The University of Texas at Austin

# Mining Relationships between US Petroleum Statistics and Global Oil & Gas Prices

Project Report

CE 395R - Data Mining w/Dr.Carlos Caldas
Spring 2018

**TEAM MEMBERS:**
Milind Siddhanti (mss4376)
Sanchit Singhal (ss84657)

## Problem Definition

The project will be focused in the petroleum domain and will attempt to come up with patterns and draw inferences between US petroleum statistics about drilling performance and rig counts to the global oil and gas prices.

## Background

Recently, the petroleum industry is going through a change, not only worldwide, but particularly in the US. Although oil prices have fallen over the last decade, America's market share in world oil production has gradually increased. With new found shale oil reserves, the United States is playing a more dominant role in controlling oil prices.

## Objective

Given the recent developments, the project will attempt to understand the relationship of the US petroleum production and how it affects the price of both oil and gas.

## Need

By evaluating the impact of the regional oil industry on oil and natural gas prices, the US government can better regulate petroleum firms to coordinate a strategy as a country that helps them gain a competitive advantage in the global oil markets. Further, companies can utilize this analysis to develop future strategies and direct their investment that maximizes growth opportunities. Lastly, this information can be used to develop economic relationships internally, between states, for a more stable trade market in the US.

## Approach

Early on in the investigation, the team decided to analyze the relationships primarily through unsupervised learning. Solving a regression/classification problem to predict or forecast either production or price might be a little out of scope - as that can involve thousands of factors - but by drawing inference in the data, the team hopes to learn about the impact of production statistics on regional prices over time. We attempt to swift through multiple lines of modeling techniques to discover implicit patterns and identify variables that significantly impact pricing.

## Data

The dataset consists of features like: Region where the rigs are available, monthly data of the rig counts, production per rig, legacy production change, and total production. The data is available for petroleum liquids, Oil and Natural Gas. Also, features with respect to the rigs such as number of drills that were dug per month, completed and drilled but uncompleted (DUC) are available regionally and these can be used to compare the production rates for completed rigs.

Data for the oil prices across the US will be merged to the dataset as an additional attribute and inferences can be made of the captured datasets.
Data Source: https://www.eia.gov/petroleum/drilling/

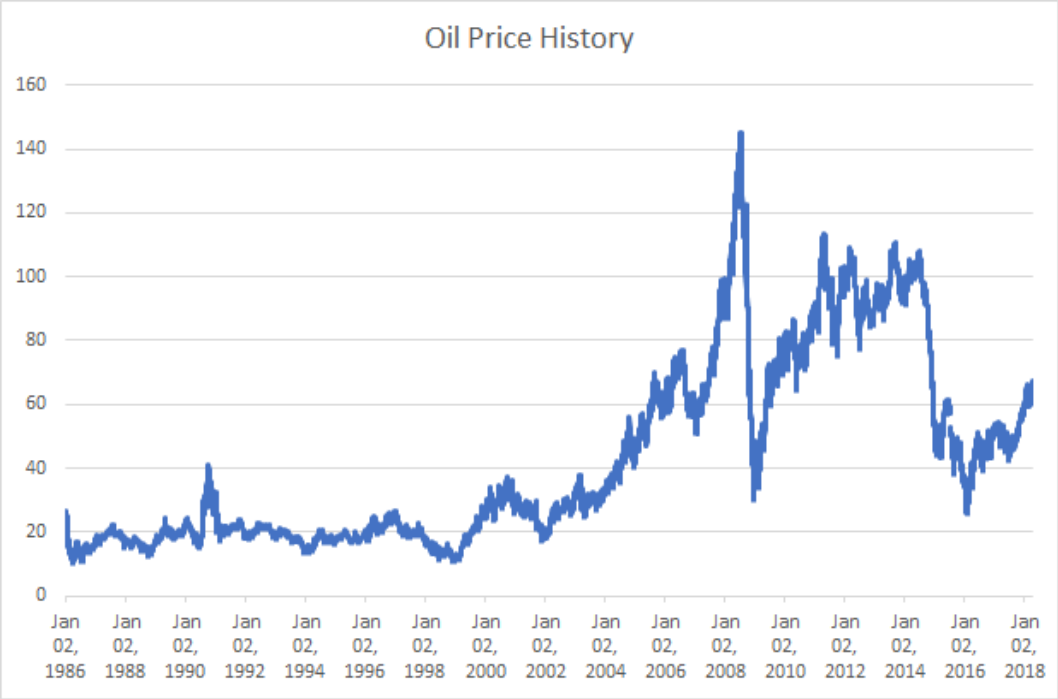
## Methodology

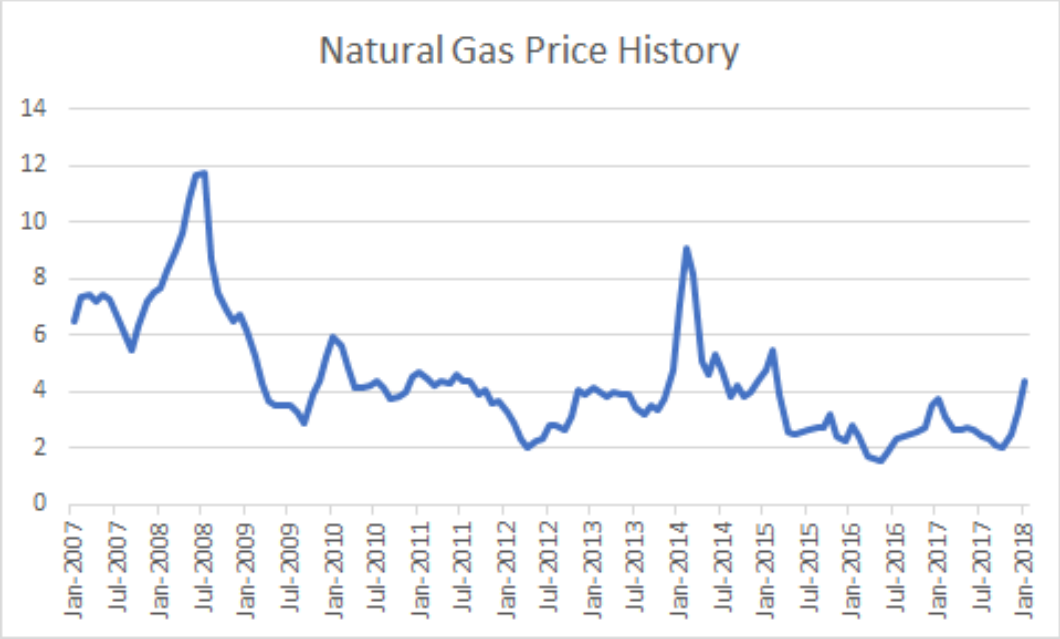
At the beginning of the project, the team began with data collection by gathering all datasets that they might need. The US Energy Information Administration (EIA) was identified as a data source. Next, the team went through a selection process in which the target dataset was finalized. This was an iterative process - we went back and forth a few times to ensure we had all data points that we needed. Once target data had been selected, the team preprocessed the data to verify that it is of high quality and in a format that is easy to use with the chosen software. The transformed data enabled the team to explore the data visually. We wanted to learn patterns in the pricing of both oil and gas prices without any predictors before applying data mining algorithms so that we had a better understanding of the fluctuations. The team also explored relationships in just the production statistics without prices too - with a similar objective. At this stage the team continued on to apply algorithms to the data with the goal of drawing inferences between the two datasets. Some of the analysis from our investigation are detailed below.

