ConocoPhillips

Problem 2: Sweet Spot Identification

Aditya Holla, Ameera Aslam, Beshoy Shaker, Elijah Flores, Jayant Bhaskaruni

What is this Project?

- The goal of the project is to determine where to drill down into the ground for oil and gas.
- Not all parts of the rocks are made equally.
- Some have more oil and gas. Some have better fluidity. “How easily liquids flow”
- Where are the rocks with the most oil and gas that have the best fluidity?

Ultimately, we are trying to find the “sweet spots” of the area through machine learning, so we can get the most oil and gas, in the most efficient way possible.

Why it matters?

Getting sweet spot identification right means:

- **Higher efficiency** – drilling fewer but more productive wells.
- **Lower cost** – avoiding bad drilling locations.
- **Better sustainability** – using fewer resources to get the same energy

Project Data

Petrophysical	Production	Spatial (Well Locations)
● Depth ● Porosity ● Permeability ● Facies	● Oil Output ● Pressure ● Locations	● Bottom-hole X/Y ● Well Depth ● 3D Data

Data Source: *ConocoPhillips*

Basic Dataset Description

- 55 different wells
- 14 different variables
- Typically everything was numeric
- Facies was the only categorical variable

VARIABLES

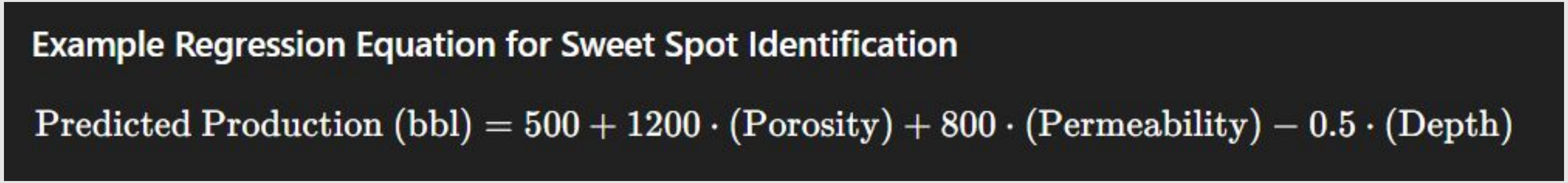
Well Number	Numbers assigned to wells
Well Name	Assigned names to the wells
Bottomhole X	x coordinates
Bottomhole Y	y coordinates
Co [MSTB]	Cumulative oil in stock tank barrel
Cw (bbl)	When hydraulically fracturing, this is the total volume (in barrels) of water or other fluid injected into the well.
POROS	Porosity of the rock
KX	Permeability in the x direction
KY	Permeability in the y direction

VARIABLES CONT.

TD(MD)	Total (measured) depth
Cg (mmcf)	Cumulative gas production measured in Million Cubic Feet (MMcf)
FACIES	The type and characteristics of rock layers
P_2020-1-6	Reservoir pressure recorded on Jan 6, 2020
P_2029-1-1	Simulated pressures predicted for Jan 1, 2029

Relevant ML Terms

- **Supervised Learning** Task
- Risk: **Overfitting** - memorizes training
- Small Dataset
- **Cross-validation -** testing how model works with new data
- **Regularization -** adds penalty to large coefficients
- **Regression**



Example Regression Equation for Sweet Spot Identification

$$\text{Predicted Production (bbl)} = 500 + 1200 \cdot (\text{Porosity}) + 800 \cdot (\text{Permeability}) - 0.5 \cdot (\text{Depth})$$

Approach

- Going to start with a simple model. Understand how the data correlates with each other
- How “good” each well is. Predict production
- Advance to more complicated models and eventually deep learning
- Be able to create a map, and find the best location for drilling

Research Questions

Geological Sweet Spots	Spatial Sweet Spots	Pressure & Production Dynamics
• Which facies, porosity, and permeability values are linked to high-production wells? • Is performance driven by stratigraphic or structural traps?	• Where on the map are the most productive wells located? • Can clusters of sweet spots be identified using production and pressure data?	• How does pressure drawdown oil production? • What trends emerge in production and pressure overtime? • Why do some wells underperform despite similar geographical settings?

Assumptions

- Features actually drive production
- The features won't change
- Data is accurate and consistent
- Wells are independent

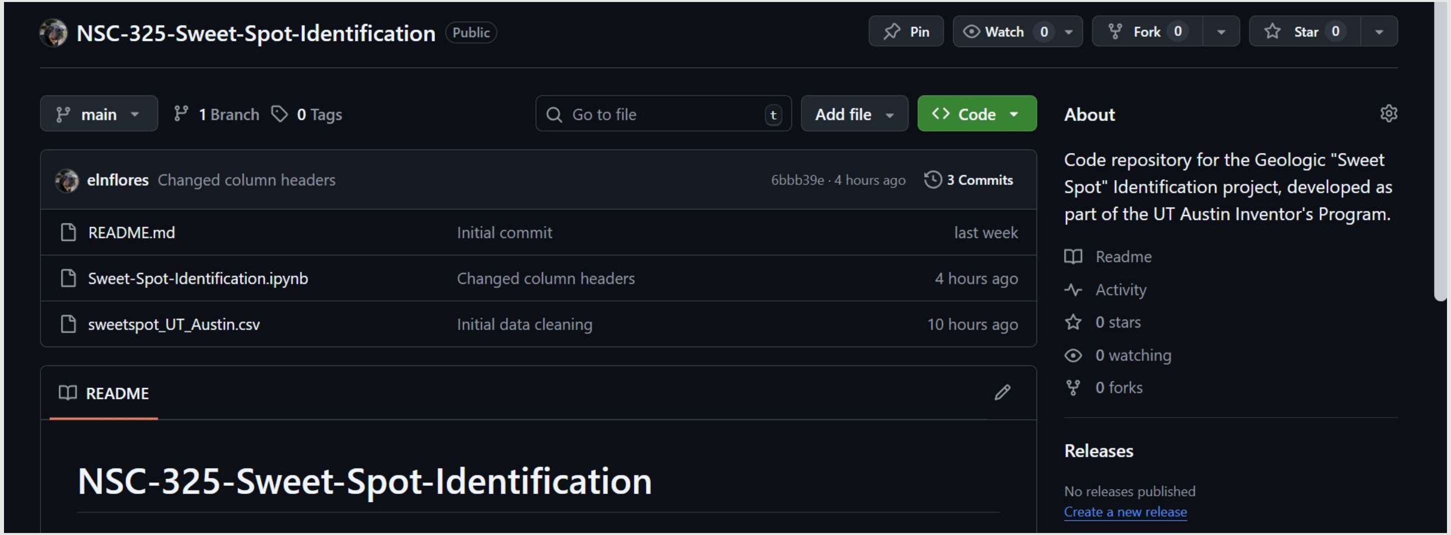
Requests to Mentor

- Reached out to Dr. Ortega
- Give a brief rundown of what we have looked at so far and the steps we have taken

- Confirm the 0’s are missing data?
- End result?
- Different types of porosity?

Public GitHub Repository

https://github.com/elnflores/NSC-325-Sweet-Spot-Identification



NSC-325-Sweet-Spot-Identification

elnflores Changed column headers 6bbb39e · 4 hours ago 3 Commits

README.md	Initial commit	last week
Sweet-Spot-Identification.ipynb	Changed column headers	4 hours ago
sweetspot_UT_Austin.csv	Initial data cleaning	10 hours ago

README

NSC-325-Sweet-Spot-Identification

About

Code repository for the Geologic "Sweet Spot" Identification project, developed as part of the UT Austin Inventor's Program.

- Readme
- Activity
- 0 stars
- 0 watching
- 0 forks

Releases

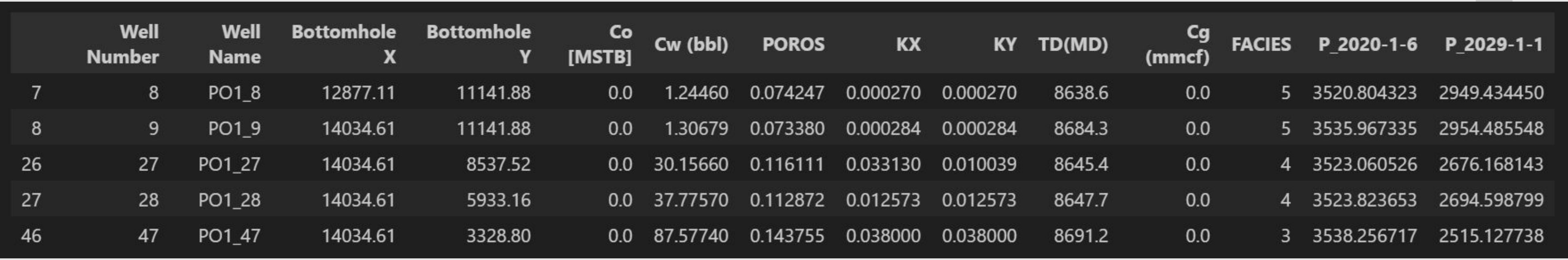
No releases published
Create a new release

Feature Engineering

- Renamed variables to be more human-readable
  - Ex. bh_x, bh_y, oil_prod_mstb
- One-hot encoded facies
  - FACIES -> facies_2, facies_3, facies_4, facies_5

Feature Imputation

- Printing the amount of 0s in the dataset revealed 5 data points with missing data, which is around 9% of the data.



