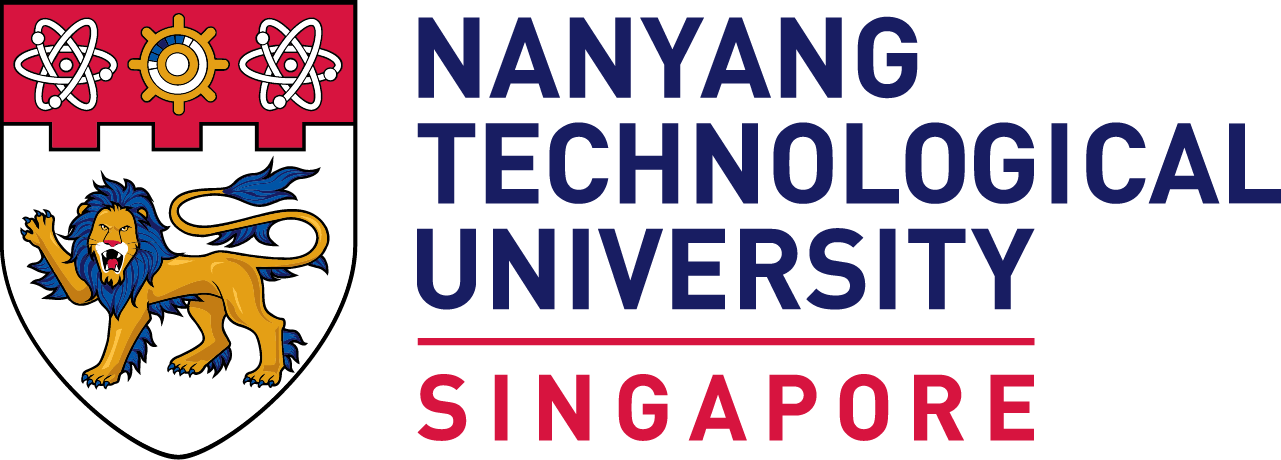
### 



**BC2406-ANALYTICS I: VISUAL & PREDICTIVE TECHNIQUES**

**A/Y 2023-2024 Semester 1**

Lynn Chia Xiao Lian U2210598G

Seng Swee Chong U2211710D

Renee D’ Angela Raymond U2210393C

Sim Yuyi U2210834B

Donovan Goh Jun Heng (U2220717H)

Nguyen Ngoc Tram Anh U2010920A

[**Executive Summary 4**](#_8013c4koqxtf)

[**1. Project Background 5**](#_si2yir8eb17a)

[1.1. Business Problems 5](#_74kpws9xpg4c)

[1.2. Existing Solutions 5](#_hjh6mlr4ku)

[1.3. Our Approach 6](#_u3cm38dxez9x)

[**2. Corrosion Dataset 7**](#_ks2eurbnt1ay)

[2.1. Introduction to Dataset 7](#_fz9unmwpy5jh)

[2.2. Data Preparation 7](#_pe8jzd5deoc8)

[2.3. Data Exploratory Analysis 7](#_ose5d6kw7qja)

[2.3.1. Correlation Matrix 7](#_1vyru97a06h7)

[2.3.2. Key Visualisations 8](#_vwz6lk95tpux)

[**3. Corrosion Predictive Modelling 9**](#_4x0ige9464lh)

[3.1. Linear Regression 9](#_1mxrce47ub43)

[3.1.1. Introduction to Linear Regression 9](#_we9snw3o2gf)

[3.1.2. Preliminary Model 9](#_m8jqh65psbfw)

[3.1.3. Linear Regression with Backward Elimination 10](#_oooa4ljvjucb)

[3.1.4. VIF Check 11](#_dzlsh0u502ly)

[3.1.5. Diagnostic Plot 11](#_oda9yon5c5ic)

[3.1.6. Model Accuracy 12](#_9wefkfnmpcrc)

[3.2. CART 13](#_hn9vmekba0c9)

[3.2.1. Introduction to CART Modeling 13](#_i97fcb8j15qk)

[3.2.2. Our Dataset 14](#_af68et3f7vzo)

[3.2.3. Model Building 14](#_cvowbkrgdyta)

[3.2.3.1. Tree growing 14](#_etaojdemm6vi)

[3.2.3.2. Tree pruning 15](#_97qh637mkqn1)

[3.2.4. Variable importance 15](#_tywvp2ywmd38)

[3.2.5. Model Accuracy 16](#_qozjcw1hfq0f)

[3.3. Model Comparison 17](#_mcgkptglamy2)

[**4. Impact Evaluation 17**](#_lkg5rfbwgyhx)

[4.1. Oil Spill Dataset Introduction 17](#_41hugtvqpzwg)

[4.2. Dataset Preparation 17](#_92halkuy35r2)

[4.2.1. Cleaning of Variables with Too Much Missing Data 17](#_a5g9jyywjnqj)

[4.2.2. Creating New Variables 18](#_ojrcveegty2o)

[4.2.3. Assembling Relevant Variables 18](#_k3hkmx7xc1xl)

[4.2.4. Handling the Remaining NA Values 18](#_jxxuviu5ovum)

[4.2.5. Changing Variable Types 18](#_aw7rz5hbz782)

[4.2.6. File Export 19](#_3ltpp9o46aoa)

[4.3. Exploratory Analysis 19](#_a4tv21qr6y9n)

[4.4. Pipeline Shutdown Prediction 21](#_cbgaveehaiay)

[**5. Recommendations and Business Applications 21**](#_dekz9avjgfp7)

[**6. Limitations and Further Considerations 23**](#_tikwko73jvmm)

[**7. Conclusion 24**](#_91qj2xxskjef)

[**8. Appendix 25**](#_ow8qchllmwb9)

[**References 34**](#_yhbn18uarebh)

### 

### Executive Summary

This report aims to leverage data analytics and machine learning to solve the problem of internal corrosion for Aramco. Given the serious impacts of internal corrosion, it is in the best interest of Aramco to reduce the costs spent on rectifying or monitoring corrosion in pipelines.

Using the simulated corrosion dataset obtained from a reliable data source, our team employed various machine learning techniques, including Linear Regression and Classification and Regression Tree (CART) to develop models and identify factors influencing pipeline corrosion.

Our analysis indicates that CO2 pressure and pH are the contributing factors to the corrosion rate of pipelines. Based on this finding, we developed 2 solutions for Aramco to target the aforementioned 2 factors.

For carbon dioxide control, we suggest equipping Aramco's pipelines with carbon dioxide removal technology. By integrating it into their oil pipelines, we can effectively reduce CO2 pressure by capturing those CO2 emissions in the pipelines. However, the adaptation of Direct Air Removal technology in oil pipelines faces challenges, including its novelty to the industry, as companies like Aramco have yet to implement it, thus potentially requiring extensive pipeline modifications and a thorough cost-benefit analysis due to its economic implications. Despite these hurdles, if this technology is adopted, it could position Aramco at the forefront of innovation, enhancing its competitive stance in the oil industry.

For pH control, an automatic monitoring system with precise chemical dosing could be implemented, with careful selection of reactants to avoid adverse chemical reactions that might affect the purity of the oil obtained. Thus, this solution would require further consultation/collaboration with a professional chemist/chemical engineer with relevant domain knowledge.

The 2 solutions can be implemented in conjunction. By identifying the influencing factors, Aramco can identify potential pipeline corrosion and adjust those factors to reduce corrosion. From here, Aramco can conduct further analysis to implement measures to conduct timely predictive maintenance.

#### 

#### 

### 1. Project Background

Aramco, officially known as Saudi Arabian Oil Company, is one of the world's leading integrated energy and chemicals companies. Aramco has the world's largest proven crude oil reserves and the world's largest daily oil-producing company. Armaco faces a multitude of challenges and we have flagged out corrosion to be one of the major issues that they are facing. (Aramco, 2019)

#### 1.1. Business Problems

Internal corrosion in pipelines is one of the greatest adversities that Aramco faces, resulting in operational inefficiency, substantial financial losses, as well as safety and sustainability concerns (Aramco, 2019). The cost of corrosion globally is estimated to be around $2.5 Trillion, indicating that this is a significant problem not only in the oil & gas industry.

The effective management of corrosion, development of new technologies, routine inspections, detections, corrections, and preventions can significantly reduce such risks and costs.

Leveraging data and analytics to pinpoint the most significant factors contributing to internal pipeline corrosion will empower Aramco to address this critical challenge, allowing them to efficiently and effectively pinpoint the key factors that result in oil spills.

#### 1.2. Existing Solutions

Some of the current existing solutions that Aramco employs are Corrosion Inhibitors, Scale Inhibitors, Biocides, Oxygen Scavengers, Coating, Cathodic Protection, and Material Selection.

| Corrosion Inhibitors | Chemicals added to the fluid passing through pipelines form a protective layer on the metal’s interior surface, preventing internal corrosion. |
| --- | --- |
| Scale Inhibitors | Chemicals are used to prevent scale build-up in pipelines, which can obstruct flow in pipelines. |
| Biocides | Chemical substances or microorganisms that kill bacteria, fungi, and algae, which can lead to corrosion if not managed properly as these microorganisms can grow within pipeline systems. |
| Oxygen Scavengers | Apparatus that reduce oxygen levels in the liquids contained within pipelines, as oxygen can lead to corrosion for pipelines containing metal parts. |
| Cathodic Protection | Either using a sacrificial metal which is more reactive to protect the pipeline, or passing electrical currents through the metal pipelines to prevent corrosion. |
| Material Selection | Material selection entails the careful selection of materials that would be resistant to corrosion, and they depend on many factors such as the environment in which the pipeline is built, the liquids passing through the pipelines, etc. Corrosion-resistant alloys or coatings are often chosen to extend the lifetime of pipelines. |

Aramco has also invested in technologies like Inductosense WAND technology (Aramco, 2023) that help monitor corrosion, such as sensors which help detect wall thickness, making the collection of data on corrosion easier and more accurate.

#### 1.3. Our Approach

Big Data and Machine Learning are potentially powerful tools for the oil industry in general, and for corrosion management in particular (Eliyan, 2023). In this project, we intend to apply 2 machine learning techniques, namely Linear Regression and CART (Classification And Regression Tree).

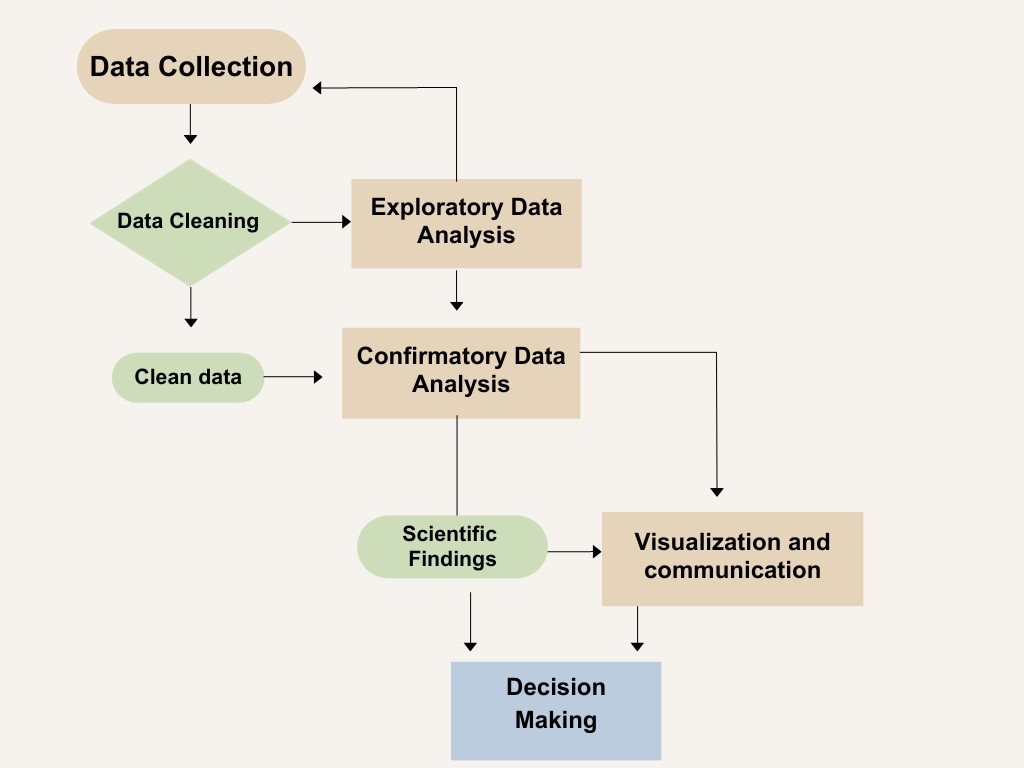


Fig. 1 Our approach to decision-making by machine learning

### 2. Corrosion Dataset

#### 2.1. Introduction to Dataset

This is a simulated dataset obtained from Mendeley data, a reliable data source.

The experimental and field measurements of internal corrosion rates are scarce due to the cost and difficulty of inspection/monitoring techniques. As such, we will be using simulated data. The corrosion dataset contains the corrosion rates of oil pipelines under harsh conditions. The Corrosion rate figures act as the target variable in this dataset. Apart from Corrosion rate, the dataset also contains 7 other variables which are: Temperature, Flow velocity, CO2 pressure, Internal pressure, Corrosion Inhibitor efficiency, Shear stress & pH. With this dataset, we aim to develop a predictive model that could predict the corrosion rates of oil pipelines with precise accuracy.

The dataset contains 243 data points and 8 continuous variables. The overview of data variables **(refer to** [**Appendix A**](#m1p7vd2nsea7)**)**

#### 2.2. Data Preparation

The data does not contain any missing or duplicated values. As a result, there is no need to clean the data. The distribution for all variables is also normal **(refer to** [**Appendix B**](#jgjitdvje0x0)**)**.

For efficient analysis and modelling, we rename the variables as follows:

“Flow velocity” -> “Flow\_velocity”, “CO2 pressure” -> “CO2\_pressure”, “Internal pressure” -> “Internal\_pressure”, “Corrosion Inhibitor efficiency” -> “Corrosion\_Inhibitor\_efficiency”, “Shear stress” -> “ Shear\_stress”, “Corrosion rate” -> “Corrosion\_rate”

#### 2.3. Data Exploratory Analysis

##### 2.3.1. Correlation Matrix

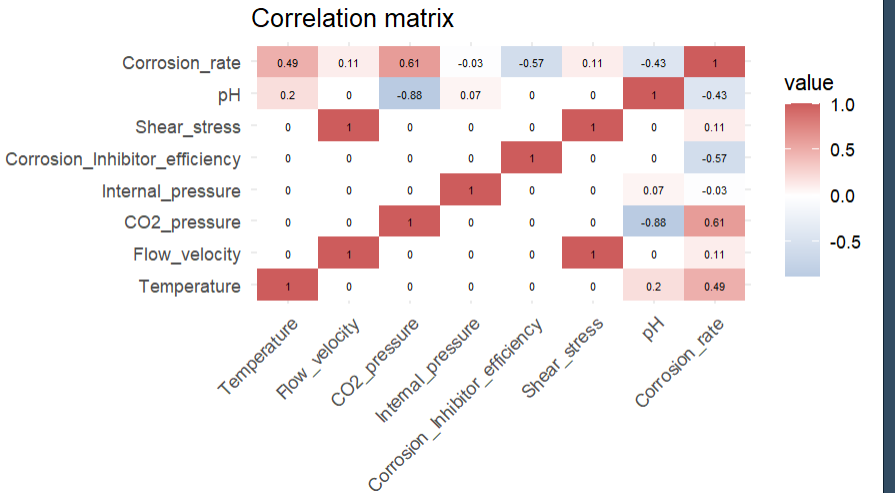


Fig. 2 Correlation Matrix

From Figure 2, we can infer that “Corrosion\_rate” has a strong positive correlation with “CO2\_pressure” and “Temperature” while it has a strong negative correlation with “Corrosion\_inhibitor\_efficiency” and “pH”. This means that when Temperature and CO2 pressure in pipes increase, the rate of corrosion also increases. Conversely, when pH and Corrosion Inhibitor efficiency decrease, the rate of corrosion increases.

Another noteworthy finding is that “Shear\_stress” and “Flow\_velocity” have a perfect correlation of 1. This suggests that “Shear\_stress” and “Flow\_velocity” can essentially be used interchangeably in a predictive model.