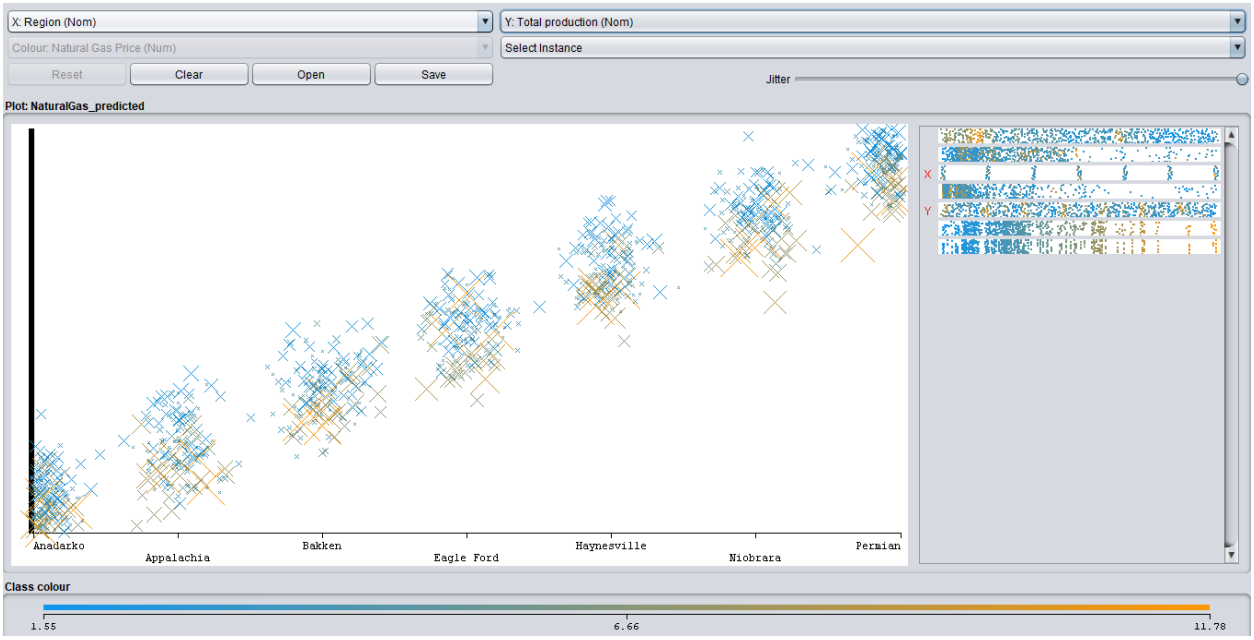
## Data Preprocessing Steps Taken

- Split Datasets into Oil and Natural gas sets
- Removed missing values
- Merged Region information as a feature
- Added Price information
- Selected last day price of month as bin value
- Discrete numerical values (Oil Price into 5 bins, Natural Gas into 4 bins)

## Data Exploration






## Data Modeling Methods Explored

## 1) Clustering models

### a) Simple K Means-

k-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

Advantages of k means clustering are - Fast, robust, and easy to understand. Gives good results when data set consists of distinct data points.

Disadvantages are - Euclidean distance measures can unequally weight underlying factors. Randomly choosing of the cluster center cannot lead us to the fruitful result. Applicable only

when mean is defined i.e. fails for categorical data. Unable to handle noisy data and outliers. Algorithm fails for non-linear data set.

### 1. b) Hierarchical Clustering-

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows us to decide the level or scale of clustering that is most appropriate for your application. The dendrogram function plots the cluster tree.

Find the similarity or dissimilarity between every pair of data points in the data set. Calculate the distance between the data points. Then, group the data points into a binary, hierarchical cluster tree. As data points are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed. At last, determine where to cut the hierarchical tree into clusters.

Advantage of Hierarchical clustering is - it is easy to implement and gives best result in some cases.

Disadvantages are - Algorithm can never undo what was done previously. Sometimes it is difficult to identify the correct number of clusters by the dendrogram. Sensitivity to noise and outliers. Difficulty handling different sized clusters and convex shapes.

## 2) Classification models

### a) Logistic Regression-

Logistic regression aims to measure the relationship between a categorical dependent variable and one or more independent variables (usually continuous) by plotting the dependent variables' probability scores. A categorical variable is a variable that can take values falling in limited categories instead of being continuous. Logistic regression uses regression to predict the outcome of a categorical dependent variable on the basis of predictor variables. The probable outcomes of a single trial are modeled as a function of the explanatory variable using a logistic function. Logistic modeling is done on categorical data which may be of various types including binary and nominal. For example, a variable might be binary and have two possible categories of 'yes' and 'no'; or it may be nominal say hair color maybe black, brown, red, gold and grey. Another objective of logistic regression is to check if the probability of getting a particular value of the dependent variable is related to the independent variable. Multiple logistic regression is used when there are more than one independent variables under study.

Advantages of Logistic regression are- Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Logistic models can be updated easily with new data using stochastic gradient descent.

Disadvantages are- Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships.

## 2. b) Naive Bayes Classifier-

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features. For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Advantages of Naive Bayes classifier are- Very simple, easy to implement and fast. If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression. Even if the NB assumption doesn't hold, it works great in practice. Need less training data.

Disadvantages are- Precision and recall is low, relies on independence assumption and will perform badly if this assumption is not met

## 2. c) Decision Tree Classifier-

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. The data is continuously split based on the feature set available. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Decision trees can handle both categorical and numerical data.

Advantages of decision tree are- It does not require any domain knowledge. It is easy to comprehend. The learning and classification steps of a decision tree are simple and fast.

Disadvantages are- Exponential calculation growth while problem is getting bigger. Need to discrete data for some particular construction algorithm.

## 3) Association rules - Apriori Rule Association

Association Rules find all sets of items (itemset) that have support greater than the minimum support and then using the large itemset to generate the desired rules that have confidence greater than the minimum confidence. The lift of a rule is the ratio of the observed support to that expected if X and Y were independent. A typical and widely used example of association rules application is market basket analysis.

Disadvantage is- suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms.

## Results and Analysis

We as a team decided to apply different models on the natural gas and oil datasets to see the variations in the distribution of the production of natural gas per rig over the years region wise across the US. With the help of the software WEKA, we targeted to run a variety of data mining technique to understand the patterns and key findings from the dataset. The data mining techniques that we implemented were the classification, association rules and clustering models.

We found significant results from each of the models executed as a function of the three data mining techniques. The results of some of the models applied will be discussed in detail with the feature set, model and properties of the model:

## Modeling Relationship between Natural Gas and US Petroleum Production

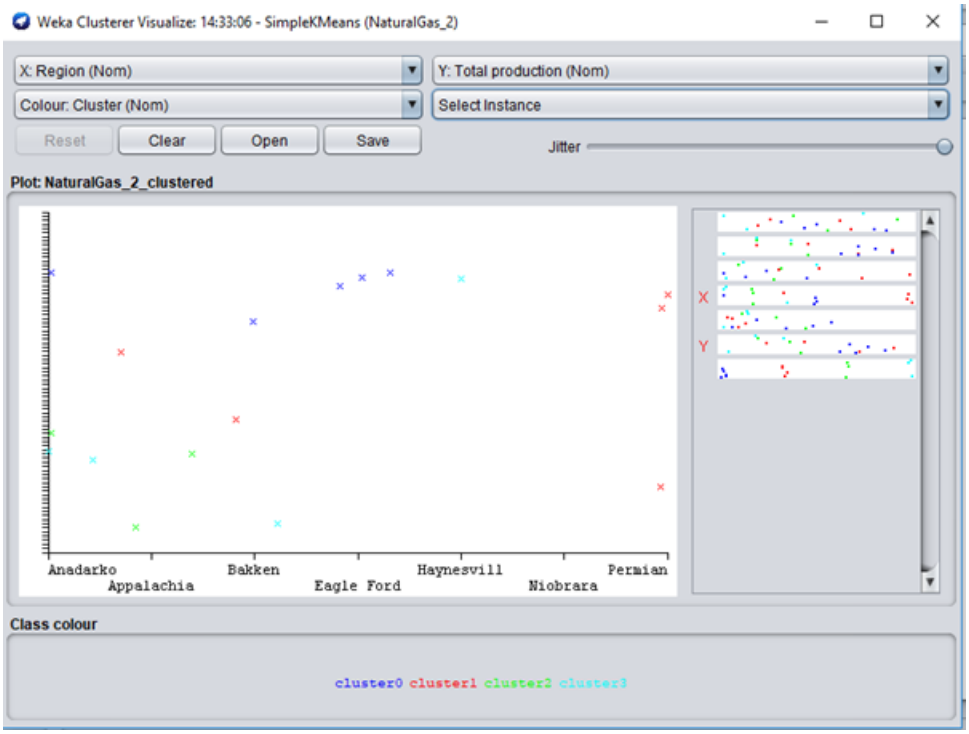
Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region

### Model: Simple K Means(Clustering)

Properties: Euclidean distance, Max iterations = 200, No of clusters = 4, Initialization method = Farthest first, Percentage split = 80

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 97.74979572692916

Initial starting points (farthest first):

Cluster 0: 2015,51,Niobrara,1955.400695,'45,25,628'
Cluster 1: 2014,537,Permian,429.1254,'56,69,913'
Cluster 2: 2018,76,Appalachia,14868.5714,'2,65,35,545'
Cluster 3: 2007,182,Haynesville,1049.962444,'37,79,979'

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute               Full Data           0           1           2           3
                           (67.0)      (21.0)      (14.0)       (9.0)      (23.0)
================================================================================
Month                   2012.3731   2014.1905   2012.3571   2013.1111   2010.4348
Rig count                140.5373     77.3333    319.3571     82.7778         112
Region                   Niobrara    Niobrara     Permian  Appalachia Haynesville
Production per rig      2741.0656    2165.215    826.2238   7604.9753   2529.1289
Total production        41,65,976   36,68,958   47,65,801   13,85,236   41,65,976



