Final Report

Project Title: Crude Price and Production Forecasting for Strategic Planning

Prepared by: Arsen Tagibekov

Date: July, 2025

This final report and financial models were built with support from ChatGPT (OpenAI), used as a co-pilot for idea structuring, Excel logic validation, and formatting refinement. The project reflects my own financial judgment, assumptions, and execution, but benefited from AI-based structuring and iterative feedback throughout.

Table of Contents

1. Executive Summary ***2***

2. Project Objectives ***2***

3. Data Overview ***3***

4. Exploratory Data Analysis (EDA) ***5***

4.1 WTI Crude Oil Prices **5**

4.2 U.S. Crude Oil Production **9**

5. Time Series Modeling and Forecasting ***13***

5.1 WTI Crude Oil Prices **13**

5.2 U.S. Crude Oil Production **16**

6. Relationship between Price and Production ***19***

7. Strategic Implications ***21***

8. Conclusion ***22***

1. Executive Summary

This project presents a time series forecasting analysis of WTI crude oil prices and U.S. crude oil production volumes, developed to support strategic planning within the energy sector. Using publicly available data from the U.S. Energy Information Administration (EIA), the analysis spans several decades of monthly records and highlights key historical trends, volatility, and seasonality within the oil market.

The objective was to build reliable models to forecast the next 24 months of oil prices and production volumes, using industry-standard techniques such as Exponential Smoothing (ETS) and ARIMA. The models were evaluated based on their in-sample accuracy, trend behavior, and interpretability for decision-makers.

To enhance usability, the project also includes an interactive Shiny dashboard, enabling dynamic exploration of forecasts and historical behavior. In addition to individual forecasts, a correlation analysis was conducted to investigate the relationship between crude prices and production levels. The results suggest a moderate positive correlation with lag-dependent dynamics, reflecting how producers adjust output in response to price signals.

This end-to-end project showcases practical applications of R programming, time series forecasting, and data storytelling to address real-world questions in oil & gas economics. It is intended as a professional portfolio piece demonstrating both technical capability and business-oriented analysis.

2. Project Objectives

The primary objective of this project is to develop robust time series forecasting models for:

1) WTI Crude Oil Spot Prices
2) U.S. Monthly Crude Oil Production Volumes

These forecasts are intended to support strategic decision-making for stakeholders in the oil & gas sector, including financial analysts, policy advisors, planners, and risk managers. In particular, the analysis seeks to:

- Identify seasonal trends and structural shifts in oil prices and production using historical data
- Apply proven forecasting methods (ETS and ARIMA) to project market conditions over a 24-month horizon

- Evaluate and compare model performance using accuracy metrics such as RMSE and MAPE
- Investigate the relationship between price and production, including potential lag effects
- Present findings in a clear and interactive format via a Shiny dashboard

This project is positioned as a practical demonstration of how statistical modeling and forecasting in R can inform operational and strategic planning in the volatile context of global energy markets.

3. Data Overview

This project is based on two primary datasets obtained from the U.S. Energy Information Administration (EIA):

WTI Crude Oil Spot Prices

- **Source:** EIA – https://www.eia.gov/dnav/pet/hist_xls/RWTCd.xls
- **Frequency:** Daily
- **Period Covered:** January 1986 – Present
- **Unit:** U.S. Dollars per Barrel (USD/bbl)

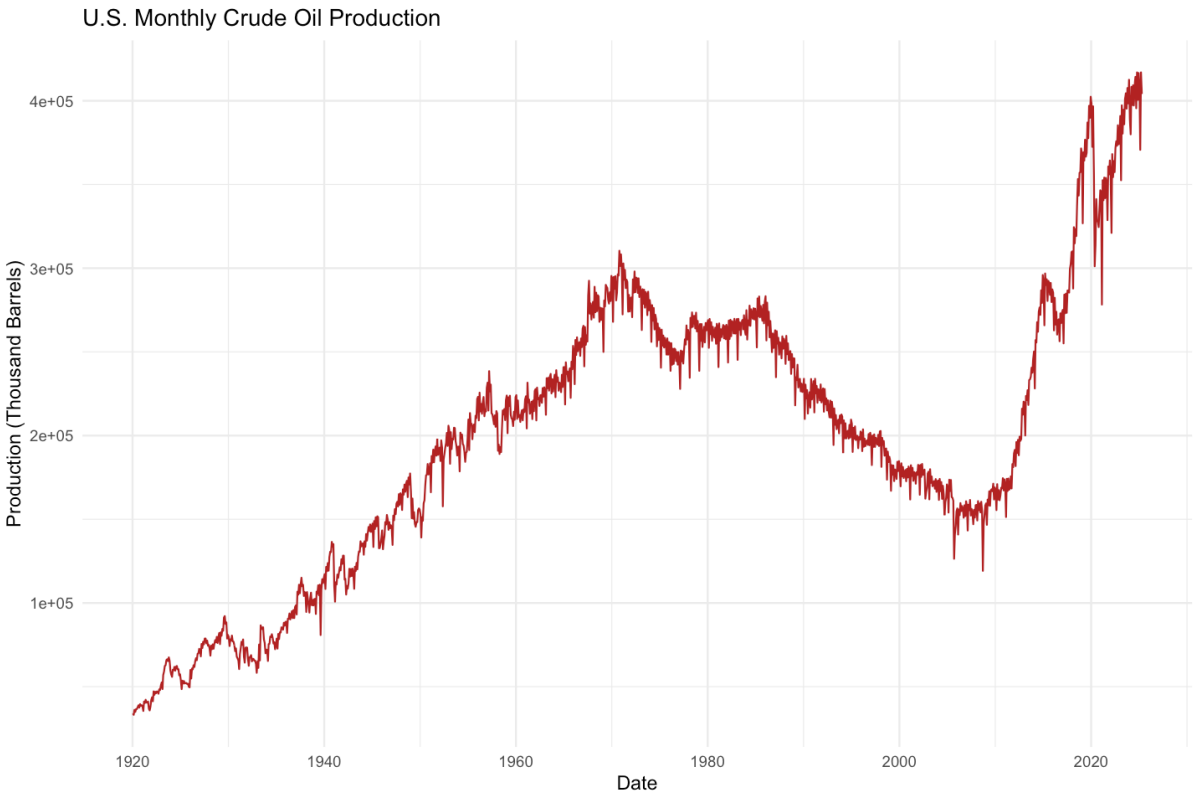
The dataset captures the daily spot price of WTI crude oil – one of the most widely recognized global benchmarks. Provided graph (Figure 1) includes key price volatility events such as the 2008 financial crisis and the historic crash in April 2020, when WTI briefly dropped below zero due to storage overflow and demand collapse during the COVID-19 pandemic.

U.S. Crude Oil Production Volumes

- **Source:** EIA – https://www.eia.gov/dnav/pet/xls/PET_CRD_CRPDN_ADC_MBBL_M.xls
- **Frequency:** Monthly
- **Period Covered:** January 1920 – Present
- **Unit:** Thousand Barrels per Month

This dataset provides historical production volume data aggregated at the national level. It reflects long-term shifts in U.S. oil output, including:

- The rise of domestic production post-1940s
- The decline after 1970s peak oil
- The sharp resurgence during the Shale Boom (post-2010)



