Datathon 2024: Well, Well, (Oil) Well

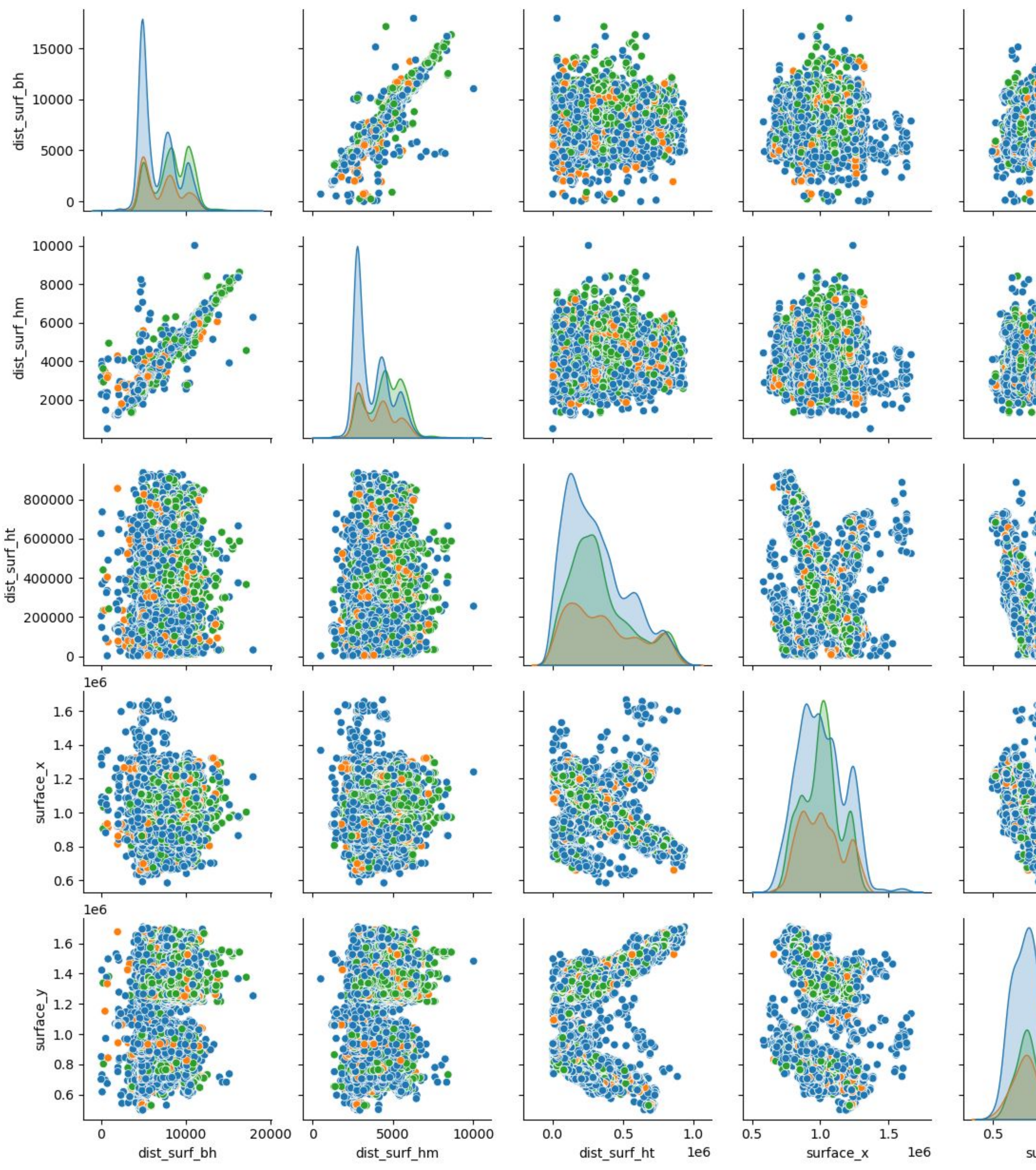
Charlie Liu, Bayzhan Mukatay, Akshay Raj, Daniel Suarez

