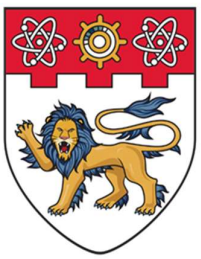


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AY2023/24 SEMESTER 1

BC2406- ANALYTICS I

Title:

Analytics for Sustainable Methane Management in Oil and Gas Operations:

A Proof of Concept for Saudi Aramco

Tutor: Mr Kevin Ngui

Seminar Class: S07

Group: 04

Date of Submission: 04/11/2023

Group members	**Matriculation Number**
Chan Kit Ho	
Chan Yi Hong	
Justinian Santos	
Tan Ji Rong	

Table of Content

Executive Summary

Methane — a potent greenhouse gas with significant contributions to global warming, has drawn increased attention from environmentalists, policymakers, and industry leaders alike. As part of the global effort to mitigate climate change, Saudi Aramco similarly seeks to proactively align with international sustainability goals and reduce its environmental footprint.

This report serves as a Proof of Concept for Aramco: where we develop an analytical framework to tackle methane emissions through data-driven predictive modelling. The goal is to identify and implement strategies that are environmentally beneficial and economically viable.

Through statistical analysis of historical data and advanced modelling techniques, such as Linear Regression, Classification and Regression Trees (CART), and Random Forest, our approach aims to forecast methane emissions and outline the actionable steps for the reduction. We hence propose a three-tiered approach to methane emission management:

1. Assessing Acceptability of Predicted Emission Level: Benchmarking to ensure alignment with industry standards and regulatory expectations.

2. Tackling Significant Variables Through Actionable Steps: Identifying and devising plans to act on significant factors.

3. Scenario Testing with Predictive Modelling: Utilising the models to test various intervention scenarios and assessing the impact on emissions, allowing for optimization of strategies.

Key findings indicate that technological upgrades, maintenance of older infrastructure, and deployment of efficient equipment like combustors are critical to reducing methane emissions.

The proposed solution seeks to offer Aramco a dual advantage: compliance with emerging global standards and an opportunity to reinforce its place in sustainable energy practices. By reducing emissions, Aramco can potentially lower its cost of capital and enhance shareholder value, proving that environmental sustainability and economic success are not mutually exclusive but are, in fact, synergistic.

1. Introduction

1.1 Background

Methane is a greenhouse gas that contributes to global warming. In fact, Methane has a stronger warming power when compared to other greenhouse gases such as Carbon Dioxide once it reaches the atmosphere *(Methane: A crucial opportunity in the Climate Fight, n.d.)*. With too much methane, it will lead to the depletion of our ozone layer, this will lead to harmful ultraviolet (UV) radiation from the sun to reach Earth, increasing the risk of skin cancer and other health problems *(Methane emissions are driving climate change. Here's how to reduce them., 2021)*.

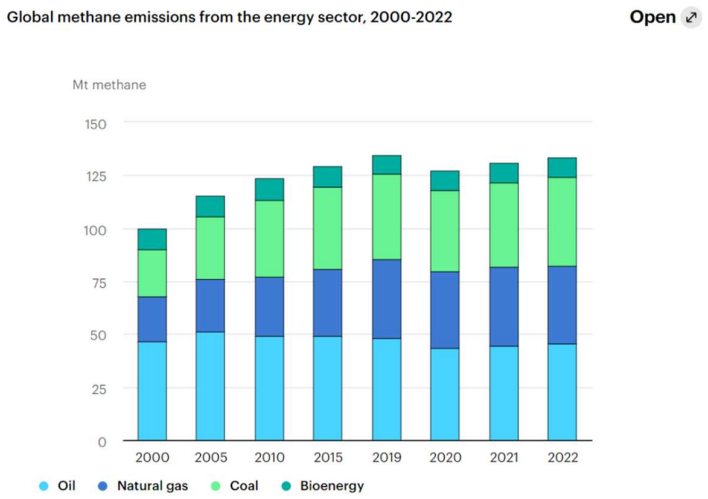


Figure 1: Global Methane Emissions from Energy Sector, 2000-2022

(IEA, 2022)

The global energy sector was responsible for nearly 135 million tonnes of methane emissions in 2022, which is about 40% of total methane emissions attributable to human activity, second only to agriculture *(Overview – global methane tracker 2023 – analysis, n.d.)*.

Furthermore, countries that made up 45% have pledged to the Global Methane Pledge that was launched at COP26 to reduce their methane emissions by 30% by 2030 (The global methane pledge – global methane tracker 2022 – analysis, n.d.). Aramco is a state-owned company in Saudi Arabia, and they are still seeking to reduce their methane emissions to sustainably increase shareholder and societal values. Reducing methane emissions is one of the most cost-effective methods to reduce short-term global warming.

1.2 Problem Statement

In the oil and gas industry, drilling, production, and their other operations has led to the release of methane, posing environmental threats and compliance challenges. These methane emissions contribute to climate change, as it is a greenhouse gas, and can potentially harm Saudi Aramco's reputation.

Our objective is to utilise data analytics and modelling techniques to devise effective methods to reduce methane emission, ensure regulatory compliance, and maintain operational efficiency and competitiveness in the industry.

1.3 Current Situation

The United States has passed the Inflation Reduction Act in 2022 which has a provision on the Methane Emissions Reduction Program, aimed to reduce methane emissions from oil and gas companies through a fee on excessive emissions over 25,000 tonnes of methane. The fee starts at $900 per ton of methane in 2024 and will increase to $1,500 per ton of methane by 2026 *(Inflation reduction act methane emissions charge: In brief - CRS reports, 2022)*. These fees could be expected to be implemented in other countries in the future as well.

Saudi Aramco takes sustainability into consideration by utilising methane that would have been flared into domestic use to generate power. Saudi Aramco also has developed Leak Detection And Repair (LDAR) programs to minimise and attend to leakages as quickly as possible *(Saudi Aramco, 2023).*

1.4 Benefits to Methane Abatement

Methane emissions from human activities account for more than half of global methane. 95% of these emissions come from the following 3 sectors: *(Kuylenstierna et al., 2021).*

- 35% from fossil fuels (23% from oil and gas, and 12% from coal)
- 20% from waste (including the waste sector, landfills and waste water)
- 40% from agriculture (32% from livestock, and 8% from rice cultivation)

Through implementing measures to reduce methane emissions, it is predicted that emissions can be reduced by almost 45% by 2030, translating to reduction of nearly 0.3°C of warming after 2040.

Furthermore, reducing ozone concentrations by reducing methane emissions could prevent the loss of 26 million tonnes of staple crops such as wheat, maize, soybeans, and rice yearly *(Kuylenstierna et al., 2021)*. The potential loss is estimated to be 1-2% of the global yield in 2020. There is an increasing demand for food to be grown to feed the growing global population. Thus, the benefit from reducing methane emissions is very important for food sustainability in the future.

2. Methodology

2.1 Approach

The approach we plan to take uses a model based on statistical analysis from historical datasets to predict methane emissions from drilling. After exploring and cleaning the dataset we will train models like linear regression, CART and Random Forest to help predict methane emissions. We will be training these models using a 70-30 train-test split to ensure our model is not overfitted to the data. We will then compare the results of all our model's predictions to the actual methane emission for accuracy. Afterwards we will consider the suitability and accuracy of the model before finalising which model we feel is best fitted to predict methane emissions. Lastly we will be using these predictions from the model to create actionable plans to reduce emissions from drilling.

2.2 Feasibility

2.2.1 Advantages

Our model uses a set of variables from drilling to accurately predict methane emission. For example, Age of well, Wellhead Pressure etc. These variables used in our model for predictions are needed in every drilling procedure meaning that they are always measured and tracked since they are required in every drilling. This consistent availability of variables allows our model to make more consistent and accurate predictions about methane emissions helping to reduce uncertainty and variance in predictions. This consistency and certainty will also help Aramco in their quest to reduce methane emission.

2.2.2 Limitations

One potential limitation of our model is that though it has the potential to make accurate predictions, it does not equate to reduction in methane emissions during drilling. This is extremely prominent to variables that cannot be changed. For example, the age of drilling infrastructure is fixed and cannot be changed without high economic costs or disruption in operation. These immutable factors will constraint the reduction of methane using our model. Hence Aramco must find a balance between economic benefits and sustainability efforts.

2.3 Desired Outcomes

The most important outcome we have planned for our model is to reduce methane emissions. By doing so, Aramco can reduce its carbon footprint while also helping to combat climate change. This reduction in methane will help Aramco economically by reducing the amount paid for carbon tax from the incoming US Inflation Reduction Act 2022. We also hope that this model can help Aramco with their pledge to the Global Methane Pledge and their other environmental sustainability values. All in all, we hope that our model can contribute to sustaining the environment while also helping Aramco.

3. Data Preparation, Cleaning and Data Exploration

In this section, we detail our data cleaning process, aimed at retaining valuable variables while eliminating any inaccuracies and anomalies. Exploration of the dataset is conducted both prior to and after the cleaning process to ensure thoroughness and integrity of the whole process.

3.1 Data Preparation

For this report, we will be using data from a Methane Emission Study done by University of Texas (Allen et al., 2013). The study provides two key datasets:

- final_SITES.csv
- final_SOURCES.csv

Our preliminary data preparation steps involve merging these two datasets using Site Name as the unique identifier. This was done using Excel. After merging, we aggregated the methane emission data from final_SOURCES.csv, calculating the total emissions attributed to each site. This consolidated dataset will then serve as the foundation for our subsequent analysis.

3.2 Data Cleaning

Through a preliminary data exploration, we noted that several data cleaning measures were deemed necessary:

3.2.1 Outliers

We found that there exist outliers in the dataset, for 4 of the variables (as seen in Appendix A). Outliers can significantly skew the analysis and since the number of outliers we found is small, 16/1079 ≈ 1.5%, we decided to remove the rows with outliers.

3.2.2 Missing Data Handling

We found that there are 78 missing values in some of the data fields, and missing data can reduce the accuracy of models, since some models auto ignore rows with NA fields.
As such, to tackle this, we took on 2 approaches: (more details in Appendix A)

1. For Continuous Variables: We used the median of the column to estimate the missing values. The choice of median is in line with the variables having whole integer values.
2. Categorical Variables: We performed logistic regression modelling and input the predicted class labels obtained from the model to estimate the missing value.

3.2.3 Categorical Variables Preparation

To facilitate interpretation and modelling in later sections, categorical variables were converted into numerical formats, making them suitable for algorithmic analysis. We factored the categorical variables to numeric levels so that it will be easier to interpret the result of our models.

3.2.4 Duplicated Rows

There were no duplicated rows detected to handle in this dataset.

3.2.5 Data Conversion

Before starting the analysis, we note that our methane emission data is in cubic feet per minute, while regulations are in metric ton per year. To address this, we will incorporate a conversion formula, which we detail in section 5.1 of our business application.

3.3 Data Exploration

3.3.1 Current Emission Status

Upon analysing the data, we found that, based on the conversion of the emissions threshold, a significant 59.8% of the wells (as depicted in Figure 2) are emitting methane at levels surpassing the established regulatory threshold. This finding raises substantial concerns regarding potential fines that Aramco may face due to non-compliance.

