Politecnico di Torino

Politecnico di Torino

Department of Environment, Land and Infrastructure Engineering

Master of Science in Petroleum and Mining Engineering

Data-Analytics Based Investigation of Controlling Parameters of CO_2 Sequestration in Depleted Unconventional Reservoirs

Supervisor:

Prof. Vera Rocca

Candidate:

Hassan Khaled Hassan Baabbad

Co-Supervisor:

Prof. Emre Artun (Istanbul Technical University)

November 2021

Thesis submitted in compliance with the requirements for the Master of Science degree in Petroleum and Mining Engineering

Abstract

Data-analytics has received a great deal of attention in recent years because it supports in improving operations and saving time especially in the oil and gas industry. For the time being, laboratory experiments and numerical reservoir simulators are used to model and discern the behavior of CO_2 sequestration. Nonetheless, these methods have high computational cost. Besides, studies have been done on the use of data-driven statistical techniques, these studies mainly focused on production optimization in unconventional reservoirs. In this study, a data-analytics based investigation was carried out to develop insights and analyze the primary variables that affect CO_2 sequestration process in unconventional reservoirs. The dataset to be utilized consists of a large number of numerical-simulation scenarios that were conducted as part of another study (Kulga, 2014). Basically, two techniques were used: an exploratory data analysis and predictive modeling. Exploratory data analysis revealed a relationship between reservoir, operational parameters, and the cumulative CO_2 injected. A considerable number of operational parameters displayed a monotonic relationship with the cumulative CO_2 injected. Stimulated reservoir volume fracture permeability was the variable which displayed the best correlation. In addition, statistical and machine-learning based predictive models were developed to predict the volume of CO_2 sequestered. Comparison of these predictive models indicated that random forest was the preferred method due to having the lowest prediction error. Lastly, variable importance was implemented to determine the most influential parameters of the CO_2 sequestration process in unconventional shale-gas reservoirs. Interestingly, the most influential parameters are the ones affecting the stimulated reservoir volume. According to our results, operational parameters are more dominant than reservoir parameters in driving high-performance and stimulated reservoir volume fracture permeability is the most important parameter in order to get high-performance. Our findings will aid in designing these sequestration projects sustainably.

Keywords: CO_2 sequestration, Unconventional reservoirs, Data-analytics, Exploratory data analysis, Predictive modeling

Dedicated to My beloved Parents

Acknowledgements

First and foremost, I want to convey my heartfelt gratitude and admiration to my co-supervisor, Prof. Emre Artun, for his countless support, steering insights continuously during our meetings and throughout this thesis. Along with Prof. Burak Kulga from Istanbul Technical University, who provided us with the dataset that was gathered at Penn State University. I am very grateful for their support and warm welcome during my time in Istanbul, albeit it was in remote mode and could not have in person discussions because of COVID-19.

I would also like to thank my internal supervisor Prof. Vera Rocca for her help and supervision during this study. Together with all my respected professors in the Department of Environment, Land, and Infrastructure Engineering, their experience played a vital role in my professional development.

I would like to give thanks to Politecnico di Torino outgoing mobility office for the scholarship they provided to perform my research abroad. Finally, a massive thank you, especially, to my family and friends who believed in me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AIC	Akaike Information Criterion
AAE	Average Absolute Error
BIC	Bayesian Information Criterion
BSCF	Billions of Standard Cubic Feet
CART	Classification and Regression Trees
CC	Correlation Coefficient
DA	Data Analytics
EDA	Exploratory Data Analysis
EOR	Enhanced Oil Recovery
E&P	Exploration and Production Companies
GBM	Gradient Boosting Machine
GDP	Gross Domestic Product
GHG	Greenhouse Gas
IOT	Internet of Things
IQR	Interquartile Range
MSE	Mean Squared Error
ML	Machine-Learning
OLS	Ordinary-Least-Squares
RCC	Rank Correlation Coefficient
RF	Random Forest
RSE	Residual Standard Error
RSS	Residual Sum of Squares
SCF	Standard Cubic Foot
TSS	Total Sum of Squares

List of Symbols

σ_{xy}	Covariance
σ_x	Standard deviation of variable x
σ_y	Standard deviation of variable y
$\bar{X}$	Mean of variable x
$\bar{Y}$	Mean of variable y
$N-1$	Degrees of freedom
x_i	Individual outcome for x
y_i	Individual outcome for y
$E[X]$	Expected value
$V[X]$	Variance
σ_x^2	Variance

Chapter 1 Introduction

1.1 Overview

The global population in the next few decades is predicted to rise by about 1.5 billion people and reach about 9.2 billion people by 2040 (United Nations Population Division, 2019). Besides, the gross domestic product (GDP) is proposed to further increase within the same time frame. This expected rise in global welfare will lift billions of individuals out of poverty and into the middle class. Many forecasts anticipate a 25% to 30% increment in global energy demand by 2040 to achieve this tremendous growth in prosperity (*BP Energy Outlook*, 2019). Along with providing reasonable, reliable energy to aid growing economies and individuals, the world must likewise focus on climate change risks and rising greenhouse gas (GHG) emissions (Armstrong et al., 2019).

Due to its radiation absorption capacity in the atmosphere, Carbon-Dioxide (CO_2) has been recognized as the most critical GHG that is targeted for emission-reduction activities. To reduce its impact on the climate, its sequestration and storage have been considered as a challenging engineering problem and named as one of the Grand Engineering Challenges of the 21st Century by the U.S. National Academy of Engineering (*NAE Grand Challenges For Engineering* ^{TM}, 2017). Sequestration into geological formations has been offered as a viable part of the solution over the years. More recently with the exploration and exploitation of unconventional resources, these resources have been identified as ideal candidates for this process due to their deep nature, large areal extent and volume, existing infrastructure for injection (horizontal wells and hydraulic fractures) and potentially induced fracture network due to hydraulic fracturing. While considered as a potential solution, the uncertainties related to its long-term operational, financial and sustainability aspects are still being investigated through modeling studies. That is why this research addresses these uncertainties and problems by developing insights regarding the operational aspects of CO_2 sequestration in unconventional namely shale reservoirs through a data-analytics based investigation.

1.2 Background of study

In recent years, the terms “big data” and “data analytics” have turned into somewhat of a buzzword, owing to several alleged uses in fields such as health and life sciences, consumer marketing and national security. As a result, many people believe that big data analytics has the potential to revolutionize oil and gas operations (Holdaway, 2014). The oil and gas sector

is looking at the possibility of mining enormous amounts of data on the subsurface, physical infrastructure, and flows to get new insights into the reservoir and improve operational efficiency (Mishra & Datta-Gupta, 2018).

The term "big data" is used to describe enormous, multivariate datasets that are described by the 3 V's: volume, variety, and velocity (Figure 1.1) (Mishra & Datta-Gupta, 2018). The term volume refers to the amount of data, that we are dealing with approximately 10^2 – 10^4 independent variables and roughly 10^3 – 10^6 observations (Mishra & Datta-Gupta, 2018). Data is now available in a variety of formats. Structured alphanumeric data is stored in traditional databases. As the digital oilfield grows its impact in the business, unstructured text documents as daily drilling reports, video, audio, e-mail, and financial transactions multiply. Governing and managing many types of these data is a demanding task the majority of Exploration & Production (E&P) companies still cope with as upstream siloed data explodes with developing digital oilfield and intelligent wells initiatives (Holdaway, 2014).

Velocity refers to the increasing prevalence of real-time streaming data from surface gauges or downhole sensors, which increases the quantity of the dataset and causes extra considerations such as re-sampling, data archival, and redundancy analysis (Mishra & Datta-Gupta, 2018). While the term "data analytics" as shown in Figure 1.1 relates to analyzing data, recognizing what the data implies and obtaining insight from the data, and developing predictions that lead to better judgments based on these data-driven insights (Hastie et al., 2008) as cited in (Mishra & Datta-Gupta, 2018).

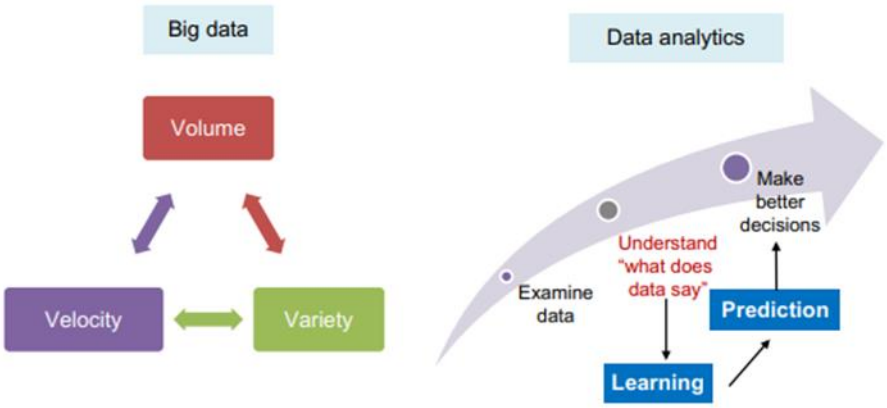

Figure 1.1 Big data and data analytics (Mishra & Datta-Gupta, 2018)

Recently the global COVID-19 pandemic has had a notable impact in the oil and gas industry. The immediate requirement for efficient and safe operations and cost cutting, as well as the long-term focus on energy transition with emerging technologies into a digital ecosystem, are inevitably linked. While many organizations began their digital transformations earlier, the interruption of 2020 has elevated digital transformation from a priority to a must. Our industry has a big chance to reinvent itself in the digital domain by focusing on integrated business transformation and discovering more dynamic working method (Feder et al., 2021) eventually the key to this transformation will be data analytics.

1.3 Structure of the thesis

Chapter 1: in this chapter the overall context of the research has been introduced including the background of the study, which will serve as an introduction to the thesis.

Chapter 2: in this chapter the literature will be examined in order to identify crucial findings regarding the application of data-analytics and building predictive models within the context of unconventional shale reservoirs and oil & gas industry in general.

Chapter 3: in this chapter the problem statement will be addressed in a clear and precise way along with the research aims, objectives and questions that will need answering by the end of the study.

Chapter 4: in this chapter the methodology will be reviewed by assessing how the research was conducted using the two main techniques which are exploratory data analysis and predictive modeling using statistical & machine learning.

Chapter 5: in this chapter, the results and discussion will be presented by reporting the main findings concisely and objectively evaluating these findings logically.

Chapter 6: the final chapter will involve the significant conclusions regarding the application of data-analytics in CO_2 sequestration process as explained in the previous chapters and recommendations for future studies will also be mentioned.

Chapter 2 Literature Review

The current development in automation in industrial processes as part of the 4th Industrial Revolution (4IR) is a key ongoing argument. Data analytics is the pivotal component of 4IR (Narayanan et al., 2020). To build a familiar literature, it is a good idea to start with basic definitions.

- **Data analytics (DA)** is the study and modeling of hidden patterns and correlations in complex, multidimensional data sets employing extensive data collection and analysis (Mishra et al., 2021).
- **Machine learning (ML)** is the process by which a model is constructed between predictors and response by employing an algorithm (commonly referred to as a black box) to deduce the underlying input/output relation from data (Mishra et al., 2021).
- **Artificial intelligence (AI)** is the process of using a predictive model to make judgments with no human interaction (and with the possibility of evaluation for model updating) (Mishra et al., 2021).

One of the most significant applications of DA in CO_2 sequestration processes for unconventional reservoirs is to optimize the performance of CO_2 sequestration by developing data-driven insights and this reduces the computational cost, since reservoir modeling and simulation in these reservoirs can be costly and infeasible. As a result, the goal of this literature review is to analyze why numerical modeling in these reservoirs can be impractical to an extent and the application of data analytics/mining for characterizing the controlling factors of the CO_2 sequestration process in shale-gas reservoir.

Firstly, reservoir modeling and simulation will be discussed followed by exploratory data analysis, then predictive input and output modeling will be investigated, and finally model evaluation and variable importance will be assessed.

2.1 Reservoir modeling and simulation

Reservoir simulation is a technology that integrates different principles such as physics, mathematics, reservoir engineering and computer programming to estimate reservoir performance under a variety of operating situations (Ertekin et al., 2001). Furthermore, in the reservoir simulation technique, a system of algebraic mathematical equations constructed from a set of PDE's (Partial Differential Equations) with proper initial and boundary conditions approximates reservoir behavior (Ertekin et al., 2001). These mathematical equations include

the most significant physical processes occurring in the reservoir system, such as fluid flow divided into three phases (oil, water, and gas), as well as mass transfer between different phases. Also, using an extended version of Darcy's law, the effects of viscous, capillary and gravity forces on fluid flow are taken into account (Ertekin et al., 2001). Likewise, if we consider CO_2 sequestration, we have to consider a model that takes into consideration the chemical composition together with the different behavior of pressure and temperature.

Application of reservoir simulation in CO_2 sequestration studies

The major tools used to execute the primary studies related to uncertainty analysis of CO_2 sequestration are reservoir simulation models. They allow for the prediction of the injection and storage process performance under various geological conditions and injection scenarios (Mohaghegh, 2018). These commercial reservoir simulators are also efficient in capturing the fluid flow behavior and manage natural gas production from unconventional resources such as shale (Boosari et al., 2015).

Several researchers have applied numerical reservoir simulation for modeling CO_2 sequestration. In their research, Yang et al. (2005) have applied reservoir simulation to model the properties of CO_2 injection in the Barrow Sub-basin field in West Australia. They observed that the direction of CO_2 migration and geological structure were two essential considerations in the selection of the optimum well pattern. This met the injection criterion, as well as demonstrating that CO_2 geological sequestration in the Barrow Sub-basin is desirable. Additionally, Ghoodjani & Bolouri (2012) built an analytical model and compared with a numerical method to estimate project performance and calculate the best rate of injection in various scenarios of CO_2-EOR and sequestration projects. They discovered that using a numerical simulator to optimize injection rate is a reliable method.

However, the longer the run time, the more sophisticated the simulation model is. Because of the huge requirements of run time and computational effort, any study involving thousands of simulation runs, such as uncertainty analysis, optimization study, or history matching, might become excessively long and impractical. These long execution durations of numerical reservoir simulation models have long been a challenge in the oil and gas sector (Mohaghegh, 2018). Moreover, for unconventional reservoirs given the requirement to simulate fluid flow in a network of induced natural fractures coupled with geomechanical effects and other phenomena, such as water blockage, non-Darcy flow in nanoscale pores, and adsorption/desorption, reservoir modeling in such systems is a tough project (Cipolla et al.,

2010); (Ding et al., 2014) as cited in (Schuetter et al., 2018). Hence, huge computational cost is the major challenge with the routine application of comprehensive physics-based simulators (Schuetter et al., 2018).

For this reason, data-driven techniques to model and understand the key parameters of CO_2 sequestration in unconventional reservoirs need to be developed and used in order to complement the numerical reservoir simulators.

2.2 Exploratory data analysis

The major purpose of Exploratory Data Analysis (EDA) is to gain a preliminary knowledge of the data in terms of individual variable qualities and the relationships between them. Other goals include identifying key variables of interest, creating questions for further investigation of the data, and selecting tools for comprehensive research (Mishra & Datta-Gupta, 2018). Multiple studies have utilized EDA for production optimization in unconventional reservoirs. In their research Schuetter et al. (2018) applied EDA by employing a matrix of scatterplots as shown in Figure 2.1 to demonstrate the relationship between all possible predictor variables and response variables, together with a histogram for all the parameters along the diagonal. This graph likewise shows substantial relationships between predictor pairs.

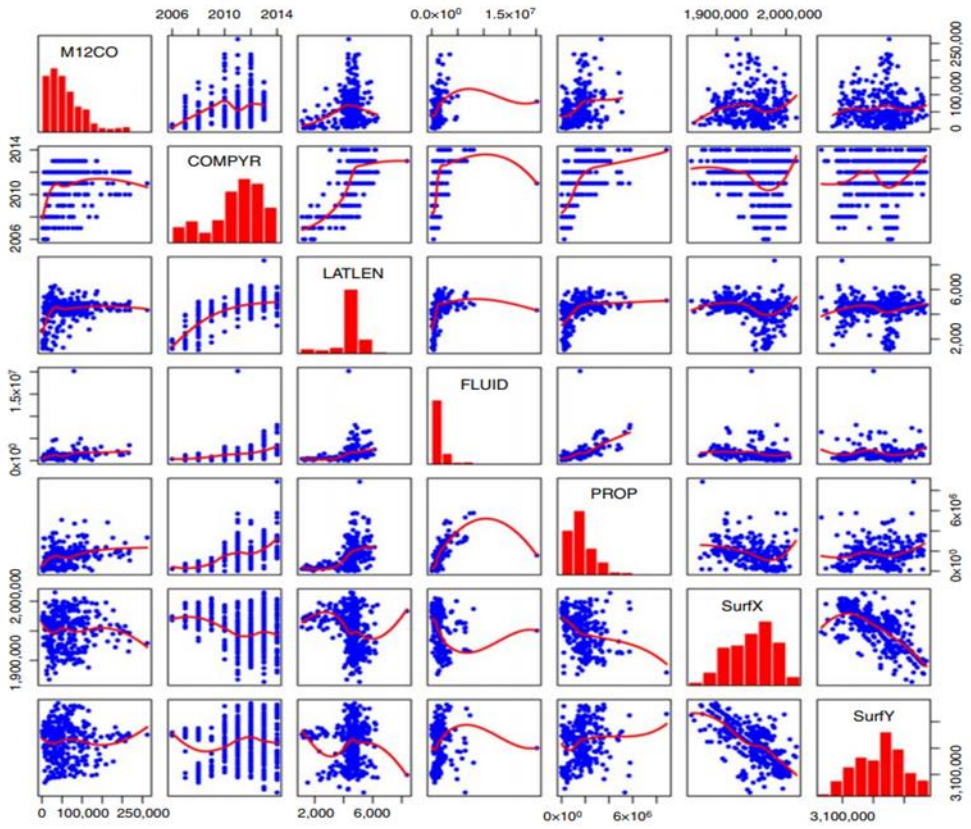

Figure 2.1 Scatterplot matrix (Schuetter et al., 2018)

Likewise, (Zhong et al., 2015) performed EDA by first applying univariate analysis such as histogram to be able to visualize continuous variables and examine each variable's features and distributions see Figure 2.2, Along with a scatterplot matrix see Figure 2.3 to be able to discover pairwise trends between variables, as well as peculiar data points such leverage points and outliers.

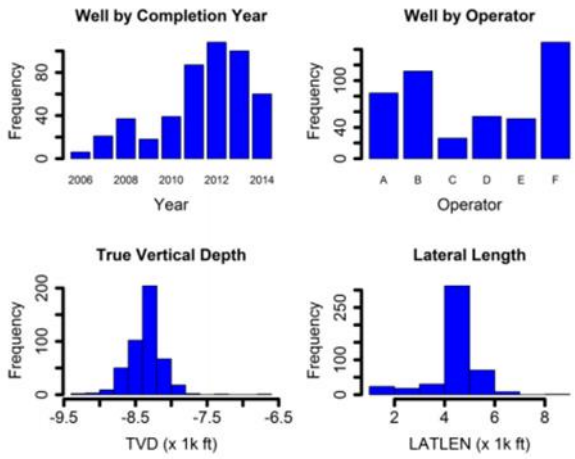

Figure 2.2 Histograms for predictor variables (Zhong et al., 2015)

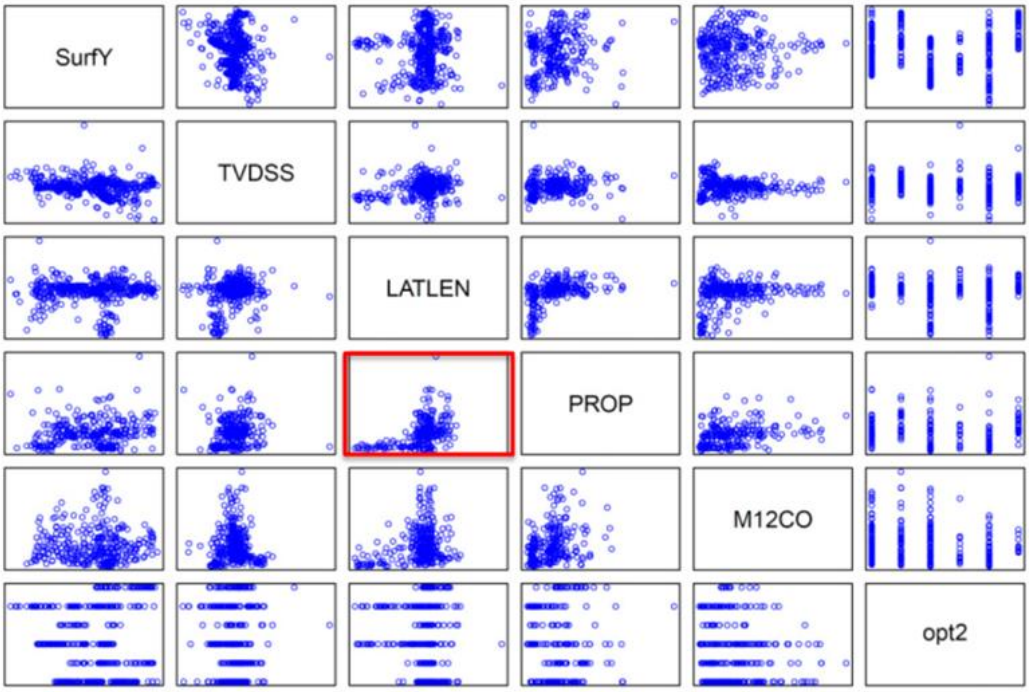

Figure 2.3 Scatterplot matrix for predictor variables (Zhong et al., 2015)

However, the EDA methodologies adopted by (Schuetter et al., 2018) and (Zhong et al., 2015) are an example of the techniques that can be used, and they are graphical in most cases. Several techniques including, examining variable distributions (for example, to discover severely skewed or non-normal patterns, such as bi-modal patterns), and assessing enormous correlation matrices for coefficients that match thresholds are examples of basic exploratory procedures. EDA for multivariate data sets comprises multivariate exploratory techniques created specifically to detect patterns in multivariate data sets (Holdaway, 2009). Most of these techniques will be discussed extensively in the methodology chapter.

2.3 Predictive input and output modeling

Kuhn & Johnson (2013) define predictive modeling as the development of a model or mathematical tool that achieves an accurate prediction. The model could be as an equation or algorithm, with one variable to predict (output) and one or more independent known predictors (inputs) (Lolon et al., 2016).

Although predictive models have been widely employed, Kuhn & Johnson (2013) pointed out that there are a few conventional reasons predictive models fail and might generate unreliable predictions and we'll go over each one in this section. The most common issues are:

- Insufficient data pre-processing
- Model validation is minimal
- Extrapolation that isn't justified
- Overfitting the model to the data that already exists

Schuetter et al. (2018) were able to point out that building predictive input/output models is a typical goal in oil and gas applications. Various empirical studies have applied predictive modeling for production optimization in unconventional reservoirs (Zhong et al., 2015); (Lolon et al., 2016); (Schuetter et al., 2018). These studies are relatively recent, with the majority occurring within the last 10 years.

While a detailed investigation by Schuetter et al. (2018) confirmed the existence of their relative strengths and weaknesses of these predictive modeling methods, which can be seen in (Schuetter et al., 2018).

In production optimization, some of the most used predictive modeling techniques for regression and classification problems will be explained in this section briefly, while a more detailed analysis will be provided in the methodology chapter.

Ordinary-Least-Squares (OLS) Regression. The response is described as a linear combination of the predictors or functions of the predictors, often known as multiple linear regression (Schuetter et al., 2018).

Classification and Regression Trees (CART). The predictor space is divided into nested rectangular sections, each with a constant value or categorical label for the response in binary decision trees (Breiman et al., 1984) as cited in (Mishra & Lin, 2017).

Random Forest Regression (RF). In this model each of the simple regression trees in the ensemble is trained with a different set of observations and predictors (Breiman, 2001) as cited in (Mishra & Lin, 2017).

Gradient Boosting Machine (GBM). In this method, each new tree aims to fix the shortcomings in predictions made by earlier trees, which are trained steadily as an ensemble of regression trees (Friedman, 2001) as cited in (Mishra & Lin, 2017).

Support Vector Machine (SVM). Transforms the data into a space where it may be modeled using a linear regression or linear classification approach (Vapnik, 1995) as cited in (Mishra & Lin, 2017).

The predictive models discussed above can have contrasting predictive ability according to different datasets. For example, in their research Zhong et al. (2015) found that tree-based approaches, RF, and GBM required less pre-processing time on the raw data, according to practice on the Wolfcamp dataset. They were also less prone to data quality difficulties and made better predictions than others.

While a study by Lolon et al. (2016) identified that although the GBM model has the lowest error when using the training set, it has the poorest prediction ability when using this specific dataset in this investigation. This is because the GBM is over-fitting and hence not being suitable as a prediction tool in this circumstance.

Thus, predictive models do not always have the same predictive ability, they can behave differently according to different datasets. However, in our literature there has been very little to almost no research done on the application of predictive models on the performance of CO_2 sequestration process. Moreover, most of the literature focuses on the application of predictive models for production optimization. That is why in this study the application of predictive models on the performance of CO_2 sequestration process will be investigated.

2.4 Model evaluation and variable importance

The evaluation of the goodness of fit (quality of fit) is an important part of model selection that is often neglected. Creating a scatterplot comparing actual response values in the training set against the predicted response using the model is a standard way to evaluate model fit (Schuetter et al., 2018). If all the scatterplot's points are near the 45-degree line, the model is well-fit to the training data see Figure 2.4. This does not, however, guarantee that the model will work for future data sets. As seen in Figure 2.4 the red curve is placing extreme insistence on replicating the training set. Nonetheless, this introduces over-fitting to the training data set, which leads to poor model predictions in the future (Schuetter et al., 2018).

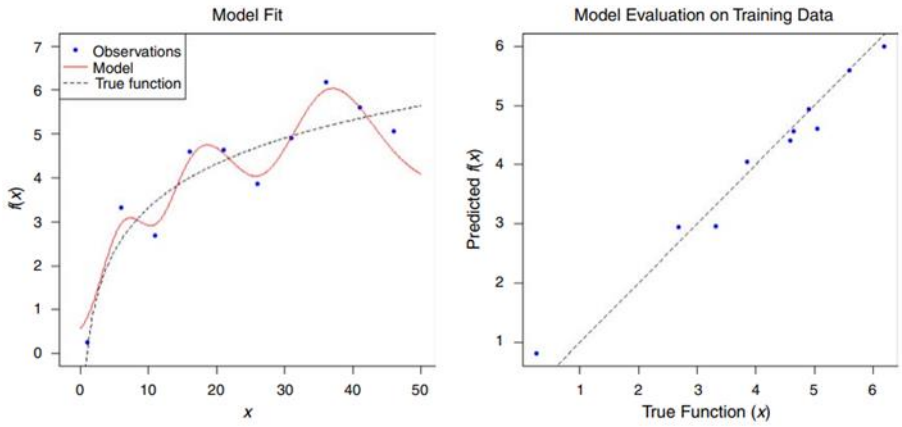

Figure 2.4 Model fit and evaluation (Schuetter et al., 2018)

Additionally, many other approaches exist in evaluating or quantifying the goodness of fit. A few studies have applied some common metrics to compare performance of different methods (Zhong et al., 2015); (Mishra & Lin, 2017); (Schuetter et al., 2018). These common metrics are:

- Average absolute error (AAE)
- Mean squared error (MSE)
- Pseudo-R^2

These three metrics are quite similar since they try to represent how closely the predictions are to the assessment of the data overall (Mishra & Datta-Gupta, 2018). In the methodology chapter, these metrics will be discussed in more details in order to see how they can quantify the goodness of fit. After model evaluation and selection, the last part is to evaluate which parameters are influencing the model and response variable, this can be achieved through variable importance.

For the most part, model-specific variable importance identification is common, and associated metrics can be expressed in absolute or relative units (Mishra & Datta-Gupta, 2018). For example, multiple studies have applied the relative importance measured by RF model (Zhong et al., 2015); (Mishra & Lin, 2017); (Schuetter et al., 2018). In this technique the model calculates the increase in Root Mean Square Error (RMSE) when a variable is permuted while the others are left unaffected to determine the strength of each variable's prediction (Breiman, 2001) as cited in (Mishra & Datta-Gupta, 2018). The reasoning behind the permutation phase is that, if the predictor variable isn't crucial to the tree-building process, rearranging its values

won't make a significant difference in prediction accuracy (Mishra & Datta-Gupta, 2018). On the contrary, other methods that can be used include relative importance for GBM and R^2- loss.

2.5 Summary

Therefore, this literature review aimed to examine the reason in which numerical reservoir simulation can be infeasible to provide insights and assessed the applications of DA in unconventional reservoirs. In summary, there is consistent evidence throughout the literature that data-driven techniques are becoming more significant in production optimization. Nonetheless, in an era in which many E&P companies are trying to transition to become net zero by 2050, application of data-driven methods in production optimization is no longer adequate. For this reason, understanding and optimizing the performance of CO_2 sequestration will be critical. Considering there is a need to mitigate climate change and for this to be possible we need the storage of CO_2 underground that is at least equivalent to the mass of oil & gas emissions we produce at this moment in time. Hence, data-driven methods for CO_2 sequestration processes should be of paramount importance if we are to transition to net zero by 2050.

Chapter 3 Problem Statement

Oil and Gas companies are aware of the challenge on climate change. One of the many things required to deal with this challenge is to store CO_2 underground. For a while now, many organizations have used laboratory experiments and numerical reservoir simulators to model and understand the behavior of CO_2 sequestration. However, these methods have high computational cost and time-consuming. As stated in Chapter 2, a few studies have been done on the use of data-driven statistical techniques, these studies mainly focused on production optimization in unconventional reservoirs. This fact emphasizes the need of giving serious thought on the application of data-driven modeling on characterizing the parameters that control CO_2 sequestration in unconventional reservoirs. In addition, if the E&P companies have the ambition to reach net zero target by 2050, this would require the need for CO_2 geologic storage underground. In order to achieve this, the pathway would involve the urgency to understand the parameters that control CO_2 sequestration.