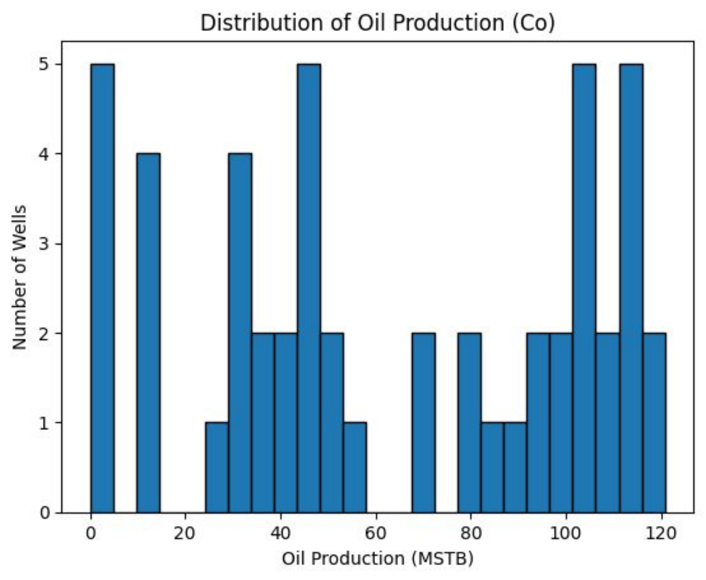
	Well Number	Well Name	Bottomhole X	Bottomhole Y	Co [MSTB]	Cw (bbl)	POROS	KX	KY	TD(MD)	Cg (mmcf)	FACIES	P_2020-1-6	P_2029-1-1
7	8	PO1_8	12877.11	11141.88	0.0	1.24460	0.074247	0.000270	0.000270	8638.6	0.0	5	3520.804323	2949.434450
8	9	PO1_9	14034.61	11141.88	0.0	1.30679	0.073380	0.000284	0.000284	8684.3	0.0	5	3535.967335	2954.485548
26	27	PO1_27	14034.61	8537.52	0.0	30.15660	0.116111	0.033130	0.010039	8645.4	0.0	4	3523.060526	2676.168143
27	28	PO1_28	14034.61	5933.16	0.0	37.77570	0.112872	0.012573	0.012573	8647.7	0.0	4	3523.823653	2694.598799
46	47	PO1_47	14034.61	3328.80	0.0	87.57740	0.143755	0.038000	0.038000	8691.2	0.0	3	3538.256717	2515.127738

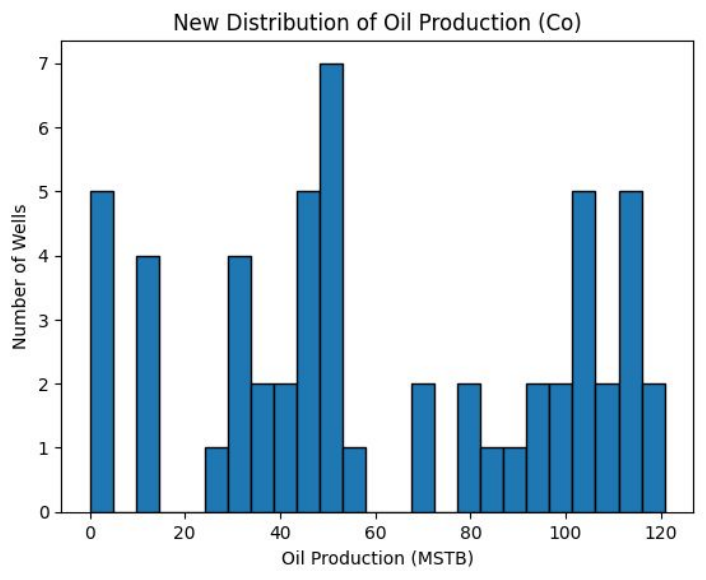
Feature Imputation

- Implemented median imputation on oil and gas production
  - Preserves dataset (very small - 55 wells)
- Wanted to start simple (between mean or median)
- Median imputation is more resistant to outliers than mean imputation because it is not affected by extreme high/low values

Feature Imputation

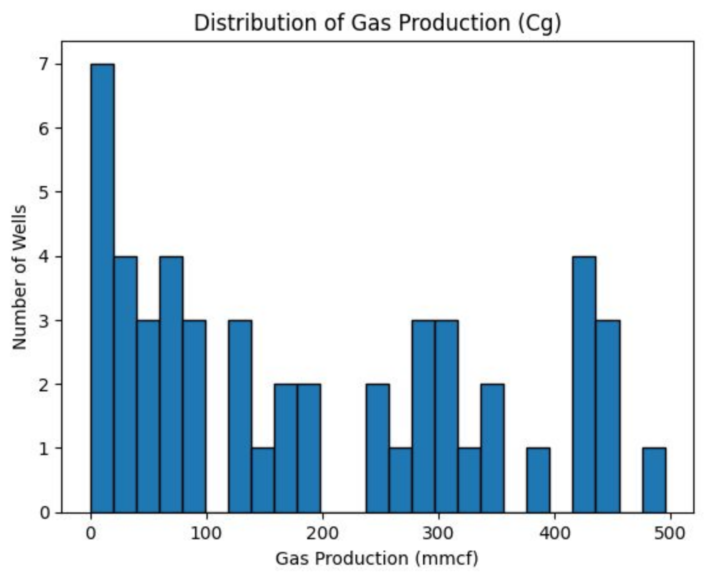
- Oil production, before and after:

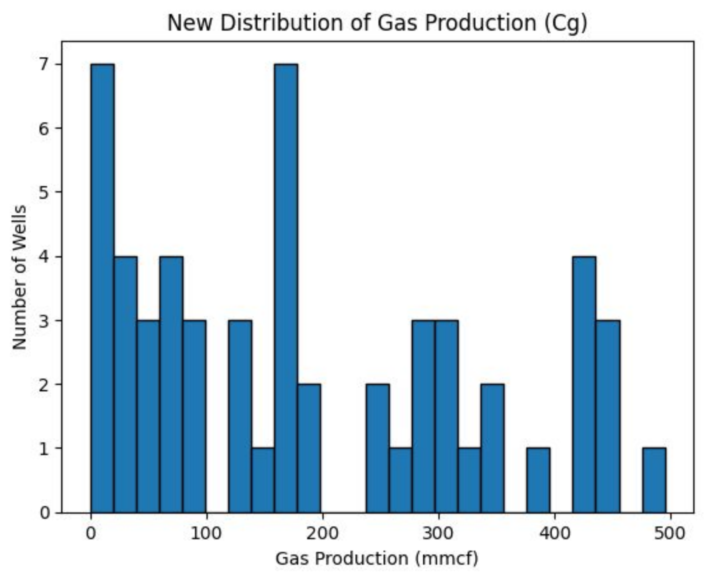




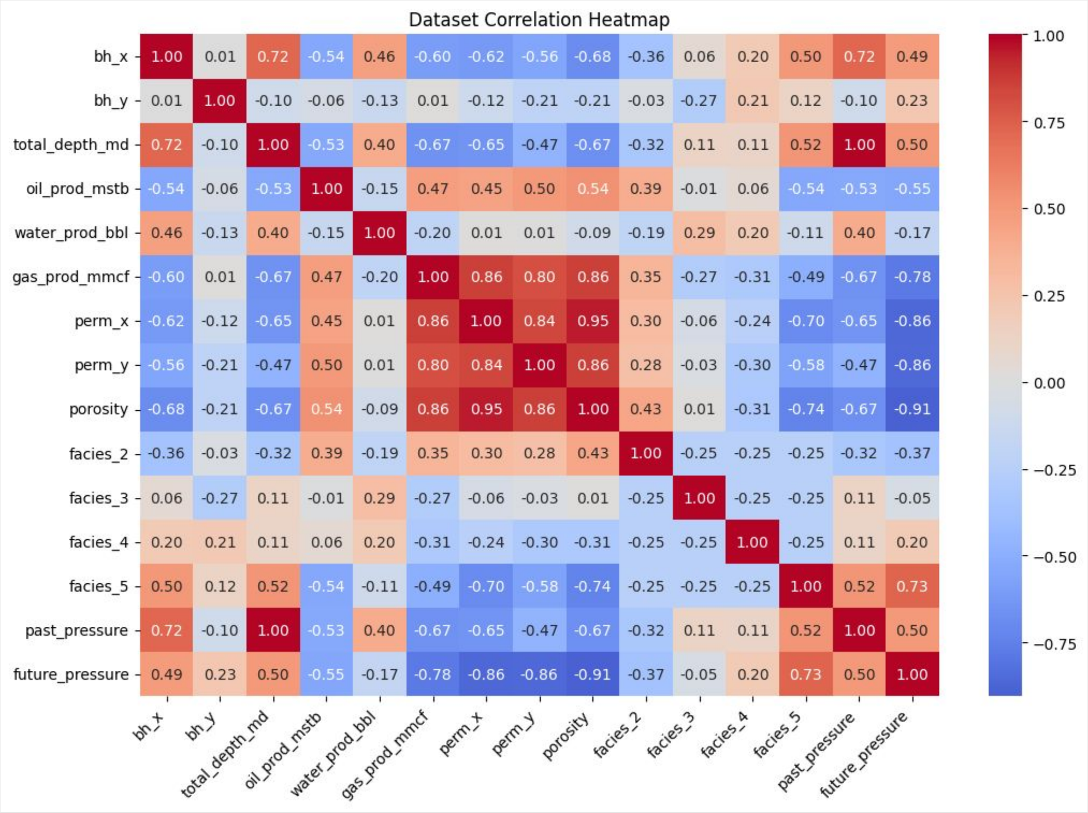
Feature Imputation

- Gas production, before and after:





Feature Correlation & Selection

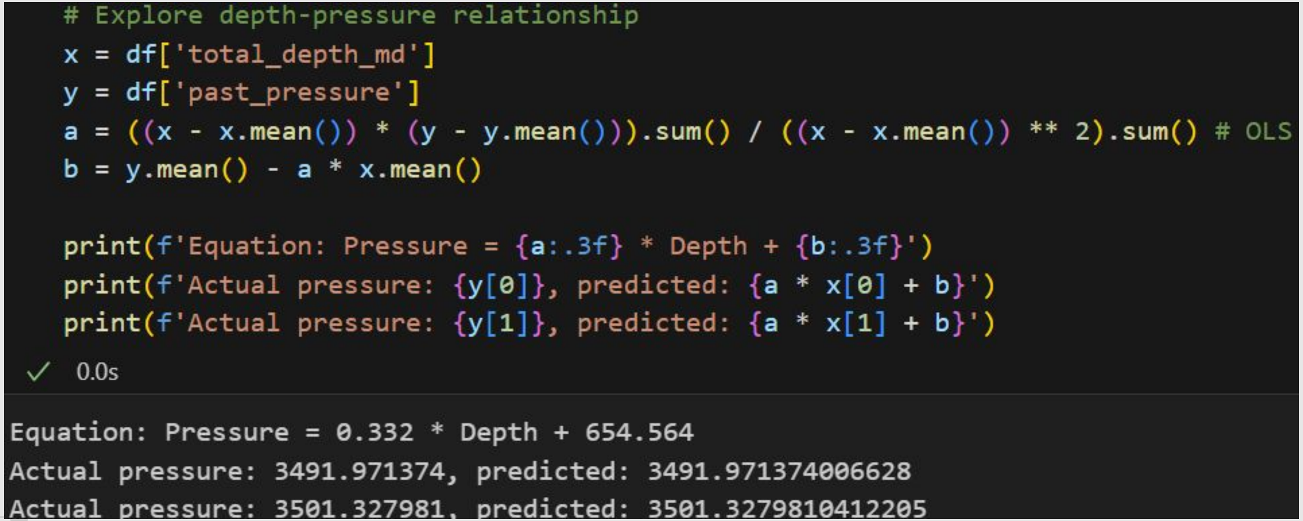


Dataset Correlation Heatmap

	bh_x	bh_y	total_depth_md	oil_prod_mstb	water_prod_bbl	gas_prod_mmcf	perm_x	perm_y	porosity	facies_2	facies_3	facies_4	facies_5	past_pressure	future_pressure
bh_x	1.00	0.01	0.72	-0.54	0.46	-0.60	-0.62	-0.56	-0.68	-0.36	0.06	0.20	0.50	0.72	0.49
bh_y	0.01	1.00	-0.10	-0.06	-0.13	0.01	-0.12	-0.21	-0.21	-0.03	-0.27	0.21	0.12	-0.10	0.23
total_depth_md	0.72	-0.10	1.00	-0.53	0.40	-0.67	-0.65	-0.47	-0.67	-0.32	0.11	0.11	0.52	1.00	0.50
oil_prod_mstb	-0.54	-0.06	-0.53	1.00	-0.15	0.47	0.45	0.50	0.54	0.39	-0.01	0.06	-0.54	-0.53	-0.55
water_prod_bbl	0.46	-0.13	0.40	-0.15	1.00	-0.20	0.01	0.01	-0.09	-0.19	0.29	0.20	-0.11	0.40	-0.17
gas_prod_mmcf	-0.60	0.01	-0.67	0.47	-0.20	1.00	0.86	0.80	0.86	0.35	-0.27	-0.31	-0.49	-0.67	-0.78
perm_x	-0.62	-0.12	-0.65	0.45	0.01	0.86	1.00	0.84	0.95	0.30	-0.06	-0.24	-0.70	-0.65	-0.86
perm_y	-0.56	-0.21	-0.47	0.50	0.01	0.80	0.84	1.00	0.86	0.28	-0.03	-0.30	-0.58	-0.47	-0.86
porosity	-0.68	-0.21	-0.67	0.54	-0.09	0.86	0.95	0.86	1.00	0.43	0.01	-0.31	-0.74	-0.67	-0.91
facies_2	-0.36	-0.03	-0.32	0.39	-0.19	0.35	0.30	0.28	0.43	1.00	-0.25	-0.25	-0.25	-0.32	-0.37
facies_3	0.06	-0.27	0.11	-0.01	0.29	-0.27	-0.06	-0.03	0.01	-0.25	1.00	-0.25	-0.25	0.11	-0.05
facies_4	0.20	0.21	0.11	0.06	0.20	-0.31	-0.24	-0.30	-0.31	-0.25	-0.25	1.00	-0.25	0.11	0.20
facies_5	0.50	0.12	0.52	-0.54	-0.11	-0.49	-0.70	-0.58	-0.74	-0.25	-0.25	-0.25	1.00	0.52	0.73
past_pressure	0.72	-0.10	1.00	-0.53	0.40	-0.67	-0.65	-0.47	-0.67	-0.32	0.11	0.11	0.52	1.00	0.50
future_pressure	0.49	0.23	0.50	-0.55	-0.17	-0.78	-0.86	-0.86	-0.91	-0.37	-0.05	0.20	0.73	0.50	1.00

Feature Correlation & Selection

- Oil production: depth, past_pressure
- Gas production: permeability and porosity
- Permeability and porosity appear to have multicollinearity
- Past pressure and depth are near perfectly collinear

