aaa*

bbb

Xiyou Wang

April 19, 2024

ccc

Contents

*Code and data are available at: https://github.com/wxywxy666/Oil-Price-in-Alberta.git.

1 Introduction

Crude oil and natural gas significantly influence the global energy portfolio, and their price linkage affects trading strategies, investment decisions, energy policies, and portfolio optimization. The Alberta oil market, crucial to both the Canadian and the global economy, requires a clear understanding of the factors determining oil prices. This paper examines the relationship between oil prices and potential predictors such as gas production and well counts in Alberta from 2005 to 2022.

Despite previous research on market dynamics, there's a need for studies that focus on the Alberta market's specific conditions. Addressing this need, this study has constructed a dataset merging oil prices with gas production and well counts from various Alberta municipalities. The approach transitions from raw data to an analytical model, aiming to reveal provincial market trends.

The study's results offer insights into the oil market dynamics of Alberta and their implications for energy economics. They highlight the influence of regional production on oil prices, providing data that can guide economic and policy decisions. The findings underscore that while gas production and well counts influence oil prices, understanding the market's intricacies requires considering a range of other factors.

The paper begin with a data section (Section 2) to visualize and further understand the measurements, sources, methods, and variables we are examining. We then introduce the model (Section 3) used to understand the relationships in the data and report the results in the results section (Section 4). Finally, we provide a discussion of the findings (Section 5), summarizing the key points and future of this research.

2 Data

2.1 Sources

There are three original data sets used in this article, analyzed in `R` (**citeR?**). `Natural gas production by municipality` and `Well count by municipality` were downloaded from Government of Alberta: Open Data. They introduce the natural gas production of each

city in Alberta from 2003 to 2022 (the measurement unit is cubic meters) and number of wells, including total development, exploratory, evaluation and experimental wells drilled including natural gas, coalbed methane, crude oil, crude bitumen and other wells. Data is collected and pubulished by "Jobs, Economy and Northern Development", updated annually. `WCS Oil Price` downloaded from Government of Alberta: Alberta Economic Dashboard, including WTI (West Texas Intermediate) crude oil prices from 1983 to 2024 and WCS (Western Canadian Select) crude oil prices from 2005 to 2024. Oil prices are recorded by the Government and measured in US dollars per barrel.

2.2 Variables

`Natural gas production by municipality` and `Well count by municipality` share the same variable name due to the same source. `CSUID` and `CSD` refer to the administrative identification codes and names of cities in the Alberta. `Period` represents the year, from 2003 to 2022. `IndicatorSummaryDescription` corresponds to natural gas production and well count respectively. `UnitOfMeasure` for gas is m^3. `OriginalValue` is the specific natural gas production and the number of wells. `WCS Oil Price` contains four variables. `Date` is the first of each month from April 1983 to February 2024. `Value` represents the price of crude oil in US dollars. `Series` distinguishes between WCS and WTI crude oil type.

Only the variables relevant to this study are introduced here. Please refer to the Section A in Appendix for details.

2.2.1 Data Glimpse

We can get a glimpse of the processed data set through Table 1. The `gas_production` and `well_counts` here are the sum of each city, so that we can discuss the overall changes in Alberta, and `oil_price` is the average price for each year. The data processing process is in Section 2.4. Currently, we only know that `gas_production` and `well_counts` have an overall downward trend.

Table 1: First ten rows of the cleaned dataset

year	gas_production	well_counts	oil_price
2005	168226730	21751	36.24
2006	167462531	21214	45.04
2007	165551523	17633	49.62
2008	157677563	16304	79.59
2009	146670842	7746	52.15
2010	139377912	10346	65.31

2011	133492878	11684	77.97
2012	125968457	10367	73.17
2013	126049561	10470	72.77
2014	131392286	9896	73.60

2.3 Measurement

This study analyzes WCS crude oil prices, and these data are recorded in US dollars per barrel. This measure gives us a standardized economic indicator that reflects the market value of crude oil between 2005 and 2024. It is important to note that these values represent average prices per year, providing a macro perspective of industry finances.

