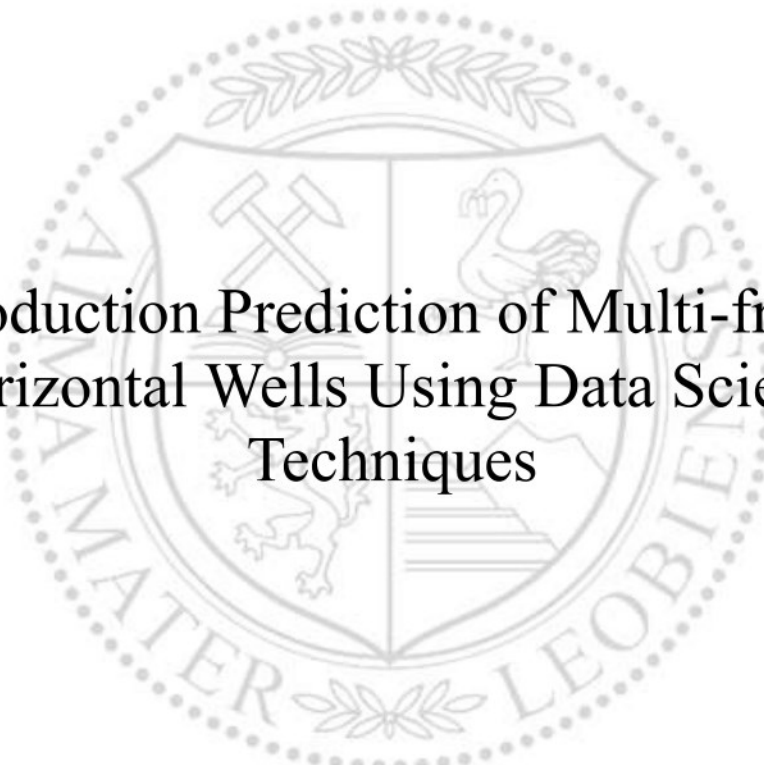




Chair of Petroleum and Geothermal Energy Recovery

Master's Thesis



Oil Production Prediction of Multi-fractured  
Horizontal Wells Using Data Science  
Techniques

Walid Bejjar, BSc

May 2021

Master Thesis

***Oil Production Prediction of Multi-fractured Horizontal Wells Using Data Science Techniques***

MONTANUNIVERSITÄT LEOBEN

**Written by:**

Walid Bejjar  
01435534

**Advisors:**

Univ.-Prof. Dipl.-Ing. Dr.mont. Herbert Hofstätter  
Dipl.-Ing. Dr.mont. Rudolf Fruhwirth

Leoben, 20.05.2021

## **EIDESSTATTLICHE ERKLÄRUNG**

Ich erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche erkenntlich gemacht habe.



**MONTANUNIVERSITÄT LEOBEN**  
www.unileoben.ac.at

**AFFIDAVIT**

I declare on oath that I wrote this thesis independently, did not use other than the specified sources and aids, and did not otherwise use any unauthorized aids.

I declare that I have read, understood, and complied with the guidelines of the senate of the Montanuniversität Leoben for "Good Scientific Practice".

Furthermore, I declare that the electronic and printed version of the submitted thesis are identical, both, formally and with regard to content.

Date 15.05.2021

A handwritten signature in black ink, appearing to read 'Walid Bejjar', written over a horizontal line.

Signature Author  
Walid Bejjar

## Kurzfassung

Die Öl- und Gasförderung aus Schieferöl- und Schiefergasspeichern hat rasch zugenommen die letzten zwei Jahrzehnte. Die Kombination von horizontalen Bohrlöchern und Hydraulic Fracturing war eine davon. Die Hauptgründe, warum die Schieferproduktion rentabel wurde. Mehrfach gebrochene horizontale Brunnen haben sich als fortschrittliches Mittel zur Steigerung der Bohrlochproduktivität in Reservoirs mit geringer Permeabilität herausgestellt. Die Wirtschaftlichkeit solcher Projekte hängt jedoch von mehreren Parametern ab. vor allem mit den Ölpreisschwankungen im letzten Jahrzehnt. Wählen Sie daher die Projekte aus mit dem höchsten Potenzial ist wichtig, um die Kapitalrendite zu maximieren.

Um die wirtschaftlichen Risiken zu verringern, die Hydraulic Fracturing-Projekte mit sich bringen, bietet Data Science Techniken können verwendet werden, um die vielversprechendsten Projekte auszuwählen. Insbesondere maschinelles Lernen Algorithmen können verwendet werden, um die Bohrlochleistung vorherzusagen und zu optimieren.

Das Ziel dieser Arbeit ist es, die Öl- und Gasproduktion mehrerer Multi-Frakturen horizontale Brunnen unter Verwendung unterschiedliche Modelle für maschinelles Lernen vorherzusagen. Diese Modelle werden mit der Leistung anderer Brunnen trainiert, die bereits im selben Gebiet gebohrt und ausgebeutet wurden. Das vielversprechendste Projekt kann daher ausgewählt werden.

## **Abstract**

The oil and gas production from shale oil and shale gas reservoirs has increased rapidly over the last two decades. The combination of horizontal wells and hydraulic fracturing was one of the main reasons shale production became profitable. Multi-fractured horizontal wells have emerged as an advanced mean for enhancing well productivity in low permeability reservoirs. However, the economic viability of such projects is dependent on multiple parameters, especially with the oil price fluctuations over the last decade. Therefore, choosing the projects with the highest potential is essential to maximize the return on investment.

To reduce the economic risks that hydraulic fracturing projects present, data science techniques can be used to choose the most promising projects. In particular, machine learning algorithms can be used to predict and optimize the well performance.

The objective of this thesis is to predict the oil and gas production of several multi-fractured horizontal wells, using different machine learning models. These models will be trained using the performance of other wells that were already drilled and exploited in the same area. The most promising project can therefore be selected.

## Table of Content

	Page
<b>1 INTRODUCTION.....</b>	<b>1</b>
<b>2 UNCONVENTIONAL RESERVOIRS.....</b>	<b>2</b>
2.1 Market Review .....	2
2.2 Hydraulic Fracturing.....	6
2.3 Environmental Challenges .....	15
<b>3 DATA SETS AND TOOLS USED.....</b>	<b>16</b>
3.1 The Permian Basin .....	16
3.2 Data sets.....	18
3.3 Python Tools and Libraries .....	20
<b>4 DATA WRANGLING AND EXPLORATORY DATA ANALYSIS.....</b>	<b>22</b>
4.1 Completeness of the Data and Data Types .....	22
4.2 Data Wrangling .....	23
4.3 Exploratory Data Analysis and Data Preparation.....	35
<b>5 MODEL CREATION AND RESULTS.....</b>	<b>52</b>
5.1 K-Fold Cross-Validation .....	52
5.2 Linear Regression.....	53
5.3 Decision Tree Regression.....	56
5.4 Random Forest Regression .....	59
5.5 Support Vector Machine Regression .....	60
5.6 Feature Importance.....	62
5.7 Final Results .....	65
<b>6 CONCLUSION .....</b>	<b>68</b>
<b>7 PUBLICATION BIBLIOGRAPHY .....</b>	<b>70</b>
<b>LIST OF TABLES .....</b>	<b>73</b>
<b>LIST OF FIGURES.....</b>	<b>75</b>
<b>ABBREVIATIONS.....</b>	<b>77</b>
<b>NOMENCLATURE .....</b>	<b>78</b>

# 1 Introduction

Unconventional reservoirs have become one of the main source of hydrocarbons in the world. Shale oil and shale gas reservoirs are the most exploited unconventional reservoirs. The market share of these reservoirs has been steadily increasing in the past years. Their importance is expected to increase even more since the technologies used to extract oil and gas from these reservoirs are improving continuously. However, unconventional reservoirs typically cost more to produce than conventional reservoirs.

The Permian basin has emerged as one of the highest producing fields in the world. The improvement in hydraulic fracturing technologies has meant that this field, which contains oil and gas producing shale formations, has reached high peaks in production over the last decade. The Permian basin is expected to continue growing, and the increasing number of new wells drilled also means that the amount of data available from this field is increasing, both in quantity and in quality.

With fluctuating oil prices, shale oil and gas exploitation can be too costly, and sometimes can represent a dangerous investment. Wells drilled to produce from shale formations are typically horizontal, which are more expensive to drill. Besides, they need to be hydraulically fractured to produce hydrocarbons. These two processes can increase the costs, and combined with low oil prices, can mean that such projects can end up losing money.

Multi-fractured horizontal wells are used to produce the largest possible volumes of hydrocarbons from shale formations. There are many methods and elements that can be used to fracture such wells. The choice of these elements, like the proppant type and the fracturing fluid, have a very high impact on the oil and gas production of these wells. The number of stages is also a parameter that needs to be determined. All these parameters, combined with the well and the formation parameters, directly influence the production, and therefore the return on investment of a multi-fractured horizontal well project. However, the increase of data provided by new wells and old wells means that relationships between these parameters and the oil and gas production of multi-fractured horizontal wells can be explored.

Data science techniques offer a way to use the data from already drilled wells to predict the performance of potential new wells. The amount of data produced by oil fields offers the chance to avoid risky projects and to focus on drilling wells with high potential. Predicting the production of new wells based on the performance of wells in the same area can reduce the risk factor associated with shale oil and shale gas projects.

The objective of this thesis is to predict the oil and gas production of some potential wells in the Permian basin area. The prediction will be based on the performance of wells already drilled and exploited. The data from these wells will be cleaned and prepared. In addition, the relationship between hydraulic fracturing data, well data, formation data, location data and oil and gas production data will be analysed. Finally, machine learning models will be trained and evaluated. The best performing models will be used to predict the oil and gas production of the target wells.



## 2 Unconventional Reservoirs

An unconventional reservoir is a reservoir with ultra-low permeability. Unconventional reservoirs are increasingly becoming a major source of oil and gas production in the world. The market share of unconventional oil and gas, especially shale gas and shale oil, is increasing. While production is mainly expanding in the United States, a lot of countries around the world have high shale oil and shale gas potential. However, hydraulic fracturing is required to produce from these reservoirs and to increase and maintain well productivity. Hydraulic fracturing refers to the process of pumping a fluid into a wellbore at high injection rate that causes the formation to fracture. These fractures enable production from low permeability formations. This process requires a very large volume of water, which can cause environmental and technical challenges. These challenges are mainly induced seismicity and water contamination.

This chapter discusses unconventional reservoirs. The first part of this chapter discusses the market share and potential of unconventional resources, especially shale oil and shale gas. The second part discusses hydraulic fracturing and the challenges it represents. The third part discusses some environmental challenges related to hydraulic fracturing.

### 2.1 Market Review

Most of the oil produced today comes from conventional reservoirs. These resources generally accumulate in favourable structural or stratigraphic traps that can be easily extracted. The conventional petroleum system is consisted of 4 essential elements (Source Rock, Reservoir Rock, Seal Rock and Overburden Rock) and processes (Trap formation, Generation-Migration-Accumulation). The conventional reservoir formations are porous and permeable but are sealed by a low permeability formation that prevents the hydrocarbons from escaping. In conventional reservoirs, no large-scale stimulation is needed to be able to produce.

On the other hand, unconventional resources are more abundant but more difficult to exploit. There are a variety of formations that are considered unconventional, including oil shales, tight gas sands, coalbed methane and gas hydrates. From a characteristics point of view; unconventional reservoirs generally present low to ultra-low permeability (generally below 1 millidarcy) and low to moderate porosity. Some unconventional reservoirs also contain high viscosity oil. This is the main reason why extracting from unconventional reservoirs is more difficult and requires different extraction techniques. These techniques differ depending on the challenges presented by the reservoir. For low permeability formations like tight oil, tight gas, gas shales and coalbed methane reservoirs, using horizontal wells and multistage hydraulic fracturing is the best method to produce economically. For formations that contain high viscosity and heavy oil, heat is used to overcome the challenge. Gas hydrate reservoirs still present a problem and new techniques are being evaluated to make extraction profitable.

The relative abundance of conventional and unconventional resources is best described by the resource triangle illustrated in Figure 1. The concept of the resource triangle is that natural resources, such as gold, silver, uranium, oil, and gas are distributed log normally in nature

(Holditch 2013). The high-quality deposits are small and difficult to find but easier to extract. A pure vein of gold is very rare to find, but once found extracting it is easy. As you go lower in the triangle, the resources become larger but more difficult to produce. The technical difficulties and bigger investments needed can be compensated by the abundant volumes found. The resource triangle concept should be applicable for each basin where oil is produced. It is therefore possible to assume that any oil or gas basin in the world that has been producing oil and gas from conventional reservoirs should have larger quantities of hydrocarbons in unconventional formations. Figure 1 shows that the high and medium quality reservoirs are less present in nature, while unconventional reservoirs are much more abundant but also more costly to exploit, especially gas hydrates and oil shales.

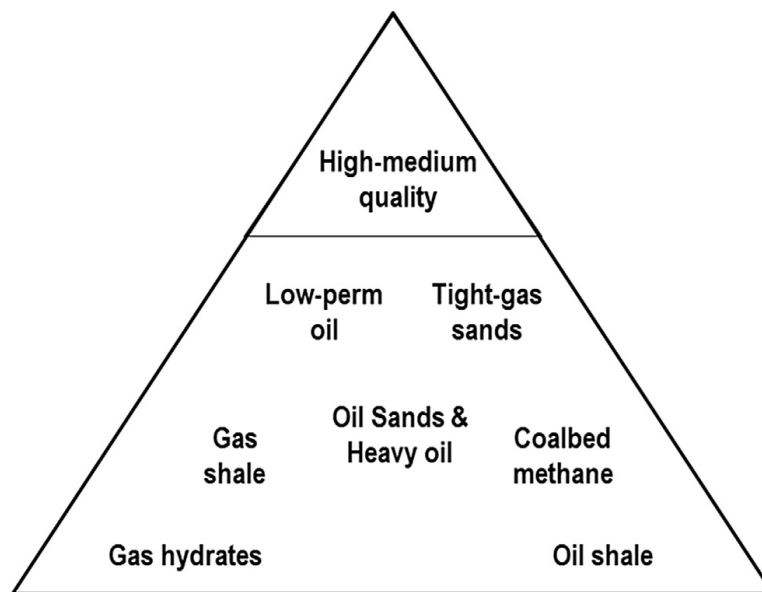


Figure 1: Resource Triangle<sup>1</sup>

Oil and Gas Production from unconventional reservoirs has seen a great increase since the start of the 2010s. With more reserves discovered and the decline of production from conventional reservoirs, it is expected that unconventional reservoirs will become the main source of oil and gas production in many countries. The main example is the United States, where tight oil, tight gas and shale gas are already the biggest source of hydrocarbon production. The steady growth is expected to continue through 2050. Figure 2 shows the history and projection of crude oil and dry gas production in the United States. It indicates that the majority of oil and gas will be produced from tight oil and tight gas formations.

The main unconventional resources being produced currently are shale gas and shale oil reservoirs. This type of shale production is typically conventional oil or gas that is produced from deeply buried shales. Shale has long been considered in the conventional petroleum system as source rock or seal due to its low permeability. However, the Barnett Shale play

---

<sup>1</sup> Holditch 2013.

demonstrated that shales can be considered as reservoirs. Shales are the most abundant sedimentary rock formation, but there are some important criteria that make a shale formation exploitable as a reservoir. These parameters describe the reservoir quality and the completion quality. Reservoir quality describes the hydrocarbon potential, volumes in place and deliverability. Completion quality describes the ability to create and maintain fracture surface area (Ma et al. 2015).

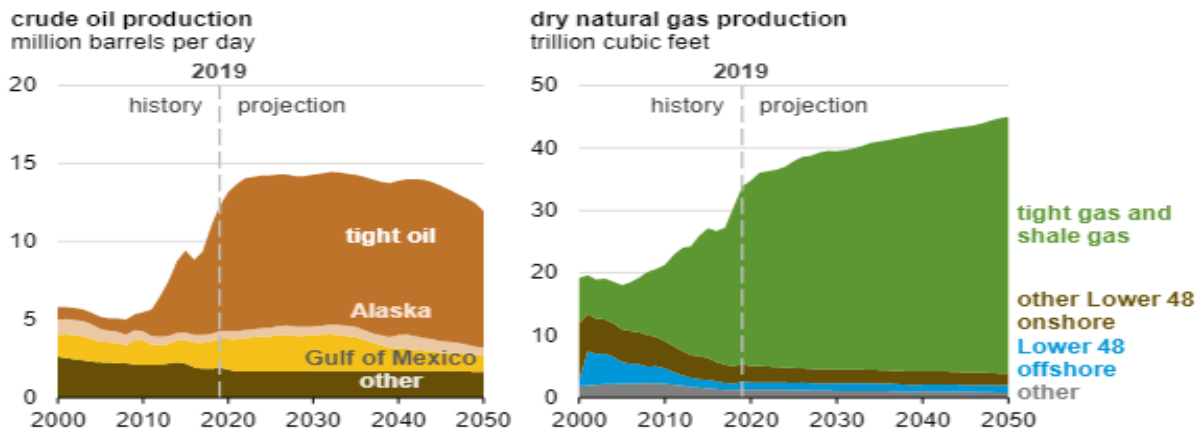


Figure 2: U.S. Crude Oil and Dry Natural Gas Production through 2050<sup>1</sup>

Table 1 shows the parameters for source rock evaluation. These parameters include geochemistry, geology, petrophysics and mineralogy data. Favourable ranges for shale formations have been identified after years of production. Total Organic Carbon (TOC) should be above 2%. For Vitrinite Reflectance (%Ro), the oil generation window is between 0.7% and 1.0%, the dry gas generation window is between 1.0% and 1.3% and the wet gas generation window is above 1.3%. Vitrinite Reflectance below 0.7% indicates immature rocks and values over 1.3% indicate over maturity. The favourable depth for shale oil and shale gas reservoirs is between 3,300 ft and 16,500 ft (Ashayeri and Ershaghi 2015).

Table 1: Source Rock Evaluation Parameters<sup>2</sup>

Geochemistry	Geology	Petrophysics	Mineralogy
TOC	Shale thickness / depth Gamma Ray /	Porosity	Clay content
%Ro	Resistivity Deposition	Permeability Young's	Water sensitivity
Tmax	Environment	Modulus	Quartz content
Kerogen Type	Seals / Barriers	Poison Ratio	Carbonate / Shale

<sup>1</sup> U.S. Energy Information Administration 5/31/2020.

<sup>2</sup> Ashayeri and Ershaghi 2015.

Shale gas and shale oil reserves have been discovered around the world. These discoveries have contributed to a big increase in worldwide proved and unproved oil and gas resources in the world. According to the U.S. Energy Information Administration (EIA), 345 billion barrels (bbl) of shale oil technically recoverable resources (TRR) and 7,299 trillion cubic feet (Tcf) of natural gas TTR exist around the world. In the U.S., Shale represents 9% of the total oil resources and 32% of the total gas resources. Outside the U.S., the main countries with recoverable shale oil resources are Russia, China, Argentina, Libya, and Venezuela. For recoverable shale gas resources, the main countries are China, Argentina, Algeria, and Canada. In total, Shale represents 10% of the recoverable oil and 32% of the recoverable gas. Table 2 summaries the proved and unproved shale and oil resources in the world.

Table 2: Technically Recoverable Shale Oil and Shale Gas Unproved Resources in the Context of Total World Resources<sup>1</sup>

	<b>Crude oil</b>	<b>Wet natural gas</b>
	<b>(billion barrels)</b>	<b>(trillion cubic feet)</b>
<b>Outside the United States</b>		
Shale oil and shale gas unproved resources	287	6,634
Other proved reserves	1,617	6,521
Other unproved resources	1,230	7,296
<b>Total</b>	<b>3,134</b>	<b>20,451</b>
<b>Increase in total resources due to inclusion of shale oil and shale gas</b>	<b>10%</b>	<b>48%</b>
<b>Shale as a percent of total</b>	<b>9%</b>	<b>32%</b>
<b>United States</b>		
EIA shale / tight oil and shale gas proved reserves	n/a	97
EIA shale / tight oil and shale gas unproved resources	58	567
EIA other proved reserves	25	220
EIA other unproved resources	139	1,546
<b>Total</b>	<b>223</b>	<b>2,431</b>
<b>Increase in total resources due to inclusion of shale oil and shale gas</b>	<b>35%</b>	<b>38%</b>
<b>Shale as a percent of total</b>	<b>26%</b>	<b>27%</b>
<b>Total World</b>		
Shale / tight oil and shale gas proved reserves	n/a	97
Shale / tight oil and shale gas unproved resources	345	7,201
Other proved reserves	1,642	6,741
Other unproved resources	1,370	8,842
<b>Total</b>	<b>3,357</b>	<b>22,882</b>
<b>Increase in total resources due to inclusion of shale oil and shale gas</b>	<b>11%</b>	<b>47%</b>
<b>Shale as a percent of total</b>	<b>10%</b>	<b>32%</b>

<sup>1</sup> Shale oil and shale gas resources are globally abundant - Today in Energy - U.S. Energy Information Administration (EIA) 5/30/2020.

Shale formations have very low permeability. For this reason, hydraulic fracturing is used to stimulate the formation to be able to produce the hydrocarbons it contains. Hydraulic fracturing is one of the main technologies that allowed the exploitation of unconventional oil and gas resources. However, hydraulic fracturing raises some concerns regarding the potential impact on the environment.

## 2.2 Hydraulic Fracturing

Hydraulic Fracturing (HF) is a process that involves pumping a fluid composed generally of water, propping agents and specific chemicals at a very high rate and pressure to break the rock containing the hydrocarbons. Without HF, shale reservoirs would not be able to produce at an economic rate. Figure 3 shows the increase in production and ultimate recovery that is allowed with HF for low permeability reservoirs. It shows that the economic limit is significantly extended with HF. The ultimate recovery is also multiplied with HF.

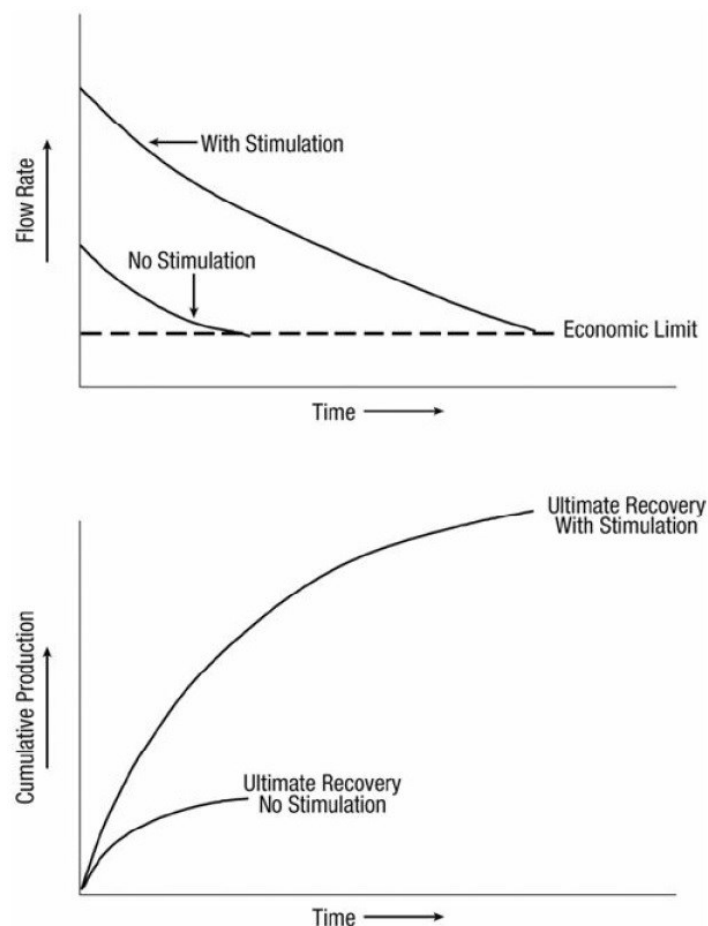


Figure 3: Production and Reserves Enhancement from HF for Low Permeability Reservoirs<sup>1</sup>

The amount of water injected is in the order of several million gallons. The large pressure associated with the injection of the fluid creates new fractures and extends existing fractures,

<sup>1</sup> Economides 2013.

which increases the production. Sand is typically used as a propping agent. The purpose of the propping agent is to hold the new and existing fractures open. The injection fluid that flows back can be reinjected to the reservoir (Aminzadeh 2020). Figure 4 shows a typical hydraulic fracturing operation. As indicated in the figure, hydraulic fracturing is typically done in horizontal wells. The combination of horizontal wells and hydraulic fracturing was one of the main reasons shale production became profitable (Smith and Montgomery 2015). The first step of production from shale formations is to hydraulically fracture the formation, which then allows for gas or oil production.

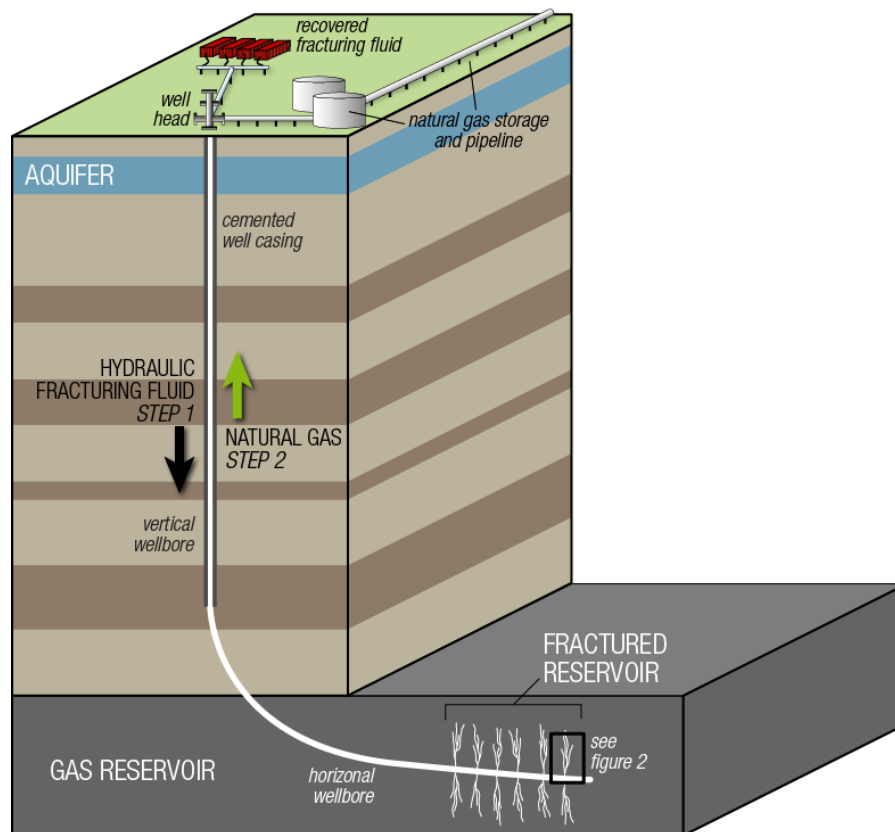


Figure 4: Schematic of a Typical Hydraulic Fracturing Operation<sup>1</sup>

The rapid increase in shale production has been accompanied with a larger use of hydraulic fracturing. According to the US EIA, 95% of the new wells drilled in the US in 2016 were hydraulically fractured (Aminzadeh 2020). The relative high cost of the HF means that its use is correlated to the oil price. Producing shale oil can cost more than 60\$/bbl. In fact, periods where oil prices have dropped saw a decline in shale related activities. The combination of HF with horizontal drilling has permitted the drilling of multiple wells from the same surface location, reducing the footprint of such projects above ground by 90% (Aminzadeh 2020).

<sup>1</sup> Hydraulic Fracturing: An Indiana Assessment 2020.

Designing the hydraulic fracturing job is essential to maximize the hydrocarbon production from the well. Hydraulic fracturing parameters are dependent on well and formation data. The most important design parameters are fracturing fluids and proppants.

### **2.2.1 Data Requirement for Hydraulic Fracturing**

The first step in designing a hydraulic fracturing treatment is to have a valid data set of different parameters needed. These parameters can be divided into reservoir data, log data, geologic data, and fracturing data (Smith and Montgomery 2015). Reservoir data needed includes porosity, permeability, reservoir pressure and temperature, reservoir fluid properties, drainage area and fluid sensitivity. Log data needed is deviation data, especially for horizontal wells, lithology, porosity, resistivity, and sonic logs. These logs can be used to determine basic properties of the formation. Geological Data needed include natural fractures and stress orientation.

Designing a hydraulic fracturing job requires information about fracture height, fracture width, fluid loss and fracture tip effects (Smith and Montgomery 2015). Fracture height is controlled by in situ stresses, while fracture width is controlled by elastic modulus of the rock. Fluid losses are dependent of the fracturing fluid parameters and reservoir parameters already discussed. Fracture tip effects relate to the formation breakdown pressure, which is the pressure required to propagate the fracture tip. Fracturing data needed includes formation young's modulus, in situ stress and fracture toughness. In addition to reservoir, geologic, log and fracture data, designing the fracturing fluid is required.

### **2.2.2 Fracturing Fluid**

The fracturing fluid is the fluid injected during the hydraulic fracturing process. In the beginning, the fluid injected does not contain propping agents. This fluid is called the pad. When the fractures are wide enough to accept proppants, they are added to the fracture fluid. Designing the fracturing fluid consists of choosing the right pad volume to provide the fracture geometry needed, and the right viscosity and density so that the fluid can be used properly. The density of the fluid should be around 8.4 pounds per gallon for water-based fluids. Viscosity is the most important design criteria of the fracturing fluid. The fluid viscosity should:

- Allow the fluid to have a good clean-up behaviour to maximize the fracture conductivity.
- Allow the fluid to provide fractures that are wide enough for the proppant to enter.
- Allow the fluid to transport the propping agent from the wellbore to the tip of the fractures.
- Prevent fluid losses into the formation.

The viscosity of the fluid should be in the range of 50 to 1000 centipoise to create a fracture width of 0.2 to 1.0 inches and transport the propping agent for distances of hundreds to thousands of feet (Lake and Fanchi 2006-2007). In addition to the volume and viscosity requirements, many other factors are important when designing the fracturing fluid. These factors include:

- The fluid should be safe and environmentally friendly to limit the risk of harm to the personnel and the environment in case of a leak off. The fluid should also be compatible with the formation, so it does not react with the formation minerals or fluids.
- The fluid should break to a low viscosity to be able to flow back and clean up the fracture.
- The fluid should be easy to mix and cost effective.

