

Quantitative Risk analysis using machine and deep learning forecasting algorithms, with applications for Oil & Gas Exploration and Production projects

**MASTER THESIS**

**Author** Luis Fernando Pérez Armas

**Contact** [lui.perezarmas@ieseg.fr](mailto:lui.perezarmas@ieseg.fr)

**Supervisor** Dr. Stefan CREEMERS

**School Year** 2017-2019

**Option** Operations Management

**MASTER THESIS**

Quantitative Risk analysis using machine and deep learning forecasting algorithms, with applications for Oil & Gas Exploration and Production projects



IESEG School of Management

3 Rue de la Digue

F-59000 Lille

*"L'IESEG n'entend donner aucune approbation ni improbation aux opinions émises dans les mémoires; ces dernières doivent être considérées comme propres à leurs auteurs"*

# Aknowledgments

Bla bla bla bla bla…

# Table of Contents

# List of Abbreviations

Bla bla bla bla bla…

# Summary

**Objective of the Thesis**

**Methodology**

**Findings**

**Value**

**1 Introduction**

**1.1 Problem Definition**

Assumptions are an inherent and unavoidable part of business and life, the sole idea that business are made via the interactions with markets and the fact that markets are time dependent and variable, makes impossible the idea of having a business without any unknowns, although this doesn’t mean that business are conducted in complete darkness and ruled merely by chance, a crucial role in any industry in any business is transforming this unknowns into something else. Assumptions are just interpretations of unknowns and it’s the responsibility of management at any given level, to correctly read these unknowns and transform them into valid information, information that can be used to create different course of actions with given uncertainties. This process is an essential part of what is known as risk management. Risk is defined as the effect of uncertainty on objectives (ISO 31000, 2009), and will be the main topic of this work.

There is an old saying, "Walking the walk is more difficult than talking the talk." or "Easier said than done." Both phrases just intend to express how difficult can be to transform an idea into a reality specially when ideas are complex and require the coordination of different resources. In business ideas are implemented in the form of projects. A project (Figure 1) is a temporary endeavor to achieve clearly defined goals and project management deals with the planning, organization, execution, monitoring (controlling) and closing of a project in order to attain the project’s objectives (Project Management Institute, 2013).

Due to the already blurry and dynamic nature of business, it is safe to assume that are also plagued by unknowns, and therefore risk is also critical aspect of project management. Traditionally at project level, project risk management has focused on scheduling efforts, Project scheduling is part of the planning phase of a project’s lifecycle, on this phase a schedule is developed which decides when to start and finish the activities in order to achieve the project’s goals. although, it is well known that the adequate planning, execution, and control of activities can severely minimize project risk, especially on environments of high uncertainty and resource constrains; certain risks, in the form of assumptions and constrains, have a less controllable nature, and therefore a quantitative determination of the level exposure becomes critical for the business continuity of projects. Normally the quantification of this assumptions is performed at the initiation phase of projects, it is part of different project documents such as the business case, benefits management plan and project charter, and it is a major concern of program and portfolio managerial levels. A program (Figure 1) is “a group of related projects managed in a coordinated way to obtain benefits and control not available from managing them individually. Programs may contain elements of work outside of the scope of the discrete projects in the program.” (Project Management Institute, 2013). A portfolio (Figure 1) is defined as “a collection of projects, programs and other work that is grouped together to facilitate the effective management of that work to meet strategic business objectives. The projects or programs of the portfolio may not necessarily be interdependent or directly related.” (Project Management Institute, 2013).

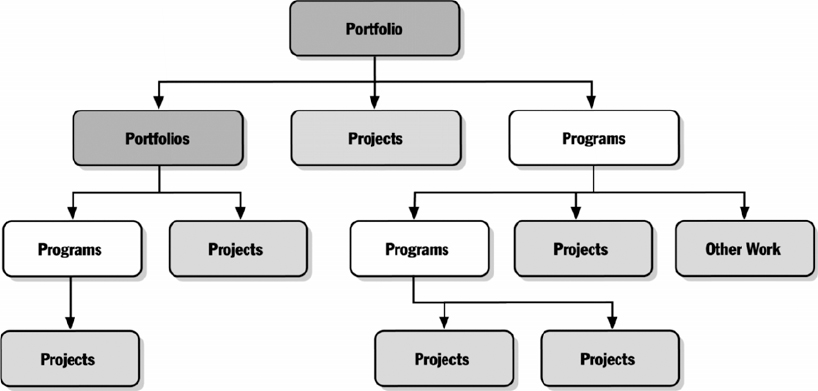


Figure 1 Project, Program and Portfolio relationships

The business world has numerous cases on which an invalid assumption at program or portfolio level has rendered projects completely unfeasible despite all efforts made on risk management at project and activity level.

The process on which assumptions are quantified is known as prediction, and when time series are involved the process is called forecasting. Forecasting is a statistical task in business, and it helps to inform decisions about the scheduling of production, transportation and personnel, and provides a guide to long-term strategic planning, forecasting is about predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events that might impact the forecasts. (Hyndman, R.J., & Athanasopoulos, G. 2018).

Despite its importance on business and projects, forecasting is usually done wrong and, on most cases, using only judgement as a basis. A large survey of 240 US corporations, revealed that only 11% reported using forecasting software. And of those who did use forecasting software, 60% indicated they routinely adjusted the forecasts based on their judgement(Sanders and Manrodt, 2003).

Although today is recognized that judgment forecasting is a valuable business tool and a good complement for statistical forecasts (Lawrence et al., 2006), it is important to be aware that Judgmental forecasts are subjective, and therefore do not come free of bias or limitations therefore it should be used as a starting point and not as a definitive prediction (Hyndman, R.J., & Athanasopoulos, G. 2018).

Thanks to several factors such as the continuous increase of computational power, massive internet connectivity, and easiness of data collection and sharing; recent years, have been characterized by the appearance and democratization of more powerful prediction and forecasting tools, that exist under the domain of machine and deep learning. Such tools are currently use on multiple domains such as image and speech recognition, fraud detection, stocks market prediction, and many other applications.

The following work intends to use some of the currently most popular machine and deep learning forecasting tools and compare the results obtained with basic naïve judgmental approaches. This work will focus on the forecast of two of the most important variables, used as part of the assumptions of Oil & Gas exploration and production (E&P) projects, programs and portfolios. These variables are the monthly oil price and active monthly drilling rigs, the study will only focus on US open data from 1973 to the end of 2017, and the data was provided by the international Agency of Energy.

**1.2 Course of investigation**

The current thesis is divided into numerous parts. The first section of this thesis will focus on the development of a theoretical and conceptual framework, known as “literature review” for the literature review we will first look and study a current outlook of the Oil & Gas industry, as well as the industry structure and business model, focusing specially on the Exploration and Productions upstream sector. Later, literature is reviewed concerning factors of influence that has traditionally affected the industry, specially oil prices. Next concepts related to project management specially prediction and forecasting in projects management will be reviewed, as well as judgmental forecasting concepts, tools and techniques. Literature concerning the general data science project pipeline is reviewed, as well as mathematical and statistical theory behind specific machine and deep learning algorithms that are going to be used on this work. Finally, the literature review section will end up with the review of previous studies concerning the forecast of future oil prices.

The development of a conceptual and theoretical framework allows us to understand the relevance of the topic, along with providing a framework of methodology, tests and metrics to use. The review of previous academic work adds value by providing insight on opportunities that can be exploited in order to improve previous works and therefore lead us on the creation of valid research questions for this thesis. As such, the literature review is a guide to determine the propositions of the current thesis and the methods to address them.

With a solid conceptual framework in place, the next step of this work is the development of the methodology section, this section will start with a fishbone analysis of the most influential explanatory variables for oil prices and rig count forecast, this analysis will be conducted based on previous studies, once variables are chosen, the next step of the methodology is related to the data collection and cleaning, so as data transformations and feature engineering. The next section of the methodology will describe different aspects that concerns a data science study, such as the chosen metrics for model performance, the train and test dataset split for performance evaluation and the comparative benchmark selection for the models.

With a defined methodology the next and third section of this work will be related to data analysis, this section will start with an exploratory data analysis of the collected data set, so as with a statistical analysis for feature confirmation of various questions derived from the previously conducted exploration. An exploratory and statistical analysis serves will contribute with the confirmation of variable importance, known as feature selection and will also help to optimize the process of model generation and optimization.

With the insights gained by the exploratory and feature selection process previously conducted the next section of this work is related to conduct statistical inference and on the generation of predictive models for monthly oil prices and monthly active US rig count, as well with the hyperparameters optimization process of both of these models, using only the training data. Once this is done, models will be evaluated on the test partition data set, these results will be compared against the “judgmental” naive baselines and results obtained on the performance of both models will be analyzed and discussed, as well as the managerial implications of these results. Finally, the the limits of the study are stated, and they will serve as a basis for a final discussion and conclusions of this work.

**2 Literature Review**

This part aims to introduce to the reader, a current outlook of the Oil & Gas exploration and production business, so as the industry structure, business model and most influential variables, it also summarizes concepts related to data science, machine and deep learning and the recent evolution and development of this area. This section ends with the review of literature related to the generation oil prices predictive models.

**2.1 2019 Outlook of the Oil & Gas industry**

The following section intends to give the reader an overall understanding of the current shape of the Oil & Gas industry, focusing on what had happened on recent years, and how past events have shaped the industry, and what to expect from the industry and the markets on the future.

**2.1.1 Understanding The 2014 Downturn**

If there is a constant that characterizes the Oil & Gas industry it is change, the industry has historically moved from cycles of high and low oil prices, mostly conditioned by transitions of over and undersupply. There are plenty of driving forces from diverse nature, that attempt to explain the behavior of the Oil & Gas markets, but despite its complexity and historical volatility, it still reacts and behaves as any other basic that follows fundamentals economic laws, this is a market on which an excess on oil supply will force prices down, and low prices will forces supply cuts, if this supply cuts are maintained for a enough time to drive the market into an state of undersupply then prices will raise, and producers will adapt to this raise by increasing production, until the point on which the market reaches an oversupply state, and the cycle will repeat (Miao et al., 2017). But let’s not trick ourselves and fall into the trap of thinking easily about the oil and gas market, on previous lines we were just explaining the industry reaction to a different market; the energy market, which is complex, always shifting and shifting fast. It reacts to factors such as the economic development of countries, new technologies, energy supply of other producers, geopolitical factors and even the life standards of human population.

On 2014 The Oil & Gas industry experienced one of its worst downturns in modern years, prices soared from average 100 USD down to a low of 26 USD, prices behave mainly responding to energy oversupply (Figure 2), but this time also reacting to profound changes in the energy market and the oil industry (iea, 2018).

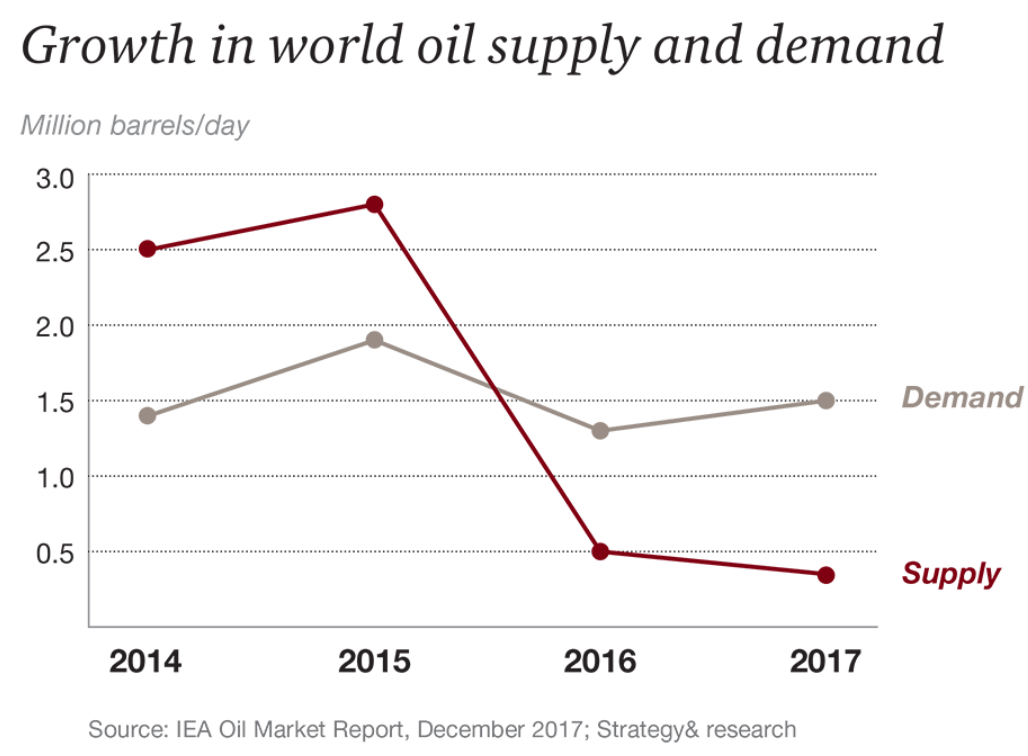


Figure 2 2014 Oil world supply and demand

Before 2008, the industry and market were driven by the Hubert’s Peak belief; being oil a natural but limited resource, it is safe to assume that eventually its flow will reach a maximum and after diminish, mainly due to the natural reservoir depletion, lack of new reservoirs to exploit, and higher energetic demands from the world market (Campbell and Laherrère, 1998). However, such panorama dramatically changed with the appearance of shale production technology.

From the end of the 19th century the oil and gas industry has been exploiting what are considered, conventional reservoirs. Oil is generated by the thermic rupture of organic matter known as kerogen, located on sedimentary non porous non permeable formations, through different physicochemical and geological process like diagenesis, catagenesis and metagenesis, the kerogen is transformed into oil and gas molecules, this transformation is characterized by an increase in temperatures and pressure, once the internal pressure of this fluids is sufficiently build up it will fracture the initial mother rock on which fluids were transformed and they will travel using permeable channels, into a sedimentary porous permeable formation on which they will get trapped once and if the fluids found a non-permeable formation ceiling, that retain them, this ceiling rock is known as a “trap” and the porous formation that contains the fluids is known as the reservoir (Guerriero et al., 2013). So normally exploration and production efforts are focused on finding hints for geological “traps” that might host under its roof an oil and gas reservoir, but on 2007 thanks to technological innovation and change in regulations, and idea came in, to disrupt the whole industry an energy markets, “What if, instead of looking for oil on conventional traps and traditional porous permeable reservoirs, we go directly to the rock mother, on which we know the fluids origin and we use hydraulics to fracture the rock therefore creating an artificial permeability that will allow us to produce from these formations?” this whole idea conceptualizes what fracking and the shale boom was. Fracking and non-conventional reservoirs added billions of oil barrels into proven reserves, and millions of barrels of new production into the marked, and they also moved further away the idea of the Hubert’s Peak (WANG and KRUPNICK, 2015). The panoramic of the traditional oil producers and importers also changed, countries considered to be net importers such as the US (Figure 3), moved to position themselves as exporters and almost completely independent on energy requirements.

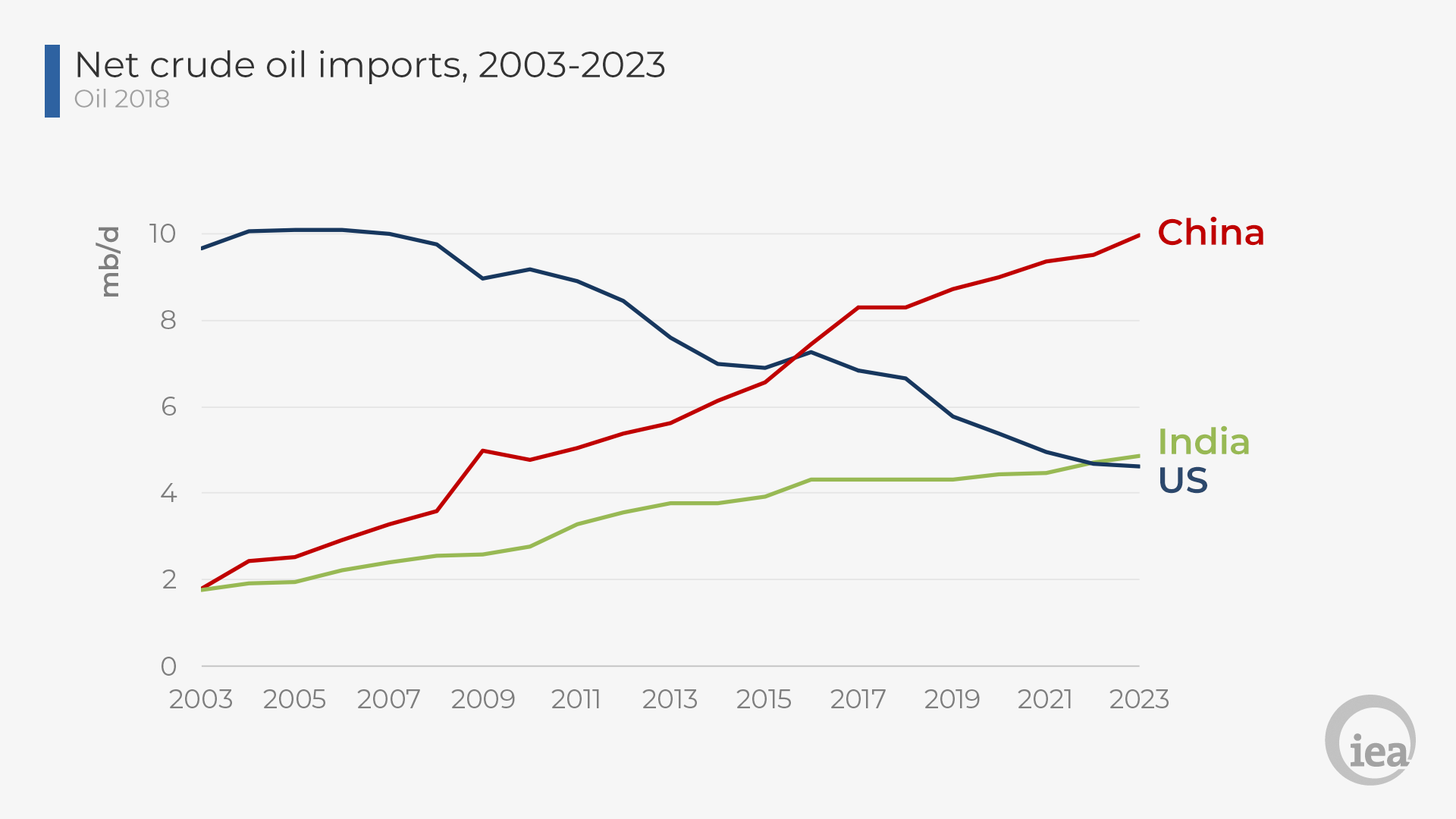


Figure 3 Net crude Oil imports, 2003-2023 – US shifting from consumer to producer

The consequences of these changes were dramatically for the oil and gas industry, especially when we consider other factors such as developments on other energy sources such as renewables and change in energetic habits. On 2014 reacting to the decrease of oil prices, the industry attempted to use its traditional formula of cutting production, specially OPEC countries, however they soon realize that this measure was not going to be sufficient and sustainable in the future, and the industry will have to change its core fundamentals to remain profitable. So besides cutting production how does the oil and gas industry reacted to these new market dynamics?

**2.1.2 The Oil & Gas industry new strategy**

After realizing that traditional measures applied in the past will not suffice to face off the new industry and market dynamics the oil and gas industry look internally and realize potential benefits from its core and fundamentals. First off, before this downturn the industry had a mentality of “overall optimism”, they will take the risk to embark themselves in no matter what development project expecting a price bullish trend, to pop up just around the corner (Maestro et al., 2019). Projects such as the development of bituminous tar sands on Canada can only break even with barrel prices above 70 USD, such projects were clearly not sustainable in time. Such an optimistic and wasteful mentality produced in the past excess in inventories, resources and frozen working capital, which created plenty of inefficiencies.

After 2014 all oil big players, moved into a more pessimistic mentality and re-evaluated their portfolios, moving resources from high break even price developments into lower cost of production areas, In January 2018, BP announced it would only approve new projects that were profitable at less than 40 USD per barrel and in 2017 Shell divested the majority of its Athabasca oil sands business on the grounds of poor economics in a lower oil price world (and perhaps with an eye to the future, given the higher emissions produced by this unconventional source). The industry decided to hold on projects with perhaps exciting new reservoir discoveries and prioritize commercial feasibility, and that can be seen on the recent downward trend in new oil and gas reservoir discoveries (Figure 4) and the slow of expenditures in exploration and production projects (Figure 5) (Maestro et al., 2019).

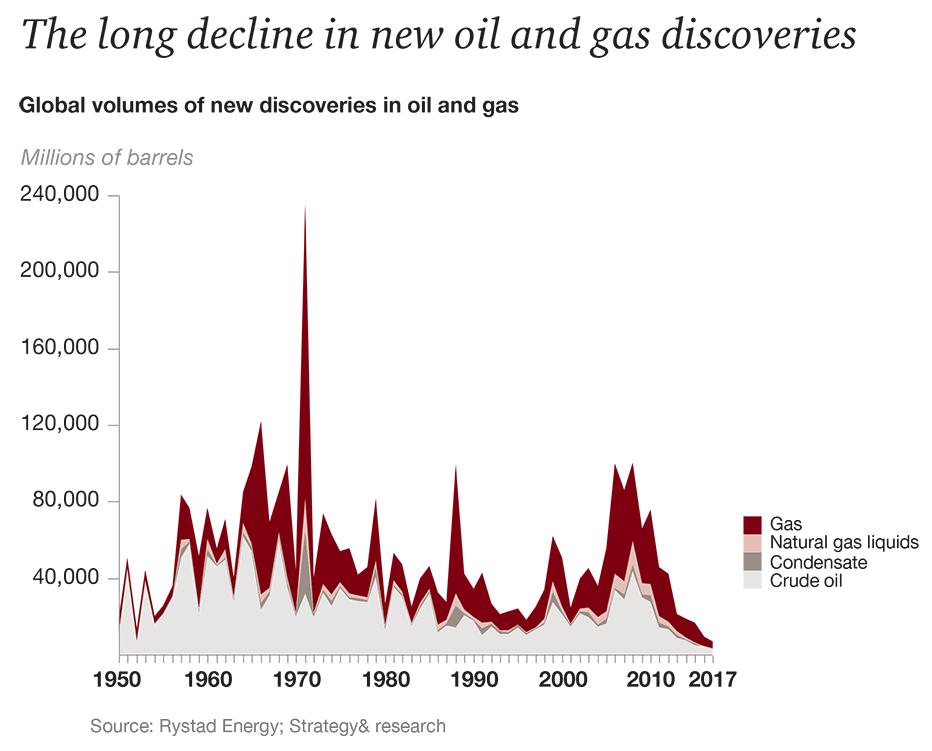


Figure 4 Volume of new discoveries of oil and gas

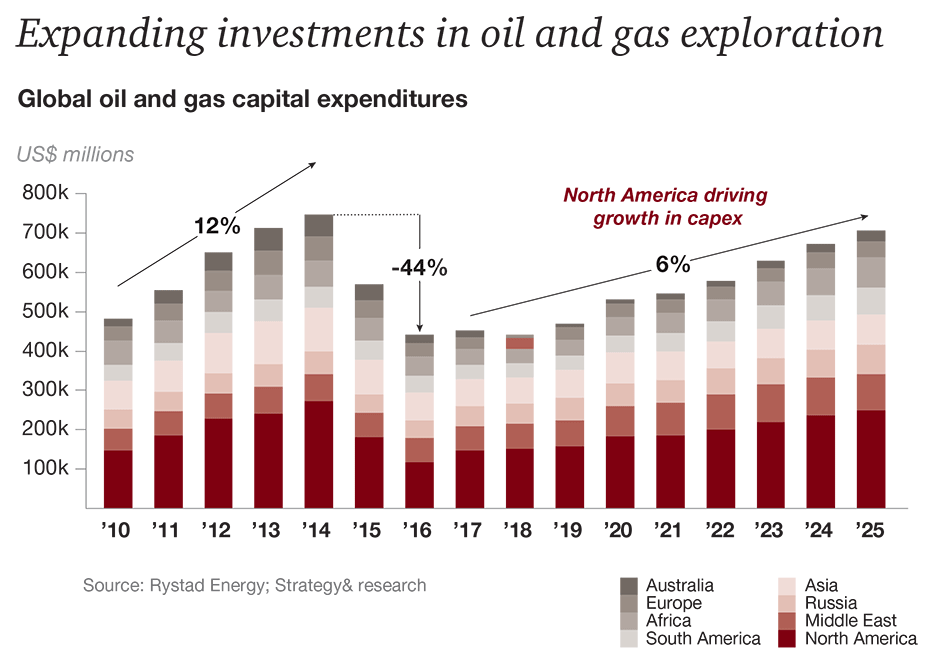


Figure 5 Investment on oil and gas exploration projects

The change of mentality did not only affect the way oil and gas companies evaluated new projects; it also affected their operative mentality. Big player took a more lean and austere vision and decided to resize themselves and delay resources expansion and maintenance for better and adequate moments. After 2014 on average most of the big operators experienced a workforce reduction of almost 10% (Figure 6) of their total number of employees, just for the sake of cost reduction (Maestro et al., 2019).

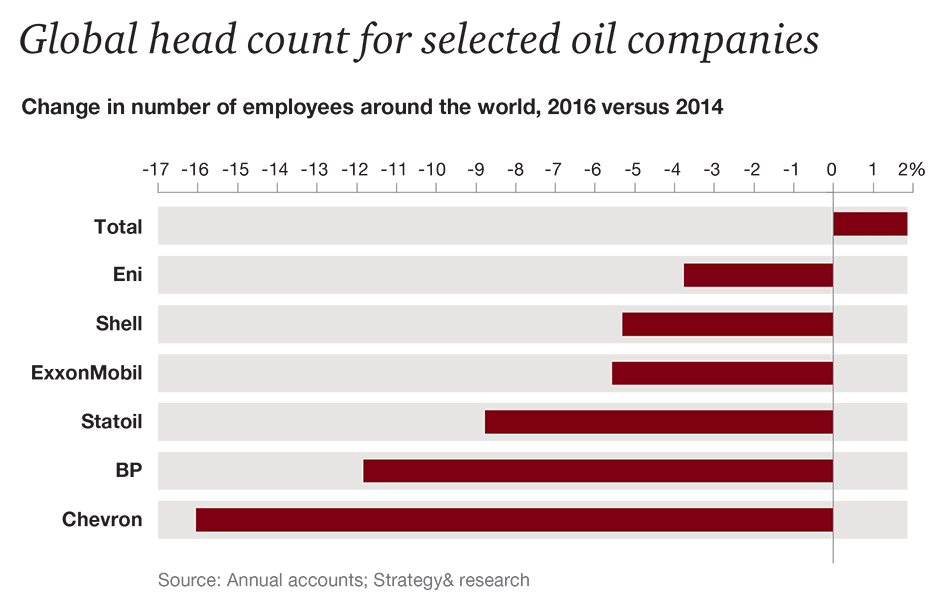


Figure 6 Global head count change of major oil IOC operators

The more austere mentality also pushed major big oil companies to back away from their vertical integration and expansive ambitions, before the crisis plenty of operators were delighted with a vision of complete vertical integration one similar to the one Henry Ford had on the early 20th century with his automotive Detroit Rouge complex. Operators had a vision on which they will not require service companies for their projects, a vision on which they will dominate and be part of every single aspect of the whole production chain, and therefore retain all revenues, however this foolish dream was only possible at incredibly high oil prices, so they could cover the inefficiencies generated the learning curve of some of these activities, activities that they were just recently performing. Looking at this vision with more perspective it is easy to realize how difficult and wasteful it would have been to continue with such a plan since some of the service companies have technology and knowledge that is being developed since the early 20th century and that requires expenditures of billions of dollars per year in research and development. Such expenditure and time gap create a very steep learning curve to catch up.