3.1 Research aim, objectives, and questions

Given the inadequacy of research regarding CO_2 sequestration in unconventional reservoirs, this study will aim to identify the most important variables that affect this process and whether reservoir or operational parameters affect more. Including, how to establish predictive models based on machine learning to predict the volume of CO_2 sequestered as well as how to create decision rules that will aid in identifying the primary variables that influence the amount of CO_2 sequestered.

A list of research objectives can be included as follows:

- Perform an exploratory data analysis and discover hidden patterns.
- Quantify the correlation between volume of CO_2 sequestered and each input variable.
- Predict the cumulative injected CO_2 volume from the numerical simulation scenarios.
- Identify the drivers of CO_2 sequestration parameters among the considerable set of predictors using a variable importance method.

A list of research questions can be included as follows:

- What are the characteristics that are critical to the data and if there are outliers?
- Is there a relationship between two or more variables?
- How accurately the predictive model will correspond to future data?

- What parameters, or combinations of parameters, influence the performance of CO_2 sequestration?

3.2 Dataset and simulator description

The dataset used to achieve these aims and objectives comprised of an enormous set of numerical-simulation scenarios (approximately 1400 scenarios) that were run using a state-of-the-art reservoir simulator which was part of another study by (Kulga, 2014). Furthermore, the reservoir simulator used was a compositional dual-permeability, dual-porosity, multi-phase reservoir simulator developed at Penn State University (PSU-SHALECOMP). The simulator integrates the effects of water presence in the micropore structure together with matrix swelling and shrinkage. In these simulations, CO_2 sequestration was performed with a constant injection rate constraint after primary gas recovery period until a specified fracturing-pressure limit is reached. Table 3.1 presents the variables and their specified ranges. These ranges were used to randomly generate uniformly distributed scenarios for each variable. A combination of these input variables constitutes a given numerical simulation scenario for which sequestered volume of CO_2 is collected. In the numerical model, the network of induced fractures is represented using the stimulated reservoir volume (SRV) approach in which the fracture network is approximated as an elliptical area around the horizontal well (Figure 3.1).

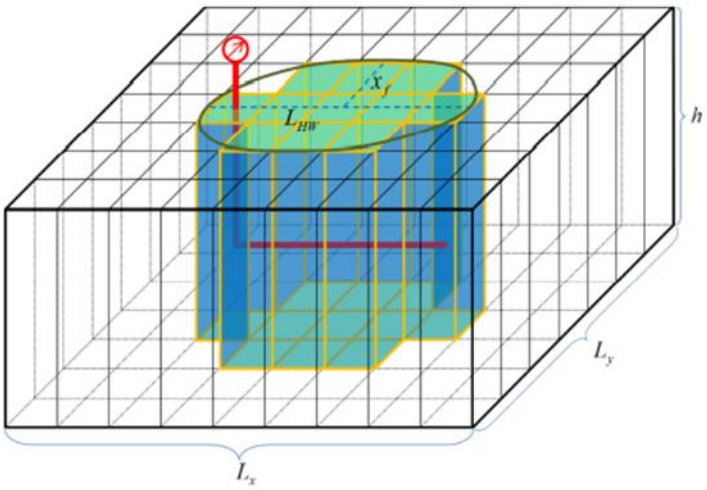

Figure 3.1 Approximation of the induced fracture network in the numerical model using the SRV approach (Kulga & Ertekin, 2018)

Table 3.1 Ranges of parameters used as inputs for the numerical simulation scenarios (Kulga, 2014)

	Minimum Value	***Maximum Value***	***Unit***
L_{hw}	2,000.86	4,998.36	ft
L_f	202.252	999.351	ft
L_x	1.201*L_{hw}	1.599*L_{hw}	ft
L_y	1.205*L_f	1.997*L_f	ft
h	100.226	299.704	ft
ϕ_m	5.0008	9.995	%
ϕ_f	0.5022	1.9999	%
SRV-ϕ_f	1.202*ϕ_f	1.499*ϕ_f	%
k_m	1.00E-06	1.00E-04	md
k_f	0.000101	0.001097	md
SRV-k_f	2.00797*k_f	11.9926*k_f	md
Δx_s	0.902	2.998	ft
SRV-Δx_s	0.401*Δx_s	0.799*Δx_s	ft
S_{wm}	5.003	13.998	%
$V_{L\text{-}CH4}$	50.33	248.99	scf/ton
$P_{L\text{-}CH4}$	201.971	998.081	psi
$V_{L\text{-}CO2}$	2.006*$V_{L\text{-}CH4}$	5.99*$V_{L\text{-}CH4}$	scf/ton
$P_{L\text{-}CO2}$	201.97	999.08	psi
P_i	3,004.92	7,997.14	psi
T_i	120.016	199.999	F
$q_{sf\text{-}prod}$	1,018,365	4,994,210	scf/d
t_{prod}	7306.09	18,233.0	days
P_{frac}	1.1*P_i	1.499*P_i	psi
$q_{sf\text{-}inj\text{-}STOPPING}$	200,429	597,457	scf/d

3.3 General workflow

The workflow that was used to answer the research problem was a data-analytics based investigation combined with statistical modeling. This workflow can be summarized in Figure 3.2.

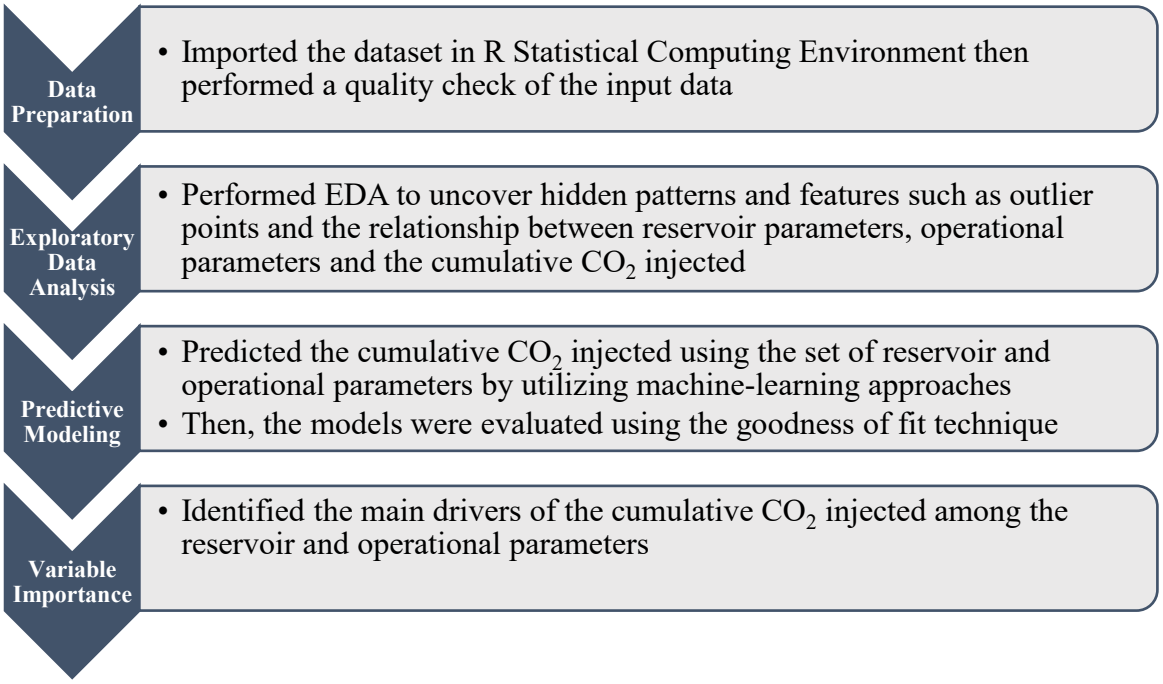

Figure 3.2 Workflow followed for data-analytics approach

Chapter 4 Methodology

4.1 Methodological approach

In this study, the aim was to identify the most important variables that affect the CO_2 sequestration process in unconventional shale reservoirs and whether reservoir or operational parameters affect more. The software used to run all the analysis was R Statistical Computing Environment (R Development Core Team, 2021). Furthermore, the approach taken to answer the research problem was a data-analytics based investigation and combining with statistical modeling. As explained by Mishra & Datta-Gupta (2018), it is more useful to consider DA and statistical modeling as part of an integrated data analysis cycle (Figure 4.1) for petroleum geoscience applications.

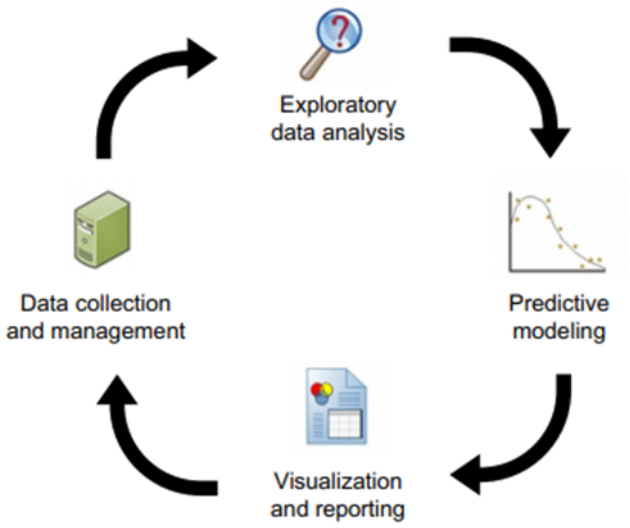

Figure 4.1 The cycle of data analysis (Mishra & Datta-Gupta, 2018)

4.2 Exploratory data analysis

Sensor measurements, events, text, photos, and videos are all examples of data sources. The Internet of Things (IoT) is generating an overload of data. Much of this data is unstructured: images comprise pixels, each of which contains RGB values (red, green, blue) information on color (Bruce et al., 2020). Numeric and categorical data are the two main types of structured data. Continuous data, such as wind speed or time period, and discrete data, such as the number of times an event occurs, are two types of numerical data. While data such as a type of TV screen or a state name are categorical data, since they can only take a fixed set of values (Bruce et al., 2020).

Bruce et al. (2020) points out that why do we need to be concerned about with a classification of data types? It turns out that the data type is critical in determining the type of visual display, data analysis, or statistical model used in data analysis and predictive modeling.

EDA was the first methodological approach taken in which the data was summarized, visualized and a more detailed analysis was performed. In this study EDA was partitioned into three main steps:

- Univariate data analysis
- Bivariate data analysis
- Multivariate data analysis

4.2.1 Univariate data analysis

The observed values of a variable are likely to differ from one another, whether we're dealing with a population or a sample. It's useful to quantify the average value, the spread around that average value, and the overall asymmetry over the entire range of observed values to investigate this intrinsic variability for a single variable numerically (Mishra & Datta-Gupta, 2018). These univariate metrics, as well as several standard graphical approaches for visually reviewing and summarizing the data, are explained here (Mishra & Datta-Gupta, 2018).

4.2.2 Measures of central tendency

The mean or expected value is the most popular measure of central tendency. The mean of a random variable X, where x_i are the individual outcomes (Mishra & Datta-Gupta, 2018), is given in **Eq 4.1:**

$$\boldsymbol{E[X] = \overline{X} = \sum_{i=1}^{N} f_i x_i = \frac{1}{N} \sum_{i=1}^{N} x_i} \qquad \textbf{Eq 4.1}$$

Where,

f_i relative frequency

X random variable

x_i individual outcomes

The weighted average of all values based on relative frequency is called the arithmetic mean (Mishra & Datta-Gupta, 2018). There are two more helpful measurements of central tendency

(a) median, this is the distribution's midpoint, and (b) mode, which is the most frequent occurring value (Mishra & Datta-Gupta, 2018). For symmetrical (or near-symmetrical) distributions, the mean, median, and mode are usually the same, but if the distribution is asymmetrical, they can be substantially different. The extreme numbers have a significant impact on the mean, whereas the median is more robust and less responsive to outliers (Mishra & Datta-Gupta, 2018). In Figure 4.2, the median lies between the mode and the mean in two situations, but the mean and mode swap positions depending on the asymmetry (i.e., left-skewed or right-skewed) (Mishra & Datta-Gupta, 2018).

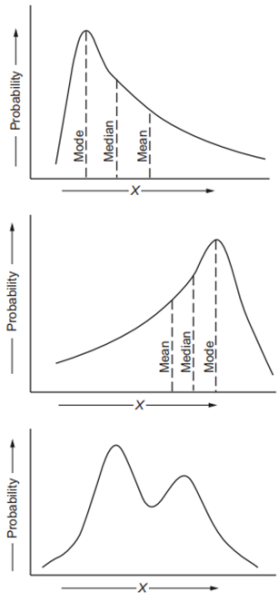

Figure 4.2 Position of mode for different category of distribution (Mishra & Datta-Gupta, 2018)

4.2.3 Measures of dispersion

For summarizing a feature, location is only one of many factors to consider. Variability, also known as dispersion, is a second dimension that determines whether the data values are clustered or spread out (Bruce et al., 2020). The variance, which measures dispersion or variability around the mean, is the most essential measure of spread (Mishra & Datta-Gupta, 2018). It's described by **Eq 4.2:**

$$\boldsymbol{V[X] = \sigma_x^2 = \sum_{i=1}^{N} f_i (x_i - E[X])^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - E[X])^2} \qquad \textbf{Eq 4.2}$$

$$\boldsymbol{V[X] = \frac{\sum x_i^2}{N} - (E[X])^2 = E[X^2] - (E[X])^2}$$

The difference between the mean of the squares and the square of the mean is the variance. The standard deviation is equal to the square root of the variance and the root-mean-square error (RMSE) (Mishra & Datta-Gupta, 2018).

4.2.4 Univariate data graphs

In this study, the univariate data graphing approached used was by plotting box plots, and histograms which are useful to explore data in one dimension.

Box plot. The box plot (Figure 4.3) (also known as the box-and-whisker diagram) is a standardized technique of depicting data distribution based on five essential features which are minimum, first quartile, median, third quartile, and maximum. In a box plot, the rectangle represented at the center spans the first quartile to the third quartile (IQR). A section inside the rectangle shows the whiskers and median below and above the box shows the position of the minimum and maximum (Kirkman, 1996).

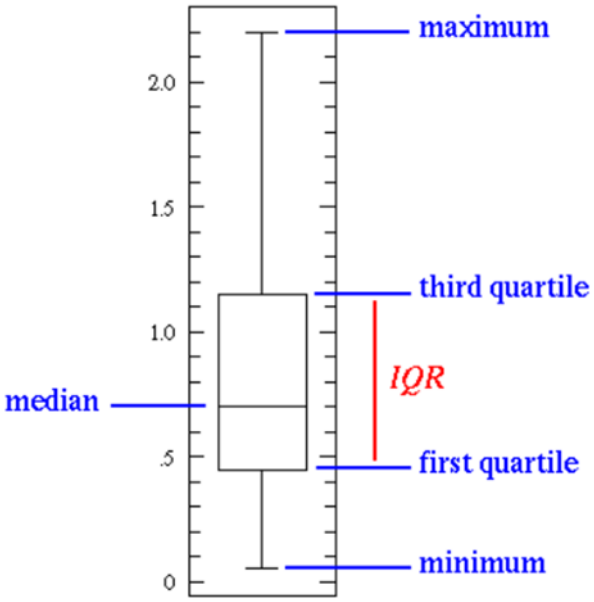

Figure 4.3 Box plot (Kirkman, 1996)

Outliers are data points that fall outside the box plot whiskers' maximum or minimum values (Kirkman, 1996). Figure 4.4 shows how the outliers can be visualized and seen in a box plot.

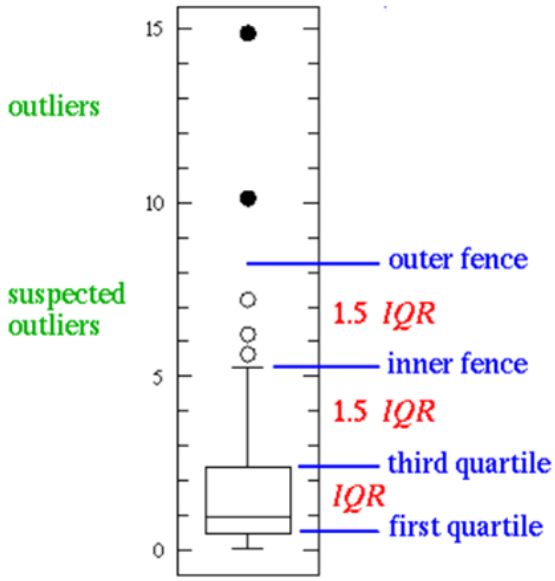

Figure 4.4 Outliers representation (Kirkman, 1996)

Histograms. The histogram's major purpose is to display the relative class frequencies in the data and, hence, provide information on the data density function. A histogram (Figure 4.5), which is essentially a bar plot of a frequency distribution grouped in intervals or classes, is a widely used graphical display of univariate data. Moreover, the central tendency, the dispersion, and the general shape of the distribution are all essential visual information that may be gained from histograms (Holdaway, 2009). It is made by splitting the observed range into many intervals (bins) and plotting the actual frequency of occurrence in each interval. The number of bins used in histograms is usually determined by trial and error. The following are some common rules of thumb that have been presented (Mishra & Datta-Gupta, 2018).

- The number of intervals k for a given sample size of N should be the smallest integer, such that $2k \geq N$ (Iman & Conover, 1986) as cited in (Mishra & Datta-Gupta, 2018).
- A suggestion given by Venables & Ripley (1996) is to use the number of bins as {3.3log (N) + 1} as a default value.

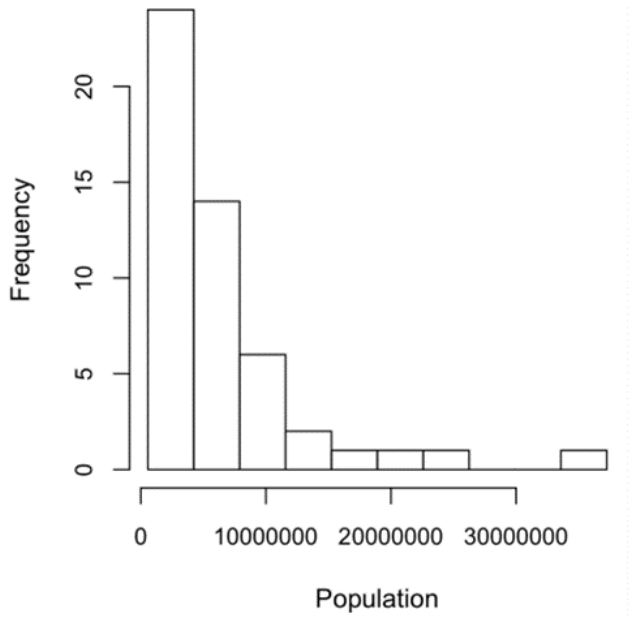

Figure 4.5 Histogram sample (Bruce et al., 2020)

The histogram's shape extremely depends on the number of intervals chosen. It will be sensitive to bin size (Figure 4.6), and hence it might not be a reliable graphic tool. Unless the analyst performs experiments of multiple bin sizes until a robust indication of shape is reached (Mishra & Datta-Gupta, 2018).

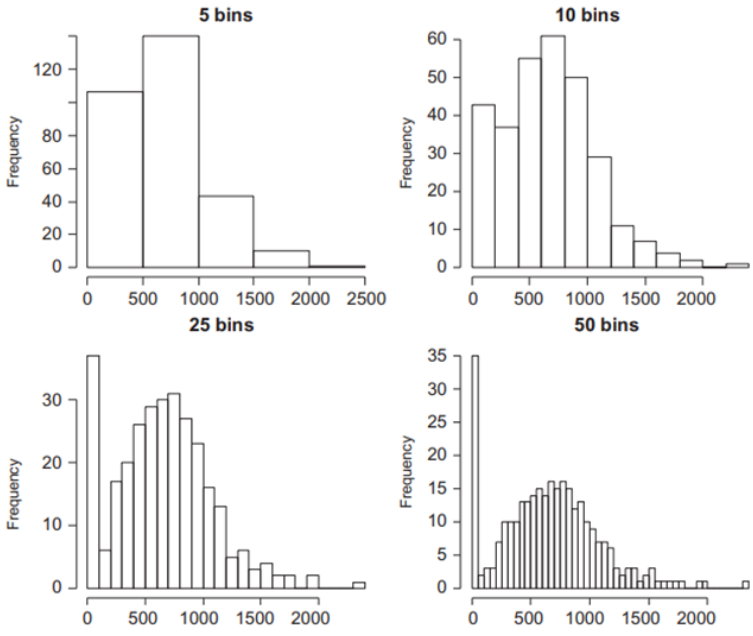

Figure 4.6 Sensitivity of bin size (Mishra & Datta-Gupta, 2018)

4.2.5 Bivariate data analysis

The main goal for bivariate data analysis is to describe the relationship between two variables. These bivariate measurements, as well as several standard graphical approaches for visually reviewing and summarizing the data, are explained here (Mishra & Datta-Gupta, 2018).

4.2.6 Correlations

Investigating the correlation among predictors and between predictors and a response variable is important in EDA and in many modeling projects (Bruce et al., 2020). The correlation coefficient (*CC*), often known as the Pearson correlation coefficient, is a measure of the strength of a linear relationship between two random variables (Mishra & Datta-Gupta, 2018). it is defined by **Eq 4.3:**

$$\boldsymbol{CC = \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{x_i - \bar{X}}{\sigma_x} \right) \left(\frac{y_i - \bar{Y}}{\sigma_y} \right)} \qquad \textbf{Eq 4.3}$$

Where,

σ_{xy}	Covariance
σ_x	Standard deviation of variable *x*
σ_y	Standard deviation of variable *y*
$\bar{X}$	Mean of variable *x*
$\bar{Y}$	Mean of variable *y*
$N-1$	Degrees of freedom
x_i	Individual outcome for *x*
y_i	Individual outcome for *y*

The *CC* value ranges between -1 and +1. whereby, a perfect negative correlation is showed by -1 and a perfect positive correlation is showed by +1. The absolute value measures the magnitude of the relationship, while the sign shows the trend's direction. It's vital to remember that the term "correlation" only relates to a monotonic relationship (Mishra & Datta-Gupta, 2018).

The rank correlation coefficient (RCC), also known as the Spearman correlation coefficient, can be employed as a more robust measure of nonlinear association if the variables of interest are associated in a nonlinear form (Mishra & Datta-Gupta, 2018). It is defined as follows by **Eq 4.4:**

$$\boldsymbol{RCC = 1 - \frac{6\sum d^2}{N(N^2 - 1)}} \qquad \textbf{Eq 4.4}$$

Where,

d difference of ranks

4.2.7 Bivariate data graphs

In this study different techniques were used to explore data in two dimensions. Mainly two graphing techniques were employed which were scatterplots and scatterplots combined with histograms.

Scatterplot. One of the simplest technique for depicting the relationship between two variables is to use a scatterplot (Mishra & Datta-Gupta, 2018). The horizontal axis displays the values of one variable, while the vertical axis displays the values of the other variable. If there is an explanatory variable (predictor variable), always plot it on the scatterplot's horizontal axis (x axis). The explanatory variable (predictor variable) is commonly referred to as x, and the response variable is referred to as y. Either variables can belong on the horizontal axis if there is no explanatory-response differentiation (Moore et al., 2018). In order to describe the overall pattern given by the scatterplot, three strategies are adopted: direction, form, and strength. The direction of the general pattern specifies whether it moves from lower left to upper right, upper right to lower left, or none of the two.

The approximate functional form is referred to as form. Is it, for example, roughly a straight line, curved, or oscillating? The strength of the plot is determined by how well the points in the plotline follow the form (Moore et al., 2018).

The absolute value of the Pearson *CC* (ρ) reflects the strength of the linear relation, whereas the sign of ρ shows whether the correlation is negative or positive. Several examples of scatter diagrams are provided in Figure 4.7, each depicting a different range of probable behavior between two generic variables, X and Y. A strong positive trend can be seen in the top-left panel (A). A very significant negative linear trend can be seen in the top-right panel (B) along with a modest negative correlation, may be seen in the bottom-left panel (C), while in the bottom-right panel (D) there is a moderate positive trend (Mishra & Datta-Gupta, 2018).

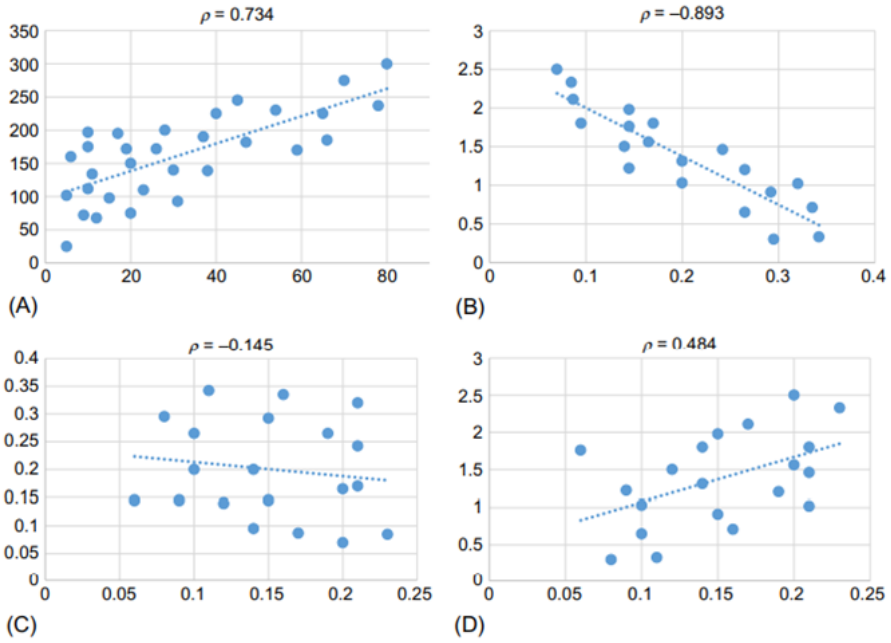

Figure 4.7 Multiple scatterplots with linear trend (Mishra & Datta-Gupta, 2018)

Scatterplot with marginal histogram. Histograms and scatterplots (Figure 4.8) can be used together to show how individual variables are distributed throughout their ranges. The marginal (individual) distributions of X and Y are represented by the histograms along the axes, whereas the scatterplot represents the combined distribution of X and Y (Mishra & Datta-Gupta, 2018).

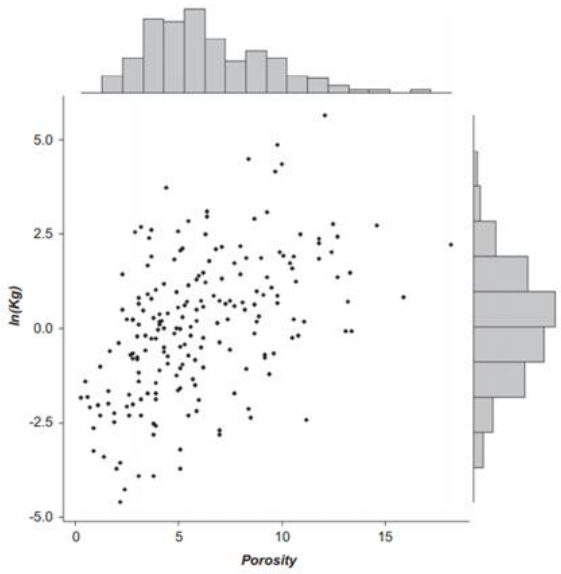

Figure 4.8 Scatterplot with histogram (Mishra & Datta-Gupta, 2018)

4.2.8 Multivariate data analysis

Correlation analysis in multivariate data extends the techniques covered earlier for bivariate data analysis. This requires computing the Pearson or Spearman CC for all variable pairs and displaying it as a correlation matrix (Figure 4.9). It suffices to show the lower or upper part of the matrix since the correlation matrix is symmetrical (Mishra & Datta-Gupta, 2018). Likewise, scatterplot matrix or a pairs plot can be used for data visualization, this is developed by incorporating different scatterplots of variable pairs to show their interaction (Venables & Ripley, 1996) as cited in (Mishra & Datta-Gupta, 2018). Each scatterplot can be colored coded to identify membership of specific data points in different groups and annotated with a smoothing line to help visualize the underlying trend. The benefit of scatterplot matrix (Figure 4.10) is that it allows you to get a quick overview of the relationships, patterns, and trends among predictor variables (independent variables) as well as between response variables (dependent variables) and predictor variables (Mishra & Datta-Gupta, 2018).

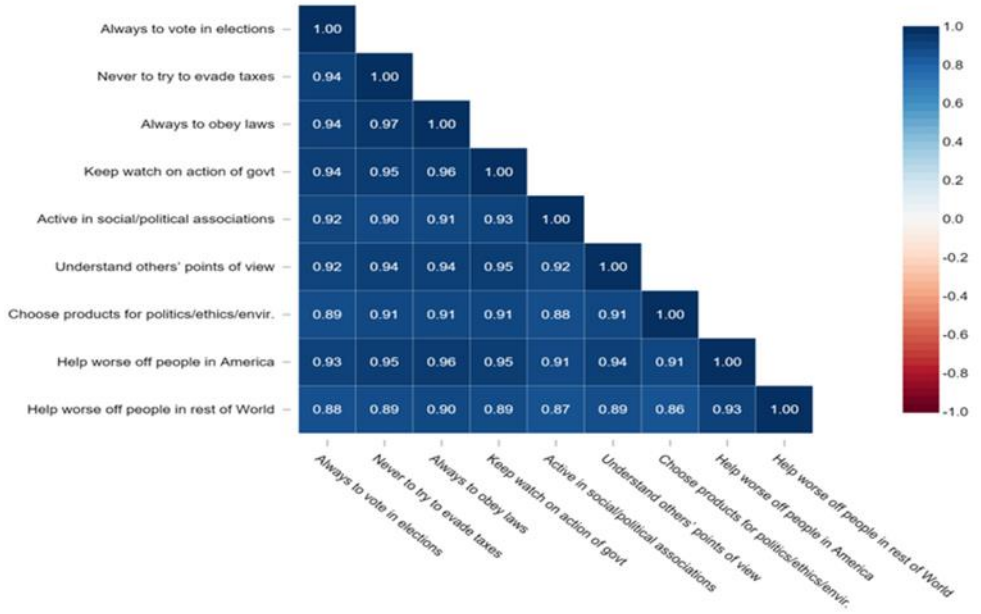

Figure 4.9 Correlation matrix (Bock, n.d.)

