Data Analysis Thesis

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 contains 15,634 rows of data whereas the Kyle Master dataset contains 27,324 rows of data. 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 contains valuable information. However, in order to ensure that the data in these datasets are ideal for a machine learning model, data cleaning and pre-processing must be done.

A) Empty Data

Analysis of Missing Data

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, the open-sourced *missingno* library was used. The *missingno* library contains a visualization technique entitled heatmap, this technique was utilized on both datasets. The heatmap automatically omits variables that are always present or always empty as these variables would not give significant insights. The correlation of missing values in the dataset ranges from -1 to 1. 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 express that if one feature is present than the other feature will unquestionably be present as well . 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. 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 which behaves differently. 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.

1) Volve Dataset

Figure 1 shows the heatmap which describes how the presence or absence of one variable affects another variable in the Volve dataset.

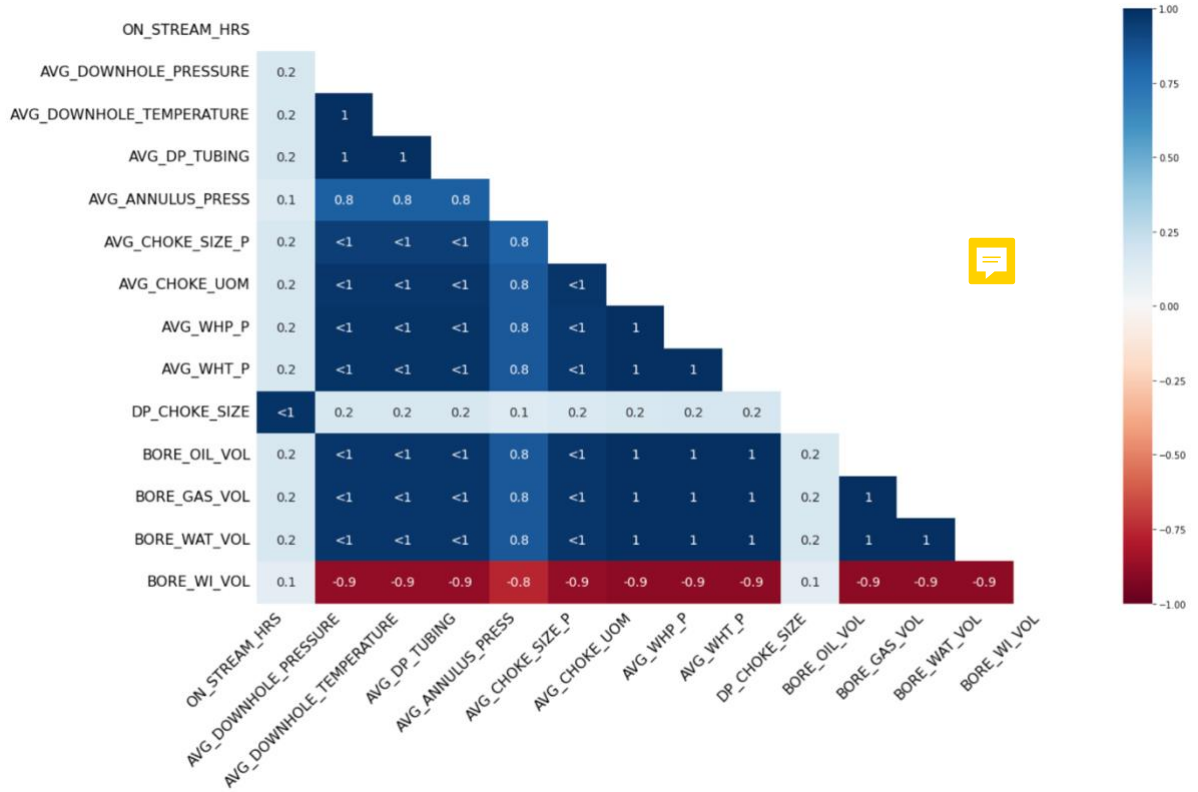

Figure 1 : Correlation of Missing Values in Volve

In Figure 1, it can be observed that most of the features are highly correlated to one another. The heatmap mostly consists of <1 correlation (28 occurrences) and 1 correlation (18 occurrences). This result suggests that the missing values in the Volve dataset follows the ~~MNAR (*Missing Not at Random)*~~ mechanism therefore the presence of these missing values are significant and could affect the model's performance.

2) Kyle Master Dataset

Figure 2 shows the heatmap that portrays how the absence or presence of one value affects another value in the Kyle Master dataset.

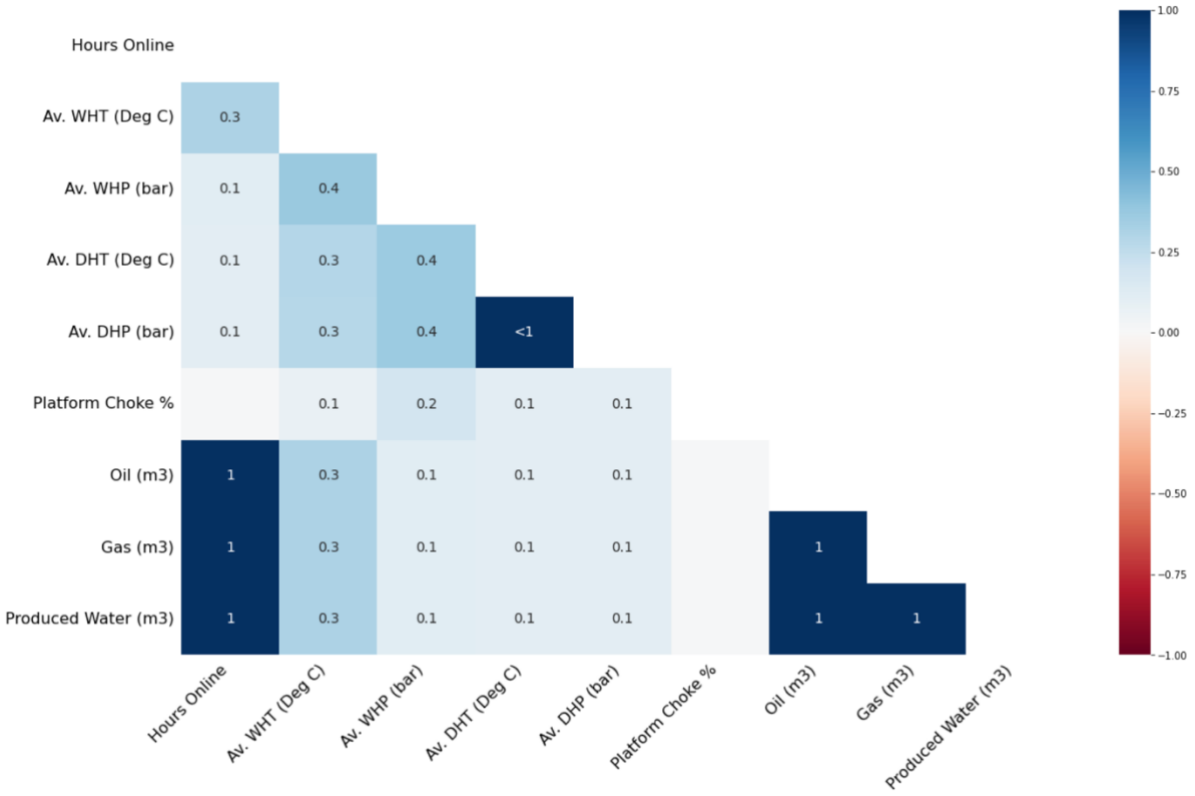

Figure 2 : Correlation of Missing Values in Kyle Master

In Figure 2, it is shown that most of the features are barely correlated as most of the correlation values near 0. The most common correlation value is 0.1 which occurs 15 times. However, there are some features that are highly correlated to another feature. For instance, *Oil (m3)* and *Gas (m3)* have a correlation value of 1 which means these missing values of these features are dependent on one another. Additionally, *Av. DHP (bar) and Av. DHT (Deg C)* also have a correlation value of 1 which suggest that they are dependent on each other. Most of the features do not show much correlation, however, there are features which are highly correlated. Therefore, this observation suggests that the Kyle Master follows the MAR (*Missing At Random)* mechanism. The MAR mechanism is not as significant as the MNAR mechanism . However, there is a slight possibility that the missing values would affect the model's performance.

B) Data Imputation

As has been mentioned in *Empty Data* section of this paper, both Volve and Kyle Master dataset contains a great deal of missing values. Additionally, the missing data mechanisms are not *MCAR,* 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 mean value of the feature.

C) Correlation in Dataset

In this section, this paper will explore the correlations between the features in dataset. Correlation is a measure which describes how one feature is related to another feature. There are different types of correlations, namely, positive, negative and no correlation. A positive correlation denotes that when the value of one feature increases, the value of the other feature increases as well. A negative correlation indicates that when the value of one feature increases, the value of the other feature decreases. On the other hand, no correlation means that the features are unrelated, thus a change in one feature would not impact the other feature.

Volve Dataset

1) Forward Filling

Figure 3 shows the Pearson correlation between the features in the Volve dataset when the Forward Filling method is used. From this heatmap we can determine which features are highly correlated. This heatmap shows that the features that are highly correlated are :

1) *BORE_OIL_VOL* and *BORE_GAS_VOL*
2) *AVG_DOWNHOLE_PRESSURE and AVG_DOWNHOLE_TEMPERATURE*
3) *AVG_WHT_P and AVG_WHP_P*

as their correlation values near 1 or -1.

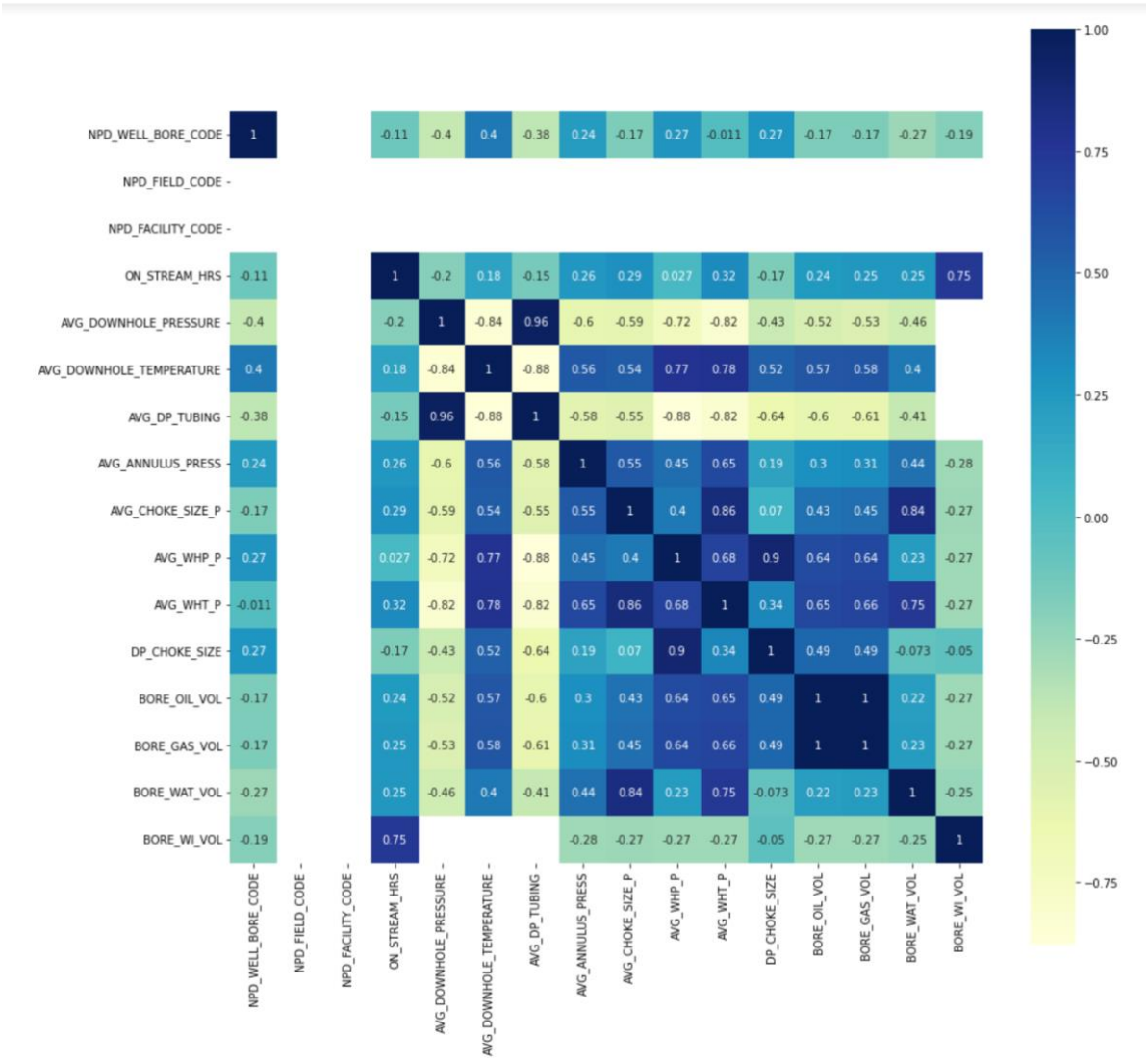

Figure 3 : Feature Correlation in Volve Dataset with Forward Filling

It is also possible to see the correlation between these features more clearly by plotting their values into a graph as shown in Figure 4, Figure 5, and Figure 6. Figure 4 shows the relationship between *BORE_OIL_VOL* and *BORE_GAS_VOL* is strong, positive and linear. Figure 5 shows that relationship between *AVG_DOWNHOLE_PRESSURE and AVG_DOWNHOLE_TEMPERATURE* is negative and linear. Whereas Figure 6 shows the

relationship for *AVG_WHT_P* and *AVG_WHP_P* is positive and linear.

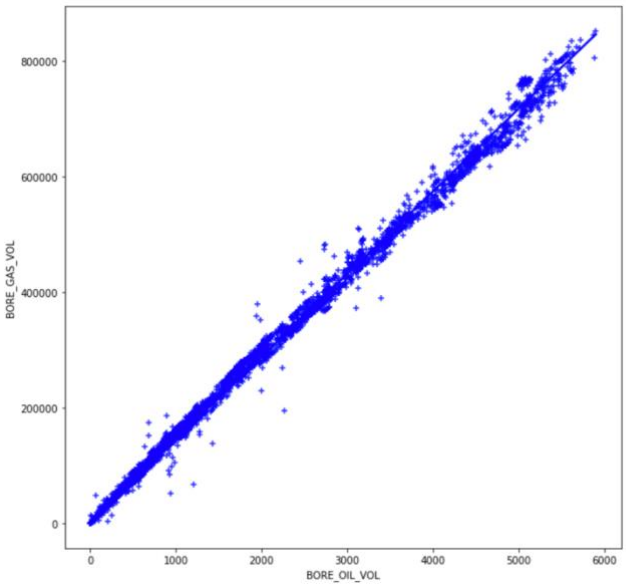

Figure 4 : Positive Linear Correlation between BORE_OIL_VOL and BORE_GAS_VOL

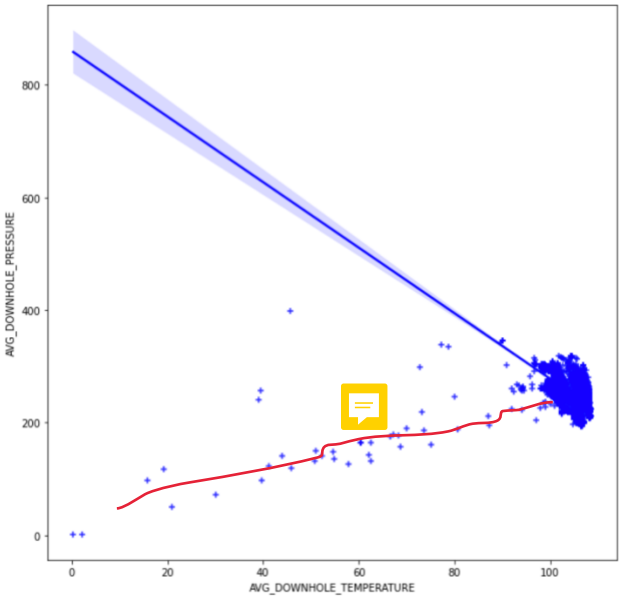

Figure 5 : Negative Linear Correlation between AVG_DOWNHOLE_PRESSURE and AVG_DOWNHOLE_TEMPERATURE

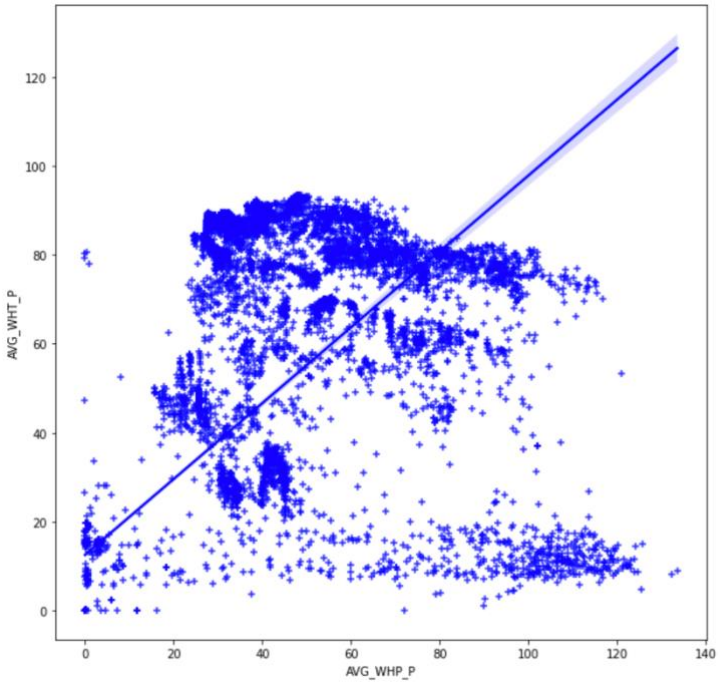

Figure 6 : Positive Linear Correlation between AVG_WHT_P and AVG_WHP_P

2) Mean Imputation

Figure 7 shows the Pearson correlation between the features in the Volve dataset when the Mean Imputation method is used. From this heatmap we can determine which features are highly correlated. The features that are highly correlated are :

1) *BORE_GAS_VOL and BORE_OIL_VOL*
2) *AVG_DOWNHOLE_PRESSURE and AVG_DP_TUBING*
3) *AVG_WHT_P and BORE_WAT_VOL*

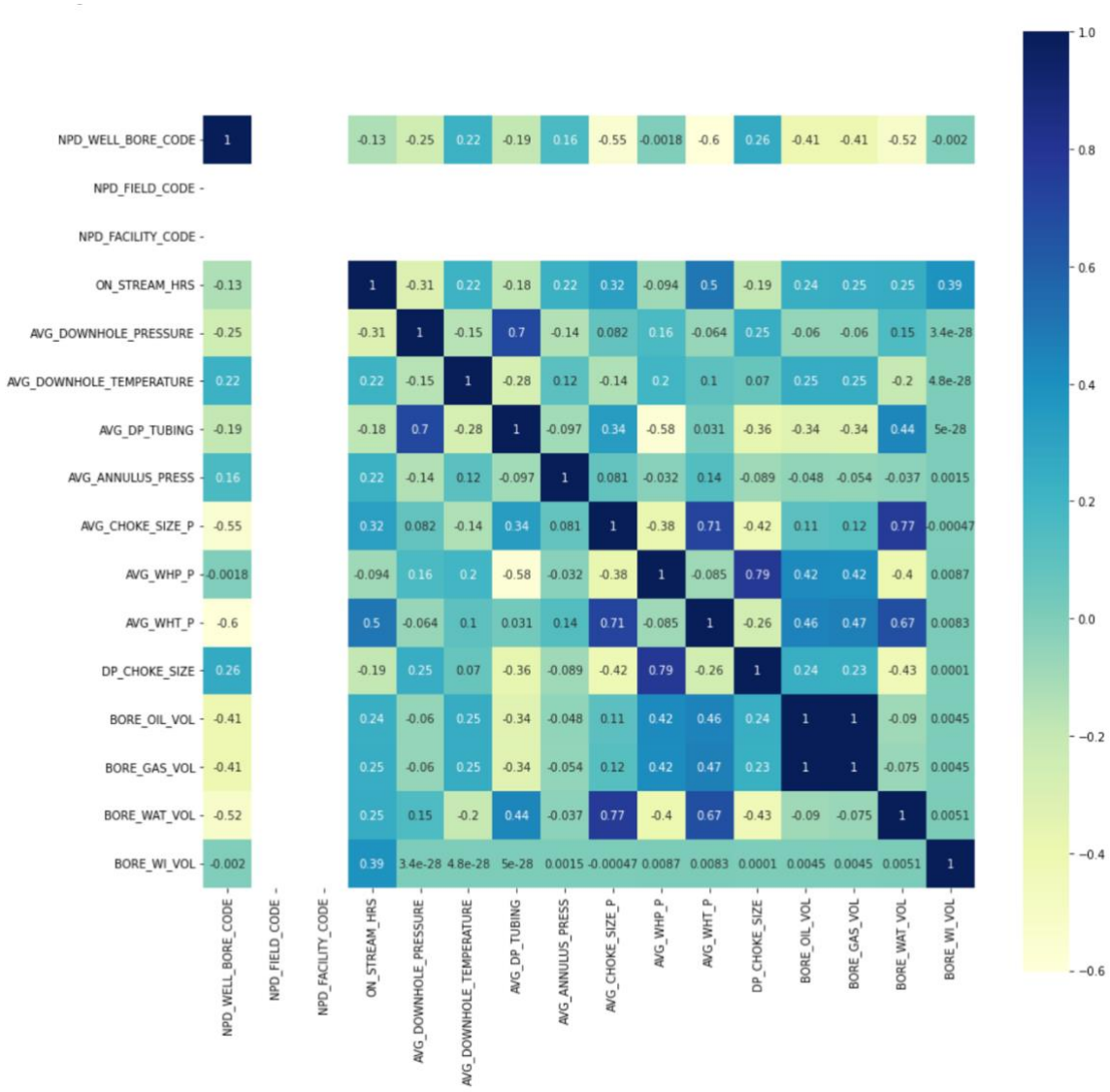

Figure 7 : Feature Correlation in Volve Dataset with Mean Imputation

Figure 8, Figure 9 and Figure 10 shows the relationship between these highly correlated features.