The transformation from raw numbers into meaningful dataset entries involved aggregating individual municipality data to provide a province-wide overview. This consolidation was pivotal to shift the focus from local to regional trends, ensuring that our analysis captures the broader economic patterns that could influence the oil prices. By utilizing these measures, we transition from observing discrete activities—such as the drilling of a new well or the extraction of a cubic meter of gas—to examining the collective impact these activities have on Alberta's oil prices. This approach helps us understand the interconnected nature of production, infrastructure, and market prices.

Through meticulous data cleaning and processing, as detailed in Section 2.4, we ensure the dataset's reliability. We filter, aggregate, and cross-reference data from multiple sources, harmonizing different measurements into one.

2.4 Data Cleaning

Raw data usually cannot be used directly in analysis articles because they originally serve different service targets. Data cleaning is to organize the raw data into a form that serves our main research goals. First, for `Natural gas production by municipality` I merged the data of different municipalities by year in order to study the trends in the province. Do the same thing to `Well count by municipality`, but only selected observes with series as `WCS`. Those raw varibales are saved as `gas_producion` and `well_counts`. And since different data sets contain different numbers of samples, I chose a compromise from 2005 to 2022 as the research object to ensure that there is no missing data in the final processed data. In the `WCS Oil Price` I deleted the `labels` variable because it represents the specific time when prices are collected each time and is useless for our research. I stored the cleaned `gas_production`, `well_counts` and `oil_price` which represent the average price in a year in 'data/edited_data' respectively, and merged them together and saved it as `merged_data.csv`. Some R packages were used in cleaning the data, such as: `dplyr` (**dplyr?**), `lubridate` (**lubridate?**), `janitor` (**janitor?**).

3 Model

3.1 Simple Linear Regression

Simple linear regression is a statistical technique that allows us to summarize and study the relationship between two continuous variables: one predictor (independent variable) and one response (dependent variable). The relationship is typically modeled with a linear equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Here, Y represents the response variable, X is the predictor variable, β_0 is the intercept of the regression line, β_1 is the slope of the regression line indicating the expected change in Y for a one-unit change in X, and ε is the error term that captures all other factors affecting Y which are not included in the model.

In the context of our study, the simple linear regression model can be applied in several ways:

- To predict *oil_price* based on *gas_production*, where Y is *oil_price* and X is *gas_production*.
- To predict *oil_price* based on *well_counts*, where Y is *oil_price* and X is *well_counts*.
- To understand the relationship between *gas_production* and *well_counts*, where Y is *well_counts* and X is *gas_production*.

Each combination provides unique insights into the factors influencing the oil market in Alberta.

The first step is to use a simple linear model to understand the relationship between the number of wells (`well_counts`) and gas production (`gas_production`). I hypothesize that the number of wells is a significant predictor of gas production.

It can be found from Figure 1 that by simulating a simple linear model, there is a positive correlation between natural gas production and the number of wells in Alberta from 2005 to 2022, and most points are distributed within the conrfidence interval, which is the shaded area in the Figure 1. The simple conclusion is that as the year progresses, both natural gas production and the number of wells decline simultaneously. This sounds obvious, but as introduced in the Section 2.2, the number of wells here includes both natural gas and oil extraction. The increase in crude oil prices will lead to an increase in oil drilling, which will lead to a reduction in natural gas drilling, potentially causing natural gas prices to rise. The price signal between two commodities is the catalyst that prompts suppliers to produce one fuel over the other to maximize profits. Therefore, I hope to introduce the multiple linear model by adding more independent variables to further analyze and predict oil prices.

Figure 2a shows that the relationship appears quite scattered with very little correlation between gas production and oil prices (correlation coefficient -0.046, obtained from Table 2).