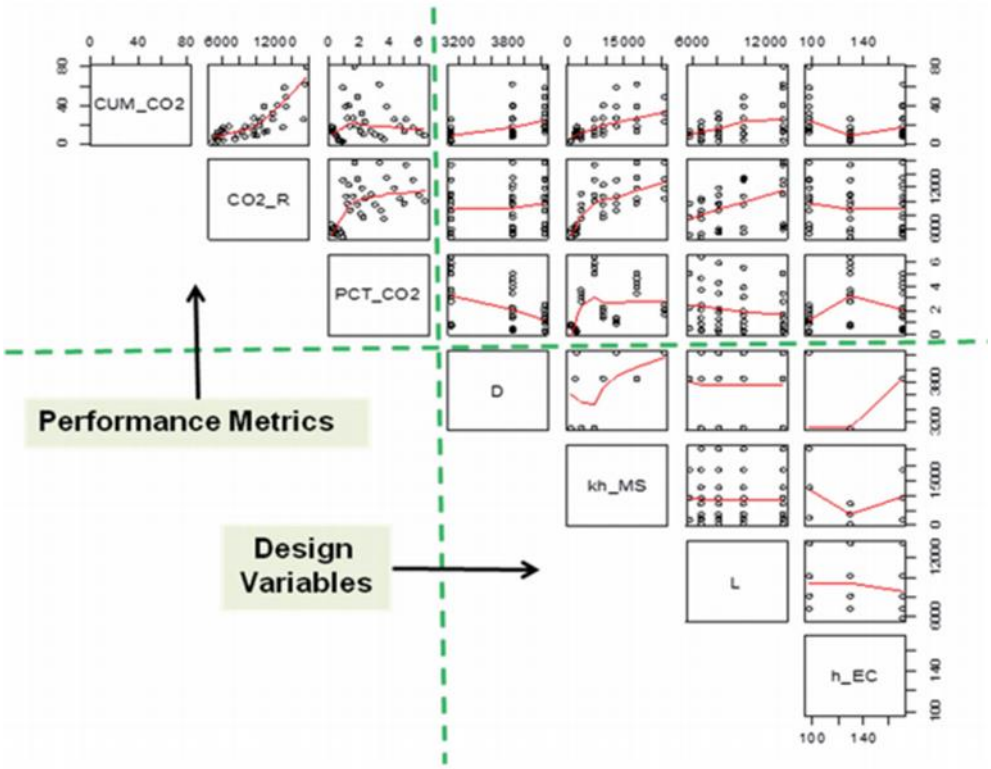

Figure 4.10 Scatterplot matrix (Mishra et al., 2014)

4.3 Predictive modeling

After developing a preliminary understanding and digging deeper into our data by using EDA, the next step in our methodology was to develop predictive models. Predictive models are important tools for understanding the relationship between response and predictor variables. The main focal point of this section is predictive statistical modeling, where statistical and machine-learning approaches will be used to discover the dependency or relation between dependent and independent variables. Al-Alwani et al. (2019) point out that the goal of using predictive analytics (predictive modeling) is to improve operations while cutting down costs and saving time. In this context, the keywords statistical learning, data mining, knowledge discovery, and data analytics are all interchangeable. Applying supervised and/or unsupervised learning, the purpose of such a scheme is to identify relevant patterns and trends and comprehend "what the data says" (Hastie et al., 2008) as cited in (Mishra & Datta-Gupta, 2018). The goal of supervised learning is to predict the value of an output measure based on a set of input measurements, while the goal of unsupervised learning is to explain the correlations and patterns among a set of input measures (Hastie et al., 2008).

This study mainly used supervised learning through techniques such as linear regression and tree-based methods such as bagging, random forests and gradient boosting machine (GBM).

4.3.1 Linear regression

One of the most extensively utilized strategies for investigating and exploiting the relationship between dependent (response) and independent (predictor) variables is regression modeling. Linear regression occurs when a relationship can be described using linear equations. It involves a single predictor and a response variable (Mishra & Datta-Gupta, 2018). While, multiple regression, also referred to as OLS, involves over one predictor variable. In this study, the concept of simple linear regression will be illustrated first then followed by multiple regression, later model selection and evaluation, and finally choosing the optimal model for the multiple regression.

4.3.2 Simple linear regression

It's a fairly simple method for predicting a quantitative response *Y* based on a single predictor variable *X*. It is presumptively assumed that *X* and *Y* have a linear relation (James et al., 2013). This linear relationship can be written mathematically by **Eq 4.5:**

$$\boldsymbol{Y \approx \beta_0 + \beta_1 X} \qquad \textbf{Eq 4.5}$$

Where,

Y	Quantitative response variable
X	Predictor variable
β_0	Intercept
β_1	Slope

In **Eq 4.5**, the intercept and slope terms are β_0 and β_1, respectively, which are two unknown constants in the linear model. The model coefficients or parameters are known as β_0 and β_1. We can forecast future data based on a particular value of the predictor variable by computing $\hat{\beta}_0$ and $\hat{\beta}_1$ (**Eq 4.6**) estimates for the model coefficients after we've used our training data to produce them (James et al., 2013).

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x} \qquad \textbf{Eq 4.6}$$

Where,

$\hat{y}$	Prediction of Y
$\hat{\beta}_0$	Coefficient estimates
$\hat{\beta}_1$	Coefficient estimates
x	X=x

β_0 and β_1 are unknown in practice. As a result, before we can use **Eq 4.5** to generate predictions, we must first estimate the coefficients using data (James et al., 2013). let

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

Denote n observation pairs, each of which comprises an X and a Y measurement. For example, given a dataset for TV advertising which comprises sales of that product for $n = 200$ markets. We want to find an intercept $\hat{\beta}_0$ and a slope $\hat{\beta}_1$that will cause a line that is as near (close) to the $n = 200$ data points as possible. Closeness can be measured in a variety of ways.

The most popular procedure, however, is to minimize the least squares criterion. Figure 4.11 shows that by minimizing the sum of squared errors, the best fit is found. Each grey line segment denotes an error, and the fit averages their squares as a compromise (James et al., 2013).

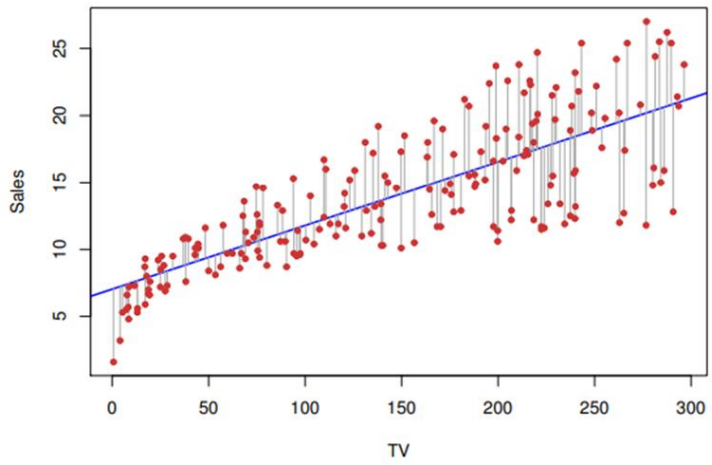

Figure 4.11 Least squares fit (James et al., 2013)

The residual sum of squares (RSS) can be defined by **Eq 4.7:**

$$\mathbf{RSS} = (\boldsymbol{y_1} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x_1})^{\boldsymbol{2}} + (\boldsymbol{y_2} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x_2})^{\boldsymbol{2}} + \cdots + (\boldsymbol{y_n} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x_n})^{\boldsymbol{2}} \quad \textbf{Eq 4.7}$$

To reduce the RSS, the least squares method chooses $\hat{\beta}_0$ and $\hat{\beta}_1$. It is possible to show that the minimizers can be given by **Eq 4.8** (James et al., 2013).

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} = \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{y_i} - \overline{\boldsymbol{y}})}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\boldsymbol{2}}} \quad \textbf{Eq 4.8}$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} = \overline{\boldsymbol{y}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\overline{\boldsymbol{x}}$$

Where,

$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ Sample mean

$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ Sample mean

For simple linear regression, **Eq 4.8** defines the least squares coefficient estimates. However, there is a chance that the genuine relationship is not linear, and that there are other variables causing variability in *Y*, and that measurement error exists. Normally, we presume that the error term is unaffected by *X*. The approach described by **Eq 4.9** illustrates the population regression line, which is the genuine relationship between *X* and *Y* and provides the best linear approximation. Whilst the least squares line (**Eq 4.6**) is defined by the least squares regression coefficient estimates (**Eq 4.8**) (James et al., 2013).

$$\boldsymbol{Y = \beta_0 + \beta_1 X + \epsilon} \qquad \textbf{Eq 4.9}$$

Where,

β_0 Intercept term

β_1 Slope

ϵ mean-zero random error term

Furthermore, after determining the least squares coefficient estimates, it's only logical to want to know how well the model fits the data. The residual standard error (RSE) and the R^2 statistic are commonly used to evaluate the quality of a linear regression fit (James et al., 2013).

Residual standard error. We know from **Eq 4.9** that each observation has an error term ϵ, so even if we had the actual regression line (β_0 and β_1), we wouldn't be able to perfectly predict *Y* from *X*. The standard deviation of ϵ is estimated using the RSE. It is expressing the average deviation of the response from the actual regression line. The **Eq 4.10** is used to calculate it (James et al., 2013).

$$\textbf{RSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n-2}}\textbf{RSS}} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n-2}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i - \hat{y}_i})^{\mathbf{2}}} \qquad \textbf{Eq 4.10}$$

Where,

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

***R^2* statistic.** The RSE is an absolute measure of the **Eq 4.9** lack of fit to the data. However, because it is expressed in the units of *Y*, it is not always apparent what defines a good RSE. The R^2 statistic is another way to assess fit. It is independent of the scale of *Y*, as well as taking on a value between 0 and 1. As a result, it is known as the proportion of variance explained (James et al., 2013). The **Eq 4.11** is used to calculate R^2 statistic.

$$\boldsymbol{R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}} \qquad \textbf{Eq 4.11}$$

Where,

TSS (total sum of squares) = $\sum (y_i - \bar{y})^2$

4.3.3 Statistical significance

Before moving into multiple linear regression, it is important to understand the concept of statistical significance. At first, the concept of establishing a null hypothesis that we wish to uncover evidence against appears unusual. Consider a criminal trial as an example. "Until proven guilty," the defendant is presumed innocent. The null hypothesis is innocence, and the prosecution must strive to disprove this hypothesis with persuasive evidence. That's exactly how statistical significance tests operate. Except in statistics, we deal with data-based evidence and apply a probability to determine how strong it is (Moore et al., 2018).

A *P-value* is a probability that quantifies the degree of evidence against a null hypothesis. The *P*-value of the test is the probability that the test statistic will take a value as severe or more extreme than that actually observed, given that the null hypothesis (H_0) is true. The lower the *P*-value, the stronger the data's proof against H_0. Small *P*-values provide evidence against H_0, since they show that the observed outcome is unlikely to occur if H_0 is correct. Large *P*-values do not provide proof against H_0. How low of a *P*-value is striking evidence against H_0? Many statisticians believe that results less than 0.05 or 0.01 are acceptable (Moore et al., 2018).

4.3.4 Multiple linear regression

For predicting a response based on a single predictor variable, simple linear regression is a helpful method. In actuality, though, we frequently have over one predictor. Running three independent linear regressions is one possibility, as the method of fitting a separate basic linear

regression model for each predictor is not highly recommendable. Rather than constructing a separate simple linear regression model for each predictor, extending the simple linear regression model (**Eq 4.9**) to directly handle multiple variables is a preferable method. In a single model, we may achieve this by assigning a distinct slope coefficient to each predictor (James et al., 2013).

Assume we have *p* unique predictors; the multiple linear regression model then assumes the following model:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon} \qquad \textbf{Eq 4.12}$$

Where,

X_j *j*th predictor

β_j quantifies association between that variable and the response

We depict β_j as the average effect of a one-unit increase in X_j on Y, with all other predictors held constant. The regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in **Eq 4.12** must be estimated since they are unknown, such as it was in the simple linear regression (James et al., 2013). We may use the **Eq 4.13** to generate predictions based on the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

$$\boldsymbol{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p} \qquad \textbf{Eq 4.13}$$

By using the least squares approach that we saw in simple linear regression, the parameters can be estimated. In order to minimize the sum of squared residuals, we choose $\beta_0, \beta_1, \ldots, \beta_p$ (James et al., 2013).

$$\boldsymbol{RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

$$\boldsymbol{RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2} \qquad \textbf{Eq 4.14}$$

The least squares regression line becomes a plane in a three-dimensional situation (Figure 4.12) with two predictors and one response. Moreover, the plane is selected so that the total of the squared vertical distances between each observation and the plane is as small as possible (James et al., 2013).

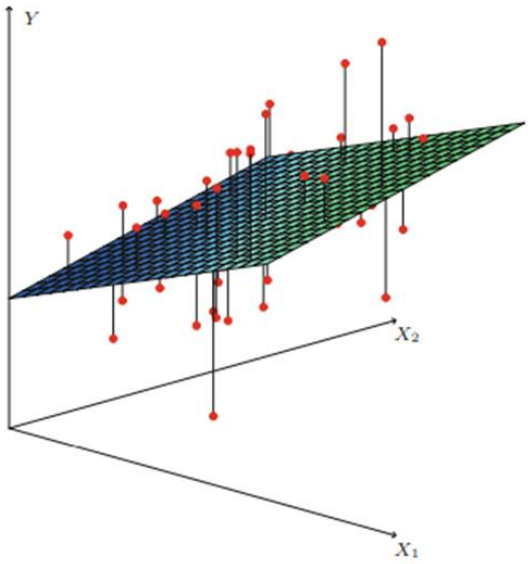

Figure 4.12 Least squares fit for multiple regression (James et al., 2013)

We generally want to find answers to a few key questions when we perform multiple linear regression (James et al., 2013):

1) Is it possible to predict the response variable using at least one of the predictors X_1, X_2,..., X_p?
2) Is it possible to use all the predictors to explain Y, or is it only possible to use a subset of them?
3) What is the model's fit to the data?

Is there an association between the response and the variables that predict it?

We need to check if all the regression coefficients are zero in a multiple regression with p predictors (in case $\beta_1 = \beta_2 = \cdots = \beta_p = 0$) (James et al., 2013).

To address the question, we use a hypothesis test. in particular, we put the null hypothesis into the test.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

against the alternative

$$H_a : \text{at least one } \beta_j \text{ is not a } 0$$

The *F- statistic* is used to do this hypothesis test (James et al., 2013).

$$\boldsymbol{F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}} \qquad \textbf{Eq 4.15}$$

In simple terms, we should expect *F* to be greater than 1 if there is a relationship, otherwise when there is no association between the response and the predictors, the F-statistic should be near to 1. What is the minimum F-statistic before we can rule out H_0 and conclude that there is a relationship? The solution turns out to depend on the values of *n* and *p*. When *n* is big, even an F-statistic somewhat larger than 1 can give evidence against H_0. If *n* is small, however, a higher F-statistic is required to reject H_0 (James et al., 2013).

Variable selection problem

The F-statistic is computed, and the accompanying p-value is examined as the first step in a multiple linear regression analysis. If we infer that at least one of the predictors is connected to the response based on that p-value, it's reasonable to speculate which are the ones that are bad. It is more often the case that the response is only related to a subset of the predictors. But it is possible that all the predictors are associated with the response. Variable selection is a technique used to determine which predictors are associated with the response in order to fit a single model that associate only those predictors (James et al., 2013). In an ideal world, we'd test out several models, each including a different subset of the predictors, to conduct variable selection. We may then choose the best model out of all the models we've evaluated after generating a model with a different selection of predictors. How can we know which model is the best? A variety of statistics may assess a model's quality. Among them are Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2 and Mallow's C_p (James et al., 2013). The concept of variable selection problem will be assessed more in detail in the next section of the methodology.

Fit of the model

The RSE and R^2, or the proportion of variance explained, are two of the most used numerical metrics of model fit. These values are calculated and interpreted in the same way as they are for simple linear regression. Normally, the square of the response and variable is given by the correlation R^2. This R^2 value, though, turns out to be equal to $\text{Cor}(Y,\hat{Y})^2$, the square of the correlation between the response and the fitted linear model, in multiple linear regression.

As a matter of fact, among all potential linear models, one feature of the fitted linear model is that it maximizes this correlation. Moreover, the model explains a substantial amount of the variance in the response variable if the R^2 value is near to 1 (James et al., 2013).

4.3.5 Selection of a linear model

It's not unusual for any or all of the variables included in a multiple regression to be unrelated to the response. Including such unimportant variables in the model results in needless complexity. We may get a more readily understood model by eliminating these variables—that is, by setting the associated coefficient estimates to zero. It's highly improbable that least squares will produce any coefficient estimates that are exactly zero (James et al., 2013). In the next section, we will see a technique for automatically performing a variable selection procedure in order to eliminate unrelated variables from a multiple linear regression model. There are many approaches for excluding irrelevant variables, but in the following section, best subset selection will be illustrated.

4.3.6 Best subset selection

We fit a separate least squares regression for each practical combination of the p predictors to achieve optimal subset selection. In other words, we fit all p models with precisely one predictor, full $\binom{p}{2} = p(p-1)/2$ models that comprise altogether two predictors, and so forth. We next examine all the resultant models in order to determine which one is the best (James et al., 2013). The steps involved in the best subset selection are given below as explained by (James et al., 2013):

1) Let M_0 stand for the null model, which is free of predictors. For each observation, this model simply estimates the sample mean.
2) Considering $k = 1, 2, \ldots p$:
    a) All $\binom{p}{k}$ models with precisely k predictors must be fitted.
    b) Choose the finest of these $\binom{p}{k}$ models and name it M_k. The best is defined as having the least RSS or, in other words, the biggest R^2.
3) Using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2, choose a single best model from among M_0,..., M_p.

Now all we have to do is choose between these p+1 alternatives to find the best model. Because the RSS of these p+1 models drops monotonically and the R^2 grows monotonically as the

number of components included in the models increases, this work must be done with caution. The difficulty is that a model with a low RSS or a high R^2 has a low training error, but we want a model with a low-test error. As a result, if these statistics are used to choose the optimal model, we will always end up with a model that includes all the variables. Hence, we put to use cross-validated prediction error, BIC, C_p, or adjusted R^2 in step 3 in order to select the optimal model among M_0, M_1,..., M_p (James et al., 2013).

4.3.7 Deciding on the best model

As seen previously, RSS and R^2 are ineffective in choosing the best model from a set of models with varying amounts of predictors. Hence, we must estimate the test error in order to choose the optimal model in terms of test error. There are two techniques that are commonly used (James et al., 2013):

- By adjusting the training error to account for the bias caused by over-fitting, we may estimate test error indirectly.
- Using either a validation set or a cross-validation technique, we may directly estimate the test error.

These approaches will be considered below.

Mallow's C_p. The C_p estimate of test MSE is obtained using the equation **Eq 4.16** for a fitted least squares model with d predictors (James et al., 2013).

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2) \qquad \textbf{Eq 4.16}$$

Whereby, the variance of the error ϵ combined with each response measurement is given by the estimate $\hat{\sigma}^2$. Normally, $\hat{\sigma}^2$ is calculated using the complete model, which includes all predictors. To compensate because the training error underestimates the test error, the C_p statistic adds a $2d\hat{\sigma}^2$ penalty to the training RSS. The penalty obviously increases as the number of predictors in the model grows; this compensates for the associated reduction in training RSS. As a result, the C_p statistic takes on a small value for models with minimal test error, thus we select the model with the lowest C_p value when deciding which of a group of models is best (James et al., 2013).

Akaike information criterion (AIC). For a large class of models fit by maximum likelihood, the AIC criterion is described. Maximum likelihood and least squares are the same thing with model **Eq 4.12** with Gaussian errors (James et al., 2013).

The AIC is defined by **Eq 4.17:**

$$\mathbf{AIC} = \frac{\mathbf{1}}{\boldsymbol{n\hat{\sigma}^2}} \left(\mathbf{RSS} + \mathbf{2}\boldsymbol{d\hat{\sigma}^2}\right) \qquad \textbf{Eq 4.17}$$

As a result, C_p and AIC are proportional to each other for least squares models.

Bayesian information criterion (BIC). BIC is developed from a Bayesian perspective, yet it resembles C_p and AIC in display. The BIC for a least squares model with d predictors is provided by **Eq 4.18** up to irrelevant constants. For a model with a low-test error, the BIC, like C_p, will take on a small value, therefore we choose the model with the lowest BIC value (James et al., 2013).

$$\mathbf{BIC} = \frac{\mathbf{1}}{\boldsymbol{n\hat{\sigma}^2}} \left(\mathbf{RSS} + \mathbf{log}(\boldsymbol{n})\, \boldsymbol{d\hat{\sigma}^2}\right) \qquad \textbf{Eq 4.18}$$

Adjusted R^2 statistic. Another frequent method for deciding amongst a collection of models with varying numbers of variables is to use the adjusted R^2. Because the RSS diminishes as more variables are added to the model, the R^2 rises. The adjusted R^2 statistic for a least squares model with d variables is obtained as follows (James et al., 2013):

$$\mathbf{Adjusted}\ \boldsymbol{R^2} = \mathbf{1} - \frac{\mathbf{RSS}/(\boldsymbol{n} - \boldsymbol{d} - \mathbf{1})}{\mathbf{TSS}/(\boldsymbol{n} - \mathbf{1})} \qquad \textbf{Eq 4.19}$$

A big value of adjusted R^2 suggests a model with a small test error, unlike C_p, AIC, and BIC, where a small value shows a model with a low-test error (James et al., 2013).

Validation set approach. It comprises splitting the set of observations into two halves at random: a training set and a validation set (or hold-out set). The training set is used to fit the model, and the fitted model is used to predict the response for the validation set observations (James et al., 2013). The test error rate is estimated using the validation set error rate, which is generally measured using MSE in the situation of a quantitative response. The validation set technique is depicted in a schematic diagram (Figure 4.13). A collection of n observations is

divided into a training set in blue and a validation set in beige at random. The training set is used to fit the statistical learning technique, and the validation set is used to evaluate its performance (James et al., 2013).

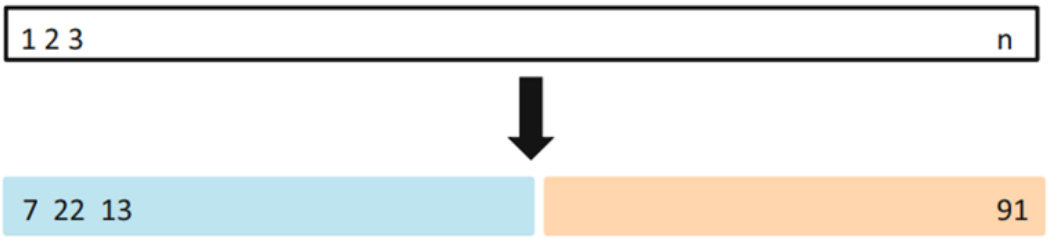

Figure 4.13 Validation set procedure (James et al., 2013)

The validation set technique is both theoretically and practically straightforward. However, there are two possible drawbacks (James et al., 2013):

- Depending on which observations are included in the training set and which observations are included in the validation set, the validation estimate of the test error rate might be extremely varied.
- Only a subset of the observations is used to fit the model in the validation method—those that are included in the training set, rather than the validation set. Given that statistical techniques perform worse when trained on fewer observations, the validation set error rate may overemphasize the test error rate for the model fit across the complete data set.

To address these potential drawbacks, a k-fold cross validation will be presented as a refinement technique.

***k*-fold cross validation.** The training dataset is randomly divided into *k* distinct groups or folds in this method, see Figure 4.14. Following that, each of the k groups is held out one at a time, and the model is trained on the other *k*-1 groups before applying to the group that was held out (Mishra & Datta-Gupta, 2018). There will be a single cross validated prediction for every observation in the dataset after cycling through all *k* groups, through which the predictions were created by adopting a model for which the training set does not include that observation. Through repeating the process with a unique set of *k* groups at random, the cross-validation technique may be expanded. Using r repeated runs of *k* randomly selected groups, a repeated cross validation will provide *r* distinct predictions for each observation. Not only may these be

used to generate statistics on goodness-of-fit measures, but they also provide valuable insight into model prediction variability as a function of the training set's properties (Mishra & Datta-Gupta, 2018).

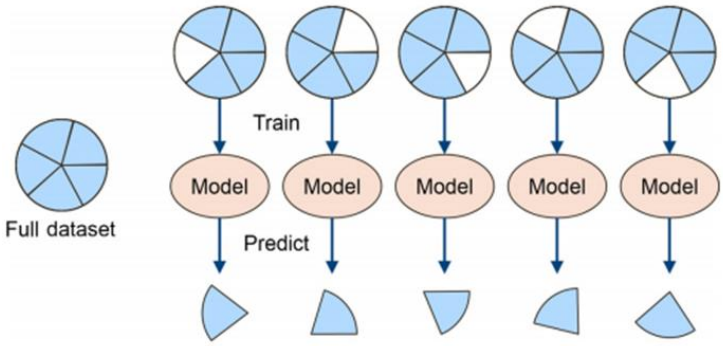

Figure 4.14 k-fold cross validation (Schuetter et al., 2018)

4.3.8 Goodness of fit

Average absolute error (AAE). The average magnitude of the difference between the real and predicted response is defined as the AAE (in other words, the average size of the residuals) given by **Eq 4.20** (Mishra & Datta-Gupta, 2018).

$$\mathbf{AAE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}|\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}| \qquad \textbf{Eq 4.20}$$

Where,

y_i	True response
$\hat{y}_i$	Predicted response

Mean squared error (MSE). MSE is identical to AAE, except instead of the absolute value, it measures the average squared difference between observations and their associated predictions (Mishra & Datta-Gupta, 2018).

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}} \qquad \textbf{Eq 4.21}$$

MSE uses the response variable's squared units, whereas AAE uses the same units as the response variable. The root-mean-square error, or RMSE, is a popular MSE variation that is just the square root of MSE. Closer to zero values are preferable since they imply fewer differences between observations and predictions (i.e., more accurate prediction) (Mishra & Datta-Gupta, 2018). Because of its well-known distributional characteristics and ability to be an adequate statistic for normally distributed processes, MSE or RMSE is generally selected over AAE (Navidi, 2008) as cited in (Mishra & Datta-Gupta, 2018).

4.3.9 Regression diagnostics

Once you've completed a regression analysis, you should always check if the model works properly for the data you're working with. Additionally, there are several assumptions about the data at hand made by linear regression. For instance, the response variable and the predictor variable have a linear relationship. This may not be the case. It's possible that the relationship is polynomial or logarithmic. Furthermore, the results of the regression could be affected because the data might contain outliers which are influential observations (Kassambara, 2017). As a result, you should do a regression diagnostic on the model you created to identify any issues and determine if the linear regression model's assumptions are satisfied or not. In order to do this, we will first define what are the fitted (predicted) values and residuals and then look at the regression assumptions (Kassambara, 2017).

Fitted (predicted) values and residuals. According to the built regression model, the fitted (predicted) values are the y-values you would estimate for the provided x-values (or simply, the best-fitting regression line). Whereas residual errors are the difference between observed (measured) values from the predicted values. As seen in Figure 4.15 these residuals are expressed by red vertical lines (Kassambara, 2017).

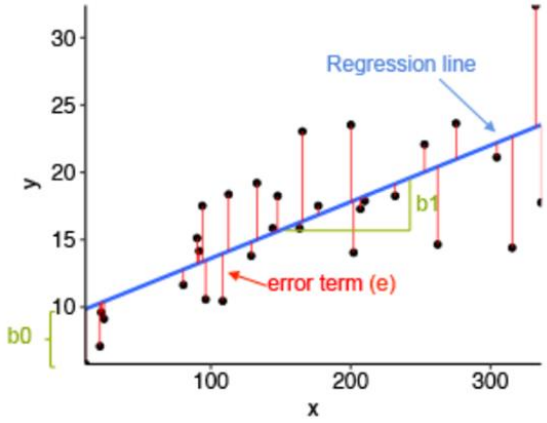

Figure 4.15 Residual errors (Kassambara, 2017)

Kassambara (2017) points out that there are several assumptions about the data made by linear regression including:

1) The data's linearity. The predictor and response variable are considered to have a linear relationship.
2) Residuals' normality. It is assumed that the residual errors are normally distributed.
3) Variance of the residuals should be homogeneous. The variance of the residuals is considered being constant (homoscedasticity).
4) Residual error terms should be independent.

Linearity of the data. The residuals vs fitted (predicted) plot (Figure 4.16) can be used to test the linearity assumption. Moreover, the red line should be approximately horizontal at zero in order for the residual plot to show no fitted pattern. Otherwise, it may show a problem with the linear model (Kassambara, 2017).

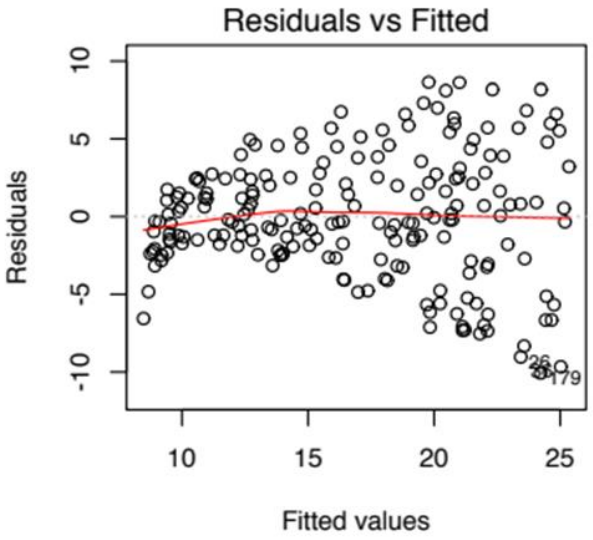

Figure 4.16 Residuals vs Fitted plot (Kassambara, 2017)

Homogeneity of variance. The scale-location plot, also known as the spread location plot, can test this assumption (Figure 4.17). This graph demonstrates if residuals are distributed evenly across predictor ranges. If you observe a horizontal line with evenly spaced points, that's a positive indicator (Kassambara, 2017).

