**Introduction**

Profitability for oil & gas companies are largely driven by commodity prices that are set by the market. These can depend on various factors such as supply and demand (both local and international), weather conditions, seasonal demand, political unrest in the Middle East etc. Weather is actually a big factor for natural gas demand because daily weather has a significant impact on power plants for electricity generation. Over the last few years, natural gas has replaced coal to become the leading source for power generation in the country. It would be very beneficial to natural gas producers if they could predict spot prices based on weather patterns and accordingly scale production back or up to not oversupply the market. The goal of this project is to build a model that would predict the effect of temperature data on natural gas spot prices at the Henry Hub natural gas distribution hub in Erath, Louisiana. Henry Hub is the leading distribution hub for natural gas in the US.

**Dataset deep dive**

The two datasets used for this project are public and accessible to the general population using Web APIs. The EIA government website provides historical data for daily natural gas spot prices. Weather Underground provides daily weather data for Erath, LA. Let us look at each dataset in more detail.

**EIA:** Using an EIA provided API key, this data was loaded into a .csv file for analysis. It is a fairly simple dataset with just 2 fields, the date in yyymmdd format and the spot price for that day. There was data all the way from 1997 through current date. This turned out to be very clean and nicely formatted and did not need any data wrangling except for filtering the data down to our dates of interest

**Weather Underground:** A similar API call was obtained from Weather Underground to extract weather data. The dataset had several weather related variables and the fields of interest were narrowed down to temperature, humidity and pressure. There was a fair amount of data wrangling done to get this dataset into the correct format. Some of the readings showed a value of -9999 which one assumes was just a placeholder for a reading not recorded or an incorrect reading. These rows were dropped from the dataset. The date field was of the text datatype and it had to be converted to an integer in the yyyymmdd format to match the EIA data. Additionally, the dataset had readings for every 20 minutes so the data had to be averaged to report at the day level. Each API call would extract data for one day and it was written to an individual csv file. A for loop in R was used to increment the API syntax to pass the following day’s date for the subsequent call. All the files were combined into one dataset prior to the data wrangling exercise.

Additionally, Weather Underground has restrictions on how much data one could extract within the scope of a free account. Only 500 days of history could be extracted each day, so multiple runs over different date ranges had to be made to fill out the training and test datasets. Another restriction was on the number of calls per minute so a pause had to be built in.

After the necessary data wrangling was done on the weather dataset, the two datasets were combined keyed on date and the resulting dataset was written out to a csv file to use in the machine learning exercise.

**Limitations**

Though weather is indeed a big factor that goes into the determination of commodity prices, there is another equally important factors that play a role. Storage levels of natural gas drives supply and demand and we’re not taking that into account for this exercise. Also, we are only looking at weather readings for the Henry Hub location, it would make the model more robust if we picked weather readings across the country and averaged them out.

**Initial Findings**

The first few models showed a poor adjusted R-squared value, the highest was 10%. This was with all three independent variables – temperature, pressure and humidity included. So, we don’t need to experiment with dropping any variables to improve the model.

**Next Steps**

There are quite a few techniques that we can use to improve our model. We could look at correlations between the variables and if needed, create a new variable which would be a combination of two or more. We could pre-process the data (centering and scaling) and observe the results. Additionally, we could plot histograms and boxplots of the data to look at data distribution and find outliers if any. If we do find outliers, excluding them should have a positive effect on the model.

As a last step, we could try to use a regression tree model instead of a linear regression one to see if it is a better fit for the dataset used in this exercise.