AGENDA

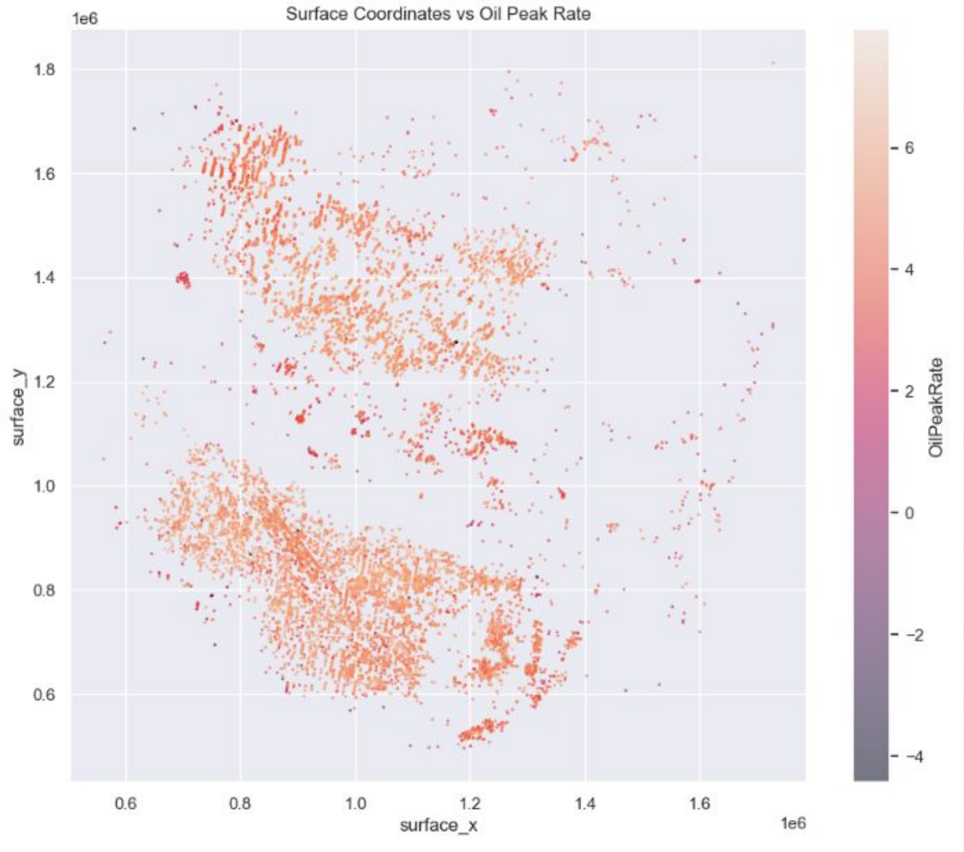
1 DATA EXPLORATION, WRANGLING, & ENGINEERING