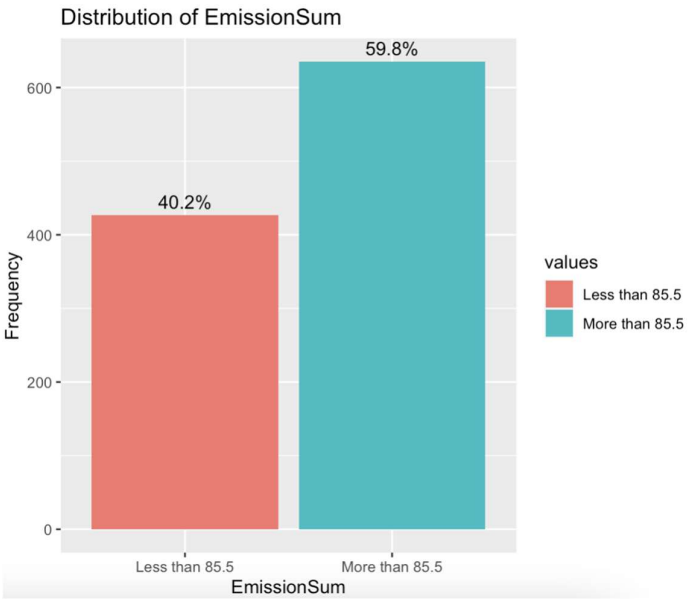


Figure 2: Distribution of EmissionSum Exceeding Threshold

3.3.2 Redundant Variables

During our examination of the dataset, we observed that the 'Fracked' variable displayed a uniform value of 'Yes' across all entries. Since a variable with a single level does not vary and contributes no informational value to our analysis or predictive models, we decided to exclude it from further consideration.

```
$ Fracked?                    : Factor w/ 1 level "Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Figure 3: Distribution of Fracked Variable

3.3.3 Age of Wells

According to Planète Énergies (n.d.), the average operational lifespan of oil wells in the industry typically ranges from 15 to 30 years. However, our analysis reveals a notable deviation from this standard. The wells in our dataset exhibit a mean and median age of 5.46 and 5.48 years, respectively.

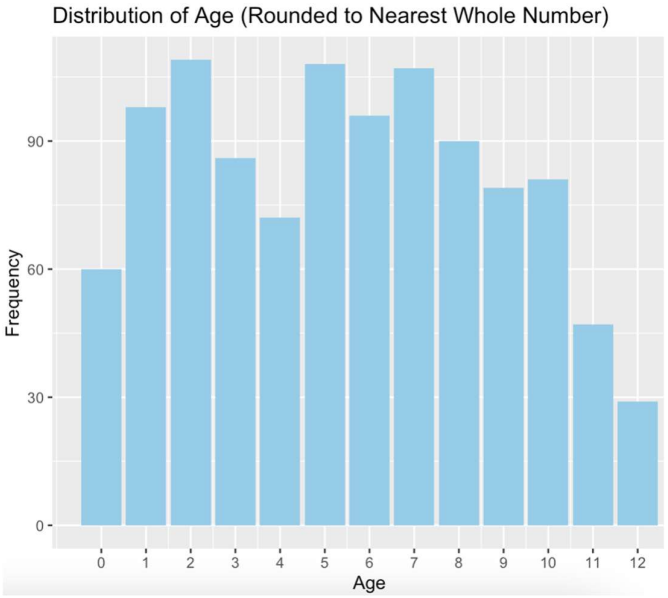


Figure 4: Distribution of Age

It is imperative to note that although the wells in this study are comparatively young, they are still contributing significantly to methane emissions: section 3.3.1 — where 59.8% of these wells are emitting methane beyond regulatory thresholds (Figure 2). This is particularly concerning as older wells will typically exhibit even higher emissions due to leaks (Adams-Heard, Z. M., 2021), underscoring an immediate need for robust methane mitigation strategies to address the current situation.

3.3.4 Correlation of Variables

From the correlation heatmap in Appendix A, we note the following:

1. WaterProduced and GasProduced have a correlation of 0.678

   This suggests that higher gas production is often associated with increased water output, which is logical since water is produced as a by-product during oil and gas drilling.

2. Age and emissionsSum has a correlation of 0.717
   This indicates that the age of the wells might be a strong predictor of methane emissions, likely due to degradation or lack of modern technology in older wells.

3. MeasuredChemicalInjectionPumps and TotalChemicalInjectionPumps has a correlation of 0.793
   This may indicate that the 2 variables may have significant overlap in its measurements. From our analysis, it appears that MeasuredChemicalInjectionPumps is likely a subset of TotalChemicalInjectionPumps. We will remove the variables if there is significant multicollinearity during the modelling phase.

4. Model Training and Evaluation

4.1 Train-Test Split

The Train-Test Split method divides the dataset into training and testing subsets. The model learns from the training data and is evaluated on the testing data, ensuring its ability to make predictions on new, unseen instances. This strategy aims to prevent overfitting and assess the model's real-world applicability. This is also known as supervised learning.

In this project, we will be using a 70-30 split, with 70% of the data as training data and 30% of the data for testing.

4.2 Linear Regression Model

Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables with assumptions of a linear relationship.

4.2.1 Feature Selection

Feature selection is the process of choosing the most relevant features or variables that contribute to the predictive power of the model. Selecting the appropriate features can improve the model’s performance by reducing overfitting, improving accuracy, and enhancing the interpretability of the model.

1. Using Stepwise Algorithm

   Stepwise algorithm is a dynamic and iterative method where variables are selectively included or excluded from the model. This algorithm uses statistical criteria, such as the Akaike Information Criterion (AIC), to add or remove variables, providing a good balance between model complexity and fit. Appendix B shows the selected variables after a backwards elimination using step() in R.

2. Multicollinearity Check

   A thorough evaluation for multicollinearity is then conducted. This ensures that the variables within the model maintain a level of independence, preventing any variable from being overly influenced by others, which could compromise the model's reliability and stability.

   Based on the VIF values in Appendix B, all the continuous variables has a GVIF of less than 5 and categorical variables has GVIF^(1/(2*Df)) of less than 2, indicating that there is no significant multicollinearity.

The linear regression model after feature selection consists of 14 variables: 10 continuous and 4 categorical.

4.2.2 Training Model

In this section, the constructed linear regression model, represented by the equation illustrated in Appendix B, was utilised for predictive analysis. The training dataset was used to calibrate the model, allowing it to learn and adapt to the underlying data patterns.

After the model was trained, the testing set was introduced to the model for prediction purposes. This phase reveals the model's ability to generalise and make accurate predictions on new, unseen data.

The performance of the model's predictions will be reviewed in the later part of this report.

4.2.3 Model Evaluation

The linear regression model assumes:

1. Linear Association between Y and Xs.
2. Errors have a normal distribution with mean 0.
3. Errors are independent of X and have constant standard deviation.

Through examining the various Model Diagnostic Plots in Appendix B, we can confirm the validity of these assumptions.

Plots	Findings
Residuals vs Fitted Plot	No distinctive pattern in the plot affirms the linearity of the model and the normal distribution of errors. (Assumption 1 & 2)
Normal Q-Q Plot	Residuals follow closely to the reference line, and do not deviate from normal distribution. (Assumption 2)
Scale-Location Plot	The evenly distributed spread of points suggests a consistent standard deviation across the range of fitted values (Assumption 3)
Residuals vs Leverage Plot	Regression line does not cross beyond the dotted lines, based on Cook's distance. This shows that there are no influential outliers.

4.3 CART Model

The Classification and Regression Trees (CART) model is a decision tree learning technique that is used for classification and regression predictive modelling. In this report, since we are predicting the continuous variable of `emissionsSum`, we will be using a Regression Tree Model.

4.3.1 Model Building

In building the CART model, we will first grow out the tree to reach the maximum number of terminal nodes. We then apply pruning techniques to streamline the model, ensuring a balance between statistical accuracy and complexity.

4.3.2 Pruning

We utilised the 1 Standard Error (SE) rule guideline in choosing the simplest tree that would be statistically equivalent with the minimum Cross-Validation (CV) error, based on a 10-fold cross validation. This approach aims to balance model simplicity with predictive accuracy, preventing overfitting and over complex trees.

With reference to the complexity parameter (CP) plot in Appendix C, the 9th tree stands out as the optimal choice, meeting the criterion of being within the 1 SE margin of the minimum CV error.

4.3.3 Model Evaluation

The pruned model consists of 8 splits, 7 intermediate nodes and 9 terminal nodes. Refer to Appendix C for the tree illustration.

4.4 Random Forest Model

Random forests are created by aggregating multiple decision trees, where each tree is built from a unique sample drawn with replacement (bootstrapping) from the dataset and uses a subset of features chosen randomly. The method ensures that each tree is influenced by different patterns in the data, which increases the overall robustness of the model (Breiman, 2001).

4.4.1 Model Building

In this analysis, our random forest model is configured with 500 trees — a number large enough to capture diverse data insights. Each tree considers a random subset of 5 features for splitting at each node, where 5 is derived from 'floor(M/3)'. Refer to Appendix D for the R implementation.

4.4.2 Model Evaluation

Plotting the Out-Of-Bag (OOB) error rate against the number of trees in Appendix D, we can see that errors would approximately be the same for any number of trees above 200. This means that the error rate has settled down, which is the rationale for selecting a value of B sufficiently large (James et al., 2017).

The variable importance is based on %IncMSE, where a larger %IncMSE value indicates that the model relies more on that variable for making accurate predictions. It is a measure of how much worse the model performs without that predictor. The plot for variable importance can be found in Appendix D.

4.5 Evaluation of Model Performance

To evaluate model performance, we chose to use more than one performance measure to evaluate the performance of a model as different measures capture different aspects of model performance, provide vital information on potential strengths and weaknesses of the models, and prevent any model from overfitting to a single measure.

We focused on three key metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the R-Squared Value. The definitions of the measures can be found in Appendix E.

	Linear Regression	**CART**	**Random Forest**
RMSE	55.892	52.408	**47.246**
MAE	43.887	40.471	**35.456**
R^2	0.67	0.71	**0.765**

Figure 5: Performance of the Models Across 3 Measures

Evaluating the measures, the random forest model seems to provide the best fit, given its lower RMSE and MAE values, and a higher R^2 value. We hence conclude that the Random Forest is the best performing model, and our business recommendations will be based on this model.

5. Business Applications

5.1 Recommendations

The prediction model can be applied in a 3-stage approach.

Stage 1: Assess Acceptability of Predicted Emission Level

Using a predictive model to estimate the methane emissions per well during the oil and gas drilling process can have operational advantages for Aramco.

Firstly, the model can predict various levels of emissions. The predicted rates can be benchmarked against current and future regulatory frameworks, such as the Inflation Reduction Act. Since the charges in the fee only apply to facilities that emit more than 25,000 metric tons of carbon dioxide equivalent per year, we can use it as a benchmark to minimise its emissions rate. Based on our calculations in Figure 6, the act will impose a fee if Aramco produces more than 85.5 Standard Cubic Feet per Minute (Brander, 2023).

$$\frac{25000\ tonnes}{29.8^{*}} \times \frac{1}{365\ Days \times 1440\ Mins} = 1.59\ \text{kg}^{\#}\ \text{per minute}$$

* 1 tonne of Methane = 29.8 tonnes of Carbon dioxide equivalent (CO2e)

\# 1 metric tonne = 1,000 kg

$$\frac{1.59\ \text{kg}}{0.657\ \text{kg/m}^3{}^{*}} \times 35.315^{\#} = 85.5\ \text{cubic feet per minute}$$

* Density of Methane is 0.657 kg/m³

\# 1 cubic meter = 35.315 cubic foot

Figure 6: Conversion of Methane Threshold from Tonnes/Yr. to SCFM

Emissions Guideline

	Acceptable	Moderately Acceptable	Unacceptable
Emission Rate	< 85.5 SCFM	85.5 to 90 SCFM	> 90 SCFM