These factors should all be considered for the design of the fluid. In the ideal case, the fracturing fluid designed is compatible with the formation rock and fluid, can generate a wide fracture, can transport the propping agent in the fracture, break to a low viscosity fluid for clean up and be cost effective and environmentally friendly. Compromises always need to be made since generally cost is the limiting factor. There are a lot of types of fracturing fluids that can be used. These fluids differ in their characteristics, costs, and impact on the environment. For most reservoirs, water-based fluids with some additives are the best choice. It is important to control the quality of the water used when using these types of fluids. It is also possible to use oil-based fluids or foams. The most commonly used fracturing fluids are water frac, linear gel, cross-linked gels, oil-based fluids, and foams/poly-emulsions (Smith and Montgomery 2015). Slickwater refers to the use of low-viscosity fluids pumped at high rates to generate narrow, complex fractures with low-concentrations of propping agent. When uncross-linked gels are used in late-slurry stages of a fracturing treatment, they are often referred to as "hybrid" fracturing treatments (PetroWiki 2020).

- **Water Frac:** Composed of water, clay control agent and a friction reducer, with the possible addition of a water recovery agent. This mixture presents low viscosity, so the transport mechanism of proppants has to be the velocity. This means that water frac is generally pumped at very high rates. Besides, the fracture width is low due to the low viscosity. However, the main advantages are the lower cost and the ease of mixing. Besides, the water can be recovered and reused.
- **Linear Gel:** Composed of water, clay control agent and gelling agent, with the possible addition of bactericides or biostats, chemical breakers and water recovery agents. The linear gel has improved but still relatively low viscosity characteristics. This means that, similarly to water frac, the fracture width is narrow. The cost is also low. However, the big disadvantage is that the water cannot be reused because it has residual breaker.
- **Cross-Linked Gels:** Composed of water, clay control agent, gelling agent, and cross-linker, with the possible addition of bactericides or biostats, chemical breakers and water recovery agents. The cross-linker is used to significantly increase the viscosity, which can go from 50 centipoises to 100s or 1000s of centipoises. The higher viscosity offers multiple advantages. It increases the fracture width and thus higher proppant concentrations can be used. Other advantages include better proppant transport, improved fluid efficiency, and reduced friction pressure.



- **Oil-Based Fluids:** Composed of Oil, gelling agent, and cross-linker. Oil-based fluids are used in water-sensitive formations that can be damaged if water-based fracturing fluids are used. The disadvantages of using these fluids are the high cost when using refined oils. Besides, these oils have to be taken from the refineries before additives are added. If crude oils are used, gelling problems can occur. Finally, these fluids can present safety issues for the personnel and can have a high environmental impact compared to water-based fluids.
- **Foams/Poly-emulsions:** Composed of water and a material that is not miscible with water, like nitrogen, carbon dioxide or a hydrocarbon such as propane, diesel, or condensate. The main concern with these fluids is the safety aspect since the fluids are pumped at high pressure and they contain gas or flammable fluids. The cost is also high for these fluids, and sometimes the gases needed for these mixtures are not available in remote areas. On the other hand, the advantages are numerous. These fluids are very clean, fluid loss is minimal, and proppant transport is good. The viscosity is controlled by changing the ratio of the gas or hydrocarbon used and the water.

In addition to these base fluids, some additives may be needed in order to improve the overall efficiency of the hydraulic fracturing job. The use of these additives is dependent on the fluid system. Additives are transported in concentrated form and diluted when pumped. Table 3 offers a summary of different chemical additives types, functions, and products. It shows that these additives can be used to kill bacteria, reduce viscosity, reduce friction, etc.

Table 3: Fracturing Fluids Chemical Additives<sup>1</sup>

Type of Additive	Function Performed	Typical Products
Biocide	Kills bacteria	Gluteraldehyde carbonate
Breaker	Reduces fluid viscosity	Acid, oxidizer, enzyme breaker
Buffer	Controls the pH	Sodium bicarbonate, fumaric acid
Clay stabilizer	Prevents clay swelling	KCl, NHCl, KCl substitutes
Diverting agent	Diverts flow of fluid	Ball sealers, rock salt, flake boric acid
Fluid loss additive	Improves fluid efficiency	Diesel, particulates, fine sand
Friction reducer	Reduces the friction	Anionic copolymer
Iron Controller	Keeps iron in solution	Acetic and citric acid
Surfactant	Lowers surface tension	Fluorocarbon, Non-ionic
Gel stabilizer	Reduces thermal degradation	MEOH, sodium thiosulphate

---

<sup>1</sup> Lake and Fanchi 2006-2007.

When looking at the overall composition of the injected fracturing fluid, it typically contains 90% water, 9.5% proppant materials and 0.5% chemicals. These chemicals are the additives to the fracturing fluid discussed in the table above. Figure 5 shows the overall composition of a typical fracturing fluid. It shows that apart from water and sand, which represent up to 99.5% of the total volume, no other element represents a percentage of more than 0.2% of the total volume.

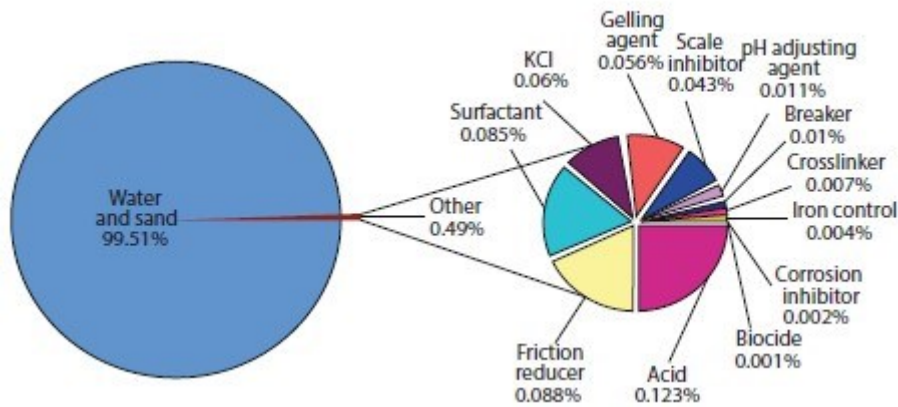


Figure 5: Overall Composition of a Typical Fracturing Fluid<sup>1</sup>

In addition to designing the adequate fracturing fluids base composition and additives, the design of the proppant is an essential design parameter for the hydraulic fracturing job.

### 2.2.3 Proppant

The purpose of hydraulic fracturing is to create fractures in the formation in order to increase production. The width profile and the fracture height area are affected by the fracturing fluid volume and properties. However, once fluid pumping is stopped, the fractures will close because of pressure loss. To avoid this, a material is included in the fracturing fluid to keep the fractures open once pressure drops. This material is the propping agent. The main design parameters when choosing the adequate proppant is the proper grain size and proppant type. The ideal propping agent is readily available, has a low cost, a low density, and a high resistance to corrosion and to crushing.

Proppants have differences in cost, availability, specific gravity, strength, and stress handling. Material strength should allow the proppant to withstand high closure pressures. Closure pressure is the pressure at which the fracture closes (Belyadi et al. 2016b). The required strength is determined by calculating the effective stress. The effective stress is defined as the difference between the formation closure and the bottom hole flowing pressure (Ma et al. 2015). An increase of the effective stress results in a decrease in fracture conductivity. The reason is that grain failure can create small fines that migrate and reduce the permeability. Typically, materials that can handle higher closure pressures have a higher specific gravity,

<sup>1</sup> Aminzadeh 2020.

which means that they are more likely to settle faster. In order to compensate for the fast settling, the fracturing fluid should be designed to allow for the proper carrying of these proppants.

The shape of the grains is also an important factor. The shape can be described by the roundness and sphericity. Roundness is a measure of the sharpness of the corners in the grain, while sphericity is a measure of how closely the grain approaches the shape of a sphere. As shown in figure 6, the shape improves with higher sphericity and roundness.

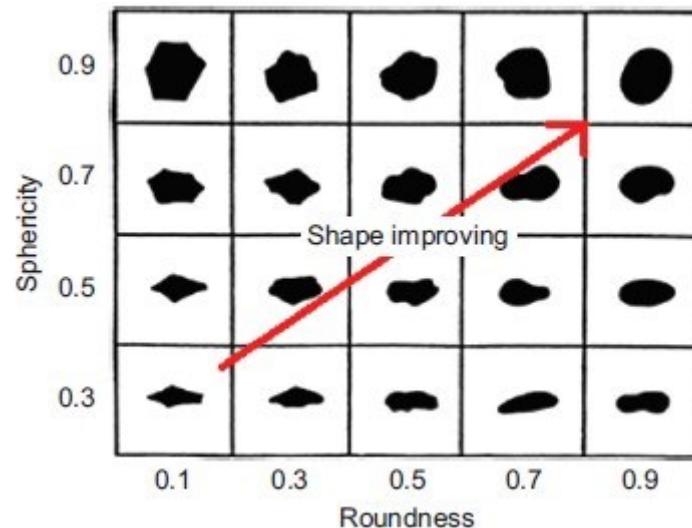


Figure 6: Estimation of Roundness and Sphericity of a Grain<sup>1</sup>

The types of proppant typically used are conventional sand, resin-coated sand, and ceramic proppants.

- **Sand:** is the cheapest and most available proppant. It has the lowest strength of all proppant types. It can withstand closure pressure up to 6000 psi. There two types of sand generally used: Ottawa sand and Brady sand. Ottawa (also known as white sand) is more expensive but is of higher quality compared to Brady (also known as brown sand). Sand typically has an irregular shape and size. Specific gravity of sand is 2.65.
- **Resin-Coated Sand:** is more expensive than sand. This type of proppant has intermediate strength. It is created by adding resin coating to sand in order to have a higher conductivity compared to normal sand. It is generally used for closure stress of 6000 to 8000 psi. Specific gravity is from 2.55 to 2.60.
- **Ceramic Proppant:** is the best quality proppant available. It is also the most expensive option. It presents a uniform shape and size. Besides, it is more thermally resistant and

<sup>1</sup> Belyadi et al. 2016b.

has a very high crush resistance. Ceramic Proppants can be divided to lightweight, intermediate strength, and high strength. Lightweight ceramic proppant can withstand pressure of 6000 to 10000 psi and has a specific gravity of around 2.72. Intermediate strength ceramic proppant can handle pressure between 8000 and 12000 psi and has a specific gravity of 2.9 to 3.3. High strength ceramic proppant is sintered bauxite. It is the strongest type of proppants used in the industry and can handle pressure of up to 20000 psi. It is generally used in deep high-pressured formations. Specific gravity is 3.4 or more.

Table 4 summarizes different proppant types and their characteristics. It shows that regular sand is the cheapest option, ceramic proppant is the highest quality proppant, while resin-coated sand offers a good compromise of cost and quality.

Table 4: Proppant Types Summary<sup>1</sup>

<b>Regular Sand</b>	<b>Resin-Coated Sand</b>	<b>Ceramic Proppant</b>
Cheapest	More expensive (compared to regular sand)	Most expensive
Lowest conductivity	Medium conductivity	Highest conductivity
Lowest strength	Medium strength	Highest strength
Irregular size and shape	Irregular size and shape	Uniform size and shape
Naturally occurring product	Manufactured product	Engineered and manufactured product

Proppants are used to prevent fractures from closing after the fracking job is finished. However, the proppant does not reach the whole length of the fracture. This means that unpropped areas will close with time and lose their conductivity. Figure 7 shows the difference between created fracture dimensions and propped fracture dimensions. It demonstrates that propped fracture dimensions represent a fraction of the created fracture dimensions. Proppant size is another important design parameter of proppants and has a big influence on propped fracture dimensions and conductivity. Figure 8 shows the effect of a smaller (40/70) and a bigger (20/40) mesh size on the fracture conductivity and the propped fracture length.. It shows that using smaller mesh proppants allows the particles to travel further inside the fracture which results in a longer propped fracture length. However, since the grains are smaller, the space between them is smaller so the fracture conductivity is lower.

---

<sup>1</sup> Belyadi et al. 2016b.

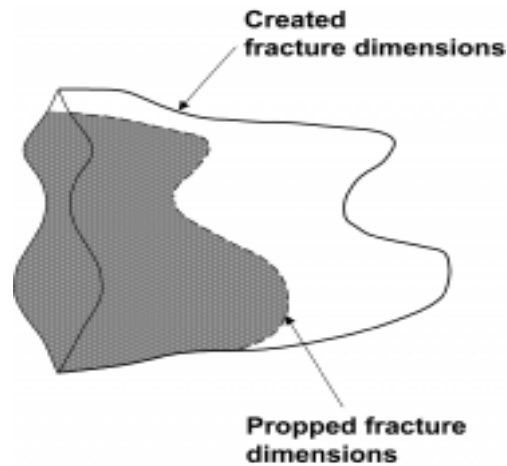


Figure 7: Difference between Created and Propped Fracture Dimensions<sup>1</sup>

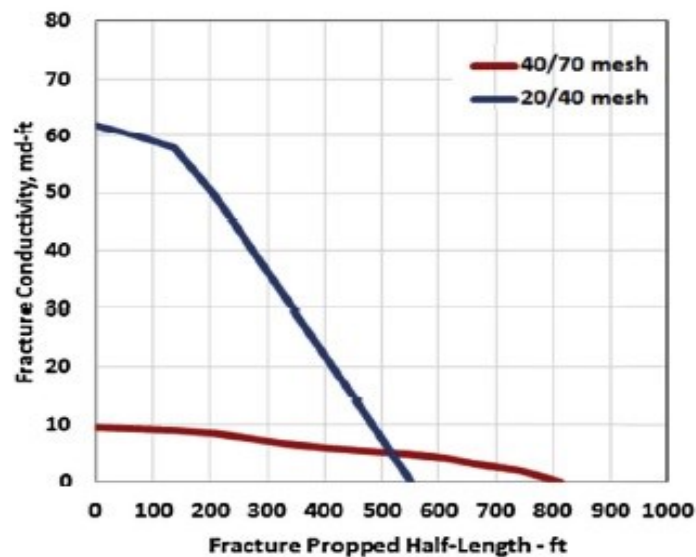


Figure 8: Effect of Mesh Size on Fractures<sup>2</sup>

Different mesh sizes are used in the hydraulic fracturing industry. The most commonly used are 100 mesh, 40/70 mesh, 30/50 mesh, and 20/40 mesh (Belyadi et al. 2016a). Most hydraulic fracturing jobs use a combination of these types.

- 100 Mesh:** is the smallest mesh size and is designed to be placed in hairline cracks of the formation. It can be used at the start of the operation to seal off microfractures and perforation erosion, to decrease leak-off and to provide a conduit for the upcoming sands.

<sup>1</sup> Lake and Fanchi 2006-2007.

<sup>2</sup> Ma et al. 2015.

- **40/70 Mesh:** is larger in size than 100 mesh. Using this mesh size creates a high fracture length and some conductivity. Using smaller mesh sizes like 100 and 40/70 provides a higher crush resistance since the stress is distributed on a bigger number of grains compared to bigger mesh sizes.
- **30/50 Mesh:** is larger than 40/70. The conductivity when using a bigger mesh size is better compared to a small mesh size.
- **20/40 Mesh:** is the largest sand size used. It is used to maximize near wellbore conductivity.

## 2.3 Environmental Challenges

The practice of hydraulic fracturing has been linked with concerns related to its impact on the environment. In general, the potential environmental impact of hydraulic fracturing can be divided into 3 categories: Impact on water cycle, air pollution and induced seismicity. The impact on the water cycle can be a result of these activities (U.S. Environmental Protection Agency 2016):

- Water withdrawals for hydraulic fracturing use in areas where groundwater resources are scarce.
- Spills of chemicals, hydraulic fracturing fluids or produced water that results in chemical substances reaching groundwater resources.
- Well integrity problems that result in contamination of groundwater resources from the injected fluid.
- Injection of fracturing fluids directly into groundwater resources.
- Discharge of treated or produced fracturing fluid in surface water resources or disposal of wastewater in unlined pits that result in contamination of groundwater resources.

Other risks to the environment include (Ahmed and Meehan 2016):

- Release of Greenhouse gases into the atmosphere.
- Micro-seismic events.
- Naturally occurring radioactive materials brought to the surface.
- Generally greater footprint and noise pollution than conventional hydrocarbon projects.

Water acquisition and disposal are one of the biggest challenges faced by hydraulic fracturing projects. However, the proper handling of the fluids used is essential in order to reduce the impact on the environment. Disclosure of the chemicals used, the water use, and its origin is also important to provide nearby habitants and authorities with sufficient information and to address their concerns. Seismic effects are generally less than minus 2 or minus 3 on the Richter scale during hydraulic fracturing (Speight 2016). In general, all these risks can be minimized by following the best practices from the industry. Standards related to well construction, spill and leak reduction and containment, water and waste disposal must be followed.

### 3 Data Sets and Tools Used

The objective of this thesis is to use data science and machine learning techniques to predict the oil and gas production from several multi-fractured horizontal wells in the Permian Basin. The data set consists of several wells with a description of the hydraulic fracturing stages, relevant well information, cumulative production, etc. The information from these wells will be used to predict the production from target wells with different attributes. Several models will be created based on the most widely used machine learning techniques. These models will then be evaluated and used to determine the oil and gas output of the target wells. The data is publicly available from the Texas Railroad Commission website. The project will be conducted using Python.

This chapter discusses the importance of the Permian basin, as well as the data sets and the tools used. The first part gives an overview on the Permian basin and the location of the wells used in this thesis. The second part presents a summary of the data sets of the thesis. The third part is a small description of the different python tools and libraries used.

#### 3.1 The Permian Basin

The Permian Basin is an oil-and-gas-producing area located in West Texas, as shown in figure 9. The Permian Basin covers an area approximately 250 miles wide and 300 miles long and is composed of more than 7,000 fields. Various producing formations such as the Yates, San Andres, Wolfcamp, etc are all part of the Permian Basin. The oil and natural gas production depths range from a few hundred feet to five miles below the surface (Railroad Commission of Texas). The wells in this thesis produce from the Wolfcamp formation.

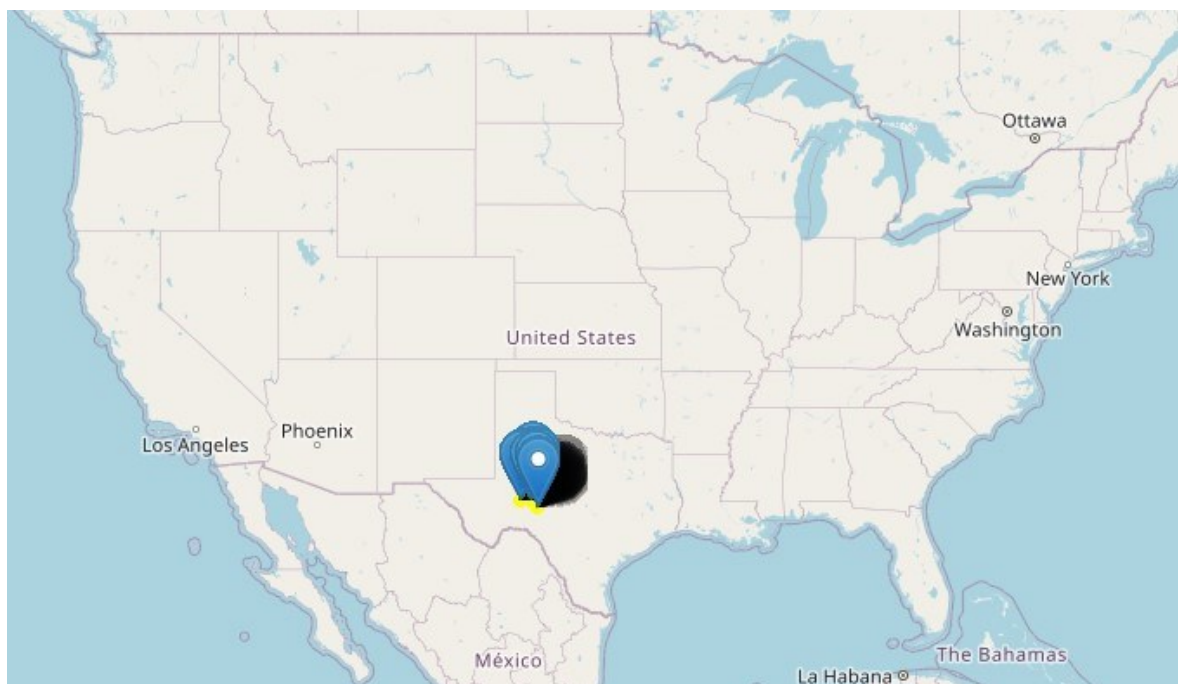


Figure 9: Location of the Permian Basin

The Permian Basin has generated hydrocarbons for about 100 years and supplied more than 33.4 billion barrels of oil and about 118 trillion cubic feet of natural gas as of September 2018. The use of hydraulic fracturing, horizontal drilling, and completion technology advancements during the past decade has reversed the production drop in the Permian, and the basin has exceeded its previous peak in the early 1970s. In 2017, it accounted for 20% of the total U.S. crude oil production and about 9% of the total U.S. dry natural gas production. For 2016, EIA estimates that the remaining proven reserves in the Permian Basin exceed 5 billion barrels of oil and 19.1 trillion cubic feet (Tcf) of natural gas, making it one of the largest hydrocarbon-producing basins in the United States and the world (U.S. Energy Information Administration). Figure 10 shows the increase of oil production in the Texas Permian Basin from 2008 to November 2020.

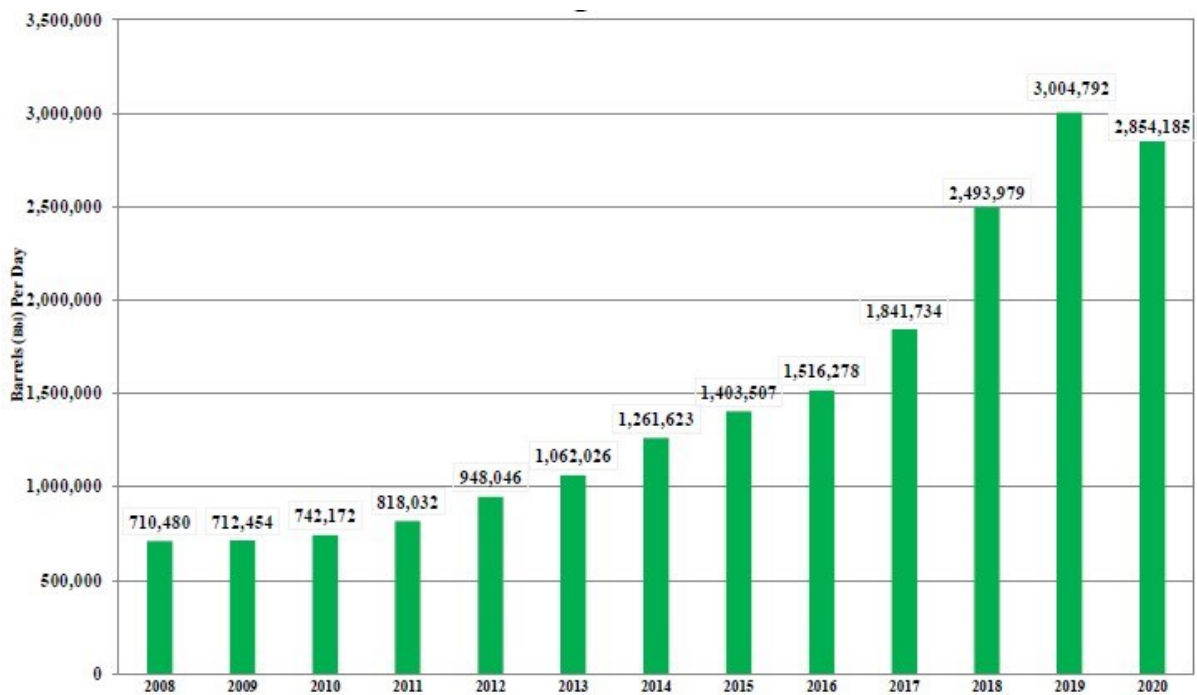


Figure 10: Average Daily Oil Production in the Texas Permian Basin through November 2020<sup>1</sup>

As shown in the figure above, the oil production saw a steady increase through the last decade. Gas and condensate production also follow the same trend. The Permian Basin has the potential to become the world's most productive oil field. The reasons are the big increase in production, the great number of drilled but uncompleted wells and the great volumes of hydrocarbons still left in the formation (Rapier 2018). Since the data provided contains the latitude and longitude coordinates of the wells, it is possible to create a map using Python's Folium library to visualize these wells. Figure 11 shows the wells used in both the training set and the target set.

---

<sup>1</sup> Railroad Commission of Texas.



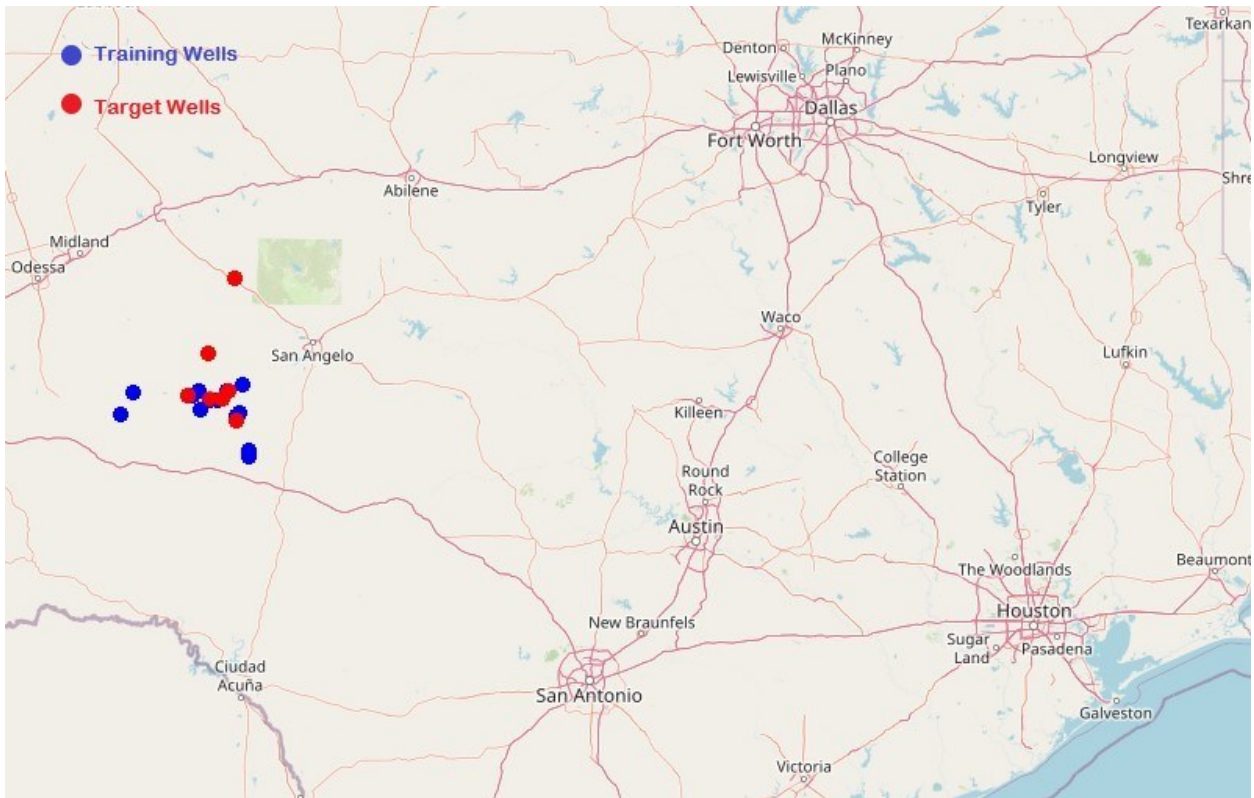


Figure 11: Location of the Wells in Texas

As seen in figure 11, the wells are located near the city of San Angelo in Texas. The wells used in the training data set are shown in blue, while the wells of the target data set are shown in red. An in-depth analysis of the impact of the location of the wells on the oil and gas production will be conducted in the exploratory data analysis chapter.

### 3.2 Data sets

The data sets consist of 27 multi-fractured horizontal wells. Similar information about the wells is given, but the oil and gas production are given for only 20 of these wells. These 20 wells will be used as a training data set for the model to predict the production of the remaining 7 target wells. The location of the wells used is shown in figure 12.

While most of the wells are very close, some wells are relatively far from each other. Most of the wells in the target set have wells from the training set close to them. While the total number of wells is 27, some are very close and therefore not all wells appear in figure 12. Most wells are located near Big Lake and Mertzon. Two training wells are located near Eldorado. Most of the target wells are located in areas that already contain training wells nearby. The only exception is a target well located in Sterling City. The oil and gas production of these wells will be visualized to determine if any relationship between hydrocarbon production and location can be established.