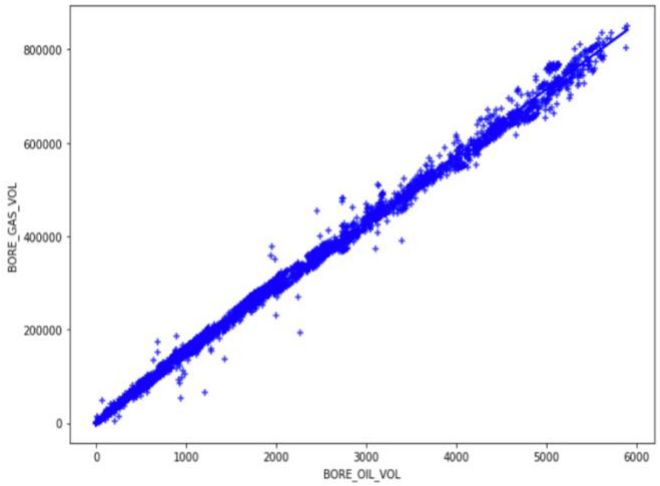

Figure 8 : Positive Linear Correlation between BORE_OIL_VOL and BORE_GAS_VOL

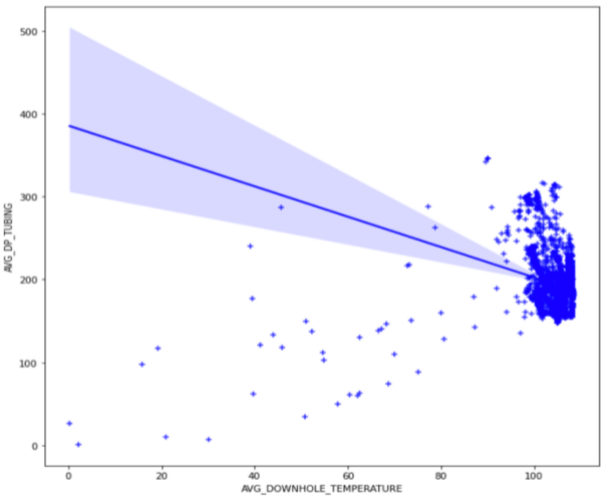

Figure 9 : Positive Linear Correlation between BORE_OIL_VOL and BORE_GAS_VOL

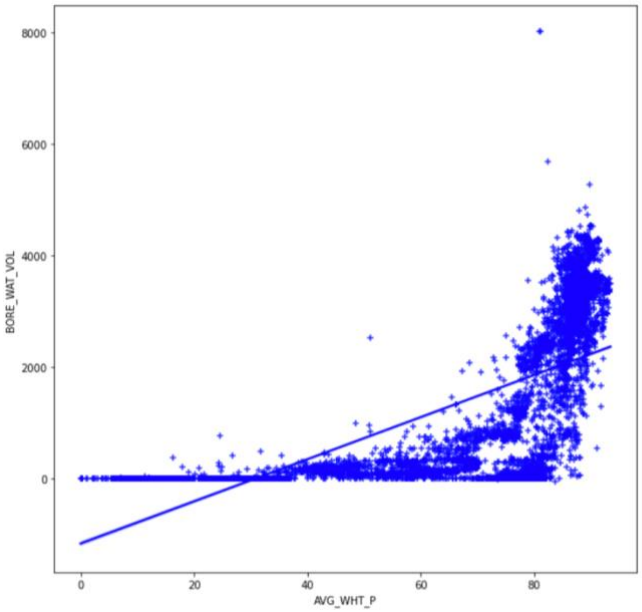

Figure 10 : Positive Correlation between AVG_WHT_P and BORE_WAT_VOL

The feature correlation in Figure 7 is not as strong as in the feature correlation Figure 3. This could suggest that Forward Filling could be better than Mean Imputation for this dataset.

Kyle Master

1) Forward Filling

Figure 11 shows the Pearson correlation between the features in the Kyle Master dataset when the Forward Filling method is used. From this heatmap we can determine which features are highly correlated. This heatmap shows that the features that are highly correlated are :

1. *Av. DHT (Deg C) and Av. DHP (bar)*
2. *Oil (m3) and Gas (m3)*
3. *Av. WHT (Deg C) and Oil (m3)*

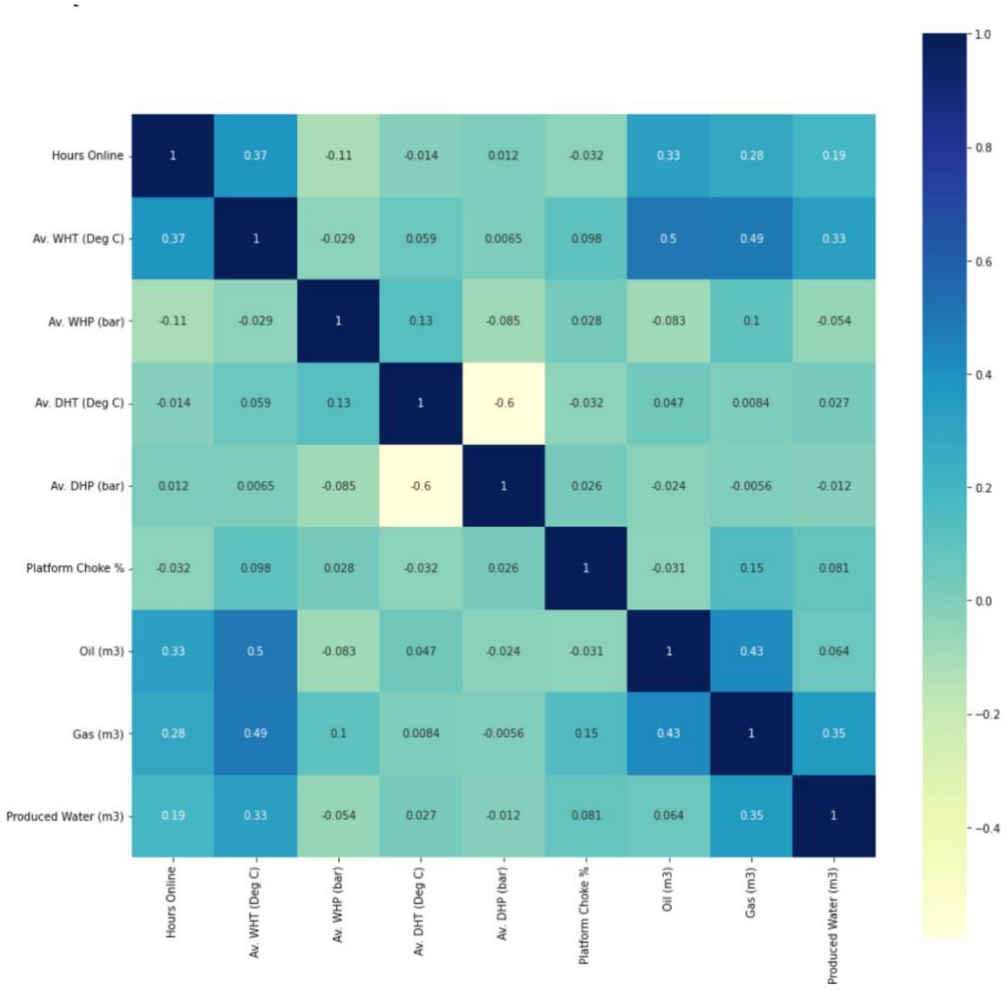

Figure 11 : Feature Correlation in Kyle Master with Forward Filling

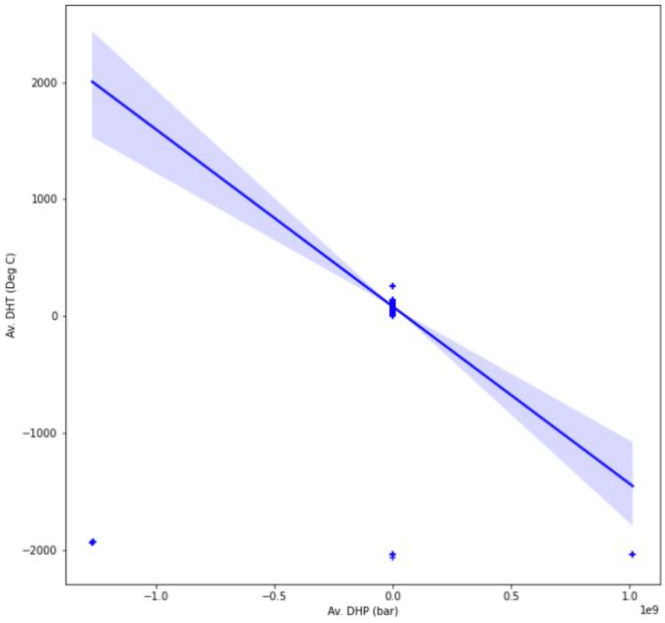

Figure 12 : Negative Linear Relationship between Av. DHT (Deg C) with Av. DHP (bar)

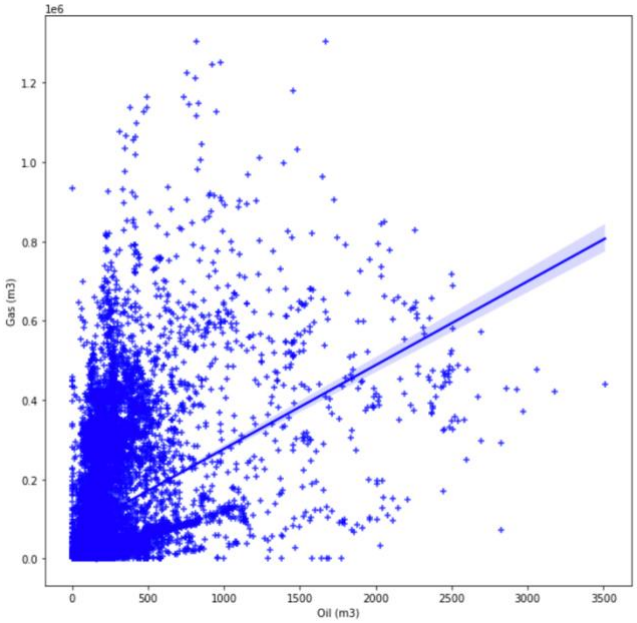

Figure 13 : Positive Linear Relationship between Oil (m3) and Gas (m3)

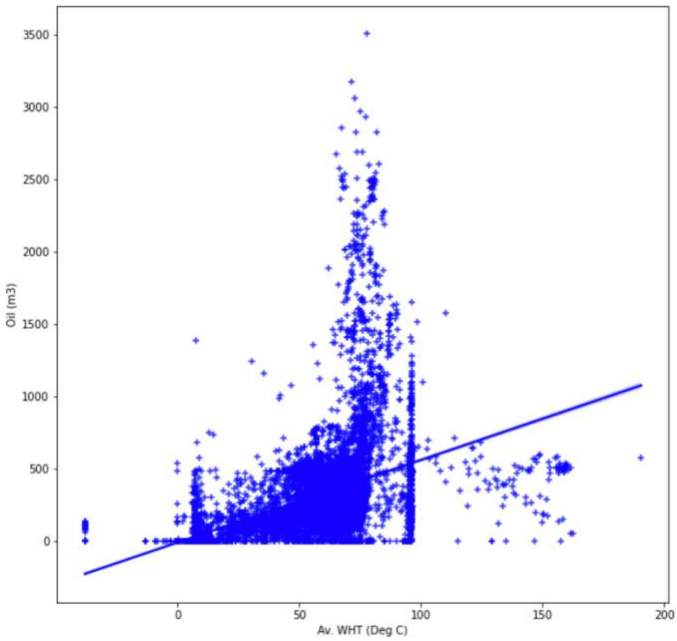

Figure 14 : Positive Linear Relationship between Av. WHT (Deg C) with Oil (m3)

Figure 12, Figure 13, and Figure 14 also shows more clearly the relationship between *Av. DHT (Deg C) with Av. DHP (bar), Oil (m3) with Gas (m3), Av. WHT (Deg C) with Oil (m3), and Av. WHT (Deg C) with Gas (m3).* Figure 12 shows that the relationship between *Av. DHT (Deg C) with Av. DHP (bar)* is negative and linear. Figure 13 shows that the relationship between *Oil*

(m3) and *Gas (m3)* is positive and linear. Figure 14 shows that the relationship between *Av. WHT (Deg C)* and *Oil (m3)* is positive and linear as well.

2) Mean Imputation

Figure 15 shows the Pearson correlation between the features in the Kyle Master when the Mean Imputation method is used. The heatmap in Figure 15 is the same as the heatmap in Figure 11.

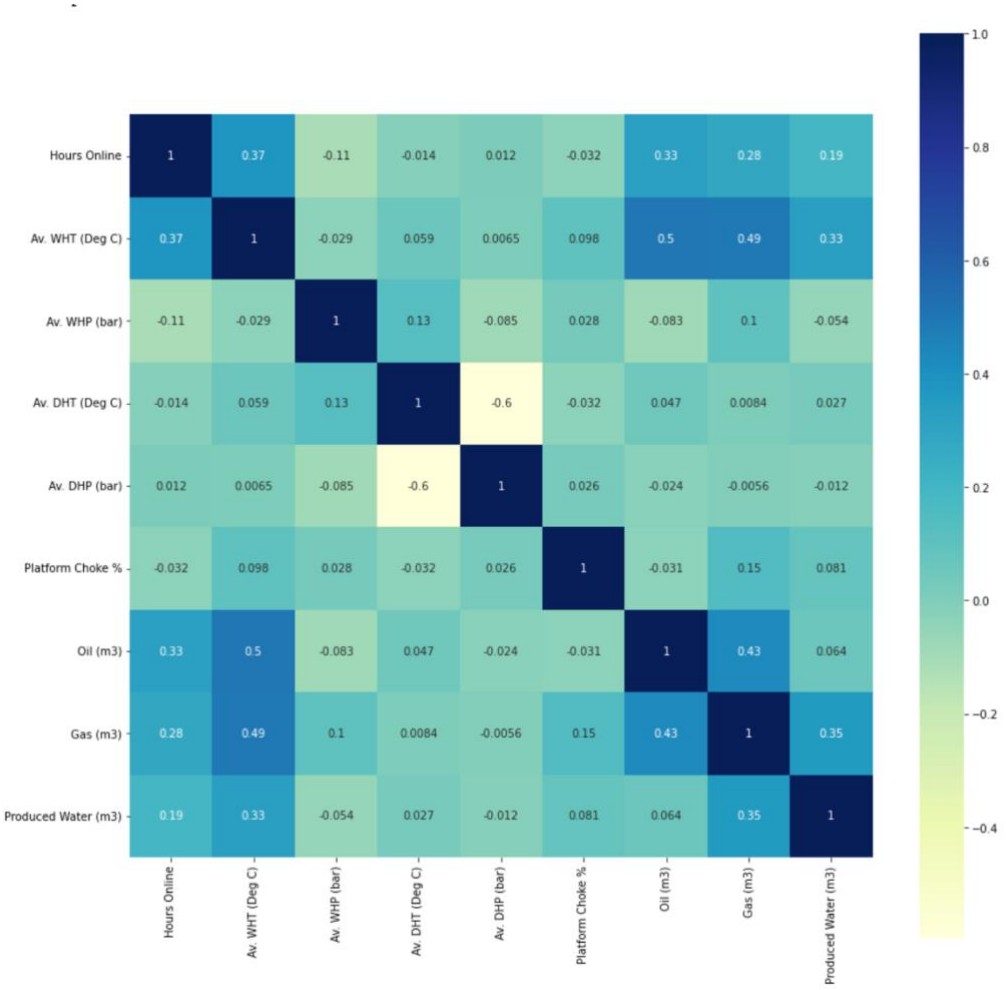

Figure 15 : Feature Correlation in Kyle Master with Mean Imputation

Hence, filling in the missing values with the mean influences the dataset the same way as filling it in with the forward filling method.

Data Imputation Method Decision

The Kyle Master dataset behaves the same way regardless of the data imputation used. Therefore, either one of these methods could be used for the dataset. However, in the Volve dataset, it would seem that the Forward Filling method would influence the dataset better. The feature correlation for the Volve dataset is not as good when the Mean Imputation method is used. Therefore, the ideal data imputation method would be Forward Filling.

D) Feature Selection

Feature Selection is the process of cutting down the input variables which will be fed into the models. This is useful as it gets rids of the noise in the dataset so that the model can focus on the useful information. 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. 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 it could reduce the model's accuracy. It would lead to lack of variation in the data or even result in data leakage which would make the model perform unrealistically good.

1) Volve Dataset

The volve dataset contains 24 columns namely :

1) DATEPRD
2) WELL_BORE_CODE
3) NPD_WELL_BORE_CODE
4) NPD_WELL_BORE_NAME
5) NPD_FIELD_CODE
6) NPD_FIELD_NAME
7) NPD_FACILITY_CODE
8) NPD_FACILITY_NAME
9) ON_STREAM_HRS
10) AVG_DOWNHOLE_PRESSURE
11) AVG_DOWNHOLE_TEMPERATURE
12) AVG_DP_TUBING
13) AVG_ANNULUS_PRESS
14) AVG_CHOKE_SIZE_P
15) AVG_CHOKE_UOM

16) *AVG_WHP_P*
17) *AVG_WHT_P*
18) *DP_CHOKE_SIZE*
19) *BORE_OIL_VOL*
20) *BORE_GAS_VOL*
21) *BORE_WAT_VOL*
22) *BORE_WI_VOL*
23) *FLOW_KIND*
24) *WELL_TYPE*

As the goal is to create a model that can predict oil and gas production, it is essential to include their production value. Thus, the first two features selected for model training are *BORE_OIL_VOL* and *BORE_GAS_VOL*. Afterwards, with respect to Figure 3, the next features selected would be *AVG_DOWNHOLE_PRESSURE* and *AVG_DOWNHOLE_TEMPERATURE* as these features have a high correlation value. Additionally, oil and gas formation are also heavily reliant on pressure and temperature, which makes these features ideal for the model's training. *AVG_WHP_P* is also added to the dataset as it shows decent correlation to *BORE_OIL_VOL* and *BORE_GAS_VOL*. In order to add variety to the dataset, other features should be added. Oil and gas production can also be improved by water injection, therefore *AVG_WHT_P* is also added to the model's training.

2) Kyle Master Dataset

The Kyle Master dataset contains 11 columns which are :

1) *Wellbore ID*
2) *Date*
3) *Hours Online*
4) *Av. WHT (Deg C)*
5) *Av. WHP (bar)*
6) *Av. DHT (Deg C)*
7) *Av. DHP (bar)*
8) *Platform Choke %*
9) *Oil (m3)*
10) *Gas (m3)*
11) *Produced Water (m3)*

The first two features selected are *Oil (m3)* and *Gas (m3)* as these features contain the production value of oil and gas. With respects to Figure 11, *Av. WHT (Deg C)* and *Av. WHP (bar)* are also included as they have decent correlation with *Oil (m3)* and *Gas (m3).* Furthermore, oil and gas production are also heavily reliant on the pressure and temperature of the reservoir, thus the features *Av. DHT (Deg C)* and *Av. DHP (bar)* are added for the model's training.

