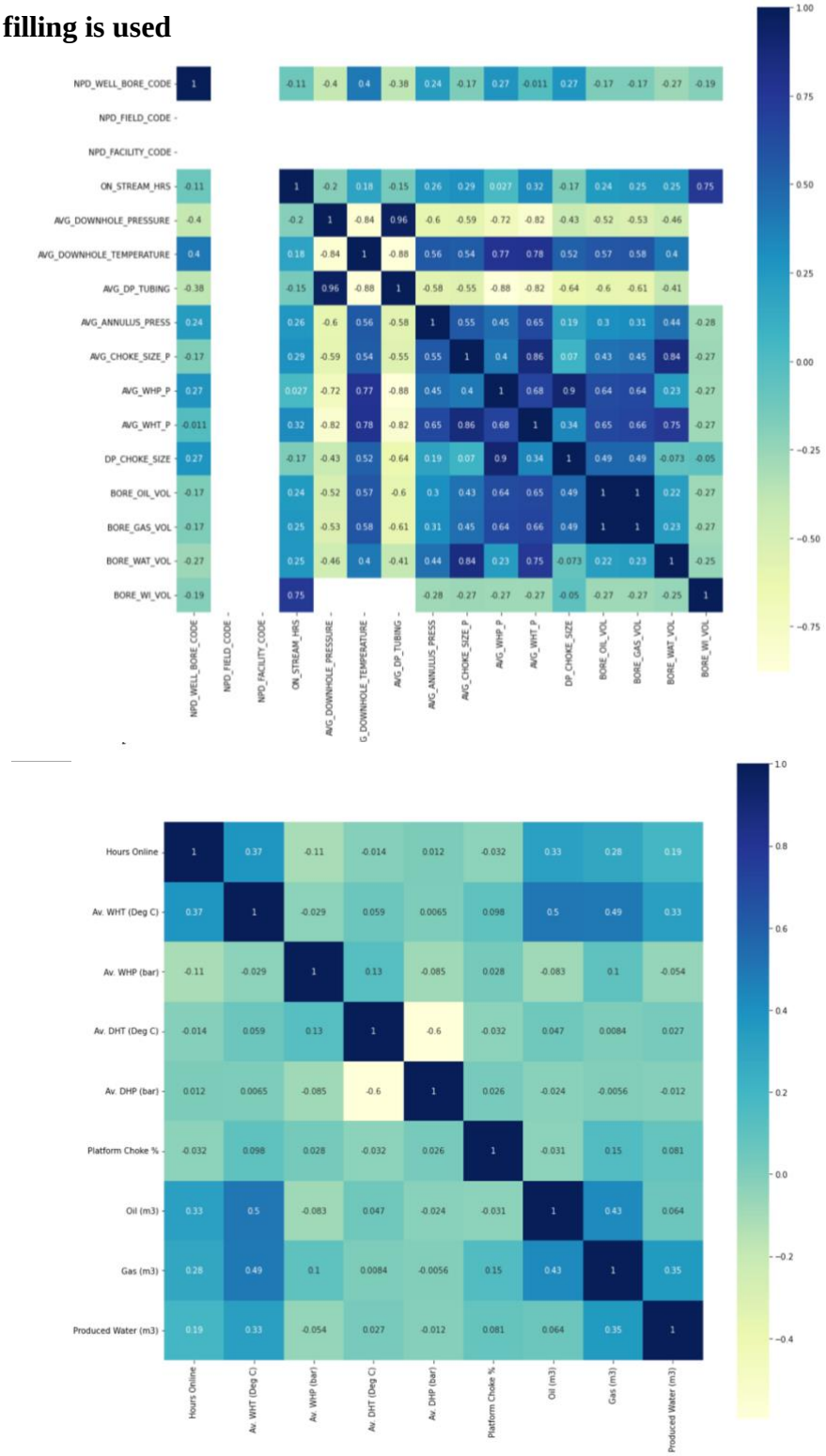


Figure A.1 : Feature correlation for Volve and Kyle Master datasets with forward filling

II. Feature correlation in Volve and Kyle Master datasets when median imputation is used

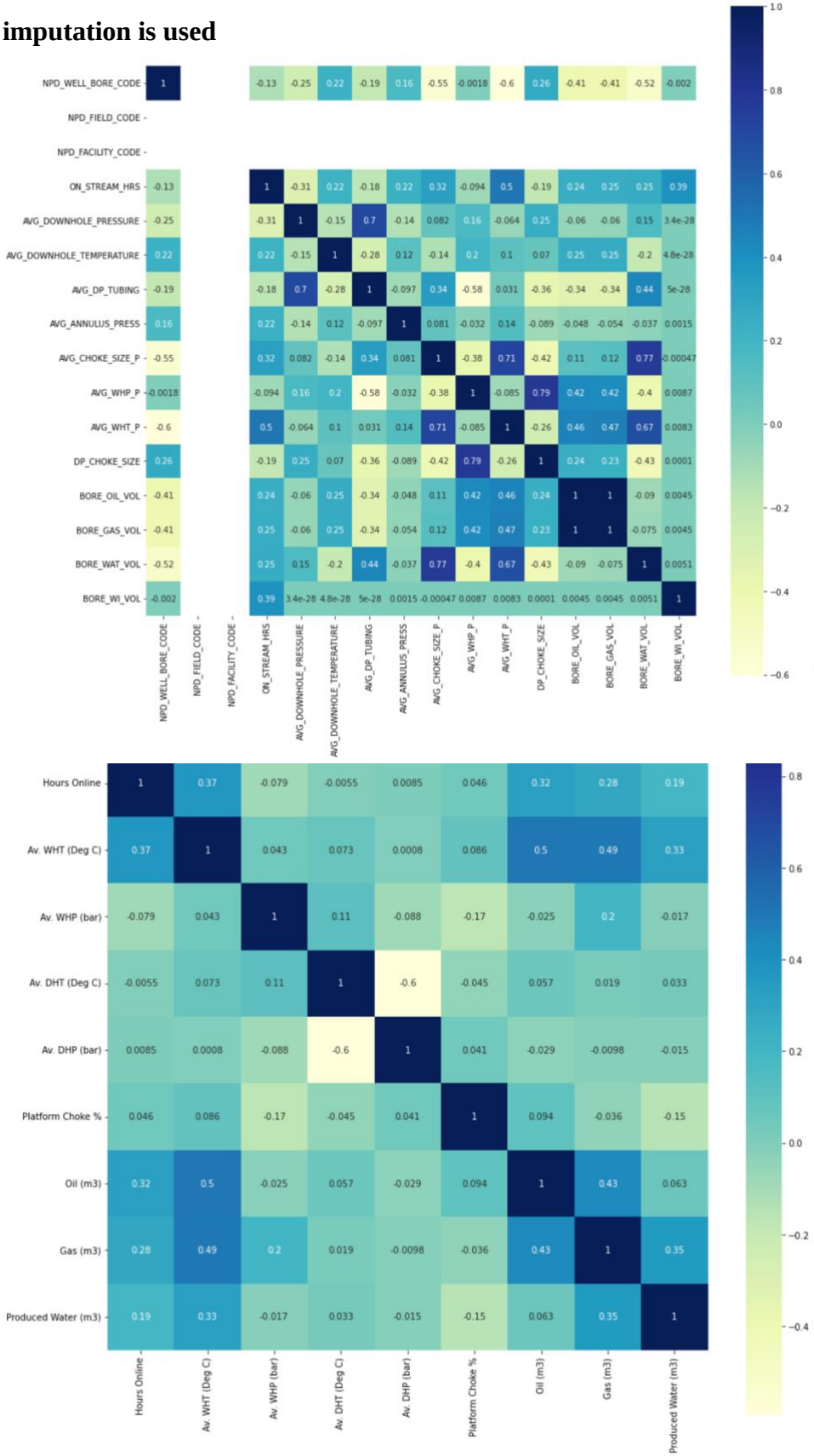


Figure A.2 : Feature correlation for Volve and Kyle Master datasets with median imputation