Time taken to build model (percentage split) : 0.01 seconds

Clustered Instances

0       5 ( 29%)
1       5 ( 29%)
2       3 ( 18%)
3       4 ( 24%)
```




The above clustering pattern shows the region on the x-axis and the total production on y-axis. It shows the split of the 4 clusters over various regions in the US. It is inconclusive to figure a pattern for the total production of natural gas over different regions since the data set consists of data points not distinct.

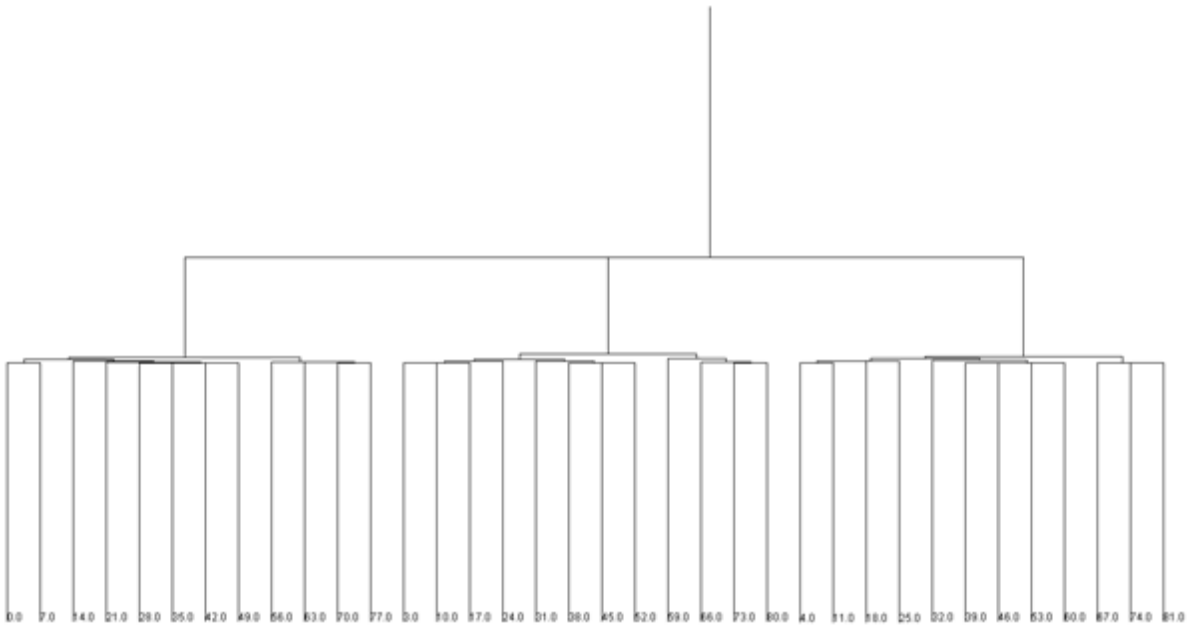
<u>Model details applied on the dataset using WEKA software:</u>

Features: Month, Rig count, Production per rig, Total production, Region

**Model: Hierarchical Clustering**

Properties: Euclidean Distance, No of clusters = 5

```
Clustered Instances

0       4 ( 15%)
1       7 ( 27%)
2       1 (  4%)
3      10 ( 38%)
4       4 ( 15%)
```



As shown in the dendrogram above, the clusters are split on the basis of the total production per rig and the clusters are grouped closer to each other which shows the similarity between the data points. The clusters shown 3 major split which could be classified as low production, medium production and high production regions.

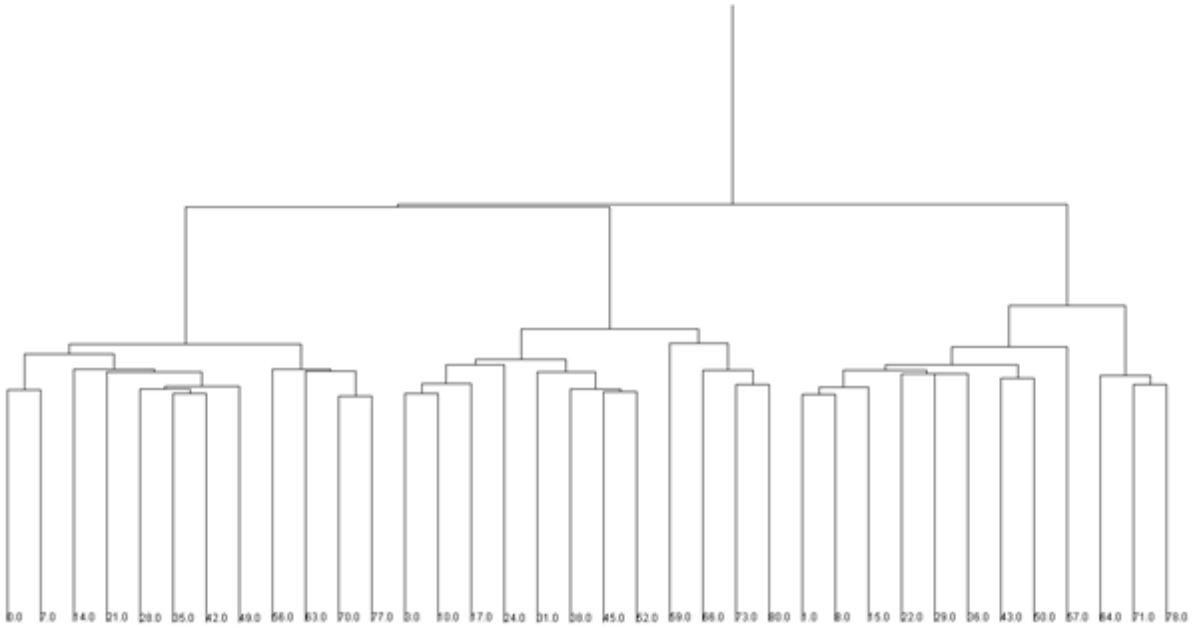
Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region

**Model: Hierarchical Clustering**

Properties: Manhattan Distance, No of clusters = 5

```
Clustered Instances

0       4 ( 15%)
1       7 ( 27%)
2       1 (  4%)
3      10 ( 38%)
4       4 ( 15%)
```



Unlike the Euclidean distance clustering of the model, the dendrogram developed by taking the mean as Manhattan distance, the clustering is more distinct and shows the dissimilarity between the points. It is always suited to apply Euclidean distance metric to calculate the k-clusters because the distortion in k-means using Manhattan distance metric is less than that of k-means using Euclidean distance metric.

<u>Model details applied on the dataset using WEKA software:</u>

Features: Month, Rig count, Production per rig, Total production
Response: Region

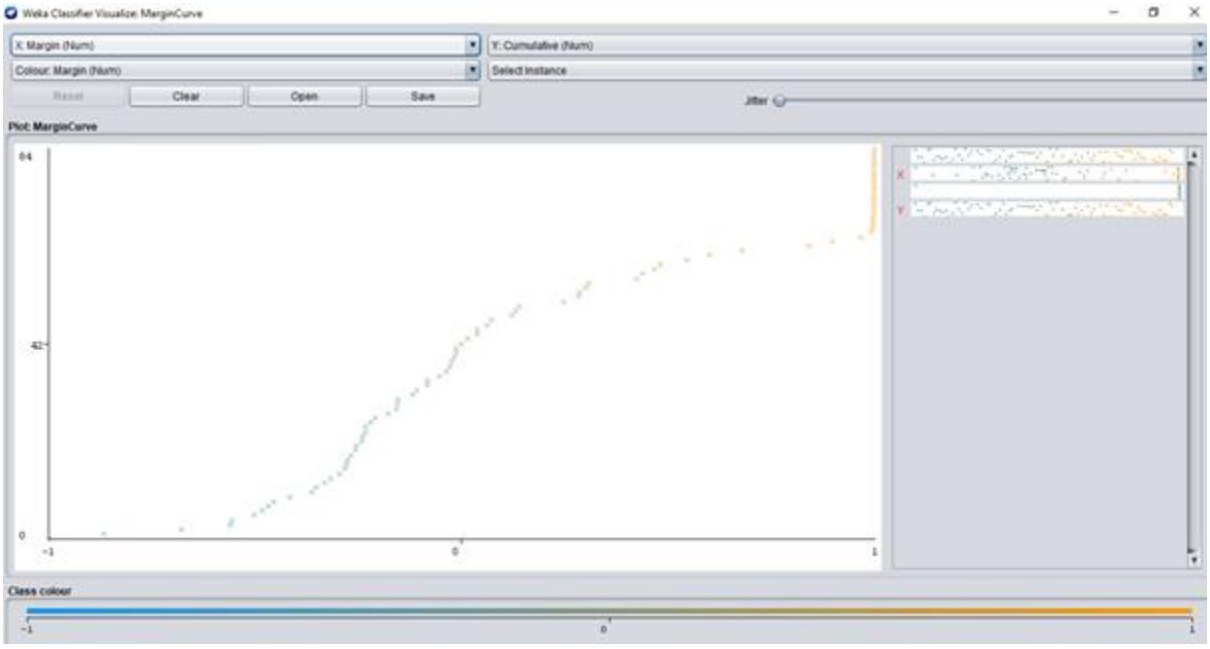
**Model: Logistic Regression (Classification)**

10-fold cross-validation

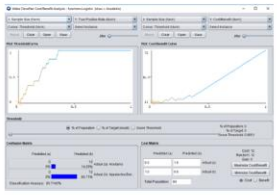
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          43               51.1905 %
Incorrectly Classified Instances        41               48.8095 %
Kappa statistic                          0.4306
Mean absolute error                      0.14
Root mean squared error                  0.277
Relative absolute error                 57.0639 %
Root relative squared error             78.9971 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.417    0.125    0.357      0.417    0.385      0.274    0.876     0.447     Anadarko
                 0.333    0.139    0.286      0.333    0.308      0.183    0.813     0.495     Appalachia
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Bakken
                 0.250    0.125    0.250      0.250    0.250      0.125    0.770     0.431     Eagle Ford
                 0.417    0.056    0.556      0.417    0.476      0.409    0.918     0.604     Haynesville
                 0.333    0.125    0.308      0.333    0.320      0.202    0.846     0.438     Niobrara
                 0.833    0.000    1.000      0.833    0.909      0.900    0.948     0.913     Permian
Weighted Avg.    0.512    0.081    0.537      0.512    0.521      0.442    0.881     0.618

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  5  0  0  5  0  2  0 |  a = Anadarko
  0  4  0  1  3  4  0 |  b = Appalachia
  0  0 12  0  0  0  0 |  c = Bakken
  5  1  0  3  1  2  0 |  d = Eagle Ford
  0  5  0  1  5  1  0 |  e = Haynesville
  2  4  0  2  0  4  0 |  f = Niobrara
  2  0  0  0  0  0 10 |  g = Permian
```