On recent years operators resigned to such dreams and embraced a new partnership mentality with their service provider, a mentality on which they can profit from individual domain expertise by letting the best of the class to work on what he is best at. Another key aspect of recent year has been a refocus on digitalization, especially on the aspects of data analytics, such new technologies have proven in other industries to boost dramatically the operational and financial efficiency (Maestro et al., 2019).

**2.1.3 Oil & Gas industry, what to expect on the short term?**

The current expectation for the short-term future of the industry are mainly optimistic, prices have been stable at a 60-70 USD and slowly on the rise thanks to worldwide production cuts. Positive expectations are also powered up by economic growth indicators such as low unemployment and a recent boost on investment specially in the US due to recent tax cuts (Deloitte, 2019). Despite the positive signals there are still some concerns on the overall shape of the industry, the delayed maintenance and asset expansions of 2014, combined with a the decrease of new oil and gas discoveries might put the industry on a critical undersupply condition that will trigger unsustainable high prices again, but perhaps only for the short term, since many operators are dramatically shifting investments to renewable energy projects that might steal a share of the fossil fuels place on the energy market. Lastly it is important not to forget that supply cuts are an artificial tool for price regulation and the major market influence factor it is still on the corner, worldwide there is more production capacity, tremendous unconventional reservoirs and new sources of energy to tackle the worldwide demand.

The abundance of moderately priced natural gas in the US, in basins such as Marcellus shale and Permian, does not receive the deserved attention as the oil sector. And yet it is enabling very material long-term change in US and global energy and geopolitical markets. Natural gas continues to grow as a source of lower-carbon power generation on the US and worldwide. The wave of new investment in petrochemical facilities would not be possible without the growing US natural gas and NGL supply. Moreover, the United States is now a major player in global LNG markets, with two facilities in operation, at Sabine Pass and Cove Point, and four more due to start up in 2019. This is expected to shape global prices, trade flows, and business models (Deloitte, 2019).

Expectations of a positive oil and gas market on the years to come are not considered unreasonable, since they have rational bases, although this position is fragile as stated before and certainly does not mean that the industry has a free pass to comeback to their previous mentality, since its position can be considered fragile. The market requires a growing energy demand, therefore a growing economy and a stable competence from other energy suppliers, and those are three very big assumptions for the years to come, which definitively cannot remain true forever.

**2.1.4 Energy and Economic development on the long term**

It is practically a law to assume that economic growth and energy demands are linked. As economies grow, energy demand increases; if energy is constrained, GDP growth pulls back in turn. That’s been the case since the dawn of the Industrial Revolution, if not long before, even at the age of pillage and military expansion as a source of economic growth, armies required resources such as food, shelter guns and transportation and every single one of these needs requires energy in other to satisfy them. Food need energy to be cooked, shelter requires, materials for construction and heating capabilities and for guns and transportation the energy requirements are beyond obvious.

But what has remained true in the past, does not necessary remains the same in the future, and recent studies suggests that we’re beginning to see a decoupling between the rates of economic growth and energy demand, which in the decades ahead will become even more pronounced (Mckinsey, 2019). But let us not to be foolish, the reasons are not because the world will require less energy. The human population will continue to use energy in their daily lives, and in the future ahead, more people will have access to more modern appliances and on-the-grid housing. Businesses will still need energy to run; economies will require it to grow. Nonetheless, new technologies and larger trends should cause the energy demand curve to flatten (Figure 7).

The decoupling of the rates of economic growth (climbing steadily) and energy demand growth (ascending, but less steeply) will largely be a function of the following four forces:

* A steep decline in energy intensity of GDP, primarily the consequence of a continuing shift from industrial to service economies in fast-growing countries such as India and China.
* A marked increase in energy efficiency, the result of technological improvements and behavioral changes.
* the rise of electrification, in itself a more efficient way to meet energy needs in many applications.
* the growing use of renewable resources that don’t need to be burned to generate power, this is a trend with the potential not only to flatten the primary energy demand curve but also to utterly change the way on which society thinks about power and energy.

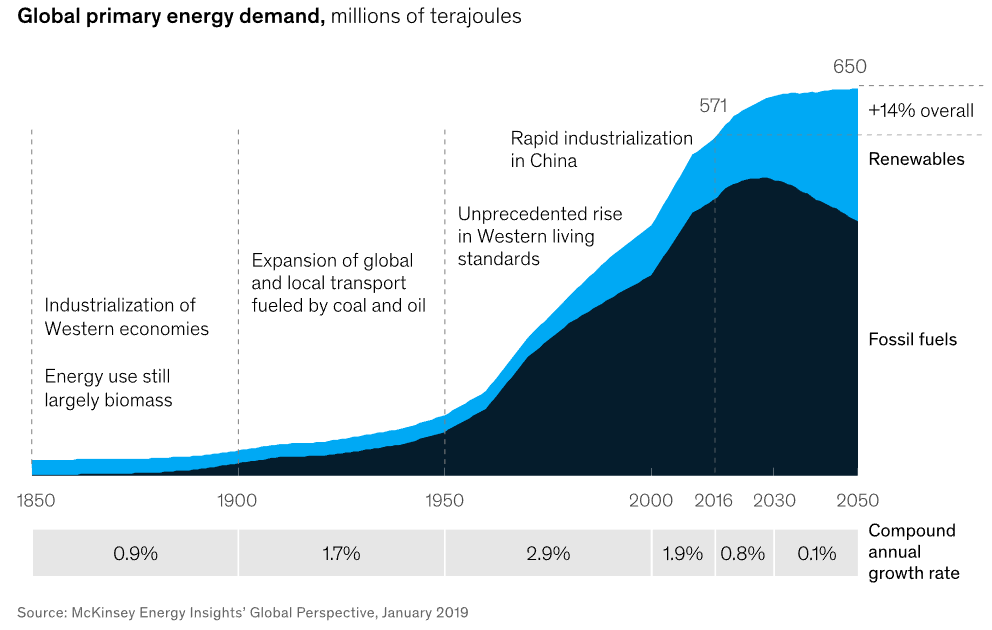


Figure 7 Expected energy demand

For the Oil & Gas industry this only means that they cannot rely only on higher prices as in the past, and even less on monopoly and cartel market interactions, competition from other energy sources has become a reality and energetic demand which is and was the main boost of the energy market is expected to slow down. The industry is forced to play in the game that other industries have played since long time ago, as in the case of electronics or information and technology, a game on which the stronger and the efficient and the one that is able to adapt, is the only that can survive, for the sake of their survival the industry has to innovate not only on technology directly associated with oil and gas development but in other functional areas such as supply chain, finance and HR, waste is not an option anymore and digitalization will be a key factor to avoid it.

**2.2 Structure and business model of the Oil & Gas industry**

Considered to be the biggest sector in the world in terms of dollar value, the oil and gas industry is a global powerhouse employing hundreds of thousands of workers worldwide as well as generating hundreds of billions of dollars globally each year.

Being oil and gas natural and limited resources found on the subsoil are normally with few exceptions such as the US, considered to be state owned resources, and their development is extensively regulated. States by themselves don’t have the technical capabilities and sometimes the financial capabilities to explore and develop these natural resources and this is where major operators known as IOC’s international oil companies come in to play, IOC’s are normally publicly traded companies of the energy sector, which perhaps started as an state owned company but later on, went public, in some cases some countries still owns certain participation of the IOC’s but without control of it, that is the case of Equinor (Previously known as Statoil) on Norway, Total in France and Repsol in Spain. Using a rent model, IOC’s receive a license to explore and develop oil and gas reservoirs from countries with the exchange of payments, they also receive the license to trade the different commodities produced by them.

After years of dealing with the expertise of IOC’s certain countries, and specially countries with massive oil and gas reserves, decide to develop by themselves their own reservoirs independently of IOCs, in some cases certain initially state owned companies decided to become public traded companies for the sake of increase financial efficiencies. State owned companies are known as NOC’s national oil companies, that is the case of PEMEX from Mexico, Saudi Aramco from Saudi Arabia and PDVSA from Venezuela, to name some. In certain countries such as Venezuela and Saudia Arabia almost the whole country economy depends on the performance of the NOC.

The oil and gas industry can be broken down into three key areas: Upstream, midstream and downstream.

The upstream component is also referred to as the E&P (exploration and exploration). This involves the search for underwater and underground natural gas fields or crude oil fields and the drilling of exploration wells and drilling into established wells to recover oil and gas.

Midstream entails the transportation, storage and processing of oil and gas. Once resources are recovered, it must be transported to a refinery, which is often in a completely different geographic region compared to the oil and gas reserves. Transportation can include anything from tanker ships to pipelines and trucking fleets.

Downstream refers to the filtering of the raw materials obtained during the upstream phase. This means refining crude oil and purifying natural gas. The marketing and commercial distribution of these products to consumers and end users in several forms including natural gas, diesel oil, petrol, gasoline, lubricants, kerosene, jet fuel, asphalt, heating oil, LPG (liquefied petroleum gas) as well as several other types of petrochemicals.

IOC’s and NOC’s work on the three major sectors of the industry and cover the entire value chain, but there is still one important player left for mentioning, which are the service companies. IOC’s and NOC’s normally focus on the holistic development of their reservoirs; however, such activities involve expertise in almost all engineering’s and science spectrum, making it almost impossible in terms of financials and practicality to cope with all aspects of it and to dominate the complete value chain. Service companies such as Schlumberger, Halliburton, Baker Hughes just to name some, are niche companies specialized on technical activities such as drilling, cementing, well testing, seismic and many others, and are hired as contractors by IOC’s and NOC’s to execute these services on their assigned or owned fields, services companies have invested billions of dollars in R&D throughout the years to become the best on the services they provide, and are almost an obligation that operators cannot avoid. In recent years some of the biggest service companies have started to offer similar services as the ones offered by the IOC’s providing a similar role as operator, however there are only few successful examples of this dynamic and it doesn’t occur very often.

The last two stakeholders of these industry are the government and society, both of them imposes limitations or opportunities for the oil and gas industry, recently and specially with powerful states such as US or Russia, governments have aggressively use energy as a geopolitical tool to achieve or gain political influence and force outcomes; therefore a governmental energy policy has the potential to shift completely the panorama of the oil industry at local level and in some cases at an international one.

**2.3 Variables of influence for the Oil & Gas industry and markets**

Since early 20th century oil has been considered one of the most important commodities with the capacity of affecting international economic growth. Certain estimates indicate that activities of exploration and production of the oil and gas specifically the drilling sector contributes to about three percent of global Gross Domestic product.

The behavior of oil prices has been shown tremendous impact on the world economy at various levels, from family budgets to corporate earnings and to the national economy (Eika and Magnussen, 2000) (Kilian, 2009). According to IMF, a 10% increase in oil prices results in a 0.2% drop in global GDP. Therefore, when oil prices spiked up to a record high of 145 USD in 2008, and then drop aggressively at the beginning 2014 reaching a 29 USD low in early 2016, this occasioned significant disruption and economic stress on many energy exporting nations such as Russia, Venezuela and Saudi Arabia, even countries not completely dependent on oil like Canada suffered from the price drop. On the other hand, the availability of cheaper oil has demonstrated to be a potent economic stimulus to many net oil importer countries such as China and India, while keeping inflation under check. Given the central role of oil in the economy there is a great deal of attention in the literature in forecasting crude prices.

One of the most watched oil price forecast is from the US Energy Information Administration (EIA), which formally constructs monthly and quarterly forecasts of the price of crude oil at horizons up to two years.3 Although EIA’s short-term forecasts help inform corporate investment and guide resource deployment decisions, their forecasts are hard to replicate and sometimes not very accurate (Miao et al., 2017).

Financial options such as futures-based forecasts provide a market-based expectation of oil prices. Although in principle the futures market should be a good predictor of future spot prices, this is not supported by empirical evidence. Alquist and Kilian (Alquist and Kilian, 2010) document that oil futures prices tend to be less accurate in the mean-squared prediction error sense than a simple no-change forecast (Assumes that oil prices will be the same tomorrow as they are today).

On the past various econometric methods have been proposed to forecast oil prices. However, such forecasting methods have not been particularly successful when compared with the naive no-change forecast (Alquist and Kilian, 2010). The overall, conclusions from prior studies suggest that changes in oil prices are on their nature complex, unpredictable and difficult to model, and as a result the current price of oil may be the best available forecast of future prices.

Being attempting to forecast oil price it is important to understand why oil prices are so hard to predict. The literature provides interesting insights (Hamilton, 2009). Oil prices behave as a consequence of a large set of dynamics and multi-dimensional factors including physical markets factors, financial markets factors and trading factors that are themselves often hard to predict (Miao et al., 2017). Perfect accuracy in forecast is also difficult due to unexpected events such as demand shifts in the global economy, supply disruptions, changes in oil production and inventory demand, geopolitical events, among other factors. These types of events create uncertainty about future supply or demand, which can lead and has led in the past to higher volatility in prices. Therefore, taking into consideration the history of past oil supply and demand shocks consequence from various events, market stakeholders and investors are always considering the possibility of future events and the impact of these events on oil prices. In addition to the size and duration of a potential disruption, investors also consider the stock levels and the capacity of oil producers to counteract a potential supply/demand situation. Sometimes, the pessimist risk approach of speculators and the quantification of speculative oil demand, shocks and other events may, at times, render rational standard econometric models completely un-useful, if it is the case that speculators respond to information not available to the econometrician attempting to disentangle demand and supply shocks based on historical data (Kilian and Lee, 2014).

**2.4 Data science, Artificial intelligence, machine learning and deep learning**

Previous studies such as (Alquist and Kilian, 2010) and (Hamilton, 2009) has shown the complexity of generating econometric models capable of outperforming baseline naïve and judgmental models such as the mean model and the no-information model, that are capable of predicting the behavior of oil prices and explain the behavior of oil prices based on the change of the root variables. Therefore, it seems that there is no value in moving beyond and use more refined methods, since they don’t generate a return of the time investment they require to produce. The complexity arrives not only from the fact that oil prices are hard to predict, but also from the expectations of participants, since not only accuracy is desired but interpretability of the generated model.

Based these previous conditions, traditionally oil and gas projects have used judgmental, naïve models as a basis for their assumptions and when challenged beyond, as much they might use an econometric multivariable linear model. Another interesting aspect to take into consideration is the historically prohibitive prices and the technical requirements of more advanced methods, such methods were only available at high tier information and technology firms, investments banks and at academic level, these sectors have historically focused the use of such methods on other applications such as image, speech, text recognition and others, and therefore making them not available and perhaps even not applicable at project levels. However as expected markets are changing and changing fast, and what seemed impracticable few years ago today is a reality, today markets are hyped with terminology such as artificial intelligence, data science, machine learning and deep learning, terms that on the past seemed to be part of a science fiction movie today and thanks to an overall phenomena of technological democratization have the daily bread of many industries. New applications of these methods are being developed at an incredible rate, and oil & gas projects don’t have to be the exception. But before moving beyond to the applications of these methods, it might be important to define the overall terminology.

**2.4.1 What is data science?**

“Big data” and “Data Science” are just part of the vogue buzzwords of this 4.0 era. However, more than the typical hype, data science comes along with it the opportunity to change business models design and day-to-day decision making. This growing combination of resources, tools, and applications has deep implications in various fields like project management (Waller and Fawcett, 2013).

Today data is becoming widely considered as a driver of better decision making and improved profitability, and this perception has some data to back it up. Based on their large-scale study, McAfee and Brynjolfsson (McAfee and Brynjolfsson, 2012) wrote, “the more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results … companies in the top third of their industry in the use of data-driven decision making were on average, 5% more productive and 6% more profitable than their competitors”.

Generally, data science is the application of quantitative and qualitative methods to solve relevant problems and predict outcomes (Waller and Fawcett, 2013).

**2.4.2 What is artificial intelligence (AI)?**

For the past few years, artificial intelligence (AI) has been the preferred subject of intense media and industry hype. Machine learning, deep learning, and AI come up in numerous articles and conversations, often outside of technology minded context. The public generally talks about a future of intelligent chatbots, self-driving cars, and virtual assistants a future sometimes painted in a grim light and other times as utopian, where human jobs will be scarce, and most economic activity will be handled by robots or AI agents (Chollet, 2018). But it’s important to be able to recognize the signal in the noise so that you can tell world-changing developments from overhyped press releases. So, what are artificial intelligence, machine learning, and deep learning? and How do they relate to each other? (Figure 8)

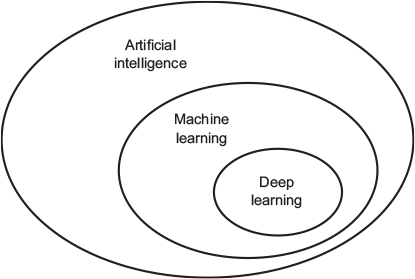


Figure 8 Artificial Intelligence, machine learning and deep learning relationship

Artificial intelligence dates back to the early 50s, a handful of pioneers from the nascent field of computer science started asking whether computers could be made to “think”. A concise definition of artificial intelligence would be “the effort to automate intellectual tasks normally performed by humans.” As such, AI is a general field that encompasses machine learning and deep learning, but that also includes many more approaches that don’t involve any learning (Chollet, 2018). For instance, early chess programs, only involved hardcode, and didn’t qualify as machine learning. For long time, many experts believed that human-level artificial intelligence could be achieved by programming a sufficiently large set of explicit rules for manipulating knowledge. This approach is known as symbolic AI, and it was the dominant paradigm in AI from the 50s to the late 80s and tt reached its peak of popularity during the expert systems boom of the 1980s (Chollet, 2018).

Although symbolic AI was suitable to solve well-defined, logical problems, such as playing chess, it turned out to be inefficient and not that intelligent to figure out explicit rules for solving more complex, fuzzy problems, such as image classification, speech recognition, and language translation. Therefore, the necessity of a new approach and that is the case of machine learning.

**2.4.3 What is machine learning?**

Machine learning arises from the question if weather or not could a computer can go beyond what we order her to do and learn on its own how to perform a specified task? Rather than programmers hardcoding and cramming complete set of direct rules, can a computer automatically derive and learn these rules by looking at data?

Such a question opened the door to a new programming paradigm. In classical programming, the paradigm of symbolic AI, humans input rules (a program) and data to be processed according to these rules, generated an outcome as an answer. With machine learning, humans input data as well as the expected answer from the data, and the machine itself derive the set of rules required to arrive to that specific outcome. These machine derived rules can then be applied to new data to produce new original answers.

A machine-learning system is trained (Figure 9) rather than explicitly programmed. It’s presented with many examples relevant to a task, and it finds statistical structure in these examples that eventually allows the system to come up with rules for automating the task (Chollet, 2018). Although machine learning only became something in the early 90s, it quickly became the most popular and most successful subfield of artificial intelligence, this was also driven by the availability of stronger computational power by new hardware and larger datasets.

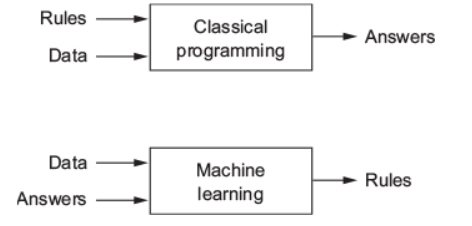


Figure 9 Classical approach vs Machine learning

**2.4.4 Deep learning**

Deep learning is a specific subfield of machine learning, it takes on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations. The deep in deep learning isn’t a reference to any kind of deeper understanding achieved by the approach; rather, it stands for this idea of successive layers of representations as such other appropriate names for the field of deep learning could have been layered representations learning or hierarchical representations learning.

In deep learning, these layered representations are almost always learned via models called neural networks (Figure 10), structured in literal layers stacked on top of each other. The term neural network is a reference to neurobiology, but although some of the central concepts in deep learning were developed in part by drawing inspiration from our understanding of the brain, deep-learning models are not modelling the brain. There’s no evidence that the brain implements anything like the learning mechanisms used in modern deep learning models (Chollet, 2018).

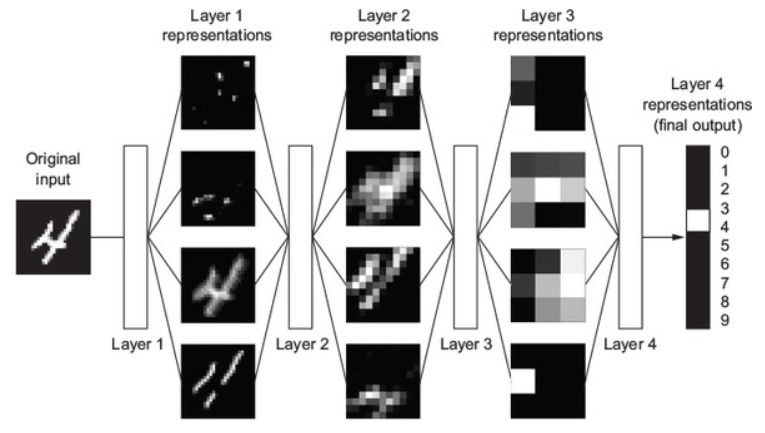


Figure 10 Image classification layers in deep learning

So, what deep learning is, technically, is a multistage way to learn data representations. It’s a simple idea but, as it turns out, very simple mechanisms, sufficiently scaled up, can end up looking like magic. Sometimes such a complex concept can be easily understood using a simple metaphor. Imagine two sheets of colored paper: one red and one blue. They are put one on top of the other. Now they are crumpled together into a small ball. That crumpled paper ball is the input data, and each sheet of paper is a class of data in a classification problem. What a deep learning algorithm (or in fact any other machine-learning model) is meant to do is figure out a transformation of the paper ball that would uncrumple it, to make the two classes cleanly separable again (Figure 11).

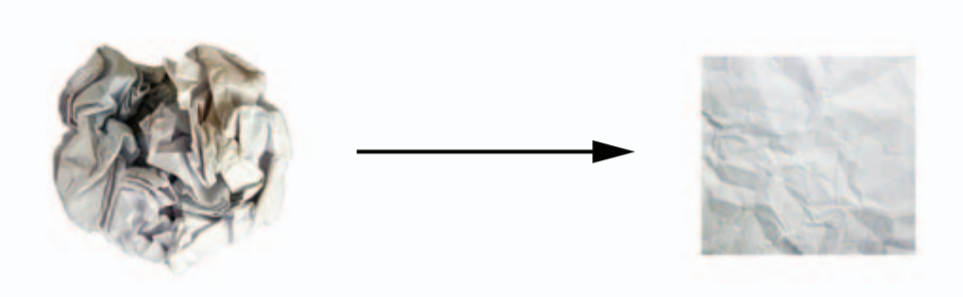


Figure 11 Understanding machine and deep learning

Uncrumpling paper balls is what machine learning is about: finding neat representations for complex, highly folded data manifolds. Machine and deep learning algorithms take the approach of incrementally decomposing a complicated geometric transformation into a long chain of elementary ones, which is most likely the strategy a human would follow to uncrumple a paper ball. Each layer in a deep network applies a transformation that disentangles the data a little and a deep stack of layers makes tractable an extremely complicated disentanglement process (Chollet, 2018).

**2.5 The data science study pipeline**

Data Science is relatively new discipline that combines expertise across a range of domains, including software development, data management and statistics, this recently new domain requires various different skills that can be summarized in three major sections (Figure 12), first we have hacking skills which is the ability to collect the required data, second we have substantive expertise which translates into computing and programming skills which are necessary to transform and work with data and lastly on third positions we have mathematics and statistic skills which allows the data science professional to make appropriate interpretations and analysis on the collected data.

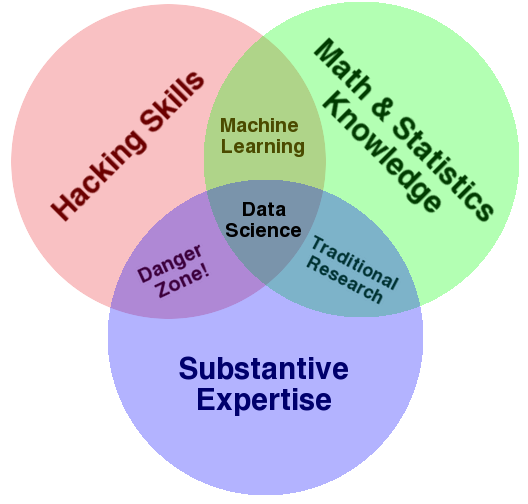


Figure 12 Data science triad of skills

Data science projects normally have the objective of identifying correlations and causal relationships, classify and predict events, identify patterns and anomalies, and infer probabilities, interest and sentiment (Das et al., 2015).

Regarding data science projects and works, the current and widely accepted descriptions of how to do data science generally adopt a task-focused approach. Jagadish (Jagadish et al., 2014) describes a process that includes various steps or phrases such as acquisition, information extraction and cleaning also known as wrangling, data integration, modeling, analysis, interpretation and deployment. Guo (Guo, 2013) shows a similar but different perspective and provided a Data Science Workflow framework. Guo’s workflow defined several high-level phases (Figure 13) such as Preparation, Analysis, Reflection, and Dissemination, with each phase having a specific series of steps that can be repeated within that phase in an iterative analysis.