Action	• Maintain Current Emission	• Maintain Emission; or • Reduce Emission (Refer to Actionable Operational Guideline)	• Reduce Emission (Refer to Actionable Operational Guideline)

Acceptable Emission Rate: To prevent going over the excessive methane threshold in the Inflation Reduction Act, the acceptable rate for Aramco would be below the threshold of 25,000 tonnes or 85.5 SCFM. This would ensure that Aramco is staying below the excessive emission and would continue to innovate to remain an industry leader in reducing its methane emissions.

The recommended action would be for Aramco to maintain its emission rate by keeping the inputs used in the Random Forest Model consistent.

Moderately Acceptable Emission Rate: The moderately acceptable emission rate is between the threshold in the act to the average methane emission from the oil sector in 2022 per operational facilities (Global Land & Offshore Oil Rig Count 2022, 2023; Overview – global methane tracker 2023, n.d.).

$$\frac{40{,}000{,}000\ tonnes}{29.8^{*}} \times \frac{1}{365\ Days \times 1440\ Mins} = 2553.8\ \text{kg}^{\#}\ \text{per minute}$$

* 1 tonne of Methane = 29.8 tonnes of Carbon dioxide equivalent (CO_2e)

\# 1 metric tonne = 1,000 kg

$$\frac{2553.8\ \text{kg}}{0.657\ \text{kg/m}^{3^{*}}} \times 35.315^{\#} \times \frac{1}{1{,}532\ facilities'} = 89.6\ \text{cubic feet per minute}$$

* Density of Methane is 0.657 kg/m^3

\# 1 cubic meter = 35.315 cubic foot

' 1,532 operational facilities based on Global Land & Offshore Oil Rig Count 2022

Figure 7: Conversion of Average Methane Emissions into SCFM

Even though it is not lower than the excessive emission threshold to keep Aramco as an industry leader in reducing methane emissions, it is still acceptable as it is within the average methane emissions emitted by the industry.

The recommended action would be for Aramco to perform a cost benefit analysis to determine if additional measures are needed to reduce the emissions by adjusting the important variables identified in stage 2 or to continue emitting at that level.

Unacceptable Emission Rate: The unacceptable rate would be any rate that exceeds the average methane emission from the oil sector in 2022 per operational facilities. This would show that the well is not meeting the industry average but is lagging in utilising more efficient practices in the drilling process.

The recommended action for Aramco would be to reduce the methane emissions by adjusting the specific variables based on the operational guideline in the second stage. The guideline would be updated regularly to be in line with the current emission levels.

Stage 2: Tackle Significant Variables Through Actionable Steps

Secondly, through methane emissions forecasting, Aramco can pinpoint specific variables that contribute to high emissions. Utilising the variable importance from the random forest model (Appendix D), Aramco can try to reduce variables to achieve an emission rate that is below the threshold.

Methods to Reduce Variables:

As stated earlier, using our model we are able to identify the variables that have the most influence on methane emission (Appendix D). Using this information we have formulated strategies to reduce these variables for abatement of methane emission.

Age:

Based on our model, age is the most significant factor that affects methane emission in oil drills. Age of well is proportionally related to methane emission. This means that as the age of the well increases, the amount of methane emission increases with it as well. This can be due to ageing causing wear and tear to pipes leading to higher leakage of methane (El Hachem, K., & Kang, M 2023). However age is not a factor that can be controlled without operational disruption or incurring very high costs. Hence our recommendation is to conduct regular maintenance, focusing on the older wells.

WellStruct:

The Structure of the well can lead to various implications when comparing Horizontal/Directional against Vertical well. They differ simply in the techniques used in drilling, horizontal and directional can access oil well at an angle while vertical can only access straight down. Horizontal/Directional are more advantages than Vertical well in terms of efficiency and effectiveness as they can reach places where Vertical well cannot (Team, W., 2023). However, it leads to higher emissions according to our analysis and is costlier. Since this factor is based on geographical location, there is little Aramco can do to mitigate this.

WellheadPressure:

The pressure being exerted at the wellhead can lead to varying levels of methane emissions. Higher well head pressure can lead to increased rates of leakage when equipment integrity is compromised. However, the wellhead pressure often follows a specific pressure range for optimal operation (Hull, K., 2023), hence Aramco will need to perform scenario analysis on the different wellhead pressures within the optimal range and determine which value will lead to the lowest emissions, this will be detailed in stage 3, scenario testing.

MeasuredEquipmentLeaks:

The higher the equipment leaks, the higher the emission of methane. Some common areas where leakages occur are wellheads, pipelines, and pneumatics (Hart Energy., 2021). Aramco can approach this by ensuring regular maintenance and timely detections to prevent leakage.

Combuster:

From our model, we have concluded that having combustors in wells helps to significantly reduce methane emission in drilling. The process of burning greenhouse gases in oil drilling using combustors is called oil flaring (World Bank, n.d.). This means that as oil is being drilled, all methane emissions and other greenhouse gases are being burnt up by these combustors hence severely reducing the methane emission in our model. Aramco should try to incorporate combustors in every drilling process to reduce methane emissions.

Actionable Operational Guideline

Variables	Suggested Actions
Age	Ensure regular maintenance, especially for older wells.
WellStruct	This is based on the geographic location of the oil, it is not within the control of Aramco.
WellheadPressure	Based on the range of pressure determined by Aramco, use our scenario testing to predict emission and adjust for the lowest value.
MeasuredEquipment Leaks	Ensure regular maintenance of equipment to prevent any leakage especially in common areas like wellheads and pipelines.
Combuster	Ensure that combusters are used during oil drilling.

Stage 3: Scenario Testing with Prediction Modelling

Aramco can perform scenario analysis with the random forest model as well to identify different scenarios that can decrease the methane emissions. These scenarios can be categorised based on their risk level and severity for Aramco to develop plans that allows them to prepare for and manage the risk exposures.

Aramco can supplement its Leak Detection And Repair programme with the predictive model to estimate the emission after its repair if there were any changes in the variables and determine if it will exceed the threshold.

5.2 Benefits of Implementing Solution

Implementing the predictive model into Aramco's business processes and risk management process can help Aramco achieve their sustainability goals as well as mitigating the potential legal risks from non-compliances of new regulations and reduce their cost of capital (Bernow et al., 2020).

5.2.1 Achieving Aramco's Sustainability Ambitions

Aramco is committed to the World Bank's "Zero Routine Flaring by 2030" and is committed to share strategies to reduce methane flaring. Furthermore, to align with Saudi Arabia's goal to have net-zero emissions by 2060, Aramco aims to achieve net-zero Scope 1 and Scope 2 greenhouse gas emissions from their operations by 2050 (Our approach to sustainability, n.d.).

By utilising the predictive model, Aramco can more accurately forecast its yearly emissions and assess Aramco progress over time. Aramco would be able to investigate areas timely and identify factors that cause the emissions to be higher than expected. Aramco can then implement targeted interventions and operational adjustments to reduce the methane emissions. With an accurate forecast, Aramco would be able to evaluate the feasibility of achieving its goal for net-zero emissions.

The use of predictive models and data driven analysis shows that Aramco is committed to quality and innovations to enhance the operational efficiencies of their processes. This would also indicate to their stakeholders that Aramco is well positioned for the future and is taking steps to increase their shareholder and societal value.

5.2.2 Mitigating Potential Legal and Financial Cost

As the United States led the "Net Zero Producers Forum" with members such as Canada, Qatar and Saudi Arabia. It could be expected that such fees from the inflation reduction act could be implemented in other countries that Aramco is operating as well. Aramco should plan ahead and implement methane reduction measures on its own terms and timeline.

By utilising the predictive model, Aramco could more accurately estimate their emissions for the year and take steps for strategic planning to maintain their emissions to fall within 25,000 tons of carbon dioxide equivalent. By proactively seeking strategies to reduce their emission it would better prepare Aramco for future regulatory change and remain a leader in reducing methane intensity while avoiding the additional financial cost of excessive emissions.

5.2.3 Reduce Cost of Capital

Companies that perform better in sustainability have a lower cost of capital by around 10 percent (Bernow et al., 2020). As investors are seeking more environmentally sustainable companies the perceived riskiness of the investment for these companies will be reduced as they are viewed to be forward looking and are taking steps to mitigate emerging risk that might threaten their operations (Grundmann et al., 2022).

The negative relationship between cost of capital and sustainability is mainly driven by the cost of equity. Aramco has a debt-to-equity ratio of 16.38% (Saudi Arabian Oil Company (2222.SR) valuation measures & financial statistics, 2023). When compared to the industry average of 50%, it indicates that Aramco utilises more equity financing than debt financing hence brand image is very important to Aramco (Oil and gas industry: Overview, financial ratios and future, 2023). Thus, by using a predictive model to minimise its methane emission, it signals to investors that Aramco is constantly finding opportunities to meet its stakeholders expectations and is including sustainability in its strategy to achieve long term sustainability (Grundmann et al., 2022).

Thus, the utilisation of the predictive model could lower Aramco's cost of equity and Aramco would have to pay a lower rate of return to its equity shareholders. Investors would perceive Aramco as less risky, and it would be easier for Aramco to raise funds through its share offerings.

5.3 Limitations

5.3.1 Predictive Modelling