The accuracy scores of the classification model shows that the regions classified correctly is 51%. The results suggest that logistic regression did not perform well to classify the regions which were 7 in number. Logistic regression runs with better results if the response was a binary variable rather than have 7 different values. As explained earlier, Logistic regression tends to underperform when there are multiple or non-linear decision boundaries as defined in this case for the response variable region.

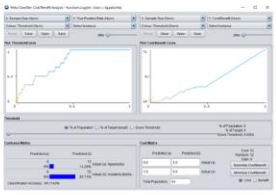
According to the confusion matrix, we can analyze that Bakken region was predicted accurately with respect to the true value with precision and ROC values being 1. The next best region classification prediction was for the region Permian.



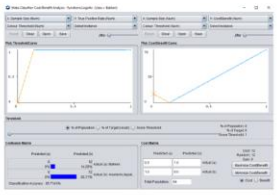

Anadarko



Appalachia

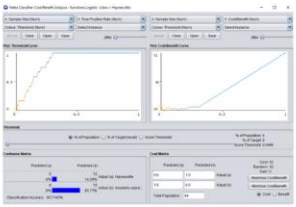


Bakken

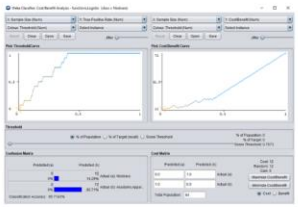


Eagle Ford

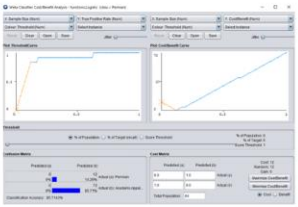
Haynesville



Niobrara



Permian



The above graphs show the performance of the model and cost and benefits calculations for all the regions in the US.

Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production

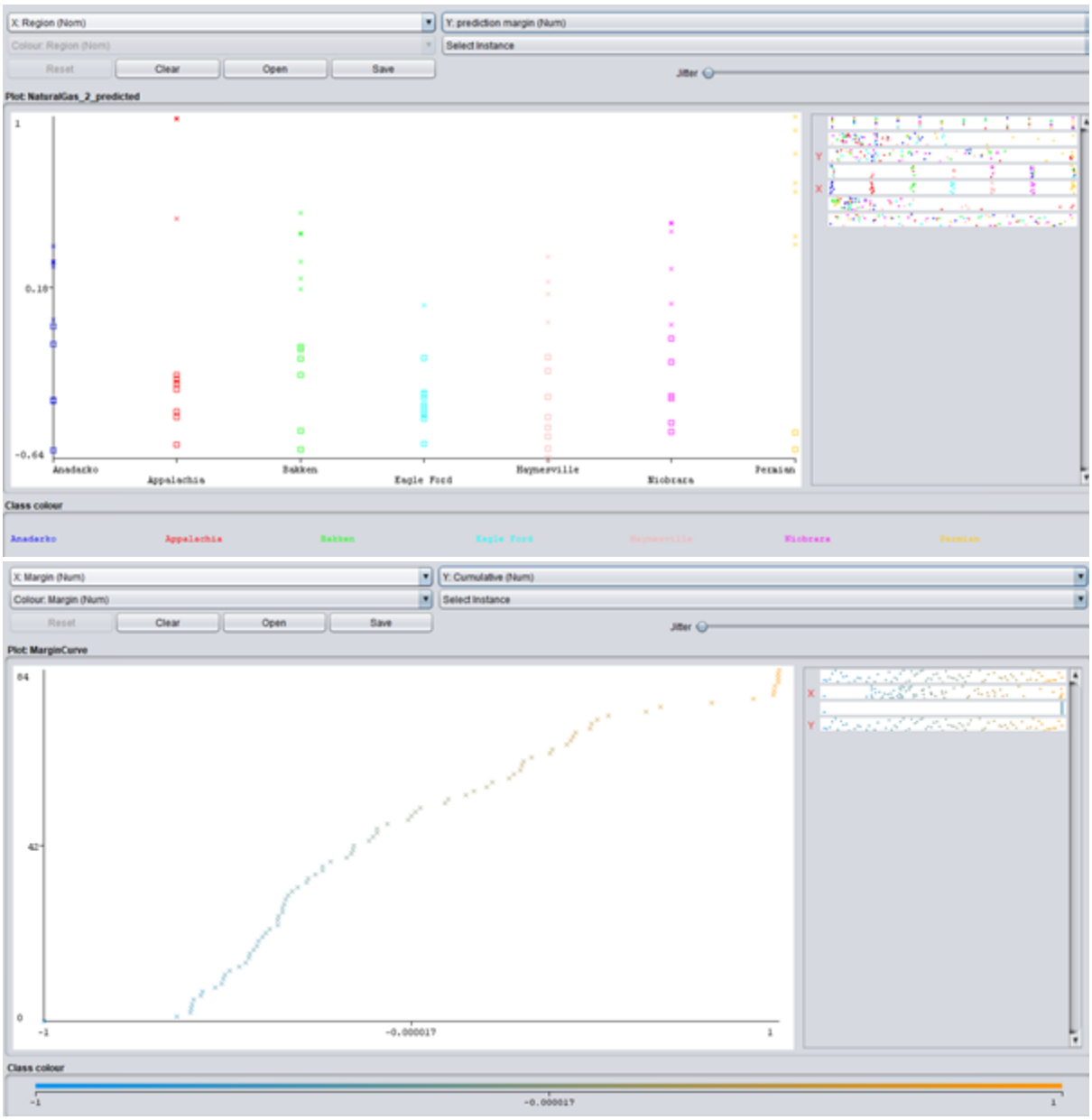
Response: Region

**Model: Naive Bayes(Classification)**

10-fold cross-validation

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          36               42.8571 %
Incorrectly Classified Instances        48               57.1429 %
Kappa statistic                          0.3333
Mean absolute error                      0.1849
Root mean squared error                  0.3215
Relative absolute error                 75.3562 %
Root relative squared error             91.7091 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.417    0.181    0.278      0.417    0.333      0.201    0.745     0.271     Anadarko
                 0.333    0.056    0.500      0.333    0.400      0.331    0.758     0.538     Appalachia
                 0.500    0.056    0.600      0.500    0.545      0.480    0.858     0.486     Bakken
                 0.083    0.028    0.333      0.083    0.133      0.105    0.444     0.153     Eagle Ford
                 0.333    0.083    0.400      0.333    0.364      0.270    0.778     0.372     Haynesville
                 0.500    0.167    0.333      0.500    0.400      0.284    0.777     0.376     Niobrara
                 0.833    0.097    0.588      0.833    0.690      0.641    0.955     0.889     Permian
Weighted Avg.    0.429    0.095    0.433      0.429    0.409      0.330    0.759     0.441