E) Feature Conversion

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 has 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 C^o (Celsius degree), while the pressure will be standardized into *bar*, and the volume will be standardized into m^3 (meter cubic).

F) Feature Statistics

In order to better understand the features in the dataset, several techniques could be employed to understand how the data is distributed.

1) Volve Dataset

a. AVG_DOWNHOLE_PRESSURE

Before Data Imputation

In order better understand this feature, a probability density function diagram was created to show the data distribution before data imputation as shown in Figure 16.

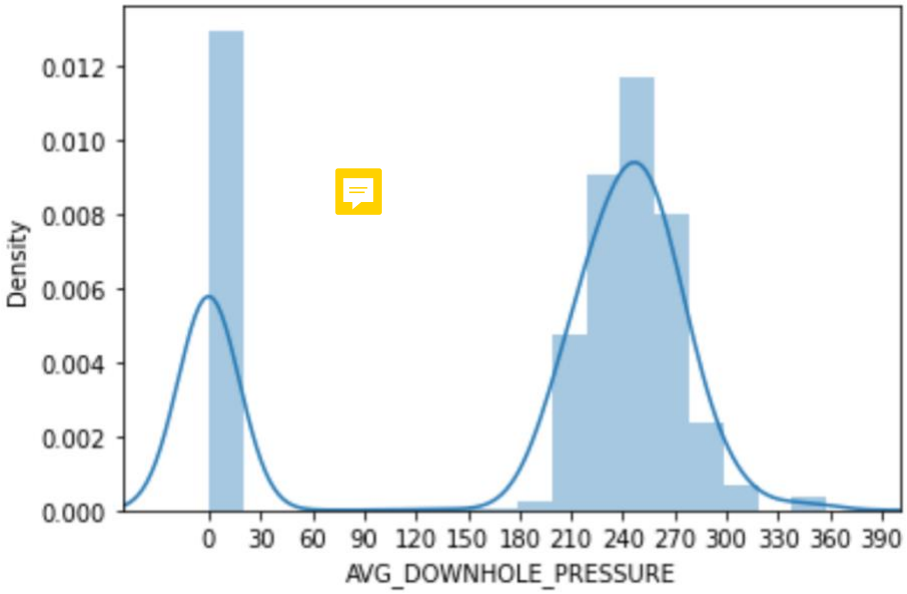

Figure 16 : Probability Density Function for AVG_DOWNHOLE_PRESSURE

Figure 16 shows that *AVG_DOWNHOLE_PRESSURE* is bimodal which means it contains two modes. From Figure 11, it can be estimated that the mode of this feature is 0 and 250. As this feature will be used to predict oil and gas production, it is not ideal for majority of the values to be 0. Therefore, it would be better to drop these data.

A boxplot was also used to understand the distribution of *AVG_DOWNHOLE_PRESSURE* as shown in Figure 17. In Figure 17, a "+" can be seen, this mark denotes the mean. Thus, from this observation, it can be deduced that the mean is around 181. Furthermore, there is no presence of outliers in this feature. Additionally, the boxplot also shows the lower quartile (Q_1) and the upper quartile (Q_3) of the feature. The lower quartile is 0 whereas the upper quartile is 255. Therefore, the Interquartile Range (IQR) of this feature is 255. This observation suggests that the middle 50% of values in this feature have a spread of 255.

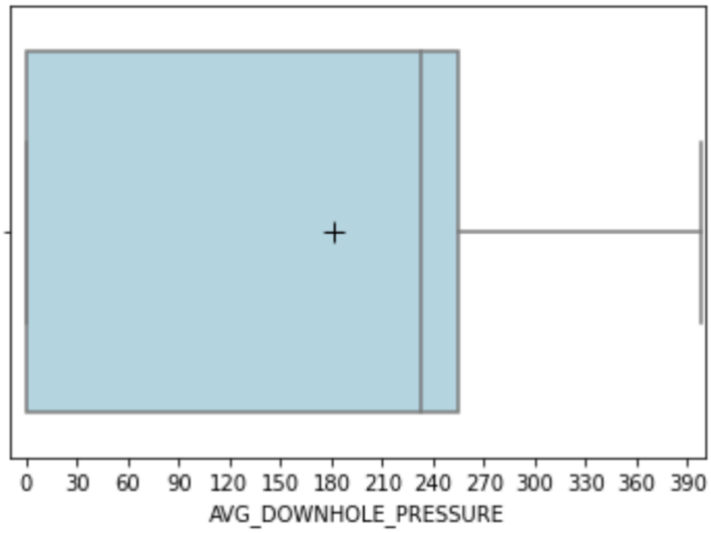

Figure 17 : Boxplot for AVG_DOWNHOLE_PRESSURE

In addition to this, the pandas inbuilt function *describe()* could be used to analyze the feature even further as shown in Figure 18.

```
count    8980.000000
mean      181.803869
std       109.712363
min         0.000000
25%         0.000000
50%       232.896939
75%       255.401455
max       397.588550
Name: AVG_DOWNHOLE_PRESSURE, dtype: float64
```

Figure 18 : Describe AVG_DOWNHOLE_PRESSURE

~~By using the *describe()* function~~, it can be determined that the mean of this feature is 181, the standard deviation is 109, the lower quartile is 0, the median is 232, whereas the upper quartile is 255.

Forward Filling

After filling in the missing values using the forward filling method, the distribution of data changes. A probability density function diagram was made as shown in Figure 19.

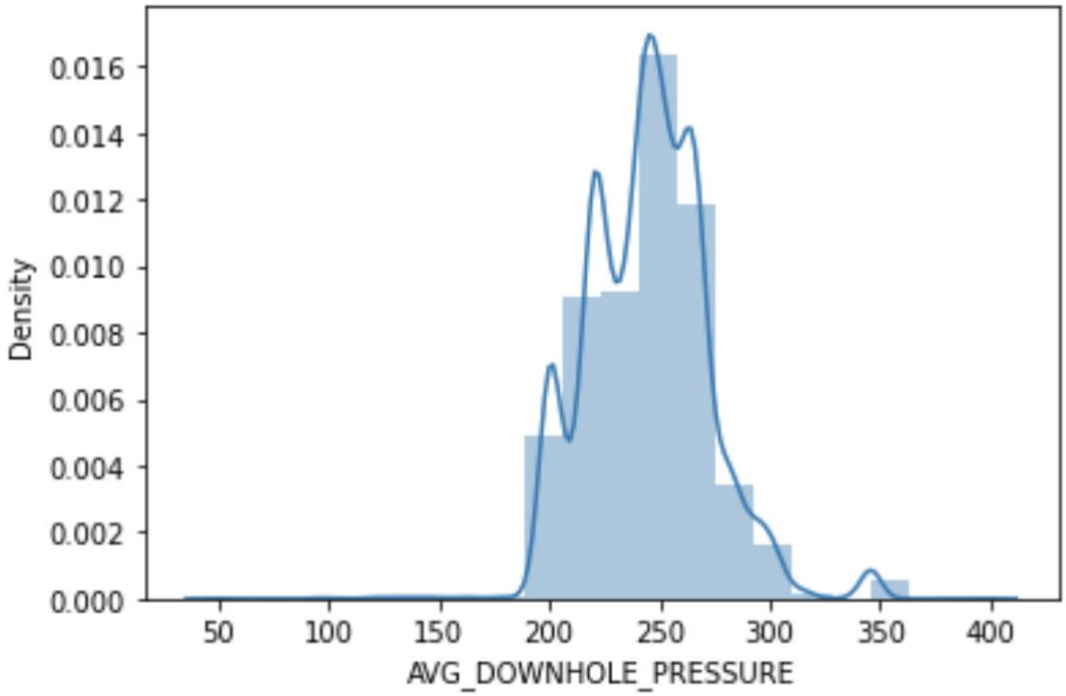

Figure 19 : Probability Density Function for AVG_DOWNHOLE_PRESSURE after Forward Filling

The density function in Figure 19 is different from Figure 16 as the missing values were filled in and the 0 values were dropped. In order to analyze the feature even further, pandas inbuilt function, *describe()* could be used as shown in Figure 20. From this, it can be seen that the mean is 244, the standard deviation is 27, the lower quartile is 223, the median is 245, and the upper quartile is 262. The data distribution can also be explained through a boxplot as shown in Figure 21. In the boxplot, the "+" denotes the mean of the feature, from this it can be deduced that the mean is around 250. Furthermore, it can be seen that this feature contains several outliers. Therefore, in order to see the distribution without the outliers, a boxplot could be created with the condition

of ignoring the outliers as shown in Figure 21. In Figure 21, it can be seen that the range for *AVG_DOWNHOLE_PRESSURE* is estimated to be from 170 to 319.

```
count    6666.000000
mean      244.913885
std        27.514695
min        49.450440
25%       223.842066
50%       245.024178
75%       262.461721
max       397.588550
Name: AVG_DOWNHOLE_PRESSURE, dtype: float64
```

Figure 20 : Describe AVG_DOWNHOLE_PRESSURE after Forward Filling

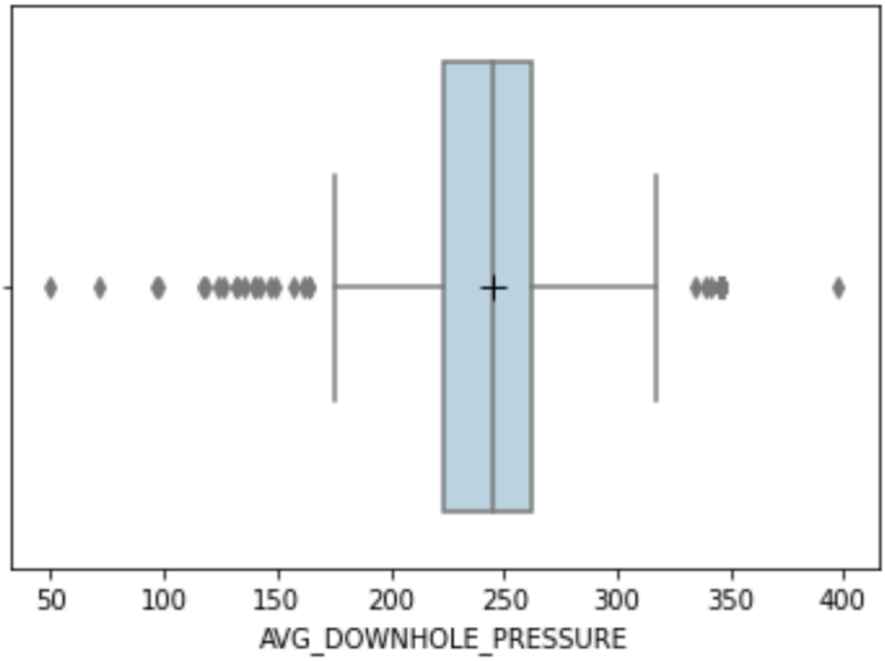

Figure 21 : Boxplot for AVG_DOWNHOLE_PRESSURE with Forward Filling

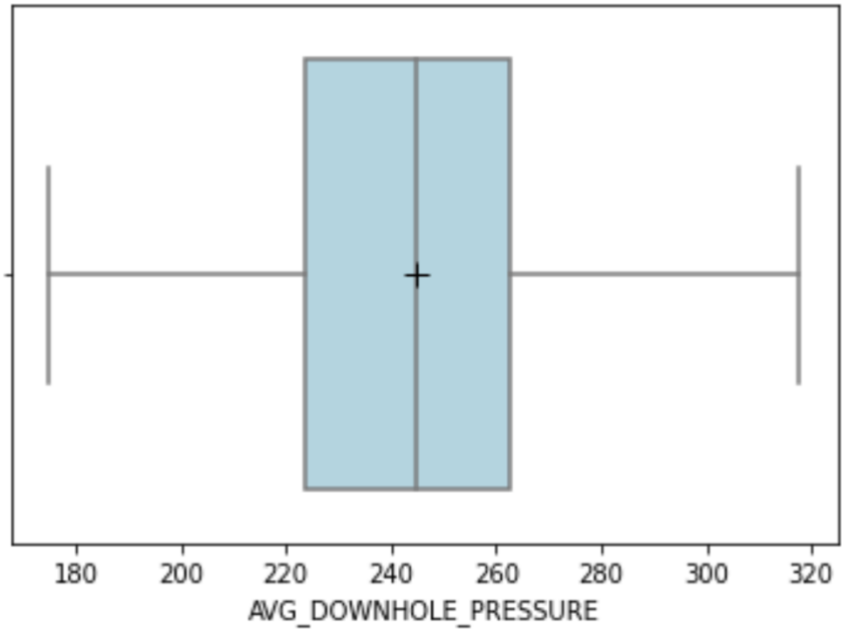

Figure 22 : Boxplot for AVG_DOWNHOLE_PRESSURE with Forward Filling and Ignoring Outliers

b. AVG_DOWNHOLE_TEMPERATURE

Before Data Imputation

A probability density function was created to understand the data distribution of this feature before data imputation as shown in Figure 23. Similar to the density function in Figure 16, there is a great number of zero values. As temperature is a feature that will be used to predict oil and gas production, these zero values should be removed.

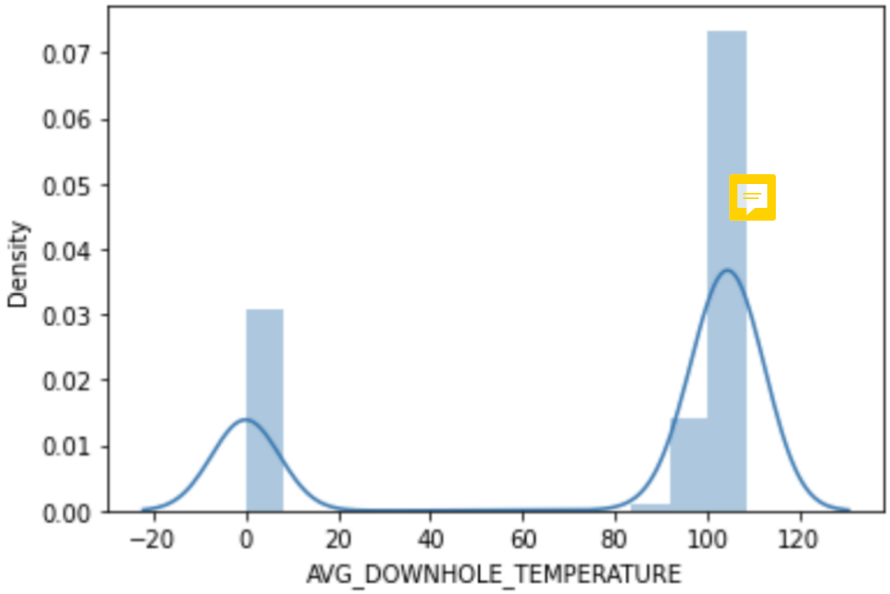

Figure 23 : Probability Density Function in AVG_DOWNHOLE_TEMPERATURE

A boxplot was also utilized to understand the data distribution as shown in Figure 24. The '+' denotes that the mean for this distribution is around 77, furthermore, there are no outliers in this distribution.

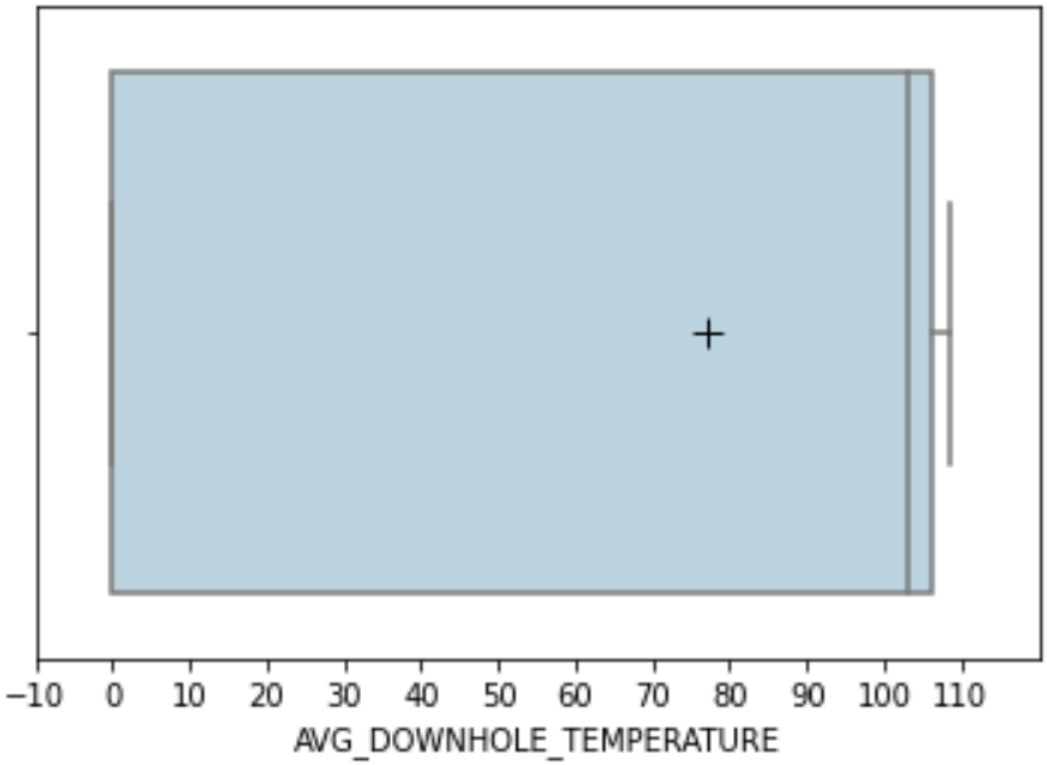

Figure 24 : Boxplot for AVG_DOWNHOLE_TEMPERATURE

```
count    8980.000000
mean       77.162969
std        45.657948
min         0.000000
25%         0.000000
50%       103.186689
75%       106.276591
max       108.502178
Name: AVG_DOWNHOLE_TEMPERATURE, dtype: float64
```

Figure 25 : Describe AVG_DOWNHOLE_TEMPERATURE

Pandas inbuilt *describe()* function can be utilized to obtain more specific statistics as shown in Figure 25. Figure 25 shows that the mean is 77, the standard deviation is 45, the lower quartile is 0, the median is 103 and the upper quartile is 106.

Forward Filling

After filling in the missing values using the forward filling method, the distribution of data shows some changes. A probability density function diagram was made as shown in Figure 26. Figure 26 shows that the data distribution after the missing values were filled in and the zero values were dropped. Figure 27 shows the specific statistics of the distribution made from pandas inbuilt function *describe()*. Figure 27 shows that the mean is 97, the standard deviation is 7, the lower quartile is 90, the median is 90, and the upper quartile is 105. Figure 28 also helps describe the data distribution of the feature. It shows that mean of this distribution is estimated to be around 93 due to the location of the "+" symbol. However, there is a significant number of outliers which should be removed so that the model will train better. Figure 29 shows that if the outliers are removed, the temperature ranges from approximately 65 to 109.