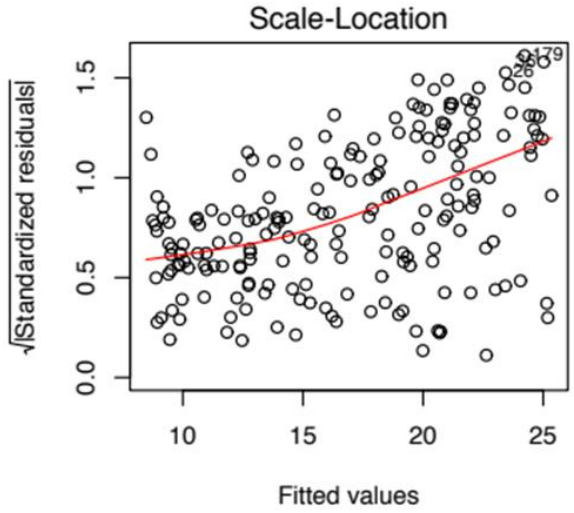

Figure 4.17 Scale location plot (Kassambara, 2017)

Normality of residuals. To visually confirm the normality assumption, take advantage of the QQ plot of residuals (Figure 4.18). The residuals normal probability plot should roughly follow a straight line (Kassambara, 2017).

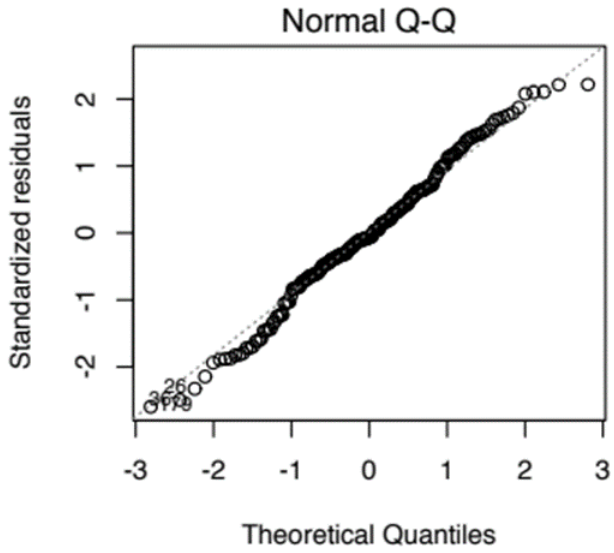

Figure 4.18 Normal QQ plot (Kassambara, 2017)

Outliers and high leverage points. The residuals versus leverage graph (Figure 4.19) may be used to spot outliers and high leverage points (Kassambara, 2017). Outliers are observations with standardized residuals higher than 3 in absolute value (James et al., 2013). For high

leverage points, the leverage statistic can identify this. A value of this statistic greater than $2(P+1)/n$ implies a high-leverage observation (Bruce et al., 2020).

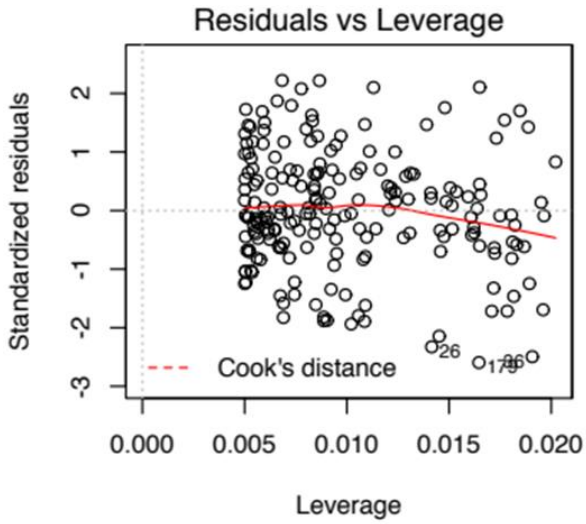

Figure 4.19 Residuals vs Leverage plot (Kassambara, 2017)

4.3.10 Tree methods

Classification and Regression Trees (CART). Tree techniques are straightforward interpretative models that illustrate how predictors influence response (Breiman et al., 1984) as cited in (Mishra & Datta-Gupta, 2018). The basic concept is to (a) divide the predictor space into nested rectangular areas, and (b) predict the response using a constant value for a regression question or a categorical label for a classification question inside each region. As displayed in Figure 4.20, the resultant binary tree (right panel) may discover structure in data and to generate prediction rules that split output into groups based on input values (Mishra & Datta-Gupta, 2018).

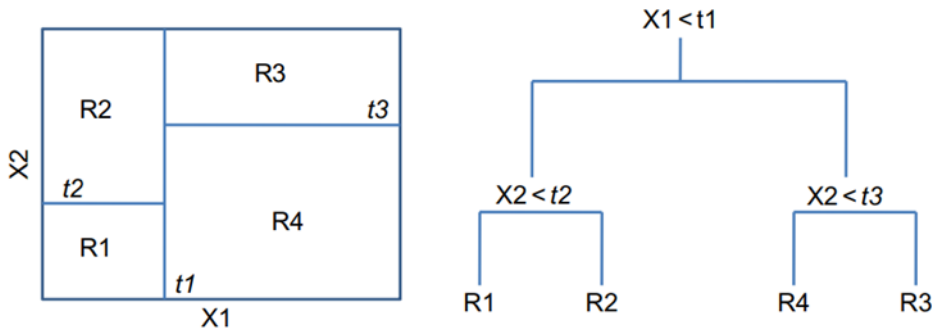

Figure 4.20 Tree based partitioning (Mishra & Datta-Gupta, 2018)

These methods are known as decision tree methods because the set of splitting criteria used to divide the prediction space may be described in a tree (James et al., 2013). For this study, a regression tree was used. In this section, the fundamental process of building regression tree will be discussed but not for classification since the study used regression trees. There are mainly two steps (James et al., 2013):

- A set of potential values X_1, X_2,..., X_p which represents the predictor space, will be branched into non-overlapping and J definite regions, R_1, R_2,..., R_J.
- The same prediction will be made for every observation that falls inside the R_j area. This prediction corresponds to the mean response values for the training observations in R_j.

How can the regions R_1,..., R_J be constructed from step 1 above? For simplicity and ease of comprehension of the resultant prediction model, we choose to partition the predictor space into high-dimensional rectangles, or boxes. The aim is to locate boxes R_1,..., R_J that have the least amount of RSS (James et al., 2013). This is given by **Eq 4.22**:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \qquad \textbf{Eq 4.22}$$

Where,

$\hat{y}_{R_j}$ mean response for training observations

However, considering every conceivable partition of the feature space into J boxes is computationally impractical. As a result, recursive binary splitting method is used, which is known as a greedy, top-down method. The method is top down because it starts at the top of the tree and separates the predictor space sequentially, with each split represented by two new branches deep down the tree. Besides, the method is greedy because, rather than looking forward and selecting a split that will lead to a better tree at a later step, this method considers the best split that is produced at each phase of the tree-building process (James et al., 2013).

If you want to perform recursive binary splitting, choose the appropriate predictor X_j and the cutpoint s so that partitioning the predictor space into the areas {X | Xj < s} and {X | Xj < s} reduces RSS as much as possible. That is, all potential cutpoint s values for each predictor as

well as all predictors X_1,..., X_p and then select the predictor and cutpoint that produces the minimum RSS tree (James et al., 2013). We define the pair of half-planes for any j and s in further detail as seen **Eq 4.23** below (James et al., 2013):

$$\boldsymbol{R_1(j,s) = \{X | X_j < s\} \; and \; R_2(j,s) = \{X | X_j \geq s\}} \qquad \textbf{Eq 4.23}$$

Then, we're looking for the j and s values that will make **Eq 4.24** as small as possible:

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \qquad \textbf{Eq 4.24}$$

Where,

$\hat{y}_{R_1}$ mean response for the training observations in $R_1(j,s)$

$\hat{y}_{R_2}$ mean response for the training observations in $R_2(j,s)$

Then we repeat the procedure, seeking for the optimal predictor and cutpoint to further separate the data and reduce the RSS within each of the resultant regions (James et al., 2013). Nonetheless, when you have a large dataset with many predictors, this entire tree, containing all predictors, seems to be highly complicated and might be difficult to comprehend. Also, it's clear that a fully developed tree would overfit the training data, perhaps resulting in poor test set performance (Kassambara, 2017). In order to avoid and overcome this issue we can reduce or stop the tree to grow. Growing the tree to nearly full size and then picking the sub-tree that optimizes some complexity criterion is a popular pruning method (Breiman et al., 1984) as cited in (Mishra & Datta-Gupta, 2018). This is usually considered comprising a summation term showing overall node impurity, as well as a penalty term combining a tuning, in other words cost complexity parameter and the number of terminal nodes (Mishra & Datta-Gupta, 2018). Additionally, the trade-off between tree size and its goodness of fit to the data is governed by this cost-complexity parameter in which large values of the parameter correspond to smaller trees and vice versa. For instance, Figure 4.21 shows the pruning chart which demonstrates this trade-off for a tree (Perez et al., 2005).

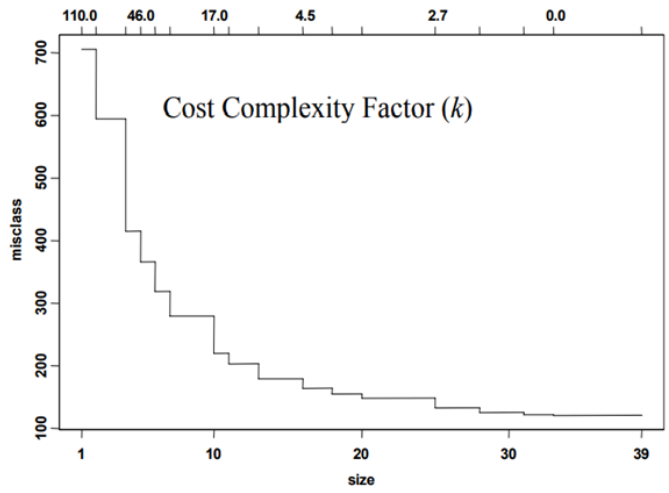

Figure 4.21 Pruning (cost complexity) graph (Perez et al., 2005)

The most significant predictors may be easily spotted at the top of the tree once the optimum tree has been formed. For example, Figure 4.22 displays a classification problem in which the most significant well logs are the photoelectric (PEF), density (DT), and neutron porosity (NPHI) (Mishra & Datta-Gupta, 2018).

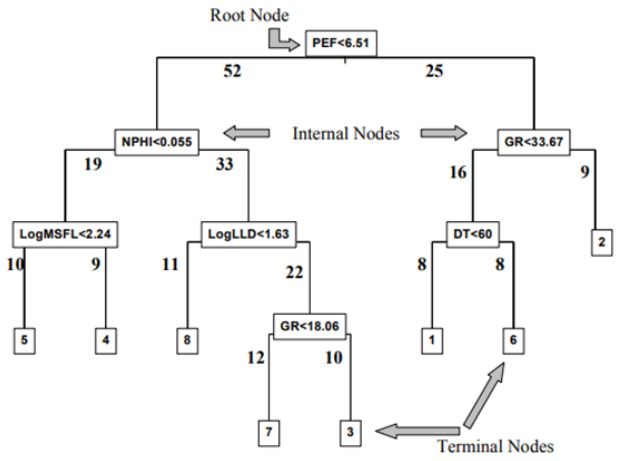

Figure 4.22 Decision tree (Perez et al., 2005)

Bagging. Bagging is a frequently used and particularly useful technique in the framework of decision trees. In addition, bagging (aggregation) is a strategy for decreasing the variance of a statistical learning method that may be used everywhere (James et al., 2013). It involves repeatedly combining several bootstrapped subsets of the data and averaging the models to create multiple distinct decision tree models from a single training dataset. Each tree is

constructed independently of the others (Kassambara, 2017). Whereas bootstrap re-sampling involves choosing a sample of n observations from the original dataset several times and evaluating the model for each iteration. After that, an average standard error is produced, and the results show the overall variance in the model's performance (Kassambara, 2017).

Random Forest. Using a bagging method, random forest regression produces an ensemble of trees to improve the performance of a single regression tree (Breiman, 2001) as cited in (Mishra & Datta-Gupta, 2018). Variety is added by using subsets of the input data and/or predictors to create many trees and therefore see the dataset from various viewpoints as an ensemble or random forest, because using the whole input dataset would always result in the same regression tree (Mishra & Datta-Gupta, 2018).

Essentially, each split considers a random subset of the predictors along with each tree in the ensemble is trained using a bootstrap sample of the training data. The regression tree focuses on moderately different aspects of the predictor-response relationship because of this randomization. Thanks to an averaging step that lowers the variation caused by individual trees' noisy nature, the trees can integrate this information into a strong prediction tool (Mishra & Datta-Gupta, 2018). A series of regression trees, each of which is constructed from random selections of data points and predictors using the regression tree-building approach outlined in the preceding section, is the starting point for creating an RF regression model (Figure 4.23) (Mishra & Datta-Gupta, 2018).

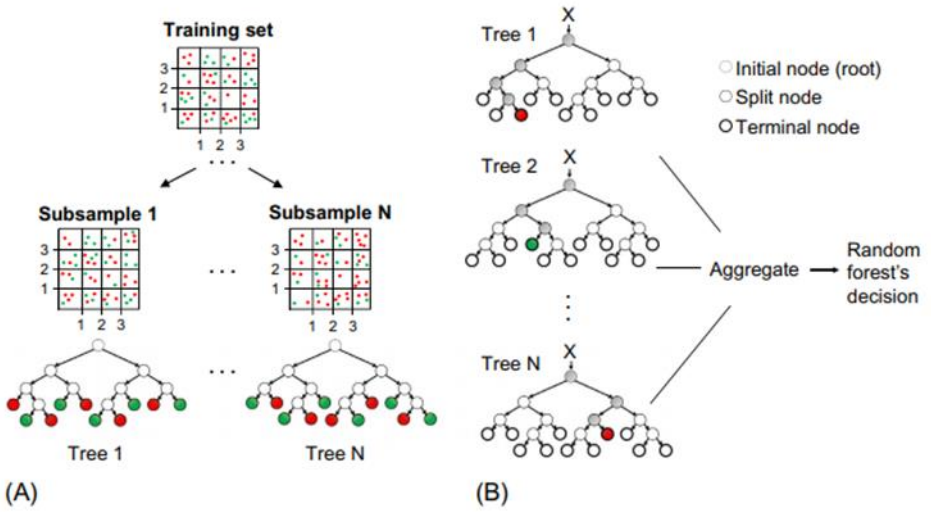

Figure 4.23 Random forest model (Mishra & Datta-Gupta, 2018)

Likewise, each new observation is run through all the trees in the ensemble for prediction, resulting in a distinct regression estimate. The average of the individual tree-level estimations is the final model prediction. The built-in cross validation feature in the RF algorithm makes it simple to validate the prediction model. Moreover, the remaining observations are referred to as out-of-bag samples, since each tree only sees a portion of the data. Those out-of-bag samples may be considered as independent test data and used to produce error rate estimates to assess model performance for that tree (Mishra & Datta-Gupta, 2018). Further details regarding RF classifier and construction of the RF model can be found in (Hastie et al., 2008).

Gradient Boosting Machine (GBM). Gradient boosting of regression trees is appropriate for mining less than clean data, also they produce highly robust, competitive, and interpretable techniques for both regression and classification (Friedman, 2001). Instead of constructing a single complicated model, the primary idea behind GBM is to gather prediction power from many small models. However, unlike the RF model, these trees are built sequentially rather than in simultaneously. To compensate for the shortcomings of the previous tree, a new tree is constructed (Mishra & Datta-Gupta, 2018). To put it another way, if the training data is fitted poorly for certain predictor values, the following tree will place a more focus on observations in that problematic region, ensuring that the predictions are more accurate. The ultimate model may be thought of as a thousand-term linear regression model, with each term being a regression tree. When the outputs of many weak models are coupled to produce a more accurate prediction, this process is normally referred to as boosting (Hastie et al., 2008) as cited in (Mishra & Datta-Gupta, 2018).

Starting with a base model (i.e., tree), the general GBM process introduces a correction term (i.e., new model) to compensate for residuals of the prior tree, as indicated by negative gradients of a squared-error loss function. The caveat for GBM models is that it models the noise and overfit when the sequential fitting process is repeated multiple times. This problem can be addressed in a number of ways (Mishra & Datta-Gupta, 2018):

- Applying a fractional multiplier or learning rate to the correction term so that the updated model improves the fit more slowly.
- Putting limitations on the fitting parameters, such as the maximum number of iterations.
- Instead of utilizing the whole dataset, employing a bootstrap sample of the data at each iteration.

Chapter 5 Results and Discussion

The major purpose of this chapter is to discuss the results obtained for recognition of the main controlling parameters of CO_2 sequestration in depleted unconventional shale reservoirs made by the use of statistical modeling and data-analytics approach. This analysis involved the following cases:

- Descriptive statistics to understand and describe the data
- Perform a visual analysis through histograms and box plots
- Achieve a visual analysis through scatterplots and scatterplots combined with marginal histograms
- Quantify correlation between the volume of CO_2 sequestered and each input variable
- Perform a supervised learning approach such as OLS and tree-based methods to predict the volume of CO_2 sequestered
- Examine the main drivers of CO_2 sequestration performance in unconventional reservoirs

5.1 Descriptive statistics

When studying datasets, you should first gain a sense of the dataset at hand by asking questions like these (Holdaway, 2009):

- Which values are the smallest and largest?
- For this set of data, what would be a suitable single representative number?
- How wide is the variance or spread?
- Is the dataset distributed evenly throughout a range of values or they are clustered around one or more values?

These questions can be answered through descriptive statistics or summary statistics for the reason that they describe the data. In this study descriptive statistics was performed for both the reservoir and operational parameters.

5.1.1 Reservoir parameters

Table 5.1 shows the summary statistics for the reservoir parameters, which contain the minimum value, maximum value, mean ($\bar{x}$), standard deviation (σ), variance (σ^2) and the skewness which were obtained from this study.

Initially, note that in Table 5.1 the mean value for fracture porosity is indeed lower than the matrix porosity (1.29% vs 7.56%), which is sensible and accurate since this is a dual porosity model which implies that there are two distinct porous media interacting in which the matrix blocks have high storativity (the fluids are mainly contained in the matrix blocks). The amount of fluids contained in the fracture is considerably negligible which entails that the fracture system has low storativity.

Furthermore, note that in Table 5.1 the mean value for fracture permeability is much higher than matrix permeability (0.00062 md vs 0.0000495 md) this is because the fractures are highly conductive and provide the total mobility (fracture openings are large than matrix pore throat dimensions). The matrix blocks supply the storage capacity, so the permeability of the fracture is high compared to the matrix.

Additionally, it can be seen in Table 5.1 that the standard deviation is larger for matrix porosity compared to fracture porosity which implies that the values of matrix porosity are more spread out compared to fracture porosity. Similarly, the standard deviation of fracture permeability is higher than matrix permeability which indicates that the values of fracture permeability are more spread out compared to matrix permeability. Finally, the overall skewness of reservoir parameters is approximately symmetric due to low values of skewness except for Langmuir Volume CO_2 which depicts moderately skewed behavior compared to other parameters. Hence, reservoir parameters seem to follow a normal distribution.

Table 5.1 Descriptive statistics for reservoir parameters

Parameter	***Min***	***Max***	$\bar{x}$	σ	σ^2	***skewness***	***unit***
Thickness (h)	100	300	204	57	3238	-0.08	ft
Matrix Porosity (ϕ_m)	5.001	9.99	7.56	1.44	2.08	-0.05	%
Fracture Porosity (ϕ_f)	0.5007	2	1.29	0.42	0.18	-0.11	%
Water Saturation in Matrix (S_{wm})	5.002	14	9.52	2.61	6.83	-0.004	%
Matrix Permeability (k_m)	1.02E-06	1.01E-04	4.95E-05	2.91E-05	8.47E-10	0.05	md
Fracture Permeability (k_f)	1.02E-04	1.10E-03	6.20E-04	2.79E-04	7.76E-08	-0.04	md
Fracture Spacing (Δx_s)	0.901	3	1.98	0.607	0.37	-0.06	ft
Initial Pressure (P_i)	3001	8000	5420	1436	2063331	0.06	psi
Initial Temperature (T_i)	120	200	160	23	527	-0.007	F
Langmuir Volume CH_4 ($V_{L\text{-}CH4}$)	50	250	148	58	3324	0.05	scf/ton
Langmuir Pressure CH_4 ($P_{L\text{-}CH4}$)	200	1000	596	232	53607	0.03	psi
Langmuir Volume CO_2 ($V_{L\text{-}CO2}$)	109	1486	586	293	85730	0.67	scf/ton
Langmuir Pressure CO_2 ($P_{L\text{-}CO2}$)	200	1000	610	231	53362	-0.06	psi

5.1.2 Operational parameters

Table 5.2 shows the descriptive statistics for the operational parameters which includes the minimum value, maximum value, mean ($\bar{x}$), standard deviation (σ), variance (σ^2) and the skewness.

First, note that in Table 5.2 the mean value of *SRV*-k_f is higher than the mean value of fracture permeability in Table 5.1 (0.00439 md vs 0.00062 md) this is sensible because of the formation being hydraulically fractured, a larger permeability in the *SRV*-zone should be expected. It is seen from Table 5.2 that the maximum values of *SRV*-k_f and *SRV*-ϕ_f are much higher than the maximum values for the fracture permeability and fracture porosity in Table 5.1, also this is due to the fact that these high values demonstrate that for formations which are hydraulically fractured fracture permeability and porosity would be normally high.

Additionally, in Table 5.2 the standard deviation of the length of the reservoir in x direction (L_x) is larger compared to length of the reservoir in y direction (L_y), which illustrates that the values for L_x are more spread out than those for L_y. Likewise, note that the *SRV*-k_f is moderately positively skewed compared to *SRV*-Δx_s which is approximately symmetric (0.77 vs 0.38). Finally, overall, the operational parameters seem to have high skewness values compared to reservoir parameters (except for Langmuir Volume of CO_2). For the reason that operational parameters seem to deviate from a normal distribution and display to an extent lognormal distribution due to being right skewed. Together with the standard deviation, operational parameters have larger values, which indicate that they are more spread out than reservoir parameters.

Table 5.2 Descriptive statistics for operational parameters

Parameter	***Min***	***Max***	$\bar{x}$	σ	σ^2	***Skewness***	***unit***
Horizontal Wellbore Length (L_{hw})	2001	4999	3590	847	717335	-0.11	ft
Hydraulic Fracture Length (L_f)	201	1000	629	226	50944	-0.12	ft
Length of Reservoir (L_x)	2457	7963	5065	1271	1614276	0.05	ft
Length of Reservoir (L_y)	260	1989	1025	400	159795	0.12	ft
SRV Fracture Porosity (SRV-ϕ_f)	0.615	2.97	1.75	0.58	0.34	-0.05	%
SRV Fracture Permeability (SRV-k_f)	3.11E-04	1.31E-02	4.39E-03	2.70E-03	7.29E-06	0.77	md
SRV Fracture Spacing (SRV-Δx_s)	0.366	2.39	1.19	0.44	0.19	0.38	ft
Total Production Time (t_{prod})	7300	18242	12748	3159	9979081	0.01	days
Fracture Pressure (P_{frac})	3164	11679	6949	1918	3679291	0.17	psi

5.2 Univariate data analysis

For univariate data analysis, the fundamental techniques that were used include graphing the box plots and histograms. The visual analysis of these methods will help us determine if the data has outliers, analyze its symmetry and the degree of skewness.

5.2.1 Box plots for reservoir parameters

In Figure 5.1, it can be observed that the median is located at the center of the interquartile range (IQR) for the reservoir parameters box plots. This implies that the sample values are equally packed between the median and the IQR, which again signifies that the reservoir parameters sample values are evenly distributed on both sides of the median. Furthermore, there are no outliers depicted since there are no points which are 1.5 IQR above the third quartile or higher than 1.5 IQR lower than the first quartile.

However, in Figure 5.2 the Langmuir Volume CO_2 box plot shows the median is at a lower position from the top half of the box plot (third quartile). The upper whisker is longer than the lower one, this implies that the data has a longer upper tail than the lower tail. The Langmuir Volume CO_2 values are pulling the box plot upward. As a result, there is more variability of the Langmuir Volume CO_2 box plot. Moreover, the box plot depicts outliers since there are points which are 1.5 IQR above the third quartile. But these outliers can be due to the lognormality behavior of Langmuir Volume CO_2.

Finally, reservoir parameters box plots for Langmuir isotherms (Figure 5.2) do not display evenly distributed sample values as compared to reservoir parameters in Figure 5.1, as they have more variability because of the slight skewness that they possess.

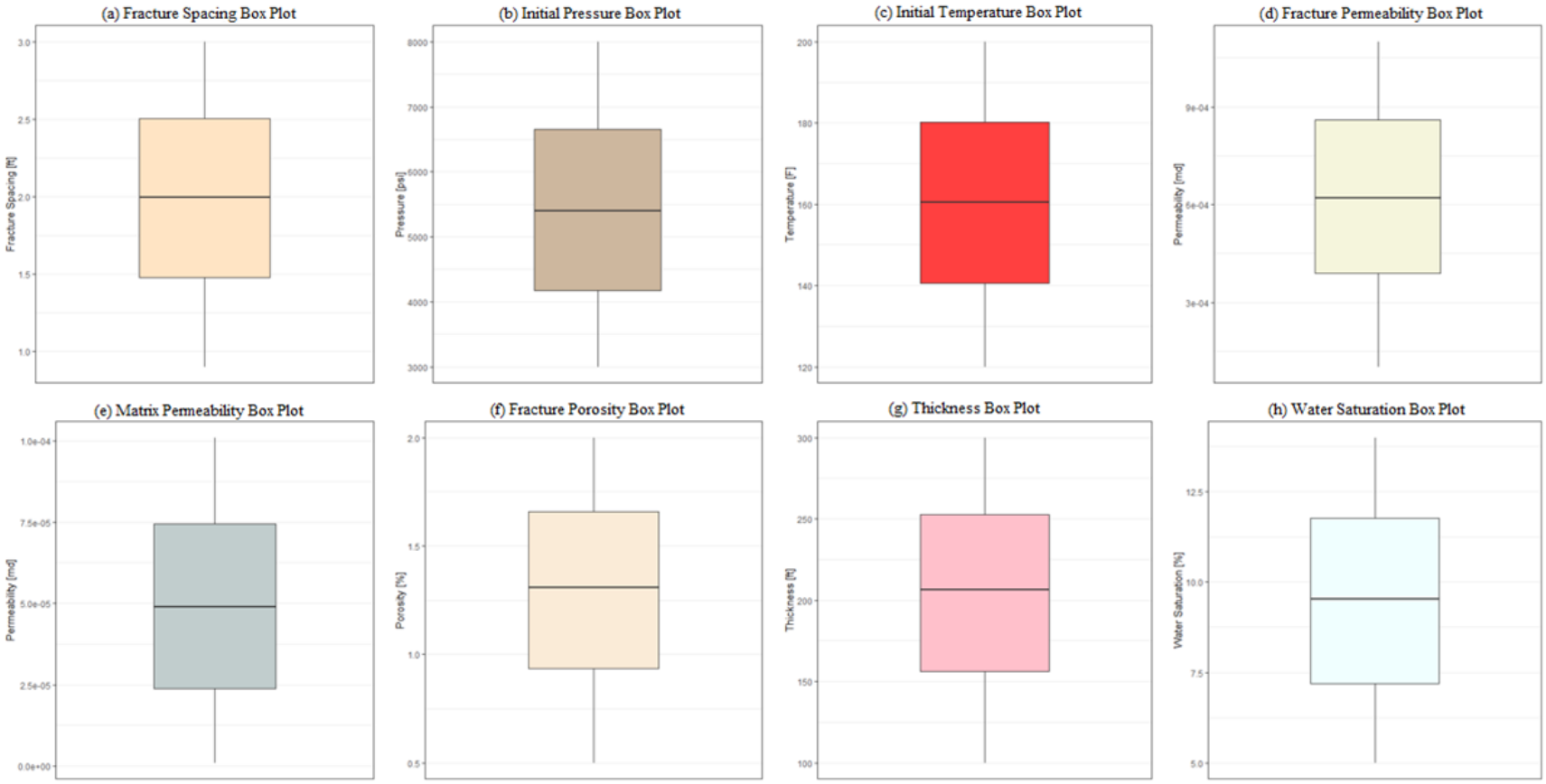

Figure 5.1 Reservoir parameters box plots: a) Fracture spacing, b) Initial pressure, c) Initial temperature, d) Fracture permeability, e) Matrix permeability, f) Fracture porosity, g) Thickness, h) Water saturation

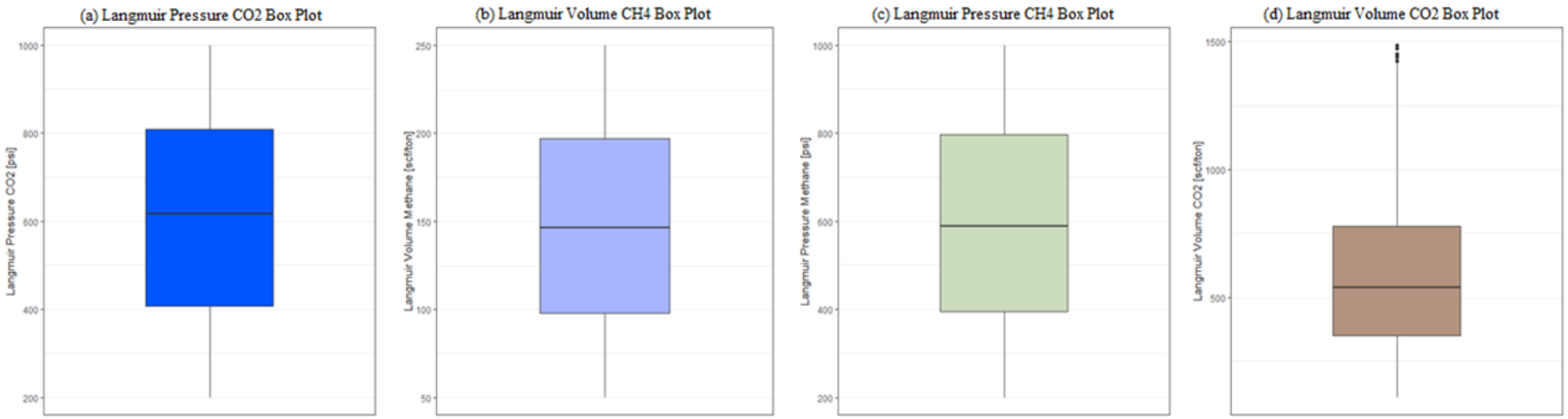

Figure 5.2 Reservoir parameters box plots for Langmuir isotherms: a) Langmuir pressure CO_2, b) Langmuir volume CH_4, c) Langmuir pressure CH_4, d) Langmuir volume CO_2

5.2.2 Box plots for operational parameters

It can be observed from Figure 5.3 that the box plots of *SRV*_xs and *SRV*_kf indicate the median is located at a lower position from the top half of the box plot (third quartile). The upper whisker is longer than the lower one for both box plots, implying that the sample data has an elongated upper tail than the lower tail. As a result, these variables are higher since they are pulling the upper part of the box, which shows more variability as well. The other operational parameters in Figure 5.3 seem to display an even distribution on both sides of the median.

