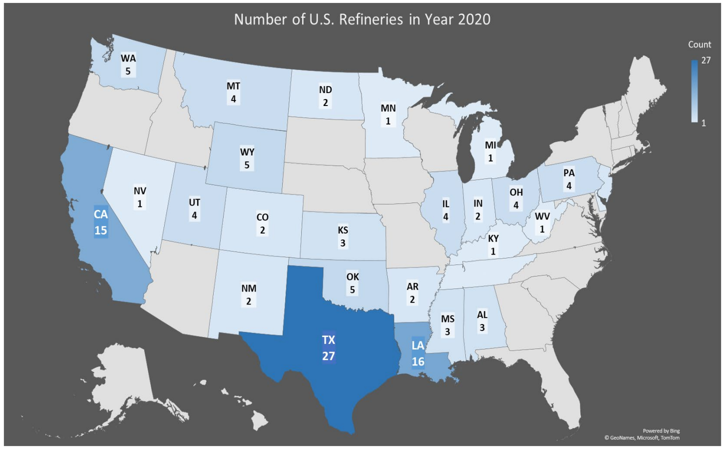
Selecting best greenfield location for a new U.S. refinery

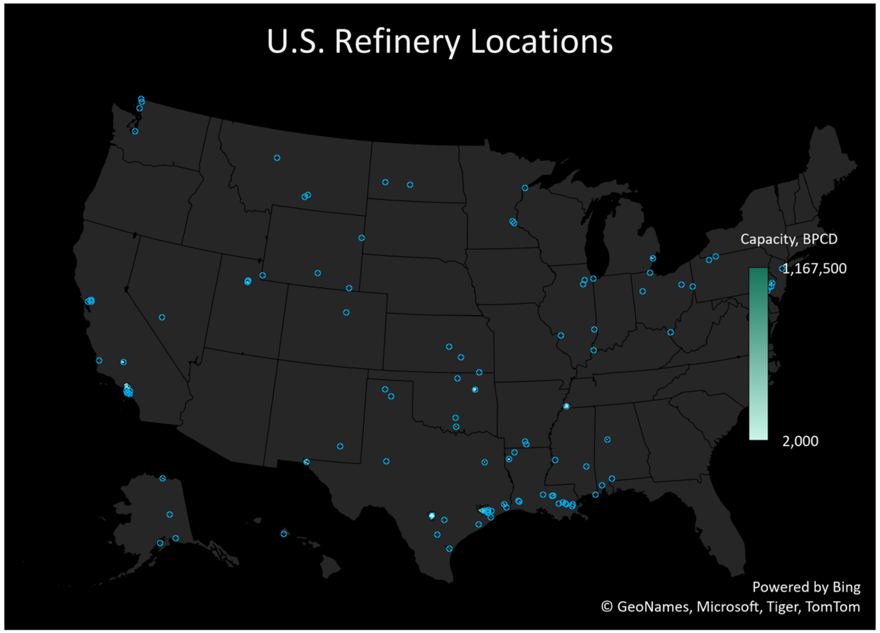
By Aibek Uraimov

What is the best spot for building a new greenfield refinery in the U.S.?

Current count of U.S. refineries by state



Locations of U.S. refineries



by operating capacity

(barrels per calendar day)

Problem Identification

Greenfield Refinery Feasibility Validation

Greenfield Refinery Problem Statement

Develop a data-driven methodology to validate economic feasibility study of building and operating a small (under 50kBpd) crude oil refinery in the U.S., using datasets from the U.S. Census and EIA.gov websites. Identify 2-3 key parameters that have high impact on the economic performance of a refinery, located in the pre-selected PAD-2-2 district, optimizing for anticipated gross margin ($/Bbl crude oil refined), based on at least 8 years of historical data

Context

Green Light Co. (GLC), a California startup, wishes to validate its preliminary economical model for building a greenfield crude oil refineries in the U.S. It recognizes that no new refinery with a significant downstream capacity has been built in the U.S. in 40 years; the existing refinery plants are relying on outdated technology to produce petroleum products.

GLC preliminary has selected several potential locations near major sources of domestic crude oil supply and theorizes that these locations will bring the highest profit margins. I developed a methodology that could be used to forecast the margins, based on the geographical location of a proposed refinery site, quality of crude oil (feedstock), plant equipment configuration, and proximity to destination markets.

Criteria for success

The criterion for success is to validate the anticipated gross profit margins in the range of $25-$30 per 1 U.S. barrel of crude oil processed, given the pre-selected sites for new refinery construction, type of crude available, and proximity to destination markets.

Scope of solution space

Historical data on energy production and consumptions - we will use 5-year datasets:

Scope of solution space

Crude Oil production + Net Imports and Crude Oil price indices - WTI for the U.S., and Brent for the global markets, to determine correlation with the prices of petroleum products produced by refineries and net imports

Historical refinery operations parameters: average plant utilization rates across the U.S., product yields by input crude and refinery configuration types, and volumes produced - to determine indicative profit margins for the existing refineries and build comparative analysis

Constraints within solution space

Some of the constraints on data availability and level of accuracy I had to accept were as follows:

