**Team 5: U.S. Crude Oil and Natural Gas Price and Rotary Rigs in Operation Analysis**

UH SPE Machine Learning Bootcamp First Project: Linear Problem and Linear Classification

Dung Nguyen, Xu Yang, Charles White, Jose Benavides

1. **Project Introduction:**

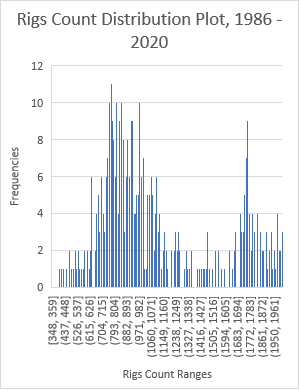
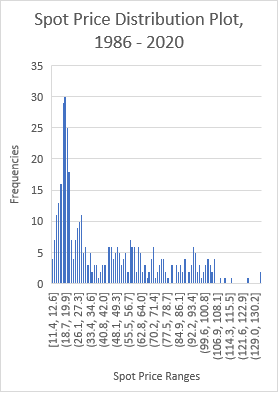
Rotary rig is an equipment commonly used for drilling purpose in most wells. Accordingly, how many rotary rigs in operation can reflect how much oil a company is getting, which in turn affects the crude oil price. The information on this relationship, if obtained, can be combined with the cost and revenue analysis from the company’s sales department to provide an insight on how to increase its profits.

The spot price will be analyzed to determine the rigs count, as it’s more commonly done by oil companies to determine the number of rotary rigs that should be in operation.

1. **Dataset**

In this report, two datasets will be obtained from <https://www.eia.gov/> using the site’s provided API and the dataset’s IDs. The two specific datasets are: “U.S. Crude Oil and Natural Gas Rotary Rigs in Operation” and “Spot Price of WIT.”

1. **Features and Pre-Processing:**
2. Features



Due to the 2008 financial crisis that introduces various complex variables to the relationship between oil price and rigs count (as indicated by the above chart) and the interests of analyzing more current, relevant data with more rig drilling efficiency being realized, only the 2010-2020 time period’s data will be selected.



As indicated by the chart, there are two significant clusters of data: [348, 1116] and [1616, 1984]. All subsequent analysis of this dataset would be done by using a model for each cluster, because there should be no significant importance to force these two clusters to correlate.

1. Pre-Processing

First, The Spot Price and Rotary Rigs Count datasets are synchronized by shifting the time series to make them match. Secondly, we sorted the dataset according to the date from latest to oldest. This is necessary so as to make sure the access of data could be done with more confidence.

In addition, in oil industry the spot prices are often recorded 3 months late compared to the rigs count, due to the fact that oil companies sign the contract quarterly. The real price should not reflect the actual rig count until next quarter agreement period. Therefore we selected the Spot Price dataset 3 rows late and the Rigs Count dataset 3 rows early. Here is the plot of Rigs Count vs Spot Price, 2010 – 2020 period with all of the pre-processing we’ve done:

As we can see that there seems to be a very obvious linear relationship between the rig count and spot prices. Also, as stated previously, there are two clusters on two ranges. For the horizontal axis (spot price): [20, 71] (lower cluster) and [71, 120] (upper cluster).

1. **Model and Techniques**

These two clusters will be analyzed using the linear regression model provided by Scikit learn package. In this report, y will be the rigs count and a function of x, the spot price:

𝑦 = 𝛽₀ + 𝛽₁𝑥₁ + ⋯ + 𝛽ᵣ𝑥ᵣ + 𝜀

With 𝜀 as the error and 𝛽₀, 𝛽₁, ⋯ 𝛽ᵣ will be iterated to minimize the error following this equation:

SSR = Σᵢ(𝑦ᵢ - 𝑓(𝐱ᵢ))²

with SSR being sum of squared residues. Note that in this case there will be only one vector of x since we did not include other features so as to predict the rig count.

After preprocessing, the following equations are obtained:

Lower Cluster:

coefficient of determination (R2): 0.429

𝑦 = 53.5 + 14.5 𝑥



Upper Cluster:

coefficient of determination (R2): 0.466

𝑦 = 712 + 11.7 𝑥



Even the coefficient of determination is not as good as what could be called a good linear fit, the linear regression modeling job has already been necessarily completed.

1. **Results and Discussion:**
2. Results testing:

We used the leading 70% of the dataset to train the linear model and test the model using the rest of 30%.

Lower cluster prediction:



Linear fit: 𝑦 = 53.5 + 14.5 𝑥

Coefficients: [[9.33002152]]

Mean squared error: 33442.16

Coefficient of determination (y test vs y predict): 0.12

coefficient of determination (x test vs y predict): -43.91960031478436

Upper cluster prediction:



Linear fit: 𝑦 = 712 + 11.7 𝑥

Coefficients: [[8.37372782]]

Mean squared error: 26432.48

Coefficient of determination (y test vs y predict): 0.44

coefficient of determination (x test vs y predict): 0.8440780702198116

1. Discussion:

The prediction results shown above has a very large error when utilizing the first 70% and test the remaining 30%. We made a hypothesis that this could be due to one factor modelling is too ambitious to achieve a good quality regression.

Therefore, we are putting a more complex model involving more than one vector feature in the next coming projects.