##### 2.3.2. Key Visualisations



Fig. 3 Boxplot of Corrosion Rate against Temperature Fig. 4 Boxplot of Corrosion Rate against CO2 Pressure

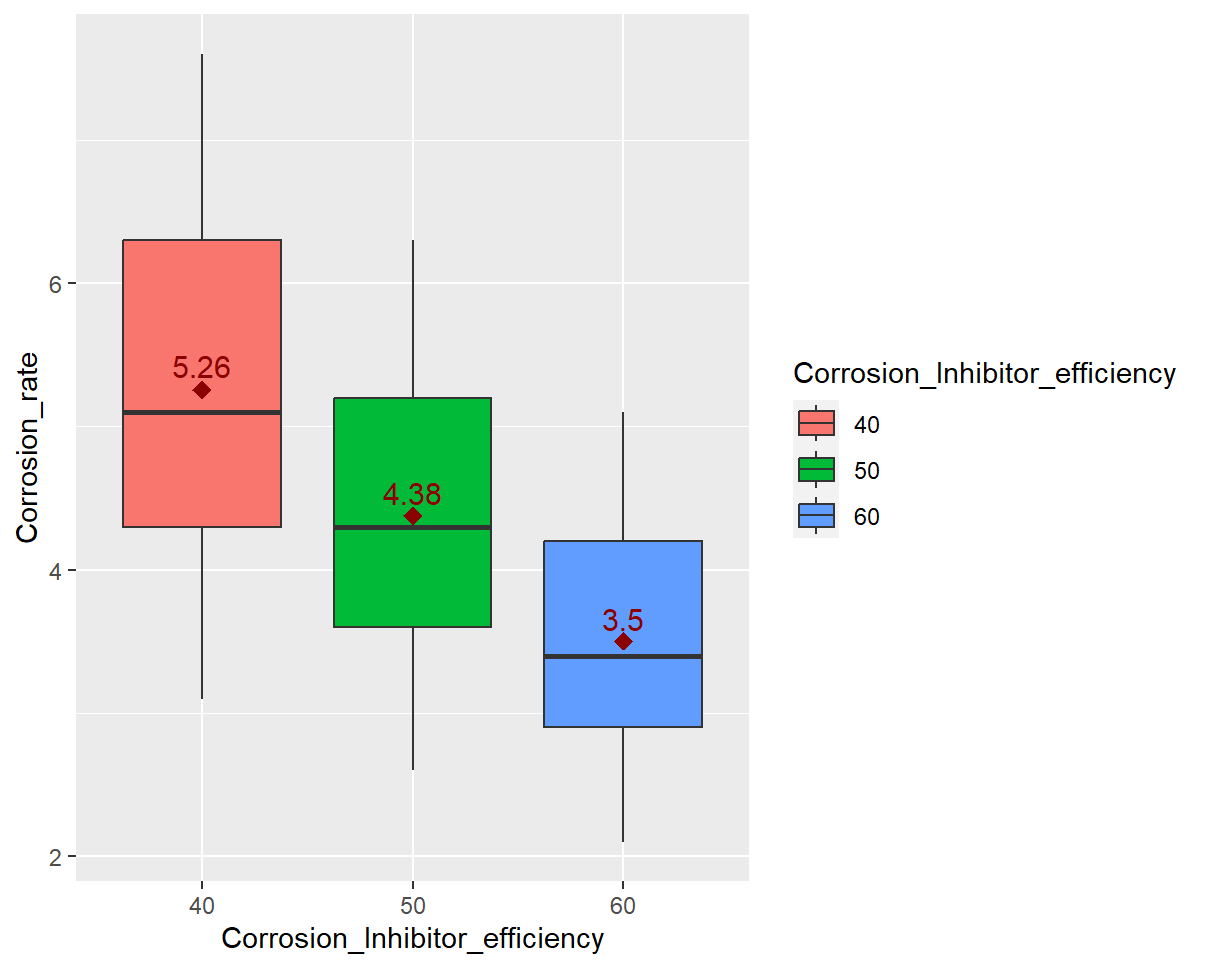
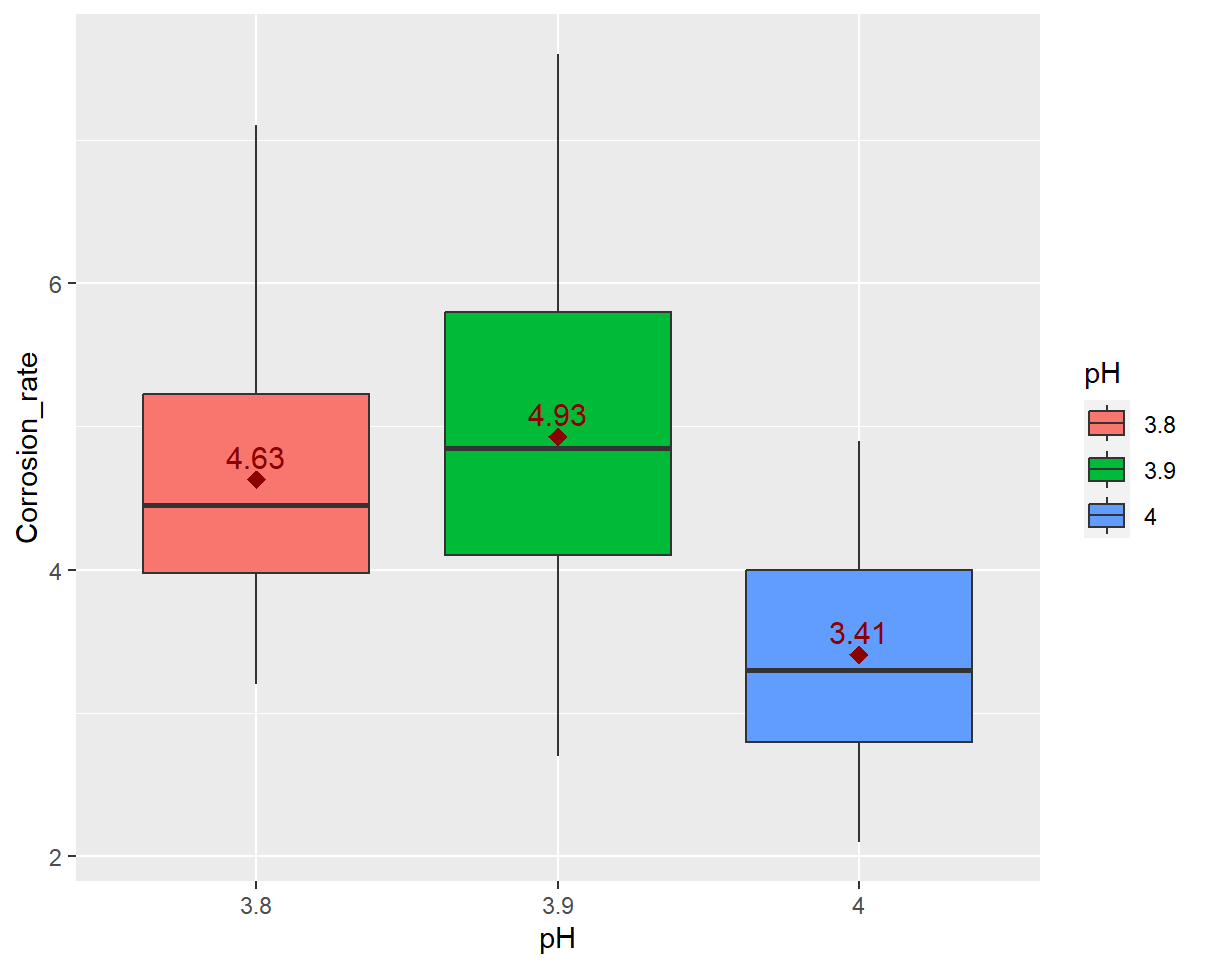


Fig. 5 Boxplot of Corrosion Rate against pH Fig. 6 Boxplot of Corrosion Rate against CO2 Inhibitor Efficiency

These boxplots provide a clearer illustration of the strong correlation between corrosion rate and each of the four aforementioned variables.

### 3. Corrosion Predictive Modelling

The dataset is divided into trainset and testset with a 70:30 ratio. Due to the random nature of data splitting, we must utilize the set.seed() function so that we get the same split every time to ensure the reproducibility of results. Here, we use set.seed(2004).

#### 3.1. Linear Regression

##### 3.1.1. Introduction to Linear Regression

Linear regression analysis is a statistical technique employed to make predictions about one variable based on the value of another. The variable we want to forecast is referred to as the dependent variable. Simultaneously, the variable used to predict the value of the dependent variable is termed the independent variable. (IBM, nd)

Mathematical relationships between the dependent and independent variables are uncovered by estimating the coefficients of a linear equation. This equation involves one or more independent variables and it aims to provide the best possible prediction of the dependent variable. To achieve this, linear regression fits a straight line or a higher-dimensional surface that minimizes the discrepancies between predicted and actual output values. (IBM, nd)

One common method for performing linear regression involves the use of a "least squares" approach, which identifies the optimal fit line or surface for a given dataset. By employing this method, one can estimate the value of the dependent variable (Y) from the independent variable (X). This introduction provides a foundation for understanding the fundamental principles of linear regression analysis and its practical applications in various fields. (IBM, nd)

##### 3.1.2. Preliminary Model

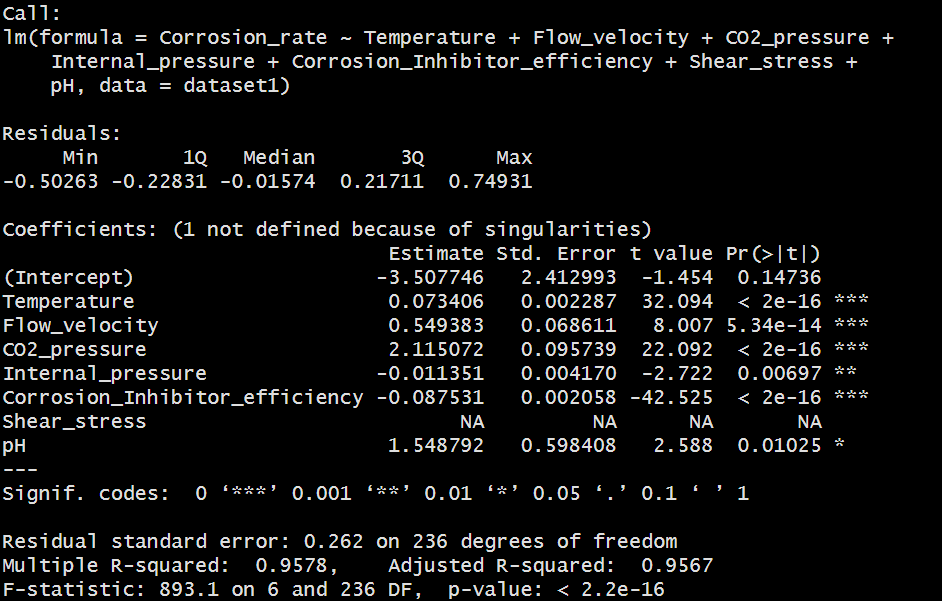


Fig. 7 Preliminary Linear Regression Model

From the correlation matrix, it can be seen that “Shear\_stress” has perfect correlation with “Flow Velocity” which can be seen through the correlation value of 1. This might explain why Shear Stress obtained NA result. Therefore, we could use backward elimination to exclude Shear stress as a factor in our analysis. However, since correlation does not imply causation, we still need to use our linear regression model to find out the significant factors affecting corrosion rate.

##### 3.1.3. Linear Regression with Backward Elimination

A model with too many predictors will train the model to follow data’s random variation (noise) and having too few predictors will produce a model that is not as accurate compared to a model with more predictors. Therefore, backward elimination is a variable selection technique used to select significant predictors by iteratively removing the least significant predictor to provide a reduced model. By eliminating irrelevant or less important predictors, this reduced model can help reduce the potential for overfitting and makes it easier for us to interpret the relationship between predictors and the response variable (Schneider et al., 2010).

This method is also useful for handling predictors which are highly correlated, like “Shear\_stress” and “Velocity\_Flow” thereby eliminating “Shear\_stress”

Backward elimination begins with a full model that includes all the potential predictors. In each step, the predictor with the highest p-value (i.e., the least significant predictor) is removed, provided its p-value exceeds a predetermined threshold (e.g., 0.05). The process stops when all predictors in the model are significant based on the chosen threshold, or when the removal of predictors doesn't lead to an improvement in a chosen model fit criterion (e.g., AIC).

AIC stands for Akaike Information Criterion, AIC evaluates the trade-off between the goodness-of-fit of the model and the complexity of the model.

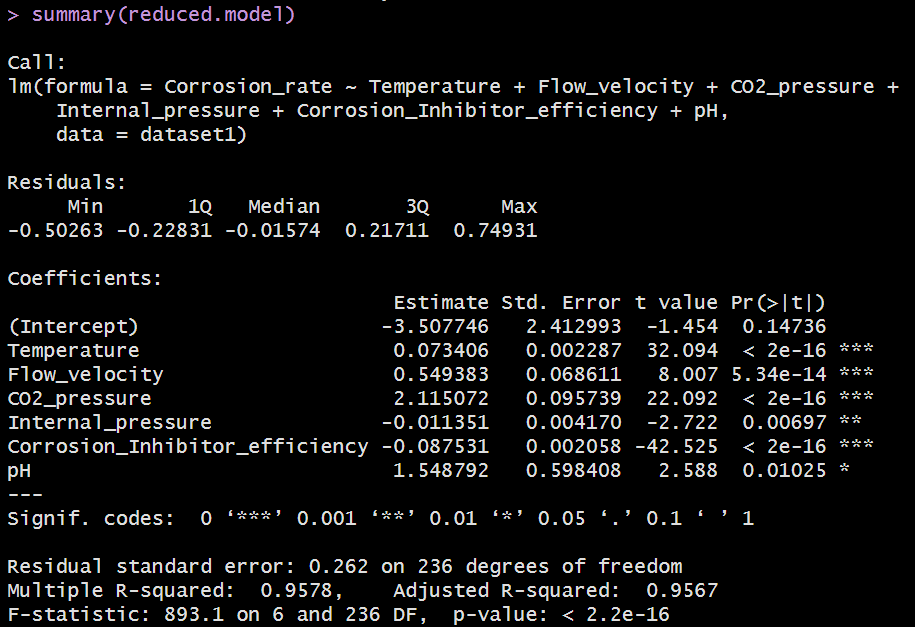


Fig. 8 Final Linear Regression Model after Backwards Elimination

The model obtained after performing backward elimination is our final model. As shown in the figure above, only important variables are kept in this model (variables with 1 to 3 stars).

##### 3.1.4. VIF Check

Variance Inflation Factor (VIF) is a tool used to quantify and detect multicollinearity in regression analysis. Multicollinearity is not very ideal as it means that it would be difficult for us to tell apart the individual effects of the predictors on the response variable, corrosion rate. The VIF values (all below 10) show that the remaining factors from the reduced model are not very highly correlated.

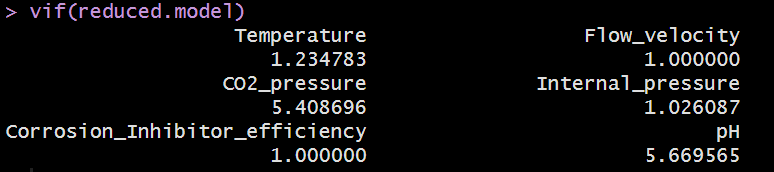


Fig.9. VIF

##### 3.1.5. Diagnostic Plot

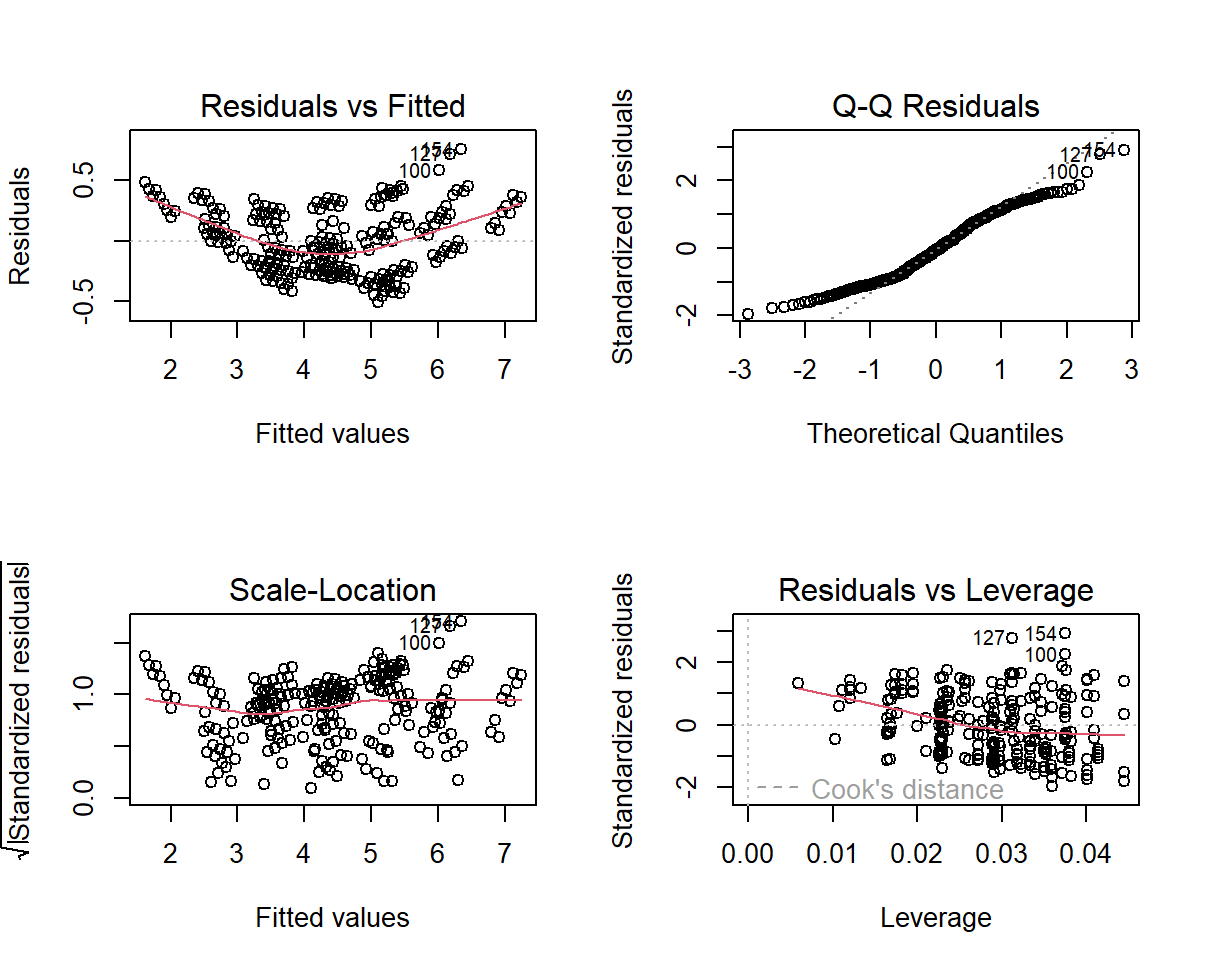


Fig.10. Diagnostic Plots

Diagnostic plots are a critical part of data analysis in statistics and machine learning, as they provide visual insight into the relationship between variables, the fit of a model if the predictions violate any of the 3 assumptions of linear regression (like homoscedasticity(constant variance of errors), normality of errors, and independence of errors) and the potential presence of outliers or anomalies.