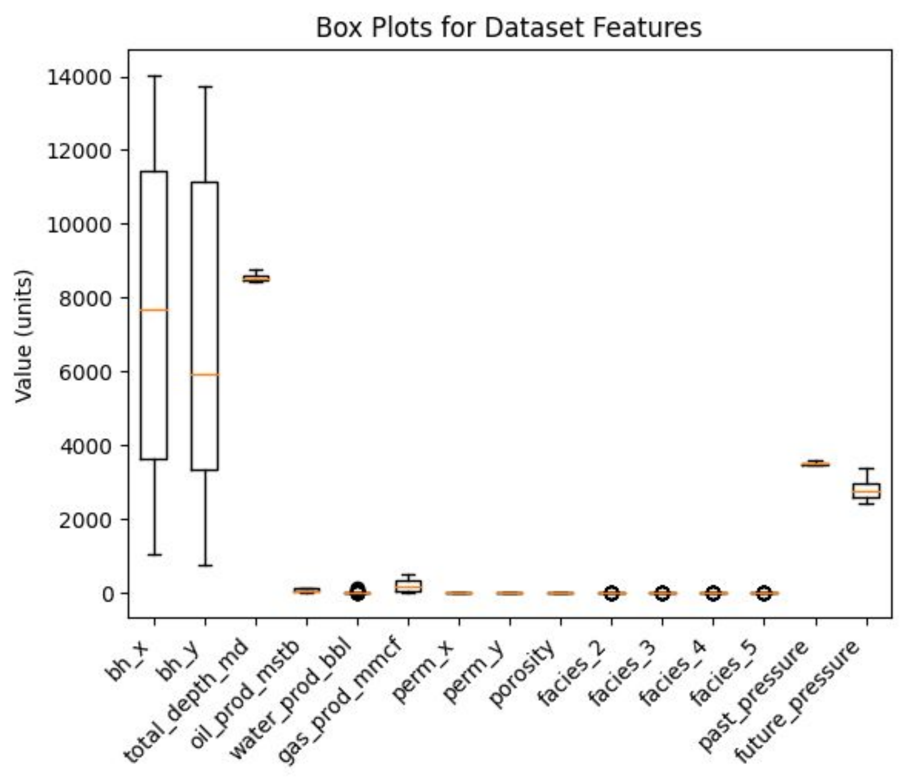


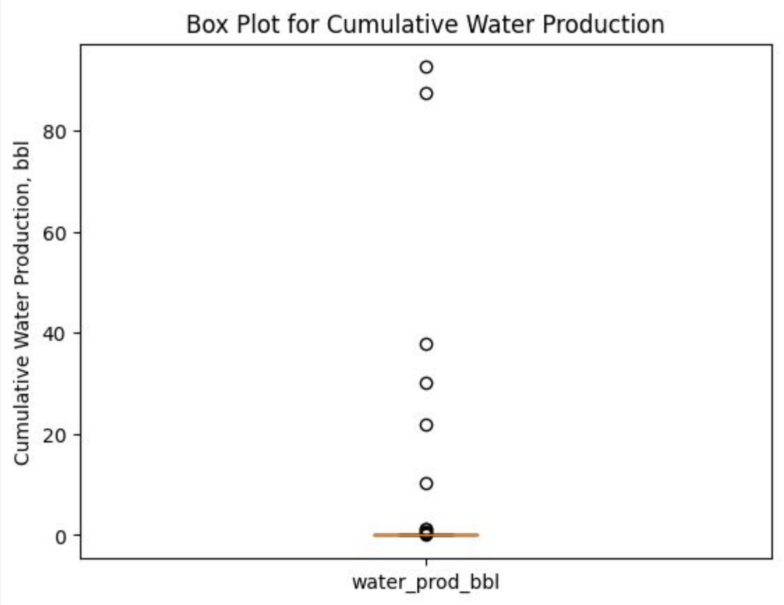
```
# Explore depth-pressure relationship
x = df['total_depth_md']
y = df['past_pressure']
a = ((x - x.mean()) * (y - y.mean())).sum() / ((x - x.mean()) ** 2).sum() # OLS
b = y.mean() - a * x.mean()

print(f'Equation: Pressure = {a:.3f} * Depth + {b:.3f}')
print(f'Actual pressure: {y[0]}, predicted: {a * x[0] + b}')
print(f'Actual pressure: {y[1]}, predicted: {a * x[1] + b}')
```

✓ 0.0s

```
Equation: Pressure = 0.332 * Depth + 654.564
Actual pressure: 3491.971374, predicted: 3491.971374006628
Actual pressure: 3501.327981, predicted: 3501.3279810412205
```

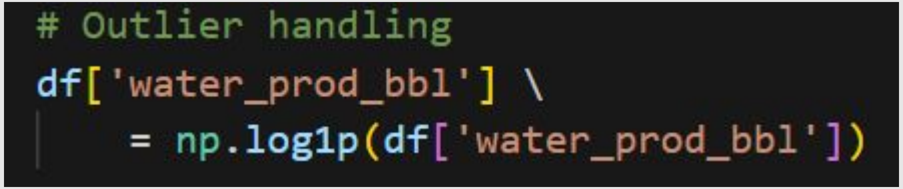
Handling Outliers



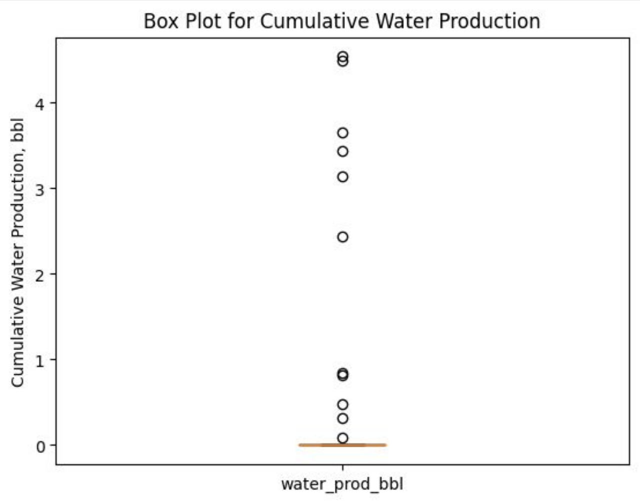


Handling Outliers

- Water production has a lot of outliers (outside IQR).
- log1 transformation was used on water production since many values were close to 0.
- Outliers still exist, but are much less severe.

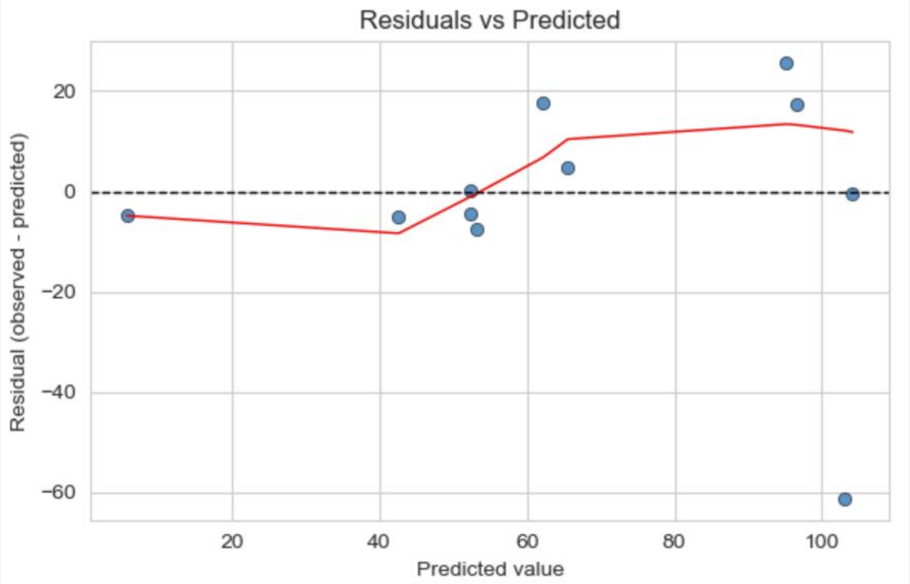


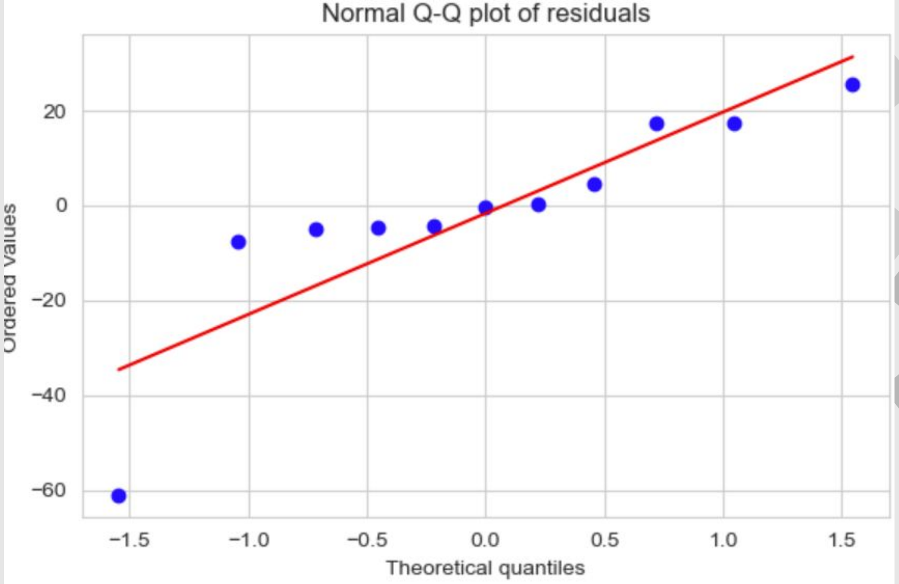
```
# Outlier handling
df['water_prod_bbl'] \
    = np.log1p(df['water_prod_bbl'])
```