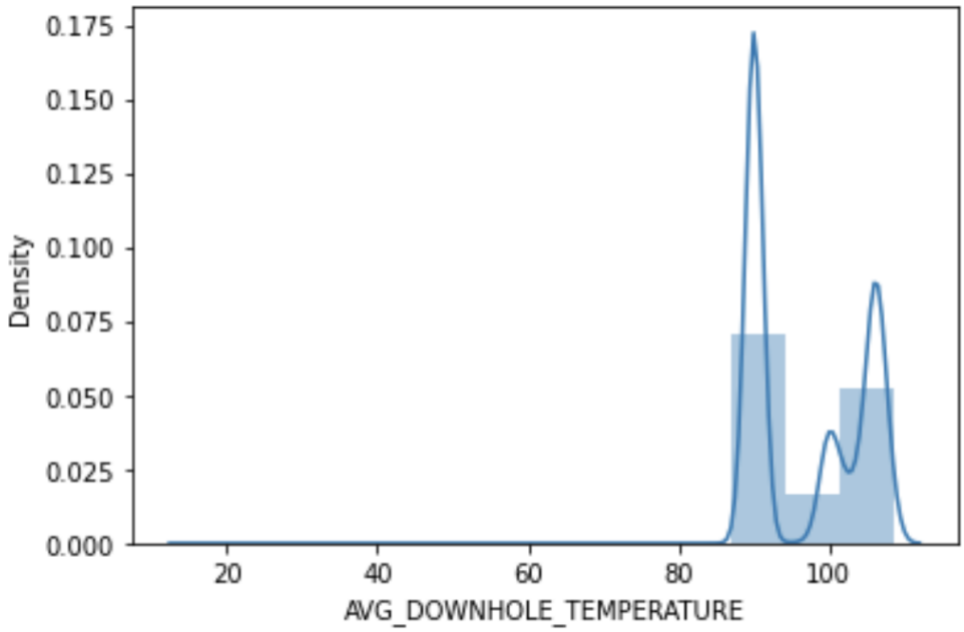

Figure 26 : Probability Density Function for AVG_DOWNHOLE_TEMPERATURE with Forward Filling

```
count    13313.000000
mean        97.010965
std          7.777522
min         15.885040
25%         90.034330
50%         90.034330
75%        105.707021
max        108.502178
Name: AVG_DOWNHOLE_TEMPERATURE, dtype: float64
```

Figure 27 : Describe AVG_DOWNHOLE_TEMPERTURE with Forward Filling

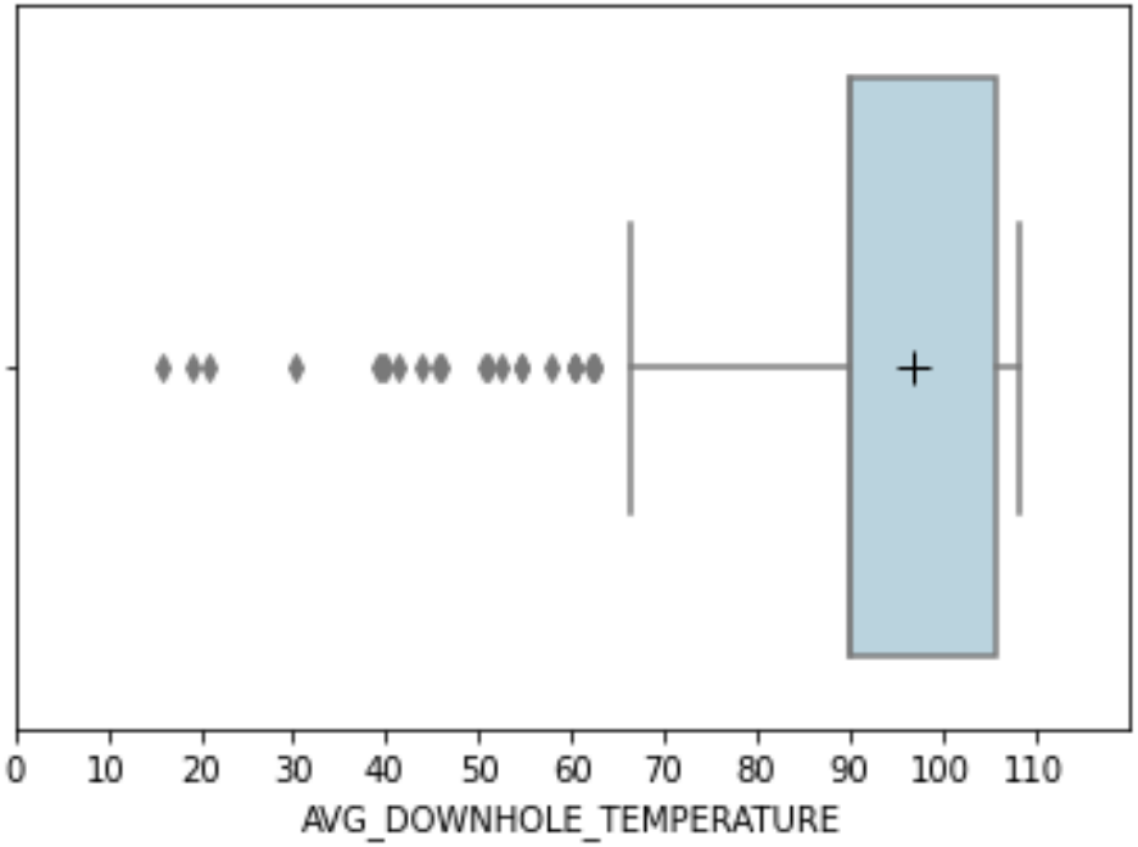

Figure 28 : Boxplot AVG_DOWNHOLE_TEMPERATUE with Forward Filling

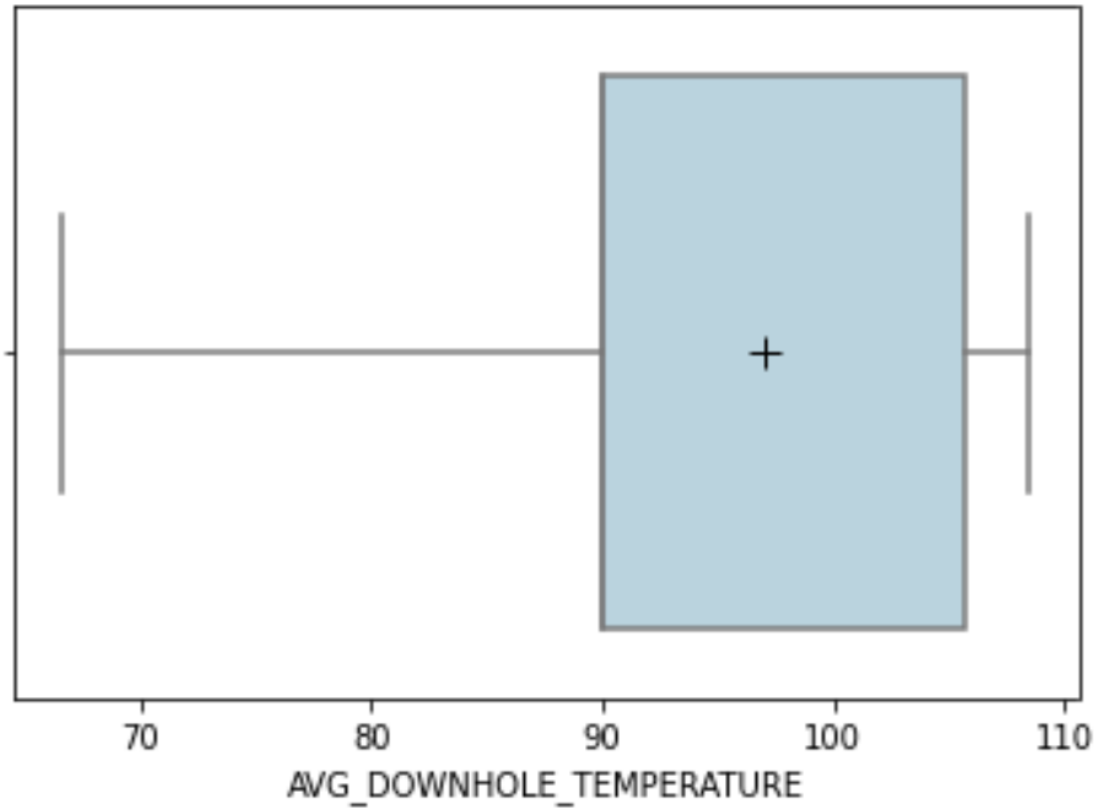

Figure 29 : Boxplot AVG_DOWNHOLE_TEMPERATUE with Forward Filling and no Outliers

c. BORE_OIL_VOL

Before Data Imputation

A probability density function was created to understand the data distribution of this feature before data imputation as shown in Figure 30. This distribution is positively skewed so most of the values are clustered around the left tail. Figure 31 shows that the mean is 1095, the standard deviation is 1323, the lower quartile is 190, the median is 557, and the upper quartile is 1345. Figure 32 shows the boxplot of this data distribution, from this it is clear that there are several outliers in the dataset. These outliers should be removed.

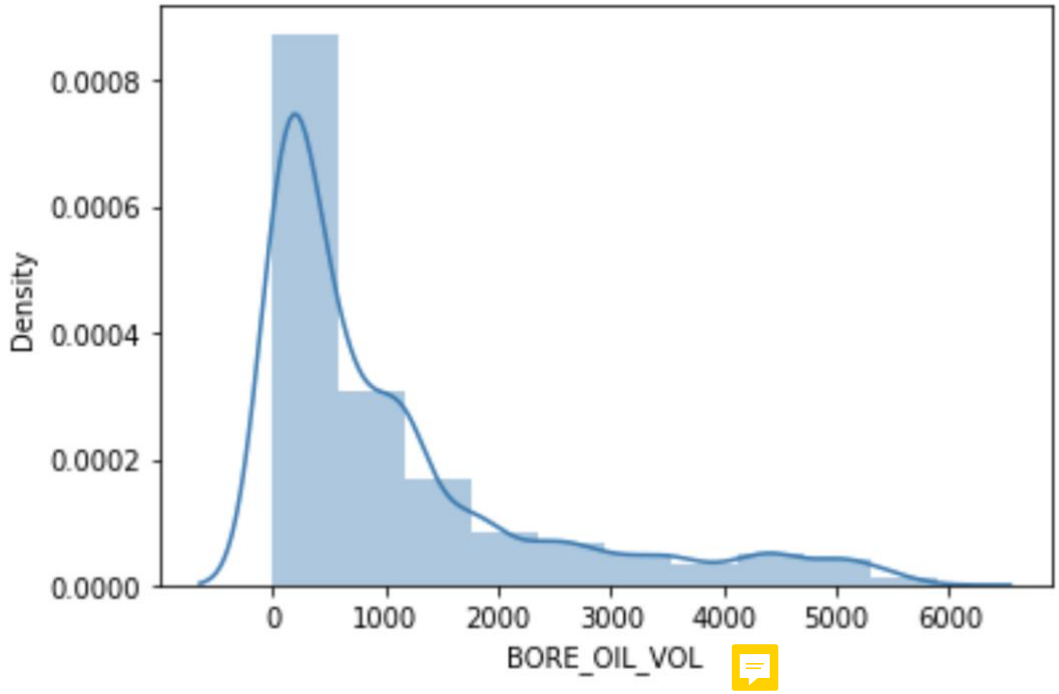

Figure 30 : Probability Density Function for BORE_OIL_VOL

```
count    9161.000000
mean     1095.631548
std      1323.538151
min         0.000000
25%       190.690000
50%       557.550000
75%      1345.200000
max      5901.840000
Name: BORE_OIL_VOL, dtype: float64
```

Figure 31 : Describe BORE_OIL_VOL

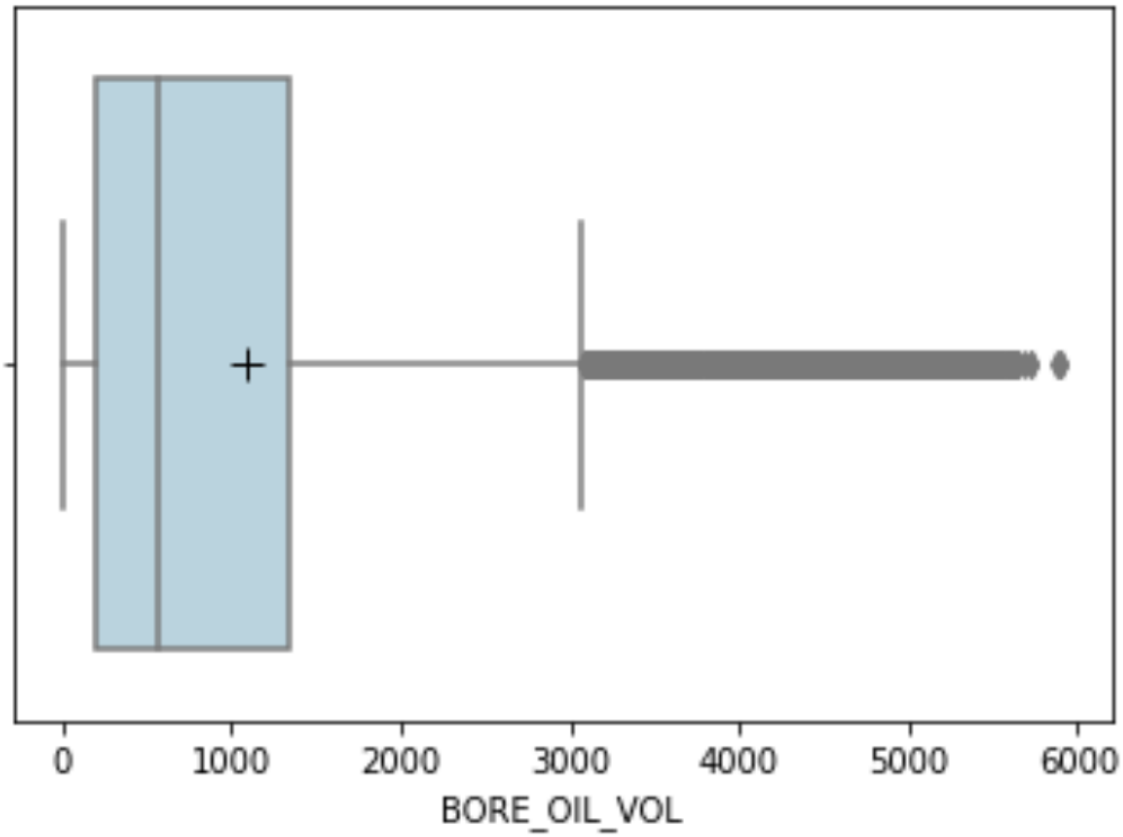

Figure 32 : Boxplot for BORE_OIL_VOL

Forward Filling

There are no changes to the data distribution for *BORE_OIL_VOL* after filling in the missing values with the Forward Filling method. However, there are changes after removing the outliers that was shown in Figure 32. After removing the outliers, the data distribution is shown in Figure 33.

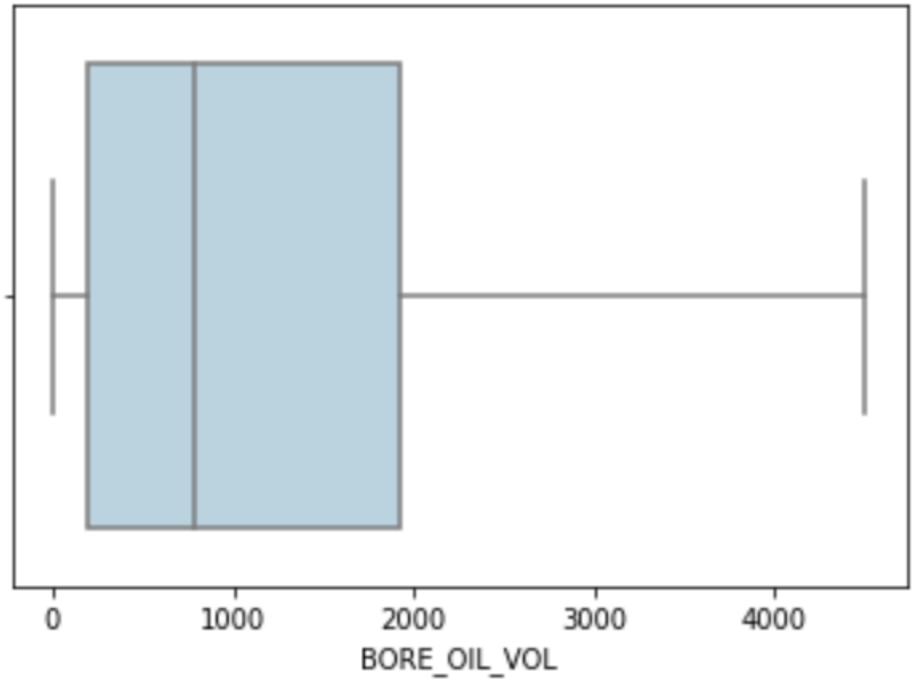

Figure 33 : Boxplot of BORE_OIL_VOL after Removing Outliers

d. BORE_GAS_VOL

Before Data Imputation

A probability density function was created to understand the data distribution of this feature before data imputation as shown in Figure 34. This distribution is positively skewed so most of the values are clustered around the left tail. Figure 35 shows the boxplot of this data distribution. The "+" in the boxplot denotes the mean of the distribution, therefore it can be assumed that the mean is approximately 160,000. Furthermore, there are several outliers in the feature which needs to be addressed. Figure 36 shows that the mean is 161,049, the standard deviation is 188,136 the lower quartile is 29,430, the median is 87,749, and the upper quartile is 20,2482.

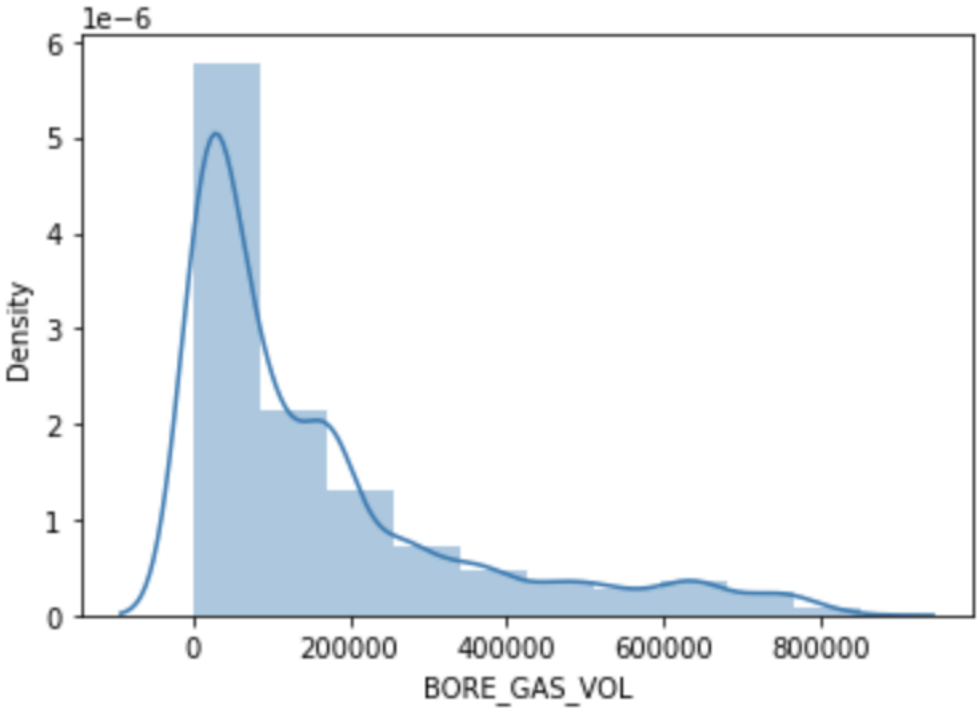

Figure 34 : Probability Density Function for BORE_GAS_VOL

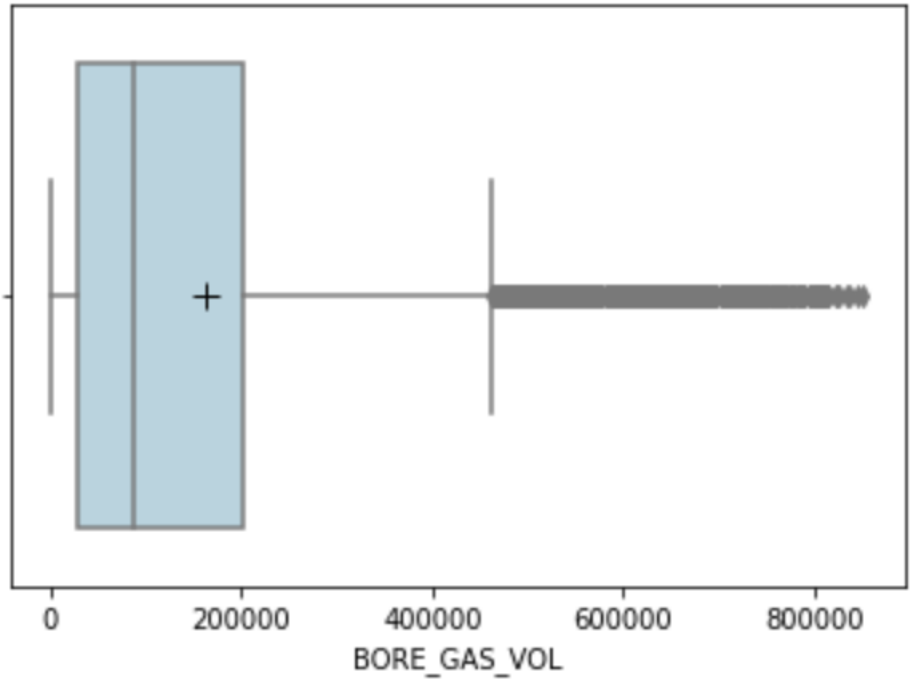

Figure 35 : Boxplot for BORE_GAS_VOL

```
count      9161.000000
mean     161049.059703
std      188136.410434
min           0.000000
25%       29430.590000
50%       87749.660000
75%      202482.300000
max      851131.520000
Name: BORE_GAS_VOL, dtype: float64
```