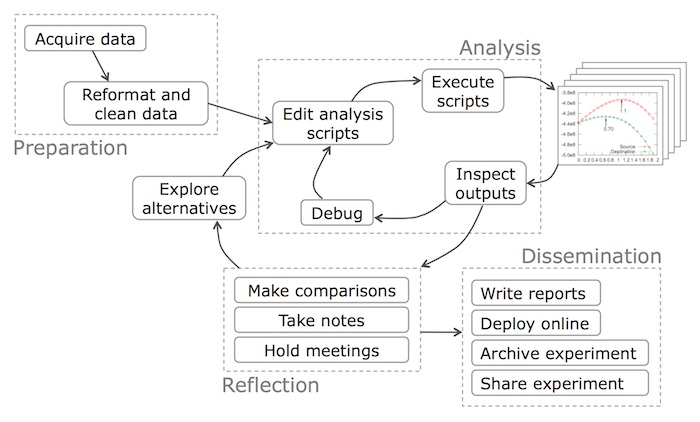


Figure 13 Guo's Data Science Workflow

Peng and Matsui (Peng and Matsui, 2016) describe the process of a data science using a similar approach, the iterative process that is applied to all steps of the data analysis can be conceived of as an epicycle which is repeated for each step along the circumference of the entire data analysis process (Figure 14). Some data analyses appear to be fixed and linear, such as algorithms embedded into various software platforms, including apps. However, these algorithms are final data analysis products that have emerged from the very non-linear work of developing and refining a data analysis so that it can be “algorithmized.” This process consists of five core activities of data analysis: Stating and refining the question, Exploring the data, building formal statistical models, Interpreting the results Communicating the results.

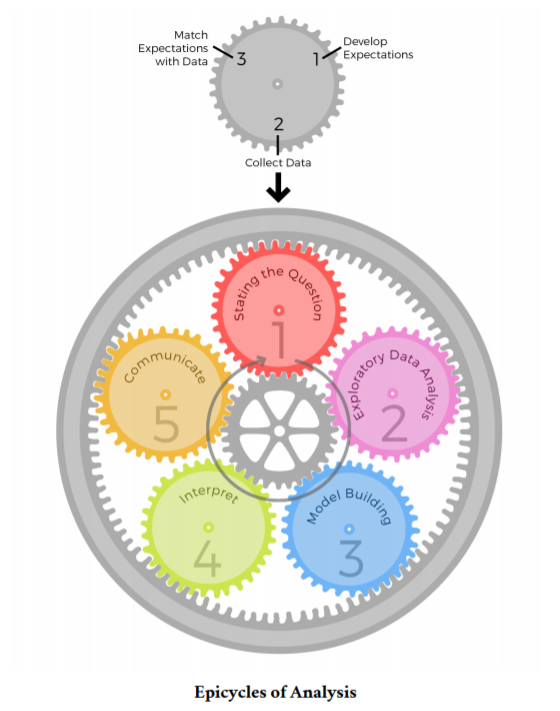


Figure 14 Peng and Matsui epiclycle of data science

For this thesis work we will follow Peng’s and Matsui approach, of the data analysis project, that will be used to derive forecasting models for monthly oil prices and US monthly active rig count.

**2.5.1 Stating the questions**

One of the most common pitfalls of data science projects is the lack of a well stablished and meditated research question or objective, Leek and Peng (Leek and Peng, 2015) highlight how massive studies and data science projects have failed because of misunderstandings about the primary statistical arguments and appropriate questions. A successful data science project requires “thinking” and the thinking begins before even looking or collecting a dataset. There are six basic types of data science questions understanding the type of question asked may be the most fundamental step taken to ensure that, in the end, the interpretation of the results is correct. The six types of questions are: Descriptive, Exploratory, Inferential, Predictive, Causal, Mechanistic. Consecutive steps of the data science pipeline will be linked to this initial question or questions, this step will set the tune of the complete data project and therefore its importance.

**2.5.2 Exploratory Data analysis**

Exploratory data analysis is the process of exploring your data, and it typically includes examining the structure and components of your dataset, the distributions of individual variables, and the relationships between two or more variables. The most heavily relied upon tool for exploratory data analysis is visualizing data using a graphical representation of the data. Data visualization is arguably the most important tool for exploratory data analysis because the information conveyed by graphical display can be very quickly absorbed and because it is generally easy to recognize patterns in a graphical display. There are several goals of exploratory data analysis, which are: To determine if there are any problems with the dataset. To determine whether the question asked can be answered by the collected data. To develop a sketch of the answer of the previously defined question or questions.

**2.5.3 Model Building**

Real life phenomena can sometimes be complex, to simplify them we use models, a model is constructed for the sake of understanding the real world. A common example is the use of an animal which mimics a human disease to help on the understanding of the disease itself, and hopefully, prevent and/or treat the disease. The same concept applies to data. As another example in the world of politics a pollster has a dataset on a sample of likely voters and the pollster’s job is to use this sample to predict the election outcome, on this example the polling data is used to construct a model to predict what will happen on election day. The process of building a model involves imposing a specific structure on the data and creating a summary of the data, so the model is essence a mathematical equation that reflects the shape or pattern of the data, and the equation allows to summarize the thousands of observations that comprises the data set.

**2.5.3.1 The Interpretability and Flexibility tradeoff**

The best model to use depends on the availability of historical data, the strength of relationships between the interest variable and any explanatory variables, and the way in which the model is intended to be used, depending on the intentions for usage there will be a tradeoff to make between model interpretability and flexibility. Model interpretability refers to how easy is to understand the relationship between explanatory and exploratory variables, model flexibility refers to the easiness on which the model is shaped to recreate the behavior of the response variable, sometimes at the expense of interpretability which is the case of various advanced machine and deep learning models, on which the interpretations of the relationship between response and explanatory is not straightforward, some of these models are known as black box models due to the complexity on the relationships and computation (James et al., 2013).

There to general used for models, predictions and inference. Predictive models are used to obtain values of a response variable which usually are not available and difficult to determine, using available data of explanatory variables, predictive models are build with the objective of minimizing the difference between the predictions and the real values of the response variable, most of the times this minimization comes at the cost of sacrificing model interpretability and maximize model flexibility. Inference models in contrast are used to generate insights and understanding on the relationships between explanatory and response variables for such models, the accuracy and error minimization is not the main goal therefore model interpretability is prioritized at the expense of flexibility.

There are hundreds of models to be used, each one with its advantages, however it is always important to consider that whichever is chosen there will always be a tradeoff to be made, and the decision for this tradeoff has to be be linked to the main investigation question made at early stages and objectives of the data science project.

**2.5.3.2 Train and Test data split tradeoff between generalization and optimization**

Imagine an utopic and unrealistic scenario on which all and absolutely all possible data is available to build up a model, on such scenario the model can be trained and will have the chance to look and adapt to all possible behaviors of the response variables and interactions with the explanatory variables, now on an scenario on which predictions are the objective we might ask, why generating a model and risk on residual errors on predictions if we have all data available an therefore we can just easily read directly the values of the response variable? This scenario is unrealistic because having “all data” is just impossible, the goal is to generate the best possible predictive model with the data available that is available for its generation. Predictive models are evaluated on their accuracy on predictions but not only on the accuracy of their predictions on the data that was used for their generation but on out of sample predictions, their true value of predictive models as it has been stated before is when they are used to model the real word and used on real scenarios that not necessarily have been seen before.

To accomplish such a task, the collected data set, hast to be divided on different partitions, on the most basic case, the data set is divided into a “train” data set and a “test” set. The train data is used to generate the model and the test data, to evaluate its performance on data that is has never seen. On more advanced data projects and on cases on which massive amounts of data is available the data can be divided on three partitions, “train”, “validation” and “test”, on this case the train and test set have the same purpose but the validation set is used to tune and tweak the model to obtain better results out of sample results.

A fundamental issue in modelling and specially in machine and deep learning is the tension between optimization and generalization. Optimization refers to the process of adjusting a model to get the best performance possible on the training data (the learning in machine and deep learning), whereas generalization refers to how well the trained model performs on data it has never seen before. The goal is to get good generalization, but generalization cannot be directly control; it is only possible to adjust the model based on its training data. At the beginning of training, optimization and generalization are correlated, the lower the loss on training data, the lower the loss on the test data. While this process is happening, the model is said to be underfit, there is still certain progress to be made; the model hasn’t yet discovered all relevant patterns in the training data. But after a certain number of iterations on the training data, generalization stops improving, and validation metrics stall and then begin to degrade: the model is starting to overfit. That is, it’s beginning to learn patterns that are specific to the training data but that are misleading or irrelevant when it comes to new data (Chollet, 2018).

On general and when there is not a time consideration, data observations are shuffled and then assigned randomly and without replacement to each one of the different data groups (train, test and validation).

**2.5.3.3 Time series data split considerations**

For certain data science projects time can be the most critical conveyor of information, on the case that we are trying to predict the future given the past (for instance, future weather, stock movements, and similar problems), the dataset should not be randomly shuffled before splitting into training and test, because by doing so it will create what is known as a temporal leak, the model will effectively be trained on data from the future, which is a situation that in reality doesn’t occur and there is incorrectly. In such situations, it is important to make sure that all data that belongs to the test set is posterior to the data in the training set (Figure 15) (Hyndman and Athanasopoulos, 2018) (Chollet, 2018) .



Figure 15 Time series data split

**2.5.3.4 Model performance metrics**

It is important to evaluate forecast accuracy using genuine predictions therefore out of sample predictions that belongs to the test set. Consequently, the size of the residuals of the training data is not a reliable indication of how large true forecast errors are likely to be. The accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model.

There are many metrics created to evaluate the performance of models, on out of sample observations, deciding which one of these metrics corresponds better with the created model, depends on the context on which the model is intended to be used and therefore the initial questions. Some of the most popular models and the ones that will be used on this thesis work are:

**2.5.3.4.1 Forecast error**

A forecast “error” is the difference between an observed value and its forecast. Here “error” does not mean a mistake, it means the unpredictable part of an observation. computed as follows:

**2.5.3.4.2 Root mean square error (RMSE):**

The root mean squared error is a frequently used measure of the differences between values predicted by a model or and the observed actual values. The RMSE represents the square root of the mean of the differences between predicted values and observed values or the quadratic mean of the forecast errors. It can be computed as follows:

**2.5.3.4.3 Directional Accuracy Ratio (DAR):**

The directional Accuracy Ratio (DAR), measures the accuracy of predicting the direction of the response variable (i.e., whether the actual value of the response variable increases or decreases in the next time slot). It can be computed as follows:

In the case of doing a random guess of the response variable direction by tossing a fair coin, the DAR would be 0.5. Thus, if the DAR of a model is greater than 0.5, then the model is better than a random guess.

**2.6 Machine and Deep learning algorithms**

There are available hundreds of different machine and deep learning algorithms to choose, on this thesis work we will focus on few of the most popular and most used ones for regression problems, based on recent surveys conducted by Kaggle (“The State of ML and Data Science 2017,” 2017) .Kaggle is a data science competition site, that offers monetary prizes for top performers teams that are able to solve specific data science problems, that are published on the site.

Before going any further and describe more deeply the selected algorithms for this thesis work, it is important to first describe the difference between supervised and unsupervised learning algorithms.

**2.6.1 Supervised and Unsupervised learning**

Most statistical learning problems fall into one of two categories: supervised or unsupervised. On supervised learning for each observation of the predictor measurements there is an associated response measurement called sometimes “labels” the main goal of supervised learning is to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (inference). Many classical statistical learning methods such as linear regression and logistic regression, as well as more modern approaches such as GAM, boosting, and support vector machines, operate in the supervised learning domain.

In contrast, unsupervised learning describes the somewhat more challenging situation in which for every observation we observe a vector of measurements but no associated response or labels, it is not possible to fit a linear regression model, since there is no response variable to predict. In this setting, we are in some sense working blind; the situation is referred to as unsupervised because we lack a response variable that can supervise our analysis. What sort of statistical analysis is possible? We can seek to understand the relationships between the variables or between the observations. One statistical learning tool that we may use in this setting is cluster analysis, or clustering. The goal of cluster analysis cluster is to ascertain, whether the observations fall into analysis relatively distinct groups (James et al., 2013).

**2.6.2 Linear regression**

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Linear regression involves using data to calculate a line that best fits that data. There are many ways one could define a line of best fit. A common way on which statisticians define the best-fit line is the one that minimizes the sum of the squared distances from the observed data to the line. This method of fitting the data line so that there is minimal difference between the observations and the line is called the method of least squares.

The goal in the method of least squares is to fit the regression line to the data by having the smallest sum of squared distances possible from each of the data points to the line. Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression and lasso (James et al., 2013).

Linear regression is a very straightforward simple approach for predicting a quantitative response and one of the first methods to be used regression problems, this mainly due to its high interpretability and easiness of computation, often the method is used as a benchmark for comparison with other more advanced and accurate methods and sometimes as part of the exploratory analysis phase.

**2.6.3 ARIMA Autoregressive Integrated Moving Average**

In a multiple regression model, we forecast the variable of interest using a linear combination of predictors. In an autoregression model, we forecast the variable of interest using a linear combination of *past values of the variable*. The term *auto*regression indicates that it is a regression of the variable against itself.

Thus, an autoregressive model of order p can be written as:

where εt is white noise. This is like a multiple regression but with *lagged values* of yt as predictors. We refer to this as an AR(p) model, an autoregressive model of order p.

Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model.

Thus, a moving average model of order q can be written as:

where εt is white noise. We refer to this as an MA(q) model, a moving average model of order q.

If we combine differencing with autoregression and a moving average model, we obtain an ARIMA model. ARIMA is an acronym for Autoregressive Integrated Moving Average (in this context, “integration” is the reverse of differencing). The full model can be written as:

Where y′t is the differenced series (it may have been differenced more than once). The “predictors” on the right-hand side include both lagged values of yt and lagged errors. We call this model an ARIMA (p, d, q) model where p is equal to the order of the autoregressive part, d is equal to the degrees of first differencing involved and q is equal to the order of the moving average part (Hyndman and Athanasopoulos, 2018).

**2.6.4 Random Forest**

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them (Breiman, 2001). Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance.

Random Forest is a supervised learning algorithm. it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. In simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction (Figure 16).

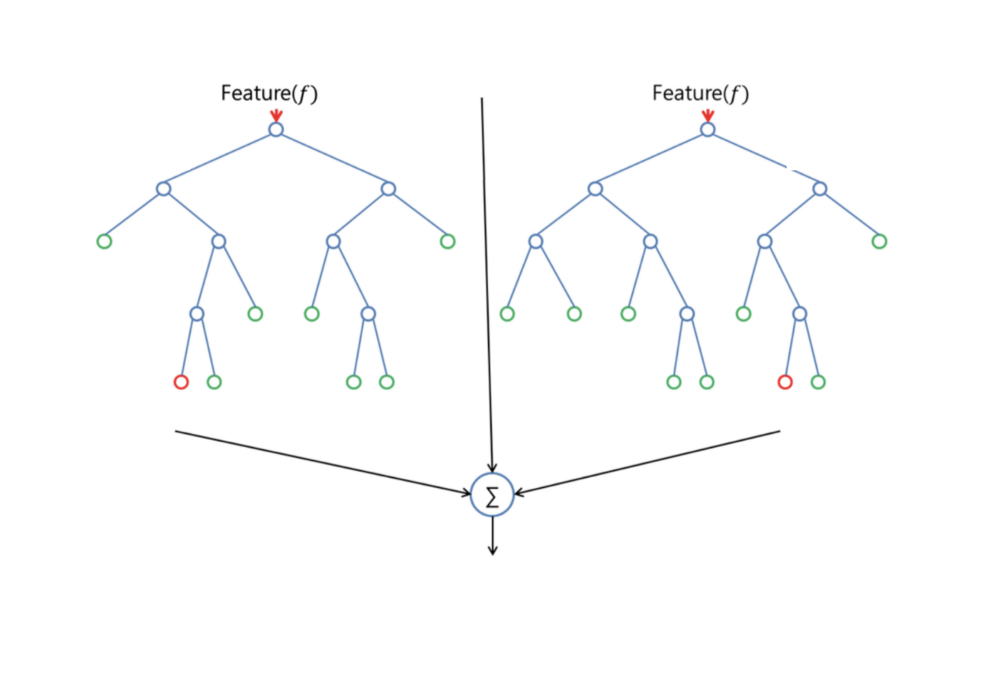


Figure 16 Random Forest feature splitting

Random Forest adds additional randomness to the built-up model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in Random Forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Trees can even be more random, by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does) (Donges, 2018a).

**2.6.5 Xgboost extreme Gradient Boosting**

Xgboost is short for eXtreme Gradient Boosting It is an efficient and scalable implementation of gradient boosting framework. The name Xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms (Chen and He, 2015).

Xgboost It is an implementation of gradient boosting machines created by Tianqi Chen, now with contributions from many developers. It belongs to a broader collection of tools under the umbrella of the Distributed Machine Learning Community or DMLC. The implementation of the algorithm was engineered for efficiency of compute time and memory resources. A design goal was to make the best use of available resources to train the model.

Xgboost is part of the boosting family algorithms. Boosting is an ensemble technique (like random forest) where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models (Brownlee, 2016).

**2.6.6 Support Vector Machines**

Support Vector Machine abbreviated as SVM is a popular and versatile machine learning algorithm can be used for both regression and classification tasks. Although in the past has been widely used in classification objectives, rather than regression problems.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N - the number of features) that distinctly classifies the data points (Figure 17) (Suykens and Vandewalle, 1999).

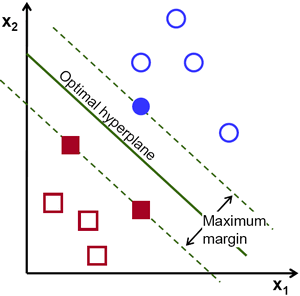
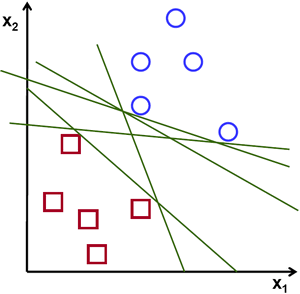


Figure 17 Support Vector Machine description

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, which means the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyperplane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build the SVM (Gandhi, 2018).

**2.6.7 Feed Forward Neural Networks (Vanilla neural network)**

A neural network can be thought of as a network of “neurons” which are organized in layers (Figure 18). The predictors (or inputs) form the bottom layer, and the forecasts (or outputs) form the top layer. There may also be intermediate layers containing “hidden neurons”.

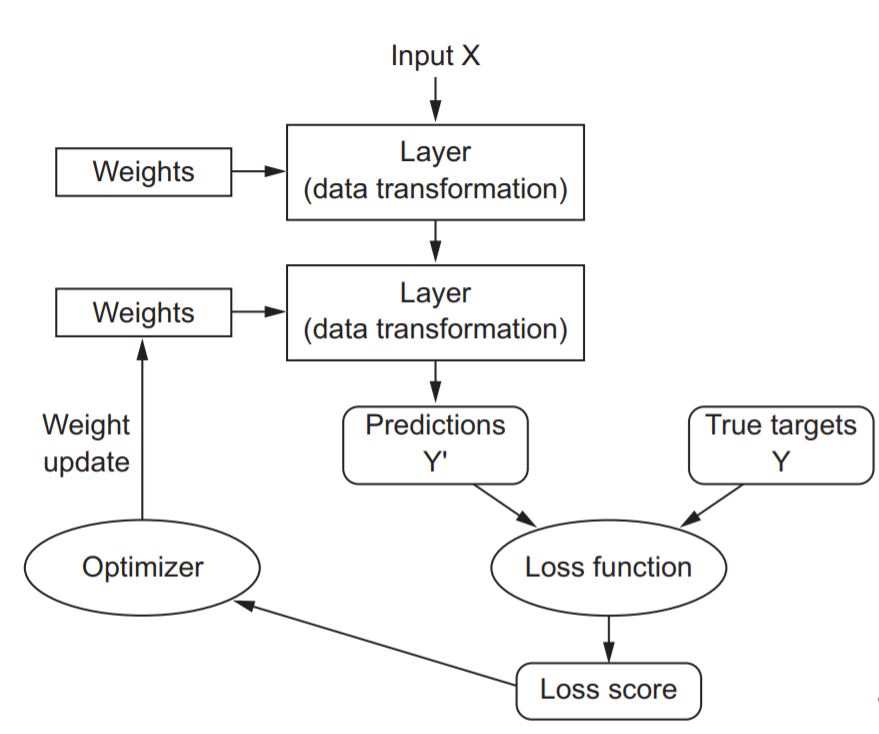


Figure 18 Neural Network process layout

The simplest networks contain no hidden layers and are equivalent to linear regressions. Figure 19 shows the neural network version of a linear regression with four predictors. The coefficients attached to these predictors are called “weights”. The predictions are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a “learning algorithm” that minimizes a “cost function” such as the MSE.

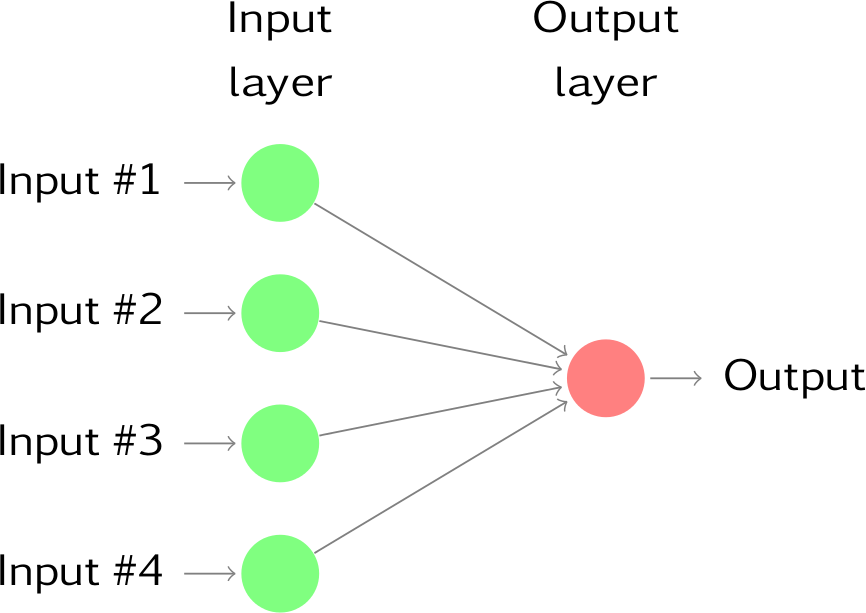


Figure 19 No hidden layer Neural Net

Once we add an intermediate layer with hidden neurons, the neural network becomes non-linear. A simple example is shown in Figure 20.

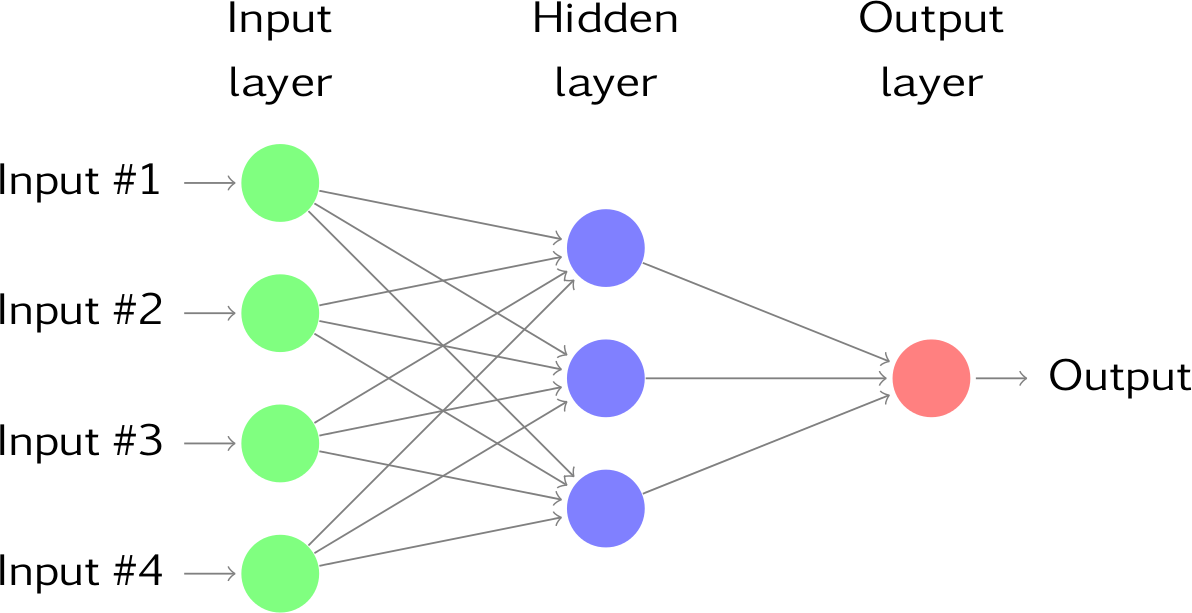


Figure 20 Feed forward Neural net with one hidden layer

This is known as a multilayer feed-forward network, where each layer of nodes receives inputs from the previous layers. The outputs of the nodes in one layer are inputs to the next layer. The inputs to each node are combined using a weighted linear combination. The result is then modified by a nonlinear function before being output. For example, the inputs into hidden neuron j in Figure 20 are combined linearly to give:

In the hidden layer, this is then modified using a nonlinear function such as a sigmoid:

to give the input for the next layer. This tends to reduce the effect of extreme input values, thus making the network somewhat robust to outliers.

The parameters b1, b2, b3 and w1,1… w4,3 are “learned” from the data. The values of the weights are often restricted to prevent them from becoming too large. The parameter that restricts the weights is known as the “decay parameter”. The weights take random values to begin with, and these are then updated using the observed data. Consequently, there is an element of randomness in the predictions produced by a neural network. Therefore, the network is usually trained several times using different random starting points, and the results are averaged. The number of hidden layers, and the number of nodes in each hidden layer, must be specified in advance (Hyndman and Athanasopoulos, 2018).

**2.6.8 Autoregressive Recurrent Neural Network (NNAR)**

Recurrent Neural Networks (RNNs) add an interesting twist to basic neural networks. A vanilla neural network takes in a fixed size vector as input which limits its usage in situations that involve a ‘series’ type input with no predetermined size.

Recurrent Neural Network remembers the past and its decisions are influenced by what it has learnt from the past. Something important to highlight is that basic feed forward networks “remember” things too, but they remember things they learnt during training.

While RNNs learn similarly while training, in addition, they remember things learnt from prior input(s) while generating output(s). It’s part of the network. RNNs can take one or more input vectors and produce one or more output vectors and the output(s) are influenced not just by weights applied on inputs like a regular NN, but also by a “hidden” state vector representing the context based on prior input(s)/output(s). So, the same input could produce a different output depending on previous inputs in the series (Figure 21).