Data Preprocessing

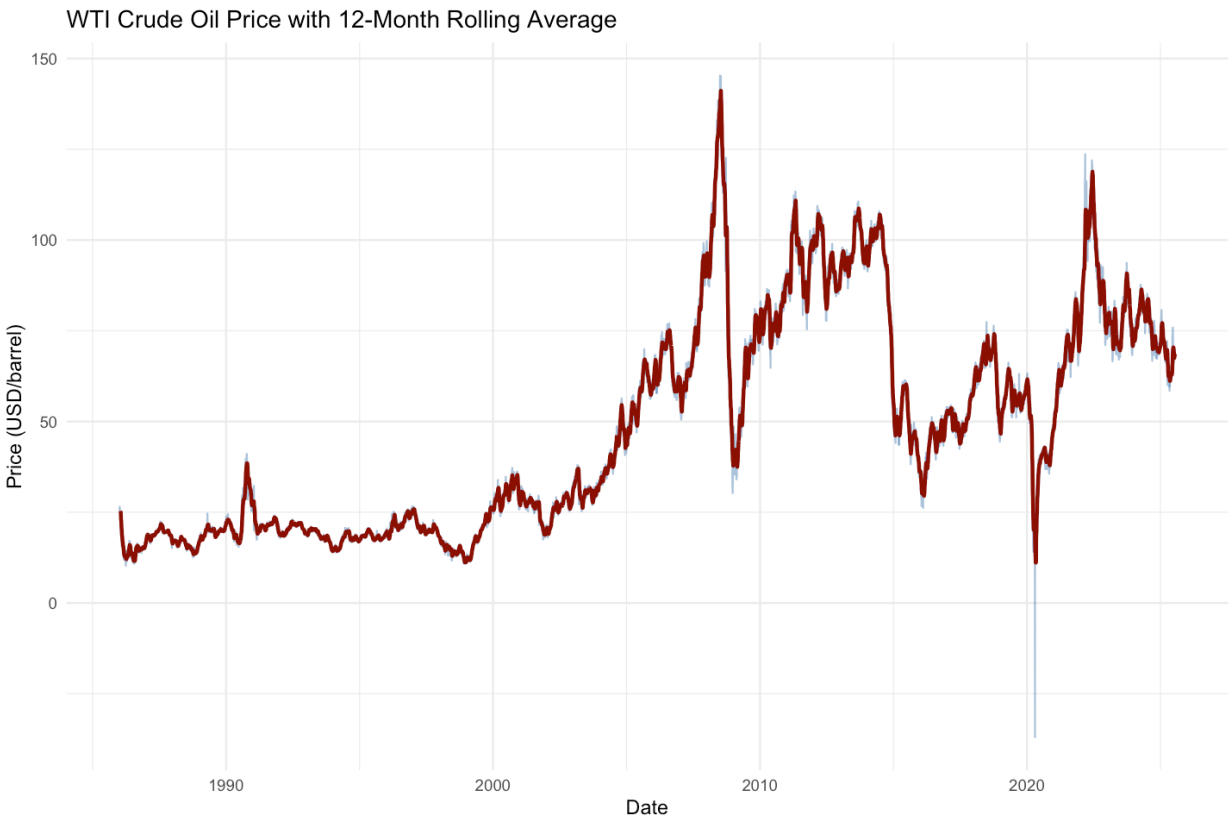
- **Date formatting:** Dates were converted from Excel formats into proper R Date objects.
- **Missing values:** Removed or forward-filled if appropriate
- **Aggregation:** WTI prices were analyzed using monthly means to match production granularity
- **Rolling averages:** 12-month rolling means were added to smooth short-term fluctuations

4. Exploratory Data Analysis (EDA)

4.1 WTI Crude Oil Prices

The exploratory analysis of WTI crude oil prices revealed significant long-term structural shifts and pronounced periods of volatility.

Trend and Rolling Average

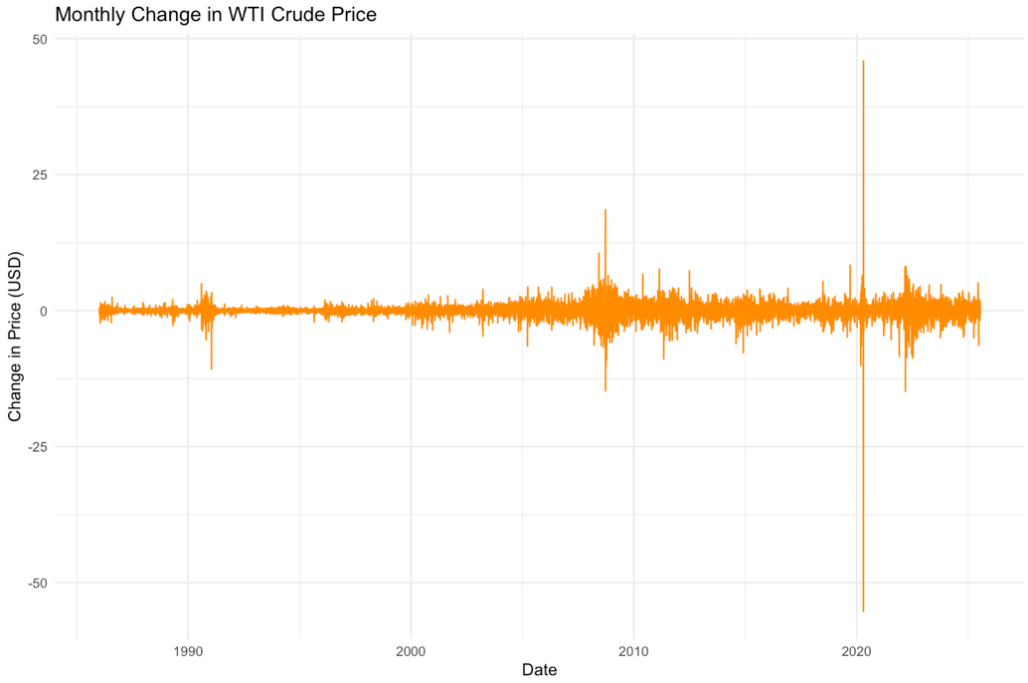


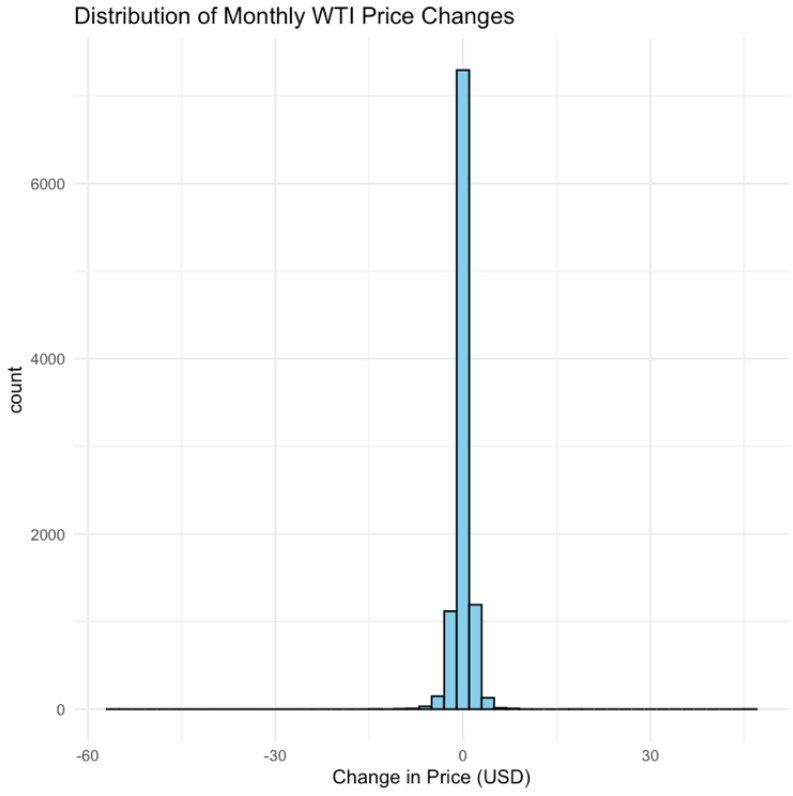
A line plot of daily WTI prices from 1986 to 2024 shows multiple major spikes and crashes, most notably:

- The 2008 commodity bubble and crash
- The 2020 negative pricing event during COVID-19

To smooth short-term noise, a 12-month rolling average was applied. This clarified the underlying macroeconomic trends and confirmed that WTI prices have gone through extended boom-and-bust cycles driven by both supply shocks and demand collapses.