- Much of the actual refinery sale prices for ready products is not publicly available data; there are agencies that sell historical petroleum pricing data, but their costs were prohibitive for an academic exercise; we will just have to rely on the EIA.gov data that is the next best thing
- There's uncertainty on the gasoline and possibly, diesel demand, not due to politics and government, but the free market, i.e., an anticipated proliferation of Electrical Vehicles within the next 10-15 years – the minimum runway for refinery project investment returns
- Finally, the COVID-19 related lockdowns have demonstrated that a significant office-based workforce might successfully transition to remote work attendance practices, leading to lower commute and possibly, lower consumption of petroleum-based fuels

Greenfield Refinery Problem Statement

Develop a data-driven methodology to validate economic feasibility study of building and operating a small (under 50kBpd) crude oil refinery in the U.S., using datasets from the U.S. Census and EIA.gov websites. Identify 2-3 key parameters that have high impact on the economic performance of a refinery, located in the pre-selected PAD-2-2 district, optimizing for anticipated gross margin ($/Bbl crude oil refined), based on at least 8 years of historical data

Constraints within solution space

- Great level of uncertainty in the future of gasoline markets due to politically driven decisions by the U.S. government, which may have a negative impact on the U.S. energy sector, e.g., the ‘Carbon Free by 2035’ plan

Key data sources

I used datasets obtained from the US Census and EIA.gov websites to identify key metrics:

- major metro areas and specific region population
- historical energy production and consumption rates
- historical energy source composition dynamics (oil & gas, solar, hydro, nuclear, etc.)
- refinery operational throughput dynamics (to determine trends over the last 10 years)
- refinery product yields and utilization rates - to assess their historical profitability
- crude oil actual (refinery) crack-spreads by type of refining configurations,
- finally, predictive price modeling for ready products, in relation to crude oil price indices

I abandoned some items, as they were either not very useful for narrowing down the solution scope, or were becoming too large of a task, i.e., outside of the stated scope:

- (abandoned) Make-up of energy produced and consumed by source - to determine where the sustainable production of crude oil based refineries are the strongest and to analyze the patterns of consumption by regions
- (abandoned) Amount of energy generated, as well as its net imports into the U.S., measured in British Thermal Units
- (abandoned) Amount of energy consumed in the U.S., also measured in British Thermal Units

The Data

Greenfield Refinery Feasibility Validation

Data Sourcing

Raw Data

After reviewing over 20+ raw datasets the narrowed down list included:

- Refinery specific: geographical location, PADD designation, operating capacity by process type and equipment; product yields by % volume, crude oil acquisition prices with implied discounts or premiums
- Markets specific: population of the U.S. states and major metro areas; ready petroleum products wholesale resale prices by U.S. refineries; domestic WTI and European Brent indices

Initial Data Preparation

Some initial data preparation included:

- Categorizing destination PADD markets as ‘major metro’ (population > 300K)
- Calculating distances from refineries to major cities
- Adding abstract measures of ‘refinery supply pressure’ – daily production volumes targeting markets and ‘market resistance’ – daily consumption caps, based on population density and the intensity of supply from competing refineries targeting the same markets

Sources and References

eia *Independent Statistics & Analysis* U.S. Energy Information Administration

United States® Census Bureau

API American Petroleum Institute

Google Maps Platform

Data Wrangling

Cleaning

- Datasets required minimal cleaning and mostly involved separating multi-level data fields from aggregate and sub-total groups, e.g., crude API data in was presented at top (U.S.), middle (PAD district), and lowest (PAD sub-district) levels and needed extracting into distinct categories to conduct proper analysis.
- Since datasets were sourced from specialized institutions, they had negligible number of missing values and were mostly of the refinery resale pricing
- The little amount of missing pricing data were imputed using pandas 'fillna()' function

Combining, melting, merging datasets

Datasets were combined and merged on the most common categorical features:

- PAD District, PAD sub-district, Time Periods

Other common features used for combining tables and merging datasets included:

- Refinery capacities by process type
- Refinery product yields
- Feedstock purchase and product resale prices
- Distances from refineries to selected market destinations

Exploratory Data Analysis

Greenfield Refinery Feasibility Validation

EDA

Target Variable

- Crack Spread is the most important feature in validating Green Light’s feasibility studies – a gross profit margin in USD per 1 US barrel of crude processed.
- **Refinery Specific Crack spread is the target variable** to be predicted in the modeling

Features values

Categorical

- Eight groups of observations per each of the 121 refineries and 1 modeled refinery case
- 5 top-level and 7 second-level grouping of refinery locations per PAD districts

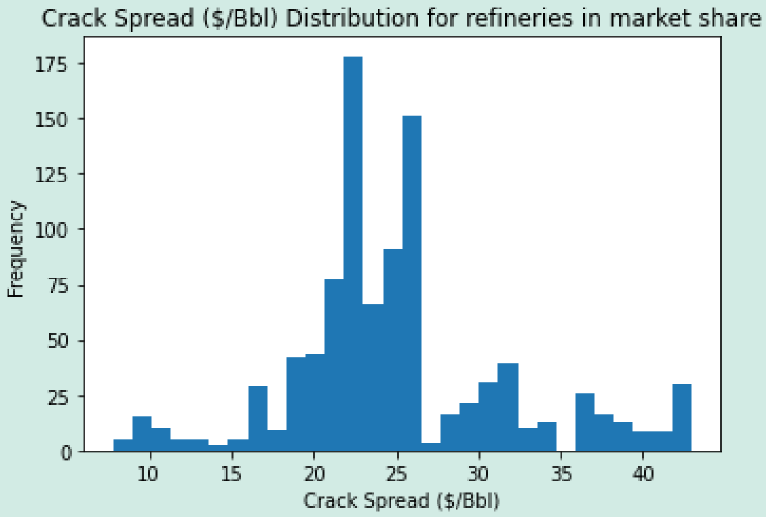
Key Numerical Features:

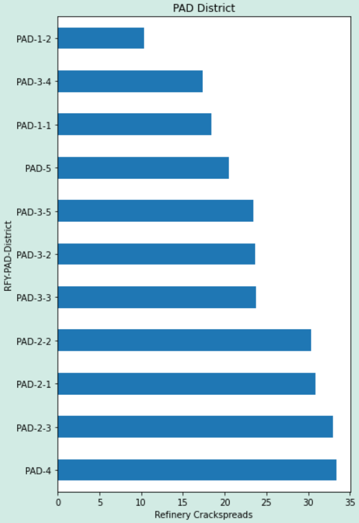
- Crude API, Product Yields (Gasoline, Diesel, Residue)
- Capacities by Process Type, Crude prices and differentials (premium or discount)
- Destination market population density, Refinery Supply Pressure, Market Resistance
- Distances from refineries to selected market destinations
- Historical Crack Spreads by PAD districts

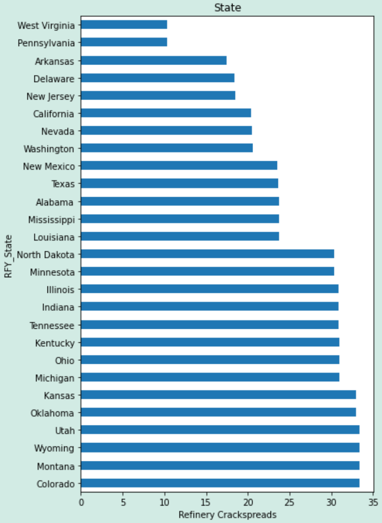
EDA

Crack Spread Distribution by PAD

- Although Texas (PAD-3) and Louisiana (PAD-3) have the highest count of refineries with biggest capacities, they ranked in the mid-range in terms of crack-spreads $/Bbl



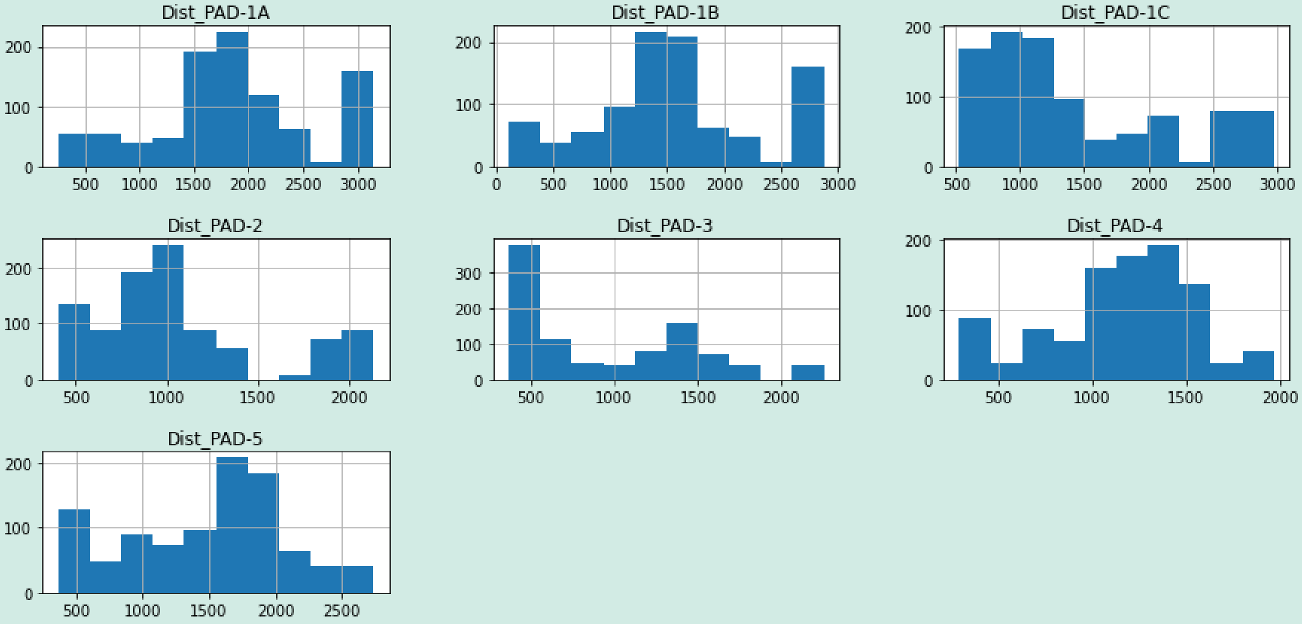




EDA

Cumulative Refinery-to-Market Distances

- PAD-3 refineries: proximity to target markets; create highest supply pressure; lower crack-spreads
- PAD-1A/1B, PAD-2, PAD-4, and PAD-5: lower concentration, low supply pressure