Throughout this Proof of Concept, we explored the use of three distinct predictive models to gauge their efficacy in forecasting methane emissions. However, with continuous innovations and research, there will be newer and more sophisticated analytical algorithms being introduced. These advanced models may have the potential to yield higher accuracy and predictive capabilities. Therefore, it is essential to acknowledge that while our current model provide valuable insights, future explorations should consider the latest developments in data analytics to further refine and enhance predictive performance in the context of methane emissions.

5.3.2 External Factors

There could be other external factors that are not within the control of Aramco. Factors such as location of the rig could influence the amount of methane being emitted and Aramco does not have the ability to control or change the location (Carbon Mapper., 2022). Resulting in variables that cannot be adjusted even if it contributes highly to methane emission.

5.3.3 Balancing Economical Gains with Environmental Responsibility

Addressing methane emissions requires a delicate balance between environmental stewardship and economic practicality. Aramco would have to weigh the long-term benefits against the associated costs of implementing predictive models for emission reduction.

Nevertheless, our team believes that while economic considerations are important, Aramco should prioritise on sustainability objectives given that it does not significantly compromise their financial performance.

5.4 Future Improvements

Based on limitations mentioned above, Aramco can utilise more advanced analytics to develop different models that can integrate additional data into the prediction models.

5.4.1 Different Models for Different Types of Drillings

The dataset that the prediction model is trained on was collected after the drilling process. Thus, to improve the decision-making process, there could be different models that are trained with data collected during different stages of the drilling process. It could be trained with data for exploratory drilling and during the drilling process to provide a more holistic view of the methane emissions.

5.4.2 Continuous Training of the Model

As the technology used during the oil and gas drilling process is always improving, the prediction model would need to be kept up to date in order to generate the most accurate prediction on methane emissions. Thus, the model should be trained continuously with new data to improve the accuracy of the prediction to be up to date with the advancement in technology.

6. Conclusion

This report demonstrated that predictive modelling serves as a useful tool for Saudi Aramco's methane emission reduction initiatives. The Random Forest model stands out as the most suitable tool for forecasting emissions levels and guiding strategic decisions. By adopting this model, Aramco can better manage its operations to stay within regulatory thresholds, such as those outlined by the Inflation Reduction Act, and support its commitment to the Global Methane Pledge.

In conclusion, the balance between economic objectives and environmental responsibility is the main consideration. While the predictive model provides a pathway to achieve sustainability goals, they must be continually refined with emerging technologies and analytics to ensure relevance and accuracy. Aramco's proactive stance on methane mitigation is expected to not only enhance its environmental standing but also provide a competitive edge by reducing capital costs and strengthening investor confidence. The pursuit of technological innovation, adherence to maintenance best practices, and investment in predictive analytics are therefore recommended as prudent steps toward a more sustainable and economically sound future for Saudi Aramco.

References

Adams-Heard, Z. M. (2021, October 12). Exposing climate threats from an empire of dying gas wells. Bloomberg.com. https://www.bloomberg.com/features/diversified-energy-natural-gas-wells-methane-leaks-2021/

Allen, D. T., Torres, V. M., Thomas, J., Sullivan, D. W., Harrison, M., Hendler, A., Herndon, S. C., Kolb, C. E., Fraser, M. P., Hill, A. D., Lamb, B. K., Miskimins, J., Sawyer, R. F., & Seinfeld, J. H. (2013). Measurements of methane emissions at natural gas production sites in the United States. Proceedings of the National Academy of Sciences, 110(44), 17768–17773. https://doi.org/10.1073/pnas.1304880110

Bernow, S., Nuttall, R., & Brown, S. (2020, May 26). Why ESG is here to stay. McKinsey & Company. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/why-esg-is-here-to-stay

Brander, M. (2023, July). Greenhouse Gases, CO2, CO2e, and Carbon: What Do All These Terms Mean?. Ecometrica. https://ecometrica.com/assets/GHGs-CO2-CO2e-and-Carbon-What-Do-These-Mean-v2.1.pdf

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Carbon Mapper. (2022, August 11). Study suggests offshore oil and gas production in Gulf of Mexico has higher methane loss rates than typical onshore production. https://carbonmapper.org/study-suggests-offshore-oil-and-gas-production-in-gulf-of-mexico-has-higher-methane-loss-rates-than-typical-onshore-production/

Downey, L. (2022, July 17). Horizontal Well: What it Means, How it Works. Investopedia. https://www.investopedia.com/terms/h/horizontalwell.asp

Gas venting (2023) Trenchlesspedia. Retrieved on 29th Sep 2023, from https://www.trenchlesspedia.com/definition/3398/gas-venting

GHG emissions management program. (2023, September 18). Saudi Aramco. Retrieved on 29th Sep 2023, from https://www.aramco.com/en/sustainability/climate-change/managing-our-footprint/ghg-emissions-management-program

Global Land & Offshore Oil Rig Count 2022. Statista. (2023, August 25). https://www.statista.com/statistics/1128408/number-of-global-oil-rigs-by-type/

Grundmann, G., Klein, Dr. F., & Josten, F. (2022, September 14). Staying ahead of the sustainability curve. Deloitte Insights. https://www2.deloitte.com/xe/en/insights/topics/strategy/sustainability-in-business-staying-ahead-of-the-curve.html

Hachem, K. E., & Kang, M. (2023). Reducing oil and gas well leakage: a review of leakage drivers, methane detection and repair options. Environmental Research: Infrastructure and Sustainability, 3(1), 012002. https://doi.org/10.1088/2634-4505/acbced

Hull, K. (2023, June 22). Wellhead Pressure Interlock Protection - SOR Controls Group. SOR Controls Group. https://www.sorinc.com/2014/05/19/wellhead-pressure-interlock-protection/

Inflation reduction act methane emissions charge: In brief - CRS reports. (2022, August 29). Retrieved on 25th Sep 2023, from https://crsreports.congress.gov/product/pdf/R/R47206

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017a). An introduction to statistical learning: With applications in R. Springer, Ch 8.2.

Kuylenstierna, J. C. I., Michalopoulou, E., & Dixon, F. (2021, May 12). Why we must reduce methane emissions now to solve the climate crisis. SEI. Retrieved on 15th Sep 2023, from https://www.sei.org/features/why-we-must-reduce-methane-emissions-now-to-solve-the-climate-crisis/

Methane: A crucial opportunity in the Climate Fight. Environmental Defense Fund. (n.d.). Retrieved on 15th Sep 2023, from https://www.edf.org/climate/methane-crucial-opportunity-climate-

fight#:~:text=Methane%20has%20more%20than%2080,by%20%20methane%20from%20human%20%20actions

Methane emissions are driving climate change. here's how to reduce them. UNEP. (2021, August 20). Retrieved on 29th Sep 2023, from https://www.unep.org/news-and-stories/story/methane-emissions-are-driving-climate-change-heres-how-reduce-them

Overview – global methane tracker 2023 – analysis. IEA. (n.d.-a). Retrieved on 29th Sep 2023, from https://www.iea.org/reports/global-methane-tracker-2023/overview

Oil and Gas Industry Should Confront Flaring, Methane Leaks | Hart Energy. (2021, July 8). Hart Energy. https://www.hartenergy.com/exclusives/oil-and-gas-industry-should-confront-flaring-methane-leaks-194984

Oil and gas industry: Overview, financial ratios and future. LinkedIn. (2023, June 1). https://www.linkedin.com/pulse/oil-gas-industry-overview-financial/

Our approach to sustainability. Aramco. (n.d.). https://www.aramco.com/en/sustainability/our-approach-to-sustainability

Overview – global methane tracker 2023. IEA. (n.d.). https://www.iea.org/reports/global-methane-tracker-2023/overview

Primary sources of methane emissions | US EPA. Primary Sources of Methane Emissions. (2023, September 15). Retrieved on 29th Sep 2023, from https://www.epa.gov/natural-gas-star-program/primary-sources-methane-emissions

Team, W. (2023, September 13). Horizontal well. WallStreetMojo. https://www.wallstreetmojo.com/horizontal-well/

The life cycle of oil and gas fields. (n.d.). Planète Énergies. https://www.planete-energies.com/en/media/article/life-cycle-oil-and-gas-fields#:~:text=Oil%20and%20gas%20fields%20generally,more%20for%20the%20%20largest%20%20deposits

The global methane pledge – global methane tracker 2022 – analysis. IEA. (n.d.-b). Retrieved on 29th Sep 2023, from https://www.iea.org/reports/global-methane-tracker-2022/the-global-methane-pledge

United Nations Environment Programme. (n.d.). How secretive methane leaks are driving climate change. UNEP. Retrieved on 23rd Sep 2023, from https://www.unep.org/news-and-stories/story/how-secretive-methane-leaks-are-driving-climate-change

What is Gas Flaring? (n.d.). World Bank. https://www.worldbank.org/en/programs/gasflaringreduction/gas-flaring-explained#why

Yahoo! (2023, November 4). Saudi Arabian Oil Company (2222.SR) valuation measures & financial statistics. Yahoo! Finance. https://finance.yahoo.com/quote/2222.SR/key-statistics?p=2222.SR

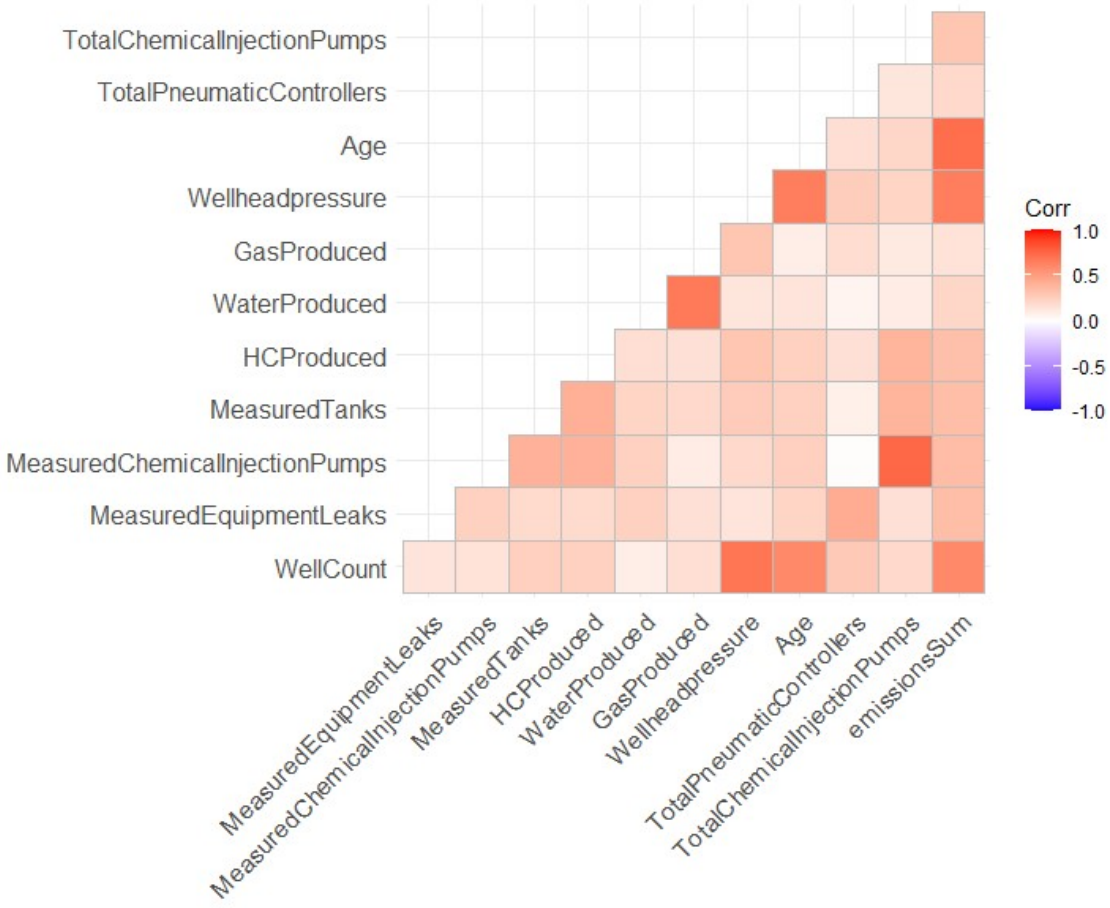
Appendix

Appendix A: Data Cleaning & Exploration

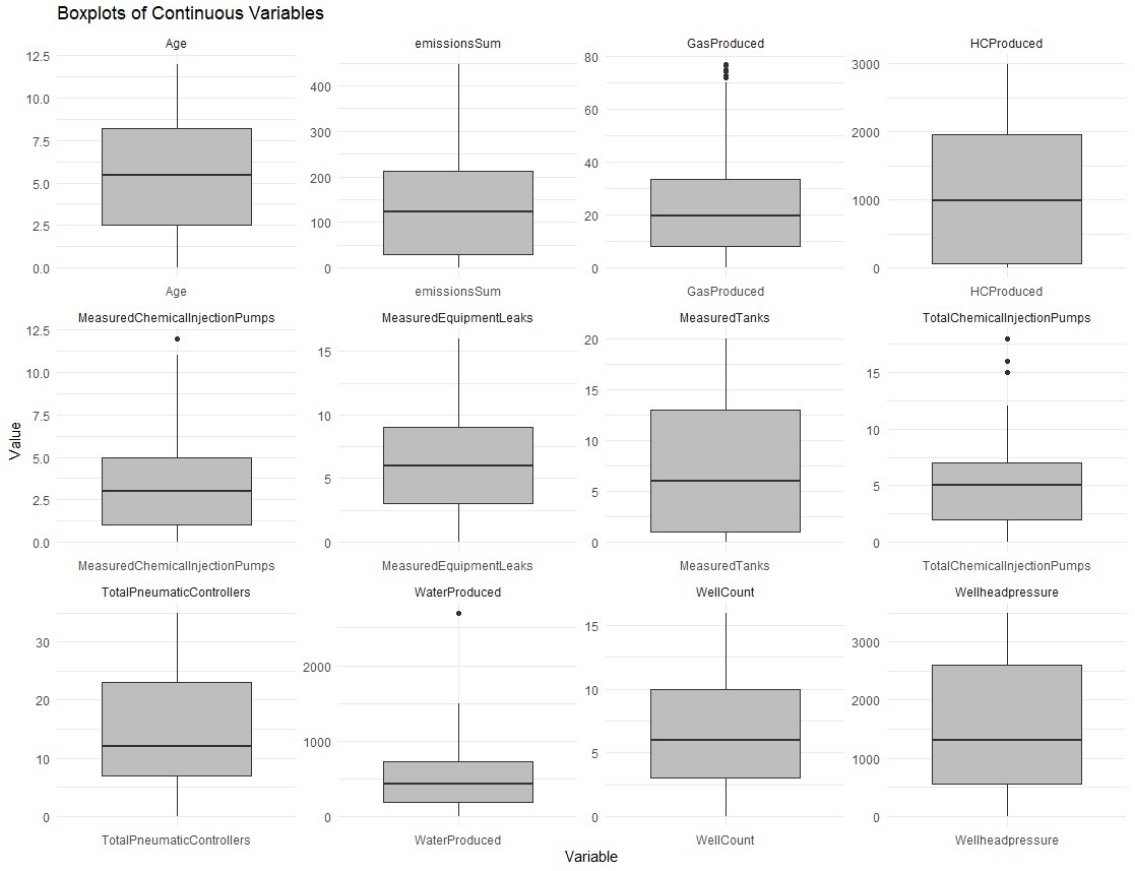
Data Dictionary for our dataset:

Variable Name	Explanation
SiteName	A unique identifier for the site where the oil and gas extraction or processing is taking place.
WellCount	The number of wells present at the site.
MeasuredEquipmentLeaks	Number of measured equipment leaks at the site.
MeasuredChemicalInjectionPumps	Number of measured chemical injection pumps at the site.
MeasuredTanks	Number of measured tanks at the site.
HCProduced (bbl/day)	Volume of hydrocarbons produced (in barrels per day)
WaterProduced (bbl/day)	Volume of water produced (in barrels per day)
GasProduced (MMscf/day)	Volume of gas produced (in million standard cubic feet per day)
WellheadPressure (psig)	Pressure at the wellhead (in pounds per square inch gauge)
Age	Age of the well (in years)
Fracked?	Whether fracking was carried out.
ProducingType	Type of geological formation (Shale gas, Other Tight Reservoir Rock & Shale, …)
WellStruct	Structure of the well (Vertical, Horizontal, Bidirectional)
TotalPneumaticControllers	Total number of pneumatic controllers.
TotalChemicalInjectionPumps	Total number of chemical injection pumps.
Combuster	Whether a combustor is present.
Tracerflux	Whether tracer flux is present.
EmissionsSum (scfm)	Methane emissions in standard cubic feet per minute.

Correlation Heatmap for continuous variables:



Boxplot of variables before cleaning:



Removal of Outliers:

```
## Outlier Handling ------------------------------------

# Identify outlier rows for each column
outlier_rows <- emissionsData.dt %>%
  select_if(is.numeric) %>%
  map(~ which(.x %in% boxplot.stats(.x)$out)) %>%
  unlist() %>%
  unique()

# Remove rows containing outliers
emissionsData.dt <- emissionsData.dt[-outlier_rows, ]
```

Detecting missing fields and replacing NA fields with mean for continuous variables:

```
## Missing Data Handling ------------------------------------------------
missing_values <- sapply(emissionsData.dt, function(x) sum(is.na(x)))
print(missing_values)

# Replace NA with mean for continuous
emissionsData.dt <- emissionsData.dt %>%
  mutate(across(-c(Combuster, tracerflux, ProducingType, WellStruct),
                ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

# Check if replaced
sapply(emissionsData.dt, function(x) sum(is.na(x)))
```

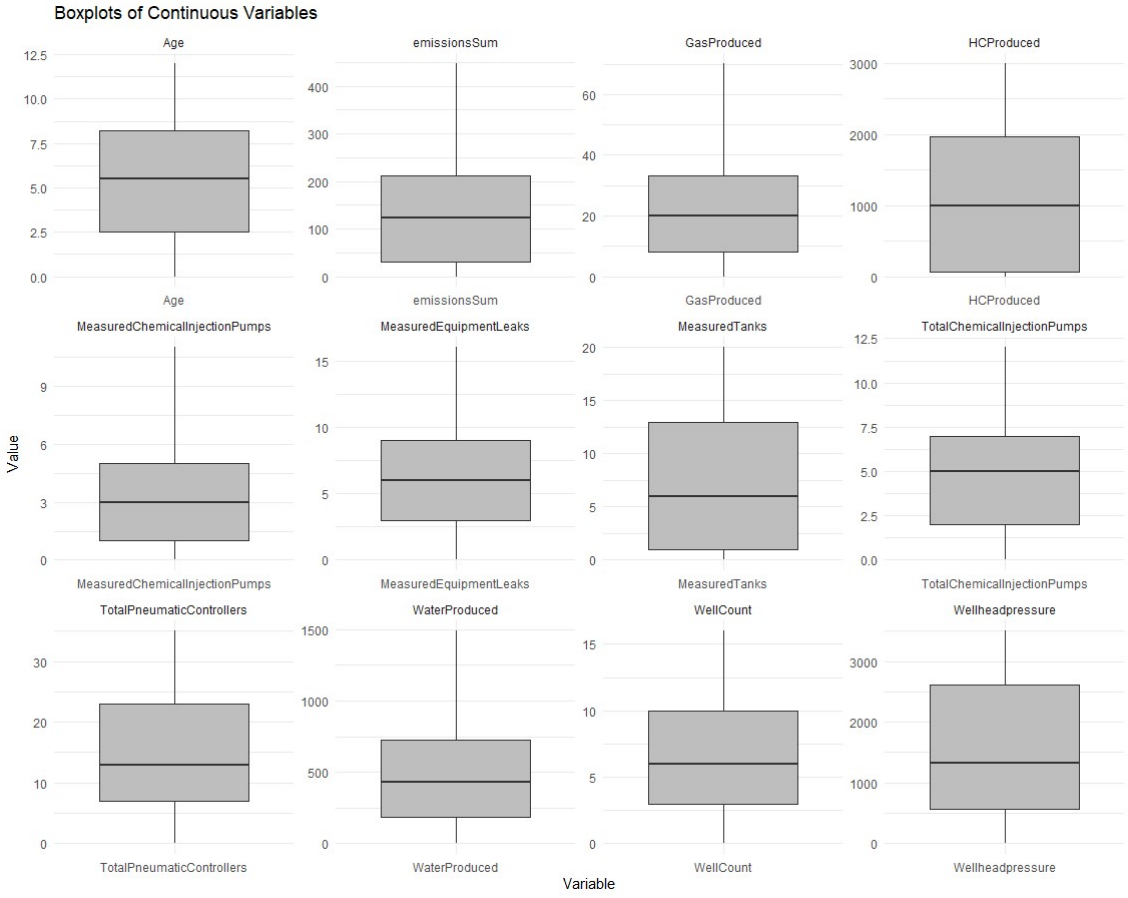
Replacing NA fields via logistic regression models for categorical variables of `Combuster`, `tracerflux` and `ProducingType`:

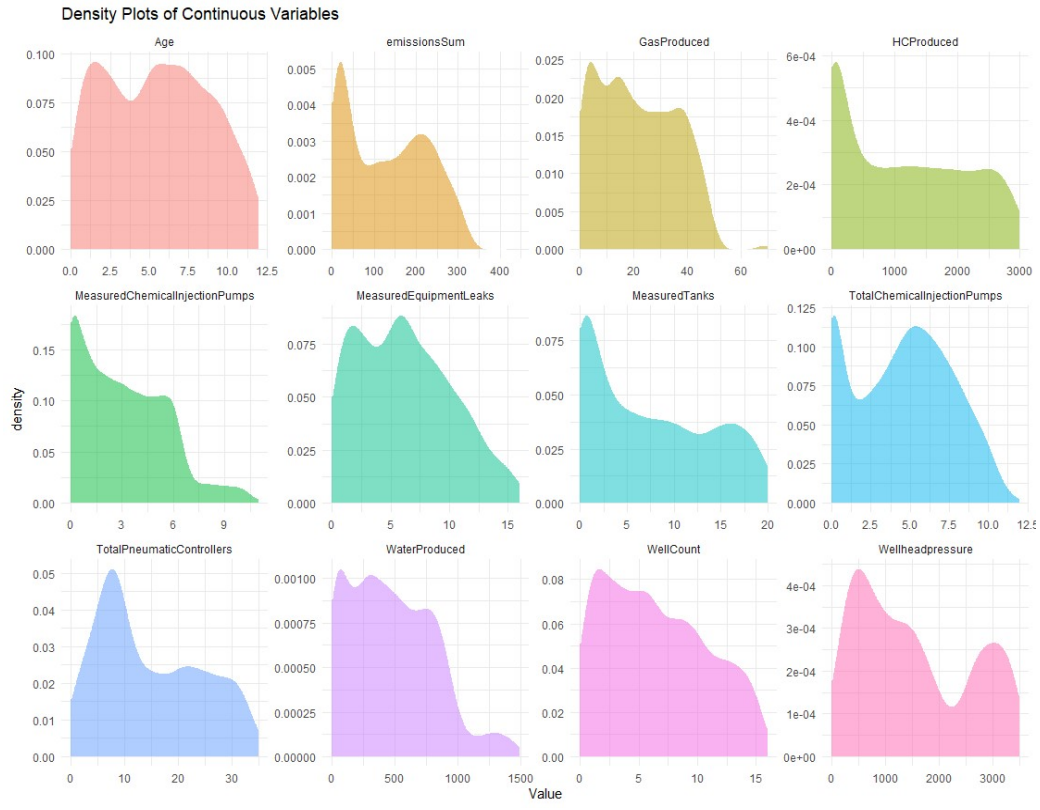
```
# Logistic Regression for `Combuster` with 2 levels
log.Combuster <- glm(Combuster ~ .,data = emissionsData.dt, family = binomial)
summary(log.Combuster)
# Predict probabilities using the logistic regression model
predicted_probs <- predict(log.Combuster, newdata = emissionsData.dt, type = "response")
# Identify indices of missing values in 'Combuster'
missing_indices <- which(is.na(emissionsData.dt$Combuster))
# Replace missing values using a threshold (e.g., 0.5 for binary classification)
emissionsData.dt$Combuster[missing_indices] <- ifelse(predicted_probs[missing_indices] > 0.5, 1, 0)

# Logistic Regression for `tracerflux` with 2 levels
log.tracerflux <- glm(tracerflux ~ ., data = emissionsData.dt, family = binomial)
# Predict probabilities using the logistic regression model
predicted_probs_tracerflux <- predict(log.tracerflux, newdata = emissionsData.dt, type = "response")
# Identify indices of missing values in 'tracerflux'
missing_indices_tracerflux <- which(is.na(emissionsData.dt$tracerflux))
# Replace missing values using a threshold
emissionsData.dt$tracerflux[missing_indices_tracerflux] <- ifelse(predicted_probs_tracerflux[missing_indices_tracerflux] > 0.5,

# Multinomial Logistic Regression for `ProducingType`
multinom.ProducingType <- multinom(ProducingType ~ . -Combuster -tracerflux -WellStruct, data = emissionsData.dt)
# Predict the class labels using the multinomial logistic regression model
predicted_labels_ProducingType <- predict(multinom.ProducingType, newdata = emissionsData.dt)
# Identify indices of missing values in 'ProducingType'
missing_indices_ProducingType <- which(is.na(emissionsData.dt$ProducingType))
# Replace missing values
emissionsData.dt$ProducingType[missing_indices_ProducingType] <- predicted_labels_ProducingType[missing_indices_ProducingType]
```

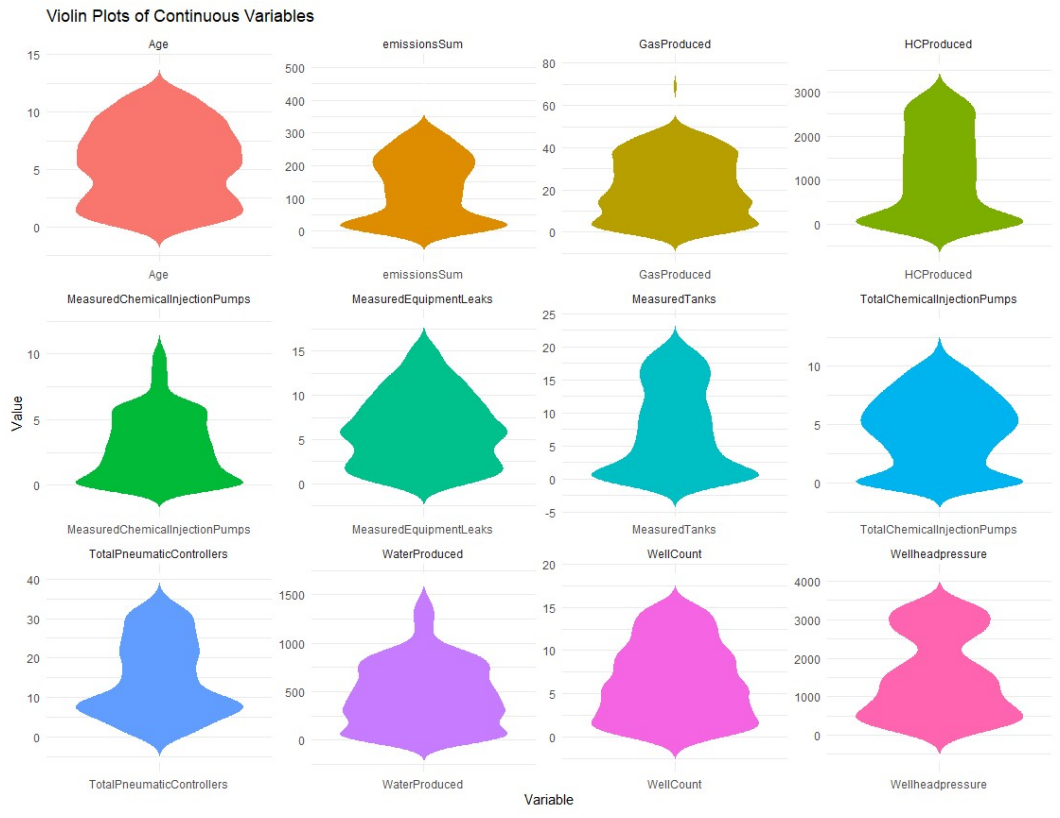
Duplicate checks:

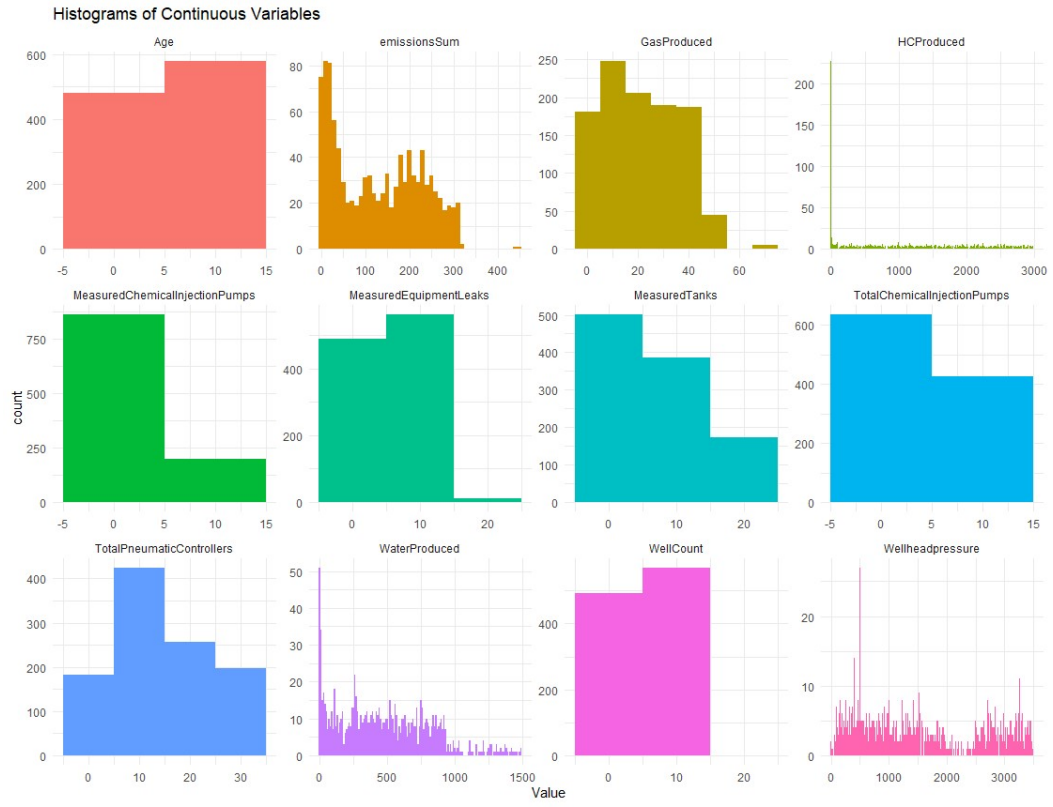
```
## Check for duplicated rows based on all columns -------------
duplicates <- emissionsData.dt[duplicated(emissionsData.dt), ]