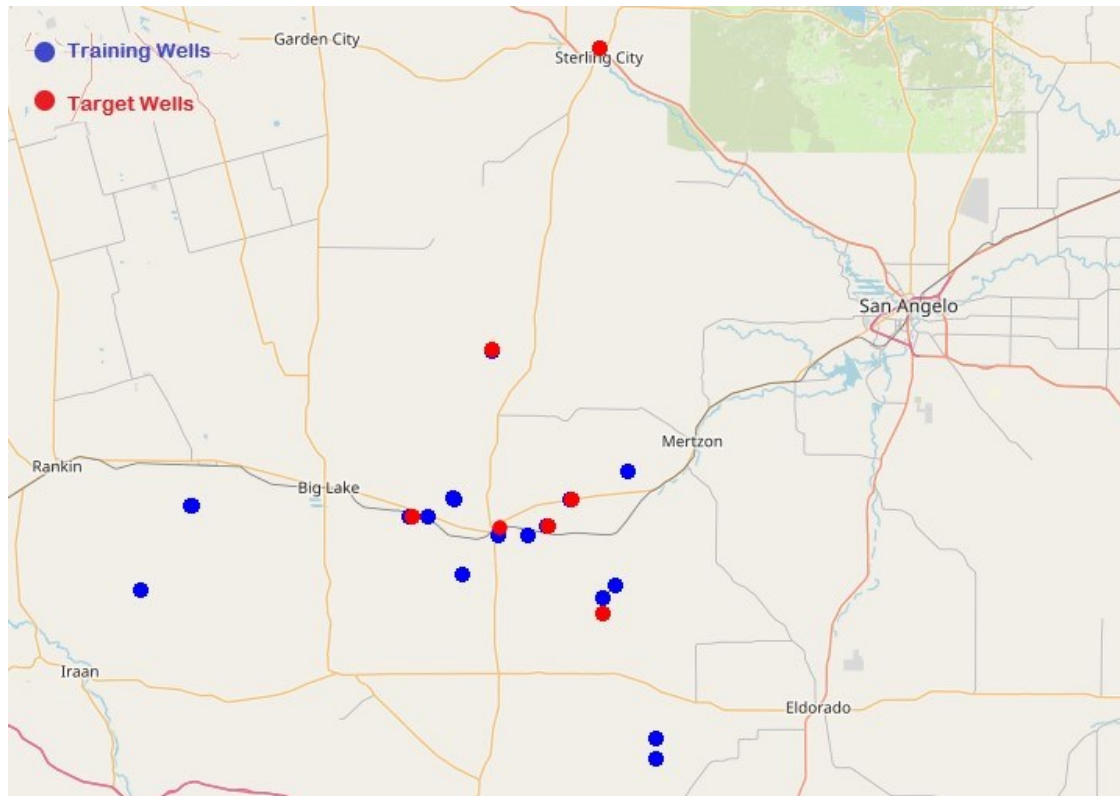


Figure 12: Location of the Training Wells and the Target Wells

The training data set consists of 1179 rows and 28 columns. The target data set consists of 338 rows and 28 columns. The following columns are the same for both data sets:

1. **WELL\_ID:** Contains the well identification number. The training data set is composed of 20 wells [ 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16, 18, 19, 21, 22, 23, 24, 25, 26] while the target data set is composed of 7 wells [ 1, 5, 9, 13, 17, 20, 27].
2. **JOB\_DESC\_STAGING:** Contains information about the hydraulic fracturing staging, the geological formation in which the wells are drilled, the fracturing fluid, the perforation measured depth range, and the day and stage number. Exp: "Day 4 Wolfcamp Frac Slickwater Stg 24".
3. **PROPPANT\_MESH\_SIZE:** Contains information about the mesh size of the proppant used. Also contains information about the type of proppant used in some cases. Exp: "Sand, White, 100 mesh".
4. **PROPPANT\_MESH\_DESCRIPTION:** Contains information about the proppant type used and the mesh size. Exp: "Sand, Brown, 40/70".
5. **PROPPANT\_MASS\_USED:** Mass of proppant used for each fracturing stage.
6. **PROPPANT\_MASS\_UOM:** Unit of measure of proppant mass (**1CWT = 112 lbs**).
7. **VOLUME\_PUMPED\_GALLONS:** Volume of fracturing fluid pumped during each fracturing stage.
8. **AVERAGE\_STP:** Average standard temperature pressure.
9. **AVERAGE\_STP\_UOM:** Unit at standard temperature & pressure (**psi**).
10. **FRACTURE\_GRADIENT:** Fracturing gradient of the formation. This column contains the same value for each well.

11. **FRACTURE\_GRADIENT\_UOM**: Unit of fracture gradient (**psi/ft**).
12. **MD\_MIDDLE\_PERFORATION**: Measured depth of the middle perforation.
13. **MD\_MIDDLE\_PERFORATION\_UOM**: Unit of measured depth (**ft**).
14. **TVD\_DEPTH**: Well true vertical depth. This column contains the same value for each well.
15. **TOP\_DEPTH**: Well measured depth. This column contains the same value for each well.
16. **WELL\_LATITUDE**: Well Latitude. This column contains the same value for each well.
17. **WELL\_LONGITUDE**: Well Longitude. This column contains the same value for each well.
18. **MIN\_STP**: Minimum STP.
19. **MIN\_STP\_UOM**: Unit at standard temperature & pressure (**psi**).
20. **MAX\_STP**: Maximum STP.
21. **MAX\_STP\_UOM**: Unit at standard temperature & pressure (**psi**).
22. **UPPER\_PERF**: Upper perforation location. This column contains the same value for each well.
23. **LOWER\_PERF**: Lower perforation location. This column contains the same value for each well.
24. **TRUE\_VERTICAL\_DEPTH**: True vertical depth. This column contains the same value for each well.
25. **WELL\_HORZ\_LENGTH**: Well horizontal length. This column contains the same value for each well.
26. **NET\_PROD\_DAYS**: Well days of production. This column contains the same value for each well.
27. **LIQ\_CUM\_BBLs**: Cumulative produced oil. This column contains the same value for each well.
28. **GAS\_CUM**: Cumulative produced gas. This column contains the same value for each well.

The data sets contain a lot of redundant and irrelevant columns. Besides, the content of many columns is not well organized. For this reason, the data needs to be cleaned and explored before starting to create predictive models.

### 3.3 Python Tools and Libraries

Python has become the preferred tool for data scientists because of its simple, easy to use syntax and the great number of modules and packages it supports. The libraries used in this project are NumPy, SciPy, Pandas, Scikit-learn, Matplotlib, Seaborn and Folium.

1. **NumPy**: NumPy (short for Numerical Python) is a Python library that provides a multidimensional array and matrix data structures. NumPy is the basis for pandas. It has many useful functions, and its advantages include speed and memory. The main data structure in NumPy is the NumPy array. A NumPy array is similar to a list. It is usually fixed in size and each element is of the same type. NumPy provides a vast

number of mathematical and statistical operations which can be performed on these arrays.

2. **SciPy:** SciPy (short Scientific Python) is a scientific computation library that uses NumPy underneath. It provides more utility functions for optimization, stats, and signal processing. While NumPy contains array data and basic operations such as sorting, indexing, etc, SciPy is the library that contains fully featured versions of these functions along with many others.
3. **Pandas:** Pandas is a Python package that provides fast, flexible, and expressive data structures. Data in pandas is often used to feed statistical analysis in SciPy, plotting functions from Matplotlib, and machine learning algorithms in Scikit-learn. Pandas can be used to import data from different file formats (comma-separated values, JSON, SQL, Microsoft Excel, etc...) into a DataFrame, handle missing data easily (represented as NaN), insert, rename, or delete columns in the DataFrame, view, explore and inspect the DataFrame, explore the relationship between continuous variables, etc.
4. **Scikit-learn:** Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is very easy to use, yet it implements many machine learning algorithms efficiently. It features various classification, regression, and clustering algorithms. Examples include linear regression, support vector machines (SVM), random forests (RF), gradient boosting, k-means and DBSCAN etc. The different machine learning models created in this thesis will be using Scikit-learn.
5. **Matplotlib:** Matplotlib is a multiplatform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It is used to create static, animated, and interactive visualizations in Python. Matplotlib is mainly deployed for basic plotting. Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots, etc.
6. **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Compared to Matplotlib, Seaborn is more comfortable in handling Pandas data frames. It uses basic sets of methods to provide beautiful graphics in python. It also uses fewer syntax and has easily interesting default themes.
7. **Folium:** Folium is a Python library used for visualizing geospatial data. Folium is a Python wrapper for Leaflet.js which is a leading open-source JavaScript library for plotting interactive maps. Folium is used in this thesis to visualize the location of the wells.

## 4 Data Wrangling and Exploratory Data Analysis

The data sets are not well organized and contain a lot of redundant columns and information. Before creating the predictive models, the data must be cleaned and processed. The first step in cleaning the data is to verify the completeness of the data and the data types when imported in Python. The second step is to change the format of different columns to make the contents clear and relevant. The third step is to explore the data sets with exploratory data analysis. Exploratory data analysis is important to examine the impact of different attributes on the oil and gas production of the training wells.

This chapter discusses all the changes that were made to the data sets to make them useable for machine learning algorithms. It also summarizes different attributes and their impact on the oil and gas production of the training wells. The first part of this chapter presents the columns that contain missing data and the data type of each column. The second part contains the different transformation that were applied to the data sets to obtain a clear and precise DataFrame. The third part presents an overview of the oil and gas production data of the training wells. It also discusses the most important categorical and numerical variables that impact the oil and gas production, as well as the transformations that were made to encode and scale the data.

### 4.1 Completeness of the Data and Data Types

It is important to verify if the data contains any missing values. Missing data can influence the performance of the model created. Data types are important because some operations can only be used with specific data types. Integer and float are number types, while object generally refers to text. Table 5 shows the missing values contained in each column in the training and target data set and the data type of each column (data types are the same for both data sets).

- **Completeness of the Data**

As shown in table 5, the training data set has 7 missing values in the column “MIN\_STP”. The target data set has no missing values apart from the oil and gas production. Determining the values of these columns is the objective of this thesis. In case of missing data, it is possible to either drop the data (drop the row or column) or replace the data (replace by mean, frequency, or based on other functions). The missing data will be replaced in the data wrangling part, using the mean value of the column “MIN\_STP” to replace the missing values. New attributes extracted from the data sets, in particular from the column “JOB\_DESC\_STAGING”, also contained missing data. This missing data will also be discussed more in the Data Wrangling part of this chapter.

- **Data Types**

When importing a data set into a pandas DataFrame, it is important to verify the data types. The data contained in both data sets is mostly numerical. Most of the object columns (columns that contain text) are the columns that contain the unit of the previous column. The types of all

the columns when imported were representative of the contents of the columns, so no changes were made. The columns that contain text (object) will be replaced by columns that contain numerical data, because machine learning algorithms generally prefer numerical values. This transformation will be discussed in the exploratory data analysis and data preparation part of this chapter.

Table 5: Missing Values and Data Types of the Data Sets Used

Column Name	Missing Values (Training Set)	Missing Values (Target Set)	Column Type
WELL_ID	0	0	int64
JOB_DESC_STAGING	0	0	object
PROPPANT_MESH_SIZE	0	0	object
PROPPANT_MESH_DESCRIPTION	0	0	object
PROPPANT_MASS_USED	0	0	int64
PROPPANT_MASS_UOM	0	0	object
VOLUME_PUMPED_GALLONS	0	0	int64
AVERAGE_STP	0	0	float64
AVERAGE_STP_UOM	0	0	object
FRACTURE_GRADIENT	0	0	float64
FRACTURE_GRADIENT_UOM	0	0	object
MD_MIDDLE_PERFORATION	0	0	float64
MD_MIDDLE_PERFORATION_UOM	0	0	object
TVD_DEPTH	0	0	int64
TOP_DEPTH	0	0	int64
WELL_LATITUDE	0	0	float64
WELL_LONGITUDE	0	0	float64
MIN_STP	7	0	float64
MIN_STP_UOM	0	0	object
MAX_STP	0	0	int64
MAX_STP_UOM	0	0	object
UPPER_PERF	0	0	int64
LOWER_PERF	0	0	int64
TRUE_VERTICAL_DEPTH	0	0	int64
WELL_HORZ_LENGTH	0	0	int64
NET_PROD_DAYS	0	0	int64
LIQ_CUM_BBLS	0	338	int64
GAS_CUM	0	338	int64

## 4.2 Data Wrangling

The data sets are not well organized and contain a lot of redundant columns and information. Some columns will be deleted, and others will be created to better organise the DataFrame. Having a precise DataFrame is essential to examine the impact of each attribute. On the other hand, columns that give no useful information for the machine learning algorithm, like columns containing units, need to be deleted to reduce the dimension of the DataFrame. The following problems need to be addressed to have clearer data sets:

1. The columns “LIQ\_CUM\_BBLs” and “GAS\_CUM” will be renamed to “CUMULATIVE\_OIL\_PRODUCTION” and “CUMULATIVE\_GAS\_PRODUCTION” respectively to have more accurate column names.
2. Two columns “TVD\_DEPTH” and “TRUE\_VERTICAL\_DEPTH” seem to contain the same information. However, for some wells, the two columns contain values that are not the same. This needs to be investigated to determine which column contains the correct true vertical depth data. The other column will be deleted. Besides, many columns contain information about units and should be removed.
3. The columns “PROPPANT\_MESH\_SIZE” and “PROPPANT\_MESH\_DESCRIPTION” contain a lot of redundant information. “PROPPANT\_MESH\_SIZE” contains a lot of proppant type information, and “PROPPANT\_MESH\_DESCRIPTION” contains a lot of proppant mesh size information. In many cases, these two columns contain exactly the same value. These two columns should be used to extract a “PROPPANT\_TYPE” column and a “PROPPANT\_MESH\_SIZE” column. This way, it will be possible to examine the impact of each of these two different attributes on the oil and gas production of the training wells.
4. The column “JOB\_DESC\_STAGING” contains information about the day number, the stage number, the fracturing fluid, and the geological basin. Since all wells are drilled in the same basin, the geological basin information is irrelevant for the model. This column should be removed, and instead columns containing the relevant information should be created. The columns “FRAC\_FLUID”, “DAY\_NUMBER”, and “STAGE\_NUMBER” will be created, and will be filled with the relevant information extracted from the “JOB\_DESC\_STAGING” column, which will be deleted.
5. The column “MIN\_STP” contains missing data. The missing data will need to be filled.

#### 4.2.1 Removing Redundant Columns

Since many columns only contain the units of the columns before them, they will be removed from both data sets since the information is useless for the model. These columns are: “PROPPANT\_MASS\_UOM”, “AVERAGE\_STP\_UOM”, “FRACTURE\_GRADIENT\_UOM”, “MD\_MIDDLE\_PERFORATION\_UOM”, “MIN\_STP\_UOM”, and “MAX\_STP\_UOM”.

The columns “TVD\_DEPTH” and “TRUE\_VERTICAL\_DEPTH” both contain information about the same parameter: True Vertical Depth of the well. Normally, these two columns should contain the same values. However, when plotting the two columns against each other, they are not exactly the same for the training data set, as shown in Figure 13. The graph clearly shows that there is a strong correlation between the two variables. This is also confirmed when calculating the correlation coefficient, which is 0.881. Most wells have the same value for these two columns. However, it is clear that there are two outliers in the graph. When the outliers from the plot in Figure 13 are removed, the correlation coefficient becomes 0.999. The resulting plot is shown in Figure 14.

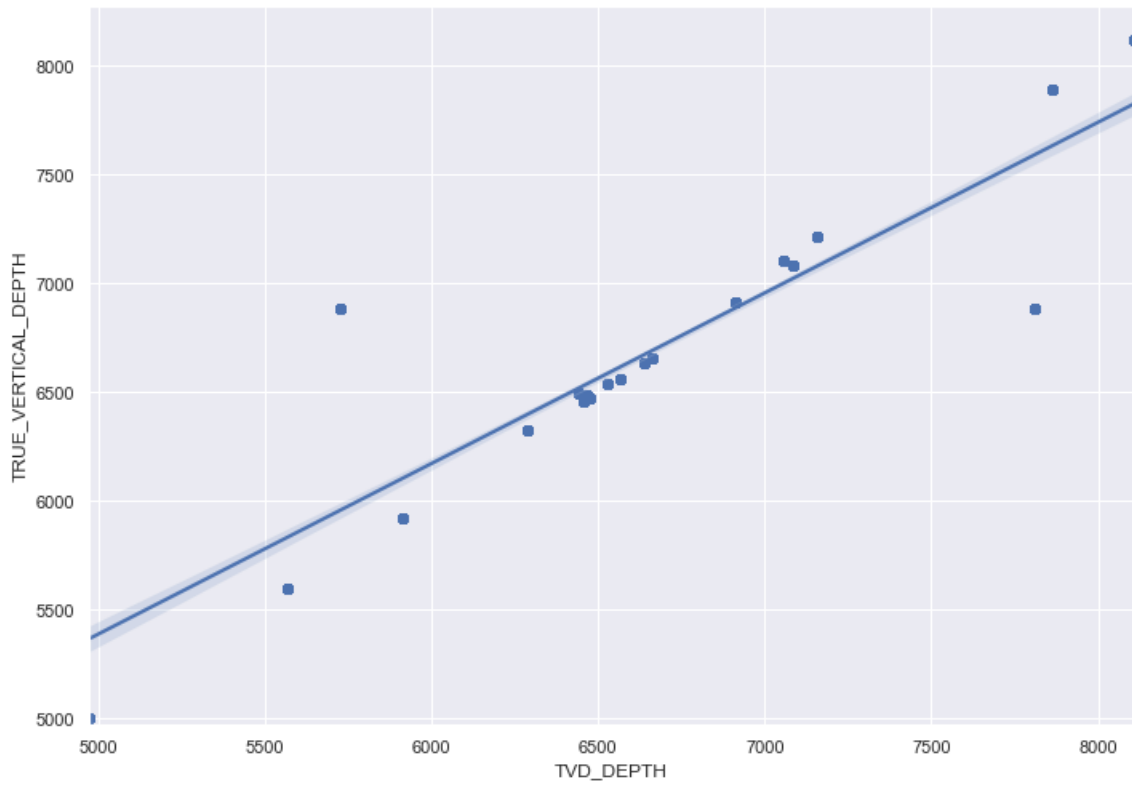


Figure 13: Plot of True Vertical Depth Column versus TVD Depth Column of the Training Set

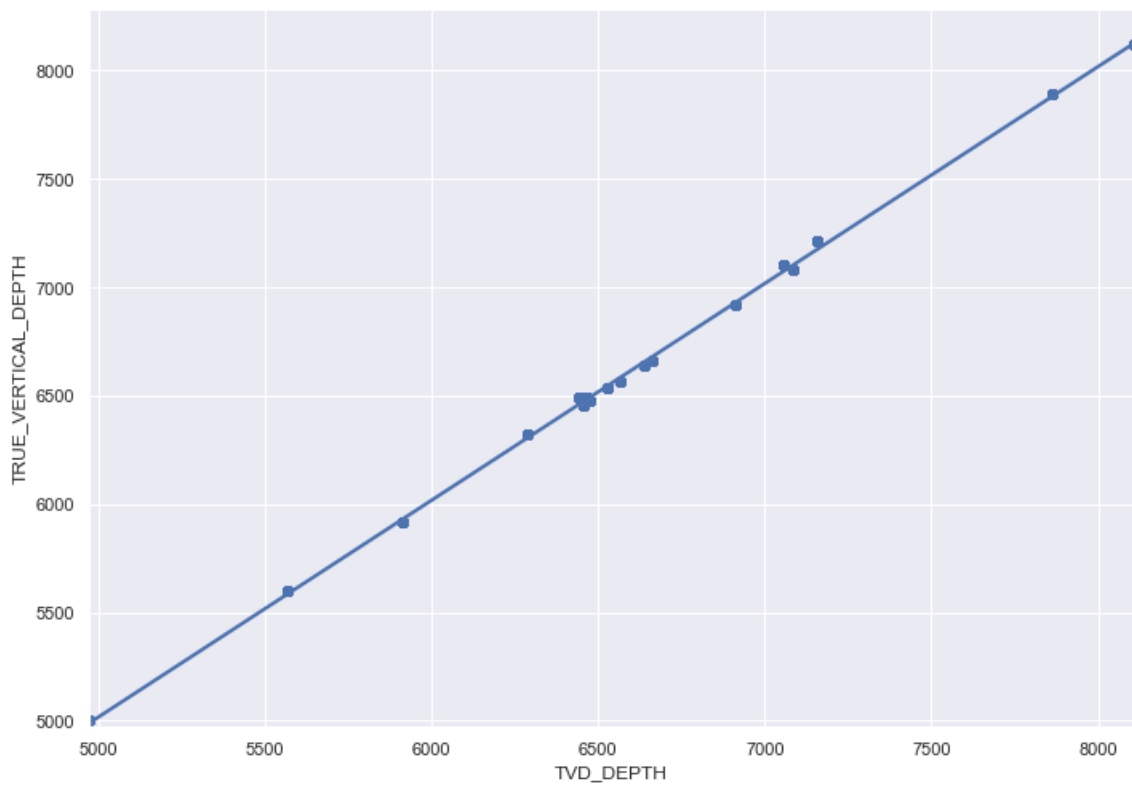


Figure 14: Plot of True Vertical Depth Column versus TVD Depth Column of the Training Set without outliers



Figure 14 confirms that both columns contain the same information, and one of them needs to be removed. The two outliers are wells 2 and 3. For well 2, “TVD\_DEPTH” is 5727 and “TRUE\_VERTICAL\_DEPTH” is 6888. For well 3, “TVD\_DEPTH” is 7809 and “TRUE\_VERTICAL\_DEPTH” is 6888. “TRUE\_VERTICAL\_DEPTH” contains the same value for two different wells, while “TVD\_DEPTH” contains different values for these two different wells. It is therefore safe to assume that the column “TRUE\_VERTICAL\_DEPTH” contains erroneous data and that it must be removed.

The correlation coefficient of the same columns for the target data set is 0.999, and the relationship is plotted in Figure 15. This also proves that these two columns are supposed to contain the same values. Therefore, the column “TRUE\_VERTICAL\_DEPTH” is therefore also removed from the target data set.

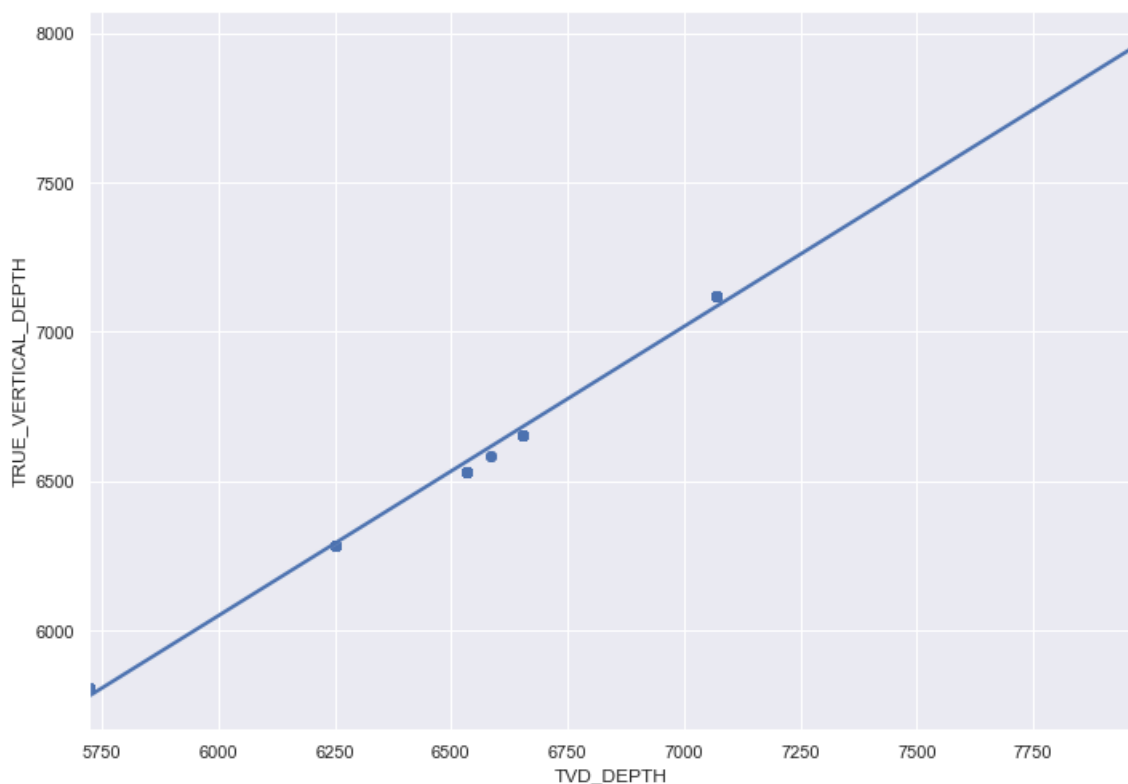


Figure 15: Plot of True Vertical Depth Column versus TVD Depth Column of the Target Set

Both data sets now contain 21 columns. The next step is to extract the relevant information from some columns that are not well organised. This information is related to the proppant type, the proppant mesh size, the fracturing fluid used, the number of days and the number of stages of the hydraulic fracturing job.

#### 4.2.2 Extracting Proppant Information

Information about proppant mesh size and type is contained in two columns: “PROPPANT\_MESH\_SIZE” and “PROPPANT\_MESH\_DESCRIPTION”. These columns sometimes contain the same information. The objective is to rename the column “PROPPANT\_MESH\_DESCRIPTION” to “PROPPANT\_TYPE” and change it so that it

contains the proppant type used. The column “PROPPANT\_MESH\_SIZE”, which contains the mesh size of the proppant used, will contain the mesh size of the proppant. The initial contents of these two columns in the training data set are presented in Table 6, while the initial contents of the same columns in the target data set are presented in Table 7:

Table 6: “PROPPANT\_MESH\_DESCRIPTION” and “PROPPANT\_MESH\_SIZE” contents in the Training Data Set

“PROPPANT_MESH_DESCRIPTION”		“PROPPANT_MESH_SIZE”	
Values in the Column	Count	Values in the Column	Count
Sand, White, 100 mesh	524	Sand, White, 100 mesh	518
Sand, White, 40/70	403	40/70	419
Sand, White, 30/50	169	30/50	172
Sand, White, 20/40	61	20/40	61
Sand, Brown, 40/70	16	S-8C, Sand, 100 mesh, bulk	7
Sand, White, 30/50 SSF Odessa	2	20/50	2
Sand, White, 20/50	2		
S-8C, Sand, 100 mesh, bulk	1		
Sand, Brown, 30/50	1		

Table 7: “PROPPANT\_MESH\_DESCRIPTION” and “PROPPANT\_MESH\_SIZE” contents in the Target Data Set

“PROPPANT_MESH_DESCRIPTION”		“PROPPANT_MESH_SIZE”	
Values in the Column	Count	Values in the Column	Count
Sand, White, 40/70	168	40/70	169
Sand, White, 100 mesh	144	100	144
Sand, White, 30/50	24	30/50	24
Sand, White, 20/40	1	20/40	1
Super LC, 40/70	1		

Using tables 6 and 7, the following observations can be made:

1. Mesh size 20/50 normally does not exist in the industry, therefore it will be considered as a mistake and changed to 30/50.
2. SSF Odessa, S-8C are special types of white sand. Super LC is Resin Coated Sand, but since there is only one use of this proppant type, it will be transformed to white sand.
3. “PROPPANT\_MESH\_DESCRIPTION” will be changed into “PROPPANT\_TYPE” and all information about mesh size will be removed.
4. All information about proppant type will be removed from “PROPPANT\_MESH\_SIZE”.

Once all these changes are made, the columns “PROPPANT\_TYPE” and “PROPPANT\_MESH\_SIZE” will contain information that is logical and that can be understood and used in the exploratory data analysis part. The result of the data processing is shown in Table 8 for the training data set. It shows a big disparity in the use of the two proppant types. The brown sand is of less quality compared to the white sand, and therefore white sand should have a better impact on the quality of the hydraulic fracturing job and therefore on the oil and gas production.

Table 8: “PROPPANT\_TYPE” and “PROPPANT\_MESH\_SIZE” contents in the Training Data Set after processing

“PROPPANT_TYPE”		“PROPPANT_MESH_SIZE”	
Values in the Column	Count	Values in the Column	Count
White Sand	1162	100	525
Brown Sand	17	40/70	419
		30/50	174
		20/40	61

For the training data set, proppant mesh size that contain smaller grain size are used more frequently than mesh sizes that contain larger grain size. The most used mesh sizes are 100 and 40/70, while 30/50 and 20/40 are used less frequently. Table 9 shows the result of the data processing for the target data set. It shows that the wells in the target data set are only fractured using white sand. The frequency of the proppant mesh size used is different compared to the training data set. The most frequently used mesh size is 40/70, followed by 100. The mesh size 30/50 is used only 7% of the time, while the mesh size 20/40 is used only once in 338 times.

Table 9: “PROPPANT\_TYPE” and “PROPPANT\_MESH\_SIZE” contents in the Target Data Set after processing

“PROPPANT_TYPE”		“PROPPANT_MESH_SIZE”	
Values in the Column	Count	Values in the Column	Count
White Sand	338	40/70	169
		100	144
		30/50	24
		20/40	1

Since the data is now cleaned and does not contain outliers, it is possible to plot it. Figure 16 shows the proppant type used in all 27 wells in the training and target sets. Figure 17 is a plot of the different mesh sizes used in the hydraulic fracturing for the 27 wells of both data sets.

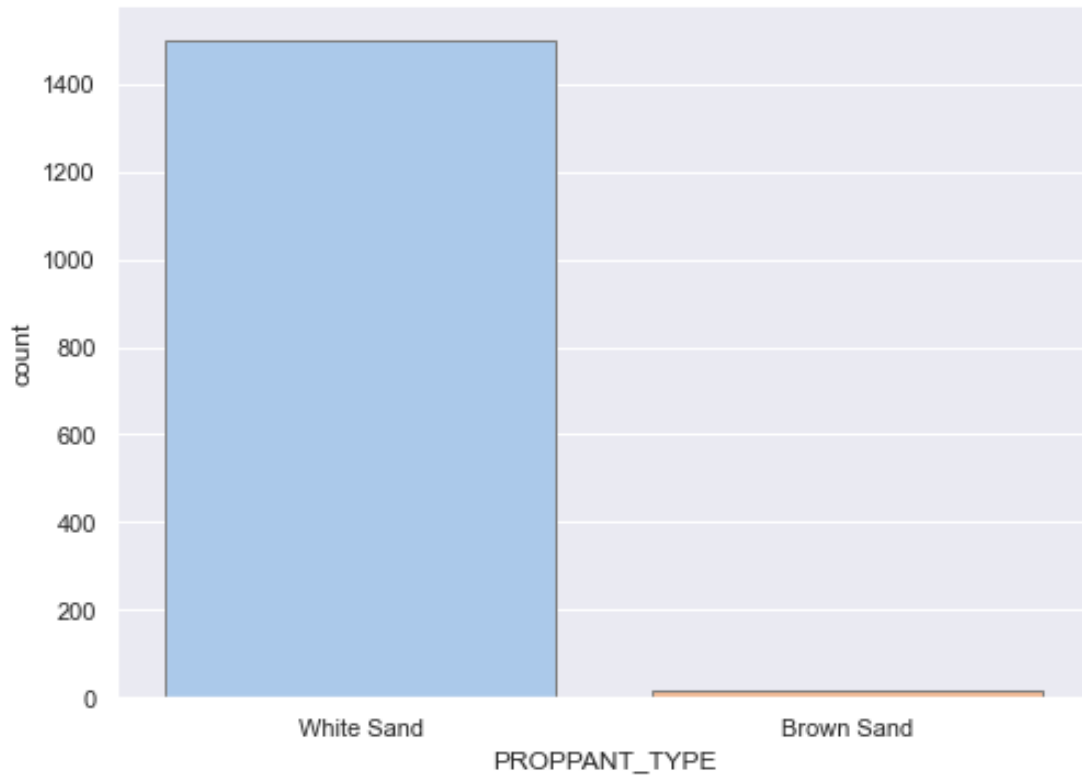


Figure 16: Proppant Type Counts of Both Data Sets after Processing

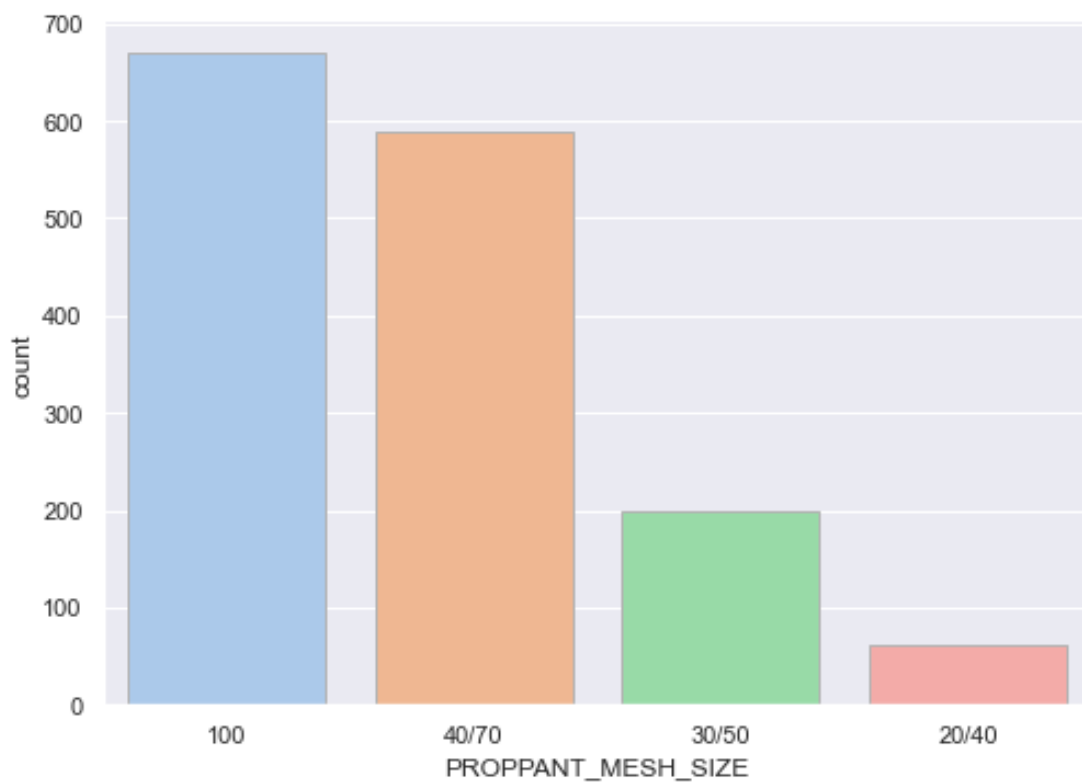


Figure 17: Proppant Mesh Size Counts in Both Data Sets after Processing

Figure 16 shows that the preferred proppant type is clearly white sand. As discussed earlier in this thesis, white sand is more expensive but is of higher quality compared to brown sand.