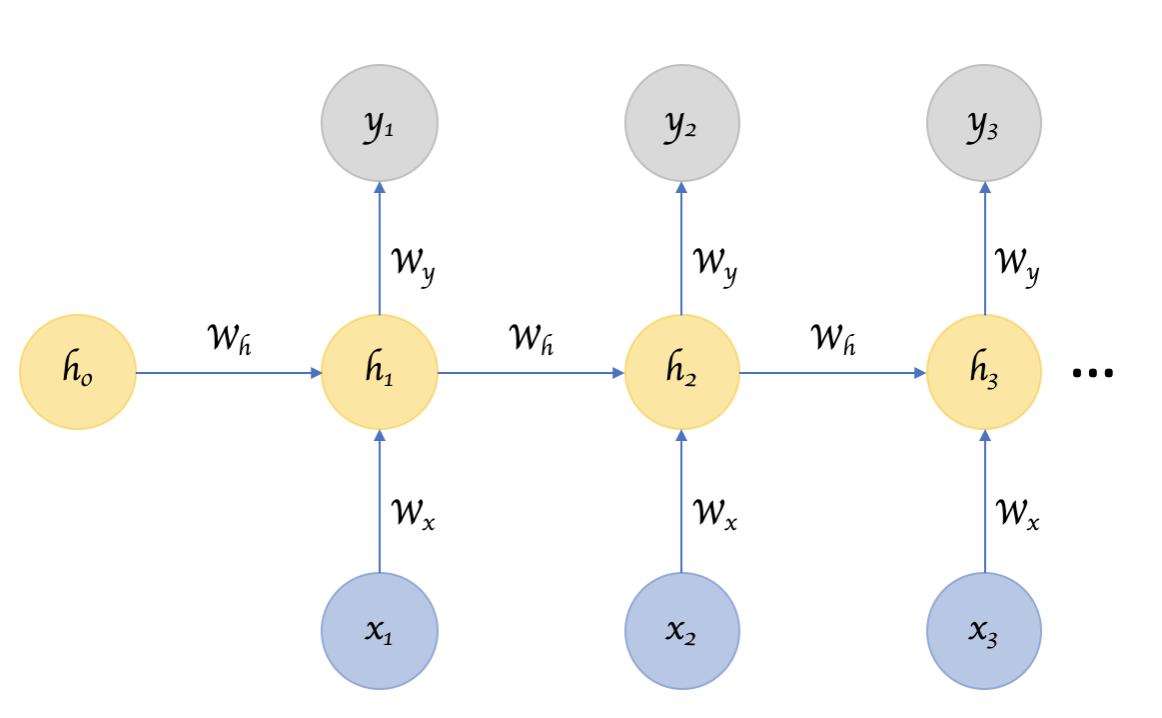


Figure 21 Schematics of a Recurrent Neural Network

In summary, in a vanilla neural network, a fixed size input vector is transformed into a fixed size output vector. Such a network becomes “recurrent” when you repeatedly apply the transformations to a series of given input and produce a series of output vectors. There is no pre-set limitation to the size of the vector. And, in addition to generating the output which is a function of the input and hidden state, the hidden state is updated itself based on the input and use it in processing the next input (Venkatachalam, 2019) (Figure 22).

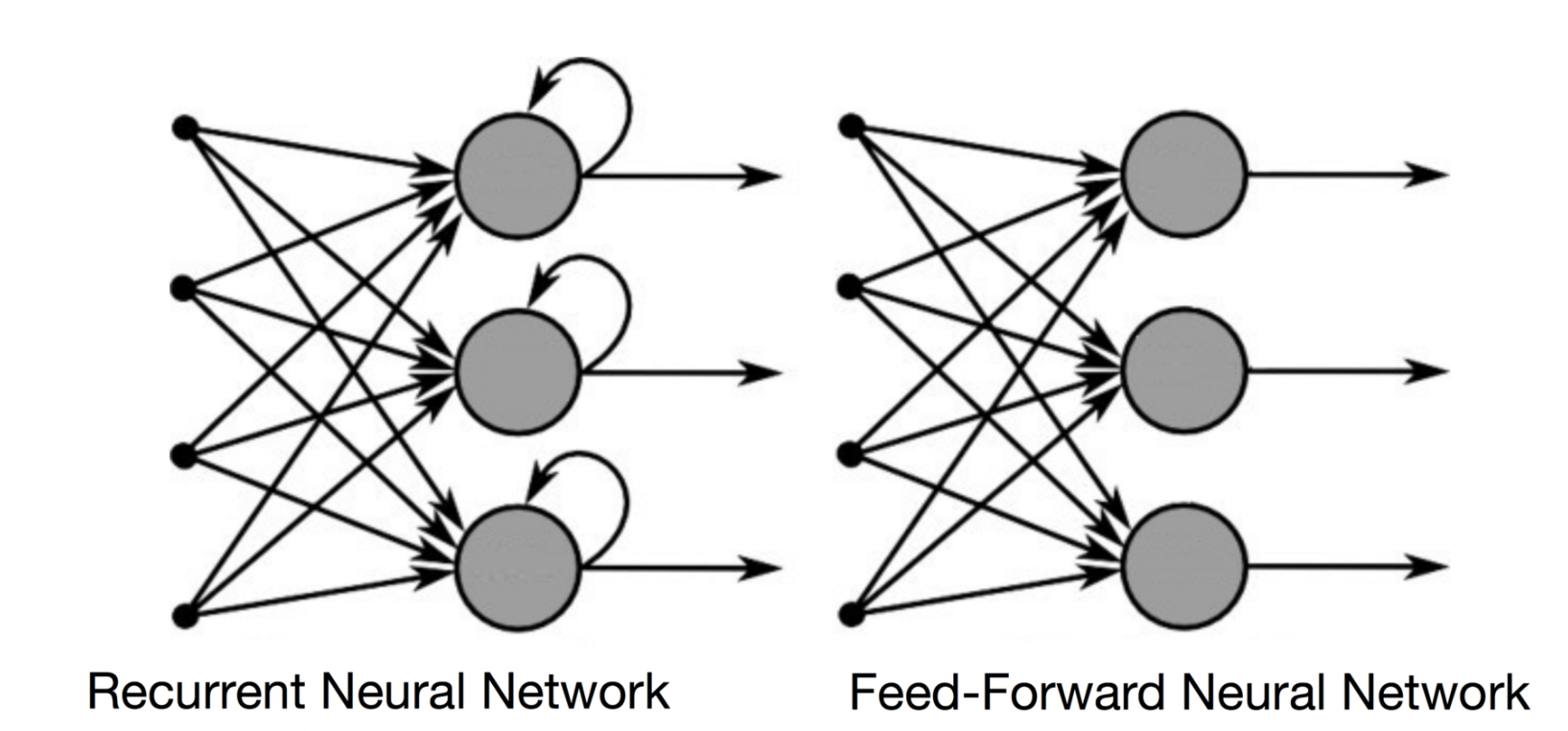


Figure 22 Comparison between NN and RNN

**2.6.9 Long-short term memory neural network (LSTM)**

There are two major obstacles RNN’s have or had to deal with. Exploding gradients and vanishing gradients. A gradient is a partial derivative with respect to its inputs. A gradient measure how much the output of a function changes, if you change the inputs a little bit. A gradient can also be imagined as the slope of a function. The higher the gradient, the steeper the slope and the faster a model can learn. But if the slope is zero, the model stops to learning. A gradient simply measures the change in all weights with regard to the change in error.

We refer to an “Exploding Gradients “when the algorithm assigns a very high importance’s to the weights, without enough reason behind. But fortunately, this problem can be solved with truncation or squash the gradients.

We refer to “Vanishing Gradients “when the values of a gradient are too small, and the model stops learning or it takes too much learning time because of the low values of assigned weights. This was a major problem in the 90s and harder to solve than the previous exploding gradients. Fortunately, it was solved through the concept of LSTM by Sepp Hochreiter and Juergen Schmidhuber (Hochreiter and Schmidhuber, 1997).

Long Short-Term Memory (LSTM) networks are an extension for recurrent neural networks, which basically extends their memory. Therefore, it is well suited to learn from important experiences that have very long-time lags in between. The units of an LSTM are used as building units for the layers of an RNN, which is then often called an LSTM network.

LSTM’s enable RNN’s to remember their inputs over a long period of time. This is because LSTM’s contain their information in a memory, that is much like the memory of a computer because the LSTM can read, write and delete information from its memory.

This memory can be seen as a gated cell, where gated means that the cell decides whether or not to store or delete information (if it opens the gates or not), based on the importance it assigns to the information. The assigning of importance happens through weights, which are also learned by the algorithm. This simply means that it learns over time which information is important and which not.

In an LSTM you have three gates: input, forget and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it isn’t important (forget gate) or to let it impact the output at the current time step (output gate).

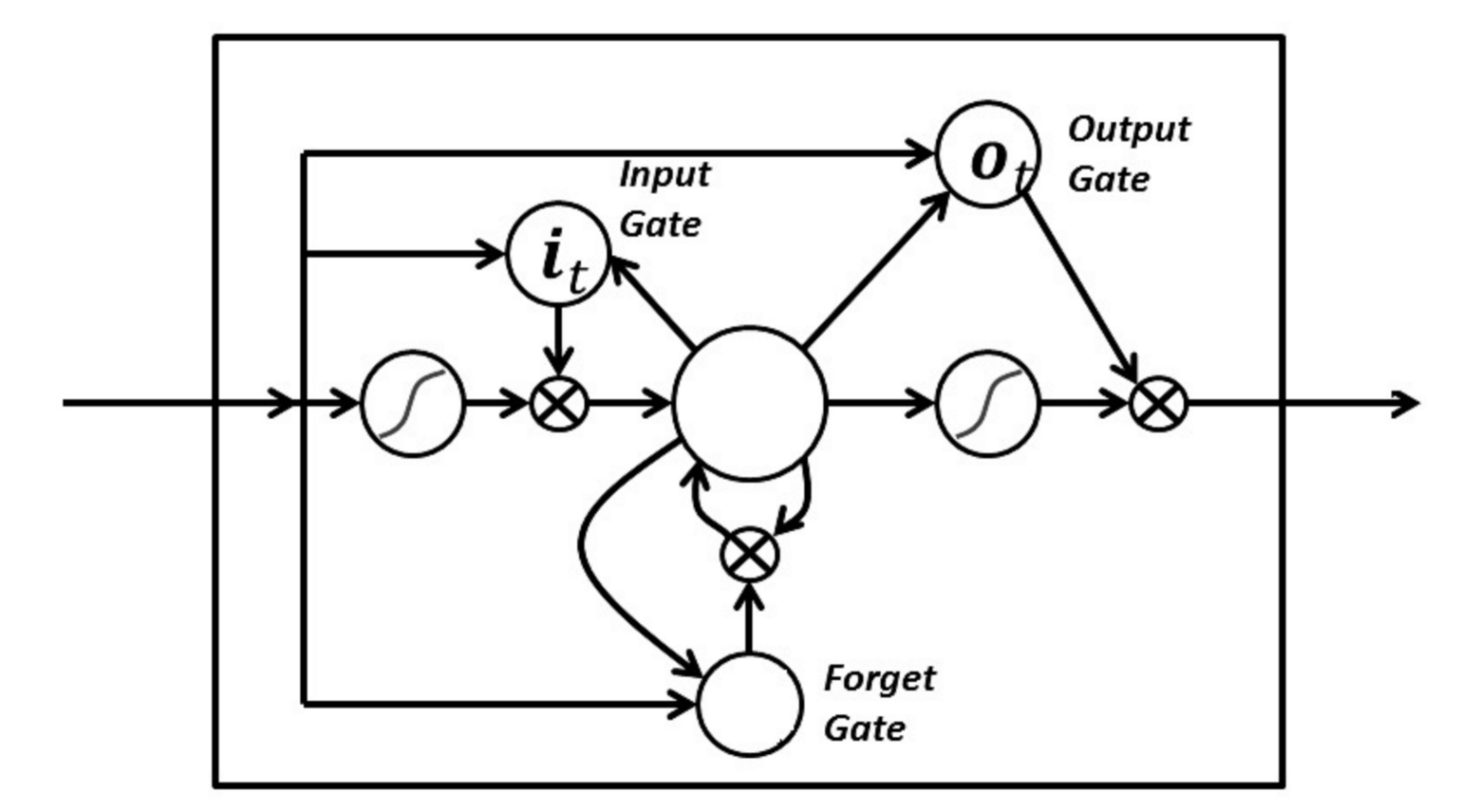


Figure 23 Schematics of the gates of an LSTM Neural Network

The gates in a LSTM are analog, in the form of sigmoid, meaning that they range from 0 to 1. The fact that they are analog, enables them to do backpropagation with it.

The problematic issues of vanishing gradients is solved through an LSTM because it keeps the gradients steep enough and therefore the training relatively short and the accuracy high (Donges, 2018b).

**2.6.10 Unsupervised learning K-means clustering**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset (Figure 24).

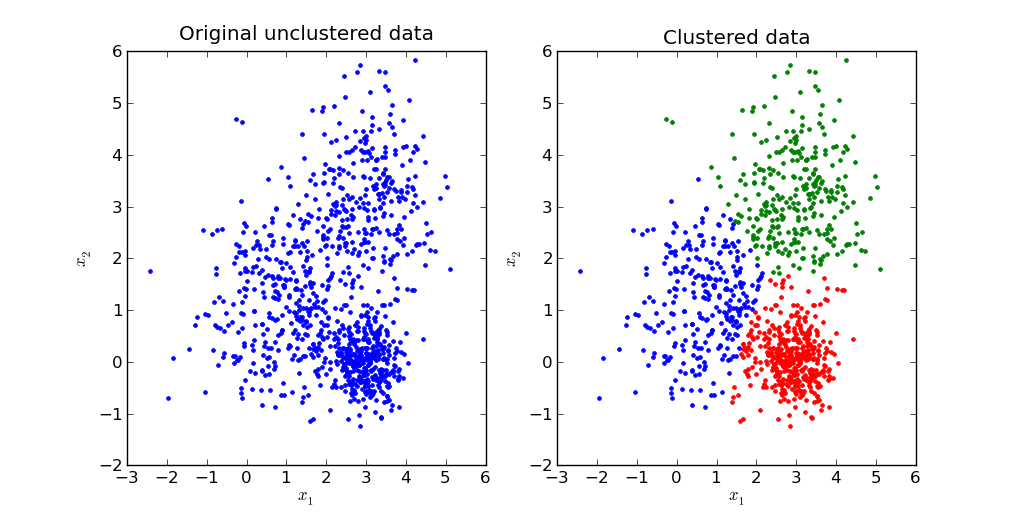


Figure 24 Clustered data example

A cluster refers to a collection of data points aggregated together because of certain similarities.

You’ll define a target number *k*, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies *k* number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The *‘means’* in the K-means refers to averaging of the data; that is, finding the centroid.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

* The centroids have stabilized — there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

**2.8 Previous work on oil price forecasting models**

Being crude oil one of the most important commodities for trading markets, and having oil such a tight relationship with countries economic development (Eika and Magnussen, 2000), it is not surprising to encounter an overwhelming amount of academic literature and research projects devoted to the forecast and understanding of oil price behavior. On this section we will summarize and analyze a sample of these studies, to generate a solid foundation for the construction of both of our forecasting models, for monthly oil price and US active rig count. It is important to mention that the contribution of some of these works, is not only related to the accuracy of the generated model, in many cases some of these models were not even able to beat the naïve no information model or the mean forecasting model, however their work revealed interesting patterns and correlations that in a sort of accretion phenomena has helped and improved further research on this topic. In general, regarding oil price forecasting researcher report success by expanding the range of explanatory factors and integrating them into their models.

Dees et al. (Dees, 2007) on his work related to factors that affect the development of the euro area decided to model oil prices as a function of oil inventories and demand, OPEC production, producer quotas and production capacity.

Ewing and Thompson (Ewing and Thompson, 2007) on their work of investigation of the effect of energy policies focused on the relationship between macroeconomic variables and crude oil prices, and found that oil prices tend to be procyclical.

Kaufmann et al. (Kaufmann and Ullman, 2009) on their study of the effect of speculation factors on oil prices, studied the potential effect of refinery utilization rate on crude oil prices and showed that lower refinery rates correspond with higher crude oil prices and added to the work an analysis on the role of speculation and forward-looking behavior in crude oil prices that has been carefully examined by various authors, and that will be essential for this thesis work.

Sari et al. (Sari et al., 2011) on their study of global risk perception and its impact on oil prices used the VIX index as a proxy for global risk and found that oil prices are influenced by related commodities.

Coleman (Coleman, 2012) on a more fundamentalist approach found out a positive relationships between crude oil prices and proxies for speculative and terrorist activities while controlling for fundamental and market parameters.

Miao et al. (Miao et al., 2017) created a LASSO regression model including 26 different parameters of distinctive nature, the obtained model was able to beat the accuracy of the naïve no information baseline, the model also reported how crude oil inventory announcements impact oil futures and options prices. There is also some evidence that forecasting models based on economic fundamentals work better at shorter horizons (up to 3 months); whereas, models based on the crack spread work better at longer horizons (between 12 and 24 months).

Baumeister and Kilian (Baumeister and Kilian, 2016) present a combination approach with six different models that marginally improves the forecasting performance in comparison to the no-change forecast.

Wang et al. (Wang et al., 2017) created a time-varying parameter (TVP) model with single predictor, and were able to obtain a model with high accuracy performance with a mean squared predictive error reduction in comparison with the naive benchmark of no-change forecast, as high as 17% and a directional accuracy ratio as high as 0.645, all of this over a 24 months’ time horizon test period.

Fan et al. (Fan et al., 2016) created a model that mixes the insights gained by an independent component analysis (ICA) with the power and accuracy of a supports vector machine regression using a second degree polynomial kernel, and manage to obtain a considerable improvement on values of RMSE in comparison with the naïve benchmark model and directional accuracy ratios with values as high as 0.88.

Zhao et al. (Zhao et al., 2017) used a deep learning approach to forecast oil prices, specifically an stacked denoising autoencoders, SDAE neural network, and a bootstrap aggregation (bagging) which generates multiple data sets for training a set of base models known as SDAEs, and managed to obtain a 26 % reduction on RMSE in comparison with the naïve no information benchmark, and also directional accuracy ratios of values as high as 0.651.

Chai et al. (Chai et al., 2018) used an stacked model approach by combining results of different models, the propose a combination that is able to capture a variety of fluctuation features in crude oil data series, including change points, regime-switching, time-varying determinants, trend decomposition of high-frequency sequences, and the possible nonlinearity of model setting. First, product partition model-K-means (PPM-KM) model is used to detect change points in the oil price sequence. Next, we apply a time-varying transition probability Markov regime switching (TVTP-MRS) model to identify the regime-switching characteristic. Then, we use Bayesian model averaging (BMA) to filtrate main determinants at each regime. Finally, the time-varying parameter structure time series model (TVP-STSM) is used to decompose the oil sequence, capture the time-variation of coefficients in “volatile upward” regime, and forecast the crude oil price.

Funk (Funk, 2018) investigated the out-of-sample performance of seven individual forecasting models and its combinations. His results show that it is possible to construct better forecasts compared to a naive no-change benchmark for horizons up to 24 months with gains in the MSPE ratio as high as 25%.

**3 Methodology**

The previous parts of the current thesis reviewed existing literature on various topics, including, the generic framework of a successful data science project, general machine and deep learning modeling techniques and considerations but specific previous studies directly related to the main objectives of this thesis which is the creation of a forecasting model for monthly oil prices and US monthly active rig count. The previous section served to stablish a baseline framework. The following part presents and justifies the methodology employed to conduct this study and arrive successfully to the desired objectives. An important point to mention at this stage is that this study will focus primarily on US oil market dynamics, since is the country with the highest amount of data available from open sources and also one of the countries that recently is disrupting and leading aggressively the oil and gas industry and markets.

**3.1 Variables selection (Fishbone analysis)**

As various previous researches such as Miao et al. (Miao et al., 2017), Coleman et al. (Coleman, 2012), Kauffman et al. (Kaufmann and Ullman, 2009) and many other, we are interested in primarily studying and gaining insight the various and possible relationships that exist between our selected response variables and explanatory variables.

Similar to the approach taken by Miao et al. we will also look for explanatory variables coming from different groups or sources, that previous research and our own hypothesis suggests that might influence the behavior of both of our response variables.

There is vast evidence of previous studies on the influence of oil stock, oil production, oil imports and many other similar industry related variables, on the behavior of oil prices, therefore the first selected group of variables, for this study will be “Oil & Gas related variables” on which we will include, both of our response variables, oil prices and US active rigs, then the US oil production, US oil stock, US oil reserves and finally US monthly drilled wells. All of these variables are included since they can directly create situations of over and under supply for the oil and gas market.

The second group is intrinsically related with the first one, since the oil and gas market is subjugated to the broader energy market, making it logical to include an “energy” related group variables on which we will include the US monthly energy consumption and the US net energy imports. These variables are included since oil and gas are primarily sources of energy, therefore it is logical to assume that changes in energy demands will affect oil and gas markets.

The third group, the “Economic” variables group is also tightly related with both of the previously defined groups, changes in group one and two will without a doubt affect what happens on this group and vice versa. On this group we will include, the US Gross domestic product, the US life expectancy and the US population. Economic variables are included due to the historical relationship between economy and energy demands, and therefore for modern times fossil fuels consumption. A stronger growing economy will most probably require more energy, more population means there is more people consuming and therefore we should expect a higher energetic demand and lastly higher life expectancies will mean more years of energetic consumption from the same population.

Due to the already found evidence of speculative dynamics for the oil and gas markets, the fourth group will directly be related to such variables and we will name it “politics” group, recently and specially after the miraculous appearance of shale oil and gas US reserves, we have seen a more aggressive us approach regarding their local and international energy policy, the us seems to have grasped into a geopolitical tool, from which they can profit and regulate the behavior of allies and countries that are not so allies, this is mainly thanks to the unhealthy economical relationship that some of the countries have with their oil and gas resources, for some of them the dependency is so strong that the oil and gas industry represents the main contributor for their economies, as examples we have Iran, Venezuela, Russia and many others. buffering the oil prices by overflowing the market with oil and gas, will harm tremendously such economies and therefore debilitating the political influence of these countries on the world. On this group we will include, the US presidential party provenance (Democrat or Republican), the US congress dominating party, the US senate dominating party and as Kauffman (Kaufmann and Ullman, 2009) did, we will include a variable related to armed conflicts but not directly terrorist events as he did but total worldwide armed active conflicts. The Senate, congress and president variables will show how changes in policy affects the oil and gas markets and also the inverse relationship. The armed conflict variable will help us understand if modern war is such an energy consumer as we tend to believe and if so, if it has the potential to severely affect the market dynamics.

The fifth group of variables is more environmentally related, it includes the monthly temperature count of degrees above the 65 F, called the CDD as per “Cooling degree days” threshold and the count of temperature degrees below the same threshold called the HDD as per “heating degree days” both variables are included based on the assumption that cold days require heating and therefore energy so in a similar way cooling days will most likely require air conditioning.

The last group and sixth group the “time” group is perhaps the less interesting one, however it is an obligatory one, on this group variables such as date, year and month are included, these variable might help to confirm if the oil market shows a seasonal or cyclical time dependency.

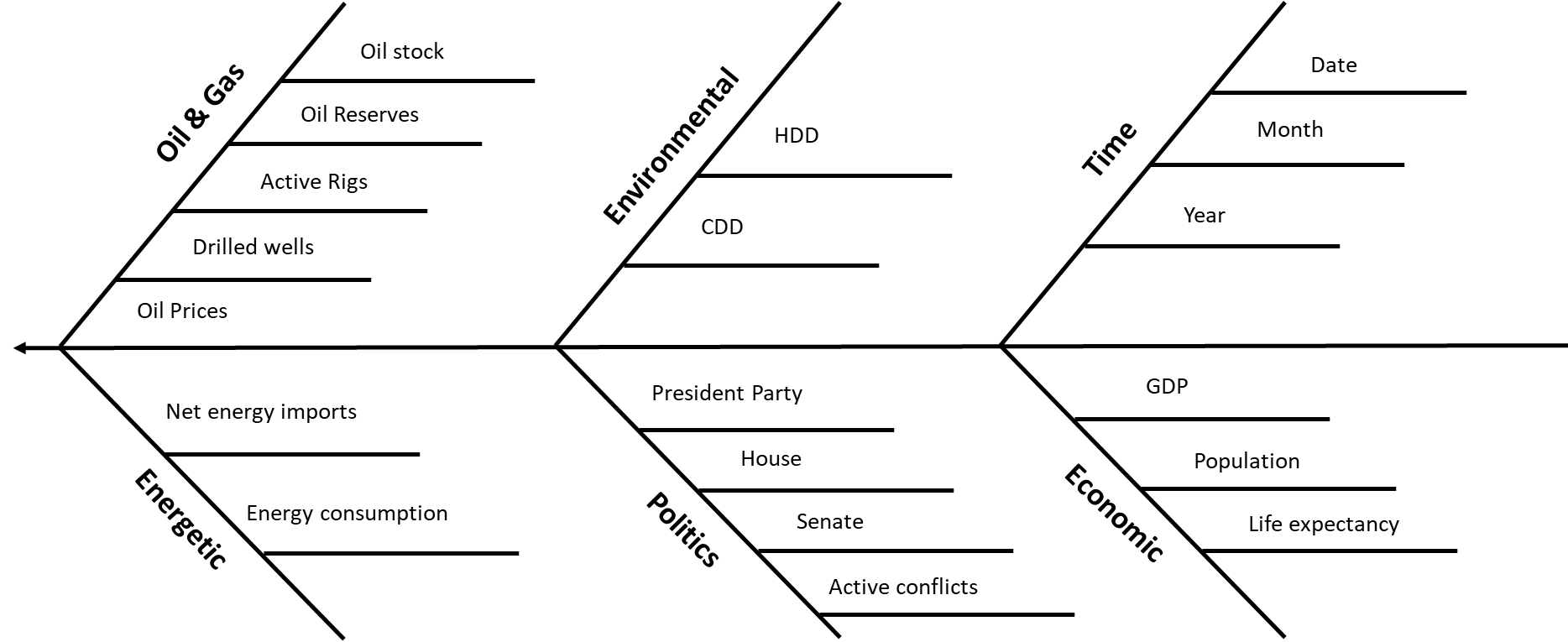


Figure 25 Fishbone analysis for variable selection

**3.2 Data collection**

With the knowledge and insight gained from the previous analysis the next step is to collect the required data, on this thesis work as it was previously stated, we will focus on US data, and its market dynamics. The selected source for data collection is the United States Energy information administration.

The U.S. Energy Information Administration (EIA) is a governmental agency of the U.S. Federal Statistical System responsible for collecting, analyzing, and disseminating energy information to promote rational and facts-based policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment. EIA programs cover data on coal, petroleum, natural gas, electric, renewable and nuclear energy. EIA is part of the U.S. Department of Energy.

One major concern with government data is the natural bias that this government has with its own interest, and therefore the suspicious that somehow that data might be contaminated by these interests, specially data with such a power, as it is in the case of energy data. That is why this organization Is independent by state mandate.

By law, EIA’s products are prepared independently of policy considerations. EIA neither formulates nor advocates any policy conclusions. The Department of Energy Organization Act allows EIA’s processes and products to be independent from review by Executive Branch officials; specifically, Section 205(d) (US Department of Energy, 1977) says:

“The Administrator shall not be required to obtain the approval of any other officer or employee of the Department in connection with the collection or analysis of any information; nor shall the Administrator be required, prior to publication, to obtain the approval of any other officer or employee of the United States with respect to the substance of any statistical or forecasting technical reports which he has prepared in accordance with law.”