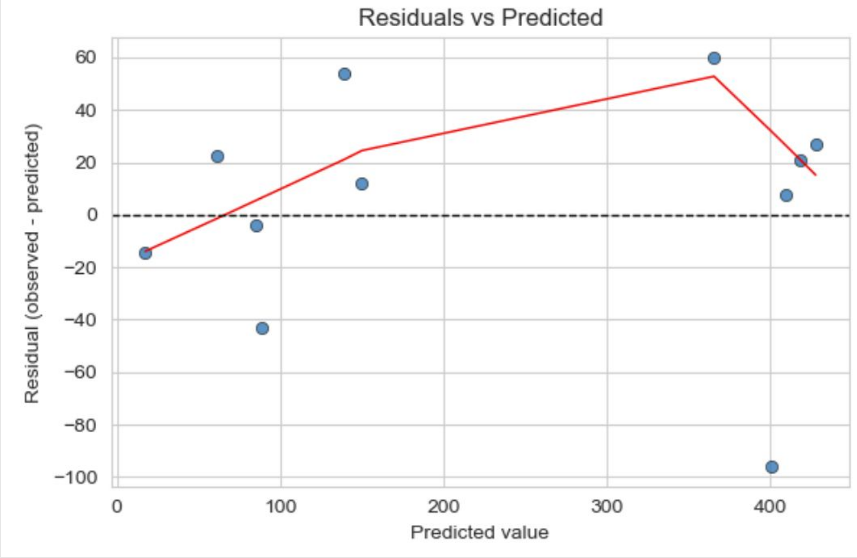
Box Plot for Cumulative Water Production

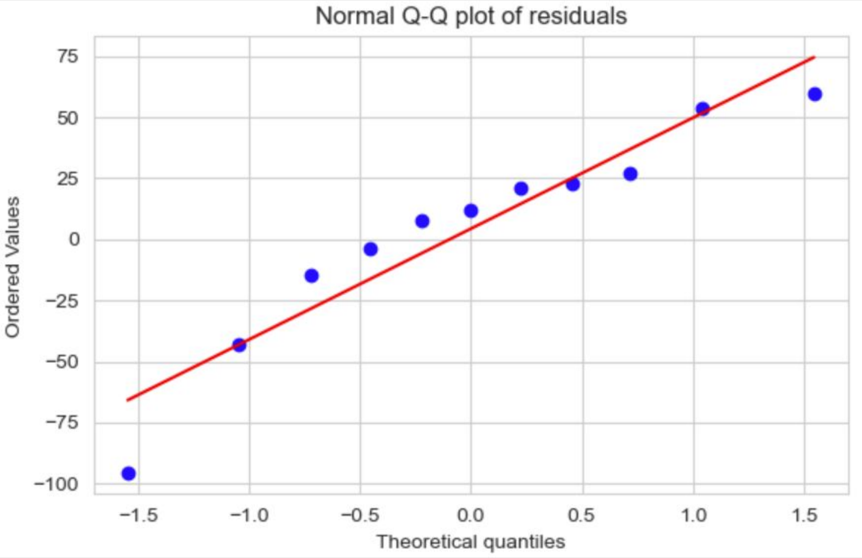
Meeting Assumptions (Oil)

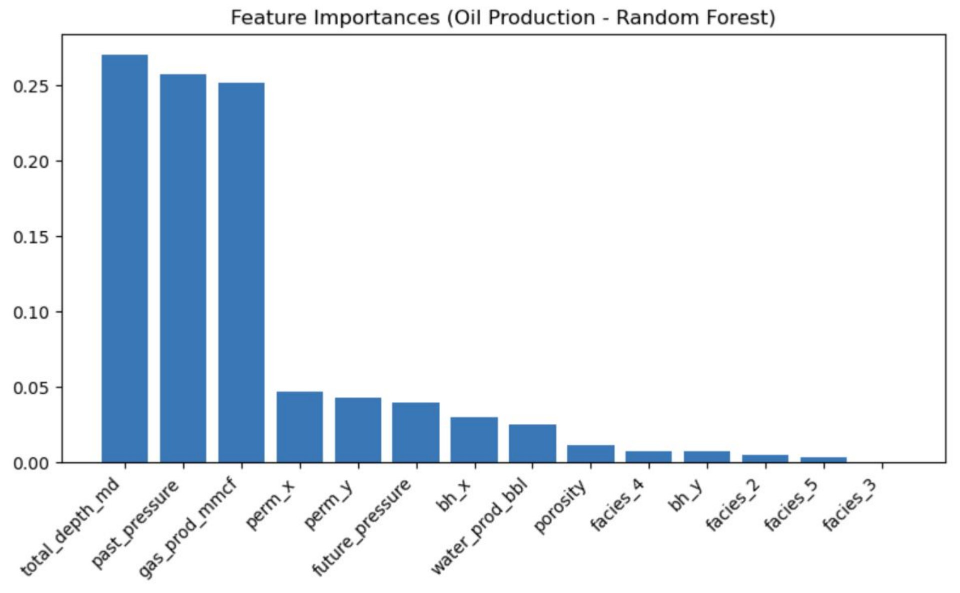


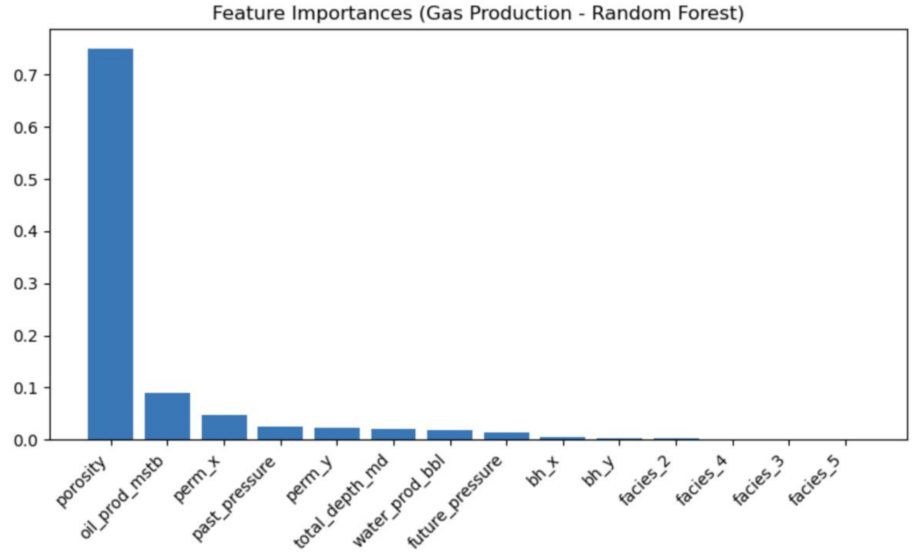


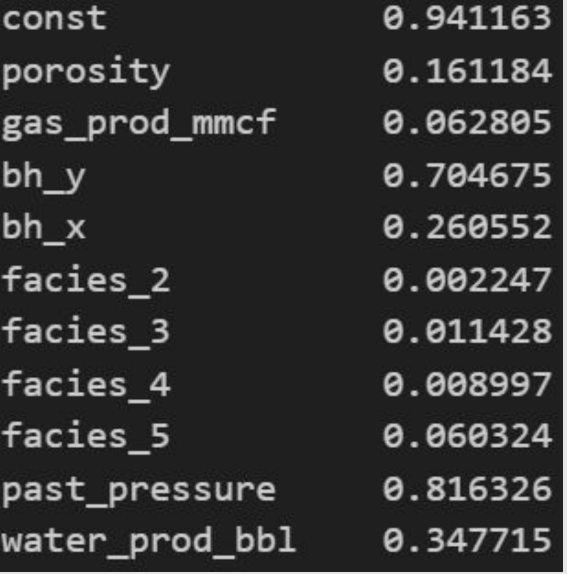
Meeting Assumptions (Gas)



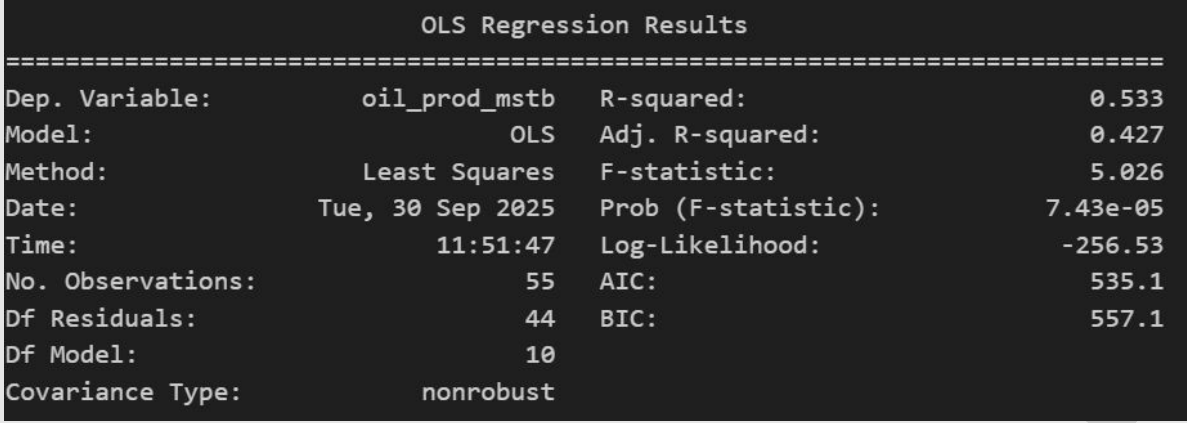



Feature Importances (Oil Production - Random Forest)
0.00
0.05
0.10
0.15
0.20
0.25
total_depth_md
past_pressure
gas_prod_mmcf
perm_x
perm_y
future_pressure
bh_x
water_prod_bbl
porosity
facies_4
bh_y
facies_2
facies_5
facies_3