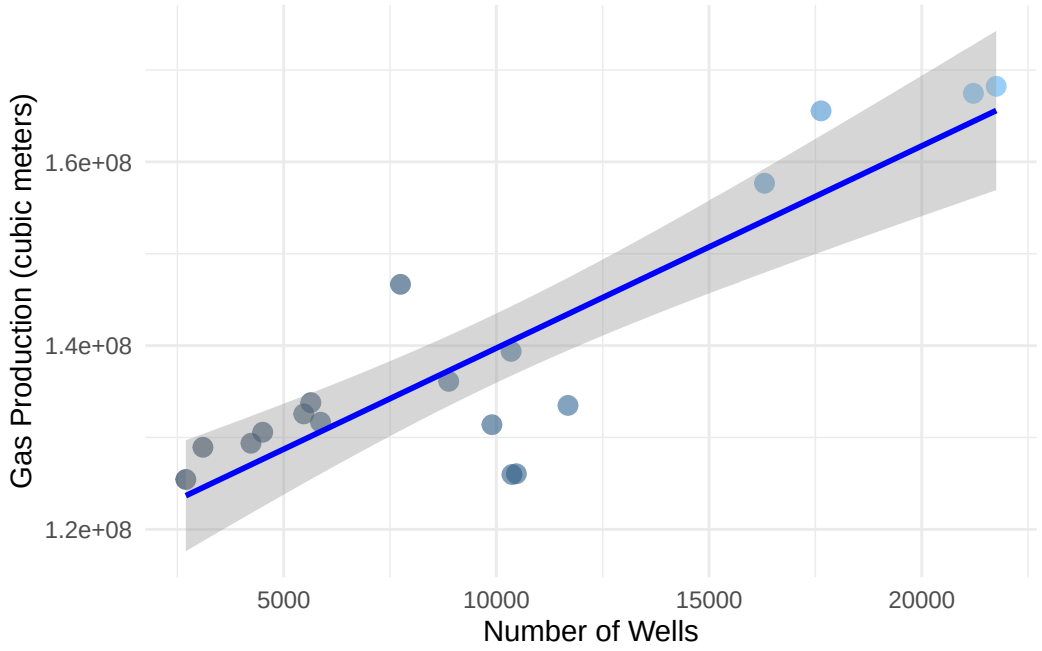
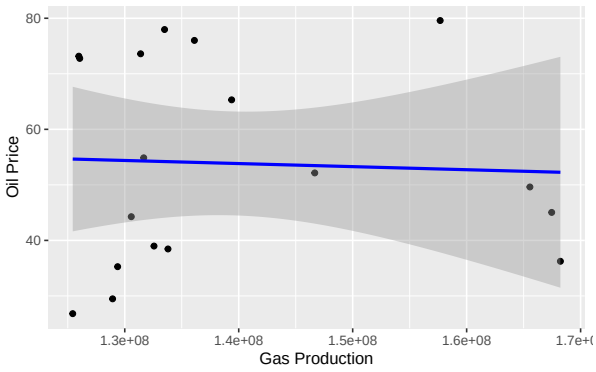
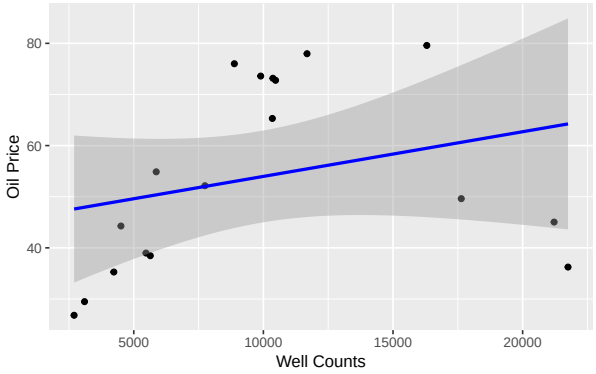


Figure 1: Linear regression between well counts and gas production in Alberta, 2005-2022



(a) Gas Production vs Oil Price



(b) Well Counts vs Oil Price

Figure 2: Relationship between each varibles and oil price in Alberta, 2005-2022

Table 2: Correlation calculation for Gas Production vs Oil Price and Well Counts vs Oil Price

	Variables		
	gas_production	**well_counts**	**oil_price**
gas_production	1.0000000	0.8712028	-0.0455136
well_counts	0.8712028	1.0000000	0.2836814
oil_price	-0.0455136	0.2836814	1.0000000

In Figure 2b, there's a visible upward trend, suggesting that higher well counts might be associated with higher oil prices, albeit the correlation (0.284) isn't very strong.