III. Feature statistics in Volve dataset

ON_STREAM_HRS

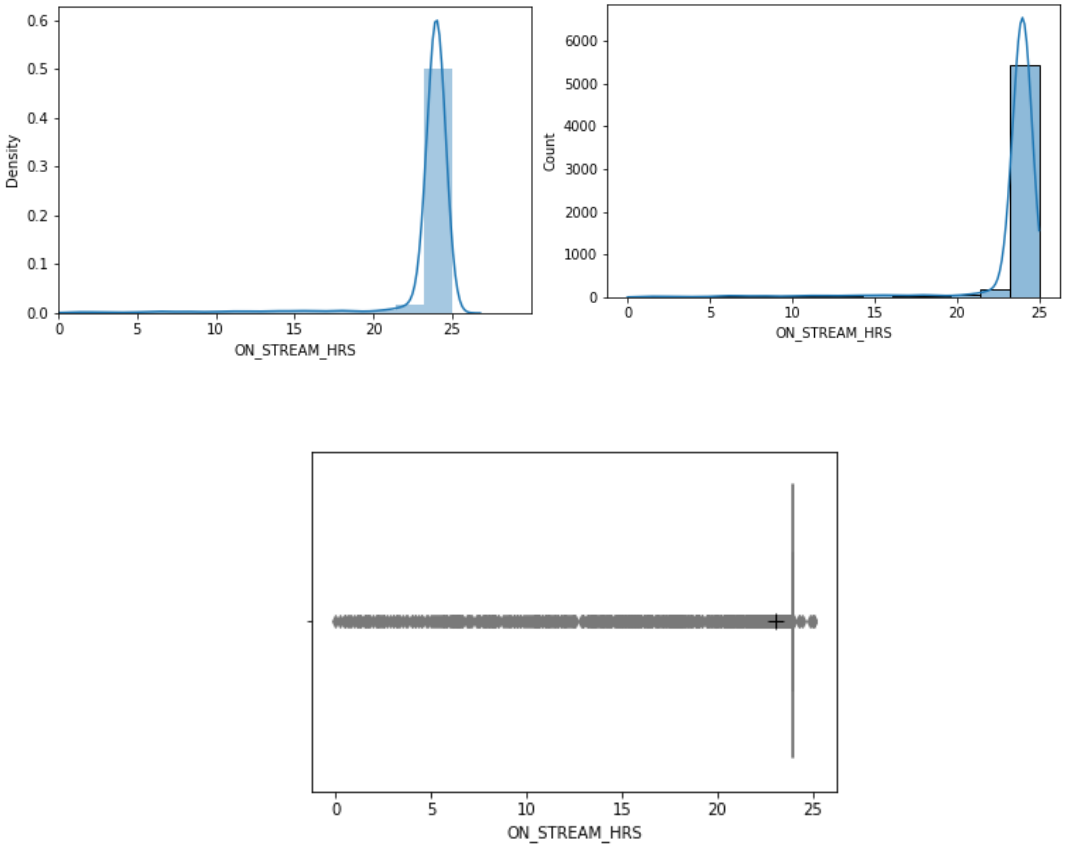
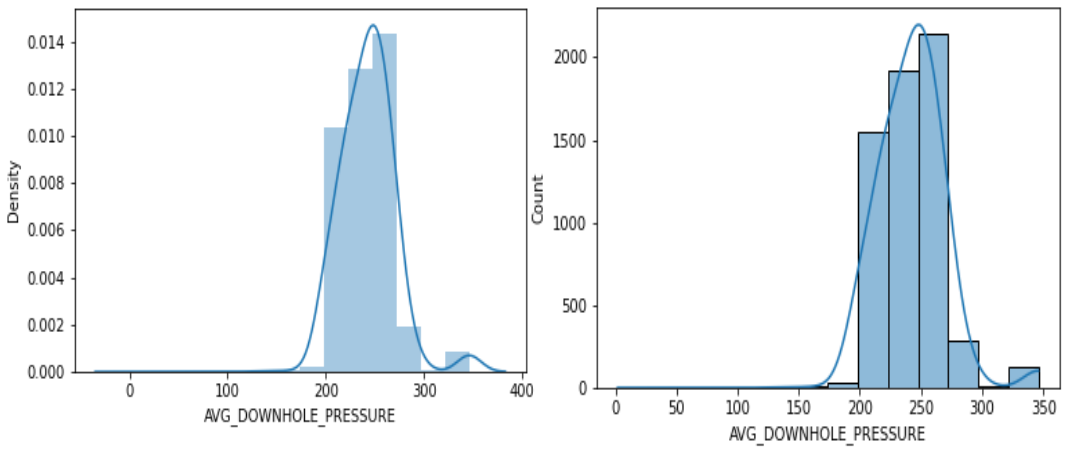


Figure A.3: Kernel Density Estimation plot, histogram, and boxplot for ON_STREAM_HRS

AVG_DOWNHOLE_PRESSURE



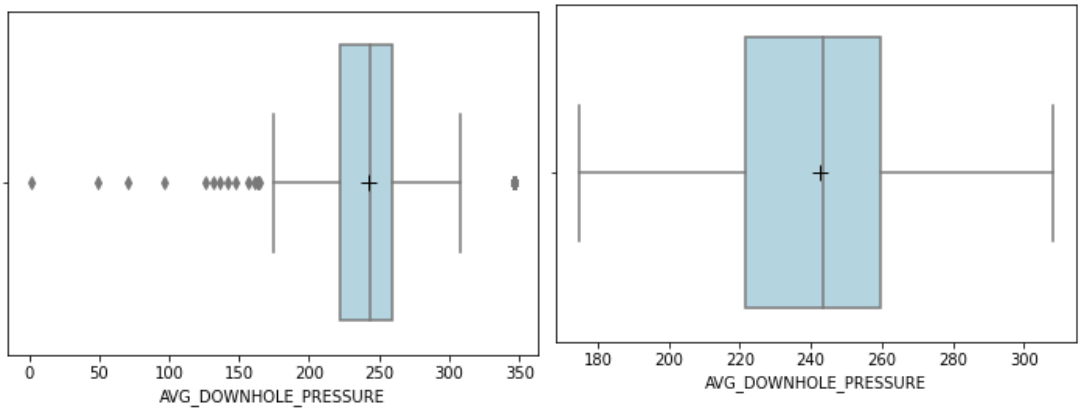
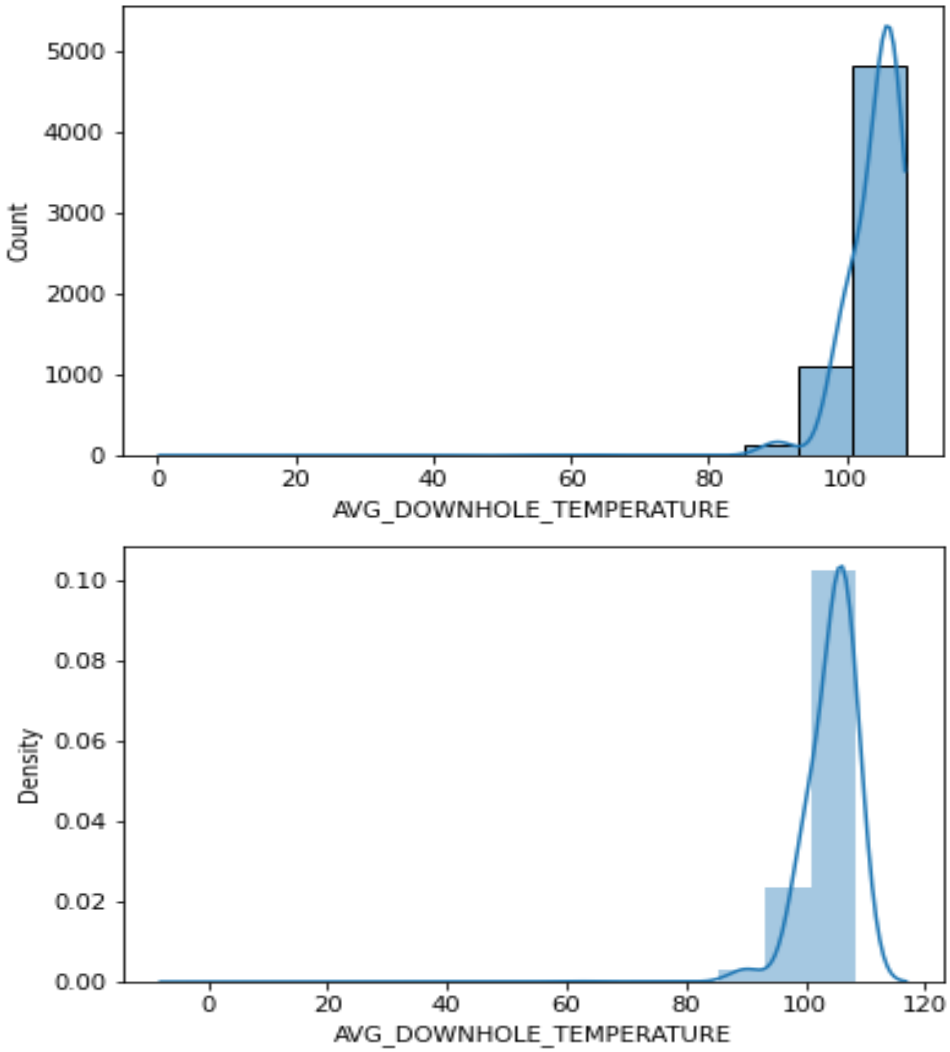


Figure A.4 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for AVG_DOWNHOLE_PRESSURE