EDA

Correlation of features

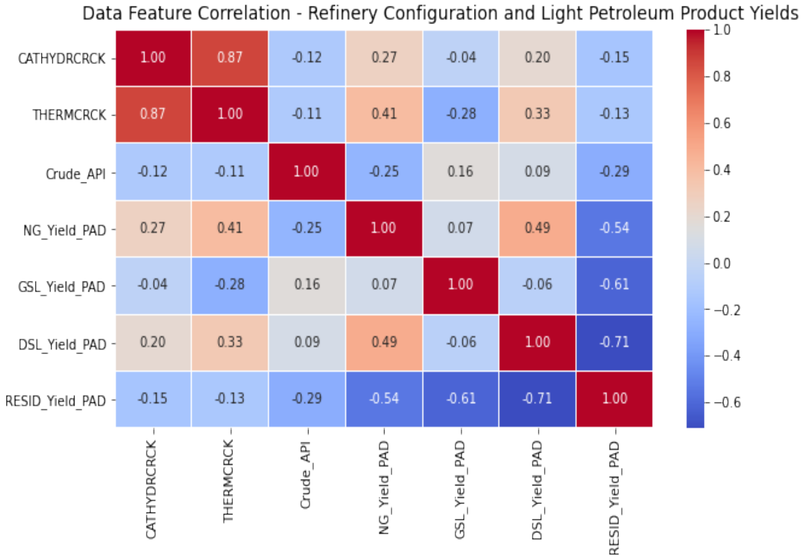
During the exploratory data analysis, it becomes clear that there is relationship between plant configuration, Crude API and the Product Yields.

The analysis showed strong correlation of light product yields with the following data features:

- *Catalytic Hydrocracking*
- *Thermal Cracking*
- *Crude API*

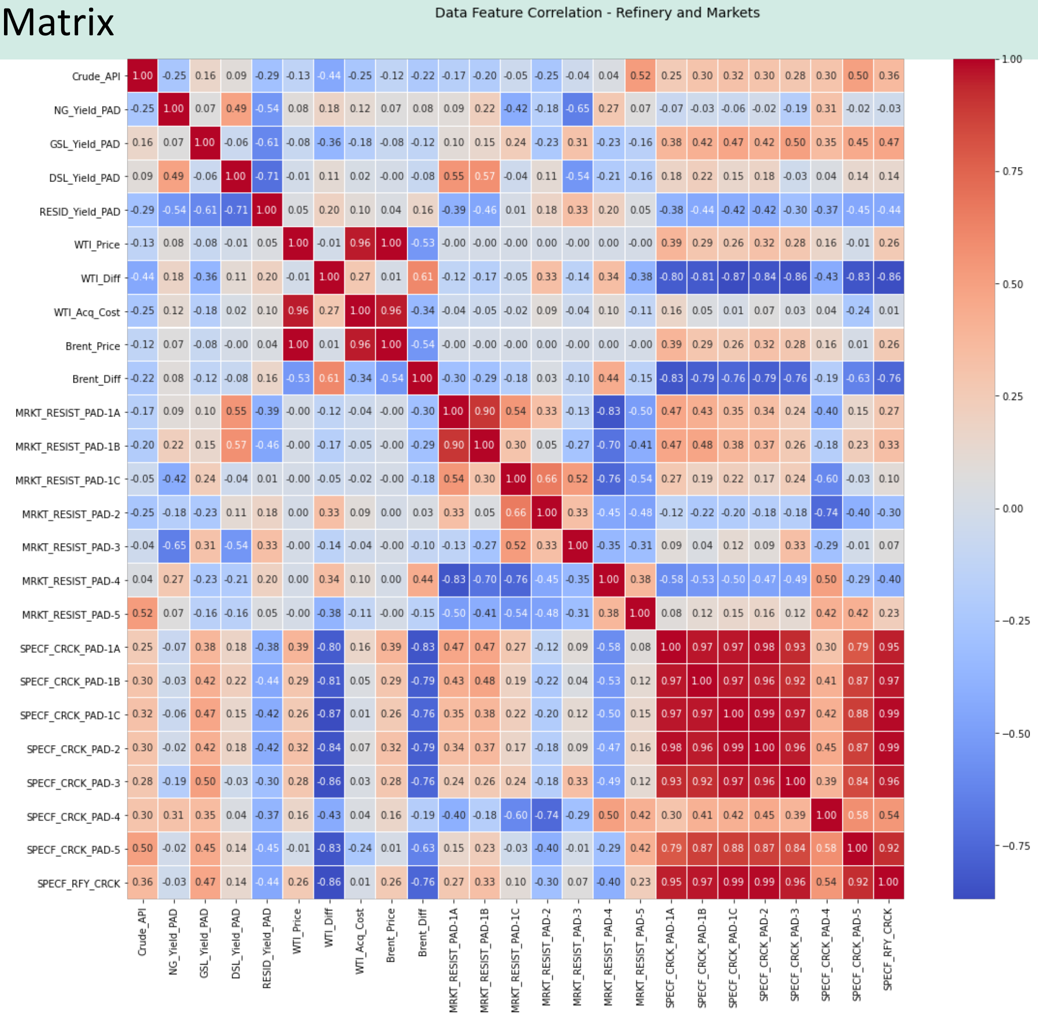
The heatmap and a series of scatter plots further confirmed this pattern.