Using sand is a cheaper but less effective option compared to resin-coated sand or ceramic proppants. Figure 17 shows that the most used proppant mesh sizes are 100 and 40/70, while 30/50 and 20/40 are used less frequently. The impact of the different proppant types and mesh sizes on the oil and gas production will be further explored in the exploratory data analysis section.

### 4.2.3 Extracting Fracturing Fluid, Day, and Stage Number:

Fracturing fluid, day, and stage number information is contained in the “JOB\_DESC\_STAGING” column. However, the information in this column is not organised. Some samples from the column “JOB\_DESC\_STAGING” from the training data set: “Day 5 Wolfcamp Frac Slickwater Stg 27”, “Day 3: Stg 11 Wolfcamp Frac (11220-11470)”, “Day 5 Stage 28: Wolfcamp (Hybrid)”, “Day 4 Stage 17: Wolfcamp @ 7733'-7915'”, etc. The information contained in this column can be divided into different categories and parameters:

1. Wolfcamp is the geological formation from which the wells are producing. Since all the wells are producing from the same formation, this information is not relevant for the model.
2. Day and Stage (or stg) refer to the day number and stage number of the fracturing job.
3. Slickwater and Hybrid are types of fracturing fluid.
4. Some rows contain information about the perforation measured depth range (11220-11470, @ 7733'-7915'). This information is already contained in the column “MD\_MIDDLE\_PERFORATION”, therefore it is not needed.

These samples prove that this column should not be used in a machine learning model in its current state. The objective is to extract the information into 3 columns: “FRAC\_FLUID”, “DAY\_NUMBER”, and “STAGE\_NUMBER”.

Fracturing fluids specified in the column “JOB\_DESC\_STAGING” are Slickwater and Hybrid. Slickwater is in some cases written as “SW”. In case no information is provided, it is assumed that water is the fracturing fluid used. A new column “FRAC\_FLUID” is created containing the type of fracturing fluid, and its contents are shown in Table 10. Figure 18 presents a count plot of the column “FRAC\_FLUID” for the wells of both the training and target wells.

Table 10: “FRAC\_FLUID” contents for the training and target data sets

“FRAC_FLUID” Training Set		“FRAC_FLUID” Target Set	
Values in the Column	Count	Values in the Column	Count
Water	837	Water	335
Slickwater	231	Slickwater	3
Hybrid	111		

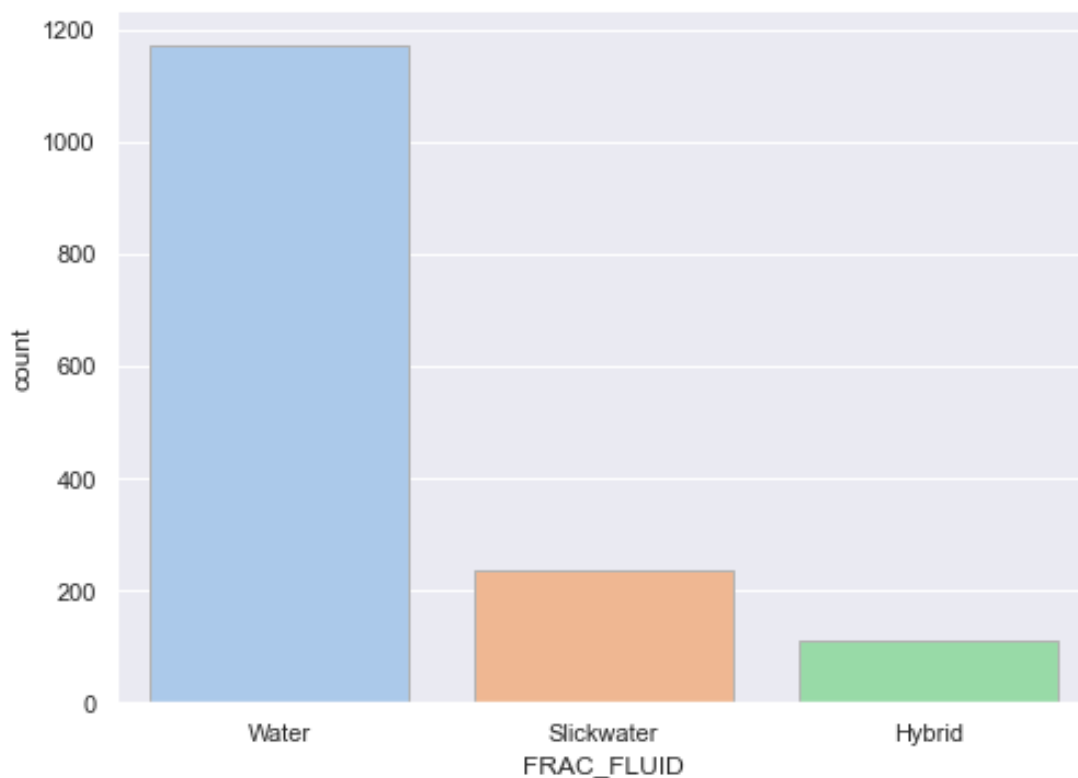


Figure 18: Fracturing Fluid Counts in both Data Sets after Processing

As shown in the table and figure above, Water is the most used fracturing fluid in both data sets. Slickwater is used a considerable number of times in the training data set, but only three times in the target data set. Hybrid is only used in the training data set. The impact of the different fracturing fluid on the oil and gas production will be further examined in the exploratory data analysis section.

Once the fracturing fluid information has been extracted, the only remaining relevant information contained in the “JOB\_DESC\_STAGING” is the day number and stage number. The number of stages and the number of days of hydraulic fracturing is important for the production of oil and gas. The day number and stage number are contained in most of the rows. However, the column “JOB\_DESC\_STAGING” presents a challenge for some wells:

1. Well 4 (Training Set): “JOB\_DESC\_STAGING” only contains a number, without specifying whether it is a date number or stage number. However, since it is increasing every 2 rows, it is safe to assume that the number refers to the stage number. Exp: “Wolfcamp Frac (Slickwater) 30”. Therefore, the number will be contained in the “STAGE\_NUMBER” column and the “DAY\_NUMBER” column will be empty for this well.
2. Well 6 (Training Set): “JOB\_DESC\_STAGING” contains the value “Stage 1 Wolfcamp” for two rows and “Stage 2-34 Wolfcamp” for the rest of the rows. The column “PROPPANT\_MASS\_USED” will be used to determine the stages 2 to 34 since it is increasing for each stage. The column “DAY\_NUMBER” will be empty.

3. Well 27 (Target Set): "JOB\_DESC\_STAGING" only contains the day number for the first stage of each day. Exp: "Day 3- Stage 11" "Stage 12". Each row in the column "DAY\_NUMBER" will therefore contain the last day number mentioned.

The resulting column "DAY\_NUMBER" contains 136 missing values in the training data set, which are the values for wells 4 and 6. The missing values will be filled in the filling missing data section.

The information that can be extracted from the column "JOB\_DESC\_STAGING" is stored in the columns "FRAC\_FLUID", "DAY\_NUMBER", and "STAGE\_NUMBER". The column can therefore be deleted. The impact of these features on the oil and gas production of the training wells will be examined in the exploratory data analysis and data preparation part of this chapter.

#### 4.2.4 Filling Missing Data

After the data processing, 2 columns contain missing data in the training data set, while the target data set contains no missing data. The column "MIN\_STP" contains 7 missing values. The contents of this column are described in Table 11:

Table 11: Statistical Description of the Column "MIN\_STP" in the Training Data Set

Missing Values	7
Count	1172
Mean	3759.52
Standard Deviation	1622.04
Minimum Value	9
25 <sup>th</sup> Percentile	2957
50 <sup>th</sup> Percentile	3660
75 <sup>th</sup> Percentile	4490
Maximum Value	32641

Most values are in the range of 2000 to 5000 psi. Some outliers exist, the minimum value being 9 psi and the maximum value more than 32000 psi. The 7 missing values will be replaced with the mean of the column.

The column "DAY\_NUMBER" contains 136 missing values, which are the values for wells 4 and 6. Since a correlation exists between stage number and day number, it is possible to use the average number of stages executed per day to fill the missing values. Figure 19 shows the number of stages per day for the wells of the training set.

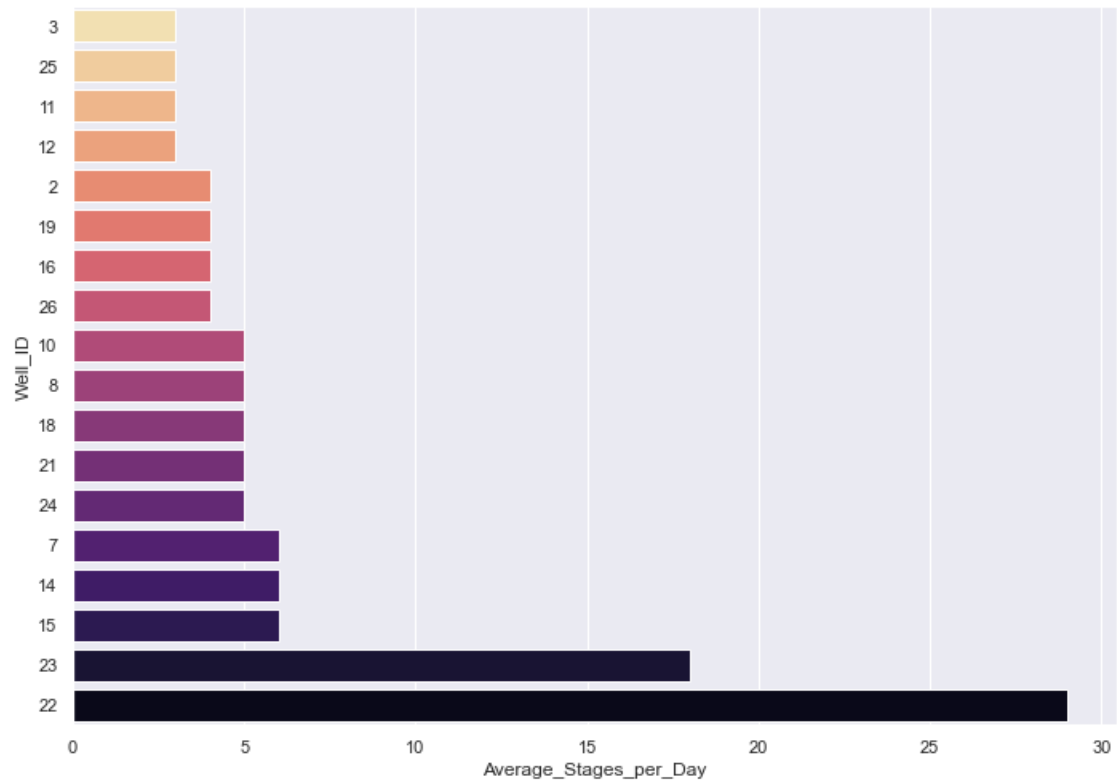


Figure 19: Average Number of Stages per Day for the Wells of the Training Set

Figure 19 shows that most wells have an average number of 3 to 6 stages per day. However, wells 22 and 23 both present values that are much bigger. This means that the mean of the list of averages would be influenced by these two wells. For this reason, the median, which is 5, will be used to fill the missing values for the wells 4 and 6. Figure 20 shows the number of hydraulic fracturing stages for the training data set.

The bars coloured in red shown in figure 20 indicate the wells 4 and 6, which are missing information about the day number of the hydraulic fracturing. Both these wells are fractured in 34 stages. Using the median of the number of stages per day (5 stages per day) as mentioned earlier, it will be assumed that these 34 hydraulic fracturing stages are spread across 7 days. Figure 21 shows the number of days of the hydraulic fracturing job for the wells of the training set, with the bars in red indicating wells 4 and 6. Figures 20 and 21 show a big difference between the wells in the number of days and number of stages. The data also shows a big variation between wells in different other categorical and numerical parameters. The next step is then to determine which of these parameters has the highest influence on the oil and gas production. Since both data sets no longer contain missing data, it is possible to do some exploratory data analysis. The final step before creating the model is to prepare the data for machine learning algorithm.



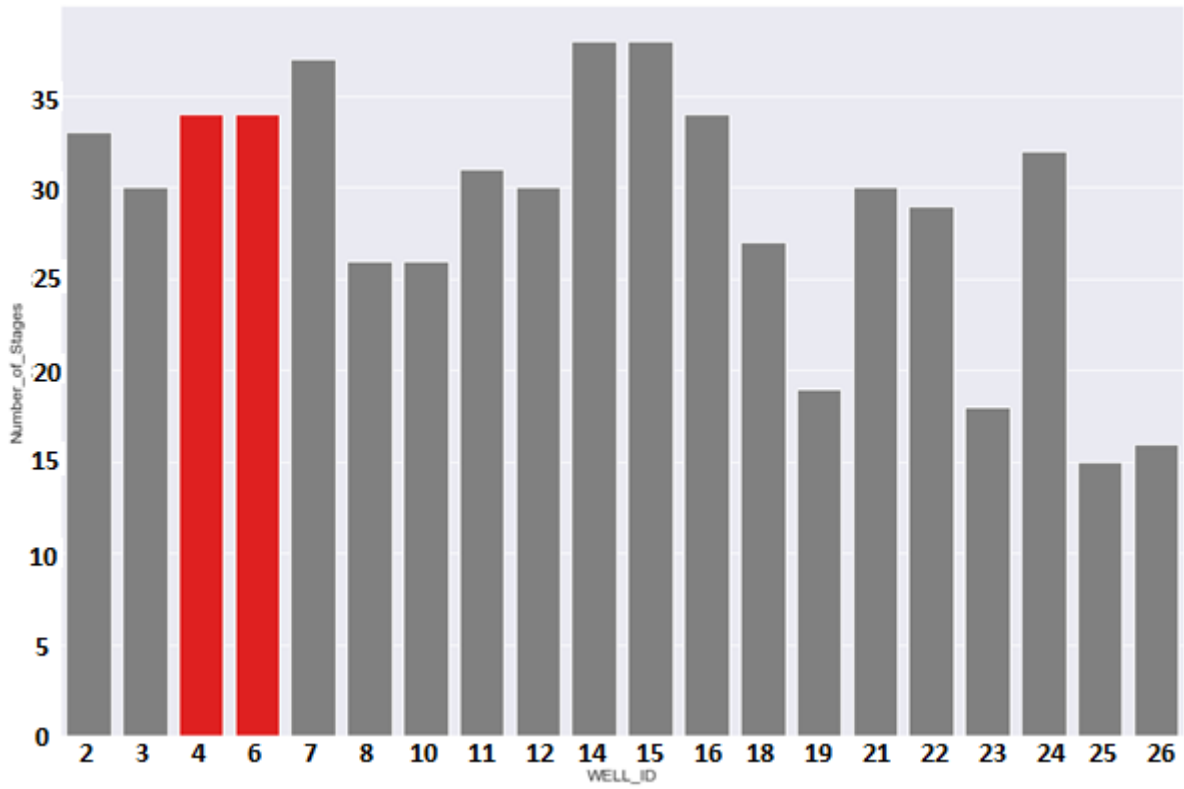


Figure 20: Number of Stages for each Well in the Training Data Set

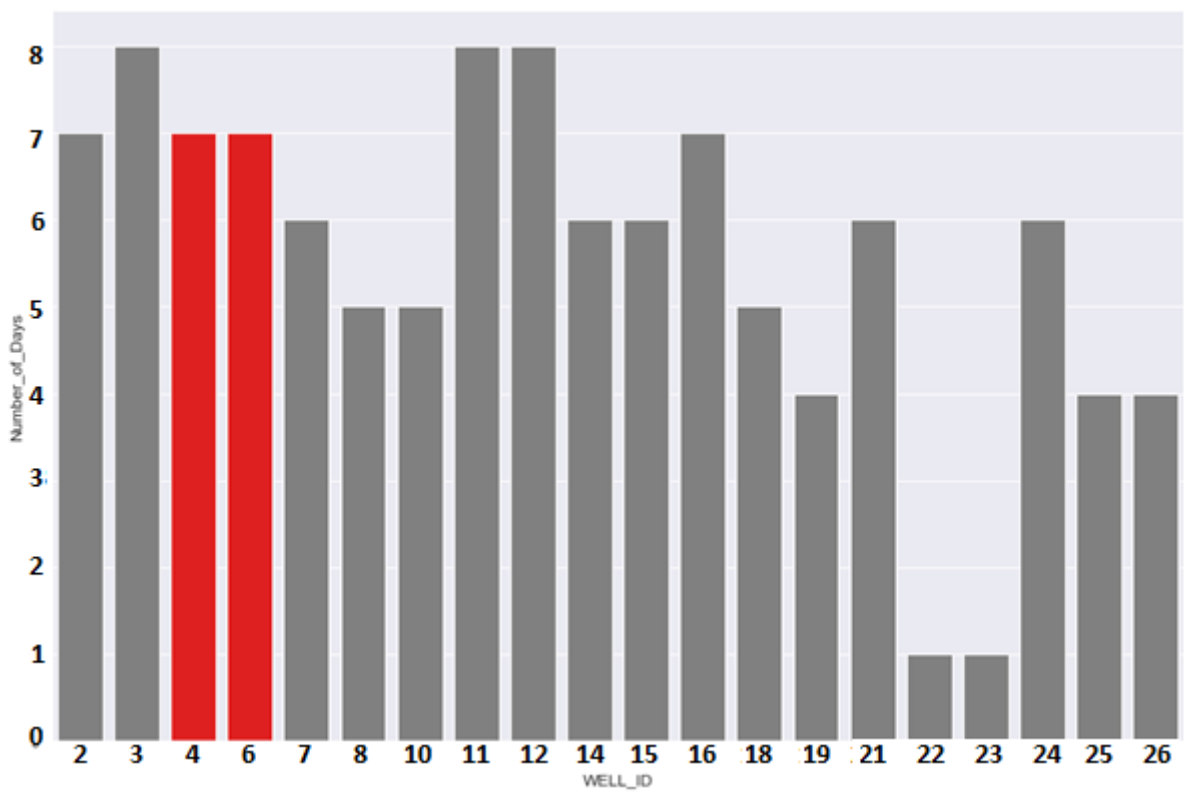


Figure 21: Number of Days for each Well in the Training Data Set

## 4.3 Exploratory Data Analysis and Data Preparation

Since the data is now processed and the missing data is filled, the training data set needs to be further explored before creating the model. Since the wells have many different varying parameters, it is essential to determine which parameters have the highest influence on the oil and gas output. Exploratory data analysis can be used to analyse and investigate data sets and summarize their main characteristics, often using data visualization methods. It helps to determine how best to manipulate data sources to discover patterns, detect outliers or anomalous events, find interesting relations among the variables, test a hypothesis, or check assumptions. Once exploratory data analysis is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

For categorical variables, box plots can be used to examine the distribution of the production with respect to each variables. For numerical variables, the linear relationship between the variable and the production can be calculated. Plotting scatterplots with fitted lines is also a helpful way to visualize the relationship between numerical variables. The data then needs to be prepared, by scaling the numerical variables and encoding the categorical variables.

### 4.3.1 Oil and Gas Production of the Training Wells

The 20 wells used in the training have varying oil and gas productions. The mean value of oil production is 49613 barrels , while the standard deviation is around 43107 barrels. Most wells have a cumulative oil production between 50000 and 120000 barrels. Figure 22 shows the cumulative oil production per well. It indicates that the highest oil producing well is well 6, with a cumulative oil production of 160000 barrels. The lowest oil producing wells are wells 7, 25 and 26. These wells have a production of 2000 barrels or less.

For the cumulative gas production, the mean value is 245236 thousand cubic feet (Mcf), and the standard variation is 211168 Mcf. The values range between 5000 Mcf to around 700000 Mcf. Figure 23 shows the cumulative gas production per well. It shows a big disparity in the cumulative gas production. Wells 3 and 21 produce negligible amounts of gas. Wells 7, 22, 24, 25 and 26 produce small amounts of gas, not exceeding 100000 Mcf. The three biggest gas producers are wells 6, 11 and 12, with gas volumes above 550000 Mcf.

Using figures 22 and 23, it is possible to determine the best and worst overall producers. Well 6 is the best overall well, having the best cumulative oil production and the third best gas production. Well 15 is also a very good producer, with a cumulative oil production of around 115000 barrels and a cumulative gas production of more than 450000 Mcf. The worst wells are wells 7, 25 and 26.

The ratio of oil to gas production is very different between the wells. The well with the lowest oil to gas ratio is well 7, which produces 0.019 barrels of oil per Mcf of gas. Other wells also present a ratio of less than 0.1 barrels per Mcf, which are wells 26, 25 and 18. Most wells

produce between 0.1 to 0.3 barrels per Mcf. Two wells produce more than 1 barrel per Mcf, which are wells 3 (1.8 barrels per Mcf) and 21 (12.8 barrels per Mcf).

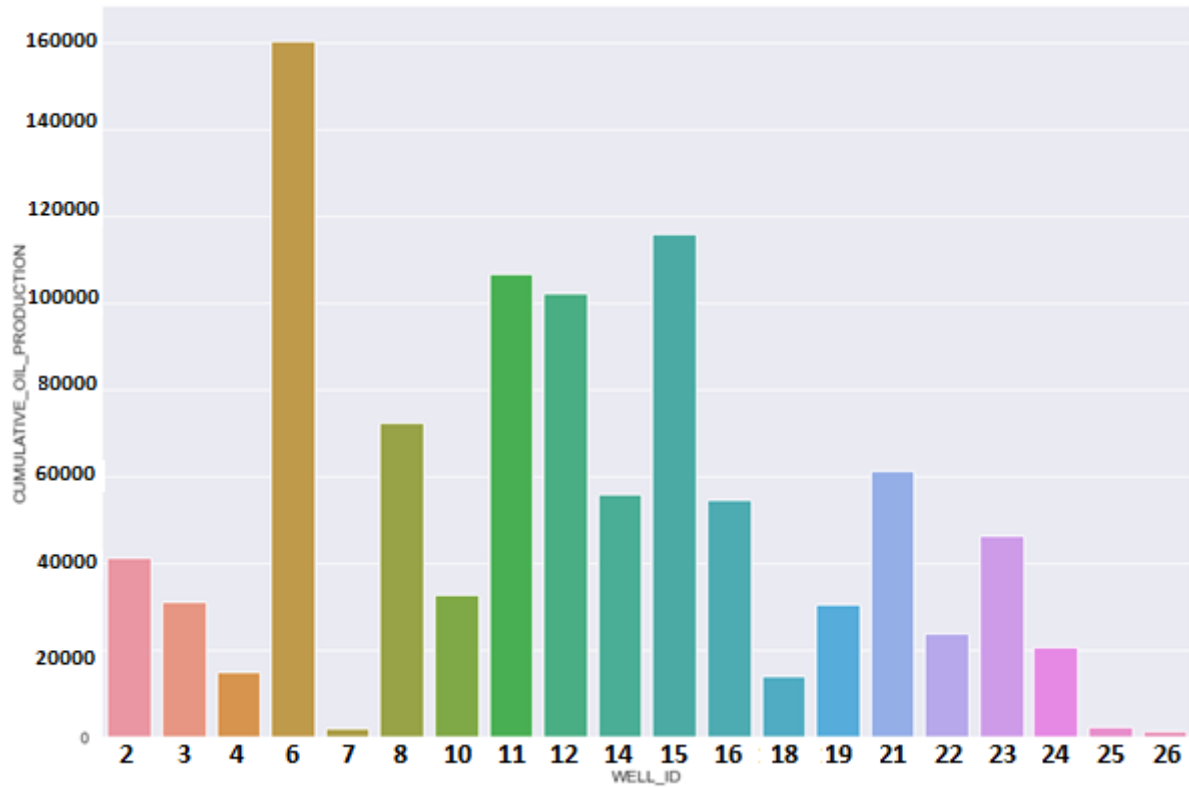


Figure 22: Cumulative Oil Production of the Training Wells in Barrels

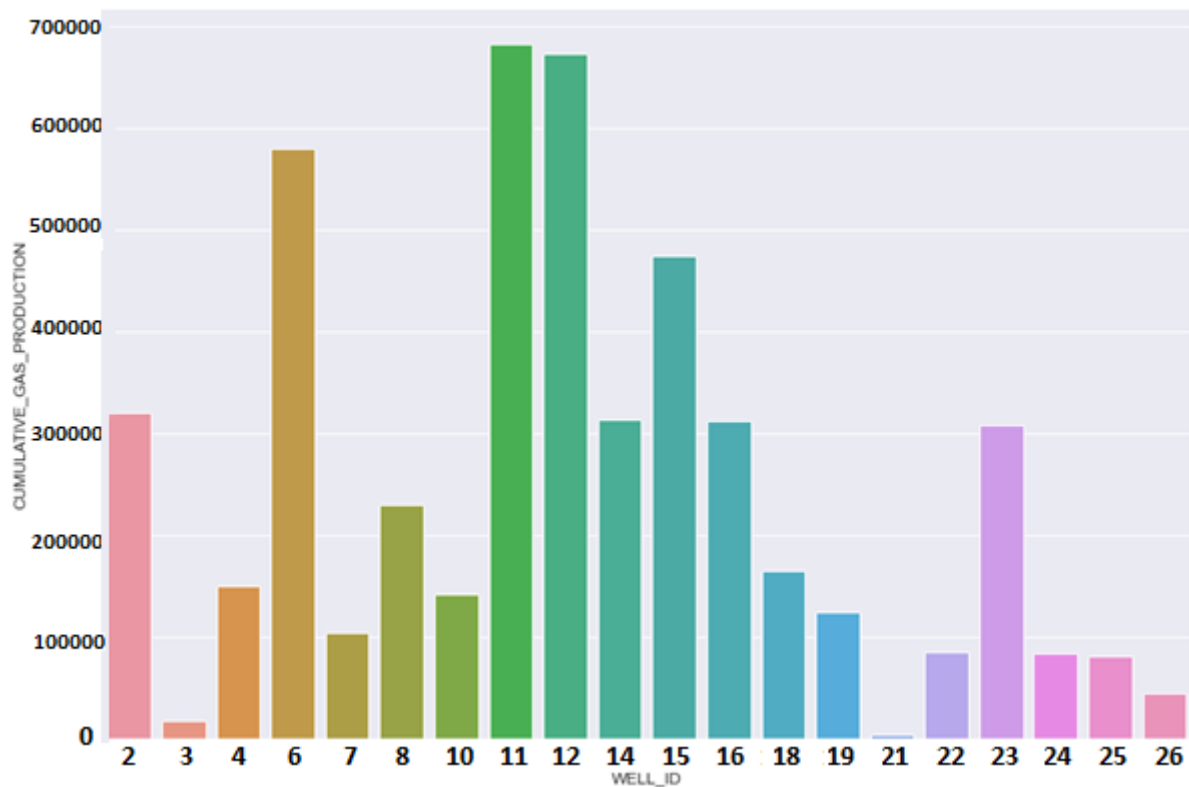


Figure 23: Cumulative Gas Production of the Training Wells in Mcf

It is important to examine the number of production days for each well to be able to compare the oil and gas production per day for each well. Figure 24 shows the number of days of production for each well.