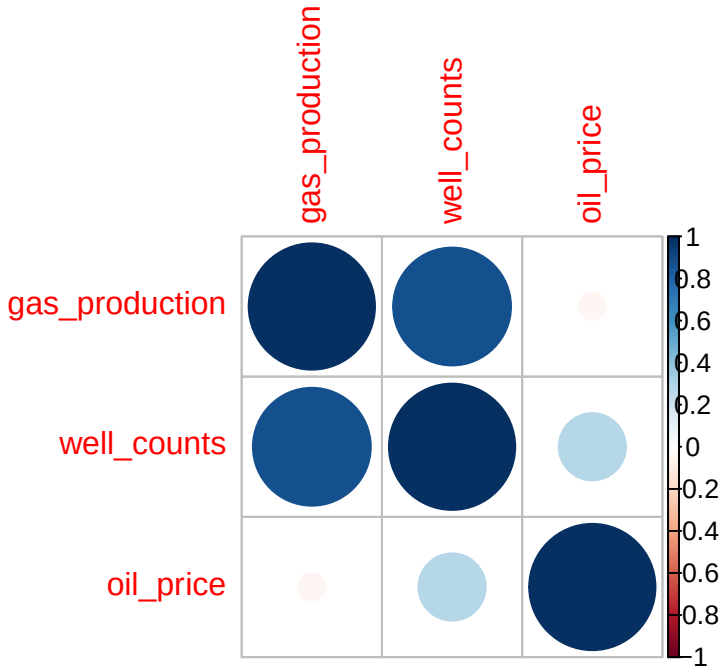


Figure 3: Visible correlation data display by correlation matrix plot

Gas Production and Well Counts are strongly correlated (0.871), which indicates multicollinearity and could affect the reliability of a regression model that includes both as predictors. The correlation between Well Counts and Oil Price is positive but moderate (0.284), suggesting some relationship.

3.2 Mutiple Linear Regression

When we extend our analysis to include more than one predictor variable, we use multiple linear regression. The model takes the form:

Table 3: Model summary

term	estimate	std.error	statistic	p.value	AIC	AICc	BIC	R2	R2_adjusted	RMSE	Sigma
(Intercept)	219.5372502	57.1114717	3.844013	0.0015933	NA	NA	NA	NA	NA	NA	NA
gas_production	-0.0000015	0.0000005	-3.073948	0.0077170	NA	NA	NA	NA	NA	NA	NA
well_counts	0.0041320	0.0012167	3.396148	0.0039892	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	152.1569	155.2338	155.7184	0.4358548	0.3606354	13.26886	14.53531

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

This allows us to investigate how multiple predictors jointly influence the response variable. For instance, we could model *oil_price* using both *gas_production* and *well_counts* simultaneously, allowing us to assess their combined effects.

- $Y = oil\_pric$, $X_1 = gas\_production$, and $X_2 = well\_counts$ to analyze the combined effect on oil prices.

By using these models, we can build a more comprehensive understanding of the dynamics in Alberta's oil market.