Economic US data was collected from the world bank database. The World Bank is an international financial institution that provides loans to countries of the world for capital projects. It comprises two institutions: The International Bank for Reconstruction and Development (IBRD), and the International Development Association (IDA). The World Bank is a component of the World Bank Group.

And lastly conflict data was collected from the most recent release of the “War and peace” study conducted by Max Roser (Roser, 2016).

The final joined and cleaned dataset, comprises 528 complete monthly observations dating from January 1973 up until December 2017.

* 1. **Data Codebook**

The description of variables, original source and units of measurements can be found on Table 2 Variables DescriptionTable 2, Table 3 and Table 3. Data was wrangled and joined using “R” programming language. All data wrangling code used on this thesis work, can be found on the following repository [Link](https://github.com/ceche1212/MBA_Thesis/tree/master/Original_data/Data_wrangling).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Source** | **Variable Name** | **Source** |
| Active\_Rigs | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec5_3.pdf) | Drilled\_Wells | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec5_4.pdf) |
| US\_Population | [Data Link](https://data.worldbank.org/indicator/SP.POP.TOTL?locations=US) | Active\_Conflicts 2 | [Data Link](https://en.wikipedia.org/wiki/List_of_ongoing_armed_conflicts) |
| President\_Party | [Data Link](https://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States) | Active\_Conflicts | [Data Link](https://ourworldindata.org/war-and-peace) |
| Senate | [Data Link](https://history.house.gov/Institution/Party-Divisions/Party-Divisions/) | Oil\_Reserves | [Data Link](https://www.eia.gov/dnav/pet/pet_crd_pres_dcu_NUS_a.htm) |
| House | [Data Link](https://history.house.gov/Institution/Party-Divisions/Party-Divisions/) | Net\_imports\_Energy | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec1_3.pdf) |
| life\_exp | [Data Link](https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=US) | Energy\_Consumption | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec1_7.pdf) |
| GDP | [Data Link](https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=US) | CDD | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec1_21.pdf) |
| Oil\_Production | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec1_5.pdf) | HDD | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec1_20.pdf) |
| Oil\_Stock | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec3_15.pdf) | Oil\_Price | [Data Link](https://www.eia.gov/totalenergy/data/monthly/pdf/sec9_3.pdf) |

Table 1 Data sources

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Date | Date of the observation |
| Month | Month of the observation |
| Year | Year of the observation |
| Active\_Rigs | Number of active drilling rigs on and offshore recorded to be working on that date |
| US\_Population | USA population estimates for the date of the observation |
| President\_Party | Identified political party for the ruling US president on the date observation |
| Senate | Political party that holds the majority of US Senate on the date of the observation |
| House | Political party that holds the majority of US congress on the date of the observation |
| life\_exp | US average life expectancy at the date of the observation |
| GDP | Gross Domestic Product of the US economy on the date of the observation |
| Oil\_Production | US Total oil production on the date of the observation |
| Oil\_Stock | US inventory of oil in stock on the date of the observation |
| Oil\_Price | US average domestic first purchase price, not corrected by inflation |
| HDD | Monthly US HDD are the number of degrees that the daily average temperature falls below 65 °F |
| CDD | Monthly US CDD are the number of degrees that the daily average temperature rises above 65 °F |
| Energy\_Consumption | The amount of energy consumed by the US market on the date of the observation |
| Net\_imports\_Energy | The Net amount of energy imported by the US market on the date of the observation |
| Oil\_Reserves | US proven oil reserves on the date of the observation |
| Active\_Conflicts | Number of world active military conflicts on the date of the observation, includes civil wars |
| Drilled\_Wells | Number of drilled wells on US on the date of the observation |

Table 2 Variables Description

|  |  |
| --- | --- |
| **Variable Name** | **Units of measurement** |
| Date | Date format Year-Month-Day |
| Month | Integer from 1-12 |
| Year | Integer 4 digits |
| Active\_Rigs | Integer each unit represents a drilling rig |
| US\_Population | Numerical in Millions of people |
| President\_Party | Categorical variable Either Democrat or Republican |
| Senate | Categorical variable Either Democrat or Republican |
| House | Categorical variable Either Democrat or Republican |
| life\_exp | Numerical life expectancy in years |
| GDP | Numerical GDP in Trillions of dollars 10^12 USD |
| Oil\_Production | Numerical Oil production in Thousand barrels per day |
| Oil\_Stock | Numerical US inventory of oil in million barrels |
| Oil\_Price | Numerical US dollars |
| HDD | Numerical Monthly US HDD measured un number of °F |
| CDD | Numerical Monthly US CDD measured un number of °F |
| Energy\_Consumption | Numerical Energy consumed in Quadrillion btu’s 10^15 |
| Net\_imports\_Energy | Numerical Net energy imported in Quadrillion btu’s 10^15 |
| Oil\_Reserves | NUmerical Oil reserves in million barrels |
| Active\_Conflicts | Numerical each unit represents an active military conflict |
| Drilled\_Wells | Numerical each unit represents a drilled well |

Table 3 Units of measurements

**3.4 Train and test dataset split**

As part of every machine learning project we have to test, the generated forecasting models with out of sample data, with data that the model did not used or saw during its training, if not the evaluation of the models is flawed and over-optimistic. To achieve this, the collected dataset will be split into two partitions, normally this partition will be done, by randomly shuffling the observations and then randomly assigning them into the two groups, however on the models we are trying to create time is a major information carrier and have to be considered, therefore we will not shuffle the data set since we will create a time leak in our models. For this thesis work we will use the Hyndman (Hyndman and Athanasopoulos, 2018) time series best practices and use continuous observations followed one by another. As many other our test set will be composed by 24 months of continuous data, for this thesis works this means 24 months observations starting from January 2016 and ending up with December 2017, the test dataset corresponds to the indexes 505 up to 528. As a direct consequence the train dataset will be comprised by observations of indexes 1 through 504, which means monthly observations from December 1973 up to December 2015.

It is also a good and recommended practice to add a third data partition called the validation split set, for this thesis we will take the approach proposed by Hyndman (Hyndman and Athanasopoulos, 2018) and use time series cross validation which is also known in econometrics as “evaluation on a rolling forecasting origin” (Figure 25). In this procedure, there are a series of test sets, each consisting of a single observation. The corresponding training set consists only of observations that occurred *prior* to the observation that forms the test set. Thus, no future observations can be used in constructing the forecast. The forecast accuracy is computed by averaging over the test sets.

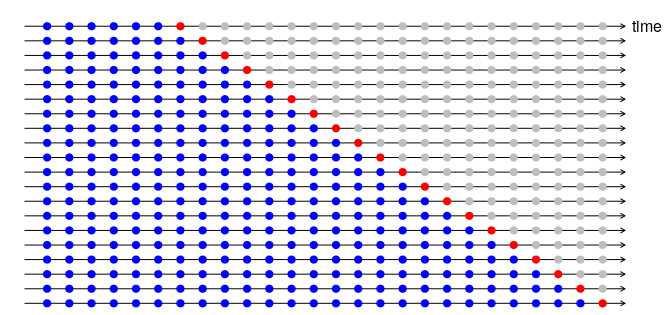


Figure 26 Time series cross validation

**3.5 R Open source libraries for machine and deep learning**

As it was mentioned on previous opportunities the computing tasks of this thesis works will be done using R programing language, R as descendent of S programming language it offers an incredibly rich statistical library of functions that will be helpful during later stages of this thesis work. Another goodness displayed by R is its open source nature, R has a prolific community that is continuously generating new functions and packages, especially on the area of data science, machine and deep learning.

The main goal of this thesis work is not to prove that there are algorithms and advanced computing methods that are able to produce better forecasts of oil price and active rig count, compared to traditional benchmarks such as the naïve model or a simple linear regression, this has already been proved on previous works, specially with complex computing algorithms like the developed on the works of Chai et al. (Chai et al., 2018) or Zhao et al. (Zhao et al., 2017). The main goal of this thesis work is to prove that forecasting and modeling methods that were considered prohibitive in the past, are considered today accessible, mainly thanks to recent trends of technological democratization and that they are at the reach of any project, program or portfolio manager that is willing to use them, this thesis intends to show how far in terms of accuracy we can arrive by using already commercially deployed open source packages compared to benchmark methods, and determine if there is a return on investment in taking the extra mile of using these tools.

For this thesis work we will use various popular R packages or libraries, directly from the central R repository. For the creation of the majority of machine learning models we will use the caret package, created by Kuhn (Kuhn, 2008) (Kuhn, 2019). For data wrangling and handling we will use the dplyr package, created by Hadley Wickham (Grolemund and Wickham, 2014). For time series forecasting we will use the forecast package created by Hyndman (Hyndman and Athanasopoulos, 2018) and lastly for deep learning models we will use the keras package created by Chollet (Chollet, 2018) with the Google’s tensor operations engine called tensorflow (Google, 2015). There are many other packages used for other tasks related to communication and visualization but the one mentioned above are primarily linked with the objective of this thesis work.

**3.6 Benchmark and performance analysis of forecasting models**

The train and test data split is already stablished and the performance metrics for evaluation are selected, the root means squared error and the DAR, although on this thesis work and similar to the work performed by Baumeister and Kilian (Baumeister and Kilian, 2016), Funk (Funk, 2018) and (Chai et al., 2018) we will test the performance of models generated by different already commercially deployed algorithms and therefore we will need a benchmark to compare them with.

For this thesis work we will select three different basic benchmarks to compare with, we consider this benchmarks as part of judgmental benchmarks due to their simplicity and their high frequency of usage outside of the academic world (Hyndman and Athanasopoulos, 2018).

Our first selected benchmark model is the Naïve model, which assumes that the last observed value of the response variable will remain constant in the future; for this specific data set this model corresponds to the value of oil price at index 504, December 2015. It is important to mention that a naïve model is not able to product a DAR evaluation since there is no change in direction on the forecasts, this model will only be used to be compared in terms of RMSE.

The second selected benchmark is the mean model, which assumes that the mean of the past observed values of the response variable will be the values of the future; for this specific data set this model corresponds to the mean of observation from indexes 1 to 504. It is important to mention that a mean model is not able to product a DAR evaluation since there is no change in direction on the forecasts, this model will only be used to be compared in terms of RMSE.

Lastly due to the inability of previous models on generating a DAR benchmark metrics, we will select a third model specifically for DAR purpose; this model shares similitudes with the mean and naïve approach. The no information model assumes that the same patterns of DAR experienced on the training dataset, will remain constant on the forecasted future. for this specific data set this model corresponds to the DAR calculated on observation from indexes 1 to 504. It is important to mention that this model will not produce any RMSE metrics to be used for comparison and its sole purpose is to be used as a DAR benchmark.

With all the benchmarks in place for comparison we need to stablish the correct mechanism for comparison, in order to not be fooled by randomness and ensure that on the case that model generate an observable difference in performance compared with the benchmarks, this difference can be considered statistically significant. To accomplish this task, we will select as test, Diebold-Mariano Test for accuracy forecasting (Diebold and Mariano, 2002).

**3.7 Exploratory Data analysis**

With all the data collected, joined and cleaned the next logical approach is to start exploring it, with the objective of finding interesting relationships, and therefore interesting questions for a posterior phase of inference, it is also critical do determine which variables seem to have more influence on our selected response variables, and use this knowledge to optimize the forecasting models that will be made on successive sections of this thesis work.

On its majority we will tackle the exploratory data analysis with visualization, and we will leave more advanced computation for the next inference section.

**3.7.1 Response variables**

The first step that we will take is to analyze each one of the distributions of our response variables.

**3.7.1.1 Monthly Oil Prices**

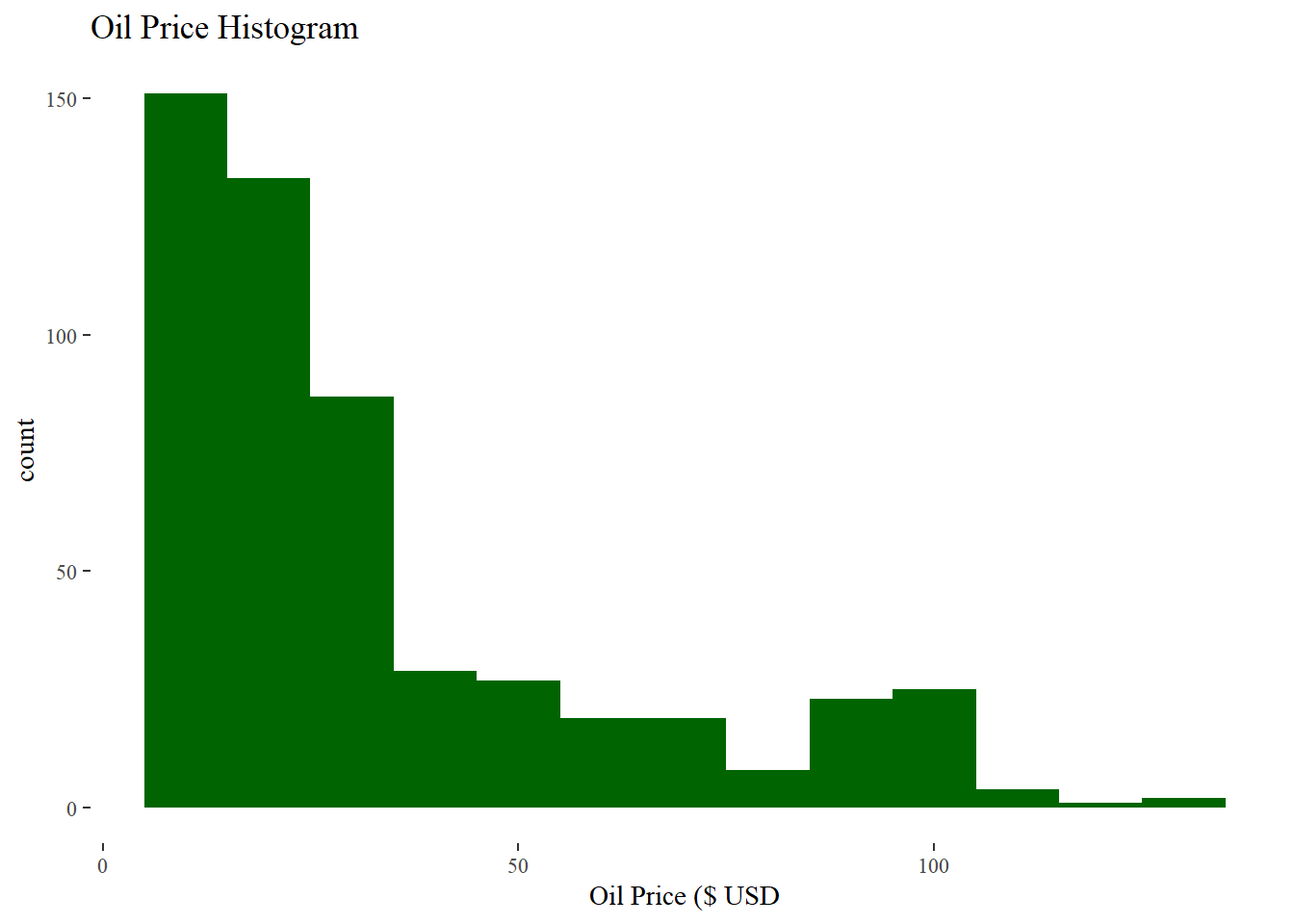


Figure 27 Oil prices histogram

The histogram of oil prices not corrected by inflation (Figure 26), shows an almost unimodal heavily right skewed distribution, with a mean of 33.28 USD, a median of 23.51 USD a minimum value of 3.89 USD a maximum value of 128 USD and an interquartile range of 32,18 USD.

**3.7.1.2 Monthly Active Rigs**

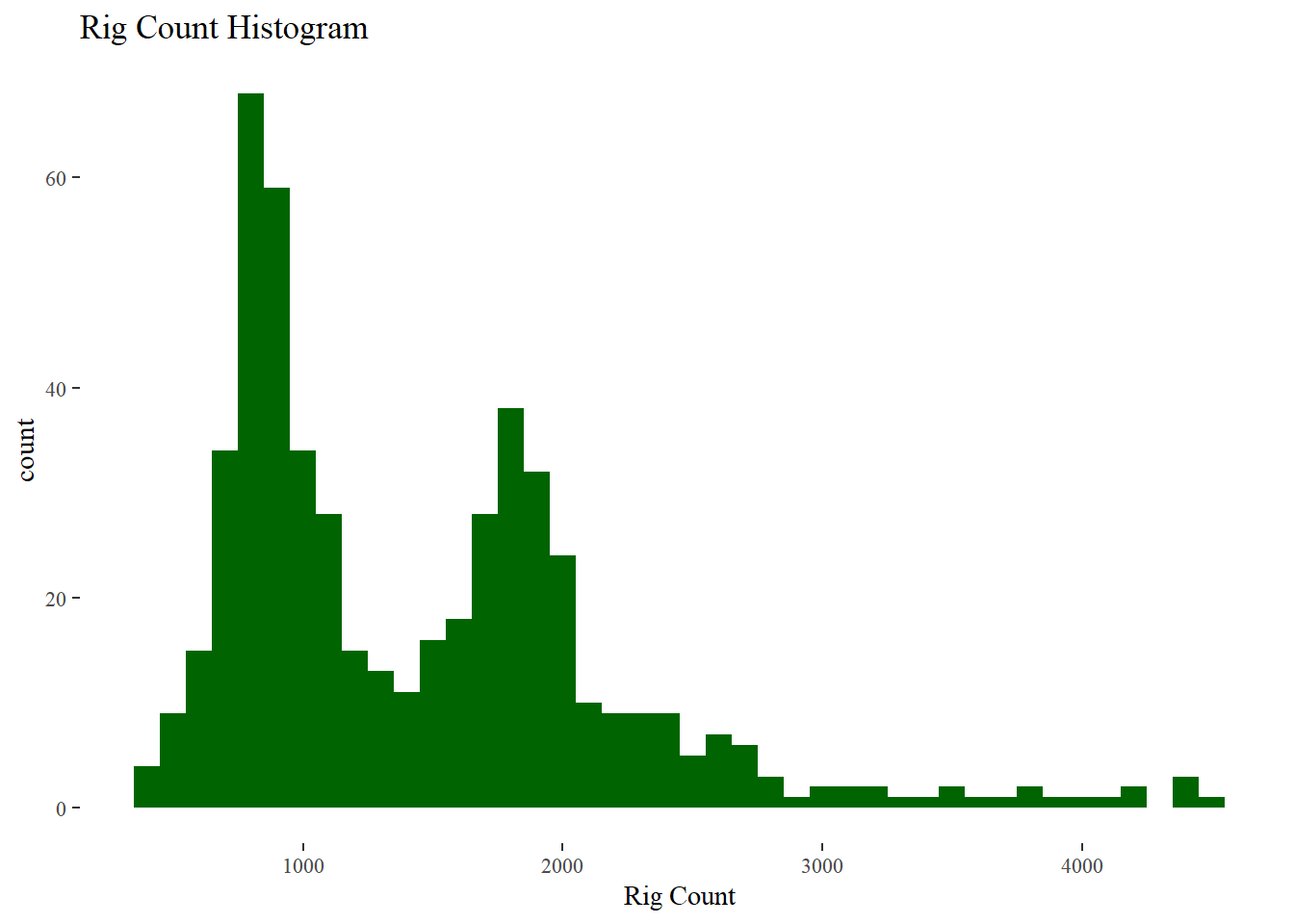


Figure 28 Active Rigs histogram

The histogram of monthly active rigs or rig count (Figure 27), shows bimodal heavily right skewed distribution, with a mean of 1435 rigs, a median of 1184.5 a minimum value of 407 rigs a maximum value of 4521 and an interquartile range of 994.7 rigs.

This bimodal pattern looks very interesting and sparks plenty of questions, one of the first ones is to see if weather or not these extreme right values correspond to specific dates? The industry of oil and gas and specially the exploration and production segments has evolved since the early 70’s, we might expect that perhaps efficiencies have increase since this date and that today the industry requires less deployed rigs for a similar amount of production. Another aspect to consider is the change of practices in the industry, on the early 70’s exploration via wildcats (exploration technique that consist on drilling a well with low geological or Petro-physics data or evidence of a reservoir just for the sake of checking if there is oil) wells was still a common practice, therefore the chances for dry wells were very high, today wildcats are the exception and even on exploration wells there is already sufficient evidence that maximizes the chances for promising oil and gas prospects, therefore we might expect less deployed rigs for similar exploration results compared with the past. All of these questions are going to be explored on further sections of this analysis.

**3.7.2 Pair plots analysis**

After studying the distribution of both of our response variables our next step will be two perform a pairwise plot analysis between our response variables and variables of each one of our previously defined groups. Pairwise plots have on their lower area scatter plots generated between all of our variables from a specific group and they result very helpful on studying the relationship between variables. On the upper area of the plots we can find the values of the Pearson correlation coefficients and on the diagonal of the plot, we have the density plot of each one of the group variables.

**3.7.2.1 Oil & Gas related variables**

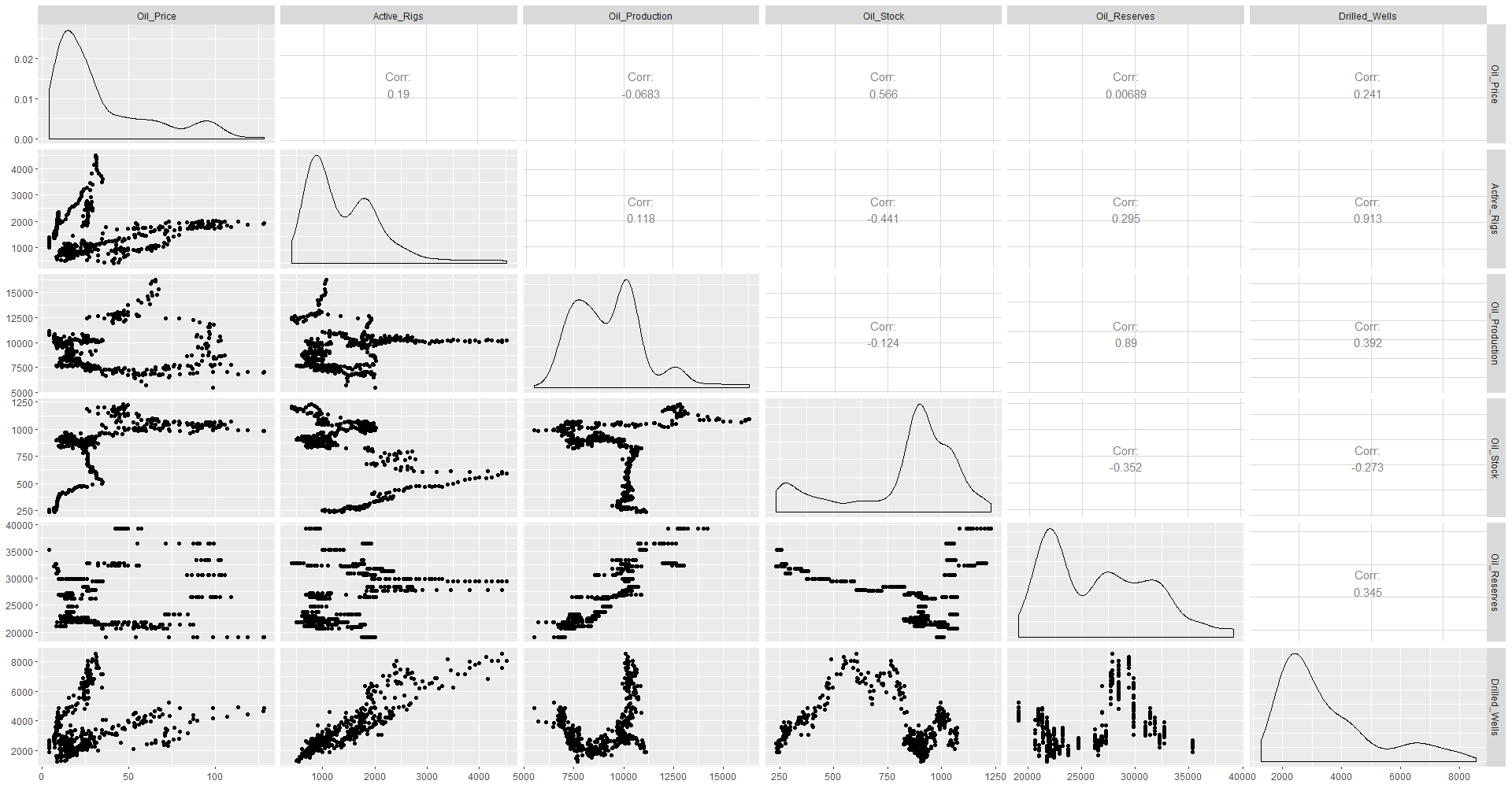


Figure 29 Oil & Gas group variables pairwise plot

The first thing that we notice from the pairs plot (Figure 28) is that oil prices and rig count have a positive not linear relationship, and it seems conformed by two different linear responses, Oil prices and oil production have an unknown relationship with a weak negative Pearson correlation coefficient, despite what common sense might suggest, active rigs and oil production don’t seem to have an strong positive relationship, we might expect that for more production we need more well and therefore more deployed rigs, but data suggest a different story. Oil prices show a medium positive relationship with oil stocks with a correlation coefficient of 0.57 such relationship is not shared with active rigs, in fact if there is a relationship is negative. Oil prices show no relationship with oil reserves at least not for US oil reserves but they do share a positive medium relationship with active rigs, such relationships seems to be logical since to increase reserves we need to conduct more exploration activities and more exploration activities require in general more deployed rigs, another ingesting relationship to highlight is the one shared between oil reserves and oil production, both have an strong positive linear relationship. Lastly, drilled wells shows a very strong positive linear relationship with active rigs, with a correlation coefficient of 0.93, this is completely logic since well can only be drilled by active drilling rigs, as a consequence almost all other variables will show a similar relationship with drilled wells to the one showed with active rigs.