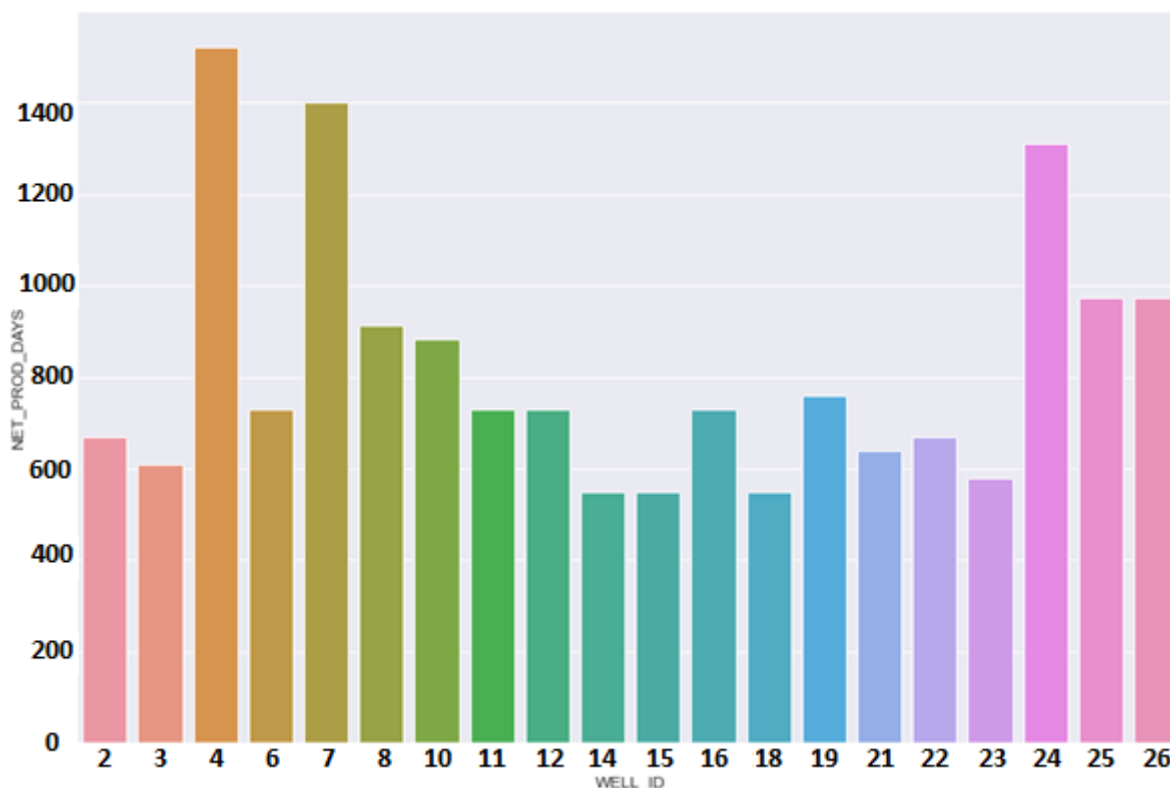


Figure 24: Net Production Day of the Training Wells in days

Figure 24 shows that the minimum number of production days is around 500. The wells that produced for the shortest period of time are wells 14, 15 and 18. The wells that have produced the longest are wells 4, 7 and 24, with well 4 producing for the longest period of days (1522). The majority of wells have produced for 600 to 800 days. Combining the cumulative oil and gas production with the number of production days gives the oil production per day per well and the gas production per day per well, which are presented in figure 25 and figure 26, respectively.

These figures show that well 6 is still the best oil producer, while well 15 is also a comparable oil producer, both reaching more than 200 barrels of oil per day. Wells 11, 12; 14 and 21 are also good producers, with 100 barrels per day or more. Wells 7, 25 and 26 produce almost no oil.

The best gas producers per day are wells 6, 11, 12 and 15 with 800 Mcf/day or more. Wells 2, 14, 16 and 23 are also good producers, reaching a daily average gas production of 400 Mcf/day or more. The worst gas producers are wells 3 and 21 since they produce almost no gas at all. The overall worst well is well 26, which produces very little amount of oil and gas. The relationship between oil and gas production and the impact of the number of days on the oil and gas production will be further explored.

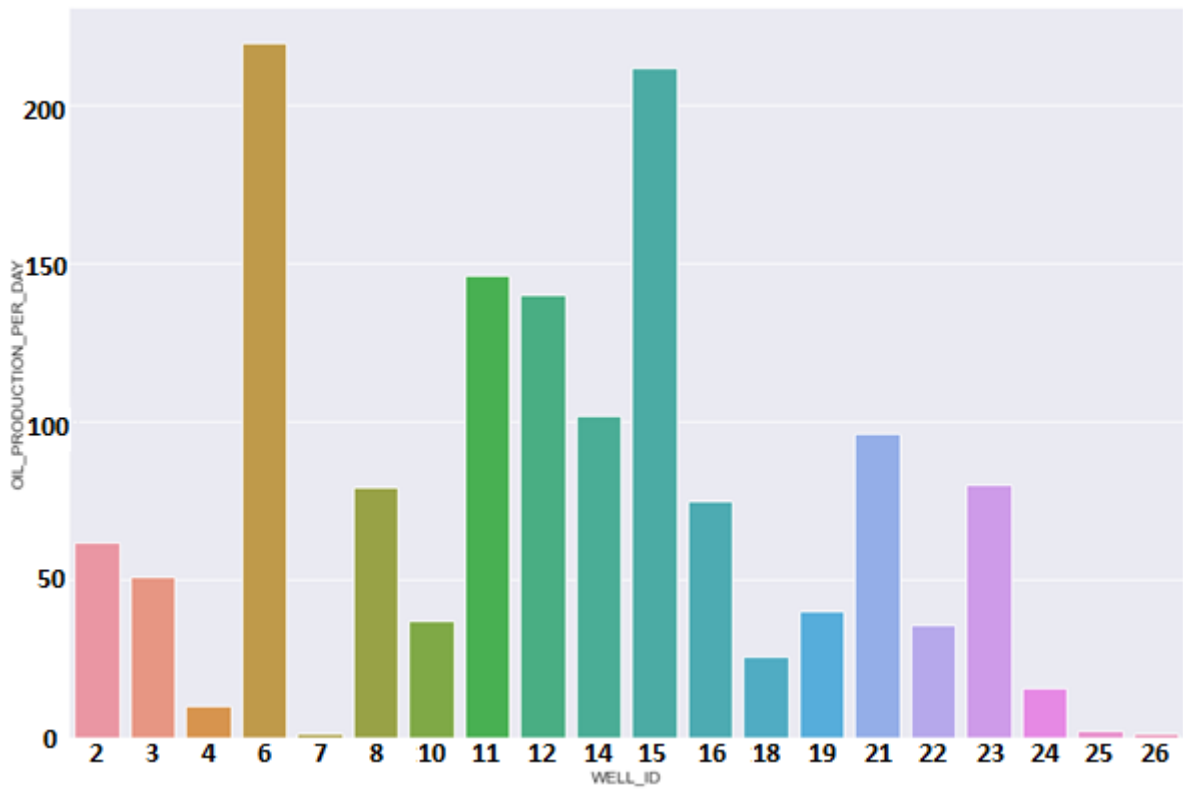


Figure 25: Oil Production per Day per Well in bbl/day

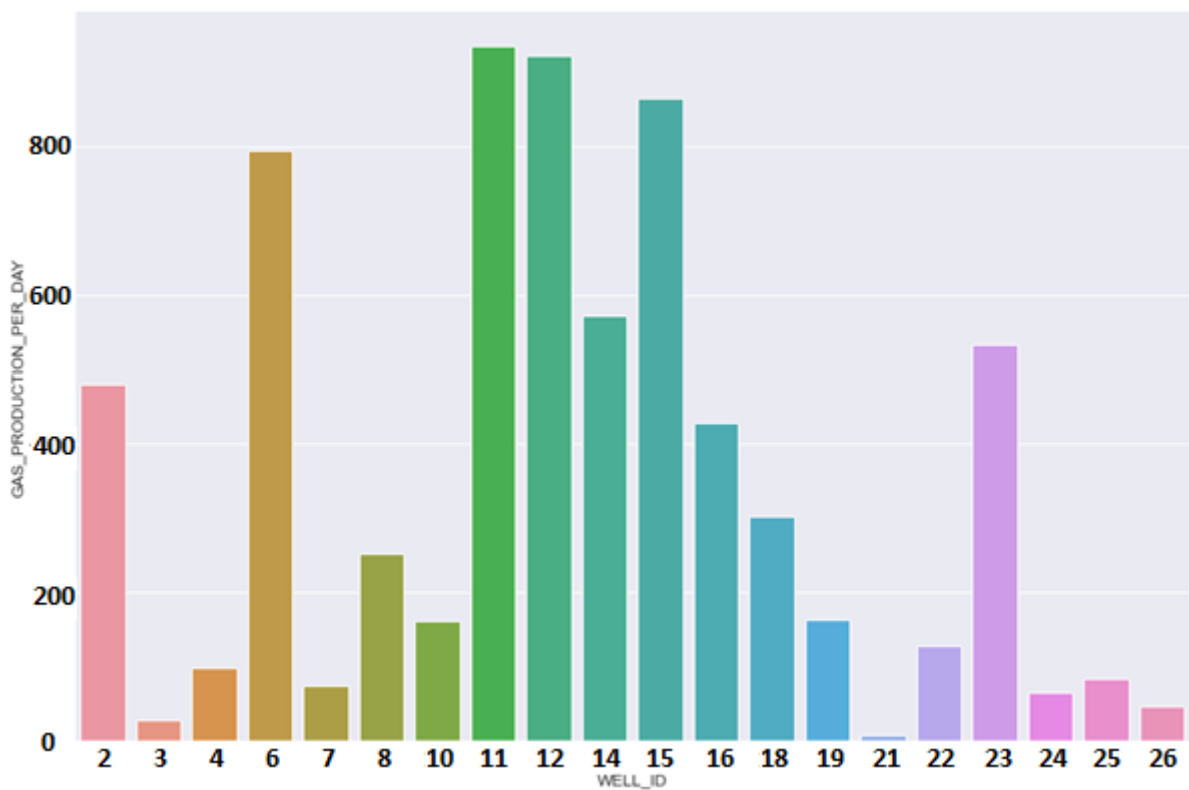


Figure 26: Gas Production per Day per Well in Mcf/day

Since the available data contains the latitude and longitude of each well, it is possible to visualize the impact of the location of the wells on the oil and gas production. Plotting the well

latitude versus well longitude can give an idea about the importance of well location. Figure 27 shows the location of the training wells and their oil and gas production.

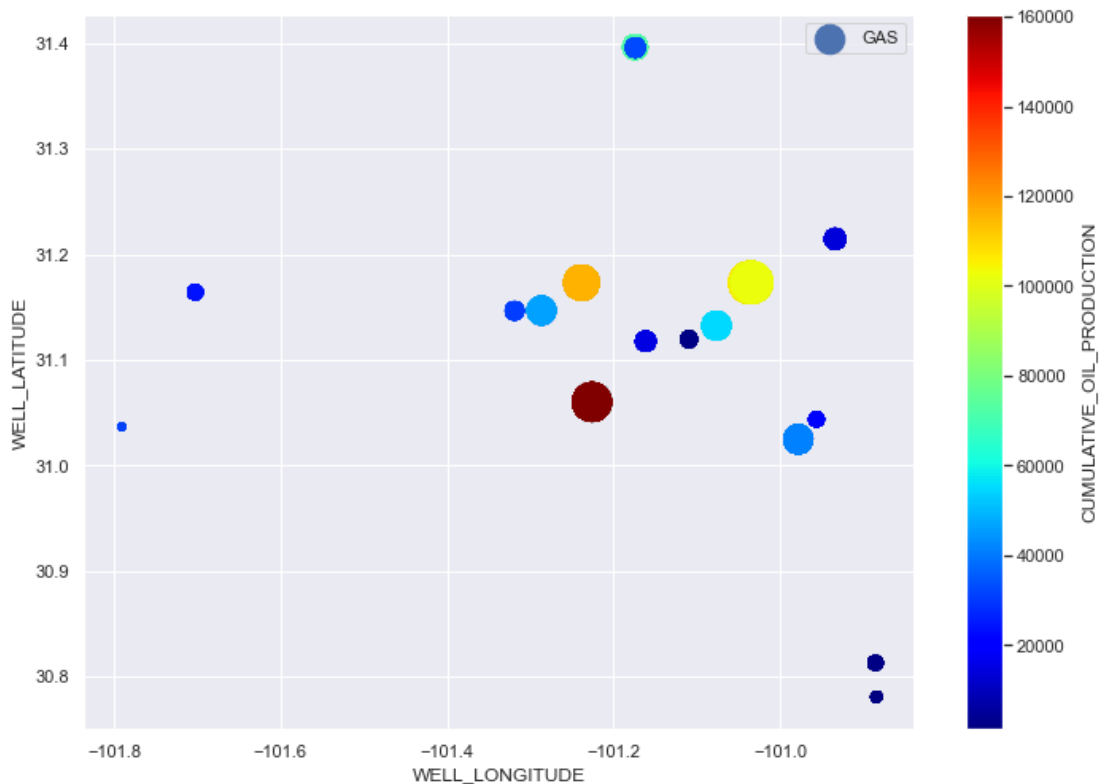


Figure 27: Effect of the Well Location on the Oil and Gas Production

In figure 27, the x axis is the well longitude, and the y axis is the well latitude. Each circle represents one of the training wells. The colour of the circle indicates the cumulative oil production of this well. Blue represents low oil production, yellow represents medium production and red represents high production, as shown in the colour bar on the right. The cumulative gas production is indicated with the size of each circle. The small circles represent wells that produce small volumes of gas, while the bigger circles represent wells that produce large volumes of gas. The dark red, large circle in the centre represents well 6, which is both a very good oil producer and a very good gas producer.

Using figure 27, it is possible to conclude that wells that are located on the edges of the map are generally bad producers. Most wells are located close to each other, and for these wells it is not possible to see a correlation between well location and oil and gas production. The relationship between oil production and gas production and well latitude and longitude will be further explored using Pearson's correlation coefficient.

The oil and gas production depend on a number of features. Before creating the model, it is important to explore which features have the highest impact on the production of the training wells. Some of these features are categorical, while most of them are numerical.

### 4.3.2 Categorical Variables

The data set contains 3 columns with categorical variables. These three columns contain information about the proppant mesh size, the proppant type, and the fracturing fluid used. Using the describe function from pandas, table 12 can be generated. “Count” refers to the number of values in each column. Since none of the columns have missing data, the count is 1179 for all 3. “Unique” is the number of unique values in each column, “Top” is the most frequent value, and “Frequency” is the frequency of the top value.

Table 12: Description of the Columns Containing Categorical Variables

	PROPPANT_MESH_SIZE	PROPPANT_TYPE	FRAC_FLUID
Count	1179	1179	1179
Unique	4	2	3
Top	100	White Sand	Water
Frequency	525	1162	837

Table 12 shows that there are 9 categorical variables. The impact of these categorical values on the cumulative oil and gas production can be examined using different plotting methods. However, these features cannot be used in their current form, and they need to be one-hot encoded to be used in the model training. The 3 columns that contain them will be replaced with 9 columns for each of these attributes.

- **Impact of Different Categorical Variables on the Oil and Gas Production**

Proppant type, proppant mesh size and fracturing fluid are all very important design elements of any hydraulic fracturing operation. These elements are expected to have an impact on the oil and gas production of any hydraulically fractured well.

The best method to analyse the impact of a categorical variable on a quantitative variable is using box plots. A box plot shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The box shows the quartiles of the data set while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” using a method that is a function of the inter-quartile range. Figure 28 explores the impact of the fracturing fluid type on oil production, while Figure 29 explores its impact on gas production.

Figures 28 and 29 show that the production when using 100 mesh proppant size ranges from low to high, which means that the use of this proppant mesh size might not be a good predictor of oil and gas production. On the other hand, the use of 40/70 and 30/50 mesh sizes can be correlated with a relatively lower oil and gas production, while the use of the 20/40 mesh size can be correlated with a higher production. In addition to the proppant mesh size, the impact of the proppant type can also be seen using the same box plots.

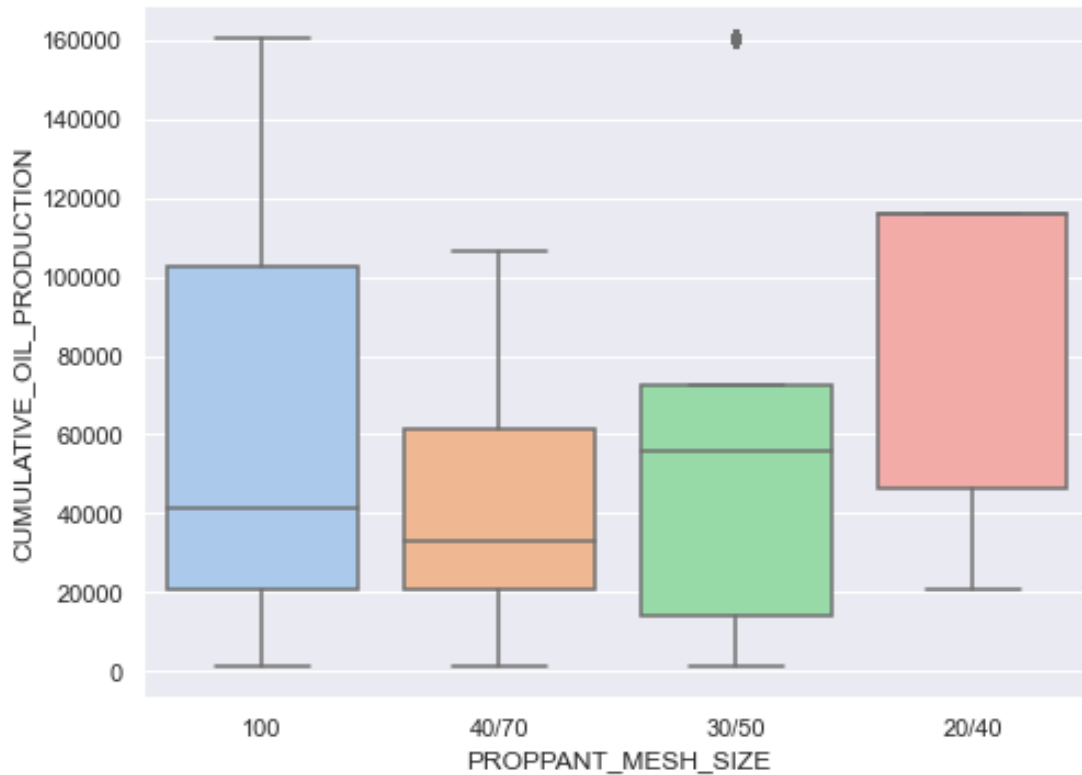


Figure 28: Impact of the Proppant Mesh Size on Oil Production

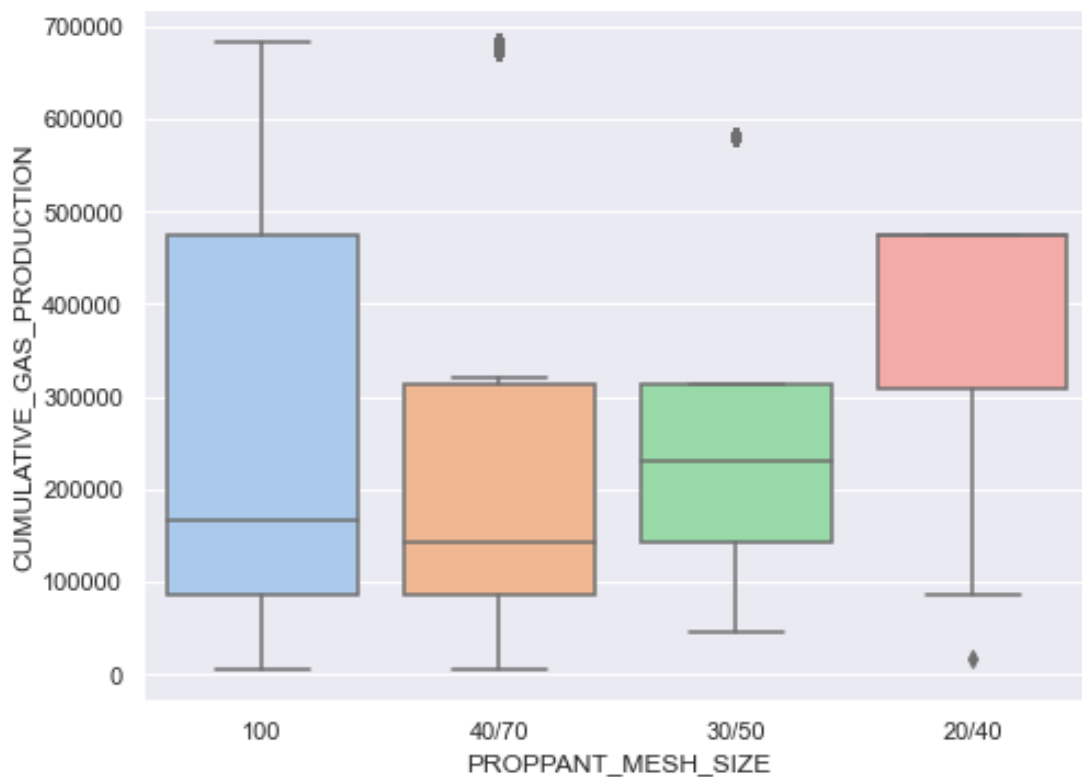


Figure 29: Impact of the Proppant Mesh Size on Gas Production

Figures 30 and 31 explore the relationship between proppant type and oil and gas production, respectively. They show that the use of brown sand is generally correlated with a lower oil and



gas production. This can be expected since brown sand is considered of lower quality than white sand.

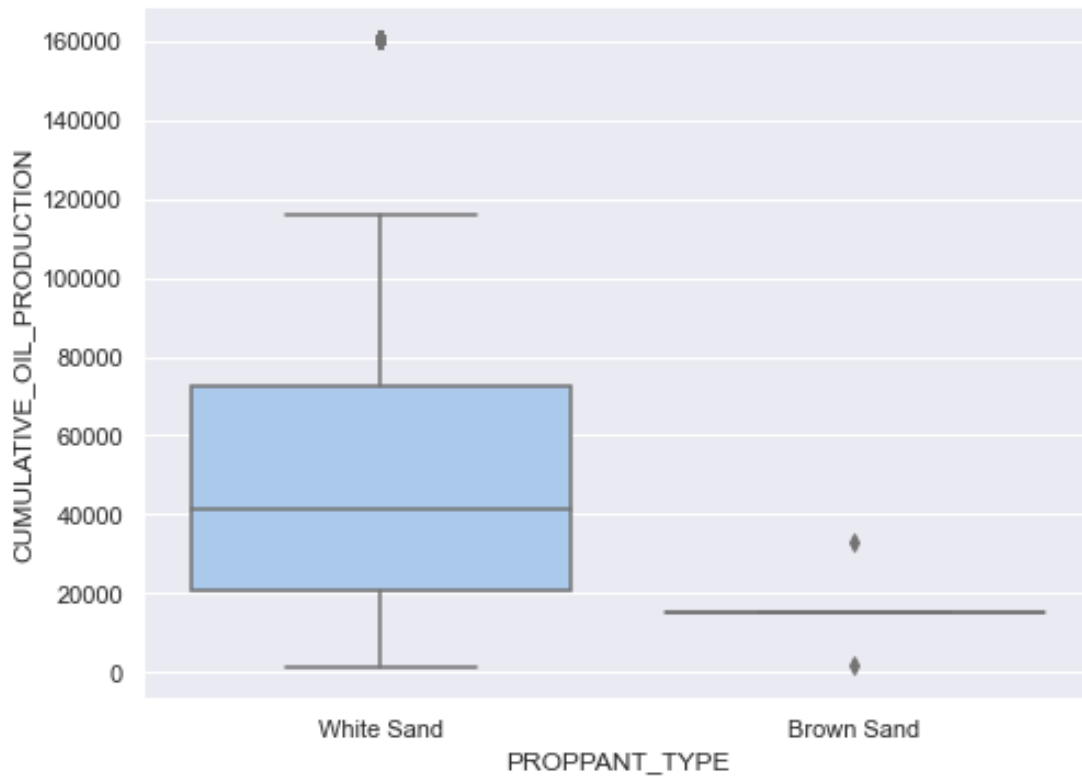


Figure 30: Impact of the Proppant Type on Oil Production

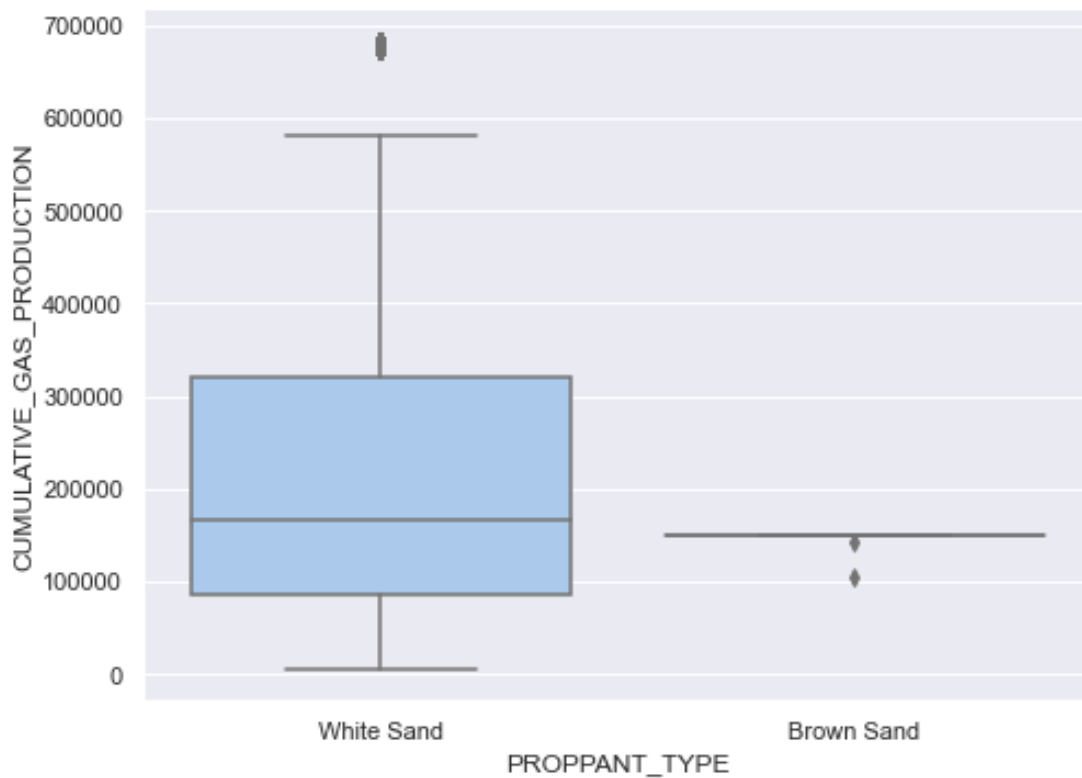


Figure 31: Impact of the Proppant Type on Gas Production

Figures 32 and 23 explore the impact of the fracturing fluid on the oil and gas production.

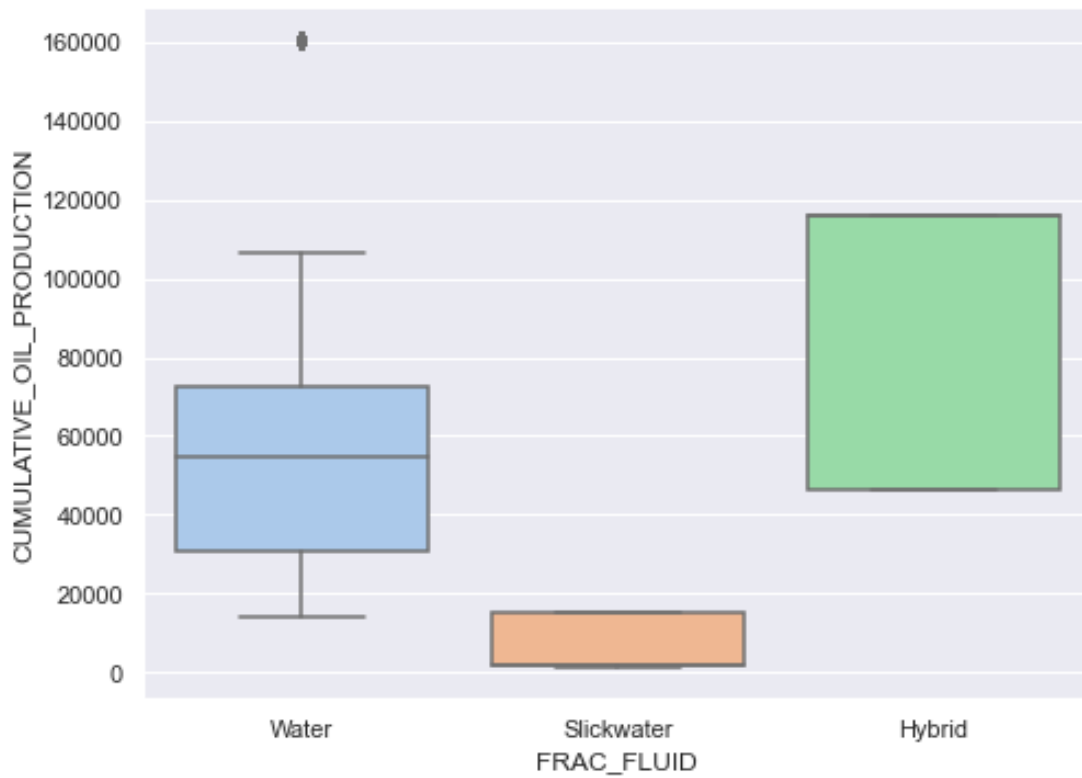


Figure 32: Impact of the Fracturing Fluid Type on Oil Production

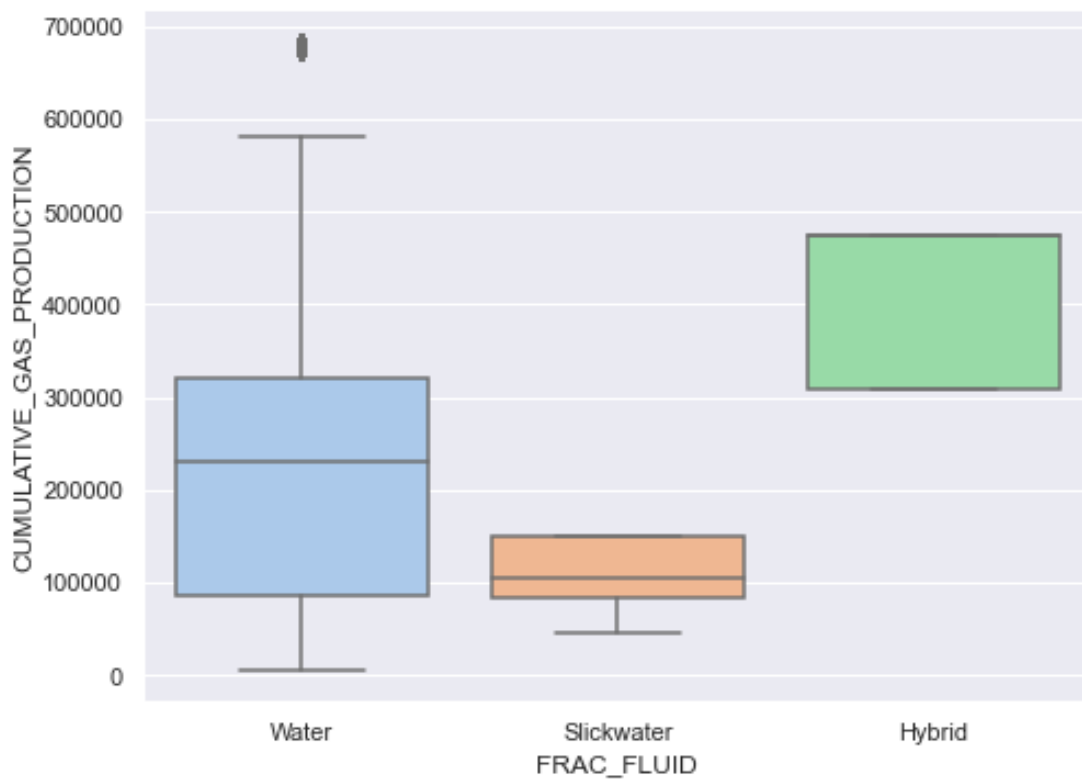


Figure 33: Impact of the Fracturing Fluid on Gas Production

Figures 32 and 33 explore the impact of fracturing fluid type on the oil and gas production. They show a clear correlation between the fracturing fluid used and the oil and gas production. The use of Slickwater results in low oil and gas production, the use of water results in a low to medium production, and the use of hybrid results in a high production.