AVG_DOWNHOLE_TEMPERATURE



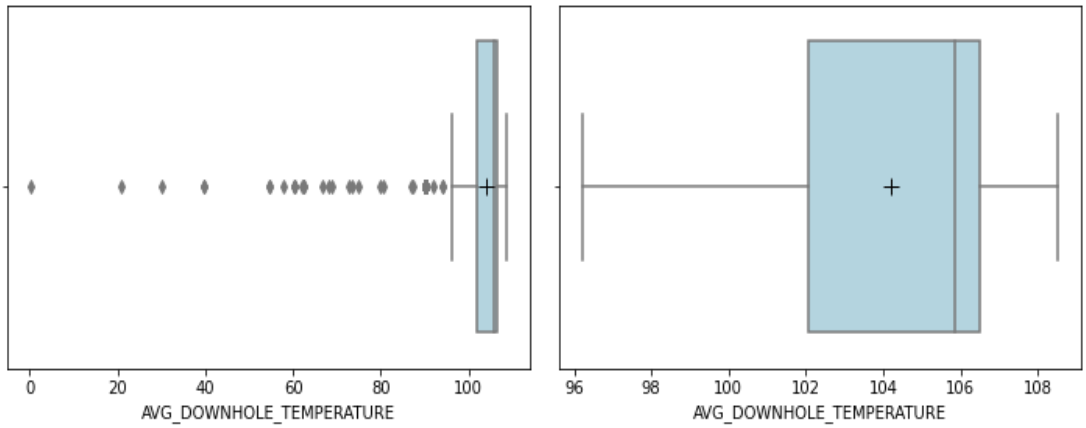
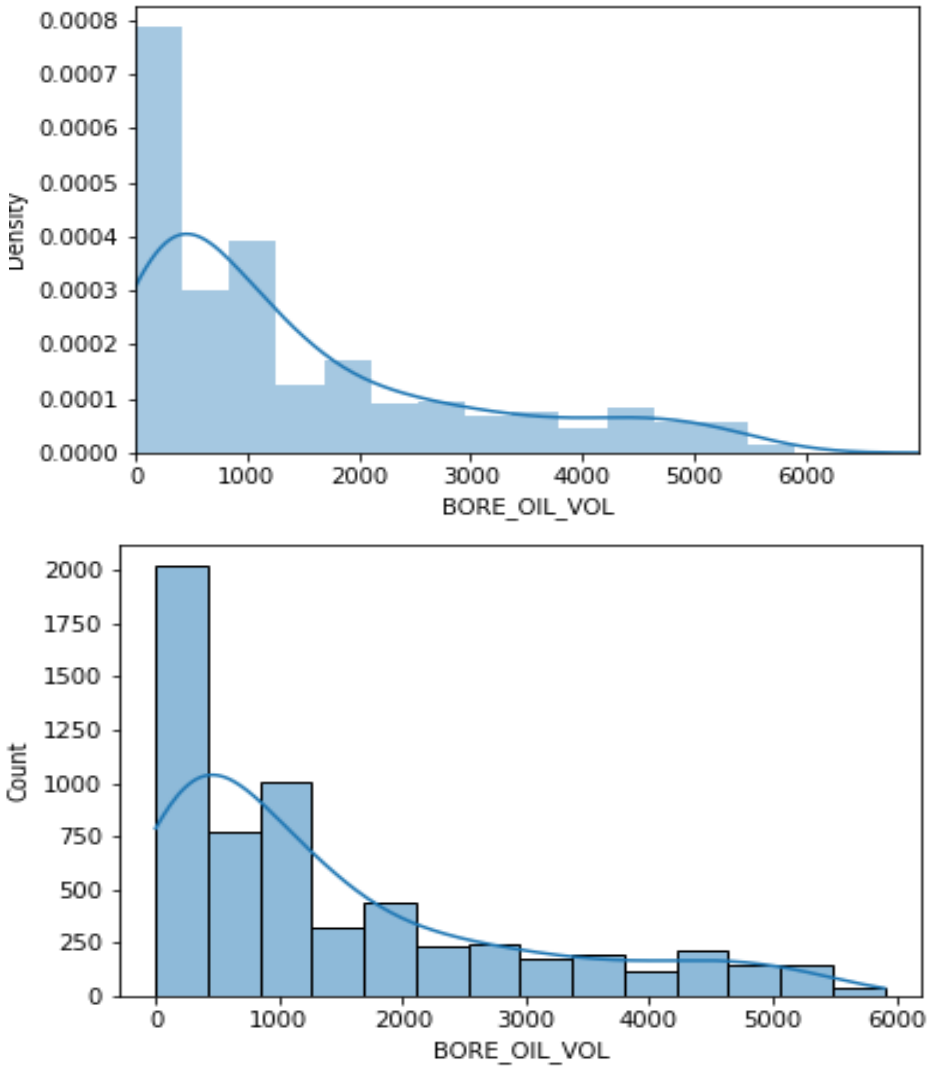


Figure A.5 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for AVG_DOWNHOLE_TEMPERATURE

BORE_OIL_VOL



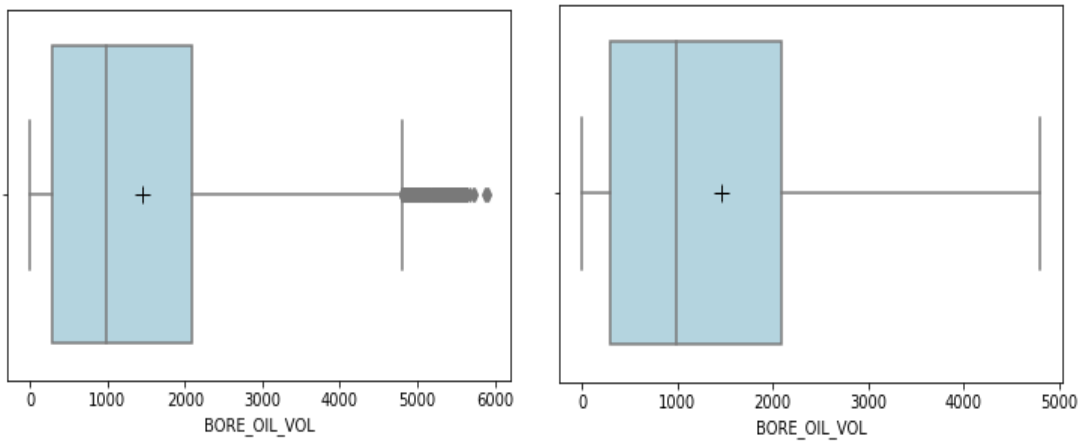
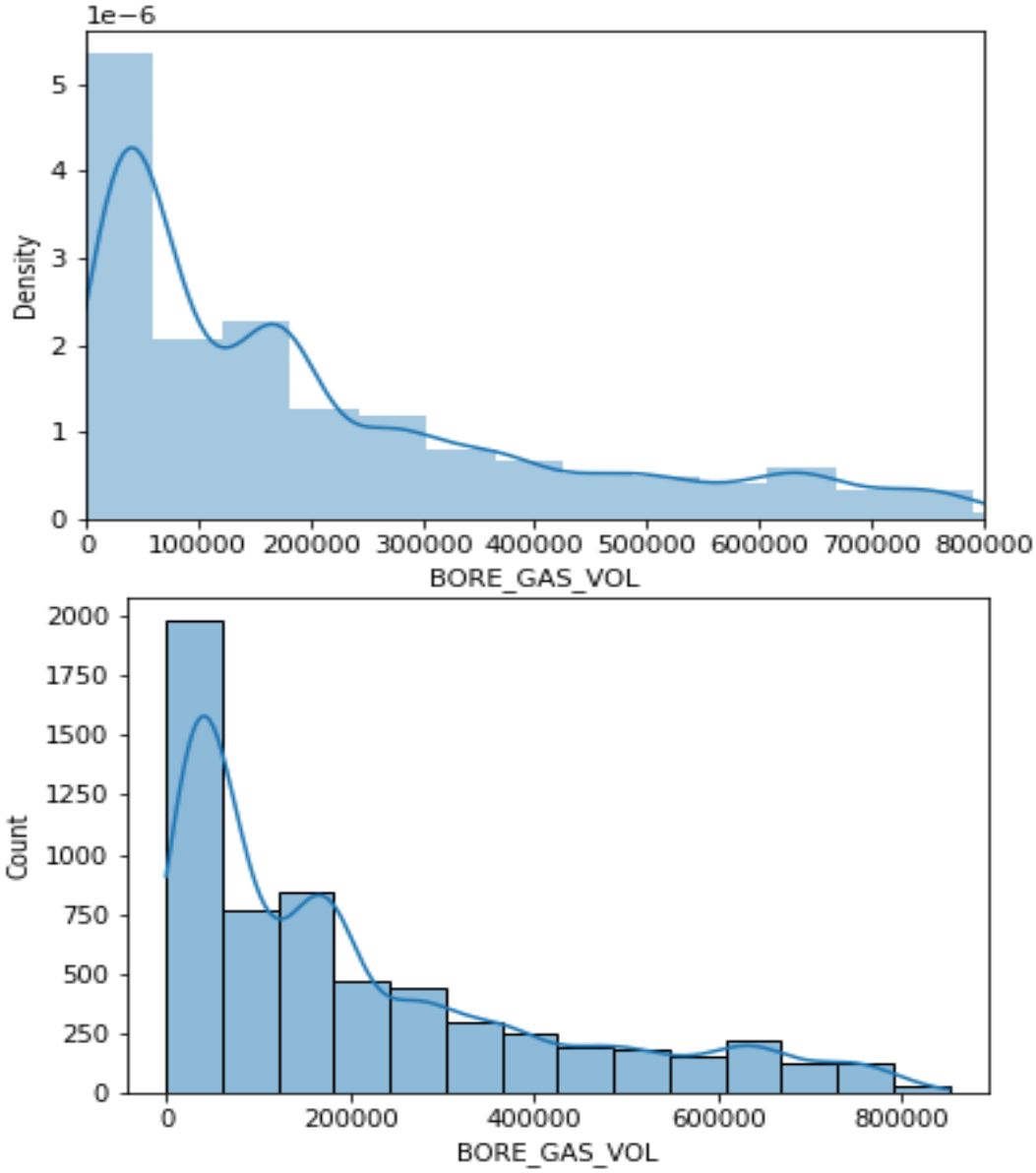