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  5  0  0  0  0  4  3 |  a = Anadarko
  2  4  3  1  2  0  0 |  b = Appalachia
  0  0  6  0  0  3  3 |  c = Bakken
  4  1  0  1  1  4  1 |  d = Eagle Ford
  3  3  0  1  4  1  0 |  e = Haynesville
  3  0  0  0  3  6  0 |  f = Niobrara
  1  0  1  0  0  0 10 |  g = Permian
```

The Naive Bayes classifier too gives a lower accuracy of 42.8% since the dataset is small and the features are correlated with similar linear differences. If there were to categorical values, logistic regression and Naive Bayes models would underperform and give bad results for the classification problem definition.




Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production

Response: Region

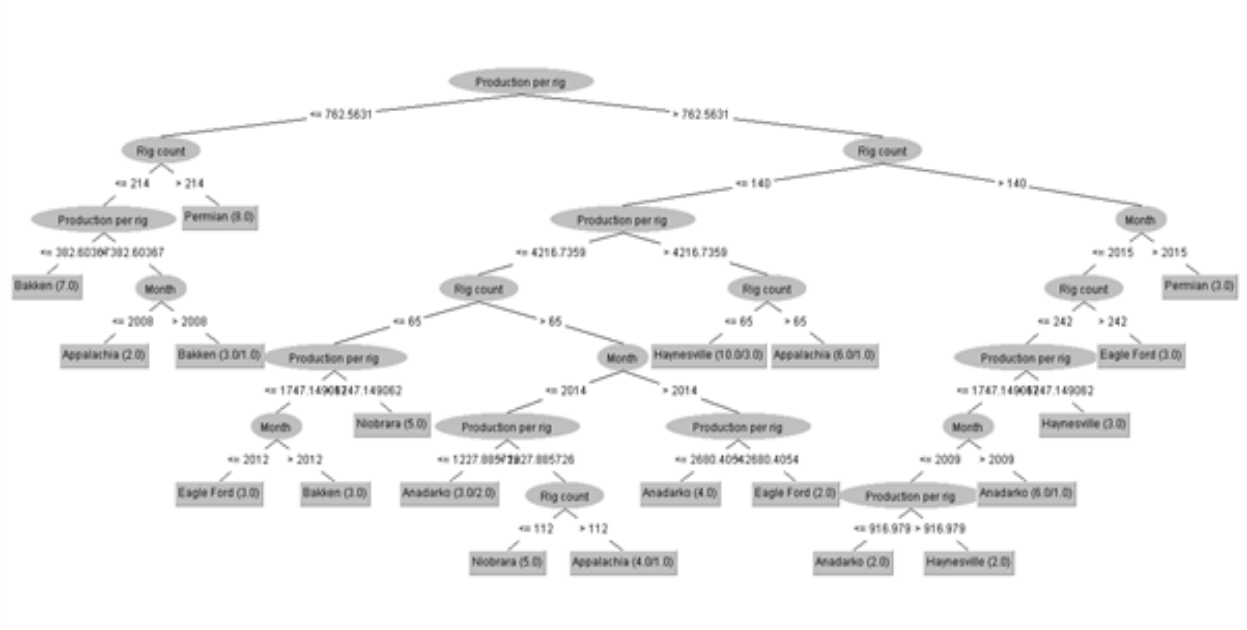
**Model: Decision Tree (Classification)**

Filter: J48

Properties: confidenceFactor = 0.25, numFolds = 3, unpruned = false

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          45               53.5714 %
Incorrectly Classified Instances        39               46.4286 %
Kappa statistic                          0.4583
Mean absolute error                      0.1404
Root mean squared error                  0.328
Relative absolute error                 57.2091 %
Root relative squared error             93.5484 %
Total Number of Instances               84

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.500    0.125     0.400      0.500     0.444     0.343     0.720     0.382     Anadarko
                 0.417    0.097     0.417      0.417     0.417     0.319     0.713     0.485     Appalachia
                 0.750    0.069     0.643      0.750     0.692     0.639     0.915     0.609     Bakken
                 0.250    0.125     0.250      0.250     0.250     0.125     0.742     0.281     Eagle Ford
                 0.417    0.083     0.455      0.417     0.435     0.346     0.706     0.300     Haynesville
                 0.583    0.042     0.700      0.583     0.636     0.585     0.798     0.492     Niobrara
                 0.833    0.000     1.000      0.833     0.909     0.900     0.909     0.857     Permian
Weighted Avg.    0.536    0.077     0.552      0.536     0.541     0.465     0.786     0.487

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  6  1  1  3  1  0  0 |  a = Anadarko
  2  5  2  0  2  1  0 |  b = Appalachia
  0  2  9  0  0  1  0 |  c = Bakken
  5  1  1  3  2  0  0 |  d = Eagle Ford
  1  2  0  3  5  1  0 |  e = Haynesville
  0  1  0  3  1  7  0 |  f = Niobrara
  1  0  1  0  0  0 10 |  g = Permian
```

The decision tree model did not perform well with the dataset since the the region had 7 classifiers. The accuracy of the model is 53.6% which is better than the other classification models on a numeric dataset. Decision tree performs better than the rest models on numeric dataset and the results are seen above.



Shown above is the decision tree classification model, whose root node is production per rig. From this model we can find a close relation between the response variable region and the predictor variable production per rig. The feature variable production per rig carries weightage and information per region which is significant compared to the other feature variables. The second-best node is the rig count in every region which explains the connection between regions and production rate at the second level of the tree.

Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region

**Model: Apriori Rule Association**

Properties: No of Rules = 5, Min Metric = 0.9

```
Apriori
=======

Minimum support: 0.1 (8 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 22

Size of set of large itemsets L(3): 5

Best rules found:

1. Rig count='(149.25-278.5]' 23 ==> Production per rig='(-inf-3790.503828]' 23    <conf:(1)> lift:(1.29) lev:(0.06) [5] conv:(5.2)
2. Month='(-inf-2009.75]' 21 ==> Production per rig='(-inf-3790.503828]' 21    <conf:(1)> lift:(1.29) lev:(0.06) [4] conv:(4.75)
3. Month='(-inf-2009.75]' Rig count='(-inf-149.25]' 14 ==> Production per rig='(-inf-3790.503828]' 14    <conf:(1)> lift:(1.29) lev:(0.04) [3] conv:(3.17)
4. Region=Appalachia 12 ==> Rig count='(-inf-149.25]' 12    <conf:(1)> lift:(1.53) lev:(0.05) [4] conv:(4.14)
5. Region=Niobrara 12 ==> Rig count='(-inf-149.25]' 12    <conf:(1)> lift:(1.53) lev:(0.05) [4] conv:(4.14)
```

The association rule when restricted to 5 rules gives not significant rules as expected. The most prominent rule that can be taken from the model is that if the year is prior to 2009 and the rig count is below 150, then there was always a production rate of less than 3780 per rig. This shows the lack of technology and resources back in time and shows the growth of production of natural gas over the recent times.

# Modeling Relationship between Natural Gas Prices and US Petroleum Production

### Simple K Means on Natural Gas vs. Price

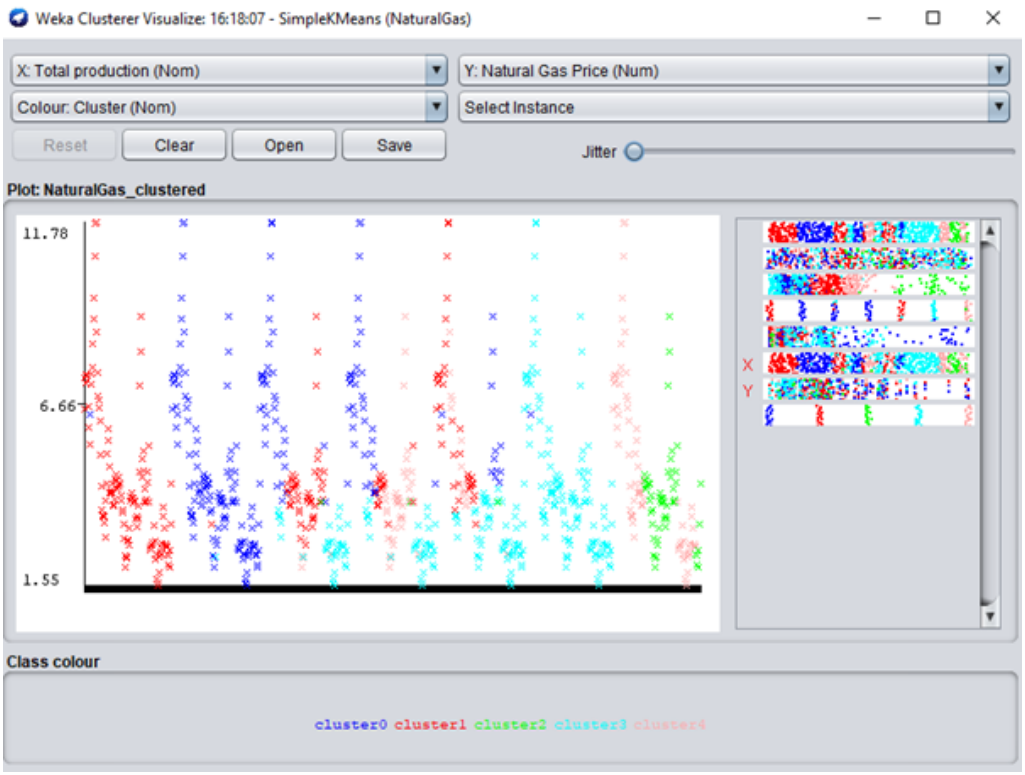
Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region, Natural Gas Price