Heatmap of features



EDA

Feature Correlation Matrix



	Crude_API	NG_Yield_PAD	GSL_Yield_PAD	DSL_Yield_PAD	RESID_Yield_PAD	WTI_Price	WTI_Diff	WTI_Acq_Cost	Brent_Price	Brent_Diff	MRKT_RESIST_PAD-1A	MRKT_RESIST_PAD-1B	MRKT_RESIST_PAD-1C	MRKT_RESIST_PAD-2	MRKT_RESIST_PAD-3	MRKT_RESIST_PAD-4	MRKT_RESIST_PAD-5	SPECF_CRCK_PAD-1A	SPECF_CRCK_PAD-1B	SPECF_CRCK_PAD-1C	SPECF_CRCK_PAD-2	SPECF_CRCK_PAD-3	SPECF_CRCK_PAD-4	SPECF_CRCK_PAD-5	SPECF_RFY_CRCK
Crude_API	1.00	-0.25	0.16	0.09	-0.29	-0.13	-0.44	-0.25	-0.12	-0.22	-0.17	-0.20	-0.05	-0.25	-0.04	0.04	0.52	0.25	0.30	0.32	0.30	0.28	0.30	0.50	0.36
NG_Yield_PAD	-0.25	1.00	0.07	0.49	-0.54	0.08	0.18	0.12	0.07	0.08	0.09	0.22	-0.42	-0.18	-0.65	0.27	0.07	-0.07	-0.03	-0.06	-0.02	-0.19	0.31	-0.02	-0.03
GSL_Yield_PAD	0.16	0.07	1.00	-0.06	-0.61	-0.08	-0.36	-0.18	-0.08	-0.12	0.10	0.15	0.24	-0.23	0.31	-0.23	-0.16	0.38	0.42	0.47	0.42	0.50	0.35	0.45	0.47
DSL_Yield_PAD	0.09	0.49	-0.06	1.00	-0.71	-0.01	0.11	0.02	-0.00	-0.08	0.55	0.57	-0.04	0.11	-0.54	-0.21	-0.16	0.18	0.22	0.15	0.18	-0.03	0.04	0.14	0.14
RESID_Yield_PAD	-0.29	-0.54	-0.61	-0.71	1.00	0.05	0.20	0.10	0.04	0.16	-0.39	-0.46	0.01	0.18	0.33	0.20	0.05	-0.38	-0.44	-0.42	-0.42	-0.30	-0.37	-0.45	-0.44
WTI_Price	-0.13	0.08	-0.08	-0.01	0.05	1.00	-0.01	0.96	1.00	-0.53	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.39	0.29	0.26	0.32	0.28	0.16	-0.01	0.26
WTI_Diff	-0.44	0.18	-0.36	0.11	0.20	-0.01	1.00	0.27	0.01	0.61	-0.12	-0.17	-0.05	0.33	-0.14	0.34	-0.38	-0.80	-0.81	-0.87	-0.84	-0.86	-0.43	-0.83	-0.86
WTI_Acq_Cost	-0.25	0.12	-0.18	0.02	0.10	0.96	0.27	1.00	0.96	-0.34	-0.04	-0.05	-0.02	0.09	-0.04	0.10	-0.11	0.16	0.05	0.01	0.07	0.03	0.04	-0.24	0.01
Brent_Price	-0.12	0.07	-0.08	-0.00	0.04	1.00	0.01	0.96	1.00	-0.54	-0.00	-0.00	-0.00	0.00	-0.00	0.00	-0.00	0.39	0.29	0.26	0.32	0.28	0.16	0.01	0.26
Brent_Diff	-0.22	0.08	-0.12	-0.08	0.16	-0.53	0.61	-0.34	-0.54	1.00	-0.30	-0.29	-0.18	0.03	-0.10	0.44	-0.15	-0.83	-0.79	-0.76	-0.79	-0.76	-0.19	-0.63	-0.76
MRKT_RESIST_PAD-1A	-0.17	0.09	0.10	0.55	-0.39	-0.00	-0.12	-0.04	-0.00	-0.30	1.00	0.90	0.54	0.33	-0.13	-0.83	-0.50	0.47	0.43	0.35	0.34	0.24	-0.40	0.15	0.27
MRKT_RESIST_PAD-1B	-0.20	0.22	0.15	0.57	-0.46	-0.00	-0.17	-0.05	-0.00	-0.29	0.90	1.00	0.30	0.05	-0.27	-0.70	-0.41	0.47	0.48	0.38	0.37	0.26	-0.18	0.23	0.33
MRKT_RESIST_PAD-1C	-0.05	-0.42	0.24	-0.04	0.01	-0.00	-0.05	-0.02	-0.00	-0.18	0.54	0.30	1.00	0.66	0.52	-0.76	-0.54	0.27	0.19	0.22	0.17	0.24	-0.60	-0.03	0.10
MRKT_RESIST_PAD-2	-0.25	-0.18	-0.23	0.11	0.18	0.00	0.33	0.09	0.00	0.03	0.33	0.05	0.66	1.00	0.33	-0.45	-0.48	-0.12	-0.22	-0.20	-0.18	-0.18	-0.74	-0.40	-0.30
MRKT_RESIST_PAD-3	-0.04	-0.65	0.31	-0.54	0.33	-0.00	-0.14	-0.04	-0.00	-0.10	-0.13	-0.27	0.52	0.33	1.00	-0.35	-0.31	0.09	0.04	0.12	0.09	0.33	-0.29	-0.01	0.07
MRKT_RESIST_PAD-4	0.04	0.27	-0.23	-0.21	0.20	0.00	0.34	0.10	0.00	0.44	-0.83	-0.70	-0.76	-0.45	-0.35	1.00	0.38	-0.58	-0.53	-0.50	-0.47	-0.49	0.50	-0.29	-0.40
MRKT_RESIST_PAD-5	0.52	0.07	-0.16	-0.16	0.05	-0.00	-0.38	-0.11	-0.00	-0.15	-0.50	-0.41	-0.54	-0.48	-0.31	0.38	1.00	0.08	0.12	0.15	0.16	0.12	0.42	0.42	0.23
SPECF_CRCK_PAD-1A	0.25	-0.07	0.38	0.18	-0.38	0.39	-0.80	0.16	0.39	-0.83	0.47	0.47	0.27	-0.12	0.09	-0.58	0.08	1.00	0.97	0.97	0.98	0.93	0.30	0.79	0.95
SPECF_CRCK_PAD-1B	0.30	-0.03	0.42	0.22	-0.44	0.29	-0.81	0.05	0.29	-0.79	0.43	0.48	0.19	-0.22	0.04	-0.53	0.12	0.97	1.00	0.97	0.96	0.92	0.41	0.87	0.97
SPECF_CRCK_PAD-1C	0.32	-0.06	0.47	0.15	-0.42	0.26	-0.87	0.01	0.26	-0.76	0.35	0.38	0.22	-0.20	0.12	-0.50	0.15	0.97	0.97	1.00	0.99	0.97	0.42	0.88	0.99
SPECF_CRCK_PAD-2	0.30	-0.02	0.42	0.18	-0.42	0.32	-0.84	0.07	0.32	-0.79	0.34	0.37	0.17	-0.18	0.09	-0.47	0.16	0.98	0.96	0.99	1.00	0.96	0.45	0.87	0.99
SPECF_CRCK_PAD-3	0.28	-0.19	0.50	-0.03	-0.30	0.28	-0.86	0.03	0.28	-0.76	0.24	0.26	0.24	-0.18	0.33	-0.49	0.12	0.93	0.92	0.97	0.96	1.00	0.39	0.84	0.96
SPECF_CRCK_PAD-4	0.30	0.31	0.35	0.04	-0.37	0.16	-0.43	0.04	0.16	-0.19	-0.40	-0.18	-0.60	-0.74	-0.29	0.50	0.42	0.30	0.41	0.42	0.45	0.39	1.00	0.58	0.54
SPECF_CRCK_PAD-5	0.50	-0.02	0.45	0.14	-0.45	-0.01	-0.83	-0.24	0.01	-0.63	0.15	0.23	-0.03	-0.40	-0.01	-0.29	0.42	0.79	0.87	0.88	0.87	0.84	0.58	1.00	0.92
SPECF_RFY_CRCK	0.36	-0.03	0.47	0.14	-0.44	0.26	-0.86	0.01	0.26	-0.76	0.27	0.33	0.10	-0.30	0.07	-0.40	0.23	0.95	0.97	0.99	0.99	0.96	0.54	0.92	1.00