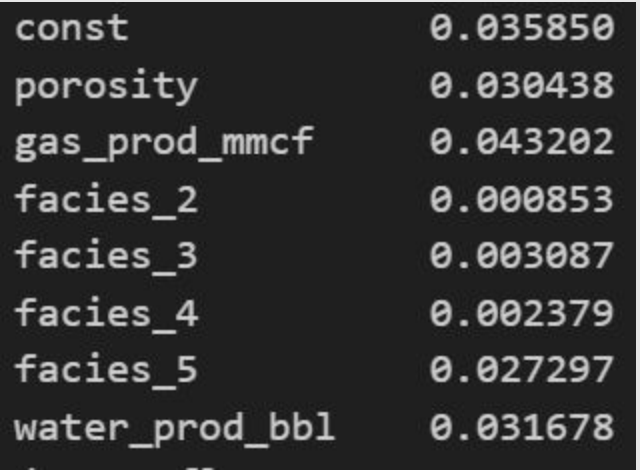

Feature Importances (Gas Production - Random Forest)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
porosity
oil_prod_mstb
perm_x
past_pressure
perm_y
total_depth_md
water_prod_bbl
future_pressure
bh_x
bh_y
facies_2
facies_4
facies_3
facies_5



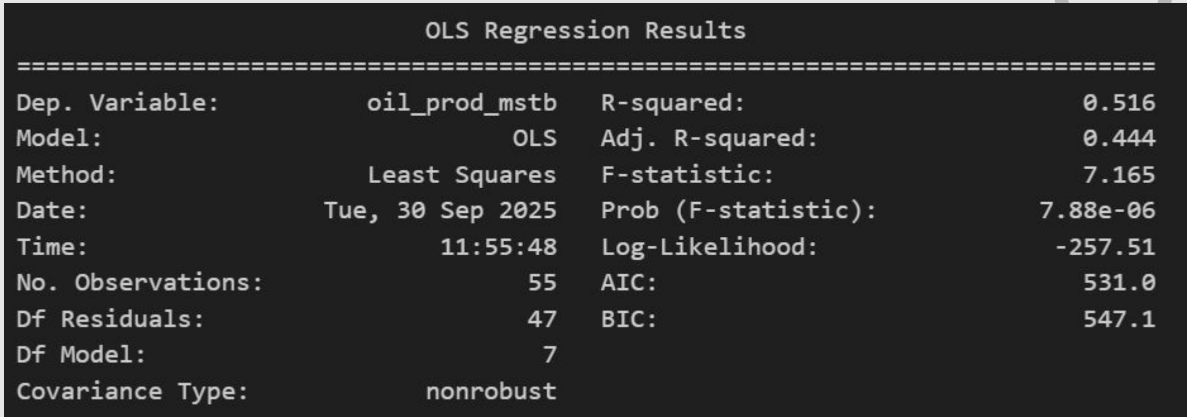
```
const             0.941163
porosity          0.161184
gas_prod_mmcf     0.062805
bh_y              0.704675
bh_x              0.260552
facies_2          0.002247
facies_3          0.011428
facies_4          0.008997
facies_5          0.060324
past_pressure     0.816326
water_prod_bbl    0.347715
```



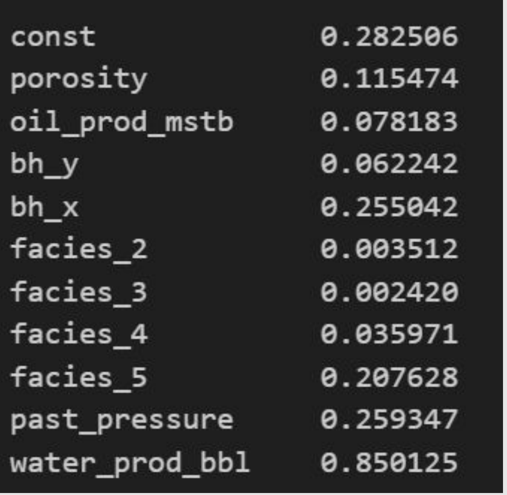
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           oil_prod_mstb   R-squared:                       0.533
Model:                             OLS   Adj. R-squared:                  0.427
Method:                  Least Squares   F-statistic:                     5.026
Date:                 Tue, 30 Sep 2025   Prob (F-statistic):           7.43e-05
Time:                         11:51:47   Log-Likelihood:                -256.53
No. Observations:                   55   AIC:                             535.1
Df Residuals:                       44   BIC:                             557.1
Df Model:                           10
Covariance Type:             nonrobust
```



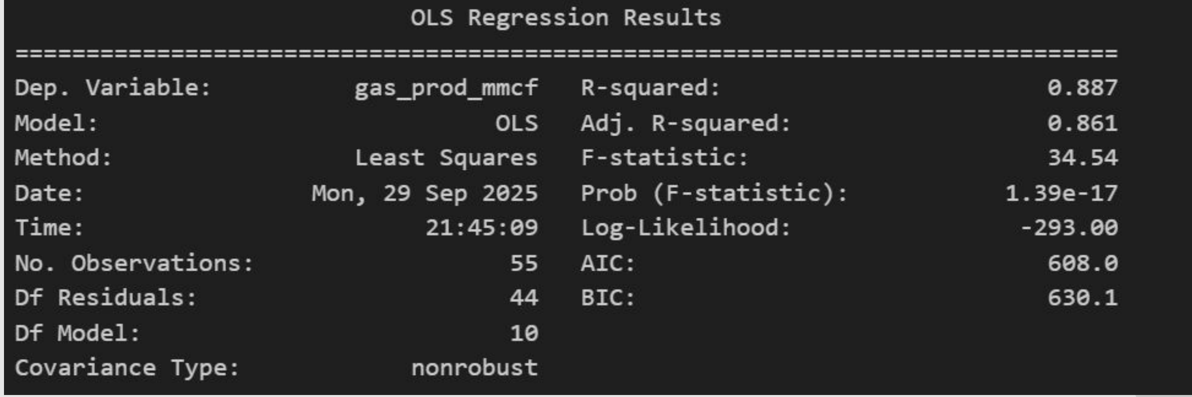
```
const             0.035850
porosity          0.030438
gas_prod_mmcf     0.043202
facies_2          0.000853
facies_3          0.003087
facies_4          0.002379
facies_5          0.027297
water_prod_bbl    0.031678
```



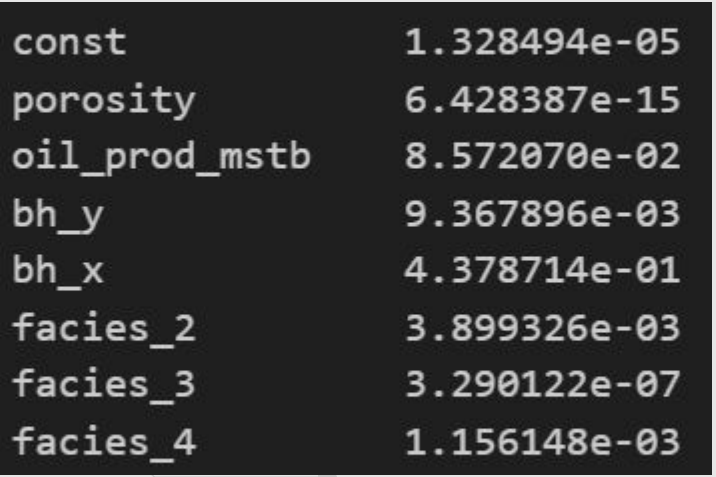
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           oil_prod_mstb   R-squared:                       0.516
Model:                             OLS   Adj. R-squared:                  0.444
Method:                  Least Squares   F-statistic:                     7.165
Date:                 Tue, 30 Sep 2025   Prob (F-statistic):           7.88e-06
Time:                         11:55:48   Log-Likelihood:                -257.51
No. Observations:                   55   AIC:                             531.0
Df Residuals:                       47   BIC:                             547.1
Df Model:                            7
Covariance Type:             nonrobust
```