2 FEATURE & MODEL SELECTION/TUNING

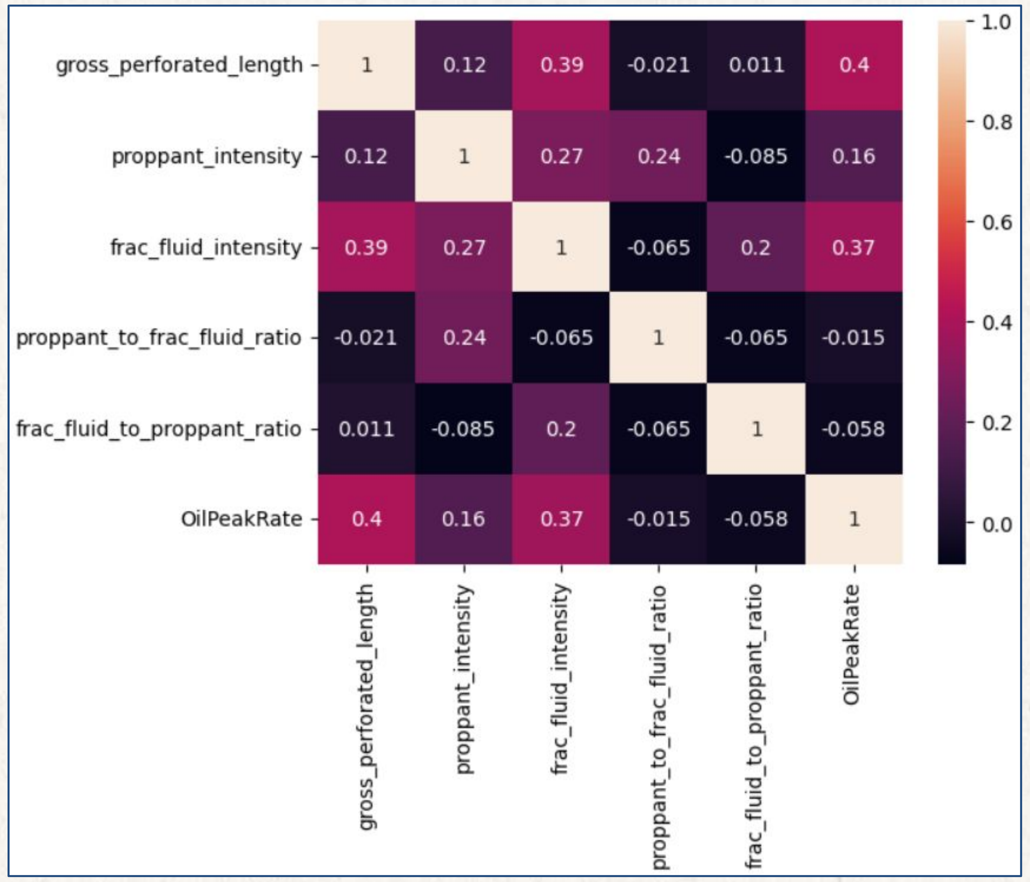
3 FINDINGS & ANALYSIS



Data Processing



Exploration



Feature Engineering

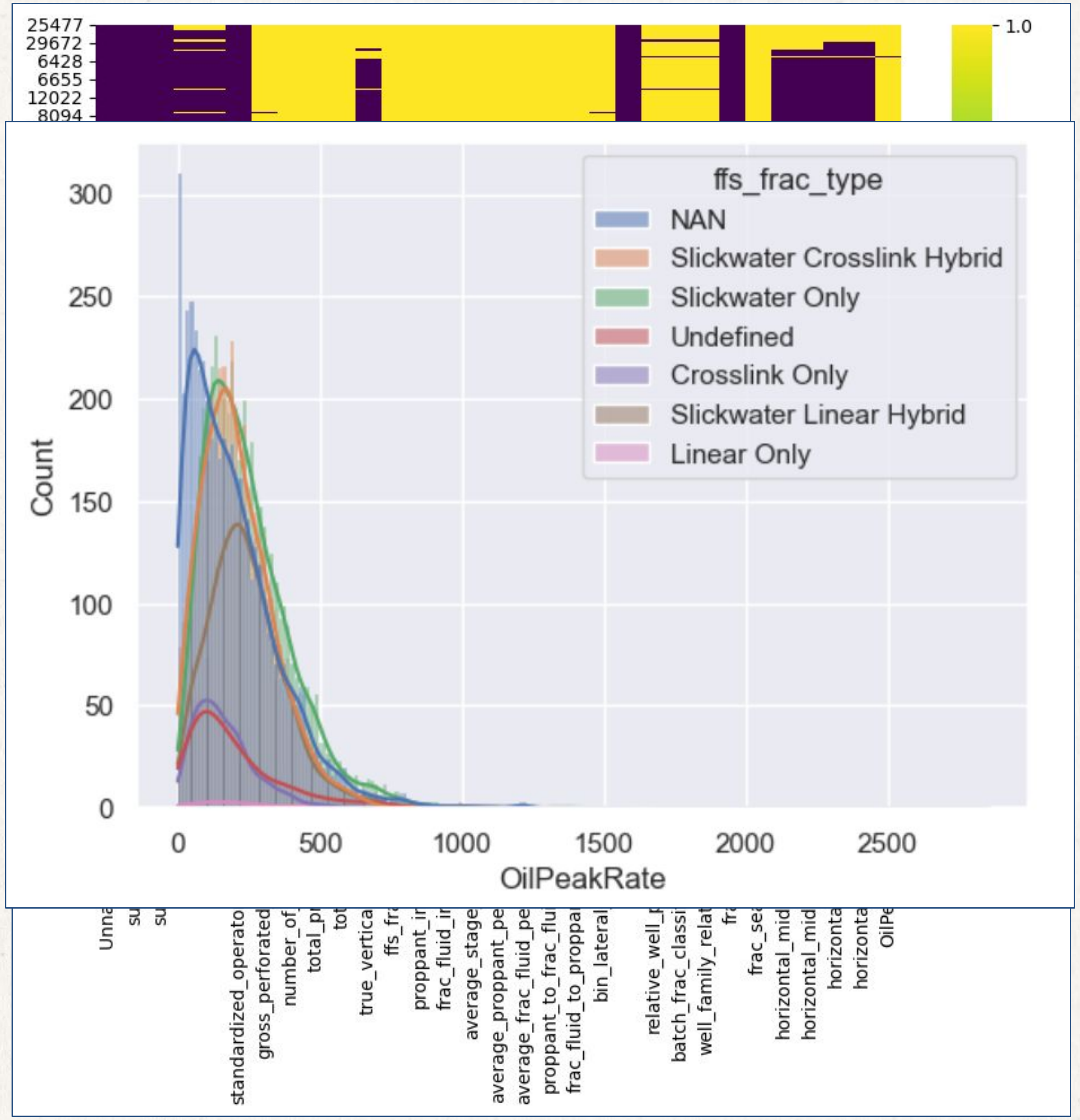
	145.2952	810.7745				0.179205	5.580189	1
Slickwater	210.5523	586.4838				0.359008	2.785454	1
Slickwater	214.7102	579.4774				0.370524	2.698882	1
	175.1102							1
	178.0678							1
	121.3453							1
	124.7229							1
Slickwater	180.6807	931.5449				0.193958	5.155753	1
Slickwater	177.6782	950.1917				0.186992	5.347823	1
Slickwater	189.9266	590.4277				0.321676	3.108715	1
Undefined	111.4279	591.2971				0.188446	5.306547	1.5
Undefined	106.5047	721.6284				0.147589	6.775555	1
	109.9499							2.5
	110.8576							2
Slickwater	216.7911							1
Slickwater	73.42521	456.0039				0.161019	6.210454	1
Slickwater	73.45889	461.6153				0.159134	6.283995	1
	144.5803	752.189	154.3166	22311.14	116075.3	0.192213	5.20257	1
Slickwater	146.3817	826.3251				0.177148	5.645004	1
Slickwater	113.2011	540.1978				0.209555	4.77202	1
Slickwater	111.3554	441.3129	195.481	21767.86	86268.3	0.252327	3.963105	1
	113.2617							1.5
	113.7462	490.5505				0.231875	4.312675	2
	34.28057	203.6173				0.168358	5.939729	1.5