Monthly Change Analysis





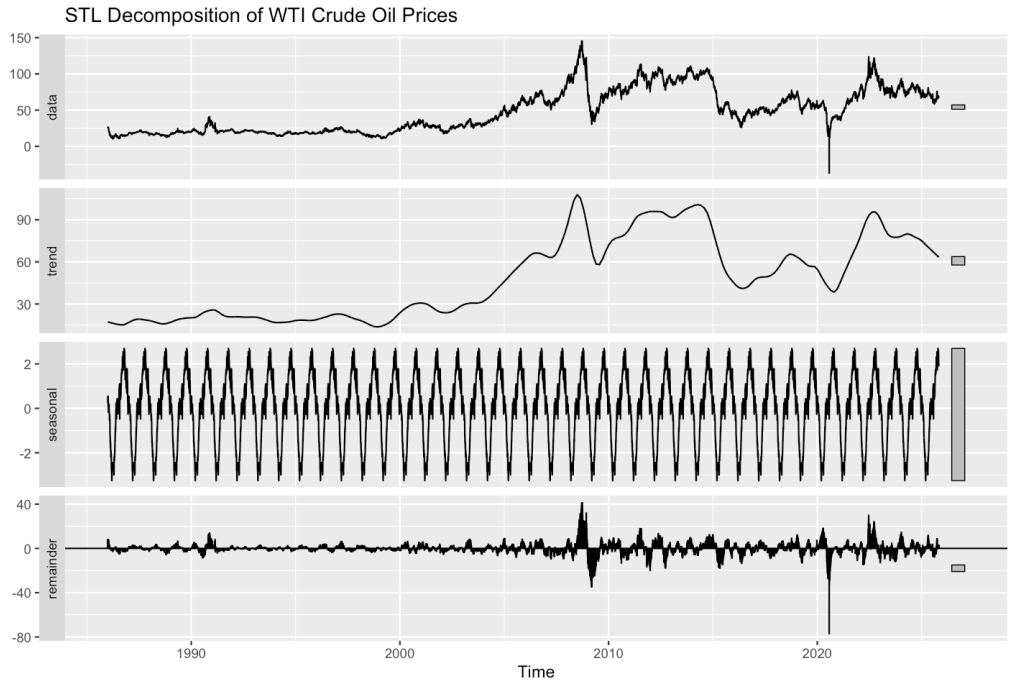
To assess volatility, we examined the first difference (monthly change) in crude prices. The chart of monthly changes revealed:

- Long periods of relative stability in the 1990s and early 2000s
- Extreme fluctuations in 2008 and 2020

- A pronounced outlier in April 2020, when prices dropped below $0

The histogram of these changes confirmed a highly leptokurtic distribution, meaning that while most changes are small, extreme, events are far more likely than in a normal distribution.

Seasonality and Decomposition (STL)



Using STL decomposition, the time series was broken into:

- A smooth long-term trend component
- A consistent seasonal pattern
- A volatile residual ("remainder") component with major spikes during crises

The seasonal component exhibits a subtle but regular pattern, indicating some cyclicality in oil prices across months.

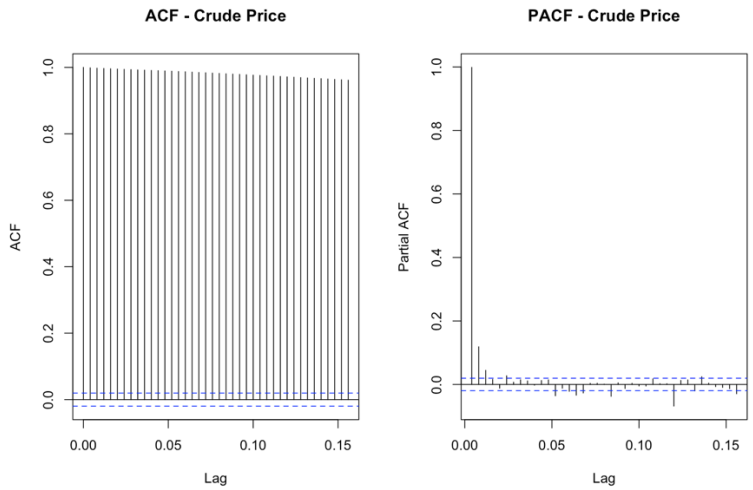
Stationarity Check (ADF-Test)

An Augmented Dickey-Fuller (ADF) test returned the following result:

- **ADF test statistic:** -3.18
- **P-value:** 0.09086

- **Conclusion:** Since the p-value > 0.05, we fail to reject the null hypothesis, meaning the WTI price series is non-stationary in levels and may require differencing before ARIMA modelling.

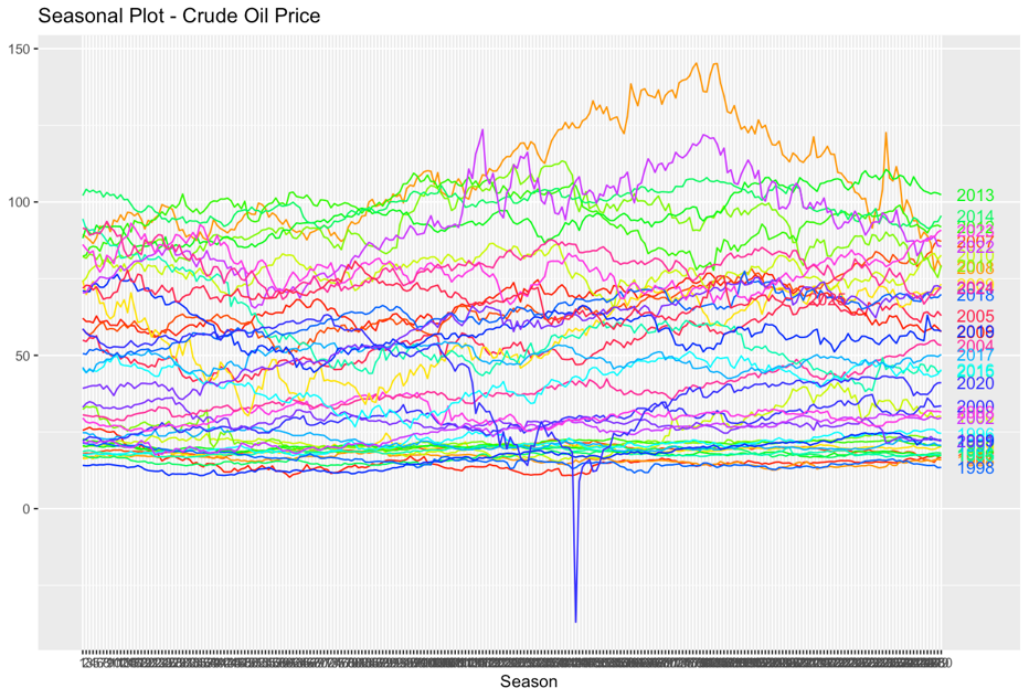
Autocorrelation (ACF/PACF)



- The ACF showed strong autocorrelation across many lags, consistent with trending behavior.
- The PACF indicated a sharp cutoff after lag 1, hinting at short-term memory.

These patterns support the need for a differencing step before applying ARIMA models.