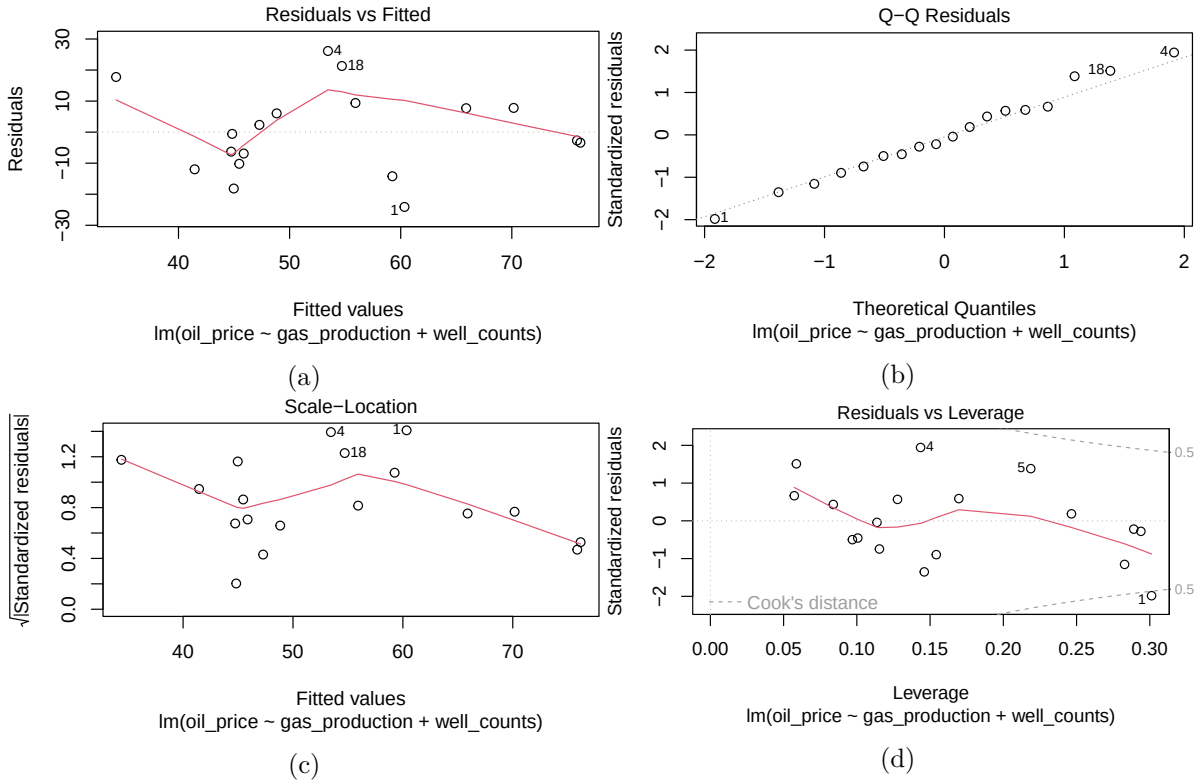
3.2.1 Model Fitting

Given the insights, a simple linear model may initially include both predictors, despite the potential multicollinearity, to see their individual effects on oil price. Therefore, only by using multiple linear regression and adding independent variables that have an impact on the resulting crude oil price can an accurate model be obtained.

From Table 3 we could know: `Intercept`: Approximately 219.537, which represents the predicted value of `oil_price` assuming both gas_production and well_counts are 0. A high value for the intercept may indicate that there are other unaccounted for factors affecting oil prices. `gas_production`: The coefficient is approximately -1.481e-06, which indicates a weak negative correlation between `gas_production` and `oil_price` For each unit increase in `gas_production`, `oil_price` is estimated to decrease slightly. `well_counts`: The coefficient is approximately 4.132e-03, which indicates a positive relationship between `well_counts` and `oil_price`. This means that with each additional oil well, `oil_price` is estimated to increase.

Figure 4a: Residuals vs. Fitted Values plot shows the distribution of residuals and fitted values. Ideally, the points should be randomly distributed around the 0 horizontal line with no obvious pattern. The residuals in this plot do not appear to exhibit a systematic pattern, indicating that there are no obvious nonlinearities in the model.

Figure 4b: Q-Q Plot shows the assumption of normal distribution of the residuals. If the residuals are perfectly normally distributed, the points should be very close to the reference



(a) (b) (c) (d)

Figure 4: Diagnostic plots for residuals

line. Although most points are close to this line, there are slight deviations at the ends, possibly indicating that the residuals deviate slightly from normality in the tails.