Figure A.6 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for BORE_OIL_VOL

BORE_GAS_VOL



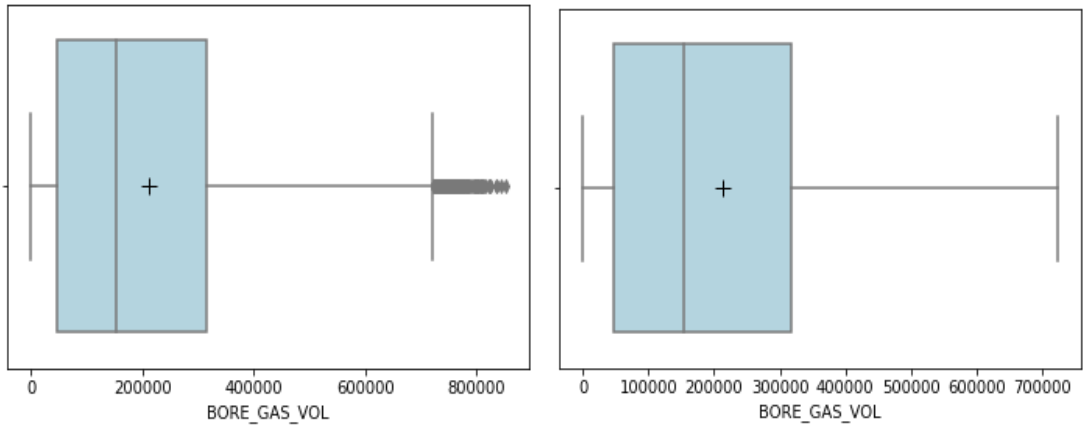
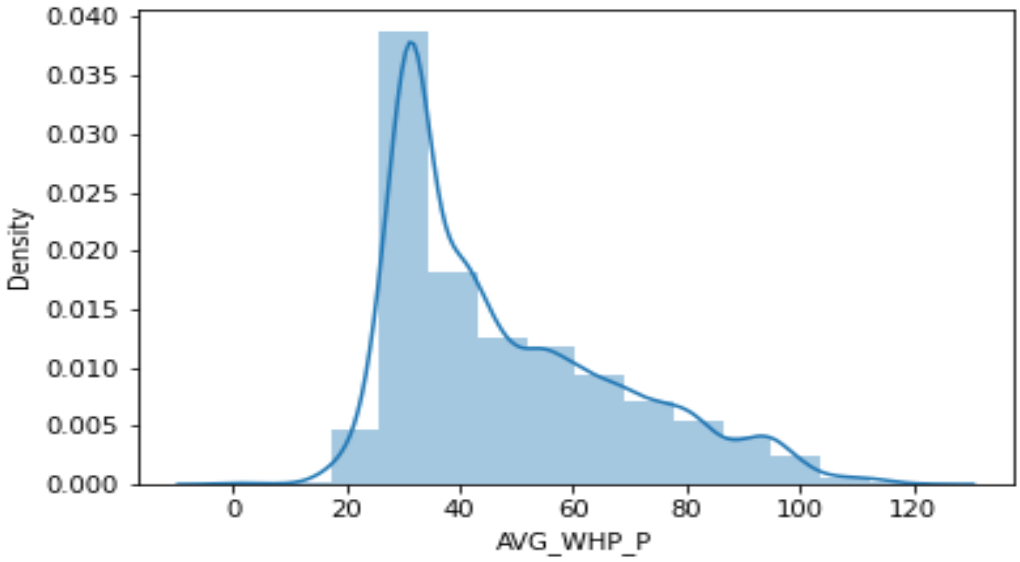
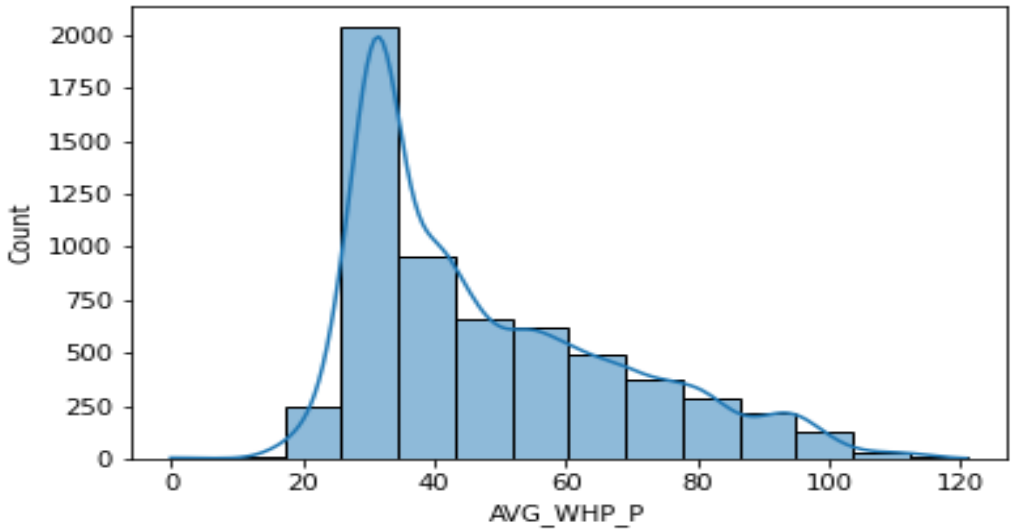


Figure A.7 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for BORE_GAS_VOL

AVG_WHP_P





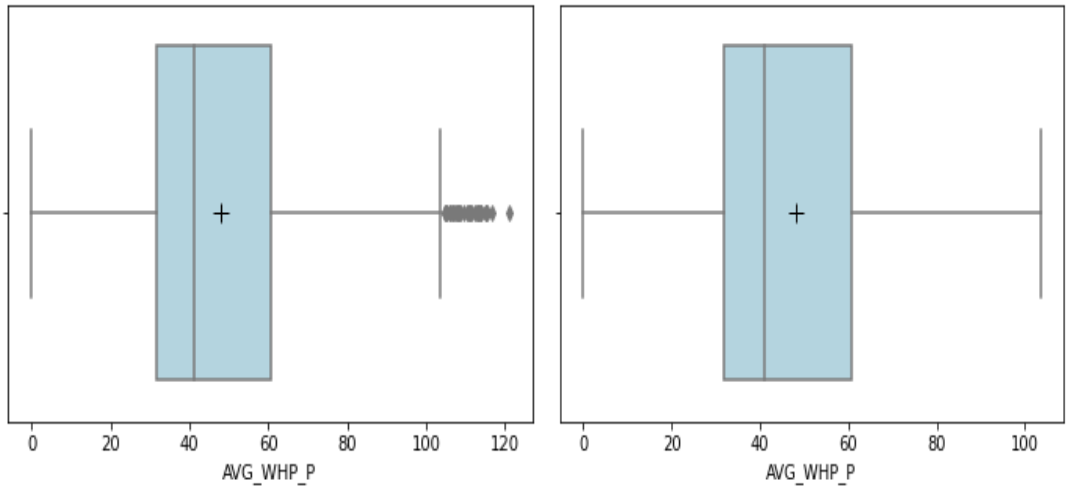
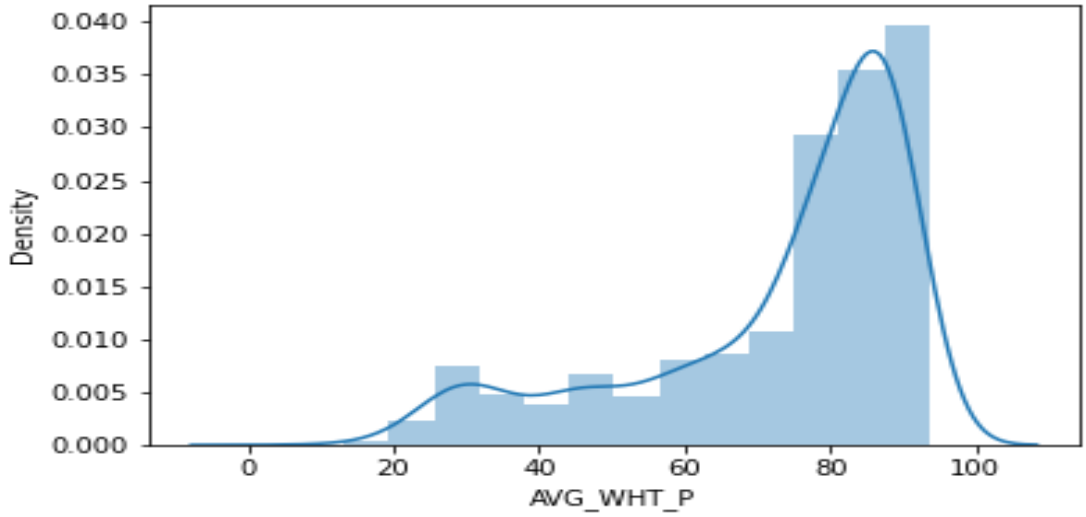
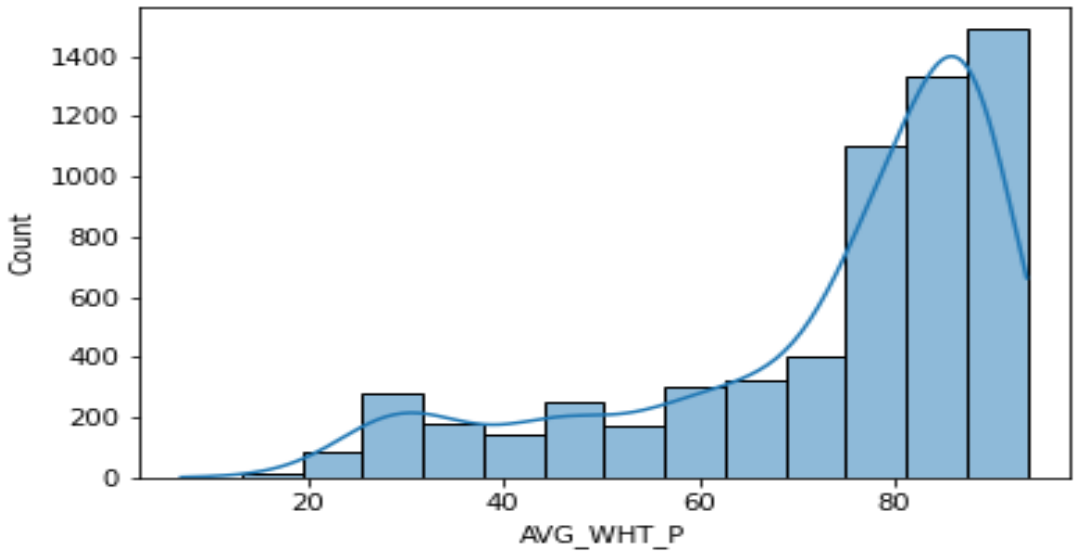