Seasonal Subseries Plot



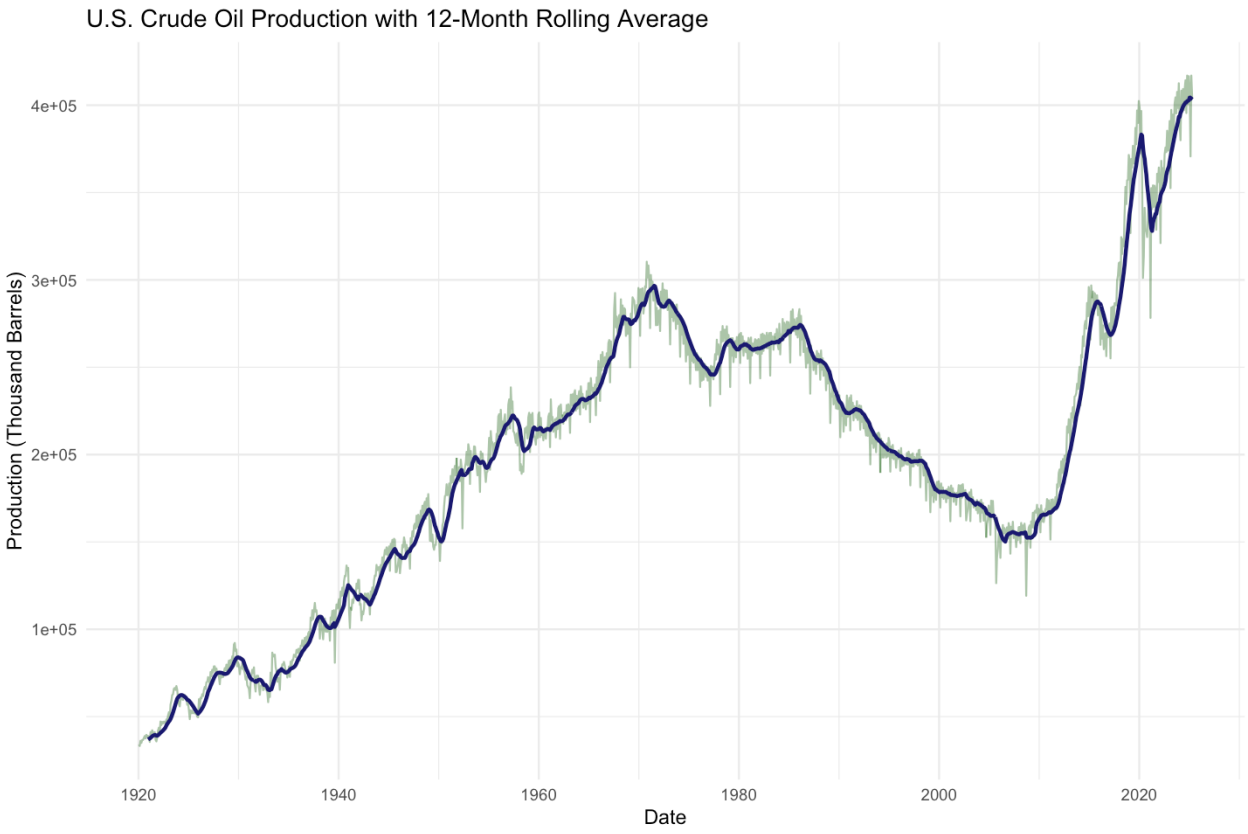
A seasonal plot comparing monthly patterns across years showed:

- Highly variable annual price paths
- No strong month-specific seasonal spike
- A few anomalous years like 2020 stood out clearly

4.2 U.S. Crude Oil Production

The time series analysis of U.S. monthly crude oil production, spanning over a century from 1920 to 2024, reveals dramatic shifts in the scale and structure of American oil output.

Trend and Rolling Average

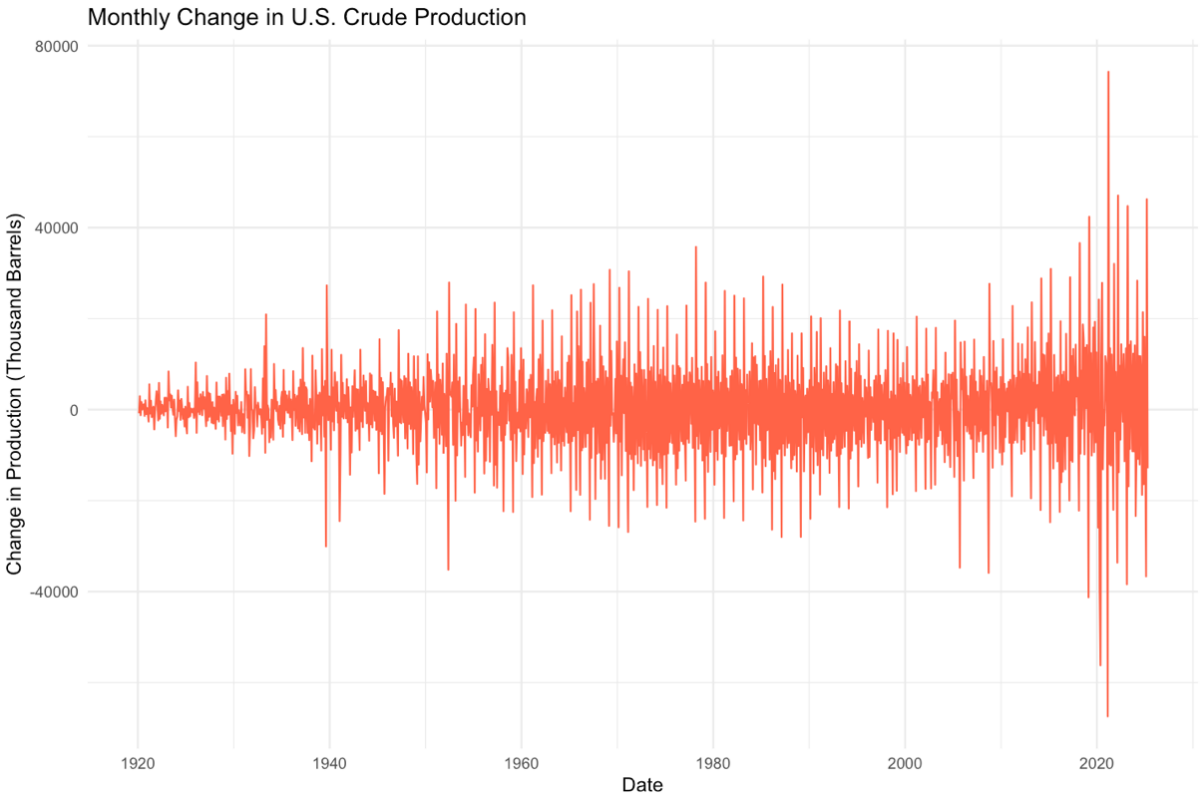


The long-term production trend shows several clear phases:

- A steady increase in production from 1920 to the mid-1970s
- A gradual decline during the post-1970s "peak oil" era
- A sharp resurgence after 2010 driven by the Shale Revolution

A 12-month rolling average smooths monthly noise and clearly highlights these major inflection points in the U.S. oil supply landscape.

Monthly Change Analysis



Monthly changes in production reveal extreme variability:

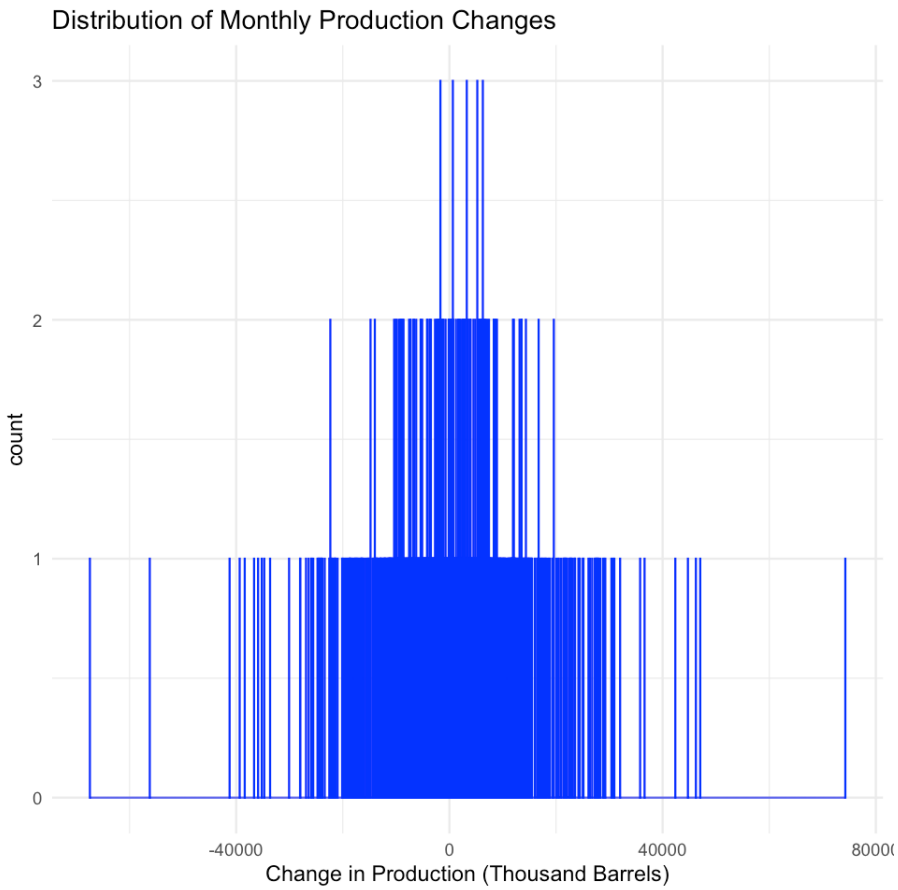
- Early decades show relatively modest month-over-month variation
- Post-2000s era exhibits much more pronounced spikes, especially in 2020-2021
- Significant outliers appear in response to price shocks and policy shifts

This change pattern underscores how modern extraction technologies and geopolitical events can rapidly alter output.