**3.7.2.2 Economic related variables**

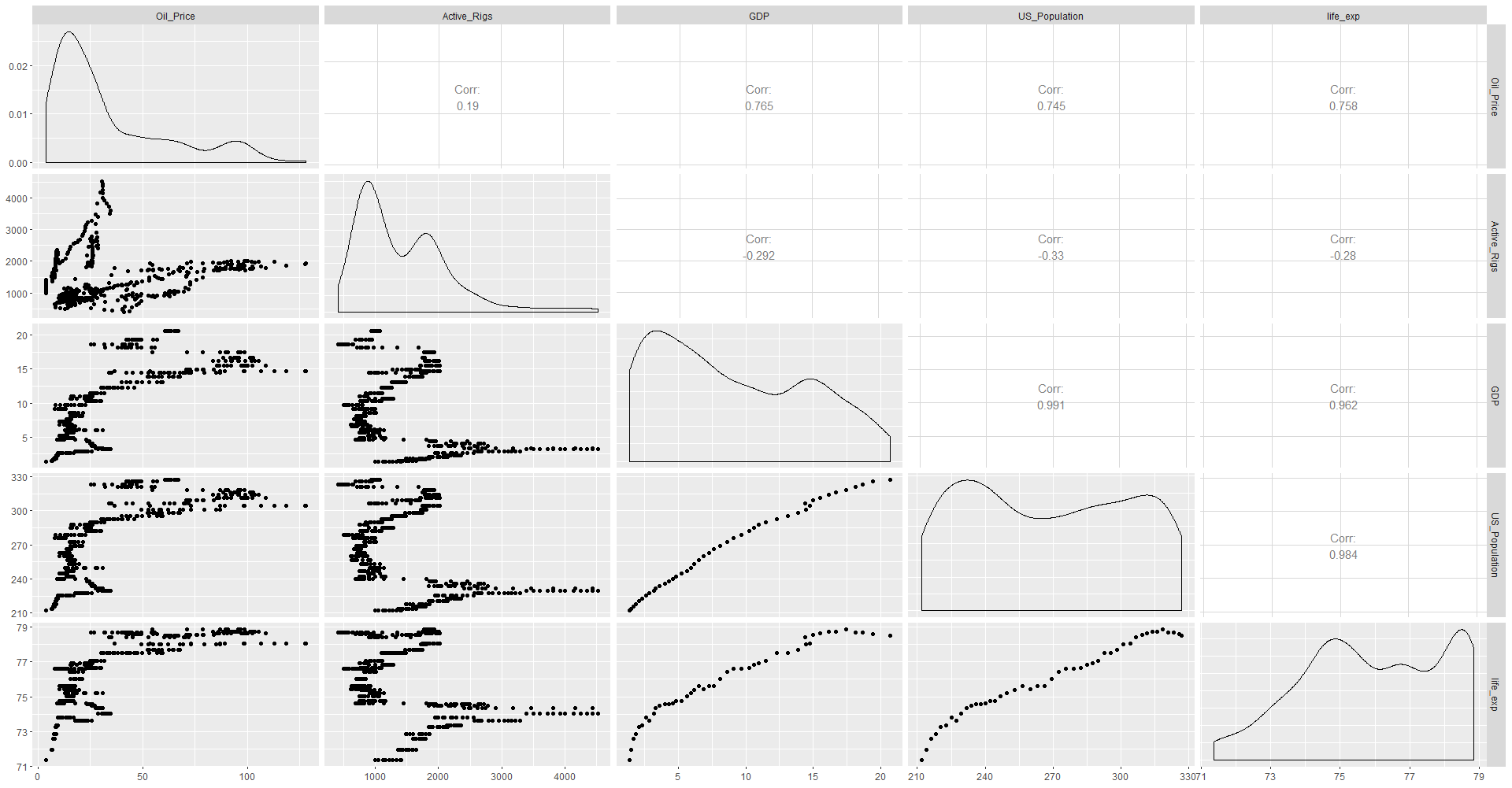


Figure 30 Economic group variables pairwise plot

The first thing we notice is the strong correlation that exist between the three economic variables (Figure 29), the three of them US GDP, US population and US life expectancy have a strong positive Pearson correlation coefficients around 0.9, there is one other particularity to highlight is the stagnation pattern that life expectancy shows at high values either of GDP and population, one might theorize and investigate on the many reasons behind this pattern, perhaps just means that income is not being translated into a better quality of life, or at least not in the majority of citizens. Oil prices share strong positive relationships with the three economic variables, with correlation coefficients in the order of 0.7. These relationships might be somehow confusing, since we expect that growing economies and populations require energy for doing so, and therefore they should benefit more from cheap energy prices rather than expensive ones, but we should also remember that the US is at the same time an oil and gas producer country, specially on recent years and also the fact that some of the most important and powerful oil and gas companies are American ones which benefit tremendously from high oil prices and can contribute to the American economy. On the other hand, the relationship seems to be inverted for active rigs, that shows a weak Pearson coefficient of -0.3 with the economic variables, this means that a growing strong economy have fewer active rigs than a weak one. At this point we can only theorize but perhaps this is more related to political moves rather than economic ones, and the US decides to promote internal reservoir exploiting with an underperforming economy, in order to promote consumption and perhaps benefit from lower salaries.

**3.7.2.3 Energy related variables**

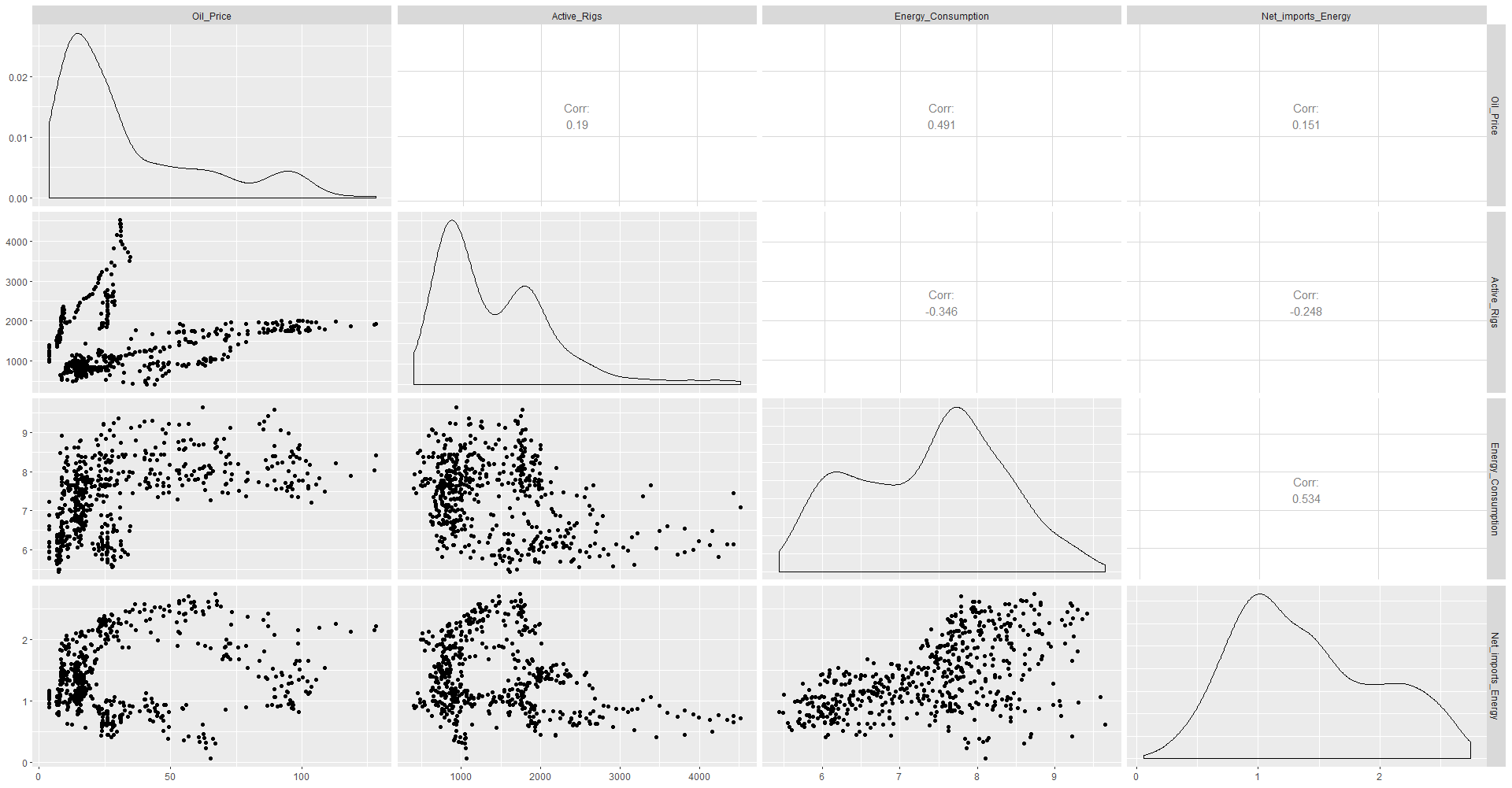


Figure 31 Energy group variables pairwise plot

Based on what the pairs plot of energy variables (Figure 30) shows, it seems that energy consumption and net energy imports shows a medium linear relationship, with a Pearson correlation coefficient of 0.5; countries have limited resources for energy production and as consumption increases the excess demand have to be supplied somehow, and as a consequence energy has to be imported into the country. Oil prices shows a medium positive relationship with energy consumption, such relationship is fairly logical since it follows the basic laws of supply and demand, the more demand of a product the higher the price would be. In regard to the net imported energy, oil prices don’t show any apparent relationship. Active rigs show a weak negative relationship with both of the energy variables.

**3.7.2.4 Environmental related variables**

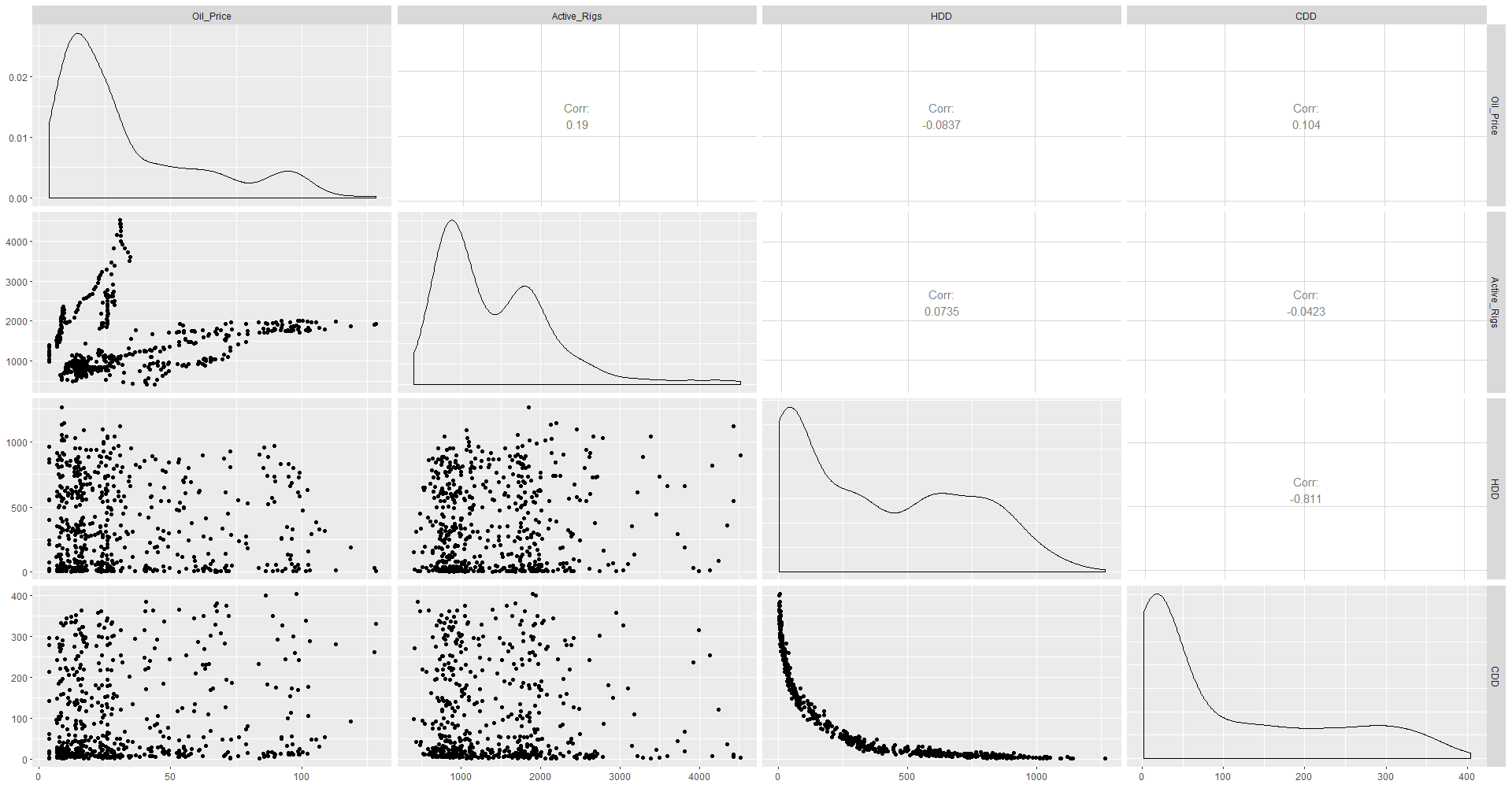


Figure 32 Environment group variables pairwise plot

The environment pairwise plot (Figure 31) shows that there is an strong negative relationship between HDD and CDD, this is fairly obvious since these two variables conform a complete set and by definition one is the complement of the other one. Neither oil prices nor active rigs have signs of any apparent relationship with the environment variables.

**3.7.2.5 Political related variables**

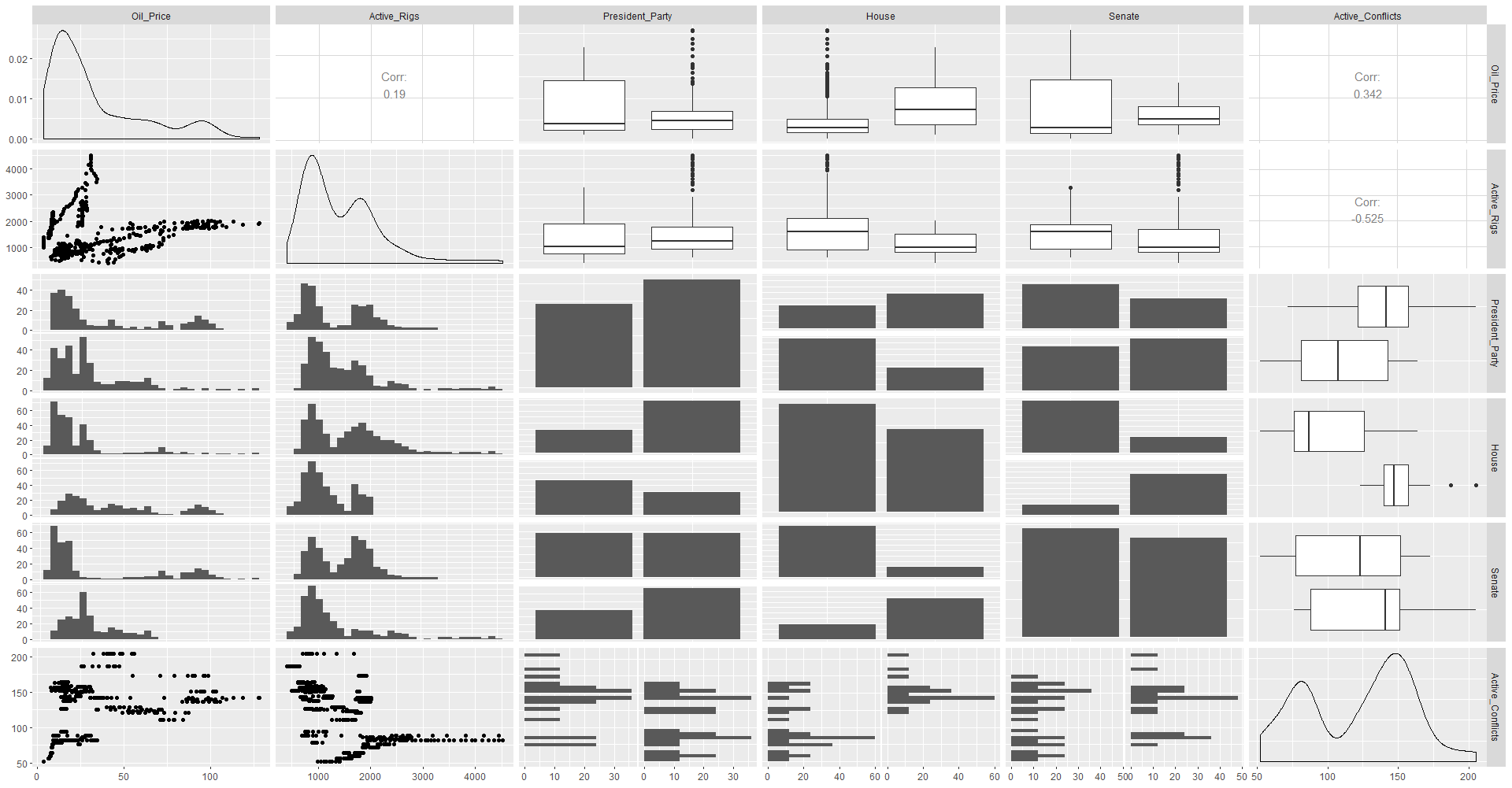


Figure 33 Politic group variables pairwise plot

The first thing that can be noticed by looking at the plot (Figure 32) is the apparent difference in oil prices whit difference dominant political parties at the three levels of government of the US. The same difference seems to exist on active rigs, although it will be confirmed or denied on later sections of this thesis work by using the appropriate statistical tests, another interesting remark is the medium relationship that both response variables show with active conflicts, as it seems war might contribute to higher oil prices and more deployed rigs.

**3.7.3 K-Means Cluster analysis**

The bimodal distribution of active rigs and the relationship that it shows with oil prices gave the impression that at certain point the oil and gas industry on the US experienced a dramatic change in its practices, since it seems it has two different lines of behavior, for those reasons our next step will be to conduct a cluster analysis on all of our variables using the K-means algorithm, to help us to evaluate the appropriate number of clusters we will use and elbow or scree plot in conjunction with a silhouette plot analysis. To facilitate the analysis and the clustering algorithm all variables including the categorical ones will be transformed to numerical and later on standardized (deduct the mean and divide by the standard deviation) this transformation will ensure that all distances are comparable, without the influence of the units or scale of the variables, after this we will run the k-mean clustering algorithm for various number of centroids and we will evaluate the total of the within squared distance of the points to find the “elbow” point, the values can be found on Table 4.

|  |  |
| --- | --- |
| **k** | **tot\_withinss** |
| 1 | 308 |
| 2 | 187.29814 |
| 3 | 124.30791 |
| 4 | 108.16505 |
| 5 | 85.18374 |
| 6 | 47.37352 |
| 7 | 43.09698 |
| 8 | 29.8266 |
| 9 | 26.98528 |
| 10 | 22.92257 |

Table 4 Centroids vs Scree plot values

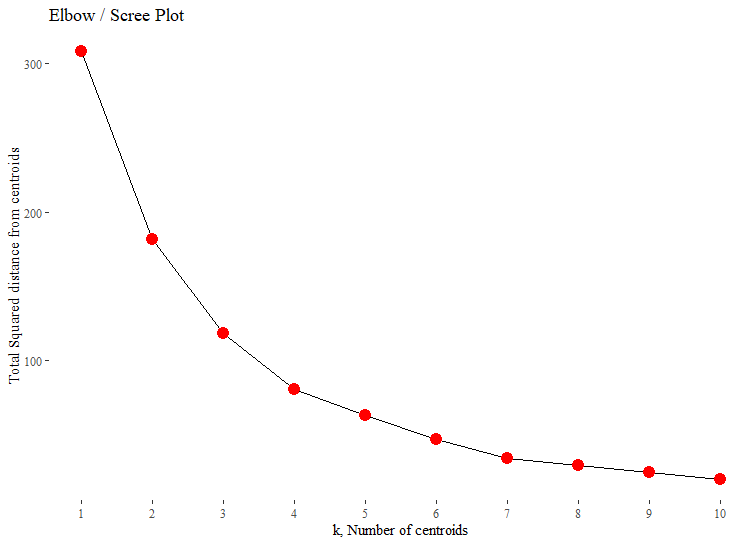


Figure 34 Scree or Elbow plot

From the analysis elbow plot (Figure 34),the identified k value for the “elbow” is two, since it shows the steepest change in total within distance. Our next step is to create a silhouette analysis and determine the ideal number of clusters via this method.

|  |  |
| --- | --- |
| **k** | **sil\_width** |
| 2 | 0.3790503 |
| 3 | 0.3665515 |
| 4 | 0.4044439 |
| 5 | 0.4094645 |
| 6 | 0.4303218 |
| 7 | 0.4655798 |
| 8 | 0.443771 |
| 9 | 0.4207146 |
| 10 | 0.4203871 |

Table 5 Silhouette analysis

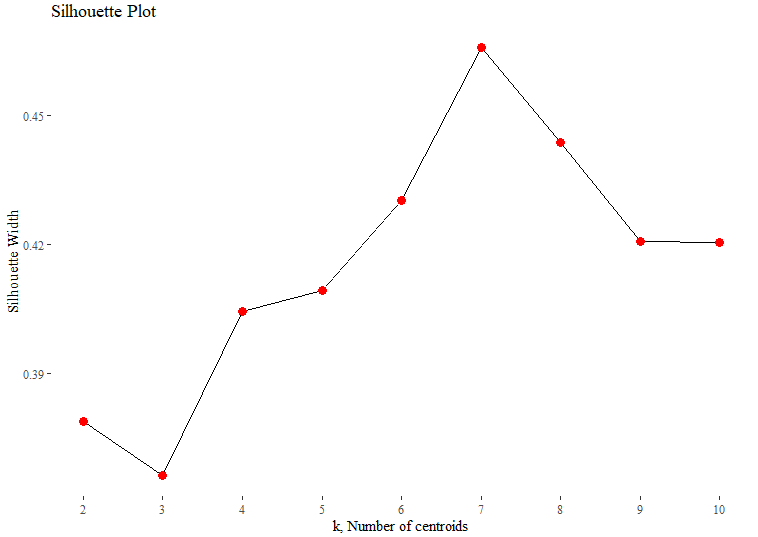


Figure 35 Silhouette plot

Based on the numerical results (Table 5) of the silhouette plot (Figure 35) it appears that the idea value corresponds its seven, however for the sake of practicality we will follow the output of the elbow analysis ,therefore we will select two as the number of clusters for our analysis.

The next step is to re-run the k-means model but using our selected number of clusters and make an analysis of the values corresponding to each one of the clusters. For this thesis work we will summarize the values of the clusters based on the median value of each of the selected variables , and we will only select the oil and gas related variables plus the year, the idea is to confirm our soupcons regarding the changes that the oil and gas may have suffered or experienced.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **cluster** | **Year** | **Active\_Rigs** | **Oil\_Price** | **Oil\_Reserves** | **Oil\_Production** | **Oil\_Stock** |
| 1 | 1983 | 1631 | 15.4 | 28416 | 10136 | 681 |
| 2 | 2006 | 1050 | 41.7 | 22398 | 7760 | 991 |

Figure 36 K-means cluster median values

As it seems, the oil and gas industry In the US changed dramatically after mid 2000’s specially regarding the number of active rigs. We will look with more detail in a graphical way by looking at the progression of the variables with time.

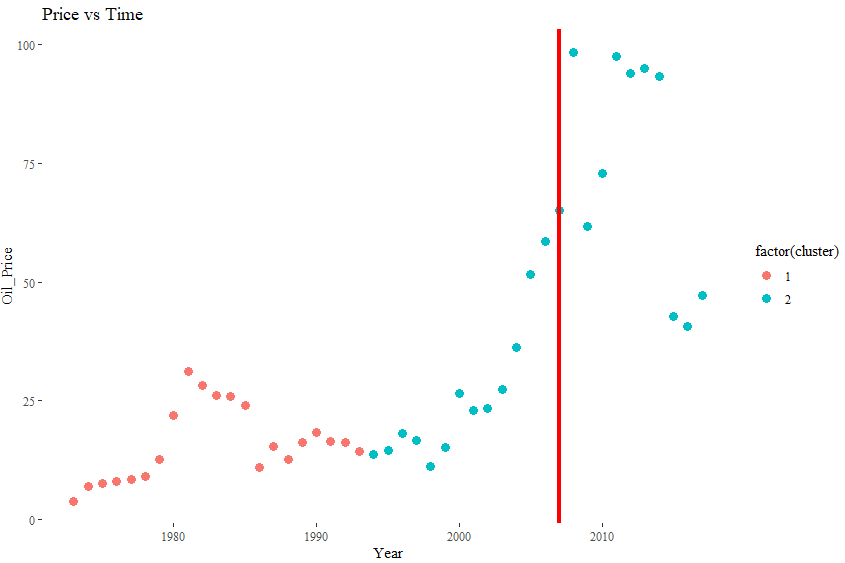


Figure 37 Clustered Oil prices time progression

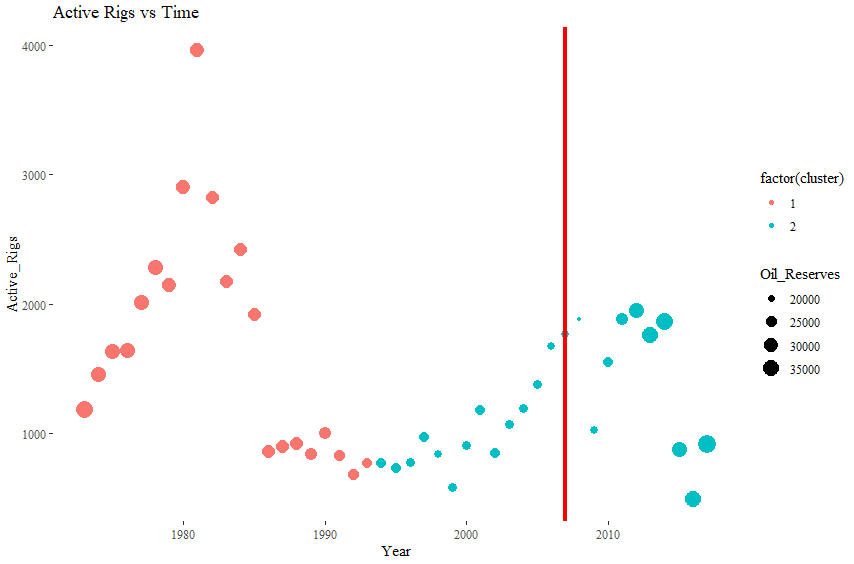


Figure 38 Active rigs clustered time progression

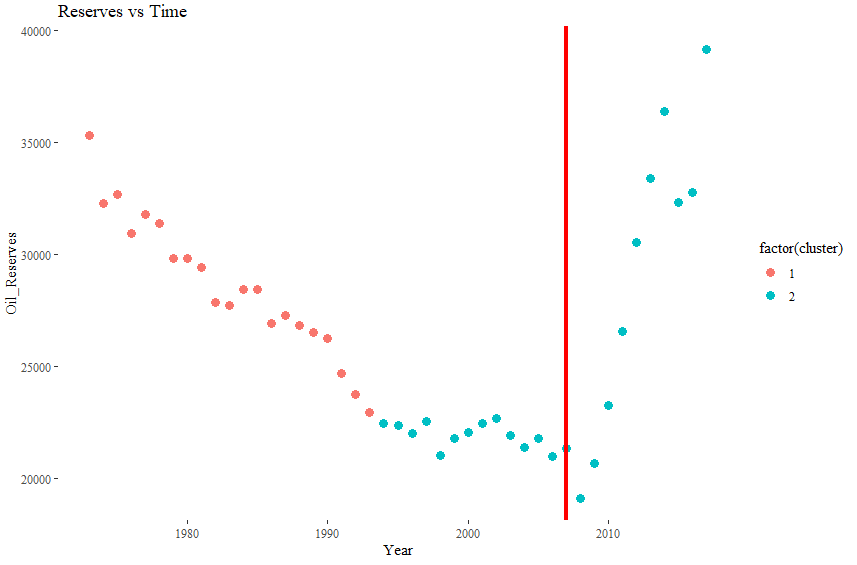


Figure 39 Oil reserves clustered time progression

With the insight gained by the cluster analysis we will like to look again at the oil prices vs active rigs scatter plot (Figure 40).