**Model: Simple K Means(Clustering)**

Properties: Manhattan distance, Max iterations = 200, No of clusters = 5

```
Number of iterations: 6
Sum of within cluster distances: 2506.8350406373097

Initial starting points (random):

Cluster 0: Oct-07,58,Appalachia,477.893621,'14,55,292',6.35
Cluster 1: Jan-12,216,Anadarko,1021.36519,'46,20,620',3.27
Cluster 2: Oct-14,559,Permian,433.145502,'59,85,170',3.87
Cluster 3: Jun-12,83,Niobrara,1587.7337,'46,62,409',2.35
Cluster 4: Nov-15,227,Permian,846.16313,'68,60,518',2.4

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute             Full Data         0          1          2          3          4
                        (931.0)   (219.0)    (235.0)     (60.0)    (278.0)    (139.0)
=============================================================================================
Month                    Jan-07     Oct-07     Jan-12     Oct-14     Jun-12     Nov-15
Rig count                   114         68        178        468         51        244
Region                 Anadarko Appalachia   Anadarko    Permian   Niobrara    Permian
Production per rig    1313.2179  1374.8728  1079.4109   357.5224  2261.5786  1072.5212
Total production      40,31,235  42,44,042  40,31,235  14,62,663  77,15,730  57,88,449
Natural Gas Price          3.94       4.75       4.24       3.91       3.17       3.96



Time taken to build model (full training data) : 0.15 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      219 ( 24%)
1      235 ( 25%)
2       60 (  6%)
3      278 ( 30%)
4      139 ( 15%)
```




**Linear Regression on Natural Gas vs. Price**

Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region

Response: Natural Gas Price

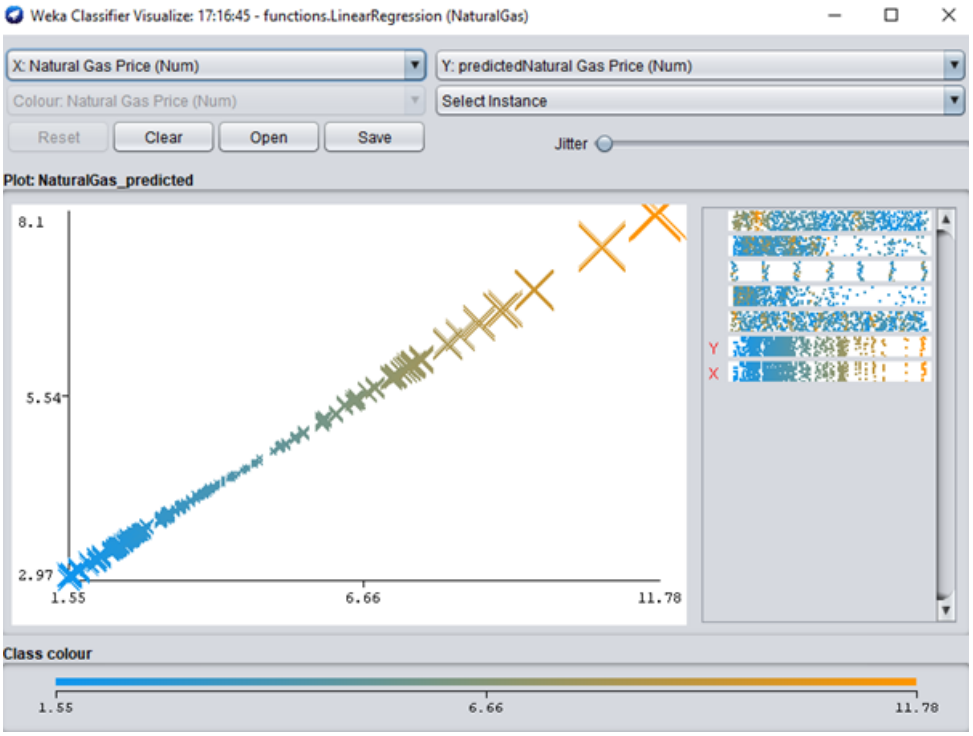
**Model: Linear Regression(Classification)**

Properties: Selection Process = M5 method

```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9998
Mean absolute error                      0.7905
Root mean squared error                  1.047
Relative absolute error                 50.7477 %
Root relative squared error             50.4609 %
Total Number of Instances              931
```

The accuracy of the model is at 50%, slightly better than the logistic regression model since the dataset consisted of numeric data. The correlation coefficient proves the collinearity between

the response and the predictor variables. With not much variance in the dataset and low record count, the model is underfit the datapoints.




**Random Forest on Natural Gas vs. Price**

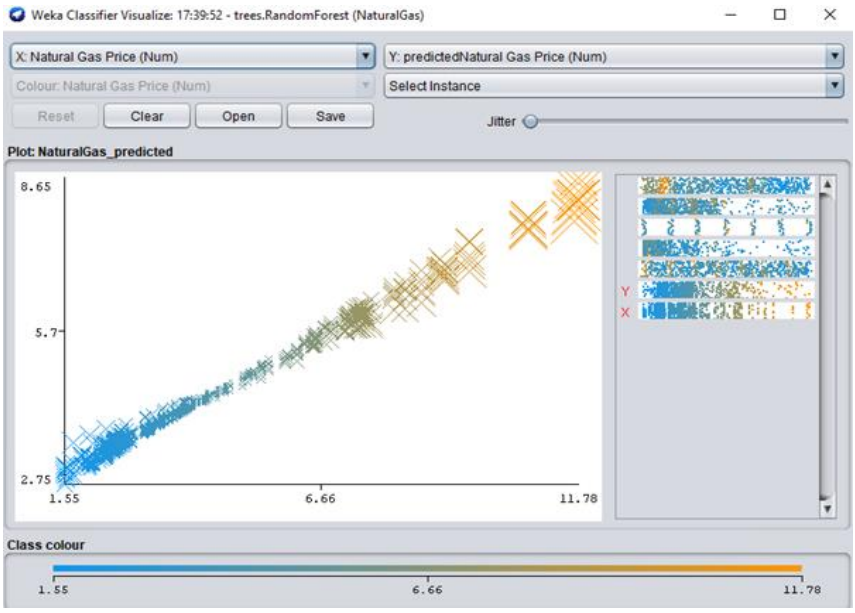
<u>Model details applied on the dataset using WEKA software:</u>

Features: Month, Rig count, Production per rig, Total production, Region

Response: Natural Gas Price

**Model: Random Forest (Classification)**

Properties: Attribute importance is true, Batch and bagging size = 100



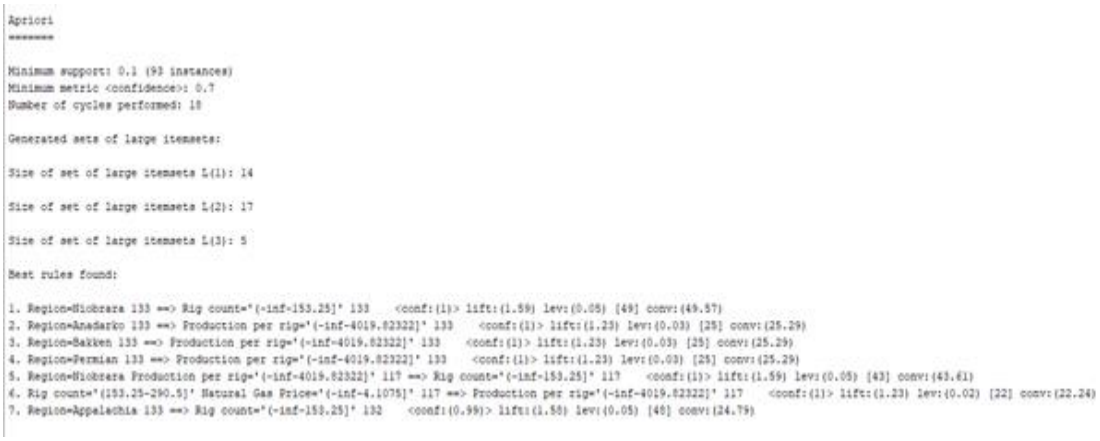
**Apriori Rule Association on Natural Gas vs. Price-**

Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region, Natural Gas Price

Model: Apriori(Association)

Properties: No of Rules = 7, Metric Type = confidence, Min metric = 0.7



```
Apriori
=======

Minimum support: 0.1 (93 instances)
Minimum metric <confidence>: 0.7
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 17

Size of set of large itemsets L(3): 5

Best rules found:

1. Region=Niobrara 133 ==> Rig count='(-inf-153.25]' 133    <conf:(1)> lift:(1.59) lev:(0.05) [49] conv:(49.57)
2. Region=Anadarko 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
3. Region=Bakken 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
4. Region=Permian 133 ==> Production per rig='(-inf-4019.82322]' 133    <conf:(1)> lift:(1.23) lev:(0.03) [25] conv:(25.29)
5. Region=Niobrara Production per rig='(-inf-4019.82322]' 117 ==> Rig count='(-inf-153.25]' 117    <conf:(1)> lift:(1.59) lev:(0.05) [43] conv:(43.61)
6. Rig count='(153.25-290.5]' Natural Gas Price='(-inf-4.1075]' 117 ==> Production per rig='(-inf-4019.82322]' 117    <conf:(1)> lift:(1.23) lev:(0.02) [22] conv:(22.24)
7. Region=Appalachia 133 ==> Rig count='(-inf-153.25]' 132    <conf:(0.99)> lift:(1.58) lev:(0.05) [40] conv:(24.79)
```

Some significant association rules are if the region is Niobrara and the production rate is less than 4020 per rig, the rig count is always less than 150 in the region. Also, if the rig count is between 150 - 290 and the natural gas price is less than 4, the production rate is below 4020 in all the regions.

