# 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***

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 I describes the observations derived from the heatmaps

***B) Data Imputation***

As has been mentioned in *Empty Data* section of this paper, both Volve and Kyle Master dataset contains 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 the dataset. Table II describes the features with the highest correlation values in Volve and Kyle Master dataset respectively when the respective data imputation methods are used. The feature correlation for Volve using mean imputation is not as strong as the feature correlation when forward filling is used. This could suggest that forward filling could be better than mean imputation for Volve dataset. However, the feature correlation for Kyle Master using forward filling is similar as when mean imputation is used. The Kyle Master dataset behaves similarly regardless of the imputation method whereas Volve shows better correlation values when forward filling is used. Thus, the ideal data imputation method would be forward filling.

***D) Feature Selection and Conversion***

This section will explain and justify which features will be used for the model's training and discuss regarding feature conversion. Table III shows the columns of the each dataset that have 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 value. In Volve, the first two features selected for model training are *BORE_OIL_VOL* and *BORE_GAS_VOL*.

*AVG_DOWNHOLE_PRESSURE* and *AVG_DOWNHOLE_TEMPERATURE* are also included as these features have a high correlation value. Additionally, oil and gas formation are also 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*. Oil and gas production can also be improved by water injection, therefore *AVG_WHT_P* is also added to the model's training. 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 they have decent correlation with *Oil (m3)* and *Gas (m3)*. Furthermore, as oil and gas production is 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. 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 $^{o}$C (degree Celsius), while the pressures will be standardized into *bar*, and the volumes will be standardized into $m^3$ (meter cubic).

*Table I : Observation for Missing Data*

| **Dataset** | **Volve** | **Kyle Master** |
|---|---|---|
| Observation | Contains mostly "<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 correlation value of 1, meaning most of the features do not show much correlation, however, there are few features which are highly correlated. |
| Missing Data Mechanism | Missing Not at Random (MNAR) | Missing at Random (MAR) |

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

| **Volve** | | | | **Kyle Master** | | | |
|---|---|---|---|---|---|---|---|
| **Forward Filling** | | **Mean Imputation** | | **Forward Filling** | | **Mean Imputation** | |
| Features | Correlation | Features | Correlation | Features | Correlation | Features | Correlation |
| BORE_OIL_VOL and BORE_GAS_VOL | 0.9988900091582152 | BORE_OIL_VOL and BORE_GAS_VOL | 0.9985636964673588 | Oil (m3) and Gas(m3) | 0.42758409089215 | Oil (m3) and Gas(m3) | 0.4269383277294414 |
| AVG_DOWNHOLE_PRESSURE and AVG_DOWNHOLE_TEMPERATURE | -0.8448797841191023 | AVG_DOWNHOLE_PRESSURE and AVG_DP_TUBING | 0.6967196755738765 | Av. DHT (Deg C) and Av. DHP (bar) | -0.5957641957 107336 | Av. DHT (Deg C) and Av. DHP (bar) | -0.5957641957 107336 |
| AVG_WHT_P and AVG_WHP_P | 0.6770906191279169 | AVG_WHT_P and BORE_WAT_VAL | 0.6743768856016769 | Av. WHT (Deg C) and Oil (m3) | 0.4998663929885465 | Av. WHT (Deg C) and Oil (m3) | 0.4998663929885465 |

*Table III : Column Headings in Volve and Kyle Master Dataset*

| **Volve Dataset** | **Kyle Master Dataset** |
|---|---|
| DATEPRD | Date |
| WELL_BORE_CODE | Wellbore ID |
| NPD_WELL_BORE_CODE | |
| NPD_WELL_BORE_NAME | |
| NPD_FIELD_CODE | |
| NPD_FIELD_NAME | |
| NPD_FACILITY_CODE | |
| NPD_FACILITY_NAME | |
| ON_STREAM_HRS | Hours Online |

| | |
|---|---|
| AVG_DOWNHOLE_PRESSURE | Av. DHP (bar) |
| AVG_DOWNHOLE_TEMPERATURE | Av. DHT (Deg C) |
| AVG_DP_TUBING | |
| AVG_ANNULUS_PRESS | |
| AVG_CHOKE_SIZE_P | Platform Choke % |
| AVG_CHOKE_UOM | |
| AVG_WHP_P | Av. WHP (bar) |
| AVG_WHT_P | Av. WHT (Deg C) |
| DP_CHOKE_SIZE | |
| BORE_OIL_VOL | Oil (m3) |
| BORE_GAS_VOL | Gas (m3) |
| BORE_WAT_VOL | Produced Water (m3) |
| BORE_WI_VOL | |
| FLOW__KIND | |
| WELL_TYPE | |

### *E) 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 IV describes the selected features of the Volve dataset whereas Table V describes the selected features for the Kyle Master dataset. In Table IV and Table V, range denotes the range of the specified feature, more specially it is 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.

*Table IV : Feature Statistics for Volve Dataset*

| Feature | Range | Outlier Count | Mean | Standard Deviation |
|---|---|---|---|---|
| AVG_DOWNHOLE_PRESSURE | 1 – 345 bar | 144 | 242 bar | 27 bar |
| AVG_DOWNHOLE_TEMPERATURE | 0.3 – 108 ºC | 156 | 104 ºC | 4 ºC |
| BORE_OIL_VOL | 1 – 5901 $m^3$ | 283 | 1458 $m^3$ | 1463 $m^3$ |
| BORE_GAS_VOL | 250 – 85113 $m^3$ | 182 | 212937 $m^3$ | 207073 $m^3$ |
| AVG_WHP_P | 0 – 120 bar | 44 | 48 bar | 20 bar |
| AVG_WHT_P | 7 – 93 ºC | 352 | 73 ºC | 18 ºC |

*Table V : Feature Statistics for Kyle Master Dataset*

| Feature | Range | Outlier Count | Mean | Standard Deviation |
|---|---|---|---|---|
| Av. DHP (bar) | 8 – 1130 bar | 3 | 111 bar | 39 bar |
| Av. DHT (Deg C) | 8 - 253 ºC | 645 | 94 ºC | 9 ºC |
| Oil (m3) | 1 – 3510 $m^3$ | 447 | 380 $m^3$ | 328 $m^3$ |
| Gas (m3) | 22.637.580 - 1.304.321.000.000 $m^3$ | 226 | 178.525.800.000 $m^3$ | 175.599.300.000 $m^3$ |
| Av. WHP (bar) | -38 – 190 ºC | 597 | 62 ºC | 18 ºC |
| Av. WHT (Deg C) | 0 – 325 bar | 48 | 57 bar | 35 bar |