The diagnostic plots show residuals in four different ways:

1. Residuals vs Fitted: Checks for a non-linear relationship between target variable Y and predictor variables X. The plot depicts a curvature possibly indicating a non-linear pattern in data.
2. Normal Q-Q: To check if residuals are normally distributed. Most of the variables in our linear regression follow the straight dash line with slight deviations at the tails. This suggests that residuals might have slight departures from normality with potential outlier
3. Scale-Location: To check if residuals are homogeneous/ spread equally along the ranges of predictors. Here we see that the residuals are evenly spread along the red line, which is good. This shows that it fulfills assumption 3 where errors are independent of X and has a constant standard deviation.
4. Residuals vs Leverage: Identifies influential cases or extreme values that can influence the regression results when included or excluded from the analysis. In this plot, we can see that there is no significant influential outlier.

Overall, the diagnostic plots are crucial for validating our linear regression model as it provides a visual means of identifying potential problems such as non-linearity, heteroscedasticity, outliers, and influential data points. Our diagnostic plot indicates that there are no major problems in the data that need to be rectified.

##### 3.1.6. Model Accuracy

The trainset and testset are assessed using Mean absolute percentage error (MAPE), Mean square error (MSE) and Root mean square error (RMSE) to gauge the model's accuracy.

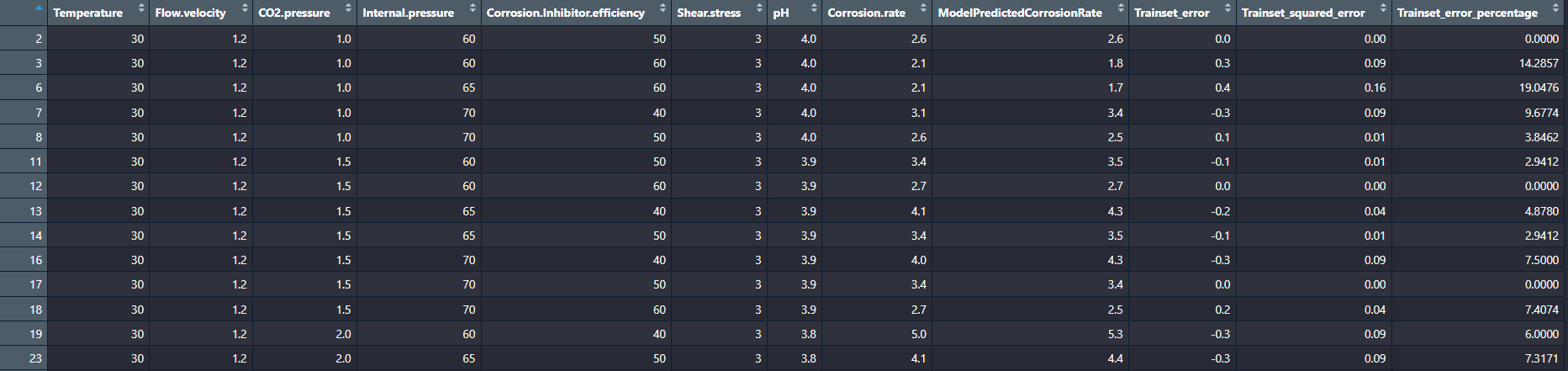


Fig.11. Trainset for Linear Model

**The percentage error of predicted values** for the trainset (Fig.11) ranges from **0% to 23.8095%** which is stable. **MAPE of 5.53%, MSE of 0.069 and RMSE of 0.262** suggest that the model demonstrates relatively low errors on average.

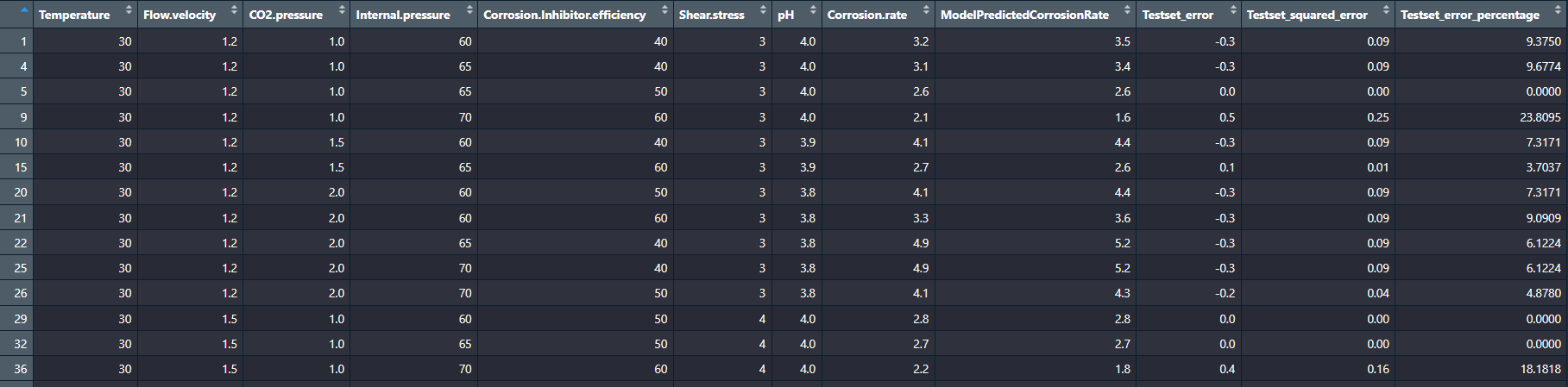


Fig.12. Testset for Linear Model

**The percentage error of predicted values** for the testset (Fig.12) ranges from **0% to 19.0476%** which is stable. **MAPE of 5.06%, MSE of 0.062 and RMSE of 0.248** suggest that the model demonstrates relatively low errors on average.

Consistently low average errors suggest that the model is proficient at making reasonably accurate predictions of the "Corrosion\_rate" based on the provided predictor variables and there was no overfitting of data.

#### 3.2. CART

##### 3.2.1. Introduction to CART Modeling

CART (Classification And Regression Tree) is a supervised Machine Learning predictive model with proven effectiveness in cause and effect analysis. It creates a decision tree by splitting the dataset multiple times according to certain cutoff values of the X-variables, which will be splitting criteria. Multiple subsets are then created, with the subsets no longer being split are called terminal nodes whereas the intermediate ones are called internal nodes. The predicted value in each terminal node is calculated as the mean value of all observations at that node. (Dutta, 2021)

There are two types of trees:

* Classification Trees: used when the target variable is categorical. The tree is used to find the "class" into which the target variable is most likely to fall.
* Regression trees: used when the target variable is continuous. The tree is used to forecast the value of this variable based on certain criteria.

To obtain the optimal Regression Tree, we need to grow the tree to the maximum, then prune it using the 1 SE rule after running 10-fold Cross-Validation. The optimal tree will avoid overfitting, while still having a low test set error.

Overall, CART’s strength lies in its readability, interpretability and its ability to handle missing values (Dutta, 2021).

##### 3.2.2. Our Dataset

The model is built on 6 out of 7 x variables (excluding “Shear stress” because it has a perfect correlation with “Flow velocity”), and the predicted variable is “Corrosion rate”. Because the target variable is continuous, we will use Regression Tree.

##### 3.2.3. Model Building

###### 3.2.3.1. Tree growing

Setting for CART model: minsplit = 20 (the minimum number of data points for each criterion split) to avoid growing the tree too big since we have nearly 250 data points, cp = 0 to grow the tree to maximum.

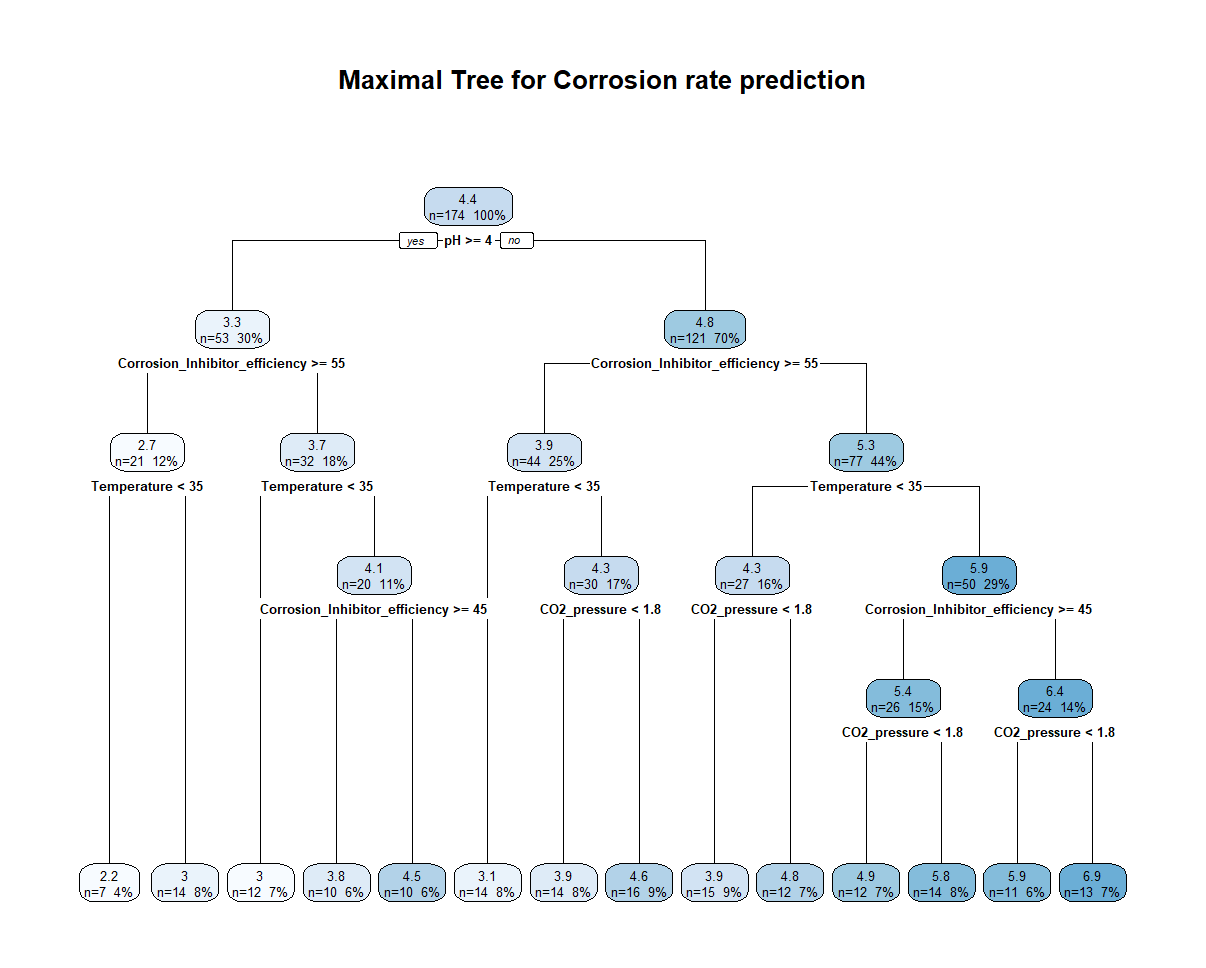
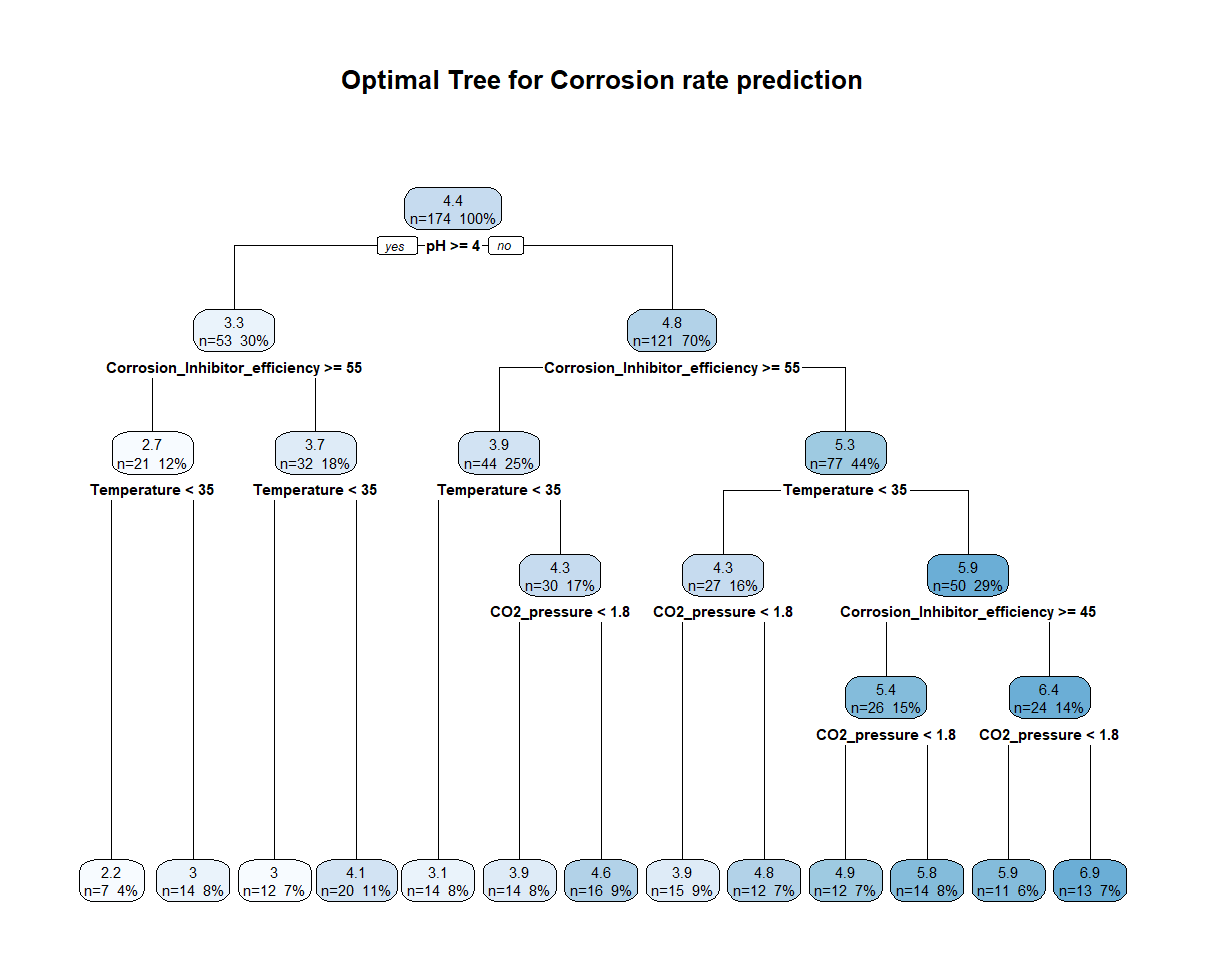


Fig.13. Maximal Tree for CART

The resulting tree has in total 27 nodes, including 1 root node, 12 intermediate nodes and 14 terminal nodes. For each split, if the answer to the criteria is “Yes”, we read the left node, and vice versa.

###### 3.2.3.2. Tree pruning

The original tree is usually not the optimal tree, so we perform pruning to make sure the optimal tree is derived as shown in Figure 13. The new optimal tree has 25 nodes in total, with 1 root node, 11 intermediate nodes and 13 terminal nodes.