Moreover, it can be noted from Figure 5.3 that *SRV*_kf box plot depicts outliers since there are points which are 1.5 IQR above the third quartile. Nevertheless, these outliers can be because of the lognormality behavior of *SRV*_kf, as shown previously in descriptive statistics, that this parameter displays a moderately positively skewed nature which causes the lognormal behavior.

Finally, note that the reservoir parameters in Figure 5.1 do not display any form of variability and outliers. The reservoir parameters are evenly distributed (except for Langmuir Volume CO_2). Nonetheless, for operational parameters (Figure 5.3) it can be observed that they display slight variability, and some parameters depict outlier points. This difference in variability between reservoir parameters and operational parameters can be due to the higher standard deviation values shown by operational parameters as compared to reservoir parameters and this standard deviation results in values being more spread out and hence the variability in the operational parameters.

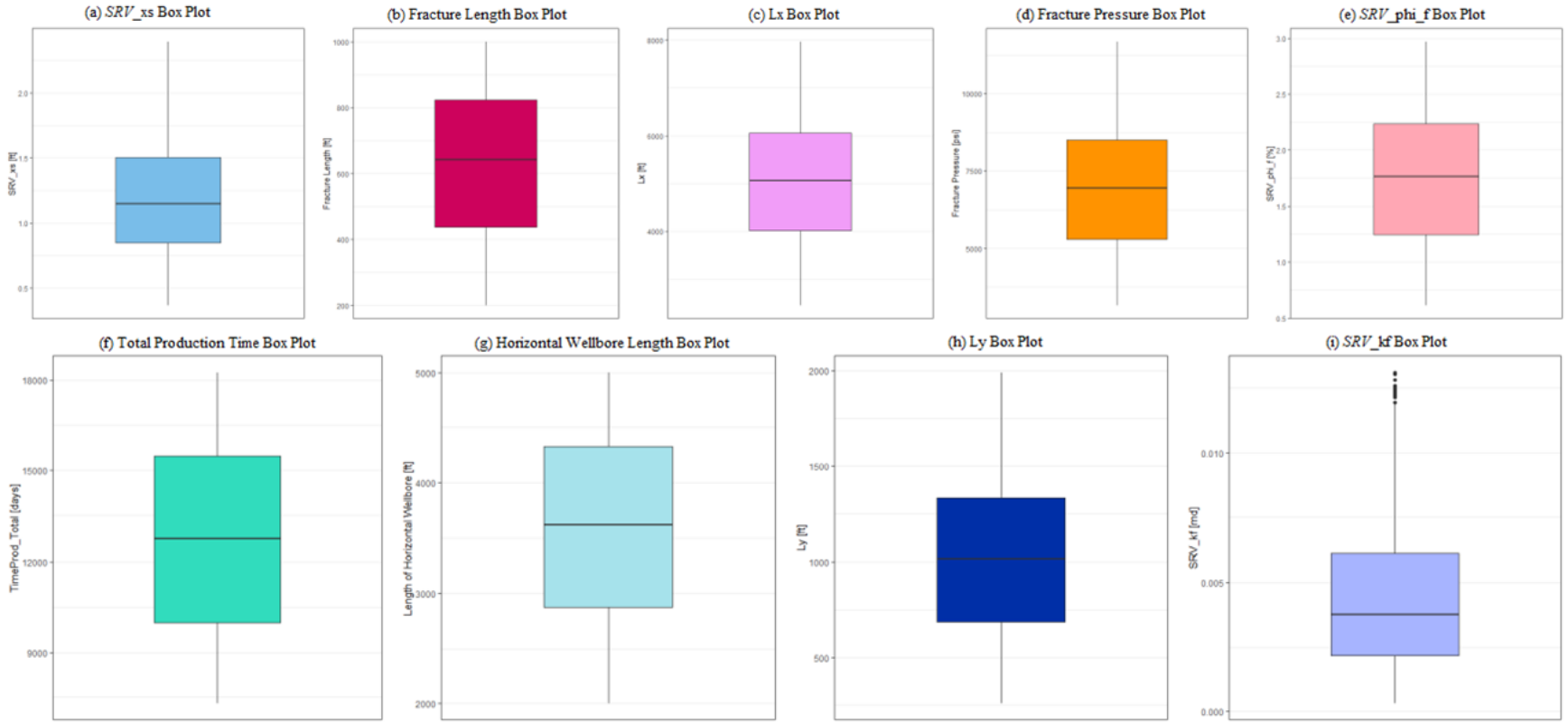

Figure 5.3 Operational parameters box plots: a) SRV_xs, b) Fracture length, c) Lx d) Fracture Pressure, e) SRV_phi_f, f) Total production time, g) Horizontal wellbore length, h) Ly, i) SRV_kf

5.2.3 Box plot for performance metric

In this study, the main performance metric was the cumulative CO_2 injected. This variable was the primary response variable analyzed in this study. It quantifies the volume of CO_2 sequestered.

It is seen from Figure 5.4 that the box plot shows the median is lower. This box plot clearly indicates the effect of variability and the upper half of the IQR is more stretched out because the values are higher at the upper end of the distribution. It can be realized also that the box plot shows outlier points, however these points can be because this box plot is highly positively (right) skewed. This skewness causes the lognormal behavior which displays the points that are 1.5 IQR above the third quartile. Hence, the performance metric (response) variable displays more variability than the input variables, which include both the reservoir and operational parameters. This variability can be because of the spread in values for the cumulative CO_2 injected.

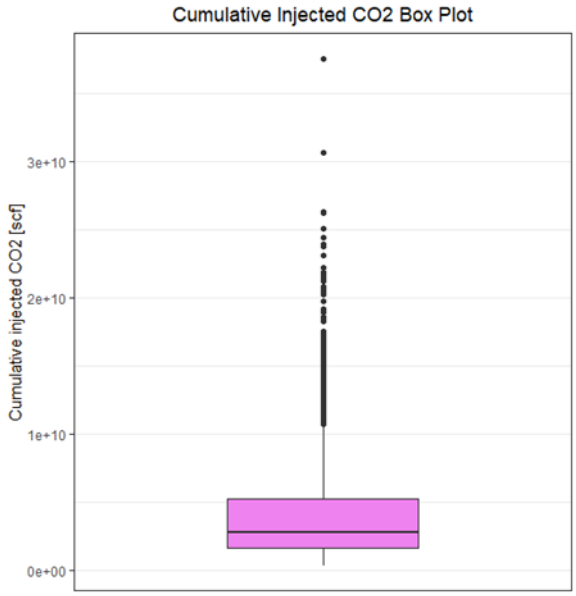

Figure 5.4 Cumulative CO_2 injected box plot

5.2.4 Histograms for reservoir parameters

First, note that in Figure 5.5, the reservoir parameters histograms display a nearly symmetric shape. This was seen in the descriptive statistics summary when the values of the skewness of reservoir parameters were close to zero. Furthermore, Figure 5.6 for Langmuir volume CO_2 histogram displays a right skewed behavior. This was also clear in the descriptive statistics summary there was a moderately positive value for the skewness. The visual analysis of the histogram confirms this behavior and verifies the variability observed in the previous box plot, and hence, the lognormal pattern in the histogram is clearly visible.

5.2.5 Histograms for operational parameters

In Figure 5.7, it can be observed that most of the operational parameters are almost symmetric in terms of the shape of the histogram. Almost all of them do not exhibit any degree of skewness except for *SRV*_kf and *SRV*_xs. These histograms clearly depict that most of the sample values are at the left and the right side of the tail is longer, hence this is a right skewed histogram and lognormal behavior. Moreover, the two parameters stimulated reservoir volume fracture permeability and fracture spacing (*SRV*_kf and *SRV*_xs) seem to display a similar pattern. The similarity in the pattern can be since these two parameters are essential to describe the hydraulic fractures for the *SRV* zone.

Finally, between reservoir and operational parameters, it can be clearly observed that reservoir parameters do not display any degree of skewness for their histograms except for Langmuir volume CO_2. Whereas, for operational parameters, *SRV*_kf and *SRV*_xs display moderate positive skewness and the other operational parameters are almost symmetric. Therefore, overall, the operational parameters have more degree of skewness and variability as compared to reservoir parameters.

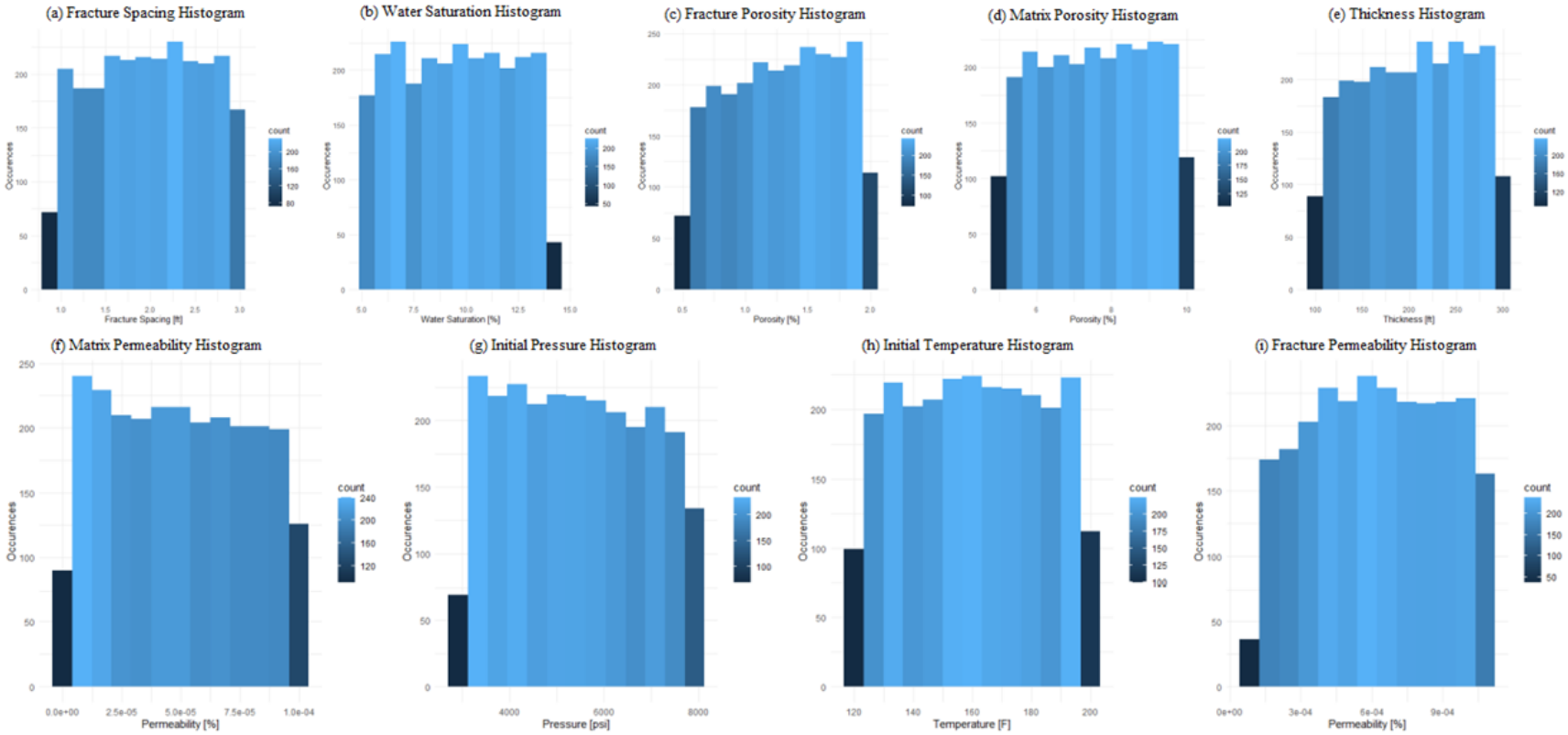

Figure 5.5 Reservoir parameters histograms: a) Fracture spacing, b) Water saturation, c) Fracture porosity, d) Matrix porosity, e) Thickness, f) Matrix permeability, g) Initial pressure, h) Initial temperature, i) Fracture permeability

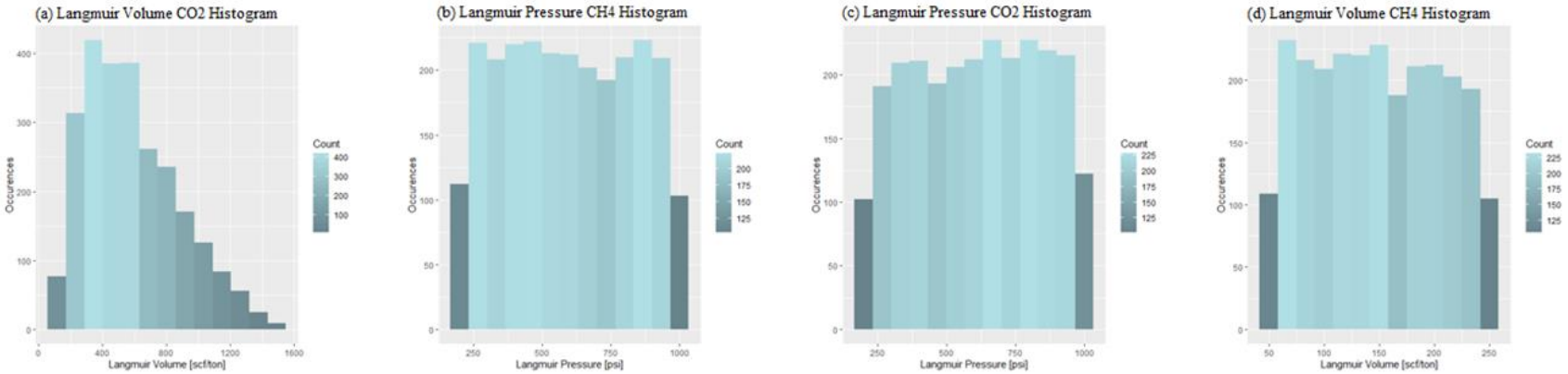

Figure 5.6 Reservoir parameters histograms for Langmuir isotherms: a) Langmuir volume CO_2, b) Langmuir pressure CH_4, c) Langmuir pressure CO_2, d) Langmuir volume CH_4

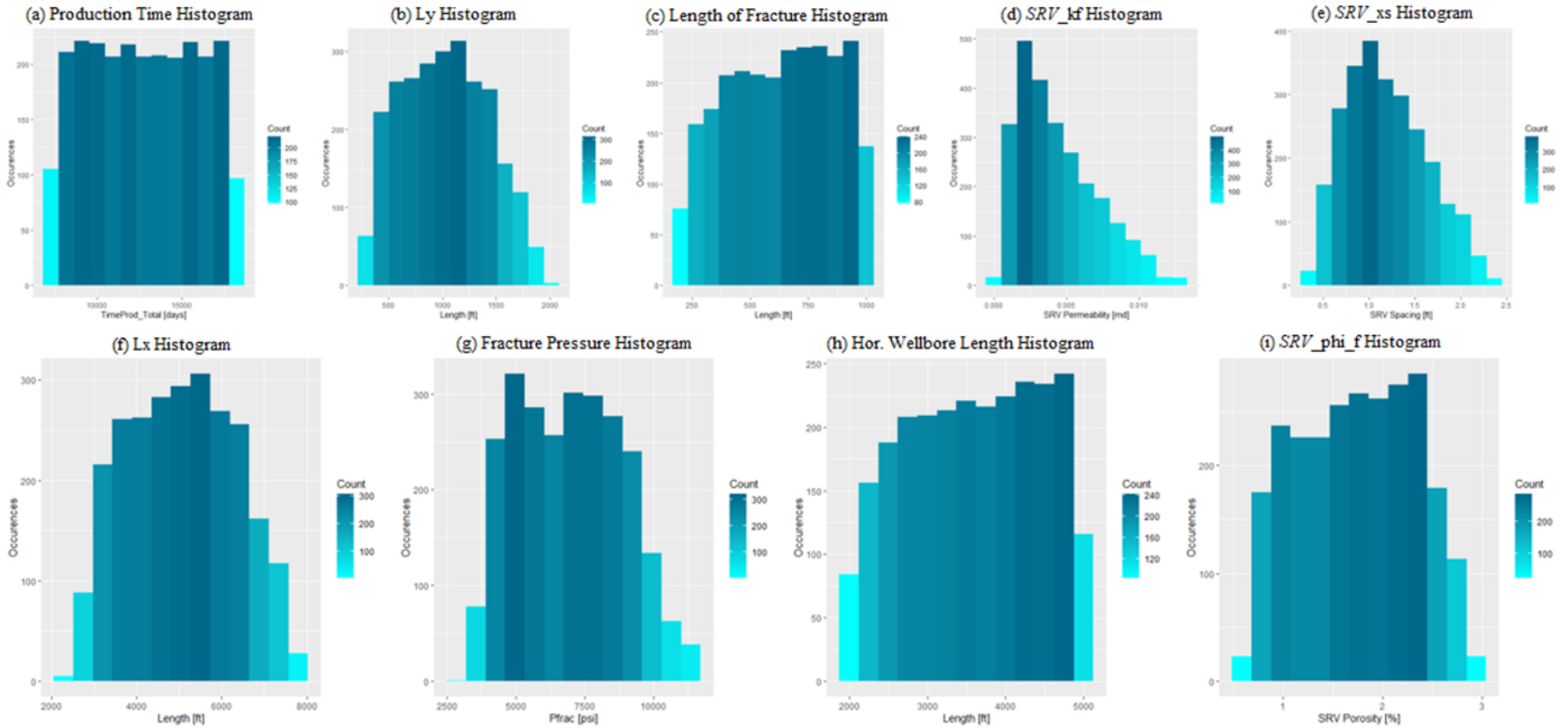

Figure 5.7 Operational parameters histograms: a) Total production time, b) Ly, c) Length of fracture, d) SRV_kf, e) SRV_xs, f) Lx, g) Fracture pressure, h) Hor. wellbore length, i) SRV_phi_f

5.2.6 Histogram for performance metric

It is seen from Figure 5.8 that the histogram is not symmetric. A histogram in which the tail on the right-hand side is long is said to be positively skewed (skewed to the right). This histogram shows that the mean value is higher than the median. Moreover, the sample data points appear to be more concentrated to the left as displayed and it is unimodal since it has only one peak. This histogram clearly depicts a lognormal behavior, as seen from previous visual analysis of reservoir parameters and operational parameters. As a result, there might be a relationship and dependency between some of the reservoir parameters and operational parameters with the cumulative CO_2 injected. Hence, the next part will be to observe this relationship and visualize the results obtained from bivariate data analysis.

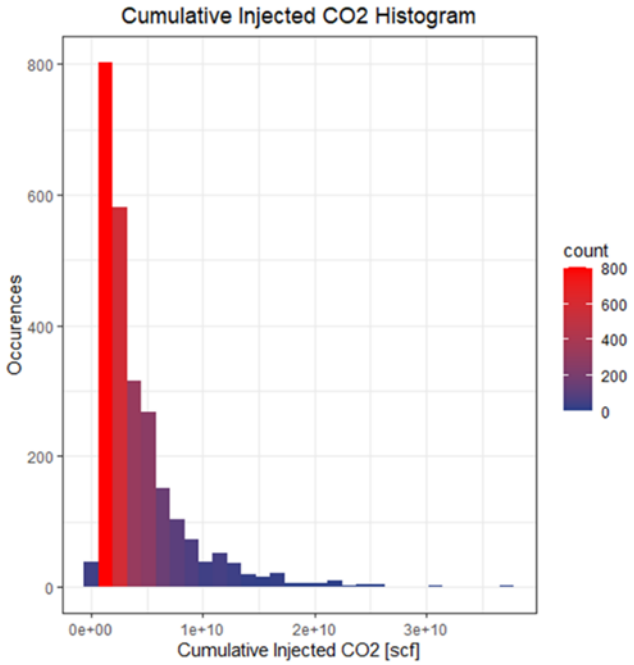

Figure 5.8 Cumulative CO_2 injected histogram

5.3 Bivariate data analysis

For bivariate data analysis, the most significant methods used were graphing the scatterplots along with scatterplots with marginal histograms to visualize the relationship between the volume of CO_2 sequestered (Cumulative CO_2 injected) and each input variable (reservoir and operational parameters). Later, the relationship between these two variables was quantified by examining the correlation between the two variables (volume of CO_2 sequestered and each input variable (predictors)).

5.3.1 Reservoir parameters scatterplots and marginal histograms

It can be realized from Figure 5.9 from the shape and pattern of the data points, there is a positive linear relationship between fracture permeability and cumulative injected CO_2 (Figure 5.9) and there is a positive linear relationship between thickness and cumulative injected CO_2 (Figure 5.9). These two scatterplots display a modest relationship. The other reservoir parameters show a nonmonotonic relationship with the cumulative injected CO_2.

It can be observed from Figure 5.11 that the Langmuir isotherms do not seem to display any visual evidence of a relationship with the cumulative injected CO_2. Furthermore, in Figure 5.10, it can be observed that all the reservoir parameters appear to have a symmetric distribution from their marginal histograms, while the cumulative injected CO_2 displays the same right skewed behavior. This visual analysis just confirms the previous analysis of histograms, but now it is clearer when it is visualized together with the performance metric.

Finally, note that in Figure 5.12 all Langmuir isotherms display a symmetric distribution except for Langmuir volume CO_2, it displays a right skewed behavior, like the cumulative injected CO_2. However, the relationship is weak and there is no obvious pattern between the two variables.

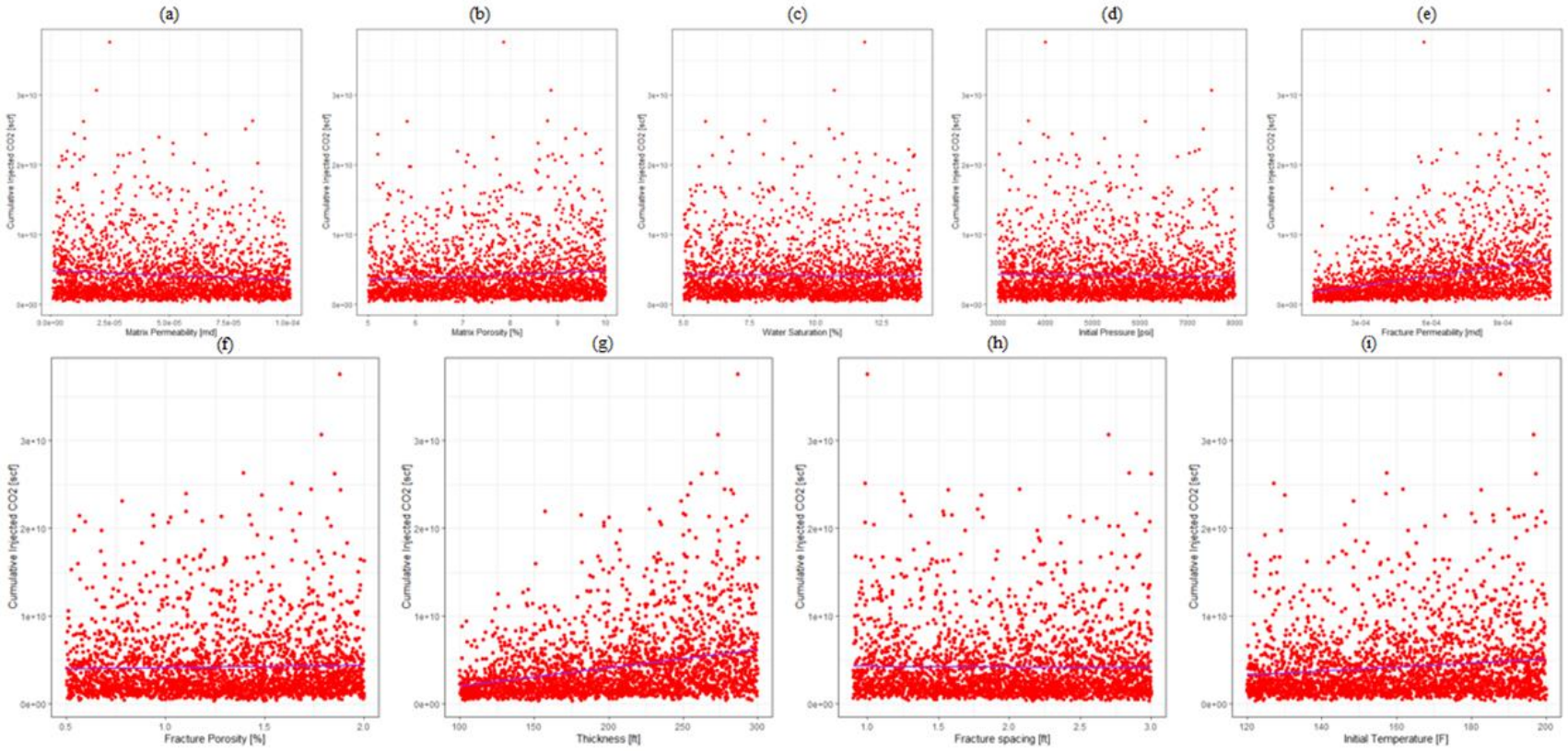

Figure 5.9 Reservoir parameters scatterplots: a) Matrix permeability, b) Matrix porosity, c) Water saturation, d) Initial pressure, e) Fracture permeability, f) Fracture porosity, g) Thickness, h) Fracture spacing, i) Initial temperature

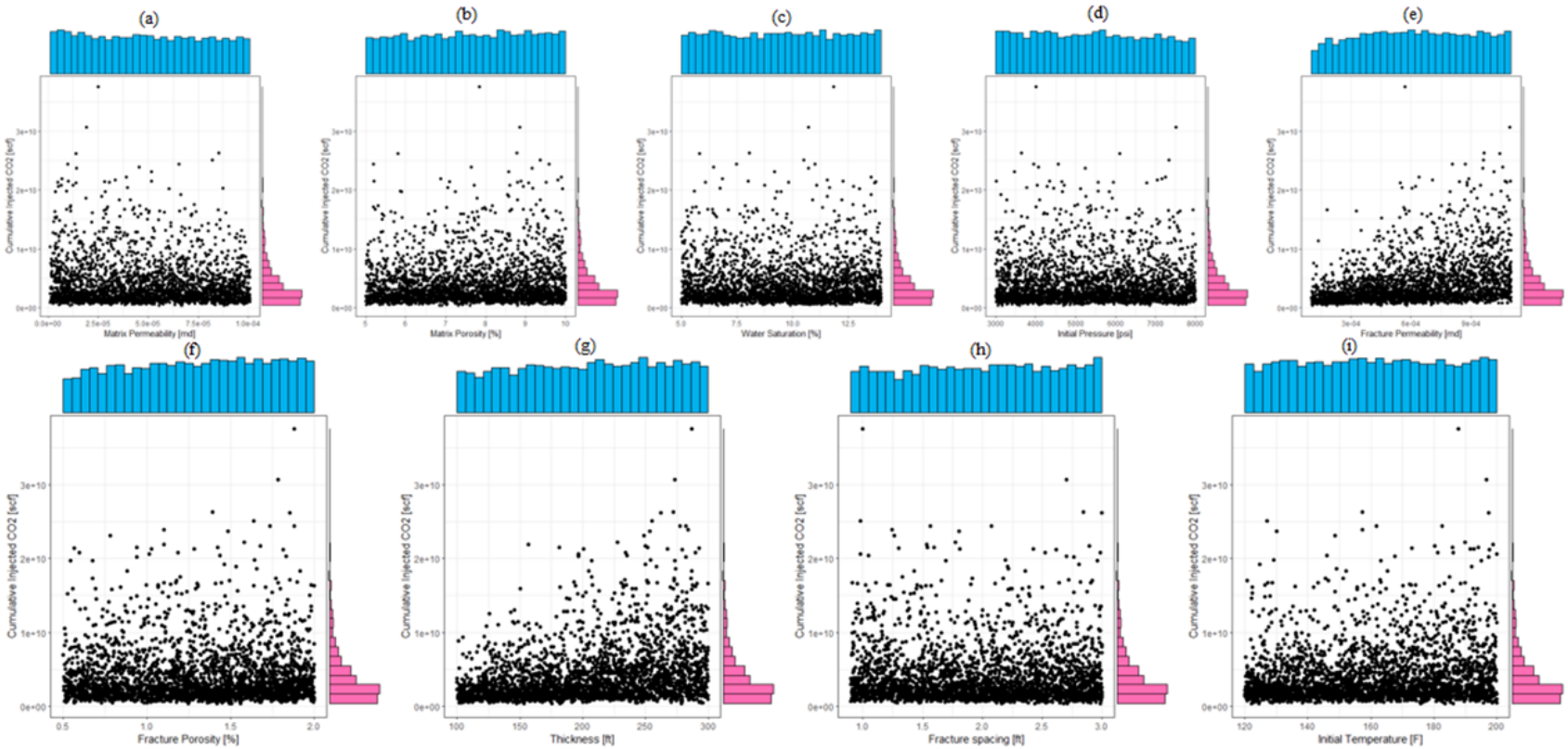

Figure 5.10 Reservoir parameters scatterplots with marginal histograms: a) Matrix permeability, b) Matrix porosity, c) Water saturation, d) Initial pressure, e) Fracture permeability, f) Fracture porosity, g) Thickness, h) Fracture spacing, i) Initial temperature