Figure 36 : Describe BORE_GAS_VOL

Forward Filling

The probability density function of *BORE_GAS_VOL* still remains the same after filling in the missing values. However, as can be seen in Figure 37, the mean of the feature has changed. The boxplot in Figure 35 showed that the feature contained several outliers, therefore these outliers were removed as showed in Figure 38.

```
count      6666.000000
mean     192098.611787
std      206705.292629
min           0.000000
25%       30552.042500
50%      118927.325000
75%      287018.292500
max      835981.330000
Name: BORE_GAS_VOL, dtype: float64
```

Figure 37 : Describe BORE_GAS_VOL after Forward Filling

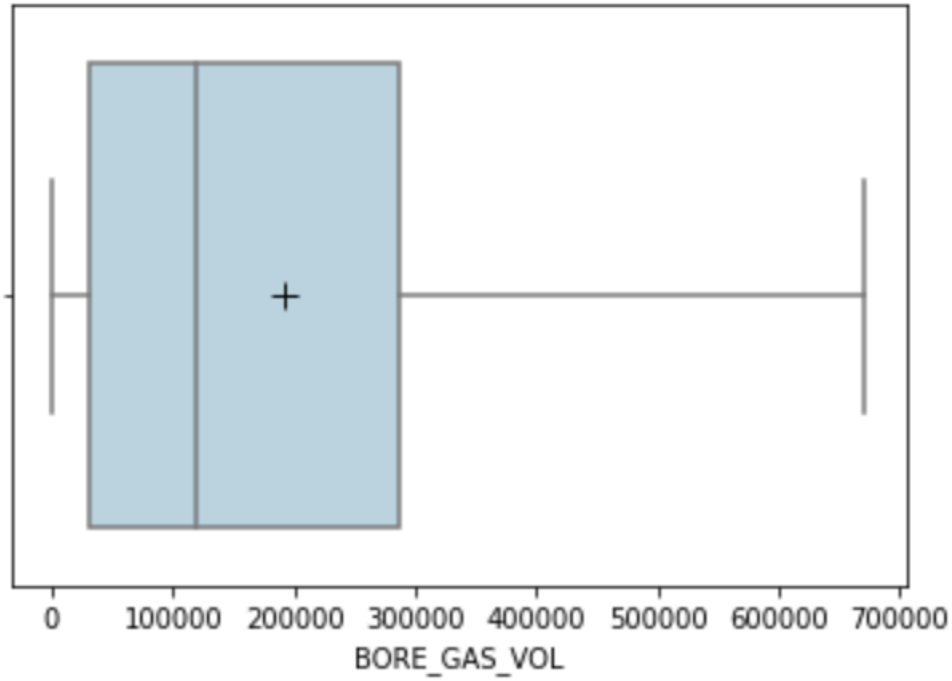

Figure 38 : Boxplot of BORE_GAS_VOL after Removing Outliers

e. AVG_WHP_P

Before Data Imputation

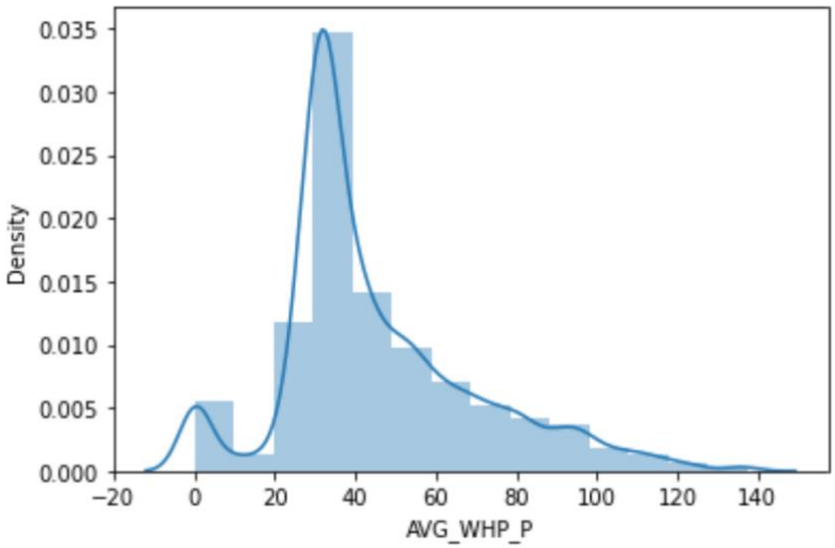

Figure 39 : Probability Density Function for AVG_WHP_P

Figure 39 shows the probability density function of this feature. From the boxplot shown in Figure 40, that there are outliers that start from 96. Additionally, the "+" shows that the mean of the distribution is around 45. In addition to this, Figure 41 shows

that the standard deviation is 24, the lower quartile is 31, the upper quartile is 57 and the median is 50.

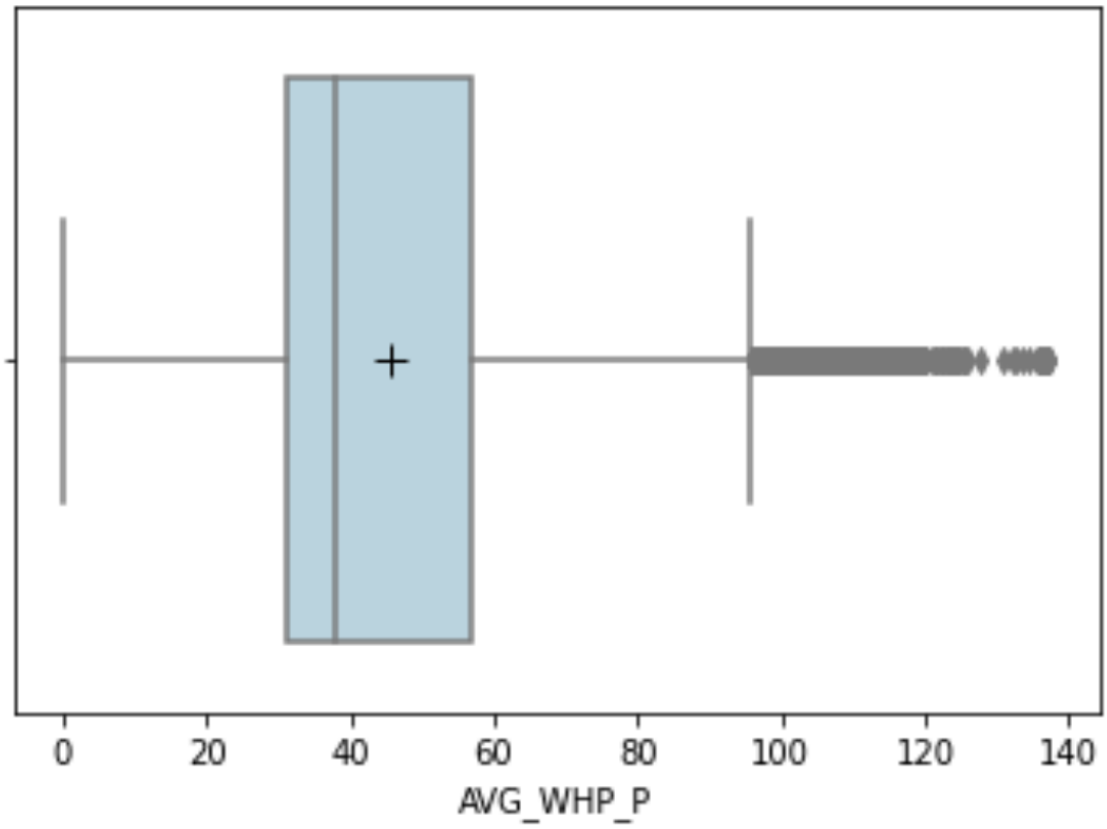

Figure 40 : Boxplot for AVG_WHP_P

```
count    9155.000000
mean       45.377811
std        24.752631
min         0.000000
25%        31.148062
50%        37.933620
75%        57.101268
max       137.311030
Name: AVG_WHP_P, dtype: float64
```

Figure 41 : Describe AVG_WHP_P

Forward Filling

The probability density function for this feature remains the same after filling in the missing values with the Forward Filling Method. However, as shown in Figure 42, the outliers in the feature has decreased. Figure 43 shows the distribution after removing the remaining outliers. Additionally, Figure 43 shows it can be seen that the mean has also changed as is supported by Figure 44. Figure 44 shows that mean is 49, the standard deviation is 24, the lower quartile is 31, the upper quartile is 64 and the median is 43.

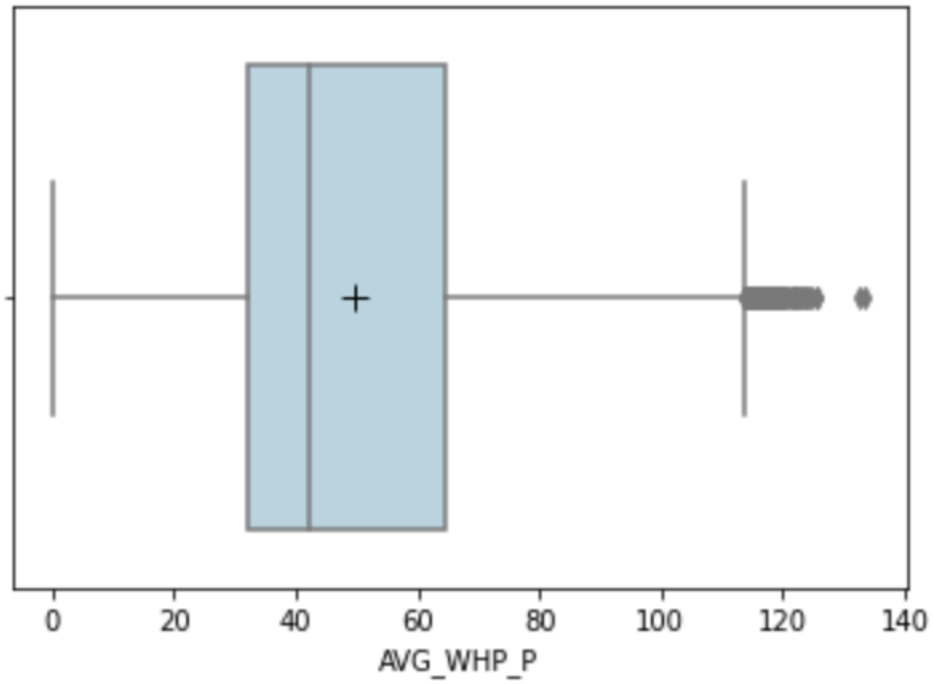

Figure 42 : Boxplot for AVG_WHP_P after Forward Filling

```
count    6666.000000
mean       49.539196
std        24.699751
min         0.000000
25%        31.862312
50%        42.070633
75%        64.557200
max       133.592510
Name: AVG_WHP_P, dtype: float64
```

Figure 44 : Describe AVG_WHP_P

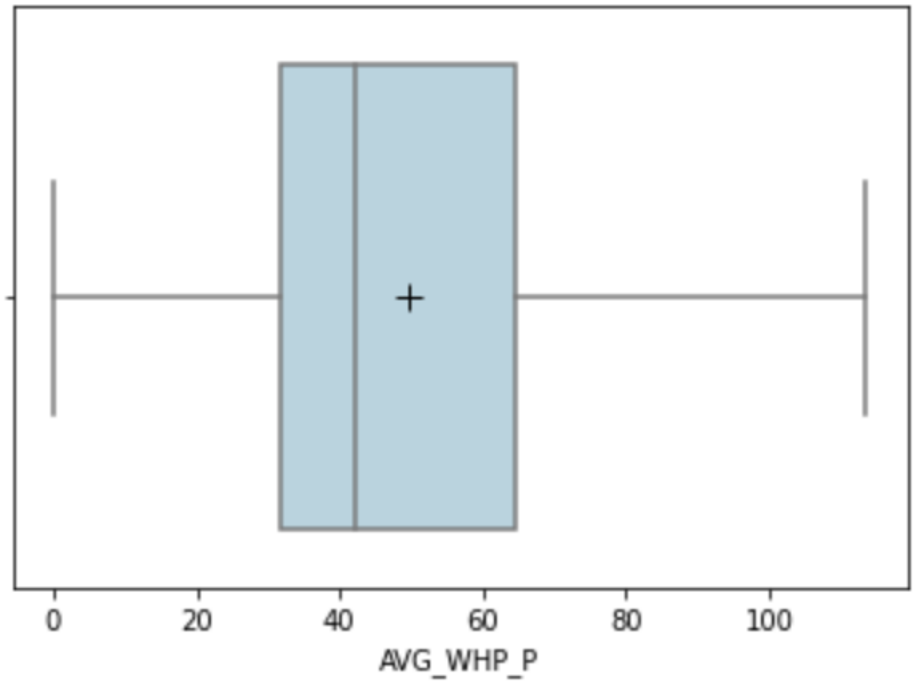

Figure 43 : Boxplot for AVG_WHP_P after Removing Remaining Outliers

f. AVG_WHT_P

Before Data Imputation

Figure 45 shows the probability density function for *AVG_WHT_P* . The distribution negatively skewed so most of the values are clustered on the right tail. Figure 46 shows that the mean is approximately 67 as denoted by the "+" on the boxplot. Figure 46 also shows that the feature has outliers near the lower limit which should be removed. Additionally, Figure 47 shows that the mean is 67, the standard deviation is 27, the lower quartile is 56, the upper quartile is 88 and the median is 80.

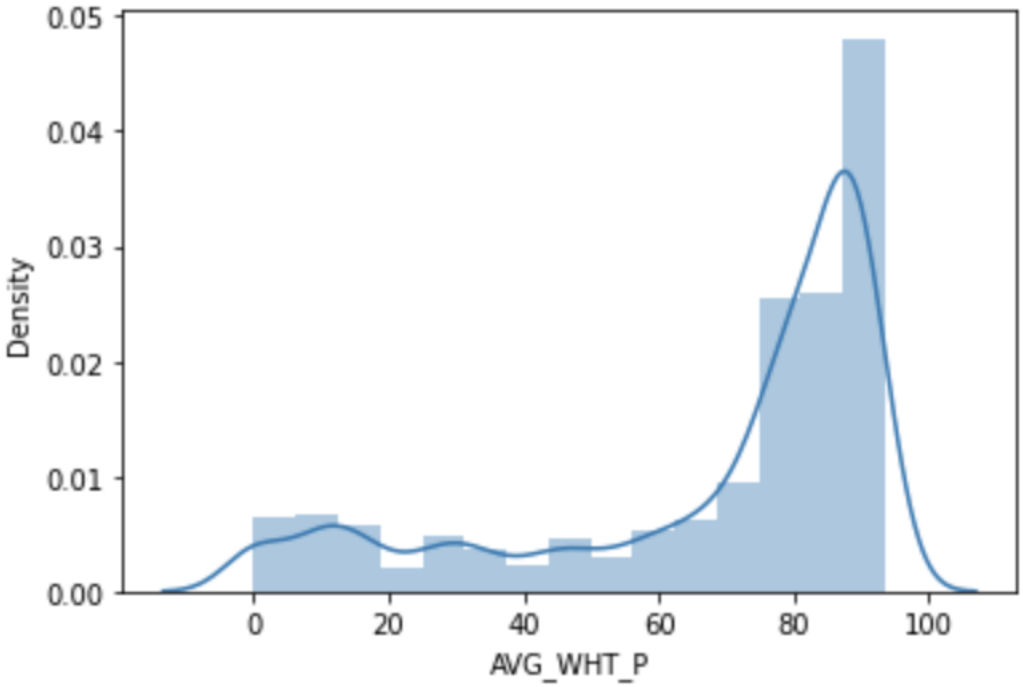

Figure 45 : Probability Density Function for AVG_WHT_P

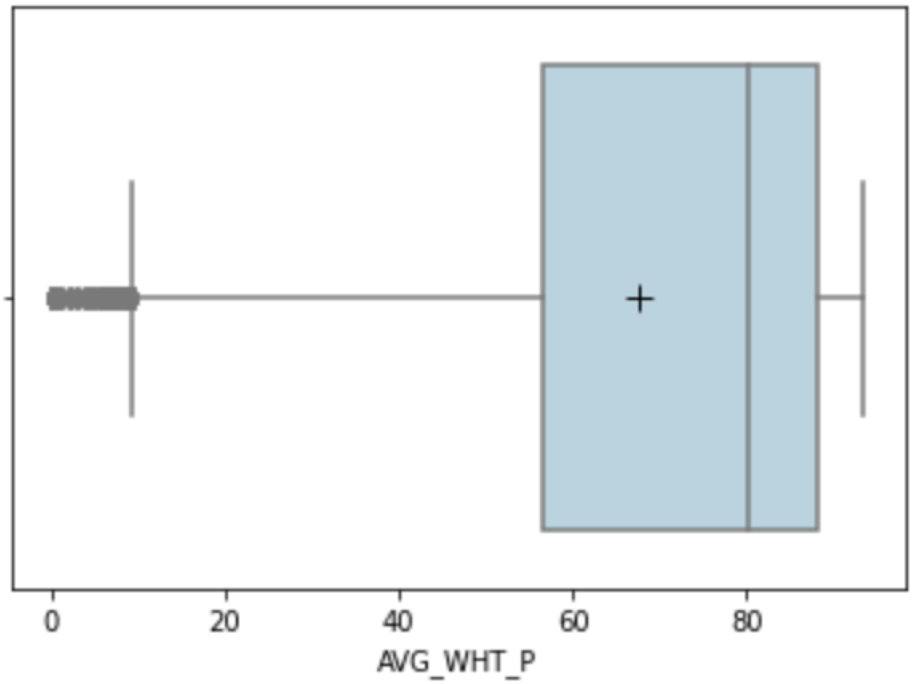

Figure 46 : Boxplot for AVG_WHT_P

```
count    9146.000000
mean       67.728440
std        27.719028
min         0.000000
25%        56.577834
50%        80.071250
75%        88.062202
max        93.509584
Name: AVG_WHT_P, dtype: float64
```

Figure 47 : Describe AVG_WHT_P

Forward Filling

The probability density function for *AVG_WHT_P* changed slightly after the missing values were filled in by the Forward Filling Method as shown in Figure 48. Furthermore, the outliers were removed as well as shown in Figure 49. Figure 49 and Figure 50 also shows that the mean has changed to 66. Figure 50 shows that the standard deviation is 26, the lower quartile is 53, the upper quartile is 86 whereas the median is 78.

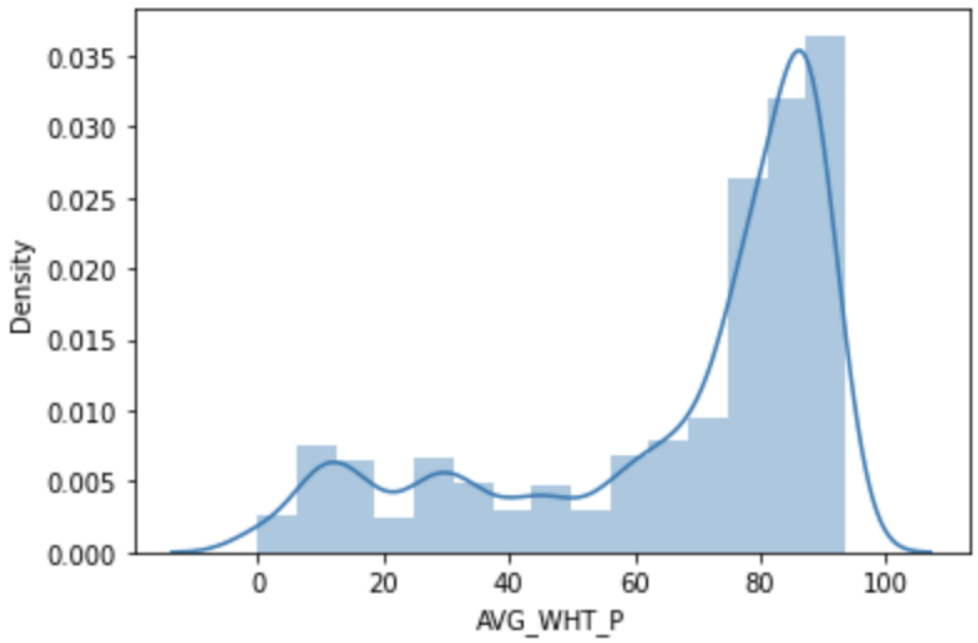

Figure 48 : Probability Density Function for AVG_WHT_P after Forward Filling

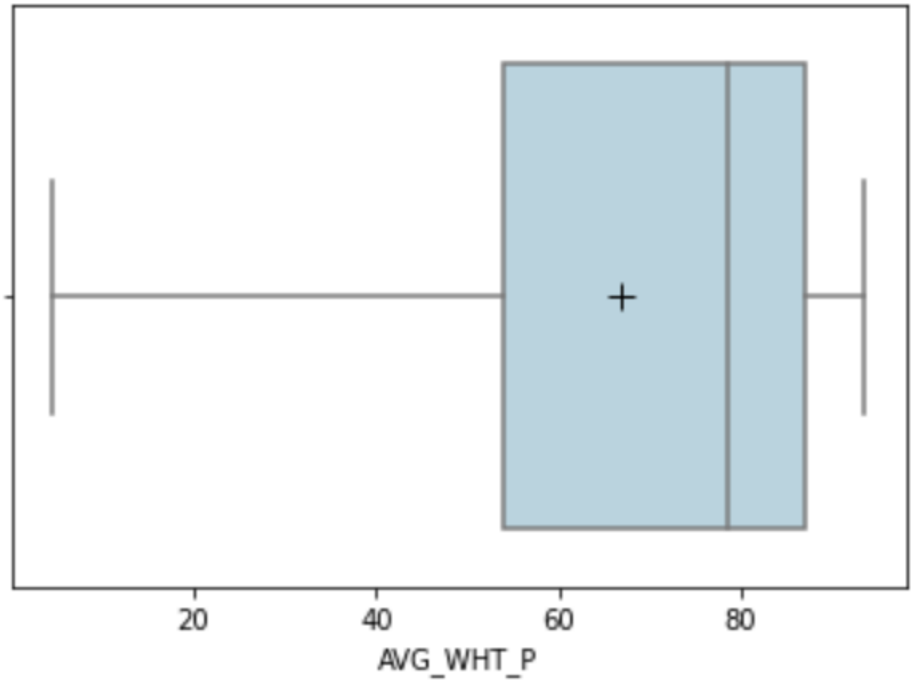

Figure 49 : Boxplot for AVG_WHT_P after removing outliers

```
count    6666.000000
mean       66.775520
std        26.284164
min         0.000000
25%        53.937156
50%        78.622468
75%        86.911547
max        93.509584
Name: AVG_WHT_P, dtype: float64
```

Figure 50 : Describe AVG_WHT_P

2) Kyle Master Dataset

a. Oil (m3)

Before Data Imputation

A probability density graph was plotted to understand the data distribution for Oil volume as shown in Figure 51. As can be seen from Figure 51, there is a significant

portion of the data that is 0. This claim is also backed by Figure 52 and Figure 53. Figure 52 also shows that a significant portion of the data is 0 and that there are several outliers. In addition to this Figure 53 shows that the lower quartile and median of the data is 0. As the goal of the model is to predict oil and gas production, it is not ideal to have a lot of 0 values for Oil volume. Therefore, these values should be dropped.