Fig.14.Optimal Tree for CART

##### 3.2.4. Variable importance

Out of the 6 x variables used, only 4 variables are considered to have a significant impact on the predictive power of the model, thus largely influencing the predicted corrosion rate. The 4 variables include: pH (29%), CO2 pressure (29%), Corrosion Inhibitor efficiency (22%) and Temperature (19%). The 4 variables are also splitting criteria for our tree. The remaining 2 variables, Flow velocity and Internal pressure did not contribute to the predictive power of this model.

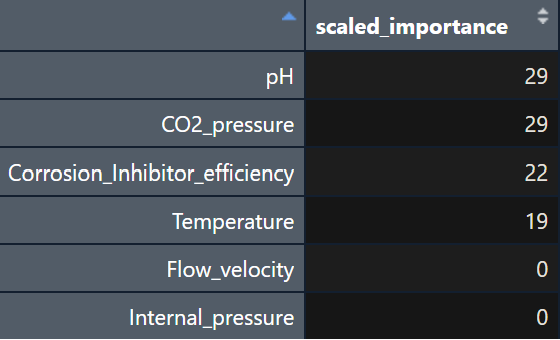


Fig.15.Variable Importance

##### 3.2.5. Model Accuracy

The train set and a test set are assessed using Mean absolute percentage error (MAPE), Mean square error (MSE) and Root mean square error (RMSE) to gauge the model's accuracy.

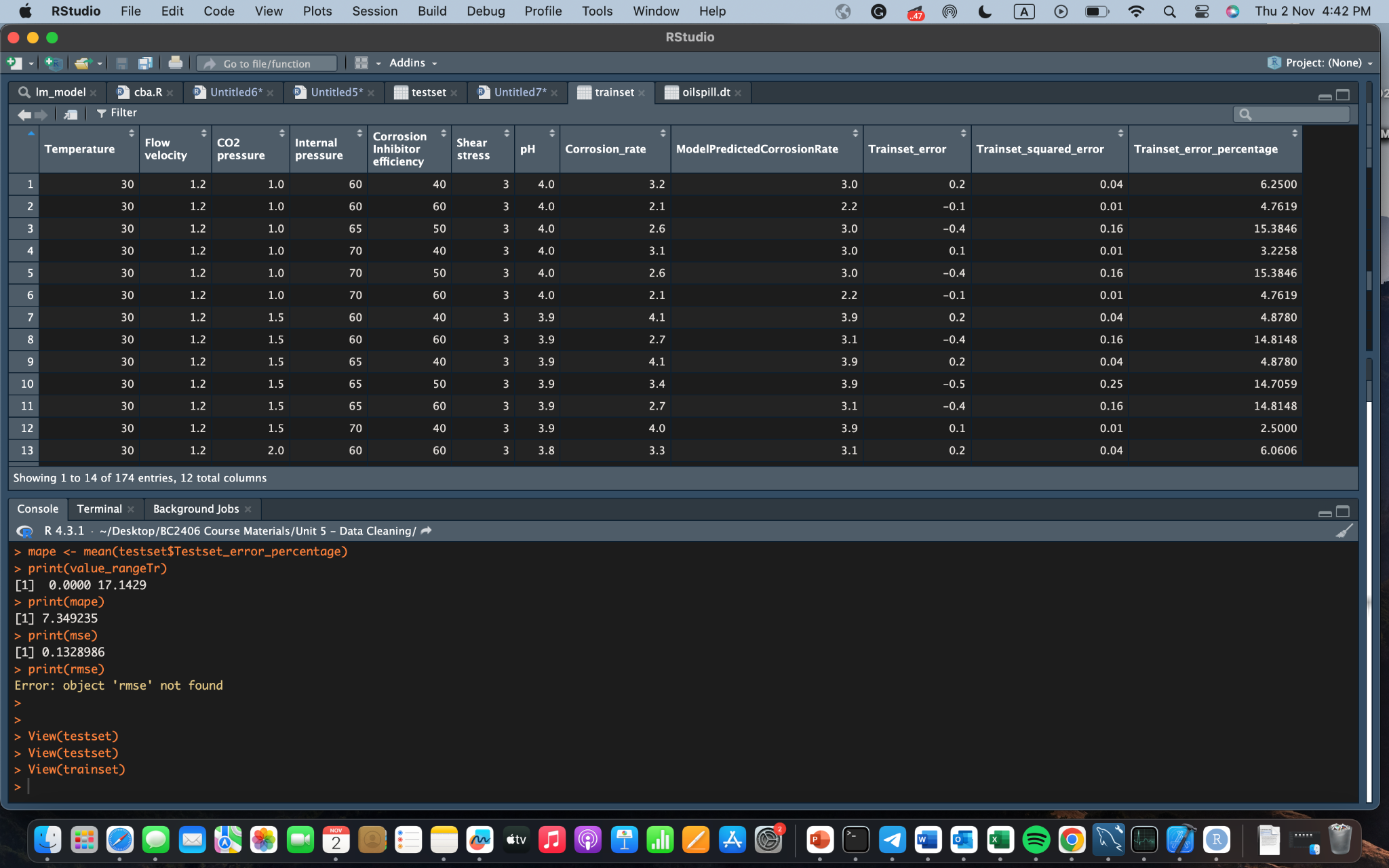


Fig.16.Trainset for CART model

**The percentage error of predicted values** for train set ranges from **0% to 20.5882%** which is stable. **MAPE of 6.12%, MSE of 0.101 and RMSE of 0.317** suggest that the model demonstrates relatively low errors on average.

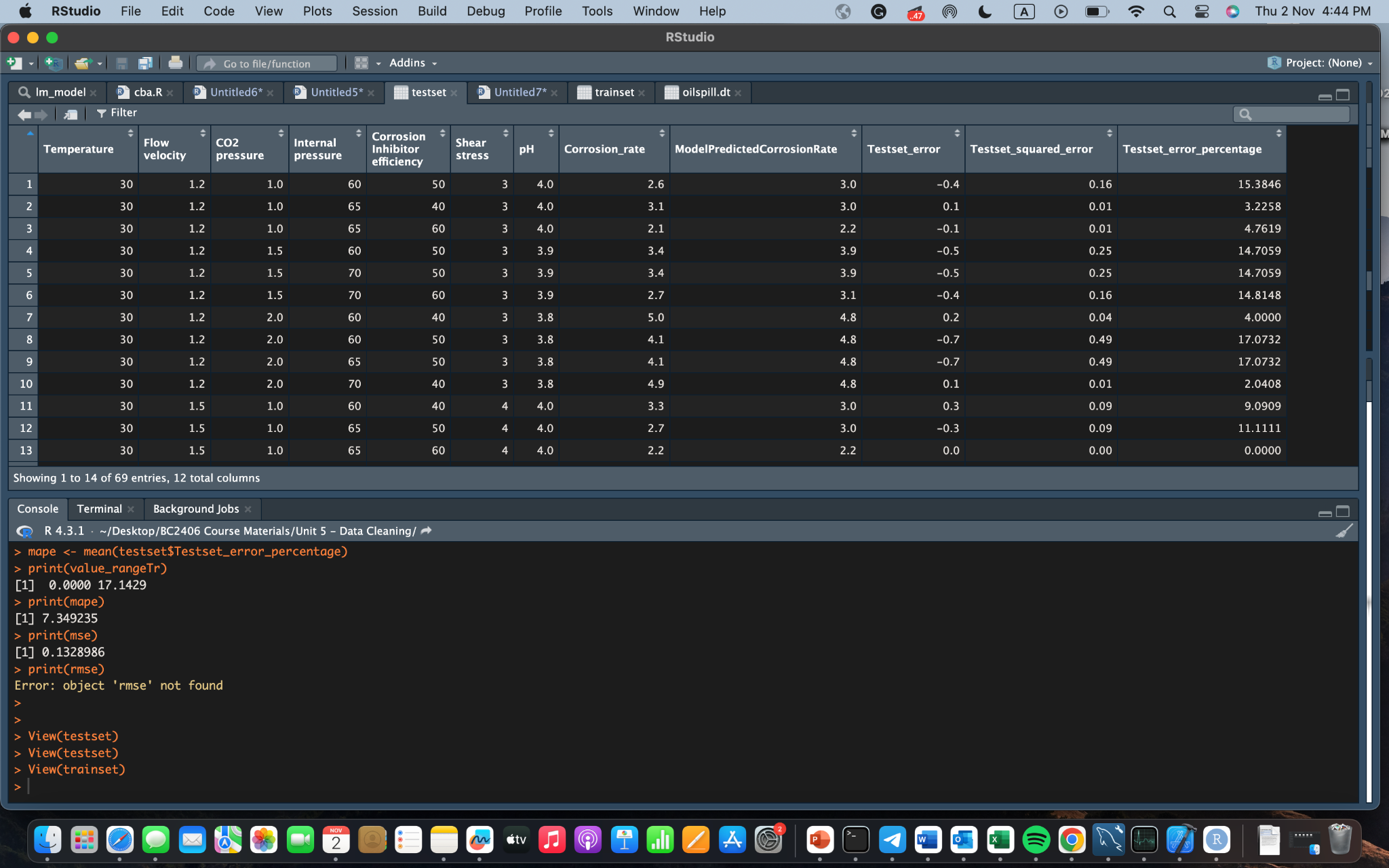


Fig.17.Testset for CART model

**The percentage error of predicted values** for test set ranges from **0% to 17.1429%** which is stable. **MAPE of 7.35%, MSE of 0.133 and RMSE of 0.365** suggest that the model demonstrates relatively low errors on average.

Consistently low average errors suggest that the model is proficient at making reasonably accurate predictions of the "Corrosion\_rate" based on the provided predictor variables and there was no overfitting of data.

#### 3.3. Model Comparison

|  | Linear regression | | CART | |
| --- | --- | --- | --- | --- |
| Train | Test | Train | Test |
| MAPE (%) | 5.53 | 5.06 | 6.120 | 7.350 |
| MSE | 0.069 | 0.062 | 0.101 | 0.133 |
| RMSE | 0.262 | 0.248 | 0.317 | 0.365 |

Fig.18.Model Comparison between Linear Regression and CART

After assessing the performance of our models, it becomes evident that the linear regression model consistently outperforms the CART model across various metrics, boasting lower Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE).

This advantageous outcome can be primarily attributed to the notably strong linear relationships that exist between our predictor variables and the target variables. Moreover, it's worth noting that linear regression holds a distinct advantage in its simplicity and ease of comprehension, making it a valuable tool in our analysis and explanation.

### 4. Impact Evaluation

#### 4.1. Oil Spill Dataset Introduction

To evaluate the potential impact of effective corrosion management, we analyzed this “Oilspill” dataset. The dataset is obtained from Kaggle, a reliable public data source. It includes a record of each oil pipeline leak or spill reported to the Pipeline and Hazardous Materials Safety Administration since 2010. The oil pipeline accident reports were collected and published by the DOT's Pipeline and Hazardous Materials Safety Administration.

There are 48 variables and 2795 data points. (**Refer to** [**Appendix C**](#wjsgeny6darf)**)**

#### 4.2. Dataset Preparation

##### 4.2.1. Cleaning of Variables with Too Much Missing Data

Some variables have too many missing values (from 457 to 2783 NAs), indicating a lack of data collected **(refer to** [**Appendix D**](#oxwdxysh6cjx)**)**. As a result, we will remove those variables from the analysis. Those variables include: “Public Evacuations”, “Operator Employee Injuries”, “Operator Contractor Injuries”, “Emergency Responder Injuries”, “Other Injuries”, “Public Injuries”, “All Injuries”, “Operator Employee Fatalities”, “Operator Contractor Fatalities”, “Emergency Responder Fatalities”, “Other Fatalities”, “Public Fatalities” and “All Fatalities”.

##### 4.2.2. Creating New Variables

We created 2 new columns: “Shutdown\_hours” and “Barrels loss (%)”.

“Shutdown\_hours” value is derived from “Pipeline Shutdown”, “Shutdown Date/Time” and “Restart Date/Time”. If “Pipeline Shutdown” indicates “YES”, meaning there was a pipeline shutdown, we calculate the shutdown durations by finding the difference between “Shutdown Date/Time” and “Restart Date/Time”.

“Barrels loss (%)” is derived from “Unintentional Release (Barrels)” and “Net Loss (Barrels)”. “Unintentional Release (Barrels)” is the total number of barrels released due to the incident, from which a part of the barrels may be recovered. “Barrels loss (%)” indicates the percentage of barrels undesirably released but can not be recovered.

##### 4.2.3. Assembling Relevant Variables

With focus on impact of corrosion, among the remaining variables, we only keep the following: “Cause Category”, “Cause Subcategory”, “Net Loss (Barrels)”, "Pipeline Location", "Pipeline Type", "Liquid Type", "Liquid Ignition", "Liquid Explosion", "Pipeline Shutdown", “Property Damage Costs”, “Lost Commodity Costs”, “Public/Private Property Damage Costs”, “Emergency Response Costs”, “Environmental Remediation Costs”, “Other Costs”, “All Costs”.

##### 4.2.4. Handling the Remaining NA Values

For the remaining variables, there are 51 NAs. A closer look reveals that the NA values are from columns 13-17, the columns regarding costs. This may indicate the lack of data on costs collected. Thus, we replace NA values with 0 to preserve data integrity and ensure the total costs are not affected.

From the summary, it can be observed that “Pipeline Type” and “Pipeline Shutdown” have some data points with “ ” value. This type of value does not affect our exploratory analysis, but it will create another level for the 2 categorical variables and cause problems when we build models. Because of that, we first change them into NA values. For preliminary visualizations, we do not handle them as they do not influence our focus on other cost analysis. Later as we try to build models to predict “Pipeline Shutdown”, we will remove the remaining NA values.

##### 4.2.5. Changing Variable Types

“Cause Category”, “Cause Subcategory”, “Pipeline Location", "Pipeline Type", "Liquid Type", "Liquid Ignition", "Liquid Explosion", and "Pipeline Shutdown" should be categorical variables. However, they are currently all string variables. Thus, we factorize them to become categorical variables.

##### 4.2.6. File Export

After data preparation, we export the cleaned dataset into a cleaned file called “oilspill\_cleaned.csv”. A snapshot of the cleaned dataset can be found in **(refer to** [**Appendix**](#4l1sp23d4qd6) **E)**.

#### 4.3. Exploratory Analysis

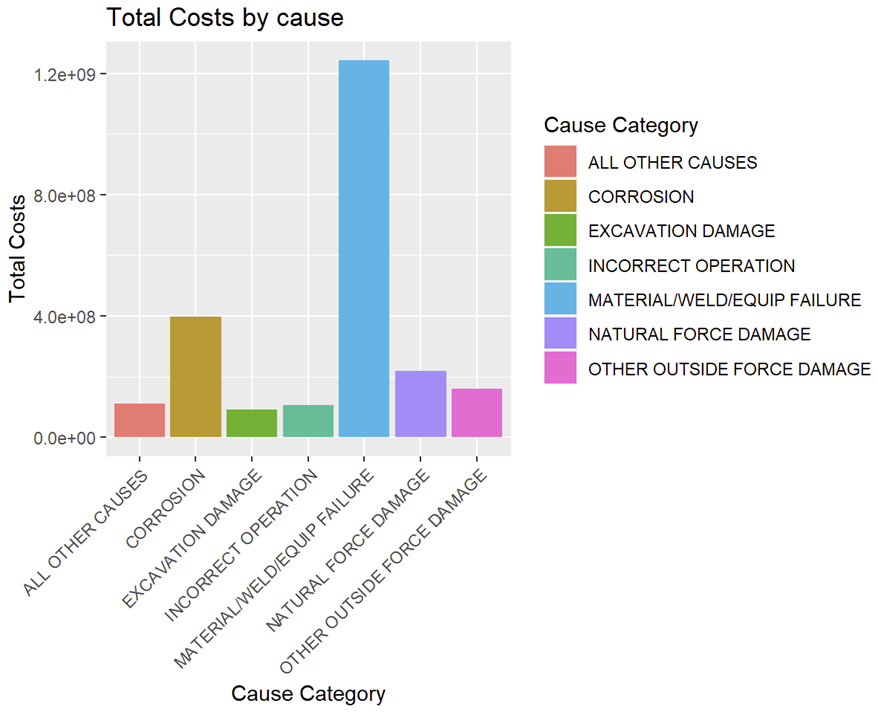
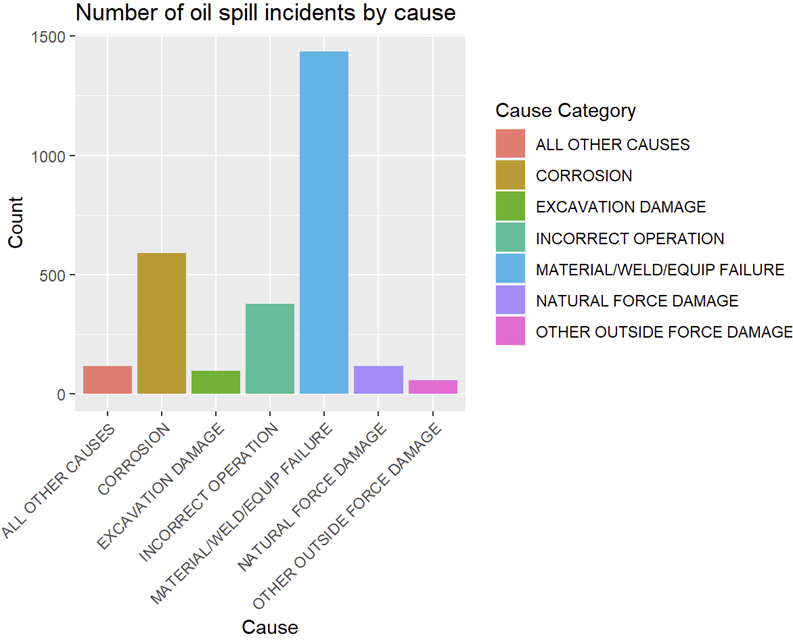


Fig.19. No.of Oil spill incidents Fig.20.Total Costs by Cause

Corrosion is the second most common cause for Oilspill (Fig.19.) and is also the second most expensive mistake (Fig.20). This proves that corrosion itself is a problem, but it also is likely to lead to more serious problems like oil spill.