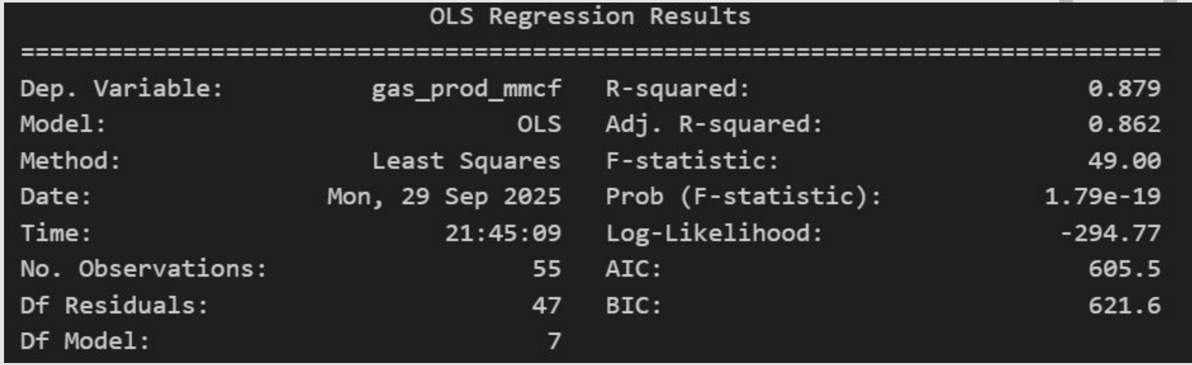
```
const            0.282506
porosity         0.115474
oil_prod_mstb    0.078183
bh_y             0.062242
bh_x             0.255042
facies_2         0.003512
facies_3         0.002420
facies_4         0.035971
facies_5         0.207628
past_pressure    0.259347
water_prod_bbl   0.850125
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           gas_prod_mmcf   R-squared:                       0.887
Model:                             OLS   Adj. R-squared:                  0.861
Method:                  Least Squares   F-statistic:                     34.54
Date:                 Mon, 29 Sep 2025   Prob (F-statistic):           1.39e-17
Time:                         21:45:09   Log-Likelihood:                -293.00
No. Observations:                   55   AIC:                             608.0
Df Residuals:                       44   BIC:                             630.1
Df Model:                           10
Covariance Type:             nonrobust
```



```
const            1.328494e-05
porosity         6.428387e-15
oil_prod_mstb    8.572070e-02
bh_y             9.367896e-03
bh_x             4.378714e-01
facies_2         3.899326e-03
facies_3         3.290122e-07
facies_4         1.156148e-03
```

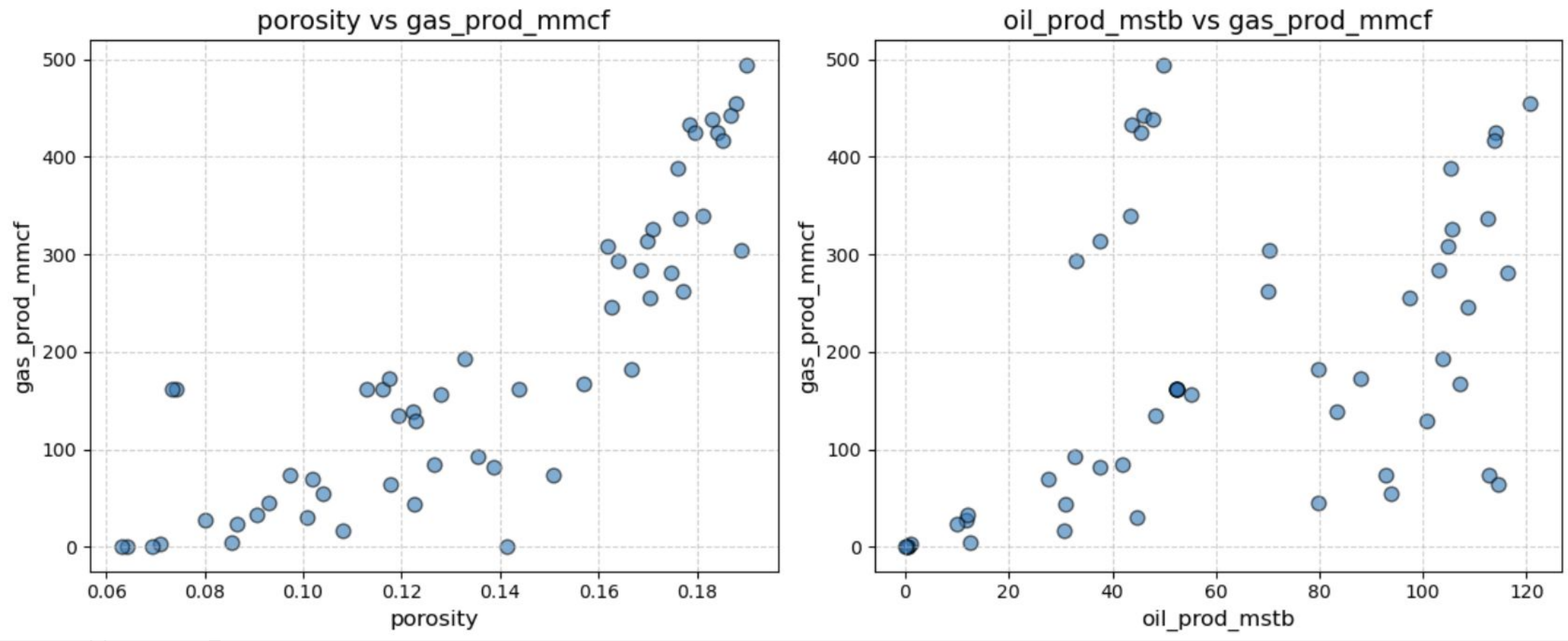


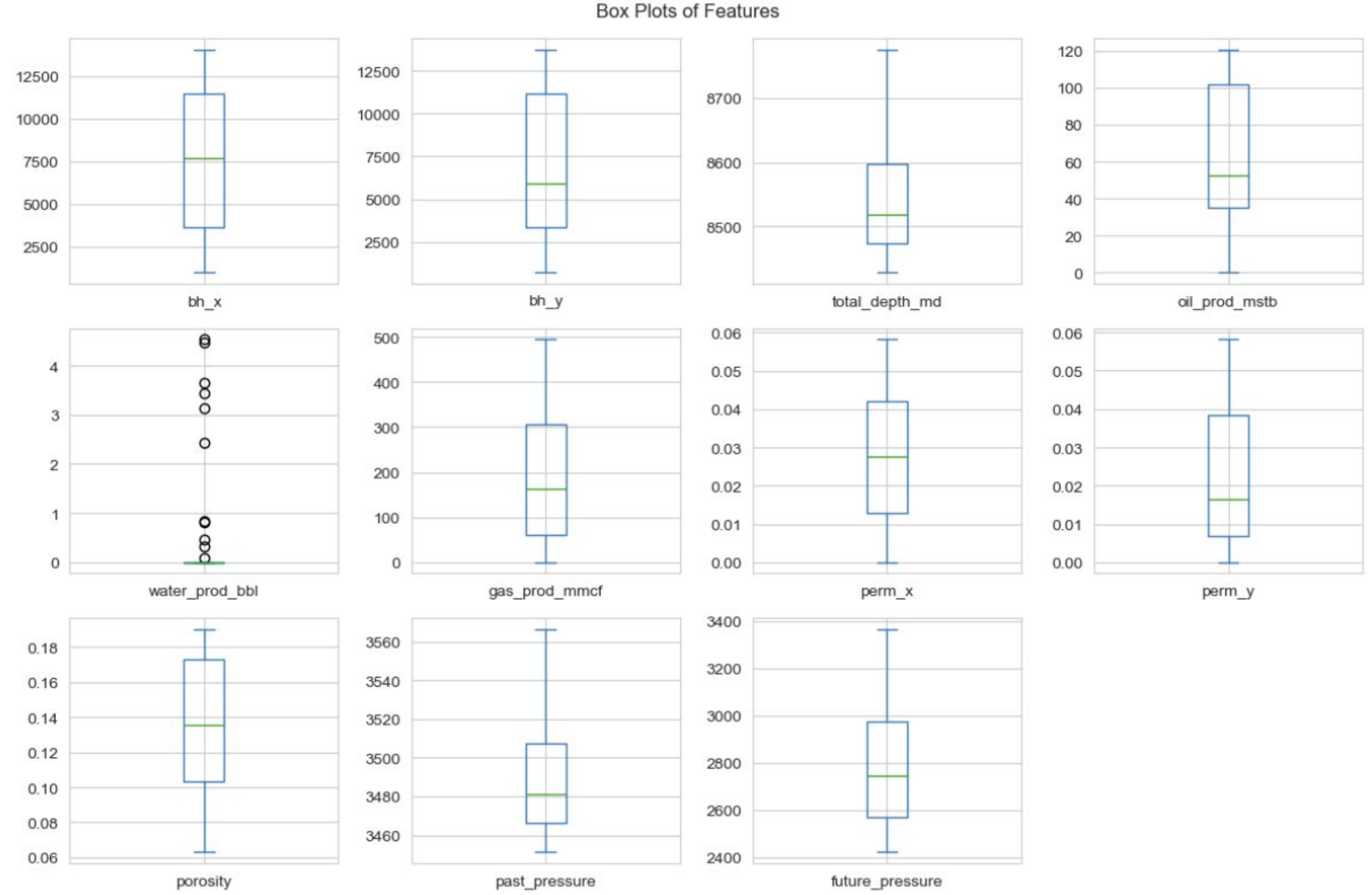
```
                            OLS Regression Results
==============================================================================
Dep. Variable:           gas_prod_mmcf   R-squared:                       0.879
Model:                             OLS   Adj. R-squared:                  0.862
Method:                  Least Squares   F-statistic:                     49.00
Date:                 Mon, 29 Sep 2025   Prob (F-statistic):           1.79e-19
Time:                         21:45:09   Log-Likelihood:                -294.77
No. Observations:                   55   AIC:                             605.5
Df Residuals:                       47   BIC:                             621.6
Df Model:                            7
```