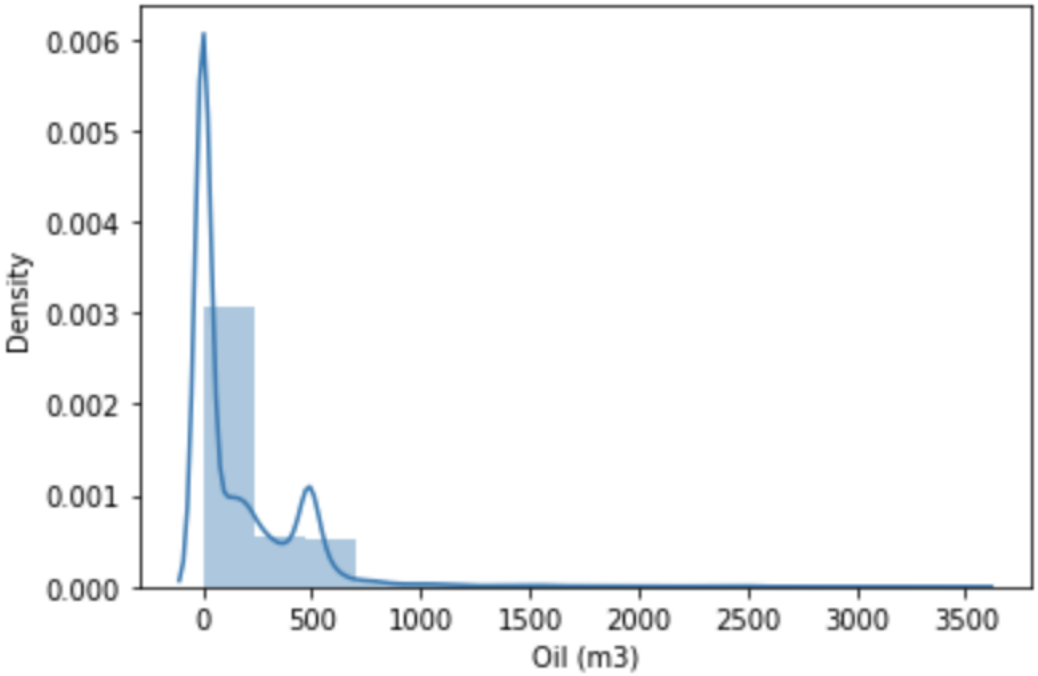

Figure 52 : Probability Density Function for Oil (m3)

```
count    27272.000000
mean       169.602447
std        278.551141
min          0.000000
25%          0.000000
50%          0.000000
75%        280.162080
max       3510.127311
Name: Oil (m3), dtype: float64
```

Figure 54 : Describe Oil (m3)

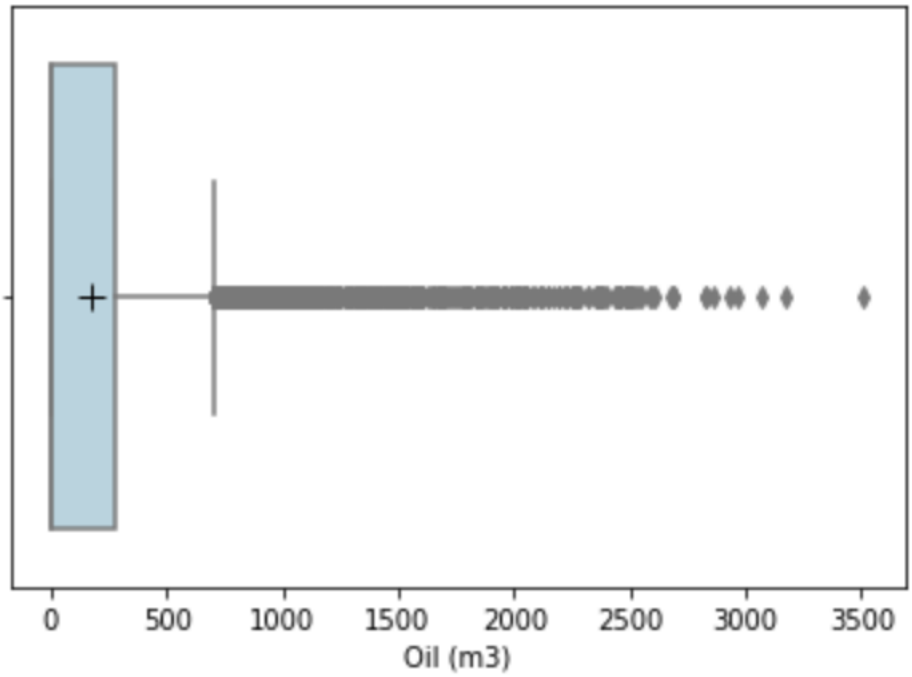

Figure 53 : Boxplot for Oil (m3)

Forward Filling

After filling in the missing values with the forward filling method and removing the 0 values from the data, there are some changes in the data distribution. Figure 54 shows the probability density function for the new distribution. As shown in Figure 55, the mean is now 371, the standard deviation is 311, the lower quartile is 167, the upper quartile is 486 and the median is 322. Additionally, the outliers have reduced as well. However, there are still some outliers, thus these outliers are removed as shown in Figure 56.

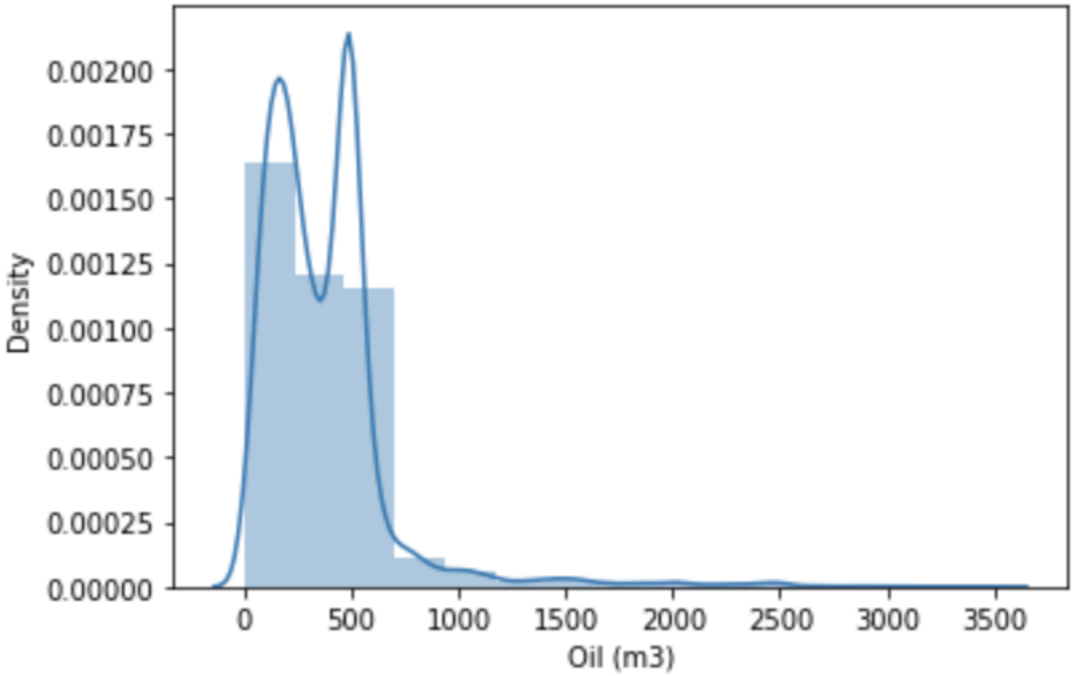

Figure 54 : Probability Density Function for Oil (m3) after Forward Filling

```
count    12271.000000
mean       371.189650
std        311.646243
min          0.000485
25%        167.738798
50%        322.623973
75%        486.254856
max       3510.127311
Name: Oil (m3), dtype: float64
```

Figure 55 : Describe Oil (m3)

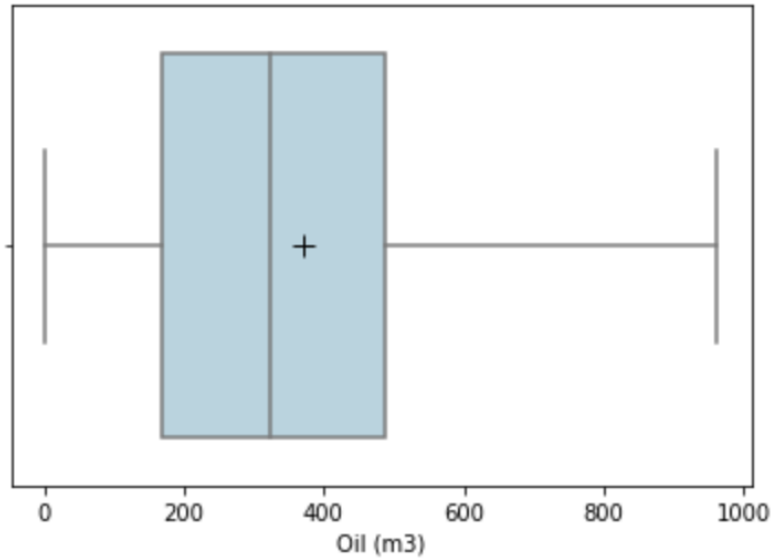

Figure 56 : Boxplot for Oil (m3) after Removing Outliers

b. Gas (m3)

Before Data Imputation

A probability density graph was plotted to understand the data distribution for Oil volume as shown in Figure 57. Figure 57 shows that a significant portion of gas volume is 0. Figure 58 also shows that a significant portion of the data is 0 and that there are several outliers. In addition to this Figure 59 shows that the lower quartile and median of the data is 0. As the goal of the model is to predict oil and gas production, it is not ideal to have a lot of 0 values for gas volume. Therefore, these values should be dropped.

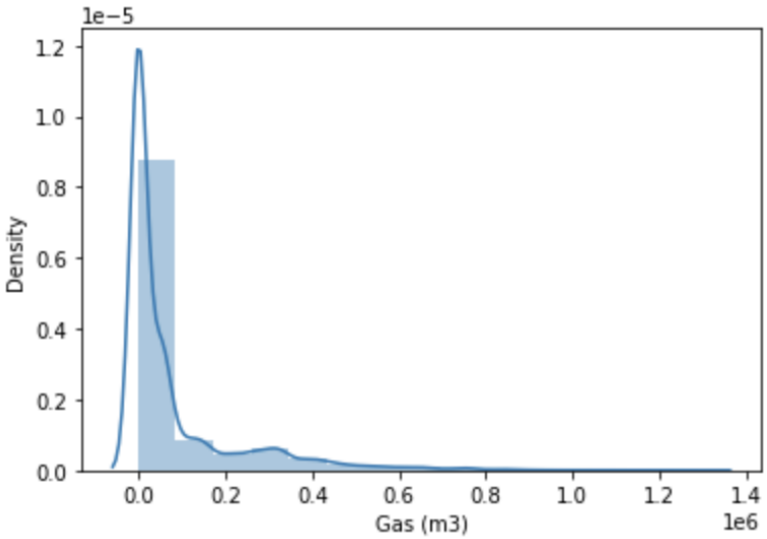

Figure 57 : Probability Density Function for Gas (m3)

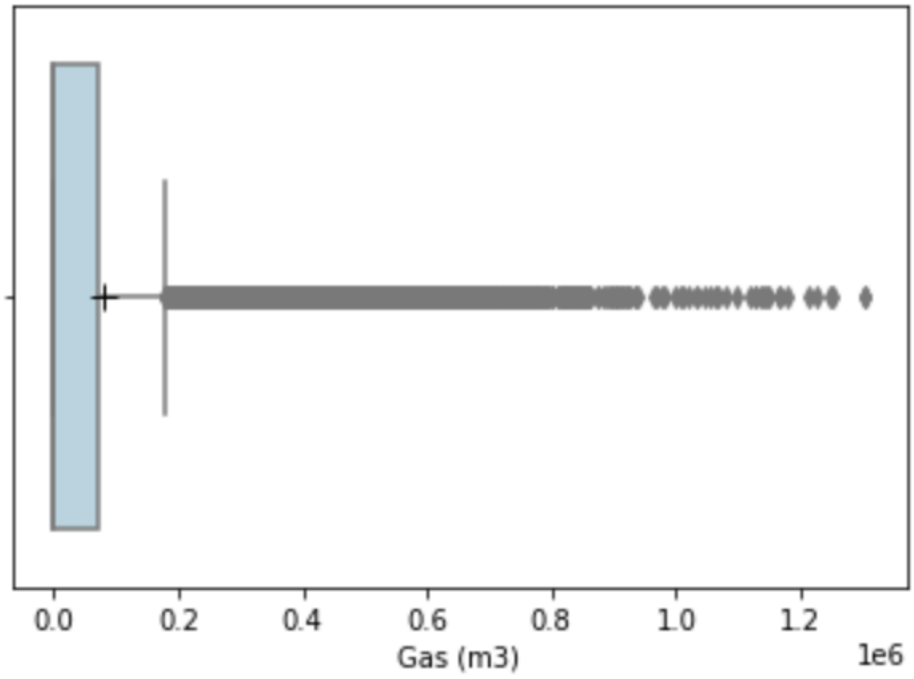

Figure 58 : Boxplot for Gas (m3)

```
count    2.727200e+04
mean     7.953906e+04
std      1.477763e+05
min     -1.670970e+03
25%      0.000000e+00
50%      0.000000e+00
75%      7.220739e+04
max      1.304321e+06
Name: Gas (m3), dtype: float64
```

Figure 59 : Describe Gas (m3)

Forward Filling

After filling in the missing values with the forward filling method and removing the 0 values from the data, there are some changes in the data distribution. Figure 60 shows the probability density function for the new distribution. As shown in Figure 61, the mean is now 173,8083, the standard deviation is 179,499, the lower quartile

is 54,420, the upper quartile is 175,105 and the median is 89,469. Additionally, the outliers have reduced as well. However, there are still some outliers, thus these outliers are removed as shown in Figure 62.

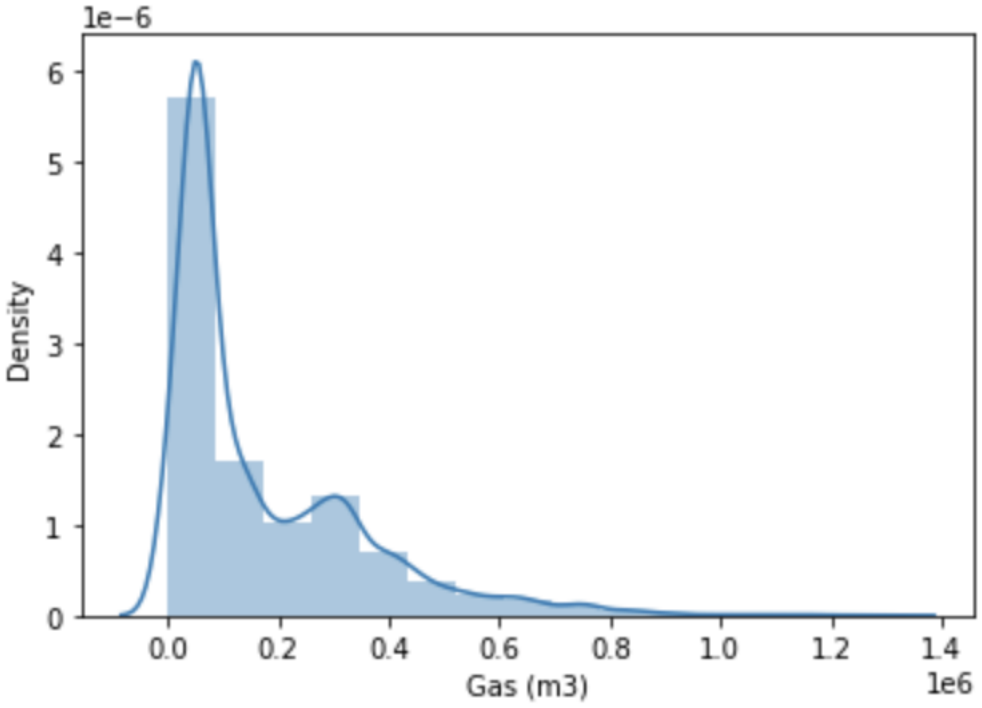

Figure 60 : Probability Density Function for Gas (m3)

```
count    1.200300e+04
mean     1.760399e+05
std      1.794431e+05
min      2.263758e+01
25%      5.442015e+04
50%      8.946996e+04
75%      2.751058e+05
max      1.304321e+06
Name: Gas (m3), dtype: float64
```

Figure 61 : Describe Gas (m3)

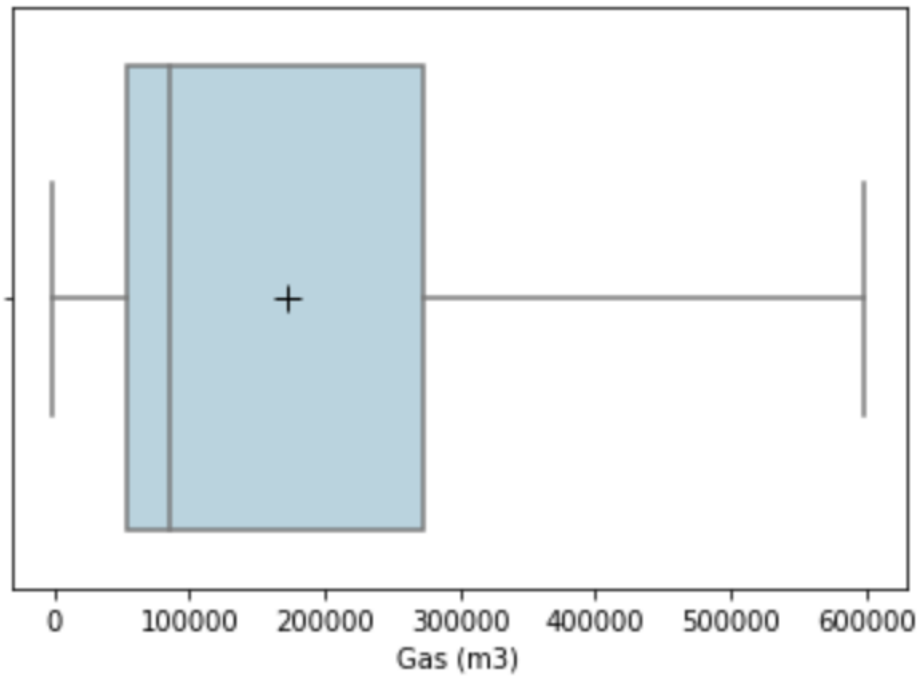

Figure 62 : Boxplot for Gas (m3) after Removing Outliers

c. Av. WHT (Deg C)

Before Data Imputation

A probability density function was made to understand the data distribution for this feature as shown in Figure 63. The distribution for *Av. WHT (Deg C)* is bimodal. As can be seen in Figure 64, the mean for this feature is approximately 30 and the distribution contains several outliers. Figure 65 shows that the mean is 28, the standard deviation 31, the lower quartile is 0, the median is 8 and the upper quartile is 60.