# Print out number of duplicated rows
cat("Number of duplicated rows: ", nrow(duplicates), "\n")
```

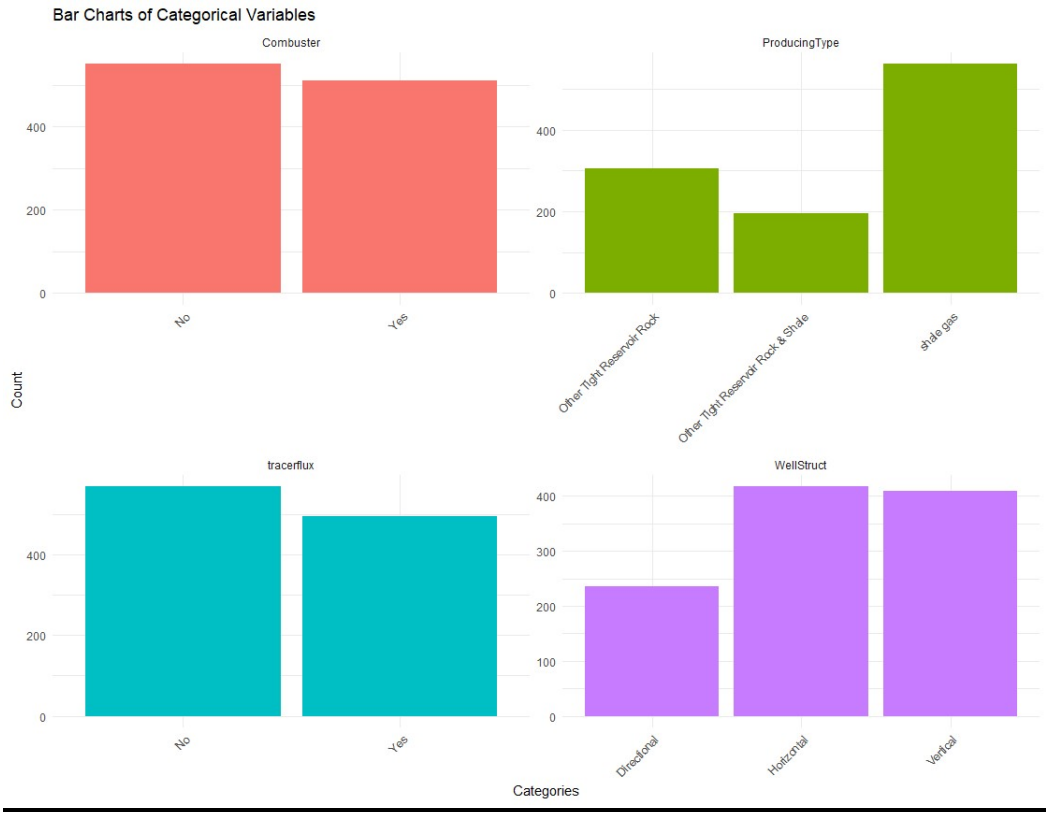
Boxplots after removal of outliers:




Density Plots of Continuous Variables
Age
emissionsSum
GasProduced
HCProduced
MeasuredChemicalInjectionPumps
MeasuredEquipmentLeaks
MeasuredTanks
TotalChemicalInjectionPumps
TotalPneumaticControllers
WaterProduced
WellCount
Wellheadpressure
density
Value


Violin Plots of Continuous Variables
Age
emissionsSum
GasProduced
HCProduced
MeasuredChemicalInjectionPumps
MeasuredEquipmentLeaks
MeasuredTanks
TotalChemicalInjectionPumps
TotalPneumaticControllers
WaterProduced
WellCount
Wellheadpressure
Value
Variable


Histograms of Continuous Variables
Age
emissionsSum
GasProduced
HCProduced
MeasuredChemicalInjectionPumps
MeasuredEquipmentLeaks
MeasuredTanks
TotalChemicalInjectionPumps
TotalPneumaticControllers
WaterProduced
WellCount
Wellheadpressure
count
Value


Bar Charts of Categorical Variables
Combuster
ProducingType
tracerflux
WellStruct
No
Yes
Other Tight Reservoir Rock
Other Tight Reservoir Rock & Shale
shale gas
Directional
Horizontal
Vertical
Count
Categories

Appendix B: Linear Regression Modeling

Set of selected variables after a backwards elimination with step() in R:

```
> linRegSelected <- step(linReg)

                            Df Sum of Sq    RSS      AIC
<none>                                    595.83 -130.009
- tracerflux                 1     2.457 598.29 -128.951
- GasProduced                1     5.718 601.55 -124.912
- HCProduced                 1     6.641 602.47 -123.774
- TotalChemicalInjectionPumps 1    8.113 603.94 -121.960
- MeasuredTanks              1     8.862 604.69 -121.039
- ProducingType              2    11.880 607.71 -119.340
- MeasuredEquipmentLeaks     1    10.998 606.83 -118.419
- TotalPneumaticControllers  1    14.219 610.05 -114.486
- WaterProduced              1    14.490 610.32 -114.156
- WellCount                  1    26.539 622.37  -99.631
- Combuster                  1    35.882 631.71  -88.560
- Wellheadpressure           1    48.466 644.30  -73.904
- WellStruct                 2    56.211 652.04  -67.026
- Age                        1   108.691 704.52   -7.510
```

Multicollinearity check via Variance Inflation Factor (VIF):

```
> vif(linRegSelected)
                                GVIF Df GVIF^(1/(2*Df))
WellCount                   2.217640  1        1.489174
MeasuredEquipmentLeaks      1.813766  1        1.346761
MeasuredTanks               1.450503  1        1.204368
HCProduced                  1.435641  1        1.198182
WaterProduced               2.093703  1        1.446963
GasProduced                 2.187548  1        1.479036
Wellheadpressure            2.762489  1        1.662074
Age                         2.159137  1        1.469400
ProducingType               1.360398  2        1.079982
WellStruct                  2.083788  2        1.201471
TotalPneumaticControllers   1.738295  1        1.318444
TotalChemicalInjectionPumps 1.437109  1        1.198795
Combuster                   1.510609  1        1.229068
```

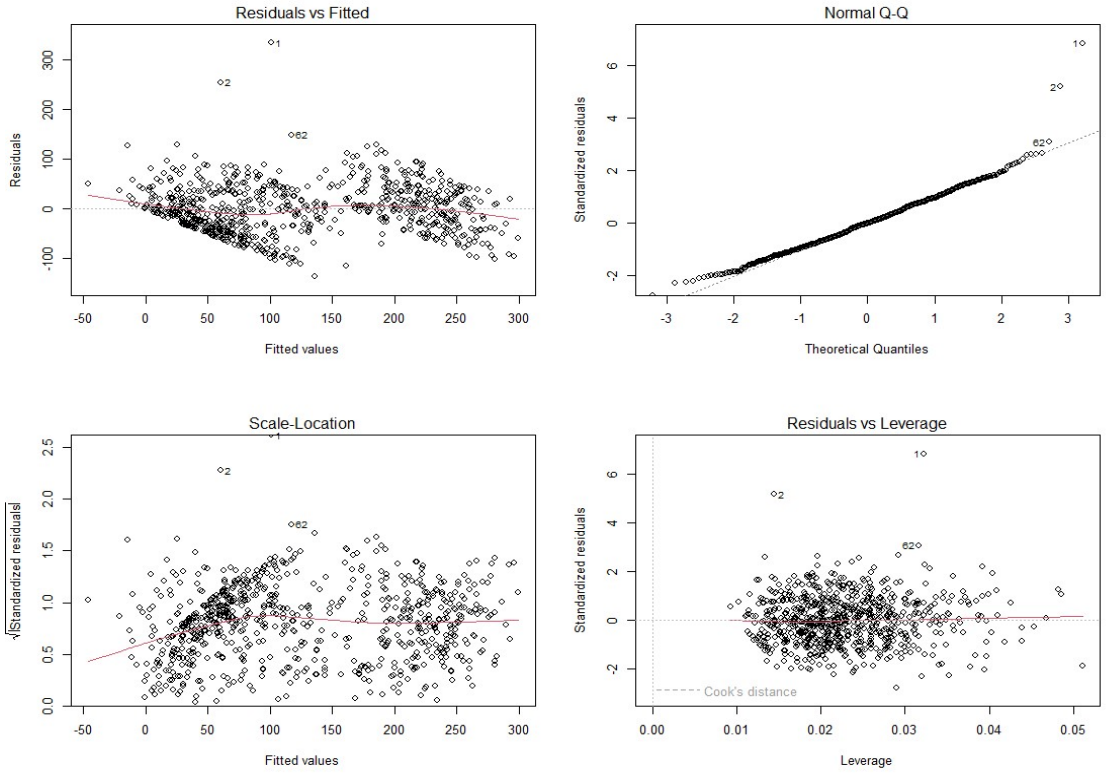
Final Linear Regression Equation:

```
Call:
lm(formula = emissionsSum ~ WellCount + MeasuredEquipmentLeaks +
    MeasuredTanks + HCProduced + WaterProduced + GasProduced +
    Wellheadpressure + Age + ProducingType + WellStruct + TotalPneumaticControllers +
    TotalChemicalInjectionPumps + Combuster + tracerflux, data = trainset)
```

Summary of Linear Regression:

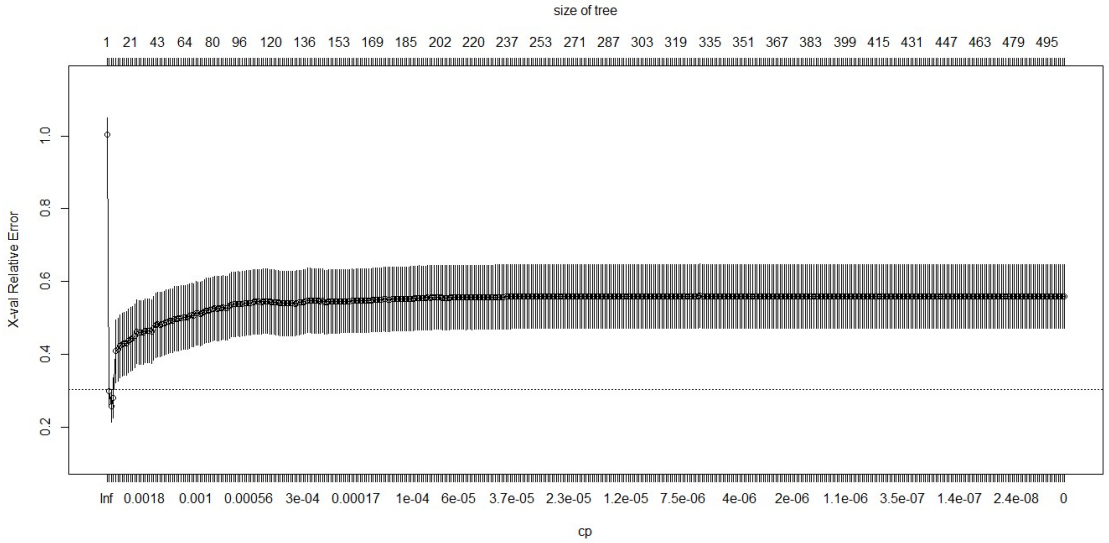
```
Call:
lm(formula = emissionsSum ~ WellCount + MeasuredEquipmentLeaks +
    MeasuredTanks + HCProduced + WaterProduced + GasProduced +
    Wellheadpressure + Age + ProducingType + WellStruct + TotalPneumaticControllers +
    TotalChemicalInjectionPumps + Combuster + tracerflux, data = trainset)