For oil spills caused by corrosion, statistically:

* $395,325,677 in total had been lost due to corrosion, with an average of $667,780 per accident. The highest total cost incurred is $142,931,884 which happened in 2015.
* Nearly 50% of all accidents caused by corrosion had oil lost (counted by barrels of oil). A total of 40,137 barrels of oil had been lost, with an average of 137 barrels for each accident. Out of 592 accidents caused by corrosion, 122 cases (more than 20%) did not manage to retrieve any of the oil unintentionally released. The accident with the highest number of oil barrels loss recorded 8000 barrels lost, equal to 100% loss.

Considering 2023 price of $95.58/barrel, an average oil spill caused by corrosion would make the company lose $13094.46 on average (oil price taken from US Energy Information Administration).

* Among all oil spill accidents, 52.2% of the accidents experienced pipeline shutdown. The total number of hours of shutdown is 96,713 hours, with an average of 313 hours (~ more than 13 days) per accident. The longest shutdown duration recorded was 12475.13 hours (~ 520 days). While pipeline shutdown does not incur direct cost, it incurs costs of reparation, manpower and poses opportunity costs to businesses.

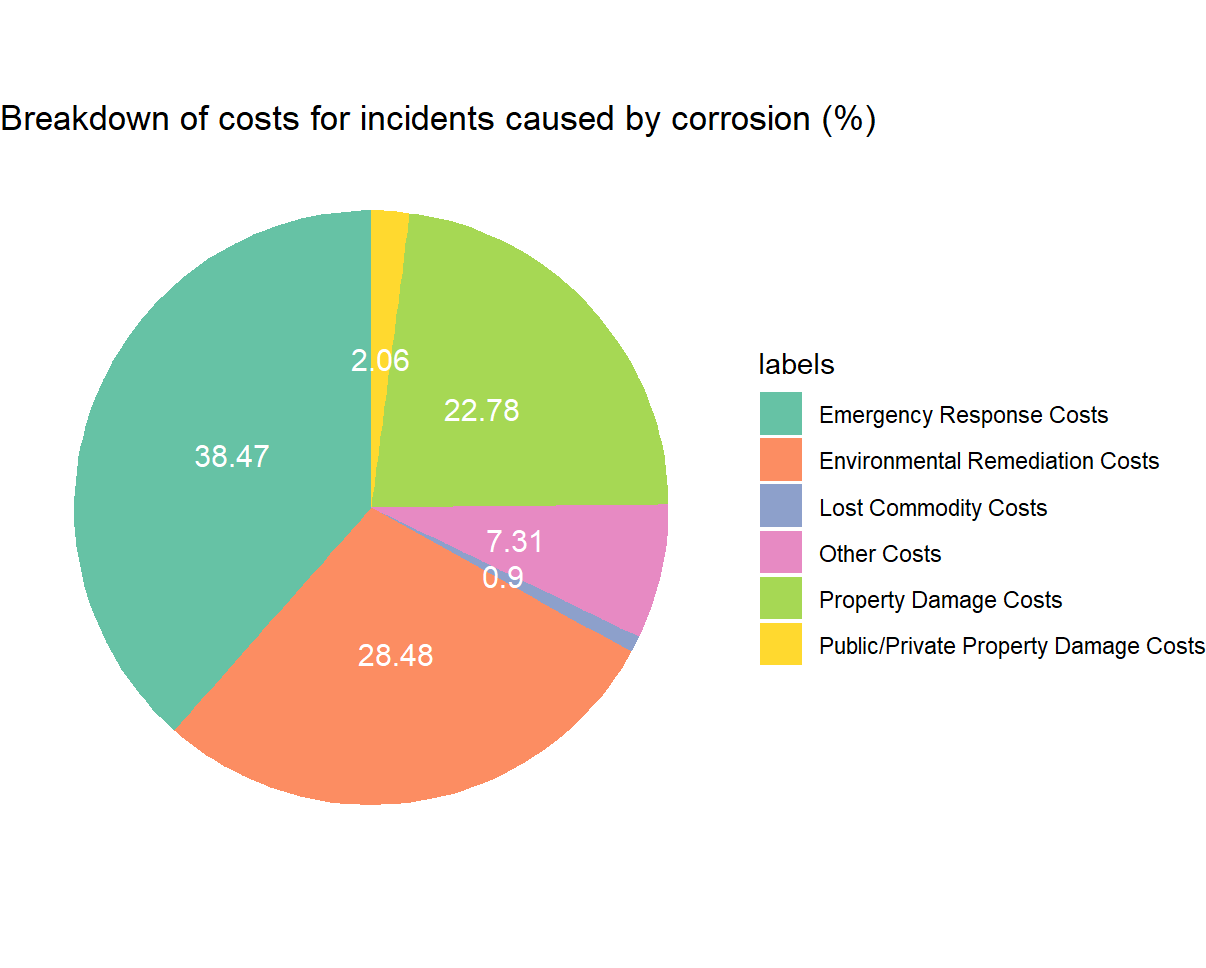


Fig.21.Breakdown of costs for incidents (Cause Category filtered to Corrosion)

Figure 21 shows the breakdown of costs businesses have to pay when an oil spill accident caused by corrosion happens. The largest percentage of costs go to emergency response costs (38.47%), environmental remediation costs (28.48%) and Property damage costs (22.78%). A more comprehensive breakdown of costs by year can be found in **(refer to** [**Appendix**](#5k5ijfh71my) **F)**.

All the costs listed above do not represent the total costs a business would have to bear following each accident. Yet, the figures are still considerable. Businesses may have some hidden costs to pay like social costs, economic damages to related parties and criminal fines to the State (Cohen, 2010).

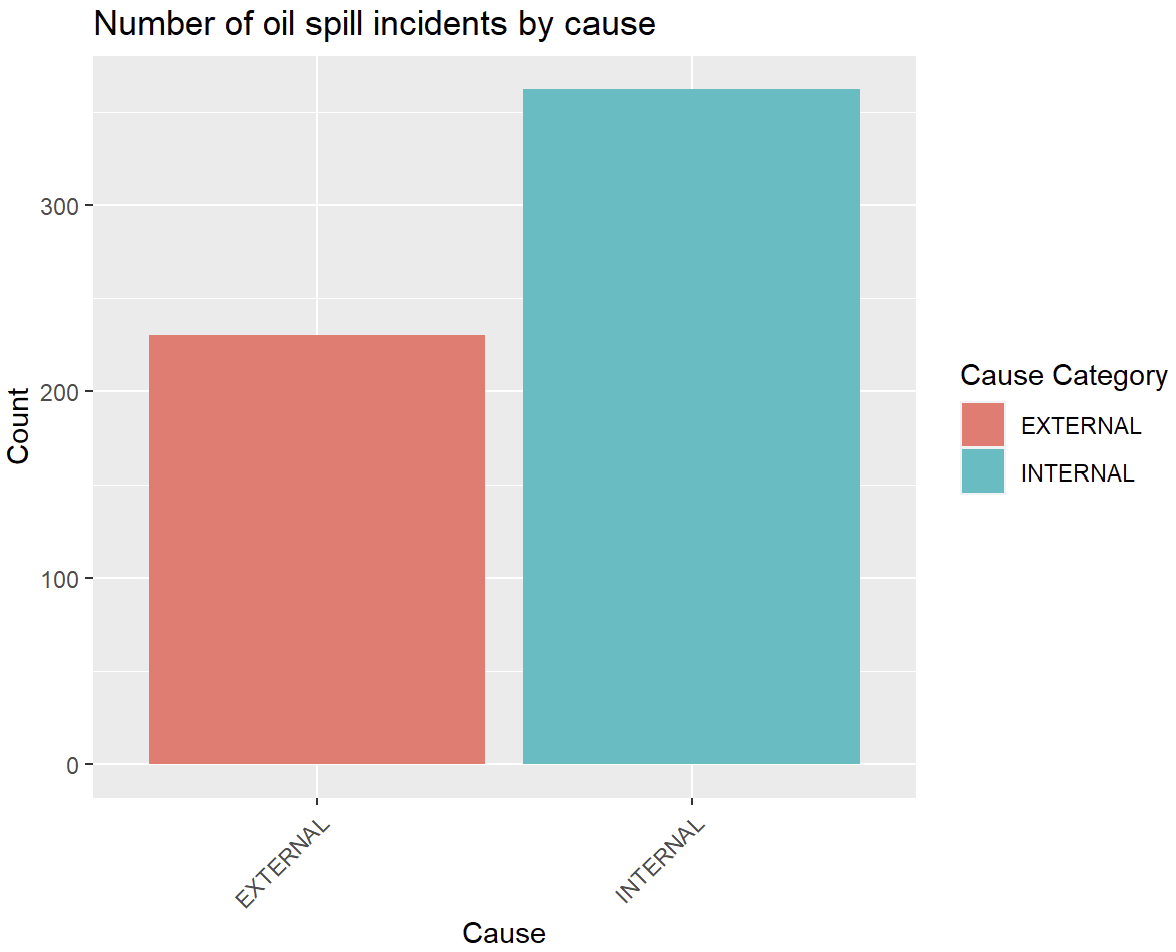


Fig.22. Number of oil spill incidents by cause

Figure 22 shows that internal corrosion is more likely to be the cause behind reported oil spills. This is because compared to external corrosion, internal corrosion is harder to detect and mitigate (O'Connor, 2020). This fact again emphasizes the importance of management for internal corrosion, which is the focus of our study.

#### 4.4. Pipeline Shutdown Prediction

As explained above, Pipeline shutdown is a huge opportunity cost for businesses. As a result, we attempted to build logistic regression and CART models to see under which condition (Pipeline Location, Pipeline Type, Liquid Type, Liquid Ignition, Liquid Explosion) that Pipeline Shutdown is most likely to happen.

The result from logistic regression is that Pipeline Type = Tank/Underground is significant. For CART, Pipeline Type is also the only significant factor. Moreover, model accuracy for both logistic regression and CART is very low (50-60%), indicating that neither of the 2 models holds sufficient predictive power for the target variable. **(Refer to** [**Appendix G**](#mkvtlczfb8ac)**)**

This could be due to the lack of data and necessary predictor variables (as this dataset is not focally aimed at Pipeline shutdown), and the fact that all Pipeline Locations for all shutdown cases are all “ONSHORE”, the result does not shed any light on factors that lead to Pipeline Shutdown due to Corrosion.

### 5. Recommendations and Business Applications

From our Linear Regression model, pH and CO2 pressure have the largest coefficient magnitudes of 1.55 and 2.12 respectively. This shows corrosion rates of pipelines are influenced the most by these variables. While linear regression can provide insights into trends and patterns, it might not be sufficient as it assumes linear relationships between variables. As a result, we can use a CART model to supplement our regression.

From our CART model, pH and CO2 Pressure are also the variables ranked as the most important, followed by Corrosion Inhibitor Efficiency and Temperature. The benefit of using a CART model is that we can see the splitting criteria for each split, and the nodes from the branch of pH > 4 had noticeably lower Corrosion Rates. This split was our first. Nodes with temperature below 35 and CO2 Pressure < 1.8 also experienced reduced Corrosion Rates compared to those without.

To validate the findings of both our models, research by domain experts was cross-referenced.   
  
Temperature:

At higher temperatures, the diffusion rate increases and synchronously the electrolyte resistance decreases, accelerating the corrosion. (Li et al., 2022) Lower temperature resulting in lower rates of corrosion is thus supported by industry experts.

pH:

The main corrosion product is FeCO3 due to the reaction of Iron and Bicarbonate (HCO3) ions (Kahyarian et al., 2017)

As pH increases, the corrosion rate decreases. At higher pH, it decreases the iron carbonate solubility and increases the iron carbonate precipitation rate to a level where iron carbonate precipitates as a dense, protective surface film which helps to reduce corrosion rates. (Nyborg, 2009)

In linear regression, every increase in pH results in a 1.55 increase in the rate of corrosion. However, our CART models show that pH > 4 results in a lower corrosion rate which ties in with our research. From our analysis of industry experts and understanding the limitations of the regression model (part 6), the CART model is more appropriate for pH.

CO2 Pressure:

An increase in CO2 pressure indicates an increase in the concentration of CO2. This results in increased formation of carbonic acid and a lower pH **(refer to** [**Appendix**](#n776ctc5ixbl) **H)**. This is supported by our correlation matrix which shows a strong negative relationship (Correlation of -0.88) between CO2 pressure and pH. Based on our research, CO2 pressure is a factor which affects pH. This shows multicollinearity and pH could have been removed from our initial model. However, since our VIF was within our threshold of 10, CO2 pressure was included in our final model.

Based on the above analysis, our business recommendations are reducing CO2 pressure and increasing the pH.

We can apply our analysis and business recommendations by targeting them individually.

**Solution 1: Carbon dioxide removal technology**

CO2 pressure can be reduced by using carbon dioxide removal technology. Oil pipelines can be equipped with carbon dioxide removal technology to directly remove carbon dioxide from pipelines.

We can leverage existing technology that combat the emission of carbon dioxide gas from industrial processes or Direct Air Removal technology which extracts CO2 from the atmosphere. This makes our solution feasible.

However, there are limitations to using Direct Air Removal technology in oil pipelines. Firstly, The oil industry, including Aramco, has not utilised such a technology in their pipelines. Significant modifications may be required to adopt it into Aramco’s pipelines.

Next, there are economic concerns. Adopting Direct Air Removal technology in oil pipelines may be costly and Aramco has to assess the cost-effectiveness of implementing Direct Air Removal technology and consider other factors such as energy requirements.

Limitations aside, Direct Air Removal technology in oil pipelines is a technological innovation breakthrough and could further cement Aramco as the leader in the oil industry, enhancing its reputation and market competitiveness.

**Solution 2: Automatic pH monitoring system**

pH can be monitored and maintained in a controlled environment. This can be done by equipping oil pipelines with an automatic pH monitoring system. Oil pipelines could be equipped with an automatic pH monitoring system which measures the pH level of the oil using a built-in probe and doses out a pH corrector as required. The pH corrector could be alkaline or acidic chemicals.

Other than the limitations experienced by the carbon dioxide removal technology, the Automatic pH monitoring system faces a critical drawback. Adding a pH corrector could result in unwanted chemical reactions. This could lead to unintended consequences that undermine its effectiveness. It could be mitigated by carefully selecting chemicals that are suitable and will not cause any unwanted chemical reactions.

**Further evaluations of solutions**

While both solutions could be used concurrently, carbon dioxide removal technology is a superior solution that can simultaneously raise pH and lower carbon dioxide pressure. This stems from the impact of carbon dioxide pressure on pH. By reducing carbon dioxide pressure, the pH level will experience a decrease, thereby effectively addressing the 2 most important variables in corrosion control.

### 6. Limitations and Further Considerations

1. The scarcity and high cost of data extraction led us to use simulated data for preliminary research. However, it is important to acknowledge that simulated data may differ from real-life data. As a result, we need more resources to extract real data, preferably from Aramco’s oil fields, to re-apply our model and conduct further analysis.
2. The corrosion data observed is not diverse in the variable value range. As a result, we can only observe behaviours for certain ranges and miss out on the other ranges. For example, temperature above 120 Celsius degrees may behave in the opposite way to temperature below 120 Celsius degree (Peng et al., 2021). The same problem can be observed in the pH variable. As explained in the Business Applications section, according to studies, a pH level of 5 or 6 induces less corrosion. However, the value range for pH in our dataset is centralised around 4, making the observation inconclusive. Again, this limitation emphasizes the need for more resources and real data.
3. We have studied the financial impact of effective corrosion control that helps eliminate the risk of oil spills caused by corrosion. Another serious consequence observed from the data is pipeline shutdown. With limited data from the oil spill dataset and the fact that the dataset is not focused on pipeline shutdown prediction, the model we built has limited predictive value. With more data, we can build better models and study the influencing factors of pipeline shutdown more extensively. Similarly, injuries and fatalities are 2 important factors that contribute to the severity of pipeline corrosion but they are not reflected in this dataset.
4. The Oilspill dataset reports oil spill accidents from 2010 to 2017. However, data for 2017 seems to be lacking, as there were only 2 records in January **(refer to** [**Appendix**](#ji100ksdmooo) **I)**. Hence, it does not reflect the correct situation in 2017, leading to the misconception that 2017 has very few oil spill accidents.
5. The accidents reported in the Oilspill dataset took place in America while Aramco oil fields are mainly in the Middle East. The difference in geographical location may lead to differences in predictor variables, thus affecting the target corrosion rate. However, it still highlights the fact that accidents caused by corrosion are costly to businesses.