Wrangling

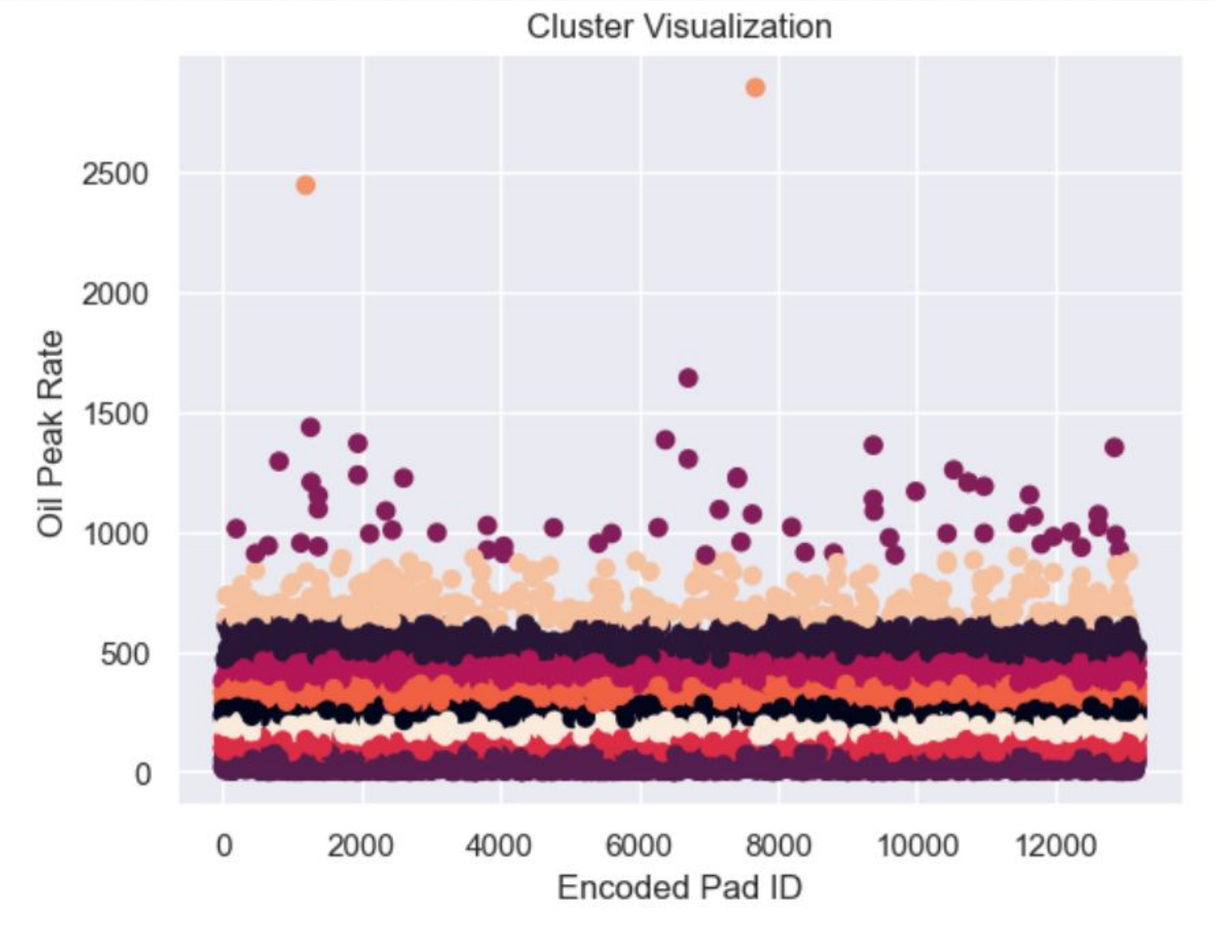
Data Exploration

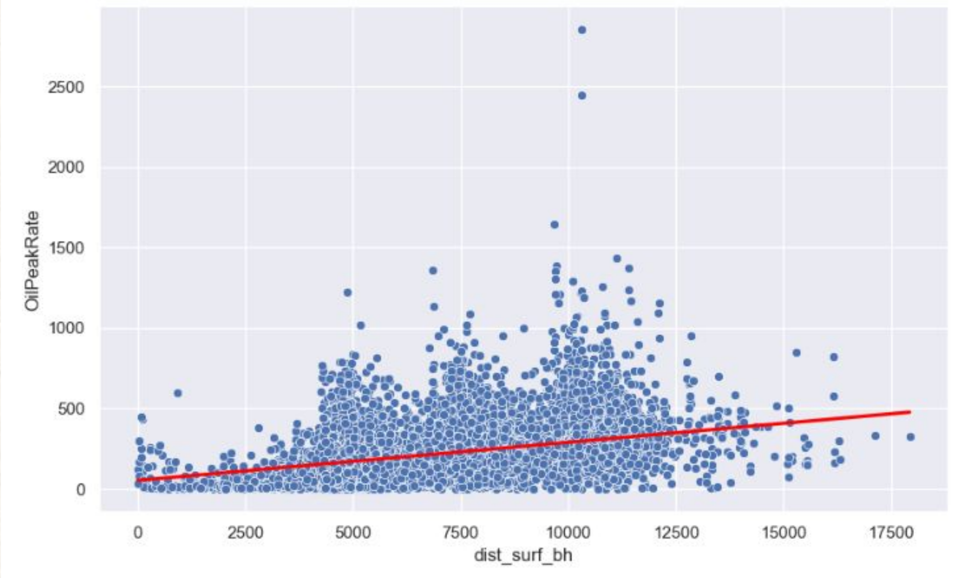
In exploring the data, we identified several areas of interest and concern and started brainstorming:

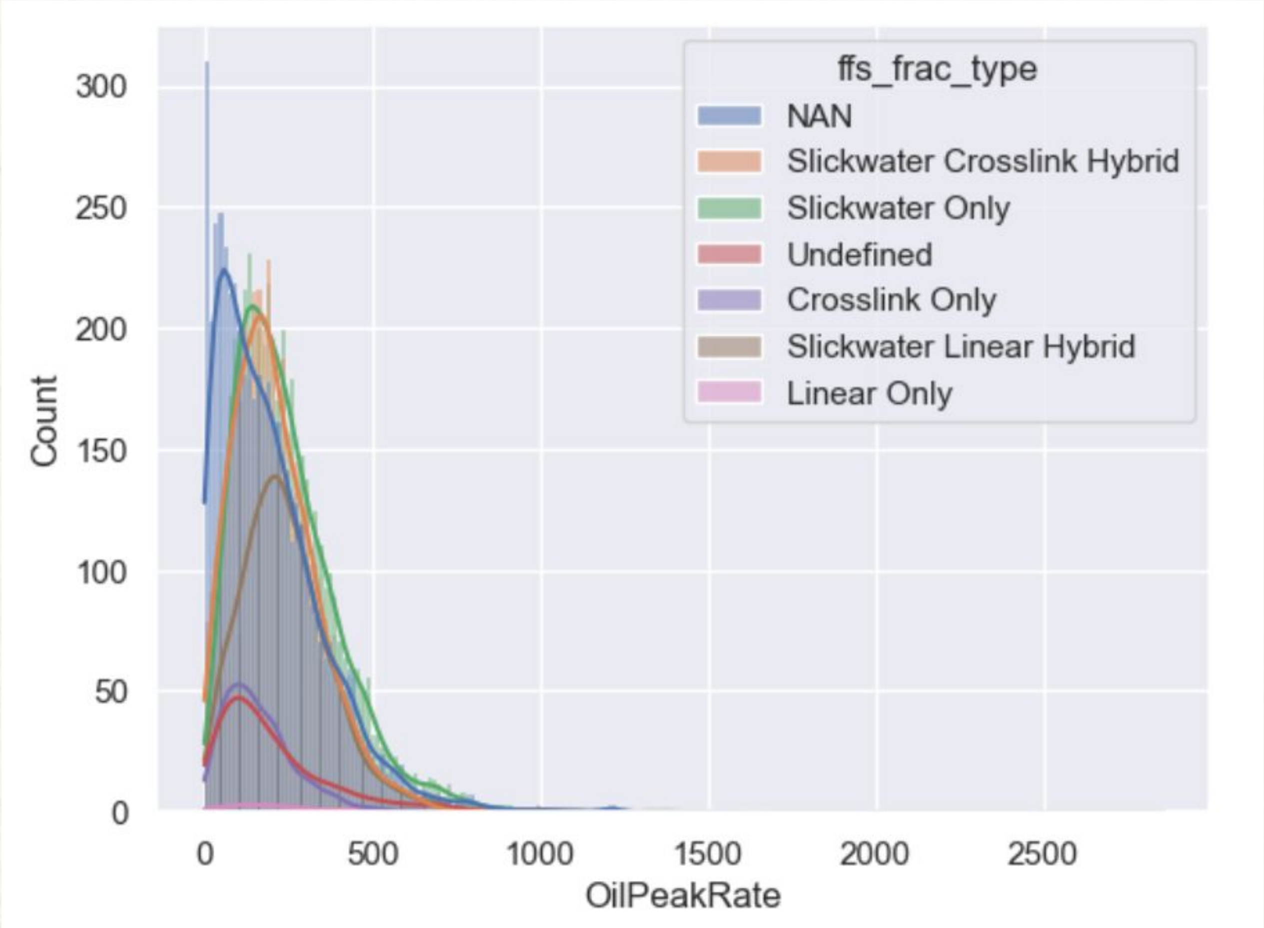
1. **Handling rows with NaN’s in the target column?**
    a. Keep these rows temporarily (may be valuable for imputation), then drop them when it's time to train
2. **Several features with too few observations (<14%)**
    a. Drop these columns entirely
3. **Working with categorical features**
    a. “Unknown” observations show a distinct distribution from NaN observations