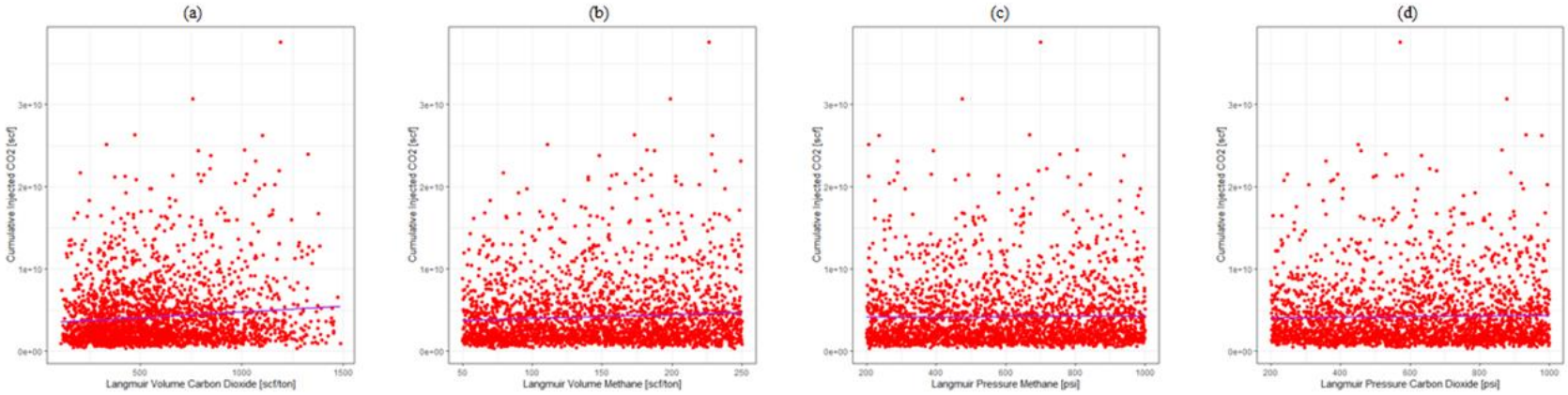

Figure 5.11 Reservoir parameters scatterplots for Langmuir isotherms: a) Langmuir volume CO_2, b) Langmuir volume CH_4, c) Langmuir pressure CH_4, d) Langmuir pressure CO_2

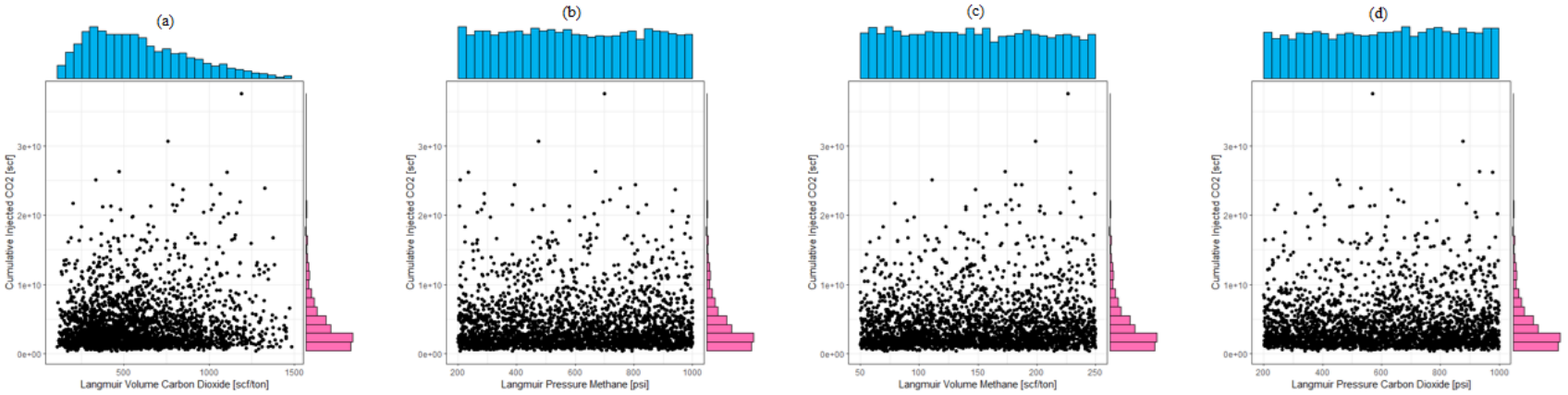

Figure 5.12 Reservoir parameters scatterplots with marginal histograms for Langmuir isotherms: a) Langmuir volume CO_2, b) Langmuir pressure CH_4, c) Langmuir volume CH_4, d) Langmuir pressure CO_2

5.3.2 Operational parameters scatterplots and marginal histograms

First, note that in Figure 5.13 from the shape and pattern of the data points there is a positive linear relationship between:

- Stimulated reservoir volume fracture permeability (*SRV*_kf) and cumulative injected CO_2
- Horizontal wellbore length (L_{hw}) and cumulative injected CO_2
- Length of reservoir in x direction (edge_x) and cumulative injected CO_2

The other operational parameters seem to display a nonmonotonic relationship with the cumulative injected CO_2. Moreover, in Figure 5.14, both marginal histograms show a moderate positively skewed pattern. This behavior confirms the previous analysis made that both stimulated reservoir volume fracture permeability (*SRV*_kf) and cumulative injected CO_2 are positively skewed. Furthermore, the presence of the previous outlier points can now be clearly explained that the behavior was mainly because of the dependency between these two variables, and this causes additional lognormality between the two parameters. Whereas the other operational parameters in Figure 5.14 show an approximately symmetric distribution.

Finally, it can be observed that, between reservoir parameters and operational parameters, operational parameters seem to display a more significance to the performance metric as compared to reservoir parameters, since more operational parameters show a monotonic relationship with the performance metric. However, the strength of this association will be elaborated more by quantifying the correlation.

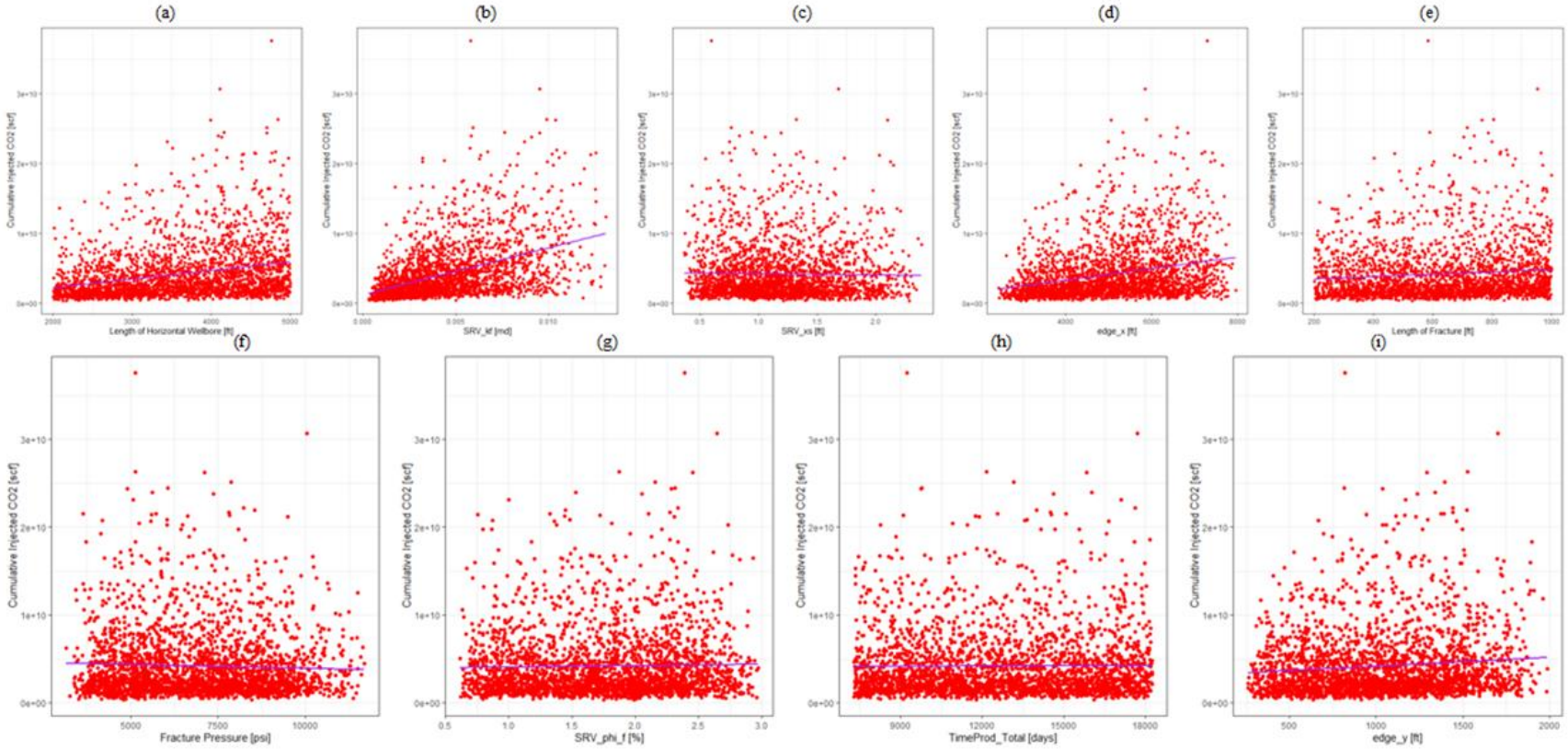

Figure 5.13 Operational parameters scatterplots: a) Hor. wellbore length, b) SRV_kf, c) SRV_xs, d) edge_x, e) Length of fracture, f) Fracture pressure, g) SRV_phi_f, h) Total production time, i) edge_y

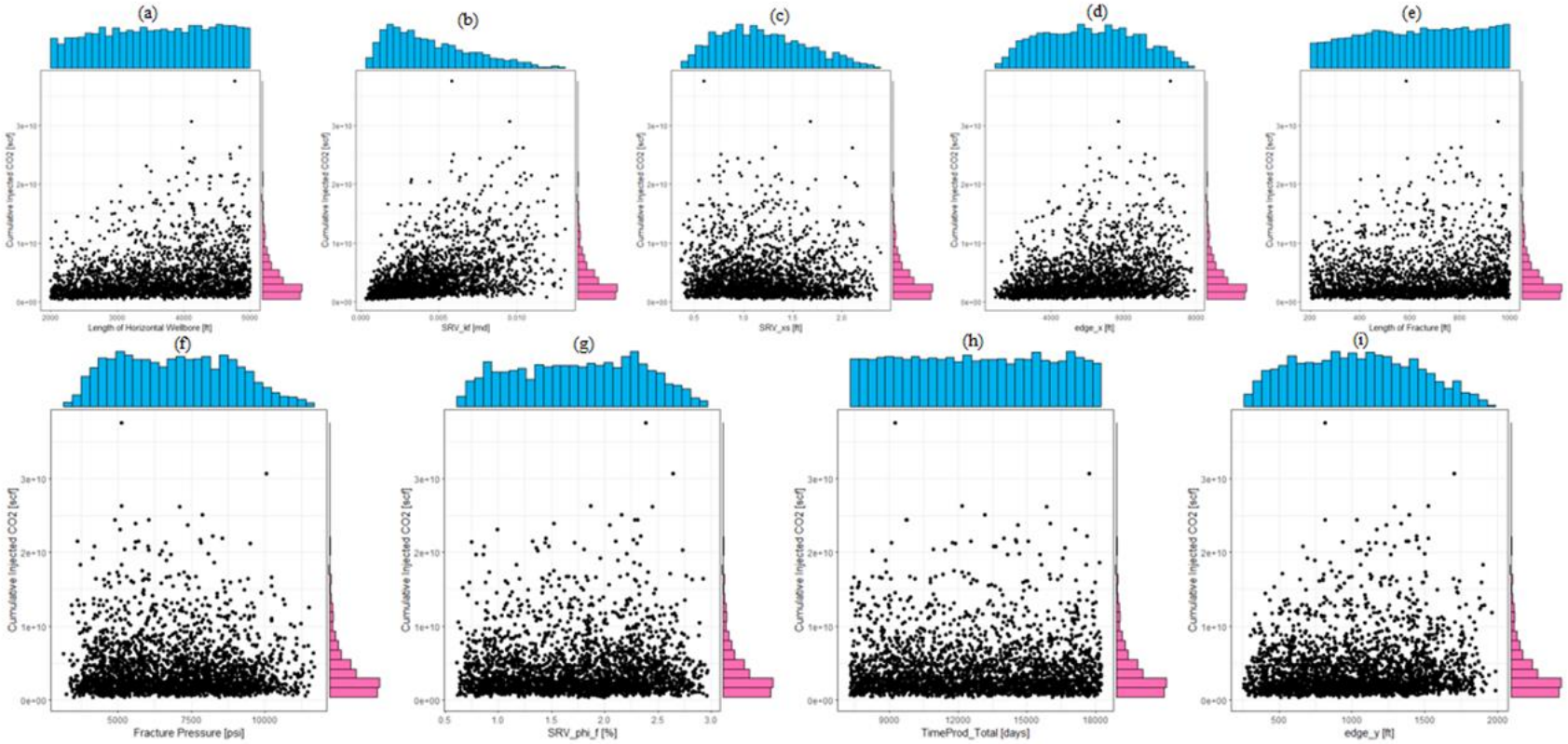

Figure 5.14 Operational parameters scatterplots with marginal histograms: a) Hor. wellbore length, b) SRV_kf, c) SRV_xs, d) edge_x, e) Length of fracture, f) Fracture pressure, g) SRV_phi_f, h) Total production time, i) edge_y

5.3.3 Correlation test

For the correlation test, two types of correlation coefficients were used, which are the Pearson correlation coefficient and the Spearman correlation coefficient. As Mishra & Datta-Gupta (2018) mention that, Spearman correlation coefficient is more robust and considers nonlinear association, whereas Pearson can be sensitive to data outliers and clusters, hence it is better to compute both measures.

It is seen from Table 5.3 that there is a modest positive correlation between thickness and cumulative injected CO_2 with a Pearson value of 0.303 and Spearman value of 0.333. Moreover, there is a modest positive correlation between fracture permeability and cumulative injected CO_2 with a Pearson value of 0.341 and Spearman value of 0.394. Also, there is a weak correlation between initial temperature and cumulative injected CO_2.

Table 5.3 Correlation between reservoir parameters and cumulative injected CO_2

Parameter	***Pearson's***	***Spearman***
Thickness (h)	0.303	0.333
Matrix Porosity (ϕ_m)	0.092	0.093
Fracture Porosity (ϕ_f)	0.030	0.031
Water Saturation in Matrix (S_{wm})	-0.011	-0.010
Matrix Permeability (k_m)	-0.095	-0.066
Fracture Permeability (k_f)	0.341	0.394
Fracture Spacing (Δx_s)	-0.013	-0.006
Initial Pressure (P_i)	-0.021	-0.008
Initial Temperature (T_i)	0.132	0.134
Langmuir Volume CH_4 ($V_{L\text{-}CH4}$)	0.072	0.058
Langmuir Pressure CH_4 ($P_{L\text{-}CH4}$)	0.012	0.021
Langmuir Volume CO_2 ($V_{L\text{-}CO2}$)	0.104	0.087
Langmuir Pressure CO_2 ($P_{L\text{-}CO2}$)	0.021	0

Furthermore, in Table 5.3, the Langmuir isotherms do not seem to show any modest correlation with the cumulative injected CO_2. However, this does not mean that there are not significant parameters in explaining the behavior of CO_2 sequestration in unconventional reservoirs they simply do not display a monotonic relationship with the performance metric, hence their relationship might be nonlinear or quadratic type. In addition, the significance of the variables

thickness and fracture permeability to the performance metric will be explained in more details in the variable importance part of the results section.

It can be observed from Table 5.4 that there is a modest positive correlation between *SRV* fracture permeability and cumulative injected CO_2 with a Pearson value of 0.465 and Spearman value of 0.536. Moreover, there is a modest positive correlation between length of reservoir (L_x) and cumulative injected CO_2 with a Pearson value of 0.270 and Spearman value of 0.307. Additionally, there is a modest positive correlation between horizontal wellbore length and cumulative injected CO_2 with a Pearson value of 0.268 and Spearman value of 0.305. The significance of these variables *SRV* fracture permeability, length of reservoir in x direction and horizontal wellbore length to the cumulative injected CO_2 will be assessed more clearly in the variable importance section when the screening will be performed, which will be explained in this results section.

It is quite clear that a reasonable number of operational parameters display a modest positive correlation with the cumulative injected CO_2. For the reason that, they are vital to describe the *SRV*-zone in which nearly all the injected CO_2 will be reserved in this zone. Finally, the results obtained from Pearson and Spearman correlations for both reservoir and operational parameters are consistent with the visual analysis through cross-plots (scatterplots) made earlier, this reveals the relevance of performing EDA.

Table 5.4 Correlation between operational parameters and cumulative injected CO_2

Parameter	***Pearson's***	***Spearman***
Horizontal Wellbore Length (L_{hw})	0.268	0.305
Hydraulic Fracture Length (L_f)	0.096	0.113
Length of Reservoir (L_x)	0.270	0.307
Length of Reservoir (L_y)	0.103	0.119
SRV Fracture Porosity (*SRV*-ϕ_f)	0.024	0.027
SRV Fracture Permeability (*SRV*-k_f)	0.465	0.536
SRV Fracture Spacing (*SRV*-Δx_s)	-0.020	-0.024
Total Production Time (t_{prod})	0.003	0.002
Fracture Pressure (P_{frac})	-0.043	-0.023

5.4 Multivariate analysis

The last part of EDA in this study comprised presenting a correlation matrix which extends the ideas discussed previously but now it will involve all variable pairs, including reservoir and operational parameters, with the performance metric.

Figure 5.15 presents the correlation matrix for all variable pairs (dependent and independent). One of the distinct features of the correlation matrix is that it is symmetrical. In this correlation matrix, it can be seen there is a dependency between *SRV* fracture permeability (SRV_kf) and fracture permeability (PermF), *SRV* fracture permeability (SRV_kf) and cumulative injected CO_2 (cum_inj). Furthermore, *SRV* fracture spacing (SRV_xs) depends on fracture spacing (xs). SRV fracture porosity (SRV_phi_f) also depends on fracture porosity (PoroF). Likewise, Langmuir volume CH_4 (Vl_ch4) depends on Langmuir volume CO_2 (Vl_co2).

The dependency between independent variables (Predictors), as well as between dependent (Response) and independent variables is the reason for observing the previous outlier points. Because this is a dataset developed from numerical simulation scenarios, the outlier points cannot be because of an incorrect input value into the dataset. Hence, this dependency causes additional lognormality, which is clear in the histograms of these variables and the box plots.

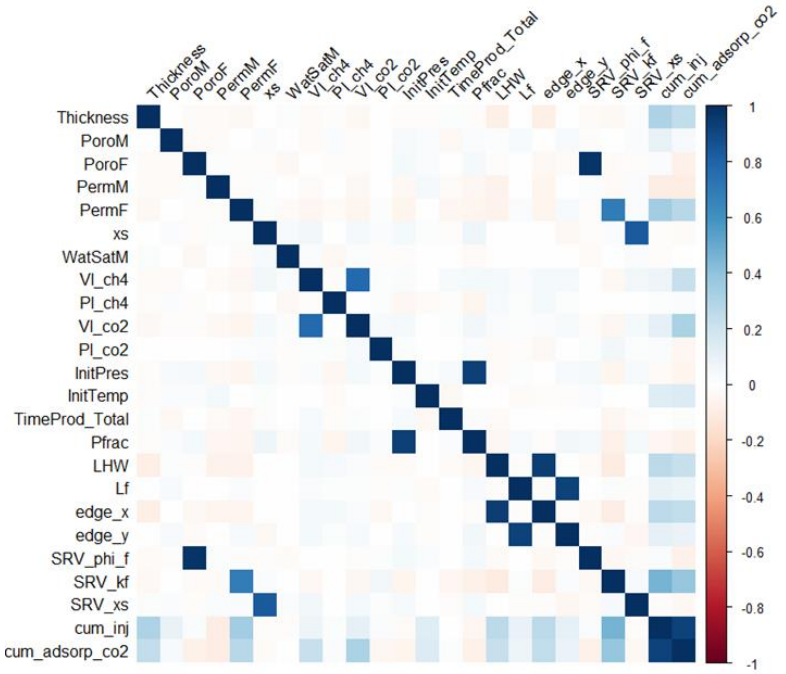

Figure 5.15 Correlation matrix

5.5 Predictive modeling

The EDA method performed in the preceding section is an essential technique used in this study to verify the parameters that have a relationship with the cumulative CO_2 injected along with determining patterns and trends in order to perform predictive modeling. In this study two fundamental techniques were applied to predict cumulative CO_2 injected which are OLS regression and tree-based methods. These predictive models are significant to provide accurate predictions of CO_2 sequestration performance using the dataset available.

5.5.1 Ordinary Least Squares Regression

In this study, since the input variables involved are over one for both reservoir and operational parameters, multiple linear regression will be used, which is also another term for OLS regression. A list of all the variables used in this study is shown in Table 5.5. The response variable (performance metric) was cum_inj, which measures the cumulative CO_2 injected in standard cubic feet (scf). The predictors which contain 22 variables include both the reservoir and operational parameters.

A typical first step in multiple linear regression is to check if at least one of the predictors Thickness, PoroM,..., Pfrac is useful in predicting the response variable (cum_inj). In order to confirm this step, the *F-statistic* was computed by first fitting a multiple linear regression for all the variables. It can be seen in Table 5.6 that the residual standard error is 2.742E+09. This value represents the standard deviation of the residual values in the model. This value shows a high standard deviation, which would imply that the residuals are not following a normal distribution. Moreover, the multiple R-Squared value is 0.5074. This value represents the goodness of fit and the variability explained by the 22-variable model. A value of R^2 corresponding to 0.5074 explains a moderate portion of the variance in the response variable.

It is seen from Table 5.6 that the F-statistic is 118.2. This value provides an appealing sign that at least one of the reservoir or operational parameters must be related to cumulative CO_2 injected. Furthermore, the p-value related to the F-statistic is 2.2E-16, which is approximately zero, hence this is significant evidence that at least one of the reservoir or operational parameters is associated with the cumulative CO_2 injected.

Table 5.5 Variables in the dataset

Description	***Variable***	***Type***
Cumulative CO_2 Injected	cum_inj	Response
Thickness (h)	Thickness	Predictor
Matrix Porosity (ϕ_m)	PoroM	
Fracture Porosity (ϕ_f)	PoroF	
Water Saturation in Matrix (S_{wm})	WatSatM	
Matrix Permeability (k_m)	PermM	
Fracture Permeability (k_f)	PermF	
Fracture Spacing (Δx_s)	xs	
Initial Pressure (P_i)	InitPres	
Initial Temperature (T_i)	InitTemp	
Langmuir Volume CH_4 ($V_{L\text{-}CH4}$)	Vl_ch4	
Langmuir Pressure CH_4 ($P_{L\text{-}CH4}$)	Pl_ch4	
Langmuir Volume CO_2 ($V_{L\text{-}CO2}$)	Vl_co2	
Langmuir Pressure CO_2 ($P_{L\text{-}CO2}$)	Pl_co2	
Horizontal Wellbore Length (L_{hw})	LHW	
Hydraulic Fracture Length (L_f)	Lf	
Length of Reservoir (L_x)	edge_x	
Length of Reservoir (L_y)	edge_y	
SRV Fracture Porosity (*SRV*-ϕ_f)	SRV_phi_f	
SRV Fracture Permeability (*SRV*-k_f)	SRV_kf	
SRV Fracture Spacing (*SRV*-Δx_s)	SRV_xs	
Total Production Time (t_{prod})	TimeProd_Total	
Fracture Pressure (P_{frac})	Pfrac	

Table 5.6 Model summary for 22 variables

Quantity	***Value***
Residual standard error	2.742E+09
Multiple R-squared	0.5074
Adjusted R-squared	0.5031
F-statistic	118.2
p-value	2.2E-16

It is quite clear that the results in Table 5.6 are corresponding to the preceding analysis of EDA that indeed there are parameters which are associated with the response variable. The next part of the analysis was to determine which subset of the predictors is associated with the response variable to fit a single OLS model using those predictors.

It can be seen in (**Appendix A1)** that the asterisk in the findings shows that a certain parameter is included in the model. For example, this report (**Appendix A1**) suggests that Thickness, LHW, and SRV_kf make up the optimal three-variable model. However, we can fit all 22-variable models and choose the best overall model. RSS and R^2 are one of the two metrics that can assess a model that has a low training error. It is observed from (**Appendix A1)** that the R^2 value increases from 22% when only one variable is included in the model, to almost 51% when all variables are included. Furthermore, Figure 5.16 shows, that as the number of variables in the model grows, RSS falls monotonically. These two metrics might not be ideal because a low RSS or high R^2 suggests a model with a low training error, but we want to choose a model with a low-test error, RSS and R^2 are not appropriate for selecting the best model from a group of models (James et al., 2013). As a result, C_p, BIC, or adjusted R^2 can be used to modify the training error to account for overfitting bias. A model with a low value for C_p and BIC is optimal, but a model with a high adjusted R^2 is acceptable.

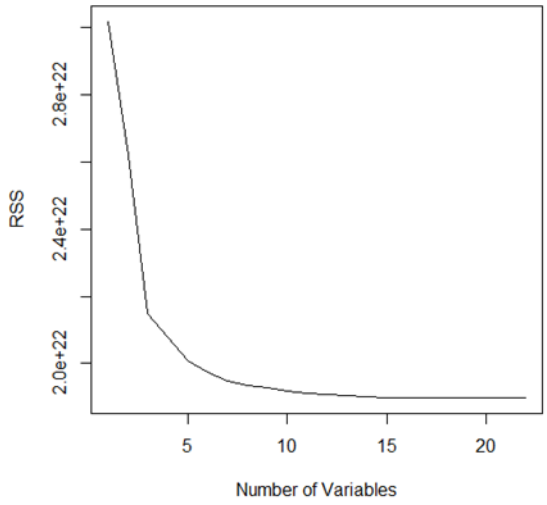

Figure 5.16 RSS plot

It is observed from Figure 5.17 that a 15 variable model would be optimal from the 22-variable model. An adjusted R^2 value of approximately 0.5 would correspond to a 15-variable model.

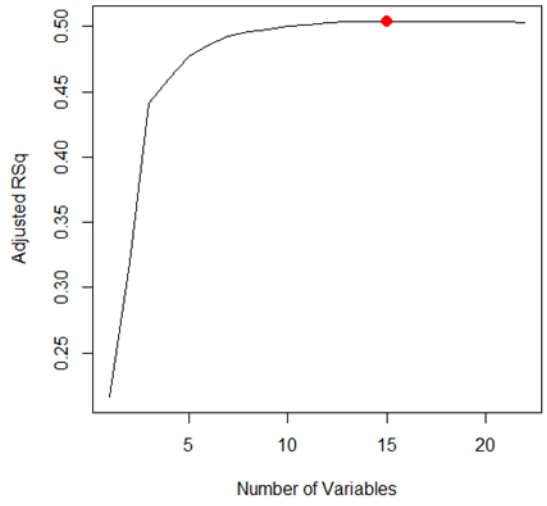

Figure 5.17 Adjusted R^2 plot

Moreover, Figure 5.18 shows the C_p with the number of variables. As explained previously, a low statistic of C_p will correspond to the optimal model. Here, C_p is approximately zero, and

this value gives an optimal model of 14 variables. Together with C_p, in Figure 5.19 a low value of BIC will also correspond to an optimal model, in this case an 11-variable model.

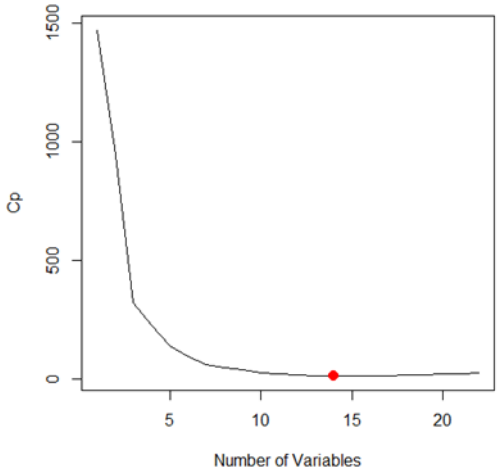

Figure 5.18 C_p plot

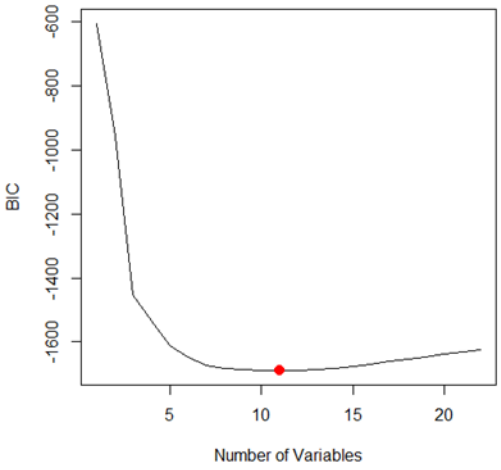

Figure 5.19 BIC plot

Therefore, between the three metrics observed, BIC statistic displayed the smallest value and a reasonable number of variables for the optimal model, which was an 11-variable model. For this 11-variable model the coefficients were estimated. The corresponding equation shows the model.

$$cum_{inj} = a_0 + a_1 Thickness + a_2 PoroM + a_3 PermM + a_4 Vl\_co2 + a_5 InitTemp + a_6 LHW + a_7 Lf + a_8 edge\_x + a_9 SRV\_phi\_f + a_{10} SRV\_kf + a_{11} SRV\_xs \quad \text{Eq 5.1}$$

The regression coefficients are:

$a_0 = -1.74 \times 10^{10}$ $a_4 = 1.76 \times 10^{6}$ $a_8 = 4.36 \times 10^{5}$

$a_1 = 2.47 \times 10^{7}$ $a_5 = 2.37 \times 10^{7}$ $a_9 = 4.11 \times 10^{8}$

$a_2 = 2.23 \times 10^{8}$ $a_6 = 9.89 \times 10^{5}$ $a_{10} = 7.51 \times 10^{11}$

$a_3 = -6.28 \times 10^{12}$ $a_7 = 1.57 \times 10^{6}$ $a_{11} = -3.59 \times 10^{8}$

Hence, this OLS regression model should not be considered as a universal model for oil and gas applications, but its applicability should be in similar circumstances to the ones we have seen in this study. Another approach that can select among a collection of models is a k-fold cross-validation method. The k-fold cross-validation can also estimate the test error or model performance. Because the adjusted R^2, C_p and BIC are computed based on training data, they might be prone to overfitting therefore the k-fold cross-validation represents a better alternative. It can be noted from Figure 5.20 that the k-fold cross-validation selects a 14-variable model based on the mean cv errors. Finally, best subset selection was performed on the full dataset to get the 14-variable model and extract its coefficients (**Appendix A2**).