### 7. Conclusion

The machine learning techniques that we have applied can greatly enhance problem-solving methodology in the oil industry.The combination of both linear and CART models ensured a robust analysis, leveraging the transparency and simplicity of linear regression alongside the flexibility of decision trees to handle complex, non-linear relationships. By integrating the findings from both models, we can identify the causes and consequences of pipeline corrosion. Our Linear Regression model can also accurately predict corrosion rates based on a combination of different variables, as well as a decision-making framework that can guide the development of preventative measures for Aramco. Consequently, this enhances Aramco’s capacity to devise effective strategies for corrosion management, ultimately leading to more resilient infrastructure systems and potentially substantial cost savings in maintenance and recovery operations.

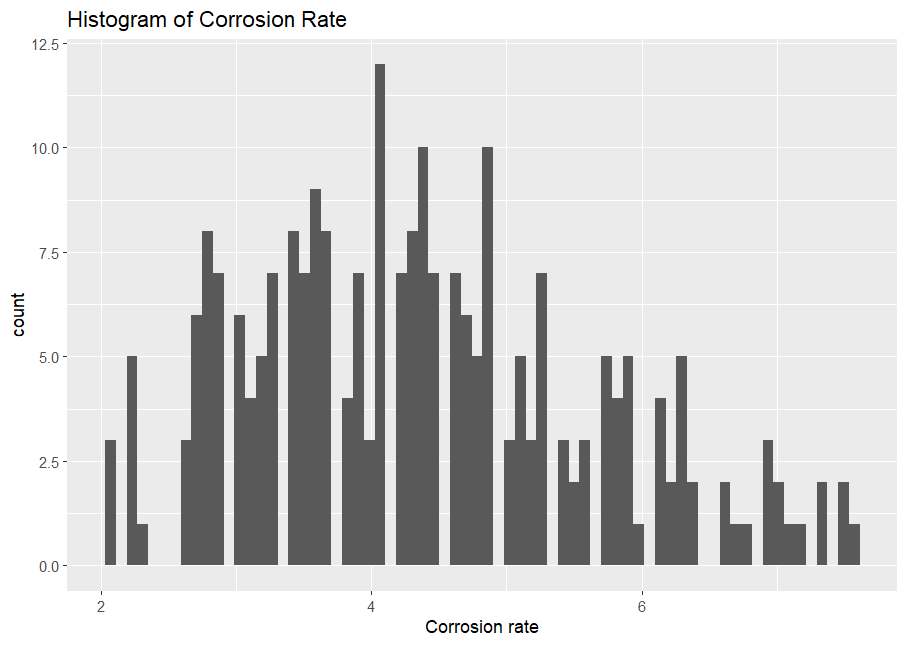
### 8. Appendix

**Appendix A - Overview of data variables**

| **Variable** | **Data type** | **Description** |
| --- | --- | --- |
| Temperature | Continuous (integer) | The temperature measured in Celsius degree |
| Flow velocity | Continuous (numeric) | The flow velocity of fluid within pipelines, measured in m/s |
| CO2 pressure | Continuous (numeric) | Pressure exerted by carbon dioxide (CO2) within a pipeline as part of a gas mixture, measured in bar |
| Internal pressure | Continuous (integer) | Internal pressure of the transmitted fluid, measured in bar |
| Corrosion inhibitor efficiency | Continuous (integer) | Efficiency of commercial corrosion inhibitors against corrosion, measured in % |
| Shear stress | Continuous (integer) | Reflects the influence of the physical properties, geometric characteristics, and velocity of the fluid on the motion characteristics of the fluid, measured in Pa |
| pH | Continuous (numeric) | pH is a scale for hydrogen ion activity in solution, i.e. the standard for measuring the acid-base degree of solution |
| Corrosion rate | Continuous (numeric) | Measured in mm/a/year |

#### 

**Appendix B**



Appendix A. Histogram of target variable: Corrosion\_rate

**Appendix C**

| **Variable** | **Data type** | **Description** |
| --- | --- | --- |
| Accident Year | Continuous (Integer) | Year of accident. (2010 - 2017) |
| Accident Date/Time | Character | Date and time of accident in the format: Month/Day/Year, Hours: Minutes |
| Cause Category | Character | Category of cause of accident |
| Cause Subcategory | Character | Specific cause of accident |
| Net Loss (Barrels) | Continuous (numeric) | The number of oil barrels lost due to oil spill that can not be recovered |
| Pipeline Location | Character | Location of pipeline |
| Pipeline Type | Character | Type of pipeline used |
| Liquid Type | Character | Type of Liquid passing through the pipelines |
| Liquid Ignition | Character | Whether liquid ignited during the oil spill |
| Liquid Explosion | Character | Whether liquid exploded during the oil spill |
| Pipeline Shutdown | Character | Whether the pipeline shut down due to oil spills |
| Property Damage Costs | Continuous (Integer) | Costs incurred from property damage |
| Lost Commodity Costs | Continuous (Integer) | Costs incurred from raw material lost |
| Public/Private Property Damage Costs | Continuous (Integer) | Costs incurred from damage to either public property or private property due to an oil spill incident. |
| Emergency Response Costs | Continuous (Integer) | Costs incurred due to emergency response to the accident |
| Environmental Remediation Costs | Continuous (Integer) | Costs incurred from environmental clean-up efforts to remove contamination in the area |
| Other Costs | Continuous (Integer) | Other costs incurred but not listed in the above categories |
| All Costs | Continuous (Integer) | Sum of all costs from the above categories |

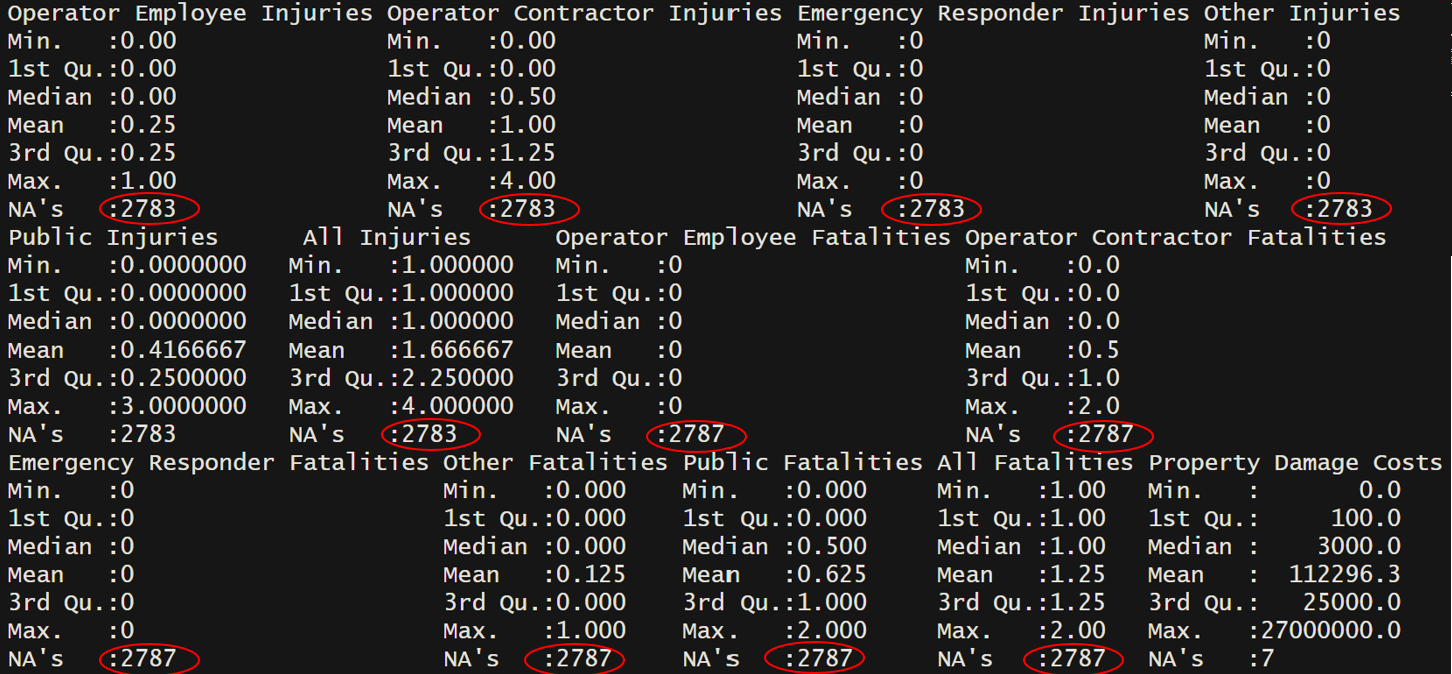
\*Take note: This is the original data type with relevant columns shown

Appendix C1. Original dataset with original data type

| **Variable** | **Data type** | **Description** |
| --- | --- | --- |
| Barrel loss (%) | Continuous (numeric) | % of barrel unintentionally released due to oil spill but can not be recovered |
| Shutdown\_hours | Continuous (numeric) | Time difference between shut down and restarting |