Figure 4c: Scale-Location plot evaluates the assumption of homoskedasticity of the residuals, i.e., different fitted values have the same residual variance. This plot shows that despite slight fluctuations, the residuals appear to be relatively evenly distributed overall, implying that the variances are relatively consistent.

Figure 4d: Residuals vs Leverage plot is used to detect whether data points have abnormal effects on the regression model. This plot shows that no data points have high leverage values or significant Cook's distances, meaning that the model's estimates are unlikely to be affected by extreme values or leverage points.

3.2.2 Model Prediction

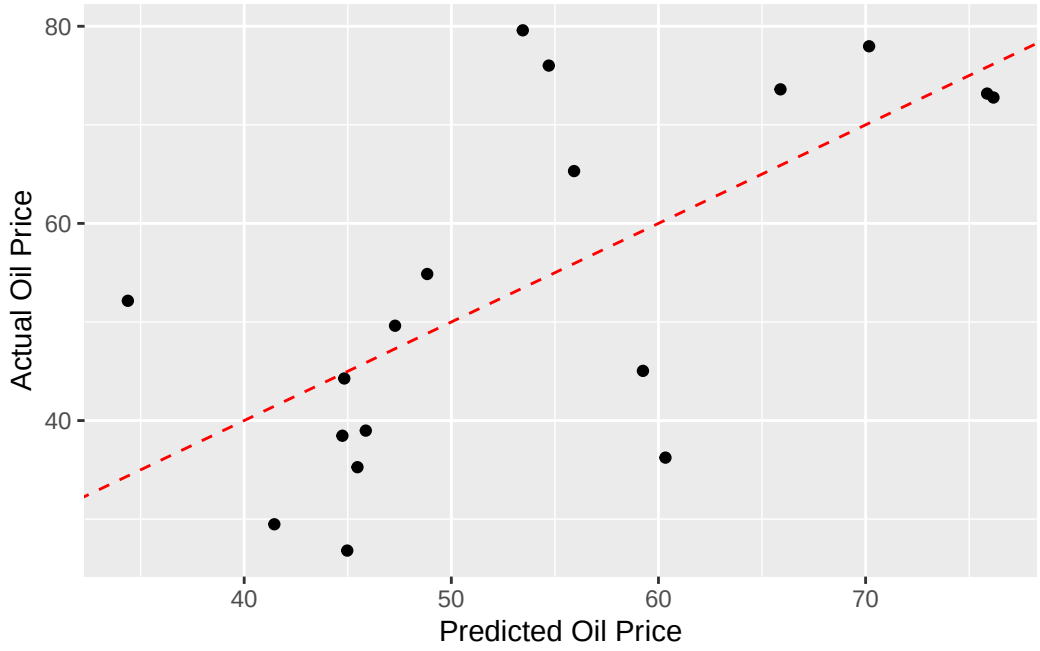


Figure 5: Predicted vs Actual oil price

The scatterplot (**fig5?**) showcases a comparison between predicted and actual oil prices, where the model's predictions are plotted on the x-axis and the actual observed prices on the y-axis. The black dots represent individual observations in the dataset, each comparing the model's prediction—based on gas production and well counts—to the observed market price of oil.

Observing the scatter of points around the red dashed line, which represents the line of perfect prediction, we can deduce that the model predictions are reasonably aligned with actual prices. This suggests that the model has captured the primary trend within the data. However, the

points do not lie exactly on the line, indicating the presence of prediction errors. The random scatter of these deviations from the red line suggests that the model does not exhibit systematic bias; in other words, it neither consistently overestimates nor underestimates the actual oil prices.

While the overall fit appears satisfactory, the points seem to disperse more as we move towards the extremities of the x-axis. This could indicate that the model's predictive accuracy varies across the range of values, potentially decreasing for very high or low predicted prices. Such a pattern might hint at additional complexities within the data that the current model does not account for, perhaps due to extreme values or inherent variability in oil prices.