## Modeling Relationship between Oil Prices and US Petroleum Production

**Simple K Means on Oil vs. Price**

Model details applied on the dataset using WEKA software:

Features: Month, Rig count, Production per rig, Total production, Region, Oil Price

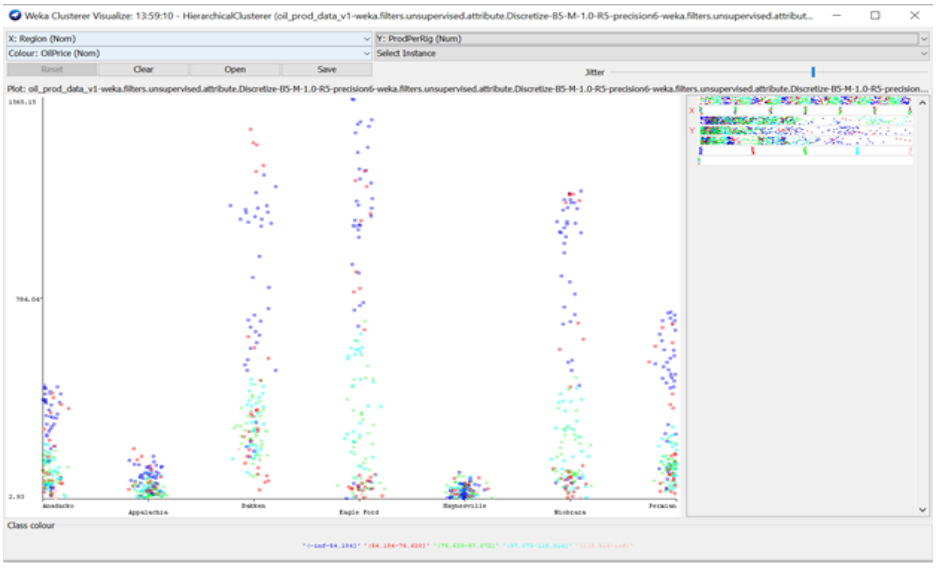
**Model: Simple K Means(Clustering)**

Properties: Manhattan distance

Max iterations = 200

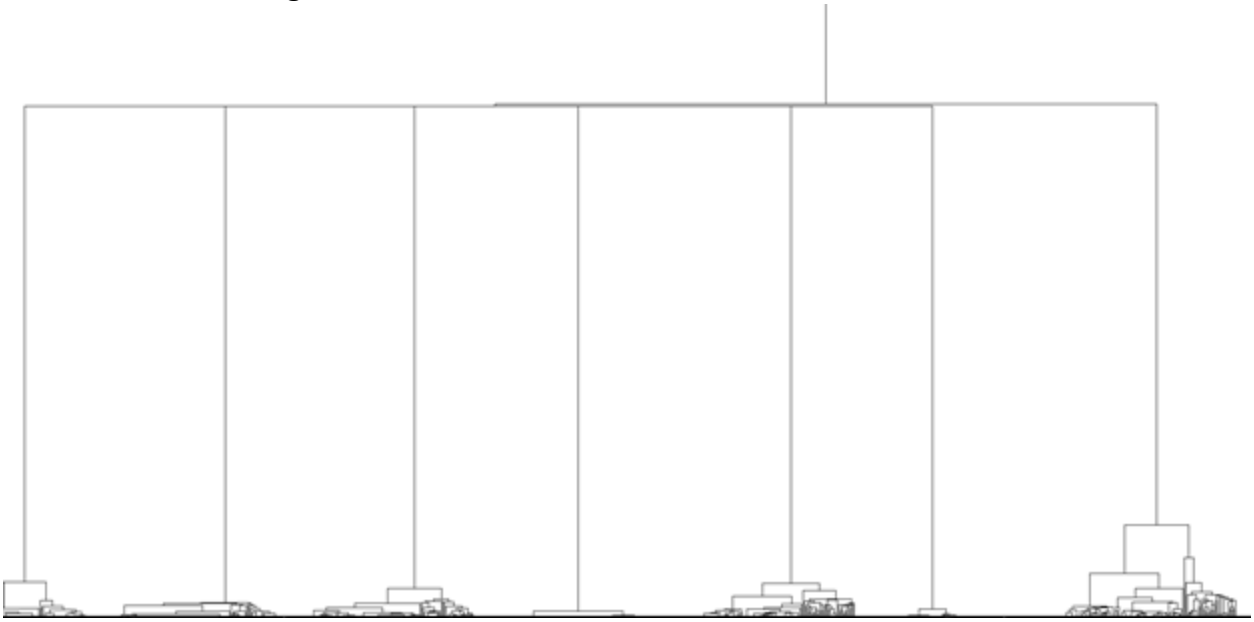
No of clusters = 5

```
kMeans
======

Number of iterations: 19
Within cluster sum of squared errors: 314.48327203969507

Initial starting points (random):

Cluster 0: Haynesville,50,23.77,42056.12
Cluster 1: Haynesville,175,7.92,62953.52
Cluster 2: Appalachia,61,11.26,19591.07
Cluster 3: Appalachia,62,12.42,21221.45
Cluster 4: Anadarko,99,223.4,463436.19

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute          Full Data          0          1          2          3          4
                     (938.0)    (160.0)    (164.0)    (229.0)    (211.0)    (174.0)
=============================================================================================
Region              Anadarko  Haynesville     Permian   Appalachia     Bakken     Anadarko
Rig Count          143.2516       107.8   317.6159      80.4279    90.5545     158.092
ProdPerRig         254.7801      19.5489   291.0785        63.44   670.0337    185.1391
TotProd          491934.0196   64555.6636 1356851.8168  81456.7725 751835.4679 294773.6874



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       160 ( 17%)
1       164 ( 17%)
2       229 ( 24%)
3       211 ( 22%)
4       174 ( 19%)


Class attribute: OilPrice
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 37 38 45 96 36 | '(-inf-54.184]'
 37 28 53 44 27 | '(54.184-75.628]'
 46 52 79 39 57 | '(75.628-97.072]'
 34 43 46 29 51 | '(97.072-118.516]'
  6  3  6  3  3 | '(118.516-inf)'

Cluster 0 <-- '(54.184-75.628]'
Cluster 1 <-- '(118.516-inf)'
Cluster 2 <-- '(75.628-97.072]'
Cluster 3 <-- '(-inf-54.184]'
Cluster 4 <-- '(97.072-118.516]'

Incorrectly clustered instances :       672.0    71.6418 %
```



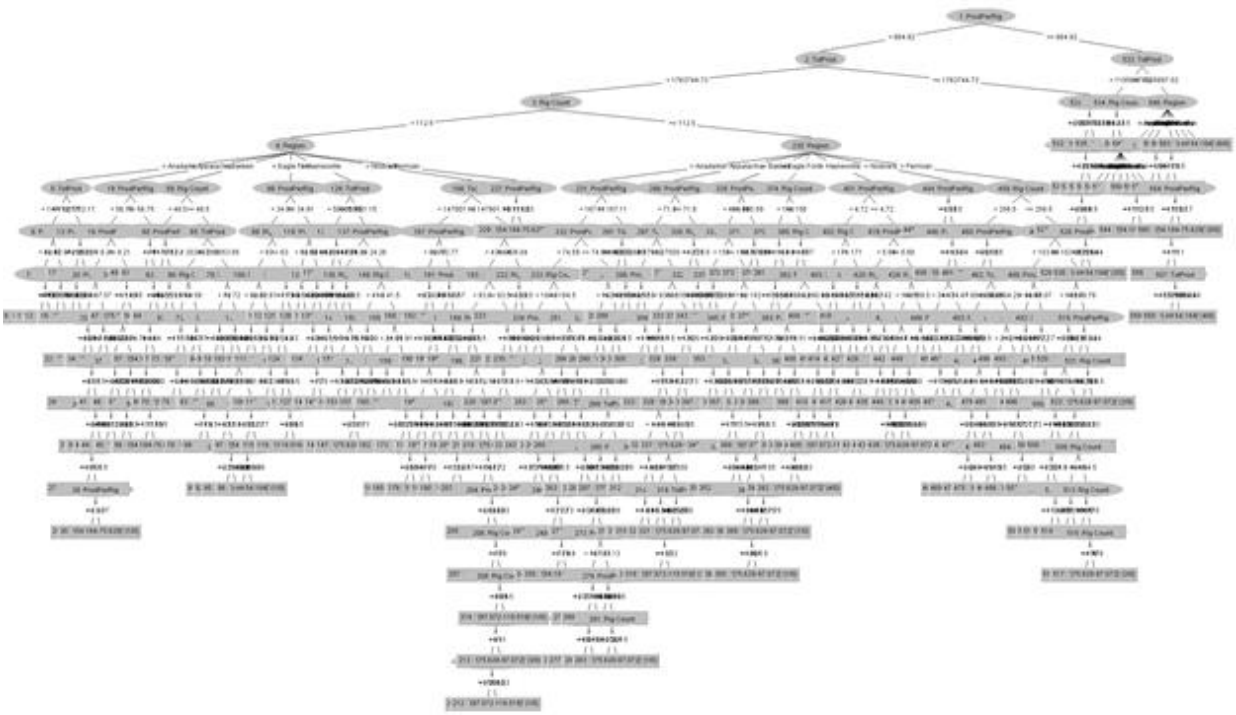

K-means clustering on the oil dataset showed that the data points are best clustered on region - all other features seem to be correlated within this space.