Figure A.8 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for AVG_WHP_P

AVG_WHT_P





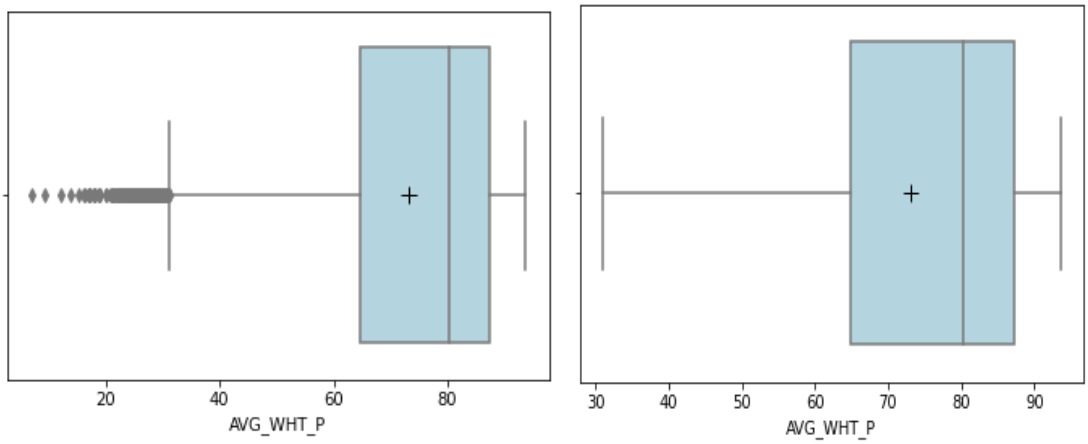
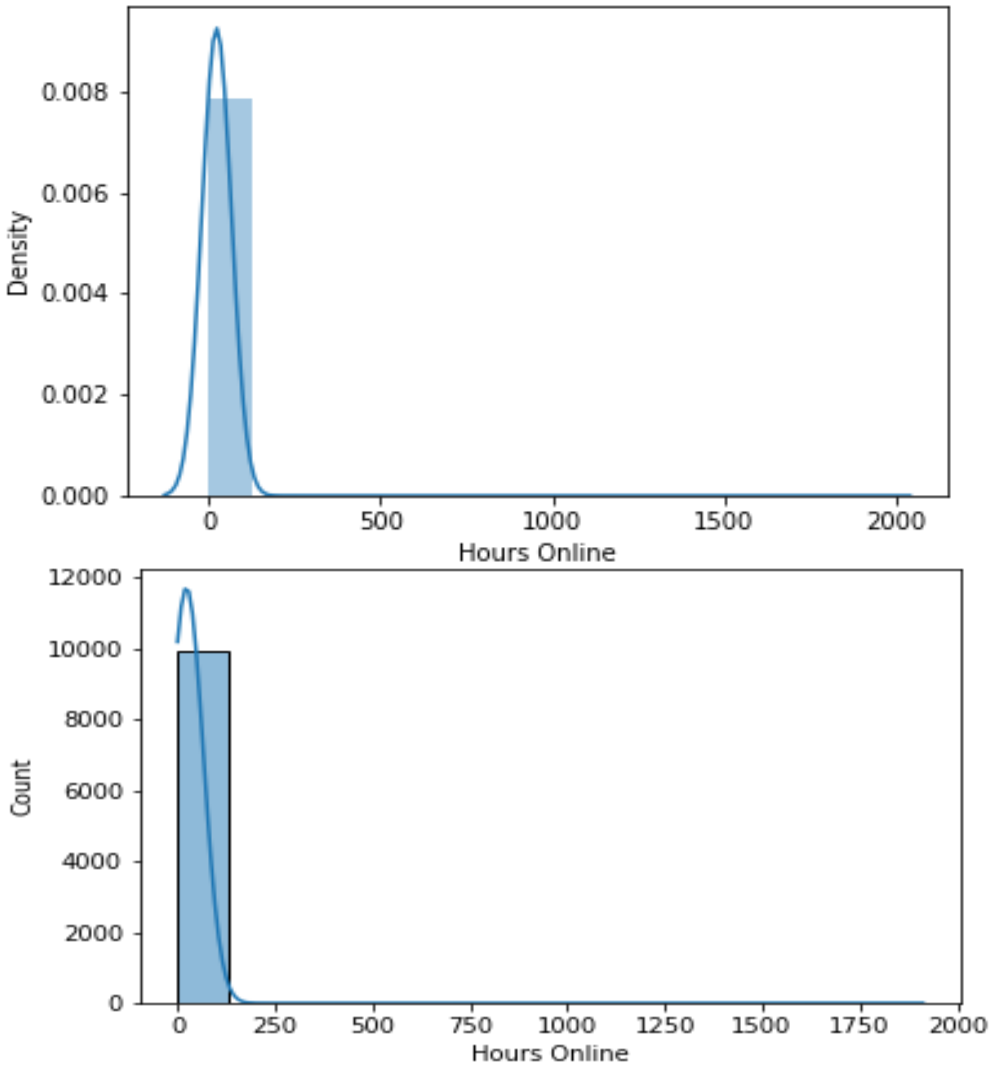


Figure A.9 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for AVG_WHT_P

IV. Feature statistics in Kyle Master dataset

Hours Online



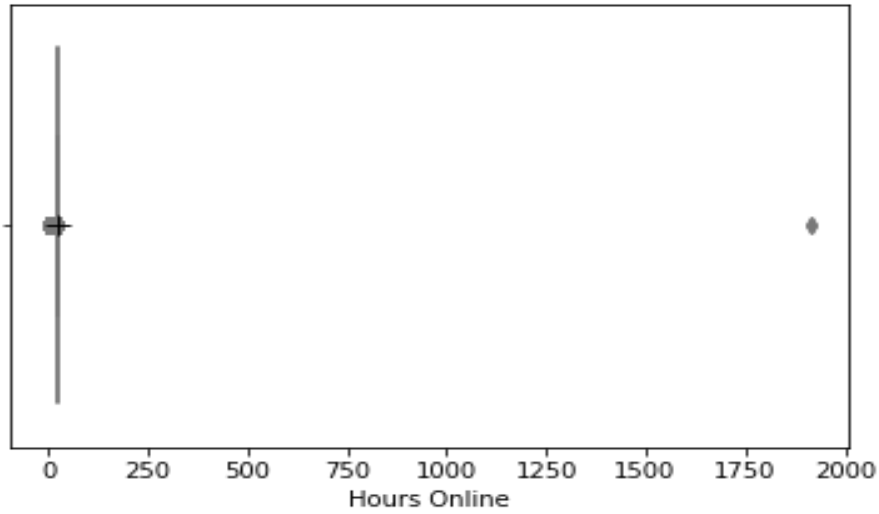
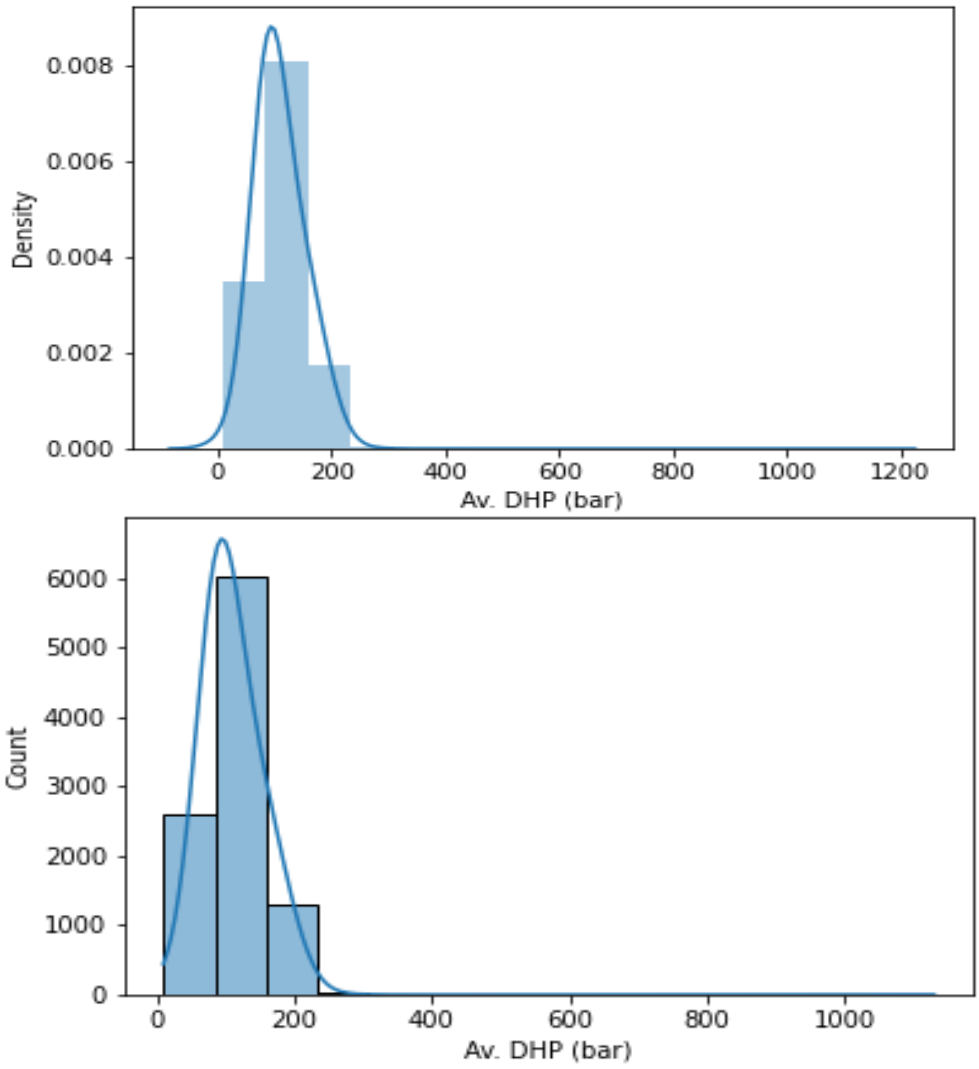