Residuals:
    Min      1Q  Median      3Q     Max
-135.76  -34.20   -1.87   33.36  334.47

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                               13.243240   6.229355   2.126 0.033845 *
WellCount                                  3.542400   0.620704   5.707 1.67e-08 ***
MeasuredEquipmentLeaks                     2.289000   0.598644   3.824 0.000143 ***
MeasuredTanks                              1.133486   0.338078   3.353 0.000842 ***
HCProduced                                 0.006339   0.002209   2.870 0.004221 **
WaterProduced                              0.033566   0.008087   4.151 3.71e-05 ***
GasProduced                               -0.499432   0.192888  -2.589 0.009812 **
Wellheadpressure                           0.021670   0.002846   7.615 8.23e-14 ***
Age                                        9.060029   0.798318  11.349  < 2e-16 ***
ProducingTypeOther Tight Reservoir Rock   16.203664   4.663719   3.474 0.000542 ***
ProducingTypeOther Tight Reservoir Rock & Shale  14.237807   5.557398   2.562 0.010609 *
WellStructDirectional                     -2.894051   5.328387  -0.543 0.587201
WellStructVertical                       -40.912613   5.145119  -7.952 7.05e-15 ***
TotalPneumaticControllers                 -0.993803   0.236524  -4.202 2.98e-05 ***
TotalChemicalInjectionPumps                2.191844   0.687655   3.187 0.001497 **
CombusterYes                             -30.872936   4.557729  -6.774 2.59e-11 ***
tracerfluxYes                              6.851619   3.793677   1.806 0.071323 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.75 on 726 degrees of freedom
Multiple R-squared:  0.7457,    Adjusted R-squared:   0.74
F-statistic:   133 on 16 and 726 DF,  p-value: < 2.2e-16
```

Linear Model Diagnostic Plots



Appendix C: CART Modeling

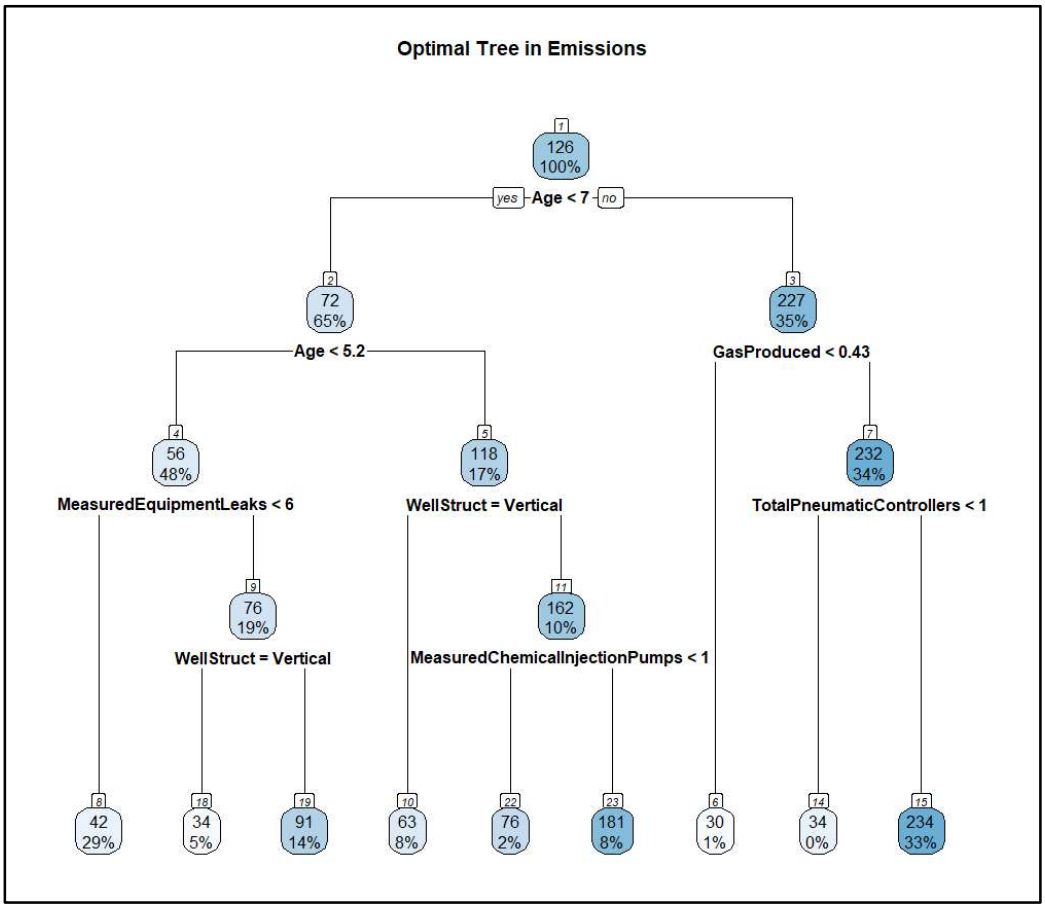
Complexity parameter (cp) plot for CART model:



Automated extraction of optimal tree based on 1SE rule:

```
# Extract the Optimal Tree via code instead of eye power
# Compute min CVerror + 1SE in maximal tree CART.train
CVerror.cap <- (CART.train$cptable[which.min(CART.train$cptable[,"xerror"]), "xerror"]
              + CART.train$cptable[which.min(CART.train$cptable[,"xerror"]), "xstd"])

# Find the optimal CP region whose CV error is just below CVerror.cap in maximal tree.
i <- 1; j<- 4
while (CART.train$cptable[i,j] > CVerror.cap) {
  i <- i + 1
}
# Get geometric mean of the two identified CP values in the optimal region.
cp.opt = ifelse(i > 1, sqrt(CART.train$cptable[i,1] * CART.train$cptable[i-1,1]), 1)

# Pruning the training model
CART.train_pruned <- prune(CART.train, cp = cp.opt)
```

Final pruned regression tree:

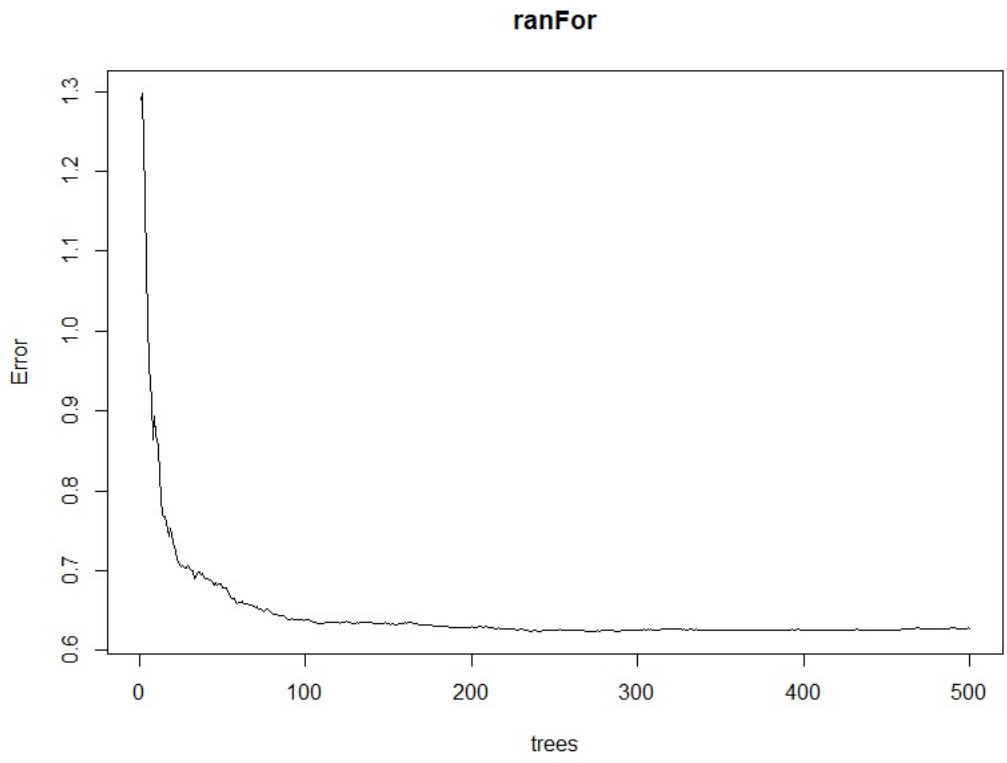


Appendix D: Random Forest Modeling

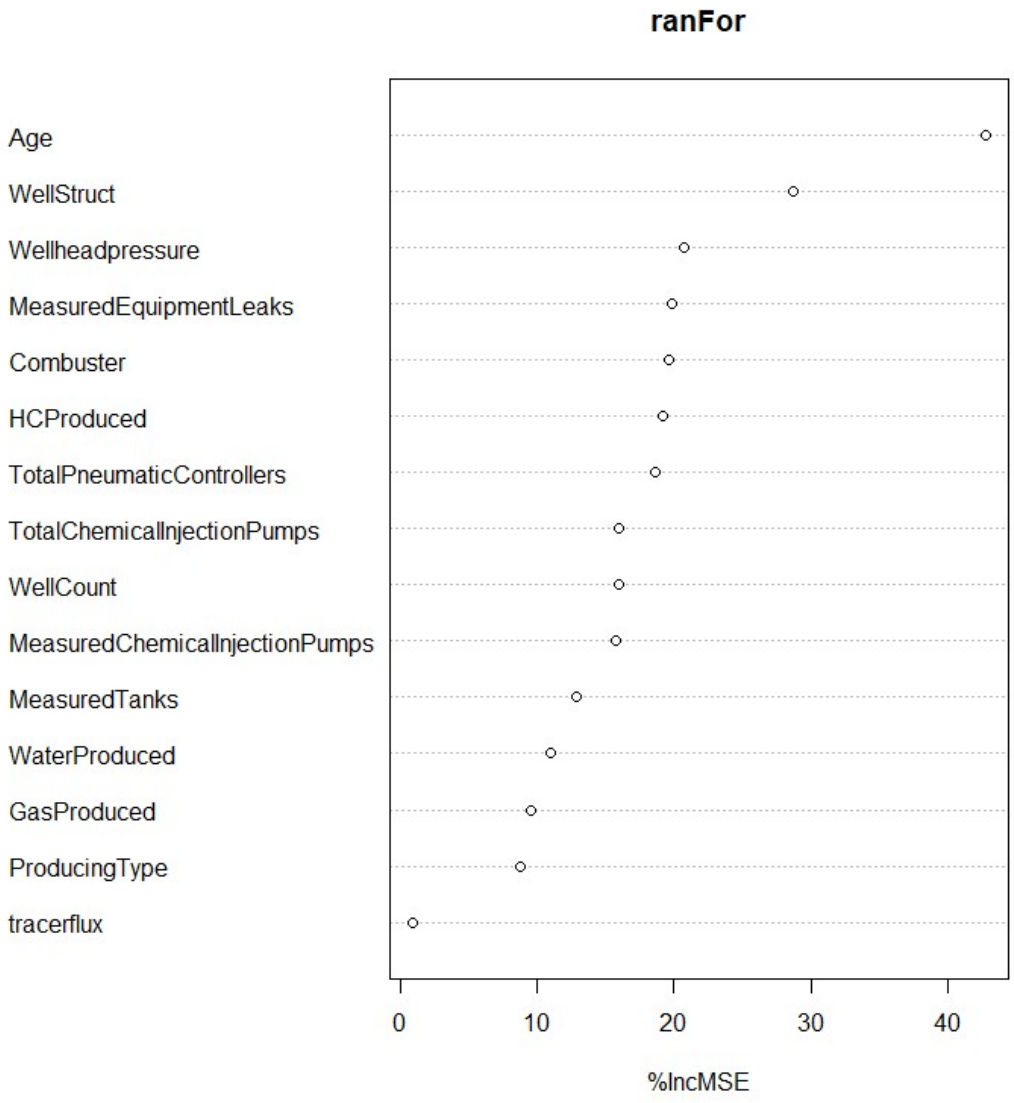
Random Forest implementation in R:

```
> randomForest(emissionsSum ~ . , data = trainset, importance = T)

Call:
 randomForest(formula = emissionsSum ~ ., data = trainset, importance = T)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 0.6262097
                    % Var explained: 80.07
```

OOB error against size of tree plot:



Variable Importance for Random Forest:



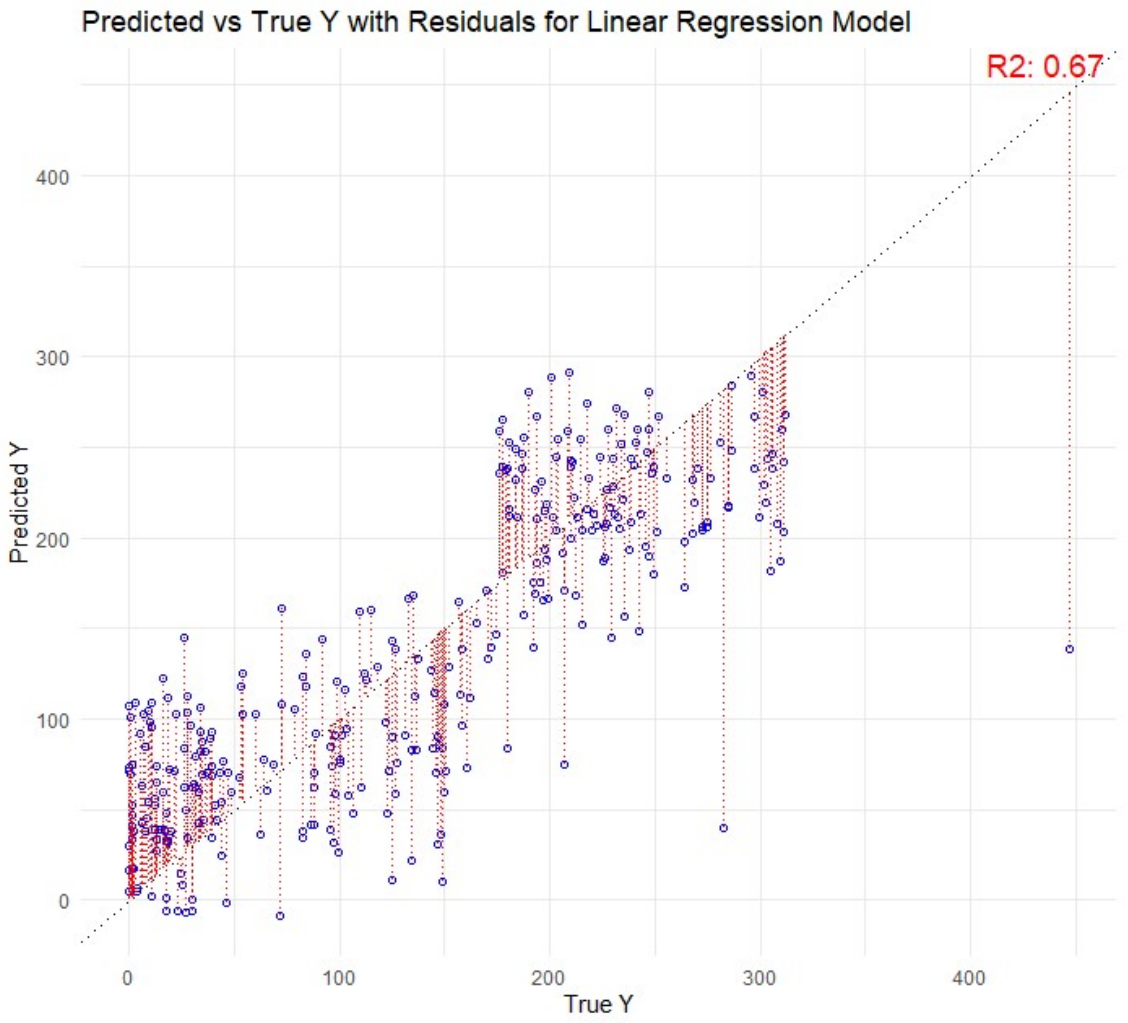
Appendix E: Model Evaluation

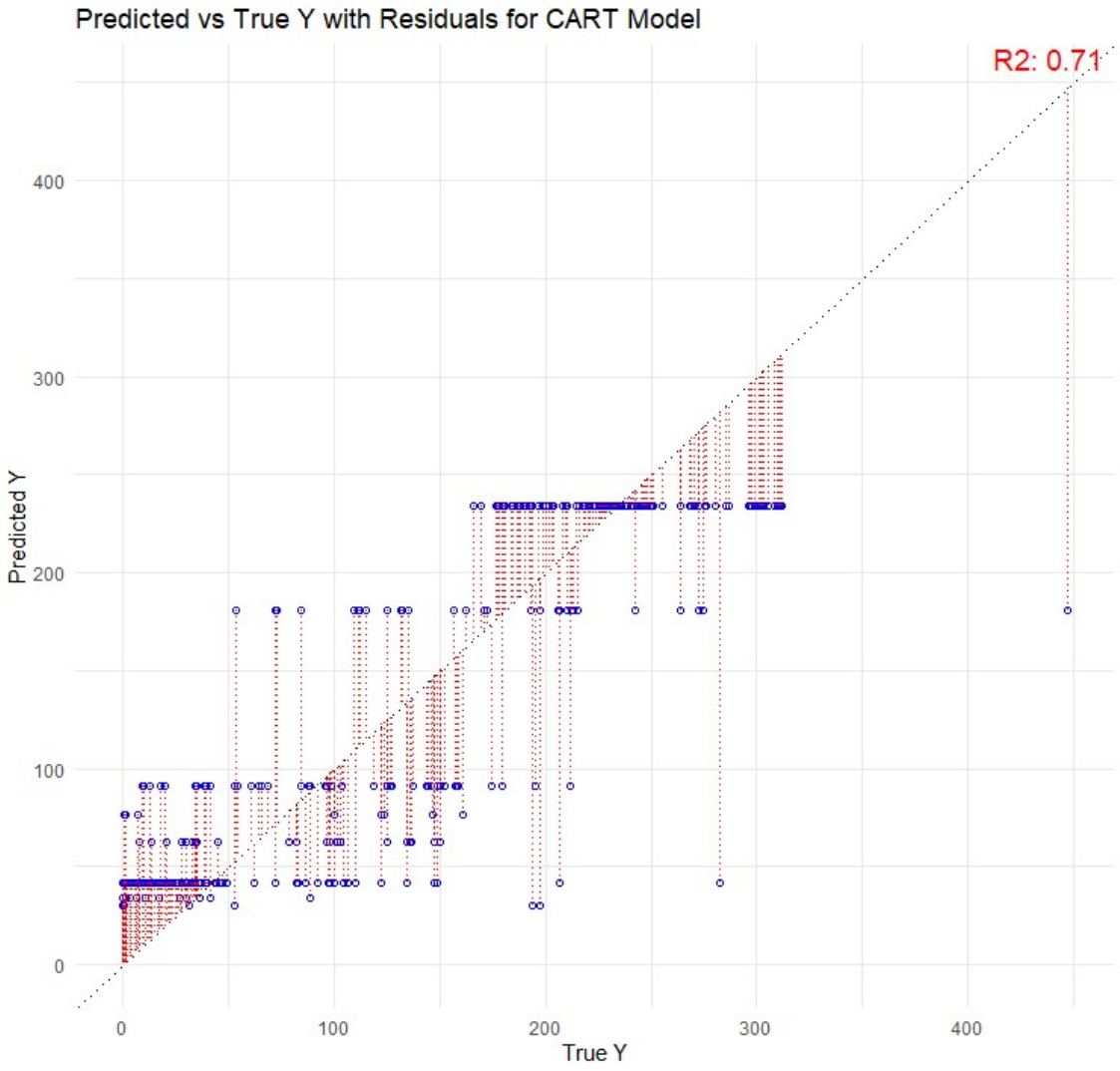
Root Mean Square Error (RMSE) measures the square root of mean squared errors between predicted and actual values. Lower RMSE values indicate better prediction accuracy.

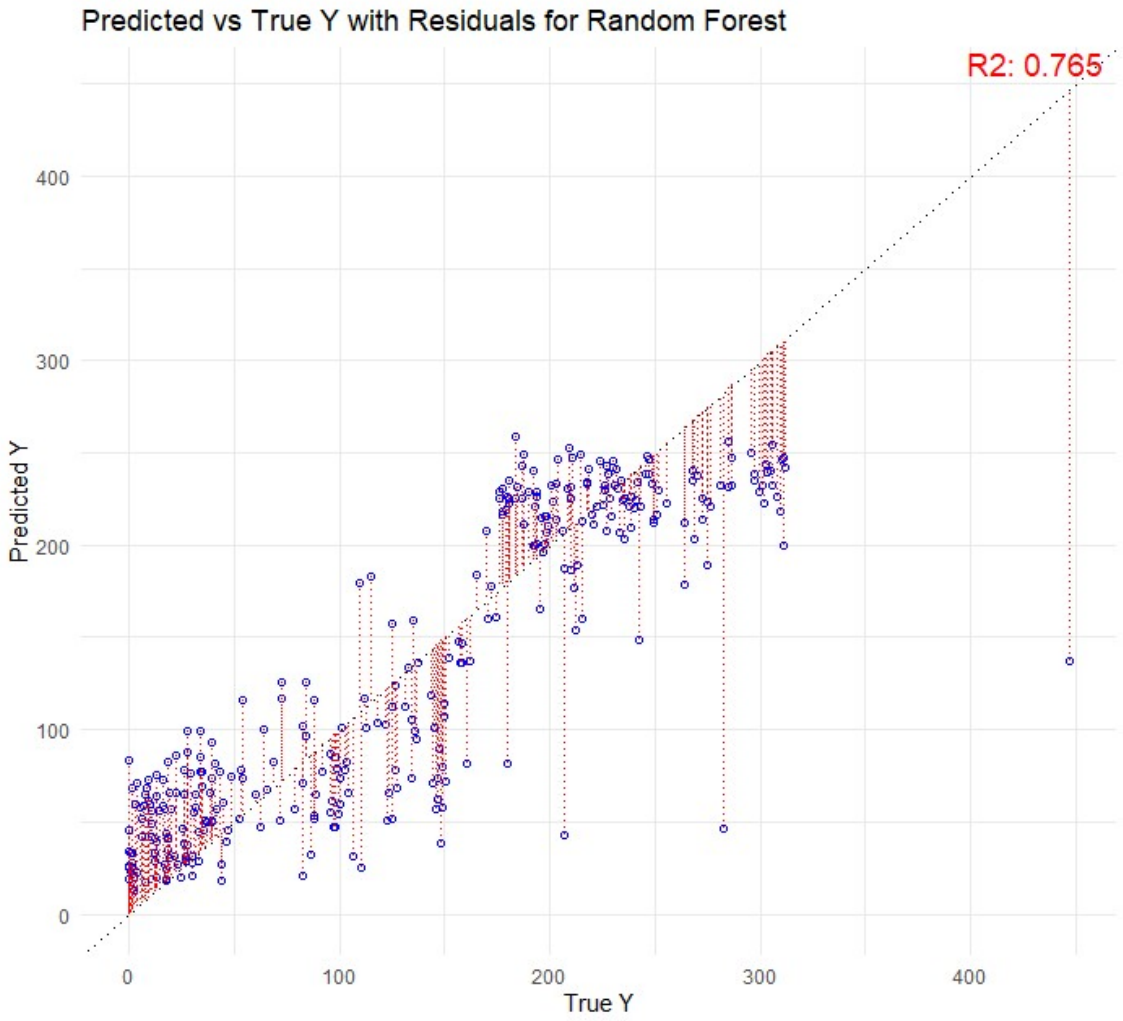
Mean Absolute Error (MAE) measures the average absolute difference between observed and predicted values, without overly penalising large errors. Lower MAE values indicate better prediction accuracy.

R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model.

The scatter plots of predicted vs true values, with indication of residuals and r^2 value:




Predicted vs True Y with Residuals for CART Model
R2: 0.71
Predicted Y
True Y
0
100
200
300
400


Predicted vs True Y with Residuals for Random Forest
R2: 0.765
Predicted Y
True Y
0
100
200
300
400