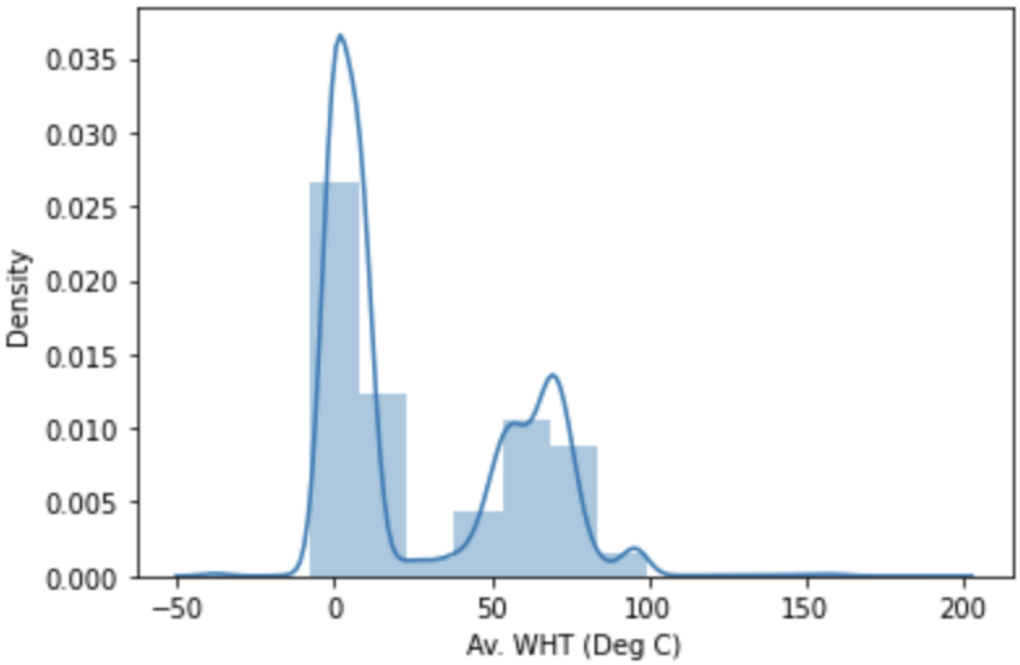

Figure 63 : Probability Density Function for Av. WHT (Deg C)

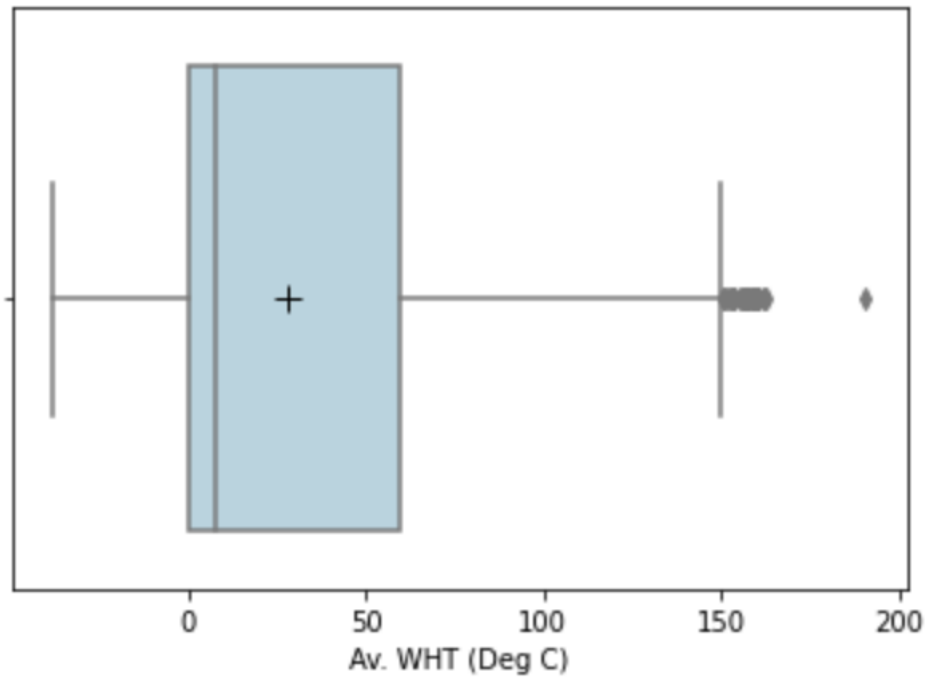

Figure 64 : Boxplot for Av. WHT (Deg C)

```
count    26964.000000
mean        28.328839
std         31.615280
min        -38.000000
25%          0.000000
50%          8.000000
75%         60.000000
max        190.453663
Name: Av. WHT (Deg C), dtype: float64
```

Figure 65 : Describe Av. WHT (Deg C)

Forward Filling

After filling in the missing values with the forward filling method, there are some changes in the data distribution. Figure 66 shows the probability density function for the new distribution. As shown in Figure 67, the mean is now 54, the standard deviation is 27, the lower quartile is 49, the upper quartile is 70 and the median is 60. However, there are still several outliers in the data as shown in Figure 62. Therefore, these outliers are removed as shown in Figure 63.

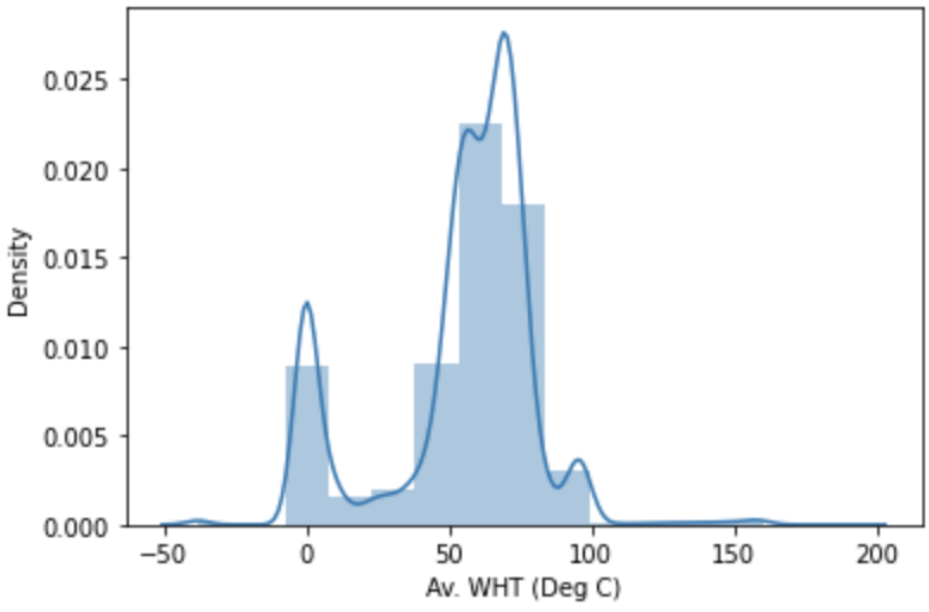

Figure 66 : Probability Density Function for Av. WHT (Deg C)

```
count    11896.000000
mean        54.374199
std         27.204354
min        -38.000000
25%         49.588200
50%         60.377025
75%         70.200419
max        190.453663
Name: Av. WHT (Deg C), dtype: float64
```

Figure 67 : Describe Av. WHT (Deg C)

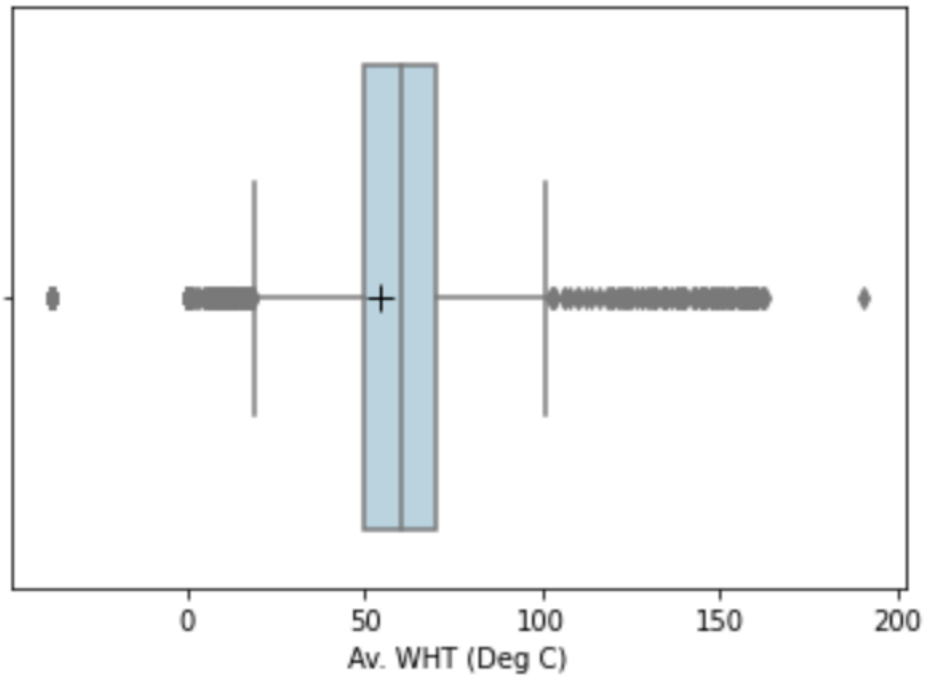

Figure 68 : Boxplot for Av. WHT (Deg C) after Forward Filling

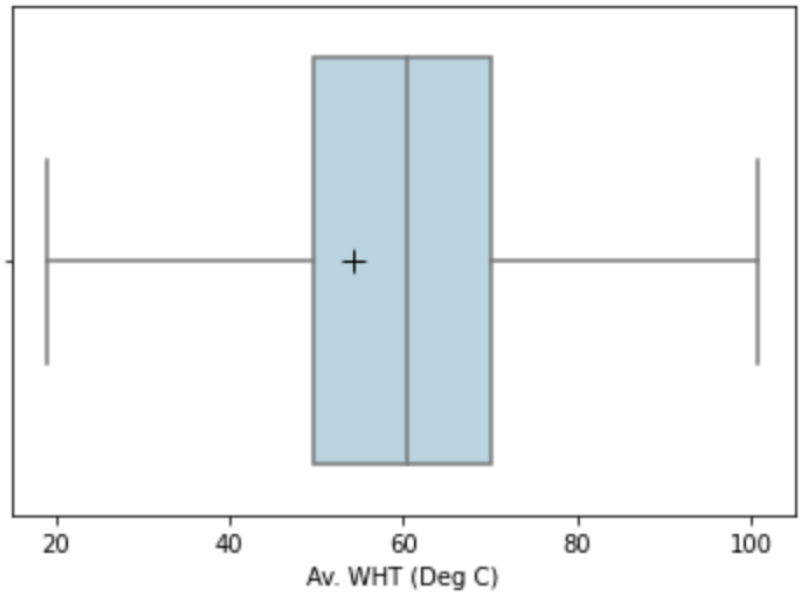

Figure 69 : Boxplot after Removing Outliers

d. Av. WHP (bar)

Before Data Imputation

A probability density function was made to understand the data distribution for this feature as shown in Figure 70. As can be seen in Figure 71, the mean for this feature is approximately 44 and the distribution contains few outliers. Figure 72 shows that the mean is 44, the standard deviation 37, the lower quartile is 0, the median is 44 and the upper quartile is 70.

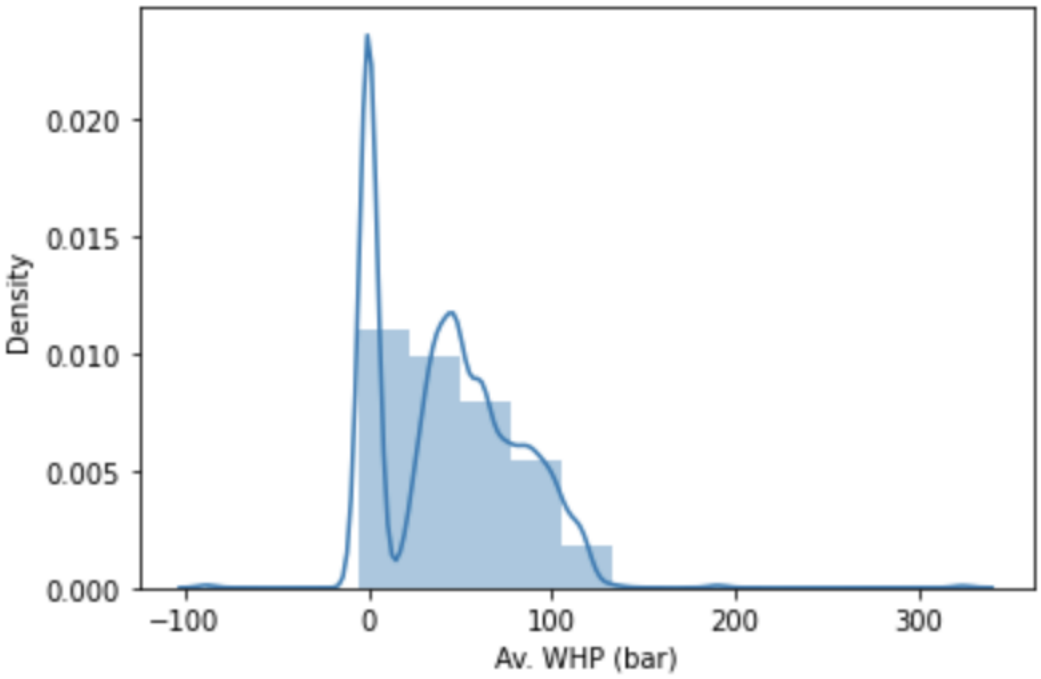

Figure 70 : Probability Density Function for Av. WHP (bar)

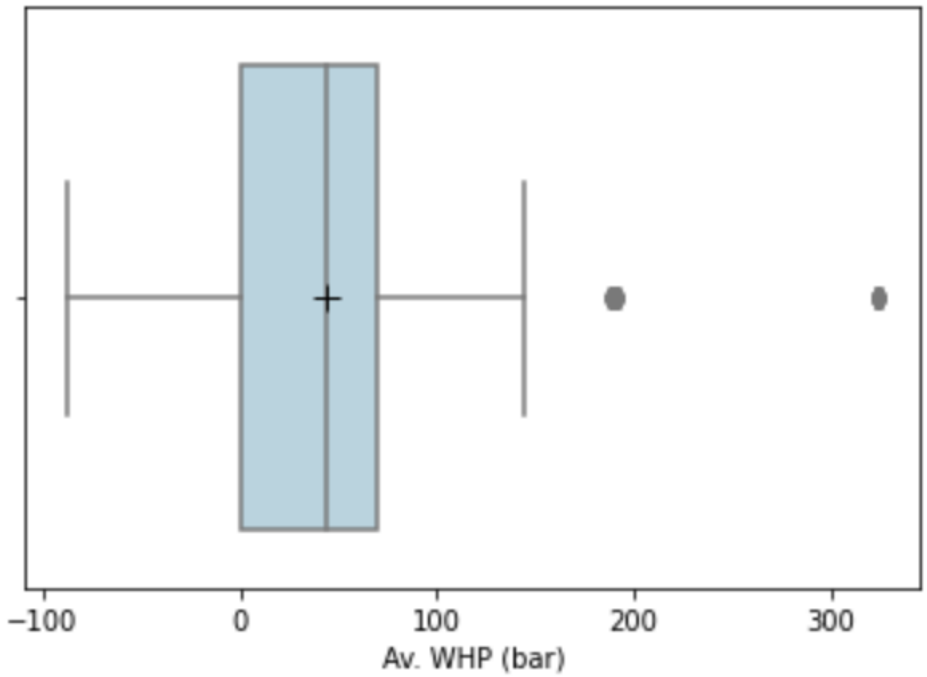

Figure 71 : Boxplot for Av. WHP (bar)

```
count    24927.000000
mean        44.197514
std         37.313958
min        -88.000000
25%          0.000000
50%         44.602351
75%         70.529973
max        325.000000
Name: Av. WHP (bar), dtype: float64
```

Figure 72 : Describe Av. WHP (bar)

Forward Filling

After filling in the missing values with the forward filling method, there are some changes in the data distribution. Figure 73 shows the probability density function for the new distribution. As shown in Figure 74, the mean is now 54, the standard deviation is 34, the lower quartile is 34, the upper quartile is 80 and the median is 48. Afterward, the outliers in the data were removed as shown in Figure 75.

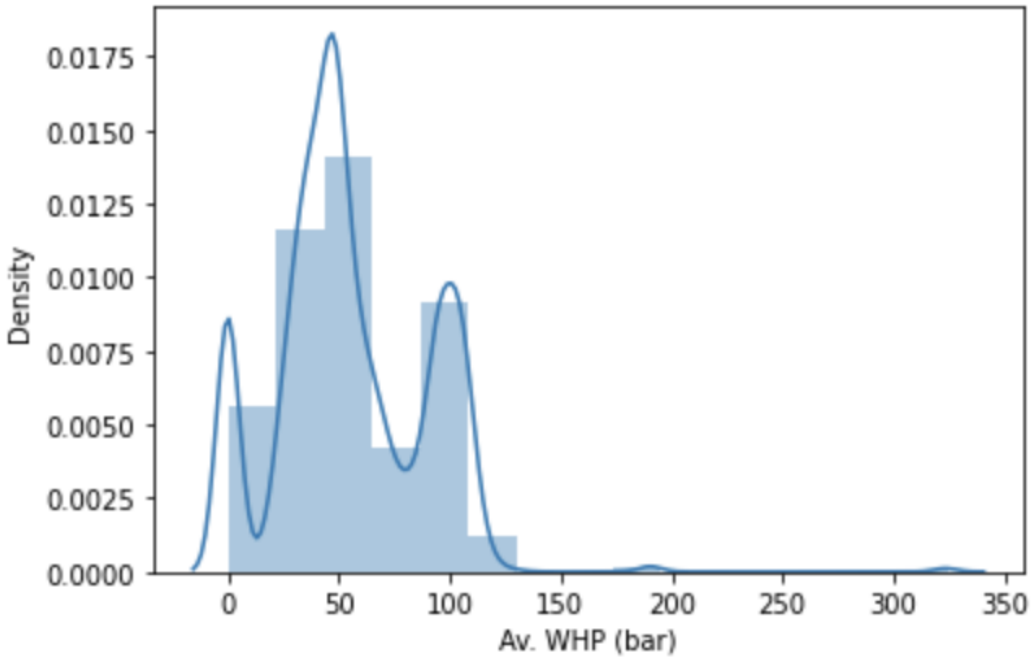

Figure 73 : Probability Density Function for Av. WHP (bar) after Forward Filling

```
count    12003.000000
mean        54.786040
std         34.095515
min          0.000000
25%         34.990782
50%         48.987094
75%         80.404520
max        325.000000
Name: Av. WHP (bar), dtype: float64
```

Figure 74 : Describe Av. WHP (bar)

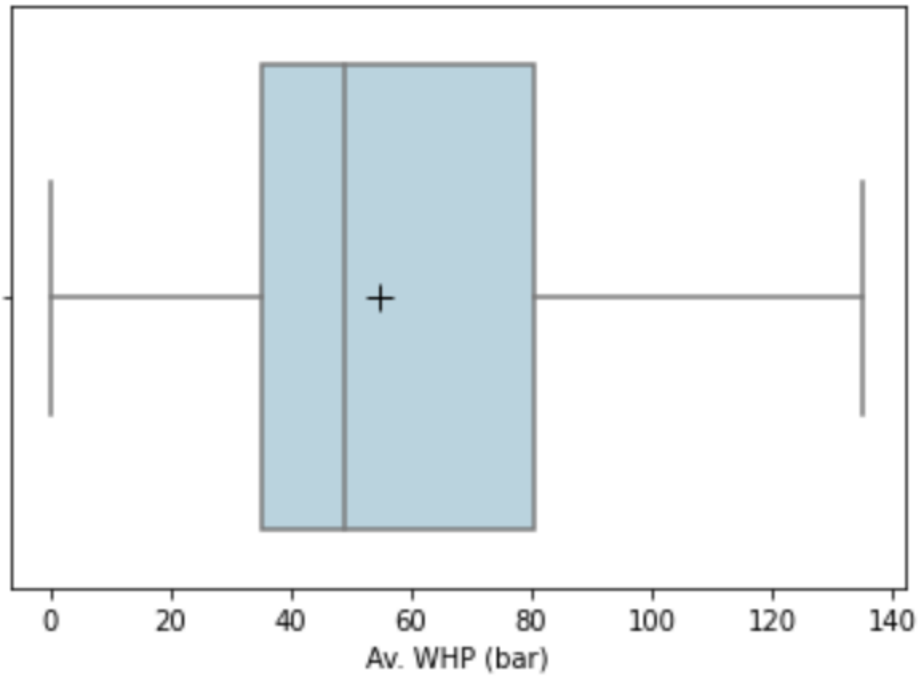

Figure 75 : Boxplot for Av. WHP (bar) after Removing Outliers

e. Av. DHT (Deg C)

Before Data Imputation

A probability density function and boxplot were plotted to help understand the data distribution as shown in Figure 76 and Figure 77. Both Figure 76 and Figure 77 shows that the data contains a lot of extreme outliers such as -2000, thus it has to

be removed so that the model can train more efficiently. Furthermore, a lot of the values in the feature are 0. As *Av. DHT (Deg C)* will be used to predict oil and gas production, these 0 values should be removed. In addition to this, Figure 78 shows that the mean is 47, the standard deviation is 148, the lower quartile is 0, the upper quartile is 96 and the median is 93.