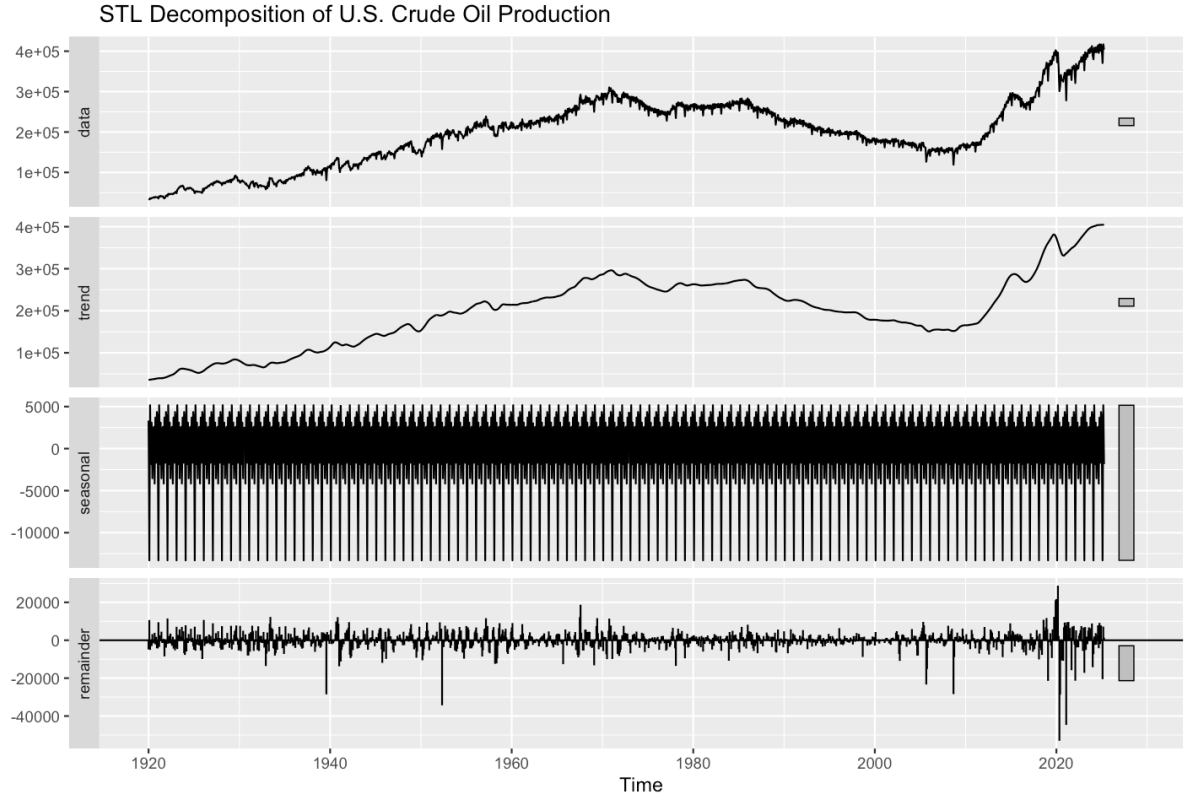
Distribution of Monthly Changes

The histogram shows a highly peaked distribution centered around zero, with frequent extreme values, particularly in recent years.

The distribution deviates substantially from normality, and the variance appears to increase over time, suggesting possible non-stationarity and heteroskedasticity.



STL Decomposition



Seasonal-trend decomposition using LOESS (STL) breaks the production time series into:

- A long-term trend component reflecting the macro-level phases of U.S. oil industry evolution
- A highly repetitive seasonal pattern, showing minor but persistent monthly cycles
- A residual component capturing volatility spikes, particularly post-2015

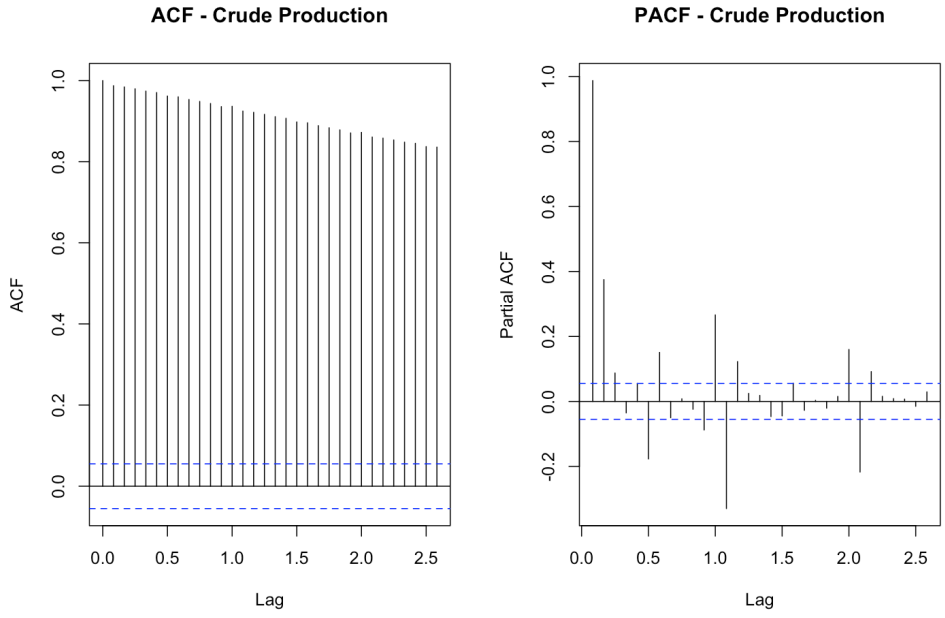
Stationarity Check (ADF Test)

The Augmented Dickey-Fuller test for stationarity yielded:

- **ADF statistic:** -1.1931
- **P-value:** 0.9076
- **Conclusion:** The series is non-stationary in levels and will require differencing prior to ARIMA modelling.

This aligns with the visual intuition of strong trend and structural breaks over time.