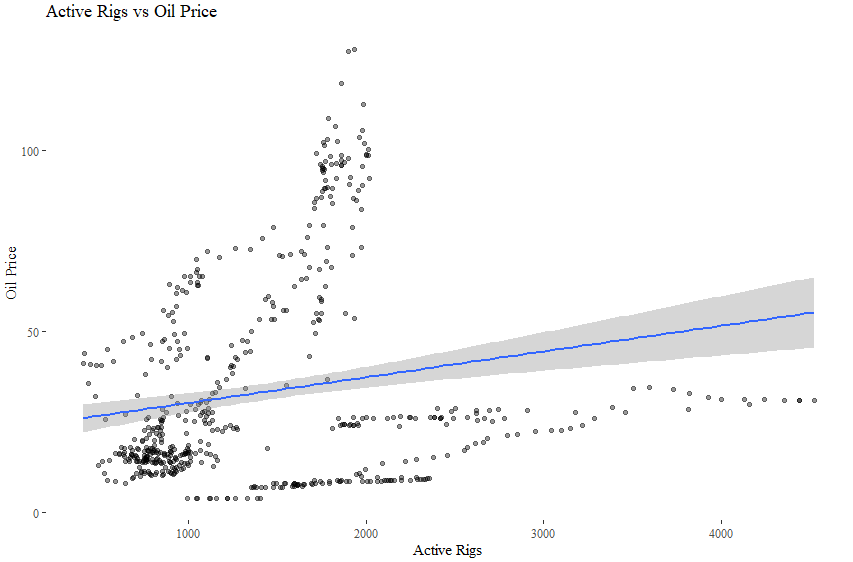


Figure 40 Oil prices vs Active rigs

Confirmed by the regression line (blue line) we notice that there is indeed a positive relationship, however data seems to be divided into two different linear trends, to look at these two variables with more detail we will first look at the year evolution of both of them but we will first scaled them, since active rigs at least one order of magnitude bigger than oil prices, therefore making it difficult to notice any significant pattern by comparing both curves .

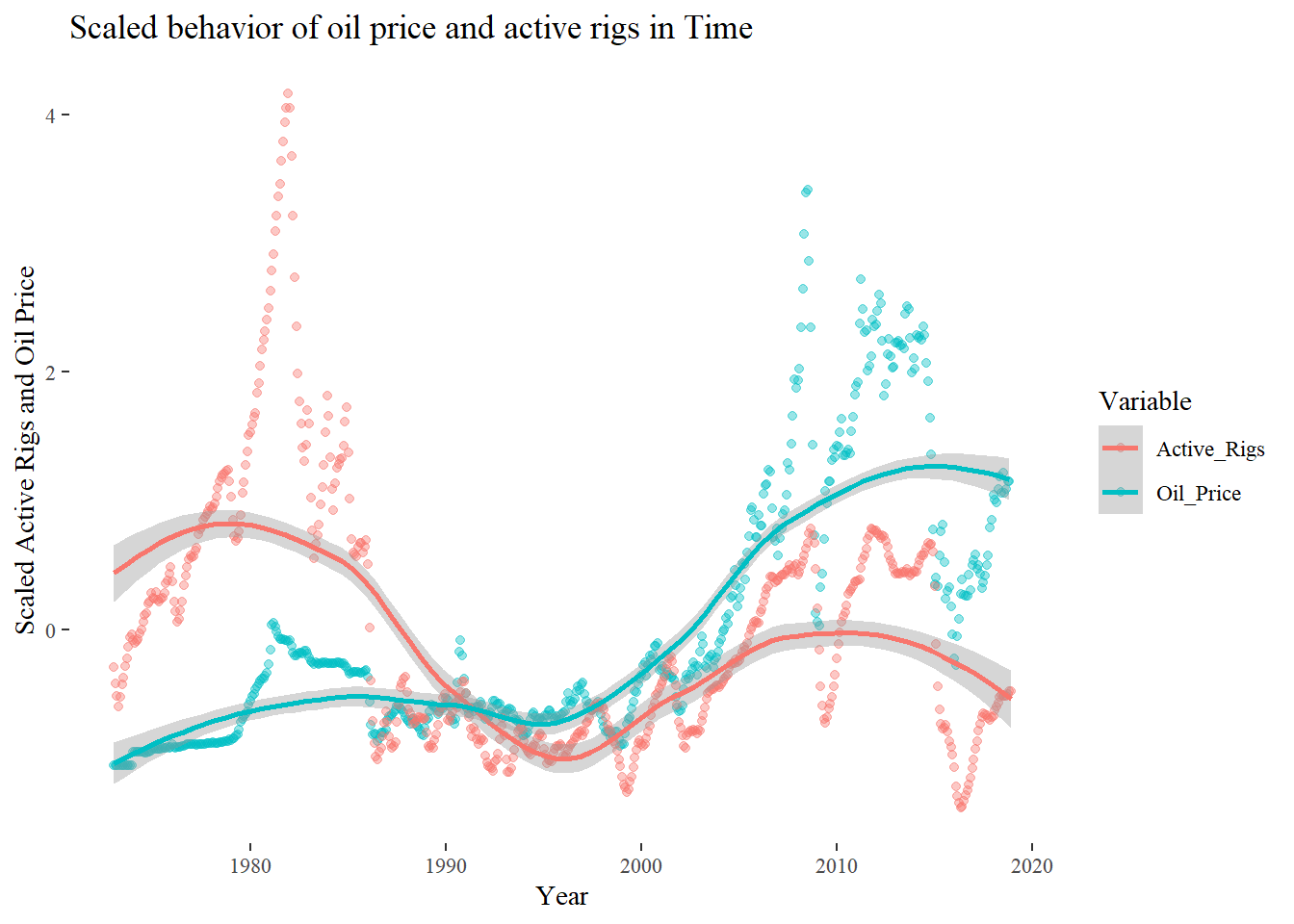


Figure 41 Time evolution of scaled values of active rigs and oil prices

At this new scale, we can notice from the graph (Figure 41) that both curves move together, and in fact are highly correlated. Lastly, we will produce again a scatter plot of oil prices and active rigs, but we will include a cluster division and see if weather or not, the k-mean algorithm was able to categorize the data into two different linear trends.

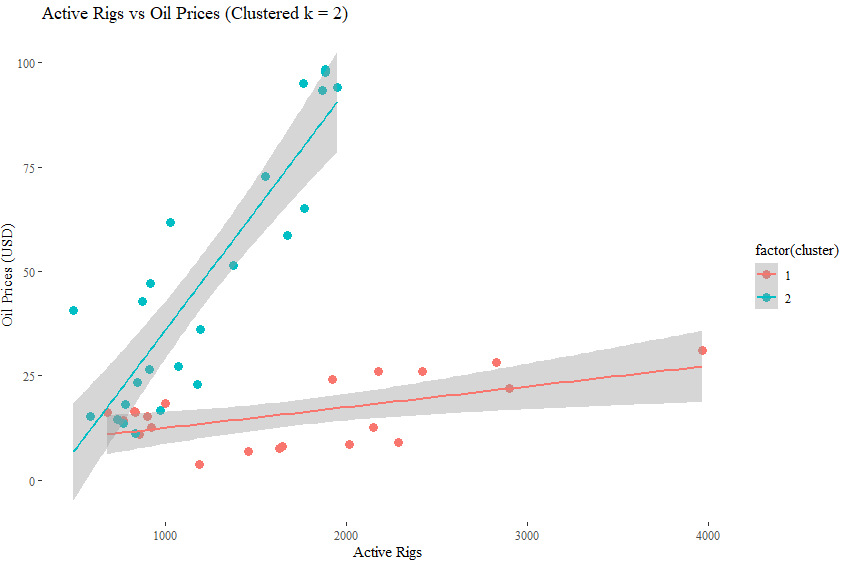


Figure 42 Active Rigs vs Oil prices scatter plot (clustered)

The k-mean clustering scatter plot of active rigs and oil prices (Figure 42) confirms that indeed there are two linear positive trends on the relationship between oil prices and active rigs, a modern and a past one, with very different slopes, on which modern times are characterized by an industry that responds more aggressively to oil prices in terms of active rigs.

**3.8 Feature Engineering**

On the previous section we have seen how the chosen explanatory variables area already generating interesting questions that we will answer properly by using statistical inference on future sections of this work and that are helping us to develop insight for the proper model’s elaboration. However in their current form they might not be able to transfer the complete amount of information that we will like, that is why we will go through the process of engineer them so we can maximize their value.

**3.8.1 Seasonality**

It is fair and part of common sense to assume that perhaps oil price and active rigs values might change due to seasonality; normally (global warming is changing the “normal” behavior of weather) winters have low temperature which require more energy, and more energy requires more fuels such as oil, and a higher demand of oil will increase prices and this will motivate more service companies and operators to drill more wells and therefore deploy more drilling rigs. In contrast we should expect that in summer the high temperatures will decrease the energy consumption, however with the peak of heat that US might experience on summer, the usage of air conditioning might be increased so as the energy consumption. To properly analyze these hypotheses, we will create a new variable by to transforming the month variable into season categorical variable, to do so we will use the season dates of the US (Table 6) and after we will create boxplot’s for both of our response variables with season.

|  |  |
| --- | --- |
| **Season** | **Dates** |
| Spring | March 1 to May 31 |
| Summer | June 1 to August 31 |
| Autumn | September 1 to November 30 |
| Winter | December 1 to February 28 |

Table 6 US season dates

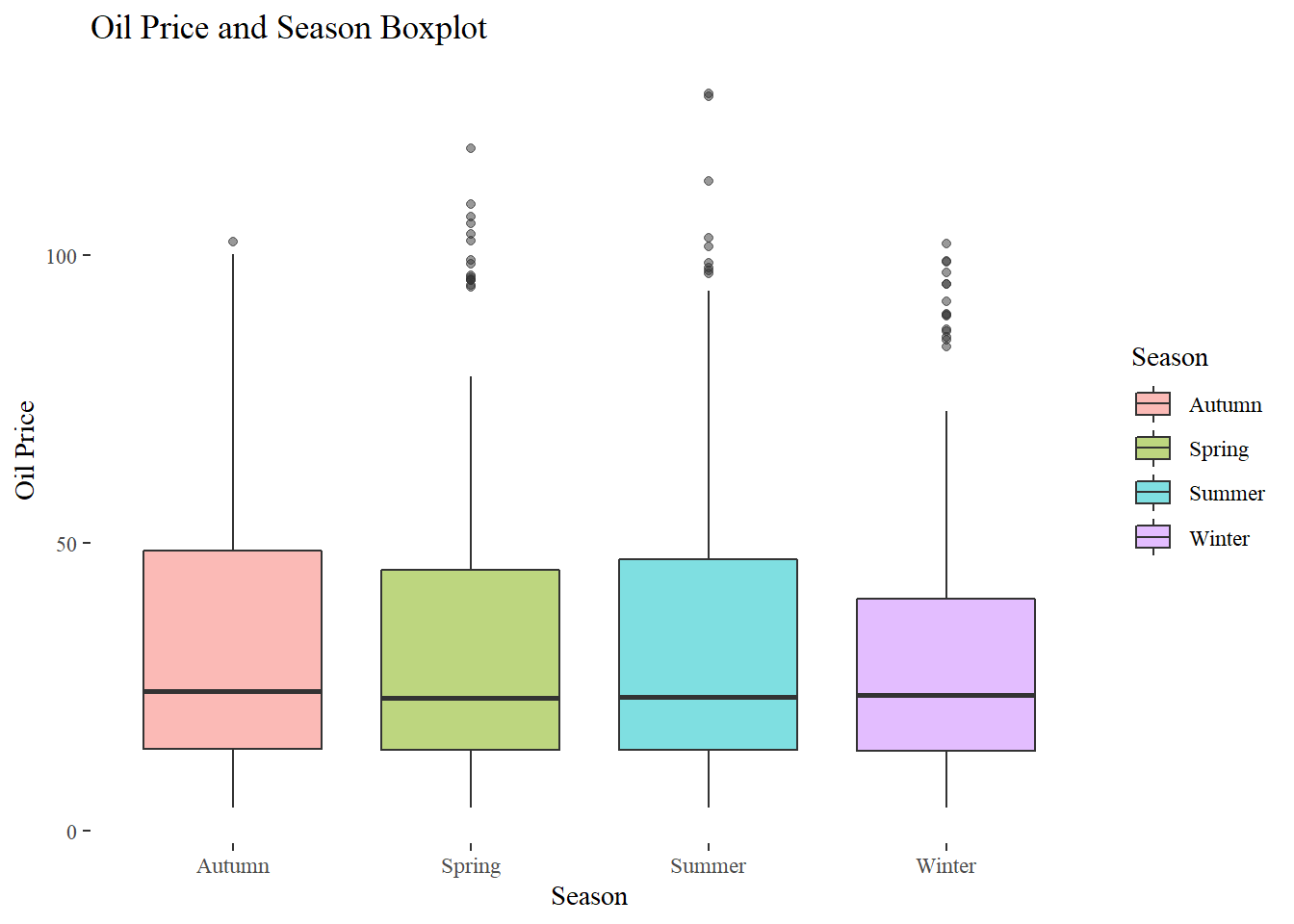


Figure 43 Oil prices and seasonality boxplot

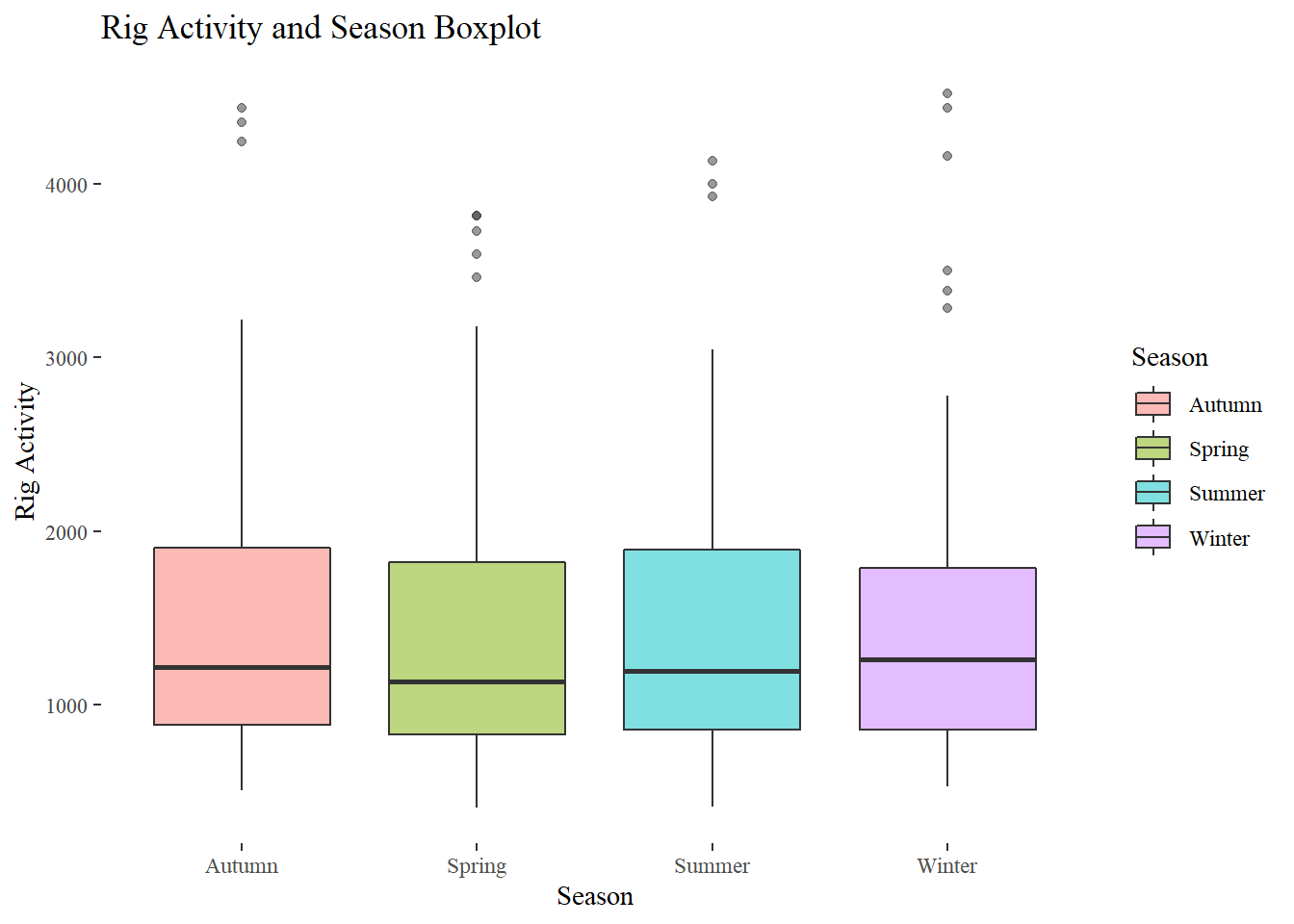


Figure 44 Active rigs and seasonality boxplot

After looking with detail a both of our seasonality graphs (Figure 43, Figure 44) it seems that there is not an strong or at least noticeable difference neither with oil prices or active rigs. Although graphs such as the ones we just analyzed have incredible communication powers, we cannot discard a significant difference until we perform the adequate statistical tests to the data.

**3.8.2 Politics**

The initial pairwise plots part of the exploratory analysis on oil prices and active rigs , in regard on US politic dynamics, showed that indeed in seems to be a difference from which party is governing at different political institutions of the US government, specifically the US Senate, the US congress or house and the executive branch represented by the US president. This can be seen more clearly on a specific boxplot form in Figure 45 and Figure 46.

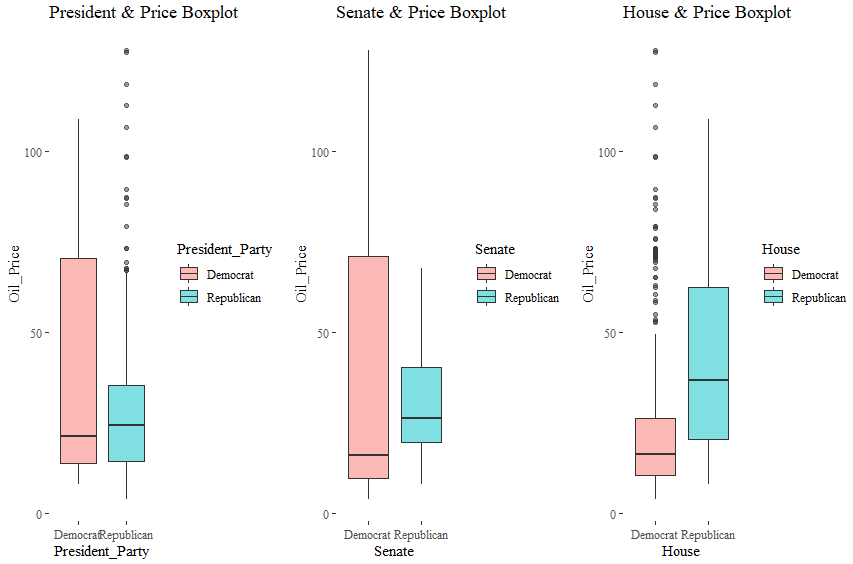


Figure 45 Oil price and politics boxplot

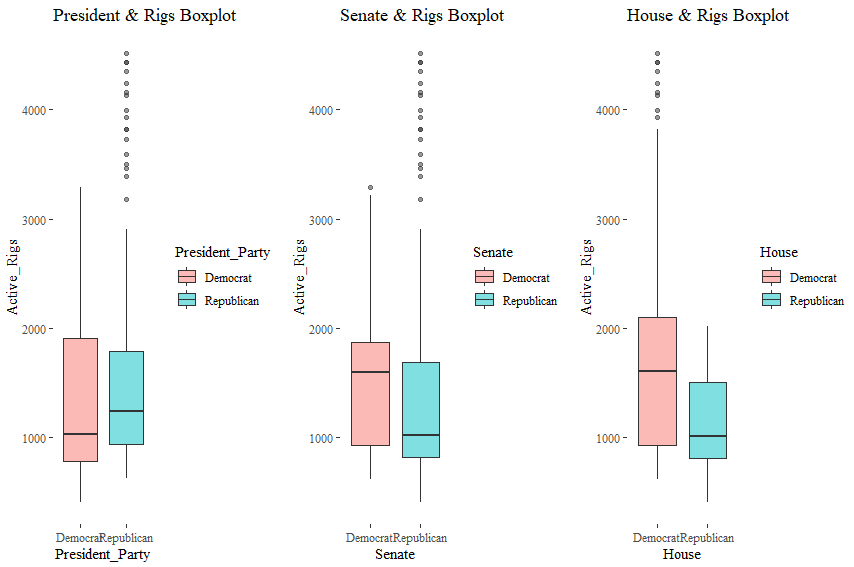


Figure 46 Active rigs and politics boxplot

But despite these graphs being quite informative, we might still need some additional information to understand further these behaviors.

Division of power is primordial for the preservation of democracy, that is why the on the US we have these three different governmental powers, if the US president doesn’t count with the majority of the house and senate he will have an opposition that will force him to negotiate and therefore to sacrifice certain aspects of its envisioned government and policies, the president’s power to rule is limited by the house and senate, the same situation applies to the other two powers. On the contrary If the president counts with the support and majority of both house and senate, this gives him practically a free pass to govern as he thinks best, of course under the constitution and the law. Knowing which party is ruling and individual government power, doesn’t necessary means that that party is absolutely governing, what we are interested in, is on the combination of the three powers, we are interested in determining if there is a difference in the oil prices or active rigs, on scenarios of total or absolute government (The same political party holds the majority on the three government power levels) in comparison with a mixed government (At least one of the government power levels is on the hand of an opposing power, limiting ruling powers of its opposition), and we are also interested in knowing if on the case of absolute government, which party is ruling and how does this influence oil prices and active rigs. To properly analyze these questions we will create two new categorical variables, on that determines if the US government is on a total or mixed government and a second one that will tell us if the government is a total government which party is ruling it, once we have both of the variables we will create the boxplots and analyze if there are noticeable differences for both of our response variables.

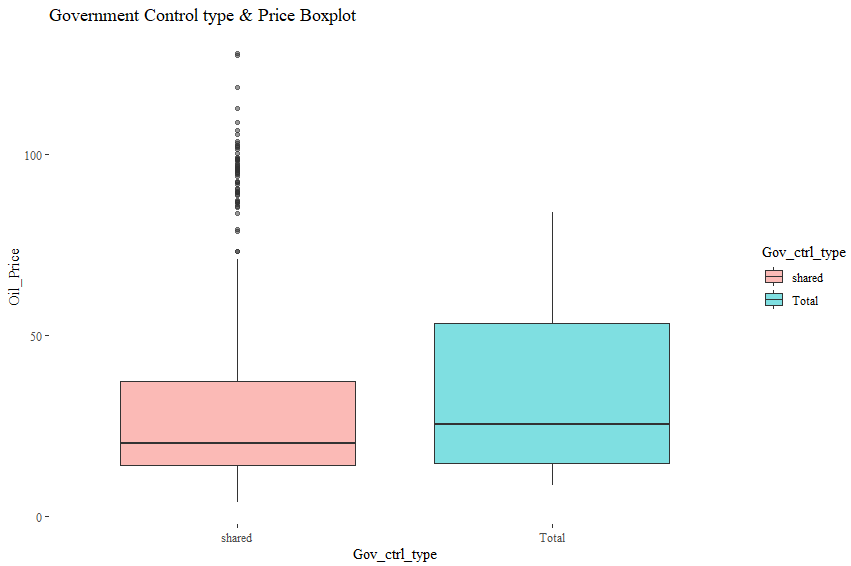


Figure 47 Oil price and government control type boxplot

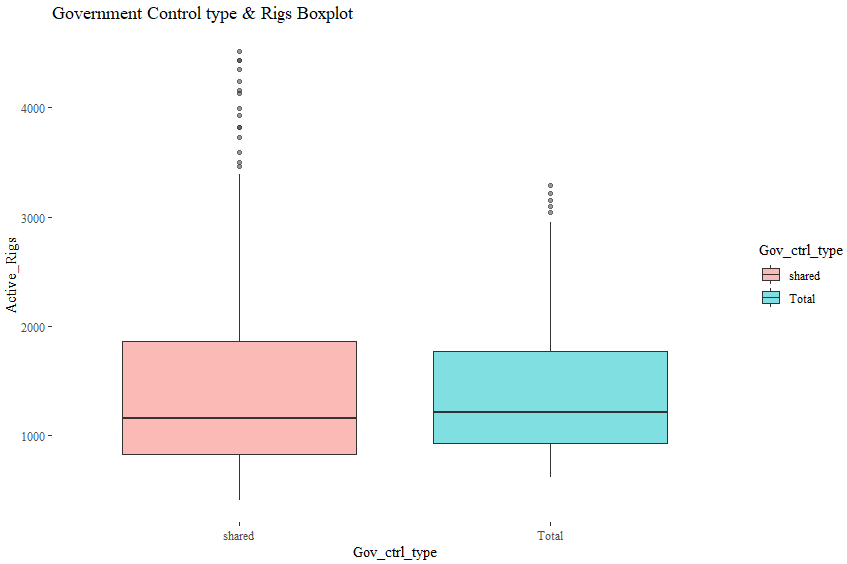


Figure 48 Active rigs and government control type boxplot

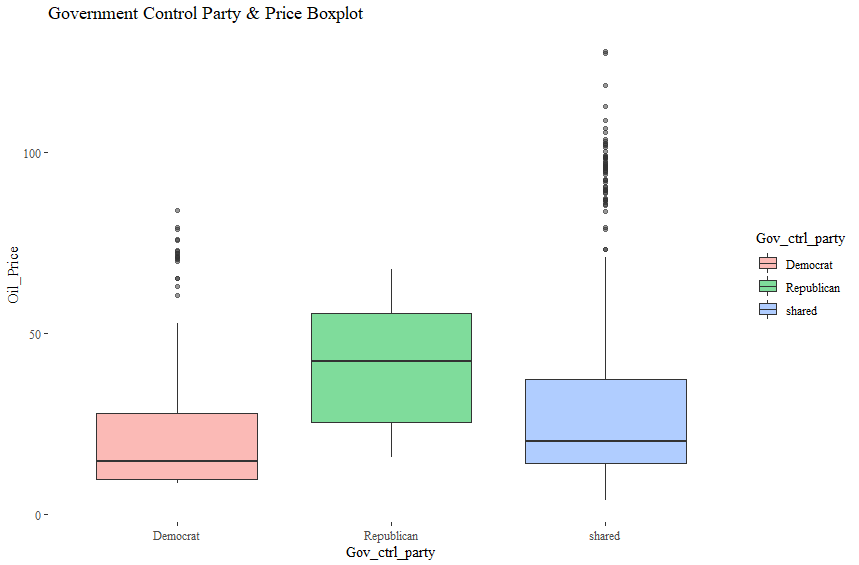


Figure 49 Oil prices and government control party boxplot

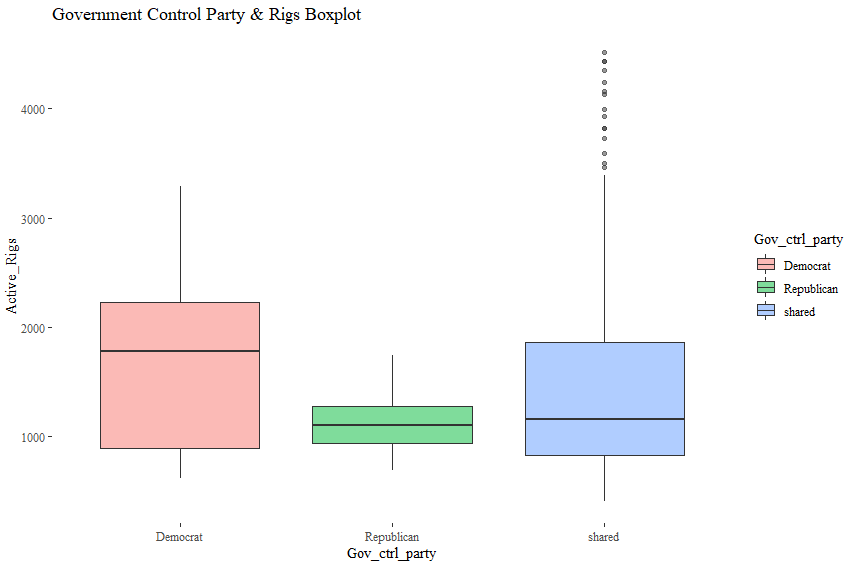


Figure 50 Active rigs and government control party boxplot

Looking at the first two boxplots Figure 47 and Figure 48 for both of our response variables, at first glance doesn’t seems to be a remarkable difference for the government control type, but when we split the government control type and assign a political party on the cases of total control (Figure 49 and Figure 50), we immediately notice that there is indeed a big difference for both of our response variables, which is a classic example of the Simpson’s paradox, on which the split or the inclusion of a new variable reverse a previous expected behavior. a total government of the republican party has experience lower levels of active rigs and higher oil prices in comparison with a total government of democrat and a mixed government. It is important to mention that currently the observed difference has only a qualitative value and can only be numerically confirmed with the use of an adequate statistical test.