Figure A.10 : Kernel Density Estimation plot, histogram, and boxplot for Hours Online

Av. DHP (bar)



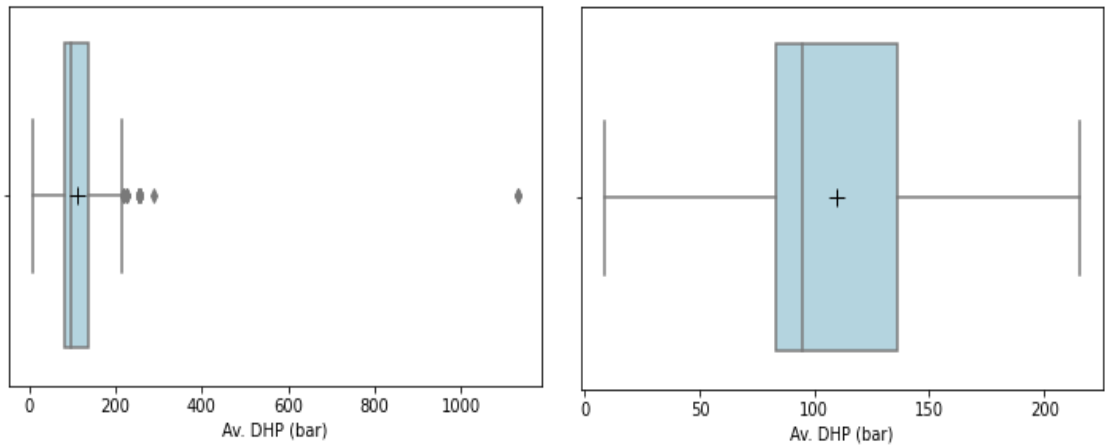
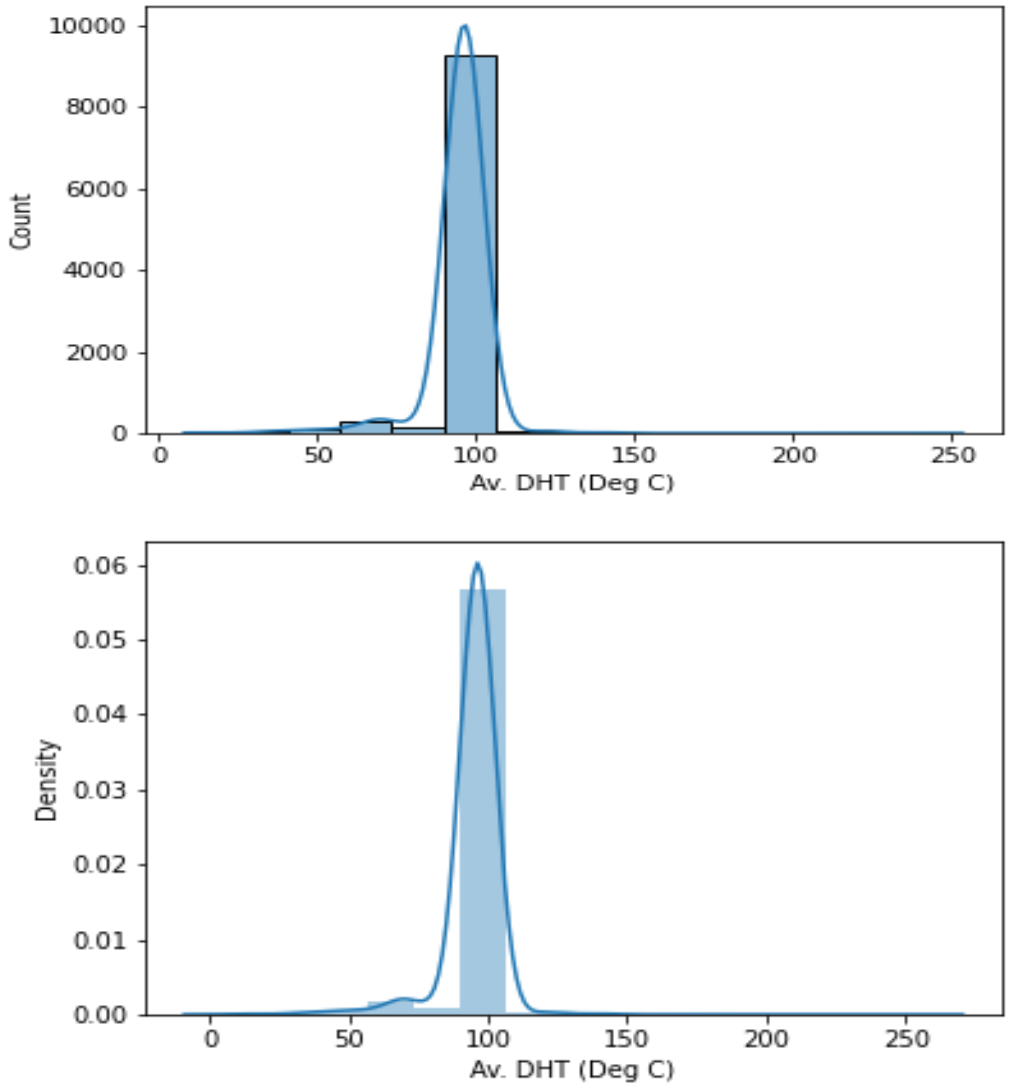


Figure A.11 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Av. DHP (bar)

Av. DHT (Deg C)



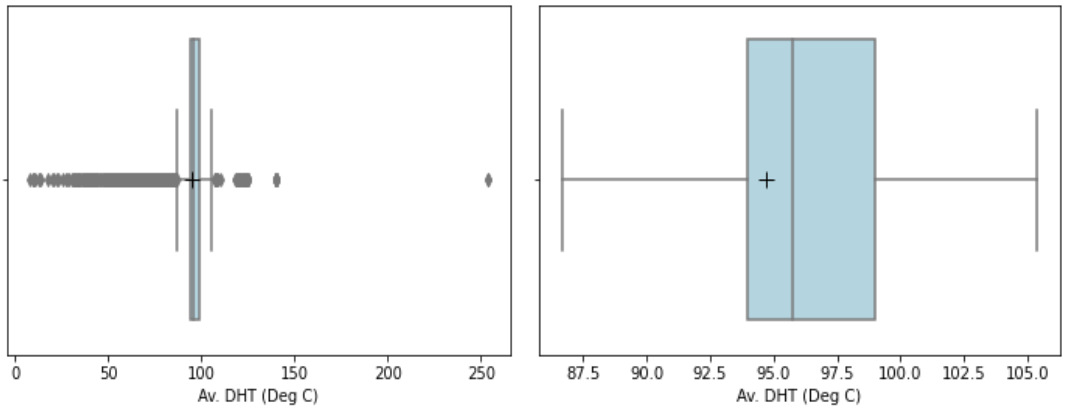
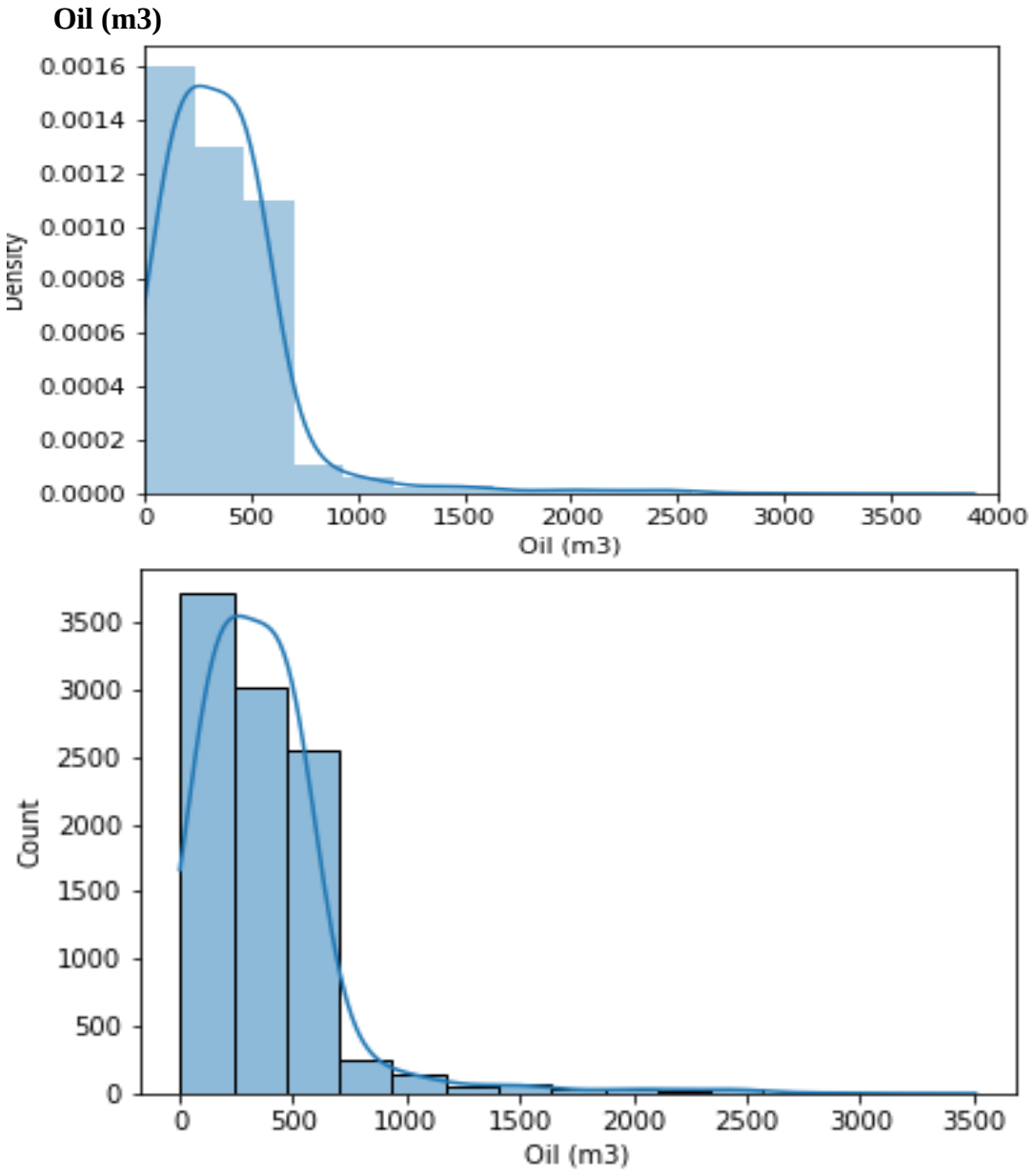