Data Exploration

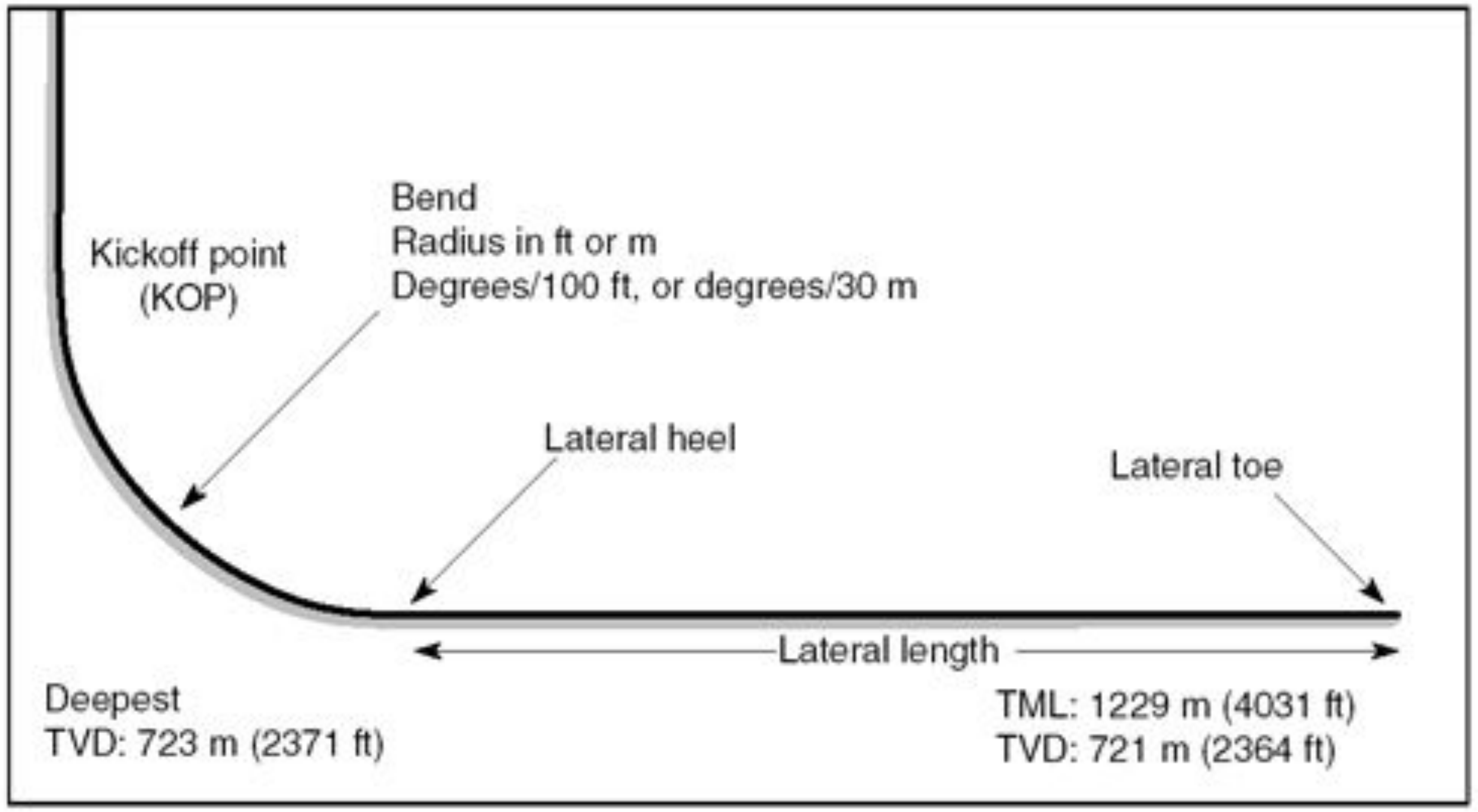






Feature Engineering

In addition to the cleaning that needs to be done, our data exploration process inspired us to engineer new and existing features in ways that best suits the provided dataset:



`projection_group` - project `'relative_well_position'` onto `'well_family_relationship'`

`dist_surf_bh` - the distance from the bottom hole to the surface hole

`dist_surf_hm` - the distance from the surface hole to the horizontal midpoint

`dist_surf_ht` - the distance from the surface hole to the horizontal toe

`log_[...]` - the natural log of certain features for which a logarithmic distribution reduces the skewness of the feature

Wrangling – Cleaning

Given what we found and had planned, we were ready to start appropriately cleaning our data:

1. **Drop** infinity observations, **drop** outliers
2. **Drop** columns that offer very little (too few observations, no variation, etc.)
3. **Relabel** certain NaN and Unknown observations for categorical features
4. **Take the natural log** (new columns) of certain numeric features
5. Create our training-test-validation set splits (70% train, 15% test, 15% validation)
6. **Drop rows** with too many (>86%) NaN observations (will be unhelpful for imputation)

Wrangling – Imputation!

Now with clean data, we were able to perform imputation to further improve our predictive power:

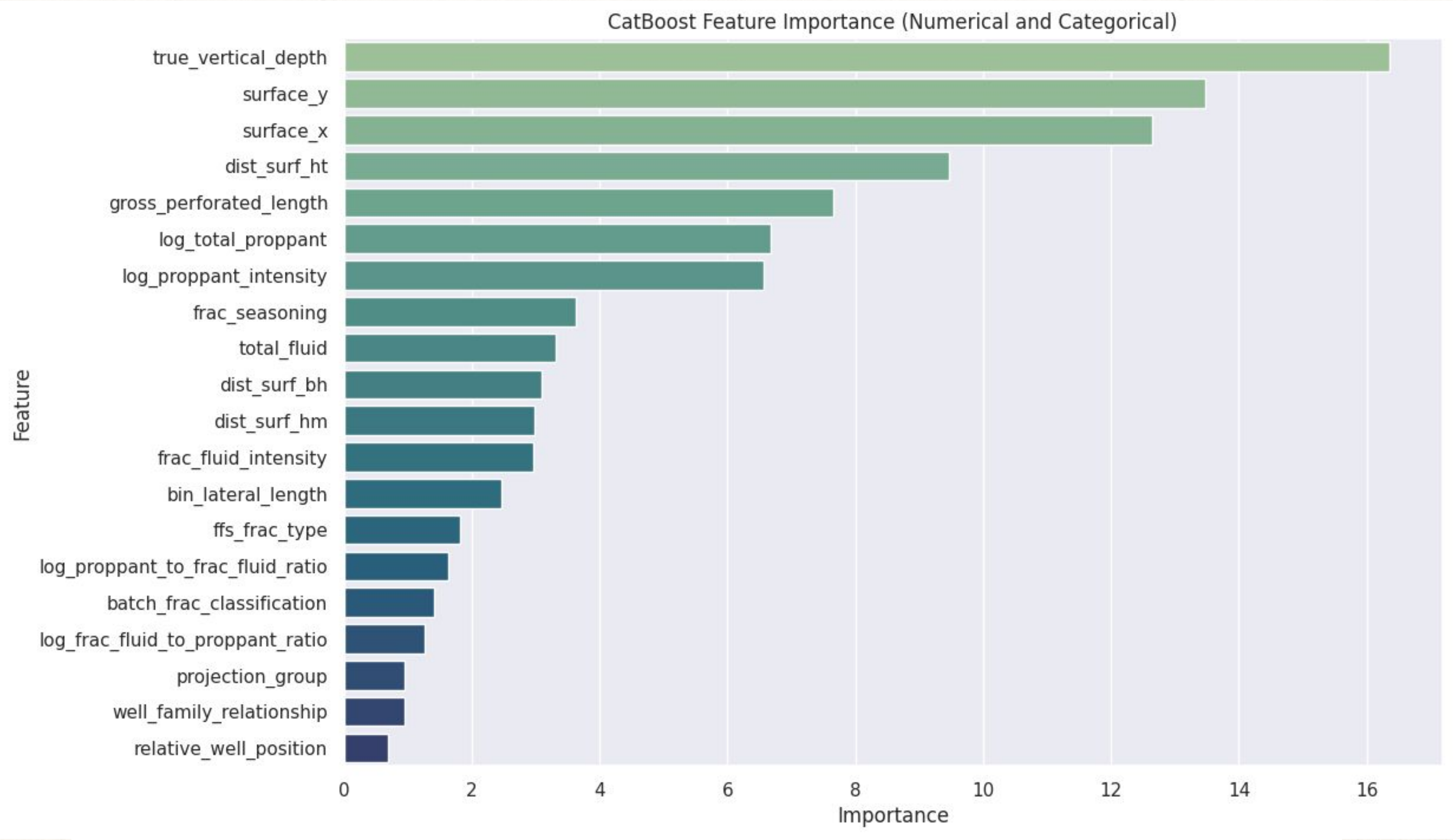
- We began with simple (mean) imputation for our new log features
- Then, we proceeded to perform predictive imputation (with a linear regression model) to impute missing values for the rest of our features
- Sets were imputed separately to avoid leakage

Finally, we were left with clean, processed data with no missing observations!

	145.2952	810.7745				0.179205	5.580189	1
Slickwater	210.5523	586.4838				0.359008	2.785454	1
Slickwater	214.7102	579.4774				0.370524	2.698882	1
	175.1102							1
	178.0678							1
	121.3453							1
	124.7229							1
Slickwater	180.6807	931.5449				0.193958	5.155753	1
Slickwater	177.6782	950.1917				0.186992	5.347823	1
Slickwater	189.9266	590.4277				0.321676	3.108715	1
Undefined	111.4279	591.2971				0.188446	5.306547	1.5
Undefined	106.5047	721.6284				0.147589	6.775555	1
	109.9499							2.5
	110.8576							2
Slickwater	216.7911							1
Slickwater	73.42521	456.0039				0.161019	6.210454	1
Slickwater	73.45889	461.6153				0.159134	6.283995	1
	144.5803	752.189	154.3166	22311.14	116075.3	0.192213	5.20257	1
Slickwater	146.3817	826.3251				0.177148	5.645004	1
Slickwater	113.2011	540.1978				0.209555	4.77202	1
Slickwater	111.3554	441.3129	195.481	21767.86	86268.3	0.252327	3.963105	1
	113.2617							
	113.7462							

1.549509793	Missing	Outer Well	Unknown	Infill Child Well
1.443013978	Slickwater Only	Inner Well	Batch-Concurrent Fr	Sibling Well
1.637200683	Missing	Outer Well	Unknown	Sibling Well
1.973418292	Slickwater Crosslink	Standalone Well	Non-Batch Frac	Standalone Well
1.996375388	Slickwater Only	Outer Well	Batch-Sequential Fr	Sibling Well
1.811831438	Slickwater Only	Inner Well	Batch-Concurrent Fr	Sibling Well
1.858592964	Slickwater Linear Hy	Standalone Well	Non-Batch Frac	Standalone Well
1.633899034	Missing	Standalone Well	Unknown	Standalone Well
2.381480762	Slickwater Linear Hy	Standalone Well	Non-Batch Frac	Standalone Well
1.463346667	Slickwater Only	Outer Well	Batch-Concurrent Fr	Sibling Well
1.659328025	Slickwater Only	Inner Well	Batch-Sequential Fr	Sibling Well
1.857269141	Missing	Standalone Well	Unknown	Standalone Well
1.325478649	Missing	Outer Well	Unknown	Sibling Well
1.938760138	Slickwater Only	Outer Well	Batch-Concurrent Fr	Sibling Well
1.832652523	Slickwater Crosslink	Inner Well	Batch-Concurrent Fr	Sibling Well
1.286981908	Slickwater Only	Inner Well	Batch-Concurrent Fr	Sibling Well
1.597587413	Slickwater Only	Inner Well	Batch-Sequential Fr	Sibling Well

Feature Selection



Model Selection

- Logistic Regression
- Lasso
- XGBoost
- CatBoost
- LightGBM
- Model Stacking

Findings!