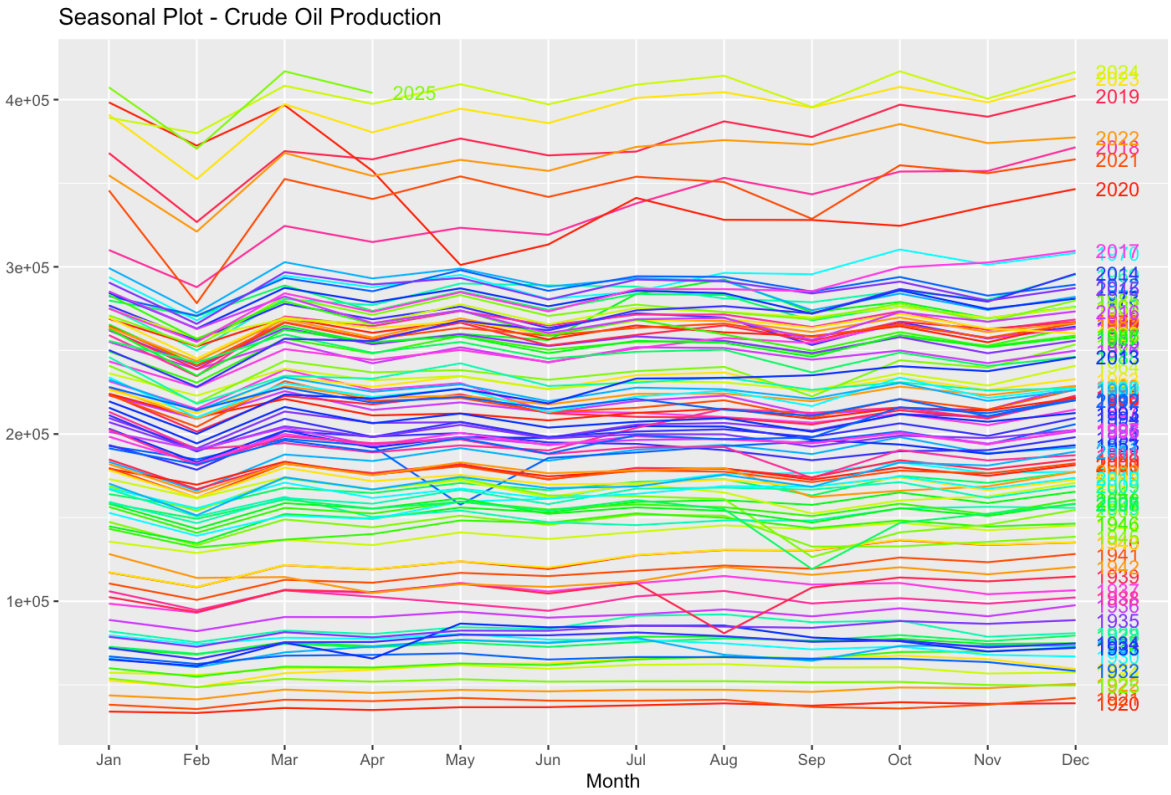
Autocorrelation (ACF and PACF)



The ACF plot shows high persistence (slow decay), reinforcing the presence of trend. The PACF plot indicates several partial autocorrelations spread across lags, implying some complex autoregressive structure.

These findings support the use of ARIMA modelling with one or more levels of differencing and potentially seasonal adjustments.

Seasonal Subseries Plot



The seasonal plot reveals that:

- Production tends to exhibit some monthly regularity across years
- However, extreme events in recent years (2020-2023) create strong visual outliers, particularly during summer months

This plot is useful for checking seasonal stationarity and informing model choices (e.g., seasonal ARIMA).

5. Time Series Modeling and Forecasting

5.1 WTI Crude Oil Prices

To forecast WTI crude oil prices, two widely used univariate time series models were implemented:

- ETS (Exponential Smoothing)
- ARIMA (AutoRegressive Integrated Moving Average)

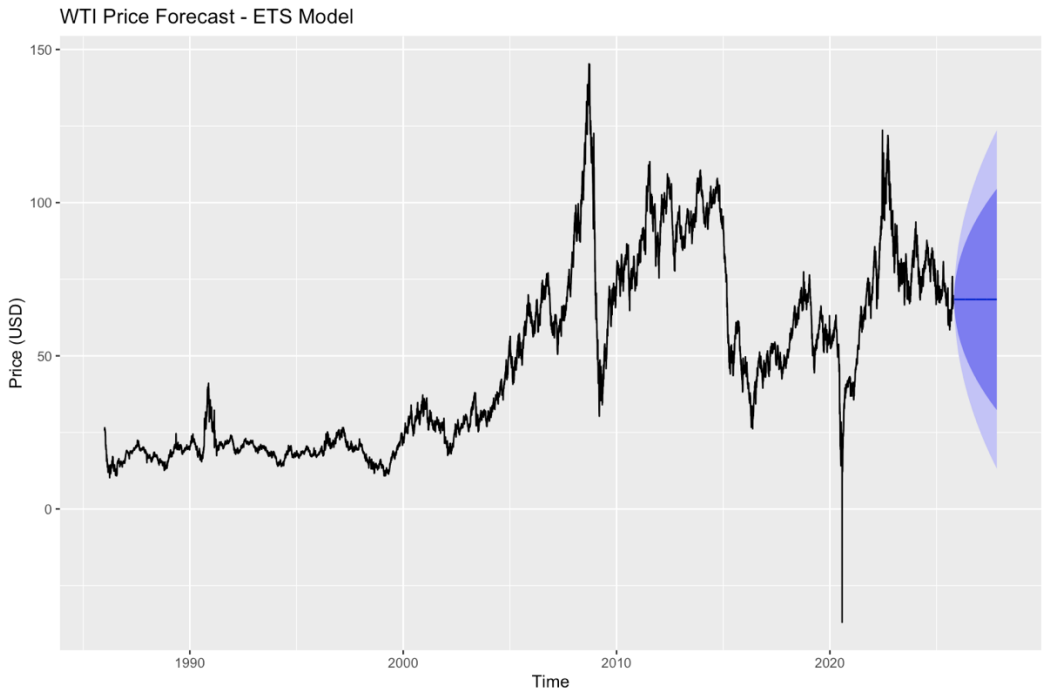
Both models were trained on monthly WTI spot price data and evaluated on in-sample performance.

ETS Model: ETS(A,N,N)

```
> summary(ets_model)
ETS(A,N,N)

Call:
ets(y = wti_ts)

  Smoothing parameters:
    alpha = 0.865

  Initial states:
    l = 25.6523

  sigma:  1.4567

     AIC     AICc      BIC
99159.00 99159.01 99180.62

Training set error measures:
                     ME      RMSE       MAE         MPE      MAPE       MASE         ACF1
Training set 0.004964107 1.456512 0.8196609 0.04119261 1.850395 0.06607663 0.004389319
```

The ETS model selected an additive error, no trend, no seasonality configuration.

- **Alpha (level smoothing):** 0.865
- **RMSE:** 1.456
- **MAE:** 0.820
- **MAPE:** 1.85%



The resulting forecast shows a mean-reverting pattern with relatively stable short-term expectations, and a confidence interval that widens into the future to reflect increased uncertainty.

ARIMA Model: ARIMA (3,1,1)

```
> summary(arima_model)
Series: wti_ts
ARIMA(3,1,1)

Coefficients:
          ar1      ar2      ar3     ma1
      -0.7950  -0.1362  -0.0515  0.6647
s.e.   0.2256   0.0314   0.0117  0.2257

sigma^2 = 2.119:  log likelihood = -17863.43
AIC=35736.85   AICc=35736.86   BIC=35772.88

Training set error measures:
                      ME     RMSE       MAE         MPE     MAPE       MASE          ACF1
Training set 0.005125777 1.455351 0.8194836 0.04024649 1.850751 0.06606234 0.0001256426
```

The best-fit ARIMA model applied first-order differencing to address non-stationarity (as indicated by the ADF test), and included three autoregressive and one moving average term.

- **ARIMA model coefficients:**
  - AR1=-0.795, AR2=-0.136, AR3=-0.051, MA1=+0.665
- **RMSE:** 1.455
- **MAE:** 0.819
- **MAPE:** 1.85%

The ARIMA forecast was visually similar to the ETS output, though slightly more responsive to short-term fluctuations.

Model Accuracy Comparison

```
> # Comparing Accuracy of the models
> accuracy(ets_model)
                      ME     RMSE       MAE         MPE     MAPE       MASE         ACF1
Training set 0.004964107 1.456512 0.8196609 0.04119261 1.850395 0.06607663 0.004389319
> accuracy(arima_model)
                      ME     RMSE       MAE         MPE     MAPE       MASE          ACF1
Training set 0.005125777 1.455351 0.8194836 0.04024649 1.850751 0.06606234 0.0001256426
```

Both models produced virtually identical accuracy metrics on the training set:

Metric	ETS Model	ARIMA Model
RMSE	1.4565	1.4554
MAE	0.8197	0.8195
MAPE	1.8504%	1.8508%
ACF1 (resid)	0.0044	~0.0000

Although both models performed equally well in terms of RMSE and MAE, the ARIMA model slightly outperformed in terms of residual autocorrelation (lower ACF1), suggesting slightly better residual independence.