Figure A.12 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Av. DHT (Deg C)



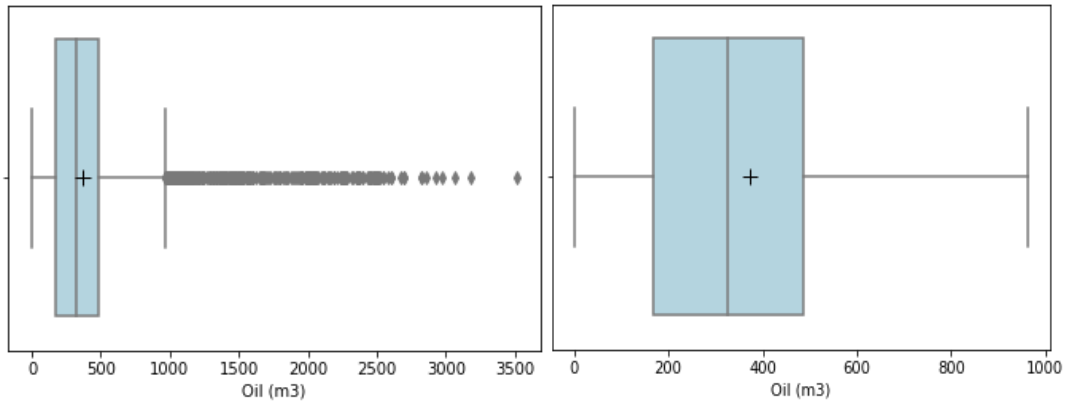
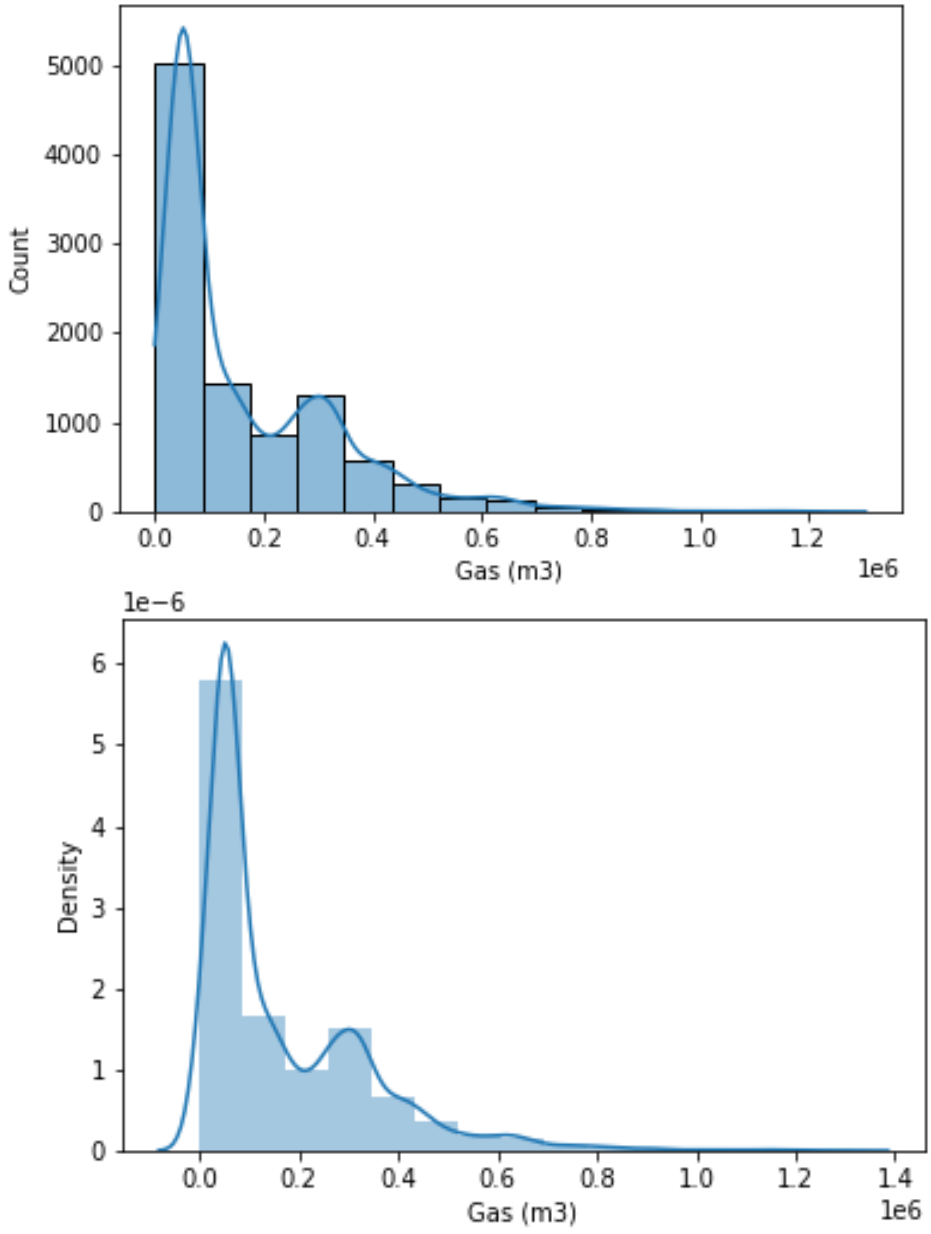


Figure A.13 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Oil (m3)

Gas (m3)