Modeling Results and Analysis

Greenfield Refinery Feasibility Validation

Modeling Results and Analysis

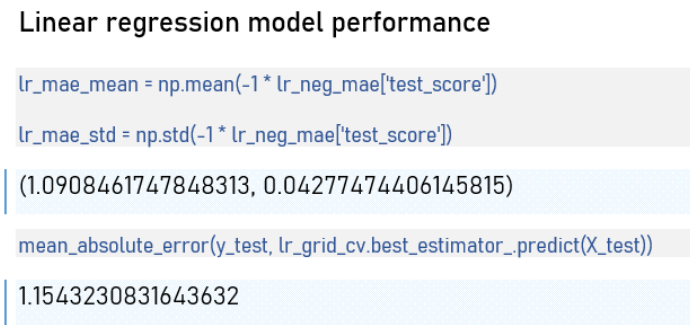
Linear Regression Model

The Linear Model further supported the patterns, initially found during the exploration.

The linear regression model found that there is a strong correlation between the Crack Spreads and the following features:

- *Brent Price*
- *Light Product Yields (Gasoline, Diesel)*
- *Crude API*
- *Crude Price Differentials*

Liner Regression Model Results



Linear regression model performance

```
lr_mae_mean = np.mean(-1 * lr_neg_mae['test_score'])

lr_mae_std = np.std(-1 * lr_neg_mae['test_score'])
```

```
(1.0908461747848313, 0.04277474406145815)
```

```
mean_absolute_error(y_test, lr_grid_cv.best_estimator_.predict(X_test))
```

```
1.1543230831643632
```

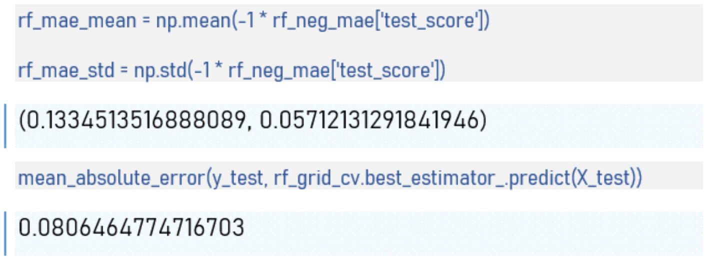
Modeling Results and Analysis

Random Forest Model

The Random Forest Model also returned top features that were common with the Linear Model:

- *Brent Price Differential*
- *Product Yield*
- *Crude API*

The Random Forest Model performance



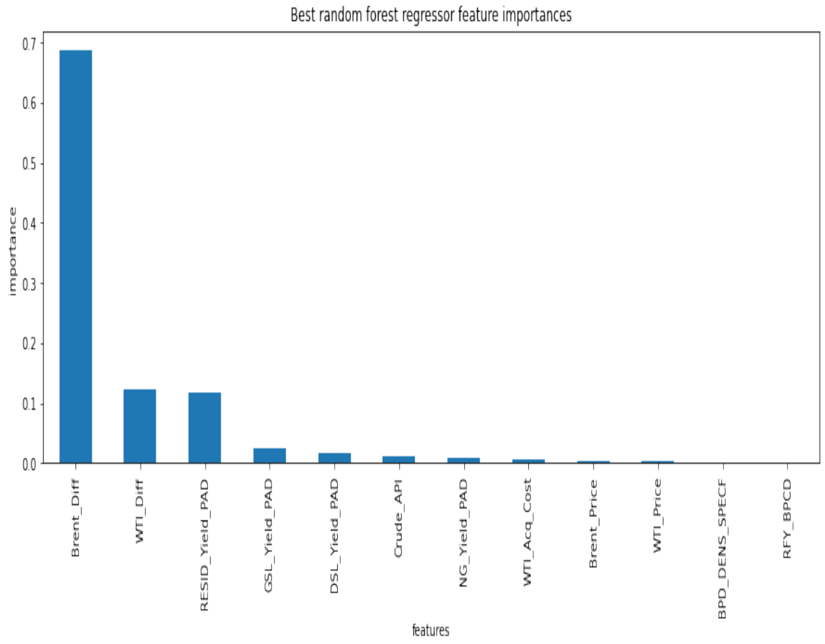
```
rf_mae_mean = np.mean(-1 * rf_neg_mae['test_score'])

rf_mae_std = np.std(-1 * rf_neg_mae['test_score'])
```

```
(0.1334513516888089, 0.05712131291841946)
```

```
mean_absolute_error(y_test, rf_grid_cv.best_estimator_.predict(X_test))
```

```
0.0806464774716703
```

Random Forest Model Results



Modeling Results and Analysis

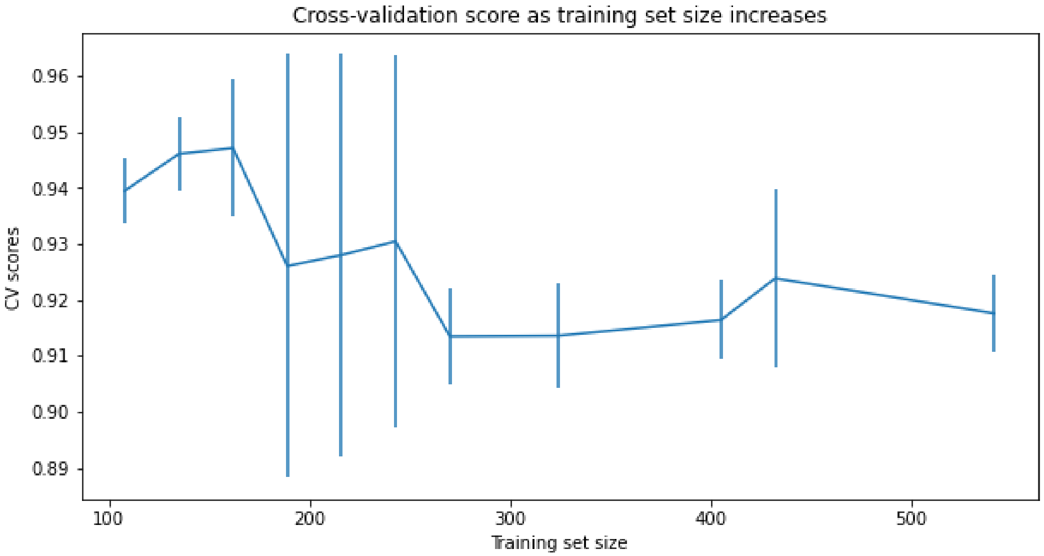
Analysis

After comparing the cross-validation MAE values for the two models, we concluded that the random forest model was a better choice, because it showed lower CV MAE and less variability. The performance on the test set was also consistent with the cross-validation results.

Data quantity assessment

Finally, we assessed whether the model would benefit from additional data collection by running the sklearn's learning_curve function.

The CV score and Training set size plot showed that no further data collection would be required. As the plot shows, the improvement in the model scores starts flatting out by around a sample size of 250



Recommendation and Key Findings

Greenfield Refinery Feasibility Validation

Key Findings and Recommendations

Key Findings

The key findings, which were determined upon analyzing the data and running models, are as follows:

- Brent Differential has a stronger impact on the economics than the domestic WTI crude price index. There is also an indication that WTI is dependent on the Brent.
- This was expected, as the heavier crude doesn't automatically mean fewer valuable products may be extracted from it than the lighter crude – it is mainly a function of refinery process type and the equipment. However, there is more lighter product yields in the lighter crude oil. It is easier to extract than from the heavy crude. This effect caps out at the lower and higher ranges, in this case, the 39-42 API range.
- The model did not rank any of the one-hot encoded 'PAD' district features very well in the 'feature importance' assessment. The reality though supports the theory that the PAD location is a very important factor in economic feasibility of a crude oil refinery – an indication that the model definitely needs further tweaking

Recommendations

Continue working to produce a useful application for the industry. There is plenty of space for improvement – work on both the data and the models.