The figures above show that the categorical variables can have a big impact on the oil and gas production. It is also possible to conclude that fracturing fluid has the most impact on production of all the categorical variables. However, it is not advised to use them as text variables, which means that they need to be encoded.

- **One-Hot Encoding**

Categorical variables are not a great choice to use in machine learning in their current form. Most machine learning algorithms prefer to work with numbers rather than words, so it is advised to transform these categorical variables from text to numbers. While it is easy to give a number for each category (Exp: “Water”: 1, “Slickwater”: 2, “Hybrid”: 3), this approach, called Ordinal Encoding, is generally not useful when using Scikit-Learn. The reason is that the package’s models make the fundamental assumption that numerical features reflect algebraic quantities. Such a solution would imply, for example, that mathematical operations or comparisons are possible, such as Water < Slickwater < Hybrid, or even that Hybrid - Water = Slickwater, which does not make sense.

To fix this issue, a common solution is to use one-hot encoding, which effectively creates extra columns indicating the presence or absence of a category with a value of 1 or 0, respectively. This is called one-hot encoding, because only one attribute will be equal to 1 (hot), while the others will be 0 (cold). Table 13 shows an example of one-hot encoding of the column proppant mesh size.

Table 13: Example of One-Hot Encoding of the Column Proppant Mesh Size

<b>Original Column Values</b>	<b>Columns Created with One-Hot Encoding</b>			
<b>PROPPANT_MESH_SIZE</b>	<b>100</b>	<b>40/70</b>	<b>30/50</b>	<b>20/40</b>
100	1	0	0	0
40/70	0	1	0	0
30/50	0	0	1	0
40/70	0	1	0	0
20/40	0	0	0	1

As shown in the table above, the goal of one-hot encoding is to create one binary attribute per category: one attribute equal to 1 when the category is “40/70” “Water”, or “White sand” (and 0 otherwise), another attribute equal to 1 when the category is “30/50”, “Slickwater” or “Brown Sand” (and 0 otherwise), and so on. The new attributes are sometimes called dummy attributes. Therefore, the 3 columns containing categorical attributes (“PROPPANT\_TYPE”, “PROPPANT\_MESH\_SIZE”, “FRAC\_FLUID”) will be transformed into 9 columns representing

each one the categorical variables (100 mesh, 40/70 mesh, 30/50 mesh, 20/40 mesh, White sand, Brown sand, Water, Slickwater, Hybrid).

The categorical variable that has the highest impact on oil and gas production is fracturing fluid. The impact of numerical variables on production is more easily determined, and the most important categorical and numerical features will be used to create the model.

### 4.3.3 Numerical Variables

The data set contains 19 numerical variables. In order to visualize their impact on the oil and gas production, it is possible to use the correlation coefficient. The most promising attributes can then be visualized using Regplots. As with categorical variables, it is advised to transform the numerical variables since they have very different scales.

- **Impact of Different Numerical Variables on the Oil and Gas Production**

The impact of the numerical attributes on the oil and gas production can be determined by calculating the correlation coefficient between each numerical variable and the oil and gas production. The correlation coefficient (Pearson's correlation coefficient) is the most familiar measure of dependence between two quantities. Mathematically, it is defined as the quality of least squares fitting to the original data (Wikipedia 2021a). Figure 34 shows some examples of plots with different correlation coefficients.

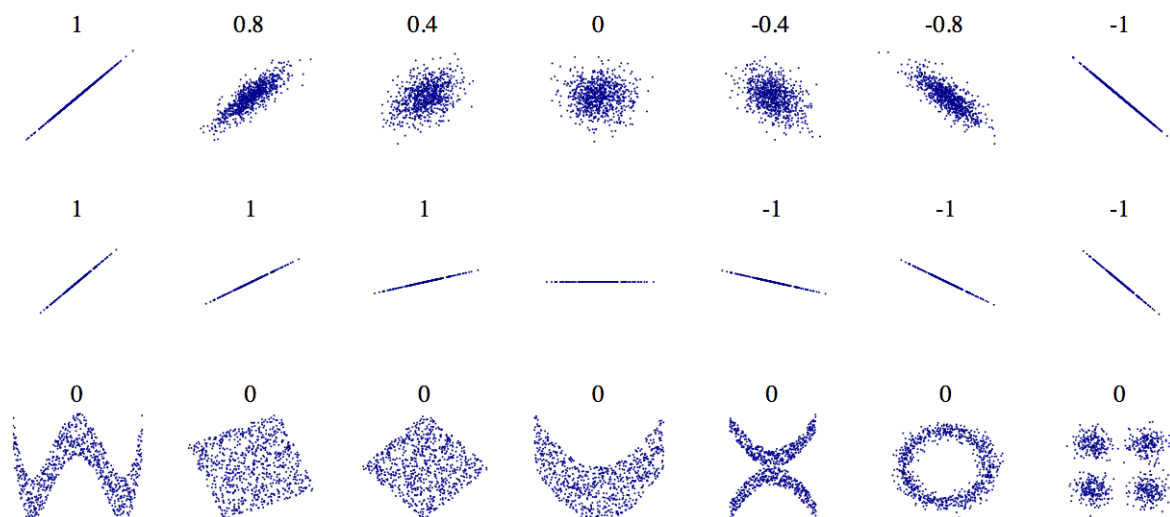


Figure 34: Standard correlation coefficient of various data sets<sup>1</sup>

The correlation coefficient only measures linear correlations (“if x goes up, then y generally goes up/down”). It may completely miss out on nonlinear relationships (e.g., “if x is close to 0,

---

<sup>1</sup> Wikipedia 2021a.

then  $y$  generally goes up”). A correlation coefficient of 0 does not mean that no relationship exists between the two variables. The first row of figure 34 shows examples of data points with different correlation coefficient. The second row presents examples where the correlation coefficient is equal to 1 or  $-1$ , which proves that the correlation coefficient is not dependent on the slope. For example, the height in inches has a correlation coefficient of 1 with the height in feet or in meter. The third row contains some plots that have a correlation coefficient of 0, even though a clear dependence exists between the variables. These plots are examples of nonlinear relationships.

Table 14 presents the correlation coefficient of all the numerical attributes with the oil and gas production. The blue colour indicates a positive correlation coefficient, while the red colour indicates a negative correlation coefficient. The columns are classified in descending order according to the absolute value of the correlation coefficient of the column with the oil production. Table 14 shows a high correlation between the oil production and the gas production; however this will not be used since no production information for the target wells is provided in the data set. The most important variables are the net production days and the proppant mass used. The average STP has also a high correlation with the oil production.

Table 14: Correlation Coefficient between Numerical Variables and Oil and Gas Production

Column	Correlation with Oil Production	Correlation with Gas Production
CUMULATIVE_OIL_PRODUCTION	1	0,835294
CUMULATIVE_GAS_PRODUCTION	0,835294	1
NET_PROD_DAYS	-0,461236	-0,339409
PROPPANT_MASS_USED	-0,293569	-0,392082
AVERAGE_STP	0,288627	0,148405
LOWER_PERF	0,278193	0,126694
TOP_DEPTH	0,269568	0,129983
MIN_STP	0,262015	0,183655
WELL_LATITUDE	0,257695	0,194977
DAY_NUMBER	0,176475	0,190545
MAX_STP	0,121172	0,090253
TVD_DEPTH	0,117474	-0,171793
STAGE_NUMBER	0,104202	0,094947
MD_MIDDLE_PERFORATION	0,102228	0,023463
FRACTURE_GRADIENT	0,094523	-0,013249
WELL_LONGITUDE	-0,071168	0,302984
UPPER_PERF	-0,029644	-0,259796
WELL_HORZ_LENGTH	0,020388	-0,014334
VOLUME_PUMPED_GALLONS	0,018814	0,0314

The relationship between these variables and the oil and gas production can be further analysed using Regplots, which plot data and a linear regression model fit. Figure 35 shows the Regplot of the cumulative oil production vs average pressure.

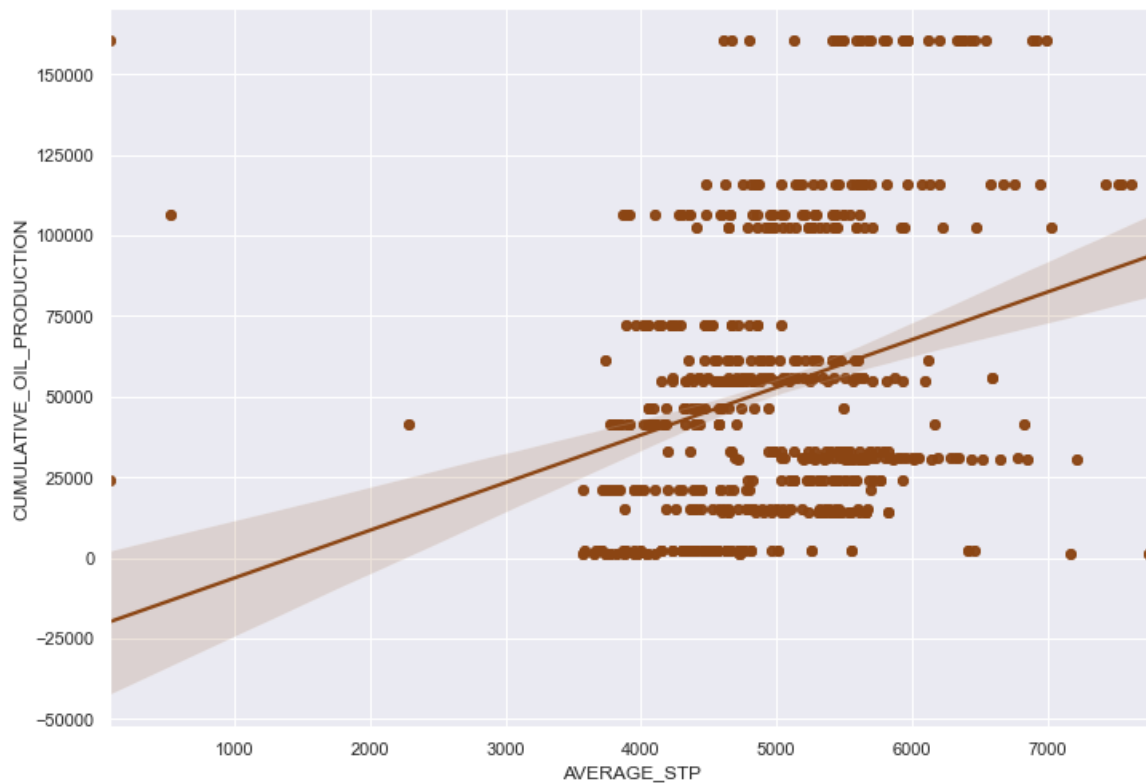


Figure 35: Cumulative Oil Production (bbl) vs Average Pressure (psi)

It shows an increase of the cumulative oil production with increasing average pressure. Since the hydraulic fracturing is multistage, the cumulative oil production does not change for each well while the average pressure changes depending on the fracturing stage. For example, well 2 has a cumulative oil production of 41307 barrels, which is constant. However, well 2 has a different average pressure for each hydraulic fracturing stage, which explains why there are multiple average pressure values for the same cumulative production value. The average pressure is not highly correlated with gas production. On the other hand, the proppant mass used is highly correlated with both the oil production and the gas production.

Figures 36 and 37 show the Regplots of the cumulative oil production vs proppant mass used and the cumulative gas production vs proppant mass used, respectively. Figure 36 shows a clear, and relatively high negative correlation between the cumulative oil production and the proppant mass used. Most wells are fractured using a proppant mass ranging from 0 to 800 cwt. The only exception is well 3, where more than 2000 cwt were used to conduct the hydraulic fracturing. Well 3 has a relatively low cumulative oil production, while well 6 has the highest cumulative oil production and a very low proppant mass used. This can explain the negative slope of the curve. Figure 37 shows that the cumulative gas production decreases with increasing proppant mass used. The slope is steeper than the slope of the cumulative oil production vs proppant mass used. This can be explained by the low cumulative gas production of well 3, and the high cumulative gas production of wells 6, and 12. These two wells have been hydraulically fractured using a mass of 250 cwt or less for each stage.

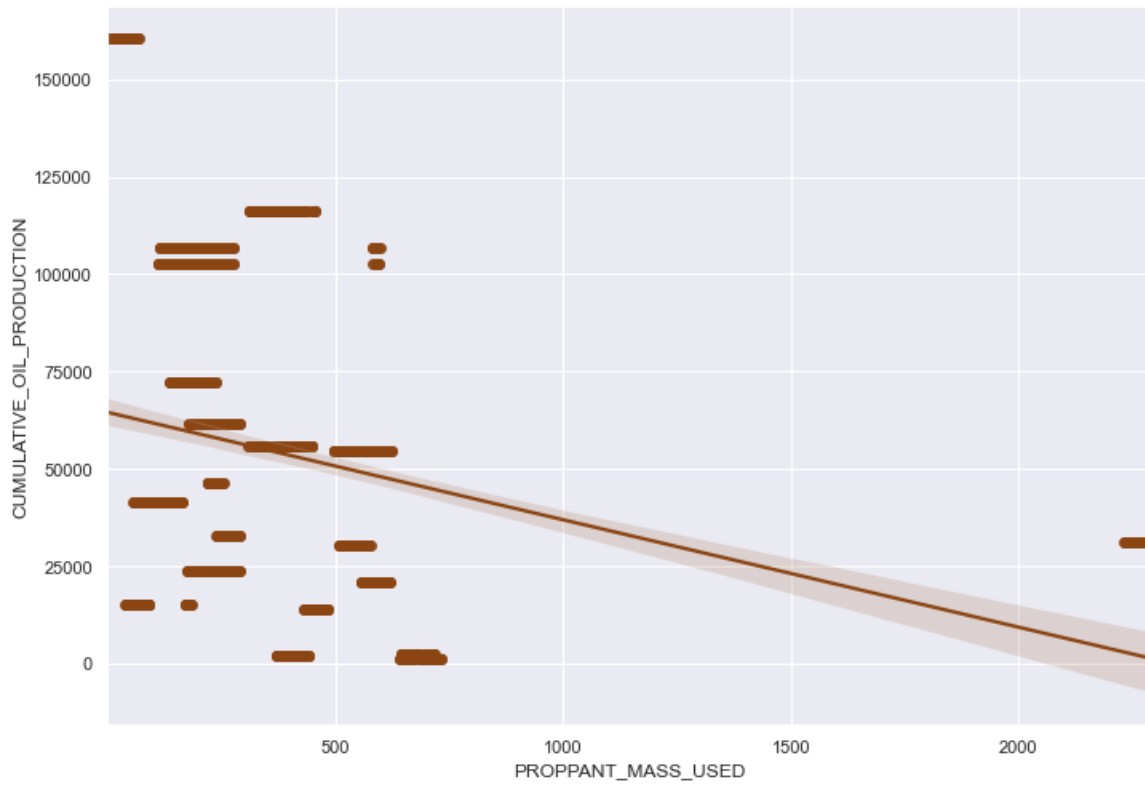


Figure 36: Cumulative Oil Production (bbl) vs Proppant Mass Used (cwt)

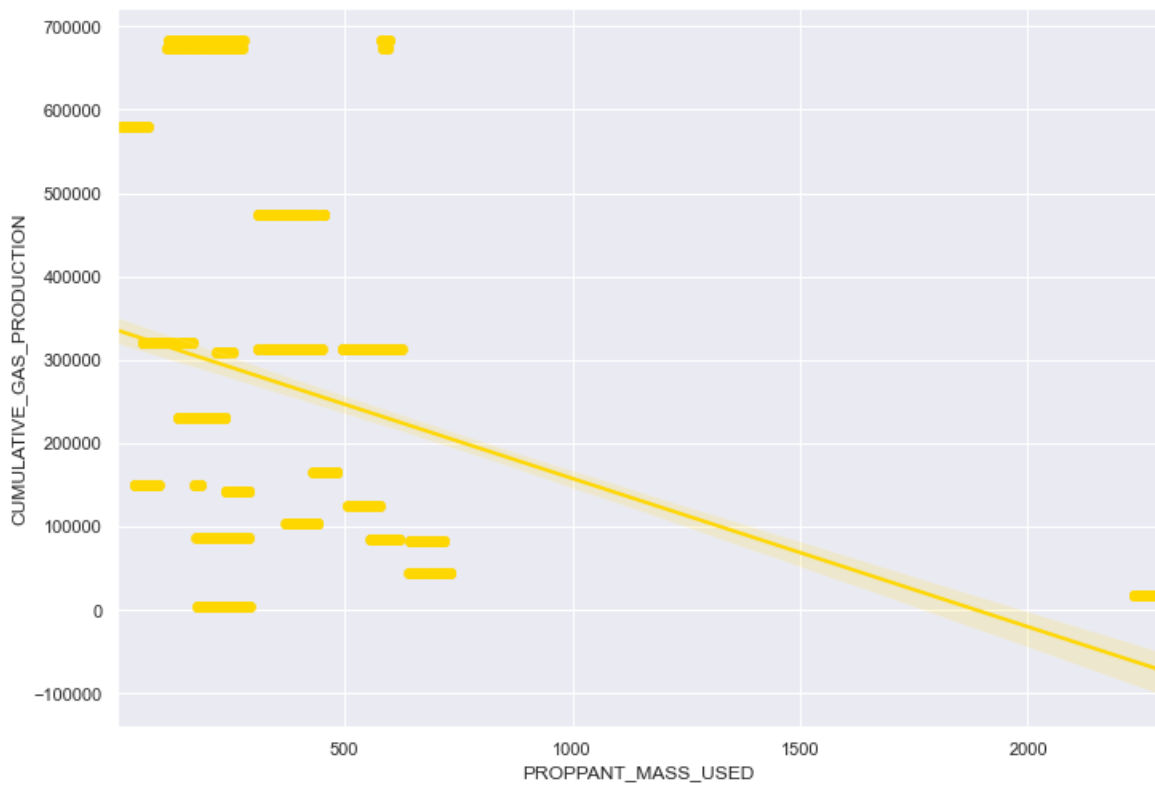


Figure 37: Cumulative Gas Production (Mcf) vs Proppant Mass Used (cwt)

Figures 38 and 39 show the Regplot of cumulative oil production vs net production days and cumulative gas production vs net production days, respectively.

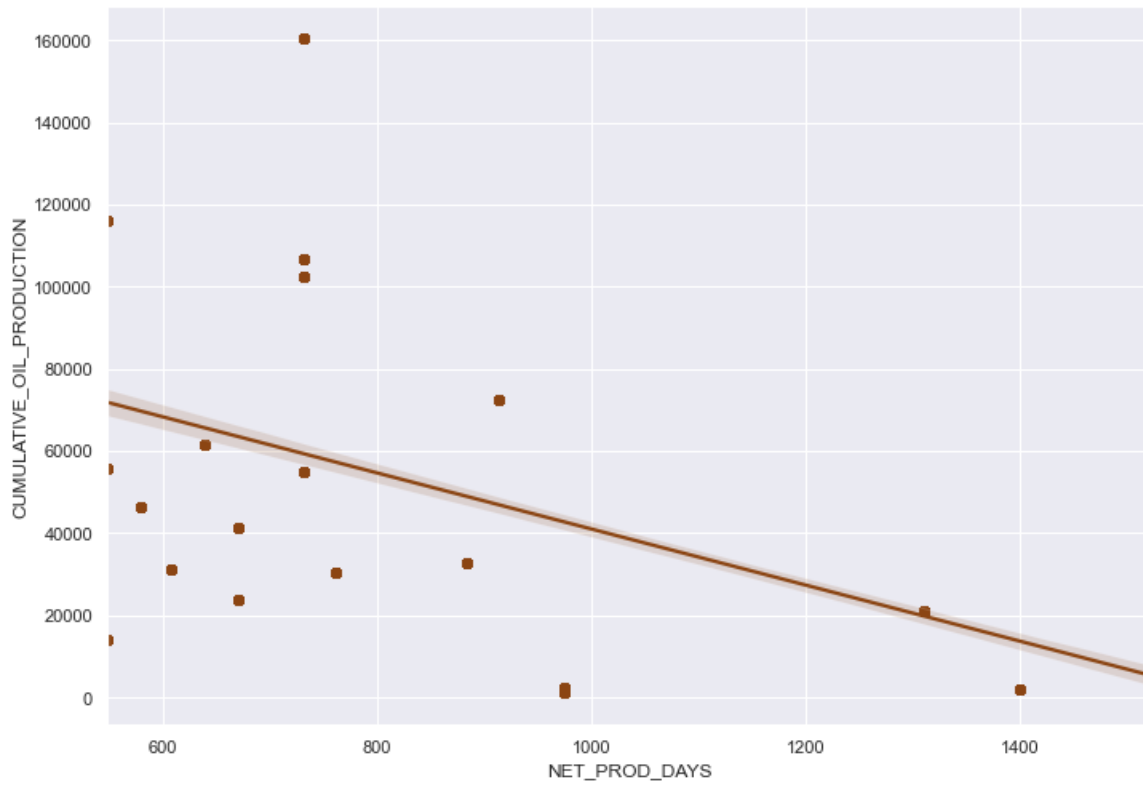


Figure 38: Cumulative Oil Production (bbl) vs Net Production Days (days)

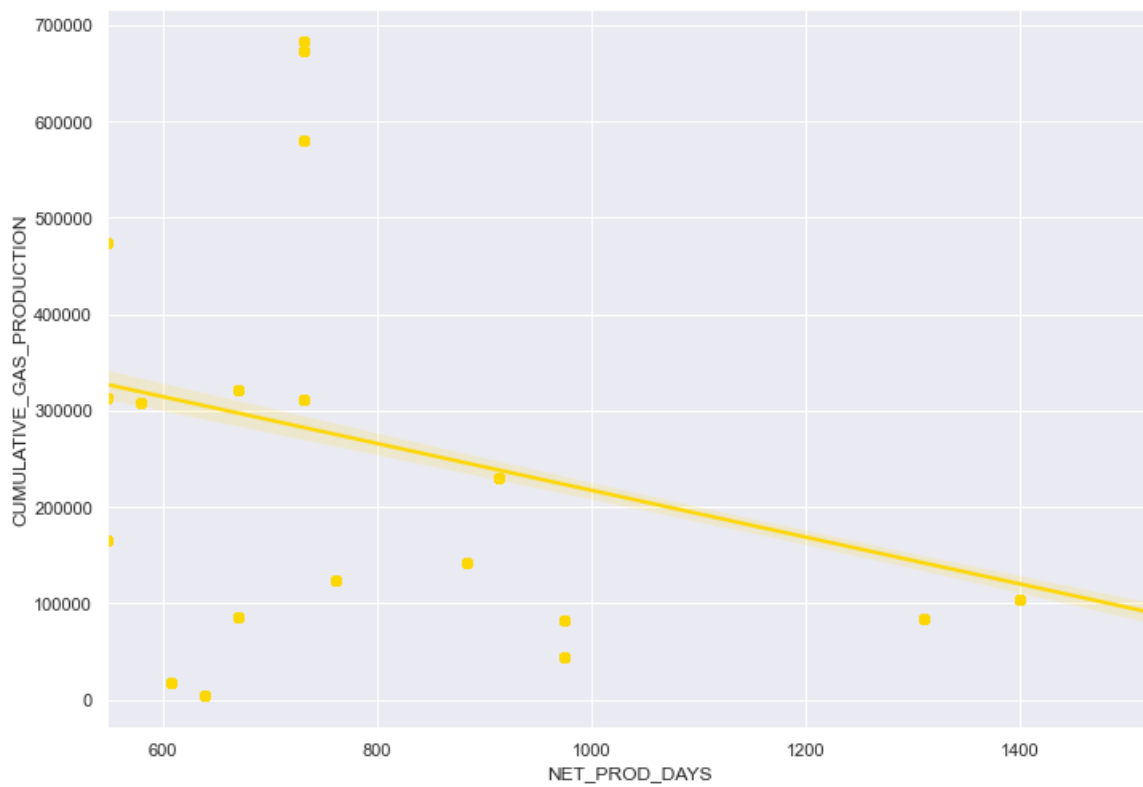


Figure 39: Cumulative Gas Production (Mcf) vs Net Production Days (days)



Figure 38 shows a negative correlation between cumulative oil production and the number of production days. The x axis starts at the value 500 days. Three wells have a big impact on the negative slope. Wells 4, 7 and 24 have a relatively low cumulative oil production and a very high number of production days. While cumulative production should normally increase with the number of production days, there can be some possible explanation for this negative slope. One possible explanation is that it makes sense to do a new hydraulic fracturing job for the wells that produce good volumes of oil, since it can be worth the investment. This may be the reason that the three best oil producers are only produced for periods of less than 750 days.

Figure 39 shows that the correlation between the cumulative gas production and the number of production days is also negative. The x axis starts at the value 500 days. The explanation of this negative slope can be the same as with figure 38. However, the slope is less steep than in figure 38.

While the correlation coefficient is an interesting data exploration techniques, it does not always mean that the features with the highest correlation coefficient will be the most impactful when the model is trained. In fact, different machine learning algorithms can find different importance for the different attributes. The most important features found with the correlation coefficient will later be compared with the most important features selected by sequential forward selection, sequential backward elimination and the feature importance calculated by the random forest regression algorithm in part 6 of Chapter 5.

- **Creation of New Attributes**

In addition to the attributes already provided, it is possible to create new attributes based on the ones already existing. There is an infinite number of attributes that can be created, but the new attributes added should have an added value. One possible addition is to create a new column with the total number of days and the total number of stages of the fracturing job for each well. The impact of the new attributes is shown in Table 15, along with the original attributes. Total days and total stages have the second and third highest correlation coefficient with the cumulative oil production. For the cumulative gas production, total days has the highest correlation coefficient, while the total stages attribute has the fourth highest correlation coefficient.

The total number of days and the total number of stages of the hydraulic fracturing operation for each well presents a higher correlation coefficient than the day number and stage number attributes. This means that the added attributes can be useful for the model training and will therefore be added to the target data set.

As with categorical variables, it is advised to make some transformations to the numerical variables before using them to predict with a machine learning algorithm. The reason is that numerical variables have very different scales.

Table 15: Correlation Coefficient between Numerical Variables and Oil and Gas Production with New Attributes

Column	Correlation with Oil Production	Correlation with GasProduction
CUMULATIVE_OIL_PRODUCTION	1	1
CUMULATIVE_GAS_PRODUCTION	0,835294	0,835294
NET_PROD_DAYS	-0,461236	-0,339409
TOTAL_DAYS	0,392714	0,435532
TOTAL_STAGES	0,360284	0,328442
PROPPANT_MASS_USED	-0,293569	-0,392082
AVERAGE_STP	0,288627	0,148405
LOWER_PERF	0,278193	0,126694
TOP_DEPTH	0,269568	0,129983
MIN_STP	0,262015	0,183655
WELL_LATITUDE	0,257695	0,194977
DAY_NUMBER	0,176475	0,190545
MAX_STP	0,121172	0,090253
TVD_DEPTH	0,117474	-0,171793
STAGE_NUMBER	0,104202	0,094947
MD_MIDDLE_PERFORATION	0,102228	0,023463
FRACTURE_GRADIENT	0,094523	-0,013249
WELL_LONGITUDE	-0,071168	0,302984
UPPER_PERF	-0,029644	-0,259796
WELL_HORZ_LENGTH	0,020388	-0,014334
VOLUME_PUMPED_GALLONS	0,018814	0,0314

- **Feature Scaling of Numerical Attributes**

Similarly to categorical variables, numerical variables generally cannot be used in their current form. They need to be transformed using feature scaling. Feature scaling is a very important step in data preparation. Machine learning algorithms generally perform better when the data is in the same scale and perform poorly when the data is in different scales (Géron 2019). There are two main ways to transform the numerical features to the same scale. The first method is called normalization or Min-Max shifting. It consists in shifting the values and rescaling them to a range from 0 to 1. This can be done by subtracting the minimum value and then dividing by the maximum value minus the minimum value. The second method is called standardization. It consists in subtracting the mean value and the dividing by the standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. Standardization does not present a specific range, as with normalization. However, the advantage is that standardization is less affected by outliers. For example, if the attribute has a range from 0 to 12 but contains one outlier which has a value of 100, normalization will mean that the outlier would become 1 and all the other values will be in the range 0 to 0.12. For this reason, standardization is the selected method for feature scaling.

Visualizing the data is not sufficient to determine the real impact of categorical variables on the prediction of the machine learning algorithm. Many methods exist to determine which variables have the higher impact on the production, which will be discussed in the feature importance part of Chapter 5.

## 5 Model Creation and Results

Once the data is scaled, it is possible to train a model and make the prediction for the oil and gas production of the target wells. Different machine learning models exist to predict the production of these wells. For this thesis, 4 models were trained, and their results are compared: a linear regression model, a decision tree regression model, a random forest regression model, and a support vector machine model.

The training data set consists of 20 wells. In order to measure the performance of the model, it is important to create a subset of wells to build the model (18 wells), called a learning set, and a smaller subset (2 wells), called a test set, to test the model. Cross validation will be used for the random forest regression model and the support vector machine regression model since they are considered more sophisticated. Cross validation is important to make sure that the model is not overfitted to the particular wells used in the training of the model.

The importance of the features used in the training of these models can be determined once the best model has been chosen. The most important features selected by sequential forward selection and sequential backward elimination will be compared to the best features selected by the random forest algorithm and to the correlation coefficient. Finally, the cumulative oil production and the cumulative gas production of the target wells will be predicted.

This chapter presents the models created and the results obtained. The first part is an overview of the cross-validation method used to fine tune the best models. The second, third, fourth and fifth part present the linear regression, decision tree regression, random forest regression and support vector machine regression models. In these parts, an overview of the models and their performance will be discussed. The sixth part discusses the feature importance as determined by different methods. The seventh part contains the final results of this thesis.