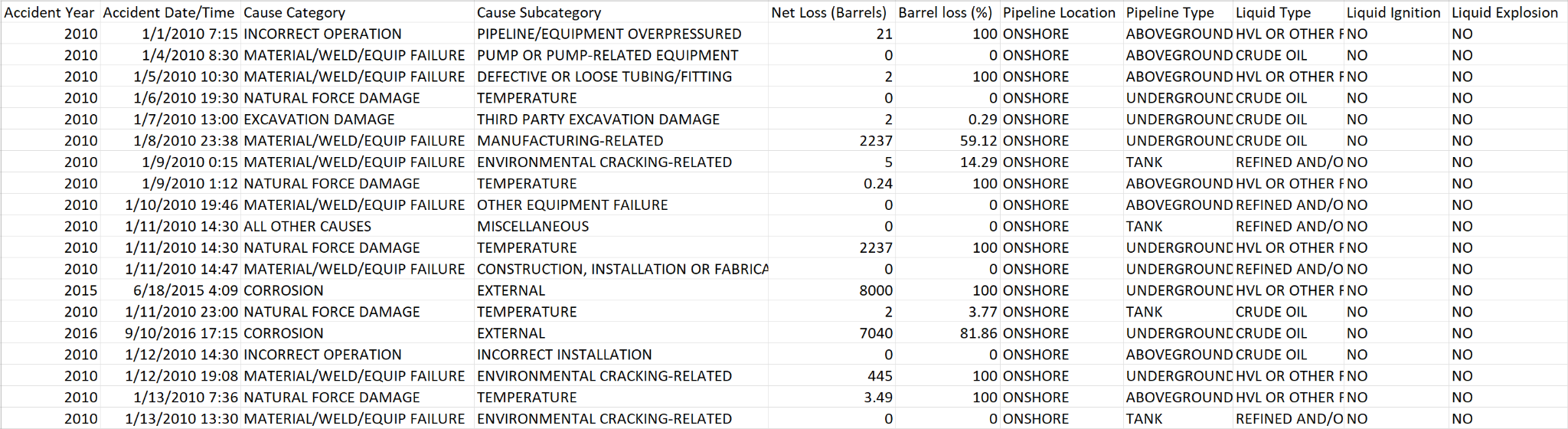
Appendix C2. New variables

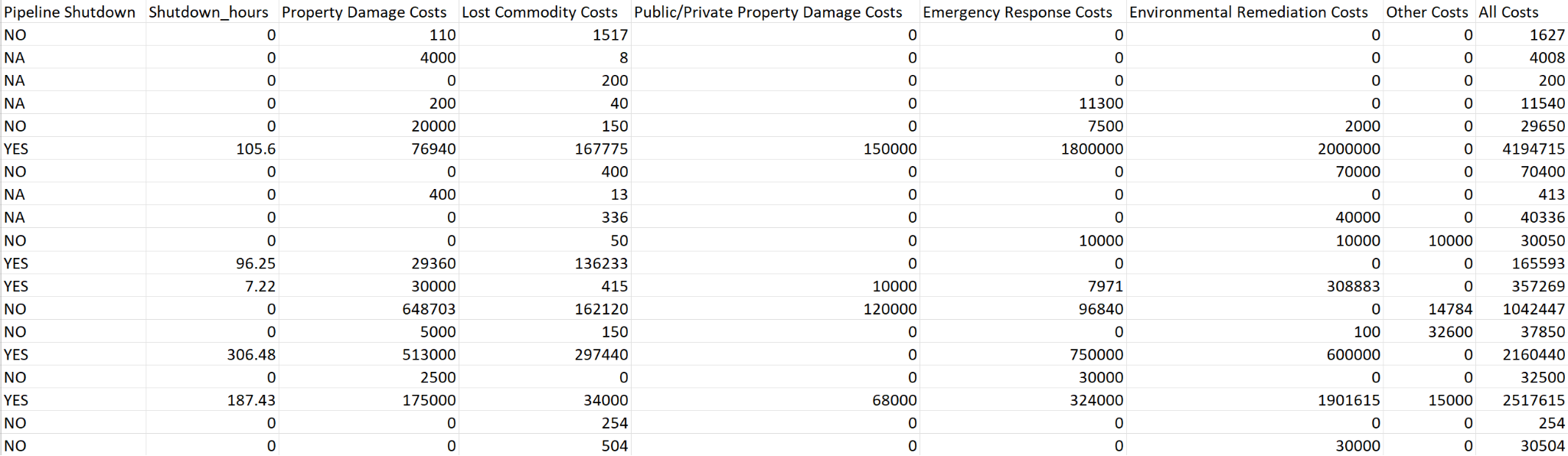
**Appendix D**



Appendix D. Variables with too many NAs

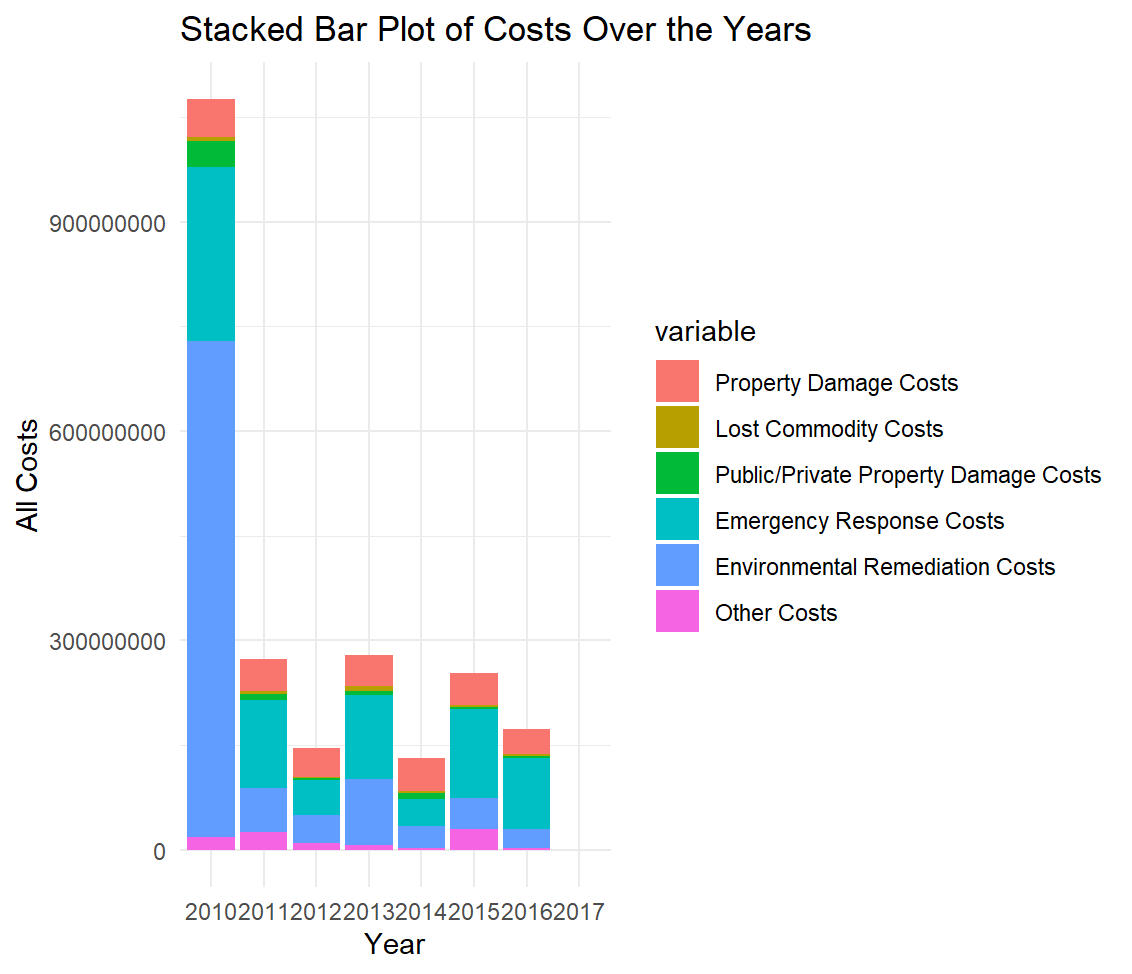
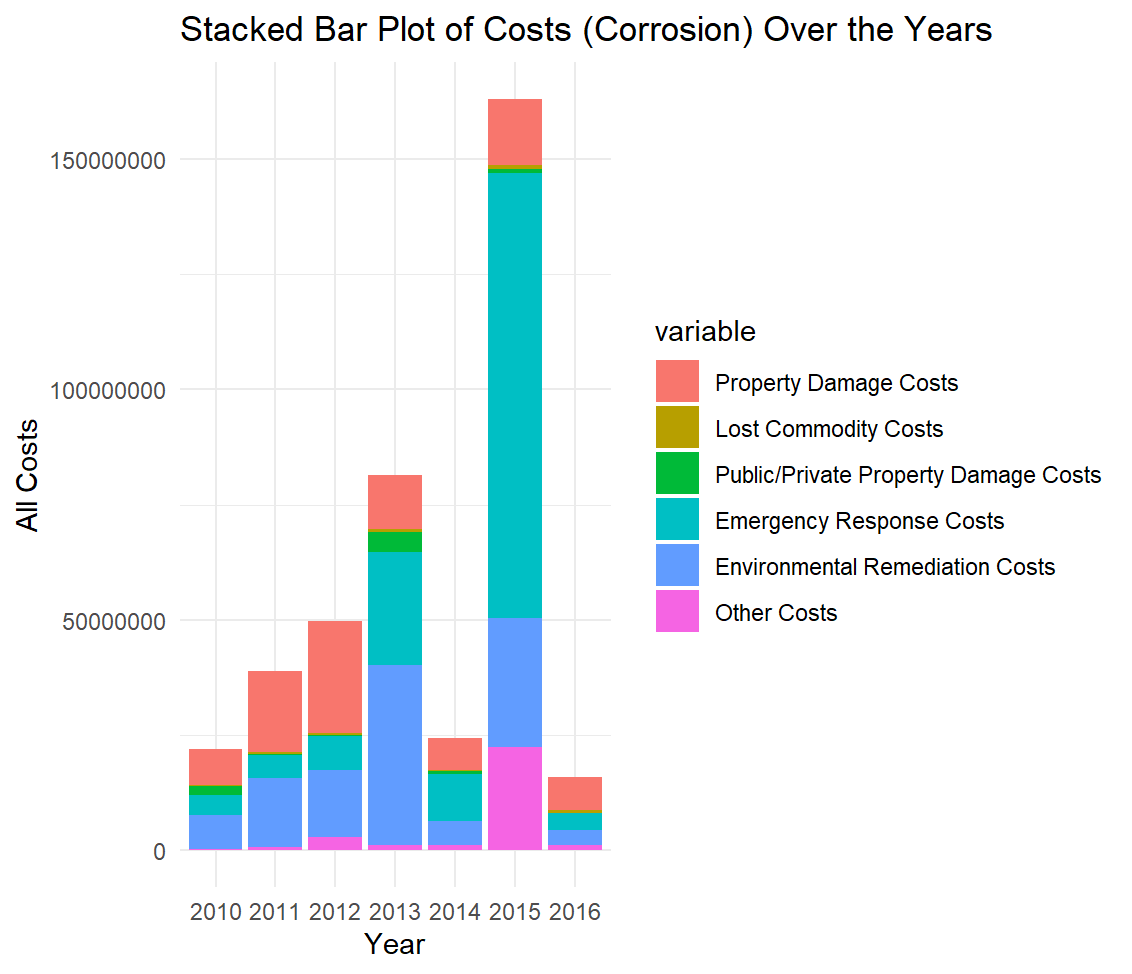
**Appendix E**

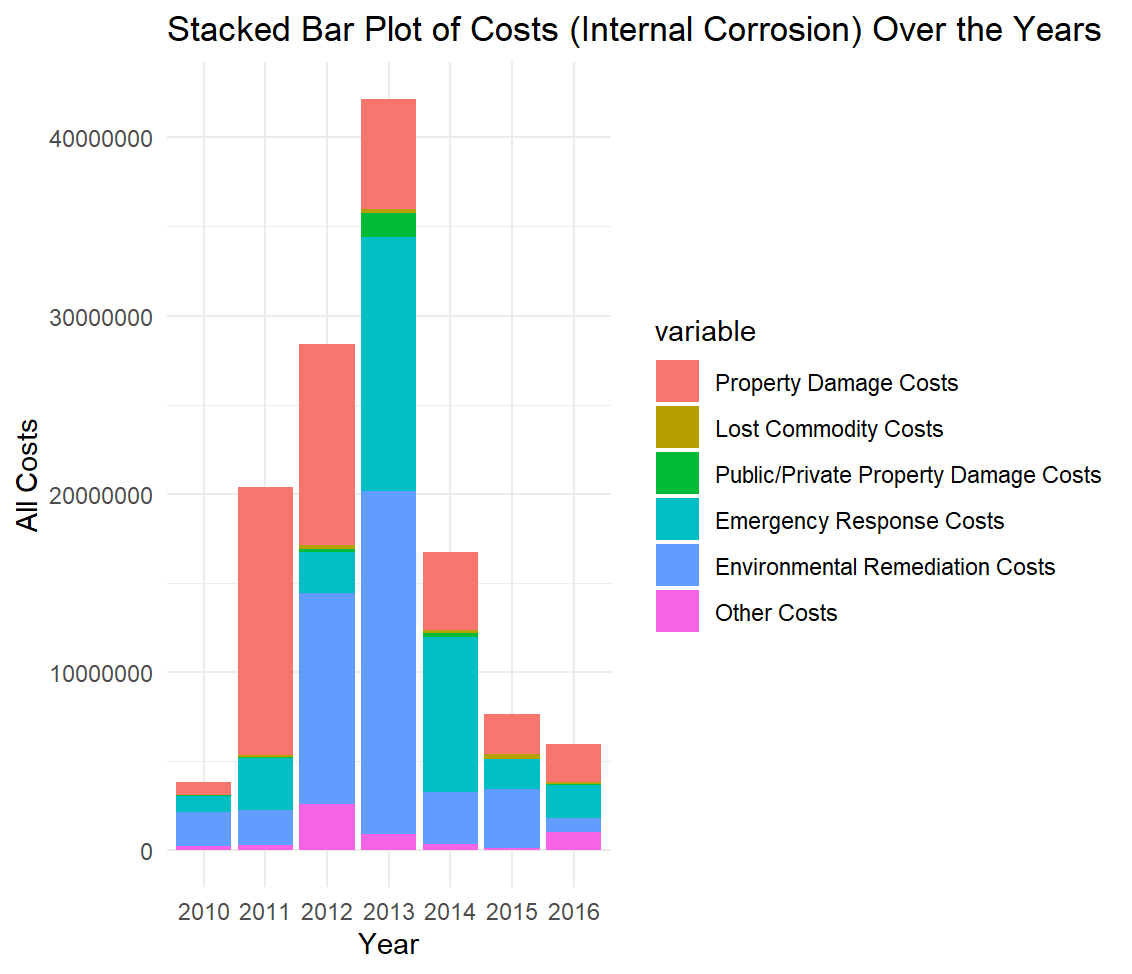




Appendix E. Relevant variables

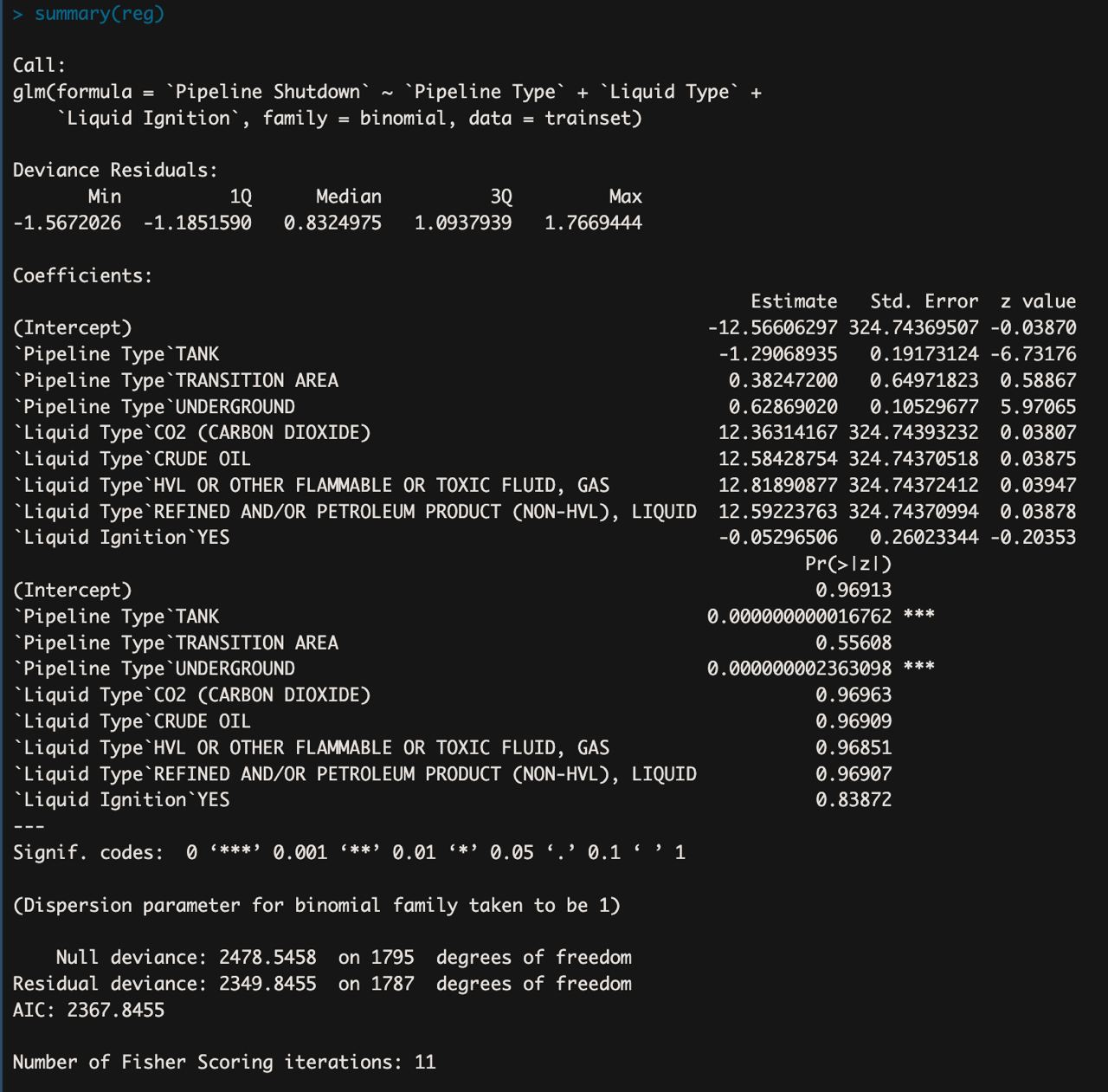
**Appendix F** - **Cost breakdown (corrosion) by year**

****

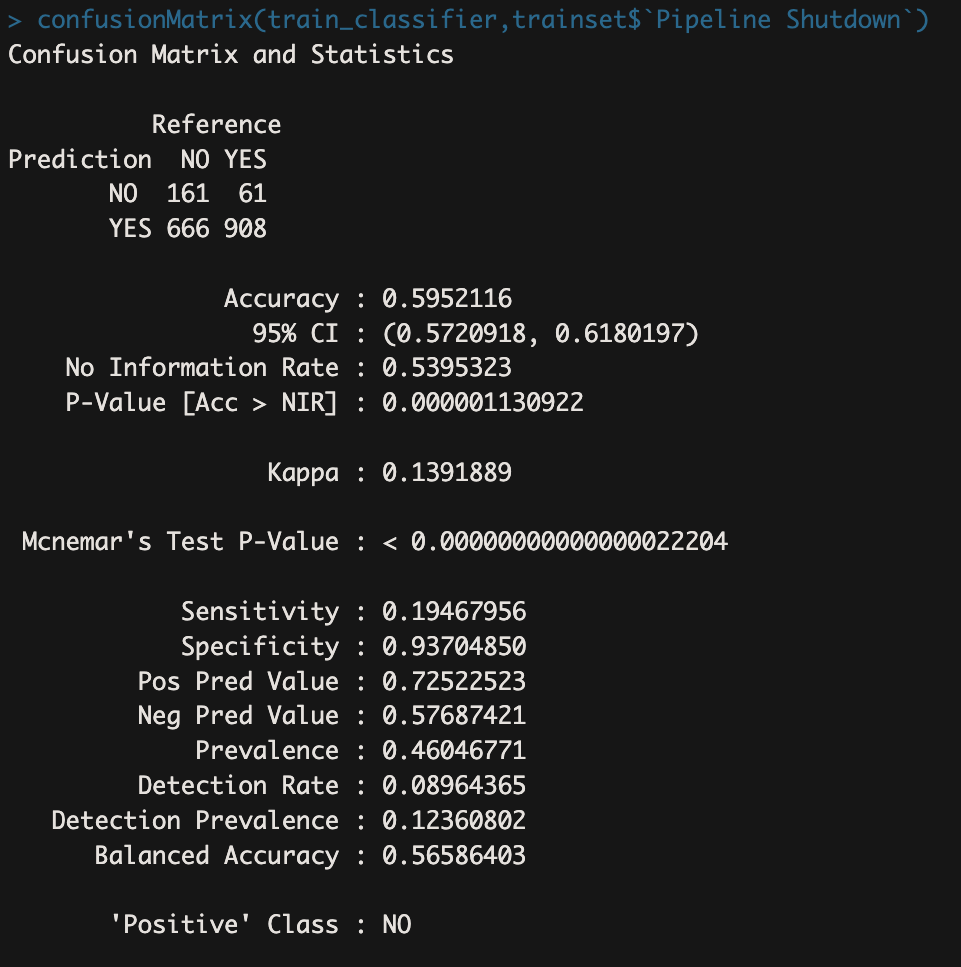
****

Appendix F. Cost breakdown over the years

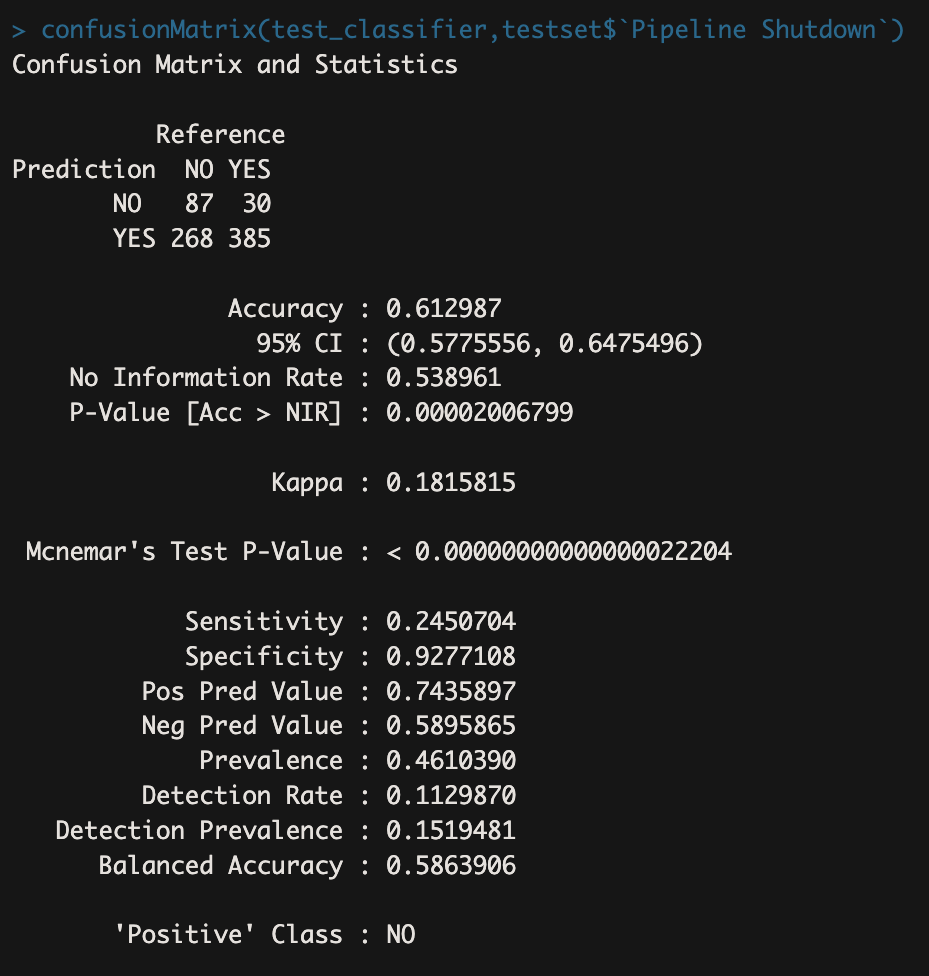
**Appendix G - Results for Pipeline Shutdown**