Interpretation

- Both models project relative stability in WTI prices over the next 24 months, with no strong upward or downward trend.
- The wide confidence intervals reflect the inherent uncertainty in oil markets, especially post-2020 volatility.
- These forecasts are suitable for baseline planning, but may require external variables (e.g., geopolitical risk, OPEC policy) for scenario-based refinement.

5.2 U.S. Crude Oil Production

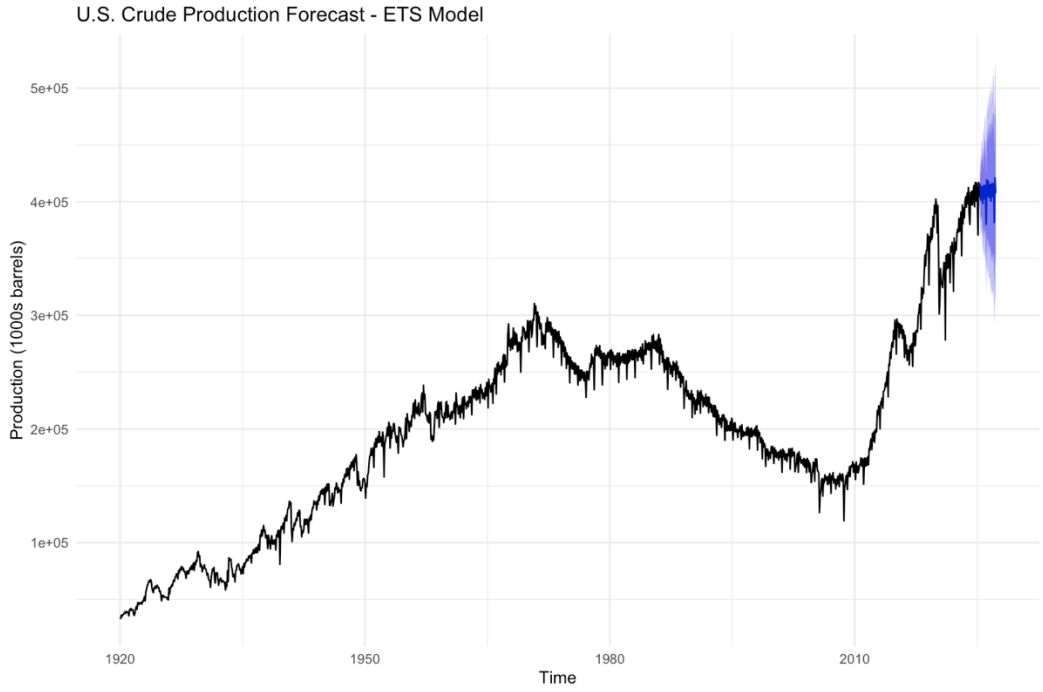
To forecast monthly U.S. crude oil production, both ETS and ARIMA models were applied to over a century of historical data. Given the non-stationarity and structural shifts identified during EDA, both models incorporated mechanisms to account for trend and seasonality.

ETS Model: ETS(M,Ad,M)

The ETS model selected a multiplicative error, additive damped trend, and multiplicative seasonality structure – a complex but well-suited configuration for energy production data.

```
ETS(M,Ad,M)

Call:
ets(y = prod_ts)

  Smoothing parameters:
    alpha = 0.7679
    beta  = 0.0042
    gamma = 1e-04
    phi   = 0.9776

  Initial states:
    l = 35561.9846
    b = 681.4967
    s = 1.0159 0.9874 1.0159 0.9784 1.0138 1.0148
           0.987 1.0228 0.9929 1.0268 0.9313 1.013

  sigma:  0.0315

     AIC     AICc      BIC
30828.72 30829.27 30921.28

Training set error measures:
                    ME     RMSE      MAE        MPE     MAPE      MASE       ACF1
Training set 291.6508 5459.964 3291.378 0.09721489 1.901867 0.2938274 0.07212681
```

- **Alpha (level smoothing):** 0.768
- **Beta (trend smoothing):** 0.004
- **Gamma (seasonal smoothing):** 0.0001
- **RMSE:** 5.459
- **MAE:** 3.291
- **MAPE:** 1.90%



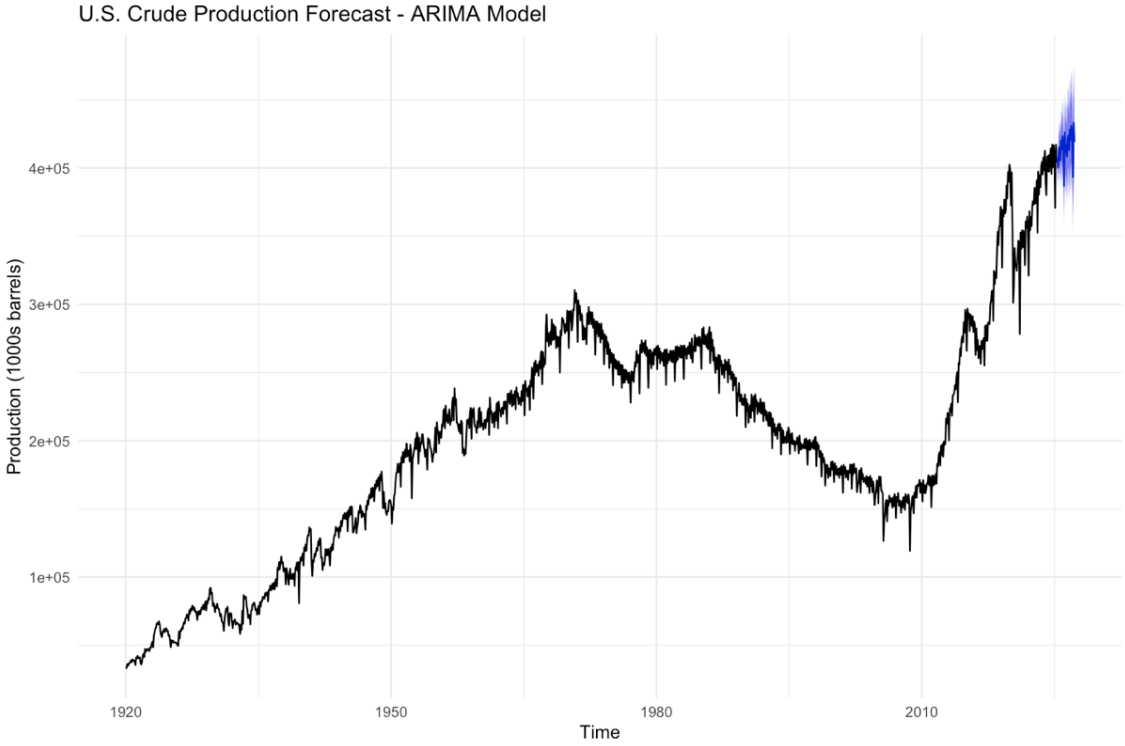
The forecast projects slight stabilization and convergence of production over the next 24 months, with a modest uncertainty band and no extreme shifts. This may reflect a dampening of recent volatility and a plateauing of shale-driven expansion.

ARIMA Model: ARIMA (2,0,0) (0,1,2) [12] with drift

```
Series: prod_ts
ARIMA(2,0,0)(0,1,2)[12] with drift

Coefficients:
         ar1     ar2     sma1    sma2     drift
      0.7967  0.1902  -0.9005  0.0986  299.4024
s.e.  0.0278  0.0280   0.0295  0.0293  196.1323

sigma^2 = 32522961:  log likelihood = -12609.53
AIC=25231.05   AICc=25231.12   BIC=25261.85

Training set error measures:
                    ME     RMSE      MAE          MPE     MAPE      MASE         ACF1
Training set 16.87991 5664.411 3459.967 -0.05655476 1.964658 0.3088776 -0.007072503
```

The ARIMA model incorporated:

- Two non-seasonal autoregressive terms (AR1, AR2)
- A seasonal differencing order of 1 to account for annual cycles
- Two seasonal MA terms (SMA1, SMA2)
- A constant drift term (~299), which acts like a trend component
- **RMSE:** 5.664
- **MAE:** 3.460
- **MAPE:** 1.96%



The ARIMA forecast closely matched the ETS trajectory, but introduced a slightly higher uncertainty band, and a more persistent upward drift due to the autoregressive and drift terms.