### 5.1 K-Fold Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. This procedure has only one parameter,  $k$ , which refers to the number of groups that the data set will be split into (Brownlee 2018).

In this thesis,  $k$  is 5, which means that the data set is divided into 5 groups. Cross-validation is primarily used in machine learning to estimate the skill of a machine learning model on unseen data. The objective is to examine the performance of the model on data not used in the training. It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. Figure 40 shows the procedure of  $k$ -fold cross validation.

As shown in figure 40, the steps of  $k$ -fold cross validation are (Brownlee 2018):

1. The data set is shuffled randomly.
2. It is then split into  $k$  groups.

3. Each group is considered as a test data set, while the remaining groups are considered as a training data set.
4. The model is fit on the training data set and then evaluated on the test data set. The performance of the model is determined using an evaluation score.
5. Each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is used in the hold out set 1 time and used to train the model k-1 times.



Figure 40: k-fold Cross Validation Procedure<sup>1</sup>

Cross validation can also be used to determine the best parameters for the machine learning algorithm. For this reason, 5-fold cross validation will be used to determine the best parameters in the decision tree regression and the support vector machine regression model.

## 5.2 Linear Regression

Regression is the process of predicting a continuous value. In regression there are two types of variables: a dependent variable and one or more independent variables. The dependent variable is the target of the prediction. The independent variables, also known as explanatory variables, can be seen as the causes of those states. The independent variables are shown conventionally by  $X$  and the dependent variable is notated by  $Y$ . A regression model relates  $Y$  to a function of  $X$ . The key point in the regression is that the dependent value should be continuous and cannot be a discrete value. However, the independent variable, or variables, can be measured on either a categorical or continuous measurement scale. The simplest form of linear regression is fitting a straight line to data. The straight-line fit is a model of the form  $y$  equals  $b$  plus  $a$  times  $x$  ( $y = b + ax$ ) where  $a$  is commonly known as the slope, and  $b$  is commonly

<sup>1</sup> Wikipedia 2021b.

known as the intercept. Figure 41 shows a random data set, and figure 42 shows the linear regression model fitted to this data.

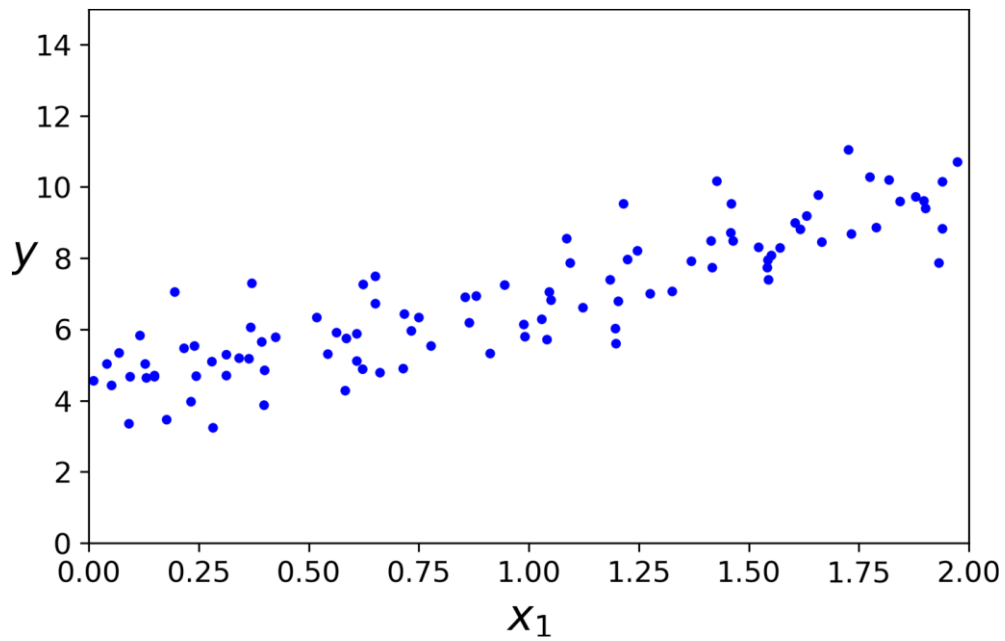


Figure 41: Random Data for Linear Regression<sup>1</sup>

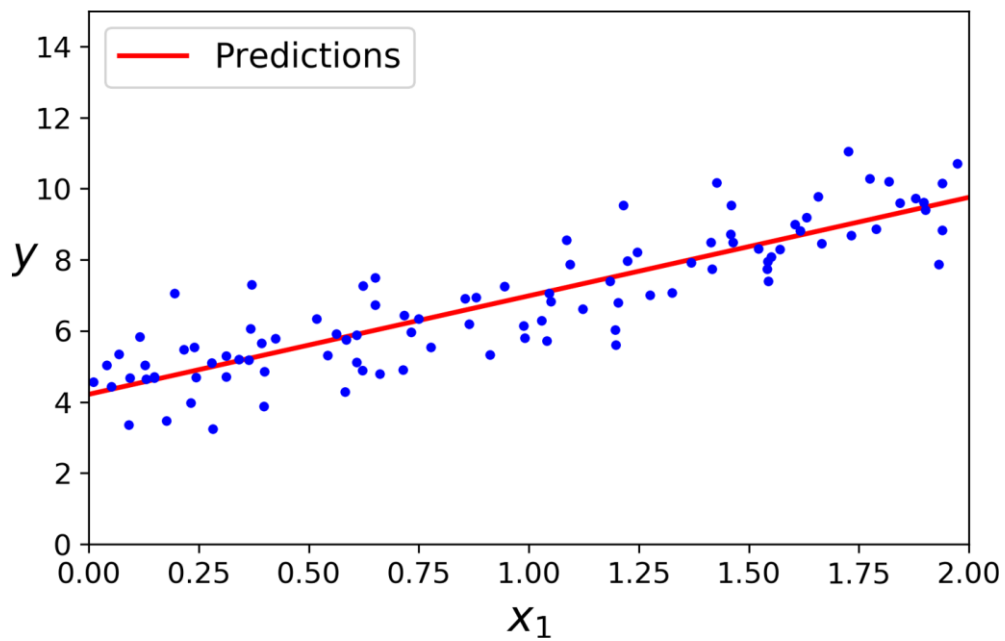


Figure 42: Data with Fitted Linear Regression Model<sup>2</sup>

---

<sup>1</sup> Géron 2019.

<sup>2</sup> Géron 2019.

Linear regression generally does not give the most accurate predictions since it is the simplest form of regression. However, since it is the simplest model, it is a good starting point to see how close the predictions are. The accuracy of a regression model is determined by comparing the actual values and the predicted values. There are many different evaluation metrics that are used to determine how accurate the model is. Typical performance measures are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared:

- Mean Absolute Error (MAE): It is the mean of the absolute value of the errors. This is the easiest of the metrics to understand since it is just average error.
- Mean Squared Error (MSE): MSE is the mean of the squared error. It is more popular than Mean Absolute Error because the focus is more towards large errors. This is due to the squared term exponentially increasing larger errors in comparison to smaller ones.
- Root Mean Squared Error (RMSE): RMSE is more sensitive to outliers than MEA.
- R-squared is not an error but is a popular metric for accuracy a model. It represents how close the data are to the fitted regression line. The higher the R-squared, the better the model fits the data. Best possible score is 1.0 and it can be negative.

Both the RMSE and the MAE are measure the distance between two vectors: the vector of predictions and the vector of target values. Various distance measures, or norms, are possible. Calculating the root of a sum of squares (RMSE) corresponds to the Euclidean norm. Euclidean norm is the length of a line segment between the two points. Computing the sum of absolutes (MAE) corresponds to the Manhattan norm because it measures the distance between two points in a city if only travel along orthogonal city blocks is possible. The Manhattan norm is less sensitive to outliers. Comparing the different evaluation metrics of each model is the best way to determine the best model to choose.

As expected, the performance of the linear regression model was very poor when predicting the cumulative oil production and the cumulative gas production of the two wells of the testing set. Table 16 presents the performance of the linear regression model on the testing set.

Table 16: Performance of the Linear Regression Model

<b>Model Performance</b>	<b>Oil Production</b>	<b>Gas Production</b>
Training Data Set Mean Value	49613 Bbl	245236 Mcf
Training Data Set Standard Deviation	43107 Bbl	211168 Mcf
MAE Test Set	30537 Bbl	121010 Mcf
RMSE Test Set	34812 Bbl	148340 Mcf
r2 Test Set	- 0.70	- 0.28

Table 16 shows that the model is not accurate in predicting both the oil production and the gas production. In fact, the model presents negative r2 scores for both the oil and gas production.

This means that the model does not follow the trend of the data, which means that it fits worse than a horizontal line.

### 5.3 Decision Tree Regression

Decision trees are versatile machine learning algorithms. They can be used for both classification and regression tasks. Tree models where the target variable is a discrete set of values are called classification trees while decision trees where the target variable can take continuous values are called regression trees. Decision trees are one of the most popular machine learning algorithms.

A Decision tree is constructed by asking a series of questions which are meant to break down the data set into smaller subsets. It uses binary splitting to determine the target values. If the decision tree is well constructed, every question cuts the number of options by half. In machine learning implementation of decision trees, each node in the tree splits the data into two groups using a cut-off value (Vanderplas 2017). Figure 43 shows some data, and figure 44 shows the different steps used by a decision tree algorithm to divide the data.



Figure 43: Random Data for Decision Tree<sup>1</sup>

As seen in figure 44, the decision tree iteratively splits the data along using a quantitative criterion, and at each level assign the label of the new region created according to the majority of data points it contains. After the first split (depth = 1), the upper part contains points that are similar, while the lower part contains points that are different. This means that in the second

---

<sup>1</sup> Vanderplas 2017.

split (depth = 2), the upper part does not need to be subdivided, while the lower part is subdivided.

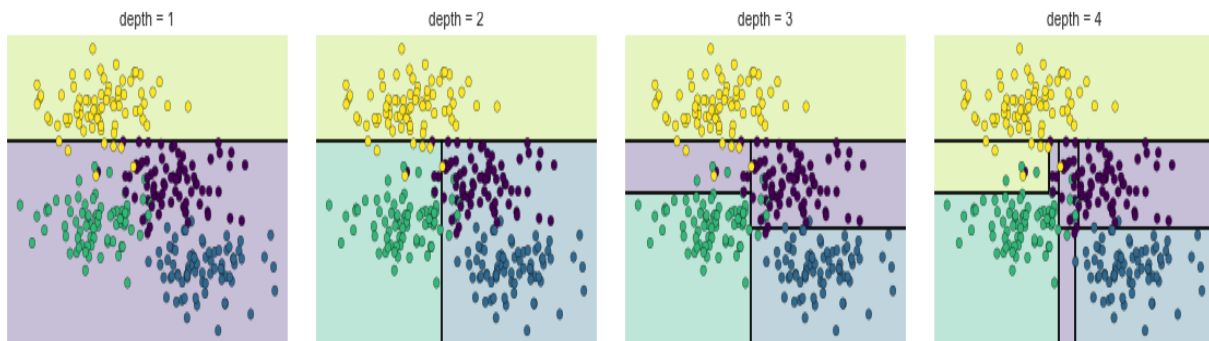


Figure 44: Different Steps of Data Splitting Using a Decision Tree<sup>1</sup>

Decision tree regression is similar to decision tree classification. The main difference is that it predicts a value in each node instead of predicting a class (Géron 2019). Figure 45 shows an example of a decision tree regressor.

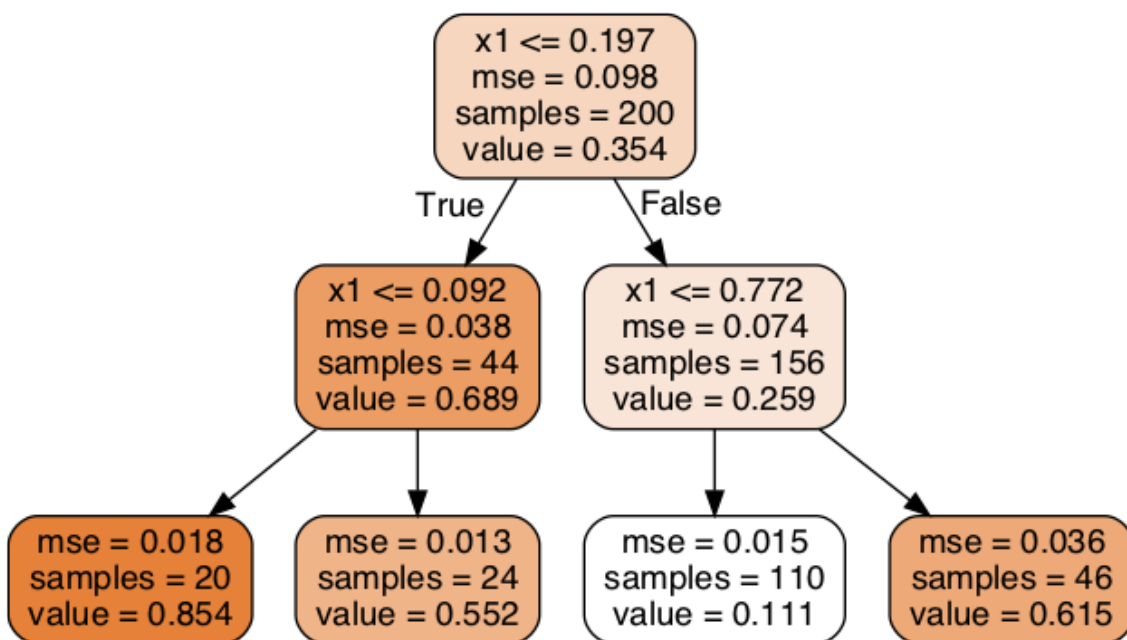


Figure 45: Decision Tree Regression<sup>2</sup>

Using the figure above, it is possible to determine the value of a new instance using the value of  $x_1$ . For example, if  $x_1=0.002$ , the predicted value is 0.854. This prediction is the average value for the 20 samples that are in this node, with a resulting MSE of 0.018.

<sup>1</sup> Vanderplas 2017.

<sup>2</sup> Géron 2019.



A high number of levels can result in an overfitted decision tree. Overfitting means that details of the particular data are used to split the data rather than the overall properties of the distributions they are drawn from. Figure 46 shows the same data set used in the figures above using two different decision tree models.

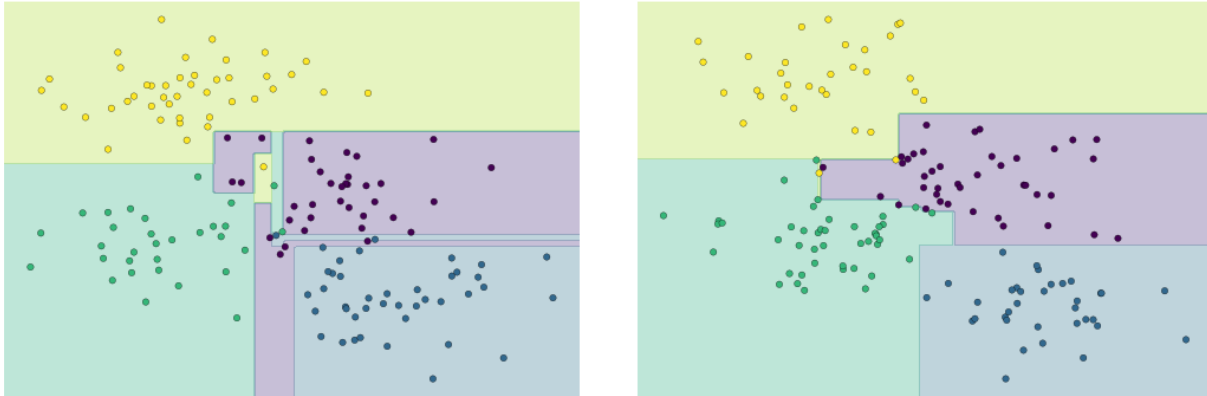


Figure 46: Overfitting in Decision Trees<sup>1</sup>

The figure above shows that in some areas, the two trees produce similar results (e.g., in the four corners), while in other areas, the two trees give very different classifications (e.g., in the regions between any two clusters). Both trees can give useful information, so using information from many decision trees at the same generally yields better results than from a single decision tree. Random forests are a machine learning algorithm that uses a number of decision trees to give better results.

The decision tree regression model performance can be misleading because of the overfitting problem. For this reason, it is simply used to be compared it with the random forest regression model. Table 17 shows the performance of the decision tree regression model.

Table 17: Decision Tree Regression Model Performance

<b>Model Performance</b>	<b>Oil Production</b>	<b>Gas Production</b>
Training Data Set Mean Value	49613 Bbl	245236 Mcf
Training Data Set Standard Deviation	43107 Bbl	211168 Mcf
MAE Test Set	3797 Bbl	113435 Mcf
RMSE Test Set	4345 Bbl	129140 Mcf
r2 Test Set	0.97	0.03

The performance of the decision tree looks very promising at first glance for the cumulative oil production. However, as discussed earlier, decision trees tend to overfit the data. Therefore, the results cannot be trusted, and it is better to use the results of a random forest to avoid the

---

<sup>1</sup> Vanderplas 2017.

overfitting problem. The performance the model when predicting the cumulative gas production of the wells in the test set is not good.

## 5.4 Random Forest Regression

To reduce the effect of overfitting, it is possible to combine a number of decision trees. It makes use of an ensemble of parallel estimators in a method called bagging. Each one of these estimators overfits the data, but the results can be averaged to find better results. A random forest is an ensemble of randomized decision trees (Vanderplas 2017). Figure 47 shows the result of a classification task when using a single decision tree (left) and when using a bagging of 500 decision trees (right).

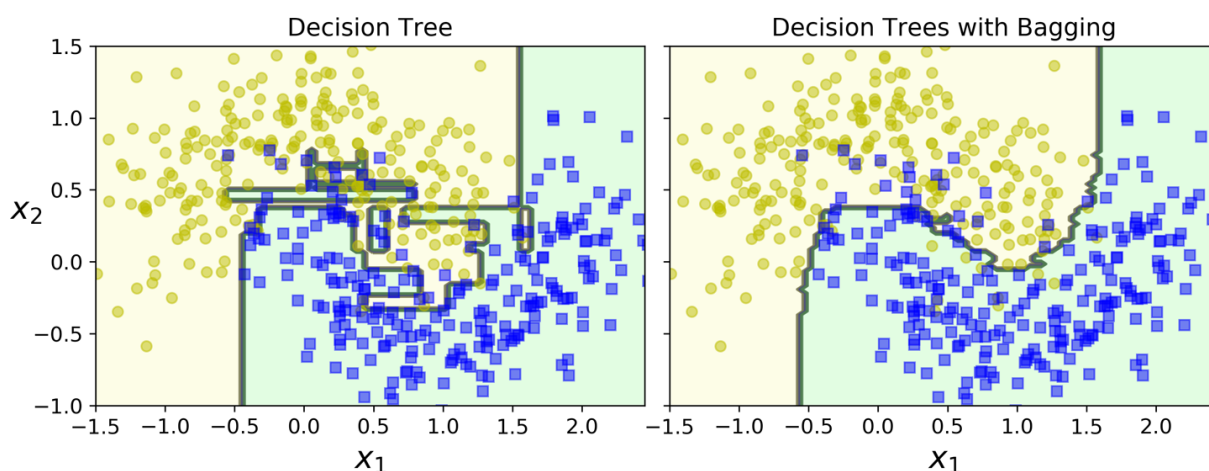


Figure 47: Single Decision Tree (Left) versus a bagging of 500 Decision Trees (Right)<sup>1</sup>

As seen in the figure above, the result of the 500 decision trees with bagging presents a smoother dividing line than the single decision trees. This result is an indication of a less overfitted model, which is one of the main reasons behind the use of random forests. Random forests are generally very fast, since they are based on simple decision trees (Vanderplas 2017).

The random forest model is controlled by a number of hyperparameters. These hyperparameters can change the performance of the model significantly. To determine the best hyperparameters for the random forest regression model, it is possible to use randomized search. Randomized search evaluates a given number of random combinations by selecting a random value for each hyperparameter at every iteration (Géron 2019). In addition to randomized search, it is possible to use cross validation to determine the best hyperparameters for the model. The cross-validation score is helpful to inspect the performance of the model before using it on the testing set. In this case, 5-fold cross validation was used. Two different models were trained, the first to predict the cumulative oil production and the second to predict the cumulative gas production. Table 18 shows the results of the

<sup>1</sup> Géron 2019.

random forest regression model. It shows that the random forest regression model did not perform very well when predicting the oil production. However, the performance of the model when predicting the cumulative gas production is very good. The  $r^2$  score is 0.84, which is excellent. The MAE and RMSE are also very good. This means that this model will be used to predict the gas production of the target wells.

Table 18: Random Forest Regression Model Performance

Model Performance	Oil	Gas
Training Data Set Mean Value	49613 Bbl	245236 Mcf
Training Data Set Standard	43107 Bbl	211168 Mcf
RMSE Cross Validation	33766 Bbl	145273 Mcf
MAE Test Set	17221 Bbl	52071 Mcf
RMSE Test Set	24183 Bbl	52684 Mcf
$r^2$ Test Set	0.18	0.84

## 5.5 Support Vector Machine Regression

A Support Vector Machine (SVM) is a versatile machine learning model. It can be used to perform classification, regression, and outlier detection. The idea behind SVMs is to create a separating margin between classes, and to make it as wide as possible. Figure 48 presents a random data set with two classes, with different possible lines separating the two classes

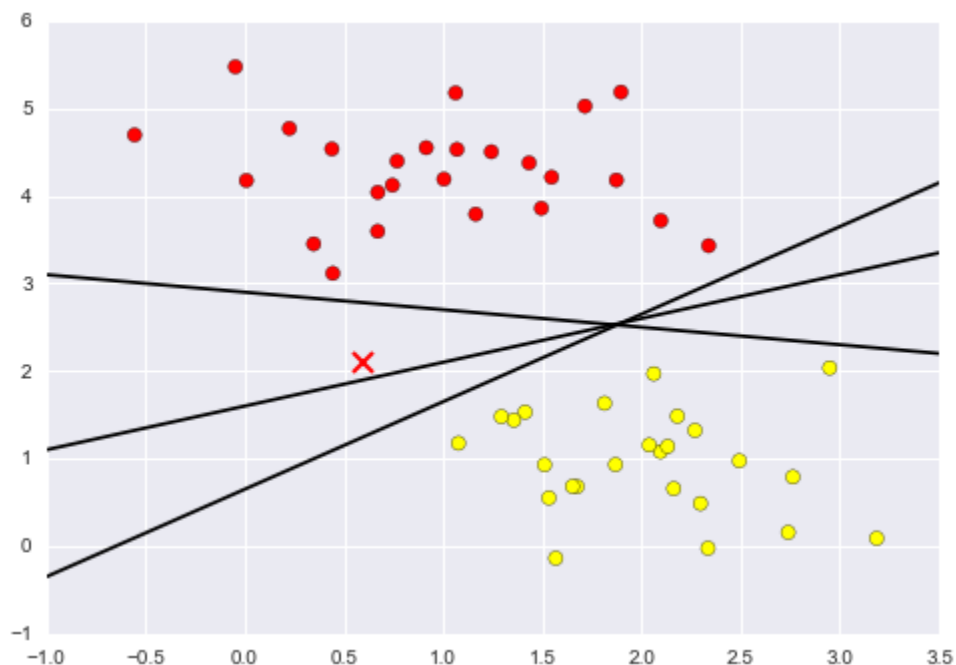


Figure 48: Random Data for Support Vector Machines<sup>1</sup>

<sup>1</sup> Vanderplas 2017.

Figure 48 shows that there are many possible lines that can be drawn to divide the two classes. A new data point, marked with the X, is classified in different classes depending on the line chosen to divide the two classes. The main idea behind support machine vectors is to draw a margin with a certain width, instead of a line, to divide the two classes. This is shown in figure 49.

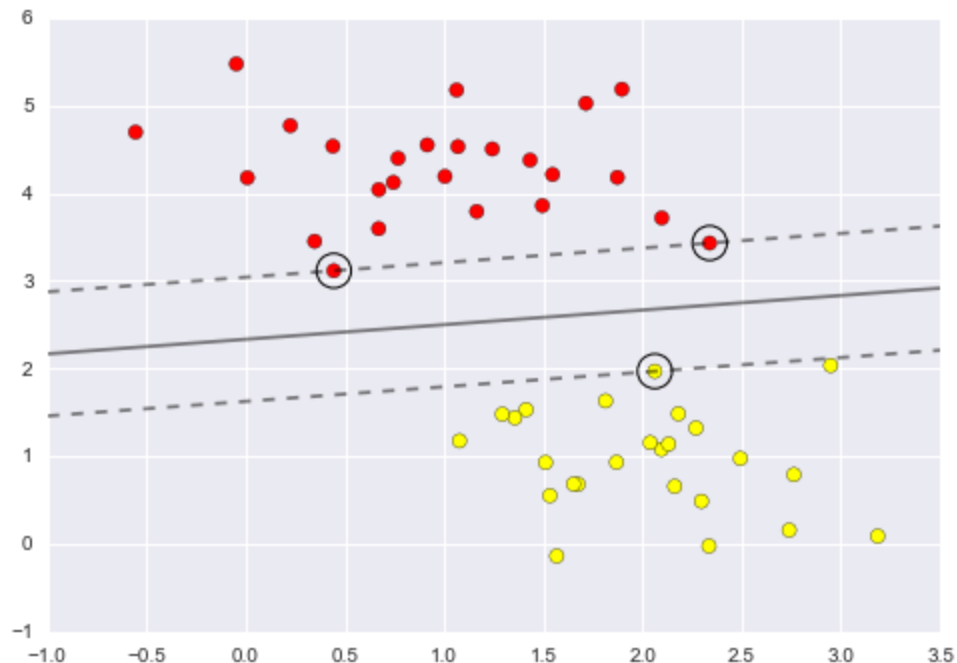


Figure 49: Support Vector Machine with Margins and Support Vectors<sup>1</sup>

As shown in the figure above, the objective of the SVM is to draw the line that maximizes the margin between the two classes. This margin is drawn with the dashed lines. The points that touch the margin are circled in black. They are called support vectors. SVMs are only affected by points near the margin.

SVM regression is based on the same principle as SVM classification. The difference is that the objective is to fit the highest number of instances on the street while minimizing instances outside the streets. Figure 50 presents an example of SVM regression. It shows the impact of the width of the margin in SVM regression. The width is controlled by  $\epsilon$ , and the value of  $\epsilon$  changes the regression result.

As with the random forest model, the support vector machine is also controlled by a number of hyperparameters. To determine the best hyperparameters for the support vector machine regression model, randomized search was also used, as well as 5-fold cross-validation, to determine the best hyperparameters for the model.

---

<sup>1</sup> Vanderplas 2017.

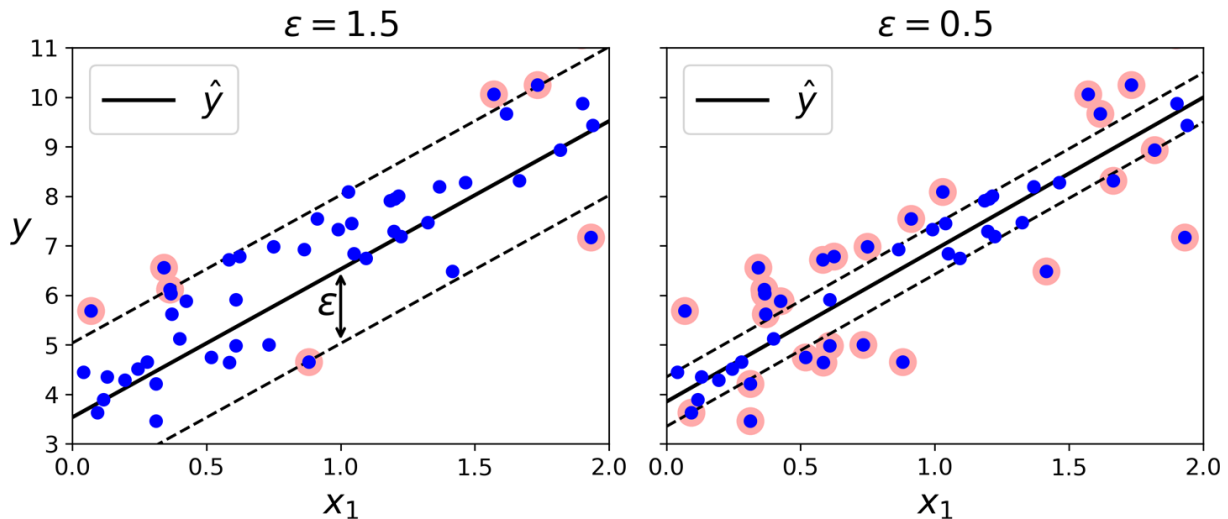


Figure 50: SVM Regression with Different Margin Width<sup>1</sup>

Two different models were trained, the first to predict the cumulative oil production and the second to predict the cumulative gas production. Table 19 presents the results of the SVM regression model. It shows that the model performed very well in predicting the oil production, with a MAE of 12931 barrels and a RMSE of 15000 barrels. The r2 score of 0.68 is very good. On the other hand, the gas production prediction was poor.

Table 19: Support Vector Machine Regression Model Performance

Model Performance	Oil Production	Gas Production
Training Data Set Mean Value	49613 Bbl	245236 Mcf
Training Data Set Standard Deviation	43107 Bbl	211168 Mcf
RMSE Cross Validation	36449 Bbl	160876 Mcf
MAE Test Set	12931 Bbl	123730 Mcf
RMSE Test Set	15000 Bbl	133594 Mcf
r2 Test Set	0.68	-0.04

## 5.6 Feature Importance

Determining the most important features can be done once the models have been trained. Knowing which features influence the production the most is important for future projects or if a change to the hydraulic fracturing plan of one of the wells is considered.

Sequential Feature Selector adds (forward selection) or removes (backward selection) features to form a feature subset in a greedy fashion. At each stage, the best feature to add or remove is chosen based on the cross-validation score of an estimator.

<sup>1</sup> Géron 2019.

This type of feature selection belongs to the family of greedy search algorithms which are used to reduce an initial  $d$ -dimensional feature space to a  $n$ -dimensional feature subspace where  $k$  is predefined and  $n < d$ . The objective is to select a subset of the features that are most relevant to the problem, which results in optimal computation efficiency while achieving reduced generalization error by filtering out irrelevant features (that act as a noise) (Ajitesh Kumar 2020).

Sequential feature selection is a greedy procedure where, at each iteration, the best new feature is selected based a cross-validation score. The selection is done using  $k$ -Fold Cross-Validation.