****

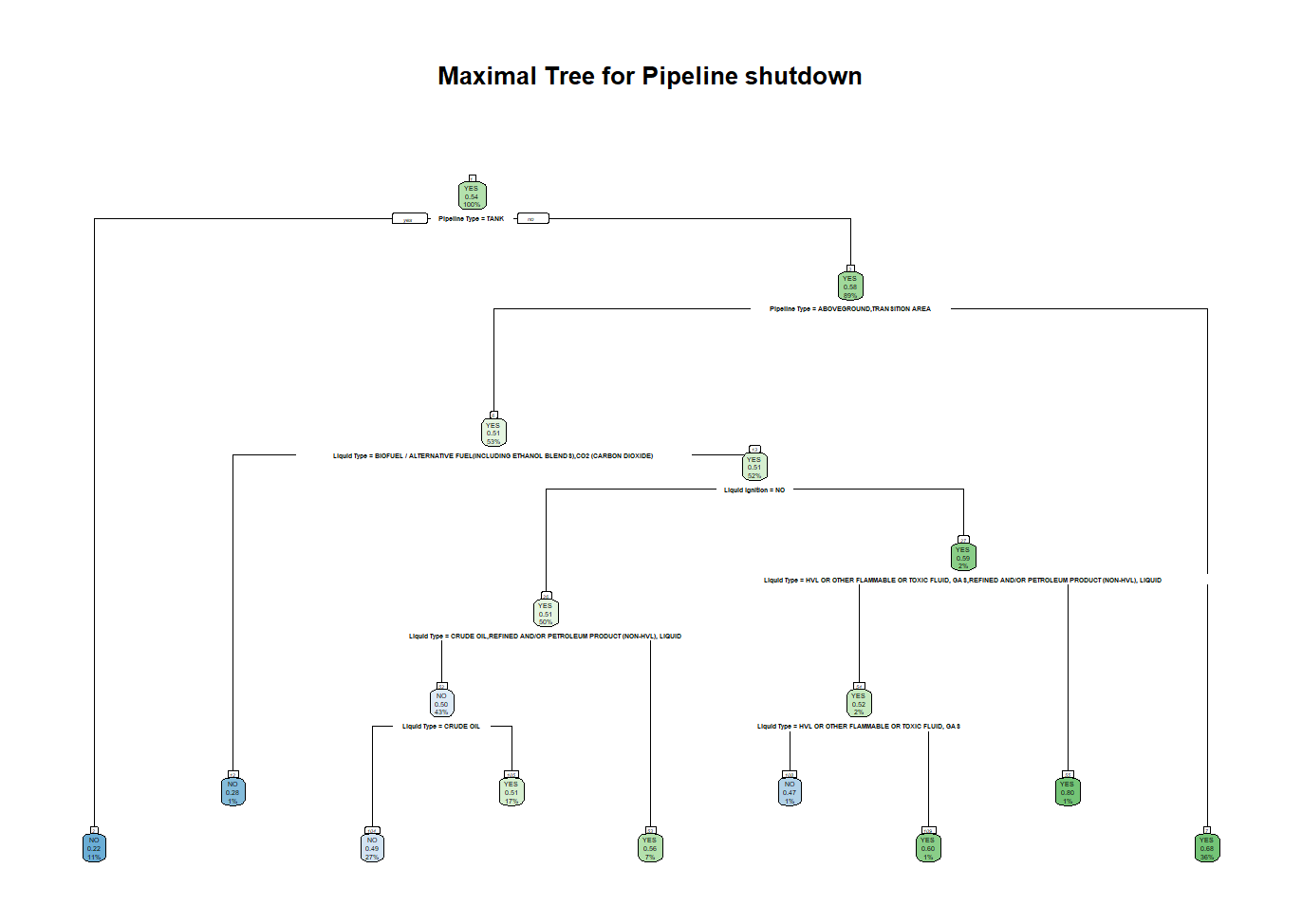
Appendix G1 - Logistic Regression

****

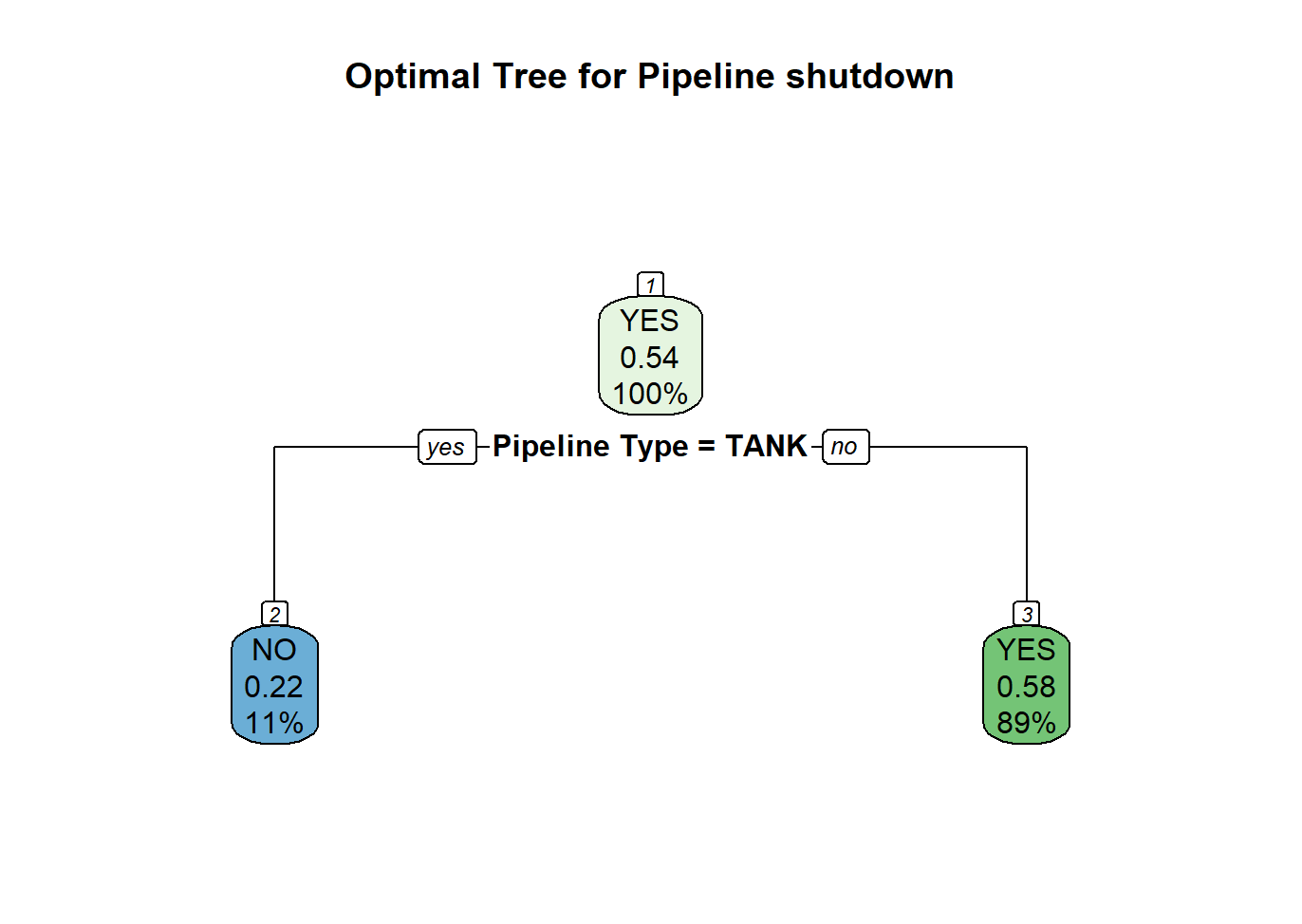
Appendix G2 - Train Set Accuracy (59.52%)

****

Appendix G3 - Test Set Accuracy (61.3%)

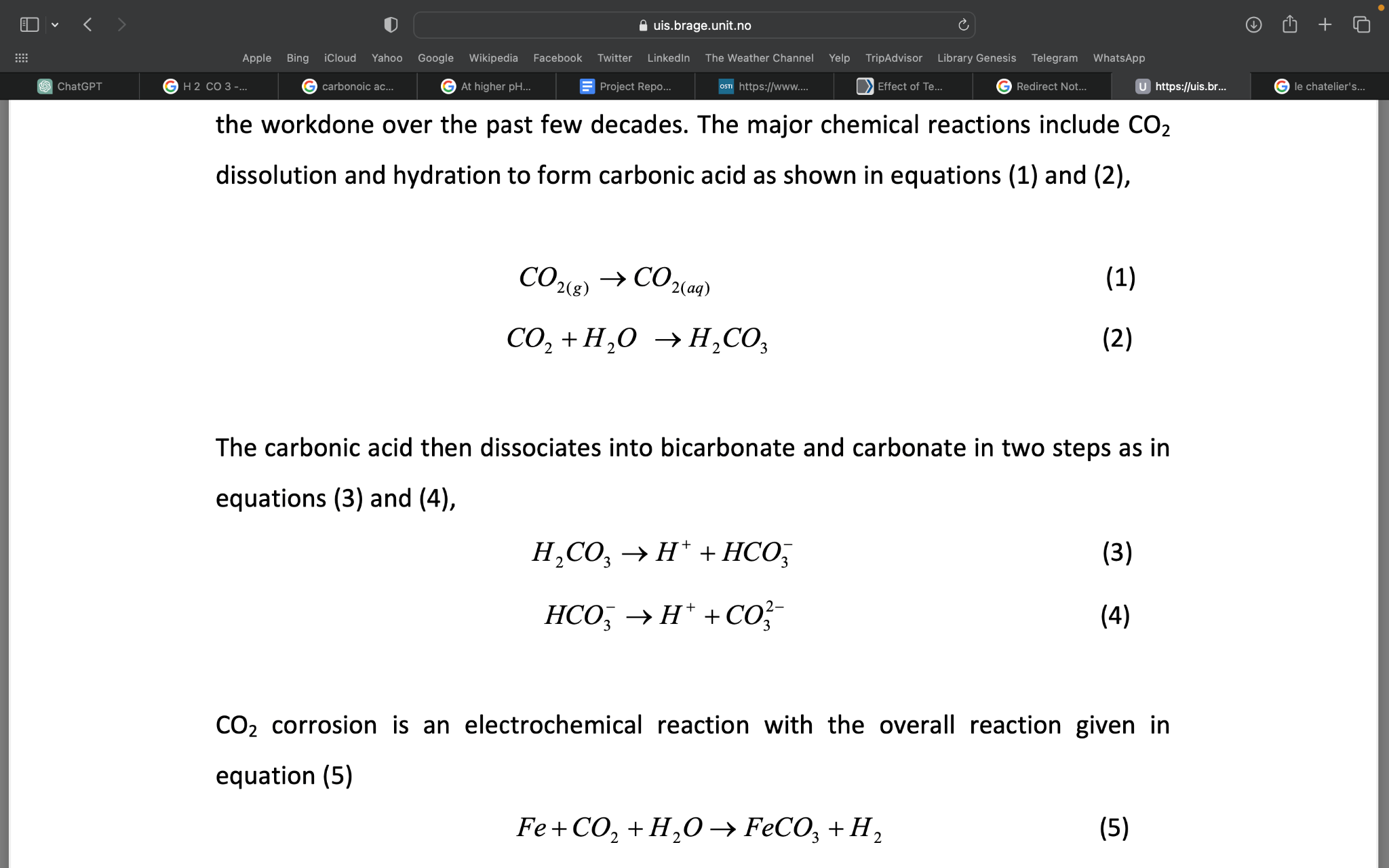
****

Appendix G4 - Maximal tree for Pipeline Shutdown prediction

****

Appendix G5 - Optimal tree for Pipeline shutdown prediction

**Appendix H**

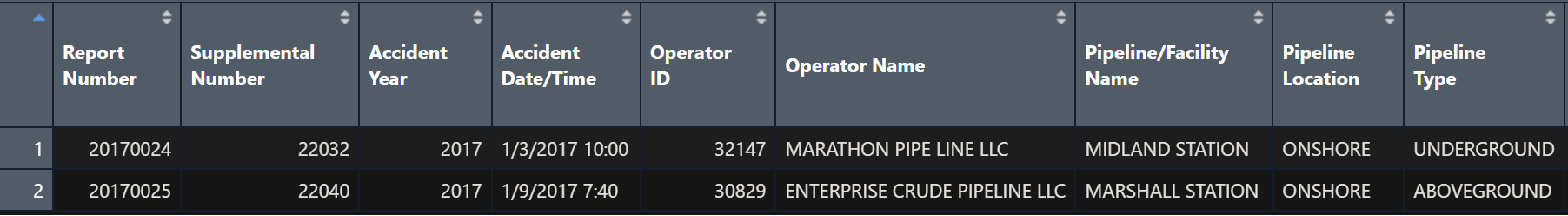


Appendix H. Chemical reactions of CO2

**Appendix I**

### 

Appendix I1.Bar Plot of all cost over the years

****

Appendix I2. Records from 2017

### References

*Aramco Ventures invests in InductoSense’s corrosion monitoring technology - Aramco Ventures*. (2023, May 21). Aramco Ventures. <https://aramcoventures.com/news/aramco-ventures-invests-in-inductosenses-corrosion-monitoring-technology/>

*Corrosion in sprinkler systems | NFPA*. (2020, August 28). <https://www.nfpa.org/News-and-Research/Publications-and-media/Blogs-Landing-Page/NFPA-Today/Blog-Posts/2020/08/28/Corrosion-in-Sprinkler-Systems>

Dutta, B. (2021). A Classification and Regression Tree (CART) Algorithm | Analytics Steps. AnalyticsSteps. <https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm>

Eliyan, F. F., PhD. (2023, August 10). *Leveraging AI for enhanced corrosion control in oil pipelines*. Corrosionpedia. <https://www.corrosionpedia.com/leveraging-ai-for-enhanced-corrosion-control-in-oil-pipelines/2/7456#early-warning-systems-and-anomaly-detection>

IBM. (n.d.). About Linear Regression | IBM. Ibm.com. Retrieved November 1, 2023, from <https://www.ibm.com/topics/linear-regression>

Kahyarian, A., Achour, M., & Nešić, S. (2017). CO 2 corrosion of mild steel. In *Elsevier eBooks* (pp. 149–190). <https://doi.org/10.1016/b978-0-08-101105-8.00007-3>

Li, Q., Zhang, Y., Cheng, Y., Zuo, X., Wang, Y., Yuan, X., & Huang, H. (2022). Effect of temperature on the corrosion behavior and corrosion resistance of Copper–Aluminum laminated composite plate. *Materials*, *15*(4), 1621. <https://doi.org/10.3390/ma15041621>

Mark A. Cohen (2010). A Taxonomy of Oil Spill Costs <https://media.rff.org/archive/files/sharepoint/WorkImages/Download/RFF-BCK-Cohen-DHCosts_update.pdf>

Nyborg, R. (2009). Pipeline corrosion prevention by pH stabilization or corrosion inhibitors. ResearchGate. <https://www.researchgate.net/publication/294566544_Pipeline_corrosion_prevention_by_pH_stabilization_or_corrosion_inhibitors>

Peng, S., Zhang, Z., Liu, E., Liu, W., & Qiao, W. (2021). A new hybrid algorithm model for prediction of internal corrosion rate of multiphase pipeline. Journal of Natural Gas Science and Engineering, 85, 103716–103716. <https://doi.org/10.1016/j.jngse.2020.103716>

Pipeline corrosion engineer. (n.d.). <https://careers.aramco.com/expat_uk/job/Pipeline-Corrosion-Engineer/852717323/>

*U.S. FOB Costs of Saudi Arabia Crude Oil (Dollars per Barrel)*. (n.d.). <https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=ISA0000004&f=A>

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear Regression analysis. *Deutsches Arzteblatt International*. <https://doi.org/10.3238/arztebl.2010.0776>

Simper, S. (2023, April 17). Aramco Ventures invests in corrosion monitoring technology. *World Pipelines*. [https://www.worldpipelines.com/business-news/17042023/aramco-ventures-invests-in-corrosion-monitoring-technology/\](https://www.worldpipelines.com/business-news/17042023/aramco-ventures-invests-in-corrosion-monitoring-technology/%5C)

*Tackling corrosion: A $2.5 trillion problem globally*. (2019, October 14). Saudi Aramco. <https://www.aramco.com/en/news-media/news/2019/tackling-corrosion>

Tackling corrosion: A $2.5 trillion problem globally. (2019, October 14). Saudi Aramco. <https://www.aramco.com/en/news-media/news/2019/tackling-corrosion>

Wu, K., & Mosleh, A. (2019). Effect of temporal variability of operating parameters in corrosion modelling for natural gas pipelines subject to uniform corrosion. *Journal of Natural Gas Science and Engineering*, *69*, 102930. https://doi.org/10.1016/j.jngse.2019.102930