Model Accuracy Comparison

```
> accuracy(ets_prod)
                    ME     RMSE      MAE         MPE     MAPE      MASE       ACF1
Training set 291.6508 5459.964 3291.378 0.09721489 1.901867 0.2938274 0.07212681
> accuracy(arima_prod)
                    ME     RMSE      MAE          MPE     MAPE      MASE          ACF1
Training set 16.87991 5664.411 3459.967 -0.05655476 1.964658 0.3088776 -0.007072503
```

Metric	ETS Model	ARIMA Model
RMSE	5459.964	5664.411
MAE	3291.378	3459.967
MAPE	1.90%	1.96%
ACF1 (resid)	0.072	-0.007

- ETS model slightly outperformed ARIMA on RMSE, MAE, and MAPE
- ARIMA model residuals were slightly less autocorrelated, as indicated by a lower ACF1

While both models provided viable short-term forecasts, ETS demonstrated better in-sample fit, and is preferable when stability and interpretability are prioritized.

Interpretation

- Forecasts suggest U.S. production will remain elevated but without the sharp year-over-year increases seen in the 2017-2019 period
- The range of future production is fairly tight, reflecting the structured seasonality captured in both models
- These projections can inform capacity planning, infrastructure investment, and policy simulations

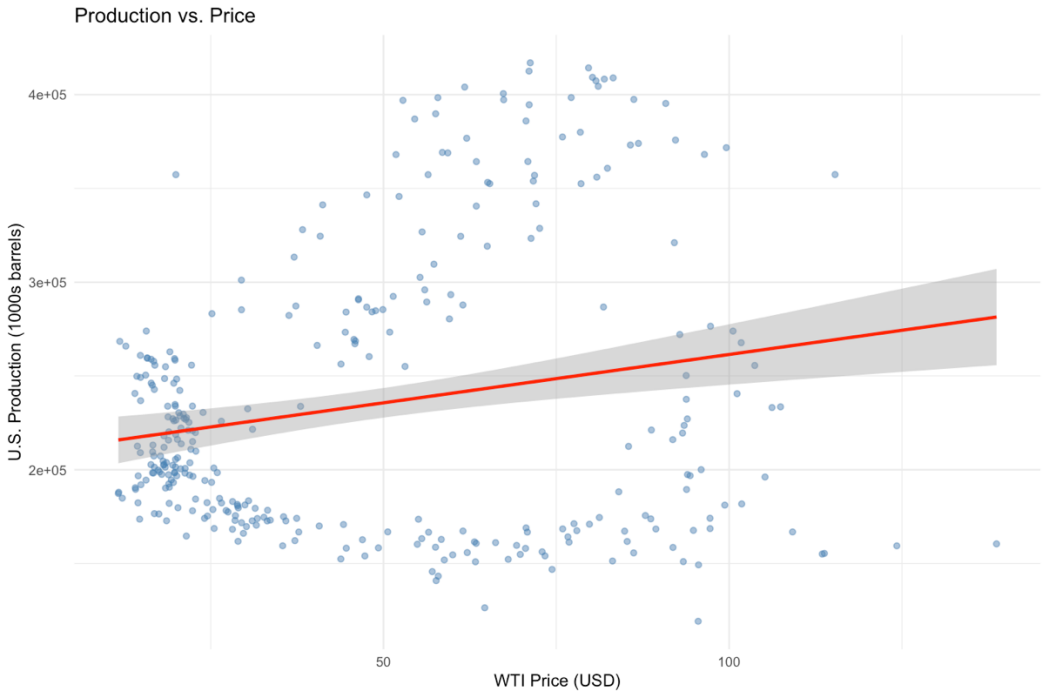
6. Relationship between Price and Production

Scatterplot Analysis

To examine the relationship between WTI crude oil prices and U.S. crude oil production, a scatterplot with a linear regression trend line was plotted. The visual pattern suggests a weak but positive relationship between the two variables.

- As WTI prices increase, production tends to increase as well, though with considerable dispersion.

- There are multiple clusters of production levels for similar price levels, indicating the influence of other factors such as technology, policy, and lag effects.



Correlation Analysis

Two Pearson correlation tests were conducted:

(a) Contemporaneous Correlation

```
> cor_test=cor.test(joined_data$Price, joined_data$Production)
> cor_test

        Pearson's product-moment correlation

data:  joined_data$Price and joined_data$Production
t = 3.7863, df = 325, p-value = 0.0001821
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09929666 0.30715295
sample estimates:
      cor
0.2055416
```

- **Pearson's r = 0.2055**
- **P-value < 0.001**
- **Interpretation:** There is a weak but statistically significant positive correlation between crude oil prices and production in the same month.

(b) Lagged Correlation

```
> cor.test(joined_data$Lagged_Price, joined_data$Production)

        Pearson's product-moment correlation

data:  joined_data$Lagged_Price and joined_data$Production
t = 4.0916, df = 324, p-value = 5.413e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1158190 0.3225151
sample estimates:
      cor
0.2216554
```

- **Pearson's r = 0.2217**
- **P-value < 0.001**
- **Interpretation:** Introducing a 1-month lag on WTI price slightly strengthens the correlation with production, supporting the idea that production responds with a short delay to price changes.

These findings suggests that while WTI prices do influence production levels, the relationship is modest, and other factors (e.g., infrastructure, regulation, geopolitics) likely play significant roles in determining output volumes.

7. Strategic Implications

- **Forecasting Insights:**
  - Forecasts provide valuable baselines for budgeting, investment decisions, and risk assessments in the oil sector.
  - ETS and ARIMA both proved useful for short- to medium-term planning.
- **Investment and Policy Planning:**
  - The moderate correlation between price and production underscores that market price alone does not drive production – operational, regulatory, and geopolitical factors must be considered.
  - Lag effects should be accounted for when designing fiscal policies or incentives for upstream investments.
- **Energy Strategy:**
  - The structural rise in production post-2010 confirms the critical role of technological advancement and unconventional extraction.

- For strategic planning, understanding the production response time to price changes is crucial – particularly when managing national reserves, taxation regimes, or climate transition timelines.

8. Conclusion

This analysis investigated the historical behavior and forecasted trends of U.S. crude oil production volumes and WTI crude oil prices using time series models and exploratory techniques. The key takeaways are:

- **Long-Term Trends:**
  - Crude production increased steadily until the 1970s, declined through the early 2000s, and has since surged due to the U.S. shale boom.
  - WTI prices remained relatively stable until the early 2000s but experienced dramatic volatility afterward, notably during the 2008 financial crisis and 2020 pandemic.
- **Volatility and Seasonality:**
  - WTI prices showed high volatility and visible seasonal components.
  - Production displayed strong long-term cyclical behavior, with notable resilience and growth post-2010.
- **Forecasting Models:** Both ETS and ARIMA models were applied to prices and production. Forecasts for both metrics suggest continued moderate growth under historical patterns, though external shocks remain possible. These models could be embedded into scenario planning tools for upstream investment budgeting, or macroeconomic sensitivity analysis.
  - **WTI Price:** ETS and ARIMA produced similar forecast accuracy, though ARIMA slightly outperformed on RMSE.
  - **Production:** ETS showed marginally better in-sample performance, but ARIMA had lower mean error and better residual structure.
- **Price-Production Relationship:**
  - Correlation analysis revealed a weak but significant positive relationship between price and production.
  - A 1-month lag in price marginally strengthened the correlation, supporting the idea of a short-term response by producers to market signals.