The aim of sequential forward selection is to search for the most important  $k$  features from the whole set of features. In this thesis, the 10 most important features are selected ( $n=10$ ). Figure 51 shows the steps in sequential forward selection (SFS):

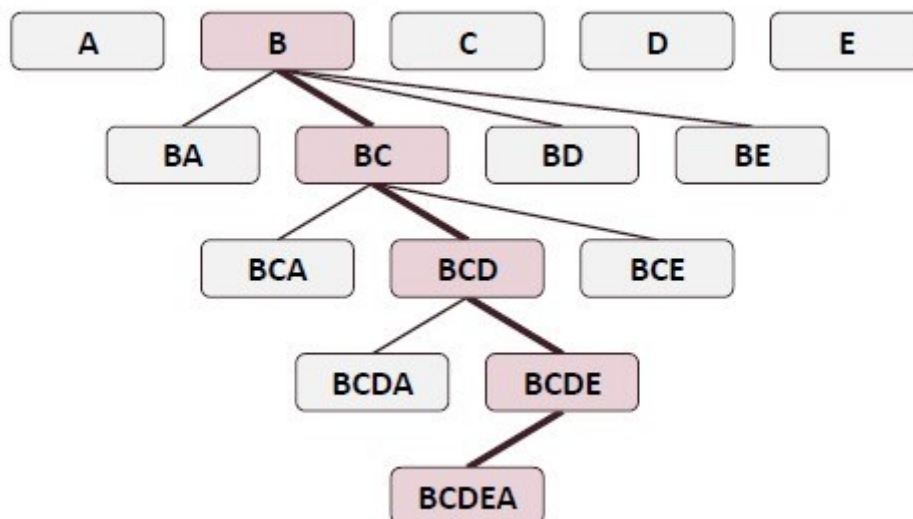


Figure 51: Sequential Forward Selection Steps<sup>1</sup>

As shown in the figure above, sequential forward selection is conducted using the following steps:

1. The first step is to determine the best feature based on a certain performance measure. In this thesis Root Mean Squared Error (RMSE) is used.
2. Once the best feature is selected (B), the next step is to form pairs using this best feature and the remaining features (A,C,D,E).
3. The best pair (BC) is selected and used to form triplets using this pair and the remaining features (A,D,E).

---

<sup>1</sup> Rudolf K. Fruhwirth 2018.

4. This procedure continues until the predefined number of features is reached.

Sequential backward elimination (SBE) is closely related. In this case, the procedure is started with the whole set of features, and one by one the least important features are removed until the desired number of features is reached.

In the case of the random forest model, it is possible to obtain the most important features considered by the model. It is therefore possible to compare what features are considered important by the SVM regression model using SFS and SBE, the features considered most important by the random forest (RF) model and the features with the highest correlation coefficient. Table 20 shows the top 10 most important features when predicting the cumulative oil production.

Table 20: Top 10 Most Important Feature in Cumulative Oil Production Prediction

<b>SVM SFS</b>	<b>SVM SBE</b>	<b>RF Feature Importance</b>	<b>Correlation Coefficient</b>
TOTAL_DAYS	TOTAL_DAYS	SLICKWATER	NET_PROD_DAYS
MESH 40/70	MESH 100	UPPER_PERF	TOTAL_DAYS
BROWN SAND	MESH 40/70	NET_PROD_DAYS	TOTAL_STAGES
VOLUME_PUMPED	PROPPANT_MASS	PROPPANT_MASS	PROPPANT_MASS
HYBRID	WHITE SAND	WELL_LATITUDE	AVERAGE_STP
SLICKWATER	HYBRID	WELL_LONGITUDE	LOWER_PERF
FRACTURE_GRADIENT	SLICKWATER	TOTAL_DAYS	TOP_DEPTH
WELL_LONGITUDE	WELL_LONGITUDE	TVD_DEPTH	MIN_STP
MIN_STP	WELL_HORZ_LENGTH	LOWER_PERF	WELL_LATITUDE
WELL_HORZ_LENGTH	NET_PROD_DAYS	MIN_STP	DAY_NUMBER

The correlation coefficient only considers the numerical features. However, the SFS, SBE and the random forest consider both numerical and categorical features. Table 20 shows that the only feature considered one of the top 10 important features by all methods is the total number of days ("TOTAL\_DAYS"). This proves that this feature, which did not originally exist in the initial data set, is very important. Fracturing fluid "Slickwater" is the most important categorical feature since it is present in the top 10 features of three calculation methods. The fracturing fluid "Hybrid" is important for the SVM regression model. Some numerical features are present in three of the four lists: "WELL\_LONGITUDE", "MIN\_STP", "PROPPANT\_MASS\_USED" and "NET\_PROD\_DAYS".

The top 10 most important features when predicting the cumulative gas production are different than for the cumulative oil production, as shown in Table 21.

Table 21: Most Important Feature in Cumulative Gas Production Prediction

<b>SVM SFS</b>	<b>SVM SBE</b>	<b>RF Feature Importance</b>	<b>Correlation Coefficient</b>
MESH 20/40	TOTAL_DAYS	WELL_LATITUDE	TOTAL_DAYS
MESH 30/50	MESH 20/40	TOTAL_DAYS	PROPPANT_MASS
PROPPANT_MASS	BROWN SAND	UPPER_PERF	NET_PROD_DAYS
BROWN SAND	WHITE SAND	NET_PROD_DAYS	TOTAL_STAGES
WHITE SAND	HYBRID	TOP_DEPTH	WELL_LONGITUDE
HYBRID	WATER	TOTAL_STAGES	UPPER_PERF
SLICKWATER	FRACTURE_GRADIENT	WELL_LONGITUDE	WELL_LATITUDE
WELL_LATITUDE	TVD_DEPTH	LOWER_PERF	DAY_NUMBER
UPPER_PERF	WELL_LATITUDE	PROPPANT_MASS	MIN_STP
NET_PROD_DAYS	LOWER_PERF	HYBRID	TVD_DEPTH

Both sequential feature selection methods (SFS and SBE) for the SVM regression model considered the categorical variables as the most important. Some categorical variables are present for both methods, like “Brown sand”, “White sand”, while “Hybrid” is present in both and also in the RF feature importance list. The most important numerical attribute is the well latitude “WELL\_LATITUDE” since it is present in all four lists. Other numerical attributes are present in three of the four lists: “UPPER\_PERF”, “NET\_PROD\_DAYS”, “PROPPANT\_MASS”, and “TOTAL\_DAYS”.

## 5.7 Final Results

Since all models have been trained and tested on the testing set, it is possible to determine the model that will be used to predict the oil and gas production of the 7 wells of the target set, which is the main objective of this thesis. The final comparison between the machine learning regression models used is shown in table. The model considered best for the cumulative oil production prediction and the model considered best for the cumulative gas production prediction will be used for the final predictions of the 7 wells of target data set. Table 22 shows the performance of different machine learning models in the prediction of the cumulative oil production of the 2 wells of the test set.

Table 22: Cumulative Oil Production Prediction Performance of the Different Models Trained

<b>Performance</b>	<b>Linear Regression</b>	<b>Decision Tree</b>	<b>Random Forest</b>	<b>SVM Reg</b>
<b>MAE</b>	30537 Bbl	3797 Bbl	17221 Bbl	<b>12931 Bbl</b>
<b>RMSE</b>	34812 Bbl	4345 Bbl	24183 Bbl	<b>15000 Bbl</b>
<b>R2</b>	- 0.70	0.97	0.18	<b>0.68</b>



While the best performance was obtained by the decision tree regression model, the result seems to be overfitted to the data used in training. This is also proved by the fact that the random forest regression model performed worse than the decision tree model, which is unusual. For this reason, the chosen model to determine the cumulative oil production of the 7 wells in the target data set is the support vector machine regression model. The final predictions of the cumulative oil production of the 7 wells of the target data set are present in figure 52 :

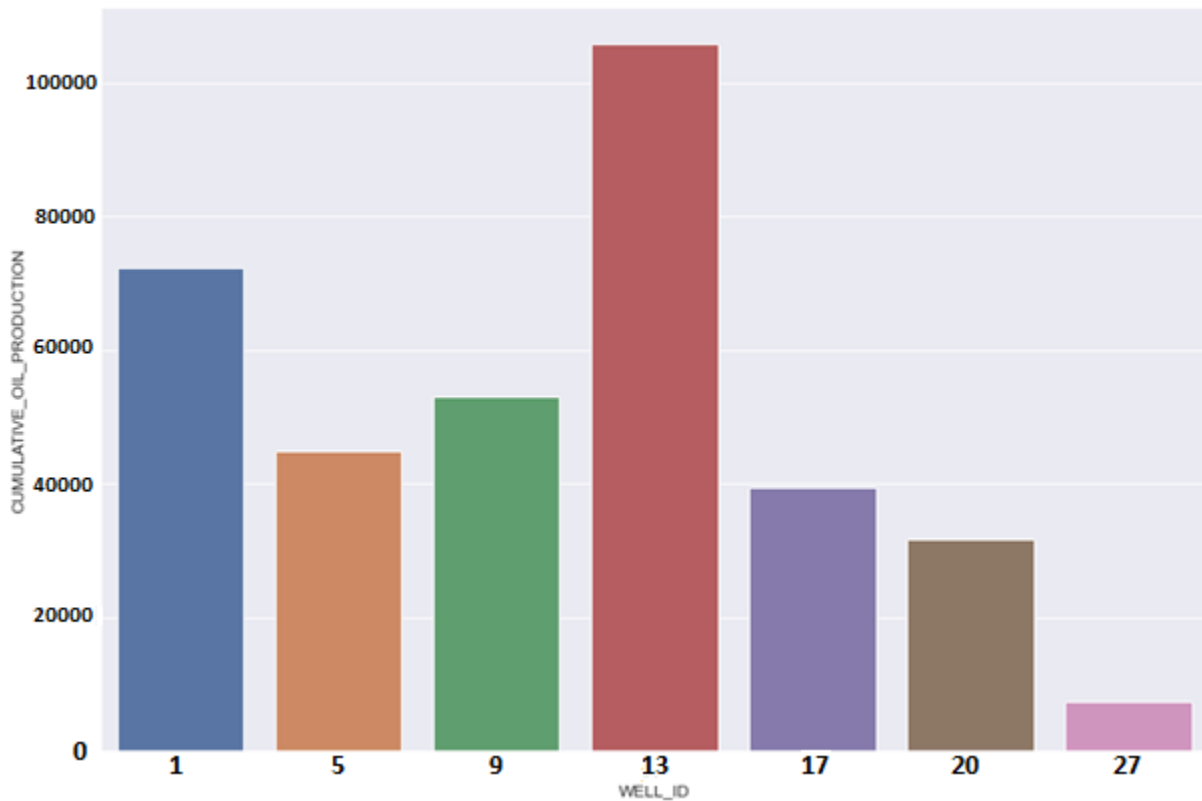


Figure 52: Final Cumulative Oil Production Prediction of the 7 Wells of the Target Data set

Figure 52 shows that most wells in the target data set are predicted to have a cumulative oil production between 35000 and 75000 barrels. The only two exception are well 13, which is predicted to be a very good producer reaching more than 100000 barrels, and well 27 which is predicted to be a bad oil producer.

The cumulative gas production predictions were calculated with different models that were optimized to predict the gas production. Table 23 shows the performance of different machine learning models in the prediction of the cumulative gas production of the 2 wells of the test set.

Table 23: Cumulative Gas Production Prediction Performance of the Different Models Trained

Performance	Linear Regression	Decision Tree	Random Forest	SVM Reg
<b>MAE</b>	121010 Mcf	113435 Mcf	<b>52071 Mcf</b>	123730 Mcf
<b>RMSE</b>	148340 Mcf	129140 Mcf	<b>52684 Mcf</b>	133594 Mcf
<b>R2</b>	- 0.28	0.03	<b>0.84</b>	-0.04

In the case of the cumulative gas production, the only model that presents good results is the random forest regression model. The  $r^2$  score of 0.84 is excellent, and the MAE and RMSE are also very good. This model will therefore be used to predict the cumulative gas production of the 7 wells of the target data set. The results are presented in figure 53.

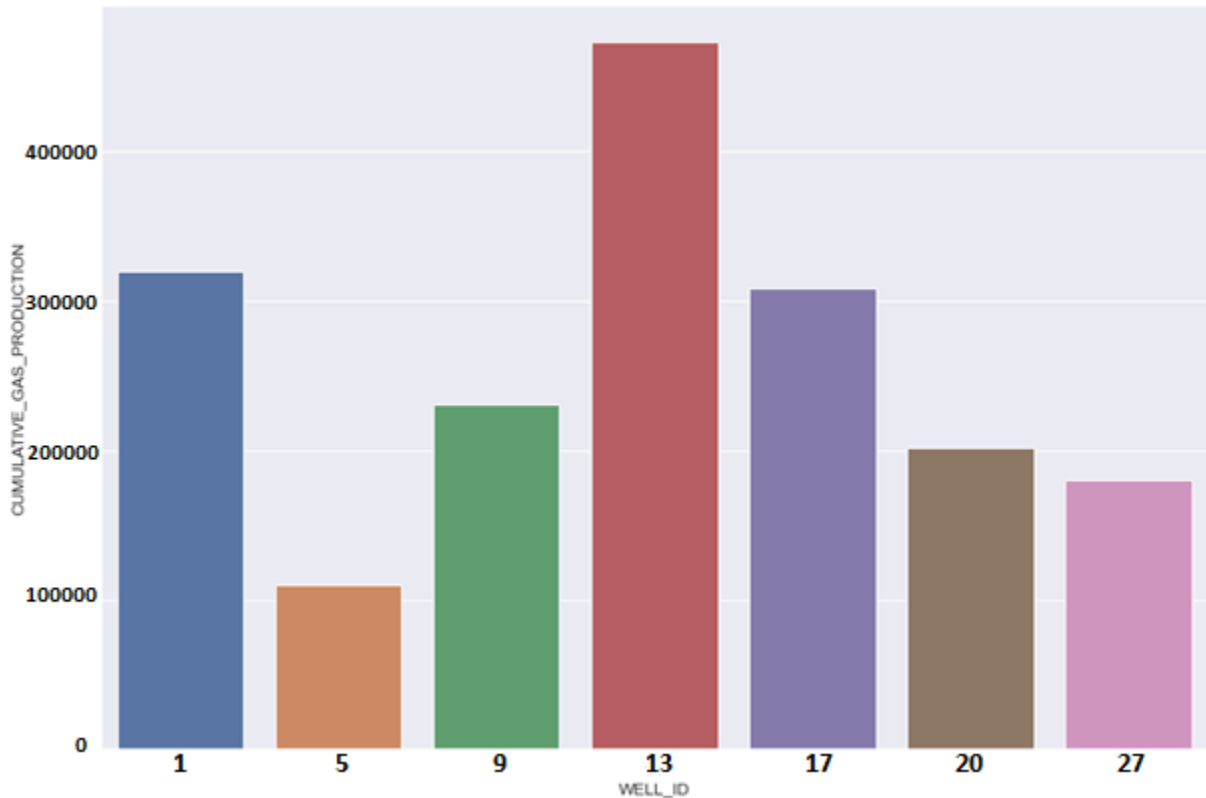


Figure 53: Final Cumulative Gas Production Prediction of the 7 Wells of the Target Data set

Figure 53 shows that most wells in the target data set are predicted to have a cumulative gas production between 320000 and 170000 Mcf. The only two exception are well 13, which is predicted to be both a very good oil producer and a very good gas producer, reaching more than 450000 Mcf, and well 5 which is predicted to produce around 100000 Mcf.

## 6 Conclusion

Machine learning has a big future in the oil and gas industry. Machine learning algorithms can be used to predict the performance of wells if enough information about similar or adjacent wells is provided. In this thesis, the cumulative oil and gas production of a number of multi-fractured horizontal wells is predicted using data from wells located in the same field.

Both data sets provided, the training data set and the target data set, contained a lot of redundant and wrong data. While the number of rows was relatively small for a machine learning problem, the number of features was important and proved to be sufficient to obtain good and logical predictions.

The biggest part of this thesis project, as with most data science and machine learning problems, was cleaning the data and preparing it for the machine learning models. A number of columns were deleted, while other columns had to be examined to extract information that turned out to be very important for the predictions. Examples of features that were not present in the correct form in the initial data set are fracturing fluid information, proppant mesh size information and proppant type information, in addition to information about the hydraulic fracturing stages.

Once the data was cleaned and prepared, exploratory data analysis was conducted. This part of the thesis was important to examine the impact of various features on the cumulative oil production and cumulative gas production. New attributes were created, which were some of the most important features used by the different machine learning models to make good predictions. The data also needed to be prepared for the different models since it was not useable in its initial form. One-hot encoding was used for categorical data, while standardization was used for numerical data.

The final part of the object was the creation of different regression models to determine the cumulative oil production and cumulative gas production of the target wells. Four models were created, which were linear regression, decision tree regression, random forest regression and support vector machine regression. Cross validation was used with the more sophisticated models, and the data was split into a learning set and a testing set to measure the performance of the different models. The most important features for different models were determined using sequential feature selection, among other techniques. These features are important to determine what can be changed in the hydraulic fracturing job in order to improve the predicted production of the wells.

The support vector machine regression model was the most accurate in predicting the cumulative oil production, while the random forest regression model was the most accurate in predicting the cumulative gas production. These models were then used to predict the cumulative oil and gas production of the target wells. The results were logical and helpful in determining the best wells for oil production and gas production.

This work was conducted using machine learning algorithms. The future direction can be to obtain more data from other adjacent wells. This data can help improve the performance and the robustness of the models used. More data can also allow for the training of neural networks,

which generally required large amounts of data. It is also possible to compare the results of the machine learning models with the predictions of some oil and gas software typically used to predict oil and gas prediction.

## 7 Publication bibliography

- 1- Ahmed, Usman; Meehan, D. Nathan (2016): Unconventional oil and gas resources. Exploitation and development / edited by Usman Ahmed and D. Nathan Meehan. 1st. Boca Raton: CRC Press (Emerging trends and technologies in petroleum engineering).
- 2- Ajitesh Kumar (2020): Sequential Forward Selection - Python Example - Data Analytics. Available online at <https://vitalflux.com/sequential-forward-selection-python-example/>, updated on 7/30/2020, checked on 4/20/2021.
- 3- Aminzadeh, Fred (Ed.) (2020): Hydraulic fracturing and well stimulation. Hoboken NJ: Wiley-Scrivener.
- 4- Ashayeri, Cyrus; Ershaghi, Iraj (2015): OPEC and Unconventional Resources. In : SPE Annual Technical Conference and Exhibition. SPE Annual Technical Conference and Exhibition. Houston, Texas, USA, 2015-09-28: Society of Petroleum Engineers.
- 5- Belyadi, Hoss; Fathi, Ebrahim; Belyadi, Fatemeh (Eds.) (2016a): Hydraulic fracturing in unconventional reservoirs. Theories, operations, and economic analysis / Hoss Belyadi, Ebrahim Fathi, Fatemeh Belyadi. Amsterdam: Gulf Professional Publishing.
- 6- Belyadi, Hoss; Fathi, Ebrahim; Belyadi, Fatemeh (2016b): Proppant Characteristics and Application Design. In Hoss Belyadi, Ebrahim Fathi, Fatemeh Belyadi (Eds.): Hydraulic fracturing in unconventional reservoirs. Theories, operations, and economic analysis / Hoss Belyadi, Ebrahim Fathi, Fatemeh Belyadi. Amsterdam: Gulf Professional Publishing, pp. 73–96.
- 7- Brownlee, Jason (2018): A Gentle Introduction to k-fold Cross-Validation. In *Machine Learning Mastery*, 5/22/2018. Available online at <https://machinelearningmastery.com/k-fold-cross-validation/>, checked on 4/20/2021.
- 8- Economides, Michael J. (2013): Petroleum production systems. 2nd ed. Upper Saddle River, NJ: Prentice Hall.
- 9- Géron, Aurélien (2019): Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. Concepts, tools, and techniques to build intelligent systems / Aurélien Géron. Second Edition. Sebastopol, CA: O'Reilly.
- 10- Holditch, Stephen A. (2013): Unconventional oil and gas resource development – Let's do it right. In *Journal of Unconventional Oil and Gas Resources* 1-2, pp. 2–8. DOI: 10.1016/j.juogr.2013.05.001.
- 11- Hydraulic Fracturing: An Indiana Assessment (2020). Available online at <https://igws.indiana.edu/OilGas/HydraulicFracturing>, updated on 6/6/2020, checked on 6/6/2020.
- 12- Lake, Larry W.; Fanchi, John R. (2006-2007): Petroleum engineering handbook. Richardson, TX: Society of Petroleum Engineers.

- 13- Ma, Y. Zee; Holditch, Stephen; Royer, Jean-Jacques (2015): Unconventional oil and gas resources handbook. Evaluation and development / Y. Zee Ma, Stephen Holditch, Jean-Jacques Royer. Amsterdam: Gulf Professional Publishing.
- 14- PetroWiki (2020): Fracturing fluids and additives - PetroWiki. Available online at [https://petrowiki.spe.org/Fracturing\\_fluids\\_and\\_additives](https://petrowiki.spe.org/Fracturing_fluids_and_additives), updated on 4/16/2020, checked on 2/19/2021.
- 15- Railroad Commission of Texas: Permian Basin Information. Available online at <https://www.rrc.texas.gov/oil-and-gas/major-oil-and-gas-formations/permian-basin/>, checked on 2/14/2021.
- 16- Rudolf K. Fruhwirth (2018): Production Data Analysis and Modelling. Montanuniversität Leoben, 2018.
- 17- Shale oil and shale gas resources are globally abundant - Today in Energy - U.S. Energy Information Administration (EIA) (5/30/2020). Available online at <https://www.eia.gov/todayinenergy/detail.php?id=11611>, updated on 5/30/2020, checked on 5/31/2020.
- 18- Smith, Michael Berry; Montgomery, Carl T. (2015): Hydraulic fracturing. Boca Raton [Florida]: CRC Press (Emerging trends and technologies in petroleum engineering).
- 19- Speight, James G. (2016): Handbook of hydraulic fracturing. Hoboken New Jersey: Wiley. Available online at <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1202083>.
- 20- U.S. Energy Information Administration (5/31/2020): EIA's Annual Energy Outlook 2020 projects consumption growing more slowly than production - Today in Energy - U.S. Energy Information Administration (EIA). Available online at <https://www.eia.gov/todayinenergy/detail.php?id=42635>, updated on 5/31/2020, checked on 5/31/2020.
- 21- U.S. Energy Information Administration: Permian Basin. Available online at [https://www.eia.gov/maps/pdf/PermianBasin\\_Wolfcamp\\_EIARreport\\_Oct2018.pdf](https://www.eia.gov/maps/pdf/PermianBasin_Wolfcamp_EIARreport_Oct2018.pdf), checked on 2/14/2021.
- 22- U.S. Environmental Protection Agency (2016): Hydraulic Fracturing for Oil and Gas: Impacts from the Hydraulic Fracturing Water Cycle on Drinking Water Resources in the United States. Executive Summary. Office of Research and Development, Washington, DC. EPA/600/R-16/236ES.
- 23- Vanderplas, Jacob T. (2017): Python data science handbook. Essential tools for working with data / Jake VanderPlas. Beijing: O'Reilly.
- 24- Wikipedia (Ed.) (2021a): Correlation and dependence. Available online at [https://en.wikipedia.org/w/index.php?title=Correlation\\_and\\_dependence&oldid=1011719331](https://en.wikipedia.org/w/index.php?title=Correlation_and_dependence&oldid=1011719331), updated on 3/12/2021, checked on 3/17/2021.

25- Wikipedia (Ed.) (2021b): Cross-validation (statistics). Available online at [https://en.wikipedia.org/w/index.php?title=Cross-validation\\_\(statistics\)&oldid=1018725954](https://en.wikipedia.org/w/index.php?title=Cross-validation_(statistics)&oldid=1018725954), updated on 4/19/2021, checked on 4/20/2021.

## List of Tables

Table 1: Source Rock Evaluation Parameters .....	4
Table 2: Technically Recoverable Shale Oil and Shale Gas Unproved Resources in the Context of Total World Resources.....	5
Table 3: Fracturing Fluids Chemical Additives.....	10
Table 4: Proppant Types Summary .....	13
Table 5: Missing Values and Data Types of the Data Sets Used.....	23
Table 6: "PROPPANT_MESH_DESCRIPTION" and "PROPPANT_MESH_SIZE" contents in the Training Data Set .....	27
Table 7: "PROPPANT_MESH_DESCRIPTION" and "PROPPANT_MESH_SIZE" contents in the Target Data Set.....	27
Table 8: "PROPPANT_TYPE" and "PROPPANT_MESH_SIZE" contents in the Training Data Set after processing .....	28
Table 9: "PROPPANT_TYPE" and "PROPPANT_MESH_SIZE" contents in the Target Data Set after processing .....	28
Table 10: "FRAC_FLUID" contents for the training and target data sets.....	30
Table 11: Statistical Description of the Column "MIN_STP" in the Training Data Set.....	32
Table 12: Description of the Columns Containing Categorical Variables .....	40
Table 13: Example of One-Hot Encoding of the Column Proppant Mesh Size.....	44
Table 14: Correlation Coefficient between Numerical Variables and Oil and Gas Production .....	46
Table 15: Correlation Coefficient between Numerical Variables and Oil and Gas Production with New Attributes.....	51
Table 16: Performance of the Linear Regression Model.....	55
Table 17: Decision Tree Regression Model Performance.....	58
Table 18: Random Forest Regression Model Performance .....	60
Table 19: Support Vector Machine Regression Model Performance.....	62
Table 20: Top 10 Most Important Feature in Cumulative Oil Production Prediction .....	64
Table 21: Most Important Feature in Cumulative Gas Production Prediction .....	65
Table 22: Cumulative Oil Production Prediction Performance of the Different Models Trained .....	65
Table 23: Cumulative Gas Production Prediction Performance of the Different Models Trained.....	66





# List of Figures

- Figure 1: Resource Triangle ..... 3
- Figure 2: U.S. Crude Oil and Dry Natural Gas Production through 2050 ..... 4
- Figure 3: Production and Reserves Enhancement from HF for Low Permeability Reservoirs 6
- Figure 4: Schematic of a Typical Hydraulic Fracturing Operation ..... 7
- Figure 5: Overall Composition of a Typical Fracturing Fluid.....11
- Figure 6: Estimation of Roundness and Sphericity of a Grain .....12
- Figure 7: Difference between Created and Propped Fracture Dimensions .....14
- Figure 8: Effect of Mesh Size on Fractures.....14
- Figure 9: Location of the Permian Basin .....16
- Figure 10: Average Daily Oil Production in the Texas Permian Basin through November 2020  
.....17
- Figure 11: Location of the Wells in Texas .....18
- Figure 12: Location of the Training Wells and the Target Wells .....19
- Figure 13: Plot of True Vertical Depth Column versus TVD Depth Column of the Training Set  
.....25
- Figure 14: Plot of True Vertical Depth Column versus TVD Depth Column of the Training Set  
without outliers .....25
- Figure 15: Plot of True Vertical Depth Column versus TVD Depth Column of the Target Set  
.....26
- Figure 16: Proppant Type Counts of Both Data Sets after Processing.....29
- Figure 17: Proppant Mesh Size Counts in Both Data Sets after Processing .....29
- Figure 18: Fracturing Fluid Counts in both Data Sets after Processing.....31
- Figure 19: Average Number of Stages per Day for the Wells of the Training Set.....33
- Figure 20: Number of Stages for each Well in the Training Data Set.....34
- Figure 21: Number of Days for each Well in the Training Data Set.....34
- Figure 22: Cumulative Oil Production of the Training Wells in Barrels .....36
- Figure 23: Cumulative Gas Production of the Training Wells in Mcf .....36
- Figure 24: Net Production Day of the Training Wells in days .....37
- Figure 25: Oil Production per Day per Well in bbl/day .....38
- Figure 26: Gas Production per Day per Well in Mcf/day .....38

Figure 27: Effect of the Well Location on the Oil and Gas Production.....	39
Figure 28: Impact of the Proppant Mesh Size on Oil Production.....	41
Figure 29: Impact of the Proppant Mesh Size on Gas Production.....	41
Figure 30: Impact of the Proppant Type on Oil Production .....	42
Figure 31: Impact of the Proppant Type on Gas Production .....	42
Figure 32: Impact of the Fracturing Fluid Type on Oil Production .....	43
Figure 33: Impact of the Fracturing Fluid on Gas Production.....	43
Figure 34: Standard correlation coefficient of various data sets.....	45
Figure 35: Cumulative Oil Production (bbl) vs Average Pressure (psi).....	47
Figure 36: Cumulative Oil Production (bbl) vs Proppant Mass Used (cwt) .....	48
Figure 37: Cumulative Gas Production (Mcf) vs Proppant Mass Used (cwt).....	48
Figure 38: Cumulative Oil Production (bbl) vs Net Production Days (days).....	49
Figure 39: Cumulative Gas Production (Mcf) vs Net Production Days (days) .....	49
Figure 40: k-fold Cross Validation Procedure .....	53
Figure 41: Random Data for Linear Regression .....	54
Figure 42: Data with Fitted Linear Regression Model .....	54
Figure 43: Random Data for Decision Tree .....	56
Figure 44: Different Steps of Data Splitting Using a Decision Tree .....	57
Figure 45: Decision Tree Regression .....	57
Figure 46: Overfitting in Decision Trees.....	58
Figure 47: Single Decision Tree (Left) versus a bagging of 500 Decision Trees (Right).....	59
Figure 48: Random Data for Support Vector Machines.....	60
Figure 49: Support Vector Machine with Margins and Support Vectors .....	61
Figure 50: SVM Regression with Different Margin Width .....	62
Figure 51: Sequential Forward Selection Steps.....	63
Figure 52: Final Cumulative Oil Production Prediction of the 7 Wells of the Target Data set	66
Figure 53: Final Cumulative Gas Production Prediction of the 7 Wells of the Target Data set .....	67

## Abbreviations

TOC	Total Organic Carbon
%Ro	Vitrine Reflectance
EIA	U.S. Energy Information Administration
TRR	Technically Recoverable Resources
Mcf	Thousand Cubic Feet
Psi	Pounds per square inch
Tcf	Trillion Cubic Feet
Bbl	Barrel
Cwt	Hundredweight
HF	Hydraulic Fracturing
Ft	Feet
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MSE	Mean Squared Error
R2	R-squared
SVM	Support Vector Machine
RF	Random Forest
SFS	Sequential Forward Selection
SBE	Sequential Backward Elimination

## **Nomenclature**

TOC Total Organic Carbon

%Ro Vitrine Reflectance