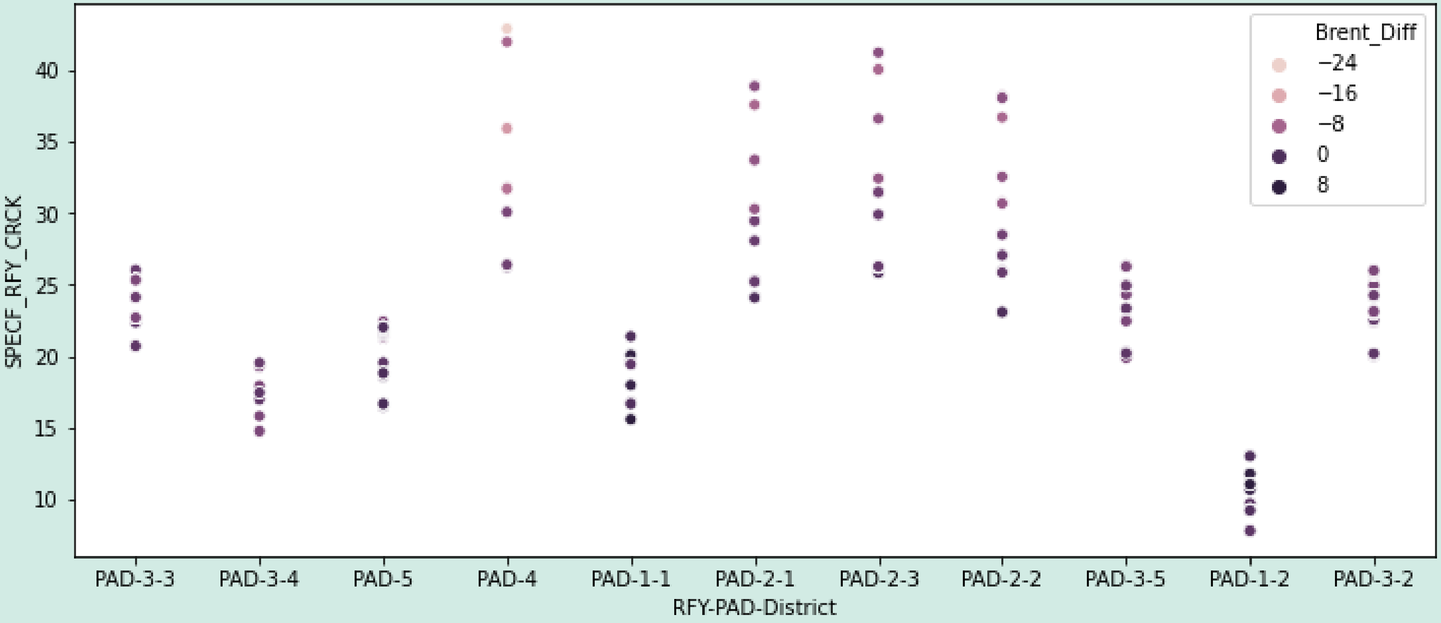
Summary and Conclusion

Greenfield Refinery Feasibility Validation

Summary and Conclusion

Importance of Refinery Location

- Here is a scatter plot that illustrates how the expected Crack Spreads are distributed by PAD district



Summary and Conclusion

Summary

The goal, as outlined in the Problem Statement was to validate the feasibility studies for building a new refinery in PAD-2-2 district by identifying 2-3 key parameters that have the highest impact on the Specific Crack Spread.
I had already confirmed the key parameters – i.e., the top features by importance.

Scenario 1:
Increasing WTI Discount from the Baseline $(6.05)/Bbl to $(7.5)/Bbl, while keeping the Brent Discount at zero, added a modest increase in the Crack Spread – from $20.58/Bbl to $22.99/Bbl, but still lower than the Baseline of $25.97/Bbl

Increasing Brent Discount from the Baseline $(2.08) to $(5.5), while keeping the WTI Discount at zero had a significant impact, increasing the Crack Spread – to $34.66/Bbl

Scenario 2: Changing the Baseline value of the Crude API from the heavier 26.11 to the lighter 39 resulted in a slightly better outcome – the Crack Spread went from $25.97/Bbl to $26.03/Bbl. Further increase of Crude API to 41 and 42 did not add anything – the Crack Spread stayed at the $26.03/Bbl level

Summary

Scenario 3: Moving the refinery location from PAD-2-2 to PAD-4 resulted in a very negligible change in the predicted Crack Spread – it went from $25.97/Bbl to $25.99/Bbl. Similarly, changing the location from PAD-2-2 to PAD-3-3 did not have any further affect

Conclusion

The conclusion here is that Brent Differential has a stronger impact on the economics than the domestic WTI crude price index. There is also an indication that WTI is dependent on the Brent.

The model did not rank any of the one-hot encoded ‘PAD’ district features very well in the ‘feature importance’ assessment. The reality though supports the theory that the PAD location is a very important factor in economic feasibility of a crude oil refinery – an indication that the model definitely needs further tweaking.