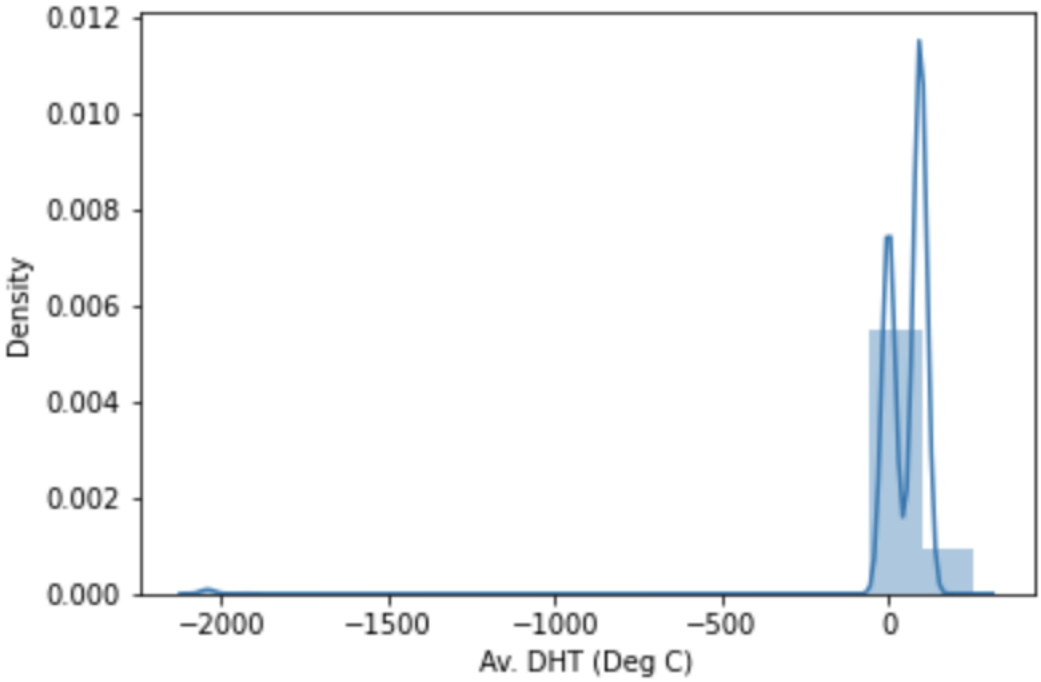

Figure 76 : Probability Density Function for Av. DHT (Deg C)

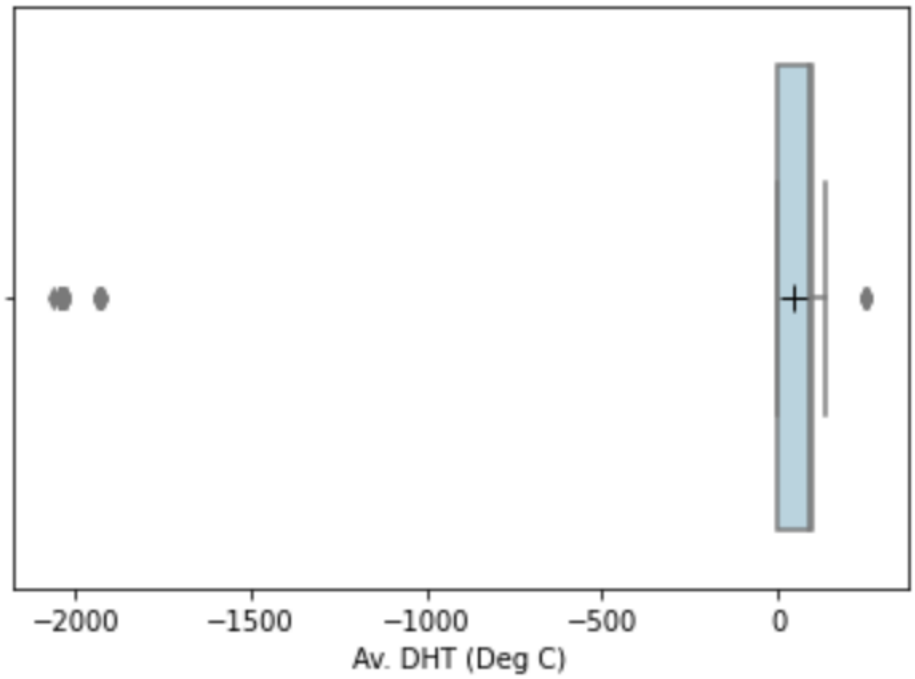

Figure 77 : Boxplot for Av. DHT (Deg C)

```
count    24571.000000
mean        47.969195
std        148.824021
min      -2064.766100
25%         0.000000
50%        93.000000
75%        96.000000
max       253.363450
Name: Av. DHT (Deg C), dtype: float64
```

Figure 78 : Describe Av. DHT (Deg C)

Forward Filling

After removing the outliers, 0 values, and filling in the missing values with the forward filling method, the data distribution changed. Figure 79 shows the probability density function of the new distribution. From Figure 80, it can be seen that the mean is now 96, the standard deviation is 2, the lower quartile is 94, the

median is 96 and the upper quartile is 99. From Figure 81, it can be seen that the range for *Av. DHT (Deg C)* is from 85 to 107.

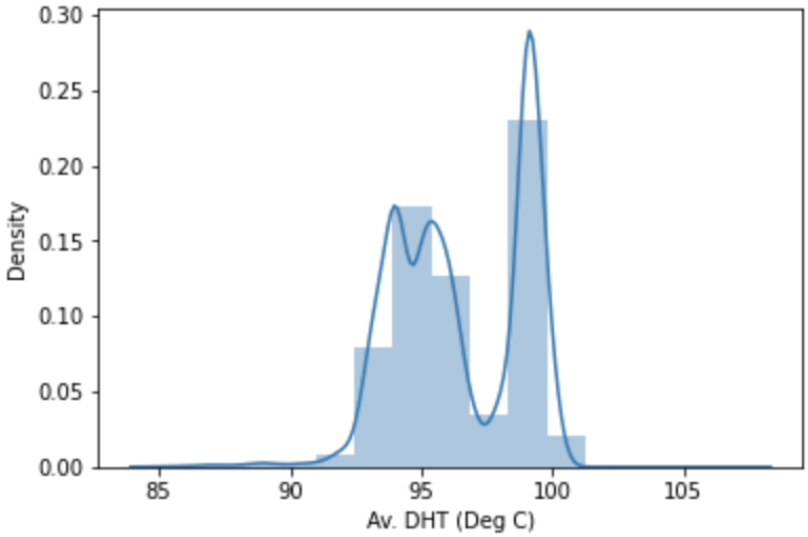

Figure 79 : Probability Density Function for Av. DHT (Deg C) after Forward Filling

```
count    9681.000000
mean       96.456462
std         2.457700
min        85.095579
25%        94.229950
50%        96.000000
75%        99.000000
max       107.126017
Name: Av. DHT (Deg C), dtype: float64
```

Figure 80 : Describe Av. DHT (Deg C)

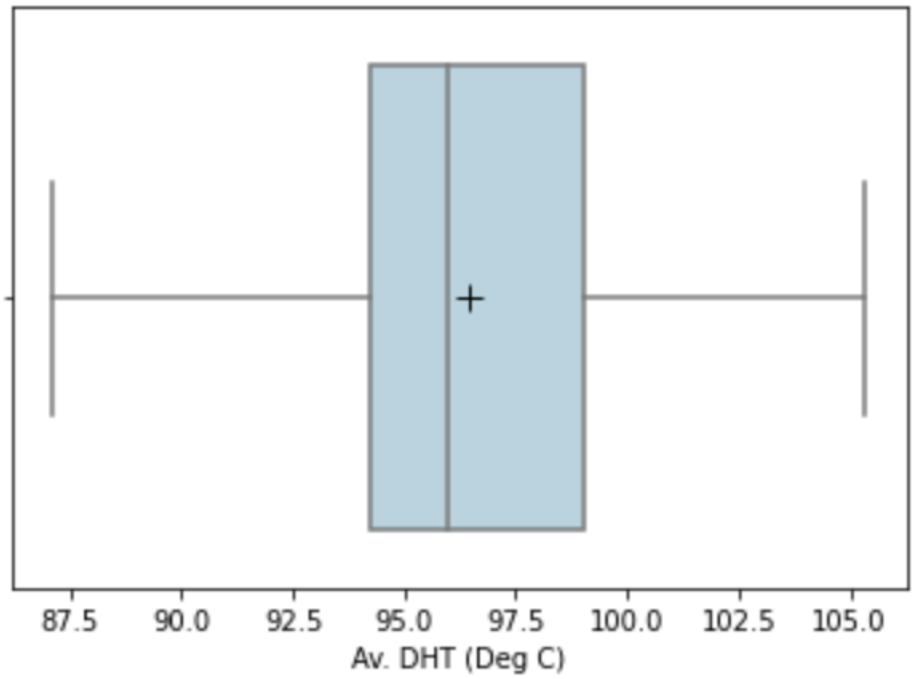

Figure 81 : Boxplot for Av. DHT (Deg C) after Removing Outliers

f. Av. DHP (bar)

Before Data Imputation

A probability density function and boxplot were plotted to help understand the data distribution as shown in Figure 82 and Figure 83. Both Figure 82 and Figure 83 shows that the data contains a lot of extreme outliers. Additionally, a lot of the values in the feature are 0. Since *Av. DHP (bar)* will be used to predict oil and gas production, these 0 values should be removed. These outliers need to be removed as they would heavily affect the model's performance.

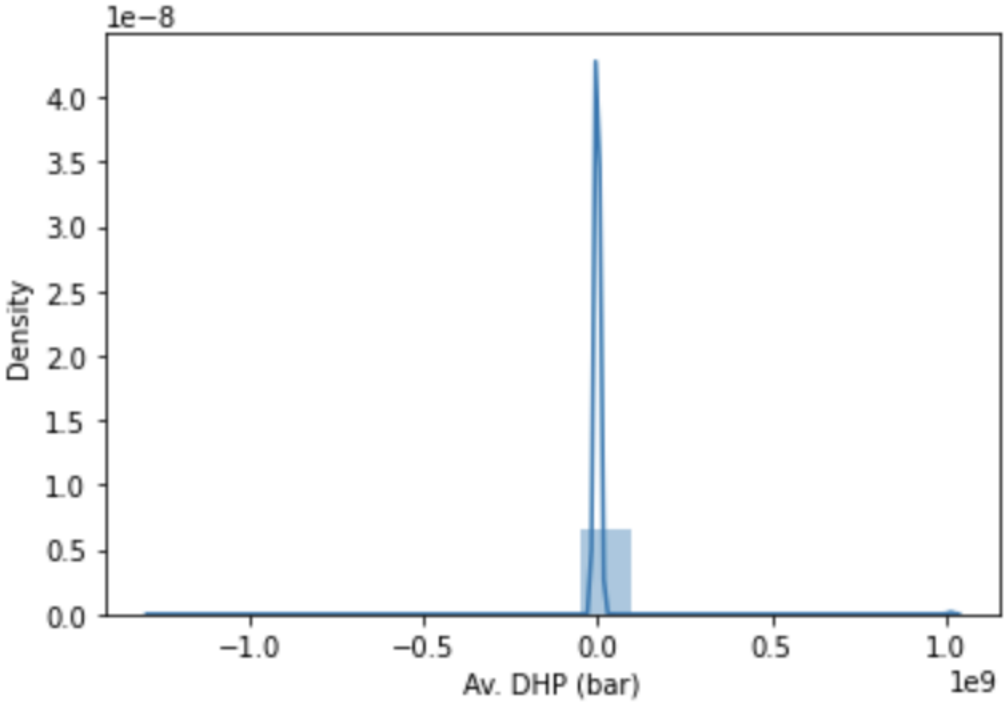

Figure 82 : Probability Density Function for Av. DHP (bar)

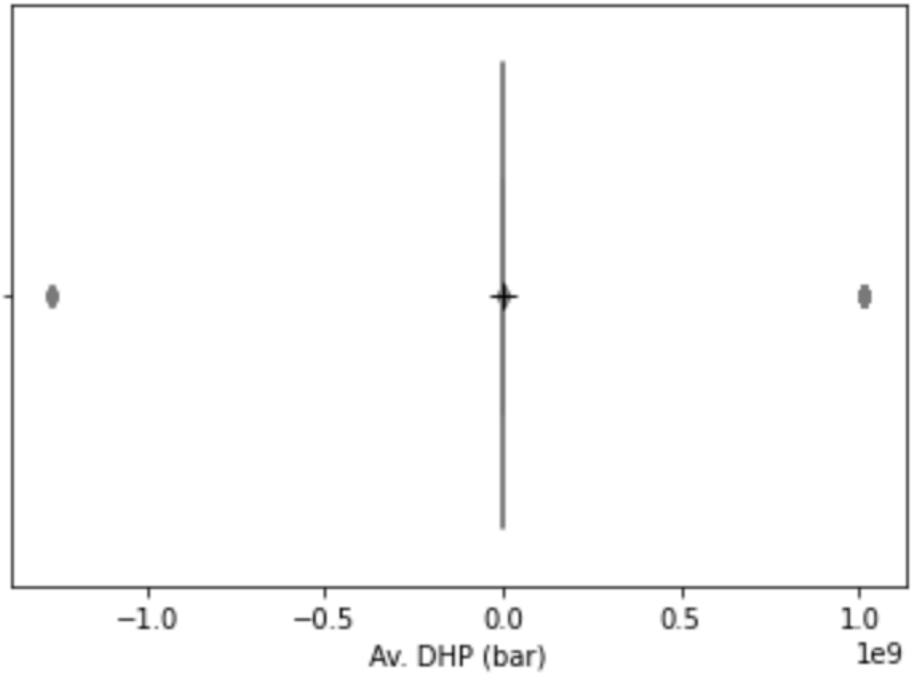

Figure 83 : Boxplot for Av. DHP (bar)

Forward Filling

After removing the outliers, 0 values, and filling in the missing values with the forward filling method, the data distribution has changed. Figure 84 shows the probability density function of the new distribution. From Figure 85, it can be seen that the mean is 107, the standard deviation is 33, the lower quartile is 83, the median is 94 and the upper quartile is 135. From Figure 86, it can be seen that the range for *Av. DHP (bar)* is from 9 to 210.

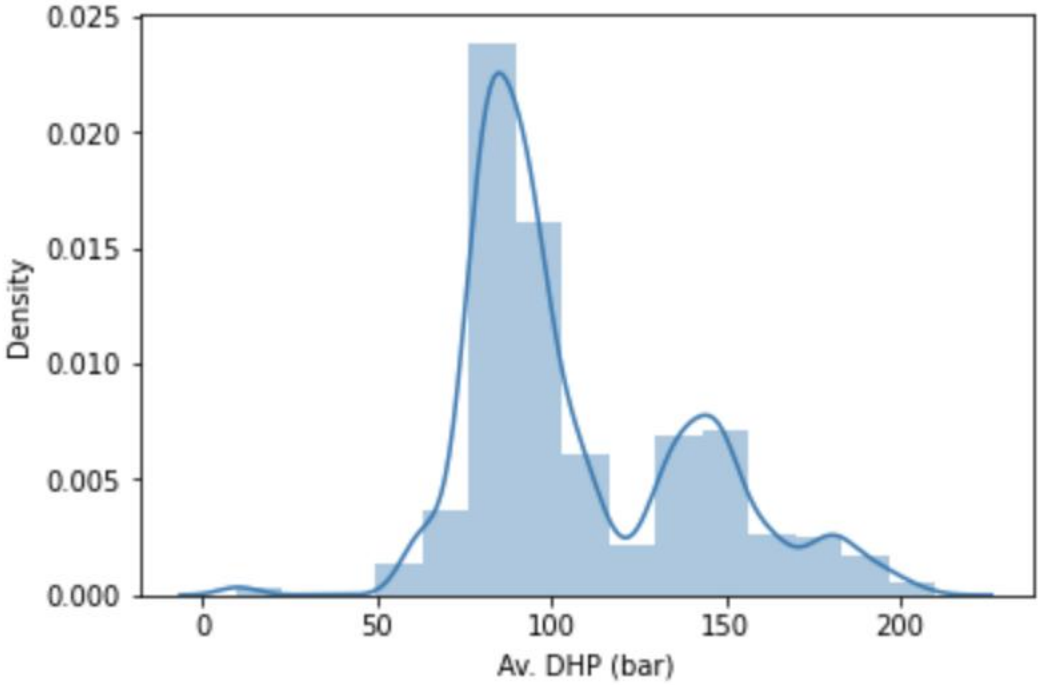

Figure 84 : Probability Density Function for Av. DHP (bar) after Forward Filling

```
count    9681.000000
mean      107.961061
std        33.506797
min         9.790524
25%        83.000000
50%        94.802611
75%       135.685912
max       209.986357
Name: Av. DHP (bar), dtype: float64
```

Figure 85 : Describe Av. DHP (bar)

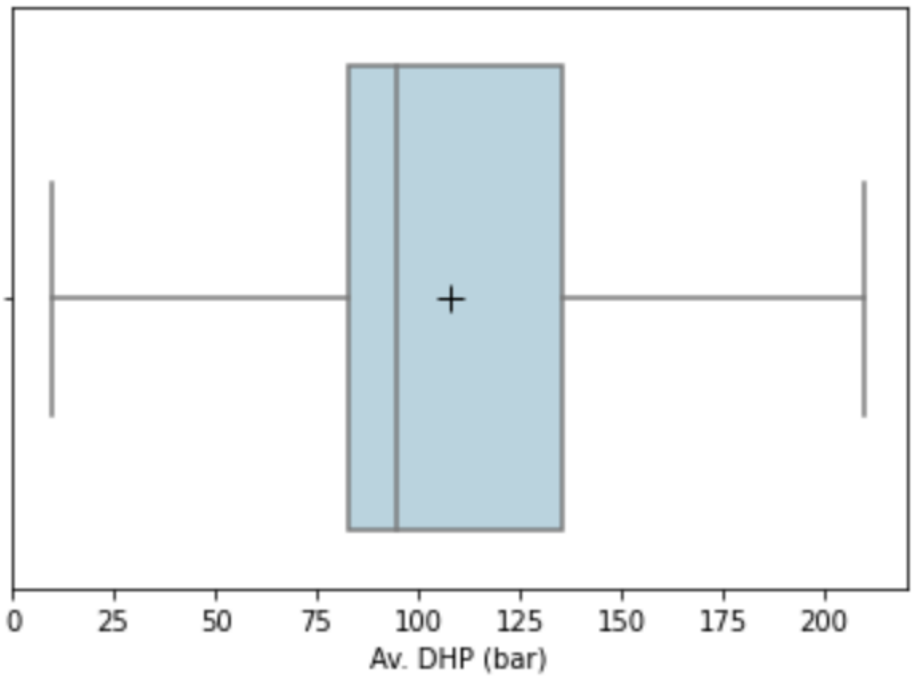

Figure 86 : Boxplot for Av. DHP (bar) after Removing Outliers