4 Results

Starting with a simple linear regression analysis, we wanted to see how the natural gas production and the number of wells in Alberta province directly affect oil prices. Our model showed a subtle but statistically significant relationship. The simple linear regression graph (see Figure 1) indicates that from 2005 to 2022, natural gas production is positively correlated with the number of wells. Over time, we can see both variables decreasing, which aligns with the industry trend of fewer wells but higher production.

Further examination through a correlation matrix plot (see Figure 3) confirmed a strong correlation (0.871) between gas production and the number of wells, indicating multicollinearity when used as predictors in the same model. The correlation between the number of wells and oil prices is positive (0.284), but not very strong, suggesting only a moderate connection between these variables.

Moving on to multiple linear regression analysis allowed for a more detailed understanding of these relationships. The regression table (see Table 3) shows that while natural gas production has an impact coefficient on oil prices of about -1.481e-06, the number of wells has a positive impact and each additional well may increase oil prices by approximately 0.0041320.

Diagnostic plots (see Figure 4) further examined the model's assumptions. There was no apparent randomness or variance bias in both residual vs fitted values plot and scale-location plot respectively. Q-Q plot and residual vs leverage plot indicated that residuals are approximately normally distributed without any single data point having too much influence on the model.

The predictive ability of our model is visualized in scatterplot (see Figure 5), comparing predicted results with actual oil prices. The closeness of data points to the perfect prediction line suggests that overall our model's predictions align with actual values, although there are some random deviations indicating no systematic bias in our predictions.

In conclusion, our results indicate that in Alberta province, natural gas production and well count partially explain oil prices; subtle differences between these variables suggest complex interactions among them. Despite both predictive indicators showing declines, our model

largely reflects observed oil prices and suggests potential improvements through incorporating other relevant variables or addressing data complexity issues.

5 Discussion

5.1 Key Findings

This study quantifies the impact of Alberta's gas production and well count on oil prices.It reveals a statistically significant but subtle link.The study finds that gas production is negatively correlated with oil prices, suggesting that as gas production increases, oil prices decline slightly.Conversely, well count shows a positive correlation, suggesting that an increase in well count is associated with higher oil prices.This interaction reflects the complex dynamics of Alberta's oil market and highlights the delicate balance between resource extraction activities and market prices.

5.2 Weaknesses and Limitations

Despite the meticulous data processing and modeling, this study has its limitations.The strong correlation between gas production and well count indicates multiple collinearities that may distort the true impact of each independent variable on oil prices.Furthermore, the multiple linear regression model appears to capture the impact of gas production and well count on crude oil prices, without showing significant nonlinear relationships. The residual distribution of the model is relatively normalized, but the normality of the residuals may need to be further explored.At the same time, the model may not capture all factors affecting crude oil prices, so further analysis may be required to explore other potential predictive variables. While the multiple linear regression model adjusts for this, it does not take into account potential external factors such as geopolitical events, global market trends or technological advances in extraction methods, which could also have a significant impact on oil prices.Taken together, the model can serve as an effective benchmark for predicting oil prices based on gas production and well count.However, further improvements, which could include additional variables, explore nonlinear relationships, or adopt cross-validation techniques – could improve its predictive performance and robustness under different market conditions.Continuous model evaluation and improvement, especially in response to new data or insights, remains an integral part of the analysis process.

5.3 Next Steps

Future research directions include incorporating more predictive variables to explain the complexity of the oil market.This could involve capturing variables such as global oil demand, energy policy shifts or renewable energy advances, which could further refine the model.Nonlinear

approaches could also be explored to explain potential threshold effects or diminishing returns in production.Furthermore, a cross-validation framework could be introduced to improve the model's predictive robustness and accuracy. Future work could extend the scope from Alberta to the global landscape, discussing how the price relationship between crude oil and natural gas evolves over time, taking into account supply and demand.

6 References

Appendix

A Dataset

aaa

B Model

aaa

C Data Sheet

aaa