Feature Deletion & EDA

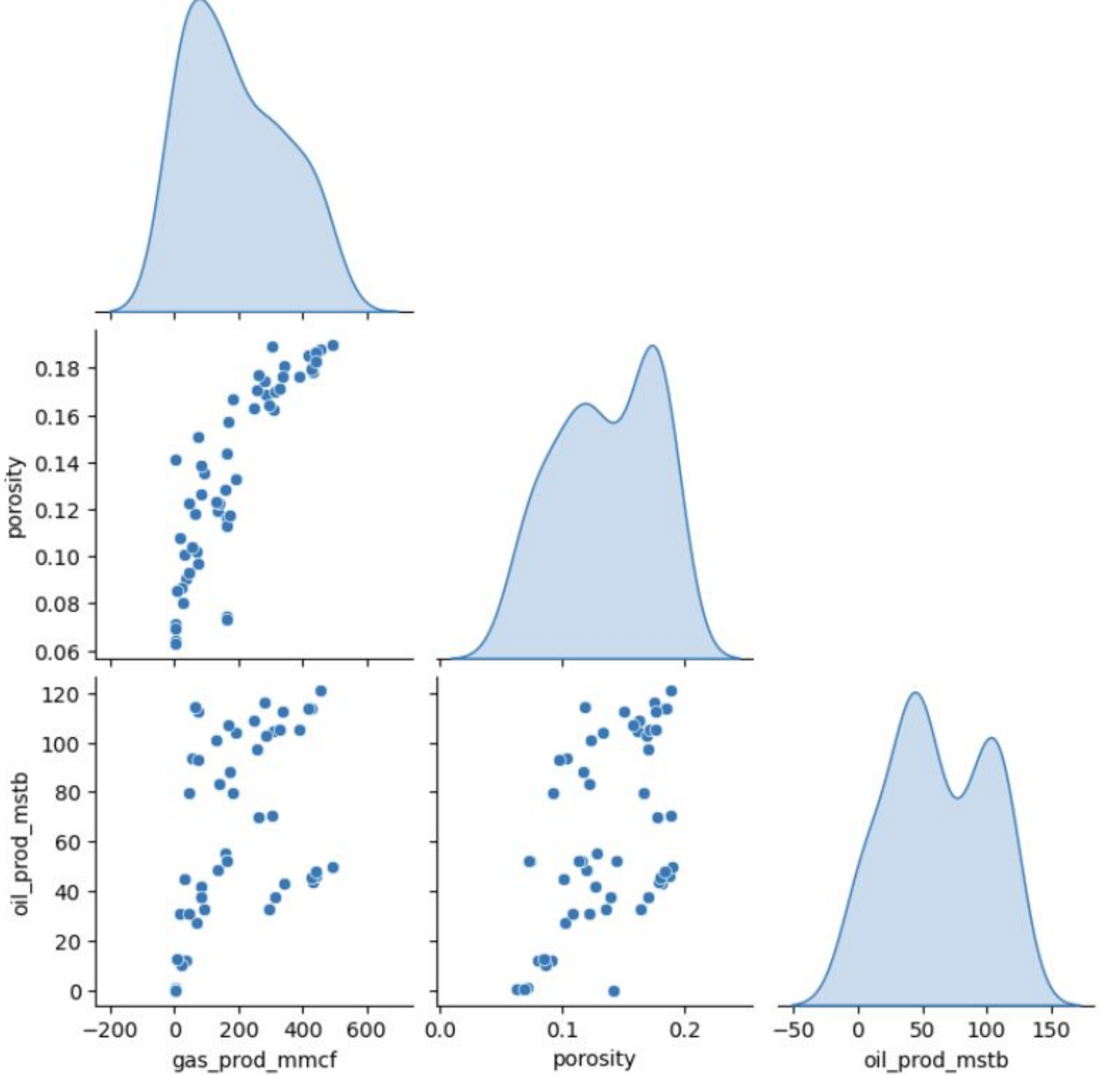
- Well Number and Well Name were both redundant as the index would suffice
- Since our model is predictive, future_pressure should also be removed as using a prediction to make a prediction which would inflate the error
- Facies_5, total_depth_md, water_prod_bbl were removed due to low p-value

Exploratory Data Analysis





Box Plots of Features
bh_x
12500
10000
7500
5000
2500
bh_y
12500
10000
7500
5000
2500
total_depth_md
8700
8600
8500
oil_prod_mstb
120
100
80
60
40
20
0
water_prod_bbl
4
3
2
1
0
gas_prod_mmcf
500
400
300
200
100
0
perm_x
0.06
0.05
0.04
0.03
0.02
0.01
0.00
perm_y
0.06
0.05
0.04
0.03
0.02
0.01
0.00
porosity
0.18
0.16
0.14
0.12
0.10
0.08
0.06
past_pressure
3560
3540
3520
3500
3480
3460
future_pressure
3400
3200
3000
2800
2600
2400


porosity
0.18
0.16
0.14
0.12
0.10
0.08
0.06
oil_prod_mstb
120
100
80
60
40
20
0
−200
0
200
400
600
gas_prod_mmcf
0.0
0.1
0.2
porosity
−50
0
50
100
150
oil_prod_mstb

Data Preprocessing & Scaling

- Other than the log transformation on water production, we did not directly normalize or standardize the dataset.
- This was due to the models used not requiring it.
- More complicated models in the future will most likely require standardization or normalization.

Challenges

- **Small dataset (55 wells, 14 features)** → risk of overfitting
- **Missing data (~9%)** → requires careful imputation to avoid bias
- **Feature redundancy/multicollinearity** → permeability & porosity highly correlated
- **Outliers in production values (very high water production)** → may skew regression
- **Facies categorical data** → limited contribution, removed from analysis

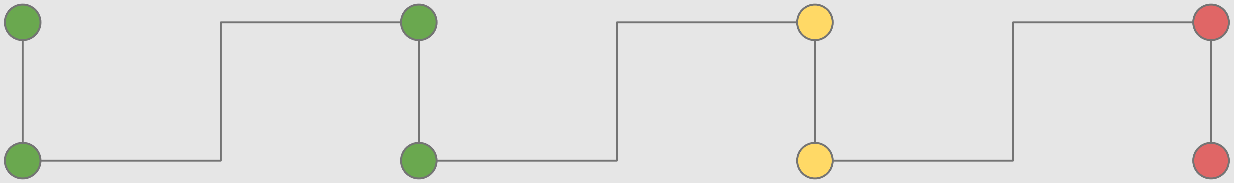
Potential Solutions

- Test multiple imputation methods for missing data
- Use cross-validation + regularization to address small dataset/overfitting
- Handle outliers carefully → detect statistically, but preserve geologically valid extremes
- Engineer new features or use dimensionality reduction to strengthen signals

Model Fine-Tuning

- Use cross-validation to tune models due to small dataset size (55 wells).
- Adjust regularization strength (alpha) in a Ridge and Lasso regression to handle multicollinearity
- Compare performance across models using R^2, RMSE, and MAE.
- Select final model based on balance of accuracy and interpretability.

Next Steps



1.Data Preparation

- Collect petrophysical, production and spatial datasets
- Clean and preprocess data
- Validate geological ranges

2.Exploratory Analysis

- Visualize distributions and correlations between variables
- Refine data preprocessing
- Identify outliers and production anomalies
- Build foundational models

3.Intermediate Research

- Model Fine-Tuning
- Assess importance of porosity, permeability, facies, and depth
- Compare oil vs. gas production patterns
- Investigate effects of pressure drawdown on performance
- Validate results with cross-validation due to small dataset size

4.Integration & Insights

- Combine geological, production, and spatial analyses
- Identify sweet spot clusters on field map
- Interpret underperformance to optimize solution

THANK YOU!

Any questions