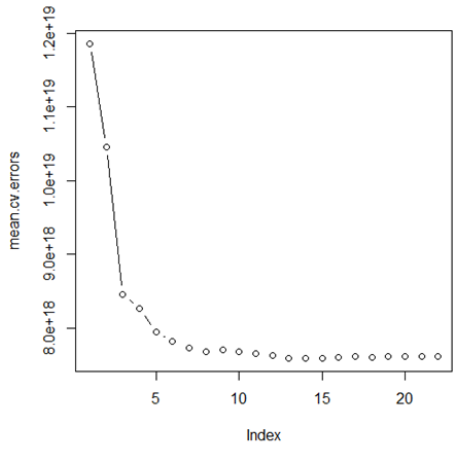

Figure 5.20 k-fold cross-validation plot

$$cum_{inj} = a_0 + a_1 Thickness + a_2 PoroM + a_3 PermM + a_4 PermF + a_5 Vl\_ch4 + a_6 Vl\_co2 + a_7 InitTemp + a_8 TimeProd\_Total + a_9 LHW + a_{10} Lf + a_{11} edge\_x + a_{12} SRV\_phi\_f + a_{13} SRV\_kf + a_{14} SRV\_xs \quad \textbf{Eq 5.2}$$

The regression coefficients are:

$a_0 = -1.8 \times 10^{10}$	$a_5 = -3.85 \times 10^{6}$	$a_{10} = 1.57 \times 10^{6}$
$a_1 = 2.47 \times 10^{7}$	$a_6 = 2.36 \times 10^{6}$	$a_{11} = 4.15 \times 10^{5}$
$a_2 = 2.26 \times 10^{8}$	$a_7 = 2.38 \times 10^{7}$	$a_{12} = 4.16 \times 10^{8}$
$a_3 = -6.09 \times 10^{12}$	$a_8 = 4.2 \times 10^{4}$	$a_{13} = 7.19 \times 10^{11}$
$a_4 = 4.91 \times 10^{11}$	$a_9 = 1.03 \times 10^{6}$	$a_{14} = -3.45 \times 10^{8}$

This 14-variable model obtained from **Eq 5.2** is only valid for the circumstances used in Table 3.1 underlying studies. As a result, this OLS model should not be seen as a general proxy model that can be used to any unconventional reservoir; rather, its application should be confined to the conditions described in this work. Finally, this 14-variable model will build a single predictive model for the full training dataset. At the same time, evaluating the goodness of fit using AAE and MSE.

Figure 5.21 is the predicted versus observed cumulative injected CO_2 (scf) for the multiple linear regression model using 14-variables. The diagonal dashed black line represents the model fit. It can be observed that not all the points lie near the 45-degree line. This shows a moderate fit to the training data.

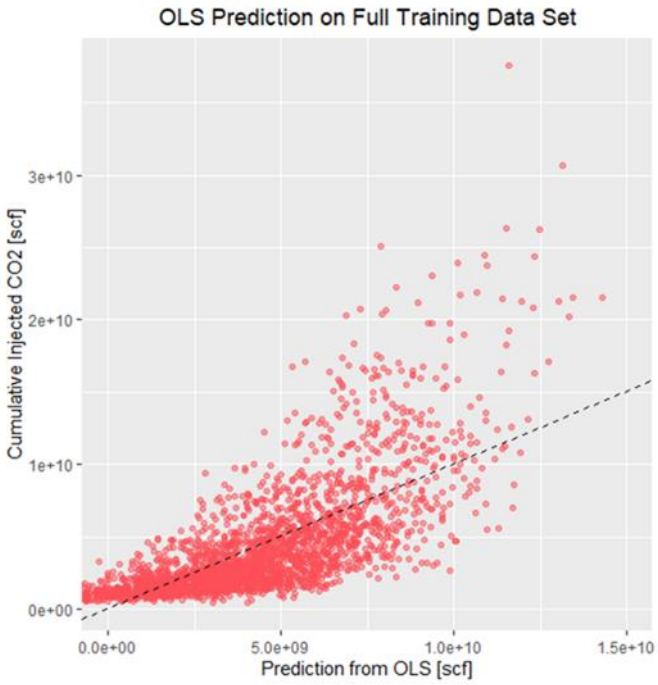

Figure 5.21 Predicted vs. observed cumulative injected CO_2 for the OLS model

Moreover, the R^2 value (**Appendix A2**) corresponding to the model fit is close to 51%. For the goodness of fit the corresponding values are:

$$\mathbf{AAE} = \mathbf{1.89}\ \boldsymbol{Bscf}$$

$$\mathbf{MSE} = \mathbf{7460655}\ \boldsymbol{kBscf^2}$$

These values will later be compared to tree-based methods to find out which statistical and machine-learning algorithm is describing the best performance for CO_2 sequestration in unconventional reservoirs. Finally, after performing multiple linear regression, the ultimate step would involve checking for potential problems and if the regression model assumptions are satisfied. This can be verified through diagnostic plots as seen in Figure 5.22.

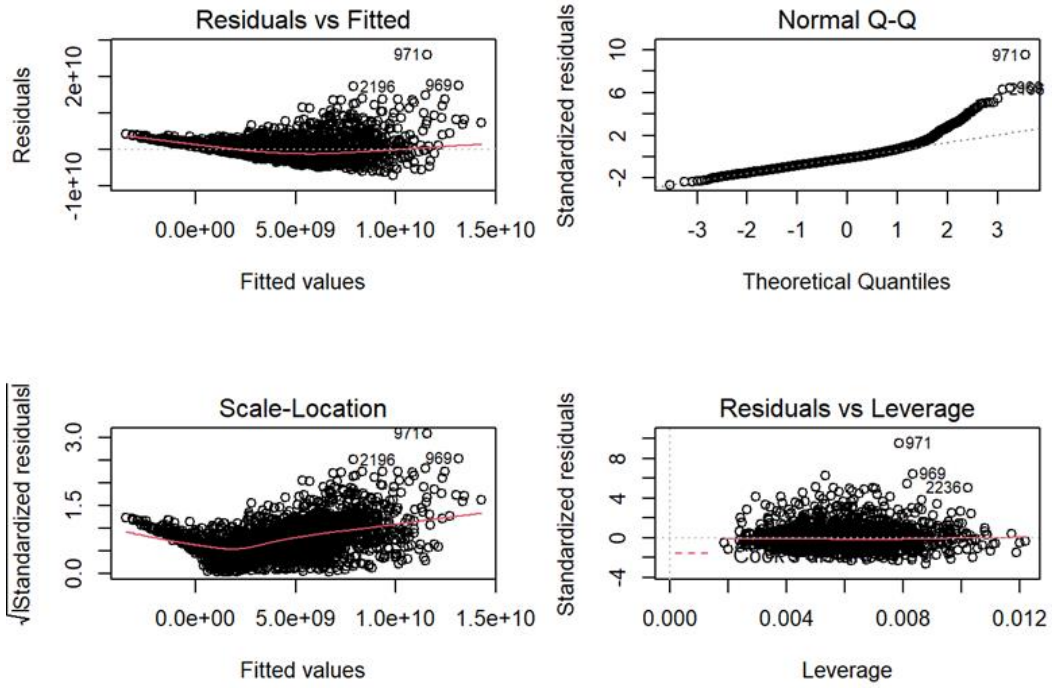

Figure 5.22 Regression diagnostic plots

It can be seen in Figure 5.22 that:

- The residuals display a U-shaped pattern in the residuals vs fitted plot, which provides an indication of non-linearity
- There is a non-constant variance (heteroscedasticity) observed in the scale location plot
- The residuals are not normally distributed, as seen in the Normal Q-Q plot
- Outlier are observed in residuals vs leverage plot

Since the OLS, regression model did not satisfy the linear regression assumptions by transforming the response variable and some predictor variables which had high skewness values will aid in modifying the diagnostic plots as well as improving the regression model. Therefore, it can be observed Figure 5.23 that after log transformation, the regression diagnostic plots seem to adhere to the linear regression assumptions. Note that in (**Appendix A2)** the R^2 value has increased from 51% to 67% which shows the importance of log transformation when you observe a non-linear relationship.

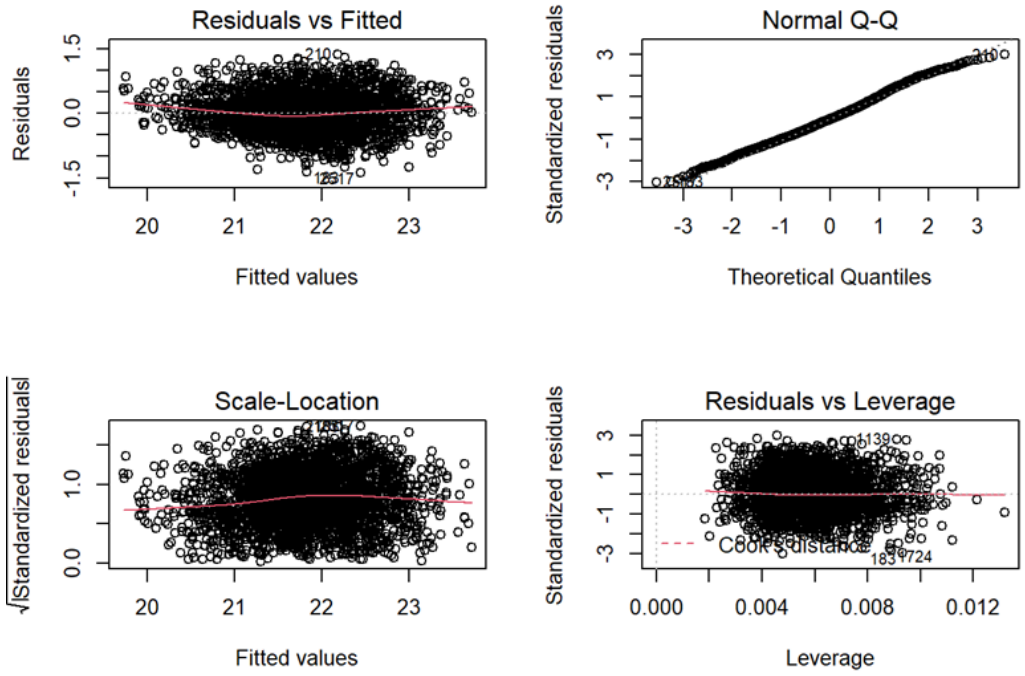

Figure 5.23 Diagnostic plots after log transformation

5.5.2 Tree-based methods

In the previous analysis which involved OLS regression modeling, the model had assumptions needed to be followed and it cannot capture nonlinear behavior directly until there is a transformation performed. However, tree-based methods do not impose any initial assumptions regarding linearity, hence they can capture nonlinear behavior and they are efficiently understandable.

It is observed from Figure 5.24 that the regression tree sections the reservoir and operational parameters into 20 regions of space. These 20 regions represent the terminal nodes for the tree. Moreover, only 10 of the 22 variables have been used in constructing the tree. The usefulness of regression trees can be displayed in Figure 5.24. From this regression tree, the main predictors which are influencing the cumulative injected CO_2 are the ones located around the top. These include SRV_kf, Thickness and LHW. The regression tree described in Figure 5.24 contains 20 regions of space and 10 variables used in the tree construction. Regression trees are normally interpretable. However, in this case, with many regions and many predictor variables, it's difficult to interpret. Also, this dataset is quite large.

Moreover, this full regression tree might overfit the training data and this leads to poor test error performance. Hence, an approach to prune the tree to make it more compact and easily interpretable would be a pleasant scenario.

A complexity parameter (cp) can be used which prunes the tree by penalizing the tree if it has too many splits. 0.01 is the default value. A larger cp value generates a smaller tree (Kassambara, 2017). It can be noted in Figure 5.25 that the cp value which would boost the accuracy of the model and prune the tree is 0.012. Finally, this value was used to provide the final version of the regression tree in a more compact form, which can be seen in Figure 5.26. The regression tree in Figure 5.26 can be interpreted as follows: the regression tree has a section with a high mean response value of cumulative injected CO_2 and a section with a low mean response value. It can be noted in Figure 5.26 that observations with SRV_kf < 0.0054 md are assigned to the left of the branch in the top split. This group is further subdivided by SRV_kf and edge_x. An example of a low-volume sequestration scenario (low mean response value) would be SRV_kf < 0.0054 md and SRV_kf < 0.0029 md which gives a mean response value of approximately 2 Bscf.

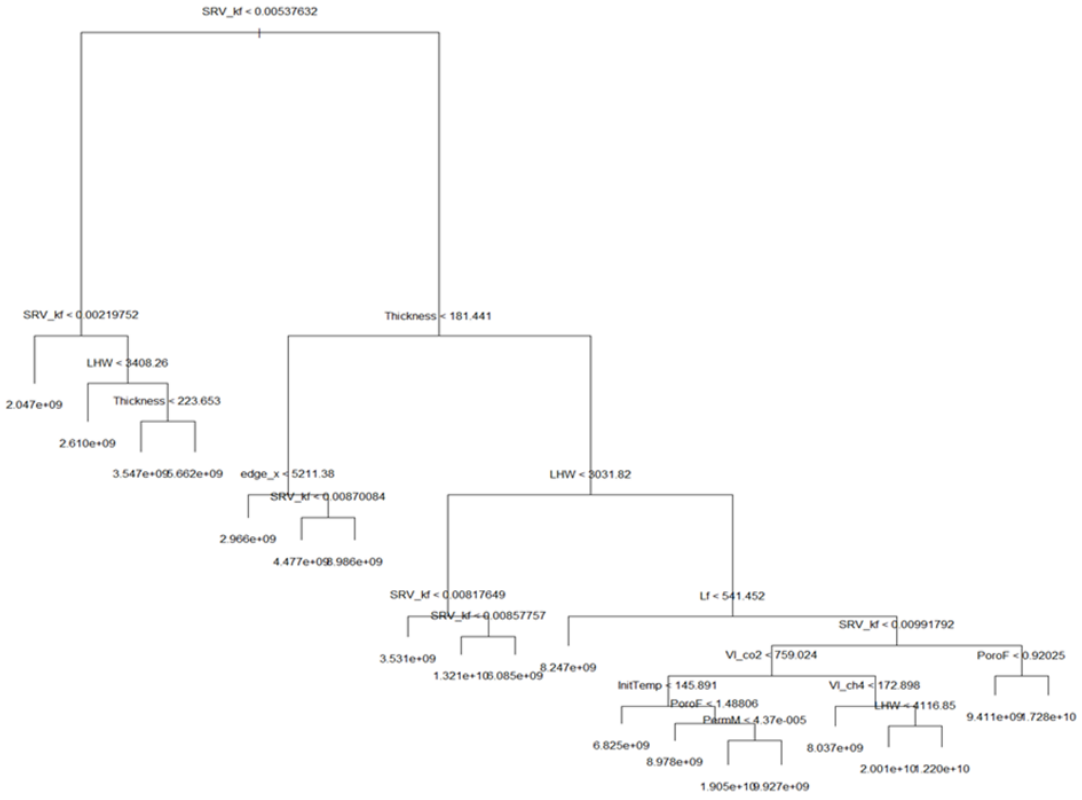

Figure 5.24 Unpruned regression tree

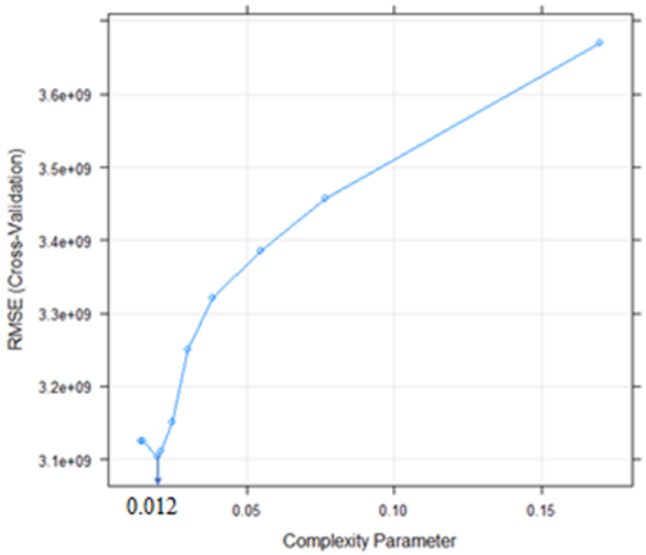

Figure 5.25 Complexity parameter (cp) plot

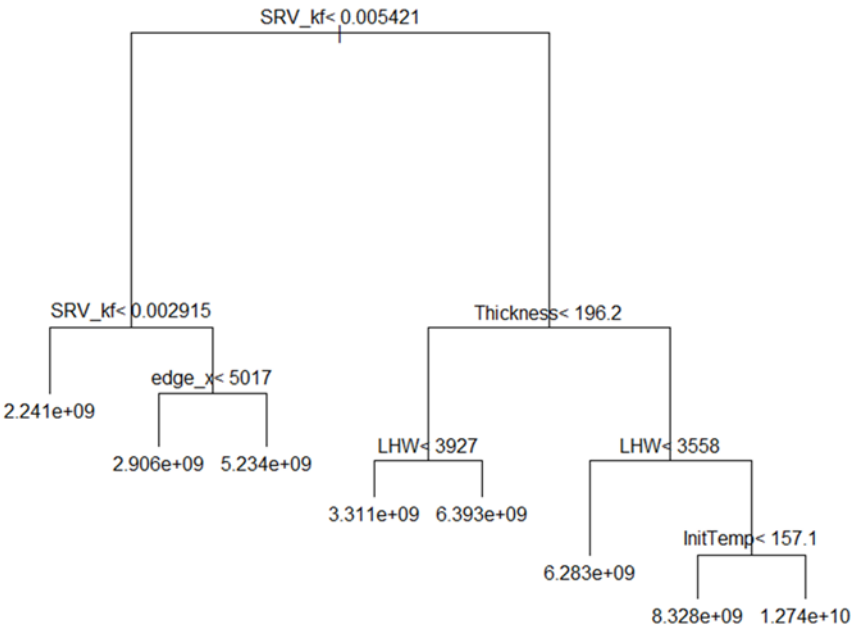

Figure 5.26 Pruned regression tree

For instance, in Figure 5.26 a high-volume sequestration case would be when:

- SRV_kf >= 0.0054 md, Thickness < 196.2 ft and LHW >= 3927 ft which would correspond to a mean response value of approximately 6.4 Bscf.

This process can be continued until all the branches are interpreted. Finally, it can be concluded from the regression tree that the most influential parameters in determining the performance of CO_2 sequestration in unconventional shale reservoirs are SRV_kf, Thickness, edge_x and LHW. Also, the model prediction error can be estimated by a cross-plot of actual and predicted values from the pruned tree. This corresponds to a prediction error of:

$$\mathbf{MSE} = \mathbf{11005778}\ \boldsymbol{kBscf^2}$$

$$\mathbf{RMSE} = \mathbf{3.32}\ \boldsymbol{Bscf}$$

$$\mathbf{Rsquare} = \mathbf{0.31} = \mathbf{31}\%$$

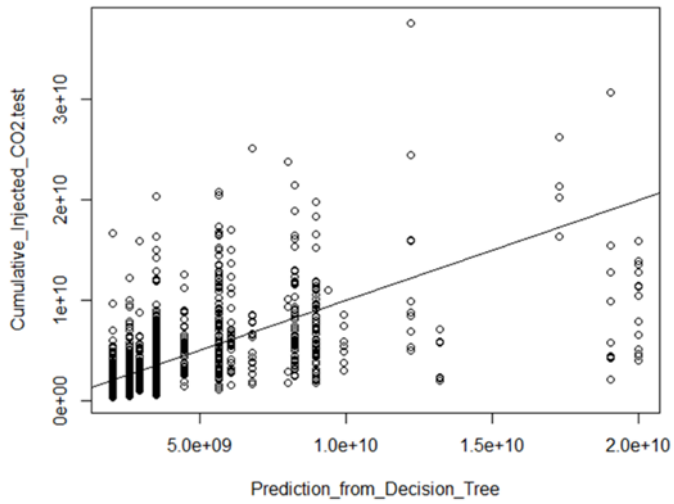

Figure 5.27 Predicted vs observed cumulative injected CO_2 for regression tree

To improve the results of the previous regression tree, more powerful techniques were employed, such as bagging, random forest and gradient-boosting machine. As explained previously in the methodology chapter, these methods aid in decreasing the variance of a statistical-machine learning algorithm as well as improving the performance of these methods. Three cross-plots were made for these methods to assess the prediction error and check if there is an improvement from the preceding method. The first cross-plot as seen in Figure 5.28 for bagging produced the following prediction error:

$$\mathbf{MSE} = \mathbf{7484000}\ \boldsymbol{kBscf^2}$$

$$\mathbf{RMSE} = \mathbf{2.74}\ \boldsymbol{Bscf}$$

$$\mathbf{Rsquare} = \mathbf{0.52} = \mathbf{52}\%$$

The second cross-plot of random forest can be seen in Figure 5.29 this technique provided the following prediction error:

$$\mathbf{MSE} = \mathbf{7382367}\ \boldsymbol{kBscf^2}$$

$$\mathbf{RMSE} = \mathbf{2.72}\ \boldsymbol{Bscf}$$

$$\mathbf{Rsquare} = \mathbf{0.54} = \mathbf{54}\%$$

The third and final cross-plot of GBM can be seen in Figure 5.30 this technique produced the following prediction error:

$$\mathbf{MSE} = \mathbf{7564000}\ \boldsymbol{kBscf^2}$$

$$\mathbf{RMSE} = \mathbf{2.75}\ \boldsymbol{Bscf}$$

$$\mathbf{Rsquare} = \mathbf{0.45} = \mathbf{45}\%$$

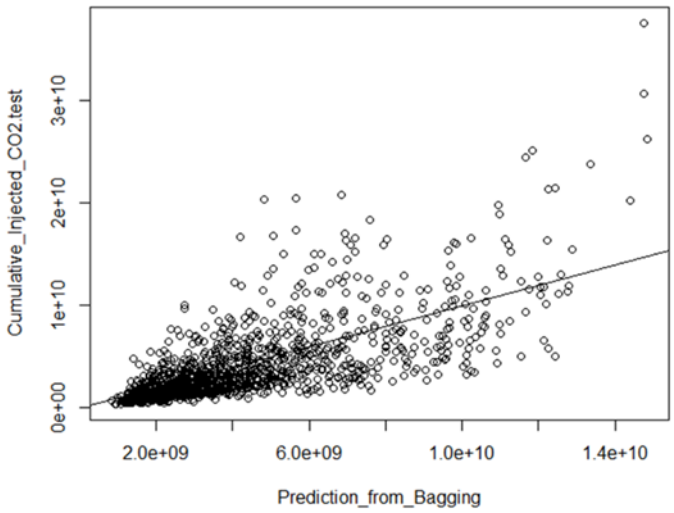

Figure 5.28 Predicted vs observed cumulative injected CO_2 for bagging

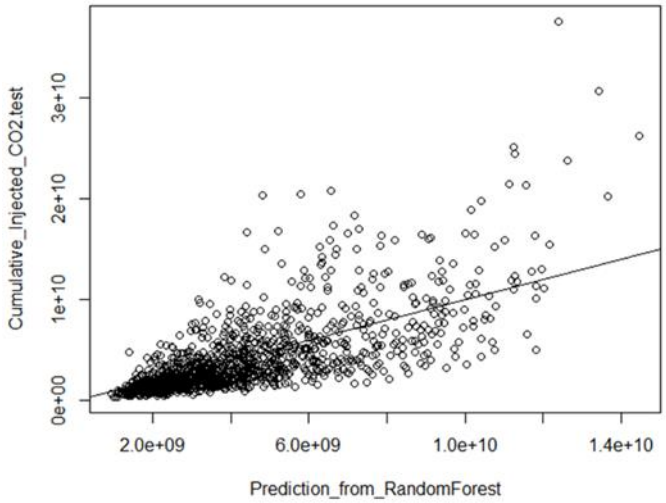

Figure 5.29 Predicted vs observed cumulative injected CO_2 for random forest

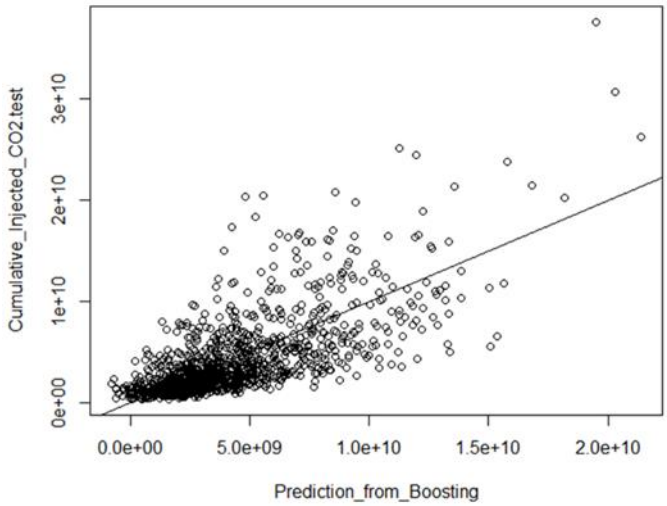

Figure 5.30 Predicted vs observed cumulative injected CO_2 for GBM

Overall, comparing the tree-based methods, it can be observed that random forest produces the minimum prediction error and hence it is the best among the tree-based techniques for prediction performance of CO_2 sequestration. Finally, a comparison of all the data-driven models to check which is the best model to predict the performance of CO_2 sequestration in unconventional shale reservoirs can be seen in Table 5.7.

Table 5.7 Comparison of data-driven models

Predictive Model	***MSE (kBscf^2)***	***RMSE (Bscf)***	***R^2 (%)***
Multiple Linear Regression	7460655	2.73	51
Regression Tree	11005778	3.32	31
Bagging	7484000	2.74	52
Random Forest	7382367	2.72	54
Gradient Boosting Machine	7564000	2.75	45

It can be noted in Table 5.7 that Random Forest outperforms all other data-driven methods with the lowest prediction error of 2.72 Bscf and the highest R^2 value of 54%. These results obtained are consistent with the theoretical background of RF, as most literature claim that it is one of the most powerful machine learning algorithms.

5.5.3 Variable importance

The last part of this study was to identify the key drivers of the CO_2 sequestration process in unconventional shale-gas reservoirs. This process is mainly managed by analyzing the response variable among a substantial set of predictor variables. In order to do this, RFs and GBMs have inbuilt functions for performing such a process to identify the most prominent predictors. For an RF model, the significance of a predictor is determined by permuting its values and calculating the percent decrease in RMSE. The notion is that if a random permutation breaks an essential variable, the accuracy will suffer considerably. Whereas for GBM, the average prediction improvement across all trees created by the boosting method represents the relative significance of a variable (Lolon et al., 2016).

It is seen from Figure 5.31 that, *SRV* Fracture Permeability (*SRV*-k_f) is the most influential predictor and has an immense impact on the performance of CO_2 sequestration followed by Thickness, Length of Reservoir (L_x), Horizontal Wellbore Length (L_{hw}), and Fracture Permeability (k_f).

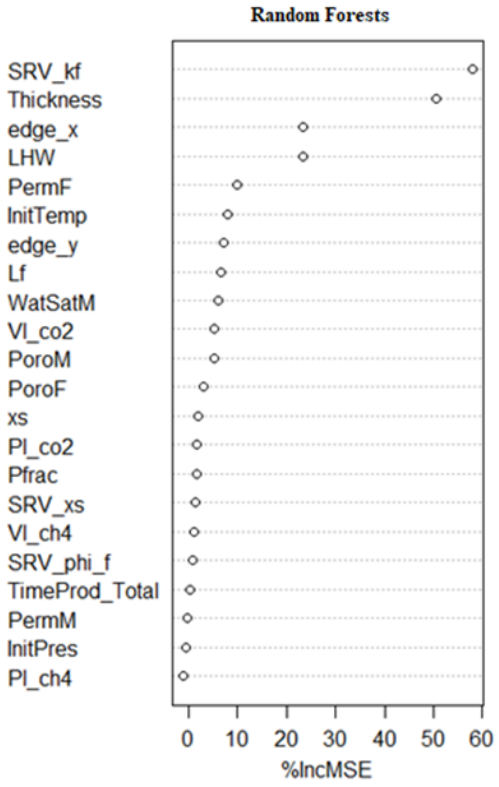

Figure 5.31 Variable importance for random forest model

Furthermore, it can be observed in Figure 5.32 that, *SRV* Fracture Permeability (*SRV*-k_f) is the most influential predictor and has an immense impact on the performance of CO_2 sequestration followed by Thickness, Horizontal Wellbore Length (L_{hw}), Length of Reservoir (L_x), and Langmuir Volume CO_2 ($V_{L\text{-}CO2}$). It can be noted that for both RF and GBM models, the top two decisive predictors (SRV Fracture Permeability and Thickness) for the shale-gas reservoirs are the same.