- Different successful models determined slightly different sets of features to be most important
  - There were some features that we determined to be very important to all our models
  - **Gross perforated length**, **true vertical depth**, and **proppant intensity**, **surface-toe distance**, among others
  - Proppant-related features were deemed important (positively correlated), significantly more so than fluid

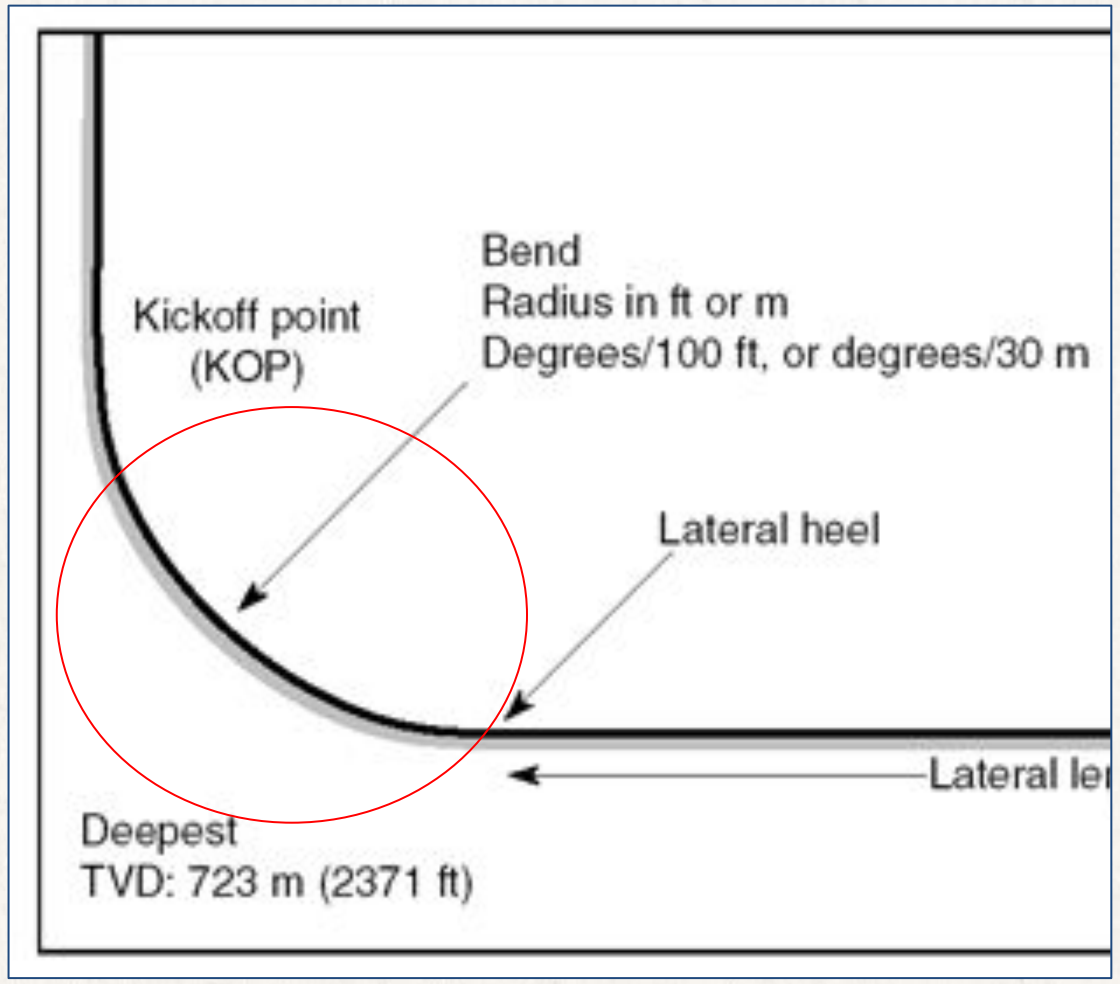
- We found overall that the well structure is a key contributor to oil production, especially considering depth

- High proppant-related metrics, especially intensity, seem to also greatly correlate with an increased peak production

Future Exploration

Engineer more features:

- Given the feature importance findings, better data collection and further exploration of additional imputation methods is valuable.
- We'd like to start observing the **angle of the well decline**
  - The importance of true vertical depth, fluid/proppant intensity, and surface-toe distance makes this clear
- We'd like to consider more geographical (potentially deanonymized) data to be able to cluster wells by location
- A deeper understanding of how Oil Wells work could provide with insight for engineering new features



THANK YOU!

2024 RICE DATATHON

Chevron