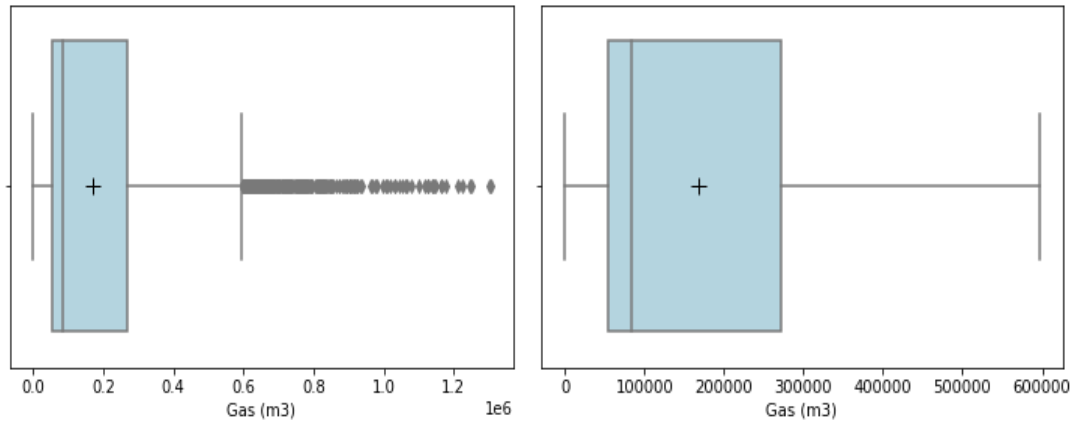
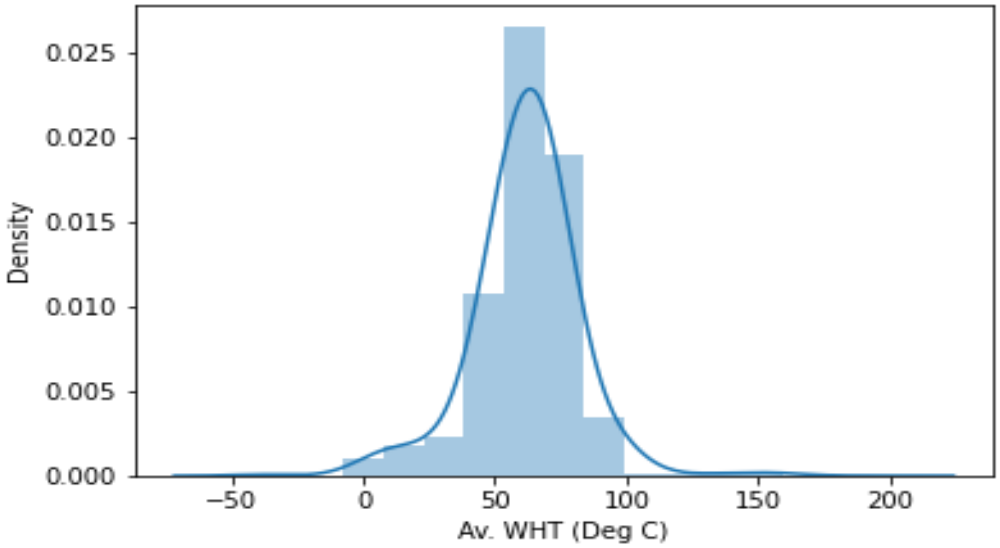
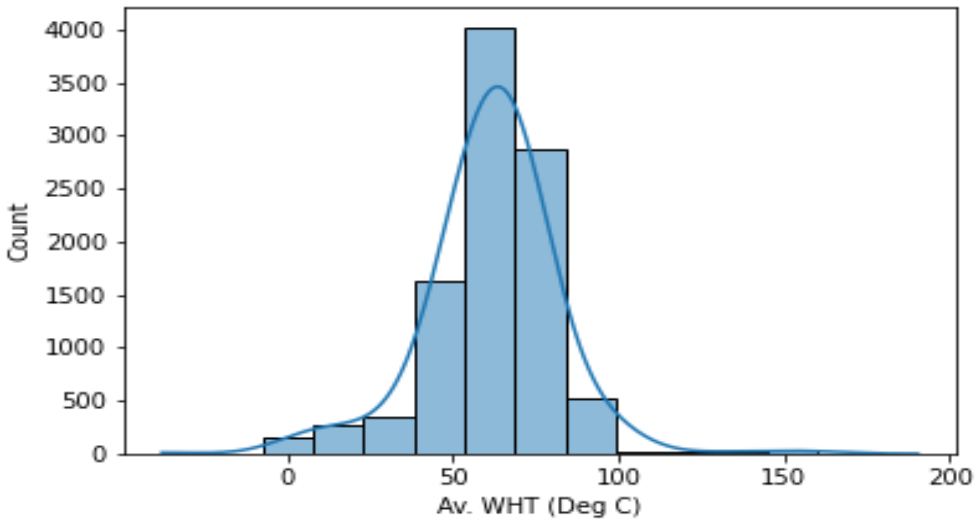


Figure A.14 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Gas (m3)

Av. WHT (Deg C)





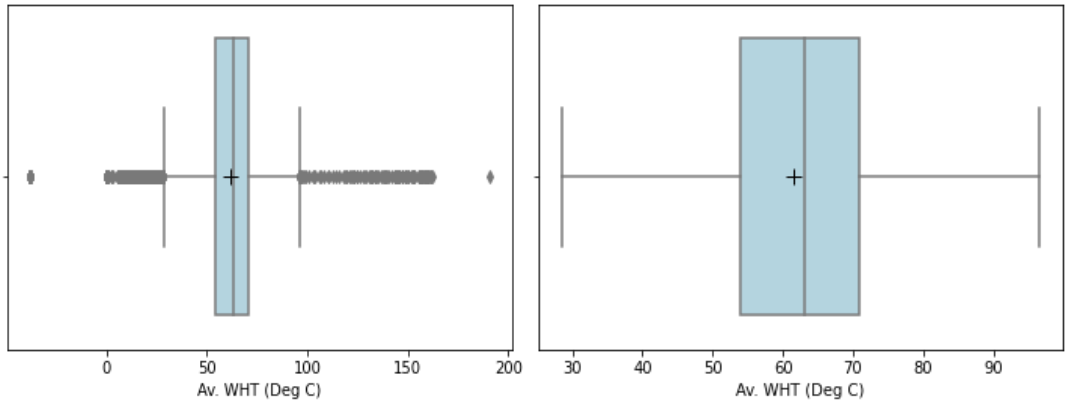
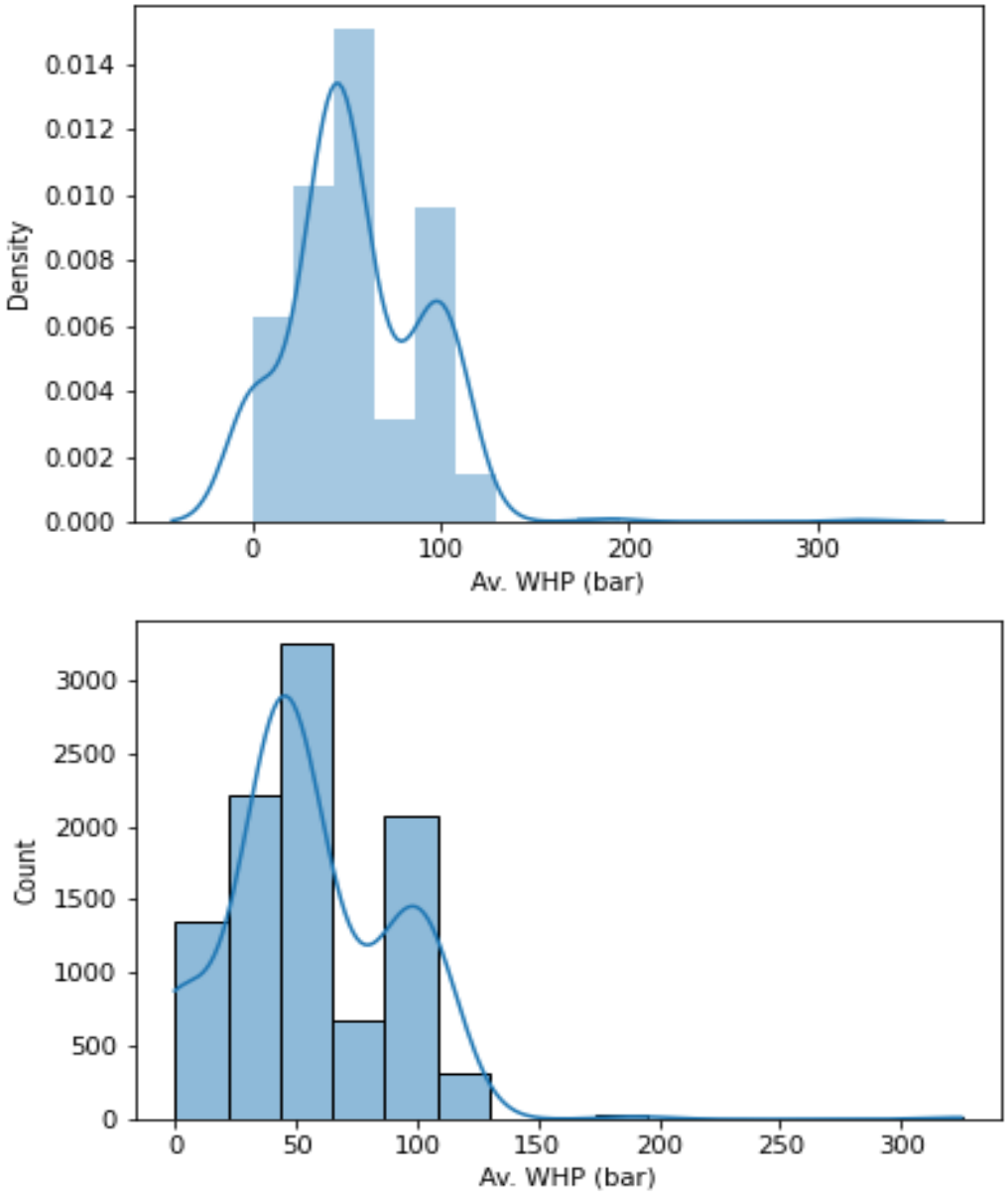


Figure A.15 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Av. WHT (Deg C)

Av. WHP (bar)



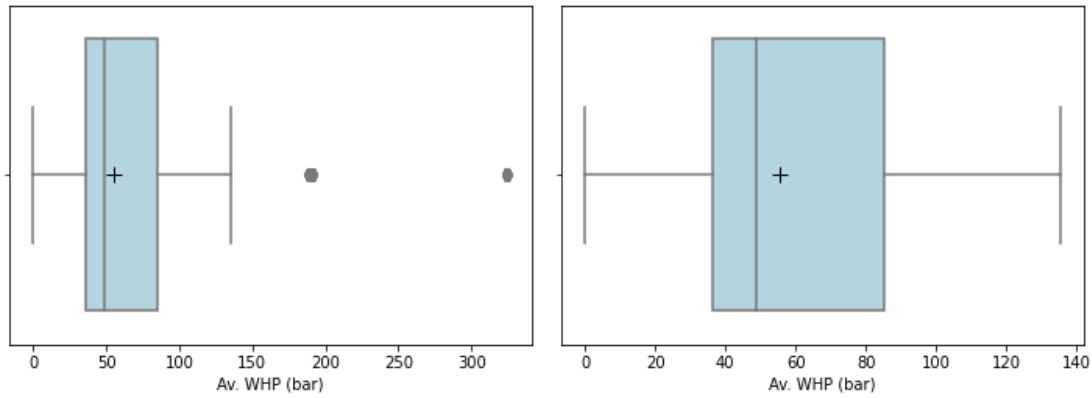


Figure A.16 : Kernel Density Estimation plot, histogram, and boxplot with and without outliers for Av. WHP (bar)

V.