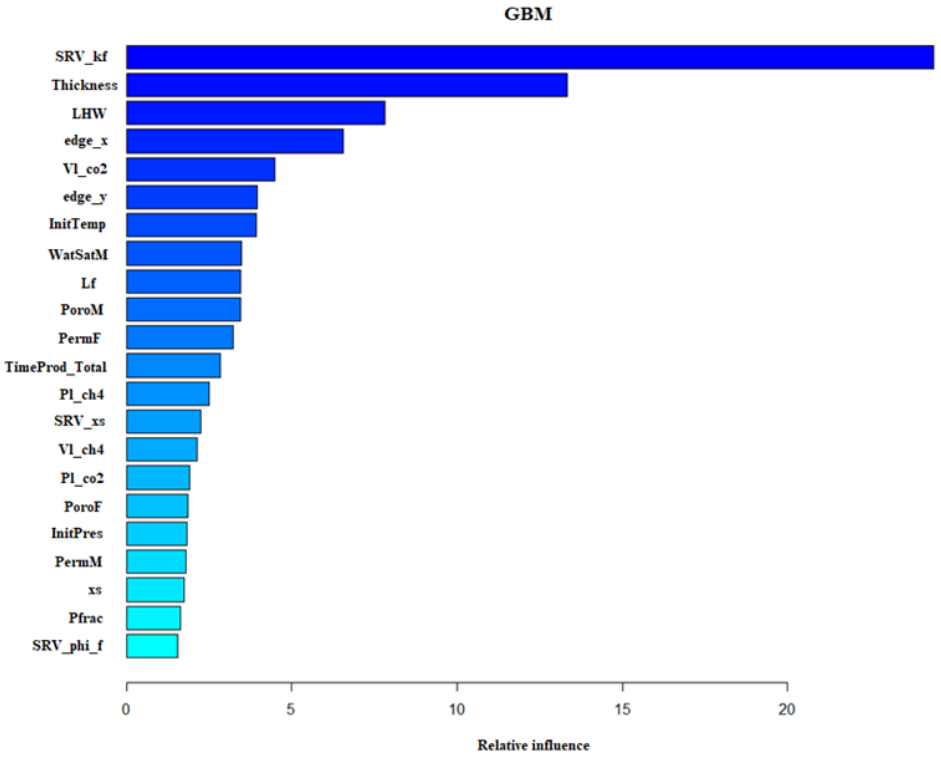

Figure 5.32 Relative influence for GBM model

It can be noted in Figure 5.33 that the models provide different rankings in terms of influence to CO_2 sequestration performance. However, the other ranking for the predictors is differently because, for instance RF ranks the most important predictors differently from the GBM and OLS. But it can be observed that the results are not significantly different between RF and GBM, as both models are reliable. The main conclusion of the parameters and in terms of their physical sense should be left to the Petroleum Engineer or Upstream Geoscientists to use the domain knowledge and interpret the significance of these variables. Nonetheless, in some cases you don't have a full understanding of CO_2 sequestration process, especially in an unconventional reservoir. We would still have some questions.

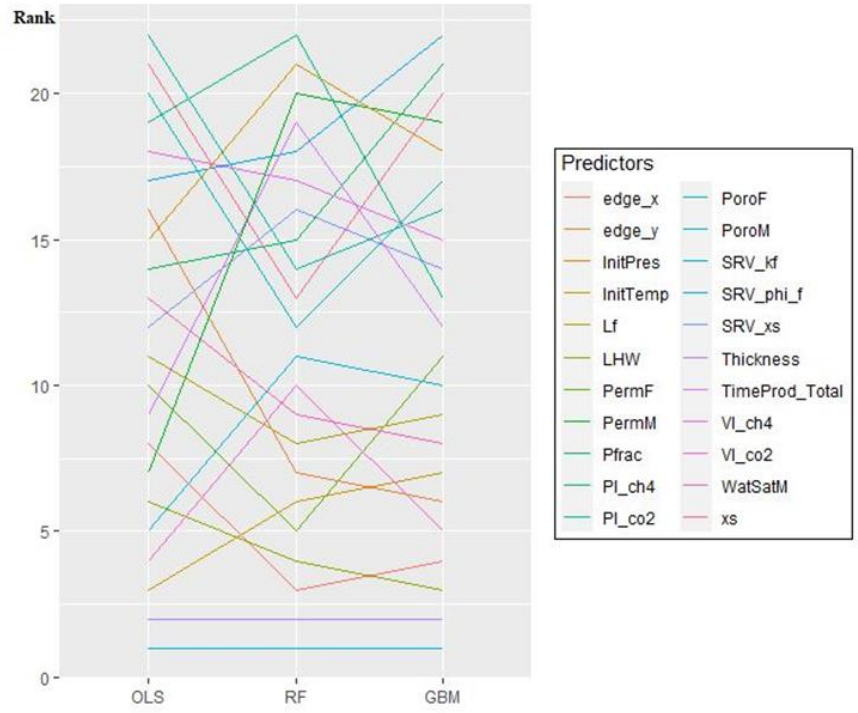

Figure 5.33 Predictor rankings for different predictive models

For this study, the significance of the predictors which drive high-performance will be assessed to see if these variables make sense from a physical standpoint.

SRV Fracture Permeability (*SRV-k_f*). Since the SRV zone is a stimulated section of the reservoir, the fracture apertures (openings) have an increased dimension and become more conductive. The total mobility and fluid flow will be more pronounced. CH_4 will be produced, and CO_2 can be injected and progress in the *SRV* zone accordingly. Hence, CO_2 sequestration performance would be high and production of CH_4 when substantial *SRV* Fracture Permeability values are attained.

Thickness, Length of Reservoir (L_x), Length of Reservoir (L_y). Reservoir thickness plays an important role in terms of the reserve capacity. Moreover, the thickness, length of reservoir (L_x) and length of reservoir (L_y) together are important because they describe the gross bulk volume of the drainage area.

Horizontal Wellbore Length (L_{hw}). The horizontal wellbore length is crucial because the well intersects the fractures which are very conductive, and this would aid in the production of CH_4 in order to inject CO_2. Besides, a long horizontal wellbore length would maximize the contact area with the *SRV* zone, and this would clearly influence the productivity index of the well.

Fracture Permeability (k_f). Fracture permeability is a key parameter by reason of the fracture openings (apertures) are much sizeable in contrast to matrix pore throat dimensions. As well as being highly conductive, hence it accounts for the overall mobility (transmissivity) inside the unconventional reservoir.

Langmuir Volume CO_2 ($V_{L\text{-}CO2}$). The Langmuir volume CO_2 is vital because it aids in controlling the reserves. The importance of this isotherm is consistent with the literature as Yu & Sepehrnoori (2019) point out that the gas volume at infinite pressure is referred to as the Langmuir volume, and it represents the maximum storage capacity for gas.

Chapter 6 Concluding Remarks

In this study, data-analytics is used to investigate the primary variables that affect CO_2 sequestration process. The study focuses on unconventional shale reservoirs. An EDA through data mining and visualization was performed to understand features and patterns within a dataset of CO_2 sequestration scenarios in shale reservoirs. This dataset that was used constituted of a significant number of numerical-simulation scenarios (close to 1400 scenarios) that were run using a state-of-the art reservoir simulator that was part of another study by (Kulga, 2014). After developing insights into the dataset, statistical and machine-learning algorithms were used to develop predictive models. For evaluating the relationship and accurately predict the process performance between reservoir parameters, operational parameters and cumulative CO_2 injected. Then, predictive efficacy of these models was assessed to see which model captures the cumulative CO_2 injected more precisely. In addition, variable importance approach was used to determine which parameters can drive high-performance for CO_2 sequestration. Consequently, these variables were checked to see if they make sense from a physical standpoint.

6.1 Conclusions

The major conclusions from this study are as follows:

1) Operational parameters are more prominent in driving high-performance CO_2 sequestration process in unconventional shale reservoirs. *SRV* Fracture Permeability (*SRV*-k_f) is the top influential parameter for long-term CO_2 sequestration process.
2) The most influential parameters that drive CO_2 sequestration performance according to RFs are:
    - *SRV* Fracture Permeability (*SRV*-k_f)
    - Thickness
    - Length of Reservoir (L_x)
    - Horizontal Wellbore Length (L_{hw})
    - Fracture Permeability (k_f)
3) The most influential parameters that drive CO_2 sequestration performance according to GBMs are:
    - *SRV* Fracture Permeability (*SRV*-k_f)
    - Thickness
    - Horizontal Wellbore Length (L_{hw})

- Length of Reservoir (L_x)
- Langmuir Volume CO_2 ($V_{L\text{-}CO2}$)

4) Random Forests have the best predictive ability since it gives the lowest prediction error of 2.72 Bscf and the highest percentage of variance explained with an R^2 value close to 54%. This result agrees with literature that RFs model is one of the most powerful machine-learning algorithms.
5) Regression trees are easily interpretable and can rank, which are the most influential parameters that influence cumulative CO_2 injected. These parameters are near the top of the tree.
6) The model accuracy of multiple linear regression increased from 51% to 67% after log-transforming some predictor variables which had high skewness.
7) It was shown that the optimal model for OLS was selected by k-fold cross validation based on test error.
8) The outlier points observed in EDA cannot be because of an incorrect input value in the dataset but because of the dependency between the predictors and response variables, which causes additional lognormality and displayed as the outlier points.

6.2 Recommendations

- On one hand, this dataset covers a wide range of shale formations with different characteristics as seen in Table 3.1. As long as the formation is within these ranges, then the models can be used to generalize the results. On the other hand, the range of applicability is quite wide based on the ranges of parameters in Table 3.1, this can be even further expanded with new simulation scenarios if needed.
- GBM is a powerful machine-learning algorithm. The tuning parameters, such as shrinkage factor, can be altered to improve the model accuracy.
- The regression tree can be converted to a classification tree in order to simplify the problem and predict whether the performance would be high or low.

References

Al-Alwani, M. A., Britt, L. K., Dunn-Norman, S., Alkinani, H. H., Al-Hameedi, A. T. T., Al-Attar, A. M., Alkhamis, M. M., & Al-Bazzaz, W. H. (2019). From data collection to data analytics: How to successfully extract useful information from big data in the oil & gas industry? *Society of Petroleum Engineers - SPE/IATMI Asia Pacific Oil and Gas Conference and Exhibition 2019, APOG 2019*. https://doi.org/10.2118/196428-ms

Armstrong, G. L., Nichols, J. L., & W.Nichols, M. (2019). *Meeting the Dual Challenge A Roadmap to At-Scale Deployment of Carbon Capture, Use, And Storage*. https://dualchallenge.npc.org/files/CCUS_V1-FINAL.pdf

Bock, T. (n.d.). *What is a Correlation Matrix?* . Retrieved August 12, 2021, from https://www.displayr.com/what-is-a-correlation-matrix/

Boosari, S. S. H., Aybar, U., & Eshkalak, M. O. (2015). Carbon Dioxide Storage and Sequestration in Unconventional Shale Reservoirs. *Journal of Geoscience and Environment Protection*, *03*(01), 7–15. https://doi.org/10.4236/gep.2015.31002

BP Energy Outlook. (2019). https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/energy-outlook/bp-energy-outlook-2019.pdf

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. https://www.taylorfrancis.com/books/mono/10.1201/9781315139470/classification-regression-trees-leo-breiman-jerome-friedman-richard-olshen-charles-stone

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists* (Second). https://www.oreilly.com/library/view/practical-statistics-for/9781492072935/

Cipolla, C. ., Lolon, E. ., Erdle, J. ., & Rubin, B. (2010). *Reservoir Modeling in Shale-Gas Reservoirs*. https://onepetro.org/REE/article-abstract/13/04/638/192587/Reservoir-Modeling-in-Shale-Gas-Reservoirs?redirectedFrom=fulltext

Ding, D. yu, Farah, N., Bourbiaux, B., Wu, Y.-S., & Wang, C. (2014). *Numerical Simulation of Low Permeability Unconventional Gas Reservoirs*.

https://onepetro.org/SPEUNCV/proceedings-abstract/14UNCV/2-14UNCV/D021S011R001/210908

Ertekin, T., Abou-Kassem, J. H., & King, G. R. (2001). *Basic Applied Reservoir Simulation*. https://store.spe.org/Basic-Applied-Reservoir-Simulation--P12.aspx

Feder, J., Palisch, T., Livescu, S., Reid, D., Petrone, A., Pearson, R., & Ozkan, E. (2021). *SPE Technical Directors' Outlook: The Industry's Transformation in 2020 and What It Means for the Future*. Journal of Petroleum Technology. https://jpt.spe.org/spe-technical-directors-outlook-the-industrys-transformation-in-2020-and-what-it-means-for-the-future

Friedman, J. H. (2001). *Greedy function approximation: a gradient boosting machine*. https://www.jstor.org/stable/2699986

Ghoodjani, E., & Bolouri, S. H. (2012). *Numerical and Analytical Optimization of Injection Rate During CO2-EOR and -Sequestration Processes*. https://onepetro.org/CMTCONF/proceedings-abstract/12CMTC/All-12CMTC/CMTC-150157-MS/567

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Holdaway, K. R. (2009). Exploratory Data Analysis in Reservoir Characterization Projects. *Society of Petroleum Engineers*, *1*(October), 1–20. https://doi.org/10.3997/2214-4609-pdb.170.spe125368

Holdaway, K. R. (2014). *Harness Oil and Gas Big Data with Analytics* (1st ed.). Wiley. https://www.perlego.com/book/997975/harness-oil-and-gas-big-data-with-analytics-optimize-exploration-and-production-with-datadriven-models-pdf

Iman, R. L., & Conover, W. J. (1986). *A Modern Approach to Statistics*. https://scholar.google.com/scholar?q=+author:R.L. Iman

James, G., Witten, D., Tibshirani, R., & Hastie, T. (2013). *An Introduction to Statistical Learning with Applications in R*. https://www.statlearning.com/

Kassambara, A. (2017). *Machine Learning Essentials Practical Guide in R*. https://www.scribd.com/document/431629248/Kassambara-Alboukadel-Machine-

Learning-Essential-Z-lib-org

Kirkman, T. . (1996). *Statistics to use*. http://www.physics.csbsju.edu/stats/

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. In *Springer* (Vol. 26). http://appliedpredictivemodeling.com/s/Applied_Predictive_Modeling_in_R.pdf

Kulga, B. (2014). *Analysis of The Efficacy of Carbon Dioxide Sequestration in Depleted Shale Gas Reservoirs* [The Pennsylvania State University]. https://etda.libraries.psu.edu/catalog/22825

Kulga, B., & Ertekin, T. (2018). Numerical representation of multi-component gas flow in stimulated shale reservoirs. *Journal of Natural Gas Science and Engineering*, *56*(June), 579–592. https://doi.org/10.1016/j.jngse.2018.06.023

Lolon, E., Hamidieh, K., Weijers, L., Mayerhofer, M., Melcher, H., & Oduba, O. (2016). Evaluating the Relationship Between Well Parameters and Production using Multivariate Statistical Models: A Middle Bakken and Three Forks Case History. *Society of Petroleum Engineers - SPE Hydraulic Fracturing Technology Conference, HFTC 2016*. https://doi.org/10.2118/179171-ms

Mishra, S., & Datta-Gupta, A. (2018). *Applied Statistical Modeling and Data Analytics: A Practical Guide for the Petroleum Geosciences*. Candice Janco. https://www.elsevier.com/books/applied-statistical-modeling-and-data-analytics/mishra/978-0-12-803279-4

Mishra, S., & Lin, L. (2017). Application of Data Analytics for Production Optimization in Unconventional Reservoirs: A Critical Review. *SPE/AAPG/SEG Unconventional Resources Technology Conference 2017*, *c*, 1060–1065. https://doi.org/10.15530/urtec-2017-2670157

Mishra, S., Oruganti, Y. D., & Sminchak, J. (2014). Parametric analysis of CO2 geologic sequestration in closed volumes. *Environmental Geosciences*, *21*(2), 59–74. https://doi.org/10.1306/eg.03101413009

Mishra, S., Schetter, J., Datta-Gupta, A., & Bromhal, G. (2021, March 1). *Robust Data-Driven Machine-Learning Models for Subsurface Applications: Are We There Yet?* Journal of Petroleum Technology. https://jpt.spe.org/robust-data-driven-machine-learning-models-for-subsurface-applications-are-we-there-yet

Mohaghegh, S. D. (2018). *Data-Driven Analytics for the Geological Storage of CO2*. https://www.routledge.com/Data-Driven-Analytics-for-the-Geological-Storage-of-CO2/Mohaghegh/p/book/9780367734381

Moore, D. S., Notz, W. I., & Fligner, M. (2018). *The Basic Practice of Statistics* (8th ed.). https://store.macmillanlearning.com/ca/product/The-Basic-Practice-of-Statistics/p/1319042570

NAE Grand Challenges For Engineering ™. (2017). www.engineeringchallenges.org.

Narayanan, A. S., Sankaran, S., & Dennett, L. (2020). *How Data Analytics Skills Can Open New Opportunities for Oil and Gas Professionals*. The Way Ahead. https://jpt.spe.org/twa/how-data-analytics-skills-can-open-new-opportunities-for-oil-and-gas-professionals

Navidi, W. (2008). *Statistics for Engineers and Scientists*. McGraw-Hill Education. http://ndl.ethernet.edu.et/bitstream/123456789/37620/1/William Navidi_2015.pdf

Perez, H. H., Datta-Gupta, A., & Mishra, S. (2005). *The Role of Electrofacies, Lithofacies, and Hydraulic Flow Units in Permeability Predictions from Well Logs: A Comparative Analysis Using Classification Trees*. https://onepetro.org/REE/article-abstract/8/02/143/112575/The-Role-of-Electrofacies-Lithofacies-and?redirectedFrom=fulltext

R Development Core Team. (2021). *R: A Language and Environment for Statistical Computing* (4.0.4). https://www.r-project.org/

Schuetter, J., Mishra, S., Zhong, M., & LaFollette, R. (2018). A data-analytics tutorial: Building predictive models for oil production in an unconventional shale reservoir. *SPE Journal*, *23*(4), 1075–1089. https://doi.org/10.2118/189969-pa

United Nations Population Division. (2019). World population prospects 2019: highlights. In *Department of Economic and Social Affairs, Population Division*. https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. https://link.springer.com/book/10.1007/978-1-4757-2440-0

Venables, W. N., & Ripley, B. D. (1996). *Modern Applied Statistics with S-Plus*. https://scholar.google.com/scholar?q=+author:W.N. Venables

Yang, Q. J., Dong, Y., & Zhou, F. (2005). *Short-Term Numerical Simulation of Geological Sequestration of CO2 in the Barrow Sub-Basin, West Australia*. 4. https://onepetro.org/SPEAPHS/proceedings-abstract/05APHS/All-05APHS/SPE-95354-MS/73092

Yu, W., & Sepehrnoori, K. (2019). *Shale Gas and Tight Oil Reservoir Simulation*. https://www.sciencedirect.com/book/9780128138687/shale-gas-and-tight-oil-reservoir-simulation#book-info

Zhong, M., Schuetter, J., Mishra, S., & LaFollette, R. F. (2015). Do data mining methods matter? : A Wolfcamp "Shale" case study. *Society of Petroleum Engineers - SPE Hydraulic Fracturing Technology Conference 2015*, 136–147. https://doi.org/10.2118/173334-ms

Appendix

A1. Best subset selection

This R code is used to perform best subset selection. The package used to perform best subset selection was leaps.

```
df<- read.csv('Regression.csv')
```

Next perform best subset selection

```
library (leaps)

## Warning: package 'leaps' was built under R version 4.0.5

regfit. full=regsubsets(cum_inj~.,df)
summary(regfit.full)

## Subset selection object
## Call: regsubsets.formula(cum_inj ~ ., df)
## 22 Variables  (and intercept)
##                Forced in Forced out
## Thickness          FALSE      FALSE
## PoroM              FALSE      FALSE
## PoroF              FALSE      FALSE
## PermM              FALSE      FALSE
## PermF              FALSE      FALSE
## xs                 FALSE      FALSE
## WatSatM            FALSE      FALSE
## Vl_ch4             FALSE      FALSE
## Pl_ch4             FALSE      FALSE
## Vl_co2             FALSE      FALSE
## Pl_co2             FALSE      FALSE
## InitPres           FALSE      FALSE
## InitTemp           FALSE      FALSE
## TimeProd_Total     FALSE      FALSE
## Pfrac              FALSE      FALSE
## LHW                FALSE      FALSE
## Lf                 FALSE      FALSE
## edge_x             FALSE      FALSE
## edge_y             FALSE      FALSE
## SRV_phi_f          FALSE      FALSE
## SRV_kf             FALSE      FALSE
## SRV_xs             FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Thickness PoroM PoroF PermM PermF xs  WatSatM Vl_ch4 Pl_ch4 Vl
_co2
## 1  ( 1 ) " "       " "   " "   " "   " "   " " " " "     " "    " "    "
"
## 2  ( 1 ) "*"       " "   " "   " "   " "   " " " " "     " "    " "    "
"
## 3  ( 1 ) "*"       " "   " "   " "   " "   " " " " "     " "    " "    "
"
```

```
## 4  ( 1 ) "*"       " "   " "   " "   " "   " " " " "     " "    " "    "
"
## 5  ( 1 ) "*"       " "   " "   " "   " "   " " " " "     " "    " "    "*
"
## 6  ( 1 ) "*"       " "   " "   " "   " "   " " " " "     " "    " "    "*
"
## 7  ( 1 ) "*"       "*"   " "   " "   " "   " " " " "     " "    " "    "*
"
## 8  ( 1 ) "*"       "*"   " "   " "   " "   " " " " "     " "    " "    "*
"
##          Pl_co2 InitPres InitTemp TimeProd_Total Pfrac LHW Lf  edge_x e
dge_y
## 1  ( 1 ) " "    " "      " "      " "            " "   " " " " " " " "    "
"
## 2  ( 1 ) " "    " "      " "      " "            " "   " " " " " " " "    "
"
## 3  ( 1 ) " "    " "      " "      " "            " "   "*" " " " " "    "
"
## 4  ( 1 ) " "    " "      "*"      " "            " "   "*" " " " " "    "
"
## 5  ( 1 ) " "    " "      "*"      " "            " "   "*" " " " " "    "
"
## 6  ( 1 ) " "    " "      "*"      " "            " "   "*" "*" " " "    "
"
## 7  ( 1 ) " "    " "      "*"      " "            " "   "*" "*" " " "    "
"
## 8  ( 1 ) " "    " "      "*"      " "            " "   "*" "*" " " "    "
"
##          SRV_phi_f SRV_kf SRV_xs
## 1  ( 1 ) " "       "*"    " "
## 2  ( 1 ) " "       "*"    " "
## 3  ( 1 ) " "       "*"    " "
## 4  ( 1 ) " "       "*"    " "
## 5  ( 1 ) " "       "*"    " "
## 6  ( 1 ) " "       "*"    " "
## 7  ( 1 ) " "       "*"    " "
## 8  ( 1 ) "*"       "*"    " "

regfit.full=regsubsets(cum_inj~.,data=df,nvmax=22)
reg.summary=summary(regfit.full)
reg.summary$rsq

##  [1] 0.2162578 0.3195239 0.4420393 0.4608364 0.4783450 0.4873174 0.4939
863
##  [8] 0.4975943 0.4997374 0.5019890 0.5036204 0.5048081 0.5059378 0.5065
767
## [15] 0.5069526 0.5070767 0.5071963 0.5073330 0.5073736 0.5073901 0.5073
963
## [22] 0.5074005
```

A2. Multiple linear regression

This R code is used to perform multiple linear regression.

First load the dataset and check the head of the file. The dataset name will be regression.

```
df<- read.csv('Regression.csv')
```

Next perform multiple linear regression

```
lm.fit=lm(cum_inj~Thickness + PoroM + PermM + PermF + Vl_ch4 +
    Vl_co2 + InitTemp + TimeProd_Total + LHW + Lf + edge_x +
    SRV_phi_f + SRV_kf + SRV_xs, data=df)
summary(lm.fit)

##
## Call:
## lm(formula = cum_inj ~ Thickness + PoroM + PermM + PermF + Vl_ch4 +
##     Vl_co2 + InitTemp + TimeProd_Total + LHW + Lf + edge_x +
##     SRV_phi_f + SRV_kf + SRV_xs, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.217e+09 -1.624e+09 -3.617e+08  1.040e+09  2.601e+10
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.801e+10  7.162e+08 -25.147  < 2e-16 ***
## Thickness       2.467e+07  9.608e+05  25.681  < 2e-16 ***
## PoroM           2.260e+08  3.776e+07   5.985 2.46e-09 ***
## PermM          -6.097e+12  1.878e+12  -3.247  0.00118 **
## PermF           4.911e+11  2.712e+11   1.811  0.07031 .
## Vl_ch4         -3.849e+06  1.510e+06  -2.549  0.01087 *
## Vl_co2          2.359e+06  2.972e+05   7.937 3.08e-15 ***
## InitTemp        2.385e+07  2.372e+06  10.053  < 2e-16 ***
## TimeProd_Total  4.224e+04  1.728e+04   2.444  0.01460 *
## LHW             1.029e+06  1.974e+05   5.213 2.01e-07 ***
## Lf              1.575e+06  2.410e+05   6.534 7.71e-11 ***
## edge_x          4.150e+05  1.312e+05   3.163  0.00158 **
## SRV_phi_f       4.156e+08  9.344e+07   4.448 9.06e-06 ***
## SRV_kf          7.195e+11  2.814e+10  25.572  < 2e-16 ***
## SRV_xs         -3.446e+08  1.241e+08  -2.777  0.00553 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.739e+09 on 2532 degrees of freedom
## Multiple R-squared:  0.5066, Adjusted R-squared:  0.5038
## F-statistic: 185.7 on 14 and 2532 DF,  p-value: < 2.2e-16
```

Moreover, we can log transform the variables with high skewness to be able to check if now the residuals are satisfying the assumptions and to see if there is a development on the global R^2 value.

```
lm.fit=lm(log(cum_inj)~Thickness+PoroM+PermM+log(PermF)+Vl_ch4+Vl_co2+Init
Temp+TimeProd_Total+LHW+Lf+edge_x+SRV_phi_f+log(SRV_kf)+SRV_xs,data=df)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = log(cum_inj) ~ Thickness + PoroM + PermM + log(PermF) +
##     Vl_ch4 + Vl_co2 + InitTemp + TimeProd_Total + LHW + Lf +
##     edge_x + SRV_phi_f + log(SRV_kf) + SRV_xs, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38096 -0.32538 -0.02209  0.29819  1.36739
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.144e+01  1.670e-01 128.372  < 2e-16 ***
## Thickness       5.807e-03  1.608e-04  36.106  < 2e-16 ***
## PoroM           4.959e-02  6.317e-03   7.851 6.05e-15 ***
## PermM          -6.908e+02  3.141e+02  -2.200 0.027931 *
## log(PermF)     -1.826e-02  2.423e-02  -0.754 0.451028
## Vl_ch4         -9.701e-04  2.526e-04  -3.841 0.000126 ***
## Vl_co2          5.126e-04  4.973e-05  10.307  < 2e-16 ***
## InitTemp        5.114e-03  3.969e-04  12.886  < 2e-16 ***
## TimeProd_Total  1.221e-05  2.893e-06   4.220 2.53e-05 ***
## LHW             2.774e-04  3.301e-05   8.402  < 2e-16 ***
## Lf              4.138e-04  4.032e-05  10.262  < 2e-16 ***
## edge_x          8.384e-05  2.194e-05   3.822 0.000136 ***
## SRV_phi_f       9.145e-02  1.563e-02   5.851 5.53e-09 ***
## log(SRV_kf)     7.357e-01  1.978e-02  37.191  < 2e-16 ***
## SRV_xs         -9.015e-02  2.076e-02  -4.342 1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4583 on 2532 degrees of freedom
## Multiple R-squared:  0.6745, Adjusted R-squared:  0.6727
## F-statistic: 374.7 on 14 and 2532 DF,  p-value: < 2.2e-16
```

We can observe that the R^2 has improved quite significantly, then we can check the diagnostic plots to see if the assumptions are now satisfied

A3. Tree methods

This R code is for predictive modeling using tree methods. The libraries and packages used for tree methods include randomForest, tree, gbm.

```
df <- read.csv("Regression.csv")
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.0.5
```

```
set.seed(10)
train=sample(1:nrow(df),nrow(df)/2)
regressiontree.df=tree(cum_inj~.,df,subset=train)
summary(regressiontree.df)
```

```
##
## Regression tree:
## tree(formula = cum_inj ~ ., data = df, subset = train)
## Variables actually used in tree construction:
##  [1] "SRV_kf"    "LHW"       "Thickness" "edge_x"    "Lf"        "Vl_co
2"
##  [7] "InitTemp"  "PoroF"     "PermM"     "Vl_ch4"
## Number of terminal nodes:  20
## Residual mean deviance:  5.876e+18 = 7.362e+21 / 1253
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -7.512e+09 -1.327e+09 -5.449e+08  0.000e+00  1.043e+09  1.295e+10
```

```
mean((yhat-df.test)^2)
```

```
## [1] 1.153401e+19
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.0.5
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
set.seed(1)
bagging.df=randomForest(cum_inj~.,data=df, subset=train, mtry=22, importan
ce=TRUE)
bag.df
```

```
##
## Call:
##  randomForest(formula = cum_inj ~ ., data = df, mtry = 22, importance =
TRUE,      subset = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 22
##
##           Mean of squared residuals: 7.297848e+18
##                     % Var explained: 52.23
```

```
mean((yhat.bagging-df.test)^2)

## [1] 7.484254e+18

set.seed(10)
rf.df=randomForest(cum_inj~.,data=df, subset=train, mtry=10, importance=TR
UE)
yhat.rf=predict(rf.df, newdata=df[-train,])
mean((yhat.rf-df.test)^2)

## [1] 7.440936e+18
```