**3.8.3 Economics**

During the exploratory data analysis and the pairwise plot analysis we noticed how economic factors such as GDP, population and life expectancy showed positive relationship with oil prices and a weak negative relationship with active rigs, however for active rigs this initial interpretation might not be conclusive since it showed a dual behavior similar to what we saw on the initial oil price and active rigs scatter plot (Figure 40), to disentangle this strange relationship we used a cluster analysis which showed a clear division between a modern and antiquate oil industry, we will conduct the same analysis for GDP a determine if weather or not we find a similar pattern (Figure 51).

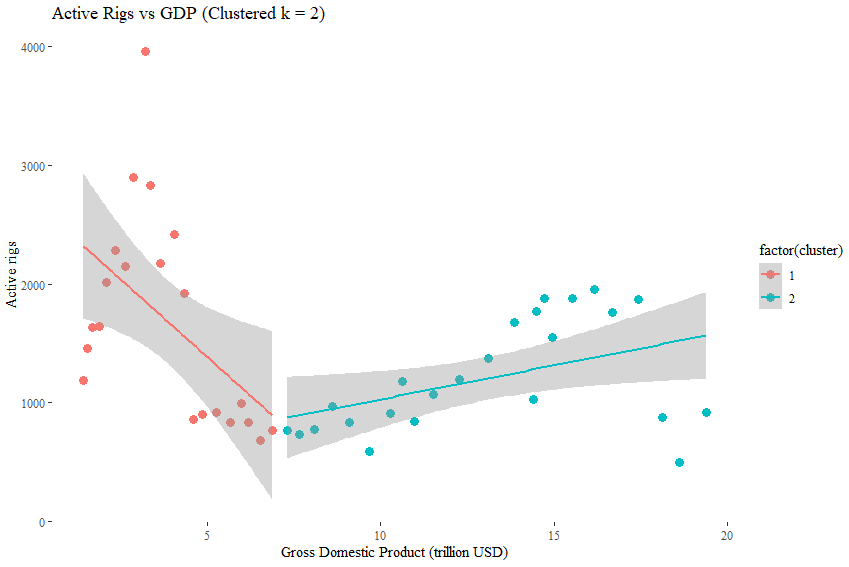


Figure 51 Active rigs vs GDP (clustered)

The old cluster, cluster number 1, shows a negative relationship of active rigs and GDP but the modern cluster, cluster number two, which corresponds with the appearance of shale oil and gas reservoirs on US soil, shows a completely opposite behavior.

We would like to explore further the influence of economy on our response variables, especially on periods of recession and high economic development, in order to do so we will create a new categorical variable that highlight periods of recession on the US economy (Table 7).

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Beginning** | **End** | **Short Description** |
| 1973 oil crisis | Nov-73 | Mar-75 | The 1973 oil crisis, a quadrupling of oil prices by OPEC. |
| 1980 recession | Jan-80 | Jul-80 | Federal Reserve, raised interest rates dramatically. |
| 1981–1982 recession | Jul-81 | Nov-82 | The Iranian Revolution & Tight monetary policy in the US. |
| Early 1990s recession | Jul-90 | Mar-91 | Federal Reserve raising interest & 1990 oil price shock. |
| Early 2000s dot-com | Mar-01 | Nov-01 | Collapse of speculative dot-com bubble September 11th attacks. |
| Great Recession | Dec-07 | Jun-09 | The subprime mortgage crisis, collapse of US housing bubble. |

Table 7 US Economic recessions

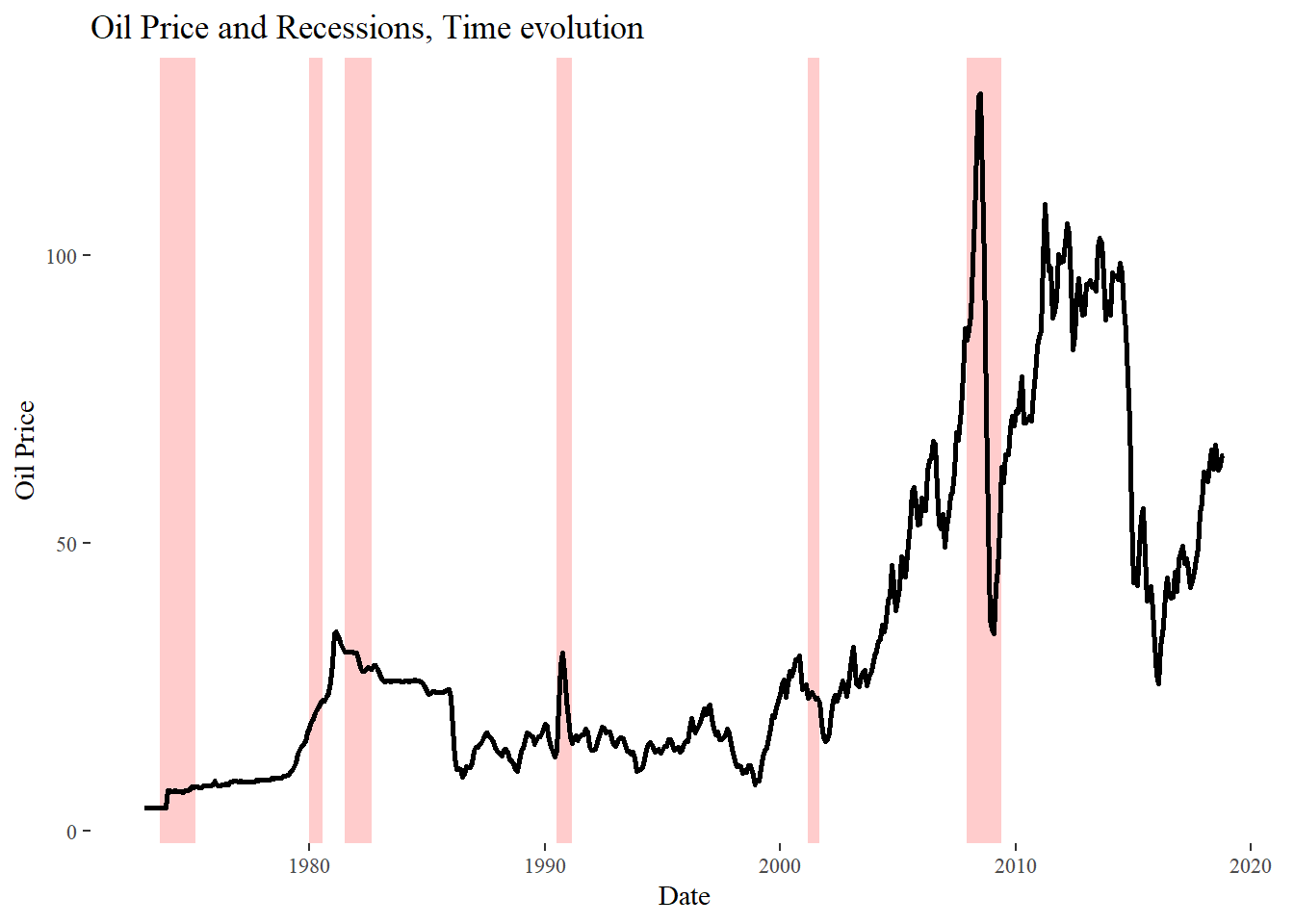


Figure 52 Oil prices and US economy recessions

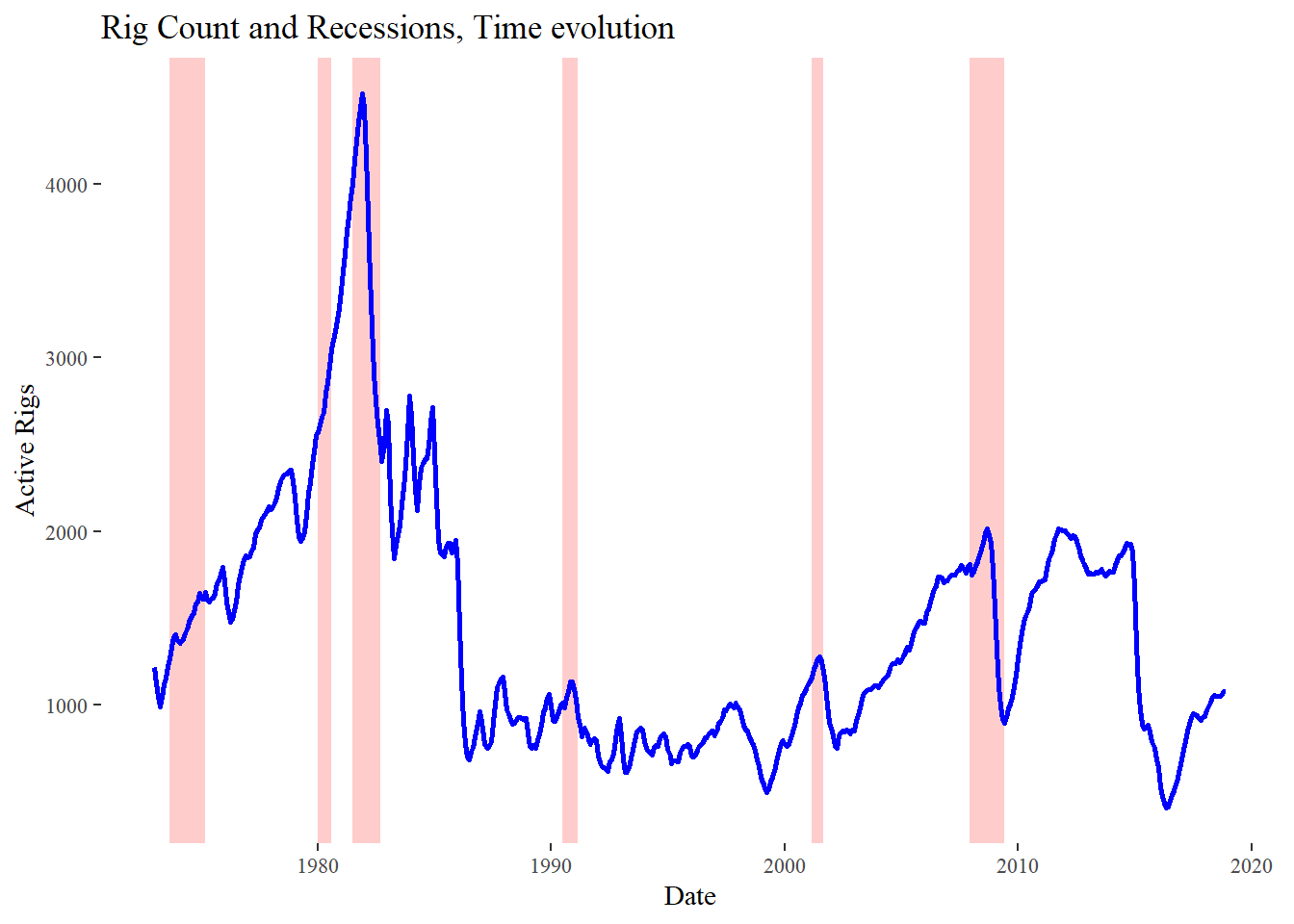


Figure 53 Active rigs and US economy recessions

Looking at the history of oil prices and active rigs including recession periods (Figure 52 and Figure 53), it seems interesting to notice that almost all economic recession are associated with peaks on oil prices and active rigs.



Figure 54 Oil price and economy recessions boxplot

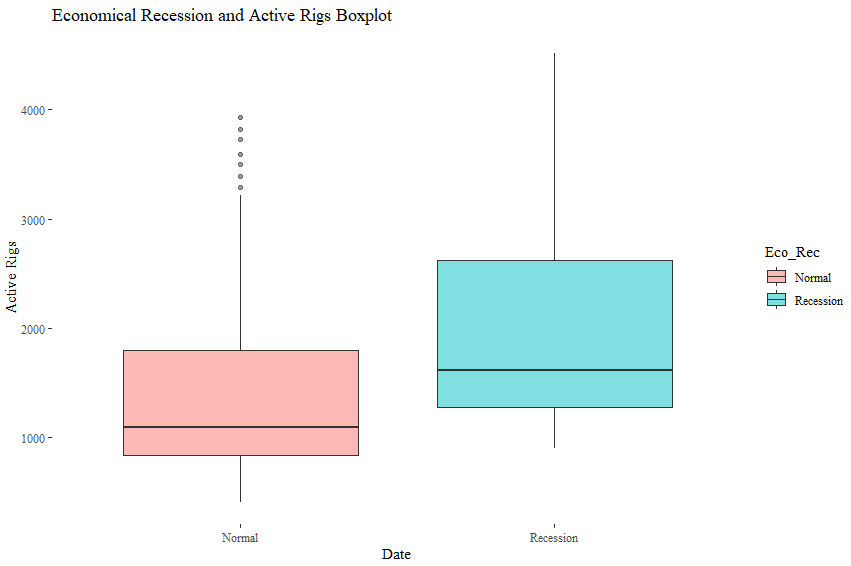


Figure 55 Active Rigs and economic recessions boxplot

By looking at the boxplots of price, rigs and economic recessions (Figure 54 and Figure 55), It seems that there is not a remarkable difference in oil prices and economic recessions, however for active rigs there is a noticeable difference.

**3.9 Feature confirmation**

Thanks to the insight gained from the exploratory and the feature engineering analysis we can move further and use this knowledge to confirm if weather or not we should include all of the variables the forecasting models or if we should drop any of them. To do so we will use the “Boruta” feature confirmation algorithm that uses the same principles of the random forest algorithm to calculate the variable importance on the variance explanation of the response variable. The results of Boruta for both of the response variables can be found on the tables (Table 8 and Table 9) and graphs (Figure 58 and Figure 59) below.

For oil prices Boruta confirmed all variables with the exception of season, for this response variable the explanatory variable with the highest importance is GDP. For active rigs Boruta confirmed all variables with only season as tentative, for this response variable the explanatory variable with the highest importance was oil prices followed by oil stock.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Median Impact** | **Decision** |
| GDP | 13.583942 | Confirmed |
| US\_Population | 13.207721 | Confirmed |
| Year | 13.116314 | Confirmed |
| Active\_Rigs | 12.602486 | Confirmed |
| Oil\_Stock | 12.566571 | Confirmed |
| life\_exp | 10.985583 | Confirmed |
| Senate | 10.465949 | Confirmed |
| Active\_Conflicts | 8.949173 | Confirmed |
| Net\_imports\_Energy | 8.597149 | Confirmed |
| Oil\_Reserves | 8.522154 | Confirmed |
| Oil\_Production | 8.50569 | Confirmed |
| Month | 6.678523 | Tentative |
| Energy\_Consumption | 6.076158 | Confirmed |
| CDD | 5.334423 | Confirmed |
| President\_Party | 5.082532 | Confirmed |
| House | 4.744941 | Confirmed |
| Eco\_Rec | 4.201994 | Confirmed |
| HDD | 2.801193 | Confirmed |
| Season | 1.878089 | Rejected |

Table 8 Boruta feature analysis for oil prices

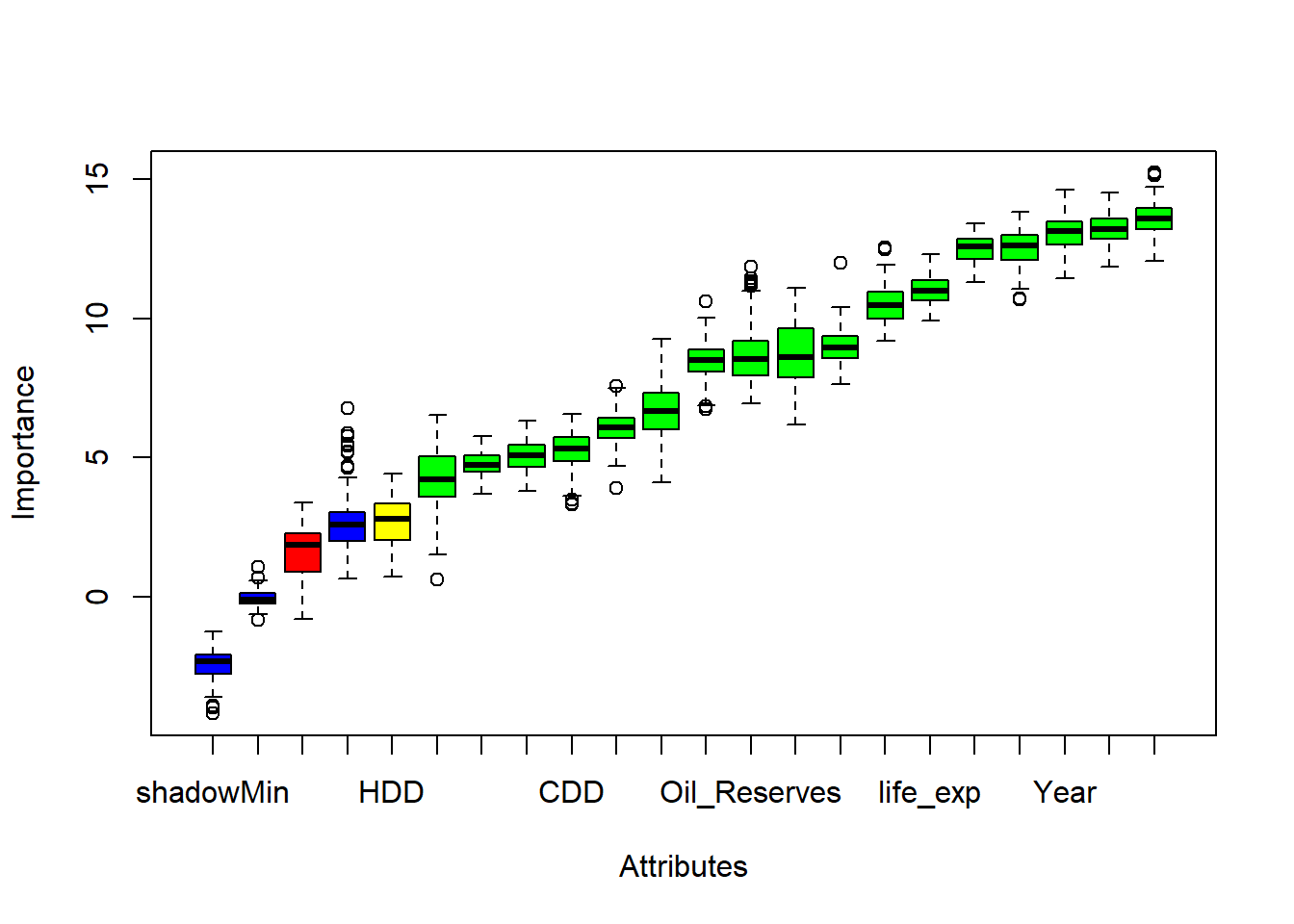


Figure 56 Boruta feature analysis for oil prices boxplot

|  |  |  |
| --- | --- | --- |
| **Features** | **Median Impact** | **Decision** |
| Oil\_Price | 21.231639 | Confirmed |
| Oil\_Stock | 14.00181 | Confirmed |
| GDP | 13.771841 | Confirmed |
| Year | 13.462641 | Confirmed |
| US\_Population | 13.449564 | Confirmed |
| life\_exp | 12.746623 | Confirmed |
| Net\_imports\_Energy | 11.438301 | Confirmed |
| Active\_Conflicts | 11.436812 | Confirmed |
| Oil\_Reserves | 11.02466 | Confirmed |
| Oil\_Production | 10.984145 | Confirmed |
| Eco\_Rec | 8.375199 | Confirmed |
| President\_Party | 8.094818 | Tentative |
| Senate | 8.037422 | Tentative |
| Energy\_Consumption | 6.580326 | Confirmed |
| Month | 5.18424 | Confirmed |
| House | 4.492152 | Confirmed |
| CDD | 1.929082 | Confirmed |
| HDD | 1.739728 | Confirmed |
| Season | 1.684771 | Tentative |

Table 9 Boruta feature analysis for Active Rigs

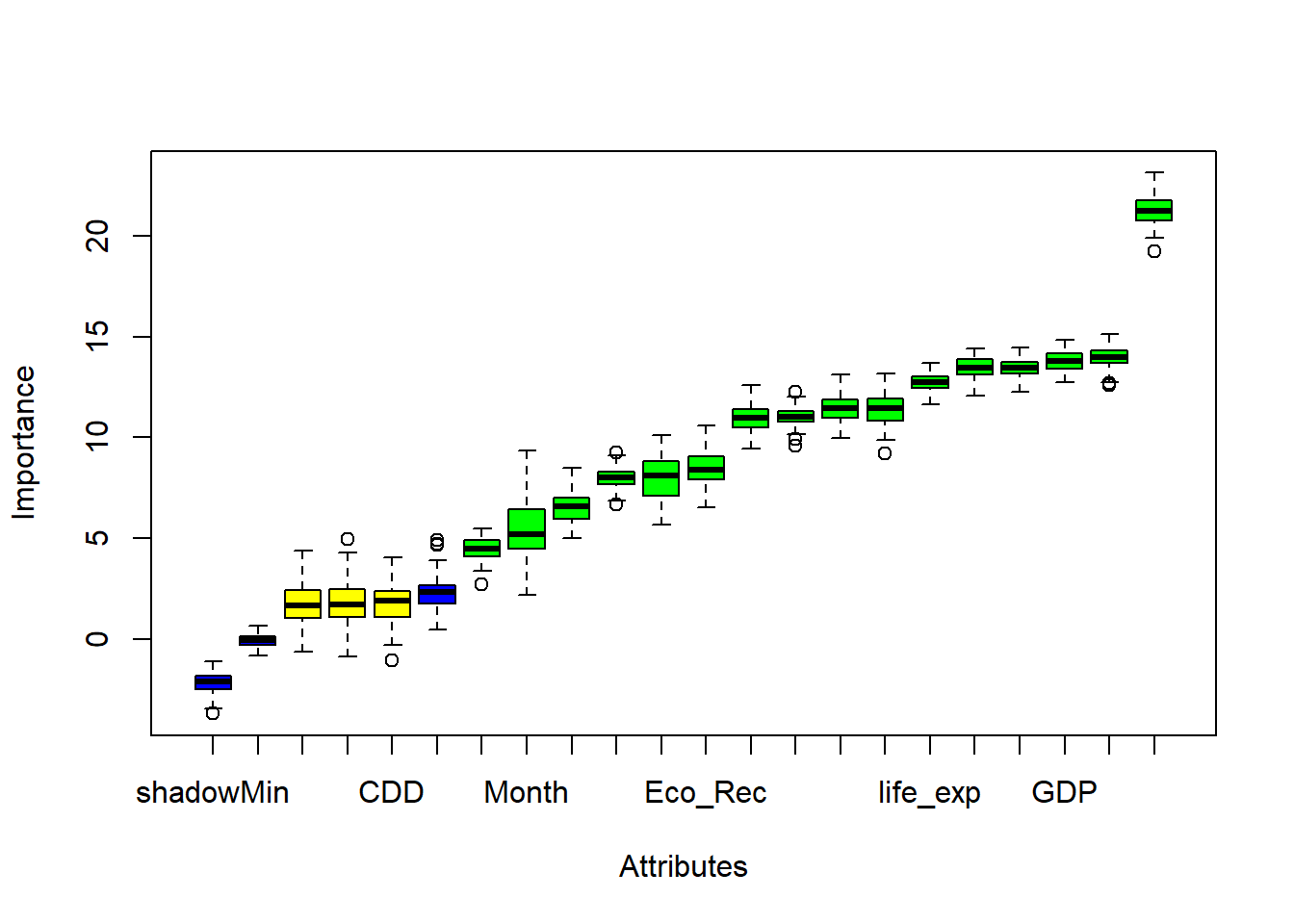


Figure 57 Boruta feature analysis for Active Rigs boxplot

**3.10 Summary of chapter**