**Hierarchical Clustering on Oil vs Price**

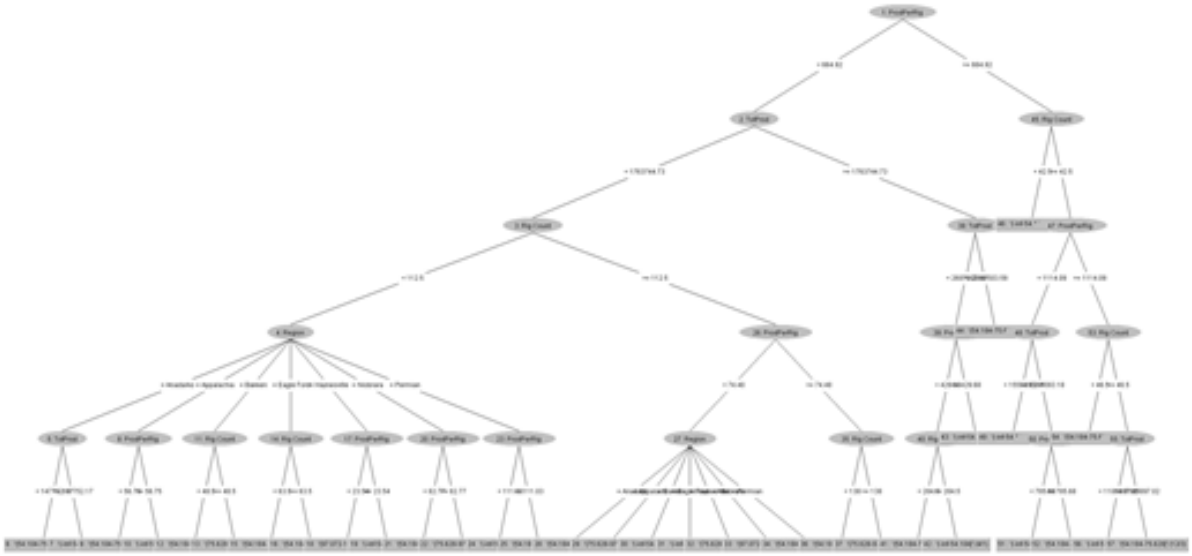


With hierarchical clustering, the results seemed to show a similar pattern where the regions were clustered together.

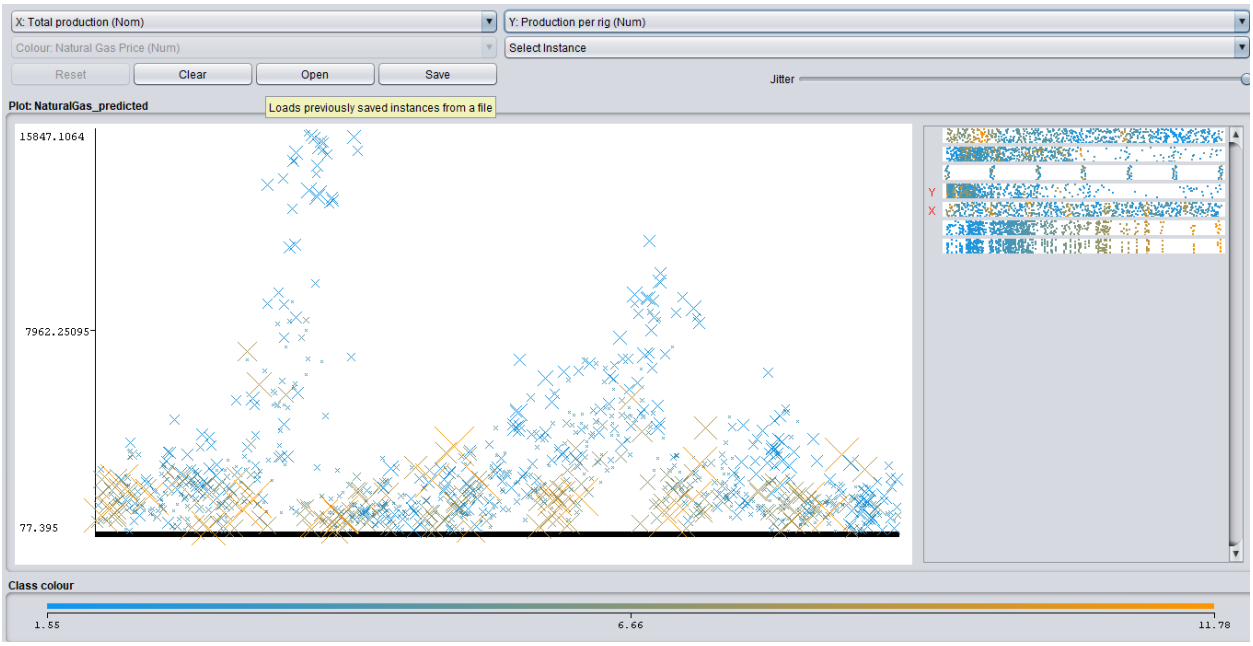
**Decision Tree on Oil vs Price**
**Unpruned tree**



**Pruned tree**



Production per rig was chosen as the root node when classifying price based on the features in the dataset.




In the figure above, it can be seen that total production of oil did not impact prices much (as the class varies uniformly across different values on the x axis), whereas the production per rig seems to define the class. Low value (on the y axis) coincides with a higher oil price, and a higher production per rig led to lower oil prices.

## Key Findings and Conclusion

The features in our dataset enabled us to get in an insight into the fluctuations in price for oil and gas. They also presented interesting patterns between the production data that helped us discover relationships. As soon as we began our analysis, It was clear that the rig counts are highly correlated with the region. This makes sense because the number of rigs should grow over time at the speed of each region's development - based on petroleum projects in the area. It allowed to learn about the similarities and differences in production cycles of each region.
In hindsight, this seems obvious but the data displayed remarkable detail into the phases of each region. During this analysis, we also saw that the rig counts and the natural gas prices was a good predictor of production per rig. Although, this association was not our focus, this provided insight into the dynamics between project funding and prices. It was also interesting to note that the production per rig - something in control of the companies profiting from these sites - acted as a lever between prices and the number of rigs. This indicated that production per rig was the best predictor of price - for both oil and gas. This was then confirmed through the decision tree where the production per rig was chosen as the root node.
This means that the production per rig is a variable that can be charted to gage the future price fluctuation of oil and gas. Although, not a definitive correlation, it should provide a picture of the relative performance of the US petroleum industry.

## Modelling Takeaways on Petroleum Production data

Most of our models performed better when a number of features were clustered, based on the region. This allowed us to analyze the impact of different production statistics on the price over time, but because each region's characteristics were unique we were able to construct and learn from the historical patterns of oil and gas.
Pruning the decision tree was also another vital technique that allowed us to increase the model's interpretability and understand each statistic feature importance. The initial implementation of the decision, although more accurate in classification rate, had no limitation on depth and therefore the large number of branches made it difficult to decorrelate the variables.
In general, the data models seemed to have a higher classification rate when prices were binned as low. This simply might be because most of the data was observed when price were not as high as peak years. This can be a bias introduced due to the dataset used.
The team does want to emphasize that the we do not want to over-generalize our findings. The results above were specifically on the petroleum data obtained.

## Challenges

Data collection and preprocessing accounted for most of the project timeline. Because the information we wanted to explore was contained in different files - using different formats, most of the first phase of the project was spent preparing the dataset to be used in WEKA. We had to correct the normalization in the databases and make sure data structures were consistent with the software.

## Recommendations & Further Research

In the future, the team would like to explore the data individually for each region. The team found it challenging to isolate variance from such a macroscopic view of the petroleum industry in the United States. Some areas of interest to continue further research would be the impact of oil quality on production. We are also curious to perform a similar analysis on another part of the world and compare the results obtained. Lastly, we suggest looking at energy consumption and its impact on petroleum production.